Statistical Inference: Project Phase I

Sarmad Zandi Goharrizi - 810199181

Question 0

a.

Student's Performance includes various information about a sample of students studying in two different *schools*.
A sense of responsibility towards one's education and academic future is a notable information which can be mined from each individual's *study time* and their rate of *going out* which has an effect on their *failures* and their *grades*.
This dataset also contains some semi-relevant factors like each student's parent's job as well as their love life.

X	school	sex	age	Fjob	Mjob	goout	internet	romantic	studytime	failures	health	absences	G1	G2	G3
0	GP	F	18	teacher	at_home	4	no	no	2	0	3	6	5.000000	7.529856	9.289229
1	GP	F	17	other	at_home	3	yes	no	2	0	3	4	5.000000	7.192039	9.424835
2	GP	F	15	other	at_home	2	yes	no	2	3	3	10	3.807703	8.000000	7.354029
3	GP	F	15	services	health	2	yes	yes	3	0	5	2	15.000000	16.373208	17.796916
4	GP	F	16	other	other	2	no	no	2	0	5	4	6.000000	12.138542	12.800024
5	GP	M	16	other	services	2	yes	no	2	0	5	10	15.000000	16.804680	18.347259
6	GP	M	16	other	other	4	yes	no	2	0	3	0	12.000000	13.691091	14.187810
7	GP	F	17	teacher	other	4	no	no	2	0	1	6	6.000000	6.794185	9.012740
8	GP	M	15	other	services	2	yes	no	2	0	1	0	16.000000	19.852952	20.000000
9	GP	M	15	other	other	1	yes	no	2	0	5	0	14.000000	17.180466	18.073614
10	GP	F	15	health	teacher	3	yes	no	2	0	2	0	10.000000	9.609179	11.950918

Figure 1: Head of the dataset

b.

We have a dataset of 395 students. Each student have 16 features (some of them where mentioned in part a).

```
> summary(StudentsPerformance)
       X          school   sex          age              Fjob            Mjob           goout        internet  romantic
 Min.   :  0.0   GP:349   F:208   Min.   :15.0   at_home : 20   at_home : 59   Min.   :1.000   no : 66   no :263
 1st Qu.: 98.5   MS: 46   M:187   1st Qu.:16.0   health  : 18   health  : 34   1st Qu.:2.000   yes:329   yes:132
 Median :197.0                    Median :17.0   other   :217   other   :141   Median :3.000
 Mean   :197.0                    Mean   :16.7   services:111   services:103   Mean   :3.109
 3rd Qu.:295.5                    3rd Qu.:18.0   teacher : 29   teacher : 58   3rd Qu.:4.000
 Max.   :394.0                    Max.   :22.0                                 Max.   :5.000
   studytime        failures          health          absences            G1               G2               G3
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   : 0.000   Min.   : 1.714   Min.   : 0.000   Min.   : 0.00
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.000   1st Qu.: 9.988   1st Qu.:10.00
 Median :2.000   Median :0.0000   Median :4.000   Median : 4.000   Median :11.000   Median :12.244   Median :13.37
 Mean   :2.035   Mean   :0.3342   Mean   :3.554   Mean   : 5.709   Mean   :10.783   Mean   :12.273   Mean   :12.64
 3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.000   3rd Qu.:15.076   3rd Qu.:16.47
 Max.   :4.000   Max.   :3.0000   Max.   :5.000   Max.   :75.000   Max.   :19.000   Max.   :20.000   Max.   :20.00
>
```

Figure 2: Summary of the dataset

c.

As *Figure 3* and *4* suggests, there were no missing values in our dataset.
If so, there are a multitude of methods to handle missing data like, list-wise deletion, estimating them using other similar variables and ... .

X	missing value	0
school	missing value	0
sex	missing value	0
age	missing value	0
Fjob	missing value	0
Mjob	missing value	0
goout	missing value	0
internet	missing value	0
romantic	missing value	0
studytime	missing value	0
failures	missing value	0
health	missing value	0
absences	missing value	0
G1	missing value	0
G2	missing value	0
G3	missing value	0

Figure 3: proportion of missing value in each feature

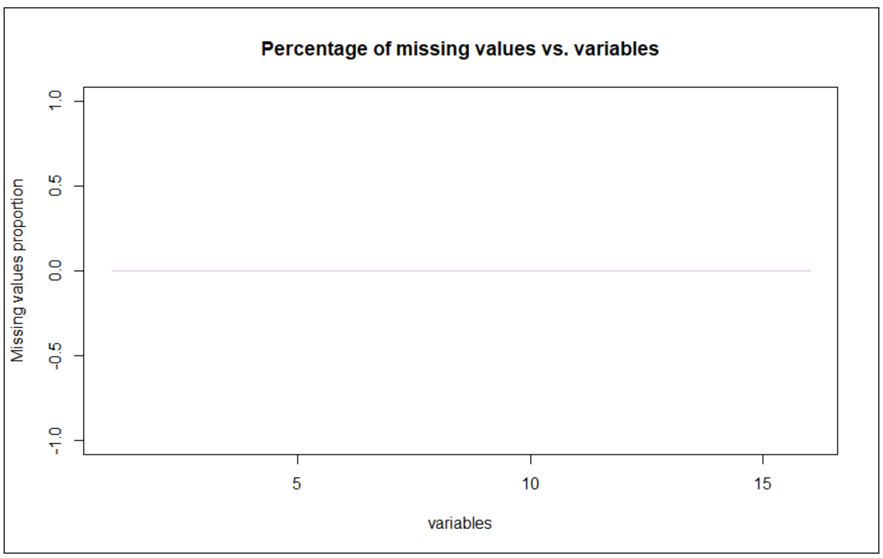


Figure 4: Line plot of missing value proportion

d.

Each student's performance is influenced highly from many different factors and cannot be decided using 3 grades, however we have to work with what we have and as was mentioned in part a, *study time* plays and important role in each individual's grades.

Question 1

Chosen Numerical Variable : *G1*

a.

The appropriate bin width is computed using *Freedman–Diaconis rule* , which leads to a normally distributed histogram of G1.

$$Bin\ Width : 2\frac{IQR(x)}{\sqrt[3]{n}}$$

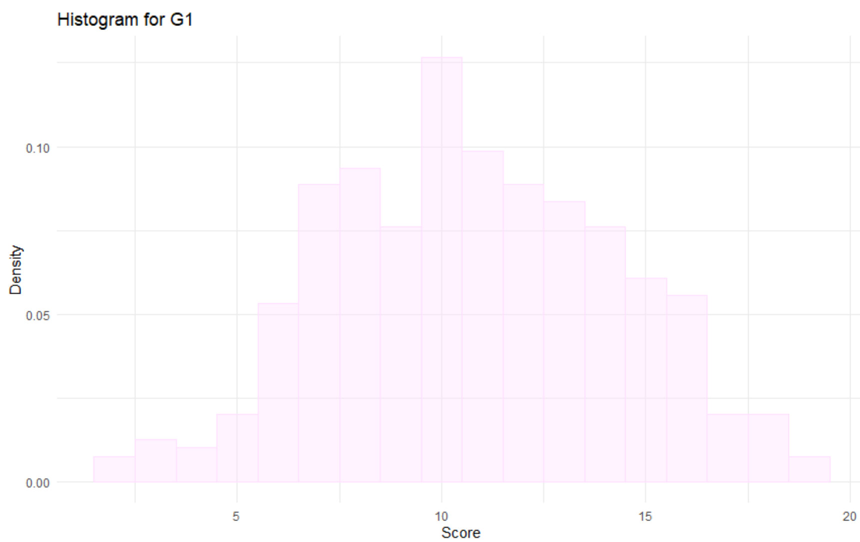


Figure 5: Histogram of G1

Figure 6 describes a *unimodal*.
A *unimodal* distribution is a distribution that has one clear peak (as can be seen in *Figure 6*). The values increase at first, rising to a single highest point where they then start to decrease. A *unimodal* distribution can either be symmetrical or non-symmetrical (more about this in part c).

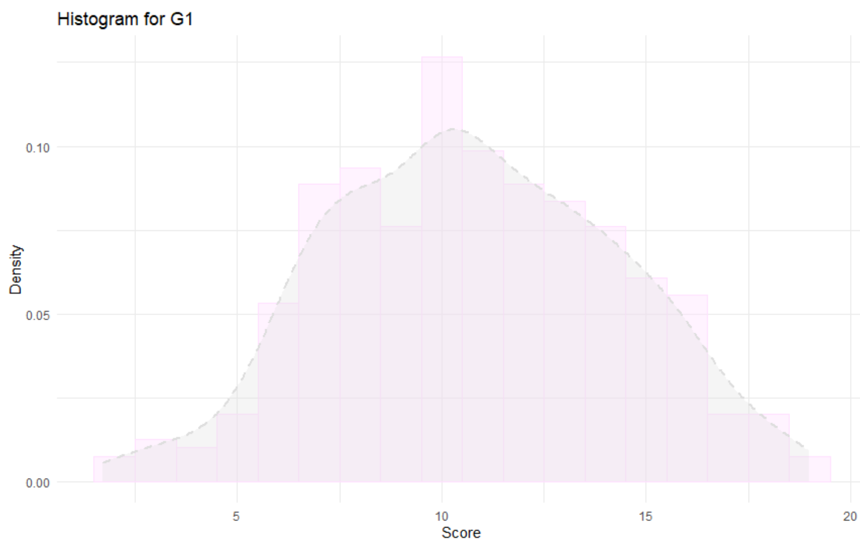


Figure 6: Histogram of G1 overlaid with density plot

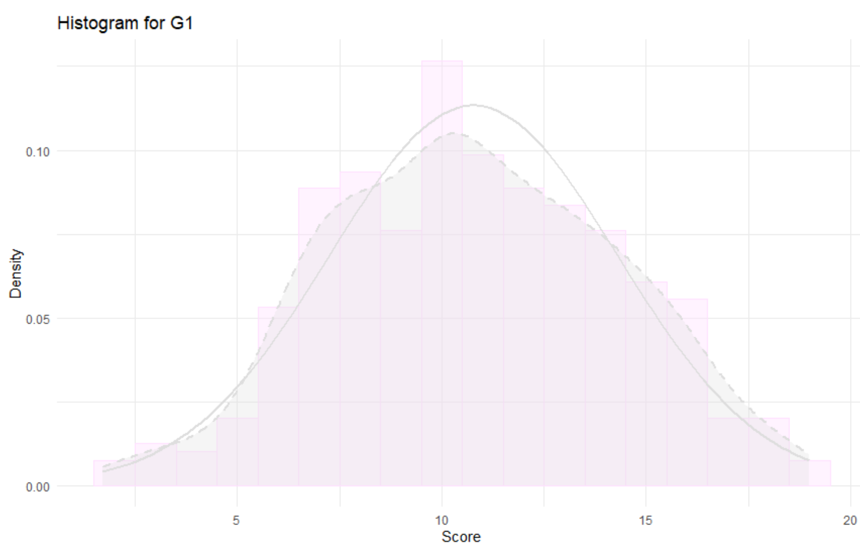


Figure 7: Histogram of G1 overlaid with fitted density plot and MLE density plot

b.

There are 3 basic properties of a distribution that we have to address: *location, spread,* and *shape*.

The *location* refers to the typical value of the distribution, such as the *mean* (10.783) or *median* (11.00).

The *spread* of the distribution is the amount by which smaller values differ from larger ones. The *standard deviation* (3.521) or *variance* (12.39) are measures of distribution spread.

The *shape* of a distribution is its pattern—peakedness, symmetry, etc. A given phenomenon may have any one of a number of distribution shapes, e.g., the distribution may be bell-shaped, rectangular-shaped, etc which in our case is nearly *bell-shaped symmetrical* (*unimodal*) as was mentioned in part a and will be discussed in part c.

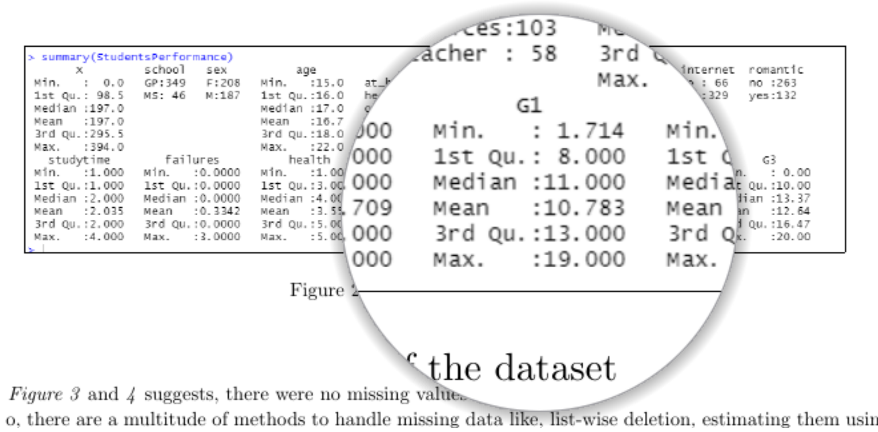


Figure 8: G1 under magnifier

It can be clearly seen that this distribution is very similar to the normal distribution but to be more precise, we use *normal Q-Q plot.*
The main purpose of a *normal probability plot (normal Q-Q plot)* is to assess normality.

A one-to-one relationship (straight line in *Figure 8*) between the data and the theoretical quantiles can be considered, so the data follow a nearly normal distribution. In other words, the closer the points to the straight line, the more confident we can be that the data follow the normal model.)

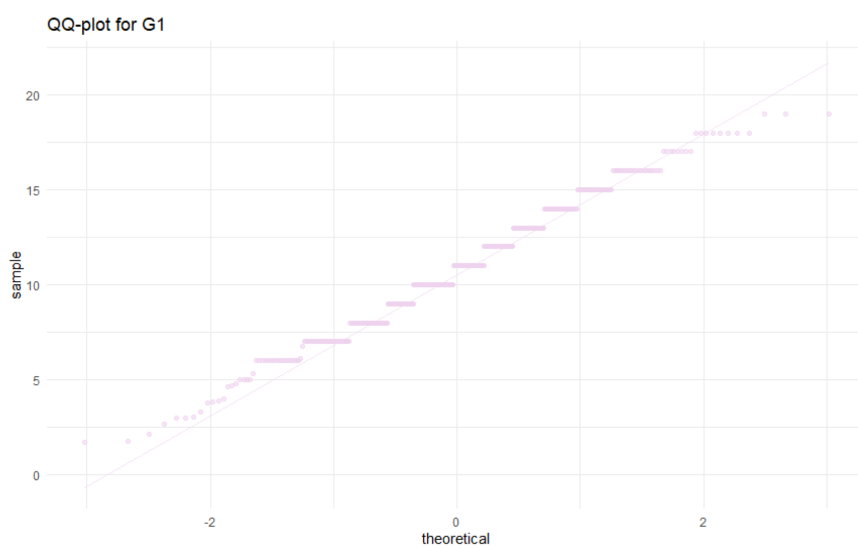


Figure 9: Normal Q-Q plot of G1

c.

Skewness is a statistical numerical method to measure the asymmetry of the distribution or data set. It tells about the position of the majority of data values in the distribution around the mean value.

$$Skewness = \frac{mean - median}{sd}$$

One method to address the skewness is to compare the mean and the median.
If :

$$1.mean > median : right\ skewed\ (negatively\ skewed)$$

$$2.mean = median : Symmetric$$

$$3.mean < median : left\ skewed\ (positively\ skewed)$$

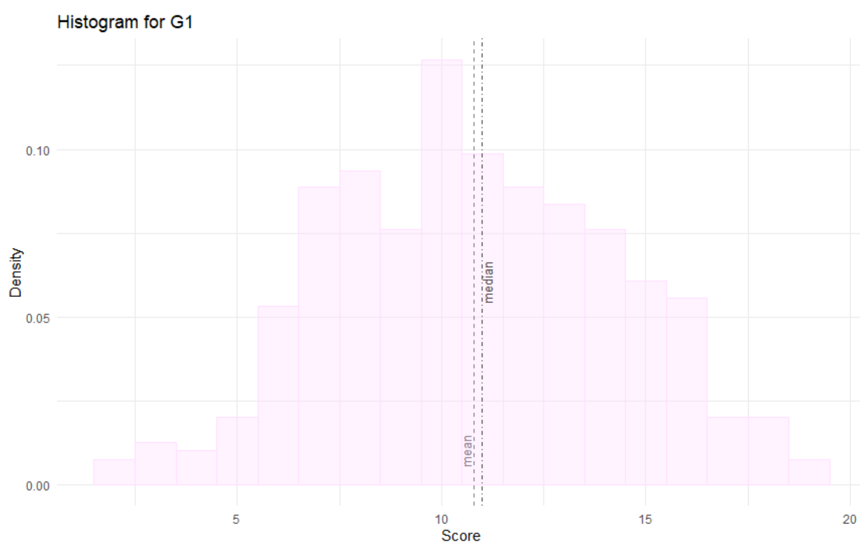


Figure 10: Median and mean marked on histogram of G1

As can be deducted from *Figure 10*, G1 (barely) falls under the third category.
This conclusion can also be supported by calculating the skewness of G1 :

```
>
> skewness(StudentsPerformance$G1)
[1] 0.01764784
>
```

Figure 11: Calculated skewness of G1

The coefficient of skewness is greater than 0, meaning the graph is positively skewed with the majority of data values less than mean. In other words, most of the values are concentrated on the left side of the graph.

d.

An outlier is a value or an observation that is distant from other observations, that is to say, a data point that differs significantly from other data points.
Boxplots provide a useful visualization of the distribution of data. Typically, Boxplots show the *median*, 1^{st} *quartile*, 3^{rd} *quartile*, *maximum datapoint*, and *minimum datapoint* for a dataset (more to it in part h) and also, last but not least, *outliers*. Fortunately, my chosen variable didn't have any outliers and the *Figures 12* and 13 below are the proof.

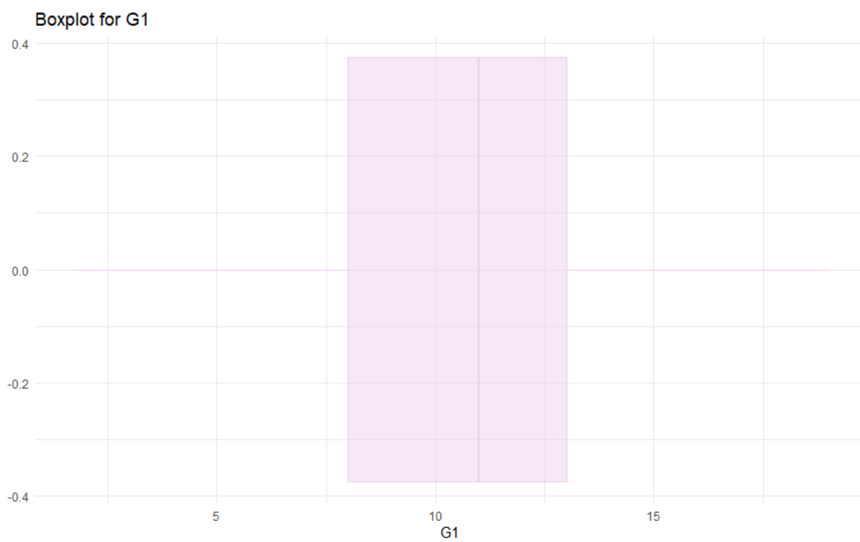


Figure 12: Boxplot of G1 to visualize outliers

```
>
> boxplot.stats(StudentsPerformance$G1)$out
numeric(0)
>
```

Figure 13: Using stats of boxplot to visualize outliers

e.

Mean : The mean identifies the average value of the set of numbers.

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median : The median identifies the midpoint or middle value of a set of numbers.

Variance : Variance measures the variability of the data set. It indicate how far individuals in the group are spread out, in the set of data from the mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$$

Standard deviation : Standard deviation measures the dispersion of the data set. A smaller standard deviation indicates less variability. tandard deviation is expressed in the same unit as the values in the dataset so it measure how much observations of the data set differs from its mean.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)}$$

```
>
> mean(StudentsPerformance$G1)
[1] 10.78285
> median(StudentsPerformance$G1)
[1] 11
> var(StudentsPerformance$G1)
[1] 12.39784
> sd(StudentsPerformance$G1)
[1] 3.521057
>
```

Figure 14: Statistics: Mean-Median-Variance-Standard Deviation

f.

The perfect description of the relationship between *mean*, *median* and *density* is that the *median* of a density curve is the point that divides the area under the curve in half, the *mean* is the point at which the curve would balance if made out of solid material.
In a perfectly symmetrical distribution, the mean and the median are the same.

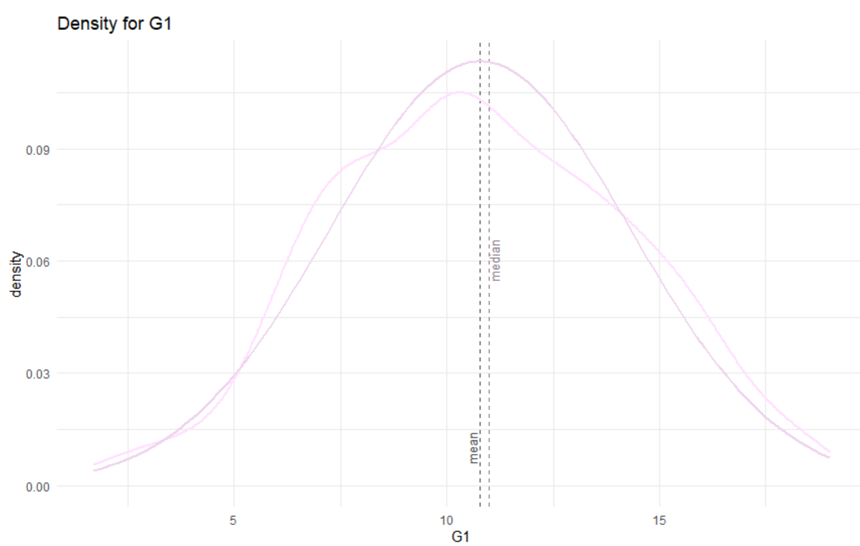


Figure 15: Median and mean marked on density of G1 - darker one is drawn using *dnorm*

g.

Pie charts are best to use when you are trying to compare parts of a whole. For this question, two different courses of action where taken :

First Method : Categorizing data by a range of values
In this approach categories are created according to logical cut-off values in the scores or measured values.

$$\begin{cases} G1 < \frac{\mu}{2} & \textit{Very Low} \\ \\ \frac{\mu}{2} < G1 < \mu & \textit{Low} \\ \\ \mu < G1 < \frac{\mu + max(G1)}{2} & \textit{High} \\ \\ G1 > \frac{\mu + max(G1)}{2} & \textit{Very High} \\ \end{cases}$$

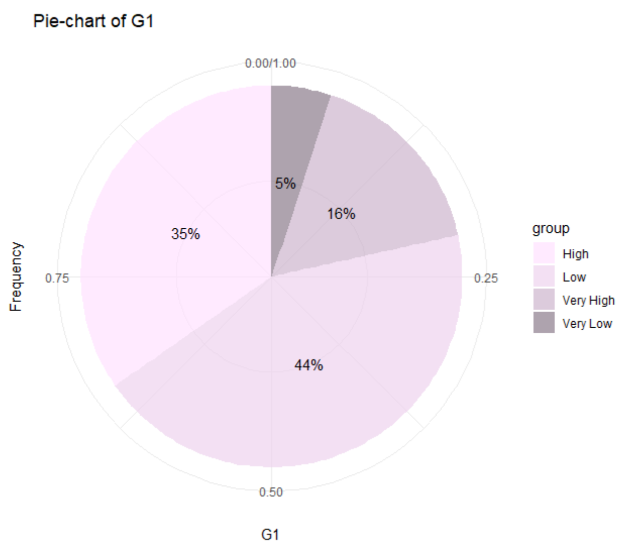


Figure 16: Piechart of G1 - 1^{st} method

Second Method : Categorizing data by percentiles (since mean and median are close)
A second approach is to use percentiles to categorize data. The advantage to this approach is that it does not rely on the scoring system being meaningful in its absolute values

$$\begin{cases} G1 < 25^{th} percentile & Very\ Low \\ 25^{th} percentile < G1 < 50^{th} percentile & Low \\ 50^{th} percentile < G1 < 75^{th} percentile & High \\ G1 > 75^{th} percentile & Very\ High \end{cases}$$

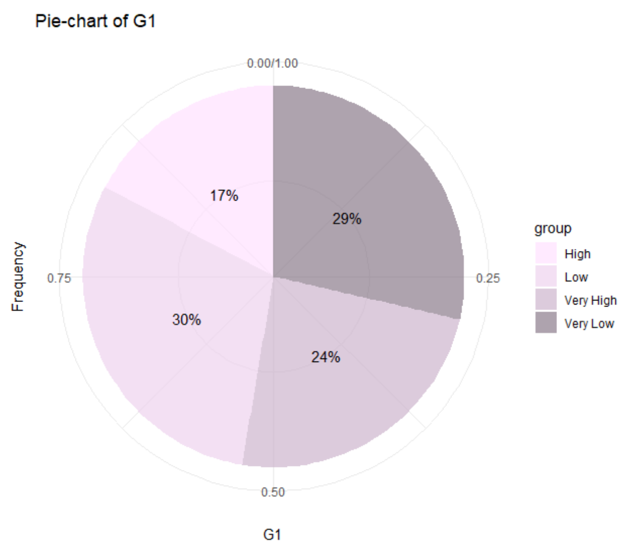


Figure 17: Piechart of G1 - 2^{nd} method

In this approach, there are approximately an equal number of respondents in each category.

h.

```
>
> G1.quant
      0%       25%       50%       75%      100%
 1.713843  8.000000 11.000000 13.000000 19.000000
>
>
```

Figure 18: 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles of G1

```
>
> G1.iqr
[1] 5
>
>
```

Figure 19: IQR of G1

Box plots are a five-number summary that includes the minimum and maximum data values, the median and lower and upper quartiles. They can be useful in understanding how is data distributed in a given set and give information about the spread of the data.

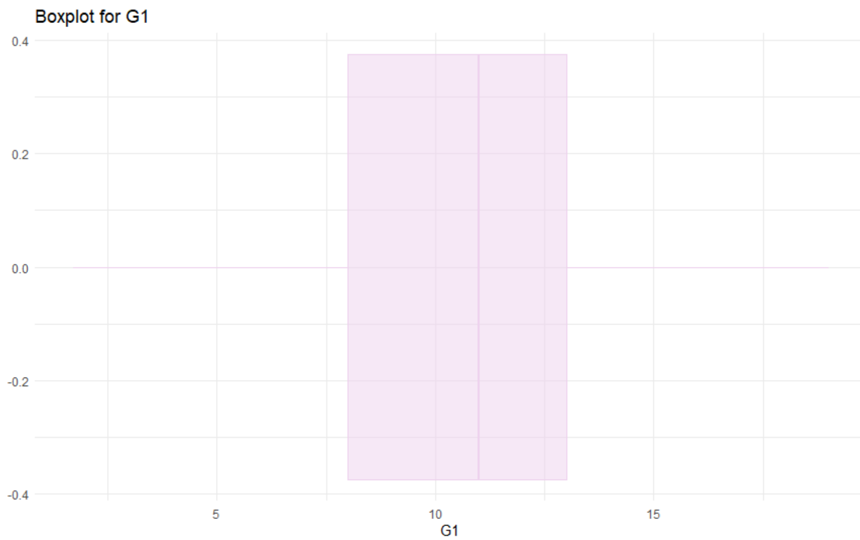


Figure 20: Boxplot of G1

```
> boxplot.stats(StudentsPerformance$G1)
$stats
[1]  1.713843  8.000000 11.000000 13.000000 19.000000

$n
[1] 395

$conf
[1] 10.60251 11.39749

$out
numeric(0)
```

Figure 21: Stats of Boxplot of G1

From *Figure 20*, G1 being (barely) LS is also clear.

Question 2

Chosen Categorical Variable : *sex*

a.

Most of them are female students.

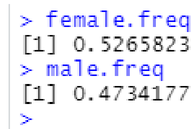


Figure 22: Frequency of each category and its percentage

b.

A stacked barplots is a variant of the bar chart.
A standard barplots compares individual data points with each other. In a stacked barplots, parts of the data are adjacent (in the case of horizontal bars) or stacked (in the case of vertical bars); each bar displays a total amount, broken down into sub amounts.
Stacked barplots are useful for visualizing conditional frequency distributions.(But in general, it is better to avoid them.)

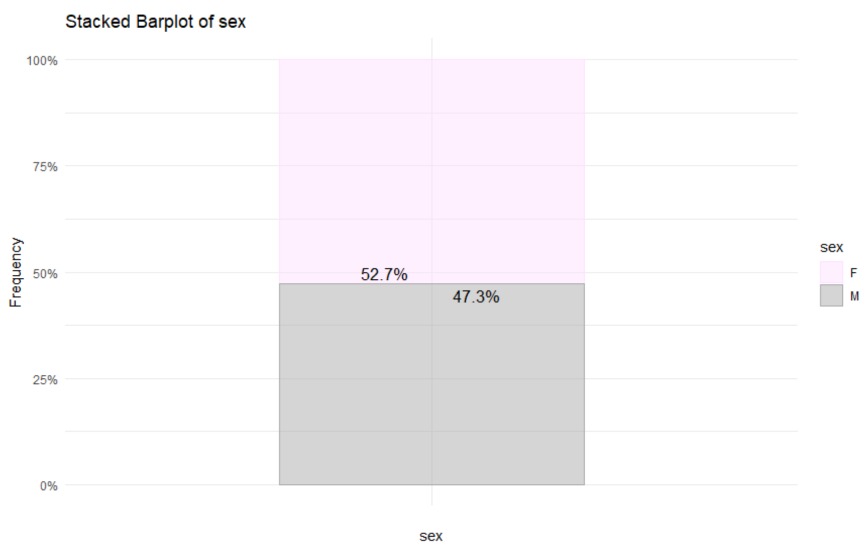


Figure 23: Stacked barplot of sex

c.

Barplots for categorical variables are like histograms for numerical variables.

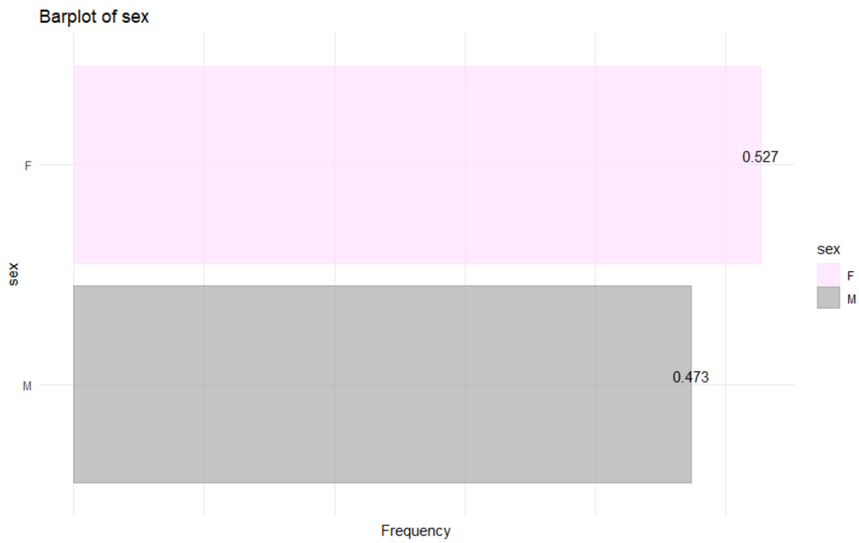


Figure 24: Horizontal barplot of sex

d.

A violinplot is a method of plotting numeric data. It is similar to a boxplot, with the addition of a rotated kernel density plot on each side.
A violinplot is more informative than a plain boxplot. While a boxplot only shows summary statistics such as mean/median and inter-quartile ranges, the violin plot shows the full distribution of the data. Wider sections of the violin plot represent a higher probability that members of the population will take on the given value; the skinnier sections represent a lower probability.
Violin plots are used to represent comparison of a variable distribution (or sample distribution) across different "categories" .
In our case, *Female* students are around 16 to 18 years old and the distribution of *Male* is wider than *Female* and continues until the age of 22 years.

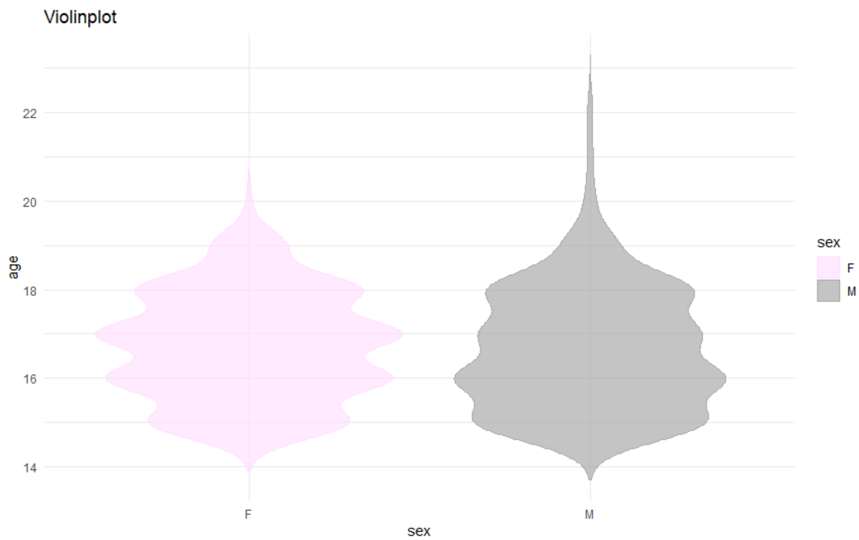


Figure 25: Violin plot of sex

Question 3

Chosen Numerical Variables : *goout* and *absences*

a.

The data points might follow an overall positive trend, the more you go out, the less you can show up to class.
My guess is a positive non-linear relationship between these two.

b.

A clear relationship cannot be described. It seams like a bell-shaped relationship, also an outlier in goout = 1 is detected.

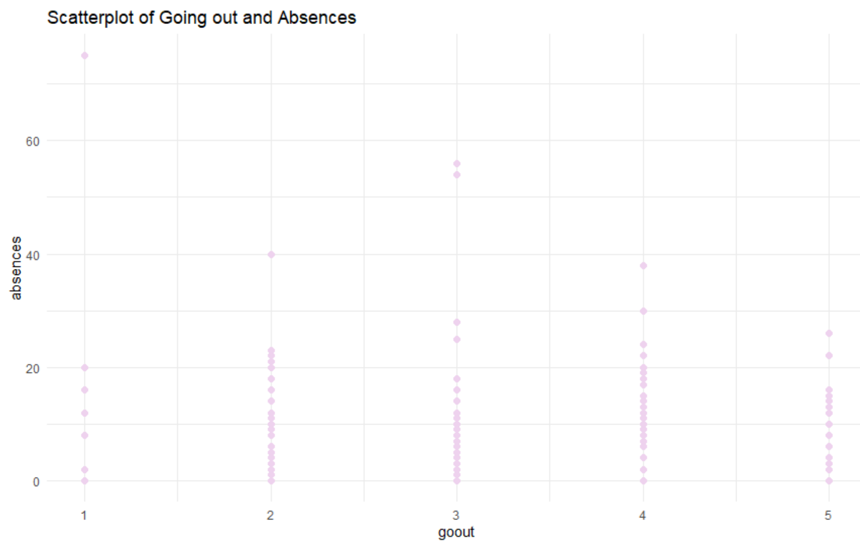


Figure 26: Scatterplot of goout and absences

c.

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables.
Correlation is computed using *Pearson correlation coefficient*.
Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}}\sqrt{\sum y_i^2 - n\bar{y}}}$$

```
>
> goout_absences.correlation
[1] 0.04430222
>
>
```

Figure 27: Correlation coefficient of goout and absences

d.

The correlation coefficient ranges from -1 to 1. A value of 1 implies that a *linear equation* describes the relationship between X and Y perfectly (a.k.a perfect positive correlation), with all data points lying on a line for which Y increases as X increases. A value of -1 implies that all data points lie on a line for which Y decreases as X increases (a.k.a perfect negetive correlation). A value of 0 implies that there is no linear correlation between the variables.

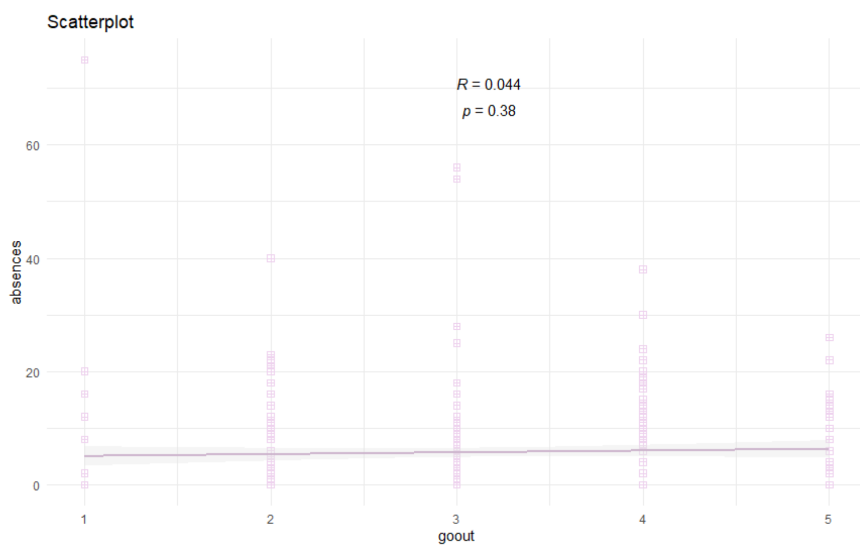


Figure 28: Scatterplot of goout and absences

In our case, $R = 0.044$ means no or negligible (positive) relationship.(So the assumption made in part a was somewhat true.)

e.

Statistical inference based on *Pearson's correlation coefficient* often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient r.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

In this part, the first aim is our target. A p-value is the probability that the null hypothesis is true. When using *Pearson's correlation coefficient*, it represents the probability that the *correlation* between x and y in the sample data occurred by chance.

In our case, ρ a.k.a *p-value* is 0.38.
A p-value of 0.38 means that there is 38% chance (!) that results from the sample occurred due to chance. Comparing to significant level of 5%, we fail to reject the null hypothesis.
We conclude that the correlation is not statically significant. Or in other words *we conclude that there is not a significant linear correlation between x and y in the population whatsoever.*

f.

Chosen Categorical Variable : *romantic*

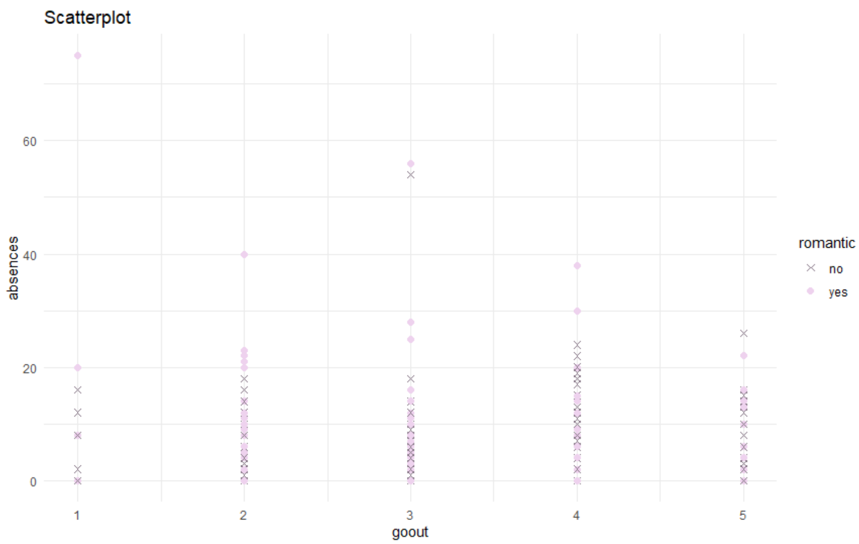


Figure 29: Scatterplot of goout and absences categorized by romantic

g.

Hexbin map uses hexagons to split the area into several parts and attribute a color to it. The graphic area is divided into a multitude of hexagons and the number of data points in each is counted and represented using a *color gradient.*

Hexbin plot is helpful in situations where :

- Creating an unbiased density distribution is needed
- Representing discrete categorical information is needed (Better than heatmaps in visualizing categorical information)
- Showing complete information by eliminating the edge effects is needed (Circle is the lowest ratio, but cannot form a continuous grid, and hexagons are the closest shape to a circle that can still form a grid.)

Hexbin plot should be avoided in situations where simplicity of definition and data storage is needed.

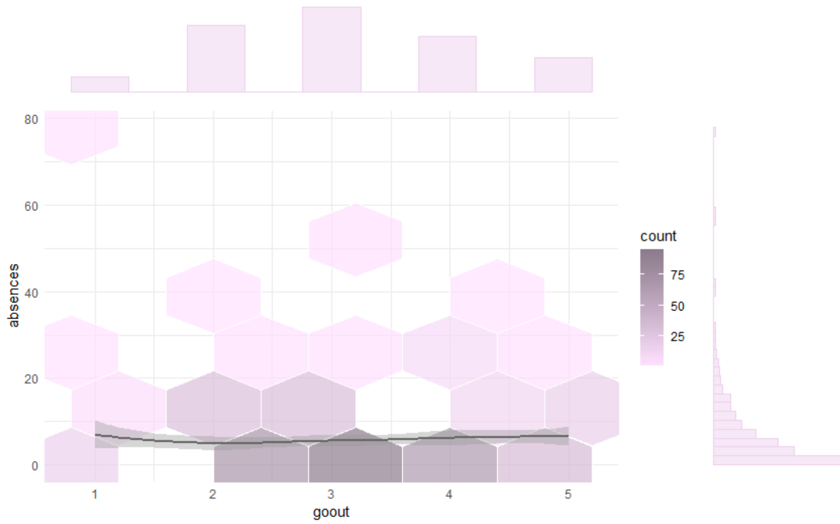


Figure 30: Hexbin plot, binsize = 5

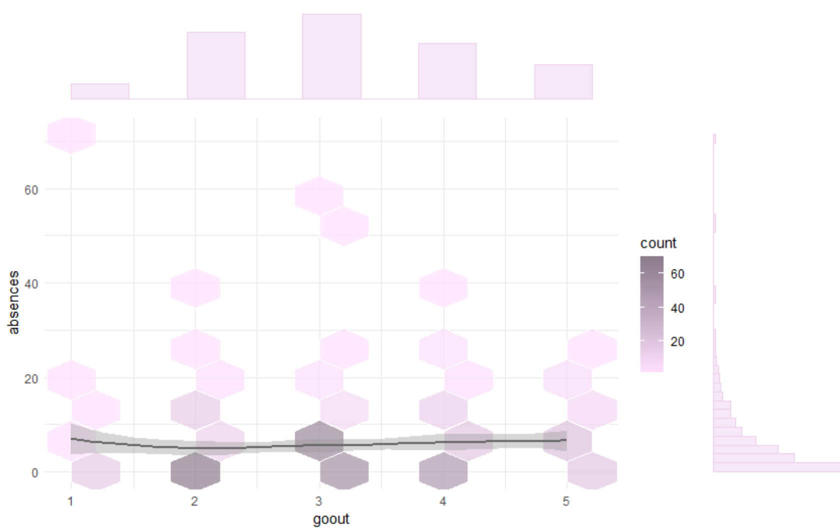


Figure 31: Hexbin plot, binsize = 10

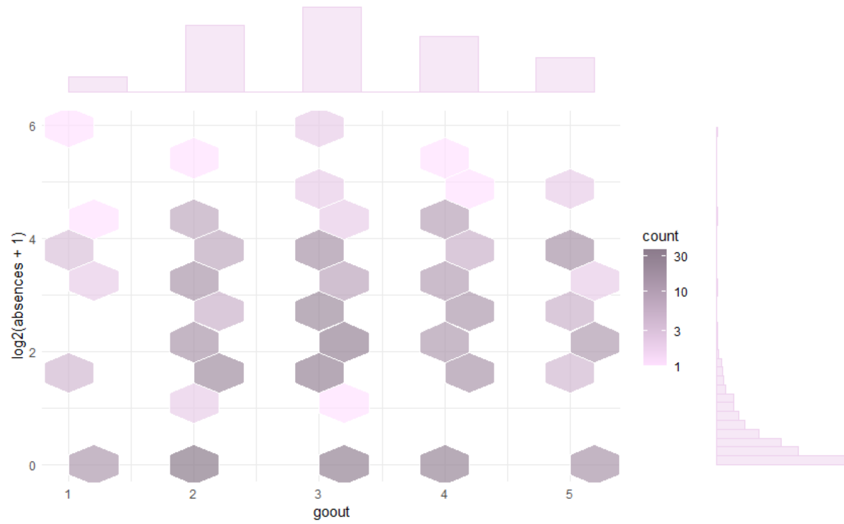


Figure 32: Hexbin plot, binsize = 10 (logarithmic)

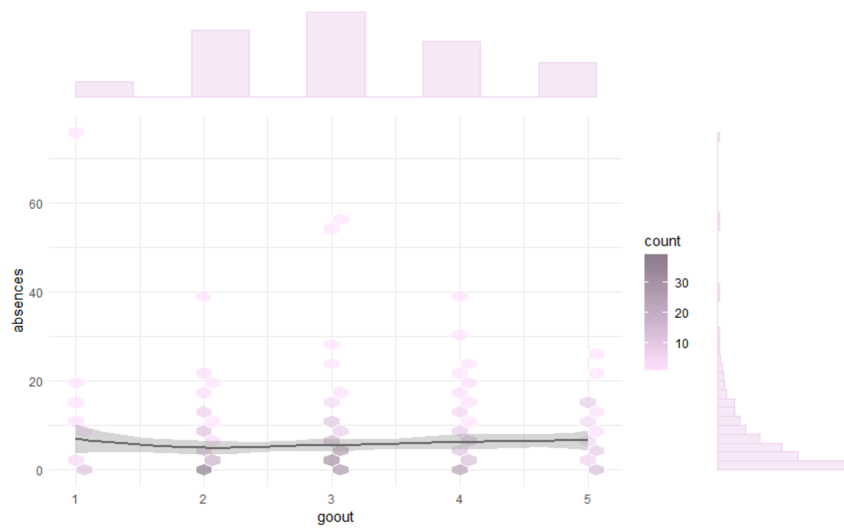


Figure 33: Hexbin plot, binsize = 30

It can be seen that by decreasing the binsize, each hexagon contains more amount of samples. Binsize about 10 is fairly good and can be informative. Bigger Binsizes will be misleading and not robust to noisy datas. Logarithmic plot was also plotted to have a better visualization.

h.

A 2D density plot displays the relationship between 2 numeric variables, where one variable is represented on the X-axis, the other on the Y axis. The number of observations within a particular area of the 2D space is counted and represented by a *color gradient* to indicate differences in the distribution of data in one region with respect to the other.
2D density plot is helpful in situations where :

- Sample size is huge and a clearer picture of the distribution is needed
- A nuanced visualization of density is needed (Better than heatmaps in visualizing categorical information)
- Visualize several distributions at once is needed

2D density plot should be avoided in situations where not enough data points are present, therefore risk of overplotting is low(using scatterplot is a more effective visualization).

The biggest disadvantage of 2D density plots and Hexbin maps are their sensitivity to bin size/bandwidth, inaccurate bin size/bandwidth and can lead to different and/or wrong conclusions.

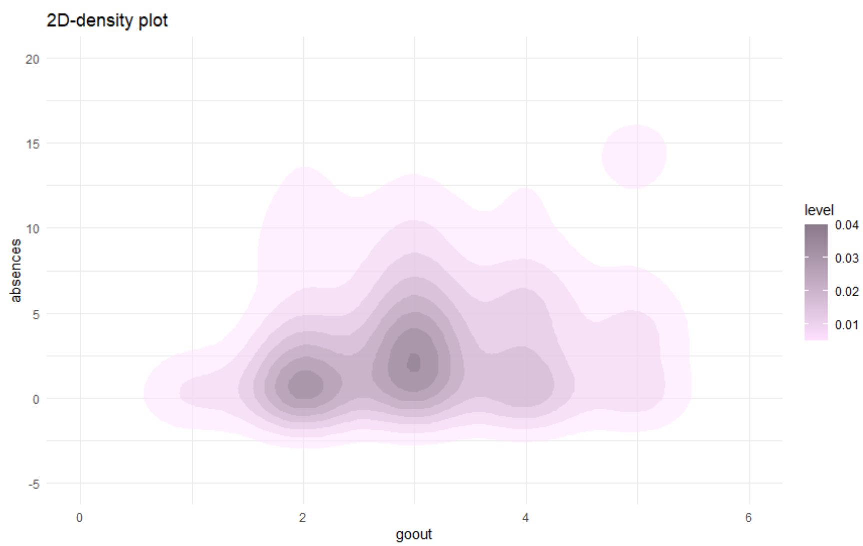


Figure 34: 2D density plot of goout and absences

As can be concluded from *Figure 30*, the densest part of the plot is when students goout 3 times and are absencent for 5 times.

Question 4

a.

Scatterplots of each pair of numeric variable are drawn on the left part of the figure. Pearson correlation is displayed on the right. Variable distribution is available on the diagonal.

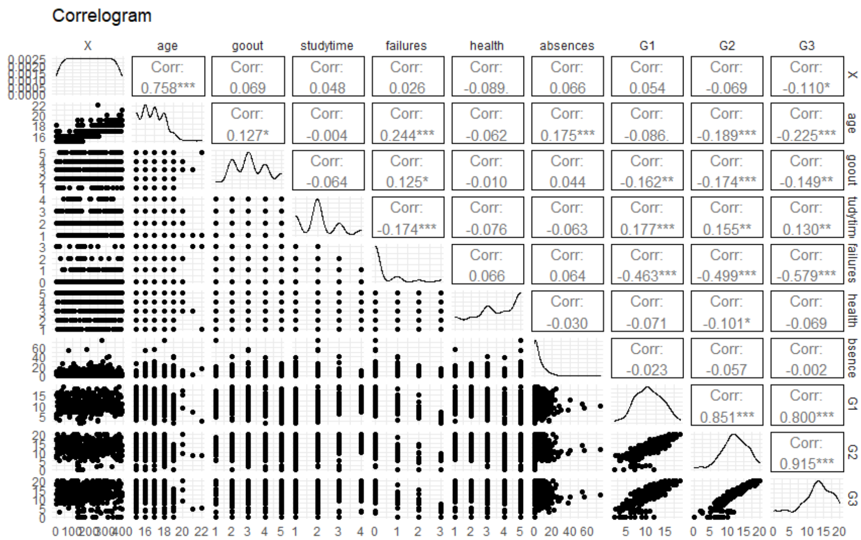


Figure 35: Bivarient Correlogram with pearson correlation

Density's bandwidth of *Failures* variable was inf, so we had to omid it in order to get a plot:

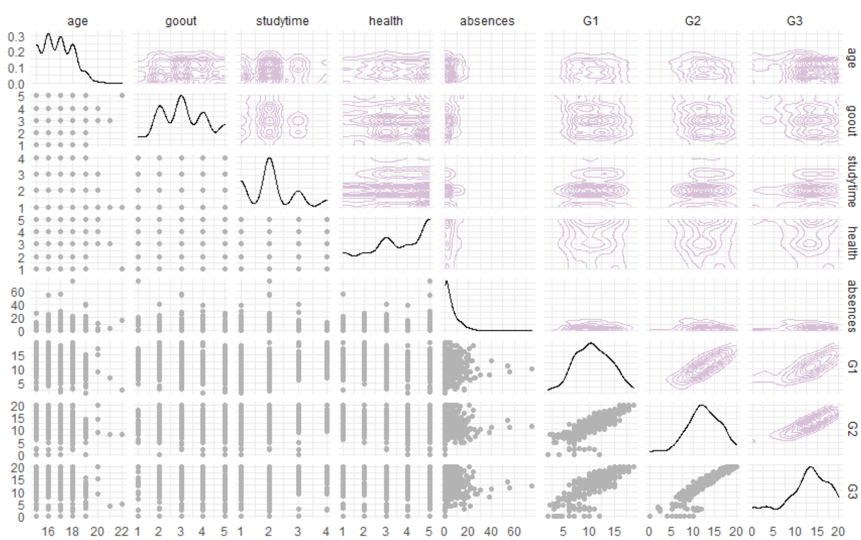


Figure 36: Bivarient Correlogram with density - scatterplot

Judging by *Figure 36*, where the scatterplot of 2 variables is dense, density plot is completely meaningful and where the scatterplot of 2 variables is not dense, density plot is not that informative and it's better to stick to scatterplot as was mentioned in part h of question3, *2D density plot should be avoided in situations where not enough data points are present, therefore risk of over-plotting is low(using scatterplot is a more* *effective visualization)*

To have the full view of all of our numerical variables, boxplot, barplot and scatterplot with linear association was also plotted :

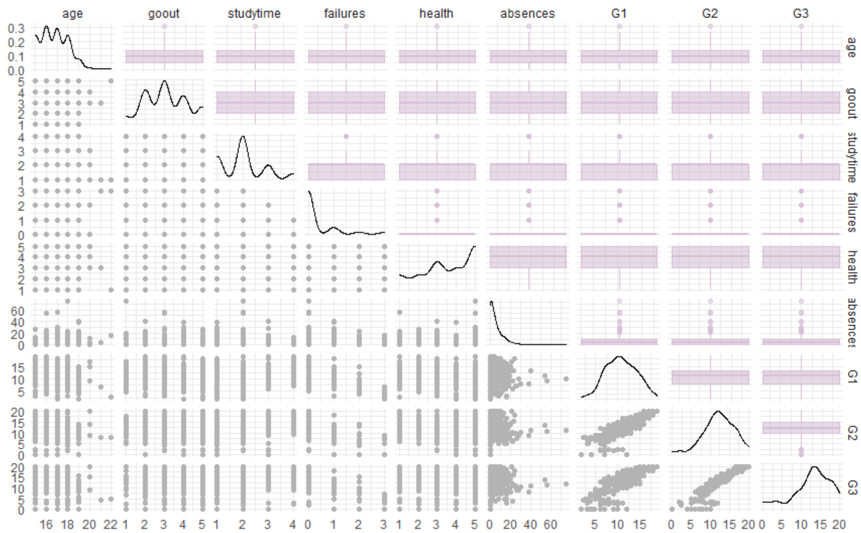


Figure 37: Bivarient Correlogram with barplot - scatterplot

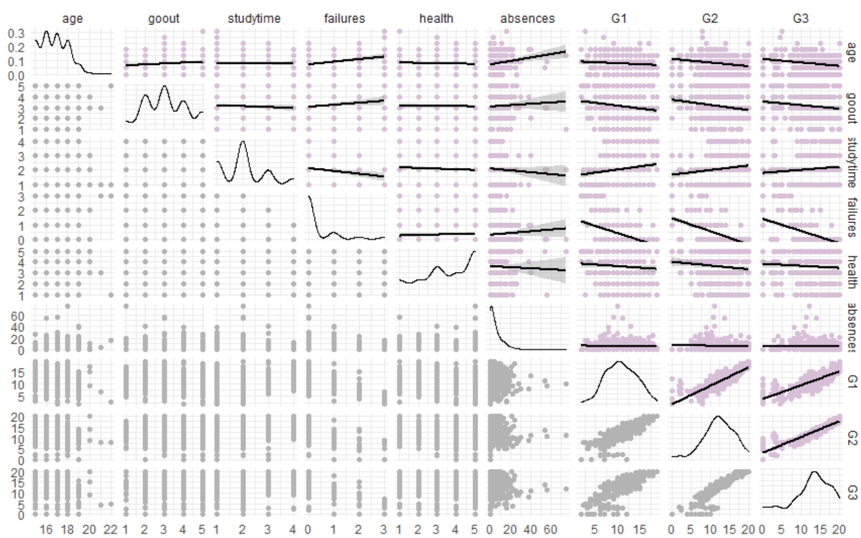


Figure 38: Bivarient Correlogram with linear assosiation - scatterplot

Judging by *Figure 38*, $G1$ and $G2$ and $G3$ have positive linear associations with each other and with *studytime* as expected. $Failure$ and *goout* both have a negative linear associations with $G1$, $G2$ and $G3$.

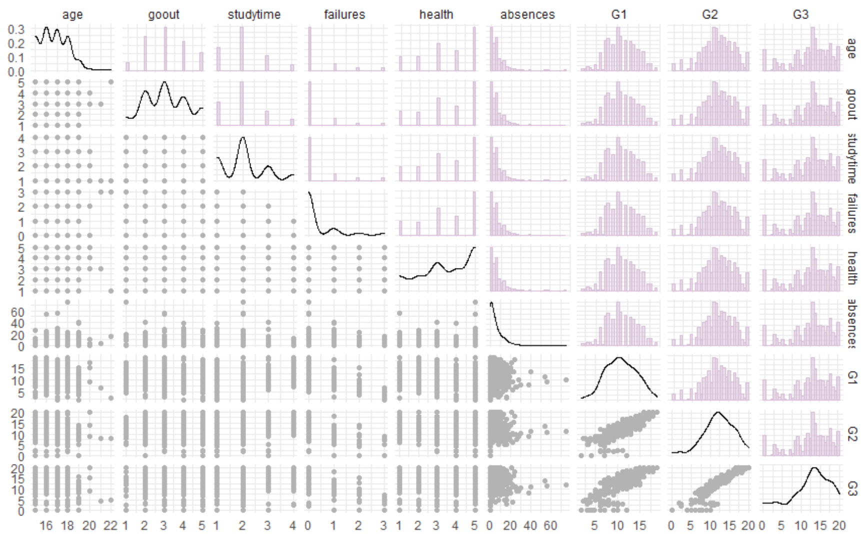


Figure 39: Bivarient Correlogram with barplot - scatterplot

b.

I used black for negative correlation and thistle for positive correlation (hope thats okay :))
significance level = 0.05 .
The cells that are crossed are rejected by p-value.
(Note : diag. correlations are omitted)

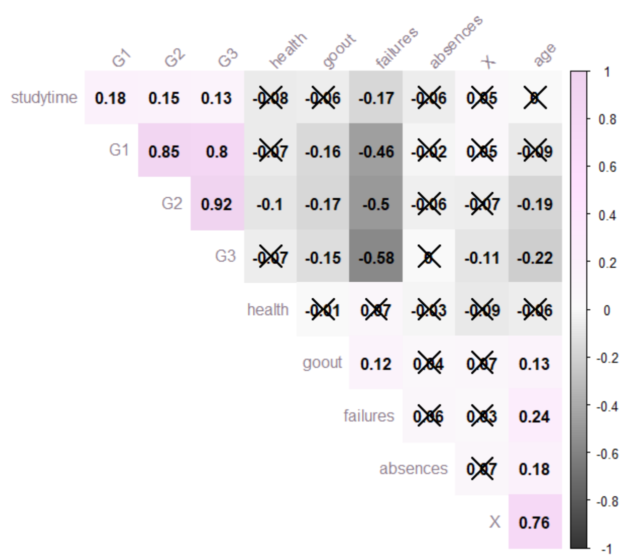


Figure 40: Heatmap correlogram of numerical values

c.

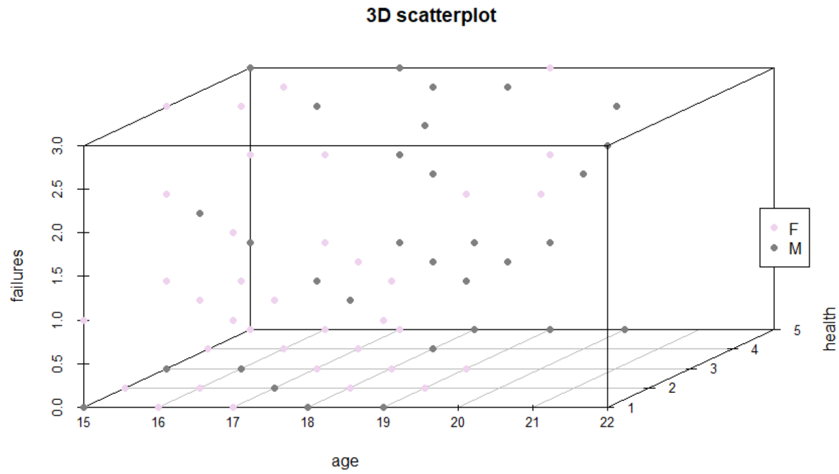


Figure 41: 3D scatterplot of age, failures and health colorized by sex

Unfortunately, it seems like there is not a specific relationship between these 3 variables; but we can see that *Females* have *Females failures* and *Males* and also, *Females* are in the younger *age* group.

Question 5

Chosen Categorical Variables : *sex* and *romantic*

a.

```
>
> print.table(table)

       F   M Sum
  no  129 134 263
  yes  79  53 132
  Sum 208 187 395
>
```

Figure 42: Frequency/ Contingency table of sex and romantic

b.

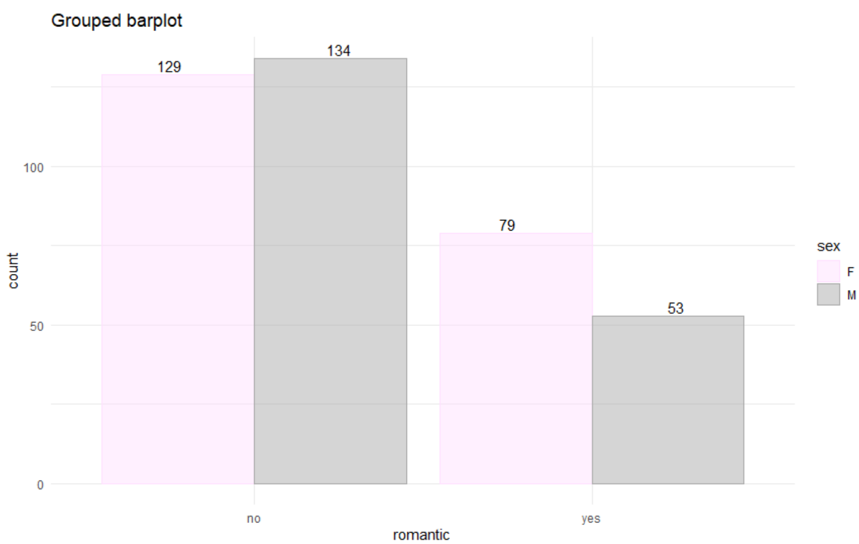


Figure 43: Grouped barplot of sex and romantic

c.

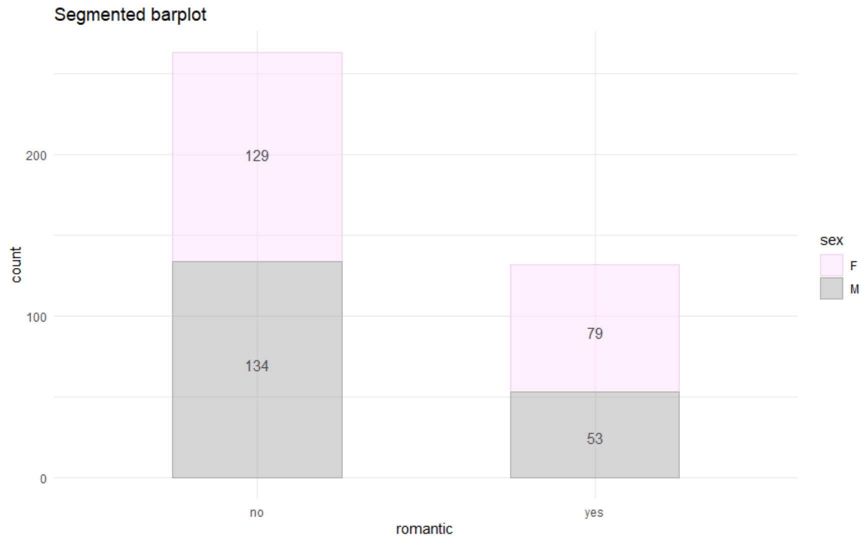


Figure 44: Segmented barplot of sex and romantic

d.

The segmented barplot does well in informing about the percent of each category within each group. The information that is missing is the size of each group.
A mosaic plot allows us to see these group sizes by scaling on the x-axis!

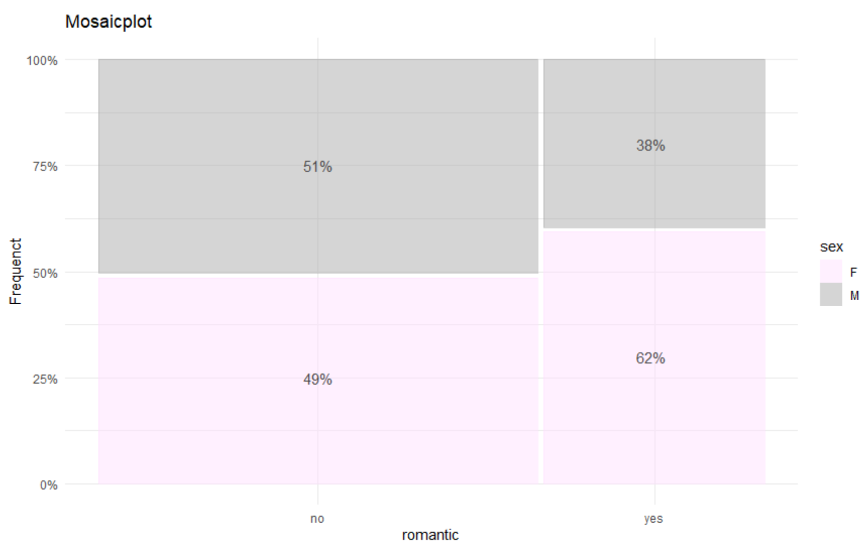


Figure 45: Mosaicplot of sex and romantic

Question 6

Chosen Numerical Variable : *goout*

Check Condition :

- Independent Observations :
  - Random sample/assigment
  - sampling without replacement, 395 < 10% all of the students
- Sample size / skew :
  - $n < 30 \rightarrow$ t-test , $n > 30 \rightarrow$ z-test
  - skewness : *Figure 46* shows no skewness and also by checking mean and median of *age* in *Figure 2*, we can see that mean and median are pretty much the same so we are good to go.

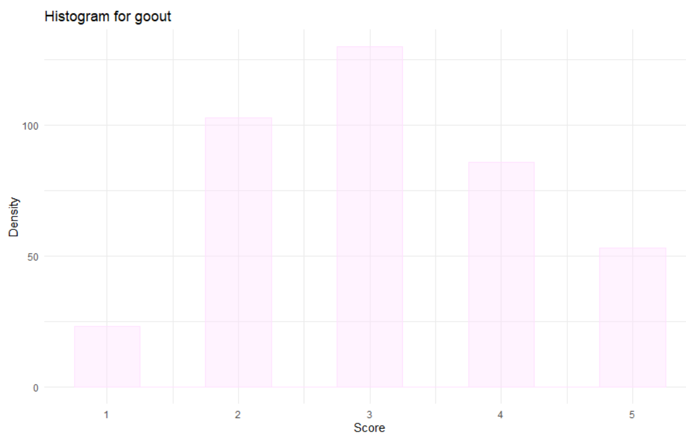


Figure 46: Histogram of goout

a.

Confidence intervals include the point estimate for the sample with a margin of error around the point estimate. The point estimate is the most likely value of the parameter and equals the sample value. The margin of error accounts for the amount of doubt involved in estimating the population parameter. The more variability there is in the sample data, the less precise the estimate, which causes the margin of error to extend further out from the point estimate.

Sample size = 25, t-test :

```
"Confidence Interval(using t-test) : ( 2.693 , 3.387 )"
```

Figure 47: Confidence Interval of goout using $\alpha = 5\%$

Sample size = 200, z-test :

```
"Confidence Interval(using z-test) : ( 2.92 , 3.23 )"
```

Figure 48: Confidence Interval of goout using $\alpha = 5\%$

b.

We are 95% confident that the the times these students goout are on average between 2.92 and 3.23 (according to *z-test*).
In other words, 95% of random samples of 395 students will yeild CIs that capture the true population mean of the times they goout.

c.

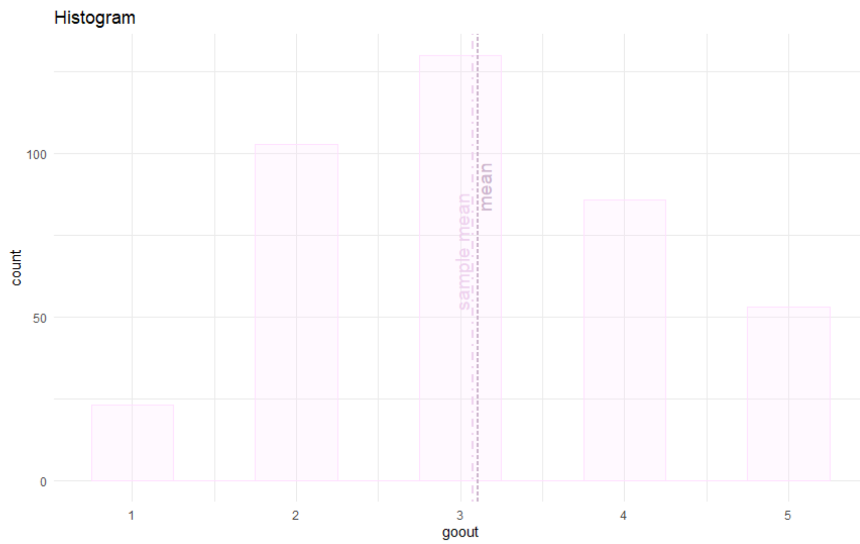


Figure 49: Histogram of goout marked with actual mean and sample mean

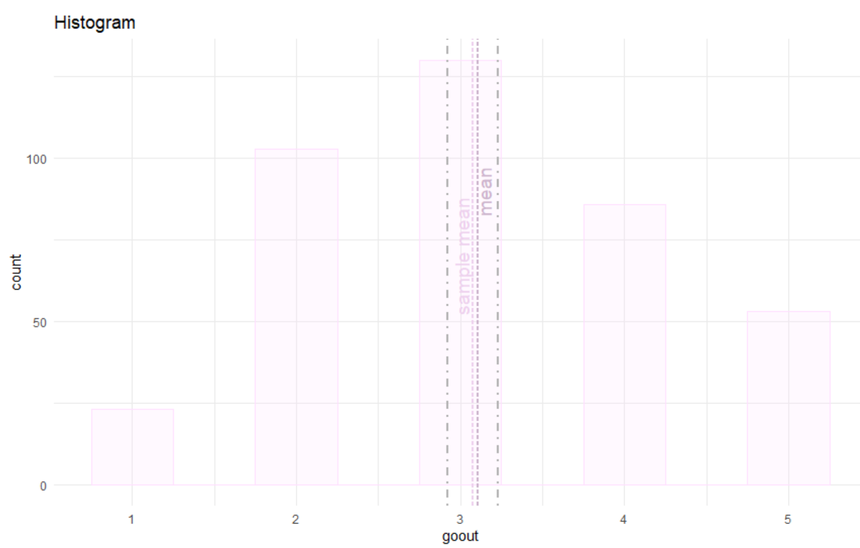


Figure 50: Histogram of goout marked with CI, actual mean and sample mean

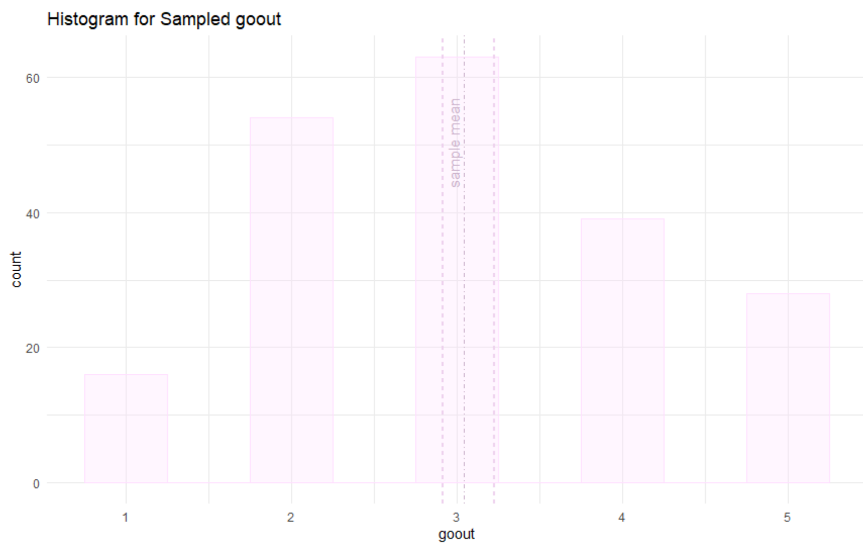


Figure 51: Histogram of sampled goout marked with CI and sample mean

d.

Hypothesis test :

$$H_0 : \mu = 2.8$$

$$H_A : \mu \neq 2.8$$

Sample size = 25, t-test :

```
>
> Hypothesis.test(goout.sampled.t, null.value = 2.8)
[1] "Null Hypothesis: mean =  2.8"
[1] "Alternative Hypothesis: mean /=  2.8"
[1] "Using t-distribution"
[1] "p-value = 0.0219829970441023"
[1] "Reject Null Hypothesis."
>
>
```

Figure 52: Hypothesis test of goout using $\alpha = 5\%$

Since *p-value* is 5% and is higher than 0.021, we should reject the null hypothesis in favor of the alternative hypothesis.

Sample size = 200, z-test :

```
>
> Hypothesis.test(goout.sampled, null.value = 2.8)
[1] "Null Hypothesis: mean =  2.8"
[1] "Alternative Hypothesis: mean /=  2.8"
[1] "Using Z-distribution"
[1] "p-value = 0.000135695892379579"
[1] "Reject Null Hypothesis."
>
>
```

Figure 53: Hypothesis test of goout using $\alpha = 5\%$

Since *p-value* is 5% and is higher than 0.00013, we should reject the null hypothesis in favor of the alternative hypothesis.
According to *Figure 52*, if the null hypothesis were true, there is only 2.1% (very tiny) chance that we would take a sample of size 25 and obtain a sample mean of 3.07.

According to *Figure 53*, if the null hypothesis were true, there is a tiny chance that we would take a sample of size 25 and obtain a sample mean of 3.07.

e.

P-value and *Confidence Interval* are two equivalent methods of interpreting results of a statistical analysis and their results *always agree*.
Both of these concepts specify a distance from the mean to a limit and these distances are precisely the same length.

f. and **g.**

The error that occurs when one accepts a null hypothesis that is actually false is the type II error. A type II error produces a false negative, also known as an error of omission.

$$\beta = P(H_0 \text{ is true } | H_0 \text{ is actually false})$$

```
>
> TypeIIerr(goout.sampled, null.value = 2.8)
[1] "TypeII error = % 2.4"
[1] "Power = % 97.6"
>
>
```

Figure 54: Power and typeII error of goout

Using **R**'s built-in function :

```
     One-sample t test power calculation

              n = 200
          delta = 0.3088608
             sd = 1.111837
      sig.level = 0.05
          power = 0.974388
    alternative = two.sided

>
```

Figure 55: Power and typeII error of goout

An effect size is closely related to a power of a statistical test because when *difference* of two groups is big, it is easy to reject the null hypothesis.
In other words, as the effect size gets larger, it is more likely to reject the null hypothesis; less likely to fail to reject the null hypothesis, thus the power of the test increases.

Question 6 continued on next page...

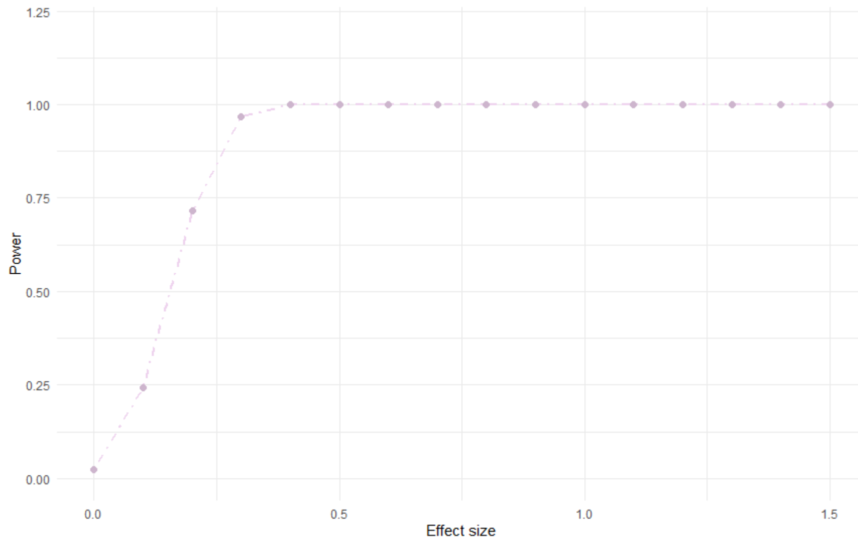


Figure 56: Relationship between effect size and power

Question 7

a.

Chosen Numerical Variable : *health* and *goout*

When two sets of observations have a special correspondence (they were chosen from one X in the dataset), they are said to be paired. To analyze paired data, it is useful to look at the difference in outcomes of each paired observation.

Check Condition :

- Independent Observations :
  - Random sample/assigment
  - sampling without replacement, $395 < 10\%$ all of the students
- Sample size / skew :
  - $n = 25 < 30 \rightarrow$ t-test. The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. Usually, t-tests are more appropriate when dealing with problems with a limited sample size .
  - skewness : As was mentioned in question6 (part a) *goout* is not skewed, health is a bit leftskewed.

```
>
> Hypothesis.test(StudentsPerformance.sampled$health, StudentsPerformance.sampled$goout, paired = TRUE)
[1] "Null Hypothesis: diff mean =  0"
[1] "Alternative Hypothesis: diff mean /=  0"
[1] "Using t-distribution"
[1] "p-value = 0.0384069445168378"
[1] "Reject Null Hypothesis."
>
>
```

Figure 57: Paired t-test between health and goout

Using **R**'s built-in function :

```
        Paired t-test

data:  StudentsPerformance.sampled$health and StudentsPerformance.sampled$goout
t = 2.1909, df = 24, p-value = 0.03841
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0463708 1.5536292
sample estimates:
mean of the differences
                    0.8
```

Figure 58: Paired t-test between health and goout

Since *p-value* is 5% and is higher than 0.038, we should reject the null hypothesis in favor of the alternative hypothesis. There is strong evidence that the null hypothesis is invalid.

Question 7 continued on next page. . .

b.

Check Condition :

- Independent Observations :
  - Random sample/assigment
  - sampling without replacement, 395 < 10% all of the students
- Sample size / skew :
  - $n = 100 > 30 \rightarrow$ z-test
  - skewness : As was mentioned in question6 (part a) *goout* is not skewed, health is a bit leftskewed but our sample size is big enough so we can ignore it.

```
>
> Hypothesis.test(health.sampled, goout.sampled)
[1] "Null Hypothesis: diff mean =  0"
[1] "Alternative Hypothesis: diff mean /=  0"
[1] "Using Z-distribution"
[1] "p-value = 0.0355069327255375"
[1] "Reject Null Hypothesis."
>
>
```

Figure 59: z-test between health and goout

Using **R**'s built-in function :

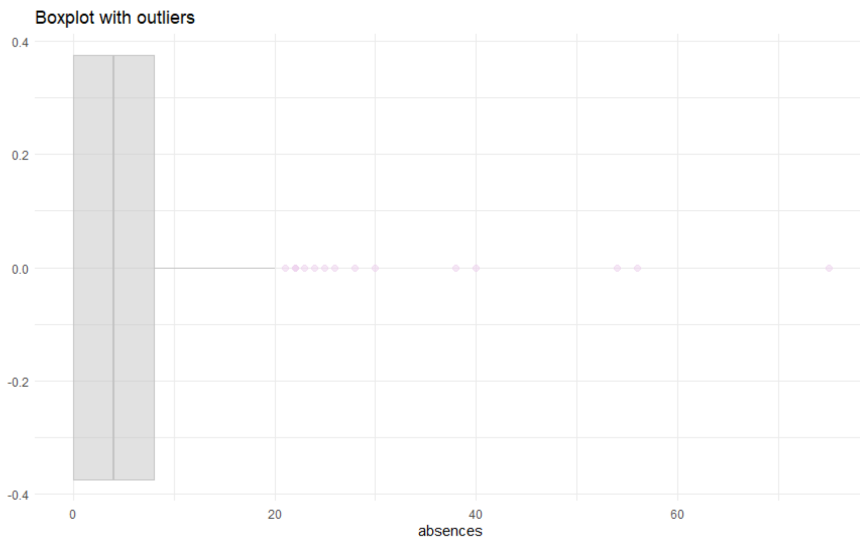
```
        Welch Two Sample t-test

data:  health.sampled and goout.sampled
t = 2.1025, df = 186.83, p-value = 0.03685
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02469156 0.77530844
sample estimates:
mean of x mean of y
     3.48      3.08

> |
```

Figure 60: z-test between health and goout

Confidence Interval : $0 \notin [0.024, 0.77] \rightarrow$ Reject the null hypothesis.
P-value and *Confidence Interval* are two equivalent methods of interpreting results of a statistical analysis and their results *always agree*.

Question 8

Chosen Numerical Variable : *absences*




```
$stats
[1]  0  0  4  8 20

$n
[1] 395

$conf
[1] 3.364013 4.635987

$out
 [1] 25 54 26 56 24 28 22 21 75 22 30 38 22 40 23
```

Figure 61: boxplot of absences with stat

Using the normal approximation might not be good in all applications where the sample size is at least 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

a.

Using quantile doesn't seem like a good idea as can be deducted from *Figure 62*, so 100 samples were chosen and replicated 1000 times, and the interval for their mean can be seen in *Figure*.

```
>
> quantile(StudentsPerformance$absences, c(0.025, 0.975))
 2.5% 97.5%
 0.00 23.15
>
>
```

Figure 62: Simple percentile method

Question 8 continued on next page. . .

```
"Confidence Interval: ( 5.18 , 7.74 )"
```

Figure 63: Percentile method

b.

A random sample with replacement was taken from the original sample.Bootstrap statistic (*mean* in our case) was computed on bootstrap samples and these steps was repeated to create a bootstrap distribution. The middle 95% of the bootstrap distribution was calculated for CI :

```
"Confidence Interval: ( 5.56 , 6.22 )"
```

Figure 64: Percentile method (bootstrapped)

c.

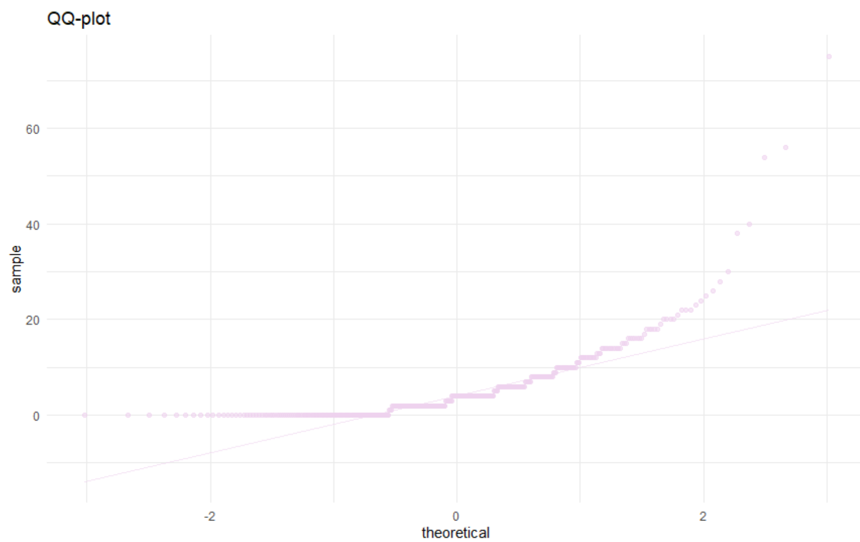


Figure 65: QQ-plot of absences

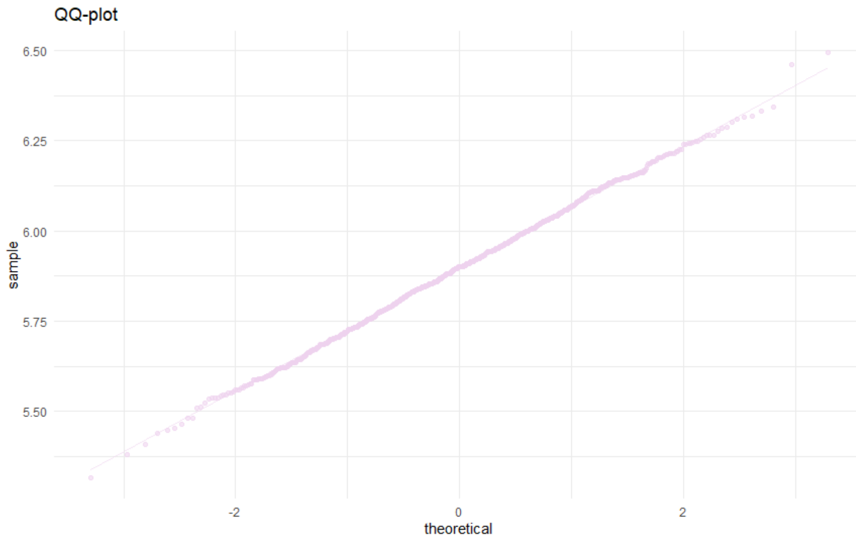


Figure 66: QQ-plot of mean of bootstrapped samples

Percentile method is a method which is sensitive to outliers; so, the calculated interval might not be as informative as we desired. Therefore, bootstrapping method was used in order to remove outliers and result a (approximately) normal distribution (*Figure 66*). (A better approach is using SD method which is more robust when facing outliers)

Knowing these facts and figures, we can conclude that *bootstrapping* is a stronger procedure and a more informative CI is the proof of it.

Question 9

In *ANOVA*, the *null hypothesis* is that there is no difference among group means. If any group differs significantly from the overall group mean, then the *ANOVA* will report a statistically significant result.

In our case :

$$H_0 : \mu_{failure=0,G_1+G_2+G_3} = \mu_{failure=1,G_1+G_2+G_3} = \mu_{failure=2,G_1+G_2+G_3} = \mu_{failure=3,G_1+G_2+G_3}$$

$$H_A : \textit{ one group differs significantly from the overall group mean}$$

Significant differences among group means are calculated using the *F statistic*, which is the ratio of the mean sum of squares (explained variable) to the mean square error (unexplained variable) .

If the *F statistic* is higher than the alpha value (0.05), then the difference among groups is deemed statistically significant.

Degrees of freedom associated with *ANOVA* :

$$df_T = n - 1 \quad , \quad df_G = k - 1 \quad , \quad df_E = df_T - df_G = n - k$$

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SSG = \sum_{j=1}^{k} n_j (\bar{x_j} - \bar{x})^2 \Rightarrow MSG = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{x_j} - \bar{x})^2$$

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x_j} - \bar{x}) = \sum_{j=1}^{k} (n_j - 1) s_j^2 \Rightarrow MSE = \frac{1}{n-k} \sum_{j=1}^{k} (n_j - 1) s_j^2$$

$$F = \frac{\textit{Variability bet. groups}}{\textit{Variability w/in groups}} = \frac{MSG}{MSE}$$

Check Condition :

- Independence :
  - within groups: sampled observations are independent
  - between groups: the groups are independent of each other (non-paired)
- Approximate normality : distributions should be nearly normal within each group → we assume they are

- Equal variance : groups should have roughly equal variability

```
>
> sd.df
  groups       sds
1 Group0 10.276562
2 Group1 10.082132
3 Group2 10.556222
4 Group3  6.172193
>
>
```

Figure 67: SD of each group

The standard deviation of group0, group1 and group2 are close to each other, but the one for group3 is different from others. Although this could happen because of the low group size, we can consider these three numbers as almost the same.

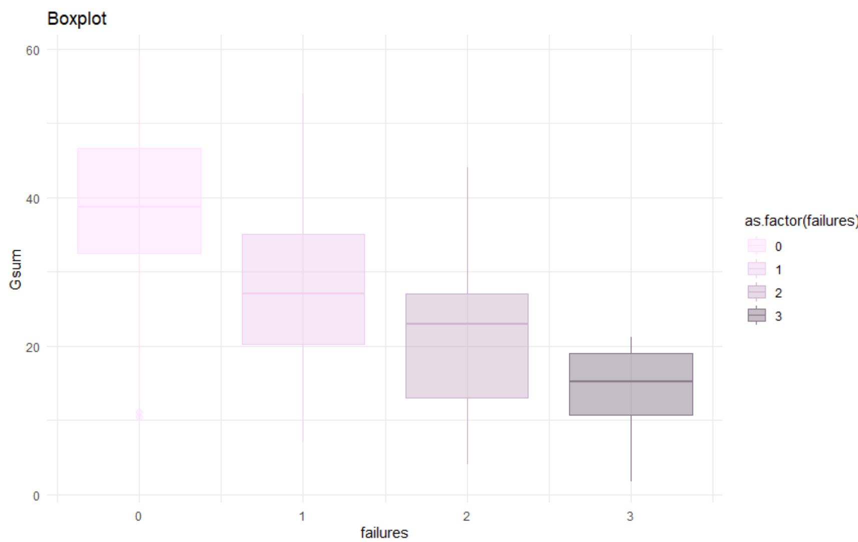


Figure 68: Boxplot grouped by number of failures

```
>
> summary(aov.Gsum_failures)
                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(failures)   3  17949    5983   58.22 <2e-16 ***
Residuals           391  40179     103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
```

Figure 69: ANOVA table

Since *p-value* is smaller than 0.05, we reject the null Hypothesis.
The data provides convincing evidence that at least one pair of population means are different from each other.

ANOVA tells us if there are differences among group means, but *not what the differences* are. To find out which groups are statistically different from one another, you can perform a *Tukey's Honestly Significant*

Difference (Tukey's HSD) post-hoc test for pairwise comparisons.

The significant groupwise differences are any where the 95% confidence interval doesn't include zero. In other words, p-value for these *pairwise differences* is < 0.05.

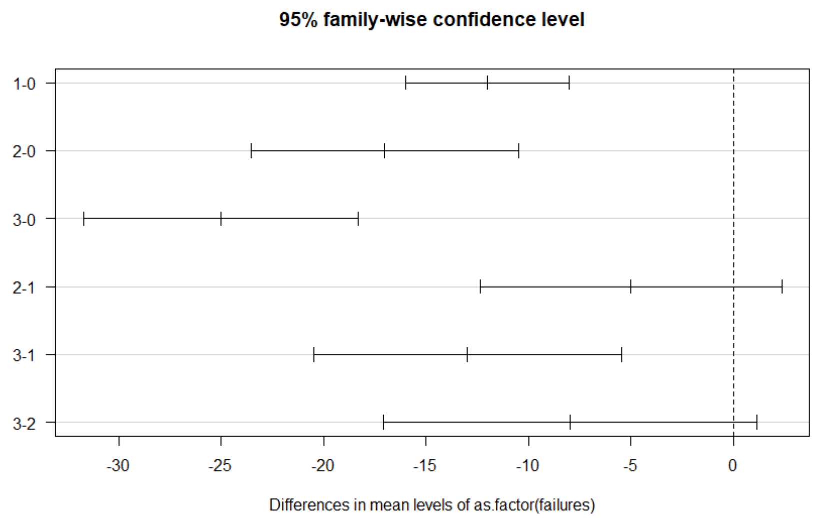


Figure 70: Pairwire confidence level(0.95%)

R Codes

```
library(magrittr)
library(ggfortify)
library(ggplot2)
library(plyr)
library(gridExtra)
require(qqplotr)
library(moments)
library(hexbin)
library(ggmosaic)
library("plot3D")
library(plotly)
library(scatterplot3d)
library(RNHANES)
library(GGally)
library(dplyr)
library(Hmisc)
require(ggpubr)
require(Hmisc)
require(corrplot)
library(patchwork)
library(ggExtra)


theme_set(theme_minimal())

summary(StudentsPerformance)

#Question 0

missingvalues <- colSums(is.na.data.frame(StudentsPerformance))
missingvalues.proporion <- missingvalues/nrow(StudentsPerformance)
plot(missingvalues.proporion , main = "Percentage of missing values vs. variables",
     xlab = "variables", ylab = "Missing values proportion" , type = 'l', col = 'thistle')

missingvalues.proporion <- data.frame("missing value", missingvalues/nrow(
    StudentsPerformance))


#QUESTION 1

#Numerical value chosen : Grade 1

StudentsPerformance$G1

#a.
breaks <- pretty(StudentsPerformance$G1, n = nclass.FD(StudentsPerformance$G1), min.n = 0)
bwidth <- breaks[2]-breaks[1]


G1_hist <- ggplot(StudentsPerformance, aes(x = G1)) +
  geom_histogram(aes(y=..density..) , binwidth = bwidth, alpha = 0.4, color="thistle1", fill
      ="thistle1") +
  geom_density(color = "gray87",linetype="dashed", fill = "gray87" , alpha = 0.3, size=1) +
  #stat_function(fun = dnorm, n = 101, args = list(mean = mu, sd = std) , color = "gray87" ,
      size=1) +
  labs(title = "Histogram for G1", x = "Score", y="Density")


G1_hist
#---
```

```
#b.
G1.qq <- ggplot(StudentsPerformance, aes(sample = G1, color = "", alpha = 0.7)) + geom_qq()
    +
  geom_qq_line() + labs(title="QQ-plot for G1")

G1.qq + theme(legend.position="none") + scale_color_manual(values=c("thistle2"))
#---

#c.
print(skewness(StudentsPerformance$G1))

G1_hist + geom_vline(xintercept = mean(StudentsPerformance$G1), linetype="dashed", color = "
    thistle4", size = 0.5) +
  geom_vline(xintercept = median(StudentsPerformance$G1), linetype="dotdash", color = "
      gray29", size = 0.5)+
  annotate("text", x = mu - .2 , label = "mean", y = 0.01, size = 3.4, angle = 90 , color =
      'thistle4') +
  annotate("text", x = median + 0.1 , label = "median", y = 0.06, size = 3.4, angle = 90,
      color = 'gray29')

#---

#d.
G1_box <- ggplot(StudentsPerformance, aes(x = G1)) + geom_boxplot(color ="thistle2", fill ="
    thistle2", alpha = 0.5) +
  labs(title="Boxplot for G1")

G1_box + theme(legend.position="none")
#---

#e.
mu <- mean(StudentsPerformance$G1)
median <- median(StudentsPerformance$G1)
var <- var(StudentsPerformance$G1)
std <- sd(StudentsPerformance$G1)
#---

#f.
G1_density <- ggplot(StudentsPerformance, aes(x = G1)) +
  geom_vline(xintercept = mu, linetype="dashed", color = "gray29") +
  geom_vline(xintercept = median, linetype="dashed", color = "thistle4") +
  geom_density(color = "thistle1", size = 1) +
  stat_function(fun = dnorm, n = 101, args = list(mean = mu, sd = std) , color = "thistle2",
       size = 1) +
  annotate("text", x = mu - .2 , label = "mean", y = 0.01, size = 3.4, angle = 90 , color =
      'gray29') +
  annotate("text", x = median + 0.1 , label = "median", y = 0.06, size = 3.4, angle = 90,
      color = 'thistle4')+
  labs(title="Density for G1")

G1_density
#---

#g.

#method1
StudentsPerformance$categorizedG1 <- ifelse(StudentsPerformance$G1 > (mu + max(
    StudentsPerformance$G1))/2, 'very high', ifelse(StudentsPerformance$G1 > mu, 'high',
    ifelse(StudentsPerformance$G1 > mu/2, 'low', 'very low')))

freq_vlow <-length(which(StudentsPerformance[17] == 'very low')) / length(
```

```
StudentsPerformance$G1)
freq_low <- length(which(StudentsPerformance[17] == 'low'))/ length(StudentsPerformance$G1)
freq_high <- length(which(StudentsPerformance[17] == 'high'))/ length(StudentsPerformance$G1
)
freq_vhigh <- length(which(StudentsPerformance[17] == 'very high'))/ length(
    StudentsPerformance$G1)


G1.categorized <- data.frame(group = c("Very Low", "Low", "High", "Very High"),
                  value = c(freq_vlow, freq_low, freq_high, freq_vhigh))



G1.pie <- ggplot(G1.categorized, aes(x="", y = value, fill = group)) +
  geom_bar(stat = "identity", alpha = 0.7) + coord_polar("y")


G1.pie  + scale_fill_manual(values = c("thistle1", "thistle2", "thistle3", "thistle4")) +
  geom_text(aes(label = paste0(round(value*100), "%")), position = position_stack(vjust =
      0.5)) +
  labs(title="Pie-chart of G1", x = 'Frequency', y = 'G1')

#method2

G1.quant <- quantile(StudentsPerformance$G1)

StudentsPerformance$categorizedG1 <- ifelse(StudentsPerformance$G1 > G1.quant[[4]], 'very
    high', ifelse(StudentsPerformance$G1 >  G1.quant[[3]], 'high', ifelse(
    StudentsPerformance$G1 >  G1.quant[[2]], 'low', 'very low')))

freq_vlow <-length(which(StudentsPerformance[17] == 'very low')) / length(
    StudentsPerformance$G1)
freq_low <- length(which(StudentsPerformance[17] == 'low'))/ length(StudentsPerformance$G1)
freq_high <- length(which(StudentsPerformance[17] == 'high'))/ length(StudentsPerformance$G1
)
freq_vhigh <- length(which(StudentsPerformance[17] == 'very high'))/ length(
    StudentsPerformance$G1)


G1.categorized <- data.frame(group = c("Very Low", "Low", "High", "Very High"),
                            value = c(freq_vlow, freq_low, freq_high, freq_vhigh))



G1.pie <- ggplot(G1.categorized, aes(x="", y = value, fill = group)) +
  geom_bar(stat = "identity", alpha = 0.7) + coord_polar("y")


G1.pie  + scale_fill_manual(values = c("thistle1", "thistle2", "thistle3", "thistle4")) +
  geom_text(aes(label = paste0(round(value*100), "%")), position = position_stack(vjust =
      0.5)) +
  labs(title="Pie-chart of G1", x = 'Frequency', y = 'G1')


#-----

#h.

boxplot.stats(StudentsPerformance$G1)

G1.quant <- quantile(StudentsPerformance$G1)
G1.iqr <- IQR(StudentsPerformance$G1)
```

```r
#-----
#------------------------------------------------------------

#QUESTION 2

#Categorical Variable chosen : sex

StudentsPerformance$sex

#a.
female.freq <- length(((StudentsPerformance %>% filter(sex == 'F'))$sex)) / length(StudentsPerformance$sex)
male.freq <- length(((StudentsPerformance %>% filter(sex == 'M'))$sex)) / length(StudentsPerformance$sex)
#-----

#StudentsPerformance.se1 <- subset(StudentsPerformance, sex == "F")


#b.

#freq <-data.frame(female.freq, male.freq)
sex.barplot <- ggplot(StudentsPerformance, aes(x = " ", color = sex, fill = sex)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), alpha = 0.5, width = 0.5) + labs(title="Stacked Barplot of sex", y = 'Frequency')

sex.barplot + scale_color_manual(values = c("thistle1", "gray67")) + xlab("sex") +
  scale_fill_manual(values = c("thistle1", "gray67")) + scale_y_continuous(labels = scales::percent) +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))),
            stat = "count", hjust = 0.5, size = 4.5, color = 'black', vjust = 1.4, position = position_dodge(width = 0.3))
#-----

#c.
categorizedsex.barplot <- ggplot(StudentsPerformance, aes(x = sex, color = sex, fill = sex)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), alpha = 0.7) + labs(title="Barplot of sex", y = 'Frequency')

categorizedsex.barplot + scale_color_manual(values = c("thistle1", "gray67")) +
  scale_fill_manual(values =c("thistle1", "gray67")) + coord_flip() + scale_x_discrete(limits=c("M", "F"))+
  geom_text(aes(y = ((..count..)/sum(..count..)), label = round(((..count..)/sum(..count..)), 3)),
            stat = "count", vjust = -0.25, size = 4, color = 'black') + theme(axis.text.x=element_blank())
#-----

#d.
sex.df <- data.frame(sex = c("F", "M"), frequency = c(female.freq, male.freq))
sex.violinplot <- ggplot(StudentsPerformance, aes(x = sex, y = age, color = sex, fill = sex)) +
  geom_violin( trim=FALSE, alpha = 0.7) + labs(title="Violinplot")

sex.violinplot+ scale_color_manual(values = c("thistle1", "gray67")) + xlab("sex") +
  scale_fill_manual(values = c("thistle1", "gray67"))
#-----
#------------------------------------------------------------

#QUESTION 3
```

```
#Numerical Variable chosen : goout and absences -> it actually depends


#b.
goout_absences.scatterplot <- ggplot(StudentsPerformance, aes(x = goout, y = absences)) +
  geom_point(color = "thistle2", size = 2)

goout_absences.scatterplot + labs(title="Scatterplot of Going out and Absences")
#----

#c.
goout_absences.correlation <- cor(StudentsPerformance$goout, StudentsPerformance$absences)
goout_absences.correlation
#----

#c.
ggscatter(StudentsPerformance, x = "goout", y = "absences", shape = 12, add = "reg.line",
    conf.int = TRUE,
          color = "thistle2", add.params = list(color = "thistle3", fill = "gray90"), cor.
              coef = TRUE,
          cor.coeff.args = list(method = "pearson", label.x = 3, label.sep = "\n")) + theme_
              minimal() +
  labs(title="Scatterplot")

#----

#f.
goout_absences_romantic.scatterplot <- ggplot(StudentsPerformance,
                                              aes(x = goout, y = absences, color = romantic,
                                                  shape = romantic)) +
  geom_point(size = 2) + labs(title="Scatterplot")

goout_absences_romantic.scatterplot +  scale_shape_manual(values = c(4, 16)) +
  scale_color_manual(values=c('thistle4','thistle2'))
#----

#g.
breaks <- pretty(StudentsPerformance$goout, n = nclass.FD(StudentsPerformance$goout), min.n
    = 0)
goout.hist <- ggplot(StudentsPerformance, aes(x = goout)) + geom_histogram(binwidth = breaks
    [2]-breaks[1],color = "thistle2", fill = "thistle2", alpha = 0.5) +theme_void()

breaks <- pretty(StudentsPerformance$absences, n = nclass.FD(StudentsPerformance$absences),
    min.n = 0)
absences.hist <- ggplot(StudentsPerformance, aes(x = absences)) +
  geom_histogram(binwidth = breaks[2]-breaks[1], color = "thistle2", fill = "thistle2",
      alpha = 0.5) + coord_flip() + theme_void()


gar.hexbinplot.log <- ggplot(StudentsPerformance, aes(x = goout, y = log2(absences + 1))) +
  geom_hex(bins = 10, color = "white", alpha = 0.7) + scale_fill_gradient(low =  "thistle1",
       high = "thistle4" ,trans="log10")
#+ geom_smooth(col = 'grey40')

goout.hist + plot_spacer() + gar.hexbinplot.log + absences.hist +
  plot_layout( ncol = 2, nrow = 2, widths = c(4, 1), heights = c(1, 4))

gar.hexbinplot <- ggplot(StudentsPerformance, aes(x = goout, y = absences)) +
  geom_hex(bins = 10, color = "white", alpha = 0.7) + scale_fill_gradient(low =  "thistle1",
       high = "thistle4") +
  geom_smooth(method = "loess", col = 'grey40')
```

```r

goout.hist + plot_spacer() + gar.hexbinplot + absences.hist +
  plot_layout( ncol = 2, nrow = 2, widths = c(4, 1), heights = c(1, 4))


#---

#h.
gar.2ddensity <- ggplot(StudentsPerformance, aes(x = goout, y = G1)) +
  stat_density2d(aes(fill = ..level..), geom = "polygon", alpha = 0.5) + lims(x = c(0,6), y
      = c(-5, 20))

gar.2ddensity + scale_fill_gradient(low = "thistle1", high = "thistle4") +
  labs(title="2D-density plot")
#---
#---------------------------------------------

#QUESTION 4

#a.
ggpairs(dplyr::select_if(StudentsPerformance, is.numeric), title = "Correlogram")


#density, without failure
ggpairs(StudentsPerformance[, c(4, 7, 10, 12, 13, 14, 15, 16)],
        upper = list(continuous = wrap("density", colour="thistle")),
        lower = list(continuous = wrap("points", colour="grey70")))

#linear relationship
ggpairs(StudentsPerformance[, c(4, 7, 10, 11, 12, 13, 14, 15, 16)],
        upper = list(continuous = wrap("smooth", colour="thistle")),
        lower = list(continuous = wrap("points", colour="grey70")))
#barplot
ggpairs(StudentsPerformance[, c(4, 7, 10, 11, 12, 13, 14, 15, 16)],
        upper = list(continuous = wrap("barDiag", colour="thistle", fill = "thistle", alpha
            = 0.5)),
        lower = list(continuous = wrap("points", colour="grey70")))
#boxplot
ggpairs(StudentsPerformance[, c(4, 7, 10, 11, 12, 13, 14, 15, 16)],
        upper = list(continuous = wrap("box_no_facet", colour="thistle", fill = "thistle3",
            alpha = 0.5)),
        lower = list(continuous = wrap("points", colour="grey70")))

#---
#b.

col <- colorRampPalette(c("grey80", "white", "thistle1", "thistle2"))
StudentsPerformance.corr <- rcorr(as.matrix(dplyr::select_if(StudentsPerformance, is.numeric
    )))
StudentsPerformance.corr.p <- StudentsPerformance.corr$P
StudentsPerformance.corr.p[is.na(StudentsPerformance.corr.p)] <- 1

M <- cor(dplyr::select_if(StudentsPerformance, is.numeric))

corrplot(M, method = "color", col = col(200), type = "upper", order = "hclust", addCoef.col
    = "black",
          tl.col = "thistle4", tl.srt = 45, p.mat = StudentsPerformance.corr.p, sig.level =
              0.05, diag = FALSE)
#---

#c.
cols <- c("thistle2", "grey50")
```

```r

with(StudentsPerformance, scatterplot3d(age, health, failures, main="3D scatterplot",
                  pch = 16, color = cols[as.numeric(StudentsPerformance$sex)]))

legend("right", legend = levels(StudentsPerformance$sex),
       col =  c("thistle2", "grey50"), pch = 16)

#--------------------------------------------------

#Question 5

#Chosen : sex and romantic
#a.
table <- addmargins(table(StudentsPerformance$romantic, StudentsPerformance$sex), c(1,2))
print.table(table)
#-----


#b.

romantic_sex.groupedbarplot <- ggplot(StudentsPerformance, aes(x = romantic,color = sex, fill = sex)) +
  geom_bar(position = "dodge", alpha = 0.5) + labs(title="Grouped barplot", x="romantic")

romantic_sex.groupedbarplot + scale_color_manual(values = c("thistle1", "gray67")) +
  scale_fill_manual(values = c("thistle1", "gray67")) +
  geom_text(aes(y = ..count.., label = ..count..), stat = "count", vjust = -0.25,  size = 4, color = 'black', position = position_dodge(width = 1))


romantic_sex.groupedbarplot <- ggplot(StudentsPerformance, aes(x = romantic,color = sex, fill = sex)) +
  geom_bar(alpha = 0.5, width = 0.5) + labs(title="Segmented barplot", x="romantic")

romantic_sex.groupedbarplot + scale_color_manual(values = c("thistle2", "gray68")) +
  scale_fill_manual(values = c("thistle1", "gray67")) +
  annotate("text", x = 1 , label = "134", y = 70, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 1 , label = "129", y = 200, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 2 , label = "53", y = 25, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 2 , label = "79", y = 90, size = 4, angle = 0 , color = 'gray29')


romantic_sex.mosaicplot <- ggplot(StudentsPerformance) +
  geom_mosaic(aes(x = product(romantic), fill = sex), alpha = 0.5) + labs(title="Mosaicplot", y = "Frequenct") +
  scale_y_continuous(labels = scales::percent) +
  annotate("text", x = 0.33 , label = "49%", y = .25, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 0.33 , label = "51%", y = .75, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 0.83 , label = "62%", y = .3, size = 4, angle = 0 , color = 'gray29') +
  annotate("text", x = 0.83 , label = "38%", y = .8, size = 4, angle = 0 , color = 'gray29')



romantic_sex.mosaicplot + scale_fill_manual(values = c("thistle1", "gray67"))

#-----
#--------------------------------------------------

#Question 6
```

```
#Chosen Numerical Variable: age

breaks <- pretty(StudentsPerformance$goout, n = nclass.FD(StudentsPerformance$goout), min.n
    = 0)
bwidth <- breaks[2]-breaks[1]


goout_hist <- ggplot(StudentsPerformance, aes(x = goout)) +
  geom_histogram(binwidth = bwidth, alpha = 0.4, color="thistle1", fill="thistle1") +
  labs(title = "Histogram for goout", x = "Score", y="Density")
goout_hist

#a.
CI.calculate <- function(data.sampled, alpha = 0.05){

  sample.len <- length(data.sampled)

  mu <- mean(data.sampled)
  s <- sd(data.sampled)
  SE <- s/sqrt(sample.len)

  if(sample.len > 30){
    print("Using Z-distribution")
    Zstar <- abs(qnorm(alpha/2))
    error.margin <- Zstar * SE}

  else{
    print("Using t-distribution")
    tstar <- abs(qt(alpha/2, df = sample.len - 1))
    error.margin <- tstar * SE }
  confidence.interval <- c(mu - error.margin, mu + error.margin)
  return(confidence.interval)

}

goout.sampled.t <- sample(StudentsPerformance$goout, 25)
confidence.interval.t <- CI.calculate(goout.sampled.t)
print(paste("Confidence Interval(using t-test) : (", round(confidence.interval.t[1], 3),",",
    round(confidence.interval.t[2],3),")"))


goout.sampled <- sample(StudentsPerformance$goout, 200)
confidence.interval <- CI.calculate(age.sampled)
print(paste("Confidence Interval(using z-test) : (", round(confidence.interval[1], 3), ",",
    round(confidence.interval[2], 3),")"))
#---

#c.

goout.hist <- ggplot(StudentsPerformance, aes(x = goout)) +
  geom_histogram(binwidth = bwidth, alpha = 0.2, color="thistle1", fill="thistle1") +
  labs(title = "Histogram", x = "goout") +
  geom_vline(xintercept = mean(StudentsPerformance$goout), color = "thistle3", linetype="21
    ", size = 0.8) +
  geom_vline(xintercept = mean(goout.sampled), color = "thistle2", linetype="dotdash", size
    = 0.8) +
  annotate("text", x = mean(StudentsPerformance$goout) + 0.03, label = "mean", y = 90, size
    = 5 , angle = 90, color = "thistle3") +
  annotate("text", x = mean(goout.sampled) - 0.07, label = "sample mean", y = 70, size = 5 ,
    angle = 90, color = "thistle2")

goout.hist
```

```


goout.hist <- ggplot(StudentsPerformance, aes(x = goout)) +
  geom_histogram(binwidth = bwidth, alpha = 0.2, color="thistle1", fill="thistle1") +
  labs(title = "Histogram", x = "goout") +
  geom_vline(xintercept = mean(StudentsPerformance$goout), color = "thistle3", linetype="21", size = 0.8) +
  geom_vline(xintercept = mean(goout.sampled), color = "thistle2", linetype="21", size = 0.8) +
  geom_vline(xintercept = round(confidence.interval[1], 3), color = "grey70", linetype="dotdash", size = 1) +
  geom_vline(xintercept = round(confidence.interval[2], 3), color = "grey70", linetype="dotdash", size = 1) +
  annotate("text", x = mean(StudentsPerformance$goout) + 0.03, label = "mean", y = 90, size = 5 , angle = 90, color = "thistle3") +
  annotate("text", x = mean(goout.sampled) - 0.07, label = "sample mean", y = 70, size = 5 , angle = 90, color = "thistle2")

goout.hist



breaks <- pretty(goout.sampled, n = nclass.FD(goout.sampled), min.n = 0)
bwidth <- breaks[2]-breaks[1]

goout.df <- data.frame(goout.sampled)
sampled.goout.hist <- ggplot(goout.df, aes(x = goout.sampled)) +
  geom_histogram(binwidth = bwidth, alpha = 0.3, color="thistle1", fill="thistle1") +
  labs(title = "Histogram for Sampled goout", x = "goout") +
  geom_vline(xintercept = mean(goout.sampled), color = "thistle3", linetype="dotdash", size = 0.5) +
  geom_vline(xintercept = confidence.interval[1], color = "thistle2", linetype="22", size = 1) +
  geom_vline(xintercept = confidence.interval[2], color = "thistle2", linetype="22", size = 1) +
  annotate("text", x = mean(goout.sampled) - 0.07, label = " sample mean", y = 50, size = 4 , angle = 90, color = "thistle3")


sampled.goout.hist


#----

#d.
Hypothesis.test <- function(data.sampled, null.value, alpha = 0.05){

  sample.len <- length(data.sampled)
  print(paste("Null Hypothesis: mean = ", null.value))
  print(paste("Alternative Hypothesis: mean /= ", null.value))

  x_bar <- mean(data.sampled)
  s <- sd(data.sampled)
  SE <- s/sqrt(sample.len)
  score <- abs((x_bar - null.value)) / SE


  if(sample.len > 30){
    print("Using Z-distribution")
    pvalue <- 2*pnorm(score, lower.tail = FALSE)}

  else{
```

```
    print("Using t-distribution")
    pvalue <- 2*pt(score, df = sample.len - 1, lower.tail = FALSE)}


  print(paste("p-value =", pvalue))

  if (pvalue < alpha)
    print("Reject Null Hypothesis.")
  else
    print("Fail to Reject Null Hypothesis.")

}
mean(goout.sampled)

Hypothesis.test(goout.sampled.t, null.value = 2.8)

Hypothesis.test(goout.sampled, null.value = 2.8)
#----

#f.and #g.
TypeIIerr <- function(data.sampled, null.value, alpha = 0.05){
  sample.len <- length(data.sampled)
  mean.actual <- mean(StudentsPerformance$goout)
  s <- sd(data.sampled)
  SE <- s/sqrt(sample.len)
  ME <- abs(qnorm((alpha/2))) * SE
  errorTypeII <- pnorm(abs(null.value + ME - mean.actual)/SE, lower.tail = F) +
    pnorm(abs(null.value - ME - mean.actual)/SE, lower.tail = F)

  print(paste("TypeII error = %", 100*round(errorTypeII,3)))
  print(paste("Power = %", 100*round(1-errorTypeII,3)))

}

TypeIIerr(goout.sampled, null.value = 2.8)

power.t.test(n = 200, delta = mean(StudentsPerformance$goout) - 2.8, sd = sd(goout.sampled),
    type="one.sample", alternative="two.sided")




differences <- seq(from = 0,to = 1.5,by = 0.1)
power.effect <- sapply(differences, function(d){power.t.test(n = 200, delta = d, sd = sd(
    goout.sampled), type="one.sample")}$power)

df <- data.frame(differences, power.effect)

ggplot(data = df, aes(x = differences, y = power.effect)) + ylim(c(0, 1.2)) +
  geom_line(linetype="dotdash", color="thistle2", size=1)+ ylab("Power") + xlab("Effect size
    ") +
  geom_point(color="thistle3", size = 2)


#----
#------------------------------------------------

#Question 7

#a. b)


```

```
StudentsPerformance.sampled <- sample_n(StudentsPerformance, 25)

Hypothesis.test <- function(data.sampled.var1, data.sampled.var2, null.value = 0, alpha = 0.05, paired = FALSE){

  sample.len <- length(data.sampled.var1)
  print(paste("Null Hypothesis: diff mean = ", null.value))
  print(paste("Alternative Hypothesis: diff mean /= ", null.value))

  x_bar <- mean(data.sampled.var1) - mean(data.sampled.var2)
  s1 <- sd(data.sampled.var1)
  s2 <- sd(data.sampled.var2)
  if (paired)
    SE <- sd(data.sampled.var1 - data.sampled.var2) / sqrt(sample.len)
  else
    SE <- sqrt((s1^2/sample.len) + (s2^2/sample.len))
  score <- abs((x_bar - null.value)) / SE

  if(sample.len > 30){
    print("Using Z-distribution")
    pvalue <- 2*pnorm(abs(score), lower.tail = FALSE)}

  else{
    print("Using t-distribution")
    pvalue <- 2*pt(score, df = sample.len - 1, lower.tail = FALSE)}


  print(paste("p-value =", pvalue))

  if (pvalue < alpha)
    print("Reject Null Hypothesis.")
  else
    print("Fail to Reject Null Hypothesis.")

}


Hypothesis.test(StudentsPerformance.sampled$health, StudentsPerformance.sampled$goout, paired = TRUE)

t.test(StudentsPerformance.sampled$health, StudentsPerformance.sampled$goout, paired = TRUE)

#----

#b.


idx.sampled <- sample(StudentsPerformance$X, 200)
health.sampled <- StudentsPerformance$health[idx.sampled[1:100]]
goout.sampled <- StudentsPerformance$goout[idx.sampled[1:100]]

Hypothesis.test(health.sampled, goout.sampled)

t.test(health.sampled, goout.sampled)

#----
#--------------------------------------------------

#Question 8
```

```r
absences_box <- ggplot(StudentsPerformance, aes(x = absences)) +
  geom_boxplot(outlier.colour="thistle2", color ="gray77", fill ="gray77", alpha = 0.5, outlier.size = 2) +
  labs(title="Boxplot with outliers")

absences_box
boxplot.stats(StudentsPerformance$absences)



#a.

quantile(StudentsPerformance$absences, c(0.025, 0.975))



bs.size <- 1000
rep.size <- 1000

absences.sample <- replicate(1, sample(StudentsPerformance$absences, size = 200, replace = FALSE))
absences.replicated <- replicate(rep.size, sample(absences.sample, size = 100, replace = FALSE))

means <- apply(X = absences.replicated, MARGIN = 2, FUN = mean, na.rm = TRUE)

means <- sort(means)
margin <- 0.025 * bs.size

print(paste("Confidence Interval: (", round(means[c(margin)], 3),",",round(means[c(bs.size - margin)],3),")"))
#---

#b.
bs.size <- 1000
rep.size <- 1000

absences.sample <- replicate(1, sample(StudentsPerformance$absences, size = 20, replace = FALSE))
absences.bootstrapped <- replicate(rep.size, sample(absences.sample, size = 1000, replace = TRUE))

means <- apply(X = absences.bootstrapped, MARGIN = 2, FUN = mean, na.rm = TRUE)

means <- sort(means)
margin <- 0.025 * bs.size

print(paste("Confidence Interval: (", round(means[c(margin)], 3),",",round(means[c(bs.size - margin)],3),")"))
#---

#c.
absences.qq <- ggplot(StudentsPerformance, aes(sample = absences, color = "", alpha = 0.7)) + geom_qq() +
  geom_qq_line() + labs(title="QQ-plot ")

absences.qq + theme(legend.position="none") + scale_color_manual(values=c("thistle2"))


m.absences.qq <- ggplot(data.frame(mean = means), aes(sample = means, color = "", alpha = 0.7)) + geom_qq() +
  geom_qq_line() + labs(title="QQ-plot ")
```

```r

m.absences.qq + theme(legend.position="none") + scale_color_manual(values=c("thistle2"))

#----
#-------------------------------------------------------


#Question 9


StudentsPerformance$Gsum <- StudentsPerformance$G1 + StudentsPerformance$G2 +
    StudentsPerformance$G3


f0.Gsum <- ((StudentsPerformance %>% filter(failures == 0))$Gsum)
f1.Gsum <- ((StudentsPerformance %>% filter(failures == 1))$Gsum)
f2.Gsum <- ((StudentsPerformance %>% filter(failures == 2))$Gsum)
f3.Gsum <- ((StudentsPerformance %>% filter(failures == 3))$Gsum)

sd.df <- data.frame(groups = c("Group0", "Group1", "Group2" , "Group3" ) ,
                    sds = c(sd(f0.Gsum), sd(f1.Gsum), sd(f2.Gsum), sd(f3.Gsum)))



aov.Gsum_failures <- aov(Gsum ~ as.factor(failures), data = StudentsPerformance)
aov.Gsum_failures

summary(aov.Gsum_failures)



test1 <- lm(Gsum ~ failures, data = StudentsPerformance)
summary(test1)

TukeyHSD(aov.Gsum_failures)

plot(TukeyHSD(aov.Gsum_failures),las = 1)


sd(Gsum ~ as.factor(failures))



box <- ggplot(StudentsPerformance, aes(x = failures, y = Gsum, group = failures)) +
  geom_boxplot(alpha = 0.5, outlier.size = 2 , color = as.factor(failures), fill = as.factor
    (failures)) +
  labs(title="Boxplot")


box + scale_color_manual(values=c("thistle1", "thistle2", "thistle3" , "thistle4")) +
  scale_fill_manual(values=c("thistle1", "thistle2", "thistle3" , "thistle4"))
```

code.R

Statistical Inference: Project Phase II

Sarmad Zandi Goharrizi - 810199181

Question 0

Student's Performance includes various information about a sample of students studying in two different *schools*.
A sense of responsibility towards one's education and academic future is a notable information which can be mined from each individual's *study time* and their rate of *going out* which has an effect on their *failures* and their *grades*.
This dataset also contains some semi-relevent factors like each student's parent's job as well as their love life.

X	school	sex	age	Fjob	Mjob	goout	internet	romantic	studytime	failures	health	absences	G1	G2	G3
0	GP	F	18	teacher	at_home	4	no	no	2	0	3	6	5.000000	7.529856	9.289229
1	GP	F	17	other	at_home	3	yes	no	2	0	3	4	5.000000	7.192039	9.424835
2	GP	F	15	other	at_home	2	yes	no	2	3	3	10	3.807703	8.000000	7.354029
3	GP	F	15	services	health	2	yes	yes	3	0	5	2	15.000000	16.373208	17.796916
4	GP	F	16	other	other	2	no	no	2	0	5	4	6.000000	12.138542	12.800024
5	GP	M	16	other	services	2	yes	no	2	0	5	10	15.000000	16.804680	18.347259
6	GP	M	16	other	other	4	yes	no	2	0	3	0	12.000000	13.691091	14.187810
7	GP	F	17	teacher	other	4	no	no	2	0	1	6	6.000000	6.794185	9.012740
8	GP	M	15	other	services	2	yes	no	2	0	1	0	16.000000	19.852952	20.000000
9	GP	M	15	other	other	1	yes	no	2	0	5	0	14.000000	17.180466	18.073614
10	GP	F	15	health	teacher	3	yes	no	2	0	2	0	10.000000	9.609179	11.950918

Figure 1: Head of the dataset

Question 1

Chosen Categorical Variables : *sex* and *Mjob*

a.

In this part we indent to compare the proportion of mothers who are *teachers* between *Male* and *Female* students.

Conditions for inference for comparing two independent proportions :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
- Sample size / skew : samples should meet the success-failure condition (at least 10 *successes* and 10 *failures*) :
  - $n_1\hat{p_1} \geq 10 \rightarrow n_1\hat{p_1} = 200 \times 0.053 = 10.6 \geq 10$
  - $n_1(1-\hat{p_1}) \geq 10 \rightarrow n_1(1-\hat{p_1}) = 200 \times 0.947 = 189.4 \geq 10$
  - $n_2\hat{p_2} \geq 10 \rightarrow n_2\hat{p_2} = 200 \times 0.126 = 25.2 \geq 10$
  - $n_2(1-\hat{p_2}) \geq 10 \rightarrow n_2(1-\hat{p_2}) = 200 \times 0.874 = 174.8 \geq 10$

All is met.

$$\textsf{Confidence Interval} : \textit{point estimate} \pm \textit{margin of error} \longrightarrow \hat{p_1} - \hat{p_2} \pm z^\star SE_{\hat{p_1}-\hat{p_2}}$$

$$SE_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = 0.047$$

$$\textsf{Confidence Interval} : (0.0655, 0.0811)$$

If we take repeated samples from this population, and make a confidence interval using each sample, we expect about 95% of the resulting confidence intervals to contain $\hat{p_1} - \hat{p_2}$.

We are 95% confident that the difference of population proportion of g*Male* and *Female* students whose mother's job is *teachers* is between 0.0655 and 0.0811.

Other confidence intervals can be computed accordingly : mothers who are *at-home* between *Male* and *Female* students.

$$\textsf{Confidence Interval} : (-0.0632, -0.0567)$$

We are 95% confident that the difference of population proportion of g*Male* and *Female* students whose mother's job is *being home* is between -0.0632 and -0.0567.

mothers who are *health* between *Male* and *Female* students.

$$\textsf{Confidence Interval} : (-0.0239, -0.016)$$

We are 95% confident that the difference of population proportion of g*Male* and *Female* students whose mother's job is *health* is between -0.0239 and -0.016.

Question 1 continued on next page. . .

mothers who are *services* between *Male* and *Female* students.

$$\text{Confidence Interval} : (-0.040, -0.019)$$

We are 95% confident that the difference of population proportion of g*Male* and *Female* students whose mother's job is *services* is between −0.040 and −0.019.

b.

To test the independence, I tested my hypothesis using 2 different methods :
H_0 : *Mother's job is independent from sex of the student. (Mother's job does not vary with the sex of the child)*
H_A : *Mother's job is dependent to sex of the student. (Mother's job varies with the sex of the child)*

Method 1 : Pooling :

$$\hat{p_{pool}} = \frac{\#\text{ success}}{\#\text{ total}} = 0.085$$

$$SE_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{\hat{p_{pool}}(1-\hat{p_{pool}})}{n_1} + \frac{\hat{p_{pool}}(1-\hat{p_{pool}})}{n_2}} = 0.039$$

Conditions for inference for comparing two independent proportions (pooling) :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
- Sample size / skew : samples should meet the success-failure condition (at least 10 *successes* and 10 *failures*) :
  - $n_1\hat{p_{pool}} \geq 10 \rightarrow n_1\hat{p_{pool}} = 200 \times 0.085 = 17 \geq 10$
  - $n_1(1-\hat{p_{pool}}) \geq 10 \rightarrow n_1(1-\hat{p_{pool}}) = 200 \times 0.947 = 183 \geq 10$
  - $n_1\hat{p_{pool}} \geq 10 \rightarrow n_1\hat{p_{pool}} = 200 \times 0.085 = 17 \geq 10$
  - $n_1(1-\hat{p_{pool}}) \geq 10 \rightarrow n_1(1-\hat{p_{pool}}) = 200 \times 0.947 = 183 \geq 10$

All is met.

Due to the fact that p-value (0.105) is larger than 0.05 ,we fail to reject the null hypothesis. ⟶ There are evidence that *Mother's job (teaching specifically) does not vary with the sex of the child .*

Method 2 : χ^2 test :

$$Expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

$$text\ statistic : \frac{point\ estimate\ \text{-}\ null\ value}{SE\ of\ point\ estimate}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad and \qquad df = (R-1)(C-1)$$

Conditions for χ^2 test :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
  - each case only contributes to one cell in the table other (non-paired)
- Sample size : Each particular scenario (i.e.cell) must have atleast 5 expected cases. $\rightarrow \times$

```
   Fjob
sex at_home health other services teacher
  F       5      6    61       33       6
  M       4      5    55       16       9
```

Figure 2: dataset table

Which will give out the following warning :

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  sp.sampled.table
X-squared = 4.6465, df = 4, p-value = 0.3255
```

Figure 3: χ^2 test

For our last condition to meet, we have to merge two columns, *at-home* and *health* :

```
      other services teacher
F 11     61       33       6
M  9     55       16       9
```

Figure 4: dataset table

So our χ^2 test won't give any warnings :

```
        Pearson's Chi-squared test

data:  sp.sampled.table.bind
X-squared = 4.6445, df = 3, p-value = 0.1998
```

Figure 5: χ^2 test

Either way, due to the fact that p-value is larger than 0.05 , we fail reject the null hypothesis. $\longrightarrow$ There are evidence that *Mother's job does not vary with the sex of the child .*

Note : It's important to mention that in hypothesis testing in categorical variables, CI approach and p-value approach might not always give out the same result.

Question 2

Chosen Categorical Variable : *romantic*

$$H_0 : p = 0.5$$
$$H_A : p < 0.5$$

Conditions for inference for comparing two independent proportions :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
- Sample size / skew : samples should meet the success-failure condition (at least 10 *successes* and 10 *failures*) :
  - $n\hat{p} \geq 10 \rightarrow n\hat{p} = 15 \times 0.53 = 5.3 \ngeq 10$
  - $n(1-\hat{p}) \geq 10 \rightarrow n(1-\hat{p}) = 400 \times 0.47 = 4.7 \ngeq 10$

Due to the fact that our conditions did not meet, we will use simulation.

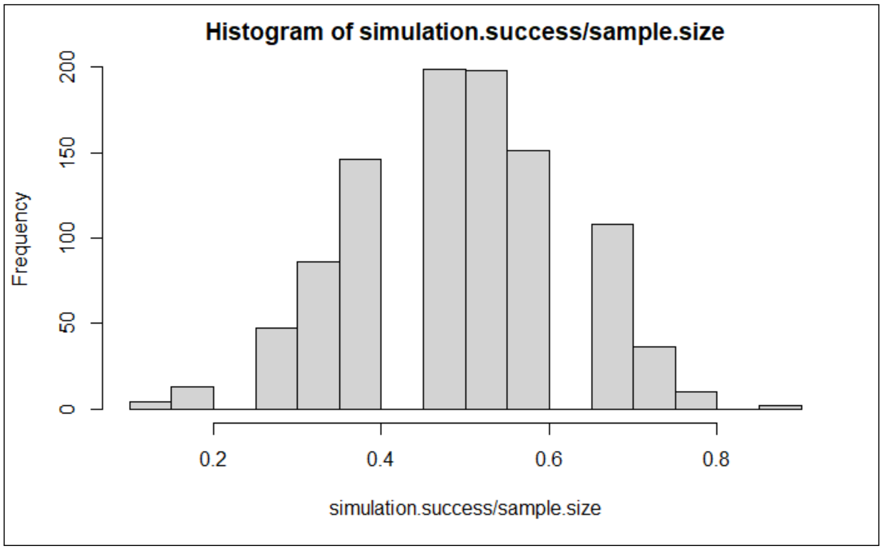


Figure 6: Histogram

Since, the p-value (0.505) is larger than 0.05, we fail to reject the null hypothesis and declare that there is not convincing evidence to accept the alternative hypothesis.
This means that each person is 50% likely to be in a romantic relationship.

Question 3

Chosen Categorical Variable : *Mjob*

```
sample.original
 at_home   health    other services  teacher
      59       34      141      103       58
```

Figure 7: Mjob

```
sample.original
 at_home   health    other services  teacher
  0.1494   0.0861   0.3570   0.2608   0.1468
```

Figure 8: Mjob - probability distribution

a.

Conditions for χ^2 test :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
  - each case only contributes to one cell in the table other (non-paired)
- Sample size : Each particular scenario (i.e.cell) must have atleast 5 expected cases.

$$H_0 : \textit{Samples are randomly chosen and there is nothing going on}$$

$$H_0 : \textit{Samples are not randomly chosen and there is something going on}$$

Randomly selected sample :

```
sample.unbiased
 at_home   health    other services  teacher
      16        9       39       18       18
```

Figure 9: 100 samples - randomly

χ^2 test :

```
        Chi-squared test for given probabilities

data:  unbiased.table
X-squared = 3.6496, df = 4, p-value = 0.4555
```

Figure 10: χ^2 test - randomly

Due to the fact that p-value (0.455) is larger than 0.05, we fail to reject the null hypothesis. There is convincing evidence to accept the null hypothesis.

Question 3 continued on next page...

Randomly selected sample with 0.6 bias through teachers :

```
sample.biased
 at_home   health    other services  teacher
      10       12       33       21       24
```

Figure 11: 100 samples - biased

χ^2 test :

```
        Chi-squared test for given probabilities

data:  biased.table
X-squared = 10.071, df = 4, p-value = 0.03924
```

Figure 12: χ^2 test - biased

Due to the fact that p-value (0.0392) is smaller than 0.05, we fail to reject the null hypothesis, there is convincing evidence that the samples are randomly chosen. (!)

b.

Chosen Categorical Variable : *Fjob*

$$H_0 : \textit{Mother's job and father's job are 2 independent variables}$$

$$H_0 : \textit{Mother's job and father's job are dependent variables}$$

$$\textit{Expected count} = \frac{\textit{row total} \times \textit{column total}}{\textit{table total}}$$

$$\textit{text statistic} : \frac{\textit{point estimate - null value}}{\textit{SE of point estimate}}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad \textit{and} \qquad df = (R-1)(C-1)$$

Conditions for χ^2 test :

- Independence :
  - random sample/assignment
  - if sampling without replacement, $n <$ 10% of population
  - each case only contributes to one cell in the table other (non-paired)
- Sample size : Each particular scenario (i.e.cell) must have atleast 5 expected cases. $\rightarrow \times$

```
          Fjob
Mjob       at_home health other services teacher
  at_home        2      2    21        7       0
  health         0      4     8        3       0
  other          4      0    51        8       4
  services       2      2    24       18       5
  teacher        1      3    12       13       6
```

Figure 13: table

Our last condition is not met, so we get a warning :

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  Mjob.Fjob
X-squared = 44.517, df = 16, p-value = 0.0001645
```

Figure 14: χ^2 test

We combine *at-home* , *health, services* and *teacher* of *Mjob* and compare it to *other* :

	[,1]	[,2]
at_home	11	21
health	7	8
other	16	51
services	27	24
teacher	23	12

Figure 15: combined table

```
        Pearson's Chi-squared test

data:  Mjob.Fjob.bind
X-squared = 20.514, df = 4, p-value = 0.0003952
```

Figure 16: χ^2 test

Both p-values indicate that we should reject the null hypothesis meaning that parent's job are dependent to each other.

Question 4

Chosen Variables : *G1 - failure* and *studytime*

a.
In phase 1, we used *pearson correlation* and *Correlogram* for our predictions.
(In order not to confuse the report of this phase with the previous phase, the question related to phase 1 of the project is also placed in the zip file of this project, although its abstract is also described here.)

Quoting phase 1 : *'Judging by Figure 34, G1 and G2 and G3 have positive linear associations with each other and with studytime as expected. Failure and goout both have a negative linear associations with G1, G2 and G3.'*

From all the variables mentioned,I chose one of the grades (G_1) as my response variable and from failures, goout and studytime i chose 2 of them that had the most correlation with G1 (absolute value of them are aimed).
(Note : Although G2 and G3 had a very high correlation with G1, i didn't pick them, because all three of these variables are scores in different classes and it is better to use other variables to better understand each person and do not estimate their G1 score only based on their other scores. (Each one of them can be a great response variable) Although in the end I built the model based on these two variables, because I do not know exactly was the exact aim of this question, to choose only based on scores or not, simply because I myself though a better model should be based on a student's other characteristics, I explained more about this.)

$$cor(G_1, failure) = -0.463$$

$$cor(G_1, studytime) = 0.176$$

$$cor(G_1, goout) = -0.161$$

Using those codes here, judging by the results, we can say *failures* is the more significant predictor :

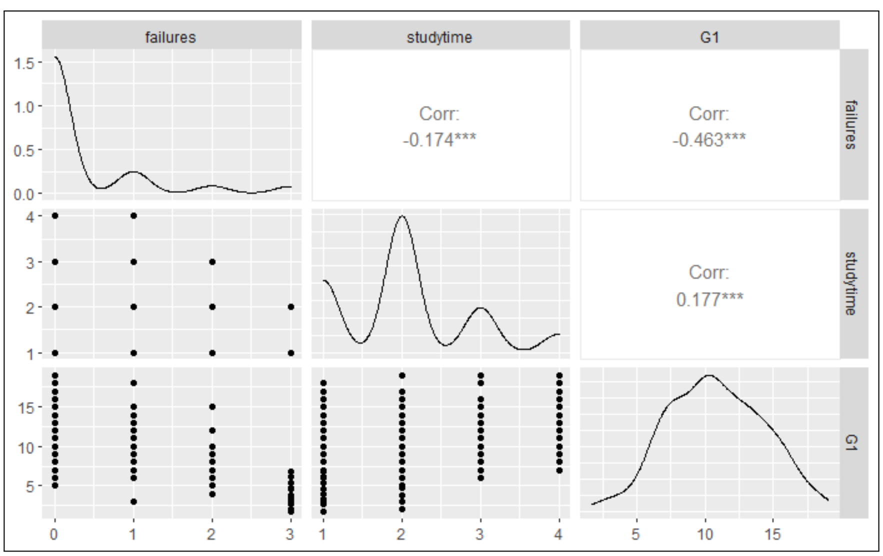


Figure 17: Correlogram

b.

Conditions for linear regression :

- Residuals vs Fitted : Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.
- Normal Q-Q : Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line.
- Scale-Location :(or Spread-Location). Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.
- Residuals vs Leverage : Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis.

failures :

```
Call:
lm(formula = G1 ~ failures, data = StudentsPerformance)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5154 -2.5154 -0.5154  2.4846  8.6768 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.5154     0.1724   66.79   <2e-16 ***
failures     -2.1922     0.2117  -10.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.125 on 393 degrees of freedom
Multiple R-squared:  0.2144,	Adjusted R-squared:  0.2124 
F-statistic: 107.2 on 1 and 393 DF,  p-value: < 2.2e-16
```

Figure 18: LM model

$$R^2 = 0.214$$

$$p - value < 2.2e - 16$$

According to R^2, 0.214 of the variability of the model is explained by failures.

According to the $p - value$, by modeling $G_1 \sim failures$, we can reject the null hypothesis that suggests there is no relationship between these two variables (slope is zero.)

```
Call:
lm(formula = G1 ~ failures, data = StudentsPerformance)

Coefficients:
(Intercept)     failures  
     11.515       -2.192  
```

Figure 19: LM model

$$G_1 = 11.515 - 2.192 \times failures$$

Intercept: When $failures = 0$, G_1 is expected to equal the intercept (11.515). (Maybe meaningless in context of the data, and only serve to adjust the height of the line.)

In our case when the student has not failed at all , their G_1 score is nearly 11 .

Slope: For each unit increase in $failures$, G_1 is expected to be 2.192 lower on average.

We also need to check whether conditions for using linear regression are met :

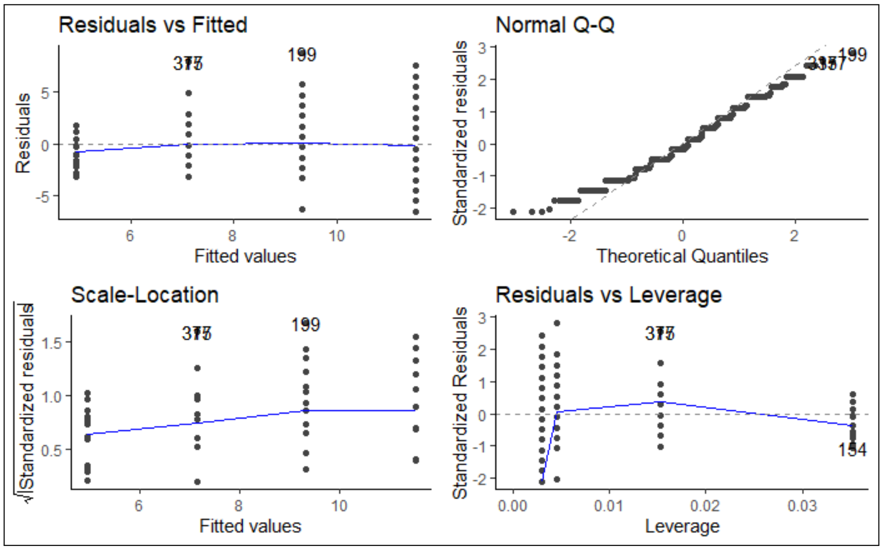


Figure 20: LM conditions - all met

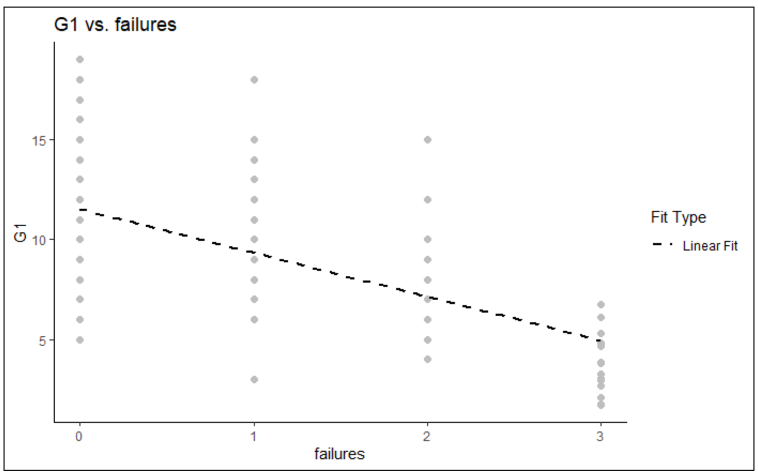


Figure 21: Scatter plot

studytime :

```
Call:
lm(formula = G1 ~ studytime, data = StudentsPerformance)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6592 -2.7566 -0.0149  2.3726  8.2434 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.2732     0.4585  20.224  < 2e-16 ***
studytime     0.7417     0.2083   3.561 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.47 on 393 degrees of freedom
Multiple R-squared:  0.03125,	Adjusted R-squared:  0.02879 
F-statistic: 12.68 on 1 and 393 DF,  p-value: 0.0004154
```

Figure 22: LM model

$$R^2 = 0.031$$

$$p - value = 0.00041$$

According to R^2, only 0.03 of the variability of the model is explained by studytime (which is alot smaller than failures).

According to the $p - value$, by modeling $G_1 \sim studytime$, we can reject the null hypothesis that suggests there is no relationship between these two variables (slope is zero.)

```
Call:
lm(formula = G1 ~ studytime, data = StudentsPerformance)

Coefficients:
(Intercept)    studytime  
     9.2732       0.7417  
```

Figure 23: LM model

$$G_1 = 9.2732 + 0.7417 \times studytime$$

intercept: When $studytime = 0$, G_1 is expected to equal the intercept (9.2732). Maybe meaningless in context of the data, and only serve to adjust the height of the line. In our case when the studentdoes not study at all , their G_1 score is nearly 9 .

slope: For each unit increase in $studytime$, G_1 is expected to be 0.7417 higher on average.

We also need to check whether conditions for using linear regression are met :

Question 4 continued on next page...

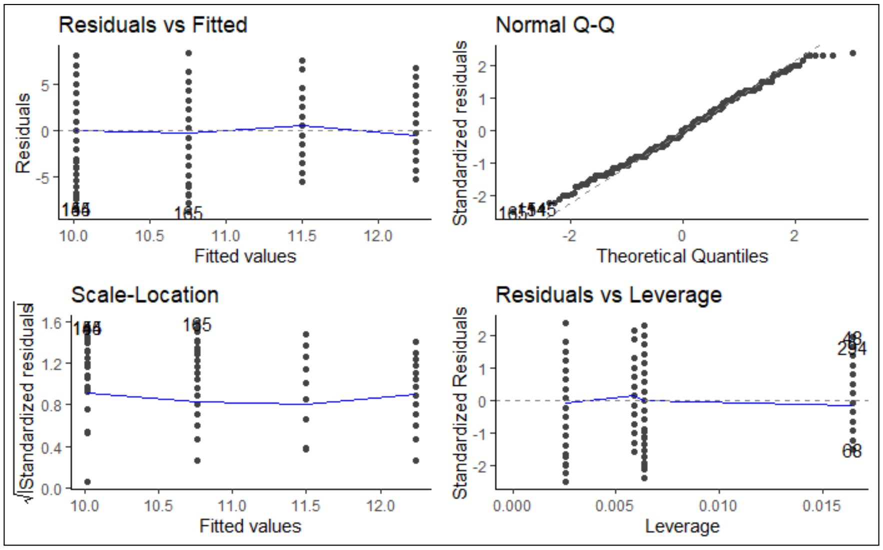


Figure 24: LM conditions - all met

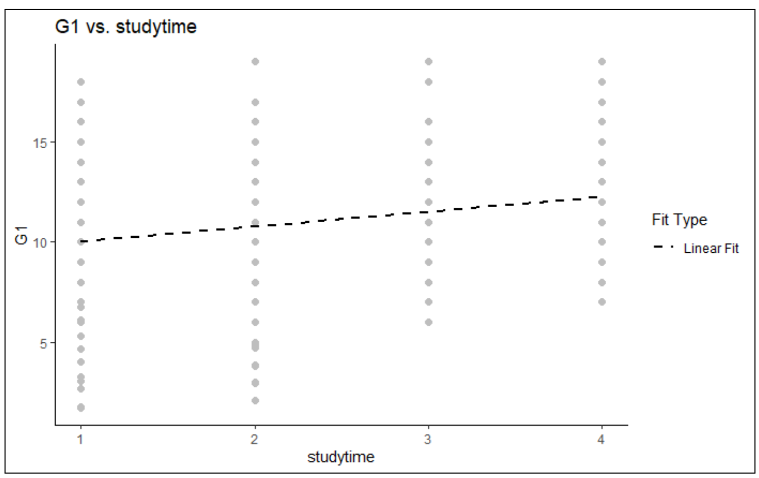


Figure 25: Scatter plot

c.

Judging by above figures, in order to pick the the more significant predictor we can use both R^2_{adj} and $p - value$:

	Adj. R-squared	p-value
failures	0.2124	2.2e-16
studytime	0.02879	0.00041

The more significant predictor is the one with the lowest $p - value$ and highest R^2_{adj}. Both of these point to failures being the best one.

Chosen Variables : *G1* - *G2* and *G3*

$$cor(G_1, G2) = 0.85$$

$$cor(G_1, G3) = 0.80$$

Using those codes here, judging by the results, we can say *G2* is the more significant predictor :

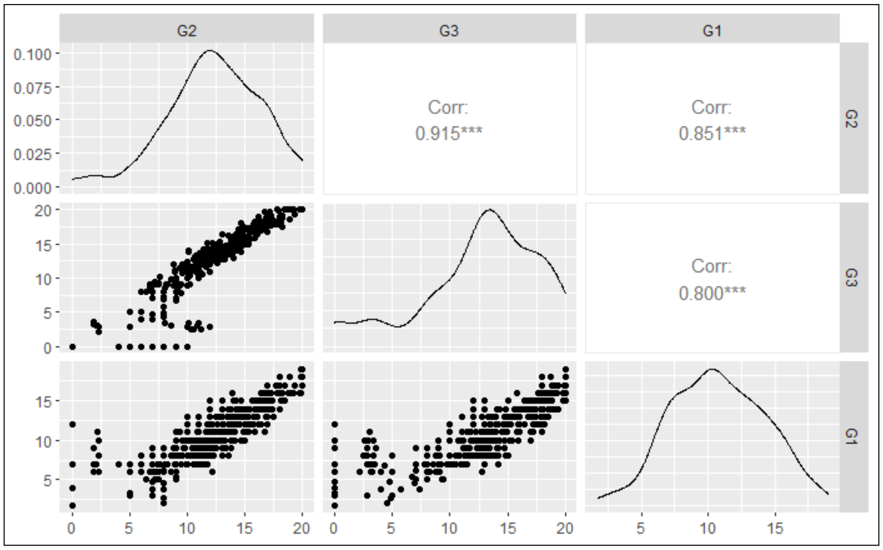


Figure 26: Correlogram

b.

Conditions for linear regression :

- Residuals vs Fitted : Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.
- Normal Q-Q : Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line.
- Scale-Location :(or Spread-Location). Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.
- Residuals vs Leverage : Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis.

G2 :

```
Call:
lm(formula = G1 ~ G2, data = StudentsPerformance)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5525 -1.1545 -0.0471  1.0380 10.2153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78475    0.29527   6.045 3.49e-09 ***
G2           0.73313    0.02283  32.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.852 on 393 degrees of freedom
Multiple R-squared:  0.7241,	Adjusted R-squared:  0.7234
F-statistic:  1031 on 1 and 393 DF,  p-value: < 2.2e-16
```

Figure 27: LM model

$$R^2 = 0.724$$

$$p - value < 2.2e - 16$$

According to R^2, 0.724 of the variability of the model is explained by failures (which is pretty good).

According to the $p - value$, by modeling $G1 \sim G2$, we can reject the null hypothesis that suggests there is no relationship between these two variables (slope is zero.)

$$G1 = 1.7847 + 0.7331 \times G2$$

Intercept: When $G2 = 0$, $G1$ is expected to equal the intercept (1.7847). (Maybe meaningless in context of the data, and only serve to adjust the height of the line.)
In our case when the student has not failed at all , their G_1 score is nearly 1.78 .

Slope: For each unit increase in $G2$, $G1$ is expected to be 0.7331 higher on average.

We also need to check whether conditions for using linear regression are met :

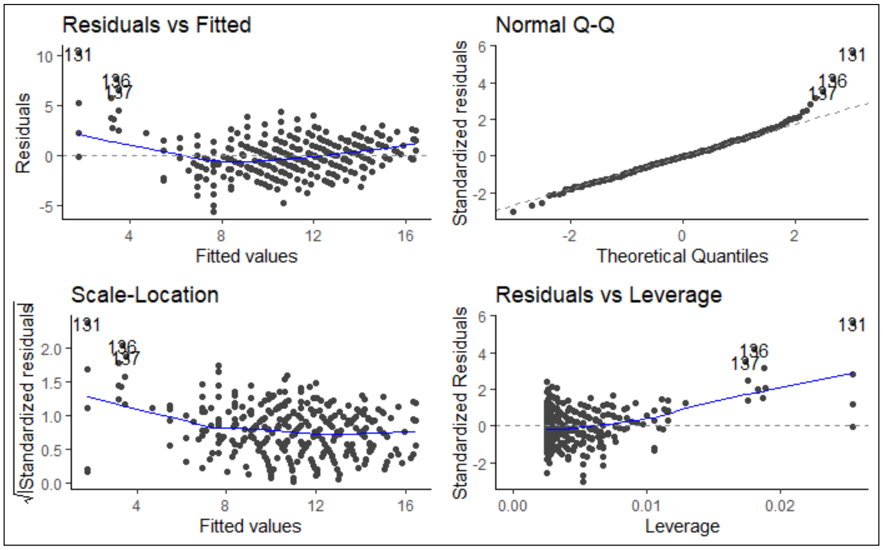


Figure 28: LM conditions - all are hardly met

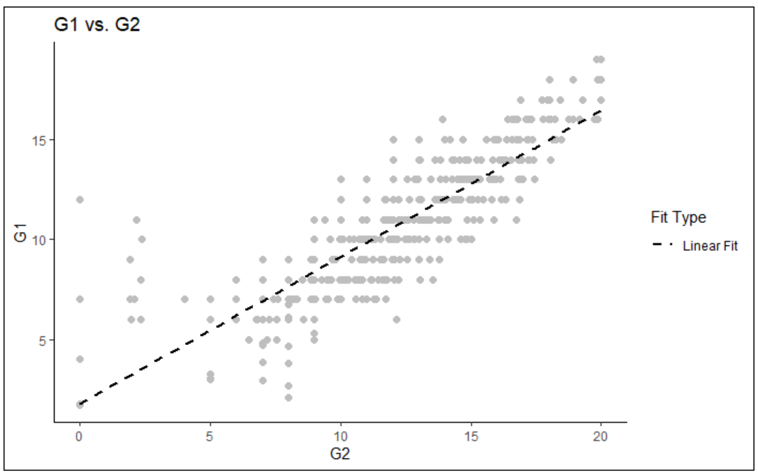


Figure 29: Scatter plot

G3 :

```
Call:
lm(formula = G1 ~ G3, data = StudentsPerformance)

Residuals:
   Min     1Q Median     3Q    Max
-4.870 -1.623 -0.080  1.338  8.140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8602     0.2825   13.66   <2e-16 ***
G3            0.5476     0.0207   26.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.114 on 393 degrees of freedom
Multiple R-squared:  0.6403,	Adjusted R-squared:  0.6394
F-statistic: 699.6 on 1 and 393 DF,  p-value: < 2.2e-16
```

Figure 30: LM model

$$R^2 = 0.64$$

$$p - value < 2.2e - 16$$

According to R^2, 0.64 of the variability of the model is explained by failures (which is pretty good).

According to the $p - value$, by modeling $G1 \sim G3$, we can reject the null hypothesis that suggests there is no relationship between these two variables (slope is zero.)

$$G1 = 3.860 + 0.5476 \times G3$$

Intercept: When $G3 = 0$, $G1$ is expected to equal the intercept (3.860). (Maybe meaningless in context of the data, and only serve to adjust the height of the line.)
In our case when the student has not failed at all , their $G3$ score is nearly 1.78 .

Slope: For each unit increase in $G3$, $G1$ is expected to be 0.5476 higher on average.

We also need to check whether conditions for using linear regression are met :

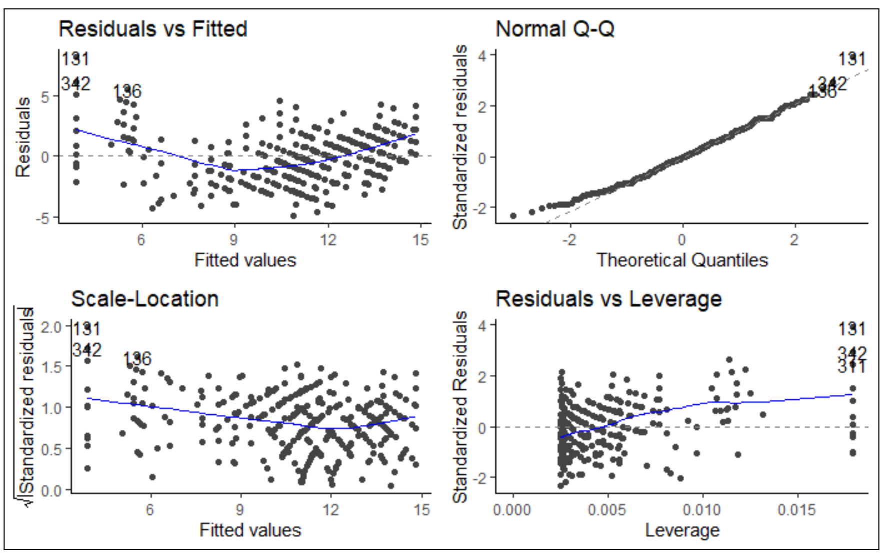


Figure 31: LM conditions - all are hardly met

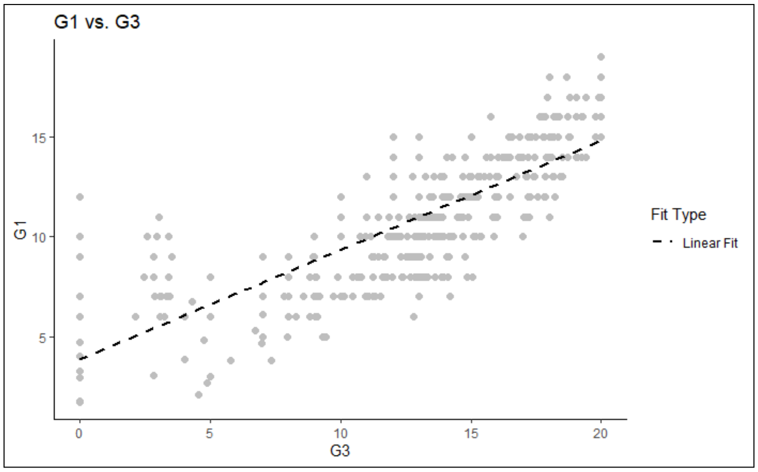


Figure 32: Scatter plot

	Adj. R-squared	p-value
G2	0.724	2.2e-16
G3	0.64	2.2e-16

The more significant predictor is the one with the lowest $p - value$ and highest R^2_{adj}. Although there is no difference in $p - value$, according to Adj. R-squared, G2 is the best one.
(Note : Between G2 and failures, G2 has a better R^2_{adj}, but it did not meet the conditions very well. ButR^2_{adj} is more important so if we have to choose one variable, we choose G2)
d.
From this part forward , i will compare both of the models i made till now :

Adj. R-squared :
As was also mentioned in part c., Comparing failure vs. studytime using R^2_{adj} will result in $G_1 \sim failures$ to be the better model.
As was also mentioned in part c., Comparing G2 vs. G3 using R^2_{adj} will result in $G_1 \sim G2$ to be the better model.
As was also mentioned in part c., Comparing failure vs. G2 using R^2_{adj} will result in $G_1 \sim G2$ to be the better model.

ANOVA table :

In order to compare my models, we first consider a base model, for example :

$$G1 \sim sex$$

```
Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq F value Pr(>F)
sex         1   28.1  28.108  2.2745 0.1323
Residuals 393 4856.6  12.358
```

Figure 33: anove

failure vs. studytime :

```
Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq  F value  Pr(>F)
sex         1   28.1   28.11   2.9056 0.08906 .
failures    1 1064.5 1064.54 110.0451 < 2e-16 ***
Residuals 392 3792.1    9.67
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 34: anove

```
Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq F value    Pr(>F)
sex         1   28.1  28.108  2.3741    0.1242
studytime   1  215.6 215.641 18.2140 2.479e-05 ***
Residuals 392 4641.0  11.839
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 35: anove

$$R^2 = \frac{SS_{reg}}{SS_{total}}$$

	Base	+ failures	+ studytime
R2	0.01	0.22	0.05

Figure 36: computed R2 base on anova

Comparing failure vs. studytime using R^2 will result in $G_1 \sim failures$ to be the better model.
G2 vs. G3 :

```
Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq   F value    Pr(>F)
sex         1   28.1    28.1    8.1791  0.004464 **
G2          1 3509.5  3509.5 1021.2354 < 2.2e-16 ***
Residuals 392 1347.1     3.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 37: anove

```
Analysis of Variance Table

Response: G1
           Df  Sum Sq Mean Sq  F value  Pr(>F)
sex         1   28.11   28.11   6.2773 0.01263 *
G3          1 3101.39 3101.39 692.6334 < 2e-16 ***
Residuals 392 1755.25    4.48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 38: anove

$$R^2 = \frac{SS_{reg}}{SS_{total}}$$

```
Analysis of Variance Table

Response: G1
           Df  Sum Sq Mean Sq  F value  Pr(>F)
sex         1   28.11   28.11   6.2773 0.01263 *
G3          1 3101.39 3101.39 692.6334 < 2e-16 ***
Residuals 392 1755.25    4.48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 39: computed R2 base on anova

Comparing G2 vs. G3 using R^2 will result in $G_1 \sim G2$ to be the better model.
Due to the fact that $n-1$ and $n-k-1$ are approximately the same, R^2 and Adjusted R^2 doesn't have a noticeable difference.

e.

When there are many possible predictors, we need some strategy for selecting the best predictors to use in a regression model.

- Adjusted R^2 : Under this criterion, the best model is the one with the highest value of Adjusted R^2 .
- Cross-validation (explained in the next question) :Under this criterion, the best model is the one with the smallest value of MSE.
- Corrected Akaike's Information Criterion :Under this criterion, the best model is the one with the smallest value of AIC.
- Schwarz's Bayesian Information Criterion :Under this criterion, the best model is the one with the smallest value of BIC.

While R^2 is widely used, and has been around longer than the other measures, its tendency to select too many predictor variables makes it less suitable for forecasting.
Many statisticians like to use the BIC because it has the feature that if there is a true underlying model, the BIC will select that model given enough data. However, in reality, there is rarely, if ever, a true underlying model, and even if there was a true underlying model, selecting that model will not necessarily give the best forecasts (because the parameter estimates may not be accurate).

f.

H_0 : *The explanatory variable is not a significant predictor of the response variable, i.e. no relationship* $\rightarrow \beta = 0$

H_A : *The explanatory variable is a significant predictor of the response variable, i.e. relationship* $\rightarrow \beta \neq 0$

(a)

	p-value	significant
studytime	0.385	×
failure	2.2e-16	✓
G2	2.2e-16	✓
G3	2.2e-16	✓

Comparing failure vs. studytime : failure is significant predictor of the response variable.

Comparing G2 vs. G3 : Both are significant predictor of the response variable.

Comparing failure vs. G2 : Both are is significant predictor of the response variable.

(b)

$$\textit{failure CI} \ : (-2.448, -1.769)$$

We are 95% confident that for each additional point on failure, G1 is expected on average to be lower by 1.769 to 2.448 points.

$$\textit{studytime CI} \ : (0.019, 0.845)$$

We are 95% confident that for each additional point on studytime, G1 is expected on average to be higher by 0.019 to 0.845 points.

$$\textit{G2 CI} \ : (0.613, 0.698)$$

We are 95% confident that for each additional point on G2, G1 is expected on average to be higher by 0.613 to 0.846985 points.

$$\textit{G3 CI} \ : (0.530, 0.565)$$

We are 95% confident that for each additional point on G3, G1 is expected on average to be higher by 0.530 to 0.565 points.

(c)

	Actual	Predicted studytime	Predicted failues	Predicted G2	Predicted G3
209	9	10.2	11.5	10.2	10.9
244	13	10.2	11.5	11.9	12.2
6	15	10.6	11.5	13.8	13.9
358	12	10.6	11.5	12.0	11.6
178	6	10.6	11.5	7.3	8.7
44	8	10.2	11.5	9.4	11.5
201	16	10.6	11.5	14.6	14.3
82	11	11.0	11.5	10.6	11.4
295	14	11.0	11.5	12.4	13.2
30	10	10.6	11.5	11.8	11.5

Figure 40: Predicted

(d)

Question 4 continued on next page. . .

	Actual	Predicted studytime	Predicted failues	Predicted G2	Predicted G3
209	0	1.2	2.5	1.2	1.9
244	0	2.8	1.5	1.1	0.8
6	0	4.4	3.5	1.2	1.1
358	0	1.4	0.5	0.0	0.4
178	0	4.6	5.5	1.3	2.7
44	0	2.2	3.5	1.4	3.5
201	0	5.4	4.5	1.4	1.7
82	0	0.0	0.5	0.4	0.4
295	0	3.0	2.5	1.6	0.8
30	0	0.6	1.5	1.8	1.5

Figure 41: Prediction error

In order to compute success rate, I accept an 0.1 error which is 2 (data range $\times$ accepted error $= 20 \times 0.1 = 2$).

If the predicted result is ± 2 of my actual value, I accept it.

	Predicted studytime	Predicted failues	Predicted G2	Predicted G3
1	40 %	40 %	100 %	80 %

Figure 42: Success rate

Comparing failure vs. studytime : no difference !

Comparing G2 vs. G3 : G2 is the best predictor of the response variable.

Comparing failure vs. G2 : G2 is the best predictor of the response variable.

G2 is by far the best predictor.

Using *min-max calculation* :

$$MinMaxAccuracy = mean\left(\frac{min(actual, predicted)}{max(actual, predicted)}\right)$$

	Predicted studytime	Predicted failues	Predicted G2	Predicted G3
1	79.99 %	79.76 %	90 %	86.93 %

Figure 43: Success rate

Comparing failure vs. studytime :studytime is the best predictor of the response variable.

Comparing G2 vs. G3 : G2 is the best predictor of the response variable.

Comparing failure vs. G2 : G2 is the best predictor of the response variable.

G2 is by far the best predictor.

Using *MAPE calculation* :

$$MeanAbsolutePercentageError = mean\left(\frac{abs(actual - predicted)}{actual}\right)$$

	Predicted studytime	Predicted failues	Predicted G2	Predicted G3	
1	79.99 %	79.76 %	90 %	86.93 %	86.93 %

Figure 44: Success rate

Comparing failure vs. studytime :studytime is the best predictor of the response variable.

Comparing G2 vs. G3 : G2 is the best predictor of the response variable.

Comparing failure vs. G2 : G2 is the best predictor of the response variable.

Question 5

Chosen Categorical Variables : *G1 - G2*, *goout* , *sex* ,*failures* ,*age* and *studytime*

a.

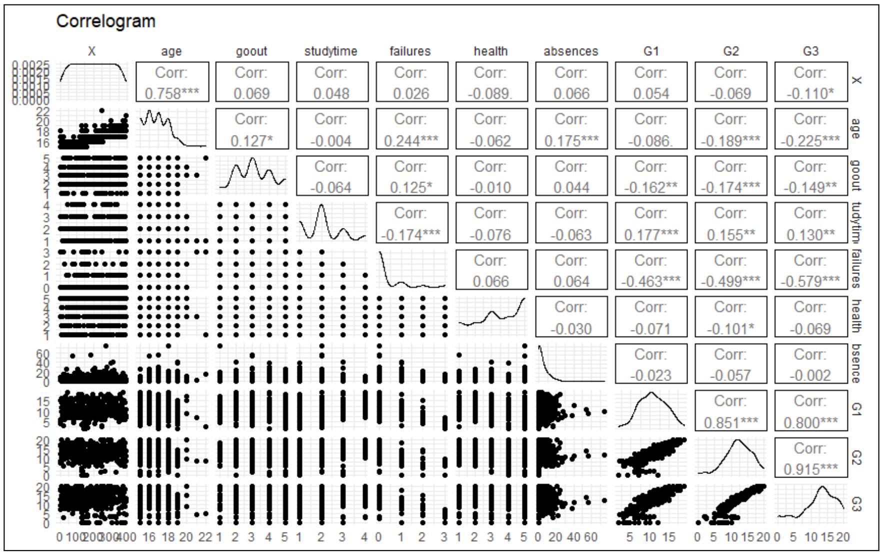


Figure 45: Correlogram

Considering all the analysis and inferences made in the previous question, for this question, from all the choices we have, we will definitely put G2 and failure among our options. Adding both G2 and G3 won't add anything new to the table since they are collinear.

We will add studytime, goout, age, and sex too, but we have to be careful not to use too many variables; we should pay attention to *occam's razor*; prefer the simplest best model!

The correlation between variables was also explained in the last question, but to summarize, as it was expected, failures and G2 have the highest correlation. Surprisingly, age has a higher correlation with G2 than studytime (considering their absolute value).

Age and sex are not correlated, just as sex and G2. sex has nothing to do with score or age, so it was probably expected.

More explanations can be found on phase 1, so to avoid lengthening the report, we move on to the next part.

Mjob and Fjob seemed hardly important to G1, so I didn't use them.

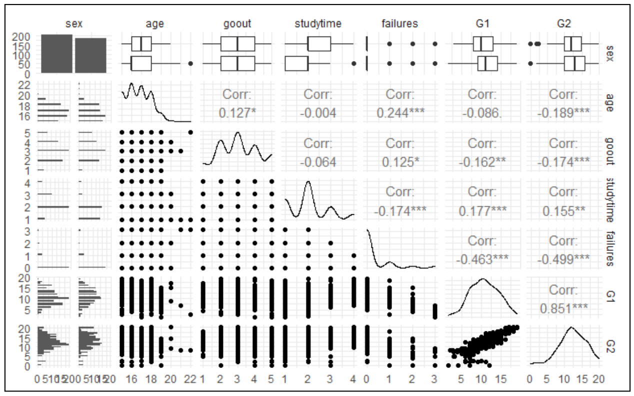


Figure 46: Correlogram

We don't want any collinearity between the variables that we chose and the below figure shows that we did a good job :

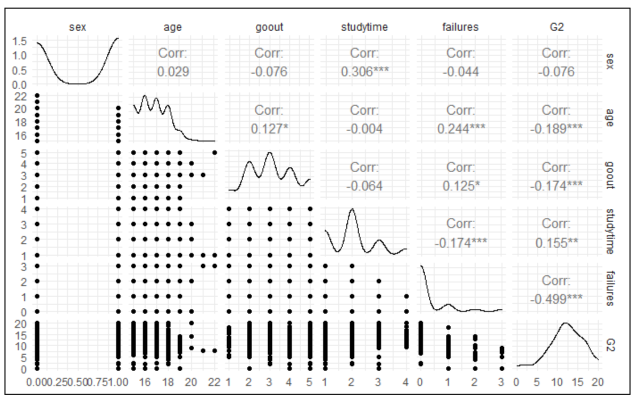


Figure 47: Correlogram - response variable omitted

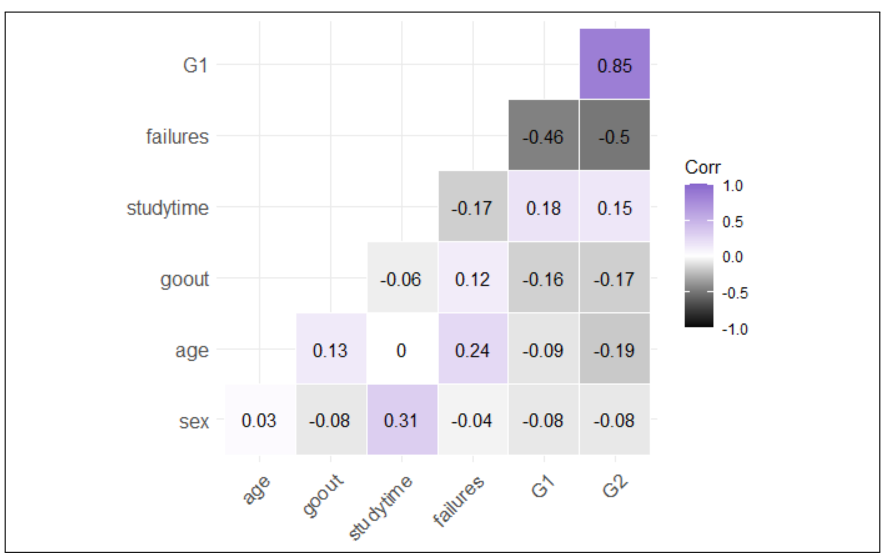


Figure 48: Correlogram

As was said multiple times in the previous question, $G2$ plays a more significant role in prediction.

b.

```
Call:
lm(formula = G1 ~ G2 + goout + failures + studytime + sex + age,
    data = StudentsPerformance)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3720 -1.1898 -0.1367  1.0890 10.6786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.97651    1.33371  -1.482   0.1392
G2           0.70774    0.02661  26.599   <2e-16 ***
goout       -0.06969    0.08454  -0.824   0.4102
failures    -0.30731    0.14614  -2.103   0.0361 *
studytime    0.20212    0.11755   1.719   0.0863 .
sex         -0.24841    0.19572  -1.269   0.2051
age          0.24627    0.07487   3.289   0.0011 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.823 on 388 degrees of freedom
Multiple R-squared:  0.7361,    Adjusted R-squared:  0.732
F-statistic: 180.4 on 6 and 388 DF,  p-value: < 2.2e-16
```

Figure 49: Correlogram

$$G1 = -1.97 + 0.7 \times G2 - 0.06 \times goout + -0.3 \times failures + 0.2 \times studytime - 0.24 \times sex : M + 0.24 \times age$$

c.

R^2 shows what percent of variability in the response variable is explained by the model. In our case nearly 74% of the variablity was explained using 6 variables out of the 15 variables available which is pretty good.

d.

Higher R^2 doesn't necessarily guarantee that the model fits the data well, we might face over-fitting if we are not careful.

Adjusted R^2 can be a good indicator of when the model fits the data well, it compares the explanatory power of regression models that contain different numbers of predictors. Adjusted R^2 is around 73% in our fitted model.

The fact that R^2 and Adjusted R^2 are this close is very good which means we don't have overfitting in our model.

Other techniques can help us know whether we have a good fit or not, for example Residuals : A good way to test the quality of the fit of the model is to look at the residuals The idea in here is that the sum of the residuals is approximately zero or as low as possible.

Although all our analyzes so far have promised a good model, the figure below shows that the model is not the best possible model.

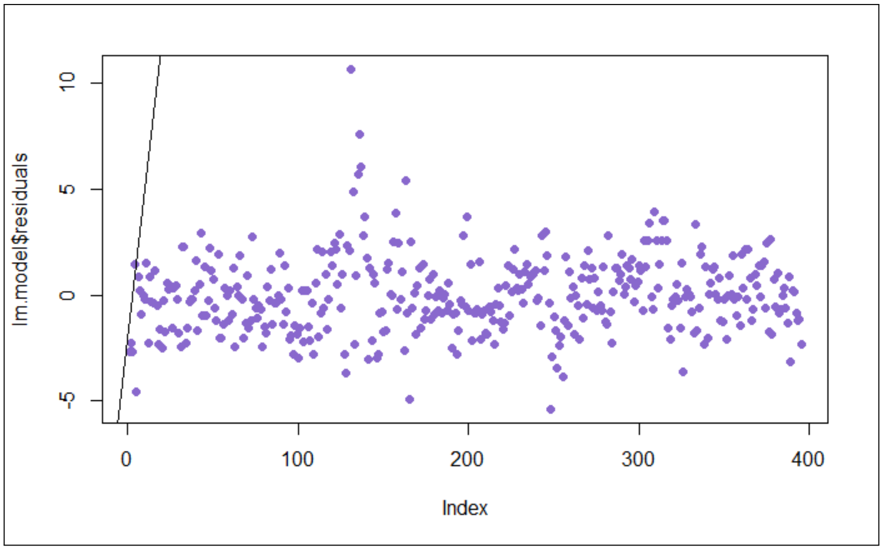


Figure 50: residuals

e.
To develop the best possible model, there are 4 different approaches :
Forward selection : start with an empty model and add one predictor at a time until the parsimonious model is reached.

(a) *p-value* :
Start with single predictor regressions of response vs. each explanatory variable (G2, G3 and failure all had pvalues smaller than 2.2 e-16 , doesn't matter which one we will choose)

Pick the variable with the lowest significant p-value
Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value

Repeat until any of the remaining variables do not have a significant p-value

For this part i used $ols_step_forward_p$, detailes can also be found in my project's file.

```
Variables Entered:

+ G2
+ age
+ failures
+ studytime
+ sex

Final Model Output
------------------

                         Model Summary
--------------------------------------------------------------
R                         0.858       RMSE                1.822
R-Squared                 0.736       Coef. Var          16.896
Adj. R-Squared            0.732       MSE                 3.319
Pred R-Squared            0.722       MAE                 1.361
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                 ANOVA
----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
----------------------------------------------------------------------
Regression    3593.549          5        718.710    216.526    0.0000
Residual      1291.200        389          3.319
Total         4884.749        394
----------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta       t        Sig      lower    upper
----------------------------------------------------------------------------------------
(Intercept)      -2.139         1.319                 -1.622    0.106    -4.731    0.453
         G2       0.711         0.026        0.825    26.946    0.000     0.659    0.762
        age       0.240         0.075        0.087     3.227    0.001     0.094    0.387
   failures      -0.309         0.146       -0.065    -2.119    0.035    -0.597   -0.022
  studytime       0.203         0.117        0.048     1.728    0.085    -0.028    0.434
        sex      -0.235         0.195       -0.033    -1.206    0.229    -0.618    0.148
----------------------------------------------------------------------------------------
```

Figure 51: Forward - p-value

Final model :

$$G1 \sim G2 + age + failures + sex + studytime$$

(b) *Adjusted R^2* :

Start with single predictor regressions of response vs. each explanatory variable Pick the model with the highest adjusted R^2
Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R^2
Repeat until the addition of any of the remaining variables does not result in a higher adjusted R^2

	best.pred	all.adj.r.squared
1	G2	0.7507277
2	G2 + age	0.7630615
3	G2 + age + failures	0.7639030
4	G2 + age + failures + studytime	0.7635379

Figure 52: Forward - Adjusted R^2

As we can see, adding the fourth variable, studytime, didn't help us with gaining a better fit for our model, so we wrapped this approach up after obtaining this model :

$$G1 \sim G2 + age + failures$$

with Adjusted $R^2 \approx 73.4$ %

Backward elimination : start with a full model (containing all predictors), drop one predictor at a time until the parsimonious model is reached.

(a) *p-value* :

Start with the full model
Drop the variable with the highest p-value and refit a smaller model
Repeat until all variables left in the model are significant

For this part i used $ols_step_backward_p$, detailes can also be found in my project's file.

```
Variables Removed:

x goout
x sex
x studytime


Final Model Output
------------------

                         Model Summary
--------------------------------------------------------------
R                        0.856       RMSE               1.825
R-Squared                0.733       Coef. Var         16.928
Adj. R-Squared           0.731       MSE                3.332
Pred R-Squared           0.723       MAE                1.363
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                 ANOVA
----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F         Sig.
----------------------------------------------------------------------
Regression    3582.019          3       1194.006    358.368    0.0000
Residual      1302.730        391          3.332
Total         4884.749        394
----------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta      t         Sig      lower    upper
----------------------------------------------------------------------------------------
(Intercept)     -1.994         1.311                 -1.521     0.129    -4.573    0.584
         G2      0.718         0.026        0.834    27.563     0.000     0.667    0.769
   failures     -0.323         0.145       -0.068    -2.225     0.027    -0.608   -0.038
        age      0.244         0.075        0.088     3.270     0.001     0.097    0.390
----------------------------------------------------------------------------------------
```

Figure 53: Backward - p-value

Final model :

$$G1 \sim G2 + age + failures$$

Forward and backward p-value approach did not gain the same result.

(b) *Adjusted* R^2 :

Start with the full model
Drop one variable at a time and record adjusted R^2 of each smaller model
Pick the model with the highest increase in adjusted R^2
Repeat until none of the models yield an increase in adjusted R^2

	best.pred	all.adj.r.squared
1	G2 + failures + studytime + sex + age	0.7630107
2	G2 + failures + studytime + age	0.7635379
3	G2 + failures + age	0.7639030

Figure 54: Backward - Adjusted R^2

Question 5 continued on next page...

As we can see, omiting either G2, failure nor age, didn't help us with gaining a better fit for our model and increasing our Adjusted R^2 , so we wrapped this approach up after obtaining this model :

$$G1 \sim G2 + age + failures$$

with Adjusted $R^2 \approx 73.4$ %

Surprisingly, both forward and backward Adjusted R^2 gained the same result.

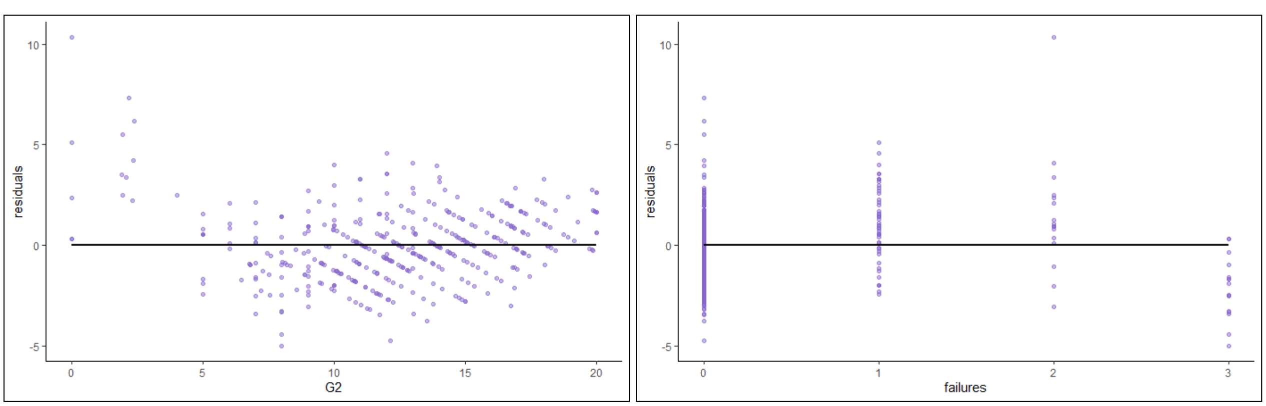
After completing these 4 methods, we see that both Adjusted R^2 approaches and the backward elimination pvalue approach all came to the same result, which is different from the result reached by forward selection pvalue.
According to the criterias mentioned in part d, $G1 \sim G2 + age + failures$ model is better than $G1 \sim G2 + age + failures + studytime + sex$, so we will use the same model in the following parts.
f.

Conditions for linear regression :

- Linear relationships between x and y :
  Each (numerical) explanatory variable linearly related to the response variable
  Check using residuals plots (e vs. x)
  Looking for a random scatter around 0
  Instead of scatterplot of y vs. x : allows for considering the other variables that are also in the model, and not just the bivariate relationship between a given x and y

- Nearly normal residuals :
  Some residuals will be positive and some negative
  On a residuals plot we look for random scatter of residuals around 0
  This translates to a nearly normal distribution of residuals centered at 0
  Check using histogram or normal probability plot

- Constant variability of residuals :
  Residuals should be equally variable for low and high values of the predicted response variable
  Check using residuals plots of residuals vs. predicted (e vs. y)
  Residuals vs. predicted instead of residuals vs. x because it allows for considering the entire model (with all explanatory variables) at once
  Residuals randomly scattered in a band with a constant width around 0 (no fan shape)
  Also worthwhile to view absolute value of residuals vs. predicted to identify unusual observations easily



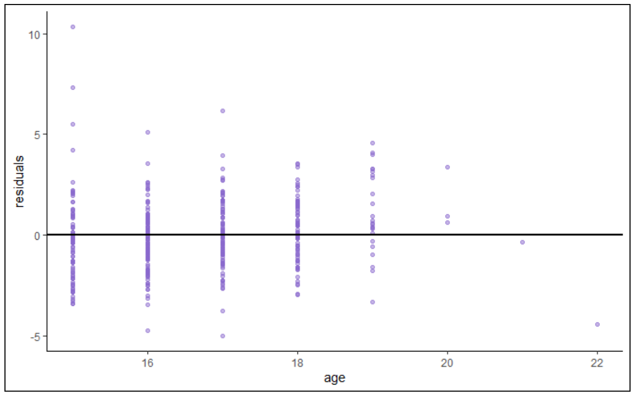


Figure 55: Linear relationship

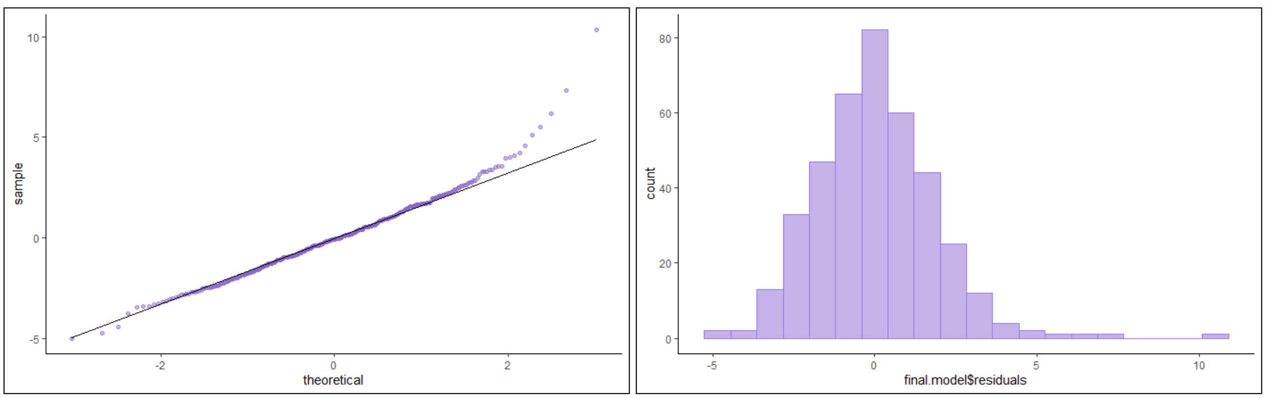


Figure 56: Nearly normal residuals

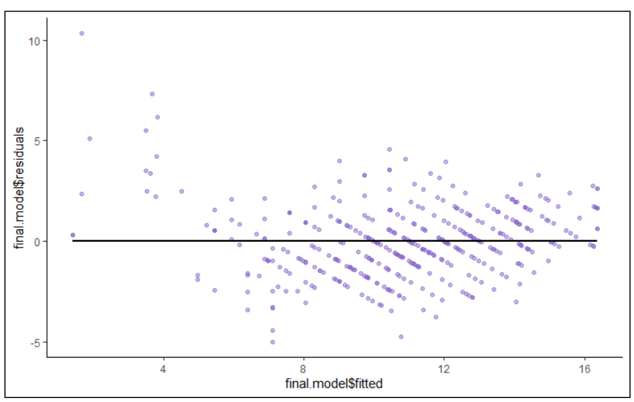


Figure 57: Constant var.

All conditions are met, not fully and perfectly, but they are met.
There are outliers that are effecting our model, if we eliminate them, we will meet them perfectly.
More on ourliers and detecting them in conditions and plots below :
(Note : this conditions were not mentioned in the slides, i checked them too just to be sure, some of them might overlap with the prev. conditions) Conditions for linear regression :

- Residuals vs Fitted : Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.
- Normal Q-Q : Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line.
- Scale-Location :(or Spread-Location). Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.
- Residuals vs Leverage : Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis.

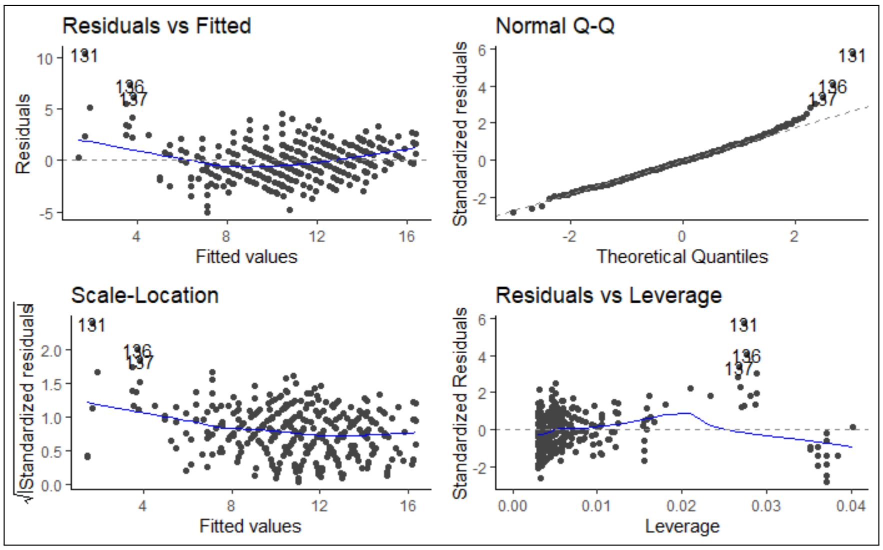


Figure 58: Conditions

All in all, we have a reliable model !

g.

The basic idea, behind cross-validation techniques, consists of dividing the data into two sets:
The training set, used to train (i.e. build) the model; and the testing set (or validation set), used to test (i.e. validate) the model by estimating the prediction error. Cross-validation is also known as a resampling method because it involves fitting the same statistical method multiple times using different subsets of the data.
The k-fold cross-validation method evaluates the model performance on different subset of the training data and then calculate the average prediction error rate. The algorithm is as follow:

- Randomly split the data set into k-subsets (or k-fold) (for example 5 subsets)
- Reserve one subset and train the model on all other subsets
- Test the model on the reserved subset and record the prediction error
- Repeat this process until each of the k subsets has served as the test set.
- Compute the average of the k recorded errors. This is called the cross-validation error serving as the performance metric for the model.

Question 5 continued on next page. . .

Root Mean Squared Error, which measures the model prediction error. It corresponds to the average difference between the observed known values of the outcome and the predicted value by the model. The lower the RMSE, the better the model.

```
Linear Regression

395 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 317, 315, 316, 316, 316
Resampling results:

  RMSE     Rsquared   MAE
  1.85564  0.7268193  1.389432

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 59: Full model

```
Linear Regression

395 samples
  3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 316, 316, 315, 317, 316
Resampling results:

  RMSE      Rsquared   MAE
  1.836361  0.7357864  1.379812

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 60: Best model

Due to the fact that RMSE is lower in the so called 'Best model', we trust what we have done till now.

	RMSE	Rsquared	MAE	Resample
1	1.637994	0.7763565	1.284930	Fold1
2	2.069837	0.6892742	1.364075	Fold2
3	1.712149	0.7383289	1.322173	Fold3
4	1.848170	0.7410741	1.559756	Fold4
5	1.912499	0.7349491	1.378221	Fold5

Figure 61: Different metrics of all 5-fold , best model

Question 6

Chosen Variables : *catG3* - *failures*, *studytime*, *G2* and *sex*

Due to the fact that my dataset lacked a good binary categorical variable, i made G_3 into a binary categorical variable $\longrightarrow$ if $G_3 < 10 : Fail(0)$ else $Pass(1)$

a.

```
Call:
glm(formula = catG3 ~ failures + studytime + G2 + sex, family = binomial(link = "logit"),
    data = StudentsPerformance)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.81360   0.00018   0.01265   0.13763   2.19874

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.5388     2.4952  -5.827 5.65e-09 ***
failures     -0.7129     0.3903  -1.827   0.0678 .
studytime    -0.2745     0.3209  -0.856   0.3922
G2            1.6466     0.2553   6.449 1.13e-10 ***
sex          -0.4254     0.5741  -0.741   0.4587
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 433.50  on 394  degrees of freedom
Residual deviance: 102.11  on 390  degrees of freedom
AIC: 112.11

Number of Fisher Scoring iterations: 8
```

Figure 62: GLM model

$$log(\frac{p}{1-p}) = -14.538 - 0.712 \times failures - 0.2745 \times studytime + 1.646 \times G2 + -0.42 \times sex : M$$

intercept : keeping all other predictors zero, the log odds ratio / odds radio of catG3 is -14.538 / exp(-14.538) = 4.85e-7

failures : keeping all other predictors constant for a unit increase in failures, the log odds ratio / odds radio of catG3 will decrease -0.712 / exp(-0.712) = 0.49

studytime : keeping all other predictors constant for a unit increase in studytime, the log odds ratio / odds radio of catG3 will decrease -0.2745 / exp(-0.2745) = 0.763

G2 : keeping all other predictors constant for a unit increase in G2 , the log odds ratio / odds radio of catG3 will increase 1.6462 / exp(1.646) = 5.15

sex : keeping all other predictors constant, the log odds ratio / odds radio of catG3 for reference point (M) is - 0.42 / exp(-0.712) = 0.657 less than F

b.

Odds ratio (OR) is a statistic that quantifies the strength of the association between two events, A and B. The odds ratio is defined as the ratio of the odds of A in the presence of B and vise versa, which, due to symmetry, is equal to the ratio of the odds of B in the presence of A and the odds of B in the absence of A.

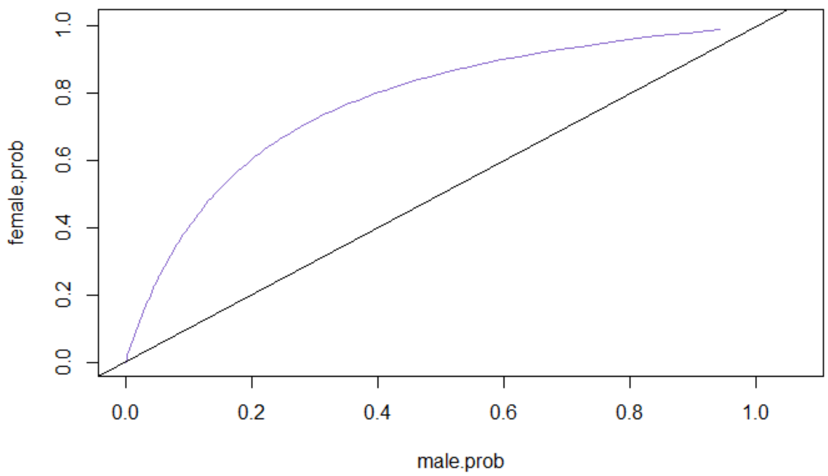


Figure 63: Odds ratio curve for sex - ref : M

This curve indicates the probability of passing G3 (catG3 = 1), for male refrence point :

$$x : P(catG3|Male) \sim y : P(catG3|Female)$$

c.

ROC stands for Receiver Operating Characteristics, and it is used to evaluate the prediction accuracy of a classifier model. ROC curve is a metric describing the trade-off between the sensitivity (true positive rate, TPR) and specificity (false positive rate, FPR) of a prediction in all probability cutoffs (thresholds).
It can be used for binary and multi-class classification accuracy checking.
To evaluate the ROC in multi-class prediction, we create binary classes by mapping each class against the other classes.

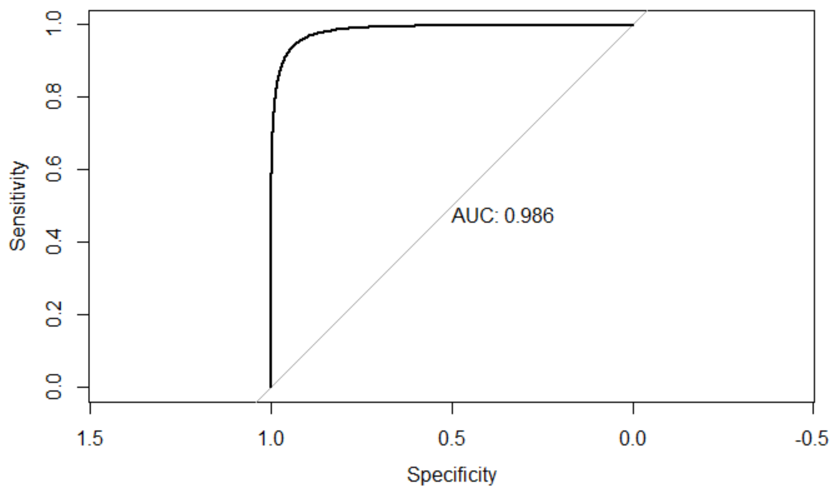


Figure 64: ROC curve - test

The AUC represents the area under the ROC curve. We can evaluate the model the performance by the value of AUC. Higher than 0.5 shows a better model performance. If the curve changes to rectangle it is perfect classifier with AUC value 1.

In our case, AUC is nearly 0.98 which is really good considering all that was mentioned.

d.
The explanatory variable with the lowest p-value in the model, plays the most significant role in the prediction.
e.
According to the summary of our model, *G2* and *failures* are the explanatory variables with the most significant contribution to the model.

```
Call:
glm(formula = catG3 ~ failures + G2, family = binomial(link = "logit"),
    data = StudentsPerformance)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.91186   0.00028   0.01588   0.14760   2.39928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.9789     2.4607  -6.087 1.15e-09 ***
failures     -0.6456     0.3882  -1.663   0.0963 .
G2            1.6030     0.2453   6.536 6.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 433.50  on 394  degrees of freedom
Residual deviance: 104.12  on 392  degrees of freedom
AIC: 110.12

Number of Fisher Scoring iterations: 8
```

Figure 65: GLM model

$$log(\frac{p}{1-p}) = -14.978 - 0.645 \times failures + 1.603 \times G2$$

intercept : keeping all other predictors zero, the log odds ratio / odds radio of catG3 is -14.978 / exp(-14.978) = 4.85e-7

failures : keeping all other predictors constant for a unit increase in failures, the log odds ratio / odds radio of catG3 will decrease -0.645 / exp(-0.645) = 0.52

G2 : keeping all other predictors constant for a unit increase in G2 , the log odds ratio / odds radio of catG3 will increase 1.603 / exp(1.603) = 4.96

Produces a table of fit statistics for multiple glm models: AIC, AICc, BIC, p-value, pseudo R-squared (McFadden, Cox and Snell, Nagelkerke).
Smaller values for AIC, AICc, and BIC indicate a better balance of goodness-of-fit of the model and the complexity of the model. The goal is to find a model that adequately explains the data without having too many terms.
BIC tends to choose models with fewer parameters relative to AIC.

Rank <dbl>	Df.res <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>	McFadden <dbl>	Cox.and.Snell <dbl>	Nagelkerke <dbl>	p.value <dbl>
5	390	114.1	114.3	138	0.7645	0.5678	0.8523	9.060e-71
3	392	112.1	112.2	128	0.7598	0.5656	0.8489	1.498e-72

Figure 66: GLM model comparison

Model analysis :

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 21  2
         1  3 73

               Accuracy : 0.9495
                 95% CI : (0.8861, 0.9834)
    No Information Rate : 0.7576
    P-Value [Acc > NIR] : 3.298e-07

                  Kappa : 0.8605

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8750
            Specificity : 0.9733
         Pos Pred Value : 0.9130
         Neg Pred Value : 0.9605
             Prevalence : 0.2424
         Detection Rate : 0.2121
   Detection Prevalence : 0.2323
      Balanced Accuracy : 0.9242

       'Positive' Class : 0
```

Figure 67: Confusion matrix and accuracy

f.

catG3 is a binary numerical variable indicating whether you pass the test or not.
A perfect regression model needs to have a low false-positive rate and a low false-negative rate.
In minimizing these factors, we face a dilemma, and we have to decide in which case it is more harmful for us to make mistakes.
It will be costly to have a large false-positive. False-positive might ruin your study plans; failing a course might have some harmful effects on your future. But having false-negative, although still bad, is not as costly as false-positive. False-negative will make you study more, although it might cause depression. :))

Outcome	Utility
True Positive	1
True Negetive	1
False positive	-80
False Negetive	-10

$$U(p) = TP(p) + TN(p) - 80 \times FP(p) - 10 \times FN(p)$$

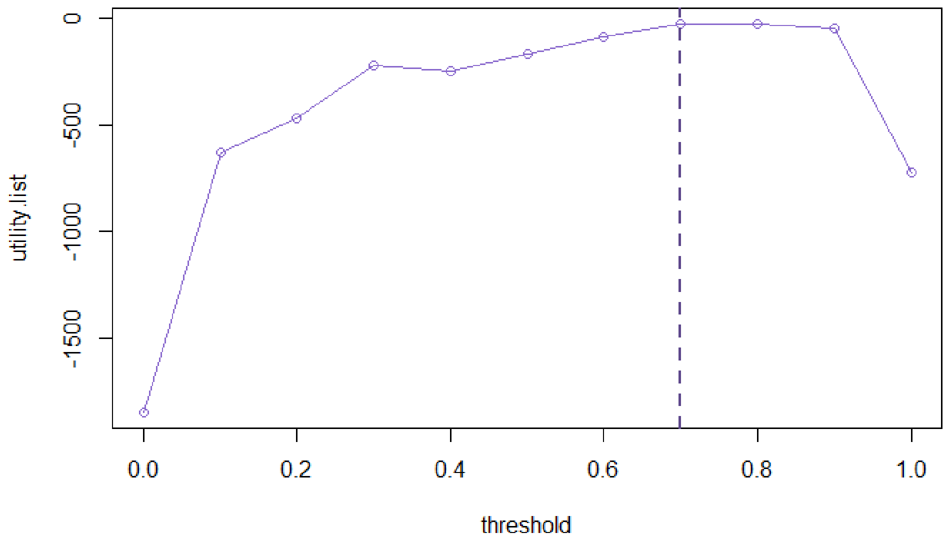


Figure 68: Utility curve

Best threshold : 0.7

Question 7

After converting the sums of Gs to a numeric binary variable :

```
Call:
glm(formula = Gsum ~ school + age + Fjob + Mjob + internet +
    romantic + health + failures + goout + studytime + absences +
    sex, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7955  -0.5313  -0.3384  -0.1397   2.9123

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.44436    3.08877  -2.086  0.03694 *
schoolMS     -0.15419    0.55473  -0.278  0.78105
age           0.28826    0.17245   1.672  0.09461 .
Fjobhealth   -0.33615    1.49316  -0.225  0.82188
Fjobother     0.54473    0.79098   0.689  0.49103
Fjobservices -0.45337    0.82518  -0.549  0.58272
Fjobteacher   0.79923    1.00316   0.797  0.42561
Mjobhealth   -1.90779    1.02396  -1.863  0.06244 .
Mjobother    -0.58612    0.53176  -1.102  0.27036
Mjobservices -0.45011    0.56543  -0.796  0.42600
Mjobteacher  -0.54123    0.75417  -0.718  0.47297
internetyes   0.25312    0.51771   0.489  0.62490
romanticyes   0.38277    0.39121   0.978  0.32788
health        0.05246    0.13736   0.382  0.70252
failures      1.80570    0.31693   5.697 1.22e-08 ***
goout         0.45450    0.17554   2.589  0.00962 **
studytime    -0.74356    0.28326  -2.625  0.00867 **
absences     -0.10105    0.03580  -2.822  0.00477 **
sexM         -0.78753    0.42097  -1.871  0.06138 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 301.62  on 296  degrees of freedom
Residual deviance: 201.84  on 278  degrees of freedom
AIC: 239.84

Number of Fisher Scoring iterations: 6
```

Figure 69: GLM model of all variables

Significant predictors are the ones with the $p-value$ smaller than 0.05 :

	p_value
(Intercept)	1
schoolMS	0
age	0
Fjobhealth	0
Fjobother	0
Fjobservices	0
Fjobteacher	0
Mjobhealth	0
Mjobother	0
Mjobservices	0
Mjobteacher	0
internetyes	0
romanticyes	0
health	0
failures	1
goout	1
studytime	1
absences	1
sexM	0

Figure 70: GLM model of all variables

According to figure 63, the variables that have significant p-value will be selected

$$Gsum \sim failures + goout + studytime + absence$$

Accuracy is 0.86, which is good enough for this model.
86% of the time, we can correctly predict whether a student will be on academic probation or not.
There are several statistics that can help us determine which predictor variables are most important in regression models. These statistics might not agree because the manner in which each one defines ”most important” is a bit different :

- P-value : Look for the predictor variable with the lowerst p-value
- Standardized regression coefficients : Look for the predictor variable with the largest absolute value for the standardized coefficient.
- Change in R-squared when the variable is added to the model last : Look for the predictor variable that is associated with the greatest increase in R-squared. (explained comprehensively in next question)

The variable with the most effect on academic probation is the variable with the least p-value, which is failures which makes sense.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 10  1
         1 12 75

               Accuracy : 0.8673
                 95% CI : (0.7838, 0.9274)
    No Information Rate : 0.7755
    P-Value [Acc > NIR] : 0.015801

                  Kappa : 0.5367

 Mcnemar's Test P-Value : 0.005546

            Sensitivity : 0.4545
            Specificity : 0.9868
         Pos Pred Value : 0.9091
         Neg Pred Value : 0.8621
             Prevalence : 0.2245
         Detection Rate : 0.1020
   Detection Prevalence : 0.1122
      Balanced Accuracy : 0.7207

       'Positive' Class : 0
```

Figure 71: Prediction accuracy

R Codes

```
---
title: "Statistical Inference"
output:
  pdf_document: default
  #html_notebook: default
---
<h1> Phase 2 </h1>
<h2> Dataset : Students Performance </h2>


### Question 0
####Refreshing the memory:
```{r}
set.seed(NULL)
StudentsPerformance <- read.csv("StudentsPerformance.csv")
head(StudentsPerformance)
summary(StudentsPerformance)

```
Question 1
Chosen varibales : *sex* and *Mjob*

a.

First we have to compute the proportions :
```{r warning=FALSE}
sample.size <- 200
sp.sample <- StudentsPerformance[sample(nrow(StudentsPerformance), sample.size), ]
sp.sampled.table <- table(sp.sample[,c("sex", "Fjob")])
sp.sampled.table

F.phat <- sp.sampled.table["F", "teacher"]/sum(sp.sampled.table["F",])
F.phat

M.phat <- sp.sampled.table["M", "teacher"]/sum(sp.sampled.table["M",])
M.phat
```

```{r}
SE <- sqrt( M.phat*(1-M.phat)/sum(sp.sampled.table["F",]) + M.phat*(1-M.phat)/sum(sp.sampled.table["M",]) )
SE
```

```{r}
(M.phat - F.phat) + c(-1, 1)*pnorm(0.975, lower.tail = F)*SE
```


b.
```{r}

p.pool <- (sp.sampled.table["F", "teacher"] + sp.sampled.table["M", "teacher"]) / (sum(sp.sampled.table["F",]) + sum(sp.sampled.table["M",]))
p.pool

SE.pool <- sqrt( p.pool * (1-p.pool)*( 1/ (sum(sp.sampled.table["F",])) + 1/ (sum(sp.sampled.table["M",]))))
```
```

```
SE.pool

p_value <- pnorm((M.phat - F.phat) / SE.pool, lower.tail = FALSE)

hypothesis.test <-function(pvalue, alpha = 0.05){
  if (pvalue < alpha){cat("Due to the fact that p-value (", round(pvalue, 3) , ") is smaller
      than", alpha, ", we reject the null hypothesis.")}
  else {cat("Due to the fact that p-value (", round(pvalue, 3) , ") is larger than " ,alpha
      , ",we fail to reject the null hypothesis.")}
  }

hypothesis.test(p_value)

```

```{r}

sp.sampled.table
sp.sampled.table.bind <- cbind(sp.sampled.table[,1] + sp.sampled.table[, 2], sp.sampled.
    table[, 3:5] )
sp.sampled.table.bind

chisq.test(sp.sampled.table , rescale.p = T)
chisq.test(sp.sampled.table.bind , rescale.p = T)

```

Question 2
Chosen varibales : *romantic*

```{r}
sample.size <- 15
romantic.sample <- StudentsPerformance[sample(nrow(StudentsPerformance), sample.size), ]$
    romantic
p.hat <- length(which(romantic.sample == 'yes' ))/sample.size
p.hat


simulation <-  data.frame(t(replicate(n = 1000, sample(levels(as.factor(StudentsPerformance$
    romantic)), size = sample.size, replace = TRUE))))

simulation.success <- apply(simulation, 1, function(x) length(which(x == 'yes')))
p_value <- length(which(simulation.success >= 8))/1000
hypothesis.test(p_value)

hist(simulation.success/sample.size)
```


Question 3
Chosen varibales : *Mjob*

a.
```{r}

sample.original <- StudentsPerformance$Mjob
round(table(sample.original) / length(StudentsPerformance$Mjob), 4)

sample.size <- 100

```

```
sample.unbiased <- sample(StudentsPerformance$Mjob, sample.size, replace = FALSE)
unbiased.table <- table(sample.unbiased)
unbiased.table


biased.prob <- ifelse(StudentsPerformance$Mjob == "teacher", 0.6, 0.4)
sample.biased <- sample(StudentsPerformance$Mjob, sample.size, prob = biased.prob)
biased.table <- table(sample.biased)
biased.table


original_prob <- c(prop.table(table(StudentsPerformance$Mjob)))
chisq.test(unbiased.table, p = original_prob)
chisq.test(biased.table, p = original_prob)
```
Chosen varibales : *Fjob*

b.
```{r}
Mjob.Fjob <- table(sp.sample[,c("Mjob","Fjob")])
Mjob.Fjob
chisq.test(Mjob.Fjob)

Mjob.Fjob.bind <- cbind(Mjob.Fjob[,1] + Mjob.Fjob[, 2] + Mjob.Fjob[, 4] + Mjob.Fjob[, 5],
    Mjob.Fjob[, 3] )
Mjob.Fjob.bind

chisq.test(Mjob.Fjob.bind,  rescale.p = T)
```
Question 4
Chosen varibales : *G1* , *failure* and *studytime*

a.

```{r}
library(ggplot2)
library("ggpubr")
library(GGally)
cor(StudentsPerformance$failures, StudentsPerformance$G1)
cor(StudentsPerformance$studytime, StudentsPerformance$G1)
cor(StudentsPerformance$goout, StudentsPerformance$G1)

ggpairs(StudentsPerformance[, c(11, 10, 14)])
```

b.
a. and b.
```{r}
#just failure
lm.G1.failure <- lm(G1 ~ failures, data = StudentsPerformance)
summary(lm.G1.failure)
lm.G1.failure
```


```{r}
#condition
library(ggplot2)
library(ggfortify)
autoplot(lm.G1.failure)+ theme_classic()
```

```

```
```{r}
#just studytime
lm.G1.studytime <- lm(G1 ~ studytime, data = StudentsPerformance)
summary(lm.G1.studytime)
lm.G1.studytime
```

```{r}
#condition
library(ggplot2)
library(ggfortify)
autoplot(lm.G1.studytime)+ theme_classic()
```

c.
```{r}
G1.failures <- ggplot(StudentsPerformance, aes(x = failures)) + geom_point(aes(y = G1), size = 2, colour = "grey") + stat_smooth(aes(x = failures, y = G1, linetype = "Linear Fit"), method = "lm", formula = y ~ x, se = F, color = "black")+ scale_linetype_manual(name = "Fit Type", values = c(2, 2)) + ggtitle("G1 vs. failures")

G1.failures + theme_classic()

G1.studytime <- ggplot(StudentsPerformance, aes(x = studytime)) + geom_point(aes(y = G1), size = 2, colour = "grey") + stat_smooth(aes(x = studytime, y = G1, linetype = "Linear Fit"), method = "lm", formula = y ~ x, se = F, color = "black")+ scale_linetype_manual(name = "Fit Type", values = c(2, 2)) + ggtitle("G1 vs. studytime")

G1.studytime + theme_classic()

```


e.
```{r}

compute.R.sqr <- function(model){
  SS.reg <- (anova(model)[[2]])[1] + (anova(model)[[2]])[2]
  SS.res <- (anova(model)[[2]])[3]
  R.sqr.f <- SS.reg / (SS.res + SS.reg)
  return(R.sqr.f)
}


base.model <- lm(G1 ~ sex, data = StudentsPerformance)
anova(base.model)
SS.reg <- (anova(base.model)[[2]])[1]
SS.res <- (anova(base.model)[[2]])[2]
R.sqr <- SS.reg / (SS.res + SS.reg)


#failure vs. studytime :
model.s.f <- lm(G1 ~ sex + failures, data = StudentsPerformance)
anova(model.s.f)
R.sqr.f <- compute.R.sqr(model.s.f)

model.s.s <- lm(G1 ~ sex + studytime, data = StudentsPerformance)
```

```
anova(model.s.s)
R.sqr.s <- compute.R.sqr(model.s.s)

R.square <- c(R.sqr, R.sqr.f, R.sqr.s)
df <- data.frame(R2 = round(R.square, 2))
df <- t(df)
colnames(df) <-c ("Base" , " + failures", " + studytime")
df

#G2 vs. G3
model.s.2 <- lm(G1 ~ sex + G2, data = StudentsPerformance)
anova(model.s.2)
R.sqr.2 <- compute.R.sqr(model.s.2)

model.s.3 <- lm(G1 ~ sex + G3, data = StudentsPerformance)
anova(model.s.3)
R.sqr.3 <- compute.R.sqr(model.s.3)

R.square. <- c(R.sqr, R.sqr.2, R.sqr.3)
df <- data.frame(R2 = round(R.square., 2))
df <- t(df)
colnames(df) <-c ("Base" , " + G2", " + G3")
df

```

e.
a.
```{r}
require(caTools)
set.seed(101)

sample.size <- 100
sp.sample <- StudentsPerformance[sample(nrow(StudentsPerformance), sample.size), ]

sample <- sample.split(sp.sample$G1, SplitRatio = 9/10)
G1.train <- subset(sp.sample, sample == TRUE)
G1.test  <- subset(sp.sample, sample == FALSE)
```


```{r}
#failues
lm.G1.failures <- lm(G1 ~ failures, data = G1.train)
summary(lm.G1.failures)
p_value <- summary(lm.G1.failures)$coefficients[8]
hypothesis.test(p_value)
```


```{r}
#studytime
lm.G1.studytime <- lm(G1 ~ studytime, data = G1.train)
summary(lm.G1.studytime)
p_value <- summary(lm.G1.studytime)$coefficients[8]
hypothesis.test(p_value)
```

```{r}
#G2
lm.G1.G2 <- lm(G1 ~ G2, data = G1.train)
summary(lm.G1.G2)
```

```
p_value <- summary(lm.G1.G2)$coefficients[8]
hypothesis.test(p_value)
```


```{r}
#G3
lm.G1.G3 <- lm(G1 ~ G3, data = G1.train)
summary(lm.G1.G3)
p_value <- summary(lm.G1.G3)$coefficients[8]
hypothesis.test(p_value)
```

b.
```{r}

calculate.CI <- function(model, alpha = 0.05){

  point.est <- summary(model)$coefficients[2]
  std.error <- summary(model)$coefficients[4]

  round(point.est + c(-1, 1) * pnorm(1 - alpha/2) * std.error, 3)
}

calculate.CI(lm.G1.failures)

calculate.CI(lm.G1.studytime)

calculate.CI(lm.G1.G2)

calculate.CI(lm.G1.G3)

```

c.
```{r}
predicted.s <-  round(predict(lm.G1.studytime, G1.test, type = "response"),1)
predicted.f <-  round(predict(lm.G1.failures, G1.test, type = "response"),1)
predicted.2 <-  round(predict(lm.G1.G2, G1.test, type = "response"),1)
predicted.3 <-  round(predict(lm.G1.G3, G1.test, type = "response"),1)



pred.actual <- data.frame(G1.test$G1, predicted.s, predicted.f, predicted.2, predicted.3)
colnames(pred.actual) <- c("Actual", "Predicted studytime", "Predicted failues", "Predicted
    G2", "Predicted G3")

```
d.
```{r}
# 0.1 * data_range =  error
error <- abs(G1.test$G1 - pred.actual)
error

succes.rate.list <- c()
for (predictor in 1:length(error)) {
  error.accepted <- length(which(error[predictor] <= 2))
  succes.rate <- paste((error.accepted / length(G1.test$G1))*100, "%")
  succes.rate.list <- c(succes.rate.list, succes.rate)
}


```

```
succes.rate <- data.frame(t(succes.rate.list[2:5]))
colnames(succes.rate) <- c("Predicted studytime", "Predicted failues", "Predicted G2", "
    Predicted G3")
succes.rate
```


```{r}
# Min-Max Accuracy Calculation
predictors <- data.frame(predicted.s, predicted.f, predicted.2, predicted.3)

mm.succes.rate.list <- c()
for (p in 1:length(predictors)) {
  actuals.preds <- data.frame(cbind(actuals = G1.test$G1, predicteds = predictors[p]))
  min.max.succes.rate <- paste(round((mean(apply(actuals.preds, 1, min) / apply(actuals.
      preds, 1, max)))*100, 2), "%")
  mm.succes.rate.list <- c(mm.succes.rate.list, min.max.succes.rate)
}

succes.rate <- data.frame((t(mm.succes.rate.list)))
colnames(succes.rate) <- c("Predicted studytime", "Predicted failues", "Predicted G2", "
    Predicted G3")
succes.rate


```


```{r}
# MAPE Calculation


mape.succes.rate.list <- c()
for (p in 1:length(predictors)) {
  actuals.preds <- data.frame(cbind(actuals = G1.test$G1, predicteds = predictors[p]))
  mape.succes.rate <- paste(round((mean(abs((actuals.preds$predicteds - actuals.preds$
      actuals))/actuals.preds$actuals) )*100, 2), "%")
  mape.succes.rate.list <- c(mm.succes.rate.list, min.max.succes.rate)
}



succes.rate <- data.frame((t(mape.succes.rate.list)))
colnames(succes.rate) <- c("Predicted studytime", "Predicted failues", "Predicted G2", "
    Predicted G3")
succes.rate



```


extra part
```{r}
library(ggplot2)
library("ggpubr")
library(GGally)
cor(StudentsPerformance$G2, StudentsPerformance$G1)
cor(StudentsPerformance$G3, StudentsPerformance$G1)


ggpairs(StudentsPerformance[, c(15, 16, 14)])
```

```
```
```{r}
#G2
lm.G1.G2 <- lm(G1 ~ G2, data = StudentsPerformance)
summary(lm.G1.G2)
lm.G1.G2

library(ggplot2)
library(ggfortify)
autoplot(lm.G1.G2)+ theme_classic()

G1.G2 <- ggplot(StudentsPerformance, aes(x = G2)) + geom_point(aes(y = G1), size = 2, colour
    = "grey") + stat_smooth(aes(x = G2, y = G1, linetype = "Linear Fit"), method = "lm",
    formula = y ~ x, se = F, color = "black")+ scale_linetype_manual(name = "Fit Type",
    values = c(2, 2)) + ggtitle("G1 vs. G2")

G1.G2 + theme_classic()

```
```{r}
#G2
lm.G1.G3 <- lm(G1 ~ G3, data = StudentsPerformance)
summary(lm.G1.G3)
lm.G1.G3

library(ggplot2)
library(ggfortify)
autoplot(lm.G1.G3)+ theme_classic()

G1.G3 <- ggplot(StudentsPerformance, aes(x = G3)) + geom_point(aes(y = G1), size = 2, colour
    = "grey") + stat_smooth(aes(x = G3, y = G1, linetype = "Linear Fit"), method = "lm",
    formula = y ~ x, se = F, color = "black")+ scale_linetype_manual(name = "Fit Type",
    values = c(2, 2)) + ggtitle("G1 vs. G3")

G1.G3 + theme_classic()
```

Question 5
Chosen response varibale : *G1*
####Chosen explanatory variables : *G2*, *goout*, *failures*, *studytime*, *sex* , *age*

a.

```{r message=FALSE, warning=FALSE}

library(GGally)
p_ <- GGally::print_if_interactive
pm <- ggpairs(StudentsPerformance[, c(3, 4, 7, 10, 11, 14, 15)], progress = FALSE) + theme_
    minimal()
p_(pm)

pm <- ggpairs(StudentsPerformance[, c(3, 4, 7, 10, 11, 15)], progress = FALSE) + theme_
    minimal()
p_(pm)


StudentsPerformance$sex <- ifelse(StudentsPerformance$sex == "F", 1, 0)
library(ggcorrplot)
ggcorrplot(cor(StudentsPerformance[, c(3, 4, 7, 10, 11, 14, 15)]) , type = "lower", lab =
    TRUE, outline.color = "white", colors = c("black", "white", "mediumpurple3"))

```

```
```

b.
```{r}
lm.model <- lm(G1 ~ G2 + goout +  failures + studytime + sex + age , data =
    StudentsPerformance)
summary(lm.model)

```

```{r}
plot(lm.model$residuals , pch = 16, col = "mediumpurple3") + abline(lm.model)
```


e.
```{r}
library(olsrr)
#forward - p-value
forward.selection.p <- ols_step_forward_p(lm.model, details = TRUE, prem  = 0.05)


#backward - p-value
backward.elimination.p <- ols_step_backward_p(lm.model, details = TRUE, prem  = 0.05)

```

```{r}
#forward - adjusted R-sqrt
library(rms)

best.pred <- c()

adj.r.square <- function(formula, dataset, k = 1) {
  n <- length(StudentsPerformance$G1)
  r.squared <- lrm(formula = formula , data = dataset)$stat["R2"]
  adjR2 <- 1 - (((n-1)/(n-k-1)) * (1-r.squared))
}

#step 1
adj.r.squared.list1 <- c()
names <- c("G2" , "goout" ,  "failures" , "studytime" , "sex" , "age")
adj.r.squared.list1 <- c(adj.r.square(G1 ~ G2, StudentsPerformance),
                         adj.r.square(G1 ~ goout, StudentsPerformance),
                         adj.r.square(G1 ~ failures, StudentsPerformance),
                         adj.r.square(G1 ~ studytime, StudentsPerformance),
                         adj.r.square(G1 ~ sex, StudentsPerformance),
                         adj.r.square(G1 ~ age, StudentsPerformance))



max.adj.r.squared <- names[which.max(adj.r.squared.list1)]
if (max(adj.r.squared.list1, 0) > 0) { best.pred <- c(best.pred, max.adj.r.squared) }
best.pred


#step 2
names <- c("G2 + goout" ,  "G2 + failures" , "G2 + studytime" , "G2 + sex" , "G2 + age")
adj.r.squared.list2 <- c(adj.r.square(G1 ~ goout + G2, StudentsPerformance, k = 2),
                         adj.r.square(G1 ~ failures + G2, StudentsPerformance, k = 2),
```

```
                         adj.r.square(G1 ~ studytime + G2, StudentsPerformance, k = 2),
                         adj.r.square(G1 ~ sex + G2, StudentsPerformance, k = 2),
                         adj.r.square(G1 ~ age + G2, StudentsPerformance, k = 2))


max.adj.r.squared <- names[which.max(adj.r.squared.list2 - max(adj.r.squared.list1))]
if (max(adj.r.squared.list2 - max(adj.r.squared.list1)) > 0) { best.pred <- c(best.pred, max
    .adj.r.squared) }
best.pred

#step 3
names <- c("G2 + age + goout" , "G2 + age +  failures" , "G2 + age + studytime" , "G2 + age
     + sex" )
adj.r.squared.list3 <- c(adj.r.square(G1 ~ goout + G2 + age , StudentsPerformance, k = 3),
                         adj.r.square(G1 ~ failures + G2 + age , StudentsPerformance, k = 3),
                         adj.r.square(G1 ~ studytime + G2 + age , StudentsPerformance, k = 3)
                         ,
                         adj.r.square(G1 ~ sex + G2+ age  , StudentsPerformance, k = 3))

max.adj.r.squared <- names[which.max(adj.r.squared.list3 - max(adj.r.squared.list2))]

if (max(adj.r.squared.list3 - max(adj.r.squared.list2)) > 0) { best.pred <- c(best.pred, max
    .adj.r.squared) }
best.pred


#step 4
names <- c("G2 + age +  failures + goout" , "G2 + age +  failures + studytime" , "G2 + age
    + failures + sex" )
adj.r.squared.list4 <- c(adj.r.square(G1 ~ goout + G2 + age +  failures ,
    StudentsPerformance, k = 4),
                         adj.r.square(G1 ~ studytime + G2 + age +  failures ,
                             StudentsPerformance, k = 4),
                         adj.r.square(G1 ~ sex + G2 + age +  failures , StudentsPerformance,
                             k = 4))

adj.r.squared.list4
max.adj.r.squared <- names[which.max(adj.r.squared.list4 - max(adj.r.squared.list4))]
adj.r.squared.list4 - max(adj.r.squared.list3)

if (max(adj.r.squared.list3 - max(adj.r.squared.list2)) > 0) { best.pred <- c(best.pred, max
    .adj.r.squared) }
best.pred


all.adj.r.squared <- c(max(adj.r.squared.list1), max(adj.r.squared.list2), max(adj.r.squared
    .list3), max(adj.r.squared.list4))

model <- data.frame(best.pred, all.adj.r.squared)
model


```
```{r}
#backward - adjusted R-sqrt
library(rms)

fullmodel.adj.r.sqr <- adj.r.square(G1 ~ G2 + goout +  failures + studytime + sex + age,
    StudentsPerformance ,k = 6)
best.pred <- c()
```

```r
#step 1
adj.r.squared.list1 <- c()
names <- c("G2 + goout +  failures + studytime + sex" , "G2 + goout +  failures + studytime
    + age" ,
           "G2 + goout +  failures + sex + age" , "G2 + goout + studytime + sex + age" ,
           "G2 +  failures + studytime + sex + age" , "goout +  failures + studytime + sex +
                age")
adj.r.squared.list1 <- c(adj.r.square(G1 ~ G2 + goout +  failures + studytime + sex,
    StudentsPerformance, k = 5),
                         adj.r.square(G1 ~ G2 + goout +  failures + studytime + age,
                             StudentsPerformance, k = 5),
                         adj.r.square(G1 ~ G2 + goout +  failures + sex + age,
                             StudentsPerformance, k = 5),
                         adj.r.square(G1 ~ G2 + goout + studytime + sex + age,
                             StudentsPerformance, k = 5),
                         adj.r.square(G1 ~ G2 +  failures + studytime + sex + age,
                             StudentsPerformance, k = 5),
                         adj.r.square(G1 ~ goout +  failures + studytime + sex + age,
                             StudentsPerformance, k = 5))



max.adj.r.squared <- names[which.max(adj.r.squared.list1 - fullmodel.adj.r.sqr)]
if ( (max(adj.r.squared.list1) - fullmodel.adj.r.sqr) > 0) { best.pred <- c(best.pred, max.
    adj.r.squared) }
best.pred


#step 2
adj.r.squared.list2 <- c()
names <- c("G2 + failures + studytime + sex" , "G2 +  failures + studytime + age" ,
           "G2 + failures + sex + age" , "G2 + studytime + sex + age",
           "failures + studytime + sex + age")
adj.r.squared.list2 <- c(adj.r.square(G1 ~ G2 + failures + studytime + sex,
    StudentsPerformance, k = 4),
                         adj.r.square(G1 ~ G2 + failures + studytime + age,
                             StudentsPerformance, k = 4),
                         adj.r.square(G1 ~ G2 + failures + sex + age, StudentsPerformance, k
                             = 4),
                         adj.r.square(G1 ~ G2 + studytime + sex + age, StudentsPerformance, k
                             = 4),
                         adj.r.square(G1 ~  failures + studytime + sex + age,
                             StudentsPerformance, k = 4))



max.adj.r.squared <- names[which.max(adj.r.squared.list2 - max(adj.r.squared.list1))]
if ((max(adj.r.squared.list2) - max(adj.r.squared.list1)) > 0) { best.pred <- c(best.pred,
    max.adj.r.squared) }
best.pred


#step 3
adj.r.squared.list3 <- c()
names <- c("G2 + failures + studytime" , "G2 + failures + age" ,
           "G2 + studytime + age", "failures + studytime + age")
adj.r.squared.list3 <- c(adj.r.square(G1 ~ G2 + failures + studytime, StudentsPerformance, k
     = 3),
                         adj.r.square(G1 ~ G2 + failures + age, StudentsPerformance, k = 3),
                         adj.r.square(G1 ~ G2 + studytime + age, StudentsPerformance, k = 3),
                         adj.r.square(G1 ~  failures + studytime + age, StudentsPerformance, k
                              = 3))
```

```
max.adj.r.squared <- names[which.max(adj.r.squared.list3 - max(adj.r.squared.list2))]
if ((max(adj.r.squared.list3) - max(adj.r.squared.list2)) > 0) { best.pred <- c(best.pred,
    max.adj.r.squared) }
best.pred


#step 4
adj.r.squared.list4 <- c()
names <- c("G2 + failures" , "G2 + age", "failures + age")
adj.r.squared.list4 <- c(adj.r.square(G1 ~ G2 + failures, StudentsPerformance, k = 2),
                         adj.r.square(G1 ~ G2 + age, StudentsPerformance, k = 2),
                         adj.r.square(G1 ~ failures + age, StudentsPerformance, k = 2))



max.adj.r.squared <- names[which.max(adj.r.squared.list4 - max(adj.r.squared.list3))]
if ((max(adj.r.squared.list4) - max(adj.r.squared.list3)) > 0) { best.pred <- c(best.pred,
    max.adj.r.squared) }
best.pred


all.adj.r.squared <- c(max(adj.r.squared.list1), max(adj.r.squared.list2), max(adj.r.squared
    .list3))

model <- data.frame(best.pred, all.adj.r.squared)
model

```


```{r}
final.model <- lm(G1 ~ G2 + failures + age , data = StudentsPerformance)
summary(final.model)
```

f.
```{r}
#linearity
data <- data.frame(G2 = StudentsPerformance$G2, residuals = final.model$residuals)
ggplot(data = data,aes(G2, residuals)) + geom_point(color = "mediumpurple3", alpha = 0.5) +
    stat_smooth(method = lm, se= F, color = "black") + theme_classic()

data <- data.frame(failures = StudentsPerformance$failures, residuals = final.model$
    residuals)
ggplot(data = data,aes(failures, residuals)) + geom_point(color = "mediumpurple3", alpha =
    0.5) + stat_smooth(method = lm, se= F, color = "black") + theme_classic()

data <- data.frame(age = StudentsPerformance$age, residuals = final.model$residuals)
ggplot(data = data,aes(age, residuals)) + geom_point(color = "mediumpurple3", alpha = 0.5) +
     geom_hline( yintercept = 0, size = 1) + theme_classic()

#nearly normal
ggplot(final.model, aes(sample = final.model$residuals)) + stat_qq(col = "mediumpurple3",
    alpha = 0.5) + stat_qq_line() + theme_classic()

ggplot(data = final.model,aes(final.model$residuals)) + geom_histogram(bins = 20, col = "
    mediumpurple2", fill="mediumpurple3", alpha = 0.5) + theme_classic()

#cons. var
```

```
ks.test(unique(final.model$residuals), "pnorm", mean=0, sd=1)
ggplot(data = final.model,aes(final.model$fitted, final.model$residuals)) + geom_point(color
    = "mediumpurple3", alpha = 0.5) + stat_smooth(method = lm, se= F, color = "black") +
    theme_classic()

```



```{r}
library(ggplot2)
library(ggfortify)
autoplot(final.model)+ theme_classic()
```

g.
```{r}
library(caret)
model <- trainControl(method = "cv", number = 5)
fullmodel.cv <- train(G1 ~ G2 + goout +  failures + studytime + sex + age, data =
    StudentsPerformance, trControl = model, method = "lm")

bestmodel.cv <- train(G1 ~ G2 + failures + age, data = StudentsPerformance, trControl =
    model, method = "lm")


fullmodel.cv
bestmodel.cv


fullmodel.cv$finalModel
bestmodel.cv$finalModel

allfolds <- bestmodel.cv$resample
```

Question 6

```{r}
StudentsPerformance$catG3 <- ifelse(StudentsPerformance$G3 < 10, 0, 1)

sample <- sample.split(StudentsPerformance$catG3, SplitRatio = 3/4)
train <- subset(StudentsPerformance, sample == TRUE)
test  <- subset(StudentsPerformance, sample == FALSE)


```


Chosen response varibale : *catG3*
####Chosen explanatory variables : *failures*, *studytime*, *G2* and *sex*

a.

```{r}
model.gml <- glm(catG3 ~ failures + studytime + G2 + sex , family = binomial(link='logit'),
     data = train)

summary(model.gml)
```

```
```

```
##### b.
```{r}

female.prob <- seq(0, 1.01, 0.01)
OR.ratio = abs(summary(model.gml)$coefficients[3])

pred.y <- function(x) {
  return ((OR.ratio*x/(1-x)) / (1 + (OR.ratio*x/(1-x))))
}
male.prob <- sapply(female.prob, pred.y)
plot(male.prob, female.prob, type = "l", col = "mediumpurple3", lwd = 1.3) + abline(a=0, b
    =1)


```

c.
```{r message=FALSE, warning=FALSE}
library(pROC)
require(ROCR)

pred <-  predict(model.gml, train  , type="response")
roc(catG3 ~ pred, data = train, plot = TRUE, print.auc = TRUE, smooth = TRUE)


pred.t <-  predict(model.gml, test  , type="response")
roc(catG3 ~ pred.t, data = test, plot = TRUE, print.auc = TRUE, smooth = TRUE)


```

e.
```{r}
library(rcompanion)
better.model.gml <- glm(catG3 ~ failures + G2 , family = binomial(link='logit'),  data =
    StudentsPerformance)
summary(better.model.gml)

compareGLM(model.gml, better.model.gml)

```

f.


```{r}
library(caret)


confusion.matrix <- function(threshold){
  prediction.probability <- predict(better.model.gml, newdata = test, type = "response")
  pos.neg <- ifelse(prediction.probability > threshold, "1", "0")
  p.class <- factor(pos.neg, levels = c("0", "1"))
  cm <- confusionMatrix(p.class, as.factor(test$catG3))
  return(cm)}

confusion.matrix(0.5)


threshold <- seq(0, 1, by = 0.1)
utility.list <- c()
for (i in 1:length(threshold)){
```
```

```
    cm <- confusion.matrix(threshold[i])

   TP <- cm$table[1]
   FP <- cm$table[2]
   FN <- cm$table[3]
   TN <- cm$table[4]

   utility <- TP + TN - 80*FP - 10*FN
   utility.list <- c(utility.list, utility)

}

plot(threshold, utility.list, type = "o", col = "mediumpurple3", lwd = 1.3) + abline(v =
    threshold[which.max(utility.list)], col="mediumpurple4", lwd = 2, lty=2)




```

Question 7
```{r}

G.sum <- StudentsPerformance$G1 + StudentsPerformance$G2 + StudentsPerformance$G3
StudentsPerformance$Gsum <- ifelse(G.sum < 25, 1, 0)


sample <- sample.split(StudentsPerformance$Gsum, SplitRatio = 3/4)
train <- subset(StudentsPerformance, sample == TRUE)
test  <- subset(StudentsPerformance, sample == FALSE)



model.gml <- glm(Gsum ~ school + age + Fjob + Mjob + internet + romantic + health  +failures
     +goout + studytime + absences + sex  , family = binomial,  data = train)

summary(model.gml)

```


```{r}

p.values <- coef(summary(model.gml))[,4]

p_value <- ifelse(p.values < 0.05, 1, 0)
significant.pvalue <- data.frame(p_value)


```


```{r}

prediction.probability <- predict(model.gml, newdata = test, type = "response")
pos.neg <- ifelse(prediction.probability > 0.5, "0", "1")
p.class <- factor(pos.neg, levels = c("0", "1"))
cm <- confusionMatrix(p.class, as.factor(test$catG3))

cm
```

````
```
````

code.Rmd
````

Forward Selection Method

Candidate Terms:

1. G2
2. goout
3. failures
4. studytime
5. sex
6. age

We are selecting variables based on p value...

Forward Selection: Step 1

+ G2

Model Summary

R	0.851	RMSE	1.852
R-Squared	0.724	Coef. Var	17.174
Adj. R-Squared	0.723	MSE	3.429
Pred R-Squared	0.719	MAE	1.383

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3537.013	1	3537.013	1031.393	0.0000
Residual	1347.737	393	3.429		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	1.785	0.295		6.045	0.000	1.204	2.365
G2	0.733	0.023	0.851	32.115	0.000	0.688	0.778

Forward Selection: Step 2

+ age

Model Summary

R	0.854	RMSE	1.834
R-Squared	0.730	Coef. Var	17.013
Adj. R-Squared	0.729	MSE	3.365
Pred R-Squared	0.722	MAE	1.348

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3565.520	2	1782.760	529.735	0.0000
Residual	1319.229	392	3.365		

Total	4884.749	394

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.956	1.318		-1.484	0.139	-4.547	0.636
G2	0.746	0.023	0.866	32.383	0.000	0.701	0.791
age	0.215	0.074	0.078	2.910	0.004	0.070	0.360

Forward Selection: Step 3

+ failures

Model Summary

R	0.856	RMSE	1.825
R-Squared	0.733	Coef. Var	16.928
Adj. R-Squared	0.731	MSE	3.332
Pred R-Squared	0.723	MAE	1.363

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3582.019	3	1194.006	358.368	0.0000
Residual	1302.730	391	3.332		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.994	1.311		-1.521	0.129	-4.573	0.584
G2	0.718	0.026	0.834	27.563	0.000	0.667	0.769
age	0.244	0.075	0.088	3.270	0.001	0.097	0.390
failures	-0.323	0.145	-0.068	-2.225	0.027	-0.608	-0.038

Forward Selection: Step 4

+ studytime

Model Summary

R	0.857	RMSE	1.823
R-Squared	0.735	Coef. Var	16.906
Adj. R-Squared	0.732	MSE	3.323
Pred R-Squared	0.723	MAE	1.361

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3588.724	4	897.181	269.98	0.0000
Residual	1296.025	390	3.323		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-2.205	1.318		-1.673	0.095	-4.796	0.387
G2	0.715	0.026	0.830	27.384	0.000	0.664	0.766
age	0.239	0.075	0.087	3.204	0.001	0.092	0.385
failures	-0.298	0.146	-0.063	-2.043	0.042	-0.585	-0.011
studytime	0.159	0.112	0.038	1.420	0.156	-0.061	0.378

Forward Selection: Step 5

+ sex

Model Summary

R	0.858	RMSE	1.822
R-Squared	0.736	Coef. Var	16.896
Adj. R-Squared	0.732	MSE	3.319
Pred R-Squared	0.722	MAE	1.361

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3593.549	5	718.710	216.526	0.0000
Residual	1291.200	389	3.319		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-2.139	1.319		-1.622	0.106	-4.731	0.453

G2	0.711	0.026	0.825	26.946	0.000	0.659	0.762
age	0.240	0.075	0.087	3.227	0.001	0.094	0.387
failures	-0.309	0.146	-0.065	-2.119	0.035	-0.597	-0.022
studytime	0.203	0.117	0.048	1.728	0.085	-0.028	0.434
sex	-0.235	0.195	-0.033	-1.206	0.229	-0.618	0.148

No more variables to be added.

Variables Entered:

+ G2
+ age
+ failures
+ studytime
+ sex

Final Model Output

Model Summary

R	0.858	RMSE	1.822
R-Squared	0.736	Coef. Var	16.896
Adj. R-Squared	0.732	MSE	3.319
Pred R-Squared	0.722	MAE	1.361

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3593.549	5	718.710	216.526	0.0000
Residual	1291.200	389	3.319		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-2.139	1.319		-1.622	0.106	-4.731	0.453
G2	0.711	0.026	0.825	26.946	0.000	0.659	0.762
age	0.240	0.075	0.087	3.227	0.001	0.094	0.387
failures	-0.309	0.146	-0.065	-2.119	0.035	-0.597	-0.022
studytime	0.203	0.117	0.048	1.728	0.085	-0.028	0.434
sex	-0.235	0.195	-0.033	-1.206	0.229	-0.618	0.148

Backward Elimination Method

Candidate Terms:

1 . G2
2 . goout
3 . failures
4 . studytime
5 . sex
6 . age

We are eliminating variables based on p value...

x goout

Backward Elimination: Step 1

Variable goout Removed

Model Summary

R	0.858	RMSE	1.822
R-Squared	0.736	Coef. Var	16.896
Adj. R-Squared	0.732	MSE	3.319
Pred R-Squared	0.722	MAE	1.361

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3593.549	5	718.710	216.526	0.0000
Residual	1291.200	389	3.319		
Total	4884.749	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-2.139	1.319		-1.622	0.106	-4.731	0.453
G2	0.711	0.026	0.825	26.946	0.000	0.659	0.762
failures	-0.309	0.146	-0.065	-2.119	0.035	-0.597	-0.022
studytime	0.203	0.117	0.048	1.728	0.085	-0.028	0.434
sex	-0.235	0.195	-0.033	-1.206	0.229	-0.618	0.148
age	0.240	0.075	0.087	3.227	0.001	0.094	0.387

x sex

Backward Elimination: Step 2

Variable sex Removed

Model Summary

```
R                        0.857       RMSE                 1.823
R-Squared                0.735       Coef. Var           16.906
Adj. R-Squared           0.732       MSE                  3.323
Pred R-Squared           0.723       MAE                  1.361
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
----------------------------------------------------------------------
Regression    3588.724          4        897.181     269.98    0.0000
Residual      1296.025        390          3.323
Total         4884.749        394
----------------------------------------------------------------------

                                  Parameter Estimates
--------------------------------------------------------------------------------------
--
      model      Beta    Std. Error    Std. Beta      t        Sig      lower
upper
--------------------------------------------------------------------------------------
--
(Intercept)    -2.205         1.318                 -1.673    0.095    -4.796
0.387
         G2     0.715         0.026        0.830    27.384    0.000     0.664
0.766
   failures    -0.298         0.146       -0.063    -2.043    0.042    -0.585    -
0.011
  studytime     0.159         0.112        0.038     1.420    0.156    -0.061
0.378
        age     0.239         0.075        0.087     3.204    0.001     0.092
0.385
--------------------------------------------------------------------------------------
--


x studytime

Backward Elimination: Step 3

 Variable studytime Removed

                        Model Summary
--------------------------------------------------------------
R                        0.856       RMSE                 1.825
R-Squared                0.733       Coef. Var           16.928
Adj. R-Squared           0.731       MSE                  3.332
Pred R-Squared           0.723       MAE                  1.363
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
----------------------------------------------------------------------
Regression    3582.019          3       1194.006    358.368    0.0000
Residual      1302.730        391          3.332
Total         4884.749        394
----------------------------------------------------------------------

                                  Parameter Estimates
--------------------------------------------------------------------------------------
--
      model      Beta    Std. Error    Std. Beta      t        Sig      lower
upper
```

```
----------------------------------------------------------------------------------------
--
(Intercept)    -1.994        1.311                  -1.521    0.129    -4.573
0.584
         G2     0.718        0.026         0.834    27.563    0.000     0.667
0.769
   failures    -0.323        0.145        -0.068    -2.225    0.027    -0.608    -
0.038
        age     0.244        0.075         0.088     3.270    0.001     0.097
0.390
----------------------------------------------------------------------------------------
--


No more variables satisfy the condition of p value = 0.05


Variables Removed:

x goout
x sex
x studytime


Final Model Output
------------------

                        Model Summary
--------------------------------------------------------------
R                       0.856       RMSE                1.825
R-Squared               0.733       Coef. Var          16.928
Adj. R-Squared          0.731       MSE                 3.332
Pred R-Squared          0.723       MAE                 1.363
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F         Sig.
----------------------------------------------------------------------
Regression    3582.019          3       1194.006    358.368    0.0000
Residual      1302.730        391          3.332
Total         4884.749        394
----------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------------------
--
      model      Beta    Std. Error    Std. Beta      t        Sig      lower
upper
----------------------------------------------------------------------------------------
--
(Intercept)    -1.994        1.311                  -1.521    0.129    -4.573
0.584
         G2     0.718        0.026         0.834    27.563    0.000     0.667
0.769
   failures    -0.323        0.145        -0.068    -2.225    0.027    -0.608    -
0.038
        age     0.244        0.075         0.088     3.270    0.001     0.097
0.390
----------------------------------------------------------------------------------------
--
```