Supplementary Material

Kevin Kosewick and Emma Hardin-Parker

2024-05-02

# 1. Overview

This document includes additional information about the data and all of our interesting EDA plots. The full code and process of creating each plot can be found in our “eda-processing-code” folder at the [GitHub Repository](https://github.com/KKosewick/KOSEWICK-HARDIN-PARKER-MADA-project).

# 2. Code & File Details and Locations

The eda.qmd file, located in the R folder and subsequently in the eda-processing-code folder contains all data exploration, processing, and transformation. The statistical-analysis.R file, found in the R folder and subsequently in the analysis-code folder holds all statistical analysis and machine learning code.

# 3. Exploratory Data Analysis

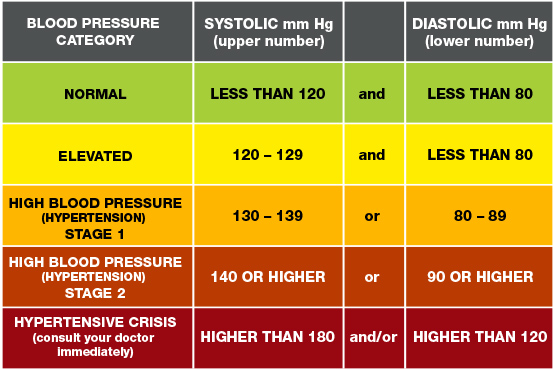
The data was obtained from kaggle.com at this [link](https://www.kaggle.com/datasets/henryshan/sleep-health-and-lifestyle/data). This page contains a detailed codebook and all available information about the data set. The following information is the codebook for this dataset:

The data set contains 373 observations for 13 variables. Some variables are not easily understandable at face value, so we went back to the data source and accessed the data dictionary. The variables are defined as follows:

Person ID: A unique identifier for each participant. Gender: The participant’s gender (Male/Female). Age: The participant’s age in years. Occupation: The job or career of the participant. Sleep Duration (hours): The daily sleep duration of the participant in hours. Quality of Sleep (scale: 1-10): A subjective assessment of sleep quality on a scale from 1 to 10. Physical Activity Level (minutes/day): The daily duration of physical activity for the participant, measured in minutes. Stress Level (scale: 1-10): A subjective assessment of the participant’s stress level on a scale from 1 to 10. BMI Category: The participant’s BMI classification (e.g., Underweight, Normal, Overweight). Blood Pressure (systolic/diastolic): The participant’s blood pressure, represented as systolic pressure over diastolic pressure. Heart Rate (bpm): The participant’s resting heart rate, measured in beats per minute. Daily Steps: The number of steps the participant takes each day. Sleep Disorder: The presence or absence of a sleep disorder in the participant (None, Insomnia, Sleep Apnea).

# 4. Data Transformation

We created a categorical hypertension variable from our current continuous blood pressure variable. We used blood pressure guidelines from the American Heart Association to do this. Our goal with this was to see how blood pressure can affect sleep quality in a more digestible way. We have attached the guidelines here: [AHA Guidelines](https://www.heart.org/en/health-topics/high-blood-pressure)



To make it easier to visualize Physical Activity, we created a new variable with categories of Low (0-30 minutes), Medium (30-60), and High (60+). We also created a new variable with groups according to the guidelines here at [10,000 steps.org](https://www.10000steps.org.au/articles/healthy-lifestyles/counting-steps/#:~:text=The%20following%20pedometer%20indices%20have%20been%20developed%20to,day%205%20Highly%20active%20is%20more%20than%2012%2C500). Nobody in the dataset goes over 10000 steps a day, however, so we won’t see any data for the “Active” or “Highly Active” groups. Lastly, we created a new variable that groups the different occupations to make analysis more succinct.

# 5. Data Visualization and Exploratory Data Analysis

We have included the plots generated by our EDA here alongside some interpretations. These are all of the preliminary findings that guided our feature engineering and variable selection steps for our final models.

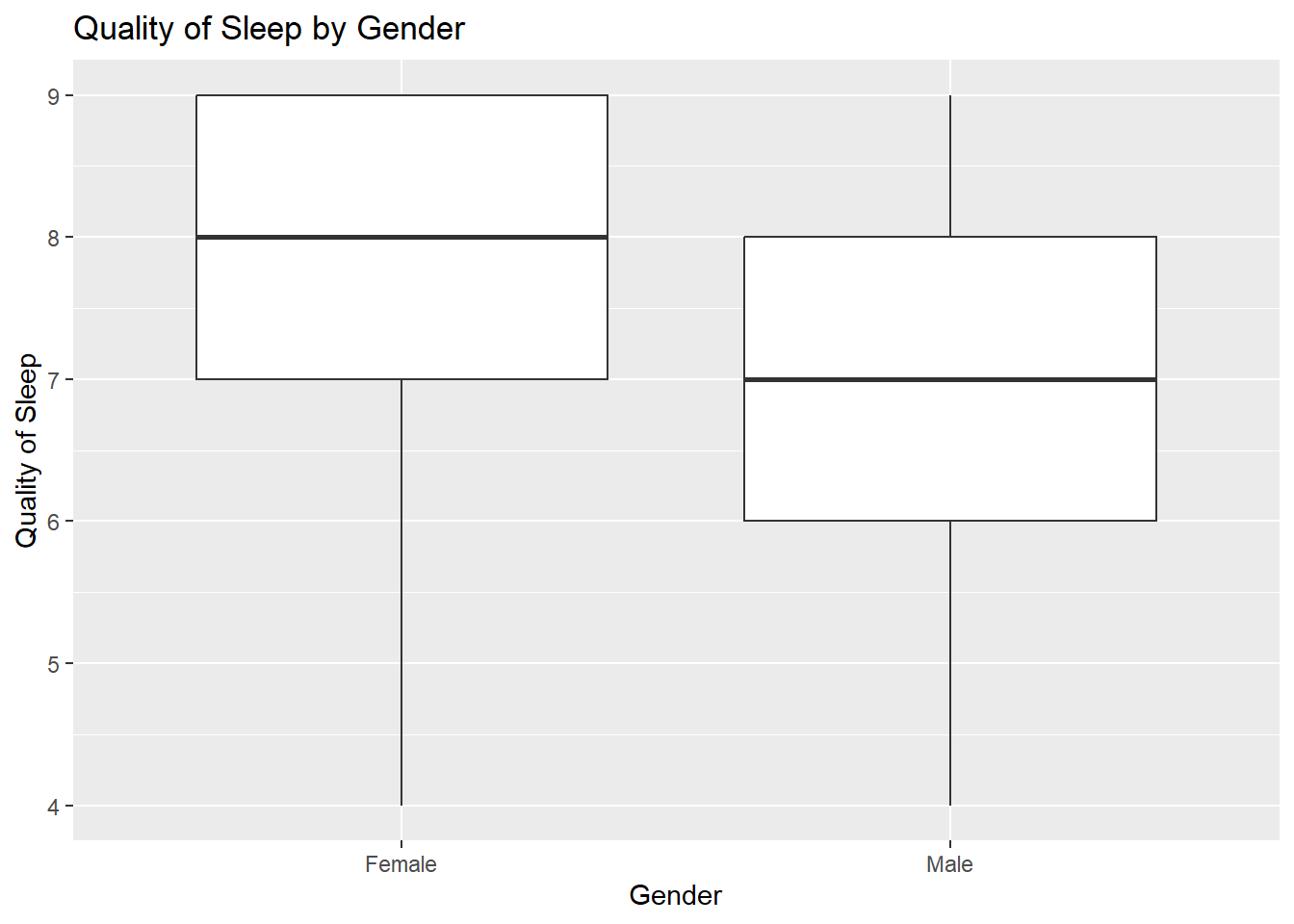


Figure 1: A boxplot comparing the average quality of sleep between males and females in our dataset.

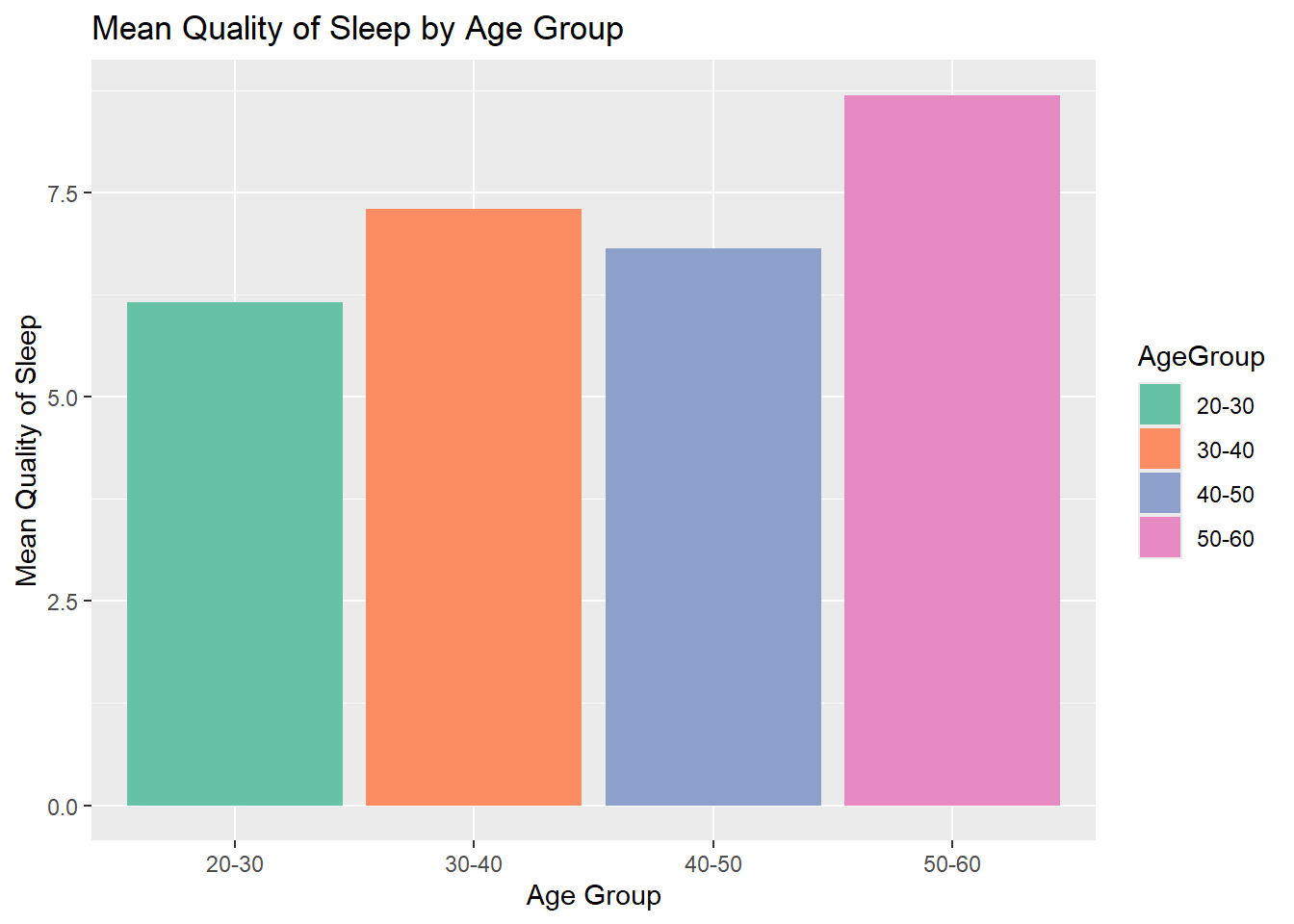


Figure 2: A bar plot showing the mean quality of sleep by age group. We eventually decided to ungroup the ages for our final model fitting processes.

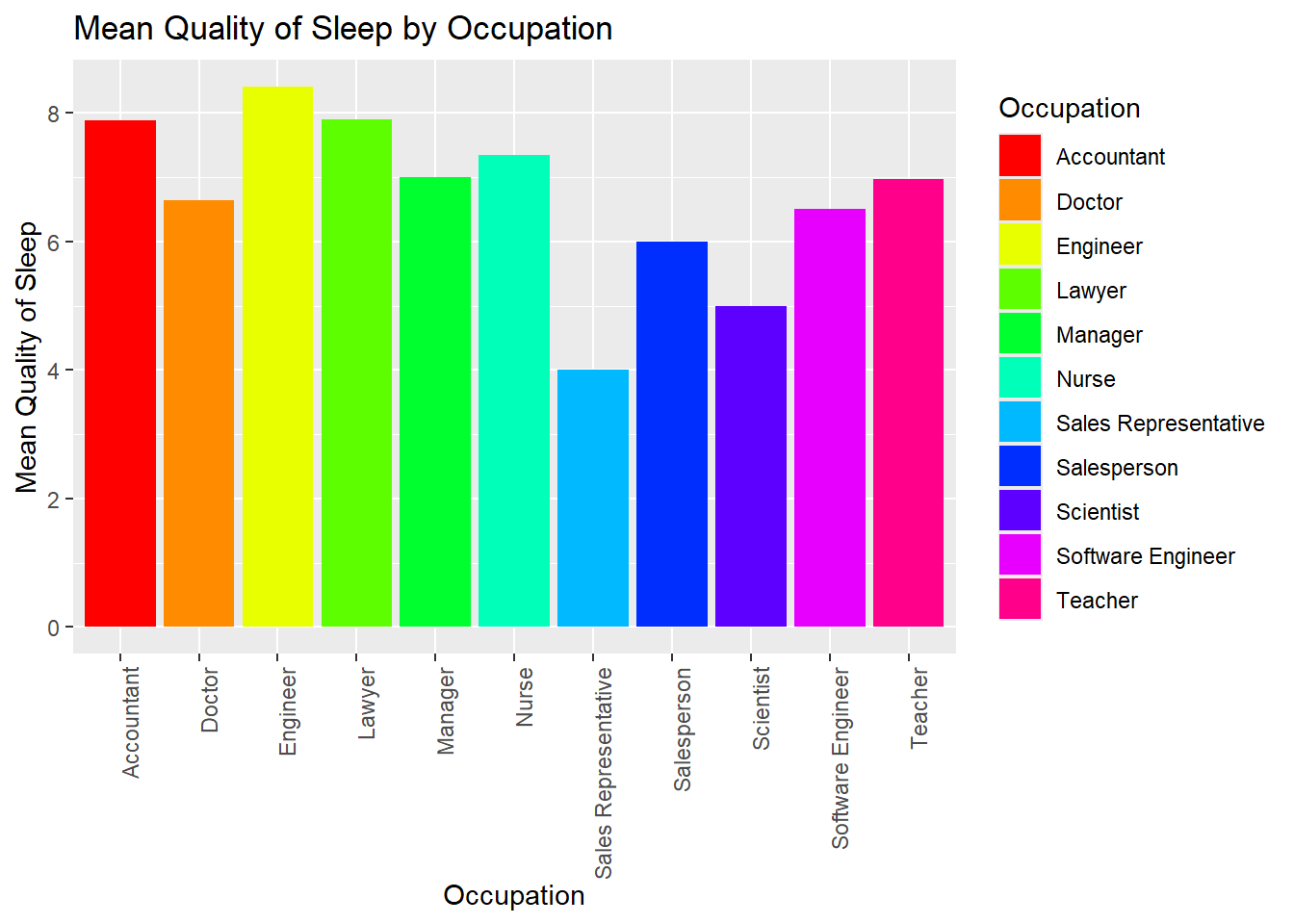


Figure 3: A bar plot showing the mean quality of sleep across the occupations in our dataset. We eventually decided to group these into broader categories for our final model fitting.

We can see some clear differences in self-reported Quality of Sleep between these different groups. Females tend to report an average of 1 higher according to the boxplot. The different age groups are similar, but the oldest (50-60) reports the highest quality of sleep by far. Our ungrouped occupations are a bit all over the place, but Sales Representatives seem to have the worst average scores by far.

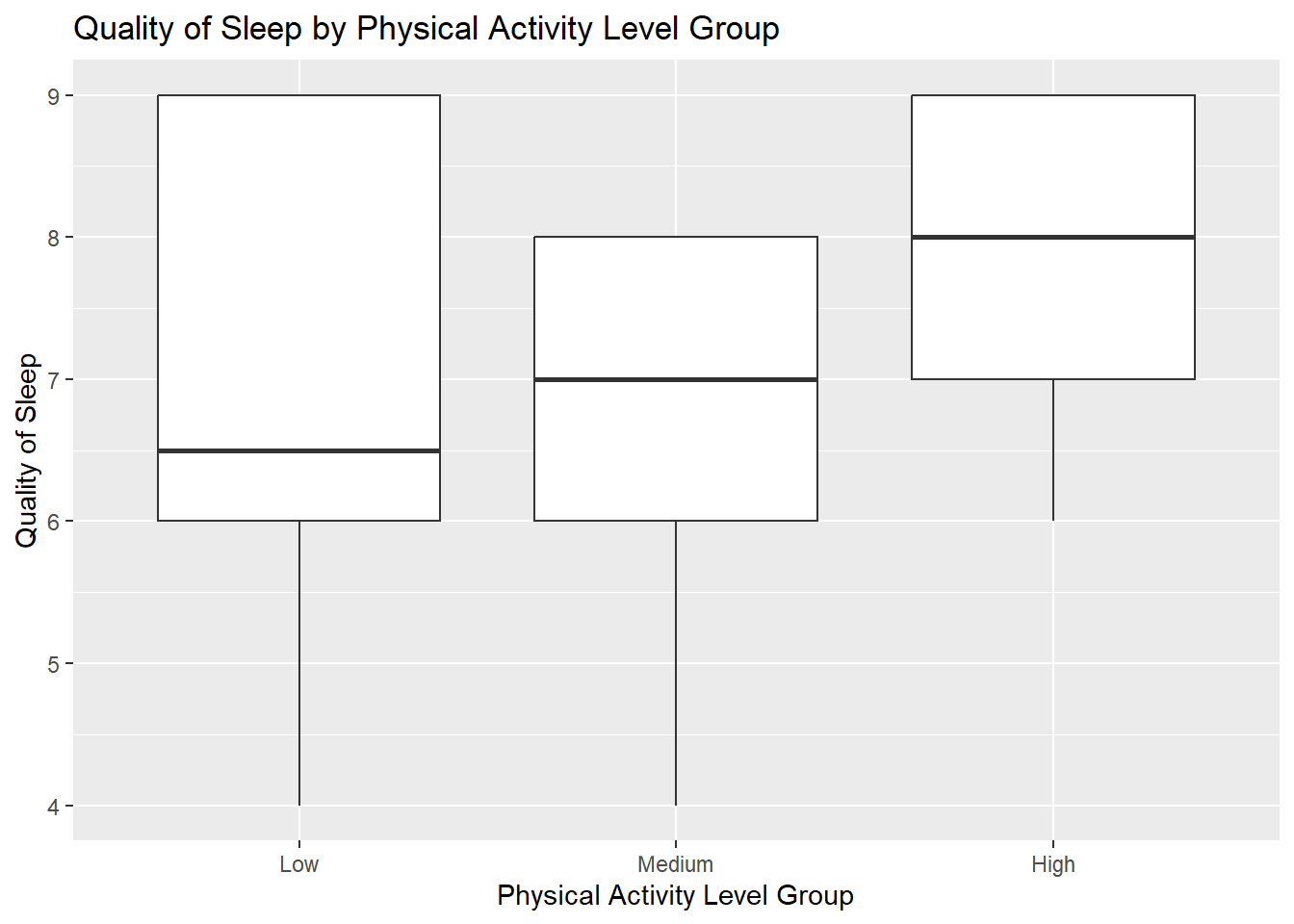


Figure 4: A boxplot of the feature engineered categorical variable for physical activity levels.

We then explored Physical Activity Level as a predictor of sleep quality using our engineered categories. We see a positive correlation between increasing physical activity and sleep quality here.

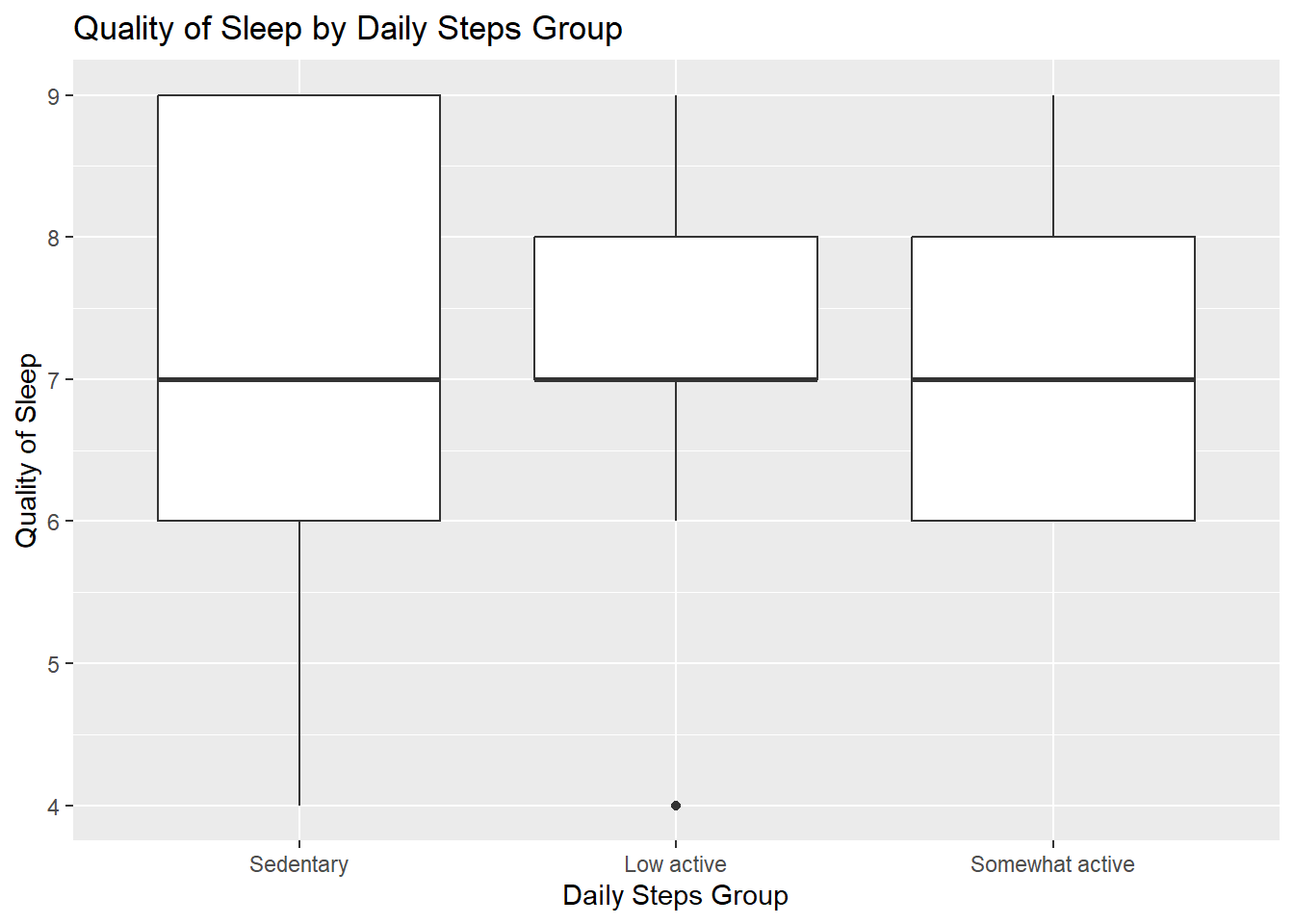


Figure 5: A boxplot of daily steps as a predictor of quality of sleep. We used the different categories as previously described.

We then explored Daily Steps with a box plot of our new engineered categorical variable. Daily Steps don’t have as clear of a trend as Physical Activity did, but there do seem to be fewer low Quality of Sleep scores in higher daily steps groups.

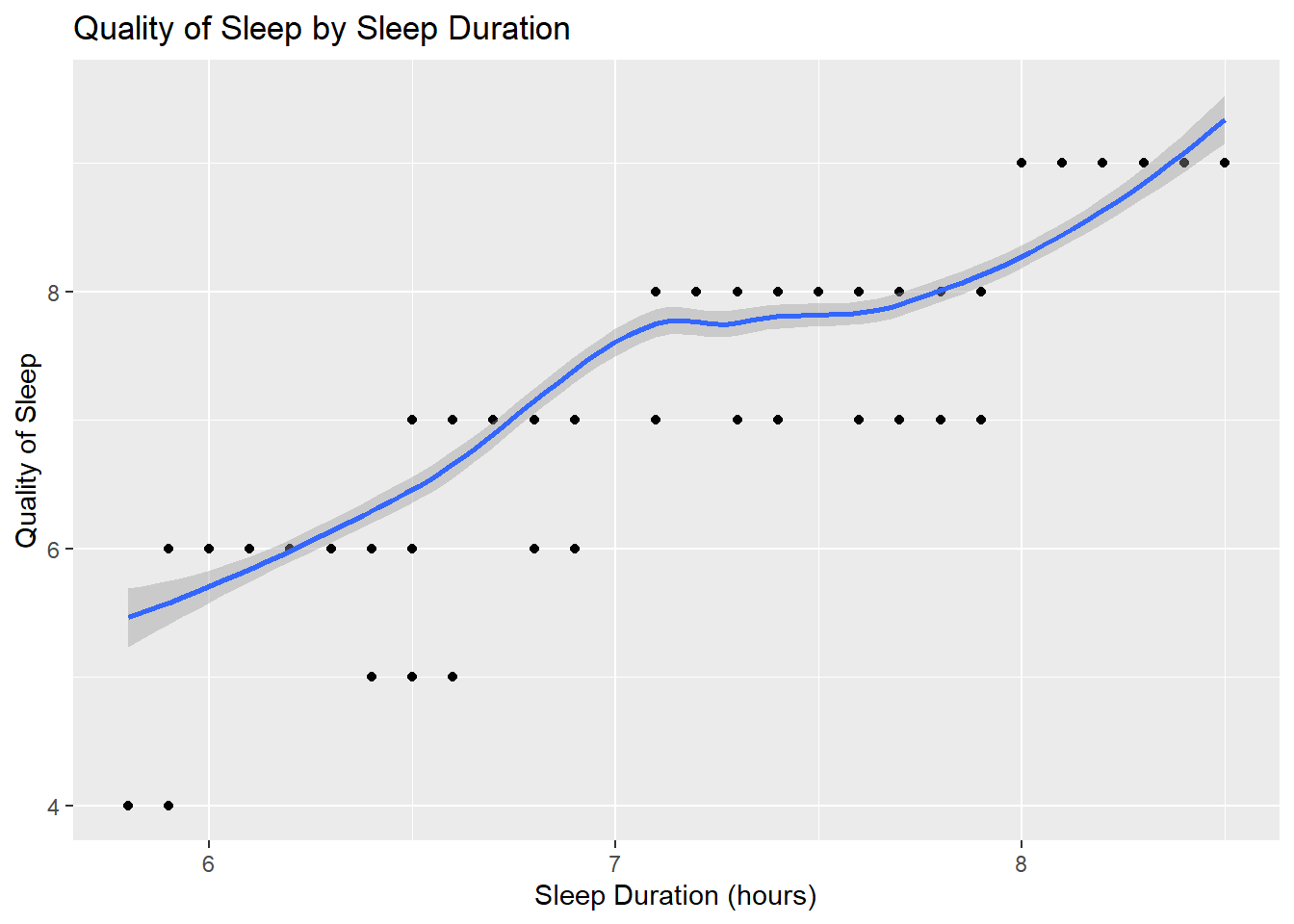


Figure 6: A scatterplot for sleep duration as a predictor of quality of sleep.

We then explored our Sleep Duration variable. According to [Medical News Today](https://www.medicalnewstoday.com/articles/sleep-cycle-stages#sleep-requirements), adults aged 18-61 are recommended 7+ hours of sleep a night. We kept this in mind when looking at our data. We can see a clear and strong positive correlation between Sleep Duration and Quality of Sleep.

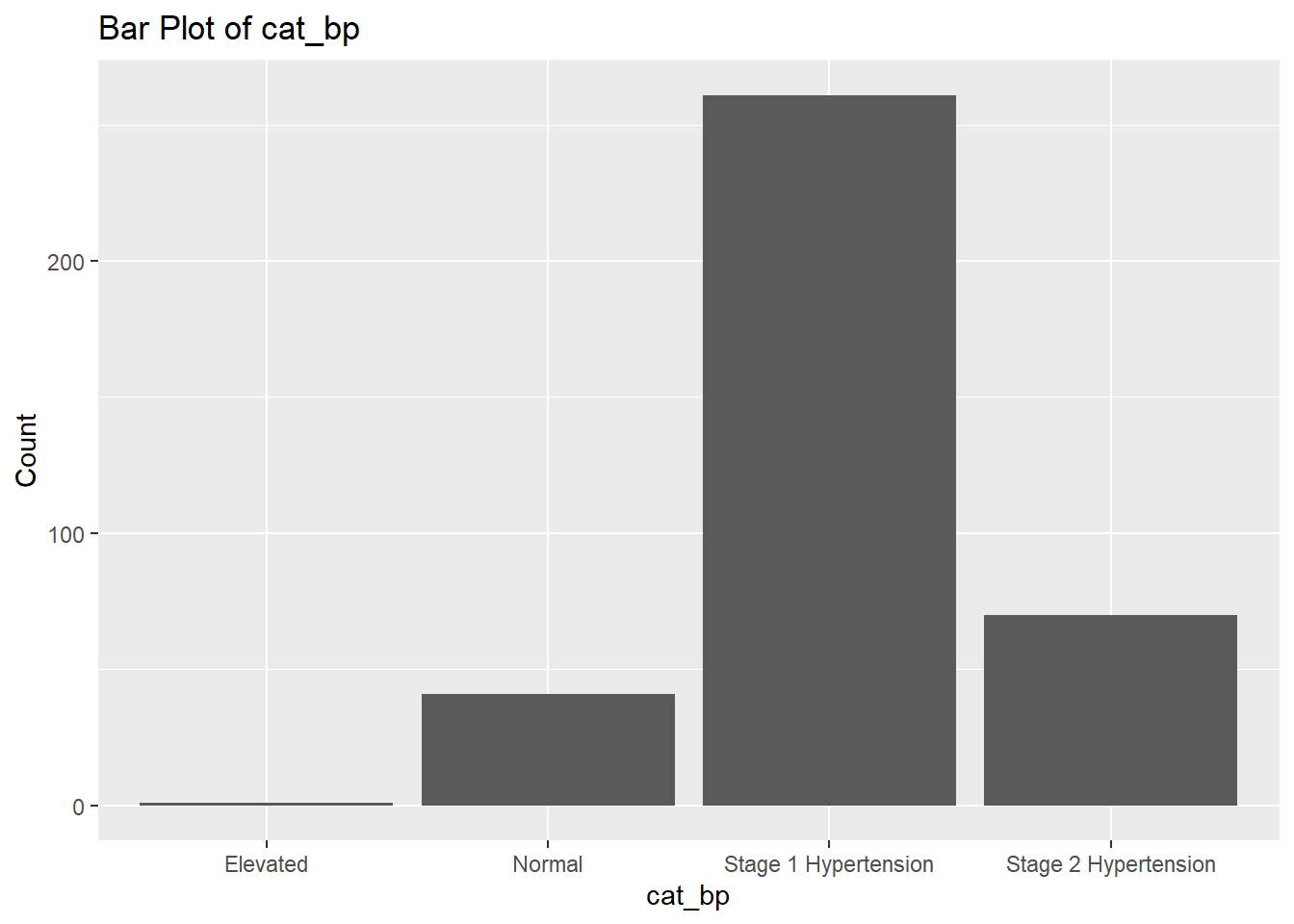


Figure 7: A bar plot of our categorical blood pressure variable.

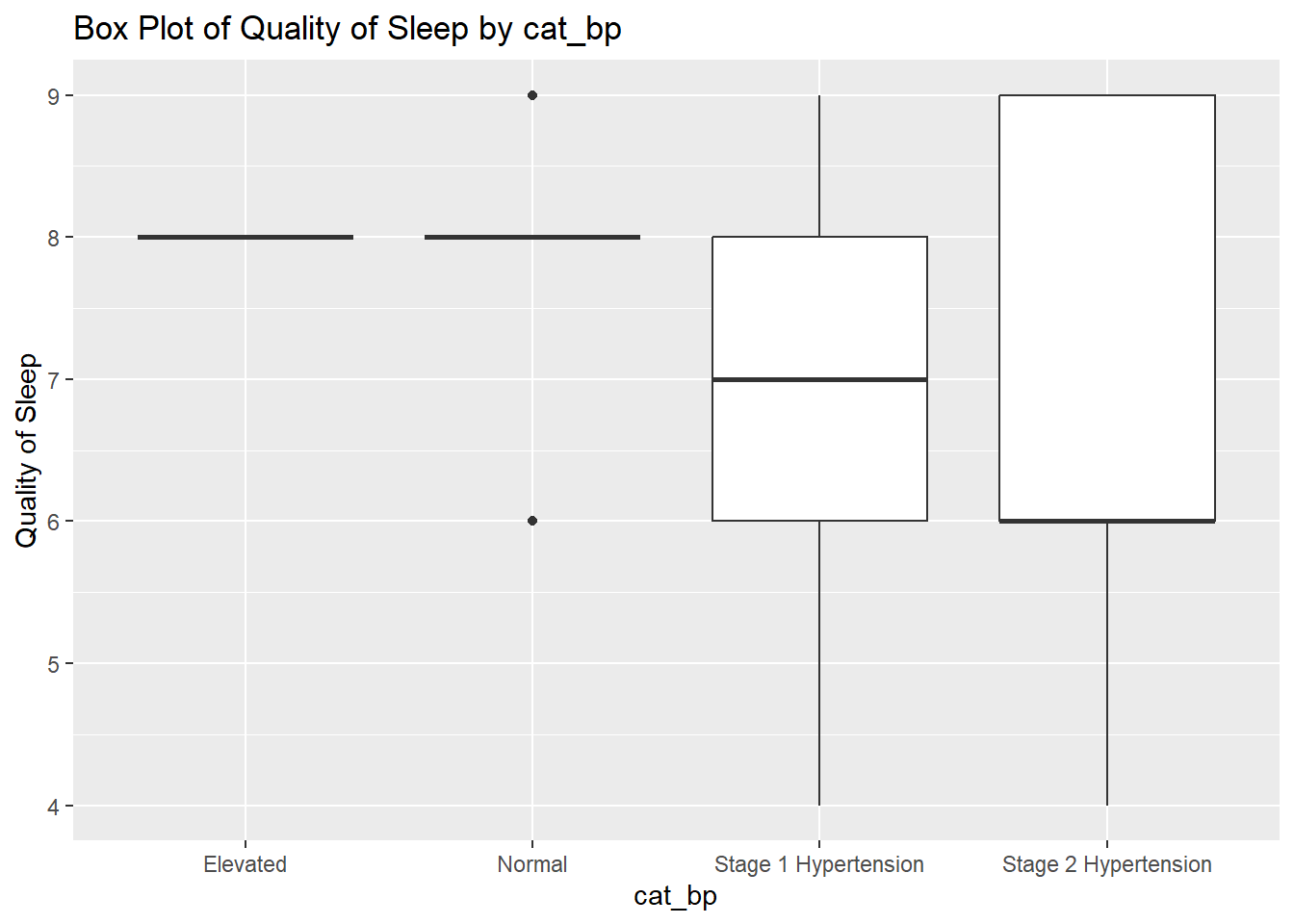


Figure 8: A box plot of hypertension category as a predictor of quality of sleep.

We then looked at the new variable we created with categories of Blood Pressure. We can see that elevated and normal blood pressure have the least observations. Most individuals fall under the category of “Stage 1 Hypertension” as defined by the AHA. We can’t see a clear relationship between our outcome and predictor based off of these plots.

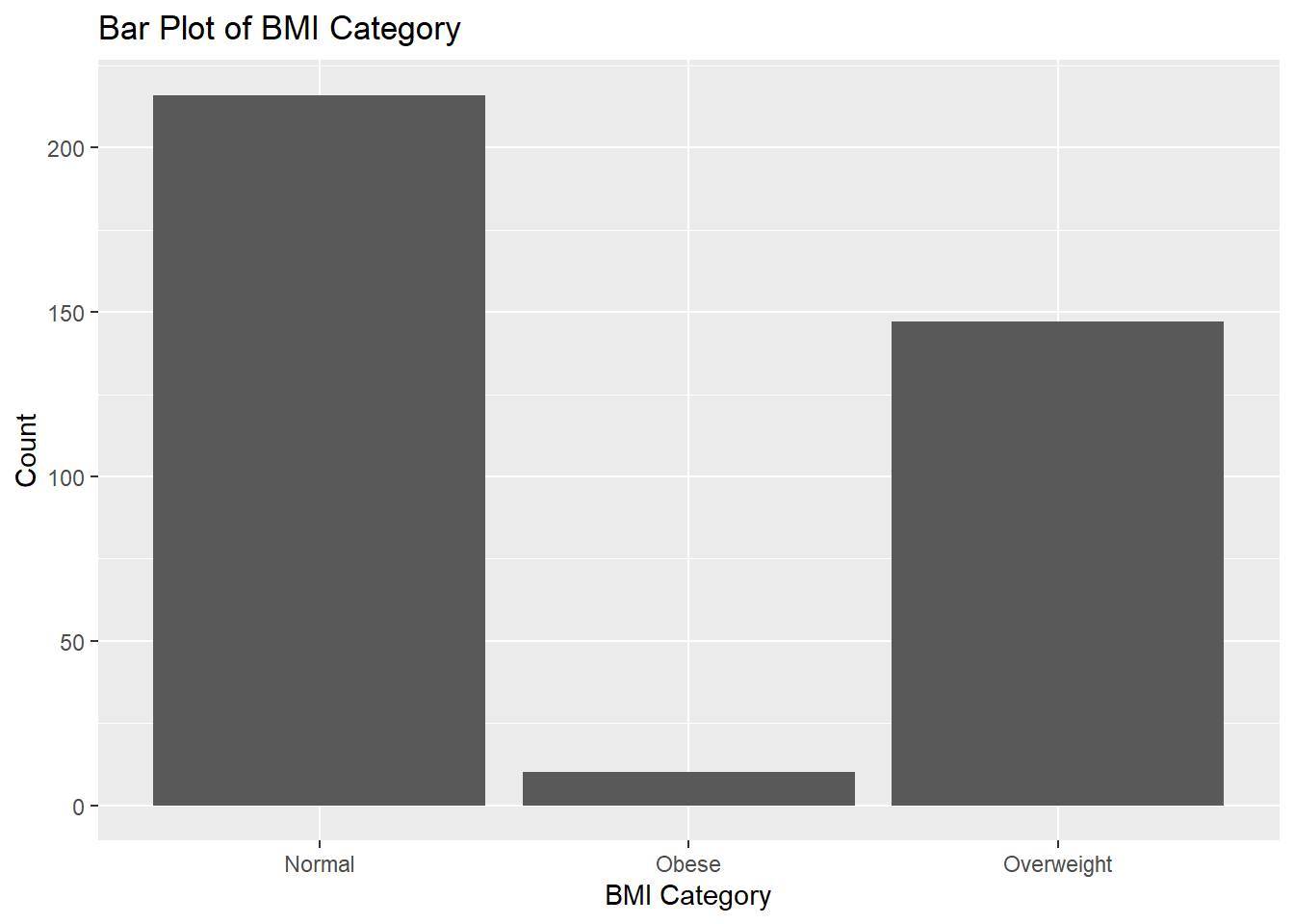


Figure 9: A bar plot showing the counts of our different BMI categories in the data.

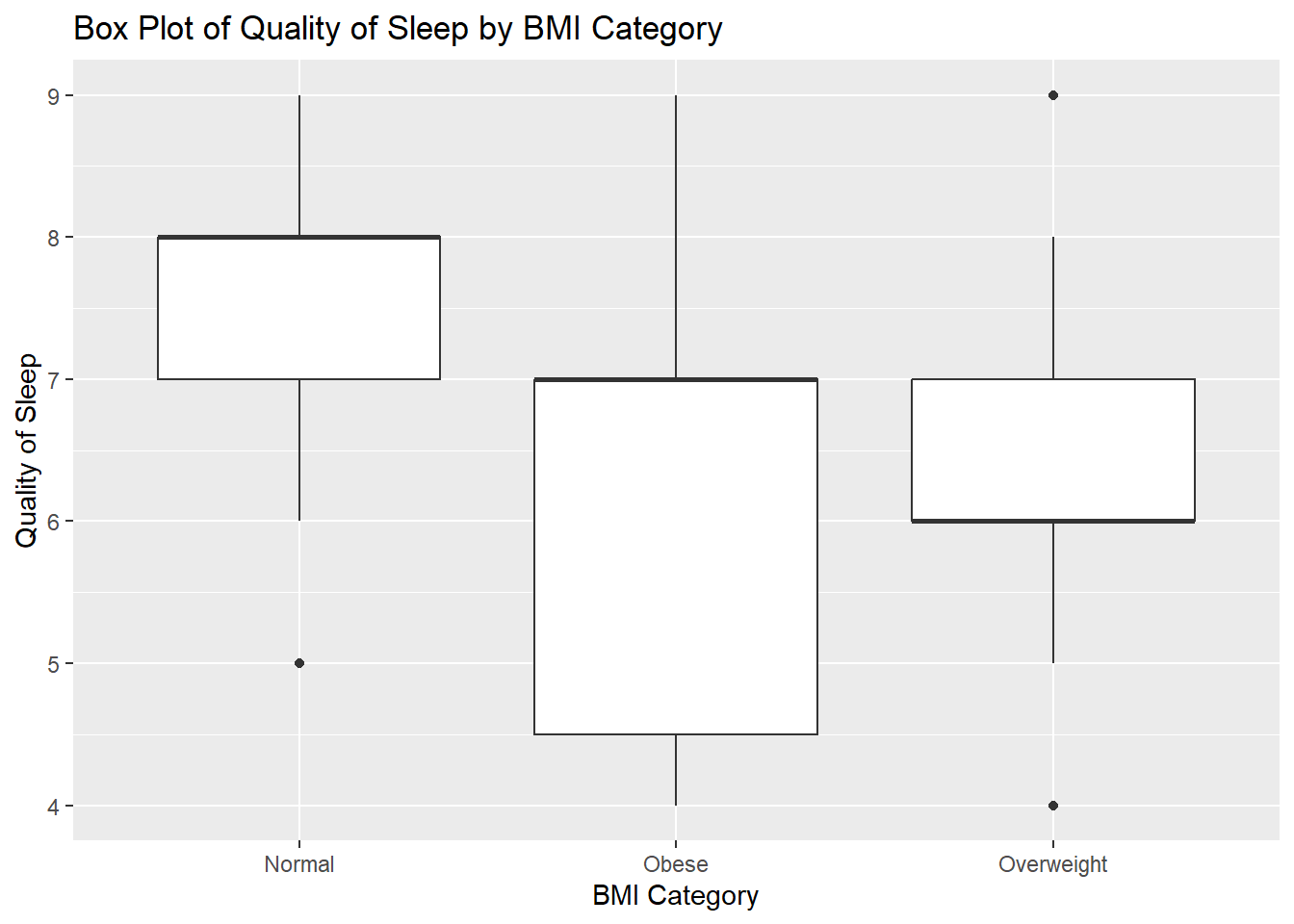


Figure 10: A box plot of BMI category as a predictor of sleep quality.

We then explored BMI with some plots. We can clearly see that Normal weights seem to have better sleep on average than the obese and overweight categories. However, there are very few observations in the obese category, so conclusions with that group may not be as easy to support.

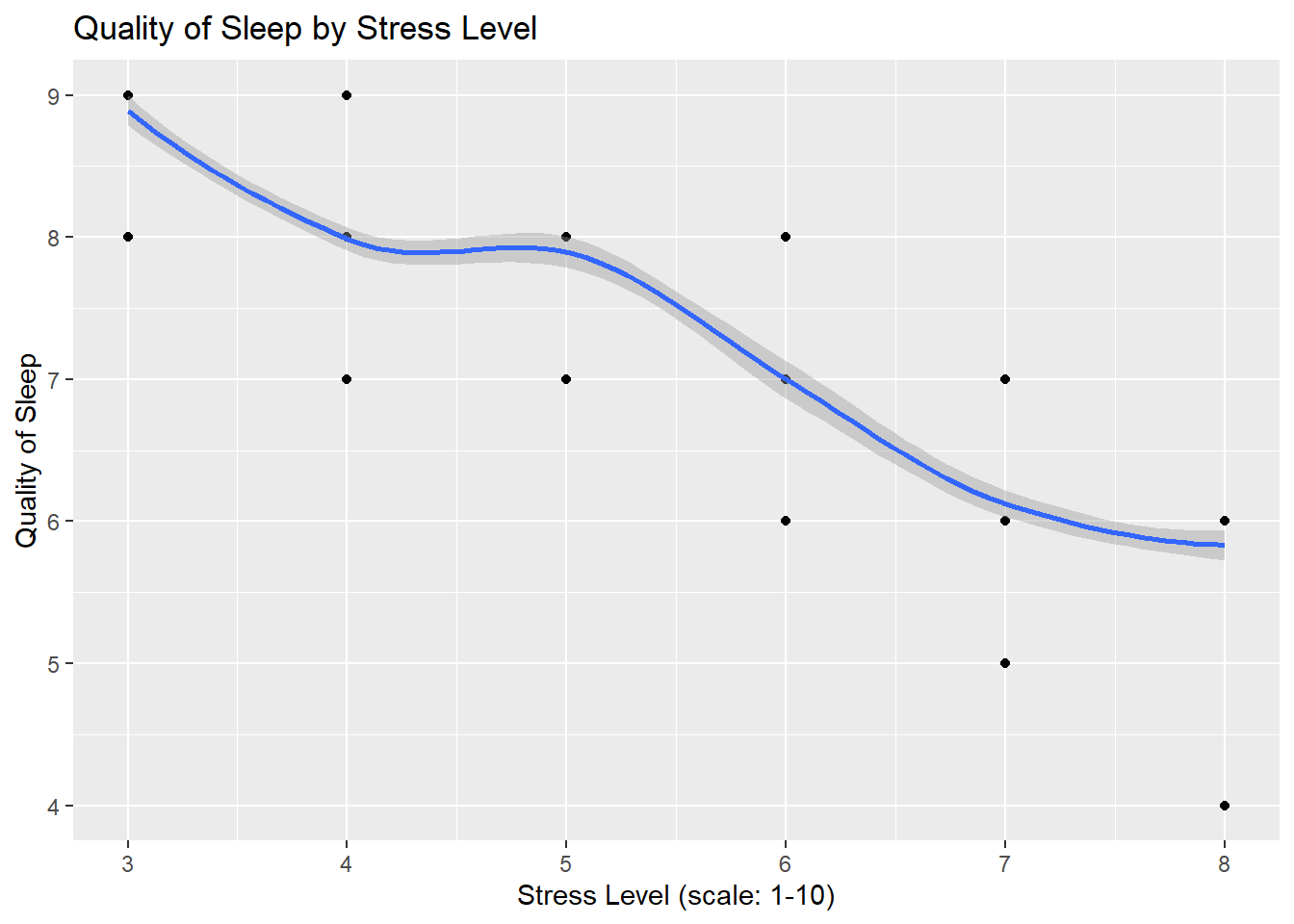


Figure 11: A scatterplot of Stress Level as a predictor for sleep quality.

We then looked at Stress Level as a predictor for quality of sleep. Our scatterplot shows a clear negative correlation between Stress Level and Quality of Sleep.

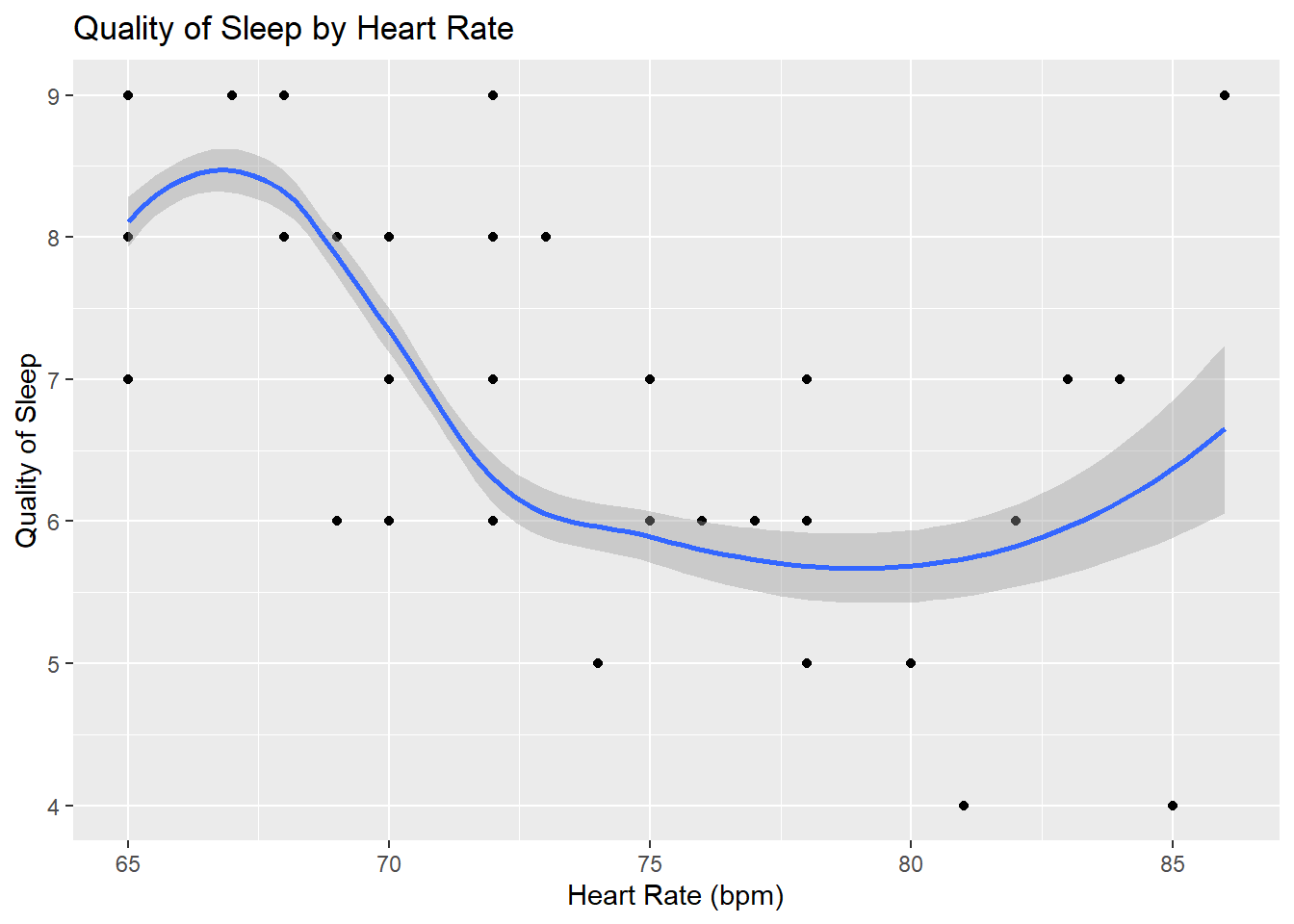


Figure 12: A scatterplot of Heart Rate as a predictor for quality of sleep.

We then looked at Heart Rate’s relationship to Quality of Sleep. We can see a bit of a skewed relationship between the two variables. Healthy Heart Rates tend to differ between each individual, making it difficult to draw conclusions about this variable.

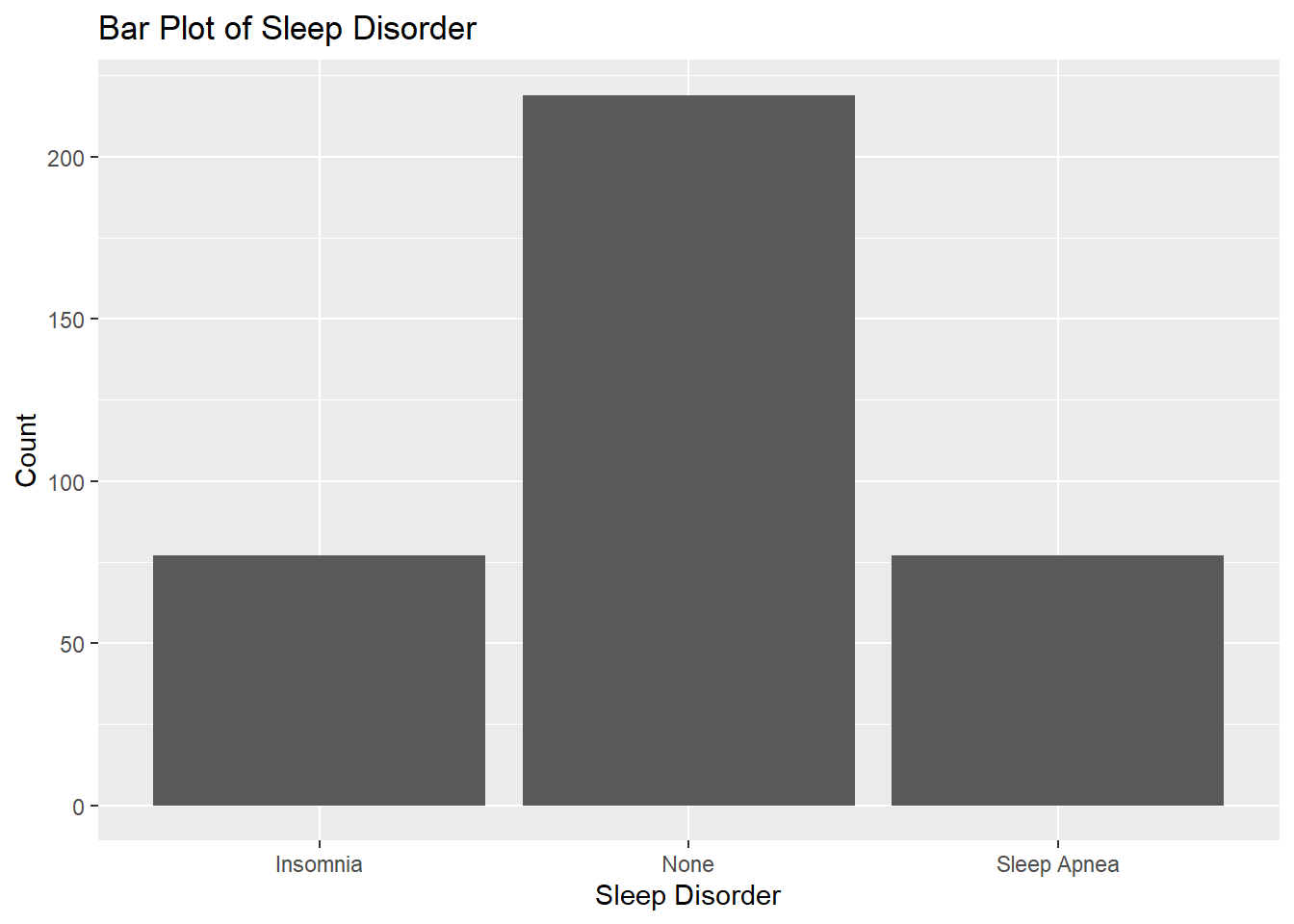


Figure 13: A bar plot showing the count of sleep disorders in our data.

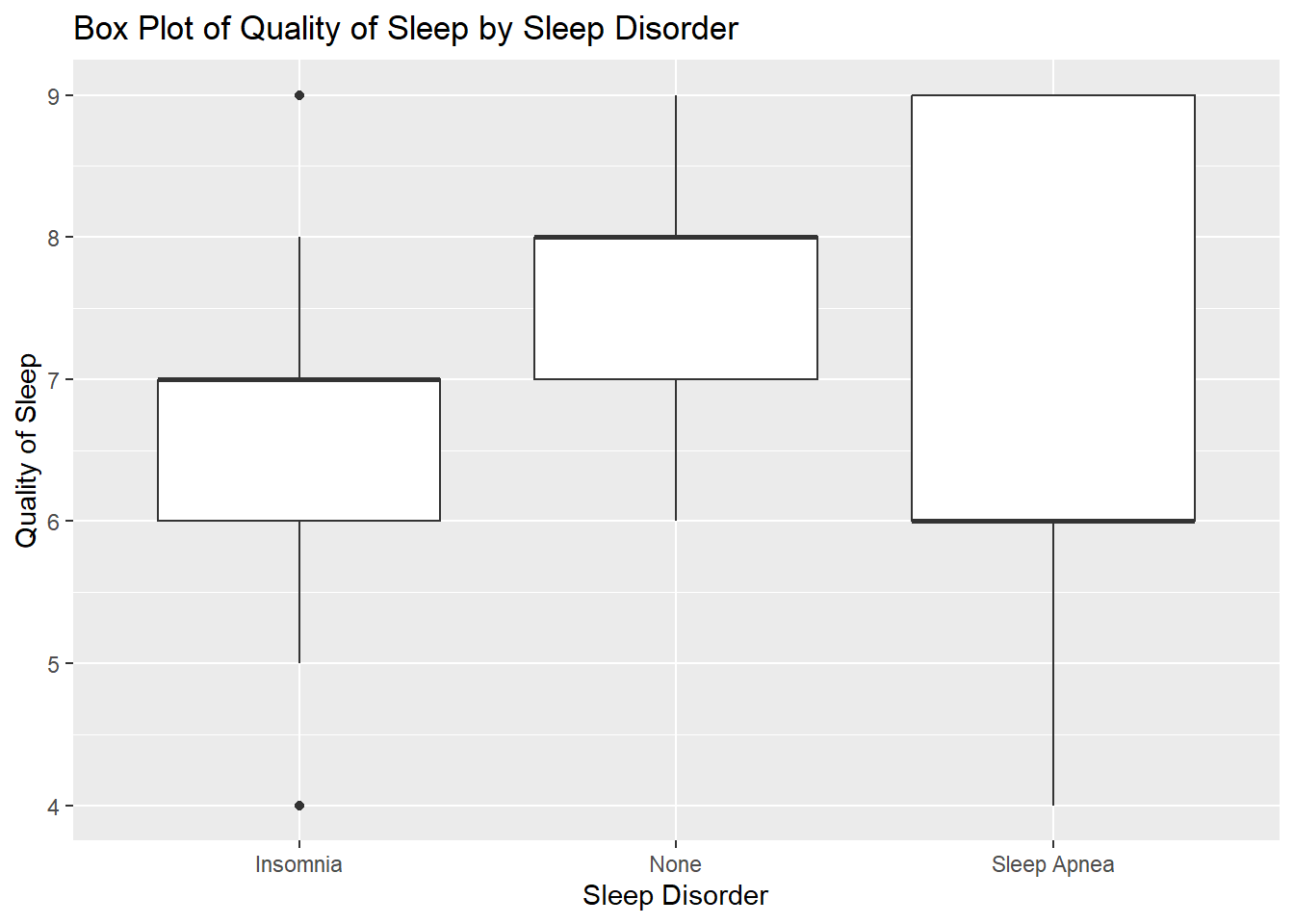


Figure 14: A box plot of Sleep Disorder as a predictor of quality of sleep.

The last variable we investigated was Sleep Disorders. We can see that individuals without a disorder tend to have higher sleep scores than those with Insomnia or Sleep Apnea, which is what we’d expect to see. There are far more observations of individuals without any disorders than of either Sleep Apnea or Insomnia.