Stat 630 Lab 3

Chris Kalra aa6389

9/6/2018

Exercise 1

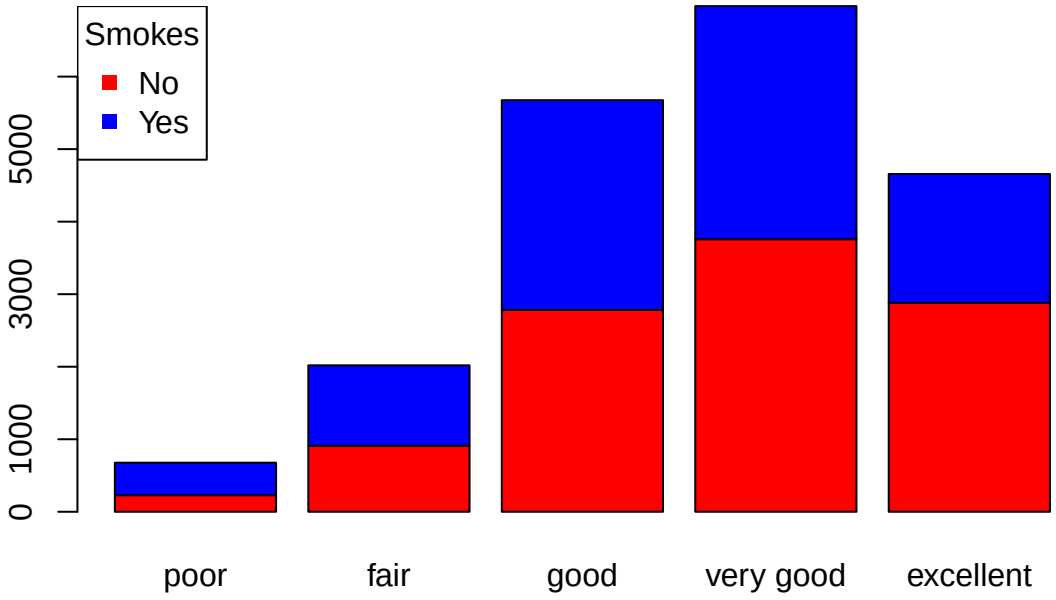
```
#Preamble
data_url <- "https://github.com/ericwfox/stat630data/raw/master/cdc.csv"
cdc <- read.csv(data_url, header = TRUE)

#1a)
class(cdc$genhlth)
```

```
## [1] "factor"
```

```
levels(cdc$genhlth)
```

```
## [1] "excellent" "fair"      "good"      "poor"      "very good"
```

```
cdc$genhlth = factor(cdc$genhlth, levels=c("poor","fair","good", "very good","excellent"))
Table=table(cdc$smoke100, cdc$genhlth)

#1b)
barplot(table(cdc$smoke100, cdc$genhlth), col=c("red", "blue"))
legend(x="topleft", c('No','Yes'), col=c('red','blue'), title="Smokes", pch=15)
```




```
#1c)
prop.table(Table, margin=1) #Marginal for those who have not smoked
```

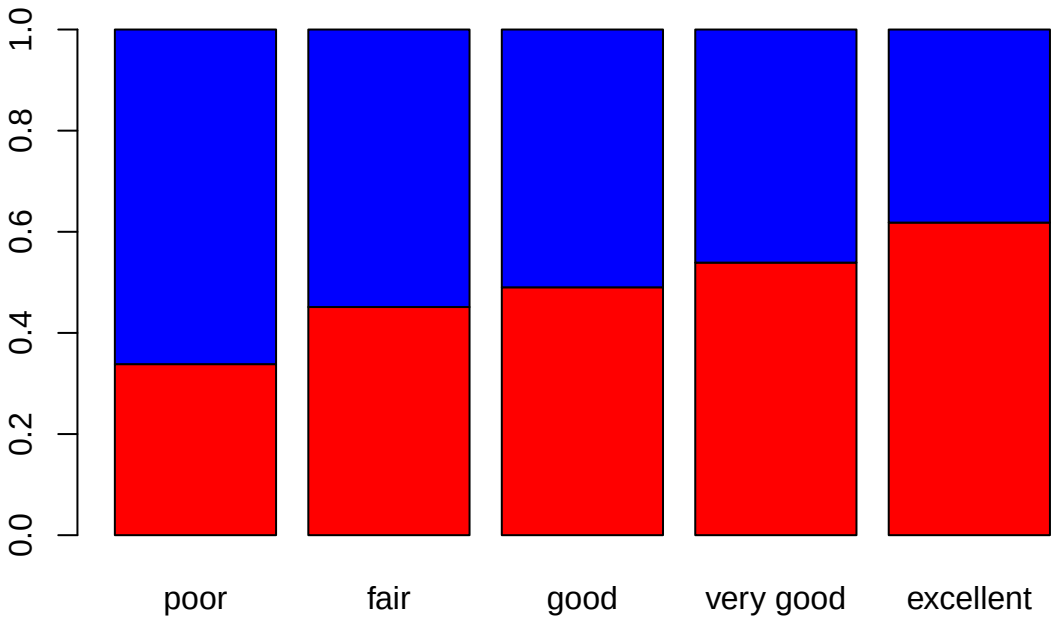
```
##
##           poor       fair       good  very good  excellent
##   0 0.02168766 0.08627711 0.26347192 0.35590492 0.27265840
```

```
##   1 0.04745260 0.11736045 0.30642940 0.34043004 0.18832751
```

```
prop.table(Table, margin=2) #Marginal based on health
```

```
##
##          poor      fair      good very good excellent
##   0 0.3382570 0.4512135 0.4902203 0.5390132 0.6182091
##   1 0.6617430 0.5487865 0.5097797 0.4609868 0.3817909
```

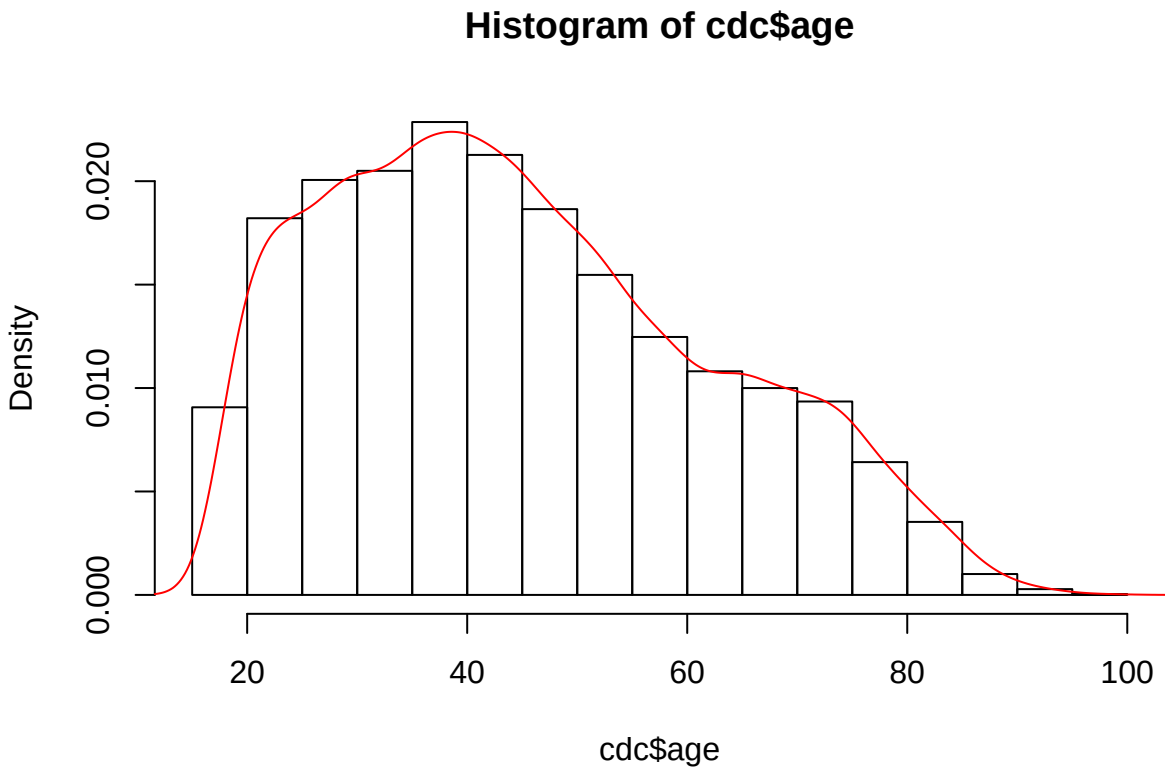
```
barplot(prop.table(Table, margin=2), col=c("red", "blue"))
```



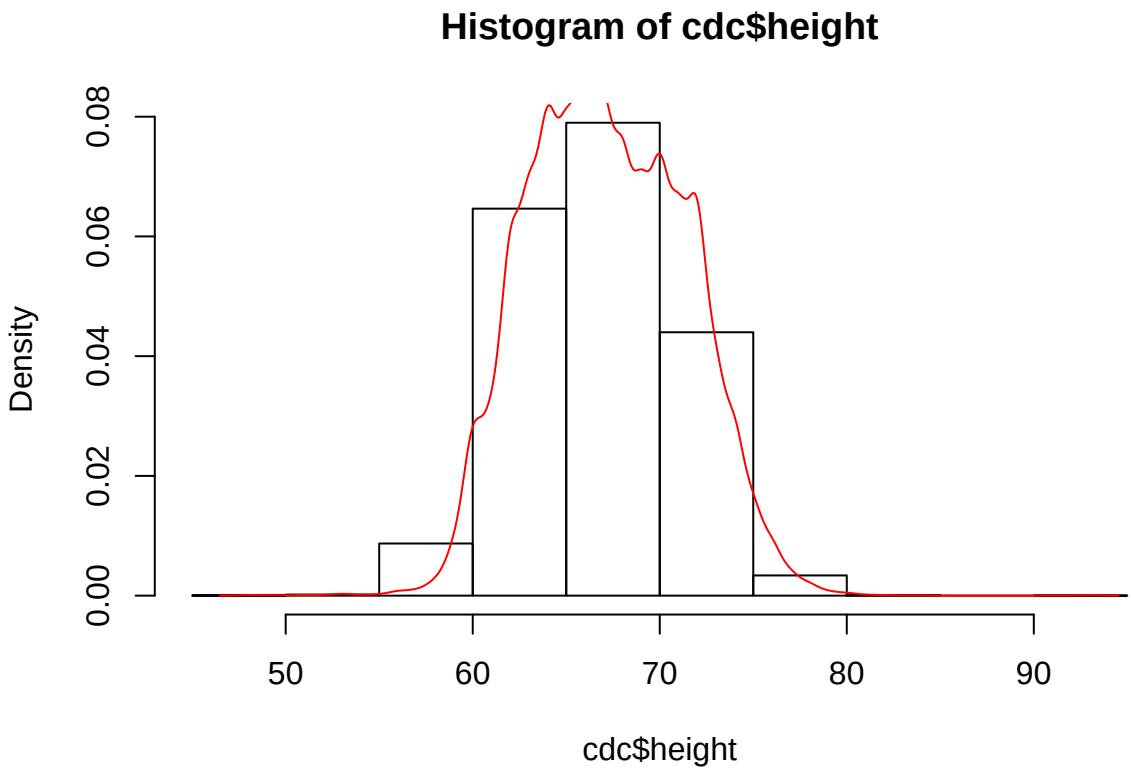
We notice that, on average, the healthier an individual claims to be, the less likely they are to smoke

Exercise 2

```
#2a)
hist(cdc$age, freq=F)
lines(density(cdc$age), col="red")
```

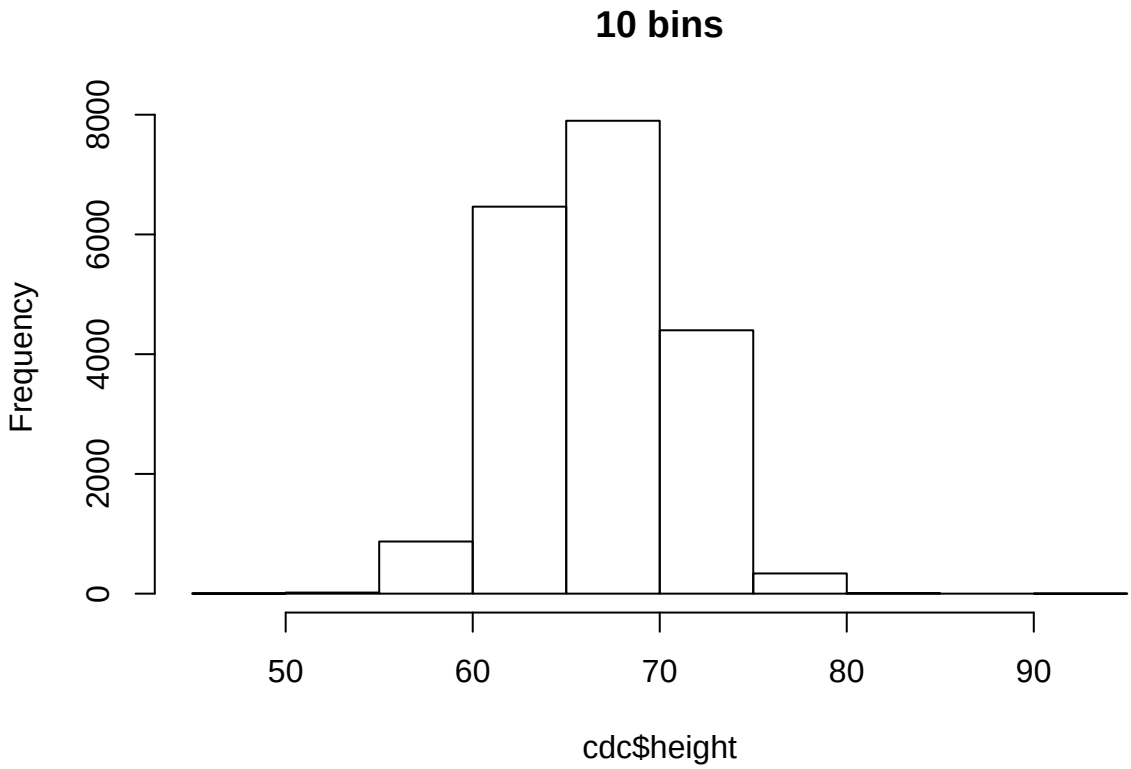



```
hist(cdc$height, freq=F)
lines(density(cdc$height), col="red")
```

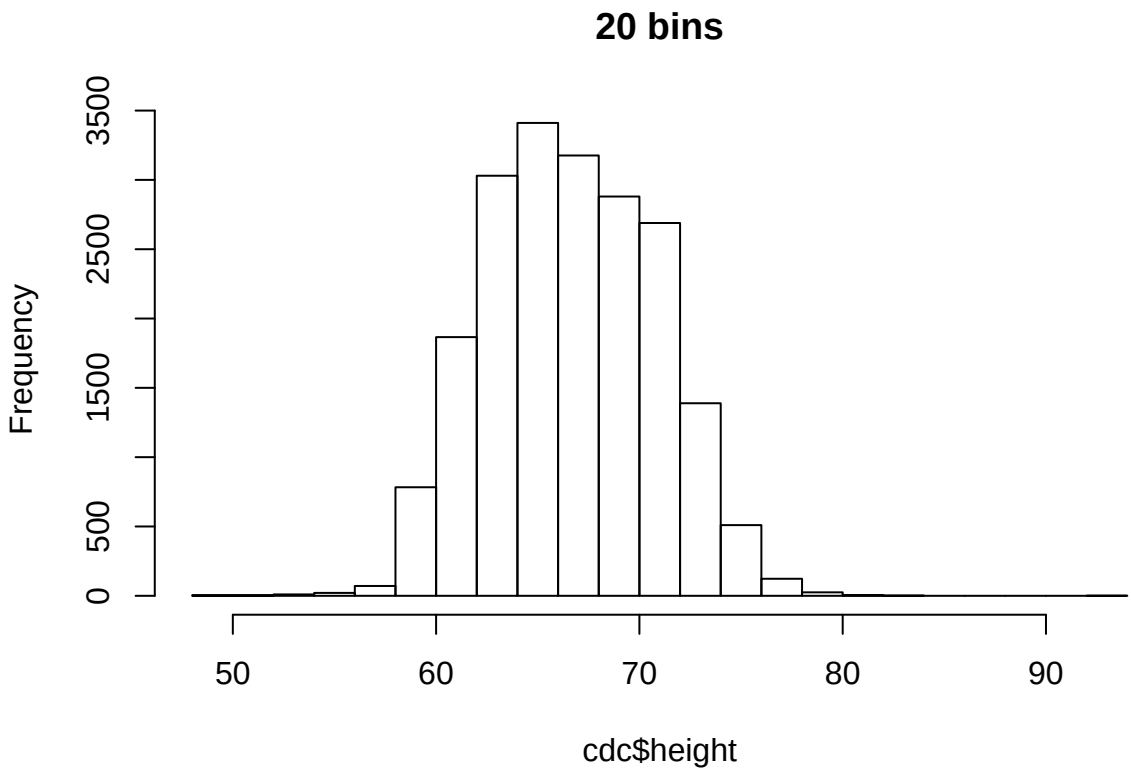


Height seems roughly normally distributed, whereas age appears to have a bit of a right skew

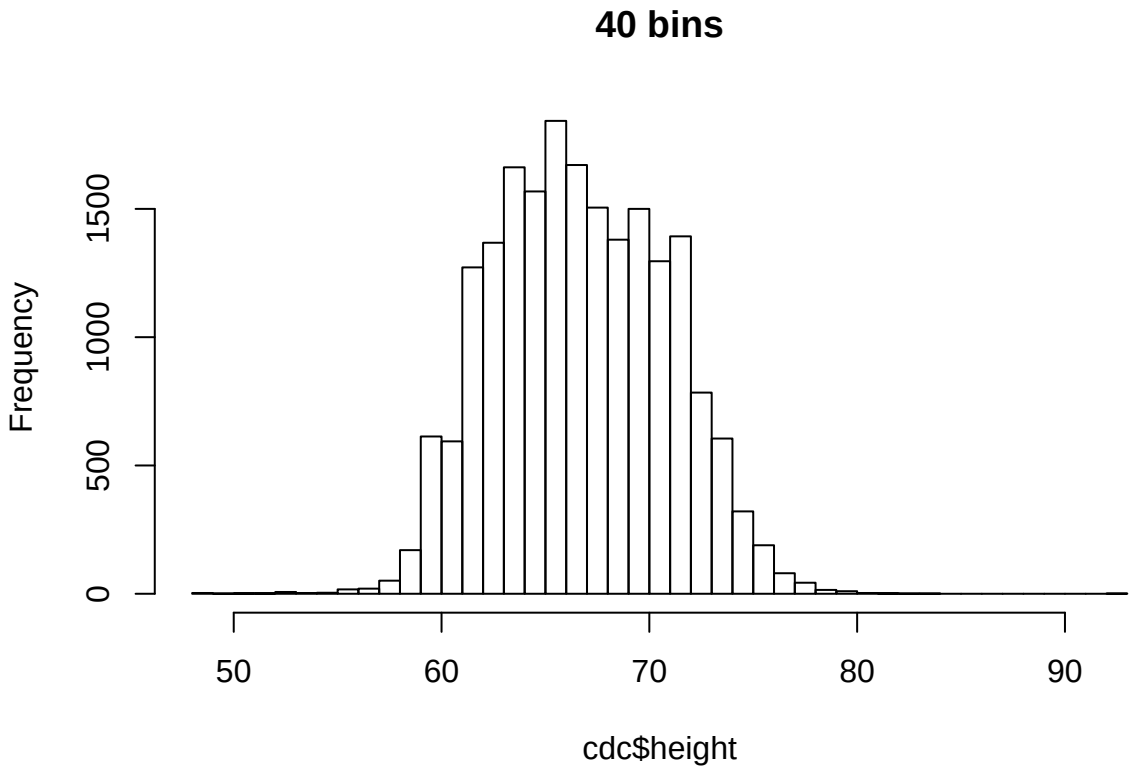
```r
#2b)
hist(cdc$height, breaks=10, main="10 bins")
```



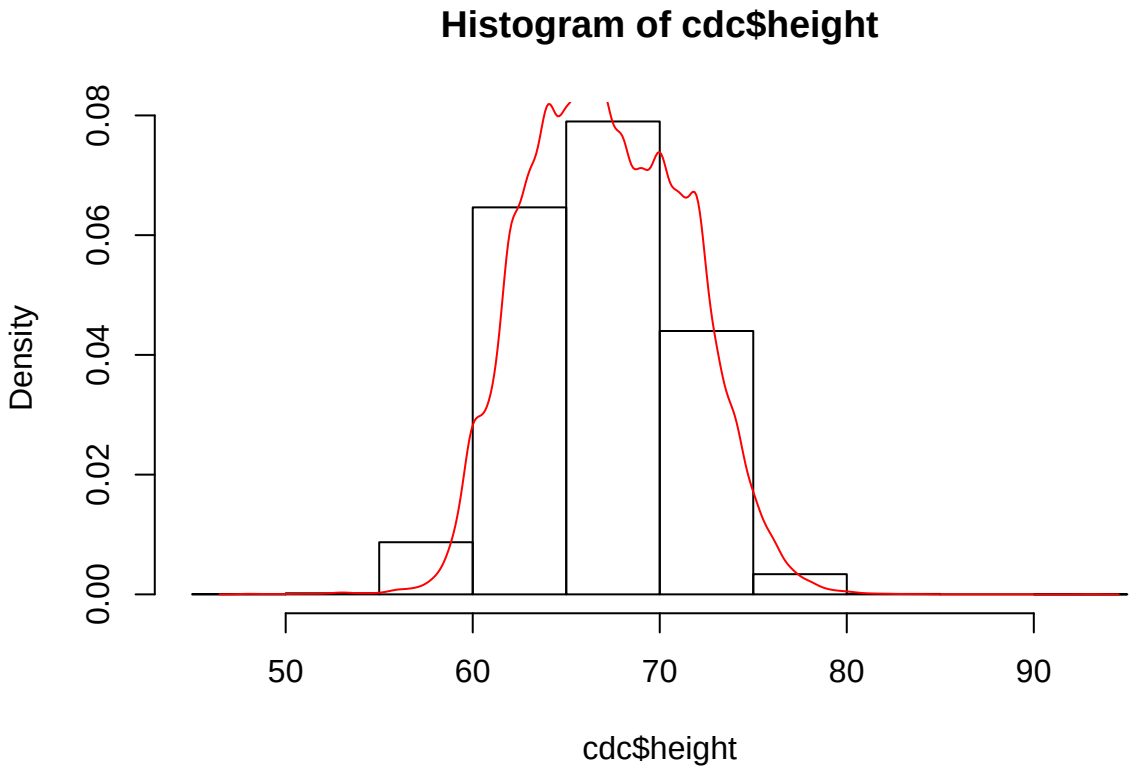

```r
hist(cdc$height, breaks=20, main="20 bins")
```



```r
hist(cdc$height, breaks=40, main="40 bins")
```



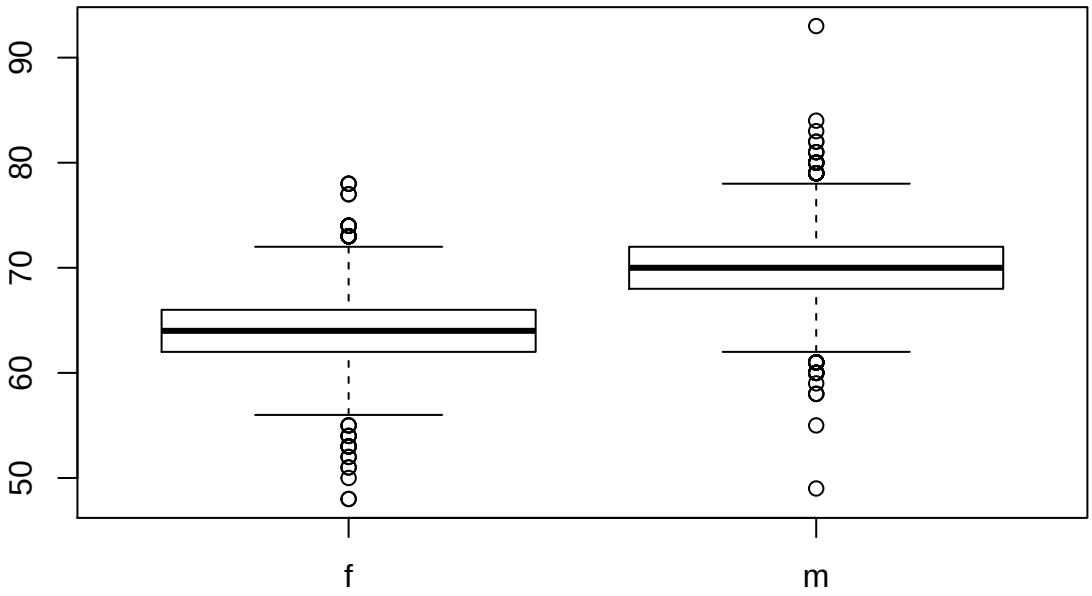

```r
#2c)
hist(cdc$height, freq=F)
lines(density(cdc$height), col="red")
```



#Exercise

3

```
boxplot(cdc$height ~ cdc$gender)
```



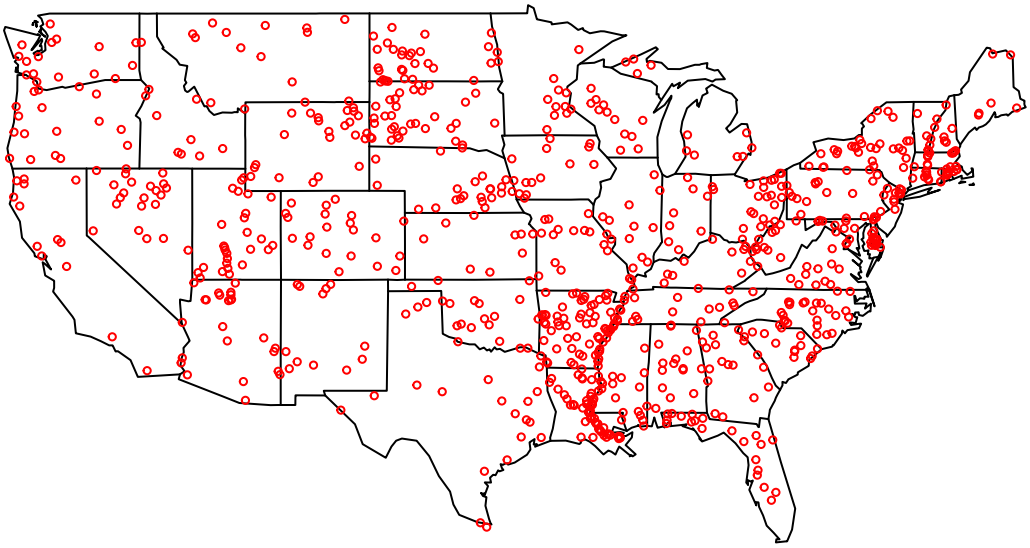
There are plenty of outliers in both directions for both genders. The distributions appear to have roughly similar shapes, with females being shifted down about 6 inches

Extra Credit

```
library(maps)
```

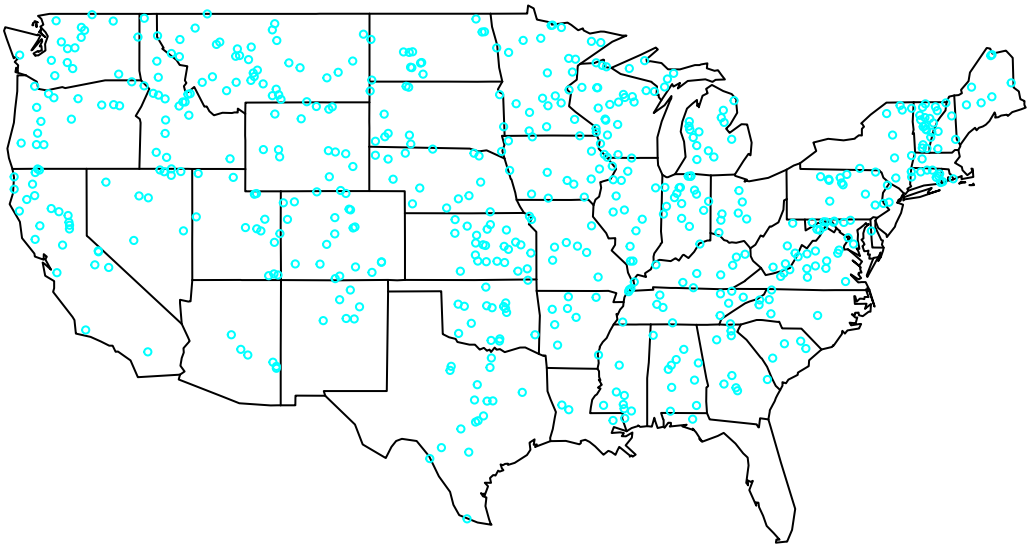
```
## Warning: package 'maps' was built under R version 3.4.4
```

```
data_url <- "https://github.com/ericwfox/stat630data/raw/master/nrsa.csv"
nrsa <- read.csv(data_url, header = TRUE)
map("state")

points(nrsa$lon, nrsa$lat, cex=0.5, col=c("red","transparent","transparent")[nrsa$cond2])
```



```
#Plotting only the stream sites that are in "Poor" condition
```

```
map("state")

points(nrsa$lon, nrsa$lat, cex=0.5, col=c("transparent","transparent","cyan")[nrsa$cond2])
```



```
#Plotting only the stream sites that are in "Good" condition
```