**MACHINE LEARNING LAB**

*PROGRAM 3 – KNN Classifier*

*Domain – Students Sleep Pattern*

Name: Merry Don

Class: 6 BCA B

Register: 2241142

**ABOUT THE DATASET**

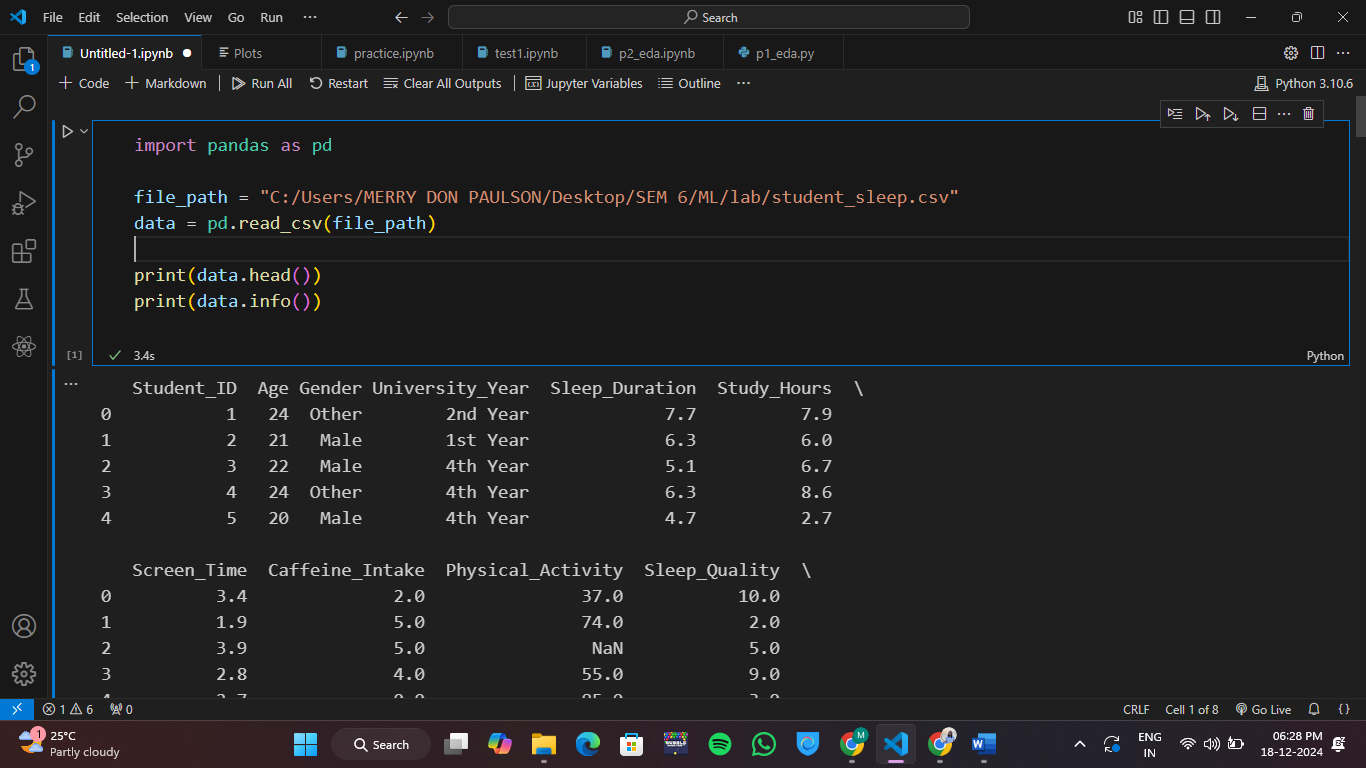
The dataset contains information about students' sleep patterns and associated lifestyle factors. It includes 500 entries with 14 columns. Each row represents an individual student, providing details about their demographics, habits, and sleep-related data.

**OBJECTIVE**

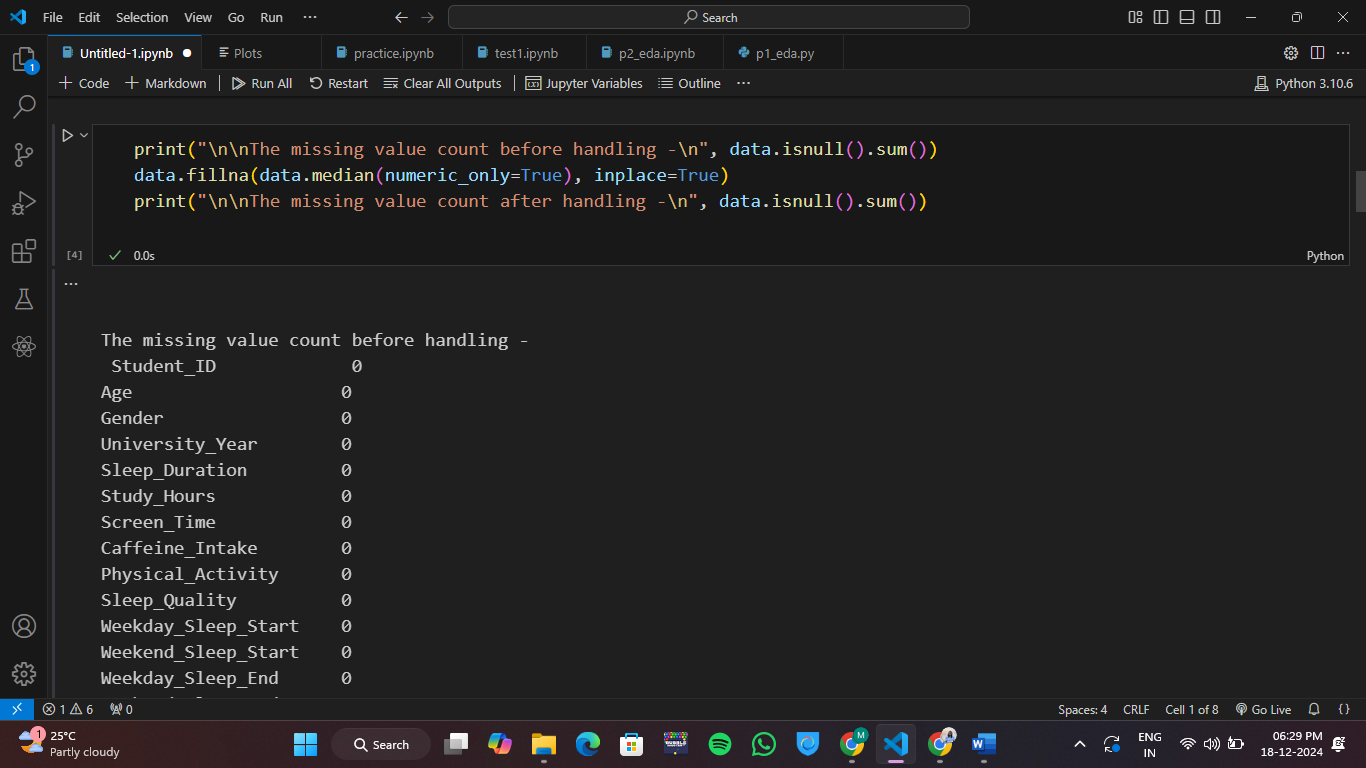
The objective of performing cross-validation on KK and displaying the performance metrics is to determine the optimal number of neighbors (KK) for a K-Nearest Neighbors (KNN) model. By evaluating the model's performance across different KK-values using metrics like RMSE (Root Mean Squared Error) or accuracy, the goal is to identify the KK that generalizes well to unseen data. This ensures the model achieves a balance between underfitting (KK too small) and overfitting (KK too large), resulting in robust predictions.

**STEPS**

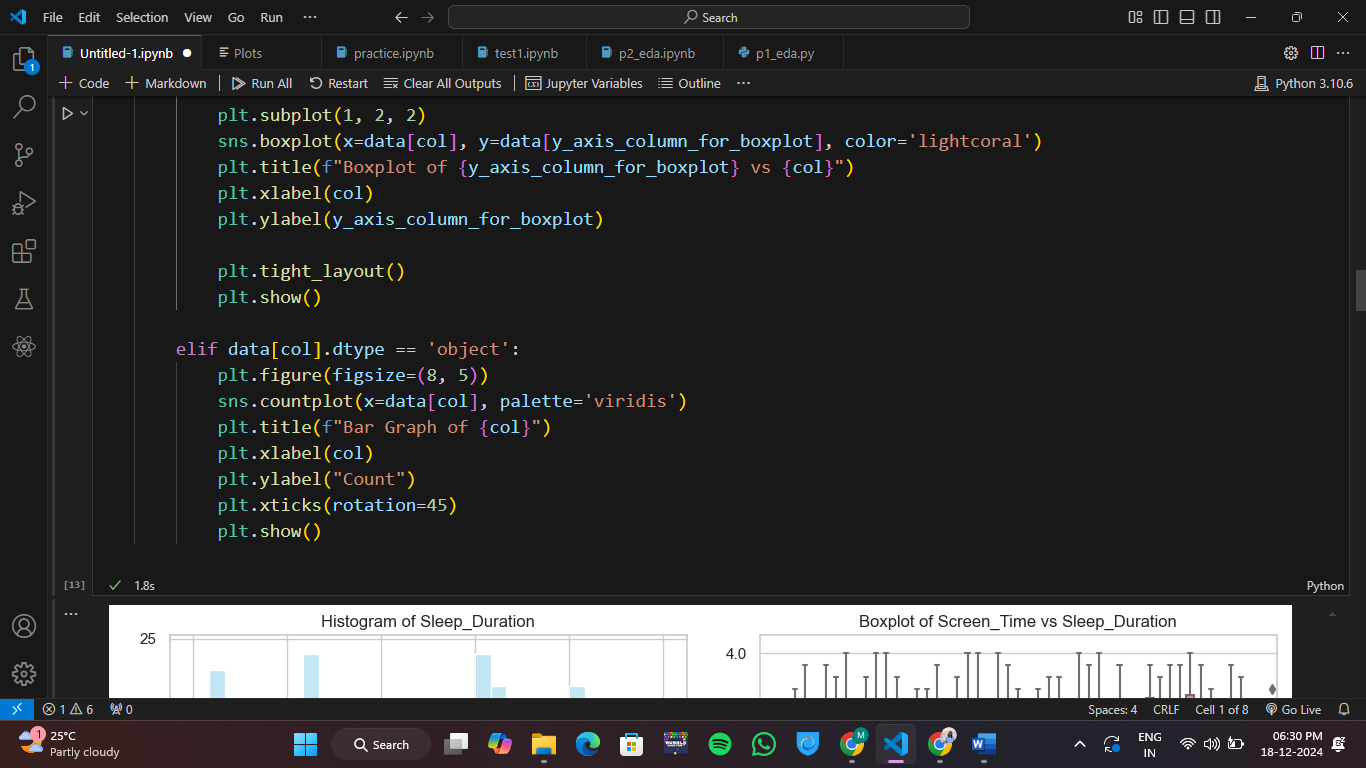
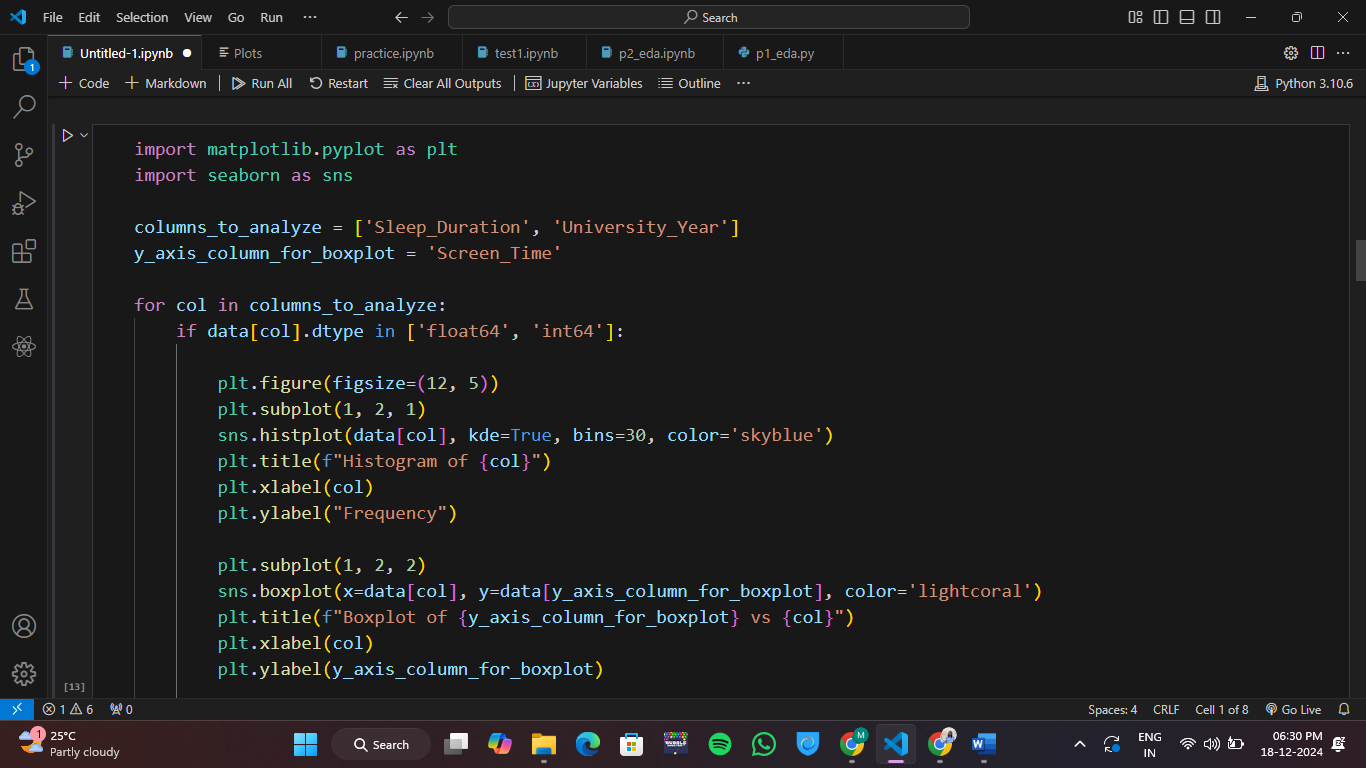
1. Importing Libraries and Loading the dataset:



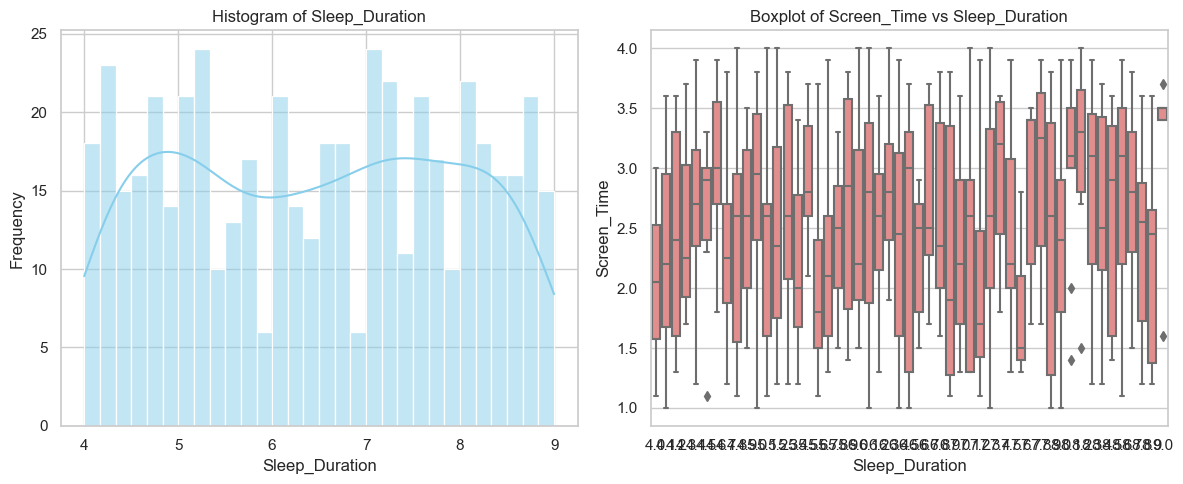
1. Handling the missing values and displaying it after :

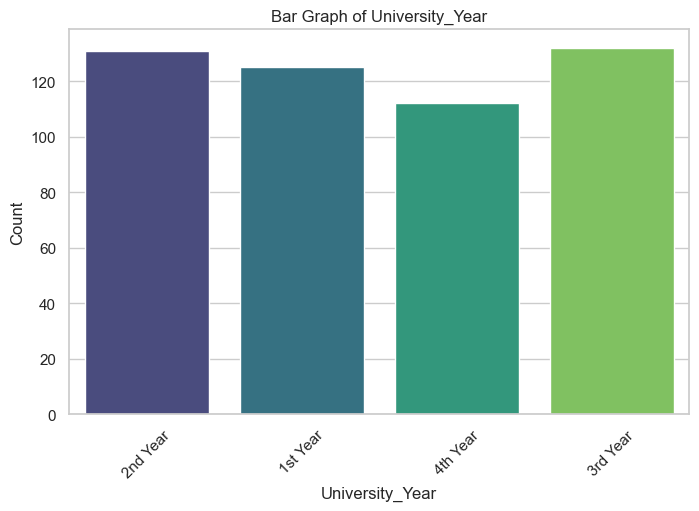


1. EDA:

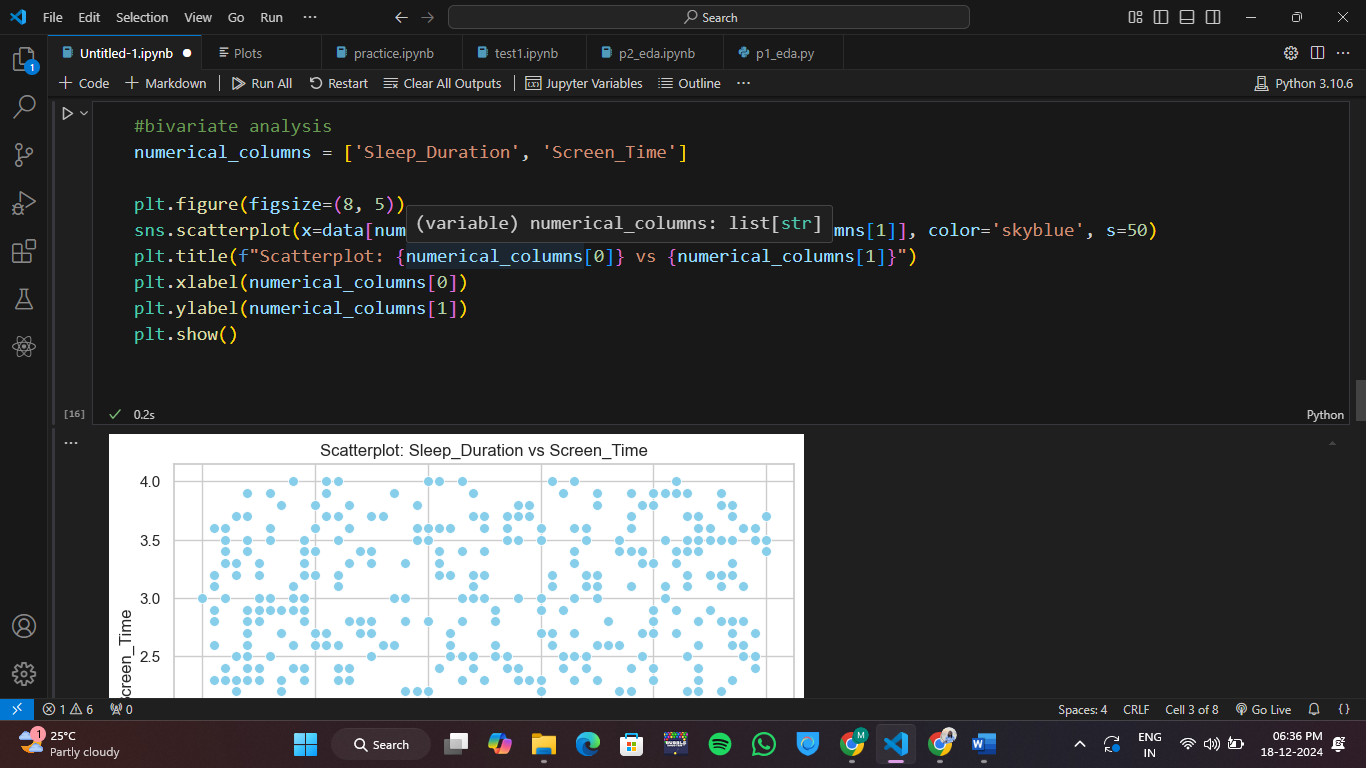


Output –

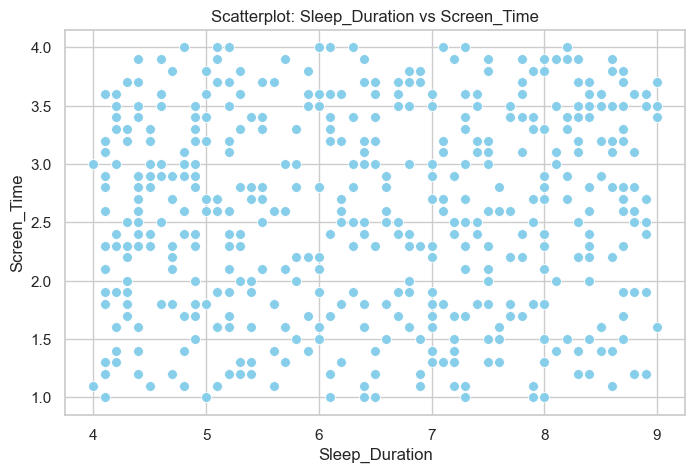




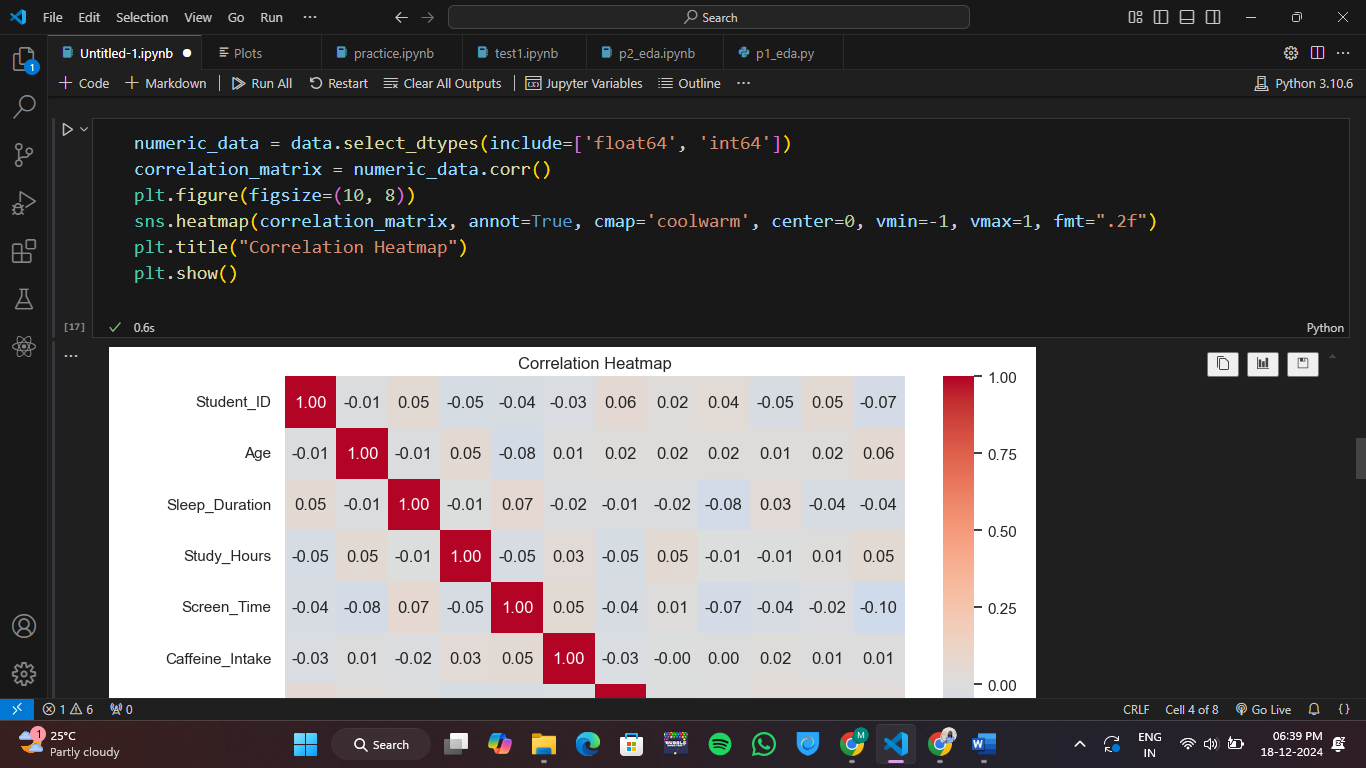
**Interpretation - The histogram of Sleep\_Duration shows a relatively uniform distribution of sleep durations between 4 and 9 hours, with slight peaks around 4.5, 6, and 8 hours. The boxplot of Screen\_Time vs Sleep\_Duration reveals that screen time generally ranges between 1 and 4 hours, with noticeable variability and a few outliers across different sleep durations. Overall, no strong trend is evident between sleep duration and screen time. The bar graph shows the distribution of students across different university years. The 3rd Year and 2nd Year have the highest counts, both exceeding 130 students, while the 4th Year has the lowest count, around 110 students. Overall, the student distribution appears fairly balanced across the years, with slight variations.**



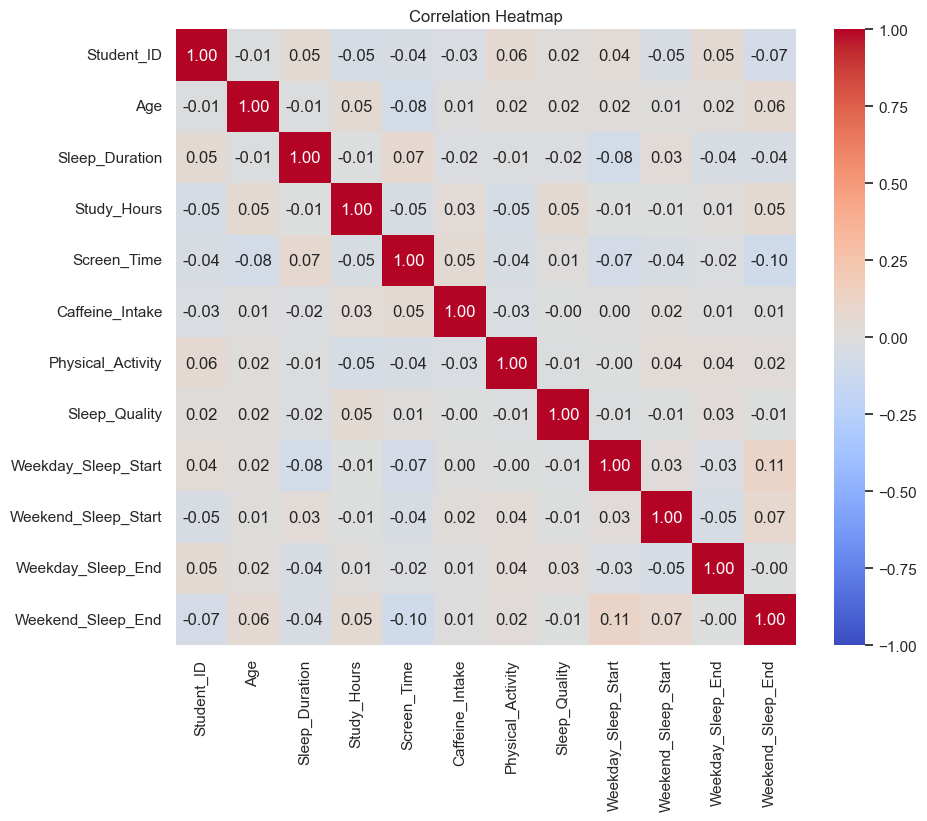
Output –



**Interpretation - The scatterplot shows no clear relationship between Sleep\_Duration and Screen\_Time, as the data points are widely scattered. Sleep duration ranges from 4 to 9 hours, while screen time varies between 1 to 4 hours. The absence of a pattern suggests no strong correlation between the two variables**.

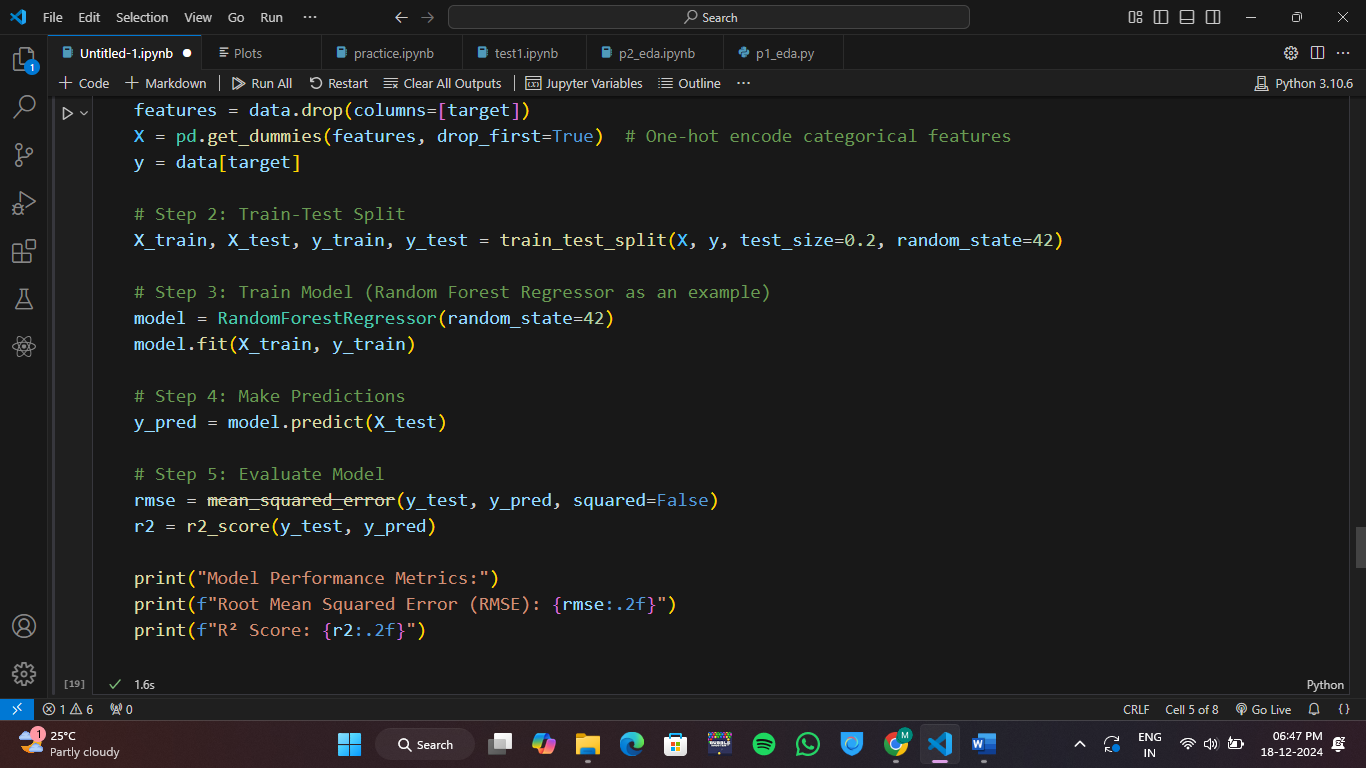
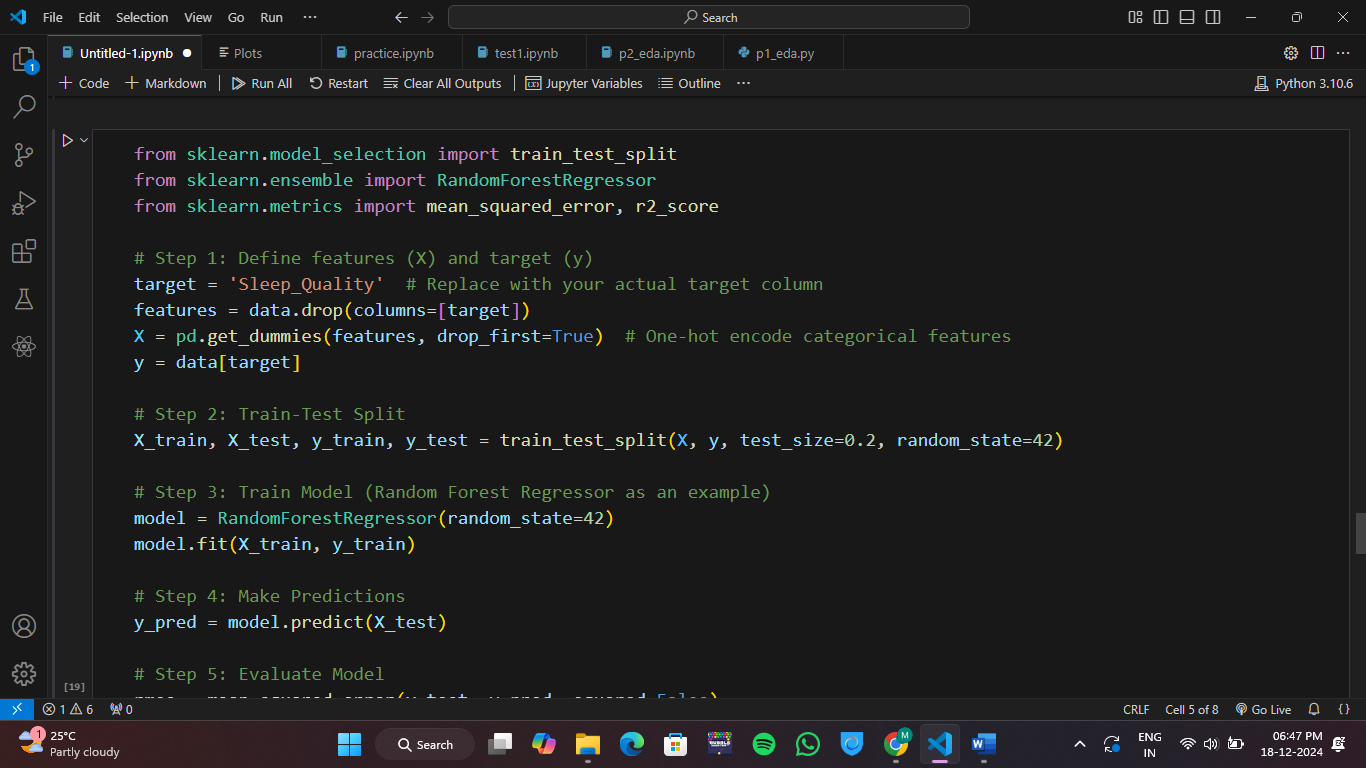


Output –

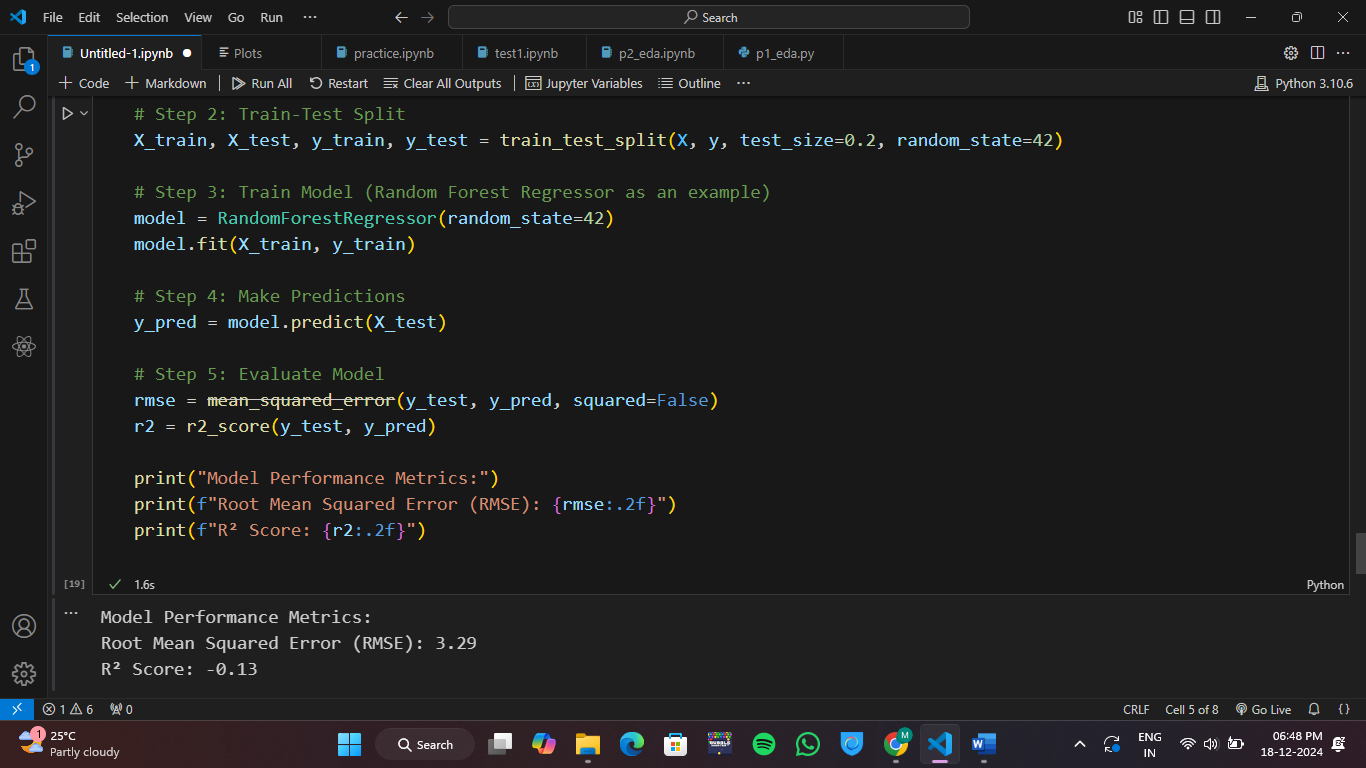
****

**Interpretation - The correlation heatmap shows weak or no linear relationships between variables. Notably, Weekday\_Sleep\_Start and Weekend\_Sleep\_End have a weak positive correlation (~0.11), while Screen\_Time has a slight negative correlation with Sleep\_Duration (-0.07). Most other variables are largely independent.**

1. Train Train-test split & Training model:



Output –

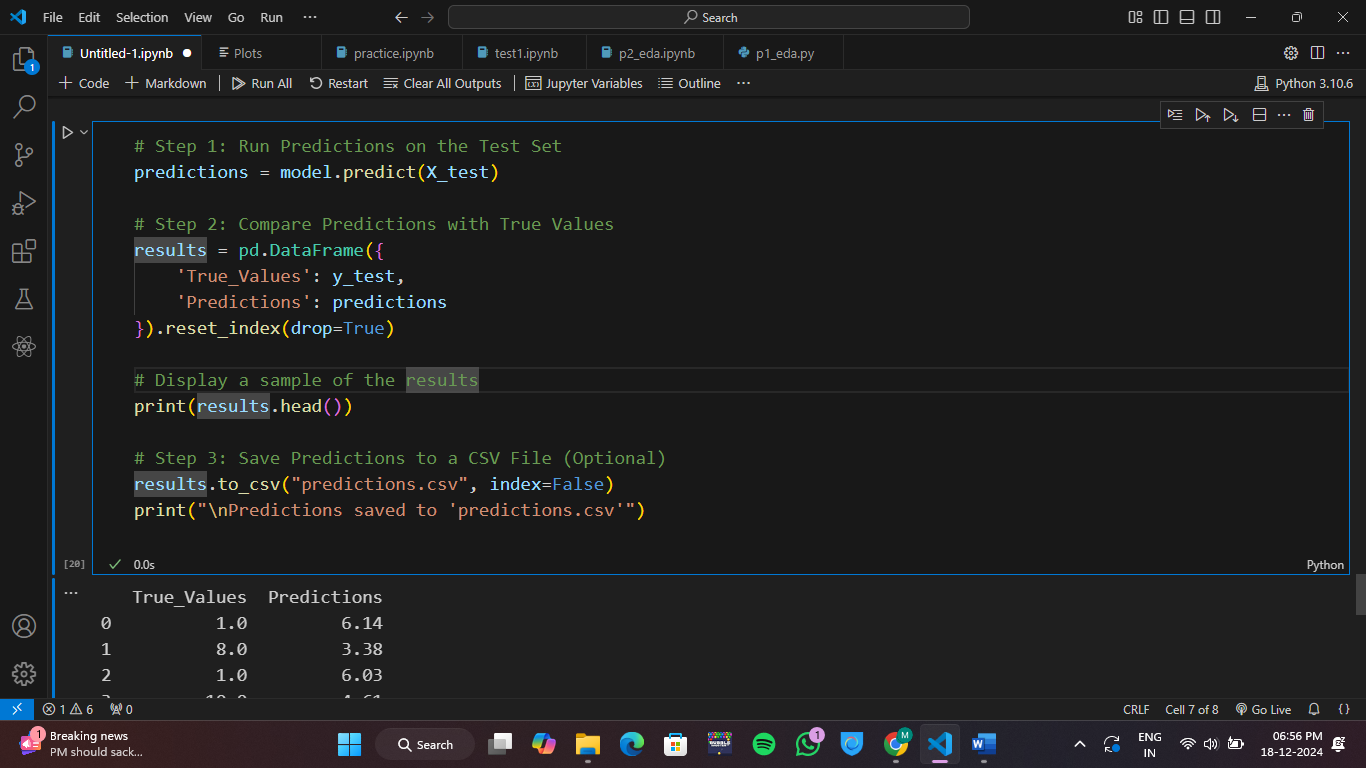


**Interpretation – The model performance metrics indicate poor performance:**

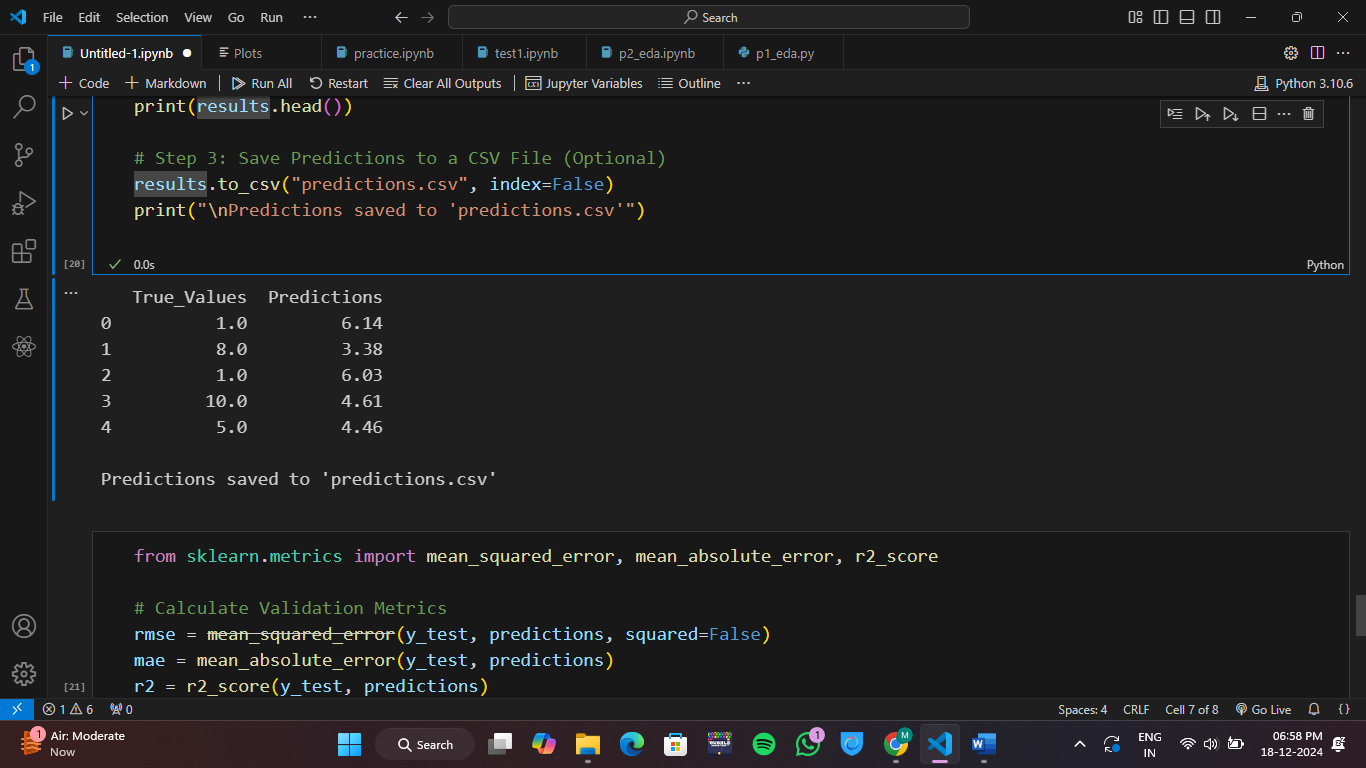
1. **RMSE (3.29): The model's predictions deviate significantly from the true values on average.**
2. **R² Score (-0.13): A negative R² score means the model performs worse than a horizontal line (mean value), showing it fails to explain the variance in the data.**

**The model likely suffers from underfitting and may need better features, tuning, or a different algorithm.**

1. Running prediction:



Output –



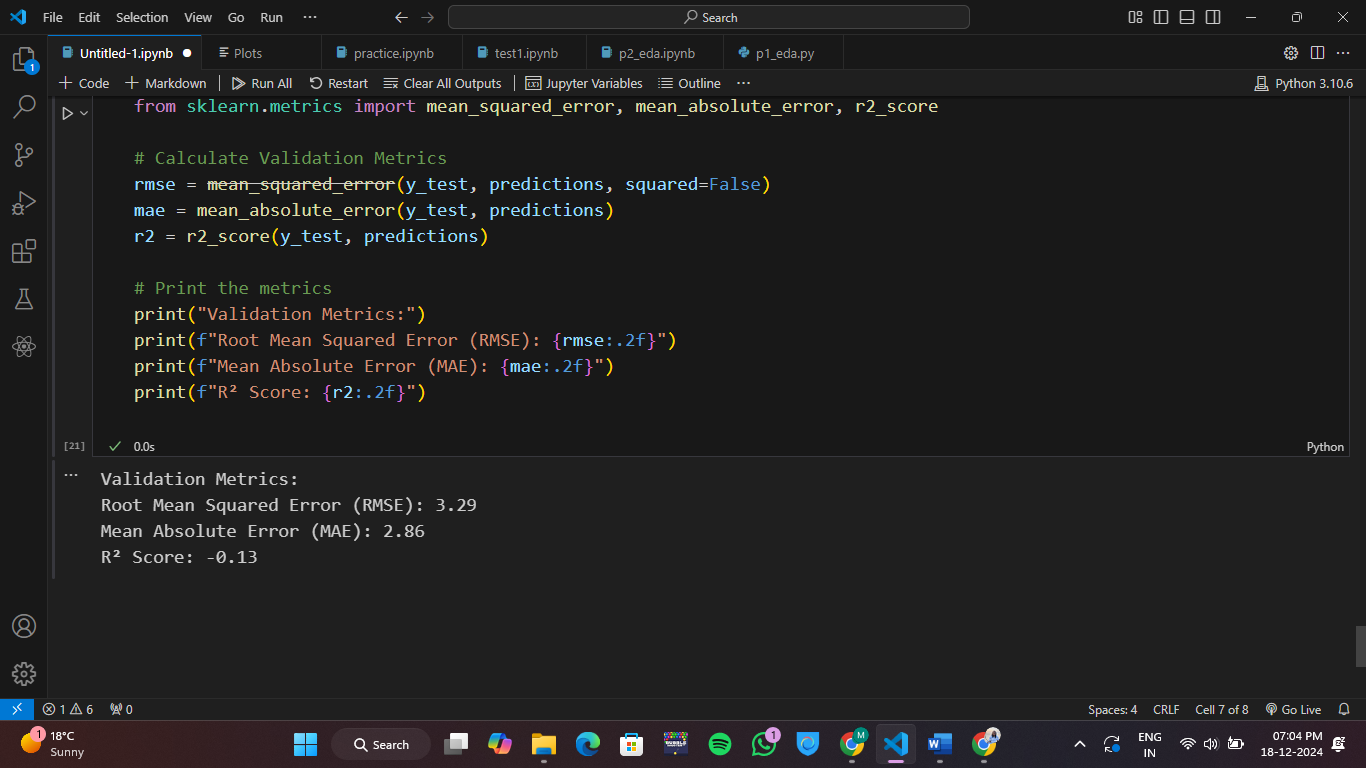
**Interpretation – 1) The predictions made by the model are noticeably different from the true values.**

**2) In most cases, the predictions tend to underestimate or overestimate the actual values:**

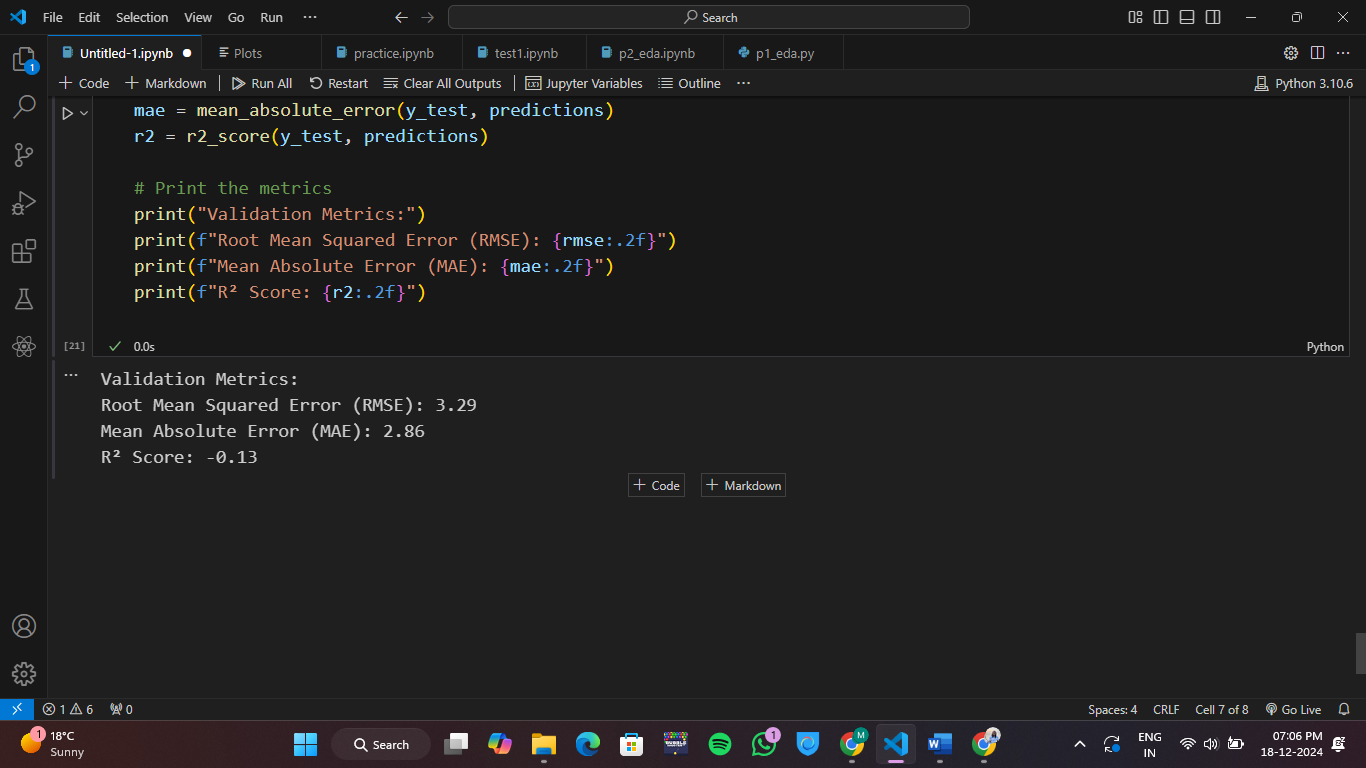
* **For index 0, the prediction (6.14) is far higher than the true value (1.0)**
* **For index 1, the prediction (3.38) significantly underestimates the true value (8.0).**
* **Similar trends are observed for the rest of the rows.**

**3) The output indicates that the predictions have been saved to a file named predictions.csv, suggesting that the results are stored for further analysis.**

1. Check validation:



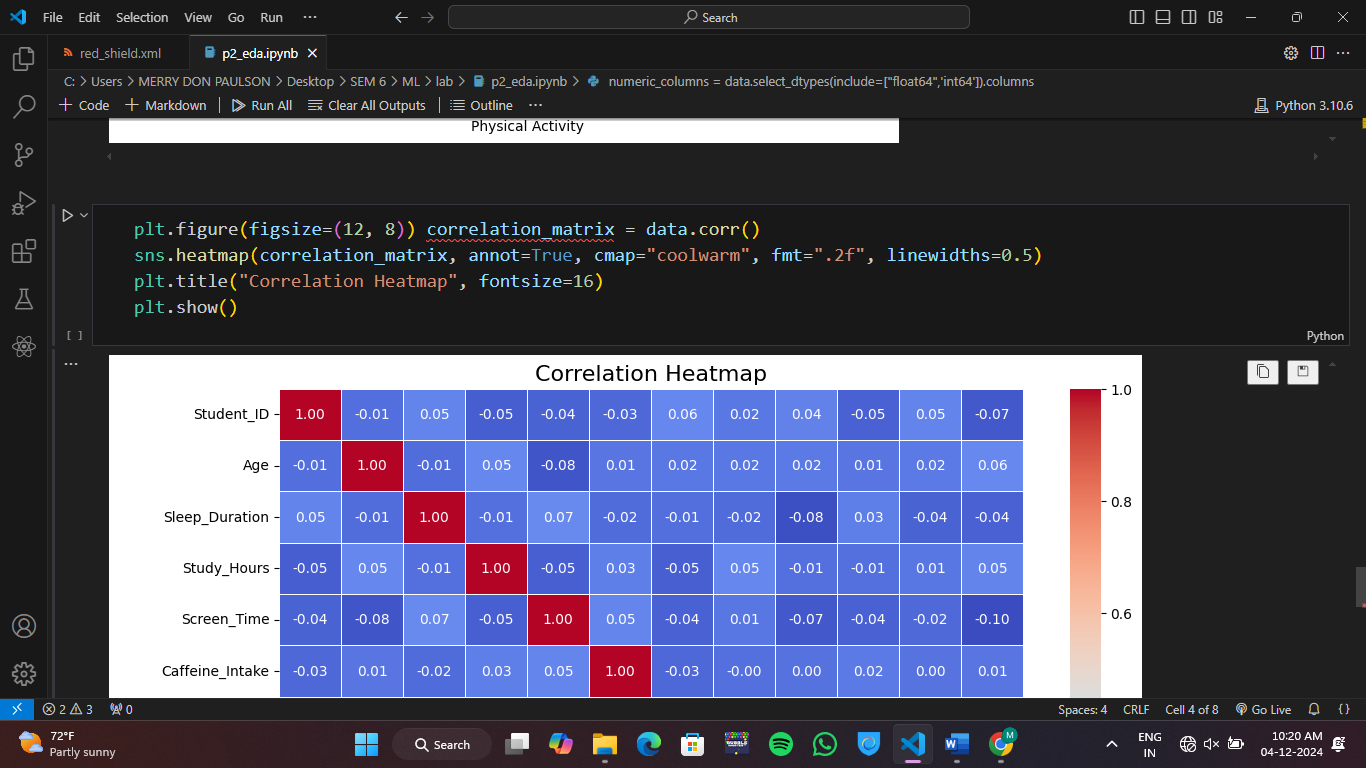
Output –



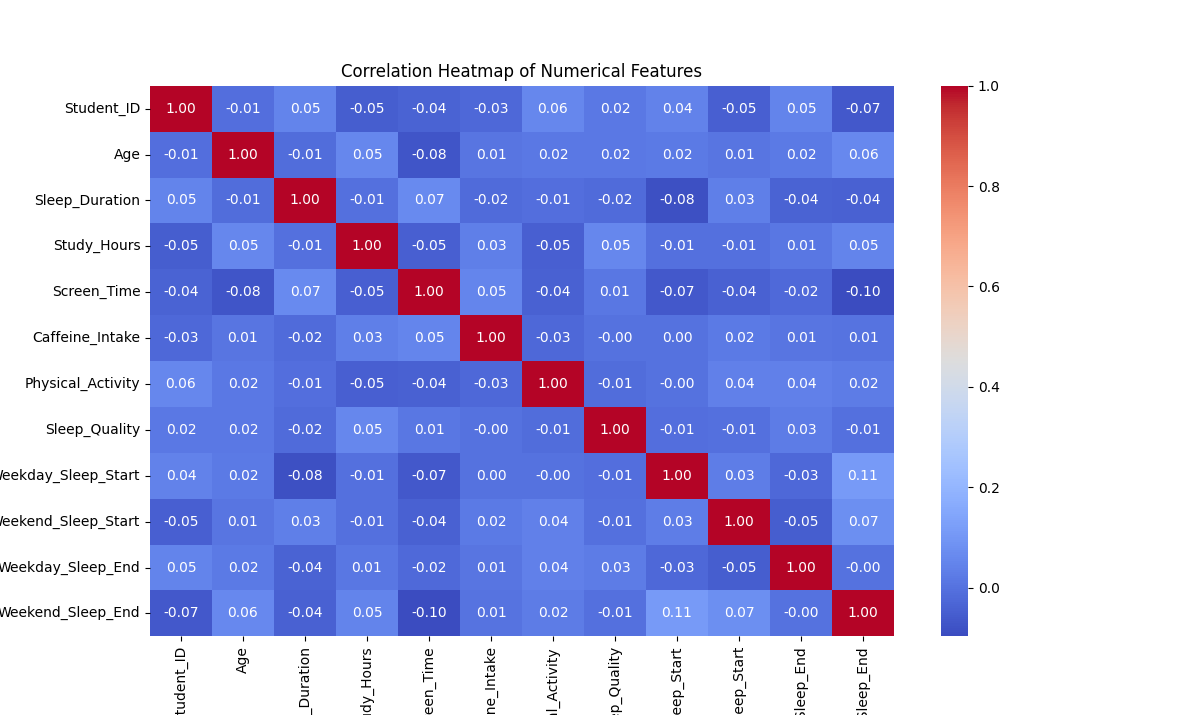
**Interpretation – Model Performance:** Poor.

* **RMSE & MAE: High, indicating large prediction errors.**
* **R² Score: Negative, suggesting the model is worse than predicting the average.**

1. Heatmap for correlation:

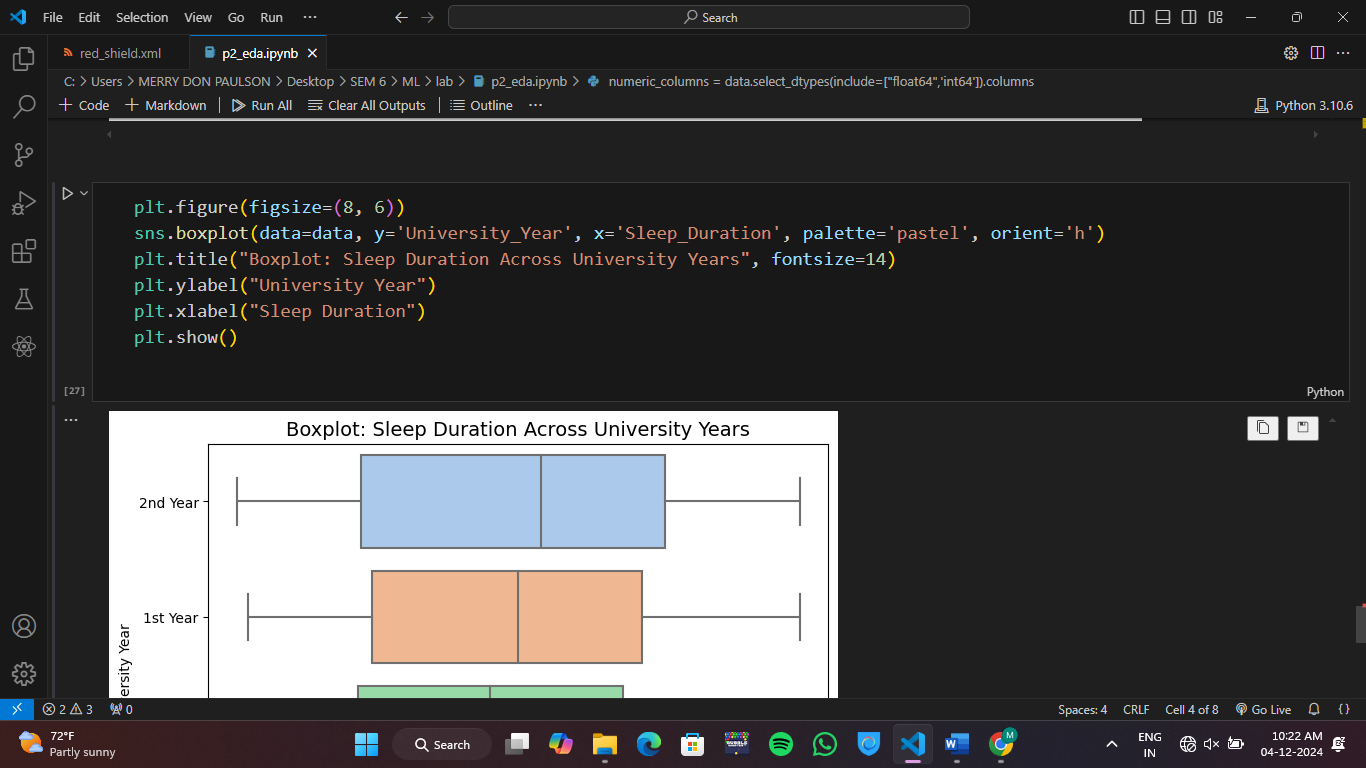


Output –

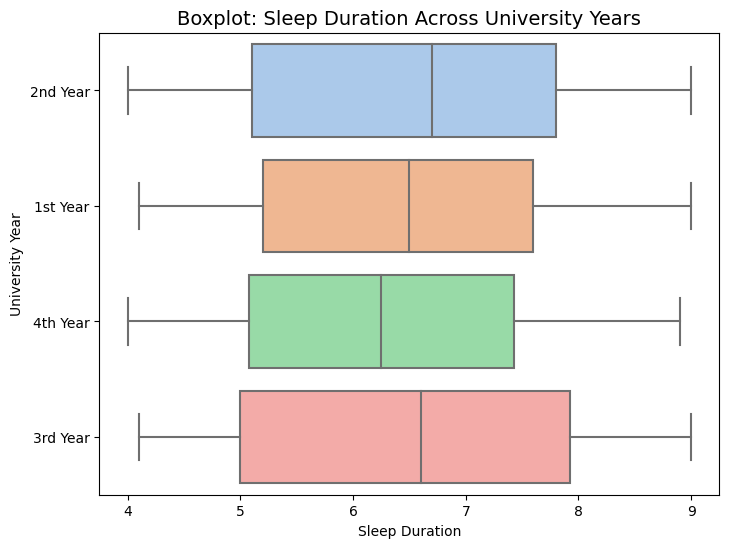


**Interpretation – Most of the correlations are weak. Sleep quality shows a low positive correlation with physical activity and sleep duration which hints that these slightly influence better sleep.**

1. Boxplot to compare values within groups:

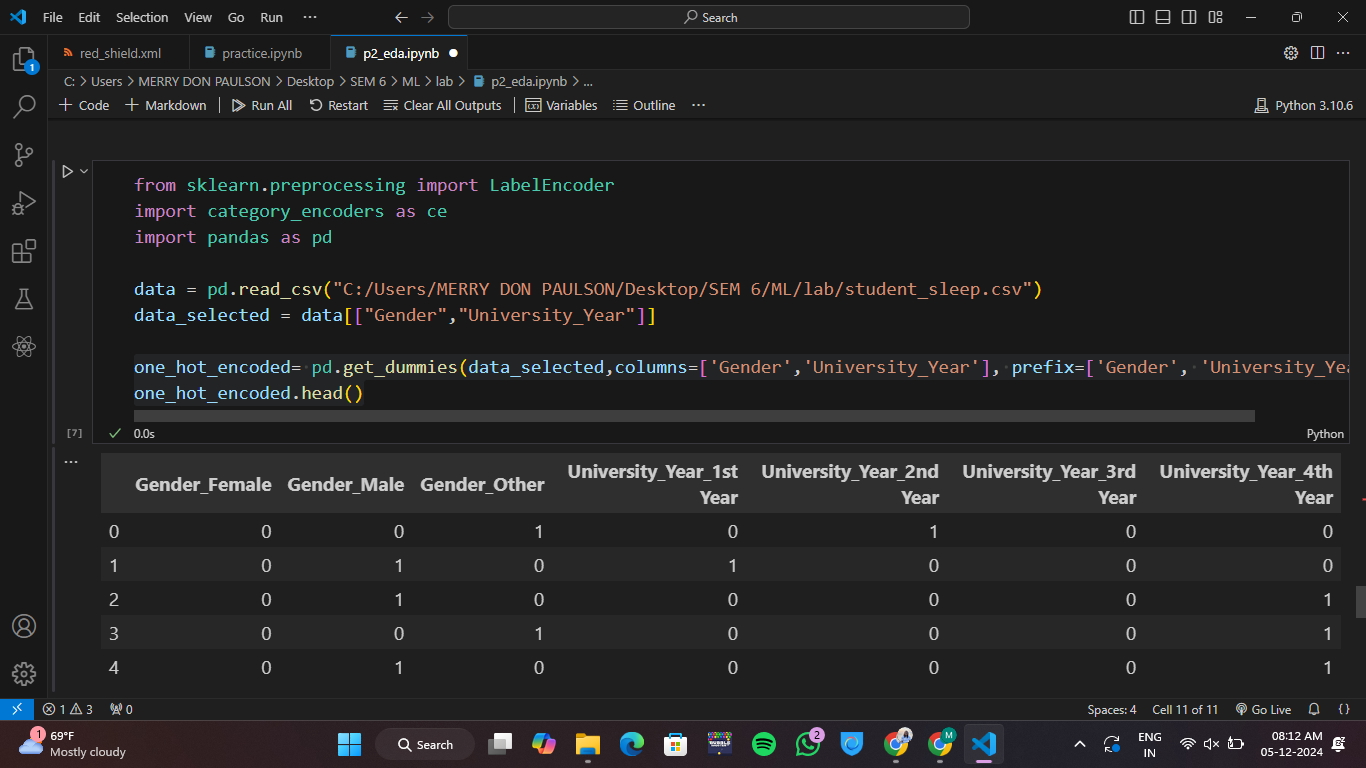


Output –

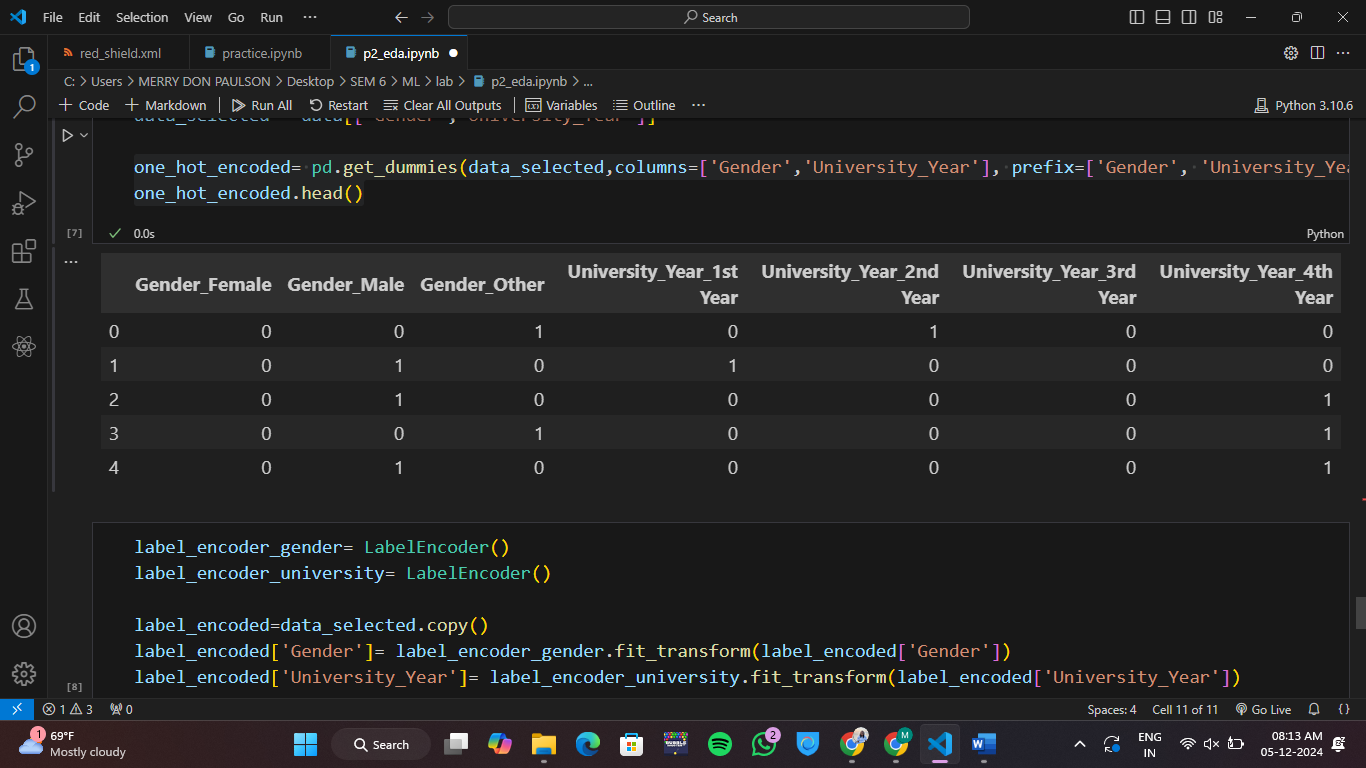


**Interpretation—First- and second-year students have similar sleep patterns, with medians around 6.5-7 hours. Third-year students show the widest range, reflecting greater variability. Fourth-year students have a slightly higher median sleep duration.**

1. One hot encoding:

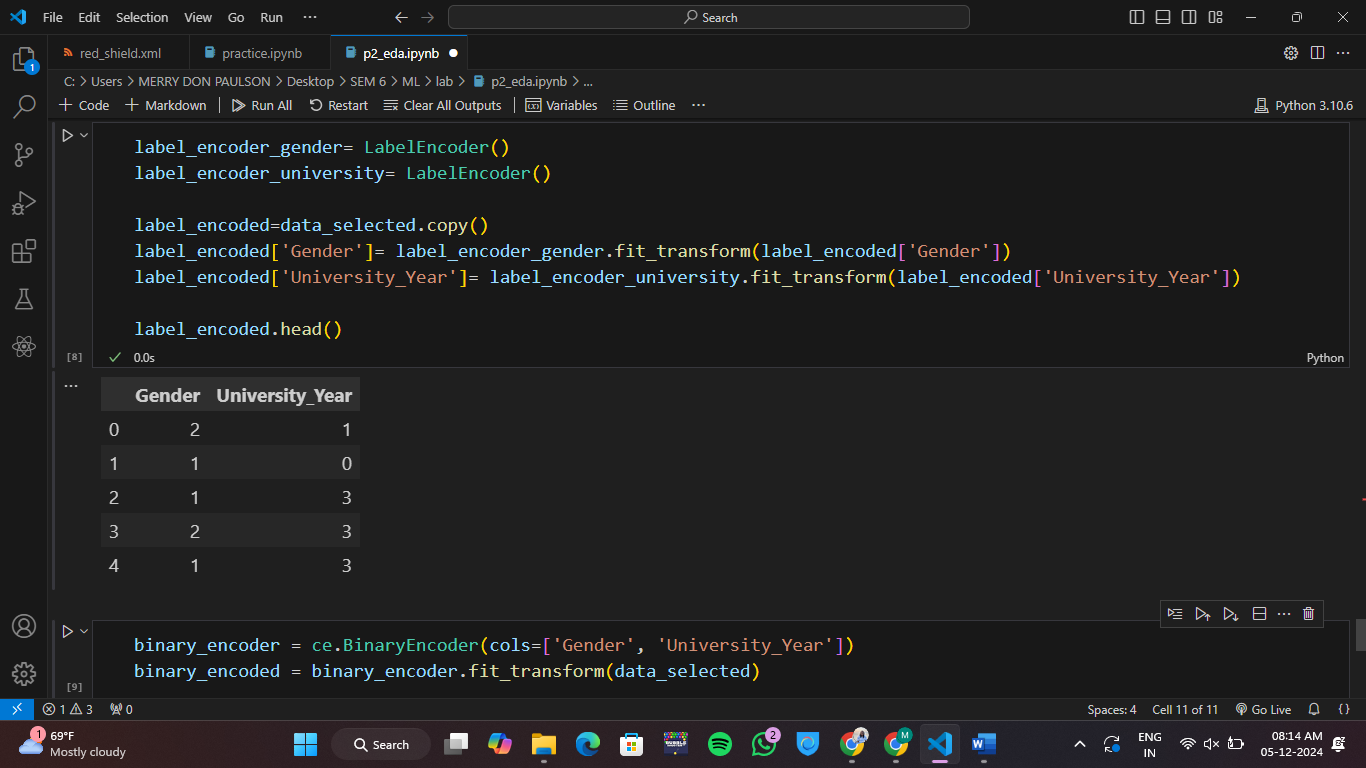


Output –

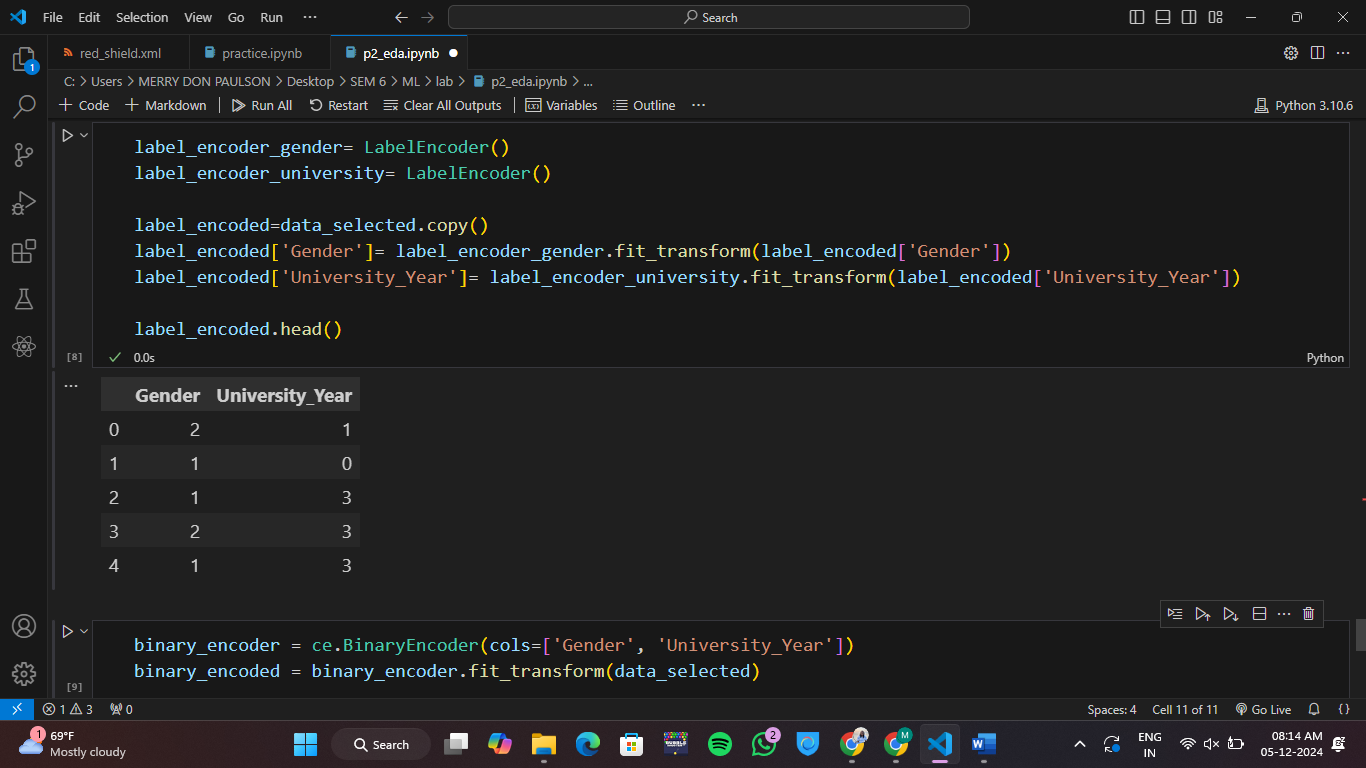


**Interpretation – The data uses one-hot encoding to represent categorical variables: gender (female, male, other) and university year (1st, 2nd, 3rd, 4th). Each row represents an individual, with 1s indicating their category and 0s for others**

1. Label encoding:

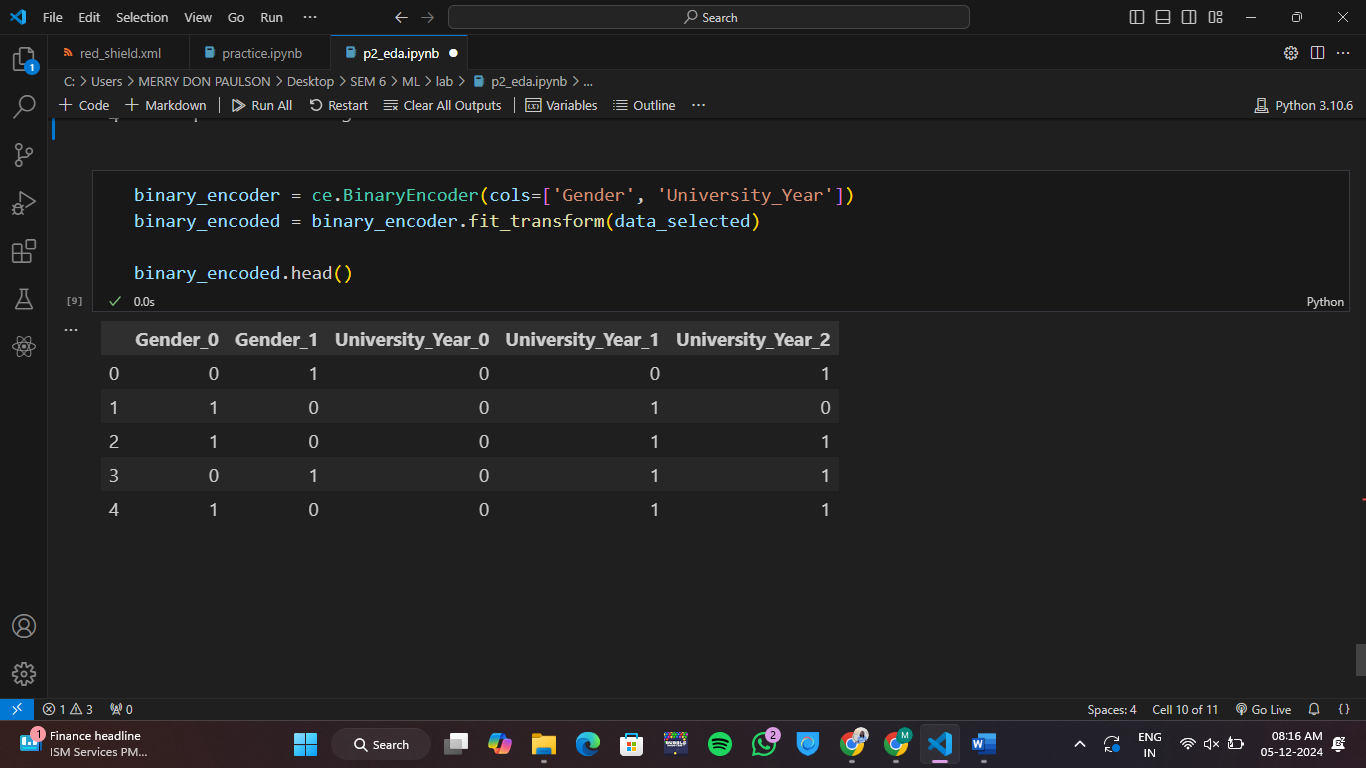


Output –

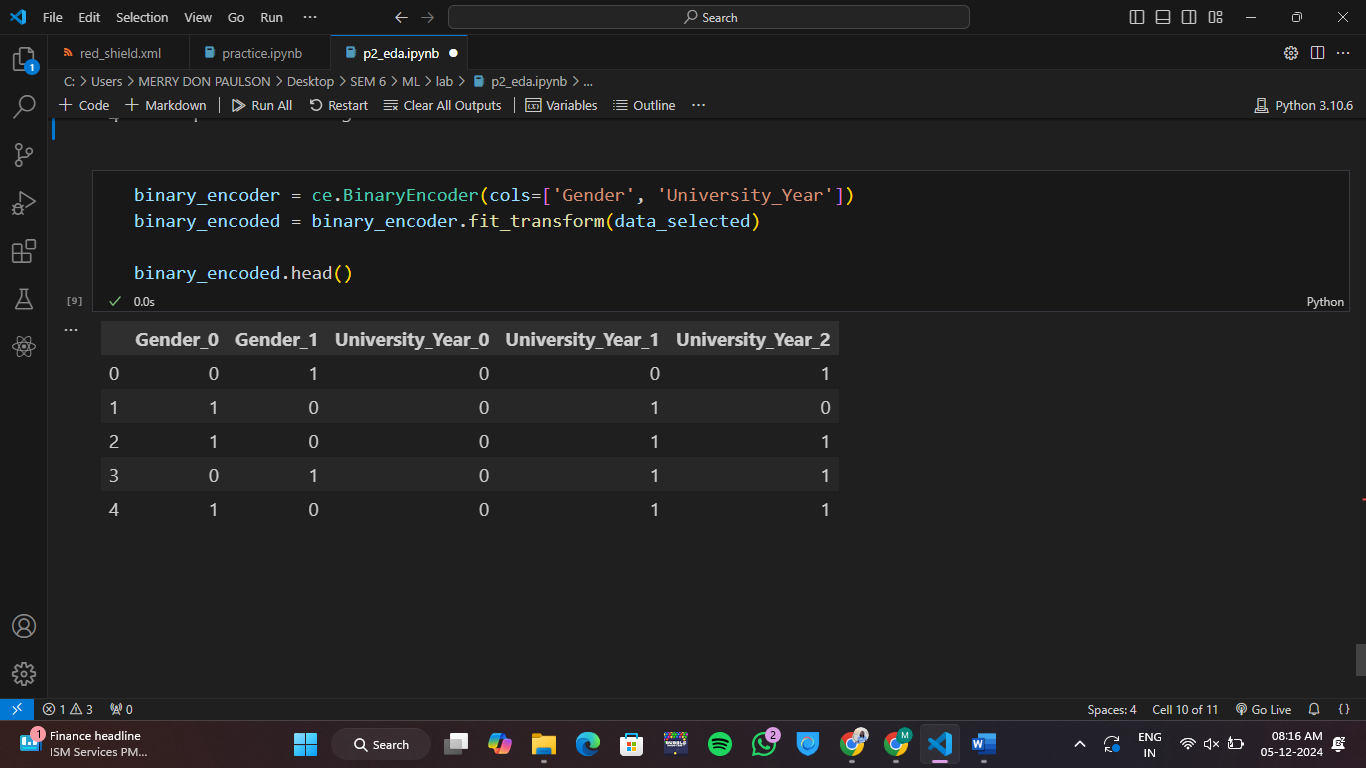


**Interpretation – Each row represents an individual, with the values in each row indicating their gender and university year. This data could be used for various analyses, such as understanding gender distribution across different years of study.**

1. Binary encoding:



Output –



**Interpretation - Each row in the table likely represents an individual. The values in each row indicate the individual's characteristics across the five categories. For example, the first row shows an individual who is female (Gender\_0 = 1), not male (Gender\_1 = 0), is in their first year of university (University\_Year\_0 = 1), and is not in their second or third year (University\_Year\_1 = 0, University\_Year\_2 = 0).**