Final Project Report

Team 22

Shuoyuan Gao: Descriptive analysis of categorical variables, Cluster analysis.
Yufei Duan: Descriptive analysis on numeric variables, statistical analysis on the relation between a numeric variable and response variable. Linear model fitting and variable selection.
Zhongwen Shen: Fitting multiple regression models to identify key predictors of student performance and selecting the optimal model based on statistical criteria.

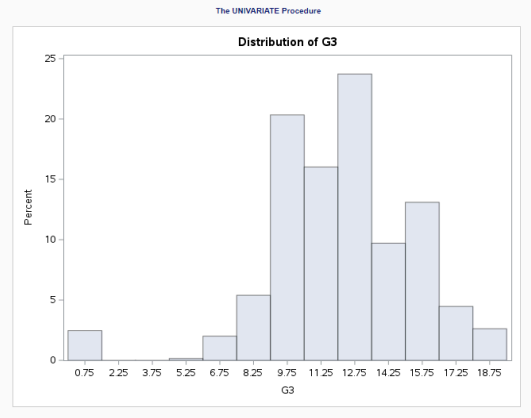
Introduction

In this paper, we delve into the analysis of the student-por data set which focuses on language courses. This data originates from two schools in Portugal and provides insights into student performance in Portuguese. The dataset encompasses aspects, including student achievements, demographic details, social factors, and school-related attributes. Cortez and Silvas 2008 study utilized these datasets for tasks like classification, quintile classification, and regression analyses. For detailed information on this dataset and the analytical approach used refer to Cortezs study titled "Student Performance" from 2014 available in the UCI Machine Learning Library at https;//doi.org/10.24432/C5TG7T.

This report aims to delve deeper into the dataset by exploring how different variables impact students' academic performance in G3. We hypothesize that several factors play a role in predicting students' final grades (G3) which are parental educational backgrounds (F_edu and M_edu) study time (studytime) aspirations for higher education (higher) failures in exams per course (failures) and school support (schoolsup). Our initial assessment involved testing G3 for normality using the Shapiro-Wilk test. The results indicated deviations from a distribution supporting our hypothesis and suggesting the need for nonlinear modeling techniques. We use the Linear Regression Model, the Over-dispersed Poisson log-linear model, and the Gamma log-linear model to support our idea.

Data and methodology

First, we use a histogram to visualize the G3 - final grade. We find many values of 0 in the histogram. They need to be deleted before analysis.



Next, we perform data exploration and descriptive analysis. The variables Medu (mother’s education) and Fedu (father’s education) represent the educational level of the students’ parents. The mean for mothers is

2.51 and for fathers is 2.31, which indicates that the average educational level of parents ranges from 5th grade to secondary education. Because in the variables Medu and Fedu are 0-4, numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 -secondary education or 4 - higher education) The educational level of the parents is Key socioeconomic factors affecting student academic performance. In the G3 mean, we see that 11 is in the middle of the pack.

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum
age	16.7442219	1.2181376	15.0000000	22.0000000
Medu	2.5146379	1.1345520	0	4.0000000
Fedu	2.3066256	1.0999309	0	4.0000000
traveltime	1.5685670	0.7486601	1.0000000	4.0000000
studytime	1.9306626	0.8295096	1.0000000	4.0000000
failures	0.2218798	0.5932351	0	3.0000000
famrel	3.9306626	0.9557169	1.0000000	5.0000000
freetime	3.1802773	1.0510926	1.0000000	5.0000000
goout	3.1848998	1.1757661	1.0000000	5.0000000
Dalc	1.5023112	0.9248344	1.0000000	5.0000000
Walc	2.2804314	1.2843800	1.0000000	5.0000000
health	3.5362096	1.4462591	1.0000000	5.0000000
absences	3.6594761	4.6407588	0	32.0000000
G1	11.3990755	2.7452651	0	19.0000000
G2	11.5701079	2.9136387	0	19.0000000
G3	11.9060092	3.2306562	0	19.0000000

The average travel time was 1.57 (range 1 to 4). Longer travel times may be related to living in less accessible areas, which may be related to socioeconomic status. Students from lower socioeconomic backgrounds may have longer commutes due to the location of affordable housing relative to school sites.

Grades for academic grades (G1, G2, G3) showed a medium range and variability, suggesting that students' final grades were not very high. It must be affected by many factors, which is also the focus of our next analysis.

After completing the numeric exploration, we continued with categorical variables.

Categorical Variables

From a demographic perspective, there are more female students than male students, which may affect class dynamics and may indicate differences in educational engagement between genders. There is even more of an impact on the romantic variable because students are in love at this stage. You will have some longing for love or start trying to fall in love. As a result, performance may be affected. However, in the love variable, fewer students are in romantic relationships, which may make them focus more on academics and personal development. This is a very good phenomenon.

The FREQ Procedure

romantic	Frequency
no	410
yes	239

sex	Frequency
F	383
M	266

address	Frequency
R	197
U	452

Mjob	Frequency
at_home	135
health	48
other	258
services	136
teacher	72

Fjob	Frequency
at_home	42
health	23
other	367
services	181
teacher	36

famsize	Frequency
GT3	457
LE3	192

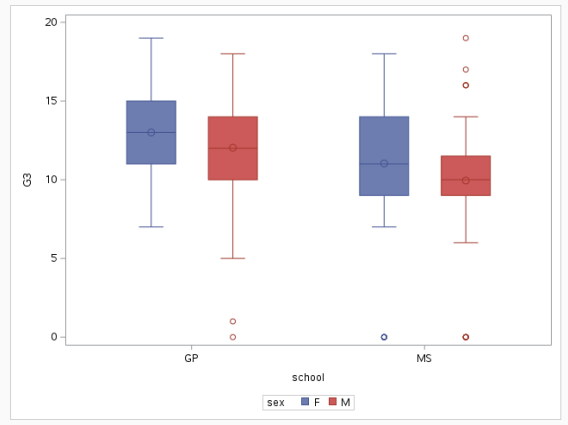
higher	Frequency
no	69
yes	580

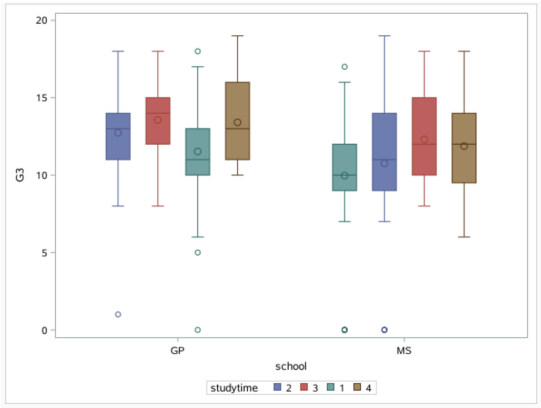
The dominance of larger family sizes at the household level may affect the economic and educational resources available to each child. With more children, parental supervision is also an issue worth considering. Most students live in urban areas, and commuting will not affect the quality of students' classes. Because thinking about it from another perspective, if students need to spend a lot of time on the road every day, then they will be sleepy in class. The diversity of parents' employment is noteworthy, with many parents working in sectors labeled "other" or "services". Mothers often worked at home or in service jobs, while fathers also worked primarily in the “other” and “service” categories. This also proves that most parents have no experience in teaching their children. In other words, as a teacher, you will have more experience teaching your children.

Judging from the school's attendance the majority of students attend Gabriel Pereira (GP), suggesting that students may be concentrated in this area or prefer this school to Mousinho da Silveira (MS). The large number of students who receive no school support at the school level highlights potential areas for educational intervention. Instead, most students have family support that facilitates their academic success. The vast majority of students aspire to pursue higher education, indicating a high educational ambition among the student body. Most students attend kindergarten, which may contribute to better preparation for formal education. In other words, most children are starting from a similar starting point and no one is left behind educationally.

The three histograms showed that the overall school grade remained consistent across the study periods. The three charts also show that overall student achievement ranges from 9.75-14.25. This indicates that the school's overall academic level is in the upper-middle range. Gender may be associated with different performance within schools, which may be due to different educational experiences or social

expectations.The Gabriel Pereira High School and the Mousinho da Silveira High School may have an impact on student achievement depending on the type of school. It also reflects the quality of academic teaching or the school's academic environment.

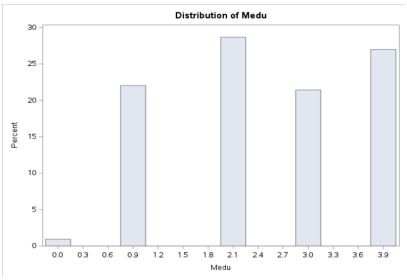


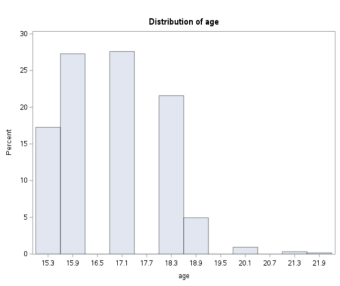


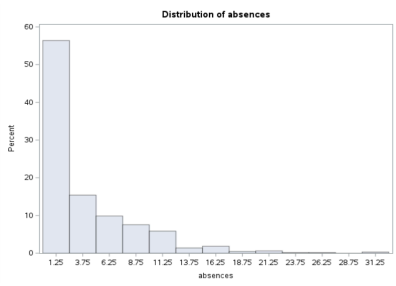
The three histograms showed that the overall school grade remained consistent across the study periods. The three charts also show that overall student achievement ranges from 9.75-14.25. This indicates that the school's overall academic level is in the upper-middle range. Gender may be associated with different performance within schools, which may be due to different educational experiences or social expectations.The Gabriel Pereira High School and the Mousinho da Silveira High School may have an impact on student achievement depending on the type of school. It also reflects the quality of academic teaching or the school's academic environment.

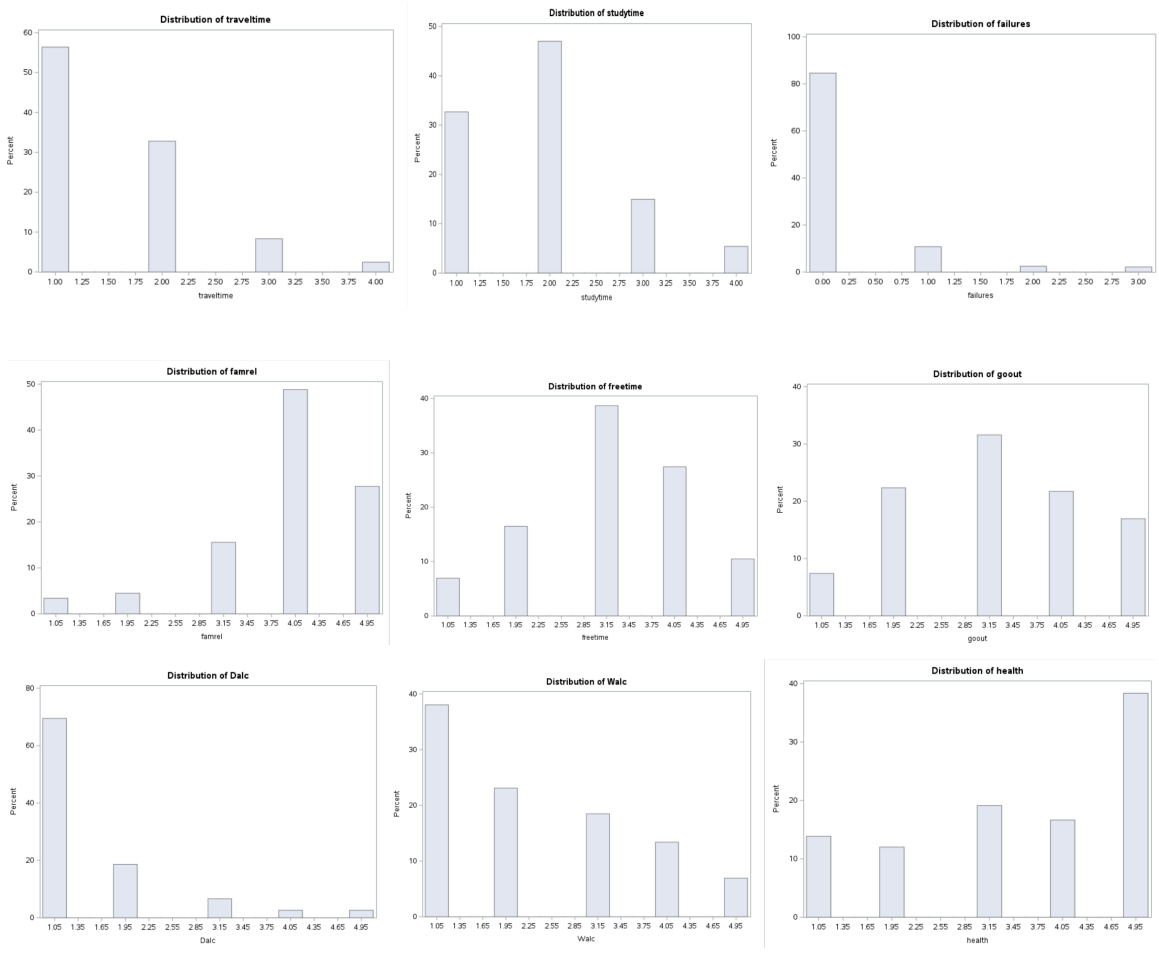
Overall, study time is an important factor. The amount of time invested in studying will be directly proportional to the overall performance in the Portuguese language course. However, in the PLOT, we found outliers. This suggests that the time invested in studying does not always lead to higher grades, which may indicate the influence of the quality of learning or the learning environment as well as other factors.

Numeric Variables









1. Age: The average age of students in the dataset is approximately 16.74 years, with a relatively narrow spread indicated by a standard deviation of 1.22 years.
2. Mother's Education (Medu) and Father's Education (Fedu): On average, both mothers and fathers have completed around 2 to 2.5 levels of education, with similar median values. The standard deviations suggest some variability in parental education levels.
3. Travel Time: The average travel time from home to school is approximately 1.57 units, with a standard deviation of 0.75, indicating relatively consistent travel times among students.
4. Weekly Study Time: Students spend an average of about 1.93 units studying per week, with a standard deviation of 0.83, suggesting some variation in study habits among students.
5. Past Class Failures: On average, students have experienced very few past class failures, with an average of only about 0.22 failures, and the majority of students having no past failures.Family Relationships (Famrel): Overall, students report relatively high levels of

satisfaction with family relationships, with an average rating of about 3.93 on a scale of 1 to 5.

6. Free Time, Going Out, Alcohol Consumption (Dalc and Walc), and Health Status: Students generally report moderate levels of free time, going out, weekday and weekend alcohol consumption, and health status. The standard deviations indicate some variability in these lifestyle factors among students.
7. Absences: On average, students have around 3.66 absences from school, with a relatively wide spread indicated by a standard deviation of 4.64, suggesting some students have higher rates of absenteeism.

Relationship with G3

The CORR Procedure

1 With Variables:	G3
13 Variables:	age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc health absences

Pearson Correlation Coefficients, N = 649
Prob > |r| under H0: Rho=0

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences
G3	-0.10651 0.0066	0.24015 <.0001	0.21180 <.0001	-0.12717 0.0012	0.24979 <.0001	-0.39332 <.0001	0.06336 0.1068	-0.12270 0.0017	-0.08764 0.0256	-0.20472 <.0001	-0.17662 <.0001	-0.09885 0.0117	-0.09138 0.0199

According to the correlation table, there are several variables showing high correlation with potential high influence on G3. Father's Education (Fedu): Moderate positive correlation (0.24015). Higher education of fathers correlates with higher final grades.

1. Study Time: Moderate positive correlation (0.21180). More time spent studying correlates with higher final grades.
2. Past Failures: Moderate negative correlation (-0.39332). More past failures correlate with lower final grades.
3. Going Out: Moderate negative correlation (-0.24979). Increased frequency of going out correlates with lower final grades.
4. Health Status: Weak negative correlation (-0.20472). Poorer health correlates with lower final grades.
5. Absences: Weak negative correlation (-0.17662). Higher absences correlate with lower final grades.

Parameter Estimates

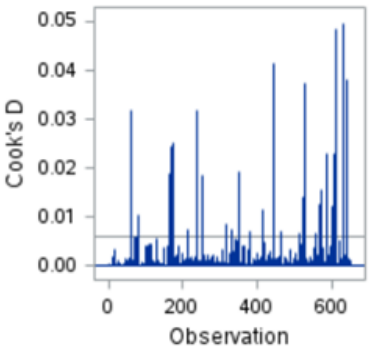
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	9.57478	1.80489	5.30	<.0001	0
age	1	0.11171	0.09816	1.14	0.2555	1.16832
Medu	1	0.31168	0.13104	2.38	0.0177	1.80631
Fedu	1	0.24403	0.13348	1.83	0.0680	1.76164
traveltime	1	-0.13587	0.15489	-0.88	0.3807	1.09896
studytime	1	0.60334	0.13917	4.34	<.0001	1.08911
failures	1	-1.74238	0.20413	-8.54	<.0001	1.19641
famrel	1	0.15146	0.11957	1.27	0.2057	1.06716
freetime	1	-0.16568	0.11468	-1.44	0.1490	1.18731
goout	1	-0.04617	0.11014	-0.42	0.6752	1.37040
Dalc	1	-0.41935	0.15428	-2.72	0.0067	1.66379
Walc	1	-0.04367	0.11868	-0.37	0.7130	1.89882
health	1	-0.17006	0.07846	-2.17	0.0306	1.05229
absences	1	-0.01132	0.02479	-0.46	0.6481	1.08142

A Variance Inflation Factor (VIF) test was conducted to assess multicollinearity among the predictors. The VIFs ranged from 1 to 2 for all variables, indicating no

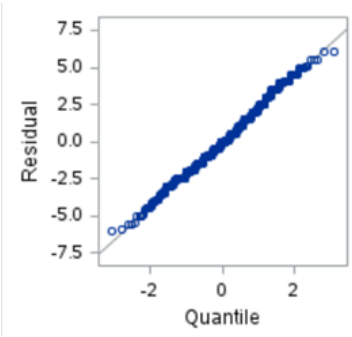
significant multicollinearity and thus retaining the validity of the regression model coefficients.

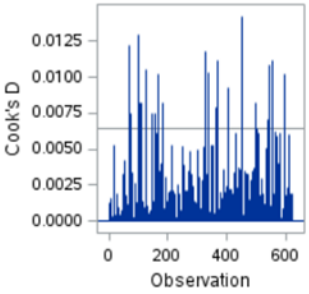
The initial regression model demonstrated an R-Squared value of 0.2441 and an adjusted R-Squared of 0.2382, indicating that approximately 24% of the variability in the final grades was explained by the model. All variables were statistically significant with p-values less than 0.05.

Upon further examination using Cook's distance, several influential data points were identified. To improve the model's accuracy, observations with a Cook's distance greater than four times the mean were removed.

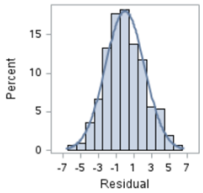


A refined model was fitted after the exclusion of these points. The new model's R-Squared improved to 0.2883, indicating a better fit than the initial model. This final model included failures, Medu, studytime, and Dalc as predictors.





The refined model's goodness of fit was quantitatively stronger, and residual analysis confirmed the assumption of normally distributed errors — a crucial consideration for the validity of a linear regression model. Both the histogram of residuals and the Q-Q plot indicated that the residuals followed a normal distribution.



The coefficients of the final model were as follows:

Association between numeric variables and G3:

Medu (Mother's Education): The positive coefficient (0.52842) suggests that an increase in the mother's education level is associated with an increase in the student's final grade. Each additional level of mother's education is predicted to increase the final grade by about 0.53 points. Studytime: The positive coefficient (0.54437) indicates that more weekly study time is associated with higher final grades. Each additional level of time spent studying per week is predicted to increase the final grade by roughly 0.54 points. Failures: The negative coefficient (-1.72054) implies a strong inverse relationship between the number of past class failures and the final grade. Each additional failure is associated with a decrease of

approximately 1.72 points in the final grade. Dalc (Workday Alcohol Consumption): The negative coefficient (-0.49221) indicates that increased alcohol consumption on workdays is correlated with lower final grades. Each additional level on the alcohol consumption scale corresponds to a decrease of about 0.49 points in the final grade.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.82202	0.35306	4620.79343	939.57	<.0001
Medu	0.52842	0.07996	214.80138	43.68	<.0001
studytime	0.54437	0.11046	119.43732	24.29	<.0001
failures	-1.72054	0.17157	494.59720	100.57	<.0001
Dalc	-0.49221	0.10523	107.60526	21.88	<.0001

Analysis of Student Performance

1. Linear Regression Model (Stepwise Selection)

The results from the linear regression stepwise selection in SAS reveal key factors influencing student performance (G3):

Age: Older students tend to perform slightly better.
Mother's Education (Medu): Higher education levels of the mother are positively associated with student grades.
Study Time: More study time correlates with better academic performance.
Failures: Previous failures have a strong negative impact on grades.
Health: Poor health negatively affects grades.
Absences: More absences lead to lower grades.
School Support: Interestingly, school support is linked with lower grades, possibly indicating that it's provided to students who are already struggling.
Aspirations for Higher Education: Strong aspirations correlate with higher grades, likely reflecting higher motivation.
Romantic Relationships: Being in a relationship is slightly negatively associated with grades.
Father's Job as a Teacher: Having a father who is a teacher is positively linked with student performance, possibly due to an academically supportive environment at home.

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.667068	1.461102	3.19
age	1	0.319657	0.079938	4.00
Medu	1	0.302721	0.090360	3.35
studytime	1	0.360866	0.110670	3.26
failures	1	-1.276008	0.164960	-7.74
famrel	1	0.168739	0.094697	1.78
goout	1	-0.242595	0.078836	-3.08
Dalc	1	-0.186453	0.106184	-1.76
health	1	-0.148839	0.061456	-2.42
absences	1	-0.067962	0.019592	-3.47
female	1	0.514404	0.197405	2.61
urban	1	0.306697	0.197748	1.55
schoolsup_yes	1	-1.084921	0.293365	-3.70
paid_yes	1	-0.583247	0.372288	-1.57
activities_yes	1	0.418016	0.178526	2.34
higher_yes	1	1.734335	0.315498	5.50
internet_yes	1	0.319411	0.221858	1.44
romantic_yes	1	-0.360695	0.186358	-1.94
Mjob_at_home	1	-0.330950	0.236672	-1.40
Fjob_teacher	1	1.152620	0.401588	2.87
Fjob_services	1	-0.353529	0.198868	-1.78

2. Over-dispersed Poisson Log-linear Model

Deviance and Pearson Chi-Square values close to their degrees of freedom (600) suggest a decent model fit. These determine the statistical significance of each predictor. Significant variables ($p < 0.05$), such as 'failures' and 'higher education aspirations' (higher_yes), strongly influence the final grades.

Type 1 Analysis: Looks at each variable's effect sequentially as they are entered into the model. This reveals how much each additional variable adds to the model.

Type 3 Analysis: Assesses the effect of each variable while controlling for all other variables, which is crucial for understanding individual predictors' contributions when other variables are accounted for. The selected predictors are:

Age: Highly significant ($P < 0.0002$), older students perform better.

Studytime: Significant ($P = 0.0011$), more study time improves grades.

Failures: Very significant ($P < 0.0001$), failures strongly decrease performance.

Goout: Significant ($P = 0.0221$), frequent socializing correlates with lower grades.

Absences: Significant ($P = 0.0018$), more absences lead to poorer outcomes.

Female: Significant ($P = 0.0088$), gender influences performance.

School Support: Significant ($P = 0.0007$), indicates support is often given to struggling students.

Higher Education Aspirations: Very significant (P < 0.0001), aspirations boost performance.
Activities: Significant (P = 0.0221), activities participation is linked to better grades.

Model Selection

1. Linear Regression Model (Stepwise Selection)

Root MSE	2.17670
Dependent Mean	12.18770
R-Square	0.3668
Adj R-Sq	0.3461
AIC	1642.90571
AICC	1644.56201
SBC	1100.39874

Root MSE: 2.17670
Dependent Mean: 12.18770
R-Square: 0.3688
Adj R-Sq: 0.3461
AIC: 1642.90571
AICC: 1644.56201
BIC: 1100.39874
This model's moderate R-Squared and Adjusted R-Square values suggest it explains about 34-37% of the variance in the dependent variable. Lower AIC and BIC values compared to other models may indicate a more effective balance of model complexity and fit.

2. Over-dispersed Poisson Log-linear Model

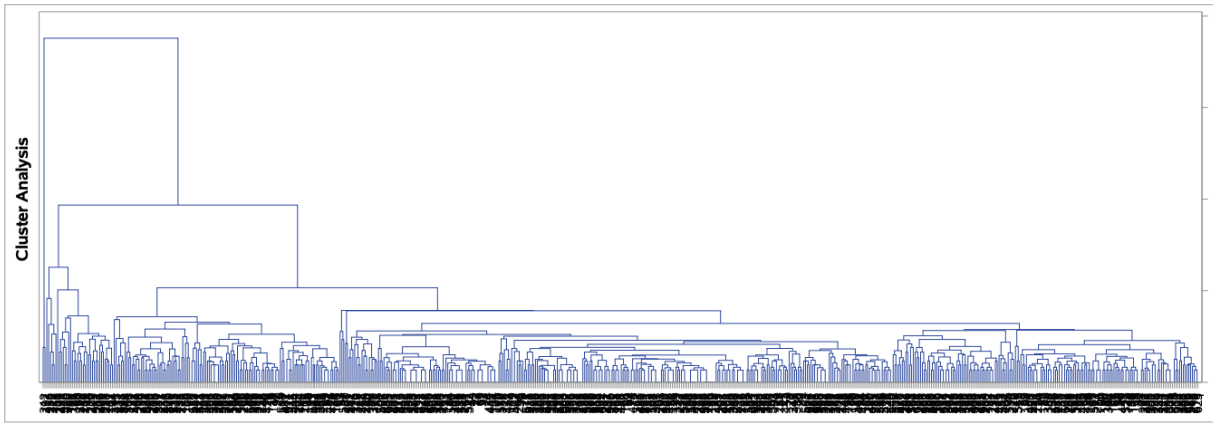
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	600	238.5421	0.3976
Scaled Deviance	600	600.0000	1.0000
Pearson Chi-Square	600	233.3364	0.3889
Scaled Pearson X2	600	586.9064	0.9782
Log Likelihood		29344.8763	
Full Log Likelihood		-1490.6378	
AIC (smaller is better)		3049.2756	
AICC (smaller is better)		3053.2489	
BIC (smaller is better)		3200.6453	

AIC: 3049.2756
AICC: 3053.2489
BIC: 3200.6453
-2 Log L: 2944.8763
This model shows the highest AIC and BIC values, indicating it might be less effective in balancing fit and complexity compared to the logistic model.

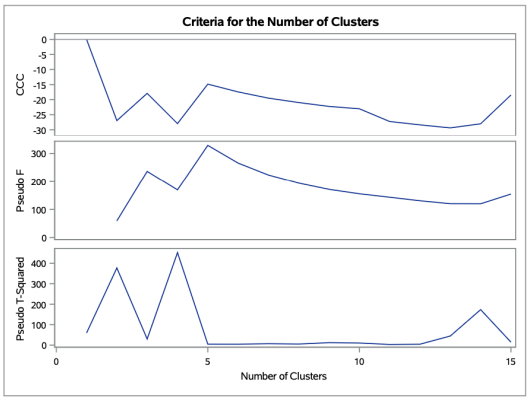
The linear model has significantly lower AIC and BIC values compared to the over-dispersed Poisson model. Lower AIC and BIC values generally suggest a model that better balances fit and complexity, making the linear model potentially more suitable for this dataset based on these criteria. Additionally, the R-Square for the linear model suggests it explains approximately 36.88% of the variance in the dependent variable, which is a useful measure of model efficacy not directly comparable via AIC/BIC in the Poisson model context but helpful for understanding model explanatory power. Based on the provided statistics, the linear model appears to be a better fit for the data than the over-dispersed Poisson log-linear model, as indicated by its lower AIC and BIC values.

Cluster Analysis

We will perform a hierarchical cluster analysis with the aim of exploring potential clusters in the dataset porr. Cluster analysis helps to identify patterns in the data to support subsequent data analysis and decision-making. We then embarked on a cluster analysis and hierarchical clustering with an average linkage approach, again trying to identify links to G3. Based on the stepwise selection of the previous large model, we selected significant variables with a p-value of less than 0.05. These included mother's education level (Medu), father's education level (Fedu), studytime, failures, daily alcohol consumption (Dalc), age, and absences.



The CLUSTER Procedure
Average Linkage Cluster Analysis



The FREQ Procedure

Frequency

Table of CLUSTER by G3

CLUSTER	G3																
	1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total
1	1	0	2	5	19	24	64	73	55	72	50	39	28	24	13	2	471
2	0	0	0	3	10	9	23	26	13	8	9	10	7	4	2	0	124
3	0	1	0	2	5	1	8	5	4	1	3	0	0	1	0	0	31
4	0	0	1	0	1	1	2	0	0	1	0	0	0	0	0	0	6
5	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2
Total	1	1	3	10	35	35	97	104	72	82	63	49	36	29	15	2	634

The results in CCC and pseudostatistics indicate that we can categorize the clusters into 5. But in Dendrogram, if we set the clustering distance to 1, we get 6 clusters. Then, frequency statistics are performed for the grades (G3) in each cluster using. This helps us to understand the specifics of the distribution of grades in different clusters, such as which clusters have students with high or low grades. Cluster 1 has high academic achievers with grades between 10 and 16. Cluster 2 has a moderate distribution of scores between 8 and 12. Cluster 3 has a more dispersed but generally low performance. Clusters 4 and 5 have small numbers of students in these two clusters and are not statistically significant.
Finally, we used two steps: In the first step, we can get a selection of variables relative to the ranked same. This allows us to know which variable has the most significant relationship with G3.
First, the variable failures were added to the model. This variable alone explains about 24.29% of the total variance, indicating that it has a significant effect on predicting G3. The F-value is 13.22 with a p-value of less than 0.0001, indicating that the addition of failures significantly improves the model's explanatory power. significantly, emphasizing the importance of failures.

Next, Dalc was introduced into the model. The addition of this variable increased the explained variance of the model by about 8.57%. The F-value for this variable is 3.86 with a p-value of less than 0.0001, proving that Dalc is also categorically significant. The Wilks Lambda decreases to 0.6922 with a p-value of less than 0.0001, continuing to show the validity of the variable.

In the third step, Medu was added to the model adding a partial R-square of 0.0865. Its F-value is 3.89 with a p-value of less than 0.0001, showing that Medu contributes significantly to the model. The Wilks Lambda for this step is 0.6323 and the p-value is still very low, further confirming the differentiating power of the variables in the model.

Subsequently, age was added to the model. Although its partial R-square contribution of 0.0593 is relatively small, the F-value of 2.58 and P-value of 0.0009 indicate that age is still statistically significant and has an important effect on the model. the Wilks Lambda of 0.5948 and the P-value of the significance test is less than 0.0001, which indicates that it is effective in differentiating between categories.

Finally, study time was added to the model. This variable added a skewed R-square of 0.0477, which is a small contribution but still shows statistical significance with an F-value of 2.05 and a P-value of 0.0107. The Wilks Lambda was further reduced to 0.5664 with a P-value of less than 0.0001, suggesting that studytime is effective in distinguishing between the different G3 categories.

Conclusion

In summary, this report offers an examination of the student dataset related to Portuguese language courses focusing on the factors influencing students' academic success. The dataset, sourced from two schools in Portugal covers aspects such as demographics, social influences, and school-related details providing valuable insights into student performance.

Our analysis reveals that several factors have an impact on a student's final grades (G3) including parental education levels, absences, age, health, romantic relationships, father's job, study time, aspirations for higher education, previous exam failures, and school support. By conducting exploratory data analysis and regression modeling we have identified predictors that affect student performance.

Notably, variables like education levels, study time, and past failures display moderate to strong correlations with final grades. Additionally, factors such as health status, absences, and involvement in relationships also show notable associations with academic achievement.

We utilized regression models like regression and over-dispersed Poisson log-linear models to assess the predictive abilities of these variables. The linear regression model especially showed a level of explanatory power by accounting for approximately 36.88% of the variability in final grades.

Furthermore, through cluster analysis, we were able to identify student groups based on their characteristics and academic accomplishments.

The study showcased the diversity among students. Offers valuable observations on how grades are spread across various groups. In essence, this report enhances our comprehension of the connections between different variables and student achievements, in Portuguese language classes. By pinpointing factors and examining their interactions teachers and policymakers can make educated choices to help students excel academically and thrive emotionally.

Reference

Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.