Sleep Analysis Summary

Andrew John

12/11/23

# Executive summary

The key questions concerning the Sleep Health and Lifestyle Dataset are focused on understanding the intricate relationships between sleep patterns, daily habits, and overall health. Key questions include: How does physical health affect sleep quality? does gender, and cardiovascular health contribute to health positively or negatively? What is the impact of stress on sleep quality and is there an underlying trend here? Do different occupations take a toll on sleep habits and quality? Do some people with occupations sleep better than others, why?

The story of my analysis of the data attempts to answer all of these questions. The way we live does impact the quality of our sleep. Our physical and mental health has a significant effect on our sleep. I discovered that people with a higher BMI tend to have a far worse sleep quality that those who are in a normal BMI range. From the data I analyzed, I found that a higher percent of the sample population of males were in a normal BMI range and had a wide range of sleep qualities, however, women tended to be in the extremes. Some women had extremely good sleep quality and other women had very poor sleep. When I dug deeper into that narative, I found that our resting heartrate has a lot to do with our sleep quality. Those who have a lower resting heart rate had far better sleep than those who had a higher resting heart rate. This explains why people with a higher BMI have trouble having a good night’s sleep because, generally, those who are overweight or obese also have a higher risk of heart disease and therefore have a higher resting heart rate. Plot number two, provides a trent hidden within the data that suggests that sleep quality and stress have an inverse relationship. As stress levels increased people tended to have poorer sleep quality. This ties in to the occupations plotting in plot number three. In this plot various occupations are plotted to show the sleep quality of the people working in that occupation. From the analysis of that plot, it is clear that doctor and nurses had the worst sleep quality. This makes sense because doctors and nurses have odd work hours and tend to deal with a lot of stress during those work hours.

# Data background

The Sleep Health and Lifestyle Dataset, which has information on 400 people, comes from a file called “The Exempt Organization Business Master File Extract” (EO BMF). This file has details about organizations that don’t pay taxes. The dataset, created by putting together six data files, tells us things about people like how much they sleep, their daily activities, and health details. It’s organized neatly with information on sleep time, quality, physical activity, stress, body weight, and more. The dataset helps us understand how sleep, lifestyle, and health are connected for these 400 individuals.

# Data cleaning

First I downloaded the data from kaggle and stored it under a data folder named data. Then I imported the data into Rstudio by using the ‘read.csv()’ function. Then I changed all the default column names using the names funciton to make them more readable and more usable. After that I filtered out certain occupations that did not have enough data points to contribute significantly to the analysis.

library(tidyverse)  
  
health\_data\_raw <- read.csv("./data/Sleep\_health\_and\_lifestyle\_dataset.csv")  
  
new\_column\_names <- c("ID", "Gender", "Age", "Occupation",  
 "Sleep\_Duration", "Sleep\_Quality", "Physical\_Activity",  
 "Stress\_Level", "BMI", "Blood\_Pressure",   
 "Heart\_Rate", "Steps", "Sleep\_Disorder")  
  
names(health\_data\_raw) <- new\_column\_names  
  
health\_data\_raw <- health\_data\_raw %>%  
 filter(!(Occupation %in% c("Manager", "Sales Representative", "Salesperson", "Scientist", "Software Engineer")))

For figure one specifically there was some more cleaning to do. I created a condensed dataset named “BMI\_summary” from the original health dataset (“health\_data\_raw”). This summary was created to use essential variables, including Body Mass Index (BMI), Sleep Quality, Gender, and Heart Rate. I arranged the dataset based on ascending Sleep Quality to facilitate further analysis. Importantly, I applied a normalization formula to the Heart Rate values, transforming them into a standardized range of 60 to 80. This normalization enhanced the comparability of Heart Rate across different visualizations. Additionally, for simplicity, I cleaned up the BMI categories by relabeling “Normal Weight” as “Normal,” while grouping BMI values identified as “Obese” or “Overweight” under the collective label called “High BMI.”

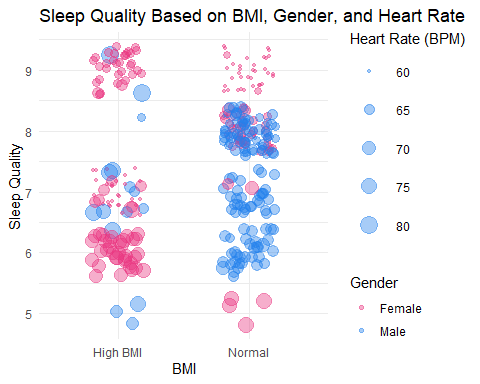
BMI\_summary <- health\_data\_raw %>%   
 select(BMI, Sleep\_Quality, Gender, Heart\_Rate) %>%   
 arrange(Sleep\_Quality)  
  
# change ranges in physical activity  
old\_min <- min(BMI\_summary$Heart\_Rate)  
old\_max <- max(BMI\_summary$Heart\_Rate)  
new\_min <- 60  
new\_max <- 80  
  
BMI\_summary$Heart\_Rate <- ((BMI\_summary$Heart\_Rate - old\_min) / (old\_max - old\_min)) \* (new\_max - new\_min) + new\_min  
  
BMI\_summary <- BMI\_summary %>%  
 mutate(BMI = ifelse(BMI == "Normal Weight", "Normal", BMI)) %>%  
 mutate(BMI = ifelse(BMI %in% c("Obese", "Overweight"), "High BMI", BMI))

# Individual figures

## Figure 1

In creating the figure (1), I used the ggplot package to make a scatter plot. I chose this figure type as it aptly visualized the relationship between Sleep Quality, BMI, Gender, and Heart Rate. For Contrast, I assigned distinct colors each gender (pink for women and blue for men), ensuring that it was easy to differentiate between the two. I coded the Heart Rate information using point size, offering an additional layer of information without any clutter. As for repetition, a consistent blue color scheme was applied throughout the visualization, with shades of blue for males and pink for females. The legend title and color were customized for coherence and alignment with the chosen color palette. I left alligned every plot carefully considering carefully placing titles and axis labels for a clean and organized appearance. The legend was also aligned to enhance readability. For the Proximity, I placed related elements close to each other, such as the legend and the points they represent. Overall layout and spacing were designed for a balanced and harmonious composition.The resulting figure effectively communicated the relationship between Sleep Quality, BMI, Gender, and Heart Rate in a visually engaging and informative manner. After importing to a vector editing software, I customized the fonts to add to the repetition and feel of the plot. I also added circles, lines and annotations to help the viewer better understand the story behind the plot.

# Plot data with points colored by Gender  
plot\_1 <- ggplot(BMI\_summary, aes(x = BMI, y = Sleep\_Quality, color = Gender, size = Heart\_Rate)) +  
 geom\_point(position = position\_jitter(width = 0.2), alpha = 0.40) +  
 labs(title = "Sleep Quality Based on BMI, Gender, and Heart Rate", x = "BMI", y = "Sleep Quality") +  
 guides(size = guide\_legend(title = "Heart Rate (BPM)", keywidth = 2, keyheight = 2, override.aes = list(color = "#2280EC")))+ # Customize legend title and color  
 scale\_color\_manual(values = c("Male" = "#2280EC", "Female" = "#E93580")) +   
 theme\_minimal()  
  
plot\_1



ggsave("plot\_1.pdf", plot\_1, width = 8, height = 6, units = "in")

## Figure 2

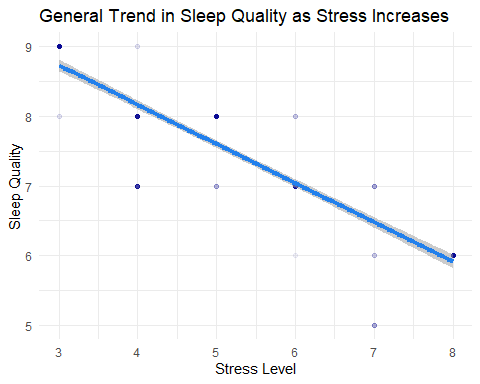
In making figure(2), I used ggplot to construct a scatter plot and a regression line which uses a linear model to portray the correlation between Stress Level and Sleep Quality. I chose graph type due to its effectiveness in illustrating trends and patterns in the given dataset. For Contrast, I set the color of the data points to “darkblue” with reduced alpha (transparency) for a subtle yet distinguishable appearance. Additionally, I added a linear regression line in a contrasting color (#2280EC) with enhanced transparency, contributing to the overall visual contrast and aiding in trend interpretation. Furthermore, for repetition I kept a consistent blue color (#2280EC) for both the regression line and data points. The title, axis labels, and minimalist theme were aligned to the chosen color palette. In terms of Alignment, titles, and labels were carefully placed for clarity and visual hierarchy. The minimal theme ensured a clean and uncluttered layout. I stuck to principles of proximity by placing related elements such as titles, axis labels, and the legend close to the relevant components. The overall layout and spacing were designed to create a well-balanced composition. Then I imported the graphic to Inkscape and added annotations to further enhance the plot. The resulting figure effectively communicated the general trend in Sleep Quality as Stress Level increases, providing a clear representation of the relationship.

plot\_2 <- ggplot(health\_data\_raw, aes(x = Stress\_Level, y = Sleep\_Quality)) +  
 geom\_point(alpha = 0.05, color = "darkblue") +  
 geom\_smooth(method = "lm", color = "#2280EC", size = 1.5, alpha = 0.5) +  
 labs(title = "General Trend in Sleep Quality as Stress Increases",  
 x = "Stress Level", y = "Sleep Quality") +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

plot\_2

## `geom\_smooth()` using formula = 'y ~ x'



ggsave("plot\_2.pdf", plot\_2, width = 8, height = 6, units = "in")

## `geom\_smooth()` using formula = 'y ~ x'

## Figure 3

In creating figure(3), I utilized ggplot to generate a violin plot complemented by points, depicting the distribution of Sleep Quality across different Occupations. I chose this graph for its ability to effectively showcase the variation in Sleep Quality within various professions. To enhance visual contrast, the violin plot was colored with a semi-transparent blue shade (#2280EC), contributing to a clear distinction between the different occupations. Simultaneously, the points were introduced with a matching color and reduced alpha for repetition and visual consistency throughout the plot. For alignment, I carefully placed titles, axis labels, and legend to ensure clarity and a structured visual hierarchy. The minimalist theme was maintained to prevent clutter and create an uncluttered layout. Principles of proximity were adhered to by strategically placing related elements, such as titles and axis labels, in close proximity to the relevant components. Upon completion, I exported the graphic as a PDF into Inkscape for further enhancement like adding colors to contrast the different sleep qualities and to seperate the points within the violin plots based on sleep quality. I also added annotations explaining what the points and the color mean.

# Final figure

For the final figure, I wanted it to be simple and not overly cluttered. I achieved this by dividing the canvas into 4 quarters and left alligning each plot within those four quarters. In the last quarter I added some text fully explaining the three different plots and what they mean. For all of the plots I added annotations about the trends, the interpretation of the clusters and shapes of the plots. I also added lines pointing to the exact region which I am talking about so that the viewers wouldn’t be confused. I chose I clean font and bolded out the titles for emphasis and contrast. I decreased the font size for descriptions for visual hierarchy. Finally I added some colors to plot three to emphasize the sleep qualities in different professions.

|  |
| --- |
| Sleep Analysis - Andrew John |

Sleep Analysis - Andrew John