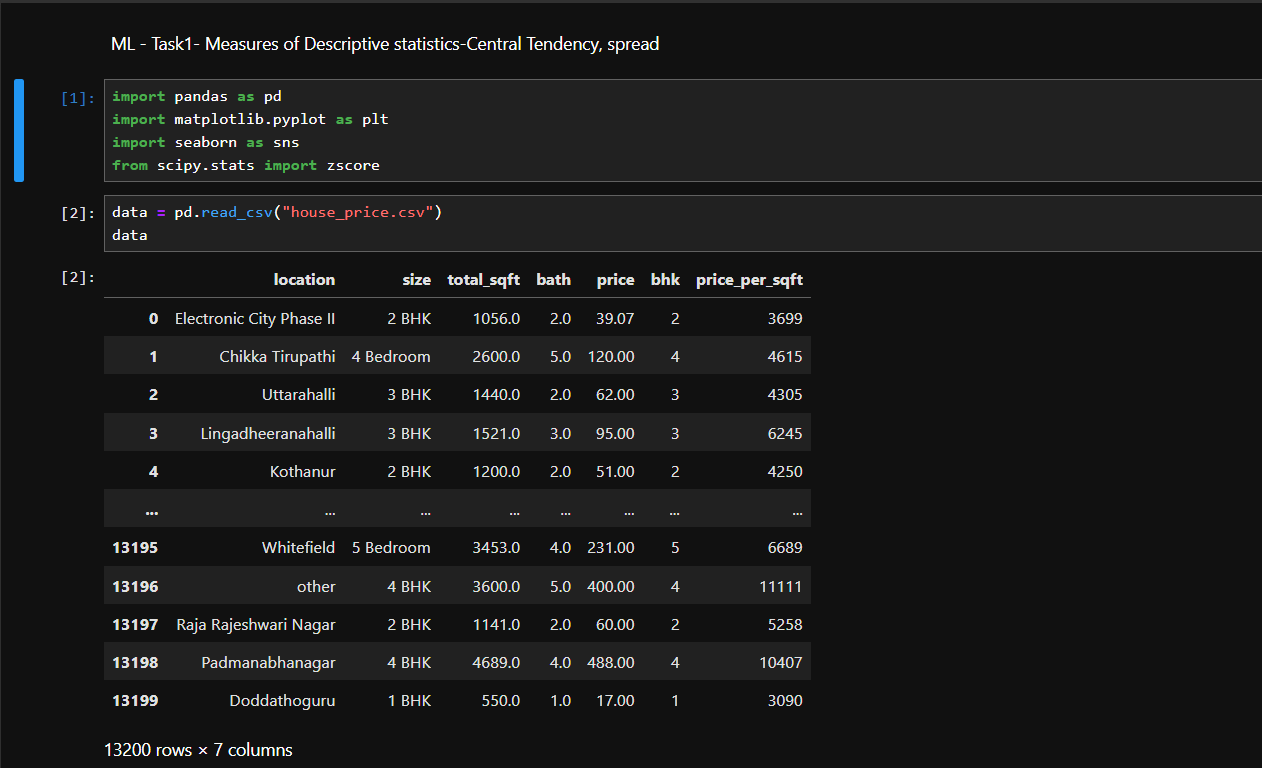
**MACHINE LEARNING ASSIGNMENTS**

**TASK:1 -**

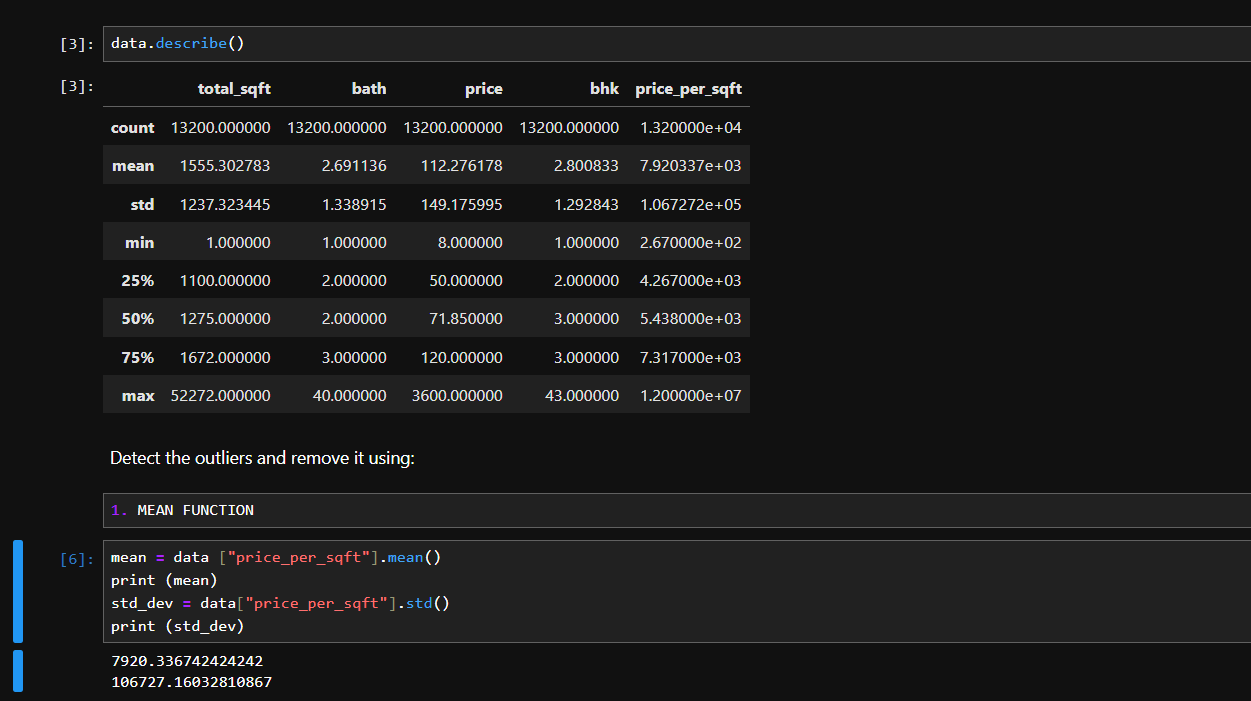
**MEASURES OF DESCRIPTIVE STATISTICS- CENTRAL TENDENCY, SPREAD**

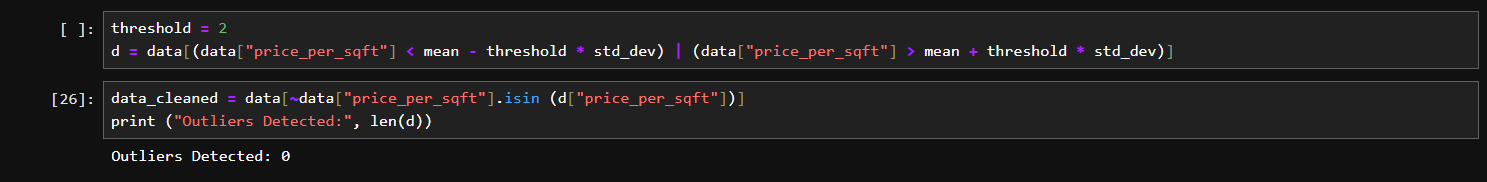
You are given house\_price.csv which contains property prices in the city of Bangalore. You need to examine price per square feet do the following:



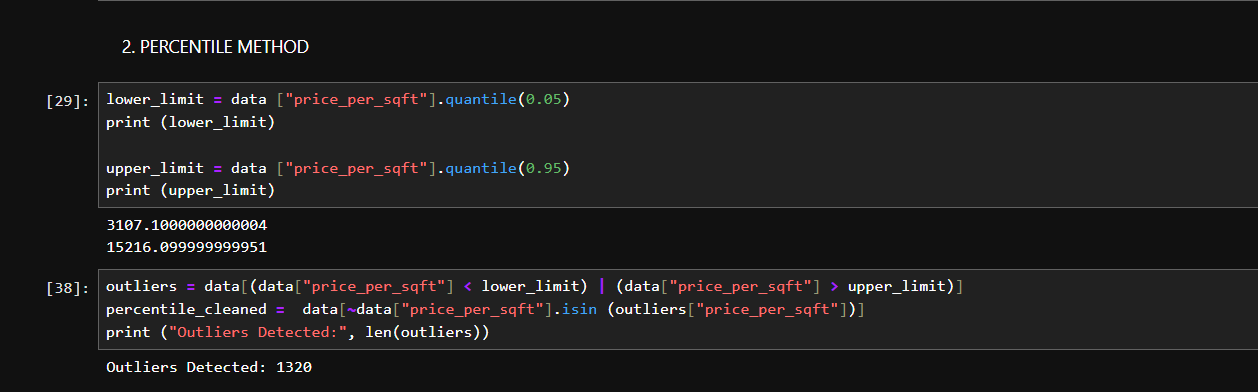
Detect the outliers and remove it using:

**1.Mean Function**





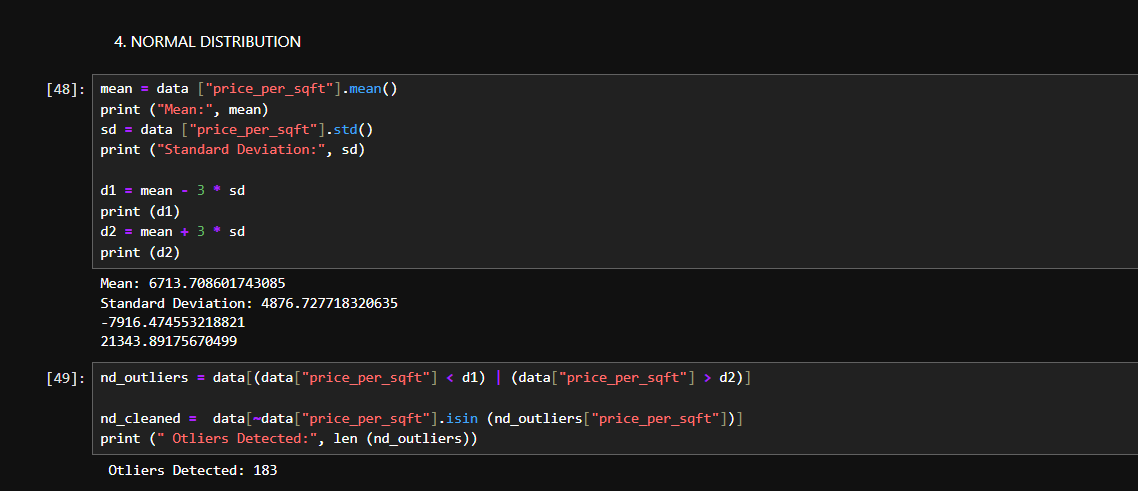
**2.Percentile method**



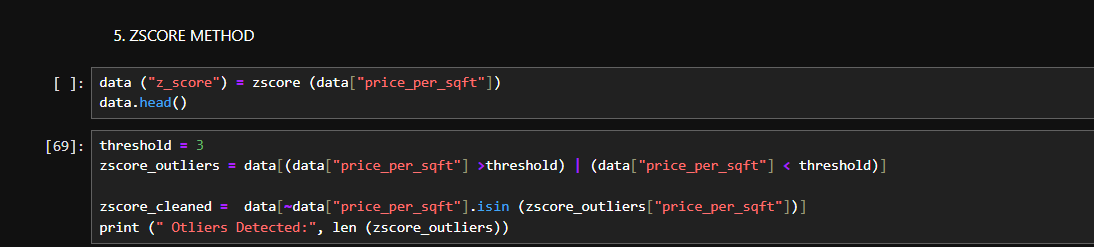
**3.IQR (Inter Quartile Range method)**

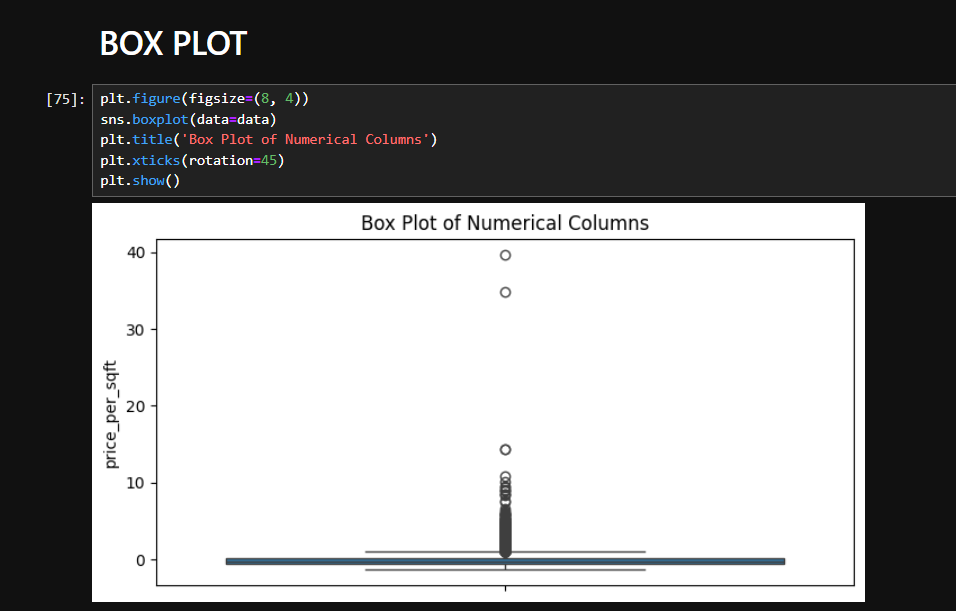


**4.Normal distribution**

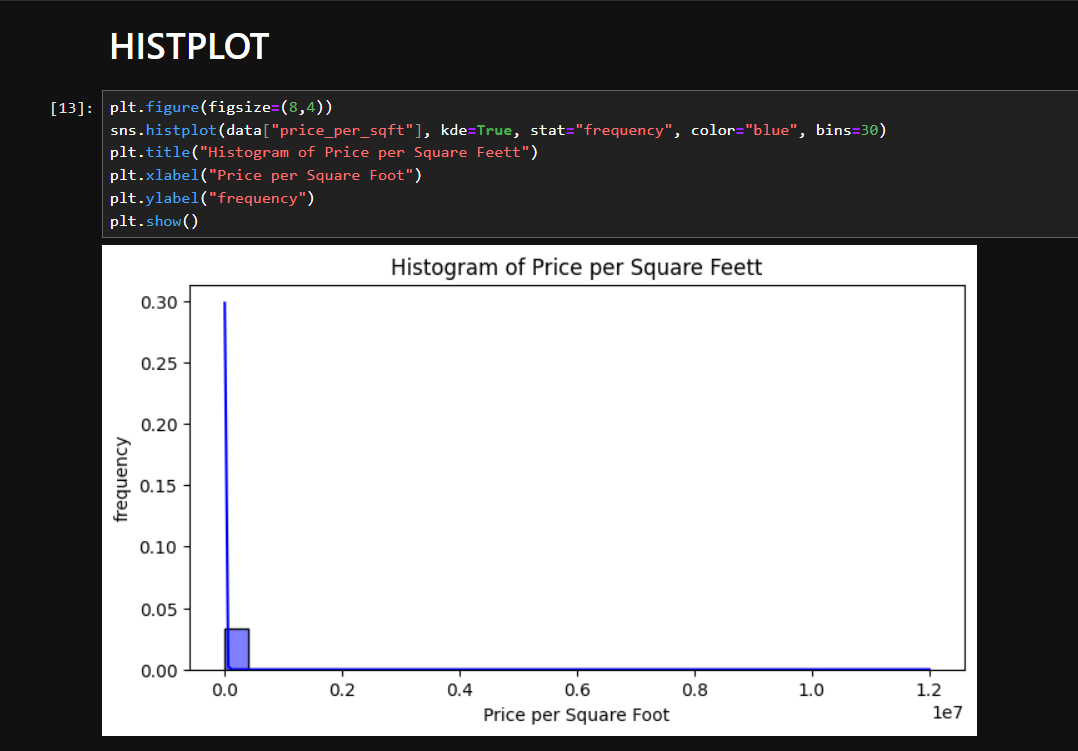


**5.Zscore method**

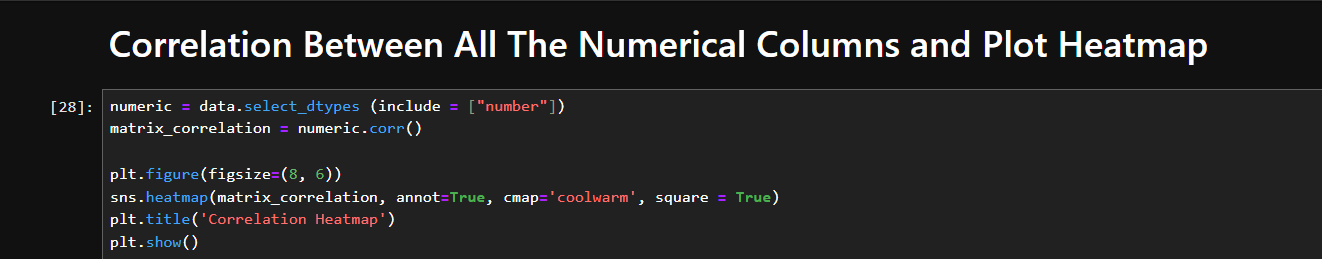
****

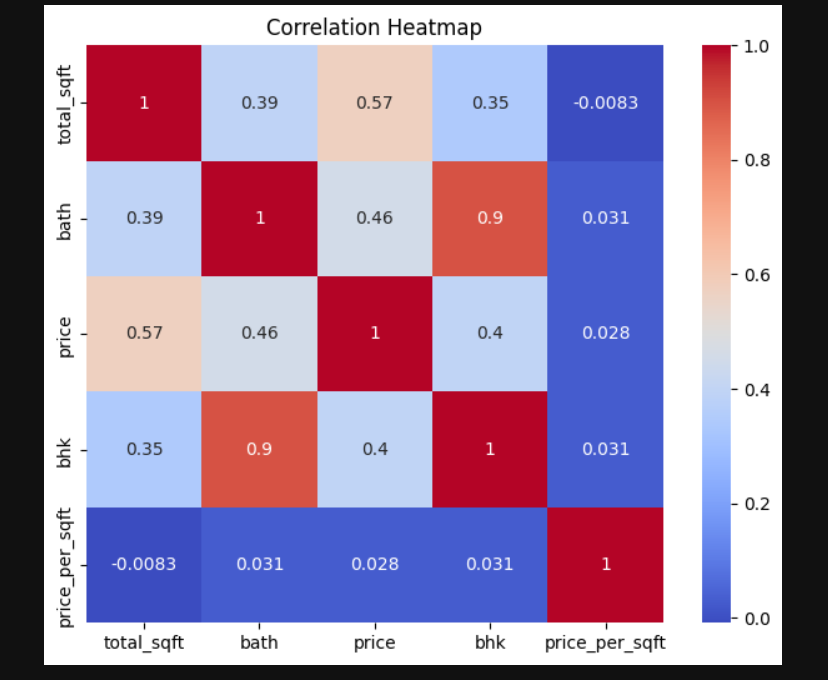
a). Plot the box plot(for all the numerical columns) 

b). Histplot(to check the normality of the column(price per sqft column).

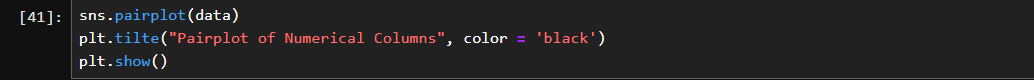


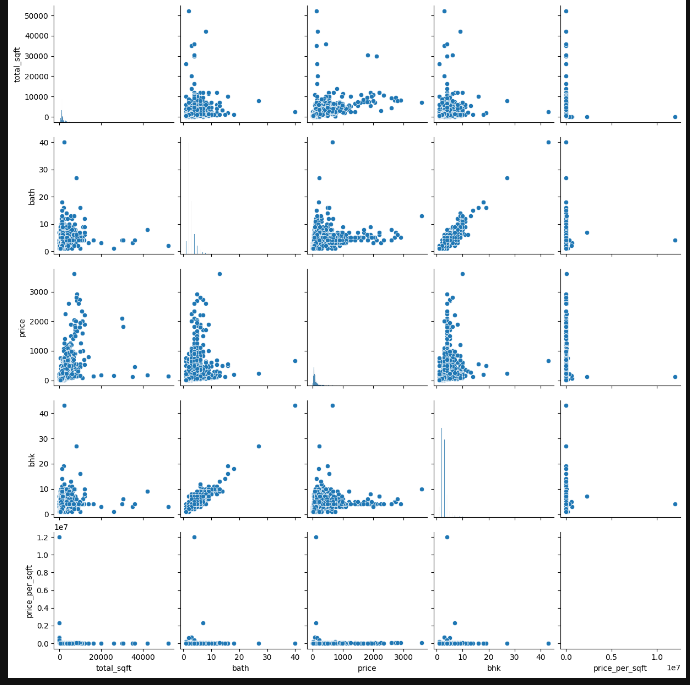
Check the correlation between all the numerical columns and plot heatmap.





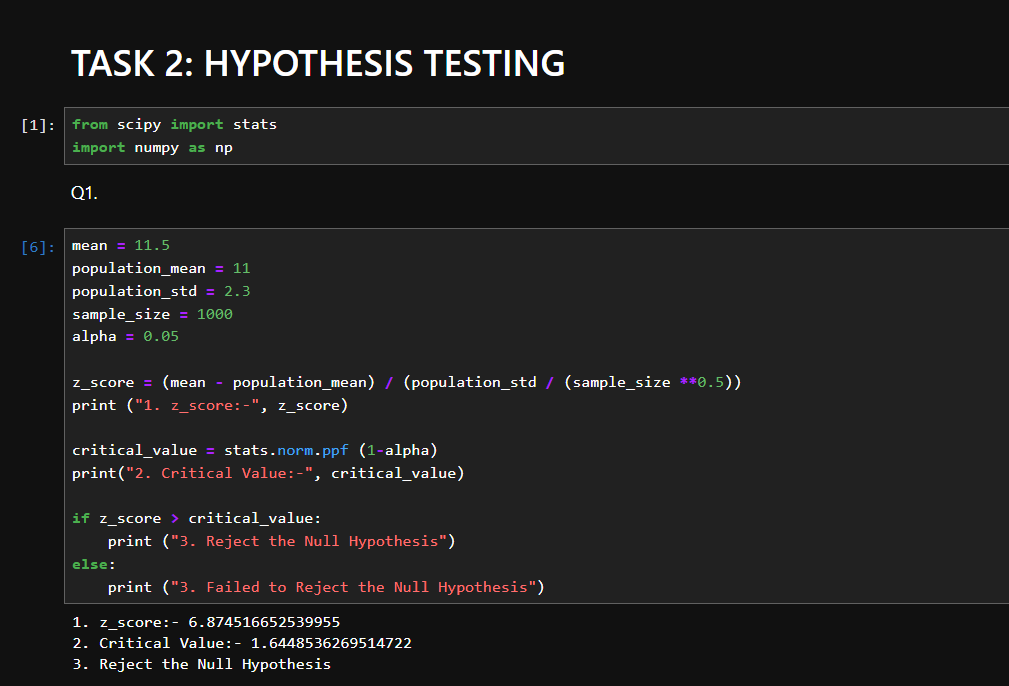
Scatter plot between the variables to check the correlation between them.



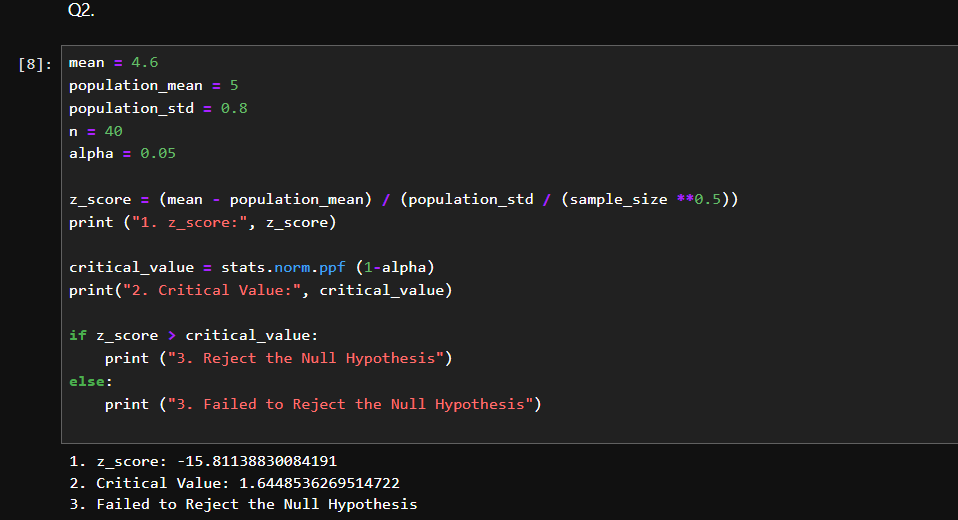


**TASK: 2 - HYPOTHESIS TESTING**

**Q1**. Suppose a child psychologist claims that the average time working mothers spend talking to their children is at least 11 minutes per day. You conduct a random sample of 1000 working mothers and find they spend an average of 11.5 minutes per day talking with their children. Assume prior research suggests the population standard deviation is 2.3 minutes.Conduct a test with a level of significance of alpha = 0.05.



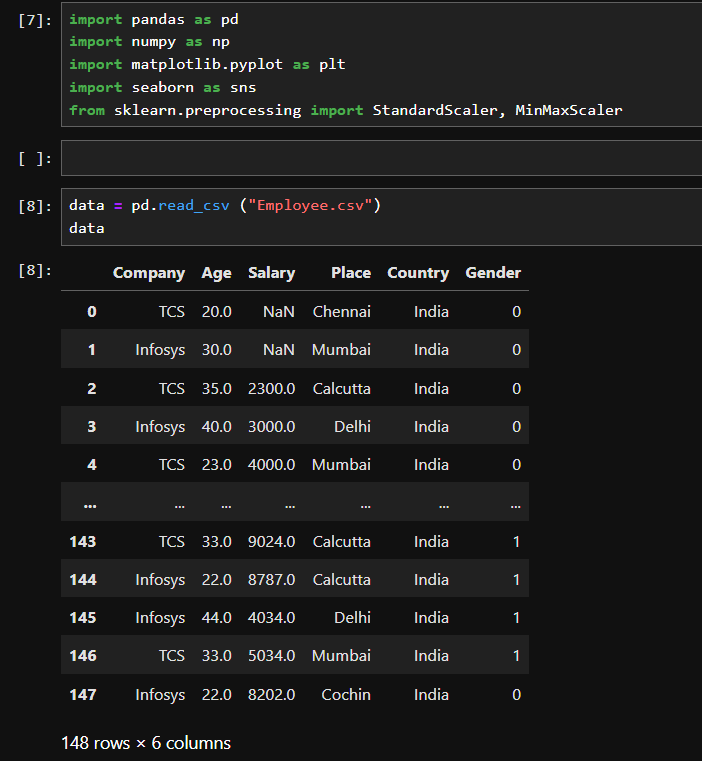
**Q2**. A coffee shop claims that their average wait time for customers is less than 5 minutes. To test this claim, a sample of 40 customers is taken, and their wait times are recorded. The sample mean wait time is found to be 4.6 minutes with a standard deviation of 0.8 minutes. Perform a hypothesis test at a significance level of 0.05 and determine whether there is enough evidence to support the coffee shop's claim.



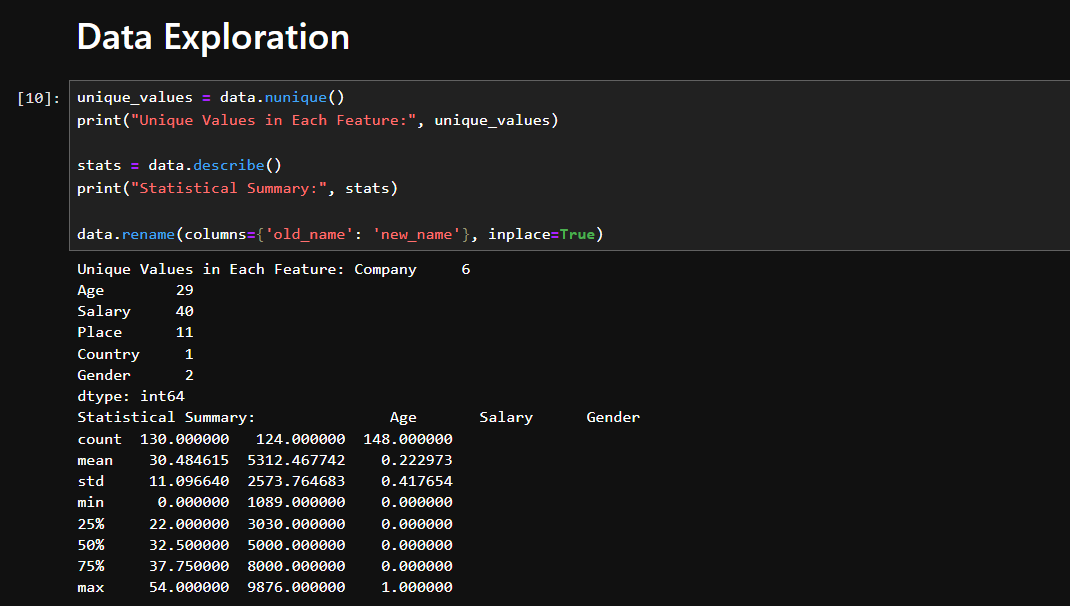
**TASK:3- DATA PROCESSING**

**Objective:**

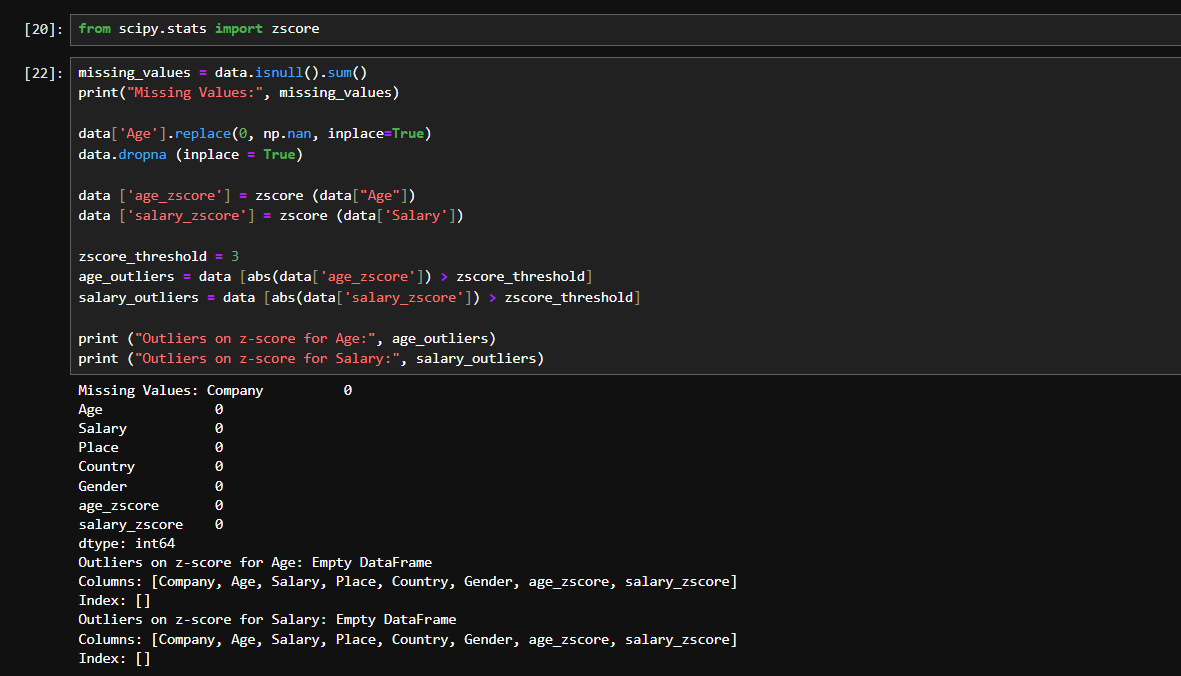
The main objective of this project is to design and implement a robust data preprocessing system that addresses common challenges such as missing values, outliers, inconsistent formatting, and noise. By performing effective data preprocessing, the project aims to enhance the quality, reliability, and usefulness of the data for machine learning.



**Data Exploration:** Explore the data, list down the unique values in each feature and find its length. Perform the statistical analysis and renaming of the columns.

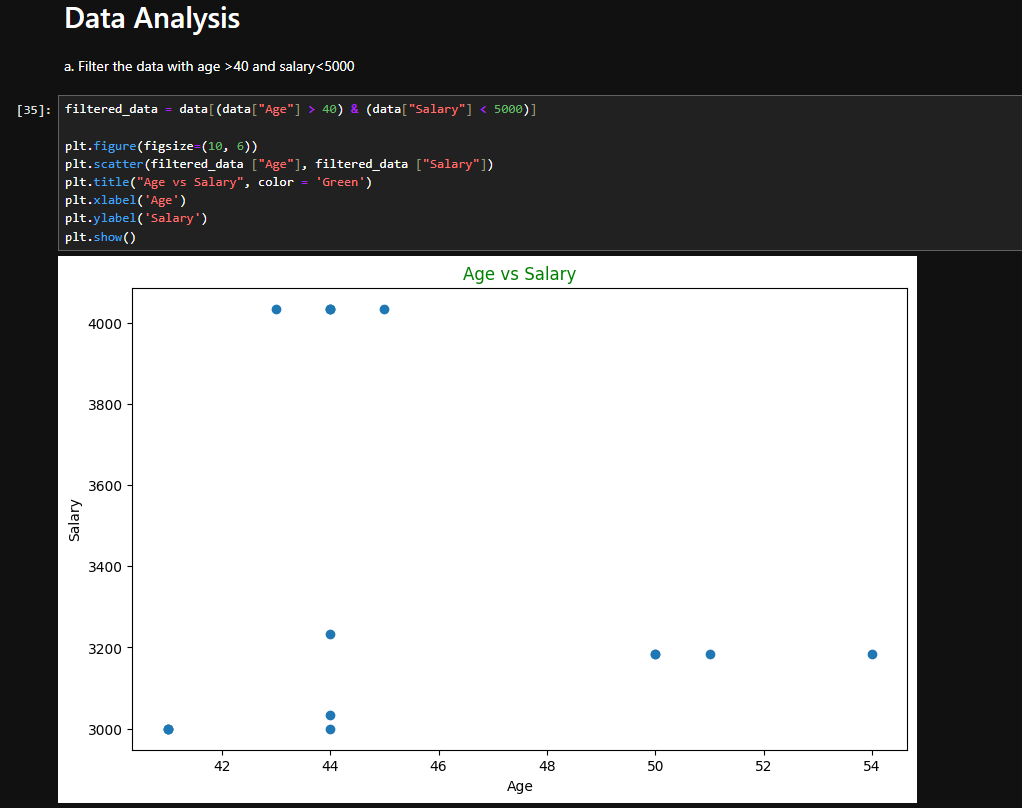


**Data Cleaning**: Find the missing and inappropriate values, treat them appropriately. Remove all duplicate rows. Find the outliers. Replace the value 0 in age as NaN Treat the null values in all columns using any measures(removing/ replace the values with mean/median/mode)

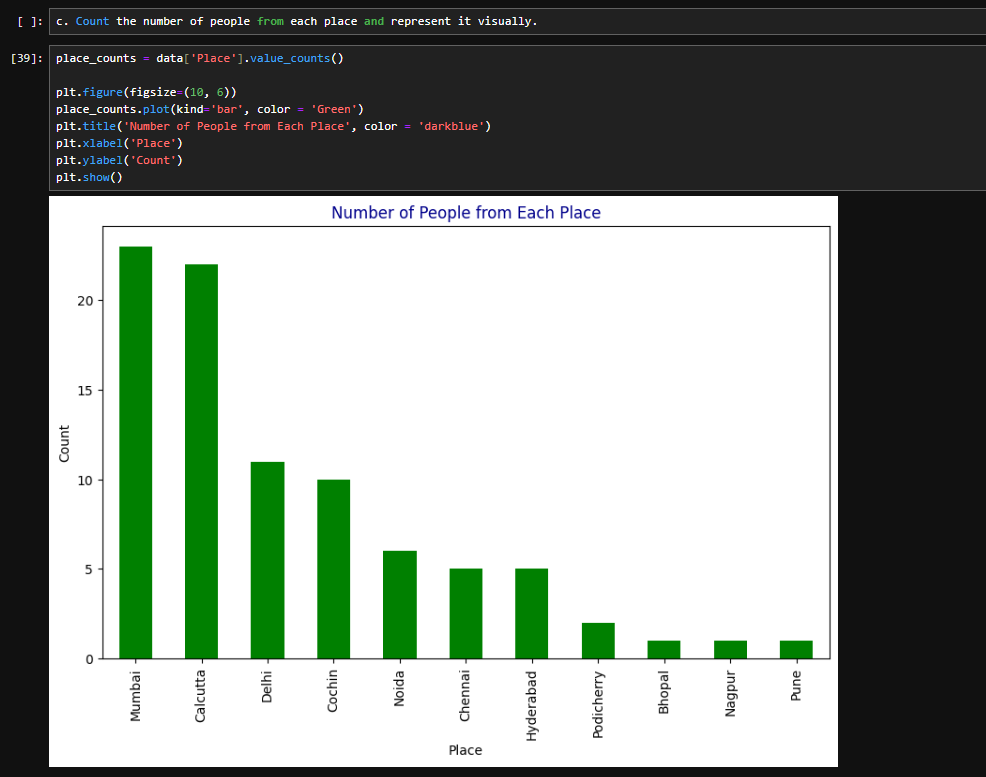


**Data Analysis:**

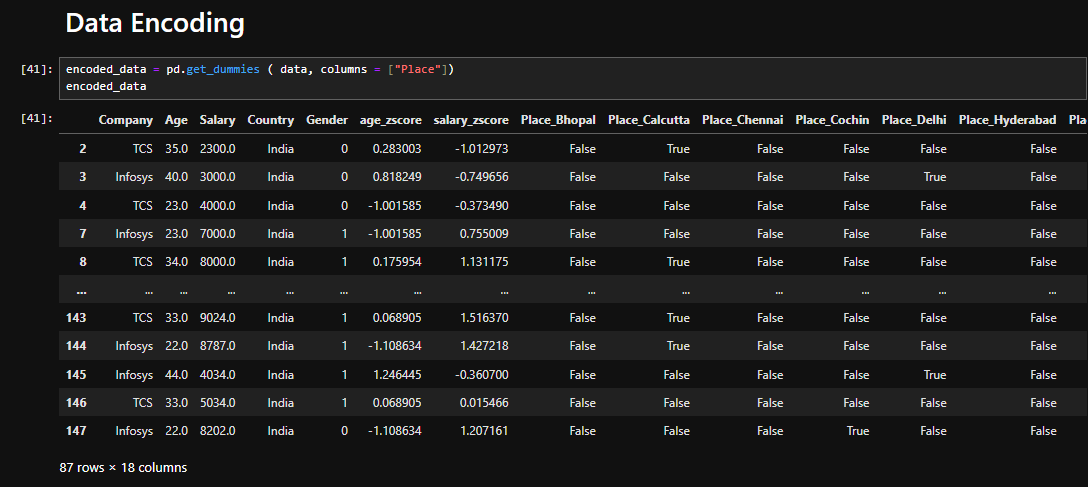
* Filter the data with age >40 and salary<5000
* Plot the chart with age and salary

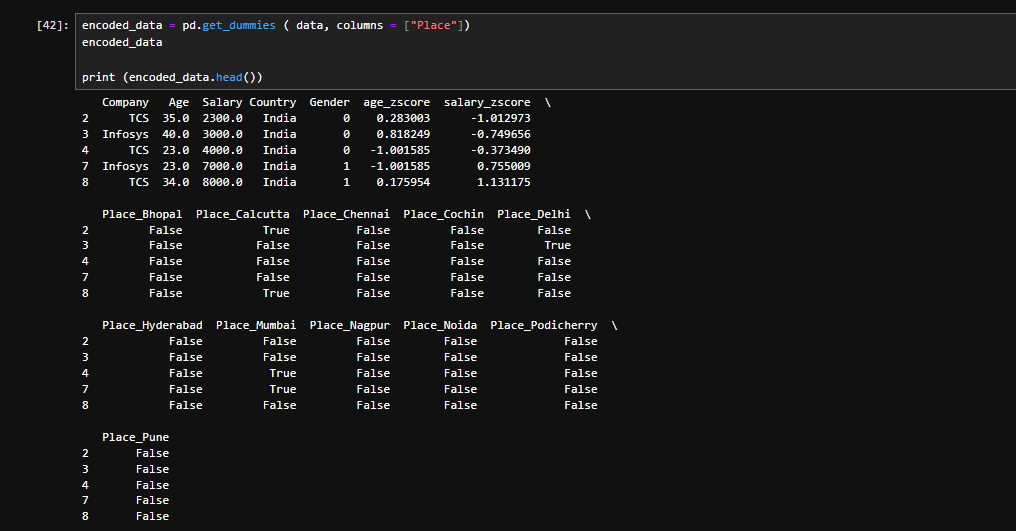


* Count the number of people from each place and represent it visually.

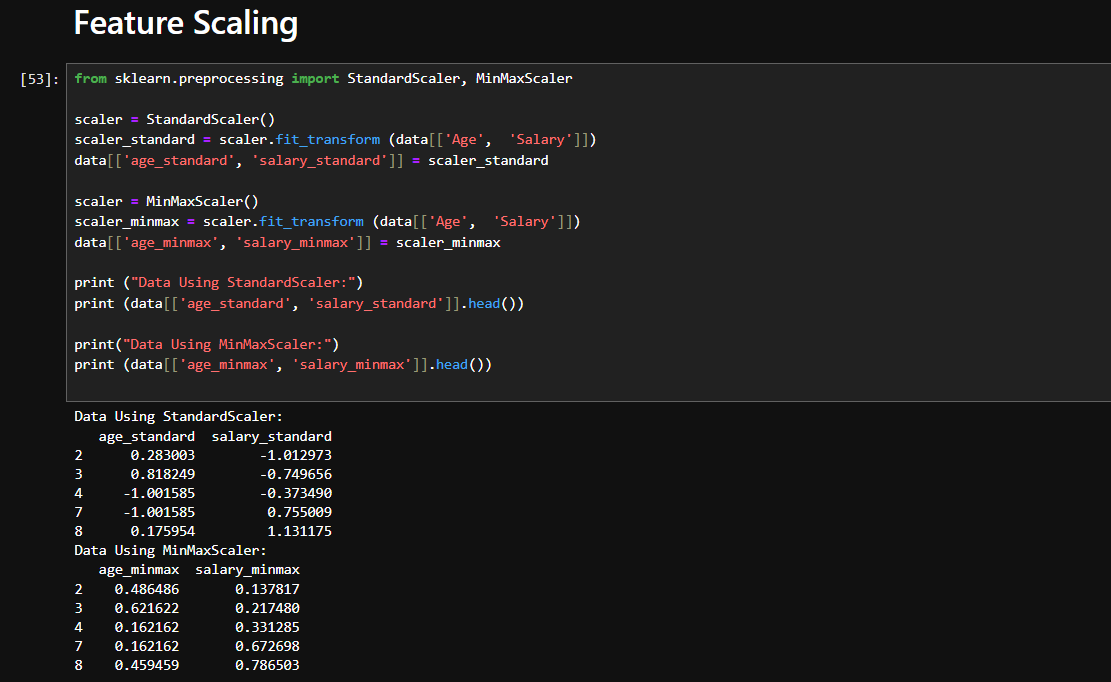


**Data Encoding:** Convert categorical variables into numerical representations using techniques such as one-hot encoding, label encoding, making them suitable for analysis by machine learning algorithms.





**Feature Scaling:** After the process of encoding, perform the scaling of the features using standardscaler and minmaxscaler.

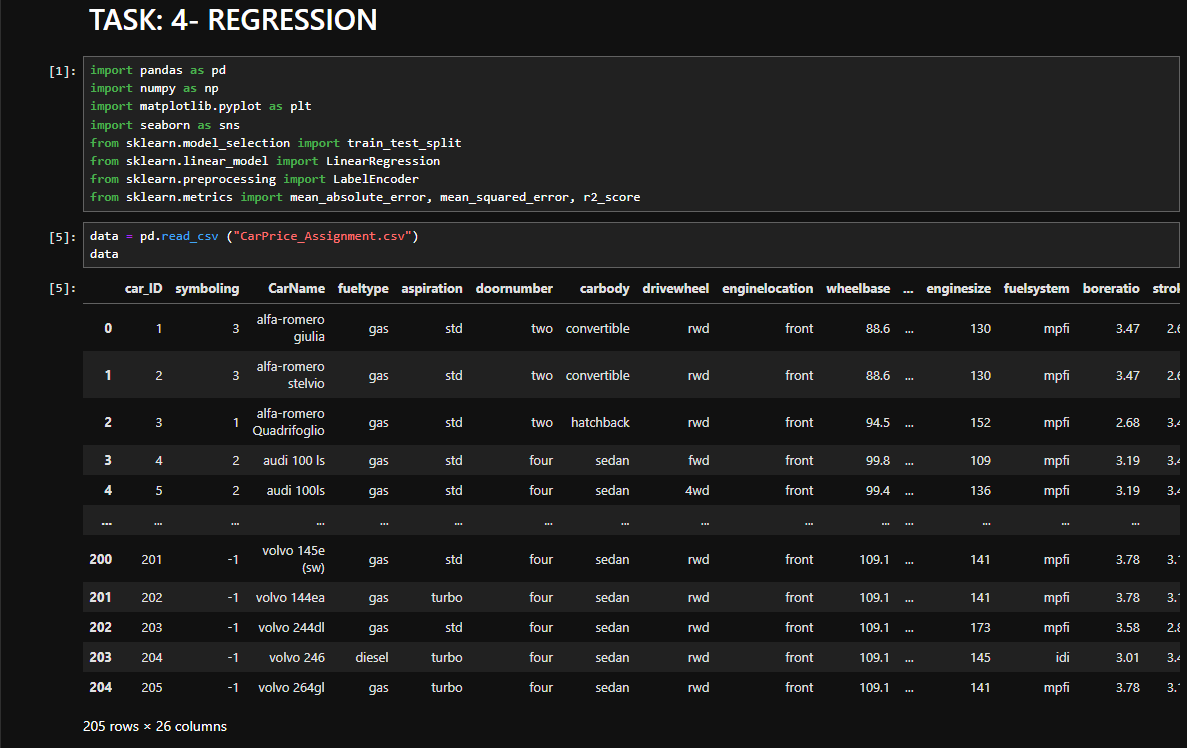


**TASK 4: - REGRESSION**

**Rootmap:**

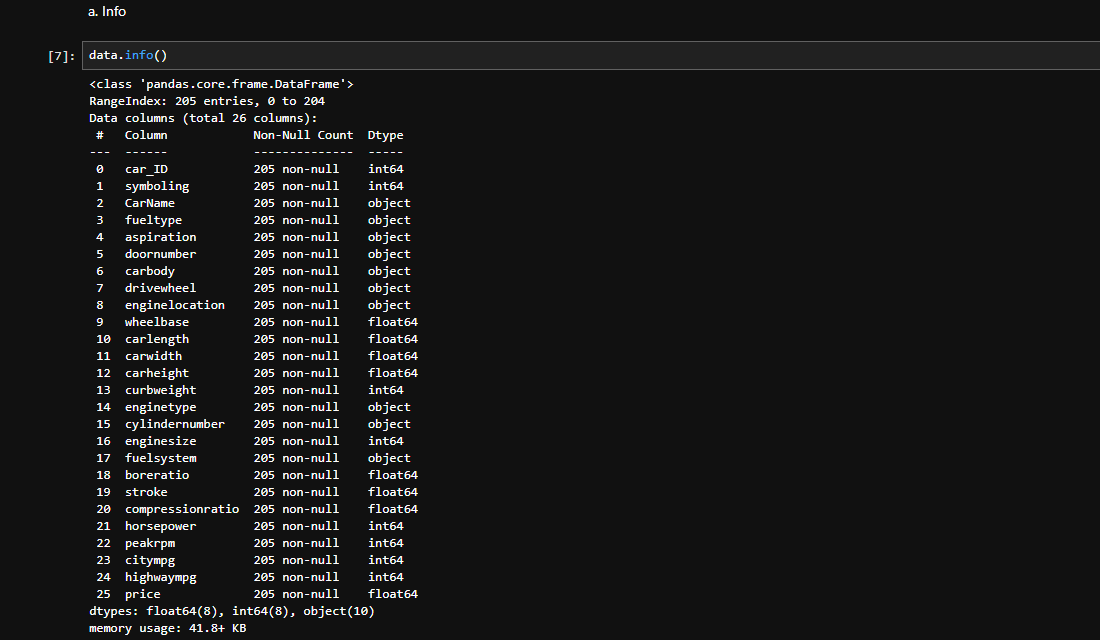
1. Understand problem statement

2. Import necessary libraries and data

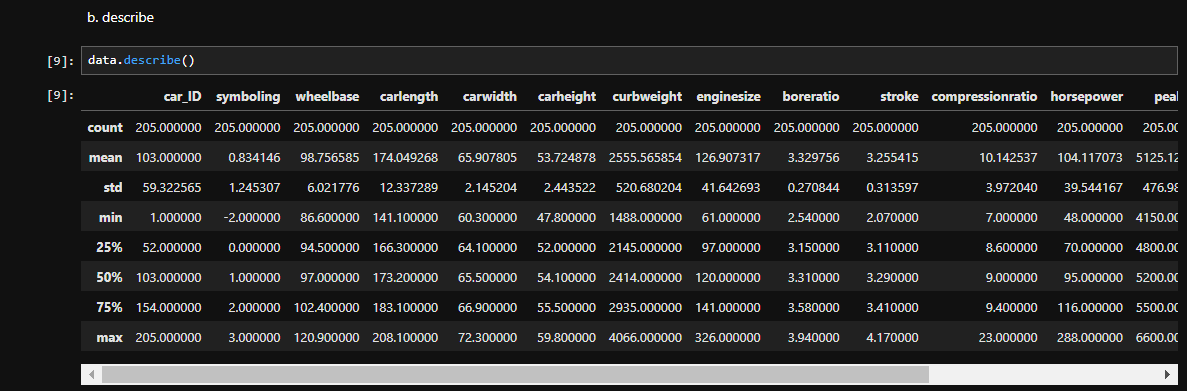


3. Check the data

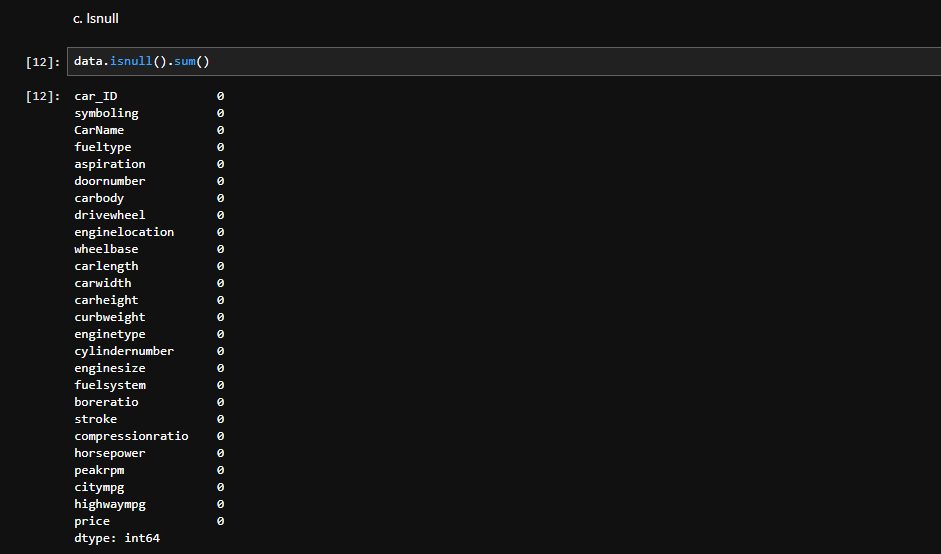
* Info()



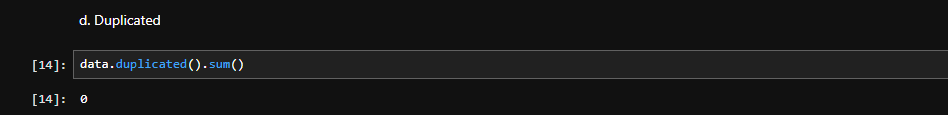
* Describe()



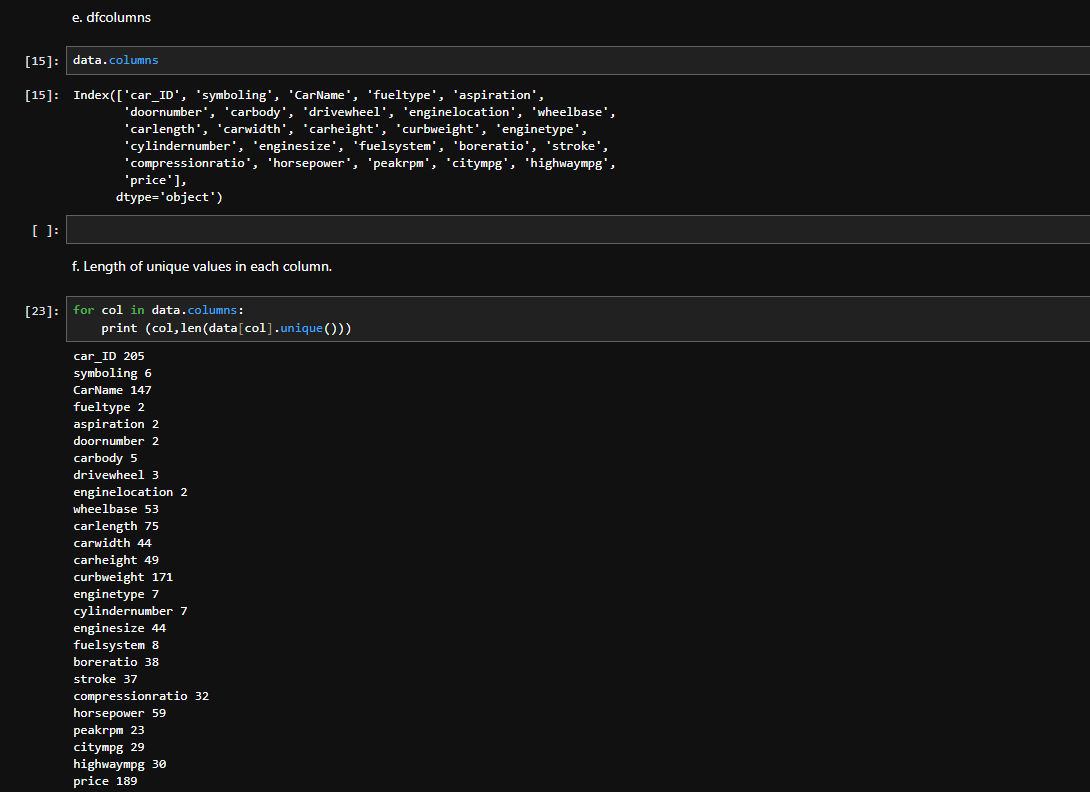
* Isnull()



* Duplicated()

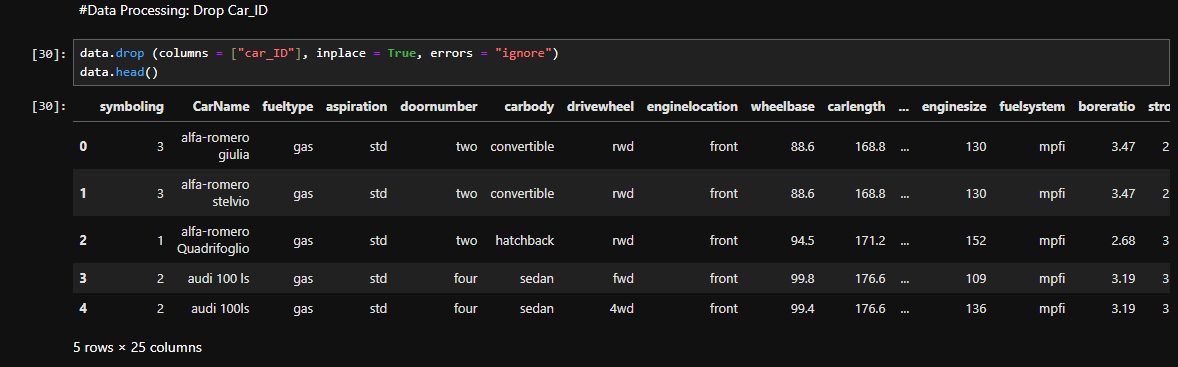


* Df. Columns
* Length of unique values in each column.

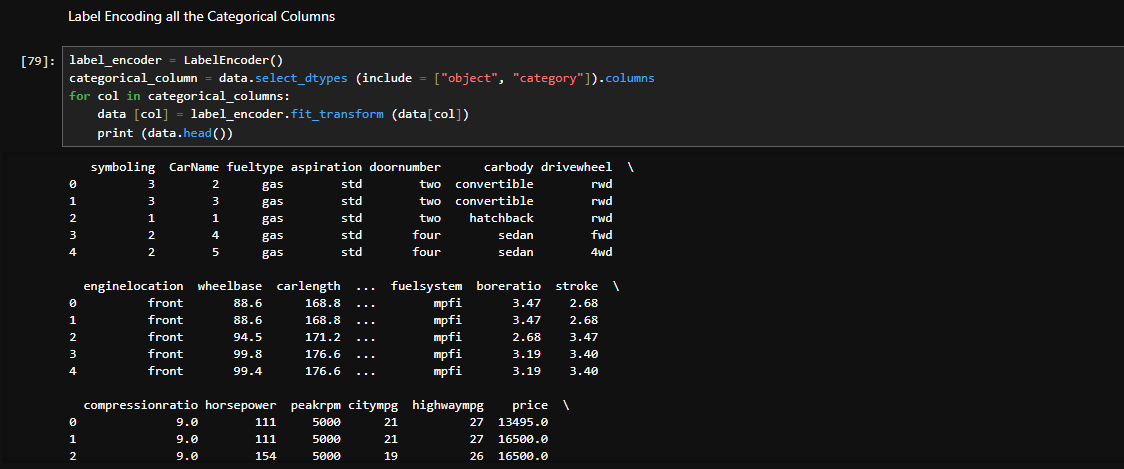
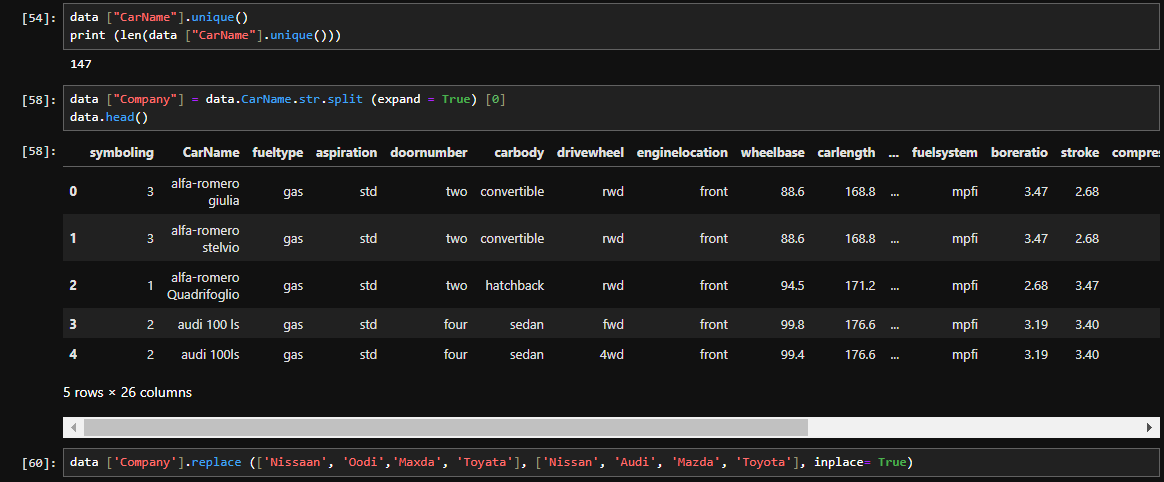
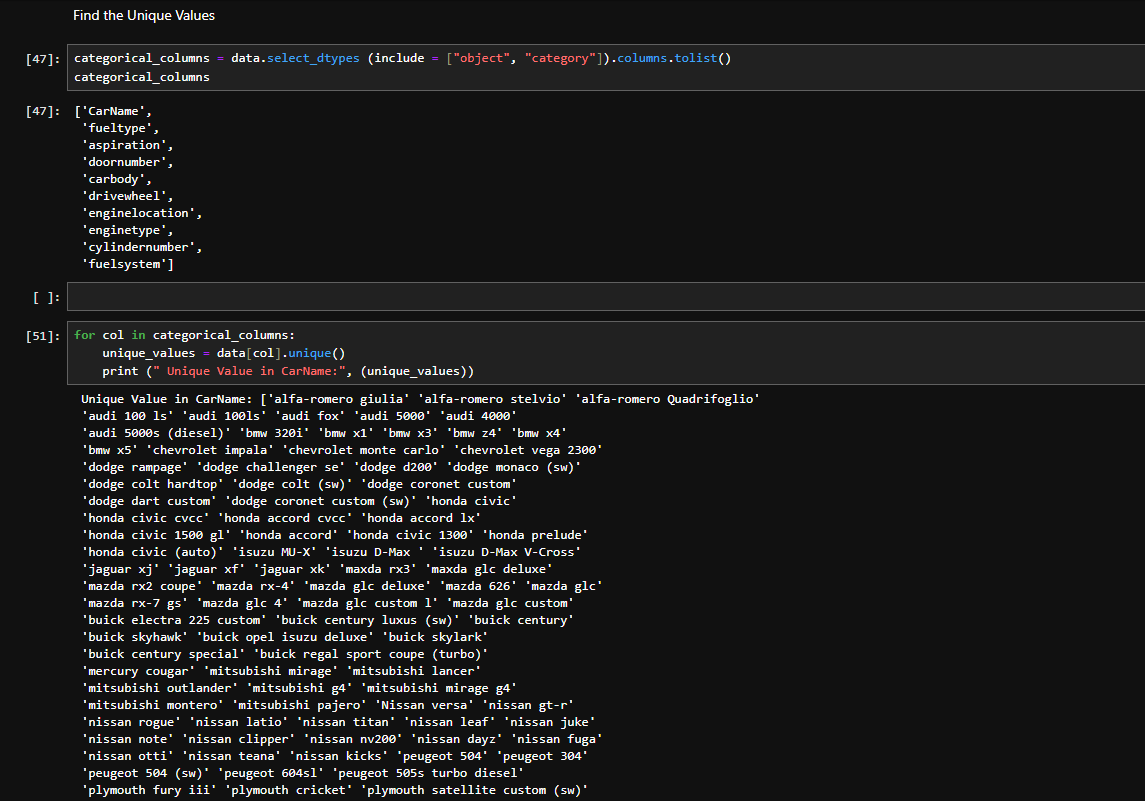


4. Data preprocessing

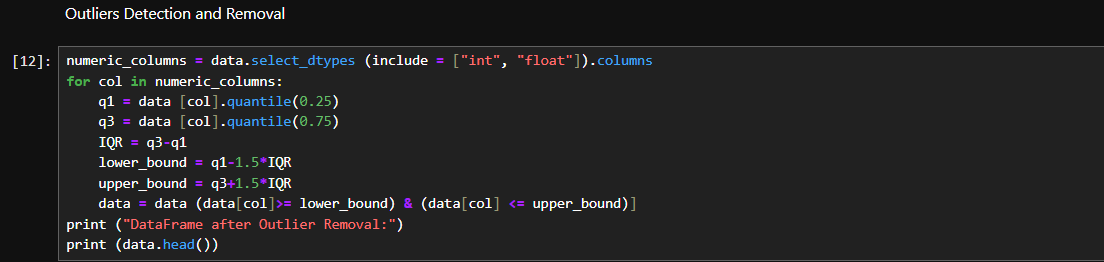
* Drop car id



* Find unique values in categorical or count plot extract company name from car name and address this new col to df also remove car name column.
* There are spelling mistakes in the company name.
* Label encoding all the categorical columns



* Outliers detection and removal( if present)



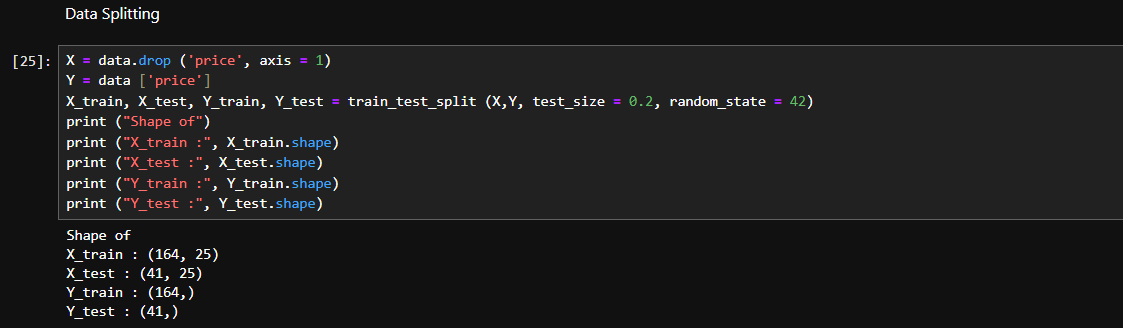
5. Feature selection

Find correlation matrix

Remove multicollinearity (remove features with High correlation .85 to 1)

6. Data splitting

Test, train



7. Model selection and implementation

8. Model evaluation

**TASK:5 - CLASSIFICATION AND CLUSTERING**

Problem Description Use sklearn.datasets iris flower dataset to train your model using logistic regression. You need to figure out the accuracy of your model and use that to predict different samples in your test dataset. In the iris dataset there are 150 samples containing the following features.

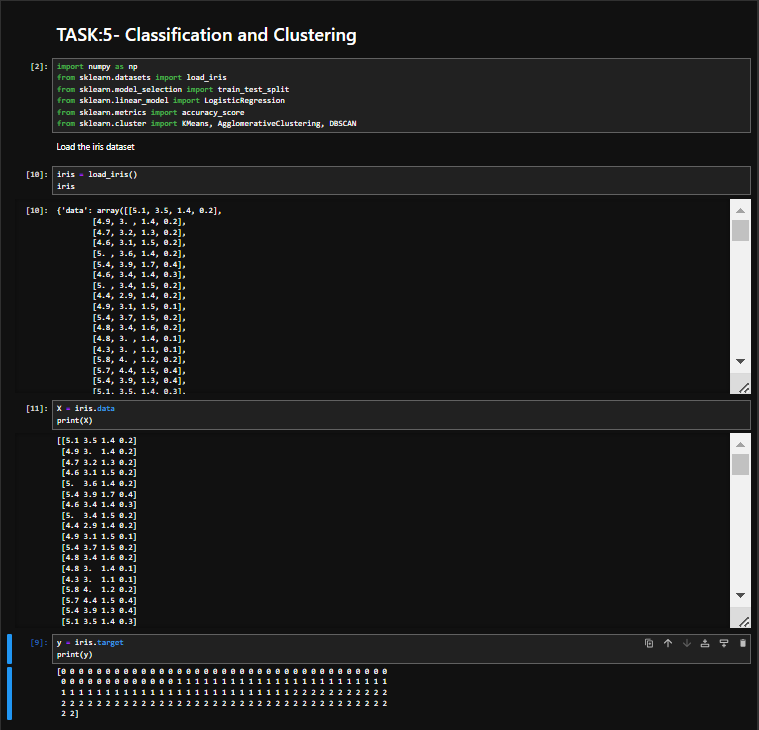
1. Sepal Length
2. Sepal Width
3. Petal length
4. Petal width

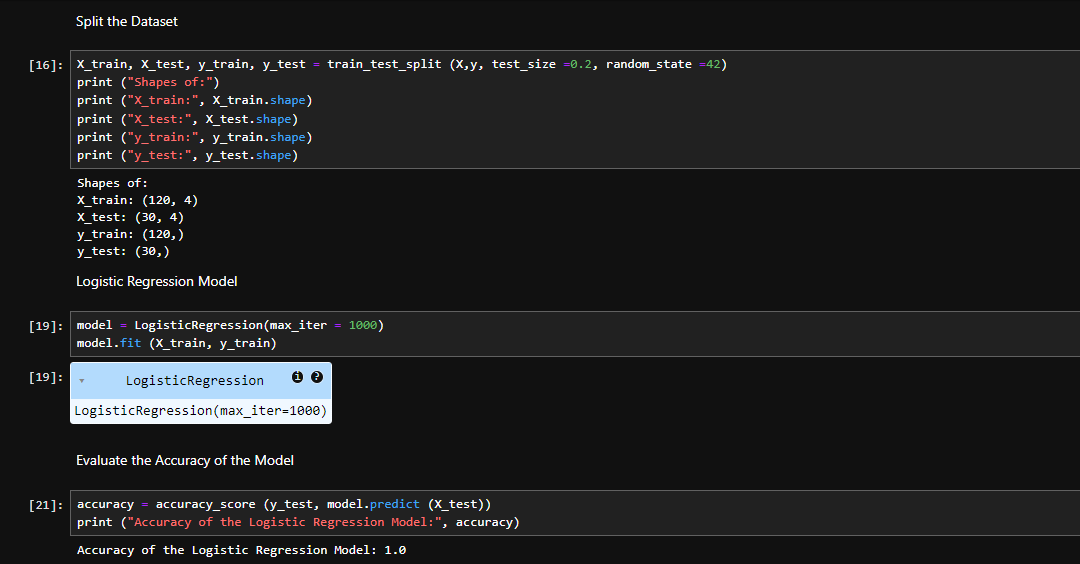
Using above 4 features you will classify a flower in one of the three categories,

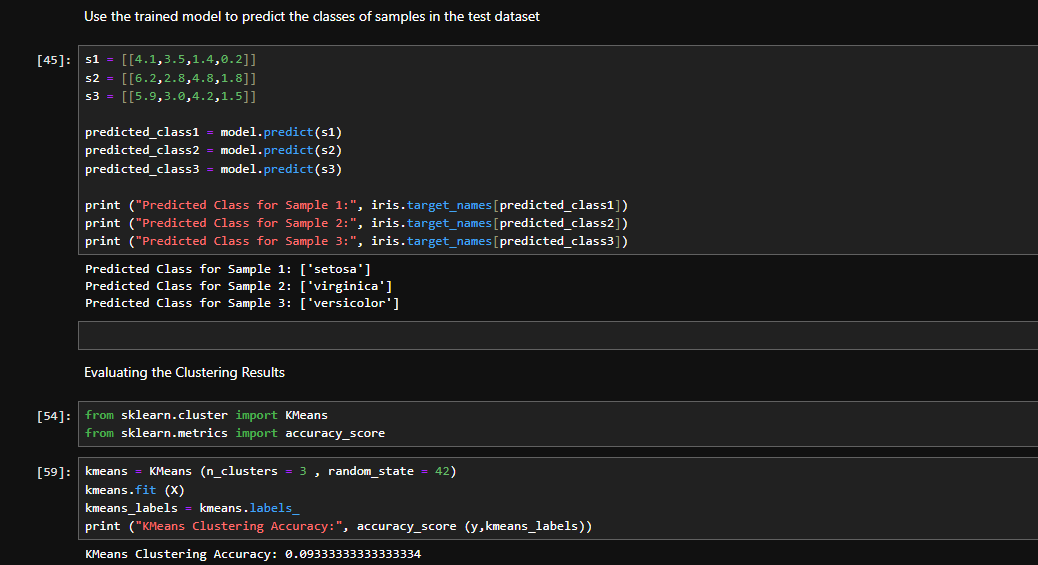
1. Setosa
2. Versicolour
3. Virginica

Download the iris dataset directly from sklearn.dataset Use the same data for different clustering(exclude target variable) and classification algorithm.

**OUTPUT:**

****

****

****