Costa Stavrianidis Homework 2

In this homework, the objectives are to

1. Work with dates in R and exploratory data visualization
2. Data Preprocessing and Transformation

Assignments will only be accepted in electronic format in RMarkdown (.rmd) files and knitted .html files. Please make sure to **print your knitted .html file into a pdf** before you submit it to the Gradescope, and you may only submit your .rmd file to Sakai. Your code should be adequately commented to clearly explain the steps you used to produce the analyses. RMarkdown homework files should be uploaded to Sakai with the naming convention date_lastname_firstname_HW[X].Rmd. For example, my second homework assignment would be named 20220922_Dunn_Jessilyn_HW2.Rmd. **It is important to note that 5 points will be deducted for every assignment that is named improperly.** Please add your answer to each question directly after the question prompt in the homework .Rmd file template provided below.

Working with Dates and EDA (25 points)

1. From the stroke dataset in the file “healthcare-dataset-stroke-data.csv”, create a new dataframe named “stroke_df” that only contains information on the subjects’ age, gender, hypertension, heart_disease, work type, BMI, smoking status, average glucose level (avg_glucose_level) as well as whether or not the patient had a stroke.

```
stroke <- read_csv("healthcare-dataset-stroke-data.csv",
                   show_col_types=FALSE)

# Selecting desired variables
stroke_df <- stroke %>% select(c(age, gender, hypertension, heart_disease,
                                 work_type, bmi, smoking_status, avg_glucose_level,
                                 stroke))
glimpse(stroke_df)
```

```
## Rows: 5,109
## Columns: 9
## $ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, …
## $ gender            <chr> "Male", "Female", "Male", "Female", "Female", "Male"…
## $ hypertension      <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1…
## $ heart_disease     <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0…
## $ work_type         <chr> "Private", "Self-employed", "Private", "Private", "S…
## $ bmi               <dbl> 36.6, NA, 32.5, 34.4, 24.0, 29.0, 27.4, 22.8, NA, 24…
## $ smoking_status    <chr> "formerly smoked", "never smoked", "never smoked", "…
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0…
## $ stroke            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
```

2. Cast all columns into the most sensible variable type, e.g., “age” should be an integer, “gender” should be a factor, “stroke” should be a logical, etc.

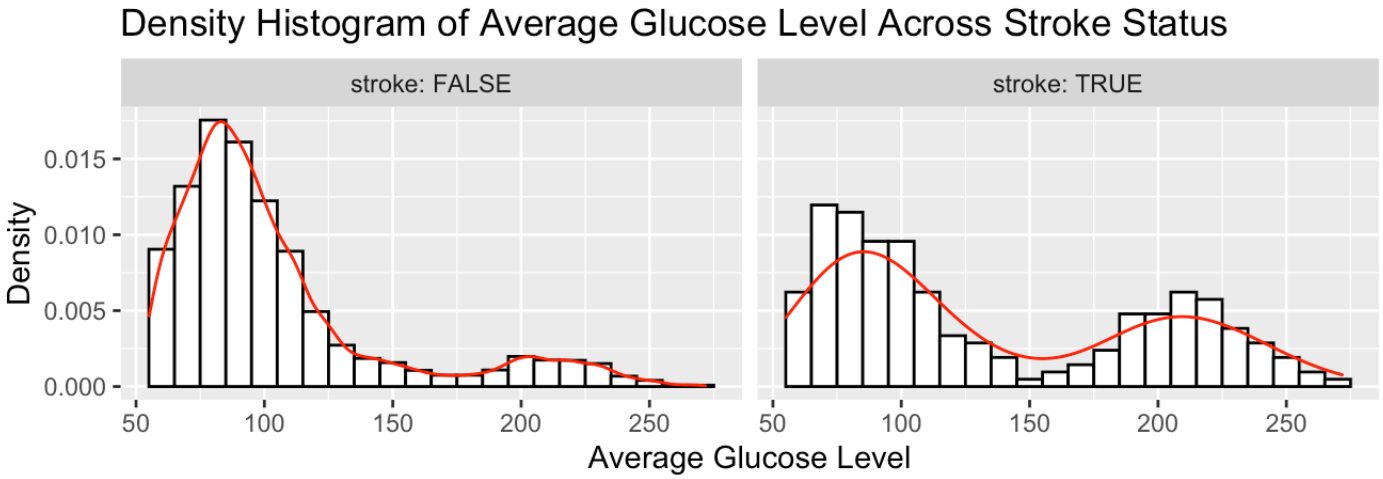
```
# Changing variables to sensible data types
stroke_df <- stroke_df %>% mutate(age = as.integer(age)) %>%
  mutate_at(c('gender', 'work_type', 'smoking_status'), as.factor) %>%
  mutate_at(c('hypertension', 'heart_disease', 'stroke'), as.logical)

glimpse(stroke_df)
```

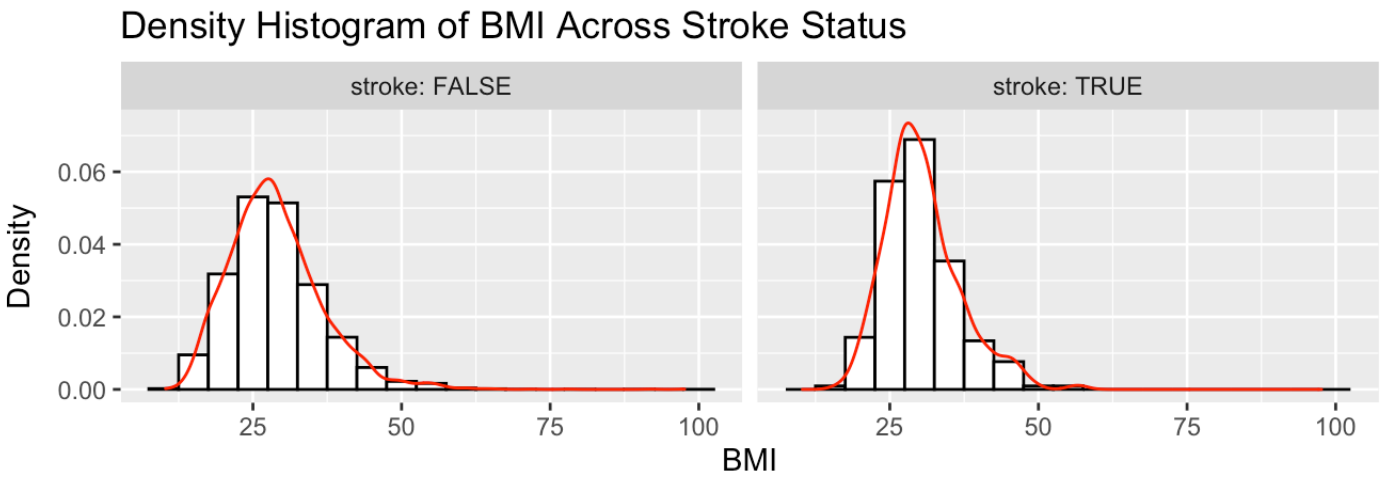
```
## Rows: 5,109
## Columns: 9
## $ age               <int> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, …
## $ gender            <fct> Male, Female, Male, Female, Female, Male, Male, Fema…
## $ hypertension      <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, TRUE, FALSE…
## $ heart_disease     <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, FALSE,…
## $ work_type         <fct> Private, Self-employed, Private, Private, Self-emplo…
## $ bmi               <dbl> 36.6, NA, 32.5, 34.4, 24.0, 29.0, 27.4, 22.8, NA, 24…
## $ smoking_status    <fct> formerly smoked, never smoked, never smoked, smokes,…
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0…
## $ stroke            <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
```

3. Using ggplot and the grid drawing method of your choice, for each stroke status, draw a density histogram with the density curve overlaid on the same plot, and wrap the plots in a 1-by-2 grid. Do this for **avg_glucose_level** and **bmi**. It is recommended to use the library “patchwork” or “gridExtra” for making the 2-plot grid. Label your axes appropriately, clearly show the legends, and include a plot title for full credit. You should also choose a suitable histogram bin width for each variable and explain why you made that choice.

Before plotting, use the following code to delete the missing terms in the column “BMI”.

```
# Removing NA values
stroke_df1 <- na.omit(stroke_df)

# Filtering for stroke and non-stroke subjects
stroke_df_stroke <- stroke_df1 %>% filter(stroke==TRUE)
stroke_df_nostroke <- stroke_df1 %>% filter(stroke==FALSE)

# Plotting density histograms
plot1 <- ggplot(stroke_df1, aes(x=avg_glucose_level)) +
  geom_histogram(aes(y=..density..), colour=1, fill="white", binwidth = 10) +
  geom_density(col = "red") +
  labs(x="Average Glucose Level", y="Density",
       title="Density Histogram of Average Glucose Level Across Stroke Status")

plot2 <- ggplot(stroke_df1, aes(x=bmi)) +
  geom_histogram(aes(y=..density..), colour=1, fill="white", binwidth = 5) +
  geom_density(col = "red") +
  labs(x="BMI", y="Density",
       title="Density Histogram of BMI Across Stroke Status")

plot3 <- plot1 + facet_wrap(~stroke, labeller = label_both)
plot4 <- plot2 + facet_wrap(~stroke, labeller = label_both)
plot3 + plot4 + plot_layout(ncol=1)
```





Bin widths of 10 and 5 were chosen to represent the glucose and BMI data. Increasing the width much more than these parameters did not fully capture the distributions, and lowering it much less than them created too many bins where a few bins were missing any observations, thus taking away from the smoothness of the distribution.

4. What are your observations? E.g. Does any feature visually show a strong difference in means or in shape of the data between the stroke groups? For each feature, which stroke group seems to have a higher average value? Clearly explain your answer to receive full credit.

For average glucose level, there appears to be a higher mean for stroke patients than for non-stroke patients. The data follow a bimodal distribution for the stroke patients and appear to be right-skewed for the non-stroke patients. The bimodal distribution in the stroke group seems to have two peaks at roughly 75 and 225 with a high percentage of individuals. The right-skewed distribution in the non-stroke group seems to have a peak around 85 with a large percentage of individuals.

For BMI, the mean for stroke patients appears to be roughly similar to the mean for non-stroke patients (perhaps slightly higher). Both stroke groups seem to have slightly right-skewed data for BMI, both with peaks around 30.

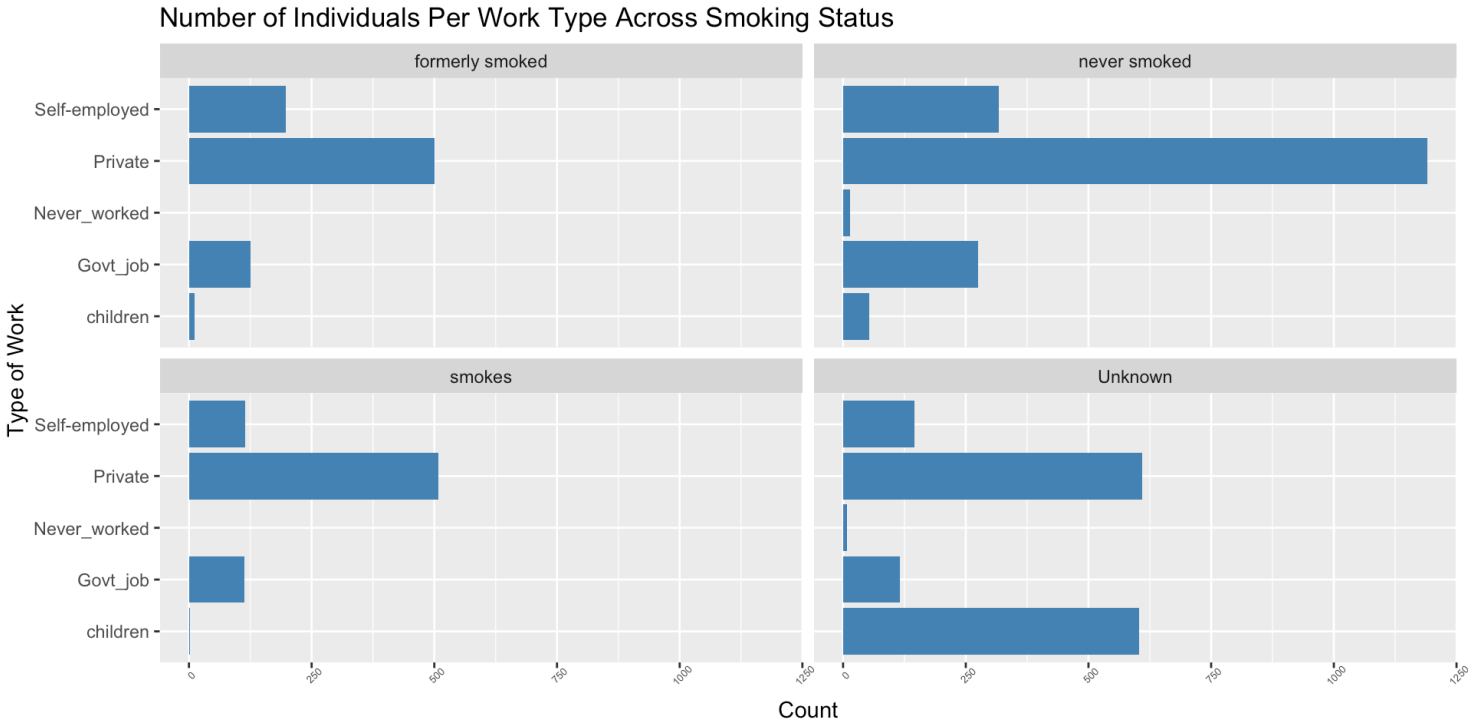
Exploratory Data Analysis and Data Preprocessing (58 points)

5. Use ggplot and facet_wrap() to plot a bar plot for the following categorical variables in this dataset:

work_type smoking_status

- When using facet_wrap(), you are encouraged to make 2 plots in one row.
- make sure to rotate the axis labels so that they are readable.

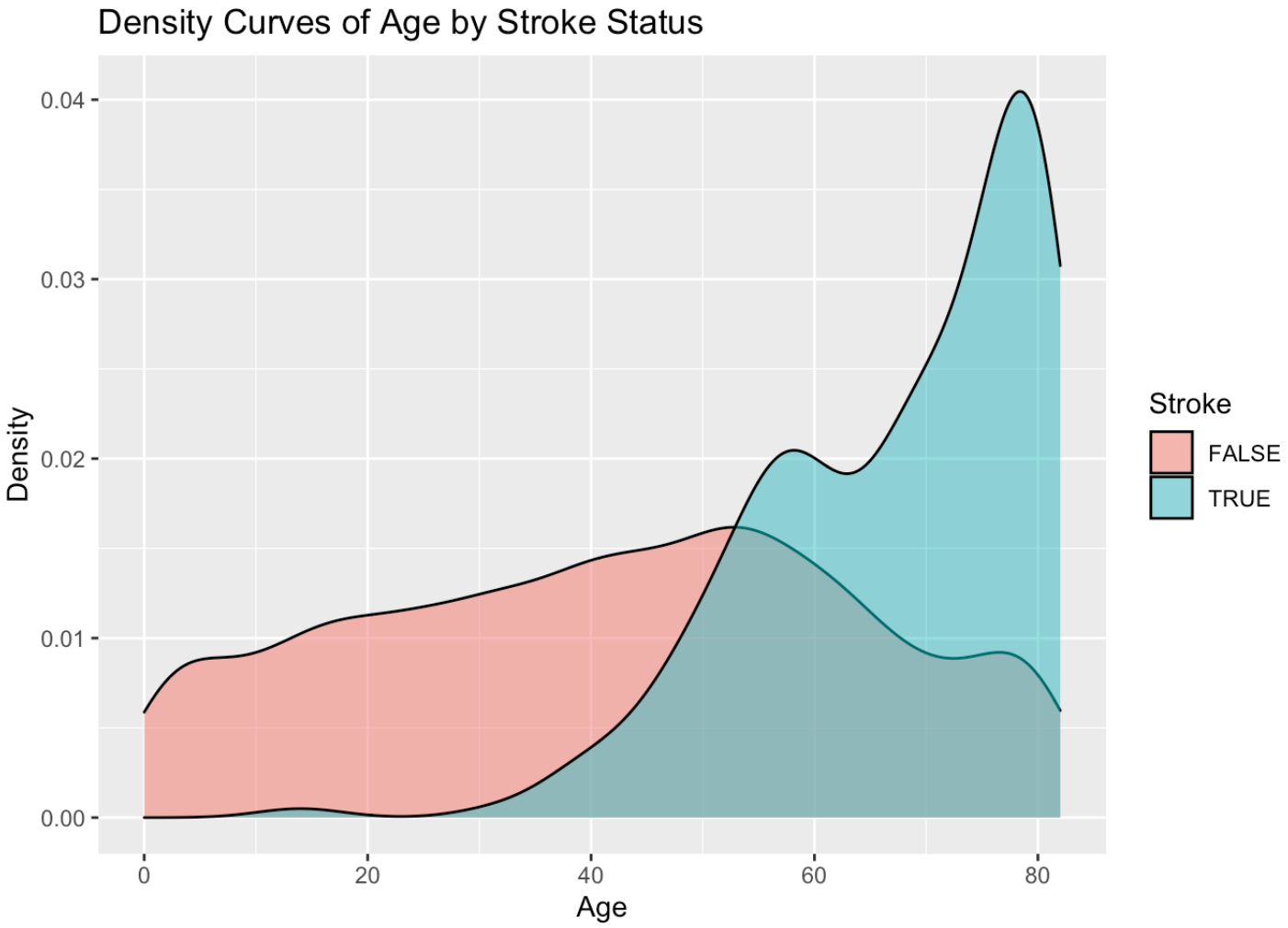
```
# Plotting bar plots
ggplot(stroke_df1, aes(x=work_type)) +
  geom_bar(fill="steelblue") +
  labs(x="Type of Work", y="Count",
       title="Number of Individuals Per Work Type Across Smoking Status") +
  facet_wrap(~smoking_status) +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 45, size = 5))
```



6. Using ggplot, draw density function curves of *age* by stroke status on the same plot. What do you observe? Which stroke group has a higher mean age?

- remember to make a title, clearly label the axes, and create a legend.

```
# Plotting density function curves of age by stroke status
ggplot(stroke_df1, aes(x=age, fill=stroke)) +
  geom_density(alpha=0.5) +
  labs(x="Age", y="Density", title="Density Curves of Age by Stroke Status",
       fill="Stroke")
```



The stroke patients have a higher mean age than the non-stroke patients.

7. Answer the following questions:

- What is the average age of the subjects in this dataset who have had a stroke?

```
mean(stroke_df_stroke$age)
```

```
## [1] 67.71292
```

The mean age of subjects who have had a stroke is ~67.7 years.

- What is the average age of the subjects in this dataset who have not had a stroke?

```
mean(stroke_df_nostroke$age)
```

```
## [1] 41.75207
```

The mean age of subjects who have not had a stroke is ~41.8 years.

- What is the average BMI (body mass index) of subjects in this dataset who had heart_disease?

```
# Filtering for heart disease patients
heart <- stroke_df1 %>% filter(heart_disease==TRUE)

mean(heart$bmi)
```

```
## [1] 30.31646
```

The mean BMI of subjects who have had heart disease is ~30.3.

- What is the average glucose level of subjects in this dataset who had heart_disease?

```
mean(heart$avg_glucose_level)
```

```
## [1] 135.3829
```

The mean average glucose level of subjects who have had heart disease is ~135.4.

- Among the population who never smoked, what is the occurrence of stroke?

```
n_smoke <- stroke_df1 %>% filter(smoking_status=="never smoked") %>% count(stroke)
print(n_smoke)
```

```
## # A tibble: 2 × 2
##   stroke     n
##   <lgl>  <int>
## 1 FALSE   1768
## 2 TRUE      84
```

There are 84 occurrences of strokes among the population who never smoked.

- Among the population who did smoke, what is the occurrence of stroke?

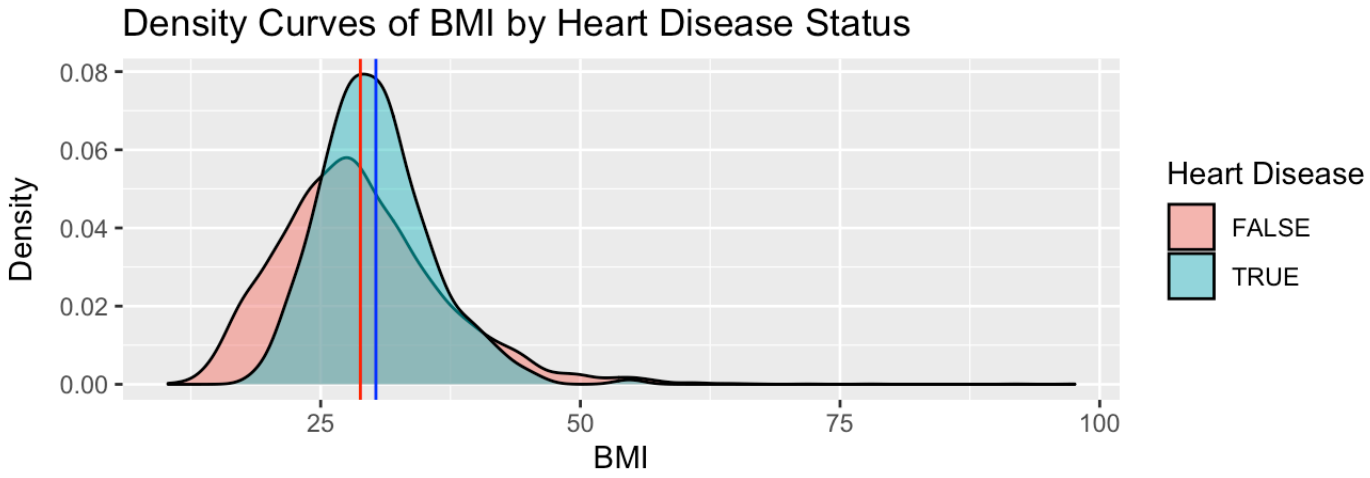
```
smoke <- stroke_df1 %>% filter(smoking_status=="formerly smoked") %>% count(stroke)
print(smoke)
```

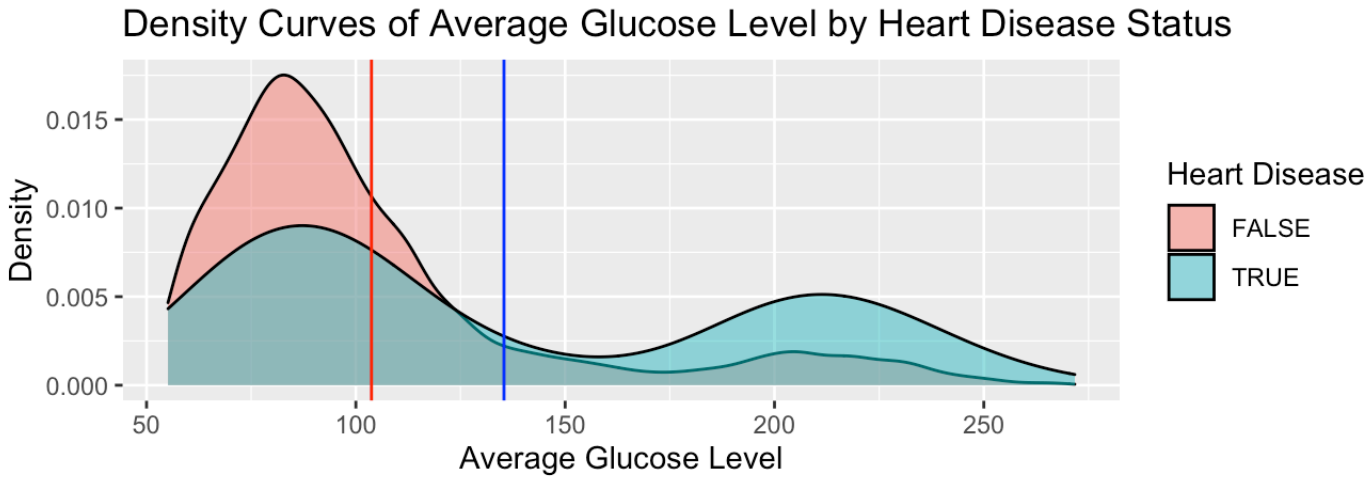
```
## # A tibble: 2 × 2
##   stroke     n
##   <lgl>  <int>
## 1 FALSE    779
## 2 TRUE      57
```

There are 57 occurrences of strokes among the population who formerly smoked.

8. Now, using a similar 1 by 2 grid, overlay two separate density curves for the two groups based on their heart disease status, for each variable *BMI* and *average glucose level*. For these figures, plot just the density curve (without the density histogram bins). Add a vertical line for each of the two density curves on each plot at the mean value for that feature for each group (grouped by heart disease status). For example, in the first entry of the grid, which is row 1 and column 1 of the 1-by-2 grid, there should be two density plots drawn in two colors based on whether subjects represented on the curve had heart disease or not.

```
# Getting mean BMI values by heart disease status
mean_bmi_noheart <- stroke_df1 %>%
  filter(heart_disease==FALSE) %>%
  pull(bmi) %>%
  mean()
mean_bmi_heart <- stroke_df1 %>%
  filter(heart_disease==TRUE) %>%
  pull(bmi) %>%
  mean()

# Plotting density curves for BMI by heart disease status
plot1 <- ggplot(stroke_df1, aes(x=bmi, fill=heart_disease)) +
  geom_density(alpha=0.5) +
  labs(x="BMI", y="Density", title="Density Curves of BMI by Heart Disease Status",
       fill="Heart Disease") +
  geom_vline(xintercept = c(mean_bmi_heart, mean_bmi_noheart), color = c("blue", "red
"))

# Getting mean glucose values by heart disease status
mean_glucose_noheart <- stroke_df1 %>%
  filter(heart_disease==FALSE) %>%
  pull(avg_glucose_level) %>%
  mean()
mean_glucose_heart <- stroke_df1 %>%
  filter(heart_disease==TRUE) %>%
  pull(avg_glucose_level) %>%
  mean()

# Plotting density curves for glucose by heart disease status
plot2 <- ggplot(stroke_df1, aes(x=avg_glucose_level, fill=heart_disease)) +
  geom_density(alpha=0.5) +
  labs(x="Average Glucose Level", y="Density",
       title="Density Curves of Average Glucose Level by Heart Disease Status",
       fill="Heart Disease") +
  geom_vline(xintercept = c(mean_glucose_heart, mean_glucose_noheart),
             color = c("blue", "red"))

plot1 + plot2 + plot_layout(ncol = 1)
```





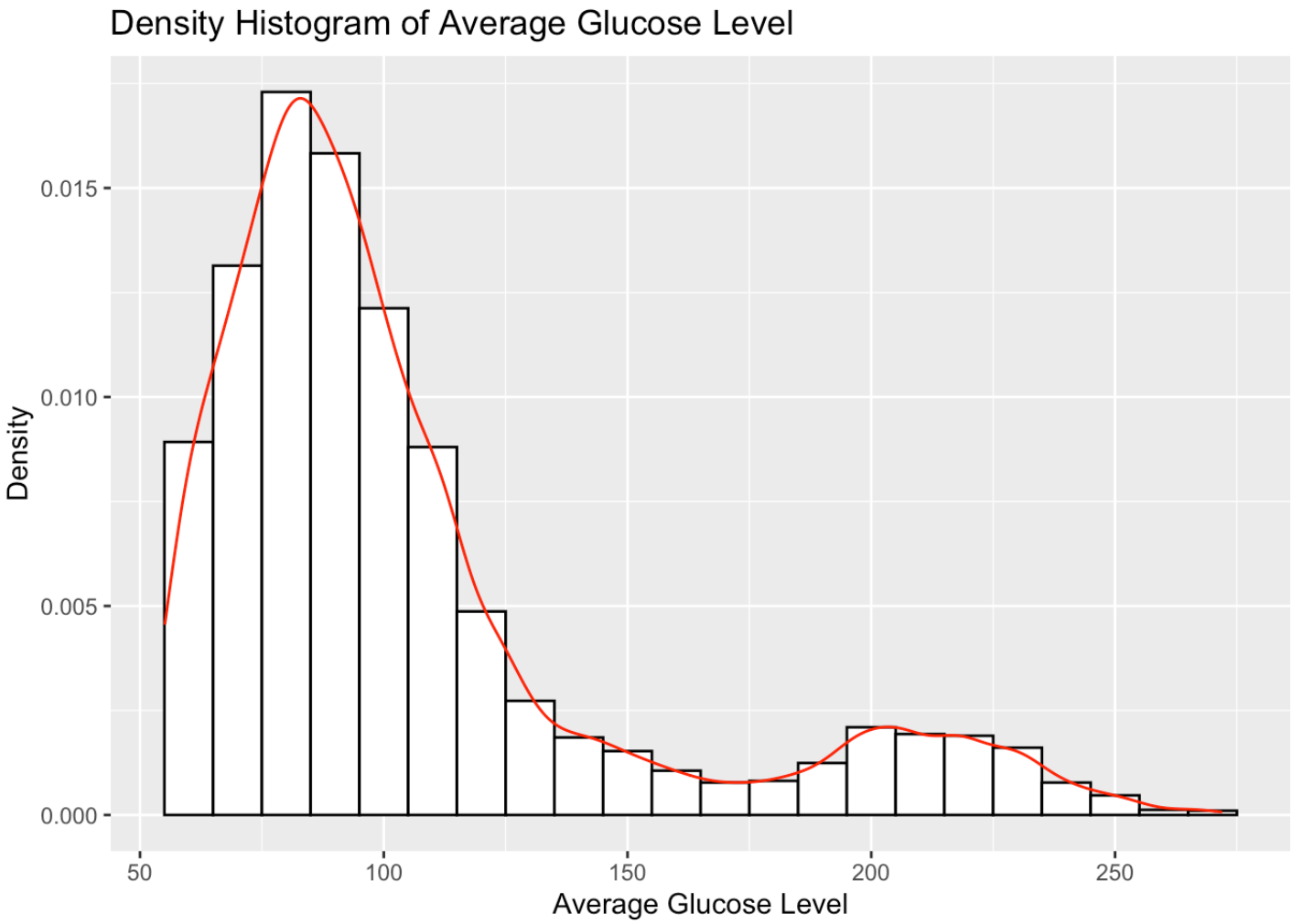
9. What are your observations from the previous plots? E.g. Do any of the features visually show a strong difference in means or in shape of the data distribution between the people who had heart disease and those who didn't? For each feature, which heart disease diagnosis group seems to have the higher average value? Clearly explain your answer to receive full credit.

For BMI, both heart disease groups seem to display a similar right-skewed distribution. It appears that both groups have very similar mean values of BMI, with the heart disease group having a slightly higher value overall.

For average glucose level, the group with heart disease appears to follow a bimodal distribution, with peaks around values of 80 and 215. The group without heart disease appears to follow a right-skewed distribution, with a peak around 80. The mean average glucose level of the group with heart disease is higher than the group without heart disease.

10. Plot the histogram and density plots of the variable "avg_glucose_level", and answer the following questions:

```
# Plotting the histogram of average glucose level
ggplot(stroke_df1, aes(x=avg_glucose_level)) +
  geom_histogram(aes(y=..density..), colour=1, fill="white", , binwidth = 10) +
  geom_density(col = "red") +
  labs(x="Average Glucose Level", y="Density",
       title="Density Histogram of Average Glucose Level")
```



a. Is the data skewed?

The data is skewed.

b. Does it have positive skewness or negative skewness?

The data has a positive skewness.

c. Compute the skewness using the definition from the lecture.

```
# Computing the skewness
skewness(stroke_df1$avg_glucose_level)
```

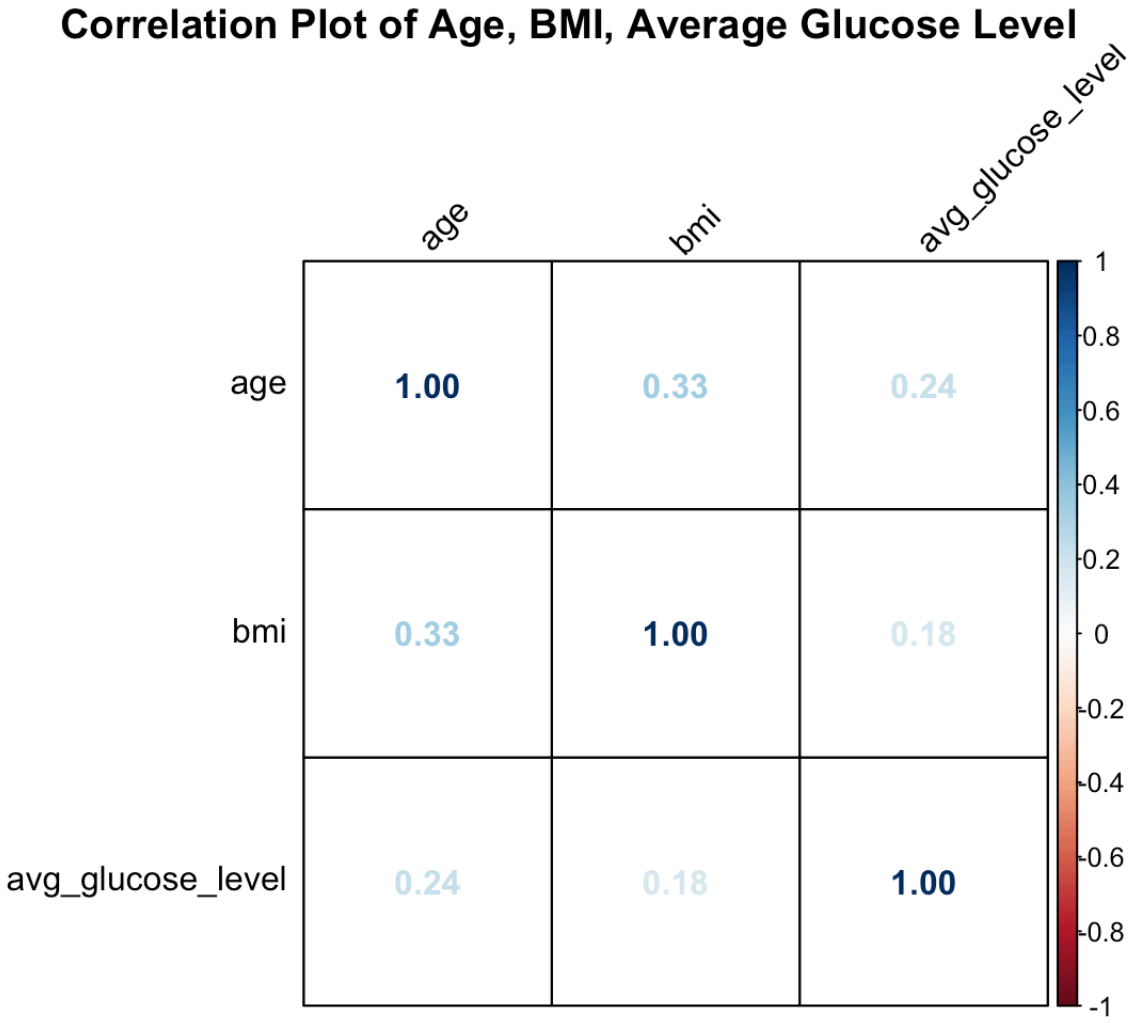
```
## [1] 1.614125
```

d. According to the criterion introduced in the lecture, is the dataset moderately skewed or highly skewed?

The data is highly skewed according to the criteria in the lecture, because the absolute value of the calculated value for skewness is above 1.

11. Correlation plots are a way to visualize multivariate relationships (between numerical variables). Using the corrplot package, make a correlation plot of the variables age, BMI and avg_glucose_level. Clearly label your axis and include a legend to receive full credit. (you should also add a title of the overall plot and coefficient values in each cell).

```
# Calculating the correlation matrix
stroke_corr <- stroke_df1 %>% select(age, bmi, avg_glucose_level)
corr <- cor(stroke_corr)

# Plotting the correlation matrix
corrplot(corr, method='number', tl.col="black", tl.srt=45, addgrid.col="black",
         cl.pos="r", title="Correlation Plot of Age, BMI, Average Glucose Level",
         mar=c(0,0,1,0))
```

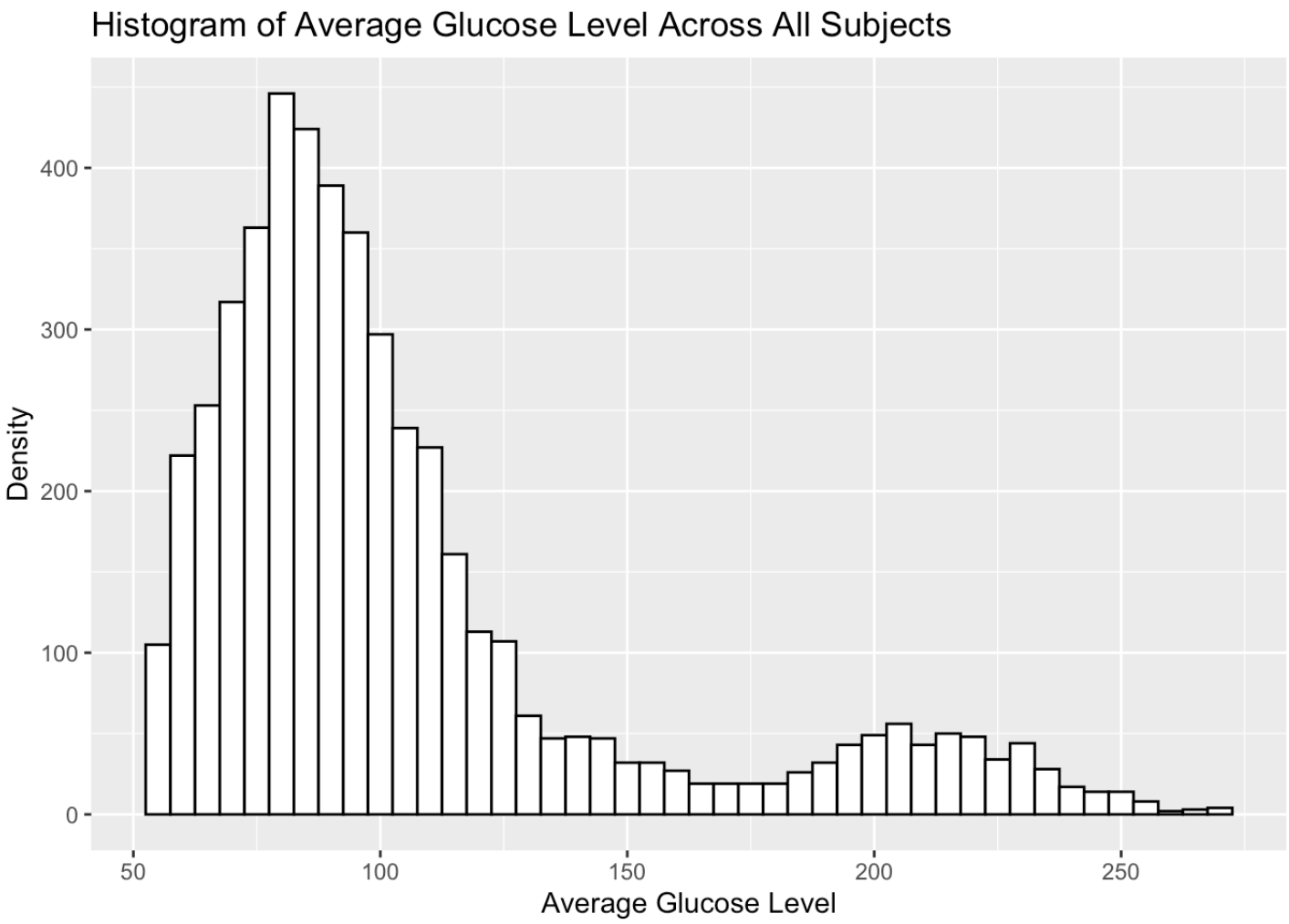


12. Calculate the z-scores of the *avg_glucose_level* and determine if there are any outliers using the definition of z-score > 3. Plot a histogram of this variable (i.e., plot histogram of avg_glucose_level). Remove the outliers (as defined by z-score>3) and plot the histogram again. What difference do you notice between the two histograms?

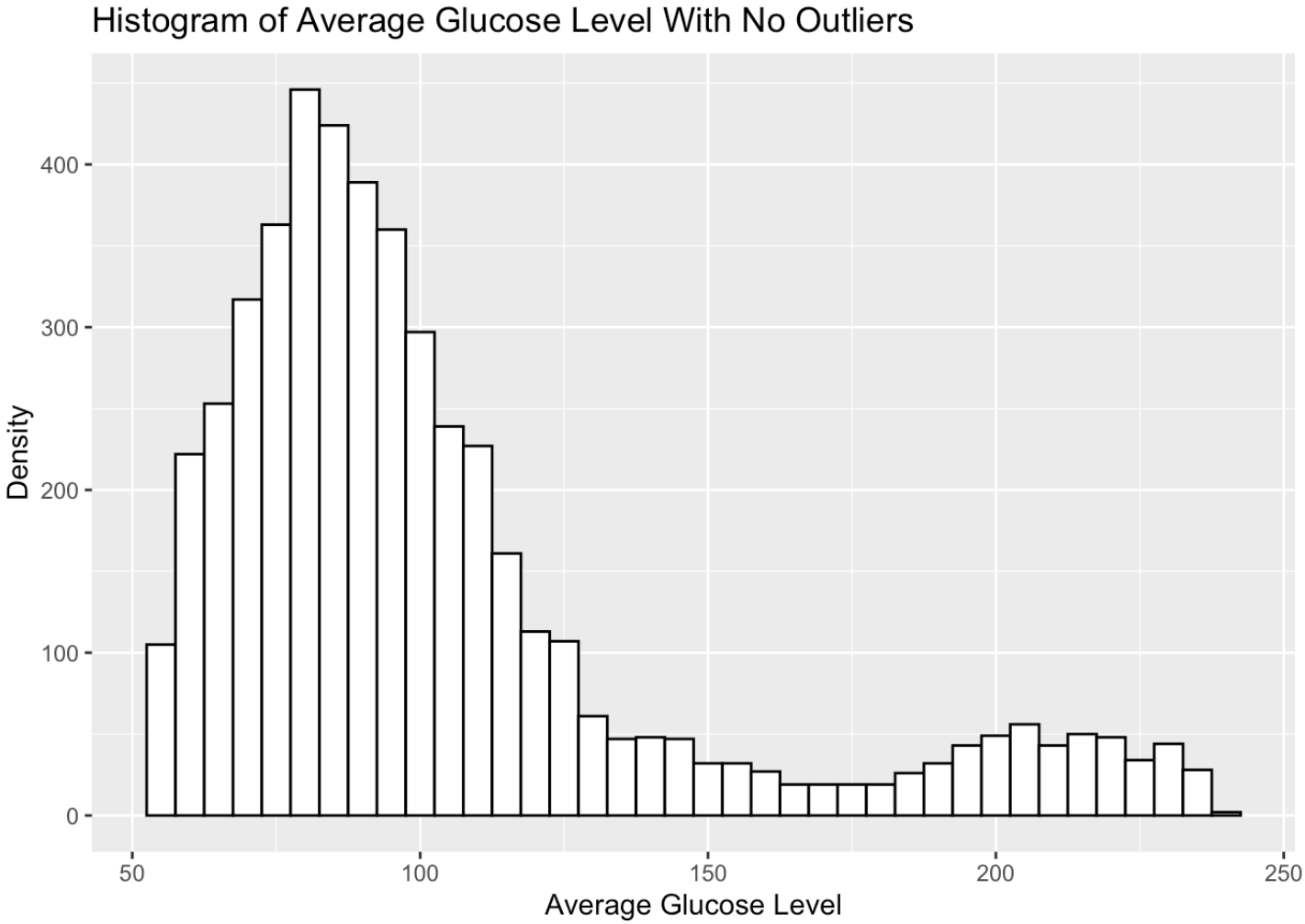
```
# Creating z-score variable for average glucose level
stroke_z <- stroke_df1 %>%
  mutate(z_scores=(avg_glucose_level - mean(avg_glucose_level)) / sd(avg_glucose_leve
l))

# Creating a dataset with only outliers and counting observations
outlier <- stroke_z %>% filter(abs(z_scores) > 3)
nrow(outlier)
```

```
## [1] 60
```

```
# Plotting histogram of glucose level across all subjects
ggplot(stroke_z, aes(x=avg_glucose_level)) +
  geom_histogram(colour=1, fill="white", binwidth = 5) +
  labs(x="Average Glucose Level", y="Density",
       title="Histogram of Average Glucose Level Across All Subjects")
```



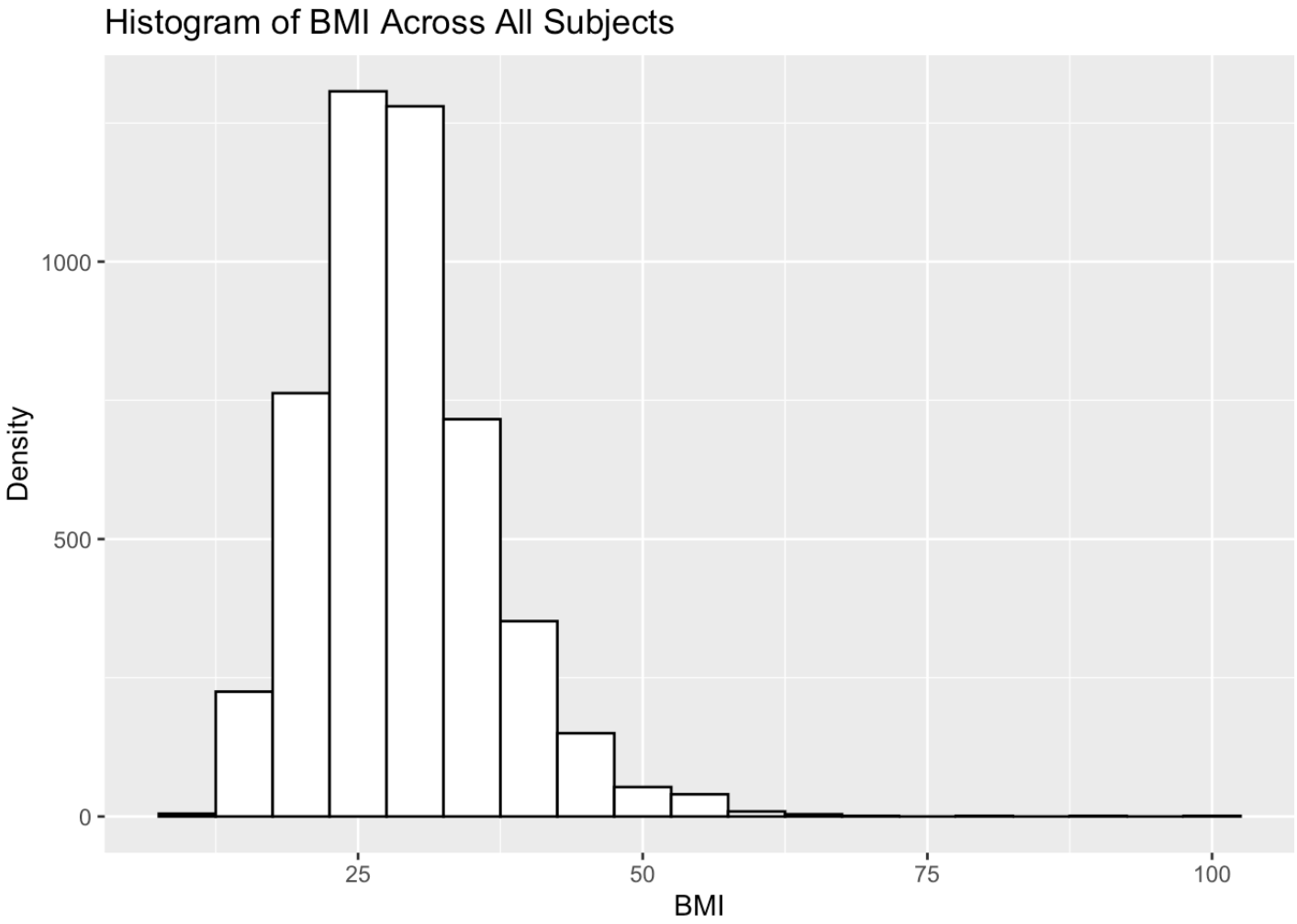

```
# Filtering for subjects who are not outliers
stroke_z <- stroke_z %>% filter(abs(z_scores) <= 3)

# Plotting histogram of glucose level across non-outliers
ggplot(stroke_z, aes(x=avg_glucose_level)) +
  geom_histogram(colour=1, fill="white", binwidth = 5) +
  labs(x="Average Glucose Level", y="Density",
       title="Histogram of Average Glucose Level With No Outliers")
```



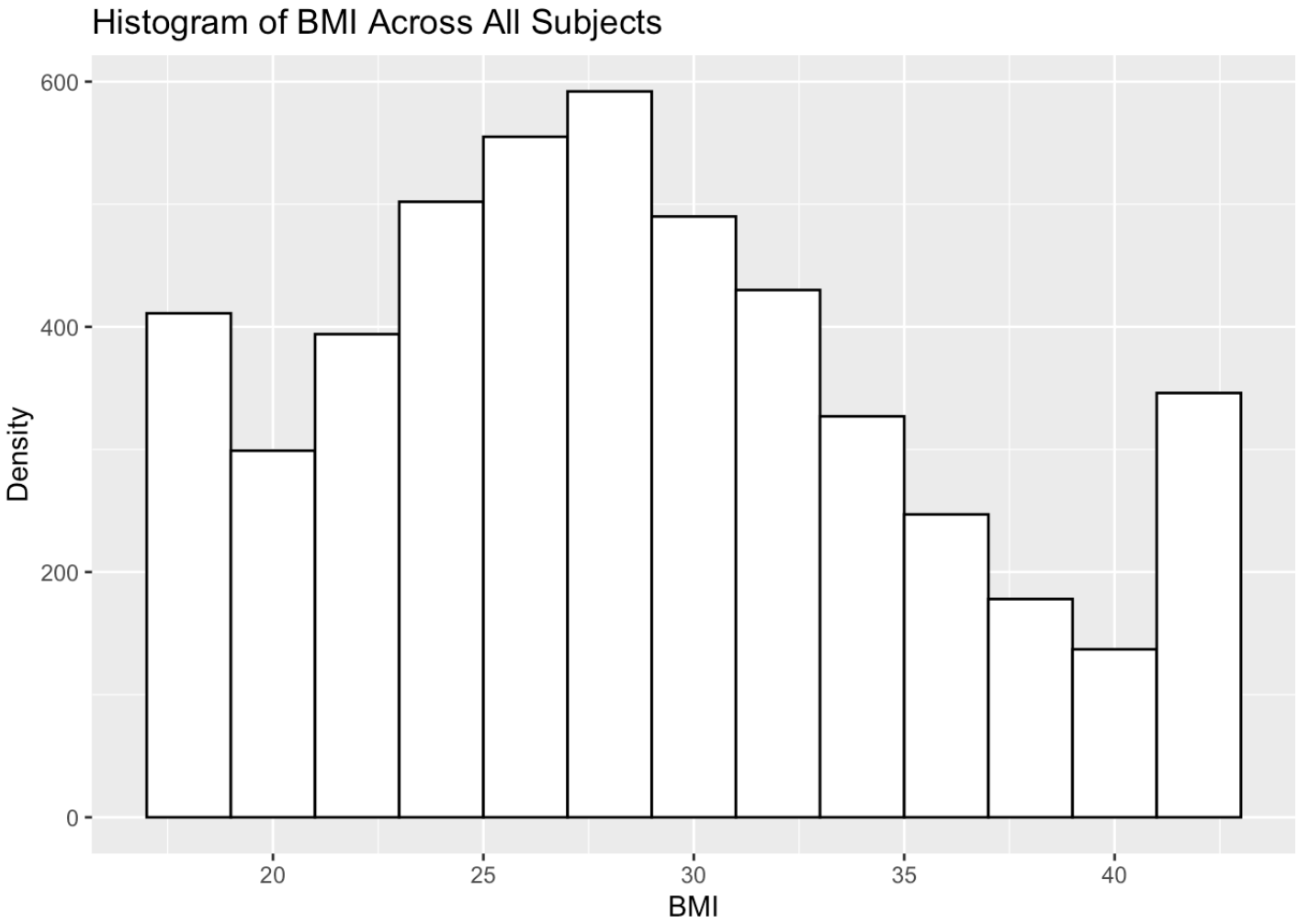
There are 60 outliers. The distribution of both histograms looks similar, as both appears to be positively skewed. However, the histogram of the average glucose levels across all subjects has a longer tail on the right side, which the outliers account for. This gives the first histogram a slightly more positive skew than the second.

13. Perform winsorization on the “BMI” variable. This is a technique that was not covered in lecture, but is another type of transformation that limits the impact of outliers in your analyses. Typically, you can decide upon a threshold (often, a specific range of percentiles) and replace all of the data points outside of the threshold with the closest value from within the threshold. An example from Wikipedia may be a helpful demonstration: “a 90% winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile.” Conduct a 90% winsorization on the “BMI” variable and plot a histogram of the winsorized data. What is your observation of the data distribution before and after winsorization?

```
# Plotting histogram of BMI across all subjects
ggplot(stroke_df1, aes(x=bmi)) +
  geom_histogram(colour=1, fill="white", binwidth = 5) +
  labs(x="BMI", y="Density",
       title="Histogram of BMI Across All Subjects")
```




```
# Creating a Winsorized variable of BMI
stroke_w <- stroke_df1 %>% mutate(bmi_w = Winsorize(bmi))

# Plotting histogram of Winsorized BMI
ggplot(stroke_w, aes(x=bmi_w)) +
  geom_histogram(colour=1, fill="white", binwidth = 2) +
  labs(x="BMI", y="Density",
       title="Histogram of BMI Across All Subjects")
```



The Winsorization has helped reduce the positive skewness of the original data and has made the data resemble more of a normal distribution. This will help mitigate the impact of outliers whenever running analyses on the BMI data. Note that the histogram after Winsorization seems to have two peaks at the lowest values and highest values on the x-axis, which we would expect with Winsorization. This is because all values below the 5th percentile were set to the 5th percentile, and all values above the 95th percentile were set to the 95th percentile, thus making the peaks.