Amani Bdair

Final project

**Quality of sleep associated with occupations, stress levels, and physical activity.**

* **Introduction:**

My project is about choosing a dataset and analyzing it using python. Focusing on the quality of sleep and its relationship with a person's occupation, stress levels and their physical activity. I eliminated the columns that were useless in my data set, and I used liner regression model to analyze the dataset using graphs and charts.

**Discuss the significance of analyzing patient readmission rates:**

Analyzing patient's readmission rates is important to understand the reasons for patient's readmission, therefore, the health organization can implement a strategy to reduce the patient readmission rate to its lowest level. The health organizations that have a high readmission rate means they are having low patients' quality services. The rate gives the health organization an indicator that they need to improve their patient's quality services. Lowering patient readmission rates is significant in reducing the cost of hospitals and other healthcare organizations. In addition, this means that the hospital is in compliance with CMS hospital readmissions initiatives and the hospital is having high patient and medical quality standards.

* **Data Collection:**

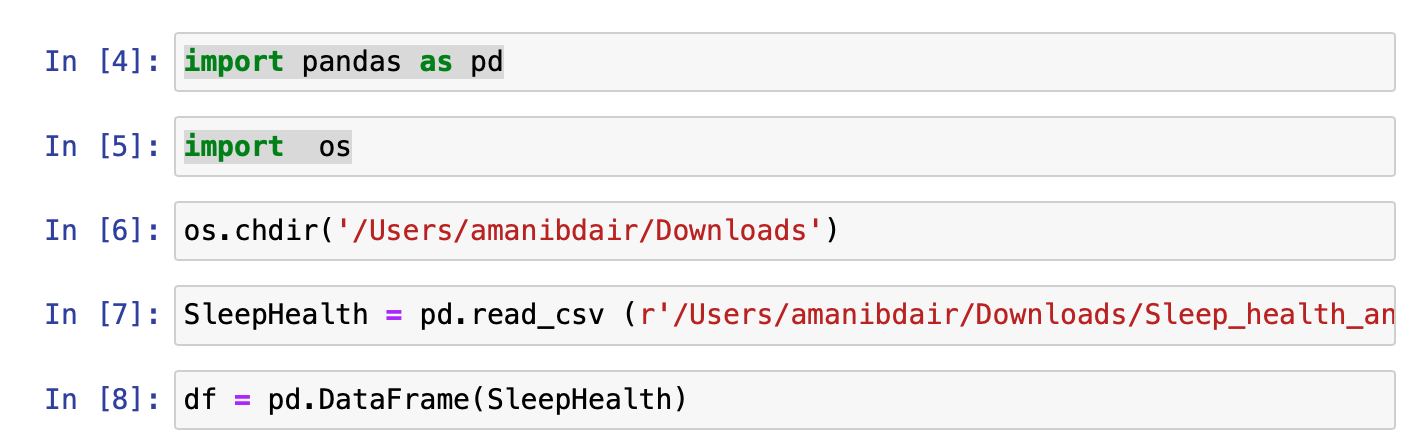
My data set is “Sleep Health and Lifestyle Dataset” from Kaggle website. The dataset contains 400 rows and 13 columns focusing on the important factors that affect quality of sleep. the data set features are:

* The Lifestyle Factors includes columns: Analyze physical activity levels, stress levels, Daily Steps and BMI categories.
* The Cardiovascular Health factors: blood pressure (“The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure”) and heart rate (“The resting heart rate of the person in beats per minute”).
* Finaly the sleep disorder factor: if a person has any sleep disorder like insomnia and sleep apnea
* The columns I need to analyze in my data set are:
* Occupation: The occupation or profession of the person. (Nurse, Doctor , Engineer, Lawyer, Teacher, Accountant, Salesperson, Software Engineer, Scientist, Sales Representative and Manager)
* Stress Level (scale: 1-10): A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
* Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.
* **Data Preprocessing:**

I dropped the data that I don't need which are: person ID, Gender, Age, sleep Duration, BMI Category, blood pressure, heart rate, Daily steps, sleep disorder.

For data processing:

* Importing pandas library as “pd”
* Importing “os” module which provides a way of using operating system dependent functionality.
* Importing my dataset to python using ‘csv’ file format

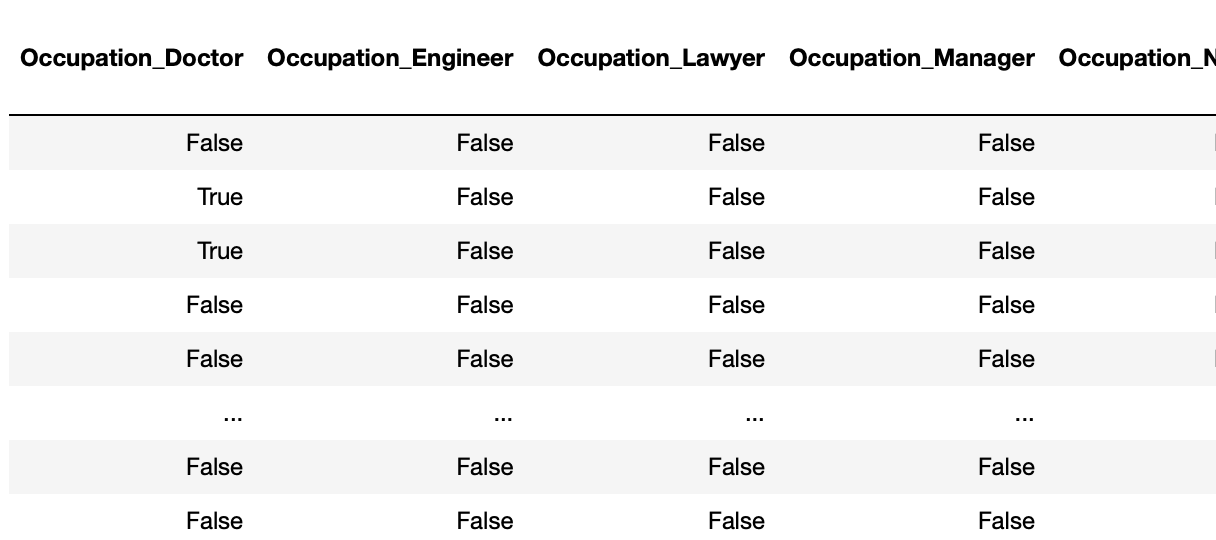


* Installing matplotlib library using pip and this for creating static, animated, and interactive visualizations in Python. I used this to create pie and bar chart
* Since my data is large, I used “import NumPy as np” which provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
* To process the “occupation” data category, I encode this column. Converting categorical variables into a numerical format that machine learning algorithms can understand.

Converting the “occupation column from this :

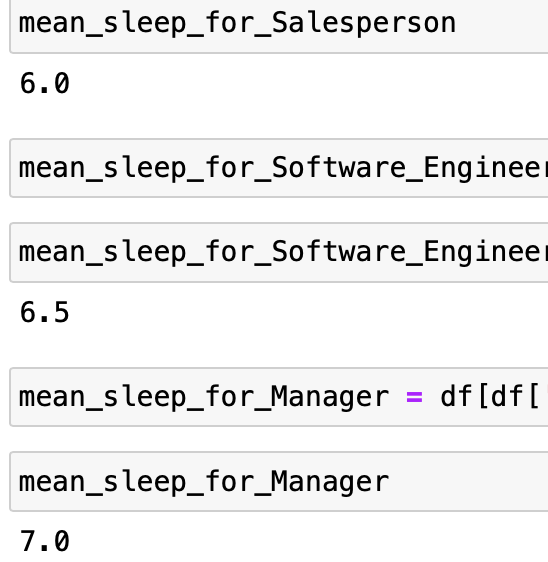


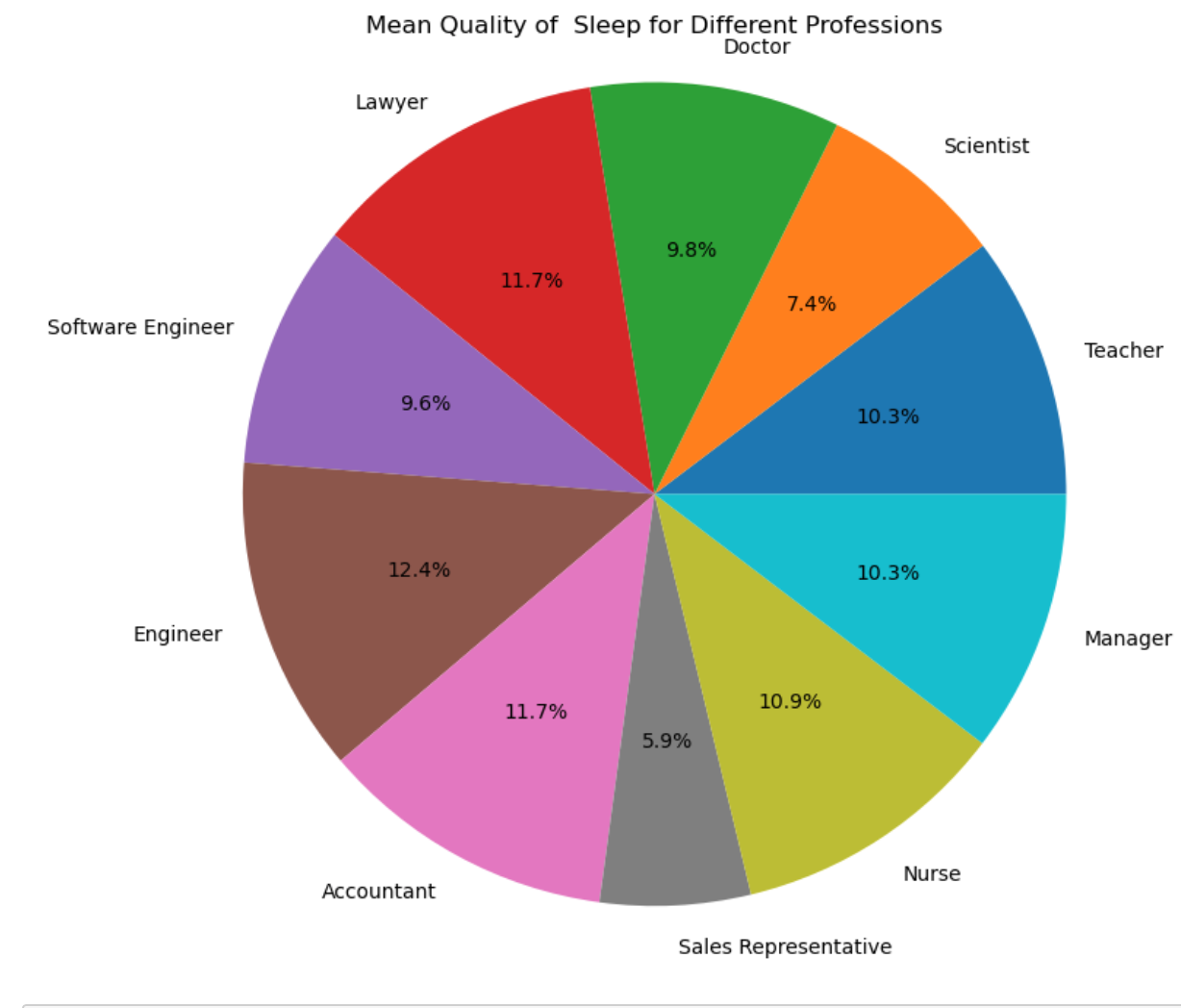
To this:



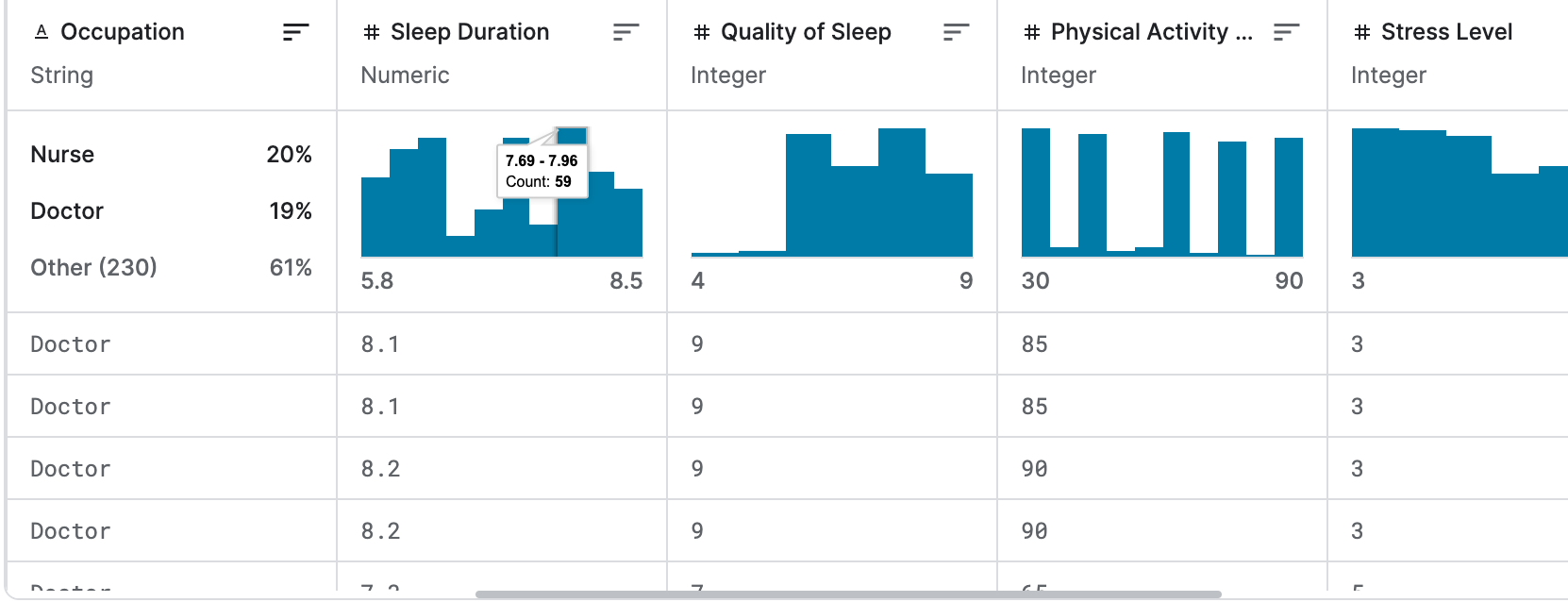
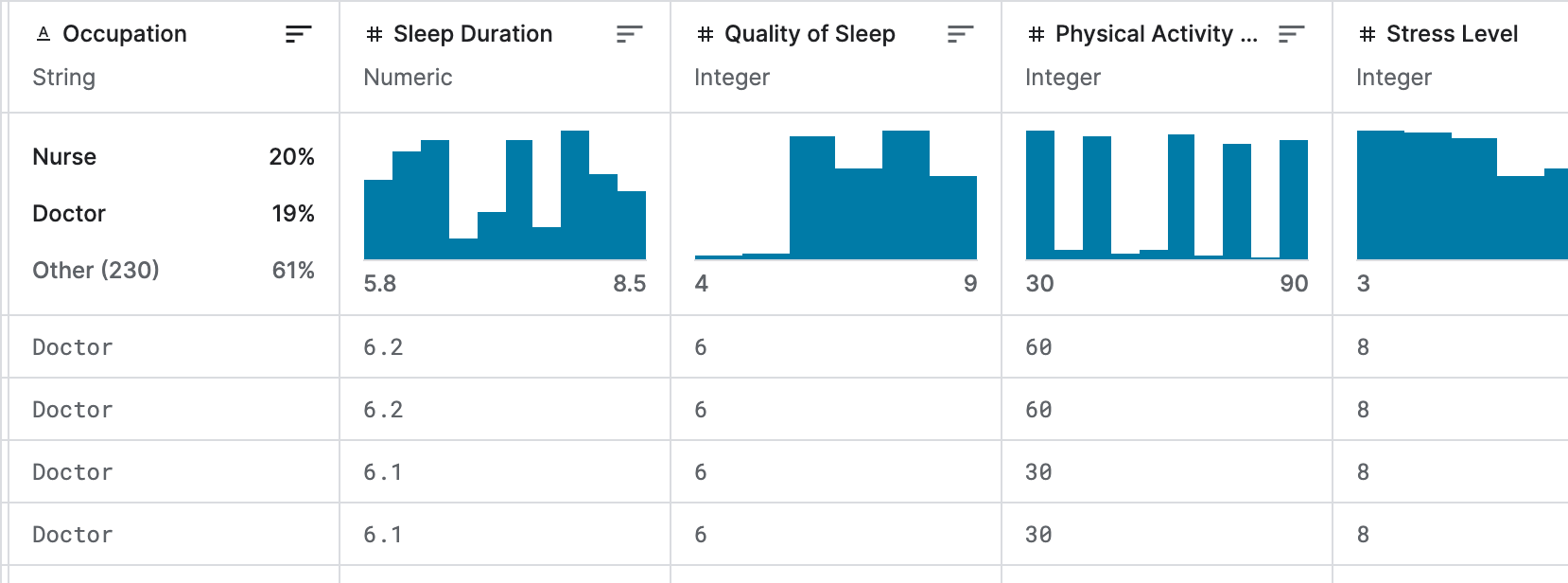
* Importing seaborn as sns . This is statistical data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics. I used it for linear Regression graphs.
* **Exploratory Data Analysis (EDA):**

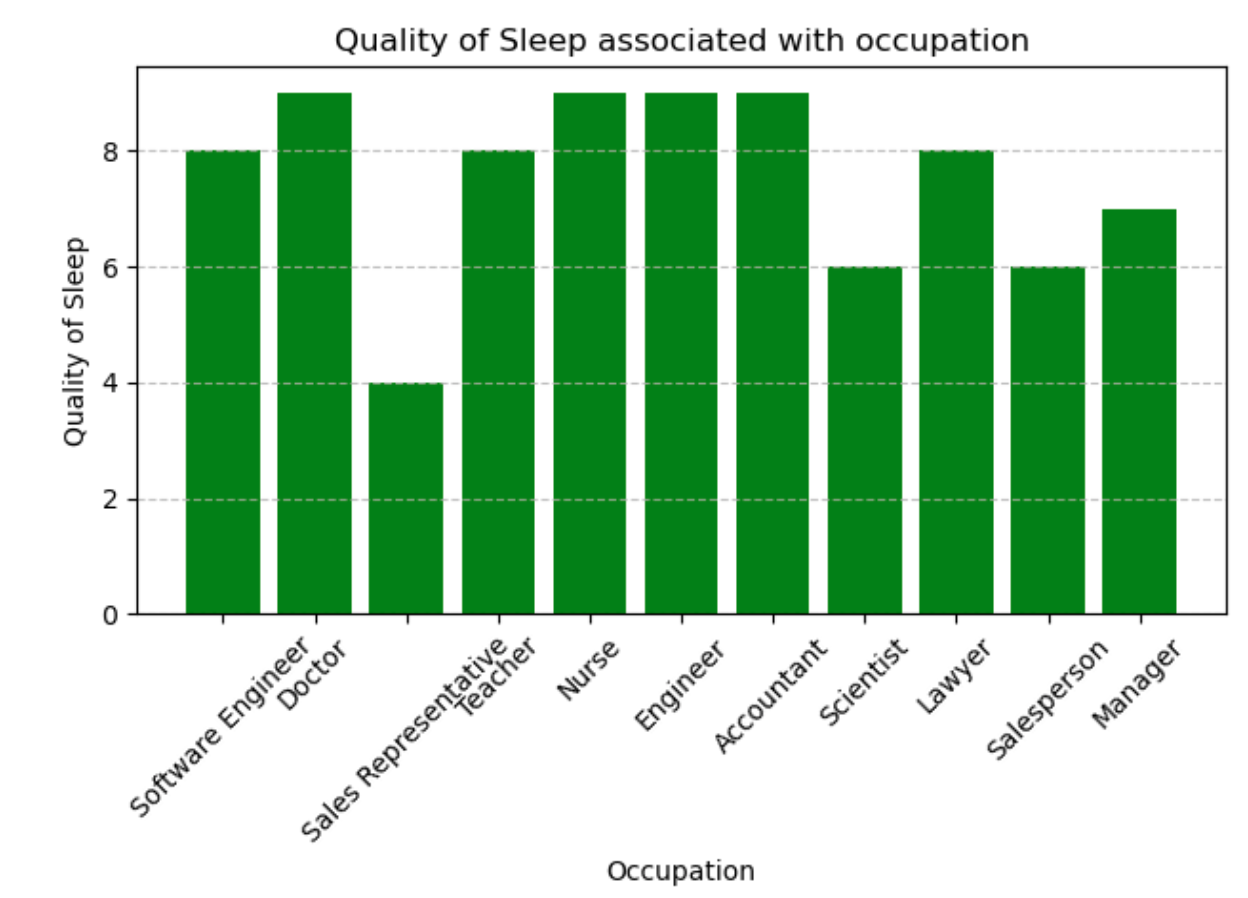
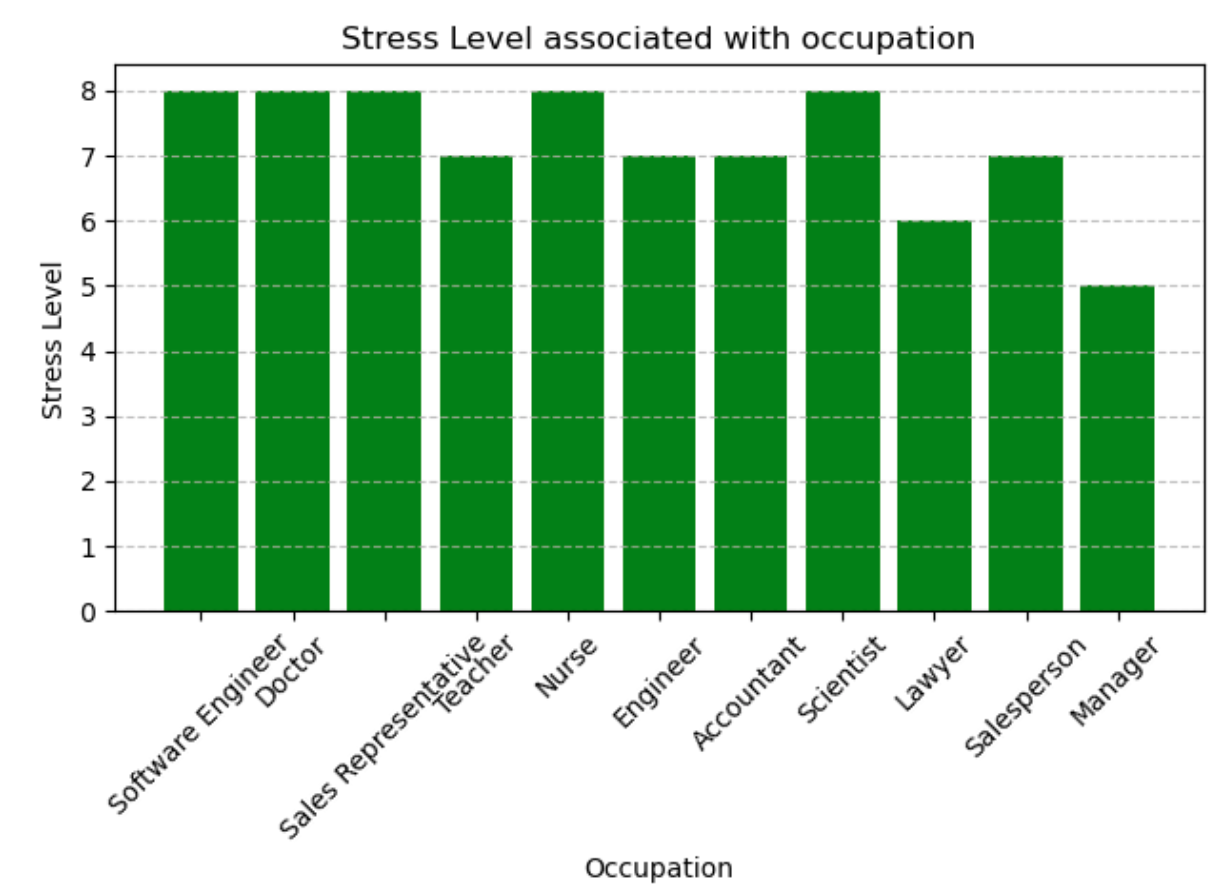
To understand the relationship between the quality of sleep and other lifestyle categories I did the following in python:

* Calculated the mean for each occupation associated with quality of sleep.
* Pie chart to visualize my data. In the chart, Engineers have the highest mean by 12.4% and the lowest is Sales Representative by 5.9%. This means that Engineers have the highest quality of sleep than other professions.



* Compared the quality of sleep with the stress level of each occupation by using bar chart: the result shows that the “stress level” of someone's occupation has no actual effects on the quality of sleep. That is because every person has their own stress level that is not associated with his profession. For example, the stress level of all doctors in the data set range from 3 to 8. Also ,The quality of sleep for sales representatives is 4 and their stress level is 8. while the lowest stress levels are managers, and their quality of sleep is more than 6 which is not the highest level of quality of sleep. The occupation that has the highest quality of sleep is engineer; however, engineer stress level is 7 which is considered average to high.





* Scatter chart to compare the quality of sleep with stress level. It shows negative liner which the lowest stress level leads to high quality of sleep.

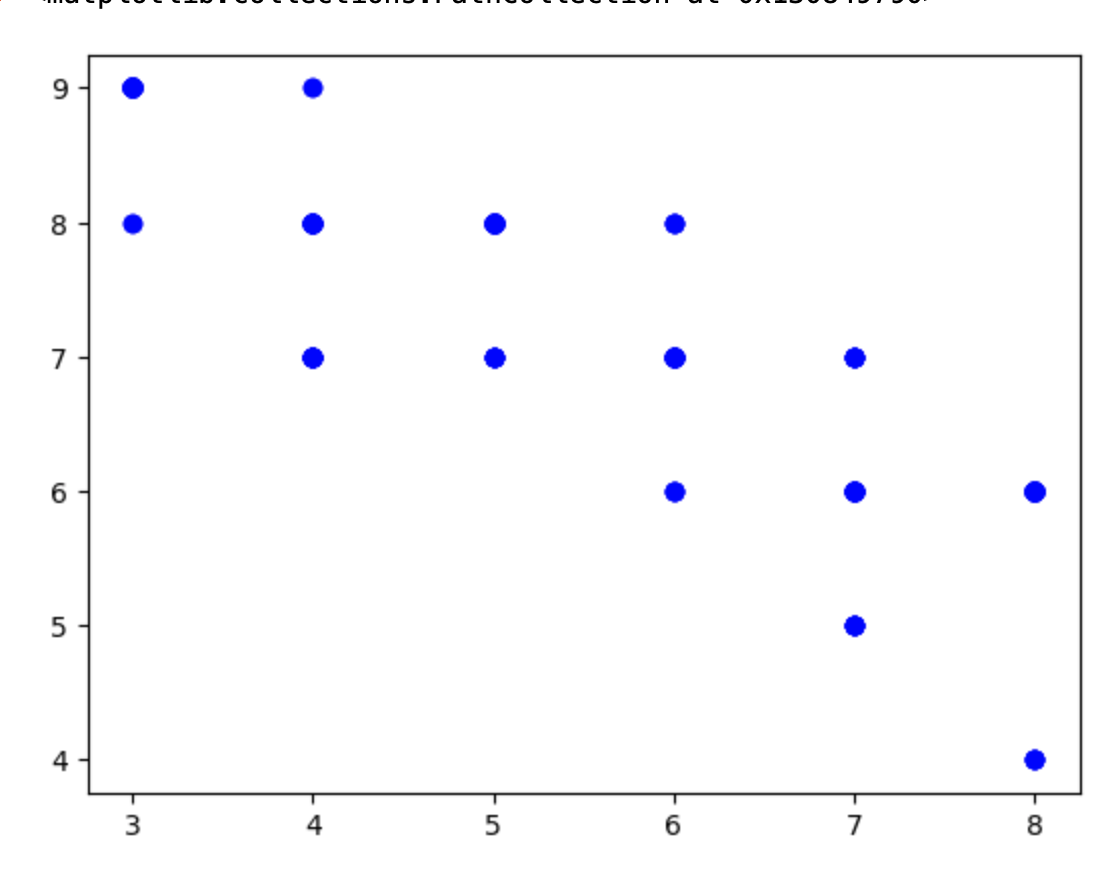
Y == ‘quality of sleep’

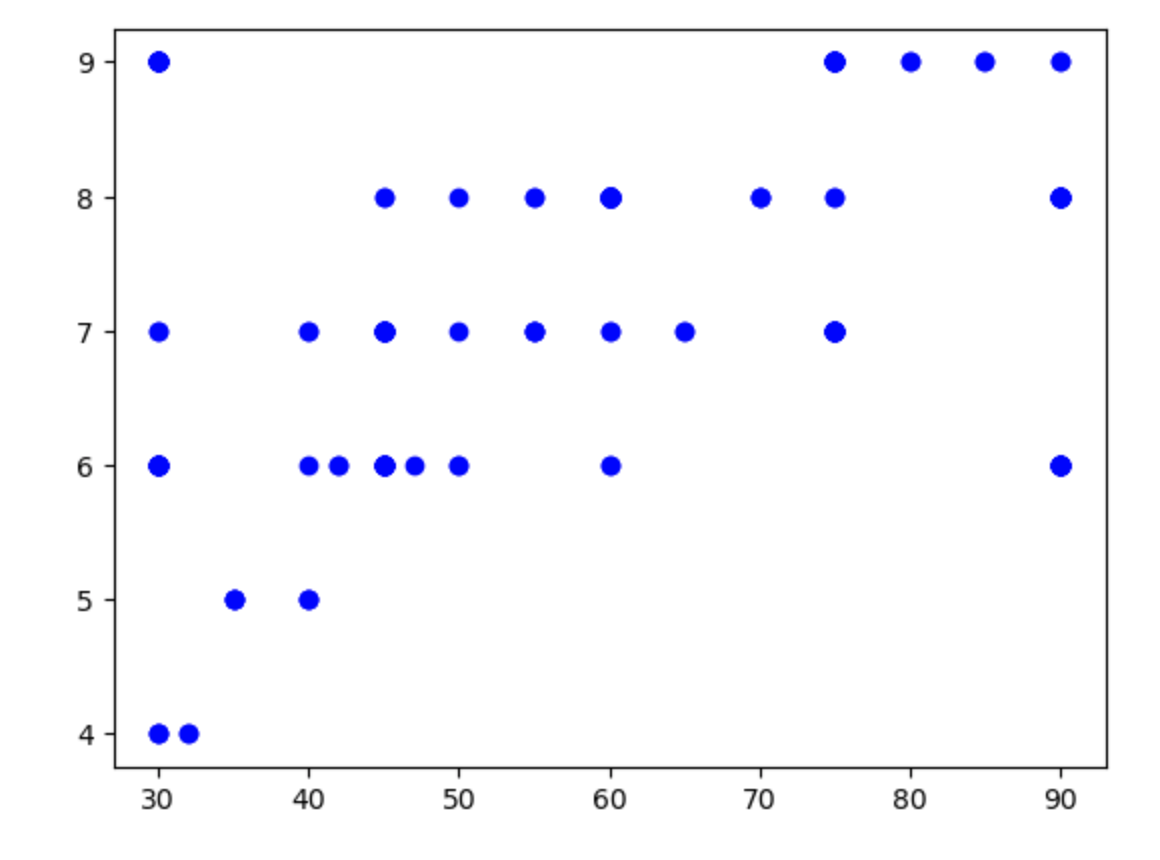
X == ‘stress level’

* The next scatter chart compares physical activity with the quality of sleep. The chart shows that physical activity does not affect the quality of sleep. People score 30 in physical activity and 4 and 9 in sleep quality.

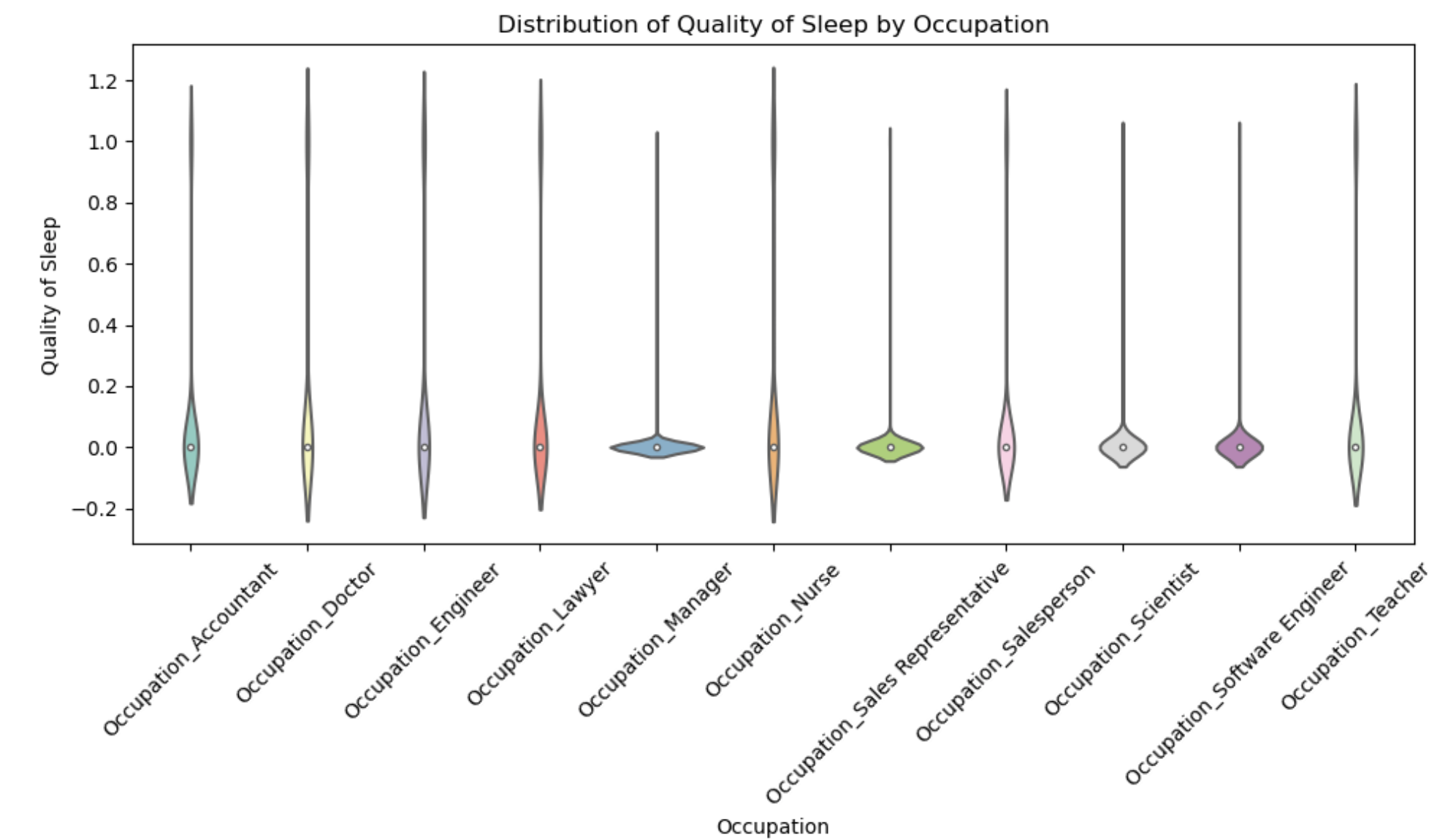
Y == ‘quality of sleep’

X == ‘physical activity’



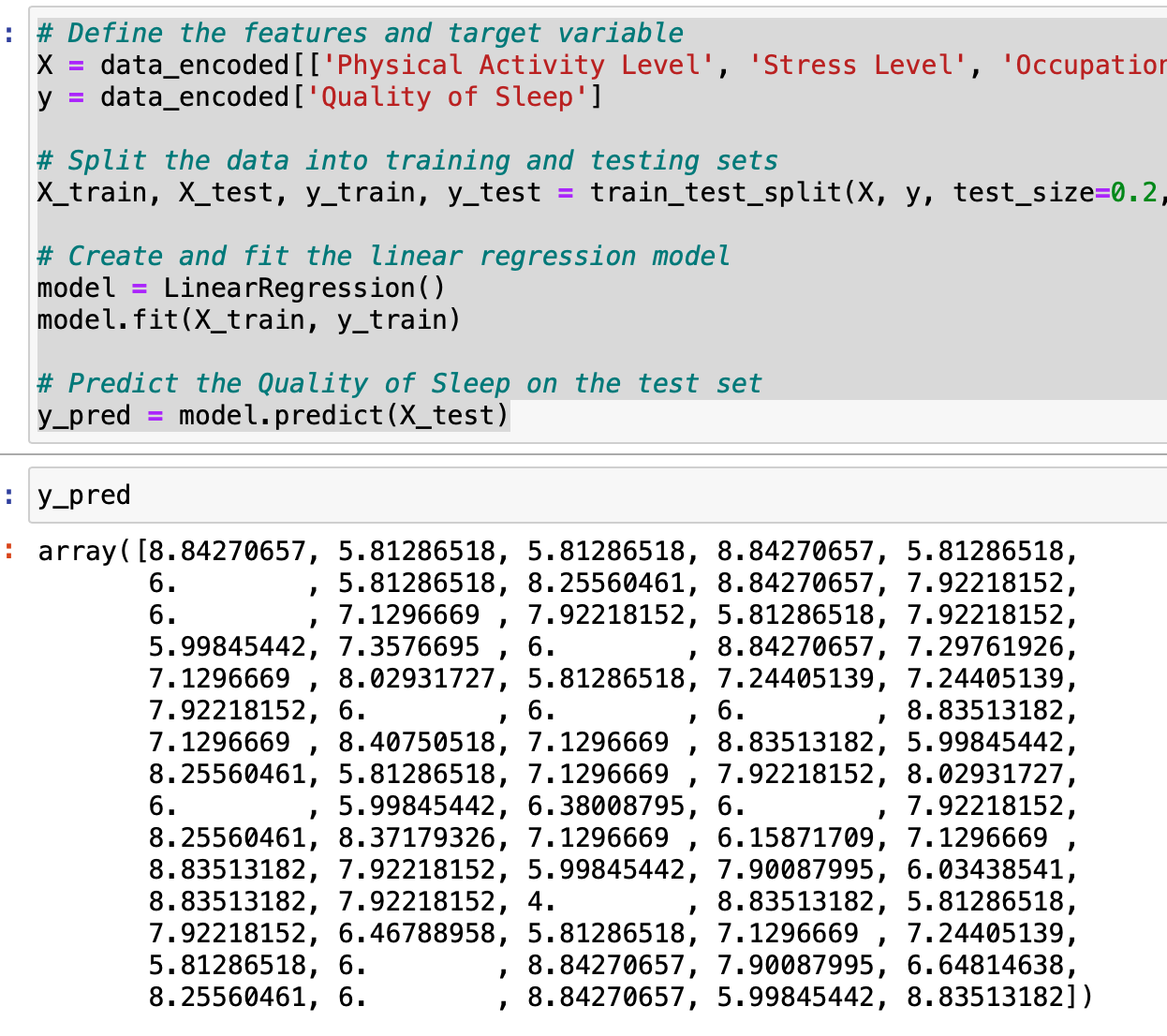


* I created a violin polt of “Distribution of Quality of Sleep by Occupation” as shown down below:

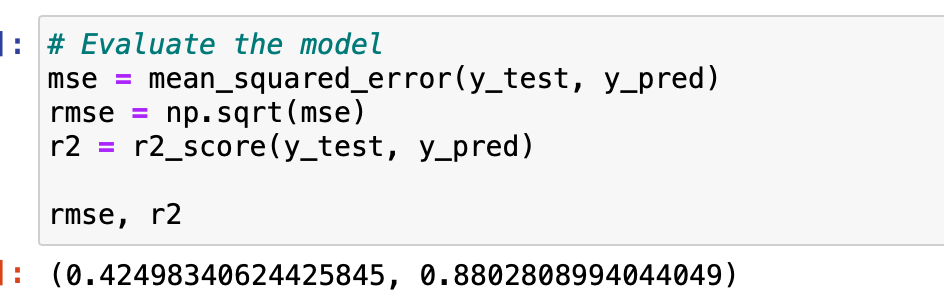


* **Model Development:**
* Splitting the dataset into training and test sets:
* I define features and target variable: X is the feature variable and y are the target variable.
* Then splitting the data into training and testing set : the test\_size is %20 and the random state is 42
* Creating and Fitting the Linear Regression Model
* Making Predictions: “y\_pred”

Down below is the result of the predicting value of the quality of sleep:

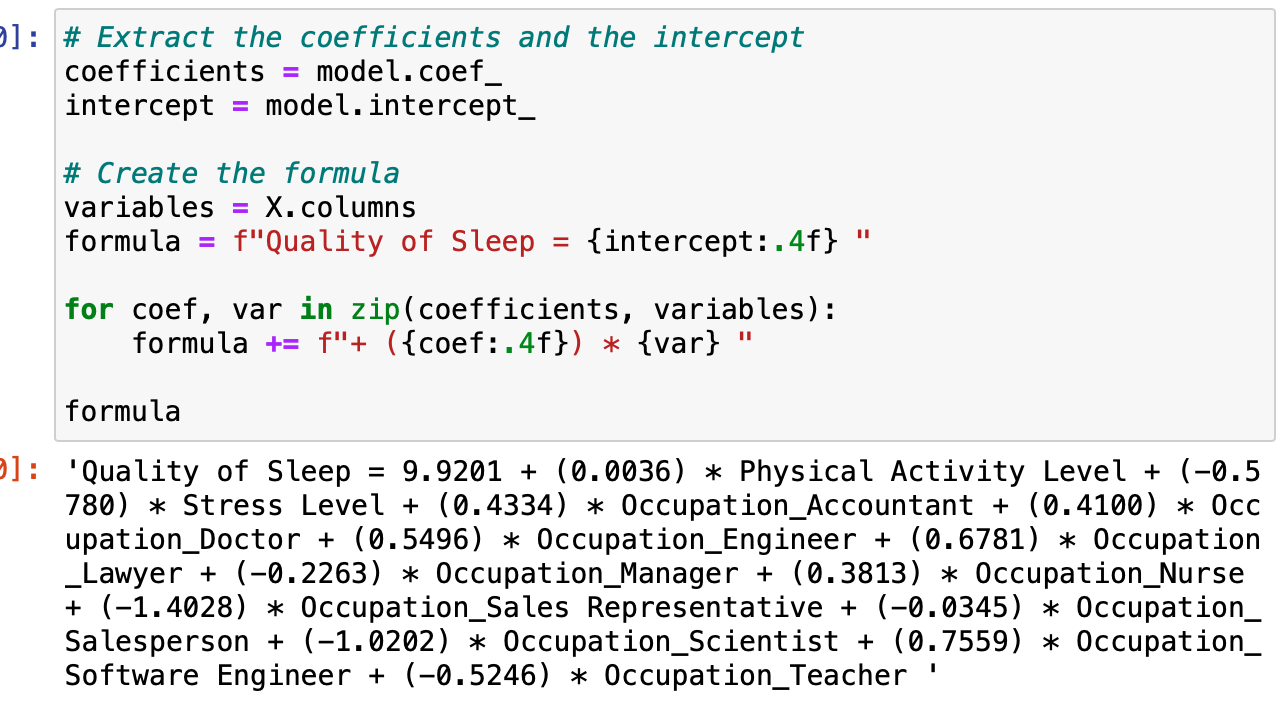


* **Model Evaluation:**
* Root Mean Squared Error (RMSE): measures the average magnitude of the errors (residuals) between the predicted values and the actual values. Which results : 0.42.
* R-squared (R2): indicates the proportion of the variance in the quality of sleep. Which results: 0.88



* **Model** **Interpretation**:
* Interpret Coefficients: The coefficients represent the change in the target variable (quality of sleep) associated with a one-unit change in the corresponding feature, holding all other features constant. Positive coefficients indicate that an increase in the feature value is associated with an increase in the quality of sleep, while negative coefficients indicate the opposite. The magnitude of the coefficients indicates the strength of the relationship between each feature and the quality of sleep.
* Intercept: The intercept represents the expected value of the quality of sleep when all features are equal to zero. It accounts for the baseline level of quality of sleep that cannot be explained by the features included in the model.
* The formula that was created epresents the linear regression model in terms of the features and their respective coefficients, along with the intercept term.

The formula and the results are down below:



* **Conclusion:**
* Sleep Health and Lifestyle Dataset was collected from Kaggle website
* Dropping columns that I don't need in my dataset
* Installing and importing libraries and functions to python
* Calculating the mean of the ‘occupations’ related to the ‘quality of sleep’ and visualizing it using the pie chart.
* Used bar graph to compare the ‘occupation’ with ‘stress level’ and ‘quality of sleep’. It shows no relation of ‘stress level’ with ‘occupation’ and there is a relation between the ‘quality of sleep’ and ‘occupation’.
* Scatter charts: negative relationship between ‘quality of sleep’ and ‘stress level’. No relation between ‘physical activity’ and ‘quality of sleep’.
* Using training and test sets to develop the model
* For model evaluation, Root Mean Squared Error and R-squared have been used.
* Using Coefficients and Intercept for Model Interpretation.

The end.

References:

Tharmalingam, L. (2023, September 18). *Sleep health and lifestyle dataset*. Kaggle. <https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>