Statistical Inference

Project Phase1

Hamidreza Aliakbary khoyi

810196514

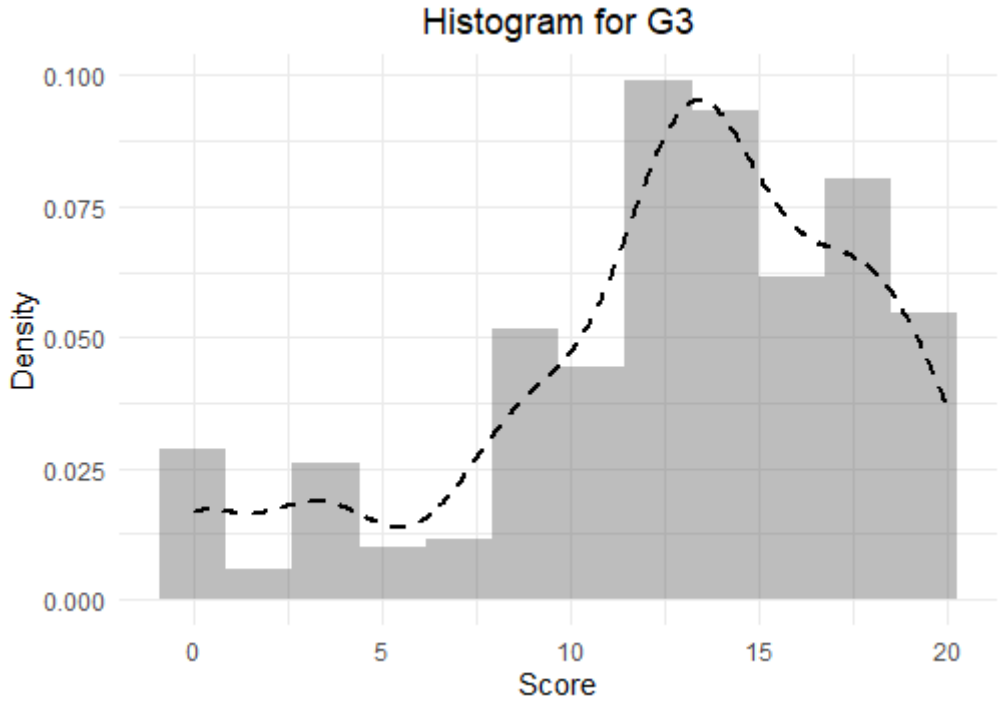
Spring 2021

Question 0

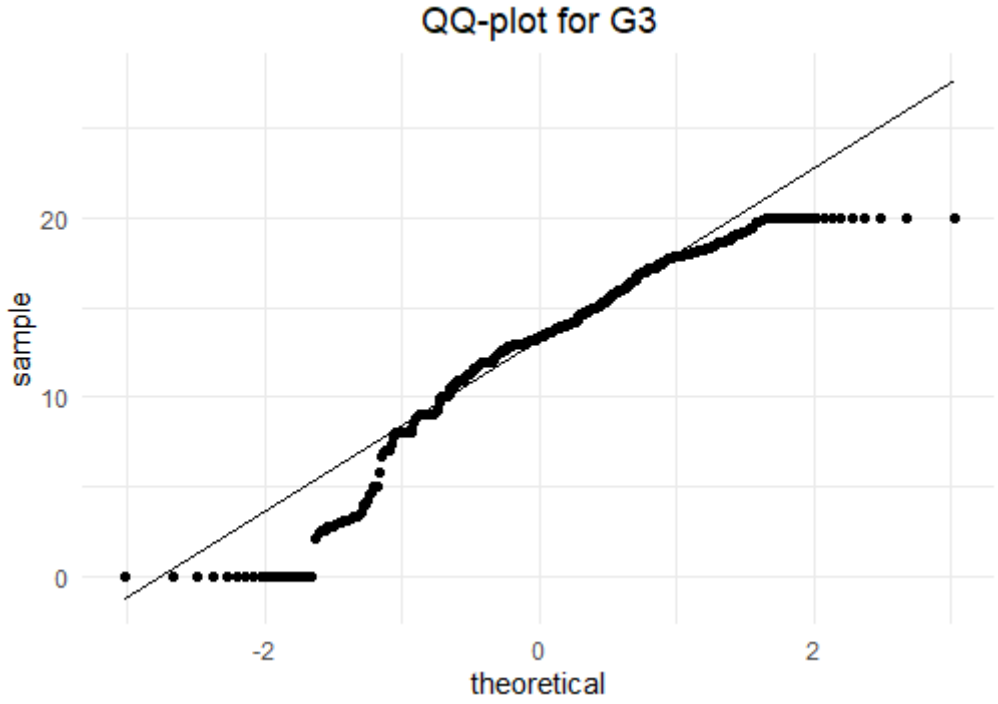
a. My data set is about students' performance that describes mane features regarding their whole performance, including 3 grades of different tests, type of school, sex, age, parents' job, their failures and absences.
   It is good to study this dataset, since we can identify main parameters that affect students performance in their academic year.
b. My dataset has 16 features.
c. I looked up my dataset for missing data, but I didn't find any, and using R's NA value counter, again, there was no such kind of variable. If there was any, I could handle it in many ways:
   - Clean whole rows including NA type.
   - Replacing NA with median or mean in their specific column.
d. With an elementary view, I guess important variables can be students grades, number of absences, health, No. failures, gout, school. From suffisticating view I guessed these variables could have more relevance to affect grades of students.

Question 1

a. After calculating efficient bin-size with this formula:
   bw <- 2 * IQR(selected.numerical) / length(selected.numerical)^(1/3)
   I got a plot for density with histogram for G3 grades:
   Plot:

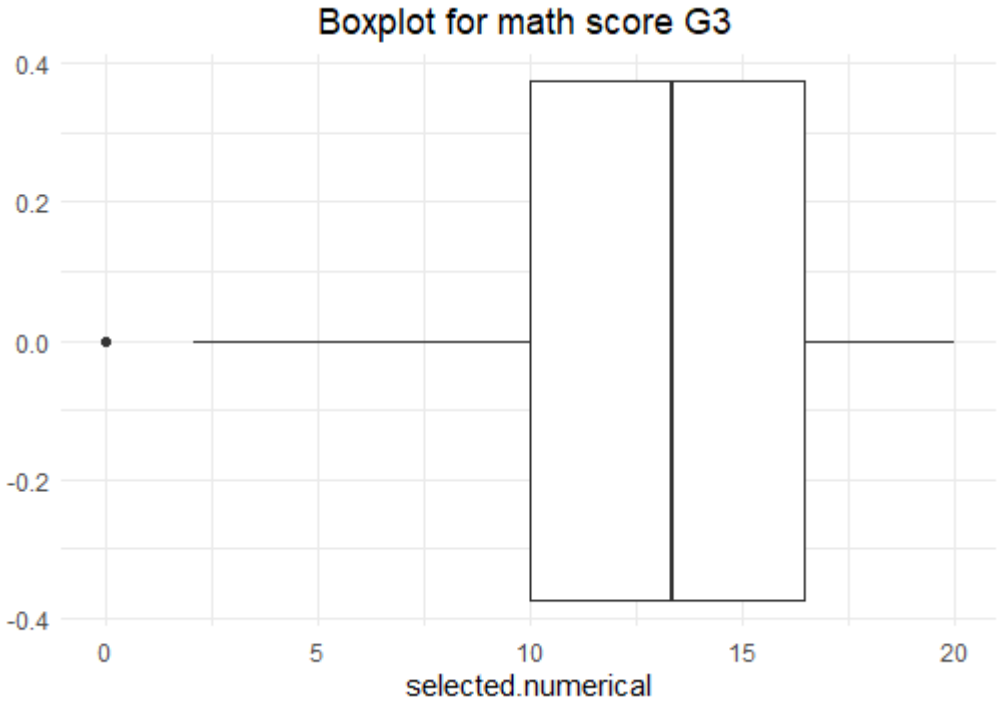


b. Based on histogram we can see that data is left-skewed and around mean, it has a quite alike behavior of Normal. Since around Grade=0 there is a look-alike local peak, we can say there is a possibility of existing outlier.
By QQ-plot we can determine howmuch data acts like normal, plot is shown below:



From QQ-plot we can wee data is left skewed and even possibility of existing outlier around 0.

c. Skewness is: -0.8343696 that is determining of being left-skewed.
d. There is one potentioal oulier in zero based on this box:

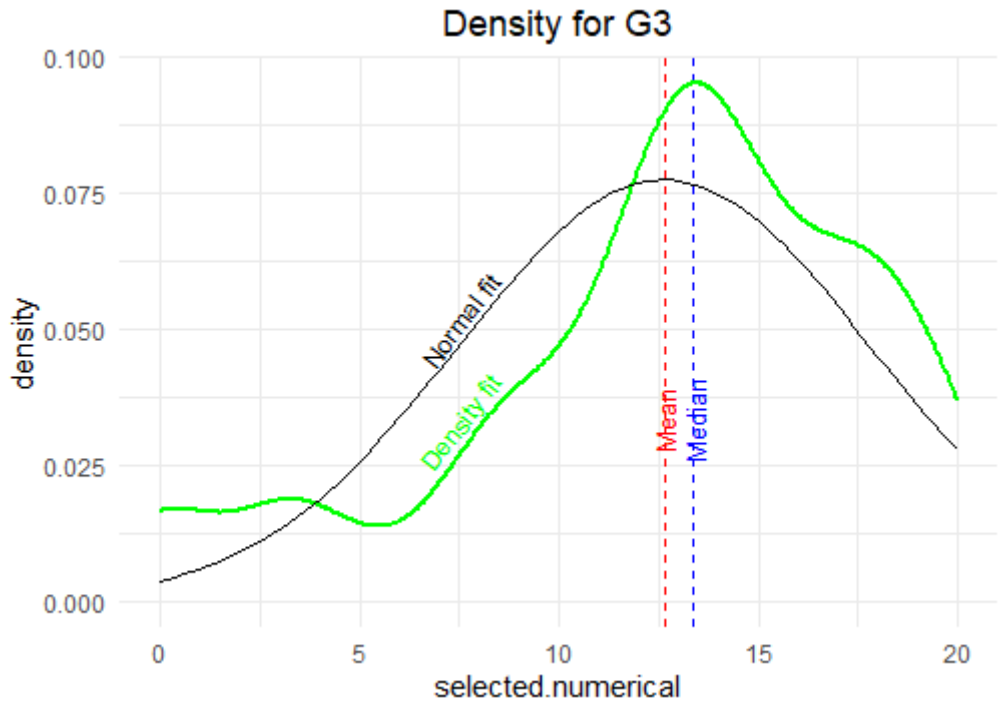


e. Calculated parameters are:

selected.numerical.median	13.3675797755583
selected.numerical.mu	12.6407603300732
selected.numerical.std	5.14487043731231
selected.numerical.var	26.4696918167302

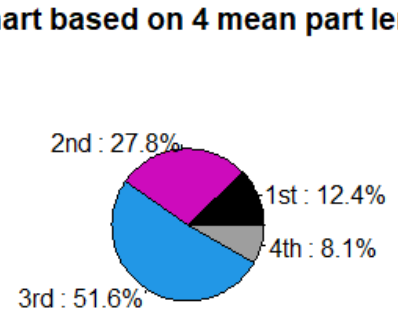
Since median is greater than mean it means data is left skewed. If data was normal, by varying 3*SE we can have our 99 percent data in range, But here if we add 3 sigma to mean we overcome 20 and it is out of range, so it is another definition of being left-skewed.

f. Plot:

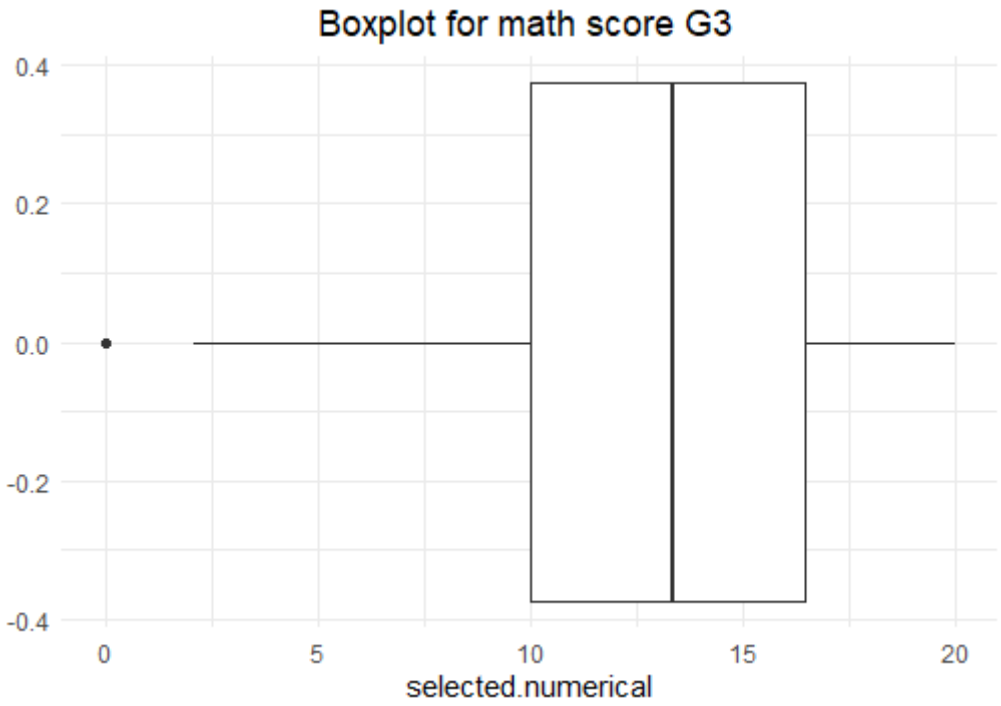


Based on median, 50% of scores in dataset is greater than that. And also mean is less than median so we are expecting to see instant drop in density when we are searching in scores less than median, and we expect to have barely drop in density when we search in greater than median.

g. For G3 grades based on 4 parts of mean we have:



h. Again by drawing boxplot we have:



And also stats about this boxplot is:

```
$stats
[1]  2.115863 10.000000 13.367580 16.473070 20.000000

$n
[1] 395

$conf
[1] 12.85298 13.88218

$out
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

There is an outlier in 0 and upper whisker is 20, lower whisker is 2.1158. 75th percentile is 16.47 and 25th percentile is 10. Median is 13.36, and confidence interval is 12.8529 to 13.8821.

Question 2

I chose sex.

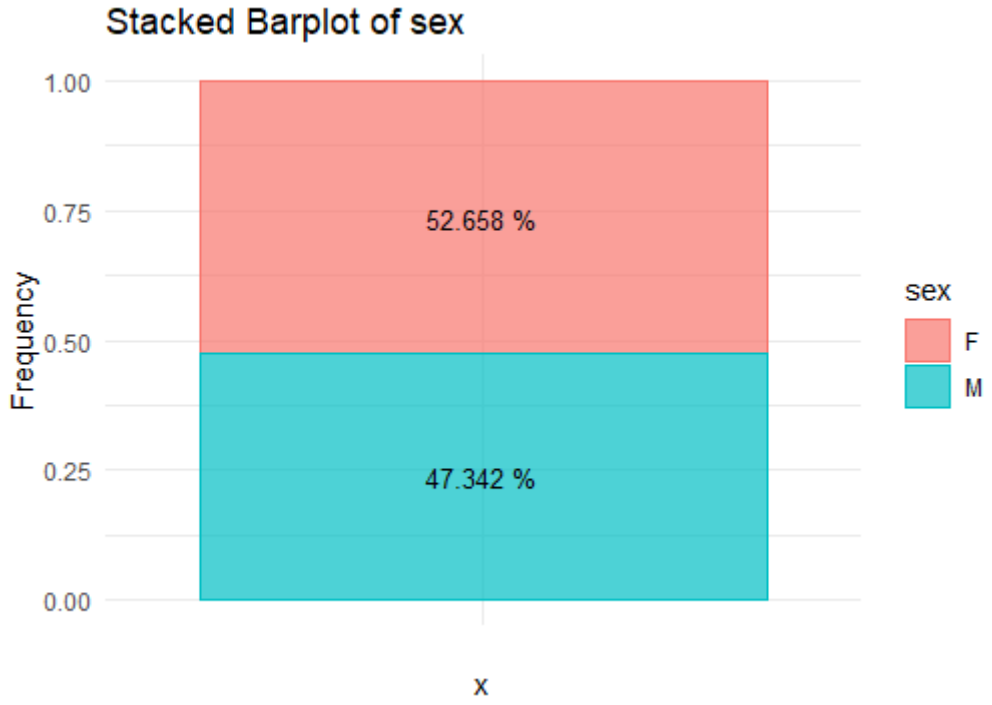
a. Frequencies are:

$$Male: 187, Female: 208$$

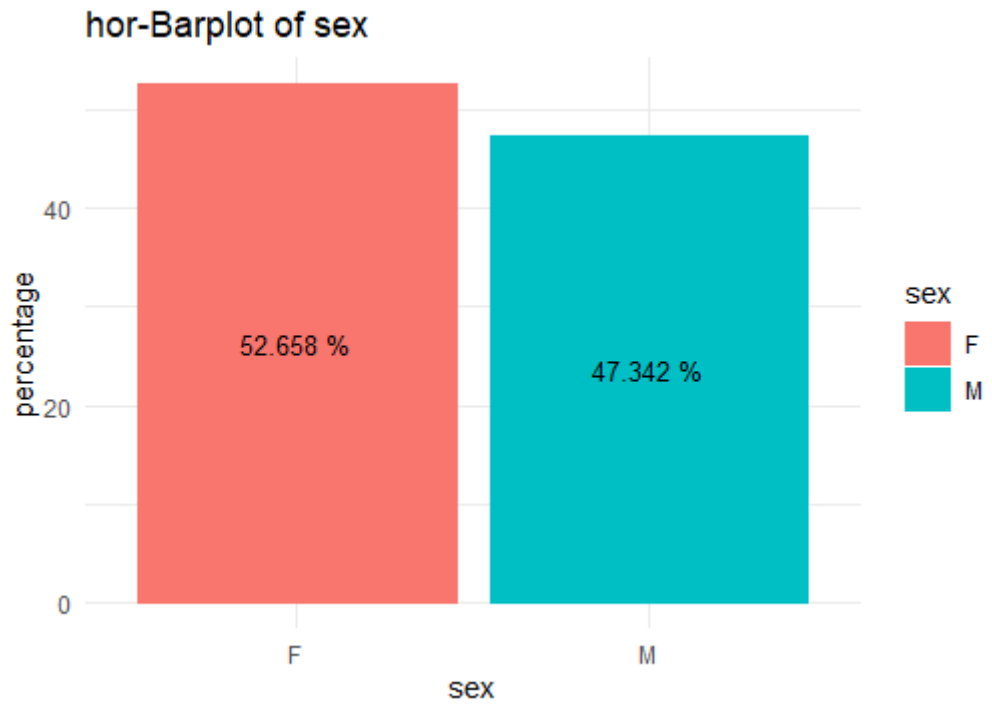
Percentage:

$$Male: 0.4734177, Female: 0.5265823$$

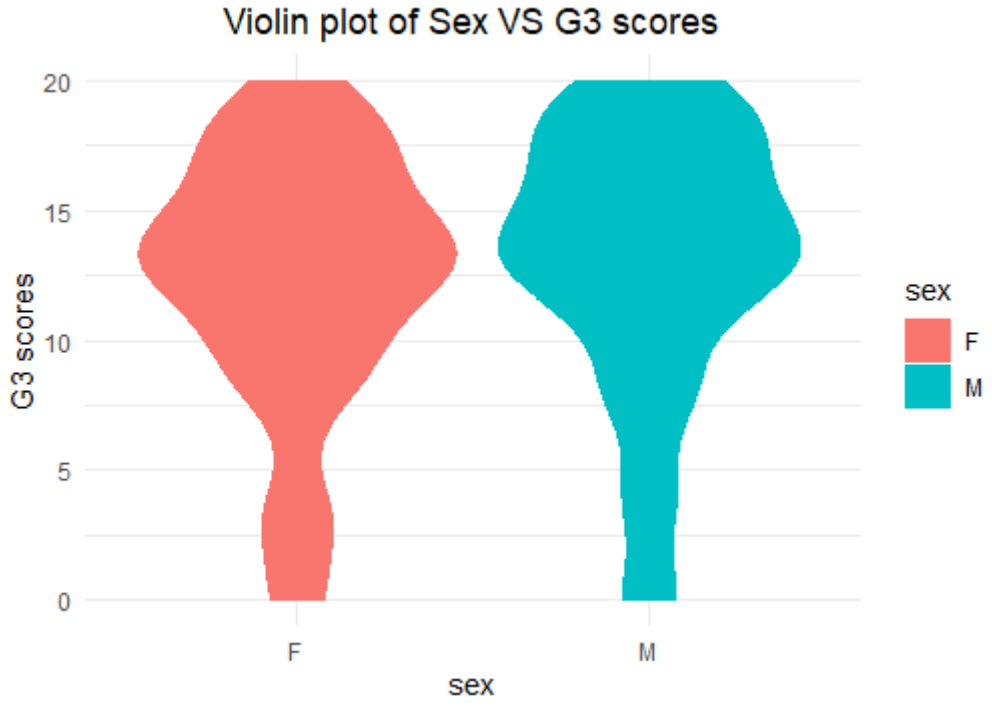
b. Plot:



c. Horizontal Bar-plot:



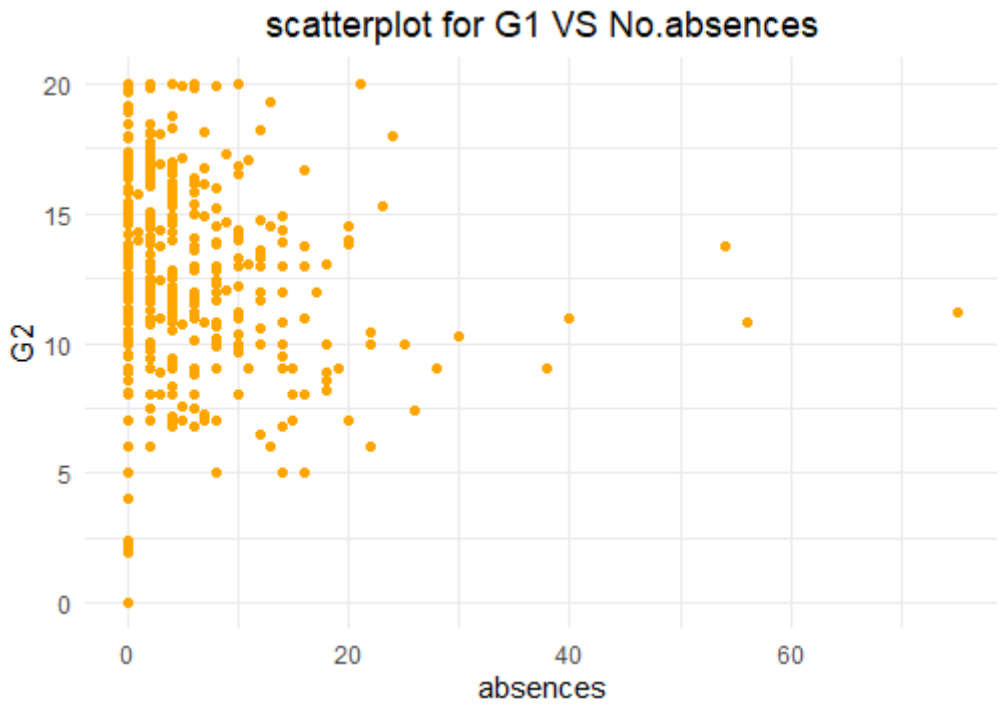
d. Violin plot:



Question 3

a. I get G2 grades with number of absences, I think there should be a negative relation between them. The higher the number of absences get, The lower the grades they get.

b. Scatter plot:



It seems there is a slightly negative relation between them.

c. Correlation coefficients is calculated with R.

$$Corr\ coeficient: -0.0574777$$

d. It looks like my guess was right. But for more statistical analysis I should check in later arts that if this correlation is enough to say they have negative relation or not.

e. With pearson method and in testing correlation and having alternative hypothesis of being corr<0.

$$H_0: corr < 0$$
$$H_A: corr \geq 0$$

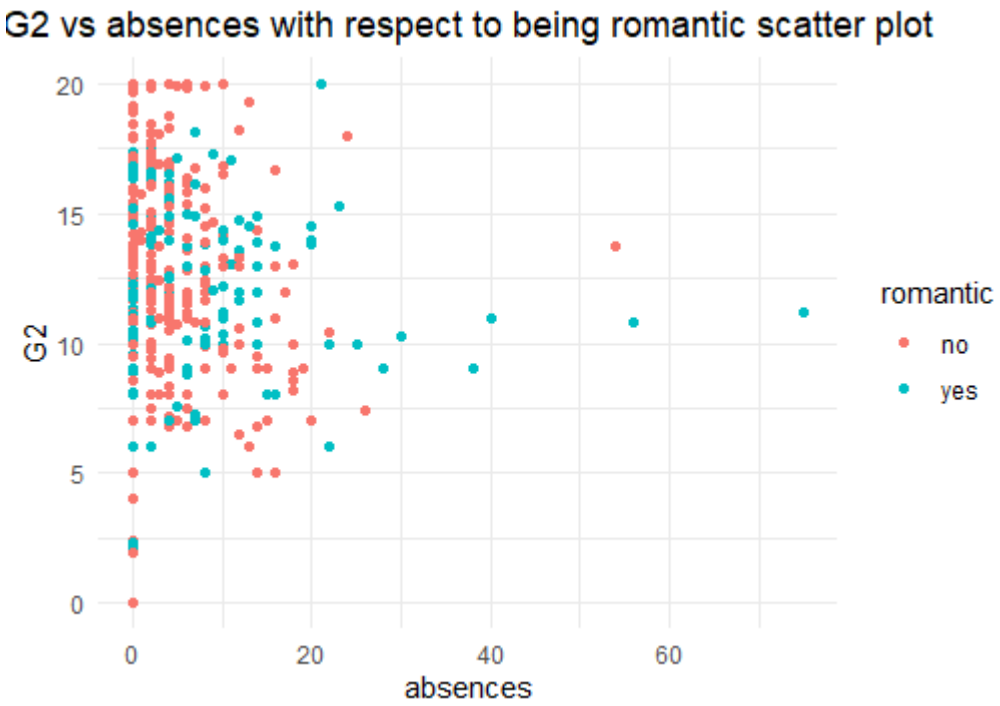
Result of test with parameters:

```
        Pearson's product-moment correlation

data:  numerical.first and numerical.second
t = -1.1413, df = 393, p-value = 0.1272
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.00000000  0.02553099
sample estimates:
       cor
-0.0574777
```

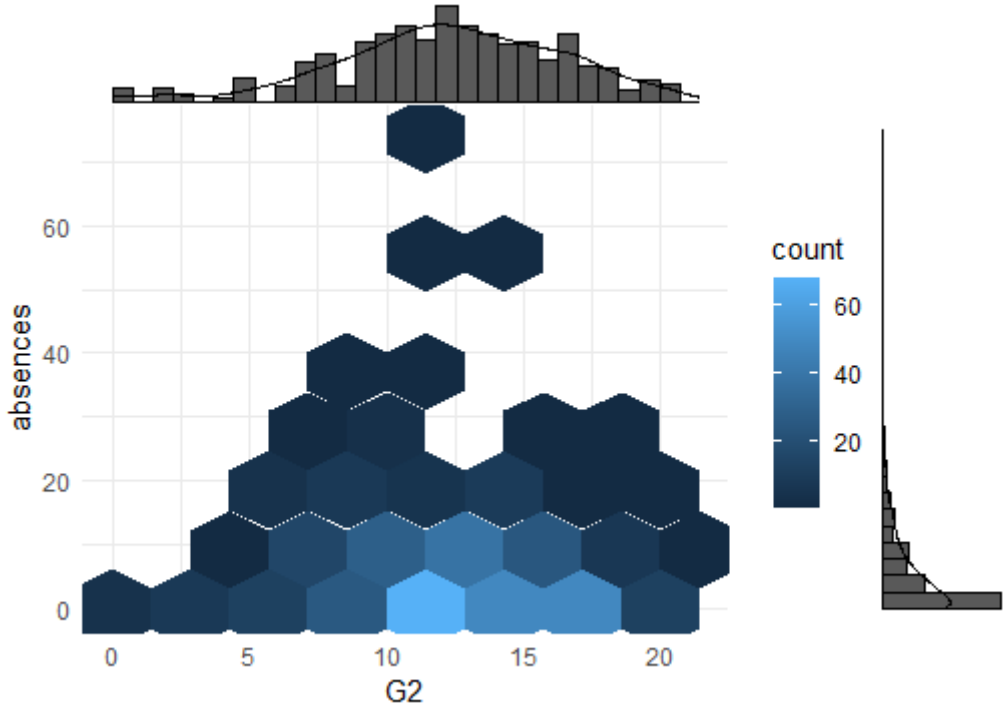
since correlation is within 95 percent confidence interval, so we can say statistically they have negative correlation. P-value shows that since it is more than significance level of 0.05, then we can say we can't reject null hypothesis.

f. Colored scatter plot:

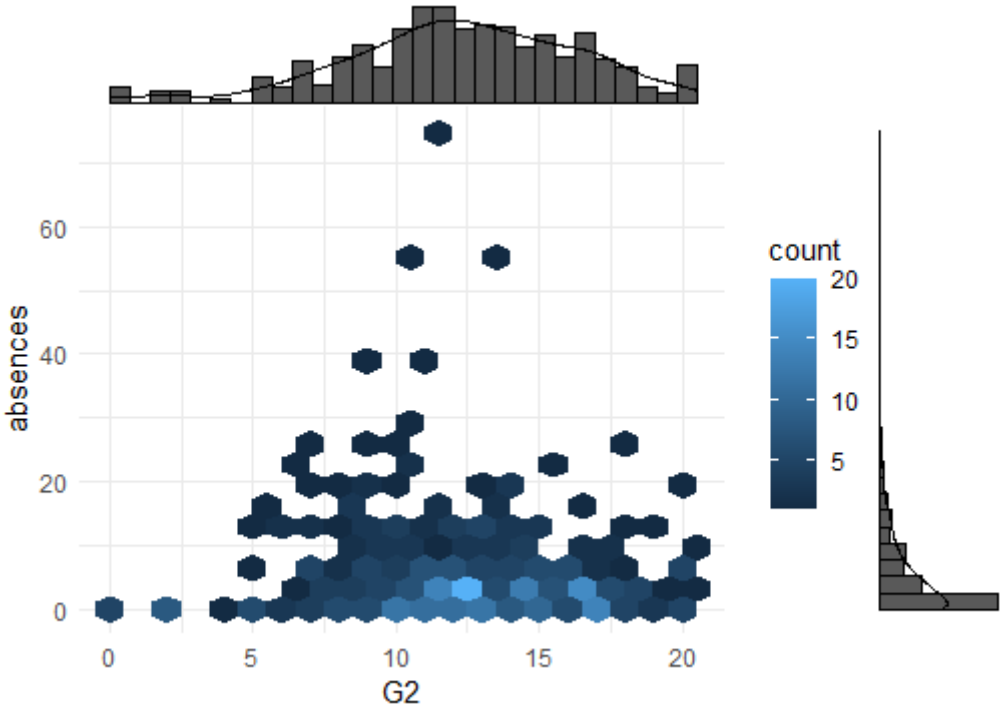


It's like students who are more romantic, do have more absences. And students who is less romantic, get better grades.

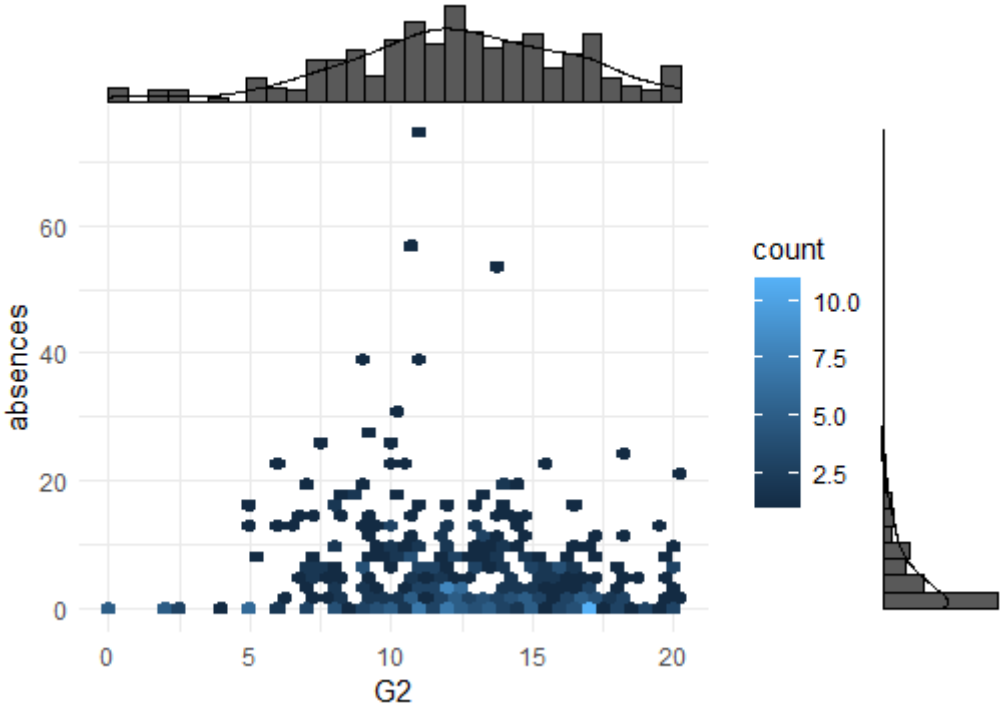
g. Plot with bin size = 7



Plot with bin size = 20



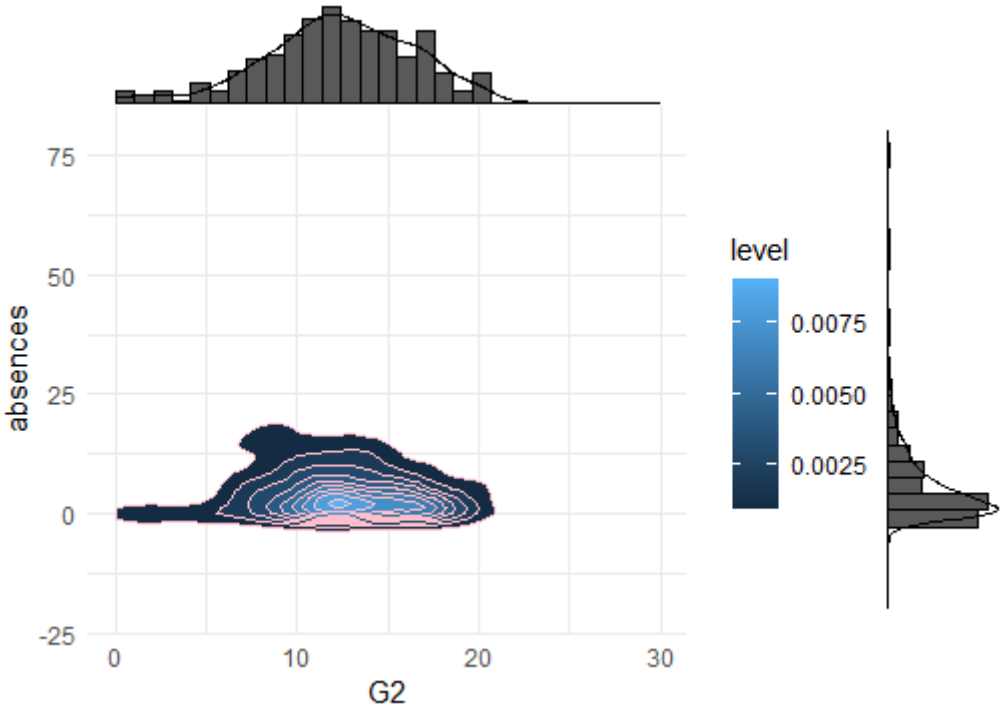
Plot with bin size=40



It is clear that for very large and very small bin sizes we lose our precision and for very small one, we have kindof generalization and on another hand for large ones. We have losing data and equally distributed hexbin plot that is not desired at all.

For plot with binsize=20 we have data's concetration around grade=15 and absences less than 5, and by getting closer to 20 we have less absences by the way we have kind of outliers around 20 that have more absences that could be explained by which they are so smart that don't need any class. And even getting in higher number of absences, tendency of getting high grades are lessen.

h. Density plot:

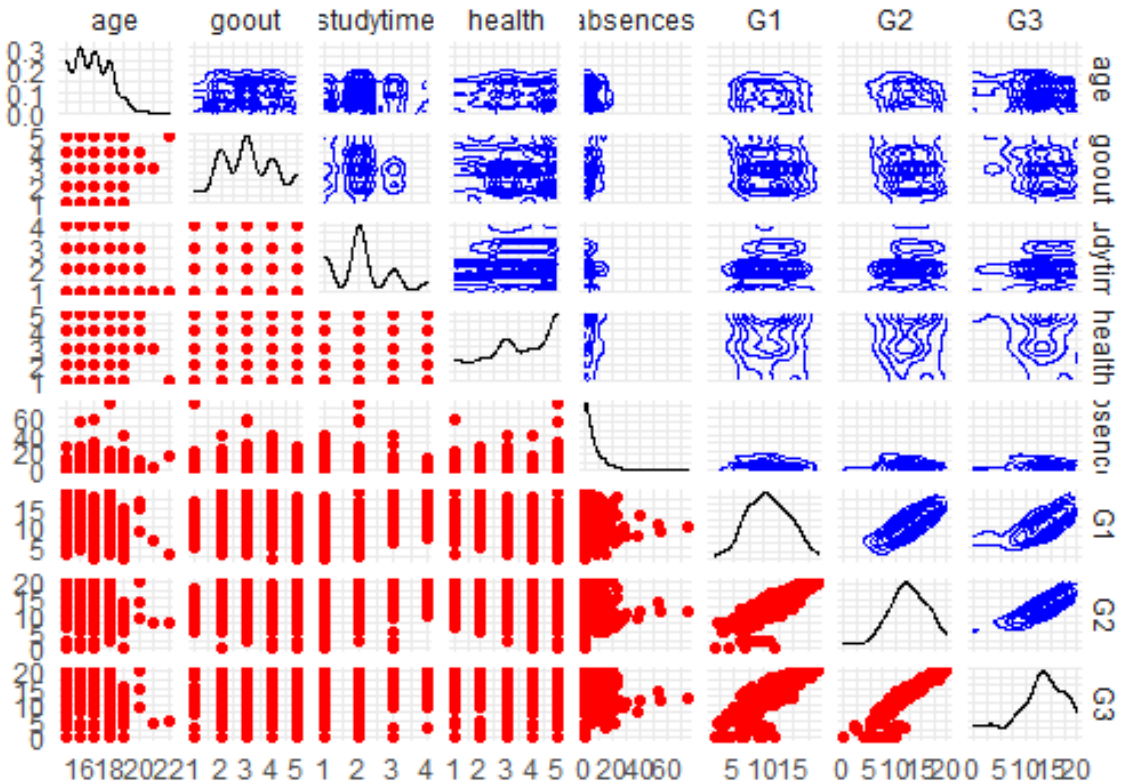


??

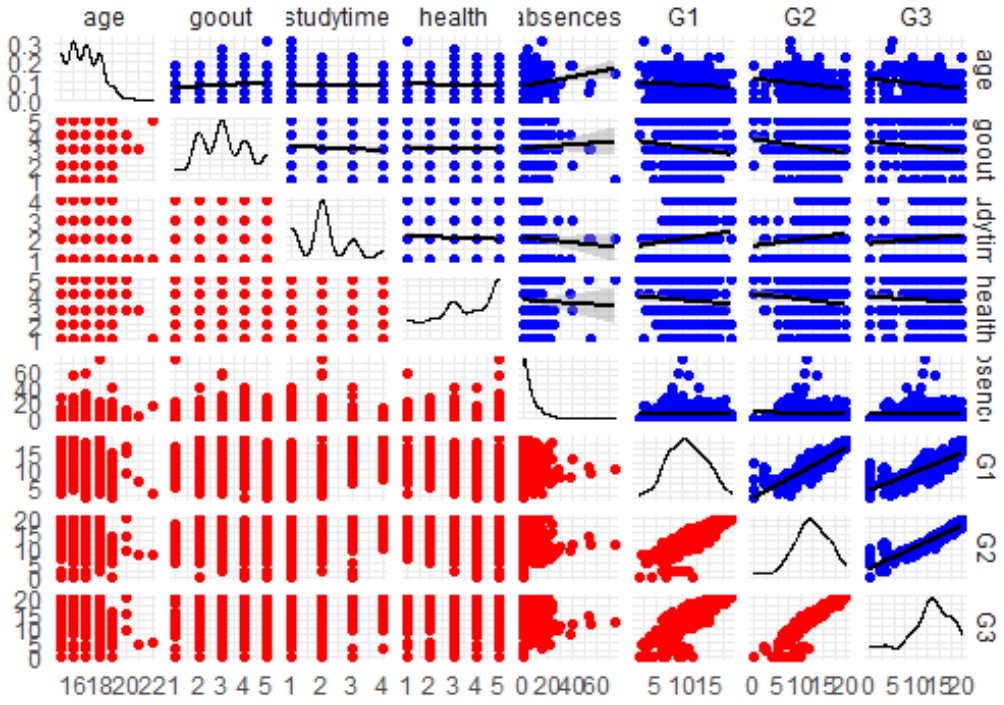
Question 4

a. Plot

Density and scatter, like what was in description:

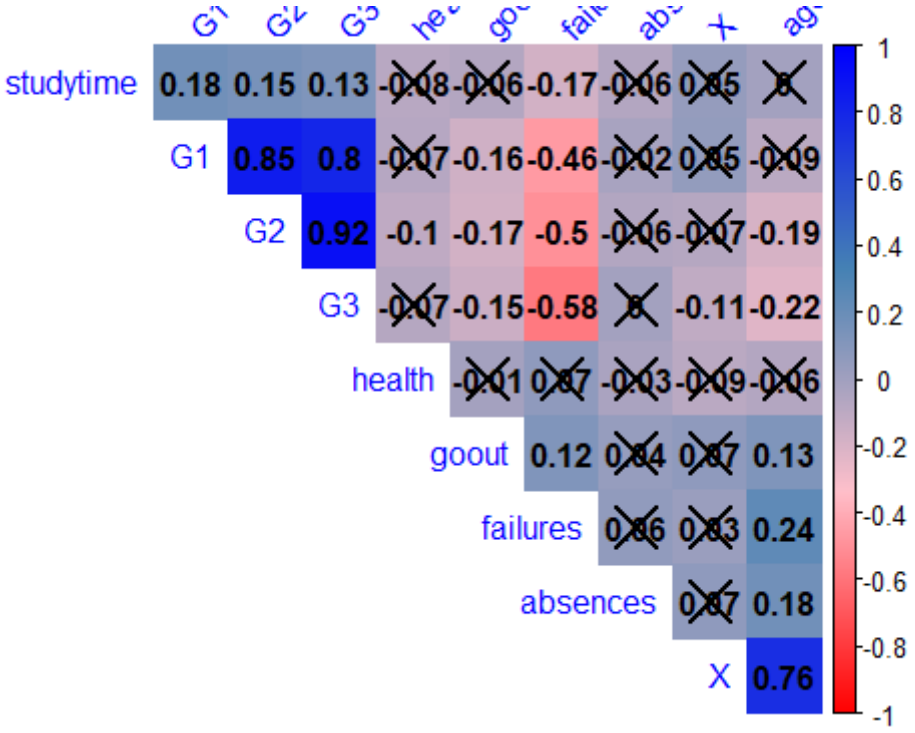


Scatter and linear relationship:

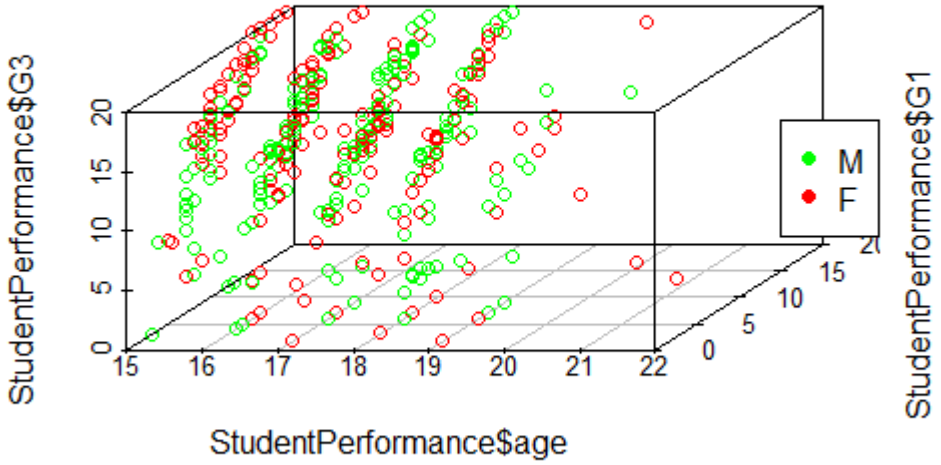


With respect to 2 above plots, I see noticeable realtion between G1 , G2, G3 grades with each other that are quietly normal. And they have mostly positive relation.

b. Plot:



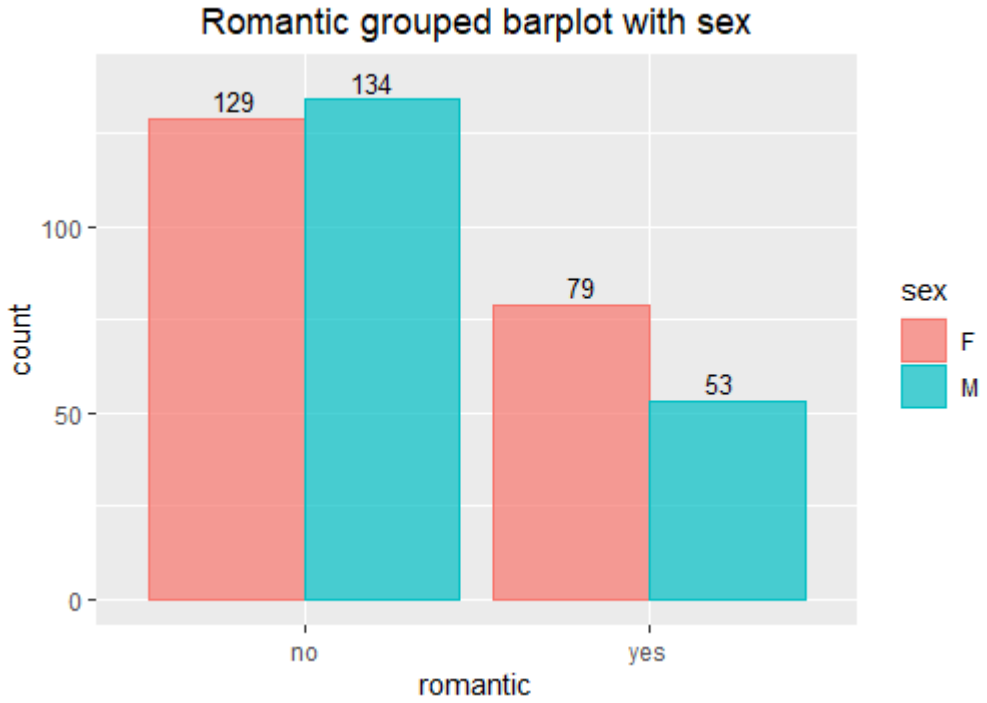
c. Plot:



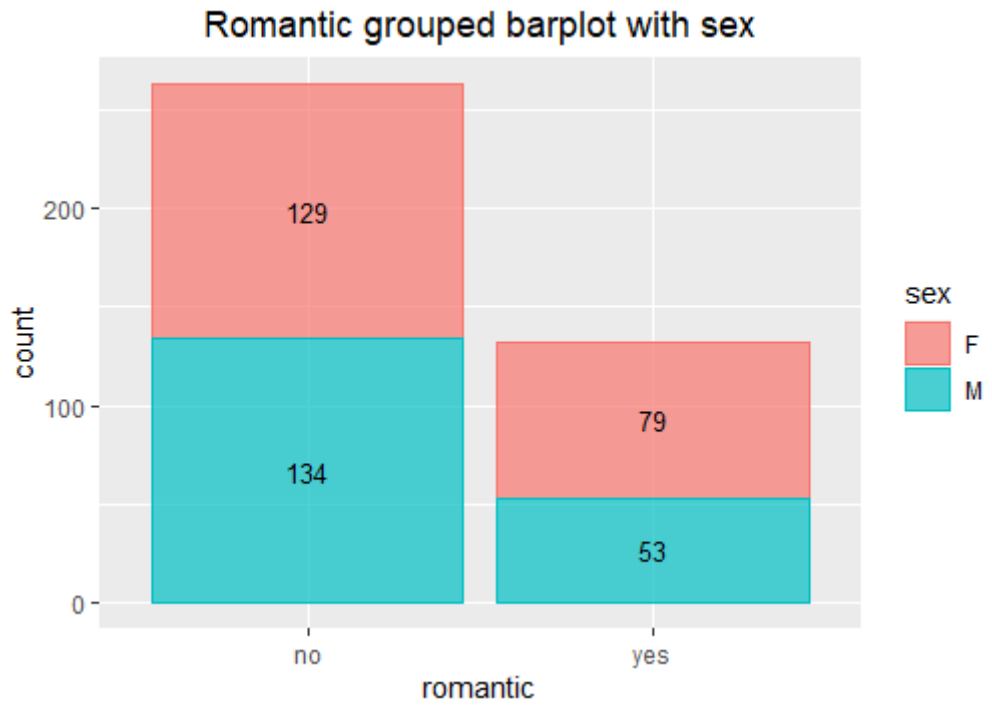
Question 5

a. Plot:

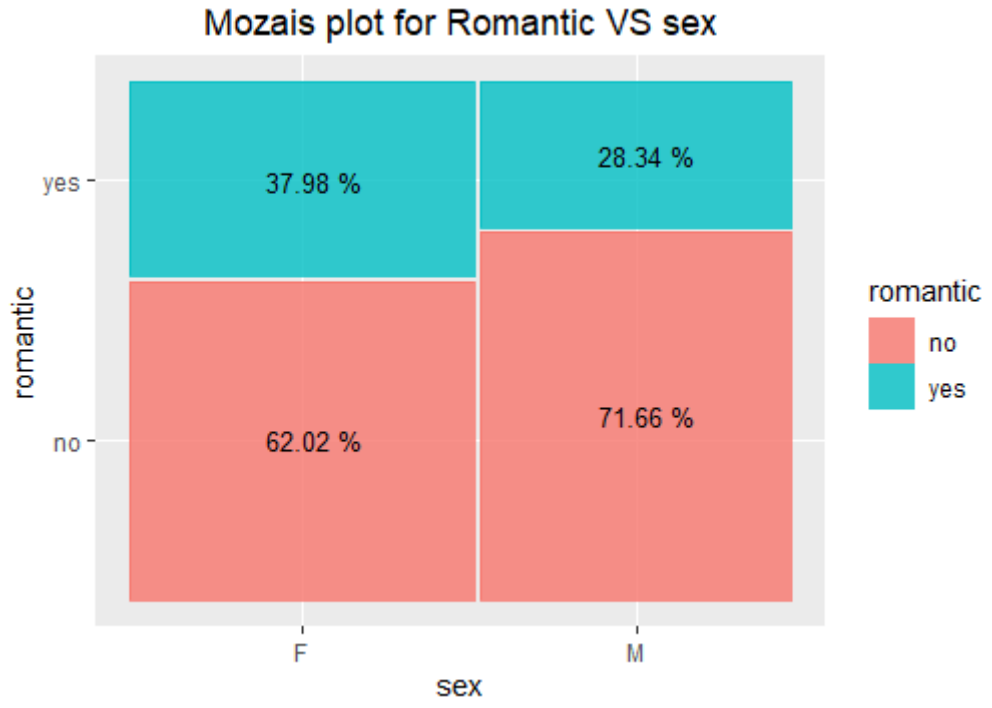
b. Plot:



c. Plot:



d. Plot:



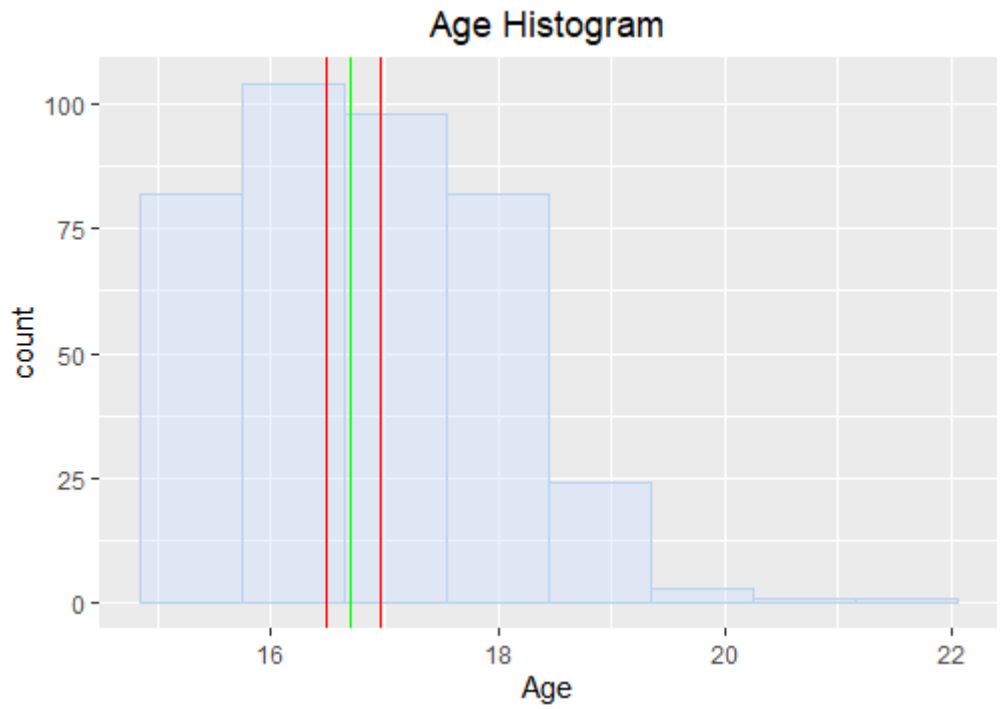
Question 6

a. Confidence interval for Age as numerical variable is:

$$Confidence\ interval: 16.36012\ to\ 16.81988$$

b. We are 95% sure that true mean of ages of school students will be in interval of 16.36012 to 16.81988.

c. Plot:



Red verticle lines are showing confidence interval and green one is mean of age.

d. Hypothesis is:

$$H_0: \ mean\ of\ age = 16$$
$$H_A: mean\ of\ age \neq 16$$

With R, p-value is calculated:

$$pValue \cong 1.984 \times 10^{-9}$$
$$since\ 0.05 > pValue \rightarrow Reject\ H_0$$

the **p-value is** the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis **is** correct. Since it is less than significance level, so we will reject null hypothesis.

e. Because H_0 that is mean is equal to 16, and not inside the interval of part 1, so it is clearly been rejected. Confidence interval and pValue result of rejecting hypothesis is equal.

f. Type II error has been calculated:

$$Type\ II\ error \cong 0.238131$$

It means probability of accepting null hypothesis, whereas it is false.

g. Power:

$$Power = 1 - type\ II\ error$$

So power is 0.76187 .

when effect size increases, the power increases

Question 7

A. Answer:
   a. Since we are not given variance and sample size is less than 30, I chose t-test with df = 25-1, ainces we have 25 samples.
   b. Hypothesis:

$$H_0: Mean_{G_2} = Mean_{G_3}$$
$$H_A: Mean_{G_2} \neq Mean_{G_3}$$

   Running t-test with this samples, in R, I got this p value:
   PValue= 0.37
   Since p_Value >significance level(0.05) so we can't reject Null hypothesis.

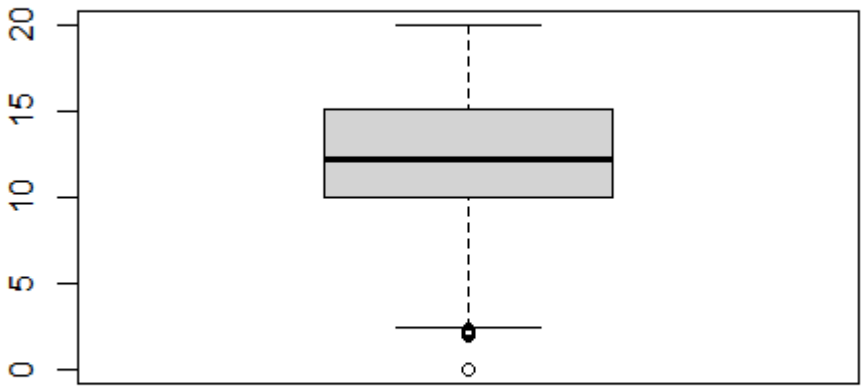
B. sampling data without replacing to have independent sampling and so because of having large number of samples that we can use z-test instead of t-test.
   Hypothesis:

$$H_0: Mean_{G_2} = Mean_{G_3}$$
$$H_A: Mean_{G_2} \neq Mean_{G_3}$$

   Running z-test with this samples, in R, I got this p value:
   PValue= 0.425574 . since we can not reject Null hypothesis (P_value >alpha), so Null hypothesis will be in 95 percent confidence interval.
   CI = -1.60 to 0.92

Question 8

   a. I choose G2 because I detected there was outlier based on boxplot:



I get 30 samples of size = 100 and calculate mean for every one of them then used quantile method to calculate 95% percentile. I didn't replace after sampling to maintain independency for our method of sampling

$$CI = 11.7\ to\ 12.8$$

b. After bootstrapping with 20 samples, I used both se method and quantile method to calculate interval, since 20 samples is less than 30 I used t score to fonde 95 confidence interval in t-dist with df=19. Final CI:

$$CI = 11.8\ to\ 12.6$$

c. In part a I used pure sampling with no replacing to maintain independency, in part b I used bootstrapping, since bootstrapping also replicates data and makes larger population than real society, so part b will give better confidence interval.

Question 9

There is 4 different failures that we can do ANOVA test for average of Grades with respect to type of failure. Assuming significant level of 0.05 and confidence interval of 95%. I run ANOVA with TukeyHSD package.

Hypothesis:

$$H_0: all\ cathegories\ have\ same\ mean$$

$$H_A: atleast\ one\ pair\ of\ cathegories\ have\ different\ mean$$

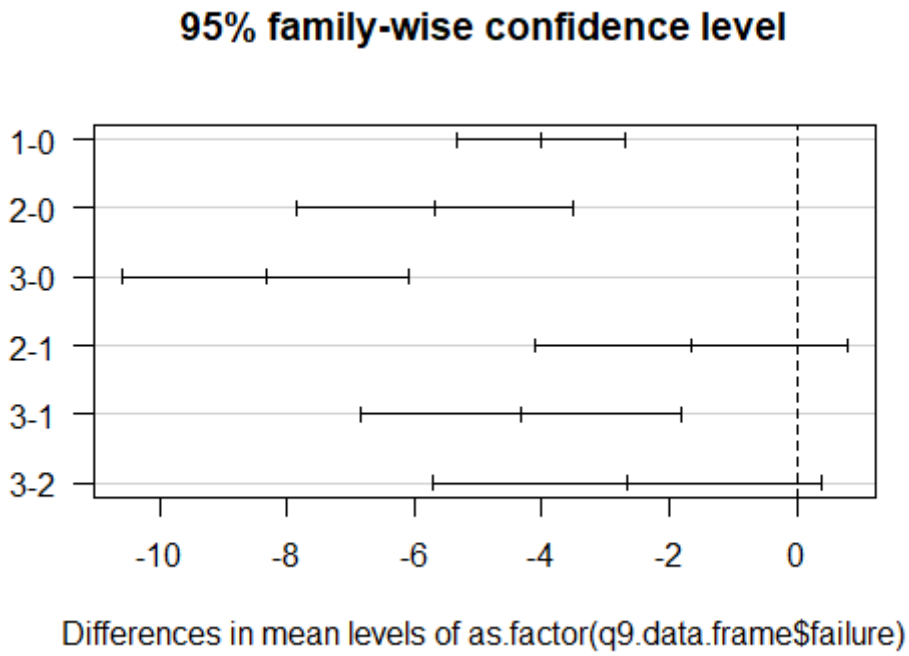
Results are shown above, and also plotting for confidence intervals, we get:

Summary of ANOVA:

Df	double [2]	3 391
Sum Sq	double [2]	1994 4464
Mean Sq	double [2]	664.8 11.4
F value	double [2]	58.2 NA
Pr(>F)	double [2]	3.9e-31 NA

Since Pr(>F) < 0.05 then we reject null hypothesis and conclude that there is at least two pairs with different means.

For Family-wise analysis, we have this plot:



For calculating improved significance level we have:

$$\alpha' = \frac{\alpha}{4 \times \frac{3}{2}} = \frac{\alpha}{6} = \frac{0.05}{6} \cong 0.0083$$

For better view on this family wise:

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

Fit: aov(formula = q9.data.frame$avg.grades ~ as.factor(q9.data.frame$failure))

$`as.factor(q9.data.frame$failure)`
         diff        lwr        upr     p adj
1-0 -4.008368  -5.336440 -2.6802956 0.0000000
2-0 -5.674642  -7.845973 -3.5033119 0.0000000
3-0 -8.337662 -10.572414 -6.1029093 0.0000000
2-1 -1.666275  -4.113972  0.7814227 0.2961363
3-1 -4.329294  -6.833423 -1.8251644 0.0000630
3-2 -2.663019  -5.699722  0.3736836 0.1086898
```

Last column shows pairwise p-values, for pairs 1-0,2-0,3-0,3-1 since their p-value is less than improved significance level, we reject them and they statistically have not equal means.

R CODES:

```
library(magrittr)
library(dplyr)
library(ggfortify)
library(ggplot2)
library(plyr)
library(gridExtra)
library(ggpubr)
require(qqplotr)
library("ggpubr")
library(moments)
library(hexbin)
library(ggmosaic)
theme_set(theme_minimal())


#Question 0
StudentPerformance <- read.csv("F:\\Semester 8\\Statistical
inference\\Project\\StudentsPerformance.csv", header = TRUE)
sum(is.na(StudentPerformance))
#.....


#Question 1
selected.numerical <- StudentPerformance$G3
##Part a
bw <- 2 * IQR(selected.numerical) / length(selected.numerical)^(1/3)
selected.numerical.hist <- ggplot(as.data.frame(selected.numerical),
                                  aes(selected.numerical)) +
  geom_histogram(aes(y=..density..) ,
                 binwidth = bw,
                 alpha = 0.4)  +
  geom_density(linetype="dashed",
```

```
            alpha = 0.3,
            size=1) +
  labs(title = "Histogram for G3",
       x = "Score",
       y="Density")+
  theme(plot.title = element_text(hjust = 0.5))
selected.numerical.hist

##part b
selected.numerical.qq <- ggplot(as.data.frame(selected.numerical),
                                aes(sample = selected.numerical))+
  geom_qq()+
  geom_qq_line()+
  labs(title="QQ-plot for G3")+
  theme(plot.title = element_text(hjust = 0.5))

selected.numerical.qq

##part c
print(skewness(selected.numerical))

##part d
selected.numerical.boxplot <- ggplot(as.data.frame(StudentPerformance),
                                     aes(x = selected.numerical))+
  geom_boxplot()+
  labs(title="Boxplot for math score G3 ",
       xlab="score")+
  theme(plot.title = element_text(hjust = 0.5))

selected.numerical.boxplot
```

```
##part e
selected.numerical.mu <- mean(selected.numerical)
selected.numerical.median <- median(selected.numerical)
selected.numerical.var <- var(selected.numerical)
selected.numerical.std <- sd(selected.numerical)

##part f
selected.numerical.density <- ggplot(StudentPerformance,
                aes(x = selected.numerical)) +
  geom_vline(xintercept = selected.numerical.mu,
            linetype="dashed",
            color = "red") +
  geom_vline(xintercept = selected.numerical.median,
            linetype="dashed",
            color = "blue") +
  geom_density(color = "green", size = 1)+
  stat_function(fun = dnorm, n = 101, args = list(mean = selected.numerical.mu,
                                    sd = selected.numerical.std))+
  annotate("text", x = 7.5 , label = "Normal fit", y = 0.052, size = 3.4, angle = 52) +
  annotate("text", x = 7.5 , label = "Density fit", y = 0.033, size = 3.4, angle = 52,
color="green") +
  annotate("text", x = selected.numerical.mu , label = "Mean", y = 0.033, size = 3.4, angle =
90, color="red") +
  annotate("text", x = selected.numerical.median , label = "Median", y = 0.033, size = 3.4,
angle = 90, color="Blue") +
  labs(title="Density for G3")+
  theme(plot.title = element_text(hjust = 0.5))

selected.numerical.density
```

```
##Part G
G3 <- StudentPerformance$G3
mu <-mean(G3)
Partial_G3 <- c(length(G3[G3<=0.5*mu]),
                length(G3[G3>0.5*mu & G3<=mu]),
                length(G3[G3>mu & G3<=1.5*mu]),
                length(G3[G3>1.5*mu & G3<=2*mu]))
percentage <- round(100*Partial_G3/sum(Partial_G3), 1)
labels <- c("1st","2nd","3rd","4th")
pie(Partial_G3, labels=paste(paste0(labels," : ", percentage, "%"), sep=" "), col = Partial_G3)
title("Pie chart based on 4 mean part length")
##part H
boxplot.stats(selected.numerical)
selected.numerical.boxplot



#-------------------------------------------------------------------------------
#Question 2
library(GGally)
Male <- dplyr::filter(StudentPerformance, sex=="M")
female <- dplyr::filter(StudentPerformance, sex=="F")
student.sex <- StudentPerformance$sex
##Part a
length(Male$sex)
length(female$sex)
Male.percentage <- length(Male$sex)/length(student.sex)
female.percentage <- length(female$sex)/length(student.sex)
Male.percentage
female.percentage
```

```
##Part b
data <- data.frame(
  sex = c("M", "F"),
  Value = c(Male.percentage*100, female.percentage*100)
)

sex.barplot <- ggplot(as.data.frame(StudentPerformance), aes(x = " ", color = sex, fill = sex)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), alpha = 0.7) +
  labs(title="Stacked Barplot of sex", y = 'Frequency')+
  annotate("text", x = 1 , label = paste(toString(round(female.percentage*100, digit=3)),"%"), y = 1-female.percentage/2 , size = 3.4)+
  annotate("text", x = 1 , label = paste(toString(round(Male.percentage*100, digit=3)),"%"), y = Male.percentage/2 , size = 3.4)

##Part c
sex.hbarplot <- ggplot(data, aes(x =sex, y=Value, color=sex, fill=sex)) +
  geom_bar(stat="identity") +
  labs(title="hor-Barplot of sex", y = 'Frequency')+
  annotate("text", x = 1 , label = paste(toString(round(female.percentage*100, digit=3)),"%"), y = female.percentage*50 , size = 3.4)+
  annotate("text", x = 2 , label = paste(toString(round(Male.percentage*100, digit=3)),"%"), y = Male.percentage*50 , size = 3.4) +
  xlab("sex")+
  ylab("percentage")

sex.barplot
sex.hbarplot

##Part d
temp <- as.data.frame(StudentPerformance)
```

```
q2.violin <- ggplot(temp, aes(x=sex, y=G3, color= sex, fill= sex)) +
  geom_violin()+
  ylab("G3 scores")+
  labs(title="Violin plot of Sex VS G3 scores")+
  theme(plot.title = element_text(hjust = 0.5))

q2.violin
#.....

#Question 3
numerical.first <- StudentPerformance$absences
numerical.second <- StudentPerformance$G2

##Part b
scatter.q3 <- ggplot(temp, aes(x= absences, y=G2))+
  geom_point(color="orange")+
  labs(title="scatterplot for G1 VS No.absences")+
  theme(plot.title = element_text(hjust = 0.5))
scatter.q3

##Part c
corr.q3 <- cor(numerical.first, numerical.second)
corr.q3



##part E
corr.test.q3 <- cor.test(numerical.first, numerical.second,
        alternative = "less",
        method = "pearson",
        conf.level = 0.95)
```

```
corr.test.q3


##Part F

twonum.and1cat.scatter <- ggplot(temp, aes(x= absences, y=G2, color= romantic, fill= romantic))+

  geom_point()+

  labs(title = "G2 vs absences with respect to being romantic scatter plot")+

  theme(plot.title = element_text(hjust = 0.5))


twonum.and1cat.scatter


##Part G

library(ggExtra)


q3.densigram.hex <- ggMarginal(ggplot(temp, aes(x = G2, y = absences)) + geom_point(col="transparent")+geom_hex(bins=20), type= "densigram", margins = "both")

q3.densigram.hex

#Part H

q3.twoddenisty.hex <- ggMarginal(ggplot(temp, aes(x = G2, y = absences)) +ylim(c(-20,80))+xlim(c(0,30))+  geom_point(col="transparent")+stat_density2d(aes(fill=..level..), geom="polygon", color="pink"), type= "densigram", margins = "both")

q3.twoddenisty.hex


#.....


#Question 4

#Part a

library(GGally)

featurePlot(x=temp[,1:8], y=temp[,8:16], plot="pairs")

ggpairs(dplyr::select_if(StudentPerformance, is.numeric), title = "Correlogram")
```

```
list.of.num <- c(4, 7, 10, 12, 13, 14, 15, 16)

#density, without failure
ggpairs(StudentPerformance[, list.of.num],
        upper = list(continuous = wrap("density", colour="blue")),
        lower = list(continuous = wrap("points", colour="red")))

#linear relationship
ggpairs(StudentPerformance[, list.of.num],
        upper = list(continuous = wrap("smooth", colour="blue")),
        lower = list(continuous = wrap("points", colour="red")))

#----
#b.
library(Hmisc)
col <- colorRampPalette(c("red", "pink", "steelblue", "blue"))
sc <- rcorr(as.matrix(dplyr::select_if(StudentPerformance, is.numeric)))
poi <- sc$P
poi[is.na(poi)] <- 1

M <- cor(dplyr::select_if(StudentPerformance, is.numeric))
library(corrplot)
corrplot(M, method = "color", col = col(400), type = "upper", order = "hclust", addCoef.col =
"black",
         tl.col = "blue", tl.srt = 45, p.mat = poi, sig.level = 0.05, diag = FALSE)

#c.



cols <- c("Green", "red")
```

```
cols <- cols[as.numeric(as.factor(StudentPerformance$sex))]
library(scatterplot3d)
scatterplot3d(StudentPerformance$age,
              StudentPerformance$G1,
              StudentPerformance$G3, color = cols)

legend("right", legend = c("M","F"),
       col =  c("green", "red"), pch = 16)

#.....
#Question 5
##Part a
q5.cat1 <- StudentPerformance$romantic
q5.cat2 <- StudentPerformance$sex
romantic <- StudentPerformance$romantic
sex <- StudentPerformance$sex
table <- table(romantic, q5.cat2)
print.table(table)
head(data.frame(table))
##Part b
combined.barplot.RS <- ggplot(temp, aes(x = romantic,color = sex, fill = sex)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title="Romantic grouped barplot with sex", x="romantic")+
  annotate("text", x = 1.2 , label = table[1,2], y = table[1,2]+5 , size = 3.4) +
  annotate("text", x = 0.8 , label = table[1,1], y = table[1,1]+5 , size = 3.4) +
  annotate("text", x = 2.2 , label = table[2,2], y = table[2,2]+5 , size = 3.4) +
  annotate("text", x = 1.8 , label = table[2,1], y = table[2,1]+5 , size = 3.4) +
  theme(plot.title = element_text(hjust = 0.5))

combined.barplot.RS
```

```
##Part c
combined.segbarplot.RS <- ggplot(temp, aes(x = romantic,color = sex, fill = sex)) +
  geom_bar(alpha = 0.7) +
  annotate("text", x = 1 , label = table[1,2], y = table[1,2]/2 , size = 3.4) +
  annotate("text", x = 1 , label = table[1,1], y = table[1,2]+table[1,1]/2 , size = 3.4) +
  annotate("text", x = 2 , label = table[2,2], y = table[2,2]/2 , size = 3.4) +
  annotate("text", x = 2 , label = table[2,1], y = table[2,2]+table[2,1]/2 , size = 3.4) +
  labs(title="Romantic grouped barplot with sex", x="romantic")+
  theme(plot.title = element_text(hjust = 0.5))

combined.segbarplot.RS

##Part d
library(ggmosaic)
mos <- data.frame(
  sex=temp$sex,
  romantic=temp$romantic
)
mosaic.plot.RS <-  ggplot(data = mos) +
  geom_mosaic(aes( x = product(romantic, sex), fill=romantic))+
  annotate("text", x = 0.78 , label =
paste(toString(round(table[1,2]/sum(table[,2]),4)*100),"%"), y = table[1,2]/sum(table[,2])/2 ,
size = 3.4)+
  annotate("text", x = 0.78 , label =
paste(toString(round(table[2,2]/sum(table[,2]),4)*100),"%"), y = 1-table[2,2]/sum(table[,2])/2
, size = 3.4)+
  annotate("text", x = 0.28 , label =
paste(toString(round(table[1,1]/sum(table[,1]),4)*100),"%"), y = table[1,1]/sum(table[,1])/2 ,
size = 3.4)+
  annotate("text", x = 0.28 , label =
paste(toString(round(table[2,1]/sum(table[,1]),4)*100),"%"), y = 1-table[2,1]/sum(table[,1])/2
, size = 3.4)+
```

```
  labs(title="Mozais plot for Romantic VS sex")+
  theme(plot.title = element_text(hjust = 0.5))


mosaic.plot.RS
#.....................................................................
#Question 6
calculate_ci <- function(sampled_data, confidence_level) {
  sample_mean <- mean(sampled_data)
  stdDev <- sd(sampled_data)

  z_value <- qnorm((1 + confidence_level)/2)
  stdError <- stdDev / sqrt(length(sampled_data))
  CI <- c(sample_mean - z_value * stdError, sample_mean + z_value * stdError)
  return(CI)
}
#Part a
sampled_data <- sample(StudentPerformance$age, 100)
age.CI <- calculate_ci(sampled_data, 0.95)

#Part c
selected.numerical <- StudentPerformance$age
bwidth <- 2 * IQR(selected.numerical) / length(selected.numerical)^(1/3)
q6.age.hist <- ggplot(StudentPerformance, aes(x = age)) +
  geom_histogram(binwidth = 0.9, alpha = 0.4, color="lightsteelblue2", fill="lightsteelblue1")
+
  labs(title = "Age Histogram", x = "Age") +
  geom_vline(xintercept =  mean(StudentPerformance$age), color = "Green") +
  geom_vline(xintercept =  age.CI[1], color = "red") +
  geom_vline(xintercept =  age.CI[2], color = "red")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
q6.age.hist


#Part d
zdist.2tail.meantest <- function(sampled_data, null_value, alpha) {
  n <- length(sampled_data)
  x_bar <- mean(sampled_data)
  S <- sd(sampled_data)
  z_score <- abs((x_bar - null_value)) / (S/sqrt(n))
  p_value <- pnorm(z_score, lower.tail = FALSE)
  print(paste("p-value =", p_value))
}
zdist.2tail.meantest(sampled_data, 16 , 0.05)


#Part f
null.value <- 16
z_value <- abs(qnorm((1-0.05)/2))
errorTypeII <- pnorm(null.value + z_value * sd(sampled_data)/sqrt(length(sampled_data)) -
mean(sampled_data))


#part g
power <- 1 - errorTypeII


#.....
#Question 7
##Part a
StudentsPerformance.sampled <- sample_n(StudentPerformance, 25)
x_bar <- mean(StudentsPerformance.sampled$G2) -
mean(StudentsPerformance.sampled$G3)
```

```
s1 <- sd(StudentsPerformance.sampled$G2)
s2 <- sd(StudentsPerformance.sampled$G3)
s <- abs((x_bar - 0)) / (sqrt((s1^2/25) + (s2^2/25)))
pvalue <- 2*pt(s, df = 24, lower.tail = FALSE)
print(paste("p-value =", pvalue))



##Part b
G2.sample <- sample(StudentPerformance$G2, 100)
G3.sample <- sample(StudentPerformance$G3, 100)
zdist.2tail.meantest(G2.sample - G3.sample, 0, 0.05)

#.....................................................................
#Question 8
library(bootstrap)
##Part a
q8.chosen.numerical <- StudentPerformance$G2
boxplot(q8.chosen.numerical)

mean.CI <- c()
for(i in 1:30){
  bootsamp <- sample(q8.chosen.numerical, 100)
  mean.CI <- c(mean.CI, mean(bootsamp))
}
q8.mean.Pa.CI <- quantile(mean.CI, c(0.025,0.975))

##Part b
get_mean <- function(x){
  mean(x)
}
```

```
boot.q8.mean <- bootstrap(x=q8.chosen.numerical, nboot=20, get_mean)
temp <- boot.q8.mean$thetastar
se <- sd(temp)
mu <- mean(temp)
t_s <- qt(0.975, df=19)
q8.mean.Pb.CI <- c(mu-t_s*se, mu+t_s*se)
q8.mean.Pb.CI <- quantile(boot.q8.mean$thetastar, c(0.025,0.975))

#.....................................................................
#Question 9
##Part a
average.grades <- (StudentPerformance$G1 + StudentPerformance$G2 +
StudentPerformance$G3)/3
q9.data.frame <- data.frame(
  avg.grades = average.grades,
  failure = StudentPerformance$failures
)
q9.scatterplot <-  ggplot(q9.data.frame, aes(x = avg.grades, y= failure))+geom_point()
q9.scatterplot

avg.f0 <- dplyr::filter(q9.data.frame, failure==0)
avg.f1 <- dplyr::filter(q9.data.frame, failure==1)
avg.f2 <- dplyr::filter(q9.data.frame, failure==2)
avg.f3 <- dplyr::filter(q9.data.frame, failure==3)
ANOVA.avg_failure <- aov(q9.data.frame$avg.grades~as.factor(q9.data.frame$failure))
summary.ANOVA <- summary(ANOVA.avg_failure)
avg_failure.tukeyHSD <- TukeyHSD(ANOVA.avg_failure,
                    ordered = FALSE,
                    conf.level = 0.95)
```

```
avg_failure.tukeyHSD
plot(avg_failure.tukeyHSD, las = 1)
```