VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
THE INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

STATISTICAL METHOD

REPORT PROJECT

Topic: Sleep Health and Lifestyle

BY GROUP 14 – MEMBER LIST

Name	ID
Nguyễn Tất Bách	ITDSIU22157
Đào Ngọc Lan Hồng	ITDSIU21088
Nguyễn Thị Ngọc Mai	ITDSIU21098
Lê Hữu An Khang	ITDSIU22134

Instructor: Pham Hoang Uyen

Table Contents

Chapter 1. Introduction 3

1. **Abstract** 4
2. **Object** 4
3. **Goal** 4

Chapter 2. Data Analysis 5

1. **Summarizing Data** 5
2. **Relation between two variable** 12
3. **Check normality for 2 quantitative variables** 18
4. **Construct Confidence Interval** 20
5. **Perform Hypothesis Testing** 20

Chapter 3. Conclusion 22

References 23

List of figures

Figure 1.1 Code for frequency bar graph 5
Figure 1.2 The bar graph for frequency Gender 6
Figure 1.3 Code for relative frequency bar graph 6
Figure 1.4 The bar graph for relative frequency Gender 7
Figure 1.5 Code for pie chart 7
Figure 1.6 The pie chart for Gender 7
Figure 1.7 Code for frequency of Stress Level 9
Figure 1.8 The bar graph frequency of Stress Level 9
Figure 1.9 Code for frequency of Quality of sleep 9
Figure 1.10 The bar graph frequency of Quality of sleep 10
Figure 1.11 Five-number summary of Stress Level 10
Figure 1.12 Five-number summary of Quality of sleep 11
Figure 1.13 Boxplot of Stress Level 11
Figure 1.14 Boxplot of Quality of sleep 12
Figure 2.1 The scatter diagram. 13
Figure 2.2 OLS Regression Results 14
Figure 2.3 The line on the scatter diagram 15
Figure 2.4 Scatter plot of residual 17
Figure 2.5 Boxplot of residual 17
Figure 2.6 Five-number summary of residual 18
Figure 3.1 Histogram for Stress Level variables 18
Figure 3.2 Histogram and KDE for Stress Level variables (range 3 – 7 level) 19
Figure 4.1 Confidence Interval algorithm for a population mean 20
Figure 4.2 Output value for CI 20

Chapter 1. Introduction

1. Abstract

Sleep plays an important role in health, directly affecting many issues such as obesity, heart disease, and mental health. The close relationship between stress levels and sleep quality is also a field that needs thorough research, as stress can cause sleep disorders and vice versa. Additionally, gender differences in these factors are noteworthy, as women often face more difficulties in maintaining quality sleep due to biological and social factors, while men often confront stress from work pressure and social responsibilities, leading to frequent sleep deprivation and deteriorating health. Analyzing this data not only provides valuable information for healthcare, education, and human resource management but also raises community awareness about the importance of sleep and stress management. Furthermore, this research contributes to the scientific foundation, opening up new research directions and promoting advanced findings, thereby building a healthier and happier society.

2. Object

- We have used many measures and performed many tests to determine the best and most suitable variables for describing people's sleep health. We have chosen these variables to analyze and execute hypothesis tests: Gender, Stress Levels, and Quality of Sleep.
- Genders are nominal variables and Stress Levels, Quality of Sleep are ordinal variables.
  - Explanatory variable: Stress Levels
  - Response Variable: Quality of Sleep

3. Goal

- Analyzing data to find patterns, trends, and the correlation between Sleep Quality and variables.
- Implementing knowledge and skills learned from the Statistical Methods course to collect and understand thoroughly the data.
- Understanding more about how datasets really work in reality.

Chapter 2. Data Analysis

1. Summarizing Data

1.1. One categorical

- Frequency table

The frequency table below summarizes the categorical data for gender.

Gender	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
Female	46	0.306667	46	0.306667
Male	104	0.693333	150	1.000000

- **Frequency**: This represents the count of occurrences for each gender.
- **Relative Frequency**: The proportion of each gender relative to the total number of observations.
- **Cumulative Frequency**: The cumulative count of observations up to and including the current category.
- **Cumulative Relative Frequency**: The cumulative proportion of observations up to and including the current category.

- Mode

The mode is the category with the highest frequency. In this dataset, the mode is "Male" with a frequency of 104.

- Graphs

To visualize the categorical data for gender, we use both a bar chart and a pie chart:

- **Bar Chart**: The bar chart displays the frequency of each gender.
- **Pie Chart**: The pie chart illustrates the proportion of each gender relative to the whole dataset.

Python code used for visualization:

```
# creating the Frequency Bar graph
df = frequency_df
print("Frequency Bar graph")
# plt.figure(figsize=(10, 5))
plt.bar(df.Gender, df.Frequency, width = 0.2)

plt.xlabel("Gender")
plt.ylabel("Frequency")
plt.title("Bar Graph for Frequency of Gender")
plt.show()
```

Figure 1.1 Code for frequency bar graph

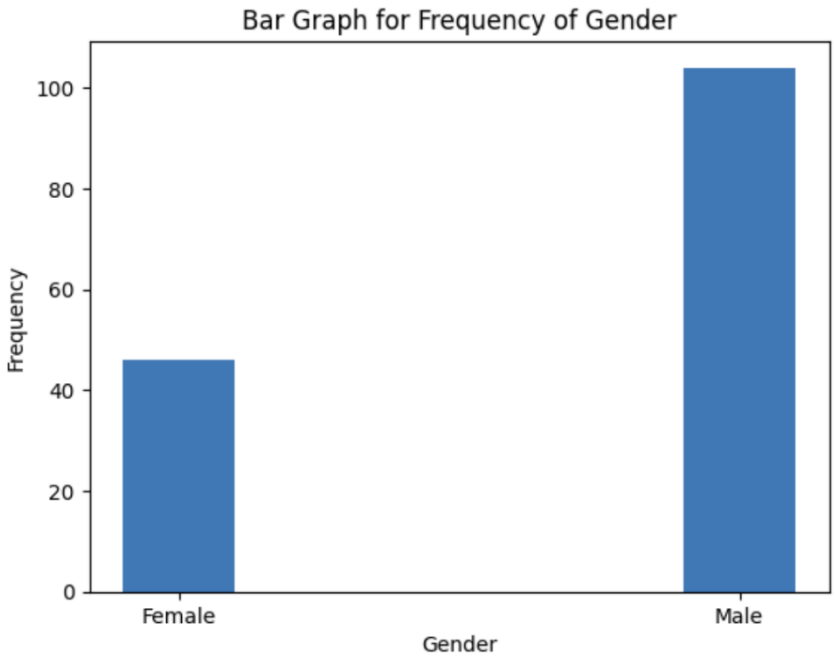

Figure 1.2 The bar graph for frequency Gender

```
# Cumulative Relative Frequency
df = frequency_df
print("Relative Frequency Bar graph")
# plt.figure(figsize=(10, 5))
plt.bar(df['Gender'], df['Cumulative Relative Frequency'], width = 0.2)

plt.xlabel("Gender")
plt.ylabel("Cumulative Relative Frequency")
plt.title("Bar Graph for Cumulative Relative Frequency of Gender")
plt.show()
```

Figure 1.3 Code for relative frequency bar graph

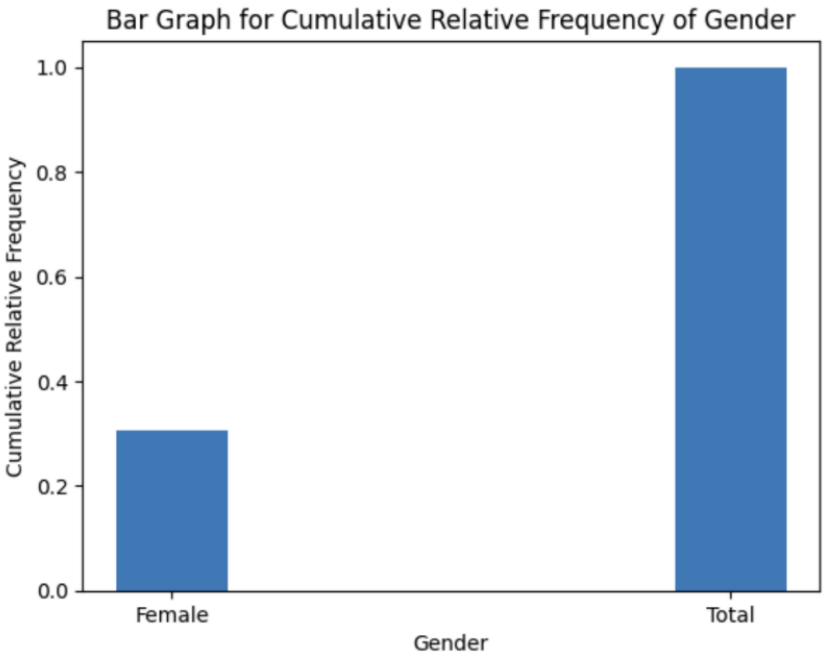

Figure 1.4 The bar graph for relative frequency Gender

```
#Pie Chart
fig, ax = plt.subplots()
ax.pie(df["Frequency"],
       labels = df["Gender"],
       autopct='%1.1f%%')
plt.show()
```

Figure 1.5 Code for pie chart

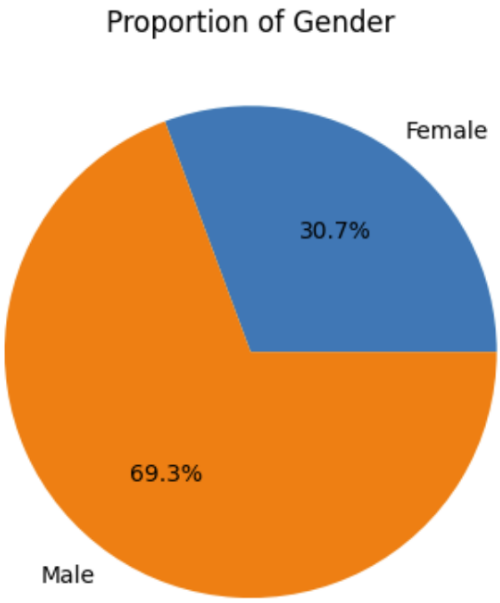

Figure 1.6 The pie chart for Gender

1.2. Two quantitative

- Frequency table

The frequency tables for StressLevel and SleepQuality are as follows:

StressLevel:

Stress Level	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
3	3	0.020000	3	0.020000
4	42	0.280000	45	0.300000
5	21	0.140000	66	0.440000
6	38	0.253333	104	0.693333
7	8	0.053333	112	0.746667
8	38	0.253333	150	1.000000

SleepQuality:

Quality of Sleep	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
4	5	0.033333	5	0.033333
5	7	0.046667	12	0.080000
6	38	0.253333	50	0.333333
7	37	0.246667	87	0.580000
8	62	0.413333	149	0.993333
9	1	0.006667	150	1.000000

- Graphs

To visualize the categorical data for gender, we use both a bar chart and a pie chart:

- **Bar Chart**: The bar chart displays the frequency of each level from stress level and quality of sleep.
- **Box Plot**: The box plot summarizes the distribution of a dataset, highlighting the median, quartiles, and potential outliers for each data of stress level and quality of sleep.

Python code used for visualization:

```
print("Frequency Bar graph")
plt.bar(df1.Stress_Level, df1.Frequency, width = 0.4)

plt.xlabel("Stress Level")
plt.ylabel("Frequency")
plt.title("Stress Level vs Frequency")
plt.show()
```

Figure 1.7 Code for frequency of Stress Level

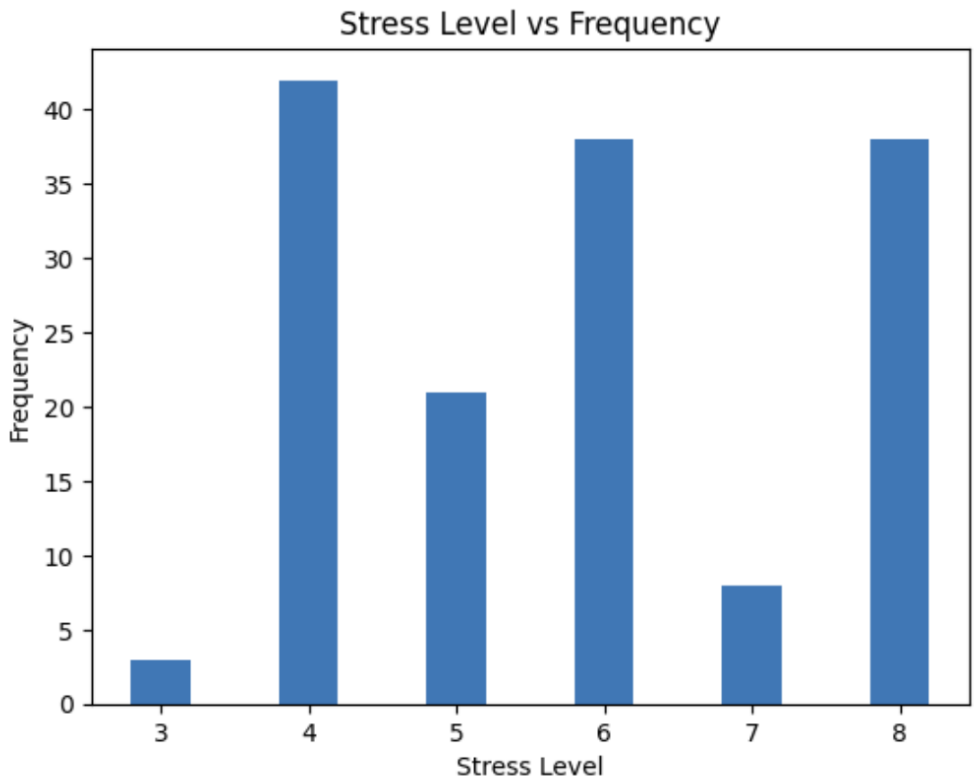

Figure 1.8 The bar graph frequency of Stress Level

Based on this chart, we can see that even though the most common stress level is 4 (which is relatively low), there are still many people who have high stress level in this sample.

```
print("Frequency Bar graph")
plt.bar(df2.Quality_of_Sleep, df2.Frequency, width = 0.4)

plt.xlabel("Quality of Sleep")
plt.ylabel("Frequency")
plt.title("Quality of Sleep vs Frequency")
plt.show()
```

Figure 1.9 Code for frequency of Quality of sleep

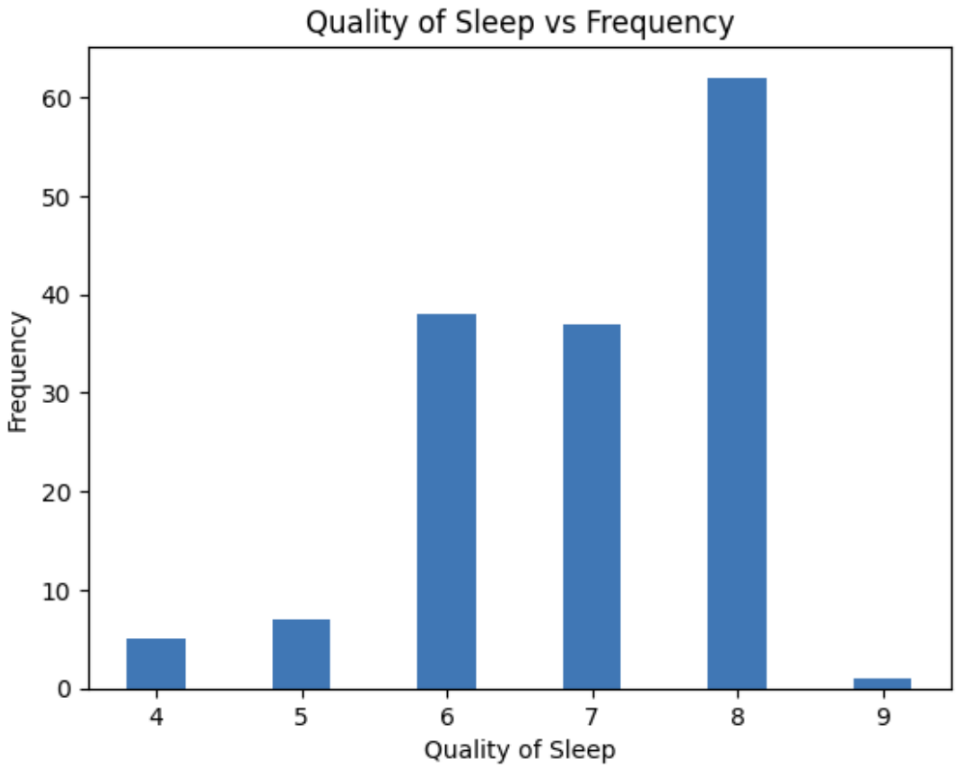

Figure 1.10 The bar graph frequency of Quality of sleep

Based on this chart, we can see most of the people in this sample have good sleep. Even though there are few appearances of 9 on the sleep quality scale, 7 and 8 are account for more than 50% of the frequency.

- All possible measurements

	Stress_Level
count	6.000000
mean	5.500000
std	1.870829
min	3.000000
25%	4.250000
50%	5.500000
75%	6.750000
max	8.000000

Figure 1.11 Five-number summary of Stress Level

	Quality_of_Sleep
count	6.000000
mean	6.500000
std	1.870829
min	4.000000
25%	5.250000
50%	6.500000
75%	7.750000
max	9.000000

Figure 1.12 Five-number summary of Quality of sleep

- Check for outliers

  This chart indicates that the data is symmetrically distributed around the median, and there seems no outlier in the data.

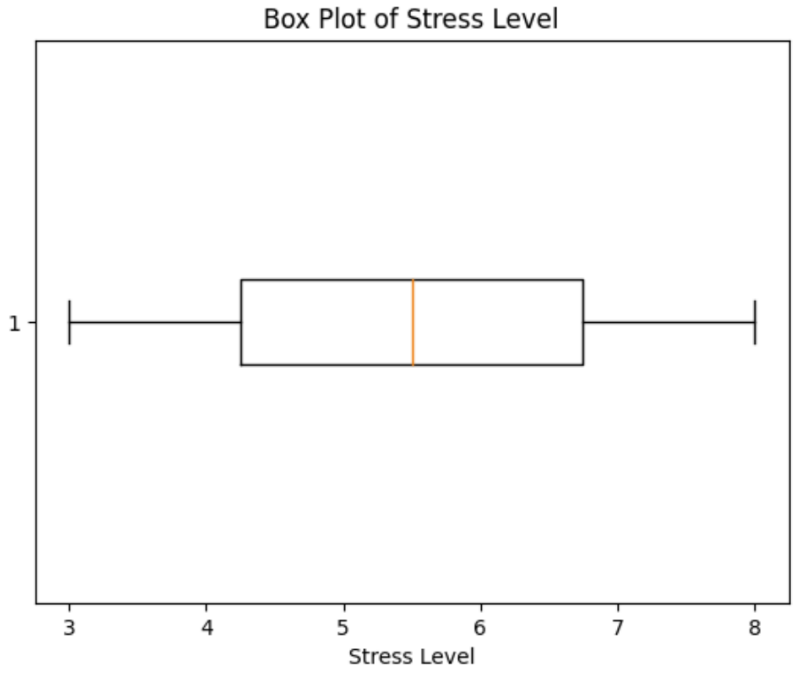

Figure 1.13 Boxplot of Stress Level

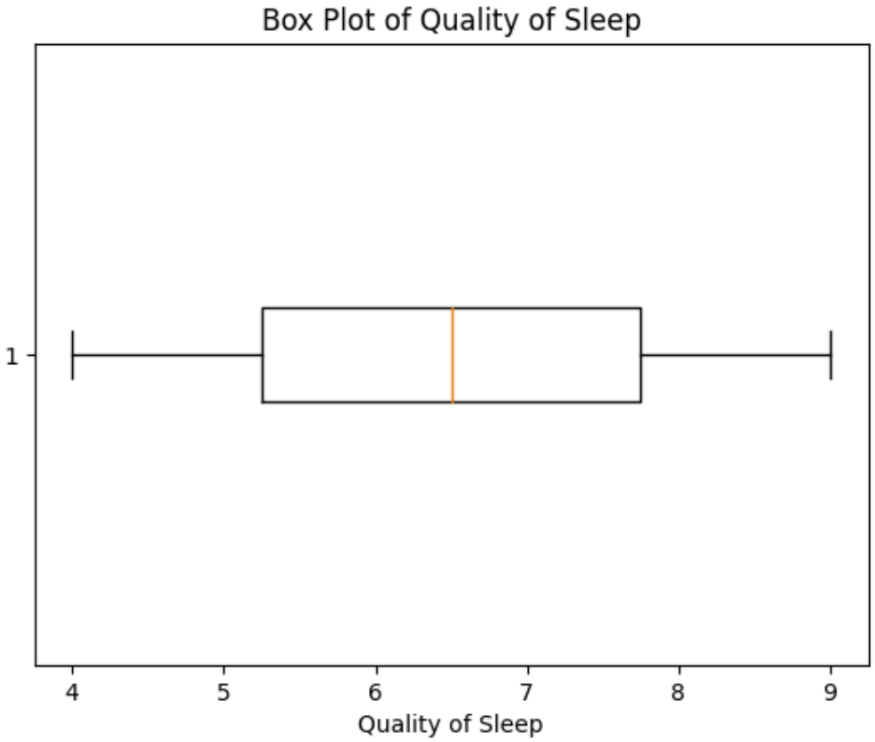

Figure 1.14 Boxplot of Quality of sleep

2. Relation between two variables

2.1. Draw and Interpret Scatter Diagrams

- Our team is currently studying the relationship between stress levels and sleep quality in men and women. In this research, stress levels are used as the explanatory variable (x-axis), while sleep quality is the response variable (y-axis). We have collected data and plotted the corresponding value pairs on a rectangular coordinate system, for example, points such as (6, 6), (8, 6), etc.

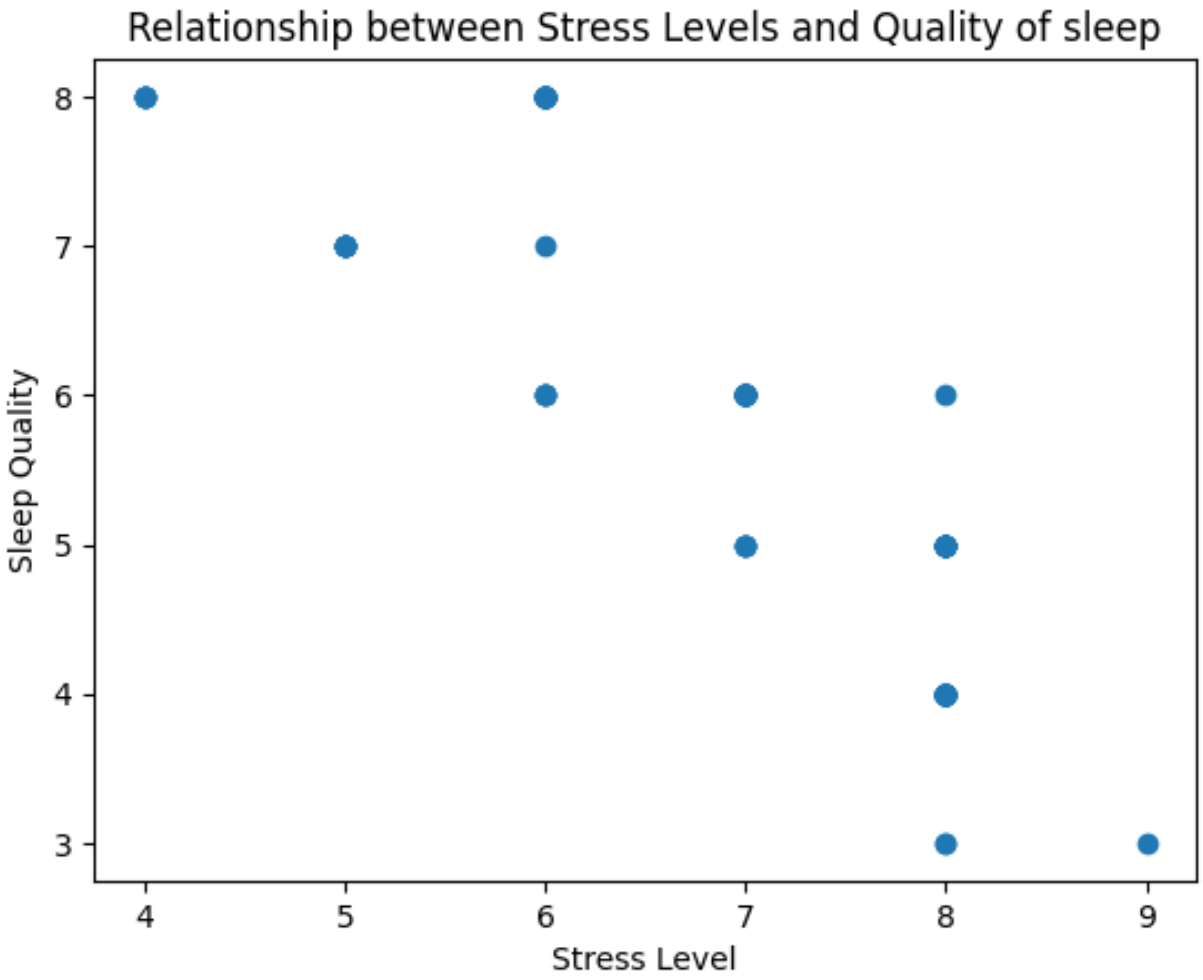

Figure 2.1 The scatter diagram.

- The scatter diagram in the above figure implies that stress level is negatively associated with the quality of sleep.

- To analyze this above, we observe a negative correlation between stress levels and sleep quality. Specifically, as stress levels increase, sleep quality tends to decrease. This is evident from the distribution of data points on the graph, showing a decreasing trend from left to right. Although the data points are not perfectly aligned along a straight line, there is still a general downward trend, suggesting that besides stress levels, other factors may influence sleep quality. The variability in the data also indicates that not all individuals with similar stress levels have the same sleep quality, demonstrating diversity in how sleep quality responds to stress levels.

2.2 Linear correlation coefficient

OLS Regression Results

Dep. Variable:	QualityofSleep	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.769
Method:	Least Squares	F-statistic:	497.0
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	3.67e-49
Time:	16:08:04	Log-Likelihood:	-114.83
No. Observations:	150	AIC:	233.7
Df Residuals:	148	BIC:	239.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.5008	0.164	64.178	0.000	10.177	10.824
StressLevel	-0.6070	0.027	-22.293	0.000	-0.661	-0.553

Omnibus:	52.049	Durbin-Watson:	1.364
Prob(Omnibus):	0.000	Jarque-Bera (JB):	102.565
Skew:	-1.603	Prob(JB):	5.35e-23
Kurtosis:	5.475	Cond. No.	23.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 2.2 OLS Regression Results

Based on the OLS table we can deduce the linear correlation coefficient using the following formula:

$$r = (sign\ b_1)\sqrt{R^2} = -\sqrt{0.771} = -0.878$$

Therefore, the correlation coefficient suggests a strong negative association between the two variables.

2.3. Testing for a Linear Relation

Step 1: The linear correlation coefficient between Stress Level and Quality of sleep is (- 0.878). So |-0.878| = 0.878.

Step 2: Table II shows the critical value with n = 150, we do not have critical values for correlation coefficient, but we know that this critical values for correlation coefficient < 0.361.

Step 3: Since |-0.878| = 0.878 > 0.361, we conclude a negative association (negative linear relation) exists between stress level and quality of sleep.

2.4. Least-Squares Regression Line

```
Correlation x and y
 [[ 1.         -0.87780061]
 [-0.87780061  1.        ]]
```

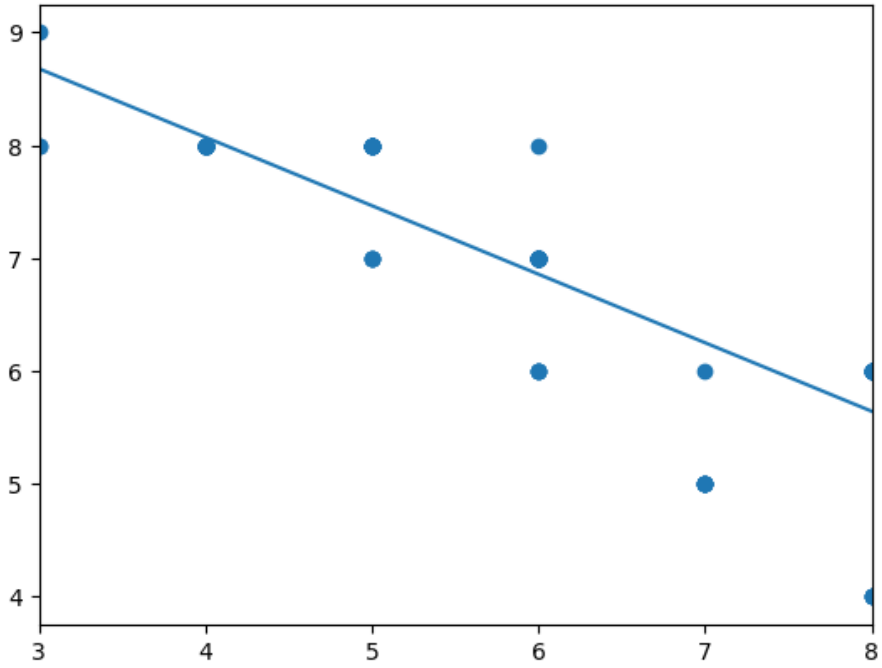

Figure 2.3 The line on the scatter diagram

Based on the OLS table we can deduce the Least-Squares Regression Line using the following formula:

- Stress Level is used as the explanatory variable (x-axis)
- Quality of sleep is the response variable (y-axis)

The Least-Squares Regression Line

$$y = 10.5008 - 0.6070x$$

Interpretation:

- The slope: for each 1 level increase in stress level then the quality of sleep decreases by 0.6070, on average.
- y-intercept: it is not appropriate because quality of sleep is not equal 0 hour.

Compute the Sum of Squared Residuals

```
## Residual Plot
from scipy.stats import norm
y_hat = -0.6070*df['StressLevel'] + 10.5008
# reg = model.fit()
yhat1 = reg.predict()
# yhat = reg.fittedvalues
residual = df['QualityofSleep'] - yhat1

data = {'residual':residual}
df1 = pd.DataFrame(data)
print(df1)

# plt.scatter(df['StressLevel'], residual)
plt.xlabel("Stress Level")
plt.ylabel("Residual")
plt.title("Stress Level vs Residual")
plt.boxplot(df1.residual, vert = False)
plt.show()
```

```
     residual
0   -0.858595
1    0.355459
2    0.355459
3   -1.644541
4   -1.644541
..        ...
145 -0.465622
146  0.534378
147 -1.251568
148  0.141405
149  0.320324
```

2.5. Compute and Interpret the Coefficient of Determination

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

explained deviation: $\hat{y} - \bar{y}$

unexplained deviation $y - \hat{y}$

Total Variation = Unexplained Variation + Explained Variation

$$(y - \bar{y})^2 = (y - \hat{y})^2 + (\hat{y} - \bar{y})^2$$

Based on the OLS table we can deduce the coefficient of determination (R^2):
R^2 = 0.771
Interpretation for R-square:
77.1% of the variability in y is explained by the least-squares regression line

2.6. Check outlier

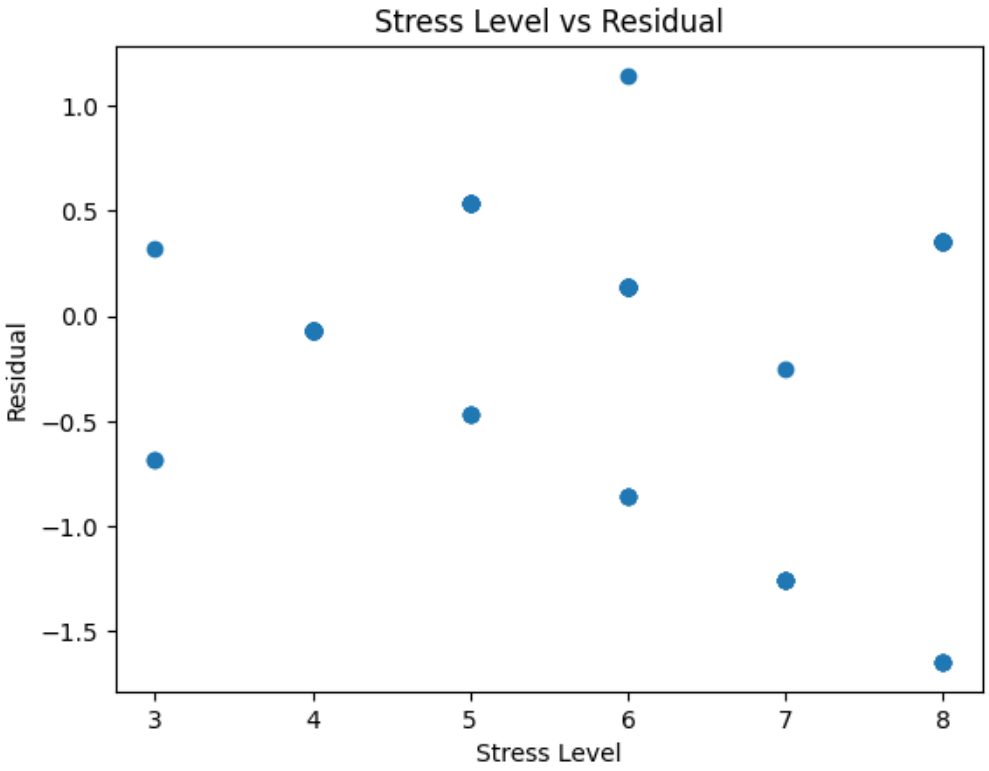

Figure 2.4 Scatter plot of residual

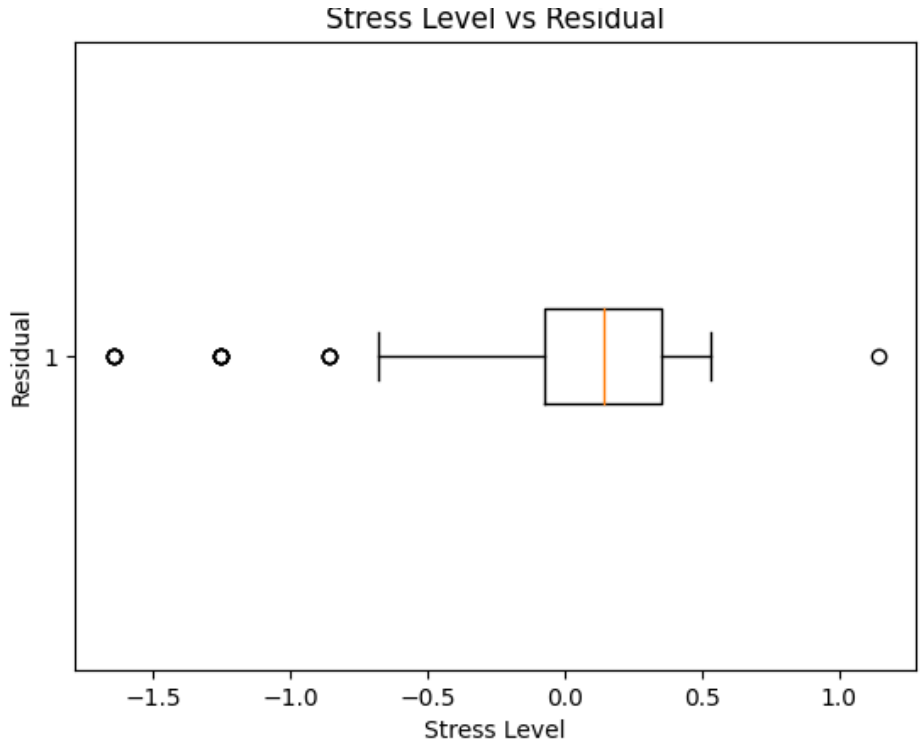

Figure 2.5 Boxplot of residual

- Yes, this plot shown no pattern → independent
- We see that the relative points lie across 0 → no mean of 0
- Points must lie between -0.7148 and 0.9972 to satisfy constant error variance. But currently we have some points that are outside this range and these points are called outlier. → no constant error variance.

Therefore, the appearance of these outlier may hide surprises in our group's data.

	residual
count	150.00000
mean	-0.00020
std	0.52201
min	-1.64480
25%	-0.07280
50%	0.14120
75%	0.35520
max	1.14120

Figure 2.6 Five-number summary of residual

IQR = Q3 - Q1 = 0.35520 - (-0.07280) =0.428
Lower fence = Q1 - 1.5(IQR) = -0.07280 - 1.5 × 0.428 = -0.7148
Upper fence = Q3 - 1.5(IQR) = 0.35520 + 1.5 × 0.428 = 0.9972

3. Check normality for 2 quantitative variables

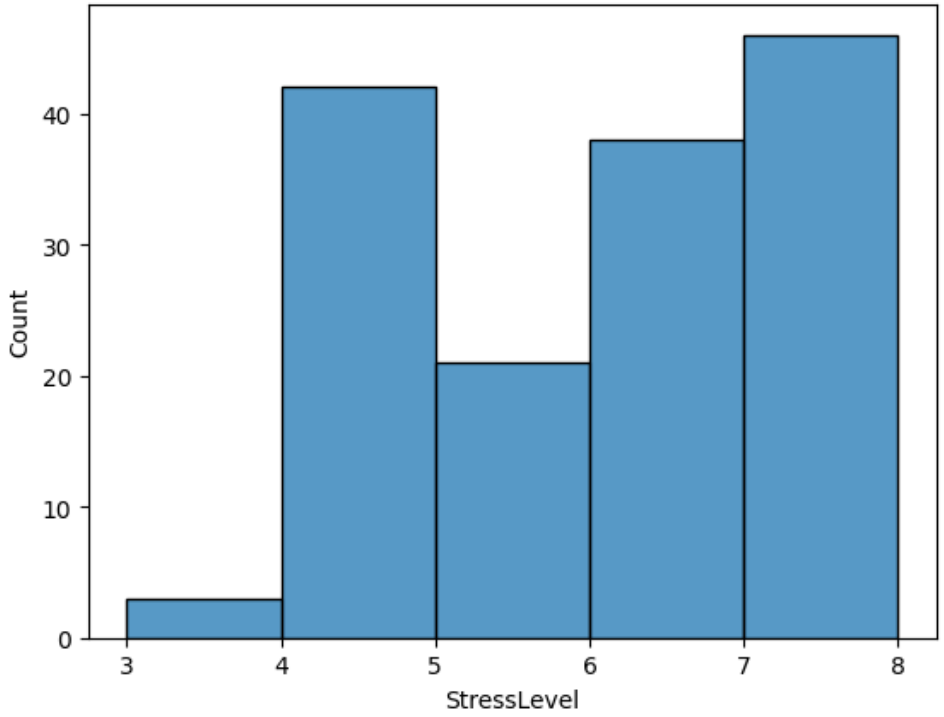

Figure 3.1 Histogram for Stress Level variables

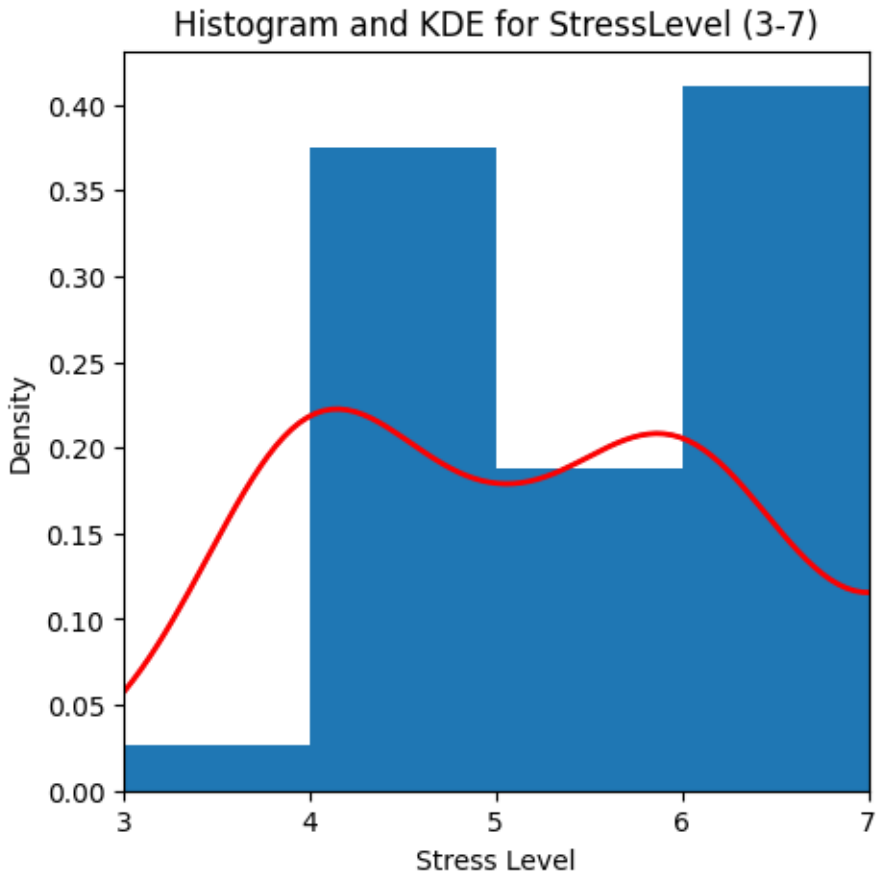

Figure 3.2 Histogram and KDE for Stress Level variables (range 3 – 7 level)

- Because the data type in our project is an ordinal data type, it does not provide enough information to determine its normality accurately.
- From the model above, it shows that this chart represents a multimodal distribution, characterized by the presence of multiple peaks instead of a single peak as commonly seen in a normal curve.[1]

Therefore, the model above is not referred to as a normal curve.

[1] https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-shape#:~:text=When%20a%20histogram%20is%20constructed,'%20or%20'bell%20curve

4. Construct Confidence Interval

```python
def confidence_interval(data, confidence=0.95):
    # Convert data to a numpy array
    data = np.array(data)

    # Calculate sample mean and standard error
    mean = np.mean(data)
    std_err = stats.sem(data)

    # Get the critical value (Z or t) based on the sample size
    n = len(data)
    if n > 30:
        critical_value = stats.norm.ppf((1 + confidence) / 2)
    else:
        critical_value = stats.t.ppf((1 + confidence) / 2, df=n-1)

    # Calculate the margin of error
    margin_of_error = critical_value * std_err

    # Construct the confidence interval
    lower_bound = mean - margin_of_error
    upper_bound = mean + margin_of_error

    return (lower_bound, upper_bound)
```

Figure 4.1 Confidence Interval algorithm for a population mean

```python
data = df['StressLevel']

# Calculate the 95% confidence interval
confidence_lvl = 0.95
ci = confidence_interval(data, confidence_lvl)

print(f"The {int(confidence_lvl*100)}% confidence interval is: {ci}")
```

```
The 95% confidence interval is: (5.547820261920501, 6.052179738079499)
```

Figure 4.2 Output value for CI

We are 95% confident that the mean stress level of both women, men is between 5.5478 and 6.0522.

This means that if we were to repeat the sampling process many times and calculate the 95% confidence interval for each sample, then 95% of these confidence intervals would contain the true average stress level for both men and women.

5. Perform Hypothesis Testing

According to a study mentioned in an article, the population mean of the stress level is 5.1.[2]

To conduct hypothesis testing, we need to use the aforementioned population mean. In the hypothesis testing section, we will determine whether the average

[2] https://www.apa.org/news/press/releases/stress/2015/snapshot

stress level in the sample data is significantly different from the hypothesized population mean.

Firstly, determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$

Note: μ_0 is the assumed value of the population mean.

In this section, use two-tailed Hypothesis Testing. Since the data is too large, use code to test.

```
# Hypothesis test for Stress Level (H0: mean = hypothesized value)
hypothesized_mean_stress = 5.1  # for example
t_statistic_stress, p_value_stress = stats.ttest_1samp(df['StressLevel'], hypothesized_mean_stress)

print(f'Test Statistics for Stress Level: {t_statistic_stress}')
print(f'P-value for Stress Level: {p_value_stress}')
```

```
Test Statistics for Stress Level: 5.440464010417555
P-value for Stress Level: 2.132692953604107e-07
```

From the results of the code, we see:

t_0 = 5.444 and p-value = 2.133e-07 < α = 0.05.

So, we reject the null hypothesis.

Given that the p-value is significantly smaller than common significance levels, we reject the null hypothesis. This suggests that there is strong evidence that the mean stress level in the sample is significantly different from the hypothesized population mean.

Chapter 3. Conclusion

Through statistical measures (Linear correlation coefficient), graphs, and hypothesis tests, results consistently demonstrate a significant correlation between stress levels and the quality of sleep individuals experience. This correlation illustrates that higher stress level leads to lower quality of sleep. High-stress levels often lead to disruptions in sleep patterns, causing difficulties in falling asleep, maintaining sleep throughout the night, and achieving restorative sleep cycles. Conversely, implementing stress-reduction techniques such as mindfulness meditation, relaxation exercises, and cognitive-behavioral therapy has been shown to improve sleep quality by alleviating tension and promoting relaxation, highlighting the intricate interplay between stress management and healthy sleep habits. Understanding and addressing this correlation are pivotal in fostering optimal sleep hygiene and promoting holistic wellness.

References

1. https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-shape#:~:text=When%20a%20histogram%20is%20constructed,'%20or%20'bell%20curve
2. https://www.apa.org/news/press/releases/stress/2015/snapshot
3. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
4. Textbook: Michael Sullivan III - Statistics_ Informed Decisions Using Data-Pearson (2012)