PROJECT REPORT - PREDICTING SLEEP EFFICIENCY

Qi Wu (72870371), Ayush Joshi (20443560), Hung Dinh (19774520)

Introduction:

Our analysis is based on the sleep efficiency dataset available on Kaggle at https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency, comprising 100 entries. This dataset was compiled in 2021 by a team of researchers in the United Kingdom, specifically associated with the University of Oxfordshire. The data gathering spanned several months and involved participants from the surrounding community. A blend of methodologies was employed to ensure comprehensive data collection, including self-reported surveys, actigraphy, and polysomnography, a technique for sleep monitoring.

Motivation behind the analysis:

Sleep is essential to our physical and mental health, yet it often takes a back seat in today's fast-paced world, with many prioritizing their careers or education over sleep quality. Numerous studies, such as the one by Kim et al. (2022), underline the health hazards linked to poor sleep. Our analysis aims to identify and understand the factors influencing sleep efficiency and their interconnections. This understanding could enable more informed choices to enhance overall sleep quality.

The research question is:

- **Which predictor play the most significant role in predicting sleep efficiency**?

Variables included in the analysis:

There are 15 variables in total in the dataset, we first select all the useful variables, then separate them into response variables and explanatory variables. All explanatory variables are potential predictors in predicting sleep efficiency.

The response variable:

1. Sleep efficiency: a numerical variable ranging from 0 to 1, measures the proportion of time in bed spent asleep.

The explanatory variables:

1. Age: a numerical variable ranging from 6 to 69.
2. Gender: a categorical variable consists of 50% male and 50% female.
3. Sleep duration: a numerical variable ranging from 5 to 10, measures the total amount of time the test subject slept (in hours)

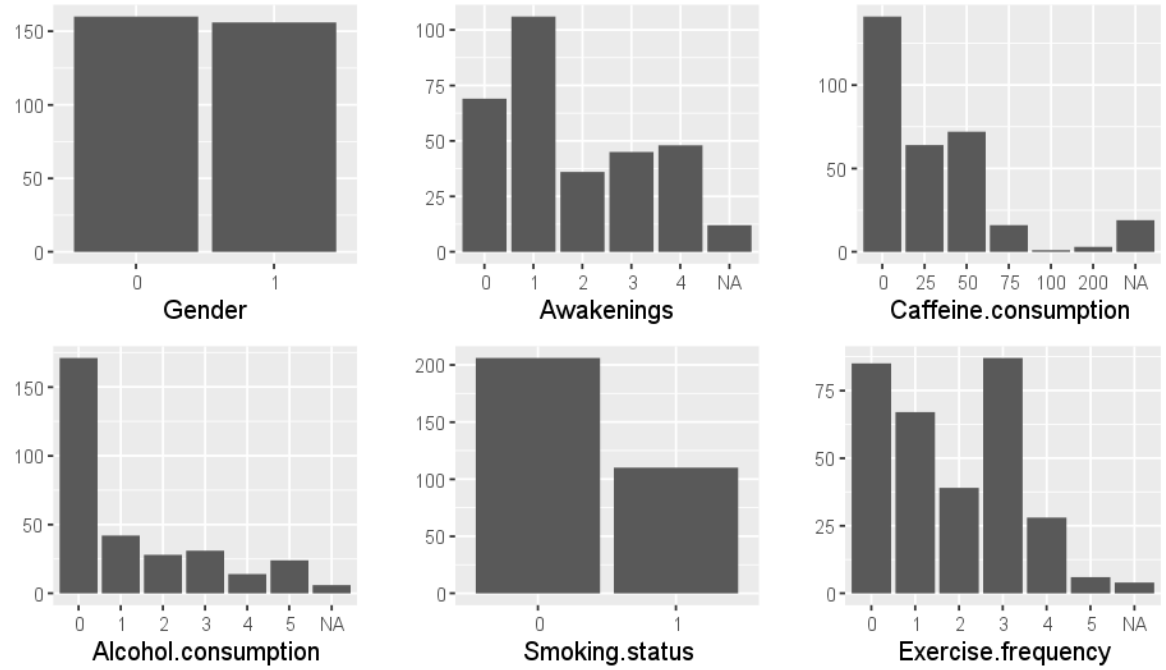
***Due to the variable "Sleep duration" = "Wakeup time" - "Bedtime", we skip the variable "Wakeup time" and "Bedtime".*

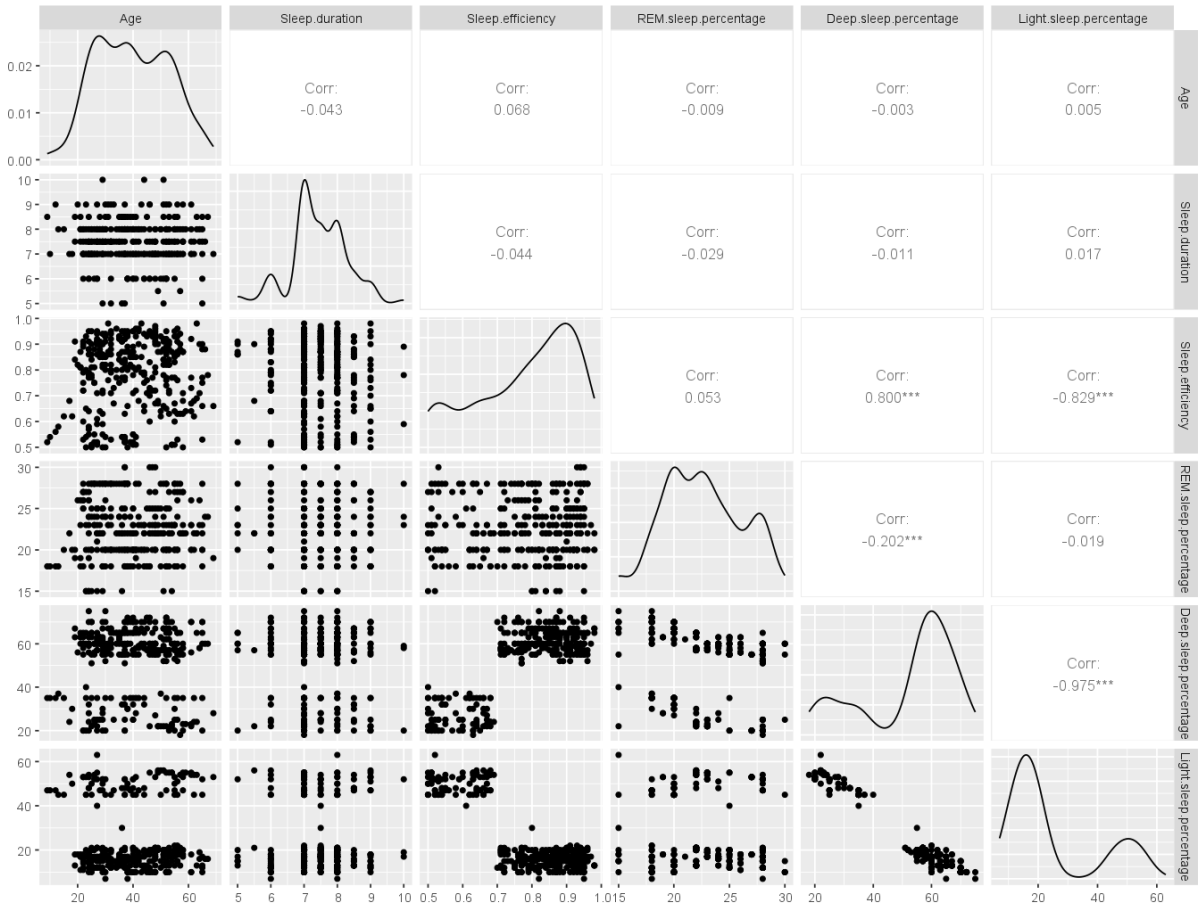
4. REM sleep percentage: a numerical variable measures the percentage of total sleep time spent in REM sleep.
   ***REM(Rapid eye movement) sleep: is the stage of sleep where most dreams happen. In REM sleep, people's brain activity is similar to when people are awake.*
5. Deep sleep percentage: a numerical variable measures the percentage of total sleep time spent in deep sleep.
6. Light sleep percentage: a numerical variable measures the percentage of total sleep time spent in light sleep.
7. Awakenings: a categorical variable measures the number of times the test subject wakes up during the night.
8. Caffeine consumption: a categorical variable measures the amount of caffeine consumed in the 24 hours prior to bedtime (in mg).
9. Alcohol consumption: a categorical variable measures the amount of alcohol consumed in the 24 hours prior to bedtime (in oz).
10. Smoking status: a categorical variable measures whether or not the test subject smokes.
11. Exercise frequency: a categorical variable measures the number of times the test subject exercises each week.

While certain variables: Alcohol and Caffeine consumption; Awakenings; Exercise frequency appear to be numerical in nature, the limited range of values they exhibit allows us to categorize and transform them into categorical variables with a few distinct levels.

Exploratory Data Analysis:

We have split the dataset into training and testing sets using 70% data for training. The EDA is performed on the training set and reveals certain properties about our categorical predictors. The gender ratio is overall very balanced (roughly 1) while about twice as many people reported that they do not smoke compared to those who smoke. All other predictor distributions seem to show some skewness with right tails (most apparent in alcohol and caffeine consumption) and exercise distribution is also bimodal.





We also use ggpairs() for all possible numeric predictors, which reveals possible collinearity between deep sleep and light sleep percentage since the absolute value of correlation coefficient is very high, of 0.975, which means that including both predictors can make our model unstable and result in lower accuracy. For this reason, we will only include deep sleep

percentage as one of the predictors. Besides that, all other variables seem to show no significant correlation with each other, and deep sleep (as well as light sleep) seems to be showing high correlation (0.8) with the response variable, sleep efficiency.

Model Building

We began the model building process by first fitting a model with all predictor variables except for Light.sleep.percentage, since it is highly correlated with Deep.sleep.percentage as seen in the EDA. This is the full model. Next, we narrowed down the predictor variables to Age, Deep.sleep.percentage, REM.sleep.percentage, Smoking.status, Awakenings, and Exercise.Frequency. We found these variables more significant than the others when predicting Sleep.efficiency.

We then tried to add some interaction terms in the form of two different models, one with Deep.sleep.percentage*Smoking.status and the other with Smoking.status*Sleep.duration. We wanted to explore how smoking affected different aspects of sleep as smoking was found to be associated with insomnia and reduced sleeping duration (PhD, A. N., et al., 2021). We found the model with the first interaction term to be the best. Below is a table with the AIC, BIC, adjusted R_sq, and RMSE scores for each of the models we tried:

model	AIC	BIC	adj R^2	RMSE
selective model with Smoking.status*Deep.sleep.percentage	-870.1925	-810.932	0.8418	0.05519
selective model with Smoking.status*Sleep.duration	-844.878	-781.9137	0.8285	0.05748
Selective model	-847.046	-791.4893	0.8286	0.05745
Full model	-779.5635	-699.9147	0.83	0.05662

We know that the highest adj R^2 score and the lowest scores for AIC, BIC, and RMSE is preferred. The selective model with the Smoking.status and Deep.sleep.percentage interaction term satisfies all these conditions, and is the model we went with. Here is the R summary of the model:

```
Call:
lm(formula = train$Sleep.efficiency ~ train$Age + train$Deep.sleep.percentage +
    train$REM.sleep.percentage + train$Smoking.status + train$Awakenings +
    train$Exercise.frequency + train$Smoking.status * train$Deep.sleep.percentage)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.165042	-0.035748	0.003101	0.037555	0.127968

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**0.4124169**	**0.0340472**	**12.113**	**< 2e-16**	***
train$Age	**0.0009596**	**0.0002598**	**3.693**	**0.000265**	***
train$Deep.sleep.percentage	**0.0047850**	**0.0003045**	**15.716**	**< 2e-16**	***
train$REM.sleep.percentage	**0.0073983**	**0.0009887**	**7.483**	**9.08e-13**	***
train$Smoking.status1	**-0.1506002**	**0.0224400**	**-6.711**	**1.04e-10**	***
train$Awakenings1	**-0.0639891**	**0.0088345**	**-7.243**	**4.11e-12**	***
train$Awakenings2	**-0.1211768**	**0.0123256**	**-9.831**	**< 2e-16**	***
train$Awakenings3	**-0.1292851**	**0.0116506**	**-11.097**	**< 2e-16**	***
train$Awakenings4	**-0.1268066**	**0.0111938**	**-11.328**	**< 2e-16**	***
train$Exercise.frequency1	**-0.0022070**	**0.0098085**	**-0.225**	**0.822134**	
train$Exercise.frequency2	**0.0229386**	**0.0114262**	**2.008**	**0.045635**	*
train$Exercise.frequency3	**0.0085910**	**0.0090889**	**0.945**	**0.345351**	
train$Exercise.frequency4	**0.0248917**	**0.0130183**	**1.912**	**0.056872**	.
train$Exercise.frequency5	**0.0129071**	**0.0242158**	**0.533**	**0.594446**	
train$Deep.sleep.percentage :train$Smoking.status1	**0.0020909**	**0.0004189**	**4.992**	**1.04e-06**	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05519 on 285 degrees of freedom

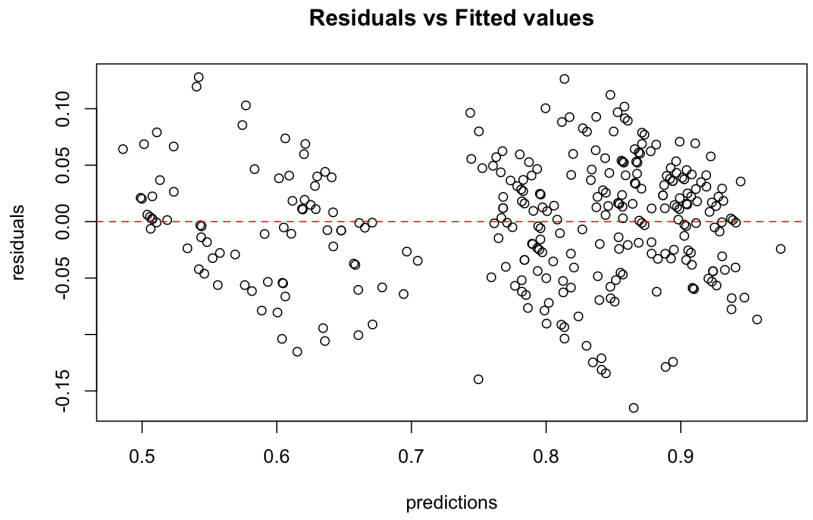
(16 observations deleted due to missingness)

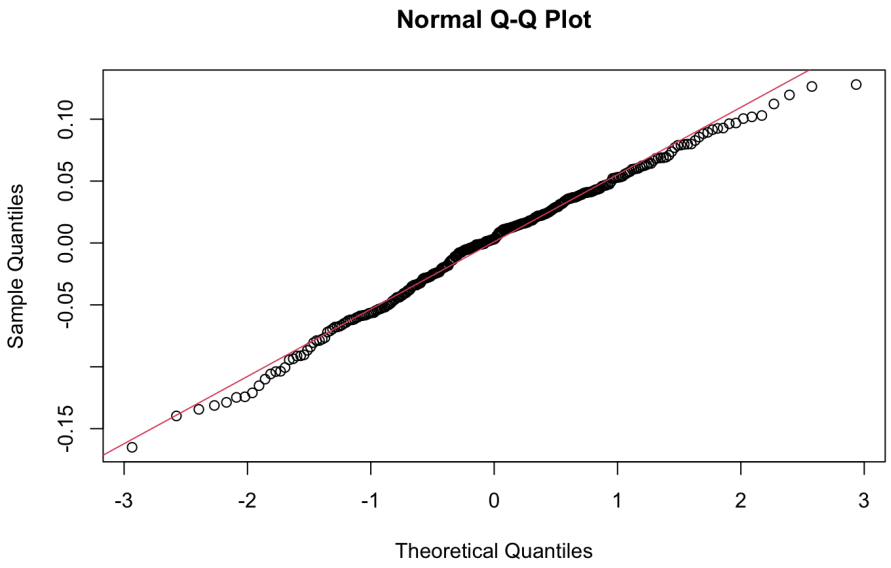
Multiple R-squared: 0.8492, Adjusted R-squared: 0.8418

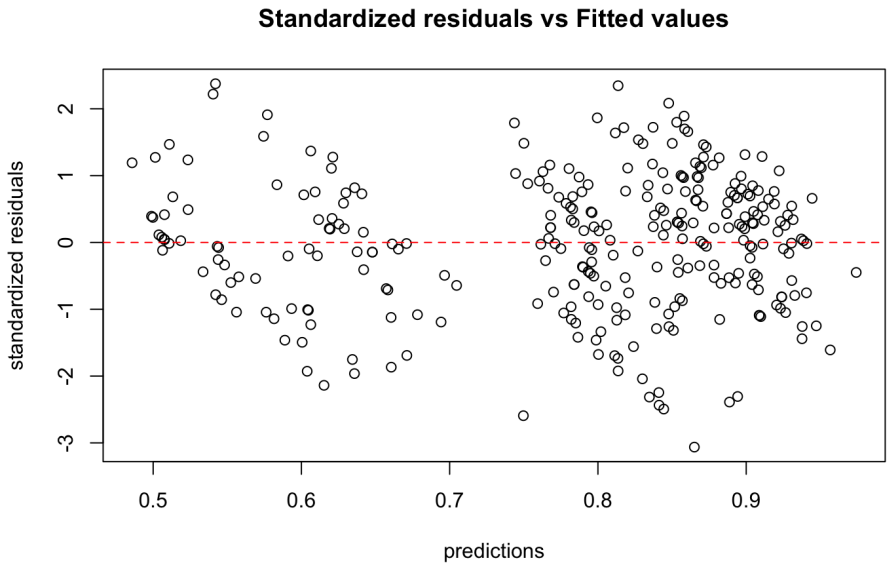
F-statistic: 114.7 on 14 and 285 DF, p-value: < 2.2e-16

Analysis

To check the model fit, we can analyze the diagnostics plots of our model:







Upon examining the residual plot, we can see that there is unequal variance throughout and the residuals seem to be clustered into two groups. However, the residuals seem to be more or less normally distributed. This is an indication that our model isn't perfect and could be improved further. This may be done through collecting information on more aspects of sleep or on more aspects of human physiology while sleeping.

Conclusion

The study demonstrated that selecting appropriate variables enhances model performance, utilizing a forward selection process to establish this. It was also discovered that among the variables considered, Gender, Sleep Duration, Caffeine Consumption, and Alcohol Consumption had lesser impact in predicting Sleep Efficiency compared to others.

What's more, we found it intriguing that alcohol consumption wasn't a key variable in predicting sleep efficiency. This might stem from its categorical classification in our model, limiting the effectiveness of a linear regression. We think that having more detailed alcohol consumption data, measured as a continuous variable to two or three decimal places, could enhance our model's precision. Incorporating alcohol as a quantitative variable might refine our regression analysis and influence the selection of variables for the model. Nonetheless, we are not surprised to find that gender had a negligible impact on the model and it was better to keep out of the model.

Generally, we were satisfied with the predictive capabilities of our model, yet we acknowledged the potential for improvement with more comprehensive data on factors like alcohol and caffeine intake to deepen our understanding of their impact on sleep efficiency. Initially, our goal was to gather insights to promote better sleep habits. Our optimal model highlighted the importance of maintaining a healthy lifestyle, avoiding smoking, and showed that variables like age, exercise frequency, and REM sleep percentage positively correlate with sleep efficiency.

References:

Kim, B., Branas, C. C., Rudolph, K. E., Morrison, C. N., Chaix, B., Troxel, W. M., & Duncan, D. T. (2022). Neighborhoods and sleep health among adults: A systematic review. Sleep Health, 701. https://doi.org/10.1016/j.sleh.2022.03.005

Professional, C. C. medical. (n.d.). *Controlled zzzs*. Cleveland Clinic. https://my.clevelandclinic.org/health/body/12148-sleep-basics

PhD, A. N., Rhee, J. U., Haynes, P., Chakravorty, S., Patterson, F., Killgore, W. D. S., Gallagher, R. A., Hale, L., Branas, C., Carrazco, N., Alfonso-Miller, P., Gehrels, J. A., & Grandner, M. A. (2021). Smoke at night and sleep worse? The associations between cigarette smoking with insomnia severity and sleep duration. *Sleep health*, *7*(2), 177–182. https://doi.org/10.1016/j.sleh.2020.10.006