Linear Modelling of Sleep Efficiency

STAT 306 Final Project (C7)

Jonathan Nah, Bob Pham, Matthew Smith, Junbo Rao

December 7, 2023

1 Introduction

Sleep is an integral to human well-being as it plays a pivotal role in learning, skill development, and general health. Given its self-evident importance, it is no surprise that sleep has been explored to great extent in previous literature. Namely, prior studies have linked poor sleep quality to poor academic performance [Oka+19], and have shown that improving sleep quality leads to better mental health [Sco+21]. Struggles with mental health are at an all-time high amongst students [Ped+14], so determining what is associated with sleep quality is ever relevant in today's society. Answering such a question better-equips college students on making lifestyle decisions regarding sleep. As such, the goal of this report is to investigate **how can one optimize their sleep efficiency**. In order to do this, we will investigate what factors are associated with sleep efficiency, and how they fit in a linear model.

Sleep efficiency, defined as the ratio of time asleep to time in bed, is considered the gold-standard in sleep quality research [RS16]. As such, we have elected to investigate the Kaggle sleep efficiency dataset for this study. This data was collected in an observational study that examined the sleep patterns of Artificial Intelligence students studying at ENSIAS in Morocco. The data collection process occurred from January 3$^{\text{rd}}$, 2021, to December 31$^{\text{st}}$, 2021. The data was collected using a combination of self-reported surveys, actigraphy for activity monitoring, and polysomnography for sleep monitoring. Given the observational nature of the study, our analysis will only allow us to explore relationships in the data, not conclude causal information on sleep efficiency. A summary of the of the columns of the data-set is present in table 1.

	Name	Format
1	Unique subject identifier	Integer
2	Age	Years
3	Gender	Factor (Male | Female)
4	Bedtime	YYYY-mm-dd HH:MM:SS
5	Wake-up time	YYYY-mm-dd HH:MM:SS
6	Sleep duration	Hours
7	REM sleep	Percentage
8	Deep sleep	Percentage
9	Light sleep	Percentage
10	Awakenings	Counts
11	Caffeine consumption	mg within last 24 hours
12	Alcohol consumption	oz within last 24 hours
13	Smoking status	Factor (True | False)
14	Exercise status	Counts of sessions in the last 7 days
15	Sleep efficiency	Percentage

Table 1: A table summarizing the 15 columns of the Kaggle ENSIAS sleep efficiency dataset. Bedtime and wakeup-time format represent the date and time of the respective occurrence. The duration threshold of an exercise session is unknown. For the purposes of this study, the response variable of interest is the sleep efficiency metric.

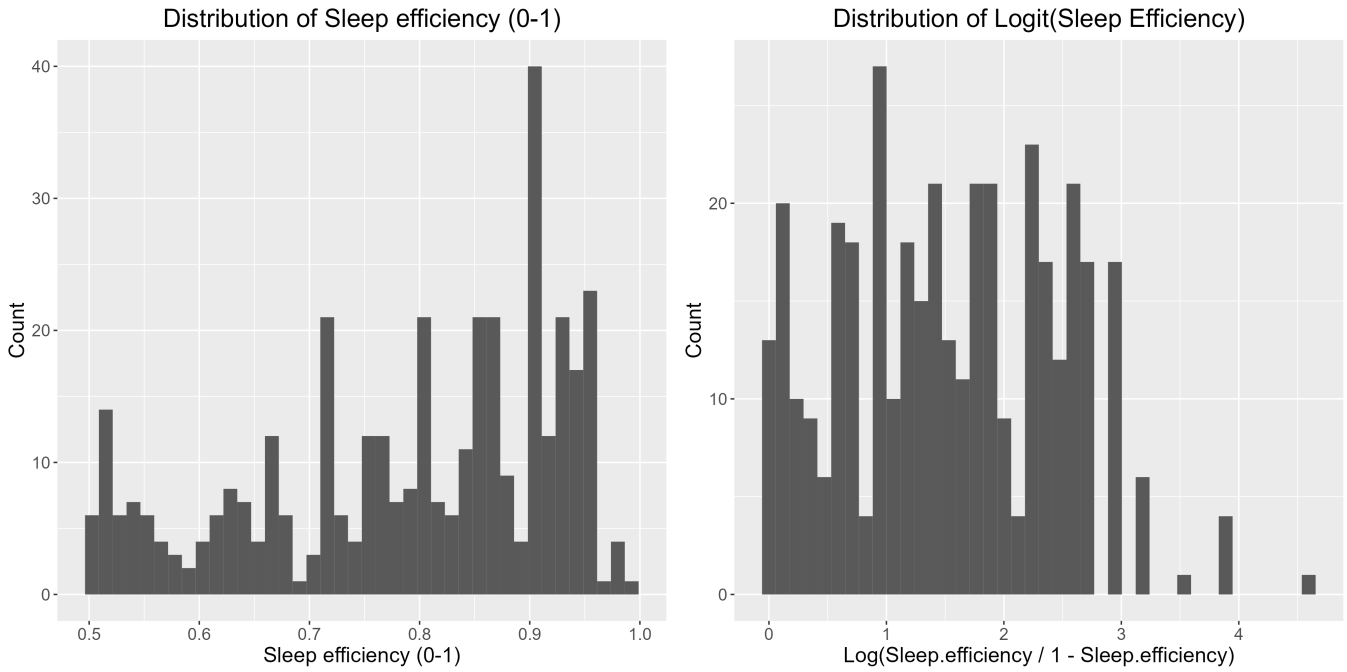
2 Analysis

Before attempting to model sleep efficiency, further investigation on the features of the explanatory variables of the data set is required.

2.1 Data Visualization

The following is an analysis on each of the columns. **All correlation and covariance statistics are against the logit transformed sleep efficiency**. We cleaned the data-set by removing any incomplete rows. Before removal, the sample size was $N = 452$, and after removal the sample size is $N = 388$.

2.1.1 Sleep efficiency



	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	SD
Not transformed	0.5000	0.7000	0.8200	0.7893	0.9000	0.9900	0.14
Transformed	0.0000	0.8473	1.5163	1.5452	2.1972	4.5951	0.92

Figure 1: Distribution and summary statistics of sleep efficiency and logit-transformed sleep efficiency.

Sleep efficiency is a calculated proportion of time spent sleeping, from 0 to 1. We see that in our data-set it is skewed towards participants having higher sleep efficiencies (closer to 1). Because sleep efficiency is contained in the interval $(0, 1)$, we have performed a logit transform so the values map to the real line. This will prove useful when investigating the association with other explanatory variables in a linear model.

2.1.2 Unique subject identifier

The unique subject identifier uniquely identifies each subject, and ranges from [1, 452]. When fitting a linear model, we will disregard this column as it is artificially manufactured. It is irrelevant to predicting sleep efficiency.

2.1.3 Age

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
9.00	29.00	41.00	40.83	52.00	69.00	13.4	0.12	1.48

Figure 2: Distribution and summary statistics of age in years.

The age of the participants is bimodally distributed, with two peaks at approximately 30 and 50 years. When examining the plot on the right, we naively note a relatively flat linear association between sleep efficiency and age with a relatively constant variance. It is however notable that this trend is only apparent beyond 20 years of age.

2.1.4 Gender

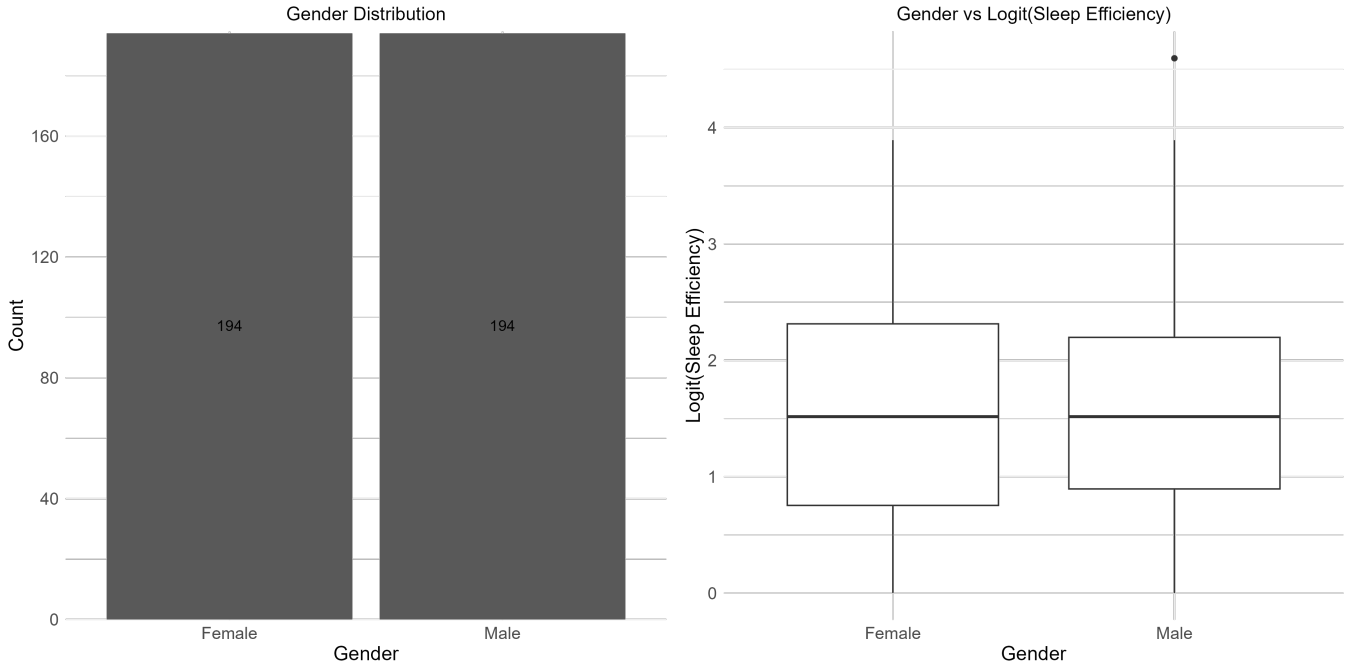
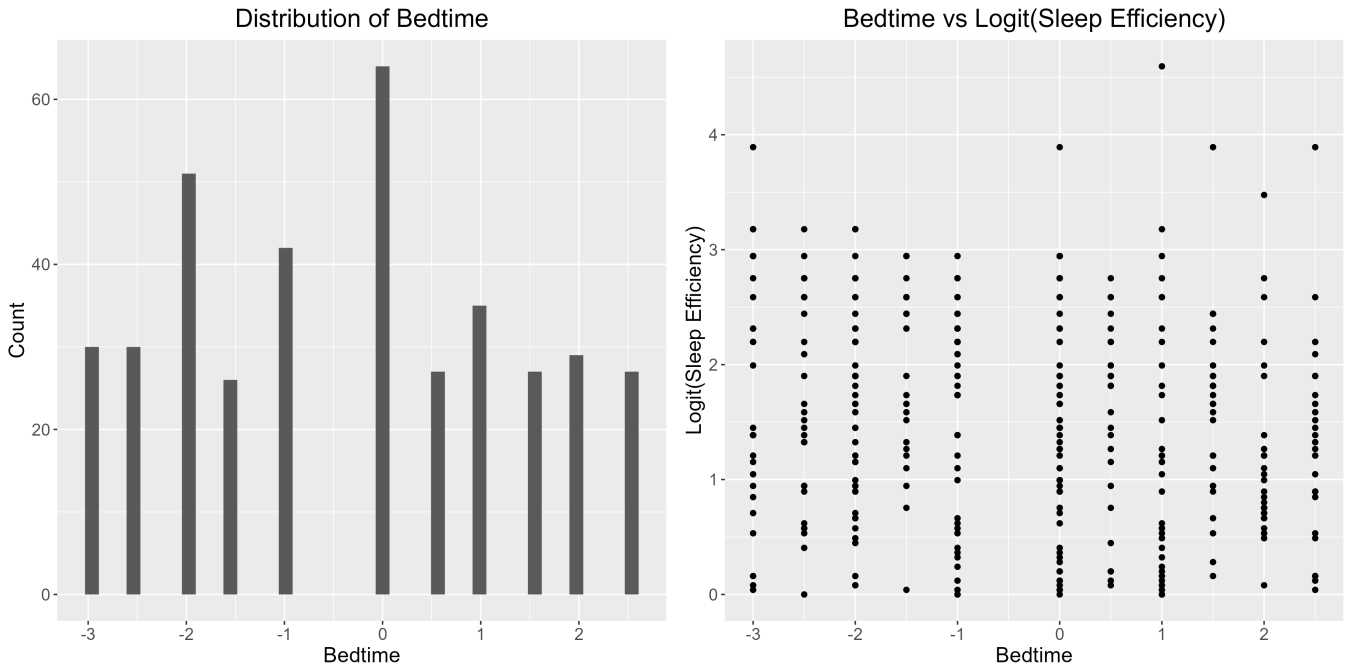


Figure 3: Count and side-by-side boxplots depicting sleep efficiency against gender.

The data-set includes exactly the same amount of female, and there appears to be very little difference between the two groups. From the boxplots, we can see the median and interquartile range are nearly identical between the two groups, but the male group does include a single outlier beyond the whiskers. Despite this, this naive analysis indicates that gender likely does not play a significant role in sleep efficiency prediction.

2.1.5 Bedtime



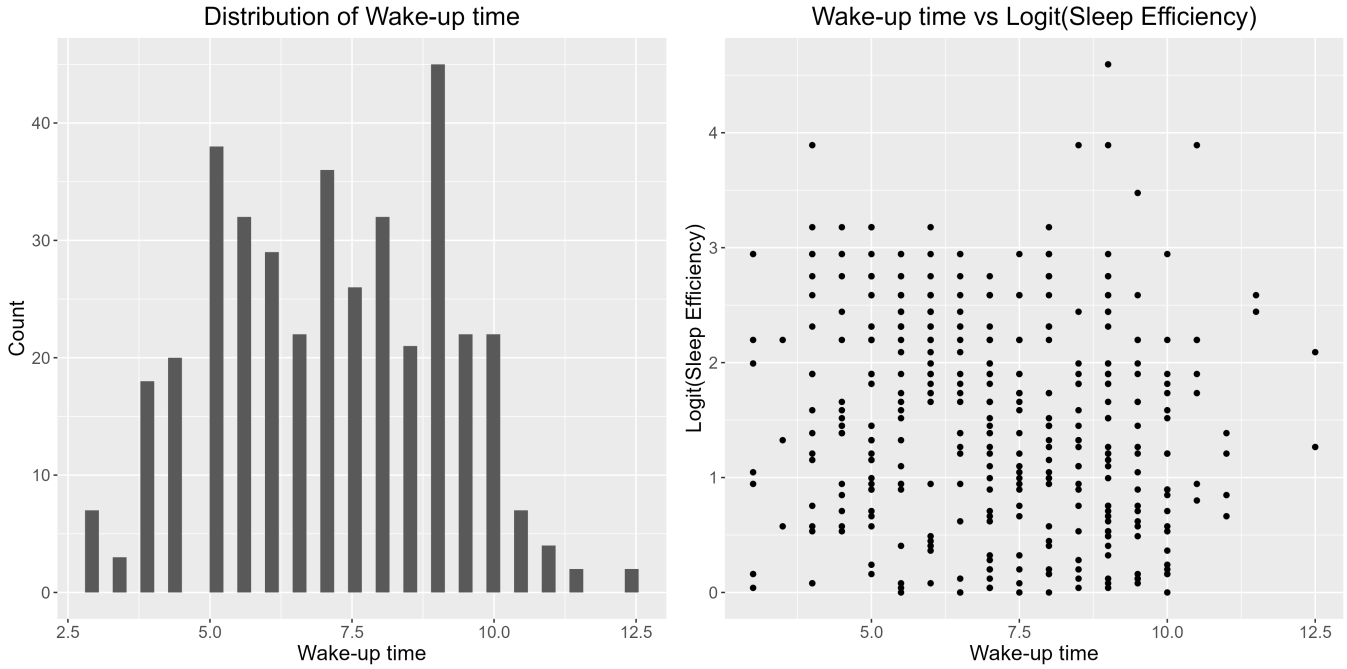
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
-3.0000	-2.0000	0.0000	-0.3441	1.0000	2.5000	1.69	-0.18	-0.28

Figure 4: Distribution and summary statistics of bedtimes.

In the data-set, bedtimes are collected as YYYY-mm-dd HH:MM:SS formatted strings. We assume that sleep efficiency is independent of month and day, so we chose to exclusively look at the time of day. We scale the times to center on midnight (12:00 am), with bedtimes being calculated as number of hours before or after midnight.

From the plots, we notice that the bedtimes are distributed approximately normally, centered at midnight.

2.1.6 Wake-up time



Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
3.000	5.500	7.000	7.107	9.000	12.500	2	-0.16	-0.29

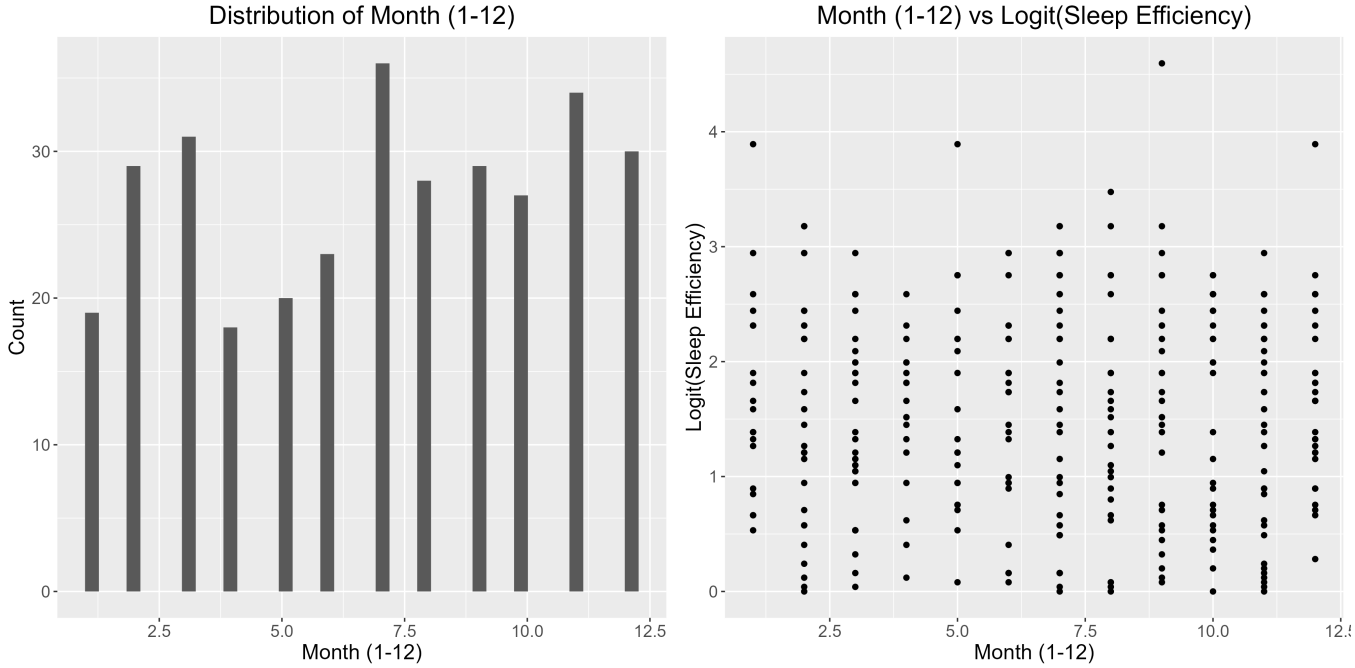
Figure 5: Distribution and summary statistics of wake-up times.

In the data-set, wake-up times are collected as Y-m-d H:M:S formatted strings. Again, we assume that sleep efficiency is independent of month and day, so we chose to exclusively look at the time of day, so we choose to look exclusively at the time of day. We scale the times to center on midnight (12:00 am), with wake-up being calculated as number of hours before or after midnight.

We notice that the wake-up times are distributed approximately normally, centered 7.5 hours after midnight (7:30 am).

It is important to note that wake-up times are collinear with bedtimes and sleep duration, since we can obtain the wake-up time by adding sleep duration to the bedtime. This linear dependence between covariates indicates that wake-up time should be excluded when fitting a linear model, as its information is already encoded in other explanatory variables.

2.1.7 Month



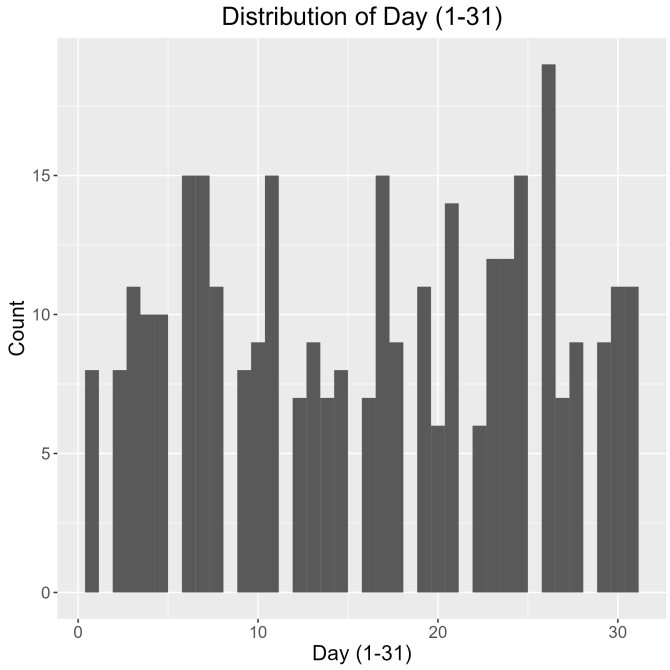
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
1.000	4.000	7.000	6.832	10.000	12.000	3.42	0.01	0.03

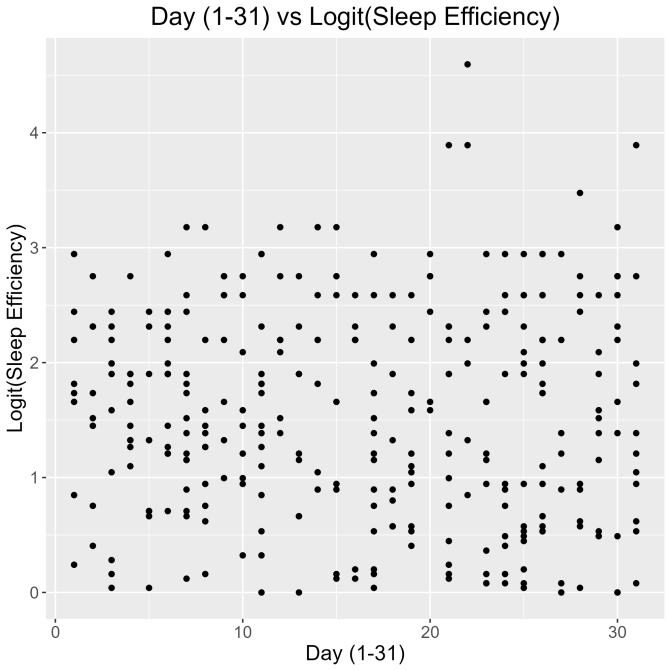
Figure 6: Distribution and summary statistics of month recorded.

In the data-set, wake-up and bedtimes times are collected as Y-m-d H:M:S formatted strings. Again, we assume that sleep efficiency is independent of month and day, so we add another variable for month, being a number 1-12 corresponding to the months January-December.

We notice that samples were taken almost uniformly across the year.

2.1.8 Day





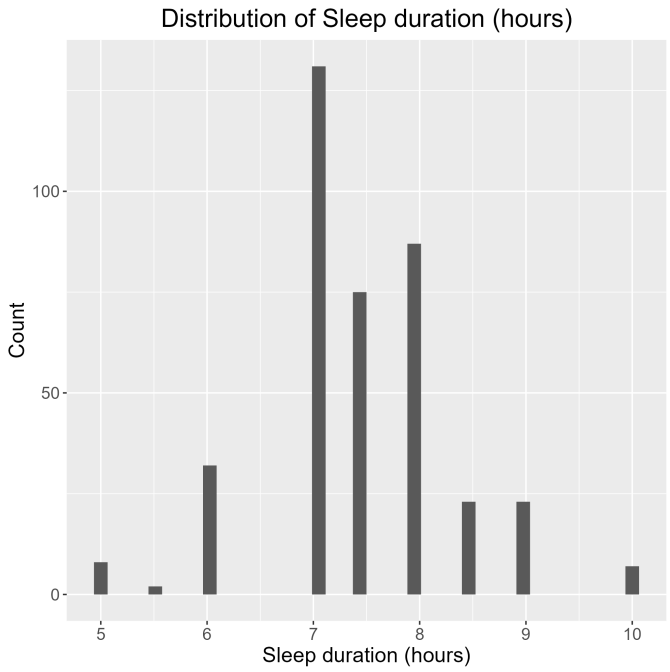
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
1.00	8.00	17.00	16.34	25.00	31.00	8.97	-0.01	-0.07

Figure 7: Distribution and summary statistics of month recorded.

In the data-set, wake-up and bedtimes times are collected as Y-m-d H:M:S formatted strings. Again, we assume that sleep efficiency is independent of month and day, so we add another variable for day, being a number 1-31 representing the day of the month

We notice that samples were taken almost uniformly across the days of the month.

2.1.9 Sleep duration



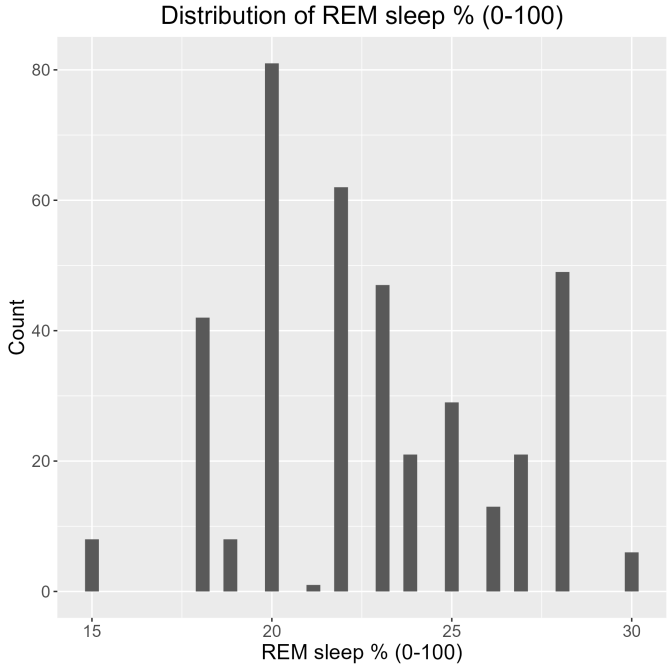
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
5.000	7.000	7.500	7.451	8.000	10.000	0.88	-0.010	-0.010

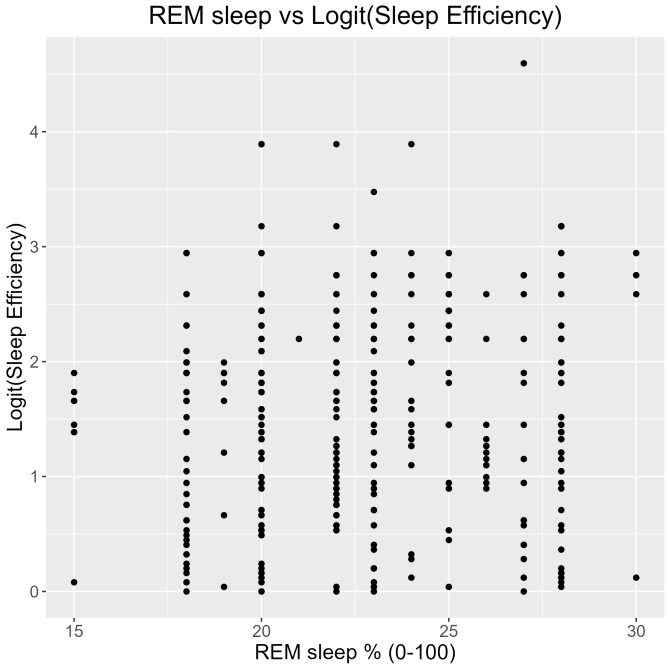
Figure 8: Distribution and summary statistics of sleep duration.

Sleep duration is captured in hours, and we see that most participants get 7 or more hours of sleep per night. The distribution is relatively normal with the main peak around 7 to 8 hours.

The previous collinearity discussion with bedtime and wake-up time applies here as usual.

2.1.10 REM sleep percentage





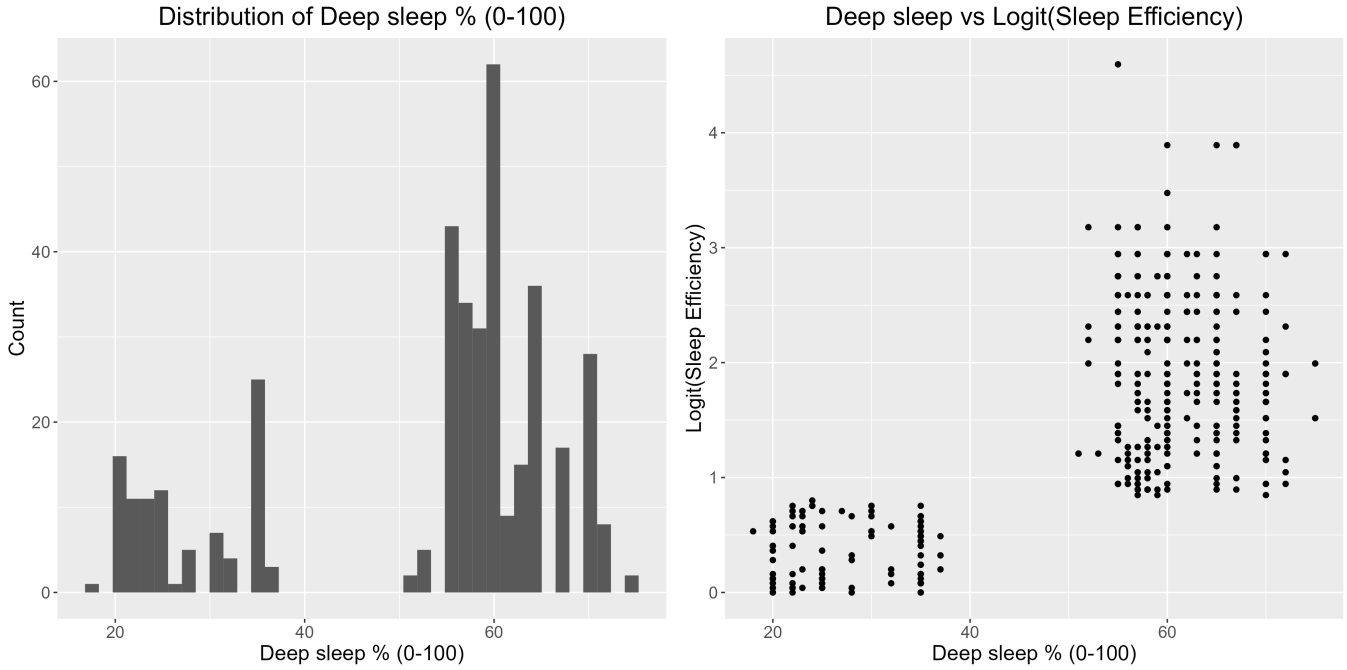
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
15.000	20.000	22.000	22.680	25.000	30.000	3.430	0.090	0.290

Figure 9: Distribution and summary statistics of REM sleep percentage.

Rapid eye movement (REM) sleep captures the percent of time the participant spends in REM sleep, which is a value between 0-1 (0 % - %100). The distribution of REM sleep is approximately normal, with the largest density around 20 - 25 %.

We note that REM sleep is collinear with Deep Sleep and Light sleep, since the sum of the 3 variables is always equal to 100%.

2.1.11 Deep sleep percentage



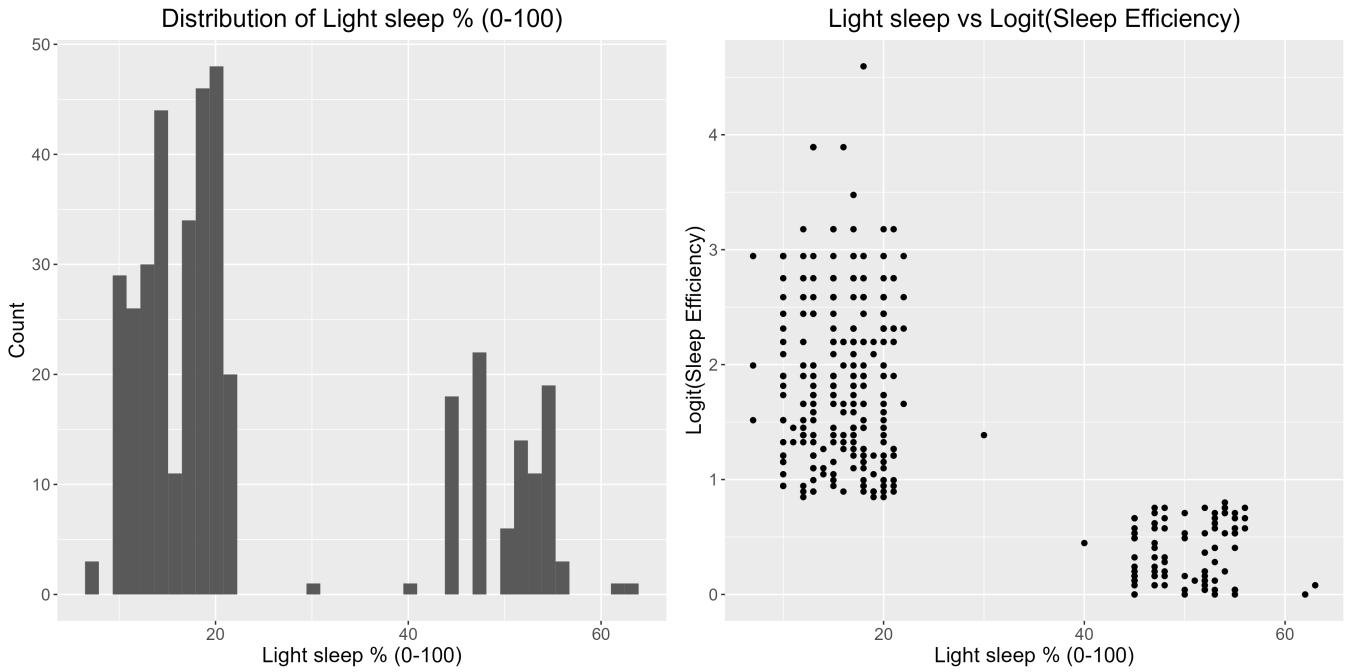
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
18.000	51.000	58.000	52.820	63.000	75.000	15.570	0.680	9.730

Figure 10: Distribution and summary statistics of deep sleep percentage.

Deep sleep captures the percent of time the participant spends in Deep sleep, which is a value between 0-1 (0 % - %100). The distribution of deep sleep is approximately bimodal, with two peaks centered at approximately 25 and 60 %.

We note that Deep sleep is collinear with REM Sleep and Light sleep, since the sum of the 3 variables is always equal to 100%.

2.1.12 Light sleep percentage



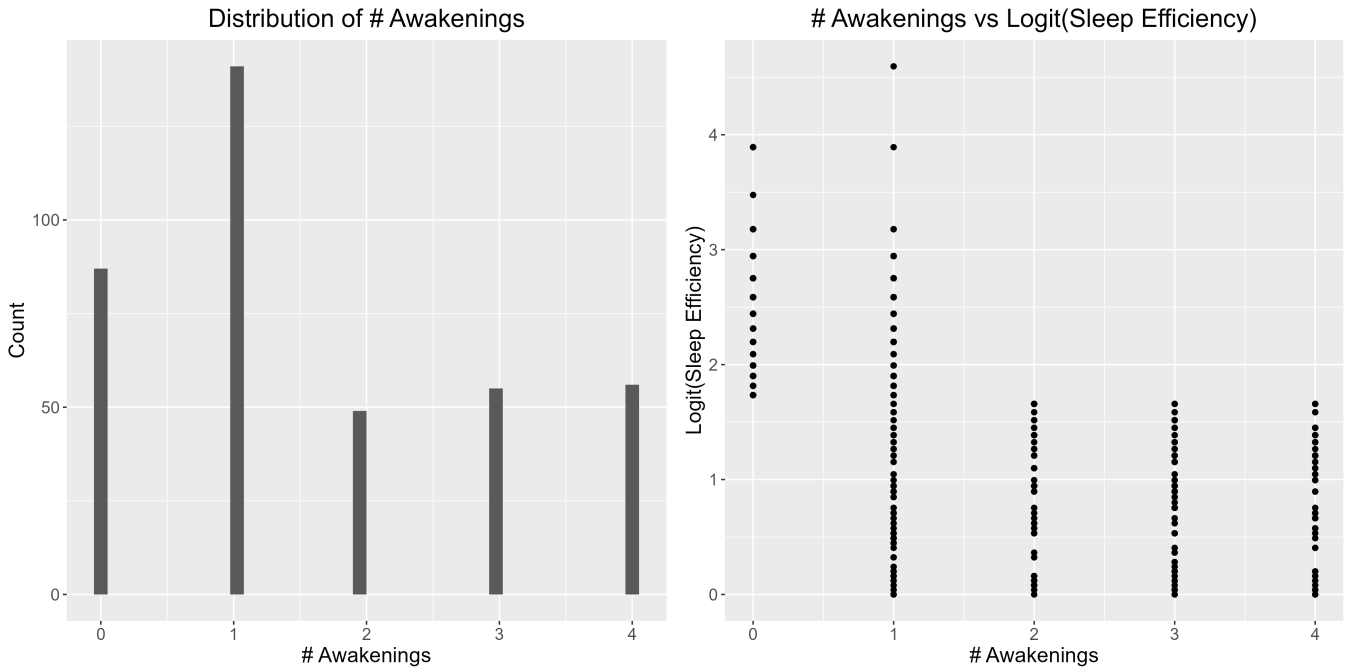
Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
7.000	15.000	18.000	24.500	24.000	63.000	15.310	-0.710	-10.020

Figure 11: Distribution and summary statistics of light sleep percentage.

Light sleep captures the percent of time the participant spends in ligth sleep, which is a value between 0-1 (0 % - %100). The distribution of light sleep is approximately bimodal, with two peaks centered at approximately 15 and 50 %.

We note that light sleep is collinear with REM Sleep and Deep sleep, since the sum of the 3 variables is always equal to 100%.

2.1.13 Awakenings

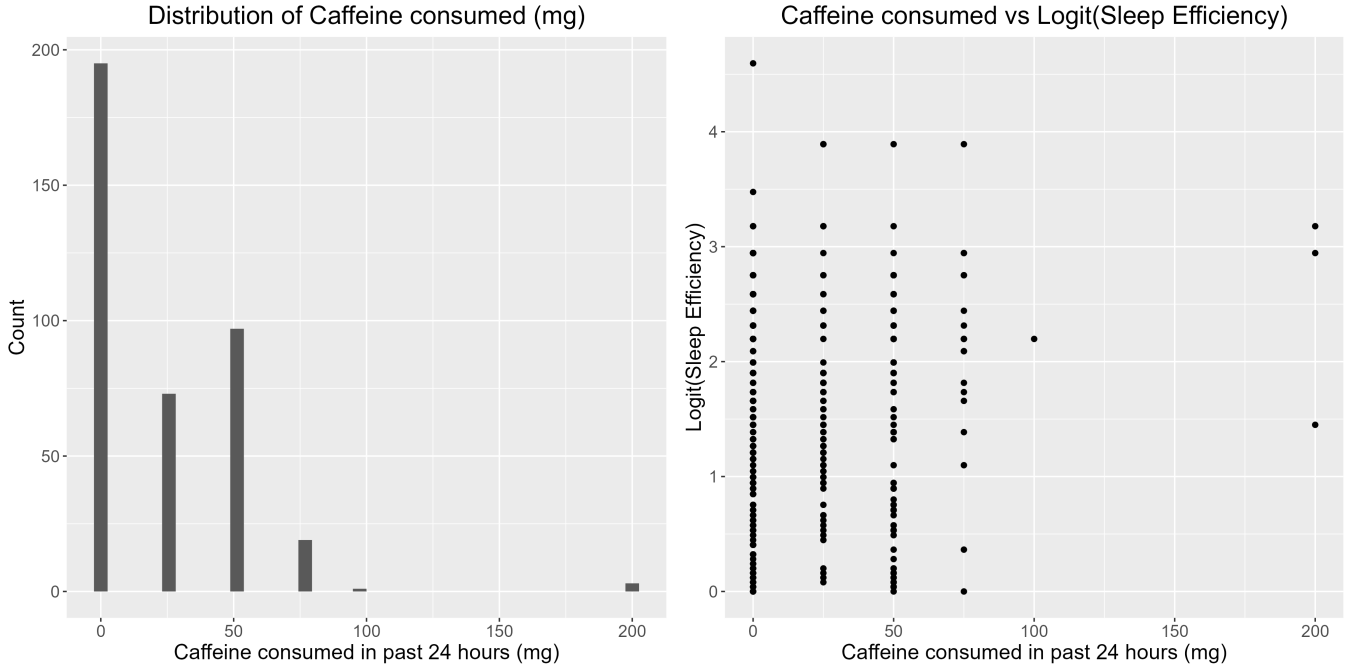


Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
0.000	1.000	1.000	1.619	3.000	4.000	1.360	-0.610	-0.760

Figure 12: Distribution and summary statistics of awakenings.

This denotes the number of times participants awoke during the night. We see that the distribution is right-skewed, with most of the awakenings being 1. Despite this, a significant number of participants did not awake at all during the night.

2.1.14 Caffeine consumption

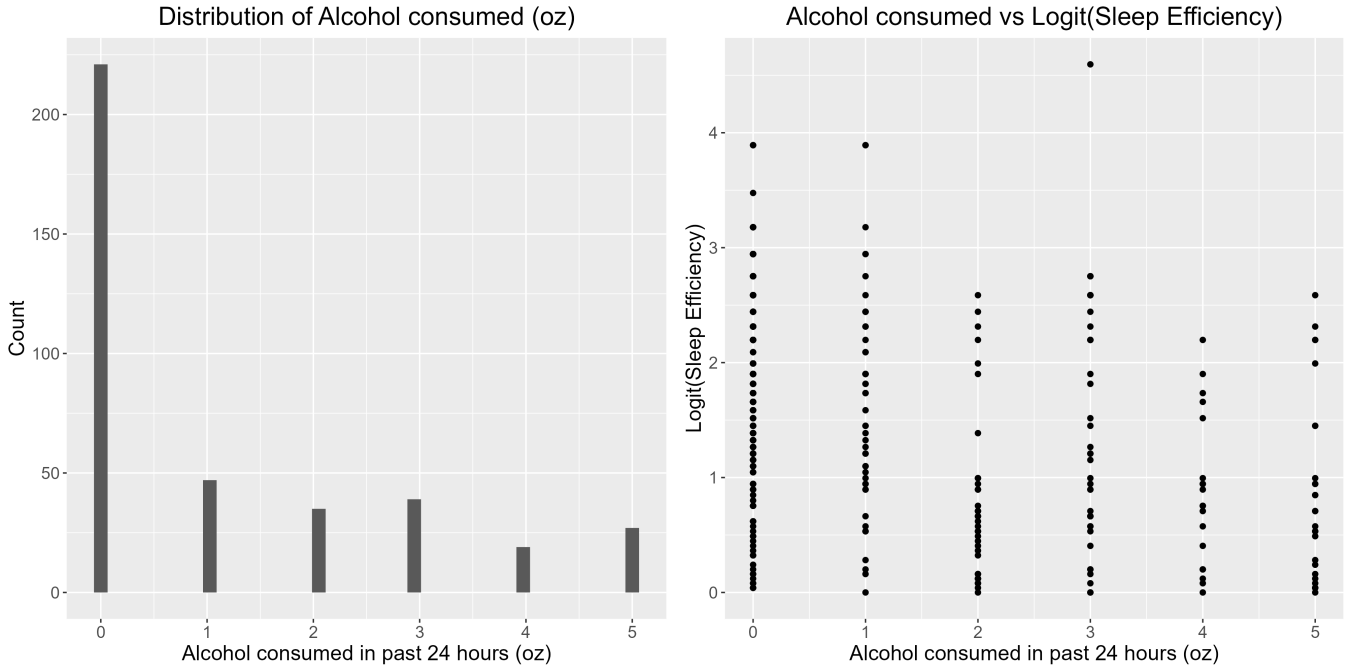


Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
0.000	0.000	0.000	22.680	50.000	200.000	29.000	0.090	2.300

Figure 13: Distribution and summary statistics of caffeine consumption.

The distribution of caffeine consumption has an evident right skew and appears to belong to an exponential family. We notice that in the data-set, most participants do not consume any caffeine in the past 24 hours, with several participants consuming exceptionally more than the others.

2.1.15 Alcohol consumption



Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
0.000	0.000	0.000	1.147	2.000	5.000	1.610	-0.360	-0.530

Figure 14: Distribution and summary statistics of alcohol consumption.

The distribution of alcohol consumption has an evident right skew, and excluding 0 oz, we see a relatively uniform distribution from 1 to 6 oz of alcohol. The 0 oz peak is quite large however, and we note that most participants did not consume any alcohol in the past 24 hours.

2.1.16 Smoking status

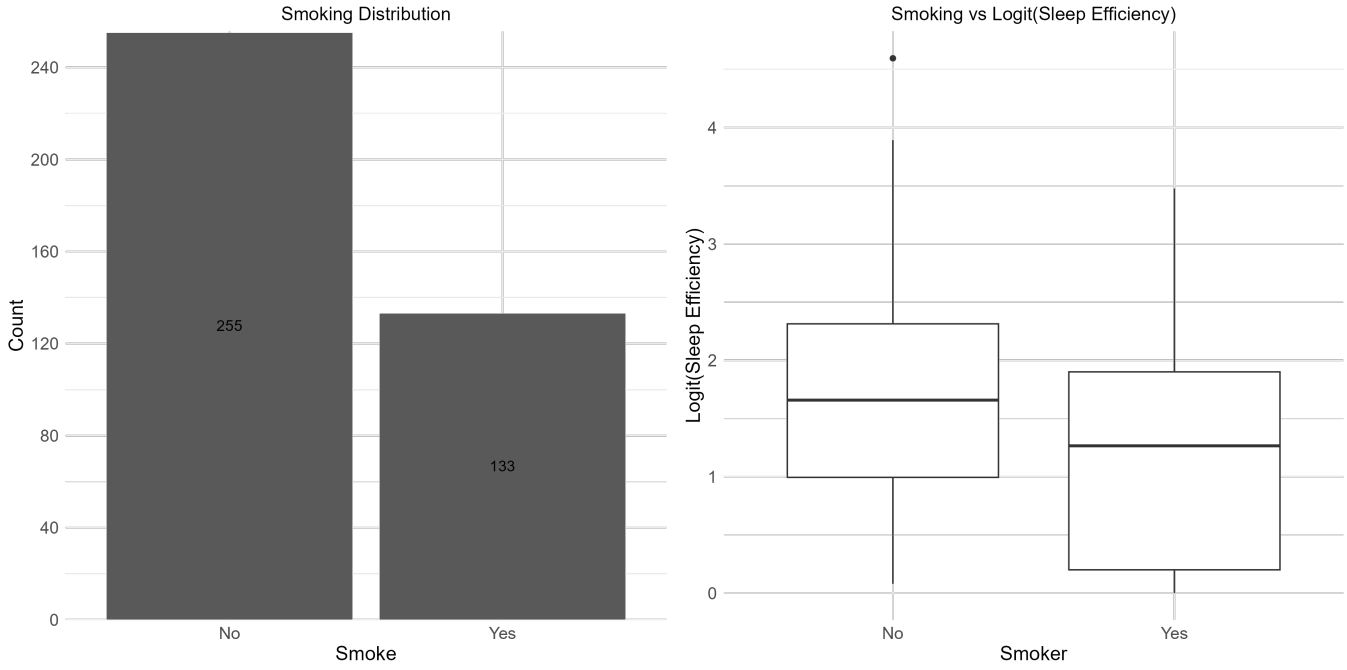
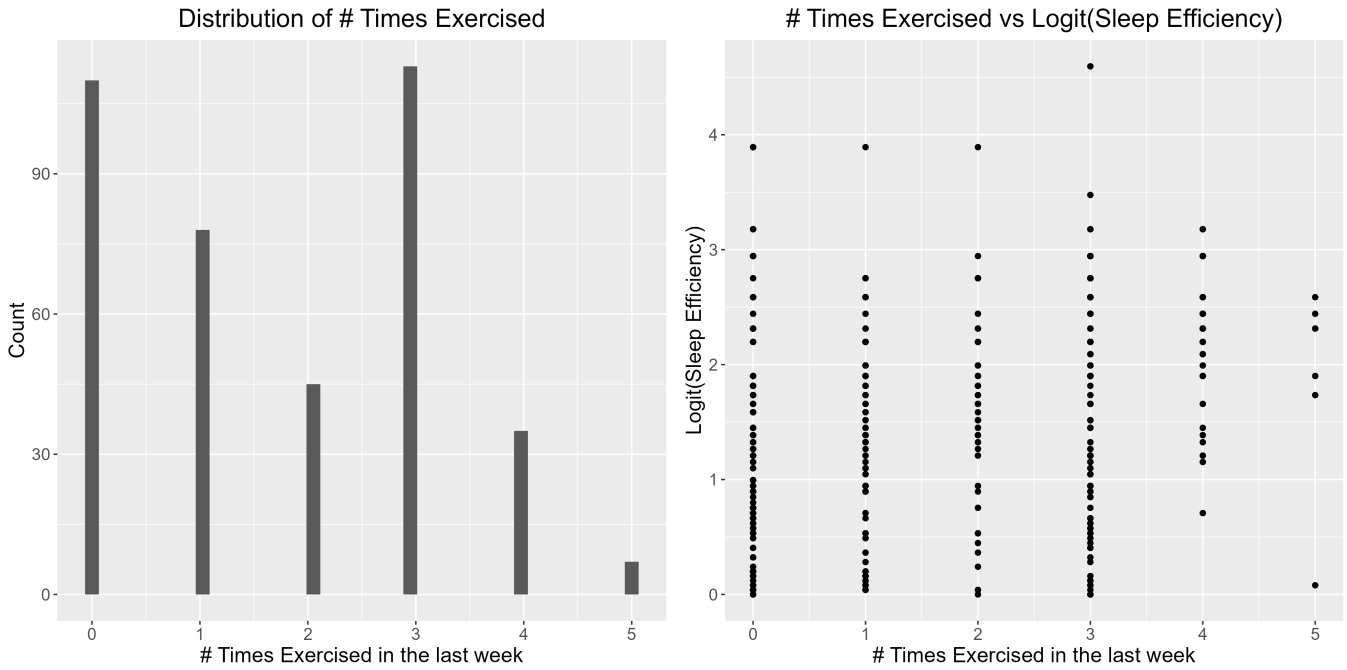


Figure 15: Count and side-by-side boxplots depicting sleep efficiency against smoking status.

We see that a majority of the participants do not smoke, with approximately a third being a smoker. From the boxplots, we note that the non-smoking group has a slightly larger median response than the smoking group. Moreover, the non-smoking group has a smaller inter-quartile range than the smoking group.

2.1.17 Exercise level



Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation	Correlation	Covariance
0.000	0.000	2.000	1.758	3.000	5.000	1.450	0.280	0.370

Figure 16: Distribution and summary statistics of exercise level.

This variable denotes the number of times the subjects exercised in the past week, and we notice that most participants exercised less than 2 or less times per week.

2.2 Methodology

Given the above visualizations and notes, we have chosen to exclude the following variables when attempting to model the response.

1. Unique subject identifier,
2. Wake-up time,
3. Deep sleep percentage.

The reason for excluding the unique subject identifier is because it is an artificially manufactured metric. It is arbitrarily chosen, so it should have no association with sleep efficiency. The reason for the exclusion of wake-up time and deep sleep percentage is due to collinearity with other explanatory variables. Namely, wake-up time can be computed by adding sleep duration to bedtime, and deep sleep percentage can be computed by subtracting REM and light sleep percentage from 100%.

It is also important to note that we aim to fit a linear model on the logit-transformed sleep efficiency values. Because sleep efficiency is a percentage value in the domain $(0, 1)$, we cannot guarantee a linear model will be constrained to this domain. In order to reconcile this problem, we perform the following transform to the response,

$$\text{logit}(y) = \log\left(\frac{y}{1-y}\right),$$

so we can fit response values to the real line.

In order to determine a reasonable model, we perform exhaustive search and pick the model that maximizes R^2 and Adjusted R^2, has a C_p closest to p, and minimizes the Bayesian Information Criterion (BIC). Figure 17 summarizes the results of the exhaustive search algorithm.

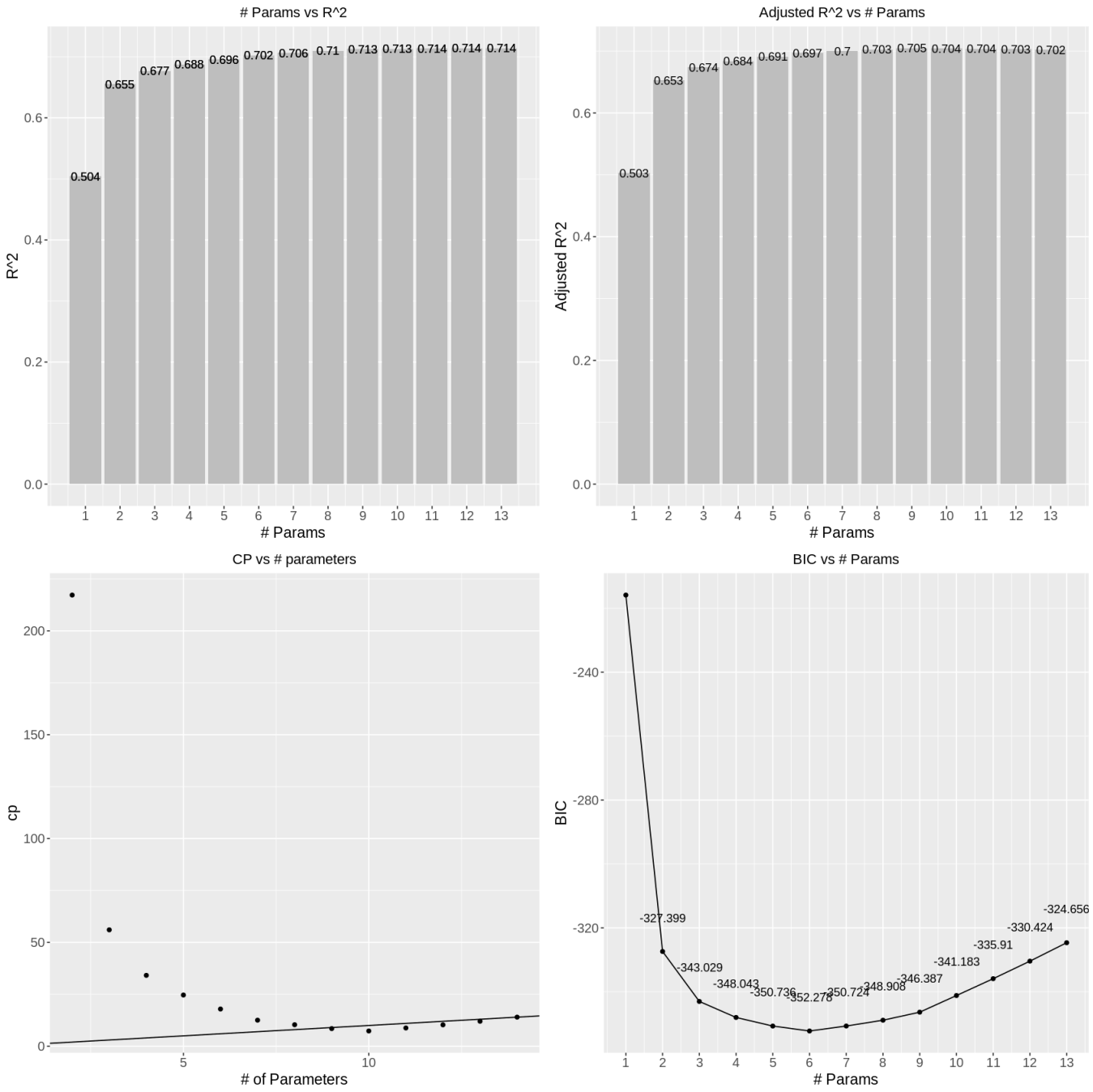


Figure 17: Four plots comparing various summary statistics of the best models given the number of parameters. For all plots except the C_p plot, the number of parameters indicates the number of coefficients in the fitted model, excluding the intercept. In the C_p plot, p includes the intercept. The top left plot indicates the R^2 statistics, in which we see diminishing returns beyond the 8 parameter model. The top right plot indicates the adjusted R^2, which also shows diminishing returns beyond the 8 parameter model. The bottom left plot compares the C_p statistics, in which the 9-parameter model appears favorable due to $C_9 \approx 9$. Finally, the bottom right plot indicates the BIC of the different models, in which favors the 6 parameter model.

Based on the results, it appears that there are a few reasonable candidates for selecting a model. Given the fact that we want to maximize adjusted R^2, it appears that any model of the models greater or equal to 8 parameters is favorable. From the C_p plot, we see that the model with 8 parameters (9 if you include the intercept), is favorable because $C_9 \approx 9$. Finally, the BIC plot favors the model with 6 parameters, albeit the trough centered around 6 parameters is not very steep. Given all of these considerations, coupled with the desire to choose a simple model, we

have chosen to use an 8 parameter model (9 with the intercept). While the 7 parameter model is also a very viable choice, the 8 parameter model's better performance in adjusted R^2 and C_p are the main driving factors in why we chose it.

2.3 Model Analysis

Denoting the i^{th} observation of sleep efficiency as y_i, we propose the following linear model for $\text{logit}(y_i)$,

$$\text{logit}(y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \beta_5 x_{5,i} + \beta_6 x_{6,i} + \beta_7 x_{7,i} + z_1 \beta_8 + \varepsilon_i, \tag{1}$$

where we assume that the error terms ε_i are independent and identically distributed with distribution $N(0, \sigma^2)$ for some unknown σ^2. Descriptions of the β terms are summarized in table 2.

Variable	Identity
β_0	Intercept term
β_1	Age parameter
β_2	REM sleep % parameter
β_3	Light sleep % parameter
β_4	Awakenings parameter
β_5	Caffeine parameter
β_6	Alcohol consumption parameter
β_7	Exercise level parameter
β_8	Smoking status parameter

Table 2: Table of β values in proposed model of $\text{logit}(y_i)$

Note that $x_{j,i}$ represents the covariate corresponding to β_j. We also define the dummy variable,

$$z_1 = \begin{cases} 1 & \text{when smoking status is true,} \\ 0 & \text{when smoking status is false.} \end{cases}$$

Given this proposed model, we can now create the fitted model,

$$\widehat{\text{logit}}(y_i) = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \widehat{\beta_3} x_{3,i} + \widehat{\beta_4} x_{4,i} + \widehat{\beta_5} x_{5,i} + \widehat{\beta_6} x_{6,i} + \widehat{\beta_7} x_{7,i} + z_1 \widehat{\beta_8}, \tag{2}$$

where $\widehat{\beta_j}$ represents the least-squares estimate from a linear regression. The values of $\widehat{\beta_j}$ are summarized in table 3.

Variable	Fitted value	p-value
$\widehat{\beta_0}$	2.107	$< 2 \times 10^{-16}$
$\widehat{\beta_1}$	0.007	6.08×10^{-4}
$\widehat{\beta_2}$	0.015	0.041
$\widehat{\beta_3}$	-0.030	$< 2 \times 10^{-16}$
$\widehat{\beta_4}$	-0.271	$< 2 \times 10^{-16}$
$\widehat{\beta_5}$	0.002	0.057
$\widehat{\beta_6}$	-0.039	0.024
$\widehat{\beta_7}$	0.052	0.005
$\widehat{\beta_8}$	-0.299	8.98×10^{-8}

Table 3: Table of fitted $\widehat{\beta}$ values in proposed model.

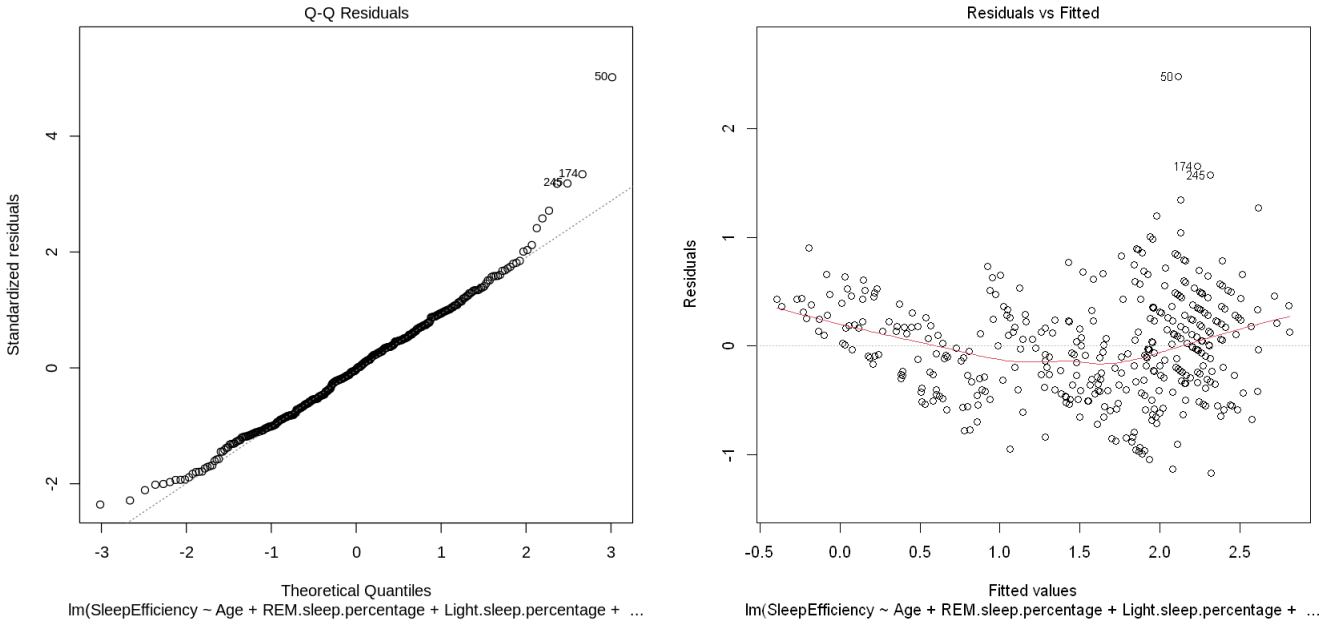
In calculating the p-values for estimates of the parameters, we utilized the mean sum of squares of the residuals to estimate the variance σ^2. This value came out to be $\text{MSS}_{residual} = s^2 = \widehat{\sigma^2} = (0.499)^2$. From the table, all parameters are significant at a 5% significance level except for $\widehat{\beta_5}$. Our model suggests that holding all other variables constant, age, REM sleep percentage, and exercise frequency have a significant positive association on sleep efficiency, while light sleep, awakenings, alcohol consumption and smoking all have negative association. Note that some coefficients are small relative to the estimated variance of the error term: the residual standard error has higher magnitude than that of smoking and comparable to awakening twice.

Our model also suggests that the time a person goes to bed, gender, caffeine consumption, and duration of sleep has an insignificant association.

We note that $\widehat{\beta_0}$ represents the baseline logit response value when the status is non-smoking, and because of the dummy variable, $\widehat{\beta_0} + \widehat{\beta_8}$ represents the baseline response value when the status is smoking. In other words by smoking, the logit response decreases by $\widehat{\beta_8} = 0.299$. For all of the other coefficients $\widehat{\beta_j}$, if all other $x_{k,i}$ with $k \neq j$ are held constant, a unit increase in $x_{j,i}$ corresponds to an increase of $\widehat{\beta_j}$ in $\text{logit}(y_i)$. How this relates directly to sleep efficiency y_i is not very difficult to interpret. For every increase in $\text{logit}(y_i)$, y_i increases by $\frac{e}{1+e}$. This fact ensures that an increase in the logit transform corresponds to an increase in the un-transformed response, so the following model interpretations apply.

Because maximizing the logit of the response is equivalent to maximizing the response itself, our model provides some insight on choices one can make to improve their sleep. Firstly, smoking has a substantial negative coefficient (comparable to about half of a standard deviation), so the model recommends, if all other factors are held constant, to not smoke. Similarly, the model also recommends if other factors are held constant, reducing alcohol consumption or minimizing the number of times awoken may also lead to higher sleeping efficiency.

2.3.1 Potential Improvements



We notice the QQ plot (left) residuals fall close to the line, with a slight right skew (potentially from outliers). As most residuals within two quantiles from 0 fall near the line, we believe the assumption of normality to be valid. However, our model may still be improved by either removing outliers that may have caused partial skew or performing a different transformation.

A trend can also be seen in the residuals vs fitted plot (right). This indicates that the data may have underlying non-linear patterns, that our model cannot capture.

There is some partial colinearity among predictors. For instance, exercise frequency is negatively correlated with light sleep percentage and awakenings. exhaustive search and ordinary least squares does not perform well with

colinearity. One potential future improvement is to perform L1 regularization or LASSO to reduce colinearity while performing variable selection.

Interactions were not included as exhaustive search recommended including the interaction term without the additive offset, and so neither terms were included.

3 Conclusion

We analyzed various predictors of sleep efficiency, and fit a model using exhaustive search. With our selected model, we found that an increase in age, proportion of REM sleep, and exercise are predicted to positively increase sleep efficiency holding all others constant, while an increase in light sleep, awakenings, alcohol consumption or smoking are predicted to decrease sleep efficiency holding all other variables constant. While we cannot conclude a causal relationship, our model suggests that being older, exercising at least a few times a week while not drinking alcohol or smoking will be ideal to improve a person's sleep efficiency. Students should consider reducing alcohol and tobacco consumption and increasing exercise to attempt to improve their sleep efficiency. We do note however that the model could have some improvements. There is potential colinearity among predictors, and interaction terms not being fit due to the additive terms not being significant. Further exploring these potential improvements could reveal more associations and potentially provide further insight on sleep efficiency.

References

[Ped+14] Paola Pedrelli et al. "College Students: Mental Health Problems and Treatment Considerations". In: *Academic Psychiatry* 39 (Aug. 2014). DOI: 10.1007/s40596-014-0205-9.

[RS16] David L. Reed and William P. Sacco. "Measuring Sleep Efficiency: What Should the Denominator Be?" In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 12 2 (2016), pp. 263–6. URL: https://api.semanticscholar.org/CorpusID:45042765.

[Oka+19] Kana Okano et al. "Sleep quality, duration, and consistency are associated with better academic performance in college students". In: *npj Science of Learning* 4 (Dec. 2019). DOI: 10.1038/s41539-019-0055-z.

[Sco+21] Alexander J. Scott et al. "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials". In: *Sleep Medicine Reviews* 60 (2021), p. 101556. ISSN: 1087-0792. DOI: https://doi.org/10.1016/j.smrv.2021.101556. URL: https://www.sciencedirect.com/science/article/pii/S1087079221001416.