1.Exploratory_Data_Analysis

October 11, 2024

0.1 Predicting Sleep Disorders Based on Lifestyle and Cardiovascular Factors

Problem Statement

This project aims to analyze the Sleep Health and Lifestyle Dataset to understand the relationships between various lifestyle factors (such as physical activity, stress levels, and BMI) and their impact on sleep quality and disorders. Specifically, we will investigate: 1. How lifestyle factors correlate with sleep duration and quality. 2. The prevalence of sleep disorders among different demographic groups and its association with cardiovascular health metrics (blood pressure and heart rate). 3. Identifying potential predictors for sleep disorders based on lifestyle and health indicators.

Potential Libraries for End-to-End Project 1. Data Manipulation and Analysis: - Pandas: For data cleaning and manipulation. - NumPy: For numerical operations and handling arrays.

2. Data Visualization:
   - Matplotlib: For basic plotting.
   - Seaborn: For statistical data visualization and enhanced visual aesthetics.
3. Statistical Analysis:
   - SciPy: For statistical tests and analysis.
   - Statsmodels: For regression analysis and advanced statistical modeling.
4. Machine Learning:
   - Scikit-learn: For implementing machine learning algorithms.
   - XGBoost: For boosting algorithms
5. Reporting and Documentation:
   - Jupyter Notebook: For documenting the analysis and findings in an interactive environment.

Feasibility of Empirical Analysis and Reporting

Conducting empirical analysis on this dataset is feasible due to the following reasons: - Sufficient Data: With 400 rows, the dataset is large enough to derive meaningful insights, especially if properly stratified. - Rich Feature Set: The variety of variables allows for comprehensive analysis, including correlations and potential causal relationships. - Clear Metrics: The presence of sleep quality, duration, and lifestyle indicators enables straightforward statistical evaluations.

We can generate reports summarizing your findings using Jupyter Notebook, which supports both code and narrative text, or by creating visual dashboards using libraries like Plotly or Tableau.

0.1.1 DataSet OverView :

The Dataset comprises 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

0.1.2 Key Features of the Dataset:

Comprehensive Sleep Metrics: Explore sleep duration, quality, and factors influencing sleep patterns. Lifestyle Factors: Analyze physical activity levels, stress levels, and BMI categories. Cardiovascular Health: Examine blood pressure and heart rate measurements. Sleep Disorder Analysis: Identify the occurrence of sleep disorders such as Insomnia and Sleep Apnea.

0.1.3 Dataset Columns:

Person ID: An identifier for each individual.
Gender: The gender of the person (Male/Female).
Age: The age of the person in years.
Occupation: The occupation or profession of the person.
Sleep Duration (hours): The number of hours the person sleeps per day.
Quality of Sleep (scale: 1-10): A subjective rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.
Stress Level (scale: 1-10): A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
BMI Category: The BMI category of the person (e.g., Underweight, Normal, Overweight).
Blood Pressure (systolic/diastolic): The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
Heart Rate (bpm): The resting heart rate of the person in beats per minute. Daily Steps: The number of steps the person takes per day.
Sleep Disorder: The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

0.1.4 Details about Sleep Disorder Column:

No: The individual does not exhibit any specific sleep disorder.
Insomnia: The individual experiences difficulty falling asleep or staying asleep, leading to inadequate or poor-quality sleep.
Sleep Apnea: The individual suffers from pauses in breathing during sleep, resulting in disrupted sleep patterns and potential health risks.

1. Import Libraries

[4]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Load the Dataset

[6]:
```python
# Load the dataset
data = pd.read_csv('./data/Sleep_health_and_lifestyle_dataset.csv')
```

```
# Display the first few rows
data.head().style.set_properties(**{'background-color': '#A04747',
                                    'color': '#E2EEF3'})
```

[6]:

```
<pandas.io.formats.style.Styler at 0x13a17e7b0>
```

3. Data Overview

[8]:

```
# Check the shape of the dataset
data.shape

# Display data types and missing values
data.info()

# Summary statistics
data.describe().style.background_gradient(cmap='viridis')
```

```
<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           374 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

[8]:

```
<pandas.io.formats.style.Styler at 0x13ddc3590>
```

4. Data Cleaning

[10]:

```
# Check for missing values
missing_values = data.isnull().sum()
print(missing_values)
# Display missing values
missing_values[missing_values > 0]
```

```
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             0
dtype: int64
```

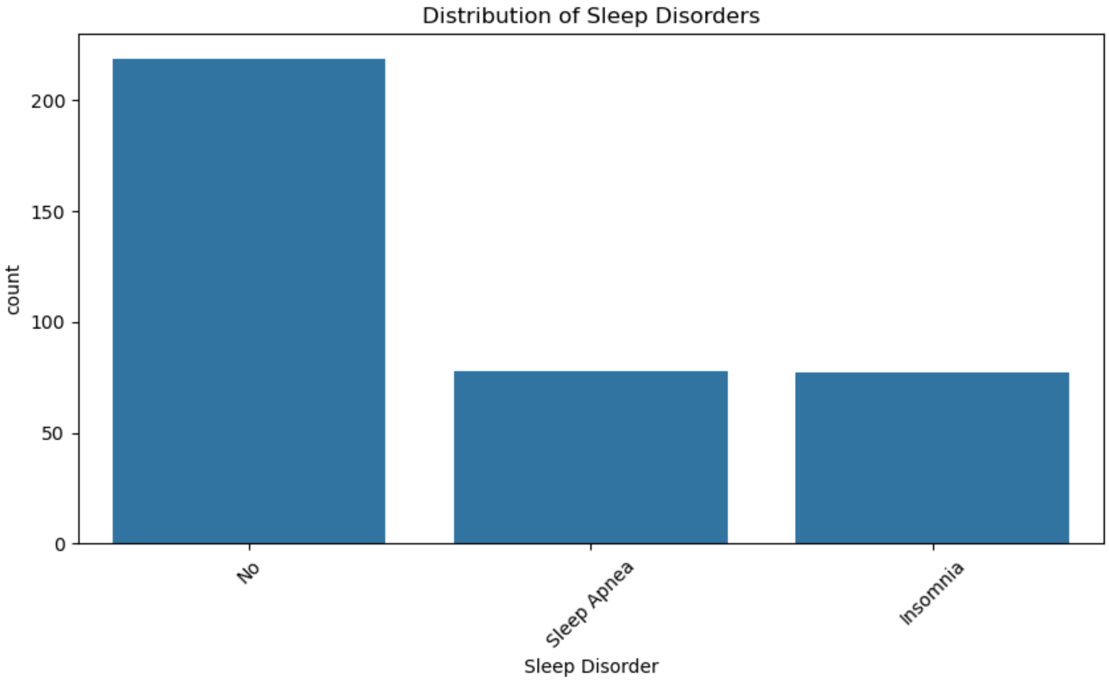
```
[10]: Series([], dtype: int64)
```

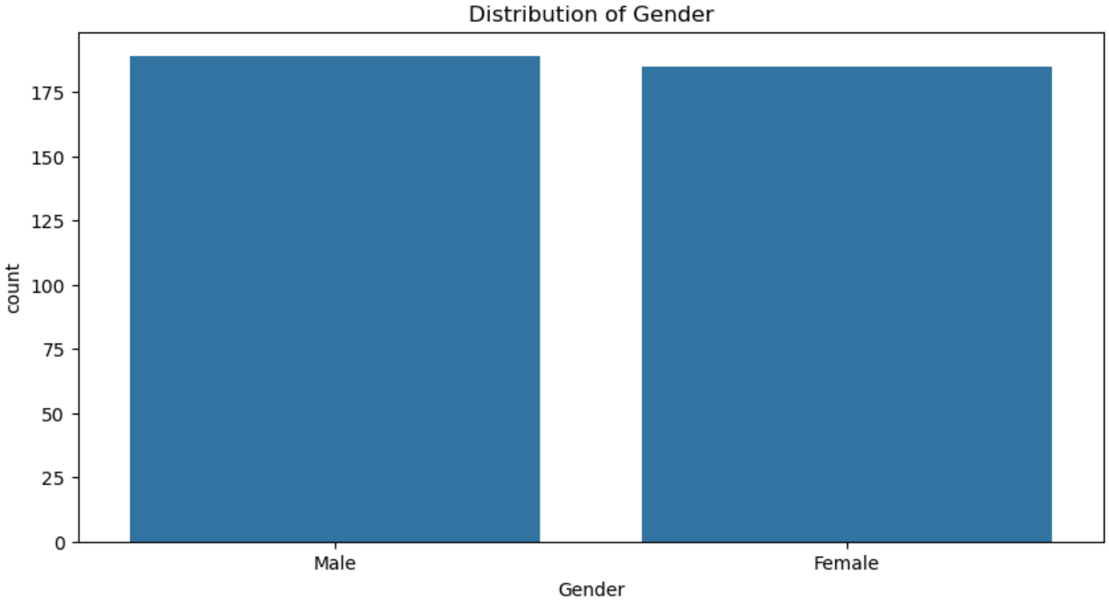
5. Univariate Analysis

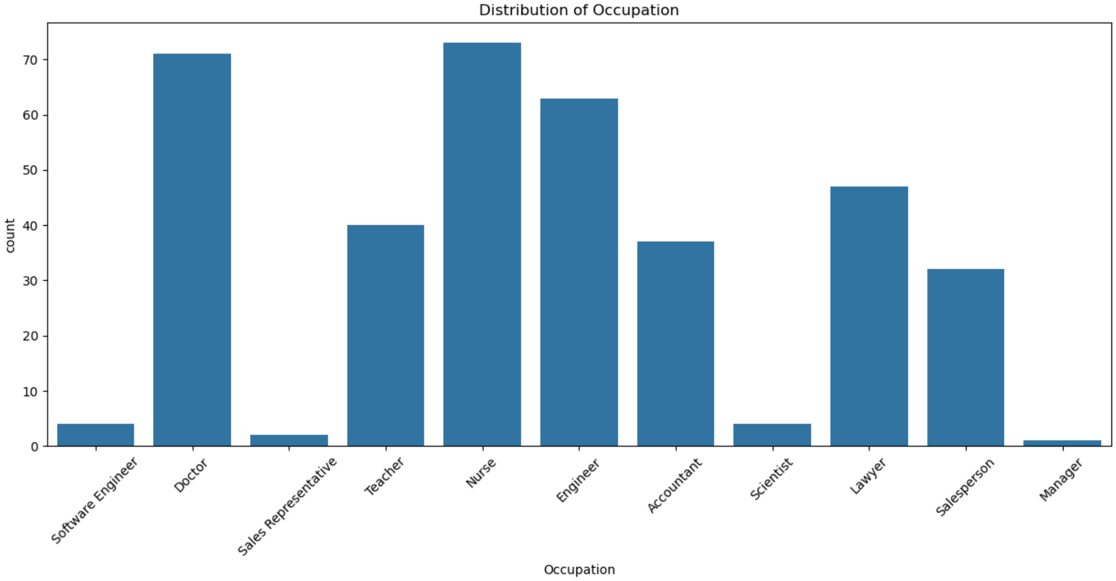
5.1. Categorical Features

[12]:

```python
# Count plot for Sleep Disorder
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='Sleep Disorder')
plt.title('Distribution of Sleep Disorders')
plt.xticks(rotation=45)
plt.show()

# Count plot for Gender
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='Gender')
plt.title('Distribution of Gender')
plt.show()

# Count plot for Occupation
plt.figure(figsize=(15, 6))
sns.countplot(data=data, x='Occupation')
plt.title('Distribution of Occupation')
plt.xticks(rotation=45)
plt.show()
```


Distribution of Sleep Disorders
count
0
50
100
150
200
No
Sleep Apnea
Insomnia
Sleep Disorder


Distribution of Gender
count
0
25
50
75
100
125
150
175
Male
Female
Gender



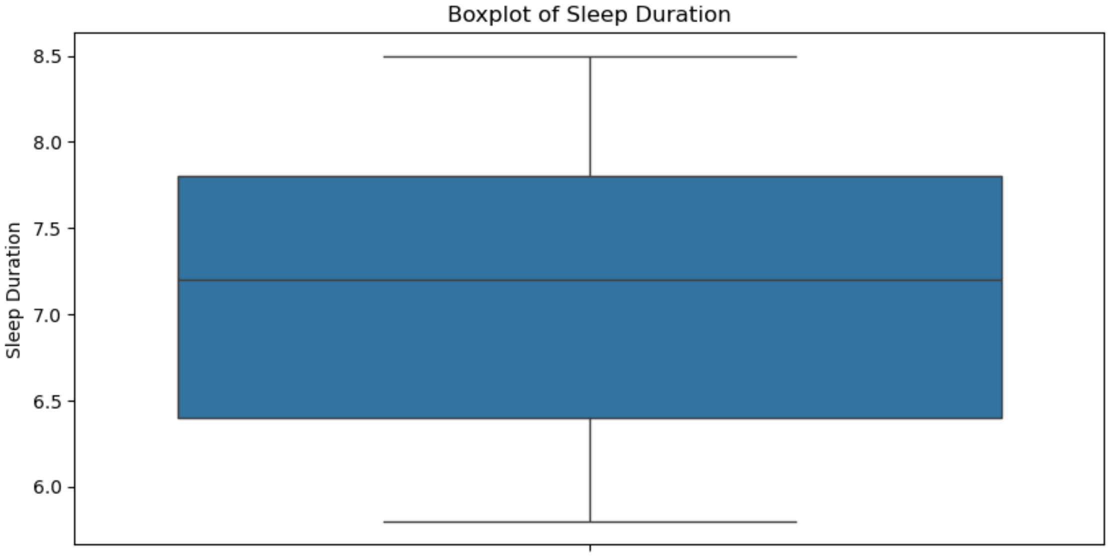
Interpretation of Charts

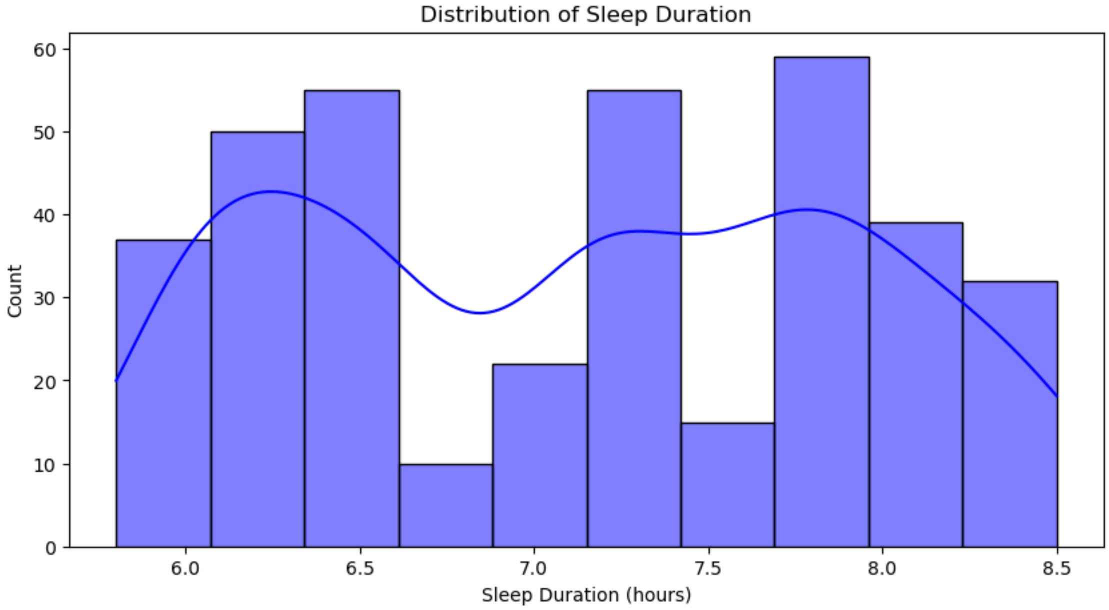
1. The First Chart, "Distribution of Sleep Disorders", shows that the dataset contains around 250+ cases of No disorder, 75+ each cases for Sleep Apnea and Insomnia
2. The Second Chart, "Distribution of Gender", shows that the dataset contains 175+ each cases for Male and Female which means dataset is evenly distributed of gender.
3. The Third Chart, "Distribution of Occupation", shows the occupation of the cases and Doctor, Nurse and Engineer being the top 3 count.

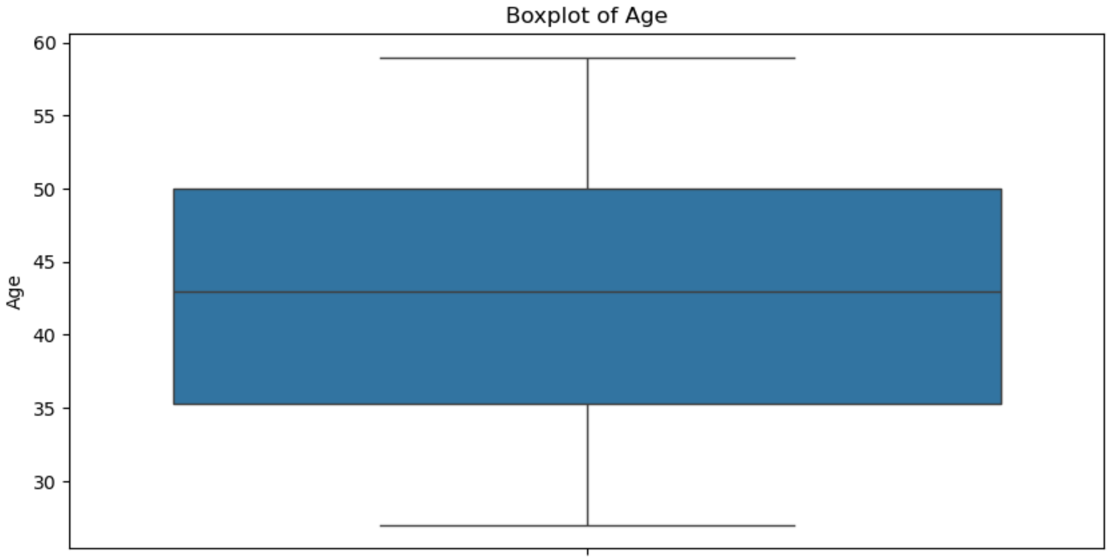
5.2. Numerical Features

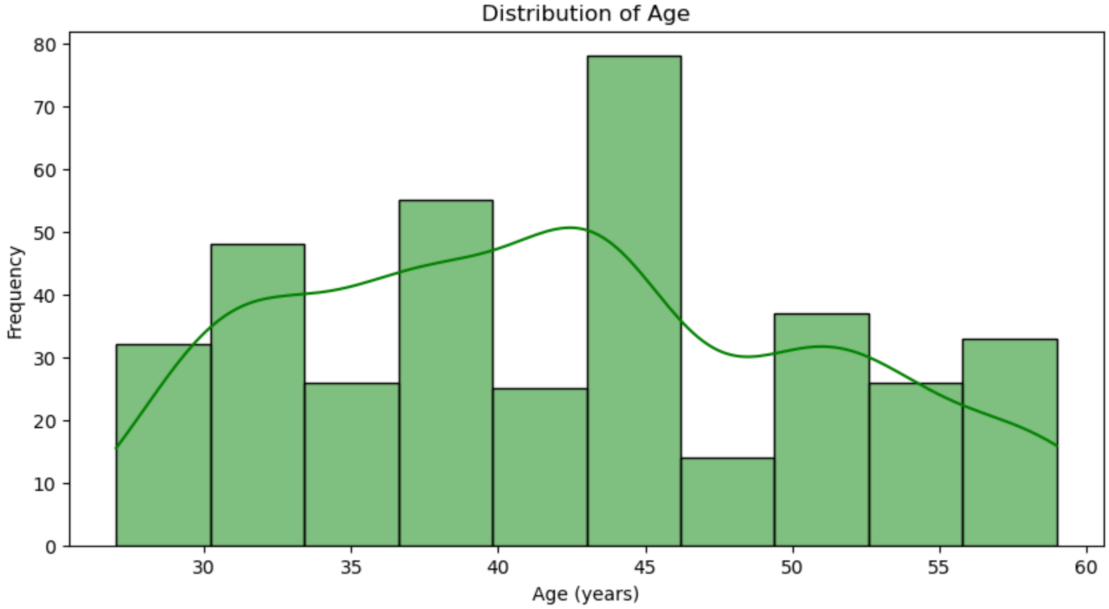
[15]:
```
# Boxplot for Sleep Duration
plt.figure(figsize=(10, 5))
sns.boxplot(data['Sleep Duration'])
plt.title('Boxplot of Sleep Duration')
plt.show()

# Histogram for Sleep Duration
plt.figure(figsize=(10, 5))
sns.histplot(data['Sleep Duration'], kde=True, color='blue')
plt.title('Distribution of Sleep Duration')
plt.xlabel('Sleep Duration (hours)')
plt.show()

# Boxplot for Age
plt.figure(figsize=(10, 5))
sns.boxplot(data['Age'])
plt.title('Boxplot of Age')
plt.show()
```

```python
#histogram for Age
plt.figure(figsize=(10, 5))
sns.histplot(data['Age'], kde=True, color='green')
plt.title('Distribution of Age')
plt.xlabel('Age (years)')
plt.ylabel('Frequency')
plt.show()
```









Interpretation of Charts

1. The First Chart, “Boxplot of Sleep Duration”, shows the spread of sleep hours among the participants. The median sleep duration is around 7.25 hrs, minimum is 6.25 hrs and maximum is 7.75 within a reasonable range, while any points outside this range are considered outliers.

2. The Second Chart, “Histogram of Sleep Duration”, displays the frequency distribution of sleep hours. The histogram provides insights into how many respondents fall into different sleep

hour categories. If there's a peak around 7 to 8 hours, it indicates that most participants are getting adequate sleep. The KDE line helps visualize the density of sleep durations, showing where the majority of values lie.

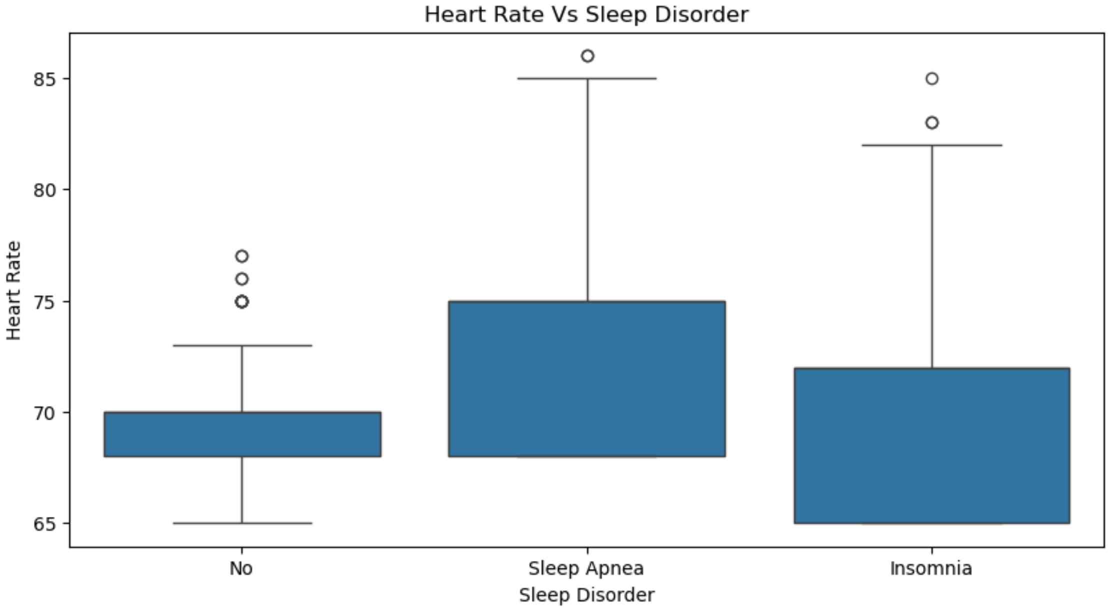
3. The Third Chart, "Boxplot of Age", illustrates the distribution of ages among participants. The median age is around 42.5 shown by the line inside the box, and the whiskers indicate the range of ages in the dataset. Any outliers represent individuals who are significantly younger or older, providing insights into the age diversity within the sample.

4. The Fourth Chart, "Histogram of Age", represents the frequency of participants across different age groups. This histogram shows how many individuals fall into each age category, with the KDE overlay providing a smoother curve to understand the overall age distribution. If the histogram peaks in a certain age range, it indicates that most participants belong to that demographic, which may influence their sleep patterns.

6. Bi-Variate Analysis

[18]:

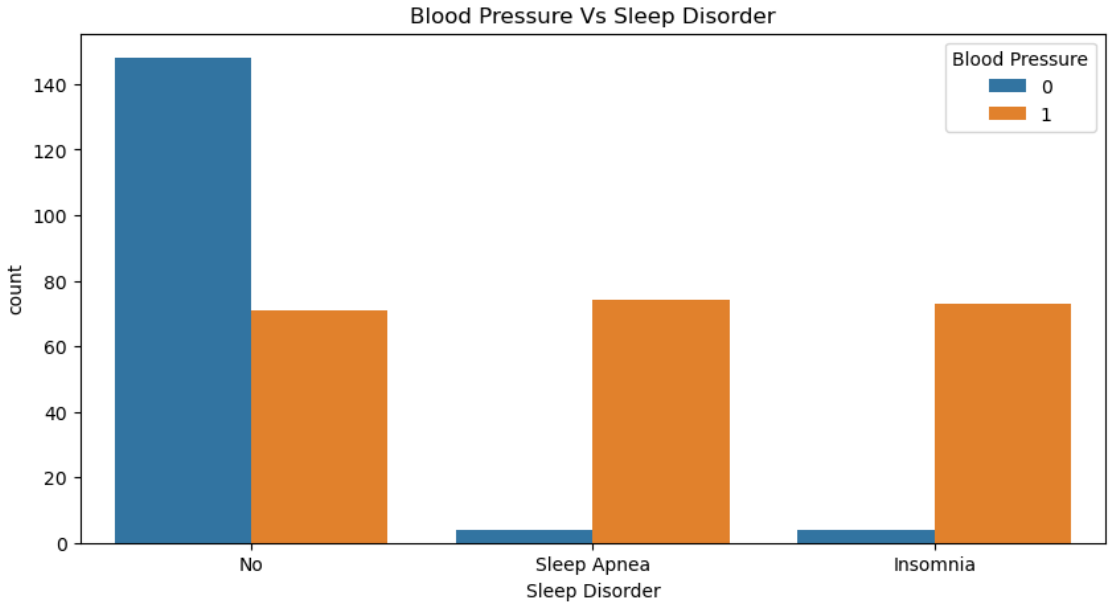
```
#Heart Rate Vs Sleep Disorder
plt.figure(figsize=(10, 5))
sns.boxplot(data=data, x='Sleep Disorder', y='Heart Rate')
plt.title('Heart Rate Vs Sleep Disorder')
plt.show()
```



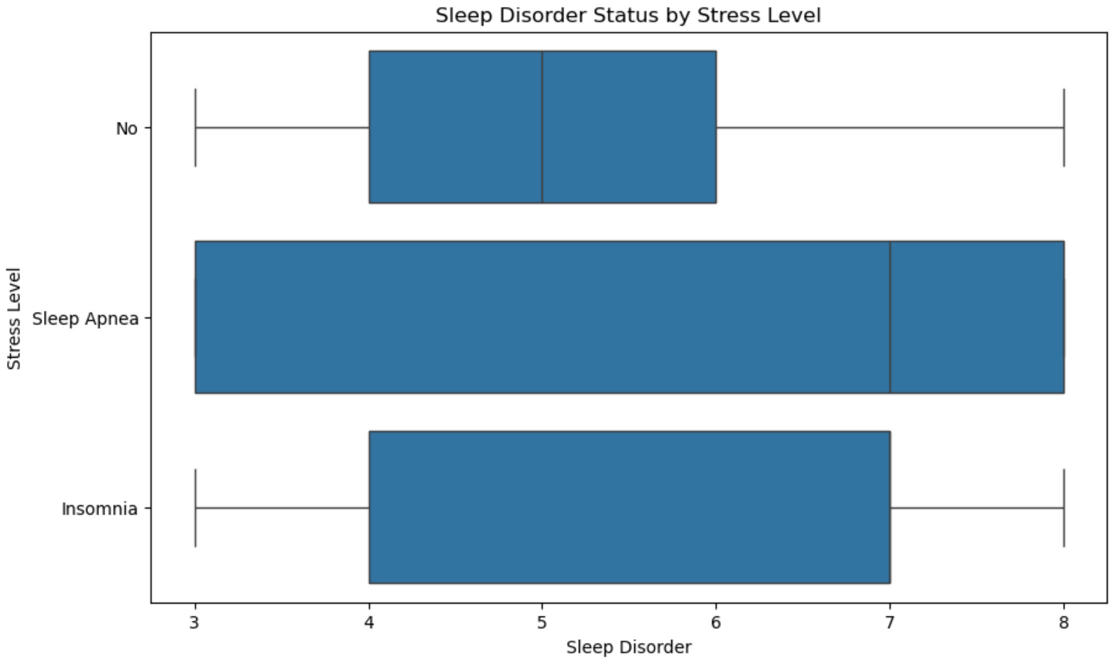
[19]:

```
#Blood Pressure Vs Sleep Disorder
#Ideal blood pressure systolic (upper number) : less than 120 , diastolic␣
 ↪(bottom number) : less than 80
```

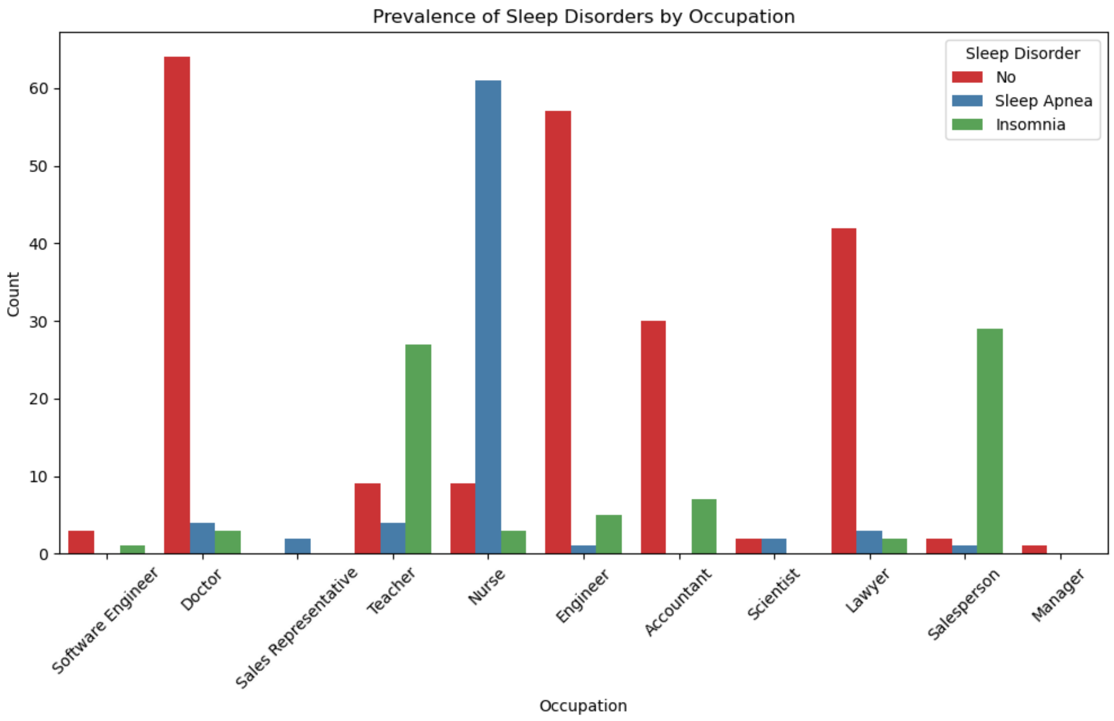
```
# Normal systolic (upper number) : in range (120 - 129) , diastolic (bottom␣
 ↪number) : in range (80 - 84)

# Otherwise, blood pressure is high
data_copy = data.copy()
data_copy['Blood Pressure']=data_copy['Blood Pressure'].apply(lambda x:0 if x␣
 ↪in ['120/80','126/83','125/80','128/84','129/84','117/76','118/76','115/
 ↪75','125/82','122/80'] else 1)
# 0 = normal blood pressure
# 1 = abnormal blood pressure

plt.figure(figsize=(10, 5))
sns.countplot(data=data_copy, hue='Blood Pressure', x='Sleep Disorder')
plt.title('Blood Pressure Vs Sleep Disorder')
plt.show()
```



[20]:
```
#Sleep Disorder
plt.figure(figsize=(10, 6))
sns.boxplot(data=data, x='Stress Level', y='Sleep Disorder')
plt.title('Sleep Disorder Status by Stress Level')
plt.xlabel('Sleep Disorder')
plt.ylabel('Stress Level')
plt.show()
```

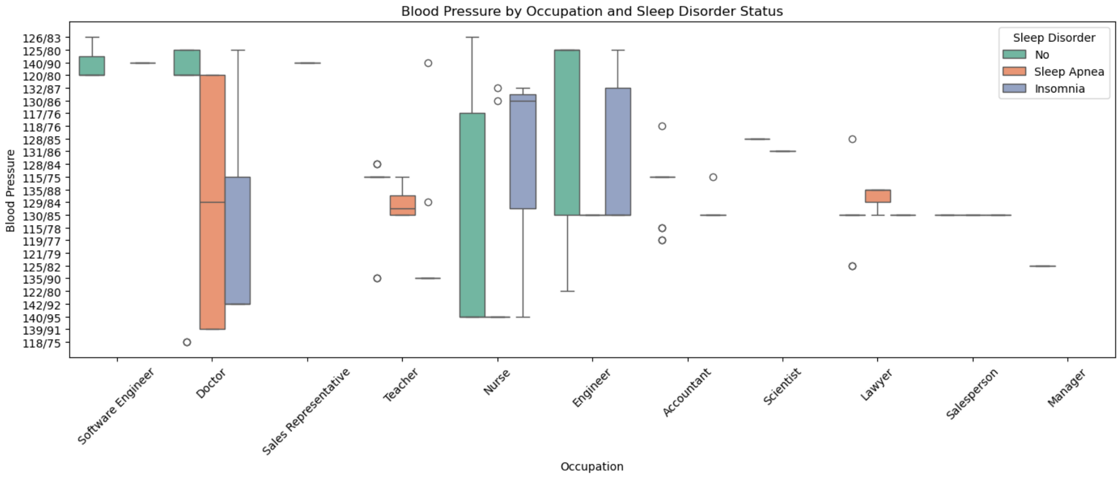


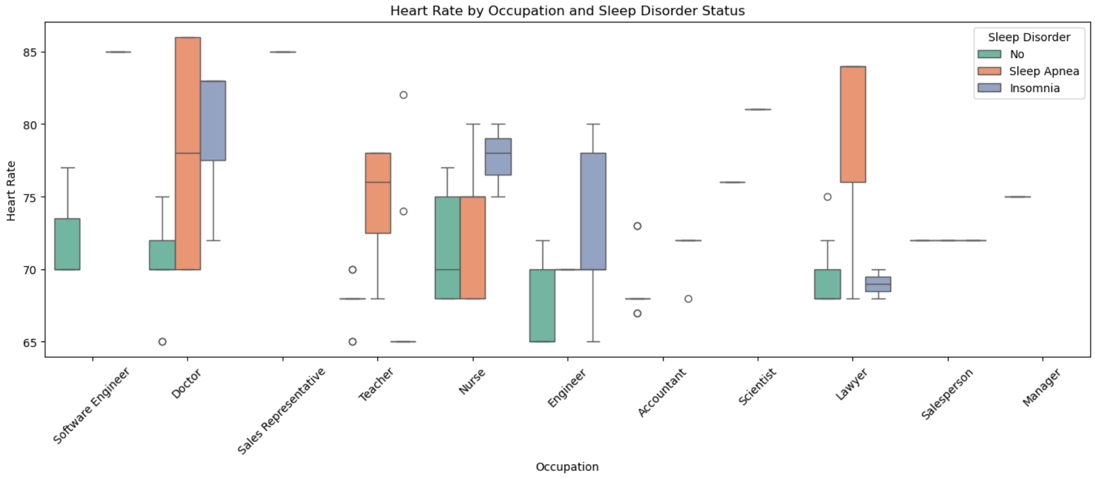
[21]:
```python
plt.figure(figsize=(12, 6))
sns.countplot(data=data, x='Occupation', hue='Sleep Disorder', palette='Set1')
plt.title('Prevalence of Sleep Disorders by Occupation')
plt.xlabel('Occupation')
plt.ylabel('Count')
plt.xticks(rotation=45)  # Rotate x labels for better visibility
plt.legend(title='Sleep Disorder', labels=['No', 'Sleep Apnea', 'Insomnia'])
plt.show()
```



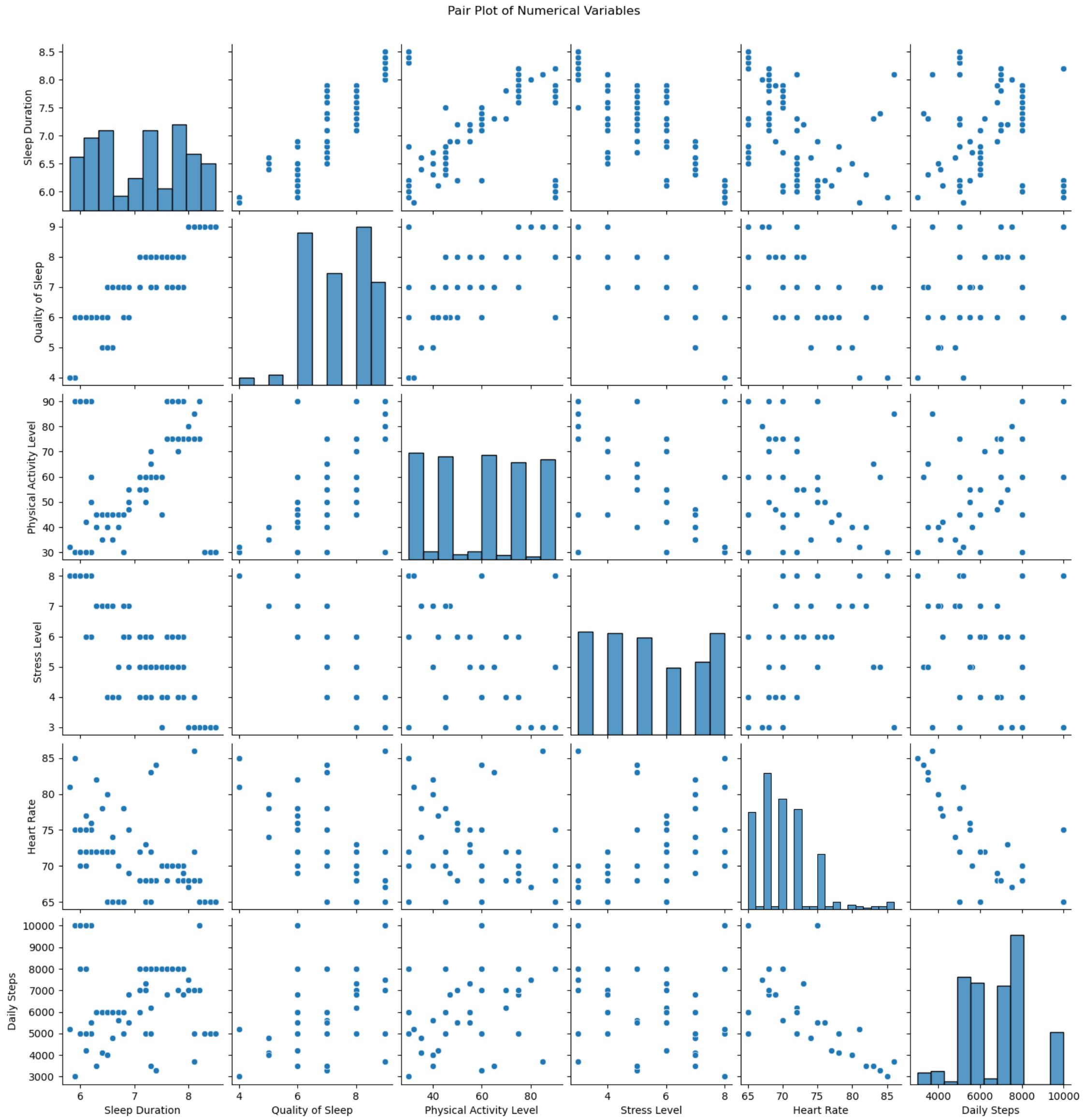
[22]:

```
plt.figure(figsize=(14, 12))

# Box plot for Blood Pressure by Occupation and Sleep Disorder status
plt.subplot(2, 1, 1)  # First plot for Blood Pressure
sns.boxplot(data=data, x='Occupation', y='Blood Pressure', hue='Sleep␣
 ↪Disorder', palette='Set2')
plt.title('Blood Pressure by Occupation and Sleep Disorder Status')
plt.xlabel('Occupation')
plt.ylabel('Blood Pressure')
plt.xticks(rotation=45)  # Rotate x labels for better visibility

# Box plot for Heart Rate by Occupation and Sleep Disorder status
plt.subplot(2, 1, 2)  # Second plot for Heart Rate
sns.boxplot(data=data, x='Occupation', y='Heart Rate', hue='Sleep Disorder',␣
 ↪palette='Set2')
plt.title('Heart Rate by Occupation and Sleep Disorder Status')
plt.xlabel('Occupation')
plt.ylabel('Heart Rate')
plt.xticks(rotation=45)  # Rotate x labels for better visibility

plt.tight_layout()
plt.show()
```





[23]:
```python
numerical_cols = ['Sleep Duration', 'Quality of Sleep', 'Physical Activity␣
 ↪Level',
                  'Stress Level', 'Blood Pressure', 'Heart Rate', 'Daily Steps']

# Create a pair plot
sns.pairplot(data[numerical_cols])
plt.suptitle('Pair Plot of Numerical Variables', y=1.02)  # Adjust title␣
 ↪position
plt.show()
```



[24]:
```
df = pd.concat([data, data['Blood Pressure'].str.split('/', expand=True)], axis=1).drop('Blood Pressure', axis=1)
df = df.rename(columns={0: 'BloodPressure_Upper', 1: 'BloodPressure_Lower'})
df['BloodPressure_Upper'] = df['BloodPressure_Upper'].astype(float)
df['BloodPressure_Lower'] = df['BloodPressure_Lower'].astype(float)
df.drop('Person ID', axis=1, inplace=True)
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df['Gender'] = label_encoder.fit_transform(df['Gender'])
df['Occupation'] = label_encoder.fit_transform(df['Occupation'])
df['BMI Category'] = label_encoder.fit_transform(df['BMI Category'])
```

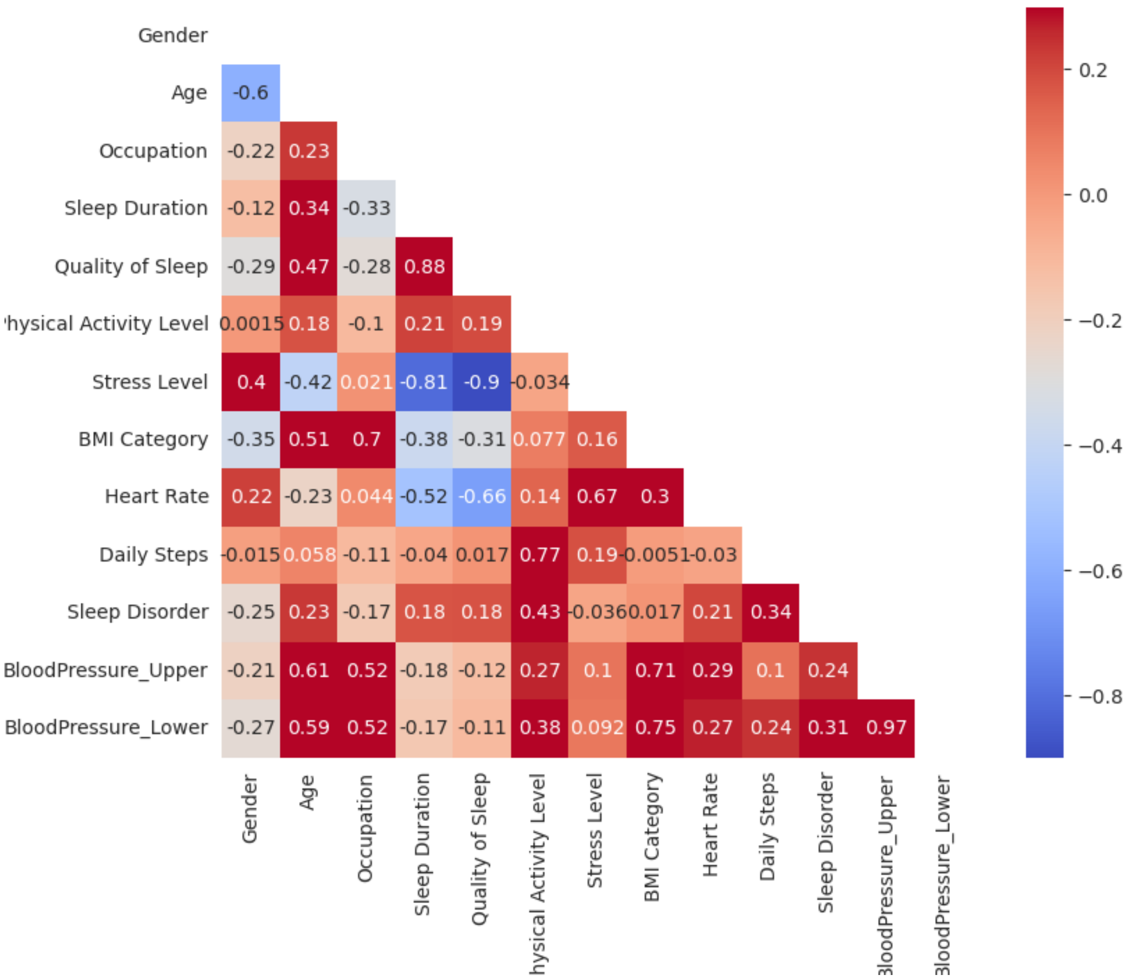
```
df['Sleep Disorder'] = label_encoder.fit_transform(df['Sleep Disorder'])
df.head()
```

[24]:

```
   Gender  Age  Occupation  Sleep Duration  Quality of Sleep  \
0       1   27           9             6.1                 6
1       1   28           1             6.2                 6
2       1   28           1             6.2                 6
3       1   28           6             5.9                 4
4       1   28           6             5.9                 4

   Physical Activity Level  Stress Level  BMI Category  Heart Rate  \
0                       42             6             3          77
1                       60             8             0          75
2                       60             8             0          75
3                       30             8             2          85
4                       30             8             2          85

   Daily Steps  Sleep Disorder  BloodPressure_Upper  BloodPressure_Lower
0         4200               1                126.0                 83.0
1        10000               1                125.0                 80.0
2        10000               1                125.0                 80.0
3         3000               2                140.0                 90.0
4         3000               2                140.0                 90.0
```

[25]:

```
def corr_vis(corr) :
    mask = np.zeros_like(corr)
    mask[np.triu_indices_from(mask)] = True
    with sns.axes_style("white"):
        f, ax = plt.subplots(figsize=(10, 7))
        g = sns.heatmap(corr, mask=mask, vmax=.3, square=True, annot=True,␣
 ↪cmap='coolwarm')
        g.set_xticklabels(g.get_xticklabels(), rotation = 90, fontsize = 10)

num_corr = df.corr()
corr_vis(df.corr())
```


Gender
Age
-0.6
Occupation
-0.22
0.23
Sleep Duration
-0.12
0.34
-0.33
Quality of Sleep
-0.29
0.47
-0.28
0.88
hysical Activity Level
0.0015
0.18
-0.1
0.21
0.19
Stress Level
0.4
-0.42
0.021
-0.81
-0.9
-0.034
BMI Category
-0.35
0.51
0.7
-0.38
-0.31
0.077
0.16
Heart Rate
0.22
-0.23
0.044
-0.52
-0.66
0.14
0.67
0.3
Daily Steps
-0.015
0.058
-0.11
-0.04
0.017
0.77
0.19
-0.0051
-0.03
Sleep Disorder
-0.25
0.23
-0.17
0.18
0.18
0.43
-0.036
0.017
0.21
0.34
BloodPressure_Upper
-0.21
0.61
0.52
-0.18
-0.12
0.27
0.1
0.71
0.29
0.1
0.24
BloodPressure_Lower
-0.27
0.59
0.52
-0.17
-0.11
0.38
0.092
0.75
0.27
0.24
0.31
0.97
Gender
Age
Occupation
Sleep Duration
Quality of Sleep
hysical Activity Level
Stress Level
BMI Category
Heart Rate
Daily Steps
Sleep Disorder
3loodPressure_Upper
3loodPressure_Lower
0.2
0.0
−0.2
−0.4
−0.6
−0.8