DESCRIPTIVE ANALYSIS OF SLEEP TIME PREDICTION DATA

ST 3082 Data Analysis Project 1

Prepared by: Group 14

16305-Pasindu Gamage
16025-Viruna Fernando
16368-Poornima Tharangani

Abstract

- Sleep is a vital physiological process that plays a crucial role in maintaining physical health, mental well-being, and overall quality of life. This report explores the findings of the exploratory data analysis conducted on the Sleep time dataset obtained from Kaggle. The exploration utilizes statistical methods and visualization techniques to understand the relationship of various daily activities and how they influence sleep time.

Contents

List Of Figures

List of Tables

Introduction

❖ Sleep is an essential function that allows the body and mind to recharge, leaving individuals refreshed and ready to face the challenges of the day. Adequate sleep is crucial for maintaining a healthy life. Insufficient or excessive sleep can lead to a range of health issues. The general recommendation for sleep duration is 7 to 9 hours per night, but it may depend on the lifestyle. Achieving optimal sleep requires more than just allocating sufficient time, it involves understanding how daily activities and routines impact sleep quality.

❖ Our findings are important for identifying how daily activities influence sleep at the end of the day. By understanding these dynamics, it can provide recommendations to help individuals achieve a balanced and healthy sleep routine.

Description of the Question

- Sleeping duration is essential for physical health, cognitive function, and productivity. However, modern lifestyles shaped by technology, work, and personal habits often disrupt consistent sleep. Understanding the factors affecting sleep can help individuals adjust their routines to improve well-being.
- To optimize sleep time, our objectives are twofold:

    1. Identifying Key Factors Influencing Sleep Duration
    2. Predicting Sleep Duration Based on Lifestyle Patterns

- By using data-driven analysis, we explore key factors impacting sleep duration, providing insights for optimizing daily habits. Understanding sleep patterns helps improve time management, health awareness, and sleep quality.

Description of the Data Set

- The Sleep Time Prediction Dataset sourced from the Kaggle presents 7 variables on 2000 observations. Here is the description of each variable in the dataset.

Variable	Type	Description
Workout Time	Quantitative	Time spent exercising. (Hours/day)
Reading Time	Quantitative	Time spent reading. (Hours/day)
Phone Time	Quantitative	Time spent on the phone. (Hours/day)
Work Hours	Quantitative	Hours spent working. (Hours/day)
Caffeine Intake	Quantitative	Caffeine consumed daily (mg/day)
Relaxation Time	Quantitative	Time spent relaxing (e.g., meditation, leisure activities). (Hours/day)
Sleep Time	Quantitative	Total hours of sleep. (Hours/day)

Table 1: Variable Description

➢ Here is how the time was allocated for the activities by individuals that are highlighted in the dataset.

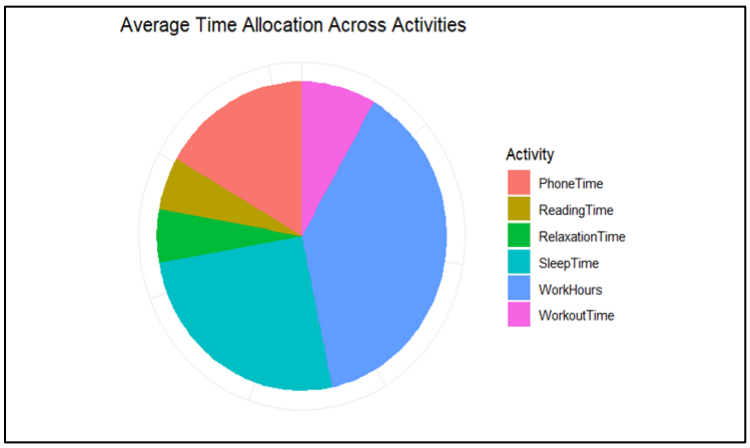

Figure 1: Pie Chart for Time Allocation

- Most individuals spend their day working and only a small amount of time has allocated for their relaxation and workouts. But significantly high time has been allocated to use the phones. We are interested in examining whether these activities have had a negative or positive impact on sleep time.

Data Preprocessing

- The data set was checked for the missing values and duplicates. But There was **not any value.**

- Handling unusual Data points: In the dataset, it has noticed that the features and target variable are independent. Usually for each observation, the total hours that are spent on each activity should be less than or equal to 24 hours (in a day). So, the observations that give the total hours for all time related variables (without caffeine Intake) as **more than 24 hours** are identified as unusual observations, and we removed them.

- Checked for Outliers: Outliers have been observed only for sleep time column. However, these outliers will not be removed as they may lead to inherent variability of the data.

Feature Engineering

- A new categorical variable, 'Sleep Pattern' was created by using the Sleep Time variable.

  - If Sleep Time < 7 Hours: Short Sleep
  - If 7<=Sleep Time <9: Optimal Sleep
  - If Sleep Time >=9: Long sleep

Note: The constraints are constructed using standards for sleep time for adults.

- Then the dataset was split into training and testing set. The training set contains 1565 observations, and the test set contains 391 observations. Descriptive analysis was conducted using the training set.

Results From Descriptive Analysis

The Distribution of The Response Variable: Sleep Time

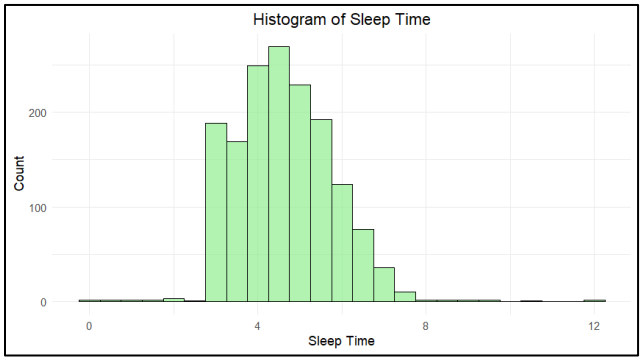

Figure 2: Histogram of Sleep Time

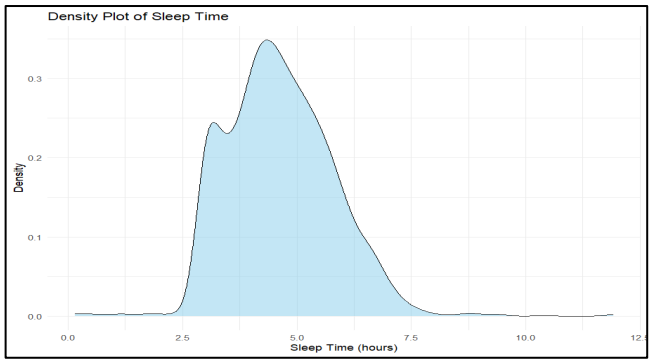

Figure 3: Density plot of Sleep Time

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	St. dev
0.15	3.82	4.55	4.631	5.38	11.92	1.176919

Table 2: summary statistics for sleep time

- From Figure 2 it is seen that the distribution of Sleep time among individuals is roughly symmetric with majority of the sleep time centered around 4-5 hours. Some outliers can be seen with sleep time is less than 1 hour and greater than 8 hours. From table 2, it is realized that the median and mean of the distribution of sleep time is approximately closer to each other so we can examine the distribution as relatively symmetric.

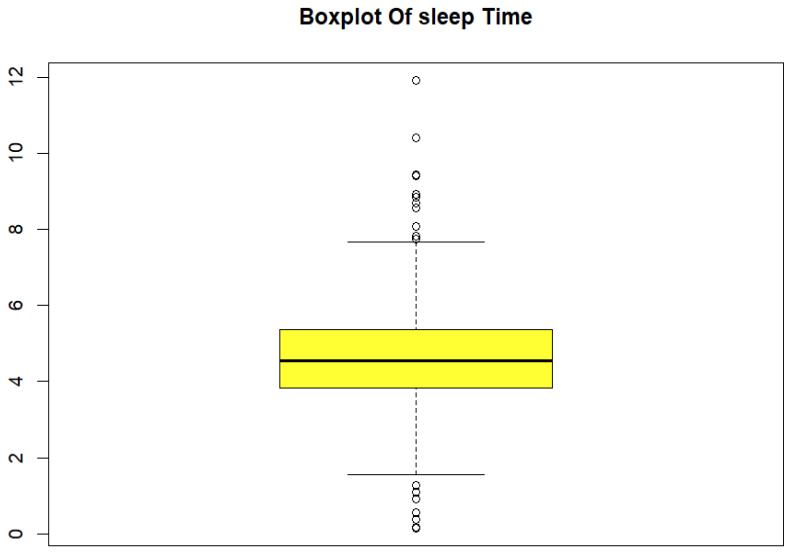

Figure 4: Boxplot of sleep time

- By Figure 3, the presence of outliers can be seen clearly as some points lie outside the interquartile range. These points are not removed from the data set as we have no reosens for the existence of these outliers.

Association Of Activities With Sleep Time

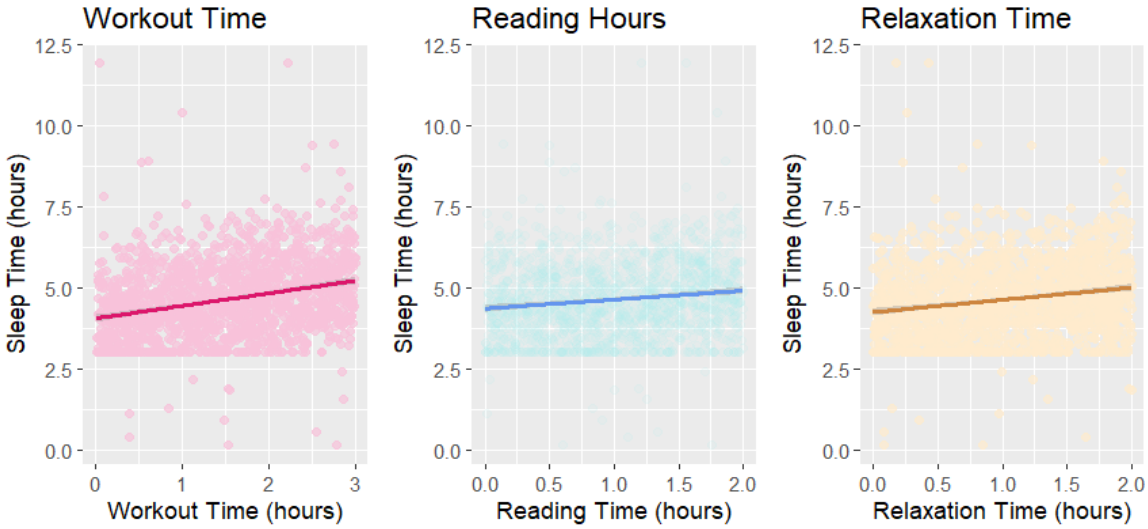

Figure 5: Scatter plot of Workout time , Reading Time and Relaxation Time

- In Figure 4 , The scatter plot shows a slight upward linear relationship between workout time and sleep time. Engaging in workouts may contribute to slightly better sleep duration, though the effect doesn't appear to be strong. There is a mild positive trend between Reading hours and sleep time such that reading time has a small positive impact on sleep duration. When compare the effect of relaxation time on sleep time the plot shows a moderate positive relationship, suggesting that relaxation activities appear to have a more noticeable positive effect on sleep time.

- All three factors exhibit a positive association with sleep time, but the magnitude of their impacts varies.Workout Time seems to have the most substantial effect on sleep time, than the two other factors , Reading time and Relaxation time. So It suggested that Workout , Reading and Relaxation make positive impact on Sleep Time.

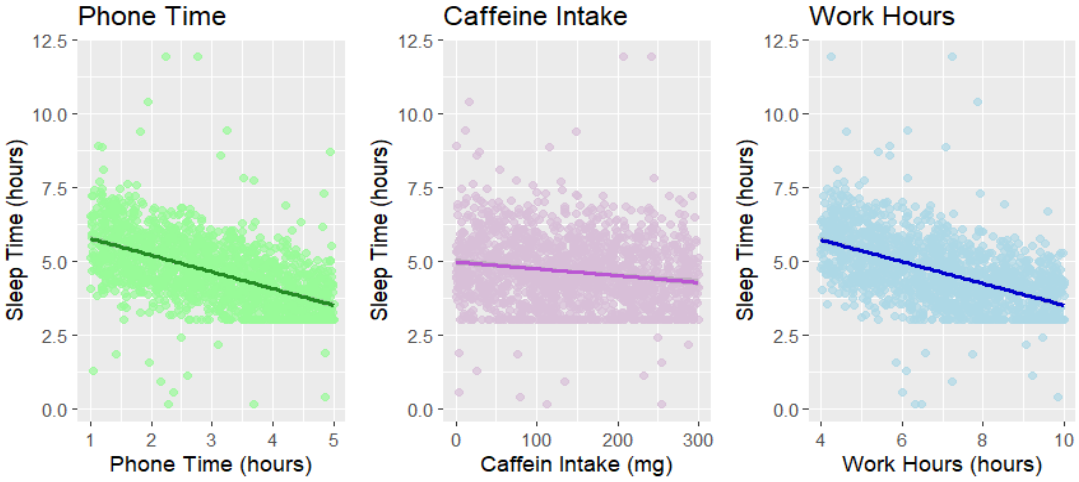

Figure 6: Scatterplot of Phone time , caffeine Intake and work hours

- The scatter plot shows a clear negative relationship between phone time and sleep time, ensuring that increased phone usage is associated with reduced sleep duration. Caffeine Intake vs Sleep time scatterplot shows a week negative slope. The scatter points are more spread, indicating a weak or negligible relationship. Higher caffeine intake might slightly reduce sleep time, but the effect is less compared to phone time or work hours. There is a strong negative relationship between work hours and sleep time. Longer work hours are associated with significantly reduced sleep time. This finding highlighted the importance of work-life balance while maintaining adequate sleep.

- All three factors show a negative impact on sleep time. But the impact of phone time and work hours are more significant than the caffeine intake.

Bivariate Analysis Between Sleep Time and Lifestyle

- Here is a bivariate analysis on how the activities are distributed over three sleep patterns Short Sleep, Optimal Sleep and Long Sleep.

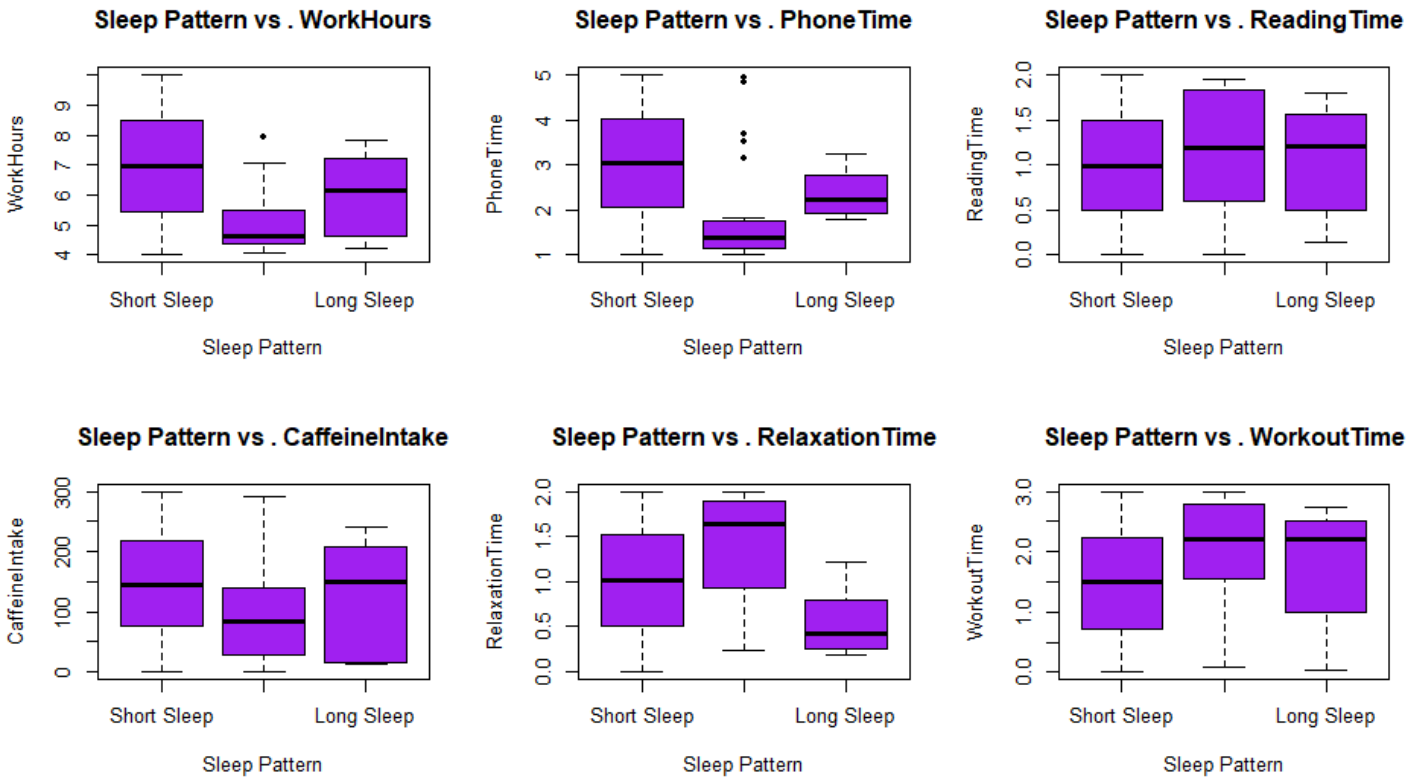

Figure 7: Boxplot of activities on sleep time

- It is Observed that for the sleep time less than 7 hours (Short sleep), high work hours, high phone time, low relaxation time has affected significantly than other factors. High Relaxation time, high work out time, less phone time and less work hours have increased the sleep time.

- Work hours, phone usage, and caffeine intake are associated with shorter sleep durations. Reading time, relaxation time, and workout time are associated with longer sleep durations.

- So the results gained from the bivariate analysis on sleep patterns and the lifestyle further ensure the results obtained by Figure 4 and Figure 5.

- For a optimal sleep it has suggested that Phone usage , Work out shedules , working hours per day be more concerned than other factor while balancing the caffeine intake per day.

Correlation Among Numerical Variables

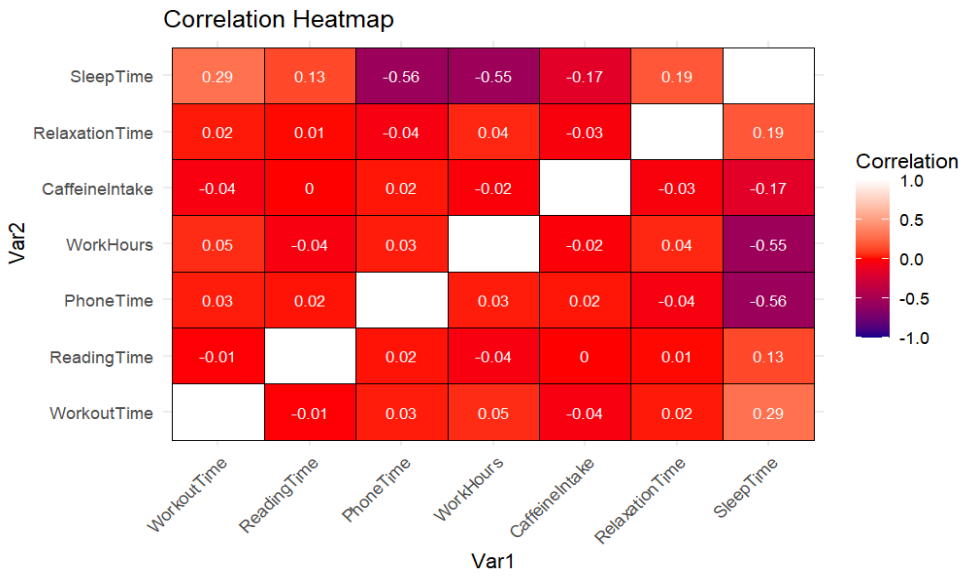

Figure 8: Correlation Heatmap on numeric variables

- The correlation heatmap represents the linear relationships between the numerical variables in our study. The response variable, Sleep Time, exhibits a moderately negative relationship with Phone Time (-0.56) and Work Hours (-0.55), indicating that increased phone usage and longer work hours are associated with reduced sleep.

- On the other hand, Workout Time shows a moderate positive relationship (0.29) with Sleep Time, suggesting that increased workout time may contribute to better sleep. The remaining predictor variables demonstrate weaker linear relationships with Sleep Time, implying a lesser direct impact.

- Additionally, it is evident that the predictor variables have very weak correlations among themselves, indicating low level of multicollinearity within them.

Further Analysis

Principle Component Analysis

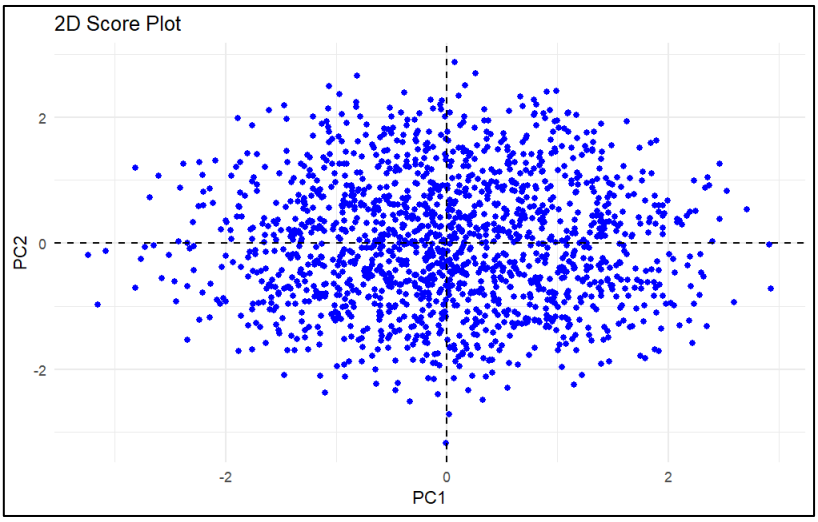

Figure 9: 2D score plot

- Principal Component Analysis was conducted on data to identify potential clusters among the observations and to visualize the presence of outliers. Before applying PCA, the categorical variable was removed from the dataset, as PCA exclusively works with quantitative data. To present the data in a two-dimensional plane, the dimensionality of x was reduced to two by selecting only two principal components x space. Consequently, the accuracy of the score plot obtained by PCA becomes questionable. Additionally, the score plot, illustrated above, provides a preliminary indication that there are no significant clusters in the data set.

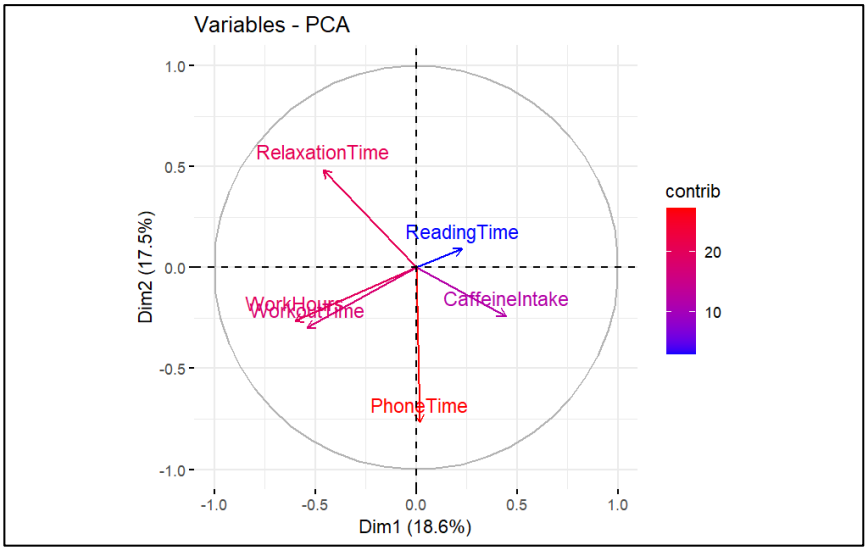

Figure 10: Loading plot of variables

- The loading plot illustrates the relationship between the original variables and the principal components (PCs), where the length of each arrow indicates the contribution of a variable to each PC. The angle between vectors and the principal component axes reflects the correlation between variables and PCs, with smaller angles indicating stronger correlations.

- **Phone Time and Relaxation Time** are strongly correlated with **PC1**, whereas **Reading Time and Caffeine Intake** are more associated with **PC2**. Additionally, the close alignment of **Phone Time and Work Hours** suggests a positive correlation, and **Reading Time and Caffeine Intake** appear somewhat related. However, the accuracy of the loading plot is questionable since the first two principal components explain only a limited proportion of variance (18.6% and 17.5%, respectively), indicating that additional dimensions may be necessary to capture more information from the dataset.

Suggestions For Advanced Analysis

Variable	VIF Value
Workout Time	1.005687
Reading Time	1.002164
Phone Time	1.004021
Workhours	1.007778
Caffeine Intake	1.003469
Relaxation Time	1.005534

Table 3: VIF Values

- Given that VIF values are less than 10, indicate little to no multicollinearity, methods like **PCA regression, Ridge, LASSO, and Elastic Net are unnecessary**.

- Since some predictors show moderate correlation with the response variable, **Multiple Linear Regression (MLR)** remains a strong baseline model.

- However, the weak correlations between some predictors and the response suggest the potential presence of **nonlinear patterns**, making advanced machine learning methods suitable. Therefore, for predictive modeling, **XGBoost, Random Forest, Support Vector Machines (SVM), and Regression Trees** can be explored to capture complex relationships.

Appendix

- **Link for the dataset:** Sleep Time Prediction

- **The R code used in our project is conveniently accessible through our GitHub repository:**

  https://github.com/pasindugamage2001/Data-Analysis-Project-1/blob/main/Project1.R