Nelson Gao
Hao Jiang
Presley Nip
Darren Tsang

STAT 306 Final Project

Introduction

Regardless of a person's age, getting consistent good quality sleep plays an essential role in one's well-being. Poor sleep is associated with lower cognitive function, reduced immune function, and an impaired response to different stresses in daily life (Williams et al., 2013). Sleep is even more important for younger children as poor quality sleep can negatively impact brain development (Jiang, 2019).

However, it is not always easy to get adequate sleep each and every day. Everyone has to face the same challenges of being able to balance social life and maintaining relationships, academic or career workload, and many other distractions each and every day. The time someone has left in a day, after accomplishing all that they want to do, may be limited and this is why it is important to ensure that the time we do have set aside for sleep is put to good use. This is known as sleep efficiency. Sleep efficiency is the ratio between the total time they have put aside for sleep, compared to the time a person is actually sleeping for. Furthermore, some studies suggest that cognitive function may be affected more by sleep efficiency rather than sleep duration (Gruber et al., 2014).

In this report, we take a look at a dataset that was collected by combining the results of surveys and sleep monitoring device tests that study participants completed in 2021 in the UK. The dataset contains observations from 452 different participants and includes information about sleep quality, quantity, and lifestyle habits. Our goal is to learn more about improving sleep quality by exploring the relationship between the lifestyle habits recorded in this dataset and sleep efficiency.

There are 4 lifestyle variables that we will explore along with sleep efficiency:

- SleepEfficiency: the time a participant was in bed divided by the time they spent sleeping
- AlcoholConsumption: the amount of alcohol consumed within 24 hours of the study
- CaffeineConsumption: the amount of caffeine consumed within 24 hours of the study
- ExerciseFrequency: number of times exercised the week leading up to the study
- SmokingStatus: a categorical variable categorizing between smokers and non-smokers

The dataset was obtained from Kaggle and originally posted by the user known as "Equilibrium". However, there is no cited source of the data, and the author claims that the data was collected as part of a study conducted at a fictitious university supposedly located in the UK. Because of this, it is quite likely that the dataset does not contain any real data, and was simply generated for further use.

Analysis

Dara Cleaning and Wrangling

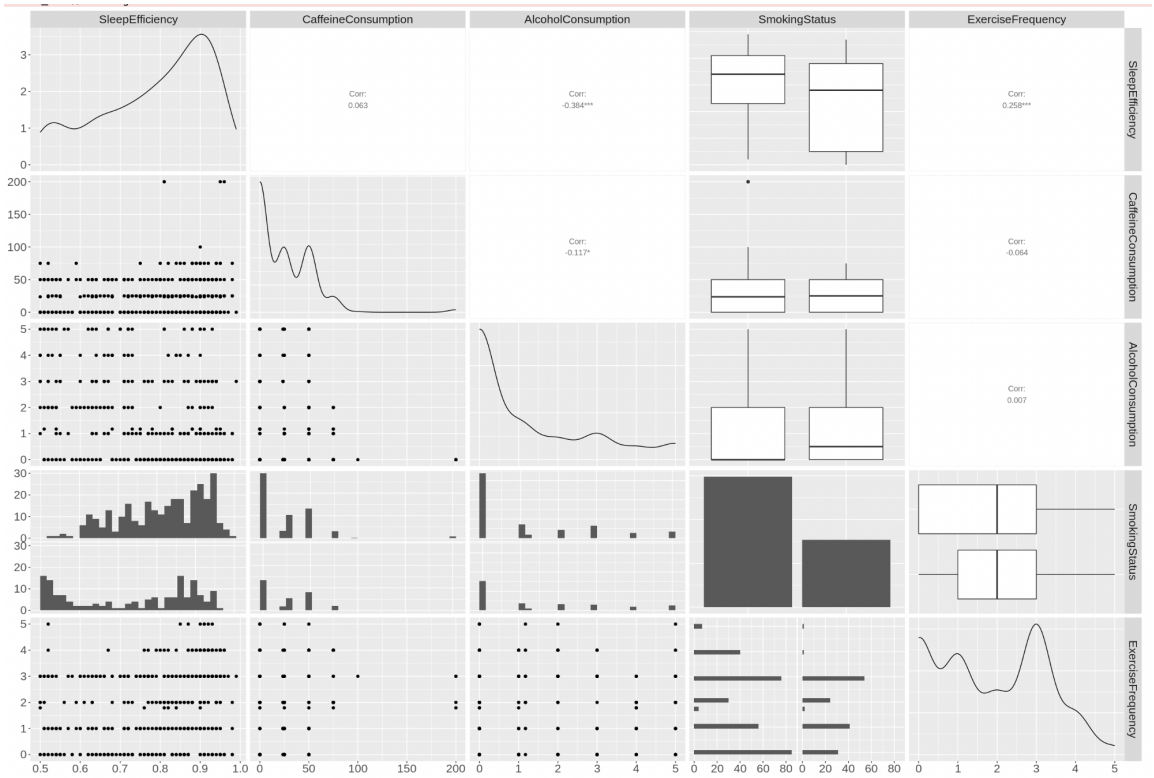
The first step of our analysis was to clean the data that we obtained. We removed the columns that were unrelated to our analysis and changed some of the variable names to be more intuitive.

	SleepEfficiency	CeffeineConsumption	AlcoholConsumption	SmokingStatus	ExerciseFrequency
	<dbl>	<dbl>	<dbl>	<fct>	<dbl>
1	0.88	0	0	Yes	3
2	0.66	0	3	Yes	3
3	0.89	0	0	No	3
4	0.51	50	5	Yes	1
5	0.76	0	3	No	3
6	0.90	NA	0	No	1

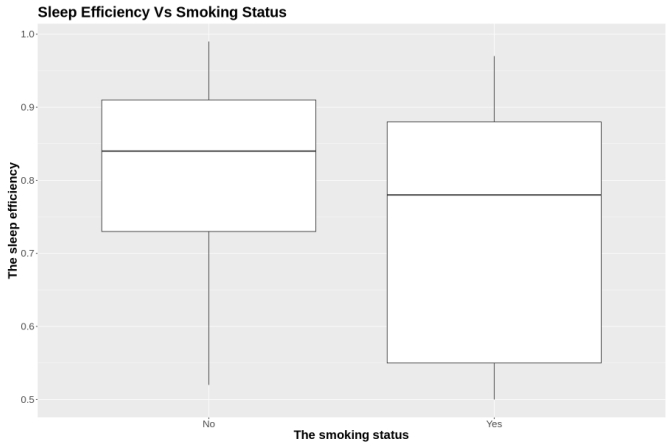
By observing the first six rows of the clean data, we found the presence of NA in the data, with a total of 45 observations, accounting for 0.0995575% of the total. It may result in the loss of statistical power and efficiency to just delete the rows with missing values, so we replaced the missing values with the mean values instead.

Associations Discovery

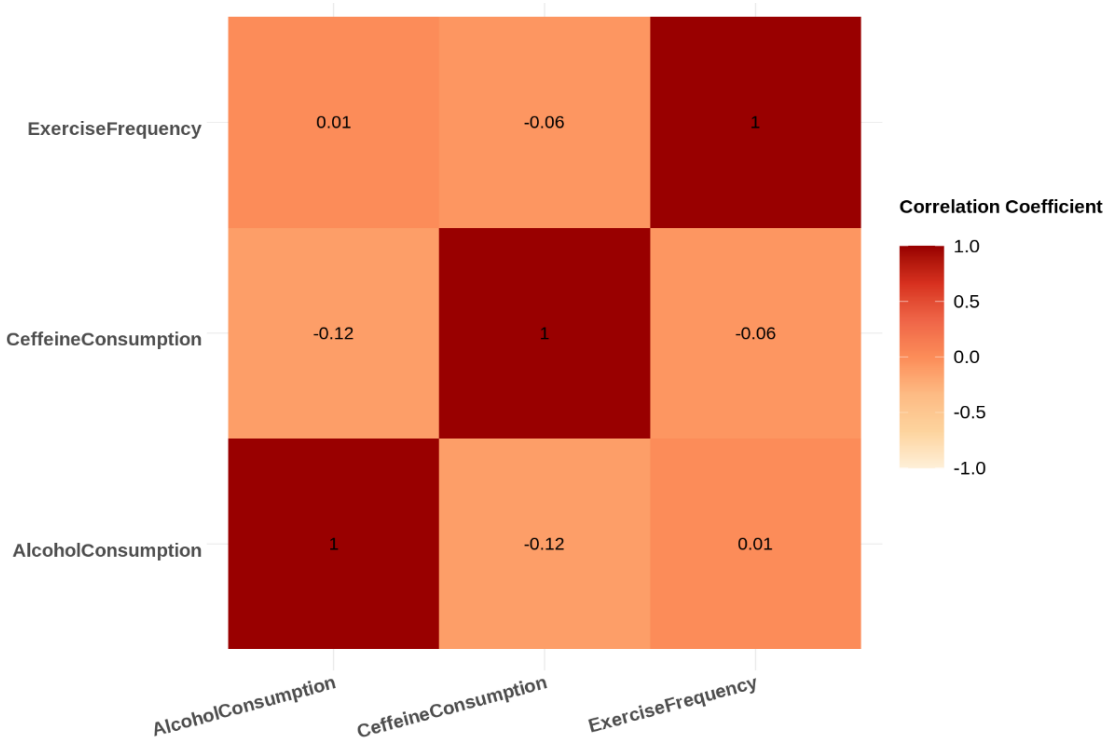
After replacing the missing values in the dataset, we explored the correlation between the response variable and the independent variables.



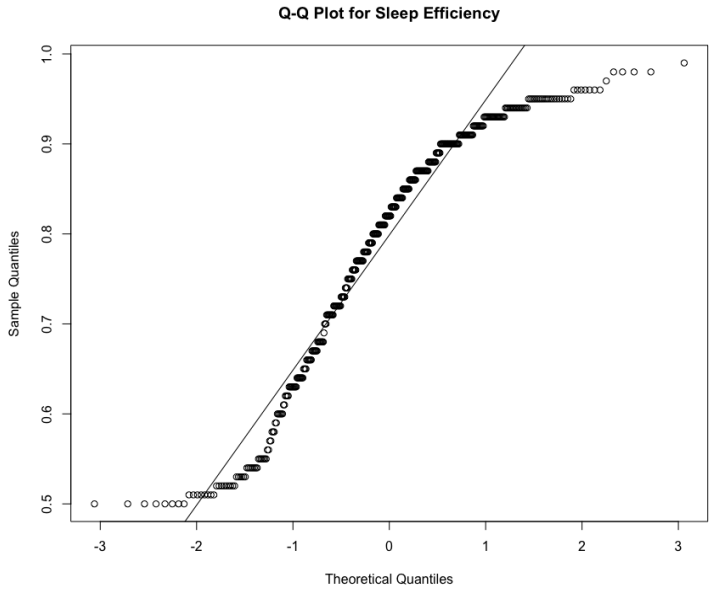
From this figure, we noticed that there is a negligible positive correlation (0.063) between SleepEfficiency and CaffeineConsumption, a low negative correlation (-0.384) between SleepEfficiency and AlcoholConsumption, and a low positive correlation (0.258) between SleepEfficiency and ExerciseFrequency. Also, we found that the median value of SleepEfficiency for non-smokers is higher than for smokers.



For our continuous variables, CaffeineConsumption, AlcoholConsumption and ExerciseFrequency, we used the correlation coefficient matrix to evaluate whether there is high multicollinearity between any two variables. Fortunately, our results are all very close to zero(0.01,-0.06,-0.12), indicating that we can incorporate all three variables into the model without the concern of multicollinearity.



From the Q-Q plot, the data points do not deviate too much from the Q-Q line. Thus, it is reasonable to assume the normality of the response variable (SleepEfficiency).



Model Selection

We split the clean data into two sets, with 70% being used as the training set for model fitting (size=316) and 30% for the testing set for model detection (size=104). We used the exhaustive model selection method to find the predictor variable subset with the best evaluation criterion and called the model fitted by the variable subset as generative_fit.

	(Intercept)	CaffeineConsumption	AlcoholConsumption	SmokingStatusYes	ExerciseFrequency
1	TRUE	FALSE	TRUE	FALSE	FALSE
2	TRUE	FALSE	TRUE	FALSE	TRUE
3	TRUE	FALSE	TRUE	TRUE	TRUE
4	TRUE	TRUE	TRUE	TRUE	TRUE

n_input_variables	RSS	ADJ.R2
<int>	<dbl>	<dbl>
1	4.934950	0.1487804
2	4.601373	0.2037827
3	4.323872	0.2494031
4	4.320209	0.2476275

We found that when we use three variables, we had the largest adjusted R^2 (0.2494031) and the smallest residual sum squared (4.323872) which is close to when we use four variables (4.320209). Thus, we chose 3 input variables (AlcoholConsumption, SmokingStatus, and ExerciseFrequency) to fit a generative_fit model with the testing set.

The **result** for the **additive model (generative_fit)** is:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.811518   0.021290  38.117  < 2e-16 ***
AlcoholConsumption -0.032463   0.007713  -4.209 5.62e-05 ***
SmokingStatusYes   -0.091609   0.022782  -4.021 0.000113 ***
ExerciseFrequency   0.025238   0.007909   3.191 0.001895 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1098 on 100 degrees of freedom
Multiple R-squared:  0.3185,    Adjusted R-squared:  0.298
```

$$Sleep\ Efficiency\ = 0.811516\ - 0.032463\widehat{\beta}1\ +\ 0.025238\widehat{\beta}2\ - 0.091609\widehat{\beta}3$$

$$Where\ \widehat{\beta}1\ =\ AlcoholConsumption,\ \widehat{\beta}2\ =\ ExerciseFrequency,\ and\ \widehat{\beta}3\ =\ SmokingStatus$$
$$\widehat{\beta}3\ =\ 1\ if\ smoker,\ 0\ otherwise$$

All p-values for the variables were below 0.05 indicating that each of the variables has a statistically significant relationship with our response variable, SleepEfficiency . AlcoholConsumption had a p-value of 5.62 e^{-05}, ExerciseFrequency had a p-value of 0.001895, and SmokingStatus had a p-value of 0.000113. The adjusted R^2 for the model provided is 0.298.

Then, we explored whether including interaction will result in a better fit by trying all combinations of adding the interaction to the generative_fit. The best interaction model with the highest adjusted R^2 (0.3633) includes the interactions both between SmokingStatus and AlcoholConsumption, and between the SmokingStatus and ExerciseFrequency.

The **result** for the **interactive model** is:

```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.7913351  0.0218908  36.149  < 2e-16 ***
AlcoholConsumption                -0.0155364  0.0093239  -1.666  0.09885 .
SmokingStatusYes                  -0.0006397  0.0398387  -0.016  0.98722
ExerciseFrequency                  0.0263006  0.0088978   2.956  0.00391 **
AlcoholConsumption:SmokingStatusYes -0.0512206  0.0155381  -3.296  0.00136 **
SmokingStatusYes:ExerciseFrequency  -0.0219769  0.0175131  -1.255  0.21251
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1046 on 98 degrees of freedom
Multiple R-squared:  0.3942,    Adjusted R-squared:  0.3633
```

$$Sleep\ Efficiency = 0.7913351 - 0.0155364\hat{\beta}1 + 0.020628\hat{\beta}2 - 0.0006397\hat{\beta}3 - 0.0512206(\hat{\beta}1 * \hat{\beta}3) - 0.0219769(\hat{\beta}2 * \hat{\beta}3)$$

$$Where\ \hat{\beta}1 = AlcoholConsumption,\ \hat{\beta}2 = ExerciseFrequency,\ and\ \hat{\beta}3 = SmokingStatus$$

$$\hat{\beta}3 = 1\ if\ smoker,\ 0\ otherwise$$

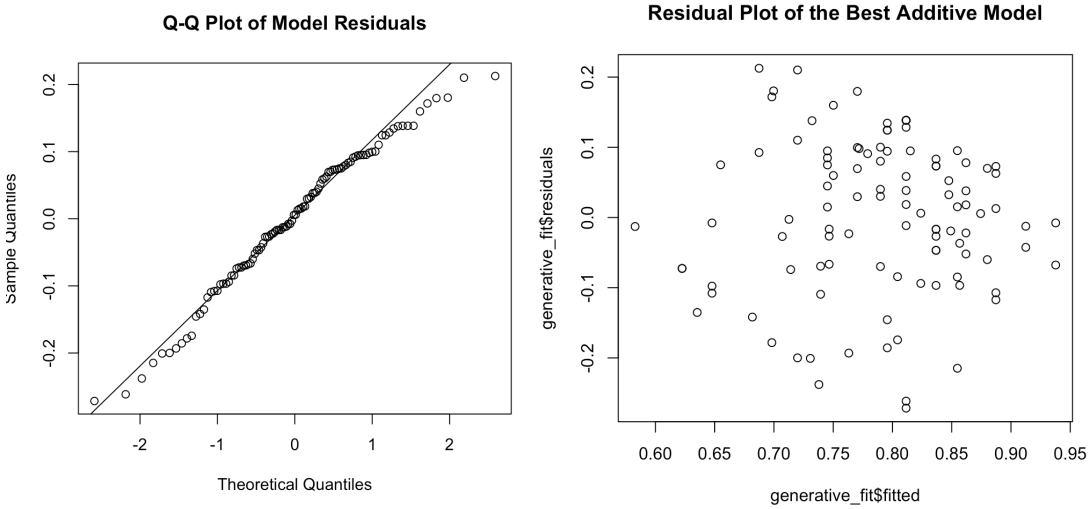
The p-values for the interactive model were not all significant, as AlcoholConsumption and SmokingStatus, and the interaction between ExerciseFrequency and SmokingStatus all had p-values higher than 0.05. Though, ExerciseFrequency and the interaction between AlcoholConsumption and SmokingStatus both had statistically significant p-values at 0.00391 and 0.00136, respectively. As well, the residuals for both the additive and interactive models were similar. Thus, it made more sense to us to continue with the additive model since we are not performing this analysis to predict sleep efficiency, but to simply determine which variables are related to it.

Multicollinearity Verification

AlcoholConsumption	SmokingStatus	ExerciseFrequency
1.017	1.026	1.041

Since all Variance Inflation Factors (VIF) values are below the threshold of 5, we believe the best generative model has low multicollinearity.

Normality and Heteroscedasticity Verifications



From the Q-Q plot of the model residuals, there is no violation of the normality assumption since there is not much deviation of the data points from the line.

However, there is a negative pattern and higher concentration of data on the right side of the Residual vs Fitted Value plot, which may indicate a violation of the linearity assumption or the presence of outliers. Though, a slight deviation from the assumptions may not have a

significant impact on the overall results of the analysis. Therefore, it is feasible to use the additive model for inference.

Conclusion and Discussion

After analysis of our dataset, we found three lifestyle habits that showed significance and may play a role in how efficiently we sleep. The three variables are AlcoholConsumption (1), Exercise Frequency (2) – which are treated as continuous variables, and SmokingStatus (3)– which is treated as a dummy variable with 1 being a smoker, and 0 otherwise.

Before choosing the additive model, we generated an interactive model where there were interactive terms between both SmokingStatus with AlcoholConsumption and SmokingStatus with ExerciseFrequency. This model had a higher adjusted R2 at 0.3633, but we decided on the additive model because we believed that it was more cohesive. The interactive model treats both the AlcoholConsumption and the SmokingStatus as non-significant terms. However, this is not quite intuitive because drinking will initially cause drowsiness but ultimately leads to more frequent awakenings during the night and early morning (Park et al., 2015), and smokers may have poorer sleep quality (Liao et al., 2019).

Our final inference model with all significant terms is:
$Sleep\ Efficiency\ = 0.811516\ - 0.032463X1\ +\ 0.025238X2\ - 0.091609X3$

From this model, we can infer that alcohol consumption and smoking will negatively impact our sleep efficiency but the frequency of exercise will improve it. Therefore, we may increase sleep efficiency with less alcohol consumption, less smoking, and more exercise frequency.

The additive model may not be a good model for inferencing since it has a low adjusted R2 and there was a pattern in the residual plot. Also, we may need to apply transformations in the future to get a better-fitting model fulfilling the randomness and the constant variance of the corresponding residual plot. Thus, the true relationships may be different than the above.

Furthermore, given that the data used in this analysis was not from an actual study, in the future, it would be interesting to see if the relationships we found between lifestyle habits are also observed in real data that comes from an actual study, or if a completely different model would fit the real data. Further exploration can be done with the three variables to examine their predictive relationship with sleep efficiency.

References

Gruber, R., Somerville, G., Enros, P., Paquin, S., Kestler, M., & Gillies-Poitras, E. (2014). Sleep efficiency (but not sleep duration) of Healthy School-age children is associated with grades in math and languages. *Sleep Medicine*, *15*(12), 1517–1525. https://doi.org/10.1016/j.sleep.2014.08.009

Jiang, F. (2019). Sleep and early brain development. *Annals of Nutrition and Metabolism*, *75*(Suppl. 1), 44–54. https://doi.org/10.1159/000508055

Liao, Y., Xie, L., Chen, X., Kelly, B., Qi, C., Pan, C., Yang, M., Hao, W., Liu, T., & Tang, J. (2019). Sleep quality in cigarette smokers and nonsmokers: findings from the general population in central China. *BMC Public Health, 19*(1). https://doi.org/10.1186/s12889-019-6929-4

Park, S. J., Oh, M. M., Lee, B., Kim, H. G., Lee, W. Y., Lee, J. H., Lim, J. H., & Kim, J. (2015). The Effects of Alcohol on Quality of Sleep. *Korean Journal of Family Medicine, 36*(6), 294. https://doi.org/10.4082/kjfm.2015.36.6.294

Williams, P. G., Cribbet, M. R., Rau, H. K., Gunn, H. E., & Czajkowski, L. A. (2013). The effects of poor sleep on cognitive, affective, and physiological responses to a laboratory stressor. *Annals of Behavioral Medicine*, *46*(1), 40–51. https://doi.org/10.1007/s12160-013-9482-x