Notebook

November 20, 2024

1 Election Analysis

Names: Shiyuan Zhang, Shuoyuan Gao

1.1 1. Introduction and Dataset Research

1. Discovering hidden patterns: For example, in the 2016 U.S. election dataset may contain multidimensional data such as voter's voting behavior, socio-economic characteristics, and geographic information. Unsupervised learning, clustering as we using in below, can help identify underlying patterns between different groups of voters, such as voters with similar voting behavior or voters with similar socioeconomic characteristics. Then as we found the patterns, we can understand better for the voter trends. This can tell us more deeper analysis and predictions of election outcomes and voter preferences."

2. Discovering outliers: In the 2016 U.S. election dataset, there may be some unusual voting patterns like precincts with unusually high or low turnout. We can identified the outliers using algorithms. This can letting people know more about the precinct characteristics. In Gaussian Mixture Models, "the GMM from sklearn calculates the score of an observation based on the density of each point's location in that space. Thus, points in higher density regions are less likely to be outliers, and vice versa."

3. Data exploration and visualization: unsupervised learning can also be used as a data exploration and visualization tool to help researchers better understand the structure and distribution of data in the 2016 U.S. election dataset. For example, cluster analysis allows for the division of voters into different clusters and the generation of visualization charts to show the relationships between these clusters. This helps researchers better understand voter behavior and attitudes, identify potential voting patterns or trends, and extract interesting insights from the dataset.

Motivation: We want to use the voting data to determine if different counties are clustered according to the state they are in. We also want to explore the distribution of candidates supported in each cluster. The specific method is: we take the information of the real state where the county is located as the pre-defined label, and then check whether the predicted clusters are related to our pre-defined Location_State by different clustering algorithms. Then we analyze each cluster in detail to explore the components with high candidate support.

Citation:

1.Delua, Julianna. "Supervised vs. Unsupervised Learning: What's the Difference?" IBM, 12 Mar. 2021, https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

2.Santos, Gustavo. "Using Unsupervised Learning to Find Outliers." Medium, Towards Data Science, 2 Nov. 2022, https://towardsdatascience.com/using-unsupervised-learning-to-find-outliers-670e07396599.

3.Verbeeck, Nico et al. "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry." Mass spectrometry reviews vol. 39,3 (2020): 245-291. doi:10.1002/mas.21602

The dataset is from Professor provided. Background is from 2016 USA president election, it contains the Democrat candidates and Republicans candidates counts.

voting information include all of the U.S. counties in the 2016 U.S. presidential election. For each county, the percentage and number of votes that went to each primary presidential candidate is listed.

[1]:
```
pip install sklearn
```

```
Requirement already satisfied: sklearn in
/Users/gj/miniconda3/lib/python3.10/site-packages (0.0.post4)

[notice] A new release of pip is
available: 23.0.1 -> 23.1.2
[notice] To update, run:
pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

1.2 2. Data Cleaning and Data Manipulation

[3]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

from sklearn.preprocessing import MinMaxScaler
from sklearn.manifold import TSNE
from sklearn.metrics import adjusted_rand_score
from sklearn.datasets import load_digits
from sklearn.cluster import KMeans
from sklearn.metrics import adjusted_rand_score
from sklearn.datasets import make_blobs
from sklearn.mixture import GaussianMixture
from sklearn.cluster import KMeans

from matplotlib.patches import Ellipse

from sklearn.metrics import silhouette_samples, silhouette_score

from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.manifold import TSNE
```

```
from sklearn.cluster import AgglomerativeClustering
from scipy.spatial.distance import pdist
from scipy.cluster.hierarchy import linkage, dendrogram, cophenet
from sklearn.preprocessing import StandardScaler

from scipy.spatial.distance import pdist, squareform
from sklearn.neighbors import NearestNeighbors
from sklearn.cluster import DBSCAN
from pyclustertend import hopkins
from sklearn.metrics import silhouette_score, calinski_harabasz_score
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

[126]:
```
#read data
df = pd.read_csv('election.csv')
df = df.drop('Location_State_Abbreviation',axis =1)
df
```

[126]:

	Location_County	Location_State	Vote_Data_Ben_Carson_Number_of_Votes	\
0	Abbeville	South Carolina	305	
1	Abbot	Maine	0	
2	Abington	Massachusetts	53	
3	Acadia	Louisiana	0	
4	Accomack	Virginia	411	
…	…	…	…	
4211	Yuma	Arizona	0	
4212	Yuma	Colorado	0	
4213	Zapata	Texas	4	
4214	Zavala	Texas	0	
4215	Ziebach	South Dakota	0	

	Vote_Data_Ben_Carson_Party	Vote_Data_Ben_Carson_Percent_of_Votes	\
0	Republican	8.3	
1	Republican	0.0	
2	Republican	2.5	
3	Republican	0.0	
4	Republican	9.5	
…	…	…	
4211	Republican	0.0	
4212	Republican	0.0	
4213	Republican	4.7	
4214	Republican	0.0	
4215	Republican	0.0	

	Vote_Data_Bernie_Sanders_Number_of_Votes	Vote_Data_Bernie_Sanders_Party	\
0	312	Democrat	
1	1	Democrat	

```
2                                          1352                        Democrat
3                                          1087                        Democrat
4                                           682                        Democrat
…                                           …                           …
4211                                       2156                        Democrat
4212                                         33                        Democrat
4213                                        685                        Democrat
4214                                        373                        Democrat
4215                                        132                        Democrat

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                          17.0
1                                         100.0
2                                          53.4
3                                          31.4
4                                          27.4
…                                           …
4211                                       31.5
4212                                       39.3
4213                                       23.6
4214                                       18.1
4215                                       58.9

      Vote_Data_Carly_Fiorina_Number_of_Votes Vote_Data_Carly_Fiorina_Party  \
0                                           0                     Republican
1                                           0                     Republican
2                                           0                     Republican
3                                           0                     Republican
4                                           0                     Republican
…                                           …                        …
4211                                        0                     Republican
4212                                        0                     Republican
4213                                        0                     Republican
4214                                        0                     Republican
4215                                        0                     Republican

      …  Vote_Data_Rand_Paul_Percent_of_Votes  \
0     …                                   0.0
1     …                                   0.0
2     …                                   0.0
3     …                                   0.0
4     …                                   0.0
…   …                                     …
4211  …                                   0.0
4212  …                                   0.0
4213  …                                   0.0
4214  …                                   0.0
```

```
4215  …                                    0.0

      Vote_Data_Rick_Santorum_Number_of_Votes Vote_Data_Rick_Santorum_Party  \
0                                          0                     Republican
1                                          0                     Republican
2                                          0                     Republican
3                                          0                     Republican
4                                          0                     Republican
…                                        …                          …
4211                                       0                     Republican
4212                                       0                     Republican
4213                                       0                     Republican
4214                                       0                     Republican
4215                                       0                     Republican

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                        …
4211                                       0.0
4212                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

      Vote_Data_Ted_Cruz_Number_of_Votes Vote_Data_Ted_Cruz_Party  \
0                                    876               Republican
1                                      0               Republican
2                                    208               Republican
3                                   1454               Republican
4                                    685               Republican
…                                  …                       …
4211                                2556               Republican
4212                                   0               Republican
4213                                  32               Republican
4214                                   0               Republican
4215                                  25               Republican

      Vote_Data_Ted_Cruz_Percent_of_Votes  \
0                                    23.9
1                                     0.0
2                                     9.9
3                                    38.2
4                                    15.9
```

```
…                                    …
4211                                 28.8
4212                                  0.0
4213                                 37.2
4214                                  0.0
4215                                 21.7

      Vote_Data_Uncommitted_Number_of_Votes Vote_Data_Uncommitted_Party  \
0                                         0                         NaN
1                                         0                         NaN
2                                         0                         NaN
3                                         0                         NaN
4                                         0                         NaN
…                                       …                           …
4211                                      0                         NaN
4212                                      0                         NaN
4213                                      0                         NaN
4214                                      0                         NaN
4215                                      0                         NaN

      Vote_Data_Uncommitted_Percent_of_Votes
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
…                                        …
4211                                     0.0
4212                                     0.0
4213                                     0.0
4214                                     0.0
4215                                     0.0

[4216 rows x 50 columns]
```

[6]:
```
df = df.fillna(0)
```

[7]:
```
#check totol vote number
sum_column = df.filter(regex='Number',axis =1).sum(axis=1) -␣
 ↪df['Vote_Data_No_Preference_Number_of_Votes']
df['sum_column'] = sum_column
df['sum_column'].sort_values(ascending=False)
```

[7]:
```
2214    1268622
677      760894
831      678313
1630     537303
```

```
2318     464471
         …
3113          0
146           0
3852          0
1492          0
4170          0
Name: sum_column, Length: 4216, dtype: int64
```

1.2.1 number extreme value which will cause inaccurate

[8]:
```
#calculate the average amont of vote of each states
df['sum_column'].mean()
```

[8]:
```
13460.871679316888
```

[9]:
```
#filter the low vote number state
df_filtered = df[df['sum_column'] > 100]
df = df_filtered.drop('sum_column', axis=1)
df
```

[9]:
```
     Location_County  Location_State Location_State_Abbreviation  \
0          Abbeville  South Carolina                          SC
2           Abington   Massachusetts                          MA
3             Acadia       Louisiana                          LA
4           Accomack        Virginia                          VA
5              Acton   Massachusetts                          MA
…                …               …                         …
4210            Yuba      California                          CA
4211            Yuma         Arizona                          AZ
4213          Zapata           Texas                          TX
4214          Zavala           Texas                          TX
4215         Ziebach    South Dakota                          SD

      Vote_Data_Ben_Carson_Number_of_Votes Vote_Data_Ben_Carson_Party  \
0                                      305                 Republican
2                                       53                 Republican
3                                        0                 Republican
4                                      411                 Republican
5                                       54                 Republican
…                                      …                          …
4210                                    0                 Republican
4211                                    0                 Republican
4213                                    4                 Republican
4214                                    0                 Republican
4215                                    0                 Republican
```

```
      Vote_Data_Ben_Carson_Percent_of_Votes  \
0                                       8.3
2                                       2.5
3                                       0.0
4                                       9.5
5                                       2.3
…                                       …
4210                                    0.0
4211                                    0.0
4213                                    4.7
4214                                    0.0
4215                                    0.0

      Vote_Data_Bernie_Sanders_Number_of_Votes Vote_Data_Bernie_Sanders_Party  \
0                                          312                       Democrat
2                                         1352                       Democrat
3                                         1087                       Democrat
4                                          682                       Democrat
5                                         2557                       Democrat
…                                         …                           …
4210                                      1730                       Democrat
4211                                      2156                       Democrat
4213                                       685                       Democrat
4214                                       373                       Democrat
4215                                       132                       Democrat

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                          17.0
2                                          53.4
3                                          31.4
4                                          27.4
5                                          46.8
…                                          …
4210                                       51.1
4211                                       31.5
4213                                       23.6
4214                                       18.1
4215                                       58.9

      Vote_Data_Carly_Fiorina_Number_of_Votes  …  \
0                                           0  …
2                                           0  …
3                                           0  …
4                                           0  …
5                                           0  …
…                                          …  …
4210                                        0  …
```

```
4211                                        0  …
4213                                        0  …
4214                                        0  …
4215                                        0  …

     Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                     0.0
2                                     0.0
3                                     0.0
4                                     0.0
5                                     0.0
…                                     …
4210                                  0.0
4211                                  0.0
4213                                  0.0
4214                                  0.0
4215                                  0.0

      Vote_Data_Rick_Santorum_Number_of_Votes  Vote_Data_Rick_Santorum_Party  \
0                                           0                     Republican
2                                           0                     Republican
3                                           0                     Republican
4                                           0                     Republican
5                                           0                     Republican
…                                           …                              …
4210                                        0                     Republican
4211                                        0                     Republican
4213                                        0                     Republican
4214                                        0                     Republican
4215                                        0                     Republican

     Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                         0.0
2                                         0.0
3                                         0.0
4                                         0.0
5                                         0.0
…                                         …
4210                                      0.0
4211                                      0.0
4213                                      0.0
4214                                      0.0
4215                                      0.0

      Vote_Data_Ted_Cruz_Number_of_Votes  Vote_Data_Ted_Cruz_Party  \
0                                    876                Republican
2                                    208                Republican
```

```
3                                   1454                Republican
4                                    685                Republican
5                                    203                Republican
…                                    …                     …
4210                                 318                Republican
4211                                2556                Republican
4213                                  32                Republican
4214                                   0                Republican
4215                                  25                Republican

     Vote_Data_Ted_Cruz_Percent_of_Votes  \
0                                   23.9
2                                    9.9
3                                   38.2
4                                   15.9
5                                    8.7
…                                    …
4210                                 8.2
4211                                28.8
4213                                37.2
4214                                 0.0
4215                                21.7

      Vote_Data_Uncommitted_Number_of_Votes  Vote_Data_Uncommitted_Party  \
0                                         0                          0.0
2                                         0                          0.0
3                                         0                          0.0
4                                         0                          0.0
5                                         0                          0.0
…                                         …                          …
4210                                      0                          0.0
4211                                      0                          0.0
4213                                      0                          0.0
4214                                      0                          0.0
4215                                      0                          0.0

     Vote_Data_Uncommitted_Percent_of_Votes
0                                       0.0
2                                       0.0
3                                       0.0
4                                       0.0
5                                       0.0
…                                       …
4210                                    0.0
4211                                    0.0
4213                                    0.0
4214                                    0.0
```

```
4215                                    0.0

[3592 rows x 51 columns]
```

[10]:
```
#delete the col including the number, we would like to use the percentage for cluster
df = df.drop(df.filter(regex='Number', axis=1).columns, axis=1)
df
```

[10]:
```
     Location_County  Location_State Location_State_Abbreviation  \
0          Abbeville  South Carolina                           SC
2           Abington   Massachusetts                           MA
3             Acadia       Louisiana                           LA
4           Accomack        Virginia                           VA
5              Acton   Massachusetts                           MA
…                …               …                          …
4210            Yuba      California                           CA
4211            Yuma         Arizona                           AZ
4213          Zapata           Texas                           TX
4214          Zavala           Texas                           TX
4215         Ziebach    South Dakota                           SD

     Vote_Data_Ben_Carson_Party  Vote_Data_Ben_Carson_Percent_of_Votes  \
0                    Republican                                   8.3
2                    Republican                                   2.5
3                    Republican                                   0.0
4                    Republican                                   9.5
5                    Republican                                   2.3
…                           …                                  …
4210                 Republican                                   0.0
4211                 Republican                                   0.0
4213                 Republican                                   4.7
4214                 Republican                                   0.0
4215                 Republican                                   0.0

     Vote_Data_Bernie_Sanders_Party  \
0                          Democrat
2                          Democrat
3                          Democrat
4                          Democrat
5                          Democrat
…                             …
4210                       Democrat
4211                       Democrat
4213                       Democrat
4214                       Democrat
4215                       Democrat
```

```
      Vote_Data_Bernie_Sanders_Percent_of_Votes Vote_Data_Carly_Fiorina_Party  \
0                                          17.0                    Republican
2                                          53.4                    Republican
3                                          31.4                    Republican
4                                          27.4                    Republican
5                                          46.8                    Republican
…                                           …                          …
4210                                       51.1                    Republican
4211                                       31.5                    Republican
4213                                       23.6                    Republican
4214                                       18.1                    Republican
4215                                       58.9                    Republican

      Vote_Data_Carly_Fiorina_Percent_of_Votes Vote_Data_Chris_Christie_Party  \
0                                          0.0                    Republican
2                                          0.0                    Republican
3                                          0.0                    Republican
4                                          0.0                    Republican
5                                          0.0                    Republican
…                                           …                          …
4210                                       0.0                    Republican
4211                                       0.0                    Republican
4213                                       0.0                    Republican
4214                                       0.0                    Republican
4215                                       0.0                    Republican

      …  Vote_Data_No_Preference_Party  \
0     …                           0.0
2     …                           0.0
3     …                           0.0
4     …                           0.0
5     …                           0.0
…   …                             …
4210  …                           0.0
4211  …                           0.0
4213  …                           0.0
4214  …                           0.0
4215  …                           0.0

     Vote_Data_No_Preference_Percent_of_Votes  Vote_Data_Rand_Paul_Party  \
0                                         0.0                 Republican
2                                         1.2                 Republican
3                                         0.0                 Republican
4                                         0.0                 Republican
5                                         0.2                 Republican
…                                          …                        …
```

```
4210                                      0.0                  Republican
4211                                      0.0                  Republican
4213                                      0.0                  Republican
4214                                      0.0                  Republican
4215                                      0.0                  Republican

     Vote_Data_Rand_Paul_Percent_of_Votes  Vote_Data_Rick_Santorum_Party  \
0                                     0.0                     Republican
2                                     0.0                     Republican
3                                     0.0                     Republican
4                                     0.0                     Republican
5                                     0.0                     Republican
…                                     …                          …
4210                                  0.0                     Republican
4211                                  0.0                     Republican
4213                                  0.0                     Republican
4214                                  0.0                     Republican
4215                                  0.0                     Republican

     Vote_Data_Rick_Santorum_Percent_of_Votes  Vote_Data_Ted_Cruz_Party  \
0                                         0.0                Republican
2                                         0.0                Republican
3                                         0.0                Republican
4                                         0.0                Republican
5                                         0.0                Republican
…                                         …                       …
4210                                      0.0                Republican
4211                                      0.0                Republican
4213                                      0.0                Republican
4214                                      0.0                Republican
4215                                      0.0                Republican

     Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                   23.9                          0.0
2                                    9.9                          0.0
3                                   38.2                          0.0
4                                   15.9                          0.0
5                                    8.7                          0.0
…                                    …                            …
4210                                 8.2                          0.0
4211                                28.8                          0.0
4213                                37.2                          0.0
4214                                 0.0                          0.0
4215                                21.7                          0.0

     Vote_Data_Uncommitted_Percent_of_Votes
0                                       0.0
```

```
2                                          0.0
3                                          0.0
4                                          0.0
5                                          0.0
…                                          …
4210                                       0.0
4211                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

[3592 rows x 35 columns]
```

[11]:
```
df_num = df.select_dtypes(include='number')
df_num
```

[11]:
```
      Vote_Data_Ben_Carson_Percent_of_Votes  \
0                                       8.3
2                                       2.5
3                                       0.0
4                                       9.5
5                                       2.3
…                                       …
4210                                    0.0
4211                                    0.0
4213                                    4.7
4214                                    0.0
4215                                    0.0

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                          17.0
2                                          53.4
3                                          31.4
4                                          27.4
5                                          46.8
…                                          …
4210                                       51.1
4211                                       31.5
4213                                       23.6
4214                                       18.1
4215                                       58.9

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
```

```
5                                          0.0
…                                          …
4210                                       0.0
4211                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
5                                          0.0
…                                          …
4210                                       0.0
4211                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                        36.9
2                                        57.9
3                                        44.5
4                                        47.9
5                                        28.2
…                                        …
4210                                     82.8
4211                                     49.5
4213                                     39.5
4214                                      0.0
4215                                     68.7

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                          81.8
2                                          44.6
3                                          53.7
4                                          72.0
5                                          52.8
…                                          …
4210                                       46.4
4211                                       63.7
4213                                       67.8
4214                                       75.5
4215                                       41.1
```

```
      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                     6.4
2                                     0.0
3                                     0.0
4                                     0.0
5                                     0.0
…                                     …
4210                                  0.0
4211                                  0.0
4213                                  0.0
4214                                  0.0
4215                                  0.0

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                       4.3
2                                      14.2
3                                       3.4
4                                       4.9
5                                      32.5
…                                      …
4210                                    5.1
4211                                    5.9
4213                                    0.0
4214                                    0.0
4215                                    9.6

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                      20.2
2                                      12.7
3                                      10.6
4                                      20.9
5                                      24.9
…                                      …
4210                                    0.0
4211                                    0.0
4213                                   12.8
4214                                    0.0
4215                                    0.0

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
5                                           0.0
…                                           …
4210                                        0.0
```

```
4211                                          0.0
4213                                          0.0
4214                                          0.0
4215                                          0.0

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                          0.0                            0.0
2                                          0.0                            0.0
3                                          0.0                            0.0
4                                          0.0                            0.0
5                                          0.0                            0.0
…                                          …                              …
4210                                       0.0                            0.0
4211                                       0.0                            0.0
4213                                       0.0                            0.0
4214                                       0.0                            0.0
4215                                       0.0                            0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                          0.0
2                                          1.2
3                                          0.0
4                                          0.0
5                                          0.2
…                                          …
4210                                       0.0
4211                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
5                                      0.0
…                                      …
4210                                   0.0
4211                                   0.0
4213                                   0.0
4214                                   0.0
4215                                   0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
2                                          0.0
```

```
3                                          0.0
4                                          0.0
5                                          0.0
…                                          …
4210                                       0.0
4211                                       0.0
4213                                       0.0
4214                                       0.0
4215                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                    23.9                          0.0
2                                     9.9                          0.0
3                                    38.2                          0.0
4                                    15.9                          0.0
5                                     8.7                          0.0
…                                     …                            …
4210                                  8.2                          0.0
4211                                 28.8                          0.0
4213                                 37.2                          0.0
4214                                  0.0                          0.0
4215                                 21.7                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes
0                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
5                                        0.0
…                                        …
4210                                     0.0
4211                                     0.0
4213                                     0.0
4214                                     0.0
4215                                     0.0

[3592 rows x 18 columns]
```

1. We first read the dataset.
2. Drop the NA value.
3. Then check totol vote number.
4. We calculate the average amont of vote of each states to drop the filter the low vote number state.
5. We filter the total vote number below the 100, which filter the extreme values out of the dataset.
6. Finally, Delete the col including the number, we would like to use the percentage for cluster.

Now, we have the percentage columns that only have the total vote number greater than 100,

"df_num".

1.3 3. Basic Descriptive Analytics

1.3.1 3.1 For numerical attributes, calculate basic summary statistics about each attribute.

[12]:
```
df_basic = df_num.describe()
df_basic
```

[12]:

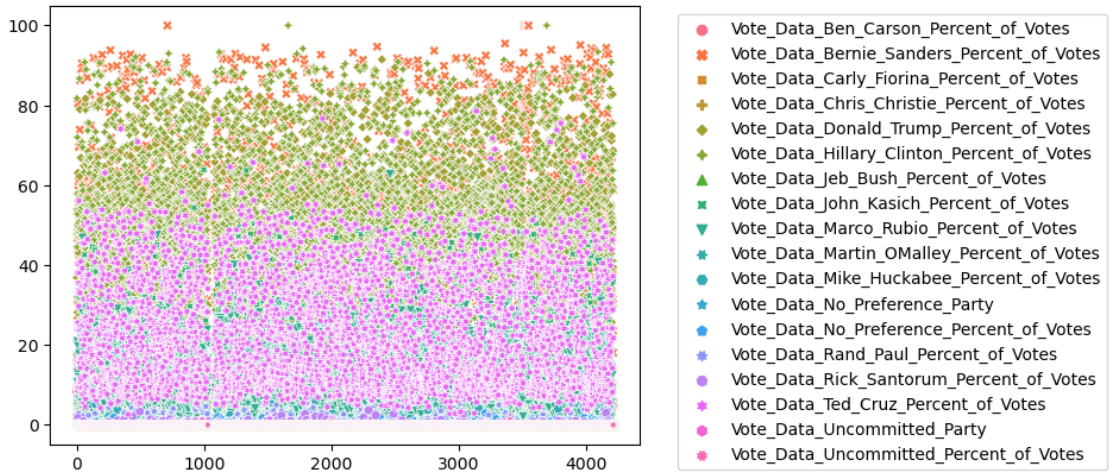
	Vote_Data_Ben_Carson_Percent_of_Votes	Vote_Data_Bernie_Sanders_Percent_of_Votes	Vote_Data_Carly_Fiorina_Percent_of_Votes	Vote_Data_Chris_Christie_Percent_of_Votes
count	3592.000000	3592.000000	3592.000000	3592.000000
mean	2.699807	47.883404	0.067054	0.053931
std	3.589121	18.395268	0.481854	0.436933
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	36.875000	0.000000	0.000000
50%	0.000000	48.600000	0.000000	0.000000
75%	5.400000	56.800000	0.000000	0.000000
max	21.700000	100.000000	11.700000	

```
max                                      8.719532

       Vote_Data_Donald_Trump_Percent_of_Votes  \
count                              3592.000000
mean                                 46.237583
std                                  15.930844
min                                   0.000000
25%                                  35.000000
50%                                  45.400000
75%                                  55.700000
max                                  91.500000

       Vote_Data_Hillary_Clinton_Percent_of_Votes  \
count                                3592.000000
mean                                   49.293099
std                                    18.057093
min                                     0.000000
25%                                    40.000000
50%                                    48.650000
75%                                    59.800000
max                                   100.000000

       Vote_Data_Jeb_Bush_Percent_of_Votes  \
count                           3592.000000
mean                              0.192129
std                               1.137126
min                               0.000000
25%                               0.000000
50%                               0.000000
75%                               0.000000
max                              12.100000

       Vote_Data_John_Kasich_Percent_of_Votes  \
count                              3592.000000
mean                                 12.111225
std                                  10.649826
min                                   0.000000
25%                                   3.900000
50%                                   8.450000
75%                                  17.500000
max                                  63.900000

       Vote_Data_Marco_Rubio_Percent_of_Votes  \
count                              3592.000000
mean                                 10.275768
std                                   9.123062
min                                   0.000000
```

```
25%                                    0.000000
50%                                   10.000000
75%                                   17.200000
max                                   62.700000

       Vote_Data_Martin_OMalley_Percent_of_Votes  \
count                                3592.000000
mean                                    0.022884
std                                     0.366343
min                                     0.000000
25%                                     0.000000
50%                                     0.000000
75%                                     0.000000
max                                    13.200000

       Vote_Data_Mike_Huckabee_Percent_of_Votes  \
count                                3592.000000
mean                                    0.066676
std                                     0.479296
min                                     0.000000
25%                                     0.000000
50%                                     0.000000
75%                                     0.000000
max                                     9.600000

       Vote_Data_No_Preference_Party  \
count                         3592.0
mean                             0.0
std                              0.0
min                              0.0
25%                              0.0
50%                              0.0
75%                              0.0
max                              0.0

       Vote_Data_No_Preference_Percent_of_Votes  \
count                                3592.000000
mean                                    0.063363
std                                     0.243077
min                                     0.000000
25%                                     0.000000
50%                                     0.000000
75%                                     0.000000
max                                     3.000000

       Vote_Data_Rand_Paul_Percent_of_Votes  \
count                           3592.000000
```

```
mean                                0.093875
std                                 0.621383
min                                 0.000000
25%                                 0.000000
50%                                 0.000000
75%                                 0.000000
max                                 9.600000

       Vote_Data_Rick_Santorum_Percent_of_Votes  \
count                               3592.000000
mean                                   0.027450
std                                    0.232522
min                                    0.000000
25%                                    0.000000
50%                                    0.000000
75%                                    0.000000
max                                    7.300000

       Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
count                          3592.000000                       3592.0
mean                             24.025030                          0.0
std                              13.759378                          0.0
min                               0.000000                          0.0
25%                              12.000000                          0.0
50%                              21.700000                          0.0
75%                              34.700000                          0.0
max                              78.600000                          0.0

       Vote_Data_Uncommitted_Percent_of_Votes
count                               3592.000000
mean                                   0.001253
std                                    0.036195
min                                    0.000000
25%                                    0.000000
50%                                    0.000000
75%                                    0.000000
max                                    1.300000
```

[13]:
```
sns.scatterplot(data=df_num)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```



1.3.2 3.2 For any categorical attributes, count up the number of observations of each type.

[14]:
```
df['Location_State'].value_counts().sort_values(ascending=False)
```

[14]:
```
Massachusetts     346
Texas             253
Vermont           229
Connecticut       169
Georgia           159
Virginia          133
Kentucky          120
Missouri          115
Illinois          103
North Carolina    100
Iowa               99
Tennessee          95
Nebraska           93
Indiana            92
Ohio               88
Michigan           83
Mississippi        82
Oklahoma           77
Arkansas           74
Wisconsin          72
Alabama            67
Florida            67
Pennsylvania       67
South Dakota       66
Louisiana          64
```

```
New York           62
California         58
Montana            56
West Virginia      55
South Carolina     46
Idaho              44
Colorado           43
Rhode Island       41
Washington         39
Alaska             38
Oregon             36
New Mexico         33
Utah               29
Maryland           24
New Jersey         21
Nevada             16
Arizona            15
New Hampshire      10
Kansas              4
Hawaii              4
Delaware            3
Wyoming             1
Maine               1
Name: Location_State, dtype: int64
```

[15]:
```
df['Location_County'].value_counts().sort_values(ascending=False)
```

[15]:
```
Washington     28
Jefferson      25
Franklin       24
Lincoln        22
Jackson        21
               ..
Seward          1
Seymour         1
Shackelford     1
Shaftsbury      1
Ziebach         1
Name: Location_County, Length: 2215, dtype: int64
```

[16]:
```
df['Vote_Data_Ben_Carson_Party'].value_counts().sort_values(ascending=False)
```

[16]:
```
Republican    3592
Name: Vote_Data_Ben_Carson_Party, dtype: int64
```

[17]:
```
df['Vote_Data_Bernie_Sanders_Party'].value_counts().sort_values(ascending=False)
```

[17]:
```
Democrat    3592
Name: Vote_Data_Bernie_Sanders_Party, dtype: int64
```

[18]:
```
df['Vote_Data_Rick_Santorum_Party'].value_counts().sort_values(ascending=False)
```

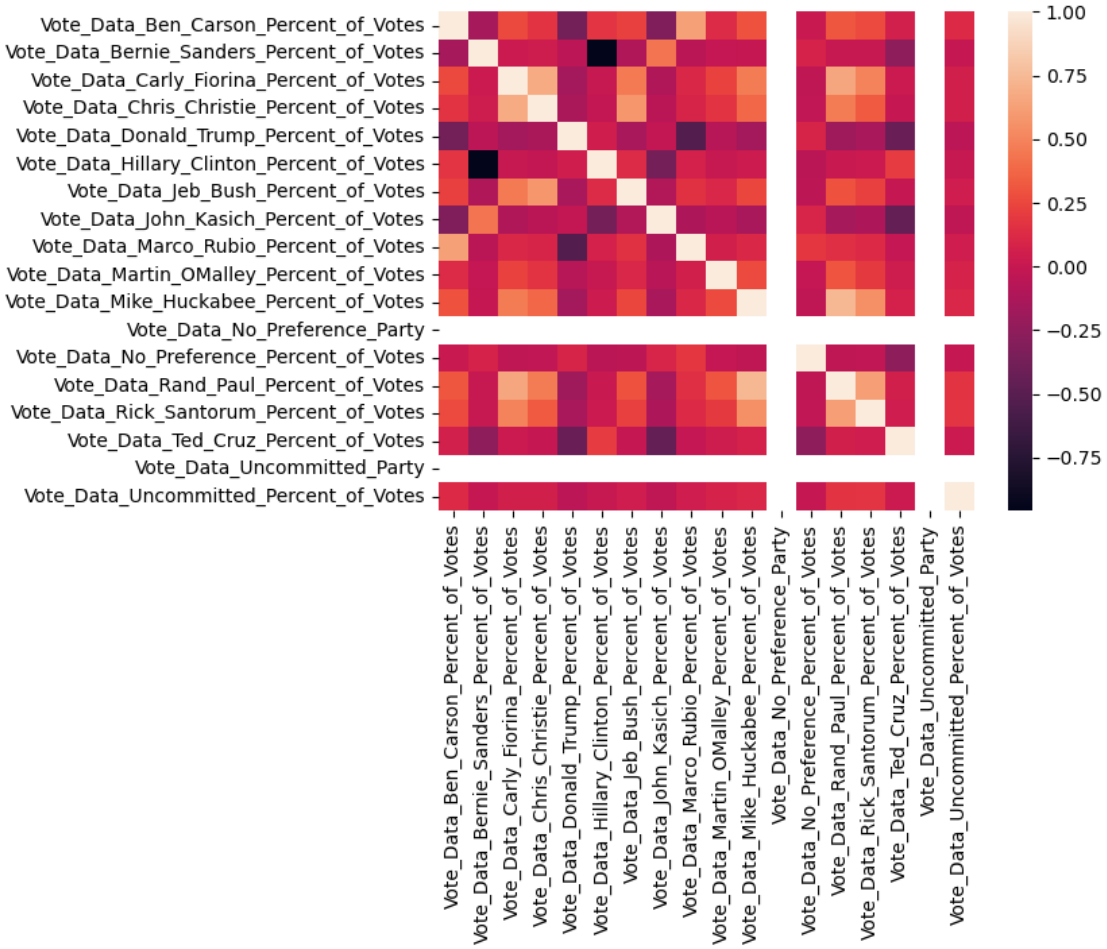
[18]:
```
Republican    3592
Name: Vote_Data_Rick_Santorum_Party, dtype: int64
```

[19]:
```
df['Vote_Data_Ted_Cruz_Party'].value_counts().sort_values(ascending=False)
```

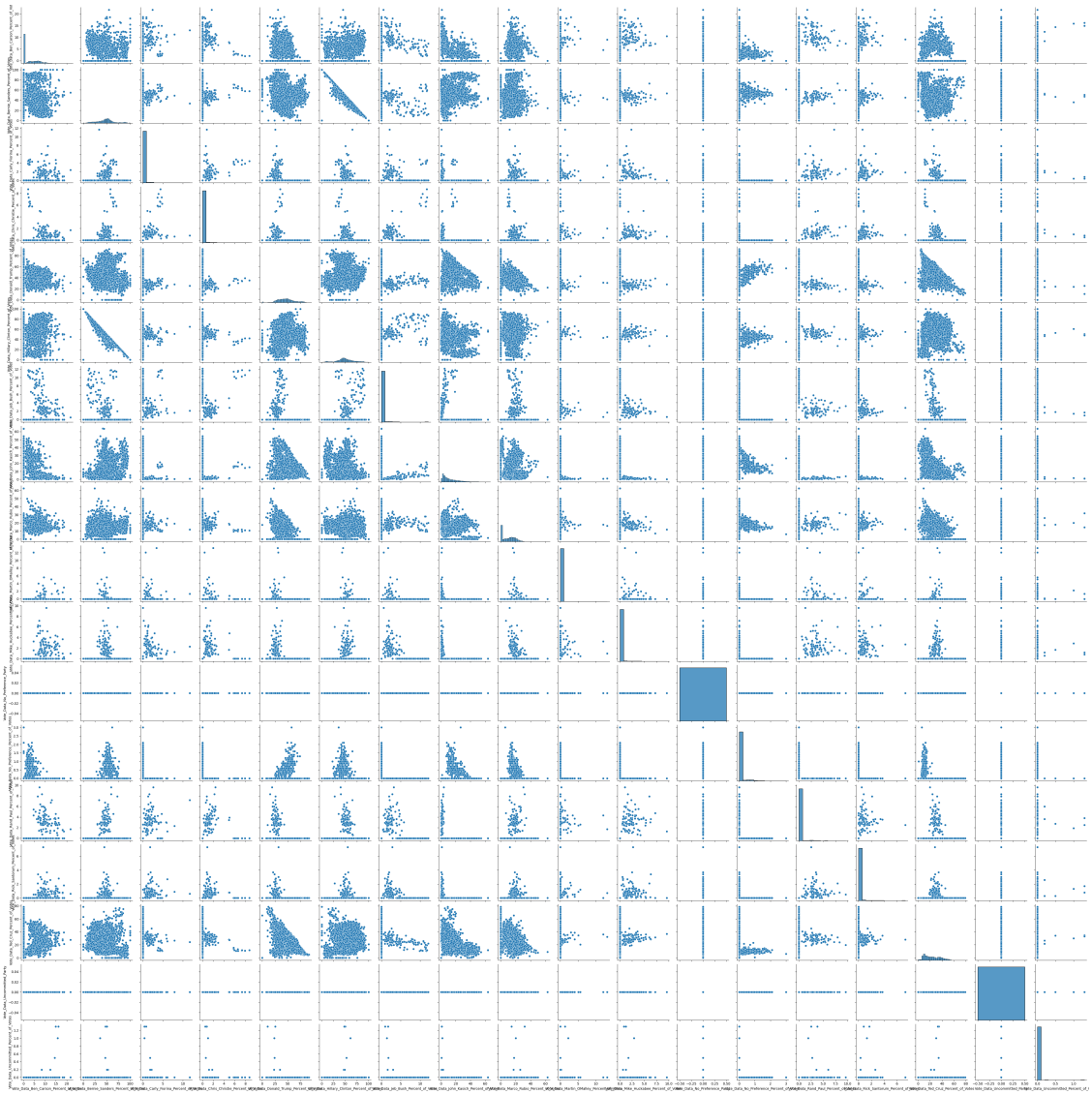
[19]:
```
Republican    3592
Name: Vote_Data_Ted_Cruz_Party, dtype: int64
```

1.3.3 3.3 Determine if there exist are any strong pairwise relationships between the variables in your dataset

[20]:
```
sns.heatmap(df_num.corr())
plt.show()
```



[21]:
```
sns.pairplot(data = df)
plt.show()
```



From the above grahp, we can see there are no strong relationship between most variable. However, the 'Vote_Data_Bernie_Sanders_Percent_of_Votes' and 'Vote_Data_Hillary_Clinton_Percent_of_Votes' have strong negative relation, which means if the extend of people want to vote for Bernie Sanders, the extend of Hillary Clinton will decrease.

1.4 4. Dataset Scaling Decisions

Scaling data using the Min-Max Scaler means scaling the data to a fixed range of values, usually [0,1] or [-1,1]. The difference between the two scaling methods using Standard Scaler and Min-Max

Scaler is that the scaling range of the data is different: Standard Scaler scales the data to a normal distribution with a mean of 0 and a standard deviation of 1, while Min-Max Scaler scales the data to a specified minimum and maximum value range.

The distribution range of the data is different: If the distribution range of the data is large, using Standard Scaler may scale the data to a smaller range, resulting in the loss of some characteristic information of the data. While using Min-Max Scaler can scale the data to a specified range and keep more information about the data features.

Outliers in the data: If there are outliers in the data, using Standard Scaler may cause the scaling range of the data to be affected by the outliers, thus affecting the performance of the model. Using the Min-Max Scaler avoids this problem by scaling the outliers to a specified range.

Model requirements: Different models have different requirements for data scaling. For example, some models (e.g., SVM) are more sensitive to data scaling, while others (e.g., decision trees) are less sensitive to data scaling. Therefore, the needs of the model used need to be considered when choosing a data scaling method.

In general, the choice of which scaling method to use should be based on the specific data set and model. If the distribution of the data is relatively large or there are outliers, it is recommended to use the Min-Max Scaler; if the distribution of the data is closer to normal or the model is not sensitive to data scaling, the Standard Scaler can be used.

[22]:
```
scale = StandardScaler()
df_scale1 = scale.fit_transform(df_num)
df_scale1
```

[22]:
```
array([[ 1.56054152, -1.67911137, -0.13917705, …, -0.00908819,
         0.        , -0.03461736],
       [-0.05567791,  0.29993386, -0.13917705, …, -1.02671774,
         0.        , -0.03461736],
       [-0.75232422, -0.89619238, -0.13917705, …,  1.0303477 ,
         0.        , -0.03461736],
       …,
       [ 0.55737084, -1.3202735 , -0.13917705, …,  0.95765988,
         0.        , -0.03461736],
       [-0.75232422, -1.61930506, -0.13917705, …, -1.74632721,
         0.        , -0.03461736],
       [-0.75232422,  0.59896542, -0.13917705, …, -0.16900141,
         0.        , -0.03461736]])
```

[23]:
```
Z = pd.DataFrame(df_scale1, columns=df_num.columns)
Z
```

[23]:

	Vote_Data_Ben_Carson_Percent_of_Votes	\
0	1.560542	
1	-0.055678	
2	-0.752324	
3	1.894932	
4	-0.111410	

```
…                                    …
3587                              -0.752324
3588                              -0.752324
3589                               0.557371
3590                              -0.752324
3591                              -0.752324

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                     -1.679111
1                                      0.299934
2                                     -0.896192
3                                     -1.113670
4                                     -0.058904
…                                         …
3587                                   0.174884
3588                                  -0.890755
3589                                  -1.320274
3590                                  -1.619305
3591                                   0.598965

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                    -0.139177
1                                    -0.139177
2                                    -0.139177
3                                    -0.139177
4                                    -0.139177
…                                        …
3587                                 -0.139177
3588                                 -0.139177
3589                                 -0.139177
3590                                 -0.139177
3591                                 -0.139177

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                     -0.123449
1                                     -0.123449
2                                     -0.123449
3                                     -0.123449
4                                     -0.123449
…                                         …
3587                                  -0.123449
3588                                  -0.123449
3589                                  -0.123449
3590                                  -0.123449
3591                                  -0.123449

      Vote_Data_Donald_Trump_Percent_of_Votes  \
```

```
0                                    -0.586214
1                                     0.732167
2                                    -0.109086
3                                     0.104367
4                                    -1.132400
…                                          …
3587                                  2.295390
3588                                  0.204815
3589                                 -0.422986
3590                                 -2.902798
3591                                  1.410192

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                       1.800480
1                                      -0.259940
2                                       0.244088
3                                       1.257681
4                                       0.194239
…                                            …
3587                                   -0.160242
3588                                    0.797964
3589                                    1.025053
3590                                    1.451538
3591                                   -0.453796

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                5.460023
1                               -0.168983
2                               -0.168983
3                               -0.168983
4                               -0.168983
…                                    …
3587                            -0.168983
3588                            -0.168983
3589                            -0.168983
3590                            -0.168983
3591                            -0.168983

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                  -0.733563
1                                   0.196160
2                                  -0.818083
3                                  -0.677216
4                                   1.914737
…                                        …
3587                               -0.658433
3588                               -0.583304
```

```
3589                                -1.137381
3590                                -1.137381
3591                                -0.235833

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                   1.087969
1                                   0.265763
2                                   0.035545
3                                   1.164709
4                                   1.603219
…                                        …
3587                                -1.126508
3588                                -1.126508
3589                                 0.276725
3590                                -1.126508
3591                                -1.126508

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                     -0.062475
1                                     -0.062475
2                                     -0.062475
3                                     -0.062475
4                                     -0.062475
…                                          …
3587                                  -0.062475
3588                                  -0.062475
3589                                  -0.062475
3590                                  -0.062475
3591                                  -0.062475

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                    -0.139132                            0.0
1                                    -0.139132                            0.0
2                                    -0.139132                            0.0
3                                    -0.139132                            0.0
4                                    -0.139132                            0.0
…                                         …                              …
3587                                 -0.139132                            0.0
3588                                 -0.139132                            0.0
3589                                 -0.139132                            0.0
3590                                 -0.139132                            0.0
3591                                 -0.139132                            0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                    -0.260707
1                                     4.676696
2                                    -0.260707
```

```
3                                    -0.260707
4                                     0.562193
…                                          …
3587                                 -0.260707
3588                                 -0.260707
3589                                 -0.260707
3590                                 -0.260707
3591                                 -0.260707

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                 -0.151096
1                                 -0.151096
2                                 -0.151096
3                                 -0.151096
4                                 -0.151096
…                                      …
3587                              -0.151096
3588                              -0.151096
3589                              -0.151096
3590                              -0.151096
3591                              -0.151096

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                    -0.118069
1                                    -0.118069
2                                    -0.118069
3                                    -0.118069
4                                    -0.118069
…                                         …
3587                                 -0.118069
3588                                 -0.118069
3589                                 -0.118069
3590                                 -0.118069
3591                                 -0.118069

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                -0.009088                          0.0
1                                -1.026718                          0.0
2                                 1.030348                          0.0
3                                -0.590591                          0.0
4                                -1.113943                          0.0
…                                     …                            …
3587                             -1.150287                          0.0
3588                              0.347082                          0.0
3589                              0.957660                          0.0
3590                             -1.746327                          0.0
3591                             -0.169001                          0.0
```

```
      Vote_Data_Uncommitted_Percent_of_Votes
0                                  -0.034617
1                                  -0.034617
2                                  -0.034617
3                                  -0.034617
4                                  -0.034617
…                                        …
3587                               -0.034617
3588                               -0.034617
3589                               -0.034617
3590                               -0.034617
3591                               -0.034617

[3592 rows x 18 columns]
```

[24]:
```
df['Location_State']
```

[24]:
```
0       South Carolina
2        Massachusetts
3            Louisiana
4             Virginia
5        Massachusetts
              …
4210        California
4211           Arizona
4213             Texas
4214             Texas
4215      South Dakota
Name: Location_State, Length: 3592, dtype: object
```

[25]:
```
df.reset_index(inplace=True)
C = df['Location_State']
frames = [C,Z]
df_label_standard = pd.concat(frames,axis = 1)
df_label_standard
```

[25]:
```
      Location_State  Vote_Data_Ben_Carson_Percent_of_Votes  \
0     South Carolina                                1.560542
1      Massachusetts                               -0.055678
2          Louisiana                               -0.752324
3           Virginia                                1.894932
4      Massachusetts                               -0.111410
…                …                                       …
3587      California                               -0.752324
3588         Arizona                               -0.752324
3589           Texas                                0.557371
```

```
3590           Texas                                -0.752324
3591    South Dakota                                -0.752324

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                     -1.679111
1                                      0.299934
2                                     -0.896192
3                                     -1.113670
4                                     -0.058904
…                                           …
3587                                   0.174884
3588                                  -0.890755
3589                                  -1.320274
3590                                  -1.619305
3591                                   0.598965

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                    -0.139177
1                                    -0.139177
2                                    -0.139177
3                                    -0.139177
4                                    -0.139177
…                                           …
3587                                 -0.139177
3588                                 -0.139177
3589                                 -0.139177
3590                                 -0.139177
3591                                 -0.139177

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                     -0.123449
1                                     -0.123449
2                                     -0.123449
3                                     -0.123449
4                                     -0.123449
…                                           …
3587                                  -0.123449
3588                                  -0.123449
3589                                  -0.123449
3590                                  -0.123449
3591                                  -0.123449

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    -0.586214
1                                     0.732167
2                                    -0.109086
3                                     0.104367
```

```
4                                    -1.132400
…                                          …
3587                                  2.295390
3588                                  0.204815
3589                                 -0.422986
3590                                 -2.902798
3591                                  1.410192

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                       1.800480
1                                      -0.259940
2                                       0.244088
3                                       1.257681
4                                       0.194239
…                                             …
3587                                   -0.160242
3588                                    0.797964
3589                                    1.025053
3590                                    1.451538
3591                                   -0.453796

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                5.460023
1                               -0.168983
2                               -0.168983
3                               -0.168983
4                               -0.168983
…                                     …
3587                            -0.168983
3588                            -0.168983
3589                            -0.168983
3590                            -0.168983
3591                            -0.168983

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                  -0.733563
1                                   0.196160
2                                  -0.818083
3                                  -0.677216
4                                   1.914737
…                                         …
3587                               -0.658433
3588                               -0.583304
3589                               -1.137381
3590                               -1.137381
3591                               -0.235833
```

```
      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                   1.087969
1                                   0.265763
2                                   0.035545
3                                   1.164709
4                                   1.603219
…                                        …
3587                               -1.126508
3588                               -1.126508
3589                                0.276725
3590                               -1.126508
3591                               -1.126508

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                     -0.062475
1                                     -0.062475
2                                     -0.062475
3                                     -0.062475
4                                     -0.062475
…                                          …
3587                                  -0.062475
3588                                  -0.062475
3589                                  -0.062475
3590                                  -0.062475
3591                                  -0.062475

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                    -0.139132                            0.0
1                                    -0.139132                            0.0
2                                    -0.139132                            0.0
3                                    -0.139132                            0.0
4                                    -0.139132                            0.0
…                                         …                              …
3587                                 -0.139132                            0.0
3588                                 -0.139132                            0.0
3589                                 -0.139132                            0.0
3590                                 -0.139132                            0.0
3591                                 -0.139132                            0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                    -0.260707
1                                     4.676696
2                                    -0.260707
3                                    -0.260707
4                                     0.562193
…                                          …
3587                                 -0.260707
```

```
3588                                -0.260707
3589                                -0.260707
3590                                -0.260707
3591                                -0.260707

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                -0.151096
1                                -0.151096
2                                -0.151096
3                                -0.151096
4                                -0.151096
…                                    …
3587                             -0.151096
3588                             -0.151096
3589                             -0.151096
3590                             -0.151096
3591                             -0.151096

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                   -0.118069
1                                   -0.118069
2                                   -0.118069
3                                   -0.118069
4                                   -0.118069
…                                       …
3587                                -0.118069
3588                                -0.118069
3589                                -0.118069
3590                                -0.118069
3591                                -0.118069

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                               -0.009088                          0.0
1                               -1.026718                          0.0
2                                1.030348                          0.0
3                               -0.590591                          0.0
4                               -1.113943                          0.0
…                                   …                          …
3587                            -1.150287                          0.0
3588                             0.347082                          0.0
3589                             0.957660                          0.0
3590                            -1.746327                          0.0
3591                            -0.169001                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes
0                                  -0.034617
1                                  -0.034617
```

```
2                                    -0.034617
3                                    -0.034617
4                                    -0.034617
…                                          …
3587                                 -0.034617
3588                                 -0.034617
3589                                 -0.034617
3590                                 -0.034617
3591                                 -0.034617

[3592 rows x 19 columns]
```

[26]:
```
scaler = MinMaxScaler()
df_scale = pd.DataFrame(scaler.fit_transform(df_num),columns=df_num.columns,);
df_num =df_scale.copy()
df_scale
```

[26]:
```
      Vote_Data_Ben_Carson_Percent_of_Votes  \
0                                  0.382488
1                                  0.115207
2                                  0.000000
3                                  0.437788
4                                  0.105991
…                                       …
3587                               0.000000
3588                               0.000000
3589                               0.216590
3590                               0.000000
3591                               0.000000

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                         0.170
1                                         0.534
2                                         0.314
3                                         0.274
4                                         0.468
…                                             …
3587                                      0.511
3588                                      0.315
3589                                      0.236
3590                                      0.181
3591                                      0.589

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
```

```
3                                                 0.0
4                                                 0.0
…                                                 …
3587                                              0.0
3588                                              0.0
3589                                              0.0
3590                                              0.0
3591                                              0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                                 0.0
1                                                 0.0
2                                                 0.0
3                                                 0.0
4                                                 0.0
…                                                 …
3587                                              0.0
3588                                              0.0
3589                                              0.0
3590                                              0.0
3591                                              0.0

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    0.403279
1                                    0.632787
2                                    0.486339
3                                    0.523497
4                                    0.308197
…                                    …
3587                                 0.904918
3588                                 0.540984
3589                                 0.431694
3590                                 0.000000
3591                                 0.750820

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                              0.818
1                                              0.446
2                                              0.537
3                                              0.720
4                                              0.528
…                                              …
3587                                           0.464
3588                                           0.637
3589                                           0.678
3590                                           0.755
3591                                           0.411
```

```
      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                0.528926
1                                0.000000
2                                0.000000
3                                0.000000
4                                0.000000
…                                   …
3587                             0.000000
3588                             0.000000
3589                             0.000000
3590                             0.000000
3591                             0.000000

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                   0.067293
1                                   0.222222
2                                   0.053208
3                                   0.076682
4                                   0.508607
…                                      …
3587                                0.079812
3588                                0.092332
3589                                0.000000
3590                                0.000000
3591                                0.150235

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                   0.322169
1                                   0.202552
2                                   0.169059
3                                   0.333333
4                                   0.397129
…                                      …
3587                                0.000000
3588                                0.000000
3589                                0.204147
3590                                0.000000
3591                                0.000000

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                           …
```

```
3587                                          0.0
3588                                          0.0
3589                                          0.0
3590                                          0.0
3591                                          0.0

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                          0.0                           0.0
1                                          0.0                           0.0
2                                          0.0                           0.0
3                                          0.0                           0.0
4                                          0.0                           0.0
…                                          …                             …
3587                                       0.0                           0.0
3588                                       0.0                           0.0
3589                                       0.0                           0.0
3590                                       0.0                           0.0
3591                                       0.0                           0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                     0.000000
1                                     0.400000
2                                     0.000000
3                                     0.000000
4                                     0.066667
…                                        …
3587                                  0.000000
3588                                  0.000000
3589                                  0.000000
3590                                  0.000000
3591                                  0.000000

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
…                                      …
3587                                   0.0
3588                                   0.0
3589                                   0.0
3590                                   0.0
3591                                   0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
```

```
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                0.304071                          0.0
1                                0.125954                          0.0
2                                0.486005                          0.0
3                                0.202290                          0.0
4                                0.110687                          0.0
…                                   …                              …
3587                             0.104326                          0.0
3588                             0.366412                          0.0
3589                             0.473282                          0.0
3590                             0.000000                          0.0
3591                             0.276081                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
…                                        …
3587                                     0.0
3588                                     0.0
3589                                     0.0
3590                                     0.0
3591                                     0.0

[3592 rows x 18 columns]
```

[27]:
```
X = pd.DataFrame(df_scale, columns=df_num.columns)
X
```

[27]:
```
      Vote_Data_Ben_Carson_Percent_of_Votes  \
0                                  0.382488
1                                  0.115207
2                                  0.000000
3                                  0.437788
```

```
4                                    0.105991
…                                       …
3587                                 0.000000
3588                                 0.000000
3589                                 0.216590
3590                                 0.000000
3591                                 0.000000

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                         0.170
1                                         0.534
2                                         0.314
3                                         0.274
4                                         0.468
…                                         …
3587                                      0.511
3588                                      0.315
3589                                      0.236
3590                                      0.181
3591                                      0.589

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                         …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                          …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
3591                                        0.0
```

```
      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    0.403279
1                                    0.632787
2                                    0.486339
3                                    0.523497
4                                    0.308197
…                                         …
3587                                 0.904918
3588                                 0.540984
3589                                 0.431694
3590                                 0.000000
3591                                 0.750820

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                          0.818
1                                          0.446
2                                          0.537
3                                          0.720
4                                          0.528
…                                          …
3587                                       0.464
3588                                       0.637
3589                                       0.678
3590                                       0.755
3591                                       0.411

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                0.528926
1                                0.000000
2                                0.000000
3                                0.000000
4                                0.000000
…                                   …
3587                             0.000000
3588                             0.000000
3589                             0.000000
3590                             0.000000
3591                             0.000000

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                    0.067293
1                                    0.222222
2                                    0.053208
3                                    0.076682
4                                    0.508607
…                                         …
3587                                 0.079812
```

```
3588                                0.092332
3589                                0.000000
3590                                0.000000
3591                                0.150235

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                   0.322169
1                                   0.202552
2                                   0.169059
3                                   0.333333
4                                   0.397129
…                                      …
3587                                0.000000
3588                                0.000000
3589                                0.204147
3590                                0.000000
3591                                0.000000

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                           …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
3591                                        0.0

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                          0.0                           0.0
1                                          0.0                           0.0
2                                          0.0                           0.0
3                                          0.0                           0.0
4                                          0.0                           0.0
…                                          …                             …
3587                                       0.0                           0.0
3588                                       0.0                           0.0
3589                                       0.0                           0.0
3590                                       0.0                           0.0
3591                                       0.0                           0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                     0.000000
1                                     0.400000
```

```
2                                    0.000000
3                                    0.000000
4                                    0.066667
…                                         …
3587                                 0.000000
3588                                 0.000000
3589                                 0.000000
3590                                 0.000000
3591                                 0.000000

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
…                                        …
3587                                   0.0
3588                                   0.0
3589                                   0.0
3590                                   0.0
3591                                   0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                0.304071                          0.0
1                                0.125954                          0.0
2                                0.486005                          0.0
3                                0.202290                          0.0
4                                0.110687                          0.0
…                                   …                              …
3587                             0.104326                          0.0
3588                             0.366412                          0.0
3589                             0.473282                          0.0
3590                             0.000000                          0.0
```

```
3591                                0.276081                                0.0

      Vote_Data_Uncommitted_Percent_of_Votes
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
…                                        …
3587                                     0.0
3588                                     0.0
3589                                     0.0
3590                                     0.0
3591                                     0.0

[3592 rows x 18 columns]
```

[28]:
```
C = df['Location_State']
frames = [C,X]
df_label = pd.concat(frames,axis = 1)
df_label
```

[28]:
```
      Location_State  Vote_Data_Ben_Carson_Percent_of_Votes  \
0     South Carolina                                0.382488
1      Massachusetts                                0.115207
2          Louisiana                                0.000000
3           Virginia                                0.437788
4      Massachusetts                                0.105991
…                …                                       …
3587      California                                0.000000
3588         Arizona                                0.000000
3589           Texas                                0.216590
3590           Texas                                0.000000
3591    South Dakota                                0.000000

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                         0.170
1                                         0.534
2                                         0.314
3                                         0.274
4                                         0.468
…                                         …
3587                                      0.511
3588                                      0.315
3589                                      0.236
3590                                      0.181
3591                                      0.589
```

```
      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                           …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
3591                                        0.0

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    0.403279
1                                    0.632787
2                                    0.486339
3                                    0.523497
4                                    0.308197
…                                       …
3587                                 0.904918
3588                                 0.540984
3589                                 0.431694
3590                                 0.000000
3591                                 0.750820

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                          0.818
1                                          0.446
2                                          0.537
3                                          0.720
4                                          0.528
…                                          …
```

```
3587                                        0.464
3588                                        0.637
3589                                        0.678
3590                                        0.755
3591                                        0.411

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                0.528926
1                                0.000000
2                                0.000000
3                                0.000000
4                                0.000000
…                                   …
3587                             0.000000
3588                             0.000000
3589                             0.000000
3590                             0.000000
3591                             0.000000

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                   0.067293
1                                   0.222222
2                                   0.053208
3                                   0.076682
4                                   0.508607
…                                      …
3587                                0.079812
3588                                0.092332
3589                                0.000000
3590                                0.000000
3591                                0.150235

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                                   0.322169
1                                   0.202552
2                                   0.169059
3                                   0.333333
4                                   0.397129
…                                      …
3587                                0.000000
3588                                0.000000
3589                                0.204147
3590                                0.000000
3591                                0.000000

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                           0.0
```

```
1                                                0.0
2                                                0.0
3                                                0.0
4                                                0.0
…                                                …
3587                                             0.0
3588                                             0.0
3589                                             0.0
3590                                             0.0
3591                                             0.0

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                          0.0                            0.0
1                                          0.0                            0.0
2                                          0.0                            0.0
3                                          0.0                            0.0
4                                          0.0                            0.0
…                                          …                              …
3587                                       0.0                            0.0
3588                                       0.0                            0.0
3589                                       0.0                            0.0
3590                                       0.0                            0.0
3591                                       0.0                            0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                     0.000000
1                                     0.400000
2                                     0.000000
3                                     0.000000
4                                     0.066667
…                                        …
3587                                  0.000000
3588                                  0.000000
3589                                  0.000000
3590                                  0.000000
3591                                  0.000000

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
…                                      …
3587                                   0.0
3588                                   0.0
3589                                   0.0
```

```
3590                                    0.0
3591                                    0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                0.304071                          0.0
1                                0.125954                          0.0
2                                0.486005                          0.0
3                                0.202290                          0.0
4                                0.110687                          0.0
…                                   …                          …
3587                             0.104326                          0.0
3588                             0.366412                          0.0
3589                             0.473282                          0.0
3590                             0.000000                          0.0
3591                             0.276081                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
…                                        …
3587                                     0.0
3588                                     0.0
3589                                     0.0
3590                                     0.0
3591                                     0.0

[3592 rows x 19 columns]
```

[29]:
```
df_cata = df.select_dtypes(include=['object', 'category'])
df_cata = [df_cata,X]
df_cata = pd.concat(df_cata,axis = 1)
```

```
df_cata = df_cata.drop(['Location_County'],axis = 1)
df_cata.head()
```

[29]:

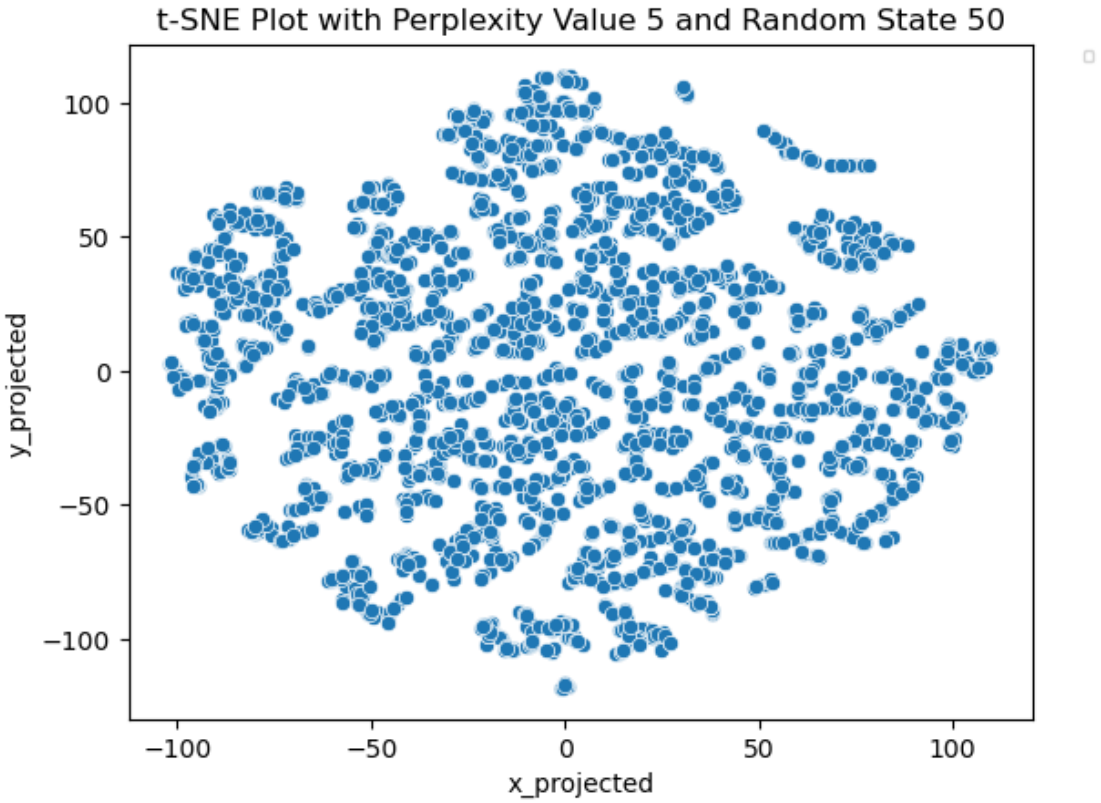
```
  Location_State Location_State_Abbreviation Vote_Data_Ben_Carson_Party  \
0  South Carolina                         SC                 Republican
1   Massachusetts                         MA                 Republican
2       Louisiana                         LA                 Republican
3        Virginia                         VA                 Republican
4   Massachusetts                         MA                 Republican

  Vote_Data_Bernie_Sanders_Party Vote_Data_Carly_Fiorina_Party  \
0                       Democrat                   Republican
1                       Democrat                   Republican
2                       Democrat                   Republican
3                       Democrat                   Republican
4                       Democrat                   Republican

  Vote_Data_Chris_Christie_Party Vote_Data_Donald_Trump_Party  \
0                     Republican                   Republican
1                     Republican                   Republican
2                     Republican                   Republican
3                     Republican                   Republican
4                     Republican                   Republican

  Vote_Data_Hillary_Clinton_Party Vote_Data_Jeb_Bush_Party  \
0                        Democrat               Republican
1                        Democrat               Republican
2                        Democrat               Republican
3                        Democrat               Republican
4                        Democrat               Republican

  Vote_Data_John_Kasich_Party  … Vote_Data_Marco_Rubio_Percent_of_Votes  \
0                  Republican  …                                0.322169
1                  Republican  …                                0.202552
2                  Republican  …                                0.169059
3                  Republican  …                                0.333333
4                  Republican  …                                0.397129

  Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                       0.0
1                                       0.0
2                                       0.0
3                                       0.0
4                                       0.0

  Vote_Data_Mike_Huckabee_Percent_of_Votes Vote_Data_No_Preference_Party  \
0                                      0.0                           0.0
```

```
1                                     0.0                           0.0
2                                     0.0                           0.0
3                                     0.0                           0.0
4                                     0.0                           0.0

  Vote_Data_No_Preference_Percent_of_Votes  \
0                                 0.000000
1                                 0.400000
2                                 0.000000
3                                 0.000000
4                                 0.066667

  Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                  0.0
1                                  0.0
2                                  0.0
3                                  0.0
4                                  0.0

   Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                       0.0
1                                       0.0
2                                       0.0
3                                       0.0
4                                       0.0

   Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                             0.304071                          0.0
1                             0.125954                          0.0
2                             0.486005                          0.0
3                             0.202290                          0.0
4                             0.110687                          0.0

   Vote_Data_Uncommitted_Percent_of_Votes
0                                     0.0
1                                     0.0
2                                     0.0
3                                     0.0
4                                     0.0

[5 rows x 34 columns]
```

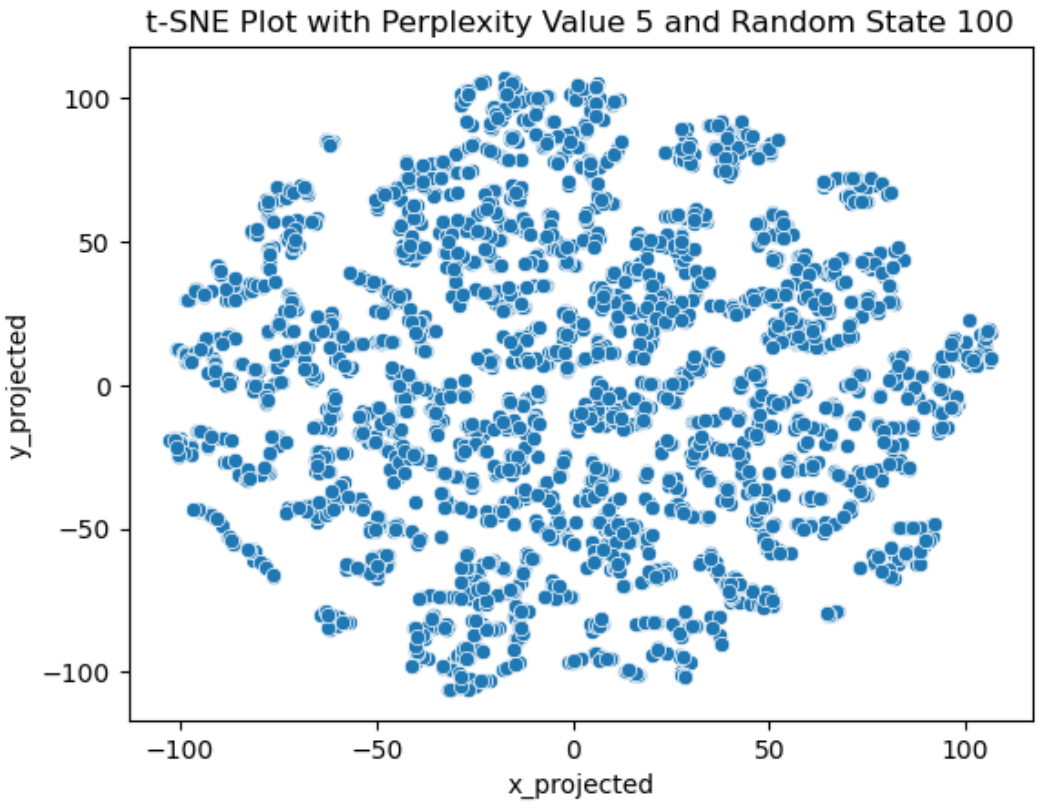
1.5 5. Clusterability and Clustering Structure Questions

[28]:
```
for perp in [5,10, 20, 30, 40, 50]:
    for rs in [50,100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', data=df_combo)
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('-----------------------------------------------')
```

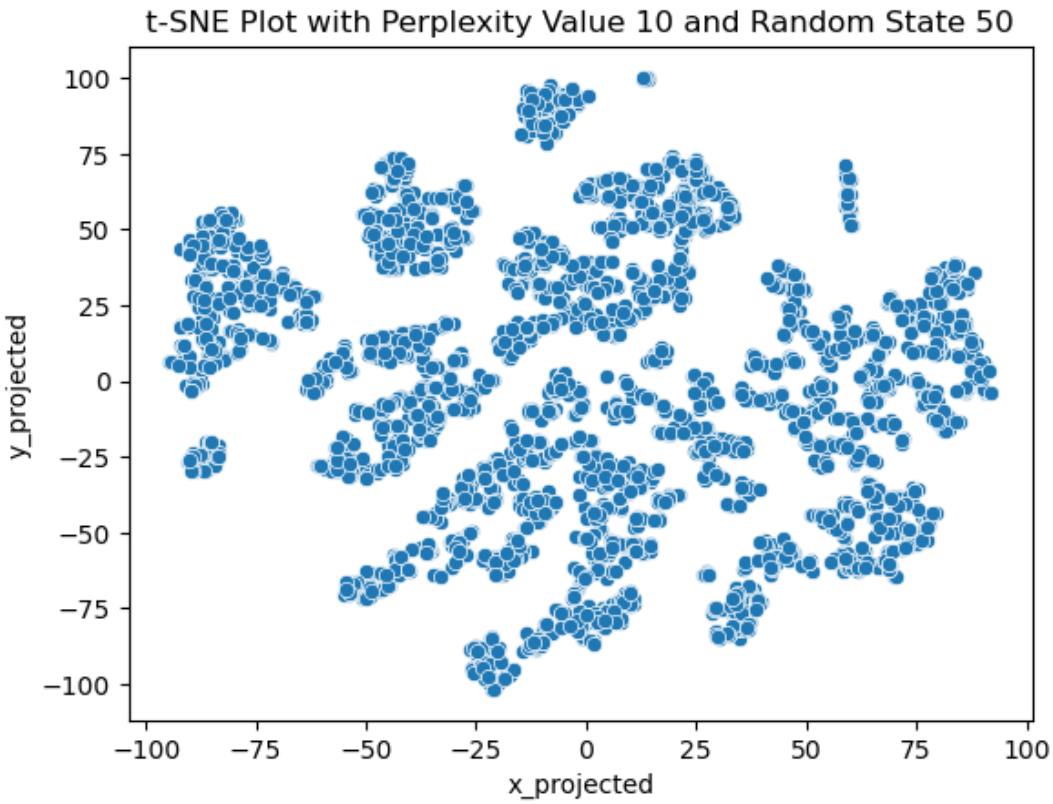
```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



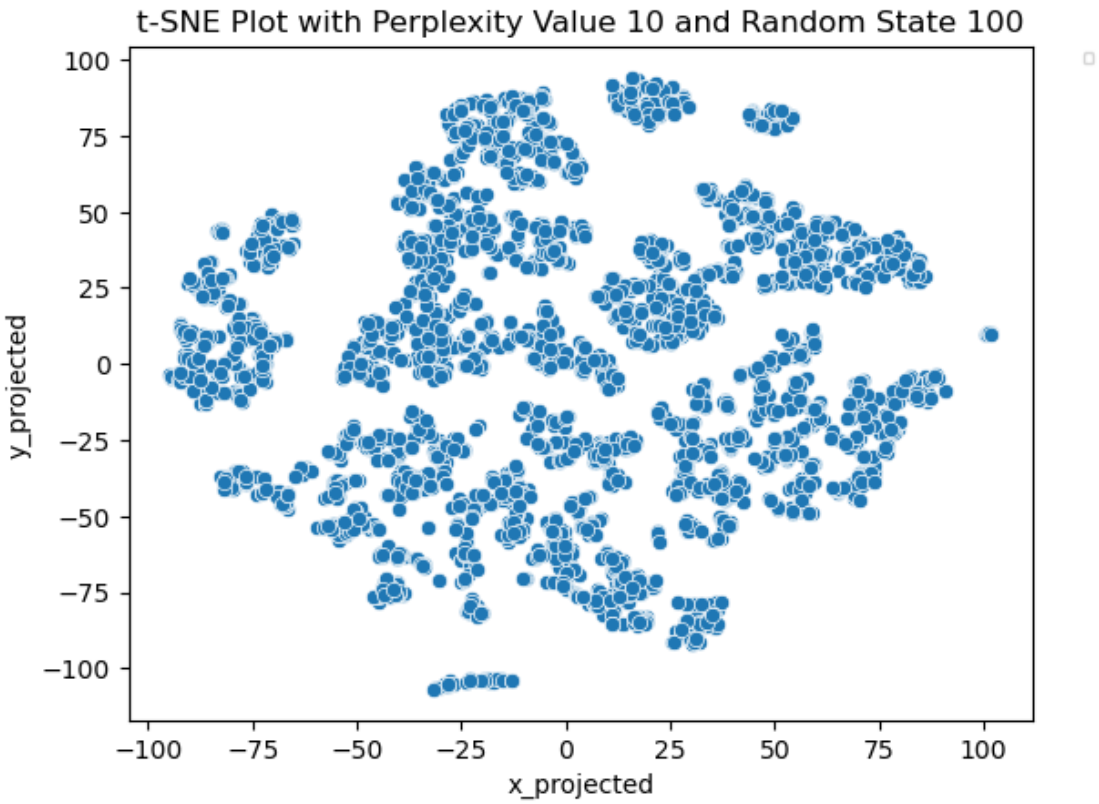

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



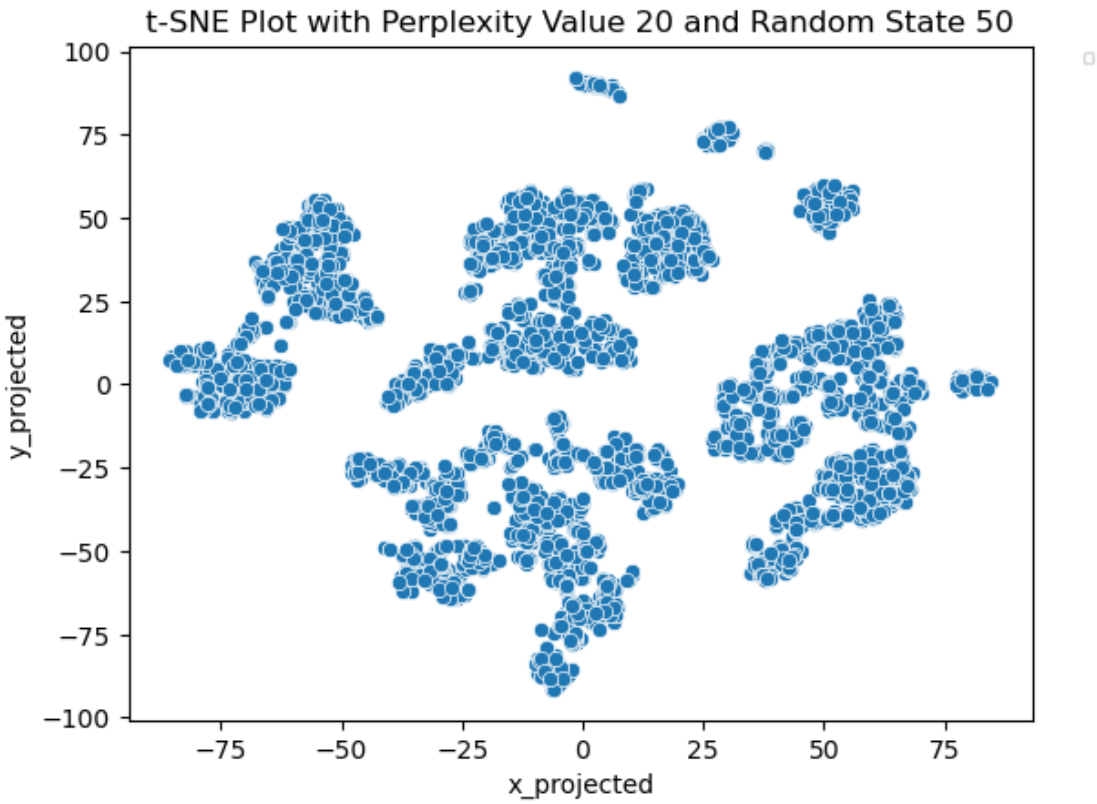

```
-------------------------------------------
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



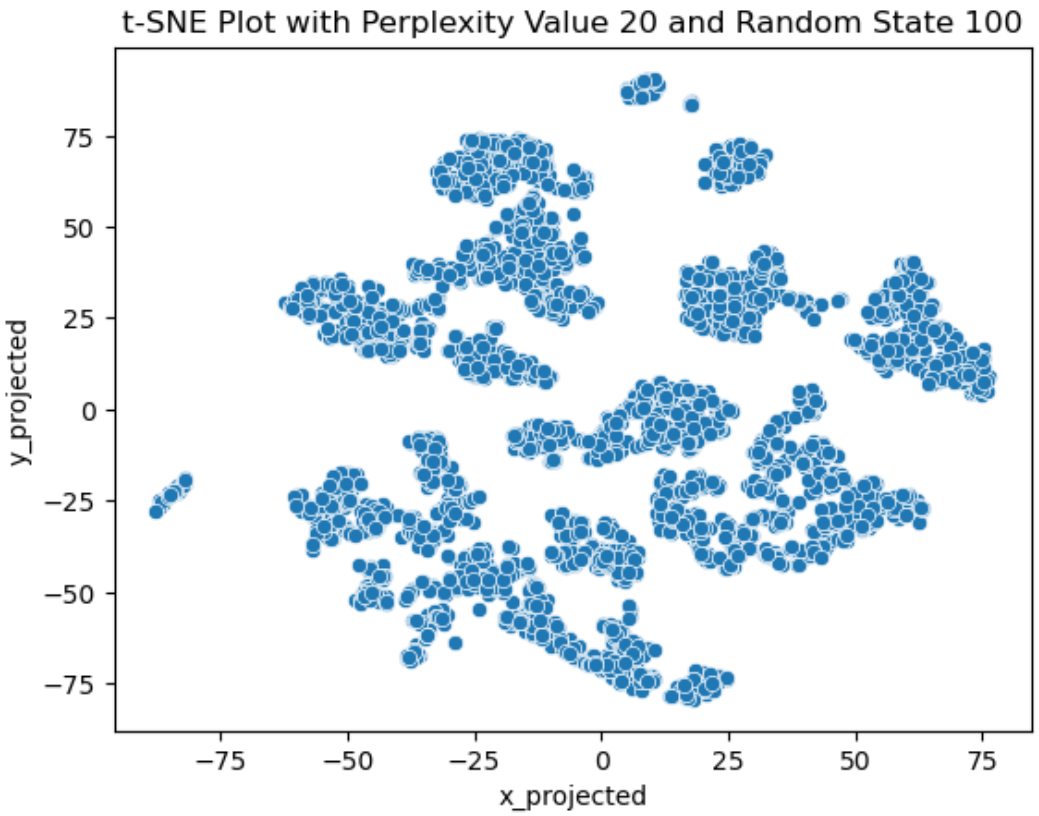

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



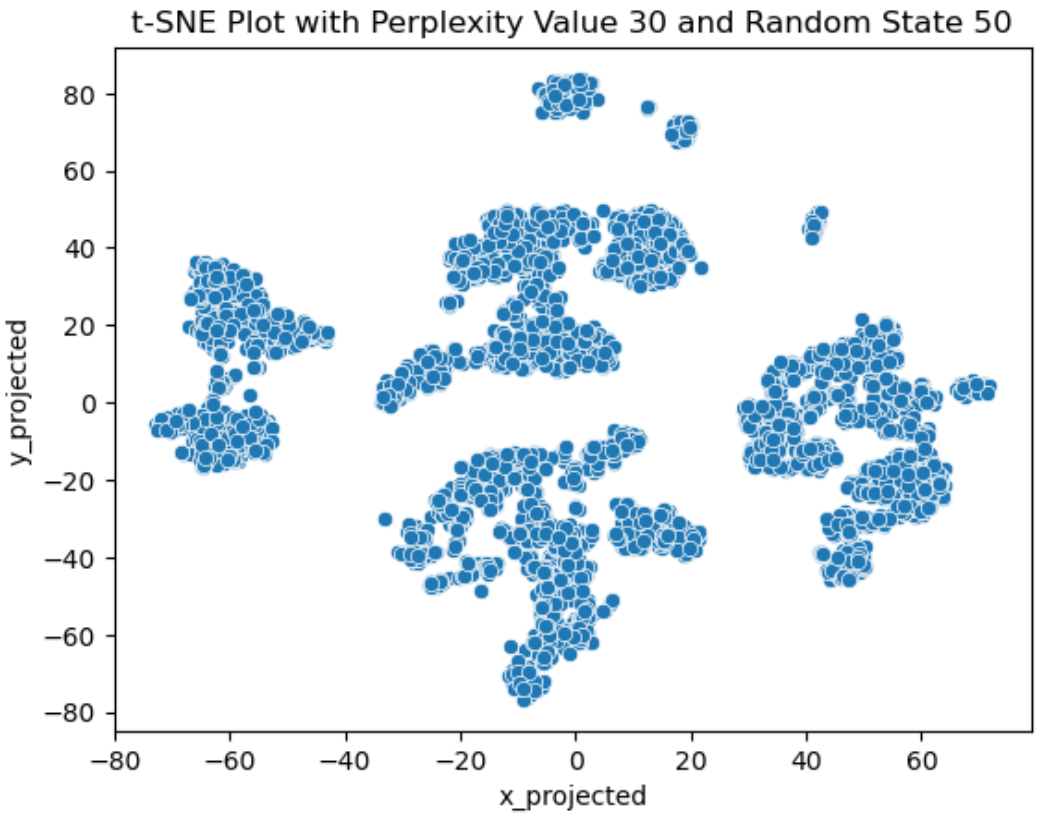

```
-----------------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



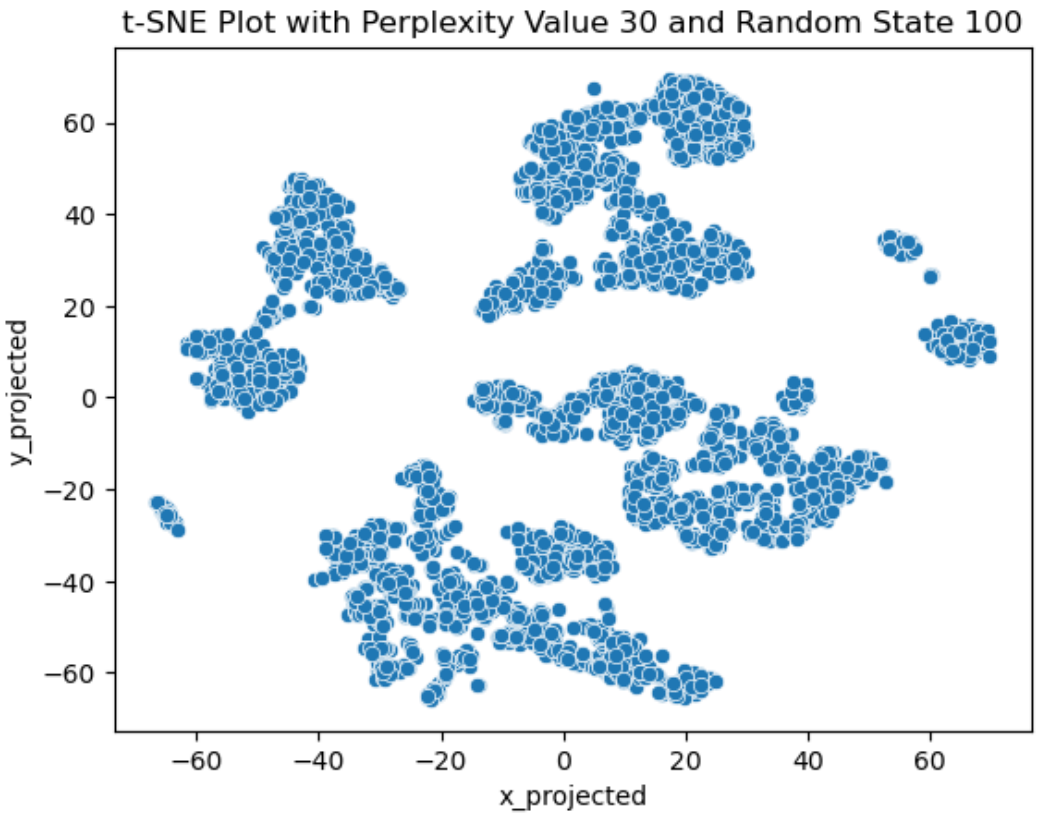

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



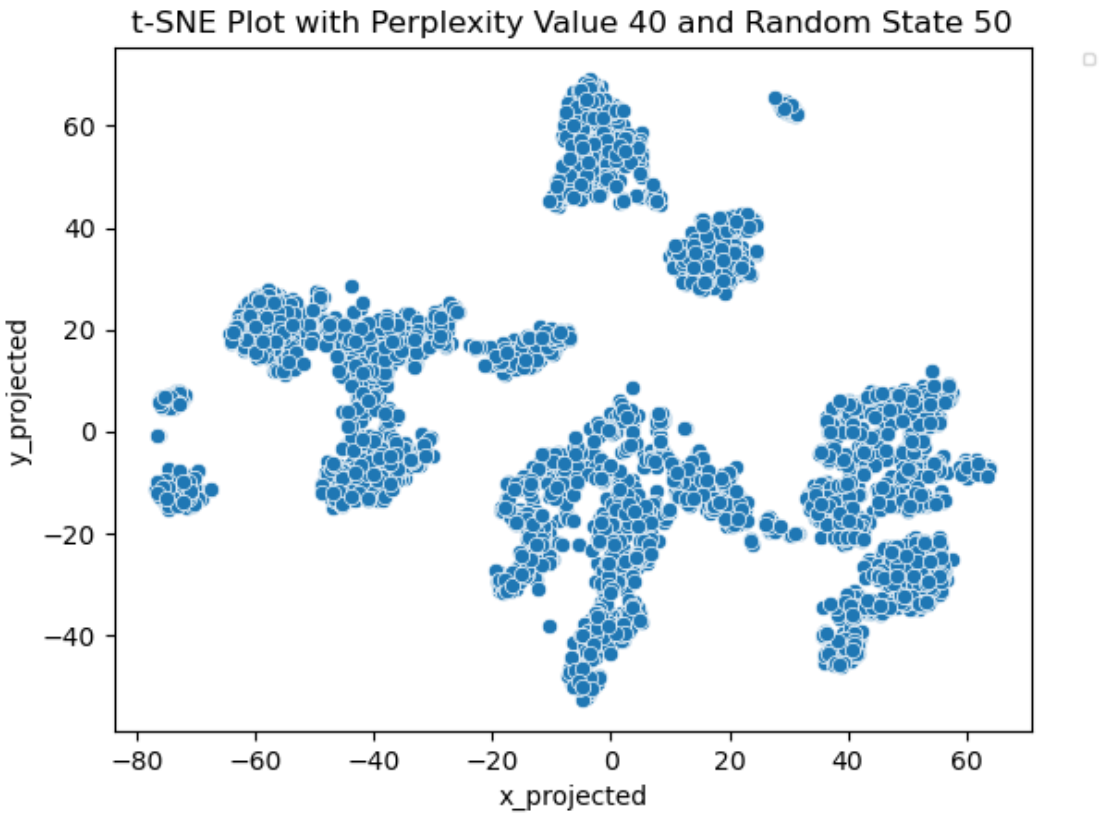

```
-------------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



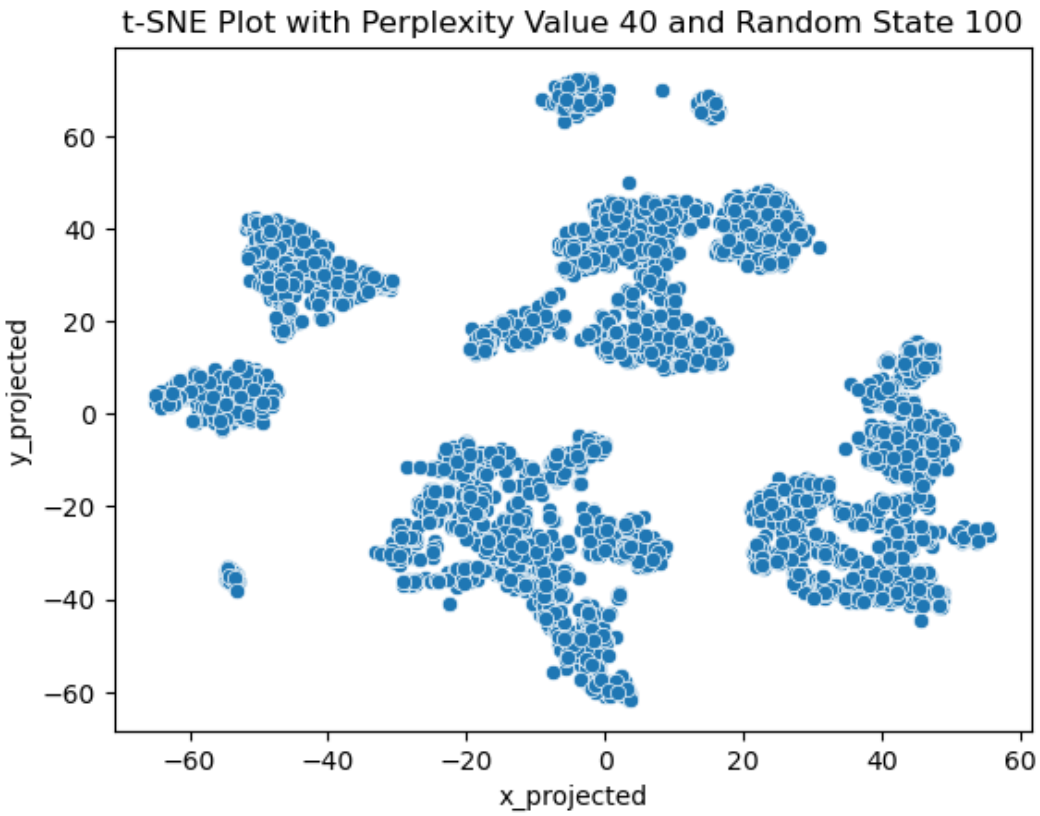

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



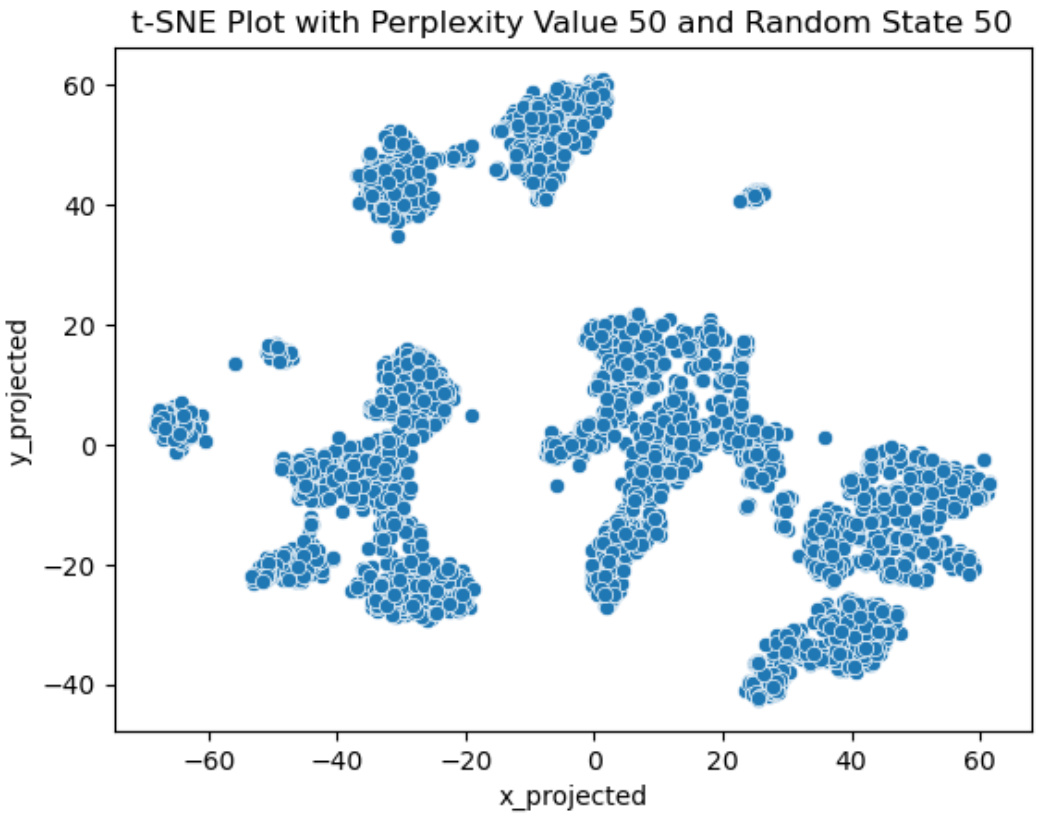

```
------------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



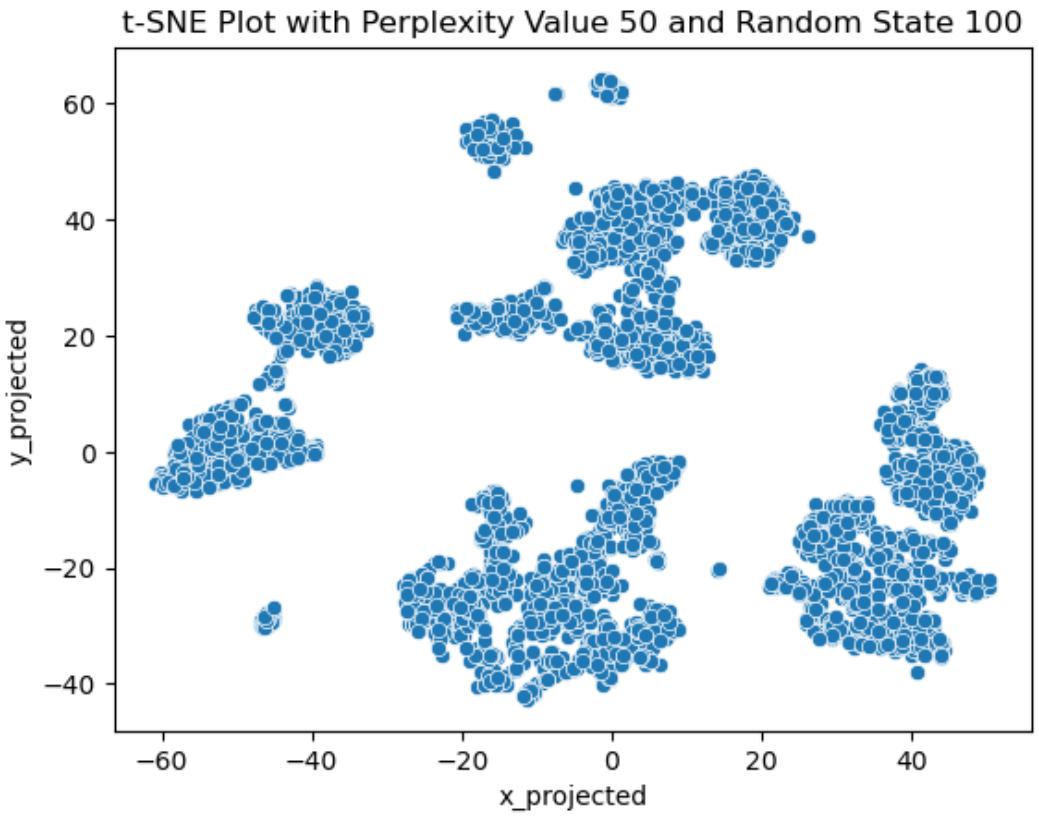

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```




```
-----------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```

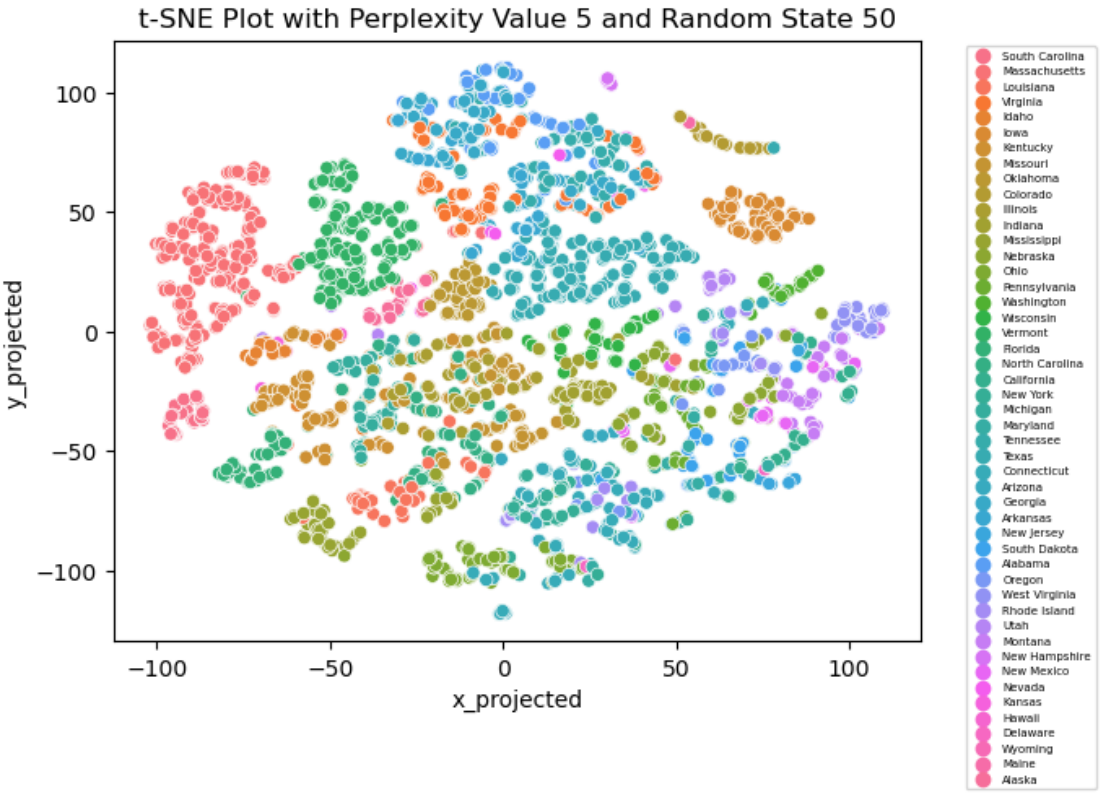


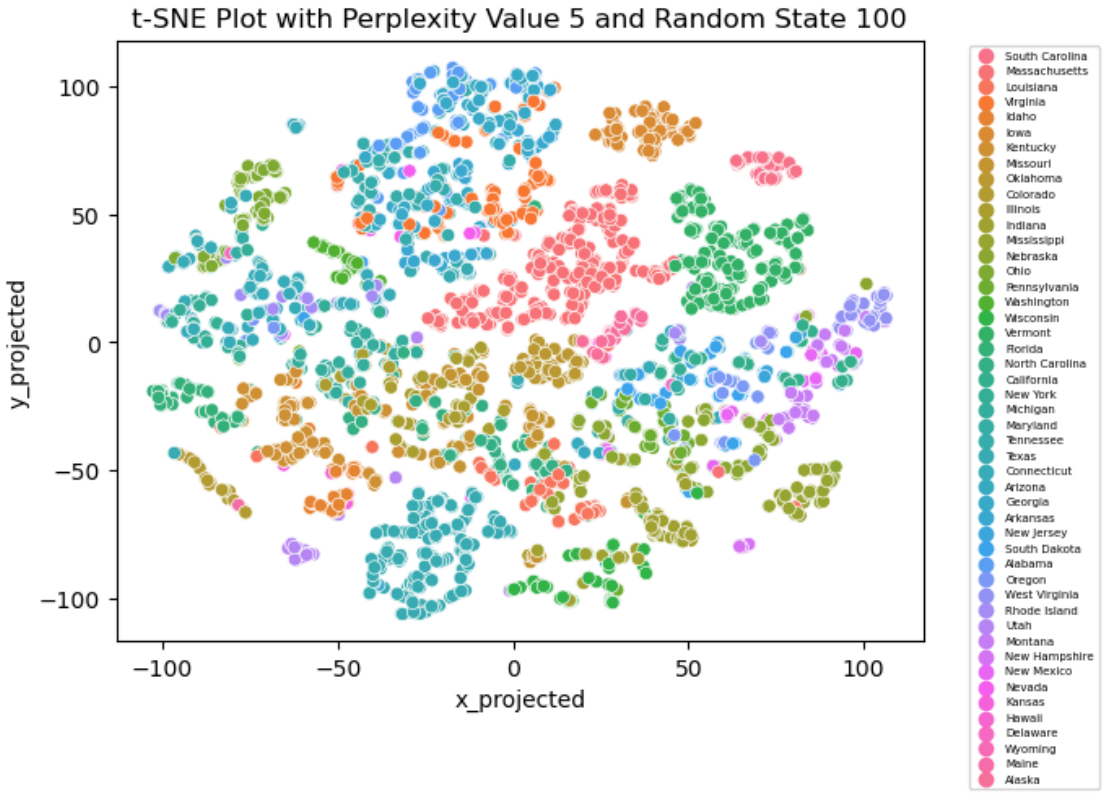

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```

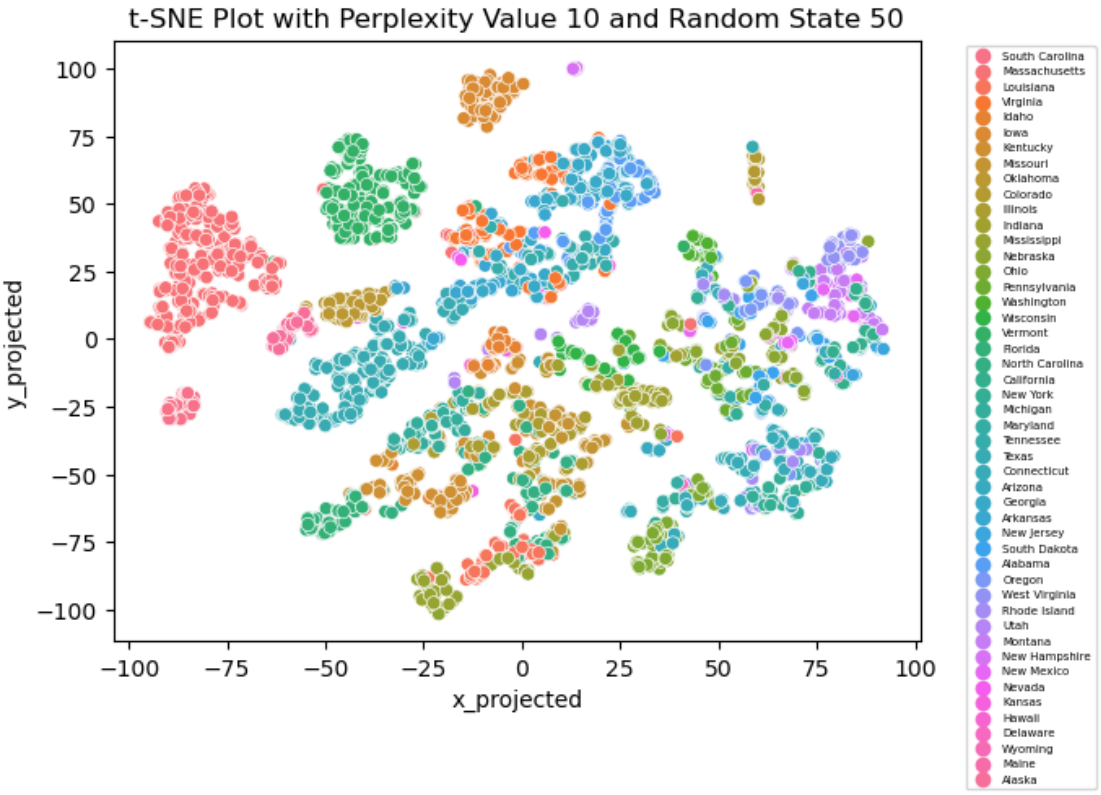


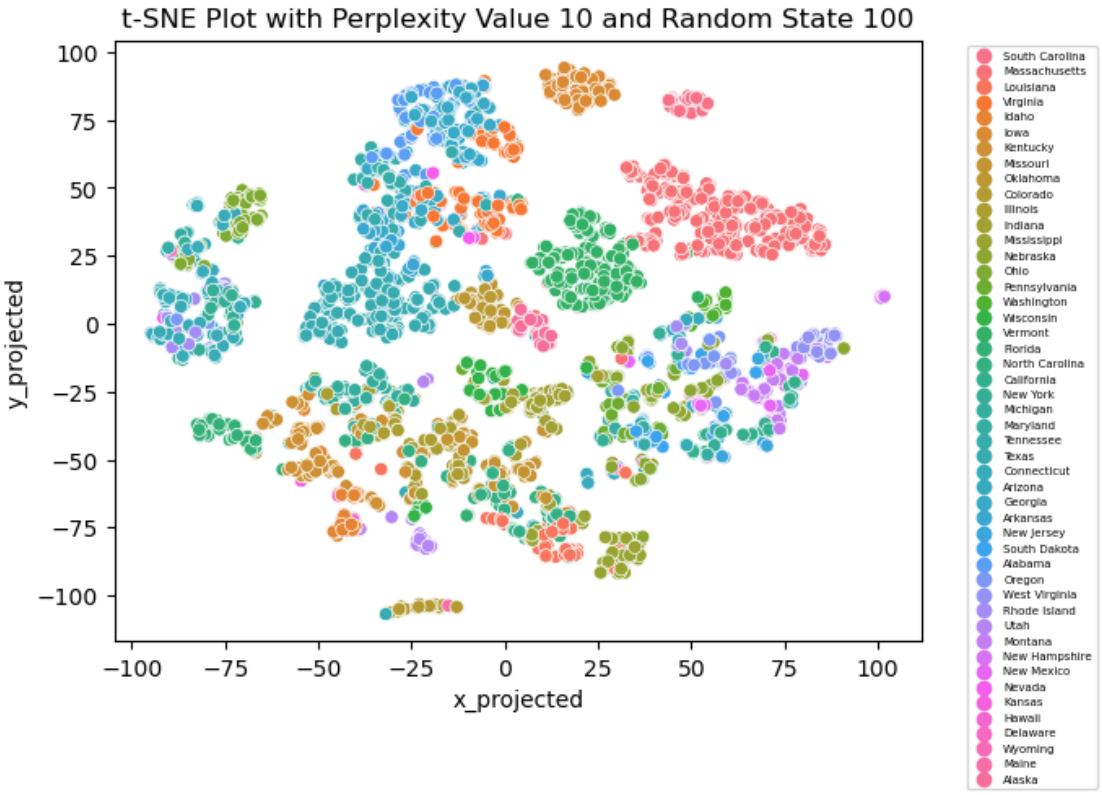
---

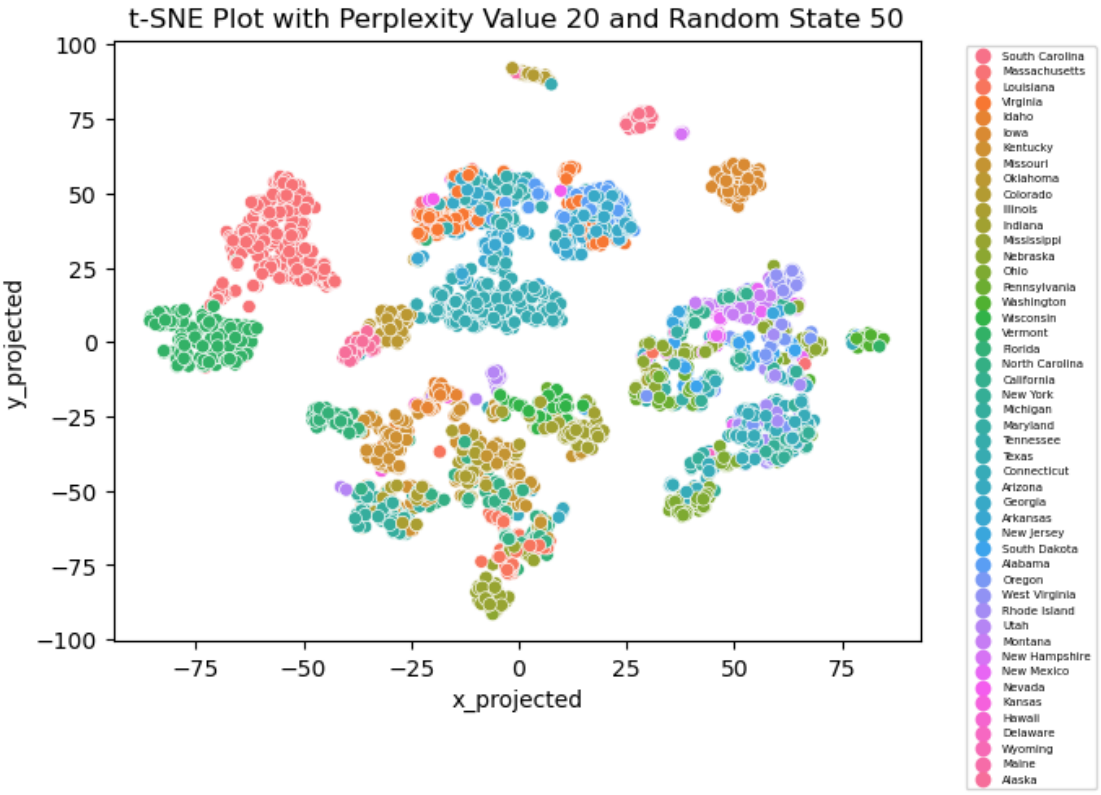
[29]:
```
# if we choose the location of state is the pre-assigned label and colorer it␣
 ↪in the t-sne plot

for perp in [5,10, 20, 30, 40, 50]:
    for rs in [50,100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
 ↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', hue =␣
 ↪"Location_State",data=df_combo)
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
 ↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('-----------------------------------------')
```

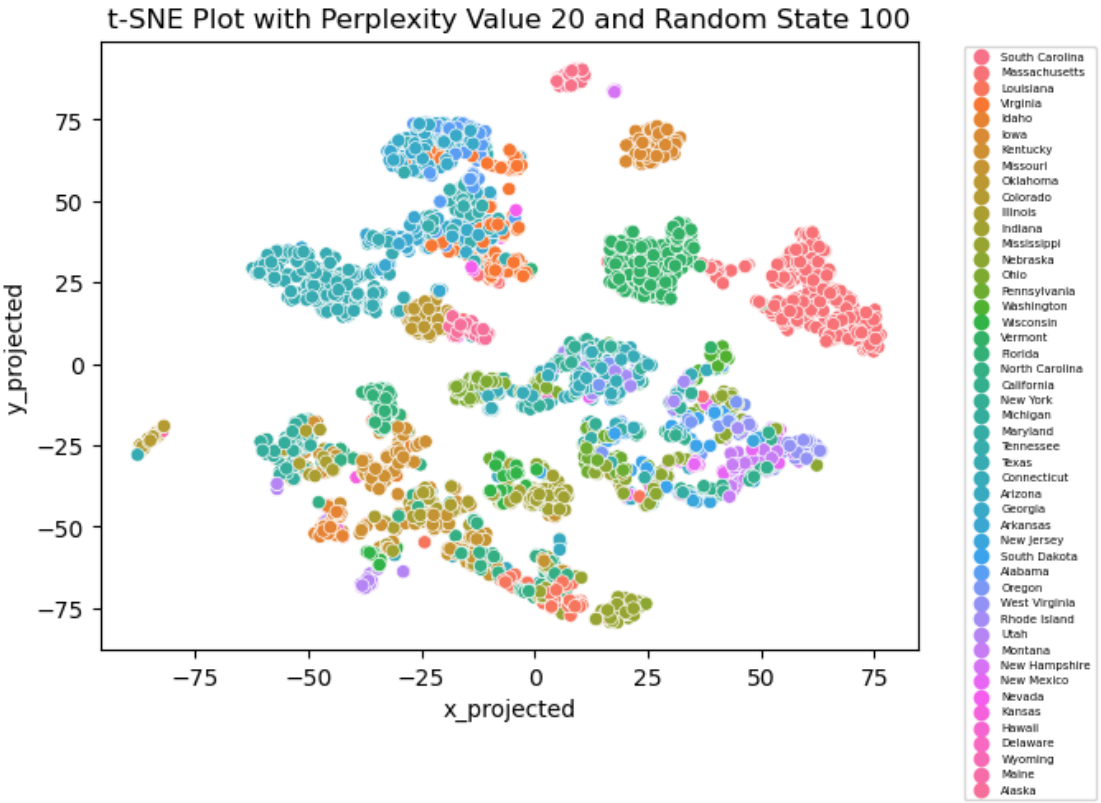

t-SNE Plot with Perplexity Value 5 and Random State 50
y_projected
x_projected
100
50
0
−50
−100
−100
−50
0
50
100
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

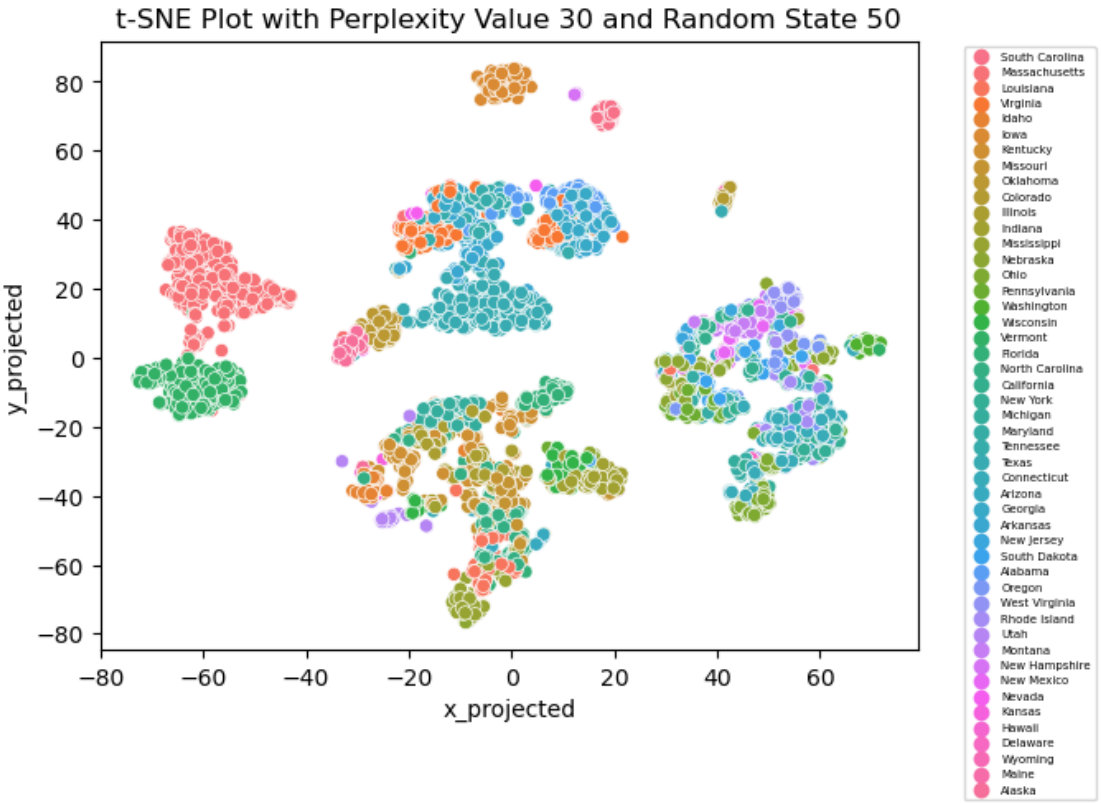

t-SNE Plot with Perplexity Value 5 and Random State 100
y_projected
x_projected
100
50
0
−50
−100
−100
−50
0
50
100
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

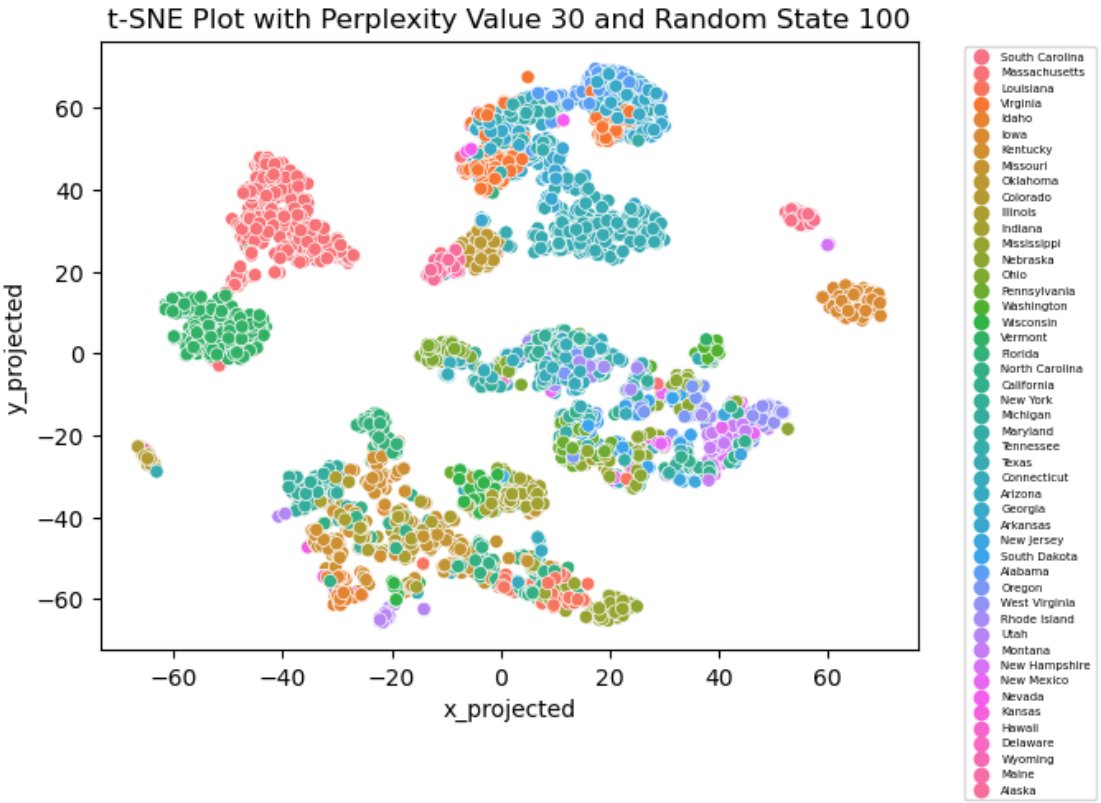

t-SNE Plot with Perplexity Value 10 and Random State 50
y_projected
x_projected
100
75
50
25
0
−25
−50
−75
−100
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

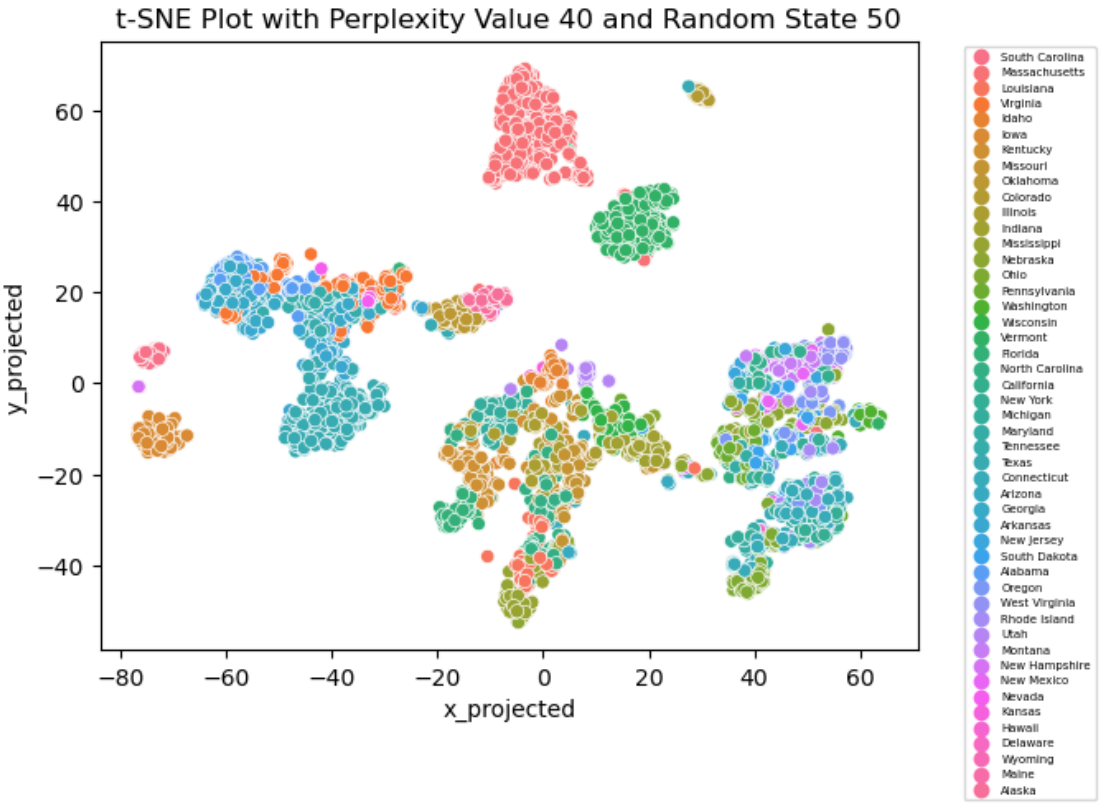

t-SNE Plot with Perplexity Value 10 and Random State 100
y_projected
x_projected
100
75
50
25
0
−25
−50
−75
−100
−100
−75
−50
−25
0
25
50
75
100
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

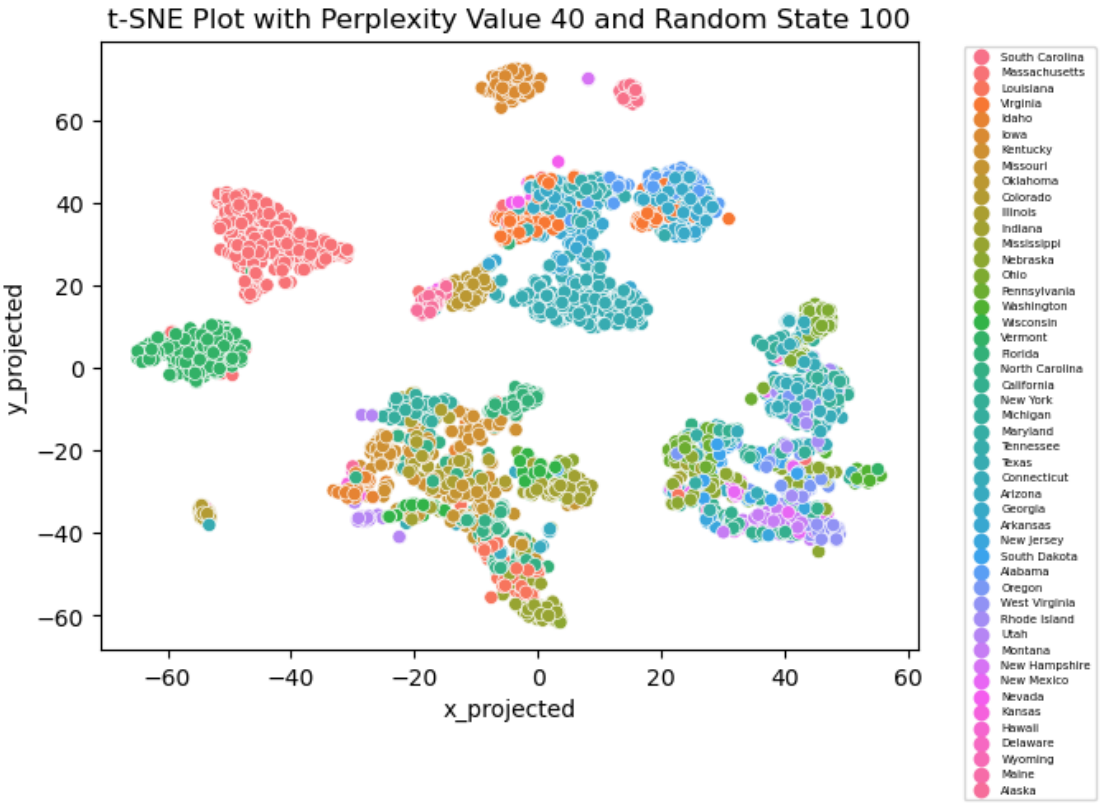

t-SNE Plot with Perplexity Value 20 and Random State 50
y_projected
x_projected
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

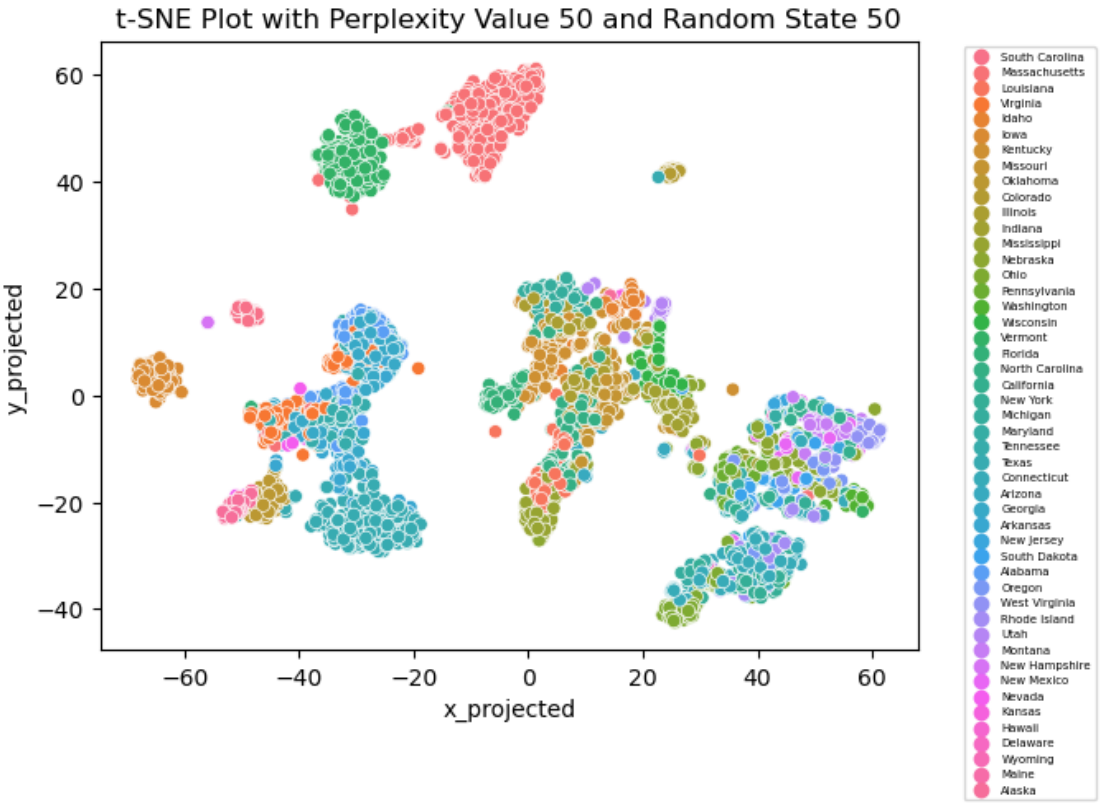

t-SNE Plot with Perplexity Value 20 and Random State 100
y_projected
x_projected
75
50
25
0
−25
−50
−75
−75
−50
−25
0
25
50
75
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

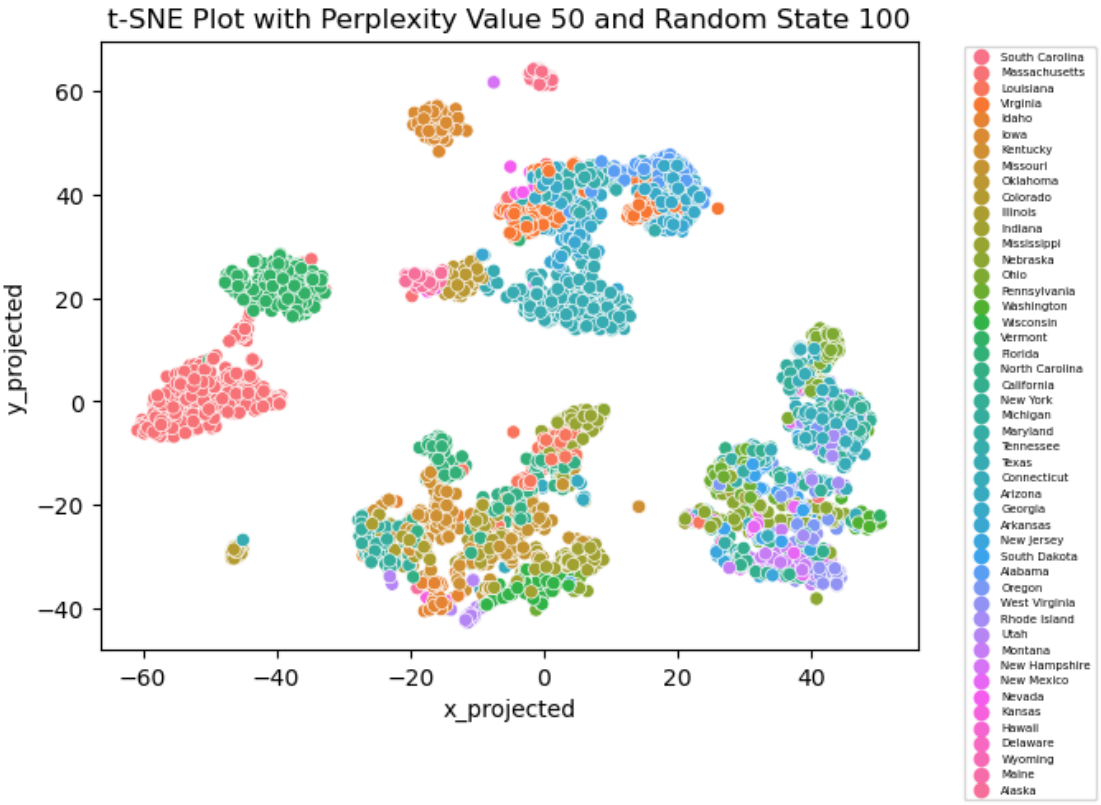

t-SNE Plot with Perplexity Value 30 and Random State 50
y_projected
x_projected
80
60
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska


t-SNE Plot with Perplexity Value 30 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska


t-SNE Plot with Perplexity Value 40 and Random State 50
y_projected
x_projected
60
40
20
0
−20
−40
−80
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska


t-SNE Plot with Perplexity Value 50 and Random State 50
y_projected
x_projected
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska



---

[30]:
```
num_trials=10
hopkins_stats=[]
for i in range(0,num_trials):
    n = len(X)
    p = int(0.1 * n)
    hopkins_stats.append(hopkins(X,p))
print(hopkins_stats)
```

```
[0.04070630045751869, 0.03929863065307637, 0.0397228717346078,
0.03817619119068465, 0.03718203368349326, 0.04121653438793144,
0.04139403559241479, 0.042184529046137245, 0.03970703705773819,
0.03955241995398281]
```

Because many of these Hopkins statistics are closer to 0 than they are to 0.5, the Hopkin's statistic suggests that the dataset is clusterable.

[31]:
```
X.shape
```

[31]: (3592, 18)

1.6 6. Algorithm Selection Motivation

We chose K-means and HAC algorithms

We perform clustering analysis for the pre-assigned label of location state, so the clustering algorithm is what we need to present. Secondly we throw out the ranks of number, so the data is presented as a percentage. The values of percentages float from 0-1, plus we removed the states with small numbers so it is not easy to have outliers.

The type variable only represents the party that the voter is running for, and given that the type variables for parties are in a repeating pattern, they are less meaningful for classification. We do not consider the analysis of type variables

By looking at our t-SNE plot, we can see that this dataset is clusterable. This is very suitable for using simple and practical clustering algorithms like k-means.

K-means is a commonly used unsupervised learning algorithm for clustering data points based on their similarity. K-means can be a useful algorithm for analyzing presidential election datasets, but its applicability depends on the specific structure and characteristics of the data and the selection of Other unsupervised learning algorithms such as hierarchical clustering, principal component analysis (PCA) and t-SNE may also be worth considering, depending on the specific structure and characteristics of the data and the selection of features used for clustering. Other unsupervised learning algorithms such as hierarchical clustering, principal component analysis (PCA) and t-SNE may also be worth considering, depending on the goals of the analysis and the nature of the data.

Hierarchical classification: This is a clustering algorithm that creates a hierarchy of clusters that can be visualized as a tree diagram, called a dendrogram. Hierarchical classification can be performed using either the aggregation method or the segmentation method. Aggregative clustering starts with each data point as its own cluster and then iteratively merges the most similar clusters until a single cluster is obtained that contains all data points. Split clustering, on the other hand, starts with all data points in one cluster and then iteratively divides that cluster into smaller clusters until each data point is in its own cluster. Hierarchical clustering is useful for identifying relationships between clusters, and for identifying subgroups within larger clusters.

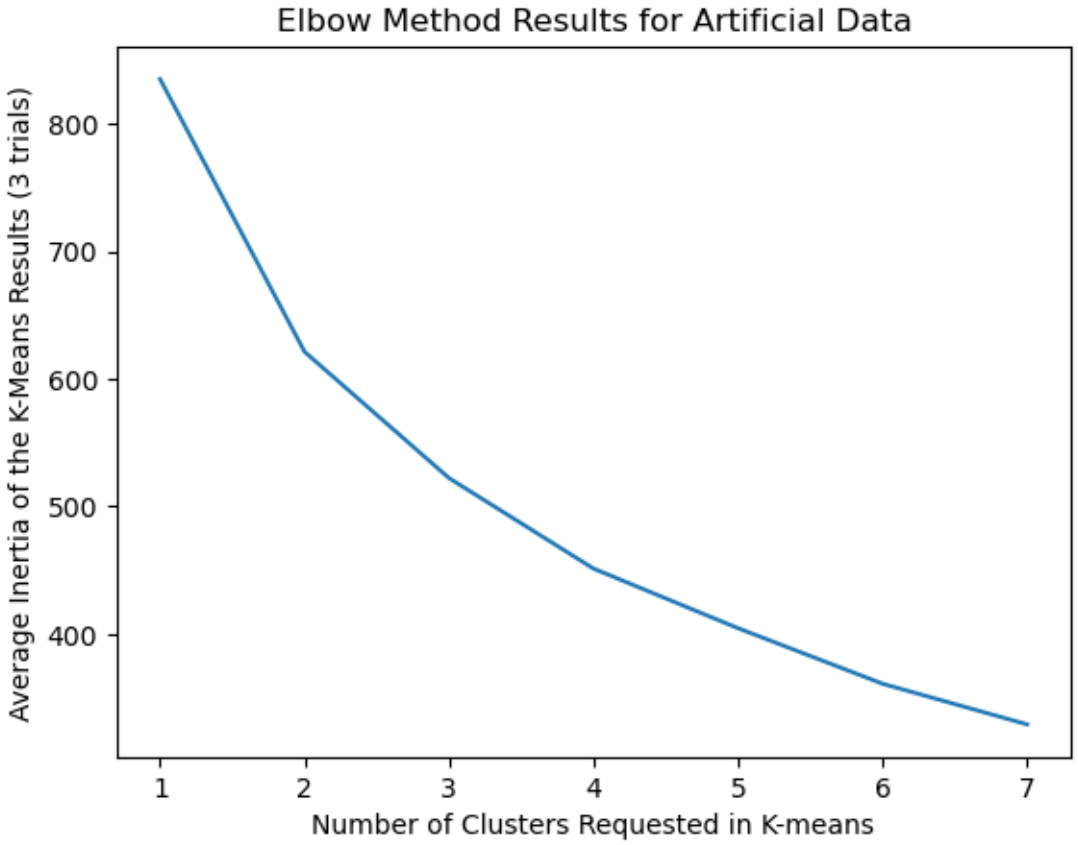
1.7 7. Clustering K-means

1.7.1 7.1. Parameter Selection

[47]:
```
cluster_num_list=range(1,8)
avg_inertia_list=[]
for k in cluster_num_list:
    print('k= '+str(k))
    sub_inertia_list=[]
    for i in range(0,3):
        kmeans=KMeans(n_clusters=k).fit(X)
        sub_inertia_list.append(kmeans.inertia_)
    avg_inertia_list.append(np.average(sub_inertia_list))

plt.plot(cluster_num_list,avg_inertia_list)
plt.xlabel('Number of Clusters Requested in K-means')
plt.ylabel('Average Inertia of the K-Means Results (3 trials)')
```

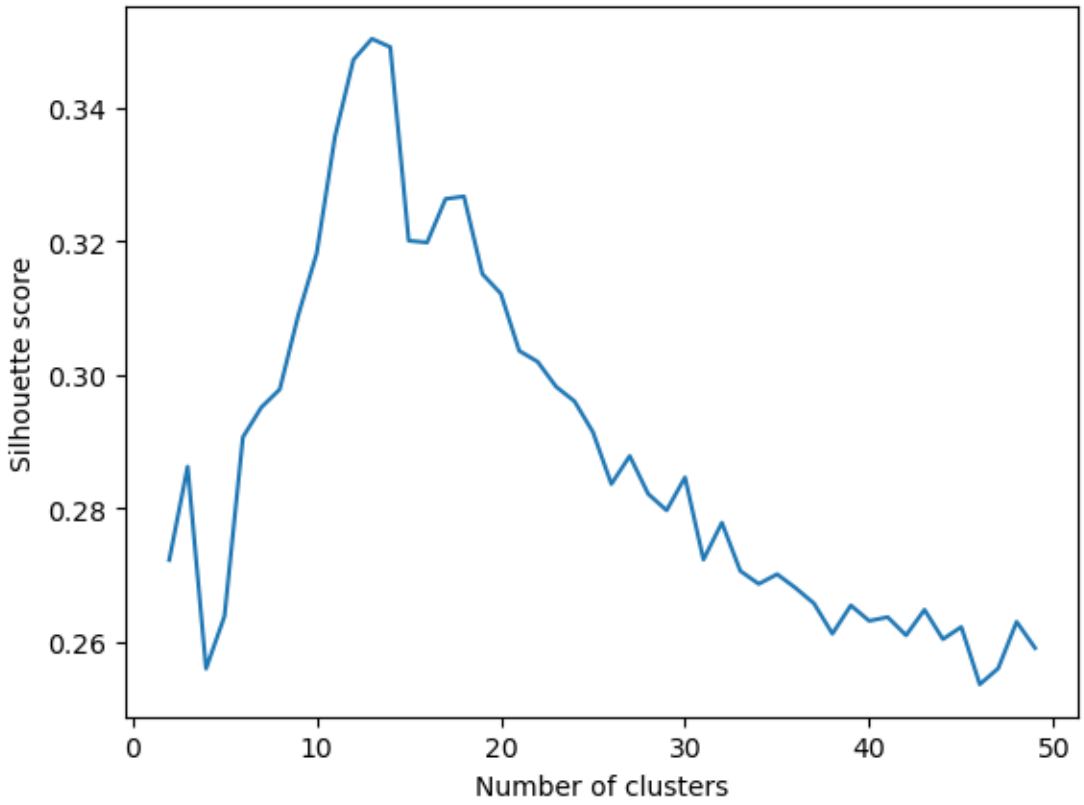
```
plt.title('Elbow Method Results for Artificial Data')
plt.show()
```

```
k= 1
k= 2
k= 3
k= 4
k= 5
k= 6
k= 7
```



[120]:
```
%matplotlib inline
silhouette_scores = []
for k in range(2, 50):
    kmeans = KMeans(n_clusters=k, random_state=0)
    labels = kmeans.fit_predict(X)
    score = silhouette_score(X, labels)
    silhouette_scores.append(score)

#
```

```
plt.plot(range(2, 50), silhouette_scores)
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette score')
plt.show()
```



We are choosing Silhouette score at around 0.35, which cluster K = 13. Because we can see the elbow plot didn't give us a good vision of the cluster. Thus, it is reasonable to choose k = 13.

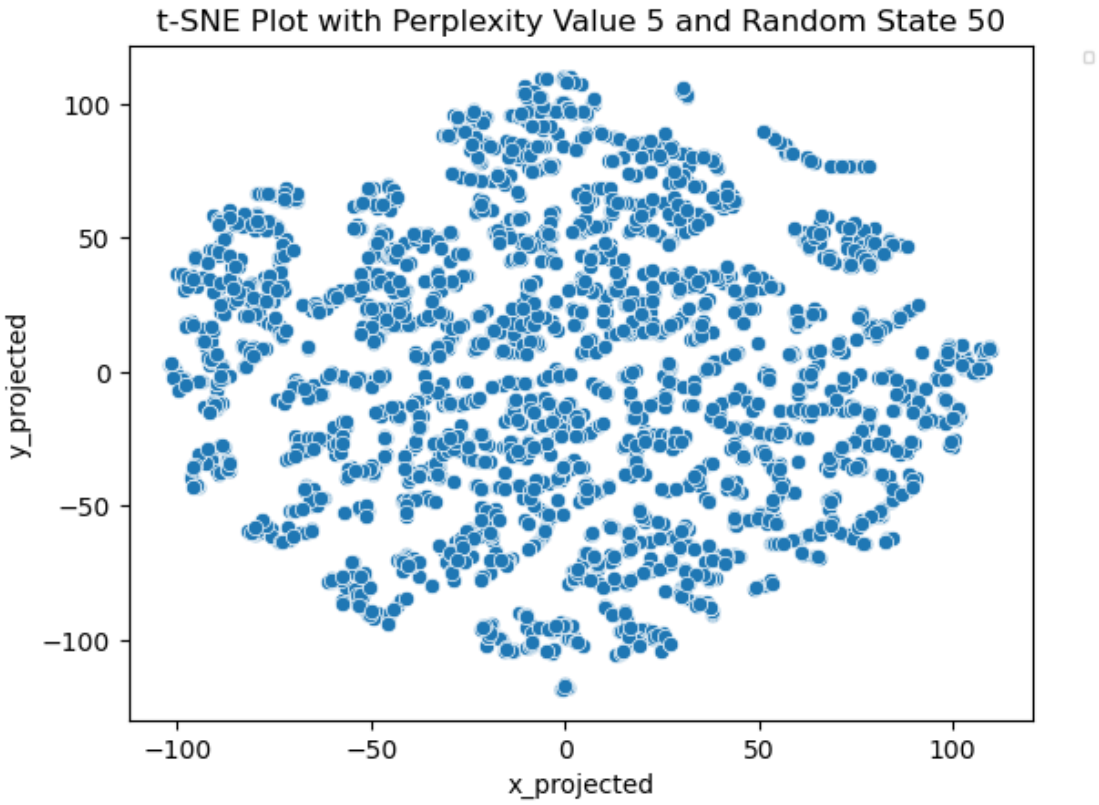
1.7.2 7.2. Clustering Algorithm

Recall the t-SNE plot, we find the perp with 40 and random state with 100, it has a pretty good clusterable structure.

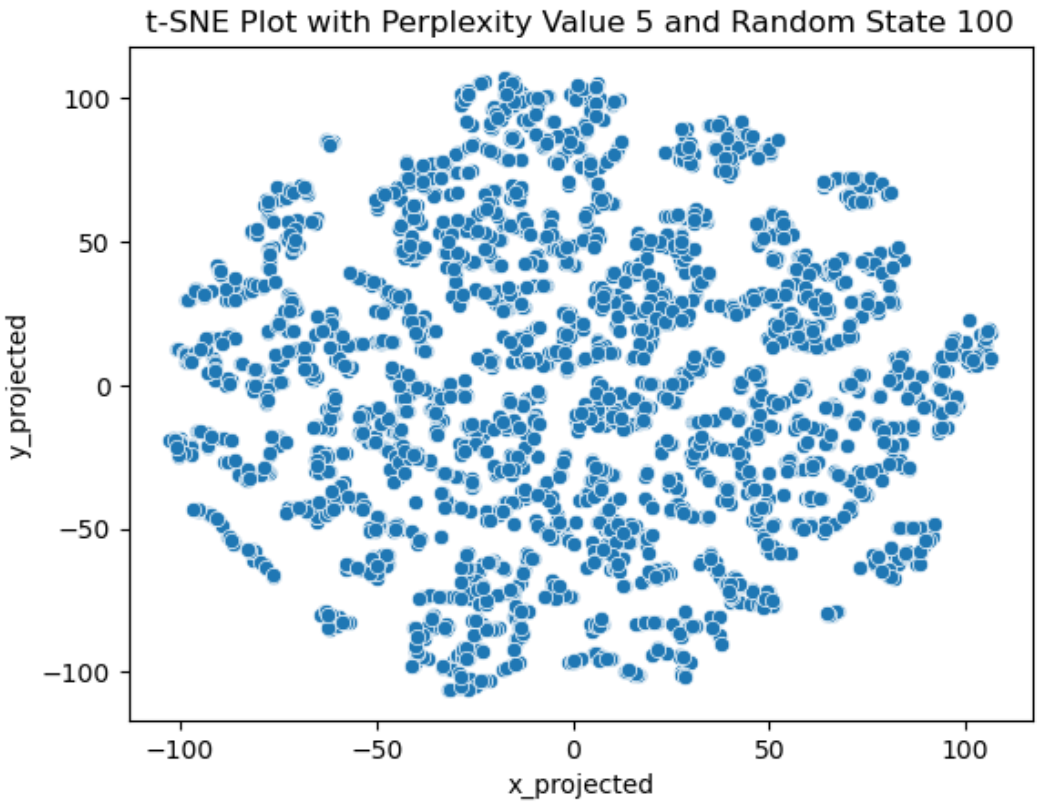
[49]:
```
for perp in [5,10, 20, 30, 40, 50]:
    for rs in [50,100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
 ↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', data=df_combo)
```

```
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('-----------------------------------------------')
```

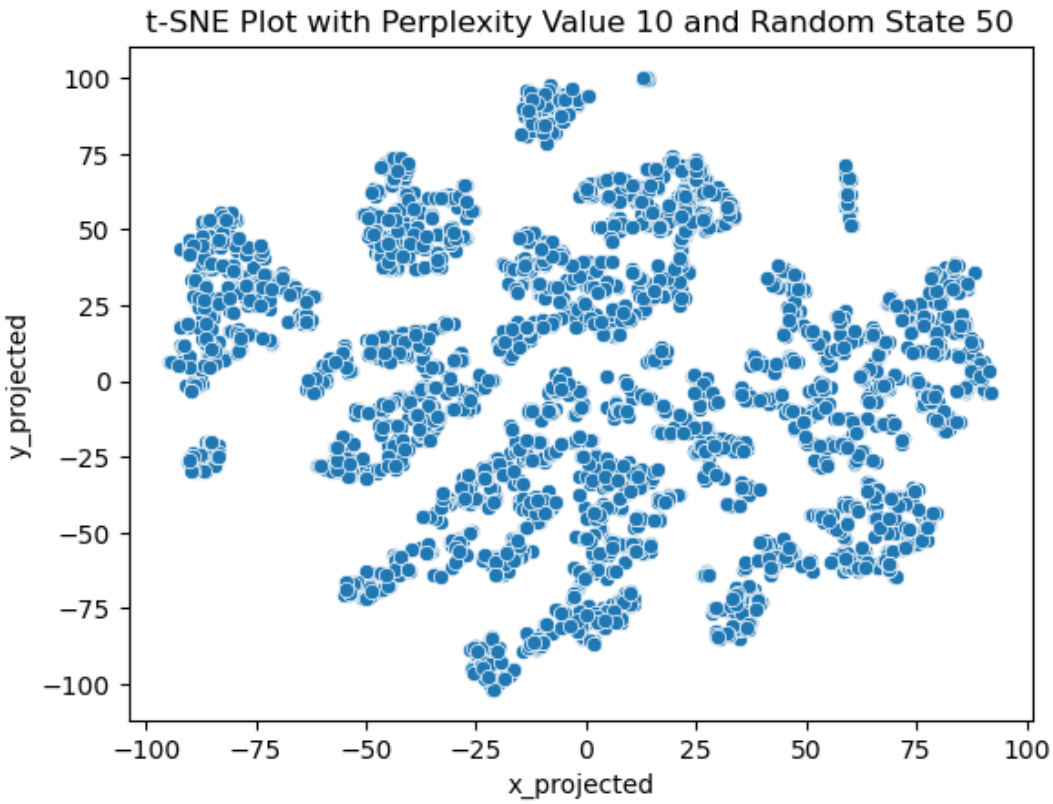
```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



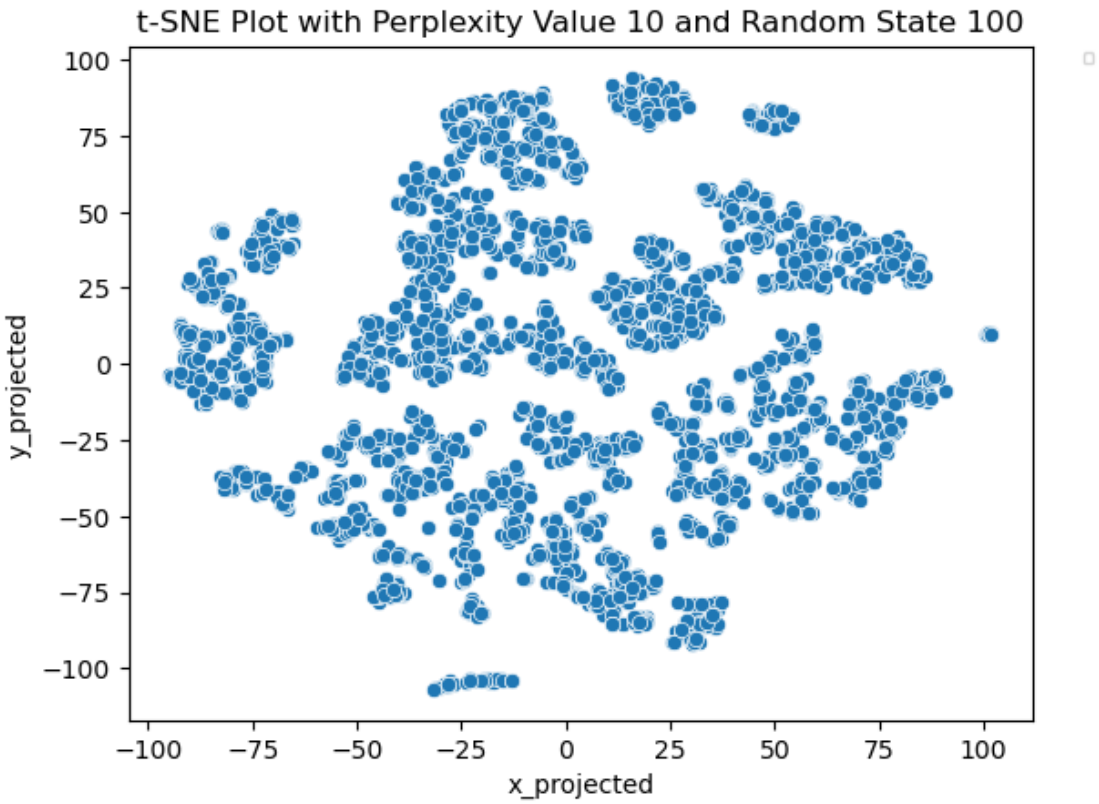

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



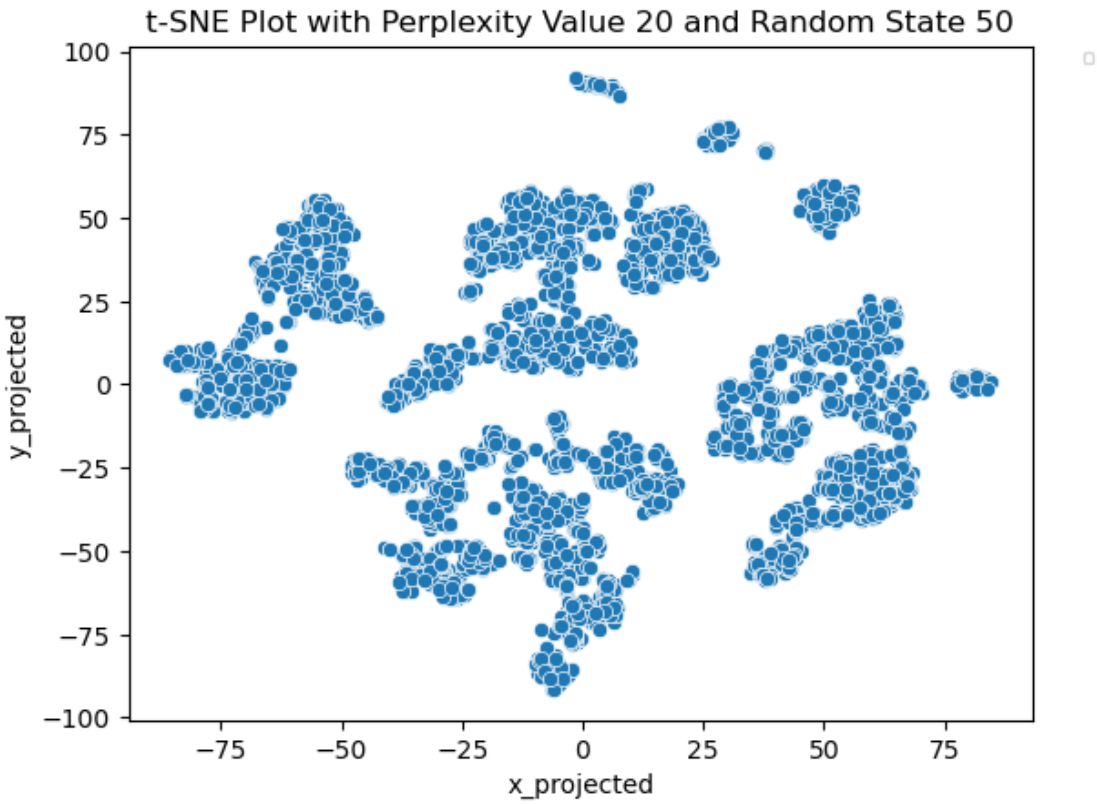

```
-------------------------------------------
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



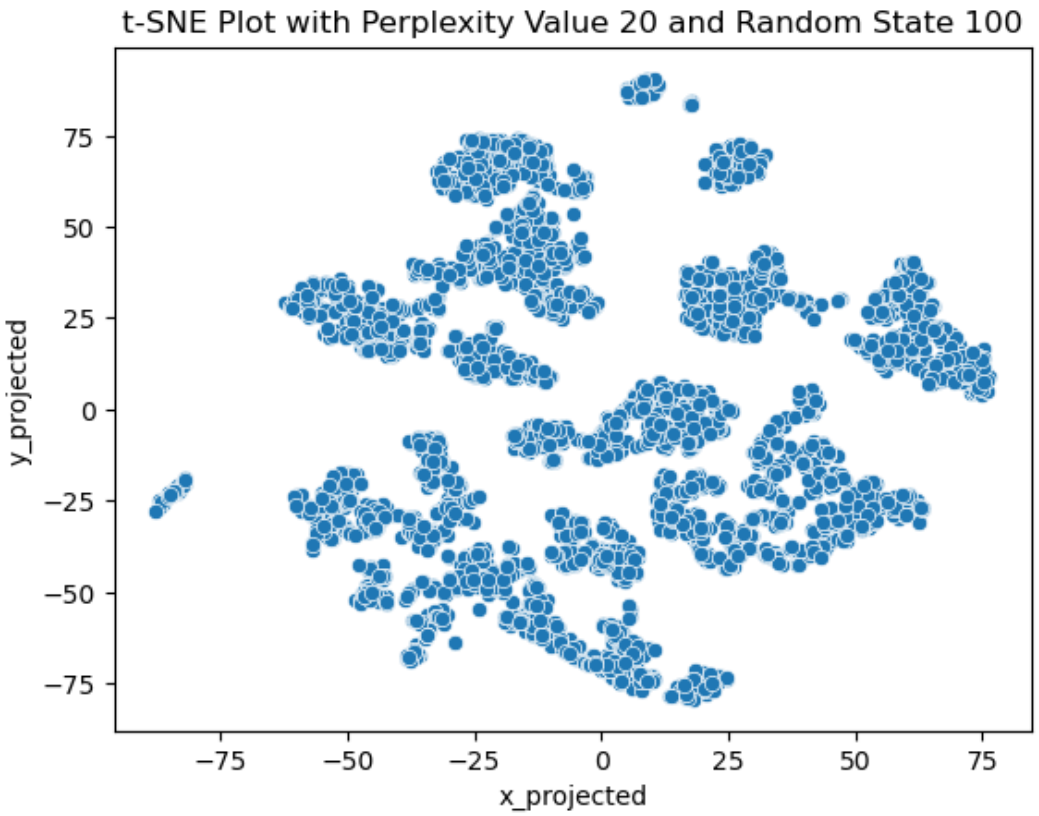

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



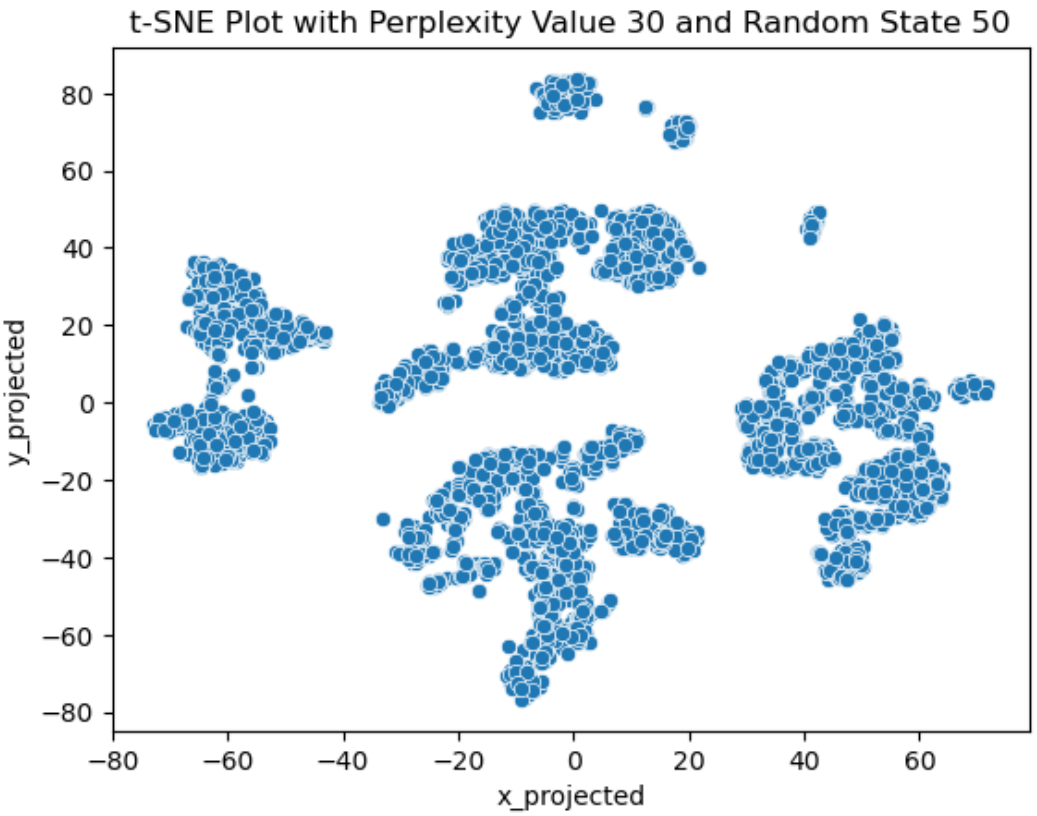

```
-----------------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



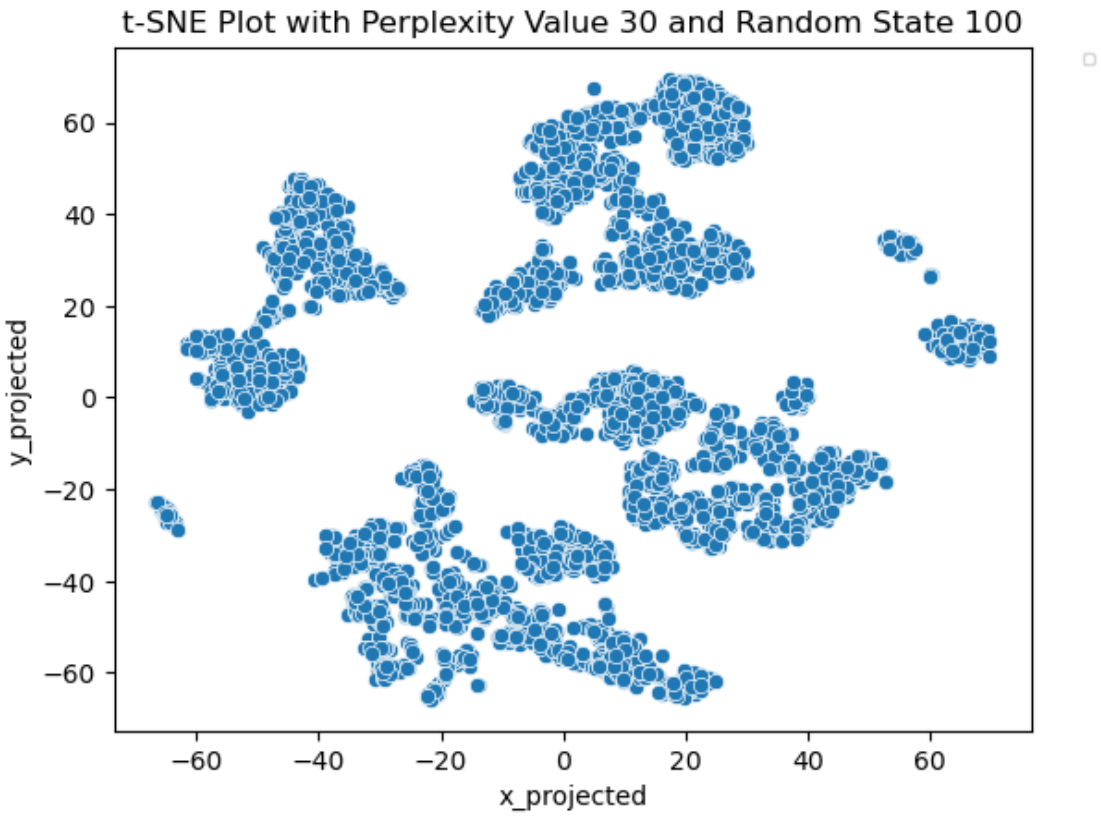

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



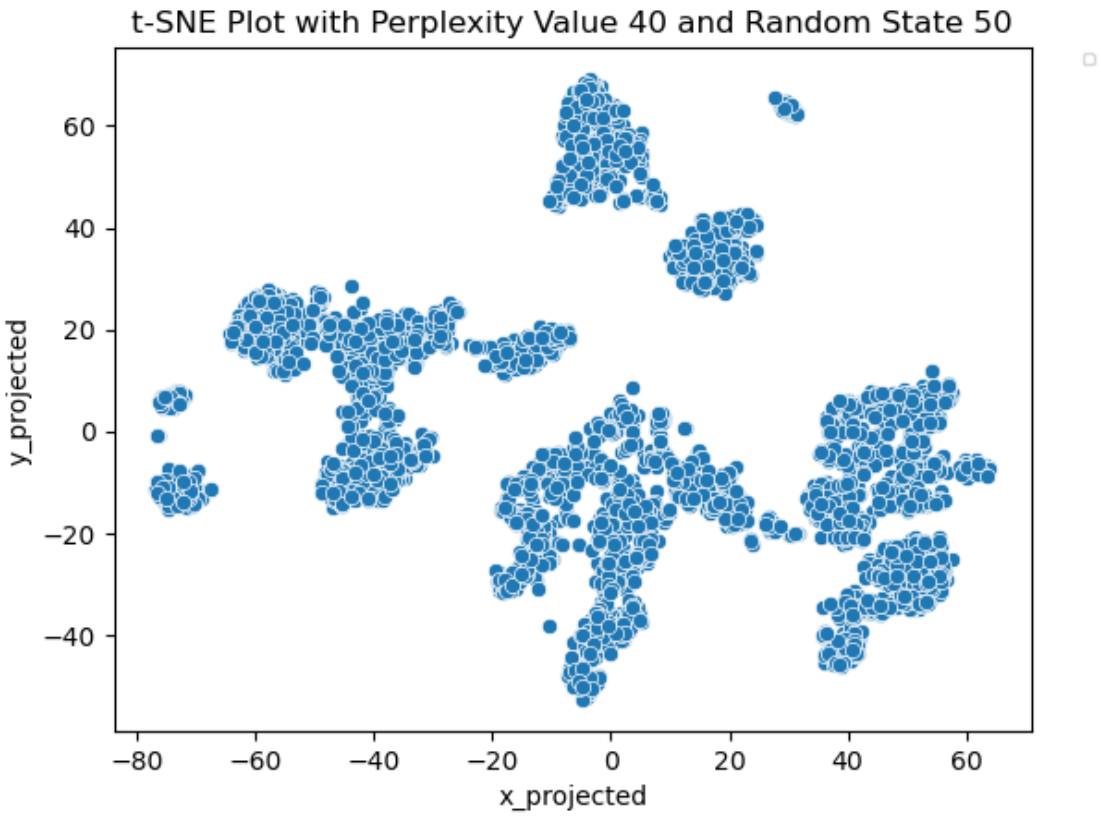

```
-----------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



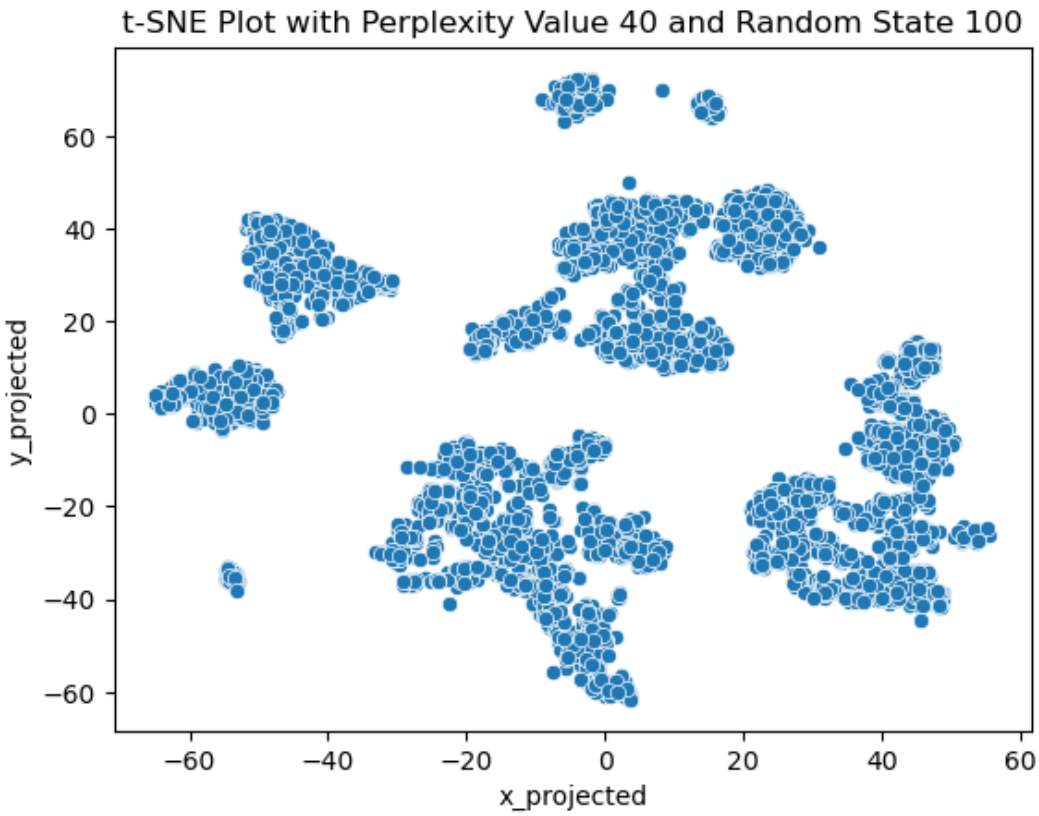

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



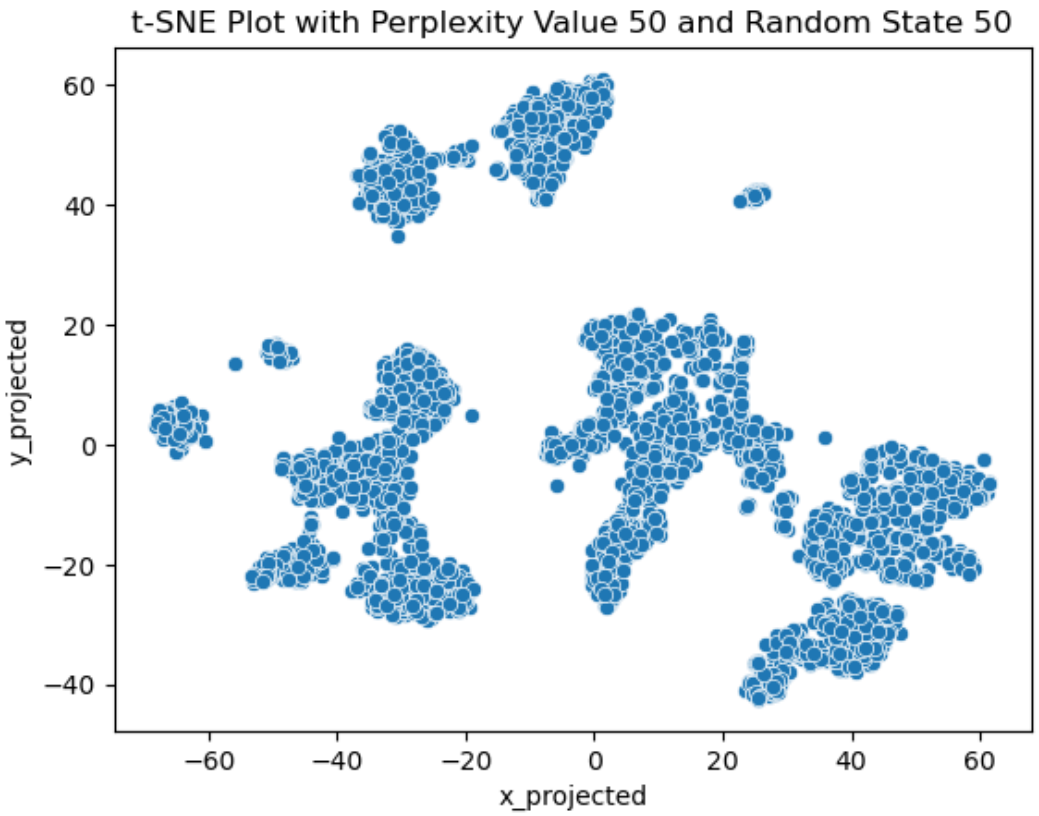

```
-----------------------------------------

No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



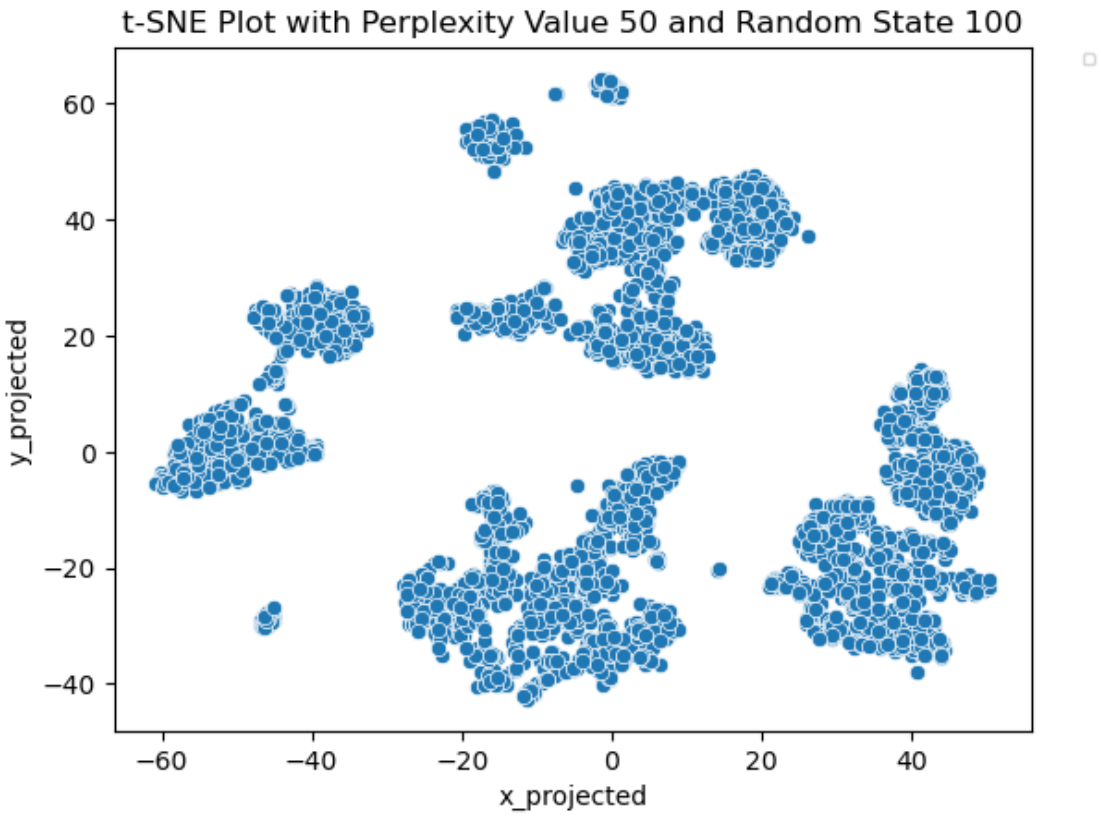

```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```




```
------------------------------------------
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```




```
No artists with labels found to put in legend.  Note that artists whose label
start with an underscore are ignored when legend() is called with no argument.
```



From the 7.1, we've already decided to use parameters of k = 13. And the random state = 100 may let the data has more clear cluster sturcture. So we would use them to perform k-means below.

[50]:
```
kmeans=KMeans(n_clusters=13).fit(X)
labels = kmeans.fit_predict(X)
df_kmeans_label = X.copy()
df_kmeans_label['label'] = labels
df_kmeans_label
```

[50]:
```
      Vote_Data_Ben_Carson_Percent_of_Votes  \
0                                  0.382488
1                                  0.115207
2                                  0.000000
3                                  0.437788
4                                  0.105991
…                                       …
3587                               0.000000
3588                               0.000000
3589                               0.216590
3590                               0.000000
```

```
3591                                0.000000

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                         0.170
1                                         0.534
2                                         0.314
3                                         0.274
4                                         0.468
…                                             …
3587                                      0.511
3588                                      0.315
3589                                      0.236
3590                                      0.181
3591                                      0.589

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                            …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                             …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
3591                                        0.0

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    0.403279
1                                    0.632787
2                                    0.486339
3                                    0.523497
4                                    0.308197
```

```
…                                       …
3587                                 0.904918
3588                                 0.540984
3589                                 0.431694
3590                                 0.000000
3591                                 0.750820

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                          0.818
1                                          0.446
2                                          0.537
3                                          0.720
4                                          0.528
…                                          …
3587                                       0.464
3588                                       0.637
3589                                       0.678
3590                                       0.755
3591                                       0.411

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                0.528926
1                                0.000000
2                                0.000000
3                                0.000000
4                                0.000000
…                                   …
3587                             0.000000
3588                             0.000000
3589                             0.000000
3590                             0.000000
3591                             0.000000

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                   0.067293
1                                   0.222222
2                                   0.053208
3                                   0.076682
4                                   0.508607
…                                      …
3587                                0.079812
3588                                0.092332
3589                                0.000000
3590                                0.000000
3591                                0.150235

      Vote_Data_Marco_Rubio_Percent_of_Votes  \
```

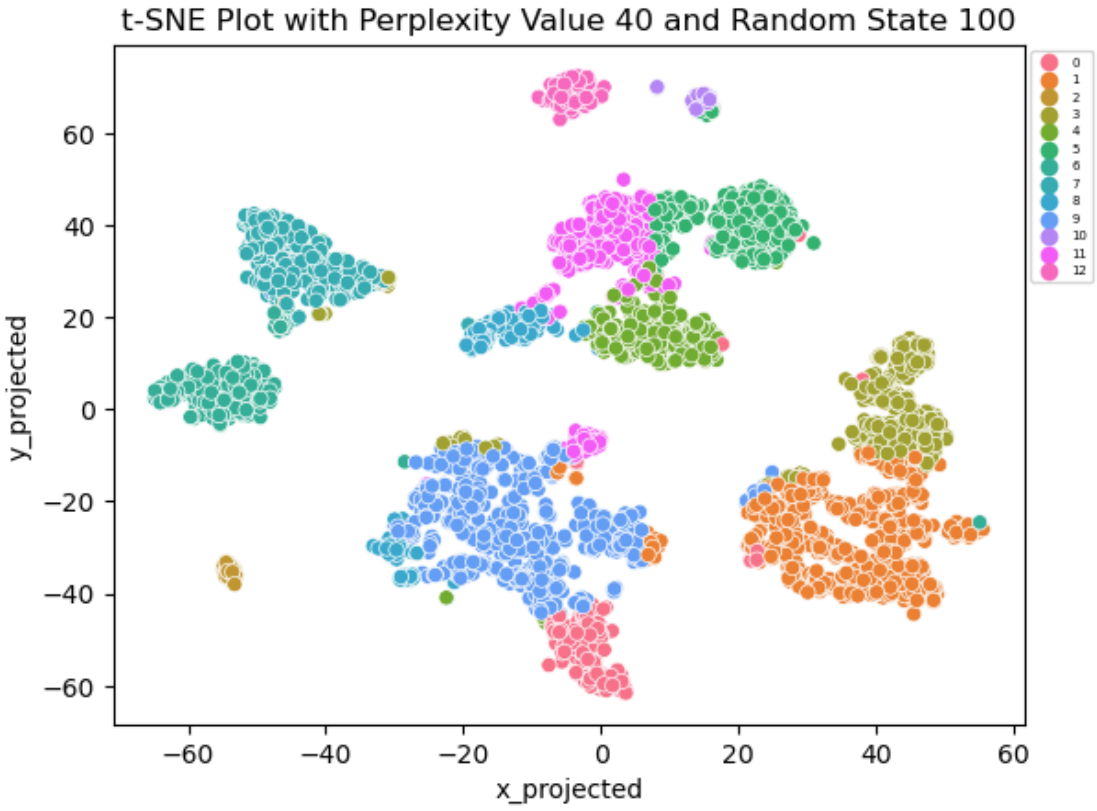
```
0                                    0.322169
1                                    0.202552
2                                    0.169059
3                                    0.333333
4                                    0.397129
…                                         …
3587                                 0.000000
3588                                 0.000000
3589                                 0.204147
3590                                 0.000000
3591                                 0.000000

      Vote_Data_Martin_OMalley_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                           …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
3591                                        0.0

      Vote_Data_Mike_Huckabee_Percent_of_Votes  Vote_Data_No_Preference_Party  \
0                                          0.0                           0.0
1                                          0.0                           0.0
2                                          0.0                           0.0
3                                          0.0                           0.0
4                                          0.0                           0.0
…                                          …                             …
3587                                       0.0                           0.0
3588                                       0.0                           0.0
3589                                       0.0                           0.0
3590                                       0.0                           0.0
3591                                       0.0                           0.0

      Vote_Data_No_Preference_Percent_of_Votes  \
0                                     0.000000
1                                     0.400000
2                                     0.000000
3                                     0.000000
4                                     0.066667
…                                          …
3587                                  0.000000
3588                                  0.000000
```

```
3589                                0.000000
3590                                0.000000
3591                                0.000000

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
…                                      …
3587                                   0.0
3588                                   0.0
3589                                   0.0
3590                                   0.0
3591                                   0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                0.304071                          0.0
1                                0.125954                          0.0
2                                0.486005                          0.0
3                                0.202290                          0.0
4                                0.110687                          0.0
…                                   …                              …
3587                             0.104326                          0.0
3588                             0.366412                          0.0
3589                             0.473282                          0.0
3590                             0.000000                          0.0
3591                             0.276081                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes  label
0                                        0.0      5
1                                        0.0     10
2                                        0.0     11
```

```
3                                        0.0       6
4                                        0.0      10
…                                        …         …
3587                                     0.0       3
3588                                     0.0      11
3589                                     0.0       4
3590                                     0.0       9
3591                                     0.0       3

[3592 rows x 19 columns]
```

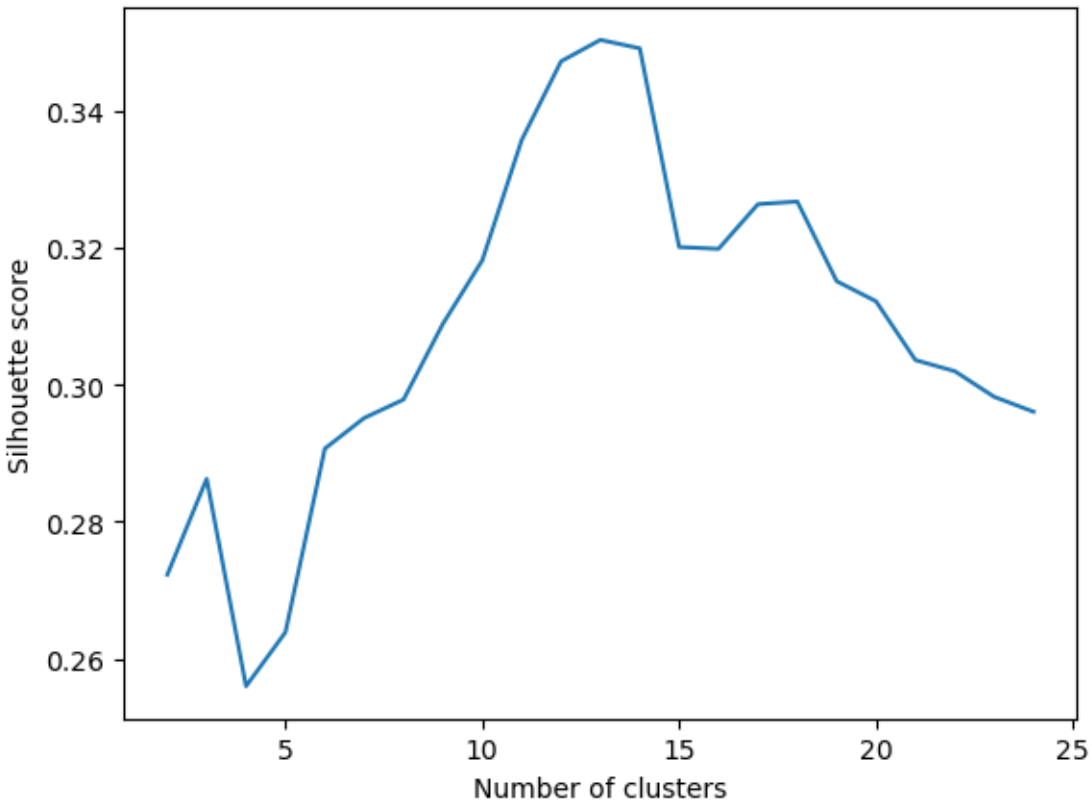
1.7.3 7.3. Clustering Algorithm Results Presentation

[121]:
```
perp = 40
rs = 100

kmeans = KMeans(n_clusters=13, random_state=100).fit(X)
df_combo["label"] = kmeans.labels_
sns.scatterplot(x='x_projected',y='y_projected',hue = "label",palette=sns.
 ↪color_palette('husl', 13), data=df_combo)
plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(perp, rs))
plt.legend(bbox_to_anchor=(1, 1), loc='upper left', fontsize = 5)
plt.show()
```



1.7.4 7.4. Assessing Clustering Separation and Cohesion

[52]:
```
silhouette_scores = []
for k in range(2, 25):
    kmeans = KMeans(n_clusters=k, random_state=0)
    labels = kmeans.fit_predict(X)
    score = silhouette_score(X, labels)
    silhouette_scores.append(score)

#
plt.plot(range(2, 25), silhouette_scores)
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette score')
plt.show()
```



[53]:
```
def show_silhouette_plots(X,cluster_labels):

    # This package allows us to use "color maps" in our visualizations
```

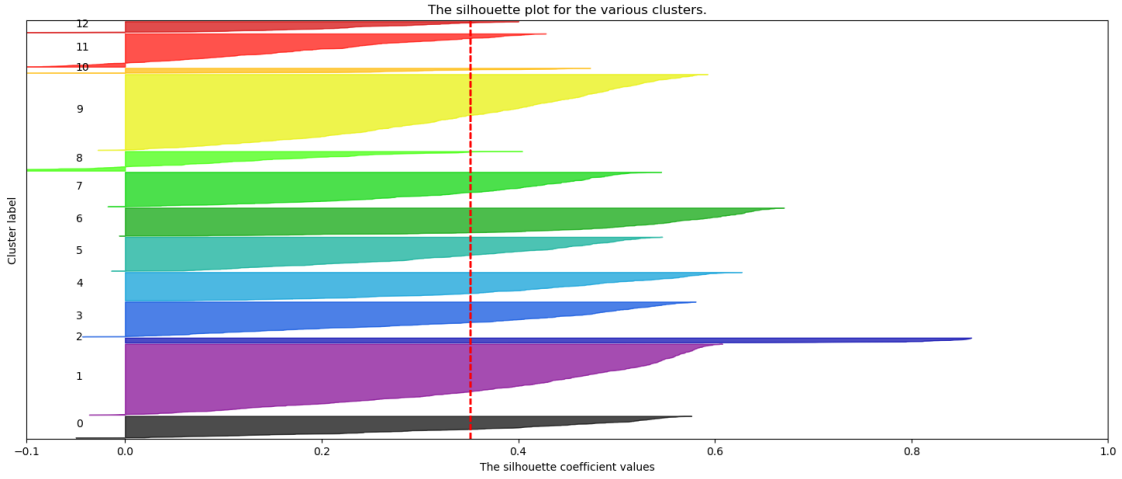
```
import matplotlib.cm as cm

#How many clusters in your clustering?
n_clusters=len(np.unique(cluster_labels))

# Create a subplot with 1 row and 2 columns
fig, ax1 = plt.subplots(1, 1)
fig.set_size_inches(18, 7)

# The 1st subplot is the silhouette plot
# The silhouette coefficient fcan range from -1, 1 but in this example all
# lie within [-0.1, 1]
ax1.set_xlim([-0.1, 1])
# The (n_clusters+1)*10 is for inserting blank space between silhouette
# plots of individual clusters, to demarcate them clearly.
ax1.set_ylim([0, len(X) + (n_clusters + 1) * 10])


# The silhouette_score gives the average value for all the samples.
# This gives a perspective into the density and separation of the formed
# clusters
silhouette_avg = silhouette_score(X, cluster_labels)
print("For n_clusters =", n_clusters,
      "The average silhouette_score is :", silhouette_avg)

# Compute the silhouette scores for each sample
sample_silhouette_values = silhouette_samples(X, cluster_labels)

y_lower = 10
for i in range(n_clusters):
    # Aggregate the silhouette scores for samples belonging to
    # cluster i, and sort them
    ith_cluster_silhouette_values = \
        sample_silhouette_values[cluster_labels == i]


    ith_cluster_silhouette_values.sort()

    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i

    color = cm.nipy_spectral(float(i) / n_clusters)
    ax1.fill_betweenx(np.arange(y_lower, y_upper),
                      0, ith_cluster_silhouette_values,
                      facecolor=color, edgecolor=color, alpha=0.7)

    # Label the silhouette plots with their cluster numbers at the middle
```

```
        ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

        # Compute the new y_lower for next plot
        y_lower = y_upper + 10  # 10 for the 0 samples

        ax1.set_title("The silhouette plot for the various clusters.")
        ax1.set_xlabel("The silhouette coefficient values")
        ax1.set_ylabel("Cluster label")

        # The vertical line for average silhouette score of all the values
        ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

        ax1.set_yticks([])  # Clear the yaxis labels / ticks
        ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])


    plt.show()

    return
```

[54]:
```
kmeans=KMeans(n_clusters=13, random_state=100).fit(X)
cluster_labels = kmeans.labels_
```

[55]:
```
show_silhouette_plots(X,cluster_labels)
```

```
For n_clusters = 13 The average silhouette_score is : 0.35128092983387776
```



a.) each of the clusters and

b.) the overall clustering. Are there are any objects that have poor cohesion with their assigned cluster? Explain.

a. We can see cluster 12 and 8 are quite same, but the score is at 0.4, which indicate they have a

poor separation and cohesion. Cluster 9, 1, 4, 0 and 3 are similar, the score is 0.6 which have a moderate separation and cohesion. CLuster 10 is at score 0.5, is a median in separation and cohesion. CLuster 6 is in a second place which have relatively good separation and cohesion. Cluster 2 have the highest score which is 0.9, well-separated and well-matched.

b. In this case, the average silhouette score of 0.351 indicates that the clusters are not fairly well-separated and data points within each cluster are not well-matched. The score is not particularly high, which suggests that there may be some overlap or ambiguity between the clusters, or that the clusters are not perfectly optimized for the data.

1.7.5 7.5. Additional Analysis

[56]:
```
from sklearn.metrics import adjusted_rand_score, homogeneity_score,␣
↪completeness_score
```

[57]:
```
ari = adjusted_rand_score(df_label["Location_State"], df_kmeans_label["label"])
ari
```

[57]: 0.3469780328175856

[58]:
```
homogeneity = homogeneity_score(df_label["Location_State"],␣
↪df_kmeans_label["label"])
homogeneity
```

[58]: 0.5464057466437389

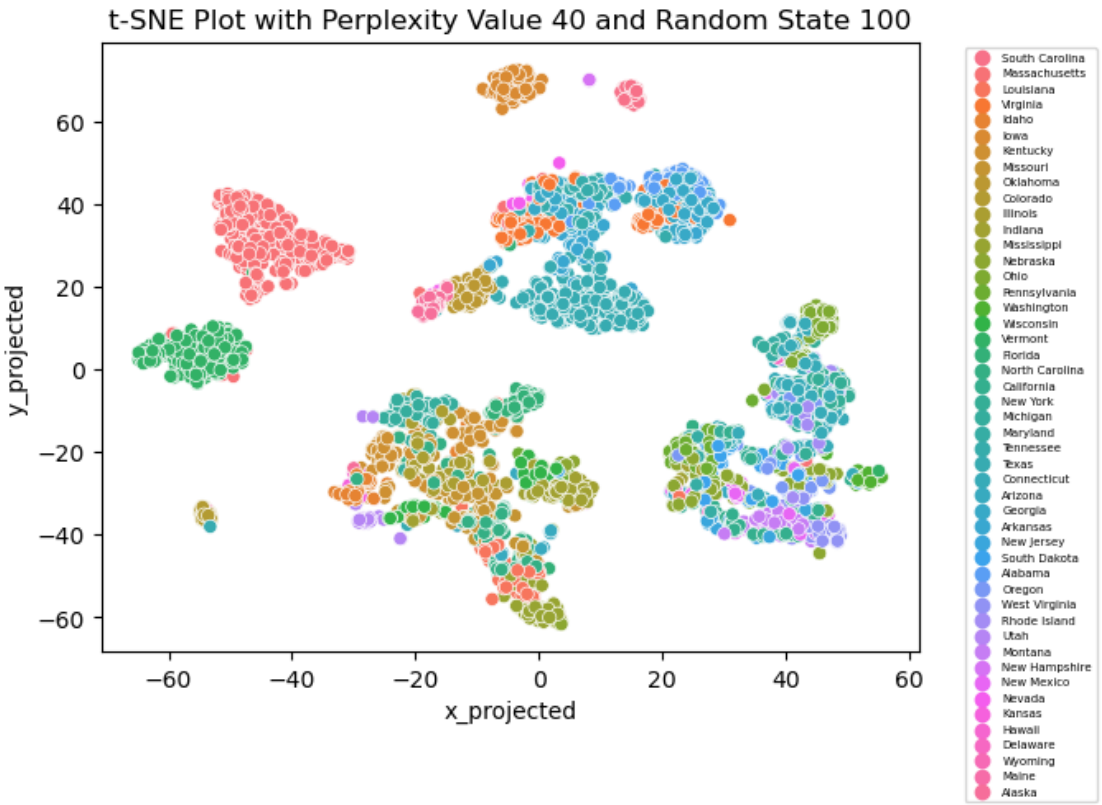
[59]:
```
completeness = completeness_score(df_label["Location_State"],␣
↪df_kmeans_label["label"])
completeness
```

[59]: 0.8269716937056959

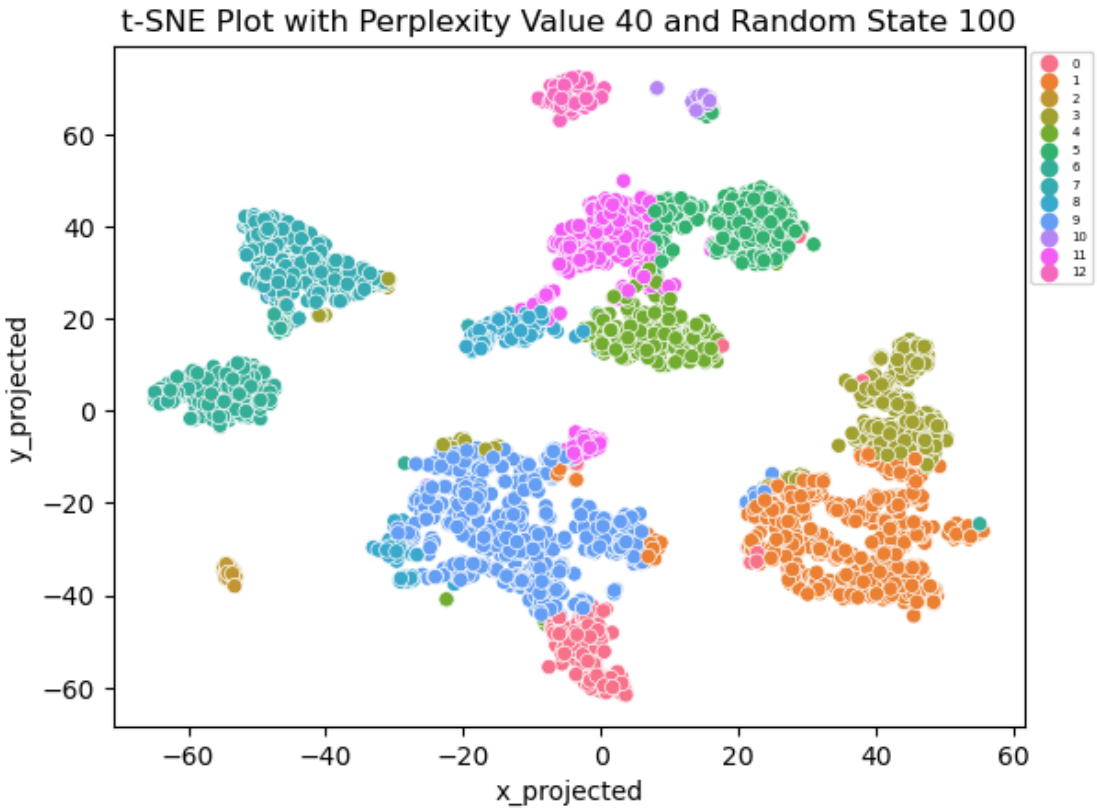
The adjusted RAND index score of 0.357 indicates a moderate level of agreement between the true and predicted labels. The homogeneity score of 0.542 suggests that each cluster contains members of multiple classes, while the completeness score of 0.811 indicates that members of each class are not all assigned to the same cluster. These results suggest that while there is some agreement between the true and predicted labels, the clustering may not be capturing all of the underlying patterns in the data. It may be worth exploring other clustering methods or tweaking the parameters of the k-means algorithm to improve the results.

[100]:
```
for perp in [40]:
    for rs in [100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', hue =␣
↪"Location_State",data=df_combo)
```

```
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
 ↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('          ----------------------True Label-----------------------')
#Color code the points in your t-sne plot by cluster labels and code the␣
 ↪"style" of the marker with your class labels.
kmeans = KMeans(n_clusters=13, random_state=100).fit(X)
df_combo["label"] = kmeans.labels_
sns.scatterplot(x='x_projected',y='y_projected',hue = "label",palette=sns.
 ↪color_palette('husl', 13), data=df_combo)
plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(perp, rs))
plt.legend(bbox_to_anchor=(1, 1), loc='upper left', fontsize = 5)
plt.show()
print('          ----------------------Predicted Label-----------------------')
```




```
          ----------------------True Label-----------------------
```

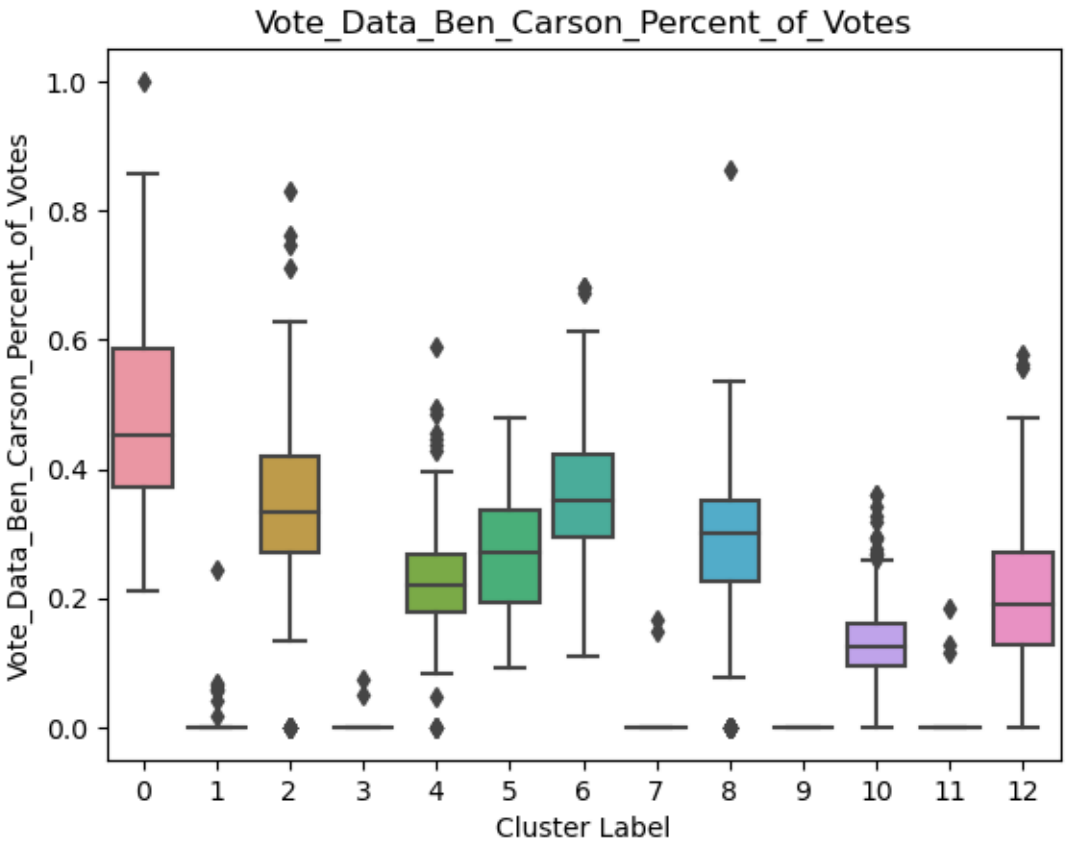


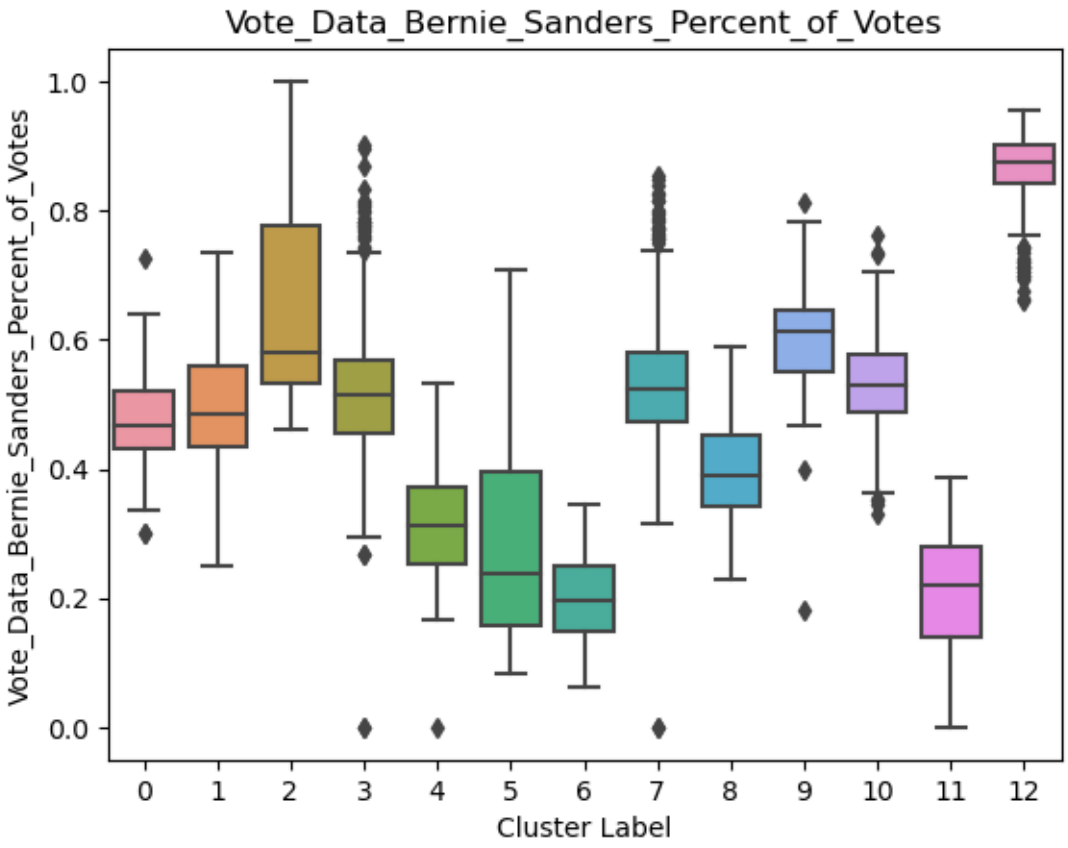
----------Predicted Label----------

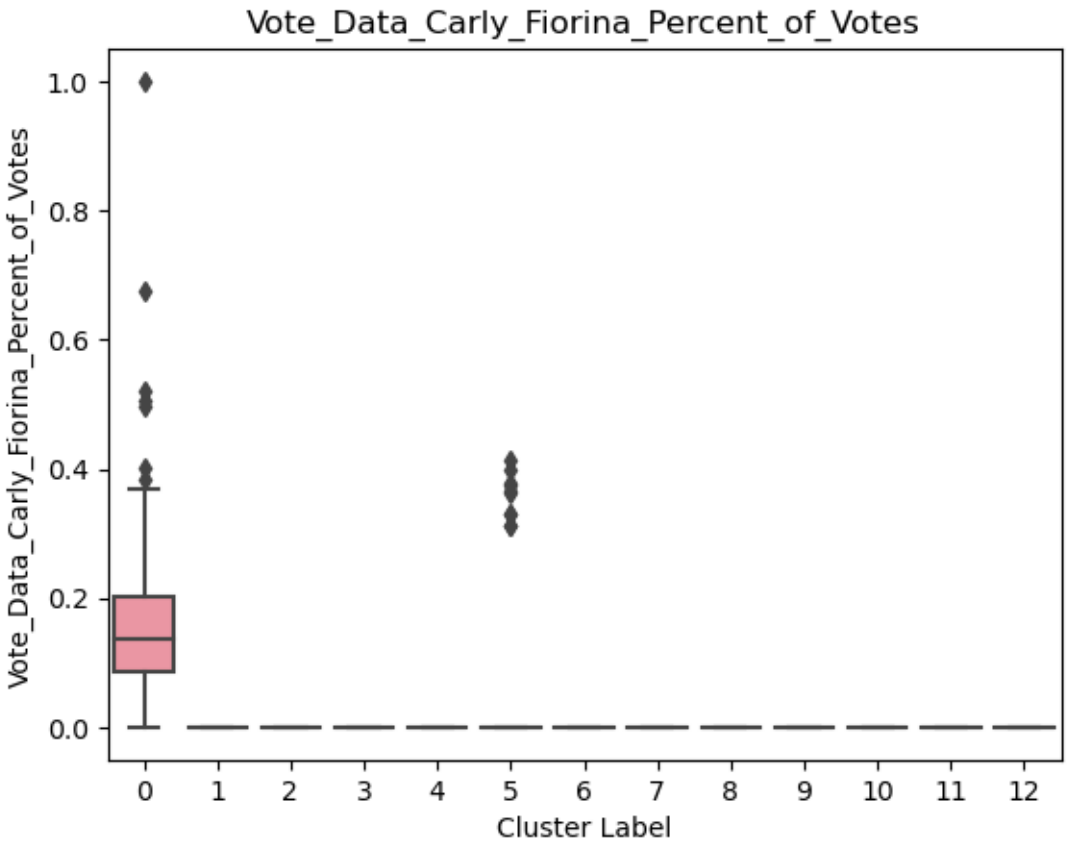
There is a bit difference between true label. We can see the left-up corner have a connect with the down one. But, the true label is not.

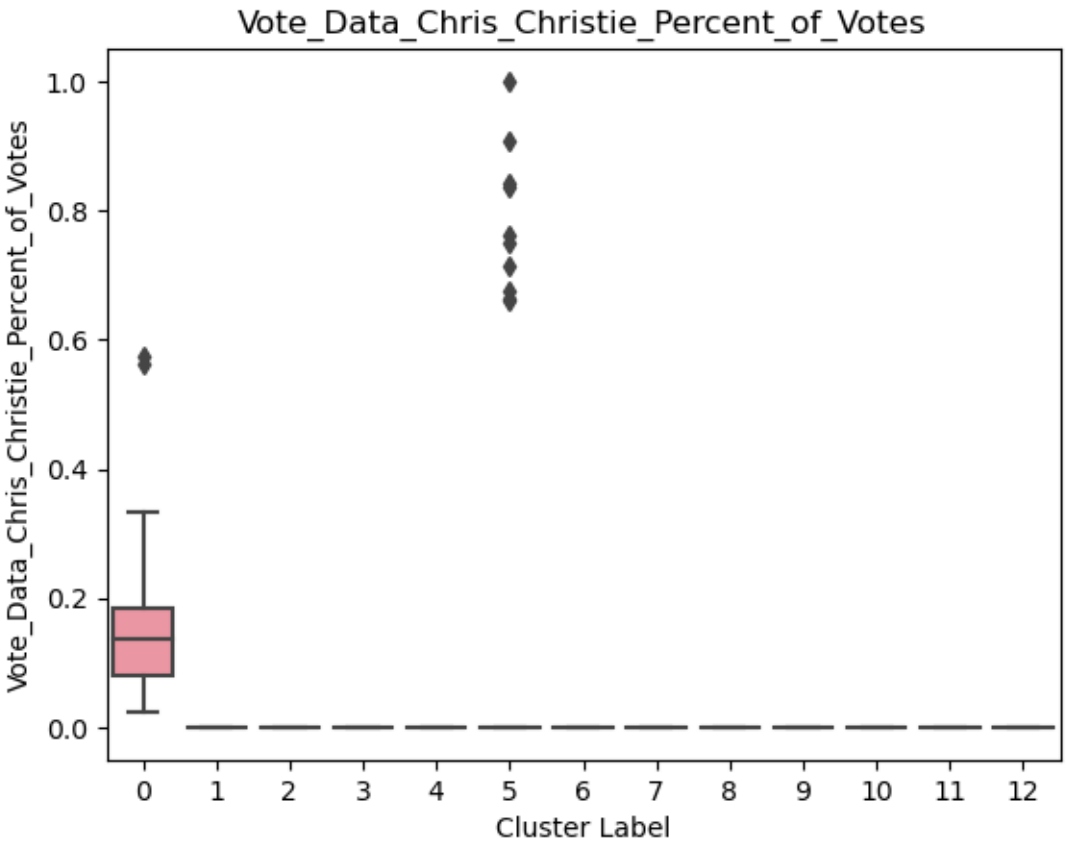
1.7.6 7.6. Describing Each of the Clusters

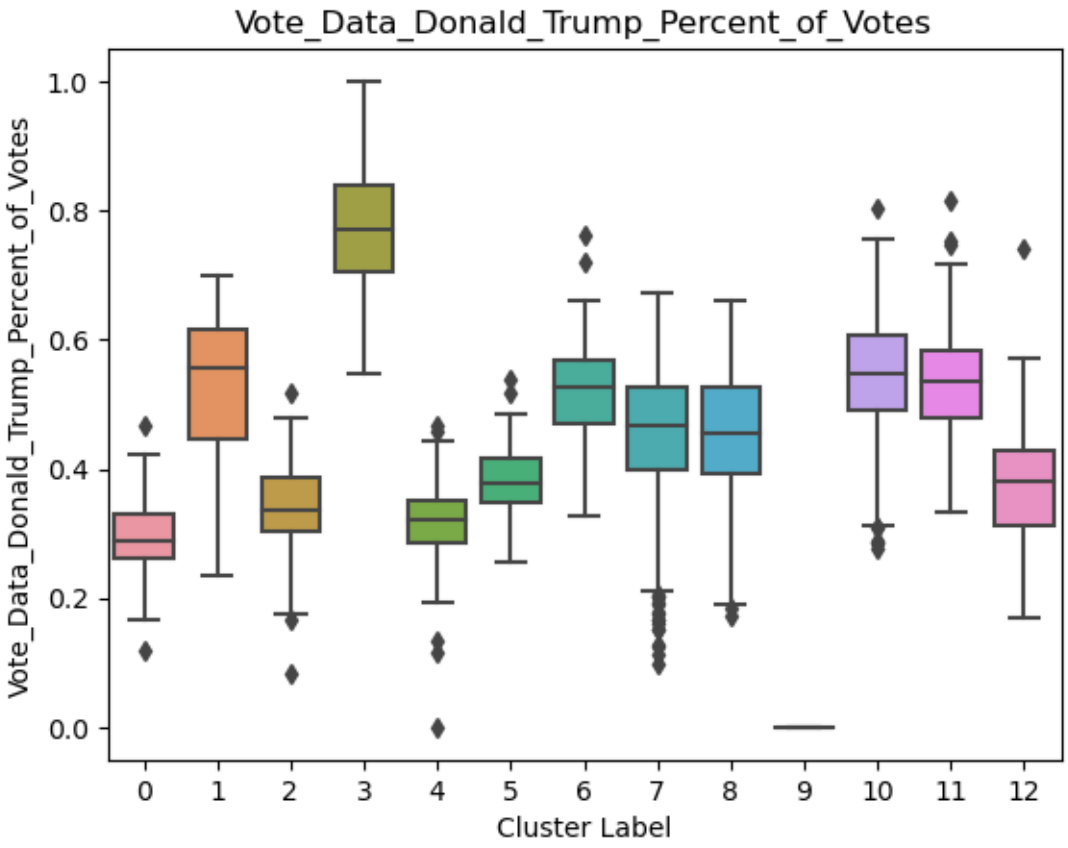
[115]:
```
for col in df_kmeans_label.columns[:-1]:
    plt.figure()
    plt.title(col)
    sns.boxplot([df_kmeans_label.loc[df_kmeans_label['label'] == label, col]␣
 ↪for label in range(13)])
    plt.xlabel('Cluster Label')
    plt.ylabel(col)
    plt.show()
```

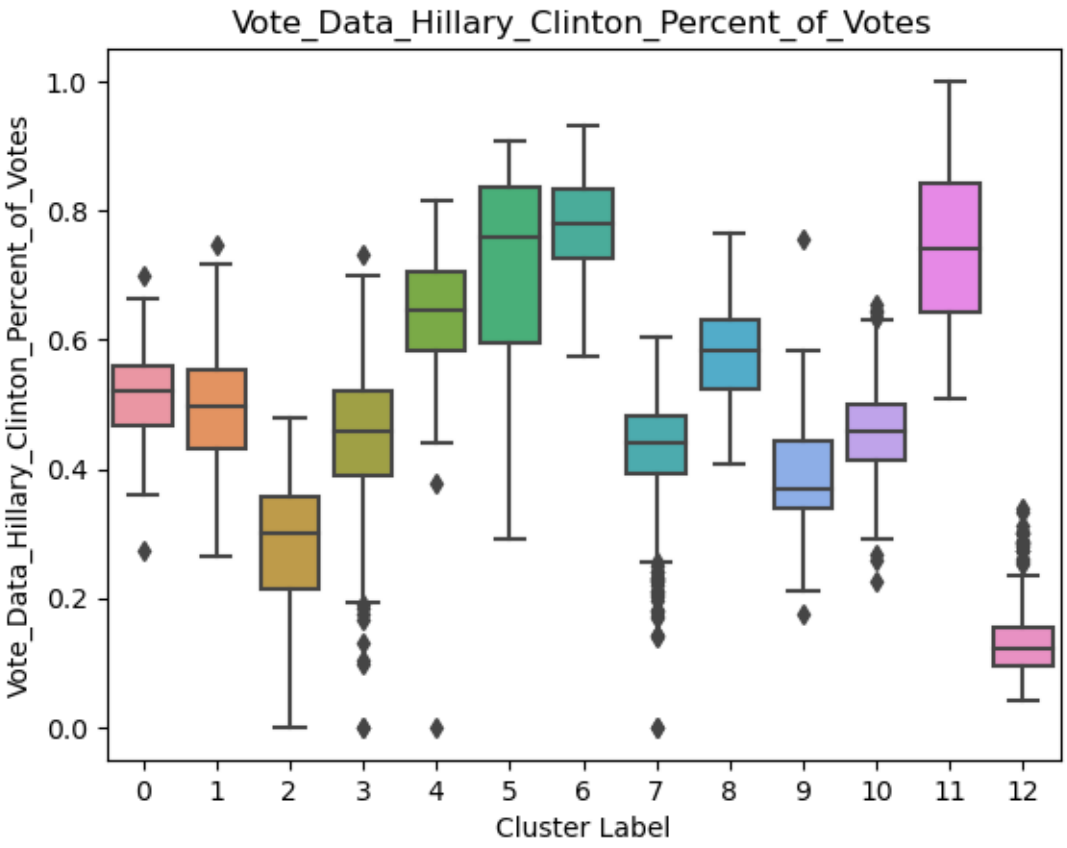

Vote_Data_Ben_Carson_Percent_of_Votes
Vote_Data_Ben_Carson_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

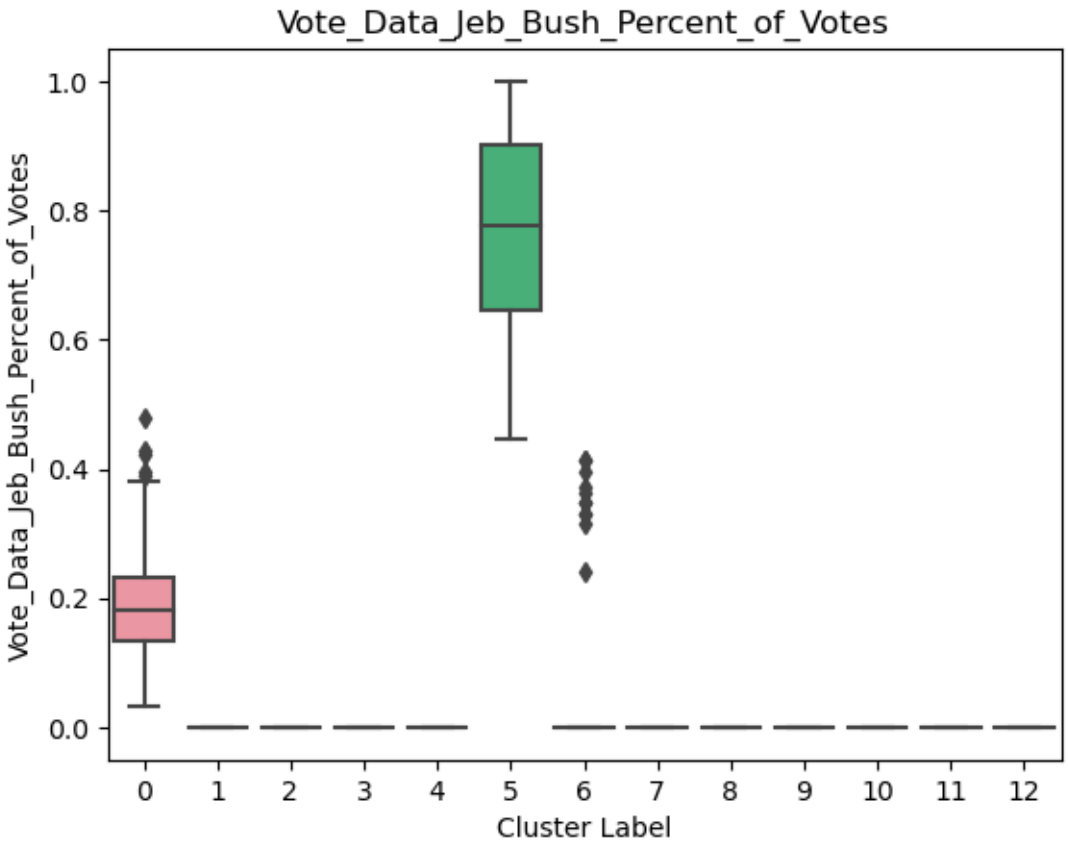

Vote_Data_Bernie_Sanders_Percent_of_Votes
Vote_Data_Bernie_Sanders_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

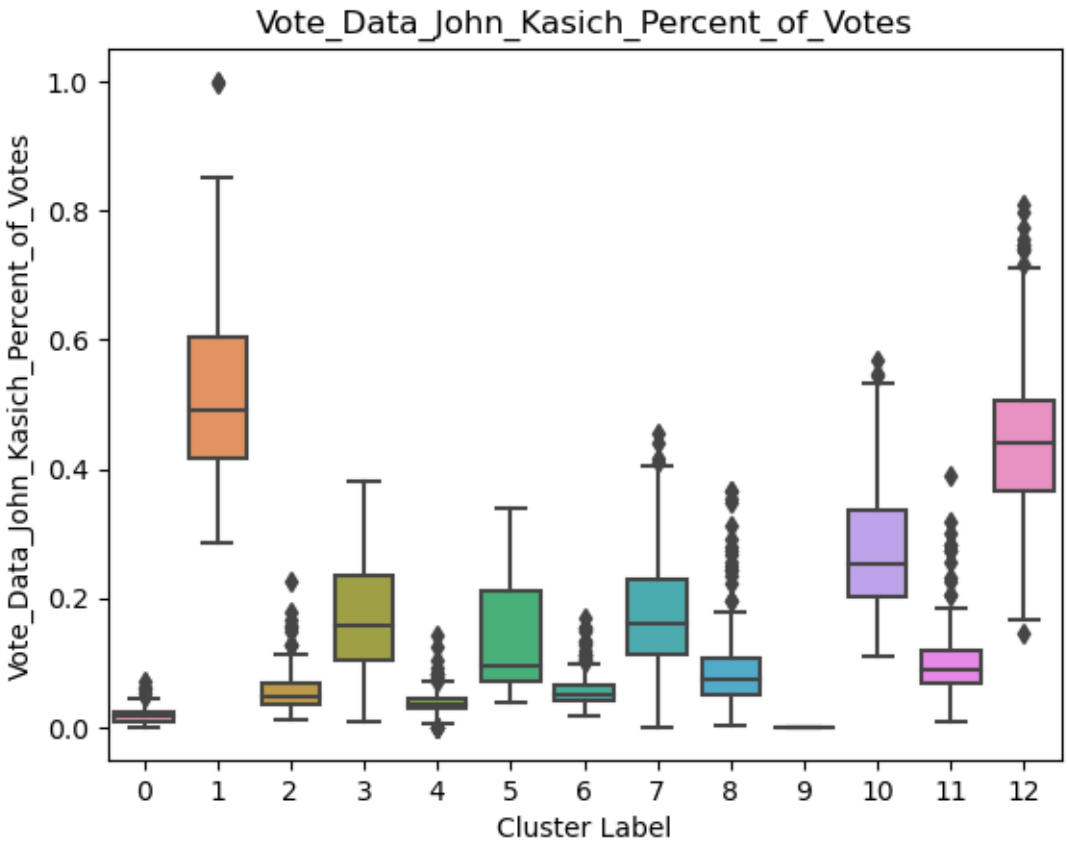

Vote_Data_Carly_Fiorina_Percent_of_Votes
Vote_Data_Carly_Fiorina_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

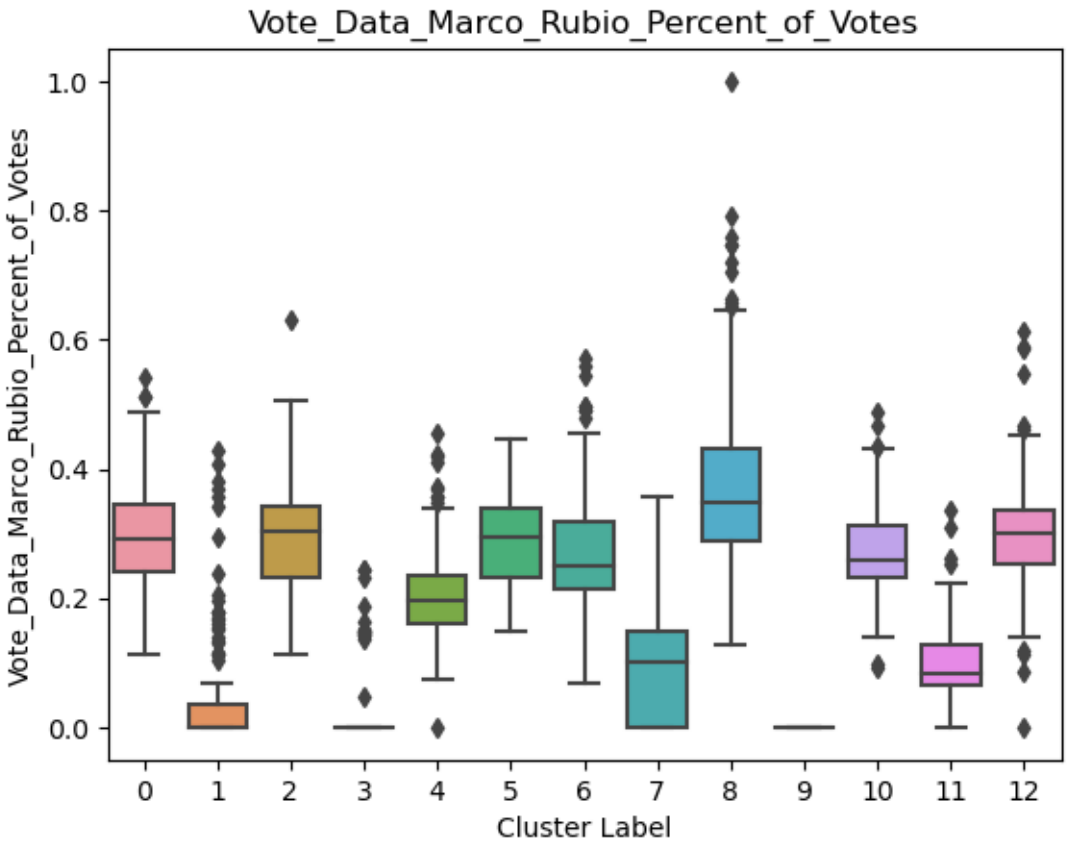

Vote_Data_Chris_Christie_Percent_of_Votes
Vote_Data_Chris_Christie_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

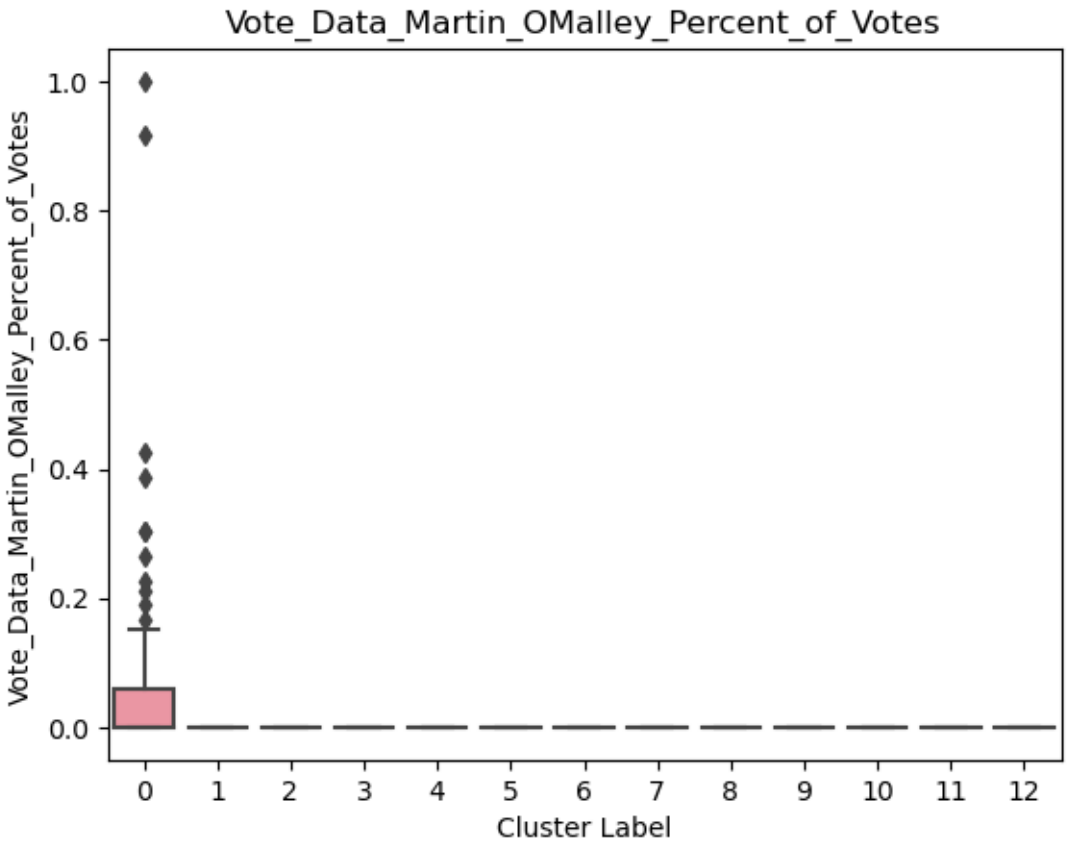

Vote_Data_Donald_Trump_Percent_of_Votes
Vote_Data_Donald_Trump_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

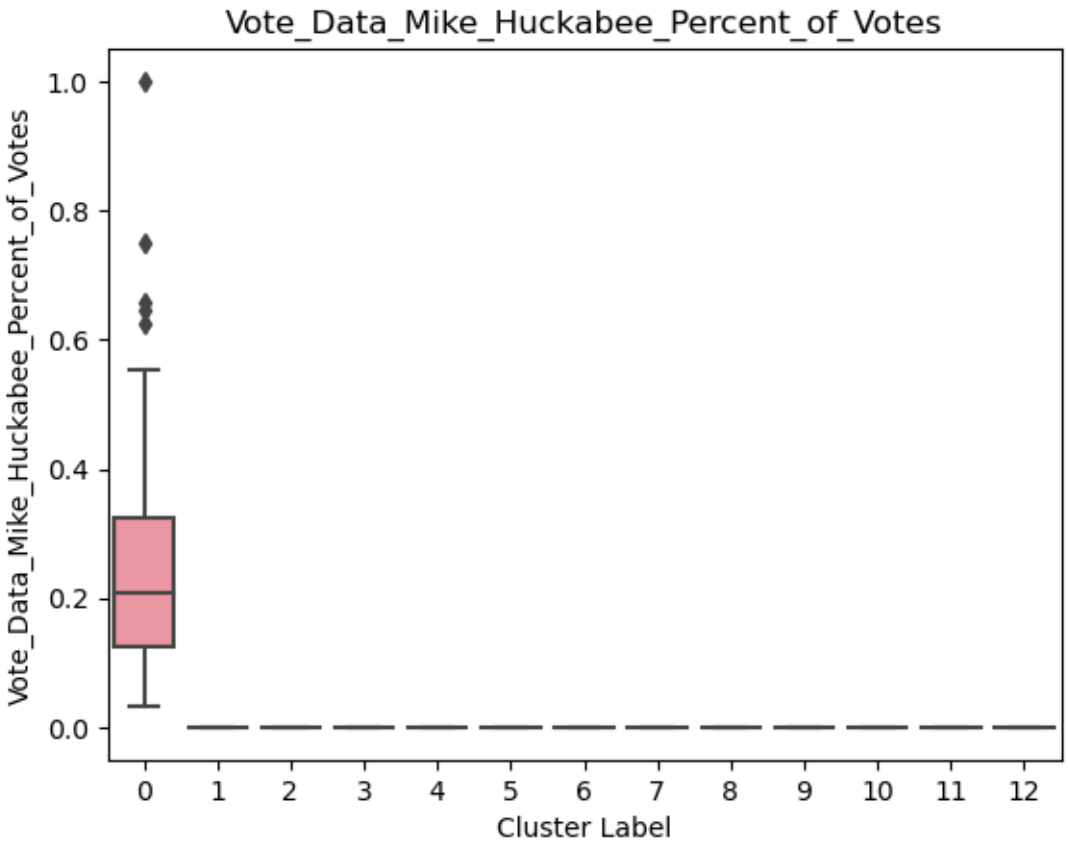

Vote_Data_Hillary_Clinton_Percent_of_Votes
Vote_Data_Hillary_Clinton_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

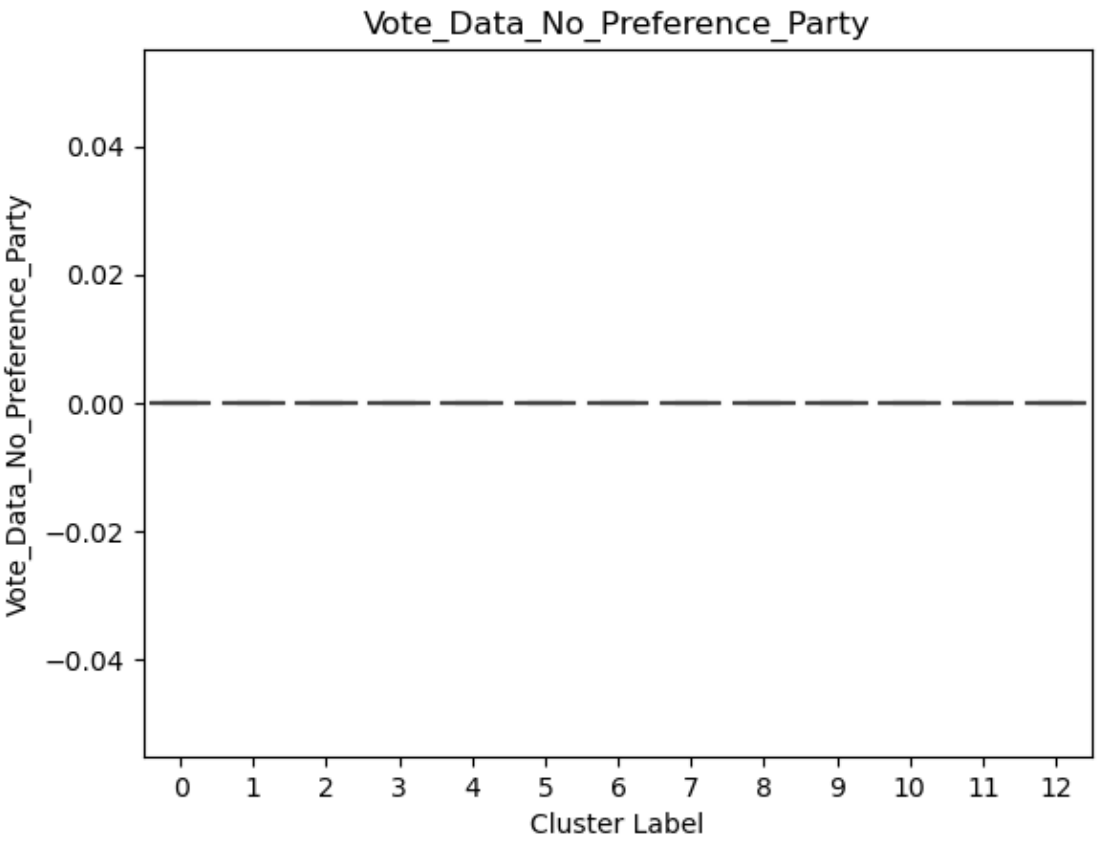

Vote_Data_Jeb_Bush_Percent_of_Votes
Vote_Data_Jeb_Bush_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

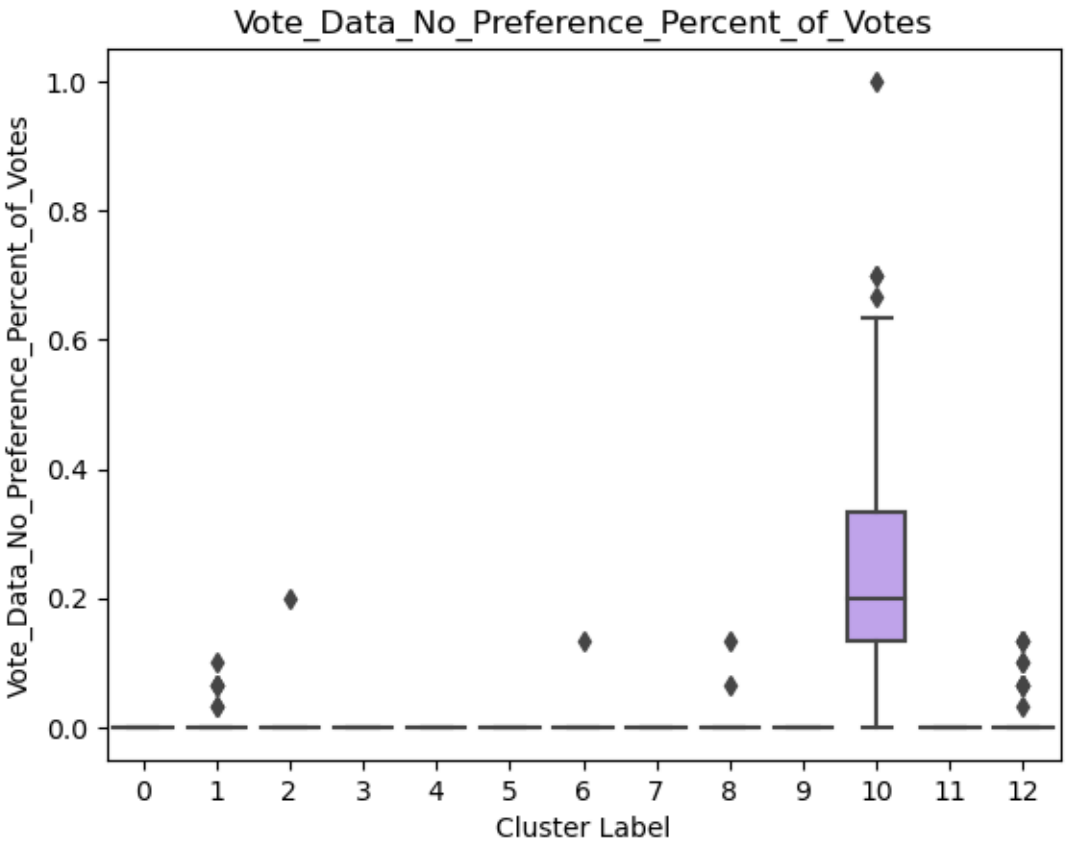

Vote_Data_John_Kasich_Percent_of_Votes
Vote_Data_John_Kasich_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

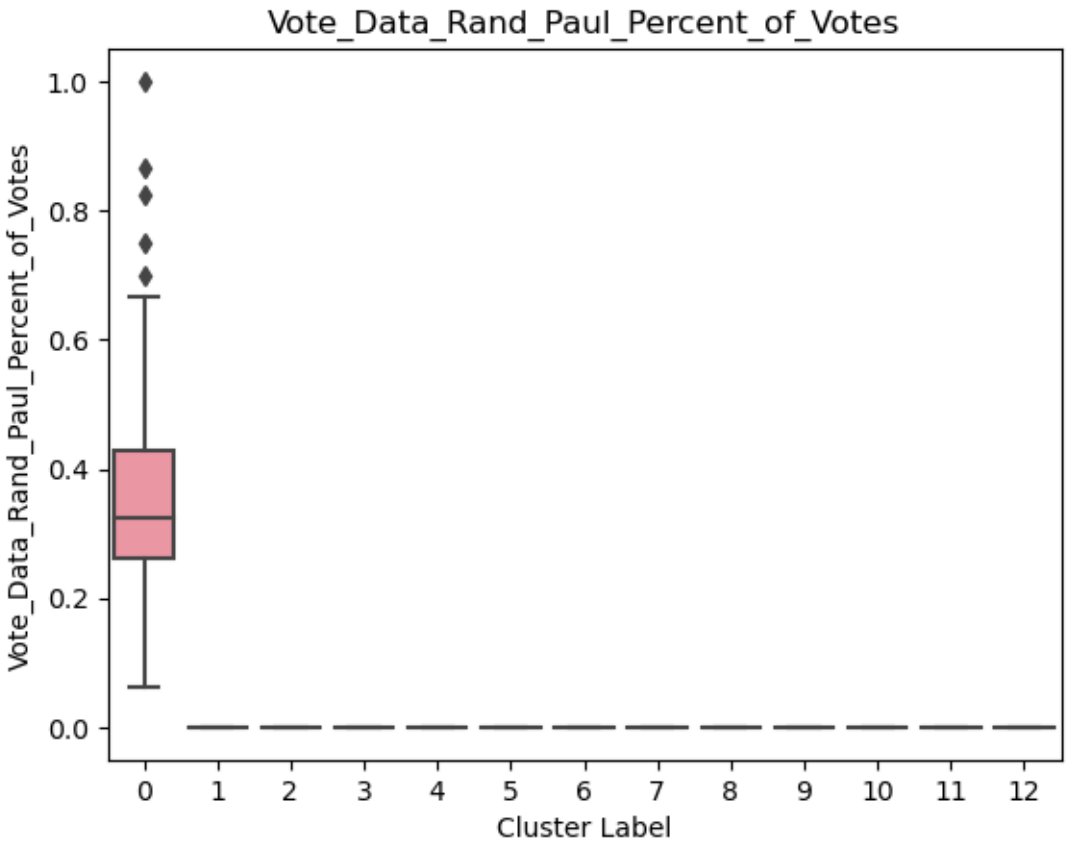

Vote_Data_Marco_Rubio_Percent_of_Votes
Vote_Data_Marco_Rubio_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

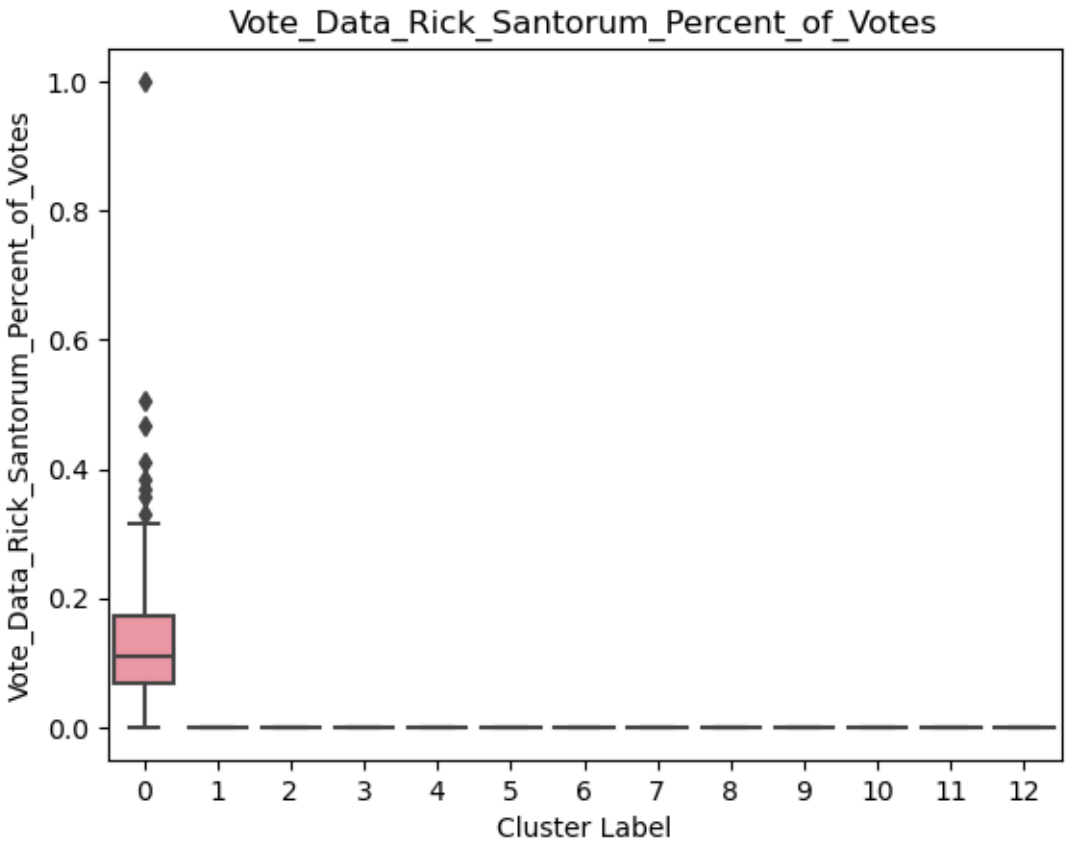

Vote_Data_Martin_OMalley_Percent_of_Votes
Vote_Data_Martin_OMalley_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

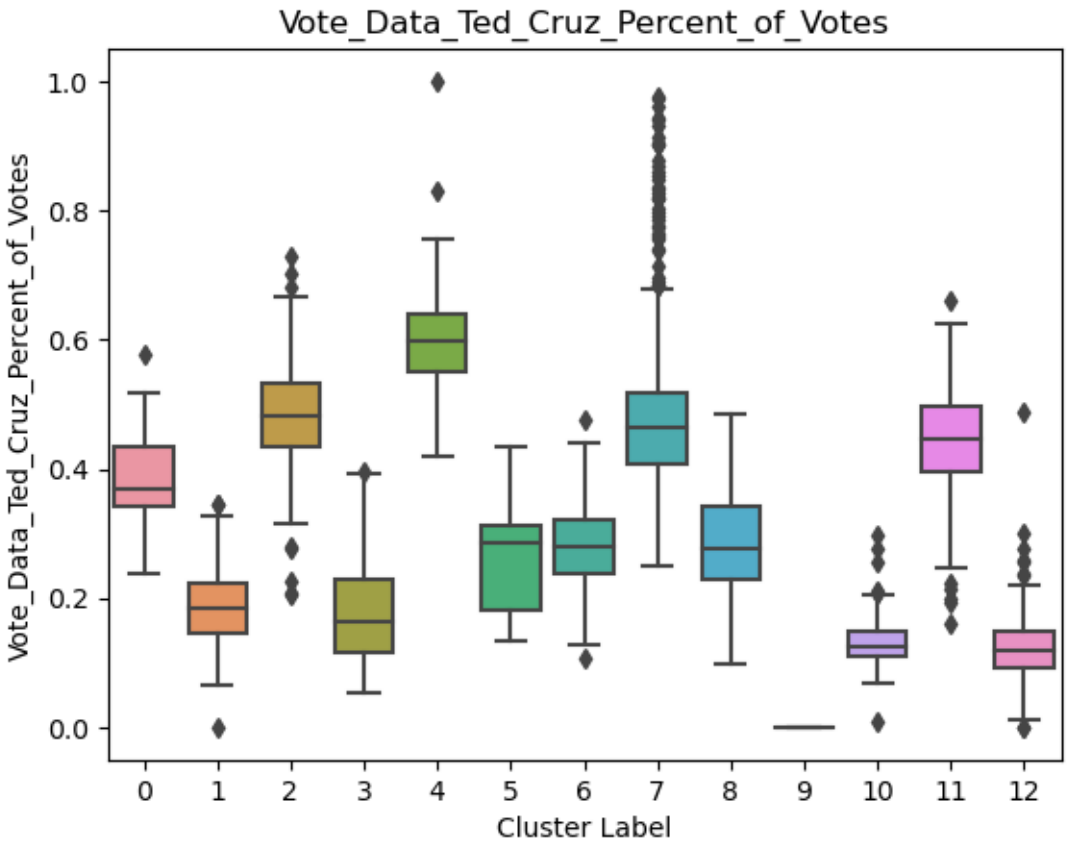

Vote_Data_Mike_Huckabee_Percent_of_Votes
Vote_Data_Mike_Huckabee_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

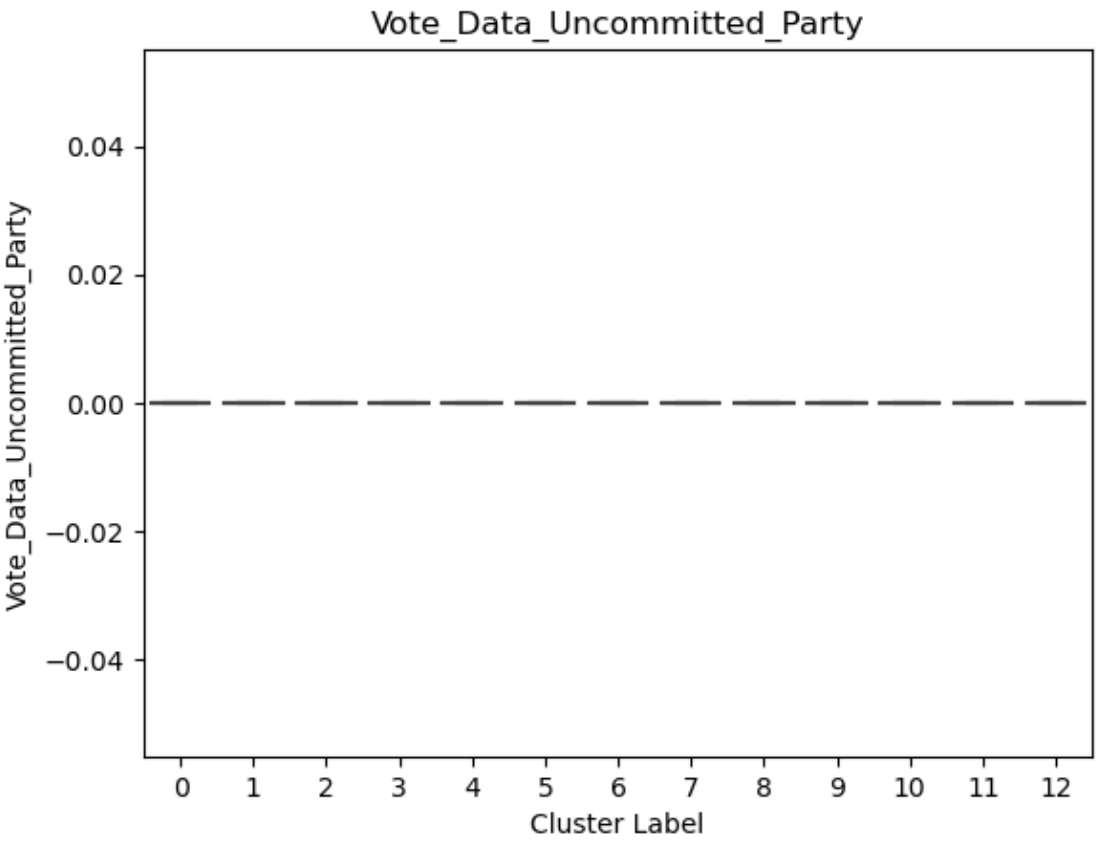

Vote_Data_No_Preference_Party
Vote_Data_No_Preference_Party
0.04
0.02
0.00
−0.02
−0.04
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

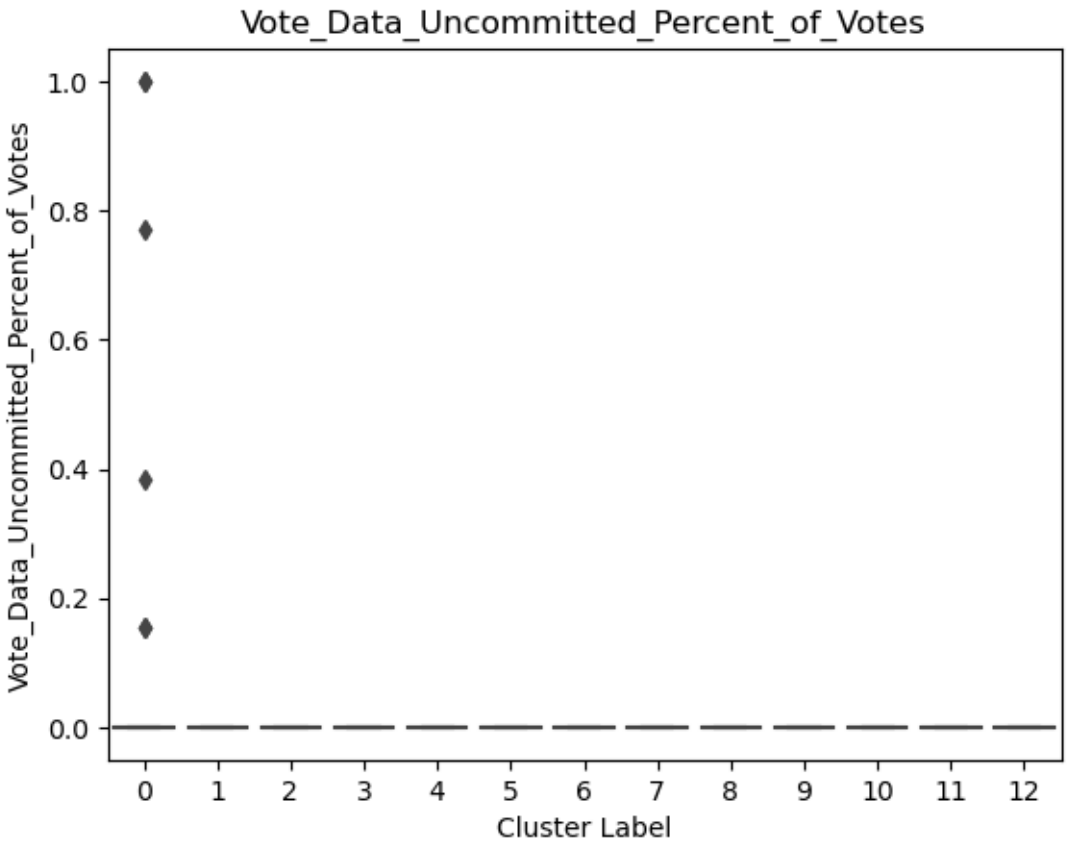

Vote_Data_No_Preference_Percent_of_Votes
Vote_Data_No_Preference_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label


Vote_Data_Rand_Paul_Percent_of_Votes
Vote_Data_Rand_Paul_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label


Vote_Data_Rick_Santorum_Percent_of_Votes
Vote_Data_Rick_Santorum_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label


Vote_Data_Ted_Cruz_Percent_of_Votes
Vote_Data_Ted_Cruz_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label

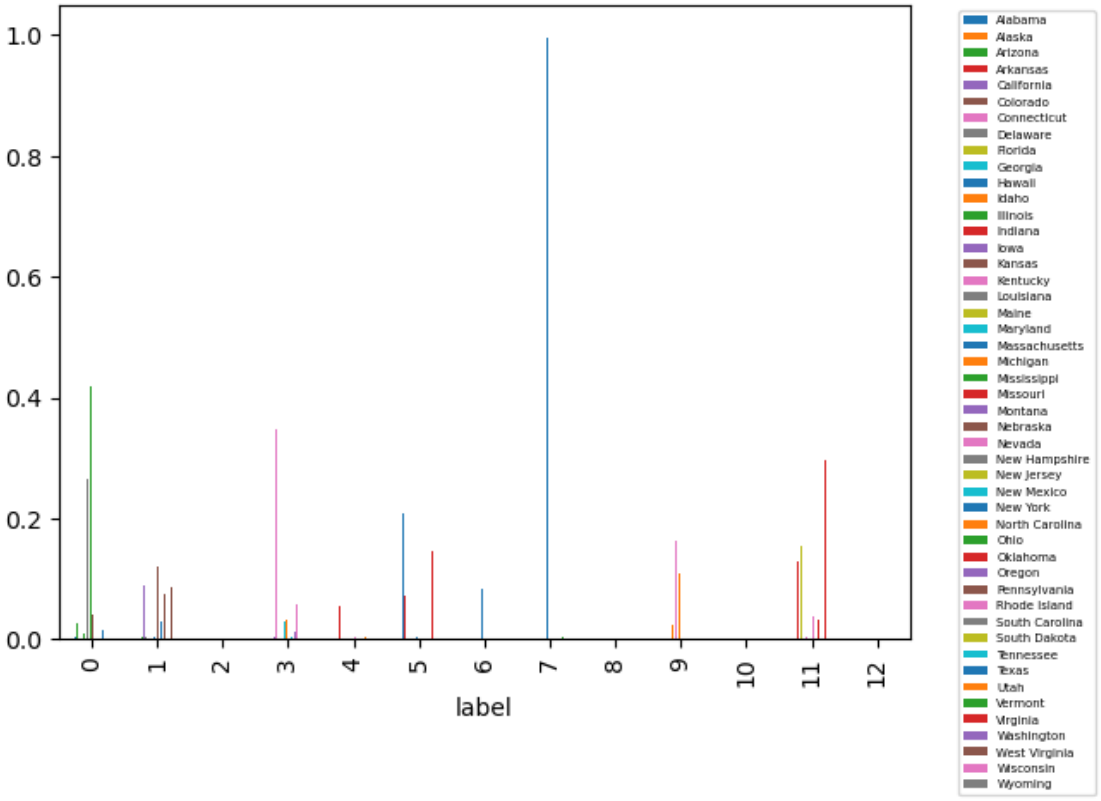

Vote_Data_Uncommitted_Party
0.04
0.02
0.00
−0.02
−0.04
Vote_Data_Uncommitted_Party
0
1
2
3
4
5
6
7
8
9
10
11
12
Cluster Label



1. Ben Carson has the highest value of 0.5 in cluster 0 and the lowest cluster is 10.
2. Bernie Sanders is distributed in cluster 12 with a maximum value of 0.9 and the lowest cluster is cluster 6.
3. Carly Fiorina is mainly distriubate in Cluster 0, have 0.1.
4. Chris Christie is mainly distriubate in Cluster 0 have 0.1.
5. Donald Trump, is mainly distriubate in Cluster 3–0.8. Then he has a very low approval rating in Cluster 0.
6. Hillary Cliton is mainly in 11, with a value of 0.7 and lowest is in 12.
7. Jeb Bush has very high support in cluster 5, with a value of 0.8 Very low support in cluster 0.
8. John Kasich's approval rating is only high in 1 and 12, with values of 0.5 and 0.5 respectively, while the rest are very low.
9. Marco Rubio has a relatively high approval rate of 8, with a value of 0.4, but the rest are quite average.
10. Mike Huckabee and Martin O'Malley are in the 0 distribution, the others are not.
11. Rand Paul's approval rating is at cluster 0.
12. Rick Santorum's approval rating is at cluster 0.
13. Ted Cruz's approval rating is highest at cluster 4 with a value of 0.6, lowest cluster is 12.

There is no centralized distribution for both unintentional and nonpartisan data, and very little data.

[103]:
```
ctab=pd.crosstab(df_combo['label'],␣
 ↪df_combo['Location_State'],normalize='index')
ctab.plot.bar()
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
plt.show()
```



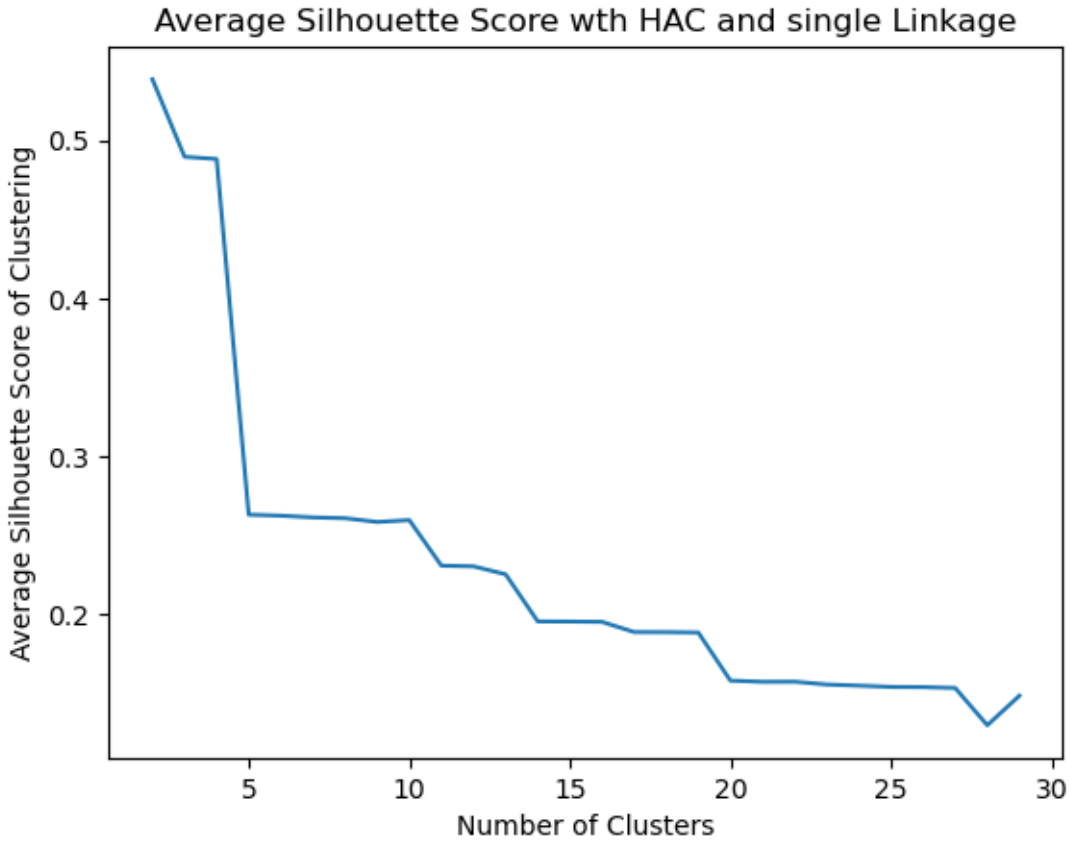
We can see that cluster 7 is composed of one state. But there are some clusters are not agree with motivation, for example, clusters 0, 1 and 11 have a lot of different states in their distribution. This shows that also the cohesion is a little bit unsatisfactory.

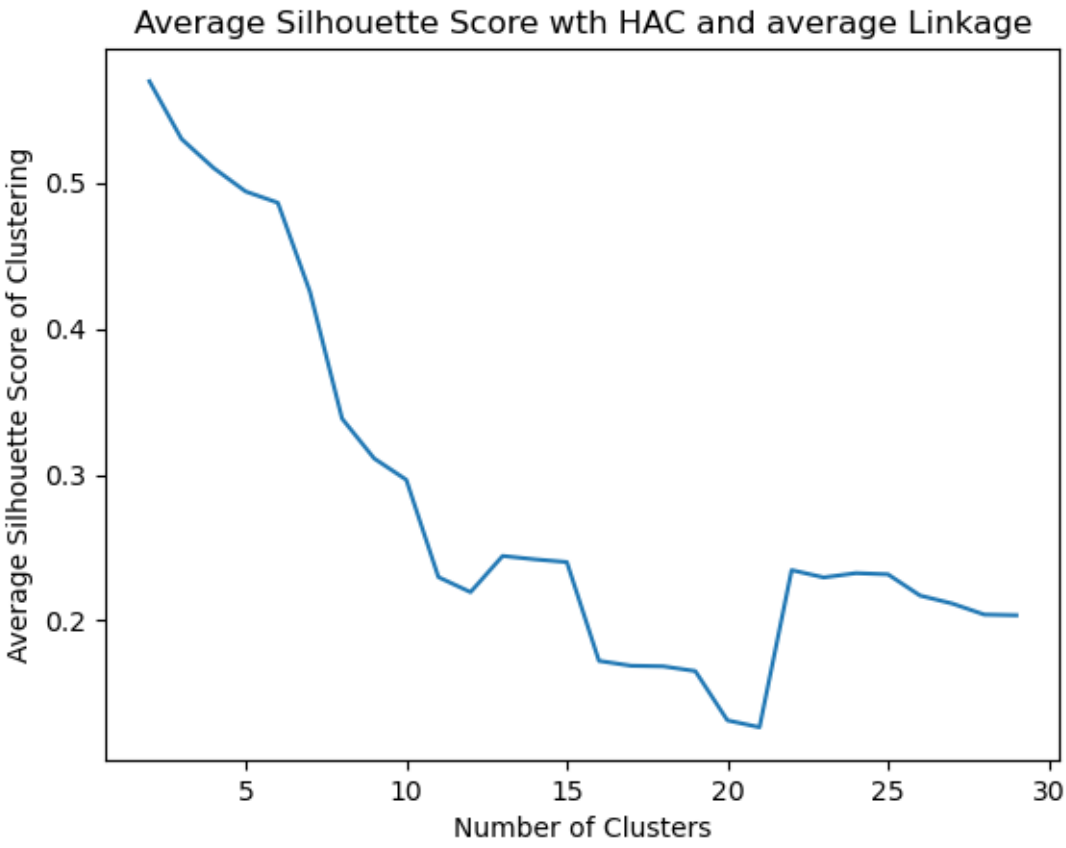
1.8 8. Clustering Algorithm 2 (change name to the algorithm you chose)

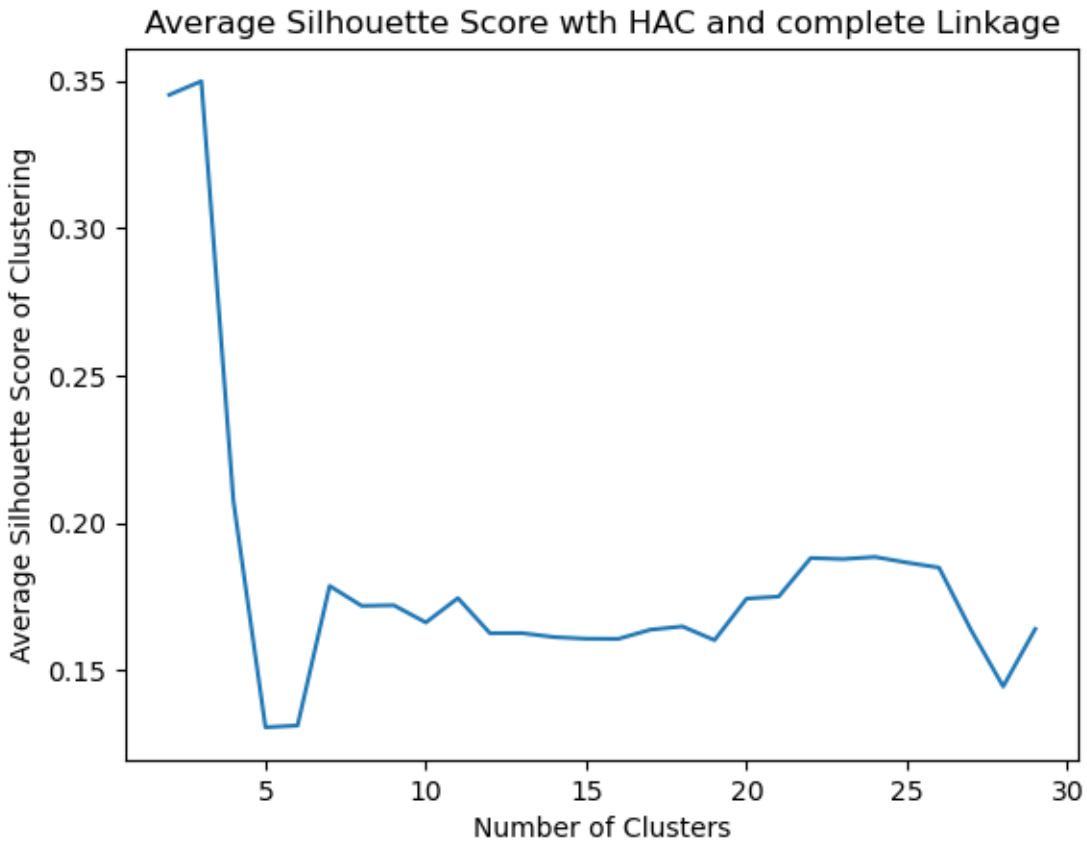
1.8.1 8.1. Parameter Selection

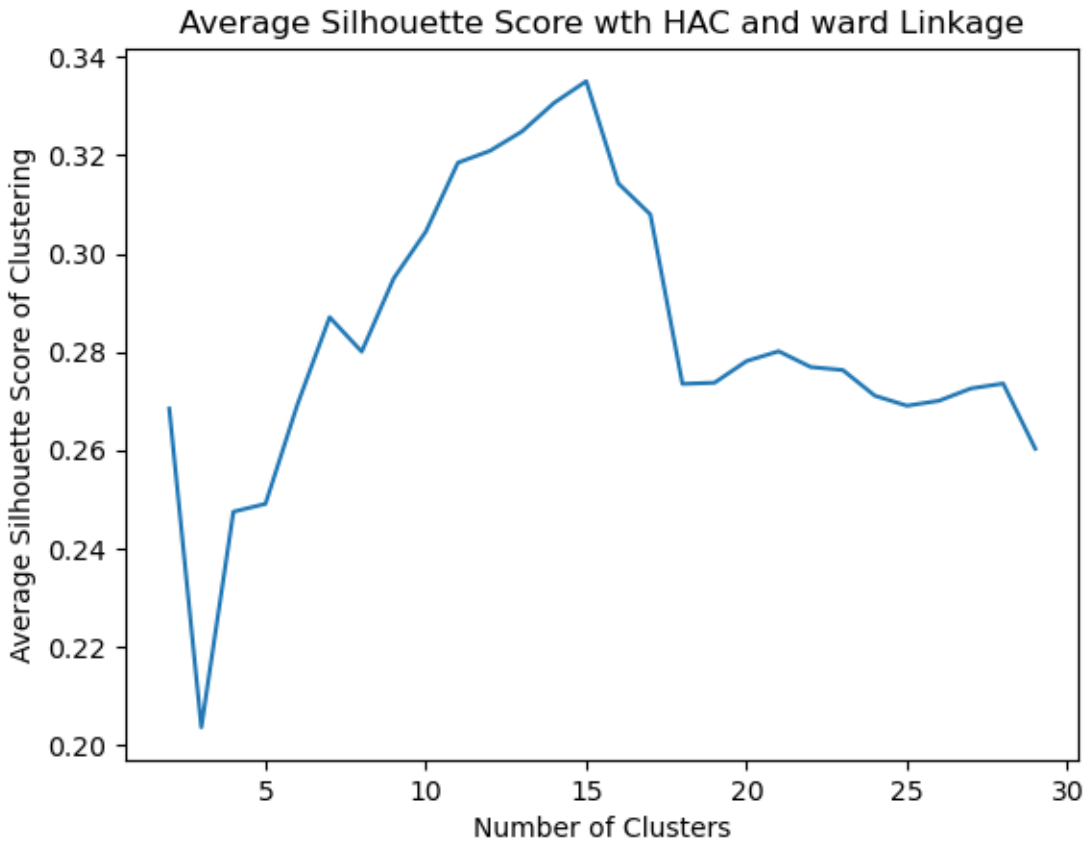
[62]:
```
for link in ['single', 'average', 'complete', 'ward']:
    avg_ss=[]
    for k in range(2,30):
        hac = AgglomerativeClustering(n_clusters=k, affinity='euclidean',␣
 ↪linkage=link)
        Y_pred = hac.fit_predict(df_num)
        avg_ss.append(silhouette_score(df_num, Y_pred))
```

```
    plt.plot(range(2,30), avg_ss)
    plt.title('Average Silhouette Score wth HAC and %s Linkage'%link)
    plt.xlabel('Number of Clusters')
    plt.ylabel('Average Silhouette Score of Clustering')
    plt.show()
```



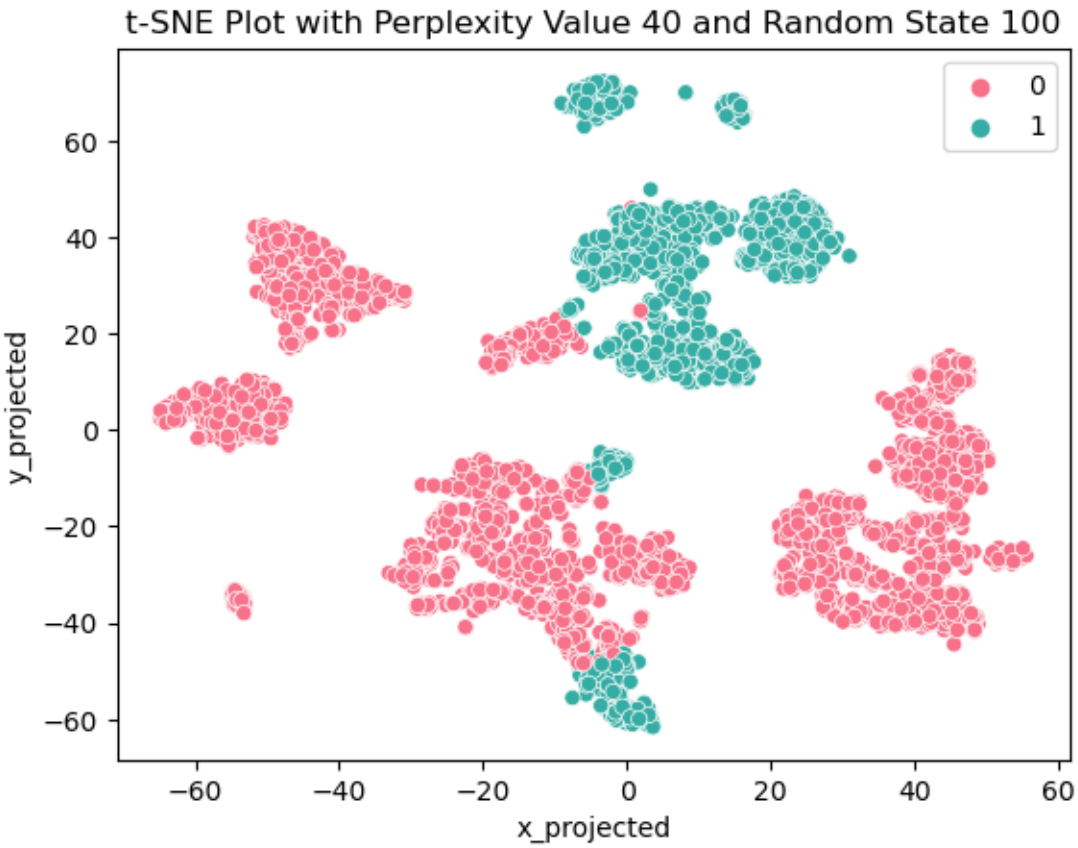

Average Silhouette Score wth HAC and average Linkage
Average Silhouette Score of Clustering
0.5
0.4
0.3
0.2
5
10
15
20
25
30
Number of Clusters

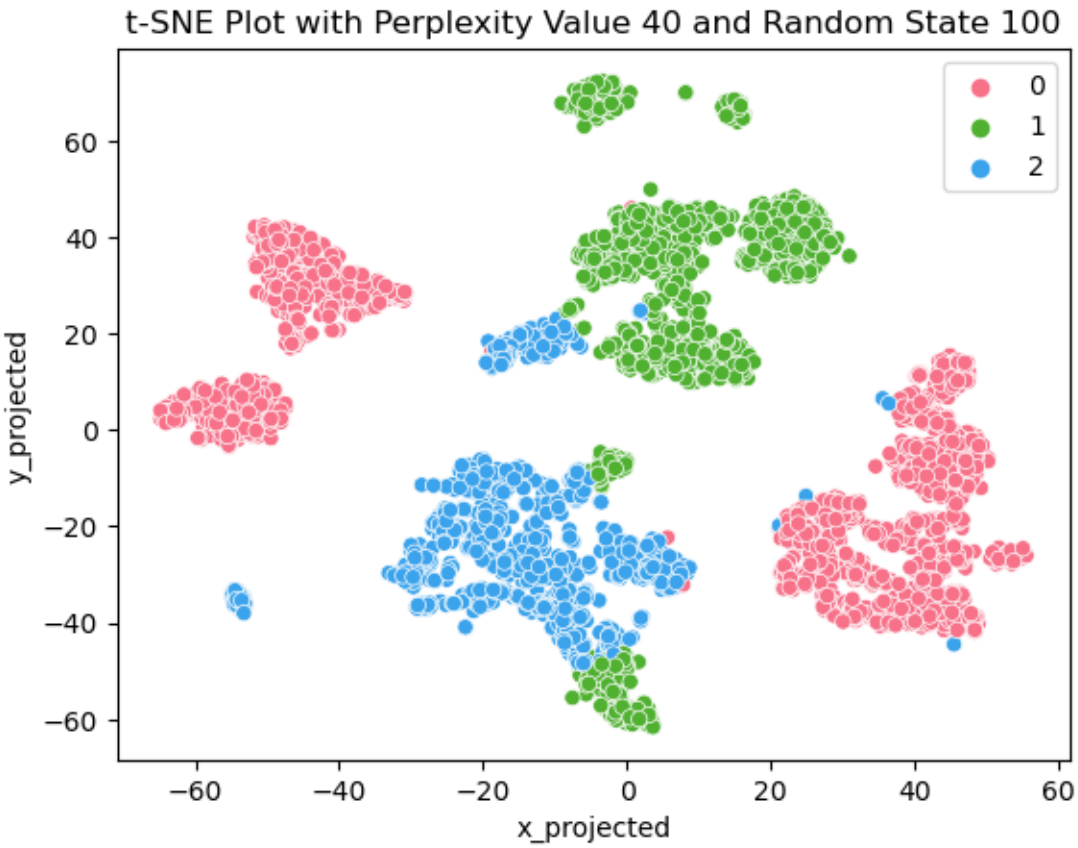

Average Silhouette Score wth HAC and complete Linkage
Average Silhouette Score of Clustering
0.35
0.30
0.25
0.20
0.15
5
10
15
20
25
30
Number of Clusters

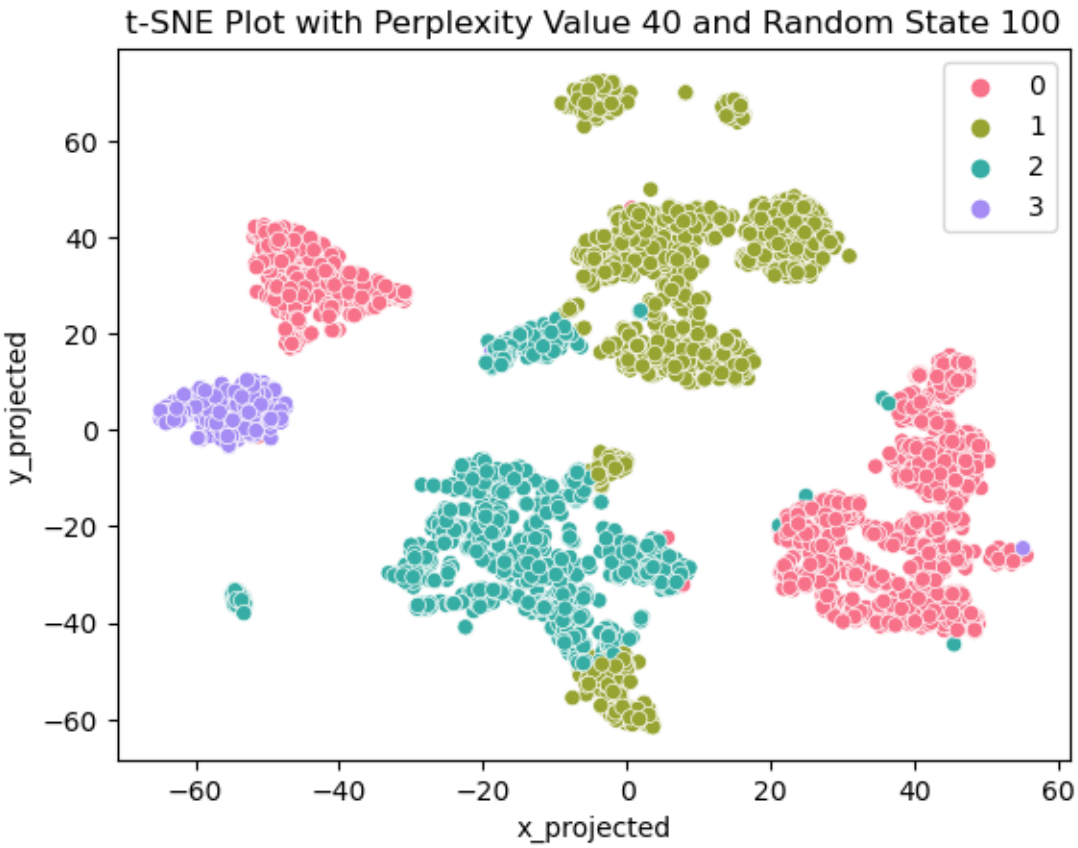


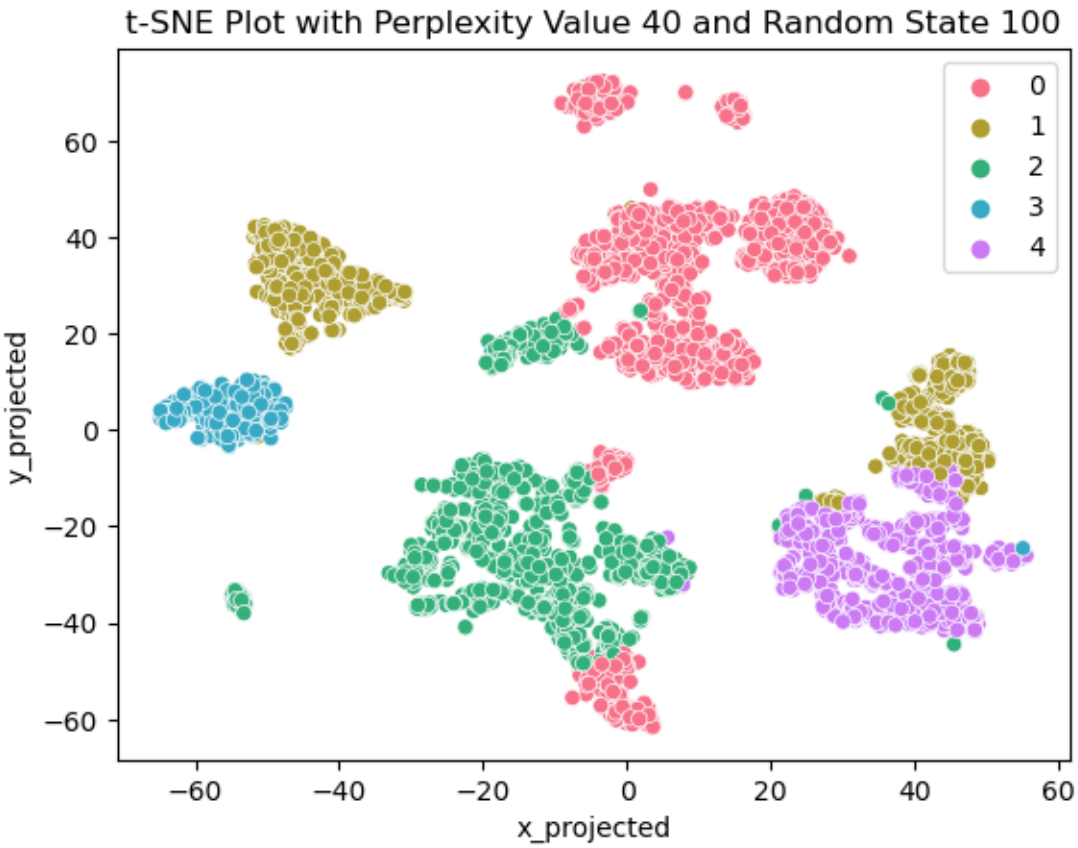
We think the cluster number should approach the number of states. If we only choose the highest Average Silhouette Score from above with low number of cluster, it is unreasonable. Thus, we selected ward linkage which K = 15. THe ward linkage have the average highest score, which the score is bigger than k = 7.

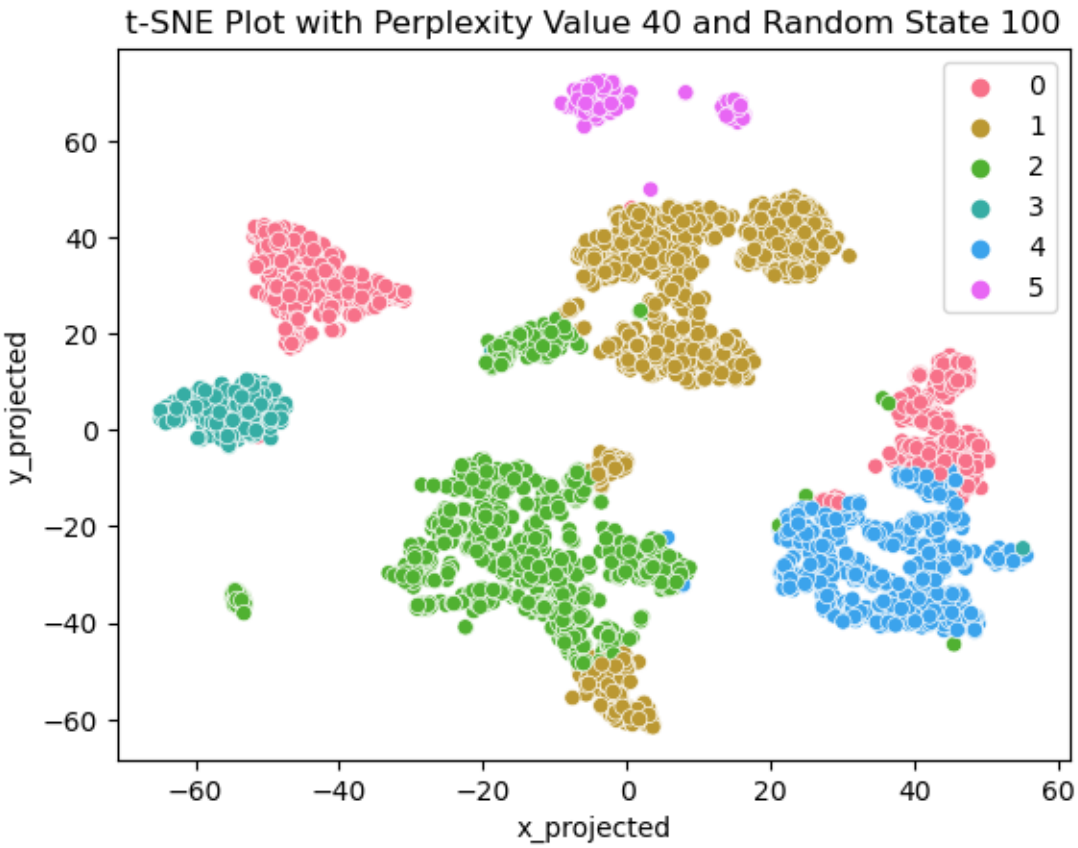
[63]:
```
for k in range(2,30):
    #Clustering from dendrogram with k clusters
    hac = AgglomerativeClustering(n_clusters=k, affinity='euclidean',␣
 ↪linkage='ward')
    df_combo['predicted_cluster'] = hac.fit_predict(df_num)

    #Map the resulting cluster labels onto our chosen t-SNE plot
    sns.scatterplot(x='x_projected',y='y_projected', hue='predicted_cluster',␣
 ↪palette=sns.color_palette("husl", k), data=df_combo)
    plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(40,␣
 ↪100))
    plt.legend(bbox_to_anchor=(1,1))
    plt.show()
```

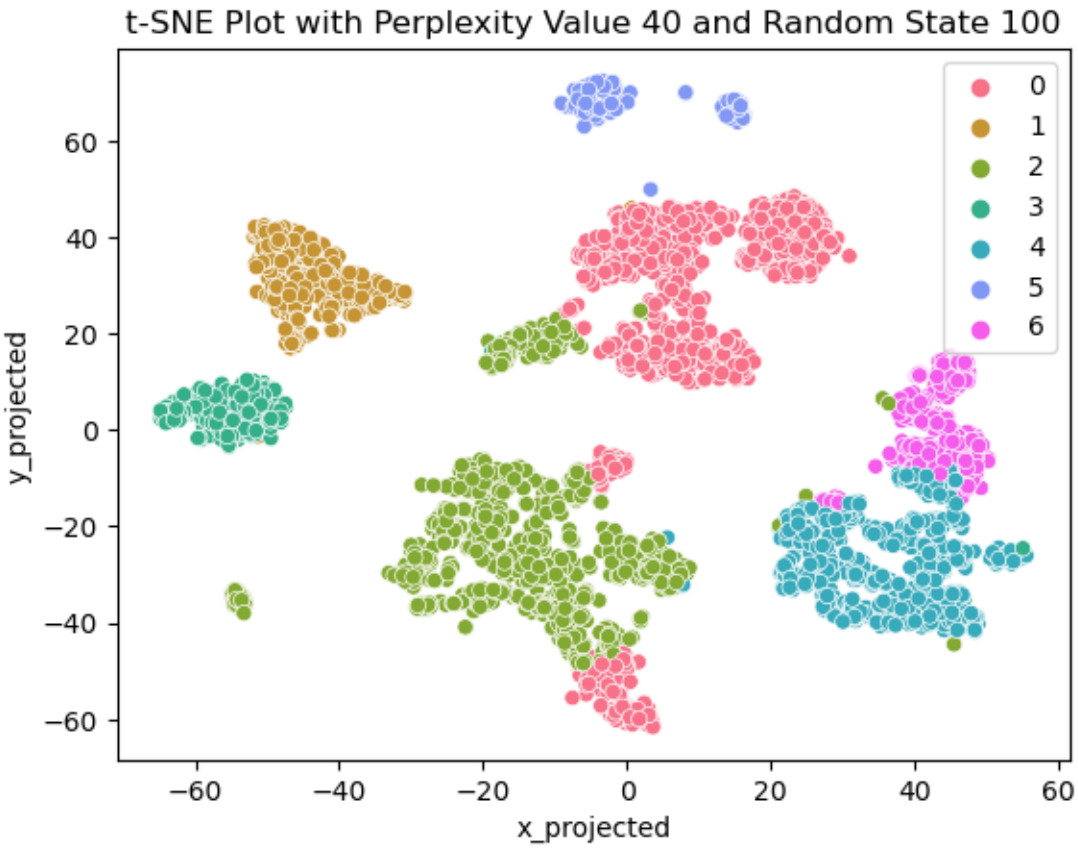

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

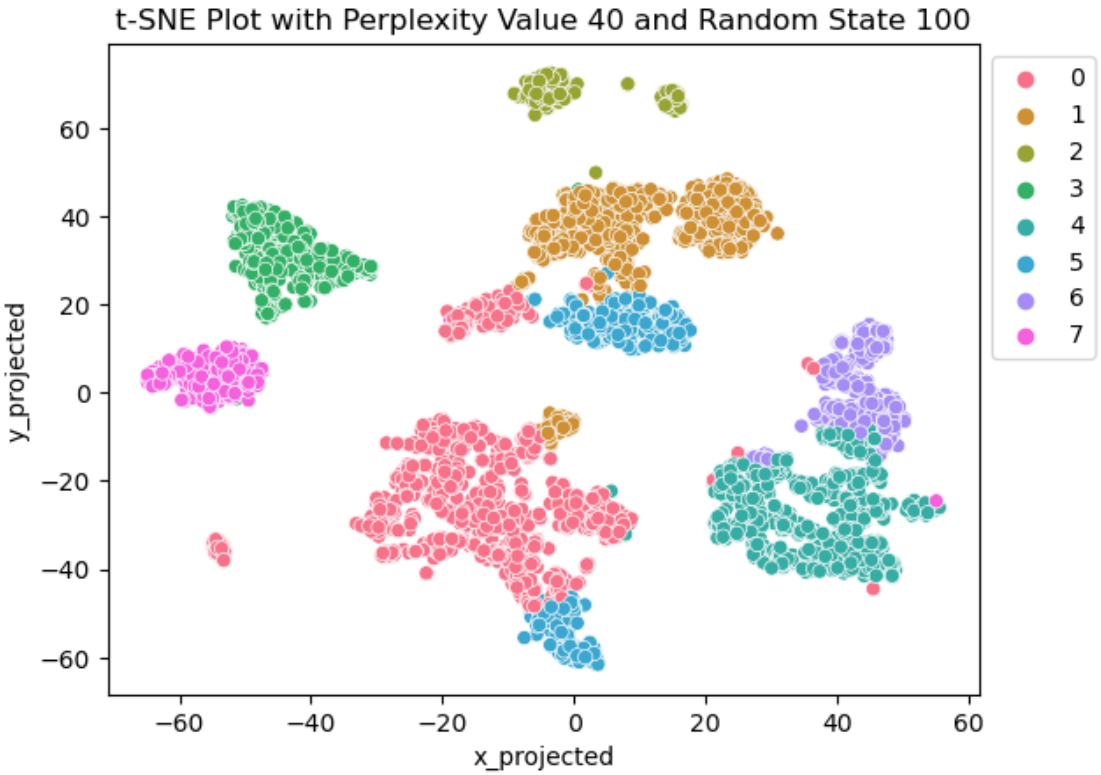

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
2
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

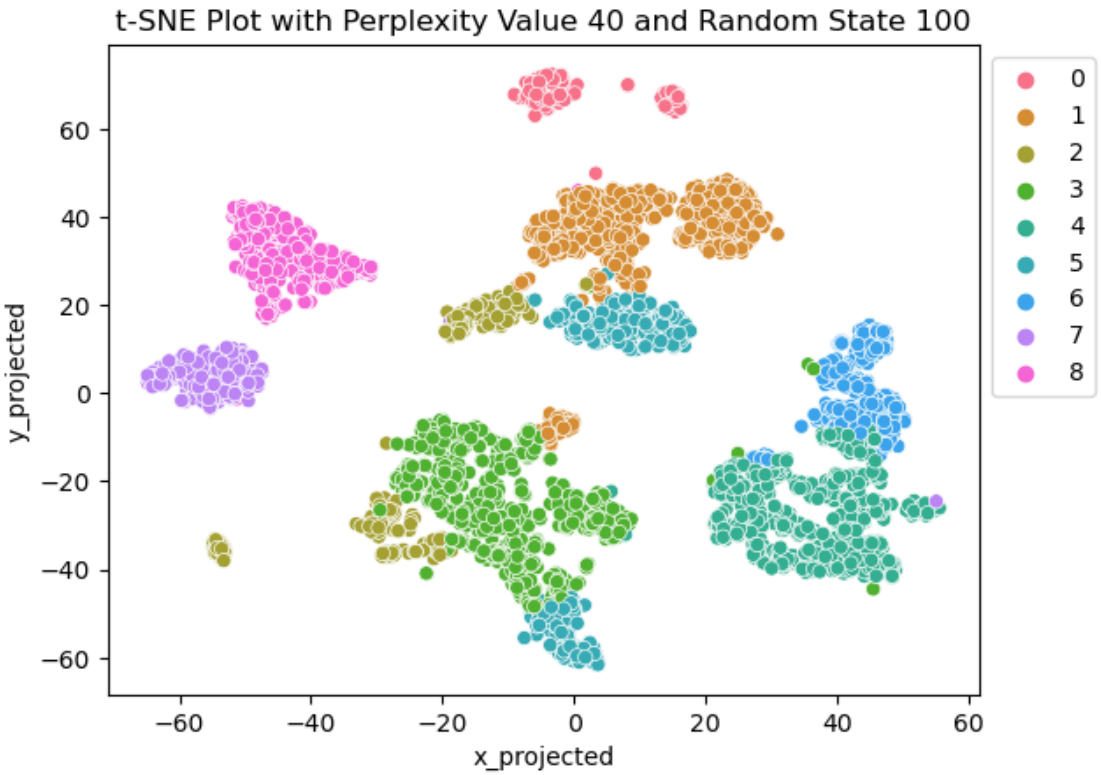

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
2
3
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

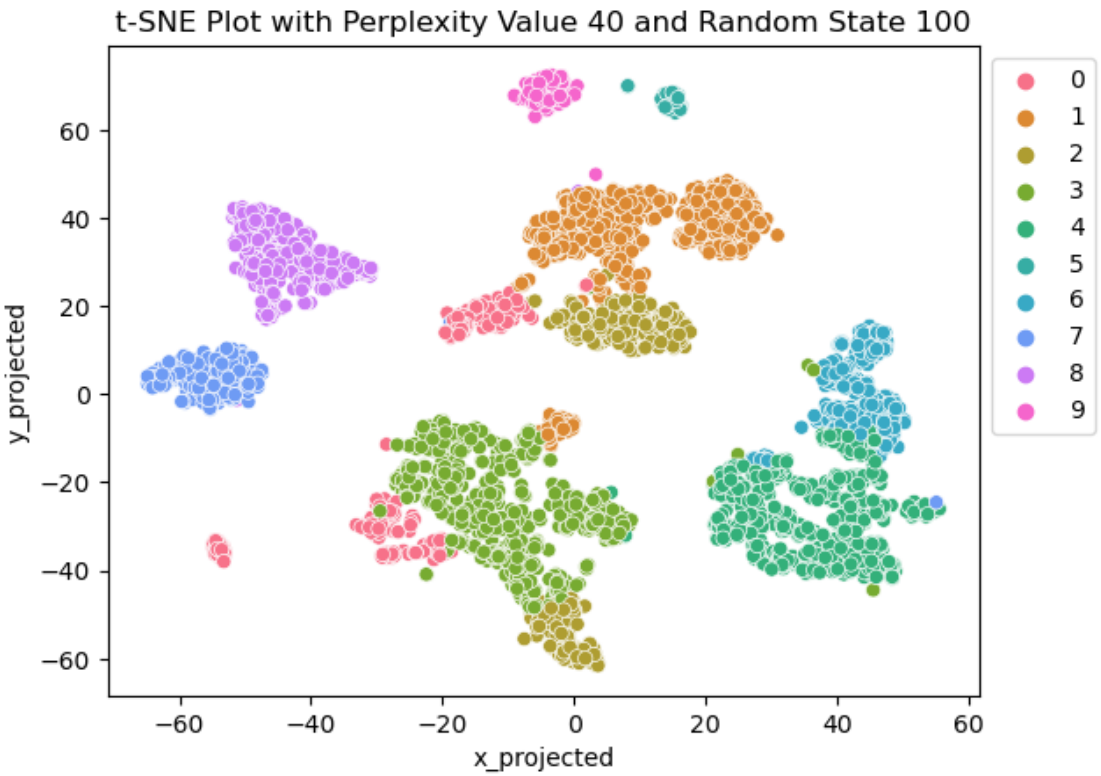

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
2
3
4
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

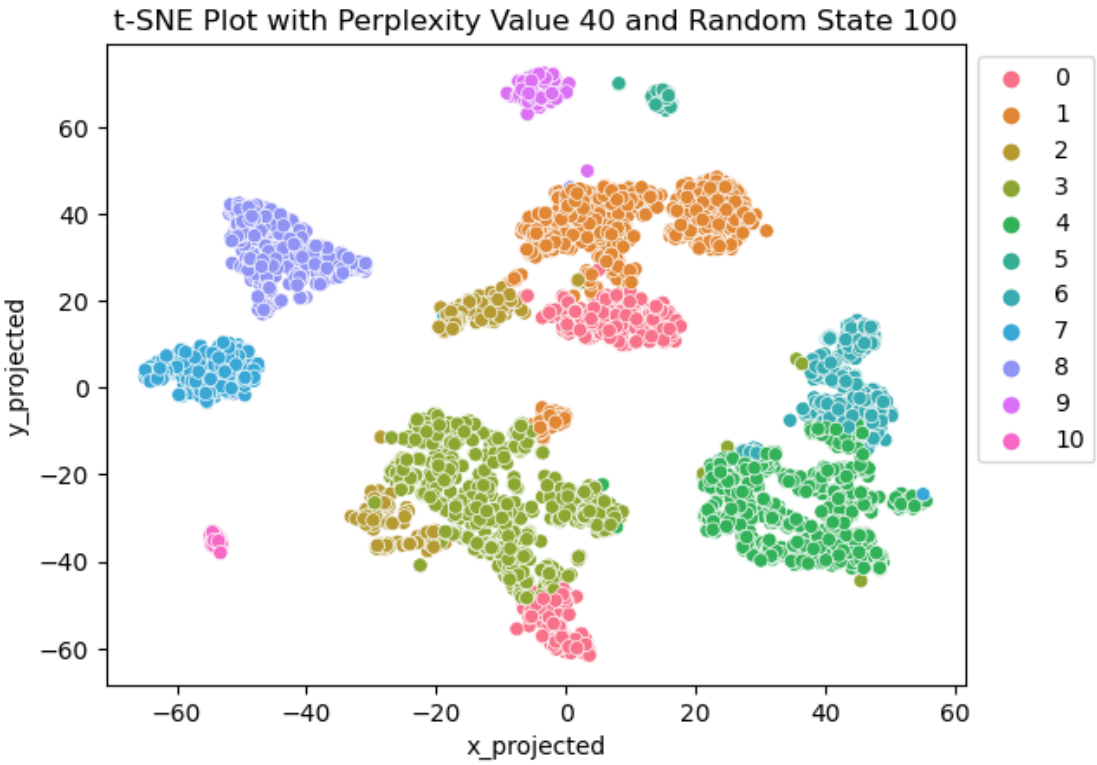

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
2
3
4
5
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

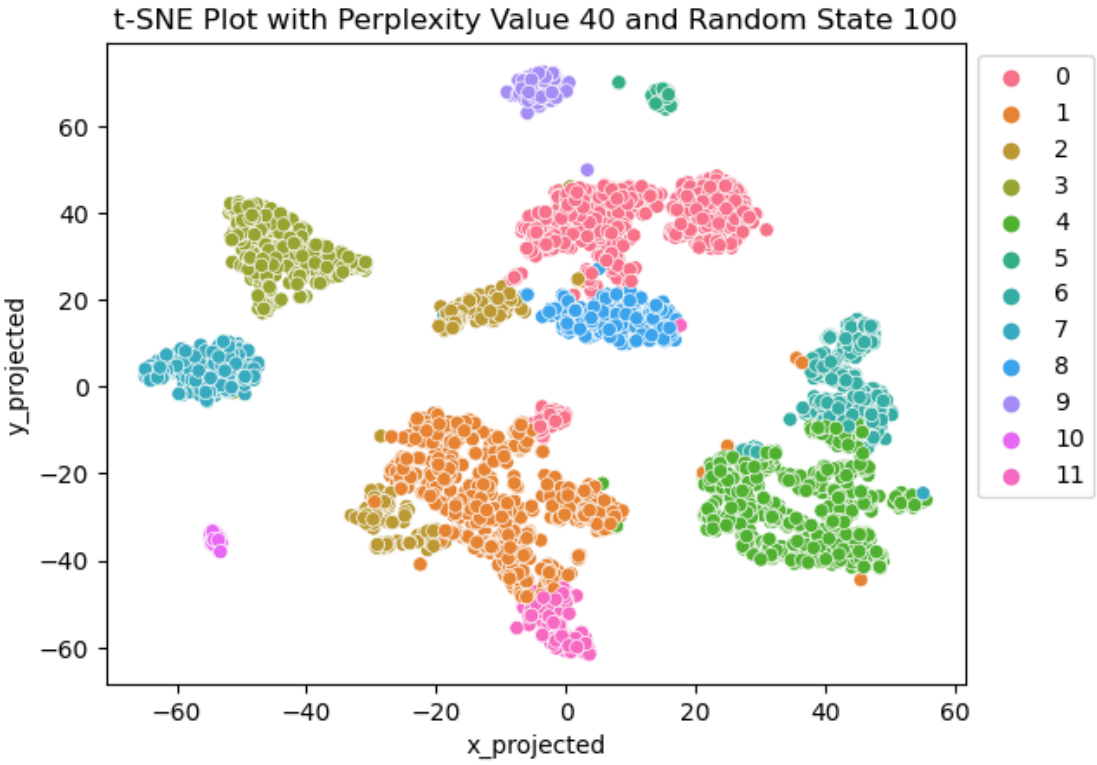

t-SNE Plot with Perplexity Value 40 and Random State 100
0
1
2
3
4
5
6
60
40
20
0
−20
−40
−60
y_projected
−60
−40
−20
0
20
40
60
x_projected

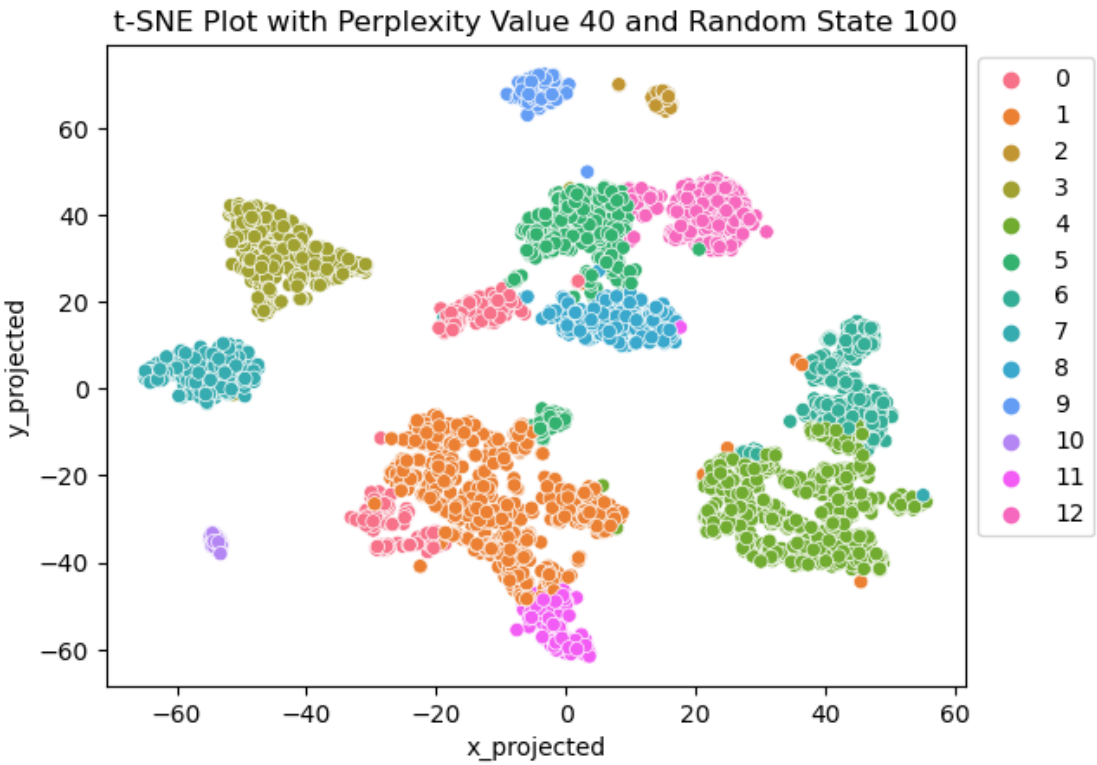

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7

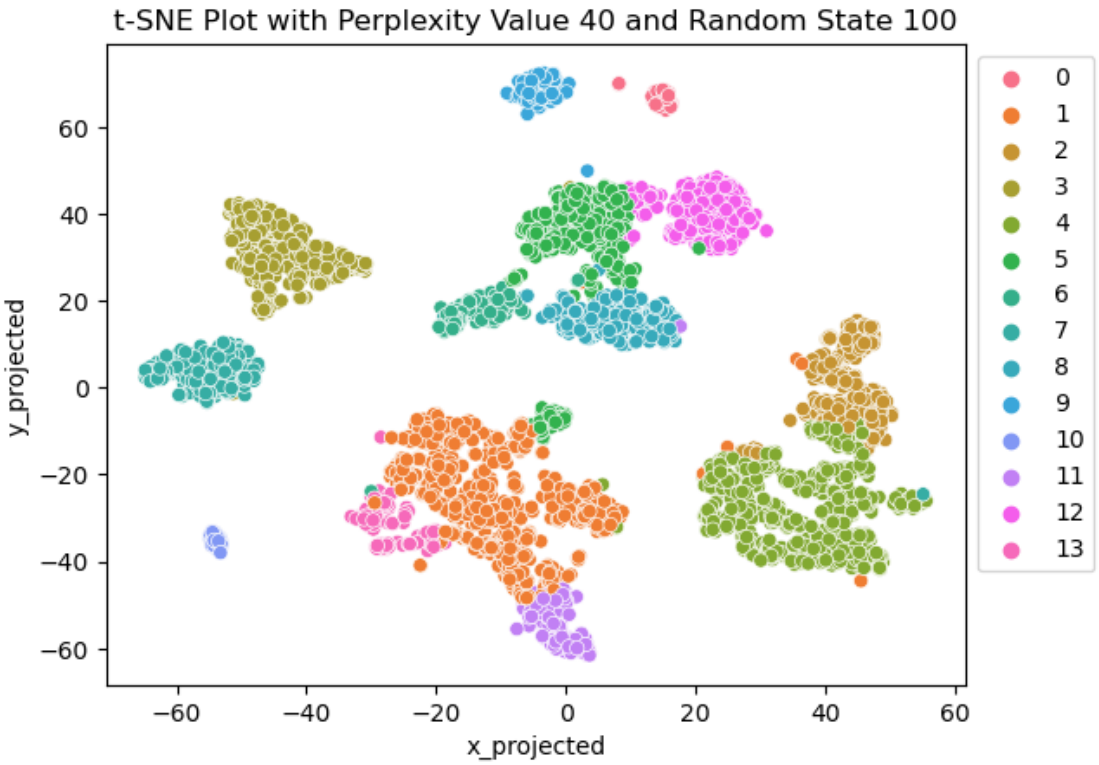

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8

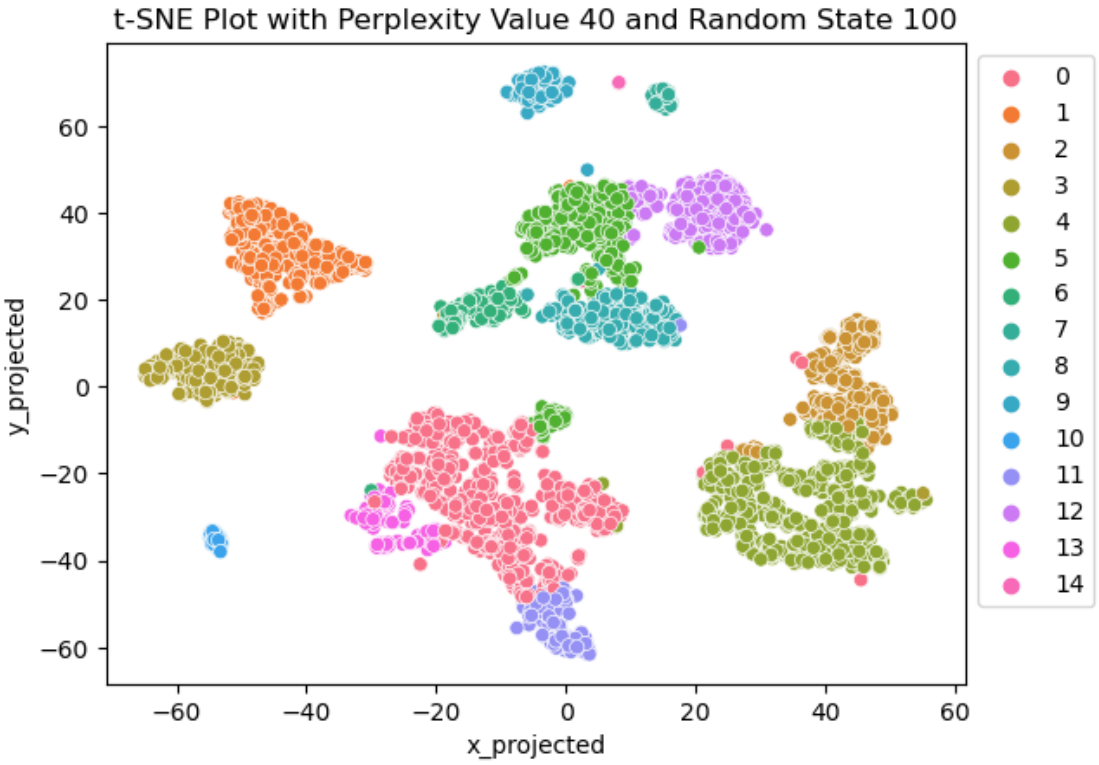

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9

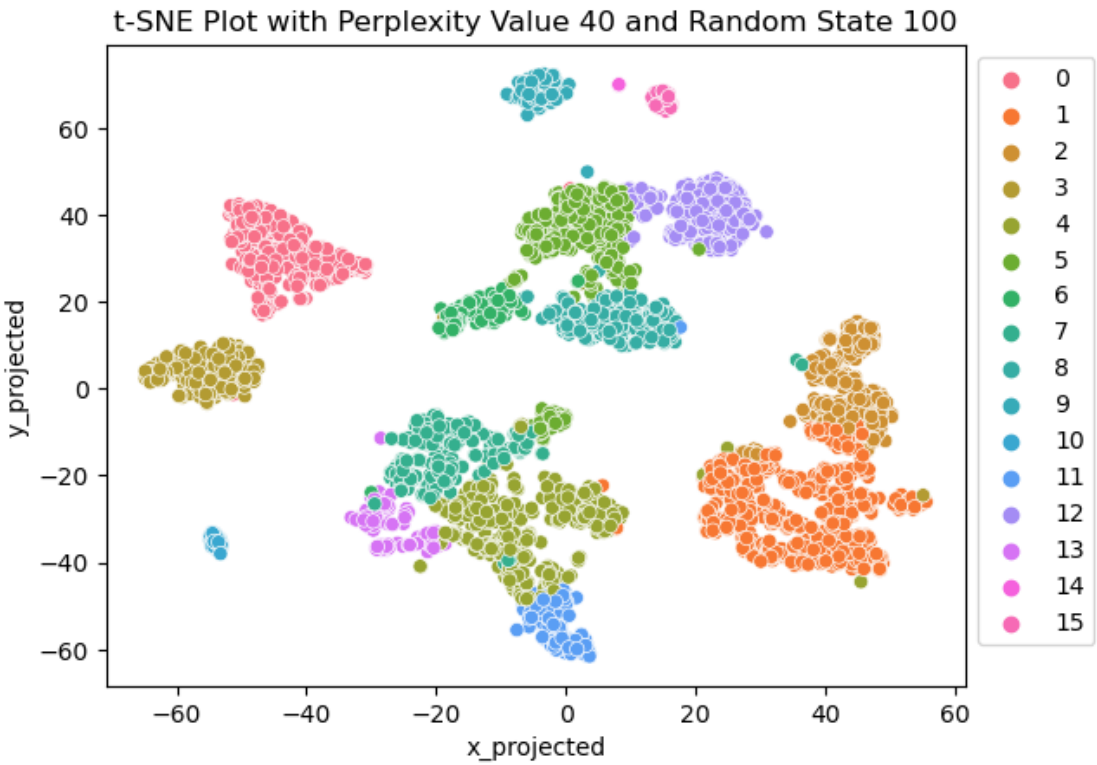

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10

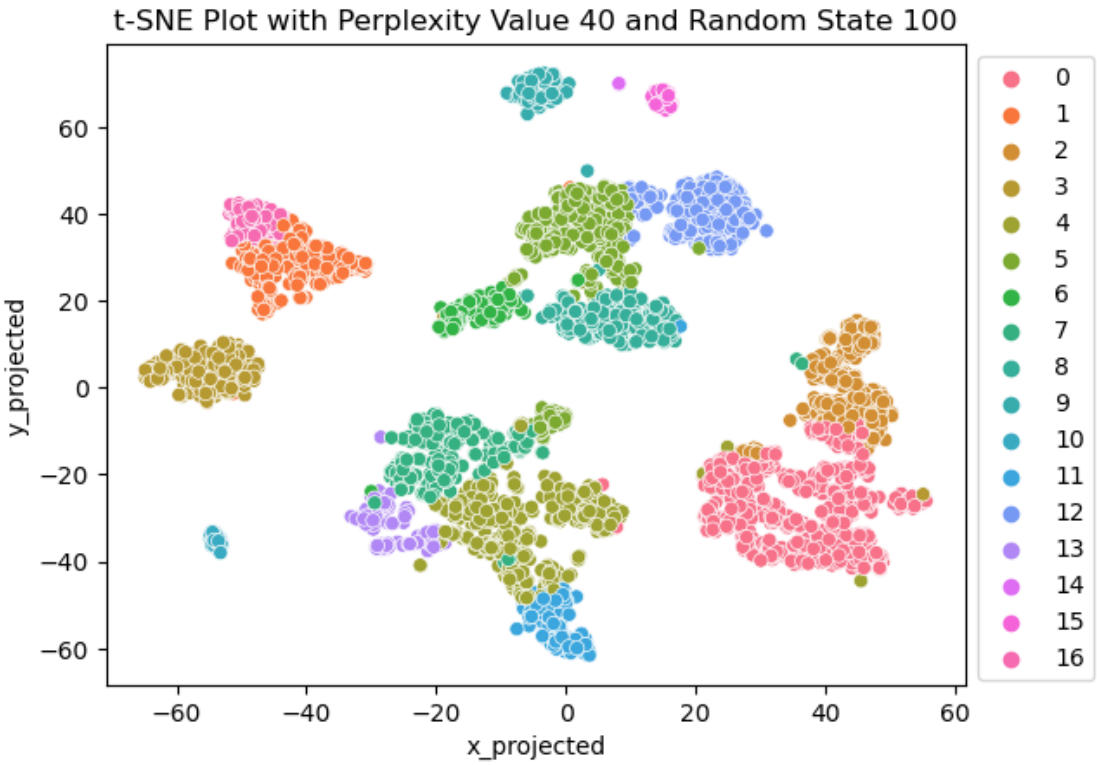

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11

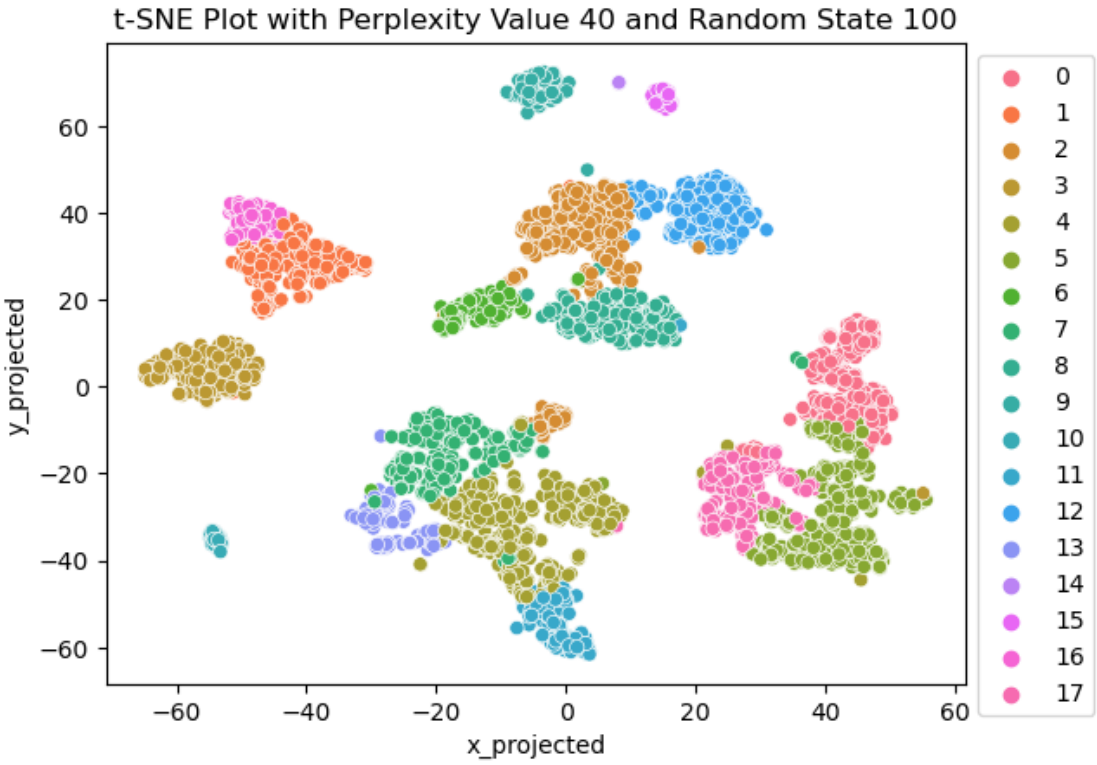

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12

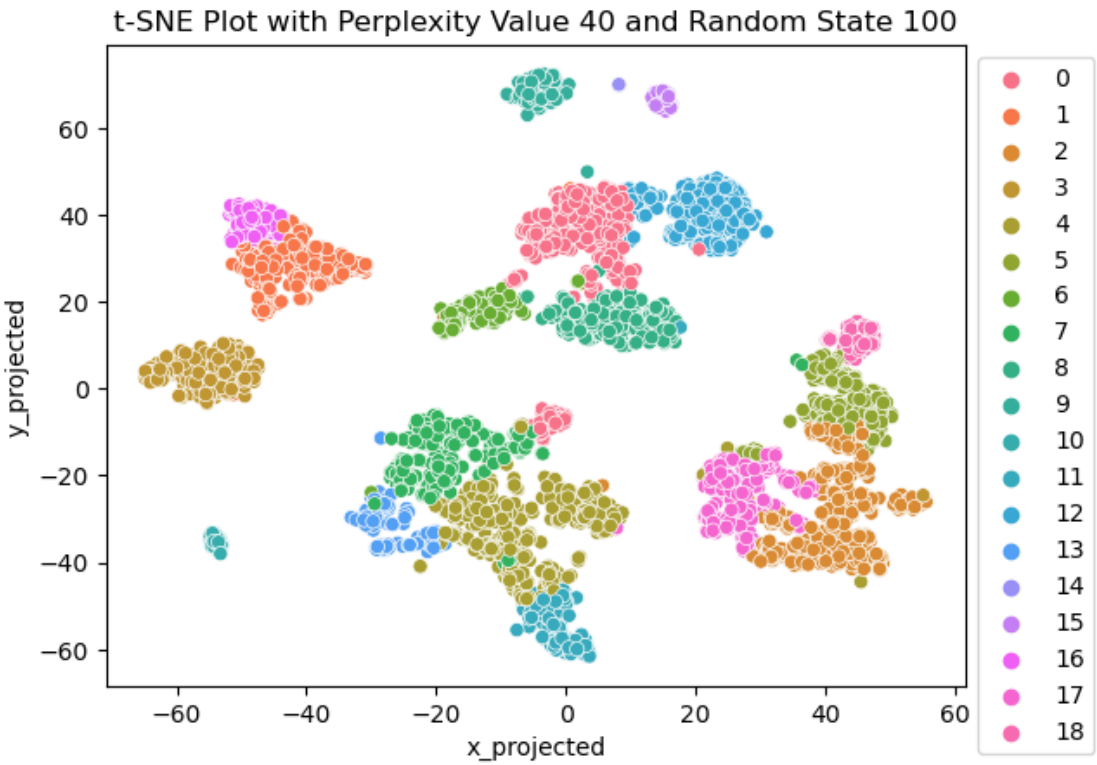

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13

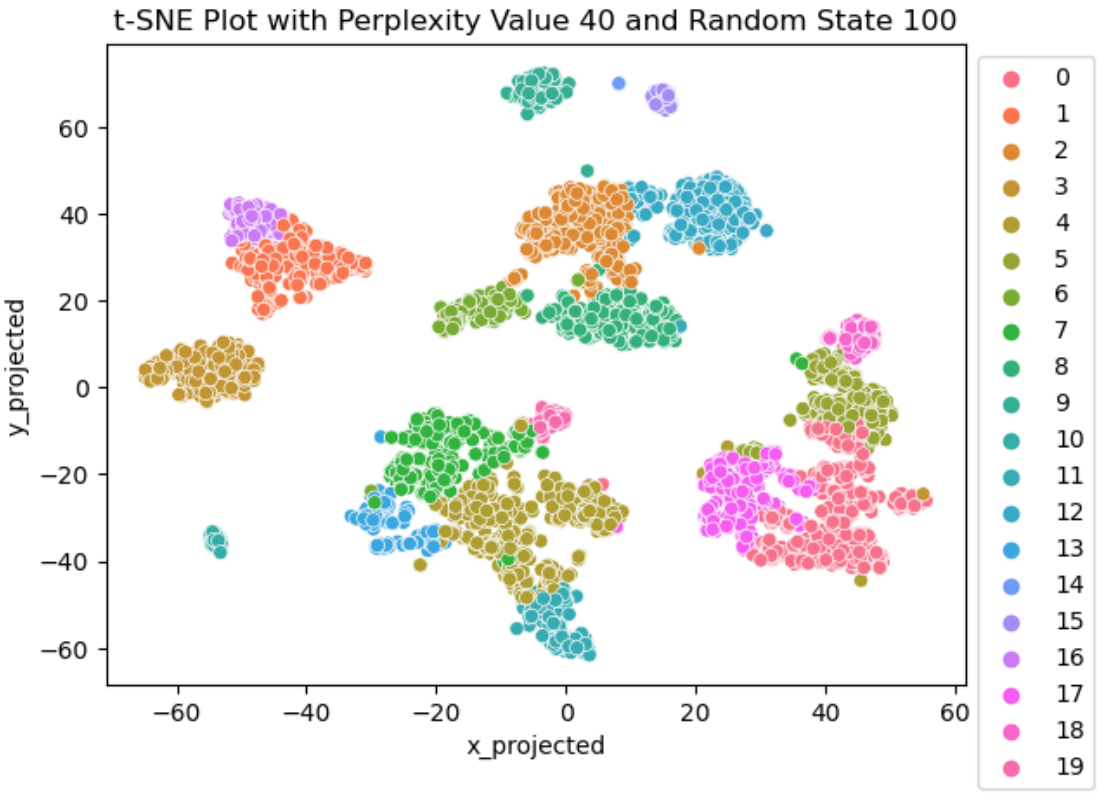

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

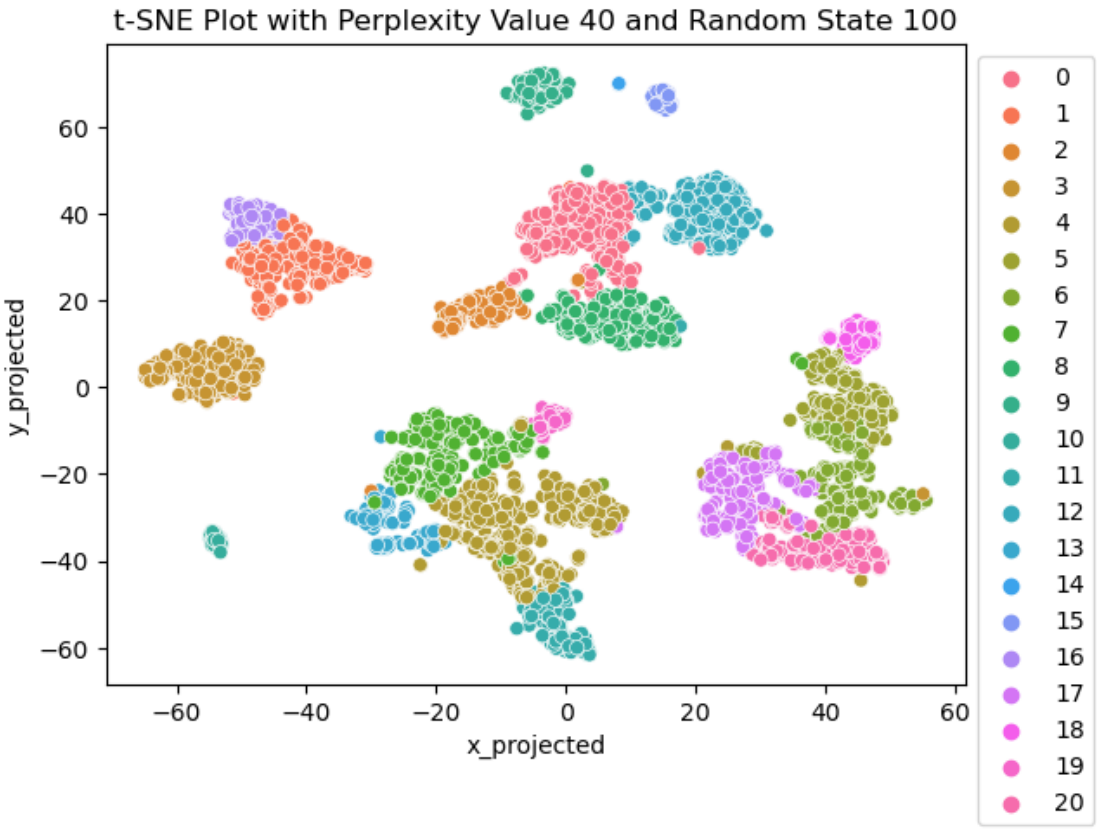

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

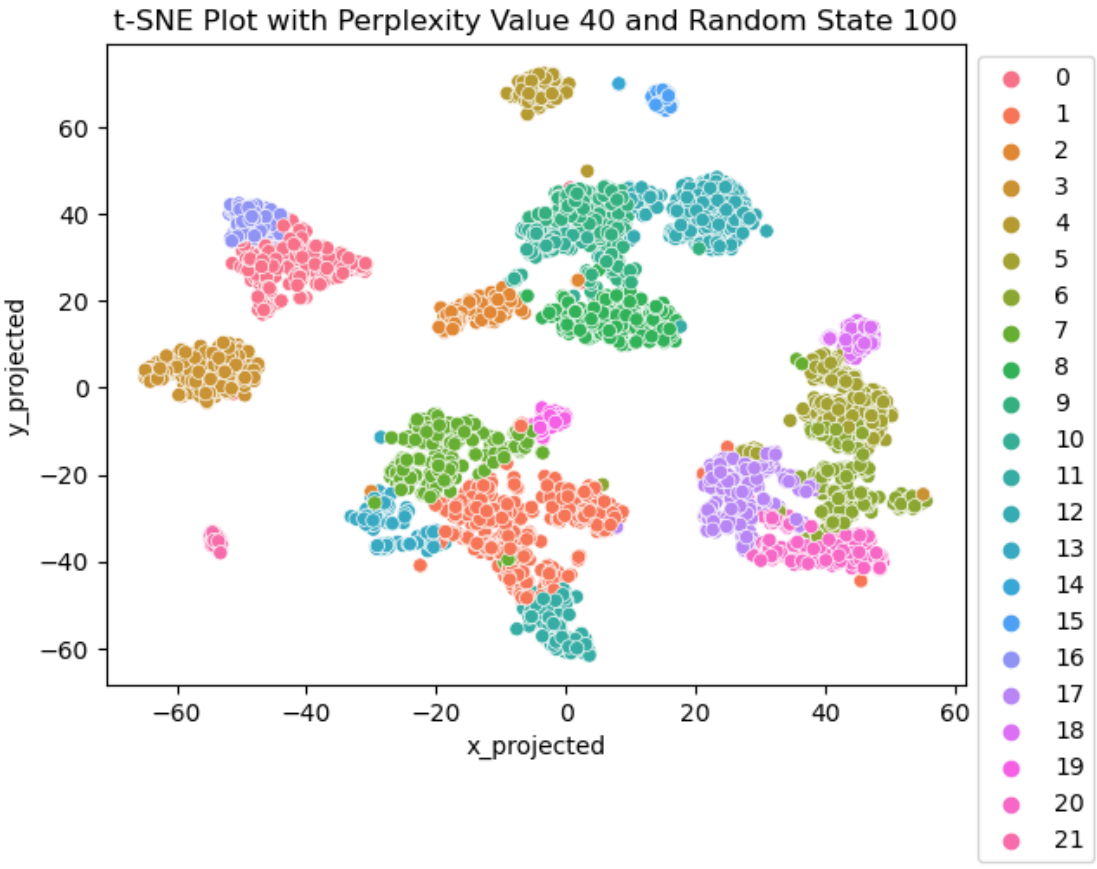

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

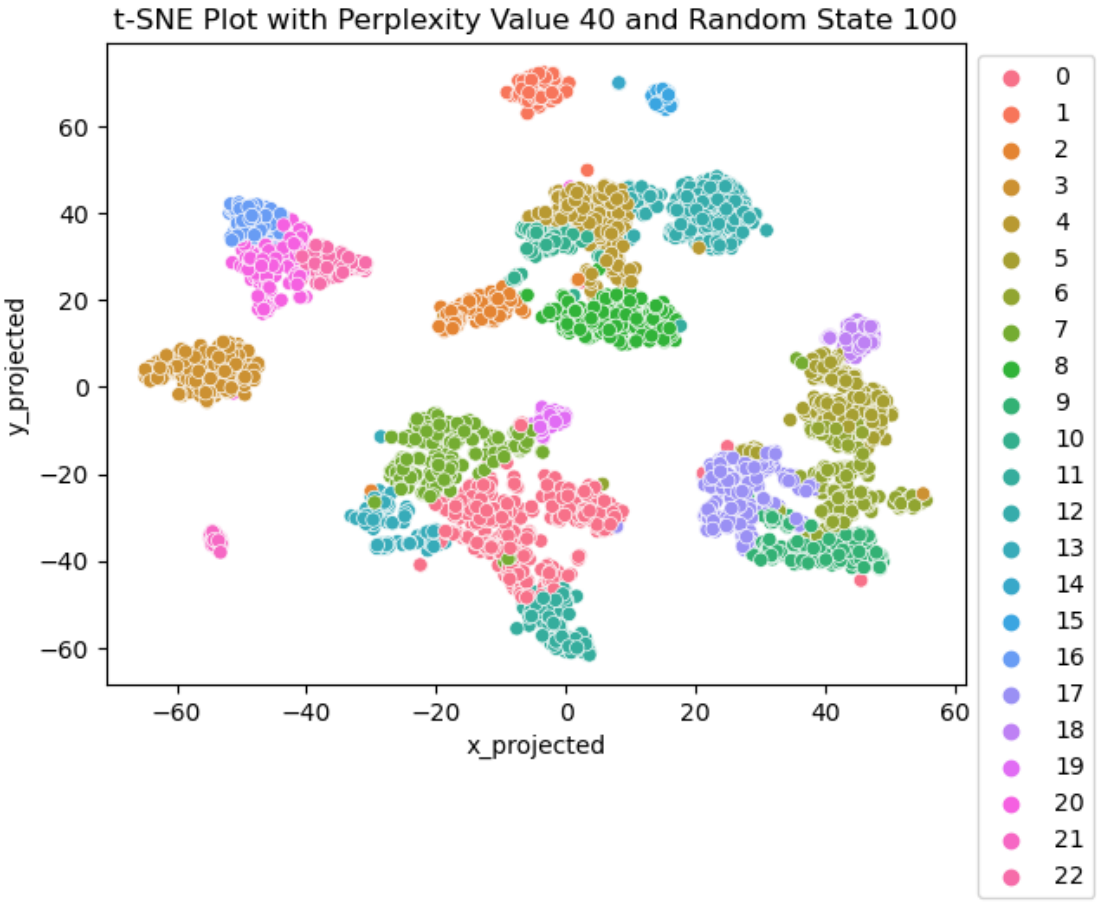

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

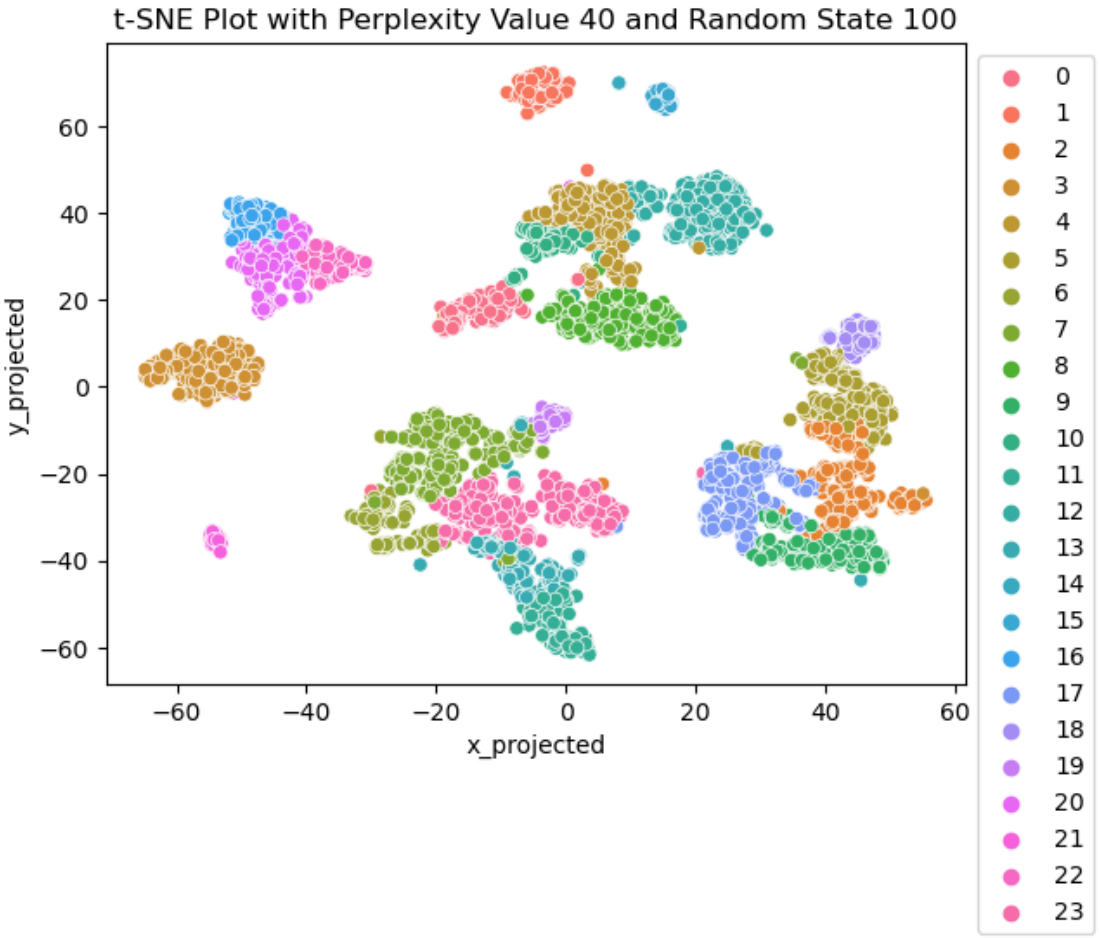

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

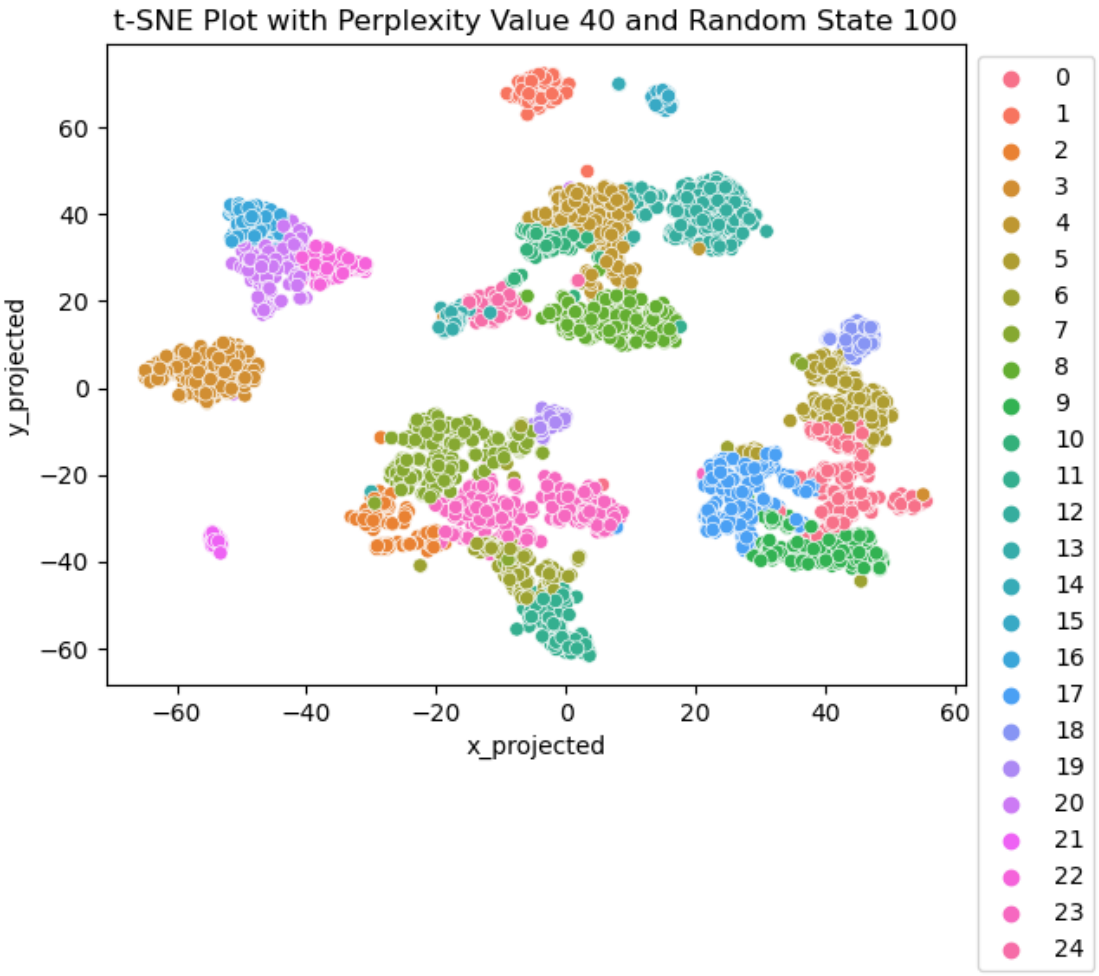

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

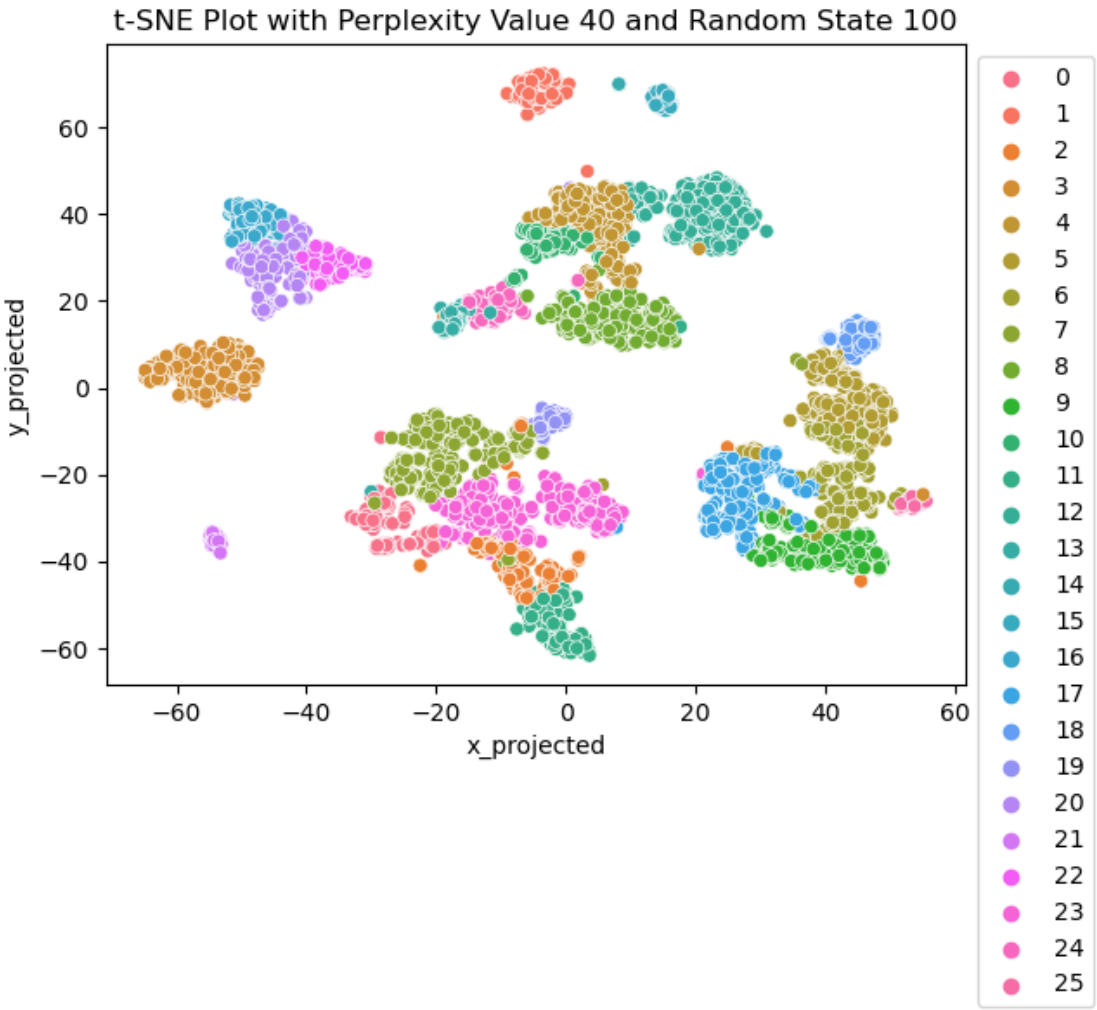

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

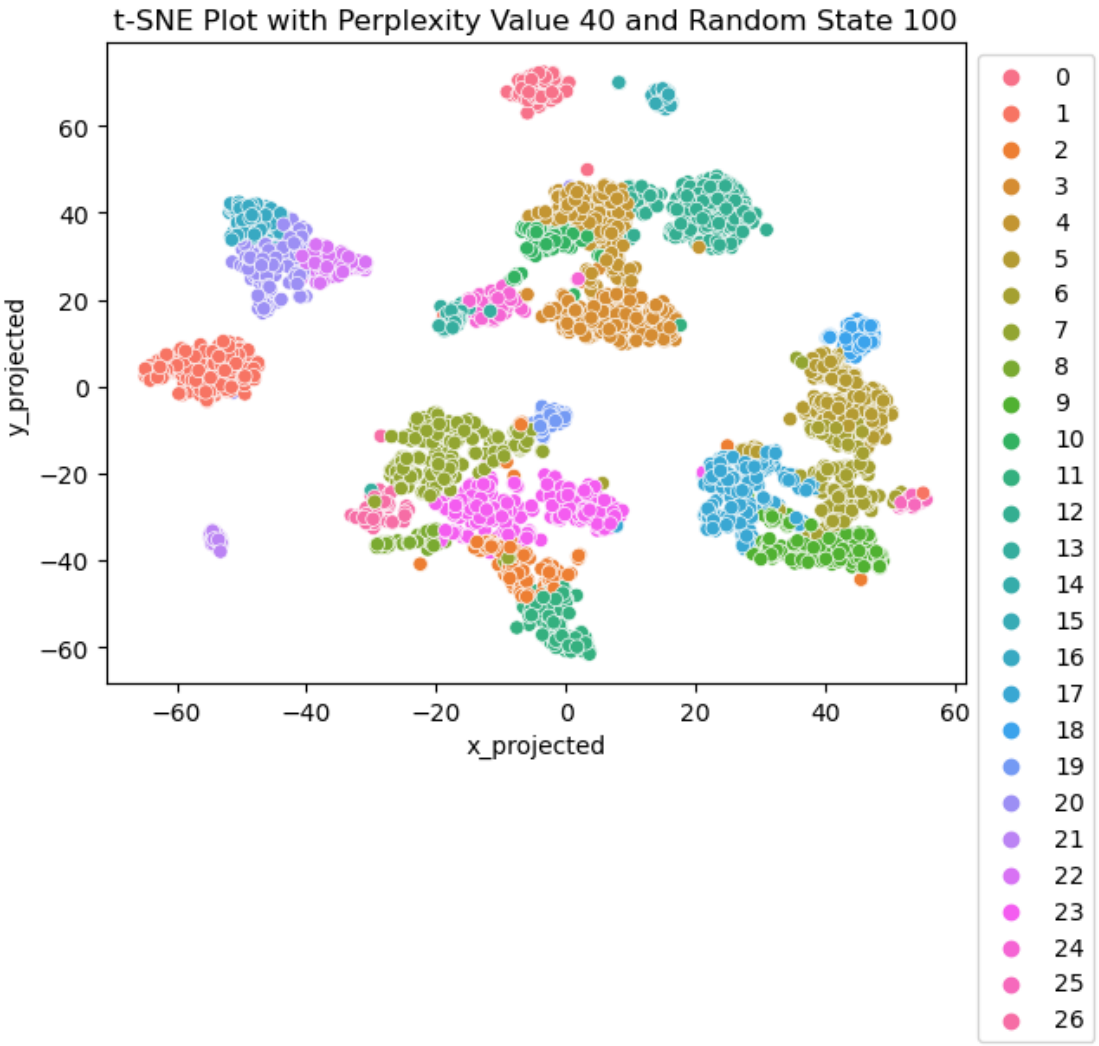

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

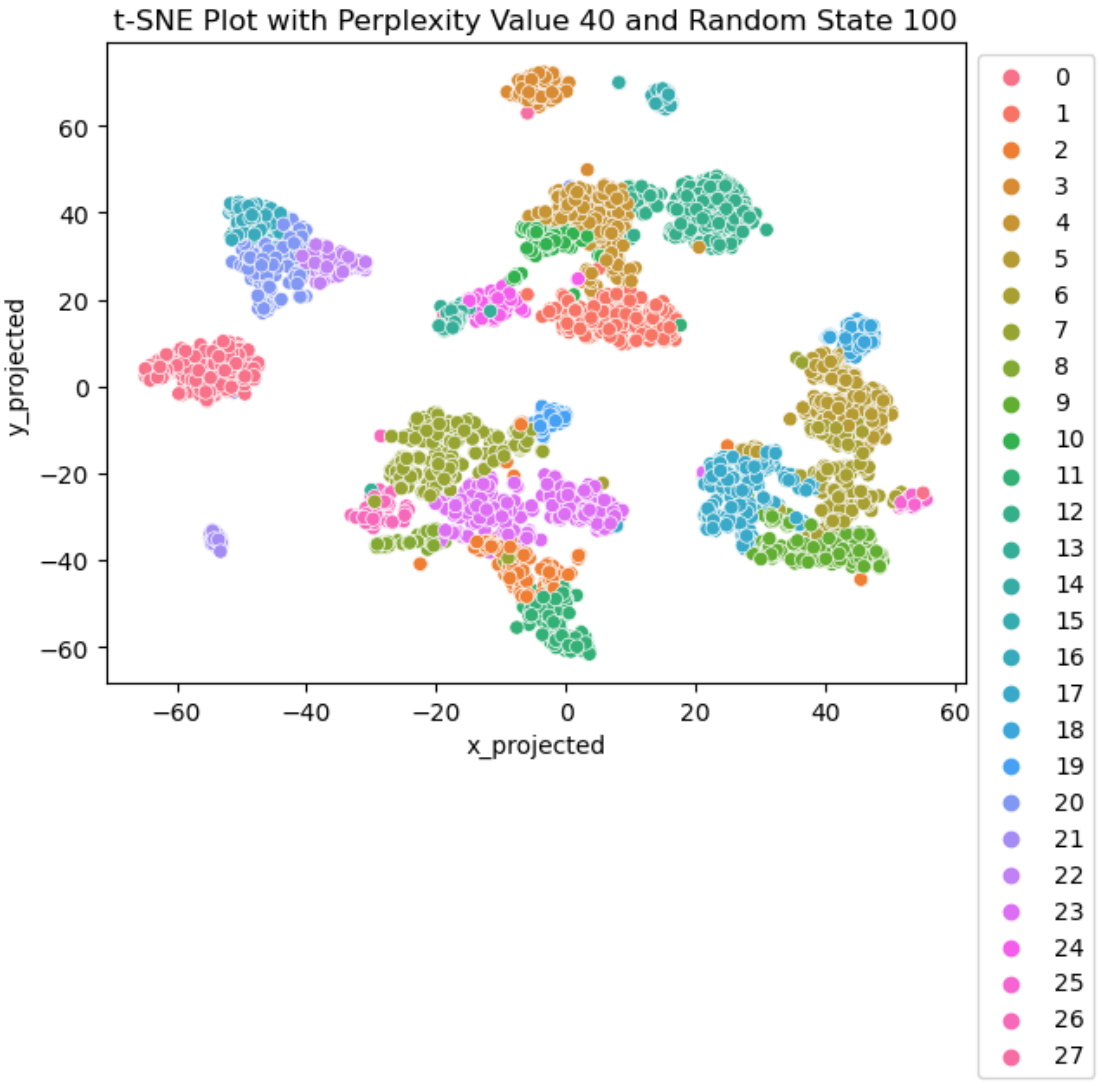

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

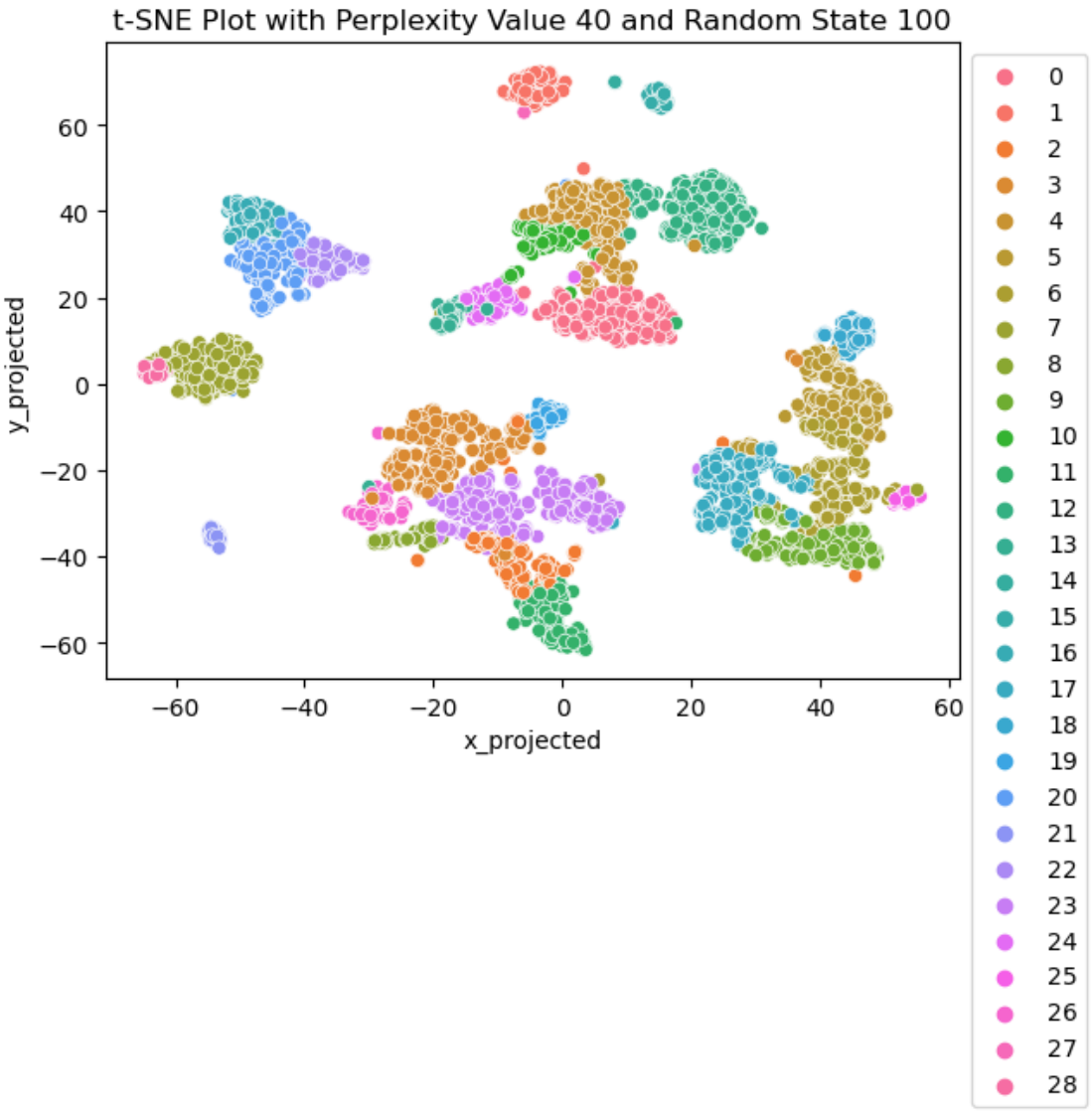

t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27



1.8.2 8.2. Clustering Algorithm

[79]:
```
#Clustering from dendrogram with 6 clusters
hac = AgglomerativeClustering(n_clusters=15, affinity='euclidean',␣
 ↪linkage='ward')
df_combo['predicted_cluster'] = hac.fit_predict(df_num)
```

[80]:
```
df_combo
```

[80]:

	Location_State	Vote_Data_Ben_Carson_Percent_of_Votes	\
0	South Carolina	0.382488	
1	Massachusetts	0.115207	

```
2          Louisiana                                  0.000000
3           Virginia                                  0.437788
4      Massachusetts                                  0.105991
…                …                                     …
3587      California                                  0.000000
3588         Arizona                                  0.000000
3589           Texas                                  0.216590
3590           Texas                                  0.000000
3591    South Dakota                                  0.000000

      Vote_Data_Bernie_Sanders_Percent_of_Votes  \
0                                         0.170
1                                         0.534
2                                         0.314
3                                         0.274
4                                         0.468
…                                         …
3587                                      0.511
3588                                      0.315
3589                                      0.236
3590                                      0.181
3591                                      0.589

      Vote_Data_Carly_Fiorina_Percent_of_Votes  \
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Chris_Christie_Percent_of_Votes  \
0                                           0.0
1                                           0.0
2                                           0.0
3                                           0.0
4                                           0.0
…                                           …
3587                                        0.0
3588                                        0.0
3589                                        0.0
3590                                        0.0
```

```
3591                                        0.0

      Vote_Data_Donald_Trump_Percent_of_Votes  \
0                                    0.403279
1                                    0.632787
2                                    0.486339
3                                    0.523497
4                                    0.308197
…                                         …
3587                                 0.904918
3588                                 0.540984
3589                                 0.431694
3590                                 0.000000
3591                                 0.750820

      Vote_Data_Hillary_Clinton_Percent_of_Votes  \
0                                          0.818
1                                          0.446
2                                          0.537
3                                          0.720
4                                          0.528
…                                          …
3587                                       0.464
3588                                       0.637
3589                                       0.678
3590                                       0.755
3591                                       0.411

      Vote_Data_Jeb_Bush_Percent_of_Votes  \
0                                0.528926
1                                0.000000
2                                0.000000
3                                0.000000
4                                0.000000
…                                   …
3587                             0.000000
3588                             0.000000
3589                             0.000000
3590                             0.000000
3591                             0.000000

      Vote_Data_John_Kasich_Percent_of_Votes  \
0                                   0.067293
1                                   0.222222
2                                   0.053208
3                                   0.076682
4                                   0.508607
```

```
…                                    …
3587                                0.079812
3588                                0.092332
3589                                0.000000
3590                                0.000000
3591                                0.150235

      Vote_Data_Marco_Rubio_Percent_of_Votes  …  \
0                                   0.322169  …
1                                   0.202552  …
2                                   0.169059  …
3                                   0.333333  …
4                                   0.397129  …
…                                      …  …
3587                                0.000000  …
3588                                0.000000  …
3589                                0.204147  …
3590                                0.000000  …
3591                                0.000000  …

      Vote_Data_No_Preference_Party  Vote_Data_No_Preference_Percent_of_Votes  \
0                               0.0                                  0.000000
1                               0.0                                  0.400000
2                               0.0                                  0.000000
3                               0.0                                  0.000000
4                               0.0                                  0.066667
…                              …                                      …
3587                            0.0                                  0.000000
3588                            0.0                                  0.000000
3589                            0.0                                  0.000000
3590                            0.0                                  0.000000
3591                            0.0                                  0.000000

      Vote_Data_Rand_Paul_Percent_of_Votes  \
0                                      0.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
…                                     …
3587                                   0.0
3588                                   0.0
3589                                   0.0
3590                                   0.0
3591                                   0.0

      Vote_Data_Rick_Santorum_Percent_of_Votes  \
```

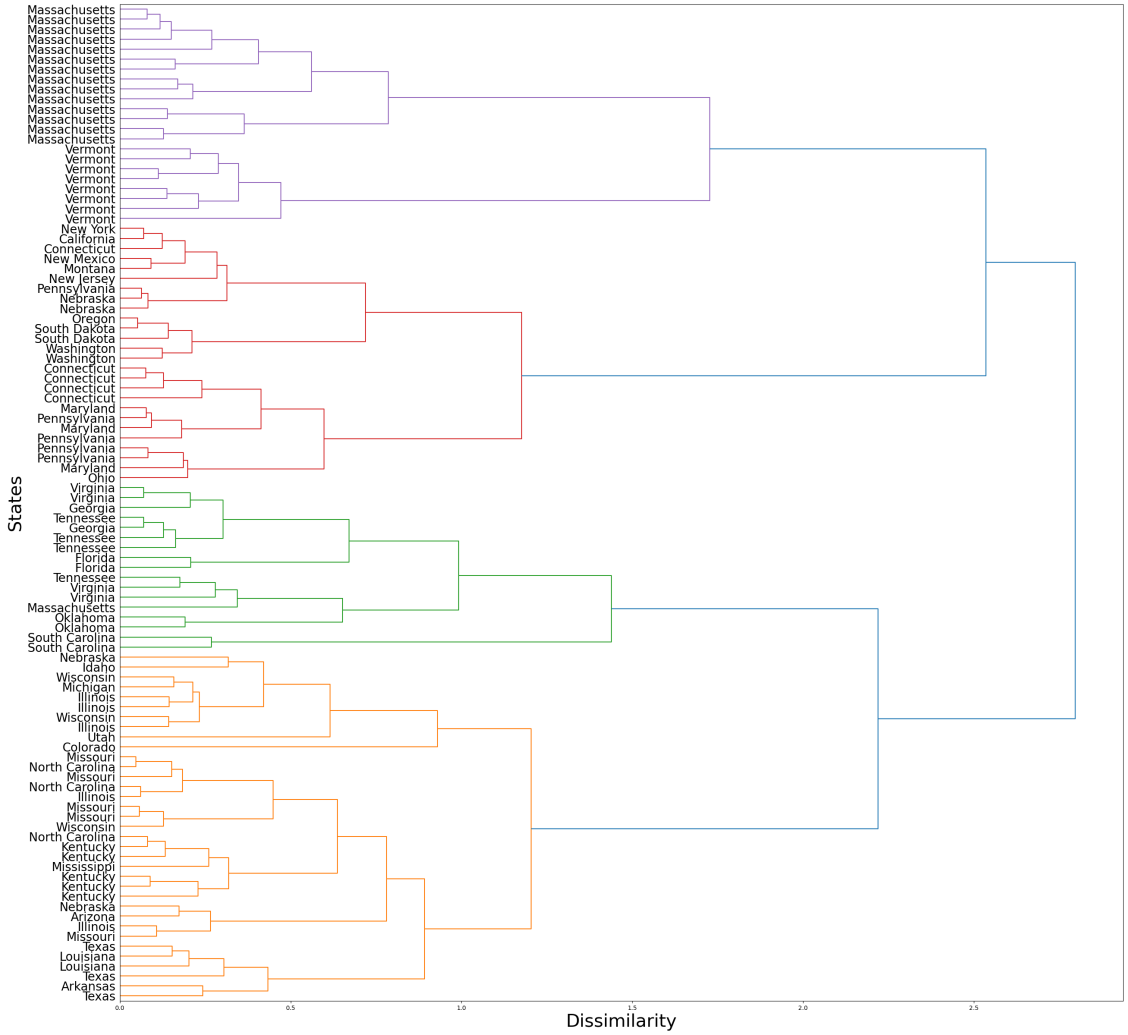
```
0                                          0.0
1                                          0.0
2                                          0.0
3                                          0.0
4                                          0.0
…                                          …
3587                                       0.0
3588                                       0.0
3589                                       0.0
3590                                       0.0
3591                                       0.0

      Vote_Data_Ted_Cruz_Percent_of_Votes  Vote_Data_Uncommitted_Party  \
0                                0.304071                          0.0
1                                0.125954                          0.0
2                                0.486005                          0.0
3                                0.202290                          0.0
4                                0.110687                          0.0
…                                   …                              …
3587                             0.104326                          0.0
3588                             0.366412                          0.0
3589                             0.473282                          0.0
3590                             0.000000                          0.0
3591                             0.276081                          0.0

      Vote_Data_Uncommitted_Percent_of_Votes  x_projected  y_projected  \
0                                        0.0    15.333004    65.792213
1                                        0.0   -48.632351    38.450615
2                                        0.0    -7.403143   -43.584099
3                                        0.0    17.125759    43.247463
4                                        0.0   -32.474995    27.306452
…                                        …           …            …
3587                                     0.0    40.149212   -39.740902
3588                                     0.0     0.413865   -43.407272
3589                                     0.0    13.933252    18.729895
3590                                     0.0   -53.272396   -38.035904
3591                                     0.0    42.094490   -25.580858

      predicted_cluster
0                     7
1                     1
2                     0
3                    12
4                     1
…                   …
3587                  4
3588                  0
```

```
3589                  8
3590                 10
3591                  4

[3592 rows x 22 columns]
```

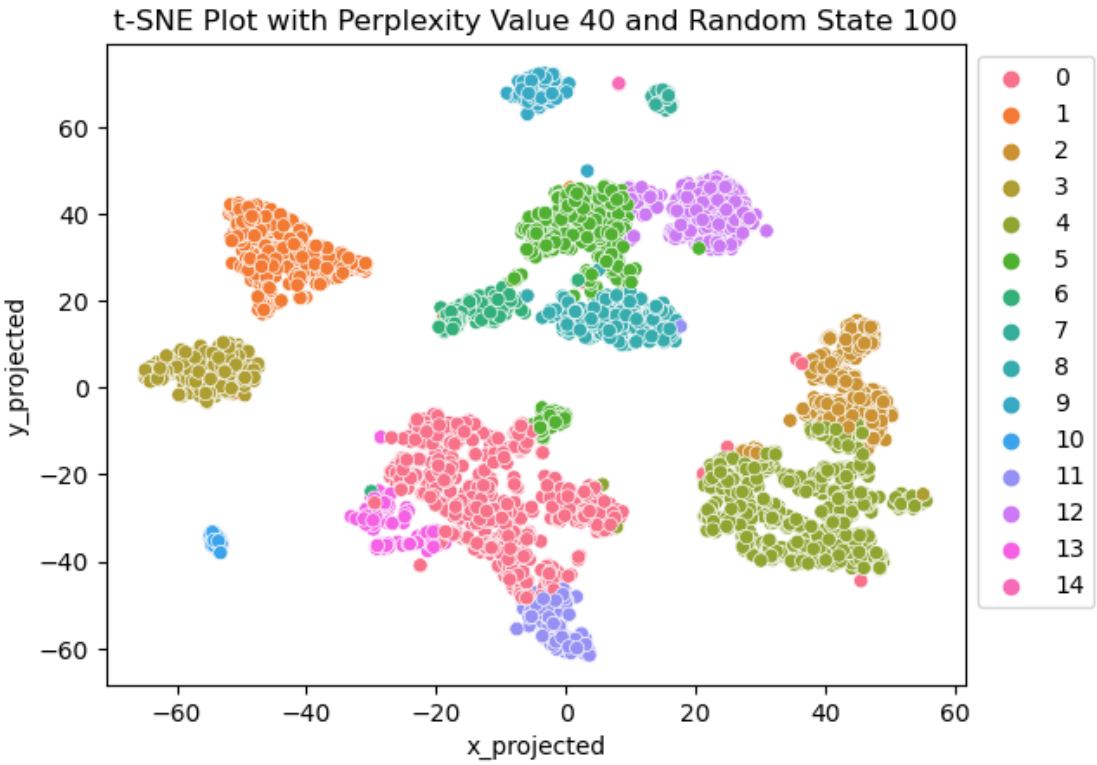
1.8.3 8.3. Clustering Algorithm Results Presentation

We randomly selected 100 values. The purpose is that because the dataset is too large, the hierarchical diagram in jupyter notebook running HAC is too intricate and complex. So we randomly selected 100 does not affect the algorithm itself, just for the aesthetics of the data visualization.

[81]:
```
df_random = df_label.sample(100).reset_index(drop=True)
df_random_num = df_random.drop(['Location_State'], axis = 1)
df_random_num

dm = pdist(df_random_num, metric='euclidean')
S = linkage(dm, method='ward')

fig, ax = plt.subplots(figsize=(30, 30))
d = dendrogram(S, orientation='right', labels=df_random['Location_State'].
 ↪array, ax=ax)
ax.set_xlabel('Dissimilarity', fontsize=30)
ax.set_ylabel('States', fontsize=30)
ax.set_yticklabels(ax.get_yticklabels(), fontsize=20)
plt.show()
```



[82]:
```
for k in [15]:
    #Clustering from dendrogram with k clusters
    hac = AgglomerativeClustering(n_clusters=k, affinity='euclidean',␣
↪linkage='ward')
    df_combo['predicted_cluster'] = hac.fit_predict(df_num)

    #Map the resulting cluster labels onto our chosen t-SNE plot
    sns.scatterplot(x='x_projected',y='y_projected', hue='predicted_cluster',␣
↪palette=sns.color_palette("husl", k), data=df_combo)
    plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(40,␣
↪100))
    plt.legend(bbox_to_anchor=(1,1))
    plt.show()
```



1.8.4 8.4. Assessing Clustering Separation and Cohesion

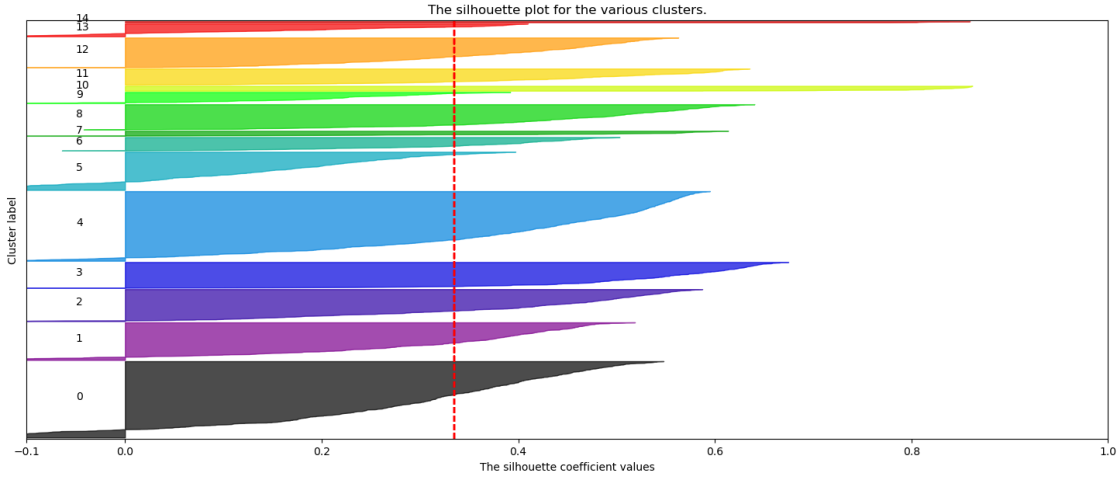
[68]:
```
def show_silhouette_plots(X,cluster_labels):

    # This package allows us to use "color maps" in our visualizations
    import matplotlib.cm as cm

    #How many clusters in your clustering?
    n_clusters=len(np.unique(cluster_labels))

    # Create a subplot with 1 row and 2 columns
    fig, ax1 = plt.subplots(1, 1)
    fig.set_size_inches(18, 7)

    # The 1st subplot is the silhouette plot
    # The silhouette coefficient fcan range from -1, 1 but in this example all
    # lie within [-0.1, 1]
    ax1.set_xlim([-0.1, 1])
    # The (n_clusters+1)*10 is for inserting blank space between silhouette
    # plots of individual clusters, to demarcate them clearly.
    ax1.set_ylim([0, len(X) + (n_clusters + 1) * 10])
```

```
    # The silhouette_score gives the average value for all the samples.
    # This gives a perspective into the density and separation of the formed
    # clusters
    silhouette_avg = silhouette_score(X, cluster_labels)
    print("For n_clusters =", n_clusters,
          "The average silhouette_score is :", silhouette_avg)

    # Compute the silhouette scores for each sample
    sample_silhouette_values = silhouette_samples(X, cluster_labels)

    y_lower = 10
    for i in range(n_clusters):
        # Aggregate the silhouette scores for samples belonging to
        # cluster i, and sort them
        ith_cluster_silhouette_values = \
            sample_silhouette_values[cluster_labels == i]


        ith_cluster_silhouette_values.sort()

        size_cluster_i = ith_cluster_silhouette_values.shape[0]
        y_upper = y_lower + size_cluster_i

        color = cm.nipy_spectral(float(i) / n_clusters)
        ax1.fill_betweenx(np.arange(y_lower, y_upper),
                          0, ith_cluster_silhouette_values,
                          facecolor=color, edgecolor=color, alpha=0.7)

        # Label the silhouette plots with their cluster numbers at the middle
        ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

        # Compute the new y_lower for next plot
        y_lower = y_upper + 10  # 10 for the 0 samples

        ax1.set_title("The silhouette plot for the various clusters.")
        ax1.set_xlabel("The silhouette coefficient values")
        ax1.set_ylabel("Cluster label")

        # The vertical line for average silhouette score of all the values
        ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

        ax1.set_yticks([])  # Clear the yaxis labels / ticks
        ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])


    plt.show()
```

```
    return
```

[104]:
```
#Clustering from dendrogram with 15 clusters
hac = AgglomerativeClustering(n_clusters=15, affinity='euclidean',␣
 ↪linkage='ward')
df_compare = df_num.copy()
df_compare = hac.fit_predict(df_num)


show_silhouette_plots(X,df_compare)
```

```
For n_clusters = 15 The average silhouette_score is : 0.3350670274512757
```



a.) each of the clusters and b.) the overall clustering. Are there are any objects that have poor cohesion with their assigned cluster? Explain.

a. We can see cluster 5 and 9 the score is at 0.4, which indicate they have a poor separation and cohesion. Cluster 1 and 6 have the score at 0.5, which said they got a moderate separation and cohesion. Cluster 7, 4 and 2 are similar, the score is 0.6 which have a moderate separation and cohesion. CLuster 11 and 8 at score 0.65, is a median in separation and cohesion. CLuster 3 is in a second place which have score at 0.7, relatively good separation and cohesion. Cluster 13 and 14 have the highest score which is 0.9, well-separated and well-matched.

b. In the case, the average silhouette score of 0.269 for n_clusters = 6 suggests that the clustering results are not very compact and well-separated. It is generally recommended to aim for a silhouette score closer to 1, indicating better clustering results. However, the choice of the number of clusters depends on the context and the specific problem you are trying to solve.

1.8.5 8.5. Additional Analysis

[70]:
```
from sklearn.metrics import adjusted_rand_score, homogeneity_score,␣
↪completeness_score
```

[71]:
```
ari = adjusted_rand_score(df_label["Location_State"], df_compare)
ari
```

[71]: 0.45900746908678663

[72]:
```
homogeneity = homogeneity_score(df_label["Location_State"], df_compare)
homogeneity
```

[72]: 0.6642166230977945

[73]:
```
completeness = completeness_score(df_label["Location_State"], df_compare)
completeness
```

[73]: 0.7704005551052056

The adjusted Rand index (ARI) is a measure of the similarity between two clusterings. It ranges from -1 (no agreement) to 1 (perfect agreement). In this case, the ARI value of 0.459 indicates a moderate agreement between the true labels and the predicted labels.

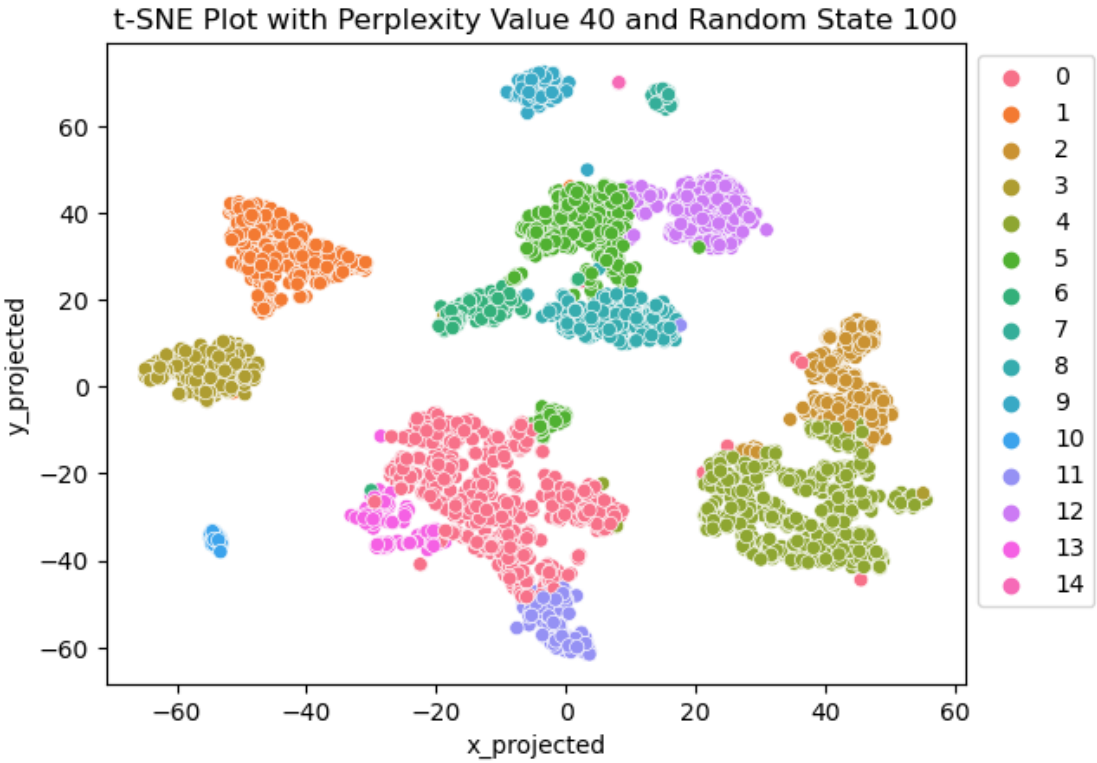
The homogeneity score measures how well each cluster contains only samples that are members of a single class. It ranges from 0 (low homogeneity) to 1 (high homogeneity). The score of 0.664 indicates a moderate level of homogeneity.

The completeness score measures how well all members of a given class are assigned to the same cluster. It ranges from 0 (low completeness) to 1 (high completeness). The score of 0.770 indicates a high level of completeness.

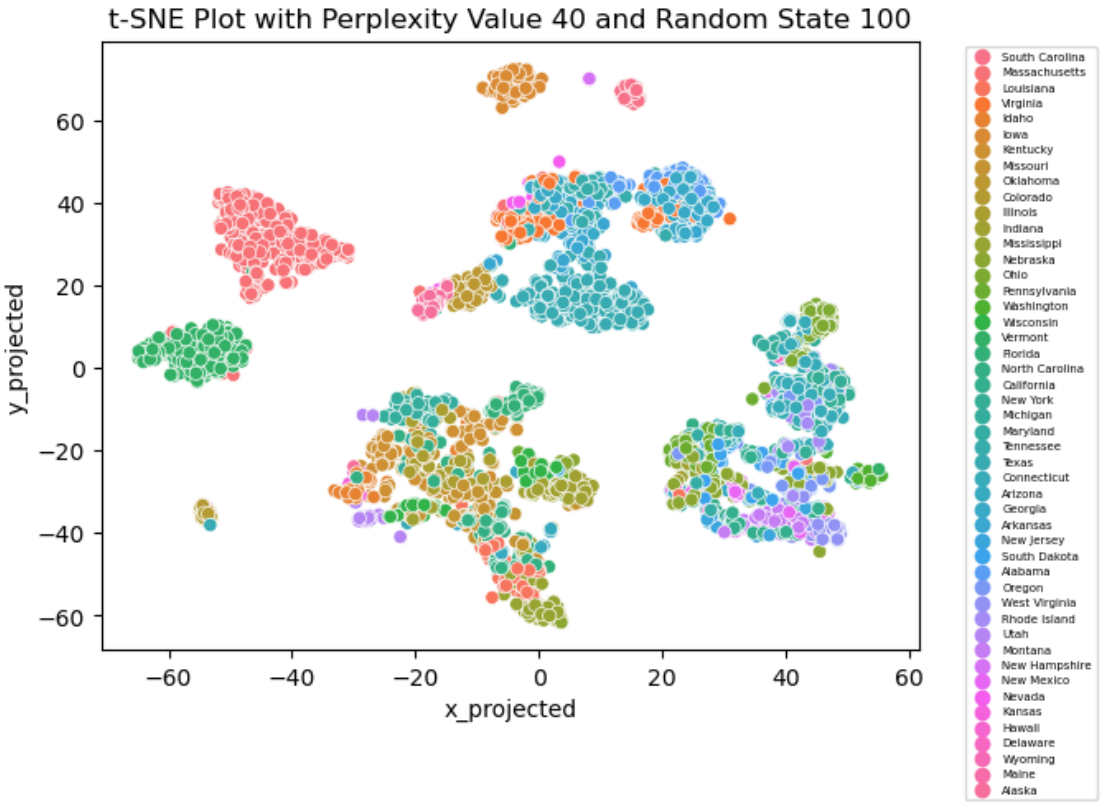
In summary, the ARI and homogeneity scores suggest that the clustering results have moderate agreement with the true labels and moderate homogeneity, while the completeness score suggests that the clustering results have high completeness.

[105]:
```
for k in [15]:
    #Clustering from dendrogram with k clusters
    hac = AgglomerativeClustering(n_clusters=k, affinity='euclidean',␣
↪linkage='ward')
    df_combo['predicted_cluster'] = hac.fit_predict(df_num)

    #Map the resulting cluster labels onto our chosen t-SNE plot
    sns.scatterplot(x='x_projected',y='y_projected', hue='predicted_cluster',␣
↪palette=sns.color_palette("husl", k), data=df_combo)
    plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(40,␣
↪100))
    plt.legend(bbox_to_anchor=(1,1))
    plt.show()
print('         ----------------------Predicted Label-----------------------')
```

```
for perp in [40]:
    for rs in [100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
 ↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', hue =␣
 ↪"Location_State",data=df_combo)
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
 ↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('          ----------------------True Label-----------------------')
```



----------Predicted Label----------



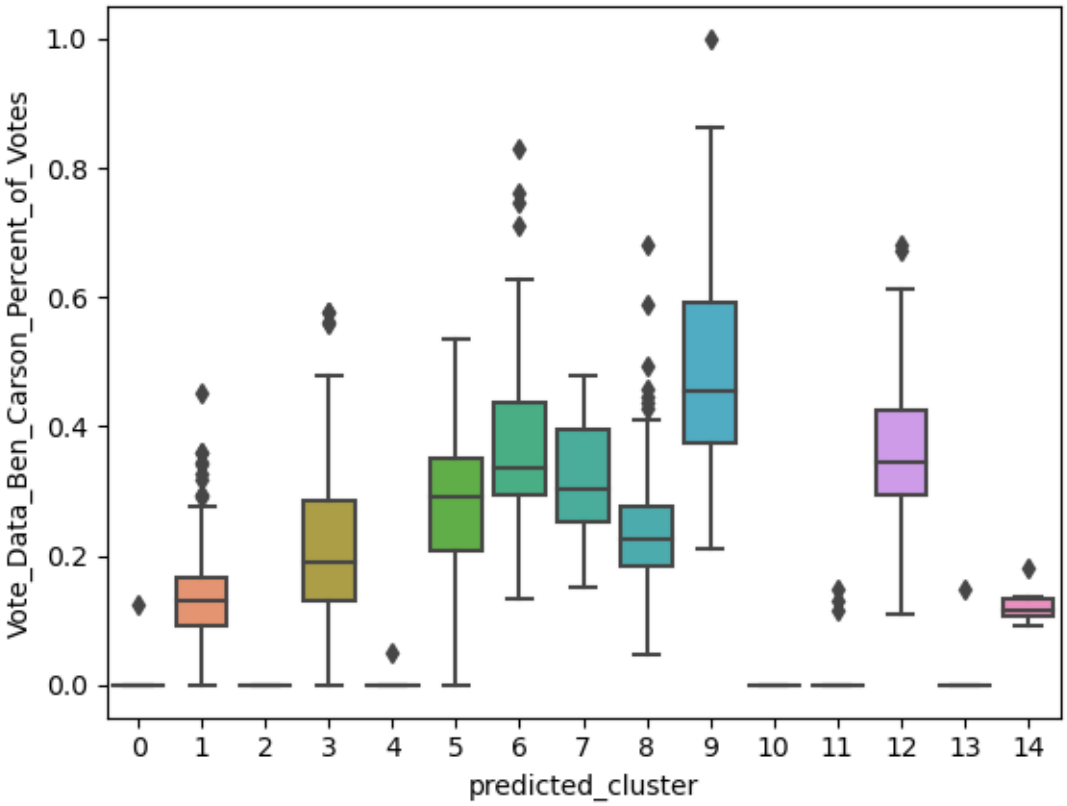
----------True Label----------

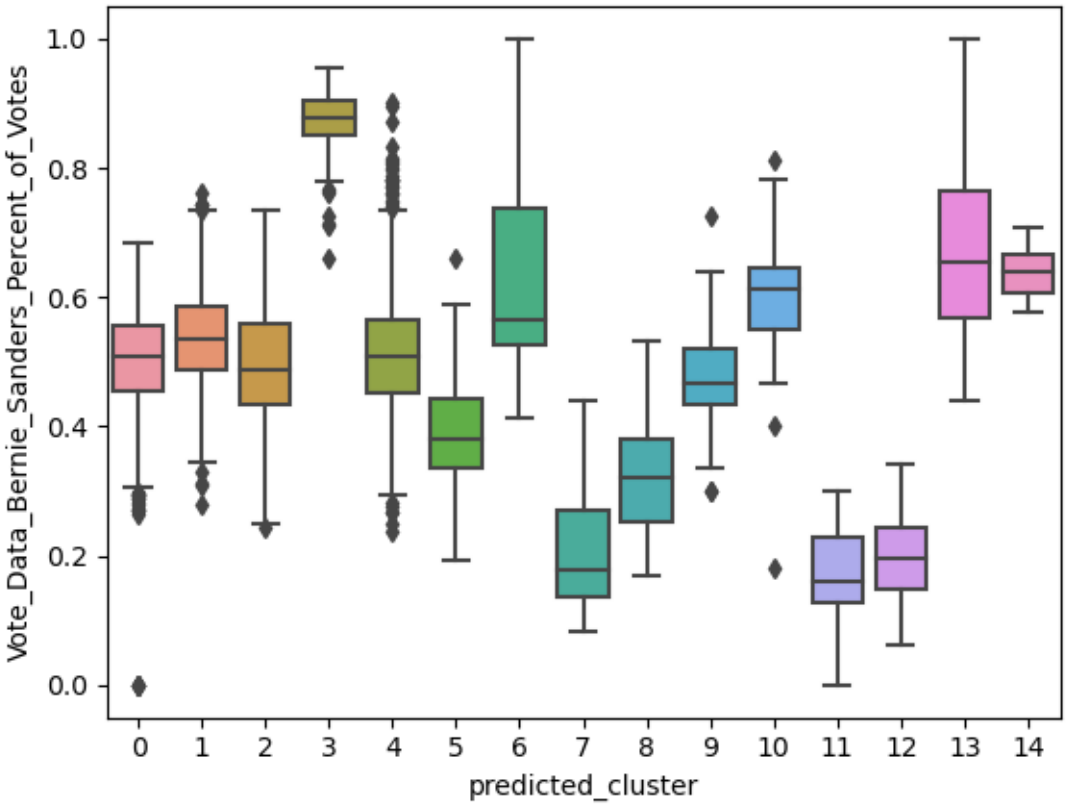
We can see there are somewhat different between the true label. For example, you can see the left up corner, there are three cluster being seperated.

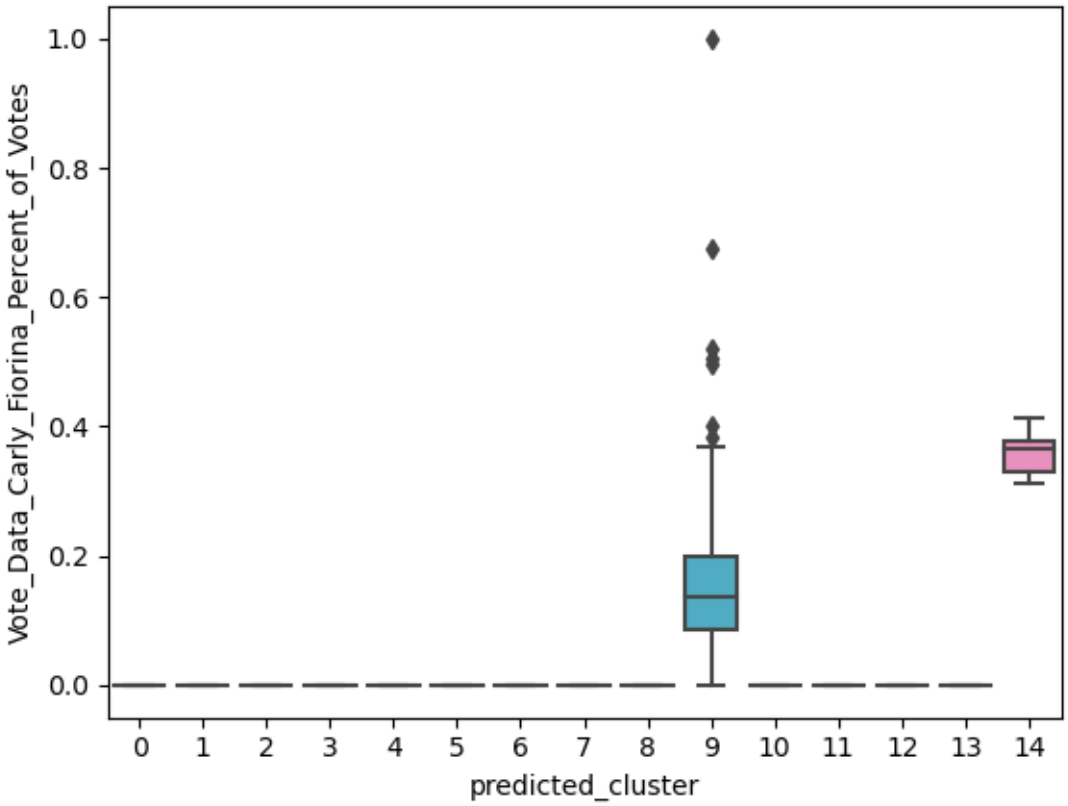
1.8.6 8.6. Describing Each of the Clusters

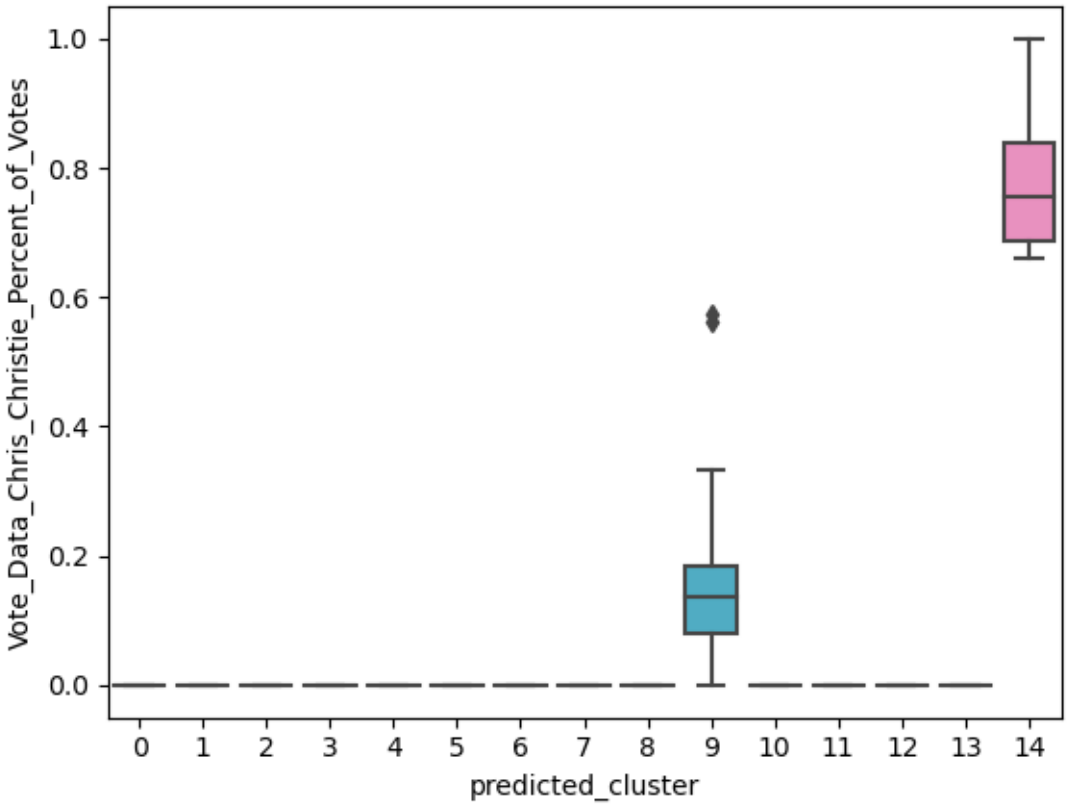
[108]:
```
#Clustering from dendrogram with k clusters
hac = AgglomerativeClustering(n_clusters=15, affinity='euclidean',␣
 ↪linkage='ward')
df_combo['predicted_cluster'] = hac.fit_predict(df_num)
```

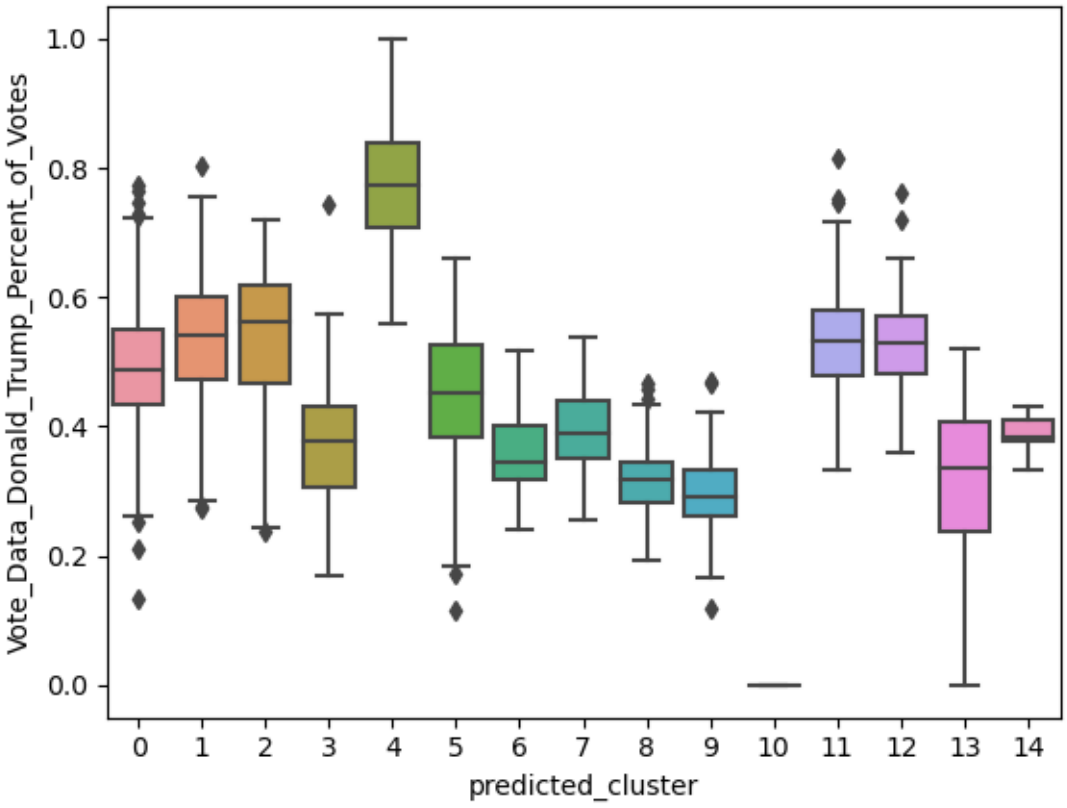
[109]:
```
for col in df_label.columns[1:18]:
    sns.boxplot(x="predicted_cluster", y=col, data=df_combo)
    plt.show()
```

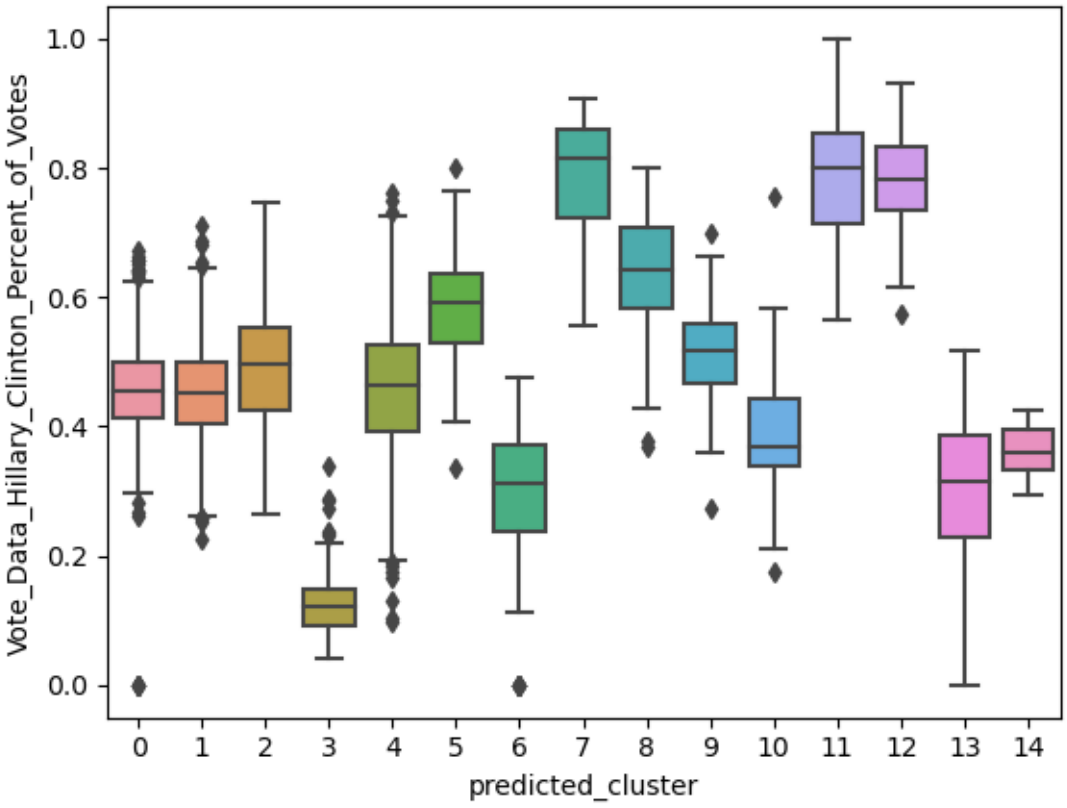

Vote_Data_Ben_Carson_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

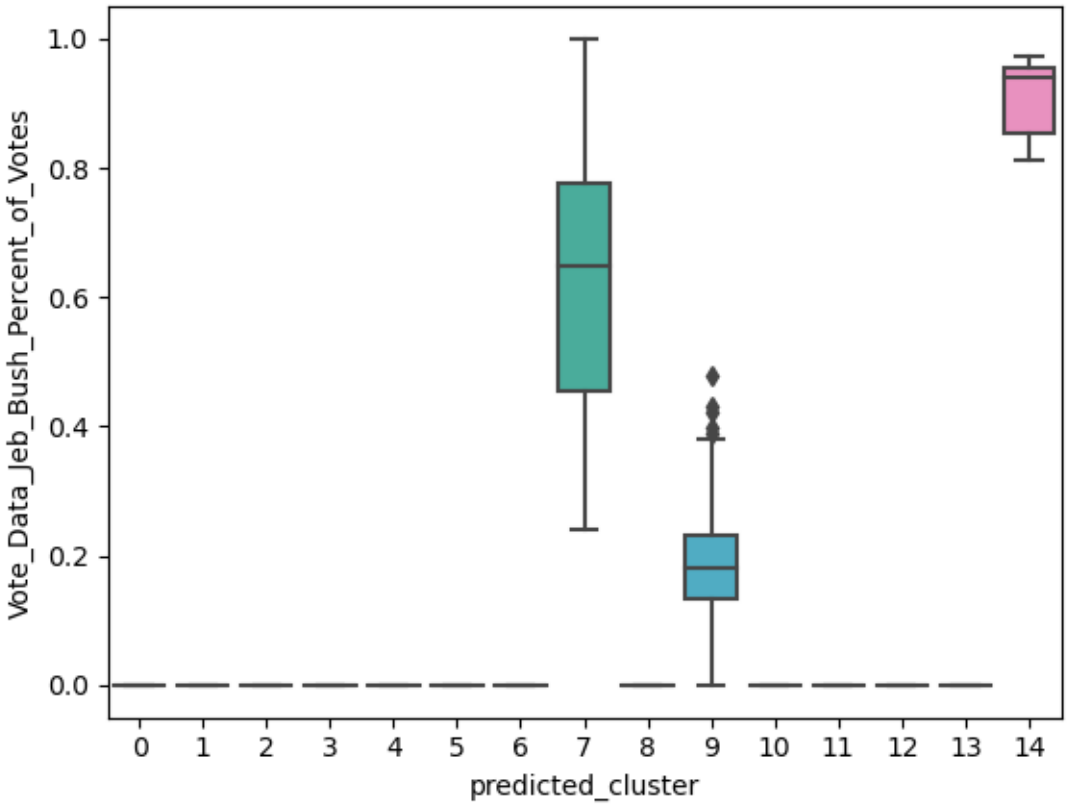

Vote_Data_Bernie_Sanders_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

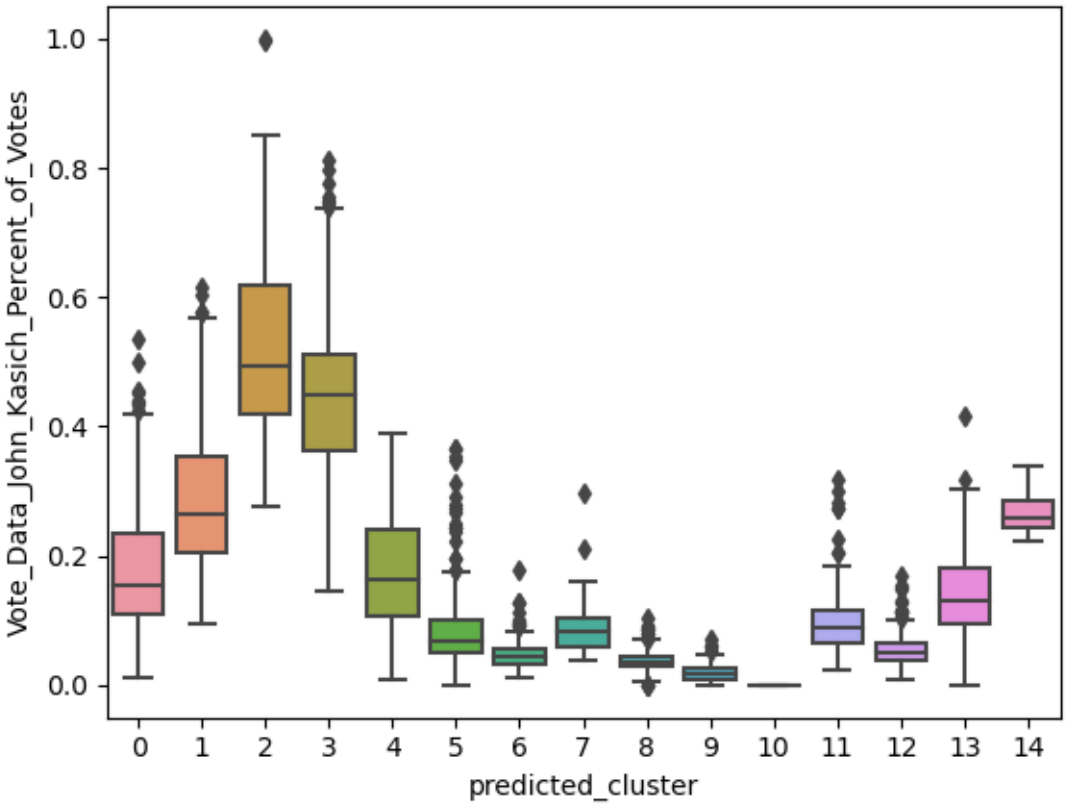

Vote_Data_Carly_Fiorina_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

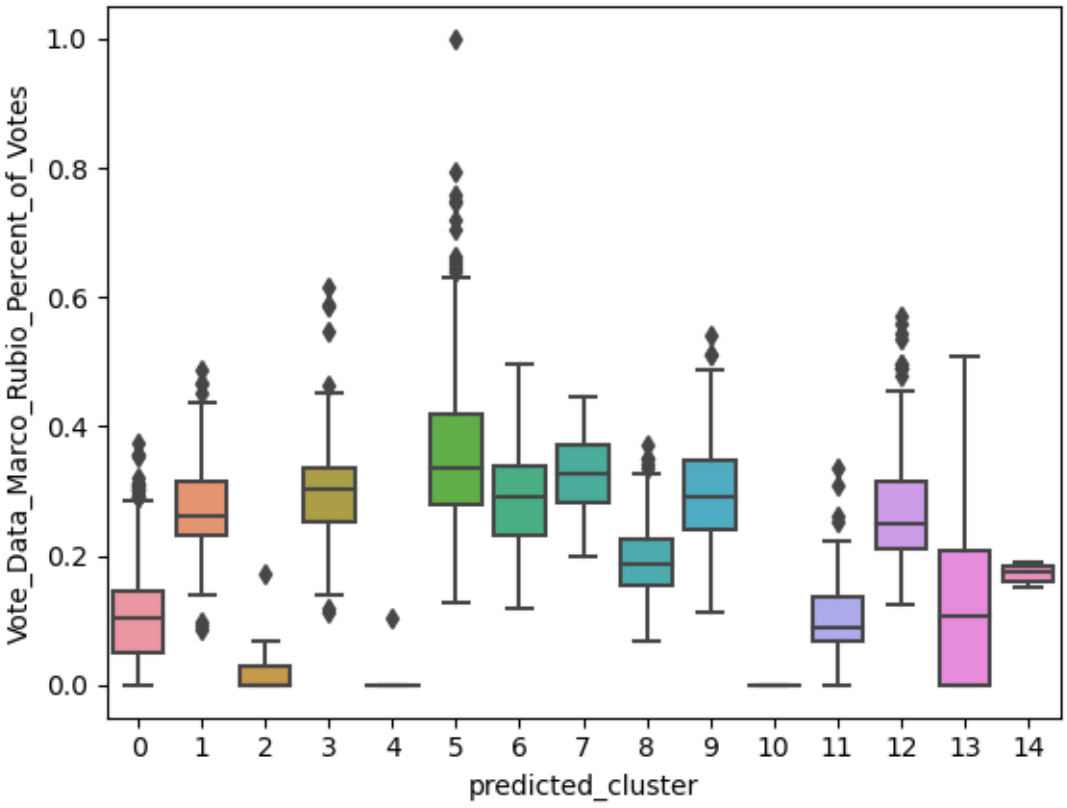

Vote_Data_Chris_Christie_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

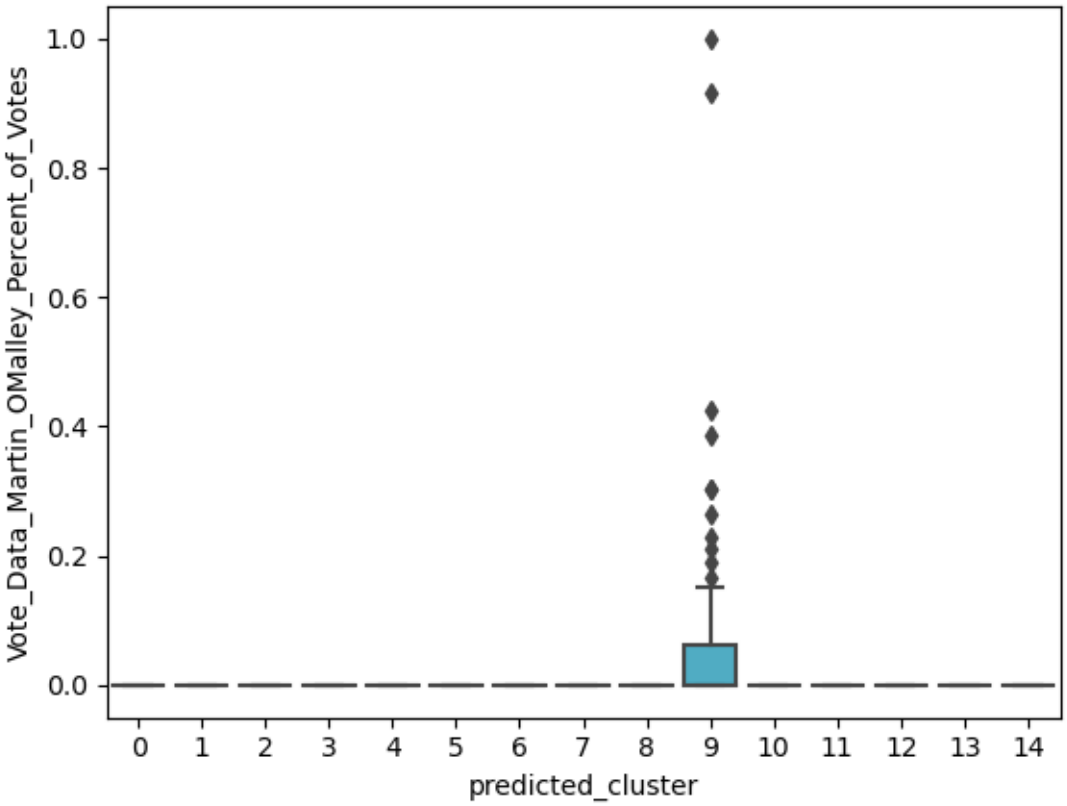

Vote_Data_Donald_Trump_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

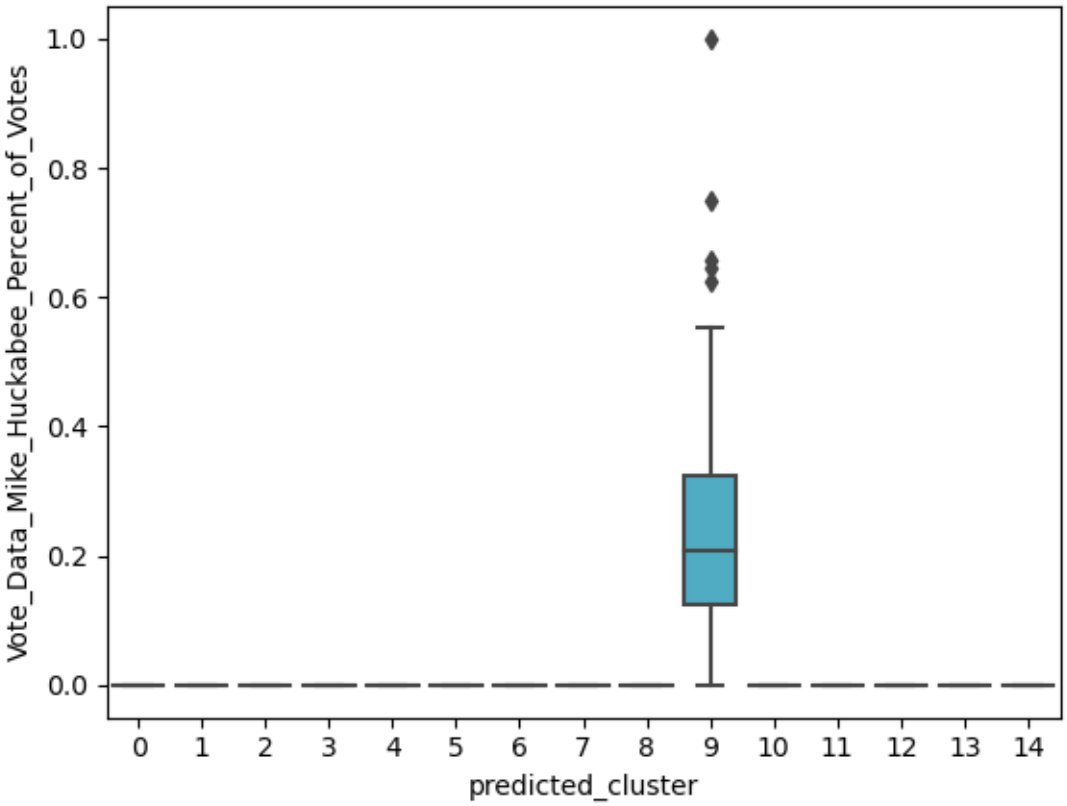

Vote_Data_Hillary_Clinton_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

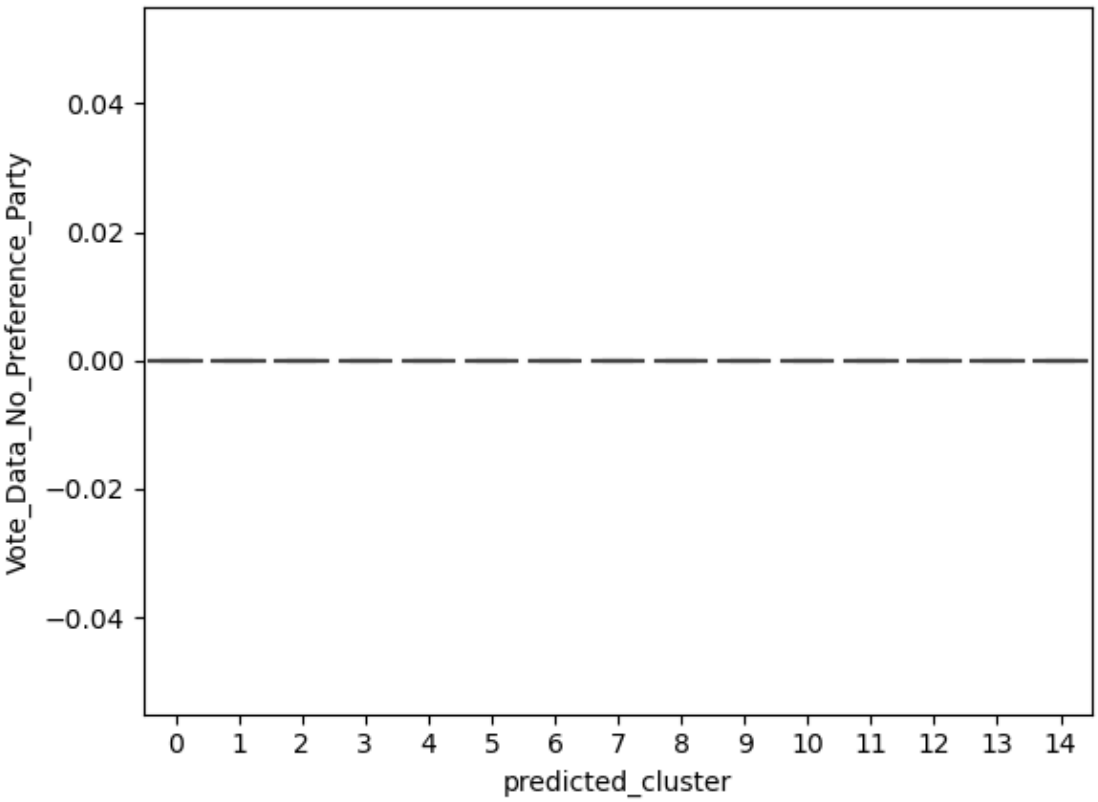

Vote_Data_Jeb_Bush_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

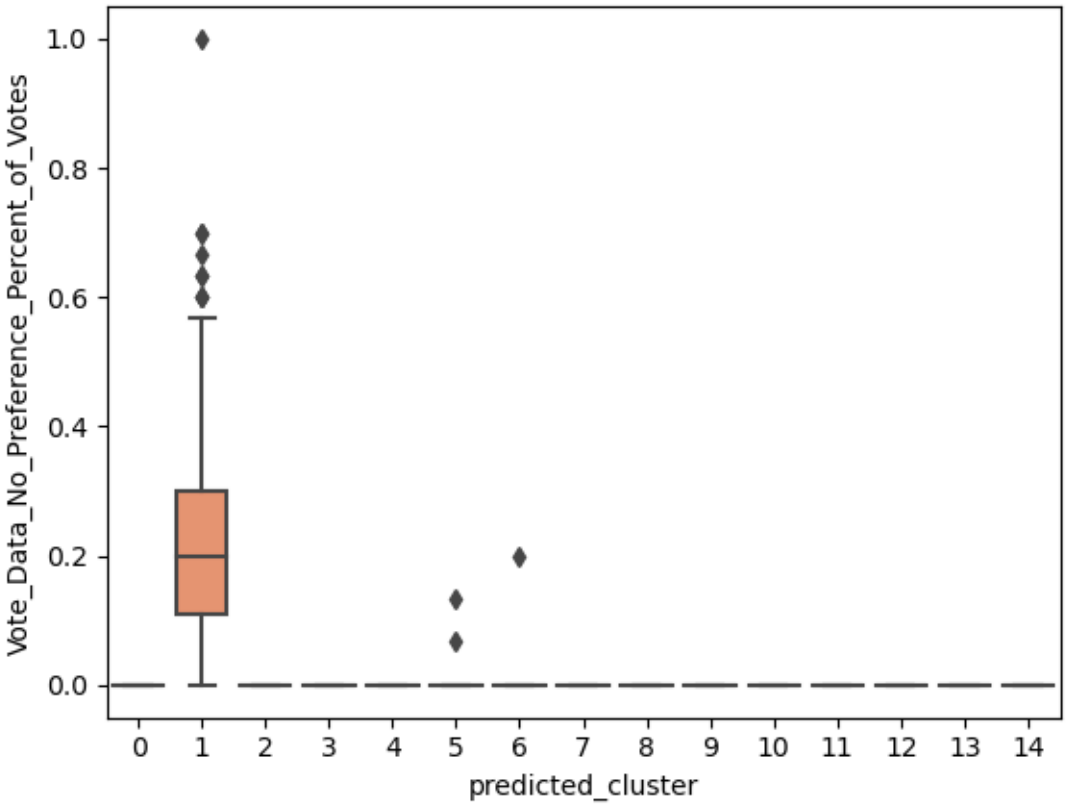

Vote_Data_John_Kasich_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

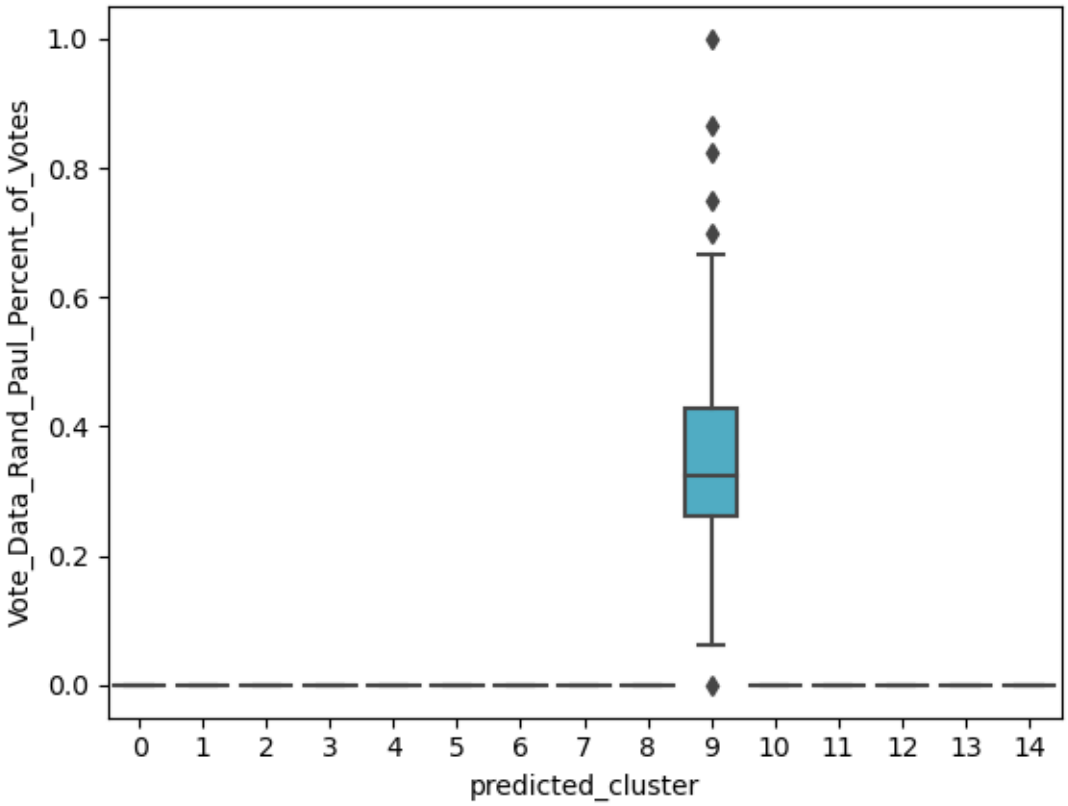

Vote_Data_Marco_Rubio_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

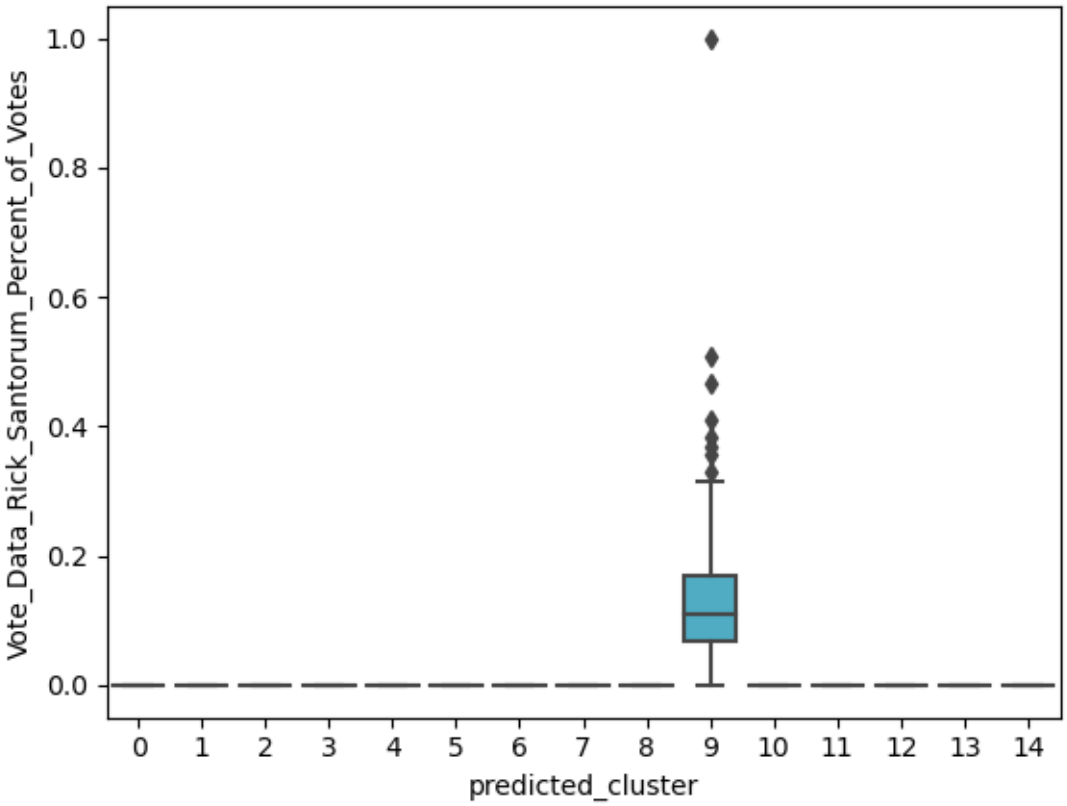

Vote_Data_Martin_OMalley_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

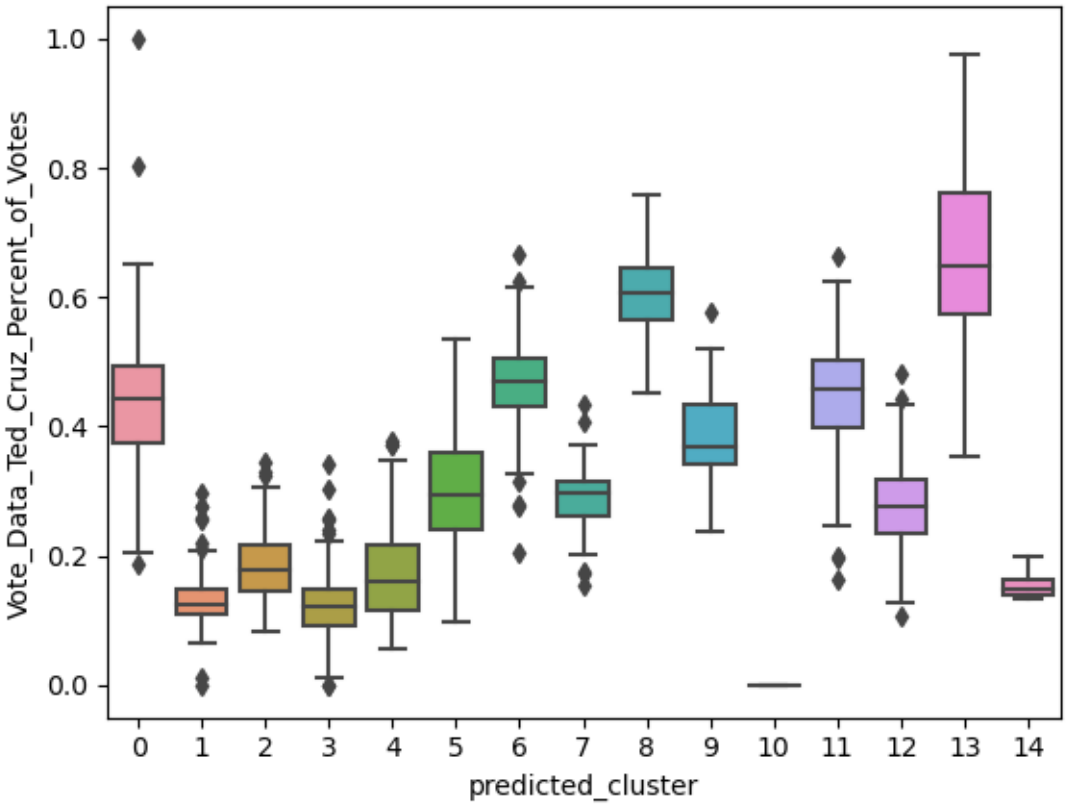

Vote_Data_Mike_Huckabee_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

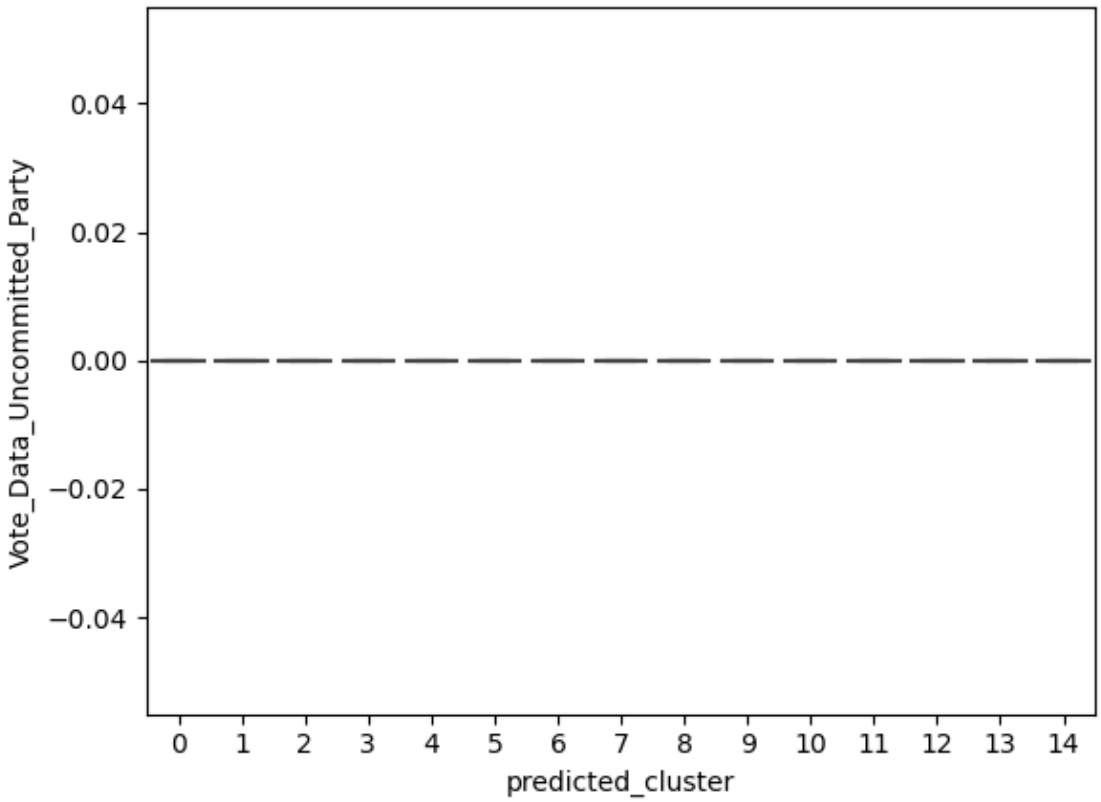

0.04
0.02
0.00
−0.02
−0.04
Vote_Data_No_Preference_Party
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster


Vote_Data_No_Preference_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster


Vote_Data_Rand_Paul_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster


Vote_Data_Rick_Santorum_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster

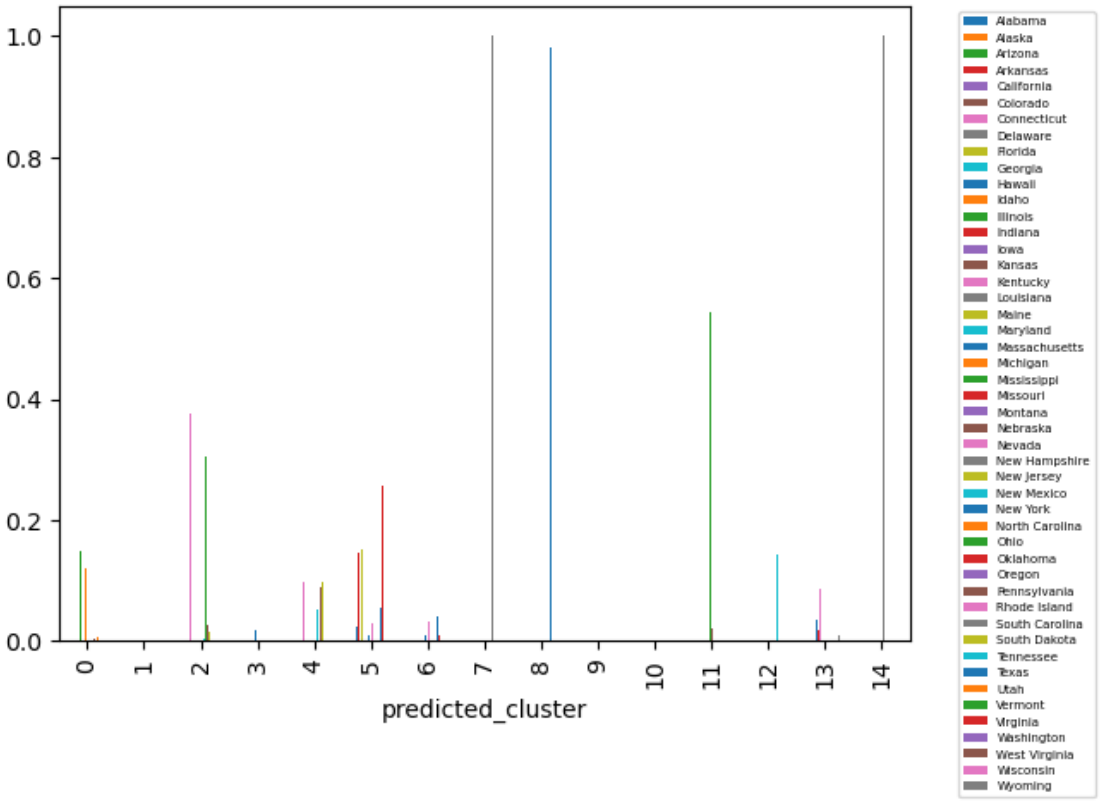

Vote_Data_Ted_Cruz_Percent_of_Votes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
predicted_cluster



1. The vote for Bernie Sanders are mainly distributed in cluster 3 and 13 which apportch to 0.9. But cluster 11 and 12 have the lowest which is 0.1.
2. For Carly Fiorina, it is mainly distriubated in Cluster 14, have 0.35.
3. For Chris Christie iy is mainly in cluster 14 have 0.75
4. For Donald Trump, mainly distributed in cluster 0, 9 and 17 which apportch to 0.8. But cluster 2 are very low.
5. HillaryCliton mostly distributed in cluster 11, 12 and 15. But less distributed in 3 and 13.
6. Jeb Bush had a lot of support on cluster 14 and 15, but cluster 1 very low support.
7. John's approval rating is only high among 18, and the others are very low.
8. Marco Rubio has a very high rating in cluster 10 but the rest are vert low.
9. Mike Huckabee and Martin OMalley are only distributed in cluster 9 and none of the others.

Unaffiliated and nonpartisan data are not centrally distributed and data are scarce.

[90]:
```
ctab=pd.crosstab(df_combo['predicted_cluster'],␣
 ↪df_combo['Location_State'],normalize='index')
ctab.plot.bar()
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
plt.show()
```



We can see from the bar plot that clusters 7, 8, 11, 12 and 14 are composed of one state. But in clusters 2 4 and 5, there are multiple states. So the bar plot of HAC also shows that the cohesion is a little bit unsatisfactory. The cluster structure goes to K=15, plus our additional analysis said the ARI and homogeneity scores suggest that the clustering results have moderate agreement with the true labels and moderate homogeneity, while the completeness score suggests that the clustering results have high completeness which fit with the bar plot.

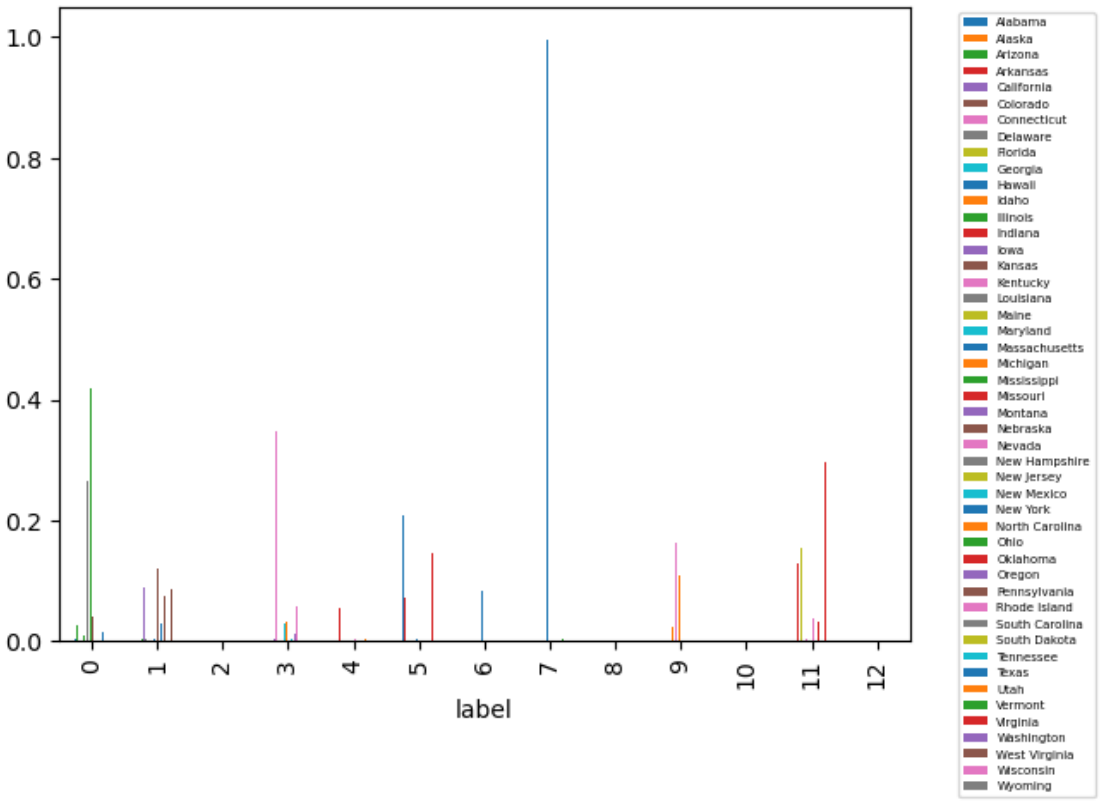
1.9 9. Analysis Summary and Conclusion

1.9.1 9.1. Algorithm Comparison Summary

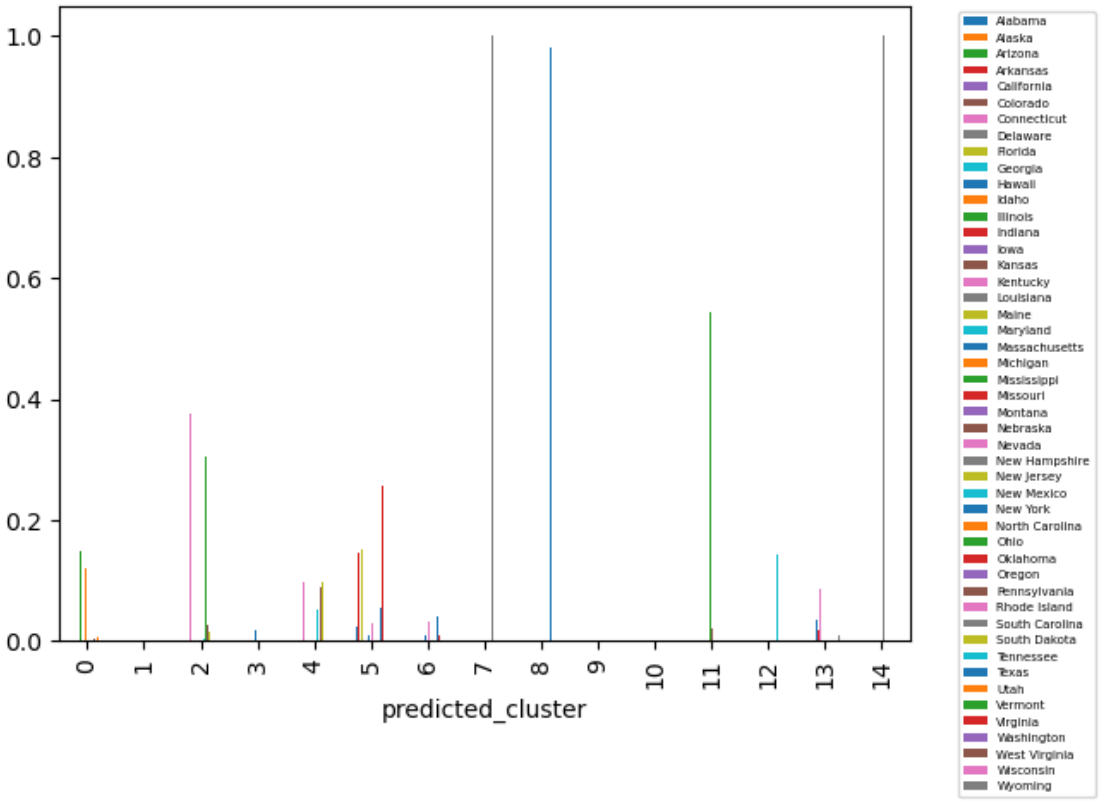
9.1.1. Comparing Algorithm Performance recall the bar plot in 7.6 and 8.6 first.

[114]:
```
kmeans = KMeans(n_clusters=13, random_state=100).fit(X)
df_combo["label"] = kmeans.labels_
ctab=pd.crosstab(df_combo['label'],␣
 ↪df_combo['Location_State'],normalize='index')
ctab.plot.bar()
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
plt.show()
print('          ----------------------Bar for Kmeans-----------------------')
```

```
ctab=pd.crosstab(df_combo['predicted_cluster'],␣
 ↪df_combo['Location_State'],normalize='index')
ctab.plot.bar()
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
plt.show()
print('          ----------------------Bar for HAC-----------------------')
```




```
----------------------Bar for Kmeans-----------------------
```



----------Bar for HAC----------

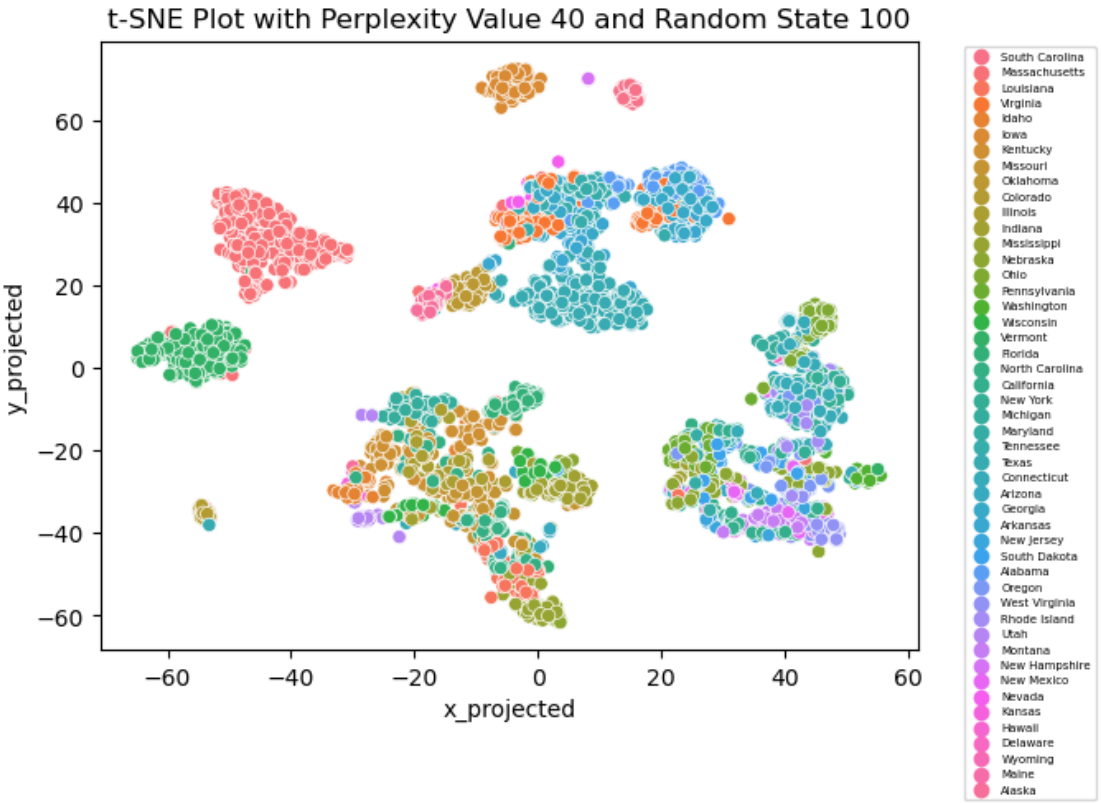
Given your research goals and motivation stated at the beginning of your analysis, compare and contrast the performance of your two clustering algorithms. recall the Motivation: We want to use the voting data to determine if different counties are clustered according to the state they are in. We also want to explore the distribution of candidates supported in each cluster. The specific method is: we take the information of the real state where the county is located as the pre-defined label, and then check whether the predicted clusters are related to our pre-defined Location_State by different clustering algorithms. Then we analyze each cluster in detail to explore the components with high candidate support.

Looking first at the k-means, we can see that cluster 7 is composed of one state. But we can see that some clusters are not conceived with motivation, for example, clusters 5 and 11 have different states in their distribution. This shows that also the separation and cohesion is a little bit unsatisfactory, especially the cohesion.

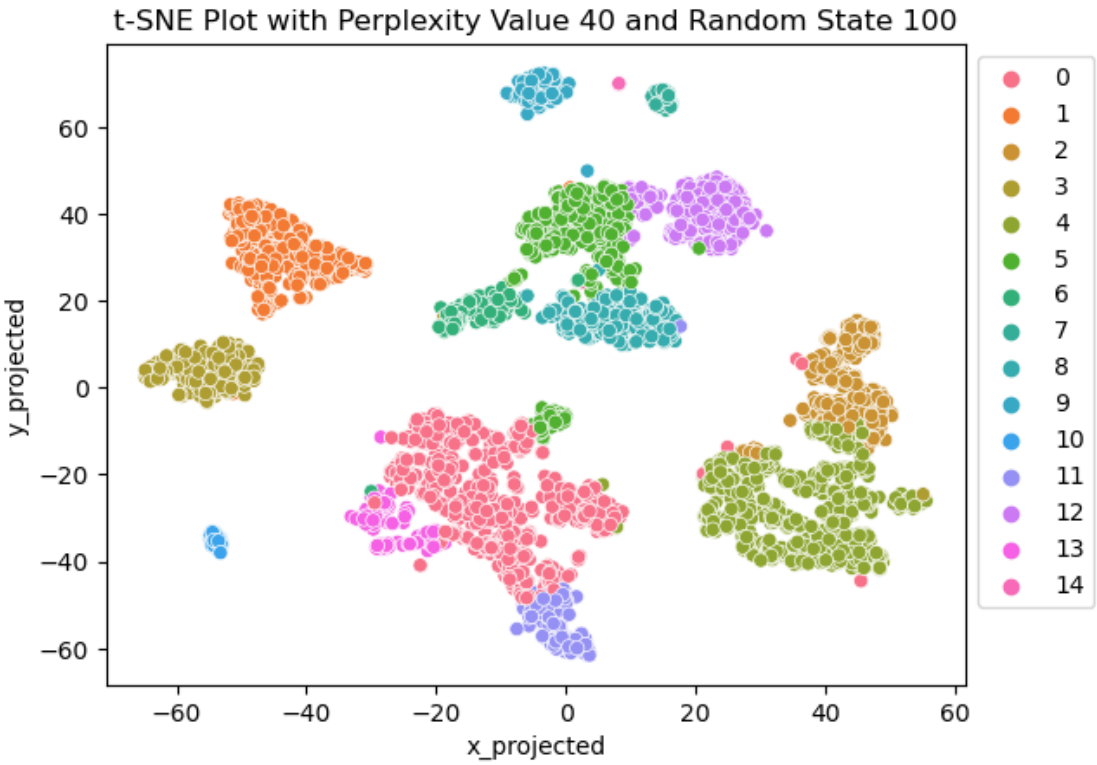
Next is the HAC, where we can see that clusters 7, 8, 11, 12 and 14 are composed of one state. But in clusters 2 and 5, there are multiple states. So the bar plot of HAC also shows that the cohesion is a little bit unsatisfactory.

9.1.2. Comparing Algorithm Results

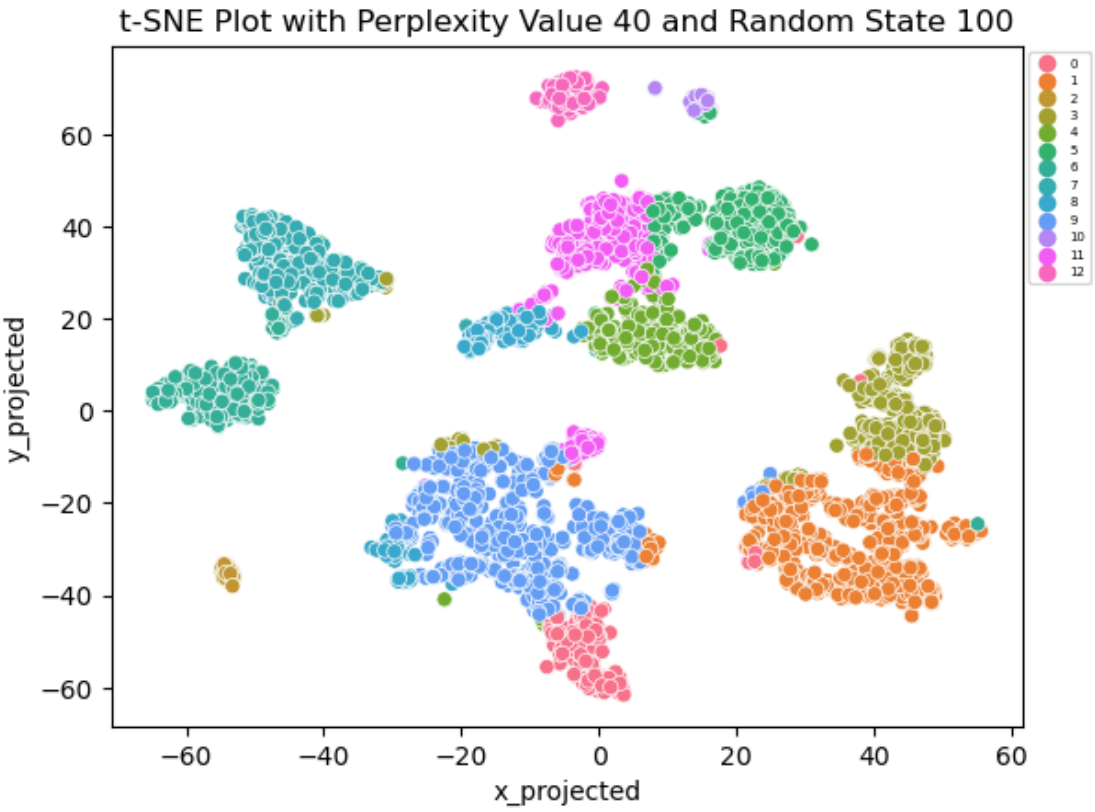
[111]:
```
for perp in [40]:
    for rs in [100]:
        tsne = TSNE(n_components=2, perplexity=perp, random_state=rs)
        data_tsne = tsne.fit_transform(X)
        df_tsne = pd.DataFrame(data_tsne, columns=['x_projected',␣
 ↪'y_projected'])
        df_combo = pd.concat([df_label, df_tsne], axis=1)
        sns.scatterplot(x='x_projected',y='y_projected', hue =␣
 ↪"Location_State",data=df_combo)
        plt.title('t-SNE Plot with Perplexity Value %s and Random State %s'␣
 ↪%(perp, rs))
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', fontsize = 5)
        plt.show()
    print('          ----------------------True Label-----------------------')

for k in [15]:
    #Clustering from dendrogram with k clusters
    hac = AgglomerativeClustering(n_clusters=k, affinity='euclidean',␣
 ↪linkage='ward')
    df_combo['predicted_cluster'] = hac.fit_predict(df_num)

    #Map the resulting cluster labels onto our chosen t-SNE plot
    sns.scatterplot(x='x_projected',y='y_projected', hue='predicted_cluster',␣
 ↪palette=sns.color_palette("husl", k), data=df_combo)
    plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(40,␣
 ↪100))
    plt.legend(bbox_to_anchor=(1,1))
    plt.show()
print('         ----------------------Predicted Label␣
 ↪HAC-----------------------')



#Color code the points in your t-sne plot by cluster labels and code the␣
 ↪"style" of the marker with your class labels.
kmeans = KMeans(n_clusters=13, random_state=100).fit(X)
df_combo["label"] = kmeans.labels_
sns.scatterplot(x='x_projected',y='y_projected',hue = "label",palette=sns.
 ↪color_palette('husl', 13), data=df_combo)
plt.title('t-SNE Plot with Perplexity Value %s and Random State %s' %(perp, rs))
plt.legend(bbox_to_anchor=(1, 1), loc='upper left', fontsize = 5)
plt.show()
print('         ----------------------Predicted Label␣
 ↪K-means-----------------------')
```


t-SNE Plot with Perplexity Value 40 and Random State 100
y_projected
x_projected
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
South Carolina
Massachusetts
Louisiana
Virginia
Idaho
Iowa
Kentucky
Missouri
Oklahoma
Colorado
Illinois
Indiana
Mississippi
Nebraska
Ohio
Pennsylvania
Washington
Wisconsin
Vermont
Florida
North Carolina
California
New York
Michigan
Maryland
Tennessee
Texas
Connecticut
Arizona
Georgia
Arkansas
New Jersey
South Dakota
Alabama
Oregon
West Virginia
Rhode Island
Utah
Montana
New Hampshire
New Mexico
Nevada
Kansas
Hawaii
Delaware
Wyoming
Maine
Alaska

----------True Label----------



----------Predicted Label HAC----------



----------Predicted Label K-means----------

Average silhouette_score is 0.35 from k-means, 0.33 from hac ward linkage. These values are quiet similar, however, the K-means are better.

And we compare the t-SNE plot, from the t-SNE plot, HAC and K-means are similar too with k=13 and k=15. Also their distribution is also similar by the t-SNE plot. In addition, those two plot have strong relation with the location_state. Both of this two algos have proved that our goal in the motivation which is the vote preference in each county may be similar according to their state.

All in all, these two algorithms are pretty similar. However, the K-means has higher average silhouette score. A higher average silhouette score indicates better clustering performance, where each data point is more similar to other data points in the same cluster than to data points in other clusters. Also, K-means have more balanced clusters. But the HAC shows more clusters, which are more close to the real number of states. The difference between them are not huge, they are both good algo in this situation as far as I am concerned.

1.9.2 9.2. Conclusion and Insights Summary

In conclusion, we have used clustering algorithms to explore the distribution of vote preference in each county and determine if they are clustered according to the state they in. The K-means and HAC clustering algorithms were used, and their performance was compared based on the average

silhouette score and t-SNE plot. The results showed that both algorithms have strong relations with the location_state and are good in this situation. However, K-means had a higher average silhouette score, indicating better clustering performance, and more balanced clusters with a greater distribution shape. On the other hand, HAC showed more clusters, which were more close to the real number of states. In general, the results suggest that the vote preference in each county may be similar according to their state, but with some variations that need to be explored further.

1.10 10. Group Contribution Report

Project Name: Election Analysis

Team Members:

[Shuoyuan Gao]

[Shiyuan Zhang]

Contributions:

[Shuoyuan Gao] - [Part1, Part2, Part3, Part4, Part8, Part9] - 60%

[Shiyuan Zhang] - [Part5, Part6, Part7, part10] - 40%

This notebook was converted with convert.ploomber.io