An Analysis of Health & Lifestyle Factors Associated with Sleep Quality

Kevin Kosewick and Emma Hardin-Parker

Invalid Date

# 1. Summary/Abstract (1)

For our project, we obtained a dataset from Kaggle.com containing 373 observations of 13 variables from a study on different factors that may affect sleep quality. Variables of interest from the study included physical activity level, age, sex, occupation, stress, and several more. In RStudio (version 4.3), we used a combination of packages to create models suitable for answering our primary question: which variables are most important for determining an individual’s sleep quality and what effect do they have? After conducting an exploratory data analysis (EDA) we hypothesized that stress, sleep duration, and occupation would have the largest impact on sleep quality. We believed that stress and sleep quality would be negatively correlated, sleep duration and quality would be positively correlated, and that jobs with longer or more variable hours would negatively affect sleep quality.

After creating linear models for each variable’s impact on sleep quality and fitting a Random Forest model with Cross-Validation to test variable importance, we determined that sleep duration, stress level, and age were the most important factors. Our hypothesis for sleep duration and stress was supported; however, we did not correctly predict that age would be one of the most important predictors for sleep quality. Our hypotheses regarding the effects of sleep duration, stress, and occupation were all supported as well. Overall, we determined that an individual’s duration of sleep, stress level, and age are all key factors that can impact sleep quality.

*Feel free to edit this Emma; I tried to include everything needed*

# 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*](https://www.heart.org/en/health-topics/high-blood-pressure). 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.

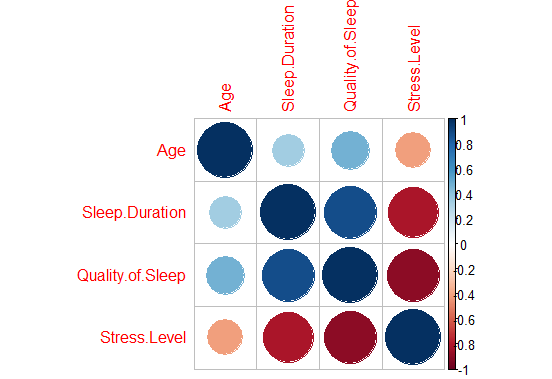
## 3.2 Variables included in Analyses

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.

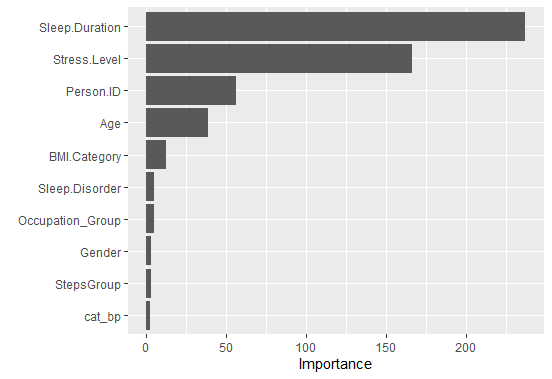
The Occupation variable was then transformed to group the various occupations into the following groups: Healthcare, Education, Engineering, Business/Finance, and Science. Prior to performing any subsequent analyses, we removed the original versions of the variables we transformed: systolic, diastolic,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

The objective of our analysis was to determine which of these variables are most important for determining an individual’s sleep quality and what specific effect they have. After conducting an (EDA) we hypothesized that stress, sleep duration, and occupation would have the largest impact on sleep quality. We believed that stress and sleep quality would be negatively correlated, sleep duration and quality would be positively correlated, and that jobs with longer or more variable hours would negatively affect sleep quality.

## 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. 

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 due to its well-studied impact on sleep; 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. 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*

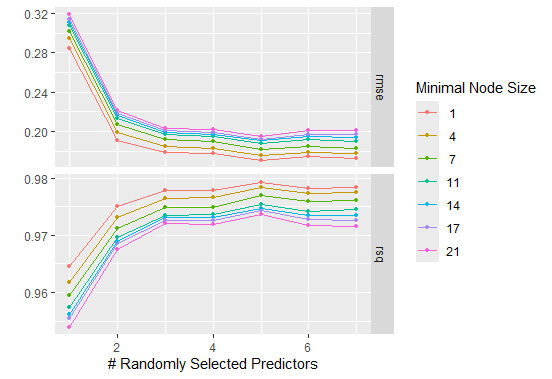
## 3.5 Evaluation of Models

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:

It is important to consider that while RMSE is scale dependent, common practice notes that low RMSE values indicate stronger model performance.

R-squared (R²) is a statistical measure used in regression analyses. It quantifies the proportion of the variance in a dependent variable that can be explained by an independent variable within a regression model and is calculated with the following formula:

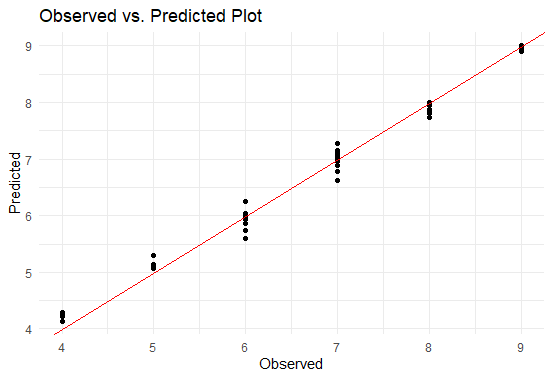
We used both RMSE and R² to evaluate the performance of our model compared to a “null model”. Null models generate predictions for an outcome of interest without using any predictors. Instead, they simply average the values of the outcome of interest to make “predictions”.



The results generated from our RF model. RMSE increases when more predictors are added to the model. The R-squared value also decreases as more predictors are added to the model.

Figure (*name/order figures after we finalize draft*) shows that as we include more predictors in our model, RMSE increases slightly while the R-squared value decreases. This indicates that some variables in our dataset are more important for predicting sleep quality than the others. Based off of our variable importance graph, we can assume that stress level, sleep duration, and age are the factors that carry the most weight when predicting sleep quality.

We went on to evaluate how well our RF model predicted values by creating an observed vs. predicted plot.



A comparison of the actual values for sleep quality in our dataset (observed) compared to the values predicted by our fitted RF model.

While the predicted values don’t align perfectly with the red “ideal value” line, they fall very close overall, indicating that our model can sufficiently predict sleep quality using unseen data.

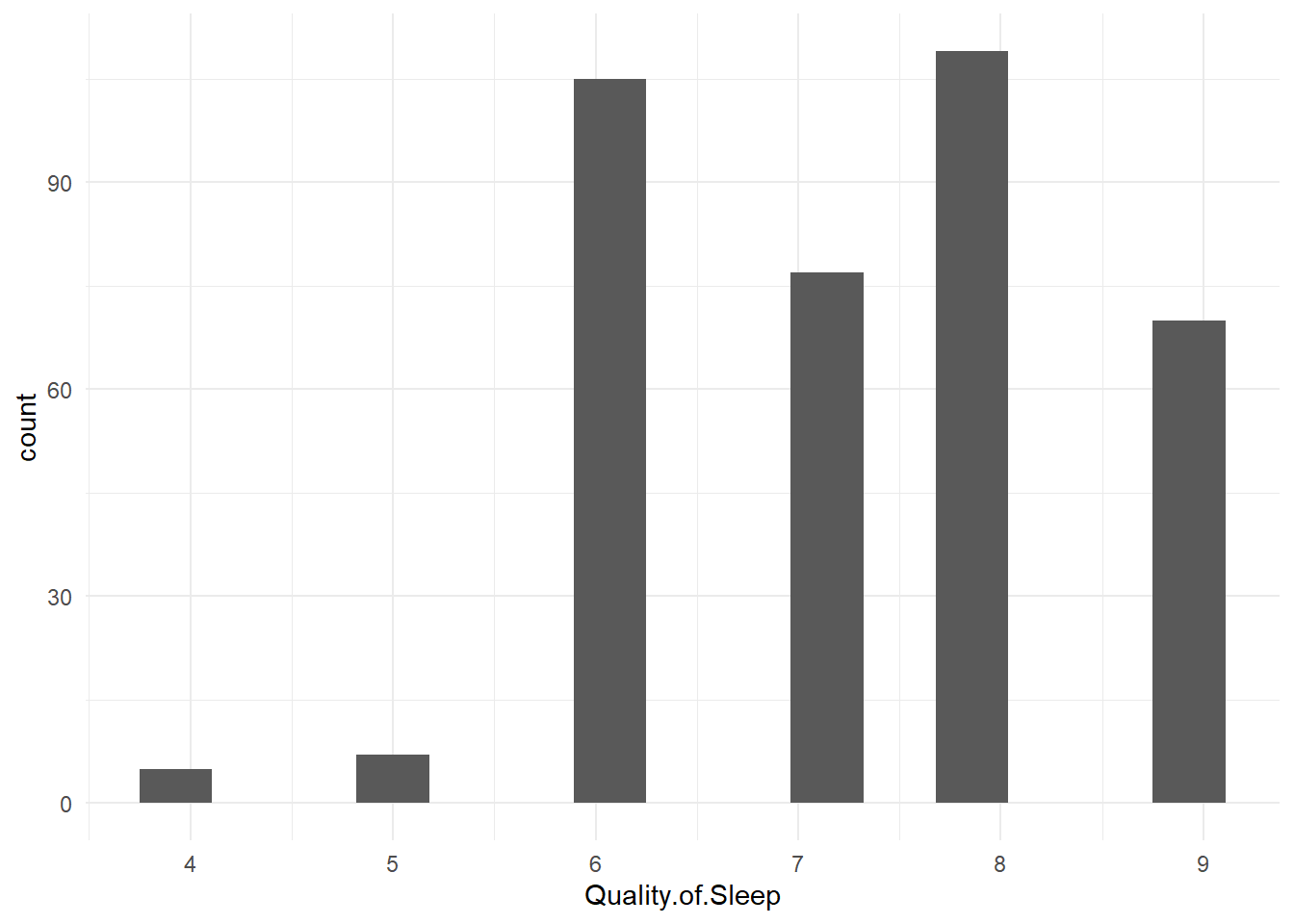
## 3.6 Software Used for Analyses

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).



Distribution of Sleep Quality Scores in the dataset.

## 4.2 Exploratory Findings

## 4.3 Machine Learning Models

*To Emma: did you want this text in both places you placed it?* 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 (5)

*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)