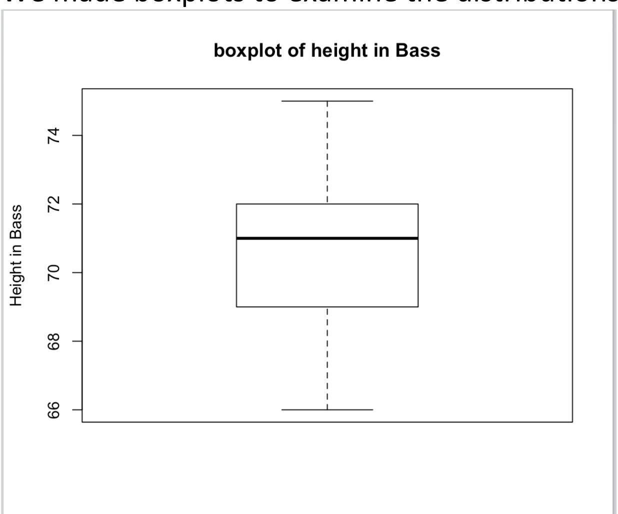
Mini Project 4 Name: You Jia, UTD-ID: 2021342433

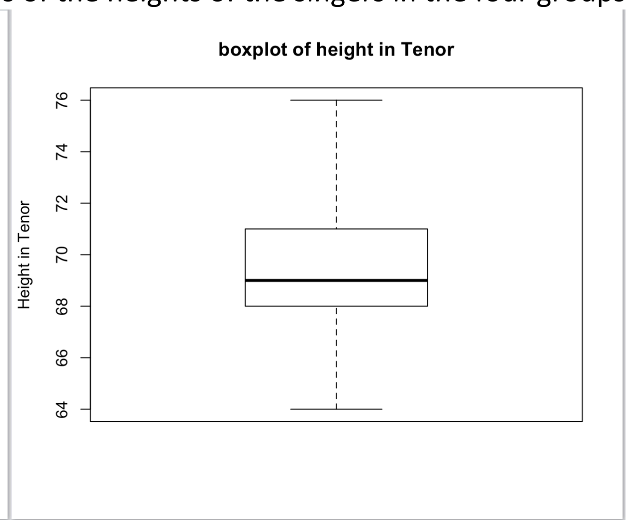
Exercise 1

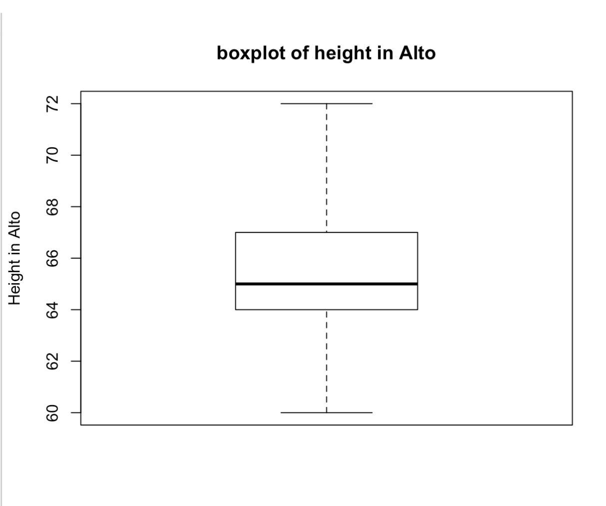
a.

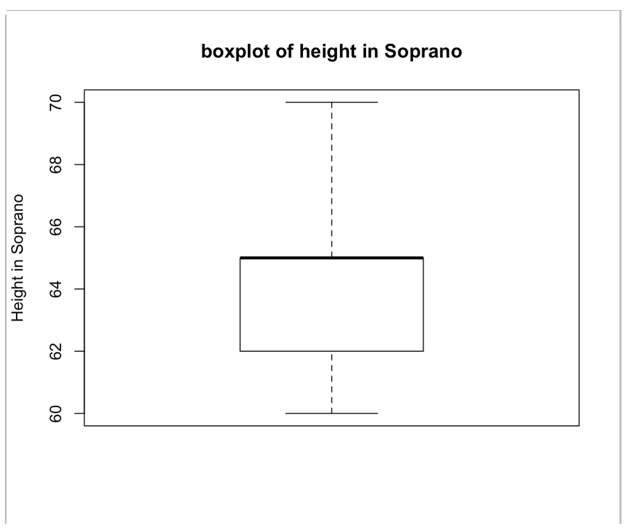
```
getwd()
setwd("/Users/youjia/Desktop")
getwd()
sing <- read.table(file = "singer.txt",fill = FALSE,sep = ",",header = TRUE)
# a) make boxplot for the four group data
Sop <- sing[sing$voice.part=="Soprano", ]
Alto <- sing[sing$voice.part=="Alto", ]
Bass <- sing[sing$voice.part=="Bass", ]
Tenor <- sing[sing$voice.part=="Tenor", ]
boxplot(Sop$height, ylab = "Height in Soprano", main = "boxplot of height in Soprano", range = 0)
boxplot(Alto$height, ylab = "Height in Alto", main = "boxplot of height in Alto", range = 0)
boxplot(Bass$height, ylab = "Height in Bass", main = "boxplot of height in Bass", range = 0)
boxplot(Tenor$height, ylab = "Height in Tenor", main = "boxplot of height in Tenor", range = 0)
qqnorm(Bass$height, main = "qqplot of height in Bass singers")
qqline(Bass$height)
qqnorm(Tenor$height, main = "qqplot of height in Tenor singers")
qqline(Tenor$height)
```

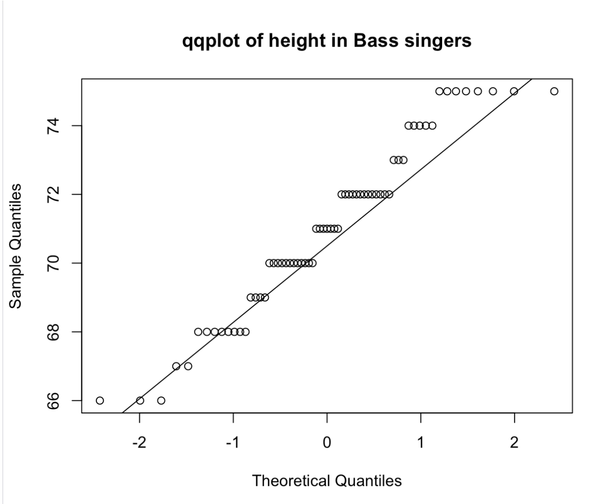
We made boxplots to examine the distributions of the heights of the singers in the four groups.

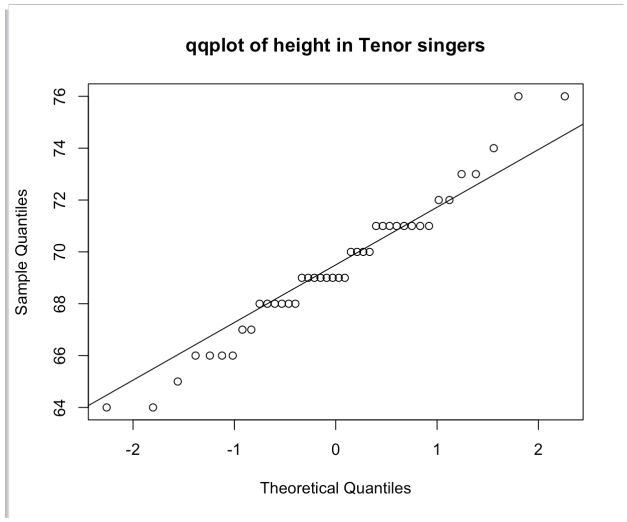












From above four boxplots, we can see that these four distributions are different. The distribution for Bass and Soprano are left-screwed while Tenor and Alto are right-screwed. And The height of Bass singers and Tenor singers are normally distributed.

b.

```
# b) hypothesis test, The height of  Bass singers and Tenor singers are normally distributed
# we perform an appropriate 5% level test to see if The Bass singers tend to be taller
# than Tenor singers, the sample size is large, delta = Tenor - Bass,
# H0: delta = 0 vs H1: delta < 0
# We use mean of height of Tenor singers as u1
u1 <- mean(Tenor$height)  # u1 = 69.40476
s1 <- sd(Tenor$height)
n1 <- length(Tenor$height)
## u2 is the height of Bass singer
u2 <- mean(Bass$height)
s2 <- sd(Bass$height)
n2 <- length(Bass$height)
zobs = (u1-u2)/sqrt(s1^2/n1+s2^2/n2)  # We get zobs = -2.974561
# compute p-value
deg <- (s1^2/n1+s2^2/n2)^2/(s1^4/(n1^2*(n1-1))+s2^4/(n2^2*(n2-1)))
pval <- pt(zobs,deg) # pval = 0.001931636
# Since pval <= 0.05, reject H0
# There is statistically significant evidence that Base singers tend to be taller than Tenor singers.
```

c.

```
# c) Conclusion in b is consistent with what I can see in a. Conclusion in b shows that Base singer tend
# to be taller than Tenor singers. From boxplot in a, we can see that it's consistent.
```

Exercise 2

```r
# Exercise 2
# a) Set up the null and alternative hypotheses, H0: mean = 10 VS H1: mean > 10

# b) choose t-test, the test statistic is the 20 examples,
# the null distribution of the test statistic is t(20-1) distribution.

# c) compute the observed value of the test statistic
X<-9.02
s<-2.22
tobs<-(X-10)/(s/sqrt(20))  # here we have tobs is approximately -1.974186

# d) compute the p-value of the test using the usual way
pvalue <- 1-pt(tobs,19)  # here we can get pvalue = 0.9684606

# e) estimate the p-value of the test using Monte Carlo simulation
# Use rt function to make random draws from the t(20-1) distribution
# then calculate the percentage of these draws is greater than tobs
# which is the estimation of the p-value.
MC <- rt(999,19)
pvalMC<- length(which(MC>tobs))/999  # here we can get pvalMC = 0.96997
# The result of d and e are very close.

# f) Here, p= 0.968 is larger than α=0.05.
# So we accepted H0 which means the mean of a normal population is 10.
```

Exercise 3

```r
# Exercise 3
# a) We construct an appropriate 95% confidence interval for the difference in mean credit limits
# of all credit cards issued in January 2011 and in May 2011 using the average credit limits and
# the sample standard deviations of these two samples. The 95% confidence interval for the difference
# is: [-302.8289, -201.1711]. This means that the mean credit limit of all credit cards is increased.
xbar = 2635
ybar = 2887
Sx = 365
Sy = 412
n1 = 400
n2 = 500
CI <- ((xbar - ybar)+c(-1,1)*(qnorm(1-0.025)*sqrt(Sx^2/n1+Sy^2/n2))) # here, CI=(-302.8289,-201.1711)

# b) We performed an appropriate 5% level test to see if the mean credit limit of all credit cards
# issued in May 2011 is greater than the same in January 2011.
# we made H0 be Jan-May=0 be and H1 be Jan-May<0. delta = Jan - May
# H0: delta = 0 VS H1: delta < 0
tobs <- (xbar-ybar)/sqrt(Sx^2/n1+Sy^2/n2)
v <- (Sx^2/n1+Sy^2/n2)^2/(Sx^4/(n1^2*(n1-1))+Sy^4/(n2^2*(n2-1)))
pvalue<-pt(tobs,v) # here pvalue = 1.395927e-21
# Since p is much less than α=0.05, we rejected H0
# which means that the mean credit limit of all credit cards issued in May 2011 is
# greater than the same in Jan 2011.
```