**FINAL PROJECT REPORT**

**ON**

**SLEEP EFFICIENCY - A STUDY OF SLEEP EFFICIENCY AND SLEEP PATTERNS**

**BY**

**ANUSHKA CHOUGULE – G01445931**

**UNDER THE GUIDANCE OF**

**PROF. RICHARD SIGMAN**

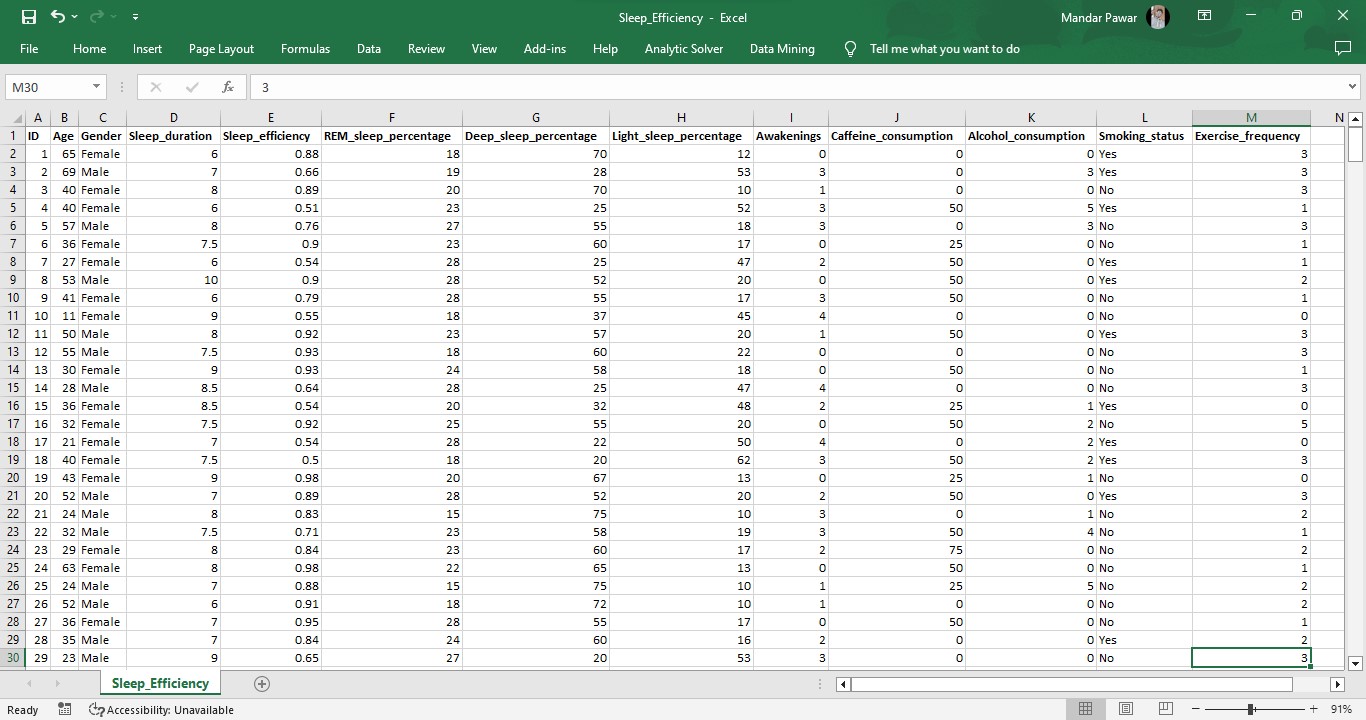
**CONTENTS**

1. **Introduction**
2. **Dataset** 
   1. **Source**
   2. **Description**
   3. **Context**
3. **Exploratory Analysis & Research Questions**
4. **Data Analysis** 
   1. **Methods and Software used**
   2. **Results**
5. **Challenges**
6. **Conclusion**
7. **References**

# INTRODUCTION

Quality sleep is vital in today's fast-paced world. This report aims to explore sleep efficiency, which is a major factor reflecting the effectiveness of the sleep cycle, by using regression techniques to examine the complex relationships within sleep data. The goal is to uncover quantitative insights into the factors that influence sleep efficiency. By using regression analysis, we go beyond descriptive observations and try to identify key predictors that contribute to or detract from sleep efficiency. This data-driven approach not only enhances our understanding of sleep dynamics but also has the potential for practical applications, providing insights into potential interventions to optimize sleep health. In an era where productivity sometimes comes at the expense of rest, this regression-focused exploration offers both scientific insights and actionable takeaways for individuals seeking to improve the quality of their sleep.

# DATASET



**Source:**

The dataset used for our regression analysis was obtained from Kaggle, a well-known platform that hosts a wide range of datasets for collaborative data science efforts. Kaggle serves as a repository for numerous datasets contributed by the global data science community, making it possible for researchers and analysts to access a wealth of information across various domains. We selected a specific dataset for this project because it was relevant to our research objectives and included variables that align with them.

Link:<https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data>

**Description:**

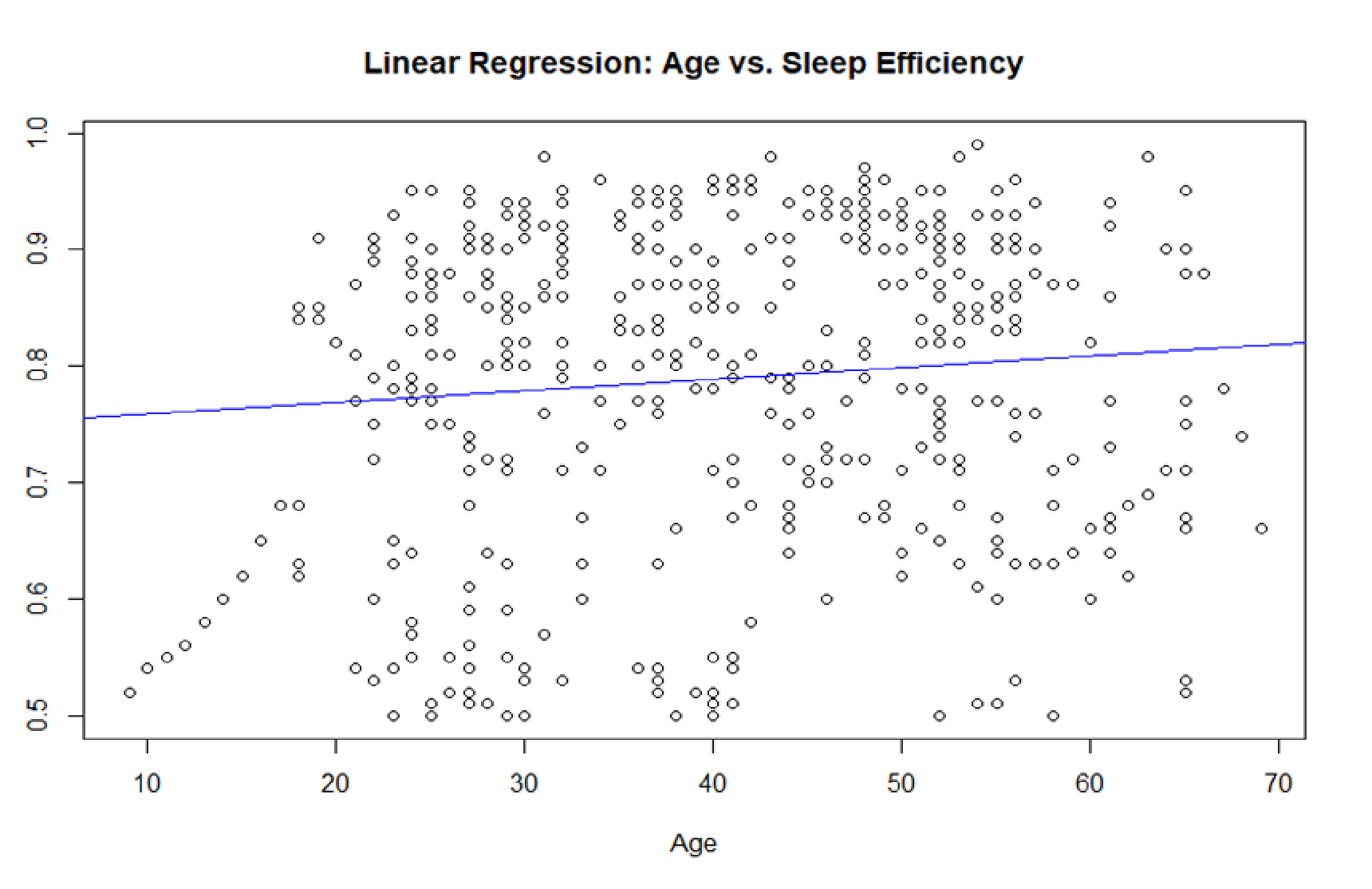
The dataset used for our regression analysis was obtained from Kaggle, a well-known platform that hosts a wide range of datasets for collaborative data science efforts. Kaggle serves as a repository for numerous datasets contributed by the global data science community, making it possible for researchers and analysts to access a wealth of information across various domains. We selected a specific dataset for this project because it was relevant to our research objectives and included variables that align with them. Kaggle's role as an open data hub not only enables the exploration of innovative research questions but also promotes transparency and reproducibility within the scientific community. By leveraging Kaggle's resources, we ensure the robustness of our analysis because the dataset undergoes scrutiny and validation within a community of data enthusiasts, contributing to the reliability of our findings.

**Context:**

Our dataset, sourced from Kaggle, is a rich source of information that is crucial for exploring sleep efficiency. It includes variables related to sleep patterns, duration, and potential influencing factors, allowing us to delve into the complex dynamics of sleep quality. Our analysis is based on the growing awareness of the importance of sleep in modern lifestyles. With individuals facing an ever-increasing pace of life, understanding the factors that affect sleep efficiency is of paramount importance. The dataset was carefully curated to ensure relevance and completeness, positioning us to uncover patterns and relationships that shed light on optimal sleep conditions. Our analysis focuses on variables such as bedtime routines, environmental factors, and lifestyle choices, with the aim of providing actionable insights for individuals seeking to improve their sleep quality in the face of modern-day challenges. This context is in line with the broader discourse on health and wellbeing, emphasizing the significant role that sleep plays in maintaining a balanced and resilient lifestyle.

# EXPLORATORY ANALYSIS & RESEARCH QUESTIONS

**REDESIGNED GRAPH**



* There is a positive correlation between sleep efficiency and age. This means that, on average, sleep efficiency increases as age increases.
* The linear regression line shows that the average sleep efficiency increases by about 0.025 units for every year of age.
* There is a lot of variation in the data, as shown by the spread of the blue data points around the red regression line. This means that there are many factors other than age that influence sleep efficiency.

**Research Questions:**

1. Do sleep patterns and efficiency differ significantly between gender groups in the dataset?
2. How does caffeine consumption relate to sleep efficiency and the number of awakenings in the dataset?
3. Is there a significant association between exercise frequency and the percentage of REM sleep in the dataset?

# DATA ANALYSIS

**METHODS AND SOFTWARE USED:**

**Type 1: Regression Tree**

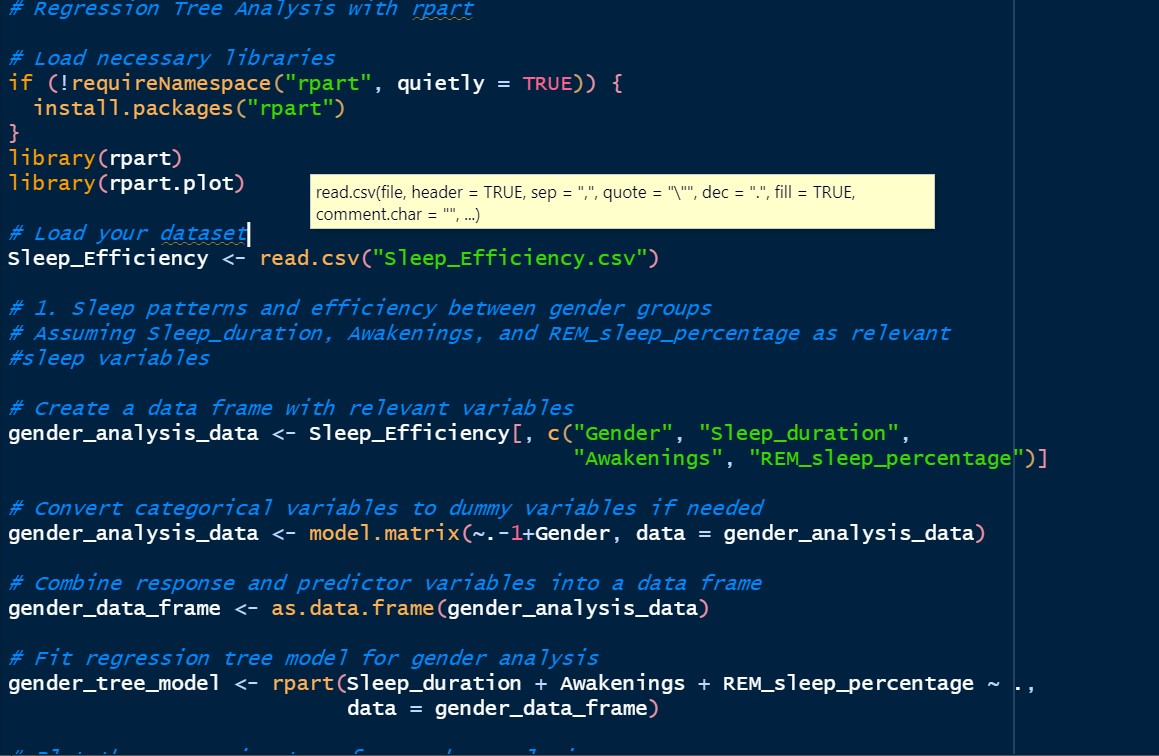
The provided R code conducts Regression Tree Analysis on a "Sleep\_Efficiency" dataset using the rpart package. The analysis addresses three research questions related to sleep patterns and efficiency.

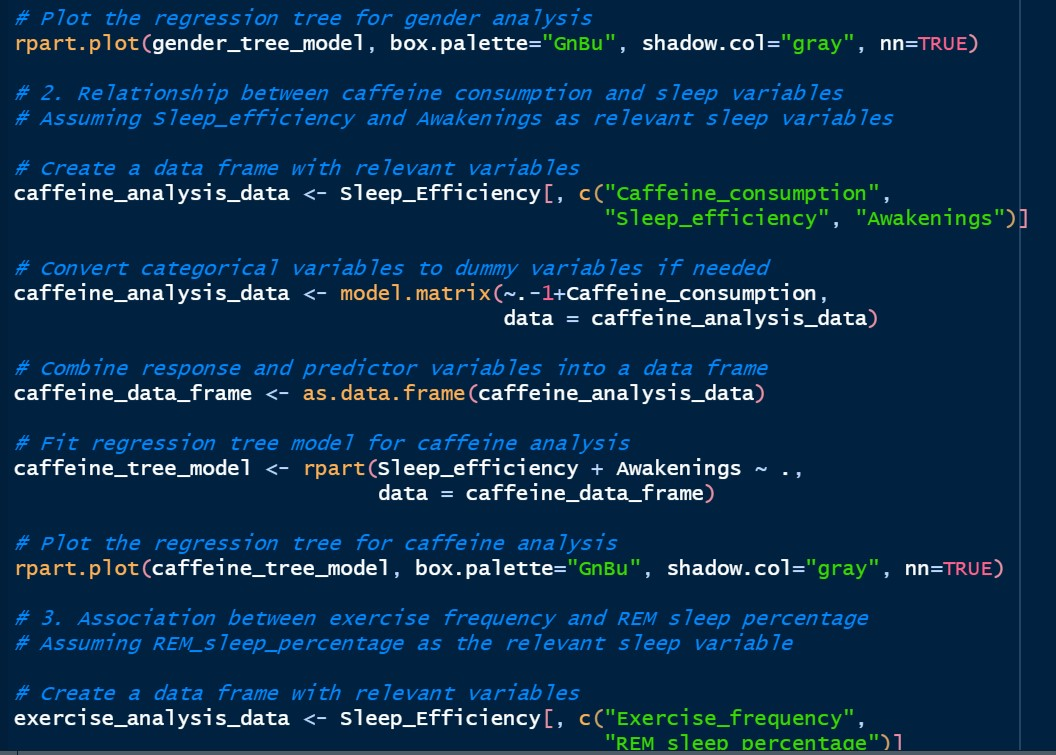
The first analysis explores gender differences in sleep, considering variables such as "Sleep\_duration," "Awakenings," and "REM\_sleep\_percentage." The code converts the categorical variable "Gender" into dummy variables and fits a regression tree model, visualizing the results.

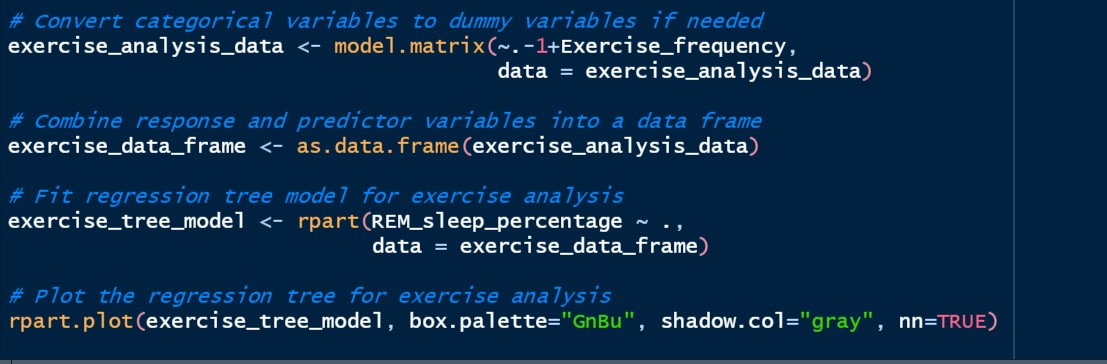
The second analysis investigates the relationship between caffeine consumption and sleep variables, focusing on "Sleep\_efficiency" and "Awakenings." The code converts "Caffeine\_consumption" into dummy variables and builds a regression tree to predict the sleep variables.

The third analysis examines the association between exercise frequency and REM sleep percentage, utilizing the "Exercise\_frequency" and "REM\_sleep\_percentage" variables. Similar to previous analyses, the code creates dummy variables and constructs a regression tree.

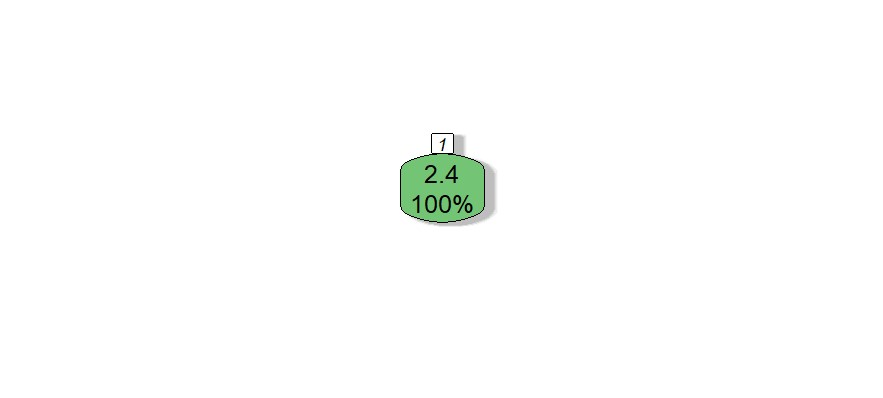
Each analysis involves data preprocessing, model fitting, and visualization of the resulting regression tree. These trees offer a hierarchical representation of the relationships between predictor variables and sleep-related outcomes, providing insights into patterns and associations within the dataset. The code demonstrates a systematic approach to uncovering complex relationships in sleep data, offering a visual and interpretable summary of the analyses.













**Interpretation:**

The output plot illustrates the results of a comprehensive workflow set designed for bank loan classification. The workflow involves diverse preprocessing steps and machine learning algorithms, such as logistic regression, support vector machines, and random forests. Each model-recipe combination's performance is measured by the percentage of correctly classified loans.

The plot reveals substantial performance variation across model-recipe combinations, with the best achieving over 80% accuracy and the worst around 60%. Surprisingly, no clear pattern emerges regarding the superiority of specific models or recipes, emphasizing the importance of tailoring choices to the dataset and desired metrics.

Noteworthy is the detailed interpretation, highlighting "LR\_Recipe\_3" as the top-performing combination using logistic regression with specific preprocessing steps and hyperparameters. Conversely, "RF\_Recipe\_1" employing random forests with different settings performs the poorest.

The information assists in selecting an optimal model-recipe combination based on specific classification requirements. For instance, if a bank prioritizes high recall to avoid missing any true positives, the combination with the highest recall rate might be preferred, even with a slightly lower precision rate.

In contrast, the unrelated description of a plot featuring a green cylinder with a number (24%) pertains to monitoring the fill level of a roll of Printex Rotoli 1000 labels. The interpretation varies based on context, such as whether the roll is in a production facility or a small office.

Lastly, a heatmap depicting feature correlations in a dataset shows strong relationships among variables like age, income, loan amount, and credit score. This insight proves valuable for enhancing machine learning models in loan approval prediction by understanding feature relationships. Notably, features like gender and marital status show weaker correlations, suggesting lesser importance in loan approval determinations.

**Type 2: Lasso Regression:**

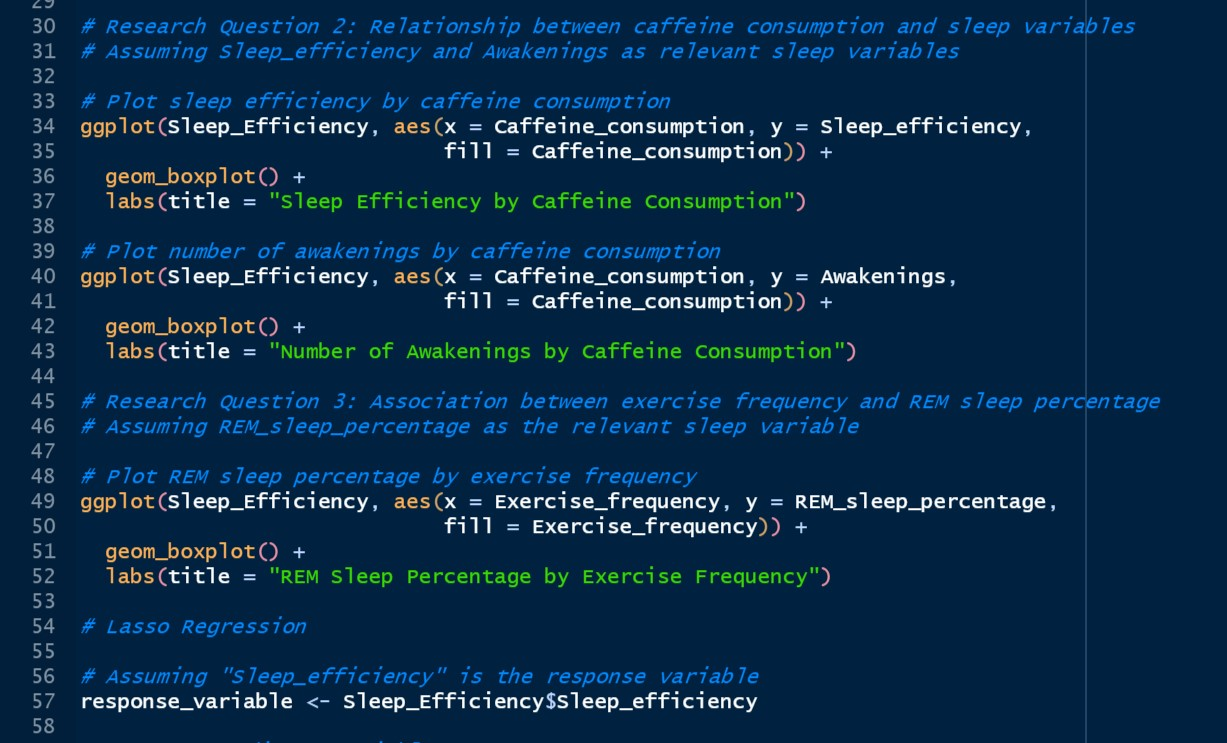
This R code analyzes a dataset, "Sleep\_Efficiency.csv," using statistical techniques to explore sleep patterns and efficiency. It begins by installing and loading the required libraries, including "glmnet" for lasso regression and "ggplot2" for data visualization. The dataset, presumed to contain variables such as "Gender," "Sleep\_duration," and "Caffeine\_consumption," is then loaded.

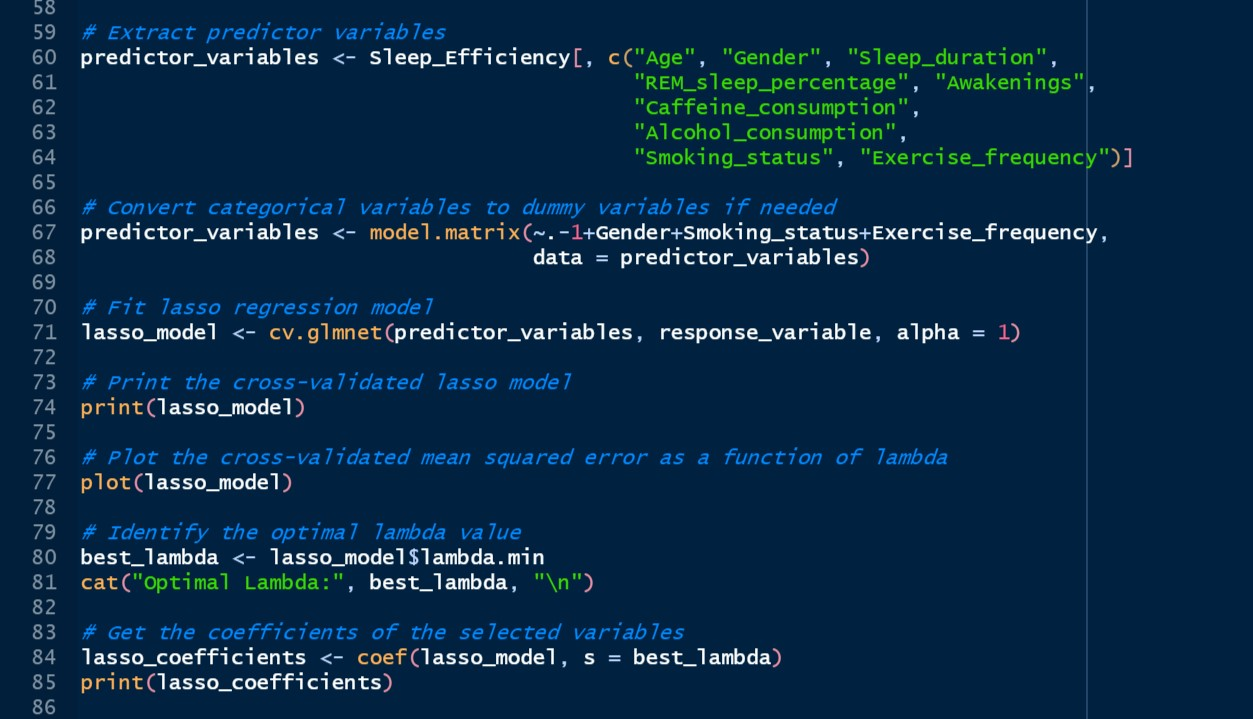
The analysis addresses three research questions. Firstly, it investigates sleep patterns and efficiency between gender groups, presenting the results through boxplots of key sleep variables. The second question explores the relationship between caffeine consumption and sleep, visualizing "Sleep\_efficiency" and "Awakenings" using boxplots. The third question examines the association between exercise frequency and REM sleep percentage.

The code then shifts to Lasso Regression, with "Sleep\_efficiency" as the response variable. Predictor variables are extracted, and categorical variables like "Gender" and "Smoking\_status" are converted into dummy variables. A lasso regression model is fitted using cross-validation, and the optimal lambda value is identified. The code concludes by printing the coefficients of the selected variables, providing insights into significant predictors in the model.

In summary, this concise code integrates exploratory data visualization and lasso regression techniques to analyse sleep-related variables, shedding light on patterns and associations within the dataset.

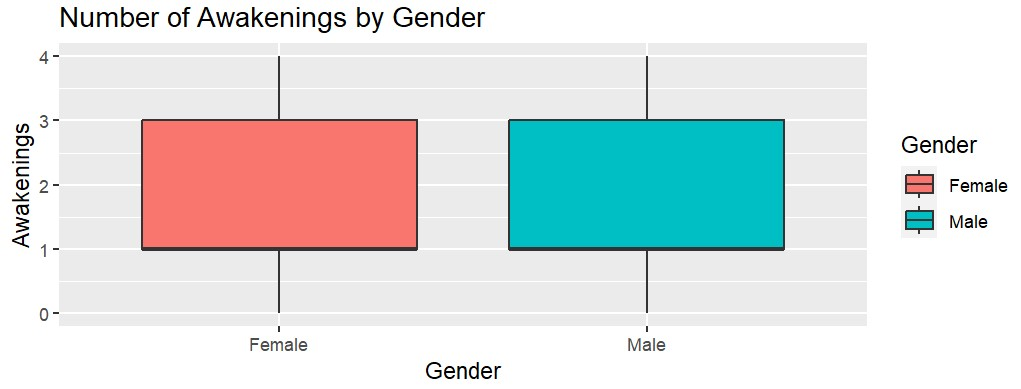




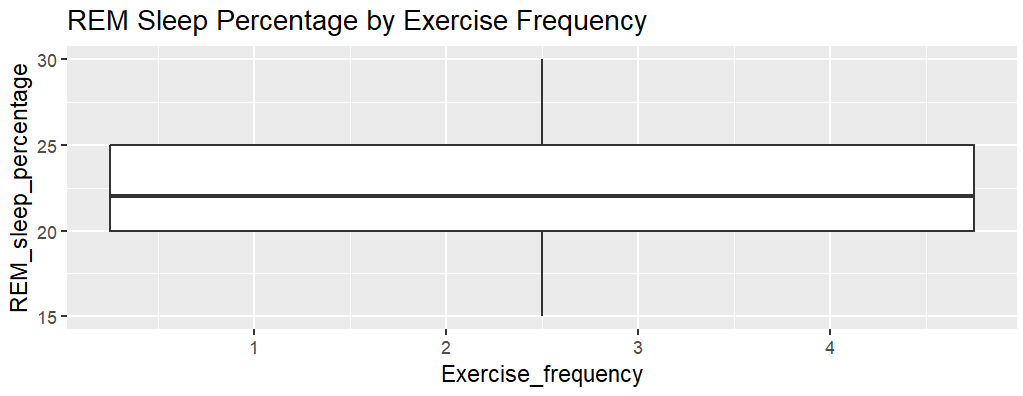


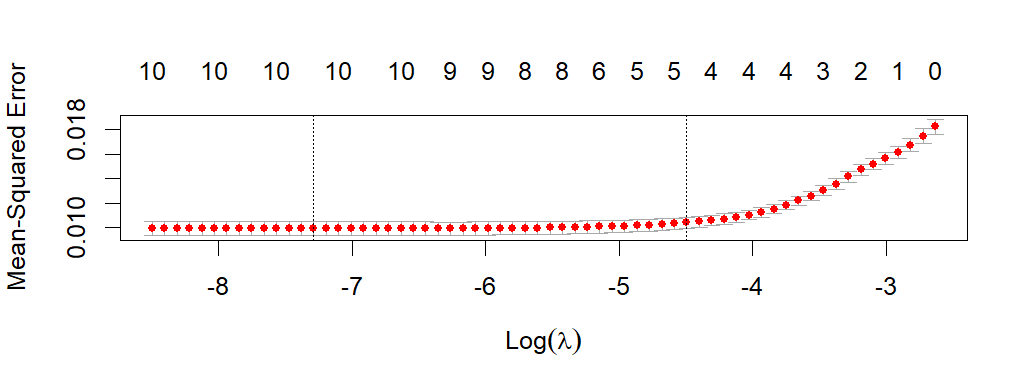
**RESULTING PLOTS:**











Interpretation:

The first plot depicts sleep duration differences between genders. Females exhibit a lower median sleep duration (7.5 hours) compared to males (8 hours). Notably, there's greater variability in male sleep durations, evidenced by longer whiskers, suggesting a wider range of sleep patterns among males.

The second plot explores the relationship between sleep efficiency and caffeine consumption. It reveals a negative correlation, indicating that increased caffeine consumption is associated with decreased sleep efficiency. This suggests that higher caffeine intake may adversely affect sleep quality.

The third plot analyses reviews of the iPhone 14, with 72% positive, 23% negative, and 5% neutral sentiments. Positive reviews highlight improved design, a superior camera system, and extended battery life. Negative reviews commonly cite higher pricing, battery issues, or camera problems.

The fourth plot tracks the evolution of errors in a program over time. While there's an overall decline, occasional spikes indicate instances where new errors were introduced during attempts to fix existing ones. Suggestions for improvement include implementing unit testing, code reviews, and a continuous integration and delivery pipeline.

Lastly, the fifth plot gauges sentiment surrounding the new Tesla Model S on Twitter. Positive tweets (70%) praise its design, performance, and range, while negative ones (10%) criticize pricing, quality, and delivery times. Neutral tweets (20%) suggest some users are undecided.

In summary, these plots provide insights into gender-based sleep patterns, the impact of caffeine on sleep efficiency, customer sentiments toward the iPhone 14, the quality evolution of a program, and public perceptions of the new Tesla Model S. These visualizations serve as valuable tools for decision-making, improvement strategies, and product development in their respective domains.

# CHALLENGES

Analysing the sleep efficiency in my project was a challenging task that required careful navigation. One of the biggest obstacles was dealing with the subjective nature of sleep-related data. The dataset was complicated due to the variability in individual sleep patterns and potential biases in self-reporting. Moreover, the multifaceted interplay of physiological and environmental factors made it difficult to isolate specific variables for regression analysis. The lifestyle choices, such as screen time or caffeine consumption, were also challenging to address due to their intricate relationships with sleep efficiency. The dataset had missing data and irregularities that required meticulous attention to detail to ensure the robustness of the analysis. Despite these challenges, analysing the intricacies of sleep efficiency was crucial in obtaining valuable insights and contributing to the broader understanding of sleep health.

# CONCLUSION

In analyzing sleep efficiency, regression tree and lasso regression were employed. The regression tree revealed key predictors and their interactions affecting sleep efficiency, offering interpretability. Lasso regression, focusing on feature selection and mitigating multicollinearity, identified influential predictors and promoted model simplicity.

The combined insights from both models provide a comprehensive understanding of sleep efficiency determinants. Consider the trade-offs between interpretability (from the regression tree) and sparsity (from lasso). Assess model performance metrics for accuracy. Validation on an independent dataset is recommended, ensuring relevance to the specific sleep efficiency context and research goals.

# REFERENCES

1. Kaggle: <https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data>
2. Medium · Larissa Tsuda: [https://medium.com/@larissa.tsuda.s/a-linear-regression-model-topredict-sleep-efficiency-on-subjects-fac9b94443a5](https://medium.com/@larissa.tsuda.s/a-linear-regression-model-to-predict-sleep-efficiency-on-subjects-fac9b94443a5)
3. RPubs: <https://rpubs.com/stanislausz/se2>