Nicholas Auyang, Philip Wu, and Daniel Sun
Professor J. Speagle, TA Lily Thao Nguyen
STA130H S LEC 0101 TUT 0107
11 April 2023

Impact of Lifestyle and Non-Lifestyle Factors on Sleep Facets

Abstract

It is general knowledge that sleep can be impacted by a variety of factors and personal choices, however, the holistic extent of these variables is not extremely established beyond basic platitudes. A poor sleep session can be reflected in the sleep efficiency which can lead to changes in mood, ability to focus, fatigue, and lower productivity. These consequences affect the quality of our lives and it is in people's best interest and desire to get a good night's sleep.

Our findings suggest that lifestyle factors are more significant factors to sleep efficiency than non-lifestyle factors; for example, caffeine consumption is more associated with sleep efficiency than age and smoking plays a major role in determining both deep sleep and sleep efficiency. A variety of statistical methods were applied to come to these conclusions, such as linear regression, bootstrapping and multivariable regression. These findings are obtained from a public source of data that was gathered by self reported surveys, actigraphy, and polysomnography. These results have important implications for future research on sleep, sleep epidemiology, as well as public health interventions aimed at improving sleep habits and by extension, our general wellbeing.

Introduction

The primary motivation for this project is to better understand the factors that contribute to healthy sleep patterns and to identify potential areas for intervention from the factors in the dataset. Research on sleep is limited in comparison to other topics and we are taking the preliminary step in the right direction, providing a basis of knowledge (on this area of sleep) for future research. Our analysis focuses on finding the correlation between lifestyle factors (caffeine intake, exercise frequency, alcohol consumption and smoking) and different aspects of sleep such as deep sleep, REM sleep and the implications of these lifestyle choices on these facets. Each of these analyses set the pace and provide meaningful insights and dispel common myths towards sleep.

Some key terminology are:

Sleep Efficiency: Generally a fraction or percentage comparing time in bed to time asleep.

Light Sleep: Short sleep stage where dreaming and drifting off begins

REM Sleep: Sleep stage with Rapid Eye Movement under eyelids and dreaming

Deep Sleep: Sleep stage with learning and emotional processing, physical recovery and detoxification

Z-score: the number of standard deviations away from the mean a data point is

Variance inflation factor (VIF): Is the measure of the amount of multicollinearity in regression analysis

Data

The dataset used in our procedures is from a public data site, named Kaggle, which is a subsidiary of Google. This website is full of a community of data scientists and machine learning practitioners. Ours is from a private author who uploaded and gathered the data from self reported surveys, actigraphy, and polysomnography. Inside the file contains 15 variables which include, ID, age, gender, bedtime, wakeup time, sleep duration, sleep efficiency, REM sleep percentage, deep sleep percentage, light sleep percentage, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency. To make this dataset reliable for our usage, we removed na values for each variable to make sure each observation is full. We also made sure that each of the 3 percentage point variables added up to 100% because they represent a full night of sleep. Some group members had extra steps, specified under their analysis.

Acknowledgements

Question 1 - Nicholas: Looking at the first question answered (i.e., Non-Lifestyle Factors), it may have been beneficial to include a multivariable regression analysis on top of the linear regression, which would have allowed us to uncover more information regarding the difference in impact between Age and Gender.

Question 1 - Nicholas

The question I seeked to answer was:

> Can a regression model predict whether the sociodemographic categories of Age and Gender of an individual affect our sleep efficiency?

This question was designed to find how Non-Lifestyle factors affect sleep efficiency. This will better allow us to compare the differences between Lifestyle and Non-Lifestyle factors.

Methods - Question 1 - Nicholas

The first step to this process was taking the data and processing much of it through selection and arrangement. Note that none of these categories had NA values, and there was no need to filter.

```
library(tidyverse)
```

```
# Read in data
data <- read_csv("NEW_Sleep_Efficiency.csv")
glimpse(data)
```

```
# Select relevant variables and arrange by age
select_data <- data %>%
  arrange(`Age`) %>%
  select (`ID`, `Sleep efficiency`, `Age`, `Gender`)
select_data %>% glimpse()
```

Next, I would use R to find the correlation between the independent and dependent variables (Age and Gender). Afterwards, I would find the slope of the plots, to see how much of an impact each of the variables has. Then, I will create null and alternative hypotheses and run a p-value test to determine whether or not we can determine if we can conclude that Age and Gender are significant or not. Last, I will create data visualisations.

Analysis - Question 1 - Nicholas

Using R, the correlation between Age and Sleep Efficiency was found to be 0.09835669. The low correlation shows that even though sleep efficiency does indeed increase with age, their relationship is relatively weak. There are likely other variables that affect sleep efficiency more directly than age.

```r
# Correlation for Age vs Sleep Efficiency
cor_age_efficiency <- cor(select_data$`Age`, select_data$`Sleep efficiency`)
cat("The correlation between Age and Sleep Efficiency is", cor_age_efficiency, "\n")
```

```
## The correlation between Age and Sleep Efficiency is 0.09835669
```

I repeated the same for Gender, and found its correlation to be 0.01483306. Again, the near-zero correlation shows that the relationship between Gender and Sleep Efficiency is very weak, and there are likely other variables that affect sleep efficiency much more directly.

```r
# Correlation for Gender vs Sleep Efficiency
```

```r
cor_gender_efficiency <- cor(select_data2$`gender_num`, select_data$`Sleep efficiency`)
cat("The correlation between Gender and Sleep Efficiency is", cor_gender_efficiency, "\n")
```

```
## The correlation between Gender and Sleep Efficiency is 0.01483306
```

Next, I ran a linear regression model for Age and Gender against Sleep Efficiency using R, and found the slope of the change of sleep efficiency. The slope with respect to Age and Gender were found to be 0.00100981 and 0.002718515 respectively.

```r
# Linear regression for Age vs Sleep Efficiency
age_model <- lm(`Sleep efficiency` ~ `Age`, data = select_data)
summary(age_model)
```

```r
slope_age <- summary(age_model)$coefficients[2, 1]
cat("The slope of the change of sleep efficiency with respect to Age is", slope_age, "\n")
```

```
## The slope of the change of sleep efficiency with respect to Age is 0.00100981
```

```r
# Linear regression for Gender vs Sleep Efficiency
gender_model <- lm(`Sleep efficiency` ~ `Gender`, data = select_data)
summary(gender_model)
```

```r
slope_gender <- summary(gender_model)$coefficients[2, 1]
cat("The slope of the change of sleep efficiency with respect to Gender is", slope_gender, "\n")
```

```
## The slope of the change of sleep efficiency with respect to Gender is 0.002718515
```

The null hypothesis would be: ‘Age and Gender have NO effect on sleep efficiency’. The alternative hypothesis would be: ‘Age and Gender HAVE an effect on sleep efficiency’.

In order to conclude firmly that the two factors have no impact, a p-value test (with an alpha value of 0.05) was run. The p-value for gender was 0.8311. which meant that we could not conclude that there is an effect.

```
summary(age_model)
```

```
##
## Call:
## lm(formula = `Sleep efficiency` ~ Age, data = select_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30680 -0.09583  0.03541  0.11355  0.20046
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.7482353  0.0204116  36.657   <2e-16 ***
## Age         0.0010098  0.0004816   2.097   0.0366 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1347 on 450 degrees of freedom
## Multiple R-squared:  0.009674,   Adjusted R-squared:  0.007473
## F-statistic: 4.396 on 1 and 450 DF,  p-value: 0.03658
```

```
summary(gender_model)
```

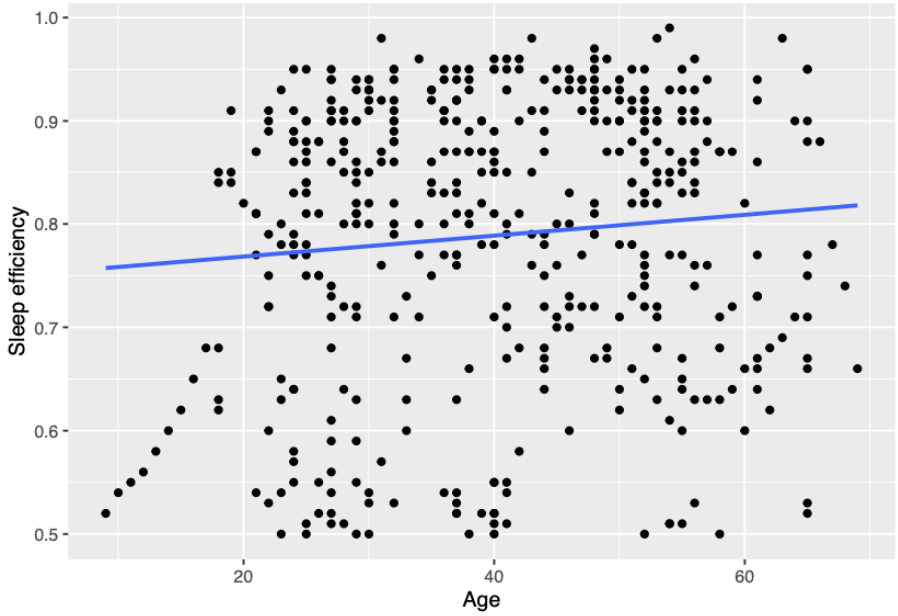
```
##
## Call:
## lm(formula = `Sleep efficiency` ~ Gender, data = select_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29026 -0.09208  0.03245  0.11245  0.19974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.787545   0.009045  87.065   <2e-16 ***
## GenderMale  0.002719   0.012736   0.213    0.831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1354 on 450 degrees of freedom
## Multiple R-squared:  0.0001012,  Adjusted R-squared:  -0.002121
## F-statistic: 0.04556 on 1 and 450 DF,  p-value: 0.8311
```

The p-value for age was actually below our alpha threshold, and was found to be 0.0366. Therefore, we can conclude that age has an effect on sleep efficiency, albeit not too significant.

Visualisations - Question 1 - Nicholas

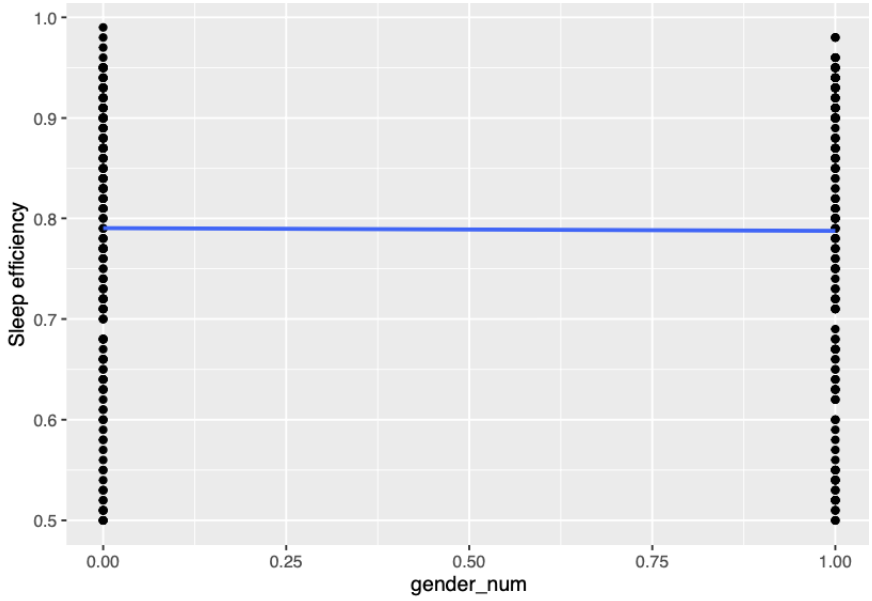
The next step was to provide visualisations and correlations for each of the findings, Age and Gender, using linear regression methods. The data visualisation tool that will be used will be two scatter plots, with the independent variables (i.e., Age and Gender) on the x-axis, and the dependent variable (i.e., Sleep Efficiency) on the y-axis:

```r
# Correlation plot for Age vs Sleep Efficiency
ggplot(select_data, aes(x = `Age`, y = `Sleep efficiency`)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```




```r
# Correlation plot for Gender vs Sleep Efficiency
ggplot(select_data2, aes(x = `gender_num`, y = `Sleep efficiency`)) +
  geom_point() +
```

```r
  geom_smooth(method=lm, se=FALSE)
```



(Note that ‘0’ represents Male, and ‘1’ represents Female).

Discussion - Question 1 - Nicholas

Overall, it is clear that there is very little correlation between Age and Sleep Efficiency, and virtually no correlation between Gender and Sleep Efficiency.

From the linear regression plot, it was found that for every increase in age, sleep efficiency generally increases by only 0.00100981. Furthermore, females only see 0.002718515 advantage over males in sleep efficiency. This shows that both age and gender have very little impact on sleep efficiency.

While it may be disappointing to see that there is such little correlation between these sociodemographic factors, it is still very important for our research. This is because having answered this question, we can eliminate these factors, and narrow down the possible factors that DO impact sleep efficiency.

Question 2 - Philip

The question I seek to answer is

> "What is a plausible range of values for the percentage of REM, deep, and light sleep."

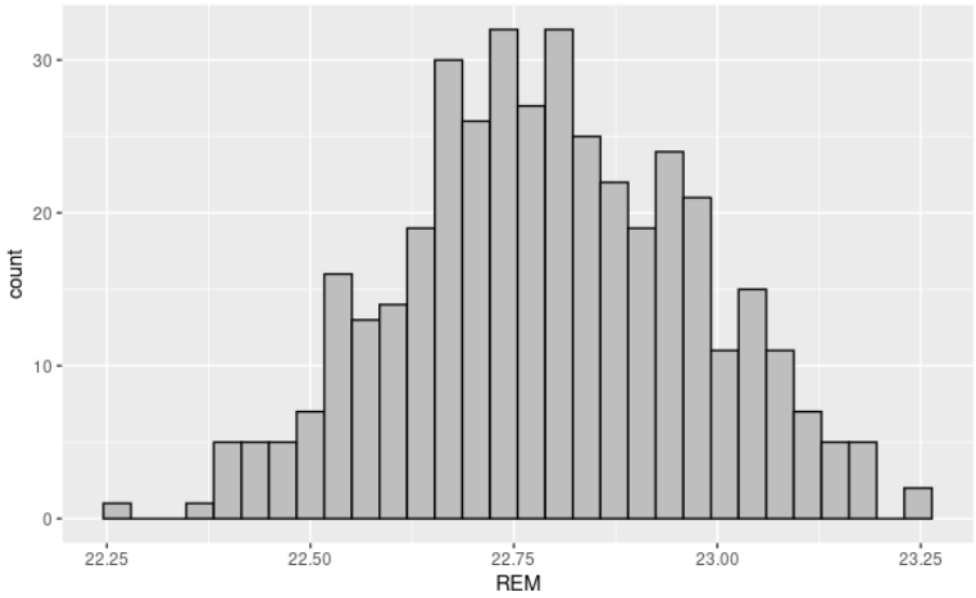
Method - Question 2 - Philip

To determine a plausible range of values for the percentage of REM, deep, and light sleep, I will use bootstrapping in R. The dataset we will be using contains ~450 observations and because bootstrapping is less effective the closer the bootstrap samples are to the actual observations, I will create 400 bootstrap samples. Using the bootstrap samples, I will calculate and store the mean values of each stage of sleep and construct sampling distributions. With the quantile function, I can calculate from the sampling distributions a 95% confidence interval the range of plausible values of the 3 stages of sleep. Overall, using this approach will provide a reliable and repeatable method for estimating a range of values for the percentage of different sleep stages.

Results & Visualizations - Question 2 - Philip

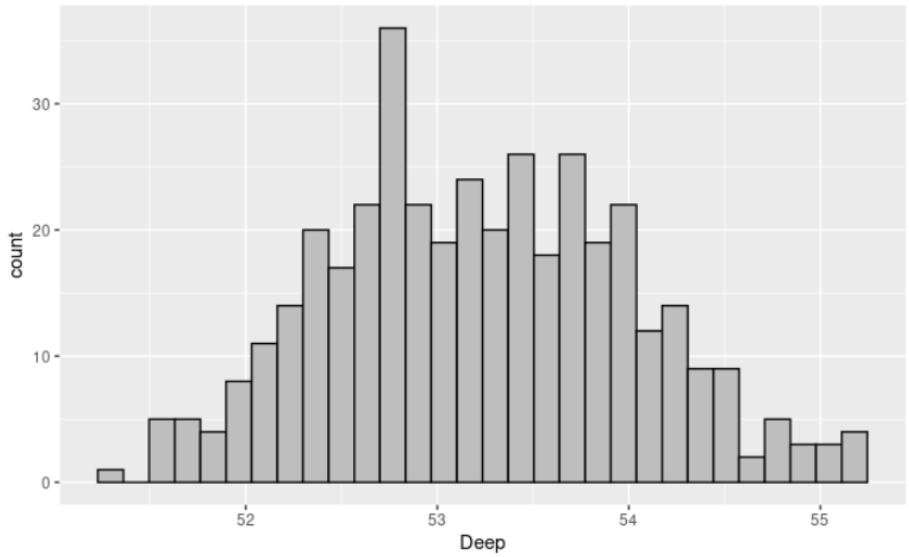
Through the use of bootstrapping in R, the following sampling distributions were made:

Sampling distribution for REM sleep:



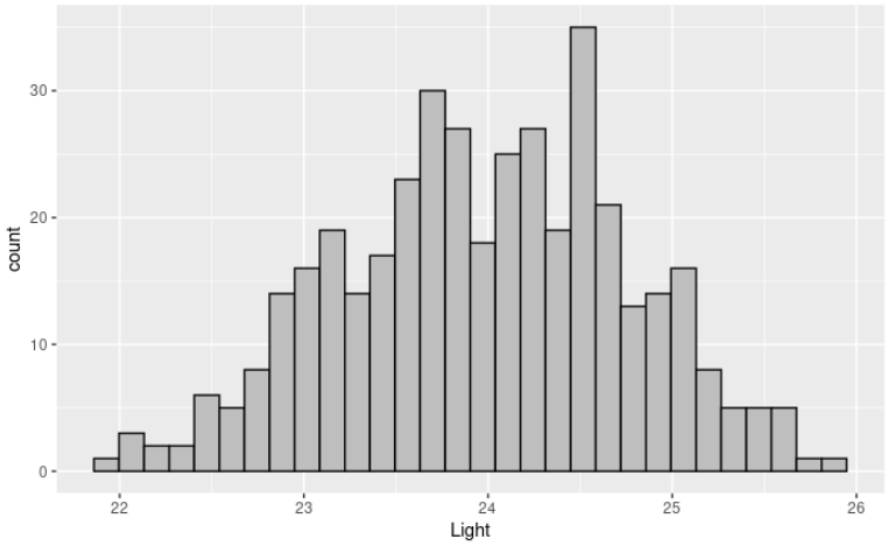
From the following histogram, I am able to compute the 95% confidence interval which gives the values, (22.45 - 23.14). The range of possible values of Rem sleep percentage is between 22.45% and 23.14% with 95% confidence.

Sampling distribution for Deep sleep:



From the following histogram, the 95% confidence interval gives the values, (51.75 - 54.83). The range of possible values of Deep sleep percentage is between 51.75% and 54.83% with 95% confidence.

Sampling distribution for Light sleep:



From the following histogram, the 95% confidence interval gives the values, (22.43 - 25.44). The range of possible values of Deep sleep percentage is between 22.43% and 25.44% with 95% confidence.

Discussion - Question 2 - Philip

Looking at sleep stages from a physical standpoint, it would make most sense for deep sleep, the recovery stage, to take up the majority of a night's sleep. This is so people can feel rejuvenated when they wake up. According to the data and results, it backs up the claim as the plausible range of values is above 50%. The values are important to uncover because it can be used as a comparison or guide to know what percentages are generally acceptable. For example, if someones deep sleep is far below the plausible range of values for percentage of Deep sleep, it can indicate that something is wrong. It is also important to mention that the data the plausible values are based on is very diverse. For example, there is a variety of people who smoke and dont smoke and people who drink heavily and don't. A better guide to compare values to is using a dataset that has the same factors as the user and generating a range of values using the same method. It is important to consider that bootstrapping assumes that the sample dataset is representative of the population. Therefore, the results I obtain from this method are valid if this condition is true. Although I know how the dataset was obtained, through surveys and sleep monitoring devices, the sample population is unknown.

Question 3 & Individual Contribution Statement - Daniel

The research question guiding my analysis is:

"Can a multiple regression model accurately predict hours in deep sleep and sleep efficiency based on smoking, exercise, caffeine and alcohol consumption and which ones are the best for predicting so?"

Data:

The data used is derived from a dataset obtained from the website Kaggle. It contains 452 observations and 17 variables; on sleep data such as deep sleep percentage, sleep efficiency, time of sleep and behaviours such as smoking, alcohol consumption, caffeine consumption and exercise frequency. To wrangle the data to answer my research question, I had to do four key things: filter out the NA values, remove outliers, change the column names and create new columns. I identified the columns I would be using and decided to filter the NA values out if they were in these columns specifically, instead of all of the columns to have as much data as possible to work with.

```
```{r}
#cleaning data
sleep <- sleep %>%
  filter(!is.na(`Sleep efficiency`) & !is.na(`Caffeine consumption`) & !is.na(`Alcohol consumption`) &
!is.na(`Smoking status`) & !is.na(`Exercise frequency`) & !is.na('Deep sleep percentage'))
```

Next, the column names in the dataset had spaces in the names; e.g. 'Sleep duration' which, after a preliminary test, caused problems in the lm() (fitting linear models) function. Thus, the spaces in the variable names were replaced with an underscore.

```
lm('Sleep duration'~ 'Awakenings', data = sleep)
```

```
Error in terms.formula(formula, data = data) : invalid term in model formula
```

```
#changed the column names because the names with two words had spaces which causes an error in lm()
function, spent a lot of time realizing I had to replace all the column names
colnames(sleep) <- c("ID", "Age", "Gender", "Bedtime", "Wakeup_time", "Sleep_duration",
"Sleep_efficiency", "REM_sleep_percentage", "Deep_sleep_percentage", "Light_sleep_percentage","Awakenings",
                     "Caffeine_consumption", "Alcohol_consumption", "Smoking_status", "Exercise_frequency")
```

Another issue was the fact that a factor in my research question was "hours in deep sleep" which is not provided. However, the dataset does provide "deep sleep percentage" and "sleep duration", from which I can solve for time in deep sleep. An issue I picked up on was the possible difficulty for viewers of the dataset to convert decimals of hours to hours;
```

for example: 1.98 hours to 1h58 minutes. I wrote a function that would calculate the more legible time and mutated the dataset by adding both of these columns for ease of reading.

```r
#since the research question is about time spent in deep sleep, which is not provided, I needed to solve for it.
#use column deep_sleep_percentage and sleep duration to find the hours in deep sleep
Deep_sleep_decimal = (sleep$Deep_sleep_percentage/100)*sleep$Sleep_duration
#create a helper function to convert a fraction of time into characters for reader legibility e.g. 4.2 hours = 4h12mins
convert_time <- function(decimal_hours) {
  hour <- floor(decimal_hours)
  minute <- round((decimal_hours - hour) * 60)
  paste(hour, minute, sep = "h")
}
#add Deep_sleep_hours for ease of reading and Deep_sleep_decimals for predictions
new_sleep<-sleep %>%
  mutate(Deep_sleep_hours = convert_time(Deep_sleep_decimal),
         Deep_sleep_decimal = Deep_sleep_decimal)
```

Finally, to account for the assumption of outliers in linear regression, I will calculate the z-scores of all columns being used to determine and get rid of extreme outliers in the data, with a threshold of 3 standard deviations away (0.27% of the data in a normal distribution). These points will then be removed from the dataset. To ensure not too much data is being lost, the size of the new dataset and the size of the cleaned dataset will be compared.

```r
#check for outliers in data due to homoscedasticity assumption in linear regression
#using Z-Score
z_scores_x1 <- abs(scale(new_sleep$Caffeine_consumption))
outliers_x1 <- which(z_scores_x1 > 3)

z_scores_x2 <- abs(scale(new_sleep$Alcohol_consumption))
outliers_x2 <- which(z_scores_x2 > 3)

z_scores_x3 <- abs(scale(new_sleep$Exercise_frequency))
outliers_x3 <- which(z_scores_x3 > 3)

z_scores_x4 <- abs(scale(new_sleep$Deep_sleep_decimal))
outliers_x4 <- which(z_scores_x4 > 3)

z_scores_x5 <- abs(scale(new_sleep$Sleep_efficiency))
outliers_x5 <- which(z_scores_x5 > 3)

outliers <- c(outliers_x1, outliers_x2, outliers_x3, outliers_x4, outliers_x5)

data_clean <- new_sleep[-outliers, ]
#compare to see how many data points are outliers
nrow(data_clean)
nrow(new_sleep)
#conclusion they're pretty similar, new_sleep has 3 more points
```

Methods/Analysis:

1. Check the response variable and the explanatory variables for multicollinearity
2. Split the data into training and testing data (80/20)
3. Run a multiple regression model with the behaviours (caffeine consumption, alcohol consumption, exercise frequency and smoking status) and deep sleep, use the deep_sleep_decimal previously created and training data as your dataset for the lm() function. Pick one of the two response variables.
4. Determine the coefficients for each explanatory variable (use summary()), and determine the significance of each independent variable using their p-values.
5. Evaluate based on the F-statistic, P-value and R-squared value if this model is a good fit.
6. Create predictions using testing data and find the RMSE to determine how well the model is able to predict the target value(s). An easy way to determine if the RMSE is adequate is by comparing it to the scale of residuals found in step 3.
7. If there are insignificant explanatory variables, repeat the steps 3-5 without them
8. Compare the new RMSE values to the previous one. A smaller RMSE is better.
9. Find the normalized feature importance of the more significant model to determine the hierarchy of the features most important in predicting our response variable.
10. Repeat for the other response variable.

Results + Discussion:

Firstly, we will take a look at deep sleep. After running the multiple regression model through the VIF function I wrote, we notice that the 0.8 threshold is not met, removing the possibility of multicollinearity disrupting the data.

	variable <chr>	VIF <dbl>
(Intercept)	(Intercept)	1.455373e-02
Caffeine_consumption	Caffeine_consumption	5.205127e-06
Smoking_statusYes	Smoking_statusYes	1.408136e-02
Exercise_frequency	Exercise_frequency	1.490952e-03
Alcohol_consumption	Alcohol_consumption	1.223216e-03

Now we can analyse the model. To begin, we notice that the residuals range over (-2.942, 3.3025) suggesting that the model is not perfect in its predictions, however, the median being 0.198 indicates that the average prediction is generally accurate. We can see that smoking and alcohol consumption slightly decreases the time spent in deep sleep while exercising marginally increases it. Caffeine also appears to decrease deep sleep, however unlike the other variables, its p-value is above the 0.05 bar implying it's not statistically significant. The small multiple R-squared value of 0.135 indicates that there is variance not captured by the model. If this model was a linear regression, it would suggest the

possibility of confounding variables. Since this is a multiple regression using all the explanatory variables provided; the possibility of unaccounted for variables still exists, although the possibility of the true relationship between the predictor variables and the response variable being non-linear is more likely. Finally the F-statistic value of 12.41 and p-value of 2.164e-09 indicates that one of the predictor variables is significantly associated with the response variables.

```
Call:
lm(formula = Deep_sleep_decimal ~ Caffeine_consumption + Smoking_status +
    Exercise_frequency + Alcohol_consumption, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9420 -0.7903  0.1980  0.7963  3.3025

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.151960   0.141513  29.340  < 2e-16 ***
Caffeine_consumption -0.003260   0.002676  -1.218  0.22405
Smoking_statusYes    -0.407443   0.139197  -2.927  0.00367 **
Exercise_frequency    0.124618   0.045294   2.751  0.00628 **
Alcohol_consumption  -0.224184   0.041026  -5.464 9.39e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.173 on 318 degrees of freedom
Multiple R-squared:  0.135,     Adjusted R-squared:  0.1241
F-statistic: 12.41 on 4 and 318 DF,  p-value: 2.164e-09
```

Now I will create predictions and find the root mean squared error (RMSE) to determine their quality.

```
#check rmse to see how well model is able to predict target value
rmse <-sqrt(mean((testData$Deep_sleep_decimal - predictions)^2))
rmse
```

```
[1] 1.015237
```

Comparing it to our residual scale, 1.015237 is pretty good. However, it takes into account what appears to be a statistically insignificant variable so I will remove it to see if the quality of my predictions change.

This time, I created multiple regression without “Caffeine_consumption”. The removal caused an increase in the maximum value, decrease in R-squared value and a higher RMSE, all suggesting two possibilities: Caffeine consumption was providing useful predictive information but there is not enough data to detect a statistically significant effect, or if it’s due to random error (or multicollinearity but we already accounted for it).

```
Call:
lm(formula = Deep_sleep_decimal ~ Smoking_status + Exercise_frequency +
    Alcohol_consumption, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8588 -0.7533  0.1697  0.8377  3.3557

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.06881    0.12406  32.797  < 2e-16 ***
Smoking_statusYes  -0.42293    0.13872  -3.049  0.00249 **
Exercise_frequency  0.13027    0.04509   2.889  0.00413 **
Alcohol_consumption -0.21848    0.04079  -5.356 1.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.174 on 319 degrees of freedom
Multiple R-squared:  0.131,     Adjusted R-squared:  0.1228
F-statistic: 16.03 on 3 and 319 DF,  p-value: 9.927e-10
```

```
[1] 1.017964
```

I decided to conduct a two sample t test to compare the two RMSE values to attempt to see the significance of the Caffeine. Unfortunately, I tried numerous manners (such as bootstrapping and folded repeated cross validation) but none of them worked due to lack of data. I generated more data but in the end I still only got one RMSE value, meaning that there isn't enough variability in the data to come to a conclusion on whether or not caffeine happens to influence the linear regression or it's just due to chance (some sample code is provided in the RMD file). Thus the impact of caffeine is inconclusive.

Following this new discovery, it would be unwise to continue with the linear model with Caffeine consumption as explanatory variable. I decided to find the normalized feature importance of the model without Caffeine consumption.

```
coefficients <- coef(no_cc_model)[-1] #removes intercept
normalized_coef <- coefficients / sum(coefficients)
normalized_coef
```

```
  Smoking_statusYes  Exercise_frequency Alcohol_consumption
          0.8274268          -0.2548652           0.4274383
```

The normalized feature importance makes a few things clear: it reaffirms the extreme F-statistic value as we can see the Yes option for Smoking_status having an extreme correlation with time in sleep. We also determine the ranking for the most important factors in predicting time in deep sleep which appears to be Smoking status, Alcohol Consumption and Exercise Frequency. Again the effect of caffeine cannot be quantified with this data.

Now, sleep efficiency. Again, we first check for multicollinearity with a given threshold of 0.8.

	variable <chr>	VIF <dbl>
(Intercept)	(Intercept)	1.455373e-02
Caffeine_consumption	Caffeine_consumption	5.205127e-06
Smoking_statusYes	Smoking_statusYes	1.408136e-02
Exercise_frequency	Exercise_frequency	1.490952e-03
Alcohol_consumption	Alcohol_consumption	1.223216e-03

Similar to the previous multiple regression model, the range of residuals (-0.28218, 0.28019) hints that the predictions are not perfect but the average prediction according to the median (0.01596) appears to be decently accurate. The p-value of Caffeine consumption is again bigger than 0.05 implying statistical insignificance. The rest of the summary is pretty similar to the previous just with different values.

```
Call:
lm(formula = Sleep_efficiency ~ Caffeine_consumption + Smoking_status +
    Exercise_frequency + Alcohol_consumption, data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28218 -0.08388  0.01596  0.08526  0.28019

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.8021770  0.0141687  56.616  < 2e-16 ***
Caffeine_consumption  0.0002112  0.0002680   0.788    0.431
Smoking_statusYes    -0.0832324  0.0139368  -5.972 6.26e-09 ***
Exercise_frequency    0.0235128  0.0045350   5.185 3.85e-07 ***
Alcohol_consumption  -0.0292321  0.0041077  -7.116 7.39e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1174 on 318 degrees of freedom
Multiple R-squared:  0.275,     Adjusted R-squared:  0.2659
F-statistic: 30.16 on 4 and 318 DF,  p-value: < 2.2e-16
```

Moving to the RMSE, an immediate observation would be that unlike the previous model, it is much smaller; however compared to its smaller range of residuals, it looks reasonable.

```
[1] 0.1068091
```

Again, I will remove the caffeine due to its large p-value to see if it has an impact on the RMSE. The general trend in change is also similar to the previous summary (although the F-statistic is much bigger) alongside the RMSE becoming bigger, the possible reasons why and the incapability to generate data to get more than one RMSE due lack of variance in the data.

```
[1] 0.1069892
```

Proceeding to determine the ranked feature importance

```
coefficients <- coef(no_cc_model_two)[-1]
normalized_coef <- coefficients / sum(coefficients)
normalized_coef
```

```
   Smoking_statusYes  Exercise_frequency Alcohol_consumption
           0.9272124          -0.2609991           0.3337867
```

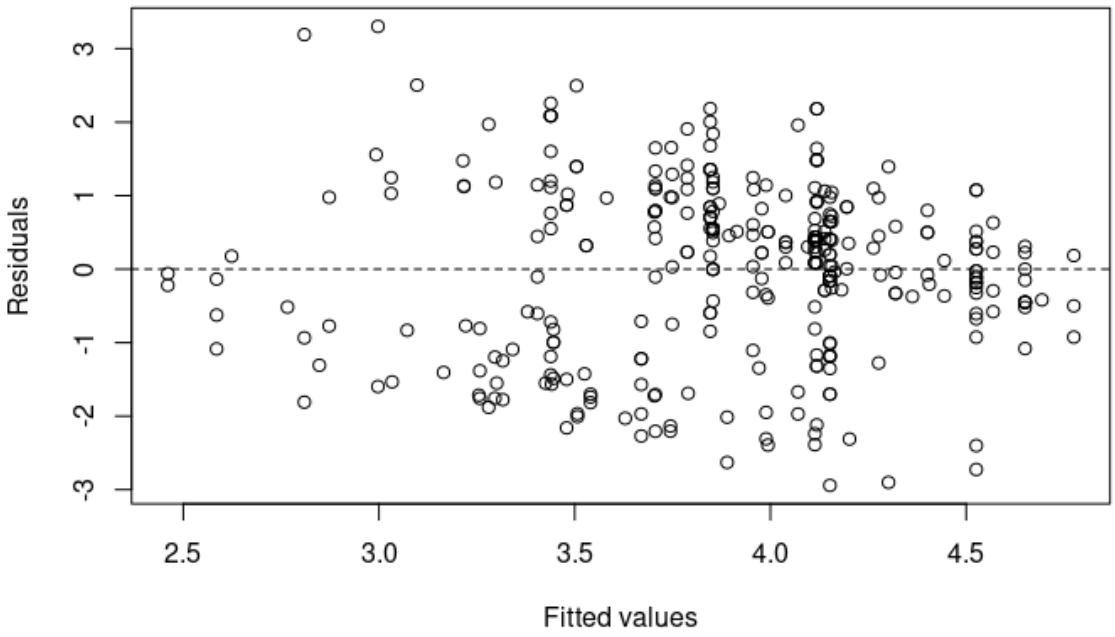
Smoking status takes a much larger importance in this model, going up by nearly 10%. The rankings for sleep efficiency are the same as deep sleep, being: smoking status first, alcohol consumption second and exercise frequency third. To restate, the true effect of caffeine cannot be quantified with this data.

Discussion (One Step Further):

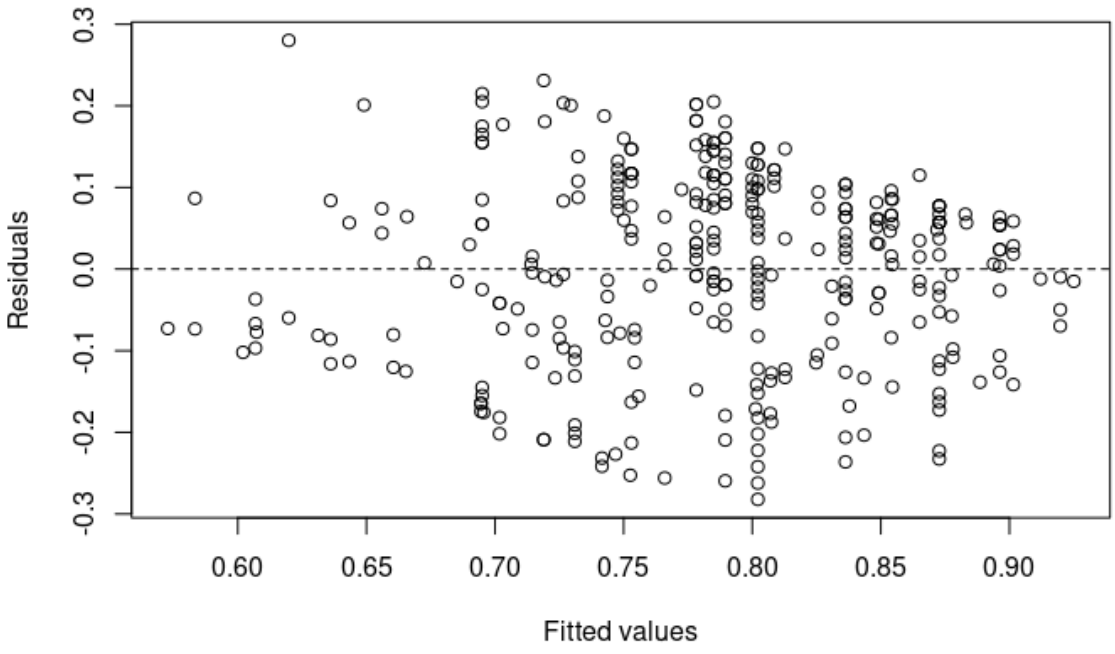
When pondering reasons why caffeine consumption did not appear statistically significant in the multiple regression model but did have an undetected impact on the model's predictive performance (since I couldn't create new data), I decided to go back to the roots and decided to see if I violated any assumptions of linear regression. Since we virtually did not touch on this at all in class, I decided to ask some TA's and the Professor for some guidance. They all unanimously agreed that for the first year level, I should worry about these four: homoscedasticity, non-linearity, missing data and outliers.

Following their advice, I had already accounted for missing data and outliers. To check for homoscedasticity, I decided to graph the residuals for all four multiple regression models:

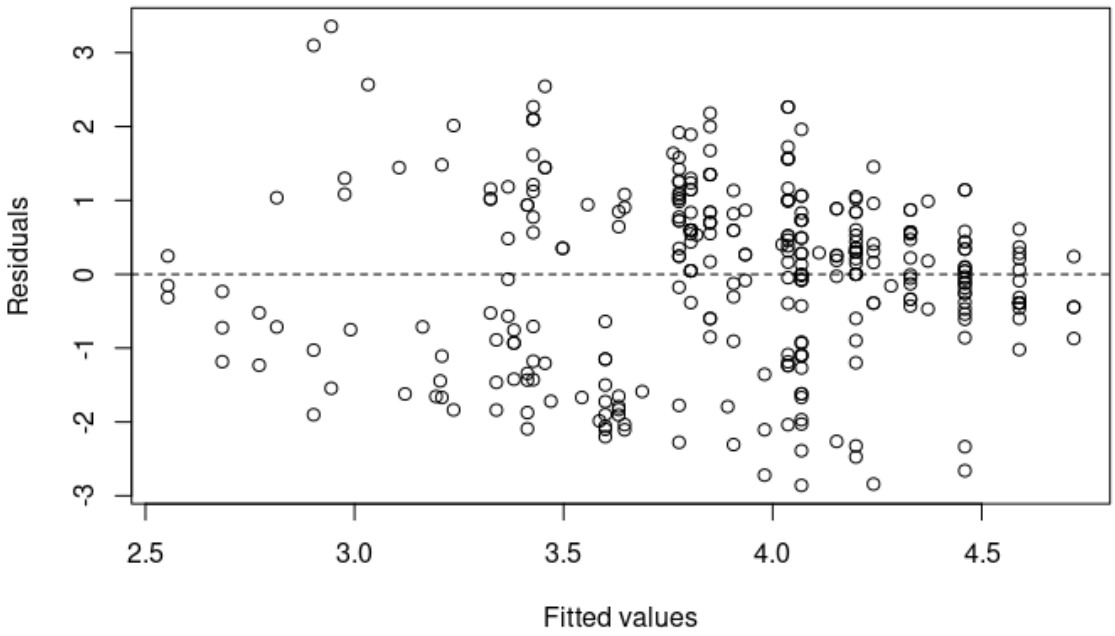
Deep Sleep with Caffeine Consumption:



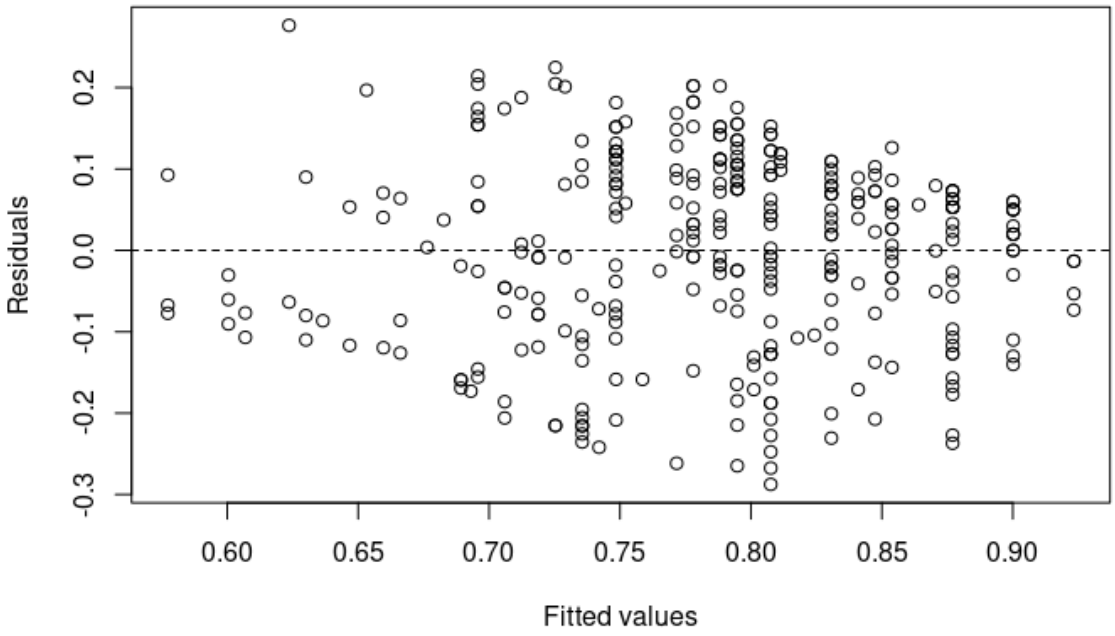
Deep Sleep without Caffeine Consumption:



Sleep Efficiency with Caffeine Consumption:

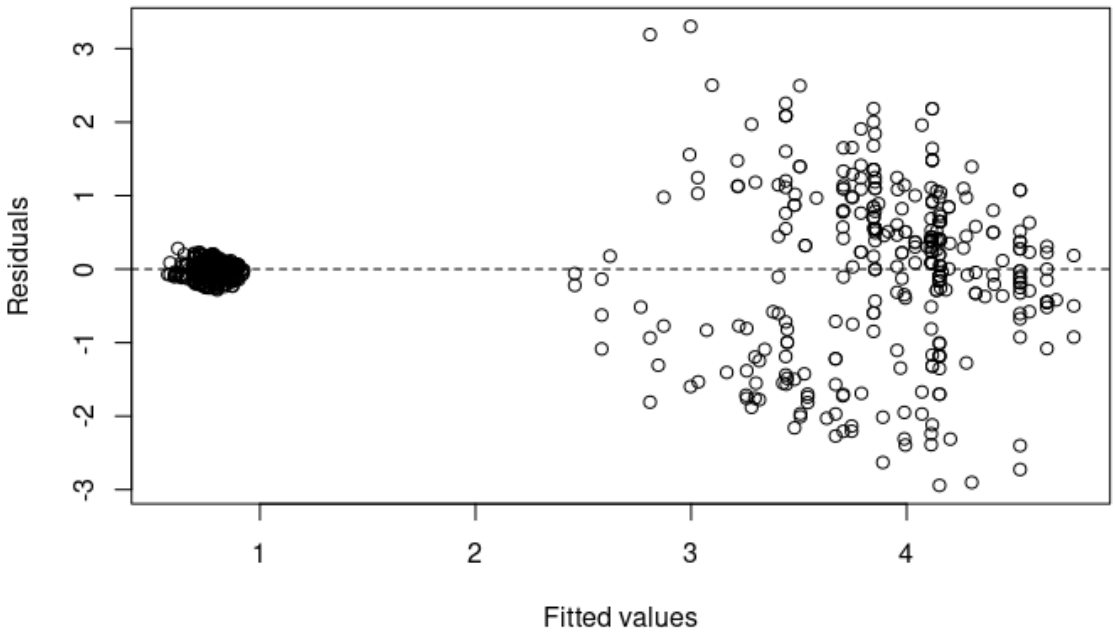


Sleep Efficiency without Caffeine Consumption:



All of which appear to be slightly biased (slight decline) and homoscedastic.

Factoring out the other three assumptions, this suggested non-linearity, however, I still wanted to come up with some proof to be able to conclude an increase in likeliness that this was true. I decided then to see what would happen if I compared all the explanatory variables to all the response variables to investigate the effects of the independent variables on both variables simultaneously. I then plotted this multivariate multiple regression:



This non-homogeneous spread of the residuals, a clear non-linear pattern, alongside the previously mentioned small multiple R-squared value and undetected impact of caffeine does imply that the relationship between the combined response variable and the explanatory variables, especially caffeine, may not necessarily be linear.

In context of the research question, we accept the null-hypothesis that a multiple regression cannot accurately predict hours in deep sleep and sleep efficiency based on caffeine consumption, alcohol consumption, exercise frequency and smoking status due likely to the potentially non-linear relationship between caffeine and the response variables. Though, it can do generally well without caffeine. The most important variables in terms of predicting both deep sleep and sleep efficiency are smoking status, then alcohol consumption, exercise frequency, with caffeine being inconclusive.

Appendix:

Daniel:

I made many decisions in an attempt to improve the flow and legibility of my final report, alongside changing parts of my research question and analysis to provide a more challenging yet rewarding experience. To begin, I will justify my choice of splitting discussion into two parts, one along with results (for placing my findings in a broader context and also interpreting my findings) and one primarily trying to determine the underlying problems with the data. When writing a draft for this report, I decided to write the results and discussion separately. The product was an extremely long and convoluted discussion which was only worsened by the fact that it was extraordinarily difficult to understand without the visualizations/results next to them. The current post-edit paper is more fluid and the logic is more traceable and thus a better read for any audience. Having the analysis of the underlying problems apart from the other discussion also places emphasis on the aspect of One Step Further as I did spend a few days learning the concepts required, instead of being in a multipage sub-category of text. Looking at the rubric for the final announcement, the concepts are grouped together further cementing my choice in doing so.

The choice to not have a visualization sub-category was to account for an already lengthy report, and they were provided to as previously mentioned for ease of understanding an otherwise wall of text.

As for changing the research question from:

“ Can a linear regression model predict hours in deep sleep and sleep efficiency based on behaviours such as smoking, exercise and caffeine and alcohol consumption?”

To:

Can a multiple regression model predict hours in deep sleep and sleep efficiency based on smoking, exercise, caffeine and alcohol consumption and which ones are the best for predicting so?

Truth be told, knowing how important this project means to our grade, I wanted to demonstrate my understanding of core concepts learned in class and apply them while also exhibiting my capability to learn and apply new concepts. When I initially started investigating my previous research question, the results didn’t allow me to show any of the qualities previously mentioned. A multiple regression is also more in line with what I actually initially wanted to do but was unaware of. The longer length of my report is my attempt at a holistic analysis and trying to cover all possible loose ends, and I do feel like I did put in quite a decent amount of time into this assignment to do so.

Philip:

A change made to question 2 between the progress report and the final report is that instead of focusing on a test statistic, it shifted towards a focus on finding a plausible range of values. This came from feedback from the progress report that highlighted what the change to the question should be which I accepted.

Conclusion

In summation, this report focuses on searching for a better understanding of the factors in the dataset that contributes to healthy sleep patterns, sleep efficiency, and to set a precedent to aid future researchers on potential areas to examine further. We found that lifestyle factors like caffeine consumption and smoking, are more significant factors to sleep efficiency than non-lifestyle factors, which include age and gender, which have little to no correlation with sleep results. In addition, the non-linear relationship between the variables in the data, primarily caffeine, should deter future analysis from using linear or multiple regression and try to determine the true impact of caffeine consumption on aspects of sleep. Furthermore, bootstrapping provided plausible values for three types of sleep stages, with deep sleep being the largest range of percentages.

Understanding our outcomes is pivotal due to the possibility of significant contributions on the future research in sleep, as well as suggestions for fellow society members to improve their overall quality of sleep, a major determinant in mood. Another possible use of this data is for businesses who can use this data to meet consumer's sleeping needs or governments to improve the aforementioned attitude, and thus happiness of its populus. To conclude, our report shows the potential of the revolution of data and its impact on the world's sleep by finding hidden patterns and relationships from both lifestyle and non-lifestyle factors.

Individual Contribution Statement - Nicholas

Since the group decided to divide the tasks by question, all aspects regarding this question, including the cleaning of data, formation of the question, visualisations, and analysis for this question were completed by me (Nicholas). Furthermore, I helped write the group portions of this report, which includes the title, the abstract, introduction, citations, and overall editing of the report.

It is important to note that the research question that I was trying to answer initially did not work out as planned due to the incompatible data. I decided to take another path, and come up with another research question, which meant pushing back much of our

project timelines. Fortunately, the original timeline allowed for some flexibility, which we took advantage of. This obstacle actually taught an important lesson, which was to be open to the idea of pivoting from the original objective. Furthermore, when coming up with ideas for the new question, I noticed an extensive number of possible questions that could be answered, given the same data set, which proved how vast data science is.

Individual Contribution Statement - Question 2 - Philip

I contributed to the entire process of answering question 2 including the formation, methods, analysis, coding, and discussion. I also wrote the first drafts of the abstract, introduction, and conclusion.

I also changed my research question using the feedback provided in the progress report. The original question had a simple test statistic calculation and was not appropriate for the project. Instead, my question changed to focus on the second part which is finding plausible values for the percentage of REM, deep, and light sleep.

Individual Contribution Statement: Daniel

I, Daniel Sun performed all aspects with question 3 e.g. wrangling the data in specific manner and analysis. I entirely revamped the abstract and introduction while improving on the conclusion. I went out of my way to help with formatting issues from my groupmates, such as titles of subcategories being at the bottom of the page with the next in the text and improving the general aesthetic of the report. I found the dataset and shared it with the group.

It should be taken into account that I had to change my research question and methodology due to a desire to challenge myself and I had to recreate, design and analyze within a time frame outlined in the project proposal. Thankfully, our original timeline allowed some leeway and I was able to complete the assignment on time. I had many personal take-aways from the project and if I had to pick one is the importance of planning ahead and factor in possible issues, otherwise I would've been unlikely to finish this report.

Citations

Equilibruimm. (January 2023). Sleep Efficiency Dataset, Version 3. Retrieved 19 March 2023 from https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency.