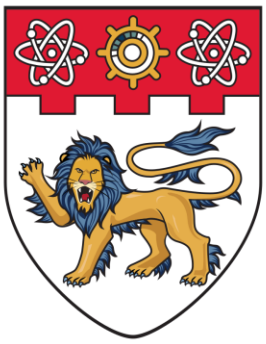


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511 Data Analysis with Computer

Group Project

Factors Affecting Daily Stress Level

Name	Matriculation Number
Guo Xiaozhen	U1940202L
Ker Xing Thong	U1940200G
Tan Li Lin	U1940820G
Jeanette Poh Wen Jun	U1940395G

Abstract:

According to recent news reports, the Covid-19 pandemic has resulted in many people experiencing increased stress due to various factors. This is a worrying trend as high levels of stress are detrimental to one's physical and mental well-being. Hence, we seek to understand how stress levels are affected by various factors, using data collected from a Global Work-Life Balance Survey. We will examine the relationship between daily stress levels and various parameters, using a range of data analysis techniques.

Table of Contents

1. Introduction

Due to the disruptions caused by Covid-19, many people have been facing increased levels of stress (Teo, 2020). For instance, a survey conducted on 1500 Americans found that 55% of them were more stressed due to the pandemic (Farr, 2020). While Covid-19 has heightened stress levels, stress often originates from a variety of factors. We aim to investigate how various parameters affect stress so that we can adjust our lifestyles accordingly to avoid high-stress levels.

To achieve our objectives, we will analyse a dataset consisting of 12,756 survey responses on daily stress levels, which consists of responses from 2015 to 2019. This dataset contains 23 attributes describing the lifestyle and behaviour of the respondents. Based on our chosen data, we aim to answer the following key questions:

1. Does age group affect daily stress level?
2. Does sleep duration affect daily stress level?
3. Does daily stress level depend on the level of support from one's close circle?
4. Does one's number of achievements influence his or her daily stress level?
5. Does the daily stress level depend on one's income sufficiency?

For each of our research questions above, we used the R language to conduct data analysis, including hypothesis tests and plotting appropriate diagrams. Thereafter, we derived appropriate conclusions which are supported by statistical explanations.

2. Data Description

The "Lifestyle_and_Wellbeing_Data" dataset was attained from Kaggle, an online data science community. It consists of a csv data frame "Wellbeing_and_lifestyle_data.csv". From this dataset, we extracted data from the "Daily_Stress" column and identified the dependency with our chosen five factors – Core_Circle, Achievement, Sleep_Hour, Sufficient_Income and Age.

As all survey questions were mandatory, respondents had to answer all questions to successfully submit their responses. Hence, there are no missing values in this dataset. However, for the Core_Circle factor, category 10 is representative of 10 or more people which gives us a range of discrete values, thus differing from the other categories which are single discrete values. Hence, we removed all rows with a value of 10 for Core_Circle to ensure a fair comparison. Likewise, we applied this filtering of data for the Achievement factor as we encountered the same scenario.

We then performed preliminary data cleaning to remove the 17 irrelevant columns, such as "Timestamp" and "Fruits_Veggies".

After conducting data cleaning, 9964 observations with 6 factors were retained for analysis:

Daily_Stress	The rating of the amount of stress the respondents experience every day.
Core_Circle	The number of family and friends that the respondents are very close to.
Achievement	The number of personal achievements over the last 12 months that the respondents is proud of.
Sleep_Hour	The respondents' sleep duration per night in hours over the past 7 days.
Sufficient_Income	An indication of whether the respondents' income is enough to cover his or her basic living expenses.
Age	The age group that the respondents belong to.

For each variable, the ratings are represented as follow:

Daily_Stress – Respondents were required to indicate, on a scale of 0 (least stress) to 5 (most stress), their average stress level over the last 12 months. Their stress could be due to various factors such as the working environment and events like divorce or illness.

Core_Circle – Respondents were required to indicate the number of close family/friends they have (0 to 9). Cases of 10 or more close family/friends were eliminated in the data analysis for Core_Circle, as mentioned above.

Achievement – Respondents were required to submit the number of personal achievements they attained over the past year. Examples of these achievements ranged from smaller tasks such as completing a marathon to major events like giving birth, opening a new business, or buying a new house. The choices given were 0, 1, 2, 3…, 9.

Sleep_Hour – Respondents were asked to indicate the number of hour(s) they typically sleep, with choices ranging from 1 hour to 10 hours per night.

Sufficient_Income – Respondents were asked how sufficient their income was to cover basic living expenses such as the cost of housing, food, health care and education. There were 2 choices where 1 represents not or hardly sufficient and 2 represents sufficient.

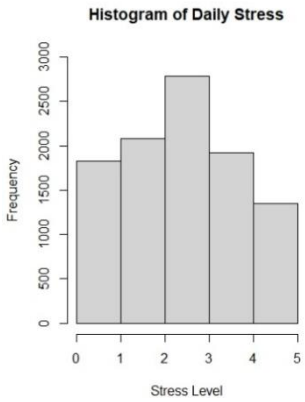
Age – Respondents were required to select their age range from four categories, namely: Less than 20, 21 to 35, 36 to 50 and 51 or more.

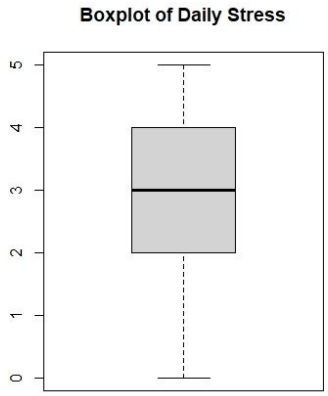
3. Description and Cleaning of Dataset

To better understand each variable, we used different data visualization tools to identify possible outliers and performed data transformation where necessary, to avoid highly skewed data. The individual statistic summaries for the respective variables are shown below.

3.1 Summary statistics for the main variable of interest, *Daily_Stress*

To understand the distribution of the variable *Daily_Stress,* we plot a histogram and a boxplot.

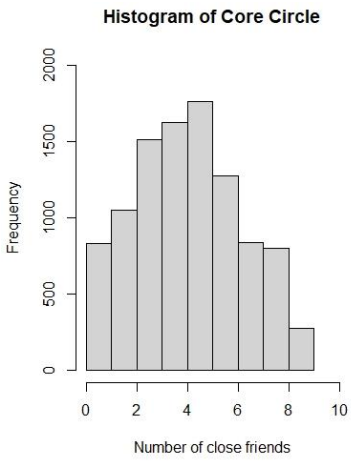


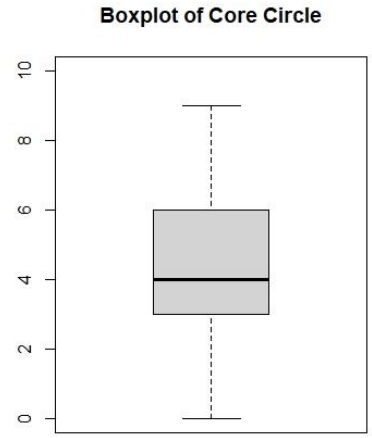


The contour of the histogram approximately gives a bell-shaped curve, which resembles a normal distribution. Hence, no data transformation was performed. No outliers were found in both the histogram and boxplot. The median stress level is 3, suggesting that most respondents experience moderate stress daily.

3.2 Summary statistics for other variables

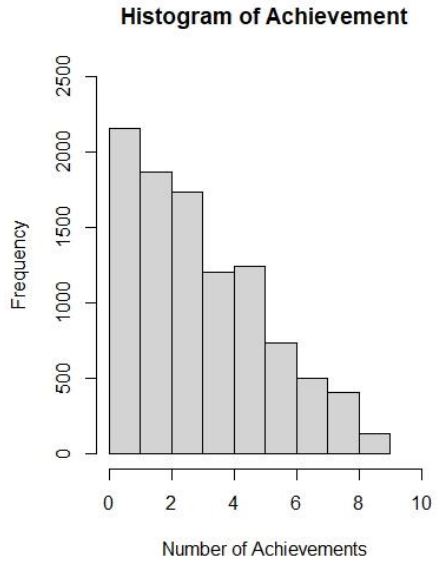
3.2.1 Core_Circle

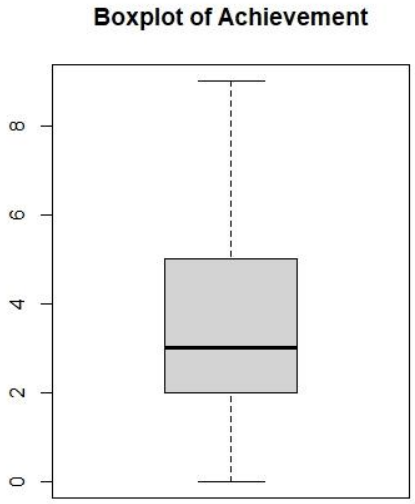


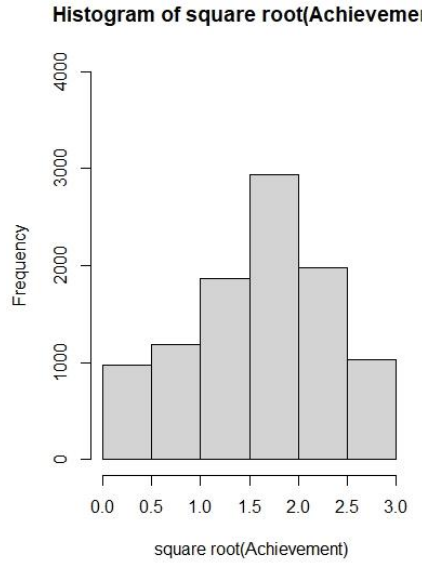


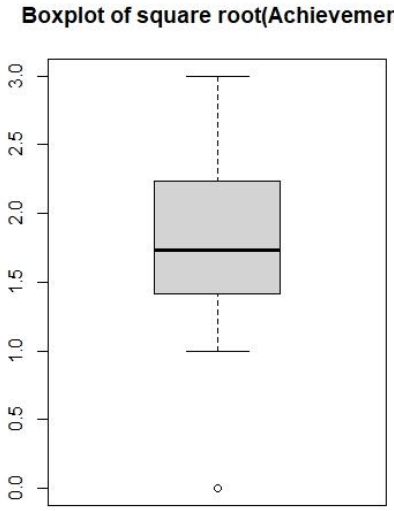
The contour of this histogram is relatively symmetrical about the mean of 4.51, suggesting that it is close to a normal distribution. No outliers were found in both the histogram and boxplot. Hence, we did not perform any data transformation.

3.2.2 *Achievement*





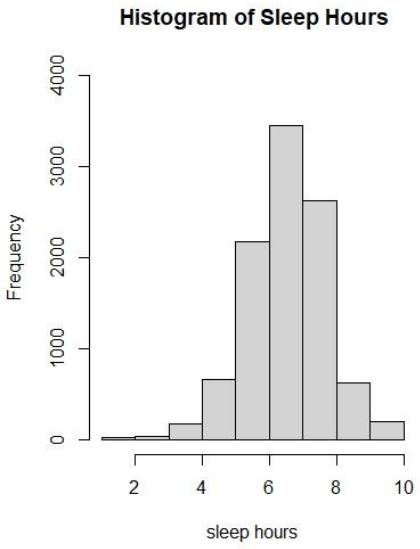


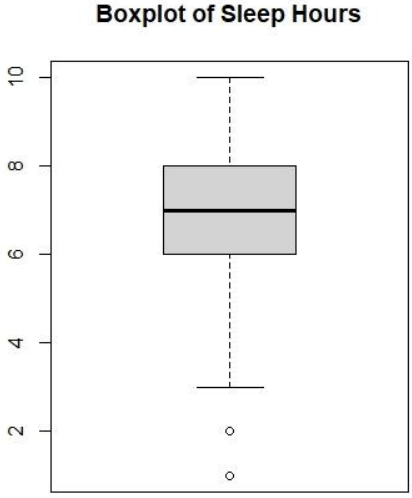


In our initial boxplot for achievement, the data appears to skew to the right. Performing log transformation would result in invalid values as there are "0"s in our data. Hence, we performed square root transformation instead to reduce data skewness.

One outlier, that is of the value 0, was found after square root transformation. However, this outlier was not removed as it is still a valid response. There could be individuals who did not manage to hit their goals, or who didn't feel proud of what they had achieved, and such factors might be related to the daily stress level. Hence, no data was removed as all data points can potentially offer valuable insights during statistical analysis.

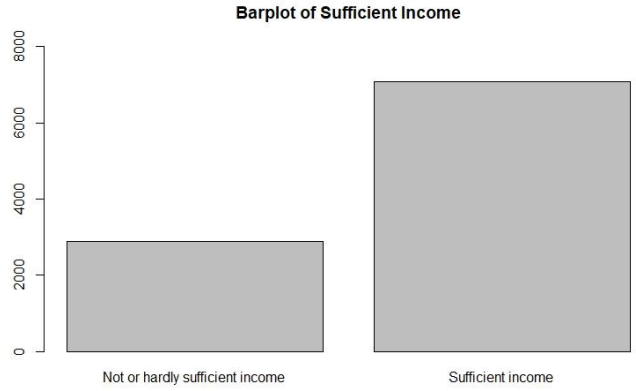
3.2.3 Sleep_Hours





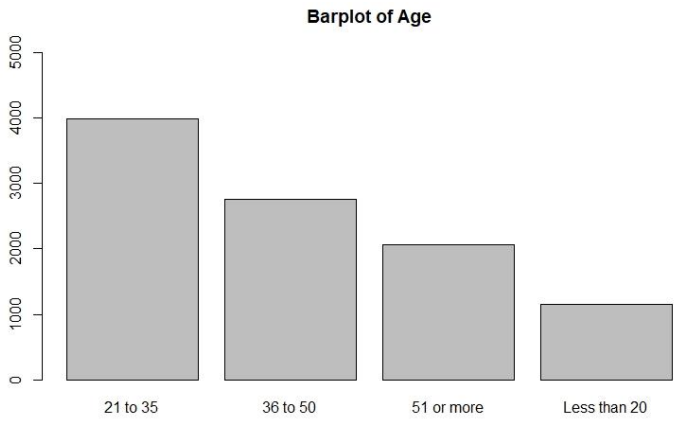
Although the data is not perfectly symmetrical about the mean, it is not largely skewed to either side. From our boxplot, we observe that the two outliers correspond to the respective categories of 1 sleep hour and 2 sleep hours. These outliers could be explained by certain circumstances such as work, resulting in some days where one has less sleep than usual. Insomnia caused by stress could also lead to abnormal sleep duration. Hence, we did not remove the outliers as our goal is to find out if stress level depends on sleep hours. Since the outliers are still valid inputs, removing them might affect the credibility of our data analysis.

3.2.4 Sufficient_Income



2877 respondents stated that their income was insufficient while the remaining 7087 respondents had sufficient income to cover their basic needs. Hence, a majority of respondents had sufficient income for their basic living expenses.

3.2.5 Age



Most respondents were between 21 to 35 years of age. Compared to the other age groups, there were fewer respondents below 20 years of age.

3.3 Final Dataset for Analysis

The final dataset consists of 9964 observations, with square root transformation being applied to Achievement.

4. Statistical Analysis

For all statistical tests, we choose the significance level (α) to be 0.05.

4.1 Relation between *Age* and *Daily_Stress*

In this section, we will answer the question "Does age group affect daily stress level?".

Chi-Square test

We conduct chi-square test to determine whether there is an association between age group and daily stress level.

H_0: There is no association between age group and daily stress level
H_1: Some association does exist between age group and daily stress level

```
> agetable <- table(data$AGE, data$DAILY_STRESS)
> agetable

                0    1    2    3    4    5
  21 to 35    126  556  857 1167  766  512
  36 to 50     73  347  587  785  566  403
  51 or more  115  359  420  536  367  263
  Less than 20  62  189  216  301  224  167
> chisq.test(agetable)

        Pearson's Chi-squared test

data:  agetable
X-squared = 82.136, df = 15, p-value = 2.836e-11
```

Figure 1 Output of the Chi-Square test results in R console.

Since p-value $= 2.836 \times 10^{-11} < \alpha$, we reject H_0 in favour of H_1, and conclude that some association exists between age group and daily stress level.

Next, we use the ANOVA test to determine the equality of the mean daily stress across the various age groups, since there are more than two age groups, and the age groups are also independent.

ANOVA test

Let μ_i be the mean of daily stress where $i = 1, 2, 3, 4$ represent the age groups less than 20 years old, 21 to 35 years old, 36 to 50 years old, and 51 years old and above respectively.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: Not all μ_i are equal across different age groups.

```
> aov(data$DAILY_STRESS~factor(data$AGE))
Call:
   aov(formula = data$DAILY_STRESS ~ factor(data$AGE))

Terms:
                factor(data$AGE) Residuals
Sum of Squares           70.401 18275.689
Deg. of Freedom              3      9960

Residual standard error: 1.354588
Estimated effects may be unbalanced
1 observation deleted due to missingness
> summary(aov(data$DAILY_STRESS~factor(data$AGE)))
                  Df Sum Sq Mean Sq F value   Pr(>F)
factor(data$AGE)    3     70  23.467   12.79 2.45e-08 ***
Residuals        9960  18276   1.835
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

Figure 2 Output of the ANOVA test results in R console.

Since p-value = $2.45 \times 10^{-8} < \alpha$, we reject H_0 and conclude that not all means of daily stress levels are equal across the age groups.

As H_0 is rejected, we perform a set of pairwise comparisons to determine which age group has different mean daily stress from the other age groups. If p-value < 0.05 when comparing 2 age groups, then there is enough evidence to infer that the mean of daily stress differs for these age groups. Hence, we conclude that the following age groups have different mean daily stress from each other:

a. 21 to 35 years old group and 36 to 50 years old group
b. 21 to 35 years old group and 51 or more years old group
c. 36 to 50 years old group and 51 or more years old group
d. 36 to 50 years old group and Less than 20 years old group

```
> pairwise.t.test(data$DAILY_STRESS, data$AGE, p.adjust.method = "none")

        Pairwise comparisons using t tests with pooled SD

data:  data$DAILY_STRESS and data$AGE

             21 to 35 36 to 50 51 or more
36 to 50     0.0053   -        -
51 or more   6.7e-05  1.2e-09  -
Less than 20 0.2525   0.0022   0.0565

P value adjustment method: none
```

Figure 3 Output of the pairwise t-test results in R console

For the comparisons of the following age groups, there is insufficient evidence to support the hypothesis that they have different mean daily stress from each other.

a. 21 to 35 years old and Less than 20 years old group
b. Less than 20 years old group and 51 or more years old group

From both the chi-square test and ANOVA test, we can infer that age group affects daily stress level.

4.2 Relation between *Sleep_Hours* and *Daily_Stress*

With a fast tempo of living, many people struggle with insufficient sleep. Hence, we would like to investigate the relation between sleep hours and daily stress.

First, we conduct a chi-square test to find out if there is an association between sleep hours and daily stress. We set up a table (sleeptable) and introduce the chisq.test function.

Chi-Square Test

H_0: There is no association between sleep hours and daily stress level
H_1: Some association does exist between sleep hours and daily stress level

```
> sleeptable
            DAILY_STRESS
SLEEP_HOURS    0    1    2    3    4    5
          1    4    0    1    3    2    6
          2    0    1    2    6    4    3
          3    3    2    9    6    7   13
          4    8   23   26   36   52   58
          5   19  104  123  225  163  196
          6   80  316  501  768  593  474
          7  152  687 1016 1322  818  456
          8  196  678  811  951  502  275
          9   61  152  168  169  143   96
          10  39   34   41   49   36   66
> chisq.test(sleeptable)

        Pearson's Chi-squared test

data:  sleeptable
X-squared = 672.21, df = 45, p-value < 2.2e-16
```

Figure 4 Output of Chi-Square test results in R console.

At a significance level of α = 0.05, we rejected H_0 as p-value = $2.2 \times 10^{-16} < \alpha$. Hence, we conclude that there is an association between sleep hours and daily stress level.

4.3 Relation between *Core_Circle* and *Daily_Stress*

In this section, we investigate if daily stress level depends on the number of people (family members/friends) that the respondents is close to.

We divide respondents into groups based on the number of close family members/friends they have and create a boxplot to show the distribution of every group.

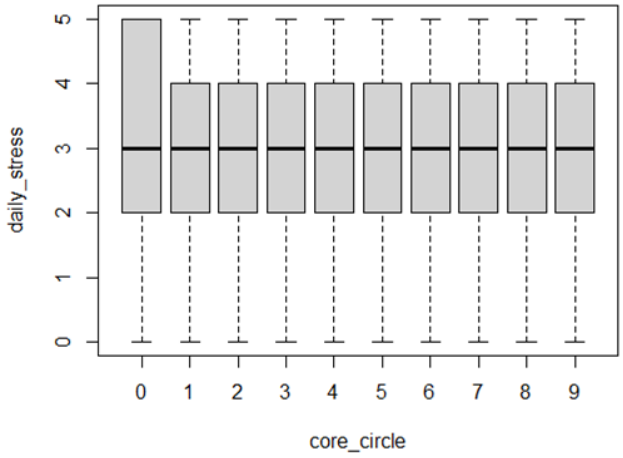


Figure 5 Boxplot of Daily_Stress by Core_Circle

The boxplot indicates that the median daily stress level is equal across all groups. The distributions of the groups are similar, except for the group which corresponds to 0 close family/friends.

Next, we perform an ANOVA test to test if the mean daily stress level is equal across all groups.

Let μ_i be the mean of daily stress level with $0 \leq i \leq 9$ across the different size of core circle.

H_0: $\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9$.

H_1: Not all μ_i are equal, where $0 \leq i \leq 9$.

```
> aov(data$DAILY_STRESS~factor(data$CORE_CIRCLE))
Call:
   aov(formula = data$DAILY_STRESS ~ factor(data$CORE_CIRCLE))

Terms:
                factor(data$CORE_CIRCLE) Residuals
Sum of Squares                  18346.09      0.00
Deg. of Freedom                        1      9962

Residual standard error: 2.41634e-13
Estimated effects may be unbalanced
> summary(aov(data$DAILY_STRESS~factor(data$CORE_CIRCLE)))
                           Df Sum Sq Mean Sq   F value Pr(>F)
factor(data$CORE_CIRCLE)    1  18346   18346 3.142e+29 <2e-16 ***
Residuals                9962      0       0
```

Figure 6 Output of the ANOVA test results in R console.

We reject the null hypothesis since the p-value ($< 2 \times 10^{-16}$) is significant and conclude that the means of daily stress across different numbers of close family/friends are unequal. Hence, we infer that daily stress level is dependent on the number of close family members/friends that the respondents have.

Next, we perform chi-square test to determine if there is any association between Core_Circle and Daily_Stress.

Chi-Square Test

H_0: There is no association between the variables Core_Circle and Daily_Stress.

H_1: The two variables, Core_Circle and Daily_Stress have some association.

```
> corecircletable=table(data$DAILY_STRESS,data$CORE_CIRCLE)
> chisq.test(corecircletable)

        Pearson's Chi-squared test

data:  corecircletable
X-squared = 256.46, df = 45, p-value < 2.2e-16
```

Figure 7 Output of the Chi-Square test results in R console.

The p-value obtained is significant as it is smaller than 2.2×10^{-16}. Hence, we reject the null hypothesis and conclude that there is some association between the number of family members or close friends that one has and his/her daily stress level.

With the results of the ANOVA test and chi-square test, we can deduce that one's core circle affects his/her daily stress level.

4.4 Relation between *Daily_Stress* and *Square root (Achievement)*

We attempt to answer the question "Does the number of achievements one is proud of influence his/her daily stress level?" in this section.

From the boxplot below, we see that the distribution of square root (Achievement) is quite similar for all 10 groups of the different numbers of achievements.

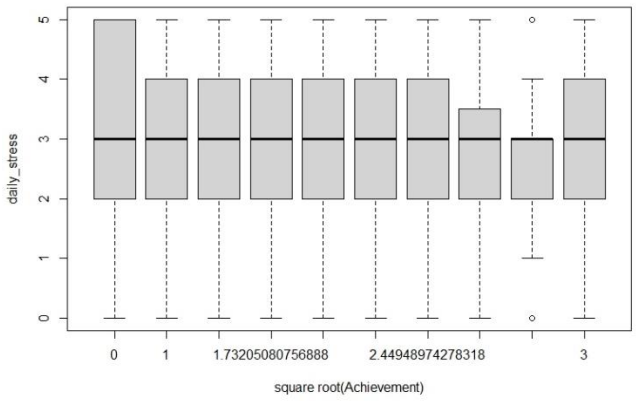


Figure 8 Boxplot of Daily_Stress by the square root of Achievement

Since there are more than two populations to be compared, and we are interested in knowing whether there is an overall effect of the factor (number of achievements that one is proud of) on the outcome (daily stress level), the ANOVA test is applicable.

Let μ_i be the mean of daily stress for the groups with $0 \leq i \leq 9$ achievements. As that the observations are independent of each other, we conducted an ANOVA test to test the following hypotheses:

H_0: $\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9$.

H_1: Not all μ_i are equal, where $0 \leq i \leq 9$.

```
> ANOVA_srt = aov(daily_stress~factor(srt))
> ANOVA_srt
Call:
   aov(formula = daily_stress ~ factor(srt))

Terms:
                factor(srt) Residuals
Sum of Squares      297.325 18048.765
Deg. of Freedom           9      9954

Residual standard error: 1.346558
Estimated effects may be unbalanced
> summary(ANOVA_srt)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(srt)    9    297   33.04   18.22 <2e-16 ***
Residuals   9954  18049    1.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9 Output of the Chi-Square test results in R console.

The ANOVA test returns a p-value of $< 2 \times 10^{-16}$, which shows that the means are significantly different. Hence, we reject H_0, and concluded that the daily stress level depends on the number of achievements one is proud of.

4.5 Relation between *Daily_Stress* and *Sufficient_Income*

In this section, we are interested in knowing whether daily stress level depends on one's income status (sufficient income, insufficient income).

The boxplot is shown below with 1 indicating insufficient income and 2 indicating sufficient income.

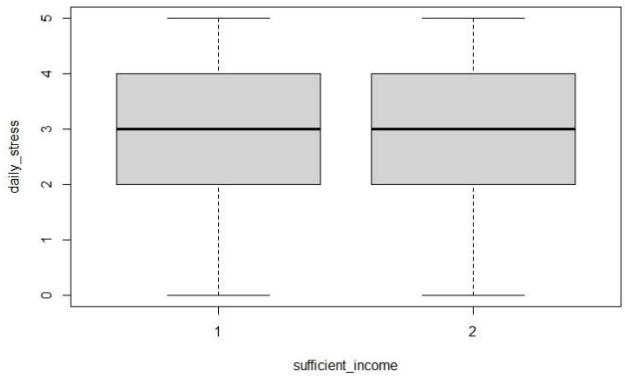


Figure 10 Boxplot of daily stress and sufficient income

It seems like both groups have the same median. We perform the chi-square test to test the following hypotheses:

H_0: The daily stress level is independent of the income sufficiency group.

H_1: The daily stress level depends on income sufficiency.

```
> incometab = table(data$DAILY_STRESS,data$SUFFICIENT_INCOME)
> incometab

     1    2
0   85  291
1  289 1162
2  490 1590
3  799 1990
4  641 1282
5  573  772
> chisq.test(incometab)

        Pearson's Chi-squared test

data:  incometab
X-squared = 234.62, df = 5, p-value < 2.2e-16
```

Figure 11 Output of the Chi-Square test results in R console.

From the above figure, the p-value is $< 2.2 \times 10^{-16}$ which is less than α. Thus, we reject H_0 and conclude that there is insufficient evidence to show that daily stress level is dependent on income sufficiency.

5. Conclusion and Discussion

Daily stress has always been a hot issue. Especially amidst the Covid-19 pandemic, Singaporeans are experiencing increased stress levels according to surveys conducted by Cigna 360 Well-Being in 2019 (Lifeskills Institute, 2020). Based on the dataset, our analysis shows that all five factors have an impact on daily stress level.

We conclude that:

1. Age affects daily stress level.
2. Sleep duration affects daily stress level.
3. Daily stress level depends on the level of support from one's close circle.
4. The number of achievements does affect the daily stress level.
5. Daily stress level is dependent on one's income sufficiency.

Some suggestions to relieve stress include exercising regularly, have sufficient sleep and spending time with your family members/friends (Health Hub, 2020).

One limitation of our report is that our dataset does not include the countries where the respondents stay. This would lead to a lack of socioeconomic information such as the cost of living of each country as the background and culture of a country could affect the amount of stress one faced. For instance, if the majority of respondents were living in a country with high living expenses, this could lead to higher stress when they had insufficient income. Thus, our result could not represent the population with lower living expenses. Without knowing where

the respondents were from and the respective proportion of people, we cannot conclude that our result is representative of the general population.

Note that the original dataset is updated with survey responses up to 14 March 2021. For further analysis, we would suggest grouping the responses by year and analyse the data. By doing so, we would obtain the trend of daily stress level throughout the years. This could be another point of view for the analysis of this dataset.

6. Appendix

The following are the codes we ran and the output in R.

```
mydata = read.csv("C:/Users/nehzo/OneDrive/Desktop/lifestyle n
stress/Wellbeing_and_lifestyle_data.csv", header = T)
str(mydata)
par(mfrow = c(1,2))
#data cleaning
data = mydata[,c(3,5,8,15,18,22)]
data = subset(data,CORE_CIRCLE < 10)
data = subset(data, ACHIEVEMENT < 10)

#converting the relevant columns into numeric class
i = c(1:5)
data[,i] = sapply(data[ , i],as.numeric)
str(data)
summary(data)

#remove the NA values
data = na.omit(data)
str(data)
summary(data)

#subsetting into individual variables
daily_stress = data[,"DAILY_STRESS"]
core_circle= data[,"CORE_CIRCLE"]
achievement = data[,"ACHIEVEMENT"]
sleep_hours = data[,"SLEEP_HOURS"]
sufficient_income = data[,"SUFFICIENT_INCOME"]
age = data[,"AGE"]

#data visualisation
hist(daily_stress,ylim = c(0, 3000), breaks = 5, main = "Histogram of Daily Stress", xlab =
"Stress Level")
boxplot(daily_stress, main="Boxplot of Daily Stress")
summary(daily_stress)

hist(core_circle, ylim = c(0,2000),xlim = c(0,10),xlab = "Number of close friends",main =
"Histogram of Core Circle",breaks = 10)
boxplot(core_circle, main ="Boxplot of Core Circle", ylim= c(0,10))
summary(core_circle)

hist(achievement, ylim = c(0,2500),xlim = c(0,10),xlab = "Number of Achievements",breaks
= 10, main = "Histogram of Achievement")
```

```
boxplot(achievement, main ="Boxplot of Achievement")

srt = sqrt(achievement)
hist(srt, breaks = 5, main = "Histogram of square root(Achievement)", xlab = " square root(Achievement)", ylim = c(0,4000))
boxplot(srt, main = "Boxplot of square root(Achievement)")
summary(srt)

hist(sleep_hours, breaks = 10, ylim = c(0,4000), main = " Histogram of Sleep Hours", xlab = "sleep hours")
boxplot(sleep_hours, main = "Boxplot of Sleep Hours")
summary(sleep_hours)

par(mfrow = c(1,1))

counts = table(sufficient_income)
barplot(counts,main = "Barplot of Sufficient Income",ylim = c(0,8000), names.arg=c("Not or hardly sufficient income", "Sufficient income"))

count2 = table(age)
barplot(count2,main = "Barplot of Age",ylim = c(0,5000),names.arg=c("21 to 35", "36 to 50", "51 or more", "Less than 20"))

#statistical test

#stress vs core circle
boxplot(data$DAILY_STRESS~data$CORE_CIRCLE,col="light gray"
      ,ylab="daily_stress",xlab="core_circle")
aov(data$DAILY_STRESS~factor(data$CORE_CIRCLE))
summary(aov(data$DAILY_STRESS~factor(data$CORE_CIRCLE)))
corecircletable=table(data$DAILY_STRESS,data$CORE_CIRCLE)
chisq.test(corecircletable)

#stress vs sleep hours
attach(data)
sleeptable = table(SLEEP_HOURS,DAILY_STRESS)
sleeptable
chisq.test(sleeptable)
detach(data)

#square root achievement vs stress
table(achievement)
summary(achievement)
```

```
boxplot(daily_stress~srt, xlab = "square root(Achievement)")
ANOVA_srt = aov(daily_stress~factor(srt))
ANOVA_srt
summary(ANOVA_srt)

#stress vs income
boxplot(data$DAILY_STRESS~data$SUFFICIENT_INCOME, xlab = "Sufficient Income",
ylab = "Daily Stress")
table(sufficient_income)

data1 = sample(sufficient_income,5000)
shapiro.test(data1)
# it is not normal, cannot use t.test

incometab = table(data$DAILY_STRESS,data$SUFFICIENT_INCOME)
incometab
chisq.test(incometab)
```

7. References

Farr, C. (1 July, 2020). *Tech*. Retrieved from CNBC: https://www.cnbc.com/2020/07/01/covid-19-pandemic-has-increased-stress-say-55percent-in-harvard-unc-study.html

Health Hub. (20 Apr, 2020). *10 Easy Ways to Manage Stress*. Retrieved from Health Hub: https://www.healthhub.sg/live-healthy/426/10easywaystomanagestress

Lifeskills Institute. (7 Nov, 2020). *Are You Part Of The 92%?* Retrieved from Lifeskills Institute: https://lifeskillsinstitute.sg/are-you-part-of-the-92-singapore-stress-report/#:~:text=According%20to%20a%20Cigna%20360,predominately%20culminates%20in%20the%20workplace.

Teo, J. (10 Aug, 2020). *Life*. Retrieved from The Straits Times: https://www.straitstimes.com/lifestyle/covid-19-can-give-rise-to-mental-health-stresses-say-experts