Mini Project 4

Names: Sathya Pooja Ramireddy(sxr176830)

Sashidhar Donthiri (sxd173730)

Contributions: Both of us have individually solved it, later discussed it and put it together.

---

1)

As asked, scatter plot for Sales against TV and Sales against radio:

```
#readData into DataFrame called MyData

MyData <- read.csv(file="C:/Users/shash/Documents/R for Stats/Mini Projects/Miniproject 4/advertising.csv", header=TRUE, sep=",")

#just checking if the data is properly read

MyData$SerialNo

MyData$TV

MyData$radio

MyData$newspaper

MyData$sales
```

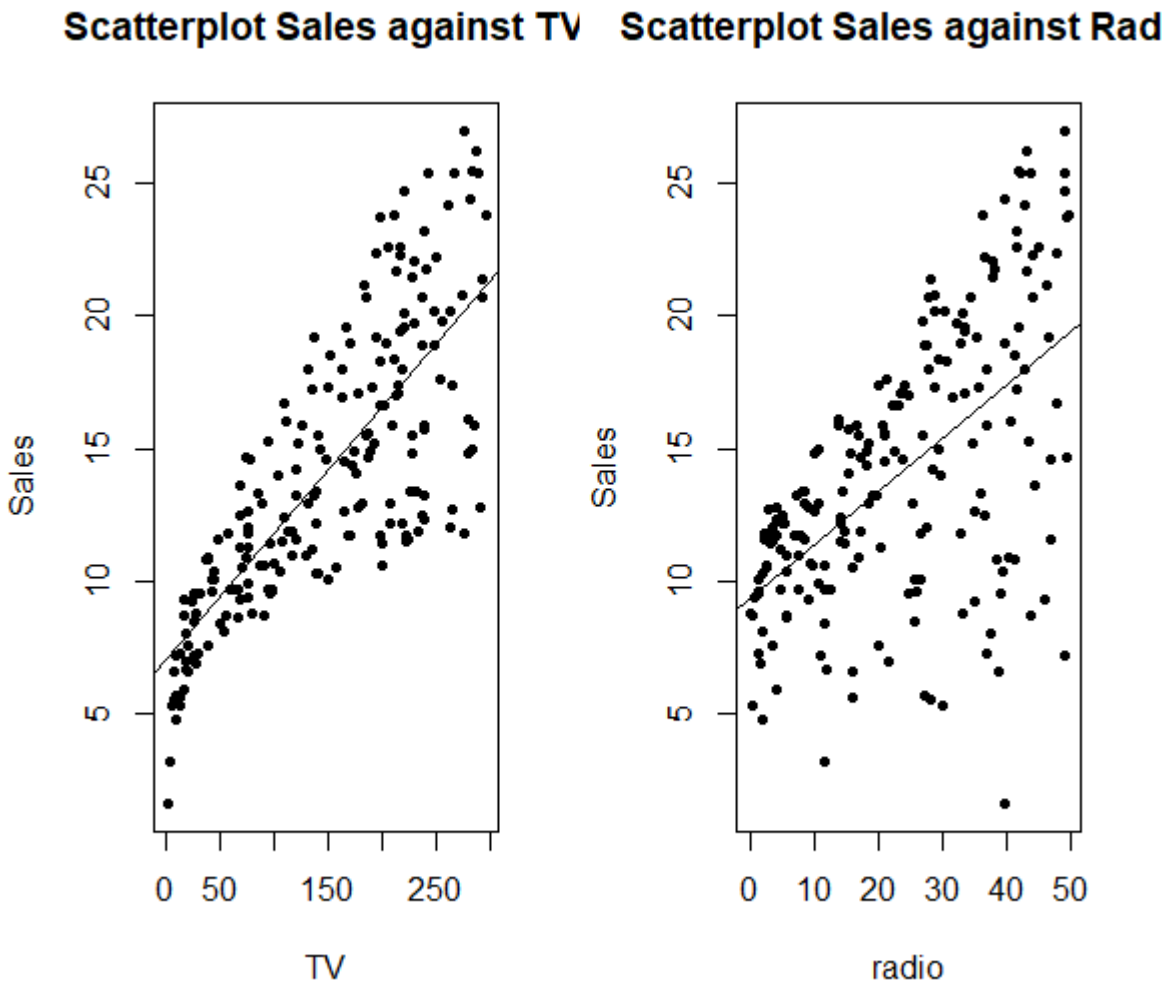
#scatter plot of Sales against TV

```
plot(MyData$TV,MyData$sales , main="Scatterplot Sales against TV",

   xlab="TV", ylab="Sales  ", pch=20)

(line.reg=lm(MyData$sales~MyData$TV))

abline(line.reg)
```

#scatter plot of Sales against radio

```
plot(MyData$radio,MyData$sales, main="Scatterplot Sales against Radio",

   xlab="radio", ylab="Sales", pch=20)

(line.reg1=lm(MyData$sales~MyData$radio))

abline(line.reg1)

#to plot multiple graphs in single sheet
```

par(mfrow=c(1,2))



- Scatter plots of the above two graphs show that, there is linear dependency in the two graphs i.e.; there is linear dependency between Sales against TV and Sales against radio which can be found out by the correlation value.
- From the graph, the Sales against TV has stronger linear dependency than Sales against radio.
- Correlation: Correlation value reflects the strength of linear dependency between X and Y ranges from -1 to 1.
- If values=0, it means there's no association between the two variables.
- A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

- A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decrease.
- It measures strength of relationship between X and Y.

#correlation(X,Y)= Cov(X,Y)/(SD(X)*SD(Y))

- Correlation is directly proportional to the Covariance between X and Y AND inversely proportional to SD of X and Y individually.

As asked Correlation can be found using cor() :

#Calculate the correlation coefficient sales against TV

Cor1=cor(MyData$TV,MyData$sales)

Cor1

#result= 0.7822244

#Calculate the correlation coefficient sales against radio

Cor2=cor(MyData$radio,MyData$sales)

Cor2

#result=0.5762226

The Correlation Coefficient value between Radio and Sales is 0.5762226

TV and sales is 0.7822244

The point estimate is considered and found using the cor function are 0.5762226 and 0.7822244 respectively.

Calculate the Bootstrap Estimate:

#install the bootstrap library called boot

install.packages("boot")

#parameter of interest is correlation coefficient here

#attach the boot library

Library(code)

#Bootstrap tech to estimate the bias of point estimate between TV and Sales

corr.npar1=function(x,indeces){

result=cor(MyData$sales,MyData$TV)

```
  return(result)
}
Corr.npar1()
#result= 0.7822244
#bootstrap estimate for sales and TV
corr.npar.boot1=boot(MyData, corr.npar, R=999, sim="ordinary", stype="i")
corr.npar.boot1

#Bootstrap tech to estimate the bias of point estimate between TV and Sales
corr.npar2=function(x,indeces){
  result=cor(MyData$sales,MyData$radio)
  return(result)
}
Corr.npar2()
#result= 0.5762226
#bootstrap estimate for sales and radio
corr.npar.boot2=boot(MyData, corr.npar, R=999, sim="ordinary", stype="i")
corr.npar.boot2

#Calculate Standard Error of the considered Point Estimate i.e.; Correlation efficient val
SE1=sqrt((1-corr.npar1^2)/(n-2)) #value of n=200
SE2=sqrt((1-corr.npar2^2)/(n-2)) #value of n=200
```

Standard Error of Correlation co-eff of point estimate **between TV and sales: 0.04427445**

Between radio and sales: 0.0588247

```
#Calculate 95% CI using percentile bootstrap:
#Percentile Bootstrap method to estimate 95% confidence interval for correlation coefficient for Radio and Sales
boot.ci(corr.npar.boot1)
#to verify Percentile Bootstrap:
```

sort(corr.npar.boot1$t)[c(25, 975)]

#Result: **Lower: 0.4681992**

Upper: 0.6748921

#Percentile Bootstrap method to estimate 95% confidence interval for correlation coefficient for TV and Sales

boot.ci(corr.npar.boot2)

#to verify Percentile Bootstrap:

sort(corr.npar.boot2$t)[c(25, 975)]

#Result: **Lower: 0.7261286**

Upper: 0.8312579

Interpretation of results obtained:

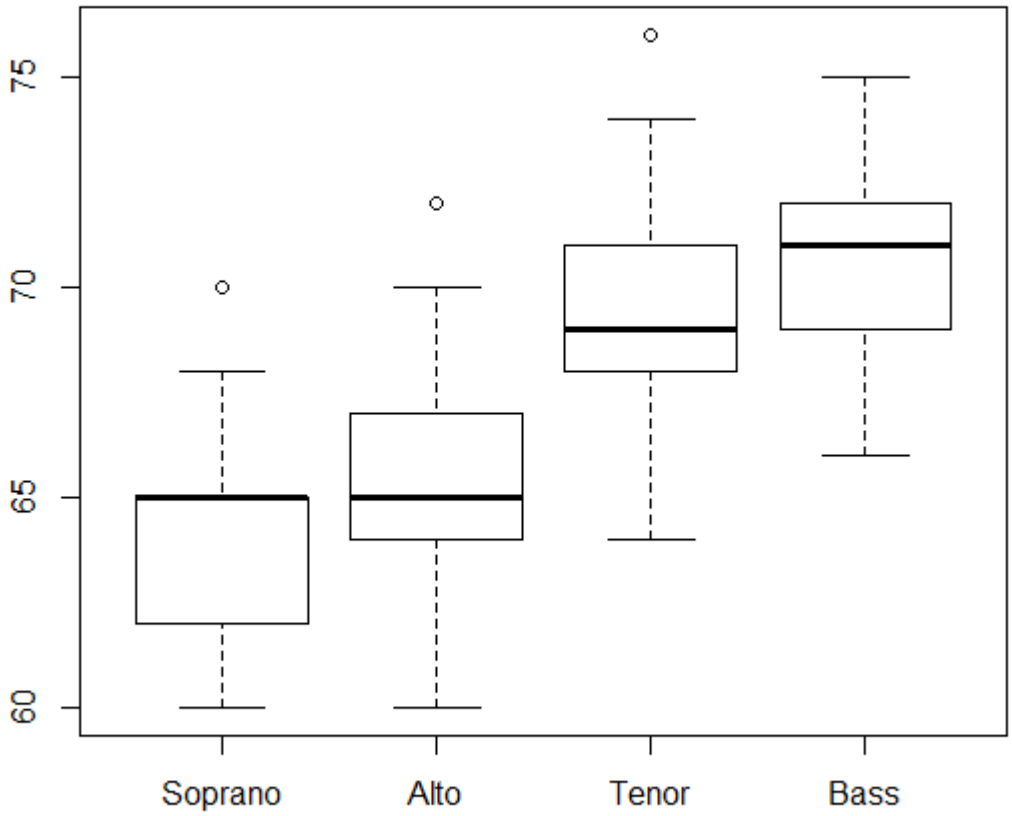
- From the Correlation coefficient values and CI boundaries obtained above, it can be said that the Correlation coefficient values lie in the Confidence interval calculated using Percentile Bootstrap method.
- The Correlation coefficient values show that they agree with the results obtained through scatter plot.
- TV and sales have stronger correlation(linear dependency) than Radio and sales.
- Bias values:
    - Sales and TV : -0.000521453
    - Sales and radio: -0.003385121
- Considering the values of Bias for both Sales, TV and Sales, radio, we can say that the values are very small and almost negligible.
- Bias (Sales and TV) < Bias (Sales and radio)
- Since the Bias values are negative, the average of all estimates done using 999 repetitions through percentile bootstrap seems to be quite small and negligible compared to obtained Correlation Coefficients.
- A statistic is **biased** if it is calculated in such a way that it is systematically different from the population parameter being estimated.
- Since both the bias are very small and negligible, it does not signify any effective difference from the Correlation coefficient being estimated.

2)

(a) An exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups.

Let's consider the distribution is normal and N is large and plotting box plots for examining the distribution of heights of singers in 4 groups:

```
data=read.csv(file="C:/Users/shash/Documents/R for Stats/Mini Projects/Miniproject 4/singers.csv",
header=TRUE, sep=",")                    #read the given data
names(data) <- NULL                      #to remove the headers since it contains the names
data                                     #display the data

boxplot(data$height,data$height2,data$height3,data$height4,
names=c("Soprano","Alto","Tenor","Bass")) #box plots for each of Soprano, Alto, Tenor and Bass
```



From the box plots,
Observations and exploratory analysis:
the height distributions are:

- Soprano: Left Skewed
- Alto: Right Skewed
- Tenor: Right Skewed

- Bass: Left Skewed

Observations:

For Right Skewed: mean>median

Left Skewed: median>mean

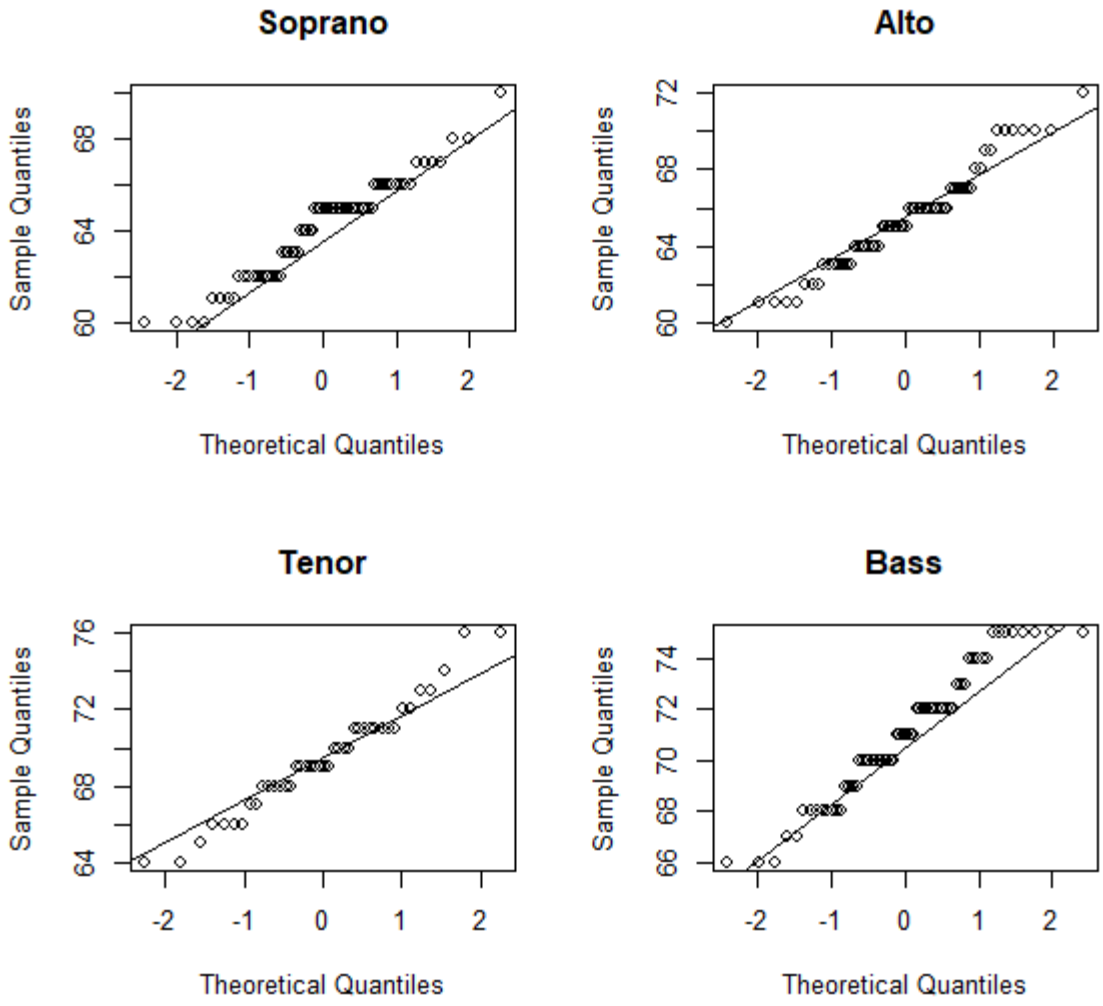
None of the data seems to be approx. normal distribution that can be inferred from the box plots.

- Consider median as a parameter to analyze: Bass singers have highest median followed by Tenor and then Alto singers whose median height is almost equal to that of Soprano singers.
- Outliers are observed in each of the singers data except that of Bass singers.
- 5 Point summary for each of the singers heights:

```
> summary(data$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   62.00   65.00   64.12   65.00   70.00
> summary(data$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   62.00   65.00   64.12   65.00   70.00
> summary(data$height2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  60.00   64.00   65.00   65.39   67.00   72.00       4
> summary(data$height3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   64.0    68.0    69.0    69.4    71.0    76.0      24
> summary(data$height4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  66.00   69.00   71.00   70.98   72.00   75.00       1
```

For verification let us consider drawing Normal QQ plots to find if the distribution of the heights of each singers is Normal distribution:

```
par(mfrow=c(2,2))
qqnorm(data$height,main = "Soprano")
qqline(data$height)
qqnorm(data$height2,main = "Alto")
qqline(data$height2)
qqnorm(data$height3,main = "Tenor")
qqline(data$height3)
qqnorm(data$height4,main = "Bass")
qqline(data$height4)
```



The four distributions don't seem to be similar.

- The IQR of each of the box plots is same i.e.; 3.000 but different mean, median and SDs
- In boxplot there is overlap between singers of different voice groups, but it is evident from the plot that singers of voice group Bass are taller than that of other groups.
- Also the above plot, shows that the distribution is normal. From above graph, we can see that there is a tie between certain number of singers in each group. But when we consider a single point value at each tie and plot a line, it would be a straight line that clearly says the distribution is normal. **All four distributions are normal.**

(b) Assumptions:

Here, we found n>30 for both Alto and Soprano which means that we may assume normality for testing the hypothesis.

Also here, sample size of Soprano ≠ alto, so we aren't assuming equal variances.

#Calculate confidence interval with Z-test at 95% confidence level

nalto=nrow(data$height2) #find number of data values

nsoprano=nrow(data$height) # find number of data values

varalto=(sd(data$height2))^2 #find variance

varsoprano=(sd(data$height))^2 #find variance

#Calculate the Confidence interval

U_limit=(mean(data$height2)-mean(data$height))+1.96*sqrt((varalto/nalto)+(varsoprano/nsoprano))
#(mean alto – mean of soprano)

L_limit=(mean(data$height2)-mean(data$height))-1.96*sqrt((varalto/nalto)+(varsoprano/nsoprano))
#(mean alto – mean of soprano)

U_limit

#result= 2.109801

L_limit

#result= 0.4219679

- It can be inferred that, the mean heights of Soprano and Alto singers are different.
- Both limits are greater than 0 and 0 does not lie between the limits of confidence interval, so we can say that the mean of Alto not equal to mean of Soprano.
- The mean heights of soprano and alto aren't equal. Also from the CI, we can say **mean of height of Alto > mean of height of Soprano singers**. (used in (c))
- From the box plot, mean(ht of Alto) is slightly greater than mean(ht of Soprano).

#Calcualte sample mean using 5 point summary calculated above

Mean(data$height2)

#65.3871

Mean(data$height)

#64.12121

The difference of means of heights of alto and soprano singers is: 1.26589

Suppose we consider applying null hypothesis on the above statement, i.e.;

H_0: Difference in mean of heights between alto and soprano singers is 0 i.e.; mean(ht of Alto) -mean(ht of soprano) =0

Alternate hypothesis:

H_1: Difference in mean of heights between alto and soprano singers is not equal to 0 i.e.; mean(ht of Alto) -mean(ht of soprano) ≠ 0

For a significant level of null hypothesis the p-value calculated using z-stat should be min(0.05), else we reject the NULL hypothesis.

#Calculate p-val:

z=(mean(data$height2)-mean(data$height))/(sqrt((varalto/nalto)+(varsoprano/nsoprano)))

p = 2*(1-pnorm(abs(z)))

p

#result= 0.003281887

Since the calculated p-value is less than 0.05, the null hypothesis doesn't hold any significance and hence we reject it.

(c) In our exploratory analysis using box plot of height distributions of soprano and alto singers, we found that Soprano singers had almost the same median height compared to Alto singers.

From the inferences and conclusions made on confidence intervals in (b), we found that **Soprano Singers had lower mean height compared to Alto Singers and the mean of height of Soprano singers is not equal to that of mean of height of alto singers**.