Secondary School Student Performance Analysis

AMS 572 Final Project

Dec 1, 2021

Introduction

For this specific study, high schoolers during the 2005 - 2006 school year from two public schools in the Alentejo region of Portugal were questioned in an effort to gain insight on variables that can predict a student's success in school. What inspired this collection of data and gives it value to analyze is the issue Portugal has with graduation rates among its population. As of 2012, only 28% of Portuguese adults over the age of 30 received a high school diploma, severely lagging behind other Western European countries. With a mix of high dropout rates and failures in high school, Portugal is left to deal with these implications. Education paves the way to improve individual lives by opening possibilities otherwise left unknown, as well as improving society as a whole since better education improves the economy and almost all other factors. That is why it is imperative to invest properly in education and improve the retention rate of schools. Which can be done looking at collected data.

In this data set, 395 high school students students from two different schools in math classes were given a questionnaire that had 37 questions to collect real world data. To accompany these questionnaires, the accompanying students' school reports were also included. This contained data such as the G1 and G2 results which were the grades at the end of 2 periods and then G3 which was the final evaluation grade. For the grades, the school system used a system of 20 points where 20/20 represented a 100% for the grade. The purpose for collecting data sets like this is to be able to see the relations between different variables in a students performance and lives that will impact their success rate in school. For places like Portugal in particular this is key if a better graduation rate is wanted.

The dependent variables that will be used for this study would be the G1, G2, and G3 results. These are the grades throughout the school year at different periods with G3 being the final end of year grade. Therefore the relation between the other variables and these final grades is what is investigated.

Demographic information on the students were collected including: which school they attend, student sex, age, living location (rural or urban), availability of internet at home, reason for choosing their current school, travel time, and current health status. These simple questions can affect a students performance due to maturity or at home tasks related to societal gender roles and expectations.

Furthermore, family information was asked about: family size, if the parents live together or separate, each parents education level along with current occupation, the students legal guardian, and the quality of relations with their family. Family is a known factor in students' success rates as the background plays a huge role in the opportunities a child has. Lastly, information on the student outside and during school was collected on: study time per week, number of past class failures, if receiving extra education support both in and out of school, participating in extracurricular activities, attended nursery school, in a romantic relationship, amount of free time after school, time spent going out with friends, workday and weekend alcohol consumption, and number of school absences.

For the dataset of interest and to be analyzed in this study, there are a total of 395 observations of students from both schools that participated in the math course. Between the two types of data collection methods (student grade reports and questionnaires), there are a total of 30 independent variables in addition to G1, G2 and G3. Three are continuous variables which are: students age, number of past failures, and the number

of class absences. All the remaining variables are categorical being either binary yes or no type, or categories ranked in a 1-5 system.

Materials and Methods

In any study, missing data can result in significant change in the results and therefore meaning of the analysis. That is why great lengths are taken to avoid large vacancies and ensure the sample collected is representative of the total population and is distributed correctly. In this analysis, two forms of missing data are investigated to observe what effects they will have on the dataset of students' performance in a math class and to see if the relationship between the different variables tells a different story.

First, missing data not at random (MNAR) exists in the dataset. With this particular dataset, it is observed that some students received zeros for either G2 or G3 or both. Those students tend to have a lower G1 score as comparing to other students. This indicates that some students did in fact drop out of the class and therefore is left with missing data. This scenario is used to analyze the MNAR. By doing this, the relationship between the final grade (G3) and the other variables could be affected due to survival bias and possible better conditions to foster the proficient grades.

Secondly, missing at completely random (MCAR) data points will also be simulated. With this scenario, there is a given proportion of data values that are not present in the set. This is not from any given reason and is randomly distributed throughout a single variable or the entire study. For this particular study, the data set is run through code that randomly imputes the missing data in an identical sample of the original unperturbed one. Then the number of missing values is accounted for to allow for an easier analysis. It is noted that almost all students (the 395 observations), will receive either 0 or 1 missing data in one of the variables they were questioned on.

For the first hypothesis, the equivalency of the G1 and G3 as well as G2 and G3 grades were to be tested. By doing so, we would be able to see if there is a change in the overall average students performance and class distribution from the first period (G1) to the final (G3). It is expected that the two grades are in fact not significantly different and that the performance of the student by G1 will result in the same pass/fails as the final G3 reports. This can serve as a tool to make early predictions within the classroom and to adapt the focus on certain students or teaching styles early on in the school year. For this analysis, a paired t-test is selected as the best fit for this hypothesis.

In addition to investigating the relationship between G1 and G3 average grades, the second hypothesis will look into what factors in a students academic and personal life will affect their final G3 grade in the math class. These results will be collected using the method of multiple linear regression. While a student's final performance may appear to be solely dependent on their other grades and perhaps even by chance, many factors can be at play that dictate how a student will end up. Life is never as simple as moving from one grade to another and in fact has many variables in circulation. That is where this study focused the majority of its time. With the combination of the student reports and personal questionnaires, its goal is to highlight any possible connections. This will allow the school and larger entities to better predict how a student body will perform and where to focus on improvements. By conducting linear regressions on the variables connected to the G3 grade, the ones showing the highest correlation will be investigated further to compute a final equation that can be used to propose an accurate determination of a students final grade with the input to a select few variables.

Finally, the last hypothesis is to analyze what factors will affect an improvement in students' score the most. Specifically, what is the main driving force to lead to a decrease in performance in final grade as comparing to the first examination. We used the logistic regression model to examine the data and results will be presented in the last section.

Summary Statistics

We first generated summary statistics for all variables. The goal is to describe the main features of numerical and categorical information with simple summaries. For categorical variables, we calculated the frequency

and proportions of each category. The P-value was calculated for each categorical variable using one-way ANOVA test to compare the means of G3 at each category. Note that ANOVA may not be suitable for some variables where the number of observations of some categories is very few (less than 5). The main goal to present P-values in this table is to provide a brief understanding on relationships between different features and G3.

For continuous variables, we presented the minimum, maximum, mean, and standard deviations of each variables.

Table 1 shows the summary statistics of all categorical variables. Most students in our dataset are from an urban setting (78%). 71% students come from a family with more than 3 people. 33% of the students' mothers have received education higher than 9th grade and 24% of students' fathers received a higher education. 50% of students study 2-5 hours every week while only 16% students study more than 5 hours per week. 13% students take extra educational school support and 61% students have extra family educational support. 95% students are planning to apply for a higher degree beyond high school and those who do not plan to go towards a higher degree has substantially lower average final grade than other students.

Deal with Missing Values

In this dataset, there are missing values in father's education and mother's education, which are labeled as 'None' in the original dataset. Students tend to have a higher final grade for those who are missing parents' education. Considering there are only 3 students missing parents' education, we will leave as what it is.

There are also missing values in G2 and G3. Here, the missing entries are considered 0's for the students who did not show for G2 exam, G3 exam, or both.

When missing values are entered in a data set, it could make the data biased. One benefit to using imputation is that we could hypothesize the data entry for any missing entries. Two kinds of missing values that we will use are MCAR and MNAR. For better organization, we transform the categorical data to have binary values. The imputing data functions will not work with categorical data.

MNAR (Missing Not at Random)

We suspecte that missing data in G2 and G3 are not at random. When a data set has MNAR, there is an external factor that is affecting the missing entries. Since our data set involves a third grade that could be influenced by the student's progress in the course, it is possible that some students withdrew from the course due to lack of performance in the class.

We plot histograms in Figure 1 and Figure 2 to show the distribution of first exam grade (G1) for who received a zero for the second exam (G2) and final exam (G3). Since most of the grades are mostly in the lower half of the distribution (less than mean of G1: 10.91), this supports our assumption that the data set includes MNAR. In addition, it is important to note that when there is a 0 in the G2 column, there is also a 0 in the G3 column. This dismisses any thought that the student may have missed G2 but is still enrolled in the class.

In order to simulate missing exam scores, we used a "for" loop to enter "NA" for the entries that are "0" within the columns G2 and G3.

```
## [1] "There are 0 missing values in G1"

## [1] "There are 13 missing values in G2"

## [1] "There are 38 missing values in G3"

## [1] "There are 13 both missing G3 and G2"
```

Table 1: Frequency Distribution of Categorical Variables

Variables	Frequency	Percentage	G3 (Mean)	P-value (G3)
School_gp	349	0.88	10.49	0.372
School_ms	46	0.12	9.85	NA
Sex_f	208	0.53	9.97	0.040
Sex_m	187	0.47	10.91	NA
Address_rural	88	0.22	9.51	0.036
Address_urban	307	0.78	10.67	NA
Famsize_gt3	281	0.71	10.18	0.106
Famsize_le3	114	0.29	11.00	NA
Pstatus_apart	41	0.10	11.20	0.250
Pstatus_livetogether	354	0.90	10.32	NA
Medu_<=4th grade	59	0.15	8.68	0.000
Medu_5th to 9th grade	103	0.26	9.73	NA
Medu_higher education	131	0.33	11.76	NA
Medu_none	3	0.01	13.00	NA
Medu_secondary education	99	0.25	10.30	NA
Fedu_<=4th grade	82	0.21	9.16	0.022
Fedu_5th to 9th grade	115	0.29	10.26	NA
Fedu_higher education	96	0.24	11.36	NA
Fedu_none	2	0.01	13.00	NA
Fedu_secondary education	100	0.25	10.66	NA
Mjob_at_home	59	0.15	9.15	0.005
Mjob_health	34	0.09	12.15	NA
Mjob_other	141	0.36	9.82	NA
Mjob_services	103	0.26	11.02	NA
Mjob_teacher	58	0.15	11.05	NA
Fjob_at_home	20	0.05	10.15	0.268
Fjob_health	18	0.05	11.61	NA
Fjob_other	217	0.55	10.19	NA
Fjob_services	111	0.28	10.30	NA
Fjob_teacher	29	0.07	11.97	NA
Reason_course	145	0.37	9.82	0.102
Reason_home	109	0.28	10.26	NA
Reason_other	36	0.09	11.17	NA
Reason_reputation	105	0.27	11.14	NA
Guardian_father	90	0.23	10.69	0.205
Guardian_mother	273	0.69	10.48	NA
Guardian_other	32	0.08	9.06	NA
Traveltime_<15 min	257	0.65	10.78	0.139
Traveltime_>1 hour	8	0.02	8.75	NA
Traveltime_15-30 min	107	0.27	9.91	NA
Traveltime_30-60 min	23	0.06	9.26	NA
Studytime_<2 hours	105	0.27	10.05	0.161
Studytime_>10 hours	27	0.07	11.26	NA
Studytime_2-5 hours	198	0.50	10.17	NA
Studytime_5-10 hours	65	0.16	11.40	NA
Schoolsup_no	344	0.87	10.56	0.100
Schoolsup_yes	51	0.13	9.43	NA
Famsup_no	153	0.39	10.64	0.438
Famsup_yes	242	0.61	10.27	NA
Paid_no	214	0.54	9.99	0.043
Paid_yes	181	0.46	10.92	NA
Activities_no	194	0.49	10.34	0.750
Activities_yes	201	0.51	10.49	NA
Nursery_no	81	0.21	9.95	0.307
Nursery_yes	314	0.79	10.54	NA
Higher_no	20	0.05	6.80	0.000
Higher_yes	375	0.95	10.61	NA
Internet_no	66	0.17	9.41	0.050
Internet_yes	329	0.83	10.62	NA
Romantic_no	263	0.67	10.84	0.010
Romantic_yes	132	0.33	9.58	NA

Table 2: Summary Statistics of Numerical Variables

Variables	Min	Max	Mean	SD
Age	15	22	16.70	1.28
Failures	0	3	0.33	0.74
Famrel	1	5	3.94	0.90
Freetime	1	5	3.24	1.00
Goout	1	5	3.11	1.11
Dalc	1	5	1.48	0.89
Walc	1	5	2.29	1.29
Health	1	5	3.55	1.39
Absences	0	75	5.71	8.00
G1	3	19	10.91	3.32
G2	0	19	10.71	3.76
G3	0	20	10.42	4.58

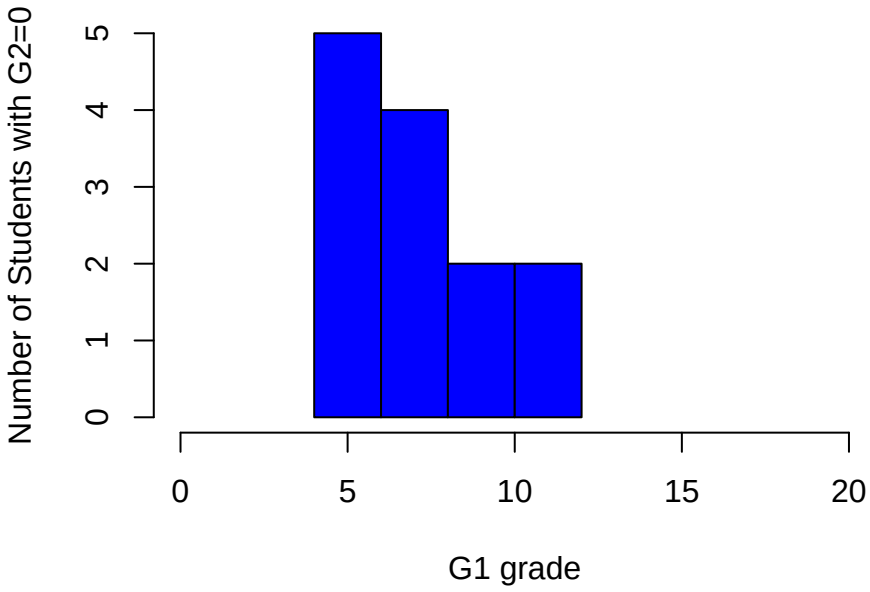

Figure 1: Distribution of G1 grade for students whose G2 equals 0

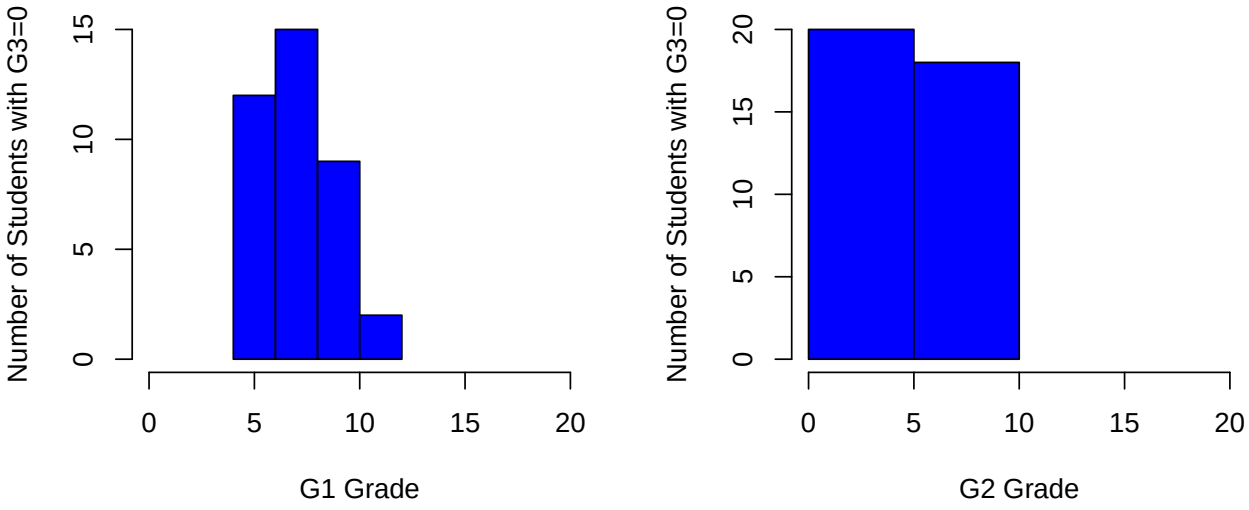

Figure 2: Distribution of G1 and G2 for students whose G3 equals 0

Imputing Missing Data

We used the package MICE (Multivariate Imputation by Chained Equations) to impute the missing data. By default, mice() calculates five (m = 5) imputed data sets. We use the final imputation for our final analysis. According to the summary data, the minimum values for G2 and G3 are all greater than 0 now.

```
##        G1              G2              G3
##  Min.   : 3.00   Min.   : 4.00   Min.   : 4.00
##  1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 9.00
##  Median :11.00   Median :11.00   Median :11.00
##  Mean   :10.91   Mean   :10.97   Mean   :11.15
##  3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :19.00   Max.   :20.00
```

MCAR (Missing Completely at Random)

When missing values are MCAR, the probability of having a missing value in each category is equal. The MICE (Multivariate Imputation by Chained Equations) package contains the function "ampute" that is used to randomly assign values in our dataset. This is done by specifying that the mechanism is MCAR in one of the parameters. Missing values are then introduced manually to the dataset from MNAR section, where missing values in G2 and G3 in the orginal dataset were imputed by multiple imputation.

Create data frame containing the number of random missing values for each column. Table 3 summarizes number of missing values for each columns that were introduced.

To deal with MCAR, if there are only a few missing data point (less than 10%), we can directly remove those records with missing values from our analysis. In this work, as missing values for each column that we introduced are less than 10%, when performing paired t-test to evaluate mean values between G1, G2 and G3 in the next section. We directly drop those records with missing values.

However, note that if there are large number of records with missing values, we may need to use mice package to impute the missing data or impute missing values with mean/mode depending on data types.

```
## Warning: Data is made numeric because the calculation of weights requires
## numeric data
```

Table 3: Summary of missing values for each columns

Features	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason
Nmissing	11	3	6	5	3	3	9	4	6	4	1
Features	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet
Nmissing	11	7	10	8	8	8	2	6	2	3	4
Features	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
Nmissing	5	6	4	7	5	7	3	3	6	7	6

Data Visualization

For this study, we are interested in what are the factors that will affect final score. According to Table 1, among all categorical variables, there are statistically significant difference in G3 between different categories for following features (p-value from ANOVA test < 0.05):

- Mother's education
- Father's education
- Mother's job
- Extra paid classes
- Willingness to take higher education
- In a romantic relationship

To visualize the variability of G3 within each group, we presented box plots for each of those variables in Figure 1 using data imputed from MNAR. The box plot shows that parents with a higher degree and the students who are willing to pursue a higher degree will tend to have a higher final grades. Students with a romantic relationship have lower final grades than other students.

Table 2 presents summary statistics of all continuous and ordinal variables. Family relationship, free-time, going out, workday alcohol, weekend alcohol, and health are ordinal variables with 1 representing the minimum and 5 representing the maximum.

According to Table 2, students' age varies from 15 to 22 with 0 to 3 failures. On average, final grades $G3$ is lower than $G1$ and $G2$. $G1, G2, G3$ range from 0 to 20 with a mean around 10. A score of 0 may be due to dropping out of the class or missing the exams. To visualize the distribution of scores at each period, Figure 2 shows the histograms of each grade. The Shapiro-Wilk normality test shows that the scores deviate significantly from normality. However, it is (a bit strongly stated) a fact that formal normality tests always reject on the huge sample sizes we work with today. Histograms shows that the grades follow bell-shaped distribution with the average score around 10.

```
##
##  Shapiro-Wilk normality test
##
## data:  df_MNAR$G3
## W = 0.97958, p-value = 2.256e-05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_MNAR$G2
## W = 0.97988, p-value = 2.624e-05
```

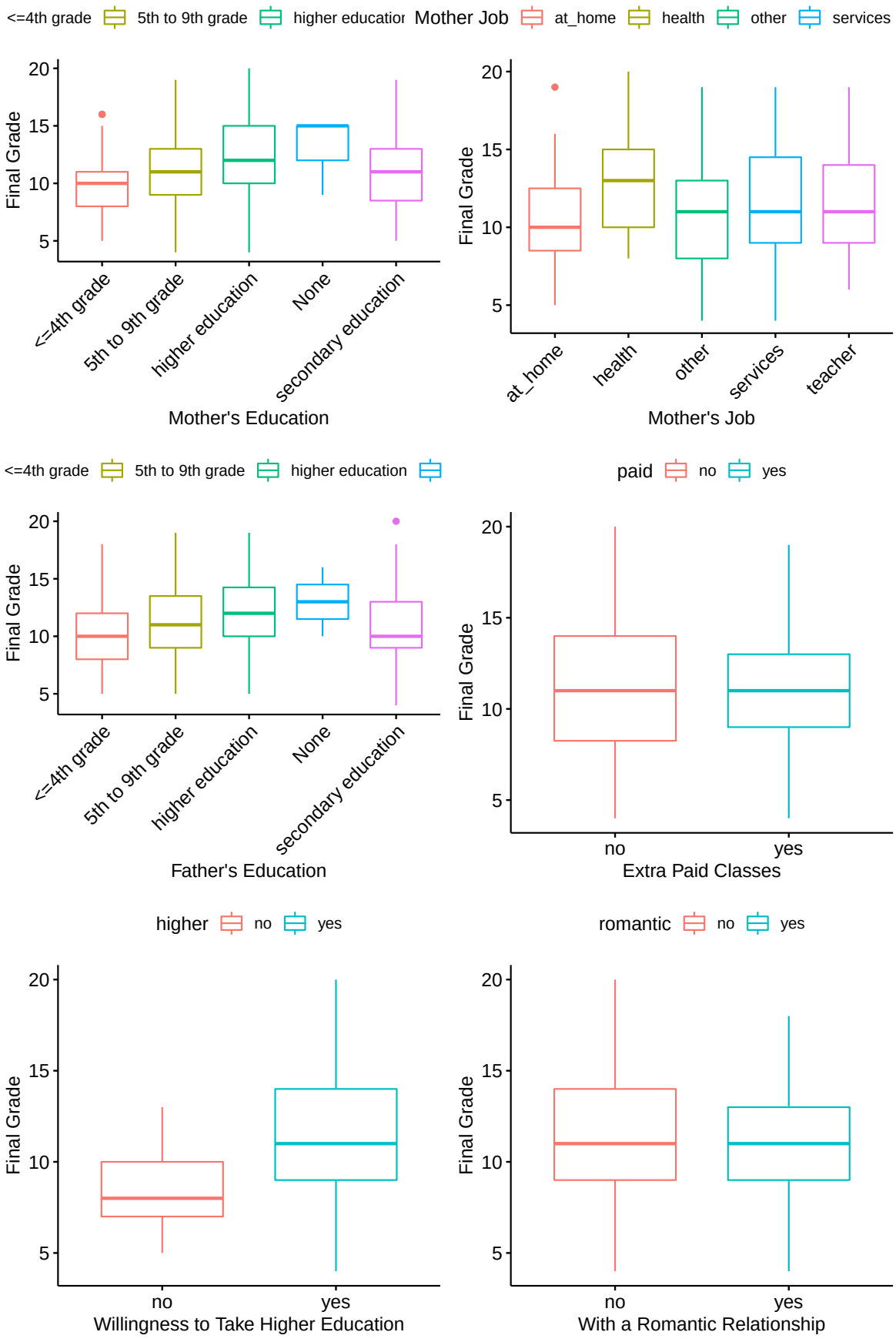

Figure 3: Boxplots for G3 by selected variables.

```
##
##  Shapiro-Wilk normality test
##
## data:  df_MNAR$G1
## W = 0.97491, p-value = 2.454e-06
```

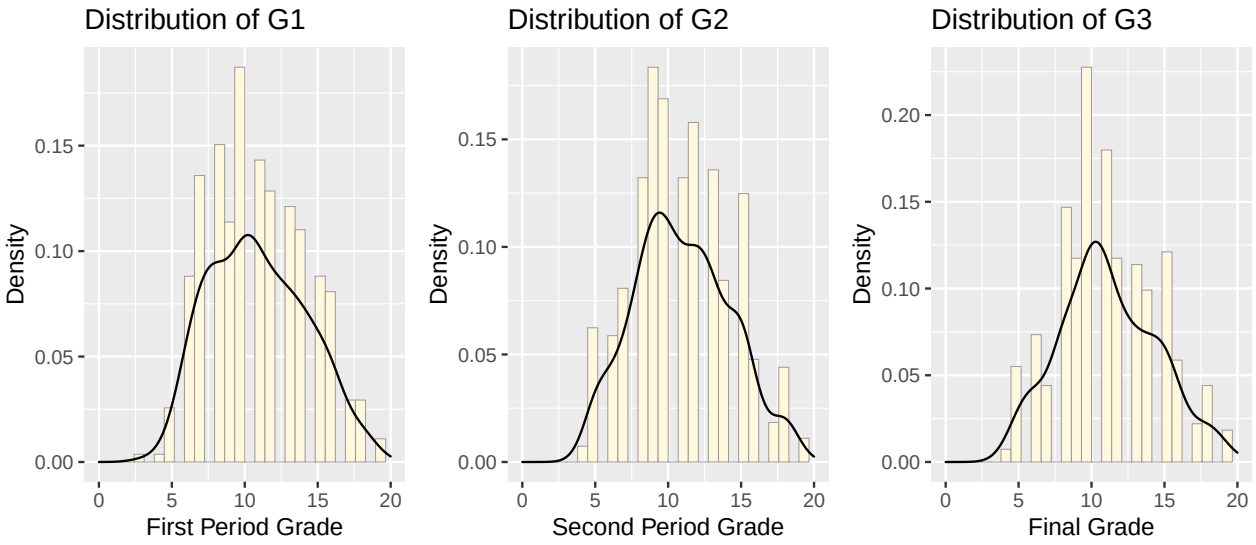

Figure 4: Distribution of grades at each period.

Paired T-test to Compare Grades at Different Periods

Each student takes exams at three different periods within the school year. We are interested in whether the mean score at each period is equal to each other. In order to evaluate whether the score differences are equal to zero, we propose two hypothesis:

Hypothesis 1:

$$H_0 : \mu_{G3} = \mu_{G1}$$
$$H_a : \mu_{G3} \neq \mu_{G1}$$

Hypothesis 2:

$$H_0 : \mu_{G3} = \mu_{G2}$$
$$H_a : \mu_{G3} \neq \mu_{G2}$$

We perform two paired t-tests to compare means of G1 versus G3 and G2 versus G3.

To apply the paired t-test to test for differences between paired measurements, the following assumptions need to hold:

- **Assumption 1: Are the two samples paired?**

Yes, since the data have been collected from recording a student's score at different period.

- **Assumption 2: Are the means normally distributed?**

Yes, the sample size is 395, which is substantially greater than 30. Even though the scores don't follow normal distribution according to Shapiro-Wilk normality test, according to Central Limit Theory, means of samples from a population with finite variance approach a normal distribution regardless of the distribution of the population.

Thus, the paired t-test is a valid test to check the two hypothesis. We first applied paired t-test to the imputed dataset, where 0's in G2 and G3 are imputed using multiple imputation.

According to the paired t-test result, we reject the hypothesis that $H_0 : \mu_{G3} = \mu_{G1}$ (p-value = 0.001) and $H_0 : \mu_{G3} = \mu_{G2}$ (p-value = 4.36e-05). By plotting histograms of the score difference in Figure 4, we see that both $G3 - G2$ and $G3 - G1$ are centralized around 0. The difference between G3 and G1 have mean of 0.248 with 95% CI of (0.1 0.39) and The difference between G3 and G1 have mean of 0.18 with 95% CI of (0.094 0.265).

```
##
##  Paired t-test
##
## data:  df_MNAR$G3 and df_MNAR$G1
## t = 3.1475, df = 394, p-value = 0.001772
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0893324 0.3866170
## sample estimates:
## mean of the differences
##               0.2379747
```

```
##
##  Paired t-test
##
## data:  df_MNAR$G3 and df_MNAR$G2
## t = 4.1767, df = 394, p-value = 3.645e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0951385 0.2643552
## sample estimates:
## mean of the differences
##               0.1797468
```

We have also performed paired t-tests with the original dataset, where those who didn't attend exams have G2 and G3 as 0's.

Consistent with the result from MNAR dataset, we reject the hypothesis that $H_0 : \mu_{G3} = \mu_{G1}$ (p-value $<$ 2.2e-16) and $H_0 : \mu_{G3} = \mu_{G2}$ (p-value = 0.002994). However, opposite to a slight increment in final grade from G2 and G1 in MNAR dataset, the final grade decreases both from G1 ($G3 - G1$ Mean (95% CI): -0.49 (-0.76, -0.22)) and from G2 ($G3 - G2$ Mean (95% CI): -0.3 (-0.49, -0.10)). The decrease is caused by missing values in G3 and G2.

Also, by plotting the histogram of score difference in Figure 5, we can see that both $G3 - G2$ and $G3 - G1$ are left skewed where some students' final grades decreased substantially from first and second periods.

```
##
##  Paired t-test
##
## data:  df$G3 and df$G1
## t = -3.5517, df = 394, p-value = 0.0004291
```

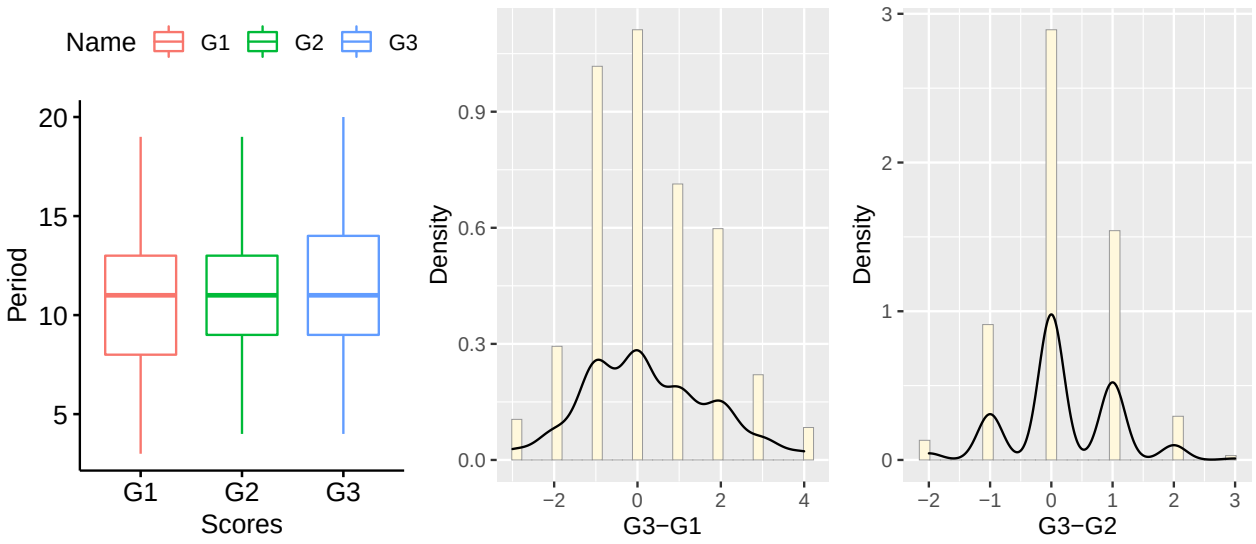

Figure 5: Distribution of G3-G1 and G3-G2 (MNAR dataset).

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7669366 -0.2204052
## sample estimates:
## mean of the differences
##               -0.4936709

##
##  Paired t-test
##
## data:  df$G3 and df$G2
## t = -2.9869, df = 394, p-value = 0.002994
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4953629 -0.1021055
## sample estimates:
## mean of the differences
##               -0.2987342
```

We have also performed a paired t-test to the dataset where we introduced complete random missing values to the MNAR dataset in the last section MCAR part. Missing values are ignored in paired t-tests. The paired t-test results are very close to that with MNAR dataset.

Conclusion

Above all, when the missing values are complete random and very few (<10%), dropping missing data won't introduce much bias in hypothesis testing we performed. However, when data is missing not completely due to randomness, if we impute all missing data using 0 as in the original data did, the test statistics will misleading. Data imputation is very essential especially when it is MNAR and there are large number of missing values in the dataset.

```
## [1] "There are 6 missing values in G1 (mcar)"

## [1] "There are 7 missing values in G2 (mcar)"
```

```
## [1] "There are 6 missing values in G3 (mcar)"
```

```
##
##  Paired t-test
##
## data:  mcar$G3 and mcar$G1
## t = 3.0742, df = 382, p-value = 0.002262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.08469336 0.38528053
## sample estimates:
## mean of the differences
##               0.2349869
```

```
##
##  Paired t-test
##
## data:  mcar$G3 and mcar$G2
## t = 4.3406, df = 381, p-value = 1.823e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1031027 0.2738607
## sample estimates:
## mean of the differences
##               0.1884817
```

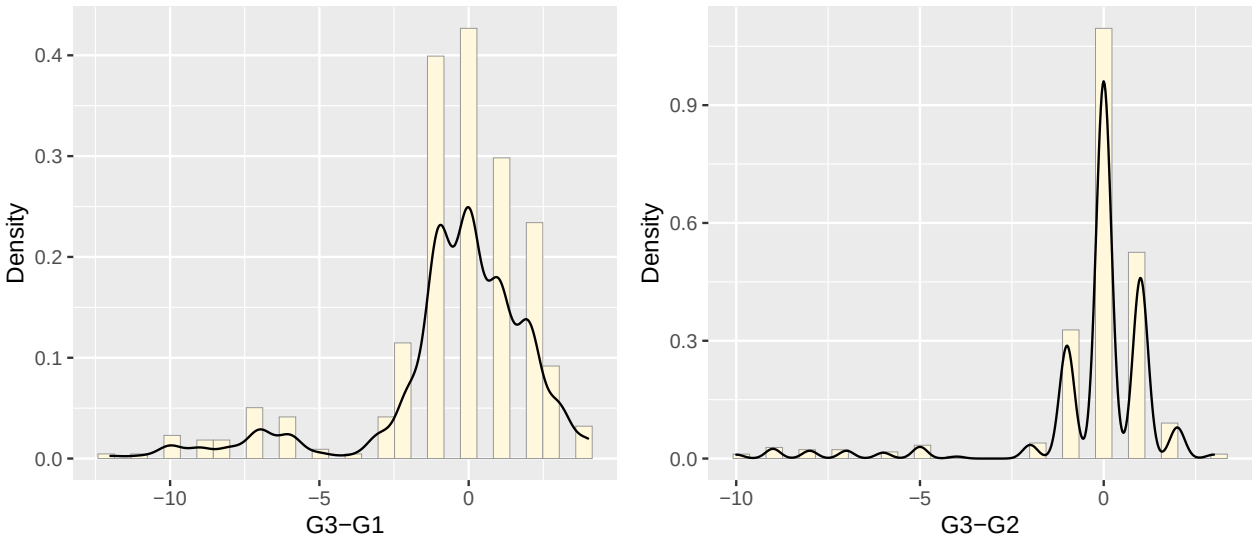

Figure 6: Distribution of G3-G1 and G3-G2 (original dataset).

Multiple Linear Regression

Paired Data Correlation

Paired data correlation is a way to understand the relationship between multiple variables and attributes. Using correlation, we can obtain insights whether one or multiple attributes are associated with each other.

Figure 6 shows the correlation plot with correlation coefficients for significantly correlated variables. The correlation coefficients shown are calculated using Spearman's rank-order correlation method since most of the continuous variables are ordinal.

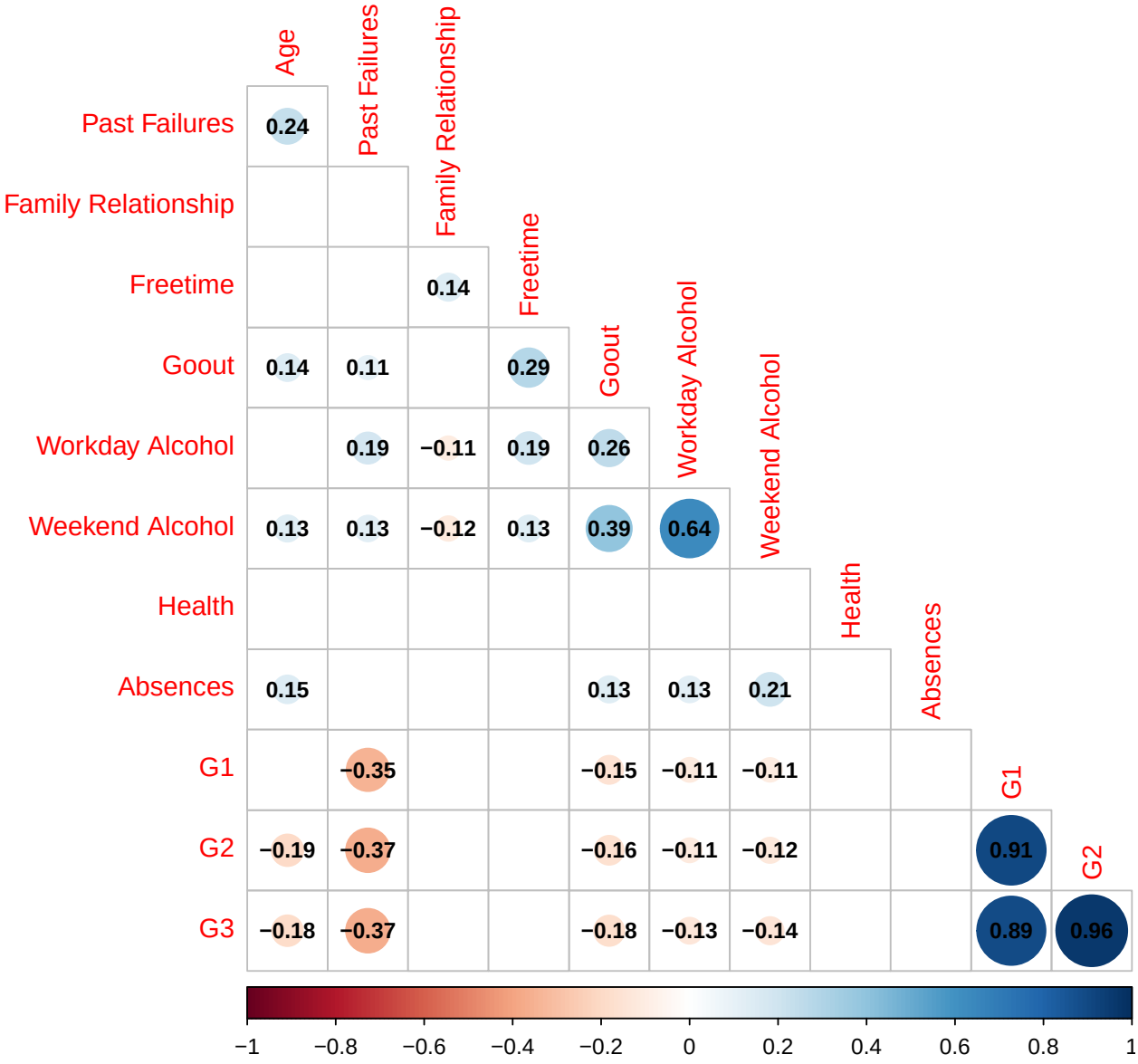

Figure 7: Correlation plot for all numerical features.

Interpretation of the correlation plot According to the correlation plot, G3 are highly correlated with G1 and G2. Number of past failures, age, alcohol consumption, and frequency going out with friends are negatively correlated with final grades. Alcohol consumption is statistically correlated with almost all other variables. Specifically, alcohol consumption is positively correlated with age, free time, past failures and frequency of hang out while negatively correlated with family relationship. Especially, number of past failures has shown highest correlation with G1, G2 and G3.

Mutiple Linear Regression

In this section, we are going to focus on using multiple linear regression to estimate final grades. We are going to perform hypothesis testing to figure out whether number of past failures are correlated with G3 adjusting by the effect of other features. The H_0 is: $\beta_{failure} = 1$ versus H_1: $\beta_{failure} \neq 0$.

There are 32 predictors available in our dataset. Since G1 and G2 are highly correlated with G3, we exclude them from our model and to find out how other predictors correlate with G3. The question we are going to answer is that whether students will perform differently in their final exam if they have different number of past failures.

We first build a linear regression with all 30 predictors using imputed MNAR dataset. Figure 7 shows residual plot and QQ plots of the fitting.

```
##
## Call:
## lm(formula = G3 ~ ., data = df_lr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8704 -1.8387 -0.1429  1.8256  6.8499
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        16.223317   3.111942   5.213 3.20e-07 ***
## schoolMS                           -0.218586   0.566130  -0.386   0.6997
## sexM                                0.896814   0.358707   2.500   0.0129 *
## age                                -0.275764   0.153747  -1.794   0.0738 .
## addressUrban                        0.388297   0.414332   0.937   0.3493
## famsizeLE3                          0.583437   0.349246   1.671   0.0957 .
## PstatusLiveTogether                -0.238977   0.510493  -0.468   0.6400
## `Mother Education`5th to 9th grade  0.343862   0.542870   0.633   0.5269
## `Mother Education`higher education  1.138404   0.740930   1.536   0.1253
## `Mother Education`None              2.898450   1.780729   1.628   0.1045
## `Mother Education`secondary education  0.661999   0.602934   1.098   0.2730
## `Father Education`5th to 9th grade -0.009747   0.463659  -0.021   0.9832
## `Father Education`higher education  0.193795   0.620771   0.312   0.7551
## `Father Education`None              0.708659   2.163668   0.328   0.7435
## `Father Education`secondary education -0.239004   0.535976  -0.446   0.6559
## `Mother Job`health                  1.094388   0.795914   1.375   0.1700
## `Mother Job`other                  -0.406454   0.502498  -0.809   0.4191
## `Mother Job`services                0.704806   0.564816   1.248   0.2129
## `Mother Job`teacher                -0.936133   0.746785  -1.254   0.2109
## `Father Job`health                 -0.278826   1.023679  -0.272   0.7855
## `Father Job`other                  -0.433920   0.730159  -0.594   0.5527
## `Father Job`services               -0.248609   0.756269  -0.329   0.7426
## `Father Job`teacher                 1.292871   0.935899   1.381   0.1680
## reasonhome                          0.290766   0.391338   0.743   0.4580
## reasonother                         0.118075   0.576479   0.205   0.8378
## reasonreputation                    0.351067   0.407791   0.861   0.3899
## guardianmother                     -0.090996   0.385256  -0.236   0.8134
## guardianother                       0.681217   0.704078   0.968   0.3340
## traveltime>1 hour                   0.509856   1.126845   0.452   0.6512
## traveltime15-30 min                -0.377961   0.362789  -1.042   0.2982
## traveltime30-60 min                 0.426668   0.699960   0.610   0.5426
```

```
## studytime>10 hours                    1.297224   0.704489   1.841   0.0664 .
## studytime2-5 hours                    0.106889   0.387399   0.276   0.7828
## studytime5-10 hours                   1.339868   0.534891   2.505   0.0127 *
## `Past Failures`                      -1.253744   0.234467  -5.347 1.63e-07 ***
## schoolsupyes                         -1.908275   0.470845  -4.053 6.26e-05 ***
## famsupyes                            -0.719770   0.337382  -2.133   0.0336 *
## paidyes                              -0.098521   0.340067  -0.290   0.7722
## activitiesyes                        -0.106099   0.314153  -0.338   0.7358
## nurseryyes                           -0.295857   0.388433  -0.762   0.4468
## higheryes                             0.804255   0.757900   1.061   0.2894
## internetyes                           0.377169   0.439356   0.858   0.3912
## romanticyes                          -0.362806   0.337025  -1.076   0.2825
## `Family Relationship`                 0.140300   0.175203   0.801   0.4238
## freetime                              0.120780   0.167349   0.722   0.4710
## goout                                -0.362556   0.158253  -2.291   0.0226 *
## `Workday alcohol`                    -0.084573   0.234756  -0.360   0.7189
## `Weekend alcohol`                    -0.090306   0.175957  -0.513   0.6081
## health                               -0.259267   0.114748  -2.259   0.0245 *
## absences                             -0.015858   0.020427  -0.776   0.4381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.869 on 345 degrees of freedom
## Multiple R-squared:  0.3553, Adjusted R-squared:  0.2637
## F-statistic:  3.88 on 49 and 345 DF,  p-value: 6.817e-14
```

Interpretation of summary table and plot

Adjusted R-square for the linear regression model is 0.26. Note that the R-square value is very low because we didn't include G1 and G2 in our model. According to the linear regression result, gender, study time, past failures, family support, school support, hanging out with friends as well as health condition are features significantly correlated with final grades, which is consistent with the paired correlation analysis. Specifically, males perform better than females. The more study time a student spends, the higher final grade they will obtain. Past failures, family support, school support, hanging out with friends as well as health condition are all negatively correlated with final grade.

We also presented residual plot and Q-Q plot in Figure 7 to check whether the data satisfy the homoscedasticity and normality assumptions of linear regression. Residual graphs allow use to check if your data shows homoscedastic, which is when residuals are equal across all values of your predicted variable. According to Figure 7, we see that the red line is pretty flat with small fluctuation and it is very close to the dashed line, which indicates that the data satisfy the homoscedasticity of linear regression.Also according to the Q-Q plot, our data form an approximate straight line and is approximately close to the dashed line (identity line), which indicates that the data meets the normality hypothesis.

Stepwise feature selection

One potential issue of the multiple linear regression we built is that we have included a lot features that are not correlated with the target variable G3. As a results, there is a high tendency to overfitting the data. Thus, we decide to the use stepwise regression to perform feature selection to find out the best model.

Stepwise regression can serve as a hypothesis generation tool, giving an indication of how many variables are likely to be useful, and identifying variables as strong candidates for predictive models. Also it involves adding or removing potential explanatory variables and testing for statistical significance after each iteration. The main goal of stepwise regression is to build the best model that gives us the predictive variable which has the largest variance in the outcome variable (adjusted r^2 or Cp or BIC). Figure 8 shows the relationship

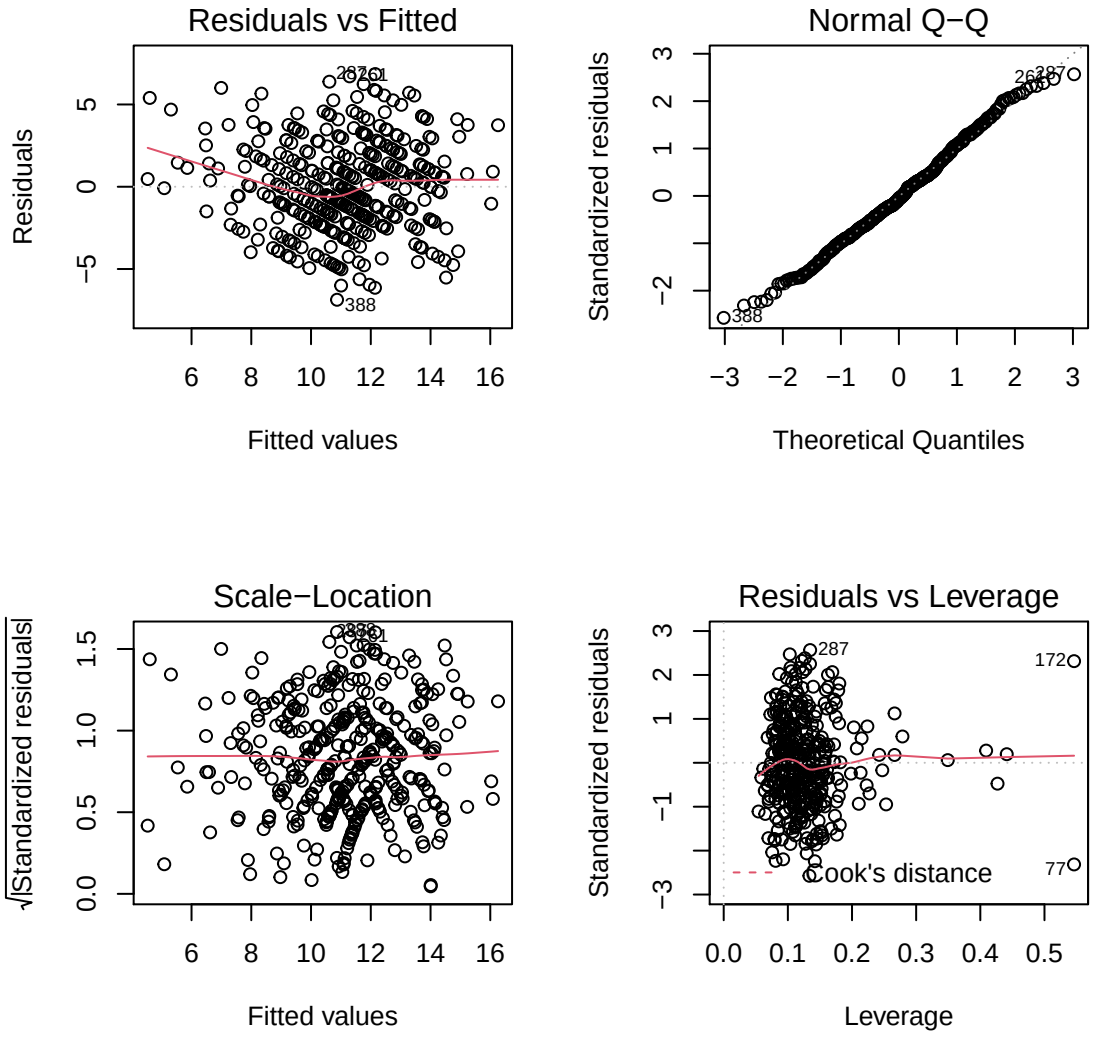

Figure 8: Multiple linear regression results

of Cp, BIC and Adjusted R-squared with number of features included based upon stepwise feature selection. Bayesian information criterion (BIC) is a criterion for model selection in a finite set of models. A lower BIC indicates a lower penalty term and is therefore a better model. R-squared computes the scatter of data points around the regression line. It is also known as the determination coefficient, or multiple determination coefficient of multiple regression. The larger R-squared, the better the regression model fits the data.

According to the plot, we choose number of features with lowest BIC and a pretty high R-squared and in the final model, we include 10 features as: sex, age, Mother Education, Mother Job, Father Job, studytime, Past Failures, schoolsup, health and goout.

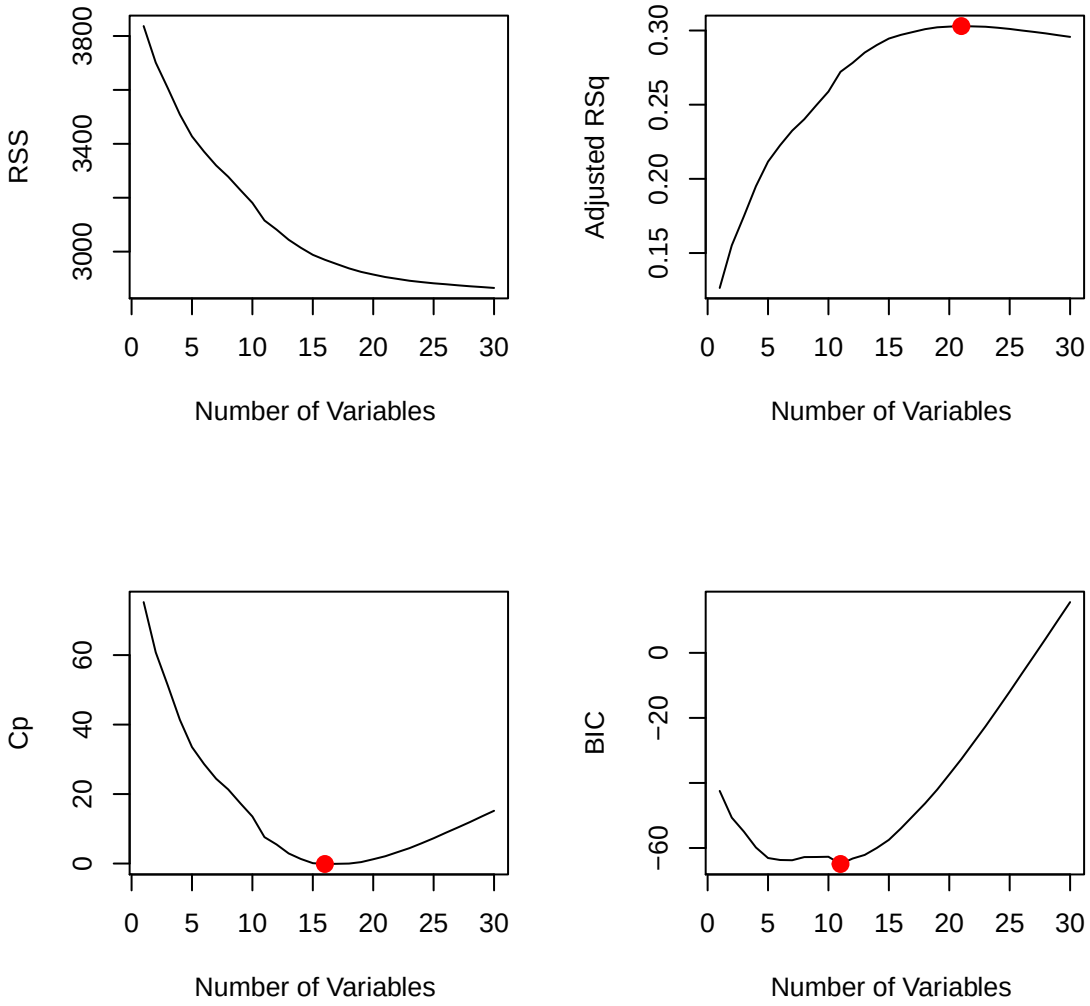

Figure 9: Ajudsted R square, Cp and BIC with number of variables according to forward feature selection.

The final linear regression model has an adjusted R-squared of 0.2724, which is slightly higher than our baseline model. Among all features, past failures shows highest correlation with final grade, that the more past failures, the lower final grade the student will have. We also noticed that school support and family support actually won't be able to help a student to achieve a higher grade. Male performs better than female in their math scores.

We have also built a linear regression model using the original data where missing scores are represented as 0's (due to limitation of report length, we didn't include the results here). The linear regression results also show that past failure is significantly negatively correlated with final grades, which is consistent with the results using imputed dataset.

Overall, we reject the H_0 that $\beta_{failure} = 0$.

```
##
## Call:
## lm(formula = I(G3) ~ (sex + age + `Mother Education` + `Mother Job` +
##     `Father Job` + studytime + `Past Failures` + schoolsup +
##     health + goout), data = df_lr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.199 -1.884 -0.272  1.849  6.885
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         18.08476    2.34357   7.717 1.10e-13 ***
## sexM                                 0.99696    0.32131   3.103  0.00206 **
## age                                 -0.31173    0.12455  -2.503  0.01275 *
## `Mother Education`5th to 9th grade   0.15483    0.49069   0.316  0.75253
## `Mother Education`higher education   0.75832    0.60052   1.263  0.20746
## `Mother Education`None               3.30728    1.70969   1.934  0.05382 .
## `Mother Education`secondary education  0.39493    0.51316   0.770  0.44202
## `Mother Job`health                   1.52579    0.73678   2.071  0.03906 *
## `Mother Job`other                   -0.09462    0.47015  -0.201  0.84061
## `Mother Job`services                 1.14983    0.52655   2.184  0.02961 *
## `Mother Job`teacher                 -0.50017    0.69472  -0.720  0.47200
## `Father Job`health                  -0.34822    0.96752  -0.360  0.71912
## `Father Job`other                   -0.61210    0.68817  -0.889  0.37433
## `Father Job`services                -0.50504    0.71313  -0.708  0.47926
## `Father Job`teacher                  1.11815    0.86375   1.295  0.19628
## studytime>10 hours                   1.29303    0.64385   2.008  0.04534 *
## studytime2-5 hours                   0.15324    0.36358   0.421  0.67365
## studytime5-10 hours                  1.36119    0.49285   2.762  0.00603 **
## `Past Failures`                     -1.30133    0.21280  -6.115 2.43e-09 ***
## schoolsupyes                        -1.98335    0.45822  -4.328 1.93e-05 ***
## health                              -0.30786    0.10791  -2.853  0.00457 **
## goout                               -0.36575    0.13310  -2.748  0.00629 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.852 on 373 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.2724
## F-statistic: 8.023 on 21 and 373 DF,  p-value: < 2.2e-16
```

Logistic Regression to Predict Sign of $G3 - G1$.

In addition to a student's performance in their final grade, we are also interested in which factors relate to an improvement in scores from first period to final exam the most. Among all students, 160 of them have a higher final grade than G1 while 235 of them have a lower or same score as G1.

Feature Selection

In this section, we perform logistic regression to predict sign of $G3 - G1$. The sign of $G3 - G1$ is represented as a indicator variable where its value is 1 if $G3 - G1 > 0$. RFE (recursive feature elimination) is then applied to all features except for G2 and G3 for feature selection before modeling.

RFE applies a backward selection process to find the optimal combination of features. First, it builds a model based on all features and calculates the importance of each feature in the model. Then, it rank-orders the features and removes the one(s) with the least importance iteratively based on accuracy. Feature importance can be computed based on random forest importance criterion.

```
## [1] "154 students have a higher score in G3 as comparing to G1"
```

```
## [1] "241 students have a lower or similar score in G3 as comparing to G1"
```

```
## [1] "Following variables are selected according to RFE:"
```

```
##  [1] "absences"            "age"                 "G1"
##  [4] "Weekend alcohol"     "school"              "address"
##  [7] "traveltime"          "schoolsup"           "famsup"
## [10] "romantic"            "higher"              "health"
## [13] "activities"          "guardian"            "Family Relationship"
## [16] "famsize"             "reason"              "Pstatus"
```

The output indicates that RFE recommends 18 features to be included in the model. Accuracy reach the maximum level when 18 features are retained in the model as shown in Figure 9. According to the feature importance generated from RFE model, absence, G1, school and alcohol correlate with improvement in score the most. We include all 18 features recommended and build a logistic regression model to find out their relationship.

Final Logistic Regression Model

Logistic regression shows that absences and G1 correlate with improvement in score the most. Specifically, more absence from school, higher G1 and being in a relationship will lead to a lower score in final grade as compared to first period. Thus, in order to achieve a higher score in his/her final grade, a student should try to ensure as few absences from class as possible.

```
##
## Call:
## glm(formula = ind ~ ., family = "binomial", data = simdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -0.9286  -0.5521   1.0568   2.5154
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            5.33255    2.25346   2.366 0.017963 *
## absences              -0.06079    0.02187  -2.779 0.005447 **
## age                   -0.23214    0.11412  -2.034 0.041943 *
## G1                    -0.14347    0.03876  -3.702 0.000214 ***
## `Weekend alcohol`     -0.05385    0.09936  -0.542 0.587829
## schoolMS              -0.87111    0.49429  -1.762 0.078009 .
## addressUrban           0.47714    0.32259   1.479 0.139120
## traveltime>1 hour      1.80584    0.91829   1.967 0.049237 *
## traveltime15-30 min   -0.50077    0.28055  -1.785 0.074261 .
## traveltime30-60 min   -0.53308    0.56157  -0.949 0.342483
## schoolsupyes          -0.19990    0.35530  -0.563 0.573688
```

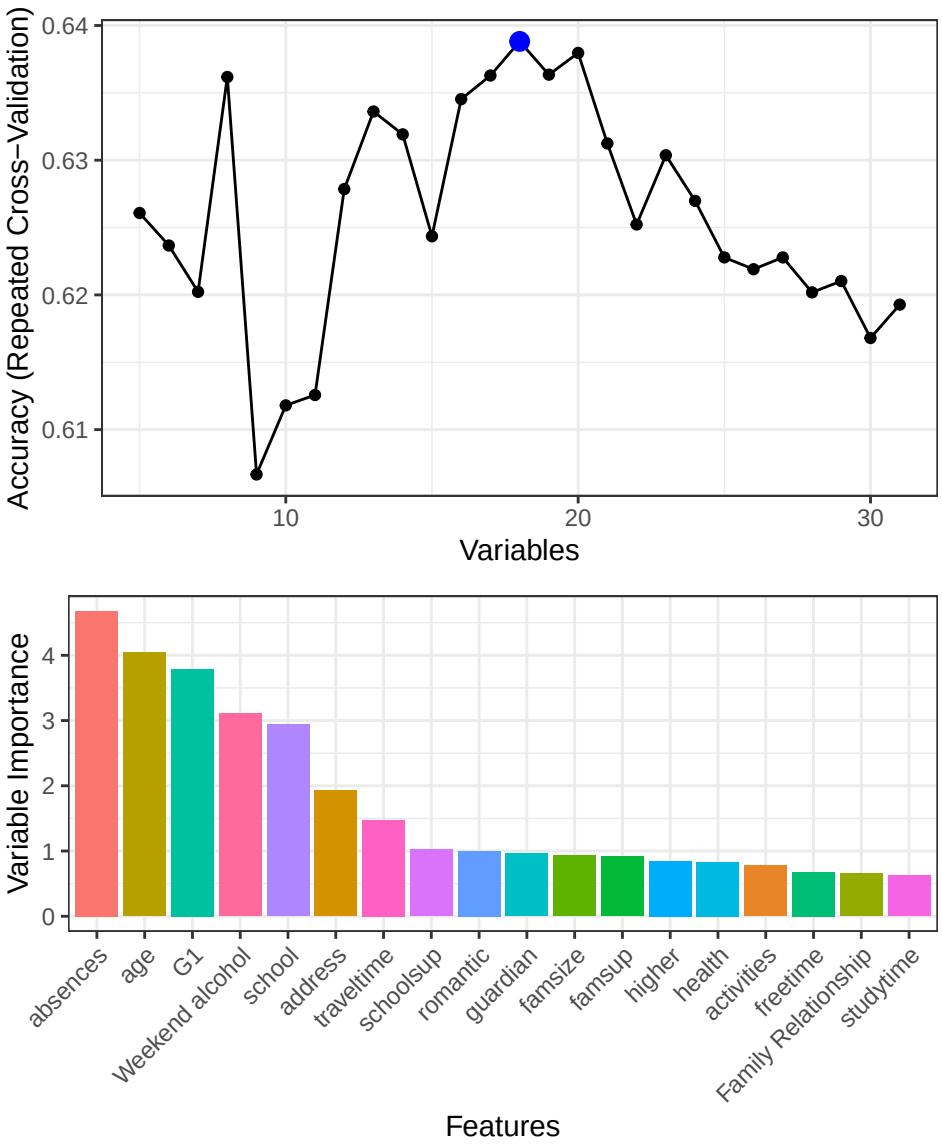

Figure 10: Accuracy of the model based on the number of features included and feature importance.

```
## famsupyes              0.12598    0.24696   0.510 0.609964
## romanticyes           -0.45645    0.25639  -1.780 0.075021 .
## higheryes             -0.21554    0.56330  -0.383 0.701985
## health                -0.06053    0.08561  -0.707 0.479564
## activitiesyes         -0.24098    0.23607  -1.021 0.307347
## guardianmother        -0.25150    0.27851  -0.903 0.366529
## guardianother         -0.03017    0.53451  -0.056 0.954989
## 'Family Relationship'  0.21175    0.13574   1.560 0.118777
## famsizeLE3            -0.04174    0.26453  -0.158 0.874611
## reasonhome            -0.04197    0.29346  -0.143 0.886282
## reasonother            1.04713    0.43762   2.393 0.016722 *
## reasonreputation      -0.03283    0.30330  -0.108 0.913795
## PstatusLiveTogether   -0.22714    0.39757  -0.571 0.567778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 528.27  on 394  degrees of freedom
## Residual deviance: 453.66  on 371  degrees of freedom
## AIC: 501.66
##
## Number of Fisher Scoring iterations: 4
```

Conclusions

Much insight was gained about this dataset and the effects that play on a student's performance in a class. It was found that the G3 grade is significantly different from the students' mothers and fathers education along with the mothers job. Being able to partake in extra paid classes, a willingness to peruse more classes, and being in a romantic relationship are also different. Through the creation of box plots, it was found that parents with a higher education degree and students are are wanting a higher degree for themselves will result in a better G3 grade. This is in contrast to students that are in a romantic relationship which sees a lower G3 grade.

In addition, the role of missing values was investigated within this dataset. Through MCAR simulations, it was shown that completely random missing values do not introduce much bias into the results. However, with non-random values from MNAR, the resulting test statistics were significantly different from the original test statistics.

Table 2 was able to show that the average G3 grade is lower than the G1 and G2 grades. This could imply that as the year continues, more factors come into play or that the students begin to feel fatigued. With the paired t-test analysis, it was found that this data set is suitable and fits the assumptions. In addition, the hypothesis of equivalence of mean in G1, G2 and g3 were rejected, therefore there is a significant difference with G3 and from both G1 and G2.

Paired T-test was performed on all three datasets: original data where missing values in scores were represented with 0's, imputed dataset based using multiple imputation and MCAR dataset (where missing values were introduced randomly to imputed dataset) where missing records were dropped from analysis. Imputed dataset and MCAR dataset yield close t-test results while t-test base upon original dataset is a little off from the other two. Thus, accurately dealing with missing values is very important in real world data analysis. Only if missing values are complete random, we can't directly drop it or impute with 0's.

The correlation plot shows that G3 is related to variables such as past failures, lack of family or school support, going out with friends, and poor health conditions which all negatively impact the G3 outcome for the student. It was also shown that males performed higher than females, possibly suggesting that traditional

gender roles could affect the performance. Lastly, more study time was shown to have a positive correlation with the G3 grades. The Q-Q plots shows that the data set is approximately normal distribution, confirming our assumption. For missing data, the MCAR was simulated in the data set whereas it is assumed that MNAR is already present due to missing G2 and G3 grades for some students.

Furthermore, with the regression analysis the final linear model to express the G3 grade was determined to contain the variables: age, mother's education (higher education), mother's job (services, health), father's job (teacher), study time (5-10 hours), past failures, school support (yes), going out with friends, and health. Here we can see that the final performance of a student depends not only on just the final exam but actually a list of several variables that were tested (there could be others included that were not investigated in this study). This reveals that the personal lives of the students are just as influential as the classes themselves. With these findings, schools and governments could create programs to prevent students from partaking in or warning against the negatively correlated variables such as going out with friends.

Lastly, based on Logistic regression modeling, whether a student able to improve their score from G1 to final exam is highly correlated the number of adscense from class. Thus, in order to improve a student's grade, we would suggest him/her to attend as many classes as possible. ## References

[1] Paulo Cortez and Alice Silva ((2014)), 'Using data mining to predict secondary school student performance'

[2] Cortez, Paulo. (2014). Student Performance. UCI Machine Learning Repository.

[3] https://www.npr.org/2012/04/09/150062919/lack-of-graduates-hampers-portugals-recovery

[4] https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

[5] https://www.gerkovink.com/miceVignettes/Convergence_pooling/Convergence_and_pooling.html

[6] https://towardsdatascience.com/effective-feature-selection-recursive-feature-elimination-using-r-148ff998e4f7

[7] https://stats.stackexchange.com/questions/2492/is-normality-testing-essentially-useless

[8] \url{http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/#:~:text=The%20stepwise%20regression%20(or%20stepwise,model%20that%20lowers%20prediction%20error