Sleep, Health and Lifestyle Analysis

Harman Singh Saggu T00727652, Raunaq Singh Dev T00737367, Nwaokenneya Precious T00727498

2023-12-01

Contents

1 Introduction

1.1 Overview

This report discusses the results of the work done in finding insights from Sleep, Health and Lifestyle data set. It aims at highlighting important key factors that affect the sleep and health patterns of subjects belonging to different occupations. Findings through various hypothesis testings are presented to emphasize on factors responsible for disrupting sleep and health patterns of an individual.

1.2 Data Description

The **Sleep, Health and Lifestyle** data set comprises of data which provides information regarding sleep duration, quality of sleep, stress level, physical activity level and sleep disorder in both male and female subjects belonging to different occupations. The data has been sourced from https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

1.3 Objective

The main of objective of the project was to use different hypothesis testing methods to deduce which key factors were responsible for change in sleep patterns and health of various individuals. It is primarily focused on sleep duration, quality of sleep and stress level of subjects to find which of these factors played a crucial role in analyzing changes in sleep and health patterns of the subjects in the data set.

1.4 Methodology

Hypothesis testing methods used for the project include one-sample tests, two-sample tests, normality checks, Kruska-Wallis test, independence test and implementation of a simple linear regression model. The confidence level for all the tests was chosen to be 95 percent(alpha = 0.05).

1.5 Tools Used

R programming language, R Studio and statistical techniques from ADSC 1000 course were used in the project.

2 Sample Description

2.1 Sample Data

The sample data has been focused on and derived from the variables *Sleep Duration, Quality of Sleep and Stress Level* of total subjects in the data set. Sleep duration provides the number of hours a subject sleeps, quality of sleep is measured on a scale of 1-10 and stress level is also measured on a scale of 1-10. These are the main variables which have been used to perform hypothesis testing on the target population.

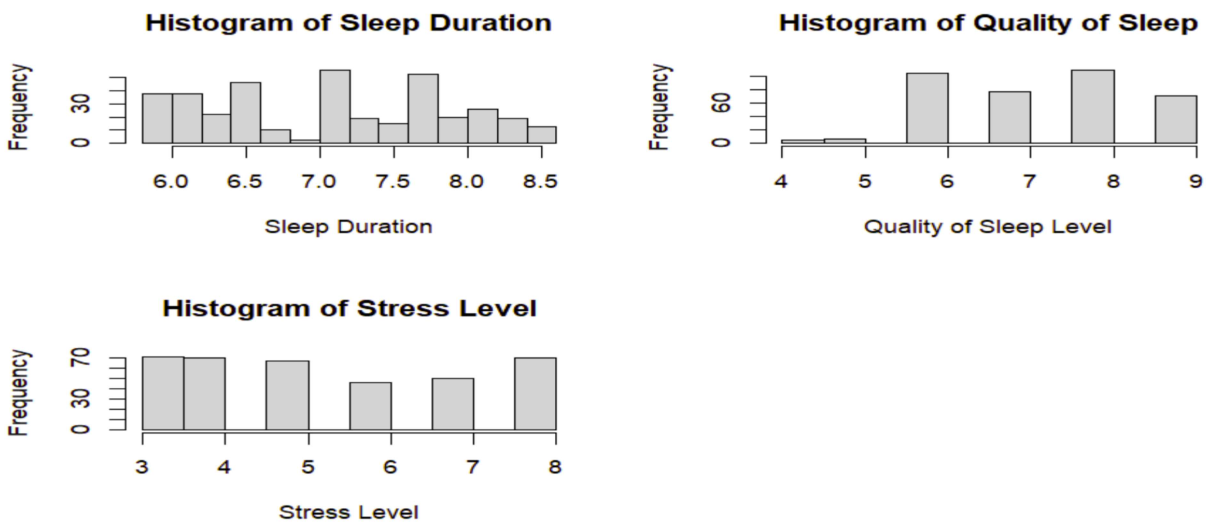

Figure 1: Histograms of Sample Data

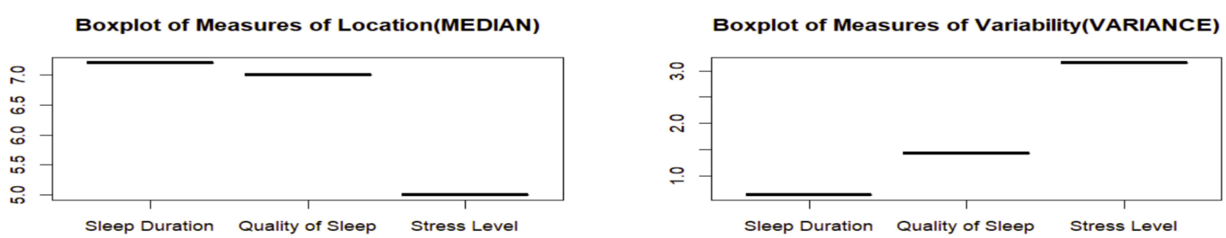

Figure 2: Measures of Location and Variability

2.2 Target Population

Target population for hypothesis testing involves subjects belonging to different occupations to infer the impact of key factors affecting their sleep health. Occupations that were chosen for this project are Engineer, Doctor, Scientist, Salesperson, Lawyer, Teacher and Nurse.

Occupation <chr>	n <int>
Accountant	37
Doctor	71
Engineer	63
Lawyer	47
Manager	1
Nurse	73
Sales Representative	2
Salesperson	32
Scientist	4
Software Engineer	4
Teacher	40

11 rows

Figure 3: Target Population

The quality of sample is good as it fairly represents the target population and does not contain any **NA** values.

3 Hypothesis Testing

3.1 One-Sample Tests

3.1.1 One-Sample Mean Tests

One-sample mean tests were done using t.test() function to check the following hypothesis:

1. The mean sleep duration of total subjects(in hours) was less than 7(alternative hypothesis). The null hypothesis(H0) was assumed that sleep duration >= 7.

2. The mean stress level for all the subjects was greater than 7(H1). H0 assumed was stress level <= 7.

3. Mean quality of sleep of total subjects is less than 7(H1). H0 assumed was quality of sleep >= 7.

3.2 Two-sample tests

Two-sample tests were focused on the target population and the factors affecting their sleep health.

3.2.1 Two-sample mean tests

Following two-sample mean tests were performed using t.test() function to check the hypothesis:

1. If stress level of Engineer was less than stress level of Scientist(H1). H0 assumed was stress level of Engineer >= Scientist.

2. If stress level of Salesperson was less than Stress Level of Doctor(H1). H0 asumed was stress level of salesperson >= doctor.

3. If quality of sleep of teachers was greater than quality of sleep of engineers(H1). H0 in this case was quality of sleep of teacher <= doctor.

4. If sleep duration of nurses was less than sleep duration of lawyers(H1). H0 in this case was assumed sleep duration of nurse >= lawyer.

3.2.2 Two-sample variance test

The two-sample variance test was done using var.test() function to check the hypothesis if there is variance in physical activity level between male and female subjects. In this case:

H0: Variance in physical activity is same for both male and female

H1: Variance in physical activity is not same for both the genders

3.2.3 Two-sample proportion test

This hypothesis test was done using prop.test() function to check if proportion of subjects with stress level greater than 6 were more than proportion of subjects with stress level less than 6.

H0: proportion of subjects with stress level < 6 is more.

H1: proportion of subjects with stress level > 6 is more.

3.3 Independence test

Independence test was done between variables using chi-squared test function to check dependency between variables. Hypothesis were as follows:

1. If quality of sleep is independent of sleep of duration
2. If stress level is independent of quality of sleep
3. If sleep disorder is independent of stress level

3.4 Normality check

The data for sleep duration, quality of sleep and stress level was checked using QQ-plots and Shapiro-Wilk tests. Shapiro-Wilk test results showed that the p-value was less than alpha(0.05) for all the data variables and suggested that the does not come from normal distribution. QQ-plots suggested the same.

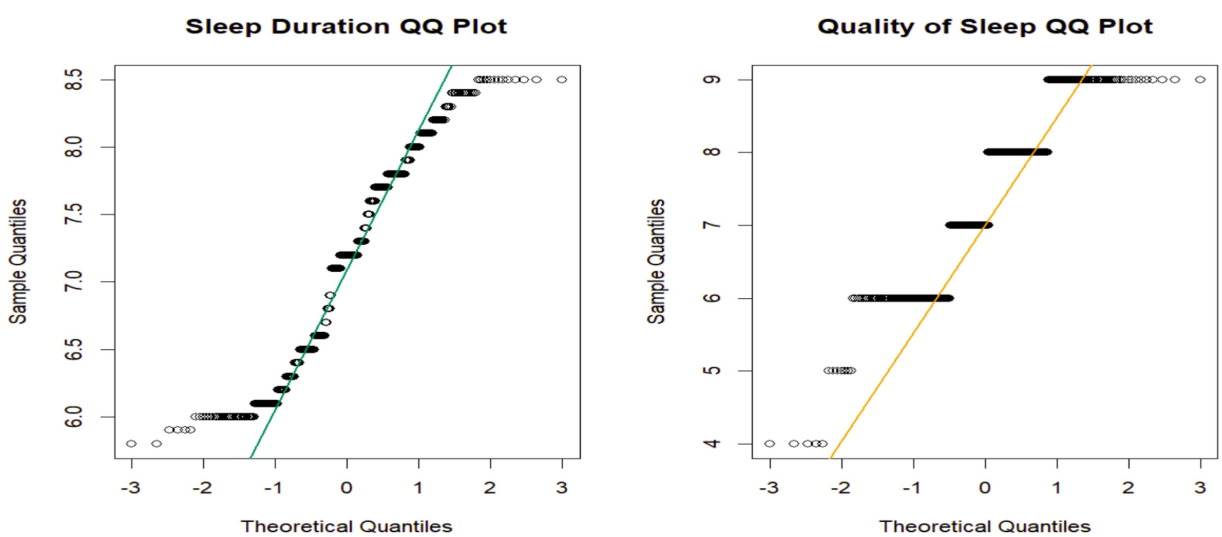

Figure 4: QQ-Plot of Sleep Duration and Quality of Sleep

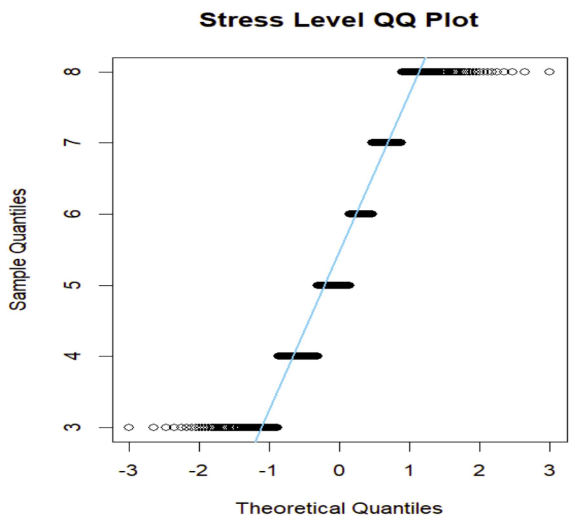

Figure 5: QQ-Plot of Stress Level

3.5 Kruskal-Wallis test

Since the sample data was not normally distributed, Kruskal-Wallis test, which is non-parametric and can be used if assumptions for ANOVA test are violated, was performed to check variances in means across different group for the following hypotheses:

1. If mean quality of sleep is same for both male and female
2. If mean sleep duration is same for both male and female
3. If mean stress level is same for both male and female

3.6 Simple Linear Regression

A simple linear regression model was made to find correlation between factors affecting sleep health of the subjects. Interaction terms were also included to see if any independent variables affect the outcome of other independent variables and AIC tests were performed to check which model was a better fit and of high quality.

4 Results

For **one-sample mean** tests, the mean sleep duration of total subjects was found to be >= 7 since the p-value was greater than alpha(0.05) and we failed to reject the H0. **Mean quality of sleep** of total subjects was also found to be >= 7 as the p-value was greater than alpha in this case as well taking H0 into consideration. **Mean stress level** was found to be <= 7 as we also failed to reject the H0 due to higher p-value.

The results of **two-sample mean** tests were as depicted in the table below:

Two Sample Mean Tests (T-test)	P-value (alpha = 0.05)	Conclusion
H0: Stress level of Engineer >= Stress level of Scientist **H1: Stress level of Engineer < Stress level of Scientist**	0.005014	We reject H0.
H0: Stress level of Salesperson >= Stress level of Doctor **H1: Stress Level of Salesperson < Stress level of Doctor**	0.9475	We fail to reject H0.
H0: Quality of sleep of Teacher <= Quality of sleep of Doctor **H1: Quality of sleep of Teacher > Quality of sleep of Doctor**	1	We fail to reject H0.
H0: Duration of sleep of Nurse >= Duration of sleep of Lawyer **H1: Duration of sleep of Nurse < Duration of sleep of Lawyer**	0.002809	We reject H0.

Figure 6: Two-sample mean tests

In **two-sample variance** test, variance in physical activity for both male and female subjects was found to be similar since p-value from the test result was greater than alpha. So, we failed to reject the H0.

In case of **two-sample proportion** test, the proportion of subjects with stress level > 6 was found to be more since p-value from the test result was very small, so the H0 was rejected and H1 was considered.

In **independence-test** results of all cases p-value was found to be very small than alpha, indicating that quality of sleep and sleep duration, stress level and quality of sleep, sleep disorder and stress level affected each others outcomes.

Kruskal-Wallis tests suggested that mean quality of sleep, mean sleep duration and mean stress level is different in both male and female since p-value was found to be very small than alpha.

For **simple linear regression**,a correlation matrix was created which clearly suggested a high positive correlation between quality of sleep and sleep duration, a high negative correlation between stress level and quality of sleep.

```
                        Sleep.Duration Quality.of.Sleep
Sleep.Duration               1.0000000        0.8832130
Quality.of.Sleep             0.8832130        1.0000000
Physical.Activity.Level      0.2123603        0.1928965
Stress.Level                -0.8110230       -0.8987520
Heart.Rate                  -0.5164549       -0.6598647
                        Physical.Activity.Level Stress.Level Heart.Rate
Sleep.Duration                       0.21236031  -0.81102303 -0.5164549
Quality.of.Sleep                     0.19289645  -0.89875203 -0.6598647
Physical.Activity.Level              1.00000000  -0.03413446  0.1369710
Stress.Level                        -0.03413446   1.00000000  0.6700265
Heart.Rate                           0.13697098   0.67002646  1.0000000
```

Figure 7: Correlation Matrix

Then, a **simple linear regression model** was made by choosing Quality of Sleep as the dependent variable(Y) and the independent variables(X) were chosen to be sleep duration, heart rate, age, physical activity level and stress level. The coefficient estimates of sleep duration and stress level suggested a high positive and high negative correlation with Y, respectively. Their significance of estimates further suggested that these independent variables have a strong impact on Y. Around 90.48 percent of variance in Y was explained by X according to adjusted R-squared value and p-value suggested that independent variables have a strong impact on Y.

Then interaction terms Stress level and Sleep disorder were added to the model to check the impact of former on the latter. The model suggests that stress level affects the outcome of sleep disorder as the coefficient estimate changes from positive to negative and the adjusted R-squared value suggests that around 92.59 percent of variance in Y is explained by X. A small p-value suggested strong impact of X on outcomes of Y.

```
Call:
lm(formula = Quality.of.Sleep ~ Sleep.Duration + Age + Heart.Rate + 
    Stress.Level * Sleep.Disorder, data = sleep_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.08015 -0.19472 -0.01441  0.16834  1.04323 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          6.328953   0.533323  11.867  < 2e-16
Sleep.Duration                       0.427028   0.043595   9.795  < 2e-16
Age                                  0.029290   0.002972   9.856  < 2e-16
Heart.Rate                          -0.035081   0.006503  -5.395 1.24e-07
Stress.Level                        -0.240722   0.030460  -7.903 3.25e-14
Sleep.DisorderNone                   0.860130   0.194296   4.427 1.26e-05
Sleep.DisorderSleep Apnea            1.136743   0.216344   5.254 2.53e-07
Stress.Level:Sleep.DisorderNone     -0.038492   0.032613  -1.180    0.239
Stress.Level:Sleep.DisorderSleep Apnea -0.139621   0.033566  -4.160 3.98e-05

(Intercept)                         ***
Sleep.Duration                      ***
Age                                 ***
Heart.Rate                          ***
Stress.Level                        ***
Sleep.DisorderNone                  ***
Sleep.DisorderSleep Apnea           ***
Stress.Level:Sleep.DisorderNone        
Stress.Level:Sleep.DisorderSleep Apnea ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3258 on 365 degrees of freedom
Multiple R-squared:  0.9275,	Adjusted R-squared:  0.9259 
F-statistic: 583.7 on 8 and 365 DF,  p-value: < 2.2e-16
```

Figure 8: Simple Linear Regression Model with Interaction Terms

Akaike Information Criterion (AIC) tests were done on both the models(with and without interaction terms) and the results were as follows:

1. Model without interaction terms: 324.1469

2. Model with interaction terms: 233.4027

The results suggested that the model with interaction terms had less loss of information and was found to be a better fit for the problem.

5 Conclusion

From the results of hypothesis tests and simple linear regression models we concluded the following:

1. **Stress level** and **sleep duration** are major factors affecting the sleep, lifestyle and health of the subjects.
2. Subjects with **greater stress level** were found to have a **lesser quality of sleep**.
3. Subjects with **greater sleep duration** experienced **better quality of sleep**.
4. Subjects with **mean stress level** $\boldsymbol{> 6}$ were found in greater proportion.
5. Sleep duration has a **positive correlation** with Quality of sleep.
6. Stress level has a **negative correlation** with Quality of sleep.

List of Figures

6 Abbreviations

H0: Null hypothesis

H1: Alternative hypothesis

Y: Dependent variable

X: Independent variables

AIC: Akaike Information Criterion

7 References

Data sourced from https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

8 Appendix1: Libraries used

dplyr

knitr

ggplot2

EnvStats

9 Appendix2: Roles of authors

One-sample tests, independence tests: Nwaokenneya Precious

Two-sample tests, normality checks: Raunaq Singh Dev

Kruskal-Wallis tests, simple linear regression and models: Harman Singh Saggu

Report: Harman Singh Saggu,Raunaq Singh Dev,Nwaokenneya Precious

10 Appendix3: Source code

```
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
library(dplyr)
library(ggplot2)
#Loading dataset
sleep_data <- read.csv("C:/Users/saggu/Desktop/project1000/Sleep_health_and_lifestyle_dataset.csv")

#Sample data
head(sleep_data)

#Measures of location
attach(sleep_data)
median(Age)
median(Sleep.Duration)
median(Quality.of.Sleep)
median(Stress.Level)
median(Heart.Rate)
median(Daily.Steps)

#Variability in data
var(Age)
var(Sleep.Duration)
var(Quality.of.Sleep)
var(Stress.Level)
var(Heart.Rate)
var(Daily.Steps)

#Visualizing measures of location and variability
par(mfrow = c(2,2))
boxplot(median(Sleep.Duration), median(Quality.of.Sleep),
```

```r
        median(Stress.Level),
        main = "Boxplot of Measures of Location(MEDIAN)",
        names = c("Sleep Duration", "Quality of Sleep",
                  "Stress Level"))
boxplot(var(Sleep.Duration), var(Quality.of.Sleep),
        var(Stress.Level),
        main = "Boxplot of Measures of Variability(VARIANCE)",
        names = c("Sleep Duration", "Quality of Sleep",
                  "Stress Level"))

#Visualizing sample data
#focusing on sleep duration, quality of sleep and stress level
par(mfrow = c(2,2)) #for plotting side by side
hist(sleep_data$Sleep.Duration, main = "Histogram of Sleep Duration",
     xlab = "Sleep Duration")
hist(sleep_data$Quality.of.Sleep, main = "Histogram of Quality of Sleep",
     xlab = "Quality of Sleep Level")
hist(sleep_data$Stress.Level, main = "Histogram of Stress Level",
     xlab = "Stress Level")





#Occupation list
select(sleep_data, Occupation) %>%
  count(Occupation)
#one sample mean

# 1. Checking the hypothesis that the mean sleep duration for total subjects is
# less than 7 (alpha = 0.05)

# H0: mean sleep duration is >= 7
```

```r
# H1: mean sleep duration is < 7

#using the t.test() function
t.test(sleep_data$Sleep.Duration, mu = 7, alternative = "less",
       conf.level = 0.95 )
# 2. Checking the hypothesis that the mean stress level for all the subjects is
# greater than 7 (alpha = 0.05)

# H0: mean stress level <= 7
# H1: mean stress level > 7

t.test(sleep_data$Stress.Level, mu = 7, alternative = "greater",
       conf.level = 0.95)
# 3. Checking the hypothesis that the mean quality of sleep for all the subjects
# is less than 7 (alpha = 0.05)

# H0: mean quality of sleep >= 7
# H1: mean quality of sleep < 7

t.test(sleep_data$Quality.of.Sleep, mu = 7, alternative = "less",
       conf.level = 0.95)
#one sample variance

# 3. Checking the hypothesis that the variance in age of subjects is not equal
# to 75 (alpha = 0.05)

# H0: variance in age == 75
# H1: variance in age != 75

#using varTest() function from EnvStats package
library(EnvStats)
```

```r
varTest(sleep_data$Age, alternative = "two.sided", sigma.squared = 75,
        conf.level = 0.95)
#two sample mean

# 4. Checking the hypothesis that the mean stress level of subjects with
# occupation Engineer is less than mean stress level of Scientist (alpha = 0.05)

# H0: mean stress level of Engineer >= Scientist
# H1: mean stress level of Engineer < Scientist

#using t.test() for two samples
t.test(sleep_data$Stress.Level[sleep_data$Occupation == "Engineer"],
       sleep_data$Stress.Level[sleep_data$Occupation == "Scientist"],
       mu = 0,alternative = "less", conf.level = 0.95, paired = FALSE,
       var.equal = FALSE)
#two sample mean

# Checking the hypothesis that the mean stress level of subjects with
# occupation Salesperson is less than mean stress level of Doctor
# (alpha = 0.05)



# H0: mean stress level of Salesperson >= Doctor
# H1: mean stress level of Salesperson < Doctor

#using t.test() for two samples
t.test(sleep_data$Stress.Level[sleep_data$Occupation == "Salesperson"],
       sleep_data$Stress.Level[sleep_data$Occupation == "Doctor"],
       mu = 0,alternative = "less", conf.level = 0.95, paired = FALSE,
       var.equal = FALSE)
# 5. Checking the hypothesis that mean quality of sleep of subjects with occupation
# Teacher is greater than occupation Engineer (alpha = 0.05)
```

```r
# H0: quality of sleep of Teacher <= Engineer
# H1: quality of sleep of Teacher > Engineer

t.test(sleep_data$Quality.of.Sleep[sleep_data$Occupation == "Teacher"],
       sleep_data$Quality.of.Sleep[sleep_data$Occupation == "Engineer"],
       mu = 0,alternative = "greater", conf.level = 0.95, paired = FALSE,
       var.equal = FALSE)
# 6. Checking the hypothesis that mean sleep duration of subjects with occupation
# Nurse is less than occupation Lawyer (alpha = 0.05)

# H0: Sleep duration of Nurse >= Lawyer
# H1: Sleep duration of Nurse < Lawyer

t.test(sleep_data$Sleep.Duration[sleep_data$Occupation == "Nurse"],
       sleep_data$Sleep.Duration[sleep_data$Occupation == "Lawyer"],
       mu = 0,alternative = "less", conf.level = 0.95, paired = FALSE,
       var.equal = FALSE)
#two sample proportion

# 7. Checking the hypothesis that proportion of subjects with stress level
# greater than 6 is more than subjects with stress level < 6

# H0: proportion of subjects with stress level < 6 is more
# H1: proportion of subjects with stress level > 6 is more

#using prop.test() function
x1 <- sum(as.numeric(sleep_data$Stress.Level < 6))
x2 <- sum(as.numeric(sleep_data$Stress.Level > 6))
n <- nrow(sleep_data)

prop.test(x = c(x1,x2), n = c(n,n), alternative = "greater",
          conf.level = 0.95, correct = FALSE)
```

```r
#two sample variance

# 8. Checking difference in variances in physical activity level between
# male and female subjects (alpha = 0.05)

# H0: physical activity level variance between male and female subjects ==
# H1: physical activity level variance between male and female subjects !=

#using var.test() function
var.test(sleep_data$Physical.Activity.Level[sleep_data$Gender == "Male"],
         sleep_data$Physical.Activity.Level[sleep_data$Gender == "Female"],
         ratio = 1, alternative = "two.sided", conf.level = 0.95)
#normality check

# 9. Using QQ plot to check for normality of sleep duration, quality of sleep
# and stress level

par(mfrow = c(1,2))

qqnorm(sleep_data$Sleep.Duration, main = "Sleep Duration QQ Plot")
qqline(sleep_data$Sleep.Duration, col = "seagreen", lwd = 2)

qqnorm(sleep_data$Quality.of.Sleep, main = "Quality of Sleep QQ Plot")
qqline(sleep_data$Quality.of.Sleep, col = "orange", lwd = 2)

qqnorm(sleep_data$Stress.Level, main = "Stress Level QQ Plot")
qqline(sleep_data$Stress.Level, col = "skyblue", lwd = 2)
#Sleep duration
shapiro.test(order(sleep_data$Sleep.Duration))

#Quality of sleep
shapiro.test(order(sleep_data$Quality.of.Sleep))
```

```r
#Stress level
shapiro.test(order(sleep_data$Stress.Level))
# 10. Using Kruskal-Wallis test to test if mean quality of sleep is same across
# both male and female (alpha 0.05)

# H0: mean quality of sleep is same across both male and female

kruskal.test(Quality.of.Sleep ~ Gender, data = sleep_data)
# Using Kruskal-Wallis test to test if mean sleep duration is same across
# both male and female (alpha 0.05)

# H0: mean sleep duration is same across both male and female

kruskal.test(Sleep.Duration ~ Gender, data = sleep_data)
# Using Kruskal-Wallis test to test if mean stress level is same across
# both male and female (alpha 0.05)

# H0: mean stress level is same across both male and female

kruskal.test(Stress.Level ~ Gender, data = sleep_data)
# 11. Using chi-squared test to check if quality of sleep is independent of
# sleep duration

# H0: quality of sleep is independent of sleep duration

chisq.test(sleep_data$Quality.of.Sleep, sleep_data$Sleep.Duration)
# Using chi-squared test to check if stress level is independent of quality of
# sleep

# H0: stress level is independent of quality of sleep

chisq.test(sleep_data$Stress.Level, sleep_data$Quality.of.Sleep)
```

```r
# Using chi-squared test to check if sleep disorder is independent of stress
# level

# H0: sleep disorder is independent of stress level

chisq.test(sleep_data$Sleep.Disorder, sleep_data$Stress.Level)
#Correlation matrix of variables
matrix1 <- select(sleep_data, Sleep.Duration, Quality.of.Sleep, Physical.Activity.Level, Stress.Level, 

cor(matrix1, method = "pearson")
#Creating a simple linear regression model

#The dependent variable is Quality of Sleep and the independent variables
#include Sleep Duration, Stress Level and Physical Activity Level

model <- lm(formula = Quality.of.Sleep ~ Sleep.Duration + Heart.Rate + Age +
            Physical.Activity.Level + Stress.Level, data = sleep_data)
summary(model)
#Including interaction terms Stress Level and Sleep Disorder
model_interaction <- lm(Quality.of.Sleep ~ Sleep.Duration + Age + Heart.Rate +
                         Stress.Level*Sleep.Disorder,data = sleep_data)
summary(model_interaction)
#Akaike information criterion (AIC) on first model
AIC(model)

#Akaike information criterion (BIC) on second model
AIC(model_interaction)
```

```r
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```