Sleep Health and Lifestyle

John Little, jlittle3@bellarmine.edu
Yeva Kramarova, ykramarova@bellarmine.edu

ABSTRACT

Our project is the final project for the Introduction of Data Science course in Bellarmine University. The main goal of the project was to combine all the knowledge and techniques that were accumulated during this semester. We first began by selecting the dataset. We picked a dataset about Sleep Health from Kaggle website. We then proceeded to the EDA by first preparing the data (cleaning it, fixing the missing values). We then proceeded to the statistical analysis of the dataset. Several graphs helped us identify some of the key features of the dataset. The final part of the analysis was the machine learning that was performed on the dataset. We first ran the linear regression model to predict the quality of sleep, then logistic regression model to predict whether a person has a sleeping disorder based on several identifications. Overall, we are satisfied with the findings because we were successful in applying the techniques learned in class.

I. INTRODUCTION

The dataset was found on Kaggle, and can be accessed here.
This dataset contains sleep information and covers a wide range of variables related to sleep and health habits. It contains data on sleep duration quality, basic health and fitness information, and information on the study's participants (age and gender). We will be using logistic regression to predict whether an individual has a sleep disorder based on given information. We will then be using linear regression to predict the sleep quality of an individual. The variables are as follows:

-Person ID: An identifier for each individual participant.
-Gender: The gender of each person (Male or Female).
-Age: The age of each person in years.
-Occupation: The occupation of each person.
-Sleep Duration: The number of hours the person sleeps per day.
-Quality of Sleep: A subjective rating of the quality of sleep, from 1 to 10.
-Physical Activity Level: The number of minutes each person engages in physical activity daily.
-Stress Level: A subjective rating of the stress level experienced by the person, from 1 to 10.
-BMI Category: The BMI category of the person (Underweight, Normal, Overweight, Obese)
-Blood Pressure: The blood pressure of the person (systolic/diastolic)
-Heart Rate: The resting heart rate of the person in beats per minute (bpm)
-Daily Steps: The number of steps the person takes per day.
-Sleep Disorder: What sleep disorder the person has (No Disorder, Insomnia, Sleep Apnea)

II. BACKGROUND

The data is synthetic and made only for illustrative purposes by the author. The author is a machine learning engineer, and this dataset was only created to demonstrate machine learning techniques. Therefore, there was no particular question that the author was trying to answer.

III. EXPLORATORY ANALYSIS

This dataset contains 374 rows and 13 columns with various data types. A complete listing of the columns and data types is shown in Table 1.

Table 1: Data Types

Variable Name	*Data Type*
Person ID	int64/Integer
Gender	Object
Age	int64/Integer
Occupation	Object
Sleep Duration	float64/ nominal

Quality of Sleep	int64/Integer
Physical Activity Level	int64/Integer
Stress Level	int64/Integer
BMI Category	Object
Blood Pressure	Object
Heart Rate	int64/Integer
Daily Steps	int64/Integer
Sleep Disorder	Object

Figure 1 is a heatmap to show the correlation between all of the numeric variables in this dataset. A correlation heatmap visually displays the strength and direction of relationships between multiple variables in a dataset. Each cell in the heatmap represents the correlation coefficient between two variables, which ranges from -1 to +1. The closer the number to +1, the stronger the relationship between the variables. The closer the number to -1, the stronger the negative relationship between the variables.

Figure 1: Correlation Heatmap

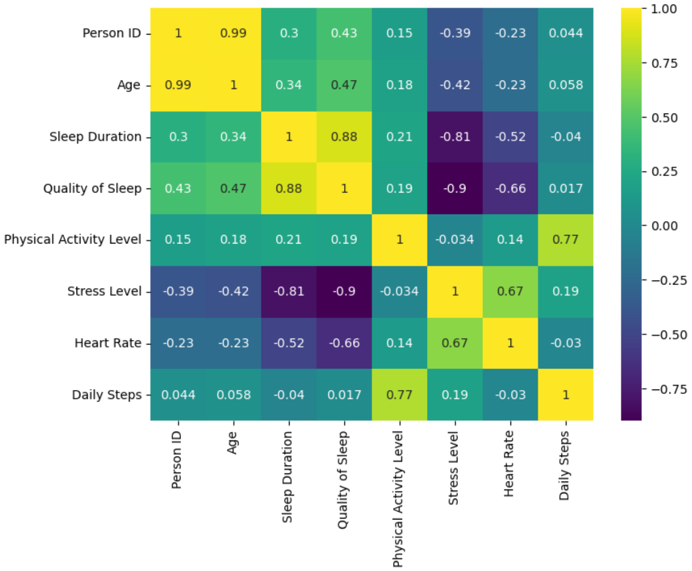


Figure 2 is a pie chart showing the distribution of sleep disorders in this dataset. Individuals are categorized into having either no disorder, sleep apnea, or insomnia. A pie chart is the best way to represent this information as we can see that the majority of individuals no not have a sleep disorder very clearly on this plot.

Figure 2: Sleep Disorder Pie Chart

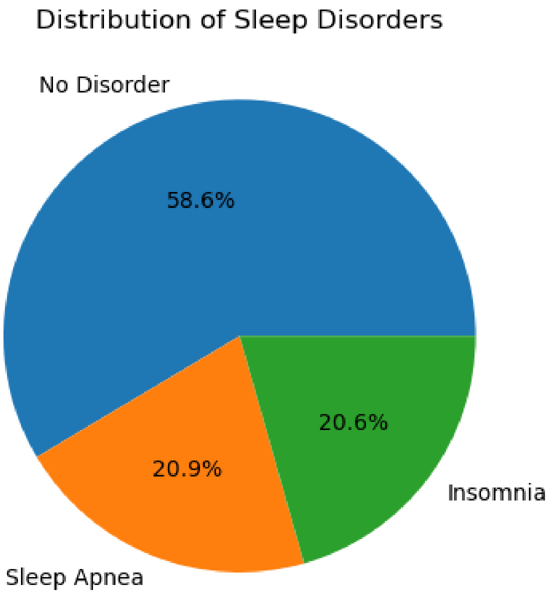


Figure 3 is a scatterplot that shows different qualities of sleep at different sleep durations. The X-axis shows the sleep duration in hours. The Y-axis shows the quality of sleep on a scale of 1-10. We can see a trend towards higher quality sleep when an individual sleeps for a longer duration.

Figure 3: Sleep Duration vs Quality of Sleep

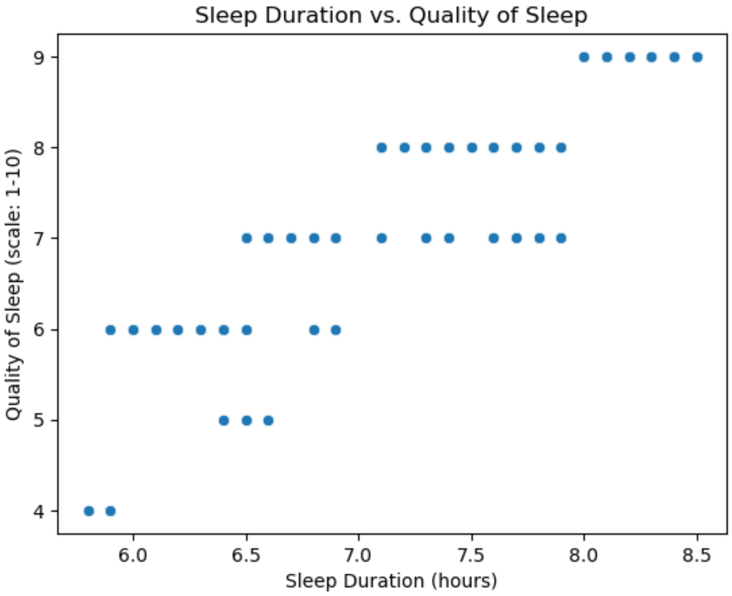


Figure 4 is a box plot that depicts the average quality of sleep for people of each sleep disorder category. There are outliers for each category, but this representation best reflects the average measures for each disorder category. Based on the box plot, people with no disorder have a higher average quality of sleep than those with insomnia. Surprisingly, people with sleep apnea have a wide range of sleep qualities, with some individuals having higher sleep quality than those without a disorder.

Figure 4: Sleep Quality by Disorders

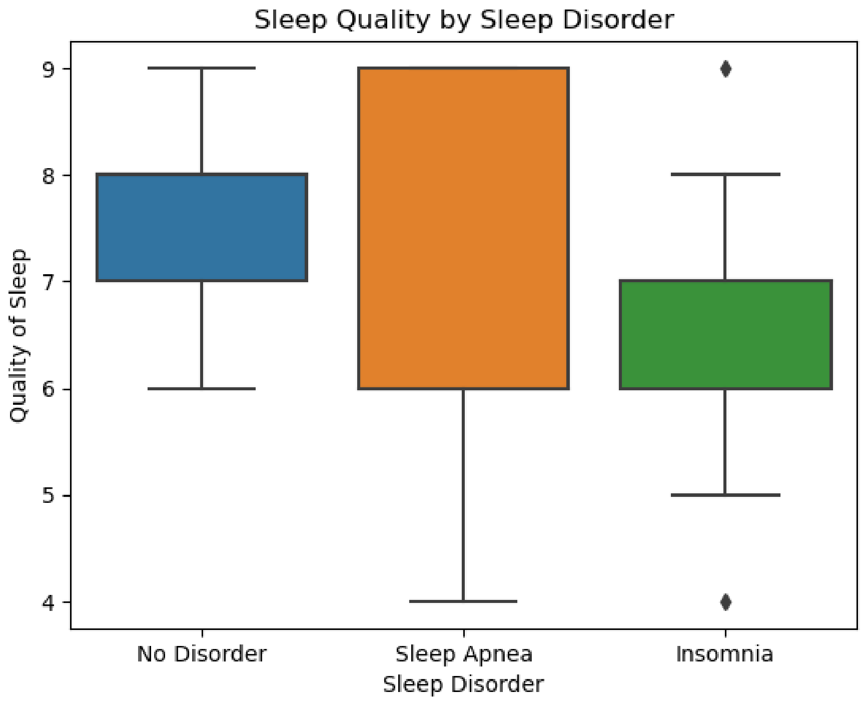
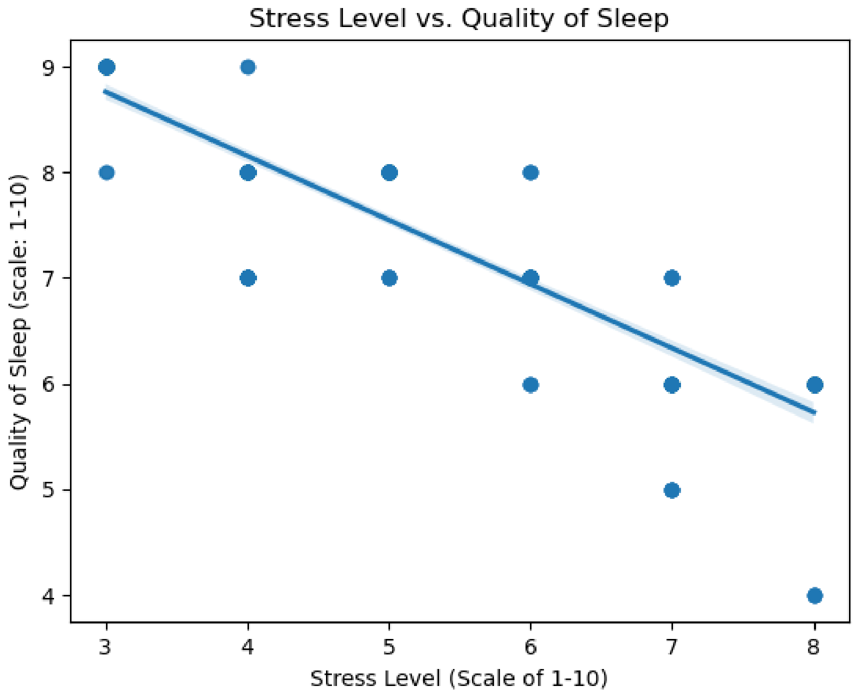


Figure 5 is a scatterplot that shows the relationship between stress levels and quality of sleep. Based on the included trend line, there is a correlation showing that people who are more stressed generally have worse quality of sleep.

Figure 5: Stress Level and Quality of Sleep



IV. METHODS

Next, we proceeded to machine learning.

A. Data Preparation

We first did a linear regression model.

First, the categorical variables such as Gender, BMI Category, and Sleep Disorder were one-hot encoded using the pd.get_dummies function. This transformed the categorical values into binary columns, allowing the model to interpret them numerically. The one-hot encoding included the drop_first=True parameter to avoid redundancy by dropping the first category for each variable.

Next, the features (X) were selected to include numerical and encoded categorical columns: Gender, Age, Sleep Duration, Physical Activity Level, Stress Level, BMI Category, Heart Rate, Daily Steps, and Sleep Disorder. The target variable (y) was set to Quality of Sleep. The dataset was then split into training and test sets using an 80-20 split with train_test_split to ensure the model could be evaluated on unseen data.

As for the logistic regression model.
For logistic regression, the primary change made to the data was encoding the Sleep Disorder column into a binary variable, where instances of "Sleep Apnea" and "Insomnia" were combined and mapped to 1 (Has a Disorder), while other values were mapped to 0 (No Disorder). Additionally, categorical variables like Gender and BMI Category were one-hot encoded to ensure compatibility with the model.

B. Experimental Design

In linear regression model:

Table X: Experiment Parameters

Experiment Number	**Parameters**
1	All four (4) raw features with 80/20 split for train, and test
2	All four (4) raw features with 70/30 split for train, and test

C. Tools Used

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1, and Patsy 0.41. Provide a brief explanation of why you chose these tools.

V. RESULTS

A. Classification Measures/ Accuracy measure

For linear regression model:
From the experiment one: Multiple Linear Regression Evaluation Metrics: (MSE:0.06544666728289363, RMSE:0.25582546253821886, R-squared:0.956618253185685).
From the experiment two: Multiple Linear Regression Evaluation Metrics: (MSE:0.07311061226367148, RMSE:0.2703897414172207, R-squared:0.7047598330187157).

For logistic regression model:
The logistic regression model demonstrates strong performance in predicting whether an individual has a sleeping disorder or not, achieving an overall accuracy of 92% on the test set. The confusion matrix indicates that the model correctly classified 40 out of 43 instances for the "No Disorder" category (0) and 29 out of 32 instances for the "Has a Disorder" category (1). There were 3 false positives (classified as "Has a Disorder" but actually "No Disorder") and 3 false negatives (classified as "No Disorder" but actually "Has a Disorder").

The classification report highlights high precision and recall for both classes. Precision for "No Disorder" (0) is 93%, meaning the model is very accurate when it predicts "No Disorder." Similarly, precision for "Has a Disorder" (1) is 91%, indicating that most predictions for this class are correct. Recall is 93% for "No Disorder" and 91% for "Has a Disorder," meaning the model successfully identifies the majority of instances for each category. The F1-scores for both classes are above 90%, reflecting a good balance between precision and recall.

B. Discussion of Results

In logistic regression, experiment two provides worse results than experiment 1. It is probably because the split was 80/20 in experiment 1, which suggests more precise results.

C. Problems Encountered

The problem that we have encountered with our data set is that we had to do a couple replacements of either missed values, or dictionary imputations to ensure that our data is ready for machine learning. It was challenging at times, but it was mostly extremely time consuming.

D. Limitations of Implementation

We consider that our data did a good job in predicting the target variable. With some data preparation we were ble to successfully predict what we intended to.

E. Improvements/Future Work

To improve the model in future work, I would explore adding more relevant variables, such as caffeine or alcohol consumption, sleep environment factors, or medical history, which may influence sleep disorders. Additionally, testing different models, such as Random Forest or Gradient Boosting, could provide better performance and handle complex interactions between features. Expanding the dataset to include more diverse demographic and health-related data would also help improve generalization and model accuracy. Finally, performing hyperparameter tuning and cross-validation could further optimize the model's performance.

VI. CONCLUSION

In conclusion, the Sleep Health and Lifestyle dataset is extensive and covers a few closely related categories – demographic, sleep, lifestyle, and health information. From the exploratory data analysis portion of this project, we can draw numerous assumptions based on the correlations shown. According to the correlation heatmap, Figure 1, many of the numerical variables have strong correlations, but the strongest relationship appears to be between stress level, duration of sleep, and quality of sleep. This correlation would suggest that the largest factor in how long someone sleeps or how well they sleep is based on their stress levels. This relationship is further proved by Figure 5, which shows a clear trend of lower sleep quality as stress levels increase. Figure 2 is beneficial in helping to understand the proportion of people with sleep disorders. Over 40% of people in this dataset have one of just two specific sleep disorders. Figure 3 suggests that sleeping for a longer time positively affects sleep quality. Figure 4 suggests that while there is a strong difference in sleep quality for people with or without insomnia, people with sleep apnea can have a very wide range of seep quality.

REFERENCES

Our data set was taken from the source below:

Laksika Tharmalingam. 2023. “Sleep Health and Lifestyle Dataset.” Kaggle.com. 2023. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.