Report #3

GROUP 9

Contents

1 EDA

1.1 Read Data

The dataset (P. Cortez 2008) have 395 records in total with 5 numerical variables and 28 categorical variables. There is no missing value in this dataset.
The numerical variables include age, number of school absences, 1st Period grade, 2nd period grade and final period grade.
Since we have too many categorical variables, for our data analysis, we will only include school, sex and number of past failures. The table below shows the first 5 records.

	School	Sex	Failures	age	absences	G1	G2	G3
1	GP	F	0	18	6	5	6	6
2	GP	F	0	17	4	5	5	6
3	GP	F	3	15	10	7	8	10
4	GP	F	0	15	2	15	14	15
5	GP	F	0	16	4	6	10	10

1.2 Summary Table

1.2.1 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

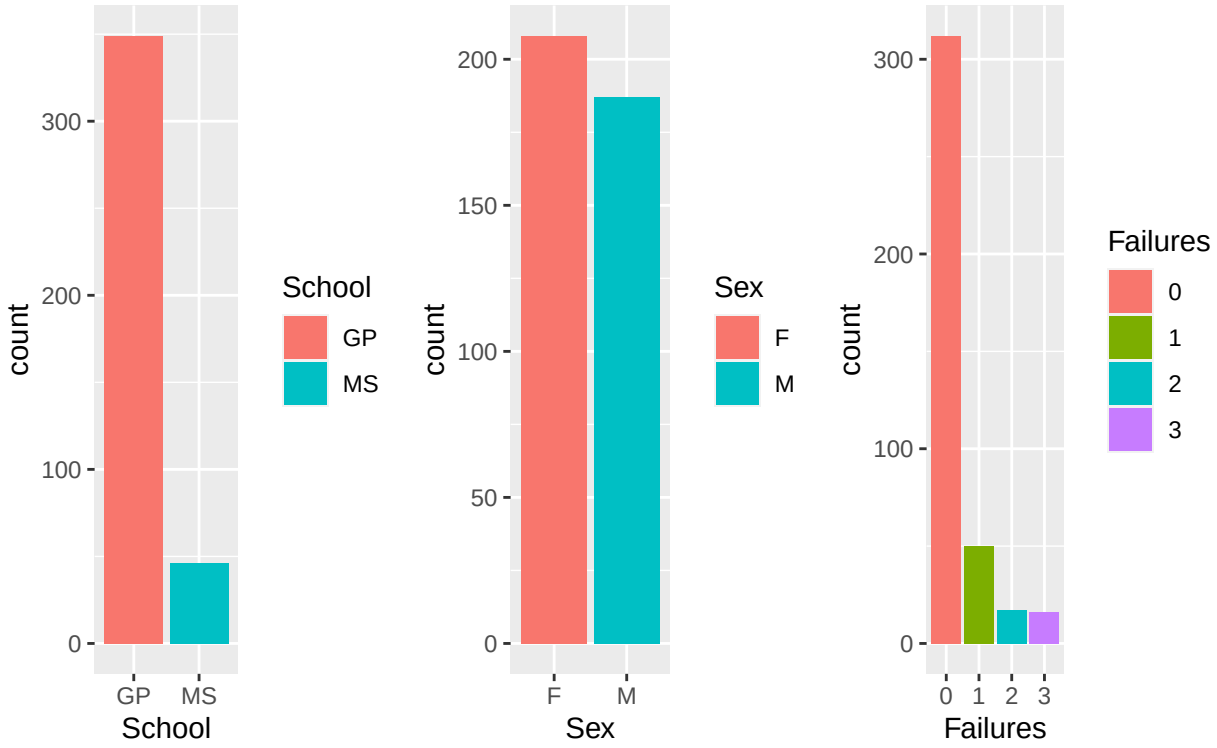

Figure 1: Barplot for Categorical Variables

1.2.2 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.

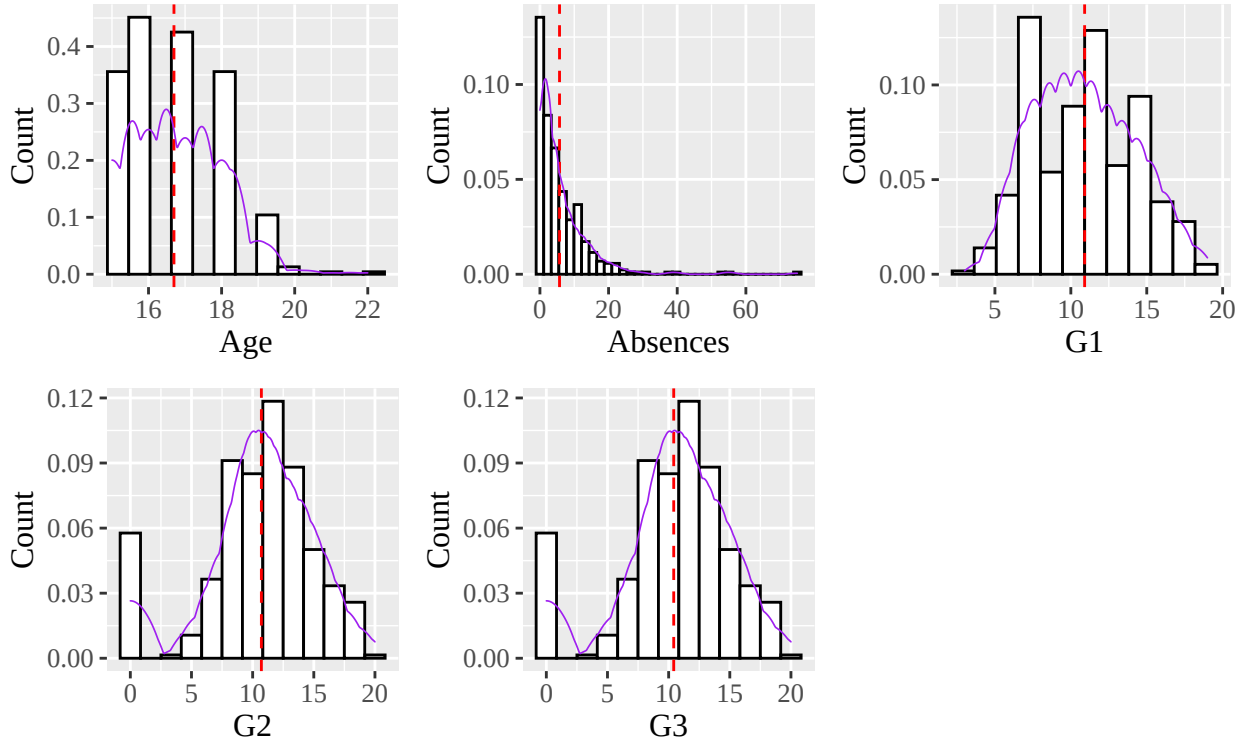

Figure 2: Barplot for Numerical Variables

1.3 Skewness and kurtosis

Skewness is a statistical numerical method to measure the asymmetry of the distribution or data set.
Kurtosis is a numerical method in statistics that measures the sharpness of the peak in the data distribution.

1.3.1 Skewness

Table 5 shows skewness of five numerical variables.
Age, absences, and G1 have positive skewness, the graph is said to be positively skewed with the majority of data values less than mean, most of the values are concentrated on the left side.
Absences skewness is the biggest among all five values, the graph is more concentrated on the left side than age and G1.
G1 skewness is closest to 0, the graph is said likely to be symmetric.

G2 and G3 have negative skewness, the graph is said to be negatively skewed with the majority of data values greater than mean, most of the values are concentrated on the right side of the graph.
G3 has a smaller skewness so the graph is more concentrated on the right side than G2.

Table 5: Skewness of Five numerical Variables

age	absences	G1	G2	G3
0.464	3.658	0.240	-0.430	-0.730

1.3.2 Kurtosis

Table 6 shows kurtosis of five numerical variables.
All five values are positive, value of G1 is smaller than 3, then the data distribution is platykurtic.
Age is approximately close to 3, then the data distribtuion is mesokurtic.
Absences has the biggest kurtosis value, it is much bigger than 3, then the data distribution is leptokurtic and shows a sharp peak on the graph.
G2 and G3 are slightly bigger than 3, then the data distribution is leptokurtic, but the sharp peak on the graph is not so obvious.

Table 6: Kurtosis of Five numerical Variables

age	absences	G1	G2	G3
2.984	24.430	2.300	3.605	3.383

1.3.3 Visualisation

For both skewness and kurtosis, Figure 2 shows the distribution of five numerical variables, and the sharpness of the peak.

1.4 Correlation Matrix and plot

Table 7 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 7: 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

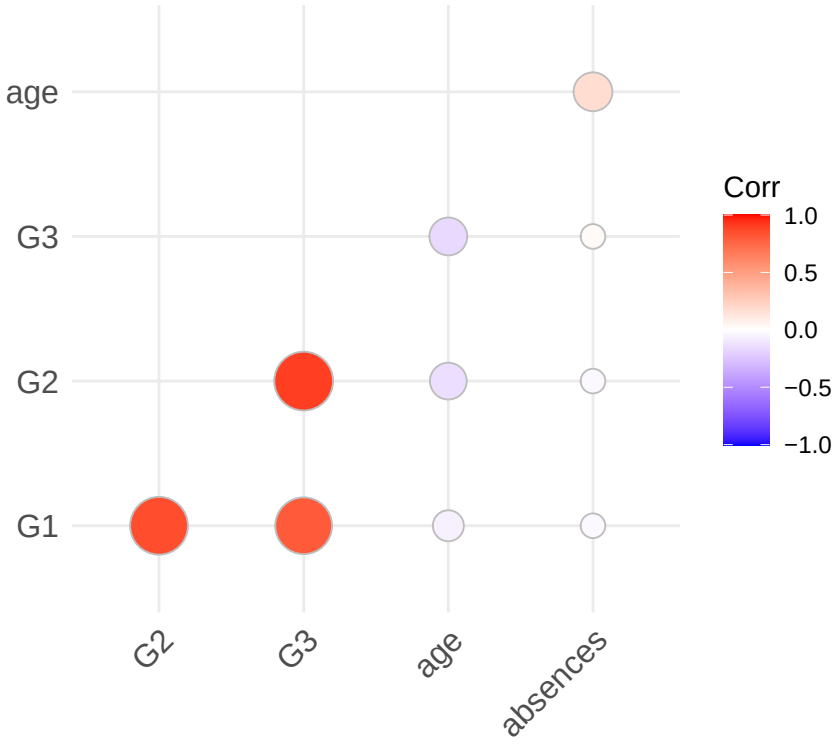

Figure 3: Correlation between five Numerical Variables

1.5 Box Plot

The Figure 4 shows the box plots of final grade, number of absences and age grouped by the three categorical variables (School, Sex and Failures). The box plots show that the final grade does not differ much among school or sex but it differs by the number of past failures. Even though there are variations, those with fewer number of failures tend to have higher final grades. All box plot for the number of absences shows that they are heavily skewed to large number. The box plot for age shows that Mousinho da Silveira school have higher median age compared to Gabriel Pereira school.

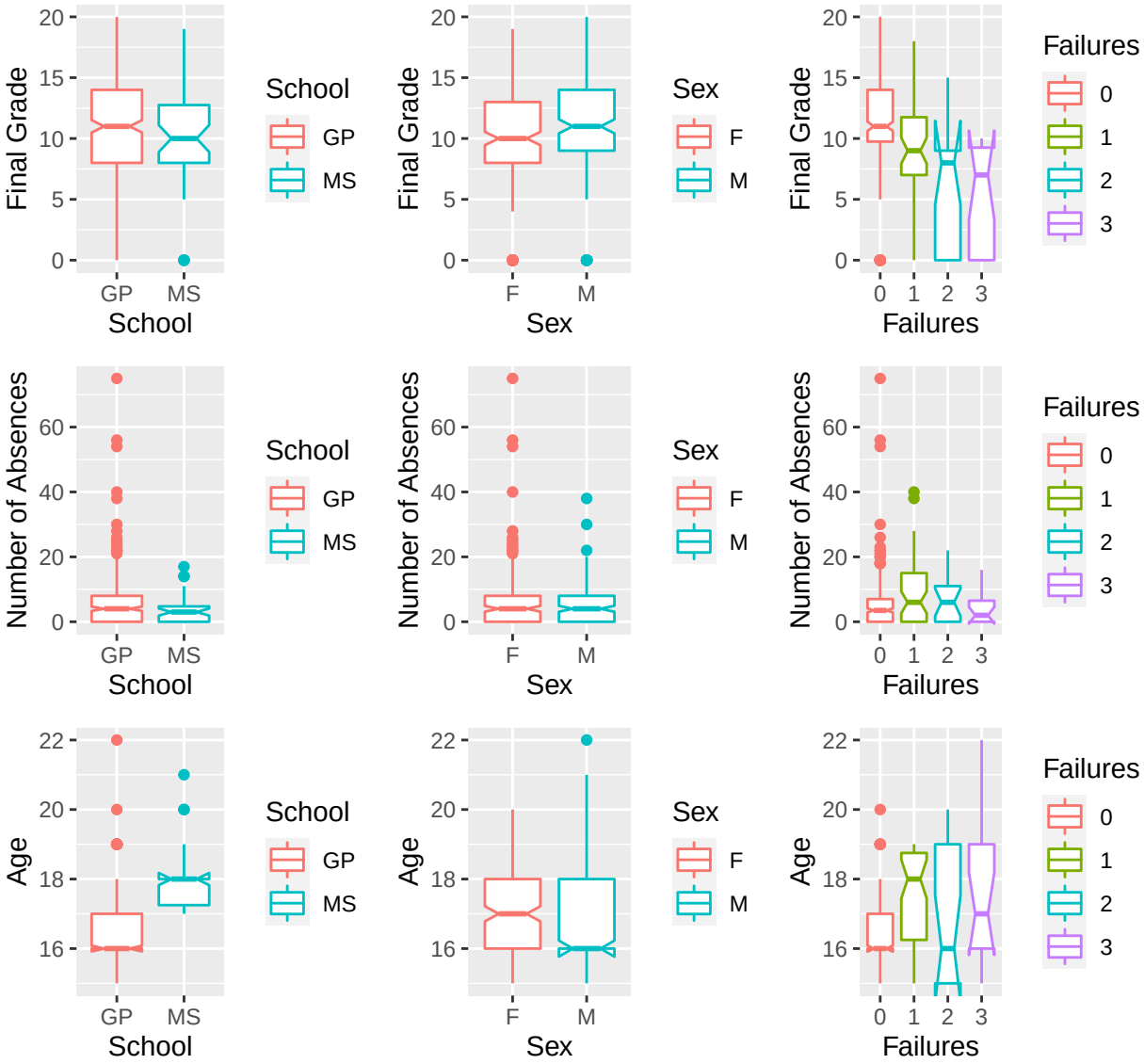

Figure 4: Boxplot of Final grade against 3 categorical variables

1.6 Pairs Plot

The Figure 5 shows the pairs plot on the numeric variables using ggpairs. As similar to Table 7 and Figure 3, the pairs plot shows strong positive correlations between G1, G2 and G3. Both age and absences have right skewed distributions.

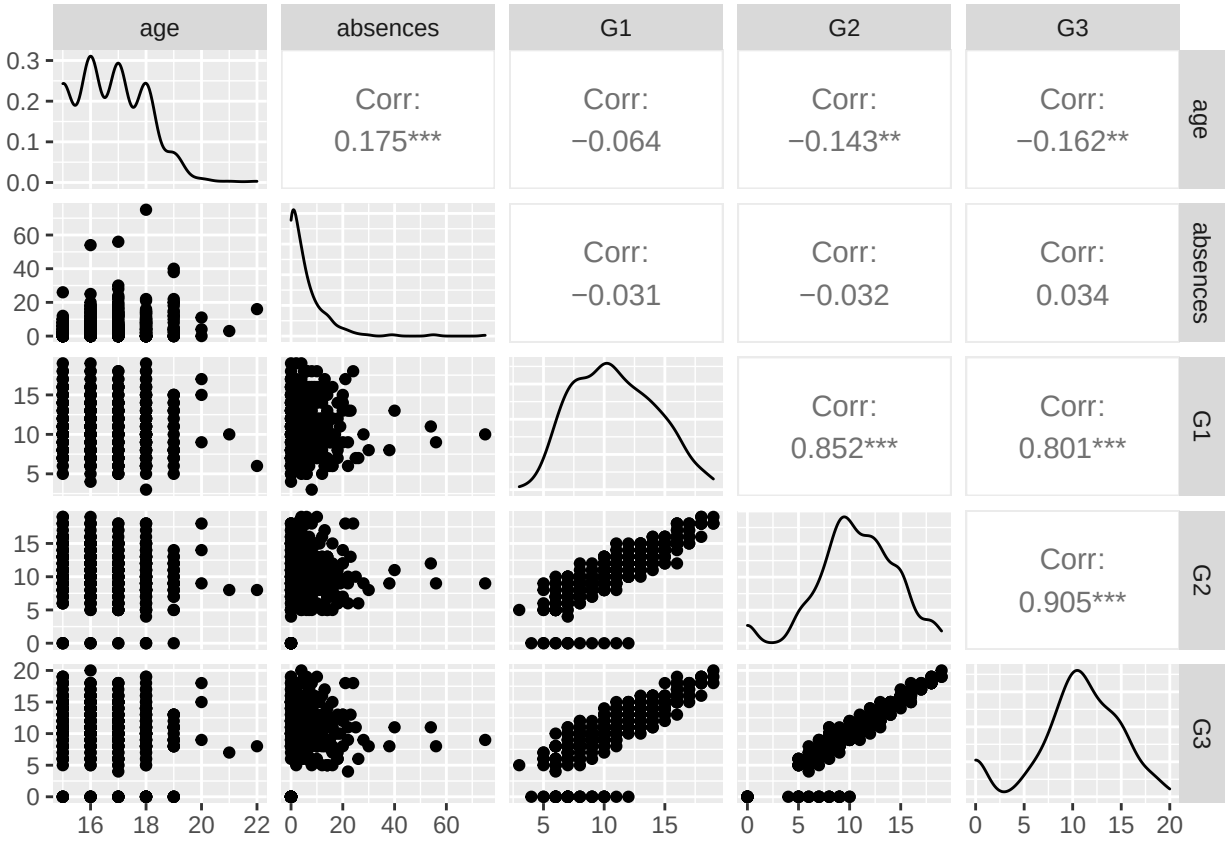

Figure 5: Pairs plot on the numerical variables

References

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