Student Grade Prediction
Exploratory Analysis

Haziq Zed, hzed@bellarmine.edu
Student Grade Prediction www.kaggle.com/dipam7/student-grade-prediction

I. INTRODUCTION

This paper outlines an investigation of the Student Grade Prediction from Kaggle. The dataset contains variables that may influence the G3 math grade for the students in the two schools located in Portugal. This data set was chosen as it contained a mix of categorical variables along with independent variables that could have an influence on a student's final math grade in secondary school, and can be applied to a broader sample of individuals, perhaps students in a more general population.

II. DATA SET DESCRIPTION

This dataset contains 395 samples with 33 columns. Sixteen columns are integer data, and seventeen columns are multi-valued discrete. A complete listing of the variable names, data types and proportion of missing data is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
school	Nominal (binary)	0%
sex	Nominal (binary)	0%
age	Ratio (integer)	0%
address	Nominal (binary)	0%
famsize	Nominal (binary)	0%
Pstatus	Nominal (binary)	0%
Medu	Ratio (integer)	0%
Fedu	Ratio (integer)	0%
Mjob	Ordinal (multi-valued)	0%
Fjob	Ordinal (multi-valued)	0%
reason	Ordinal (multi-valued)	0%
guardian	Ordinal (multi-valued)	0%
traveltime	Ratio (integer)	0%
studytime	Ratio (integer)	0%
failures	Ratio (integer)	0%
schoolsup	Nominal (binary)	0%
famsup	Nominal (binary)	0%
paid	Nominal (binary)	0%
activities	Nominal (binary)	0%
nursery	Nominal (binary)	0%
higher	Nominal (binary)	0%
internet	Nominal (binary)	0%
romantic	Nominal (binary)	0%
famrel	Ratio (integer)	0%
freetime	Ratio (integer)	0%
goout	Ratio (integer)	0%
Dalc	Ratio (integer)	0%
Walc	Ratio (integer)	0%
health	Ratio (integer)	0%
absences	Ratio (integer)	0%
G1	Ratio (integer)	0%
G2	Ratio (integer)	0%

G3	Ratio (integer)	0%

III. Data Set Summary Statistics

This data set contains 16 continuous variables. A note to notice when looking at the continuous variables is that some of the variables are based on a scale which may not identify as a true ratio data that can be used for analysis. For example, Medu and Fedu, Mother's Education and Father's Education, is based on a scale between 0-4 based on a binary system.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
age	395.0	16.696203	1.276043	15.0	16.0	17.0	18.0	22.0
Medu	395.0	2.749367	1.094735	0.0	2.0	3.0	4.0	4.0
Fedu	395.0	2.521519	1.088201	0.0	2.0	2.0	3.0	4.0
traveltime	395.0	1.448101	0.697505	1.0	1.0	1.0	2.0	4.0
studytime	395.0	2.035443	0.839240	1.0	1.0	2.0	2.0	4.0
failures	395.0	0.334177	0.743651	0.0	0.0	0.0	0.0	3.0
famrel	395.0	3.944304	0.896659	1.0	4.0	4.0	5.0	5.0
freetime	395.0	3.235443	0.998862	1.0	3.0	3.0	4.0	5.0
goout	395.0	3.108861	1.113278	1.0	2.0	3.0	4.0	5.0
Dalc	395.0	1.481013	0.890741	1.0	1.0	1.0	2.0	5.0
Walc	395.0	2.291139	1.287897	1.0	1.0	2.0	3.0	5.0
health	395.0	3.554430	1.390303	1.0	3.0	4.0	5.0	5.0
absences	395.0	5.708861	8.003096	0.0	0.0	4.0	8.0	75.0
G1	395.0	10.908861	3.319195	3.0	8.0	11.0	13.0	19.0
G2	395.0	10.713924	3.761505	0.0	9.0	11.0	13.0	19.0
G3	395.0	10.415190	4.581443	0.0	8.0	11.0	14.0	20.0

Table 3: Proportions for school (n=395.0)

Category	*Frequency*	*Proportion (%)*
GP	*349*	*88%*
MS	*46*	*12%*

Table 4: Proportions for sex (n=395.0)

Category	*Frequency*	*Proportion (%)*
F	*208*	*53%*
M	*187*	*47%*

Table 5: Proportions for address (n=395.0)

Category	*Frequency*	*Proportion (%)*
R	*88*	*78%*
U	*307*	*22%*

Table 6: Proportions for famsize (n=395.0)

Category	*Frequency*	*Proportion (%)*
GT3	*281*	*71%*
LE3	*114*	*29%*

Table 7: Proportions for Pstatus (n=395.0)

Category	*Frequency*	*Proportion (%)*
A	*41*	*90%*
T	*354*	*10%*

Table 8: Proportions for Mjob (n=395.0)

Category	*Frequency*	*Proportion (%)*
at_home	*59*	*15%*
health	*34*	*9%*
other	*141*	*35%*
services	*103*	*26%*
teacher	*58*	*15%*

Table 9: Proportions for Fjob (n=395.0)

Category	*Frequency*	*Proportion (%)*
at_home	*20*	*5%*
health	*18*	*5%*
other	*217*	*55%*
services	*111*	*28%*
teacher	*29*	*7%*

Table 10: Proportions for reason (n=395.0)

Category	*Frequency*	*Proportion (%)*
course	*145*	*37%*
home	*109*	*28%*
other	*36*	*9%*
reputation	*105*	*26%*

Table 11: Proportions for guardian (n=395.0)

Category	*Frequency*	*Proportion (%)*
father	*90*	*23%*
mother	*273*	*69%*
other	*32*	*8%*

Table 12: Proportions for schoolsup (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*344*	*87%*
yes	*51*	*13%*

Table 13: Proportions for famsup (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*153*	*39%*
yes	*242*	*61%*

Table 14: Proportions for paid (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*214*	*54%*
yes	*181*	*46%*

Table 15: Proportions for activities (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*194*	*49%*
yes	*201*	*51%*

Table 16: Proportions for nursery (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*81*	*21%*

yes	*314*	*79%*

Table 17: Proportions for higher (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*20*	*5%*
yes	*375*	*95%*

Table 18: Proportions for internet (n=395.0)

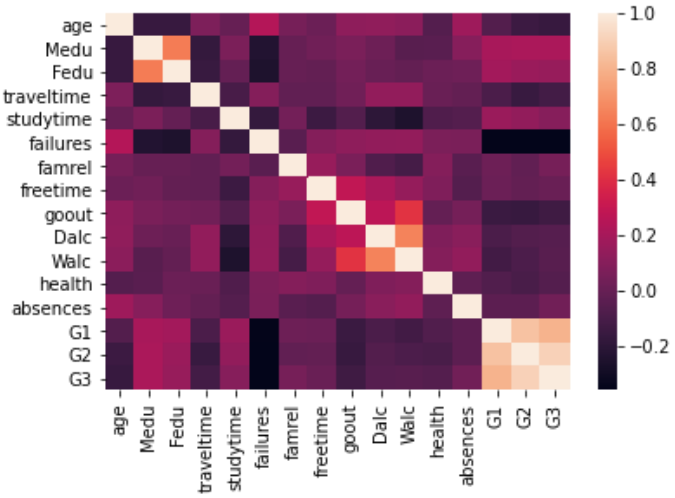
Category	*Frequency*	*Proportion (%)*
no	*66*	*17%*
yes	*329*	*83%*

Table 19: Proportions for romantic (n=395.0)

Category	*Frequency*	*Proportion (%)*
no	*263*	*67%*
yes	*132*	*33%*

Table 4: Correlation Table/Tables

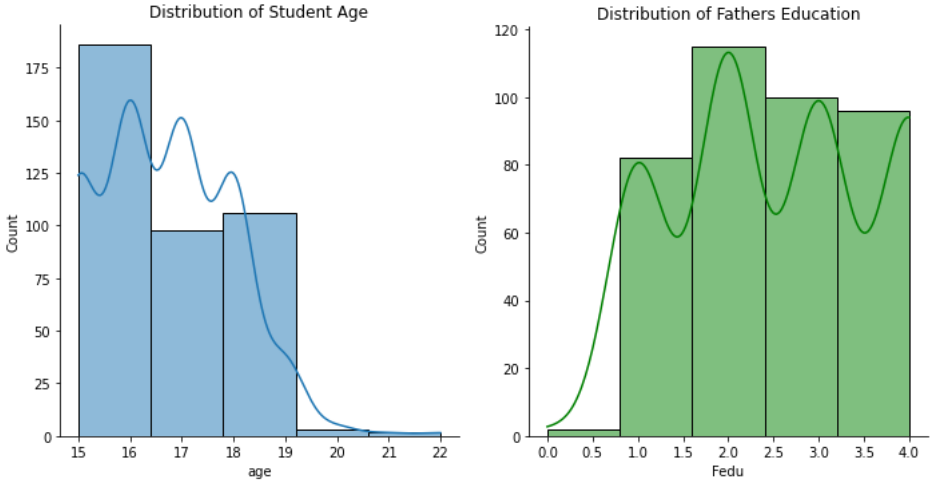
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	1.00	-0.16	-0.16	0.07	-0.00	0.24	0.05	0.02	0.13	0.13	0.12	-0.06	0.18	-0.06	-0.14	-0.16
2	-0.16	1.00	0.62	-0.17	0.06	-0.24	-0.00	0.03	0.06	0.02	-0.05	-0.05	0.10	0.21	0.22	0.22
3	-0.16	0.62	1.00	-0.16	-0.01	-0.25	-0.00	-0.01	0.04	0.00	-0.01	0.01	0.02	0.19	0.16	0.15
4	0.07	-0.17	-0.16	1.00	-0.10	0.09	-0.02	-0.02	0.03	0.14	0.13	0.01	-0.01	-0.09	-0.15	-0.12
5	-0.00	0.06	-0.01	-0.10	1.00	-0.17	0.04	-0.14	-0.06	-0.20	-0.25	-0.08	-0.06	0.16	0.14	0.10
6	0.24	-0.24	-0.25	0.09	-0.17	1.00	-0.04	0.09	0.12	0.14	0.14	0.07	0.06	-0.35	-0.36	-0.36
7	0.05	-0.00	-0.00	-0.02	0.04	-0.04	1.00	0.15	0.06	-0.08	-0.11	0.09	-0.04	0.02	-0.02	0.05
8	0.02	0.03	-0.01	-0.02	-0.14	0.09	0.15	1.00	0.29	0.21	0.15	0.08	-0.06	0.01	-0.01	0.01
9	0.13	0.06	0.04	0.03	-0.06	0.12	0.06	0.29	1.00	0.27	0.42	-0.01	0.04	-0.15	-0.16	-0.13
10	0.13	0.02	0.00	0.14	-0.20	0.14	-0.08	0.21	0.27	1.00	0.65	0.08	0.11	-0.09	-0.06	-0.05
11	0.12	-0.05	-0.01	0.13	-0.25	0.14	-0.11	0.15	0.42	0.65	1.00	0.09	0.14	-0.13	-0.08	-0.05
12	-0.06	-0.05	0.01	0.01	-0.08	0.07	0.09	0.08	-0.01	0.08	0.09	1.00	-0.03	-0.07	-0.10	-0.06
13	0.18	0.10	0.02	-0.01	-0.06	0.06	-0.04	-0.06	0.04	0.11	0.14	-0.03	1.00	-0.03	-0.03	0.03
14	-0.06	0.21	0.19	-0.09	0.16	-0.35	0.02	0.01	-0.15	-0.09	-0.13	-0.07	-0.03	1.00	0.85	0.80
15	-0.14	0.22	0.16	-0.15	0.14	-0.36	-0.02	-0.01	-0.16	-0.06	-0.08	-0.10	-0.03	0.85	1.00	0.90
16	-0.16	0.22	0.15	-0.12	0.10	-0.36	0.05	0.01	-0.13	-0.05	-0.05	-0.06	0.03	0.80	0.90	1.00



IV. DATA SET GRAPHICAL EXPLORATION

This section of the exploratory analysis displays selected visualizations of this data set. **Figure 1** shows the distributions for the demographic variables present in the data set. **Figure 2** shows the relationship between the independent continuous variables against the target output/variable (G3). **Figure 3** shows the bar charts for the different categorical variables in the data set. **Figure 4** shows the box plot chart for a selected categorical variable (address) to identify its relationship with the target variable. **Figure 5** shows the relationship between selected categorical variables and the target variable/output (G3).

A. Distributions



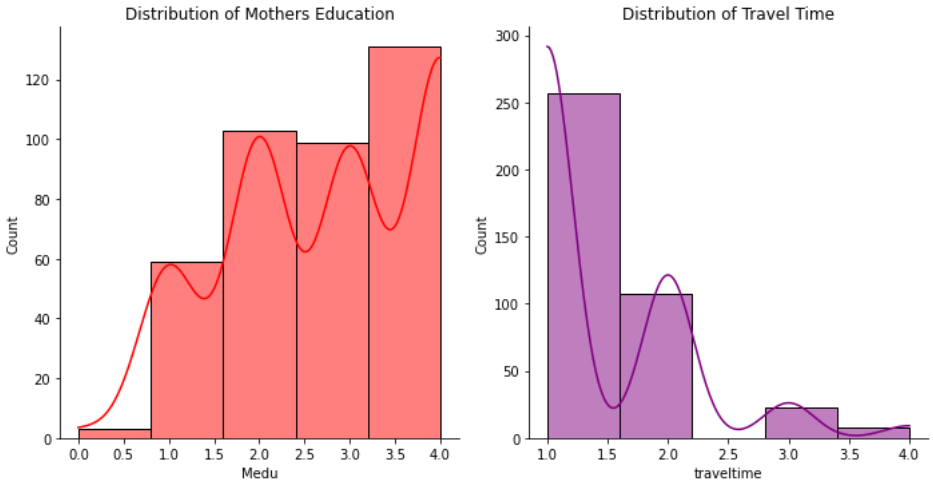


Figure 1: Distribution of Demographic Variables

B. ScatterPlots / Pairwise Plots (continuous variables)

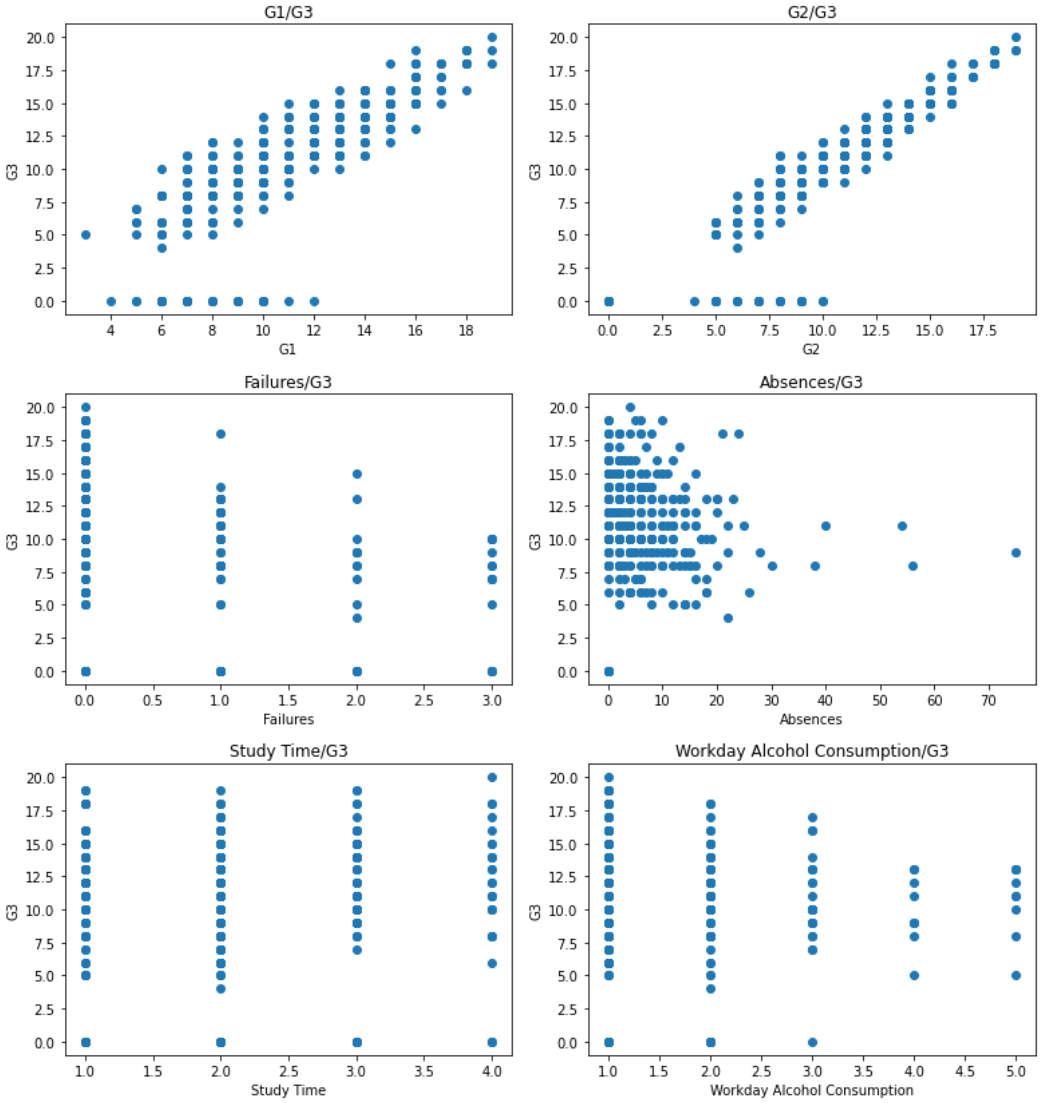
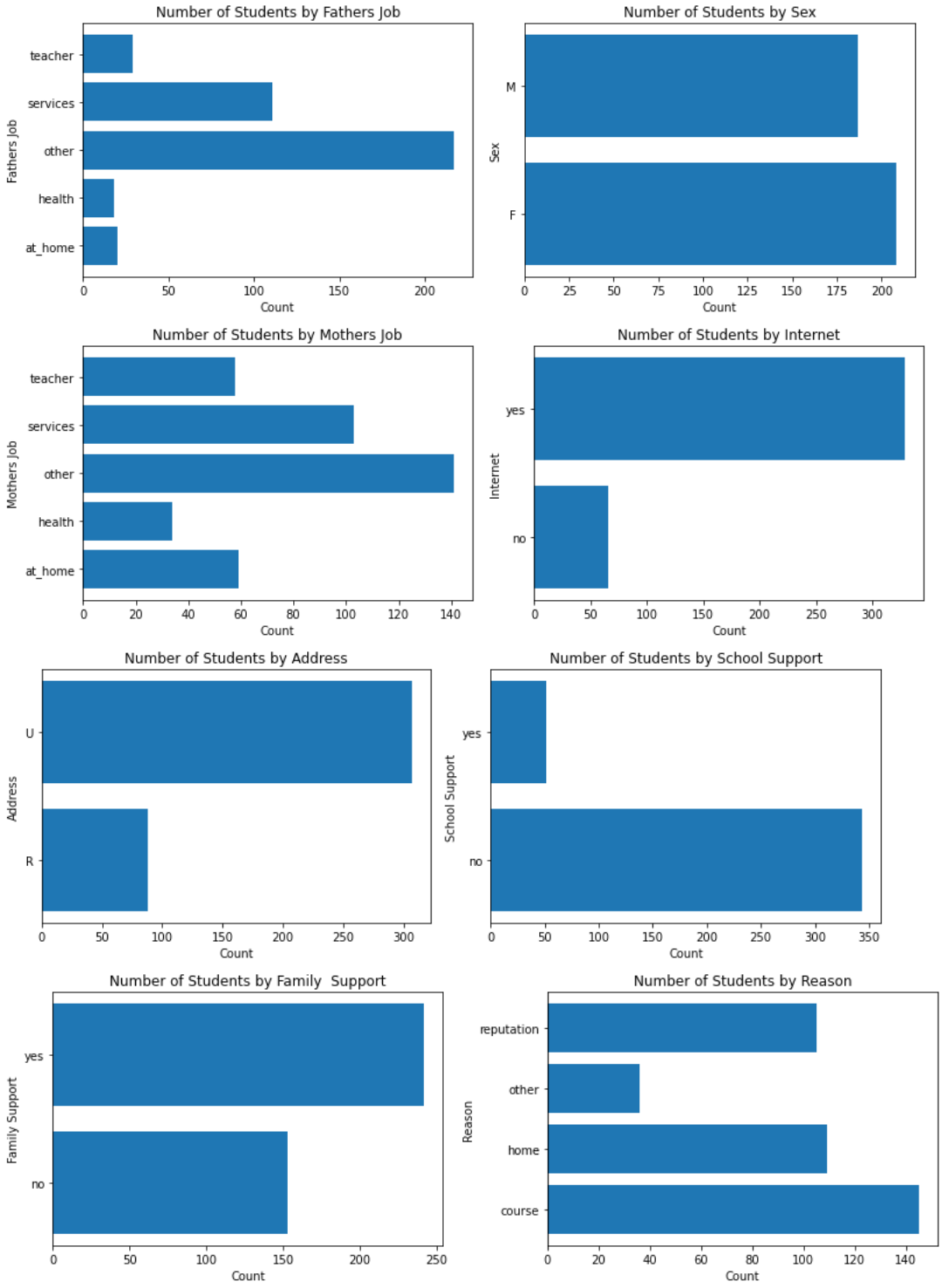


Figure 2: Scatterplots for the independent continuous variables plot against the target variable (G3)

C. Barcharts (categorical variables)



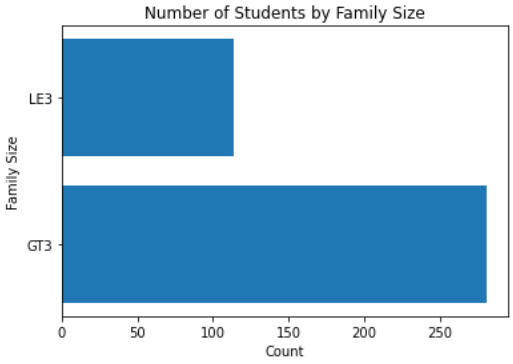


Figure 3: Bar Charts for the Categorical Variables

D. Other Plots

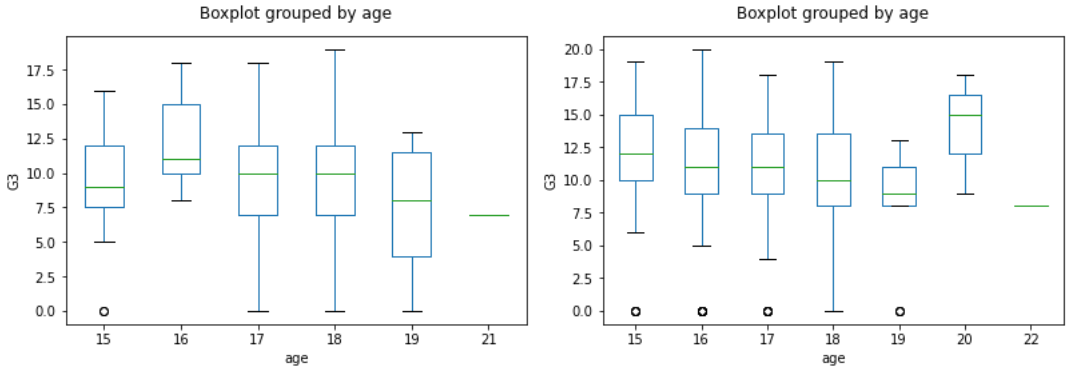


Figure 4: Box Plot Charts of G3 grouped by Age in (a) Rural Address (b) Urban Address

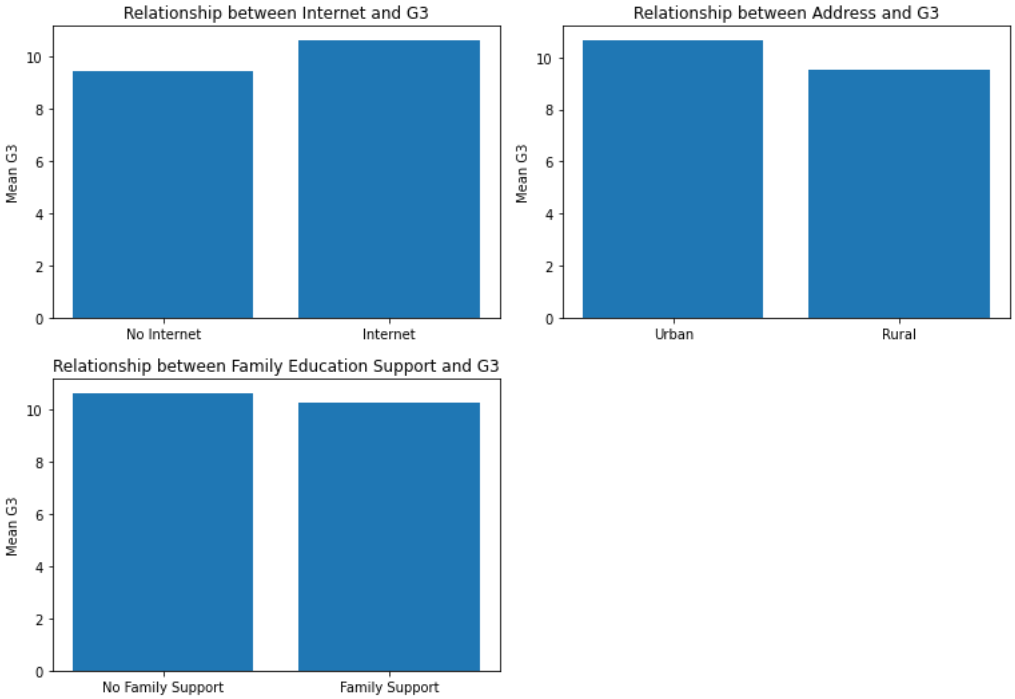


Figure 5: Bar Chart for selected Categorical Variables and its relationship with the Target Variable.

V. SUMMARY OF FINDINGS

This exploratory analysis was conducted on a dataset that contained a grade prediction for the year based on the student's grades in the first two periods of the year, alongside other independent continuous variables and categorical variables that may have had an influence on the student's performance. Firstly, looking at the distributions of the demographical variables showed that majority of the students in the data set were between the ages of 15-19 with a few outliers being older than 19 years of age. Also, the kernel density plots for the distribution of students by Mother's Education and Father's Education both follow a similar shape, perhaps implying that generally both parents had a similar education level. There existed a few students where parents had the lowest levels of education. The travel time density plot suggested that majority of the students were two or less hours away from the school, with a couple outliers of students who had to travel a longer time.

The scatter plots showed the relationships between independent continuous variables and the target variable (G3) which was the final target grade. G1 and G2 both showed a strong, positive correlation with the target variable as they were both used to project the given G3 grade, which is supported by having a correlation of > 0.8 shown in the correlation chart. The relationship between student absences and G3 was interesting as majority of the students recorded very few to none absences throughout the school year; thus, failing to show a true correlation. Another significant variable is Failures and its weak, negative correlation with G3. This relationship is supported by its correlation of -0.36 displayed in the correlation chart. All other variables displayed in Figure 2 didn't seem to have a significant correlation with G3.

The box plot charts in Figure 4 showed a specifically selected relationship between a categorical variable and the target variable. There was a difference in the number of students that lived in rural areas compared to students that lived in urban areas, so the box plot graphs displayed the G3 scores in their respective addresses by age. From analyzing the graph, the box plots of the students in the urban addresses recorded higher scores relative to the box plots of the students in the rural addresses. Lastly, a few selected categorical variables were analyzed to explore its relationships with the target variable in Figure 5. The relationship between the internet and G3 showed that students that had access to the internet recorded a slightly higher mean G3 compared to the students that did not have access to the internet. Also, the relationship between the address and G3 showed that the students that lived in urban areas recorded a slightly higher mean G3 relative to the students that lived in rural areas.