Lab 3

Jordan Lian

10/18/2020

Problem Statment

My understanding of the assignments comes from the first few chapters, where we worked with descriptive statistics. I mainly used the packet to solve the tasks for the assignment, which is what I usually do to complete the lab assignments.

Output

Task 1

Import the Health Dataset and save it as a data frame.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```
health_df <- read_csv('Health_Dataset.csv')
```

```
## Parsed with column specification:
## cols(
##   Age = col_double(),
##   Avg_Glucose_Level = col_double(),
##   BMI = col_double(),
##   Stroke = col_logical(),
##   Hypertension = col_logical(),
##   Heart_Diseases = col_logical()
## )
```

```r
str(health_df)
```

```
## tibble [1,000 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Age              : num [1:1000] 18 32 20 60 61 18 28 27 34 11 ...
##  $ Avg_Glucose_Level: num [1:1000] 107.4 67.2 83.5 102.8 96.1 ...
##  $ BMI              : num [1:1000] 31.4 31.2 23.6 29.7 24.4 25.7 38.7 49.2 23.5 19.5 ...
##  $ Stroke           : logi [1:1000] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Hypertension     : logi [1:1000] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Heart_Diseases   : logi [1:1000] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Age = col_double(),
##   ..   Avg_Glucose_Level = col_double(),
##   ..   BMI = col_double(),
##   ..   Stroke = col_logical(),
##   ..   Hypertension = col_logical(),
##   ..   Heart_Diseases = col_logical()
##   .. )
```

Task 2 (30 points)

a) Find the frequency of people in each of the BMI categories depicted in the image below.

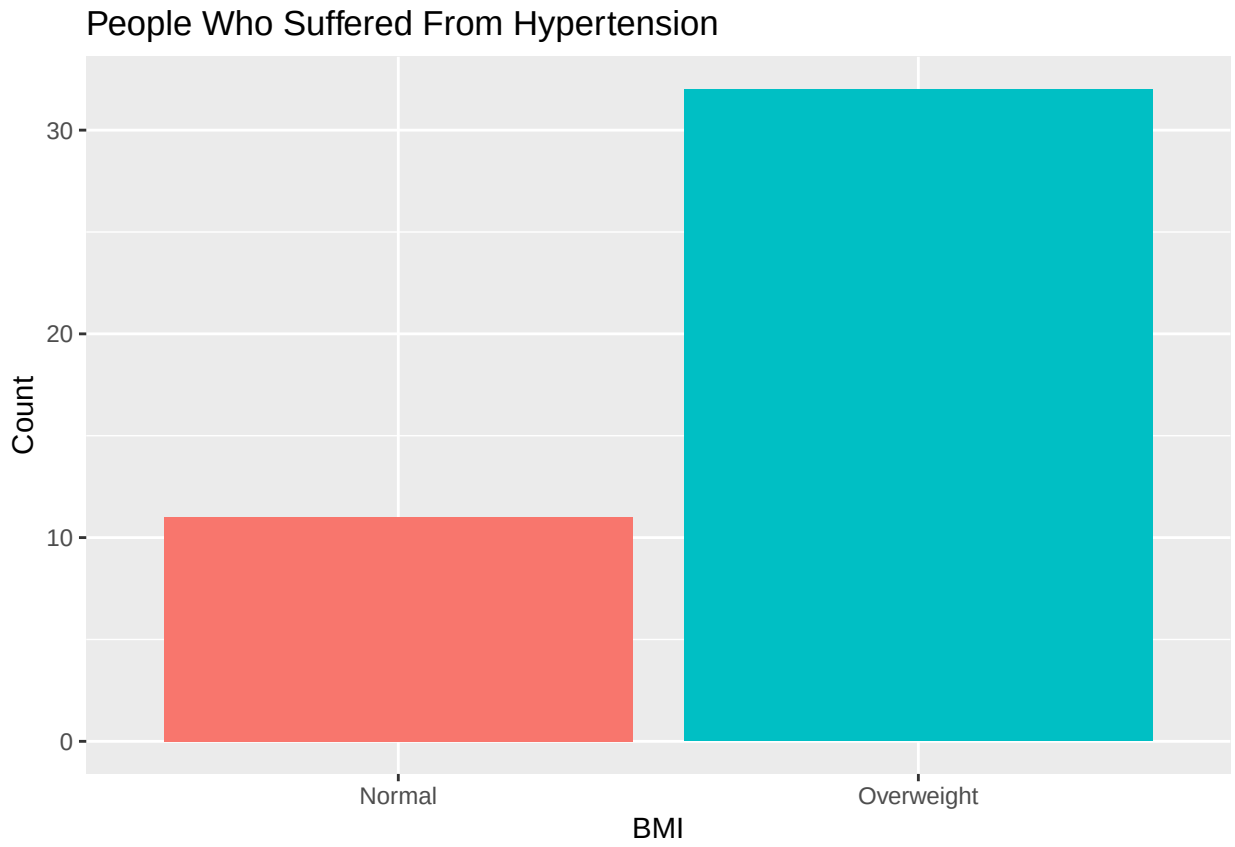
Category	BMI
Underweight	BMI < 18.5
Normal	18.5 <= BMI < 25
Overweight	25 <= BMI < 30
Obese	BMI >= 30

```r
# Initialize New Column
health_df$BMI_Level <- health_df$BMI

# Conditions
health_df$BMI_Level[health_df$BMI < 18.5] <- "Underweight"
health_df$BMI_Level[health_df$BMI >= 18.5 & health_df$BMI < 25] <- "Normal"
health_df$BMI_Level[health_df$BMI >= 25 & health_df$BMI < 30] <- "Overweight"
health_df$BMI_Level[health_df$BMI >= 30] <- "Obese"

# Print Table
table(health_df$BMI_Level)
```

```
##
##      Normal       Obese  Overweight Underweight
##         244         429         263          64
```

b) Find the number of people under the normal and overweight category who suffered from Hypertension (Hypertension == TRUE) and compare their results. Use barplot to illustrate your answer. What can you interpret from the barplot? Hint: Use geombar() to get the barplot

```r
# Set up data
HT_df <- health_df[health_df$BMI >= 18.5 & health_df$BMI < 30 & health_df$Hypertension == TRUE, ]
HT_table <- as.data.frame(table(HT_df$Hypertension, group_by = HT_df$BMI_Level))
HT_table <- HT_table[, -1]
names(HT_table) <- c("BMI", "Count")

# Plot
ggplot(HT_table, aes(BMI, Count, fill = BMI)) +
  geom_col() +
  ggtitle("People Who Suffered From Hypertension") +
  theme(legend.position = "none")
```



c) Calculate the following statistics for Normal BMI.

- Mean
- Median
- Range
- Interquartile Range
- Variance
- Standard Deviation

```r
# Normal Data frame
normal_df <- health_df[health_df$BMI_Level == "Normal", ]
```

```r
# Mean
mean(normal_df$BMI)
```

```
## [1] 22.14508
```

```r
# Median
median(normal_df$BMI)
```

```
## [1] 22.45
```

```r
# Range
range(normal_df$BMI)
```

```
## [1] 18.5 24.9
```

```r
# Interquartile Range
quantile(normal_df$BMI)
```

```
##     0%    25%    50%    75%   100%
## 18.500 20.500 22.450 23.825 24.900
```

```r
# Variance
var(normal_df$BMI)
```

```
## [1] 3.701416
```

```r
# Standard Deviation
sd(normal_df$BMI)
```

```
## [1] 1.923906
```

Task 3 (20 points)

a) Find the percentage of people when each of following conditions are met,

Stroke	Hypertension
TRUE	TRUE
TRUE	FALSE
FALSE	TRUE
FALSE	FALSE

Obtain a table as shown below.

Stroke	Hypertension	Percentage
TRUE	TRUE	%
TRUE	FALSE	%

Stroke	Hypertension	Percentage
FALSE	TRUE	%
FALSE	FALSE	%

```r
# True, True
TT <- count(health_df[health_df$Stroke == TRUE & health_df$Hypertension == TRUE, ])

# True, False
TF <- count(health_df[health_df$Stroke == TRUE & health_df$Hypertension == FALSE, ])

# False, True
FT <- count(health_df[health_df$Stroke == FALSE & health_df$Hypertension == TRUE, ])

# False, False
FF <- count(health_df[health_df$Stroke == FALSE & health_df$Hypertension == FALSE, ])

# Total Sum for Percentages
total_sum <- TT + TF + FT + FF
TT <- TT / total_sum
TF <- TF / total_sum
FT <- FT / total_sum
FF <- FF / total_sum

# Table -- Original Matrix
A <- matrix(
  c("TRUE", "TRUE", "TRUE", "FALSE", "FALSE", "TRUE", "FALSE", "FALSE"),
  nrow = 4,
  ncol = 2,
  byrow = TRUE
)
colnames(A) <- c("Stroke", "Hypertension")

# 3rd Column
B <- matrix(
  c(TT, TF, FT, FF),
  nrow = 4,
  ncol = 1,
  byrow = TRUE
)
colnames(B) <- "Percentage"

# Combine
C <- cbind(A, B)

# Final Table
C
```

```
##      Stroke  Hypertension Percentage
## [1,] "TRUE"  "TRUE"       0.003
## [2,] "TRUE"  "FALSE"      0.01
## [3,] "FALSE" "TRUE"       0.084
## [4,] "FALSE" "FALSE"      0.903
```

Task 4 (30 points)

a) Calculate the following for the Avg_Glucose_Level

- Coefficient of Variation
- Skewness
- Kurtosis

```
# Library
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.0.3
```

```
# Coefficient of Variation
sd(health_df$Avg_Glucose_Level) / mean(health_df$Avg_Glucose_Level) * 100
```

```
## [1] 41.92766
```

```
# Skewness
skewness(health_df$Avg_Glucose_Level)
```

```
## [1] 1.637986
```

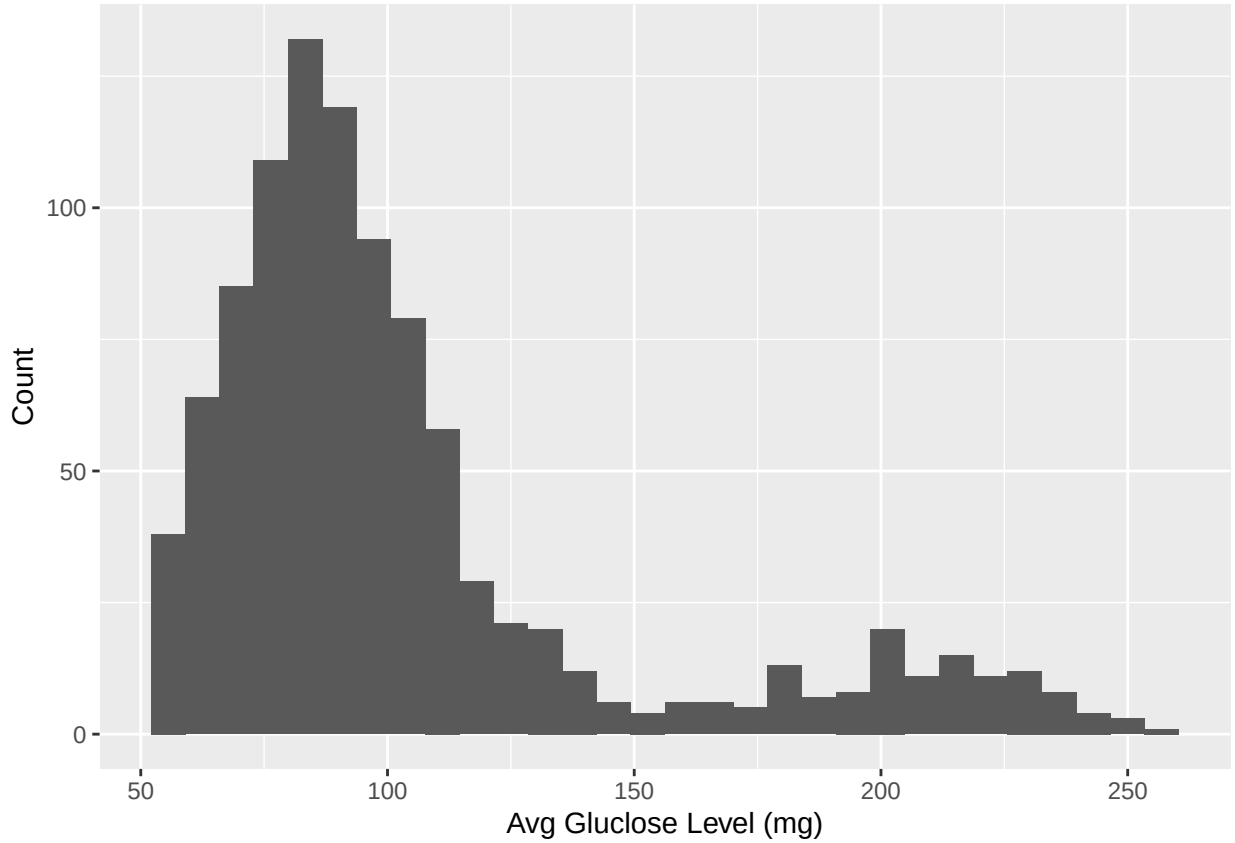
```
# Kurtosis
kurtosis(health_df$Avg_Glucose_Level)
```

```
## [1] 1.925254
```

b) Plot a histogram of Avg_Glucose_level.

```
ggplot(data = health_df) +
  geom_histogram(mapping = aes(x = Avg_Glucose_Level)) +
  xlab('Avg Gluclose Level (mg)') +
  ylab("Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



c) Compare the histogram and the results obtained in part (a) and interpret the results.

The coefficient of variation is the ratio of the standard deviation to the mean. A 41% coefficient value indicates a solid amount of variability in the data, which is shown in the highly skewed histogram.

The skewness is greater than 1, which is shows that the data is highly skewed. In this case, it is highly skewed to the left. This indicates that the median is greater than the mode, while it is less than the mean.

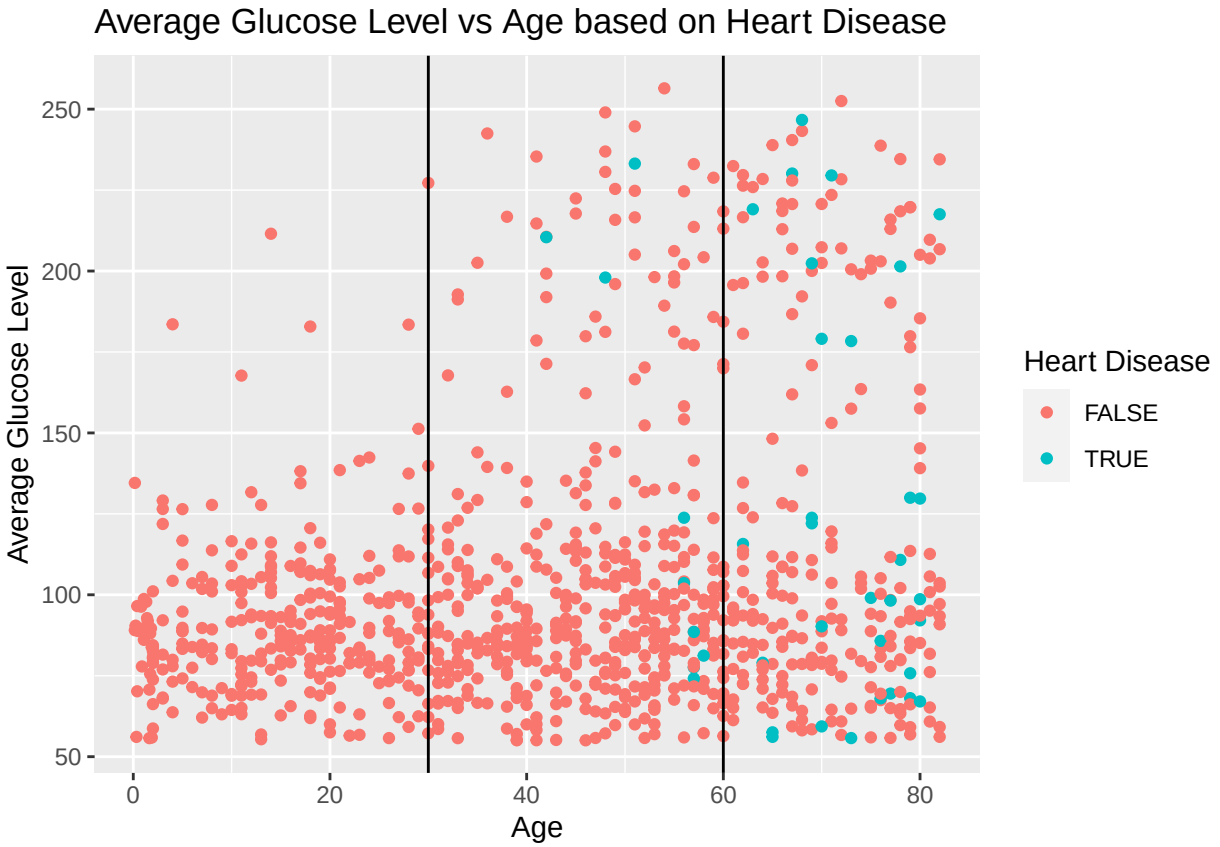
The Kurtosis value is positive, which means distribution has longer tails, which indicates a heavier presence of outliers. This is indicative in the histogram which is heavily skewed to the left.

Task 5 (10 points)

a) Plot a scatter plot for Age and Avg_Glucose_Level for people who suffered from Heart_diseases (Heart_diseases = TRUE) and those who did not (Heart_diseases = FALSE) geomvline() and geomhline() functions so data can be categorized in the following age groups

- Age < 30
- $30 <$ Age < 60
- Age > 60

```
ggplot(data = health_df) +
  geom_point(mapping = aes(x = Age, y = Avg_Glucose_Level, color = Heart_Diseases)) +
  geom_vline(xintercept = 30) +
  geom_vline(xintercept = 60) +
  ylab("Average Glucose Level") +
  ggtitle("Average Glucose Level vs Age based on Heart Disease") +
  labs(color = "Heart Disease")
```



What can you infer from the scatter plot?

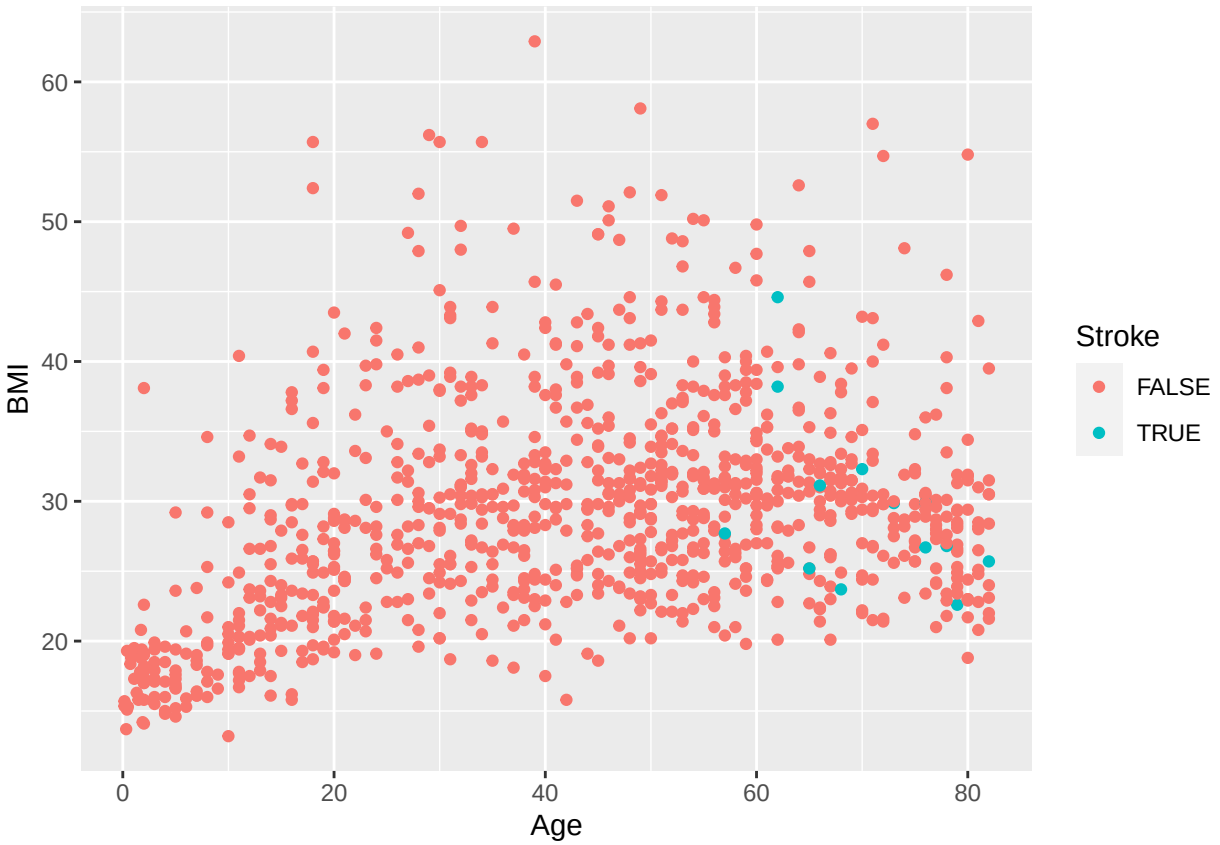
Firstly, I can infer that most of the people in the dataset did not suffer from heart disease. However, those that did (the ones in blue) were older on average, as there was a rightward skew of the blue data points. The majority of them were over 60, as shown by the X intersect line. The X intersect line of X = 30 does not reveal much.

Task 6 (10 points)

a) Analyze the data and discover any interesting facts (at most 2) about the dataset and relationship between variables.

Note: The analysis should be unique and not exceed more than one page.

```
ggplot(data = health_df) +
  geom_point(mapping = aes(x = Age, y = BMI, color = Stroke))
```



I created a scatter plot contrasting age and BMI (instead of Average Glucose Level), while seeing how many had strokes. I found that most people who got strokes were on the older side, just as it was with heart diseases. I thought that the correlation might be potentially different, but it was not meant to be. Most of the conclusions I found were pretty standard when it came to health facts. Older people are more vulnerable to strokes, hypertension, and heart disease, and the higher your BMI, the higher your chances are of getting those problems. However, the strongest correlation was age, which I found slightly surprising, although I am ignorant on the subject of health data.

Conclusion

Most of my conclusions are stated in the output section. However, the one thing I noticed over the entire lab was that the data was widely skewed, in that the majority of the sample set of people consisted of middle-aged people and senior people, who were far more vulnerable to the problems like strokes, heart disease, and hypertension. This explains why a lot of the visuals did show any normal distributions, rather skewed distributions. Overall this lab helped me understand R better while also applying my statistics knowledge from class, which is the ultimate goal for these lab assignments.