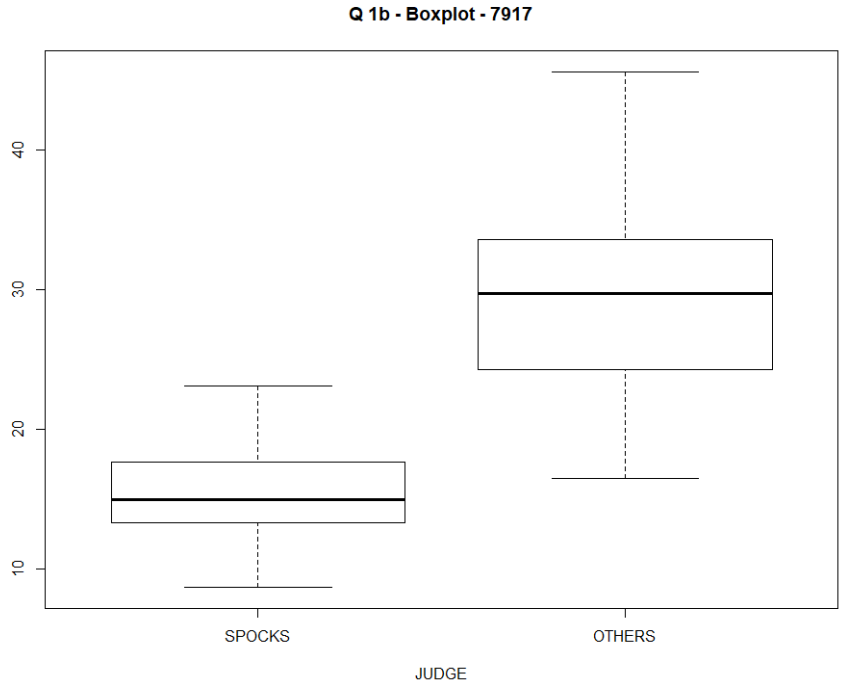
Question 1:

1a)
There is one data point below the Q1-1.5IQR threshold for Spock, a value of 6.4. There are no data points above the Q3+1.5IQR threshold for Spock.

There is one data point above the Q3+1.5IQR threshold for Others, a value of 48.9. There are no data points below the Q1-1.5IQR threshold for Others.

	Q1	Q2	Q3	Q4	IQR	Q1-1.5IQR	Q3+1.5IQR
Spock	13.3	15	17.7	23.1	4.4	6.7	24.3
Others	24.3	29.7	33.6	48.9	9.3	10.35	47.6

1b)



1c)
The boxplot which identifies outliers gives much more information about the dataset than the one without. This is especially true given the maximum and minimum whisker in R is given by the formula max_whisker = min(max(data), Q3 + 1.5*IQR), min_whisker = max(min(data), Q1 – 1.5*IQR). In the boxplot with outliers, one is able to actually see the outliers, plus understand which values the whiskers represent (be it real data points or Q3+1.5IQR or Q1-1.5IQR). If you do not have outliers on the chart you are unable to see this information.

Question 2:

2a)
The data was obtained from an observational study. A key limitation of observational studies is the relationship between variables may be due to just coincidence instead of one factor causing the other.

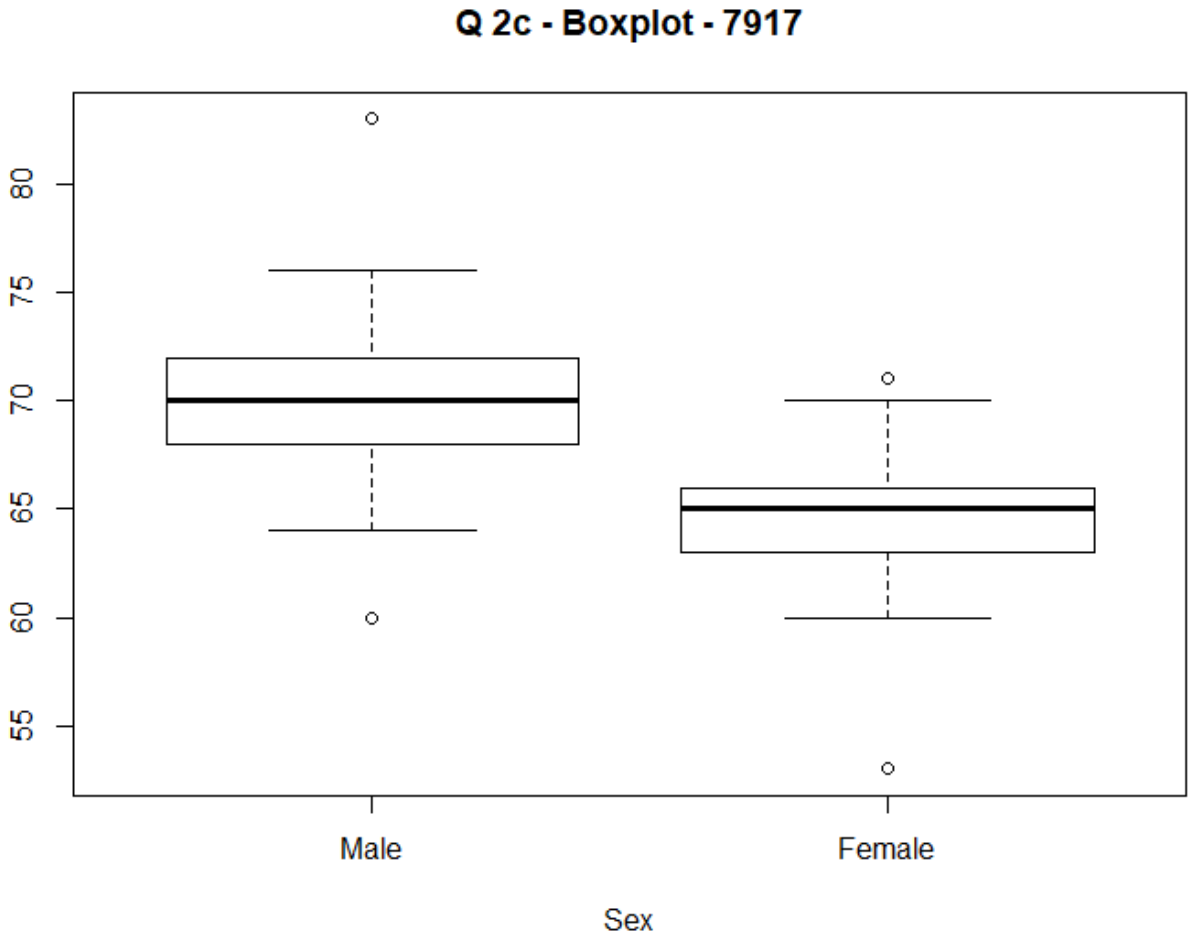
In this case we run a risk that the association we find between height and sex may be just coincidence. An experiment would allow one to better test for causation.

2b)

Sex is a categorical. Levels are Male/Female.

2c)

i) Side-by-side boxplots



ii) Null and Alternative Hypothesis
Null Hypothesis: $\mu_{male} - \mu_{female} = 0$
Alternative Hypothesis: $\mu_{male} - \mu_{female} \neq 0$

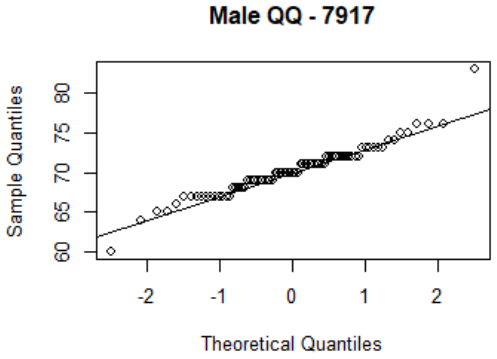
iii) A test statistic and its distribution
The pooled t-test returned a test statistic of 12.076. This test statistic comes from a t-distribution with 164 degrees of freedom.

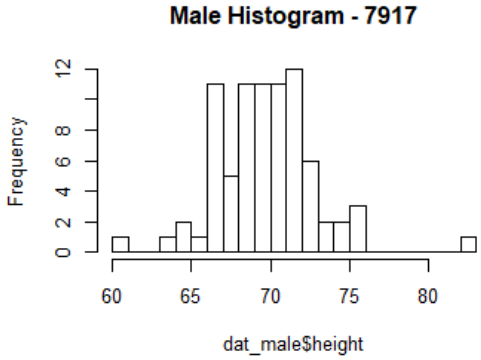
iv) Test assumptions
Assumptions for a two-sample t-test such as this include: the two samples are independent of each other, the two samples have equal variance, and the two samples are iid from approximately normal.

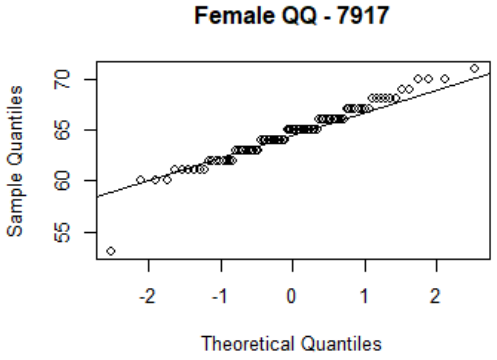
v) Test diagnostics
Clearly the two samples are independent (since they contain different individuals in each sample).

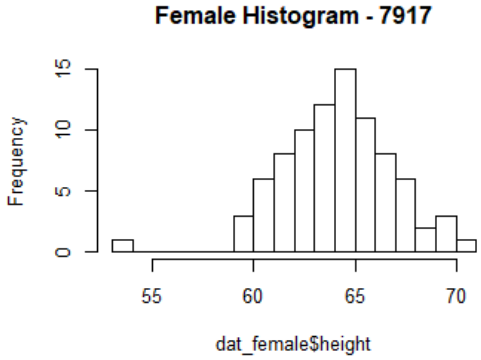
To test equal variance we can use an F-test. Given a high p-value of .2471 from the F-test there is evidence that the variances are equal.

To test normality we observe the Shapiro-Wilk normality test for male and female heights return p-values of 0.002513 and 0.00488 respectively suggesting the data is not normally distributed. However, the QQ plots and histograms show the data appears to be normally distributed. Potentially the large positive outlier for males and large negative outlier for females are skewing the result of Shapiro-Wilk test. In general, t-procedures are robust against assumptions of normality (valid even if assumption of normality is violated) so the conclusions of the two sample t-test should still be valid.









vi) P-Value
The p-value associated with the test statistic of 12.076 is $2.2e^{-16}$.

vii) Results(brief discussion and conclusion)
Given the very low p-value of $2.2e^{-16}$ there is enough evidence to reject the null hypothesis that the height of males and females is equal. There is clear evidence that there is an association between sex and height. It appears that females are in general shorter than men.

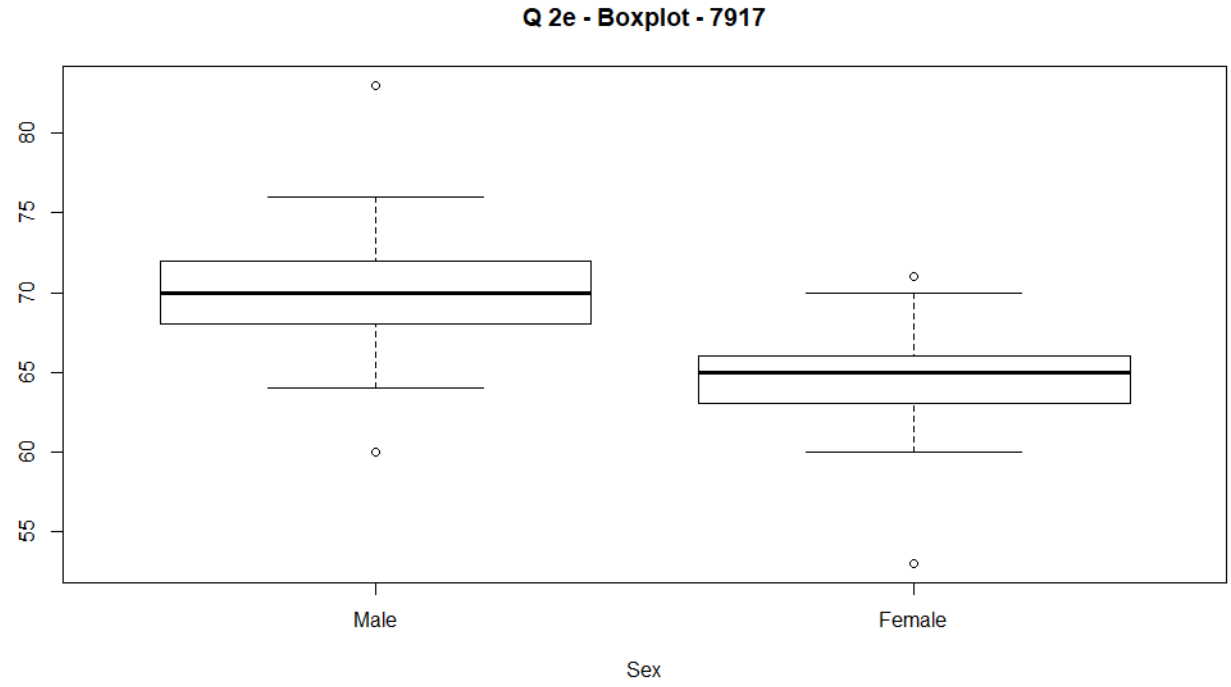
2d)

i) One statistical method equivalent to the method used in part C is One-way ANOVA where one tests if $\beta_{male} = 0$ (where $\beta_{male} = 1$ if an observation is male and 0 if an observation is female).

ii) Another statistical method equivalent is a Bonferroni Pairwise T-test.

2e)

i) Side-by-side Boxplots



ii) Null and Alternative Hypothesis
Null Hypothesis: $\mu_{male} - \mu_{female} = 0$
Alternative Hypothesis: $\mu_{male} - \mu_{female} \neq 0$

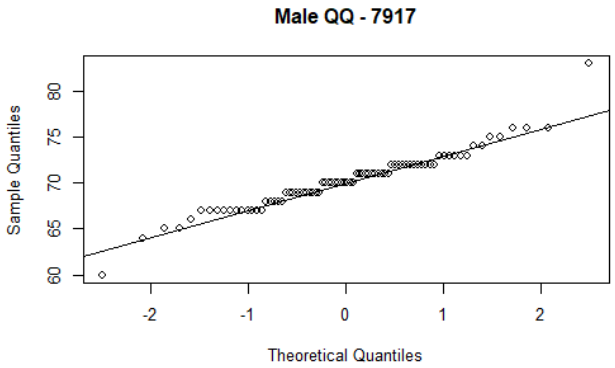
iii) A test statistic and its distribution
The pooled t-test returned a test statistic of 11.987. This test statistic comes from a t-distribution with 162 degrees of freedom (two less than previously which is unsurprising given I deleted two data points).

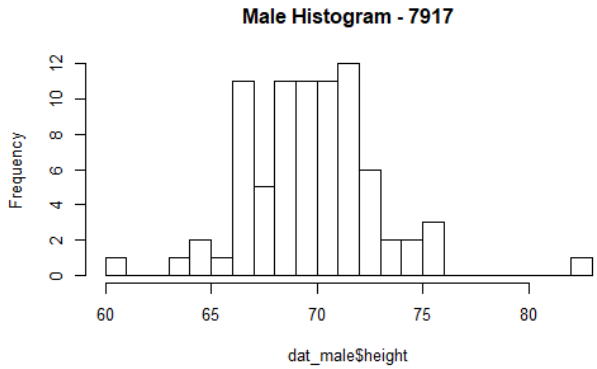
iv) Test assumptions
Assumptions for a two-sample t-test such as this include: the two samples are independent of each other, the two samples have equal variance , and the two samples are iid from approximately normal.

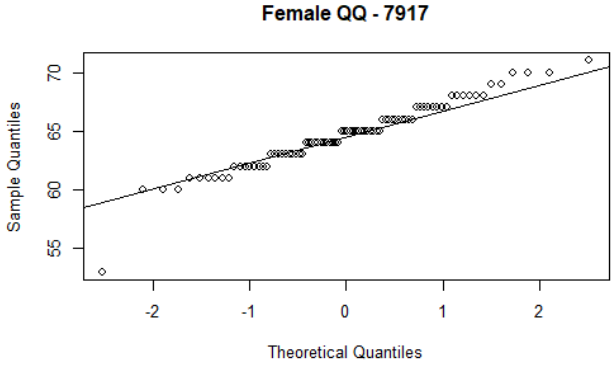
v) Test diagnostics
Clearly the two samples are independent (since they contain different individuals in each sample).

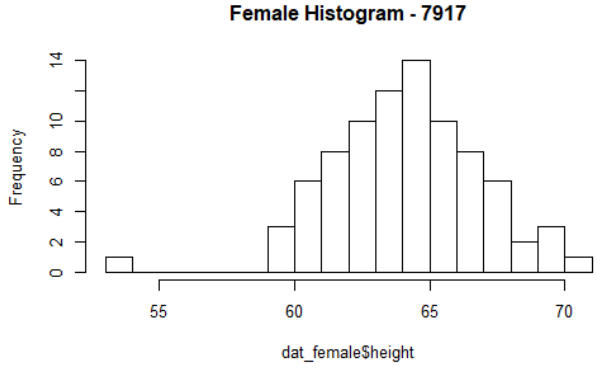
To test equal variance we can use an F-test. Given a high p-value of .2906 from the F-test there is evidence that the variances are equal.

To test normality we observe the Shapiro-Wilk normality test for male and female heights return p-values of 0.002513 and 0.00635 respectively suggesting the data is not normally distributed. Potentially the large positive outlier for males and large negative outlier for females are skewing the result of Shapiro-Wilk test. In general, t-procedures are robust against assumptions of normality (valid even if assumption of normality is violated) so the conclusions of the two sample t-test should still be valid.









vi) P-Value
The p-value associated with the test statistic of 11.987 is $2.2e^{-16}$.

vii) Results(brief discussion and conclusion)
Given the very low p-value of $2.2e^{-16}$ there is enough evidence to reject the null hypothesis that the height of males and females is equal. There is clear evidence that there is an association between sex and height. It appears that females are in general shorter than men.

2f)

It appears the two observations removed (observation 17 and 117) were not influential. This is evident in the boxplots looking largely the same, very similar test statistics (11.987 vs. 12.076 originally), similar values for the F-Test and Shapiro-Wilk tests, and identical P-Values of $2.2e^{-16}$.

Appendix A - R Output

Question 1

PART A

R Output

```
> quantiles_spock
 25%  50%  75% 100%
13.3 15.0 17.7 23.1
> IQR_spock
     75%
4.400001
> outlier_spock_high_threshold
 75%
24.3
> outlier_others_low_threshold
  25%
10.35
> outliers_spock_high
[1] PERCENT JUDGE
<0 rows> (or 0-length row.names)
> outliers_spock_low
  PERCENT  JUDGE
1     6.4 SPOCKS


> quantiles_others
 25%  50%  75% 100%
24.3 29.7 33.6 48.9
> IQR_others
     75%
9.299999
> outlier_others_high_threshold
  75%
47.55
> outlier_others_low_threshold
  25%
10.35
> outliers_others_high
   PERCENT JUDGE
14    48.9     A
> outliers_others_low
[1] PERCENT JUDGE
<0 rows> (or 0-length row.names)
```

Question 2

Part C

iii) R Output

```
> t.test(dat_male$height,dat_female$height,var.equal=T)

Two Sample t-test

data:  dat_male$height and dat_female$height
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
```

v) R output

```
> var.test(dat_male$height,dat_female$height)

F test to compare two variances

data:  dat_male$height and dat_female$height
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

> shapiro.test(dat_male$height)

Shapiro-Wilk normality test

data:  dat_male$height
W = 0.94758, p-value = 0.002513

> shapiro.test(dat_female$height)

Shapiro-Wilk normality test

data:  dat_female$height
W = 0.95556, p-value = 0.00488
```

PART D – R Output

```
> summary(aov(dat$height~dat$sex))
             Df Sum Sq Mean Sq F value Pr(>F)
dat$sex       1   1304  1303.8   145.8 <2e-16 ***
Residuals   164   1466     8.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> pairwise.t.test(dat$height,factor(dat$sex,labels=c("Male","Female")),
p.adj="bonf")
Pairwise comparisons using t tests with pooled SD

data:  dat$height and factor(dat$sex, labels = c("Male", "Female"))

       Male
Female <2e-16

P value adjustment method: bonferroni
```

Part E

iii) R Output

```
> t.test(dat_male$height,dat_female$height,var.equal=T)

Two Sample t-test

data:  dat_male$height and dat_female$height
t = 11.987, df = 162, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.702328 6.557195
sample estimates:
mean of x mean of y
 70.22500  64.59524
```

v) R Output

```
> var.test(dat_male$height,dat_female$height)

F test to compare two variances

data:  dat_male$height and dat_female$height
F = 1.2653, num df = 79, denom df = 83, p-value = 0.2906
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8168815 1.9645602
sample estimates:
ratio of variances
           1.26529


> shapiro.test(dat_male$height)

Shapiro-Wilk normality test

data:  dat_male$height
W = 0.94758, p-value = 0.002513

> shapiro.test(dat_female$height)

Shapiro-Wilk normality test

data:  dat_female$height
W = 0.95649, p-value = 0.00635
```

Appendix B – R Code

Question 1

```
setwd("C:/Users/David/Google Drive/Documents/UofT/NonDegree/303/")
juries_dat <- read.csv("juries(1).csv")

# separate data into two groups
spock_group <- juries_dat[juries_dat$JUDGE == "SPOCKS",]
others_group <- juries_dat[juries_dat$JUDGE != "SPOCKS",]

# Check for Outliers in Spocks group and other groups
# IQR rule is Q3 + 1.5IQR or Q1 - 1.5IQR

# IQR spock group
quantiles_spock <- quantile(spock_group$PERCENT,seq(.25,1,.25))
IQR_spock <- (quantiles_spock[3]-quantiles_spock[1])
outlier_spock_high_threshold = quantiles_spock[3] + 1.5*IQR_spock
outlier_spock_low_threshold = quantiles_spock[1] - 1.5*IQR_spock

outliers_spock_high <- subset(spock_group, PERCENT > outlier_spock_high_threshold)
outliers_spock_low <- subset(spock_group, PERCENT < outlier_spock_low_threshold)

# IQR of OTHERS
quantiles_others <- quantile(others_group$PERCENT,seq(.25,1,.25))
IQR_others <- quantiles_others[3]-quantiles_others[1]
outlier_others_high_threshold = quantiles_others[3] + 1.5*IQR_others
outlier_others_low_threshold = quantiles_others[1] - 1.5*IQR_others

outliers_others_high <- subset(others_group, PERCENT > outlier_others_high_threshold)
outliers_others_low <- subset(others_group, PERCENT < outlier_others_low_threshold)

# Create boxplot
boxplot(spock_group$PERCENT, others_group$PERCENT,
        xlab="JUDGE",names=c("SPOCKS","OTHERS"),
        main="Q 1b - Boxplot - 7917",outline=FALSE)
```

Question 2

```
setwd("C:/Users/David/Google Drive/Documents/UofT/NonDegree/303/")
dat <- read.csv("assign1datamodified.csv")
dat_male <- dat[dat$sex == "Male",]
dat_female <- dat[dat$sex == "Female",]

# Draw Boxplot
boxplot(dat_male$height, dat_female$height,
        xlab="Sex",names=c("Male","Female"),
        main="Q 2e - Boxplot - 7917")

# Conduct a Pooled T-Test
t.test(dat_male$height,dat_female$height,var.equal=T)

# Test diagnostics (checking model assumptions)

# F test for equal variances
var.test(dat_male$height,dat_female$height)

# Test Normality Histogram & QQ plot
par(mfrow=c(2,2))
qqnorm(dat_male$height,main="Male QQ - 7917")
qqline(dat_male$height)
hist(dat_male$height, main="Male Histogram - 7917", breaks=20)

qqnorm(dat_female$height,main="Female QQ - 7917")
qqline(dat_female$height)
hist(dat_female$height, main="Female Histogram - 7917", breaks=20)

# Test Normality - Shaprio
shapiro.test(dat_male$height)

shapiro.test(dat_female$height)
```