Homework 5

Grace Okamoto

29/9/2021

Problem 1

Create an account on GitHub (https://github.com) and create a repository for your personal dataset project. Submit the corresponding URL as for this problem.

https://github.com/not-caught-yet/DATA115

Problem 2

In your own words, write brief definitions of: ### (a) Mean The mean is the average of all the data values for a particular variable.

(b) Median

The median is the middlemost value of all the data values for a particular variable.

(c) IQR

IQR, or Inter-Quartile Range, is when the data values of a particular variable are divided into quartiles The difference between the 1st quartile and the third quartile is the IQR.

(d) Variance

The variance is the difference between the sum of each value and the mean squared and then divided by the number of values. This can be then un-squared to obtain the standard deviation.

(e) Skewness

Skewness is when the data is not symmetrically distributed horizantally-wise, but instead has larger values on one or the other end of the graph.

Problem 3

Load the provided COL.csv dataset into R.

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
col <- read.csv("COL.csv")
head(col)
```

```
##        City Cappuccino Cinema  Wine Gasoline Avg.Rent Avg.Disposable.Income
## 1  Lausanne       3.15  12.59  8.40     1.32  1714.00               4266.11
## 2    Zurich       3.28  12.59  8.40     1.31  2378.61               4197.55
## 3    Geneva       2.80  12.94 10.49     1.28  2607.95               3917.72
## 4     Basel       3.50  11.89  7.35     1.25  1649.29               3847.76
## 5     Perth       2.87  11.43 10.08     0.97  2083.14               3358.55
## 6 Nashville       3.84  12.00 13.50     0.65  2257.14               3089.75
```

```
summary(col)
```

```
##      City             Cappuccino        Cinema            Wine
##  Length:216         Min.   :0.460   Min.   : 1.810   Min.   : 2.130
##  Class :character   1st Qu.:1.320   1st Qu.: 4.397   1st Qu.: 4.260
##  Mode  :character   Median :2.085   Median : 6.540   Median : 6.540
##                     Mean   :1.981   Mean   : 6.776   Mean   : 7.080
##                     3rd Qu.:2.490   3rd Qu.: 7.850   3rd Qu.: 8.473
##                     Max.   :4.480   Max.   :79.490   Max.   :26.150
##     Gasoline         Avg.Rent      Avg.Disposable.Income
##  Min.   :0.0700   Min.   : 120.7   Min.   :  12.0
##  1st Qu.:0.7350   1st Qu.: 609.0   1st Qu.: 528.9
##  Median :0.9500   Median : 980.6   Median :1535.4
##  Mean   :0.9989   Mean   :1093.0   Mean   :1408.0
##  3rd Qu.:1.3200   3rd Qu.:1388.1   3rd Qu.:2053.8
##  Max.   :1.6900   Max.   :5052.3   Max.   :4266.1
```

(a)

Decide which rows are outliers in this data and describe and justify how you determined their outlier status.

I used the summary to provide the quartiles I needed to determine the IQR, and then determined the lower and upper outliers by subtracting 1.5 x IQR plus or minus the lower and upper quartiles. I know there are other, more graphical ways of calculating this, but I chose the method I learned in statistics purely because looking at graphs is scary. Via this method, I determined that New York, Singapore, Sydney, Geneva, London, San Francisco, Stavanger, Brighton, Manama, Tehran, and Jakarta all had data values that were considered to be outliers.

(b) && (c)

For each row you identified, if you were performing EDA on this dataset, would you include its values in your analysis and plots? Why or why not? Honestly I would keep all of the values. If I were performing something such as finding the average cost of a cinema ticket, for instance, I might wish to remove the outliers, but since there isn't a clear purpose to the exploration of the data, I might wish to even consider the outlier values as more important, as it might enable us to inspect other variables that could cause these outlying values to be so deviant. For instance, perhaps one city's rent is so high because its geograhpical proximity to other areas causes it to be a prime target for real estate investment, thus pushing the other rent prices up. Such oddities such as this, I think, are important and shouldn't be ignored in the data.

Problem 4

Load the Height_Weight_Age_Sex.csv data into R.

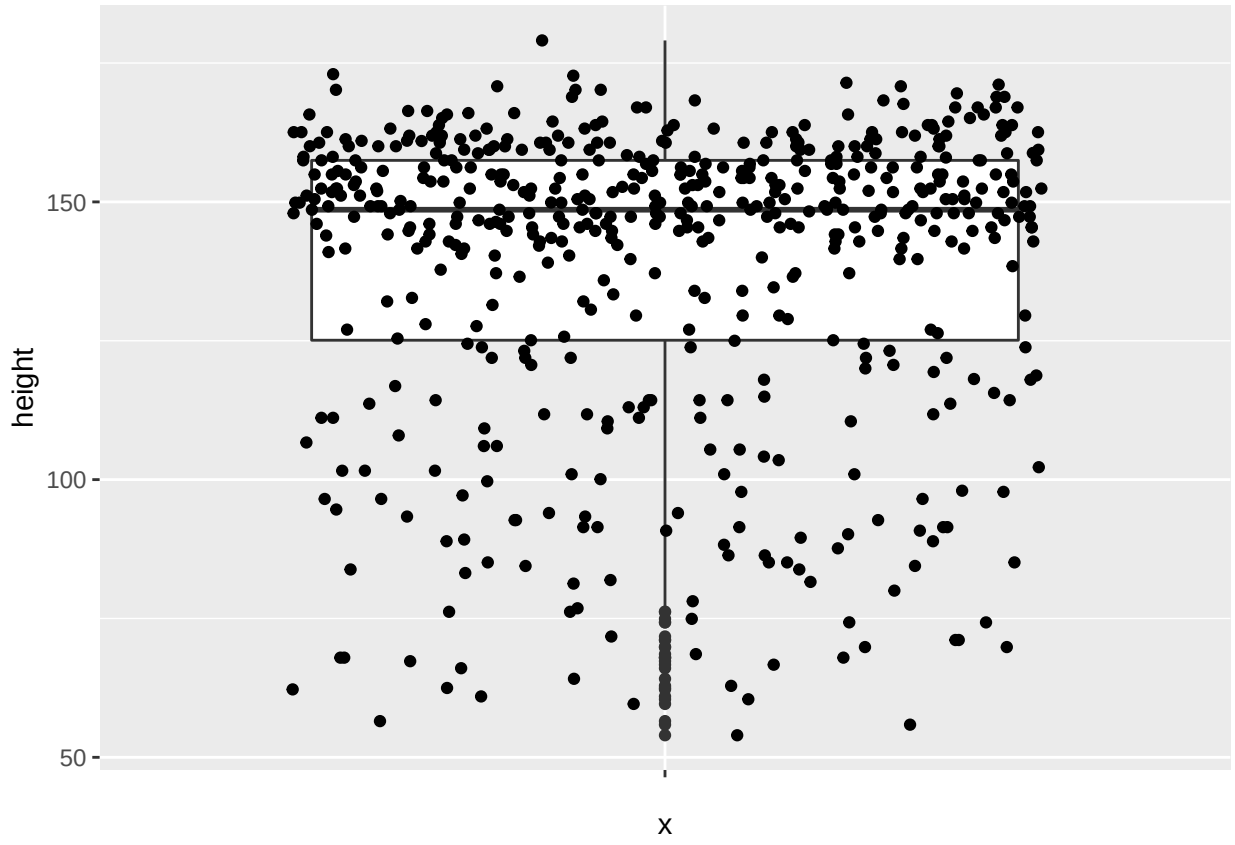
```
HWAS <- read.csv("Height_Weight_Age_Sex.csv")
head(HWAS)
```

```
##    height   weight age male
## 1 151.765 47.82561  63    1
## 2 139.700 36.48581  63    0
## 3 136.525 31.86484  65    0
## 4 156.845 53.04191  41    1
## 5 145.415 41.27687  51    0
## 6 163.830 62.99259  35    1
```

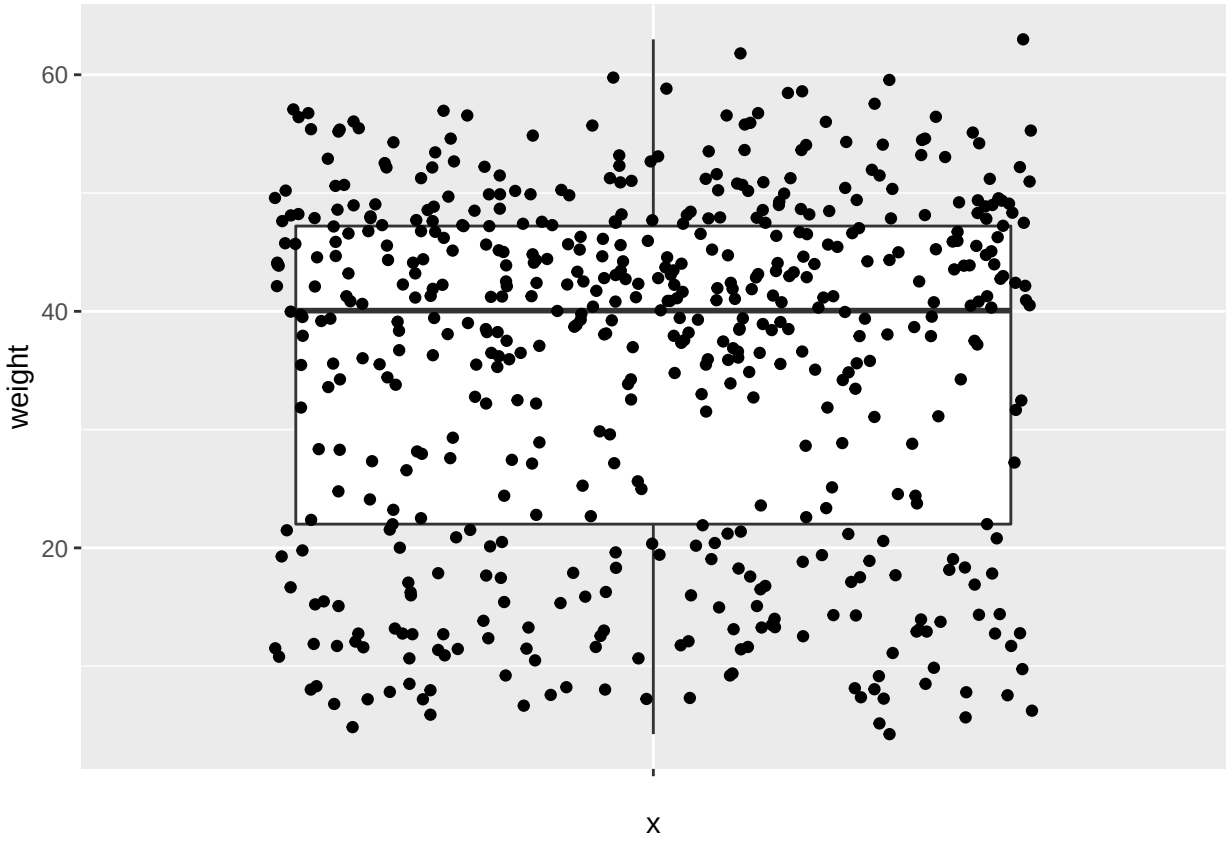
(a)

Create boxplots for the height and weight columns separately. Comment on the symmetry and sknewness, if any, for their distributions using these plots.

```
ggplot(data = HWAS, mapping = aes(x="", y=height)) + geom_boxplot() + geom_jitter()
```




```
ggplot(data = HWAS, mapping = aes(x="", y=weight)) + geom_boxplot() + geom_jitter()
```



For both height and weight, the boxplots were asymmetric and were instead skewed toward higher values. This observation makes sense, as since there is a cutoff point in which people generally stop growing (my cutoff point came a bit too early), even if the data was equally taken from all ages, the majority of the values would be at around that cutoff point, therefore skewing the mean upwards. However, since there are a fair number of underdeveloped specimens, the IQR would be larger, explaining the longer tail.

(b)

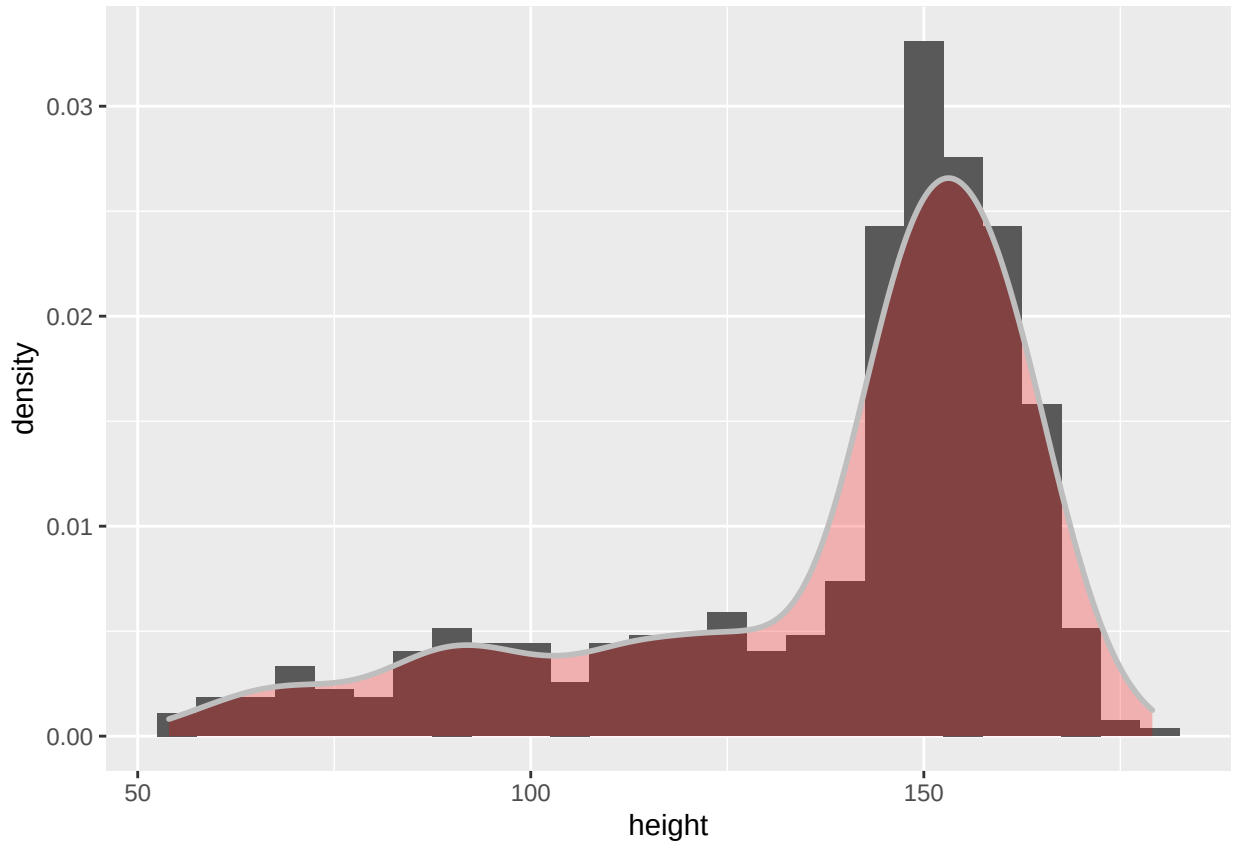
Create histograms for the height and weight columns separately. Comment on the symmetry and sknewness, if any, for their distributions using these plots. Are your conclusions based on the boxplots in (a) consistent with those based on densities?

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

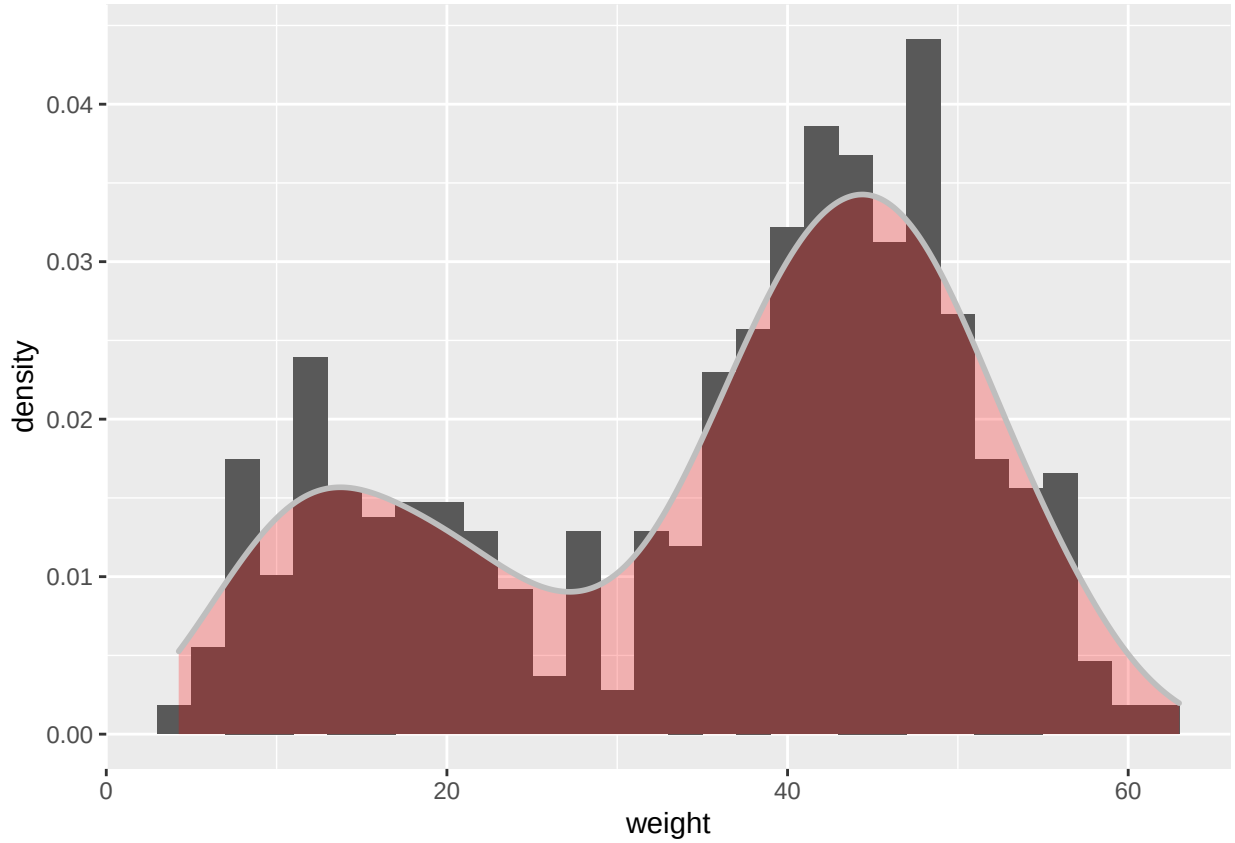
```
## Warning: package 'ggplot2' is in use and will not be installed
```

```
library("ggplot2")
```

```
ggplot(HWAS, aes(x=height)) +
  geom_histogram(aes(y = ..density..), binwidth=5) +
  geom_density(color="gray",fill="red",alpha=.25,size=1)
```




```
ggplot(HWAS, aes(x=weight)) +
  geom_histogram(aes(y = ..density..), binwidth=2) +
  geom_density(color="gray",fill="red",alpha=.25,size=1)
```

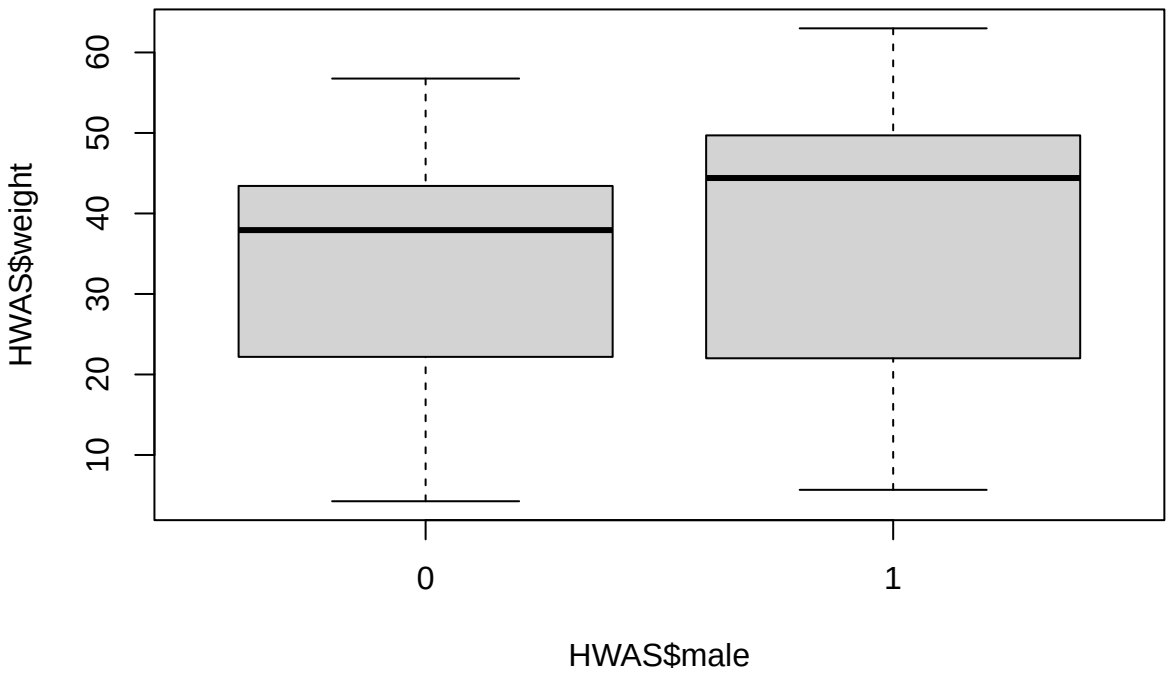


Although the general conclusion of these graphs is concurrent with that of the boxplots, interestingly, we are able to make some observations that weren't as apparent in the boxplot data. For instance, what looked like (to me, at least) a bunch of random dots in the boxplot now shows as a considerable hump and subsequent divot in the histogram.

(c)

Create separate boxplots for the weight data separated by the Male variable. What do you observe about the two distributions?

```
boxplot(HWAS$weight ~ HWAS$male)
```



Althought the average for the weight lands in approximately the same area for both boxplots with respect to the quartiles; however the third quartile and the maximum are lower in value for the female boxplot.

(d)

Add a BMI column and an underweight column to the data frame:

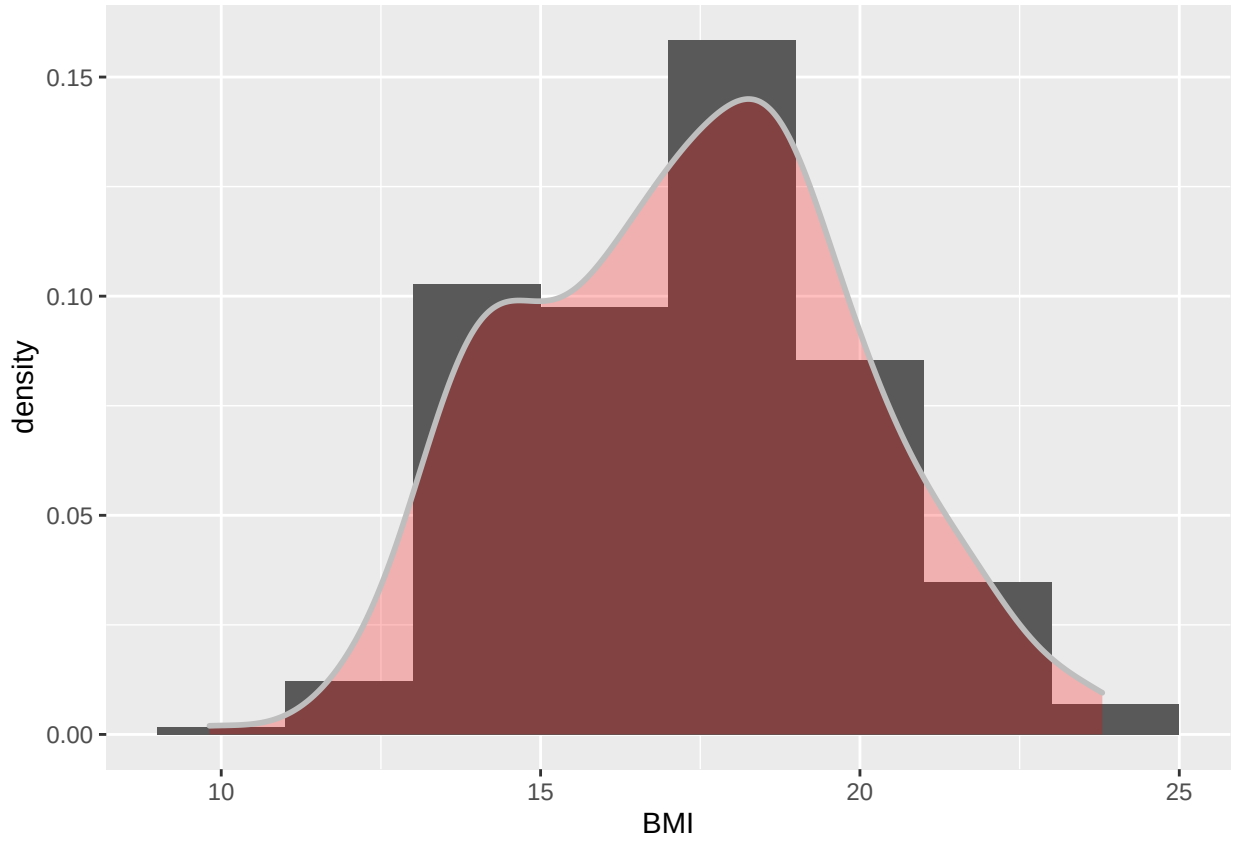
```
HWAS <- read.csv("./Height_Weight_Age_Sex.csv")

HWAS$BMI <- HWAS$weight/((HWAS$height/100)**2)

HWAS$underweight <- HWAS$BMI <18.5

head(HWAS)
```

```
##    height   weight age male      BMI underweight
## 1 151.765 47.82561  63    1 20.76430       FALSE
## 2 139.700 36.48581  63    0 18.69524       FALSE
## 3 136.525 31.86484  65    0 17.09572        TRUE
## 4 156.845 53.04191  41    1 21.56144       FALSE
## 5 145.415 41.27687  51    0 19.52038       FALSE
## 6 163.830 62.99259  35    1 23.46943       FALSE
```

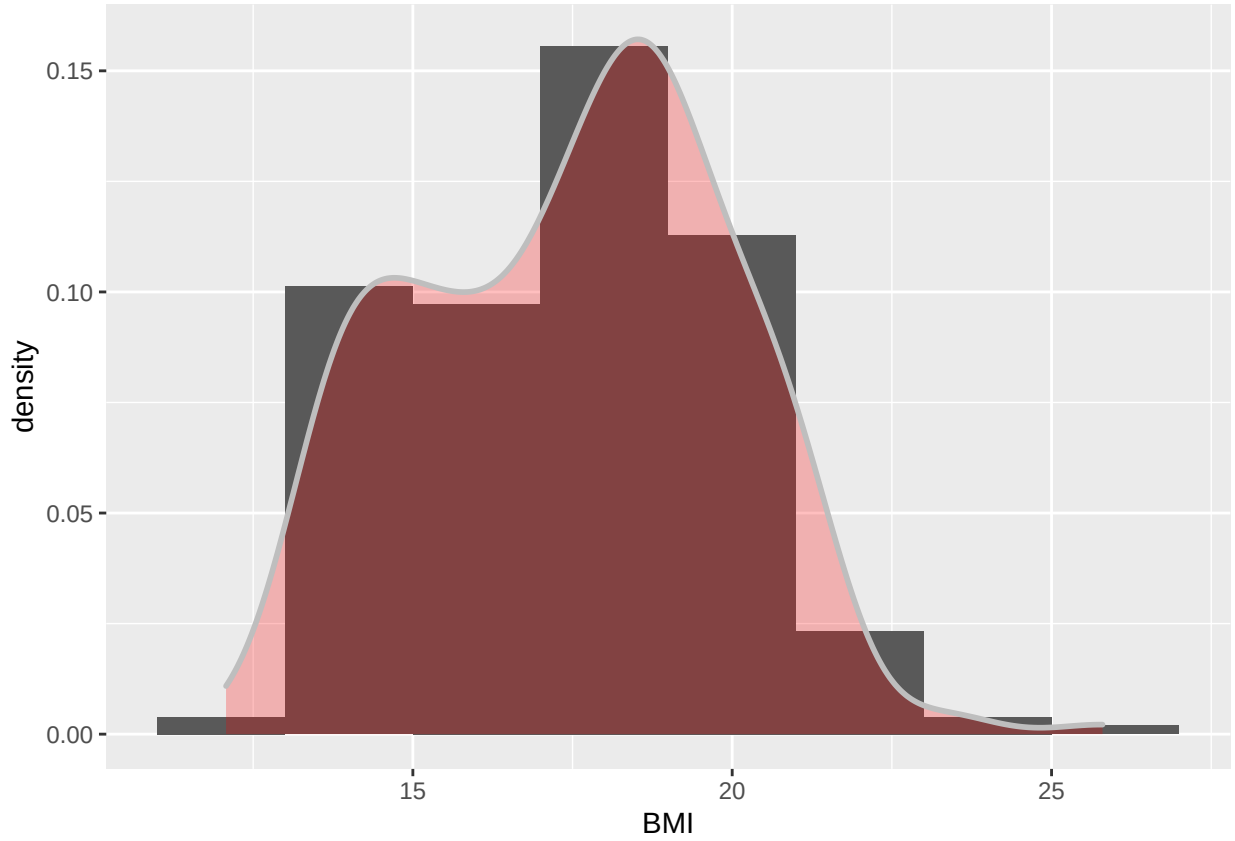
(e)

Create separate histograms for the BMI column separated by the Male variable. What do you observe about the two distributions?

```
ggplot(HWAS[HWAS$male==0,], aes(x=BMI)) +
  geom_histogram(aes(y = ..density..), binwidth=2) +
  geom_density(color="gray",fill="red",alpha=.25,size=1)
```




```
ggplot(HWAS[HWAS$male==1,], aes(x=BMI)) +
  geom_histogram(aes(y = ..density..), binwidth=2) +
  geom_density(color="gray",fill="red",alpha=.25,size=1)
```

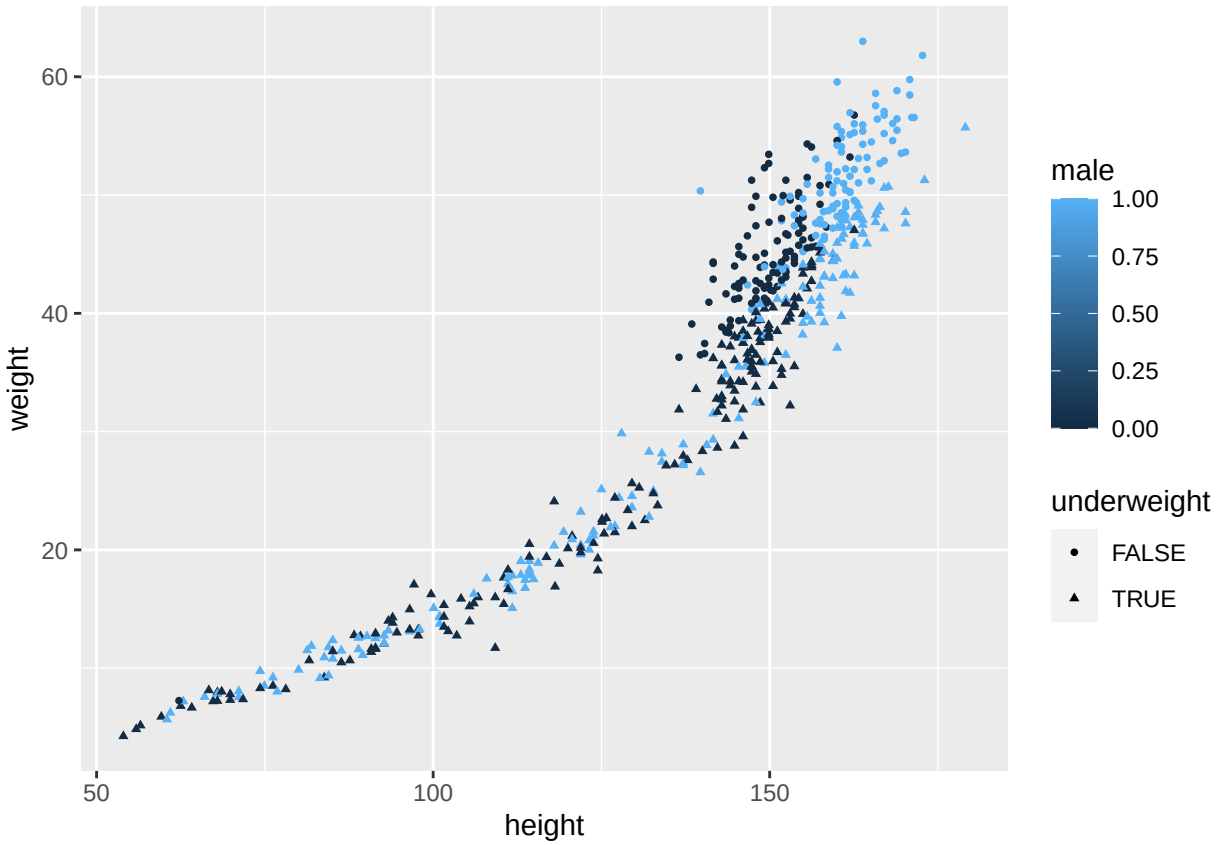


In general, it seems as if females have a higher mean BMI than men.

(f)

Make a scatterplot of height vs. weight for the full dataset that distinguishes both the Male variable and the under variable. What do you observe

```
ggplot(data = HWAS, mapping = aes(x=height, y=weight,  color = male, shape = underweight)) + geom_point
```



On average, it appears as though males weigh more and are taller. Understandably, as weight increases, there is less underweight people.

Problem 5

Read the following examples about Simpson's Paradox: How to lie with statistics? (you may also find the wikipedia page on the topic or the additional readings uploaded to Blackboard to be helpful resources). Fill in the following table with ratios of hits to attempts so that player A has a higher batting average in both season 1 and season 2 but player B has a higher overall batting average for the two seasons combined.

hits/attempts	Season 1	Season 2
Player A	25/100	1/2
Player B	49/200	99/200