Statistical Inference Course Project

Amin Asadi Sarijalu

810196410

Spring 2021

University of Tehran

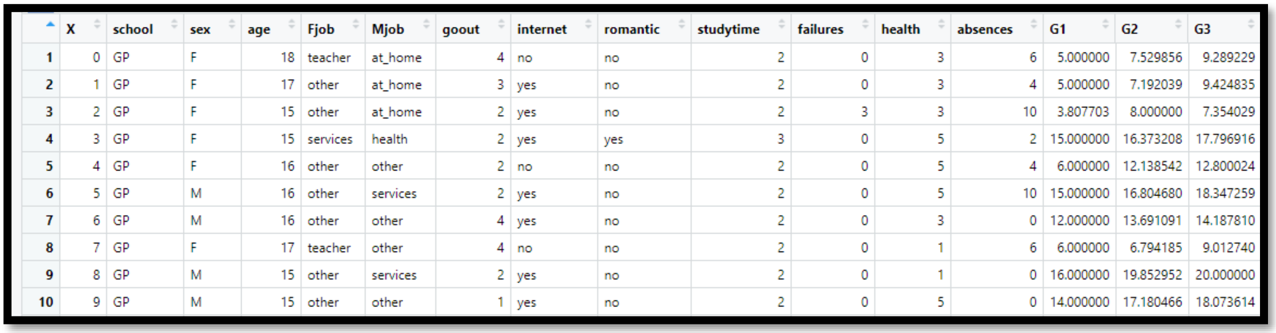
ECE Department

Contents

Question 0

0.A

Dataset: Student's performance



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

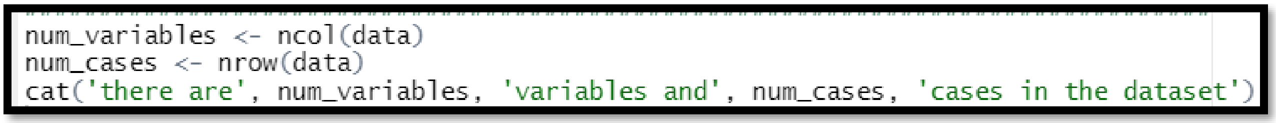
Figure 1 First 10 rows of the dataset

This dataset includes some features including personal information and education-related performances of several students from two different schools. The personal information consists of sex, age, internet access, romantic, etc. The educational information consists of how many times they have been absent, their grades and etc.

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 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.

0.B

Code:

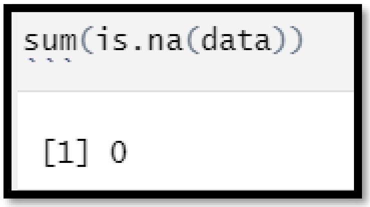


```
num_variables <- ncol(data)
num_cases <- nrow(data)
cat('there are', num_variables, 'variables and', num_cases, 'cases in the dataset')
```

There are 15 variables and 395 cases in the dataset

6 variables are categorical and 9 of them are numerical.

0.C



~~~~
sum(is.na(data))
```

 [1] 0
~~~~

As we can see, there are not missing values in the dataset. Also I searched for special characters which might indicated missing values such as ‘-‘ or ‘NA’ or ‘NAN’ but none existed. If existed we could handle them by methods such as: deletion, replacing with mean, median or mod and etc.

0.D

As an explanatory variable, “School” is important since it will reveal the influence of school at a student’s performance. Besides, personal characteristics like sex, age, romantic are important as well for the same reason and that we can make conclusion about the consequences and impact of these variables on the students’ educational performances.

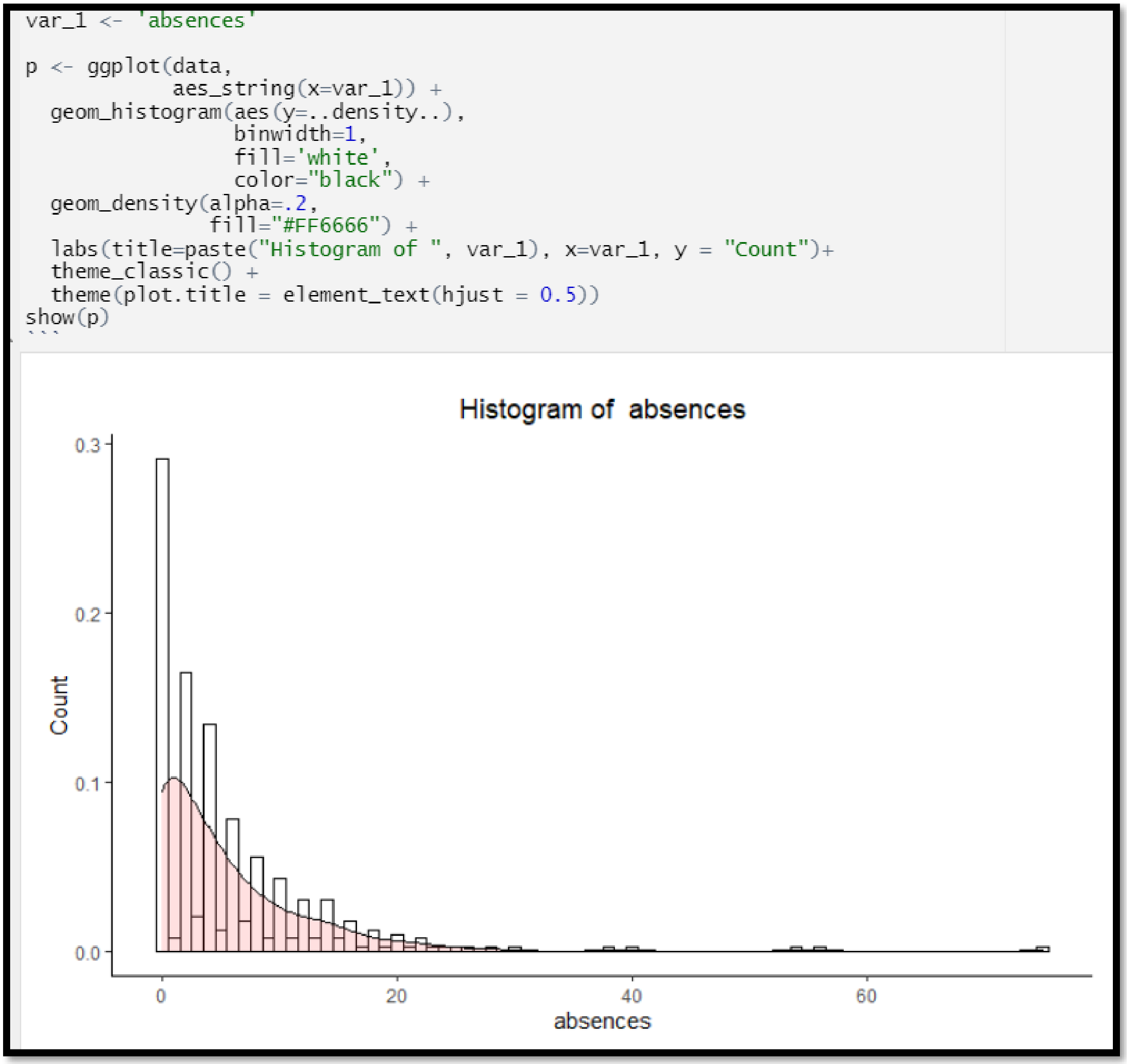
Students’ scores are very crucial as response variable because we want to find the impact of the characteristics on these performances.

Other variables like parent’s jobs or internet access can be important (for example in one of the sections we will see whether each student’s parents are intended to have the same job or not) but not as important as the ones mentioned above.

Question 1

Chosen Variable: Absences

1.A



```
var_1 <- 'absences'

p <- ggplot(data,
            aes_string(x=var_1)) +
  geom_histogram(aes(y=..density..),
                 binwidth=1,
                 fill='white',
                 color="black") +
  geom_density(alpha=.2,
               fill="#FF6666") +
  labs(title=paste("Histogram of ", var_1), x=var_1, y = "Count")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
show(p)
```

Description:

We use bin size = 1 because absences is a discrete variable and natural number.

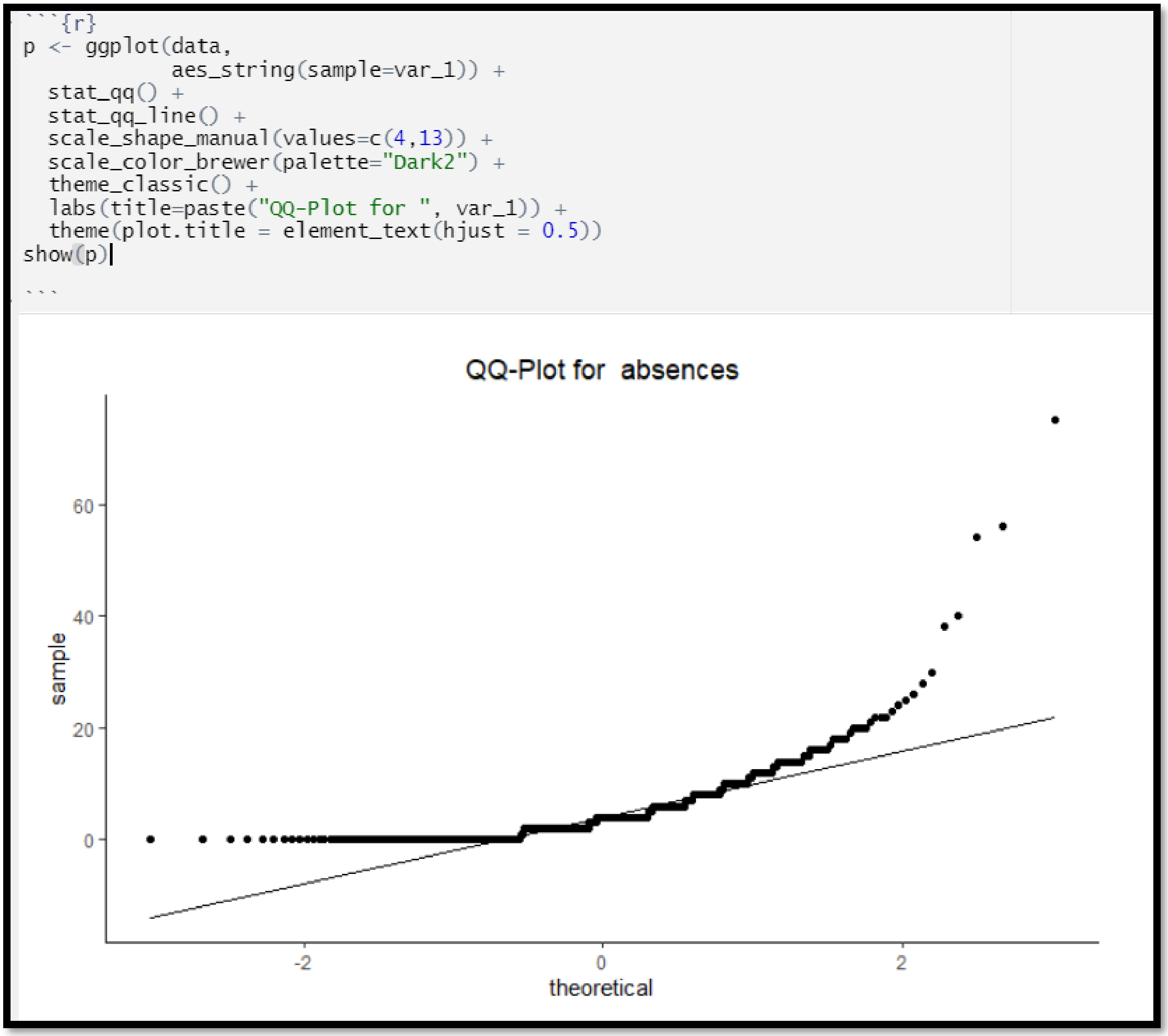
As we can see in the plot above, the is one maximum in absences. So, the modality = 1

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.



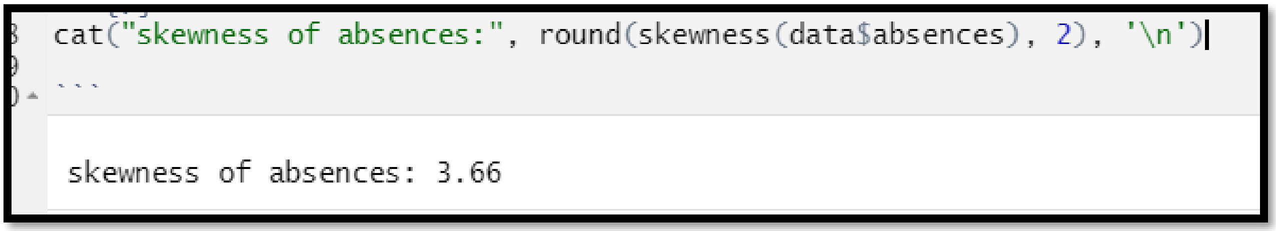
~~~
```{r}
p <- ggplot(data,
            aes_string(sample=var_1)) +
  stat_qq() +
  stat_qq_line() +
  scale_shape_manual(values=c(4,13)) +
  scale_color_brewer(palette="Dark2") +
  theme_classic() +
  labs(title=paste("QQ-Plot for ", var_1)) +
  theme(plot.title = element_text(hjust = 0.5))
show(p)

```
~~~

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

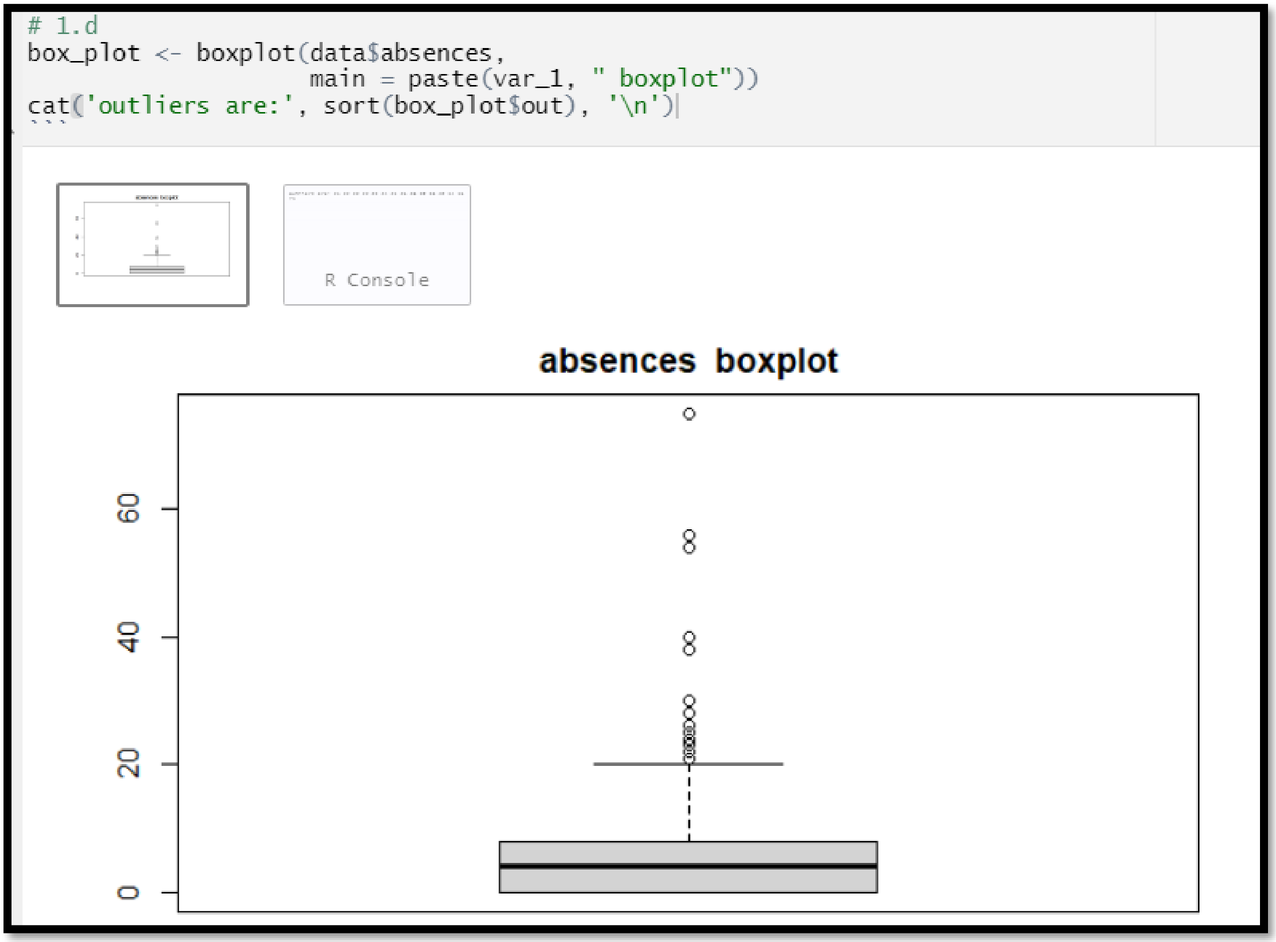


```
cat("skewness of absences:", round(skewness(data$absences), 2), '\n')
```
```

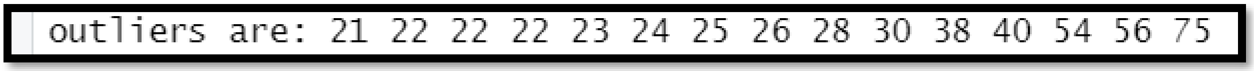
skewness of absences: 3.66
```

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



```
1.d
box_plot <- boxplot(data$absences,
                    main = paste(var_1, " boxplot"))
cat('outliers are:', sort(box_plot$out), '\n')
```
```
```

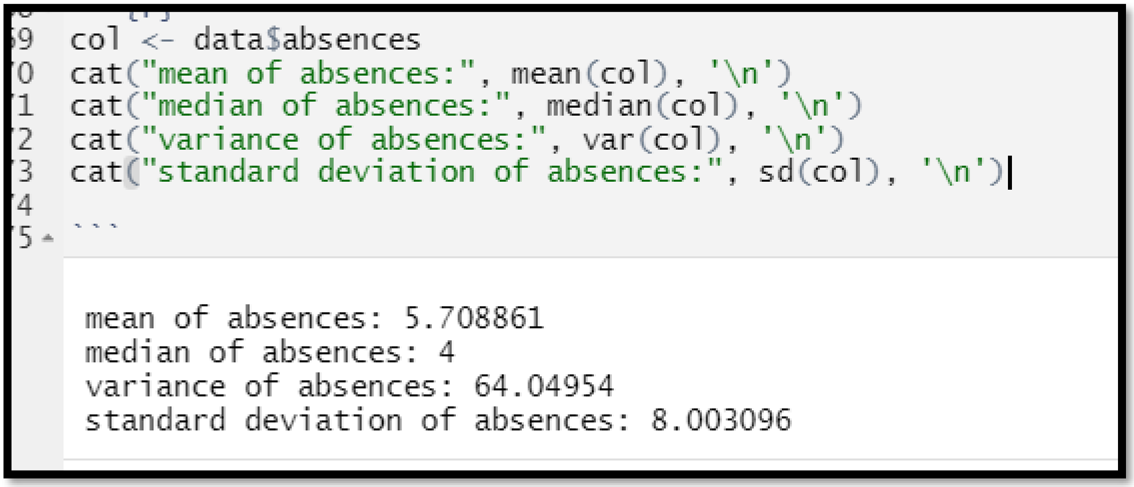


```
outliers are: 21 22 22 22 23 24 25 26 28 30 38 40 54 56 75
```

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

This may have several 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 decided to take leave of absence for a semester and therefore didn’t participate in the classes thereafter. Also, these numbers can also be human error, i.e., the person who was responsible for entering the sum of absences may have entered them incorrectly.

1.E



~~~~
col <- data$absences
cat("mean of absences:", mean(col), '\n')
cat("median of absences:", median(col), '\n')
cat("variance of absences:", var(col), '\n')
cat("standard deviation of absences:", sd(col), '\n')

```
~~~~

```
mean of absences: 5.708861
median of absences: 4
variance of absences: 64.04954
standard deviation of absences: 8.003096
```

Mean is the sum of absences divided by size of the population. The problem with mean in skewed distributions is that mean is sensitive to outliers (like the distribution of incomes). The median is the point which half of the datapoints are smaller than it and half of them are greater than it.

We can see that although most of the students have never been absent, the mean is 5.7 which can be a misinformation.

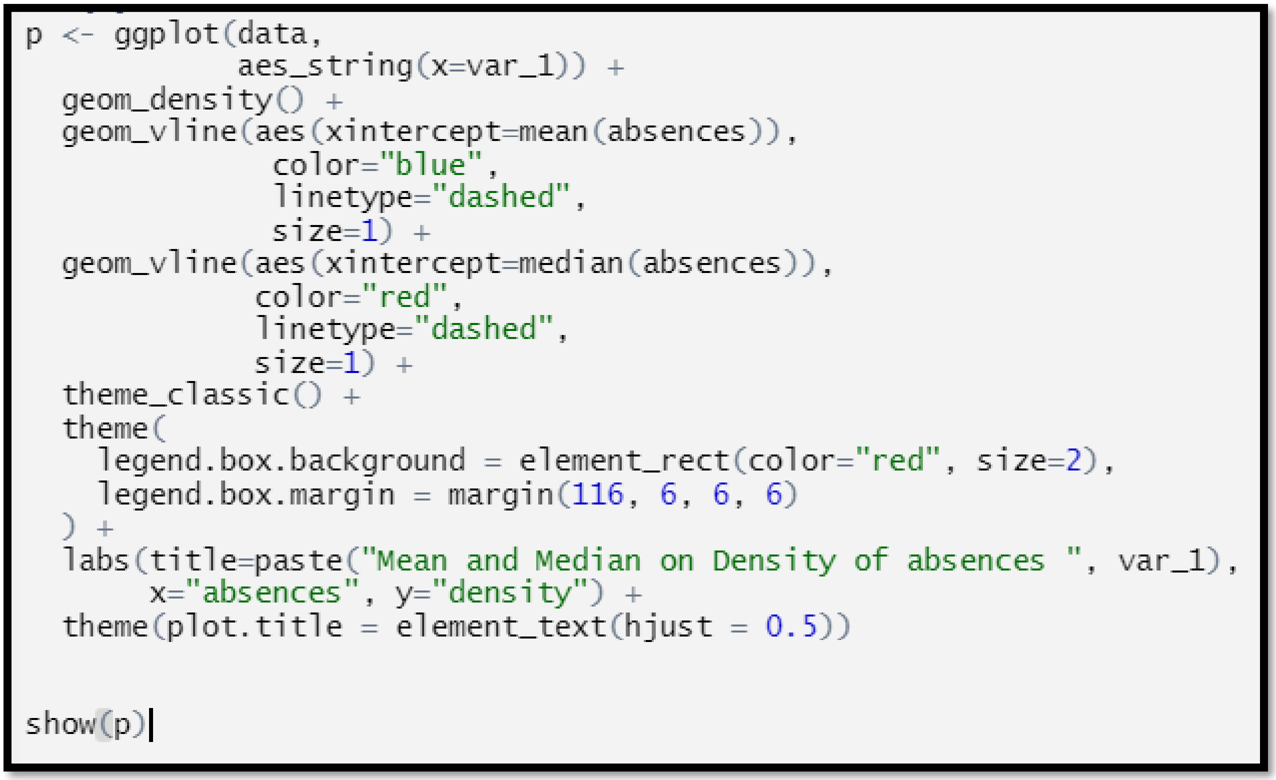
We can see that mean and median are not close. Mean is larger than median so the distribution is right-skewed.

Standard deviation and variance show spread of data. The difference between them is that the unit of variance is square of the unit of data but standard deviation has the same unit as data.

Std is 8 means the root square of the sum of differences from mean is 8.

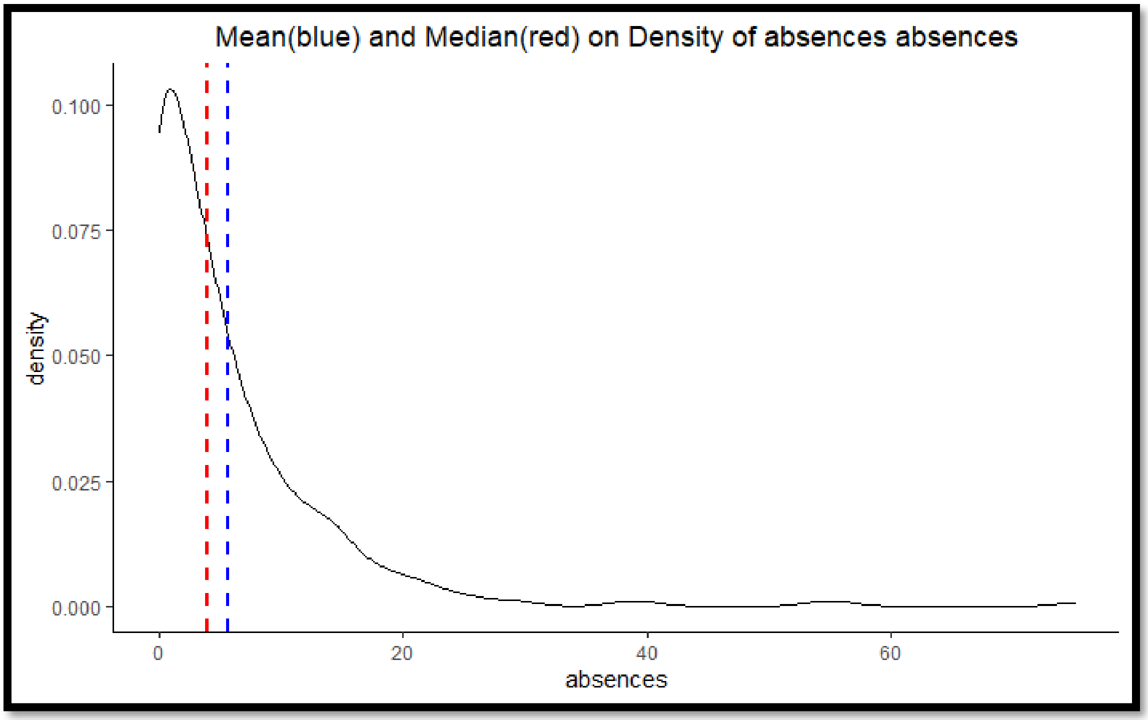
For a symmetric normal distribution It also means that about 68% of data are within the 8 units of distance of the mean, however we can not say it here because the distribution is not symmetric but skewed.

1.F



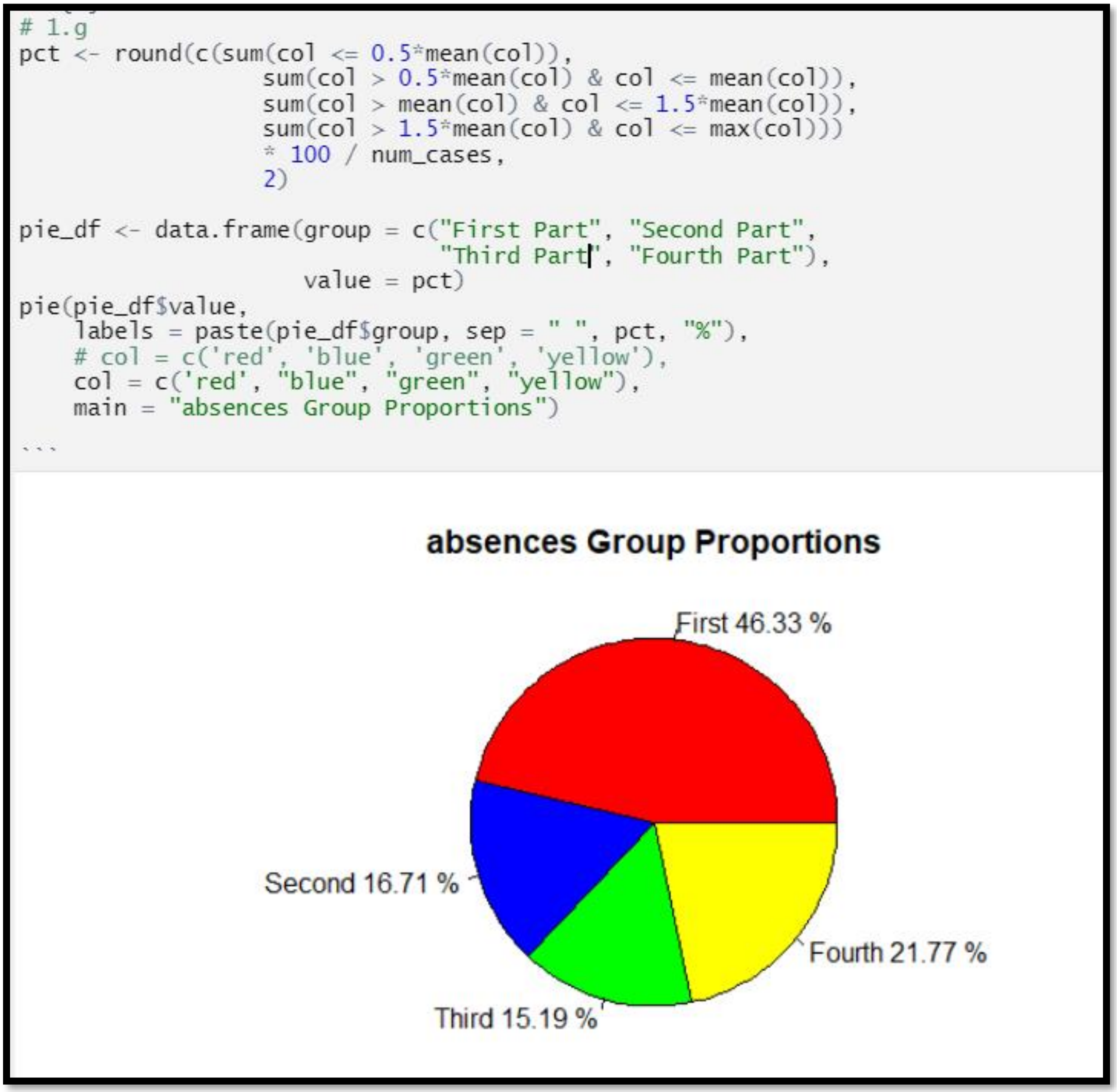
```
p <- ggplot(data,
            aes_string(x=var_1)) +
  geom_density() +
  geom_vline(aes(xintercept=mean(absences)),
             color="blue",
             linetype="dashed",
             size=1) +
  geom_vline(aes(xintercept=median(absences)),
            color="red",
            linetype="dashed",
            size=1) +
  theme_classic() +
  theme(
    legend.box.background = element_rect(color="red", size=2),
    legend.box.margin = margin(116, 6, 6, 6)
  ) +
  labs(title=paste("Mean and Median on Density of absences ", var_1),
       x="absences", y="density") +
  theme(plot.title = element_text(hjust = 0.5))


show(p)
```



As we stated in previous parts, mean(blue) is greater than the median(red).

1.G



```
# 1.g
pct <- round(c(sum(col <= 0.5*mean(col)),
                  sum(col > 0.5*mean(col) & col <= mean(col)),
                  sum(col > mean(col) & col <= 1.5*mean(col)),
                  sum(col > 1.5*mean(col) & col <= max(col)))
                  * 100 / num_cases,
                  2)

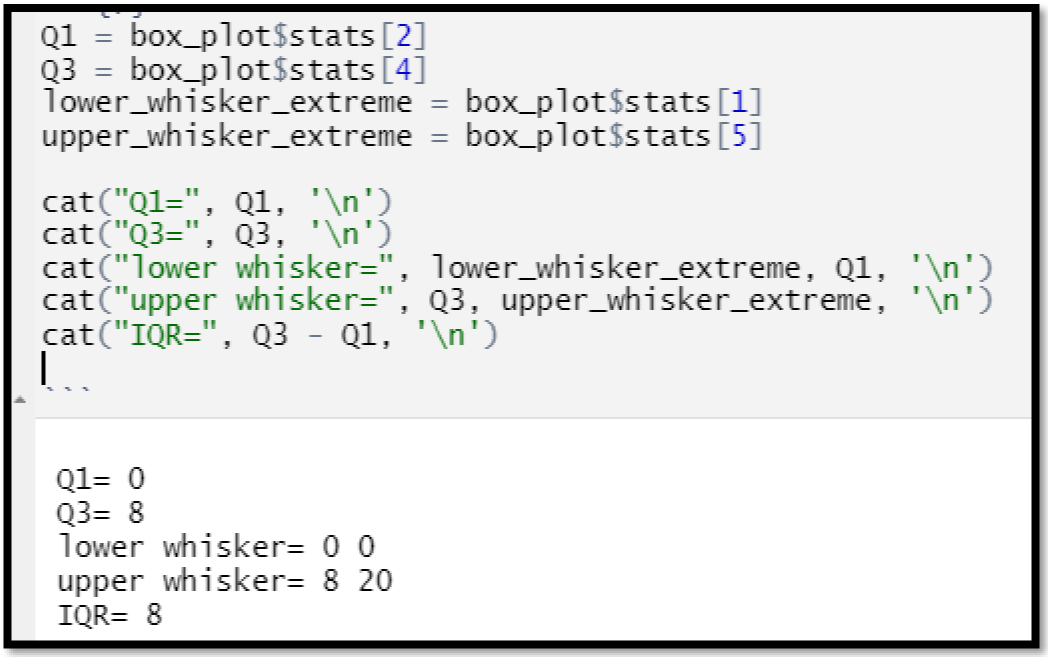
pie_df <- data.frame(group = c("First Part", "Second Part",
                               "Third Part", "Fourth Part"),
                     value = pct)
pie(pie_df$value,
    labels = paste(pie_df$group, sep = " ", pct, "%"),
    # col = c('red', 'blue', 'green', 'yellow'),
    col = c('red', "blue", "green", "yellow"),
    main = "absences Group Proportions")

```
```

The red section of the pie chart is for the first quarter of the number of absences. So, we conclude that about half (46.33 %) of the students have been absent at most as ½ of the mean. The rest of the students have equally distributed into the rest of the quarters (blue, green and yellow).
```

1.H

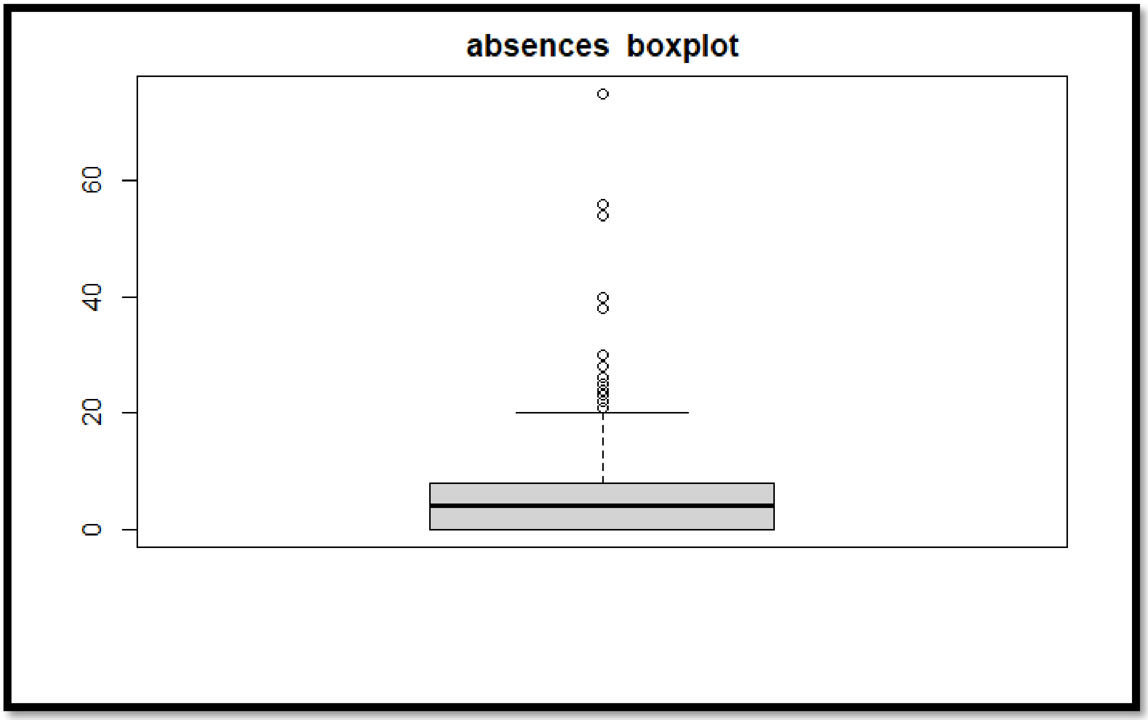
Box plot is drawn in part D



```
Q1 = box_plot$stats[2]
Q3 = box_plot$stats[4]
lower_whisker_extreme = box_plot$stats[1]
upper_whisker_extreme = box_plot$stats[5]

cat("Q1=", Q1, '\n')
cat("Q3=", Q3, '\n')
cat("lower whisker=", lower_whisker_extreme, Q1, '\n')
cat("upper whisker=", Q3, upper_whisker_extreme, '\n')
cat("IQR=", Q3 - Q1, '\n')
```

```
Q1= 0
Q3= 8
lower whisker= 0 0
upper whisker= 8 20
IQR= 8
```

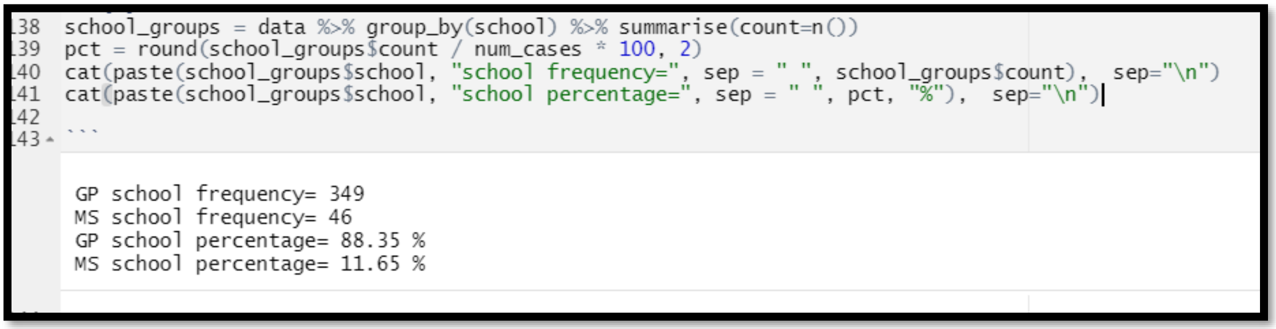


We can see that first Quartile is equal to the minimum. It means that at least 25% of the students have never been absent. So, the lower whisker is of length 0. Also, upper whisker is between 8 and 20 (20 is not max of absents) which means 1.5 * std + Q3 is 20 and the rest of upper part are outliers. IQR is 8 which means half of the students have been absent between 0 (Q1) and 8 (Q3) times.

Question 2

Chosen Variable: **School**

2.A



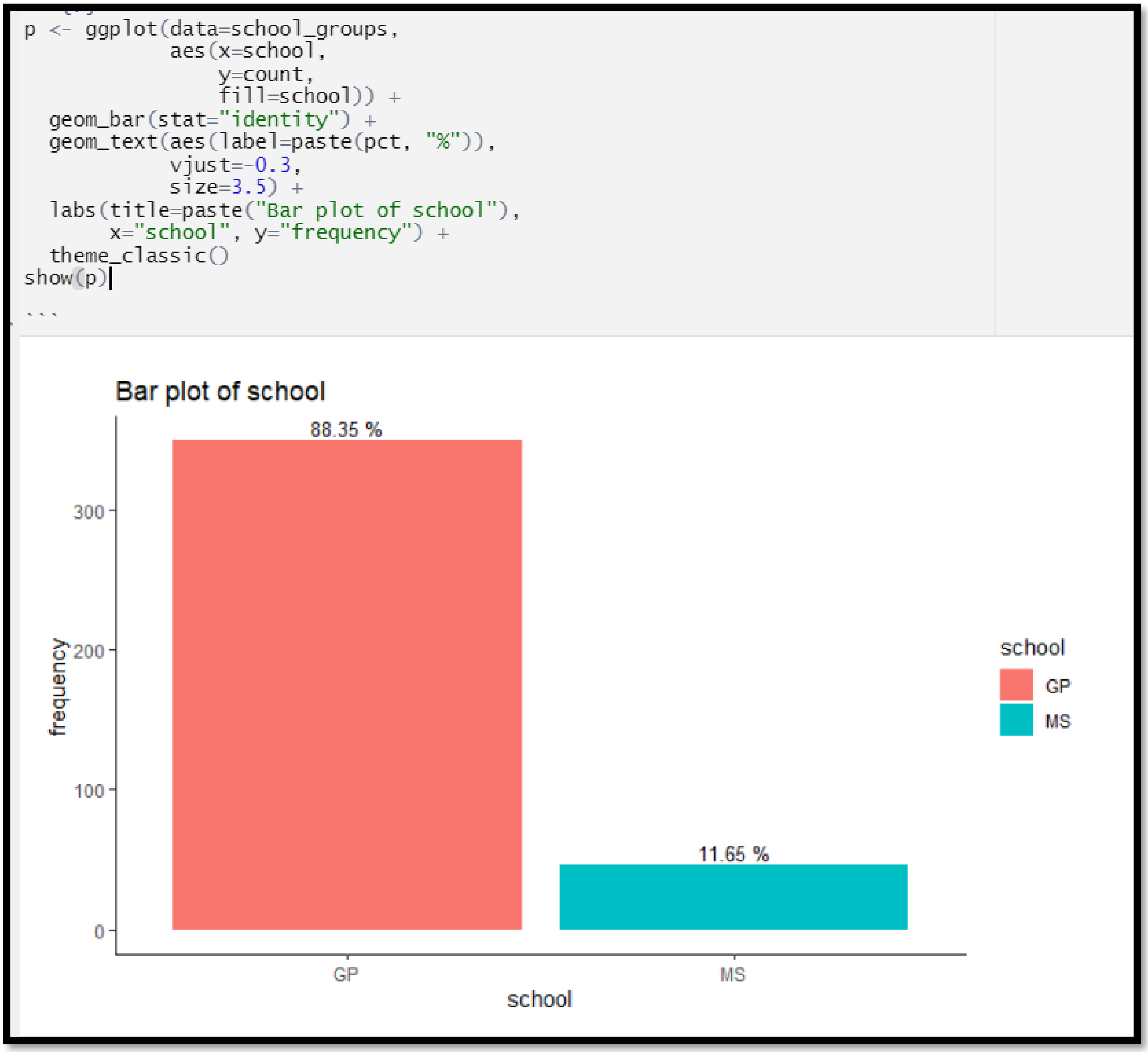
```
school_groups = data %>% group_by(school) %>% summarise(count=n())
pct = round(school_groups$count / num_cases * 100, 2)
cat(paste(school_groups$school, "school frequency=", sep = " ", school_groups$count),  sep="\n")
cat(paste(school_groups$school, "school percentage=", sep = " ", pct, "%"),  sep="\n")

```
```

```
GP school frequency= 349
MS school frequency= 46
GP school percentage= 88.35 %
MS school percentage= 11.65 %
```

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 to avoid potential bias from occurring.
```

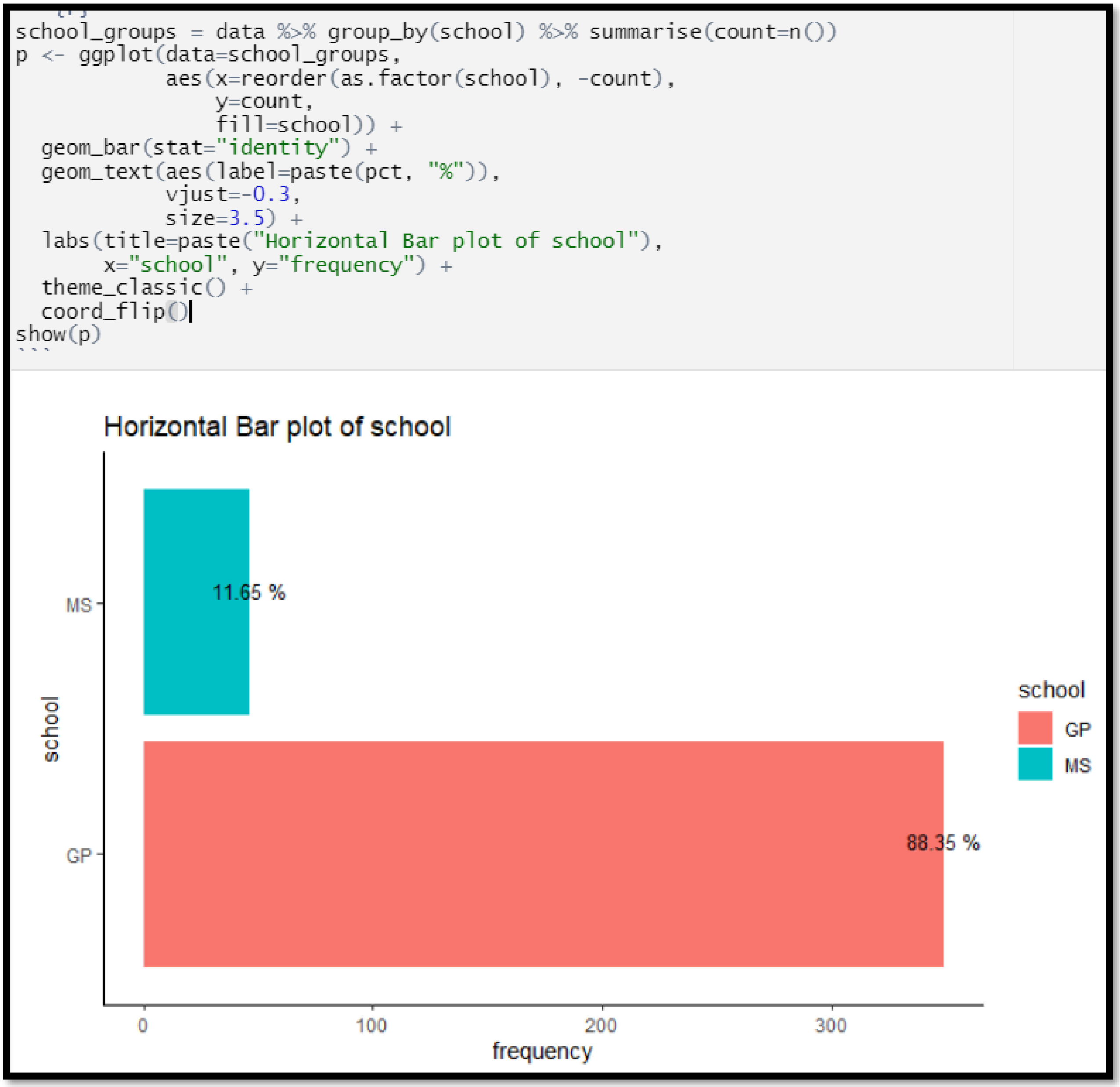
2.B



```
p <- ggplot(data=school_groups,
            aes(x=school,
                y=count,
                fill=school)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=paste(pct, "%")),
            vjust=-0.3,
            size=3.5) +
  labs(title=paste("Bar plot of school"),
       x="school", y="frequency") +
  theme_classic()
show(p)
```
```

The results are exactly the same as previous part.
```

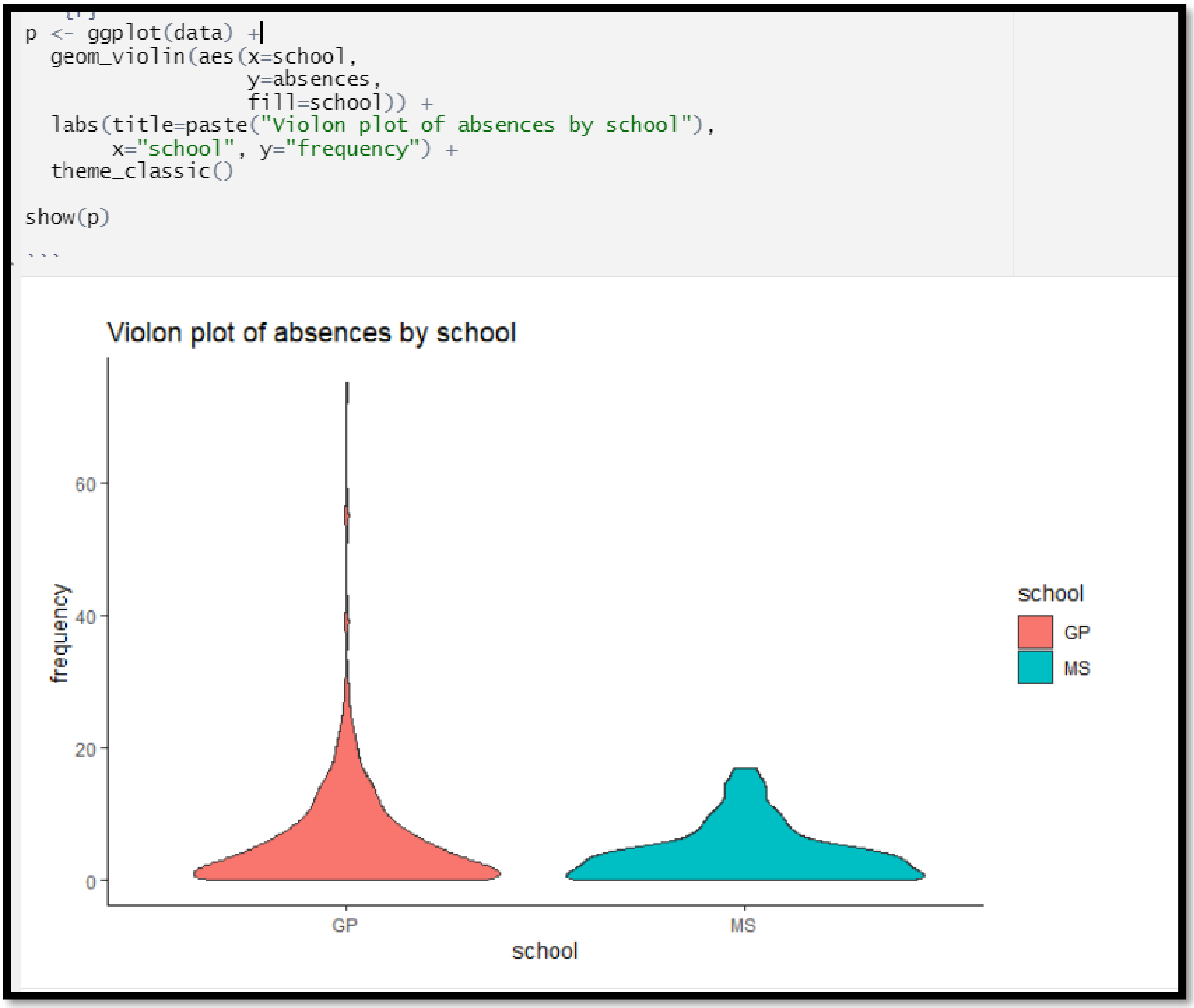
2.C



```
school_groups = data %>% group_by(school) %>% summarise(count=n())
p <- ggplot(data=school_groups,
            aes(x=reorder(as.factor(school), -count),
                y=count,
                fill=school)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=paste(pct, "%")),
            vjust=-0.3,
            size=3.5) +
  labs(title=paste("Horizontal Bar plot of school"),
       x="school", y="frequency") +
  theme_classic() +
  coord_flip()
show(p)
```

The results are exactly the same as previous part.

2.D



~~~
p <- ggplot(data) +
  geom_violin(aes(x=school,
                  y=absences,
                  fill=school)) +
  labs(title=paste("Violon plot of absences by school"),
       x="school", y="frequency") +
  theme_classic()

show(p)

```
~~~

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 demonstrates what the box plot shows such as what parts of the distribution can be considered as outliers.

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.

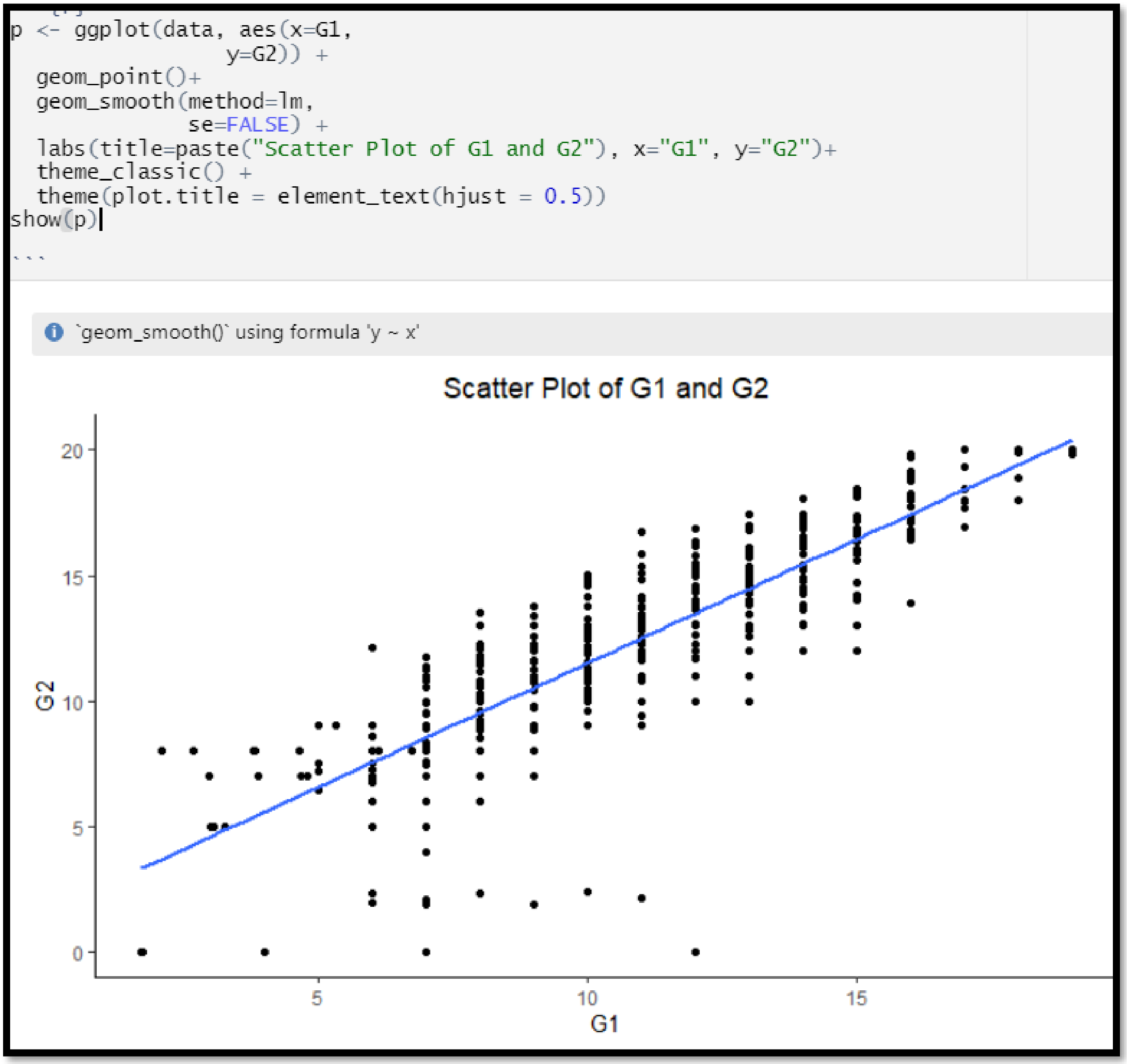
Question 3

Chosen Variables: G1 and G2

3.A

I guess G1 and G2 should have a high correlation because if a student has a higher G1 score, so in general they are intended to be studious, therefore they probably have high score in G2 too.

3.B



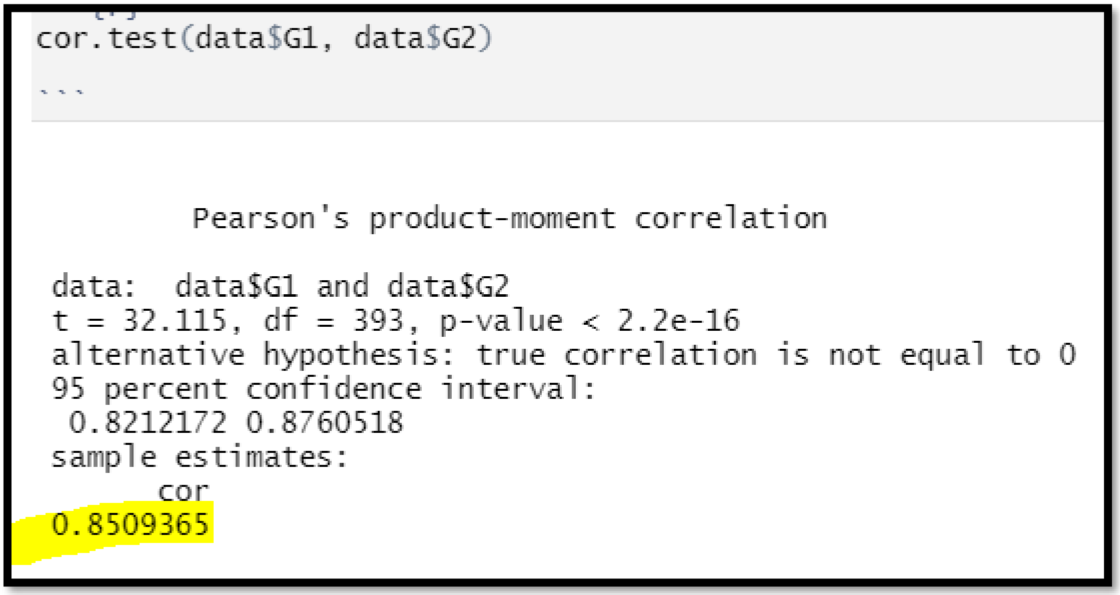
```
p <- ggplot(data, aes(x=G1,
                      y=G2)) +
  geom_point()+
  geom_smooth(method=lm,
              se=FALSE) +
  labs(title=paste("Scatter Plot of G1 and G2"), x="G1", y="G2")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
show(p)

```
```

`geom_smooth()` using formula 'y ~ x'
```

We see here that apparently there is a strong linear correlation with these two variables because as G1 increases, G2 increases too.

3.C



```
cor.test(data$G1, data$G2)

```

```

        Pearson's product-moment correlation

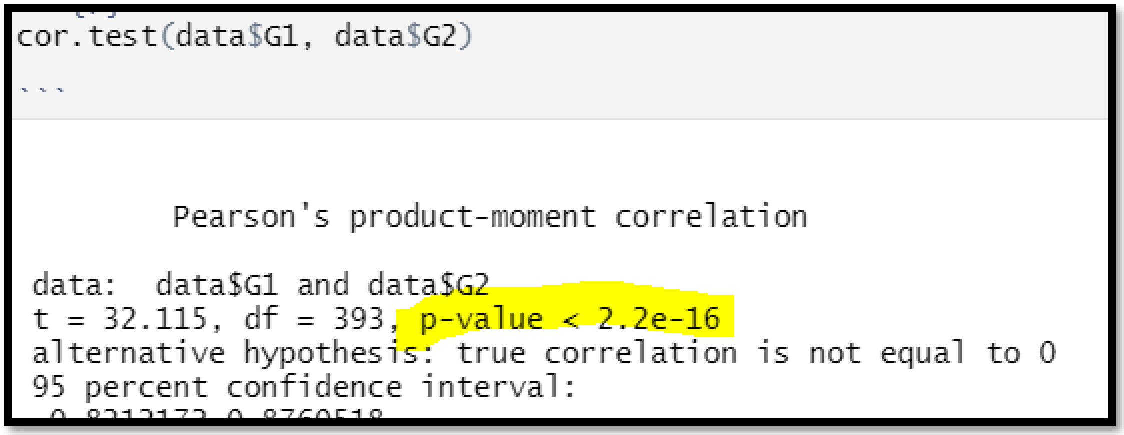
 data: data$G1 and data$G2
 t = 32.115, df = 393, p-value < 2.2e-16
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  0.8212172 0.8760518
 sample estimates:
       cor
 0.8509365
```

3.D

The correlation of these two variables is 0.85 which is very high (maximum correlation is 1) and hence agrees with my guess at part A.

3.E

The purpose of hypothesis test is to determine whether the linear relationship in the data is sufficiently strong to use to model the relationship in the population.



~~~
cor.test(data$G1, data$G2)

```


        Pearson's product-moment correlation

 data: data$G1 and data$G2
 t = 32.115, df = 393, p-value < 2.2e-16
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  0.8313173 0.8760518
~~~

Test:

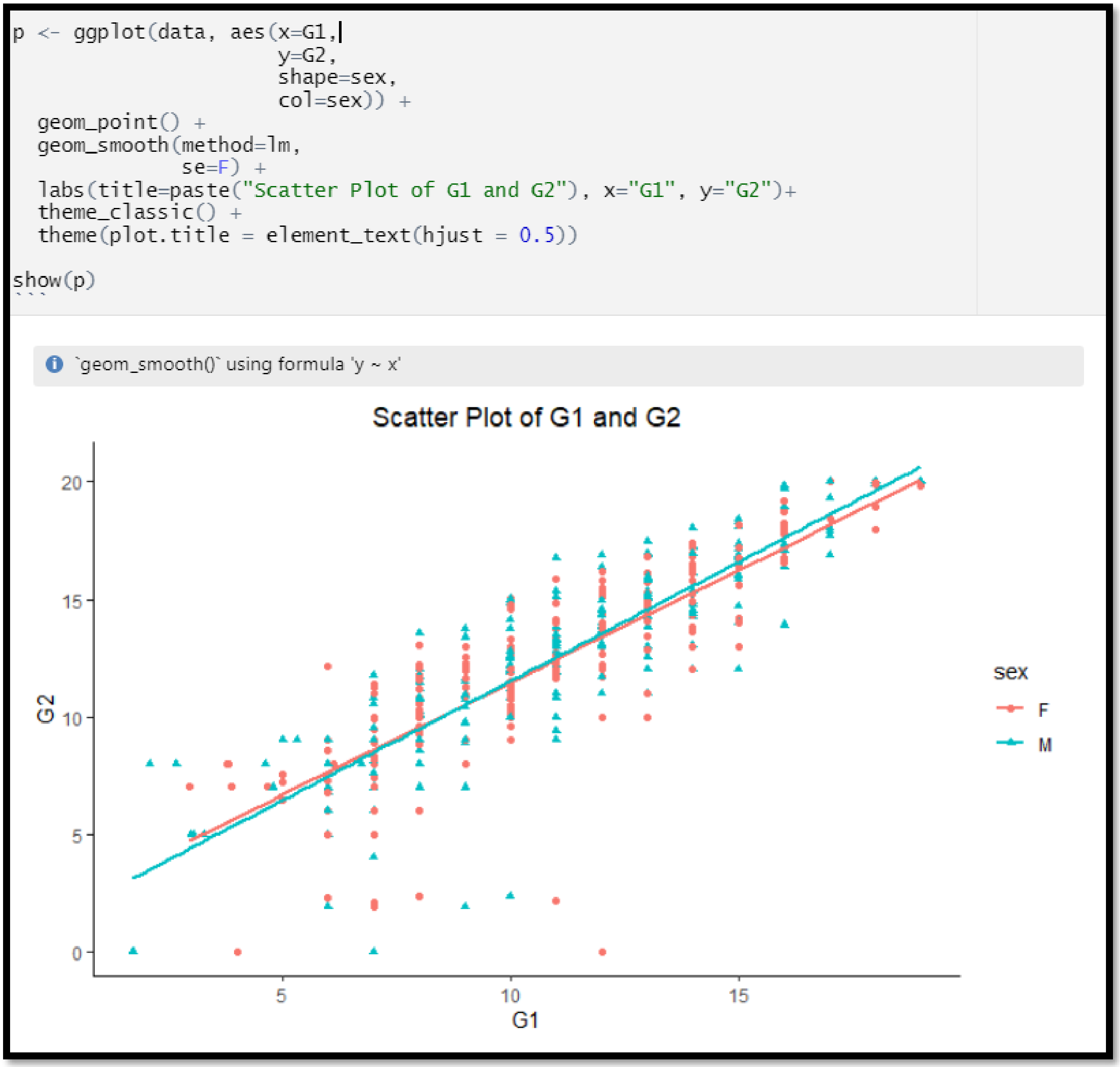
Null Hypothesis:	True correlation is equal to 0
Alternative Hypothesis:	True correlation is not equal to 0

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 strong highly correlation between these two variables as we guessed in part A.

3.F

Selected Categorial Variable is: Sex



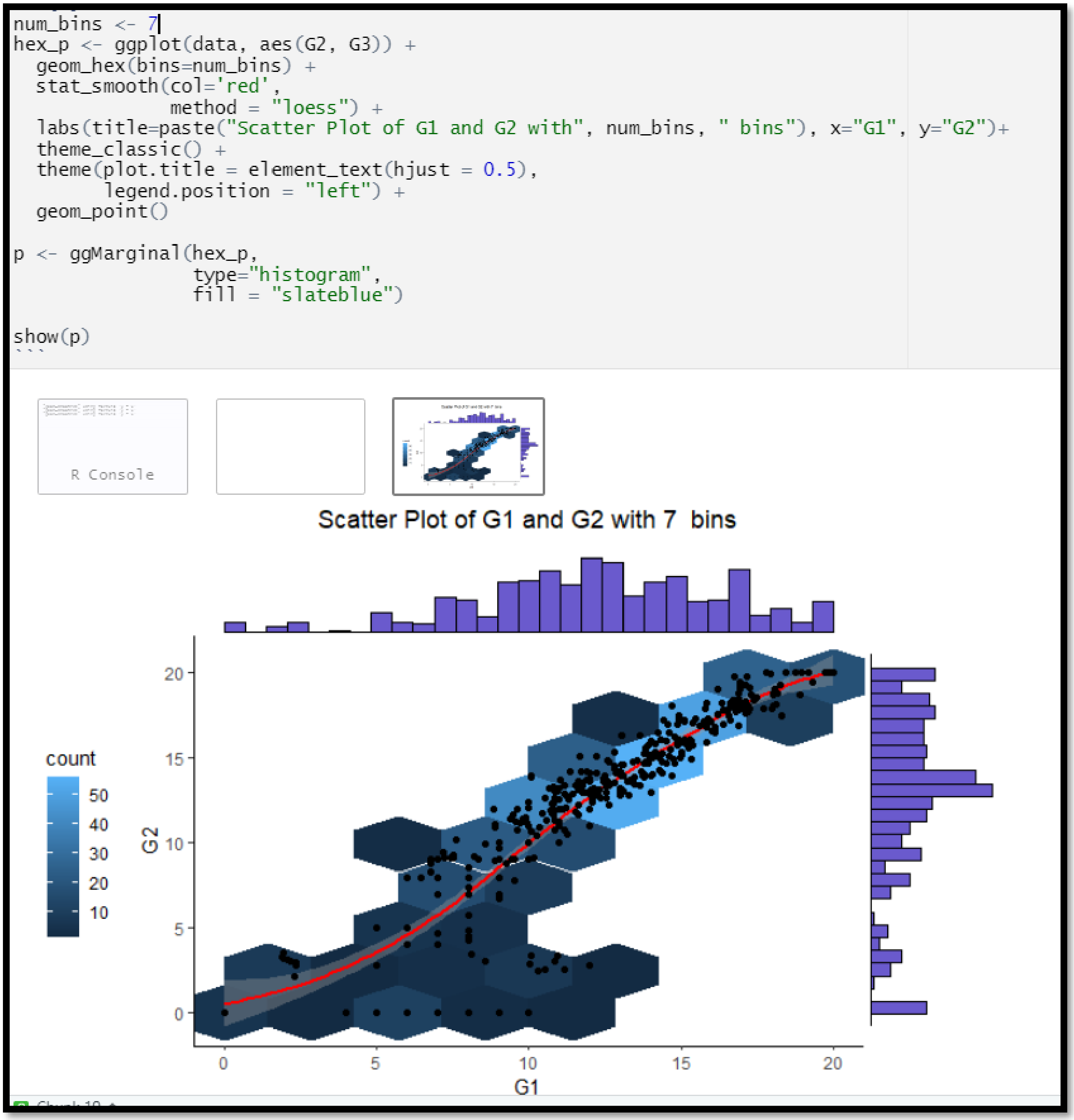
```r
p <- ggplot(data, aes(x=G1,
                      y=G2,
                      shape=sex,
                      col=sex)) +
  geom_point() +
  geom_smooth(method=lm,
              se=F) +
  labs(title=paste("Scatter Plot of G1 and G2"), x="G1", y="G2")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

show(p)
```

```
`geom_smooth()` using formula 'y ~ x'
```

As we can see, the strong correlation exists for both males and females.

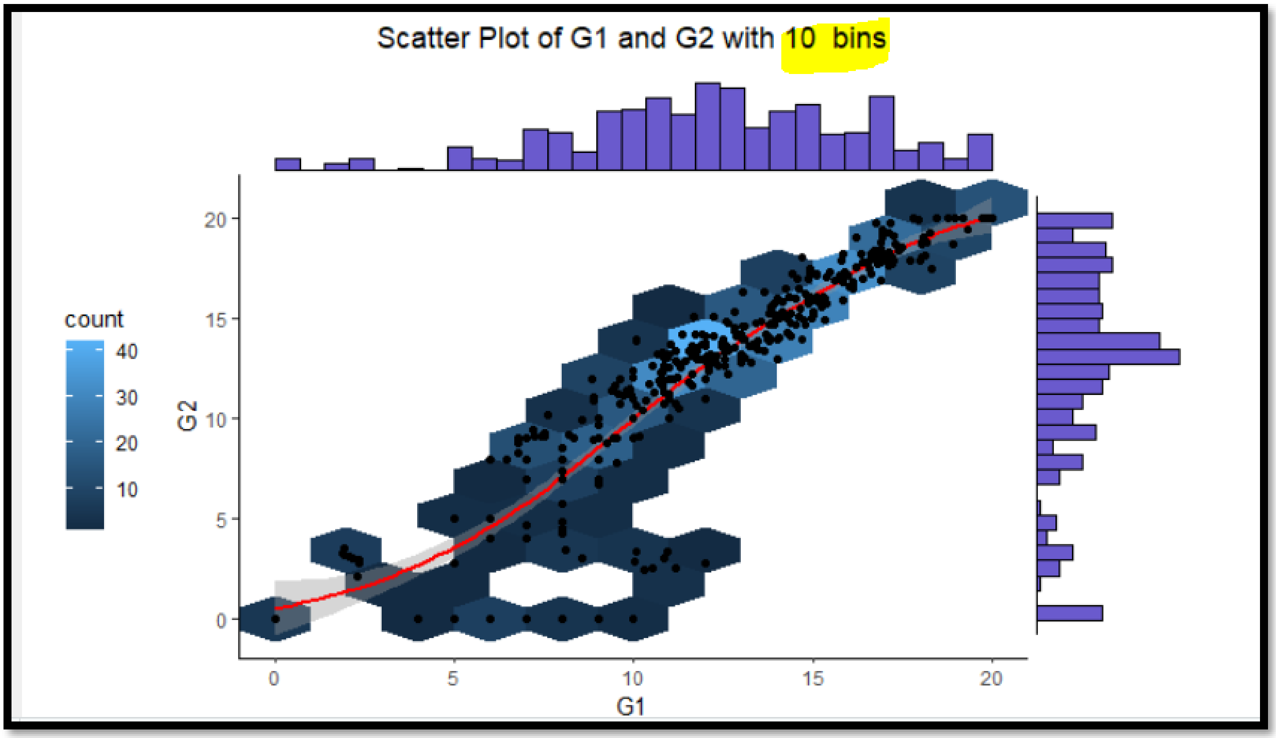
3.G

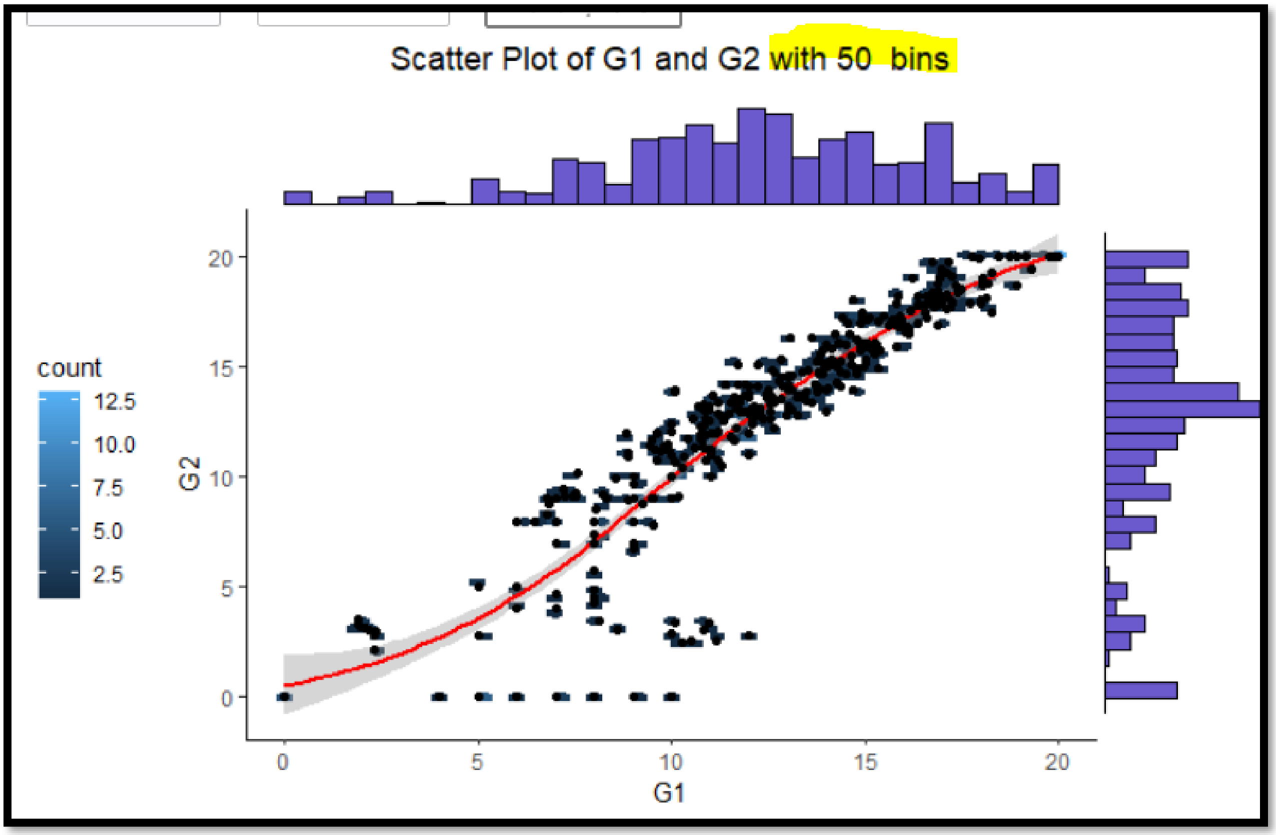


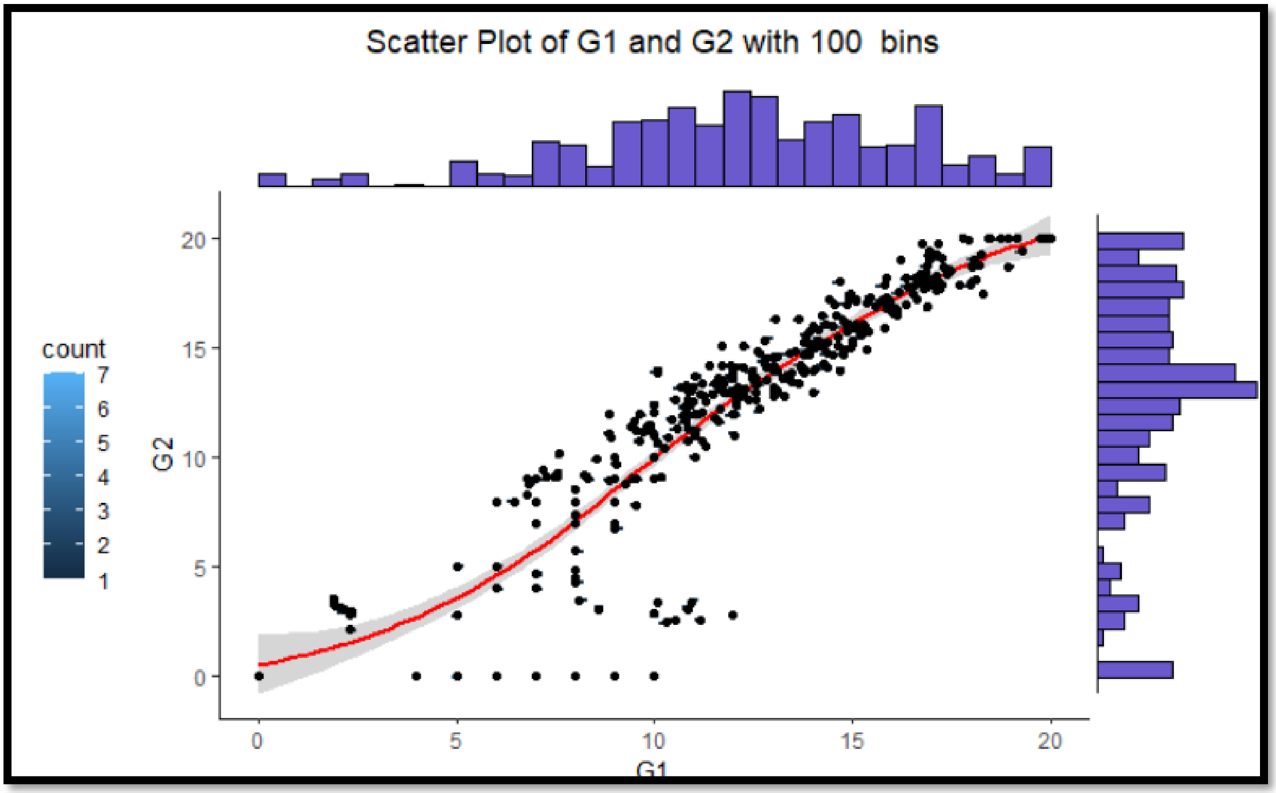
```
num_bins <- 7
hex_p <- ggplot(data, aes(G2, G3)) +
  geom_hex(bins=num_bins) +
  stat_smooth(col='red',
              method = "loess") +
  labs(title=paste("Scatter Plot of G1 and G2 with", num_bins, " bins"), x="G1", y="G2")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "left") +
  geom_point()

p <- ggMarginal(hex_p,
                type="histogram",
                fill = "slateblue")

show(p)
```


Scatter Plot of G1 and G2 with 10 bins
count
40
30
20
10
G2
20
15
10
5
0
0
5
10
15
20
G1


Scatter Plot of G1 and G2 with 50 bins
count
12.5
10.0
7.5
5.0
2.5
G2
20
15
10
5
0
0
5
10
15
20
G1



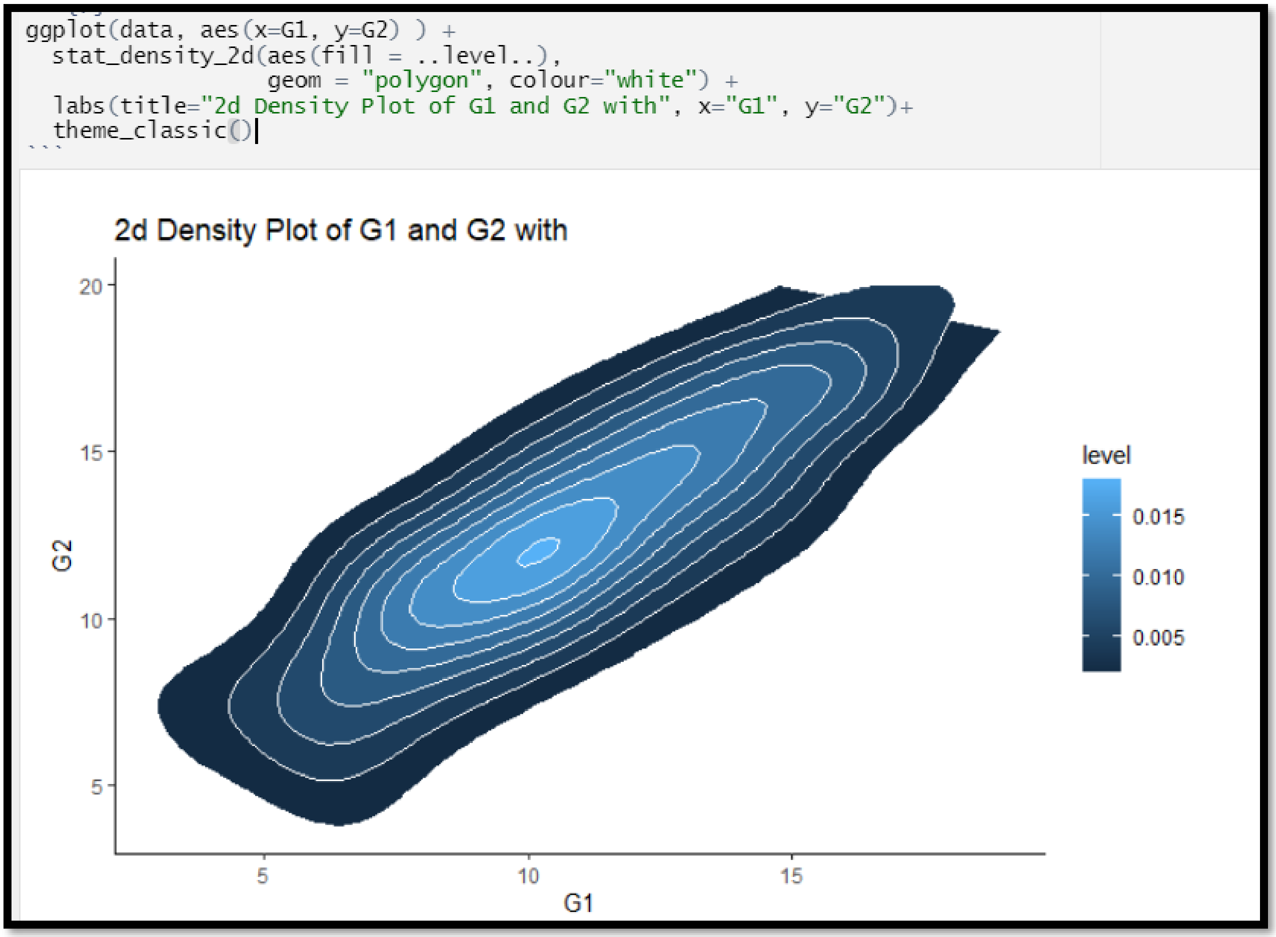
When we increase number of bins, bin size decreases and the bins become sparser. When bin size decreases, less points are located in each bin and the visualization will resemble a scatter plot. If bins be very large then too many points will be inside same bin so it will not be informative as well. So bin size ought to be set neither too big nor too small to have a desirable visualization. The interpretation of the above plot is that the density at the regions with lighter color has more density (a regular scatter plot wouldn't be able to indicate this). The marginal histogram also show the density more accurately. Also the fitted curve shows the association between G1 and G2.

3.H

The interpretation of the below plot is that the density at the regions with lighter color has more density (a regular scatter plot wouldn’t be able to indicate this).

One advantage of hexbin plot is that we can fit a curve to association of the variables. Another advantage of hexbin plot is that it divides the whole map into sections and we can use these sections to count the points falling into each section. Furthermore, because each section has equal region so it removes bias of the size of region. One drawback for hexbin plot is that we can’t use it on for example world or country maps because the way that it divides the region into hexagonal regions differs from real region division and it might be confusing.

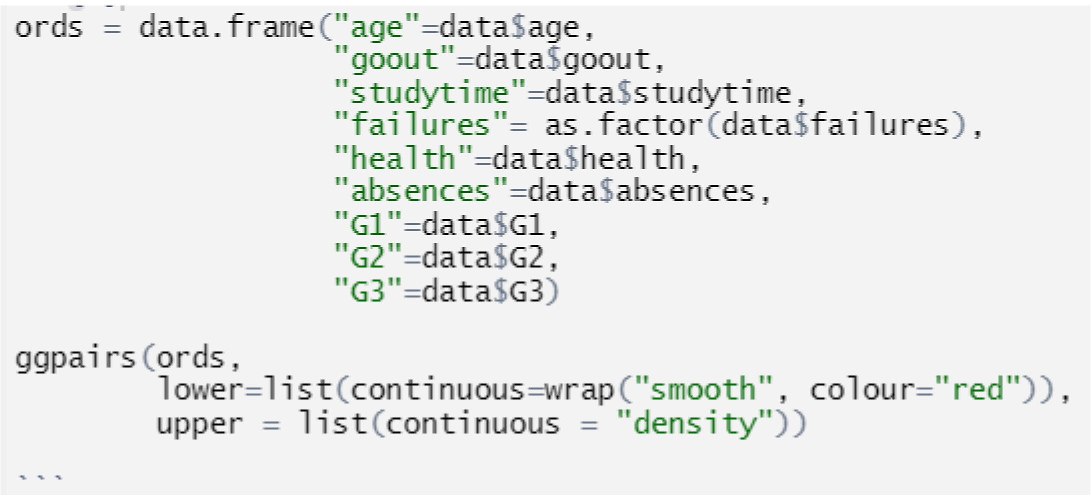
One advantage of A 2d-Density plot is that it is useful when we have a large number of points. If we use scatter plot then the result would not be informative at all if we have large number of samples. In this situation we use 2d-Density plot. 2d distribution are very useful to avoid overplotting in a scatterplot. One drawback for it is that it is not suitable if sample size is small.



```
ggplot(data, aes(x=G1, y=G2) ) +
  stat_density_2d(aes(fill = ..level..),
                  geom = "polygon", colour="white") +
  labs(title="2d Density Plot of G1 and G2 with", x="G1", y="G2")+
  theme_classic()
```

Question 4

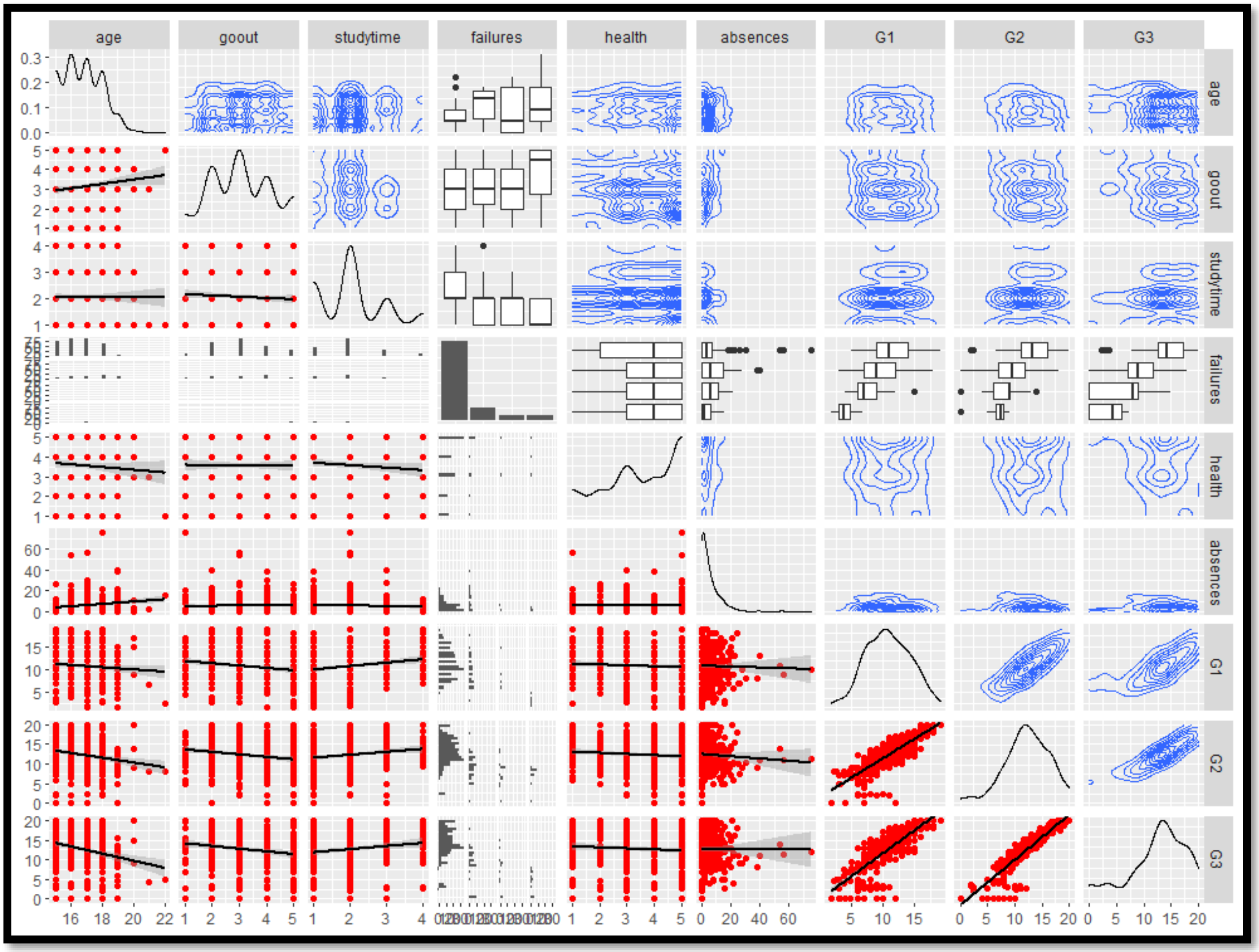
4.A



~~~
ords = data.frame("age"=data$age,
                  "goout"=data$goout,
                  "studytime"=data$studytime,
                  "failures"= as.factor(data$failures),
                  "health"=data$health,
                  "absences"=data$absences,
                  "G1"=data$G1,
                  "G2"=data$G2,
                  "G3"=data$G3)


ggpairs(ords,
        lower=list(continuous=wrap("smooth", colour="red")),
        upper = list(continuous = "density"))

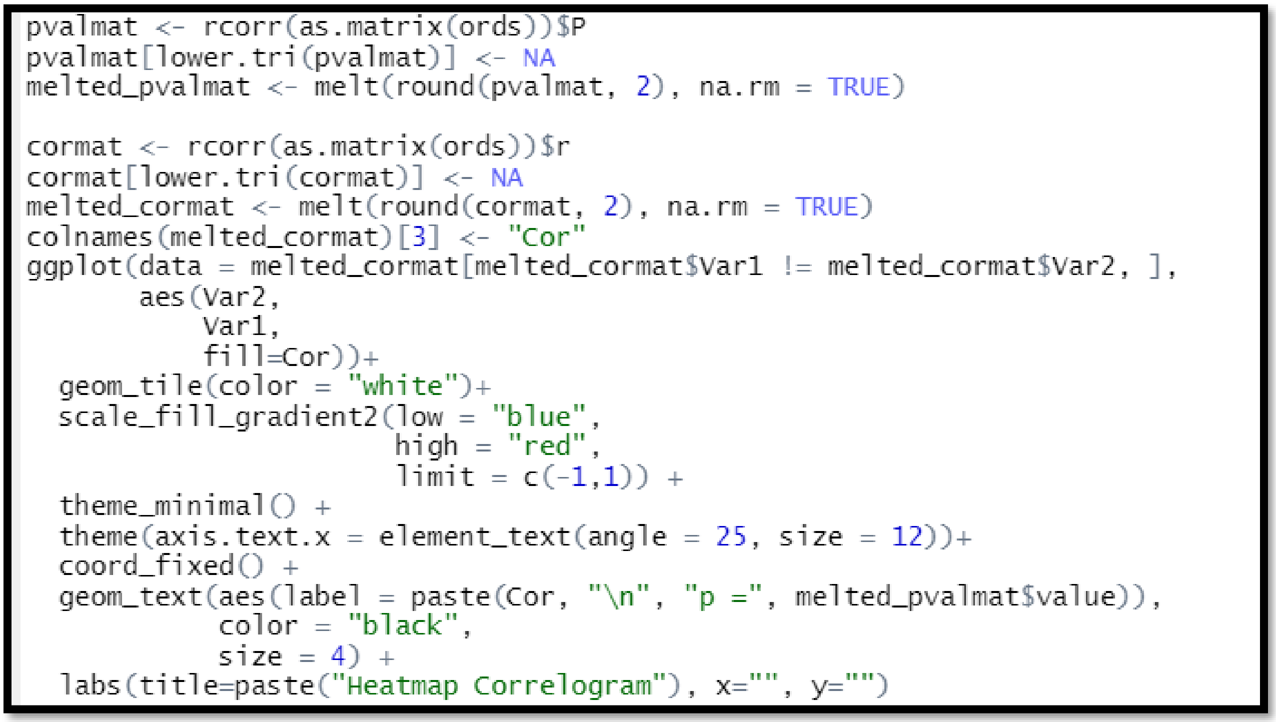
```
~~~



It is apparent from this plot that there is a strong positive linear correlation between G1, G2 and G3 which is because if someone is hard worker, they have usually high grades all courses.

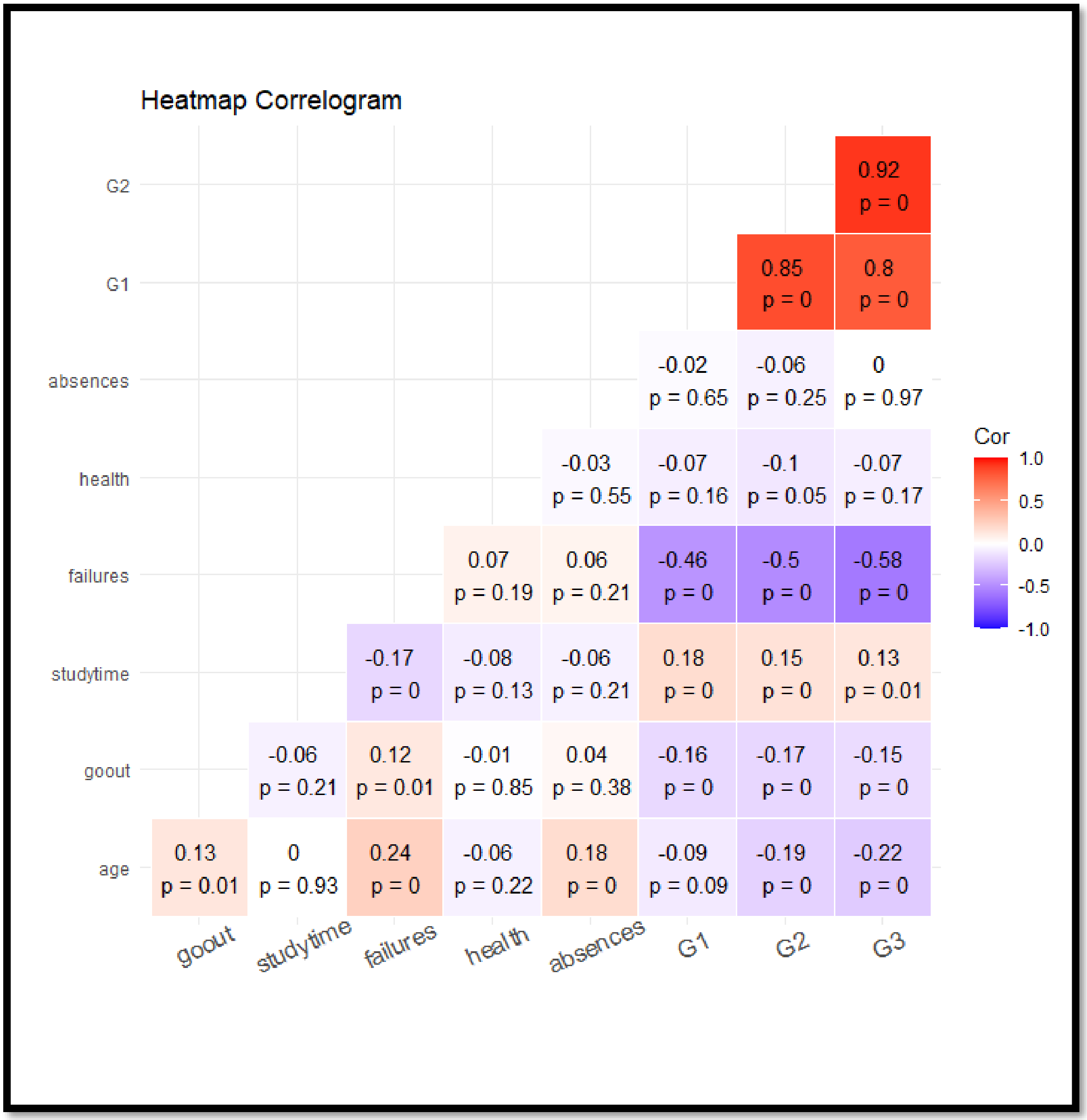
Moreover, we can see that there is a high negative correlation between failures and G1, G2 and G3. This is reasonable because if someone has good grades, they study a lot and fail less.

4.B



```
pvalmat <- rcorr(as.matrix(ords))$P
pvalmat[lower.tri(pvalmat)] <- NA
melted_pvalmat <- melt(round(pvalmat, 2), na.rm = TRUE)

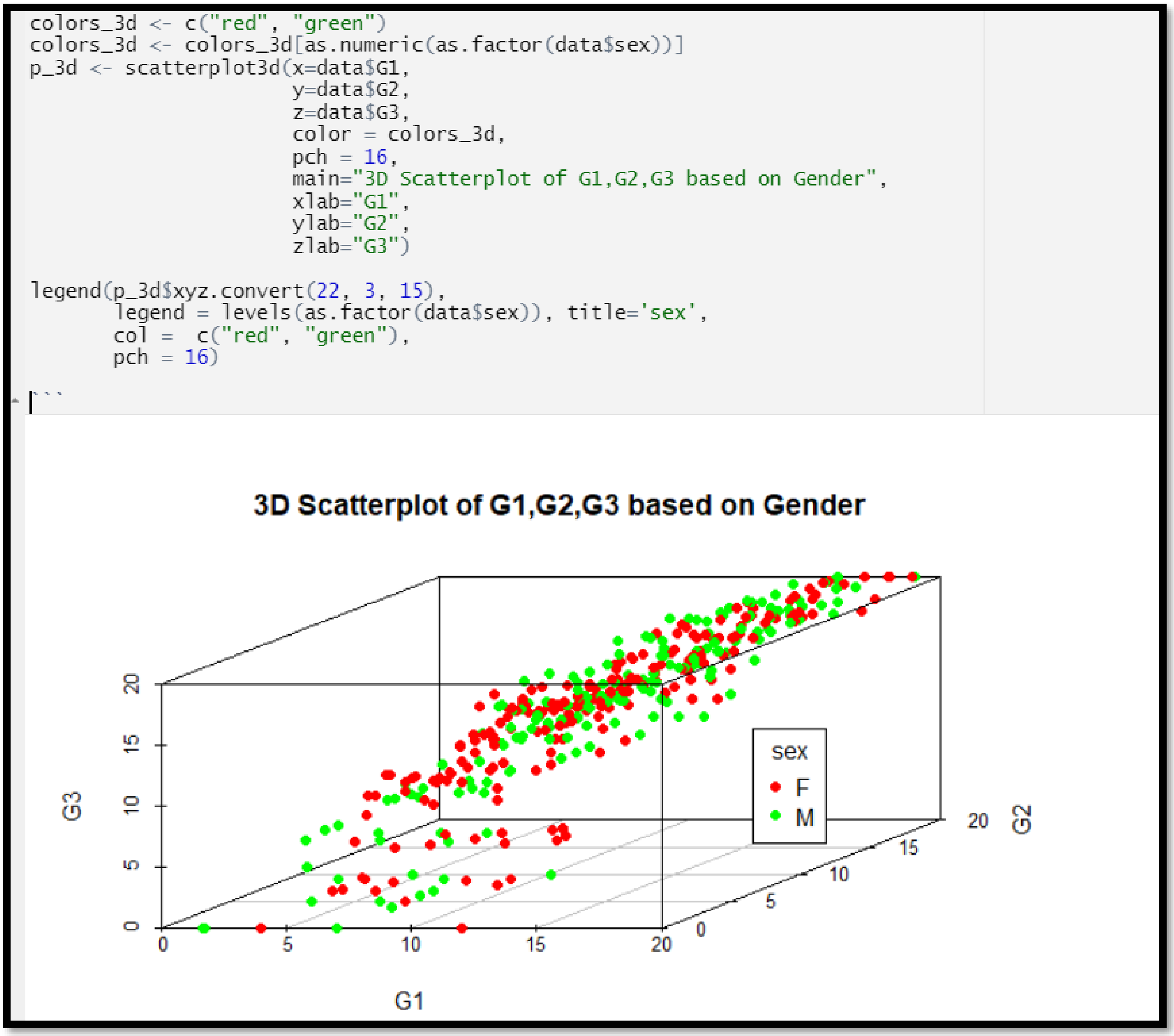
cormat <- rcorr(as.matrix(ords))$r
cormat[lower.tri(cormat)] <- NA
melted_cormat <- melt(round(cormat, 2), na.rm = TRUE)
colnames(melted_cormat)[3] <- "Cor"
ggplot(data = melted_cormat[melted_cormat$Var1 != melted_cormat$Var2, ],
       aes(Var2,
           Var1,
           fill=Cor))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue",
                       high = "red",
                       limit = c(-1,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, size = 12))+
  coord_fixed() +
  geom_text(aes(label = paste(Cor, "\n", "p =", melted_pvalmat$value)),
            color = "black",
            size = 4) +
  labs(title=paste("Heatmap Correlogram"), x="", y="")
```



	goout	studytime	failures	health	absences	G1	G2	G3
G2								0.92 p = 0
G1							0.85 p = 0	0.8 p = 0
absences						-0.02 p = 0.65	-0.06 p = 0.25	0 p = 0.97
health					-0.03 p = 0.55	-0.07 p = 0.16	-0.1 p = 0.05	-0.07 p = 0.17
failures				0.07 p = 0.19	0.06 p = 0.21	-0.46 p = 0	-0.5 p = 0	-0.58 p = 0
studytime			-0.17 p = 0	-0.08 p = 0.13	-0.06 p = 0.21	0.18 p = 0	0.15 p = 0	0.13 p = 0.01
goout		-0.06 p = 0.21	0.12 p = 0.01	-0.01 p = 0.85	0.04 p = 0.38	-0.16 p = 0	-0.17 p = 0	-0.15 p = 0
age	0.13 p = 0.01	0 p = 0.93	0.24 p = 0	-0.06 p = 0.22	0.18 p = 0	-0.09 p = 0.09	-0.19 p = 0	-0.22 p = 0

This heatmap correlation plot better shows the correlations between the variables and it verifies the results we obtained in part 4.A for example G1 and G2 and G3 are red among eachother.

4.C



~~~
colors_3d <- c("red", "green")
colors_3d <- colors_3d[as.numeric(as.factor(data$sex))]
p_3d <- scatterplot3d(x=data$G1,
                      y=data$G2,
                      z=data$G3,
                      color = colors_3d,
                      pch = 16,
                      main="3D Scatterplot of G1,G2,G3 based on Gender",
                      xlab="G1",
                      ylab="G2",
                      zlab="G3")

legend(p_3d$xyz.convert(22, 3, 15),
       legend = levels(as.factor(data$sex)), title='sex',
       col =  c("red", "green"),
       pch = 16)

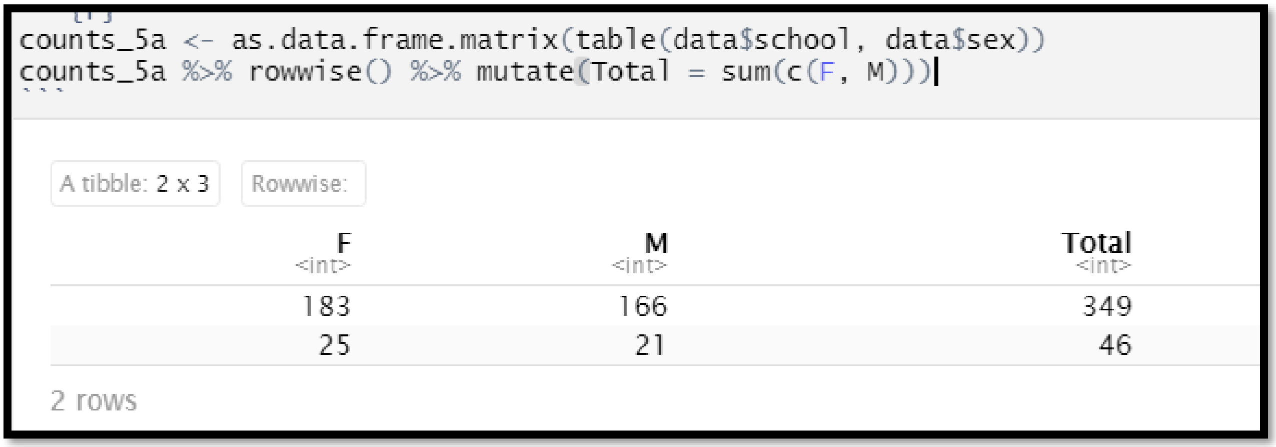
```
~~~

I chose G1, G2 and G3 as numerical and sex as categorical. We can see in the 3d scatter plot that the three grade variables (G1 and G2 and G3), not matter what is the gender of the students, have linear positive association as stated in the previous parts.

Question 5

5.A

Chosen variables: School ~ Sex



```
counts_5a <- as.data.frame.matrix(table(data$school, data$sex))
counts_5a %>% rowwise() %>% mutate(Total = sum(c(F, M)))
```

```
A tibble: 2 x 3    Rowwise:

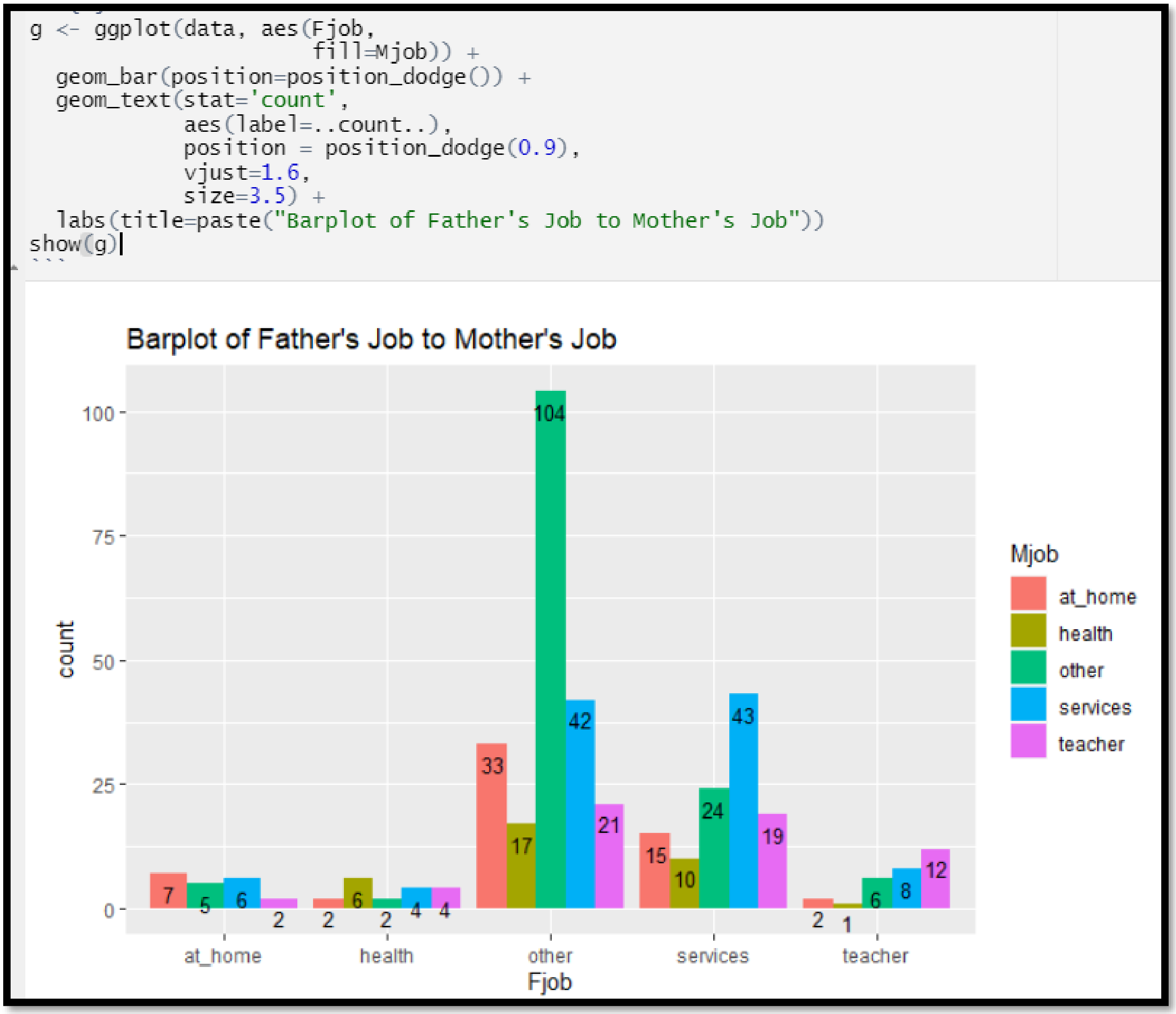
          F            M            Total
      <int>        <int>            <int>
        183          166              349
         25           21               46
2 rows
```

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

We can see that at both schools females are a bit more than males. Also the two genders comprise equal proportion of the students at both schools.

5.B

Chosen variables: Fjob ~Mjob

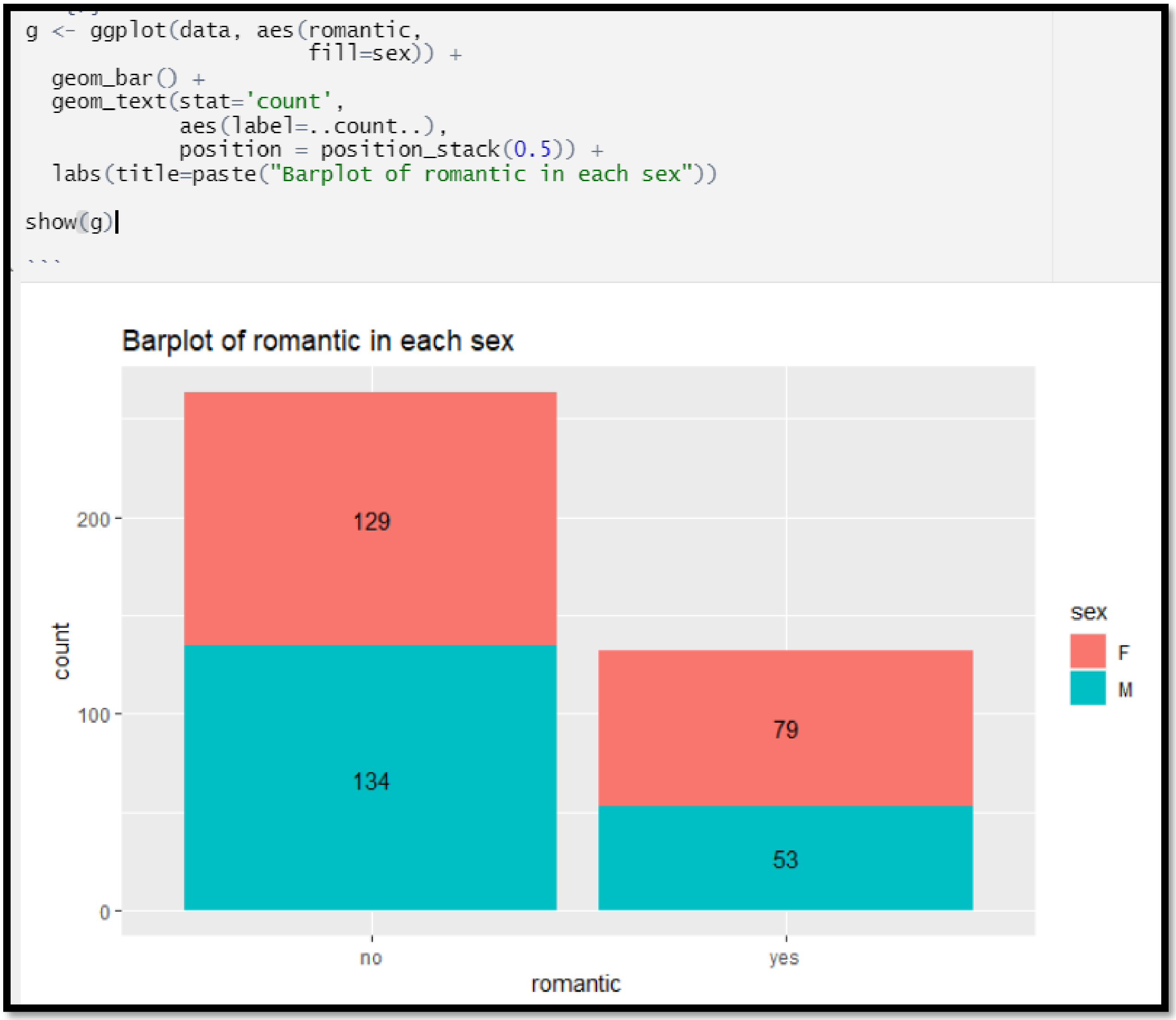


```
g <- ggplot(data, aes(Fjob,
                      fill=Mjob)) +
  geom_bar(position=position_dodge()) +
  geom_text(stat='count',
            aes(label=..count..),
            position = position_dodge(0.9),
            vjust=1.6,
            size=3.5) +
  labs(title=paste("Barplot of Father's Job to Mother's Job"))
show(g)
```

We can see that parents usually marry someone who has job at the same category. For example fathers who work at health category, their wives are more probable to work at health category.

5.C

Chosen variables: romantic ~ sex



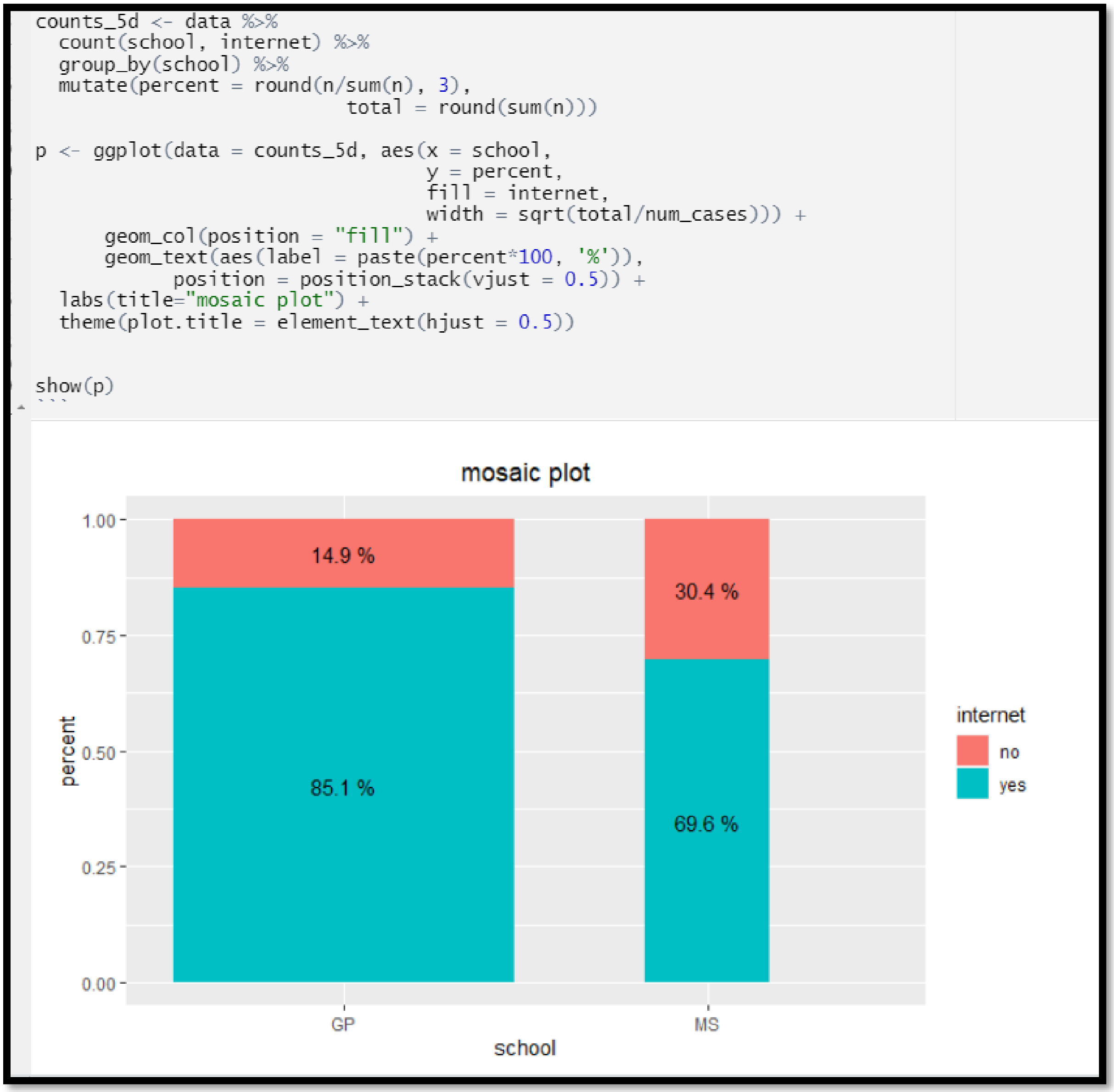
```
g <- ggplot(data, aes(romantic,
                      fill=sex)) +
  geom_bar() +
  geom_text(stat='count',
            aes(label=..count..),
            position = position_stack(0.5)) +
  labs(title=paste("Barplot of romantic in each sex"))

show(g)
```
```

In the plot above, first of all we see that most of the population have not romantic relationship. Moreover, females are more in romantic relationships that males.
```

5.D

Chosen variables: school ~ internet access



```
counts_5d <- data %>%
  count(school, internet) %>%
  group_by(school) %>%
  mutate(percent = round(n/sum(n), 3),
                          total = round(sum(n)))

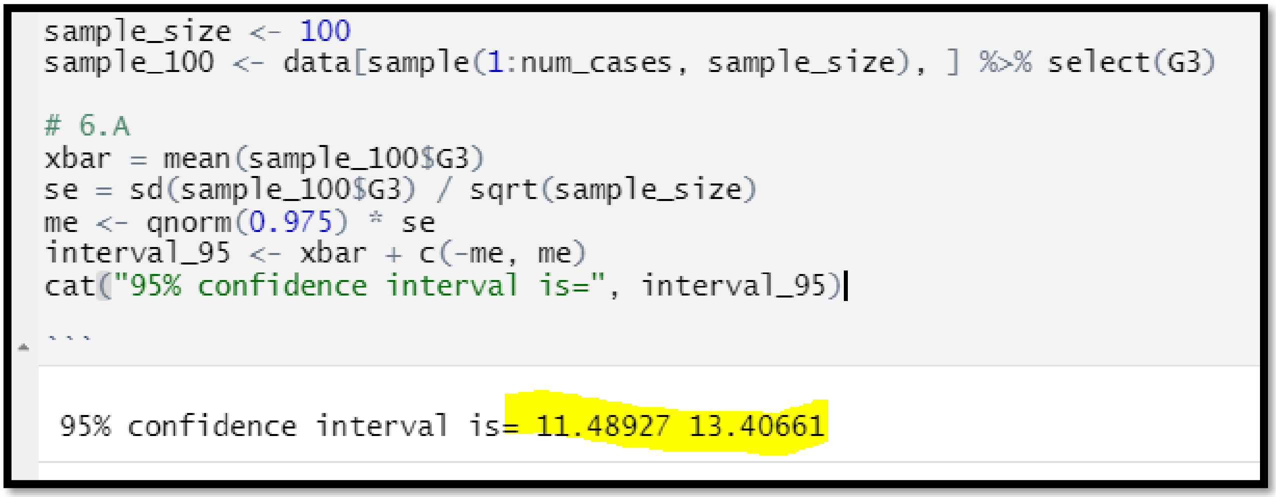
p <- ggplot(data = counts_5d, aes(x = school,
                                  y = percent,
                                  fill = internet,
                                  width = sqrt(total/num_cases))) +
      geom_col(position = "fill") +
      geom_text(aes(label = paste(percent*100, '%')),
            position = position_stack(vjust = 0.5)) +
  labs(title="mosaic plot") +
  theme(plot.title = element_text(hjust = 0.5))


show(p)
```

First of all number of students from GP are by far more than that from MS. Also at GP, about 85% of students have internet access compared to MS at which about 70% percent have internet access. This might mean that in general students at GP are from higher-income families that can more provide them with internet access.

Question 6

6.A



```
sample_size <- 100
sample_100 <- data[sample(1:num_cases, sample_size), ] %>% select(G3)

# 6.A
xbar = mean(sample_100$G3)
se = sd(sample_100$G3) / sqrt(sample_size)
me <- qnorm(0.975) * se
interval_95 <- xbar + c(-me, me)
cat("95% confidence interval is=", interval_95)

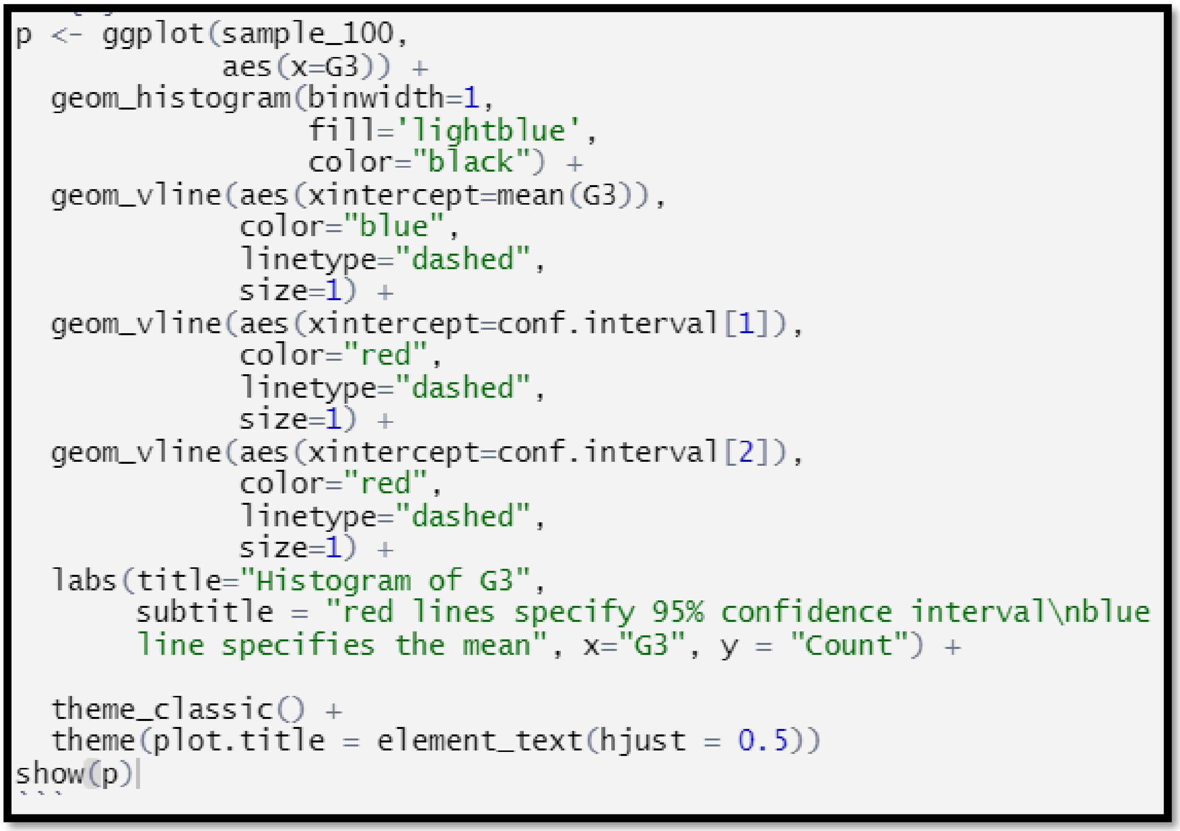
```
```

```
95% confidence interval is= 11.48927 13.40661
```

6.B

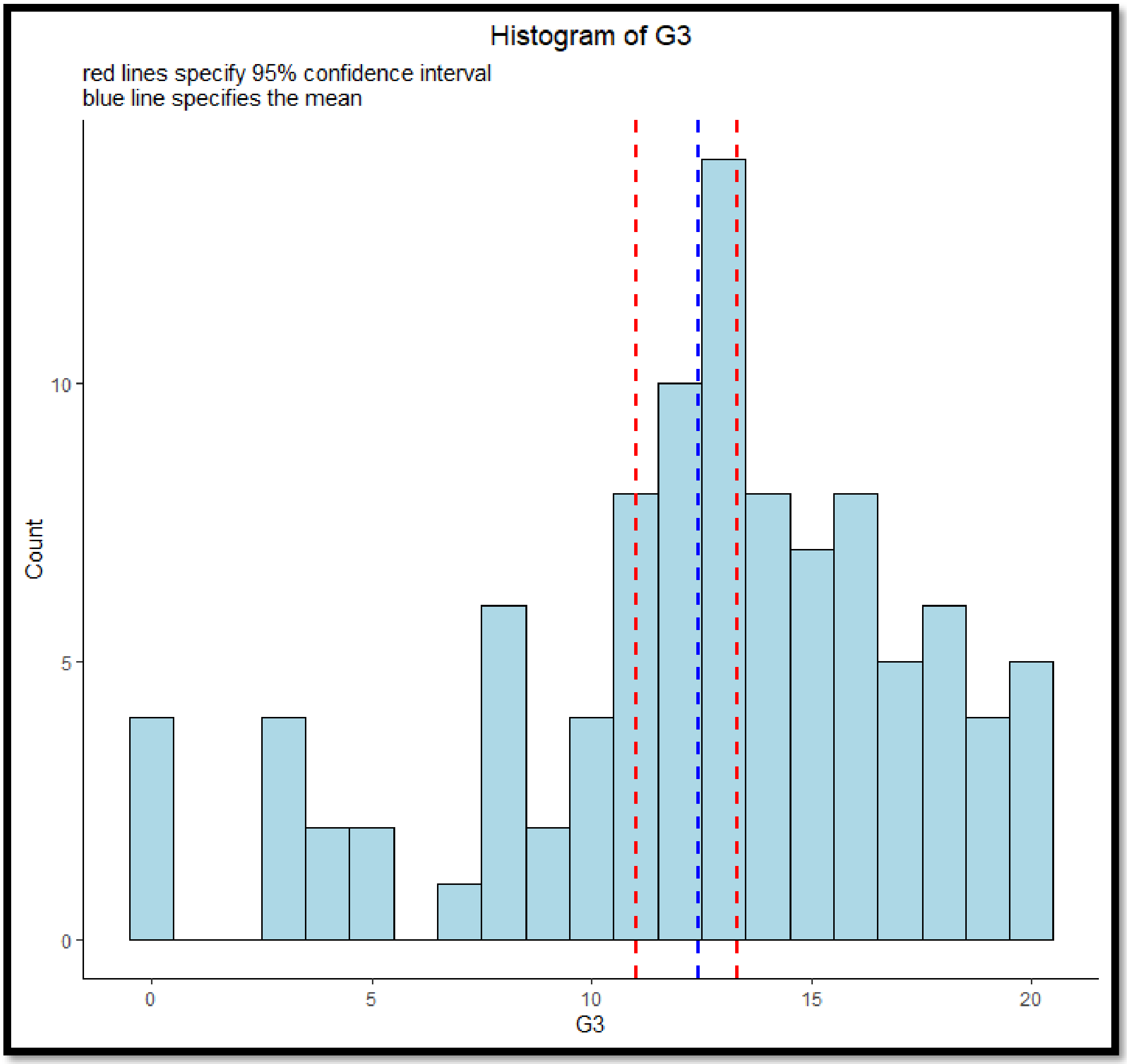
- This means that if we take many samples of size 100 students (like the one we took) and build 95% confidence intervals, about 95% of them will capture the true mean of the population.
- So, there is 95% chance for our calculated confidence interval to capture the true mean of the population.
```

6.C



```
p <- ggplot(sample_100,
            aes(x=G3)) +
  geom_histogram(binwidth=1,
                 fill='lightblue',
                 color="black") +
  geom_vline(aes(xintercept=mean(G3)),
             color="blue",
             linetype="dashed",
             size=1) +
  geom_vline(aes(xintercept=conf.interval[1]),
             color="red",
             linetype="dashed",
             size=1) +
  geom_vline(aes(xintercept=conf.interval[2]),
             color="red",
             linetype="dashed",
             size=1) +
  labs(title="Histogram of G3",
       subtitle = "red lines specify 95% confidence interval\nblue
       line specifies the mean", x="G3", y = "Count") +

  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
show(p)
```

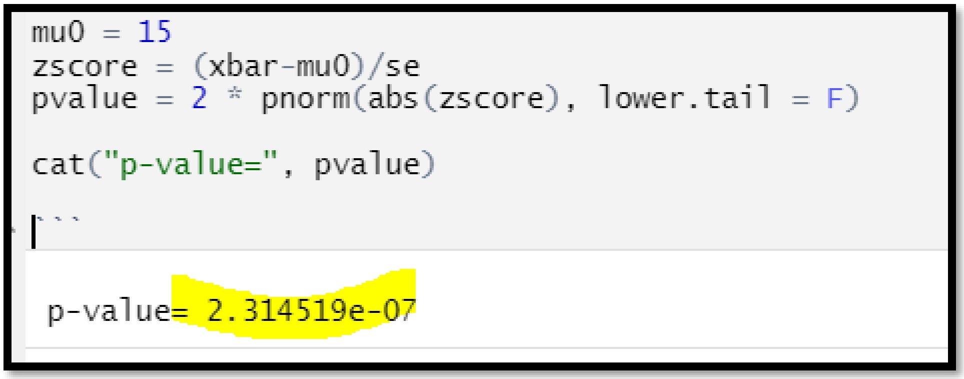


6.D

We guess that the mean of the G1 for population is not equal to 14. In order to test this hypothesis, we create a null hypothesis that states that the mean is 14.

$$\boldsymbol{H_0}: \quad \boldsymbol{\mu_0} = \mathbf{15}$$

$$\boldsymbol{H_A}: \quad \boldsymbol{\mu_0} \ \boldsymbol{is\ not\ eual\ to}\ \mathbf{15}$$



```
mu0 = 15
zscore = (xbar-mu0)/se
pvalue = 2 * pnorm(abs(zscore), lower.tail = F)

cat("p-value=", pvalue)

```

 p-value= 2.314519e-07
```

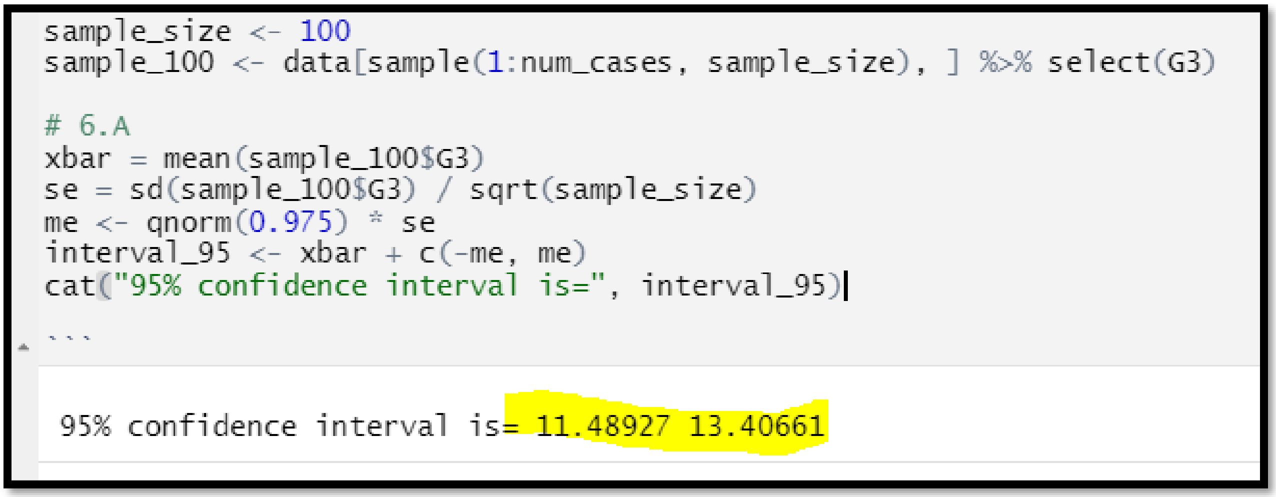
Significance level : 5%

So, because the p-value is smaller than significance level, the null hypothesis is rejected in favor of the alternative one.

Therefor we conclude that the mean is not equal to 15.

6.E

The calculated interval in part A was:



```
sample_size <- 100
sample_100 <- data[sample(1:num_cases, sample_size),] %>% select(G3)

# 6.A
xbar = mean(sample_100$G3)
se = sd(sample_100$G3) / sqrt(sample_size)
me <- qnorm(0.975) * se
interval_95 <- xbar + c(-me, me)
cat("95% confidence interval is=", interval_95)

```

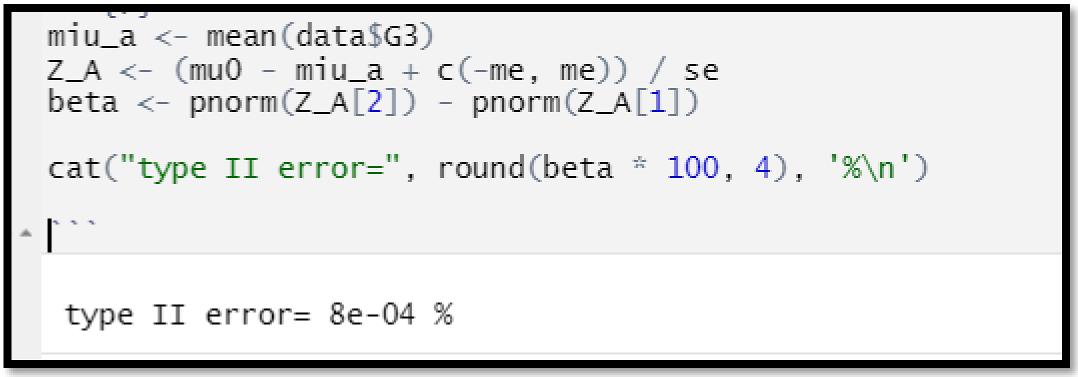
 95% confidence interval is= 11.48927 13.40661
```

Yes, because 15 doesn't fall into the confidence interval, null hypothesis is rejected and hence it supports our conclusion at the previous part.

This is reasonable that the result of these two methods are the same because if a point A is far from B, then point B will also be far from point A. Similarly, if μ_0 doesn't fall in the 95% interval around sample mean, the sample mean will not fall into 95% interval around μ_0 ,hence the p-value will be smaller that

5%. So, these two methods (CI and P-value) will always yield exactly the same results.

6.F



```
miu_a <- mean(data$G3)
Z_A <- (mu0 - miu_a + c(-me, me)) / se
beta <- pnorm(Z_A[2]) - pnorm(Z_A[1])

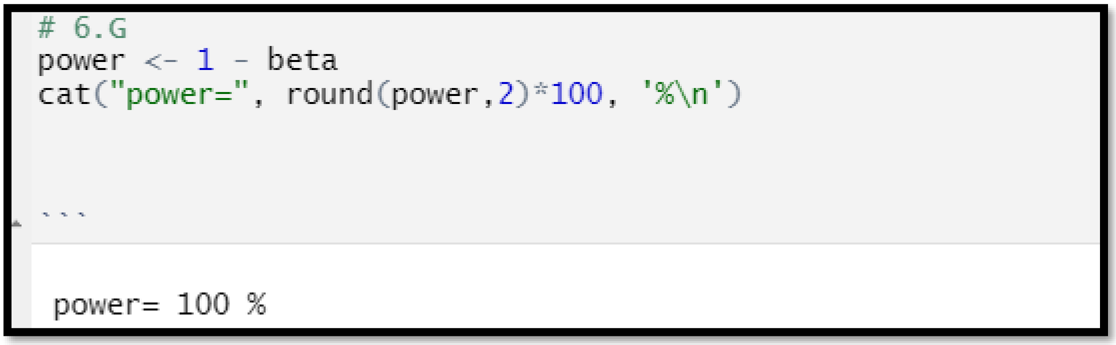
cat("type II error=", round(beta * 100, 4), '%\n')

```
```

```
type II error= 8e-04 %
```

This means that the probability of (Wrongly) not rejecting H0 if μ_A is True is about 0.0008 %.

6.G



```
# 6.G
power <- 1 - beta
cat("power=", round(power,2)*100, '%\n')

```
```

```
power= 100 %
```

This means that the probability of (Correctly) rejecting H0 if μ_A is True is about 100 %.

The larger the Effect Size, the greater the Power.

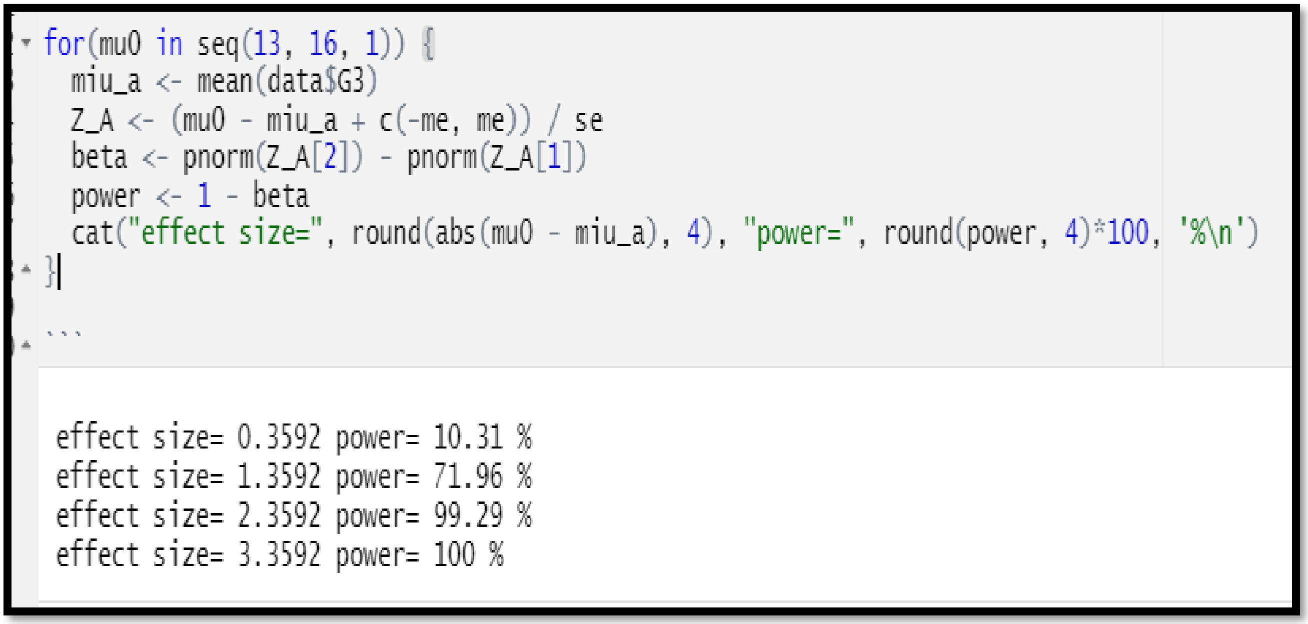
In my test the true mean was 12.64 so the effect size was :

Effect size = 15 - 12.64 = 2.36 ------> Power= 98 %

If the hypothesis mean was 14 instead of 15:

Effect size = 14 - 12.64 = 1.36 ------> Power= 65 %

And so on:



~~~
for(mu0 in seq(13, 16, 1)) {
  miu_a <- mean(data$G3)
  Z_A <- (mu0 - miu_a + c(-me, me)) / se
  beta <- pnorm(Z_A[2]) - pnorm(Z_A[1])
  power <- 1 - beta
  cat("effect size=", round(abs(mu0 - miu_a), 4), "power=", round(power, 4)*100, '%\n')
}

```
~~~

```
effect size= 0.3592 power= 10.31 %
effect size= 1.3592 power= 71.96 %
effect size= 2.3592 power= 99.29 %
effect size= 3.3592 power= 100 %
```

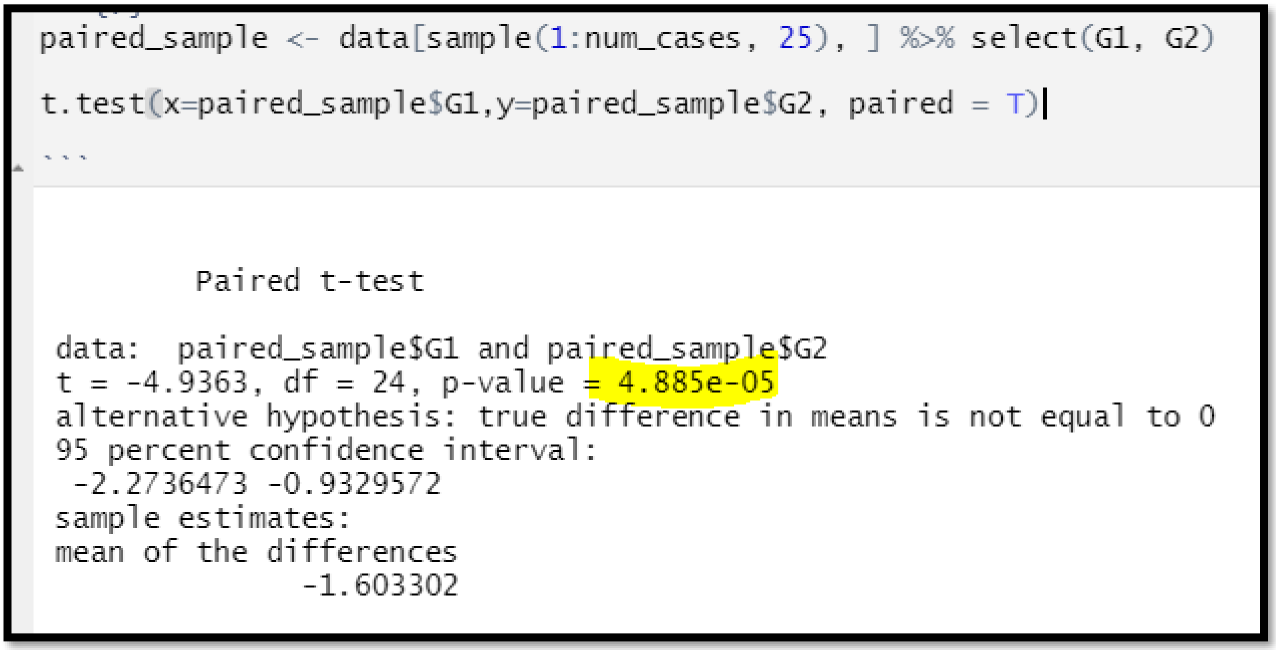
Question 7

Chosen Variables: *G1 and G2*

7.A

a. We should use t-test because the sample size is smaller than 30. This is because if we approximate the standard deviation of the population with the standard deviation of the sample distribution and because the sample size is small, then there is uncertainty about this approximation. We use this standard deviation in calculating the standard error of the mean distribution. So the uncertainty will be also in calculating the p-value. So we use t-distribution which is more cautious about the data points in the tails of the distribution to reduce the error.

b.



~~~
paired_sample <- data[sample(1:num_cases, 25), ] %>% select(G1, G2)

t.test(x=paired_sample$G1,y=paired_sample$G2, paired = T)

```


        Paired t-test

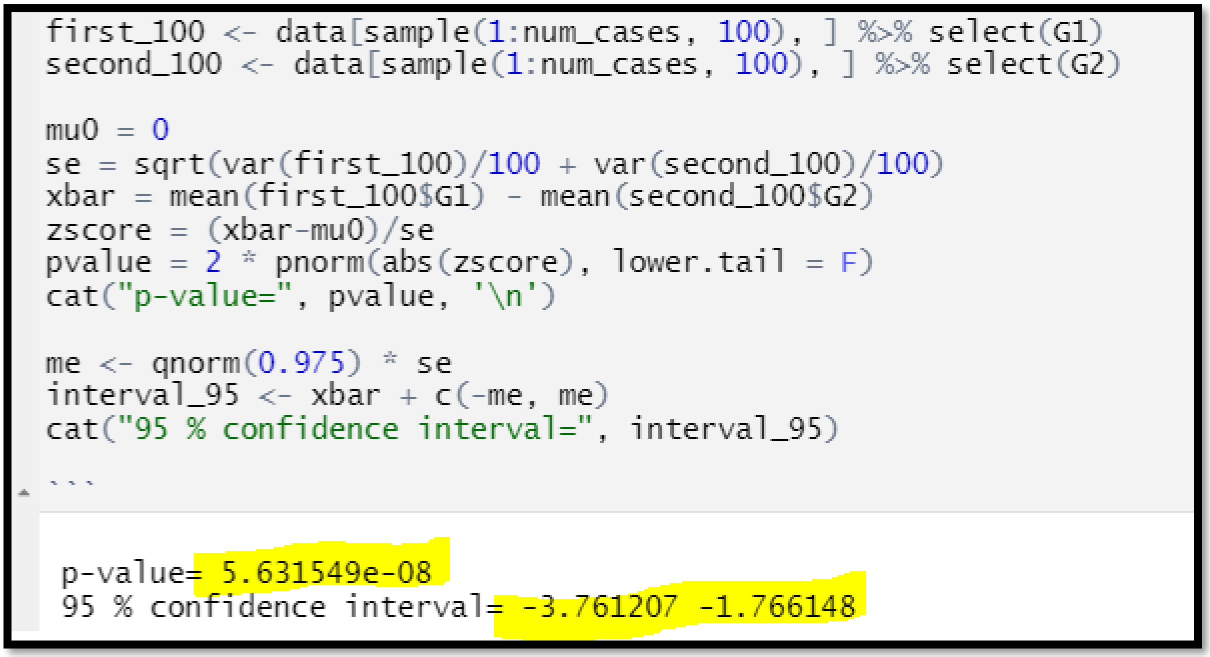
data:  paired_sample$G1 and paired_sample$G2
t = -4.9363, df = 24, p-value = 4.885e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.2736473 -0.9329572
sample estimates:
mean of the differences
              -1.603302
~~~

Because the p-value is smaller than significance level (5%) so the null hypothesis is rejected in favor of the alternative. So, there is a statistically significant difference between the mean of the two variables.

7.B

H_0: *The means of two groups are equal.*

H_A: *There is a difference between the means of two groups.*



```
first_100 <- data[sample(1:num_cases, 100), ] %>% select(G1)
second_100 <- data[sample(1:num_cases, 100), ] %>% select(G2)

mu0 = 0
se = sqrt(var(first_100)/100 + var(second_100)/100)
xbar = mean(first_100$G1) - mean(second_100$G2)
zscore = (xbar-mu0)/se
pvalue = 2 * pnorm(abs(zscore), lower.tail = F)
cat("p-value=", pvalue, '\n')

me <- qnorm(0.975) * se
interval_95 <- xbar + c(-me, me)
cat("95 % confidence interval=", interval_95)

```
```

```
p-value= 5.631549e-08
95 % confidence interval= -3.761207 -1.766148
```

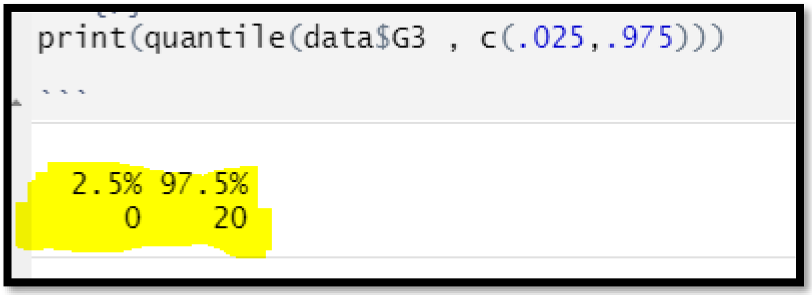
We can see that p-value is smaller than significance level (5%) so we fail to reject H0.

The hypothesized difference of means was zero which doesn’t fall into the obtained 95% confidence interval so this method fails to reject H0, too.

Hence the two methods (CI and p-value) agree on the result.
```

Question 8

8.A



~~~
print(quantile(data$G3 , c(.025,.975)))

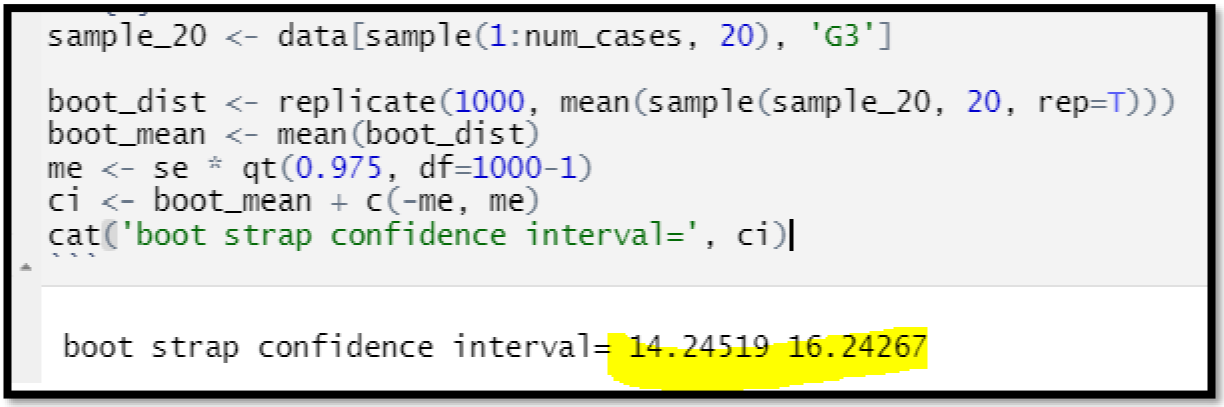
```

   2.5% 97.5%
      0    20
~~~

the confidence interval is **[0, 20]**

As we can see, based on percentile method, the confidence interval is between 0 and 20, i.e., between the smallest and largest possible value for this variable (grade); so, this interval is useless and uninformative because this interval is the whole range.

8.B



```
sample_20 <- data[sample(1:num_cases, 20), 'G3']

boot_dist <- replicate(1000, mean(sample(sample_20, 20, rep=T)))
boot_mean <- mean(boot_dist)
me <- se * qt(0.975, df=1000-1)
ci <- boot_mean + c(-me, me)
cat('boot strap confidence interval=', ci)

 boot strap confidence interval= 14.24519 16.24267
```

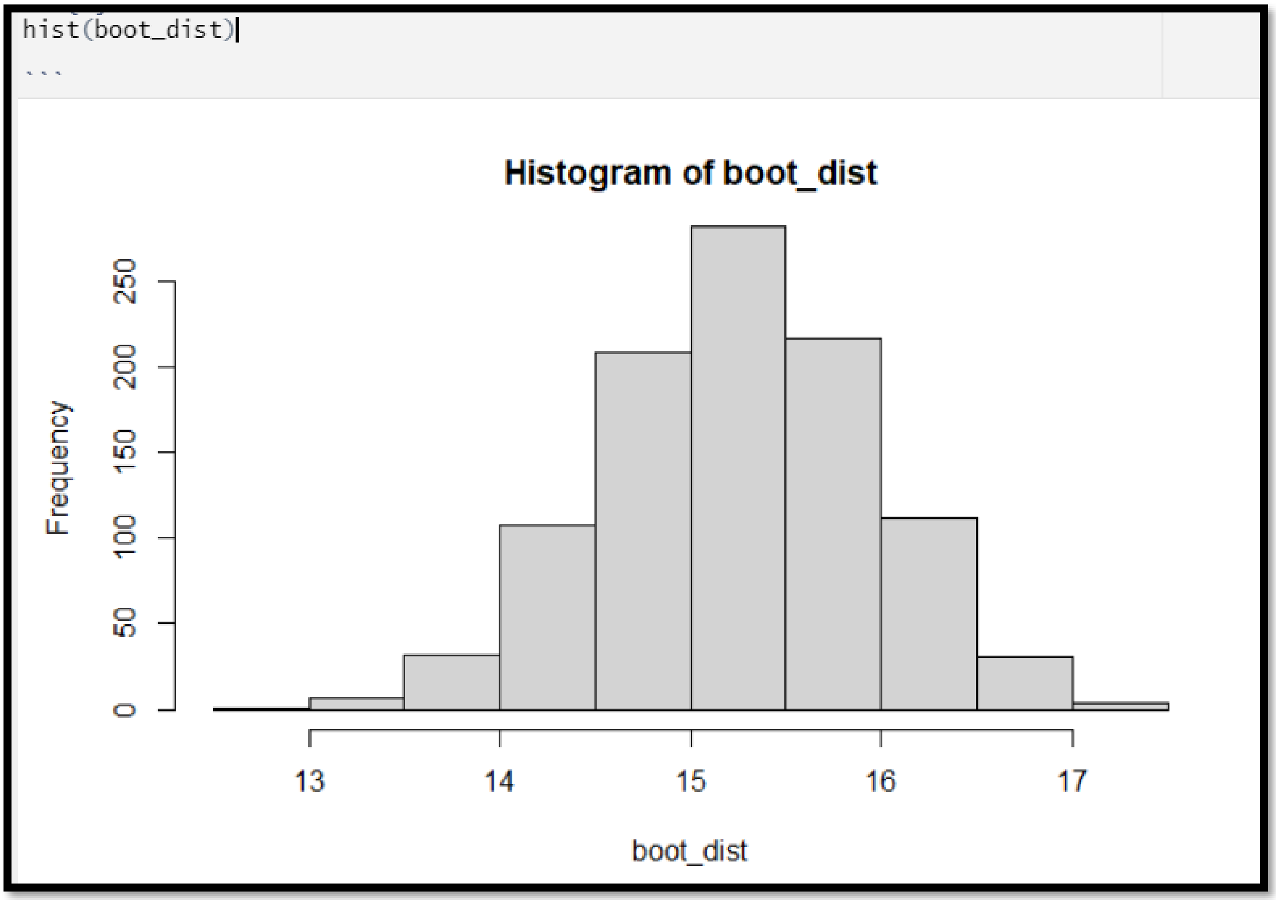
the confidence interval is **[14.24,   16.24]**

Here the confidence interval is much better and more informative and gives us more precise range.

8.C

Yes, there is a noticeable difference between these two intervals. This is because in part A we use the actual distribution which has a lot of outliers and it is left skewed. This skewness influences (in a negative way) the confidence interval when we use percentile method (as we can see the lower and upper boundaries on the interval are the minimum and maximums of the data). However, in part B we use bootstrapping which eliminates the skewness and has a more normal-shaped distribution. Beside this, in part B we use standard error method which is more precise than the percentile method.

Here is the distribution of bootstrap sampling which is pretty symmetric normal distribution and has not skewness:

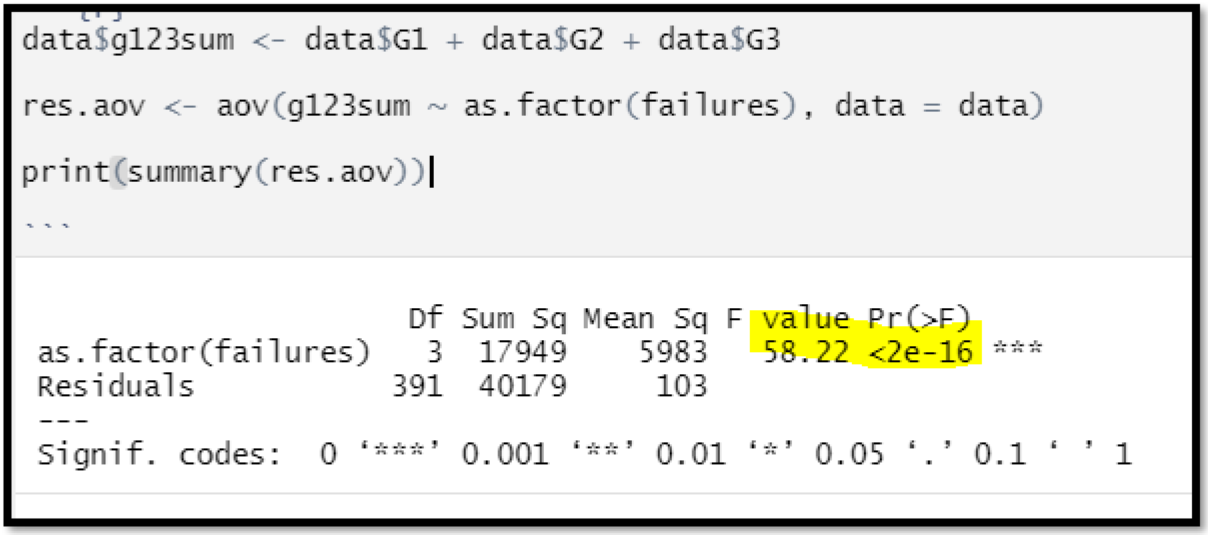


```
hist(boot_dist)
```
```
```

Question 9

Here want to test whether there is a difference between mean of G1+G2+G3 of the four groups of failure: 0, 1, 2 or 3 failures.

We use ANOVA for this purpose. The result is:



~~~
data$g123sum <- data$G1 + data$G2 + data$G3

res.aov <- aov(g123sum ~ as.factor(failures), data = data)

print(summary(res.aov))

```
~~~

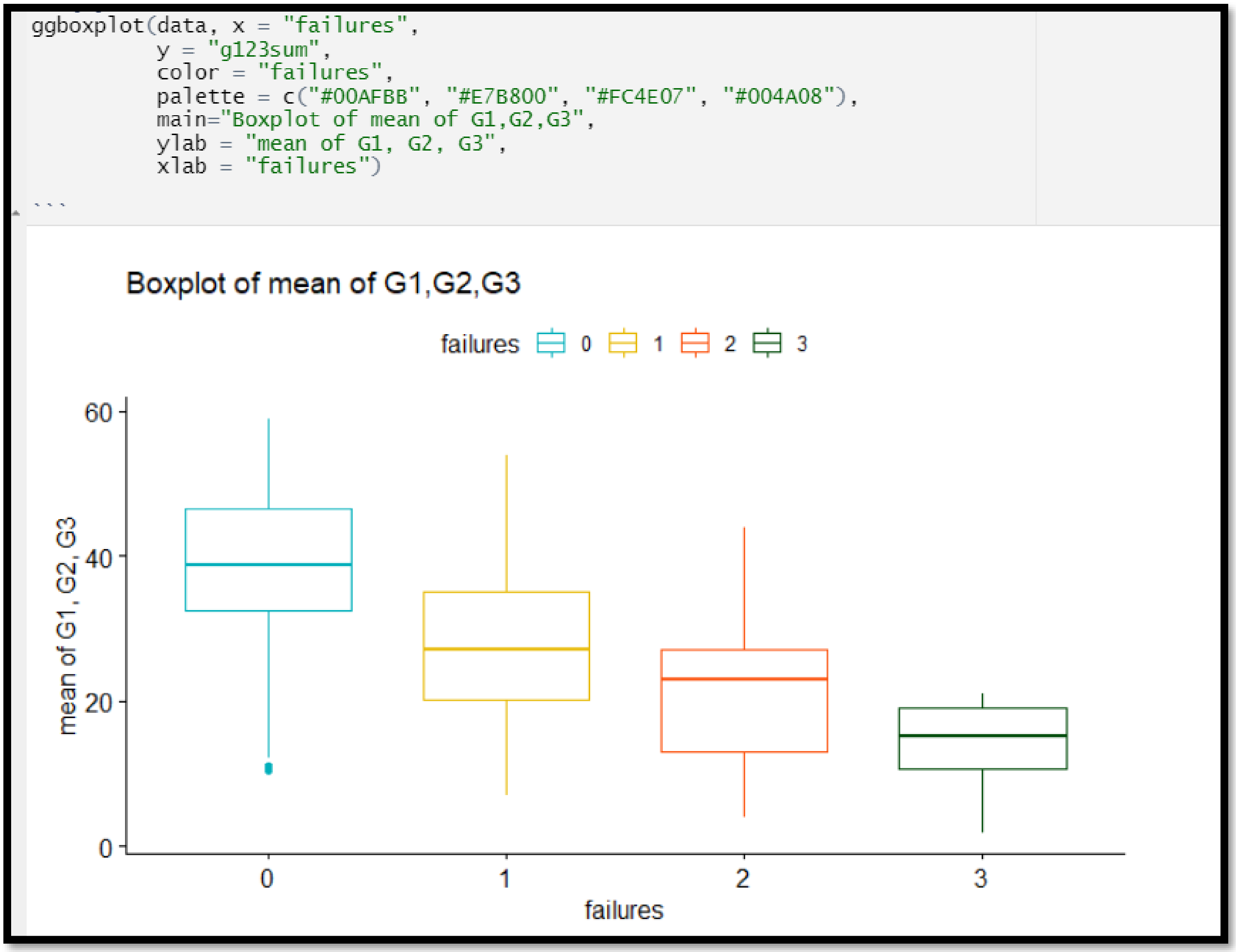
```
                     Df Sum Sq Mean Sq F value Pr(>F)    
as.factor(failures)   3  17949    5983   58.22 <2e-16 ***
Residuals           391  40179     103                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because p-value is smaller than significance level, we reject the null hypothesis.

Conclusion:

This means that there is a difference between at least two of the groups.

Here is the boxplot associated with these four groups:

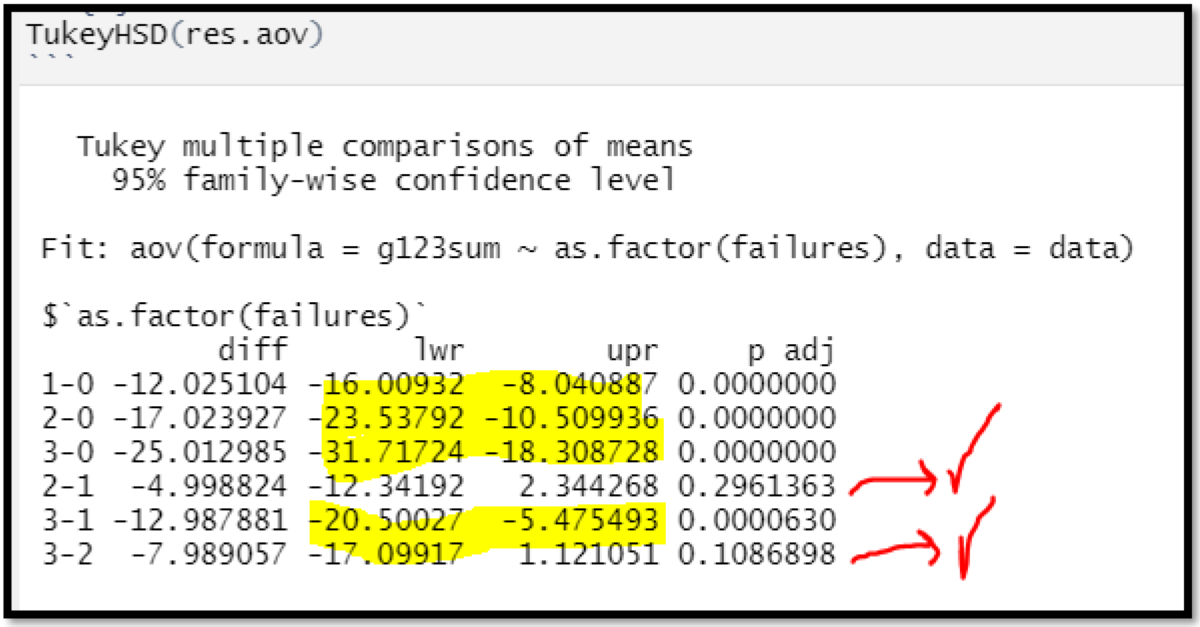


```
ggboxplot(data, x = "failures",
          y = "g123sum",
          color = "failures",
          palette = c("#00AFBB", "#E7B800", "#FC4E07", "#004A08"),
          main="Boxplot of mean of G1,G2,G3",
          ylab = "mean of G1, G2, G3",
          xlab = "failures")
```

As we can see in the group-boxplot above, it seems that the mean of group 0 is different from the rest. Groups 1 and 3 are also seem to have different means.

In purpose of making sure that our guess is actually right or not we use pairwise comparison.

Here are the results:



~~~
TukeyHSD(res.aov)
```

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = g123sum ~ as.factor(failures), data = data)

$`as.factor(failures)`
          diff       lwr        upr     p adj
1-0 -12.025104 -16.00932  -8.040887 0.0000000
2-0 -17.023927 -23.53792 -10.509936 0.0000000
3-0 -25.012985 -31.71724 -18.308728 0.0000000
2-1  -4.998824 -12.34192   2.344268 0.2961363 → ✓
3-1 -12.987881 -20.50027  -5.475493 0.0000630
3-2  -7.989057 -17.09917   1.121051 0.1086898 → ✓
~~~

Conclusion:

Because the confidence intervals of 1-0 , 2-0 , 3-0 and 3-1 do not contain 0 so we conclude that the mean of G1+G2+G3 is different for these pairs, as suggested by the boxplot.