PS2 Group16

Shailendra Patil, Mingyang Ma, Apurva Gupta

January 20, 2018

Question 1

Installing NHANES package and ggplot Libraries

```
#install.packages("NHANES")
#install.packages("ggplot2")
```

Activating NHANES package and ggplot libraries

```
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 3.4.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

Creating a data frame which contains only the people sample aged 18 and up

```
Adults = subset(NHANES, Age >= 18)
head(Adults)
```

```
## # A tibble: 6 x 76
##      ID SurveyYr Gender   Age AgeDecade AgeMonths  Race1  Race3
##   <int>   <fctr> <fctr> <int>    <fctr>     <int> <fctr> <fctr>
## 1 51624  2009_10   male    34     30-39       409  White   <NA>
## 2 51624  2009_10   male    34     30-39       409  White   <NA>
## 3 51624  2009_10   male    34     30-39       409  White   <NA>
## 4 51630  2009_10 female    49     40-49       596  White   <NA>
## 5 51647  2009_10 female    45     40-49       541  White   <NA>
## 6 51647  2009_10 female    45     40-49       541  White   <NA>
## # ... with 68 more variables: Education <fctr>, MaritalStatus <fctr>,
## #   HHIncome <fctr>, HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <fctr>, Work <fctr>, Weight <dbl>, Length <dbl>,
## #   HeadCirc <dbl>, Height <dbl>, BMI <dbl>, BMICatUnder20yrs <fctr>,
## #   BMI_WHO <fctr>, Pulse <int>, BPSysAve <int>, BPDiaAve <int>,
## #   BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>, BPSys3 <int>,
## #   BPDia3 <int>, Testosterone <dbl>, DirectChol <dbl>, TotChol <dbl>,
## #   UrineVol1 <int>, UrineFlow1 <dbl>, UrineVol2 <int>, UrineFlow2 <dbl>,
## #   Diabetes <fctr>, DiabetesAge <int>, HealthGen <fctr>,
## #   DaysPhysHlthBad <int>, DaysMentHlthBad <int>, LittleInterest <fctr>,
## #   Depressed <fctr>, nPregnancies <int>, nBabies <int>, Age1stBaby <int>,
## #   SleepHrsNight <int>, SleepTrouble <fctr>, PhysActive <fctr>,
## #   PhysActiveDays <int>, TVHrsDay <fctr>, CompHrsDay <fctr>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fctr>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fctr>, Smoke100 <fctr>,
## #   Smoke100n <fctr>, SmokeAge <int>, Marijuana <fctr>,
## #   AgeFirstMarij <int>, RegularMarij <fctr>, AgeRegMarij <int>,
## #   HardDrugs <fctr>, SexEver <fctr>, SexAge <int>, SexNumPartnLife <int>,
## #   SexNumPartYear <int>, SameSex <fctr>, SexOrientation <fctr>,
## #   PregnantNow <fctr>
```

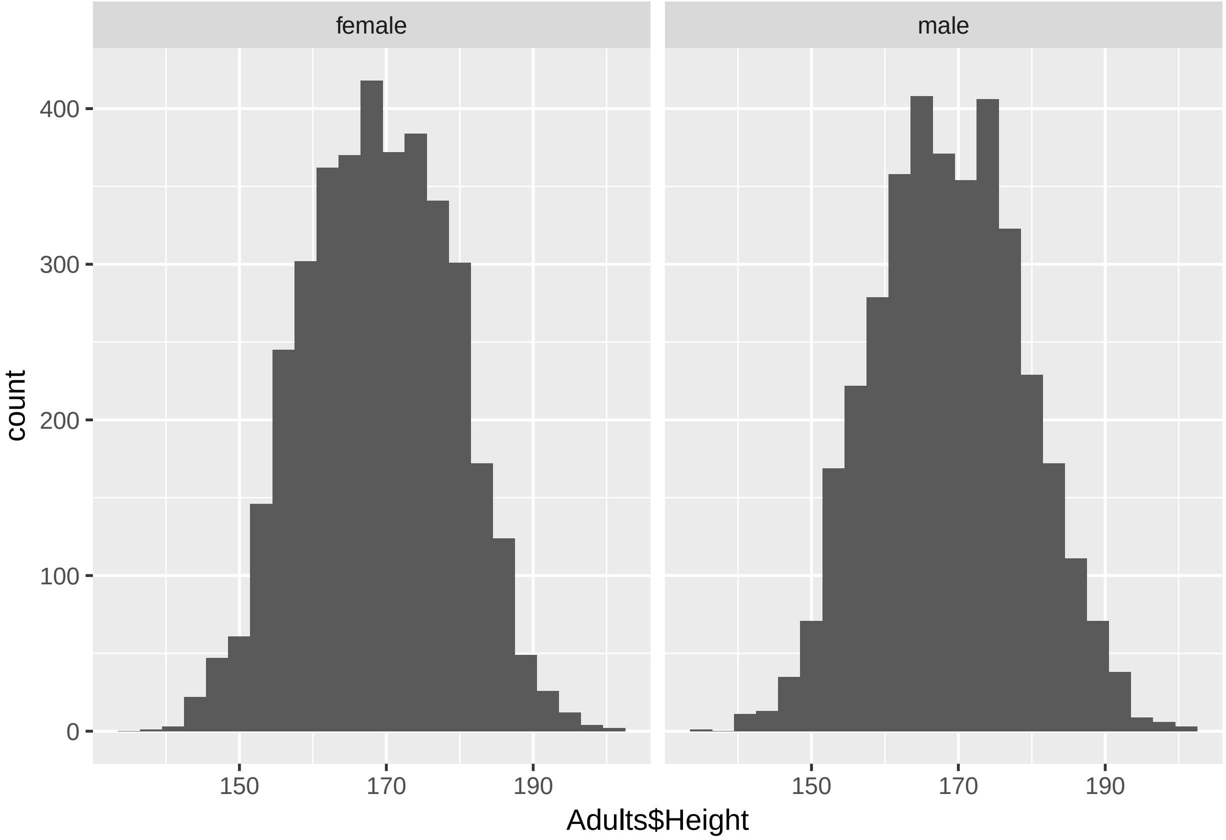
Creating an object of ggplot Library

```
ggobj = ggplot(Adults, aes(x=Adults$Height))
```

Plotting Histogram for Male and Female Heights

```
ggobj + geom_histogram(binwidth=3)+facet_grid(~Adults$Gender)
```

```
## Warning: Removed 57 rows containing non-finite values (stat_bin).
```



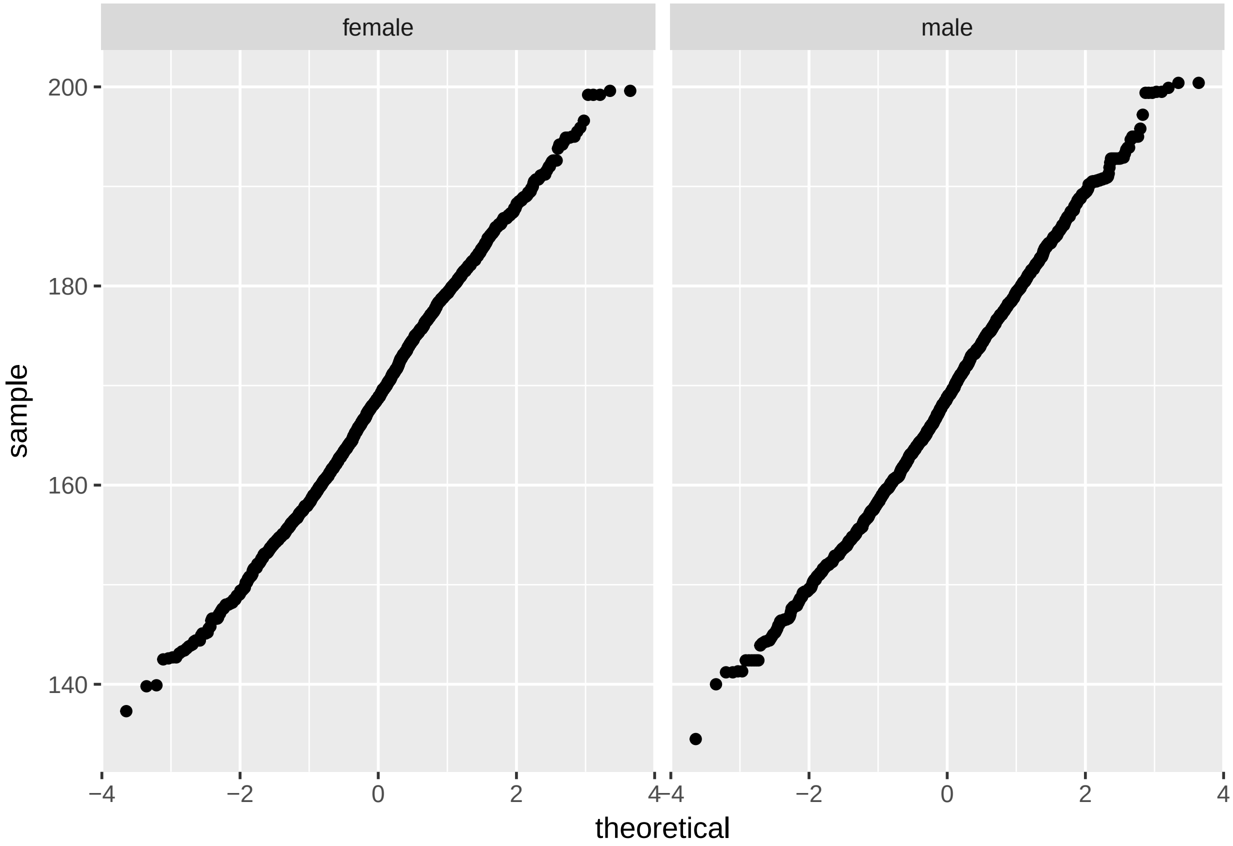
Observation

The histogram shows that both male and female heights have approximately same center, spread and distribution.

Checking Normality by plotting qqplot graphs.

```
 ggplot(Adults, aes(sample=Adults$Height)) + stat_qq() +facet_grid(~Adults$Gender)
```

```
## Warning: Removed 57 rows containing non-finite values (stat_qq).
```

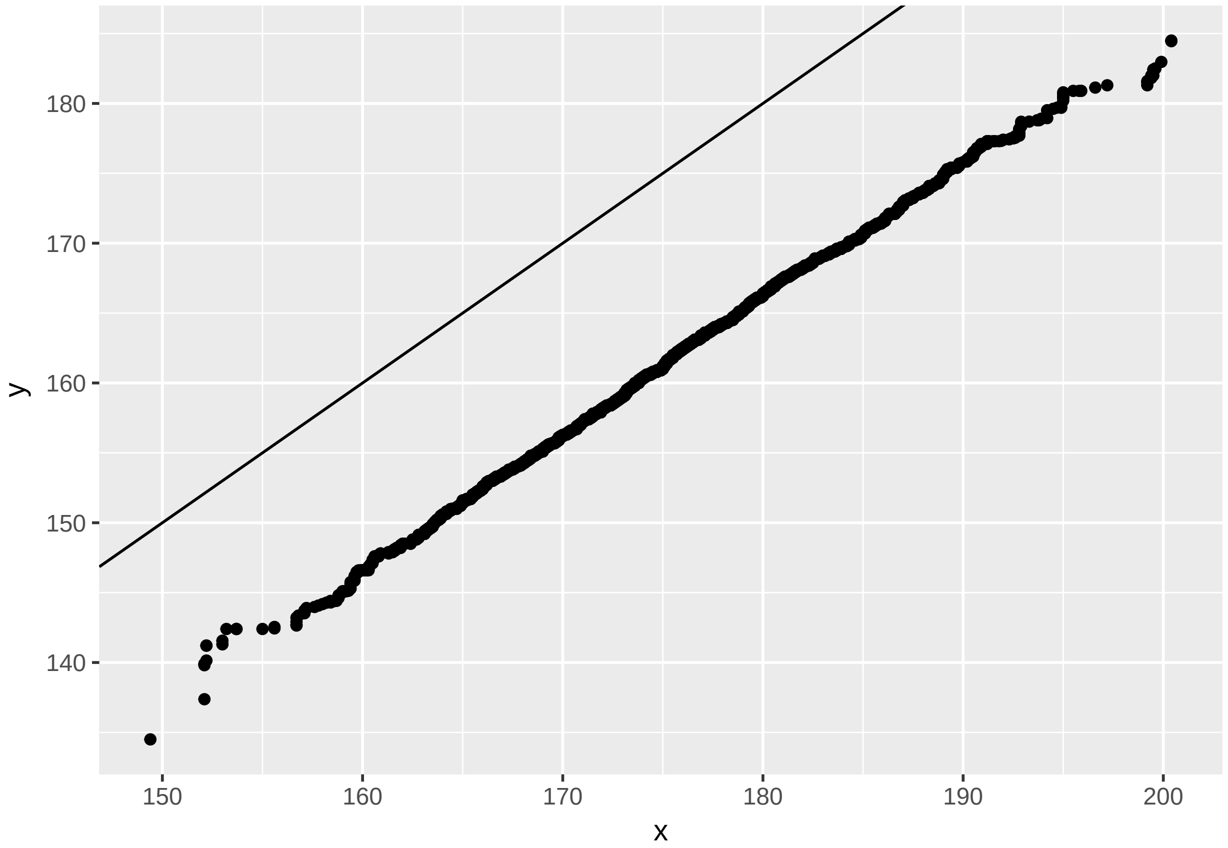


Observation

We can observe that qqplots of male heights and female heights are approximately straight lines. Hence, they are approximately normal.

Creating scatter plot to find if it is an additive/multiplicative shift.

```r
Maleheight=Adults$Height[Adults$Gender=="male"]
Femaleheight=Adults$Height[Adults$Gender=="female"]
qq.df=as.data.frame(qqplot(Maleheight,Femaleheight,plot.it = FALSE))
ggplot(qq.df,aes(x=x,y=y))+geom_point()+geom_abline()
```



Observation

We can observe that it is a straight line parallel to reference line. Therefore, Male and Female heights are in additive shift.

Quantitative Expression of Male and Female Heights

```
Adults[is.na(Adults$Height),]$Height <-mean(Adults$Height,na.rm =T)
mean_diff = mean(Adults$Height[Adults$Gender=="male"]) - mean(Adults$Height[Adults$Gender=="female"])
mean_diff
```

```
## [1] 13.70921
```

Observation

Quantitatively, We can say that difference in mean heights of male and female is about 14 units.

Question 2

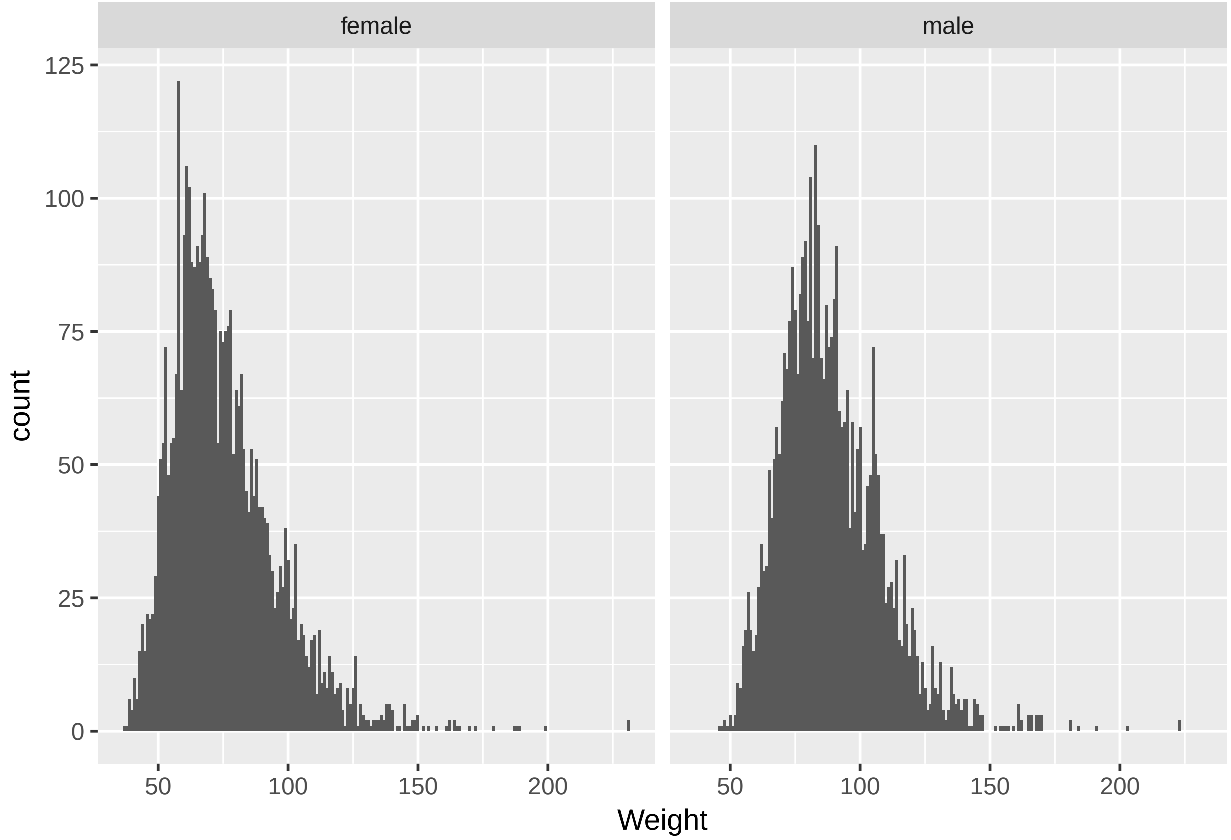
Creating an object of ggplot Library

```
ggobj1 = ggplot(Adults, aes(x=Weight))
```

Plotting Histogram for Male and Female Weights

```
ggobj1 + geom_histogram(binwidth=1)+facet_grid(~Adults$Gender)
```

```
## Warning: Removed 61 rows containing non-finite values (stat_bin).
```



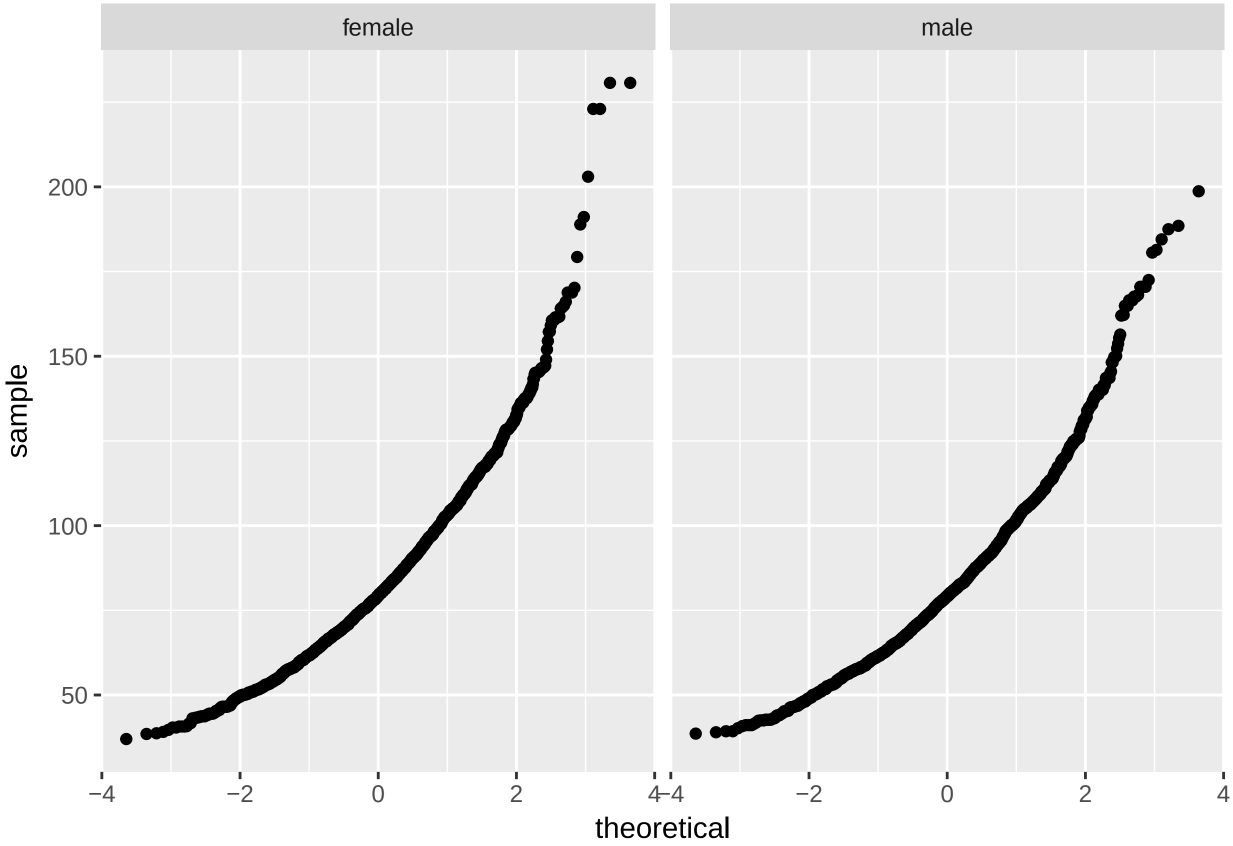
Observation

The histogram shows that both male and female weights have approximately same distribution with a slightly shifted center.

Checking Normality by plotting qqplot graphs.

```
ggplot(Adults, aes(sample=Adults$Weight)) + stat_qq() +facet_grid(~Adults$Gender)
```

```
## Warning: Removed 61 rows containing non-finite values (stat_qq).
```



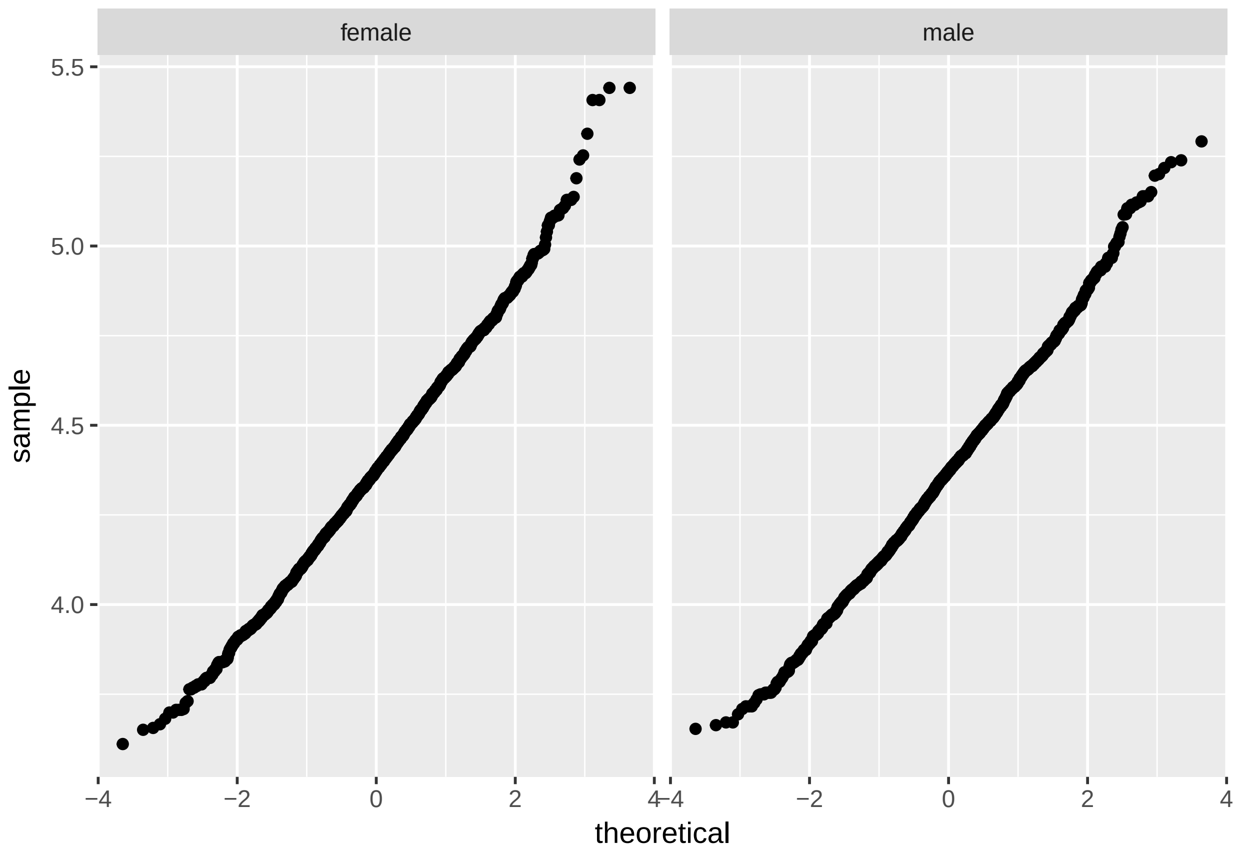
Observation

We can observe that qqplots of male Weights and female Weights are curved lines. Hence, we cannot conclude that they are approximately normal.

Checking Normality by plotting qqplot graphs of log weights.

```
ggplot(Adults, aes(sample=log(Adults$Weight))) + stat_qq() +facet_grid(~Adults$Gender)
```

```
## Warning: Removed 61 rows containing non-finite values (stat_qq).
```

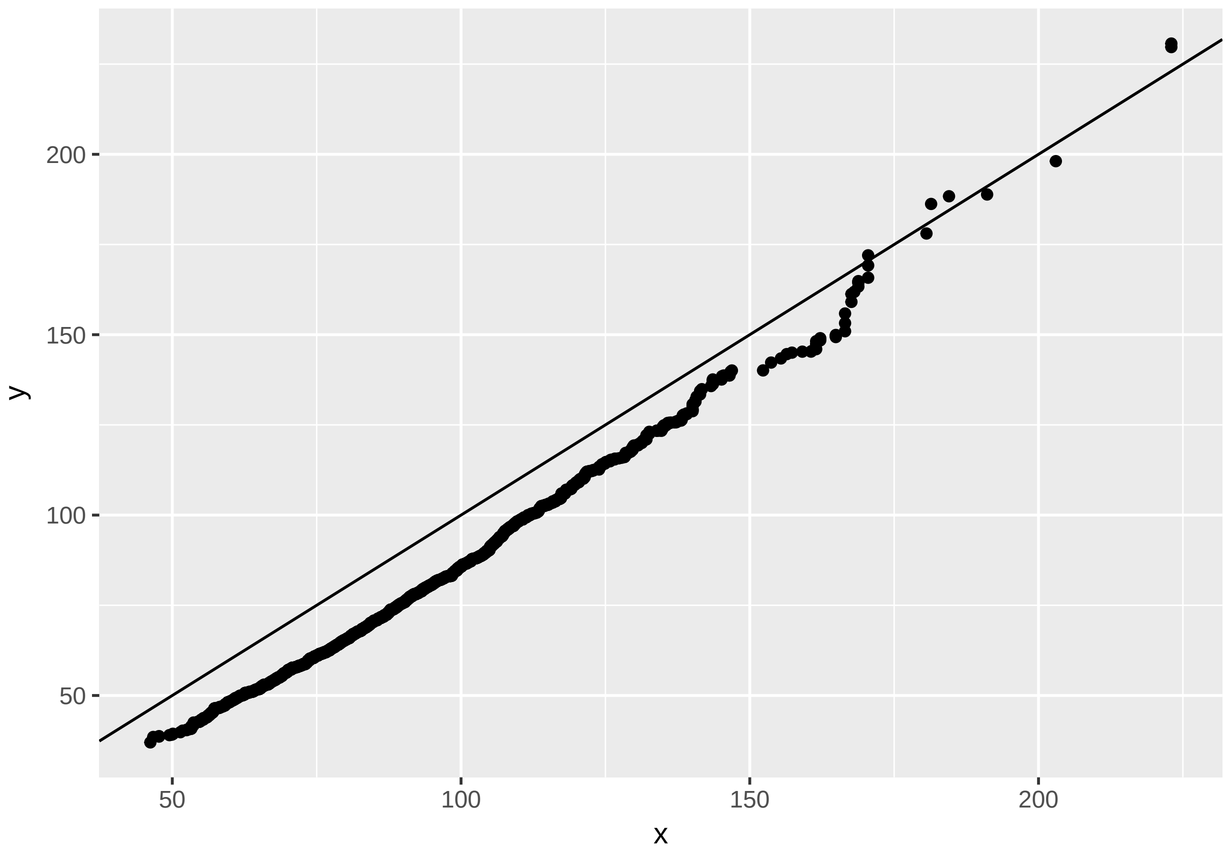


Observation

We can observe that qqplots of log male and female Weights are appromimately straight lines. Hence, we can conclude that log values are approximately normal.

Creating scatter plot to find if it is an additive/multiplicative shift.

```
Maleweight=Adults$Weight[Adults$Gender=="male"]
Femaleweight=Adults$Weight[Adults$Gender=="female"]
qq.df=as.data.frame(qqplot(Maleweight,Femaleweight,plot.it = FALSE))
ggplot(qq.df,aes(x=x,y=y))+geom_point()+geom_abline()
```



Observation

We can observe that it is approximately a straight line parallel to reference line. Therefore, we can say that Male and Female weights are in additive shift.

Quantitative Expression of Male and Female Weights

```
Adults[is.na(Adults$Weight),]$Weight <-mean(Adults$Weight,na.rm =T)
mean_diff = mean(Adults$Weight[Adults$Gender=="male"]) - mean(Adults$Weight[Adults$Gender=="female"])
mean_diff
```

```
## [1] 13.39489
```

Observation

Quantitatively, We can say that difference in mean weights of male and female is about 13 units.

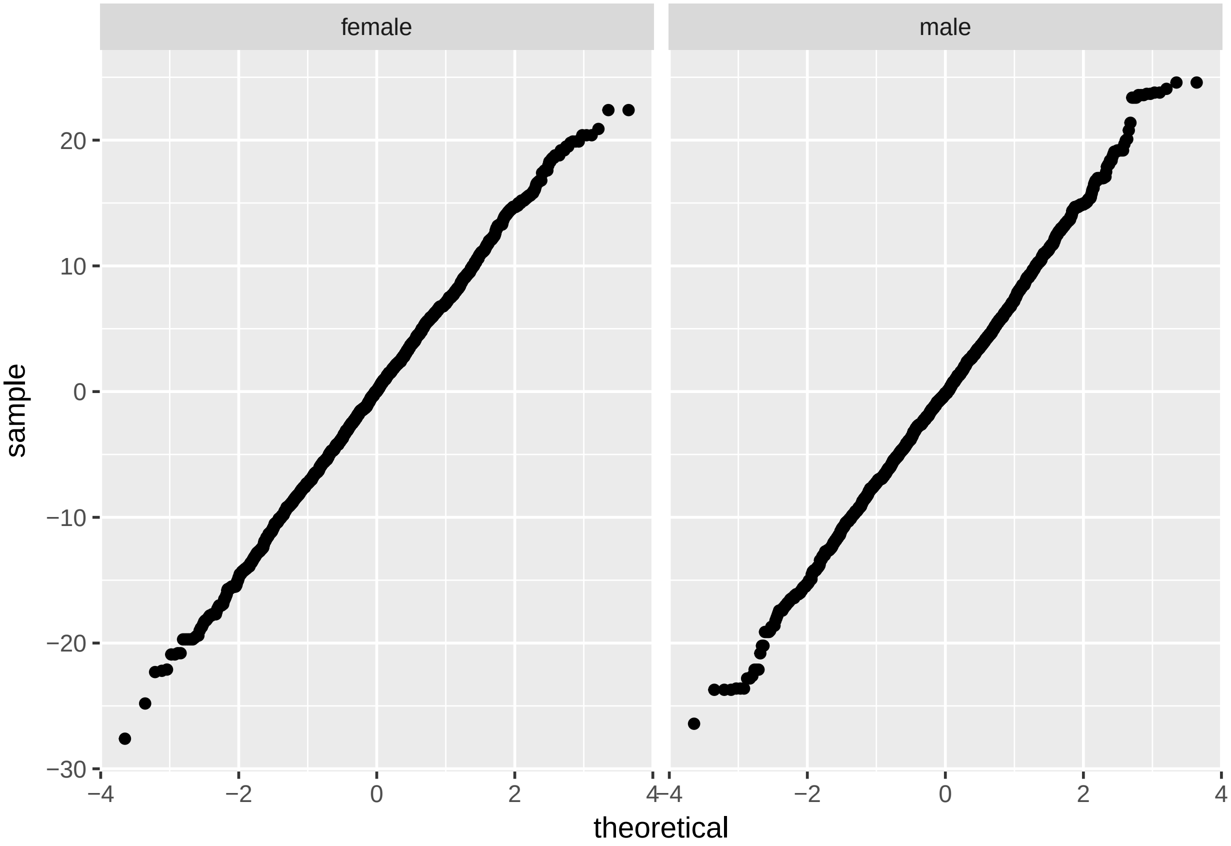
Question 3

Fitting a linear model to predict Heights from Gender

```
Adult.lm=lm(Height~Gender,data=Adults)
Adult.res=data.frame(Gender=Adults$Gender,residual=residuals(Adult.lm))
Adult.fitted = sort(fitted.values(Adult.lm)) - mean(fitted.values(Adult.lm))
Adult.residuals = sort(residuals(Adult.lm))
```

Checking normlaity of residuals

```
ggplot(Adult.res, aes(sample = residual)) +
  stat_qq() + facet_wrap(~Gender, ncol=2)
```



Observation

We can observe that residuals are approximately normal.

Creating a data frame with fiited values and residuals.

```
n = length(Adult.residuals)
f.value = (0.5:(n - 0.5)) / n
Adult.fit = data.frame(f.value, Fitted=Adult.fitted, Residuals=Adult.residuals)
```

We will take several variables and gathers them together, so you have more observations but fewer columns.

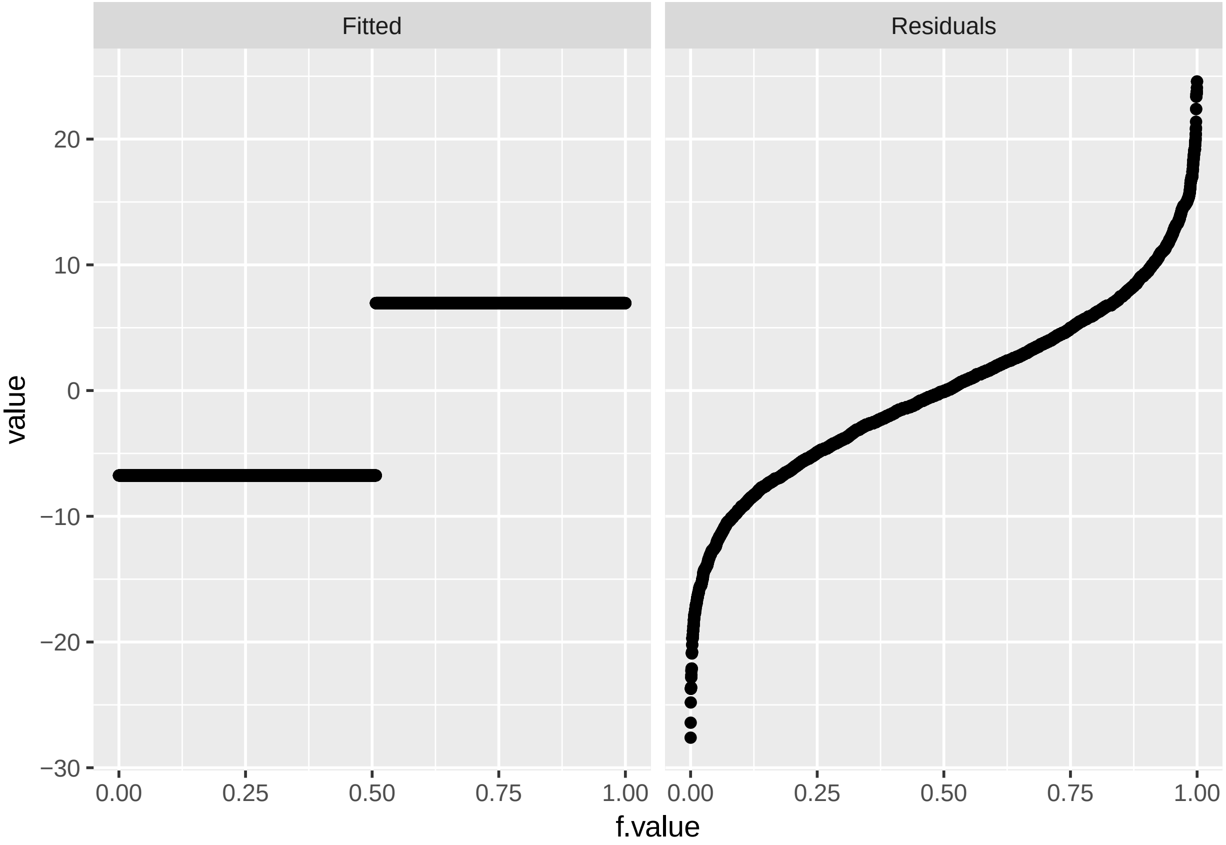
```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
adult.fit.long = Adult.fit %>% gather(type, value, Fitted:Residuals)
```

Residual fit plot

```
ggplot(adult.fit.long, aes(x=f.value, y=value)) +
  geom_point() + facet_wrap(~type)
```



Observation

The fitted values are close together compared to the residuals. While the model may be useful,it only accounts for a fraction of the variation in the data.We can check the results quantitatively.

Quantifying the model fit

```
var(Adult.fitted) / var(Adults$Height)
```

```
## [1] 0.4635624
```

Observation

the model captures 47% of the variance of the data." The remaining 53% is in the residuals.