Student Performance Analysis Report

GROUP 9

Contents

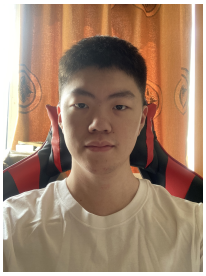
1 Authors

Fugo Takefusa
takefufugo@myvuw.ac.nz
ORCID: 0000-0001-7373-8389

Rhys Lewis-Woodley
lewisrhys@myvuw.ac.nz
ORCID: 0000-0001-5414-0832

Soumya Banerjee
banerjsoum@myvuw.ac.nz
ORCID: 0000-0001-5678-5095

Vivian Dong
dongting@myvuw.ac.nz
ORCID: 0000-0001-8356-8598



Yi Zhang
zhangyi11@myvuw.ac.nz
ORCID: 0000-0003-2959-5290

2 Introduction

The data is about student academic grades for math course of two Portuguese secondary public school during the 2005/2006 school year (P. Cortez 2008). Student grades and their economic, social and demographic variables are included in the data. The data is based on the reports of two public schools (i.e., the grades of students and the number of absence days) and a questionnaire answered by the students. All of the demographic, economic and social variables were collected using this questionnaire. The dataset has 395 records in total with 5 numerical variables and 28 categorical variables. The numerical variables include age, number of school absences, 1st Period grade, 2nd period grade and final period grade. The categorical variables include sex, address(urban/rural), parents highest qualification, family size, commuting time, study time and so on.

We have two main leading questions for our analysis:

- Can Principal Component Analysis - PCA reduce the number of dimensions to model the students' final grade?

- Can we have a discriminant model to classify students whether they pass or fail based on the other features(e.g., first test grade, age, etc)?

Since we have too many categorical variables, for our data analysis, we will only include three categorical variables, which is school, sex and number of past failures.

3 Exploratory Data Analysis - EDA

3.1 Categorical Summary

Three categorical variables:
School - student's school ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
Sex - student's sex ('F' - female or 'M' - male)
Failures - number of past class failures (n if 1<=n<3, else 4)

The table 1 and Figure 1 show the count of each categorical variables. There are more students from Gabriel Pereira school than Mousinho da Silveira school. There are slightly higher number of female students compared to male students. Most of the students never failed the other courses before.

Table 1: Summary of Three Categorical Variables

School	Sex	Failures
GP:349	F:208	0:312
MS: 46	M:187	1: 50
		2: 17
		3: 16

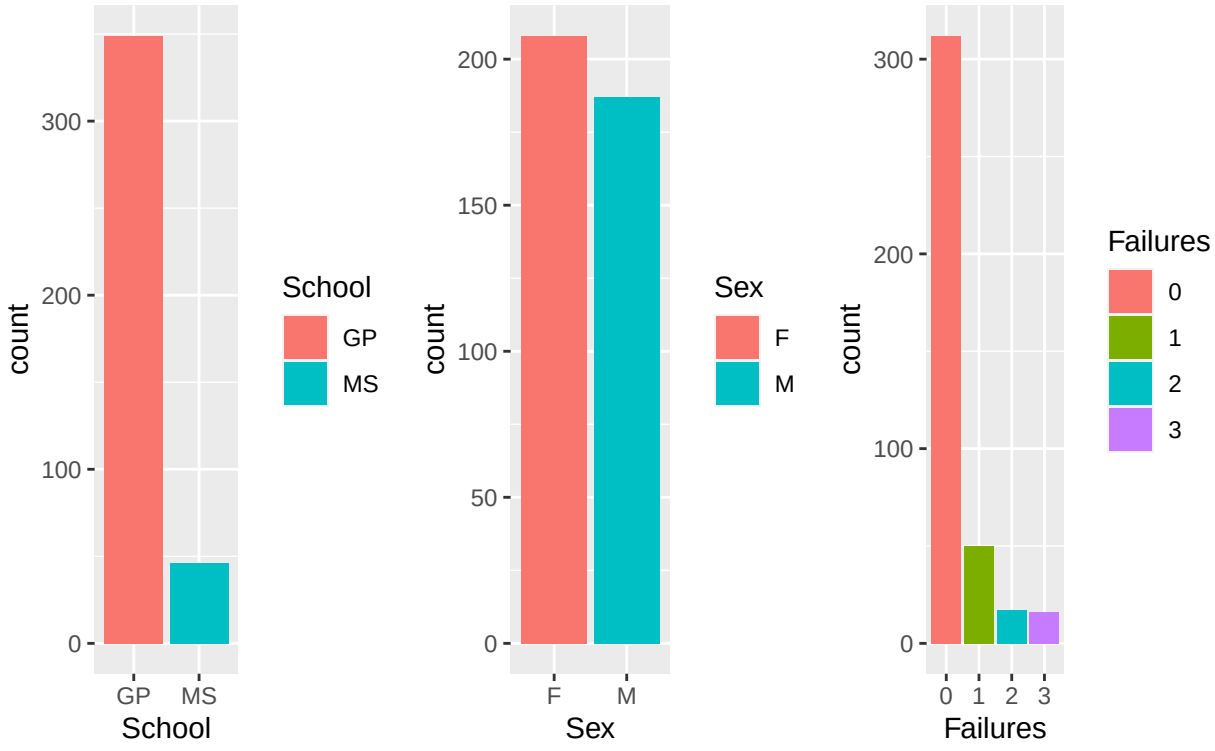


Figure 1: Barplot for Categorical Variables

3.2 Numerical Summary

Five numerical Variables:
Age - student's age (from 15 to 22)
Absences - number of school absences (from 0 to 93)
G1 - first period grade (from 0 to 20)
G2 - second period grade (from 0 to 20)
G3 - final grade (from 0 to 20, output target)

The table 2 shows five numeric variable statistic summary includes: minimum, first quartile, median, third quartile, and maximum.
The table 3 and table 4 show five numeric variable means and covariance.
The covariance between age and absences, G1 and G2 and G3, G3 and absences are positive, this means both variables will tend to move upward or downward in value at the same time. For example, when a student have a higher mark in G1, and this student may have higher mark in G3.
The covariance between age and G1 and G2 and G3, absences and G1 and G2 and G3 are negative, this means variables will move away from each other. For example, a student have more number of absences, this student may get lower mark in G1, G2 and G3.

Table 2: Summary of Five numerical Variables

	age	absences	G1	G2	G3
Min.	15.0	0.0	3.0	0.0	0.0
1st Qu.	16.0	0.0	8.0	9.0	8.0
Median	17.0	4.0	11.0	11.0	11.0
3rd Qu.	18.0	8.0	13.0	13.0	14.0
Max.	22.0	75.0	19.0	19.0	20.0

Table 3: Means of Five numerical Variables

age	absences	G1	G2	G3
16.696	5.709	10.909	10.714	10.415

Table 4: Covariance of Five numerical Variables

	age	absences	G1	G2	G3
age	1.628	1.790	-0.271	-0.689	-0.945
absences	1.790	64.050	-0.824	-0.957	1.256
G1	-0.271	-0.824	11.017	10.639	12.188
G2	-0.689	-0.957	10.639	14.149	15.594
G3	-0.945	1.256	12.188	15.594	20.990

Figure 2 shows the count of five numerical variables. Red dashed line as mean and a fitted density in purple.

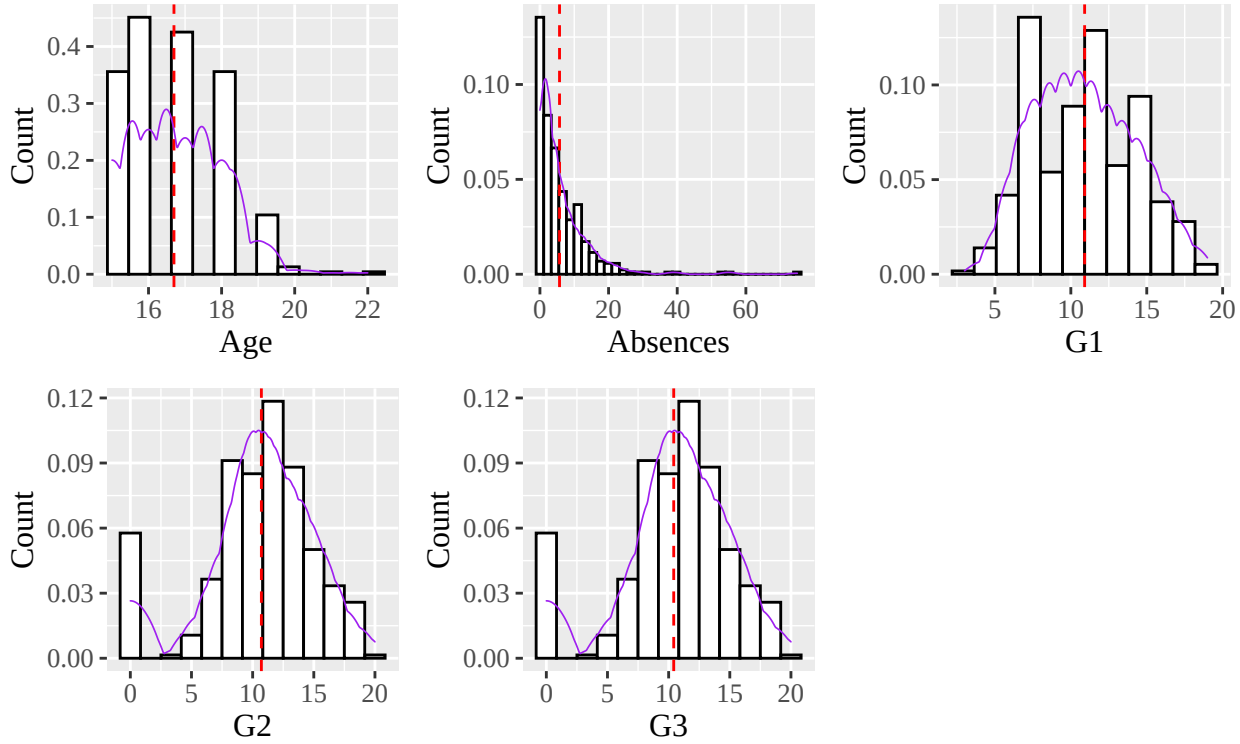


Figure 2: Barplot for Numerical Variables

3.3 Correlation Matrix and plot

Table 5 and Figure 3 shows the correlation between all five numerical variables.
From the table and figure we get that G1, G2 and G3 are have a strong positive correlation.
Between G2 and G3 have the strongest positive correlation, which is 0.905.
Between age and G1, G2, G3, they have a low negative correlation.

Table 5: Correlation Matrix of Five numerical Variables

	age	absences	G1	G2	G3
age	1.000	0.175	-0.064	-0.143	-0.162
absences	0.175	1.000	-0.031	-0.032	0.034
G1	-0.064	-0.031	1.000	0.852	0.801
G2	-0.143	-0.032	0.852	1.000	0.905
G3	-0.162	0.034	0.801	0.905	1.000

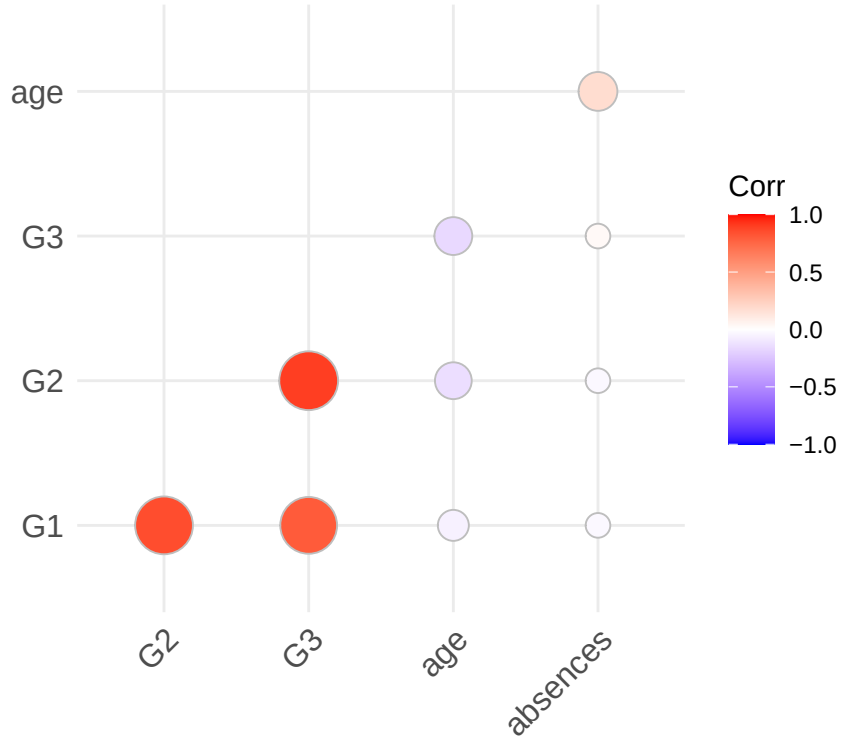


Figure 3: Correlation between five Numerical Variables

4 Methodology

We have used two methods to analyse the student performance dataset. Firstly, we have used Principle Component Analysis to see if we can reduce the number of variables to model the students' final grade. We have four numerical predictors(First test score, Second test score, Number of absences and Age). we would like to reduce the by using PCA. Secondly, we have used Linear Discriminant Analysis to classify the students whether they pass or fail the math course based on the other predictors. We have used 70% of the data to train LDA and 30% of the data to validate the model.

5 Principal Component Analysis - PCA

5.1 PCA Details

We have conducted PCA on 4 numeric variables (G1, G2, Age, Number of Absences) to see if we can perform dimentionality reduction. The table 6 shows summary of PCA results with proportion of variances explained by each component and the cumulative proportion. The first and the second principle components explained 96% of total variance. Figure 4 shows the visualization of proportion of variance explained by each principle components. It seems two principle components are enough to explain most of the variance.

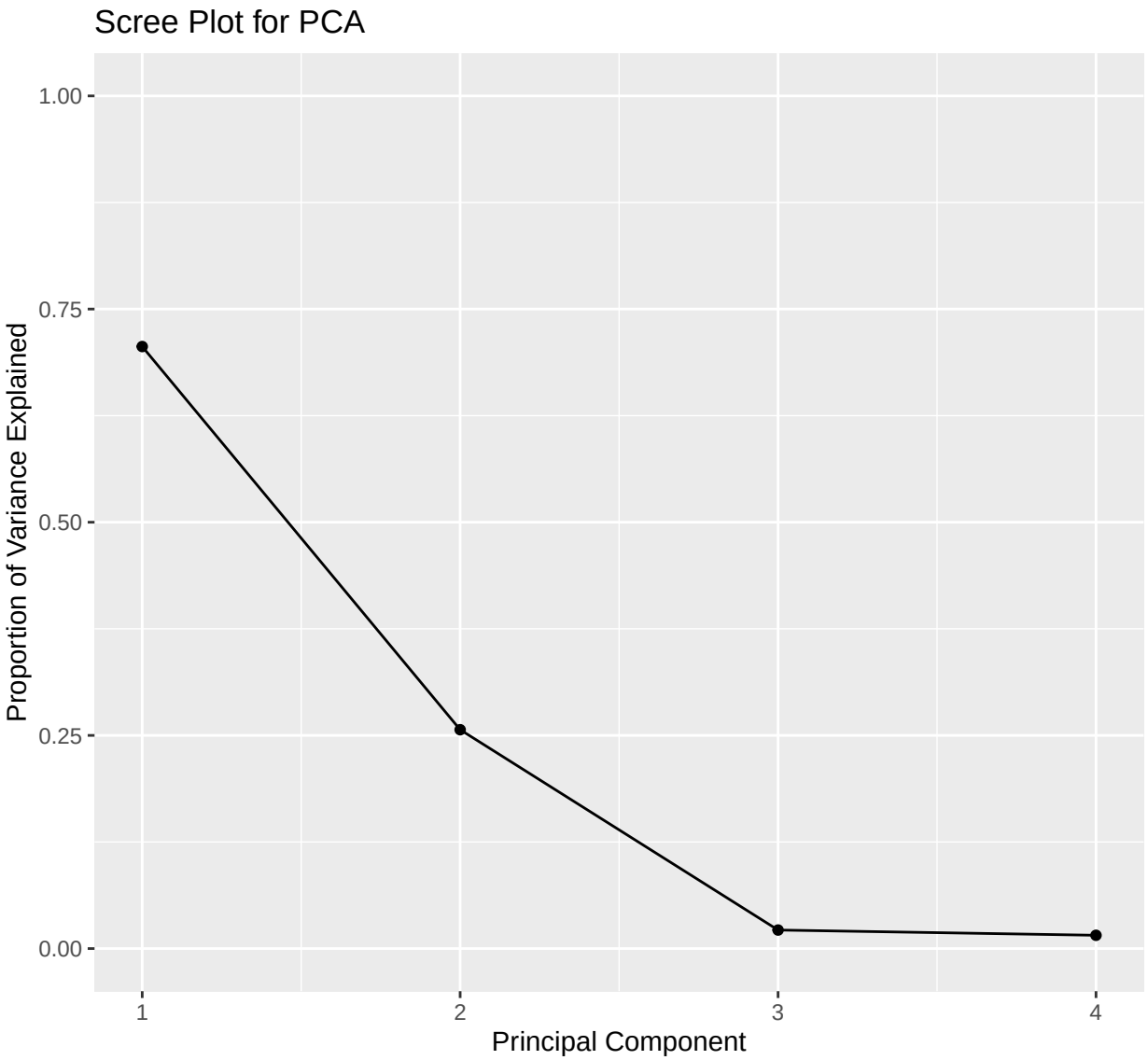


Figure 4: Scree Plot of PCA in %

Table 6: Summary of PCA

	PC1	PC2	PC3	PC4
Standard deviation	8.0088	4.8287	1.4047	1.1888
Proportion of Variance	0.7061	0.2567	0.0217	0.0156
Cumulative Proportion	0.7061	0.9627	0.9844	1.0000

5.2 Fit Linear Regression

We fitted three Linear Regression models. A model which has all numerical predictors in the model, a model that contains the first and second principal components and a last model includes the first, second and third principal components. The model equations are:

$$Full\ model : G3_i = \beta_0 + \beta_1 G1_i + \beta_2 G2_i + \beta_3 Age_i + \beta_4 Absence_i + \epsilon_i$$

$$PC12\ model : G3_i = \beta_0 + \beta_1 PC1_i + \beta_2 PC2_i + \epsilon_i$$

$$PC123\ model : G3_i = \beta_0 + \beta_1 PC1_i + \beta_2 PC2_i + \beta_3 PC3_i + \epsilon_i$$

Figure 5, 6 and 7 show the diagnostic plots of each model. It seems all models show violation of Normality and errors have some patterns. PCA could not fix the violation assumptions.

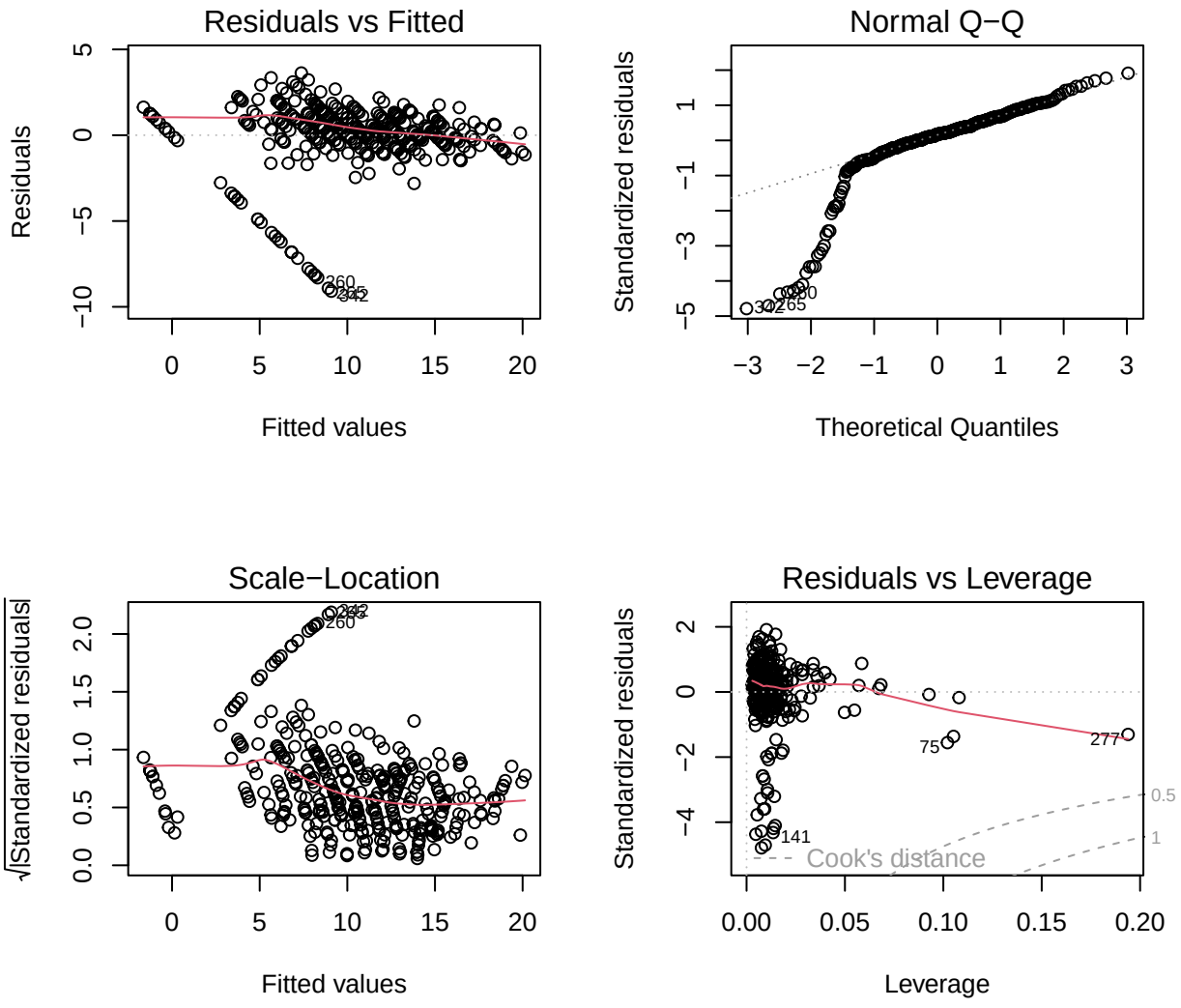


Figure 5: Diagnostic Plot for Full Model

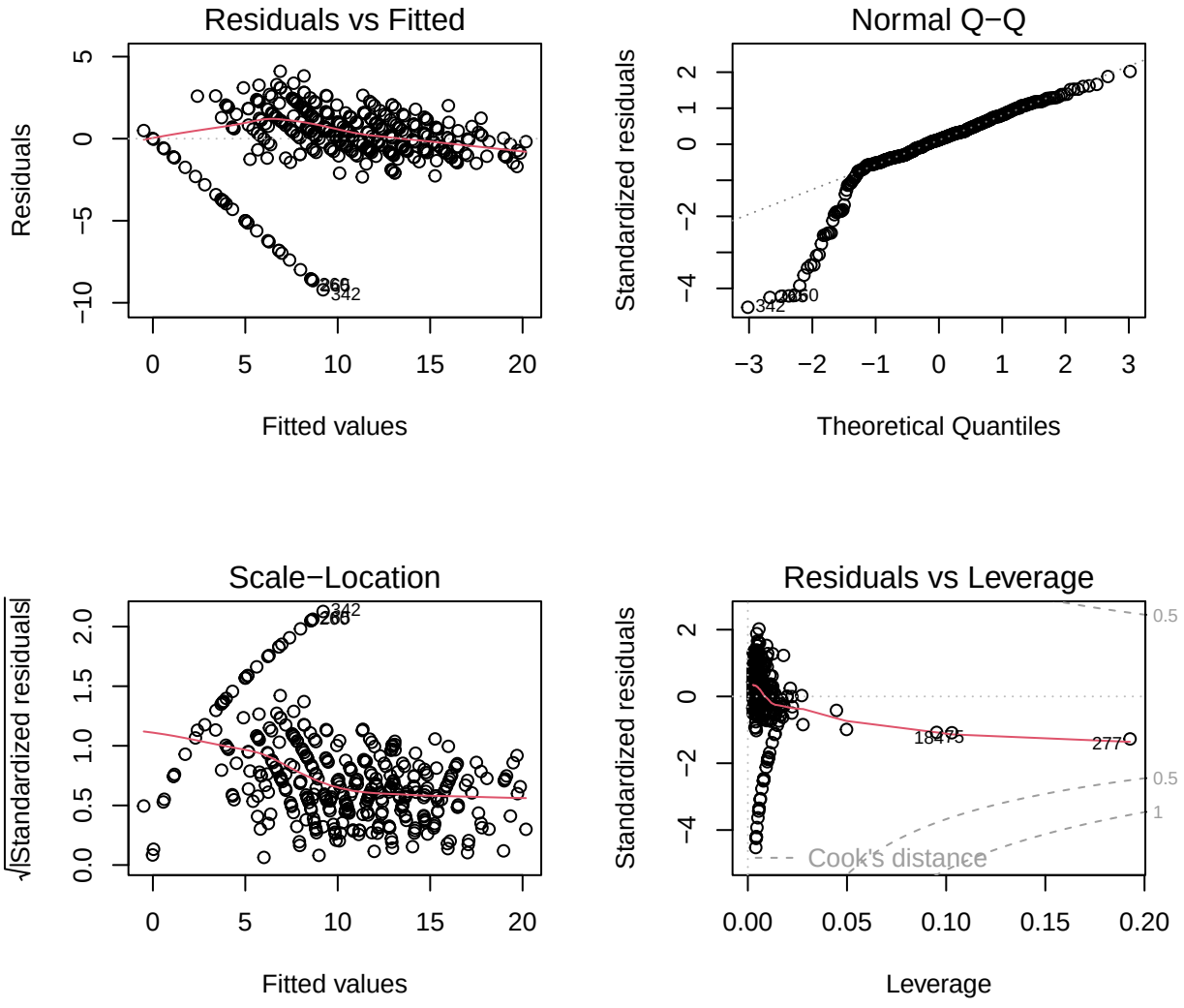


Figure 6: Diagnostic Plot for Model with PC1 and PC2

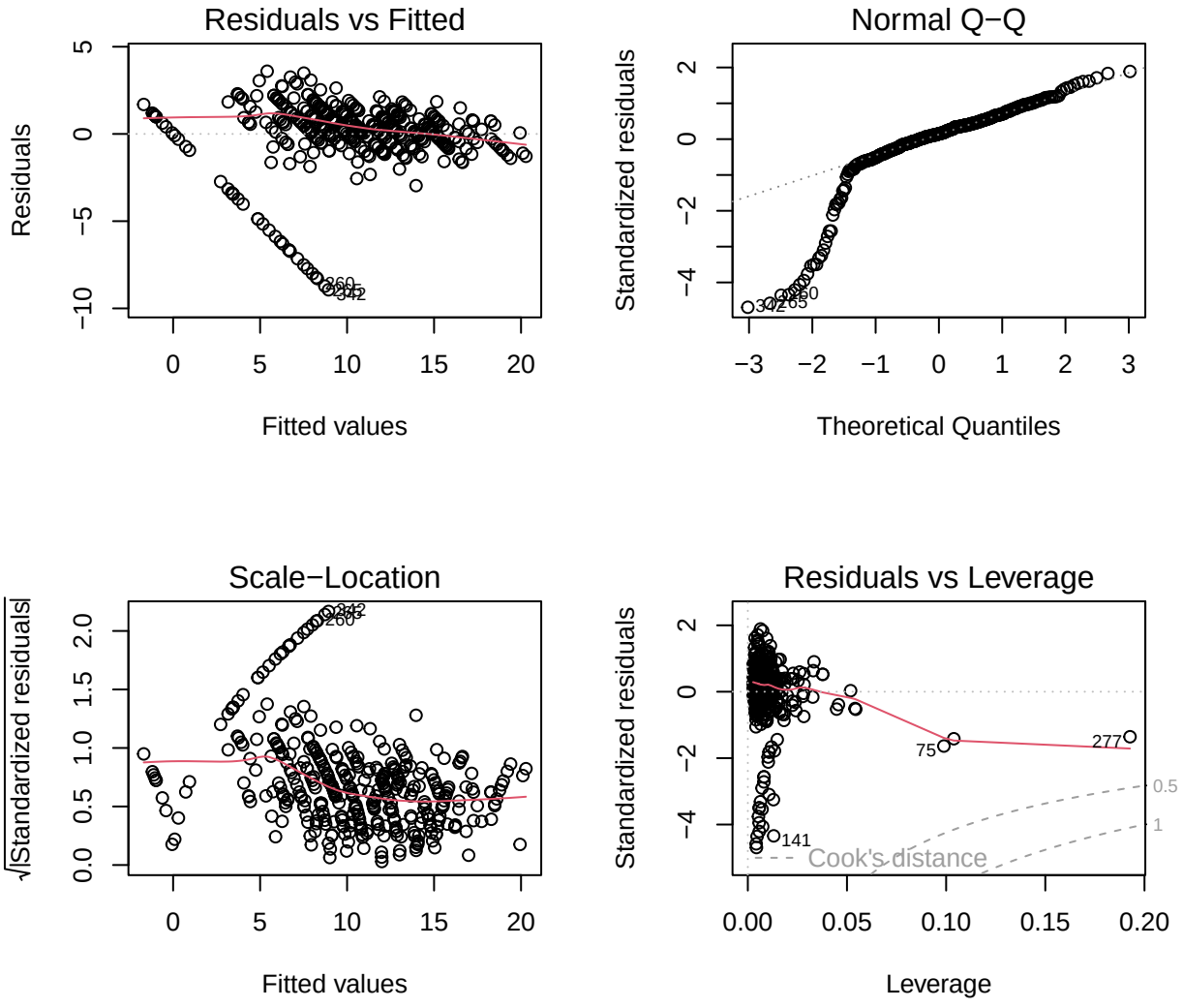


Figure 7: Diagnostic Plot for Model with PC1, PC2 and PC3

5.3 Check AIC, BIC and Adjusted R^2

We have checked the model fit by AIC, BIC and Adjusted R^2. The table 7 shows the results. In contrast with what we have seen in Figure 4 scree plot before, the model with the first and second principle components have the highest AIC, BIC and lowest Adjusted R^2 value. AIC from full model and AIC from the model with three principle components do not differ much. Model with three principle components have the lowest BIC. Thus, PC123 model is the 'best' model so far. We can conclude that by using PCA, we can reduce the number of variables to 3 instead of using 4 variables to predict G3.

Table 7: Information Criterion

	AIC	BIC	Adjusted R^2
Full Model	1637.29	1661.16	0.83
ModelPC12	1689.13	1705.05	0.80
ModelPC123	1637.31	1657.20	0.83

6 Linear Discriminant Analysis - LDA

After EDA and PCA, we are familiar with our dataset, and we want to actually use this dataset to predict whether a student will Fail or Pass the math course, if we have all the information except the final grade G3 of this student.
From our previous analyse in EDA, there might have a way to define G3 by a linear combination of the input variables. Therefore, LDA will be applied.

6.1 Define Label

Since G3 will be the class label, and the full mark of G3 is 20, therefore, grades that below 10 will be labelled as Fail, and equal and above 10 will be Pass.
The table shows the total students in each labels, there are 265 students pass the course, and 130 students who failed.

Fail	Pass
130	265

6.2 LDA Underlying hypotheses

There are two main hypotheses that need to check before any further analyse.
- The classes are linearly separable.
- The covariance matrices are not too different.

6.2.1 Pairs plot

Figure 8 shows the pairs plot of 4 numerical varialbes with two class labels. The red dots are Fail, and the green dots are Pass.
From this plot, we have correlation coefficients, in scatterplots, we can see G1 and G2 are good to separate between two groups, and the two classes are kind of linearly separable, in other cases, there is an overlapping and not a clear separation between two groups.

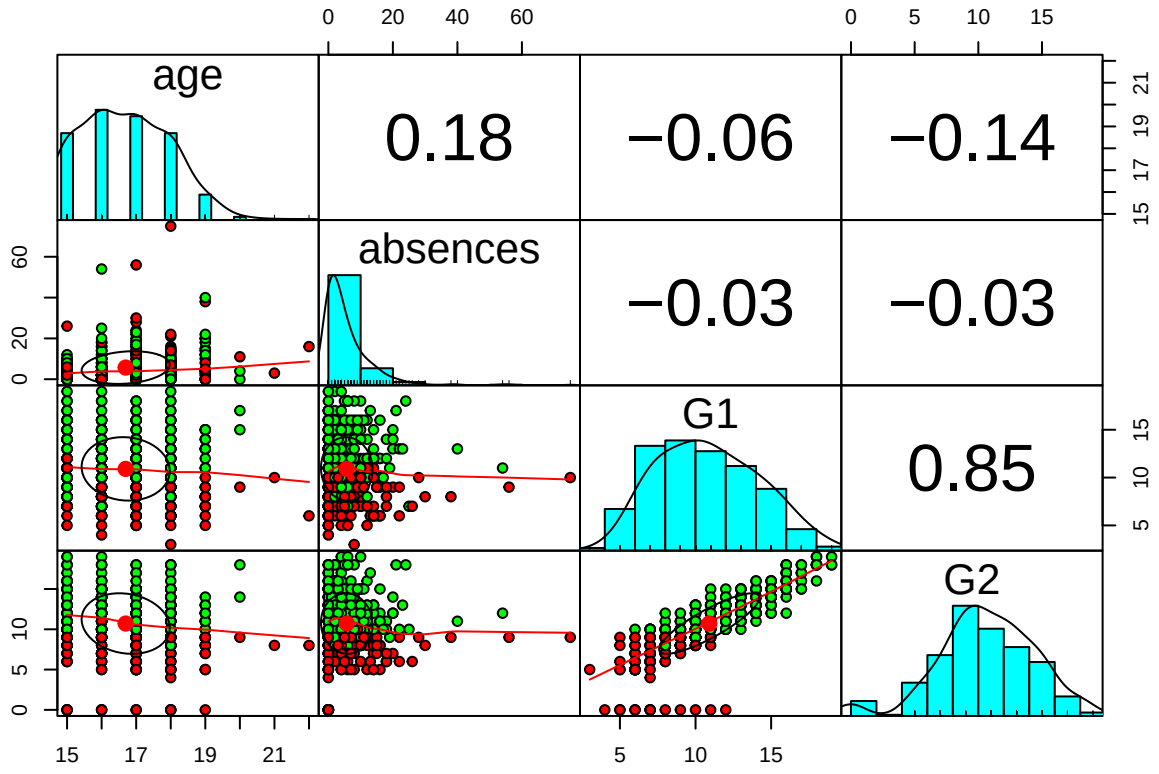


Figure 8: Pairs plot of 4 numerical variables with two class label colors

6.2.2 Covariance matrix

Table 9 and table 10 are the covariance matrices of two groups.
We have conducted two-sample HD test to check equal covariance matrices. Table 11 show the result of the test indicating unequal covariance matrices. From the equal covariance matrices test result, it shows that LDA might not work well. However, by looking at the covariance matrices, it seems they are practically not much different with each other. Thus, we will continue with LDA.

Table 9: Covariance Matrix of Fail Group

	age	absences	G1	G2
age	1.91	1.61	-0.14	0.35
absences	1.61	109.76	2.41	6.41
G1	-0.14	2.41	2.90	1.70
G2	0.35	6.41	1.70	7.36

Table 10: Covariance Matrix of Pass Group

	age	absences	G1	G2
age	1.42	1.63	0.42	-0.27
absences	1.63	41.14	0.03	-1.55
G1	0.412	0.03	7.76	6.06
G2	-0.27	-1.55	6.06	6.41

Table 11: Two-Sample HD test Result

Test statistics	P value	Alternative hypothesis
7.519	0 * * *	two.sided

6.3 LDA Histogram

We split the dataset into training data and test data randomly, we use 70% of the original dataset as training data, and 30% as test data.
We first want to train LDA with the training dataset.
Table 12 shows there are 34.0% students in training data are Fail, and 66.0% are Pass. This result is kind of what we expected, since we have 130 Fail and 265 Pass.

Table 12: Prior probabilities of groups

Fail	Pass
0.340	0.659

Figure 9 shows that the separation between two groups, but there is an overlapping which is not great. It seems for LDA, it is hard to separate students into two groups for students who have a G3 mark around 10.

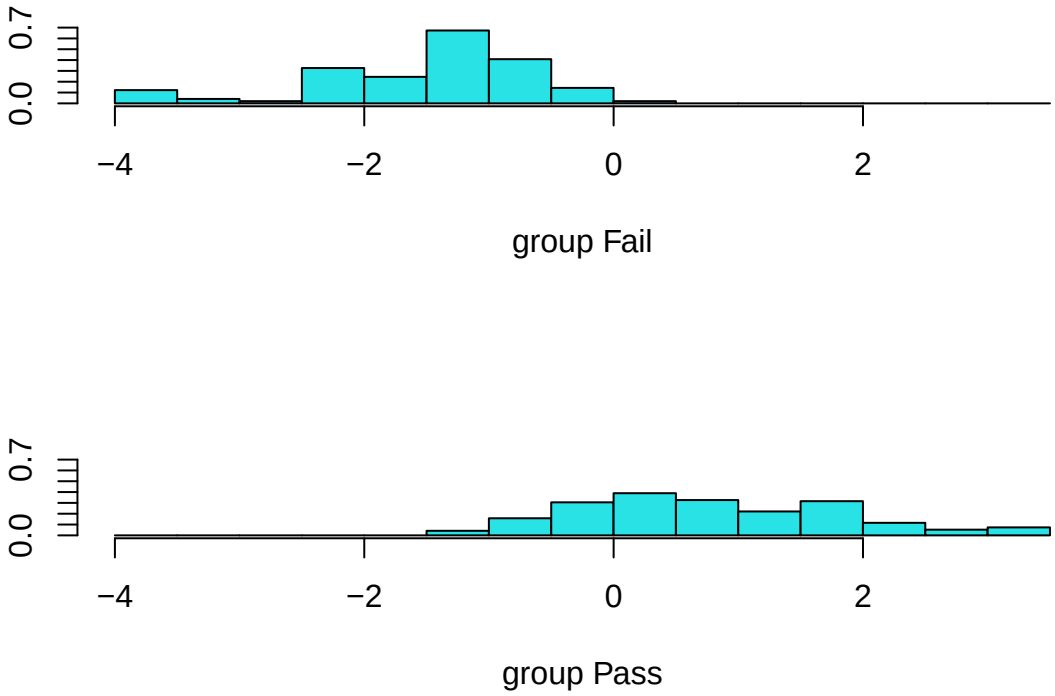


Figure 9: LDA histogram of two groups

6.4 Partition Plot

Figure 10 is the partition plot. In general, we have a really nice result with small error rate. The smallest error rate is 0.108 which is the plot of G1 and G2, this result is what we expected, since G1 and G2 have the highest correlation with G3. In the pairs plot above, this also shows that this plot may give us the best result.

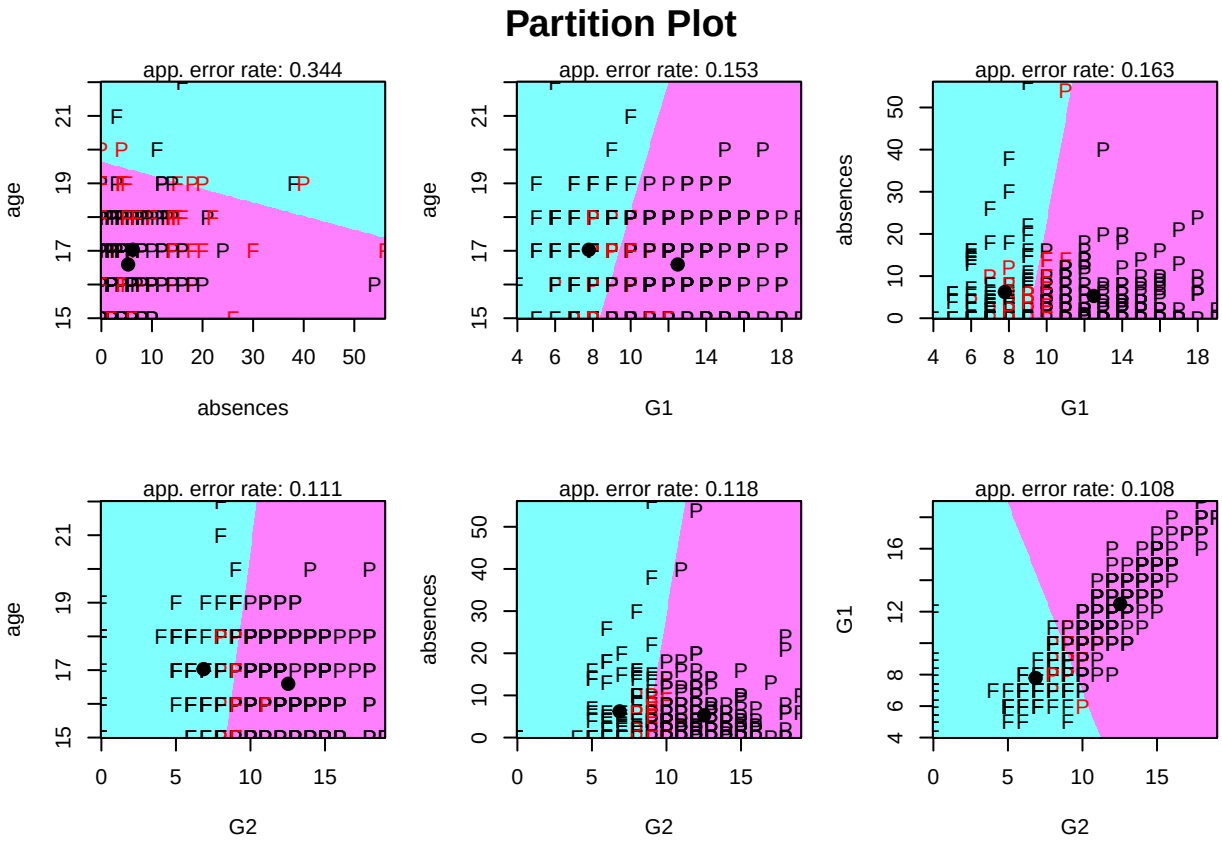


Figure 10: Partition plot of training data

6.5 Confusion matrix

Table 13 and table 14 are the confusion matrices. From the the matrices, training data has an accuracy of 0.89 and test data has accuracy of 0.92, both give us a nice result.
However, the total students in two groups are not equal, therefore, we need to check Recall and Precision as well.
In training data, we have precision 0.84 and recall 0.82, and in test data we have precision 0.87 and recall 0.84. The result here in both training data and test data gives a good result. In general, LDA gives a quite nice model, and the dataset can fit LDA well.

Table 13: Confusion matrix of training data

	Fail	Pass
Fail	80	15
Pass	18	175

Table 14: Confusion matrix of test data

	Fail	Pass
Fail	27	4
Pass	5	71

7 Discussion

7.1 Problems

- We tried using a Naive Bayes model to see if it gives us a better model, but the result is not as good as LDA, and we do not show the results here.

- There are many other variables in the original dataset. They may have some potential relationship between these variables that effect G3 and may give us a totally different result, or explain any anomalies present within the data. We did not have access to these variables hence we do not know the impact it may have on our analysis.

- The dataset used is only for students in 2005/2006, whereas it is currently 2022. A dataset for 2021/2022 or within the last 5 years it would be of more benefit towards the current generation of high school students as the curriculum has and will change over time. It would of even better benefit if we could acquire data consisting of results from schools in other countries to provide context as to how other countries compare against each other.

7.2 Conclusion

- We believe our analysis of student academic grades for math course of two Portuguese secondary public school collected in 2006 was quite successful. As our two main analysis techniques of PCA and LDA gave us results that aligned with our reasons for using them.

- With the PCA we successfully reduced the number of variables to 3 instead of 4. Shown by the favorable AIC, BIC and adjusted R squared scores of the PC123 model.

- After verifying assumptions our LDA gave us nice results as well. The G1 and G2 variables (first and second period scores) seemed to be the best indicator for classifying G3. This was represented nicely in the partition plots and later verified in the confusion matrix. This was a very satisfying insight as it aligns with common sense that the performances in other periods would best classify G1 performance.

- Although very satisfied in our analysis of the data. It was only in reviewing the collection of this data that we began to question the relevance of this analysis. As mentioned in our problems it is difficult to see how data from 2005/2006 from Portugal holds relevance in 2021/2022. A more modern collection of data would be great next step to analyse especially after the fantastic insights we gained here.

References

P. Cortez, A. Silva. 2008. "Using Data Mining to Predict Secondary School Student Performance." https://archive.ics.uci.edu/ml/datasets/Student+Performance.