Statistical Inference: Project Phase I

Narjes Noorzad - 810196626

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

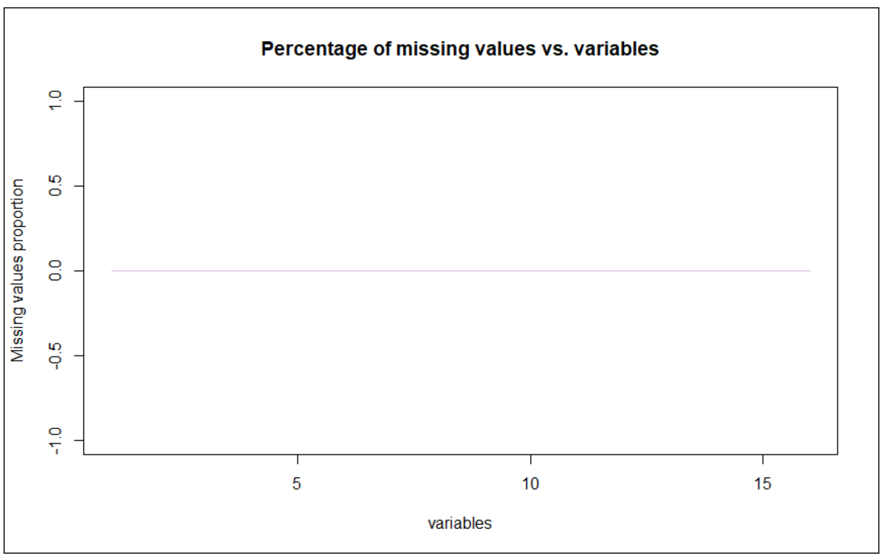


Figure 4: Line plot of missing value proportion

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.

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

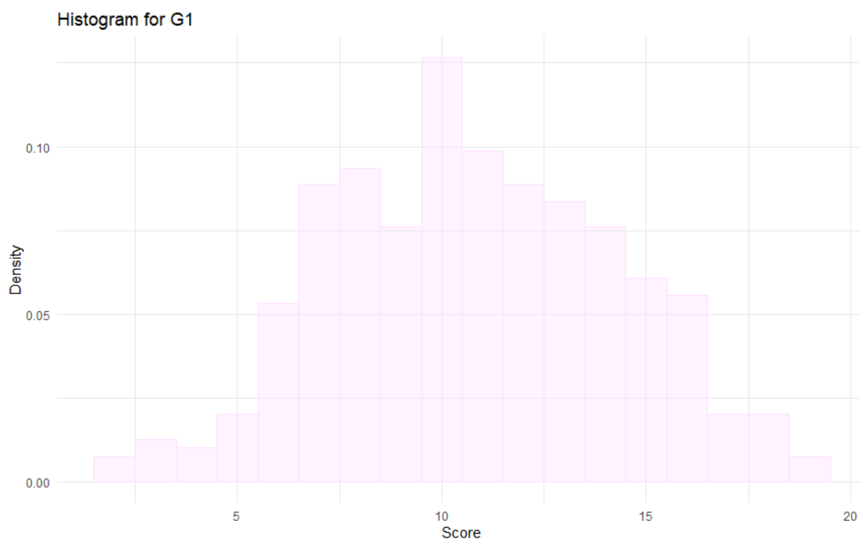


Figure 5: Histogram of G1

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).

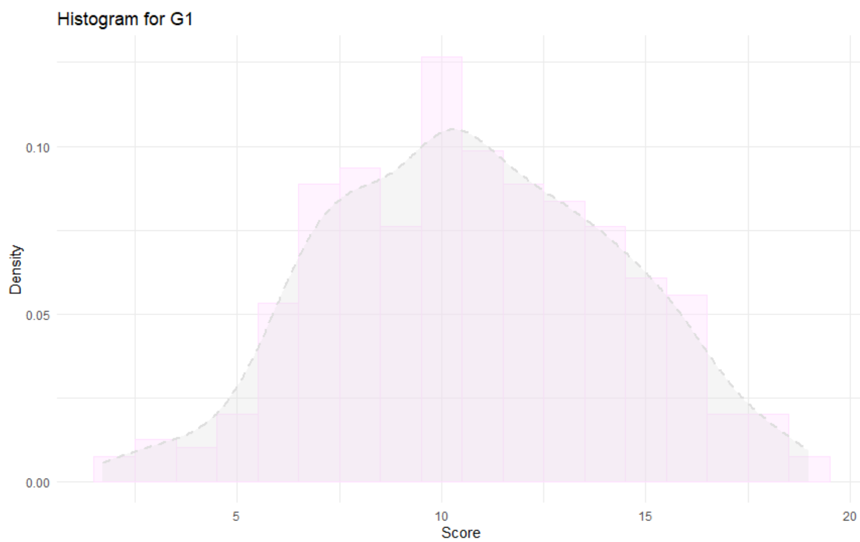


Figure 6: Histogram of G1 overlaid with density plot

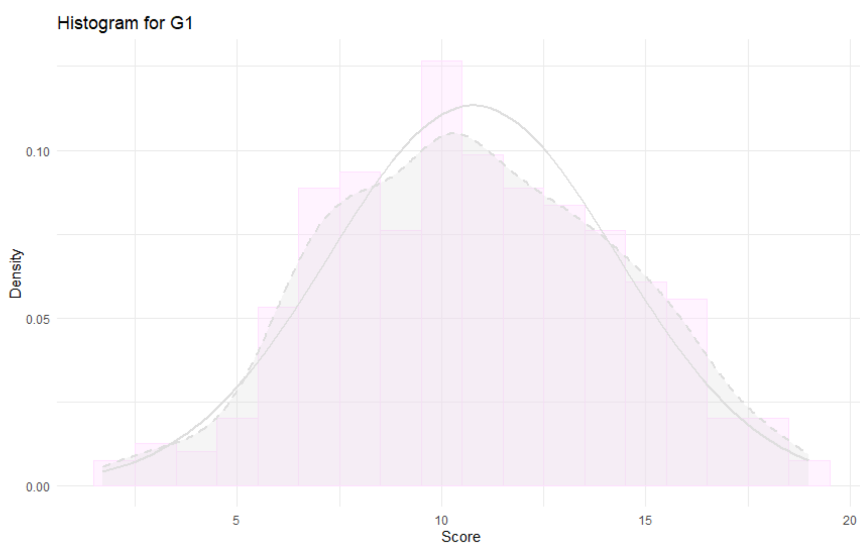


Figure 7: Histogram of G1 overlaid with fitted density plot and MLE density plot

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.

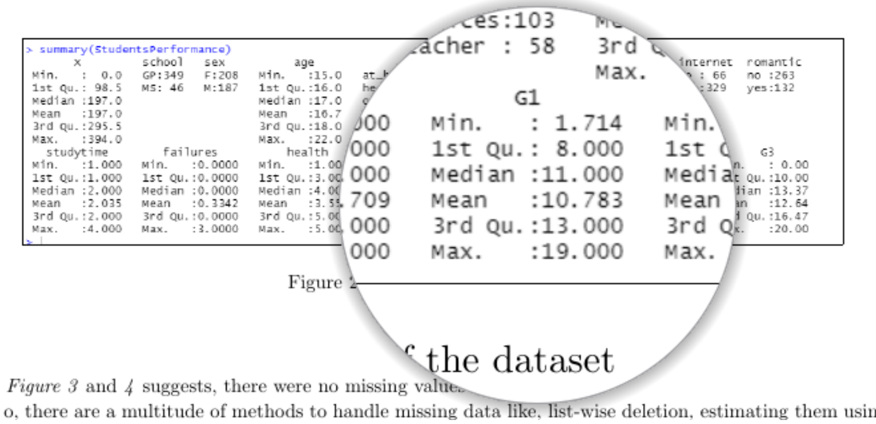


Figure 8: G1 under magnifier

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.)

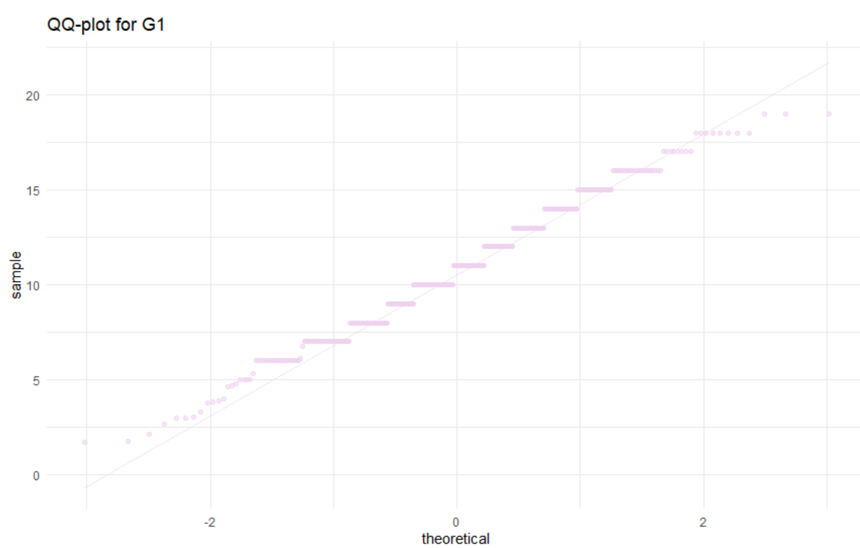


Figure 9: Normal Q-Q plot of G1

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 :

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

$$2. mean = median : Symmetric$$

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

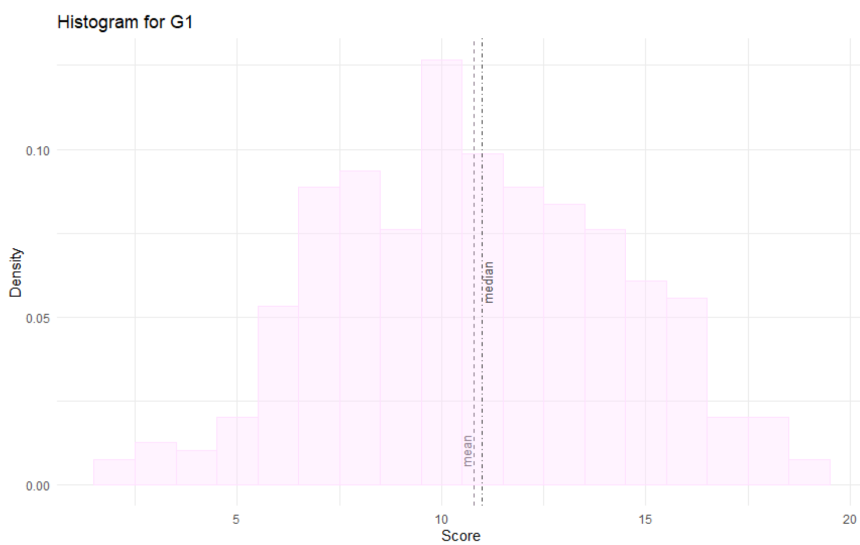


Figure 10: Median and mean marked on histogram of G1

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.

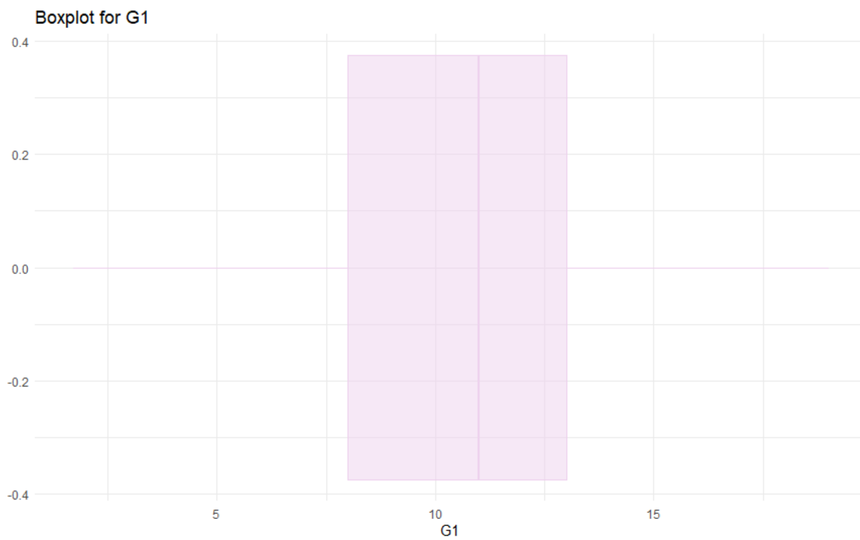


Figure 12: Boxplot of G1 to visualize outliers

```
>
> 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.

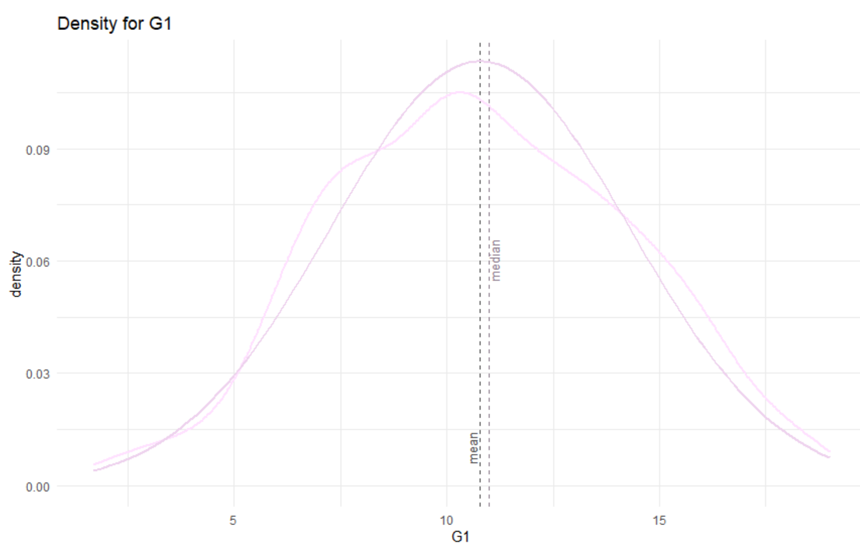


Figure 15: Median and mean marked on density of G1 - darker one is drawn using *dnorm*

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.

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

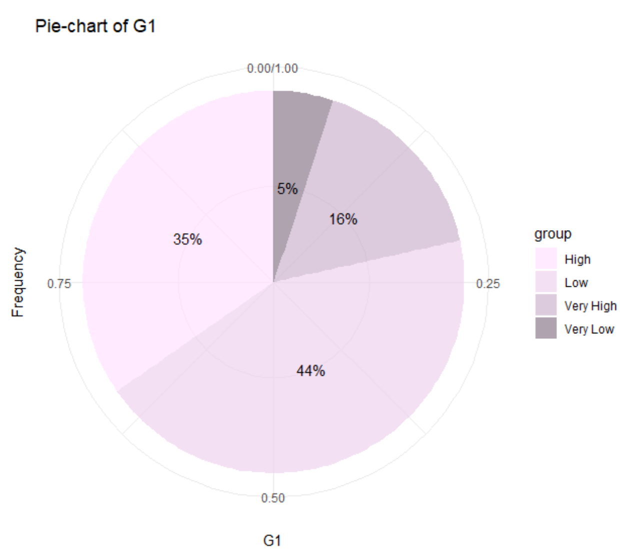


Figure 16: Piechart of G1 - 1^{st} method

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

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

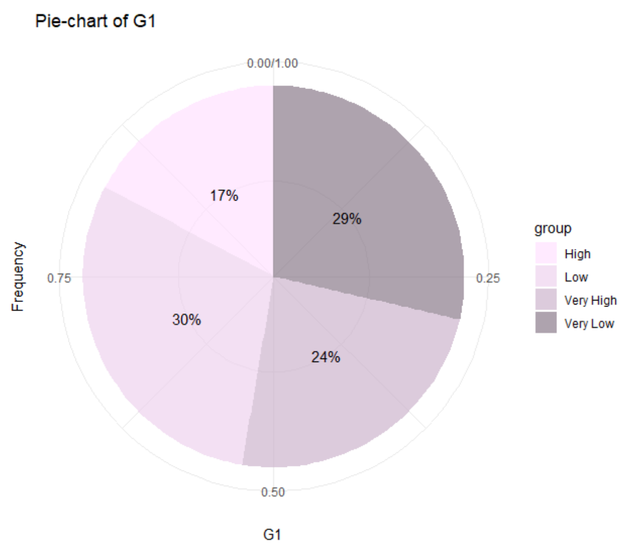


Figure 17: Piechart of G1 - 2^{nd} method

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.

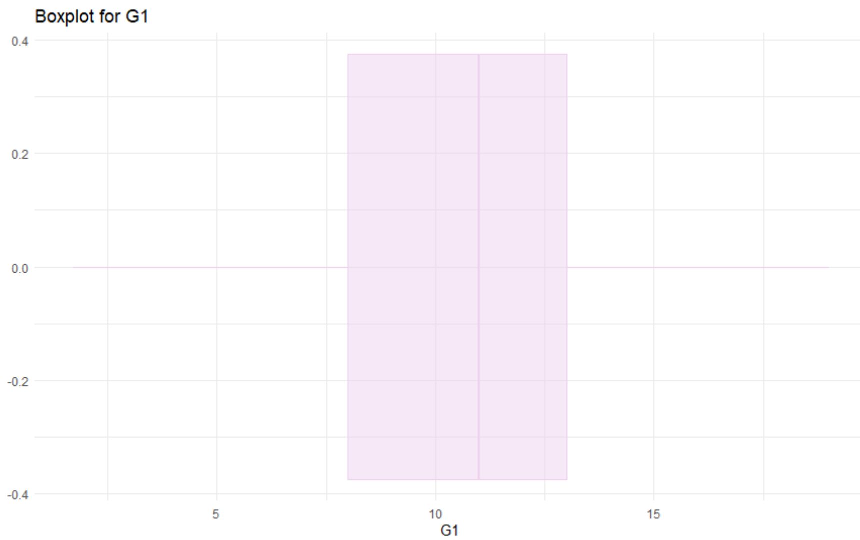


Figure 20: Boxplot of G1

```
> 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.

```
> female.freq
[1] 0.5265823
> male.freq
[1] 0.4734177
>
```

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.)

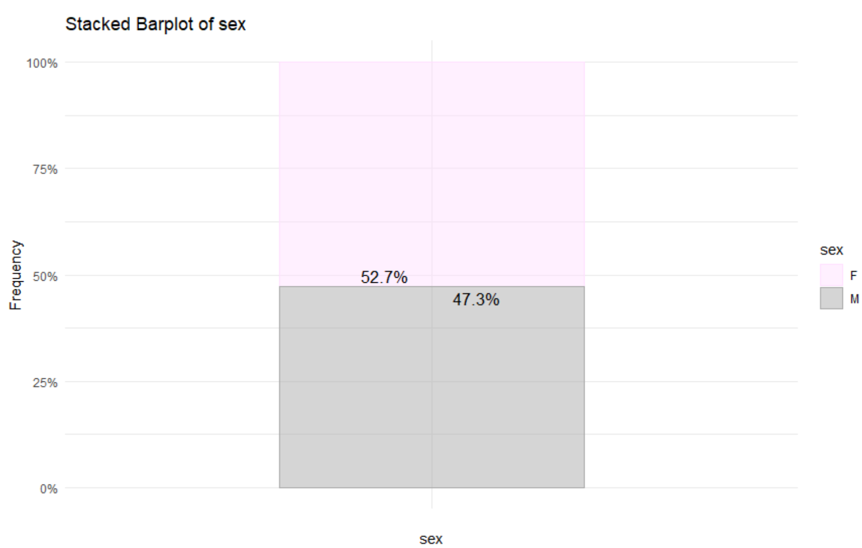


Figure 23: Stacked barplot of sex

c.

Barplots for categorical variables are like histograms for numerical variables.

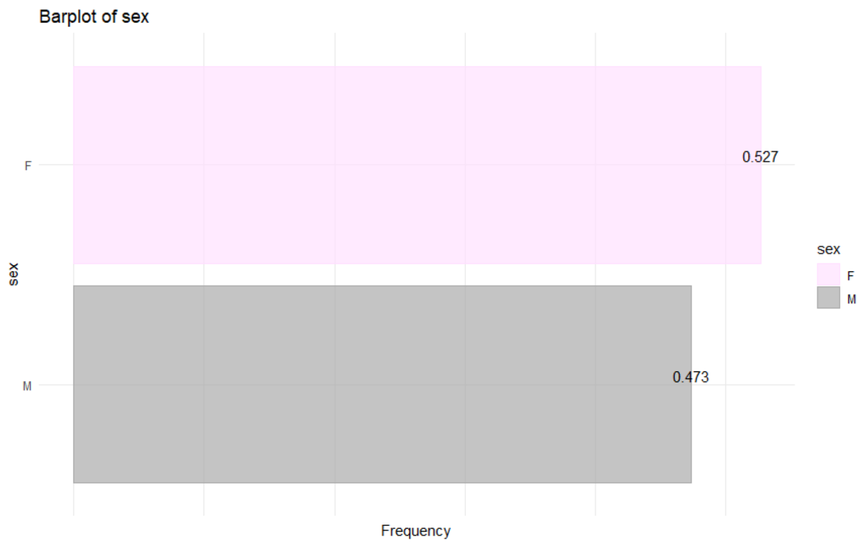


Figure 24: Horizontal barplot of sex

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.

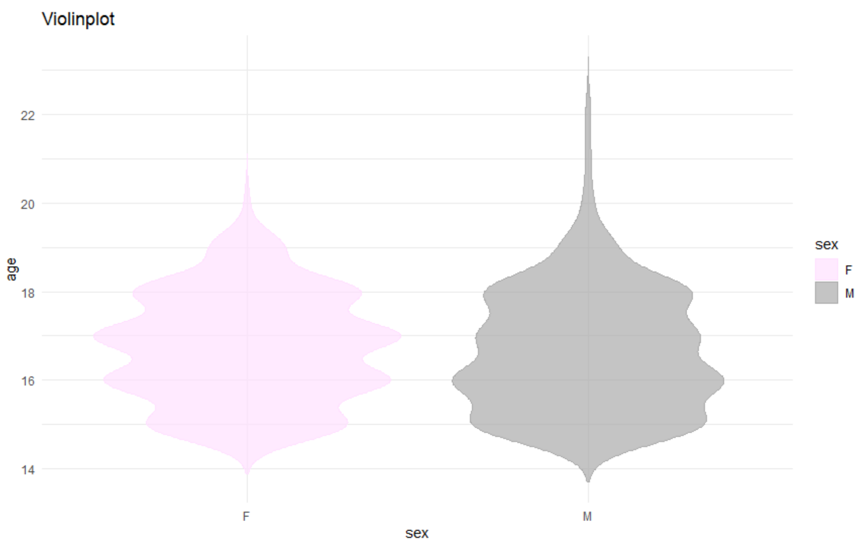


Figure 25: Violin plot of sex

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.

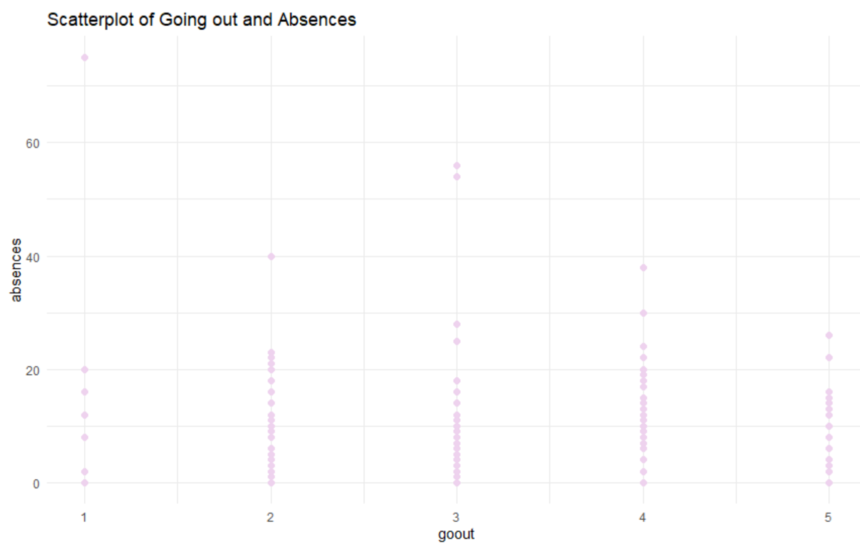


Figure 26: Scatterplot of goout and absences

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.

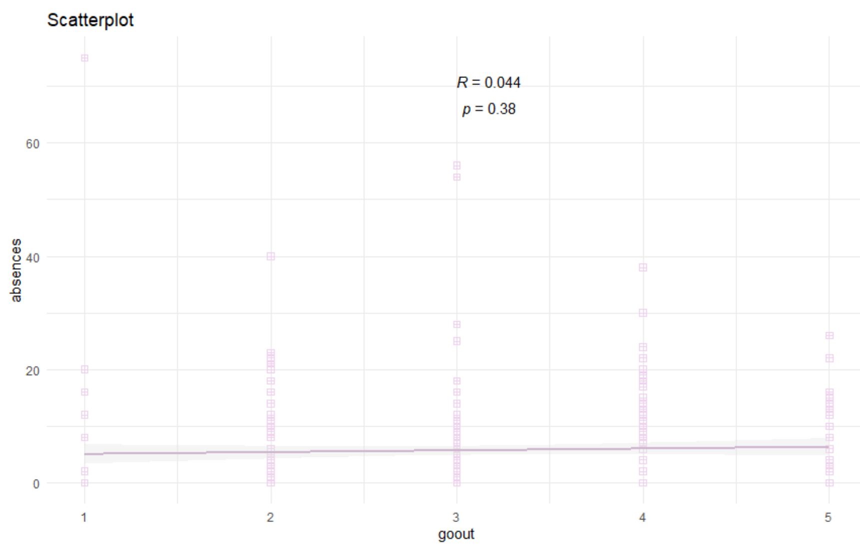


Figure 28: Scatterplot of goout and absences

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*

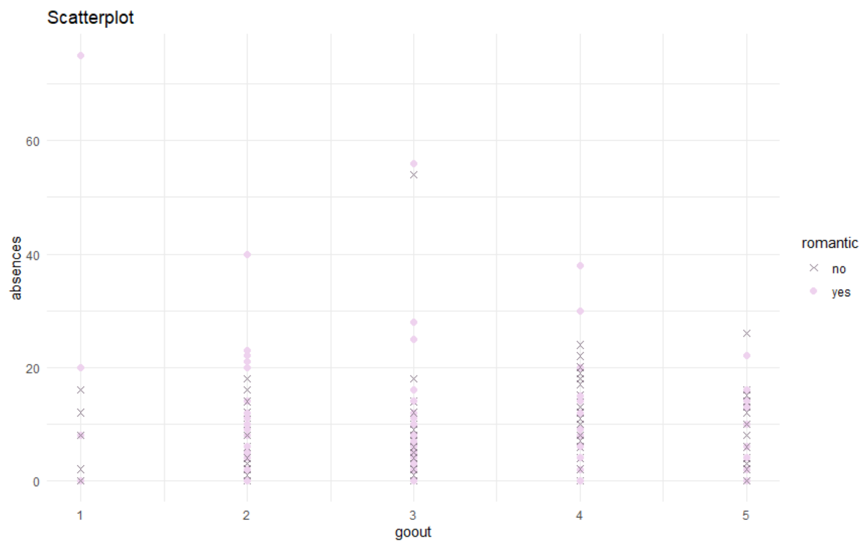


Figure 29: Scatterplot of goout and absences categorized by romantic

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.

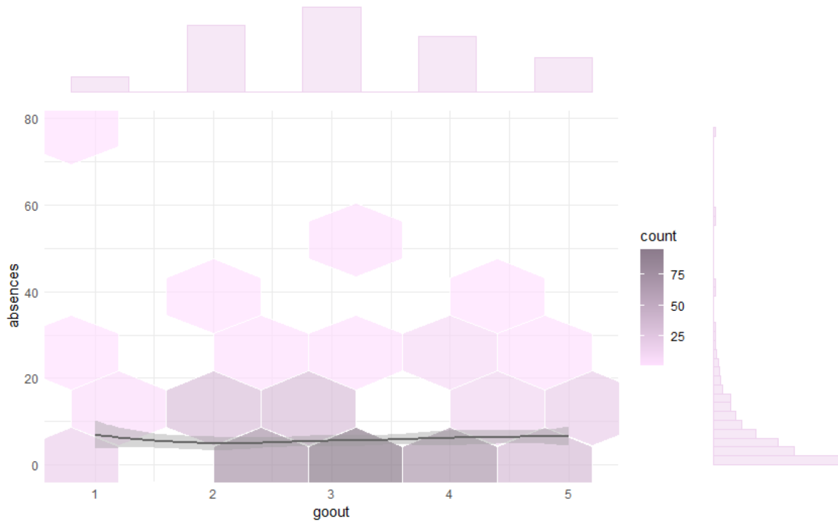


Figure 30: Hexbin plot, binsize = 5

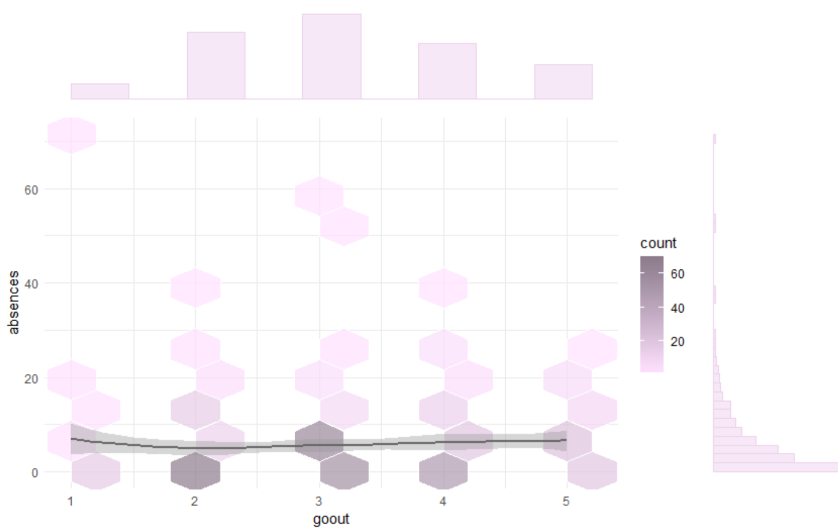


Figure 31: Hexbin plot, binsize = 10

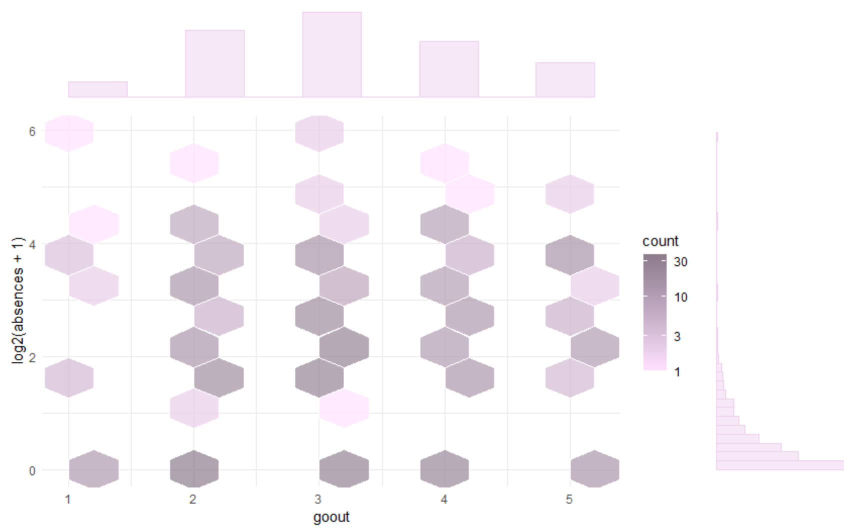


Figure 32: Hexbin plot, binsize = 10 (logarithmic)

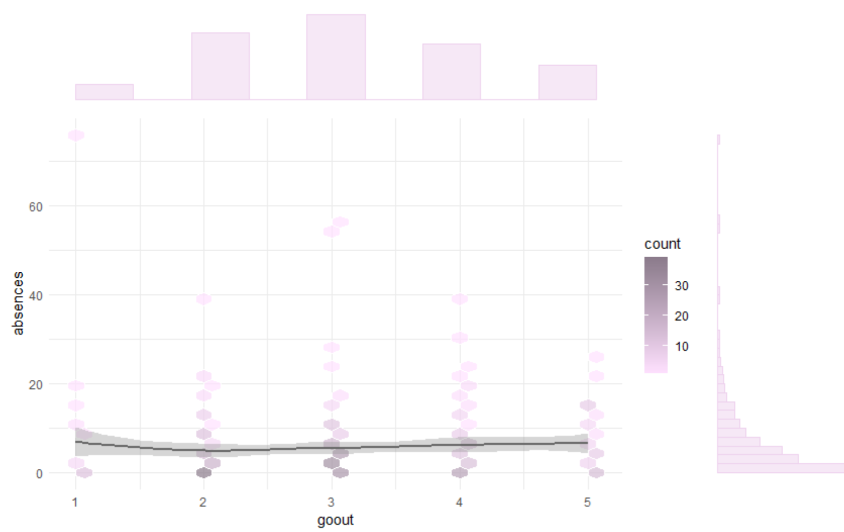


Figure 33: Hexbin plot, binsize = 30

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.

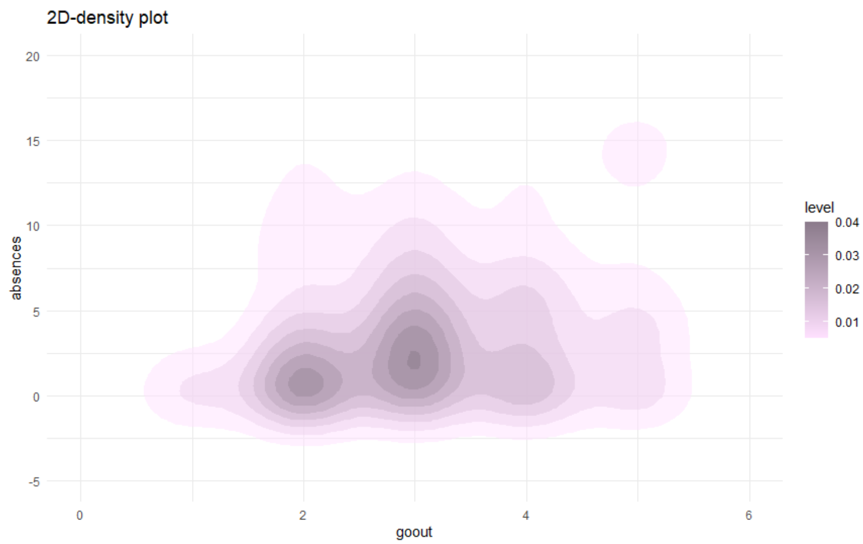


Figure 34: 2D density plot of goout and absences

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.

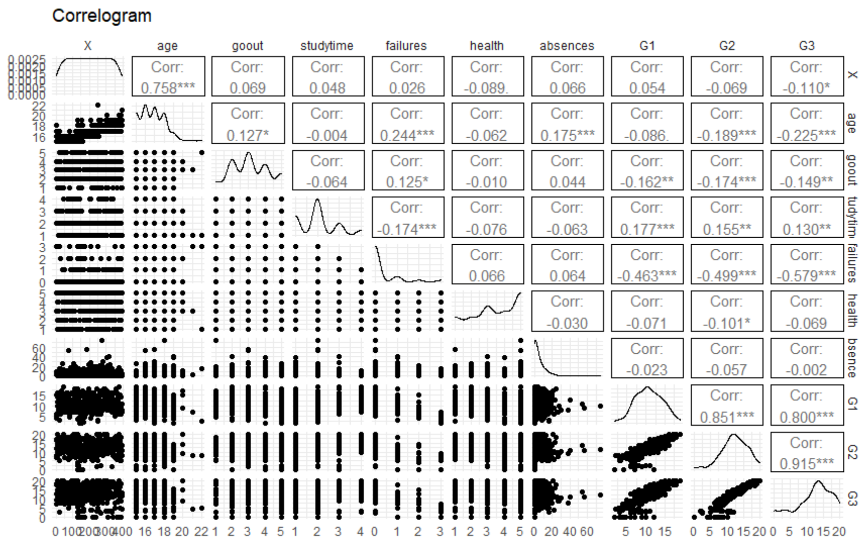


Figure 35: Bivarient Correlogram with pearson correlation

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

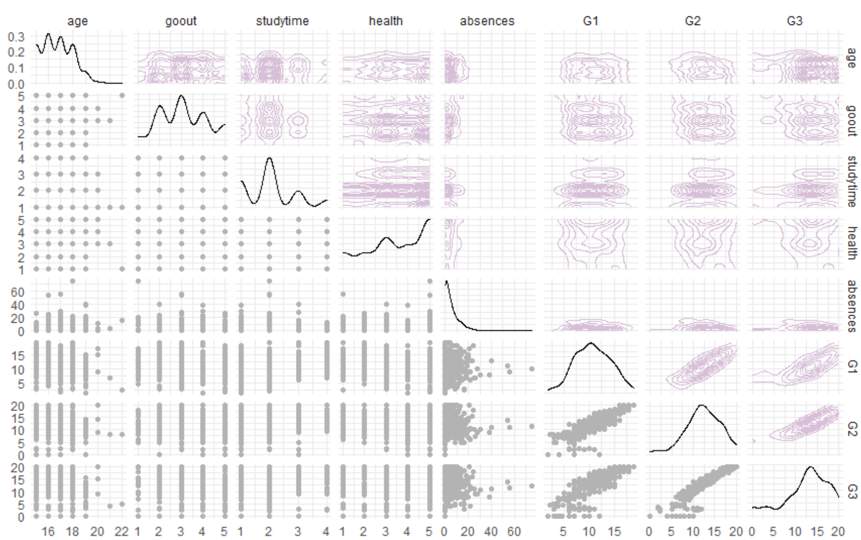


Figure 36: Bivarient Correlogram with density - scatterplot

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 :

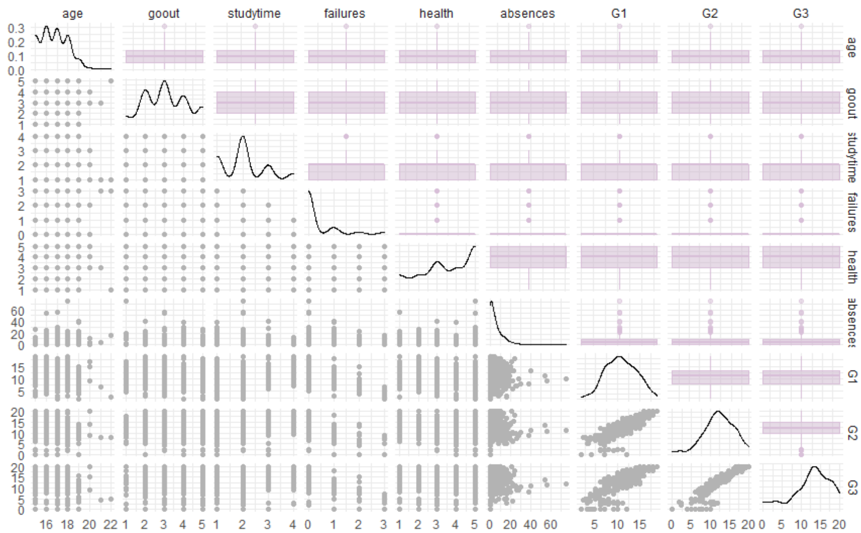


Figure 37: Bivarient Correlogram with barplot - scatterplot

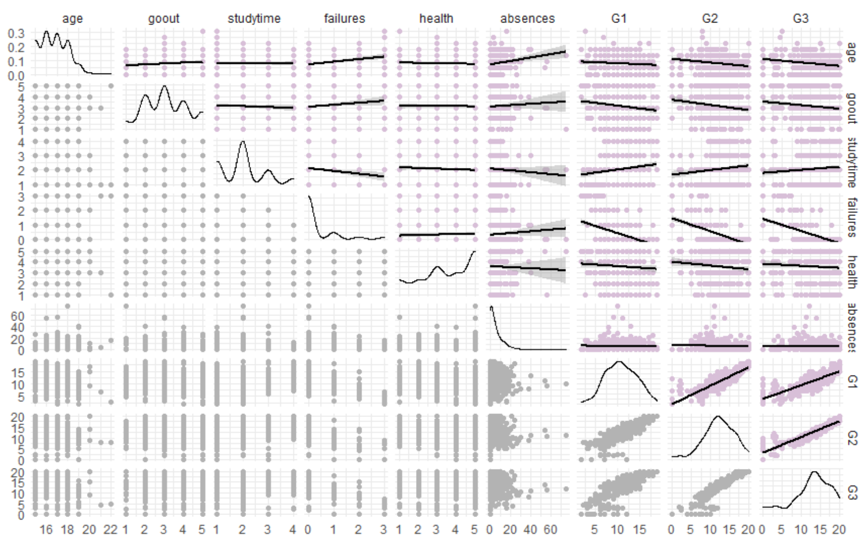


Figure 38: Bivarient Correlogram with linear assosiation - scatterplot

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 $G3$.

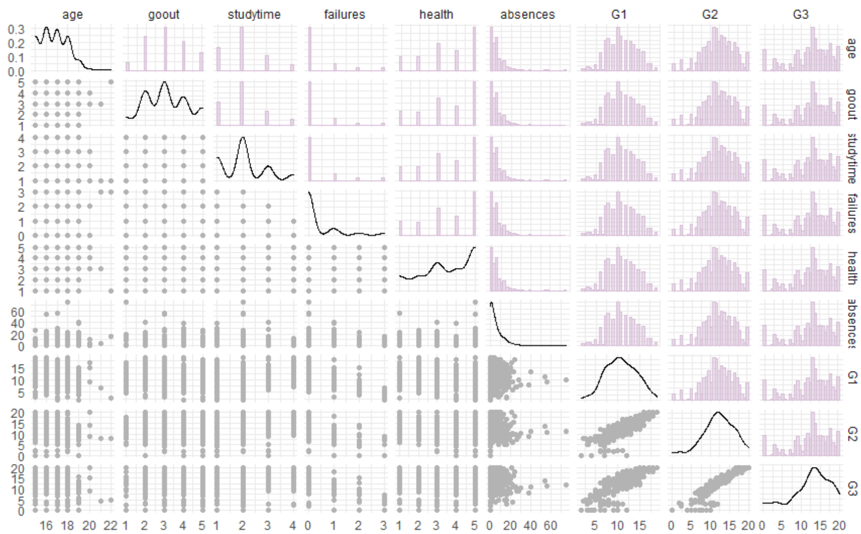


Figure 39: Bivarient Correlogram with barplot - scatterplot

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)

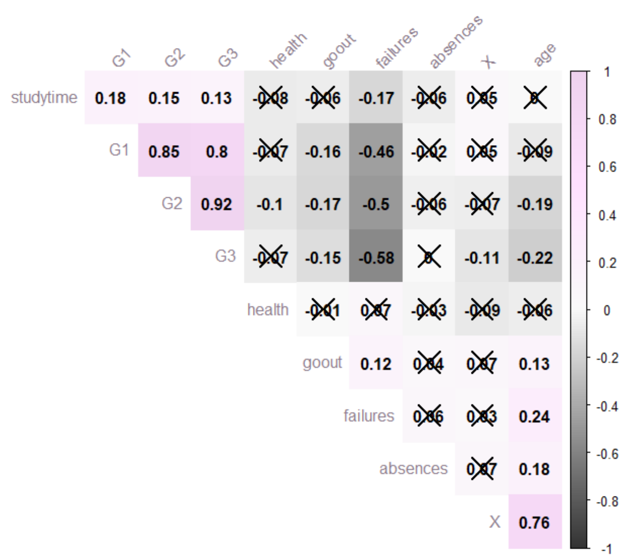


Figure 40: Heatmap correlogram of numerical values

c.

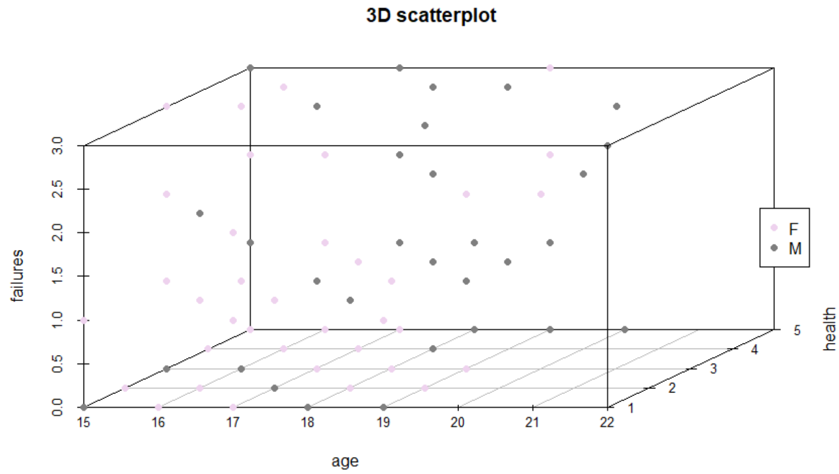


Figure 41: 3D scatterplot of age, failures and health colorized by sex

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.

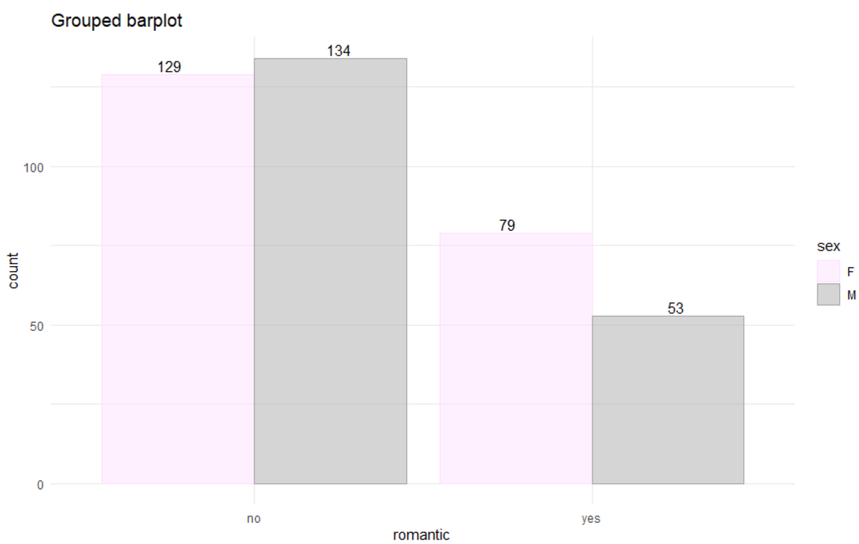


Figure 43: Grouped barplot of sex and romantic

c.

Question 5 continued on next page. . .

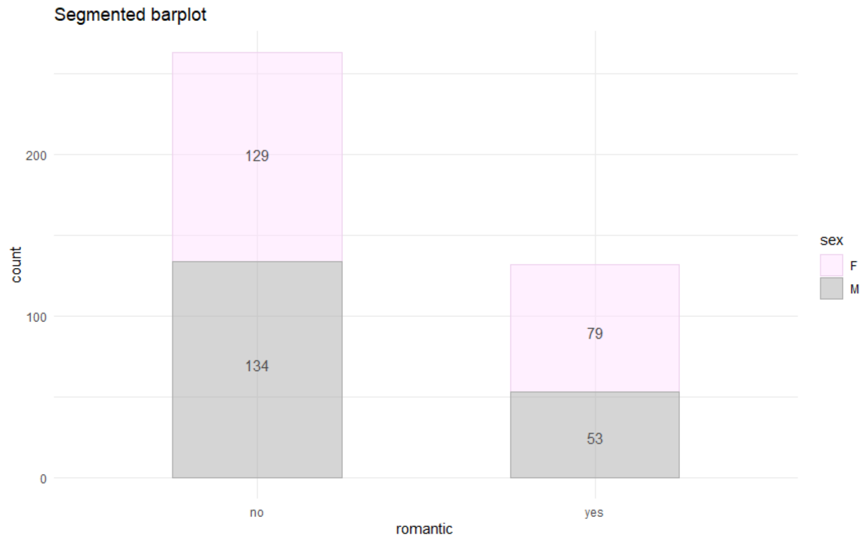


Figure 44: Segmented barplot of sex and romantic

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!

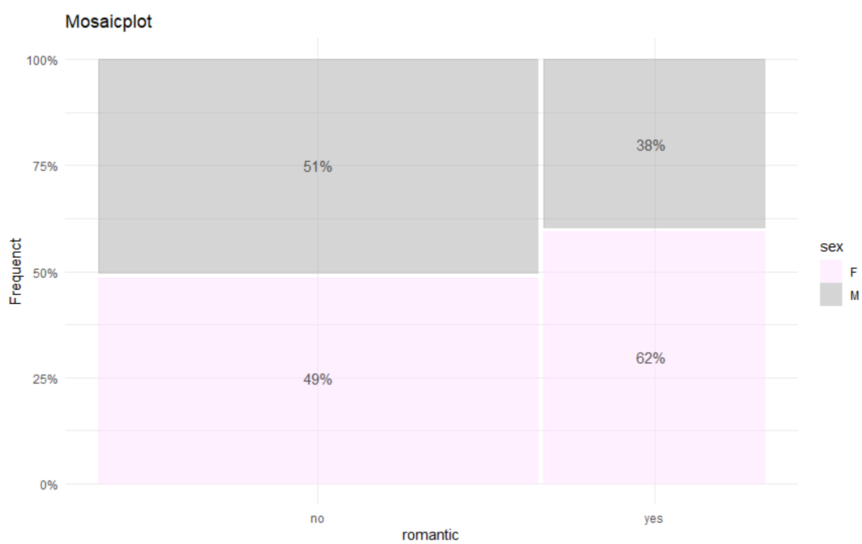


Figure 45: Mosaicplot of sex and romantic

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.

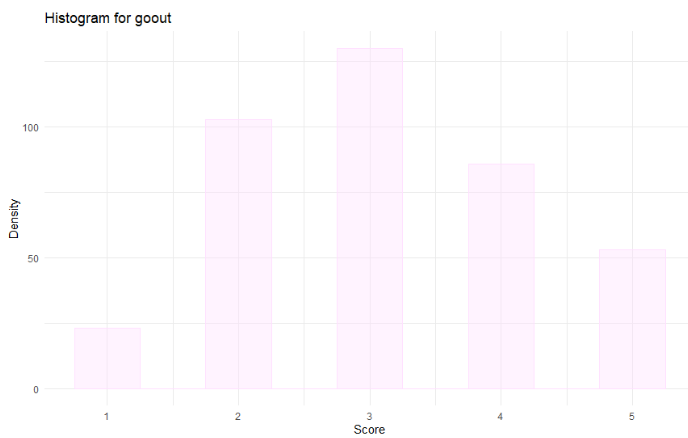


Figure 46: Histogram of goout

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.

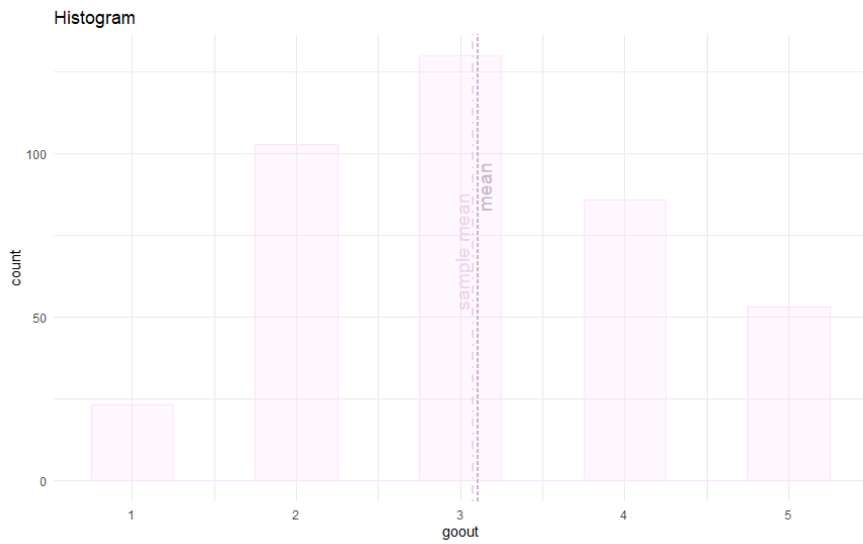


Figure 49: Histogram of goout marked with actual mean and sample mean

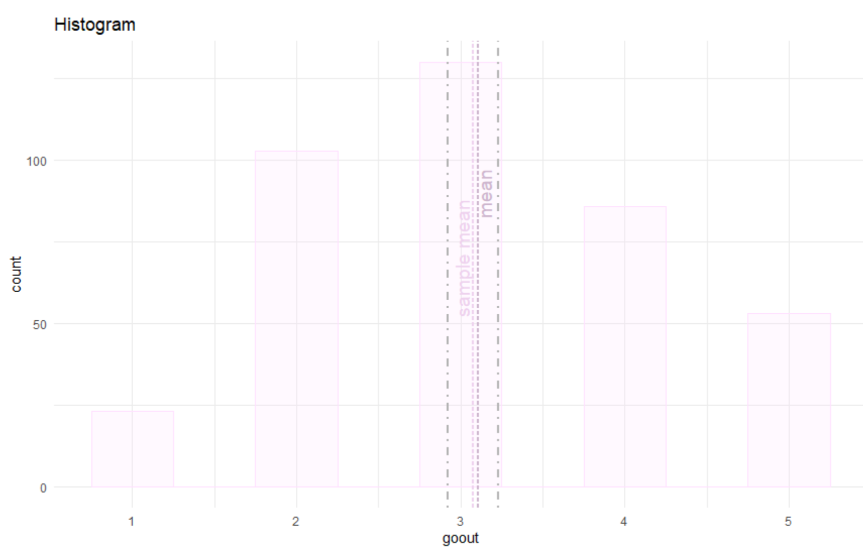


Figure 50: Histogram of goout marked with CI, actual mean and sample mean

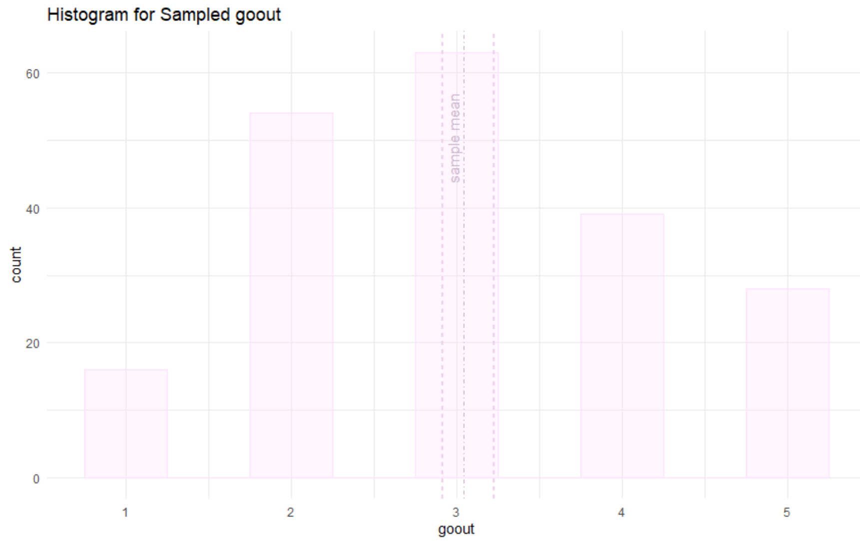


Figure 51: Histogram of sampled goout marked with CI and sample mean

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.

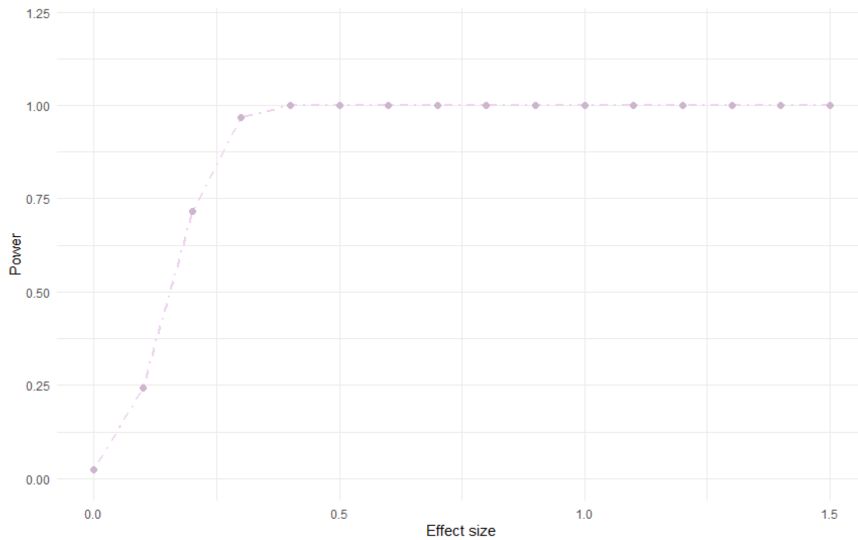


Figure 56: Relationship between effect size and power

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 :

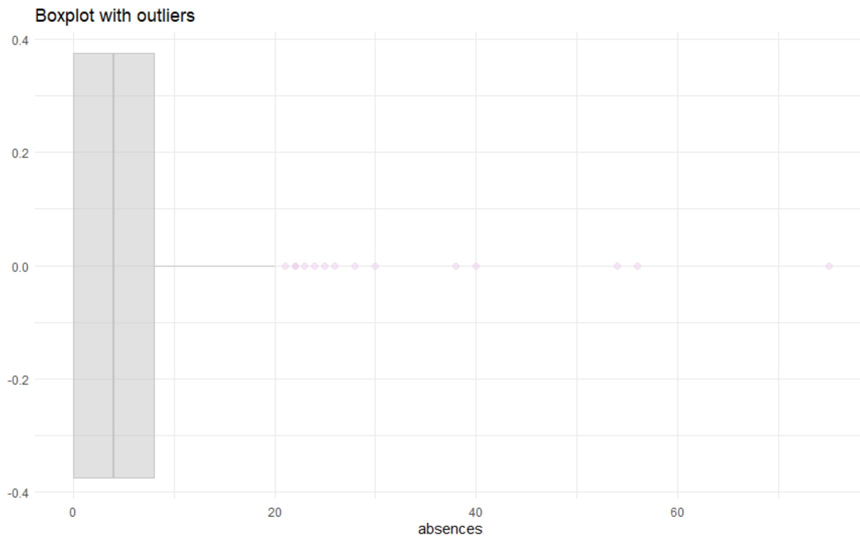
```
        Welch Two Sample t-test

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

```
"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.

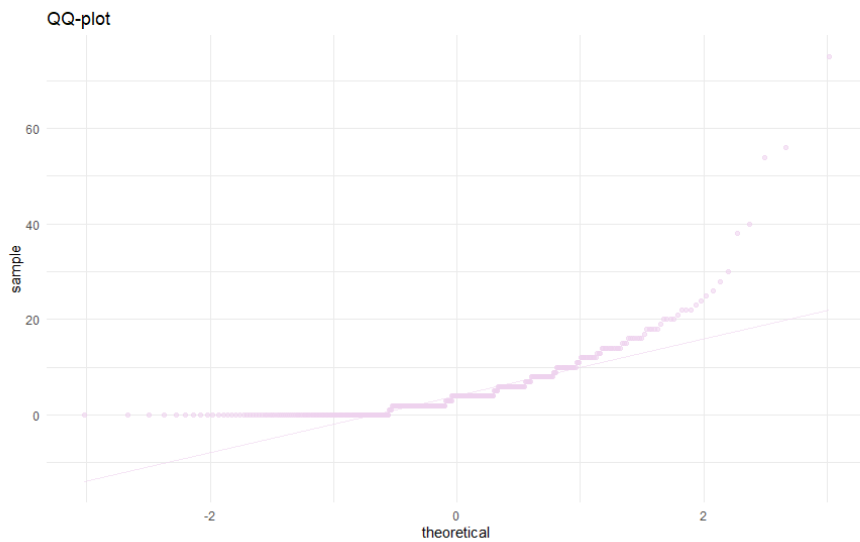


Figure 65: QQ-plot of absences

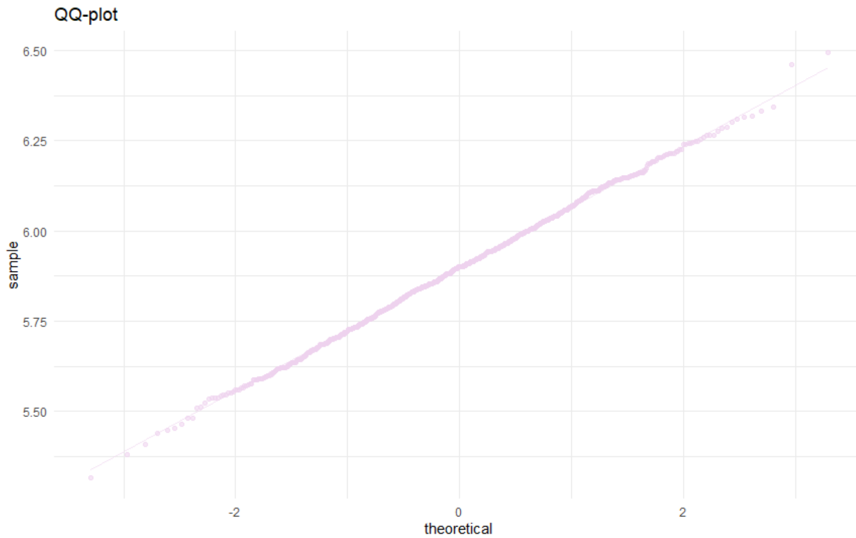


Figure 66: QQ-plot of mean of bootstrapped samples

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.

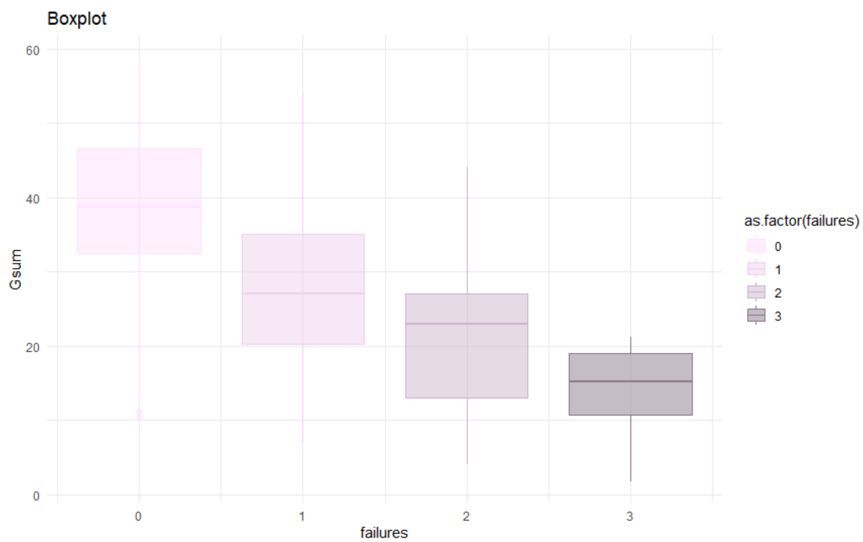


Figure 68: Boxplot grouped by number of failures

```
>
> 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.

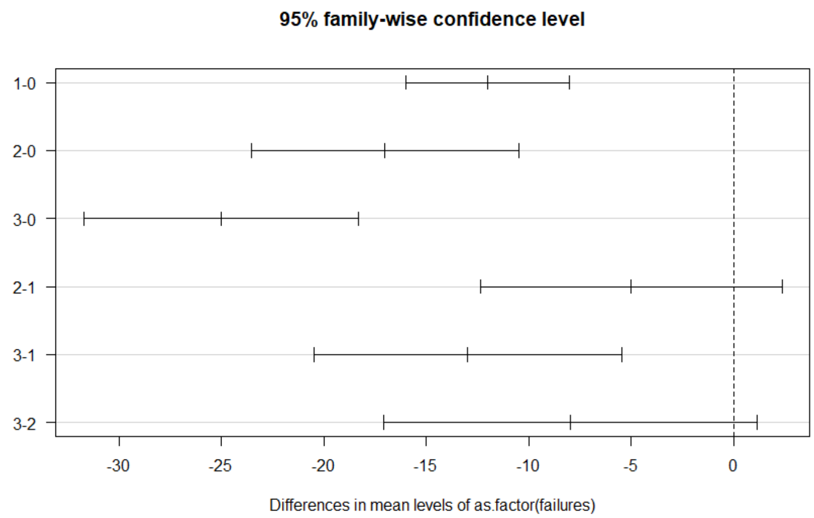


Figure 70: Pairwire confidence level(0.95%)

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')
```

```
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")
```

```
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
```

```r
#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

```

```
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 ")
```

```
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