

Statistical Inference

Course Project

Amin Asadi Sarijalu

810196410

Spring 2021

University of Tehran

ECE Department

# Question 0

## A. Briefly describe your dataset and why studying your dataset can be interesting?

Dataset: Student’s performance

This dataset includes some features including personal information and performance results of several students from two different schools. These features comprise several factors such as sex, age, and internet access. The performances consist of the scores of students in three different courses.

Studying and analyzing the information in in this database not only lets us discover the most influential factors on the students’ performances at school, but also enables us to compare the impact of the features such as sex, school, study time, etc. on the students’ performances. Moreover, we can study any two columns and see if there is an interesting statistical relationship between them. Also studying the statistical distribution of each column potentially reveals important information about the population.

## B. How many variables (features) and cases does your dataset have?

There are 15 variables and 395 cases in the dataset, 6 of which are categorical and 9 are numerical.

## C. Is there any missing value in your data? Provide a summary of a portion of missing values for each variable (feature) and describe how you handle these missing values for each variable (on what basis).

D. Using this elementary view of your dataset, which variables do you think maybe the most relevant (contain some important information)? Why?

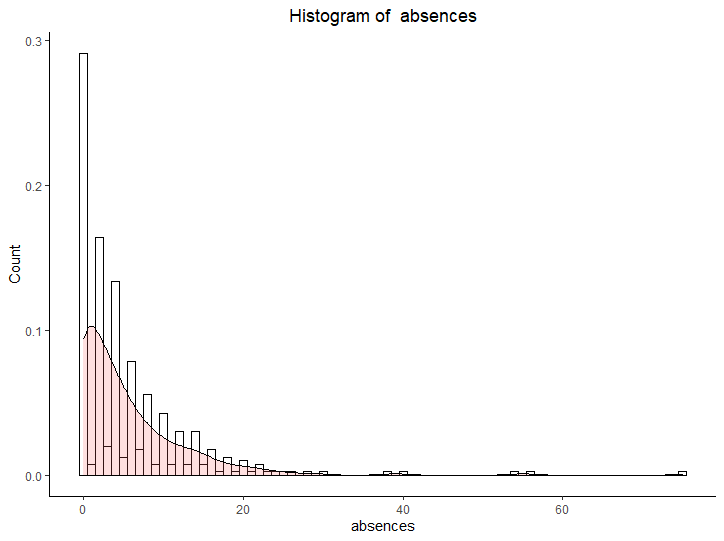
School is important as an explanatory variable, because it will show the influence of school at an student’s performance. Also personal characteristics like sex and age are important as well for the same reason. Students’ scores are very crucial as response variable because we want to find the impact of the characteristics on these performances.

# Question 1

Chosen Variable: Absences

## 1.A

Plot:



Description:

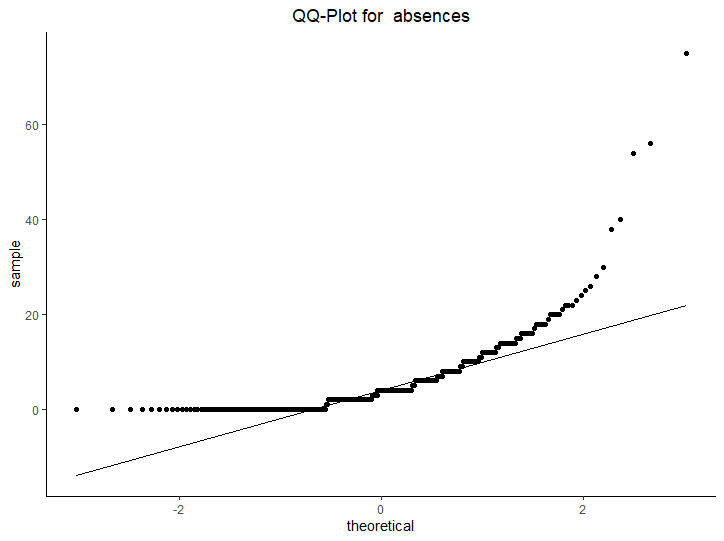
As we can see in the plot above, the is one maximum in absences. So the modality = 1

Mode is 0 which means that most of the students have never been absent.

## 1.B

We can see in the histogram that the distribution is highly right-skewed normal and has outliers in the right-most parts of it.

In order to more accurately compare it to a normal distribution we use Quantile-Quantile-Plot.



The result shows that the tails are upper than the normal qq-plot so this proves that the distribution is indeed right-skewed.

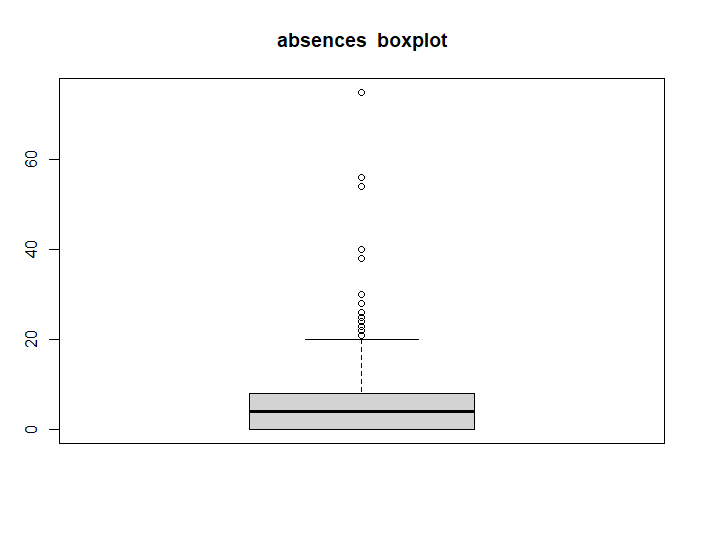
Result:

## 1.C

skewness of absences: 3.66

As we can see, the skewness of this variable is largely positive which shows that “mean” is larger than “median” which agrees with our previous result that the distribution of this variable is indeed highly right-skewed.

## 1.D

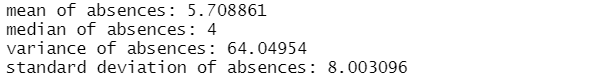




As we can see here, there are 15 outliers in absences which all are more than 20.

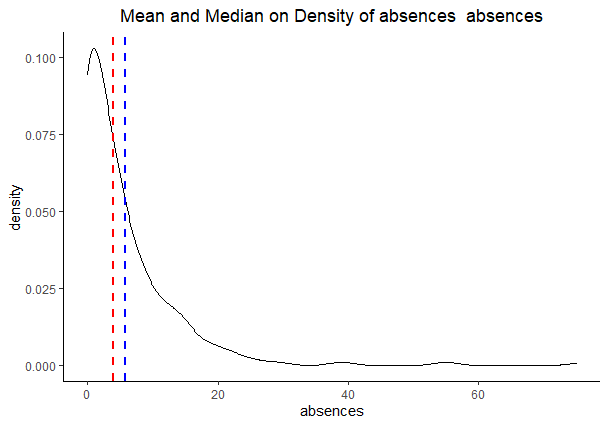
This may have many reasons. One possible reason is that these students in fact didn’t participate in most of their classes. Maybe attending in their classes were not mandatory for them or maybe they intended to take leave of absence for a semester.

## 1.E

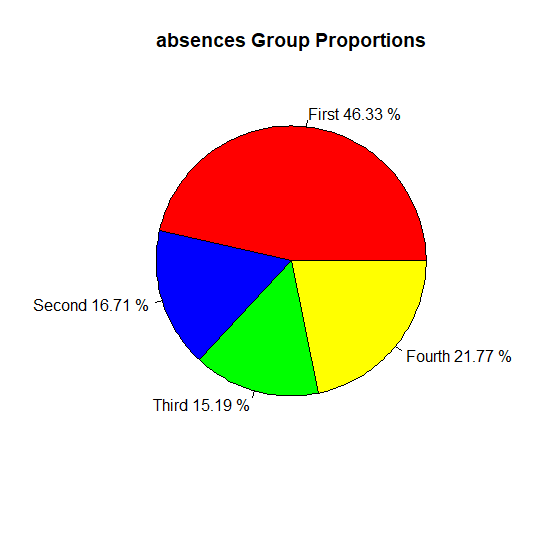


We can see that mean and median are not close. Mean is larger than median so the distribution is right-skewed.

## 1.F



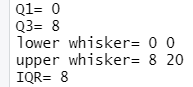
## 1.G



We can see that about half of the students have been absent less that ¼ of the means of the absences.

## 1.H

Box plot is drawn in part D



# Question 2

Chosen Variable: **School**

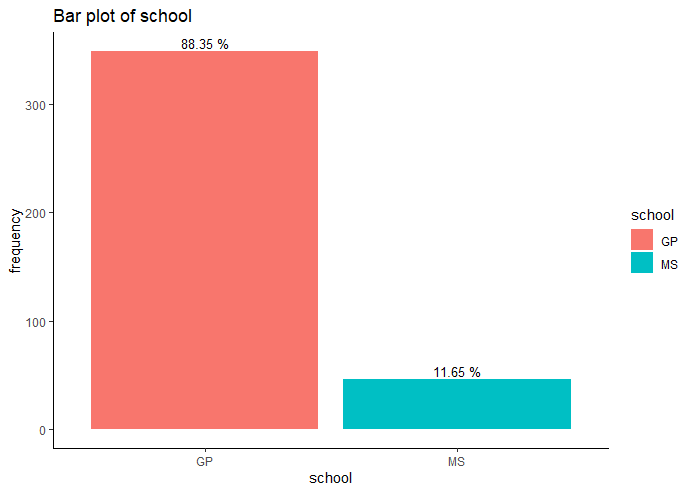
## 2.A



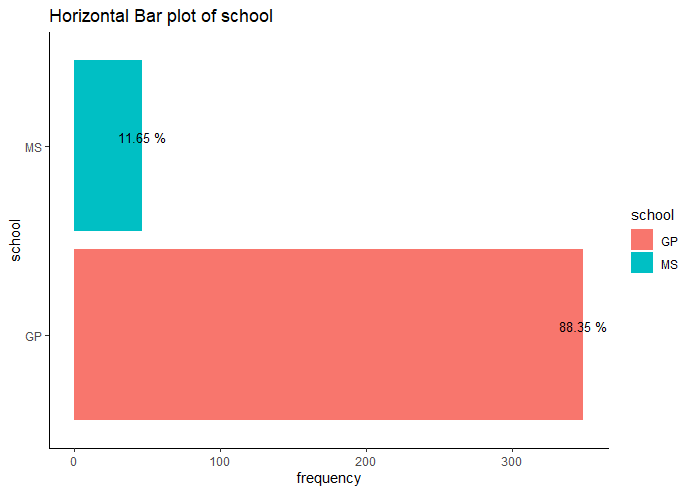


As we can see, most (88%) of students are from the “GP” school. This means that if we want to test a hypothesis that we think school is influential on the response variable, we should use stratified sampling, so that control group and treat group have same number of students from each of the schools.

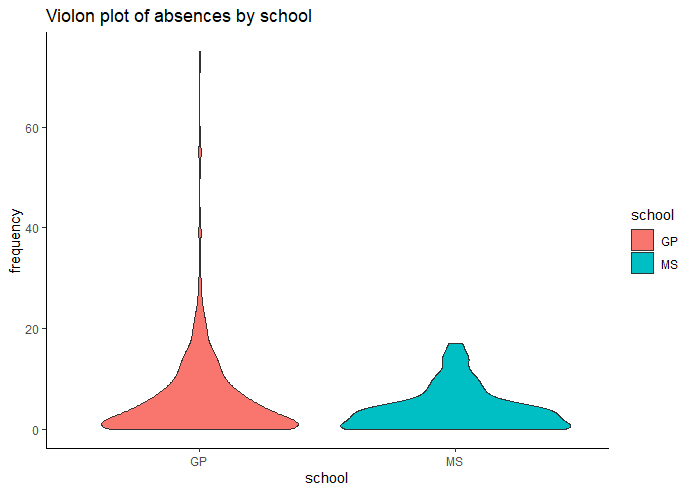
## 2.B



## 2.C



## 2.D



### Violin Plot is the combination of density plot and box plot, i.e. it both demonstrates which parts of the data have more or less probability to occur and also it shows roughly what the box plot shows.

As we can see here, the violin plot for both schools is very wide at the bottom because, as we discussed in the previous question, most of the students have never been absent(0 absences).

Also a noteworthy difference between the violin plot for these two schools is that in “MS” school there are no outliers, maybe because the students are not allowed to have more that a maximum number of absences are maybe they can not take leave of absence so they can’t decide not to attend classes after starting a semester.

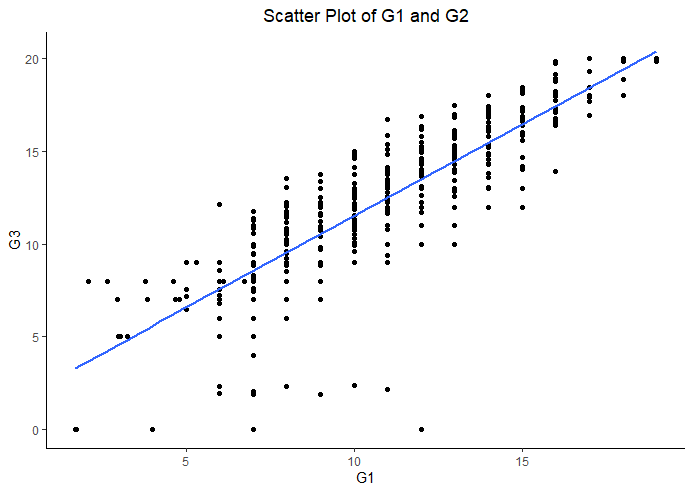
# Question 3

## Chosen Variables : G1 and G2

## 3.A

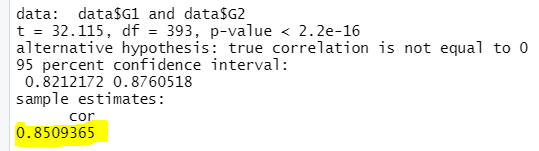
I guess that these two variables should have a high correlation because if a student has a higher G1 score, so in general they are intended to be [studious](https://abadis.ir/entofa/s/studious/), therefore they probably have high score in G2 too.

## 3.B



We see here that apparently there is a strong linear correlation with these two variables.

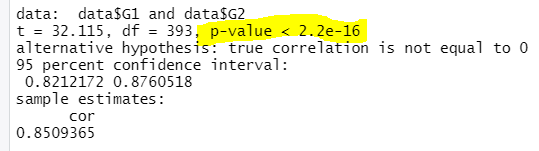
## 3.C



## 3.D

The correlation of these two variables is 0.85 which is very high and agrees with my guess at part A.

## 3.E



Null Hypothesis: True correlation is equal to 0

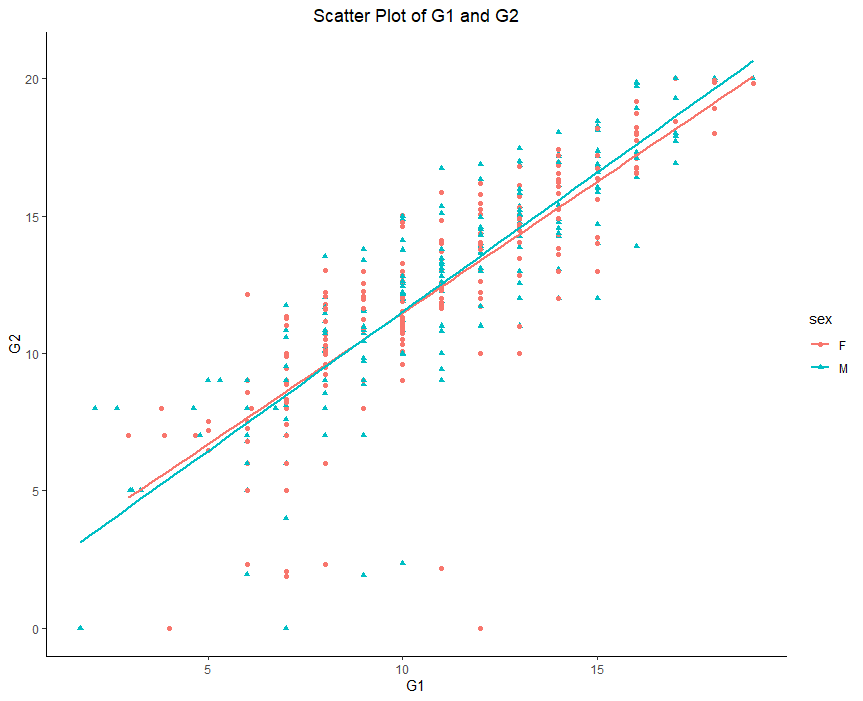
Alternative Hypothesis: True correlation is not equal to 0

p-value is nearly zero so we reject the null hypothesis in favor of the alternative.

Result: we are highly confident that there is a correlation between these two variables as we guessed in par A.

## 3.F

Selected Categorial Variable : Sex

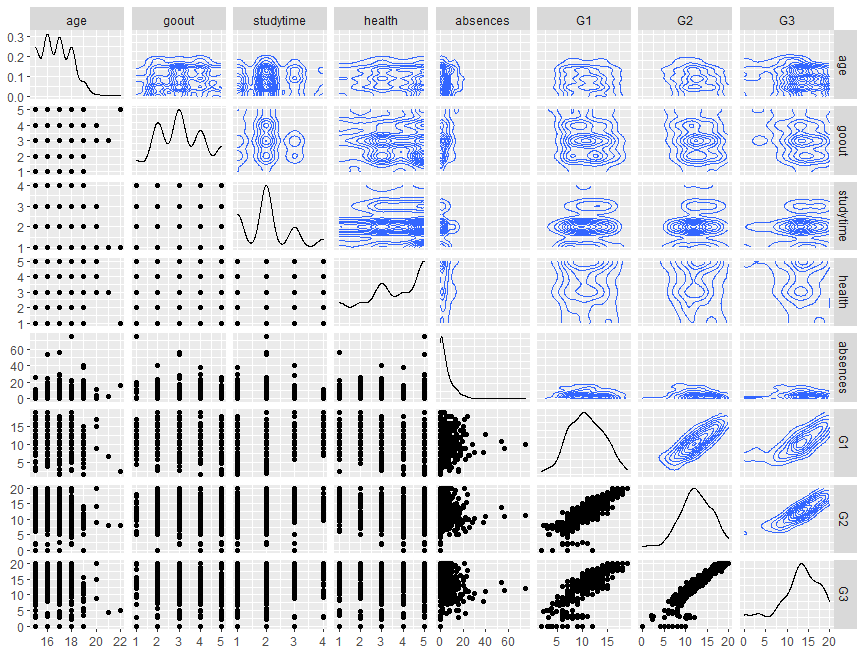


## 3.G

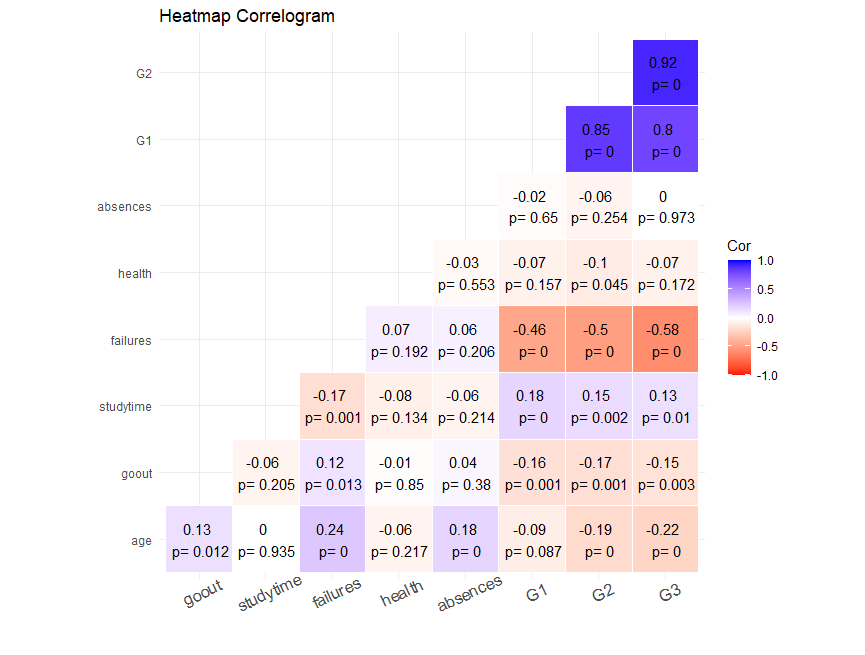
### 3.H

# Question 4

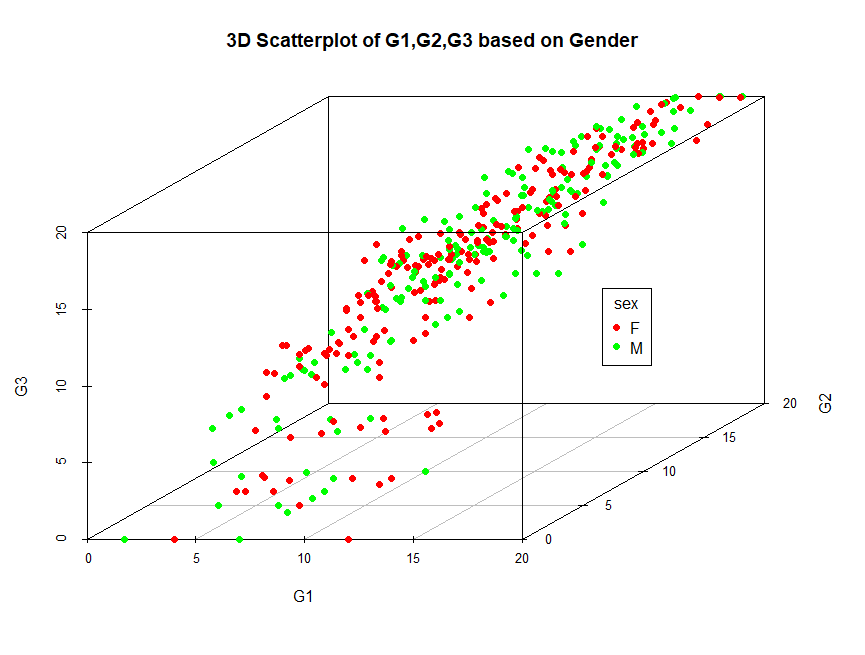
## 4.A



## 4.B



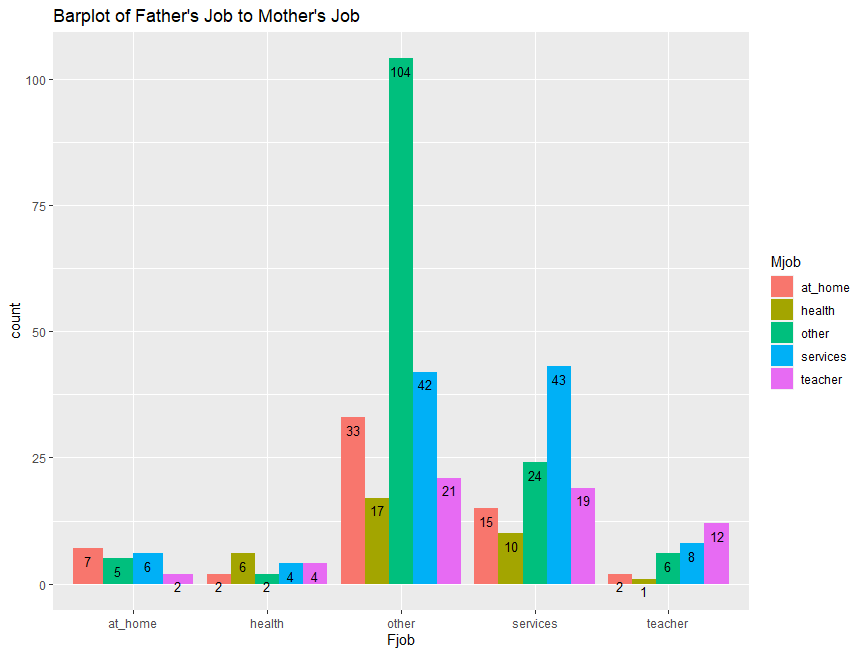
## 4.C



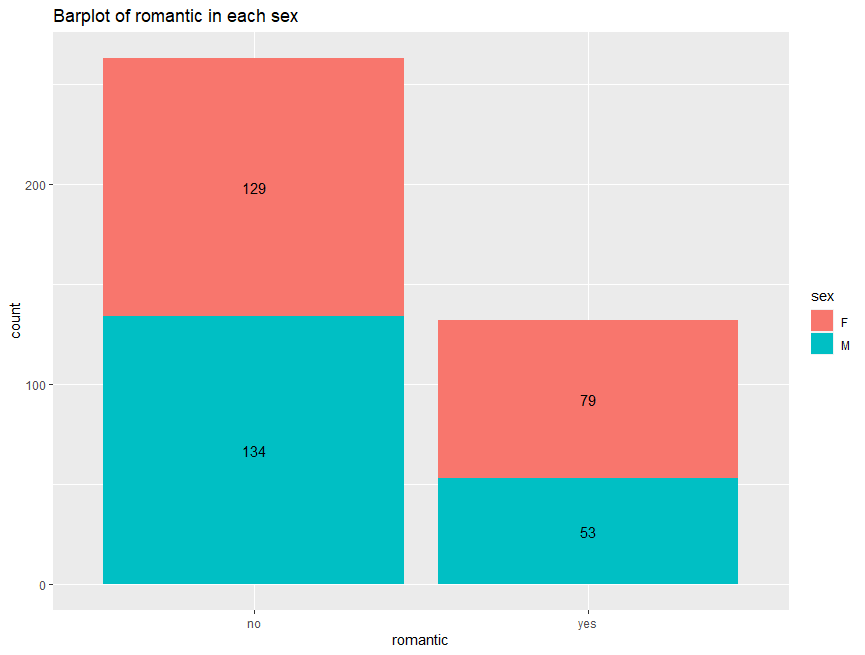
## 5.A

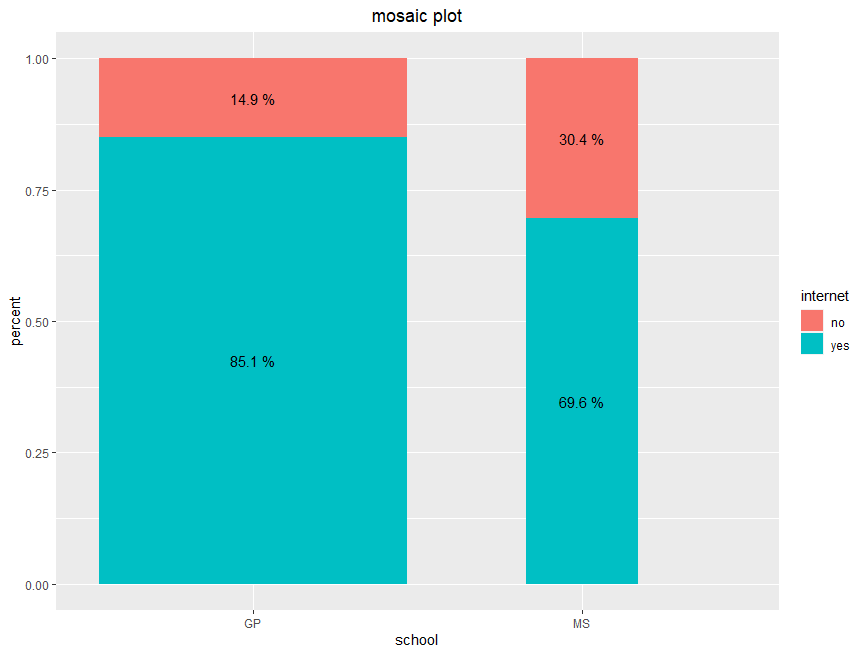
|  |  |  |  |
| --- | --- | --- | --- |
|  | Female | Male | Total |
| GP School | 183 | 166 | 349 |
| MS School | 25 | 21 | 46 |

## 5.B



## 5.C





# Question 6

## 6.A