Manuscript for MADA Project

Kevin Kosewick and Emma Hardin-Parker

2024-03-14

# 1. Summary/Abstract

\*\* NEED TO INSERT \*\*

# 2. Introduction

## 2.1 General Background Information

\*\* NEED TO INSERT \*\*

## 2.2 Description of data and data source

Project Part 1: This dataset was obtained from Kaggle.com at this [link](https://www.kaggle.com/datasets/henryshan/sleep-health-and-lifestyle/data) and was created by Mark Otto and Andrew Fong at MIT. The set has 373 observations for 13 variables covering a broad spectrum associated with sleep and daily routine.

#load in the data and check it out  
library(here)  
library(tidyverse)  
library(dplyr)  
library(skimr)

## 2.3 Questions/Hypotheses to be addressed

We want to explore the influence of each variable on sleep quality and identify which ones seem to have the most impact (both positive and negative). We also wish to explore if certain variables affect male sleep quality more than females. If time allows, we’d also like to see if certain ages seem to be affected by certain variables more often than others. We can focus on every predictor in the dataset since there are only 13.

We can likely use GLMs and t-tests/ANOVA tests to see the significance that variables may have in different combinations on an individual’s sleep quality. Heat maps and various plots will also help with exploratory data analysis. We aren’t exceptionally experienced with statistical analyses, but these are what we can think of at the moment.

# 3. Methods

\*\* NEED TO INSERT \*\*

## 3.1 Data import

We’ll load in the cleaned data here. Please see the eda.qmd file for full exploratory results and the cleaning process.

#load the cleaned data  
sleepdata<- readRDS(here("data","processed-data", "sleepdataprocessed.rds"))

## 3.2 Statistical analysis

We’ll begin to look at some of the relationships between our variables now. Sleep Quality is our outcome of interest for all of these. We’ll start with Phyiscal Activity level.

############################  
#### Fitting a model using Quality of Sleep as the outcome and Physical Activity Level as a predictor  
  
lmfit\_quality\_activity <- lm(Quality.of.Sleep ~ Physical.Activity.Level, sleepdata)   
  
# Placing results from lmfit\_quality\_activty into a data frame with the tidy function  
lmtable\_quality\_activity <- broom::tidy(lmfit\_quality\_activity)  
  
# Viewing the results from the first model fit   
print(lmtable\_quality\_activity)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 6.66 0.183 36.4 2.38e-124  
2 Physical.Activity.Level 0.0109 0.00292 3.74 2.12e- 4

# Saving the lmfit\_quality\_activity results table   
lmtable\_quality\_activity = here("results", "tables", "lmfit1table.rds")  
saveRDS(lmtable\_quality\_activity, file = lmtable\_quality\_activity)

The intercept value indicates that Quality of Sleep will be 6.66 if physical activity is at 0. This has a relatively low standard error and high significance.The coefficient for our variable indicates that as Physical activity level increases by one unit, Quality of Sleep will increase by 0.0109. Our t-statistic is indicated as significant by the p-value but is much lower than the intercept’s.This means that Physical Activity Level does have a measurable impact on Quality of Sleep, but it’s relatively small.

Next, we’ll look at Sleep Duration as a predictor.

############################  
#### Fitting a model using Quality of Sleep as the outcome and Sleep Duration as a predictor  
  
lmfit\_quality\_duration <- lm(Quality.of.Sleep ~ Sleep.Duration, sleepdata)   
  
# Placing results from lmfit\_quality\_duration into a data frame with the tidy function  
lmtable\_quality\_duration <- broom::tidy(lmfit\_quality\_duration)  
  
#Viewing the results from the second model fit   
print(lmtable\_quality\_duration)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) -2.15 0.263 -8.18 4.51e- 15  
2 Sleep.Duration 1.33 0.0367 36.2 9.70e-124

# Saving the lmfit\_quality\_duration results table   
lmtable\_quality\_duration = here("results", "tables", "lmfit2table.rds")  
saveRDS(lmtable\_quality\_duration, file = lmtable\_quality\_duration)

For our sleep duration model, we can see that our intercept’s negative value indicates that Quality of Sleep would be very poor if individuals got no sleep. This is rational and is supported by a strong p-value and a decent t-statistic, although the standard error is a bit high. For our variable coefficient, we can see that as sleep duration increases by one unit, Quality of sleep also increases by about 1. This has a very strong p-value and t-statistic, indicating a strong relationship between this predictor and our outcome of interest.

Next, we’ll look at Stress Level as our predictor.

############################  
####Fitting a model using Sleep Quality as the outcome and Stress Level as a predictor.  
  
lmfit\_quality\_stress <- lm(Quality.of.Sleep~ Stress.Level, sleepdata)  
  
#Placing results from lmfit\_quality\_stress into a data frame with the tidy function  
lmtable\_quality\_stress <- broom::tidy(lmfit\_quality\_stress)  
  
#Viewing the results from the sixth model fit   
print(lmtable\_quality\_stress)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 10.6 0.0873 121. 2.42e-300  
2 Stress.Level -0.606 0.0154 -39.4 1.60e-134

# Saving the lmfit\_quality\_stress results table   
lmtable\_quality\_stress = here("results", "tables", "lmfit6table.rds")  
saveRDS(lmtable\_quality\_stress, file = lmtable\_quality\_stress)

According to our model, Stress has a strongly defined relationship with sleep quality. We can see that a value of 0 on the stress score leads to a 10 unit increase in quality of sleep. The Stress Level coefficient indicates that as stress increases by one unit, sleep quality decreases by about 0.6 of a unit. The t-statistics and p-values for the values in this model are extremely strong, indicating a well-defined relationship between stress and sleep quality.

Finally, we’ll look at Age as our predictor.

############################  
####Fitting a model using Sleep Quality as the outcome and Age as a predictor.  
  
lmfit\_quality\_age <- lm(Quality.of.Sleep~ Age, sleepdata)  
  
#Placing results from lmfit\_quality\_gender\_age into a data frame with the tidy function  
lmtable\_quality\_age <- broom::tidy(lmfit\_quality\_age)  
  
#Viewing the results from the seventh model fit   
print(lmtable\_quality\_age)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 4.57 0.273 16.8 2.60e-47  
2 Age 0.0650 0.00634 10.3 6.75e-22

# Saving the lmfit\_quality\_gender\_age results table   
lmtable\_quality\_age = here("results", "tables", "lmfit8table.rds")  
saveRDS(lmtable\_quality\_age, file = lmtable\_quality\_age)

Our model shows a strong relationship between age and quality of sleep. As Age increases, sleep quality seems to improve. Each year increase in age is predicted to have a 0.065 increase in sleep quality score. The intercept here is also interesting as an age of “0” is predicted to have a sleep score of 4.57, which is relatively poor. The t-statistics and p-values for these are both strong.

# 4. Results

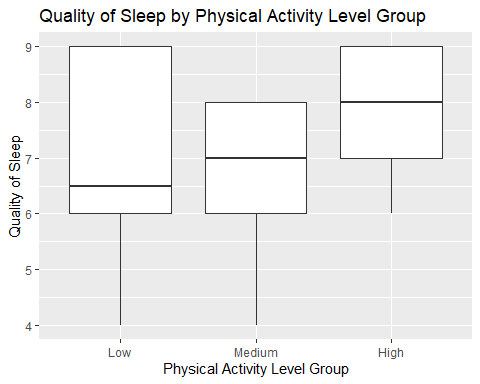
## 4.1 Exploratory/Descriptive analysis

Here are the most interesting/distinct plots from the EDA:

Physical Activity plot:

We can clearly see that Physical Activity level has a positive correlation with Quality of Sleep.

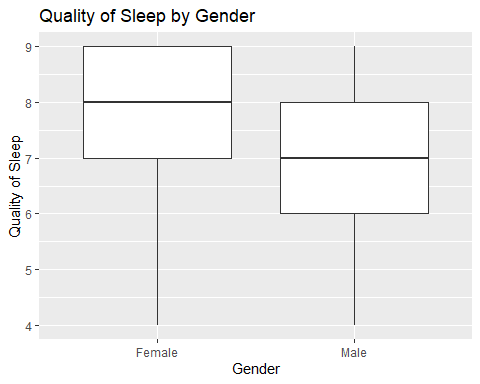
#load plot package  
library(ggplot2)  
  
# Box plot for Quality of Sleep by Physical Activity Level Group  
ggplot(sleepdata, aes(x = PhysicalActivityGroup, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Quality of Sleep by Physical Activity Level Group", x = "Physical Activity Level Group", y = "Quality of Sleep")



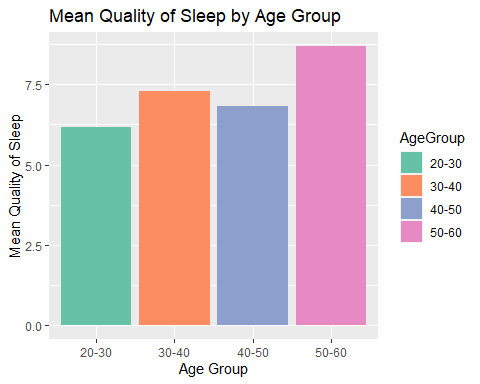
Gender, Age, and Occupation:

We can see some clear differences in self-reported Quality of Sleep between these different groups already. 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 occupations are a bit all over the place, but Sales Representatives seem to have the worst average scores by far.

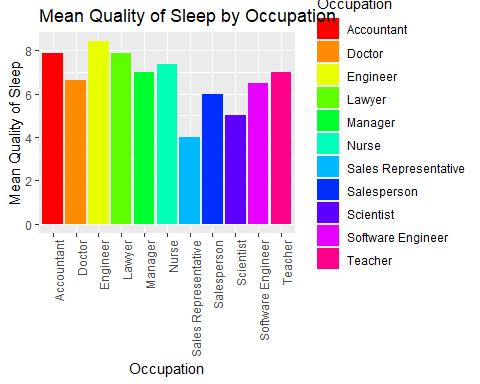
# Boxplot for Quality of Sleep by Gender  
ggplot(sleepdata, aes(x = Gender, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Quality of Sleep by Gender", x = "Gender", y = "Quality of Sleep")



# Create bins for Age in order to create a clean boxplot.  
sleepdata$AgeGroup <- cut(sleepdata$Age, breaks = c(20, 30, 40, 50, 60), labels = c("20-30", "30-40", "40-50", "50-60"), include.lowest = TRUE)  
  
# Calculate mean Quality of Sleep for each Age Group to plot with means for cleaner visuals  
sleepdata\_summary <- sleepdata %>%  
 group\_by(AgeGroup) %>%  
 summarise(MeanQuality = mean(Quality.of.Sleep, na.rm = TRUE))  
  
# Bar plot for Mean Quality of Sleep by Age Group with colorblind-friendly palette  
ggplot(sleepdata\_summary, aes(x = AgeGroup, y = MeanQuality, fill = AgeGroup)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_brewer(palette = "Set2") +  
 labs(title = "Mean Quality of Sleep by Age Group", x = "Age Group", y = "Mean Quality of Sleep")

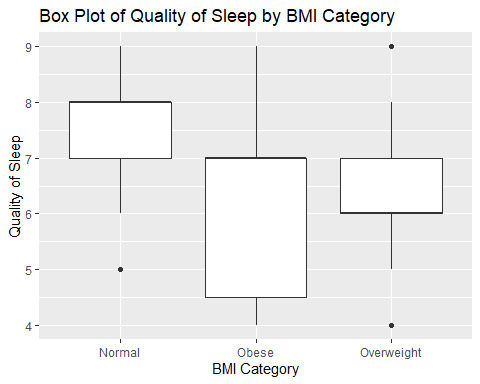


# Calculate mean Quality of Sleep for each Occupation in order to create a bar plot.  
sleepdata\_summary <- sleepdata %>%  
 group\_by(Occupation) %>%  
 summarise(MeanQuality = mean(Quality.of.Sleep, na.rm = TRUE))  
  
# Bar plot for Mean Quality of Sleep by Occupation with manually specified colors  
ggplot(sleepdata\_summary, aes(x = Occupation, y = MeanQuality, fill = Occupation)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_manual(values = rainbow(length(unique(sleepdata\_summary$Occupation)))) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 labs(title = "Mean Quality of Sleep by Occupation", x = "Occupation", y = "Mean Quality of Sleep")



BMI: 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 in the obese category, so conclusions with that group may not be as supported as others.

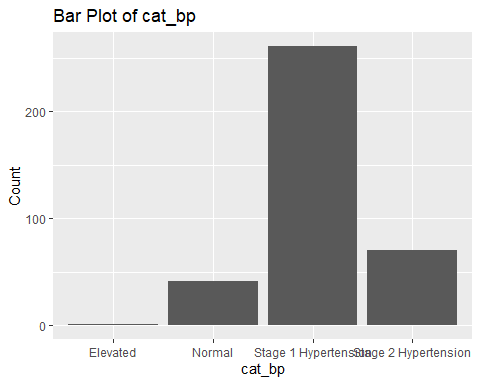
# Box plot for Quality of Sleep by BMI Category  
ggplot(sleepdata, aes(x = BMI.Category, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Quality of Sleep by BMI Category", x = "BMI Category", y = "Quality of Sleep")



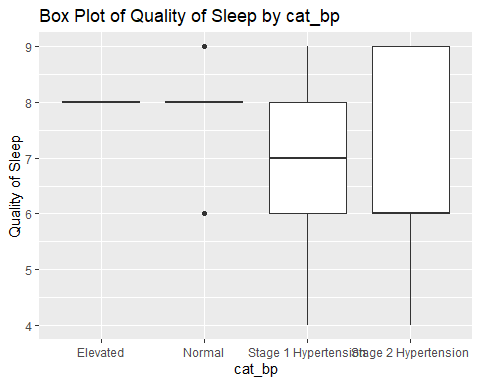
Blood Pressure:

The boxplots look a bit strange for the Elevated and Normal levels, but we can see that it’s likely due to a shortage of data for these categories. We can see, however, that Quality of Sleep does seem to be negatively correlated with increasing stages of hypertension.

# Bar plot for cat\_bp  
ggplot(sleepdata, aes(x = cat\_bp)) +  
 geom\_bar() +  
 labs(title = "Bar Plot of cat\_bp", x = "cat\_bp", y = "Count")



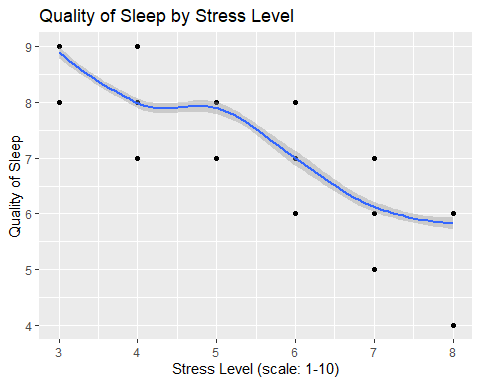
# Box plot for Quality of Sleep by cat\_bp  
ggplot(sleepdata, aes(x = cat\_bp, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Quality of Sleep by cat\_bp", x = "cat\_bp", y = "Quality of Sleep")



And Stress Level: Our scatterplot shows a clear negative correlation between Stress Level and Quality of Sleep. However, this is a subjective scale, so data should be taken with a grain of salt for now.

# Scatterplot for Quality of Sleep by Stress Level  
ggplot(sleepdata, aes(x = Stress.Level, y = Quality.of.Sleep)) +  
 geom\_point() +  
 geom\_smooth(method = "loess") +  
 labs(title = "Quality of Sleep by Stress Level", x = "Stress Level (scale: 1-10)", y = "Quality of Sleep")

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



The other plots were either self explanatory (e.g. Sleep Duration meant higher sleep quality) or had less defined trends (e.g. heart rate and daily steps).

End of Project Part 2.

## 4.2 Basic statistical analysis

**Included in the analysis-code folder for now**

## 4.3 Full analysis

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References