SLEEP HEALTH INFLUENCED BY LIFESTYLE FACTORS

Professors:

Petre Caraini
Daniel Traian Pele
Wolfgang Hardle

Research Methods - Project
IMBA, 2nd Year, 2023-2024

Team Members:

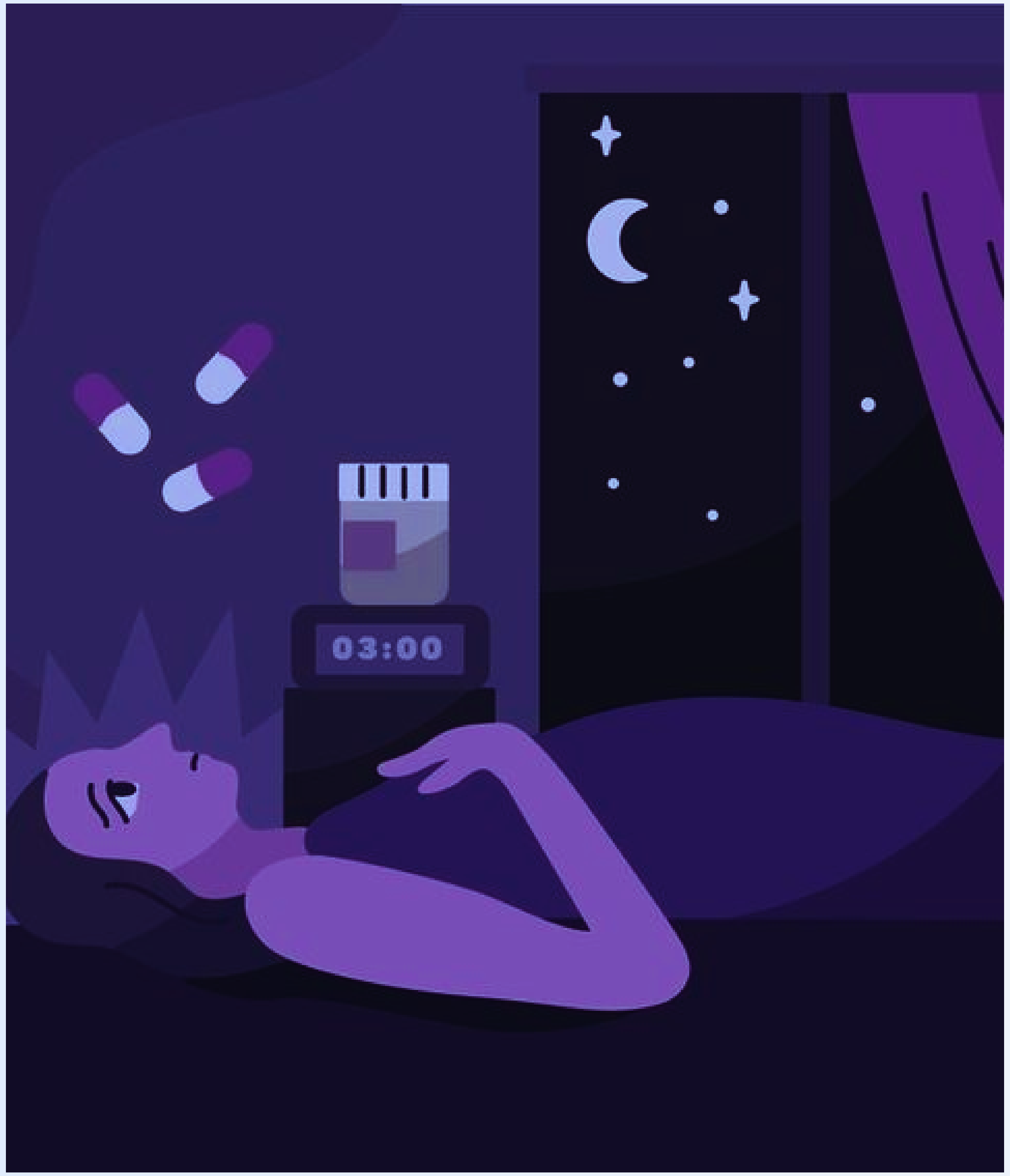
Torjescu Ana-Maria
Dumitru Laura-Alexandra
Costea Cristina-Bianca

INTRODUCTION

Sleep is a fundamental determinant of health, yet its quality is often compromised by lifestyle factors. ***This study aimed to investigate the relationship between sleep health and lifestyle factors, specifically sleep duration, physical activity level, and stress.***

Using a ***dataset of a sample of 374 , we employed descriptive analytics, visualized through scatterplots, to explore initial associations between the variables****. Moreover, an OLS regression model was developed to quantify these relationships and predict the quality of sleep from the selected lifestyle factors.*

Keywords: **Sleep Quality, Lifestyle Factors, Physical Activity, Stress, Regression,Sleep Duration**



DATASET

The Sleep Health and Lifestyle Dataset comprises 374 rows and 13 columns, covering a wide range of **variables related to sleep and daily habits:**

- 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 disorder.

Main Sleep Metrics:

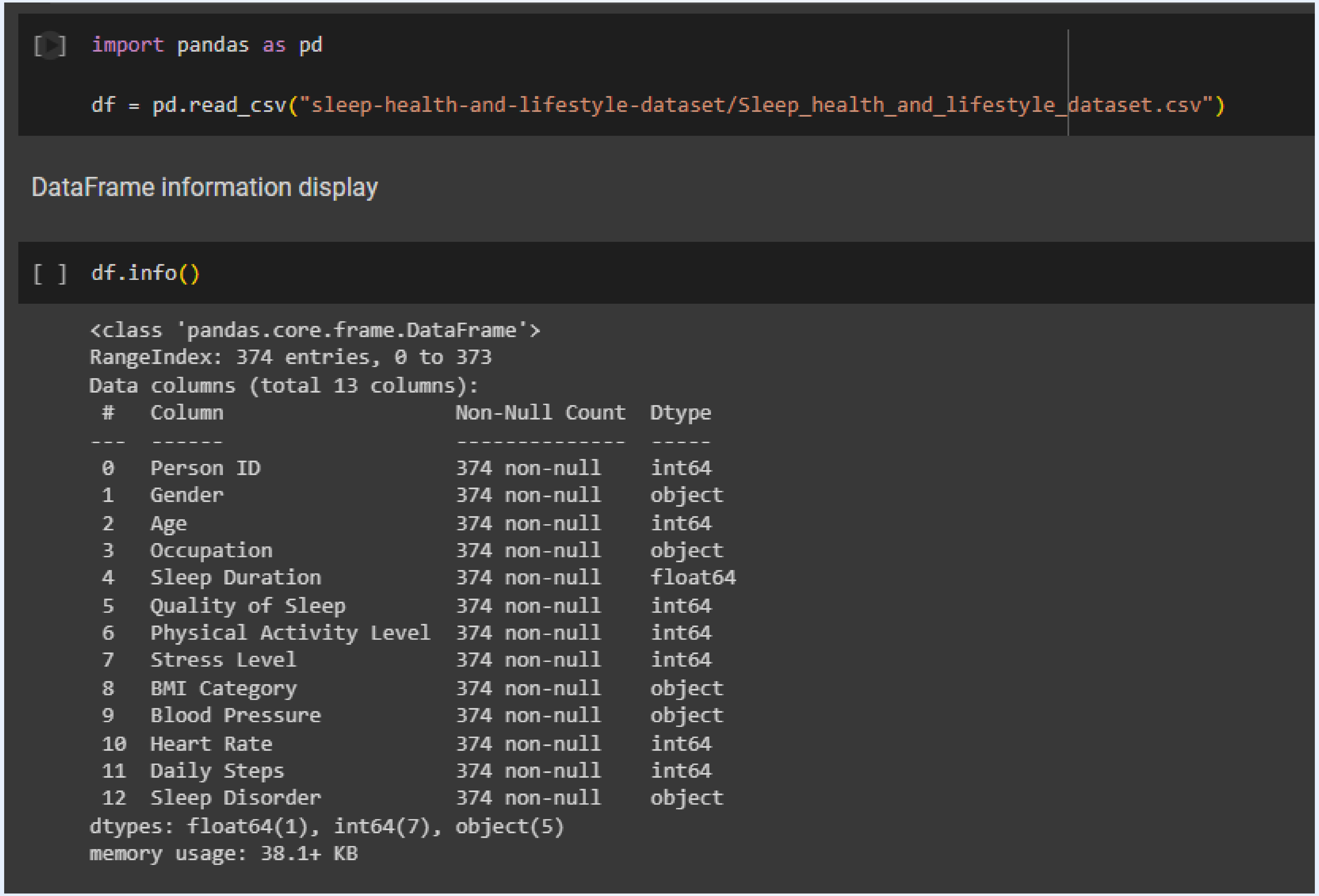
1. **Sleep Duration,**
2. **Quality, and**
3. **Factors influencing Sleep Patterns:**
   - **Lifestyle factors**(physical activity levels, stress levels, and BMI)
   - **Cardiovascular factors** (blood pressure and heart rate)
   - **Sleep disorder factors** (insomnia and sleep apnea).

RESEARCH QUESTIONS

Based on these data, **four research questions** have been developed:

- **What is the correlation between the main factors of the dataset?**
- **What is the relationship/correlation between Sleep - Physical Activity Level? What about Sleep Duration and Stress Level?**
- **What is the linear regression between quality of sleep and sleep duration and to what extent the predicted values differ from actual values?**
- **To what extent does the data analyzed represent “a good fit” for the model?**

1) IMPORTING DATA



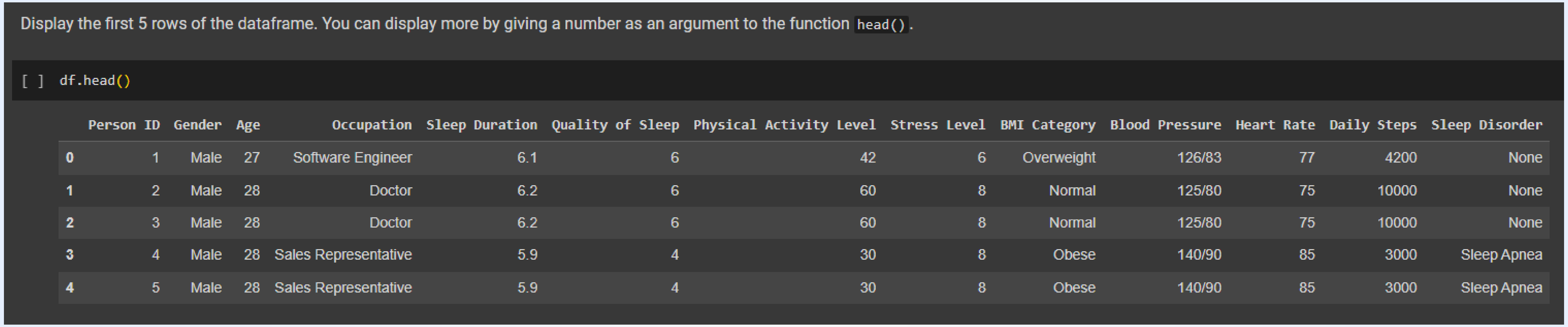
```
[ ] import pandas as pd

    df = pd.read_csv("sleep-health-and-lifestyle-dataset/Sleep_health_and_lifestyle_dataset.csv")
```

DataFrame information display

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

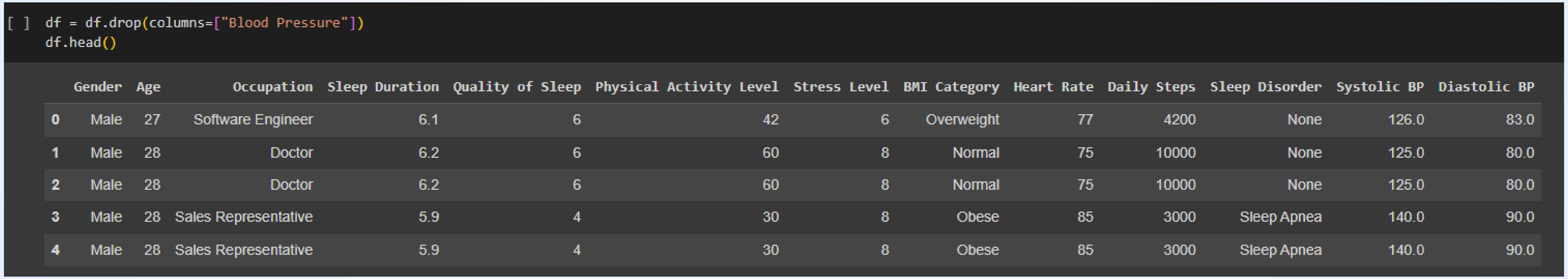
1) DATA



Display the first 5 rows of the dataframe. You can display more by giving a number as an argument to the function `head()`.

```
[ ]  df.head()
```

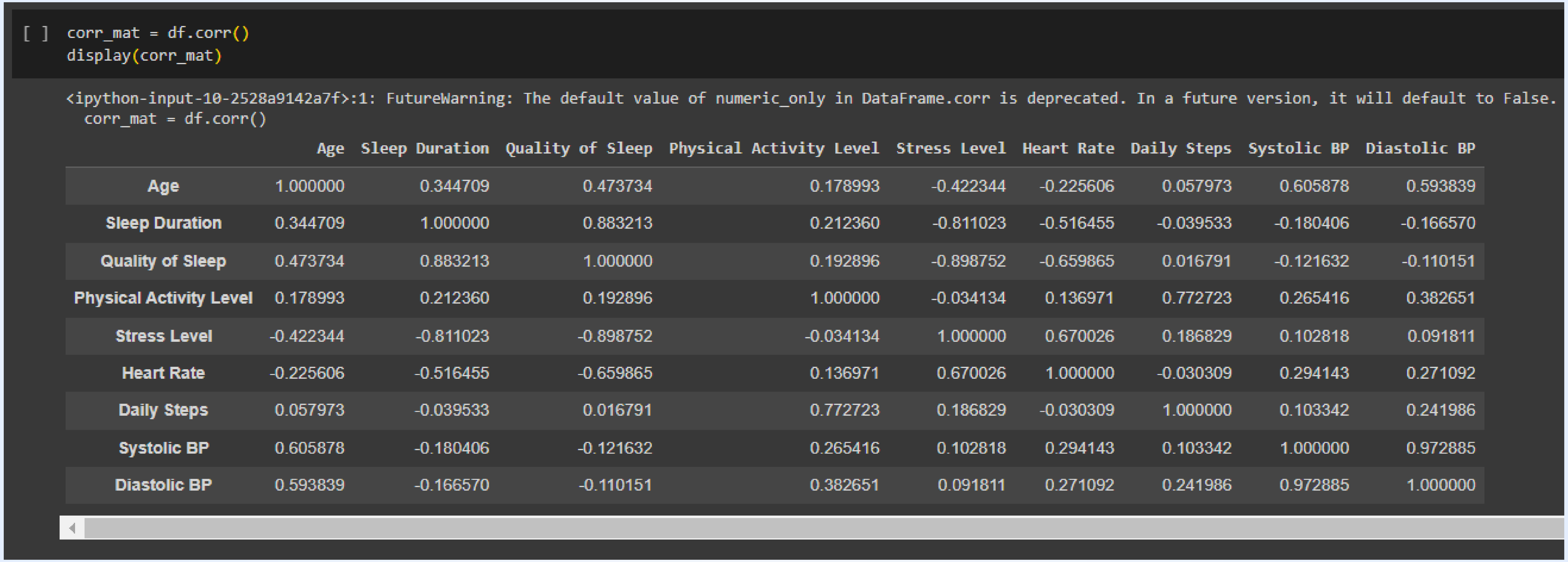
	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea



```
[ ]  df = df.drop(columns=["Blood Pressure"])
     df.head()
```

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Systolic BP	Diastolic BP
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	77	4200	None	126.0	83.0
1	Male	28	Doctor	6.2	6	60	8	Normal	75	10000	None	125.0	80.0
2	Male	28	Doctor	6.2	6	60	8	Normal	75	10000	None	125.0	80.0
3	Male	28	Sales Representative	5.9	4	30	8	Obese	85	3000	Sleep Apnea	140.0	90.0
4	Male	28	Sales Representative	5.9	4	30	8	Obese	85	3000	Sleep Apnea	140.0	90.0

2) CORRELATION MATRIX

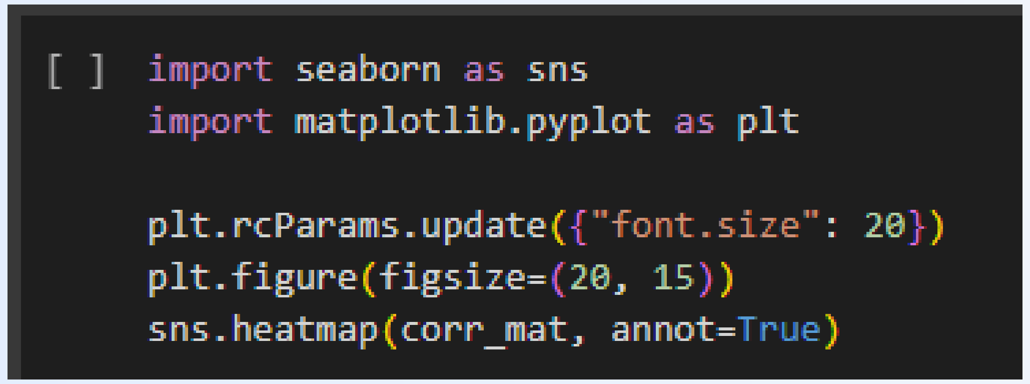


```
[ ]  corr_mat = df.corr()
     display(corr_mat)
```

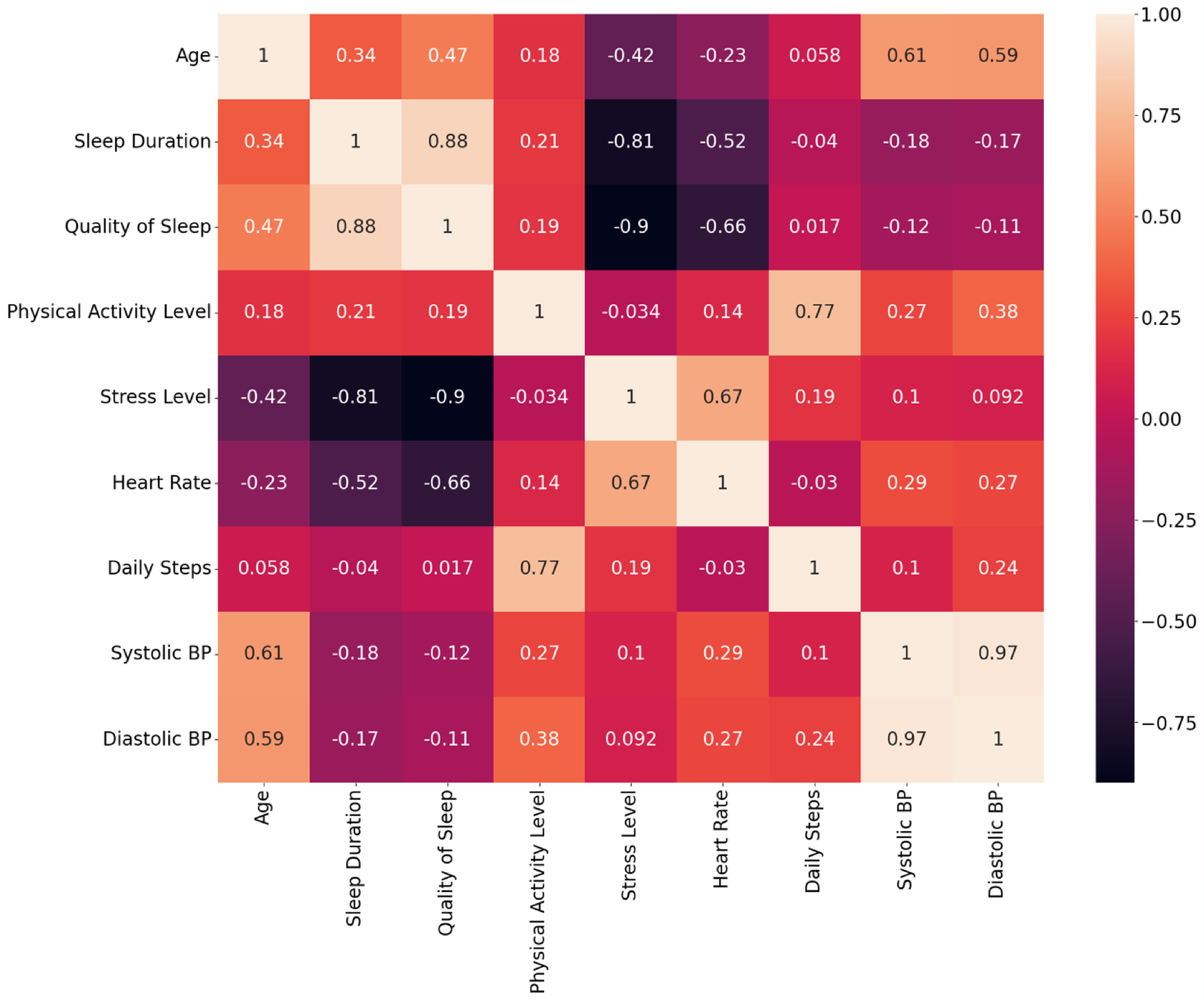
```
<ipython-input-10-2528a9142a7f>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False.
  corr_mat = df.corr()
```

	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps	Systolic BP	Diastolic BP
Age	1.000000	0.344709	0.473734	0.178993	-0.422344	-0.225606	0.057973	0.605878	0.593839
Sleep Duration	0.344709	1.000000	0.883213	0.212360	-0.811023	-0.516455	-0.039533	-0.180406	-0.166570
Quality of Sleep	0.473734	0.883213	1.000000	0.192896	-0.898752	-0.659865	0.016791	-0.121632	-0.110151
Physical Activity Level	0.178993	0.212360	0.192896	1.000000	-0.034134	0.136971	0.772723	0.265416	0.382651
Stress Level	-0.422344	-0.811023	-0.898752	-0.034134	1.000000	0.670026	0.186829	0.102818	0.091811
Heart Rate	-0.225606	-0.516455	-0.659865	0.136971	0.670026	1.000000	-0.030309	0.294143	0.271092
Daily Steps	0.057973	-0.039533	0.016791	0.772723	0.186829	-0.030309	1.000000	0.103342	0.241986
Systolic BP	0.605878	-0.180406	-0.121632	0.265416	0.102818	0.294143	0.103342	1.000000	0.972885
Diastolic BP	0.593839	-0.166570	-0.110151	0.382651	0.091811	0.271092	0.241986	0.972885	1.000000

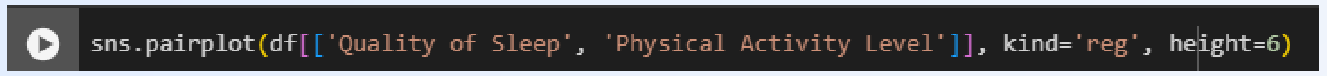
3) CORRELATION HEAT MAP



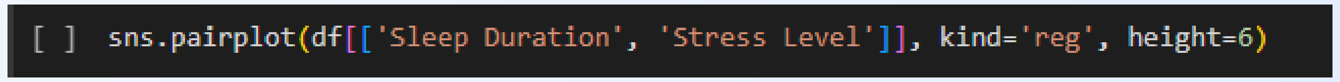
```
[ ]  import seaborn as sns
     import matplotlib.pyplot as plt

     plt.rcParams.update({"font.size": 20})
     plt.figure(figsize=(20, 15))
     sns.heatmap(corr_mat, annot=True)
```



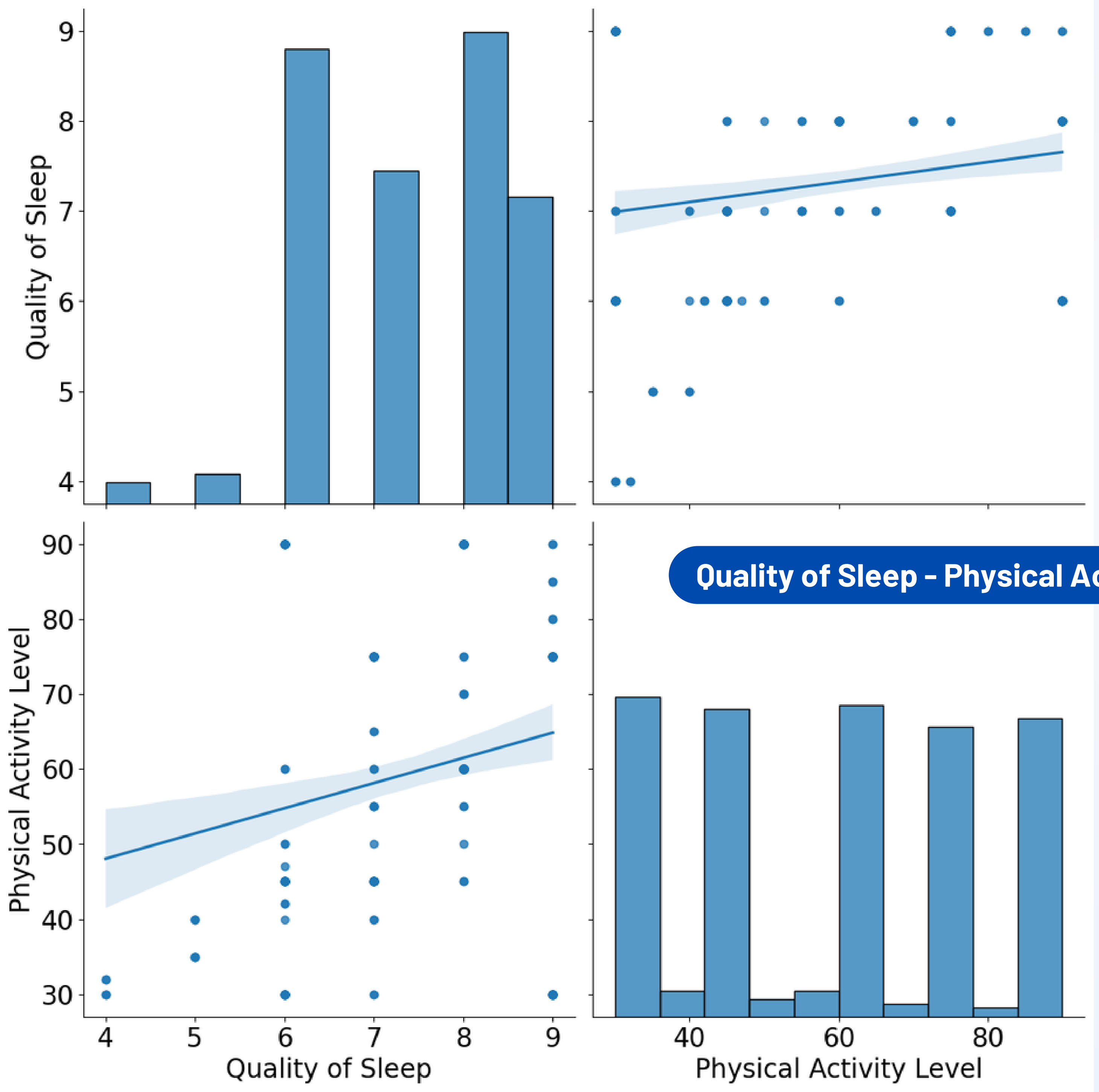
4) PAIR PLOTS



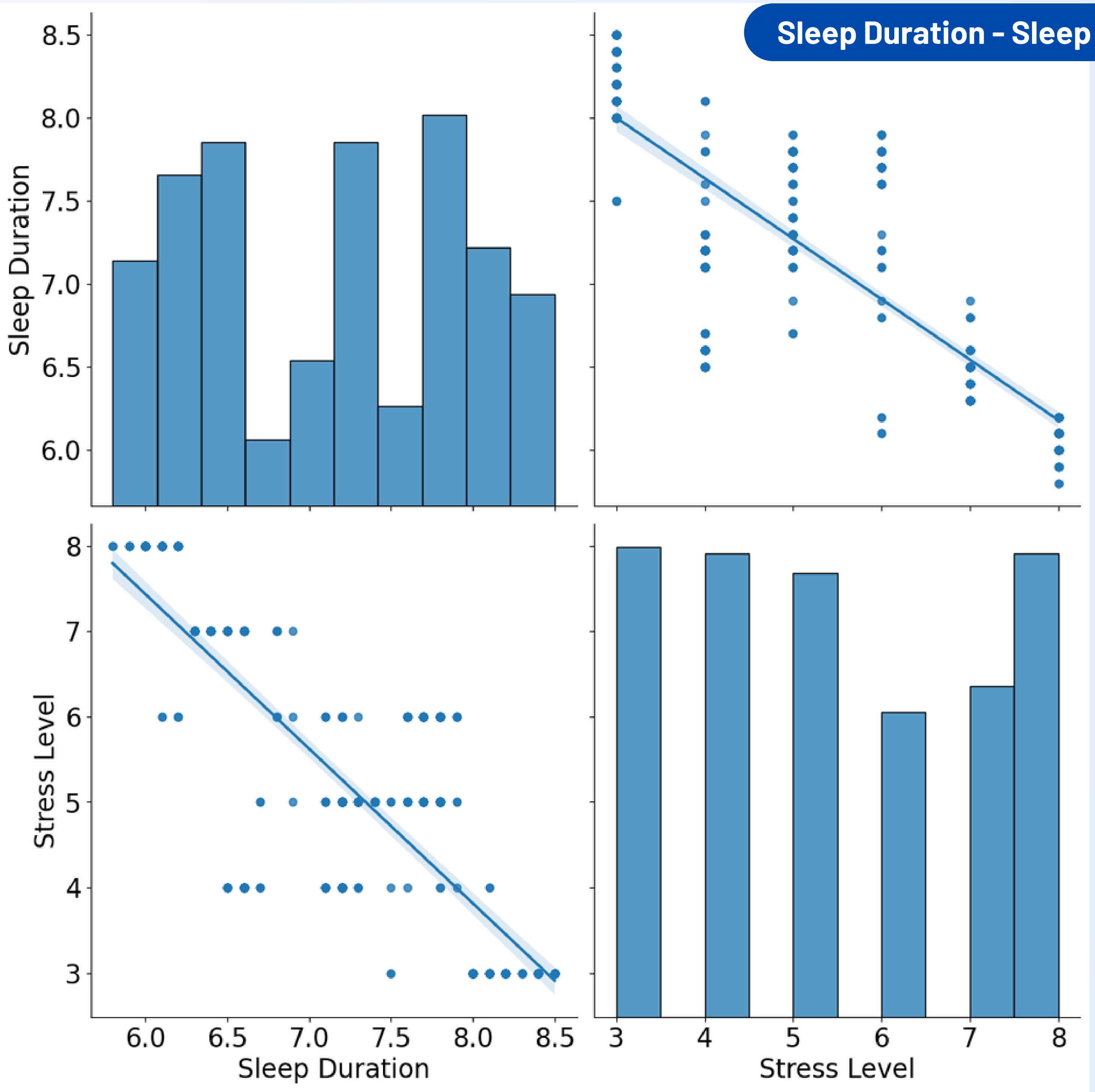
```
sns.pairplot(df[['Quality of Sleep', 'Physical Activity Level']], kind='reg', height=6)
```



```
[ ]  sns.pairplot(df[['Sleep Duration', 'Stress Level']], kind='reg', height=6)
```

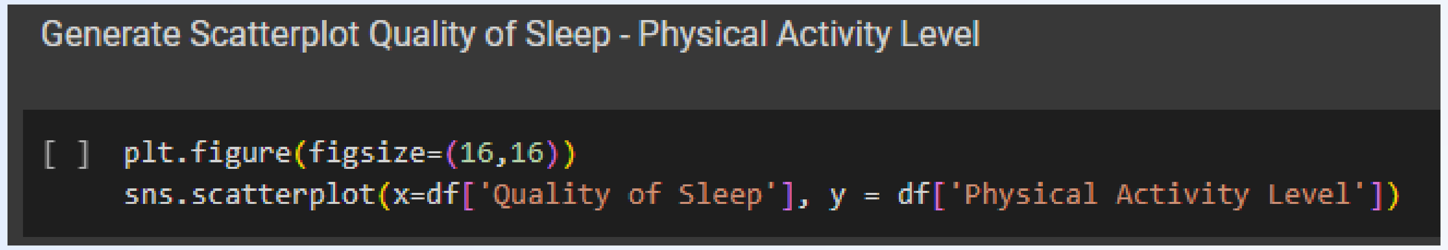


Quality of Sleep - Physical Activity Level



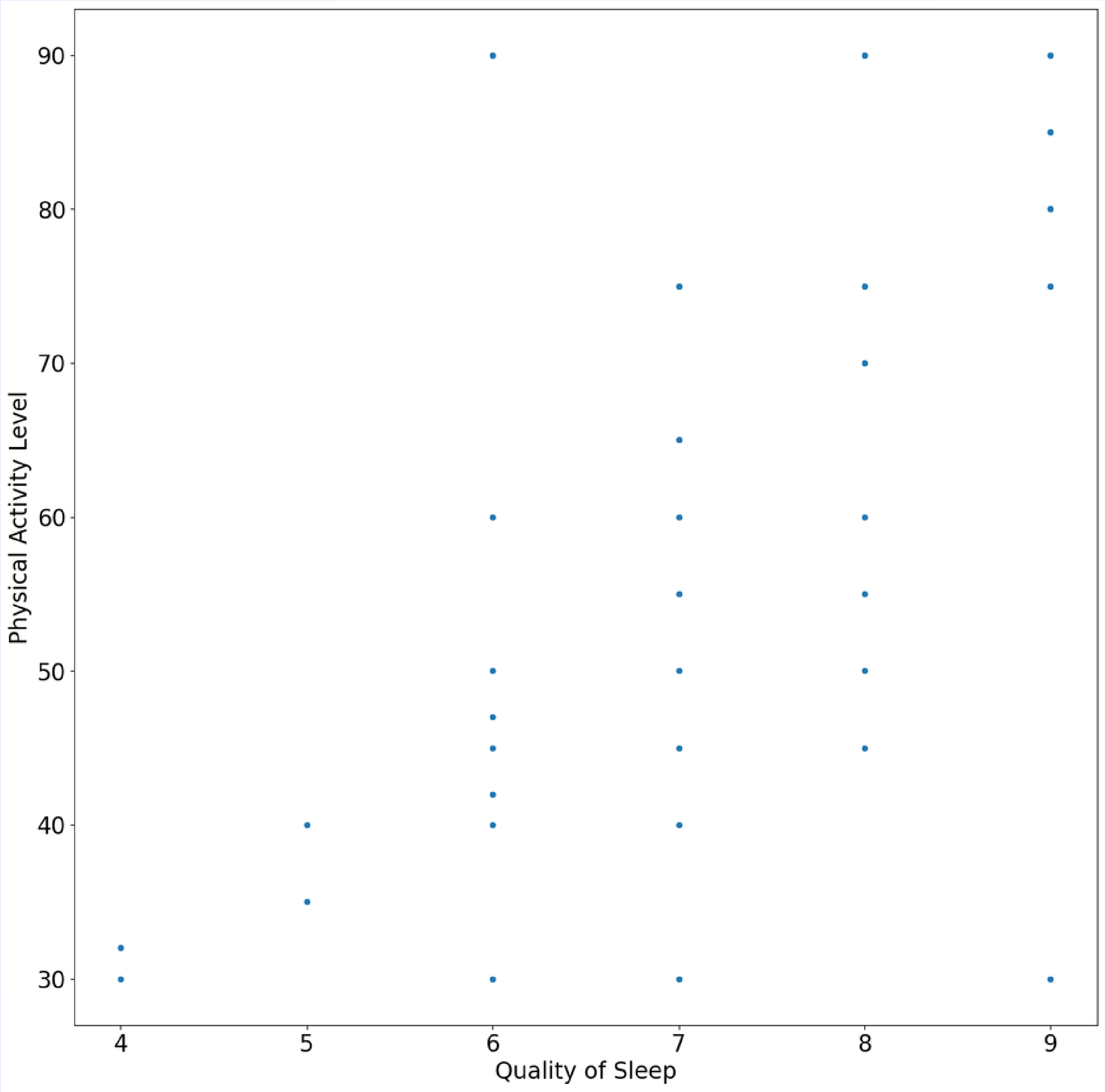
Sleep Duration - Sleep Level

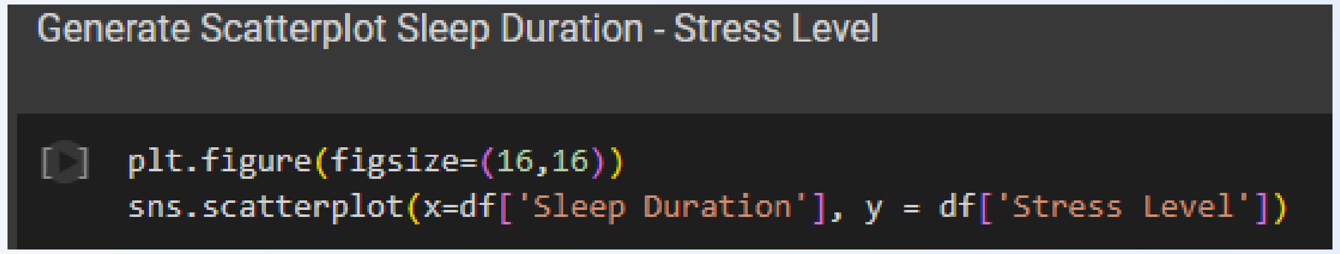
5) SCATTERPLOTS



Generate Scatterplot Quality of Sleep - Physical Activity Level

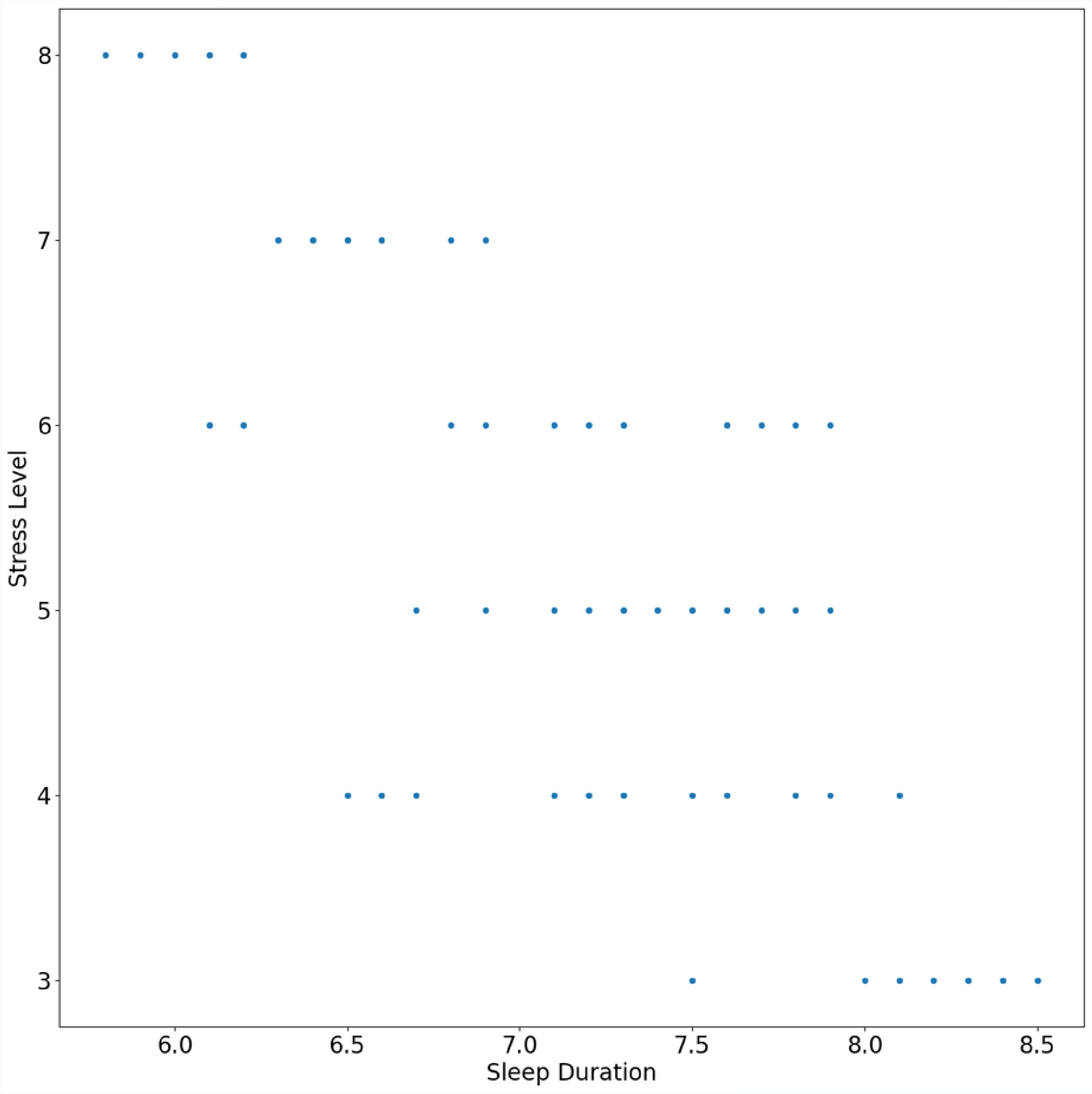
```
[ ] plt.figure(figsize=(16,16))
    sns.scatterplot(x=df['Quality of Sleep'], y = df['Physical Activity Level'])
```

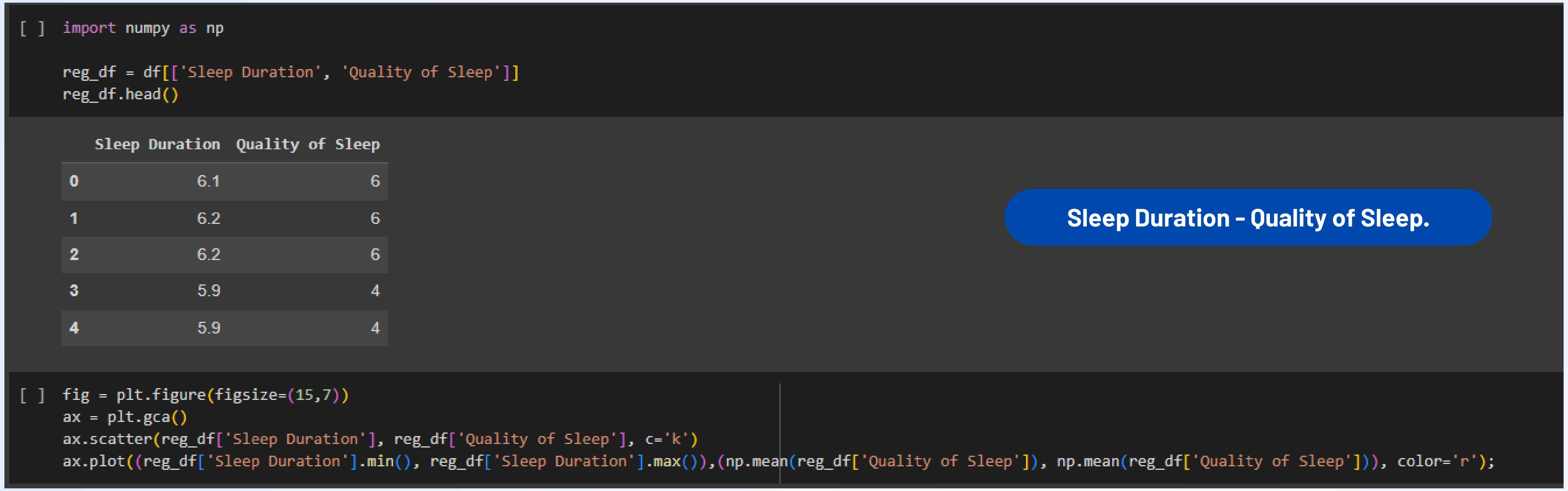




Generate Scatterplot Sleep Duration - Stress Level

```
plt.figure(figsize=(16,16))
sns.scatterplot(x=df['Sleep Duration'], y = df['Stress Level'])
```



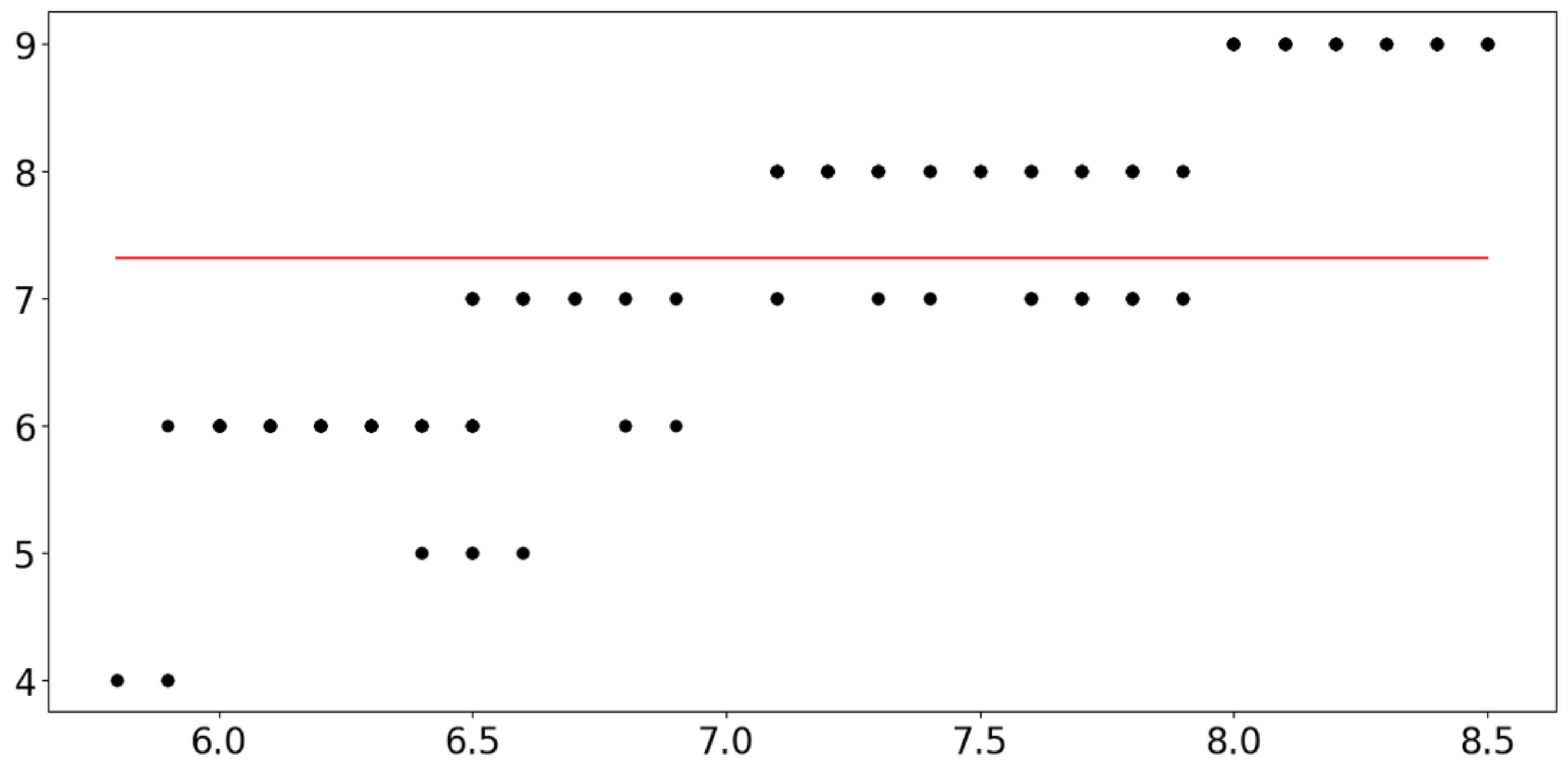


```
[ ]  import numpy as np

     reg_df = df[['Sleep Duration', 'Quality of Sleep']]
     reg_df.head()
```

	Sleep Duration	Quality of Sleep
0	6.1	6
1	6.2	6
2	6.2	6
3	5.9	4
4	5.9	4

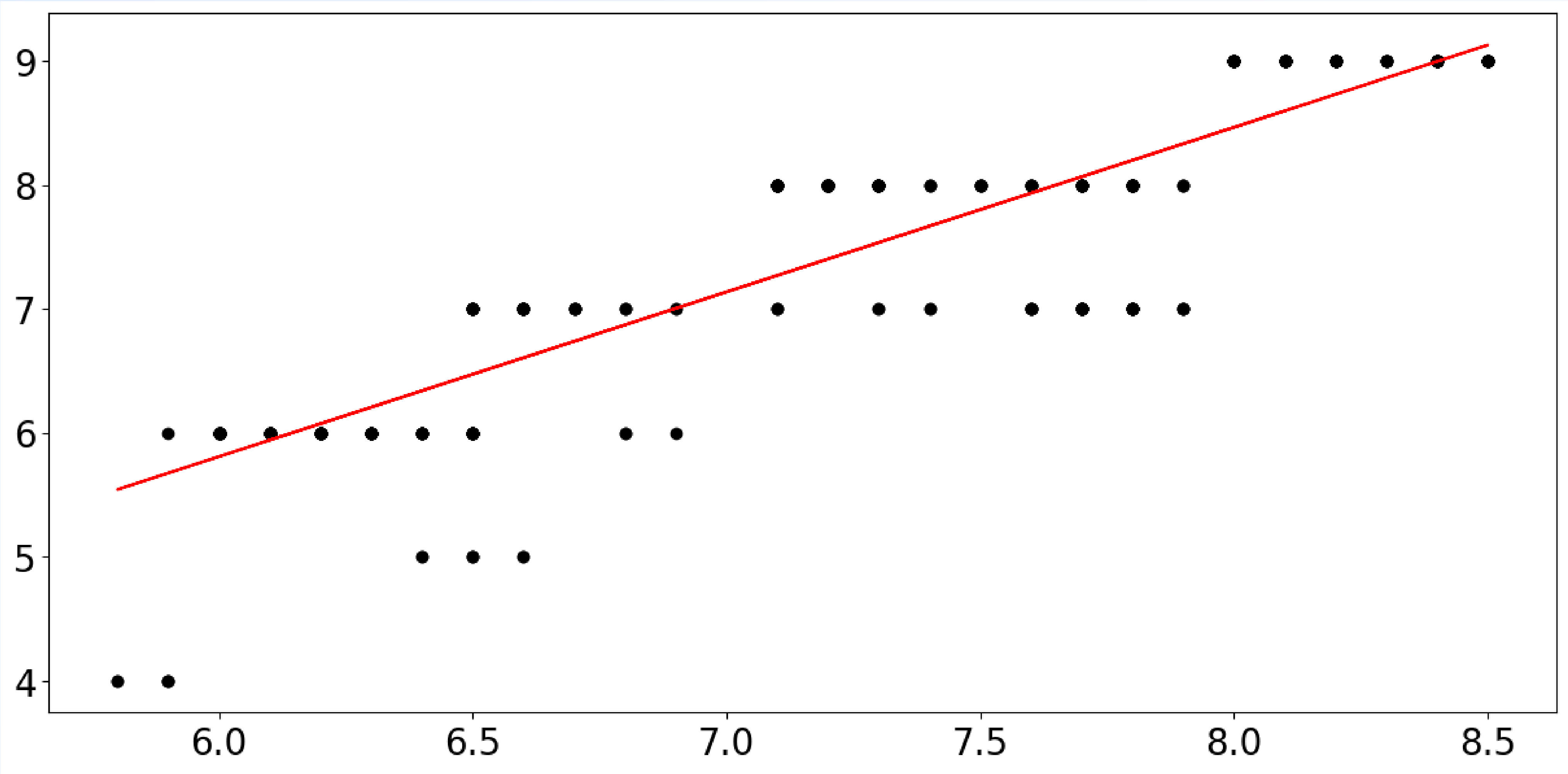
Sleep Duration - Quality of Sleep.

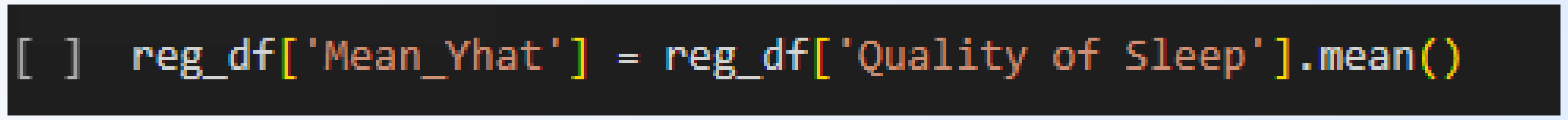
```
[ ]  fig = plt.figure(figsize=(15,7))
     ax = plt.gca()
     ax.scatter(reg_df['Sleep Duration'], reg_df['Quality of Sleep'], c='k')
     ax.plot((reg_df['Sleep Duration'].min(), reg_df['Sleep Duration'].max()),(np.mean(reg_df['Quality of Sleep']), np.mean(reg_df['Quality of Sleep'])), color='r');
```



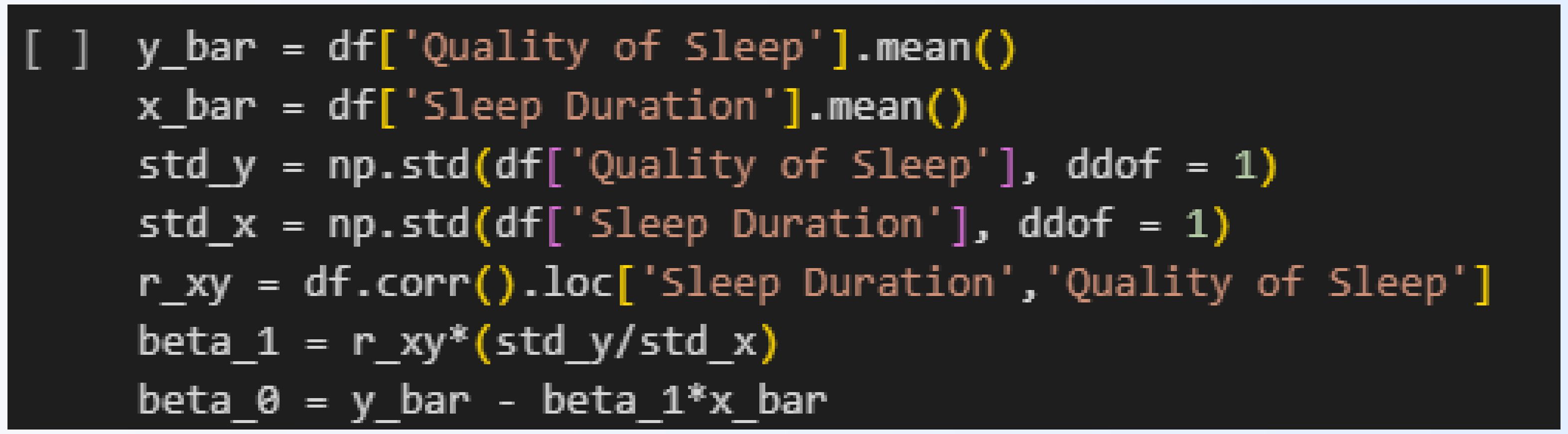
6) BASELINE PREDICTOR

7) SIMPLE LINEAR REGRESSION MODEL

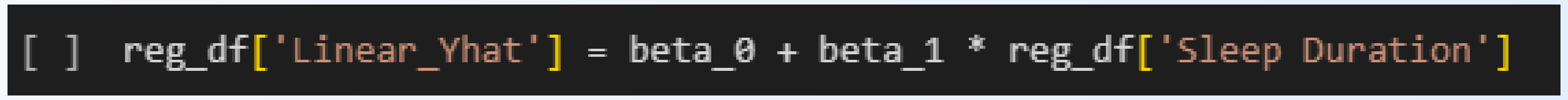




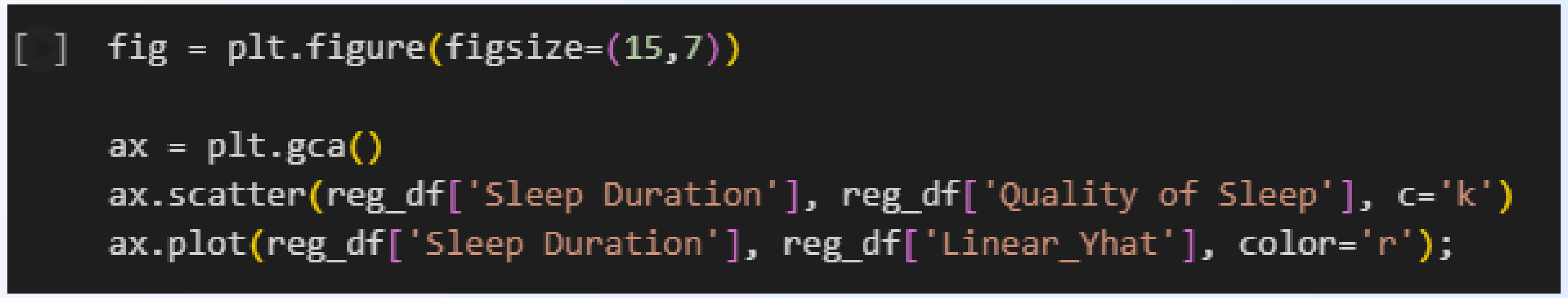
```
[ ]  reg_df['Mean_Yhat'] = reg_df['Quality of Sleep'].mean()
```



```
[ ]  y_bar = df['Quality of Sleep'].mean()
     x_bar = df['Sleep Duration'].mean()
     std_y = np.std(df['Quality of Sleep'], ddof = 1)
     std_x = np.std(df['Sleep Duration'], ddof = 1)
     r_xy = df.corr().loc['Sleep Duration','Quality of Sleep']
     beta_1 = r_xy*(std_y/std_x)
     beta_0 = y_bar - beta_1*x_bar
```

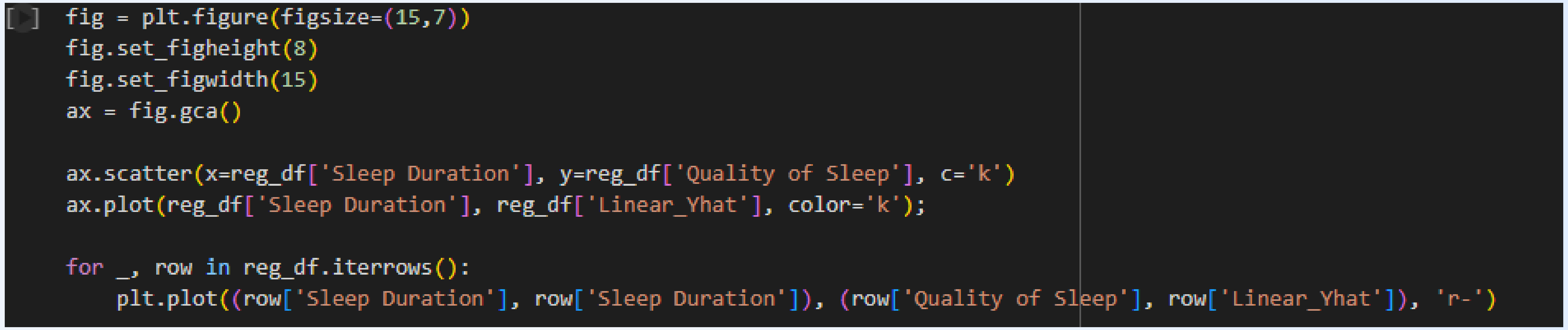


```
[ ]  reg_df['Linear_Yhat'] = beta_0 + beta_1 * reg_df['Sleep Duration']
```

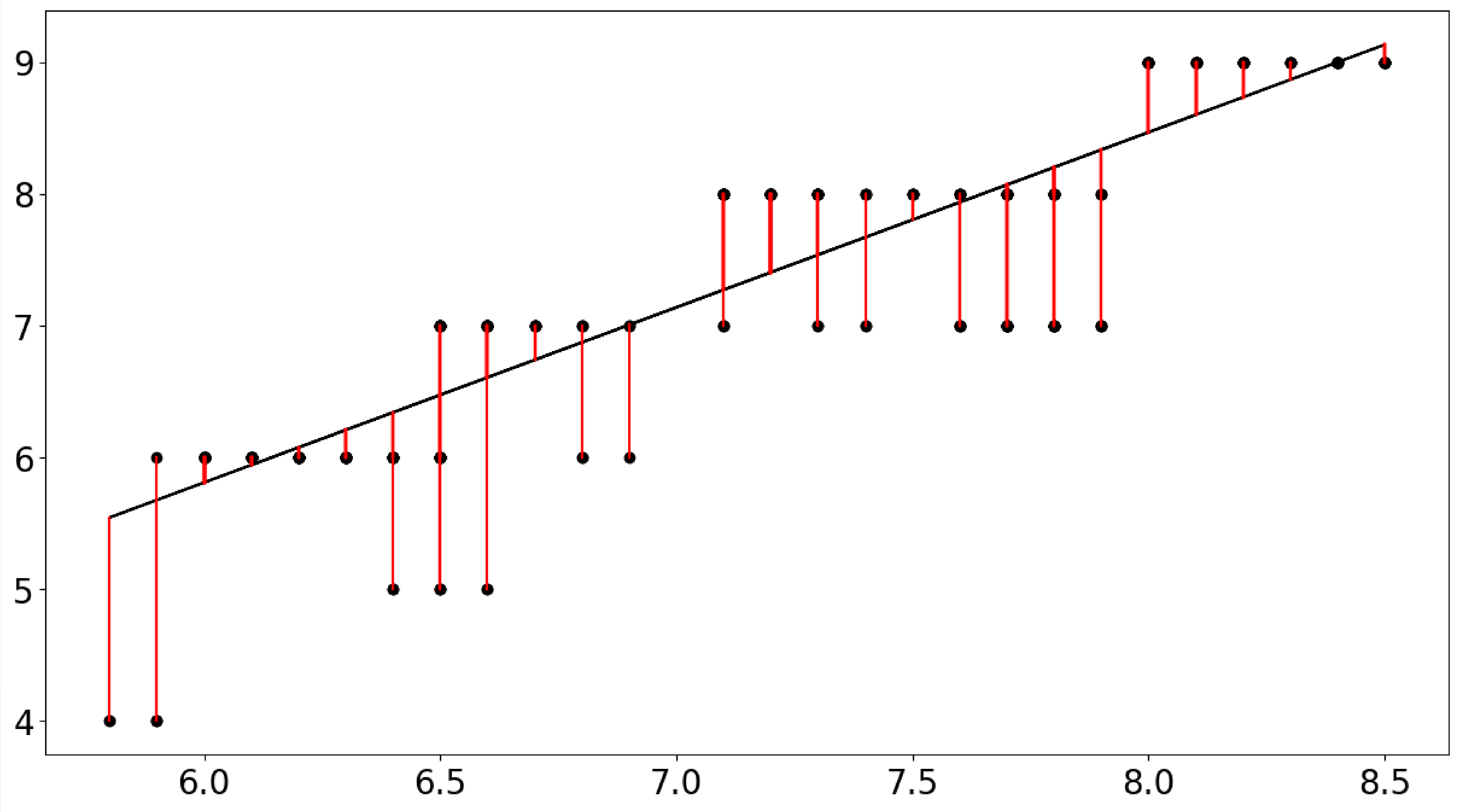


```
[ ]  fig = plt.figure(figsize=(15,7))

     ax = plt.gca()
     ax.scatter(reg_df['Sleep Duration'], reg_df['Quality of Sleep'], c='k')
     ax.plot(reg_df['Sleep Duration'], reg_df['Linear_Yhat'], color='r');
```

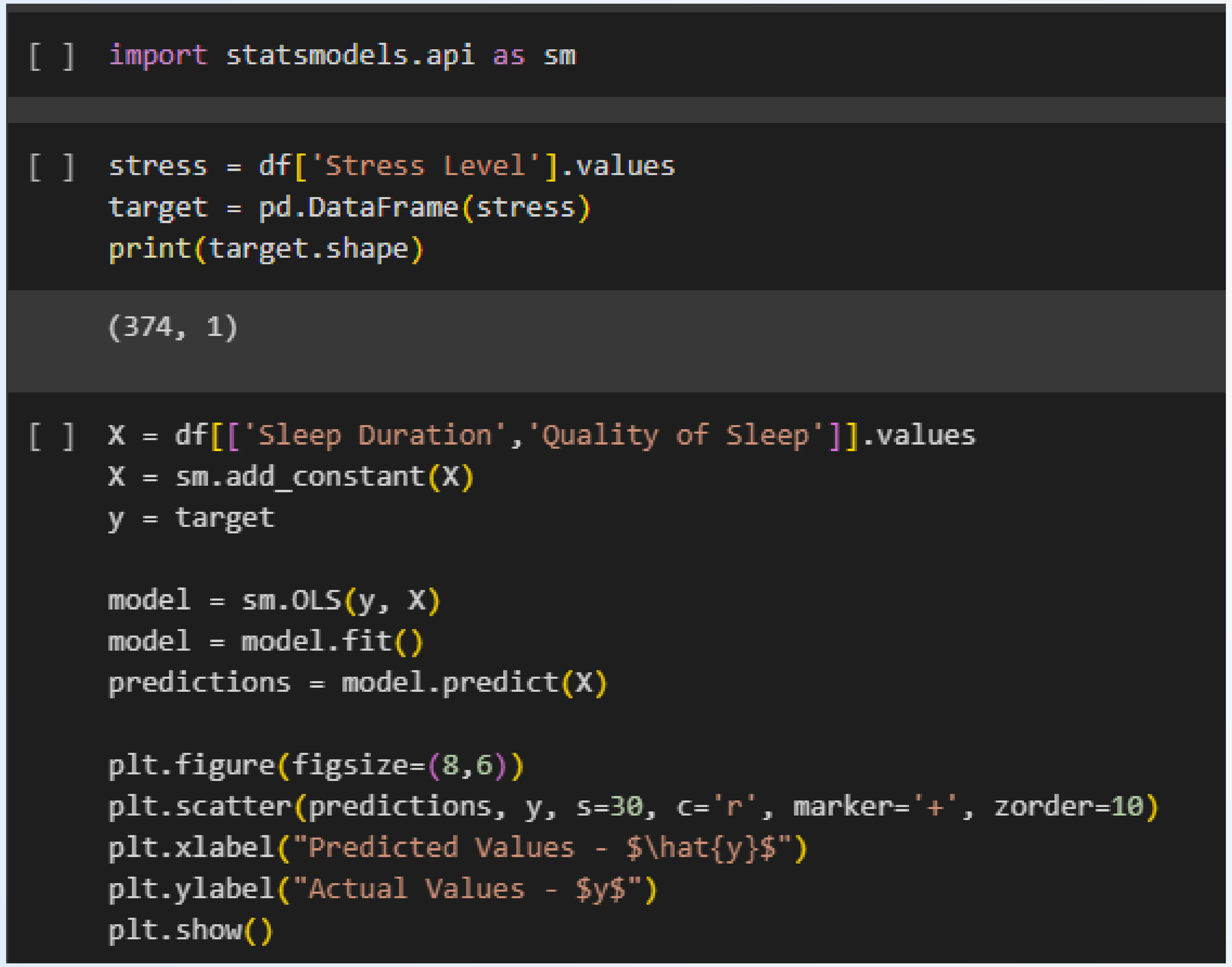
7) SIMPLE LINEAR REGRESSION MODEL



```
fig = plt.figure(figsize=(15,7))
fig.set_figheight(8)
fig.set_figwidth(15)
ax = fig.gca()

ax.scatter(x=reg_df['Sleep Duration'], y=reg_df['Quality of Sleep'], c='k')
ax.plot(reg_df['Sleep Duration'], reg_df['Linear_Yhat'], color='k');

for _, row in reg_df.iterrows():
    plt.plot((row['Sleep Duration'], row['Sleep Duration']), (row['Quality of Sleep'], row['Linear_Yhat']), 'r-')
```



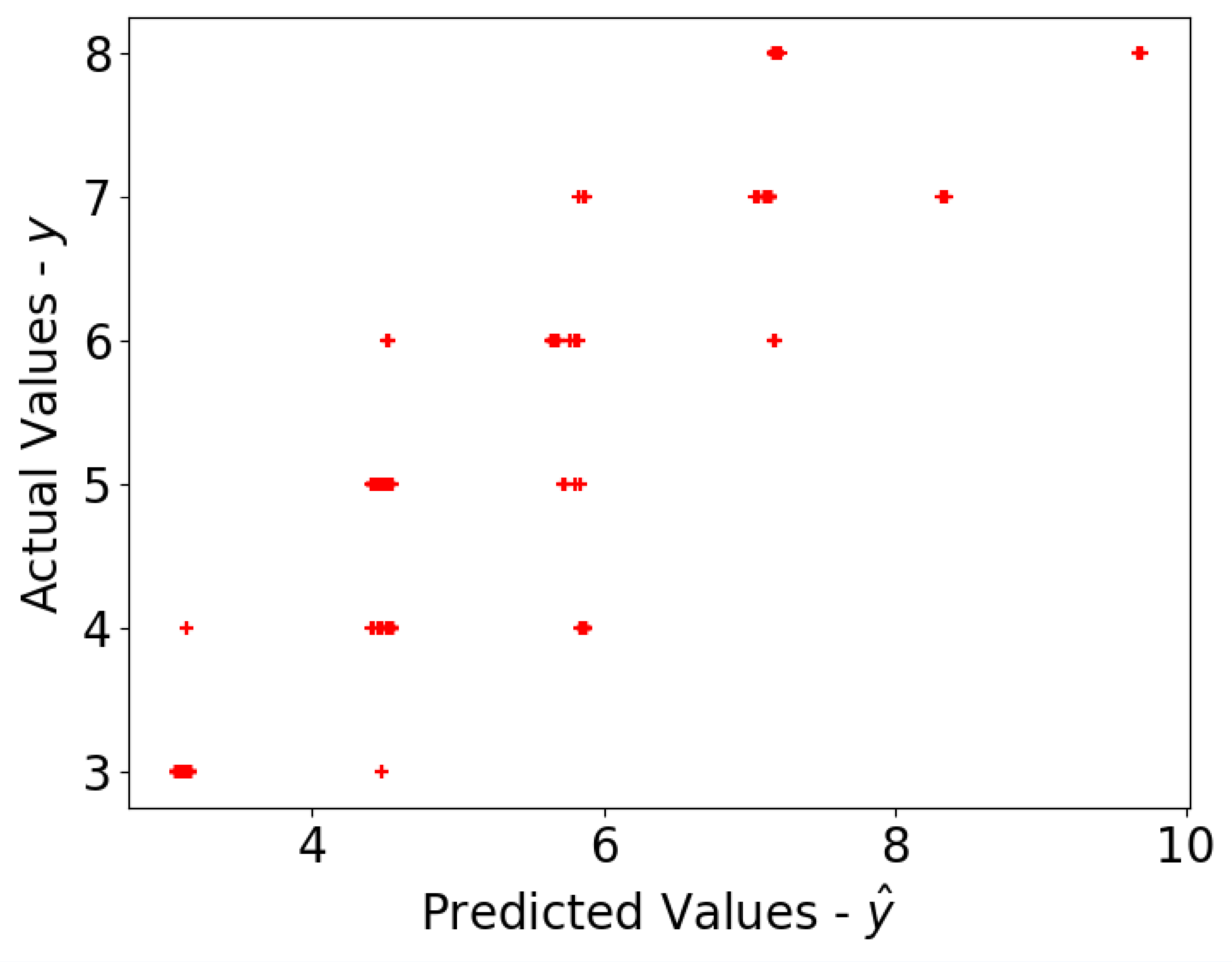
8) OLS REGRESSION MODEL



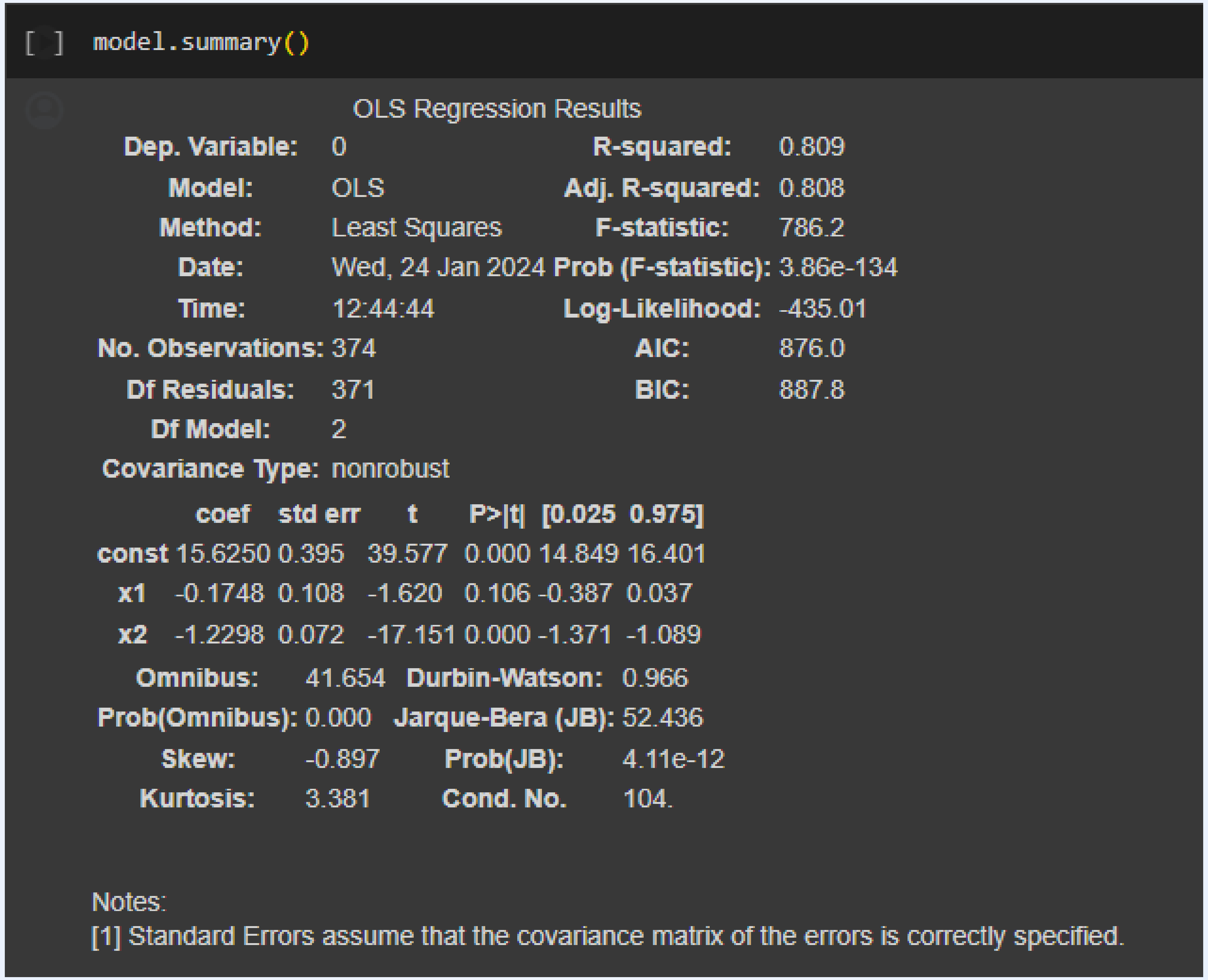
```
[ ]  import statsmodels.api as sm
```

```
[ ]  stress = df['Stress Level'].values
     target = pd.DataFrame(stress)
     print(target.shape)

     (374, 1)
```

```
[ ]  X = df[['Sleep Duration','Quality of Sleep']].values
     X = sm.add_constant(X)
     y = target

     model = sm.OLS(y, X)
     model = model.fit()
     predictions = model.predict(X)

     plt.figure(figsize=(8,6))
     plt.scatter(predictions, y, s=30, c='r', marker='+', zorder=10)
     plt.xlabel("Predicted Values - $\hat{y}$")
     plt.ylabel("Actual Values - $y$")
     plt.show()
```



8) OLS REGRESSION MODEL SUMMARY



```
[ ]  model.summary()
```

```
                        OLS Regression Results
  Dep. Variable:     0                        R-squared:      0.809
      Model:         OLS                  Adj. R-squared:     0.808
     Method:         Least Squares           F-statistic:     786.2
      Date:          Wed, 24 Jan 2024 Prob (F-statistic): 3.86e-134
      Time:          12:44:44             Log-Likelihood:     -435.01
No. Observations:    374                         AIC:         876.0
  Df Residuals:      371                         BIC:         887.8
    Df Model:        2
Covariance Type:     nonrobust

           coef    std err    t      P>|t|  [0.025   0.975]
const 15.6250  0.395   39.577  0.000 14.849 16.401
  x1   -0.1748  0.108   -1.620   0.106 -0.387  0.037
  x2   -1.2298  0.072   -17.151 0.000 -1.371  -1.089

     Omnibus:      41.654  Durbin-Watson:   0.966
Prob(Omnibus): 0.000   Jarque-Bera (JB): 52.436
       Skew:       -0.897       Prob(JB):     4.11e-12
     Kurtosis:     3.381        Cond. No.     104.


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

RESULTS & FINDINGS

Main Findings

*The main findings suggest that **the model demonstrated that approximately 80.9% of the variance in sleep quality could be explained by sleep duration and stress level.** Stress level was found to be a significant predictor of sleep quality.*

*This study's findings underscore the significant impact of stress on sleep quality, highlighting the **need for stress management interventions as part of a healthy lifestyle.***

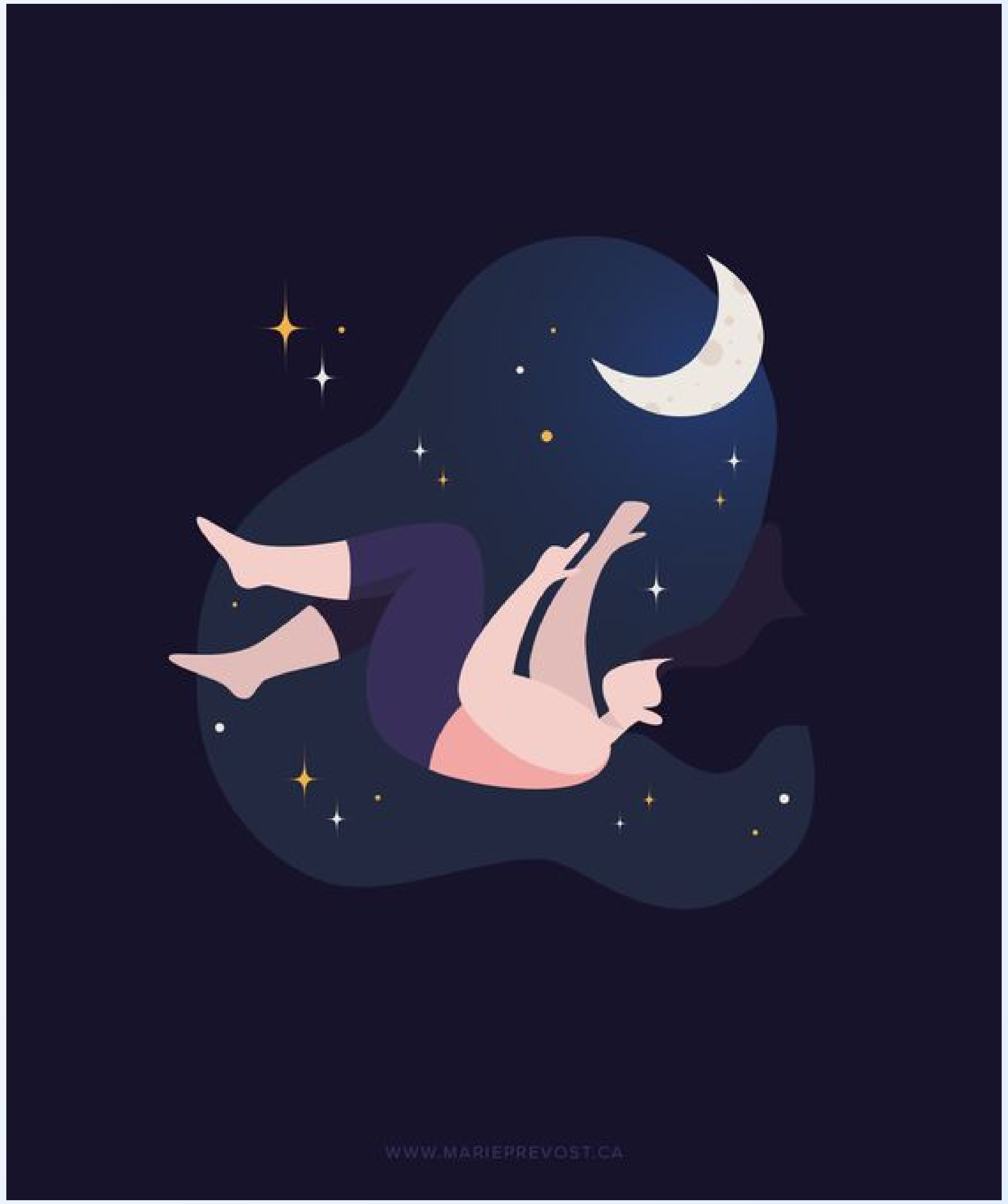
The pairplot indicates that sleep quality and physical activity level are generally positively correlated, with higher sleep quality being linked to higher levels of physical activity

RESULTS & FINDINGS

Limitations

Despite the strong model fit, **the lack of normality in the residuals suggests that future research should incorporate a broader range of variables** to fully capture the determinants of sleep quality.

The limitations of the research are mainly on the limited access to data and the lack of previous numerous researches on this topic, which needs to be further investigated due to the fact that human nature, behavior and lifestyle is changing due to the rapid changes in the environment.



THANK YOU!

Professors:

Petre Caraini
Daniel Traian Pele
Wolfgang Hardle

Research Methods - Project
IMBA, 2nd Year, 2023-2024

Team Members:

Torjescu Ana-Maria
Dumitru Laura-Alexandra
Costea Cristina-Bianca