An Analysis of Health & Lifestyle Factors Associated with Sleep Quality

Kevin Kosewick and Emma Hardin-Parker

2024-04-04

*ANYTHING BELOW CODE LINE 172 IS A DRAFT AND IS NOT TO BE CONSIDERED A PART OF THE FINAL MANUSCRIPT*

# 1. Summary/Abstract (1)

*NEED TO INSERT*

*WILL NEED COLLAB FOR PART 5*

# 2. Introduction (2)

## 2.1 Background Information

The data was obtained from kaggle.com at this [link](https://www.kaggle.com/datasets/henryshan/sleep-health-and-lifestyle/data). It contains data from a study on individuals of different sexes and age regarding their sleep quality and different variables that may affect this. We are unsure of where this data set came from and how it was collected; the publisher on Kaggle has yet to respond to a comment with this question posted on March 5th 2024. Regardless, it contains interesting variables that have been shown in previous studies to have an effect on sleep quality.

For instance, in Sun et al. (2015), a sample of Chinese individuals showed a correlation between obesity and worsened sleep quality in men but not women. We are interested in seeing if this trend holds up for our dataset; given that most of the observations are of men, we expect to see a correlation between lower sleep scores and obesity in our data. Furthermore, it has been shown that age and gender have a notable effect on sleep quality (Madrid-Valero et al. 2016). Women seem to experience a deterioration in sleep quality as they age; however, while this trend is still present in men, it is notably less consistent and can vary dramatically between individuals. We expect to see similar trends in our dataset, but as stated previously, our observations contain significantly more men than women. Finally, we wish to test whether or not certain occupational industries have different effects on sleep quality. It has been shown that managerial positions tend to have the worst sleep quality among civilian sector workers while 24 jobs that have rotating shifts tend to have the worst sleep quality (Luckhaupt et al. 2010). Our dataset contains observations of mostly white-collar/service-based jobs, so we expect to find similar trends upon analysis.

*NEED TO INSERT*

*EMMA’S SOURCES ARE LINKED HERE:*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8345503/ –> blood pressure and sleep https://www.apa.org/news/press/releases/stress/2013/sleep –> sleep duration & stress on sleep quality https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10503965/#:~:text=Regular%20physical%20activity%20can%20lead,managing%20sleep%20disorders%20like%20insomnia. –> physical activity and sleep quality

# 3. Materials and Methods (3)

## 3.1 Data and Processing

The dataset originally contained 373 observations for 13 variables covering a broad spectrum health and lifestyle variables associated with sleep quality. The curators of the dataset did not provide information on how or where this data was collected, so we solely relied on the codebook for variable definitions. There were no missing values or erroneous variables in the dataset that we needed to remove. The “Blood.Pressure” variable was transformed to reflect the systolic and diastolic blood pressure measurements from each subject. We then chose to add an additional variable, labeled ‘cat\_bp’ to reflect the categorical blood pressure status of a subject based on the American Heart Association guidelines *INCLUDE REFERENCE HERE (can be found in eda.qmd file)* . Additionally we created a variable, ‘StepsGroup’ to determine the categorical activity level of subjects in the dataset based on these guidelines provided by 10,000 steps [*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)*.* Finally, we created an additional categorical variable, ‘PhysicalActivityGroup’ based on 30 minute increments, to categorically represent the level of physical activity reported by a subject. The outcome of interest is noted as “Quality.of.Sleep” that we will reference as Sleep Quality or Quality of Sleep throughout this report.

*Unsure if we need this to be here? Or if it will even show up given the FALSE arguments we added*

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

## 3.2 Variables included in Analyses *Include Questions / Hypotheses Here*

Given that there were only 13 original variables, and 17 variables after final data transformation; simple linear regression models were performed to determine baseline associations between variables and Sleep Quality, the outcome of interest. These regression models were created to determine which variables had the largest impact on Sleep Quality and drove the rest of the analyses. Simple linear regression models were fitted for the following variables: BMI, cat\_bp, Stress Level, Gender, Age, and Occupation.

The Occupation variable was then transformed to group the various occupations into the following groups: Healthcare, Education, Engineering, Business/Finance, and Science. Prior to performing any subsequent analyses, we removed the original versions of the variables we transformed: systolic, diastolic, daily steps, occupation, Physical.Activity.Level, Heart.Rate, and AgeGroup.

## 3.3 Model Development

Random Forest Models and Cross-Validation were considered to be the “best” fit for our outcome of interest, Quality of Sleep. The data were not split into train/test subsets as the data contains less than 400 observations and many of the values were unique in comparison to the rest of the data. Prior to constructing this model, a collinearity plot was constructed to determine any presence of collinearity in our data. *INCLUDE PLOT HERE* Stress and Sleep Quality had an absolute value of 0.9 on the correlation scale. Sleep Duration and Stress were also seen to be strongly correlated. Stress was still included in our models; however, we found it important to note the strong linear relationship between Stress and Sleep Quality.

A variable importance plot (VIP) was constructed to determine which variables had the strongest impact in our Random Forest Model. *INCLUDE PLOT HERE* Sleep Duration, Stress Level, and Age had the highest levels of importance in our model. Physical Activity Level was excluded entirely, which was to be expected given our exploratory data analysis and simple linear regression models.

## 3.4 Defining the Models

*DEFINE DECISION TREES. DEFINE BAGGING AND BOOTSTRAPPING. DEFINE RANDOM FOREST MODELS* *DEFINE CROSS-VALIDATION* *DEFINE IMPORTANCE OF NULL MODEL TO COMPARE RESULTS*

*INCLUDE CITATIONS OF CHAPTER 9,10,11 OF HMLR LINKED IN “MANY TREE MODELS” SECTION OF CONTENT FOLDER 11*

## 3.5 Evaluation of Models

Root Mean Squared Error (RMSE) is a common metric used to evaluate regression models. The RMSE is formally defined as the square root of the mean square of all error and is defined by the following formula: *Include LaTeX equation of RMSE?* . It is important that while RMSE is scale dependent, common practice notes that low RMSE values indicate stronger model performance.

## 3.6 Software Used for Analyses

This analysis was conducted under R version 4.3.1 on a MacOS operating system *cite R*. The following R packages were used in the development of these analyses: *R PACKAGES USED: here, skimr, broom, tidyverse, ggplot2, dplyr, corrplot, ranger, vip* All processing and analysis code can be found in the Supplementary Material file.

# 4. Results (4)

## 4.1 Outcomes of Interest

The ‘Quality of Sleep’ variable has been selected as our outcome of interest. It is important to note that this variable is a subjective measure of a participant’s self-reported sleep quality. The following plot shows that most subjects reported a sleep quality score of 8 (out of 10). *REFERENCE SLEEP\_DIS PLOT FROM EDA.QMD FILE*

## 4.2 Exploratory Findings

## 4.3 Machine Learning Models

A variable importance plot (VIP) was constructed to determine which variables had the strongest impact in our Random Forest Model. *INCLUDE PLOT HERE* Sleep Duration, Stress Level, and Age had the highest levels of importance in our model. Physical Activity Level was excluded entirely, which was to be expected given our exploratory data analysis and simple linear regression models.

*Include comparison of RF Model and null model*

# 5. Discussion

*PART 5*

## 5.1 Summary

*PART 5*

## 5.2 Strengths and Limitations

*PART 5*

*Include that all parameters were self-reported and thus subjected to bias*

## 5.3 Conclusion

*PART 5*

# 6. References

*INCLUDE INCREMENTALLY – FINALIZE*

https://link.springer.com/article/10.1007/s11325-015-1193-z –> obesity and sleep quality (affects men but not women in China)

https://www.scielosp.org/article/gs/2017.v31n1/18-22/en/?uid=680b54f39b# –> effect of age and gender on sleep quality (women worse than men, sleep quality deteriorates with age in women but is a bit all over the place for men)

https://academic.oup.com/sleep/article/33/2/149/2454438#77919376 –> sleep quality by occupation/industry (managerial positions had the worst score)

*ANYTHING BELOW THIS COMMENT IS FROM PRIOR TO THE OVERHAUL AND IS STILL TO BE IGNORED / MOVED TO ANOTHER SECTION OF THE PROJECT*

## 6.1 Questions/Hypotheses to be addressed

We intend 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.

## 6.2 Data import

We’ll load in the cleaned data here. Please reference the eda.qmd and Supplementary-Material.qmd files for full exploratory results and the cleaning process.

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

## 6.3 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 Physical 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.

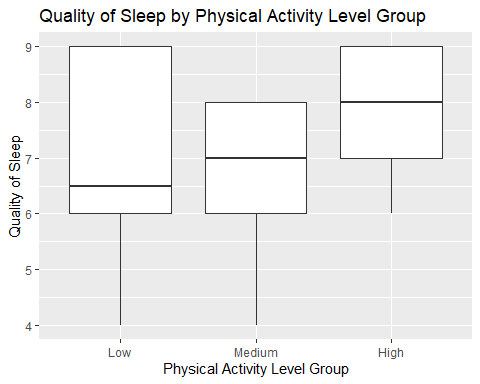
## 6.4 Exploratory/Descriptive analysis

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.

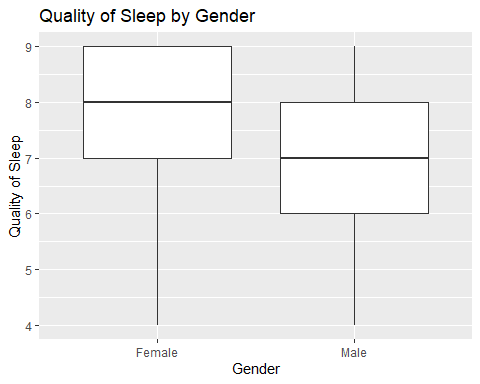
#load plot package  
library(ggplot2)  
  
# Box plot for Quality of Sleep by Physical Activity Level Group  
ggplot(sleepdata, aes(x = PhysicalActivityGroup, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Quality of Sleep by Physical Activity Level Group", x = "Physical Activity Level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



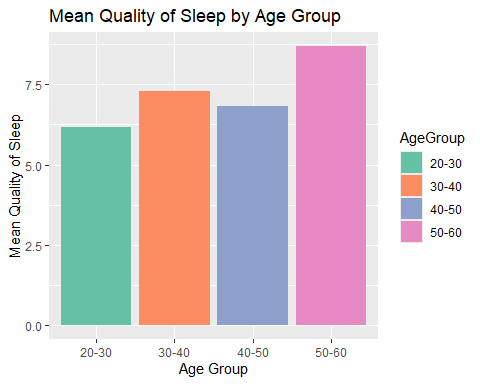
Gender, Age, and Occupation:

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.

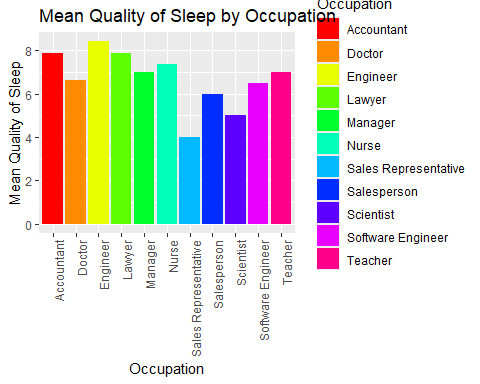
# Boxplot for Quality of Sleep by Gender  
ggplot(sleepdata, aes(x = Gender, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Quality of Sleep by Gender", x = "Gender", y = "Quality of Sleep")



# 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")

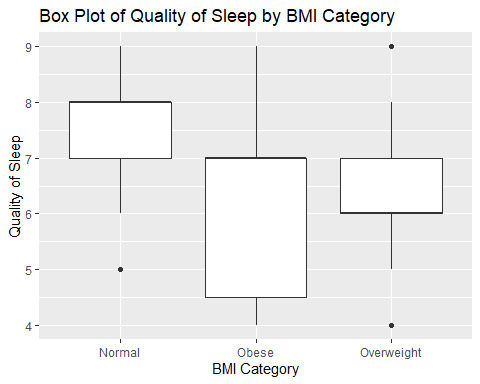


# Calculate mean Quality of Sleep for each Occupation in order to create a bar plot.  
sleepdata\_summary <- sleepdata %>%  
 group\_by(Occupation) %>%  
 summarise(MeanQuality = mean(Quality.of.Sleep, na.rm = TRUE))  
  
# Bar plot for Mean Quality of Sleep by Occupation with manually specified colors  
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.

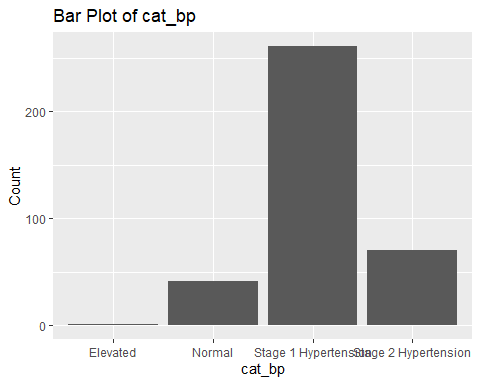
# Box plot for Quality of Sleep by BMI Category  
ggplot(sleepdata, aes(x = BMI.Category, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Quality of Sleep by BMI Category", x = "BMI Category", y = "Quality of Sleep")



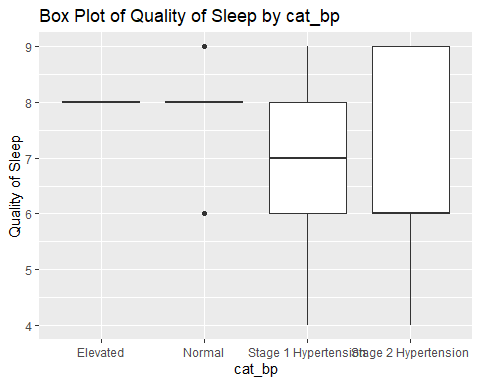
Blood Pressure:

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")



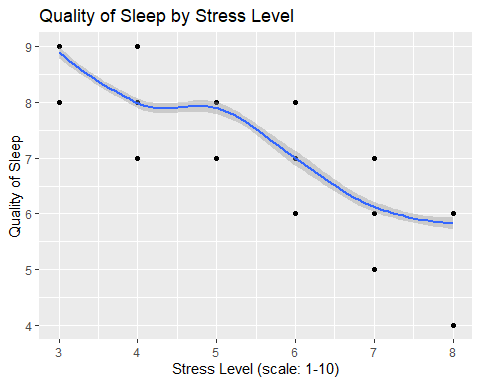
# Box plot for Quality of Sleep by cat\_bp  
ggplot(sleepdata, aes(x = cat\_bp, y = Quality.of.Sleep)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Quality of Sleep by cat\_bp", x = "cat\_bp", y = "Quality of Sleep")



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).