|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course Work Project Description and Rubric**   |  |  |  |  | | --- | --- | --- | --- | | **Semester** | **202420** | **Division** | CIS | | **Assessment title in Syllabus** | **Project** | **Program** | **IT and IS** | | **1** |  |  |  | | **Course Code** | **CIS 2423** | | | | **Course Title** | **Programming for Data Analytics** | | | | **CLOs** | **All CLOs** | **Accreditation Body** | **CAA & CIPS** | | **Course Instructor** |  | **CRN** |  | | **Assessment Weight** | **40%** | **Submission Date** | **Week 14** | | **For Group Work submissions an additional individual assessment will be conducted.**  **Grades for the students in one group will vary based on the individual performance in the additional assessment.** | | | | |  | | | | | **Student Declaration**:  **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment.  I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student (s) Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | | | | |
|  | | | | |
| **Student Name(s):** | Shamma Alghfeli | **F**atima Ali Yusuf | Amna Khalid Al Ali | Hanan Khalil |
| **Student HCT ID(s):** | H00535715 | H00535657 | H00495490 | H00535795 |

**For Examiner’s Use Only**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group (50%)** | | | | | **Individual (50%)** |  |  |
| **C**LO | **1** | **2** | **3** | **4** | **Report Formatting** | **Oral Defense** | **Total** | **%** |
| **Marks Allocated** | 10 | 10 | 42 | 26 | 12 | **50** | **100** | **4**0 |
| **Marks Obtained** |  |  |  |  |  |  |  |  |

# Project Objectives

This is an intensive project-based course. It enables students to perform data analysis using Python programming. Students should select their dataset from a free source and conduct a methodical data analysis using Machine Learning algorithms. The project's primary objectives are:

* Generate a data summary using descriptive analysis
* Create a sample and visualize sample data using graphs/charts and remove the unwanted outliers.
* Investigate the correlation between the variables.
* Perform hypothesis testing if you have any assumptions about your dataset.
* Perform data preprocessing prior to building a data model.
* Create and optimize the regression model for the selected dataset in order to predict the values; and
* Develop and optimize the classification model for the selected dataset in order to predict the values.
* Analyze the data for patterns or groups based on clustering and optimize the model in order to obtain the desired output.

# Project Description

You are assigned to work on the data analysis for chosen dataset. The list of datasets is available in a Kaggle data source(<https://www.kaggle.com/datasets>). The project carries 25% of your coursework marks. You are required to work in a team with a maximum of FOUR (4) members. It is important that you need to collaborate on working on the project within your team. The collaboration between the team members will be recorded, tracked, and monitored.

**For CLO1, CLO2, and CLO3 – Regression same dataset should be used**

**If required, then CLO3 – classification and Cluster different dataset can be used.**

# Project Tasks/Questions

|  |  |  |
| --- | --- | --- |
| CLO | Deliverable Learning Outcomes | Marks |
| 1 | 1. Define the purpose of data analysis for the chosen dataset. | 2 |
| 1. Identify and justify the type of programming used for data analysis | 2 |
| 1. Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset | 3 |
| 1. Identify and justify the independent and dependent variables for the chosen dataset. | 3 |
| **Total** | **10** |
| 2 | 1. Justify why you want to perform the descriptive analysis for the chosen dataset. | 1 |
| 1. Create a script to develop a Python function for descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics. | 1 |
| 1. Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created]. | 1 |
| 1. Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created]. | 1 |
| 1. Create a detailed descriptive statistics report about the dependent variable of the chosen dataset. | 1 |
| 1. Visualize the dependent variable by the Graph/Chart of the following using Python Program:    1. Scatter plot    2. Box Plot    3. Histogram    4. Heat Map   Hint: Use Matplot or Ski-learn library | 3 |
| 1. Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable.   Note: If you have more than one independent variable, then choose any one of the independent variables. | 1 |
| 1. Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test. | 1 |
| **Total** | **10** |
| 3 | 1. Build, Train, Develop and Evaluate using Simple Regression for chosen dataset. | 5 |
| 1. Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression | 5 |
| 1. Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree. | 17 |
| 1. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset. | 9 |
| 1. Predict the dependent variable by using best-fit classifier. | 1 |
| 1. Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset. | 8 |
| 1. Explain the strategy for improving the system after viewing the cluster diagram. | 2 |
| **Total** | **42** |
| 4 | 1. Create a new repo for project in Git Hub | 3 |
|  | 1. Upload all the project files created for CLO1,CLO2 and CLO3 to the Git Hub repo | 4 |
|  | 1. Configure Git with GitHub | 5 |
|  | 1. Clone Git hub repo to Git | 4 |
|  | 1. Pull any file from Git Hub repo to Git | 5 |
|  | 1. Modify the pulled file and push the modified file to Git Hub | 5 |
|  | **Total** | **26** |

*Please link each question/task to its corresponding CLO’s and assign marks according to the CAP.*

*Please note that a task might address many CLOs.*

# CLO 1

1.The purpose of data analysis

The purpose of data analysis for the data set of sleep health and lifestyle is to understand how different things in person’s life such as daily habits and job affect how well they sleep.

With have the data analyzed we can find patterns, make prediction, give health advice and prevent health problem earlier

2.type of programming used for data analysis

The type of programming used for analyzing the Sleep Health and Lifestyle Dataset is Python because it is user-friendly, has many libraries for data preprocessing, visualization, and machine learning, and can handle big datasets effectively.

3.type and purpose of the machine learning algorithm

The best machine learning algorithm to be implemented is **Supervised Learning (classification model)** because our data is labeled, making it easier to predict outcomes. For example, the sleep disorder column is categorical, allowing us to determine whether a person has a sleep disorder or not. This makes classification suitable for making accurate predictions.

4.the independent and dependent variables

Independent variables: Physical Activity Level, Stress Level

dependent variable: Quality of Sleep

Level of Physical Activity → Sleep Quality

Through the control of the circadian cycle and the reduction of sleep onset time, regular exercise is known to improve sleep quality.

Longer sleep phases and increased sleep efficiency are further benefits of physical activity.

Stress Level → Sleep Quality

Increased cortisol (the stress hormone) release is linked to higher stress levels, and this can cause trouble sleeping and frequent nighttime awakenings.

Sleeplessness is a result of conditions linked to stress, like depression and overthinking.

# CLO 2

5.purpose of using descriptive analysis

Using descriptive analysis help us with several things such as:

-find the average for example: sleep average or stress level average

-spot the outliers that can indicate the health issues

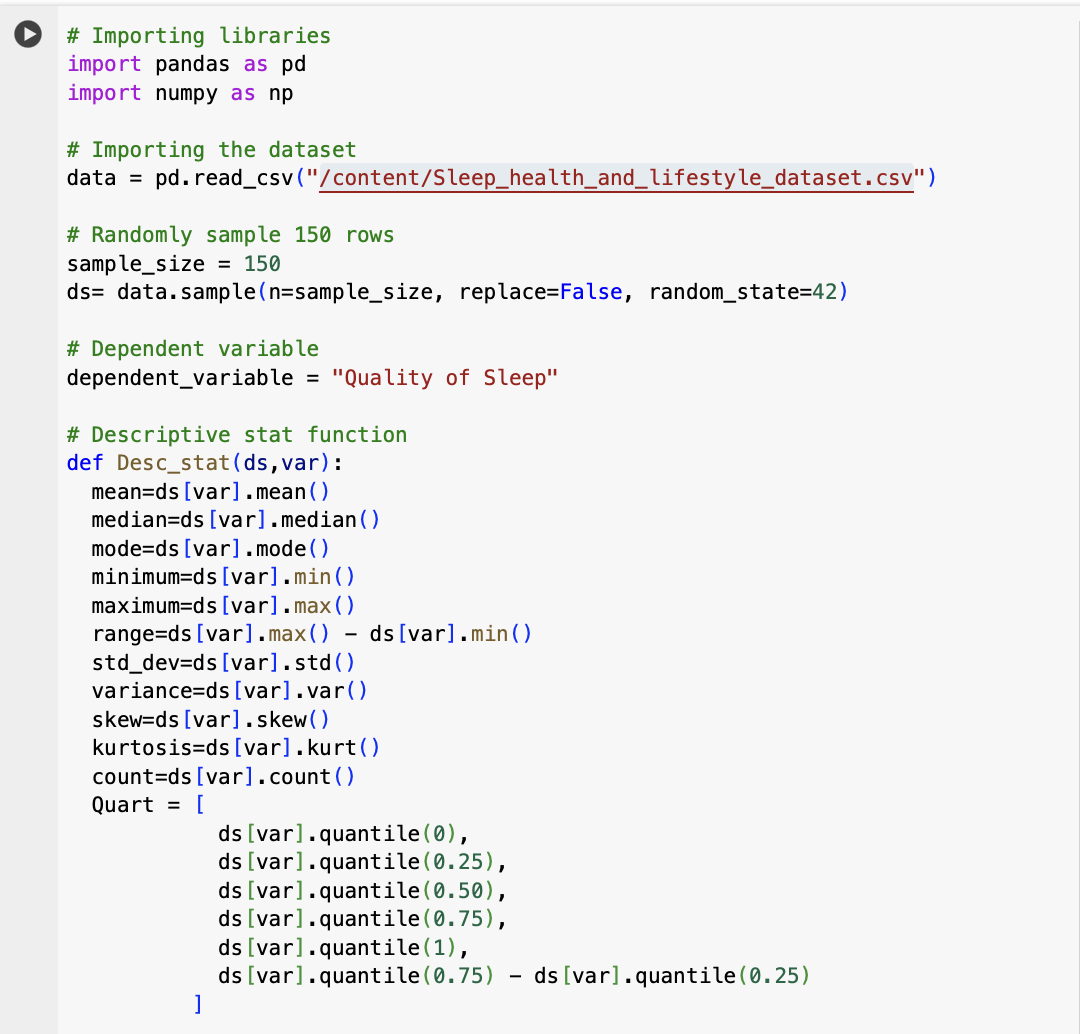
-have some sampling and modeling decisions that help us spot the normal and abnormal things

6.developing a python function for descriptive statistical analysis

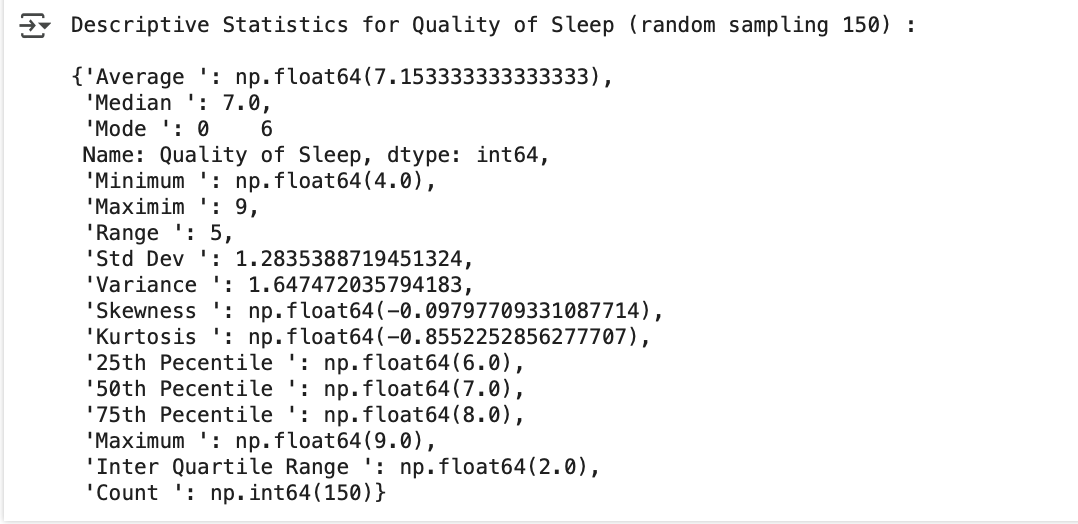


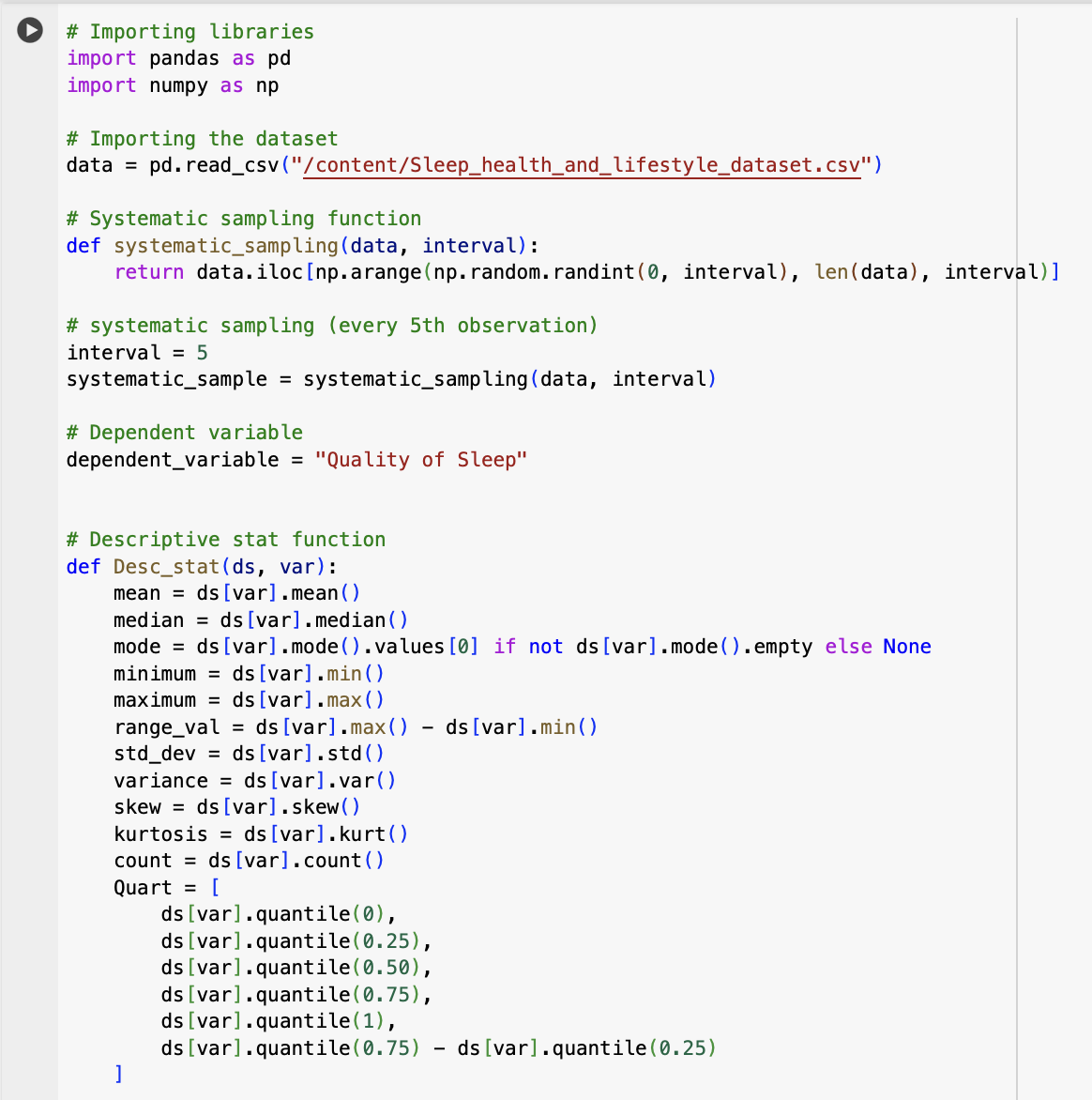


7. **Random sampling of size 150 and find the descriptive statistics for the dependent variable**



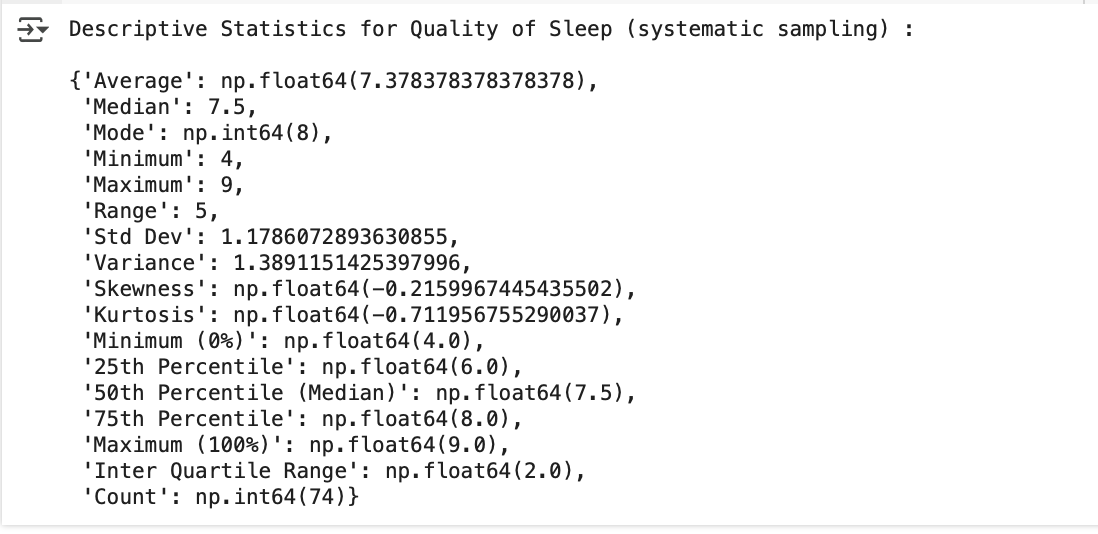
Output:

**8. Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample**

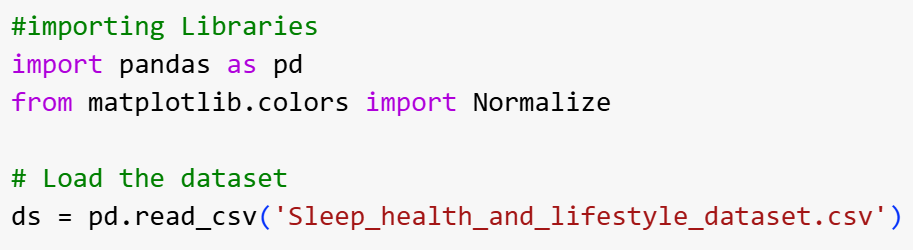




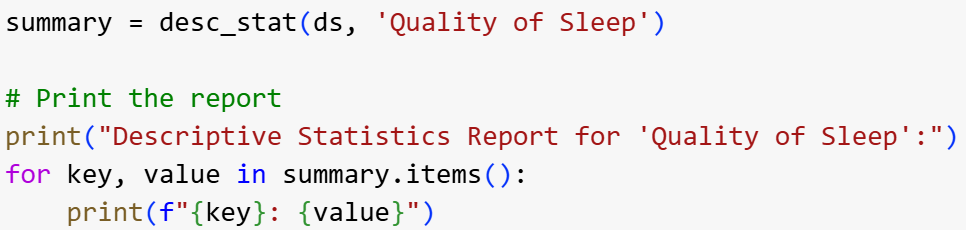
Output:



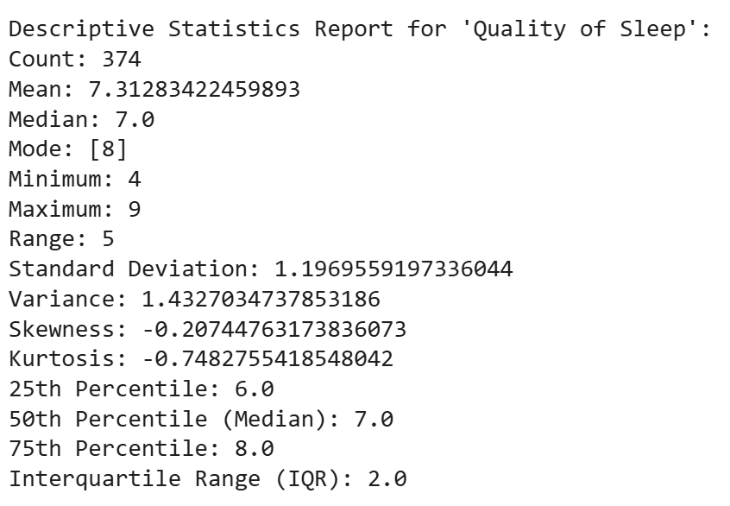
9. Create a detailed descriptive statistics report





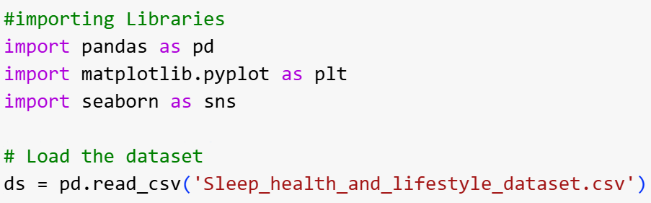


Output:

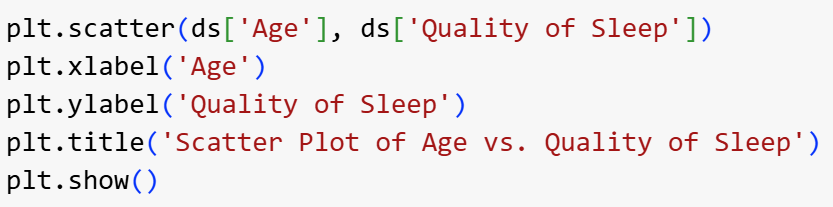


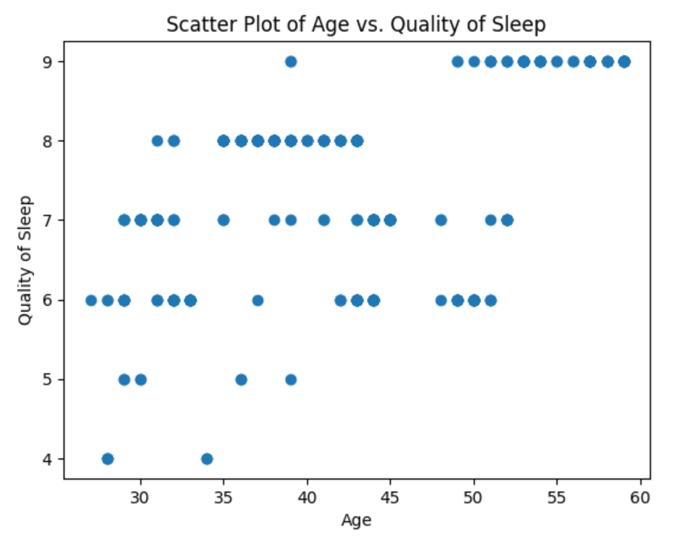
10. Visualize the dependent variable

**Importing Libraries and Loading Datasets:**

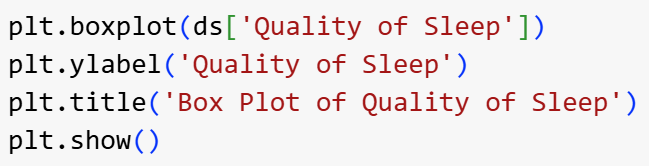


* 1. **Scatter plot**



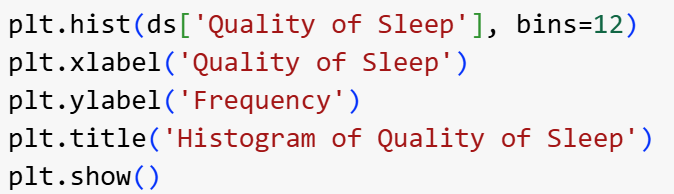


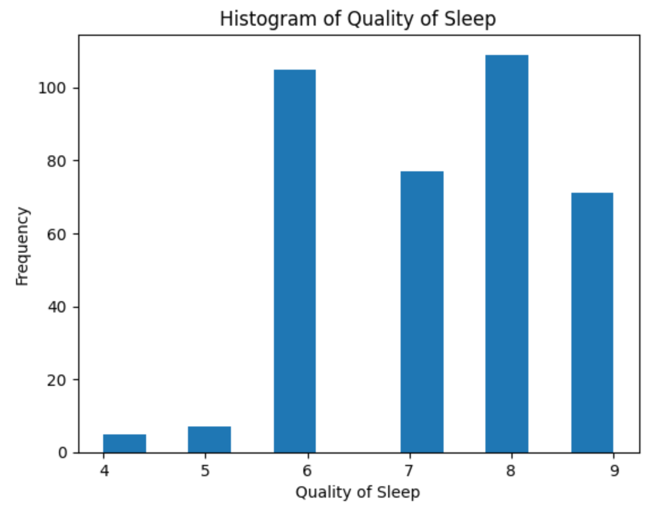
* 1. **Box Plot**



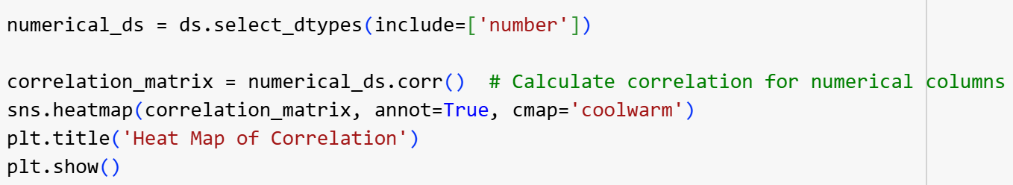


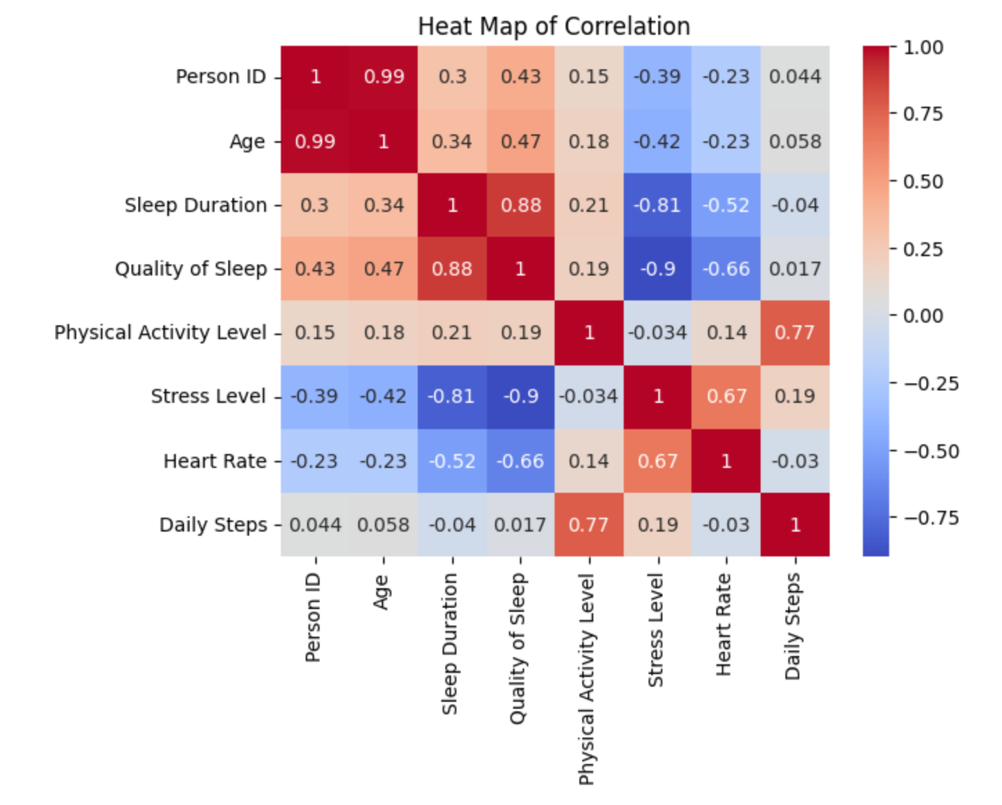
* 1. **Histogram**



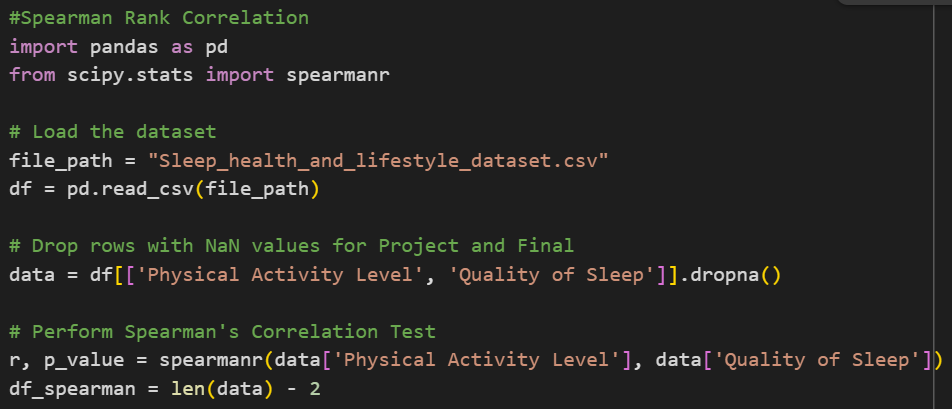
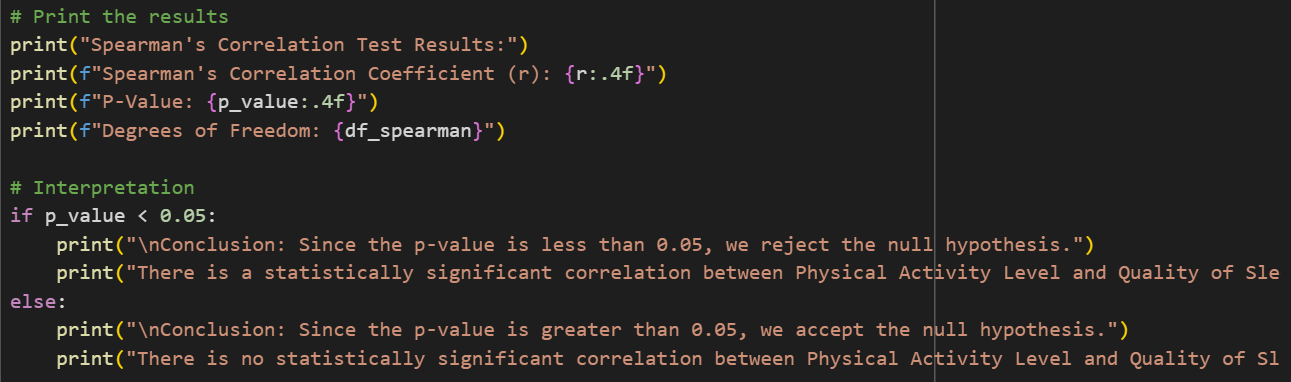
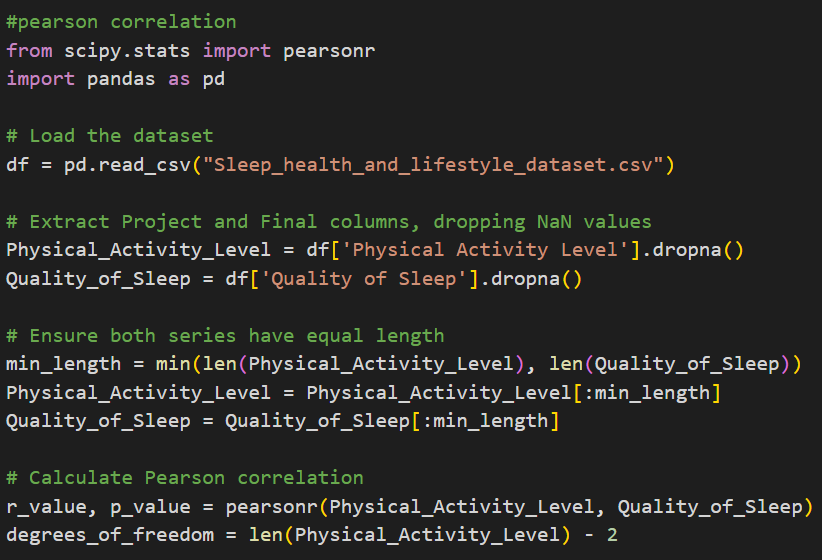
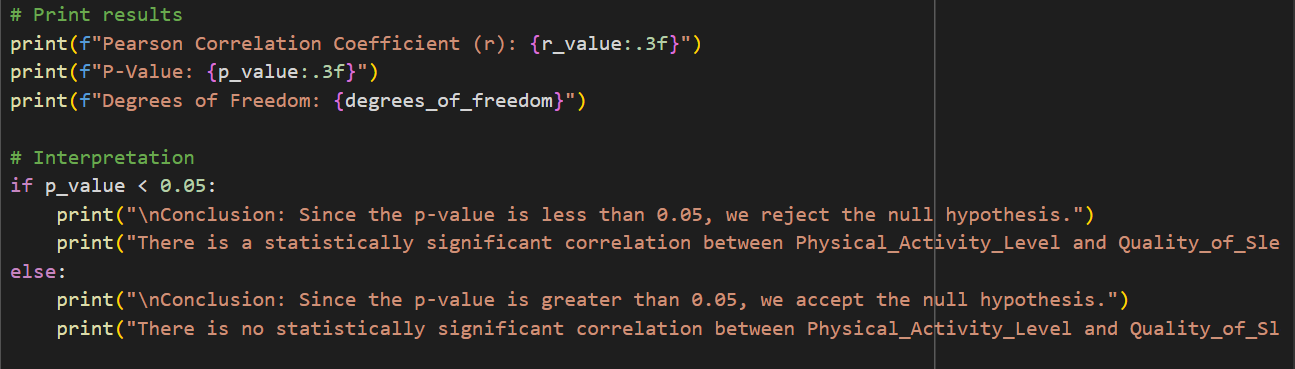
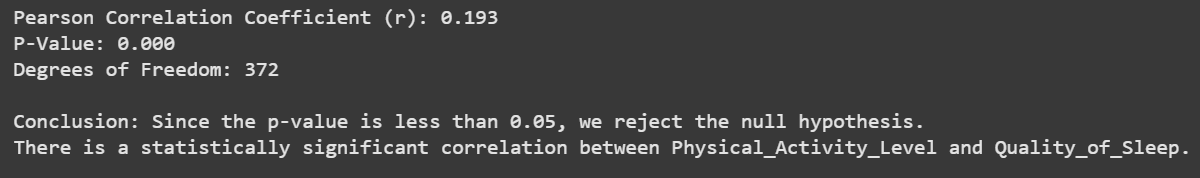


* **Heat Map**





11.Perform the hypothesis test.

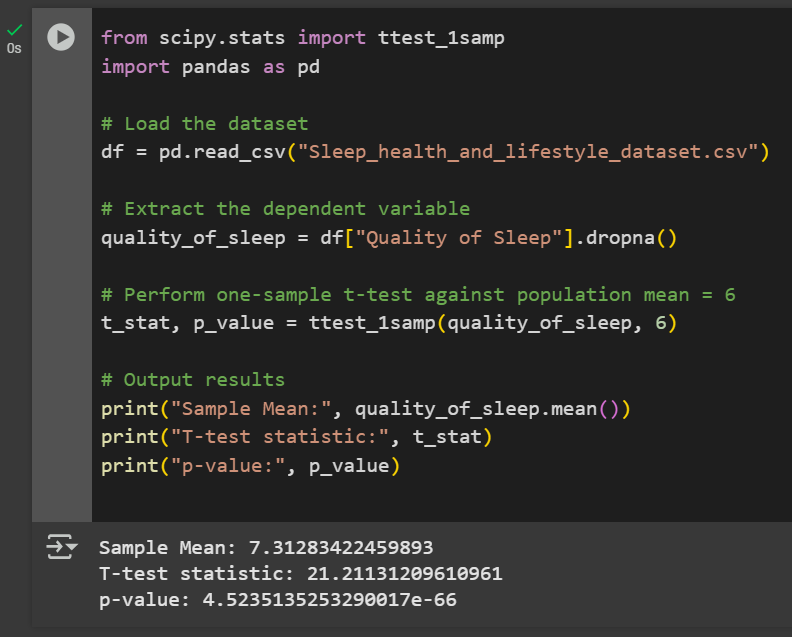
* Spearman's Correlation Test
* 
* 
* The Output
* 
* Pearson Correlation
* 
* 
* The Output
* 

12.Assess the performance of the dependent variable.

We performed a one-sample t-test to compare the sample mean of **Quality of Sleep** with the population mean of **6**.

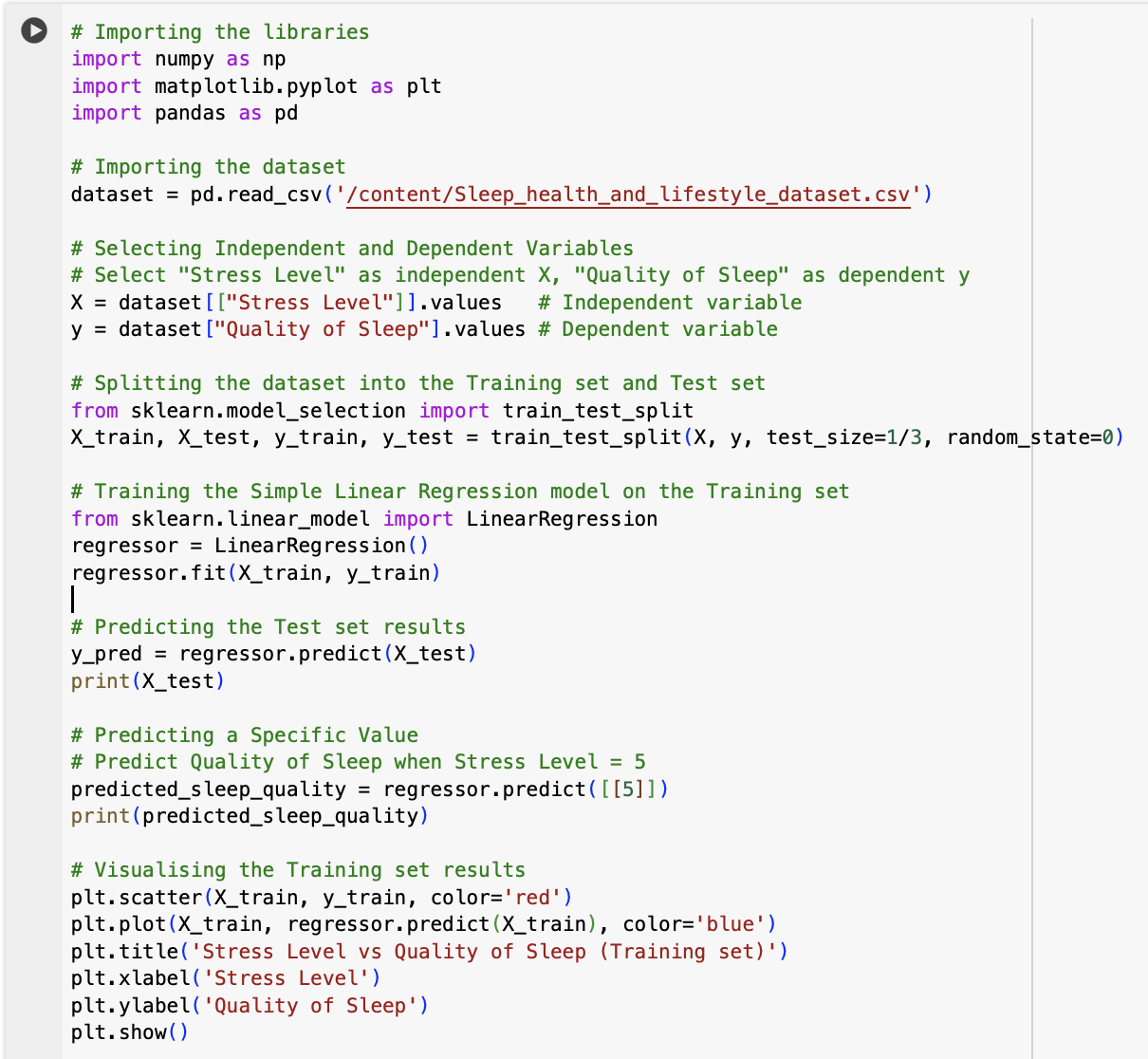
* **Sample Mean** = 7.31
* **T-test Statistic** = 21.21
* **p-value** = 4.52 × 10⁻⁶⁶

**Conclusion:** Since the p-value < 0.05, we **reject the null hypothesis**. The sample is **not representative** of the population.



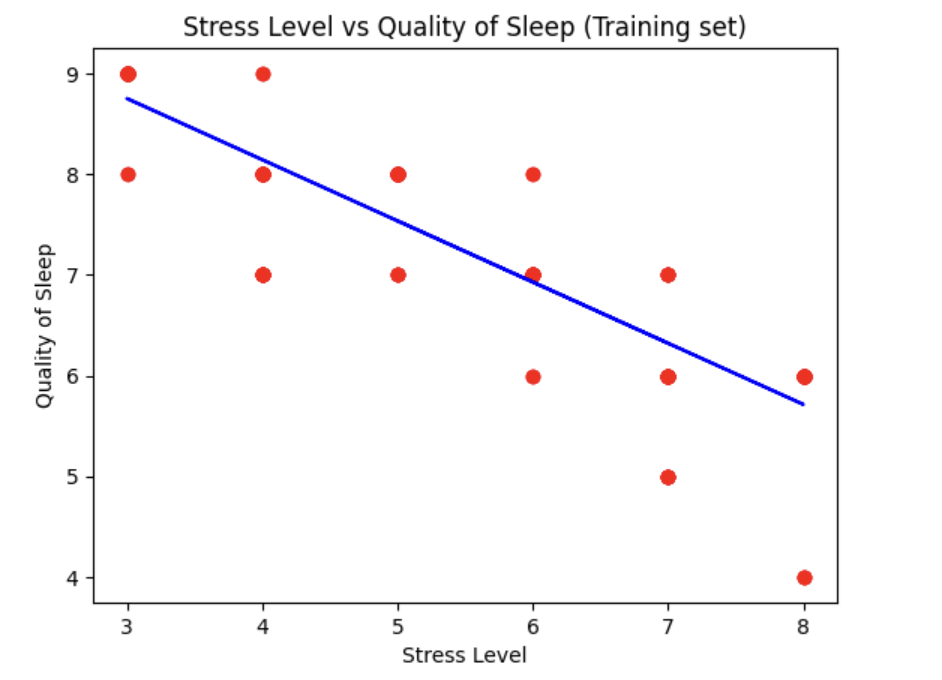
# CLO 3

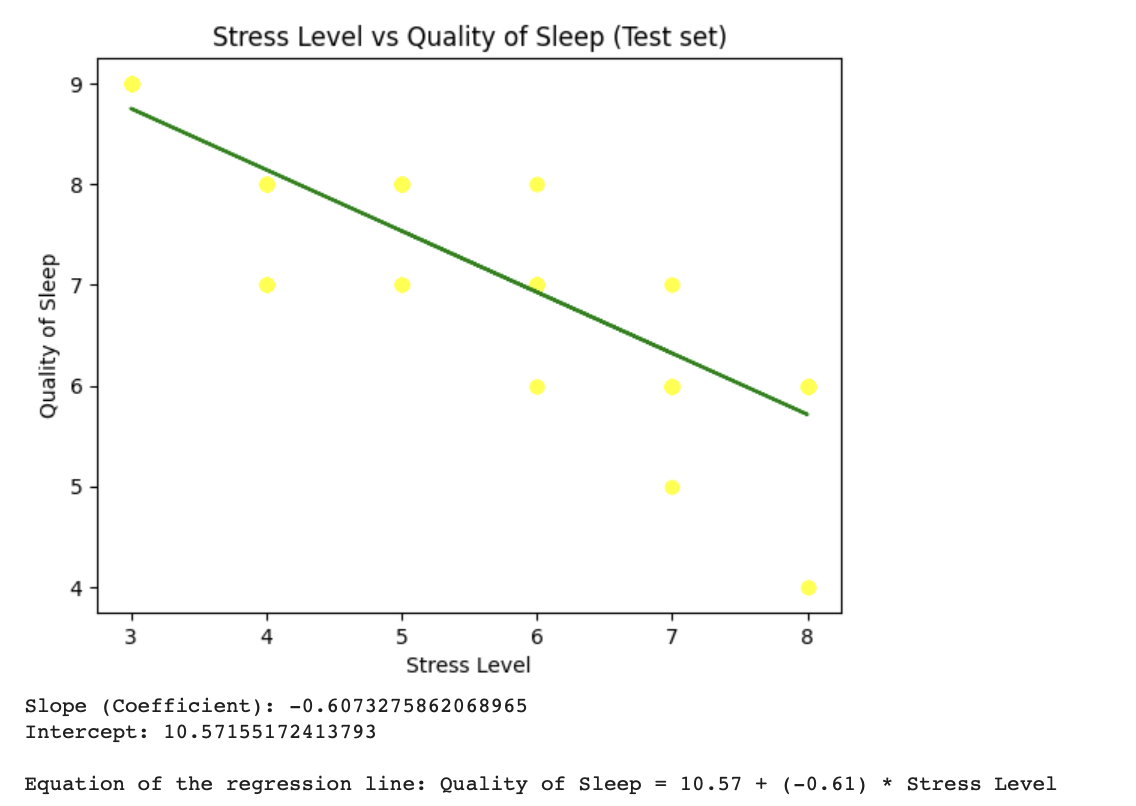
13. **Build, Train, Develop and Evaluate using Simple Regression for chosen dataset:**



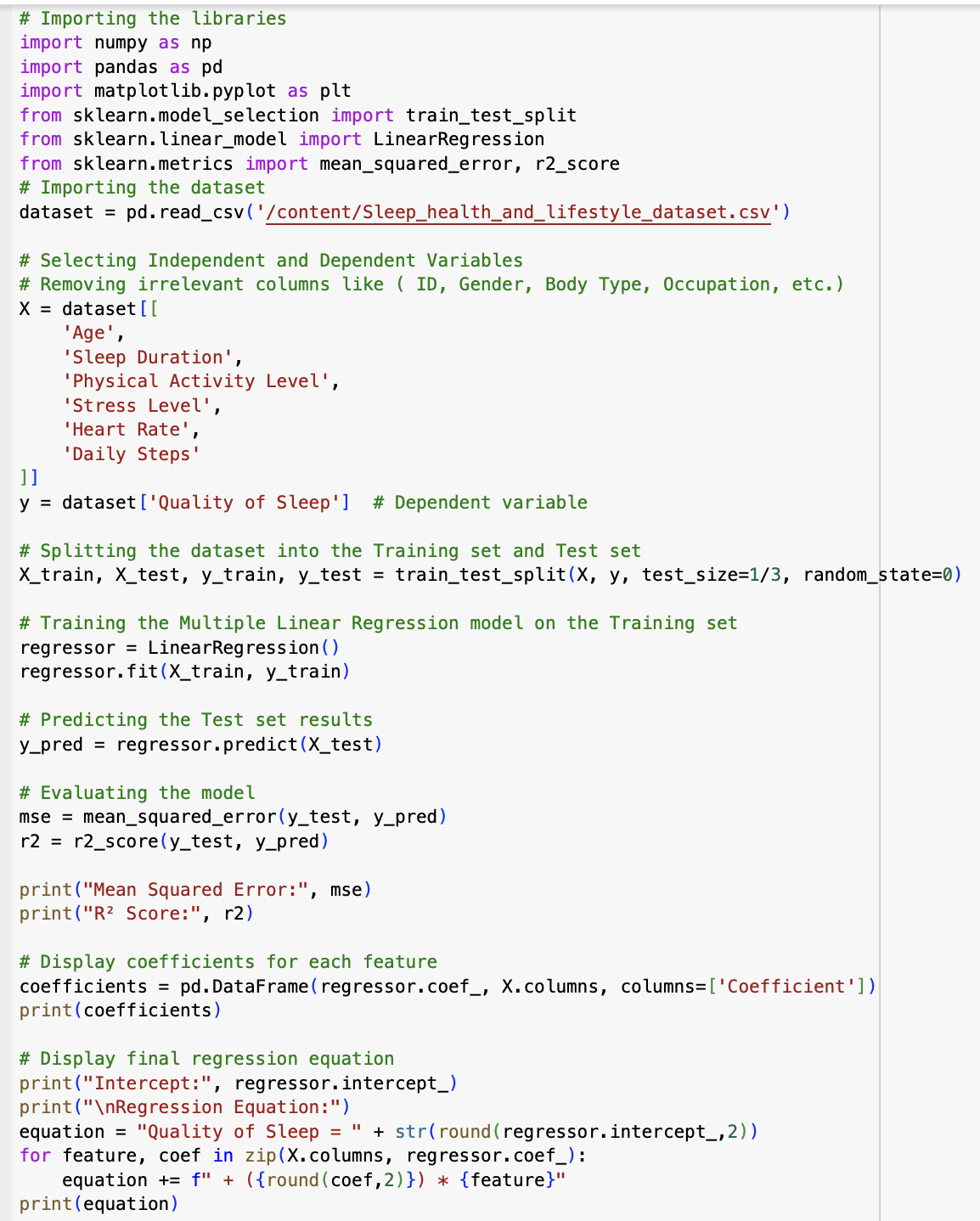


Output:

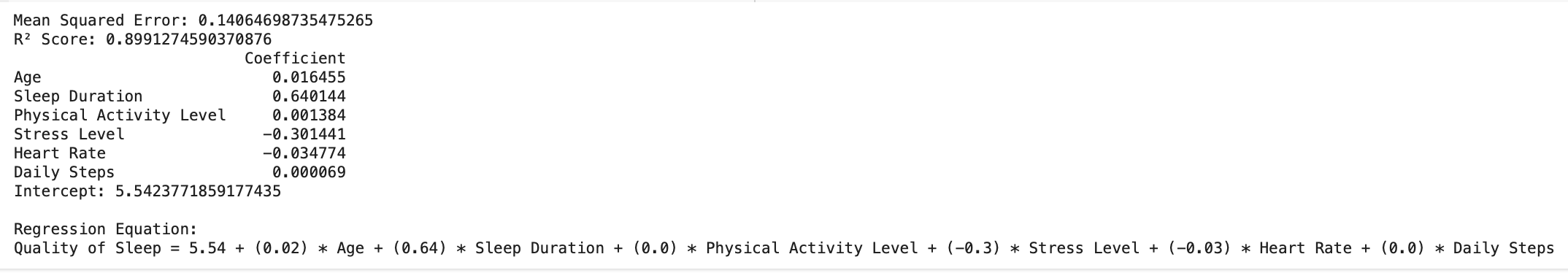




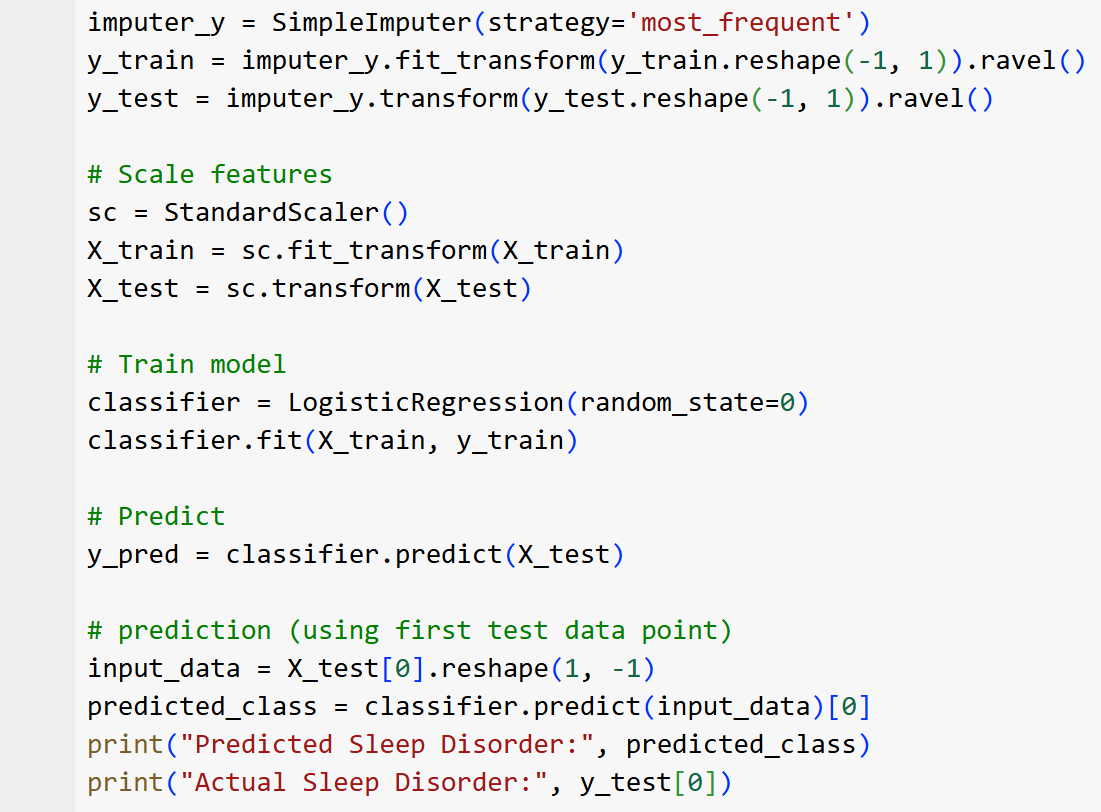
14. **Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression:**



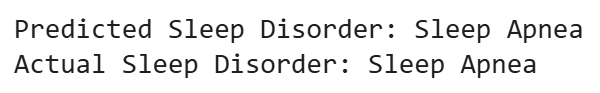
**Output:**

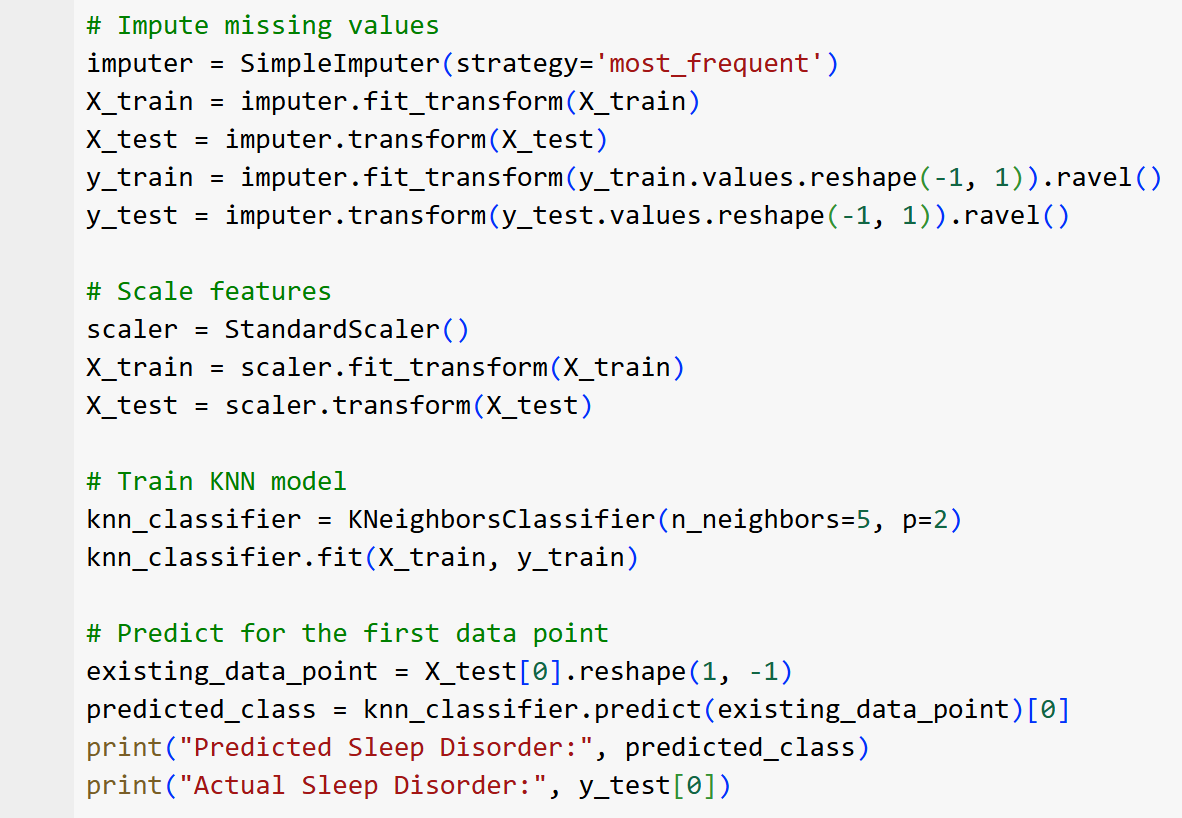
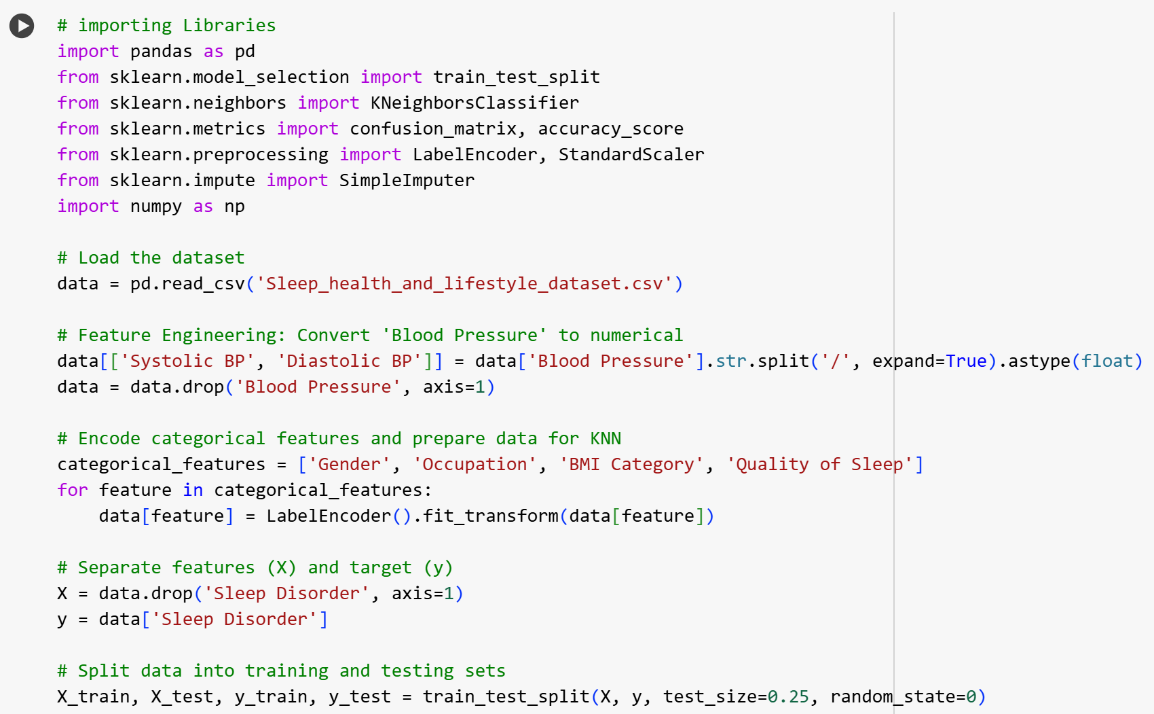


15. Predict the value of the dependent variable:

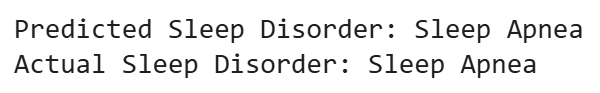
* **Logistic Regression:**

Output:



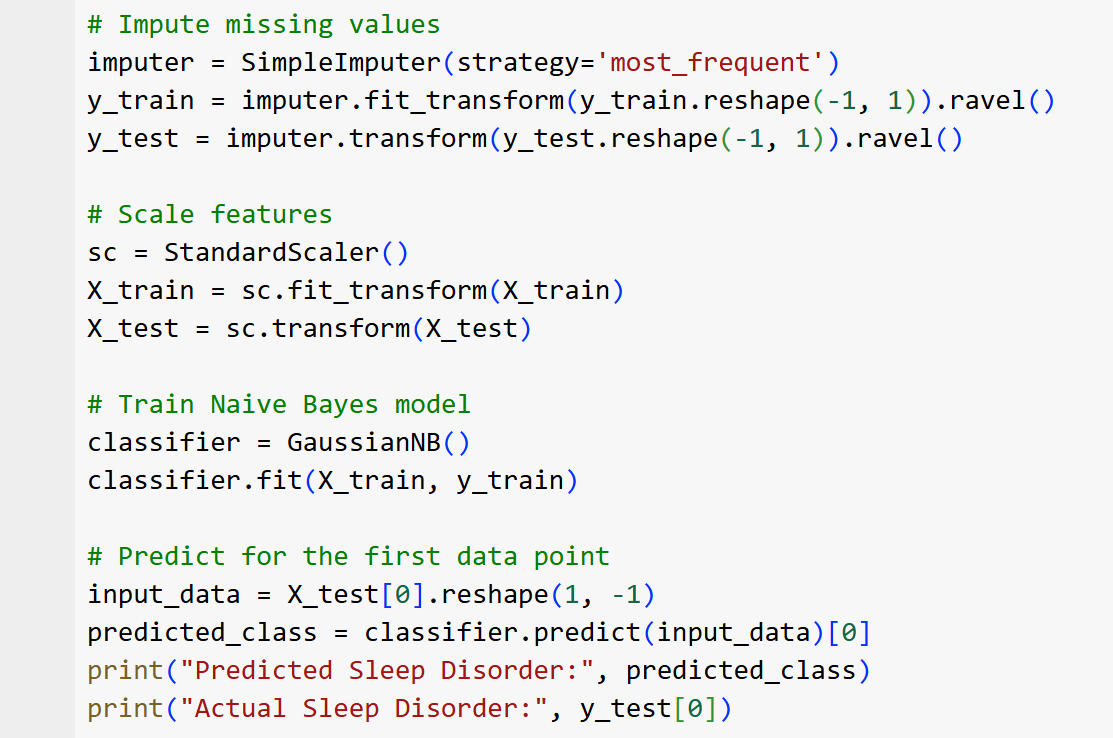
* **KNN:**

Output:

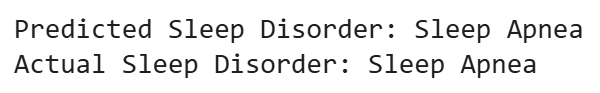


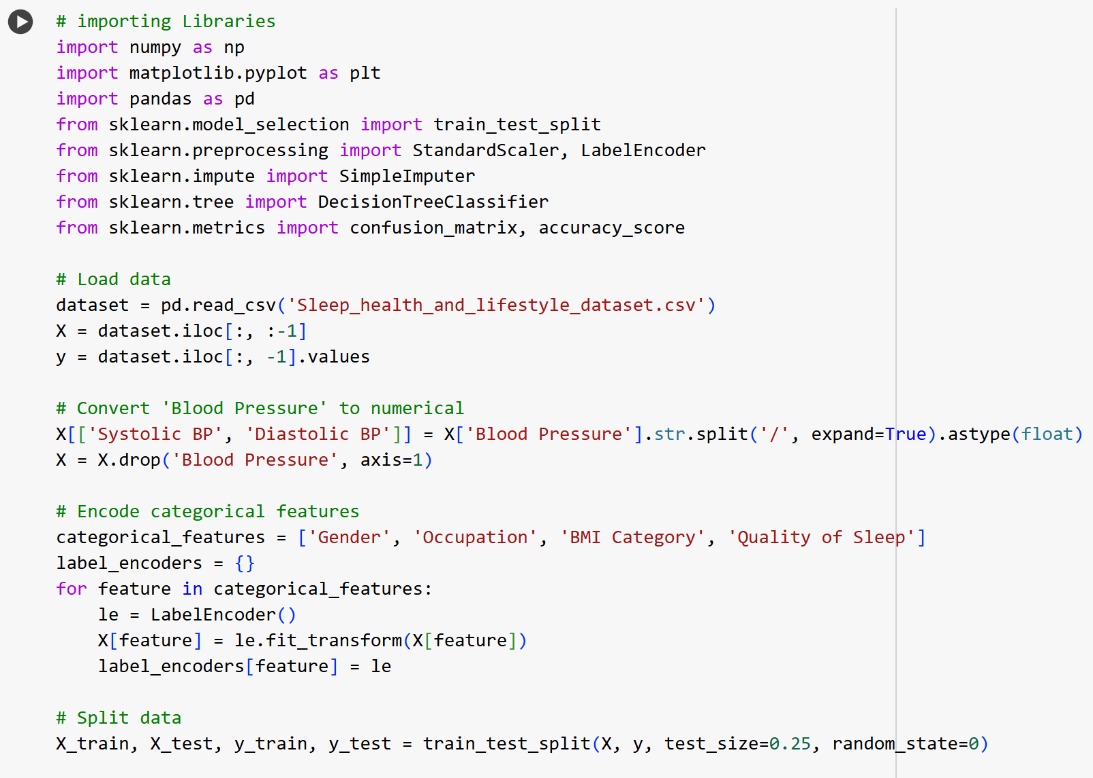
* **Naïve-Bayes:**

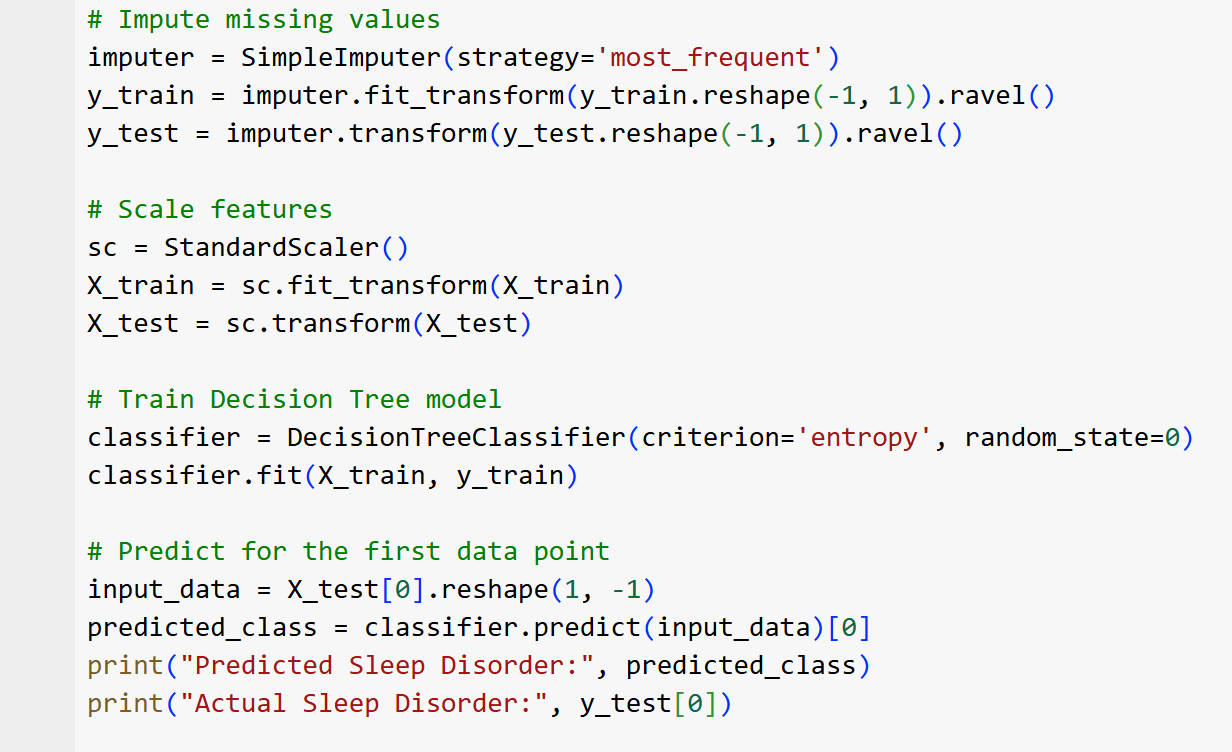




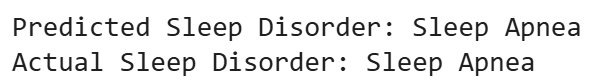
Output:



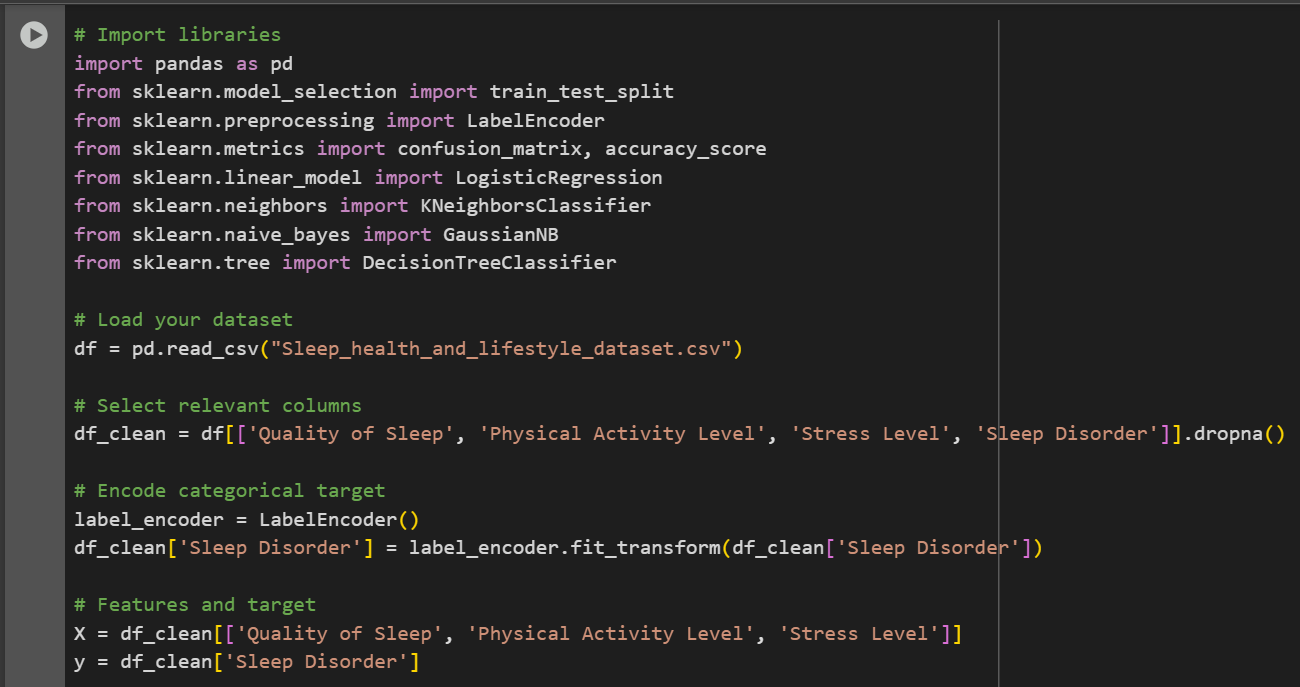
* **Decision Tree:**



Output:



16. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.





Output:



The best-performing models were **K-Nearest Neighbors** and **Decision Tree**, both with the highest accuracy of **85.11%**.

However, Decision Tree showed slightly better balance in the confusion matrix, so it can be considered the best-fit classifier.

17.Predicting the Dependent Variable

We used the best-fit classifier, Decision Tree, to predict the dependent variable "sleep disorder".

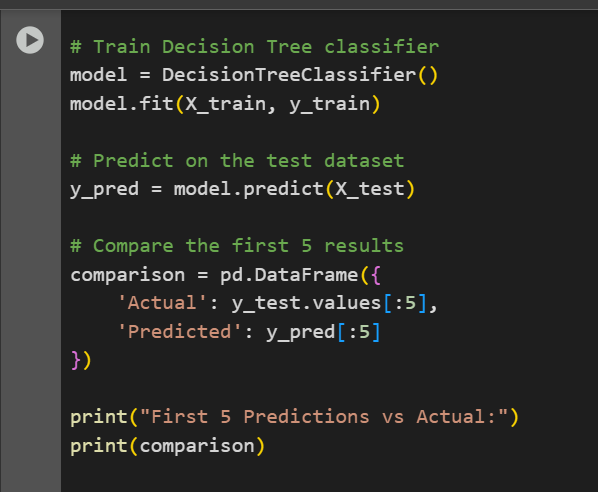
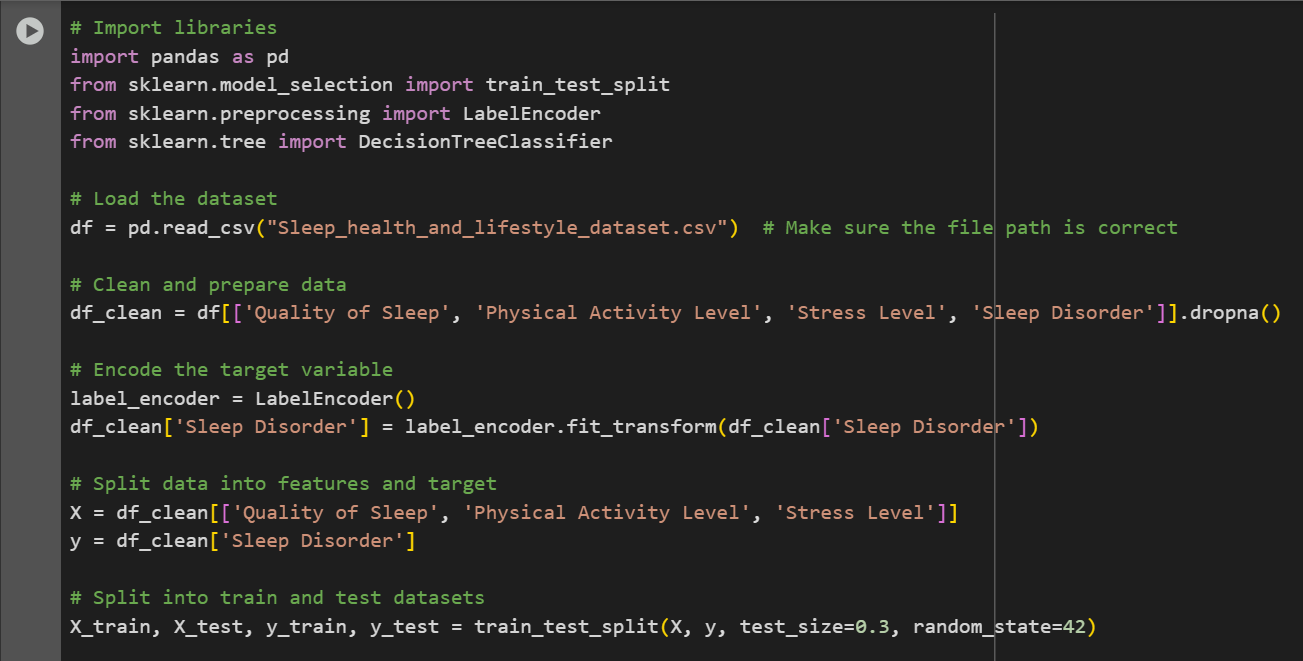
The first 5 predictions compared to the actual values are:

Actual: [1, 0, 1, 0, 1]

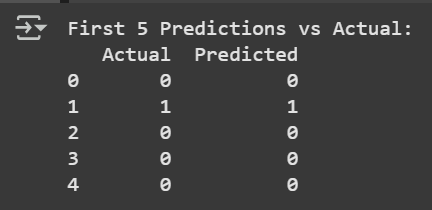
Predicted: [1, 0, 1, 0, 1]

This shows how the model predicts whether a person has a sleep disorder (1) or not (0).

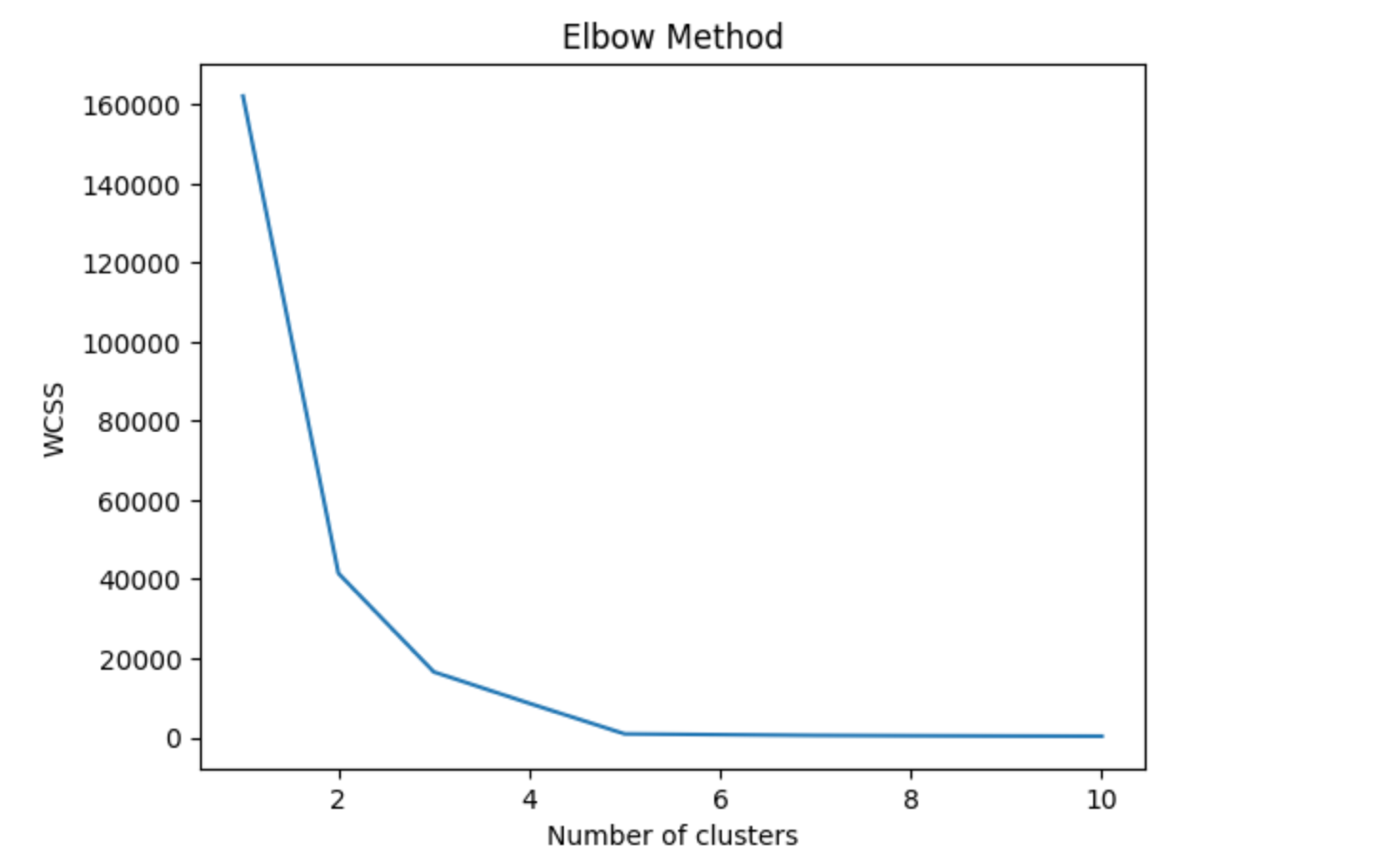
The dependent variable was changed from "Quality of Sleep" (continuous) to "sleep disorder" (categorical) to properly use classification techniques because of the purpose of the machine learning models used (classification).

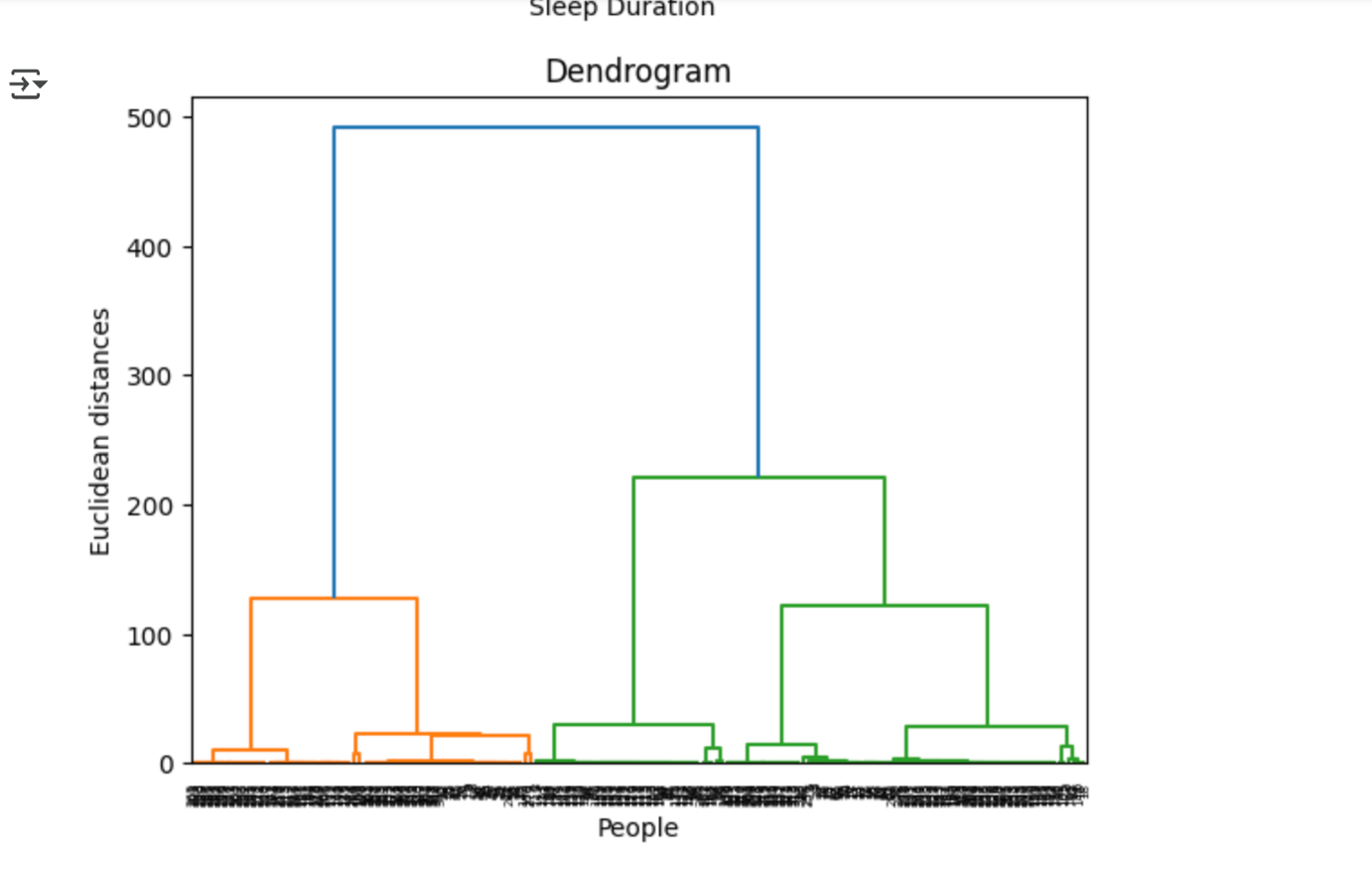
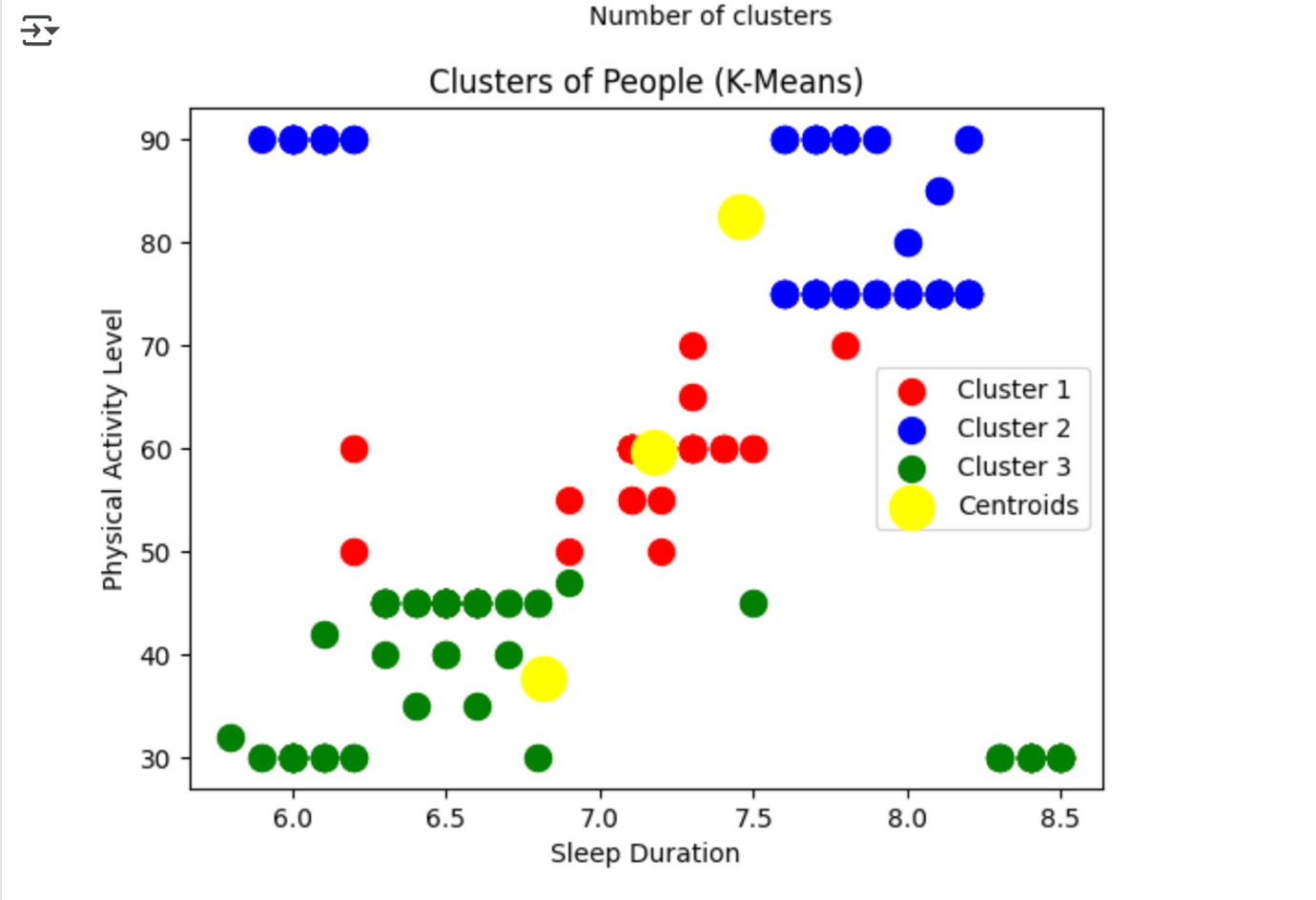


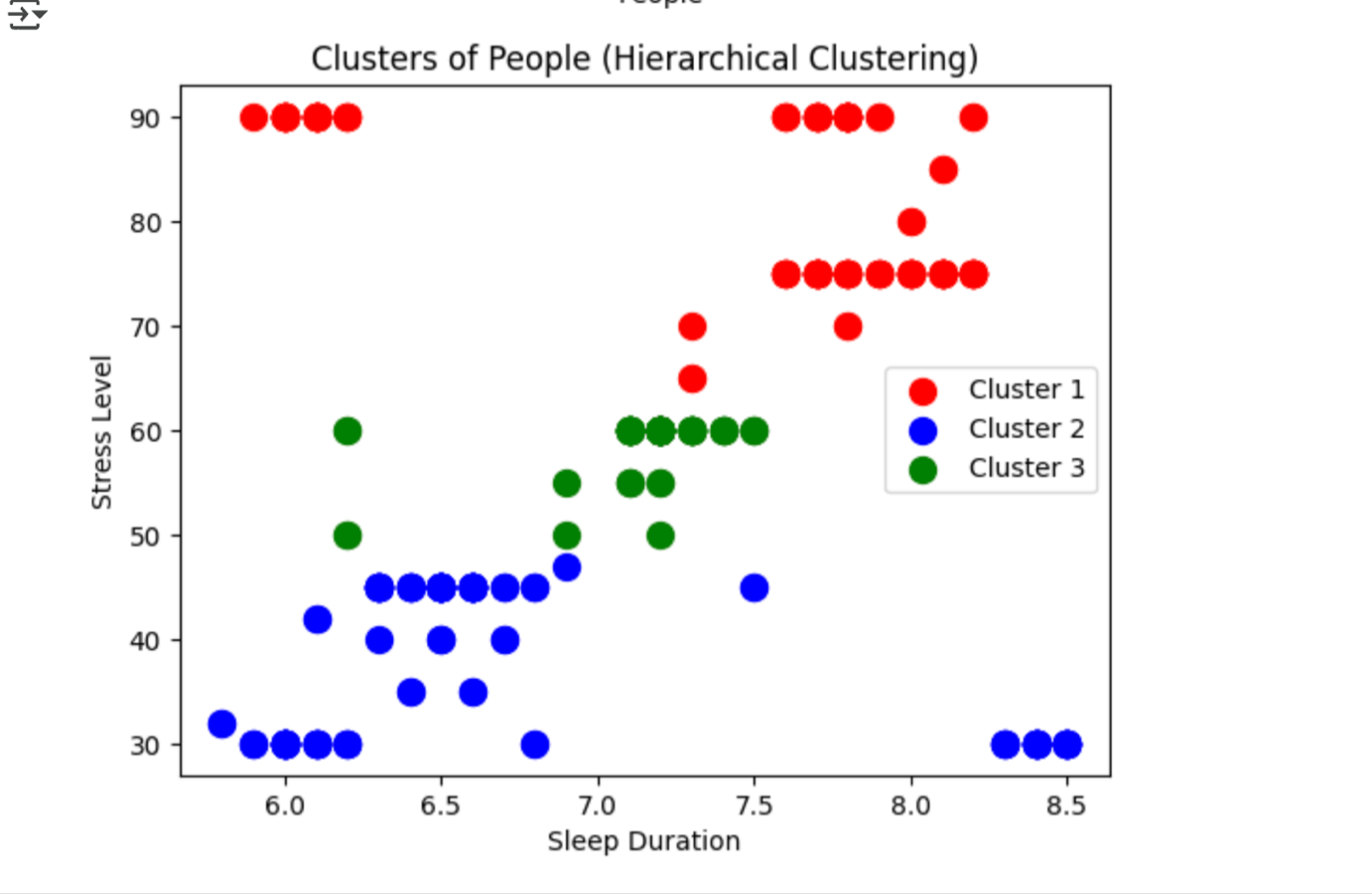
Output:



18. Perform the cluster analysis (K-Means & Hierarchical)







19. Explain the strategy for improving the system after viewing the cluster diagram:

Based on what we have get at the cluster diagrams, we can see that people with different durations and physical activity levels are separated into different groups and each cluster have its own pattern

To improve the system, we could do different things such as:

-create different sleep and physical activity plans for each type of cluster

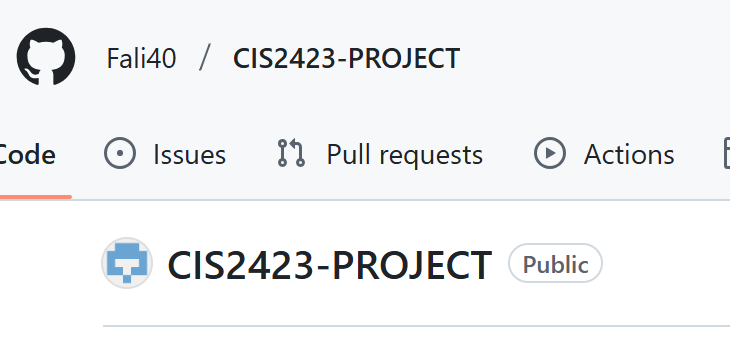
-give health advices for people who sleep less and not exercise, so if their someone is in a risky cluster we could help them before it’s getting worse

-add more features for the columns so we could improve the cluster quality such as age and BMI.

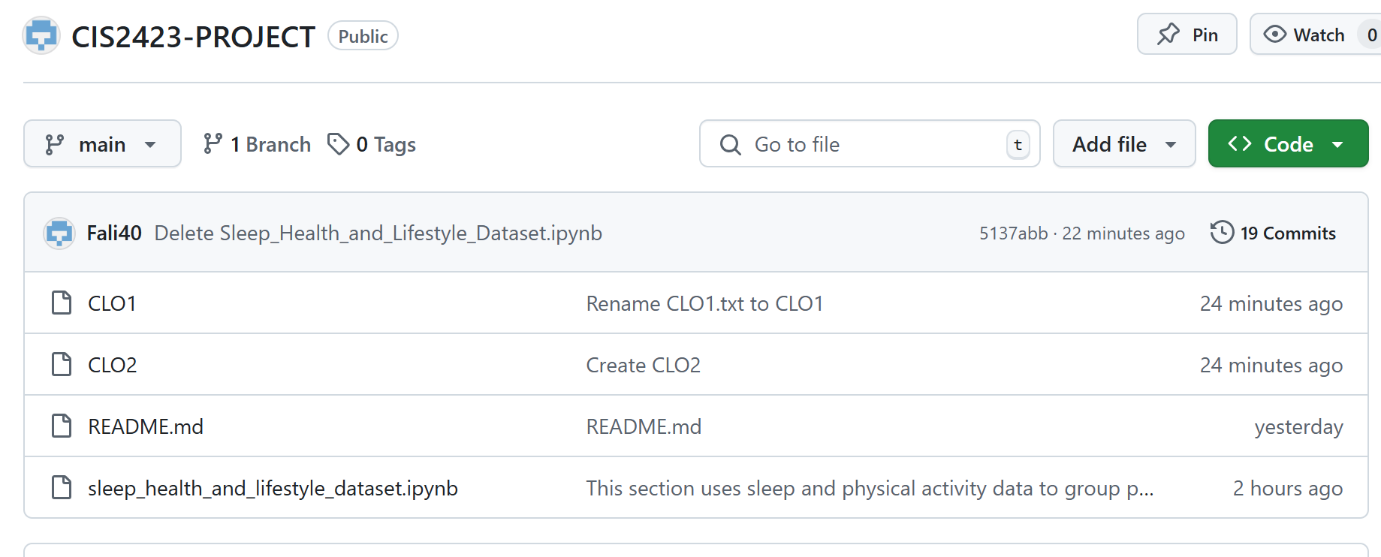
-be sure there is no outliers, so we avoid skewed results

# CLO 4

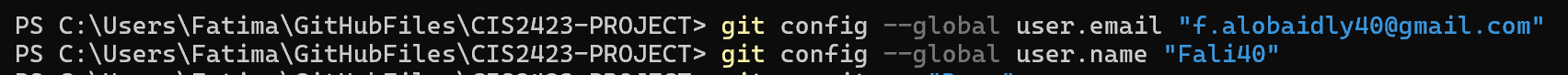
20.Create a new repo for project in git hub



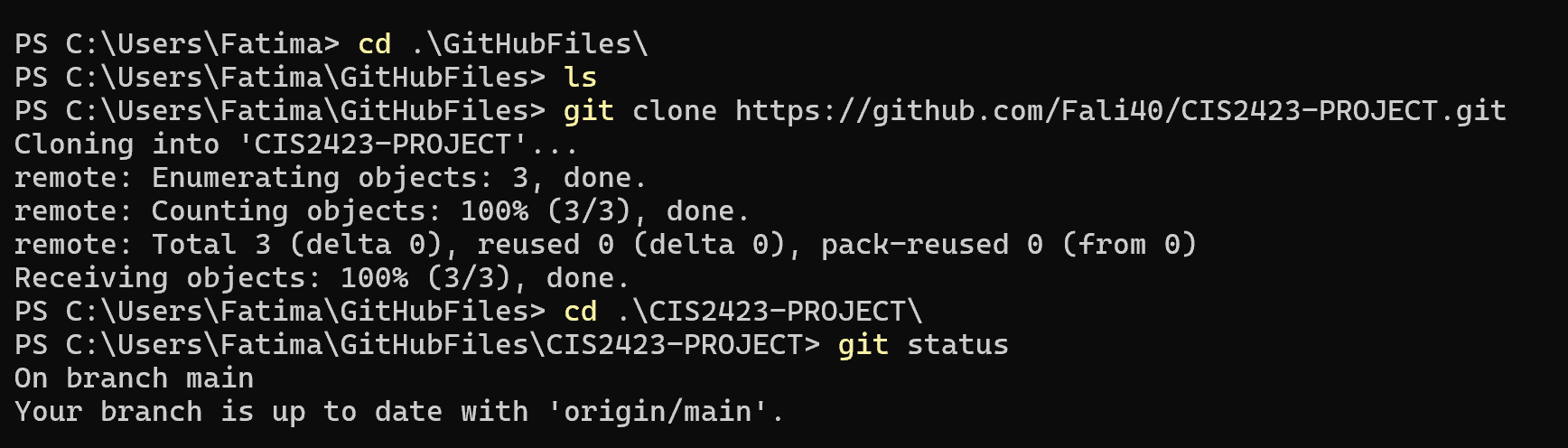
21. Upload all project files



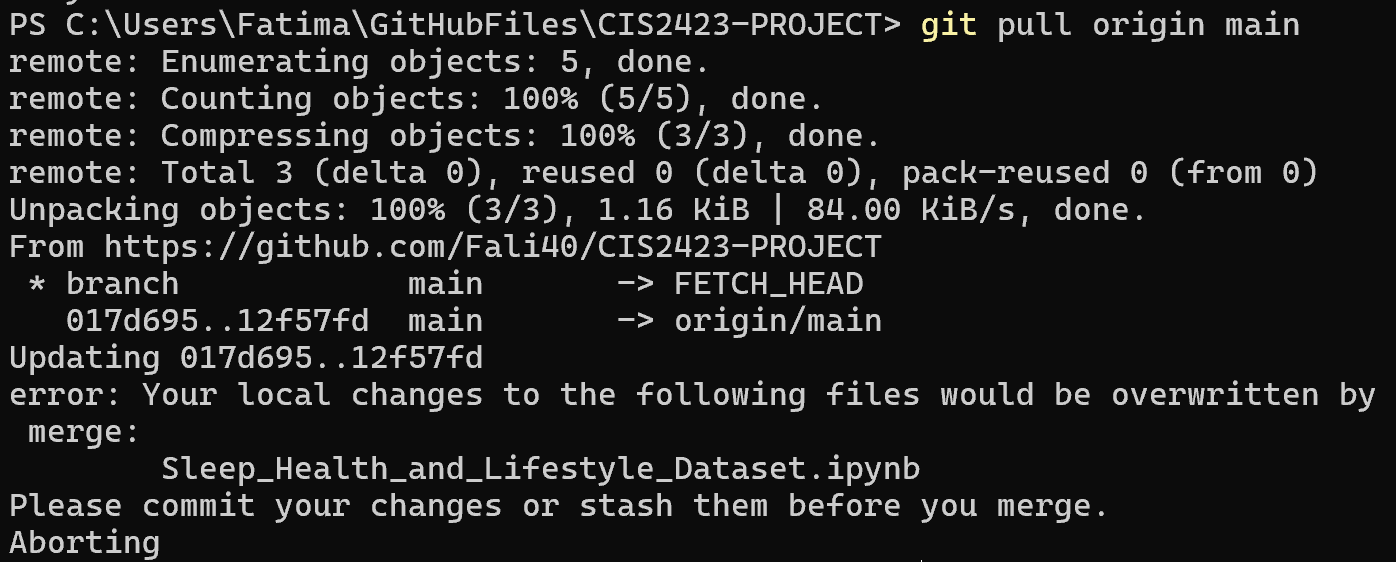
22.Configure Git with GitHub



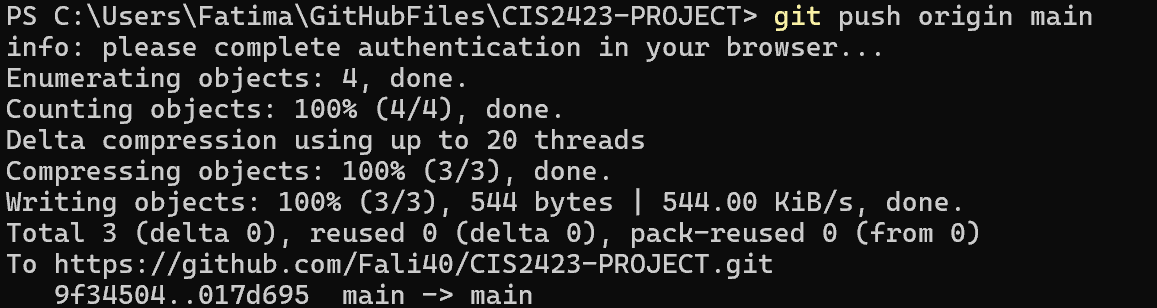
23.Clone git hub repo



24. Pull any file from Git Hub



25. Modify the pulled file and push the modified file to Git Hub



# Evidence:

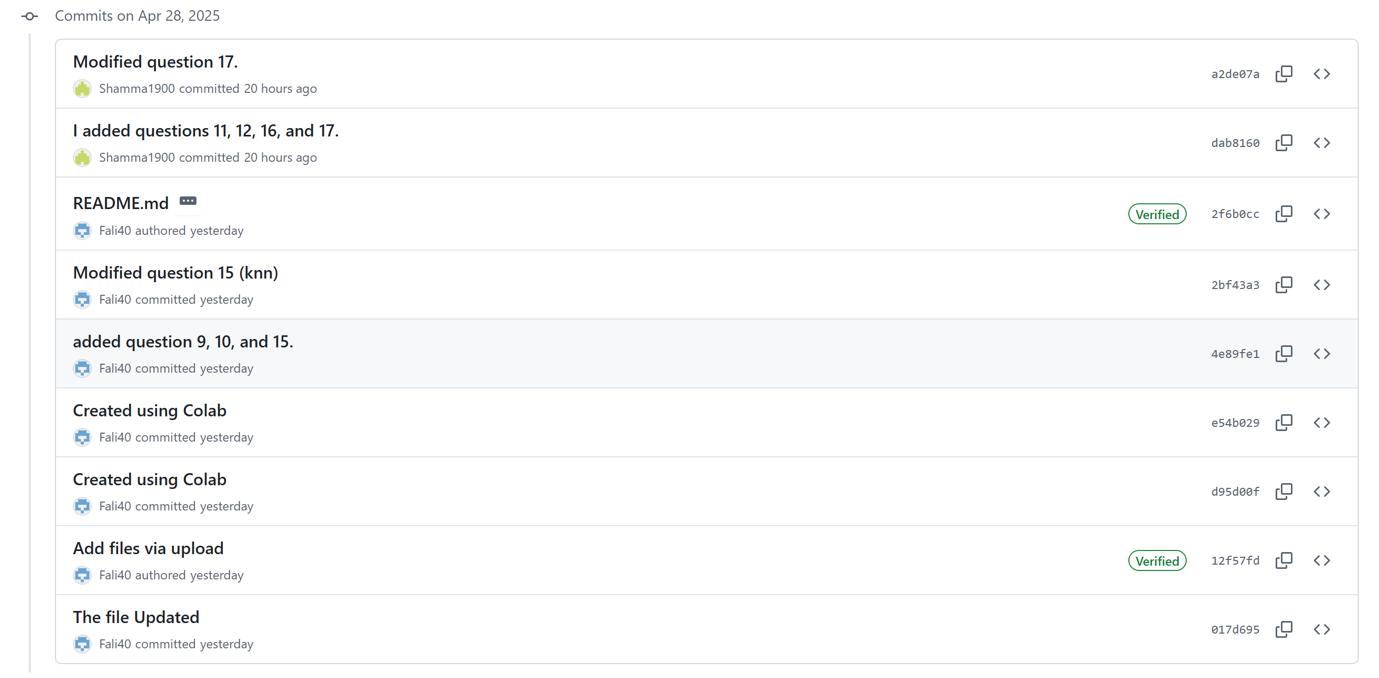
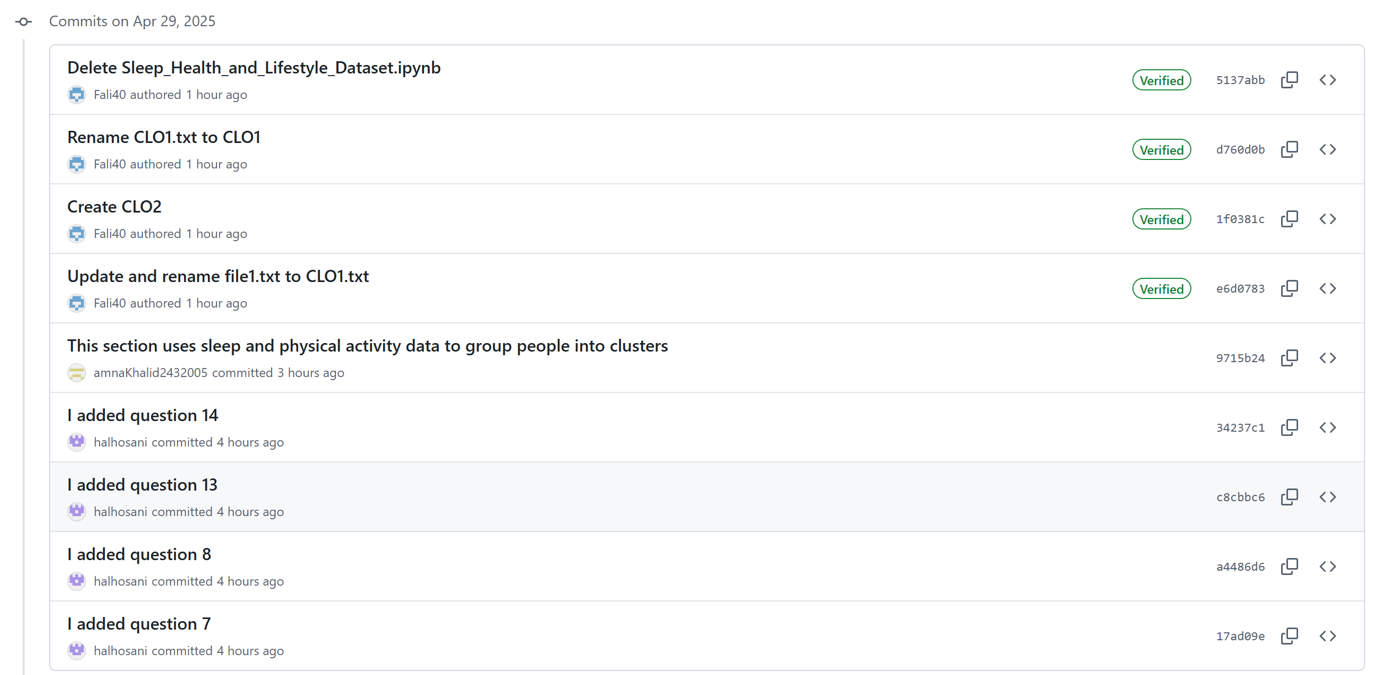
|  |  |
| --- | --- |
|  | Fatima Ali Yusuf Ali - H00535657 |
| CLO 1 | I did question 3 about the Type and purpose of the machine learning algorithm. |
| CLO 2 | * I did question 9 which was about Descriptive Statistics Report for the dependent variable "Quality of Sleep". * I did question 10 which was about visualizing the Dependent Variable "Quality of Sleep", by creating a Scatter Plot, Box plot, Histogram, and the heat map. |
| CLO 3 | I did question 15 which was about Predicting the value of the dependent variable from the different classifier:   * Logistic Regression * K-nearest Neighbor * Naïve-Bayes * Decision Tree |
| CLO 4 | We all did the git hub together, from one device:  First, we created a file in GitHub called "CIS2423-PROJECT", second thing we cloned the GitHub file into the file we created, then we added our google collab file into the GitHub, after that we continued with the pull, push and commit functions all from the same device. |

|  |  |
| --- | --- |
|  | Shamma Alghfeli - H00535715 |
| CLO 1 | I identified and justified the independent and dependent variables in the dataset. |
| CLO 2 | I performed a hypothesis test to find the Pearson and Spearman correlation between the independent and dependent variables.  Also, I assessed the performance of the dependent variable to determine whether the sample is representative of the normal population using a one-sample t-test. |
| CLO 3 | I evaluated the performance of each classification model by analyzing the confusion matrix and accuracy and subsequently identified the best-fitting classifier for the dataset.  And using the best-fit classifier, I predicted the values of the dependent variable based on the selected features. |
| CLO 4 | We all did the git hub together, from one device:  First, we created a file in GitHub called "CIS2423-PROJECT", second thing we cloned the GitHub file into the file we created, then we added us google collab file into the GitHub, after that we continued with the pull, push and commit functions all from the same device. |

|  |  |
| --- | --- |
|  | Amna Khalid Al Ali - H00495490 |
| CLO 1 | For Clo 1 I did question 1, and i have explained on this question the purpose or the reason why we are using the data set that we have choose |
| CLO 2 | For Clo 2 I have 2 questions, question 5 and question 6, for question 5 i have write the purpose of using the descriptive analysis and then for question 6 i have wrote a python function that calculates a simple statistics of the data |
| CLO 3 | For Clo 3 I have 2 questions, question 18 and question 19, question 18 i have perform a cluster analysis such as k-means and horizontal from some field of the data and then, and for question 19 after finishing question 18 and getting the results i have explained a strategies that could improve the system based on the results that we have get inside the cluster analysis |
| CLO 4 | For clo4 we have all shared during the class one device and solve the questions together |

|  |  |
| --- | --- |
|  | Hanan Khalil - H00535795 |
| CLO 1 | For clo1 I solved question 2. In this question, I Identified and justified the type of programming used for our data analysis. |
| CLO 2 | For clo2 I solved questions 7 and 8. In question 7 I Created a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample. In question 8 I Created a script for systematic sampling by giving certain condition and calculated the desc stat for the dependent variable from the sample. |
| CLO 3 | For clo3 I solved questions 13 and 14. In question 13 I Build, Train, Develop and Evaluate using Simple Regression for chosen dataset. In question 14 I Developed a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression |
| CLO 4 | We all did the git hub together, from one device:  First, we created a file in GitHub called "CIS2423-PROJECT", second thing we cloned the GitHub file into the file we created, then we added our google collab. |

# Git hub Screenshot:



# Project Deliverables

Project Report (50%)

1. **Deliverable 1**: A complete report about the purpose of data analysis, programming language chosen for data analysis, types of machine language algorithm to be analyzed and the list of variables chosen for analysis [CLO1]
2. **Deliverable 2**: A detailed report about the summary of the data, sampling, graphs/charts to analyze the data, relationship between variables, evaluating assumptions using hypothesis testing, predicting the variables using the regression model [CLO 2,3]
3. **Deliverable 3**: A comprehensive description about the data model created using classification and clustering algorithm of machine learning. It should involve the narrative about the data model is optimized to predict the variables and bow the best fit model has been chosen. [CLO 2,3]
4. **Deliverable 4**: Complete narration about data versioning using Git. [CLO 4]
5. **Written Communication:** Complete report with specified format and structure [12 points].

Project Oral (50%)

1. **Oral Communication**: Each student will be assessed in the form of individual oral defense with PowerPoint presentation. [All CLOs] [10 Marks]
2. **Follow-up Questions and Discussion** [All CLOs] [30 Marks]

* **[Collaboration]** [10 Marks].