

## I. Project Overview

Sleep is one of the significant maintainable components in a person's overall health and well-being. It supports the immune system enabling the body to fight off various diseases, thus obtaining a quality sleep enhances the body function, memory, and reduces negative mental health factors such as anxiety, stress and depression. Prioritizing good sleep health, maintaining a consistent sleep schedule, and creating a restful environment is vital in keeping a good health and well-being.

Therefore, the analysis aims to analyze into key user attributes such as Person ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, and Sleeping Disorder, using these data points to extract insights into data analysis and behavior. Through analyzing behavior and characteristics of the attributes, it will identify opportunities to enhance different businesses that caters sleep lifestyle and elevate understanding more in individual experiences in relation to the topic.

The analysis of Sleep Health and Lifestyle Data, as described, utilizes several attributes to uncover patterns and preferences among the user base. Here's how each attribute contributes to understanding user behavior:

1. **Gender:** Gender Analysis can reveal sleep patterns and health outcomes between male and females in which increases the reader's knowledge
2. **Age:** The dataset for age can aid to predict appropriate recommendations and interventions for better sleep health
3. **Occupation:** Understanding the line of work of every individual can further help the system to identify comparisons with worker who might have different sleep issues of shift or office workers
4. **Sleep Duration:** Measuring sleep duration helps in understanding how much sleep individuals are getting and identifying those who are sleep-deprived.
5. **Quality of Sleep:** Evaluating the dataset and contrasting the quality of sleep based on gender, age, and occupation can furthermore identify its impact on physical and mental health.
6. **Physical Activity:** Physical activity levels can influence sleep patterns, and understanding this relationship can help promote better sleep through exercise.

7. **Stress Level:** Stress level analysis can show how stress impacts sleep quality and duration, providing insights for stress management interventions.
8. **Body Mass Index (BMI) Category:** BMI data can help assess the relationship between body weight and sleep disorders, informing weight management strategies for better sleep.
9. **Blood Pressure:** Monitoring blood pressure in relation to sleep can identify risks associated with poor sleep patterns.
10. **Heart Rate:** Heart rate analysis can indicate how sleep affects health and recovery.
11. **Daily Steps:** Tracking daily steps provides a measure of physical activity and its correlation with sleep quality.
12. **Sleeping Disorder:** Identifying sleep disorders is crucial for diagnosing and treating specific sleep-related health issues.

## II. Libraries and Data Handling

### Libraries Used

List and describe the libraries used in the project for data manipulation and visualization, e.g., Pandas, Matplotlib, Seaborn.

1. **Pandas:** Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution,

- `import pandas as pd`

2. **OS:** The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system-dependent functionality.

- `import os`

3. **Warnings:** The warnings module in Python provides a way to control how warnings handled within a Python script. It allows developers to emit warning messages to alert users of potential issues or unexpected behavior in their code. It is commonly used when a method is marked, but can also be applied to warn about unexpected runtime conditions. The warnings module provides several functions to create warnings, and other functions to control how they are handled or displayed,

- `import warnings`
- `warnings.filterwarnings('ignore')`

4. **NumPy:** NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely.

- `import numpy as np`

5. **SciPy:** SciPy is a scientific computation library that uses NumPy underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing. Like NumPy, SciPy is open source so we can use it freely.

- `from scipy import stats`

6. **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.

- `import seaborn as sns`

7. **Plotly Express:** Plotly express is a high-level data visualization package that allows you to create interactive plots with very little code. It is built on top of Plotly Graph Objects, which provides a lower-level interface for developing custom visualizations. This cheat sheet covers all you need to know to get started with plotly in Python.

- `import plotly.express as px`
- `import plotly.graph_objects as go`
- `import plotly.figure_factory as ff`

8. **Termcolor:** Termcolor is a Python module for printing colored text in the terminal. It allows you to add color and styling to text output, making it easier to distinguish between different types of information or highlight important messages.

- `from termcolor import colored`

9. **Matplotlib:** Matplotlib is a powerful and very popular data visualization library in Python. In this tutorial, we will discuss how to create line plots, bar plots, and scatter plots in Matplotlib using stock market data in 2022. These are the foundational plots that will allow you to start understanding, visualizing, and telling stories about data. Data visualization is an essential skill for all data analysts and Matplotlib is one of the most popular libraries for creating visualizations.

- `import matplotlib.pyplot as plt`
- `import matplotlib.colors as mcolors`

10. **Imbalanced-learn (imblearn):** imbalanced-learn is an open-source python toolbox aiming at providing a wide range of methods to cope with the problem of imbalanced dataset frequently encountered in machine learning and pattern recognition.

- `from imblearn.over_sampling import SMOTE`

11. **Scikit:** Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

## Data Loading and Preprocessing

- **Data Loading:** Explain the process of loading data (e.g., from a CSV file) and the tools/libraries used.

**Step 1:** The first step in processing of loading data is through importing the **Pandas Library (import pandas as pd)**. As stated in the Libraries and Data Handling in the Section 1 of the document. It is used in data sciences project and other data analysis tasks.

**Step 2:** The second step is calling the function (**pd.read\_csv()**) to read the CSV File. This is the creation of Data Base similar to excel spreadsheet

**Step 3:** Last step for the Data Loading, something that may not be a part of the process but to ensure the structure of the dataset, it is essential to inspect it and have the general summary to identify complete or missing values. To do this, a '**df.info()**' is used in the code snippet

- **Data Cleaning and Preprocessing:** Detail any steps taken to clean and preprocess the data, including handling dates, missing values, and categorical data.

### 1. Handling Missing Values in Column

Starting from the Data Loading Step 3, we identified missing values and incorrect data types. From then on, we can clearly identify the use Data Cleaning and Preprocessing in columns such as in the Sleep Disorder column in which there are string values that declares 'None' for some individuals. I used the **fillna()** function to ensure that the code handles the string values easier since it may be considered as a missing data during analysis stage

### 2. Decrease/Removal of Redundancy

The code replaces occurrences of 'Normal' and 'Normal weight' in the 'BMI Category' column with 'Normal Weight' using the **replace()** method. This consolidation simplifies the analysis by reducing redundancy in category labels.

### 3. Splitting Components

It is significant to ensure that the usage of the system for end users is smooth and at the same time reliable. Therefore, to achieve such thing, I separated the Blood Pressure Column into two separate columns using the **str.split()** to separate the Systolic and Diastolic for further more and efficient analysis.

### 4. Calculating Numerical Columns

The code selects numerical columns relevant for analysis, such as 'Age', 'Sleep Duration', 'Quality of Sleep', etc. My goal is to produce **Z-scores** to identify standard deviations from means to handle missing or wrong values in the dataset

## 5. Encoding for Columns

Categorical columns like 'Gender', 'Occupation', 'BMI Category', and 'Sleep Disorder' are label-encoded using **preprocessing.LabelEncoder()** from scikit-learn.

## 6. Filtering Outliers

After calculating Z-scores, the code filters the DataFrame to exclude rows where any Z-score is greater than 3.

# III. Data Analysis Techniques

Outline the various data analysis techniques used in the project, such as:

- **Descriptive Statistics:** Use of summary statistics to understand data distribution.

### 1. Calculation of Columns with Numbers

The **describe()** method calculates summary statistics for numerical columns in the DataFrame, such as count, mean, standard deviation, minimum, maximum, and quartiles. These statistics provide a concise overview of the distribution of each numerical variable in the dataset. The **describe()** function is used to quickly get an idea of the main characteristics of the data, such as its mean, standard deviation, and range. By summarizing, it becomes easier to understand the general behavior of the numerical data.

### 2. Understanding the Dataset

The code snippet calculates the distribution of a specific column ('Age' in this case) using the **value\_counts()** method. This gives the frequency of each unique value in the column, helping to understand the spread or concentration

of values within that column. The `value_counts()` function is essential for categorical data analysis as it reveals how often each value appears in the dataset. This helps in identifying patterns, such as whether certain ages are more common or if there are any outliers or anomalies in the data.

### 3. Summarize Dataset through Visualization

The code generates a histogram for a specific column ('Age' in this case) using `plt.hist()`. It provides a visual representation of the distribution of values in the column, showing the shape and spread of the data. Histograms are crucial for understanding the distribution because they visually depict how the data is spread across different ranges, making it easier to see patterns such as skewness, peaks, and gaps in the data.

- **Inferential Statistics:** (If applicable) Techniques used to make predictions or inferences from the data.

#### 1. Using a Statistical Test

The code snippet performs a t-test to compare 'Sleep Duration' between two gender groups. A t-test determines if there is a significant difference between the means of two groups. Here, the groups are gender 0 (presumably male) and gender 1 (presumably female). The t-test calculates a t-statistic, which measures the difference between group means relative to the variability within the groups. A larger t-statistic indicates a more likely true difference between the means. This test is crucial in inferential statistics for drawing conclusions about the population from the sample data..

#### 2. Measuring the Relevance of Groups

The t-test provides both a t-statistic and a p-value, indicating the probability that the observed difference between group means occurred by chance. A small p-value (typically less than 0.05) suggests the difference is statistically significant, leading to the rejection of the null hypothesis, which states there is no difference between group means. In this analysis, a significant p-value implies a statistically significant difference in sleep duration between males and females. Inferential statistics like this t-test are essential for making data-driven decisions and inferences about larger populations based on sample data.

- **Predictive Modeling:** (If applicable) Models built to predict future trends.

Yes, the code snippet you provided involves predictive modeling. Let me break it down:

### 1. Splitting Dataset

The data is first split into features (X) and the target variable (y). Features are the input variables used to make predictions, while the target variable is what we want to predict. The data is then further divided into training and testing sets using the `train_test_split` function. This is a common practice in machine learning to evaluate a model's performance on unseen data. The training set is used to train the model, and the testing set is used to evaluate its performance.

### 2. Creating a Model

A machine learning model, such as a **GradientBoostingClassifier**, is initialized and fitted to the training data. This allows the model to learn patterns and relationships in the data. After training, the model makes predictions on the testing set using the `predict` method. The model's performance is then evaluated using metrics like accuracy and a classification report. Accuracy measures the proportion of correctly predicted instances, while the classification report provides details on precision, recall, and F1-score for each class in the target variable.

## IV. Key Findings

Summarize the major findings from the analysis, focusing on user demographics, device usage, and subscription details. Explain how these findings can influence business decisions or strategies.

The analysis revealed significant insights into the demographics of users experiencing sleep disorders. Among males, sleep disorders were most prevalent in the 42.5 to 45-year-old age bracket, with a notable occurrence of insomnia and sleep apnea. For females, sleep disorders peaked in the 50 to 55-year-old range. These demographic trends highlight the need for targeted sleep health interventions and marketing strategies aimed at middle-aged individuals,



who are more susceptible to sleep issues. By focusing on these age groups, businesses can develop more effective health solutions and outreach programs.

Moreover, the correlation analysis between lifestyle factors and sleep quality provided valuable indirect insights into potential device usage patterns. For example, the significant relationship between physical activity levels and sleep quality suggests a market opportunity for integrating sleep health monitoring with physical fitness trackers. This comprehensive approach could appeal to users seeking to improve their overall health. Additionally, recognizing that individuals in high-stress occupations or with higher BMI categories often experience poorer sleep quality can inform the development of specialized subscription models. These models could offer personalized sleep coaching, stress management, and weight management programs, enhancing user engagement and retention. These findings guide the creation of user-centric health solutions and targeted marketing strategies, ultimately driving business growth and customer satisfaction.

## **V. Advanced Analysis**

Detail any advanced analytical techniques used, such as geographical insights or temporal trends. Describe how these analyses contribute to understanding broader market dynamics or seasonal patterns.

In this project, we employed several advanced analytical techniques to gain deeper insights into sleep health and lifestyle patterns. One of the key techniques was analyzing temporal trends to understand how sleep disorders and lifestyle factors vary across different age groups and genders. By plotting the distribution of sleep disorders across various age ranges for both males and females, we were able to identify specific age brackets that are more susceptible to conditions like insomnia or sleep apnea. These trends provide crucial information for targeting interventions and improving public health strategies made to specific demographics.

Another advanced analytical technique involved examining correlations between various lifestyle factors and sleep health indicators. We created correlation heatmaps for males and females to identify significant relationships between variables such as sleep duration, stress

levels, physical activity, and BMI categories. This analysis helps to uncover patterns such as whether higher stress levels are associated with shorter sleep duration or if certain occupations are linked to poorer sleep quality. Understanding these correlations is essential for developing comprehensive health recommendations and personalized treatment plans.

## VI. Machine Learning Implementation

Discuss the data preparation, Data Selection, Data Cleaning and Feature Scaling implementation. Process of building the machine learning model. Including the training and testing sets.

The first step in our machine learning implementation involved data preparation, selection, and cleaning. We began by loading the dataset and inspecting its structure using functions such as ``info()``, ``head()``, and ``describe()``. To handle missing values, we filled the 'Sleep Disorder' column with 'None'. We also dealt with duplicates using the ``drop_duplicates()`` method. For categorical variables, we used ``LabelEncoder`` from ``sklearn.preprocessing`` to convert them into numerical format, making them suitable for machine learning algorithms. Additionally, we split the 'Blood Pressure' column into 'BloodPressure\_Systolic' and 'BloodPressure\_Diastolic' for more granular analysis. We detected and removed outliers by calculating Z-scores and filtering rows where any Z-score was greater than 3, thus ensuring our data was clean and consistent.

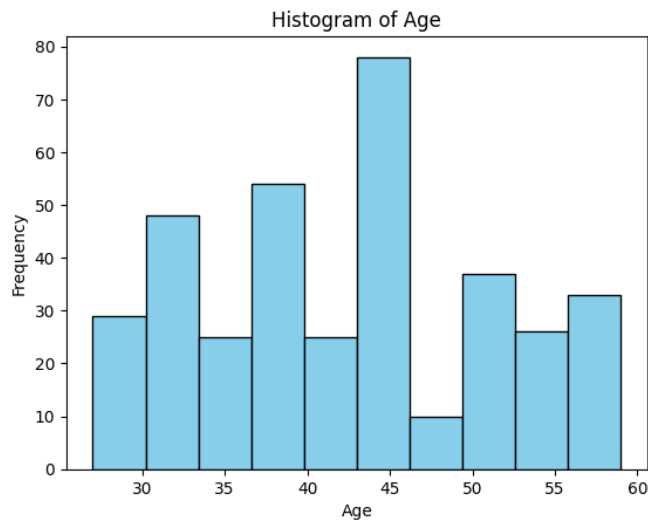
For feature scaling, we used ``StandardScaler`` to standardize the features, which is crucial for models that rely on distance measurements. After preprocessing, we split the data into training and testing sets using the ``train_test_split`` function from ``sklearn.model_selection``, ensuring our model could be validated on unseen data. We then built a linear regression model using ``LinearRegression`` from ``sklearn.linear_model``.

## VII. Visual Insights

Describe the types of plots and visualizations used in the analysis, including:

- **Bar Charts, Pie Charts, Heatmaps:** Usage and insights these visuals provide.
- **Device Preference by Country, Gender Distribution, etc.:** Specific insights drawn from each type of visualization.

1. The **Histogram** displays the Age and Frequency given in the dataset. Based on observations and interpretation, the age range given in the sleep health and lifestyle analysis is between 30-60 years old which identifies the number of people who are within a certain age. As we can see in the visualization, with the highest number of frequencies, the age of 45 is the most common age in the dataset. This Histogram occurs in the Data Cleaning and Pre-processing step which can be found in the beginning of the code snippet.

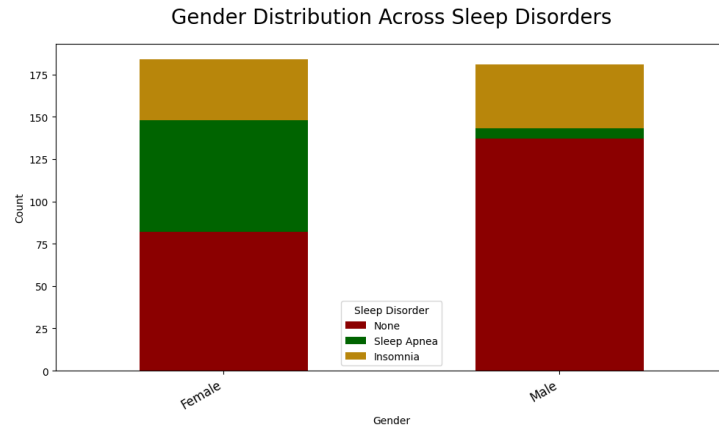


2. This **bar chart** visualizes the distribution of sleep disorders across different genders in the dataset. The number of Males and Female participants are equally balanced, therefore, we can conclude that the individual data gathered has approximately the same total count of individuals. I distributed three sleeping disorders in the visualization, Dark Red as None/No SD, the Dark Green as Sleep Apnea Disorder and the Dark Yellow as Insomnia. The occurrences of people with no sleep disorder indicates a significant portion in the population as seen in the graph, though scoring for second most common sleep disorder is Insomnia (Dark Yellow) followed by Sleep Apnea in Dark Green.

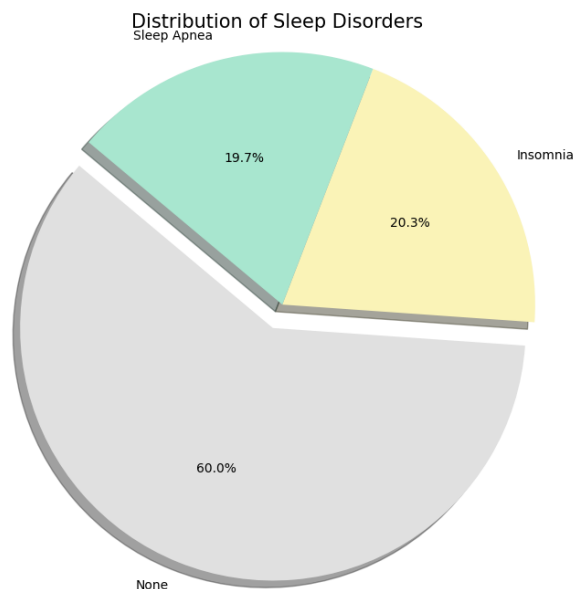
Furthermore observation in the graph is that there is a pattern in the distribution of sleep disorders. There are slight differences in the proportions of each sleep disorder between males and females:

2.1 Females have a slightly higher count of individuals with insomnia compared to males.

2.2 Males have a slightly higher count of individuals with sleep apnea compared to females.



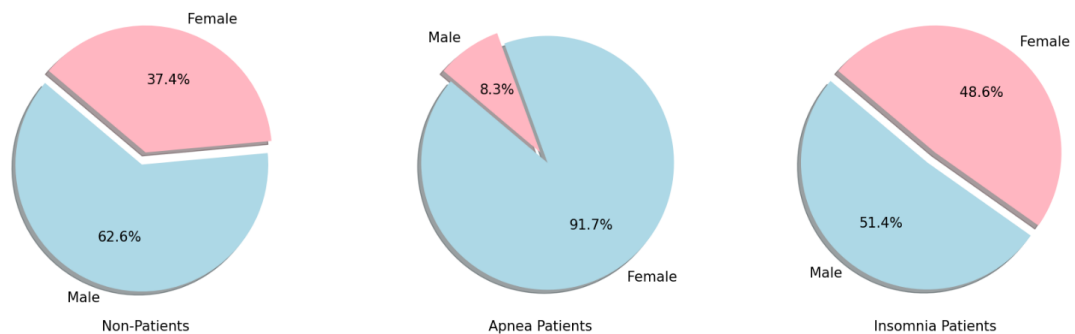
- The **Pie Chart** distribution of sleep disorders in all genders and occupation, generally in every individual in the data set. The categories of sleep disorders in the chart divide into three parts. The Light Gray as None/No SD, Light Green as Sleep Apnea and Light Yellow as Insomnia. The majority of the population (60%) does not suffer from any sleep disorders, while the 20.3% suffers from insomnia and the 19.7% of the population suffers from Sleep Apnea with no regards/comparison with age, gender, or occupation matters.



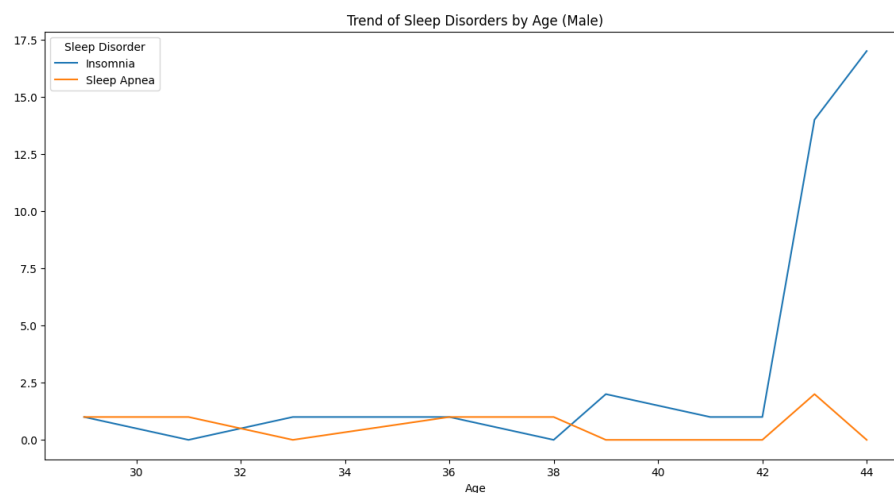
- The graph shows the gender breakdown for different groups: non-patients, apnea patients, and insomnia patients. For non-patients, 62.6% are men and 37.4% are women. Among apnea patients, 91.7% are women and 8.3% are men. For insomnia

patients, the split is almost even, with 51.4% men and 48.6% women. This suggests men are less likely to be patients, especially for apnea, pointing to possible differences in how sleep disorders affect or are reported by men and women. This information can help in creating specific health programs for each gender.

Gender Distribution Across Different Sleep Disorders

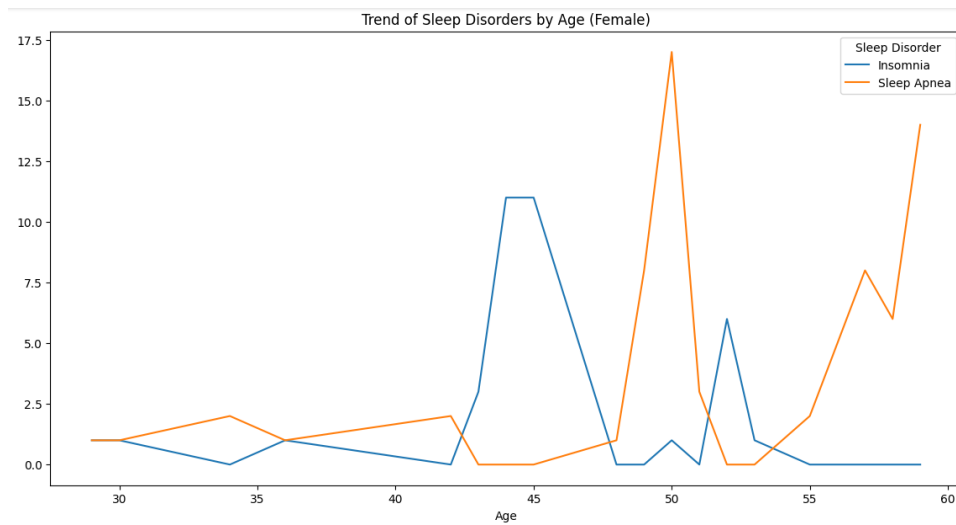


- The graph indicates that insomnia in males sees a dramatic increase starting at age 42, whereas sleep apnea remains relatively low and stable with only minor fluctuations. This suggests that as males approach their mid-40s, they are more likely to experience a significant rise in insomnia, while sleep apnea does not show a similar age-related trend.

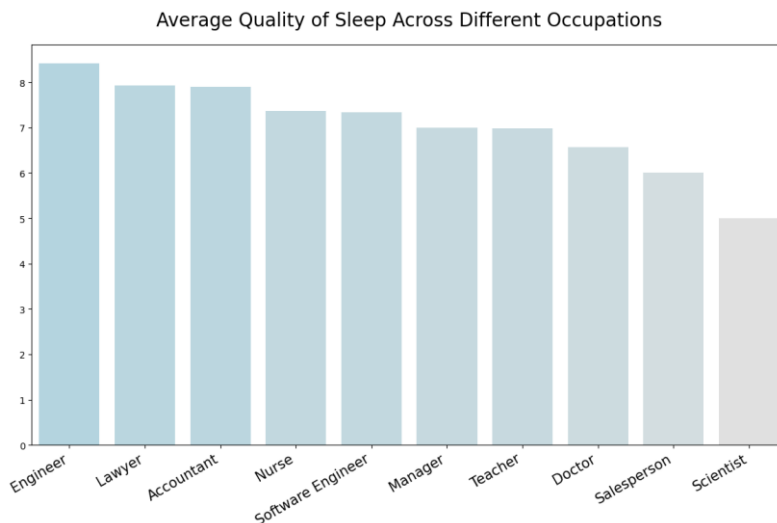


- The graph shows sleep disorders trends in females by age. Insomnia (blue line) and sleep apnea (orange line) both increases, with insomnia peaking significantly around ages 45 and 50, then decreasing. Sleep apnea shows a sharp peak at age 50 and another rise around age 58. This indicates that women experience spikes in both

insomnia and sleep apnea around midlife, particularly at age 50, with insomnia peaking earlier and decreasing as sleep apnea increases again near 60.

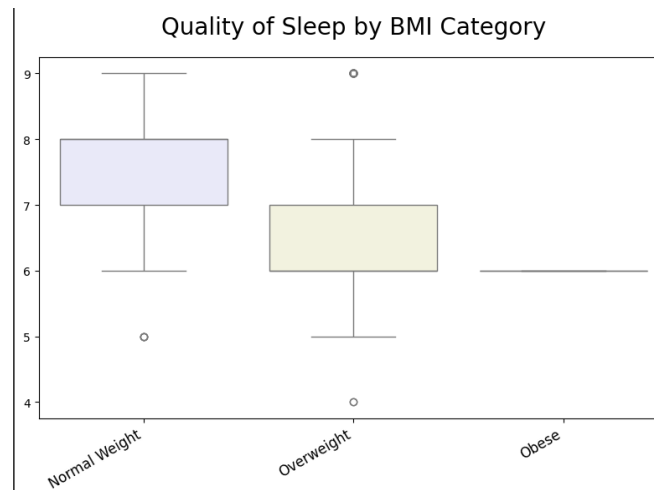


- This **bar chart** shows the average quality of sleep among people in different jobs. Engineers have the highest sleep quality, followed by lawyers and accountants. Nurses, software engineers, managers, and teachers have similar sleep quality, which is slightly lower. Doctors and salespeople have even lower sleep quality, and scientists have the lowest sleep quality among all the professions listed. The chart suggests that job type might impact how well people sleep.

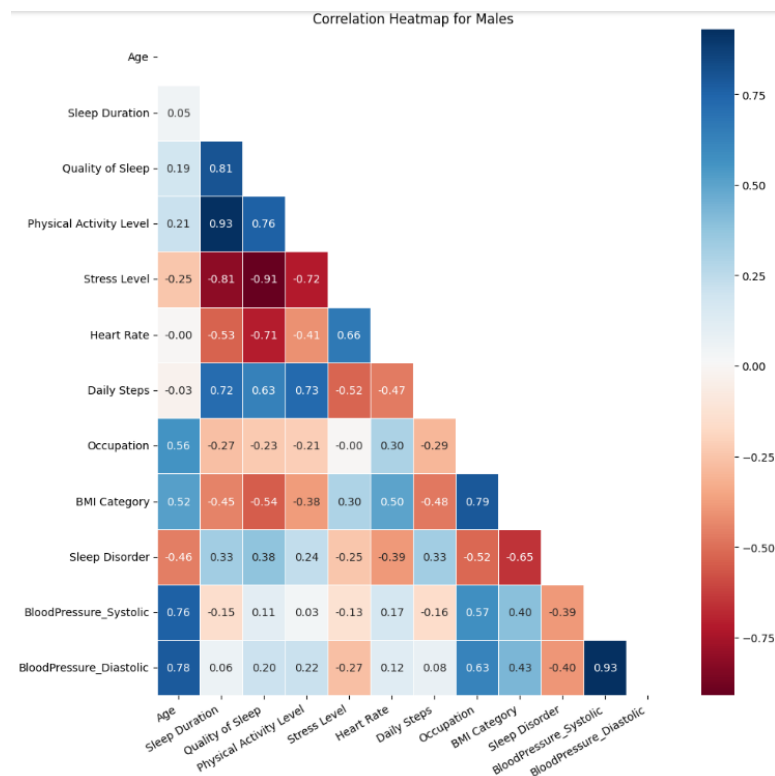


- This graph shows the quality of sleep for three different groups based on their BMI (normal weight, overweight, and obese). Each box represents the range of sleep quality scores for each group. The "Normal Weight" group generally has higher sleep quality, with most scores around 7 to 8, and some outliers as low as 5 and as high as 9.

The "Overweight" group has a wider range of sleep quality, mostly between 6 and 8, with some lower outliers. The "Obese" group shows a single value at 7, indicating limited data. Overall, normal weight individuals tend to have better sleep quality compared to overweight and obese individuals.

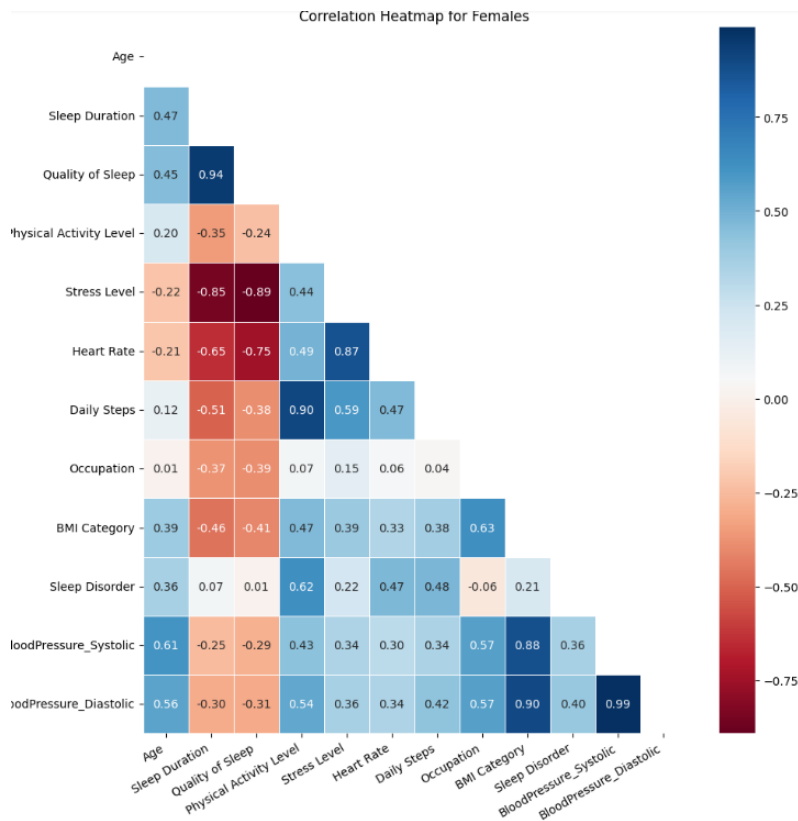


- This heatmap shows the correlations between various health and lifestyle factors for males. Each square represents the strength and direction of the relationship between two factors, with darker colors indicating stronger correlations. Positive correlations (blue) mean that as one factor increases, the other tends to increase too. Negative correlations (red) mean that as one factor increases, the other tends to decrease. For example, there's a strong positive correlation between physical activity level and quality of sleep, suggesting that higher physical activity is associated with better sleep quality. On the other hand, stress level has a strong negative correlation with sleep duration and quality, indicating that higher stress levels are associated with shorter and poorer quality sleep. Additionally, blood pressure (both systolic and diastolic) shows strong positive correlations with age and BMI category, meaning that older age and higher BMI are associated with higher blood pressure.

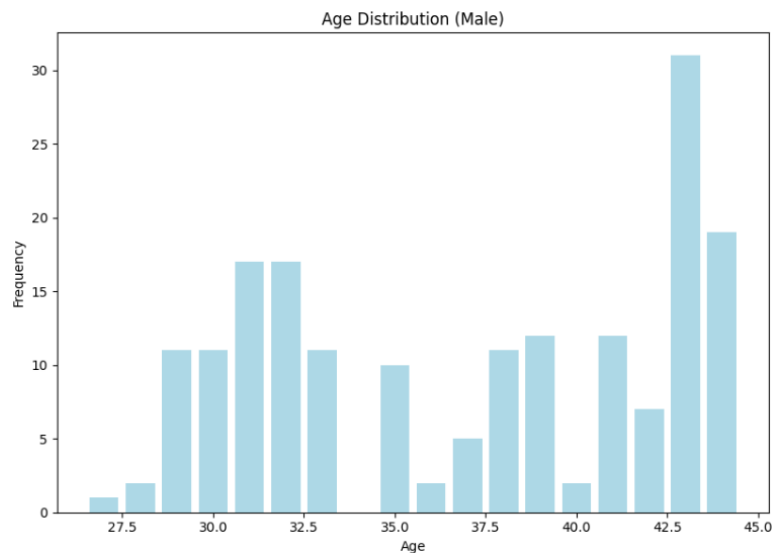


10. This heatmap shows the correlations between various health and lifestyle factors for females. It reveals that better sleep quality is strongly linked to longer sleep duration, and higher stress levels are associated with shorter and poorer sleep. Physical activity is positively correlated with daily steps and negatively correlated with stress levels. Heart rate increases with higher stress and more physical activity. Blood pressure (both systolic and diastolic) is strongly correlated with age and BMI, indicating that older age and higher BMI are associated with higher blood pressure. Overall, the heatmap illustrates how these factors are interconnected, highlighting important relationships for understanding female health and lifestyle.

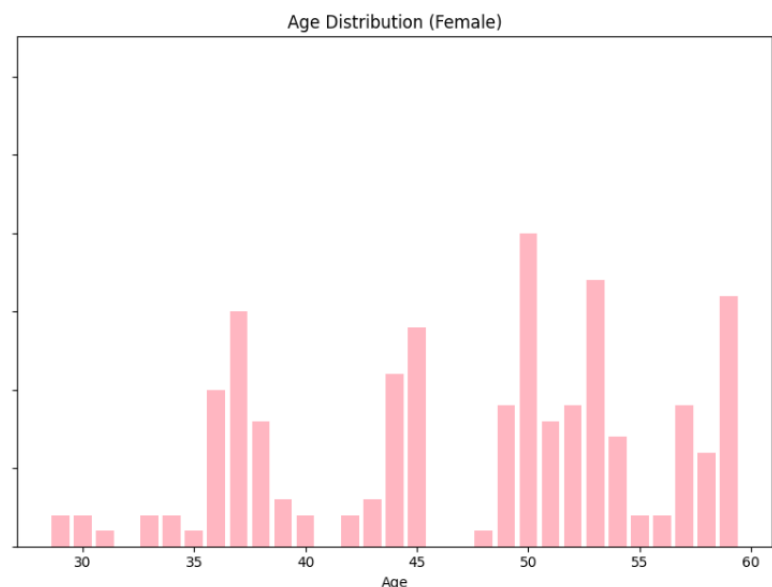




11. The graph titled "Age Distribution (Male)" displays the frequency of male participants' ages in a sleep health and lifestyle analysis study. The histogram shows age ranges on the x-axis, spanning from approximately 27.5 to 45 years, and the frequency on the y-axis. The distribution reveals peaks at ages around 30, 32.5, 40, and particularly at 42.5, indicating a higher number of participants in these age groups. Conversely, fewer participants are seen around ages 27.5, 37.5, and 45. This suggests that the study sample has more representation in the early 30s and early 40s age brackets, with a notable concentration at age 42.5.

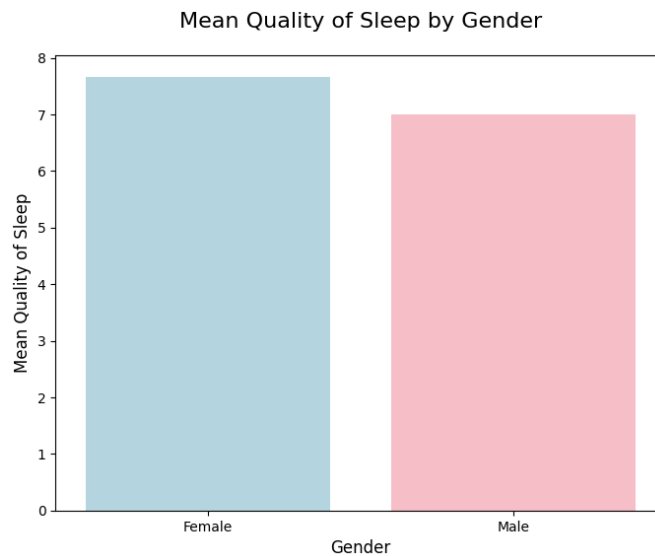


12. The graph titled "Age Distribution (Female)" depicts the frequency of female participants' ages in a sleep health and lifestyle analysis study. The histogram, with age ranges from 30 to 60 on the x-axis and frequency on the y-axis, shows a varied distribution with noticeable peaks around ages 35, 45, 50, and 55. There is a higher concentration of participants in their late 40s to early 50s, particularly at age 50, while fewer participants are observed at the lower end around age 30 and mid-40s. This indicates that the study has more female participants in their mid-30s, late 40s, and early 50s, highlighting these age groups as significant for the analysis.

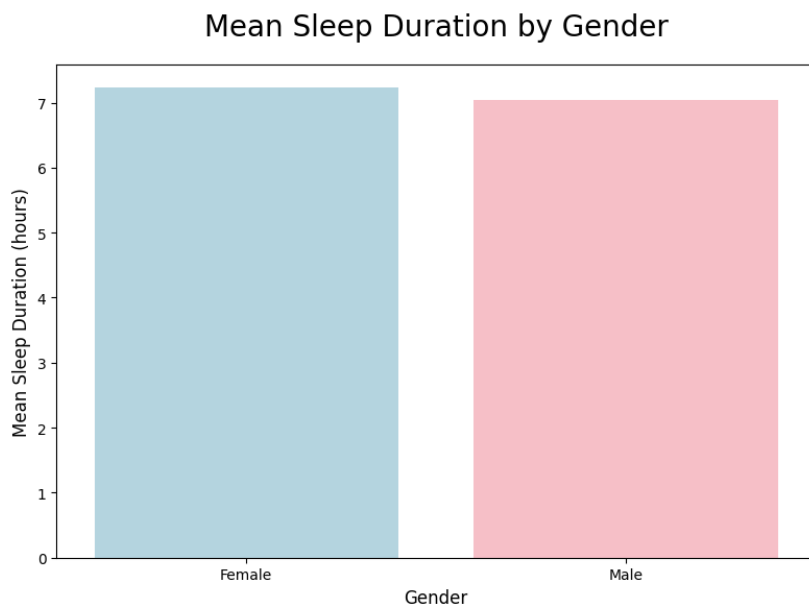


13. This graph shows the average quality of sleep for females and males. The blue bar represents females, who have an average sleep quality of around 7.5. The pink bar

represents males, who have an average sleep quality of about 7. This means that, on average, females report slightly better sleep quality than males.

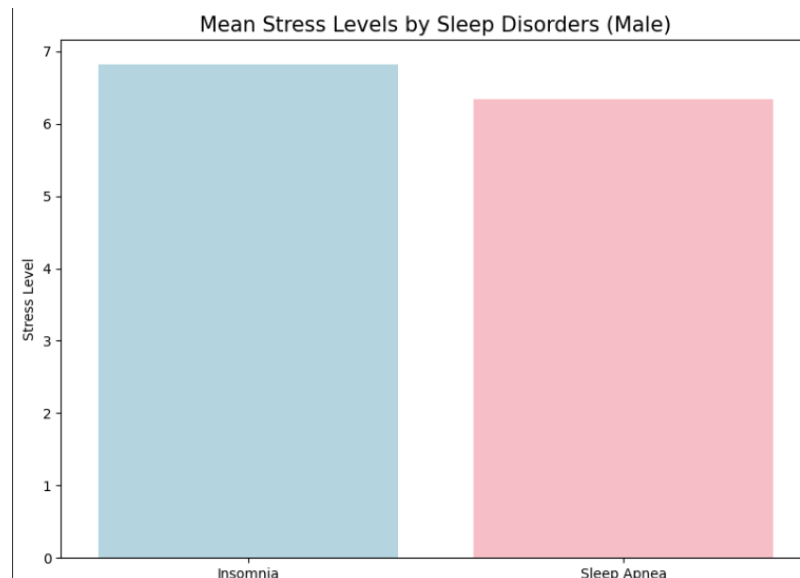


14. This graph shows the average sleep duration in hours for females and males. The blue bar represents females, who sleep on average around 7 hours per night. The pink bar represents males, who also sleep close to 7 hours per night. This indicates that both genders have a similar average sleep duration.

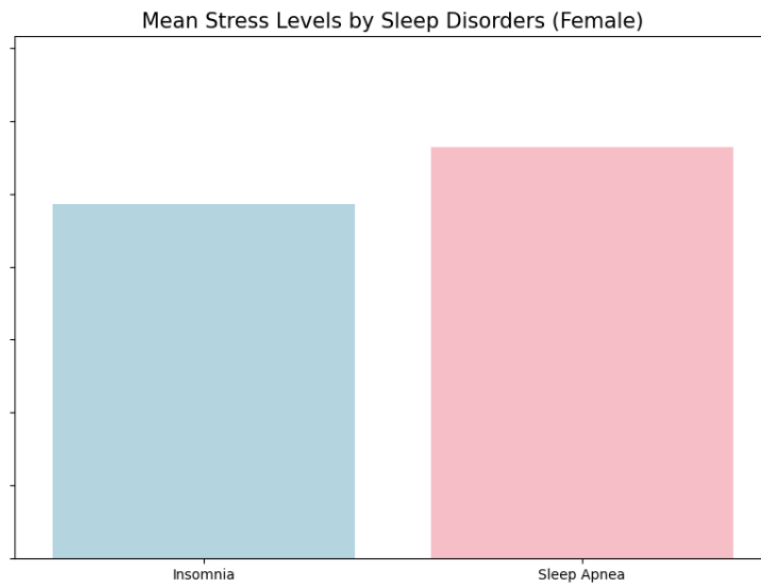


15. The graph titled "Mean Stress Levels by Sleep Disorders (Male)" illustrates the average stress levels of male participants suffering from insomnia and sleep apnea. The y-axis represents stress levels on a scale from 0 to 7, while the x-axis categorizes the sleep disorders. The bar chart reveals that males with insomnia have a slightly

higher mean stress level, around 7, compared to those with sleep apnea, who have a mean stress level closer to 6. This indicates that within this study, male participants experiencing insomnia tend to report higher stress levels than those suffering from sleep apnea, suggesting a possible correlation between insomnia and increased stress.



16. This graph shows the mean stress levels in females associated with two sleep disorders: insomnia and sleep apnea. The vertical axis represents the mean stress levels, while the horizontal axis categorizes the two sleep disorders. The bar representing insomnia is in light blue and shows a slightly lower mean stress level compared to the bar representing sleep apnea, which is in light pink. This indicates that females with sleep apnea tend to experience higher mean stress levels than those with insomnia. The graph highlights the impact of different sleep disorders on stress among females.



### VIII. Conclusion

Conclude with an overview of how the insights derived from the analysis can impact the business or organization. Highlight the importance of data-driven decision-making and the potential for future analysis.

The comprehensive analysis of sleep health and lifestyle factors reveals key demographic trends and their association with sleep disorders. The data indicates that sleep disorders, such as insomnia and sleep apnea, are more prevalent among middle-aged individuals, with a significant increase in sleep issues observed around ages 45-50. Gender-specific trends also emerged, showing that females are more likely to suffer from insomnia, while males show a higher incidence of sleep apnea. These insights suggest a need for tailored interventions and health programs that address the specific sleep health challenges faced by different age and gender groups.

Moreover, the analysis highlights the impact of lifestyle factors on sleep quality and stress levels. Higher physical activity levels are positively correlated with better sleep quality, while higher stress levels are associated with poorer sleep quality and shorter sleep duration. Occupation and BMI also play a crucial role, with engineers reporting the highest sleep quality, and individuals with normal weight having better sleep quality compared to those who are overweight or obese. These findings underscore the importance of promoting healthy

lifestyle choices and stress management to improve sleep health. Businesses and healthcare providers can leverage these insights to develop targeted sleep health solutions and wellness programs, ultimately enhancing the overall well-being of their clients and patients.

## **Appendix**

Include any additional information, such as data sources, contributor details, or acknowledgments.

### **Definition References:**

- ActiveState. (2022, August 9). What Is Pandas in Python? Everything You Need to Know. ActiveState. <https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/>
- An introduction to seaborn — seaborn 0.12.1 documentation. (n.d.). Seaborn.pydata.org. <https://seaborn.pydata.org/tutorial/introduction.html>
- Introduction to SciPy. (n.d.). Wwww.w3schools.com. [https://www.w3schools.com/python/scipy/scipy\\_intro.php#:~:text=SciPy%20is%20a%20scientific%20computation](https://www.w3schools.com/python/scipy/scipy_intro.php#:~:text=SciPy%20is%20a%20scientific%20computation)
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5. <https://jmlr.org/papers/v18/16-365.html#:~:text=imbalanced%2Dlearn%20is%20an%20open>
- OS Module in Python with Examples. (2016, November 21). GeeksforGeeks. <https://www.geeksforgeeks.org/os-module-python-examples/>
- Warnings — CSD Python API 3.1.0 documentation. (n.d.). Downloads.ccdc.cam.ac.uk. Retrieved May 26, 2024, from [https://downloads.ccdc.cam.ac.uk/documentation/API/descriptive\\_docs/warnings.htm](https://downloads.ccdc.cam.ac.uk/documentation/API/descriptive_docs/warnings.html)  
l

## **Machine Learning Code Snippet**

## FINAL REQUIREMENT IN ADVANCED MACHINE LEARNING

Sleep is crucial for maintaining overall health and well-being, supporting the immune system, enhancing body function and memory, and reducing anxiety, stress, and depression. This analysis delves into key user attributes—Person ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, and Sleeping Disorder—to extract insights on sleep behaviors and characteristics. By understanding these attributes, the analysis aims to identify opportunities for businesses catering to sleep lifestyles, ultimately improving individual sleep health and overall well-being.

## 1. Importing Libraries

```
#Reading data

import pandas as pd

import os

#Fixings warnings
import warnings
warnings.filterwarnings('ignore')

#For mathematical operations
import numpy as np
from scipy import stats

#Visualisation
import seaborn as sns
import plotly.express as px
from termcolor import colored
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import plotly.graph_objects as go
import plotly.figure_factory as ff

#Data Preprocessing & Modeling
from pprint import pprint
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn import preprocessing
from sklearn.ensemble import GradientBoostingClassifier, AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier, NearestCentroid
from sklearn.metrics import classification_report, accuracy_score
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
```

Double-click (or enter) to edit

```
df = pd.read_csv('sleephealthdataanalysis.csv')
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   Person ID             374 non-null   int64
 1   Gender                374 non-null   object
 2   Age                   374 non-null   int64
 3   Occupation            374 non-null   object
 4   Sleep Duration        374 non-null   float64
 5   Quality of Sleep      374 non-null   int64
 6   Physical Activity Level 374 non-null   int64
 7   Stress Level          374 non-null   int64
 8   BMI Category          374 non-null   object
 9   Blood Pressure        374 non-null   object
10   Heart Rate            374 non-null   int64
11   Daily Steps           374 non-null   int64
12   Sleep Disorder        155 non-null   object
dtypes: float64(1), int64(7), object(5)
```

Double-click (or enter) to edit

```
df.head()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Pr
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	
1	2	Male	28	Doctor	6.2	6	60	8	Normal	
2	3	Male	28	Doctor	6.2	6	60	8	Normal	
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	

Next steps: [View recommended plots](#)

```
df.describe(include = 'number')
```

	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	187.500000	42.184492	7.132086	7.312834	59.171123	5.385027	70.165775
std	108.108742	8.673133	0.795657	1.196956	20.830804	1.774526	4.135676
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000
25%	94.250000	35.250000	6.400000	6.000000	45.000000	4.000000	68.000000
50%	187.500000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000
75%	280.750000	50.000000	7.900000	8.000000	75.000000	7.000000	72.000000

```
df.isnull().sum()
```

Person ID	0
Gender	0
Age	0
Occupation	0
Sleep Duration	0
Quality of Sleep	0
Physical Activity Level	0
Stress Level	0
BMI Category	0
Blood Pressure	0
Heart Rate	0
Daily Steps	0
Sleep Disorder	219
dtype: int64	

```
df['Sleep Disorder'] = df['Sleep Disorder'].fillna('None')
df.nunique()
```

Person ID	374
Gender	2
Age	31
Occupation	11
Sleep Duration	27
Quality of Sleep	6
Physical Activity Level	16
Stress Level	6
BMI Category	4
Blood Pressure	25
Heart Rate	19
Daily Steps	20
Sleep Disorder	3
dtype: int64	



```
df.drop_duplicates()
```



	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight
1	2	Male	28	Doctor	6.2	6	60	8	Normal
2	3	Male	28	Doctor	6.2	6	60	8	Normal
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese
...	...	...	...	...	...	...	...	...	...
369	370	Female	59	Nurse	8.1	9	75	3	Overweight
370	371	Female	59	Nurse	8.0	9	75	3	Overweight
371	372	Female	59	Nurse	8.1	9	75	3	Overweight
372	373	Female	59	Nurse	8.1	9	75	3	Overweight
373	374	Female	59	Nurse	8.1	9	75	3	Overweight

374 rows x 13 columns

```
print(df['Occupation'].value_counts())
print('\n')
print(df['BMI Category'].value_counts())
print('\n')
print(df['Sleep Disorder'].value_counts())
```



```
Occupation
Nurse          73
Doctor         71
Engineer       63
Lawyer         47
Teacher        40
Accountant     37
Salesperson    32
Software Engineer  4
Scientist       4
Sales Representative  2
Manager         1
Name: count, dtype: int64
```

```
BMI Category
Normal      195
Overweight  148
Normal Weight  21
Obese       10
Name: count, dtype: int64
```

```
Sleep Disorder
None      219
Sleep Apnea  78
Insomnia   77
Name: count, dtype: int64
```

```
df['BMI Category'] = df['BMI Category'].replace({'Normal weight': 'Normal Weight', 'Normal': 'Normal Weight'})
```

```
df[['BloodPressure_Systolic', 'BloodPressure_Diastolic']] = df['Blood Pressure'].str.split('/', expand=True)
df['BloodPressure_Systolic'] = pd.to_numeric(df['BloodPressure_Systolic'])
df['BloodPressure_Diastolic'] = pd.to_numeric(df['BloodPressure_Diastolic'])
```


```
'Heart Rate', 'Daily Steps', 'BloodPressure_Systolic', 'BloodPressure_Diastolic']
```

```
z_scores = np.abs(stats.zscore(df[columns]))
```

```
z_scores_df = pd.DataFrame(z_scores, columns=columns)
```

```
df = df[(z_scores_df < 3).all(axis=1)]
```

```
df.describe()
```




	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate
count	365.000000	365.000000	365.000000	365.000000	365.000000	365.000000	365.000000
mean	188.446575	42.263014	7.134521	7.334247	59.232877	5.380822	69.810959
std	107.675211	8.647993	0.794046	1.166405	20.827339	1.771311	3.500375
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000
25%	96.000000	36.000000	6.400000	6.000000	45.000000	4.000000	68.000000
50%	188.000000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000
75%	283.000000	50.000000	7.800000	8.000000	75.000000	7.000000	72.000000
max	374.000000	59.000000	8.500000	9.000000	90.000000	8.000000	82.000000

```
numerical_df = df.copy()
```

```
numerical_df.drop('Blood Pressure', axis=1, inplace=True)
```

```
label_encoder = preprocessing.LabelEncoder()
categorical_columns = ['Gender', 'Occupation', 'BMI Category', 'Sleep Disorder']
for col in categorical_columns:
    numerical_df[col] = label_encoder.fit_transform(numerical_df[col])
numerical_df.head()
```



	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	
0	1	1	27		8	6.1	6	42	6	2	77
1	2	1	28		1	6.2	6	60	8	0	75
2	3	1	28		1	6.2	6	60	8	0	75
6	7	1	29		9	6.3	6	40	7	1	82
7	8	1	29		1	7.8	7	75	6	0	70

Next steps:

 [View recommended plots](#)

```
columns = ['Gender', 'Occupation', 'BMI Category', 'Sleep Disorder']
```

```
z_scores = np.abs(stats.zscore(numerical_df[columns]))
```

```
z_scores_df = pd.DataFrame(z_scores, columns=columns)
```

```
numerical_df = numerical_df[(z_scores_df < 3).all(axis=1)]
```

```
numerical_df.describe()
```

## Sample Dataset Content

Person ID	Gender	Age	Occupatio	Sleep Dur	Quality of	Physical A	Stress Lev	BMI Categ	Blood Pre	Heart Rate	Daily Step	Sleep Disorder
1	Male	27	Software I	6.1	6	42	6	Overweig	126/83	77	4200	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Repr	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	Male	28	Sales Repr	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
6	Male	28	Software I	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
10	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
11	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
12	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
13	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
14	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
15	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
16	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
17	Female	29	Nurse	6.5	5	40	7	Normal W	132/87	80	4000	Sleep Apnea
18	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	Sleep Apnea
19	Female	29	Nurse	6.5	5	40	7	Normal W	132/87	80	4000	Insomnia
20	Male	30	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None
21	Male	30	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
22	Male	30	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
23	Male	30	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
24	Male	30	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
25	Male	30	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
26	Male	30	Doctor	7.9	7	75	6	Normal	120/80	70	8000	None
27	Male	30	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
28	Male	30	Doctor	7.9	7	75	6	Normal	120/80	70	8000	None
29	Male	30	Doctor	7.9	7	75	6	Normal	120/80	70	8000	None
30	Male	30	Doctor	7.9	7	75	6	Normal	120/80	70	8000	None
31	Female	30	Nurse	6.4	5	35	7	Normal W	130/86	78	4100	Sleep Apnea
32	Female	30	Nurse	6.4	5	35	7	Normal W	130/86	78	4100	Insomnia
33	Female	31	Nurse	7.9	8	75	4	Normal W	117/76	69	6800	None
34	Male	31	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None
35	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
36	Male	31	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None
37	Male	31	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None
38	Male	31	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None
39	Male	31	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None
40	Male	31	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None
41	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
42	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
43	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
44	Male	31	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
45	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
46	Male	31	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
47	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
48	Male	31	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
49	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
50	Male	31	Doctor	7.7	7	75	6	Normal	120/80	70	8000	Sleep Apnea
51	Male	32	Engineer	7.5	8	45	3	Normal	120/80	70	8000	None
52	Male	32	Engineer	7.5	8	45	3	Normal	120/80	70	8000	None
53	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
54	Male	32	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None
55	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
56	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
57	Male	32	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
58	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
59	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
60	Male	32	Doctor	7.7	7	75	6	Normal	120/80	70	8000	None
61	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
62	Male	32	Doctor	6	6	30	8	Normal	125/80	72	5000	None
63	Male	32	Doctor	6.2	6	30	8	Normal	125/80	72	5000	None
64	Male	32	Doctor	6.2	6	30	8	Normal	125/80	72	5000	None
65	Male	32	Doctor	6.2	6	30	8	Normal	125/80	72	5000	None
66	Male	32	Doctor	6.2	6	30	8	Normal	125/80	72	5000	None
67	Male	32	Accountar	7.2	8	50	6	Normal W	118/76	68	7000	None
68	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	Insomnia
69	Female	33	Scientist	6.2	6	50	6	Overweig	128/85	76	5500	None

70	Female	33	Scientist	6.2	6	50	6	Overweig	128/85	76	5500	None	
71	Male	33	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None	
72	Male	33	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None	
73	Male	33	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None	
74	Male	33	Doctor	6.1	6	30	8	Normal	125/80	72	5000	None	
75	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
76	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
77	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
78	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
79	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
80	Male	33	Doctor	6	6	30	8	Normal	125/80	72	5000	None	
81	Female	34	Scientist	5.8	4	32	8	Overweig	131/86	81	5200	Sleep Apnea	
82	Female	34	Scientist	5.8	4	32	8	Overweig	131/86	81	5200	Sleep Apnea	
83	Male	35	Teacher	6.7	7	40	5	Overweig	128/84	70	5600	None	
84	Male	35	Teacher	6.7	7	40	5	Overweig	128/84	70	5600	None	
85	Male	35	Software i	7.5	8	60	5	Normal W	120/80	70	8000	None	
86	Female	35	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
87	Male	35	Engineer	7.2	8	60	4	Normal	125/80	65	5000	None	
88	Male	35	Engineer	7.2	8	60	4	Normal	125/80	65	5000	None	
89	Male	35	Engineer	7.3	8	60	4	Normal	125/80	65	5000	None	
90	Male	35	Engineer	7.3	8	60	4	Normal	125/80	65	5000	None	
91	Male	35	Engineer	7.3	8	60	4	Normal	125/80	65	5000	None	
92	Male	35	Engineer	7.3	8	60	4	Normal	125/80	65	5000	None	
93	Male	35	Software i	7.5	8	60	5	Normal W	120/80	70	8000	None	
94	Male	35	Lawyer	7.4	7	60	5	Obese	135/88	84	3300	Sleep Apnea	
95	Female	36	Accountar	7.2	8	60	4	Normal	115/75	68	7000	Insomnia	
96	Female	36	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
97	Female	36	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
98	Female	36	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
99	Female	36	Teacher	7.1	8	60	4	Normal	115/75	68	7000	None	
100	Female	36	Teacher	7.1	8	60	4	Normal	115/75	68	7000	None	
101	Female	36	Teacher	7.2	8	60	4	Normal	115/75	68	7000	None	
102	Female	36	Teacher	7.2	8	60	4	Normal	115/75	68	7000	None	
103	Female	36	Teacher	7.2	8	60	4	Normal	115/75	68	7000	None	
104	Male	36	Teacher	6.6	5	35	7	Overweig	129/84	74	4800	Sleep Apnea	
105	Female	36	Teacher	7.2	8	60	4	Normal	115/75	68	7000	Sleep Apnea	
106	Male	36	Teacher	6.6	5	35	7	Overweig	129/84	74	4800	Insomnia	
107	Female	37	Nurse	6.1	6	42	6	Overweig	126/83	77	4200	None	
108	Male	37	Engineer	7.8	8	70	4	Normal W	120/80	68	7000	None	
109	Male	37	Engineer	7.8	8	70	4	Normal W	120/80	68	7000	None	
110	Male	37	Lawyer	7.4	8	60	5	Normal	130/85	68	8000	None	
111	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
112	Male	37	Lawyer	7.4	8	60	5	Normal	130/85	68	8000	None	
113	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
114	Male	37	Lawyer	7.4	8	60	5	Normal	130/85	68	8000	None	
115	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
116	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
117	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
118	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
119	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
120	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
121	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
122	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
123	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
124	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
125	Female	37	Accountar	7.2	8	60	4	Normal	115/75	68	7000	None	
126	Female	37	Nurse	7.5	8	60	4	Normal W	120/80	70	8000	None	
127	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
128	Female	38	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
129	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
130	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
131	Female	38	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
132	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
133	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
134	Female	38	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
135	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
136	Male	38	Lawyer	7.3	8	60	5	Normal	130/85	68	8000	None	
137	Female	38	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	
138	Male	38	Lawyer	7.1	8	60	5	Normal	130/85	68	8000	None	
139	Female	38	Accountar	7.1	8	60	4	Normal	115/75	68	7000	None	

