

**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

MH3511 Data Analysis with Computer

Group Project

Sleep Efficiency

Name	Matriculation Number
Ho Inn Jong, Jereme	U1940727B
Lim Jun Guang	U2040237B
Poh Zi Jie Isaac	U2140416E
Stephen Andrew Gudijanto	U2140974G
Woon Zhing Wen	U2140025B

(All students equally contributed to each aspect of the project)

Abstract:

Sleep plays a pivotal role in maintaining our physical and mental health. Sleep deprivation often results in impulsive decision making, lower capability to learn and retain information, poor emotional health and more. Though various studies have been done to investigate factors that could affect sleep efficiency, the significance of association between sleep efficiency and those factors remain inconclusive. Therefore, we would like to examine the relationship between sleep efficiency and other biological or behavioural factors through basic data analysis techniques.

Table of Contents

1. Introduction	1
2. Data Description	2
3. Description and Cleaning of Dataset	3
3.1 Summary Statistics for the main variable of interest, sleep efficiency	3
3.2 Summary statistics for the other variables	4
3.2.1 Age	4
3.2.2 Gender	4
3.2.3 Bedtime	4
3.2.4 Wake-up Time	5
3.2.5 Sleep Duration	5
3.2.6 Awakenings	5
3.2.7 Caffeine Consumption	6
3.2.8 Alcohol Consumption	6
3.2.9 Smoking Status	6
3.2.10 Exercise Frequency	7
3.3 Final Dataset for Analysis	7
4. Statistical Analysis	8
4.1 Correlations between sleep efficiency and other continuous variables	8
4.2 Statistical Tests	9
4.2.1 Relationship between sleep efficiency and gender	9
4.2.2 Relationship between sleep efficiency and age	11
4.2.3 Relationship between sleep efficiency and sleep duration	12
4.2.4 Relationship between sleep efficiency and awakenings	13
4.2.5 Relationship between sleep efficiency and alcohol consumption	14
4.2.6 Relationship between sleep efficiency and caffeine consumption	15
4.2.7 Relationship between sleep efficiency and smoking status	16
4.2.8 Relationship between sleep efficiency and exercise frequency	17
4.2.9 Relationship between sleep efficiency and bedtime	20
4.2.10 Most important factor affecting sleep efficiency	21
4.3 Multiple Linear Regression	21
5. Conclusion and Discussion	23
6. References	24
7. Appendix	25

1. Introduction

Sleep is an essential part of a healthy lifestyle, and the quality of sleep is often measured by the sleep efficiency or the total sleep duration. However, many factors such as gender, age, bedtime, wake-up time, REM sleep percent, deep sleep percent, light sleep percent, caffeine consumption, alcohol consumption, smoking status, and exercise frequency affect our quality of sleep.

In our project, a dataset containing the sleep efficiency of participants is used, including other variables such as consumption of caffeinated and alcoholic drinks, sleep duration, quality of sleep and lifestyle habits. Based on the dataset, we wish to understand the following questions regarding sleep efficiency:

1. Does gender affect sleep efficiency?
2. Does age affect sleep efficiency?
3. Is sleep duration a good indicator of sleep efficiency?
4. Is awakenings a good indicator of sleep efficiency?
5. Does alcohol consumption directly cause reduced sleep efficiency?
6. Does caffeine consumption directly cause reduced sleep efficiency?
7. Does smoking affect sleep efficiency?
8. Is there a relation between the number of times a person exercises in a week and sleep efficiency?
9. Does early bed time translate into better sleep efficiency?
10. What is the most important factor that significantly affects sleep efficiency?

The report incorporates data descriptions and analyses using the R language. Statistical analysis is conducted on each research objective and conclusions were supported by the most suitable approach, explanations and elaborations.

2. Data Description

The dataset for the sleep patterns of a group of test subjects is obtained from Kaggle, an online community of data scientists and machine learning engineers. The original dataset, titled "Sleep_Efficiency.csv", contains 452 observations (test subjects) of 15 variables (categorical and numerical). The dataset was collected as part of a study conducted in the UK by a research team from The University of Oxfordshire. Currently, the dataset is open to the public for further studies and research.

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns were eliminated, i.e. ID, REM sleep percentage, light sleep percentage and deep sleep percentage were removed as we only focus on the sleep efficiency.
- Redundant information were removed, i.e. the date under bedtime and wakeup time were removed as we only need the time for analysis.
- Age groups with less than 10 subjects (i.e. children and teenagers) are treated as unrepresentative anomalies and excluded.
- Only adults' (age 18 and above) data are included in our dataset.
- Bedtime and wakeup time are converted into numeric types in hours.

After data preparation, 379 observations (test subjects) with 11 variables are retained for analysis:

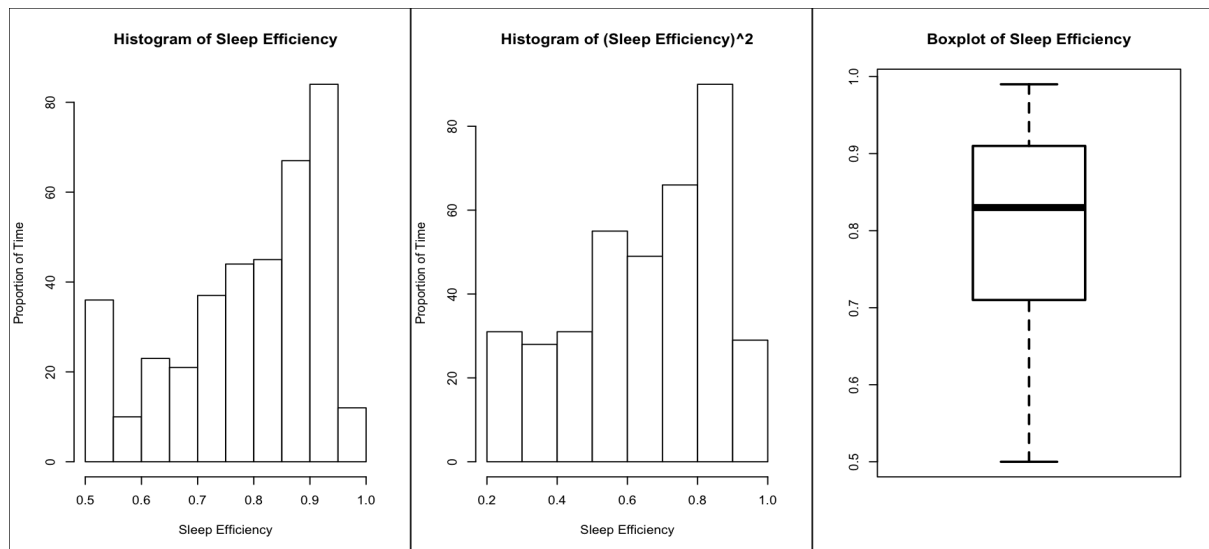
1. Age: age of the test subject
2. Gender: gender of the test subject
3. Bedtime: the time each subject goes to bed each day
4. Wakeup.time: the time each subject wakes up each morning
5. Sleep.duration: the total amount of time each subject slept in hours
6. Sleep.efficiency: the proportion of time spent in bed that is actually spent asleep
7. Awakenings: the number of times each subject wakes up during the night
8. Caffeine.consumption: the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
9. Alcohol.consumption: the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
10. Smoking.status: whether or not the test subject smokes
11. Exercise.frequency: the number of times the test subject exercises each week

3. Description and Cleaning of Dataset

In this section, we examined the data in greater detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data if applicable.

3.1 Summary Statistics for the main variable of interest, *sleep efficiency*

The following plots describe the overall distribution of the main variable *sleep efficiency*:



It appears that the main variable *sleep efficiency* is highly skewed to the left, hence we applied squared-transformation to the variable. However, the distribution of *sleep efficiency* is still skewed after transformation. Thus we will use the original data of *sleep efficiency* for the subsequent analyses. No outlier is detected from the original data of *sleep efficiency* as shown in the boxplot above. The summary statistics of original *sleep efficiency* is shown below.

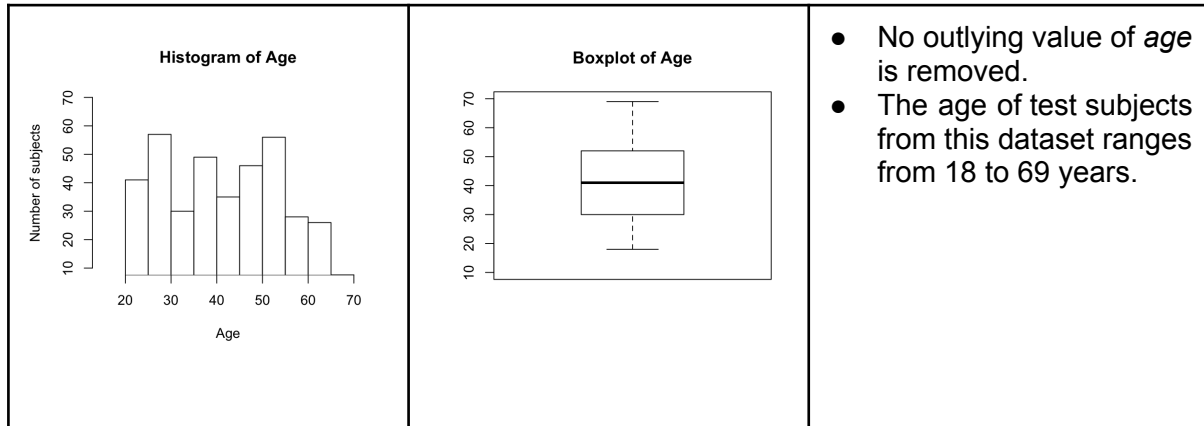
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.500	0.710	0.830	0.794	0.910	0.990

By Central Limit Theorem, we can assume that the sample mean of our main variable *sleep efficiency* is approximately normally distributed since our dataset has a large sample size of 379 (greater than 30).

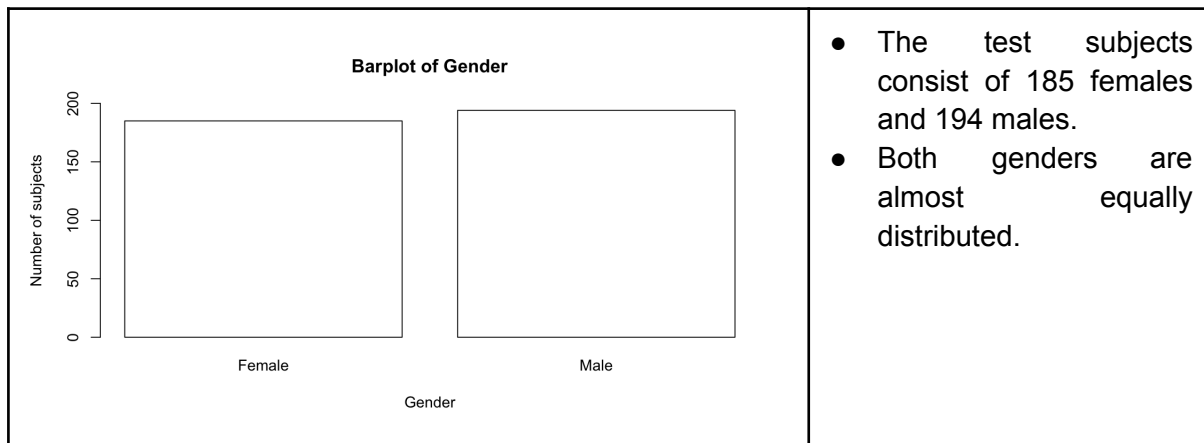
3.2 Summary statistics for the other variables

The histogram, the barplot, the boxplot applied and the outliers identified from the variables are tabulated in the following subsections. Some summary statistics and observations about each of the variables are also recorded.

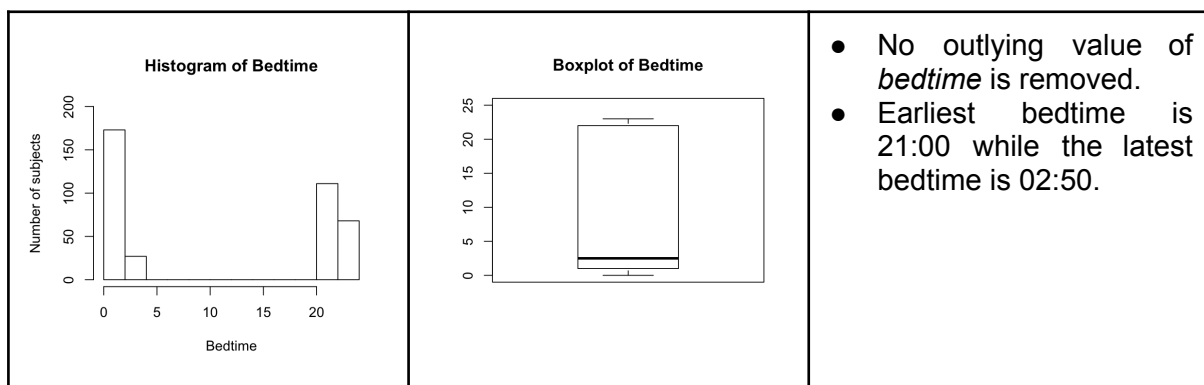
3.2.1 Age



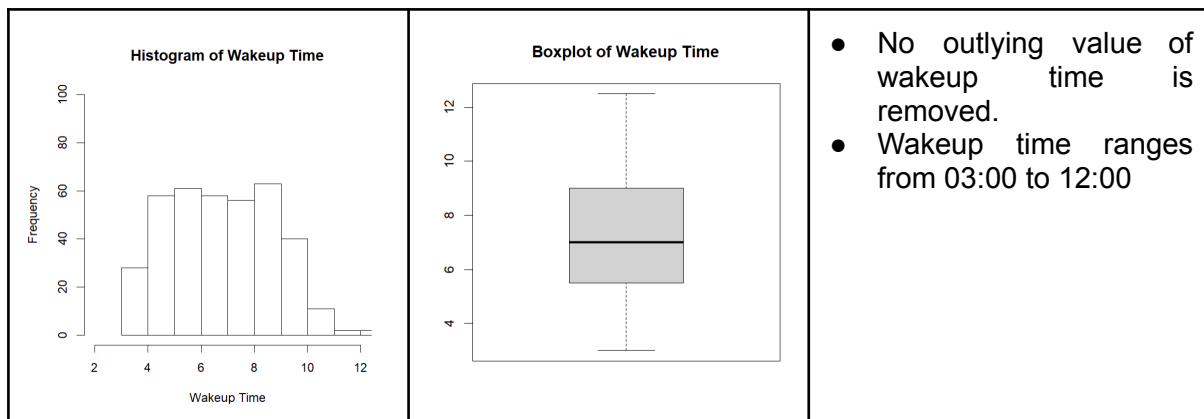
3.2.2 Gender



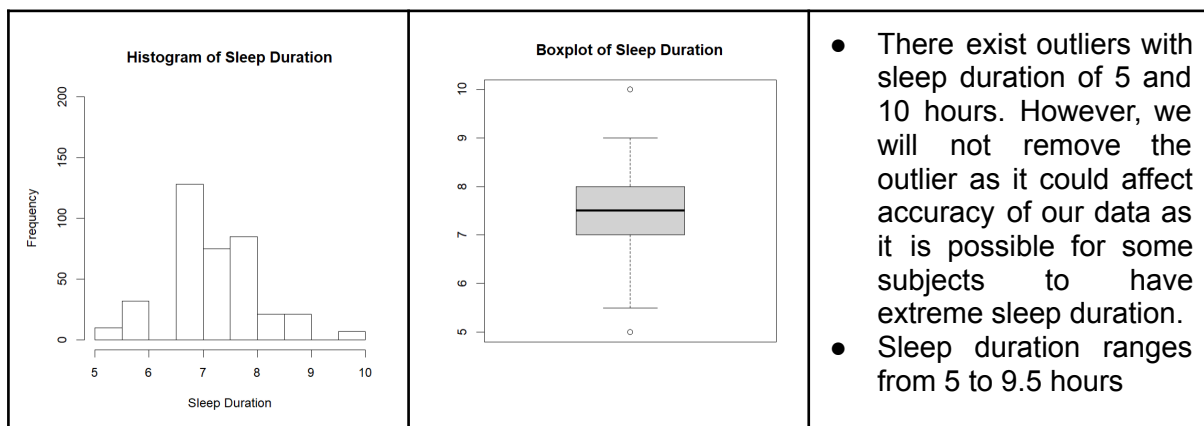
3.2.3 Bedtime



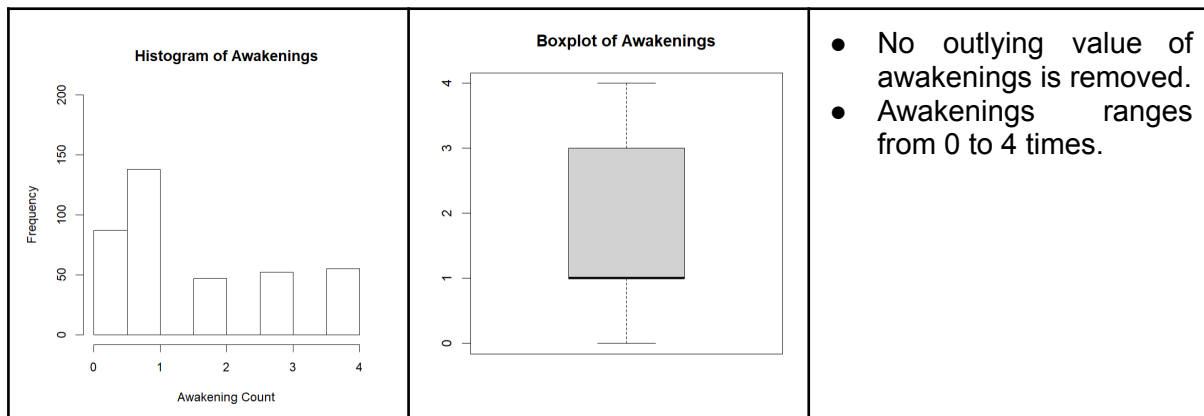
3.2.4 Wake-up Time



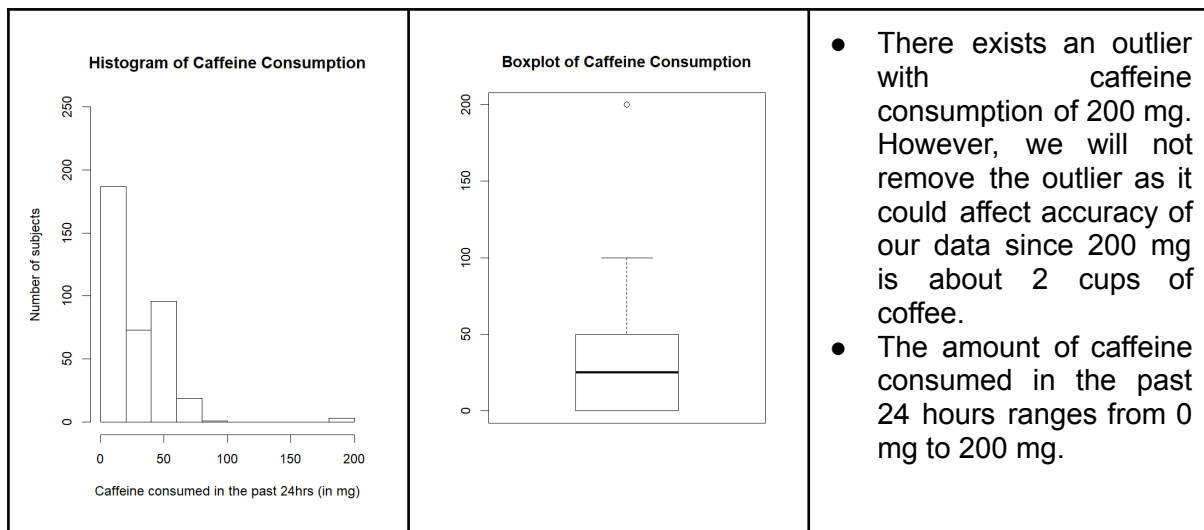
3.2.5 Sleep Duration



3.2.6 Awakenings



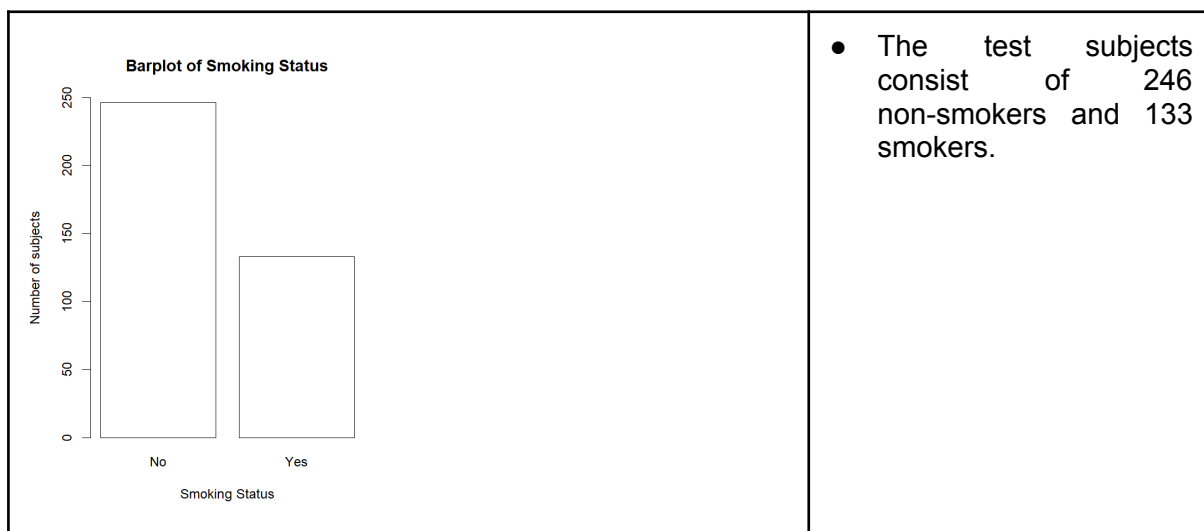
3.2.7 Caffeine Consumption



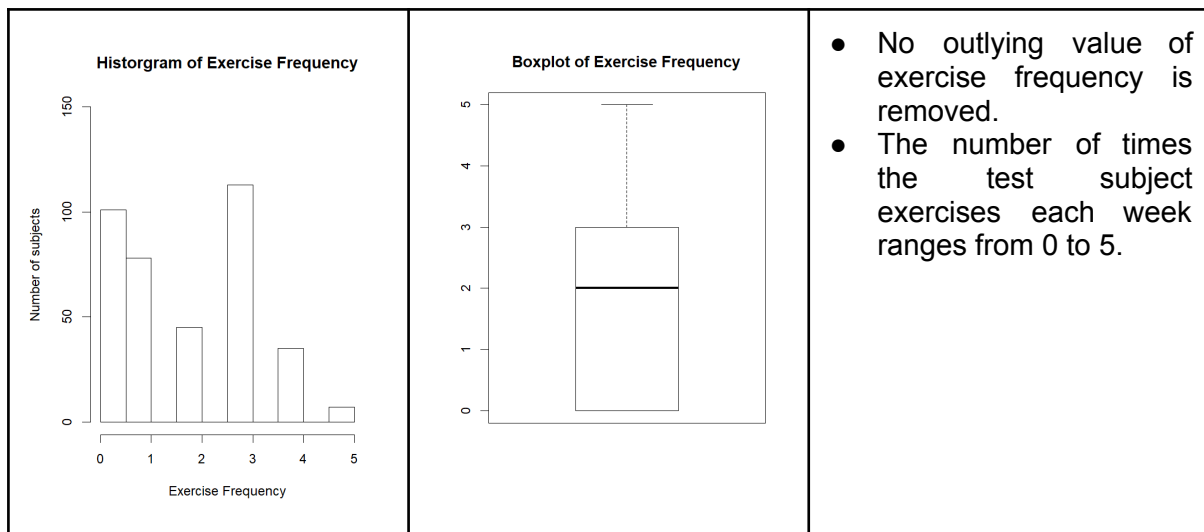
3.2.8 Alcohol Consumption



3.2.9 Smoking Status



3.2.10 Exercise Frequency

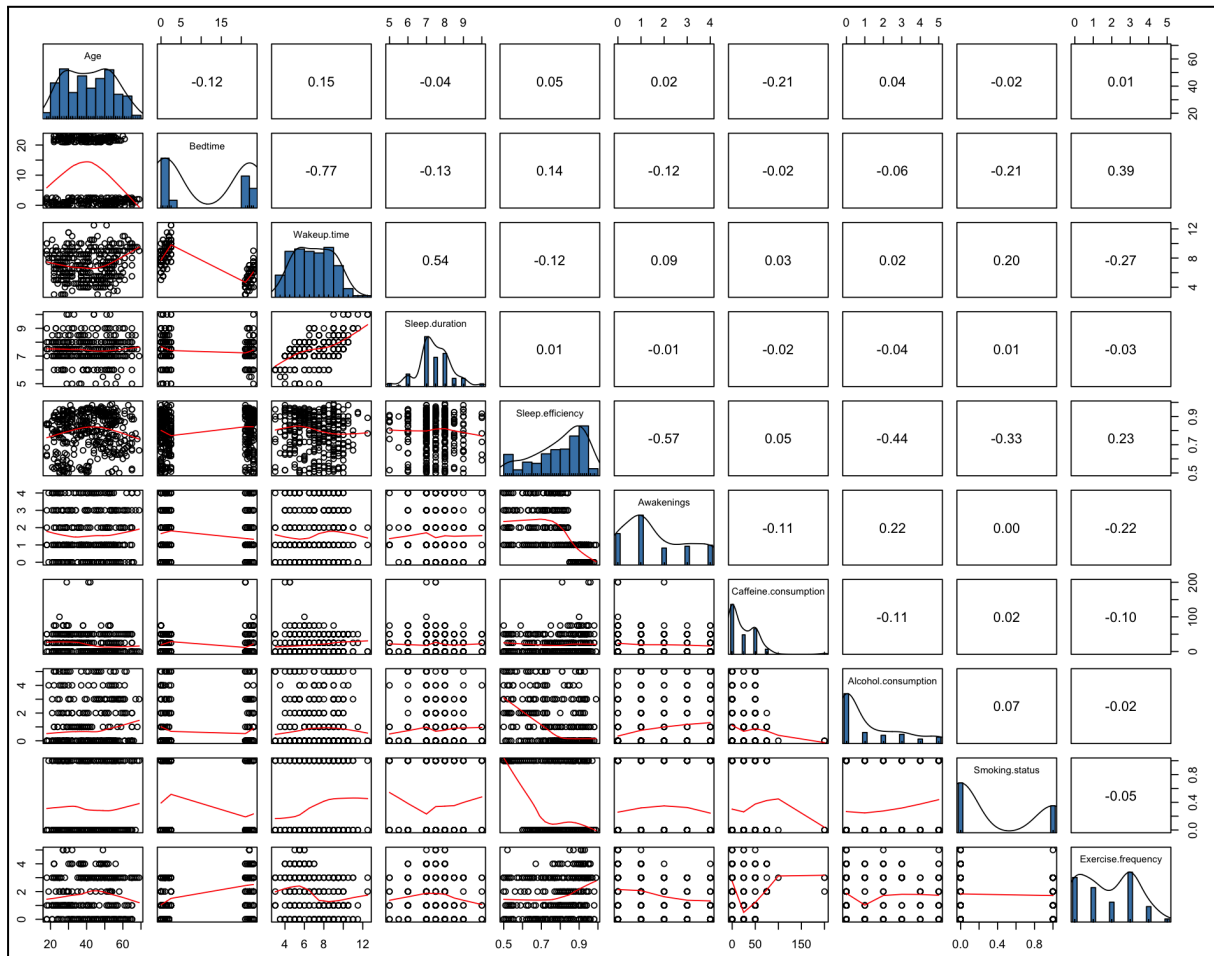


3.3 Final Dataset for Analysis

Based on the above analysis, no outlying value is removed from any of the variables in the dataset. Thus the number of observations remains unchanged, which is still 379 observations.

4. Statistical Analysis

4.1 Correlations between sleep efficiency and other continuous variables



Scatter plots and correlation coefficients are useful in studying the possible linear relationships between a subject's sleep efficiency and other variables.

Based on the correlation plot above, we observed that *sleep efficiency* is not highly correlated to any other variables except for highly negatively correlated to *awakenings*.

The following are some interesting observations among the other variables of interest from this plot:

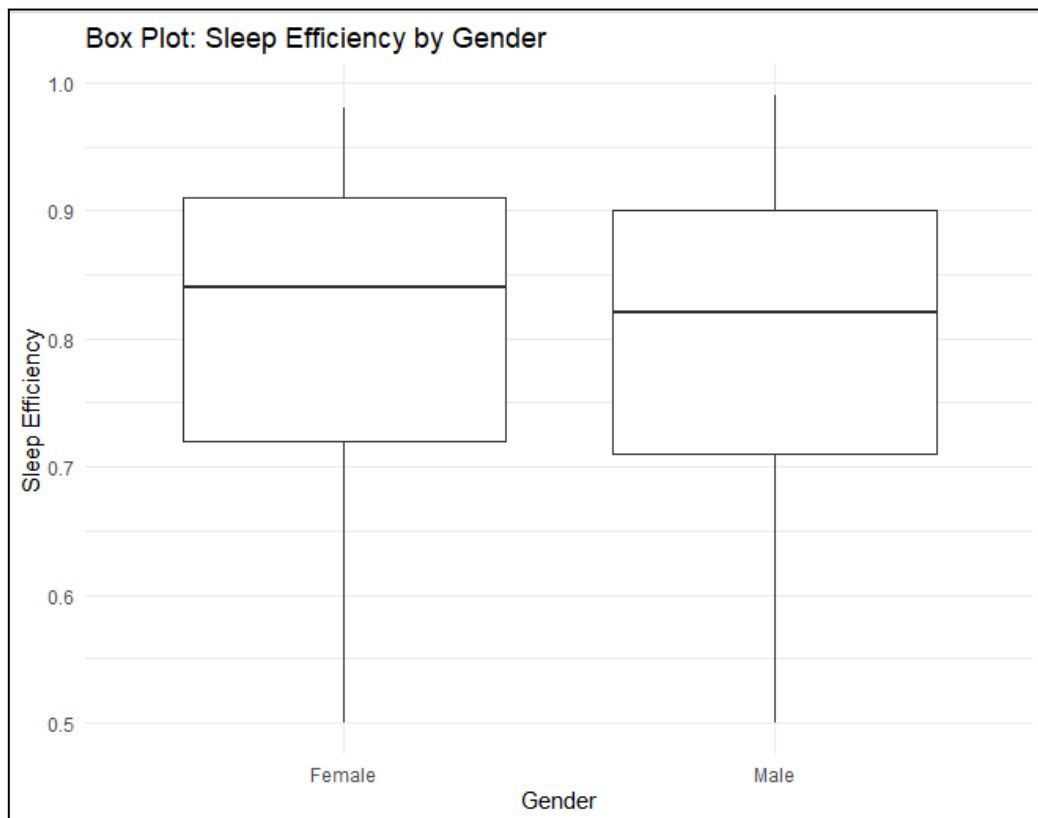
- Wake up Time and Sleep Duration are positively correlated ($r = 0.54$)
- Bedtime is highly negatively correlated to Wake Up Time ($r = -0.77$)
- Sleep Efficiency and Awakenings are negatively correlated ($r = -0.57$)

Next, we shall perform some statistical tests to confirm these observations.

4.2 Statistical Tests

4.2.1 Relationship between *sleep efficiency* and *gender*

In this section, we determine whether there is a significant difference in the sleep efficiency between Males and Females. We ensure that the data follows the assumptions of being normally distributed and having a similar variance before comparing the 2 datasets using `t.test()`.



F-test was performed using `var.test` to compare the variances of Sleep Efficiency between Males and Females.

H_0 : Variances of males' and females' sleep efficiency are equal

H_1 : Variances of males' and females' sleep efficiency are equal

```
F test to compare two variances

data:  sleep.efficiency by Gender
F = 1.118, num df = 184, denom df = 193, p-value = 0.444
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8399144 1.4895317
sample estimates:
ratio of variances
 1.117962
```

A variance ratio of 1.1 is reasonably close to 1, indicating that the variances of the two groups are not substantially different.

To test for normality in the sleep efficiency data for males and females separately, the `shapiro.test()` function was used for each gender.

```
> # For males
> male_data <- clean_sleep[clean_sleep$Gender == "Male",]
> shapiro.test(male_data$sleep.efficiency)

      shapiro-wilk normality test

data:  male_data$sleep.efficiency
W = 0.93168, p-value = 6.73e-08

> # For females
> female_data <- clean_sleep[clean_sleep$Gender == "Female",]
> shapiro.test(female_data$sleep.efficiency)

      shapiro-wilk normality test

data:  female_data$sleep.efficiency
W = 0.89534, p-value = 4.053e-10
```

Since the $p\text{-value} < 0.05$, we can assume that both sets of data are normally distributed. This is aligned with our assumptions based on the Central Limit Theorem (as mentioned in Section 3.1), whereby the sample mean of *sleep efficiency* is approximately normally distributed as our dataset has a large sample size ($n=379$).

Assuming that the data observed is independent, the other assumptions of normality and equal variances are met. The test is used to determine the difference in sleep efficiency between both genders. The Welch two-sample t-test was conducted to create a 95% confidence interval for the differences in mean of sleep efficiency. The confidence interval is $[-0.01450, 0.3903]$.

```
> t.test(sleep.efficiency ~ Gender, data = clean_sleep)

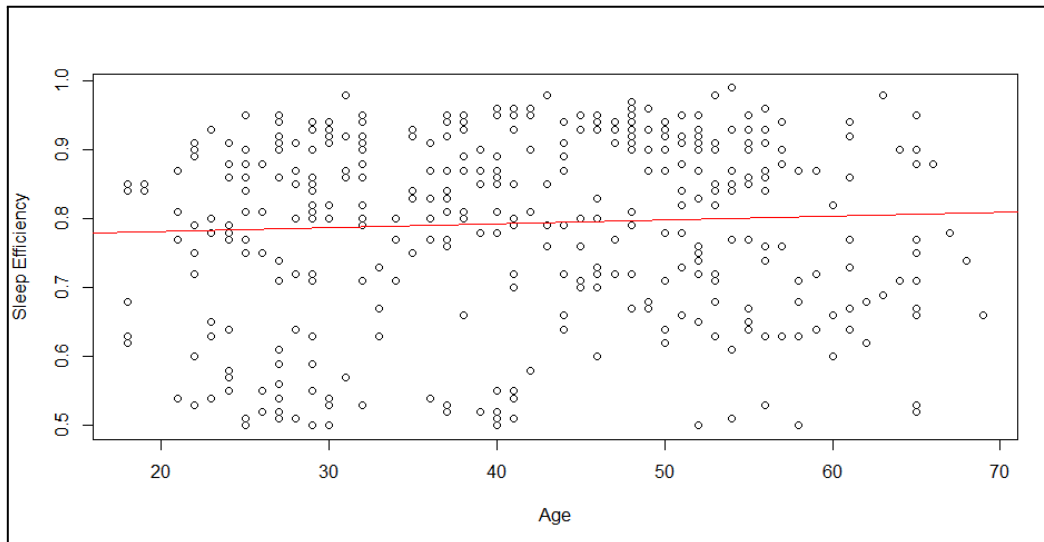
      welch Two Sample t-test

data:  sleep.efficiency by Gender
t = 0.87474, df = 373.02, p-value = 0.3823
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -0.01499742  0.03903308
sample estimates:
mean in group Female    mean in group Male
      0.8001622           0.7881443
```

The $p\text{-value}$ 0.3823 is greater than the significance level of 0.05. This means that the evidence is not strong enough to reject the null hypothesis, so we fail to reject it. The interpretation is that there is no statistically significant difference in Sleep Efficiency between the Female and Male groups, based on the data provided. Thus, gender is not a significant factor that affects sleep efficiency.

4.2.2 Relationship between *sleep efficiency* and *age*

In this section, we will try to answer the question “Is age an indicator of sleep efficiency”. We perform a single linear regression between Sleep efficiency and Age.



```
Call:
lm(formula = sleep.efficiency ~ Age, data = clean_sleep)

Residuals:
    Min       1Q   Median       3Q      Max
-0.30323 -0.08737  0.03297  0.11297  0.19185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7708283   0.0232156  33.203  <2e-16
Age           0.0005587   0.0005346   1.045   0.297

(Intercept) ***
Age
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

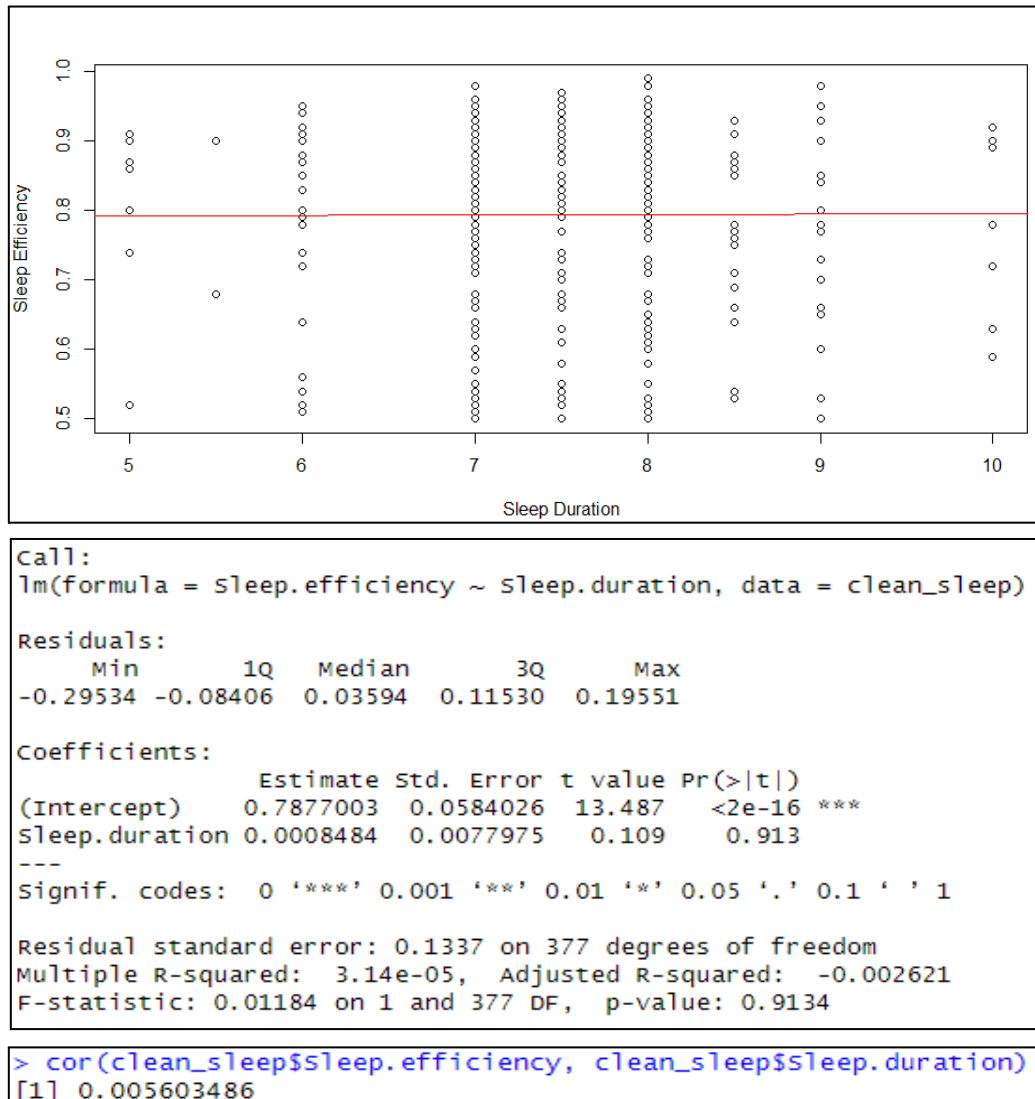
Residual standard error: 0.1335 on 377 degrees of freedom
Multiple R-squared:  0.002889, Adjusted R-squared:  0.0002443
F-statistic: 1.092 on 1 and 377 DF, p-value: 0.2966

> cor(clean_sleep$sleep.efficiency, clean_sleep$Age)
[1] 0.05375133
```

The correlation coefficient (0.05375133) shows a weak positive correlation between sleep efficiency and age. As it is close to 0, the correlation between the 2 variables is negligible. Furthermore, the multiple R-squared value of 0.002889 suggests that only about 0.29% of the variation in sleep efficiency can be explained by age. This indicates that the linear model does not fit the data well, thus, age and sleep efficiency are not linearly related.

4.2.3 Relationship between *sleep efficiency* and *sleep duration*

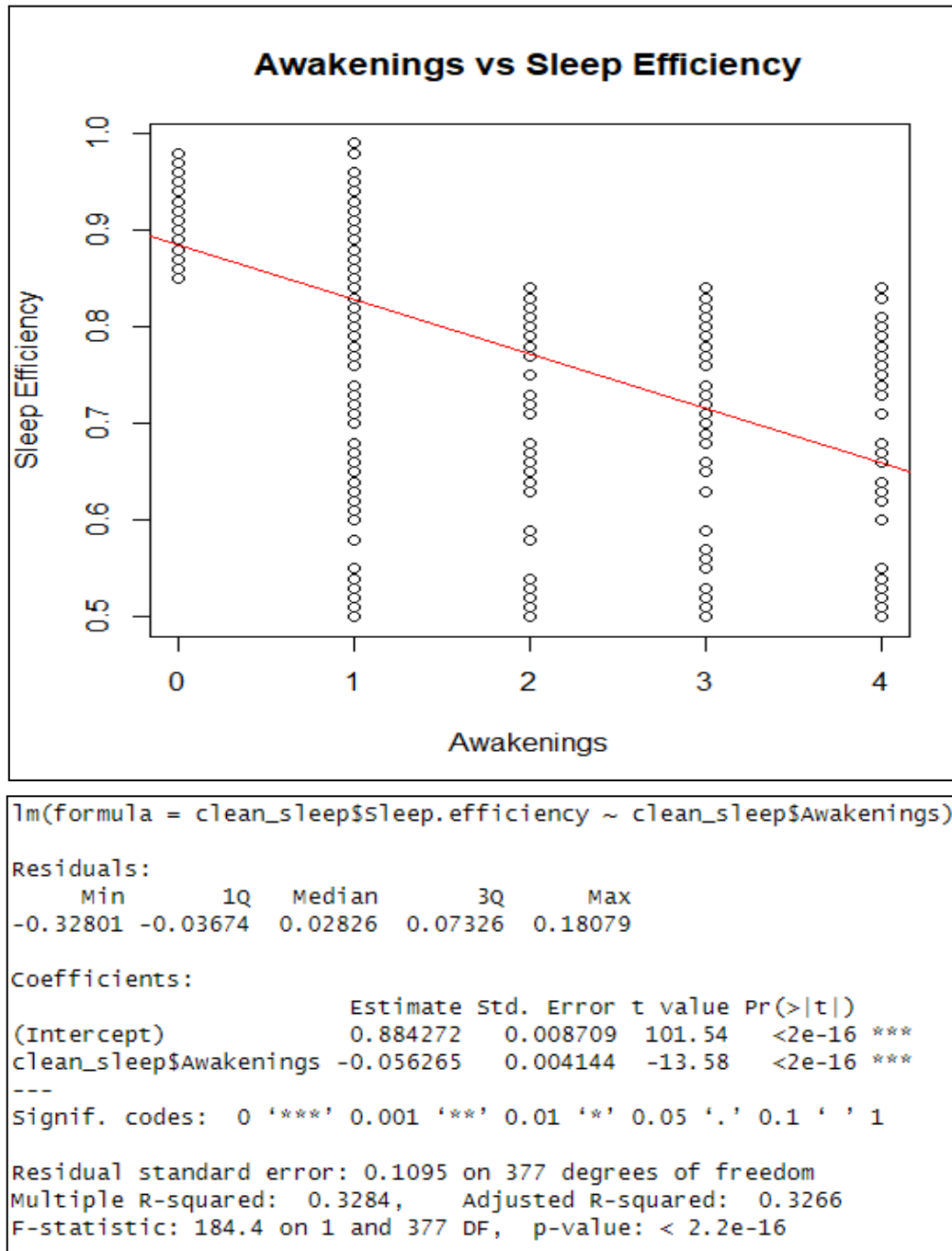
In this section, we are interested to find out whether sleep duration is a good indicator of sleep efficiency. Similar to the previous section, we perform a single linear regression between Sleep efficiency and Sleep duration.



The correlation coefficient (0.005603486) shows a weak positive correlation between sleep efficiency and sleep duration. As it is close to 0, the correlation between the 2 variables is negligible. Furthermore, the multiple R-squared value of 3.14×10^{-5} suggests that only about 0.00314% of the variation in sleep efficiency can be explained by sleep duration, suggesting that the linear model does not fit the data well and that sleep efficiency and age are not linearly related.

4.2.4 Relationship between *sleep efficiency* and *awakenings*

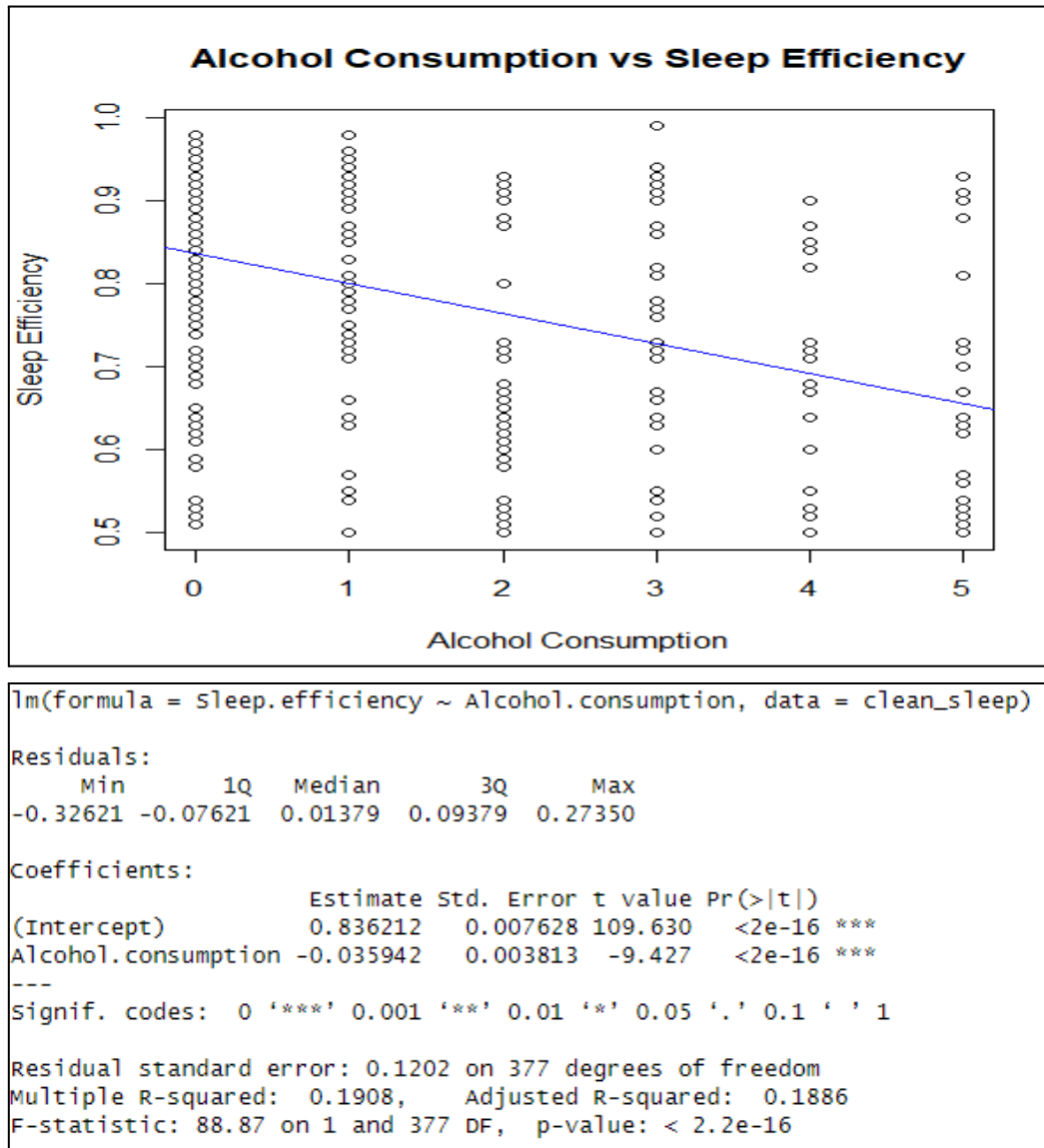
In this section, we study the effect of the number of sleep awakenings on sleep efficiency. We performed a single linear regression between awakenings and sleep efficiency.



Based on the p-value $< 2.2e-16$ (< 0.05), there is strong evidence to reject the null hypothesis. This highlights a statistically significant relationship between awakenings and sleep efficiency. Furthermore, the R Squared value of 0.3284 indicates that Awakenings and sleep efficiency are moderately correlated. This supports that awakenings do have an effect on reducing sleep efficiency.

4.2.5 Relationship between *sleep efficiency* and *alcohol consumption*

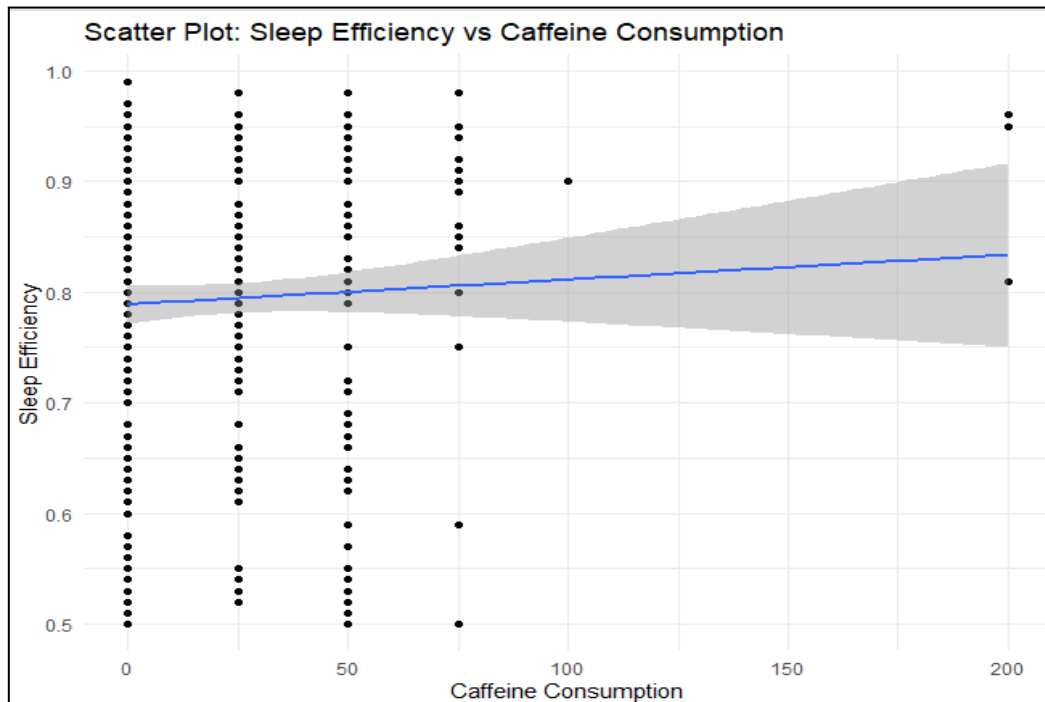
In this section, we conduct a single linear regression between sleep efficiency and alcohol consumption to answer the question of “Does alcohol consumption directly cause reduced sleep efficiency?”.



Based on the p-value <2.2e-16 (< 0.05), there is a strong evidence to reject the null hypothesis. This highlights a statistically significant relationship between alcohol consumption and sleep efficiency. Furthermore, the multiple R-squared value of 0.1908. means that approximately 19.08% of the variability in sleep efficiency can be explained by the model that includes alcohol consumption as an independent variable.

4.2.6 Relationship between *sleep efficiency* and *caffeine consumption*

In this section, we conduct a single linear regression between sleep efficiency and caffeine consumption to determine whether caffeine consumption will directly lead to reduced sleep efficiency.



```
Call:
lm(formula = sleep.efficiency ~ Caffeine.consumption, data = clean_sleep)

Residuals:
    Min       1Q   Median       3Q      Max
-0.30562 -0.08664  0.03438  0.11115  0.20115

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7888485   0.0087555  90.097  <2e-16 ***
Caffeine.consumption 0.0002236   0.0002358   0.948   0.344
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1335 on 377 degrees of freedom
Multiple R-squared:  0.002379, Adjusted R-squared:  -0.0002671
F-statistic: 0.899 on 1 and 377 DF, p-value: 0.3436
```

```
> cor(clean_sleep$sleep.efficiency, clean_sleep$Caffeine.consumption)
[1] 0.04877573
```

Based on the p-value 0.3436 (> 0.05), there is lack of strong evidence to reject the null hypothesis. Furthermore, the R-squared value 0.002379 ($< 1\%$) shows an extremely weak linear correlation between caffeine consumption and sleep efficiency. Therefore, we can conclude that caffeine consumption does not affect sleep efficiency.

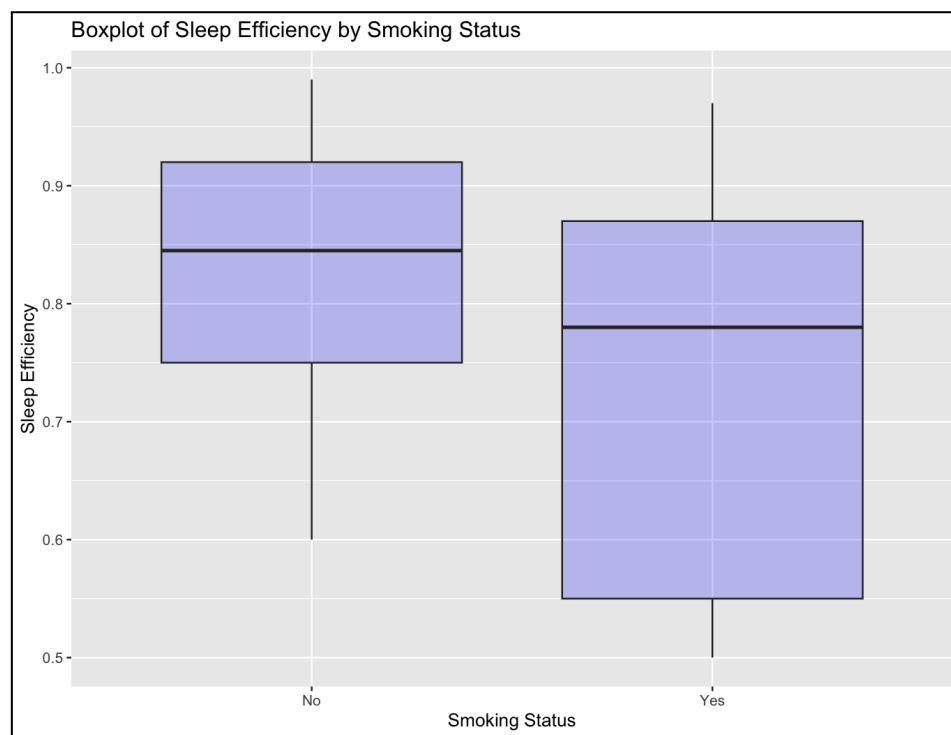
Based on the correlation coefficient of 0.04877 and the scatterplot with a best-fit line with a positive gradient, there appears to be a weak positive correlation between caffeine consumption and sleep efficiency. This finding contradicts the widely accepted understanding that caffeine intake impairs sleep efficiency. Upon closer examination, a

critical factor to consider is the timing of caffeine consumption prior to bedtime other than just looking at the amount of consumption (Centers for Disease Control and Prevention, 2022; University of Michigan, 2020).

4.2.7 Relationship between *sleep efficiency* and *smoking status*

Smoking is known to impose negative consequences to our health, hence we are interested in investigating its effect on sleep efficiency as well.

The box plot below shows the relationship between smoking status and sleep efficiency.



At first glance, it seems that subjects who smoked experienced relatively lower sleep efficiency, and this is verified once we perform the Wilcoxon rank sum test.

H_0 : There is no significant difference in sleep efficiency comparing those who smoke and those who do.

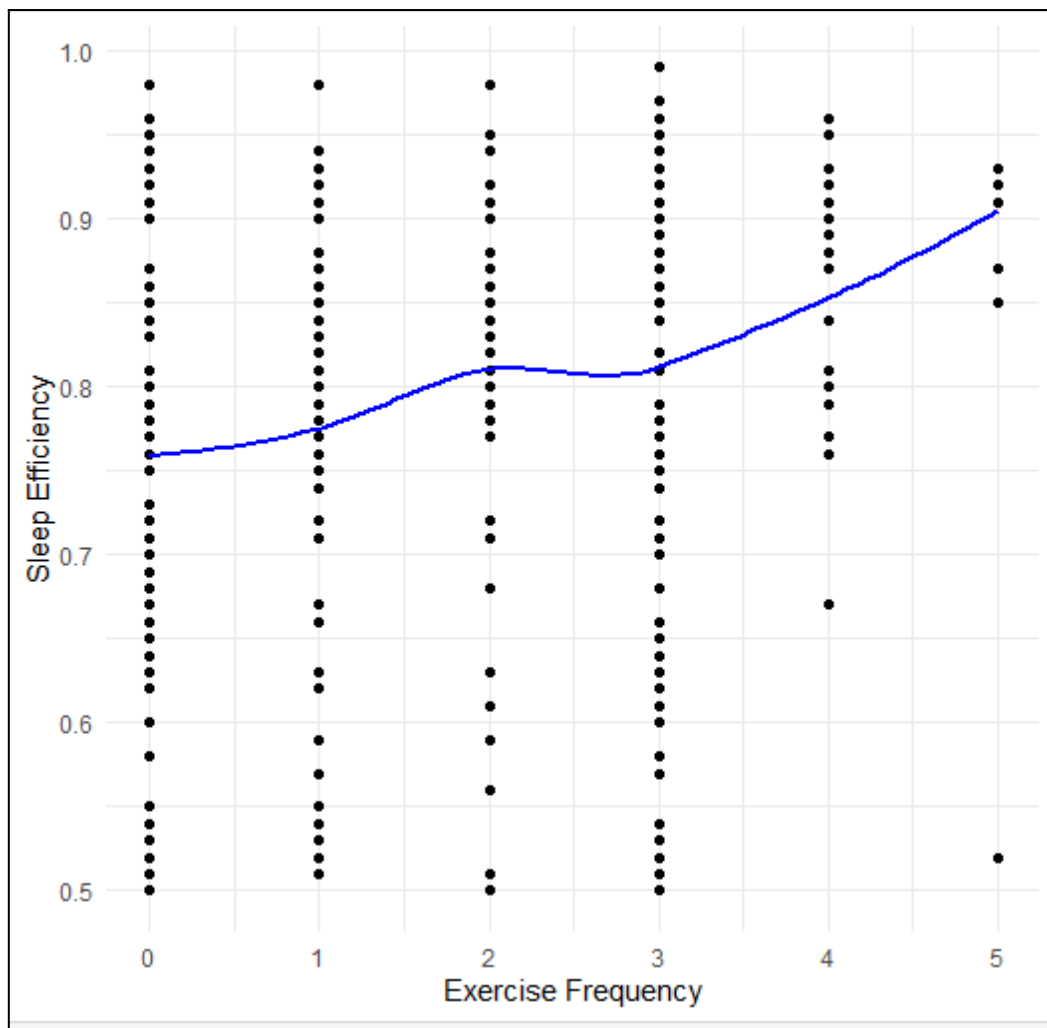
H_1 : There is a significant difference in sleep efficiency comparing those who smoke and those who do.

```
wilcoxon rank sum test with continuity correction
data:  sleepesmoke and sleepenosmoke
w = 11040, p-value = 1.717e-07
alternative hypothesis: true location shift is not equal to 0
```

As seen above, performing the Wilcoxon rank sum test gives us a p-value of 1.717×10^{-7} , letting us reject the null hypothesis at a 5% significance level.

4.2.8 Relationship between *sleep efficiency* and *exercise frequency*

It is a known fact that exercise will improve our overall health. Hence we are interested to find out will exercise also improve sleep efficiency.



From the scatterplot, we observed an increasing trend of sleep efficiency as exercise frequency increases.

```
> kruskal.test(sleep.efficiency ~ Exercise.frequency, data = clean_sleep)

    kruskal-wallis rank sum test

data:  sleep.efficiency by Exercise.frequency
kruskal-wallis chi-squared = 34.102, df = 5, p-value = 2.272e-06
```

With a p-value of 2.272×10^{-6} , it is much smaller than the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant difference in Sleep Efficiency between at least 2 of the Exercise Frequency levels.

The Kruskal-Wallis test only indicates that there is a difference between at least two groups but does not specify which groups are different. To determine which specific groups have a significant difference, we shall perform the Dunn test using the `dunn.test()` function.

		Comparison of x by group (Bonferroni)				
Col	Mean-	0	1	2	3	4
Row	Mean					
1		-0.729770 1.0000				
2		-2.035315 0.3136	-1.361054 1.0000			
3		-2.653336 0.0598	-1.720870 0.6396	0.008284 1.0000		
4		-5.409795 0.0000*	-4.674832 0.0000*	-3.089567 0.0150*	-3.607069 0.0023*	
5		-1.893332 0.4374	-1.596695 0.8275	-0.923473 1.0000	-0.967064 1.0000	0.775536 1.0000

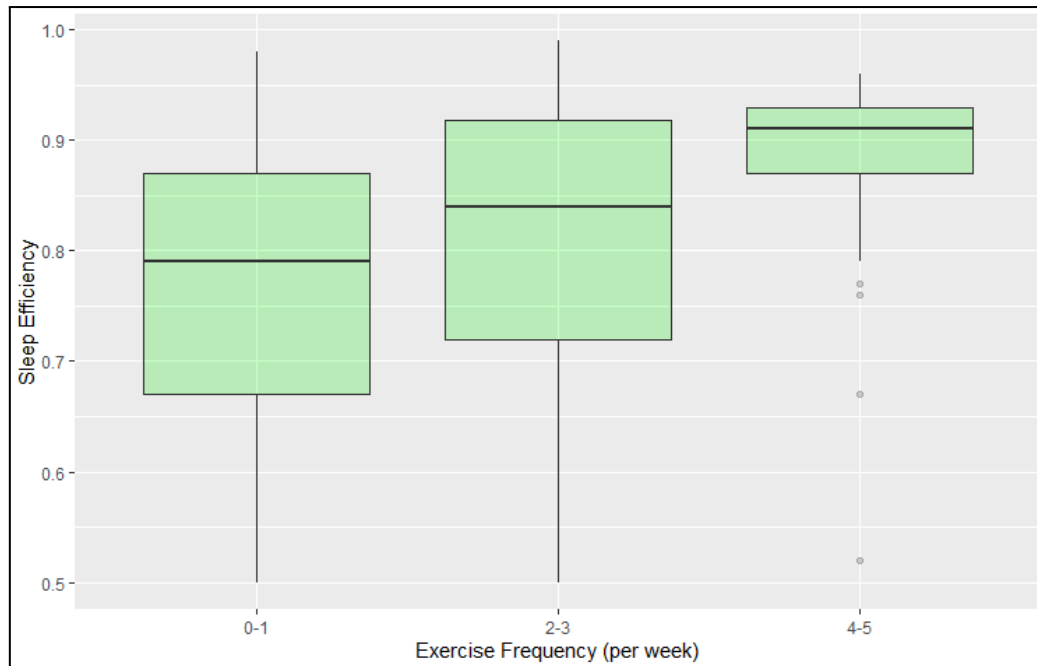
alpha = 0.05
Reject Ho if p <= alpha/2

Pairwise hypothesis testing

H_0 : There is no difference in the distribution of sleep efficiency among the groups being compared (i.e. exercise frequency groups).

H_1 : There is at least one pair of exercise frequency groups with a significant difference in the distribution of the dependent variable (exercise frequency).

We first decided to categorise the exercise frequencies further, as some categories had too few entries to be statistically significant. The entries were then grouped into three factors based on the exercise frequency of "0-1", "2-3" and "4-5", indicating low, moderate and high frequency of exercise respectively. The number of subjects in each group are 179, 158 and 42 for "0-1", "2-3" and "4-5" respectively. We then generated a boxplot of these three categories and performed further statistical analysis as seen below.



We then formulated the following null and alternative hypotheses before performing a Kruskal- Wallis rank sum test:

H_0 : Different frequencies of exercise will not cause a significant difference in a subject's sleep efficiency.

H_1 : At least one of the frequencies of exercise above will cause a significant difference in a subject's sleep efficiency.

```
kruskal-wallis rank sum test
data: clean_sleep$sleep.efficiency and clean_sleep$ExFreqGroup
kruskal-wallis chi-squared = 32.968, df = 2, p-value = 6.934e-08
```

Performing the test gives a p-value of $6.934 \times 10^{-8} < 0.05$, which means that we can reject the null hypothesis at the 5% significance level. Therefore, we can say that exercise frequency affects a person's sleep efficiency.

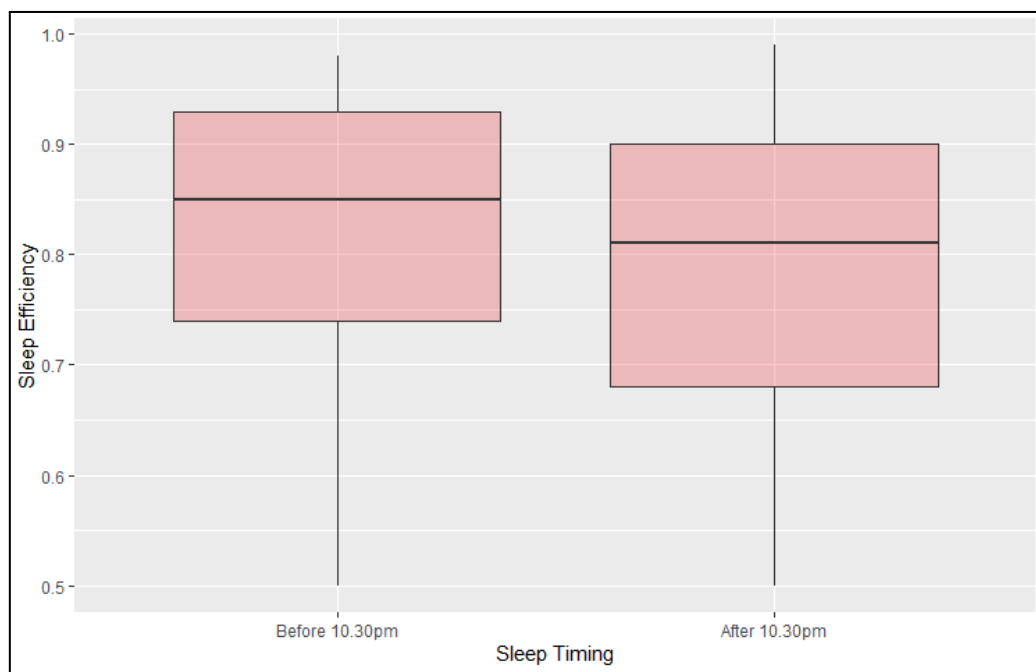
```
Pairwise comparisons using wilcoxon rank sum test with continuity correction
data: clean_sleep$sleep.efficiency and clean_sleep$ExFreqGroup
      0-1      2-3
2-3 0.00453 -
4-5 1.1e-08 0.00037
```

Performing a pairwise comparisons test shows that there are differences in sleep efficiency at all levels of exercise frequency. Judging from the boxplot above,, the higher your exercise frequency, the better your sleep efficiency.

4.2.9 Relationship between *sleep efficiency* and *bedtime*

In this section, we try to answer the question of “*Does early bed time translate into better sleep efficiency?*”.

We first decided to categorise the sleep timings into two groups. According to Peters (2022), a board-certified neurologist and sleep medicine specialist, the recommended bedtime for adults is about 10.30pm. Hence we decided to split the timings into 2 different factors - “Before 10.30pm” for early bedtime and “After 10.30pm” for late bedtime. The category of “Before 10.30pm” consists of all timings before and inclusive of 10.30pm (137 entries) while the other category, “After 10.30pm”, consists of timings after 10.30pm (242 entries). We then generated a boxplot of these two categories and performed further statistical analysis as seen below.



We then formulated the following null and alternative hypotheses before performing a Wilcoxon rank sum test:

H_0 : Sleeping before or after 10.30pm will not affect a subject's sleep efficiency.

H_1 : Sleeping before or after 10.30pm will affect a subject's sleep efficiency.

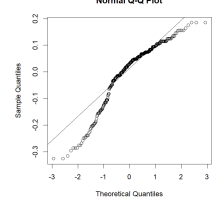
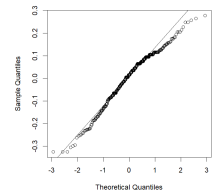
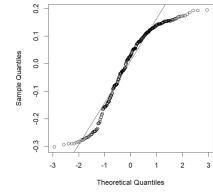
```
wilcoxon rank sum test with continuity correction
data:  sleepbefore2230 and sleepafter2230
w = 19745, p-value = 0.001981
alternative hypothesis: true location shift is not equal to 0
```

Performing the test gives a p-value of $0.001981 < 0.05$, which means that we can reject the null hypothesis at the 5% significance level. Therefore, we can say that the time a subject goes to sleep does affect their sleep efficiency.

4.2.10 Most important factor affecting *sleep efficiency*

In section 4.1, we observe that awakenings, bedtime and sleep duration are relatively strongly correlated to sleep efficiency. We perform a single linear regression to determine the strongest factor that affects sleep efficiency that could be used to model sleep efficiency linearly.

Sleep efficiency = $\beta_0 + \beta_1 * X + \epsilon$, where X is the most important factor. The analysis summary table is illustrated below. By comparing the R-squared and residual plot, we conclude that awakenings has the most significant impact on sleep efficiency.

Variable (X)	Fitted mode where Y is Sleep efficiency	P-value	R-squared	qq-plot of residuals
Awakenings	$\hat{Y} = 0.8825 + (-0.05690)X$	$< 2.2e-16$	0.3385	
Bedtime	$\hat{Y} = 0.7732 + 0.001523X$	0.04116	0.01378	
Sleep Duration	$\hat{Y} = 0.7290 + 0.008144X$	0.3625	0.002756	

4.3 Multiple Linear Regression

We tried to build a multiple linear model for Sleep Efficiency based on the variables excluding smoking status and exercise frequency which are categorical by nature. Backward stepwise regression is used to select the most appropriate model. The result is shown in the R output below.

Based on the findings, we conclude that Age, Alcohol consumption and Awakenings are the significant measures that can be used to model Sleep Efficiency while Gender, Caffeine consumption, Sleep duration, Wake up time and Bedtime are not significant.

The fitted model is:

Sleep.efficiency = $0.8626758 + 0.0009726 * \text{Age} + (-0.0265067) * \text{Alcohol.consumption} + (-0.0492081) * \text{Awakenings}$

```
> model2 = lm(Sleep.efficiency ~ Age + Wakeup.time + Caffeine.consumption
+ Alcohol.consumption + Awakenings + Sleep.duration + Bedtime + Gender,
data = train.data)
> step(model2, direction="backward")
Start: AIC=-1371.76
Sleep.efficiency ~ Age + Wakeup.time + Caffeine.consumption +
  Alcohol.consumption + Awakenings + Sleep.duration + Bedtime +
  Gender
```

	Df	Sum of Sq	RSS	AIC
- Gender	1	0.00417	3.0908	-1373.3
- Caffeine.consumption	1	0.00562	3.0922	-1373.2
- Bedtime	1	0.01347	3.1001	-1372.4
<none>			3.0866	-1371.8
- Sleep.duration	1	0.02465	3.1112	-1371.3
- Wakeup.time	1	0.03491	3.1215	-1370.3
- Age	1	0.06016	3.1467	-1367.9
- Alcohol.consumption	1	0.55464	3.6412	-1323.7
- Awakenings	1	1.24265	4.3292	-1271.2

```
Step: AIC=-1373.35
Sleep.efficiency ~ Age + Wakeup.time + Caffeine.consumption +
  Alcohol.consumption + Awakenings + Sleep.duration + Bedtime
```

	Df	Sum of Sq	RSS	AIC
- Caffeine.consumption	1	0.00373	3.0945	-1375.0
- Bedtime	1	0.01317	3.1039	-1374.1
<none>			3.0908	-1373.3
- Sleep.duration	1	0.02496	3.1157	-1372.9
- Wakeup.time	1	0.03445	3.1252	-1372.0
- Age	1	0.05627	3.1470	-1369.9
- Alcohol.consumption	1	0.55428	3.6450	-1325.4
- Awakenings	1	1.25533	4.3461	-1272.1

```
Step: AIC=-1374.99
Sleep.efficiency ~ Age + Wakeup.time + Alcohol.consumption +
  Awakenings + Sleep.duration + Bedtime
```

	Df	Sum of Sq	RSS	AIC
- Bedtime	1	0.01340	3.1079	-1375.7
<none>			3.0945	-1375.0
- Sleep.duration	1	0.02654	3.1210	-1374.4
- Wakeup.time	1	0.03615	3.1306	-1373.5
- Age	1	0.06642	3.1609	-1370.5
- Alcohol.consumption	1	0.55055	3.6450	-1327.4
- Awakenings	1	1.25794	4.3524	-1273.6

```
Step: AIC=-1375.68
Sleep.efficiency ~ Age + Wakeup.time + Alcohol.consumption +
  Awakenings + Sleep.duration
```

	Df	Sum of Sq	RSS	AIC
- Sleep.duration	1	0.01476	3.1226	-1376.2
<none>			3.1079	-1375.7
- Wakeup.time	1	0.02890	3.1368	-1374.9
- Age	1	0.05914	3.1670	-1372.0
- Alcohol.consumption	1	0.53939	3.6473	-1329.2
- Awakenings	1	1.25731	4.3652	-1274.7


```

Step:  AIC=-1376.24
Sleep.efficiency ~ Age + Wakeup.time + Alcohol.consumption +
Awakenings

              Df Sum of Sq    RSS    AIC
- Wakeup.time    1   0.01532  3.1380 -1376.8
<none>                        3.1226 -1376.2
- Age             1   0.05361  3.1763 -1373.1
- Alcohol.consumption 1   0.55474  3.6774 -1328.7
- Awakenings      1   1.28917  4.4118 -1273.5

Step:  AIC=-1376.76
Sleep.efficiency ~ Age + Alcohol.consumption + Awakenings

              Df Sum of Sq    RSS    AIC
<none>                        3.1380 -1376.8
- Age             1   0.04585  3.1838 -1374.4
- Alcohol.consumption 1   0.55084  3.6888 -1329.8
- Awakenings      1   1.33778  4.4757 -1271.2

Call:
lm(formula = Sleep.efficiency ~ Age + Alcohol.consumption + Awakenings,
    data = train.data)

Coefficients:
            (Intercept)              Age  Alcohol.consumption
Awakenings
            0.8626758              0.0009726              -0.0265067
            -0.0492081

```

5. Conclusion and Discussion

Sleep efficiency has an impact on physical and psychological health. Low sleep efficiency , also known as poor sleep quality, is associated with diseases, major depressive disorder, bipolar disorder, generalised anxiety disorder, and physical disorder (Kohyama, 2021). Therefore, this report intended to identify factors associated with sleep efficiency and attempt to answer some of the questions related to sleep efficiency based on the dataset from a research study.

We conclude that:

- Factors without correlation or association with sleep efficiency are gender, age, sleep duration, and caffeine consumption.
- Awakening is highly correlated to sleep efficiency and does have a negative impact on sleep efficiency.
- Alcohol consumption does affect sleep efficiency.
- Smoking does affect sleep efficiency as compared to those who do not.
- Exercise frequency is associated with sleep efficiency as statistics showed that frequent exercise improves sleep efficiency.
- Bedtime does affect sleep efficiency. Late bedtime, in particular, sleeping after 10.30pm will decrease sleep efficiency.

Additionally, the variables age, alcohol consumption and awakenings can be used to model the sleep efficiency using the multiple linear regression model. The remaining variables do not seem to have a strong correlation with sleep efficiency.

6. References

Boston College. (n.d.). *Transformations: an introduction*. Retrieved April 5, 2023, from

<http://fmwww.bc.edu/repec/bocode/t/transint.html>

Centers for Disease Control and Prevention. (2022, April 26). *Caffeine & long work hours*.

Retrieved April 10, 2023, from

<https://www.cdc.gov/niosh/emres/longhourstraining/caffeine.html>

Kohyama, J. (2021, June 24). *Which is more important for Health: Sleep Quantity or sleep quality?* MDPI. Retrieved April 14, 2023, from

<https://www.mdpi.com/2227-9067/8/7/542>

Peters, B. (2022, September 15). *How to choose the perfect time to go to bed*. Verywell

Health. Retrieved April 10, 2023, from

<https://www.verywellhealth.com/what-time-should-you-go-to-sleep-4588298#:~:text=School%2Dage%20children%20should%20go,00%20and%2011%3A00%20p.m>

University of Michigan. (2020, December 16). *When to stop drinking alcohol, water or*

caffeine before bed for better sleep. Retrieved April 12, 2023, from

<https://medicine.umich.edu/dept/psychiatry/news/archive/202012/when-stop-drinking-alcohol-water-or-caffeine-bed-better-sleep>

7. Appendix

```
## R Codes
```

```
library(dplyr)
```

```
library(ggplot2)
```

```
library(repr)
```

```
library(lubridate)
```

```
library(psych)
```

```
library(dunn.test)
```

```
## Read data from file
```

```
sleep <- read.csv("Sleep_Efficiency.csv")
```

```
str(sleep)
```

```
## Data Cleaning
```

```
sleep[4] = substring(sleep$Bedtime, 11) # remove date from Bedtime
```

```
sleep[5] = substring(sleep$Wakeup.time, 11) # remove date from Wakeup.time
```

```
clean_sleep <- sleep %>% filter(Age >= 18) %>% select(-c(ID, REM.sleep.percentage,  
Deep.sleep.percentage, Light.sleep.percentage)) %>% na.omit() # select adults' data; omit  
'ID' column & NAs
```

```
head(clean_sleep)
```

```
str(clean_sleep)
```

```
# convert Bedtime to numeric (in hours)
```

```
format(clean_sleep[,3], format = "%H:%M:%S", tz = "", usetz = FALSE, digits =  
getOption("digits.secs"))
```

```
clean_sleep[,3] <- sapply(strsplit(clean_sleep[,3], ":"), function(x) {  
  x <- as.numeric(x)  
  x[1] + x[2]/60 + x[3]/3600 })
```

```
# convert Wakeup.time to numeric (in hours)
```

```
format(clean_sleep[, 4], format = "%H:%M:%S", tz = "", usetz = FALSE, digits =  
getOption("digits.secs"))
```

```
clean_sleep[, 4] <- sapply(strsplit(clean_sleep[, 4], ":"), function(x) {  
  x <- as.numeric(x)  
  x[1] + x[2]/60 + x[3]/3600 })
```

```
str(clean_sleep)
```

```
summary(clean_sleep)
```

```
# Set plotting size
```

```
options(repr.plot.height = 4, repr.plot.width = 6)
```

```
## 3.1 MAIN VARIABLE: SLEEP EFFICIENCY
```

```
par(mfrow=c(1,3))
```

```
efficiency <- clean_sleep[, 'Sleep.efficiency'];
```

```
# x1 <- seq(-10, 100, by=0.1); n1 <- dnorm(x1, mean(efficiency), sd(efficiency))
```

```
hist(efficiency, main='Histogram of Sleep Efficiency', xlab='Sleep Efficiency', ylab='Proportion of Time', col='white')
box("figure", col="black")
```

```
# x1 <- seq(-10, 100, by=0.1); n1 <- dnorm(x1, mean(log_efficiency), sd(log_efficiency))
hist((efficiency)^2, main='Histogram of (Sleep Efficiency)^2', xlab='Sleep Efficiency',
ylab='Proportion of Time', col='white')
box("figure", col="black")
```

```
boxplot(efficiency, main='Boxplot of Sleep Efficiency', lwd=2, col='white')
box("figure", col="black")
```

```
summary(efficiency)
```

3.2.1 AGE

```
par(mfrow=c(1,1))
hist(clean_sleep$Age, col='white', main='Histogram of Age', ylim=c(10,70), xlab='Age',
ylab='Number of subjects', cex=4)
boxplot(clean_sleep$Age, col='white', main='Boxplot of Age', ylim=c(10,70))
summary(clean_sleep$Age)
# box("figure", col="black", lwd = 1)
```

3.2.2 GENDER

```
barplot(table(clean_sleep$Gender), ylab = 'Number of subjects', xlab="Gender",
ylim=c(0,200), col='white', main='Barplot of Gender')
```

3.2.3 BEDTIME

```
hist(clean_sleep$Bedtime, col='white', main='Histogram of Bedtime', xlab='Bedtime',
ylab='Number of subjects', ylim=c(0,200))
boxplot(clean_sleep$Bedtime, col='white', main='Boxplot of Bedtime', ylim=c(0,25))
summary(clean_sleep$Bedtime)
```

3.2.4 WAKEUP TIME

```
hist1 = hist(clean_sleep$Wakeup.time, xlim=c(2,12), ylim=c(0,100), col="white",
xlab="Wakeup Time", main="Histogram of Wakeup Time")
boxplot(clean_sleep$Wakeup.time, main="Boxplot of Wakeup Time")
```

3.2.5 SLEEP DURATION

```
hist2 = hist(clean_sleep$Sleep.duration, ylim=c(0,200), col= "white", xlab="Sleep Duration",
main="Histogram of Sleep Duration")
boxplot(clean_sleep$Sleep.duration, main="Boxplot of Sleep Duration")
```

3.2.6 AWAKENINGS

```
hist3 = hist(clean_sleep$Awakenings, ylim=c(0,200), col="white", xlab="Awakening Count",
main="Histogram of Awakenings")
boxplot(clean_sleep$Awakenings, main="Boxplot of Awakenings")
```

3.2.7 CAFFEINE CONSUMPTION

```
hist(clean_sleep$Caffeine.consumption, main="Histogram of Caffeine Consumption",
col='white', xlab="Caffeine consumed in the past 24hrs (in mg)", ylab="Number of subjects")
boxplot(clean_sleep$Caffeine.consumption, main="Boxplot of Caffeine Consumption",
col='white')
```

3.2.8 ALCOHOL CONSUMPTION

```
hist(clean_sleep$Alcohol.consumption, main="Histogram of Alcohol Consumption",
col="white", xlab="Alcohol consumed in the past 24hrs (in oz)", ylab="Number of subjects")
boxplot(clean_sleep$Alcohol.consumption, main="Boxplot of Alcohol Consumption",
col='white')
```

3.2.9 SMOKING STATUS

```
barplot(table(clean_sleep$Smoking.status), col="white", main="Barplot of Smoking Status",
ylab="Number of subjects", xlab="Smoking Status")
nrow(clean_sleep %>% select(Smoking.status) %>% filter(Smoking.status=="No"))
nrow(clean_sleep %>% select(Smoking.status) %>% filter(Smoking.status=="Yes"))
```

3.2.10 EXERCISE FREQUENCY

```
hist(clean_sleep$Exercise.frequency, col="white", main="Histogram of Exercise
Frequency", xlab="Exercise Frequency", ylab="Number of subjects")
boxplot(clean_sleep$Exercise.frequency, col="white", main="Boxplot of Exercise
Frequency")
```

4.1 Correlation Plot

```
# convert Smoking into numeric data
```

```
clean_sleep$Smoking.status <- as.factor(clean_sleep$Smoking.status)
clean_sleep$Smoking.status <- ifelse(clean_sleep$Smoking.status=="Yes",1,0)
```

```
num_var <- clean_sleep %>% select(-Gender)
pairs(num_var, main = "Sleep Efficiency and Other Variables")
pairs.panels(num_var, method = "pearson", hist.col = "steelblue", pch = 21, density = TRUE,
ellipses = FALSE)
```

4.2

4.2.1 Relationship between Sleep Efficiency and Gender

```
# boxplot of Sleep Efficiency by Gender
```

```
ggplot(clean_sleep, aes(x=Gender, y=Sleep.efficiency)) +geom_boxplot(fill="white",
alpha=0.2) + xlab("Gender") + ylab("Sleep Efficiency") + ggtitle("Box Plot: Sleep Efficiency
by Gender") + theme_bw()
```

```
# compare variance of males' and females' sleep efficiency: H0: var(M)=var(F); H1:
var(M)≠var(F)
var.test(Sleep.efficiency ~ Gender, data = clean_sleep)
```

```
# test for normality in the sleep efficiency data for males and females separately
```

```
# For males
```

```
male_data <- clean_sleep[clean_sleep$Gender == "Male",]
shapiro.test(male_data$Sleep.efficiency)
```

```

# For females
female_data <- clean_sleep[clean_sleep$Gender == "Female",]
shapiro.test(female_data$Sleep.efficiency)

# t.test for differences in mean of sleep efficiency of Males & Females at 95% of confidence interval
t.test(Sleep.efficiency ~ Gender, data = clean_sleep)

## 4.2.2 Relationship between Sleep Efficiency and Age
# linear regression between Sleep Efficiency and Age
plot(clean_sleep$Age, clean_sleep$Sleep.efficiency, xlab = "Age", ylab = "Sleep Efficiency")
age_efficiency_model <- lm(Sleep.efficiency ~ Age, data = clean_sleep)
abline(age_efficiency_model, col = "red")
summary(age_efficiency_model)

# correlation between Sleep Efficiency and Age
cor(clean_sleep$Sleep.efficiency, clean_sleep$Age)

## 4.2.3 Relationship between Sleep Efficiency and Sleep Duration
# linear regression between Sleep Efficiency and Sleep Duration
plot(clean_sleep$Sleep.duration, clean_sleep$Sleep.efficiency, xlab = "Sleep Duration", ylab = "Sleep Efficiency")
sd_efficiency_model <- lm(Sleep.efficiency ~ Sleep.duration, data = clean_sleep)
abline(sd_efficiency_model, col = "red")
summary(sd_efficiency_model)

# correlation between Sleep Efficiency and Sleep Duration
cor(clean_sleep$Sleep.efficiency, clean_sleep$Sleep.duration)

## 4.2.4 Relationship between Sleep Efficiency and Awakenings
# linear regression between Sleep Efficiency and Awakenings
plot(clean_sleep$Awakenings, clean_sleep$Sleep.efficiency, xlab = "Awakenings", ylab = "Sleep Efficiency", main = "Awakenings vs Sleep Efficiency")
awakenings_model <- lm(Sleep.efficiency ~ Awakenings, data = clean_sleep)
abline(awakenings_model, col = "red")
summary(awakenings_model)

## 4.2.5 Relationship between Sleep Efficiency and Alcohol Consumption
# linear regression between Sleep Efficiency and Alcohol Consumption
plot(clean_sleep$Alcohol.consumption, clean_sleep$Sleep.efficiency, xlab = "Alcohol Consumption", ylab = "Sleep Efficiency", main = "Alcohol Consumption vs Sleep Efficiency")
sleep_efficiency_model <- lm(clean_sleep$Sleep.efficiency ~ clean_sleep$Alcohol.consumption)
abline(sleep_efficiency_model, col = "blue")
summary(sleep_efficiency_model)

## 4.2.6 Relationship between Sleep Efficiency and caffeine consumption
# scatterplot of Sleep Efficiency and Caffeine Consumption

```

```
ggplot(clean_sleep, aes(x=Caffeine.consumption, y=Sleep.efficiency)) + geom_point() +
geom_smooth(method=lm) +
  labs(title="Scatter Plot: Sleep Efficiency vs Caffeine Consumption",
x="Caffeine.consumption", y="Sleep Efficiency") + theme_bw()
caff_model <- lm(Sleep.efficiency ~ Caffeine.consumption, data = clean_sleep)
summary(caff_model)
```

```
## 4.2.7 Relationship between Sleep Efficiency and Smoking Status
clean_sleep$Smoking.status = as.factor(clean_sleep$Smoking.status)
```

```
# boxplot of Sleep Efficiency by Smoking Status
ggplot(clean_sleep, aes(x=Smoking.status, y=Sleep.efficiency)) +geom_boxplot(fill="blue",
alpha=0.2) + xlab("Smoking Status") + ylab("Sleep Efficiency") + ggtitle("Boxplot of Sleep
Efficiency by Smoking Status")
```

```
# wilcoxon rank sum test
sleepesmoke = clean_sleep$Sleep.efficiency[clean_sleep$Smoking.status == "1"]
sleepenosmoke = clean_sleep$Sleep.efficiency[clean_sleep$Smoking.status == "0"]
wilcox.test(sleepesmoke, sleepenosmoke)
```

```
## 4.2.8 Relationship between Sleep Efficiency and Exercise Frequency
# scatterplot of Sleep Efficiency vs Exercise Frequency
ggplot(clean_sleep, aes(x = Exercise.frequency, y = Sleep.efficiency)) +
  geom_point() +
  geom_smooth(method = "loess", span = 0.75, se = FALSE, color = "blue") +
  xlab("Exercise Frequency") +
  ylab("Sleep Efficiency") +
  theme_minimal()
```

```
# Kruskal-Wallis test
kruskal.test(Sleep.efficiency ~ Exercise.frequency, data = clean_sleep)
```

```
# Dunn test
dunn.test(clean_sleep$Sleep.efficiency, clean_sleep$Exercise.frequency, method =
"bonferroni")
```

```
# Categorise exercise frequency into 3 groups
clean_sleep$ExFreqGroup[clean_sleep$Exercise.frequency==0|clean_sleep$Exercise.frequ
ency==1]= "0-1"
clean_sleep$ExFreqGroup[clean_sleep$Exercise.frequency==2|clean_sleep$Exercise.frequ
ency==3]= "2-3"
clean_sleep$ExFreqGroup[clean_sleep$Exercise.frequency==4|clean_sleep$Exercise.frequ
ency==5]= "4-5"
clean_sleep$ExFreqGroup = ordered(clean_sleep$ExFreqGroup, c("0-1","2-3","4-5"))
#179 0-1 entries
length(clean_sleep$ExFreqGroup[clean_sleep$ExFreqGroup == "0-1"])
#158 2-3 entries
length(clean_sleep$ExFreqGroup[clean_sleep$ExFreqGroup == "2-3"])
```

```

#42 4-5 entries
length(clean_sleep$ExFreqGroup[clean_sleep$ExFreqGroup == "4-5"])
# boxplot of Sleep Efficiency by Exercise Frequency
ggplot(clean_sleep, aes(x=ExFreqGroup, y=Sleep.efficiency)) +geom_boxplot(fill="green",
alpha=0.2) + xlab("Exercise Frequency") + ylab("Sleep Efficiency")

# Kruskal- Wallis rank sum test
kruskal.test(clean_sleep$Sleep.efficiency, clean_sleep$ExFreqGroup)

# pairwise comparison between the sleep efficiency of the 3 groups of exercise frequency
pairwise.wilcox.test(clean_sleep$Sleep.efficiency, clean_sleep$ExFreqGroup,
p.adjust.method = "none")

## 4.2.9 Relationship between Sleep Efficiency and Bedtime
# categorise sleep timings into 2 groups
clean_sleep$BedtimeGroup[clean_sleep$Bedtime==21|clean_sleep$Bedtime==21.5|clean_
sleep$Bedtime==22|clean_sleep$Bedtime==22.5] = "Before 10.30pm"
clean_sleep$BedtimeGroup[clean_sleep$Bedtime==23|clean_sleep$Bedtime==23.5|clean_
sleep$Bedtime==0|clean_sleep$Bedtime==0.5|clean_sleep$Bedtime==1|clean_sleep$Bedti
me==1.5|clean_sleep$Bedtime==2|clean_sleep$Bedtime==2.5] = "After 10.30pm"
clean_sleep$BedtimeGroup = as.factor(clean_sleep$BedtimeGroup)
clean_sleep$BedtimeGroup = ordered(clean_sleep$BedtimeGroup, c("Before
10.30pm", "After 10.30pm"))
# boxplot of Sleep Efficiency and Bedtime
ggplot(clean_sleep, aes(x=as.factor(BedtimeGroup), y=Sleep.efficiency))
+geom_boxplot(fill="red", alpha=0.2) + xlab("Sleep Timing") + ylab("Sleep Efficiency")
# Wilcoxon rank sum test of Sleep Efficiency and Bedtime
sleepbefore2230 = clean_sleep$Sleep.efficiency[clean_sleep$BedtimeGroup == "Before
10.30pm"]
sleepafter2230 = clean_sleep$Sleep.efficiency[clean_sleep$BedtimeGroup == "After
10.30pm"]
wilcox.test(sleepbefore2230, sleepafter2230)

# 4.2.10 Most important factor affecting Sleep Efficiency
set.seed(100)
training.idx=sample(1:nrow(clean_sleep),nrow(clean_sleep)*0.8)
train.data=clean_sleep[training.idx, ]
test.data=clean_sleep[-training.idx, ]
lmodel=lm(Sleep.efficiency~.,data=train.data)
summary(lmodel)

modela = lm(Sleep.efficiency~ Awakenings, data = train.data)
summary(modela)
res = resid(modela)
plot(fitted(modela), res)
abline(0,0)
qqnorm(res)
qqline(res)

```



```

modelb = lm(Sleep.efficiency~ Bedtime, data = train.data)
summary(modelb)
res = resid(modelb)
plot(fitted(modelb), res)
abline(0,0)
qqnorm(res)
qqline(res)
modelc = lm(Sleep.efficiency~ Sleep.duration, data = train.data)
summary(modelc)
res = resid(modelc)
plot(fitted(modelc), res)
abline(0,0)
qqnorm(res)
qqline(res)

```

4.3 Multiple Linear Regression

```

model2 =lm(Sleep.efficiency~ Age + Wakeup.time + Caffeine.consumption +
Alcohol.consumption + Awakenings + Sleep.duration + Bedtime + Gender, data = train.data)
step(model2, direction="backward")

```