Homework 5

Rhea Toves

9/28/2021

_

Problem 1 - Create an account on GitHub (https://github.com (https://github.com)) and create a repository for your personal dataset project. Submit the corresponding URL as for this problem.

(https://github.com/rheatoves/DATA-115.git (https://github.com/rheatoves/DATA-115.git))

_

Problem 2 - In your own words, write brief definitions of:

(a) Mean

The sum of the values given, divided by the total number of integers in the dataset.

(b) Median

After listing all the values from least to greatest, the median will be the number that falls in the middle of the dataset.

(c) IQR

This abbreviation stands for Interquartile Range. The IQR is calculated by the total difference between Q3 - Q1.

(d) Variance

This measures dispersion, and takes the mean/average into account when calculating how far the span of numbers are, from the mean/average.

(e) Skewness

This is when asymmetry of the probability gets taken into account and this measures the depth of the bell shaped curve.

_

Problem 3 - Load the provided COL.csv dataset into R.

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

The rows, “Cappuccino”, “Cinema”, and “Wine” are outliers in this data set. I chose these rows as outliers because the dataset provides more important information pertaining to each city, such as, the city, average rent, disposable income, and gasoline (gives us information on how much cost of living and travel time). The relevant rows listed above help us make sense of the average rent, disposable income, and cost of living - for each city.

(b) - For each row you identified, if you were performing EDA on this dataset, would you include its values in your analysis and plots?

For the rows I have identified, I would not use them if I were to perform EDA on this dataset.

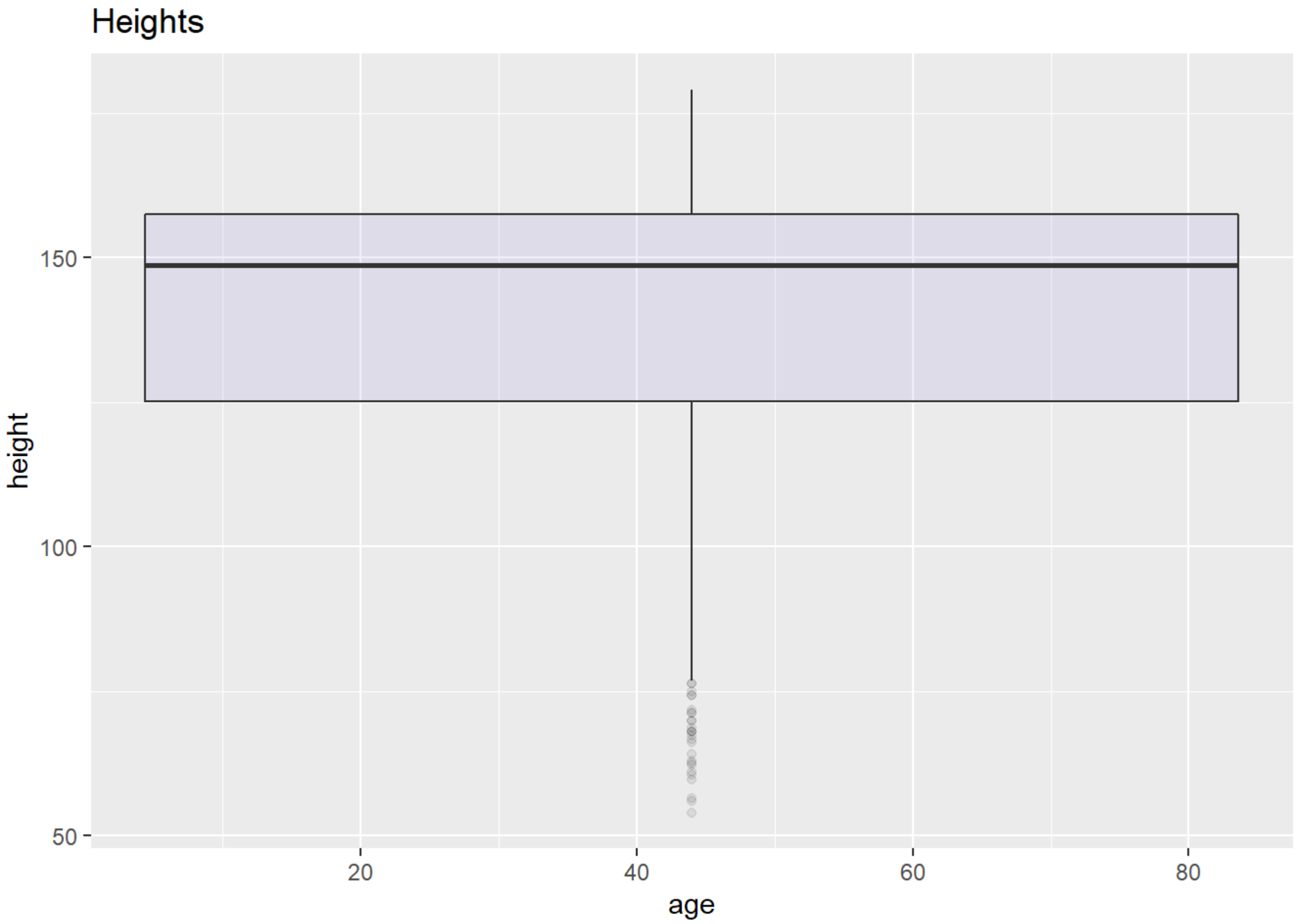
(c) - Why or why not?

I would choose not to include this in EDA because the rows, “Cappuccino”, “Cinema”, and “Wine” seem to not have significance pertaining to the average rent, disposable income, and cost of living aspects of this dataset. If I were to leave out the said rows, this would help portray my idea better and clearer.

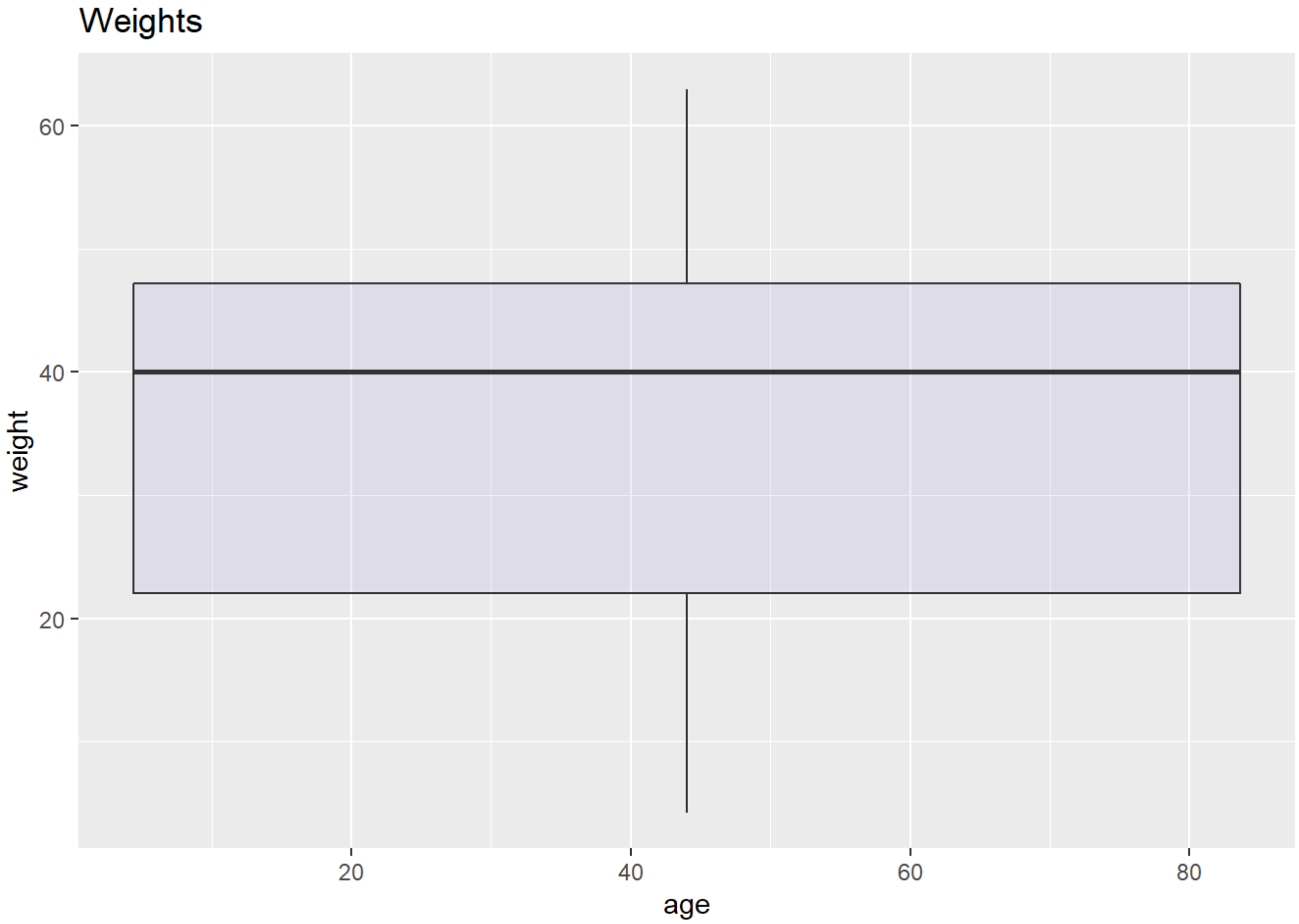
_

Problem 4 - Load the Height_Weight_Age_Sex.csv data into R.

(a) - Create boxplots for the height and weight columns separately. Comment on the symmetry and skewness, if any, for their distributions using these plots.

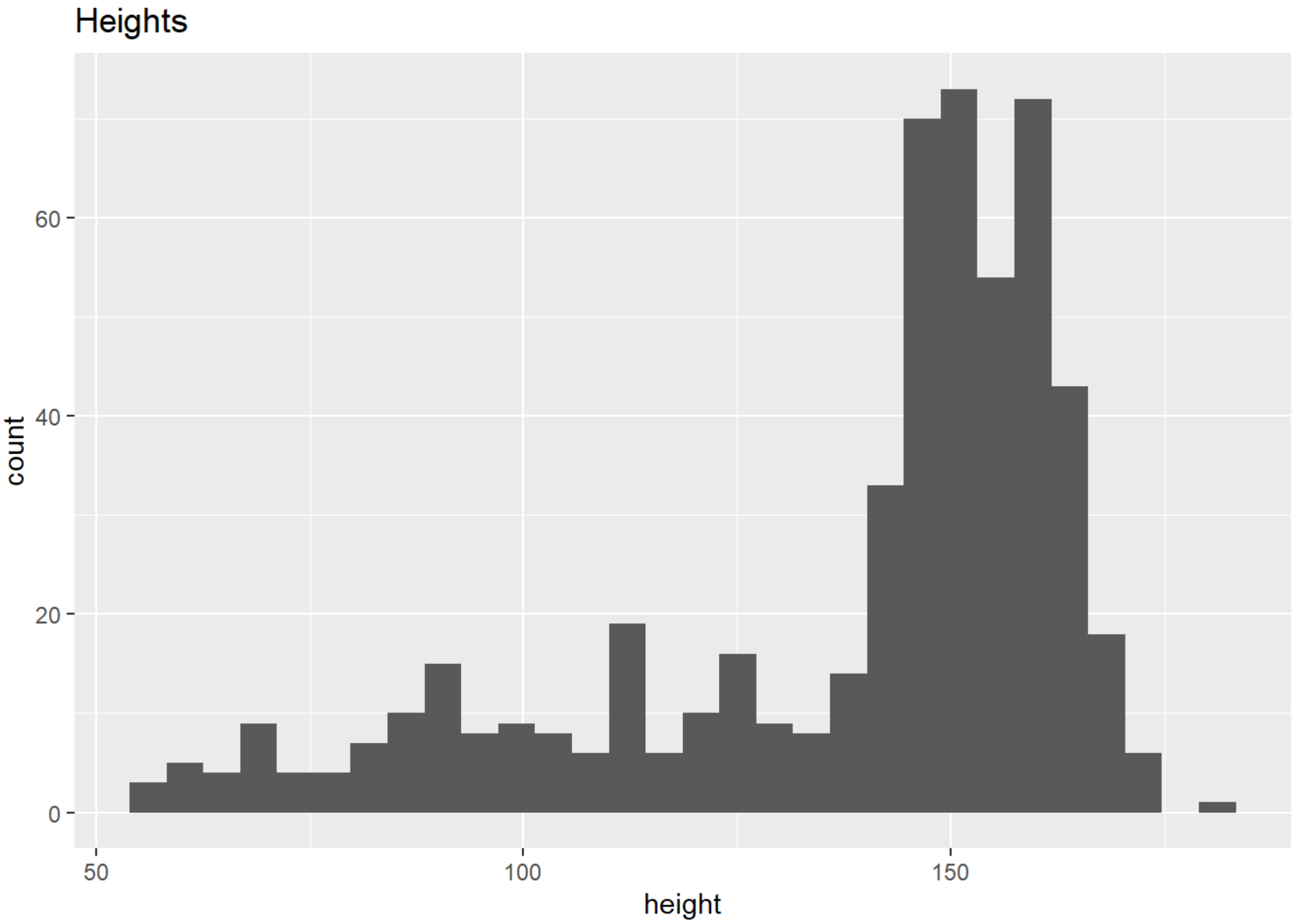


It was interesting to see the median placed so close to the Q1 and the Q3 fall far from the median. This kind of shows how dispersed the data is. Since the data is closer together on one side of the plot, the skewness of the graph does not seem to be high.

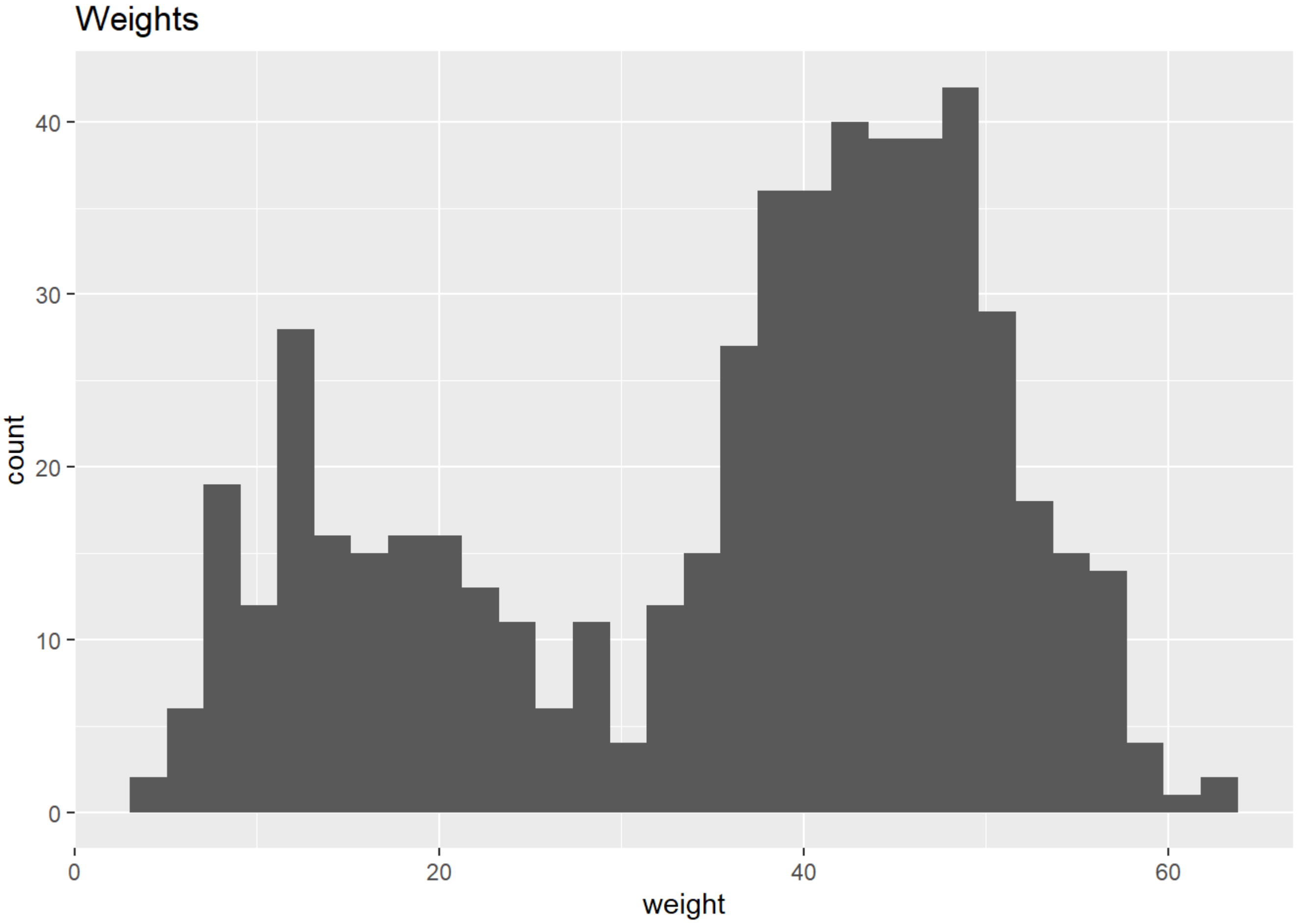


For the weight boxplot, the data was more dispersed but same thing, the distance between the median and Q1 is shorter than the distance between Q3 and the median. Since the data is closer together on one side of the plot, the skewness of the graph does not seem to be high.

(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?

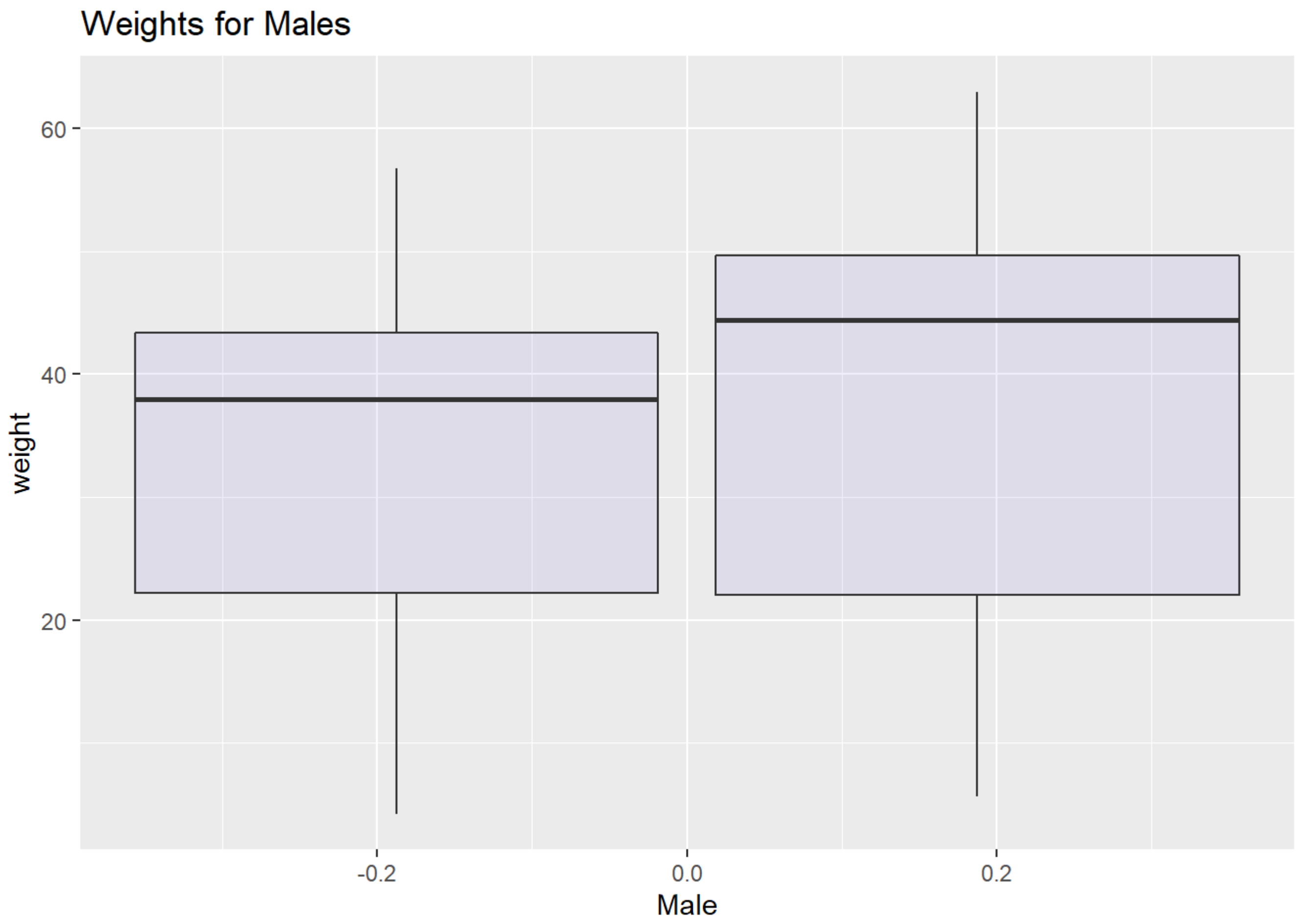


For the Heights histogram, there is one huge increase in the dataset towards the far right side. The data seems to creep up the graph for a while, trailing behind the large bell-shaped curve. The skewness of the curve seems to be relatively symmetric.



For the Weights histogram, there are two bell-shaped humps in the data. The weight seems to peak in the beginning, tank in the middle of the dataset and then peak again. There is symmetry within the graph itself (two bell shapes), the skewness appears to be quite wide for both bell-shaped curves.

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

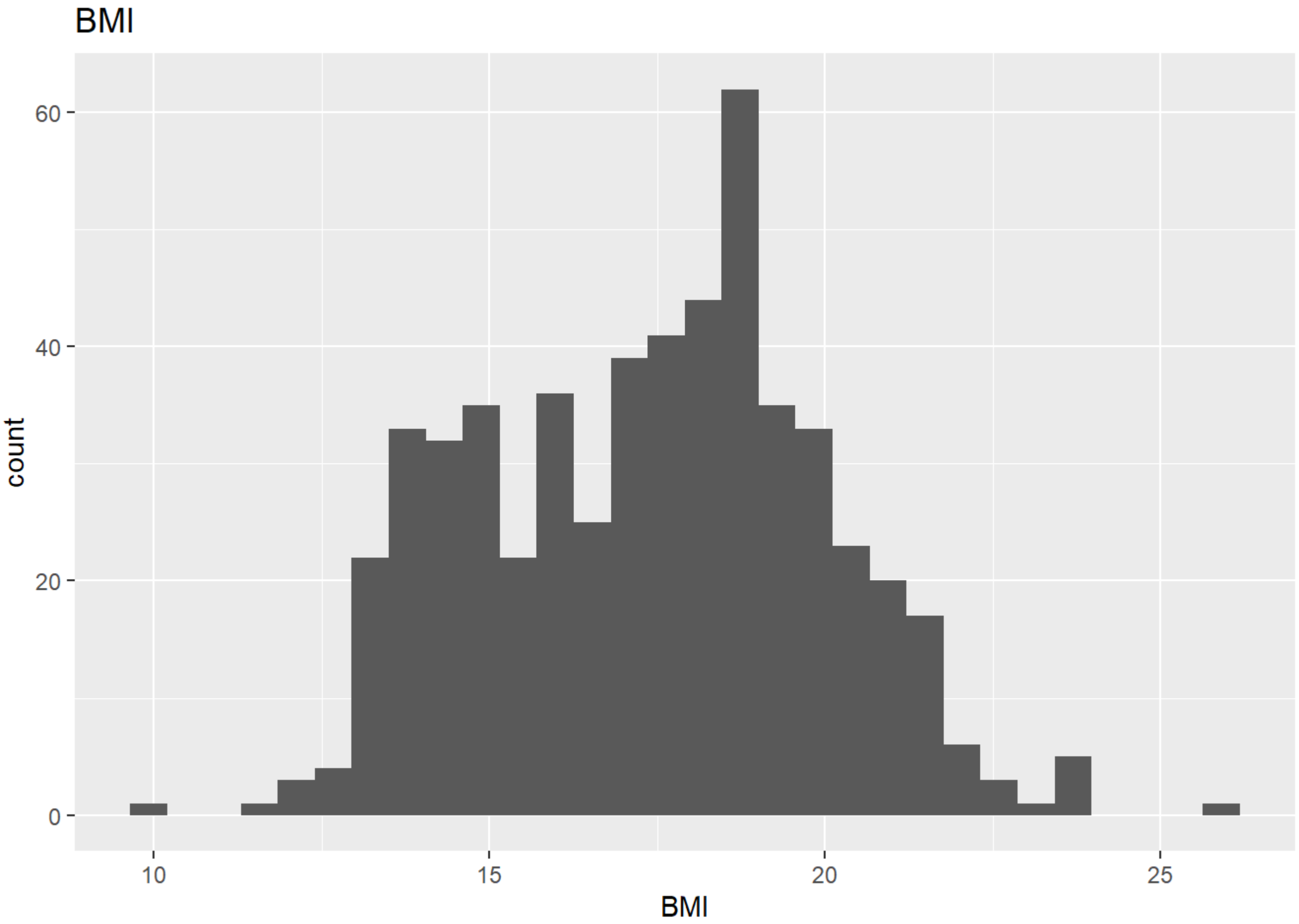


Both boxplots seem to have an uneven amount of distribution between Q1 and the median and Q3 and the median. Both plots carry more distance between Q3 and the median, rather than Q1 and the median. This hints to a wide bell-shaped curve.

(d) - Add a BMI column and an underweight column to the data frame:

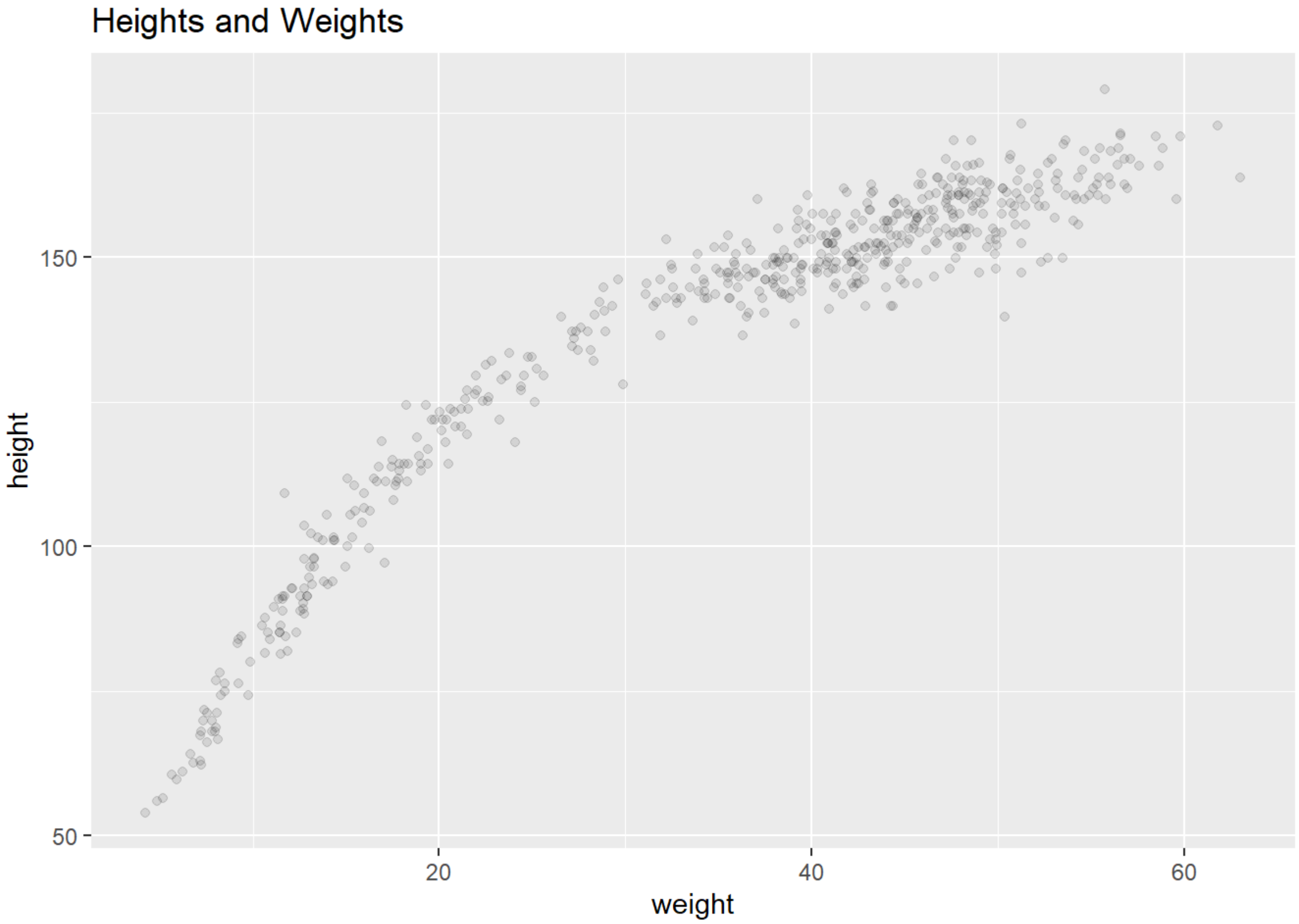
```
##    height   weight age male      BMI underweight
## 1 151.765 47.82561  63    1 20.76430       FALSE
## 2 139.700 36.48581  63    0 18.69524       FALSE
## 3 136.525 31.86484  65    0 17.09572        TRUE
## 4 156.845 53.04191  41    1 21.56144       FALSE
## 5 145.415 41.27687  51    0 19.52038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



This histogram appears to be very centered in the middle of the graph. The dataset creates a very wide bell-shaped curve, meaning that the dataset is quite dispersed.

(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?



The plotting on this graph seems to start low and slowly peak its way up. I added a color to the chart to clearly observe any overlapping with the data points. Both the heights and weights datasets are increasing and moving in the same direction. Points seem to disburse towards the right-hand side of the graph, possibly because the data points are growing.