STATISTICAL METHODS FOR DATA SCIENCE

MINI PROJECT #4

SPRING 2018

NAMES OF GROUP MEMBERS:

ADRITA DUTTA(axd172930)

NEETHU ANTONY(nxa171330)

Contribution of team members:

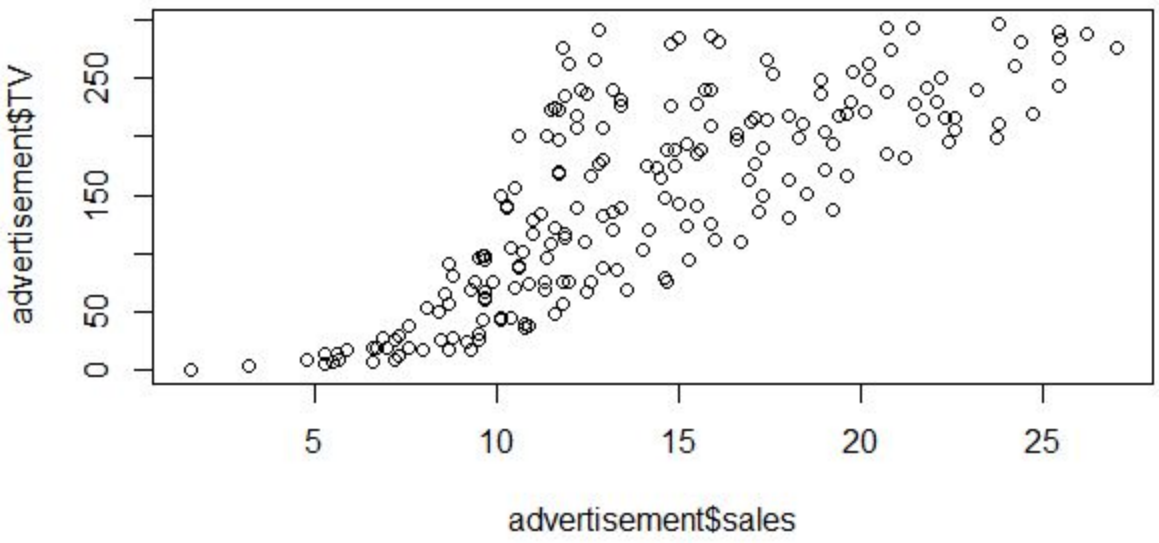
ADRITA: Question 1

NEETHU: Question 2

SECTION 1:

1Q.

ii)



iv)Correlation(Sales Vs TV):0.7822244

#Call to find estimate ,bootstrap bias and SE for correlations between sales and TV

->Bootstrap Statistics :

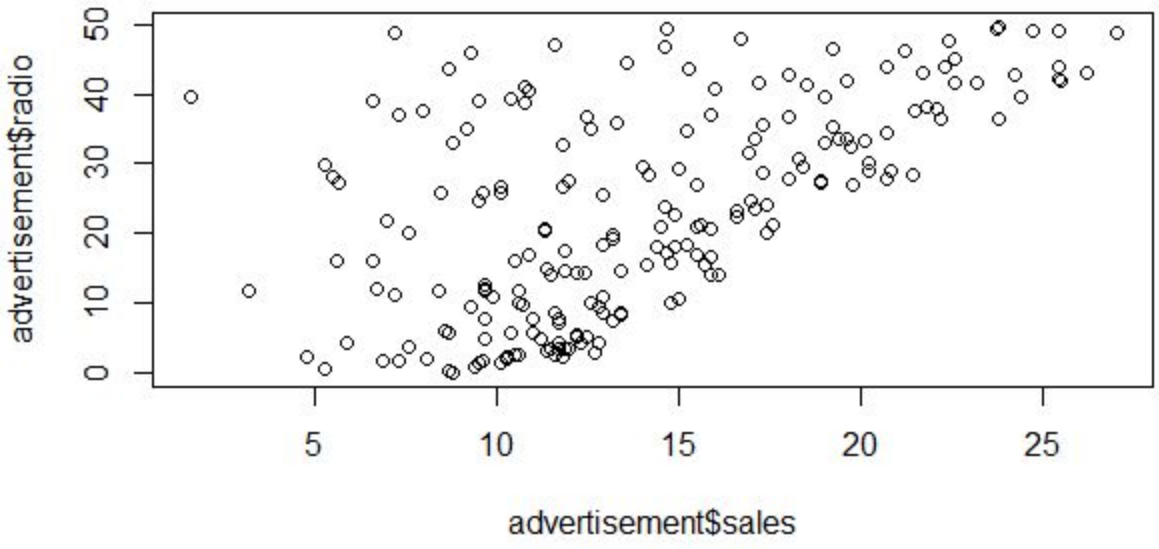
original bias std. error

t1* 0.7898647 -0.007755302 0.02662005

#95% confidence interval computed using percentile bootstrap for correlations between sales and TV

->0.7222676 0.8322526

iii)



v)Correlation(Sales Vs Radio):0.5762226

#Call to find estimate ,bootstrap bias and SE for correlations between sales and radio

Bootstrap Statistics :

	original	bias	std. error
t1*	0.5824666	-0.006761013	0.05353071

#95% confidence interval computed using percentile bootstrap for correlations between sales and radio

-> 0.4627779 0.6805818

Interpret the results:

In statistics, the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatter plot. The value of r is always between +1 and –1. To interpret its value, see which of the following values your correlation r is closest to:

- **Exactly –1.** A perfect downhill (negative) linear relationship
- **–0.70.** A strong downhill (negative) linear relationship
- **–0.50.** A moderate downhill (negative) relationship
- **–0.30.** A weak downhill (negative) linear relationship
- **0.** No linear relationship
- **+0.30.** A weak uphill (positive) linear relationship
- **+0.50.** A moderate uphill (positive) relationship
- **+0.70.** A strong uphill (positive) linear relationship
- **Exactly +1.** A perfect uphill (positive) linear relationship

From the results we can see that:

- Correlation(Sales Vs TV):

Population Correlation:0.7822244 which comes in the category (**+0.70.** A strong uphill (positive) linear relationship)

Sample Correlation:0.7898647 which also comes in the category(**+0.70.** A strong uphill (positive) linear relationship)

Confidence Interval of correlation:0.7222676 0.8322526 which is also in the category(**+0.70.** A strong uphill (positive) linear relationship)

Thus we can say with confidence that the correlation between Sales and TV is a strong uphill(Positive) Linear Relation.

- Correlation(Sales Vs Radio):

Population Correlation:0.5762226 which comes in the category (+**0.50.** A moderate uphill (positive) relationship)

Sample Correlation:0.5824666 which also comes in the category(+**0.50.** A moderate uphill (positive) relationship)

Confidence Interval of correlation:0.4627779 0.6805818 which includes 2 categories (+**0.30.** A weak uphill (positive) linear relationship) and (+**0.50.** A moderate uphill (positive) relationship)

Thus we can say that the correlation between Sales and TV is a uphill(Positive) Linear Relation, but we cannot make a confirm statement about whether it is a weak or moderate relation, it can be either of the two.

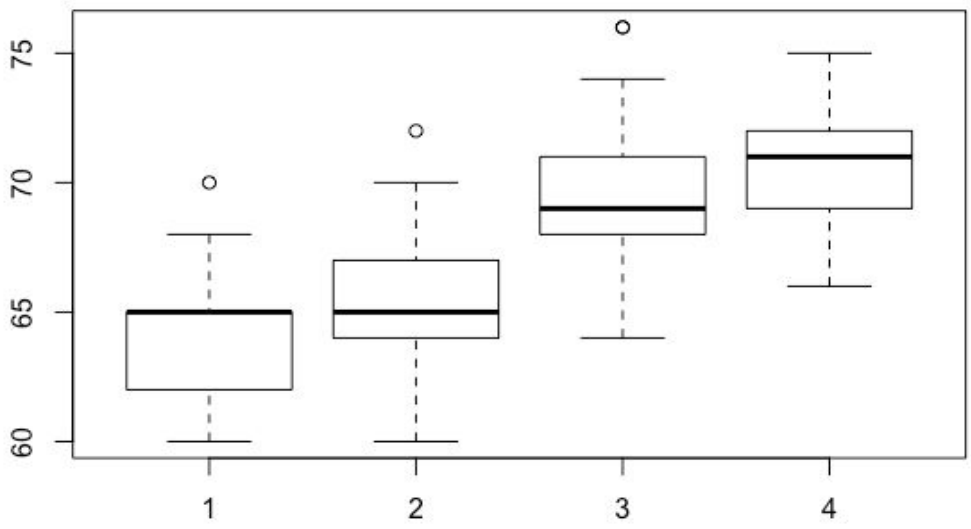
Thus overall we can see that both (Sales and TV) and (Sales and Radio) have uphill positive linear relations but (Sales and TV) has a stronger correlation than (Sales and Radio). This can also be observed in the scatterplots.

2Q.

(a).

(i). *Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you see.*

side-by-side boxplots of vSoprano, vAlto, vTenor, vBass:-

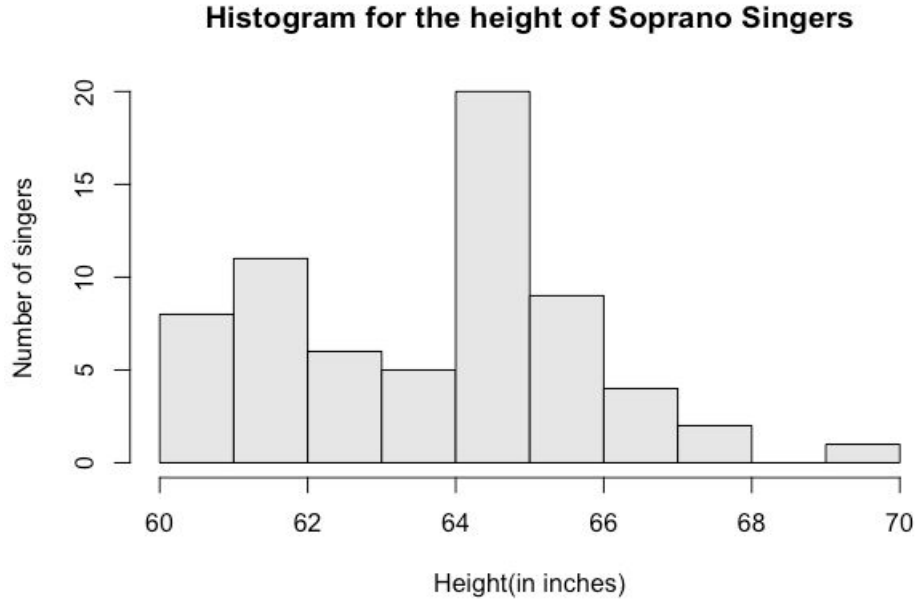


1. From the below side-by-side boxplot, we can infer that the Bass and Tenor Singers are taller compared to the Soprano and Alto Singers. Most probably, Soprano and Alto singers are female singers. And Tenor and Bass songs may be sung by male singers.
2. There is a noticeable difference in height of all the singers. However Soprano and Alto singers have clearly less height than Tenor and Bass singers. We can also say that Tenor and Bass singers belong to almost the same height group
3. Taller singers have a low pitch voice part and shorter singers have a high pitch voice part.Thus, Soprano singers have the least height and highest pitch and Bass singers are the tallest with the lowest pitch. We can represent the comparison of the heights between the groups as below:-

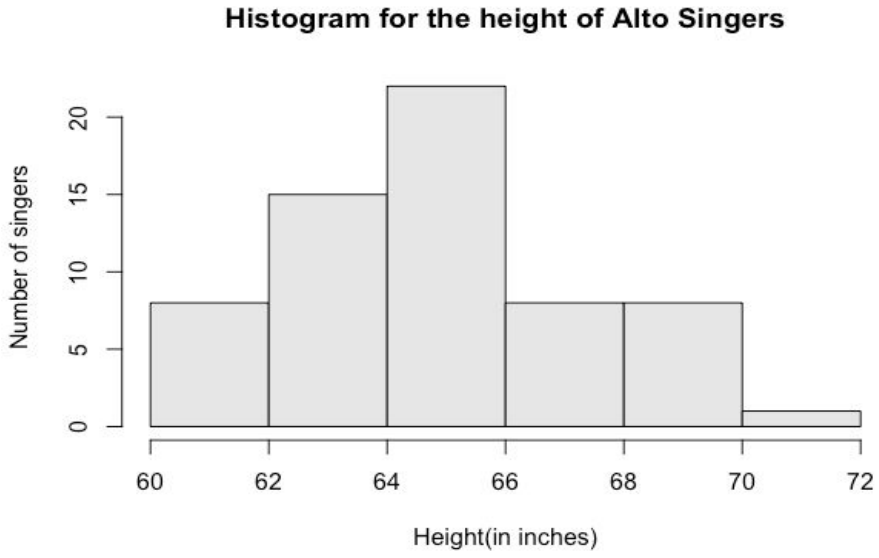
   Soprano < Alto < Tenor < Bass
4. We can also see that there is overlap between the distribution, but the distributions are different for each of the singers group.
5. It is clear that the median of Soprano and Alto singers in same.
6. Observing the IQR of the distributions, we can say that all the four distributions have almost same IQR and hence similar variability.

(ii) *Do the four distributions seem similar? Justify your answer.*

As we already mentioned, even though there is an overlap between the statistics of different singer groups, the distributions are not similar. This can be justified using the histograms of each distribution plotted below.

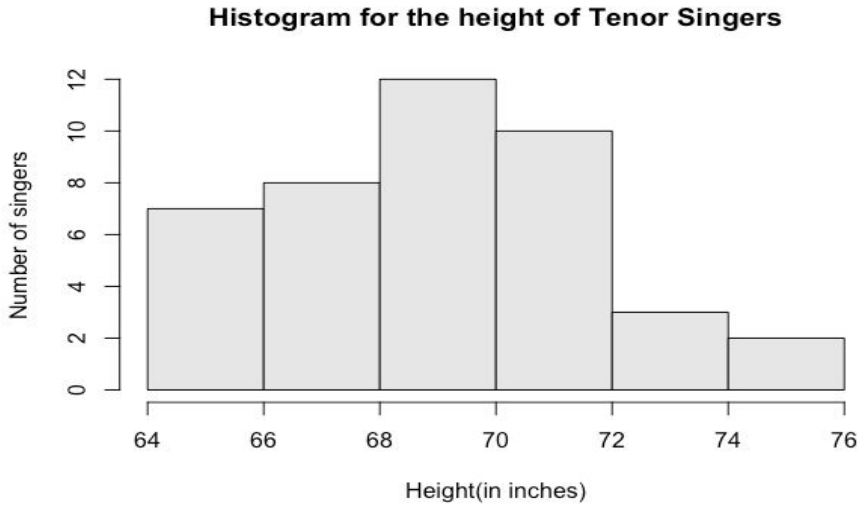


The above is the histogram for heights of the singers with the Soprano voice part. This seems to follow a bimodal distribution.

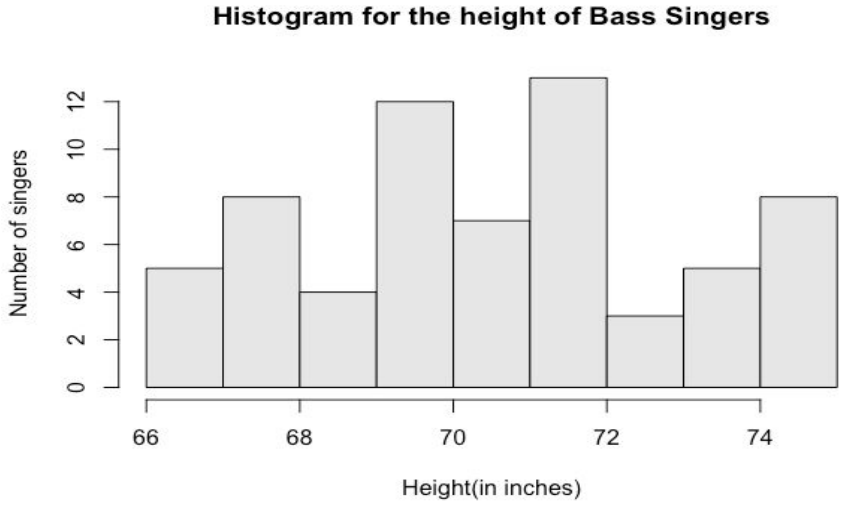


The above is the histogram for heights of the singers with the Alto voice part. This seems to

follow an approximately normal distribution. The distribution seems to be right skewed. This can be explained by the presence of a group of people with less than average height. And a few large values to the right. Could be due to the presence of males and females in the group.



The above is the histogram for heights of the singers with the Tenor voice part. This seems to follow an approximately normal distribution. The distribution seems to be right skewed. This can be explained by the presence of a group of people with less than average height. And a few large values to the right. Could be due to the presence of males and females in the group.



The above is the histogram for heights of the singers with the Bass voice part. This seems to follow an approximately normal distribution. The distribution seems to be right skewed. This can be explained by the presence of a group of people with less than average height. And a few large values to the right. Could be due to the presence of males and females in the group.

Generalisation:- The heights of the singers in general seems to follow a right skewed distribution. This can be explained by the presence of two groups within each sample that follow separate normal distributions with different mean and variance. Superimposing these two gives us the graphs we are observing.

(b). *Summarizing the data first*

First we read the singers.txt file in R and converted into a vector of voice part and height.

Using the R code "summary(singers)" we got the summary of the Singers as:-

height

Min. :60.0

1st Qu. :65.0

Median :67.0

Mean :67.3

3rd Qu. :70.0

Max. :76.0

voice.part

Soprano:66

Alto :62

Tenor :42

Bass :65

(i). *Is there any difference in the mean heights of Alto and Soprano singers? If yes, how much is the difference?*

Now we use the **subset** function to divide the data set into corresponding parts and find the summary of each of the voice parts.

Or else, you could use the by function by(singers$height, singers$voice.part, summary)
It is summarized as below:-

singers$voice.part: Soprano

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
60.00	62.00	65.00	64.12	65.00	70.00

---

singers$voice.part: Alto

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
60.00	64.00	65.00	65.39	67.00	72.00

---

singers$voice.part: Tenor

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
64.0	68.0	69.0	69.4	71.0	76.0

---

singers$voice.part: Bass

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
66.00	69.00	71.00	70.98	72.00	75.00

By looking at the mean heights, we can see that Soprano and Alto singers have less heights compared to Tenor and Bass singers.

Also, if we consider the difference in the mean heights of Soprano and Alto singers from the above summary information, we can say that the mean heights of Soprano singers is less than the Alto singers.

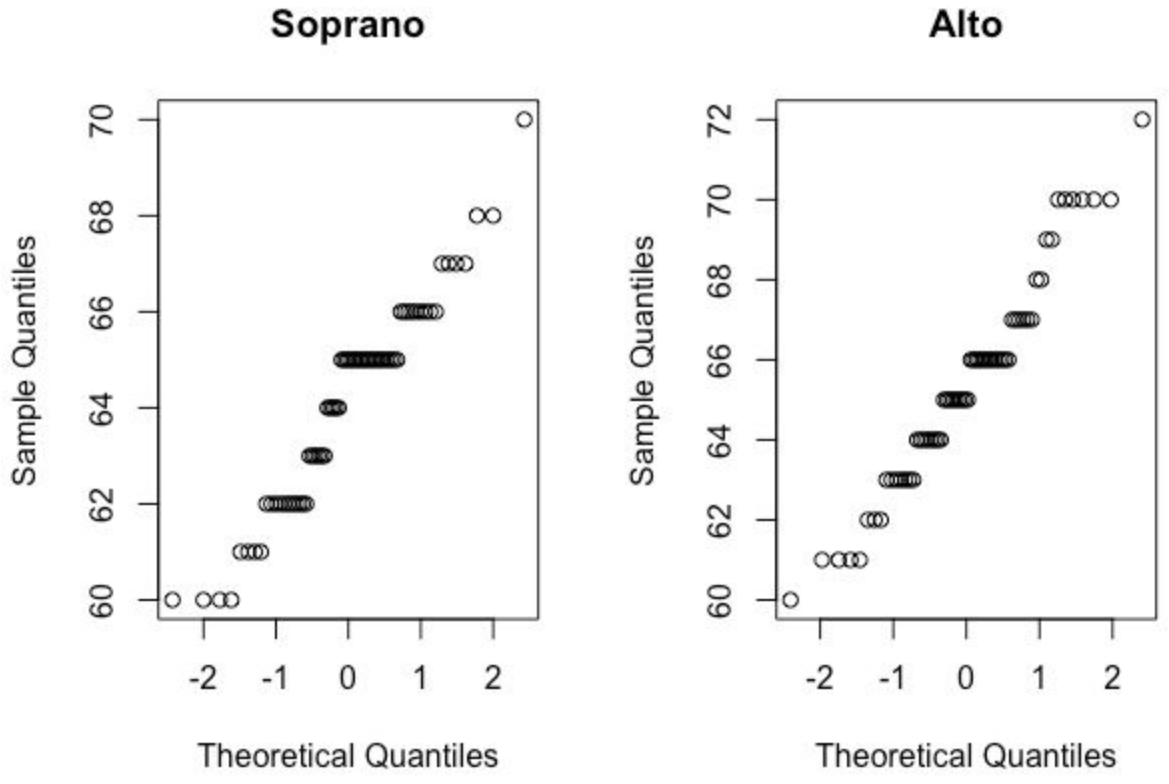
This can be proved more clearly by constructing an appropriate confidence interval for the difference in mean heights of Alto and Soprano singers.

(ii). *Is there any difference in the mean heights of Alto and Soprano singers? If yes, how much is the difference?*

Answer these questions by constructing an appropriate confidence interval. Clearly state the assumptions, if any, and be sure to verify the assumptions.

Assumption:- To find the confidence interval for mean difference, we will assume normality for the populations(here it is heights of Alto and Soprano singers).

Verification:- To verify if the assumption of normality is reasonable, we can check if we can model the data using a normal distribution. We can use normal Q-Q plots of heights of the singers, separately for Soprano and Alto. Figures below displays the same:-



From the above plots we can see that they have a "patchy" appearance and this is because of the ties in the data. But even then the points more or less form a straight line. Hence we can be sure of the fact that the normality assumption for the distribution of heights of Soprano and Alto singers in true.

Now, let's construct a 95% confidence interval for the difference in the mean heights of Alto and Soprano singers.

We are not taking any assumptions regarding the variances of both the population and hence "var.equal = FALSE" by Satterthwaite's approximation.

We can make use of the t.test function since we have the raw data.

The result of the t.test function is given below:-

```
> t.test( vAlto, vSoprano, alternative = "two.sided", conf.level = 0.95, var.equal = FALSE)

        Welch Two Sample t-test


data:  vAlto and vSoprano
t = 2.94, df = 118.32, p-value = 0.003948
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4132633 2.1185060
sample estimates:
mean of x mean of y
 65.38710  64.12121
```

From the above results we can see that the true difference in means is not equal for a 95% confidence interval for the mean differences.

Since the entire confidence interval is above 0, we can say that,

Mean of the heights of Alto singers - Mean of the heights of Soprano singers > 0

Hence, Mean of the heights of Alto singers > Mean of the heights of Soprano singers.

Also, the plausible values for the difference in the mean heights of Alto and Soprano singers is between 0.4132633 and 2.1185060 mostly.

(c). *How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?*

From our exploratory analysis in (a) we expected that Soprano singers have the least height and highest pitch and Bass singers are the tallest with the lowest pitch.

Soprano < Alto < Tenor < Bass

From (b), we concluded that,
Mean of the heights of Alto singers > Mean of the heights of Soprano singers

These two result goes hand in hand as the conclusion in (b) is what we expected from (a).

SECTION 2:

1Q.

```
# use install.packages("boot") to first install
# the package and then load it
library(boot);


#i)read file college.csv
advertisement<-read.csv("C:/Users/adrit/Desktop/utd/sem2/stats for
ds/project/PROJ4/Advertising.csv");


#ii)scatterplots of sales vs TV
plot(advertisement$sales,advertisement$TV)


#iii)scatterplots of sales vs radio
plot(advertisement$sales,advertisement$radio)


#iv)population correlation of sales vs TV
c1<-cor(advertisement$sales,advertisement$TV)


#finding bootstrap samples for correlations between sales and TV
x<-function(advertisement,X)
{
  y<-sample(advertisement$X,replace=T)
result1<-advertisement$TV[y]
result2<-advertisement$sales[y]
c3<-cor(result2,result1)
return(c3)
}
```

```
correlation.npar.boot<-boot(advertisement,x,R=999,sim='ordinary',stype="i")
#Call to find estimate ,bootstrap bias and SE for correlations between sales and TV
boot(data = advertisement, statistic = x, R = 999, sim = "ordinary", stype = "i")

#Percentile bootstrap method for correlations between sales and TV
sort(correlation.npar.boot$t)[c(25, 975)]



#v)population correlation of sales vs radio
c2<-cor(advertisement$sales,advertisement$radio)
#finding bootstrap samples for correlations between sales and TV
x1<-function(advertisement,X)
{
  y1<-sample(advertisement$X,replace=T)
  result2<-advertisement$sales[y1]
  result3<-advertisement$radio[y1]
  c4<-cor(result2,result3)
return(c4)
}
correlation1.npar.boot<-boot(advertisement,x1,R=999,sim='ordinary',stype="i")

#Call to find estimate, bootstrap bias and SE for correlations between sales and radio
boot(data = advertisement, statistic = x1, R = 999, sim = "ordinary", stype = "i")

#Percentile bootstrap method for correlations between sales and radio
sort(correlation1.npar.boot$t)[c(25, 975)]
```

2Q.

(a).

(i).

```
singers<-read.table("/Users/neethuantony/Documents/MS-2nd
SEM/Stat/Projects/MiniProject4/singer.txt",header = TRUE, sep="," )

boxplot(vSoprano, vAlto, vTenor, vBass)
```

(ii).

```
Soprano<-subset(singers,voice.part=="Soprano",select=height)

Alto<-subset(singers,voice.part=="Alto",select=height)

Tenor<-subset(singers,voice.part=="Tenor",select=height)

Bass<-subset(singers,voice.part=="Bass",select=height)


vSoprano <- Soprano[,1]

hist(vSoprano,probability=F,col=gray(.9),xlab="Height(in inches)",ylab="Number of
singers",main="Histogram for the height of Soprano Singers")


vAlto <- Alto[,1]

hist(vAlto, probability=F,col=gray(.9),xlab="Height(in inches)",ylab="Number of
singers",main="Histogram for the height of Alto Singers")


vTenor <- Tenor[,1]

hist(vTenor, probability=F,col=gray(.9),xlab="Height(in inches)",ylab="Number of
singers",main="Histogram for the height of Tenor Singers")


vBass <- Bass [,1]

hist(vBass, probability=F,col=gray(.9),xlab="Height(in inches)",ylab="Number of
singers",main="Histogram for the height of Bass Singers")
```

(b). summary(singers)

(i).

summary(Soprano)

summary(Alto)

summary(Tenor)

summary(Bass)

OR

by(singers$height, singers$voice.part, summary)

(ii).

par(mfrow = c(1, 2))

qqnorm(vSoprano, main = "Soprano")

qqnorm(vAlto, main = "Alto")

t.test(vAlto, vSoprano, alternative = "two.sided", conf.level = 0.95, var.equal = FALSE)

Note:- var.equal = FALSE --- Satterthwaite's approximation.