Analyzing Sleep Health Using Lifestyle Data

Group Number: 5

Group Members:

Tejaswini Mode: z1966081
Nagashri Rao: z1966150
Dhaval Shah: z1977916

Index

Data Source

The dataset used for this project is the "Sleep Health and Lifestyle Dataset" obtained from Kaggle. It contains information about various factors related to sleep health and lifestyle habits.

1) The dataset was obtained from the Kaggle website, a popular platform for data science competitions and datasets.

2) No, the dataset was a single file and did not require combining multiple data sources.

3) The dataset contains 374 observations (rows) and 13 variables (columns), including a mix of categorical and numerical variables.

4) No, the dataset is publicly available and does not require confidentiality.

Dataset Description

The dataset includes information about individuals' sleep duration, quality of sleep, physical activity levels, stress levels, BMI categories, blood pressure, heart rate, daily steps, and whether they have a sleep disorder or not. The variables are a mix of numerical (e.g., age, sleep duration, daily steps) and categorical (e.g., gender, occupation, BMI category).

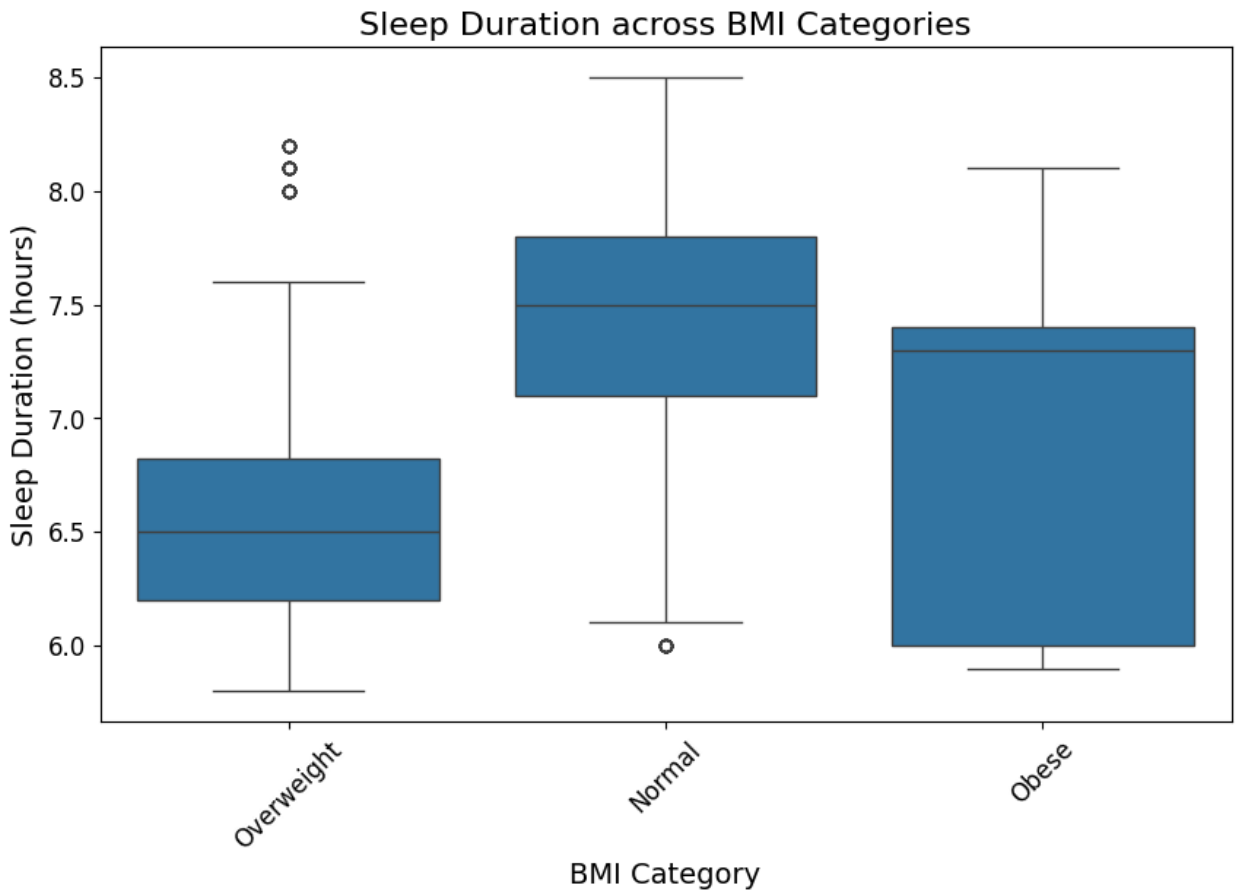
Key Variables

Sleep Duration: The average number of hours an individual sleeps per night, ranging from 5.8 to 8.5 hours.
Quality of Sleep: A rating of sleep quality on a scale of 4 to 9, with higher values indicating better sleep quality.
Physical Activity Level: A measure of an individual's physical activity level, ranging from 30 to 90.
Stress Level: A rating of stress levels on a scale of 3 to 8, with higher values indicating higher stress.

Visualizations

Several visualizations were created to explore the distributions of key variables and their relationships:

Box plot: Comparing the distribution of sleep duration across different BMI categories.



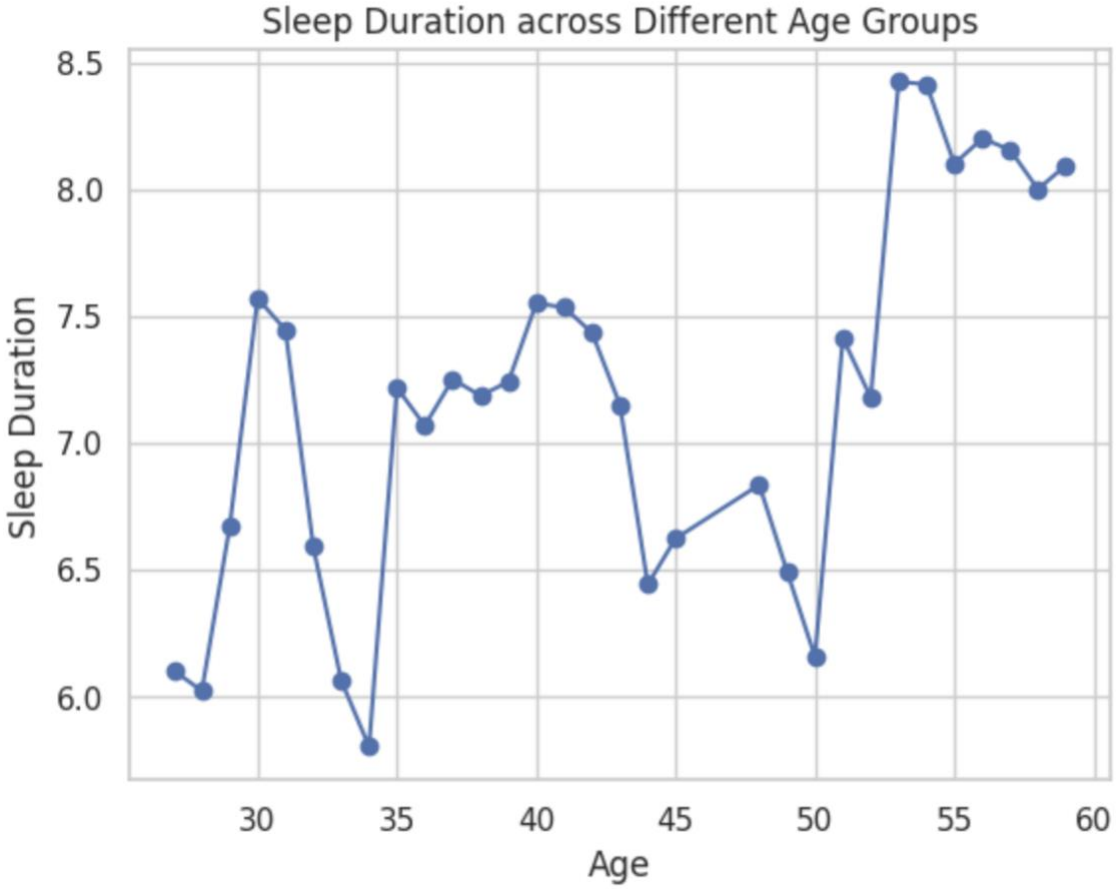
Analysis:

Overweight: This group's median sleep duration is around 6.5 hours, with most data concentrated between approximately 6.25 and 6.75 hours. There are very few outliers in this category.

Normal: This group's median is slightly higher than the Overweight group, approximately 7.5 hours, with a wider IQR stretching from about 7.25 to 7.75 hours. There are a few outliers suggesting some individuals get significantly less sleep.

Obese: The Obese category shows the highest median sleep duration, close to 7.25 hours, and a tight concentration of values, mostly ranging from 6 to 7.5 hours.
This visualization helps compare the central tendency and variability of sleep durations across different BMI categories. The data suggests that people categorized as Obese tend to sleep slightly longer than those in the other categories, on average.

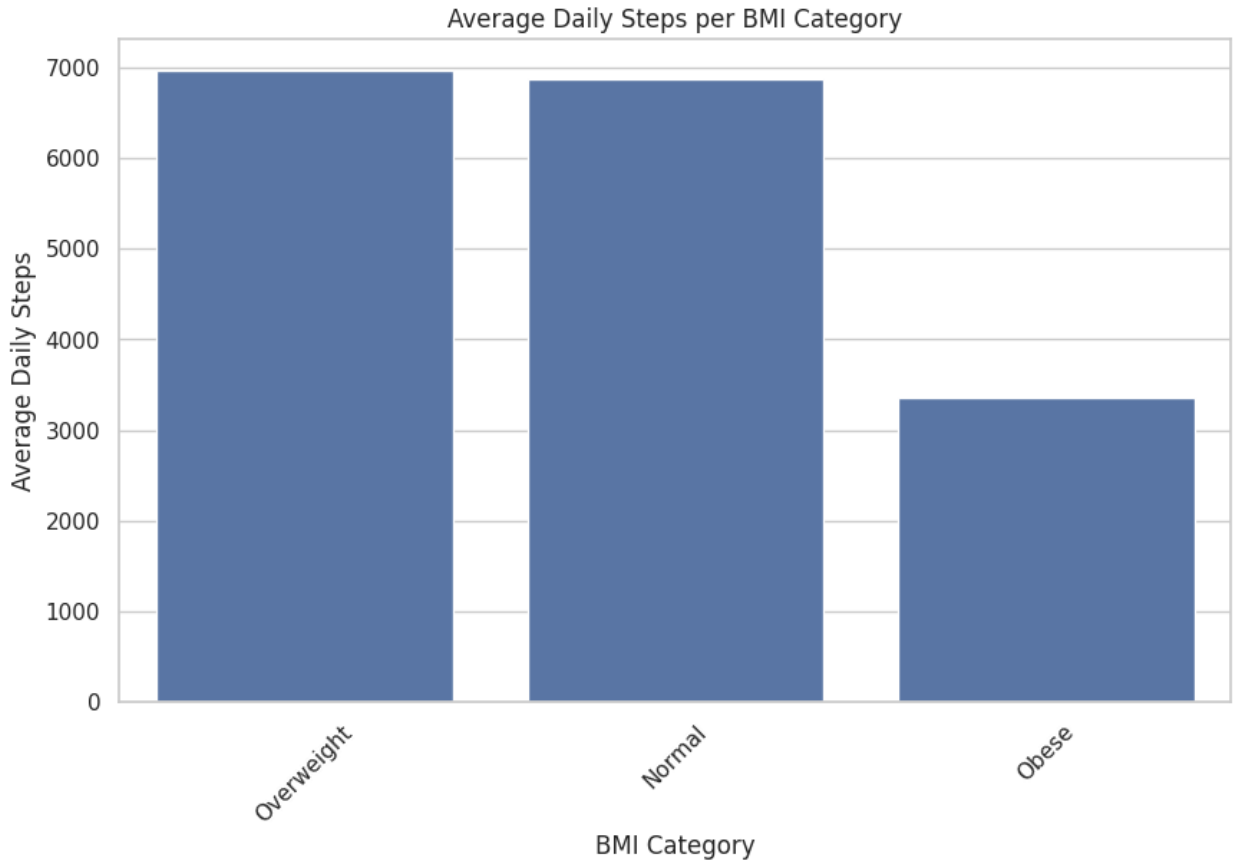
Line Plot: Comparing sleep duration across different age groups.



Analysis:

The sleep duration across various age groups shows significant fluctuation without a clear linear trend. The data points suggest variability in sleep duration with age.

Bar plot: Displaying the average daily steps for each BMI category.



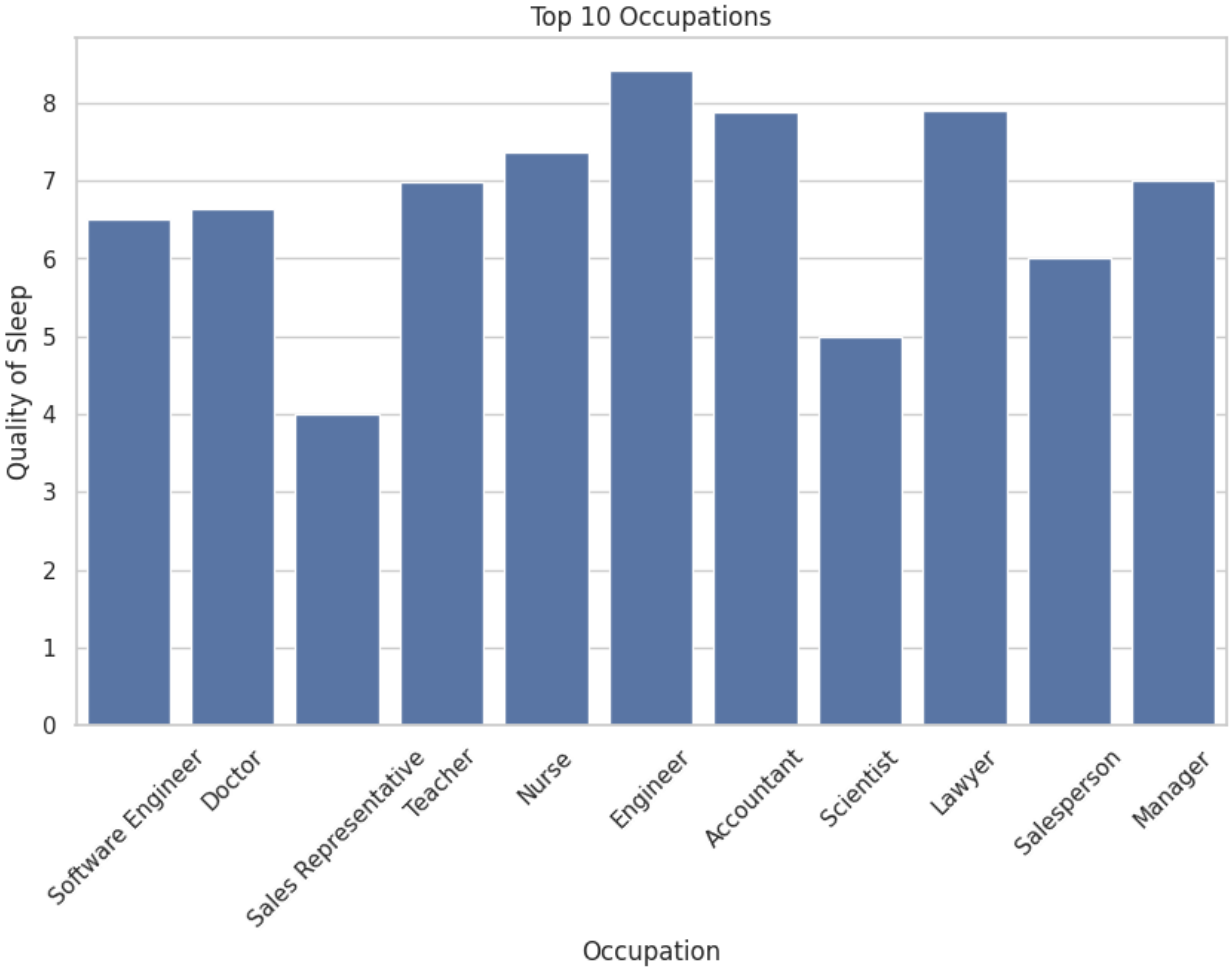
Overweight: Individuals in this group average around 7,000 daily steps.

Normal: This group has the highest average, slightly above 6,500 daily steps, indicating that individuals with a "Normal" BMI tend to be more active in terms of walking compared to the other groups.

Obese: Individuals in the Obese category show a significant decrease in average daily steps, around 3,500, which is nearly half of what individuals in the Normal category average.

The plot illustrates a clear trend whereas BMI increases from Normal to Obese, the average number of daily steps decreases. This could suggest a correlation between lower physical activity levels and higher BMI categories.

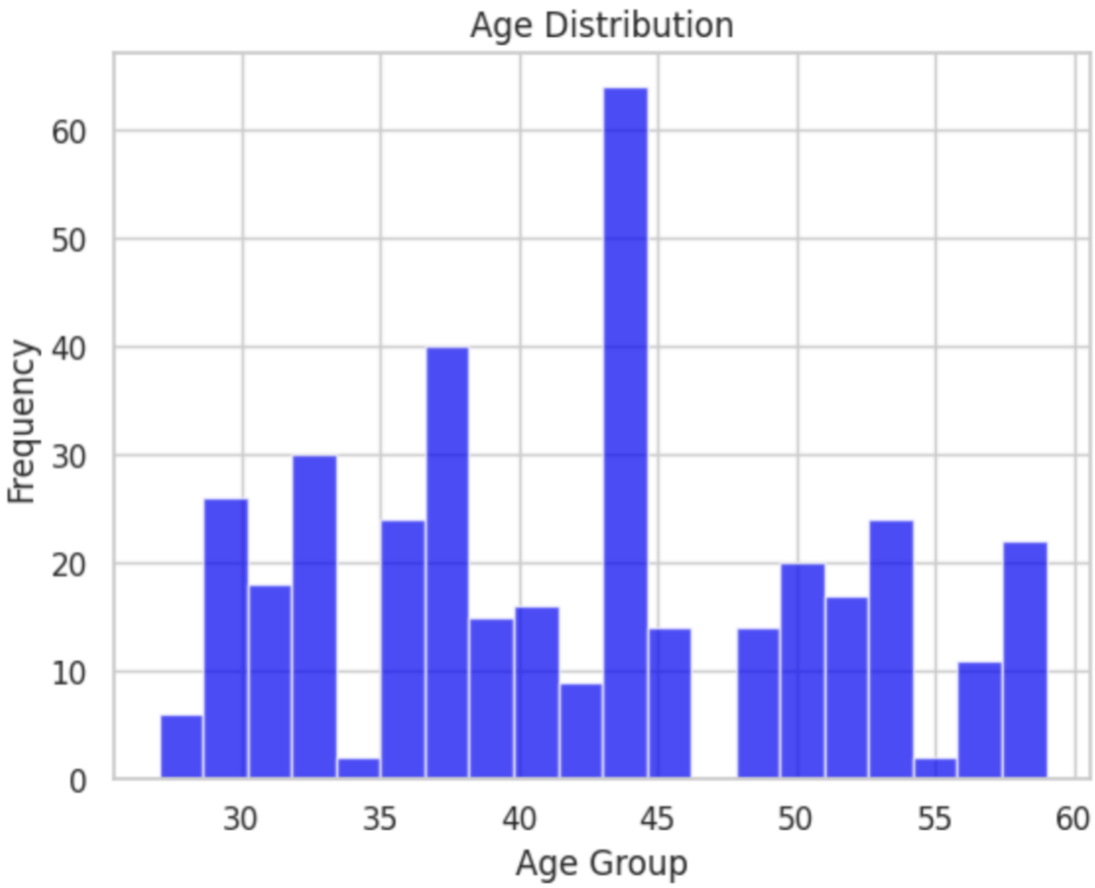
Bar plot: Showing the quality of sleep for the top 10 occupations.



- Engineers, Nurses, Accountants, and Lawyers appear to have the highest quality of sleep among the occupations listed, all scoring above 7 on the quality scale.
- Software Engineers, Doctors, Teachers, and Managers have moderately high sleep quality, roughly between 6 and 7 on the scale.
- Sales Representative report the lowest sleep quality, with scores clearly below 4.

This chart could reflect the varying stress levels, work hours, and job demands associated with each occupation, which are factors known to impact sleep quality. For instance, the lower sleep quality observed in sales representatives and scientists might be attributed to shift work, high stress, and irregular working hours common in these professions.

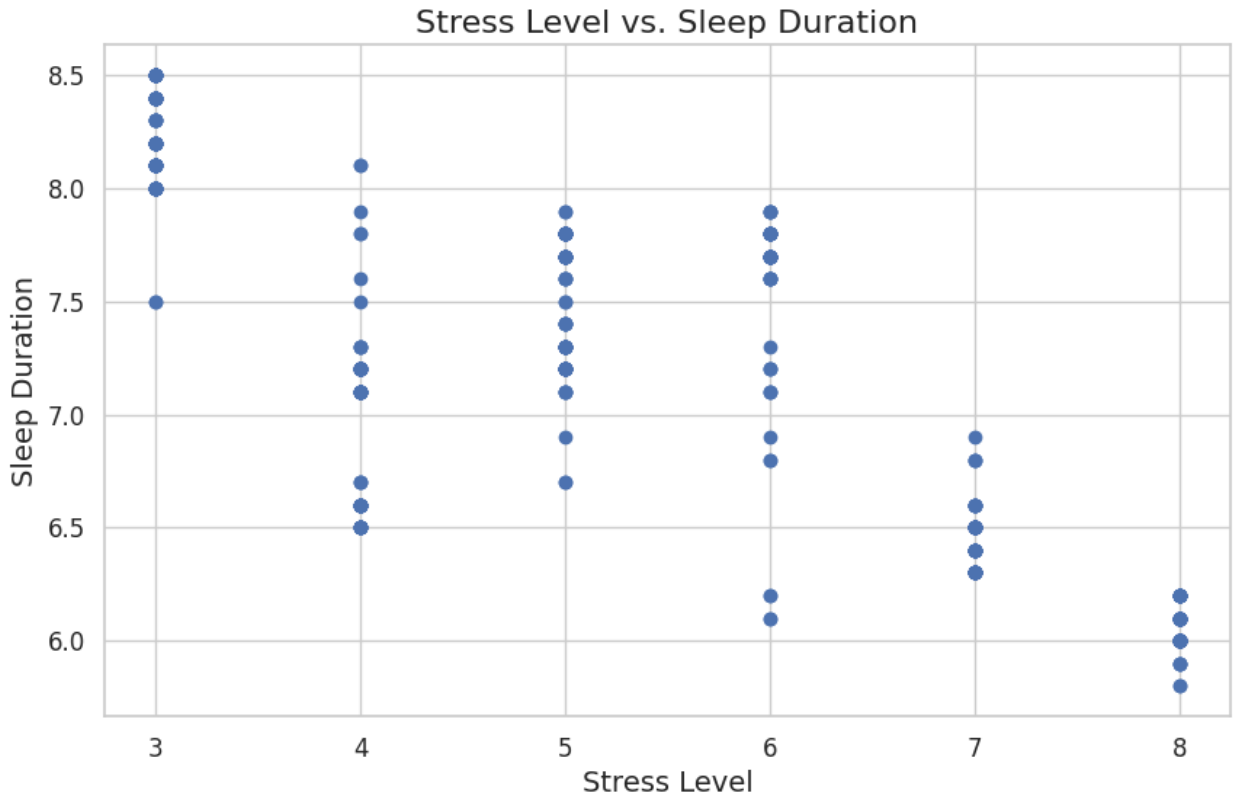
Histogram: Visualizing the distribution of ages in the dataset.



- The x-axis represents age groups in intervals, presumably in 5-year increments while the y-axis represents the frequency of individuals within each age group.
- The tallest bar is at the 45-year age group, indicating this is the most common age range in the sample.
- There's a general increase in frequency from the 30s to the 45-year age group, suggesting an accumulation of individuals up to this peak.
- Post 45, the frequency decreases, particularly noticeable in the group immediately following (likely 50-54), and then appears to increase slightly in the later age groups, suggesting a smaller resurgence of frequency as the population ages.

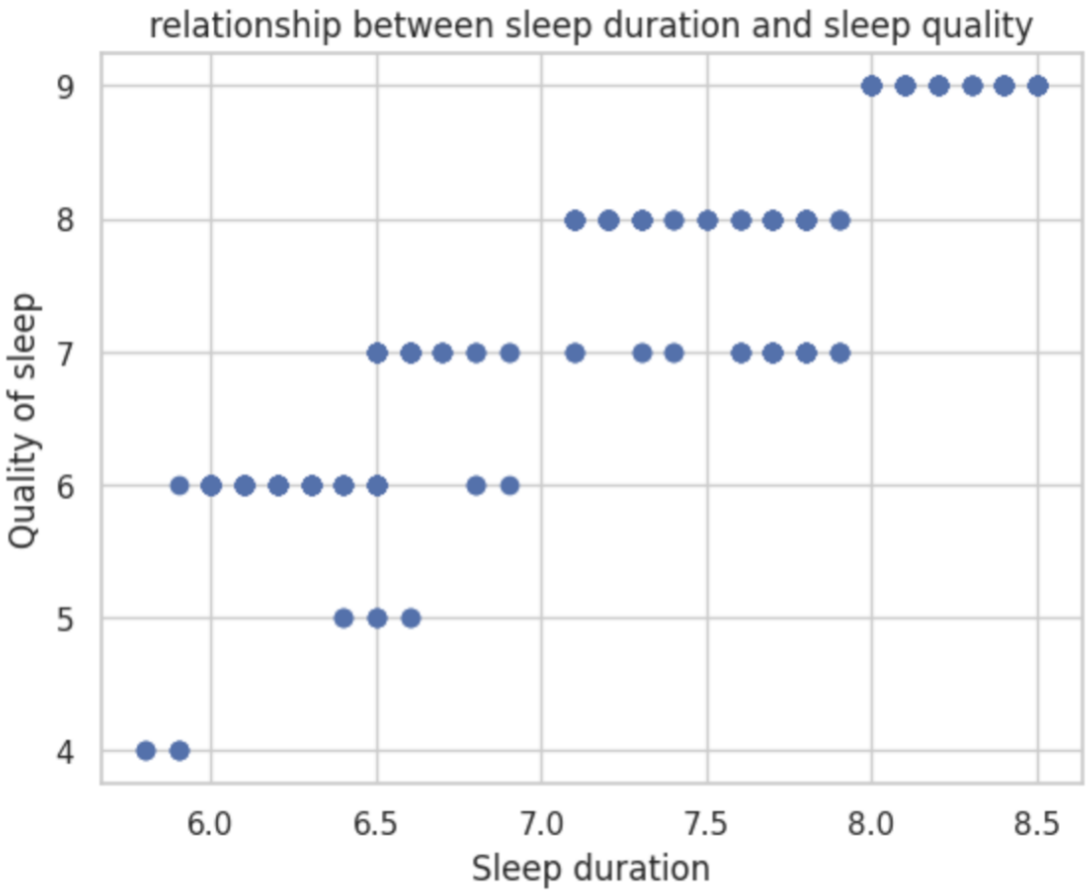
This histogram is useful for understanding the age distribution of a population or sample, indicating where the majority of individuals fall within age brackets, which can be critical for planning in areas like healthcare.

Scatter plot: Exploring the relationship between stress levels and sleep duration.



- Lower Stress Levels (3 to 4): At lower stress levels, there's a wide variation in sleep durations, ranging from about 6.5 to 8.5 hours. Generally, these levels show the potential for longer sleep durations.
- Moderate Stress Levels (5 to 6): In these mid-range stress levels, the spread of sleep durations is slightly narrower but still shows a considerable range. Most data points are clustered around 7 to 8 hours.
- Higher Stress Levels (7 to 8): At the highest stress levels shown, there is a notable trend towards shorter sleep durations. The scatter of points is denser at lower sleep durations, mostly between 6 and 7 hours, indicating that higher stress might be associated with reduced sleep.

As the stress level increases from 3 to 8, there is a general decrease in sleep duration. Most individuals with lower stress levels (3 and 4) have higher sleep durations.



- The plot indicates a clear trend where increased sleep duration is associated with higher sleep quality.
- Individuals who sleep for 7 to 8 hours or more tend to report the highest quality of sleep.
- Sleep quality ratings improve significantly as sleep duration increases beyond 7 hours.

Machine Learning Techniques

Linear Regression

A linear regression model was built to analyze the relationship between sleep duration and various factors such as age, gender, stress level, and physical activity level. The model had an adjusted R-squared value of 0.744, indicating a good fit.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("Sleep Duration")   R-squared:                       0.747
Model:                             OLS   Adj. R-squared:                  0.744
Method:                  Least Squares   F-statistic:                     272.4
Date:                 Fri, 03 May 2024   Prob (F-statistic):          1.02e-108
Time:                         16:59:07   Log-Likelihood:                -187.66
No. Observations:                  374   AIC:                             385.3
Df Residuals:                      369   BIC:                             404.9
Df Model:                            4
Covariance Type:             nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        8.1439      0.177     45.928      0.000       7.795       8.493
C(Gender)[T.Male]                0.4740      0.053      8.886      0.000       0.369       0.579
Age                              0.0116      0.003      3.657      0.000       0.005       0.018
Q("Stress Level")               -0.3902      0.013    -29.526      0.000      -0.416      -0.364
Q("Physical Activity Level")     0.0061      0.001      5.939      0.000       0.004       0.008
==============================================================================
Omnibus:                        2.367   Durbin-Watson:                   0.484
Prob(Omnibus):                  0.306   Jarque-Bera (JB):                2.368
Skew:                           0.150   Prob(JB):                        0.306
Kurtosis:                       2.752   Cond. No.                         647.
==============================================================================

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

Key findings:
- Stress level had a significant negative impact on sleep duration, with higher stress levels associated with shorter sleep durations.
- Physical activity level had a positive impact on sleep duration, with higher activity levels associated with longer sleep durations.
- Age and gender also had significant effects on sleep duration.

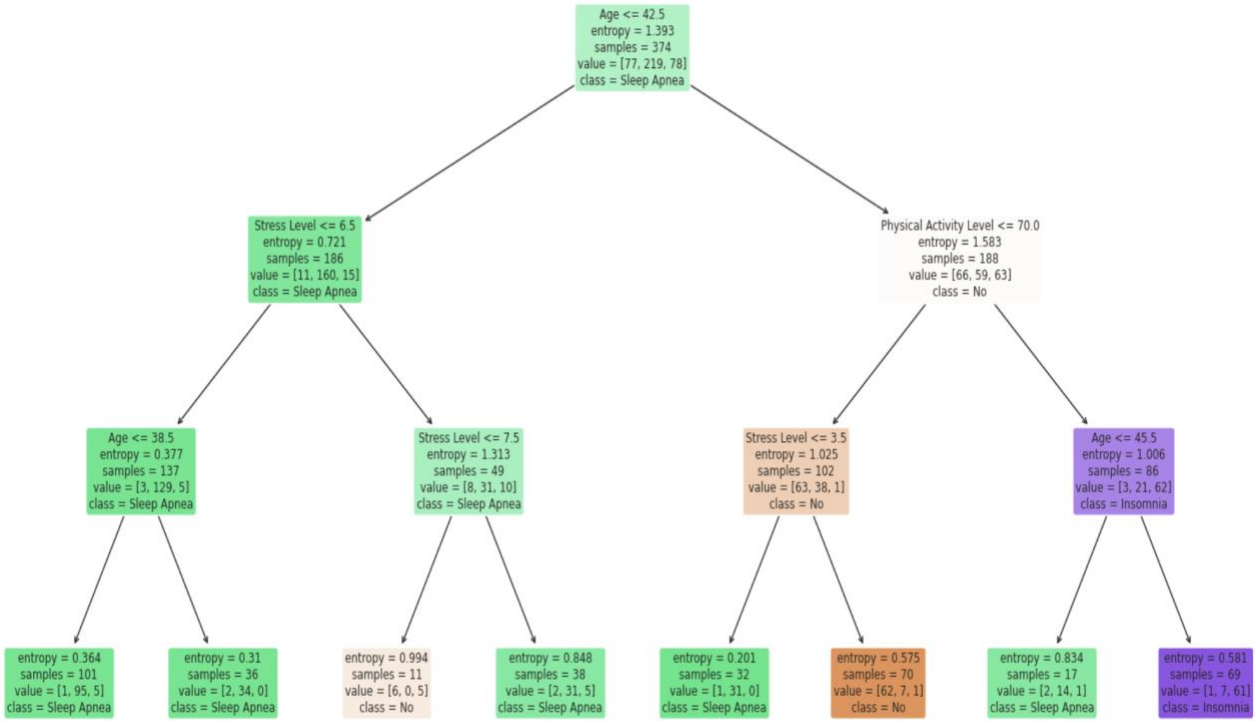
- Intercept (8.1439): The expected sleep duration when all other predictor variables are zero.
- Gender (Male: 0.4740): Being male is associated with a statistically significant increase in sleep duration by about 0.47 hours compared to females, holding all other variables constant.
- Age (0.0116): For each additional year of age, sleep duration increases by about 0.012 hours, suggesting a slight positive relationship between age and sleep duration.

- Stress Level (-0.3902): Each unit increase in stress level is associated with a decrease in sleep duration by about 0.39 hours, indicating a significant negative impact of stress on sleep duration.
- Physical Activity Level (0.0061): Each unit increase in physical activity level corresponds to an increase in sleep duration by about 0.006 hours, showing a positive but very modest effect.
- R-squared: 0.747 - This indicates that about 74.7% of the variance in sleep duration is explained by the independent variables included in the model. This is a relatively high value, suggesting a good fit of the model.

In summary, the model indicates significant relationships between sleep duration and the predictors gender, age, stress level, and physical activity, with stress level showing the strongest negative influence on sleep duration.

Decision Tree Classification

A decision tree classifier was trained to predict sleep disorders based on age, physical activity level, and stress level. The model achieved good accuracy in classifying individuals into three categories: no sleep disorder, sleep apnea, and insomnia.



Key findings:

- The decision tree identified stress level as the most important factor for predicting sleep disorders, with higher stress levels increasing the likelihood of sleep disorders.
- Physical activity level and age were also influential factors in determining sleep disorder risk.

Decision rules/Predictions: Individuals with a stress level of ≤ 6.5 and age ≤ 38.5 are classified as having "Sleep Apnea" with relatively high certainty. In contrast, individuals with very low-stress levels (≤ 3.5) generally do not have sleep apnea or insomnia unless other factors (like age > 45.5) come into play.

Recommendations

Based on the analysis, the following recommendations can be made:

1. Implement stress management programs to help individuals reduce their stress levels, as high stress was found to be detrimental to sleep duration and a significant predictor of sleep disorders.

2. Encourage regular physical activity, as higher activity levels were associated with longer sleep durations and may help mitigate the risk of sleep disorders.

3. Provide targeted interventions for individuals in high-risk groups, such as those with high-stress levels, low physical activity, or specific age groups, to address their sleep health needs.

4. Promote awareness about the importance of maintaining a healthy lifestyle, including adequate sleep, regular exercise, and stress management, to improve overall well-being.

Summary

This project aimed to conduct descriptive analytics on the "Sleep Health and Lifestyle Dataset" obtained from Kaggle. The dataset contained information about various factors related to sleep health and lifestyle habits, including sleep duration, quality of sleep, physical activity levels, stress levels, BMI categories, and more.

Through exploratory data analysis and visualizations, we identified key variables and explored their distributions and relationships. Linear regression and decision tree classification techniques were applied to analyze the impact of factors such as age, gender, stress level, and physical activity level on sleep duration and sleep disorders.

The linear regression model revealed that stress level had a significant negative impact on sleep duration, while physical activity level had a positive impact. Age and gender also played a role in determining sleep duration. The decision tree classifier identified stress level as the most important factor for predicting sleep disorders, followed by physical activity level and age.

Based on these findings, recommendations were made to implement stress management programs, encourage regular physical activity, provide targeted interventions for high-risk groups, and promote awareness about the importance of maintaining a healthy lifestyle for better sleep health.

Overall, this project demonstrated the value of data analysis and machine learning techniques in understanding the complex relationships between various lifestyle factors and sleep health. The insights gained can inform strategies and interventions to improve sleep quality and overall well-being.