Assignment 1

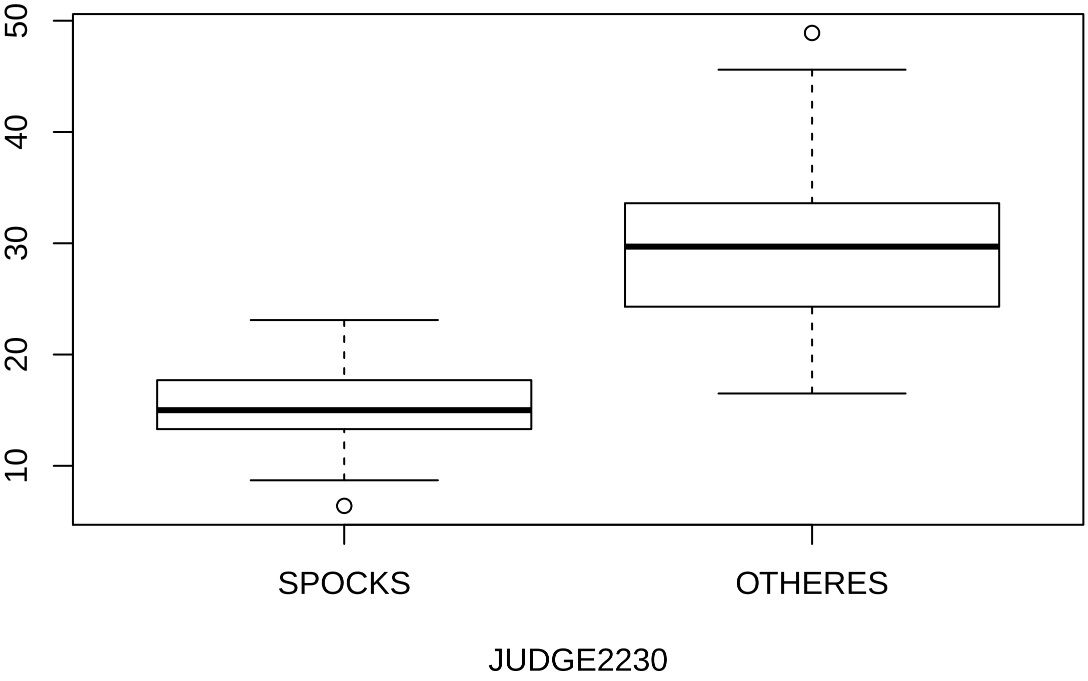
Last name: Du
First name: Min
Student ID: 1002602230

Question1

(a)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.40   13.30   15.00   14.62   17.70   23.10 
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   16.50   24.30   29.70   29.49   33.60   48.90 
```



According to above summaries of each group.

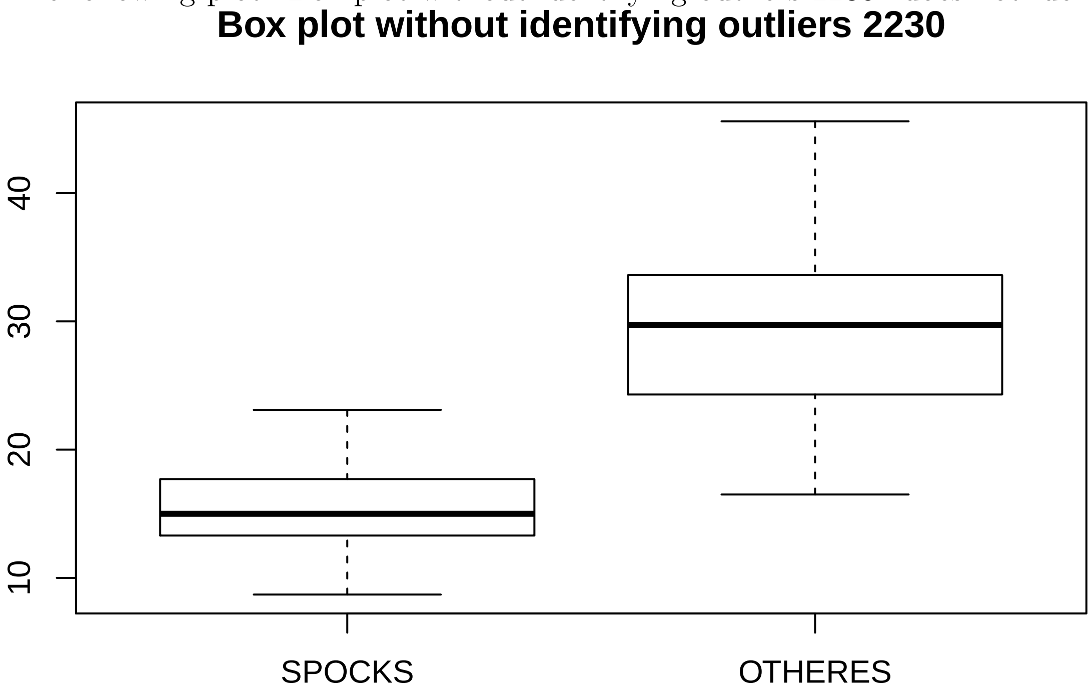
For SPOCKS, IQR = Q_3 - Q_1= 17.7 - 13.3 = 4.4 . Outliers are identified if > 24.3 (larger than Q_3 + 1.5IQR = 17.7 + 6.6), or < 6.7 (smaller than Q_1-1.5 IQR = 13.3 - 6.6). According to the Summary and Boxplot“JUDGE2230”, The SPOCKS boxplot has a small outlier=6.40, which is below 6.7.

For the others, IQR = Q_3 - Q_1 = 33.6 - 24.3 = 9.3. Outliers are identified if > 47.55 (larger than Q_3 + 1.5IQR = 33.6 + 13.95), or < 10.35 (smaller than Q_1-1.5 IQR = 24.3 - 13.95).

According to the Summary and Boxplot“JUDGE2230”, there is a large outlier=48.9, which is above 47.55.

(b)

The following plot “Box plot without identifying outliers 2230” does not identify outliers.



(c)

The regular schematic box plot includes all data points, conceal outliers. The whiskers start at minimum value, end at maximum value.

The modified box plot plot outliers as isolated points, the whiskers are only extended to largest(smallest) values that are not outliers.

Question2

(a)

It was based on oservational study. Observational studies do not feature random selection, so generalizing from the results of an observational study to a larger population can be a problem. This set of data only contain people from a class, students all around age 20, generalizing the result to all Female and Male may be a problem.

(b)

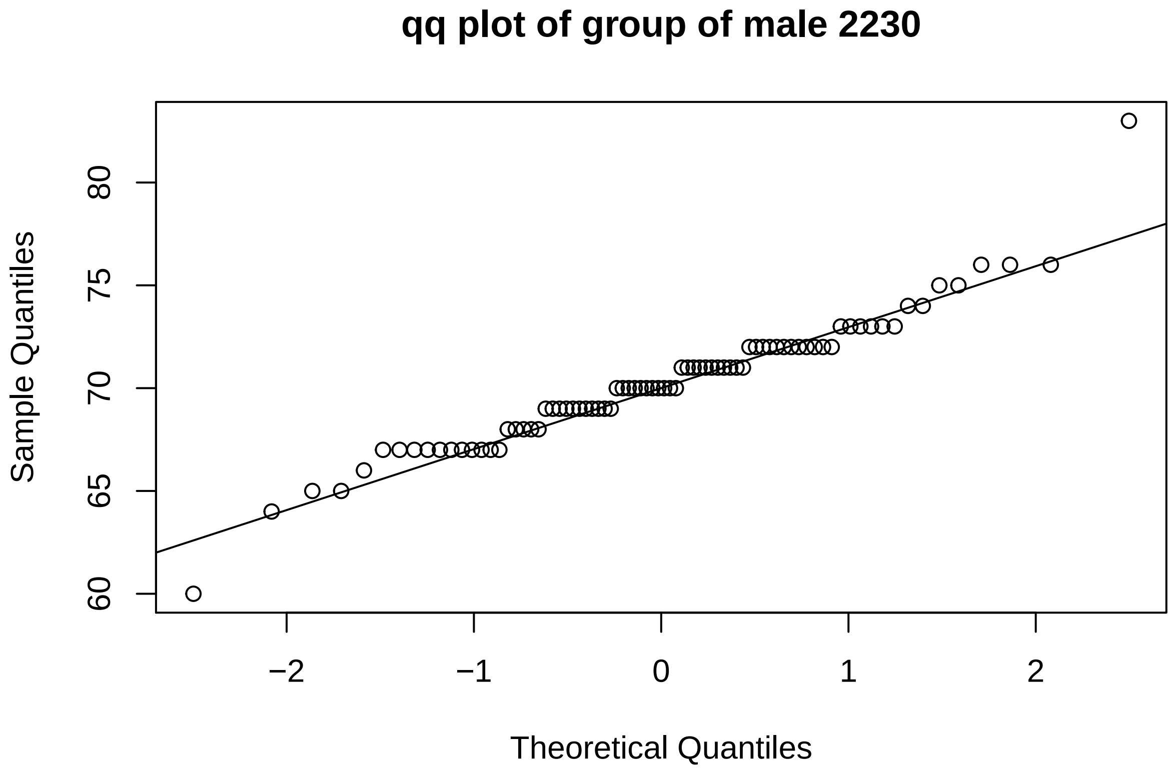
Gender is a categorical variable, there are 2 levels: Female and Male. Id is also a categorical variable, it has 166 levels.

(c)

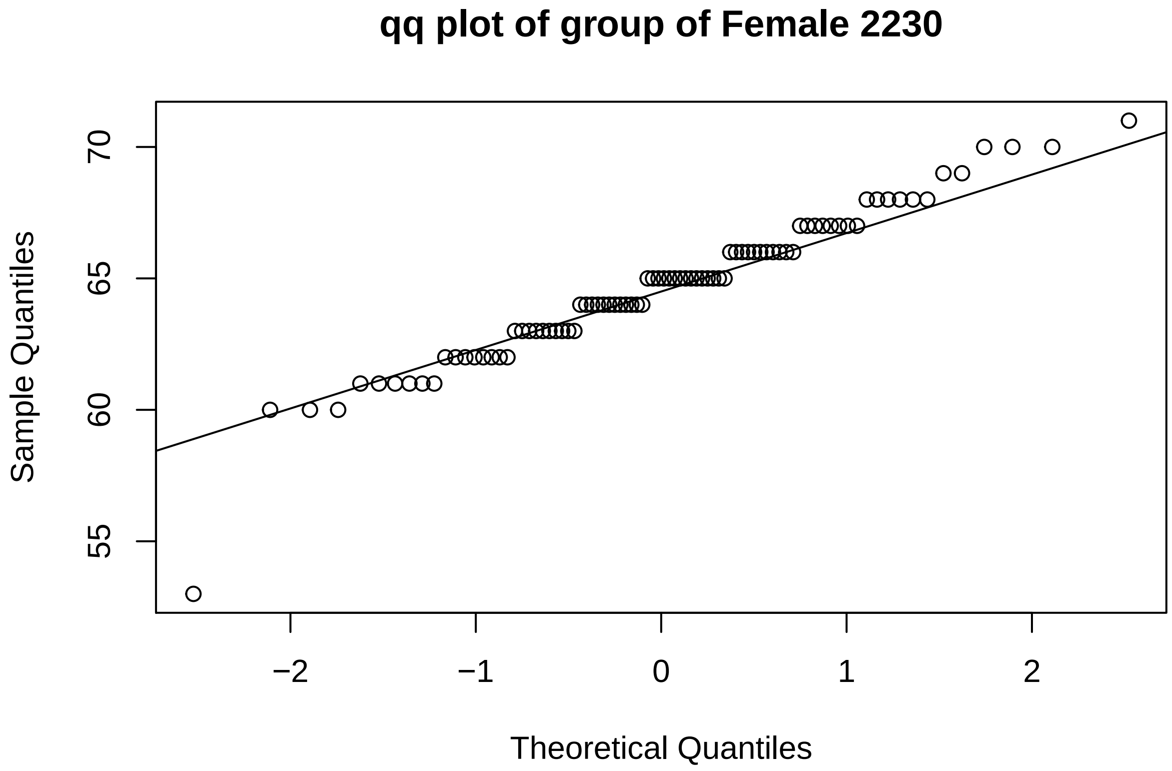
In order to carry two-sample t-test, we need to satisfy the following assumptions:

•The heights are independent, a person's height does not affect others. •Data are approximately normal for each group. •Variance of each group are equal.

• **Check data is approximately normal for each group.**



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   60.00   68.00   70.00   70.22   72.00   83.00 
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   53.00   63.00   65.00   64.62   66.00   71.00
```

From the summary of groups of male and female, we can see the median and mean of each group is very close. Also, according to the above normal qq plots for both groups, most observations lie along the 45-degree line in the QQ-plots, so we may assume that normality holds here.

•Check if variance are equal.

```
##
##  F test to compare two variances
##
## data:  male and Female
## F = 1.2917, num df = 79, denom df = 85, p-value = 0.2471
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8365331 2.0015544
## sample estimates:
## ratio of variances
##           1.291697
```

Want to check whether the variance of froup of female and male is equal, we use variance F-test, $H_0 : \sigma_1^2/\sigma_2^2 = 1, H_a : \sigma_1^2/\sigma_2^2 \neq 1$,we get p-value=0.2471, which is greater than 0.05, fail to reject H_0, the variance of the group of female and male are equal.

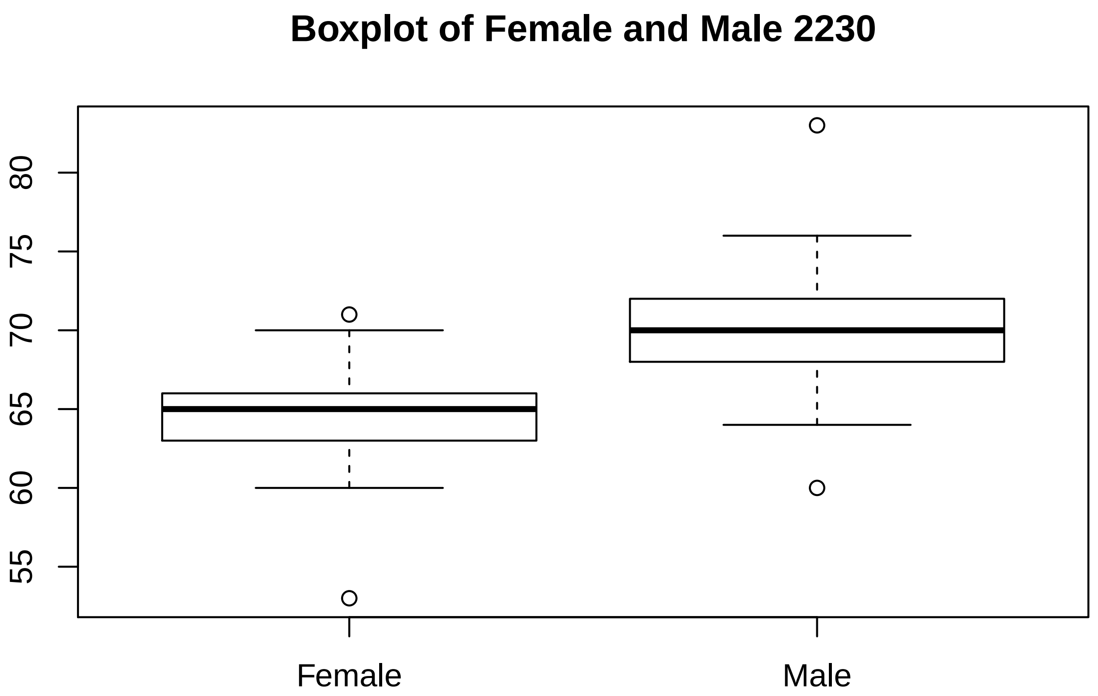
•Two sample t-test.

```
##
##  Two Sample t-test
##
## data:  male and Female
## t = 12.076, df = 164, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.691630 6.525811
## sample estimates:
## mean of x mean of y
##  70.22500  64.61628
```

Since the above assumptions are satisfied, then we can carry two sample pooled t-test.H_0 : $\mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$, p-value < 2.2e-16, which is smaller than 0.05, Then we reject H_0, the mean height of female and male is different.

t=$\bar{Y_1}$ - $\bar{Y_2}$/ $S_p\sqrt{1/n_1 + 1/n_2} \sim t_{n1+n2-1} = 12.076$

•Side by side boxplots



The right boxplot is higher than the left one, which suggests difference between the groups of Female and male. The median of male is siginificantly larger than Female. Half of female's height are between 63 and 66 inches, the maximum is 70 inches, minimum is 60 inches, and the boxplot of female is slightly left skewed. Half of male's height are between 68 and 72 inches, maximum is 75 inches, minimum is 64 inches.

Conclusion: There are two random data samples in the data, we want to do two sample t-test since it satifies the following assumption: height is independent, female and male groups have the same variance, data are approximately normal for each group. Then, by two-sample t test, we get result that the mean value of two groups is different. Moreover, according to side-by-side boxplots, the mean value of male is larger than the mean value of Female.

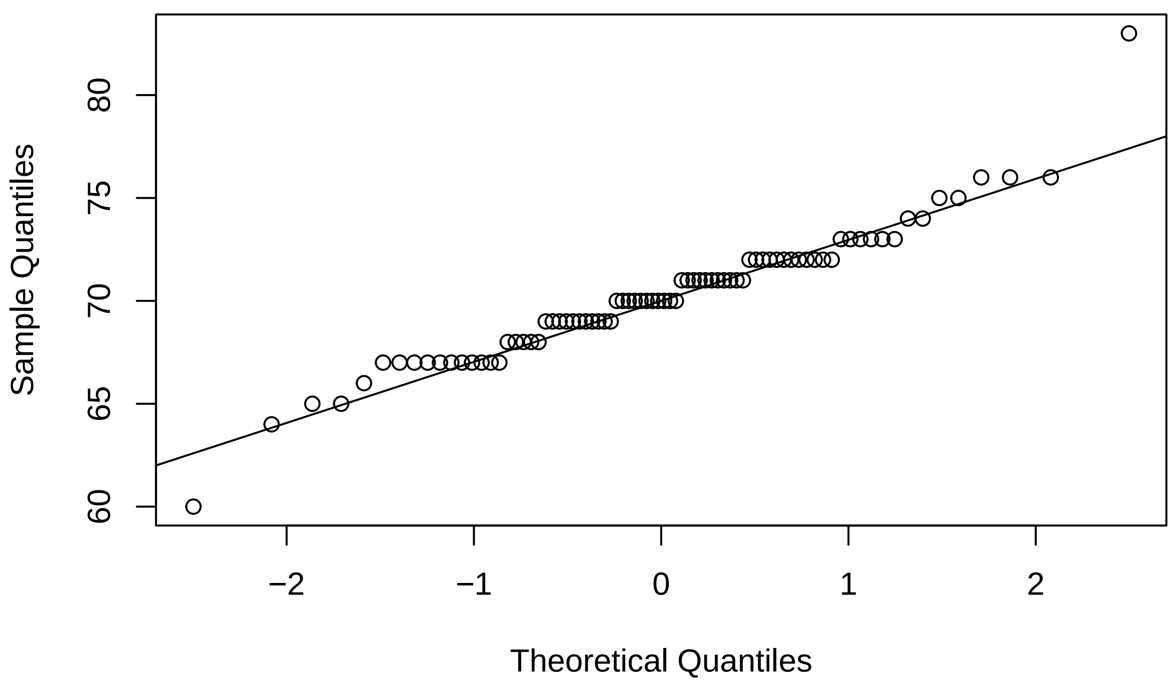
(d)

One way Anova with G=2, Simple linear Regression model with 1 dummy predictor variable.

(e)

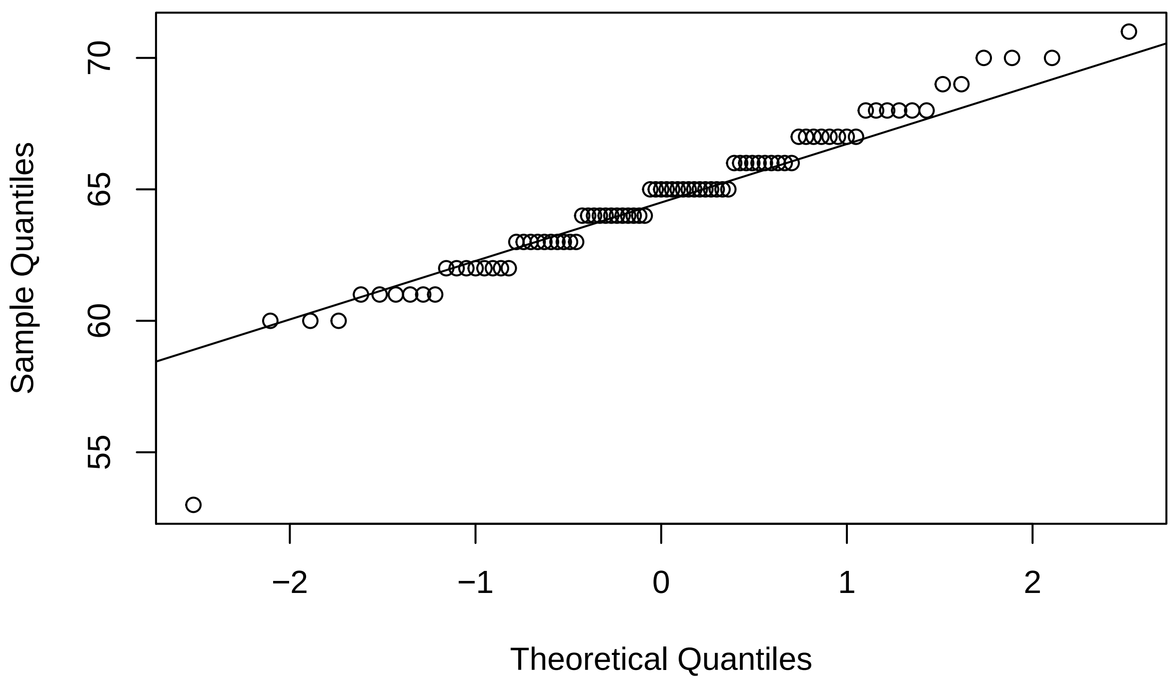
•Check data is approximately normal for each group.

qq plot of subset group of male 2230



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   60.00   68.00   70.00   70.22   72.00   83.00 
```

qq plot of subset group of Female 2230



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    53.0    63.0    65.0    64.6    66.0    71.0
```

From the summary of groups of female and male, we can see the median and mean of each group is very close. Also, according to the above normal qq plots for both groups, most observations lie along the 45-degree line in the QQ-plots. Therefore, we can conclude that they form approximately normal populations.

•Check if variance are equal.

```
##
##  F test to compare two variances
##
## data:  male2 and Female2
## F = 1.2802, num df = 79, denom df = 84, p-value = 0.2655
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8278249 1.9857384
## sample estimates:
## ratio of variances
##           1.280222
```

Want to check whether the variance of female and male is equal, we use variance F-test, $H_0 : \sigma_1^2/\sigma_2^2 = 1, H_a : \sigma_1^2/\sigma_2^2 \neq 1$, we get p-value=0.2655, which is greater than 0.05, fail to reject H_0, the variance of the group of female and male are equal.

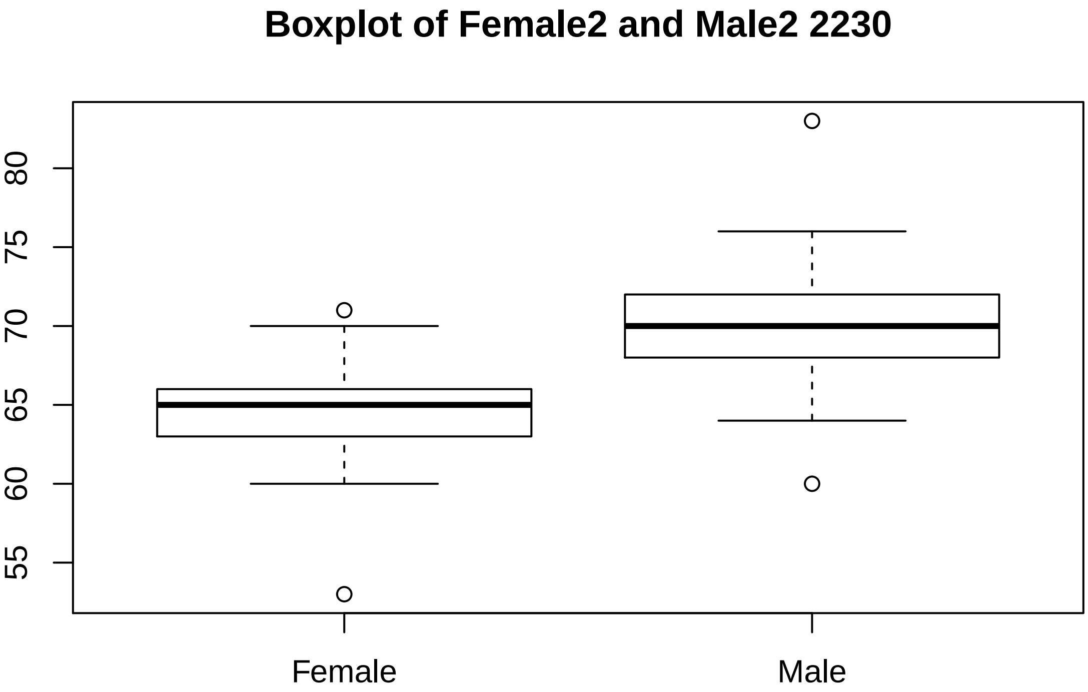
•Two sample t-test

```
##
##  Two Sample t-test
##
## data:  male2 and Female2
## t = 12.048, df = 163, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.703064 6.546936
## sample estimates:
## mean of x mean of y
##    70.225    64.600
```

Since the variance of female and male are equal, then we carry two sample pooled t-test.$H_0 : \mu_1 = \mu_2$, p-value $<$ 2.2e-16, which is smaller than 0.05, Then we reject H_0,the mean height of female and male is different.

Its distribution is t=$\bar{Y_1}$ - $\bar{Y_2}$/ $S_p\sqrt{1/n_1 + 1/n_2} \sim t_{n1+n2-1} = 12.048$.

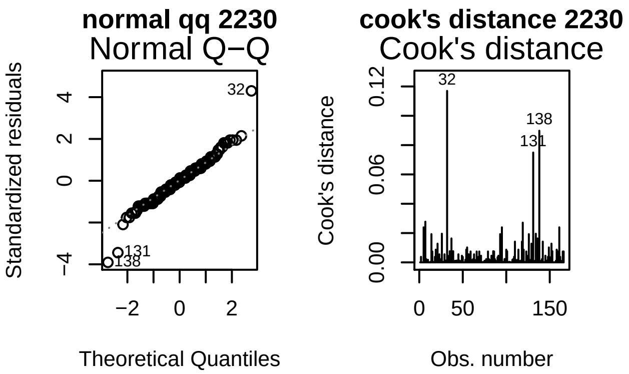
•Side by side boxplot.



The right boxplot is higher than the left one, which suggests the median of male is larger than Female. Half of female's height are between 63 and 66 inches, the maximum is 70 inches, minimum is 60 inches, and the boxplot of female is slightly left skewed. Half of male's height are between 68 and 72 inches, maximum is 75 inches, minimum is 64 inches.

Conclusion: There are two random data samples in the data, so I still choose to do two sample t-test since it satifies the following assumption: height is independent, female and male groups have the same variance, data are approximately normal for each group. Then, we get result that the mean value of two groups is different, which is the same result as before. Moreover, according to side-by-side boxplots, the mean value of male is larger than the mean value of Female.

(f)



According to normal qq-plot and cook's distance plot, we can see there are 3 influential points(138,131,32), they are not the point removed.

After removing 1 observation, there is no change in the group of male, since the data point removed is from female group. The variance of two groups are still equivalent by doing variance test. The mean value of two groups is different, which is the same result as before. The mean value of group of women is 64.6, 0.02 inches smaller than before, since the data point removed is 66, which is larger than the mean value. Moreover, the boxplots also looks the same. Therefore, the obeservations removed was not influential.

Appendix

Question1

(a)

Summary of groups of Female and Male.

```
juries = read.csv("/Users/mindu/Desktop/STA303/Assignment1/juries.csv",header=TRUE)
groupS = juries$PERCENT[juries$JUDGE== "SPOCKS"]
groupNS = juries$PERCENT[juries$JUDGE!= "SPOCKS"]
summary(groupS)
#IQR = 4.4
summary(groupNS)
#IQR = 9.3
```

(a)

Boxplot“JUDGE2230”

```
attach(juries)
boxplot(groupS,groupNS,xlab="JUDGE2230",names = c("SPOCKS","OTHERES"))
```

(b)

“Box plot without identifying outliers 2230”

```
groupS = juries$PERCENT[juries$JUDGE== "SPOCKS"]
groupNS = juries$PERCENT[juries$JUDGE!= "SPOCKS"]
boxplot(groupS,groupNS,outline = FALSE,names = c("SPOCKS","OTHERES"),main="Box plot with
```

Question2

(c)

Summaries of groups of Female and Male.

QQ-norm plot and QQ-line of Male,“qq plot of group of male 2230”.

QQ-norm plot and QQ-line of Female,“qq plot of group of Female 2230”.

```
#check the data are approximately normal for each group
heightt = read.csv("/Users/mindu/Desktop/STA303/Assignment1/assign1data.csv",header=TRUE
Female=heightt$height[heightt$sex == "Female"]
male=heightt$height[heightt$sex == "Male"]
qqnorm(heightt$height[heightt$sex == "Male"],main="qq plot of group of male 2230")
qqline(heightt$height[heightt$sex == "Male"])
summary(male)
qqnorm(heightt$height[heightt$sex == "Female"],main="qq plot of group of Female 2230")
qqline(heightt$height[heightt$sex == "Female"])
summary(Female)
```

Variance test of Female and Male.

```
#check if variance are equal
var.test(male,Female)
```

Two sample t-test

```
#two sample t-test
t.test(male,Female,var.equal = T)
```

Side by side boxplots of Female and Male,“Boxplot of Female and Male 2230”

```
#side by side boxplots
boxplot(heightt$height~heightt$sex, main="Boxplot of Female and Male 2230")
```

(e)

Summaries of groups of subset of Female and Male.

QQ-norm plot and QQ-line of subset of Male,“qq plot of subset group of male 2230”.

QQ-norm plot and QQ-line of subset of Female,“qq plot of subset group of Female 2230”.

```
heightt_subset = subset(heightt,id!=30)
Female2=heightt_subset$height[heightt_subset$sex == "Female"]
male2=heightt_subset$height[heightt_subset$sex == "Male"]
qqnorm(heightt_subset$height[heightt_subset$sex == "Male"],main="qq plot of subset group
qqline(heightt_subset$height[heightt_subset$sex == "Male"])
summary(male2)
qqnorm(heightt_subset$height[heightt_subset$sex == "Female"],main="qq plot of subset gro
qqline(heightt_subset$height[heightt_subset$sex == "Female"])
summary(Female2)
```

Variance test of subset of Female and Male.

```r
#check if variance are equal
var.test(male2,Female2)
```

Two sample t-tset of subset of Female and Male

```r
#two sample t-tset
t.test(male2,Female2,var.equal = T)
```

Side by side boxplots of Female and Male,“Boxplot of Female and Male 2230”

```r
#boxplot
boxplot(heightt$height~heightt$sex, main="Boxplot of Female2 and Male2 2230")
```

(f)

“normal qq 2230” “cook’s distance 2230”

```r
#figure out influential points
influential_plot = lm(heightt$height~heightt$sex)
par(mfrow=c(2,4))
plot(influential_plot,which=2,main="normal qq 2230")
plot(influential_plot,which=4,main="cook's distance 2230")
```