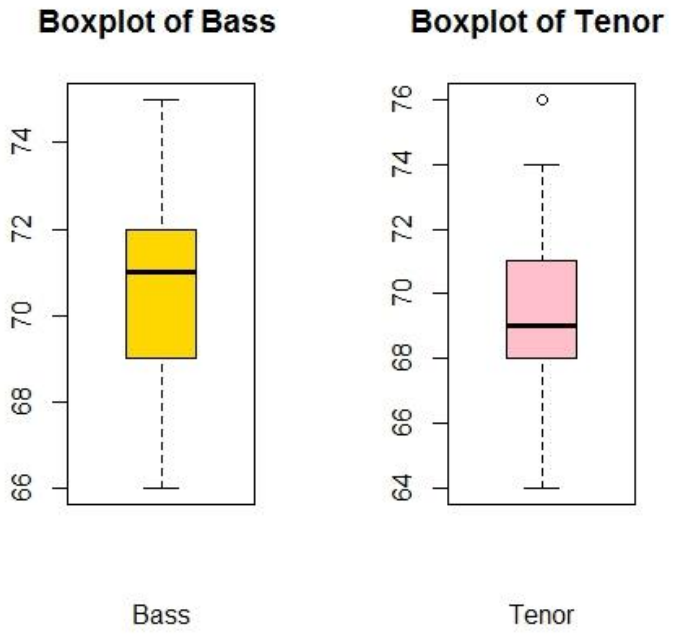
MINI PROJECT 4

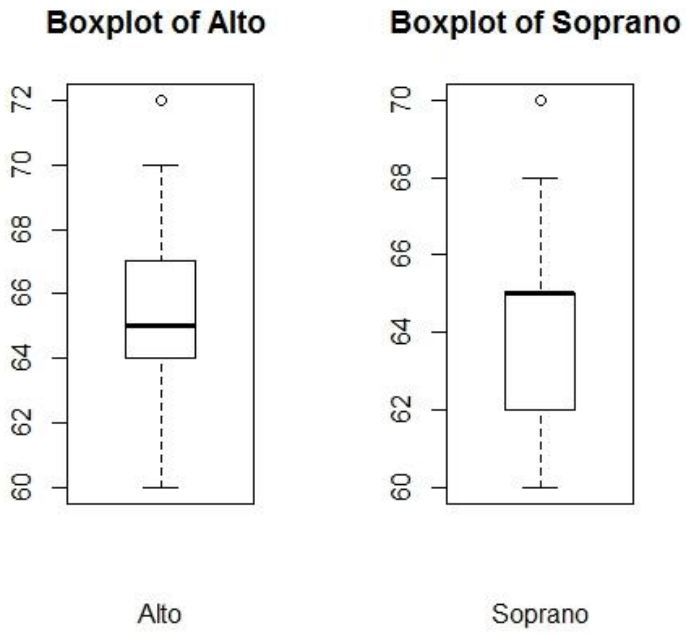
By- Raghav Mathur (rxm162130), Deepak Shanmugam (dxs161930)

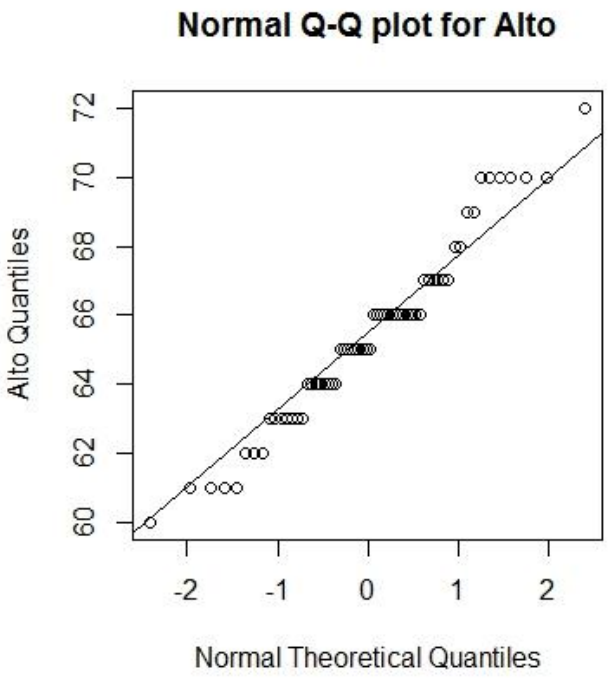
Contribution to the Project: Each member contributed equally to the project. We worked together to analyse the distributions, determine CI, and formulating as a test of hypothesis.

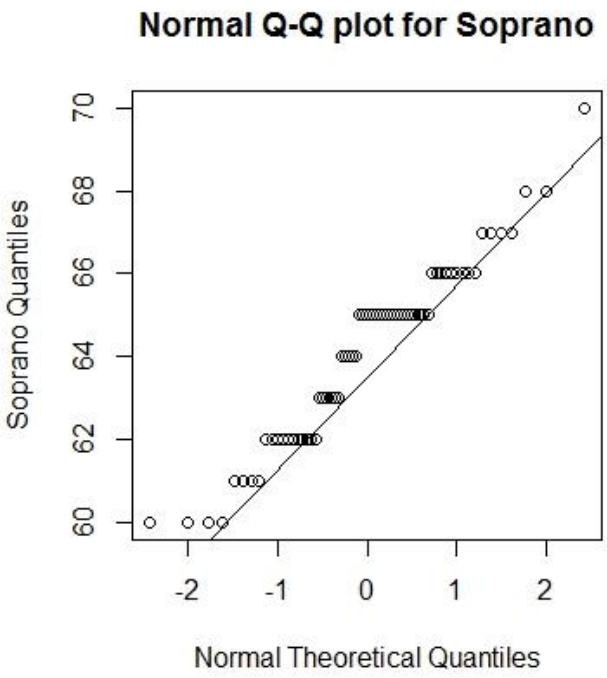
EXERCISE 1:

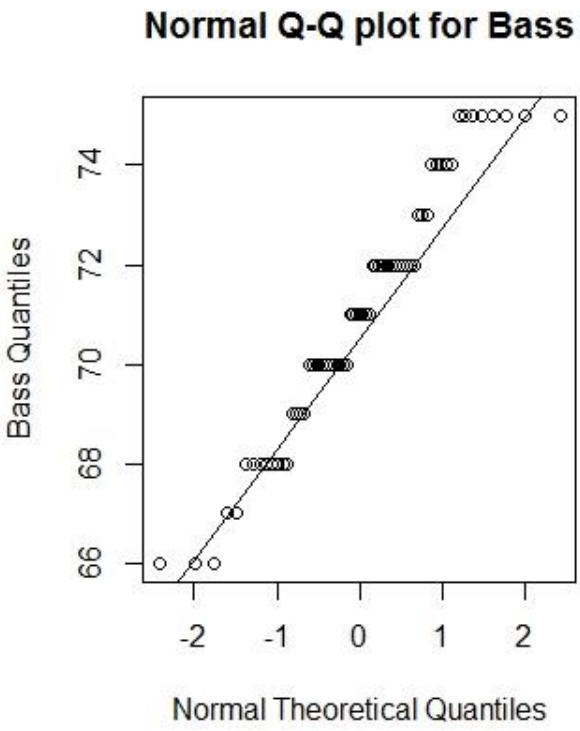
Section 1:
(a)

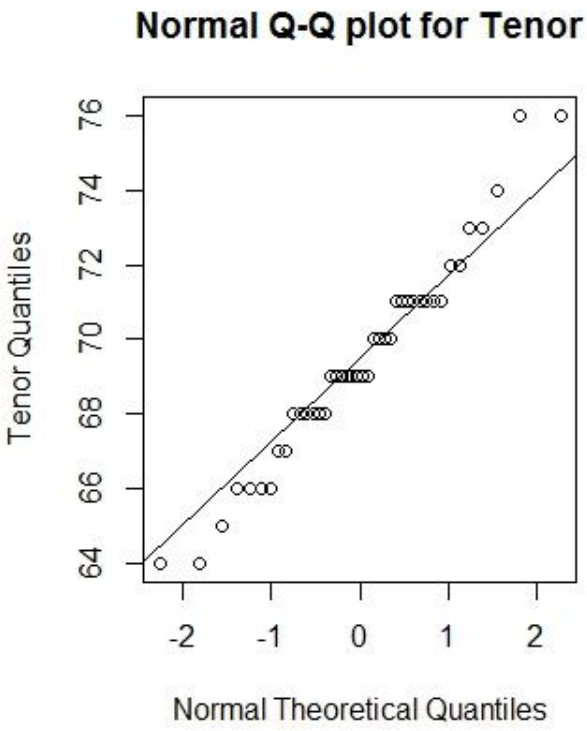


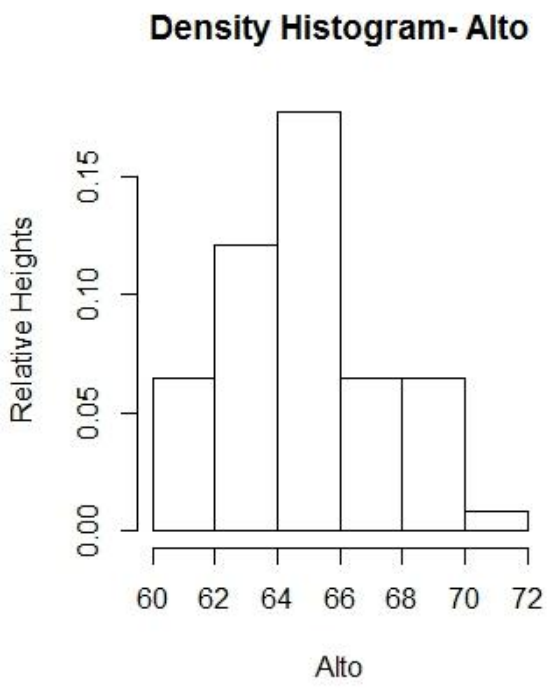


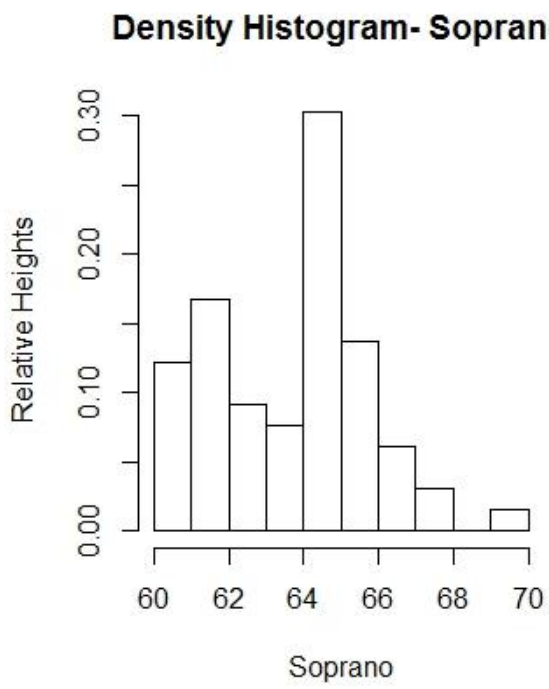

Normal Q-Q plot for Alto
Alto Quantiles
Normal Theoretical Quantiles
60
62
64
66
68
70
72
-2
-1
0
1
2

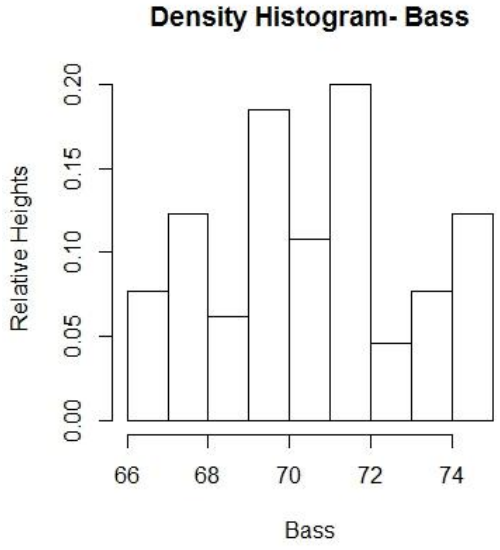

Normal Q-Q plot for Soprano
Soprano Quantiles
Normal Theoretical Quantiles
60
62
64
66
68
70
-2
-1
0
1
2

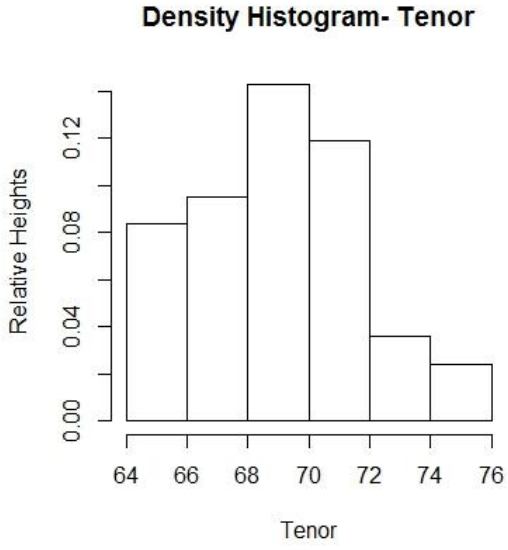

Normal Q-Q plot for Bass
Bass Quantiles
Normal Theoretical Quantiles
66
68
70
72
74
-2
-1
0
1
2


Normal Q-Q plot for Tenor
Tenor Quantiles
Normal Theoretical Quantiles
64
66
68
70
72
74
76
-2
-1
0
1
2









COMMENTS ON THE DISTRIBUTIONS:

Bass: Though the points seem to be deviated from the QQline but considering the size of the population (65) it can be approximated as normal. Also, the extent of deviation is not that high to be considered as not normal. The histogram showing the relative heights of Bass singers have some discreteness but can be considered somewhat normal w.r.t. the shape and size of population (By Central Limit theorem). From the boxplot, we can see that the median is greater than the mean.

Tenor: Majority of points in QQplot have reasonable normal distribution as they lie very close to the QQline. The upper tail points have greater variability(outliers). The histogram showing the relative heights of Tenor singers have some discreteness appears to have normal distribution but it is slightly right skewed (which is evident from boxplot) due to presence of outlier. From the boxplot, we can see that the mean is greater than the Median.

Alto: Majority of points in QQplot have reasonable a normal distribution as they lie very close to the QQline. The upper tail points have greater variability. From the boxplot, we can see that the mean is greater than the Median. Also, we can see that there are outliers at the upper whisker. From the histogram, we can see that it is slightly right skewed.

Soprano: Though the points seem to be deviated from the QQline but considering the size of the population (65) it can be approximated as normal. Also, the extent of deviation is not that high to be considered as not normal. Also, the upper tail points have greater variability. From the boxplot, we can see that the median is greater than the mean. Also, we can see that there are outliers at the upper whisker. It is also right skewed based on the histogram and boxplot.

SIMILARITY:

The distributions seem somewhat similar as they are normal but have differences with respect to their mean heights, presence of outliers and skewness, which has been described above. Mean of population increase in the order of Soprano, Alto, Tenor and Bass.

(b)
The hypothesis can be setup as follows:
H_0 = Mean height of Bass singers is equal to Mean height of Tenor singers
H_1 = Mean height of Bass singers is greater than mean height of Tenor singers

Assumptions:
The tenor distribution seems to be normal. But, bass distribution is not perfectly normal but considering the large sample size which is 65, it can be approximated as normal distribution (By Central Limit theorem). So, we used the t-test with unknown population variance to test the hypothesis.
The result turns out to be

```
                    Welch Two Sample t-test
data:  bass$height and tenor$height
t = 2.9746, df = 81.159, p-value = 0.001932
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6961488       Inf
sample estimates:
mean of x mean of y
 70.98462 69.40476
```

Since, p-value is less than alpha (Level of Significance) = 0.05, we reject the H_0 hypothesis and accept the H_1 hypothesis. It can be verified using CI results.

```
> CI
[1] 0.5388742 2.6208328
```

As 0 is not included in the CI and the interval is positive, Mean height of Bass singers is greater than Mean height of Tenor singers.

(c)
In part (a) we expected the Mean height of Bass singers to be greater than Mean height of Tenor singers which is proven to be correct in part (b) by accepting the hypothesis H_1.

Section 2: (R Code)

(a)

```
singer <- read_csv("C:/Users/mastr/Desktop/singer.txt")

bass=subset(singer, singer$voice.part =="Bass" )
alto=subset(singer, singer$voice.part =="Alto" )
tenor=subset(singer, singer$voice.part =="Tenor" )
Soprano=subset(singer, singer$voice.part =="Soprano" )

summary(singer)
summary(bass)
summary(tenor)

##Boxplots
par(mfrow=c(1,2))
boxplot(bass[,1],range=1.5, main = "Boxplot of Bass ",xlab = "Bass")
boxplot(tenor[,1],range=1.5,main = "Boxplot of Tenor ",xlab = "Tenor")
boxplot(alto[,1],range=1.5,main = "Boxplot of Alto ",xlab = "Alto")
boxplot(Soprano[,1],range=1.5,main = "Boxplot of Soprano ",xlab = "Soprano")

##Histograms
par(mfrow=c(1,2))
hist(bass$height,freq=FALSE,xlab="Bass", ylab=" Relative Heights", main="Density Histogram- Bass")
hist(tenor$height, freq=FALSE,xlab="Tenor", ylab="Relative Heights", main="Density Histogram- Tenor")
hist(alto$height, freq=FALSE,xlab="Alto", ylab="Relative Heights", main="Density Histogram- Alto")
hist(Soprano$height, freq=FALSE,xlab="Soprano", ylab="Relative Heights", main="Density Histogram-
Soprano")

##QQ Plots
par(mfrow=c(1,2))
qqnorm(bass$height,main = "Normal Q-Q plot for Bass", xlab = "Normal Theoretical Quantiles", ylab =
"Bass Quantiles")
qqline(bass$height)
qqnorm(tenor$height,main = "Normal Q-Q plot for Tenor", xlab = "Normal Theoretical Quantiles", ylab =
"Tenor Quantiles")
qqline(tenor$height)
qqnorm(alto$height,main = "Normal Q-Q plot for Alto", xlab = "Normal Theoretical Quantiles", ylab = "Alto
Quantiles")
qqline(alto$height)
qqnorm(Soprano$height,main = "Normal Q-Q plot for Soprano", xlab = "Normal Theoretical Quantiles",
ylab = "Soprano Quantiles")
qqline(Soprano$height)
```

(b)

```
x.bar=mean(bass$height)
y.bar=mean(tenor$height)
alpha=0.05
t.test(bass$height, tenor$height, alternative = "greater", conf.level = (1-alpha))
#Performs one and two sample t-tests on vectors of data
#reject H0 since pval < alplha, we reject H0
CI=(x.bar-y.bar) +c(-1,1)*qnorm(1-alpha/2)*denom
```

EXERCISE 2:

Section 1:

(a)

The hypothesis can be setup as follows:

H_0 = Mean of a normal population is 10.

H_1 = Mean of a normal population is greater than 10.

(b) We used One sample t-test for a normal population with unknown variance.

Test Statistic: $T = (\bar{x} - \mu_0) / (S/\sqrt{n})$

Null Distribution: If H_0 is true, $T \sim t_{n-1}$ (n-1 degree of freedom) where n=20

(c) The observed value comes out to be:

t.stat = -1.974186

(d) The p value comes out to be:

Pvalue = 0.9684606

(e) The p value using Monte Carlo comes out to be:

MC.pvalue= 0.9667

The value comes to be similar and greater than level of significance (alpha)

(f) Since Pvalue and MC.pvalue both are greater than alpha (5% level of significance), we accept H_0.

Section 2: (R Code)

```
mu.zero=10
x.bar=9.02
x.sigma=2.22
n.x=20
se=x.sigma/sqrt(n.x)
#test statistic
t.stat=(x.bar-mu.zero)/se
#  random generation for the t distribution with df degrees of freedom
pvalue=1-pt(t.stat, df=n.x-1)

#Monte Carlo Simulation
randvalues = replicate (10000, rt(1,df=n.x-1))
#MC.pvalue
mean(randvalues>=t.stat)
#critical.point=0.05
# since p-value and pvalue1 both are greater than critical point, we accept H0
```

EXERCISE 3:

Section 1:

(a)

μ_x = Mean credit limit of credit cards issued in January 2011 ($2635)

μ_y = Mean credit limit of credit cards issued in May 2011 ($2887)

CI = -302.8289 -201.1711

Result: The difference between μ_x and μ_y will always lie in the CI range = -302.8289 -201.1711.

As 0 is not included in the CI and the interval is negative, Mean credit limit of credit cards issued in January 2011 is less than Mean credit limit of credit cards issued in May 2011.

(b)

H_0 = Mean credit limit of credit cards issued in January 2011 is equal Mean credit limit of credit cards issued in May 2011

H_1 = Mean credit limit of credit cards issued in January 2011is less than Mean credit limit of credit cards issued in May 2011

We chose the test statistic as Two-sample test for difference in means of the non-normal populations, considering as large sample populations.

Alpha= 0.05

p-value= 1.274297e-22 (which is less than alpha)

Since p-value is less than alpha (5% level of significance), we reject the null hypothesis (H_0)

Section 2: (R Code)

```
mu.x=2635
mu.y=2887
sd.x=365
sd.y=412
n.x=400
n.y=500

alpha=0.05
se=sqrt(((sd.x^2)/n.x) + ((sd. y^2)/n.y))

CI= (mu.x-mu.y)+c(-1,1)*qnorm(1-alpha/2)*se
# -302.8289 -201.1711

delta=0
z.stat=(mu.x-mu.y-delta)/se
#pvalue
pnorm(z.stat)
```