Sleep Health and Lifestyle Project Report

Sara Akbarzadeh

March 2025

Abstract

Sleep plays a crucial role in maintaining physical, mental, and emotional well-being. In this study, we examine the connections between sleep health and various lifestyle factors using the Sleep Health and Lifestyle Dataset from Kaggle. By conducting exploratory data analysis (EDA) and performing hypothesis testing, we uncover significant relationships between sleep quality and factors like stress levels, physical activity, and BMI. The insights gained from our findings can guide strategies to enhance sleep health and point to areas that warrant further investigation.

1 Introduction

The relationship between sleep health and lifestyle factors is crucial for understanding and improving overall well-being. In this project, we utilize the Sleep Health and Lifestyle Dataset from Kaggle, which consists of 374 rows and 13 columns, encompassing a range of variables that affect sleep quality and daily habits. These variables include gender, age, occupation, sleep duration, sleep quality, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

The primary goal of this study is to analyze the interplay between various lifestyle factors and sleep health. To achieve this, we first conducted an exploratory data analysis (EDA) to understand the structure of the dataset, identify any missing or invalid values, and explore the data through visualization techniques. Subsequently, we tested several hypotheses to uncover significant correlations between sleep health and lifestyle factors such as stress levels, physical activity, and BMI.

The hypotheses examined include the distribution of sleep duration in women, the effect of daily steps on sleep quality, the variation of stress levels across different occupations, the relationship between BMI categories and blood pressure, and whether individuals with sleep disorders have higher heart rates compared to those without. Additionally, we proposed and tested new hypotheses to gain deeper insights, earning bonus points for extending the analysis beyond the primary objectives.

Through these comprehensive analyses, we aim to provide valuable insights into factors that contribute to sleep health and suggest areas for further research.

2 Dataset

The Sleep Health and Lifestyle Dataset comprises 374 rows and 13 columns, representing various aspects of sleep and lifestyle habits. The dataset contains the following features:

Feature Name	Data Type	Description
PersonID	Categorical	Unique identifier for each individual (not useful for training).
Gender	Categorical	Gender of the individual (e.g., Male, Female).
Age	Numerical (integer)	Age of the individual in years.
Occupation	Categorical	Profession or occupation (e.g., Student, Engineer).
Sleep Duration	Numerical (hours)	Number of hours of sleep per night.
Quality of Sleep	Numerical (scale from 1 to 10)	Quality of sleep rated on a scale from 1 to 10.
Physical Activity Level	Numerical (minutes per day)	Amount of physical activity performed daily.
Stress Level	Numerical (score)	Stress level typically represented as a score or scale.
BMI Category	Categorical	BMI classification (e.g., Normal, Overweight).
Blood Pressure	Numerical (mmHg)	Blood pressure measured in millimeters of mercury.
Heart Rate	Numerical (beats per minute)	Heart rate of the individual.
Daily Steps	Numerical (count)	Number of steps taken per day.
Sleep Disorder	Categorical	Presence or absence of a sleep disorder.

Table 1: Description of Features in the Sleep Health and Lifestyle Dataset

3 Visualization

We visualized the distributions of both numerical and categorical features.

3.1 Distribution Examples

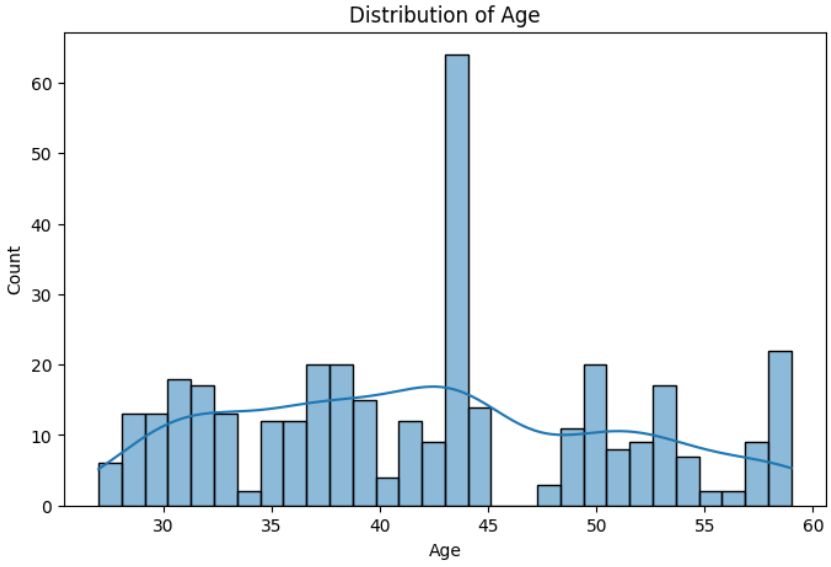

Figure 1: Distribution of Age

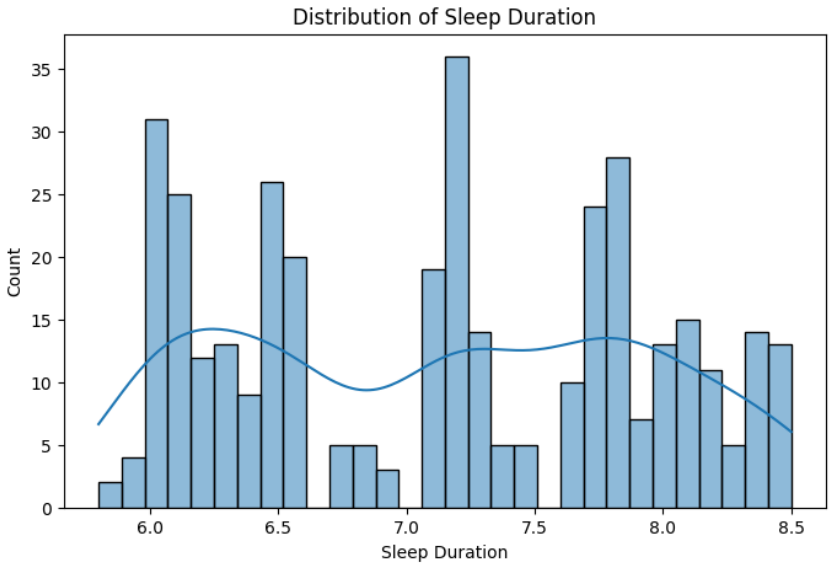

Figure 2: Distribution of Sleep Duration

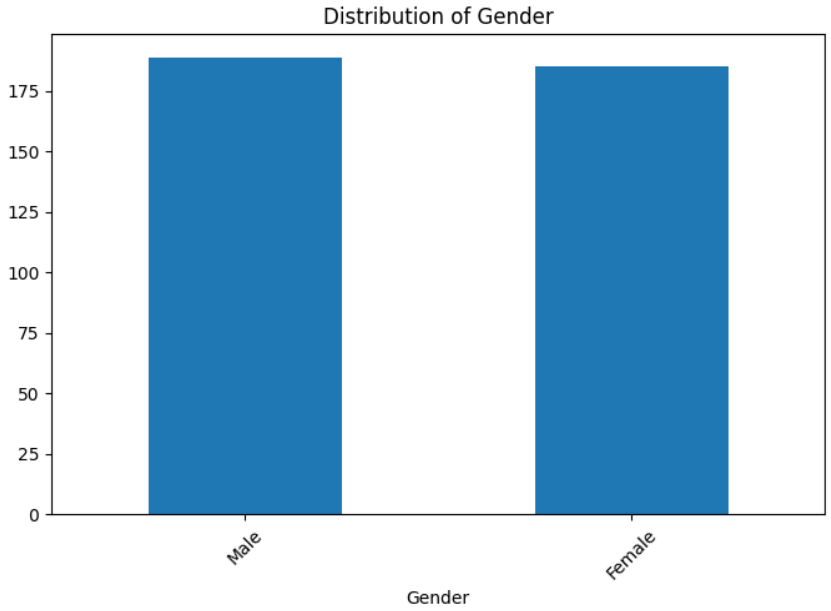

Figure 3: Distribution of Gender

4 Exploratory Data Analysis (EDA)

During the Exploratory Data Analysis (EDA) phase, the following significant findings were discovered:

- Women are more likely to suffer from sleep apnea.
- Individuals with normal sleep patterns typically sleep between 7 to 8 hours per night.
- There is a strong negative correlation between stress levels and sleep quality.

These findings emphasize the significance of gender-specific and stress-related factors in sleep health. The visual representations of the data are provided in Figures 4, 5, and 6.

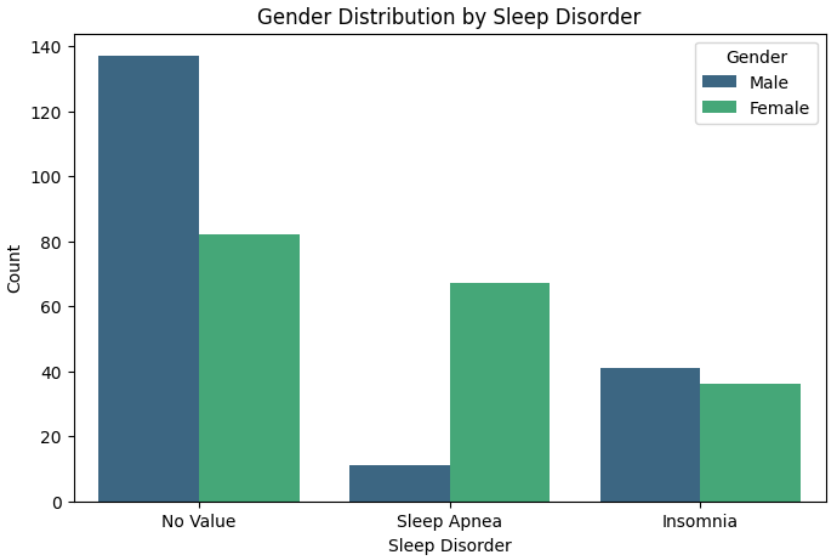

Figure 4: Gender Distribution by Sleep Disorder

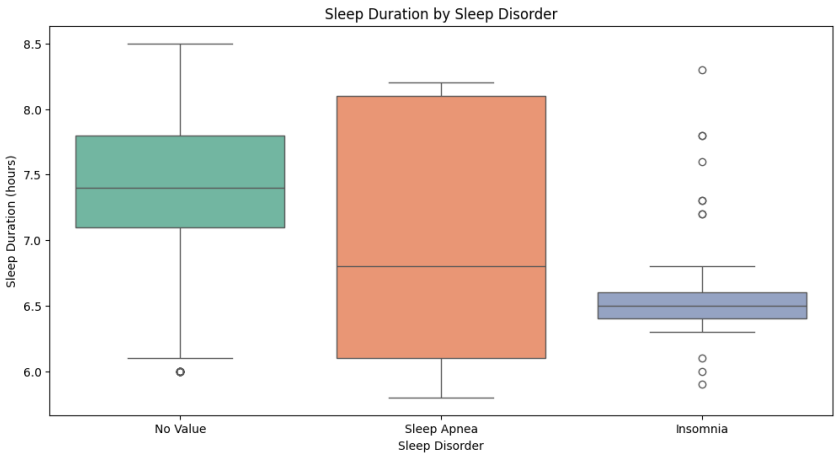

Figure 5: Sleep Duration by Sleep Disorder

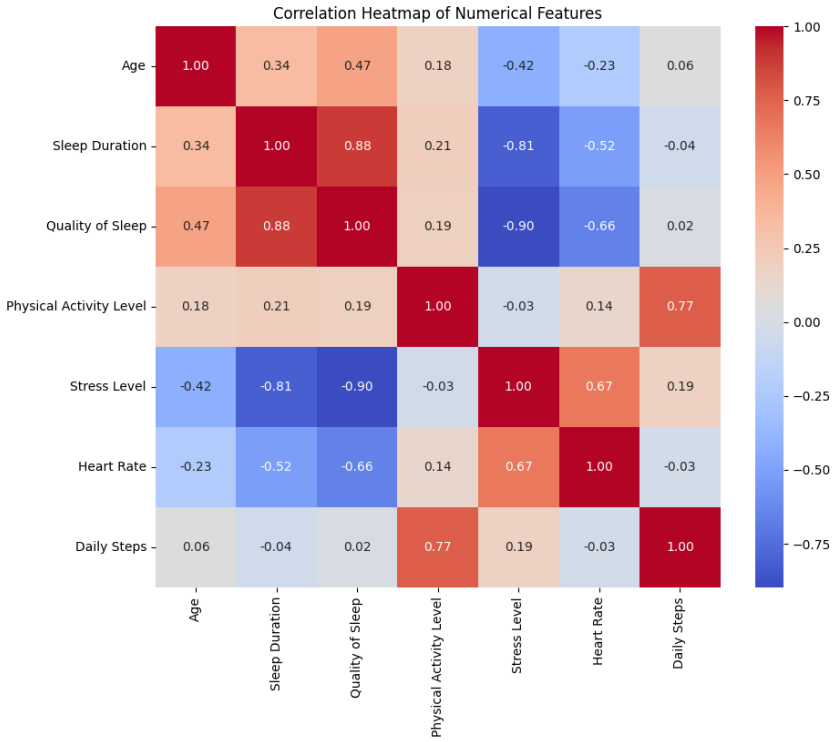

Figure 6: Correlation Heatmap of Numerical Features

5 Hypothesis Testing

5.1 Hypothesis a: Does women's sleep duration follow a normal distribution?

- **Null Hypothesis (H_0)**: The sleep duration of women follows a normal distribution.
- **Alternative Hypothesis (H_1)**: The sleep duration of women does not follow a normal distribution.

We used the Shapiro-Wilk Test to evaluate normality because it is effective for small to medium-sized datasets.

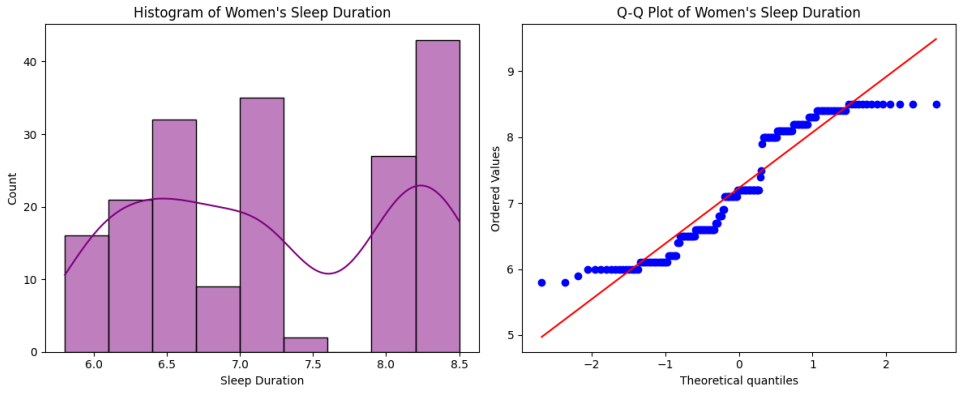

Figure 7: Women's Sleep

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis (data is normal).
- IIf the p-value is less than 0.05, we reject the null hypothesis (data is not normal).

Result:

- Shapiro-Wilk Test Statistic: 0.8985768556594849, p-value: 6.360978965780362e-10
- Reject the null hypothesis: The sleep duration does not follow a normal distribution.

5.2 Hypothesis b: Is having higher daily steps a contributing factor into better sleep?

- **Null Hypothesis (H_0)**: There is no significant correlation between Daily Steps and Quality of Sleep.
- **Alternative Hypothesis (H_1)**: There is a significant correlation between Daily Steps and Quality of Sleep.

We used the Pearson correlation coefficient to quantify the strength and direction of the linear relationship between the two variables.

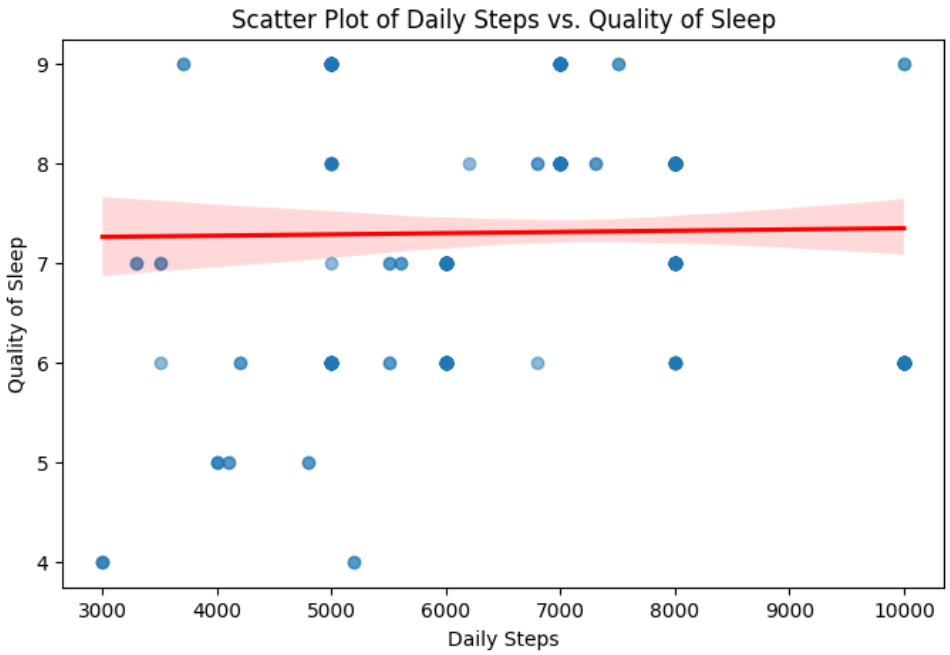

Figure 8: Daily Steps and Quality of Sleep

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- Pearson Correlation Coefficient: 0.02 and P-value: 0.746190665296111
- There is no significant correlation between daily steps and quality of sleep.

5.3 Hypothesis c: Is stress level different among different occupations?

- **Null Hypothesis (H_0)**: The mean stress level is the same across all occupations.
- **Alternative Hypothesis (H_1)**: The mean stress level differs among occupations.

The appropriate statistical test for comparing means across multiple groups is the one-way ANOVA (Analysis of Variance).

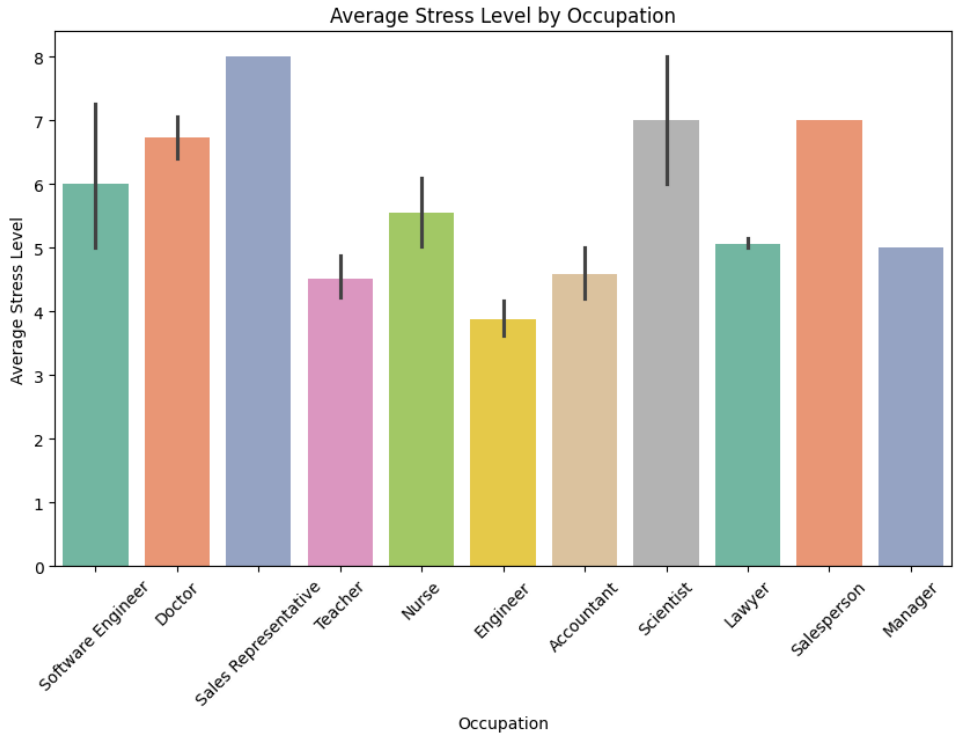

Figure 9: Average Stress by Occupation

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- ANOVA result: F=21.63598878521177, p=1.355091231304278e-31
- Reject the null hypothesis: The mean stress level differs among occupations.

5.4 Hypothesis d: Are different BMI categories significantly different given their blood pressure?

- **Null Hypothesis (H_0)**: There is no significant difference in the mean blood pressure (systolic and diastolic) between different BMI categories.
- **Alternative Hypothesis (H_1)**: There is a significant difference in the mean blood pressure (systolic and/or diastolic) between at least one pair of BMI categories.

We applied a One-Way ANOVA to test if there are statistically significant differences in the mean blood pressure across the different BMI categories.

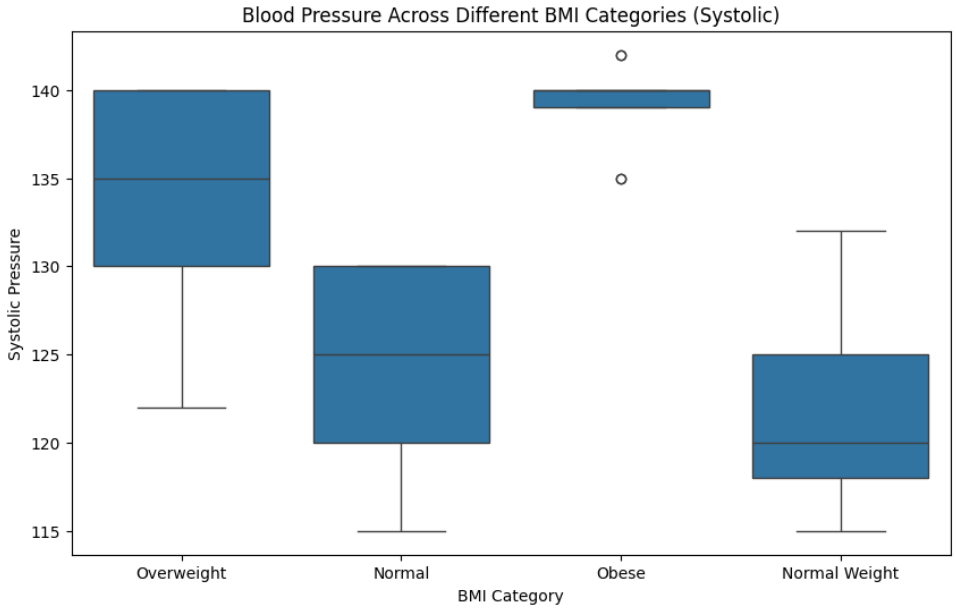

Figure 10: Systolic

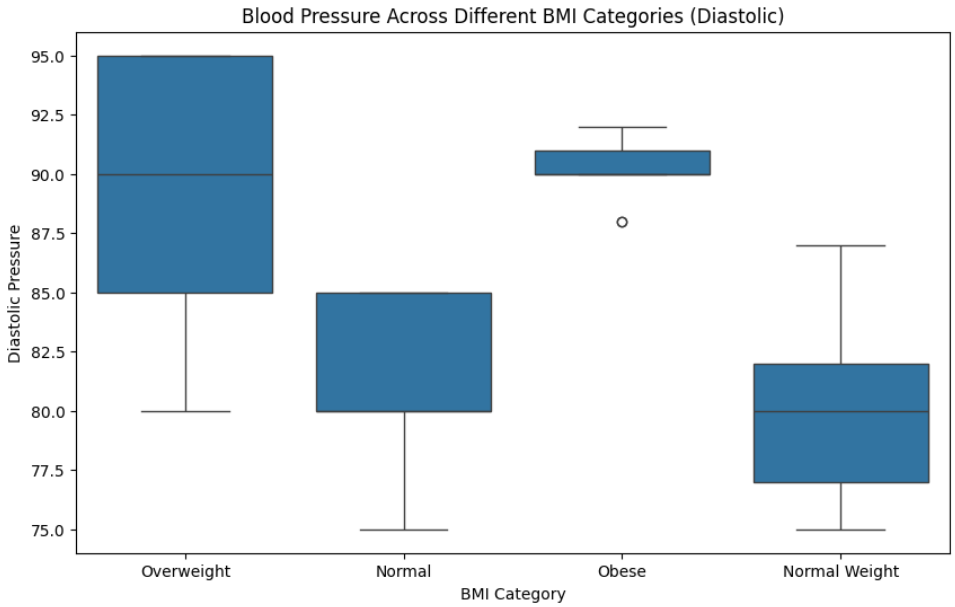

Figure 11: Diastolic

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- Systolic Pressure - Statistic: 163.545195439236, p-value: 1.743876808083431e-67
- Diastolic Pressure - Statistic: 179.09700388935, p-value: 1.0189415858614994e-71
- Reject the null hypothesis: Systolic blood pressure is significantly different among BMI categories.
- Reject the null hypothesis: Diastolic blood pressure is significantly different among BMI categories.

5.5 Hypothesis e: Do people with sleep disorders have higher heart rates than those without any sleep disorder?

- **Null Hypothesis (H_0)**: There is no difference in mean heart rates between people with and without sleep disorders.
- **Alternative Hypothesis (H_1)**: People with sleep disorders have higher mean heart rates than those without.

We performed an Independent t-test to compare the mean heart rates between two groups: people with sleep disorders and people without sleep disorders

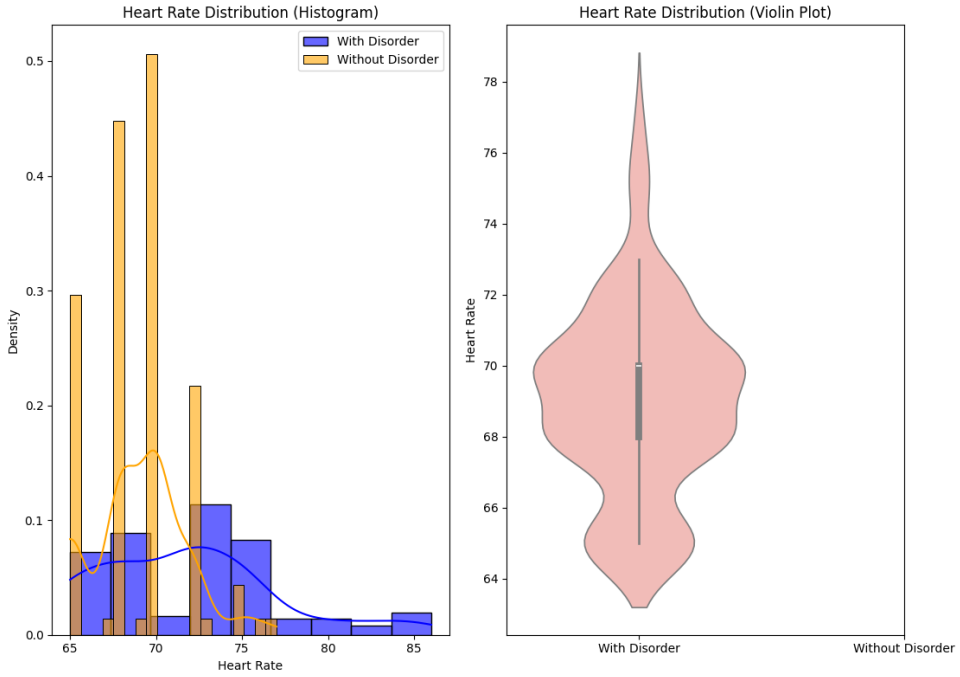

Figure 12: Heart Rate Distribution

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- T-statistic: 6.74833933997991 and P-value: 2.8747938607376974e-11
- Reject the null hypothesis: People with sleep disorders have significantly higher heart rates.

We have designed some bonus tests to further evaluate the robustness and validity of our findings.

5.6 Hypothesis f: Is there a significant negative correlation between sleep quality and stress level?

- **Null Hypothesis (H_0)**: There is no correlation between sleep quality and stress level.
- **Alternative Hypothesis (H_1)**: There is a negative correlation between sleep quality and stress level.

We utilized Pearson correlation analysis to assess the strength and direction of relationships

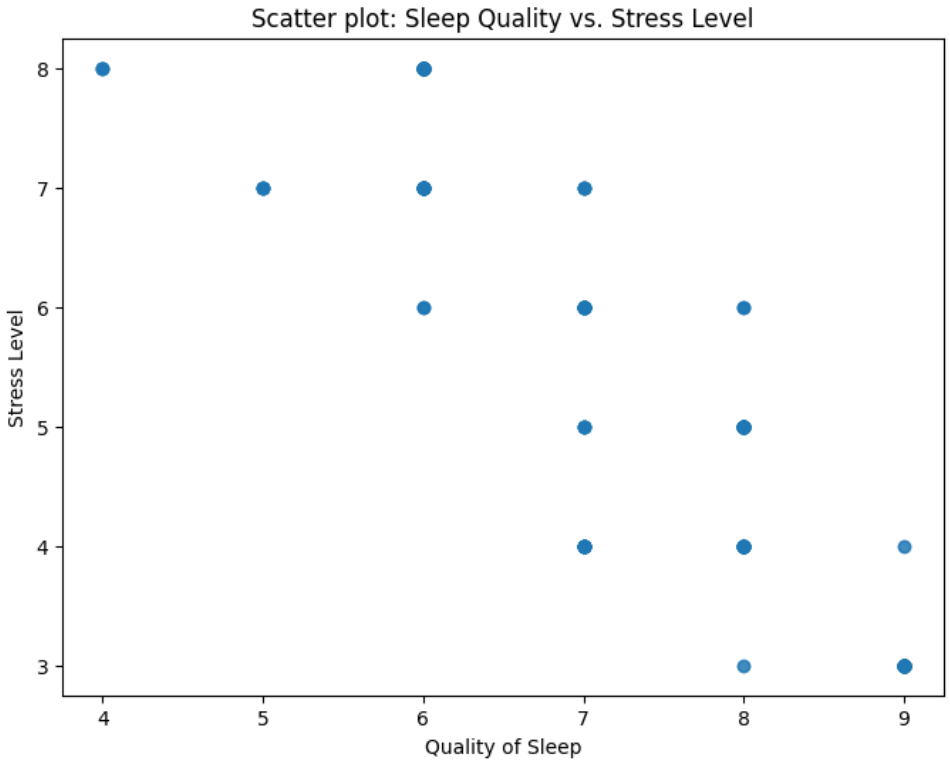

Figure 13: Heart Rate Distribution

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- Correlation: -0.8987520310040419, P-value: 2.880124240941155e-135
- Reject the null hypothesis: There is a significant negative correlation between sleep quality and stress level.

5.7 Hypothesis g: Is sleep disorder occurrence independent of gender?

- **Null Hypothesis (H_0)**: Sleep disorder occurrence is independent of gender.
- **Alternative Hypothesis (H_1)**: Sleep disorder occurrence is dependent on gender.

We employed the Chi-squared (^{2}) test to examine the association between them.

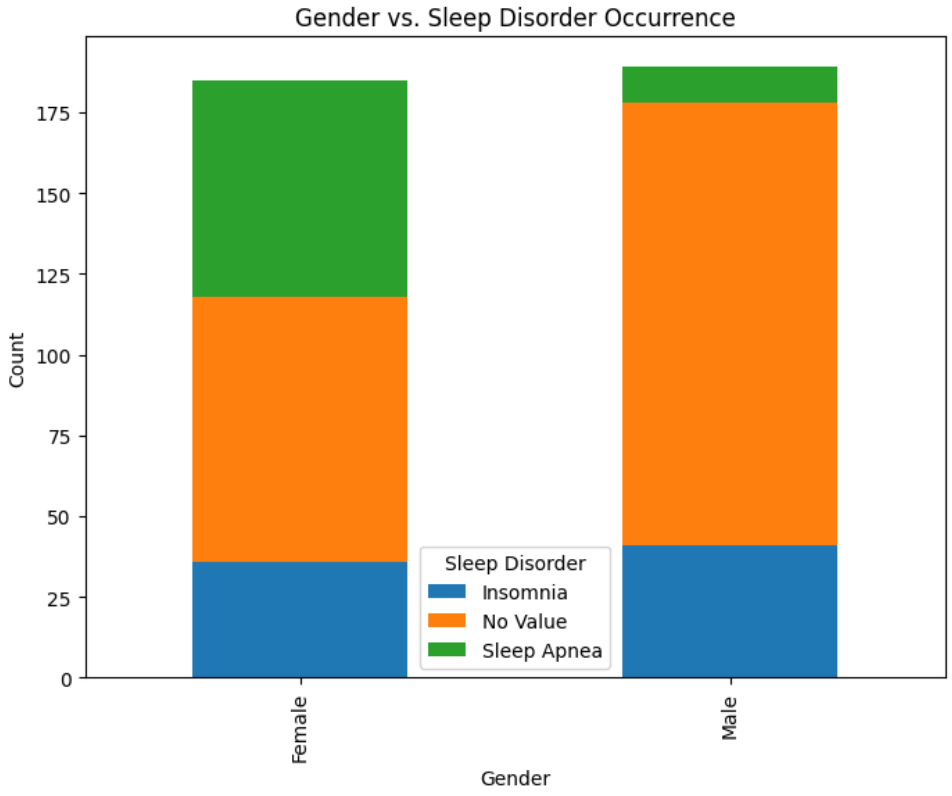

Figure 14: Gender vs Sleep Disorder

Interpretation:

- If the p-value is more than 0.05, we fail to reject the null hypothesis.
- IIf the p-value is less than 0.05, we reject the null hypothesis.

Result:

- Chi-Square Statistic: 54.30602007353474, P-value: 1.6128633524576768e-12
- Reject the null hypothesis: Sleep disorder occurrence is dependent on gender.

6 Conclusion

This study revealed significant associations between sleep health and factors such as age, gender, stress levels, and BMI. Notably, we observed a negative correlation between age and sleep duration, a higher prevalence of sleep apnea among women, and a substantial impact of stress on sleep quality. These findings highlight the need for personalized interventions aimed at enhancing sleep health. Future research could focus on analyzing longitudinal data to establish causality and explore the influence of other lifestyle factors.

7 Future Work

In future work, incorporating machine learning models to predict sleep quality based on demographic and lifestyle factors could provide valuable insights. Additionally, exploring longitudinal data may help identify trends over time and assess the long-term impact of interventions. Including more diverse datasets from different populations could enhance the generalizability of the findings.

8 References

- Kaggle. "Sleep Health and Lifestyle Dataset" Available at: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.
- NumPy Documentation. "NumPy: The Fundamental Package for Scientific Computing with Python." Available at: https://numpy.org/doc/.
- Matplotlib Documentation. "Matplotlib: Python Plotting." Available at: https://matplotlib.org/stable/contents.html.
- Seaborn Documentation. "Seaborn: Statistical Data Visualization." Available at: https://seaborn.pydata.org/.