Project Phase 1 Report

Erfan Mirhaji

ID : 810196568

Question 0 :

a) The dataset is about university admissions . In this dataset , multiple variables and feature are included regarding students who have submitted an admission for a specific university . These variables are of different types , some are the score of each individual in a specific exam , some are based on previous researches done or the completion of internship whether done abroad or not .
    a. Analyzing this dataset will give us insights about certain traits and abilities needed for a volunteer to maximize their chances of admission . It also shows to some extent how each person will likely perform in a certain exam based on other data available .

b) In this dataset , we have 10 features

c) Yes , some values are missing from the dataset , for example , some participants do not have any CGPA score assigned to them . For these N/A values , we have a few options , some of which are :
    1. Cleaning the row which has missing values and analyzing the remaining rows
    2. Changing the N/A values with the mean or median value of the corresponding feature for all the N/A values .

   Because the CGPA score is one of the most important variables , I deleted all the row with missing CGPA values .

   Also , I deleted the A & B columns , as they provided no extra information about the dataset and only kept the Serial No. column for index reference .

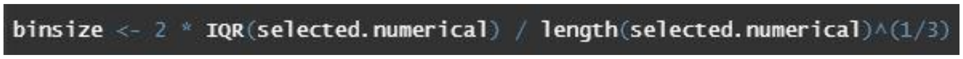
d) Based on an elementary view of the dataset , we will guess the following features to be the most relevant and contain the most information about the case :
    1. University rating
    2. CGPA
    3. Research

   The reason is :

    1. The chance of admission seems to increase as these values increase ; not all the time but for a majority of the cases .
    2. Other features seem to follow these qualities , for example if these 3 variables are considered “good” for a person , other features will generally be “good” too .

Question 1 :

a) Bin-size is calculated in the code and is equal to : 0.114393



```
binsize <- 2 * IQR(selected.numerical) / length(selected.numerical)^(1/3)
```

Plot of histogram and density curve is as follows :

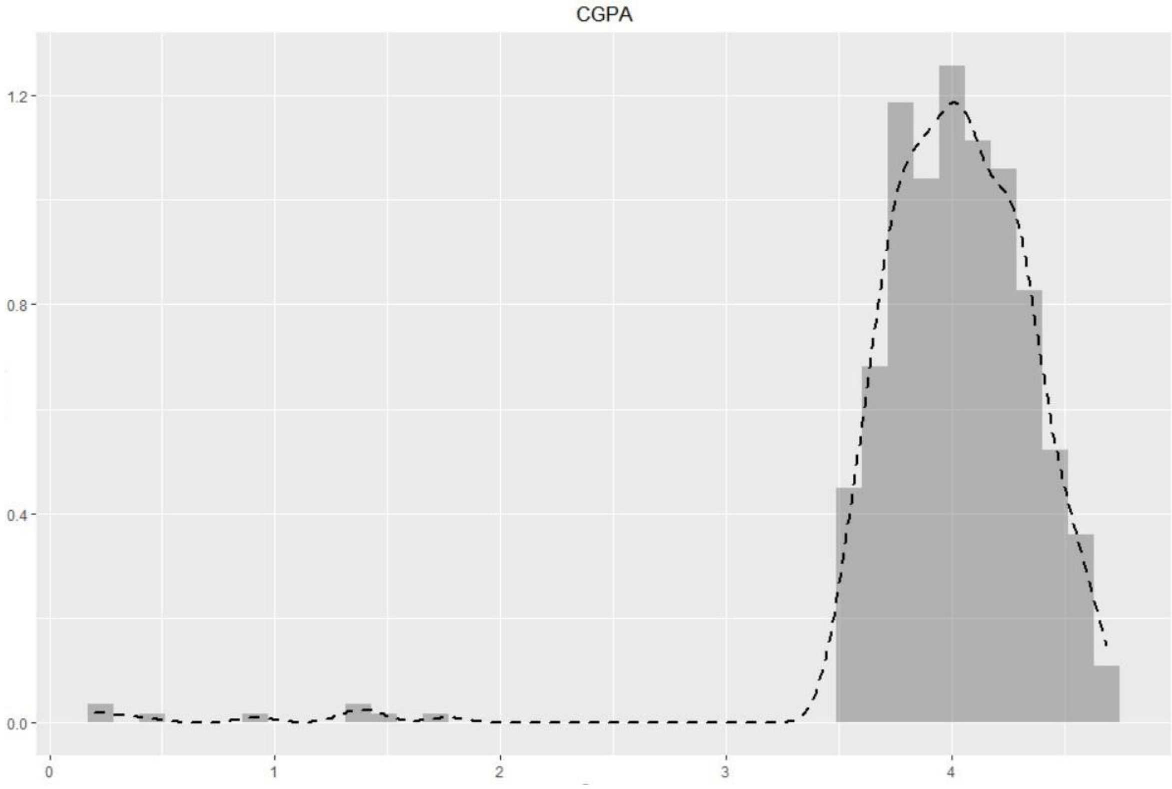


Figure 1 : CGPA histogram and density curve , density in vertical axis and CGPA score in horizontal

b) According to figure 1 , the data is a bit left-skewed with thick tails . The shape of density curve , especially around it's mean , is very similar to that of a normal distribution . There are some local peaks between 0 and 2 , so there is a possibility of existing outliers We can use QQ-Plot to more accurately compare the behavior of the density curve to a normal distribution :

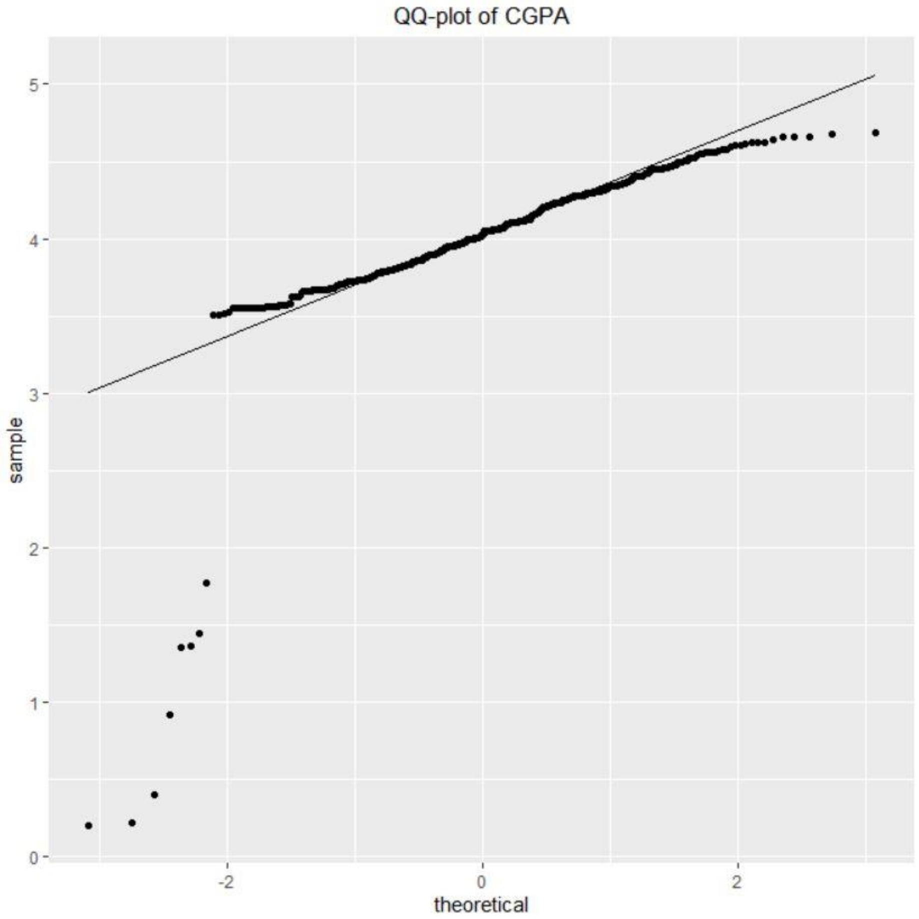


Figure 2 : QQ-Plot of CGPA

As we can see , the QQ-Plot shows a thick-tails behavior compared to the normal distribution .

c) The R script attached to thos report calculates a skewness of -4.313256

d) There are around 8 potential outliers in (0,2) interval as the boxplot depicts :

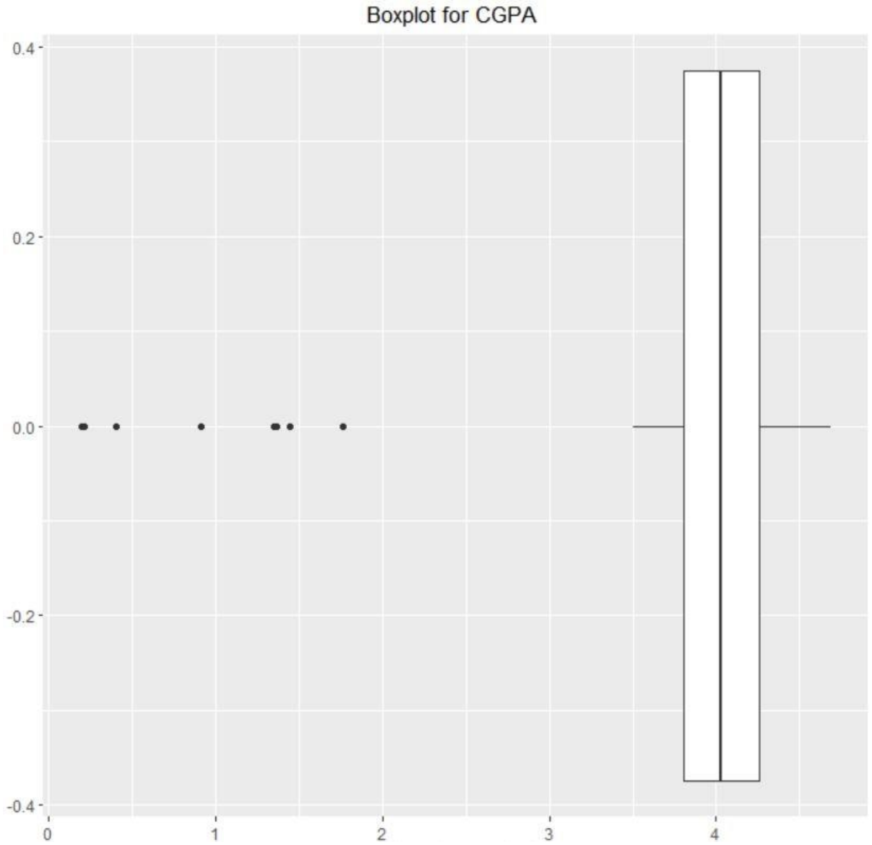
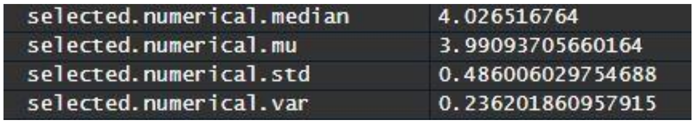


Figure 3 : Boxplot of CGPA

e) According to the R script , the parameters are :



```
selected.numerical.median    4.026516764
selected.numerical.mu        3.99093705660164
selected.numerical.std       0.486006029754688
selected.numerical.var       0.236201860957915
```

Median value is a small amount bigger than mean value , so the data is slightly left-skewed .

Median value is the value around which are the most data . Mean value is an average of the data and Standard deviation looks at how spread out a group of numbers is from the mean. Variance measures the average degree to which each point differs from the mean .

f) The plot is shown in figure 4 :

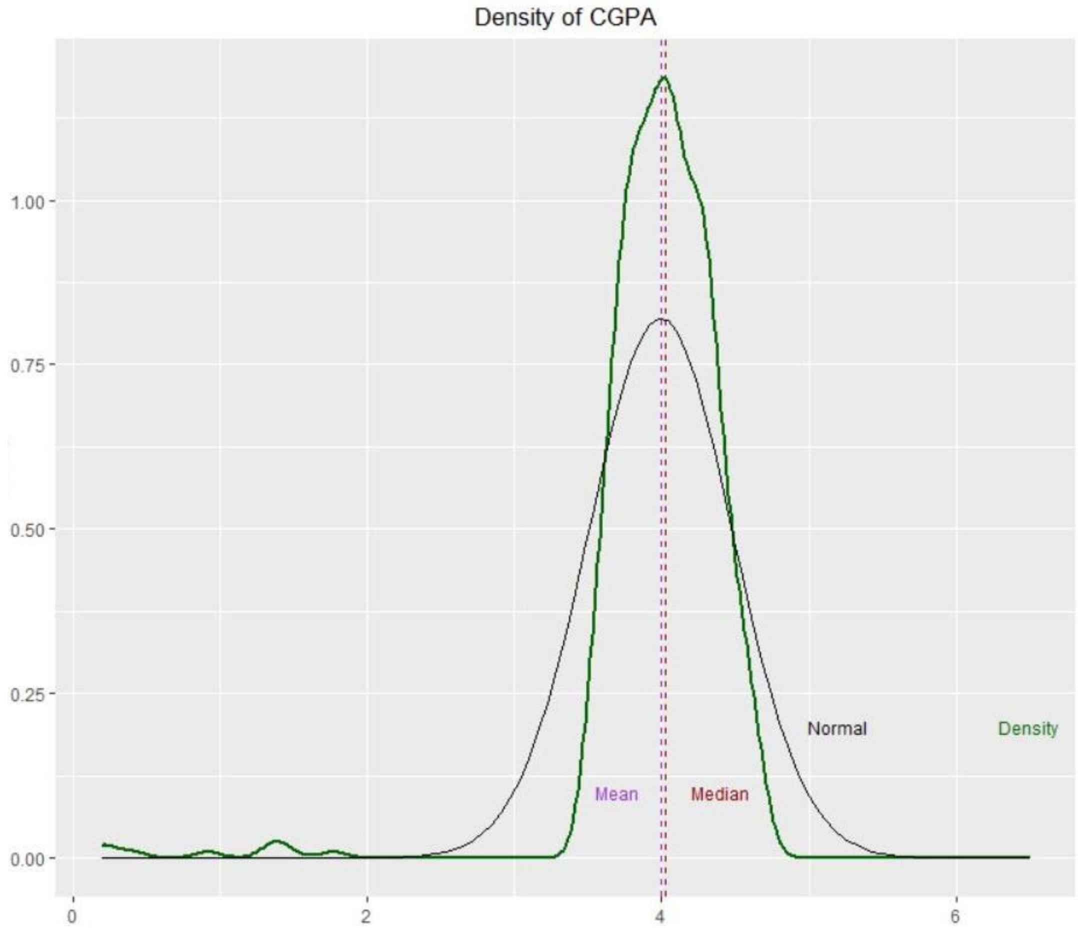


Figure 4 : density plot of CGPA

Based on median , 50% of the CGPA scores are greater than median . Mean is slightly smaller than Median , so we expect to see a drop in density when searching scores less than median . Also , we expect barely any drop in density , searching scores greater than median .

g) For CGPA with 4 mean-length parts , we get :

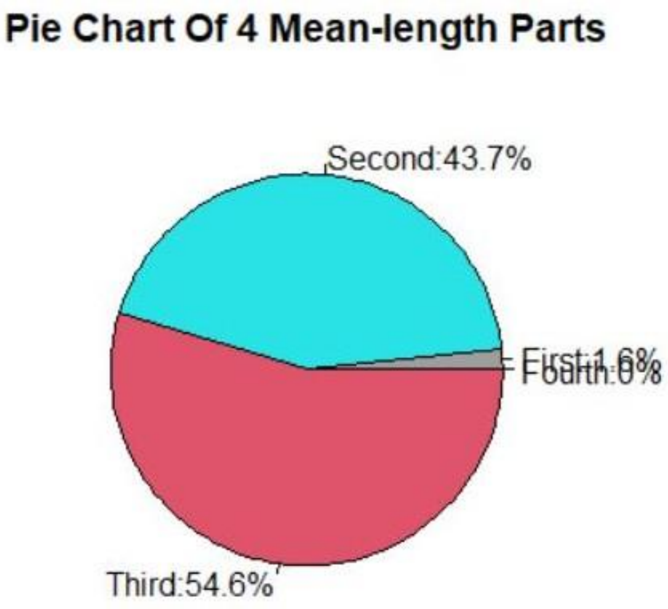
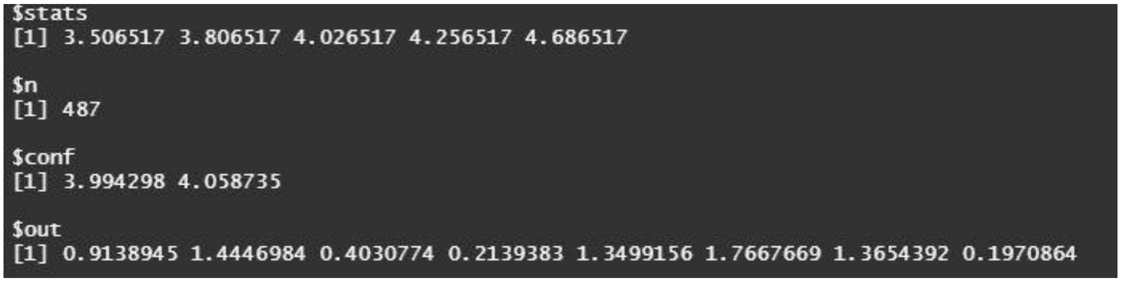


Figure 5 : Pie chart of 4 mean-length parts

h) Boxplot is shown in figure 3 . According to R calculations , we have :



```
$stats
[1] 3.506517 3.806517 4.026517 4.256517 4.686517

$n
[1] 487

$conf
[1] 3.994298 4.058735

$out
[1] 0.9138945 1.4446984 0.4030774 0.2139383 1.3499156 1.7667669 1.3654392 0.1970864
```

lower whisker is 3.506517 , 25th percentile is 3.806517 , median is 4.026517 , 75th percentile is 4.256517 & upper whisker is 4.686517 .

Question 2 :

a) In this part we choose internship abroad , 1 means “Yes” and 0 means “No”

Frequencies are : Yes : 144 & No : 343

So the percentages are : Yes : 29.56879% & No : 70.43121%

b) For the stacked bar-plot , we have :

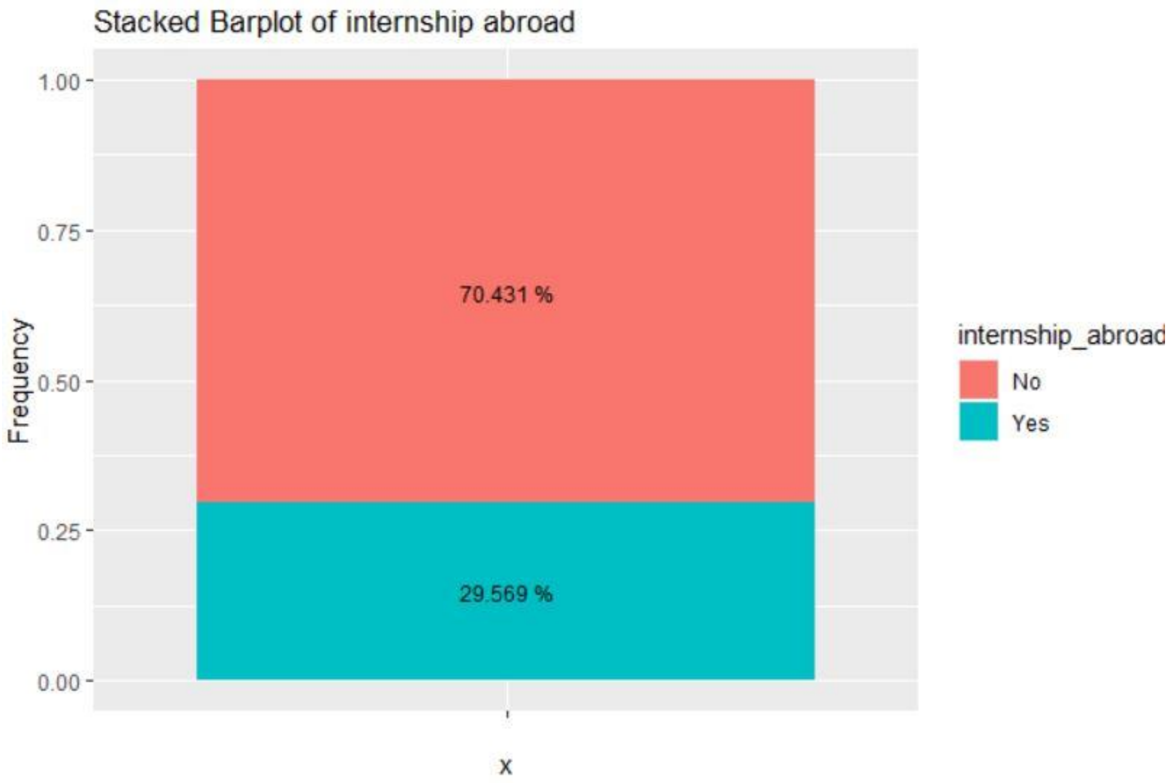


Figure 6 : Stacked barplot of internship abroad

c) Bar-plot :

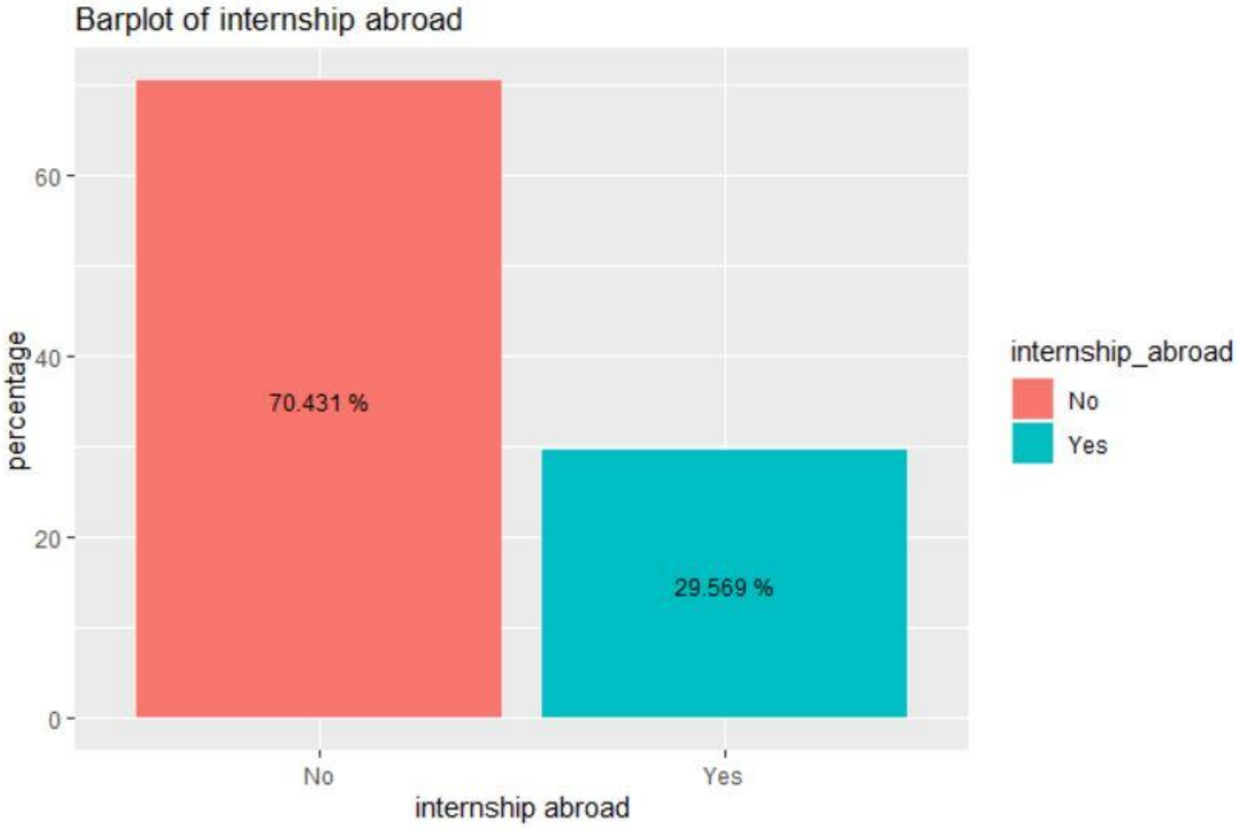


Figure 7 : Bar-plot

d) Violin plot :

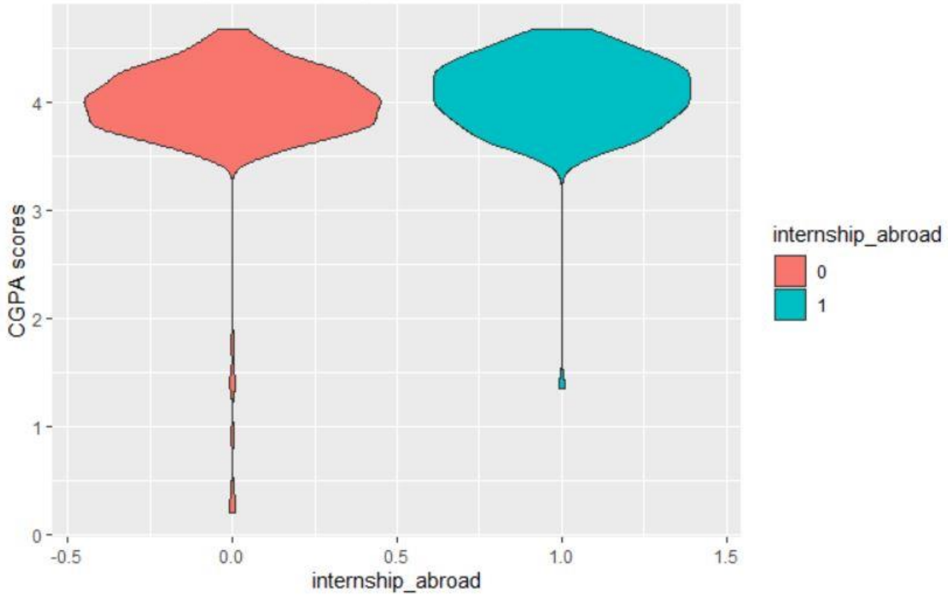


Figure 8 : Violin plot of internship abroad VS CGPA scores

Question 3 :

a) We choose GRE score and TOEFL score . We guess they have a positive correlation , because they have many subjects in common , and doing well in one will most likely mean doing well in the other .
b) Scatter plot :

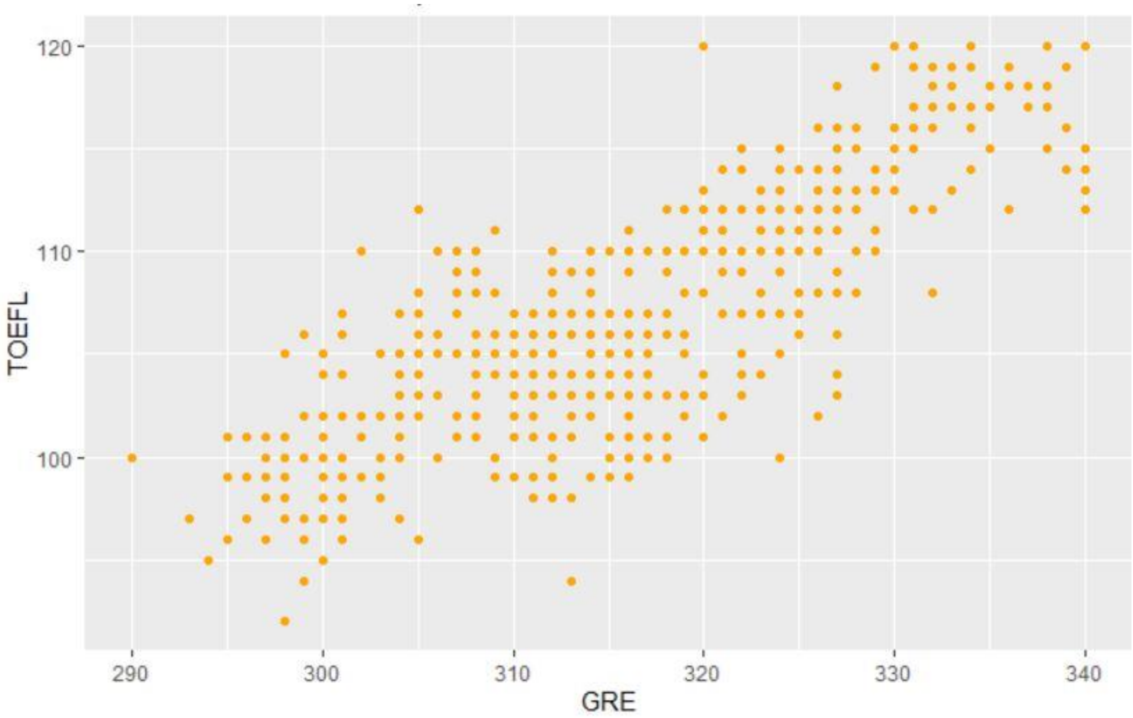
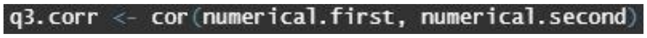


Figure 9 : scatterplot for TOEFL Score VS GRE Score

Based on the plot , we conclude that there is a positive correlation between GRE and TOEFL scores of participants .

c) Based on the R script attached to this report , correlation is equal to 0.8239124



```
q3.corr <- cor(numerical.first, numerical.second)
```

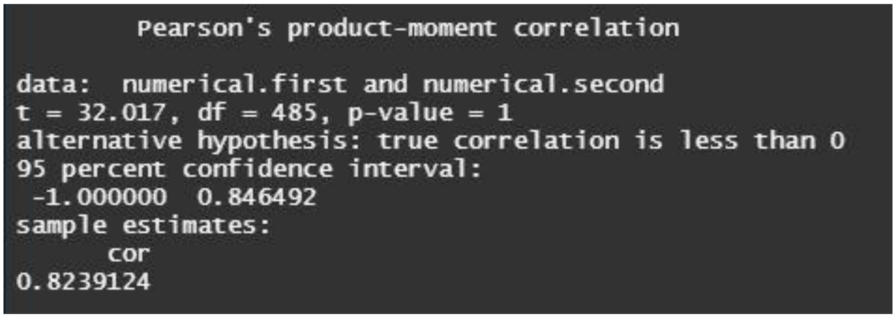
d) As we guessed , the correlation between TOEFL and GRE scores is close to 1 , which means they both move in the same direction with a strong correlation .

e) With Pearson method and in testing correlation and having alternative hypothesis of corr < 0 :

H_0 : Corr < 0

H_A : Corr ≥ 0

Result of the test :



```
        Pearson's product-moment correlation

data:  numerical.first and numerical.second
t = 32.017, df = 485, p-value = 1
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.000000  0.846492
sample estimates:
      cor
0.8239124
```

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

f) We chose Research as the categorical variable :

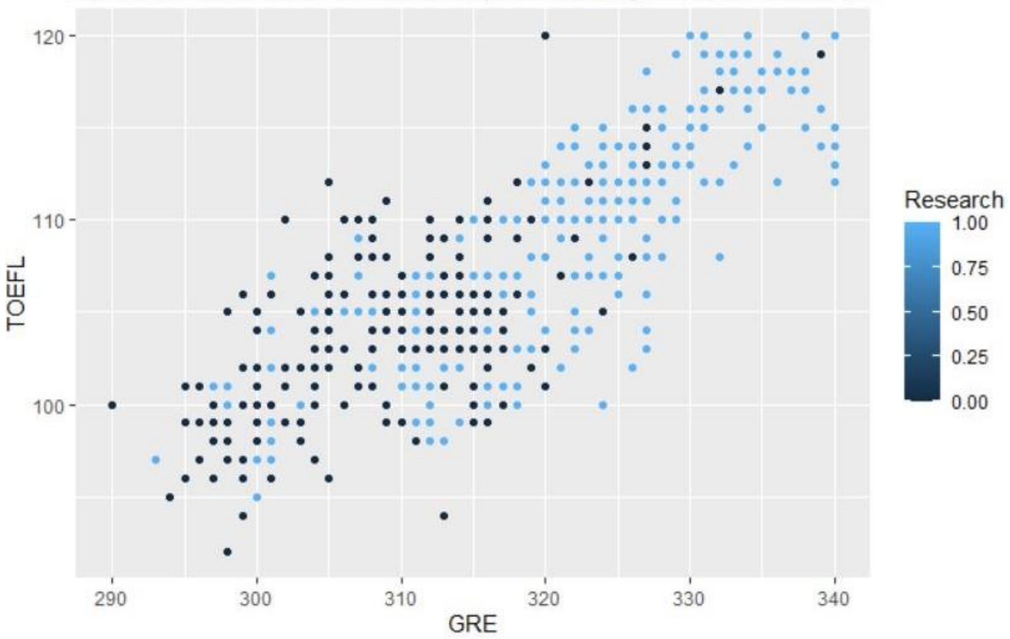


Figure 10 : TOEFL Score vs GRE Score with respect to having research scatter plot

It seems like students who have higher scores in TOEFL and GRE exams , are more likely to have done research , which makes sense .

g) Hex-bin plots for 5 , 15 & 45 bin sizes :

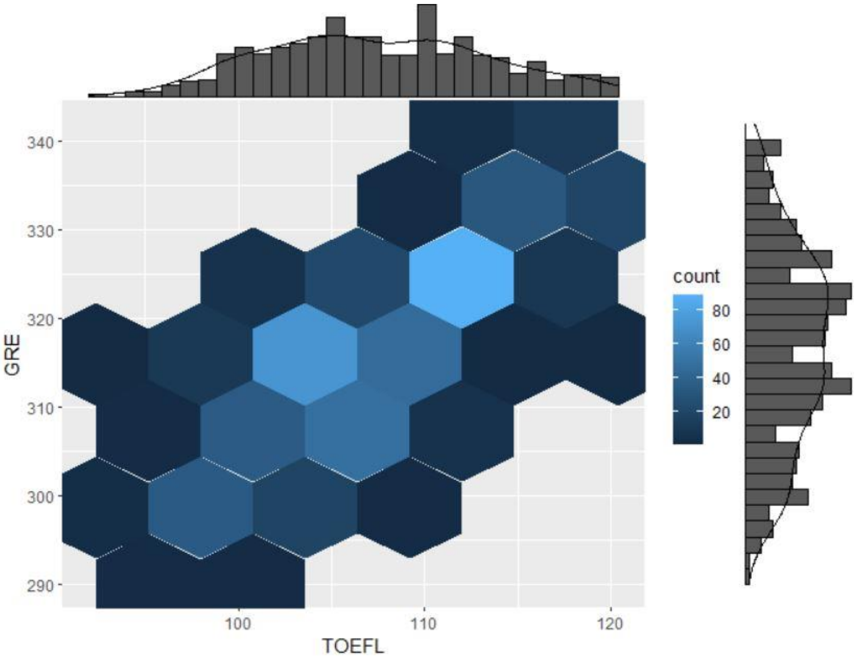


Figure 11 : Hex-bin plot for a bin size of 5

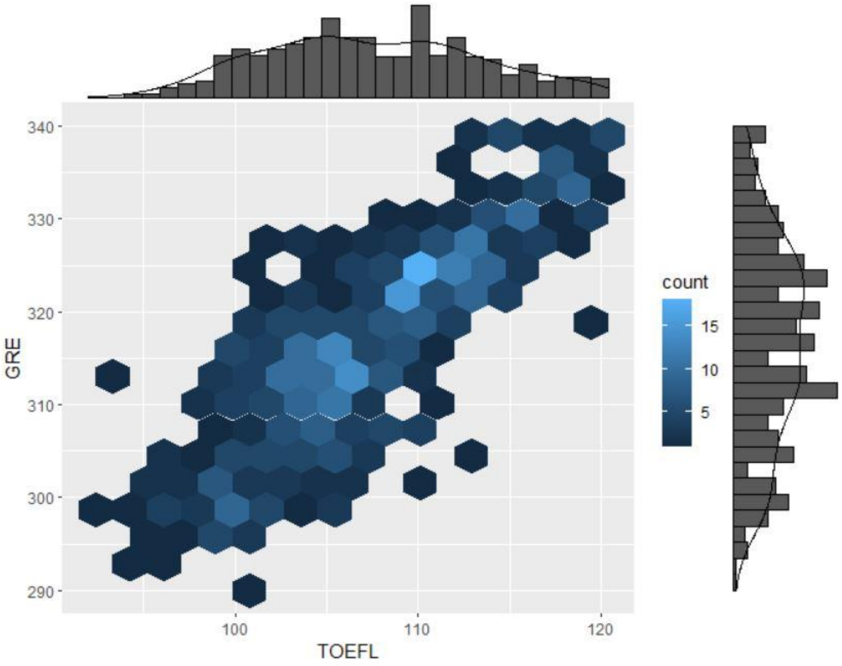


Figure 12 : Hex-bin plot for a bin size of 15

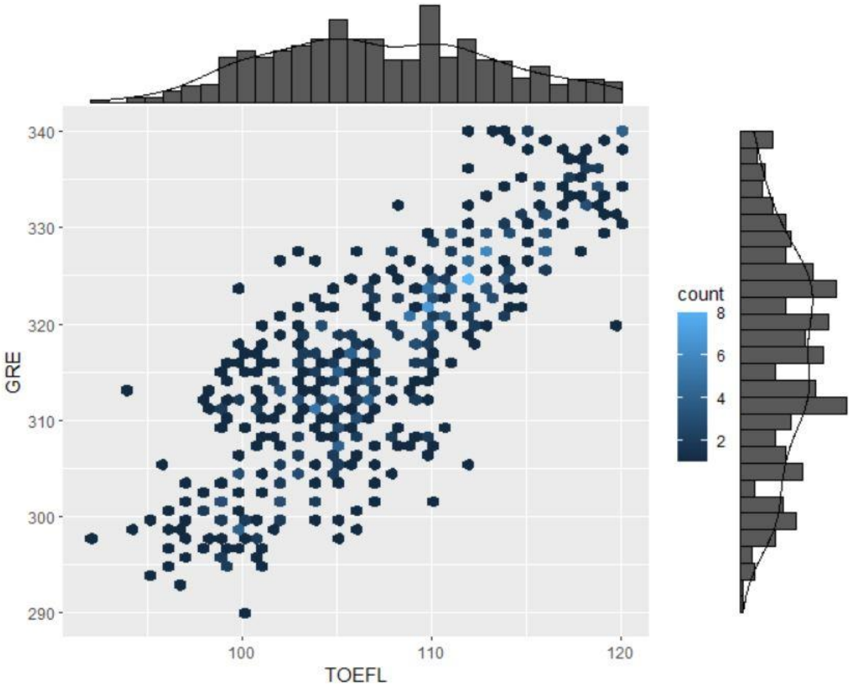


Figure 13 : Hex-bin plot for a bin size of 45

From the figures above , we can conclude that , for a large or small bin-size , we cannot have a good understanding of the data and its behavior . For small bin-sizes , we have generalized the data by a lot , and for big bin-sizes , we are being very specific and categorizing too much . The best explanation for the data , is an appropriate bin-size which is not too specific and also not too general . In these plots , the bin-size of 15 gives the best explanation about the data and we can clearly see the effects and the meaning behind the data .

h) 2D density plot :

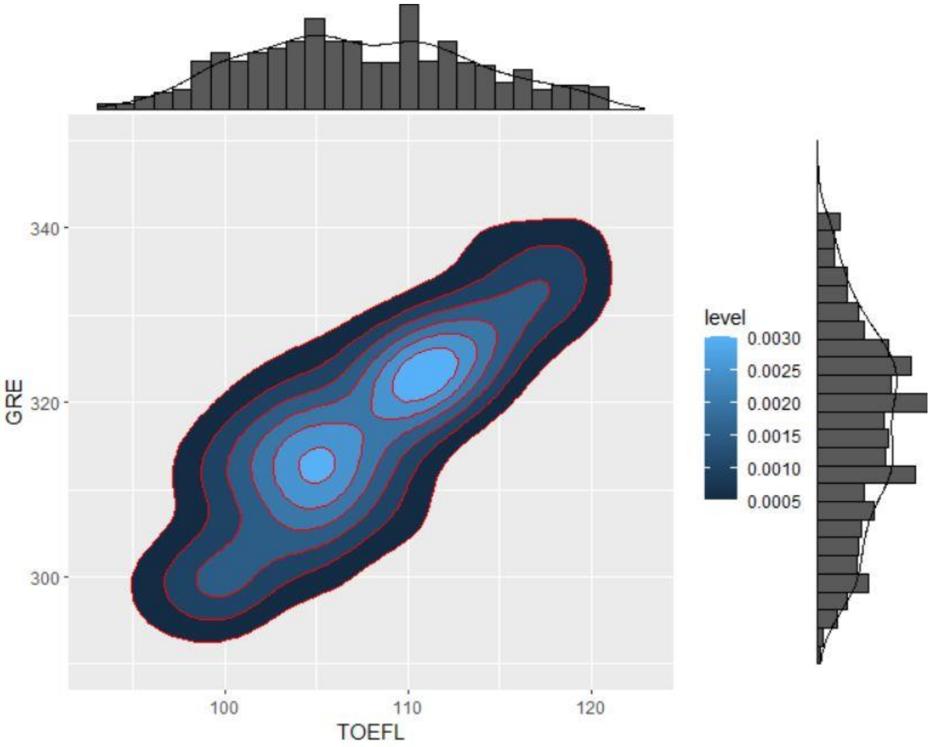


Figure 14 : 2D density plot

2d density plot , gives us the concentration of data and the rate of its change (depicted by the change of intensity in color) , while hex-bin plot gives more information about groupings of data in a certain region . 2d plot can be used as a continuous plot , while hex-bit is discrete in the neighboring areas of each neighborhood . Hex-bin plot is more agile and can change the information of a specific region , based on specific needs , while 2d density plot is less flexible .

Question 4 :

a) We can conclude a lot meaningful pattern , some of which are :
    1. TOEFL and GRE scores are very positively correlated
    2. CGPA score and Chance of admission are slightly correlated
    3. LOR and TOEFL scores are not correlated that much

   We can see that most of the variables are correlated , some are very strongly correlated and some slightly . But we can conclude that in general , students who have a good chance of admission , have done well enough in most if not all of the categories .

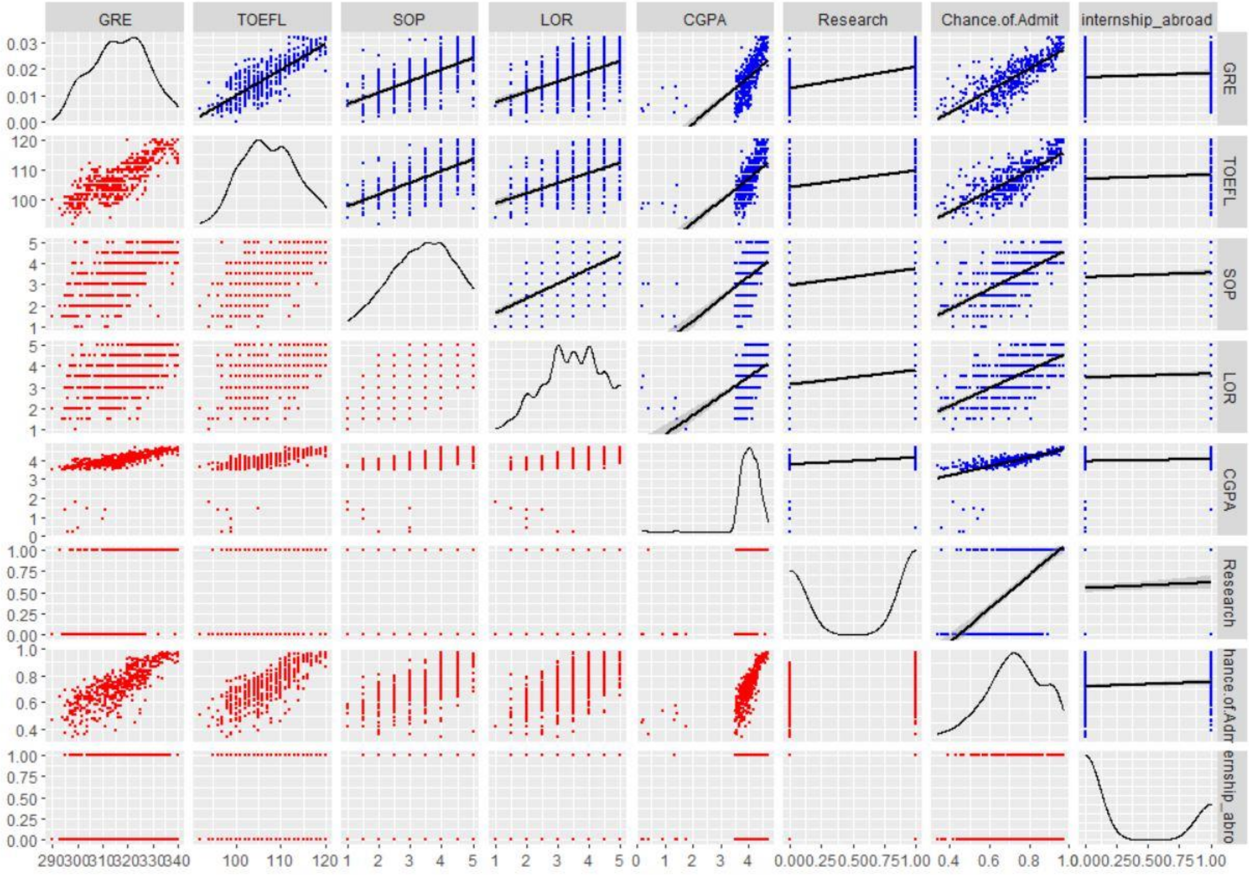


Figure 15 : All the bivariant relations

b) Plot is shown in below :

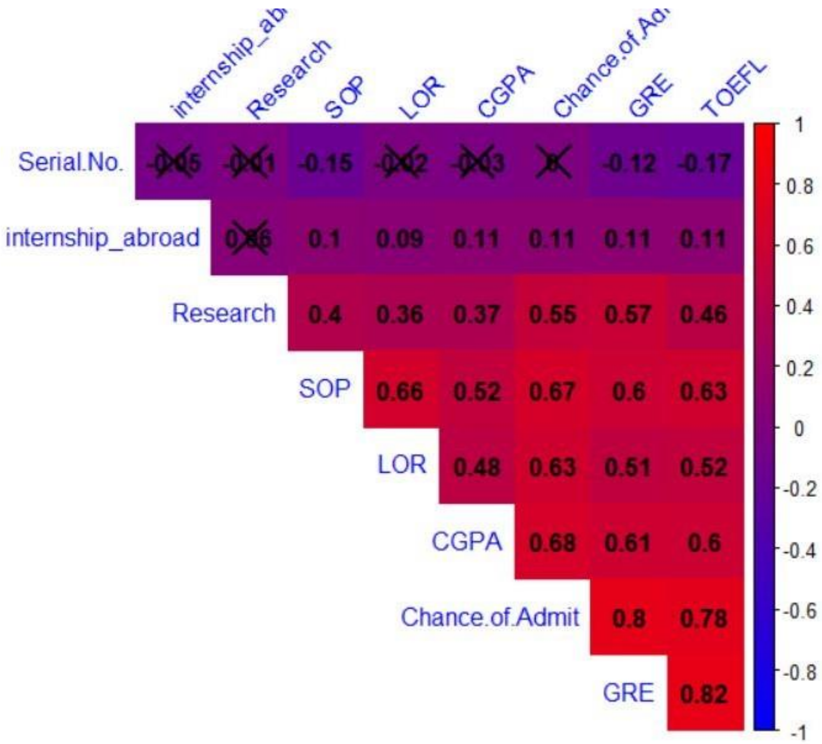


Figure 16 : heatmap correlogram , red is positive and blue is negative correlation

c) We chose TOEFL , LOR & SOP scores for the numerical and university for categorical .

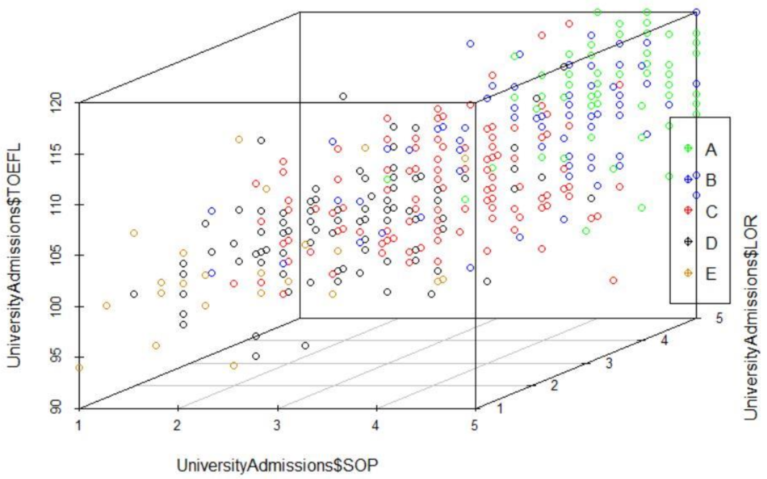
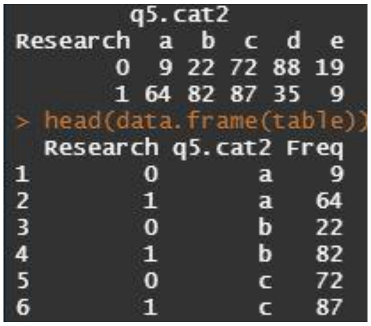


Figure 17 : 3D scatterplot

As we can see , students from type “A” universities are more concentrated on the top right corner , which translates to higher numerical scores (TOEFL , SOP & LOR) , and students from “E” type universities are more common in the bottom left corner , which means worse LR , SOP & TOEFL scores . Other students are in between , and the better the university , the higher and more to the right can they be found on the 3D scatterplot and vice versa .

Question 5 :

a) Plot :



```
           q5.cat2
Research  a  b  c  d  e
       0  9 22 72 88 19
       1 64 82 87 35  9
> head(data.frame(table))
  Research q5.cat2 Freq
1        0       a    9
2        1       a   64
3        0       b   22
4        1       b   82
5        0       c   72
6        1       c   87
```

b) Plot :

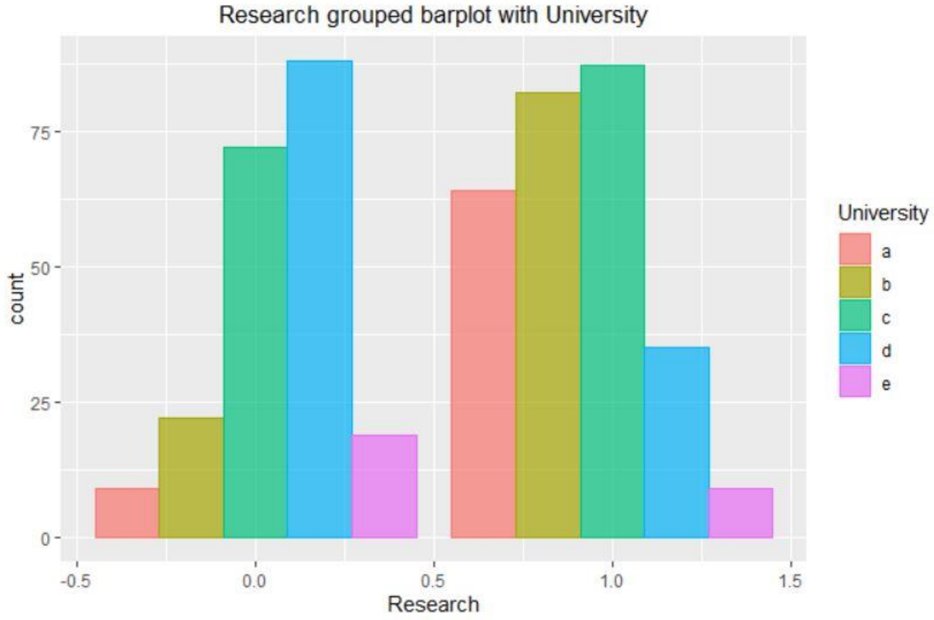


Figure 18 : Grouped bar chart

c) Plot :

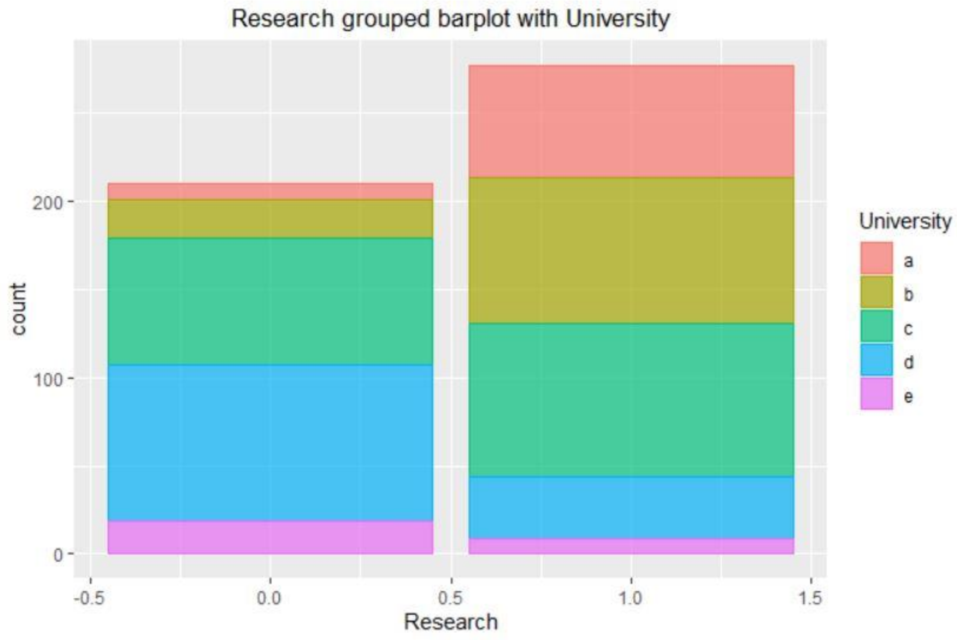


Figure 19 : Segmented bar plot

Question 6 :

a) Confidence interval for TOEFL score as the numerical variable is 105.7947 to 108.3653

```
> TOEFL.CI
[1] 105.7947 108.3653
```

b) We are 95% sure that true mean of TOEFL score of schools students will be in interval of 105.7947 to 108.3653 .

c) Plot :

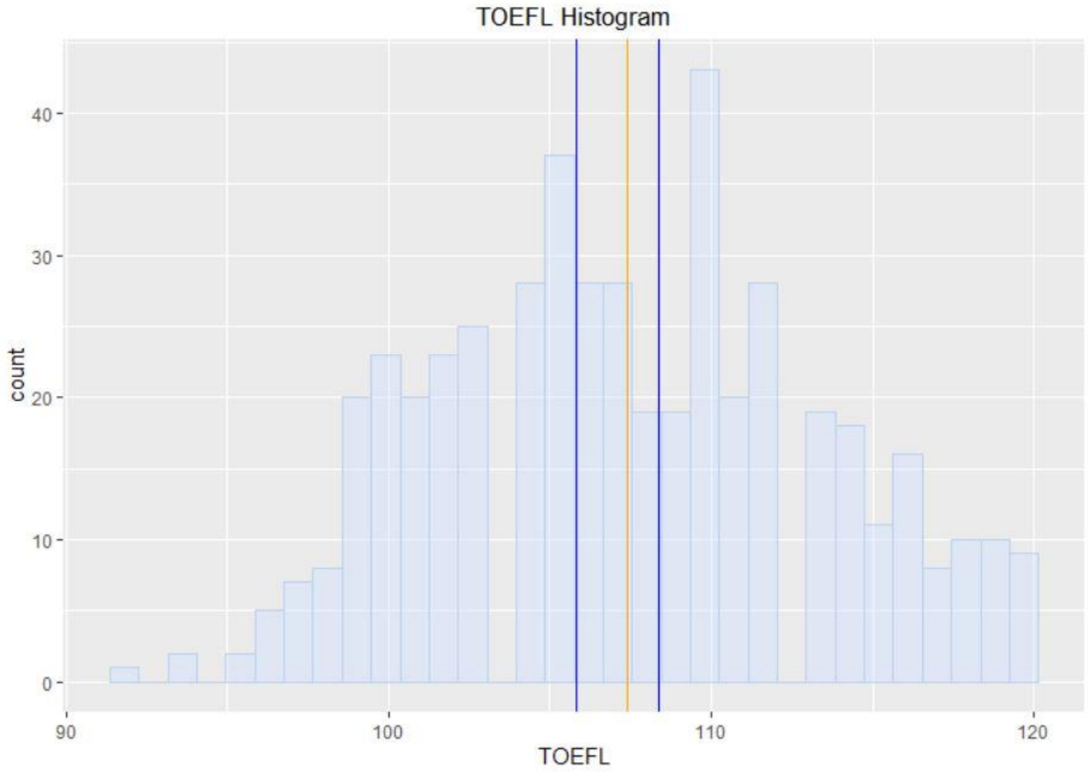


Figure 20 : TOEFL histogram , Orange line shows the mean of TOEFL scores and blue lines show the confidence boundaries

d) Hypothesis :

H_0 : mean of TOEFL scores = 105

H_A : mean of TOEFL scores != 105

pValue is equal to : 0.0007573922 . because 0.05 > pValue , we reject *H_0* . The pValue 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 , we will reject null hypothesis .

e) Because H_0 is 105 , and not inside the interval of part a , it is rejected . Confidence interval and pValue result of rejecting hypothesis is equal .

f) Type 2 error is equal to : 0.02073107 . It means probability of accepting null hypothesis , whereas it is false .

g) Power is equal to type 2 error subtracted from 1 . , so power is equal to 0.9792689 .

Question 7 :

A. We have :

a) Because we don't have the variance an sample size is less than 30 , we choose t-test & accordingly , df=25-1 .

b) Hypothesis :

$$H_0 : Mean_{CGPA} = Mean_{sop}$$
$$H_A : Mean_{CGPA} != Mean_{sop}$$

We run the t-test with R . pValue is calculated as 0.0910448149 .
Since pValue < significance level (0.05) , we can reject null hypothesis .

B. Sampling data without replacing to have independent sampling . because we have a large number of samples , we use z-test instead of f-test .
Hypothesis :

$$H_0 : Mean_{CGPA} = Mean_{sop}$$
$$H_A : Mean_{CGPA} != Mean_{sop}$$

We run the z-test in R , and we get the pValue of 2.9457581e-7 . Here pValue is smaller than alpha , so null hypothesis is rejected .

Question 8 :

a) We choose CGPA score , as it has outliers :

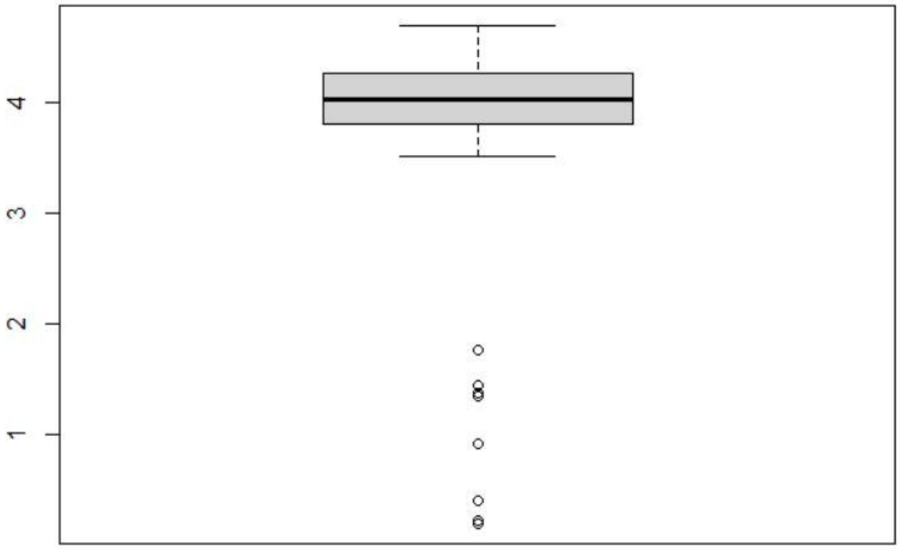


Figure 21 : outliers of CGPA

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

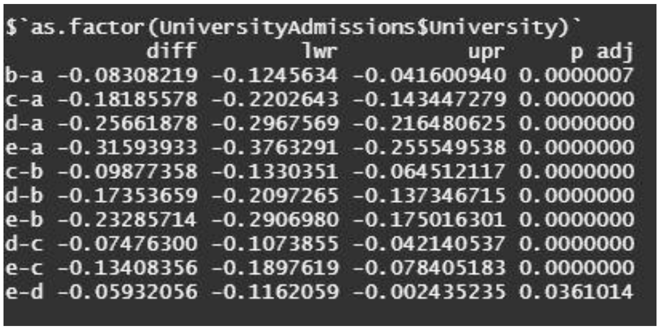
Confidence interval is : 3.827527 to 4.104011

b) After bootstrapping with 20 samples , I used both se method and quantile method to calculate the interval . since 20 samples are less than 30 , I used to score to find 95% confidence interval in t-dist with df = 20-1 = 19 .
Confidence interval is : 3.902564 to 4.055285

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

Question 9 :

Because the pValue is small , they have different means .



```
$`as.factor(UniversityAdmissions$University)`
           diff        lwr          upr     p adj
b-a -0.08308219 -0.1245634 -0.041600940 0.0000007
c-a -0.18185578 -0.2202643 -0.143447279 0.0000000
d-a -0.25661878 -0.2967569 -0.216480625 0.0000000
e-a -0.31593933 -0.3763291 -0.255549538 0.0000000
c-b -0.09877358 -0.1330351 -0.064512117 0.0000000
d-b -0.17353659 -0.2097265 -0.137346715 0.0000000
e-b -0.23285714 -0.2906980 -0.175016301 0.0000000
d-c -0.07476300 -0.1073855 -0.042140537 0.0000000
e-c -0.13408356 -0.1897619 -0.078405183 0.0000000
e-d -0.05932056 -0.1162059 -0.002435235 0.0361014
```

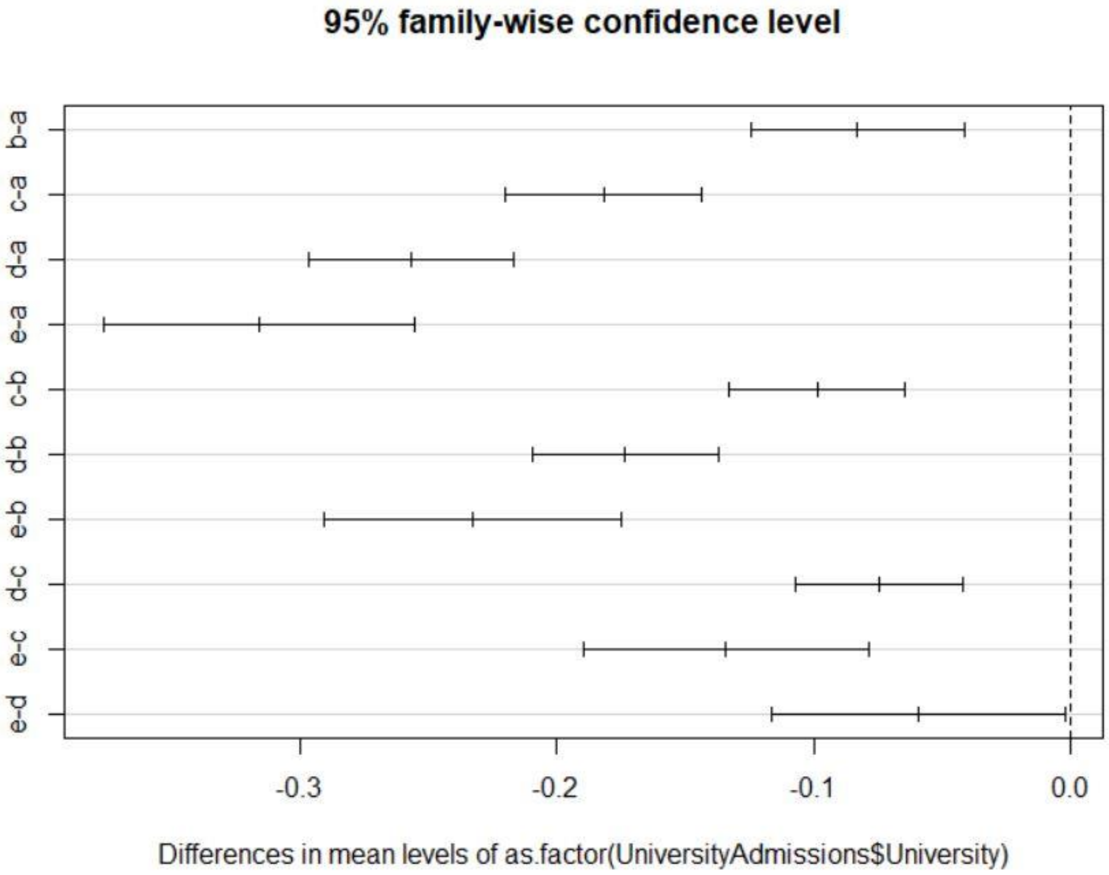


Figure 22 : ANOVA analysis

R code used for this project :

```
library(magrittr)

library(dplyr)

library(ggfortify)

library(ggplot2)

library(plyr)

library(gridExtra)

library(ggpubr)

require(qqplotr)

library("ggpubr")

library(moments)

library(hexbin)

library(ggmosaic)

library(GGally)


#####Question 0


UniversityAdmissions <-
read.csv("F:\\Downloads\\UniversityAdmissions.csv", header = TRUE)



#Question 1
```

```
selected.numerical <- UniversityAdmissions$CGPA


#A


binsize <- 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 = binsize, alpha = 0.4)
+ geom_density(linetype="dashed", alpha = 0.3, size=1) + labs(title =
"CGPA", x = "Score", y="Density")+ theme(plot.title =
element_text(hjust = 0.5))


selected.numerical.hist


#B


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


selected.numerical.qq


#C
```

```
print(skewness(selected.numerical))


#D


selected.numerical.boxplot <-
ggplot(as.data.frame(UniversityAdmissions),
                                     aes(x = selected.numerical))+
  geom_boxplot()+
  labs(title="Boxplot for CGPA ",
       xlab="score")+
  theme(plot.title = element_text(hjust = 0.5))


selected.numerical.boxplot


#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)


#F


selected.numerical.density <- ggplot(UniversityAdmissions,
                                     aes(x = selected.numerical)) +
  geom_vline(xintercept = selected.numerical.mu,
             linetype="dashed",
             color = "darkorchid3") +
  geom_vline(xintercept = selected.numerical.median,
             linetype="dashed",
             color = "darkred") +
  geom_density(color = "darkgreen", size = 1)+
  stat_function(fun = dnorm, n = 101, args = list(mean =
selected.numerical.mu,
                                                  sd =
selected.numerical.std)  )+
  annotate("text", x = 5.2 , label = "Normal", y = 0.2, size = 3,
angle = 0 , color="black") +
  annotate("text", x = 6.5 , label = "Density", y = 0.2, size = 3,
angle = 0, color="darkgreen") +
  annotate("text", x = 3.7 , label = "Mean", y = 0.1, size = 3, angle
= 0, color="darkorchid3") +
  annotate("text", x = 4.4 , label = "Median", y = 0.1, size = 3,
angle = 0, color="darkred") +
  labs(title="Density of CGPA")+
  theme(plot.title = element_text(hjust = 0.5))


selected.numerical.density


#G


CGPA <- UniversityAdmissions$CGPA
mu<- mean(CGPA)
Partial_CGPA <- c(length(CGPA[CGPA<=0.5*mu]) ,
                  length(CGPA[CGPA>0.5*mu & CGPA <= mu]) ,
                  length(CGPA[CGPA>mu & CGPA <= 1.5*mu]) ,
                  length(CGPA[CGPA>1.5*mu & CGPA <=2*mu]))
```

```
percentage<-round(100*Partial_CGPA/sum(Partial_CGPA) , 1 )

labels<- c("First" ,"Second" , "Third" , "Fourth")


pie(Partial_CGPA , labels=paste(paste0(labels , ":" , percentage ,
"%") , sep= " ") , col = Partial_CGPA)

title ("Pie Chart Of 4 Mean-length Parts")


#H


boxplot.stats(selected.numerical)

selected.numerical.boxplot



#Question 2



internshipYes <- dplyr::filter(UniversityAdmissions,
internship_abroad=="1")

internshipNo <- dplyr::filter(UniversityAdmissions,
internship_abroad=="0")

student.internship_abroad <- UniversityAdmissions$internship_abroad


##Part a


length(internshipYes$internship_abroad)

length(internshipNo$internship_abroad)
```

```
internshipYes.percentage <-
length(internshipYes$internship_abroad)/length(student.internship_abro
ad)

internshipNo.percentage <-
length(internshipNo$internship_abroad)/length(student.internship_abroa
d)

internshipYes.percentage

internshipNo.percentage


##Part b


data <- data.frame( internship_abroad = c("Yes", "No"),

  Value = c(internshipYes.percentage*100,
internshipNo.percentage*100))


internship_abroad.barplot <- ggplot(data , aes( fill=internship_abroad
, x = " " , y = Value )) +

  geom_bar(position = "fill" , stat="identity") +


  labs(title="Stacked Barplot of internship abroad", y = 'Frequency')+

  annotate("text", x = 1 , label =
paste(toString(round(internshipNo.percentage*100, digit=3)),"%"), y =
1-internshipNo.percentage/2 , size = 3)+

  annotate("text", x = 1 , label =
paste(toString(round(internshipYes.percentage*100, digit=3)),"%"), y =
internshipYes.percentage/2 , size = 3)


internship_abroad.barplot
```

```
##Part c



internship_abroad.hbarplot <- ggplot(data, aes(x =internship_abroad,
y=Value  , fill=internship_abroad)) +

  geom_bar( stat="identity") +

  labs(title="Barplot of internship abroad", y = 'Frequency')+

  annotate("text", x = 1 , label = paste(toString(round(
internshipNo.percentage*100, digit=3)),"%"), y =
internshipNo.percentage*50 , size = 3 ) +

  annotate("text", x = 2 , label =
paste(toString(round(internshipYes.percentage*100, digit=3)),"%"), y =
internshipYes.percentage*50 , size = 3 ) +

  xlab("internship abroad")+

  ylab("percentage")

internship_abroad.hbarplot





##Part d


temp <- data.frame(UniversityAdmissions)

q2.violin <- ggplot(temp , aes( x=internship_abroad, y = CGPA,  fill =
internship_abroad)) +

  geom_violin(aes ( fill = factor(internship_abroad)))+

  ylab("CGPA scores")+

  labs(title="Violin plot of internship abroad VS CGPA scores")+

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

```
q2.violin



#Question 3



numerical.second <- UniversityAdmissions$TOEFL
numerical.first <- UniversityAdmissions$GRE



##Part b

temp <- data.frame(UniversityAdmissions)

q3.scatter <- ggplot(temp, aes( x = GRE, y = TOEFL ))+
  geom_point(color="orange")+
  labs(title="scatterplot for TOEFL Score VS GRE Score")+
  theme(plot.title = element_text(hjust = 0.5))

q3.scatter



##Part c

q3.corr <- cor(numerical.first, numerical.second)
q3.corr
```

```
##part E

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


q3.test.corr


##Part F


twonum.and1cat.scatter <- ggplot(temp, aes(x= GRE, y=TOEFL, color=
Research , fill= Research ))+
  geom_point()+
  labs(title = "TOEFL Score vs GRE Score with respect to having
research 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 = TOEFL, y = GRE)) +
geom_point(col="transparent")+geom_hex(bins=45), type= "densigram",
margins = "both")

q3.densigram.hex





#Part H

temp <- data.frame(UniversityAdmissions)

q3.2ddenisty.hex <- ggMarginal(ggplot(temp, aes(x = TOEFL, y = GRE)) +

                    ylim(c(290,350))+xlim(c(93,123))+

                    geom_point(col="transparent")+

                    stat_density2d(aes(fill=..level..),
geom="polygon", color="red"), type= "densigram", margins = "both")


q3.2ddenisty.hex








#Question 4
```

```
#a

library(GGally)

featurePlot(x=temp[,1:5], y=temp[,5:10], plot="pairs")
ggpairs(dplyr::select_if(UniversityAdmissions, is.numeric), title = "Correlogram")
list.of.num <- c(2, 3, 5, 6, 7, 8 , 9, 10)

#density, without failure
ggpairs(UniversityAdmissions[, list.of.num],
        upper = list(continuous = wrap("density", colour="blue" , size = 0.5 )),
        lower = list(continuous = wrap("points", colour="red" , size = 0.5 )))

#linear relationship
ggpairs(UniversityAdmissions[, list.of.num],
        upper = list(continuous = wrap("smooth", colour="blue", size = 0.5 )),
        lower = list(continuous = wrap("points", colour="red", size = 0.5 )))
```

```
#b


library(Hmisc)
col <- colorRampPalette(c("blue", "red"))
UniversityAdmissions.corr <-
rcorr(as.matrix(dplyr::select_if(UniversityAdmissions, is.numeric)))
UniversityAdmissions.corr.p <- UniversityAdmissions.corr$P
UniversityAdmissions.corr.p[is.na(UniversityAdmissions.corr.p)] <- 1


M <- cor(dplyr::select_if(UniversityAdmissions, 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 =
UniversityAdmissions.corr.p, sig.level = 0.05, diag = FALSE)



#c.


cols <- c("Green", "Blue" , "red" , "black" , "orange3")
cols <- cols[as.numeric(as.factor(UniversityAdmissions$University))]
library(scatterplot3d)
scatterplot3d(UniversityAdmissions$SOP ,UniversityAdmissions$LOR ,
UniversityAdmissions$TOEFL , color = cols )
legend("right" , legend = c("A" , "B" , "C" , "D" , "E") ,  col =
c("Green", "Blue" , "red" , "black" , "orange3") , pch = 10)
```

```
#Question 5
```

```
##Part a

q5.cat1 <- UniversityAdmissions$Research

q5.cat2 <- UniversityAdmissions$University

Research <- UniversityAdmissions$Research

University <- UniversityAdmissions$University

table <- table(Research, q5.cat2)

print.table(table)

head(data.frame(table))
```

```
##Part b

combined.barplot.RS <- ggplot(temp, aes(x = Research,color =
University, fill = University)) +

  geom_bar(position = "dodge", alpha = 0.7) +

  labs(title="Research grouped barplot with University",
x="Research")+

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

```
combined.barplot.RS


##Part c
combined.segbarplot.RS <- ggplot(temp, aes(x = Research,color =
University, fill = University)) +
  geom_bar(alpha = 0.7) +
  labs(title="Research grouped barplot with University",
x="Research")+
  theme(plot.title = element_text(hjust = 0.5))


combined.segbarplot.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(UniversityAdmissions$TOEFL, 100)
TOEFL.CI <- calculate_ci(sampled_data, 0.95)
TOEFL.CI



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



q6.TOEFL.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, 105 , 0.05)


#Part f

null.value <- 105

z_value <- abs(qnorm((1-0.05)/2))

errorTypeII <- pnorm(null.value + z_value *
sd(sampled_data)/sqrt(length(sampled_data)) - mean(sampled_data))

errorTypeII


#part g

power <- 1 - errorTypeII

power
```

```
#Question 7




#Part a

UniversityAdmissions.sampled <- sample_n(UniversityAdmissions, 25)

x_bar <- mean(UniversityAdmissions.sampled$CGPA) -
mean(UniversityAdmissions.sampled$SOP)

s1 <- sd(UniversityAdmissions.sampled$CGPA)

s2 <- sd(UniversityAdmissions.sampled$SOP)

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

CGPA.sample <- sample(UniversityAdmissions$CGPA, 100)

SOP.sample <- sample(UniversityAdmissions$SOP, 100)

zdist.2tail.meantest(CGPA.sample - SOP.sample, 0, 0.05)




#Question 8
```

```
library(bootstrap)
#Part a
q8.chosen.numerical <- UniversityAdmissions$CGPA
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))

calculate_ci(q8.mean.Pa.CI , 0.95)



#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))


calculate_ci(q8.mean.Pb.CI , 0.95)




#Question 9




result <- aov(UniversityAdmissions$Chance.of.Admit~as.factor(UniversityAdmissions$University))

print(summary(result))

thsd <- TukeyHSD(result, conf.level = 0.95)

plot(thsd)

print(thsd)
```