**Project 8: Customer Segmentation using Data Science**



**Phase -3: Loading And Pre-Processing Of Dataset**

KAMESHWARAN BL

KASIM SHERIFF H

KISHORE A

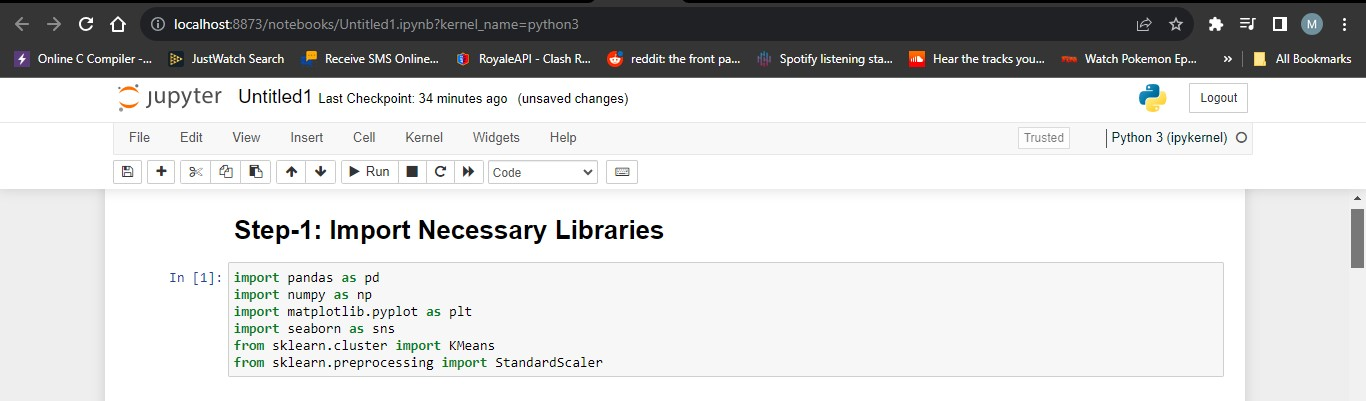
MANOJH PA

JEEVAN BABU R

**Introduction:**

In the rapidly evolving landscape of business and technology, the importance of understanding and catering to customer needs cannot be overstated. As we are college students passionate about the intersection of data science and marketing, we embark on a project aimed at leveraging data-driven insights to enhance marketing strategies. The focus is on customer segmentation, a pivotal aspect that enables businesses to tailor their approaches, fostering a more personalised and satisfying customer experience.

**1) Importing Necessary Libraries:**



**1. Numpy:** Imports the numpy library and assigns it the alias np. Numpy is used for numerical operations and working with arrays.

**2. Pandas:** Imports the pandas library and assigns it the alias pd. This library is commonly used for data manipulation and analysis.

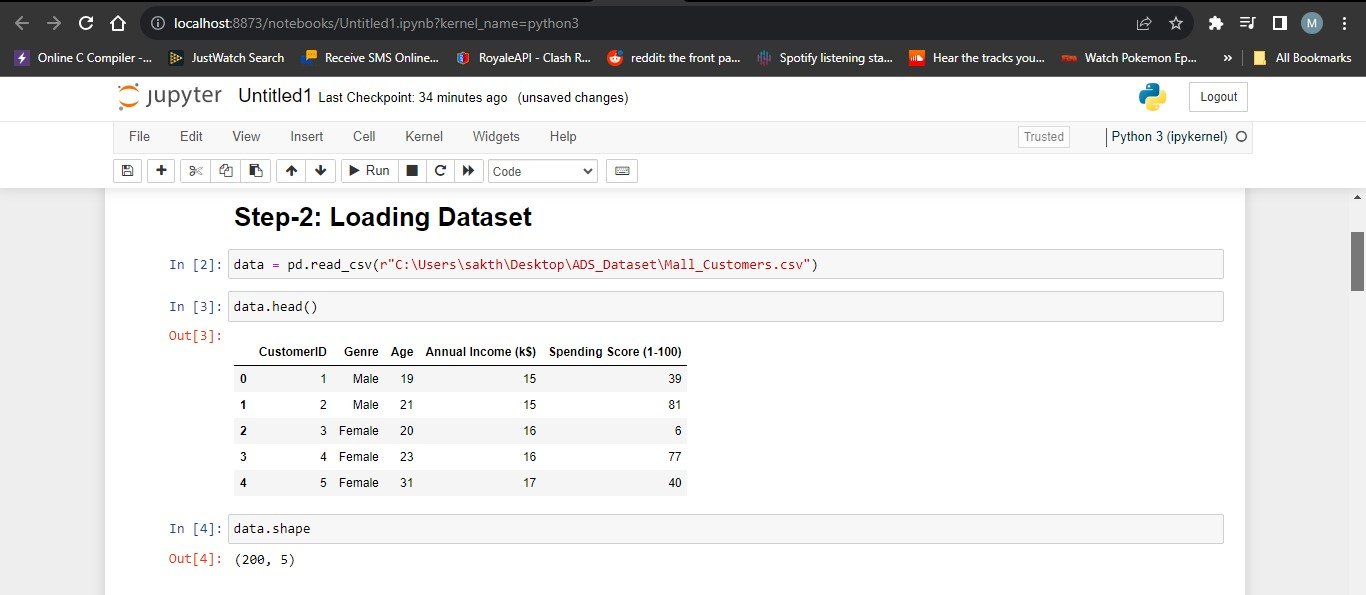
**3. Matplotlib.pyplot:** Imports the matplotlib.pyplot module and assigns it the alias plt. This library is used for creating visualizations such as plots and charts.

**4. Seaborn:** Imports the seaborn library and assigns it the alias sns. Seaborn is a data visualization library based on Matplotlib and provides a high-level interface for drawing attractive statistical graphics.

**5. Sklearn.cluster (KMeans):** Imports the KMeans class from the sklearn.cluster module. This class is used to perform K-Means clustering, a common technique for segmentation.

**6. Sklearn.preprocessing (StandardScaler):** Imports the StandardScaler class from the sklearn.preprocessing module. This class is used for standardizing features by removing the mean and scaling to unit variance.

**2) Loading Dataset:**

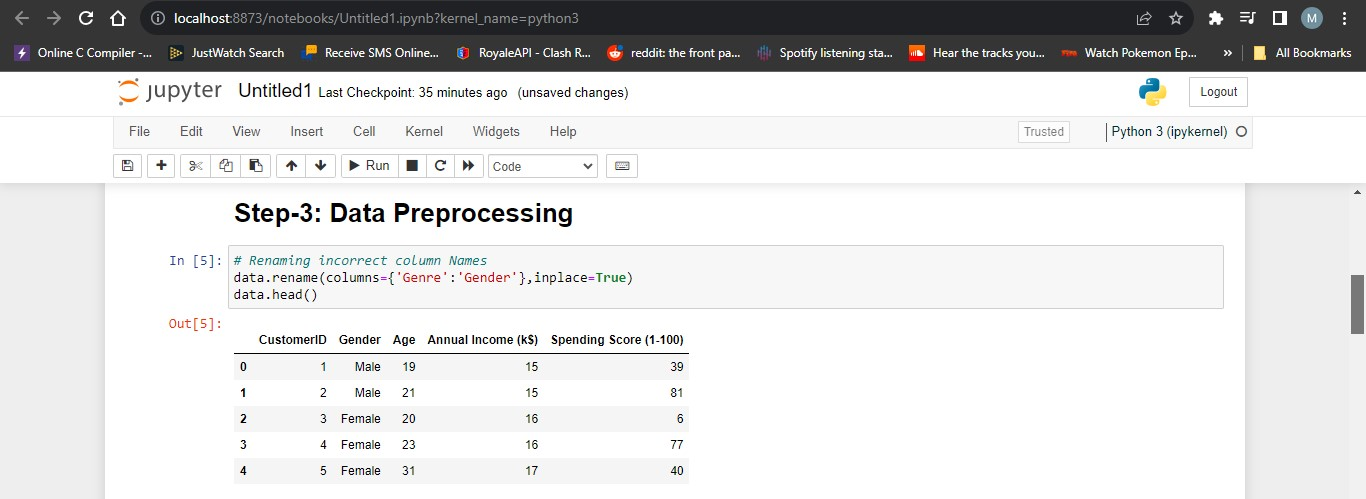


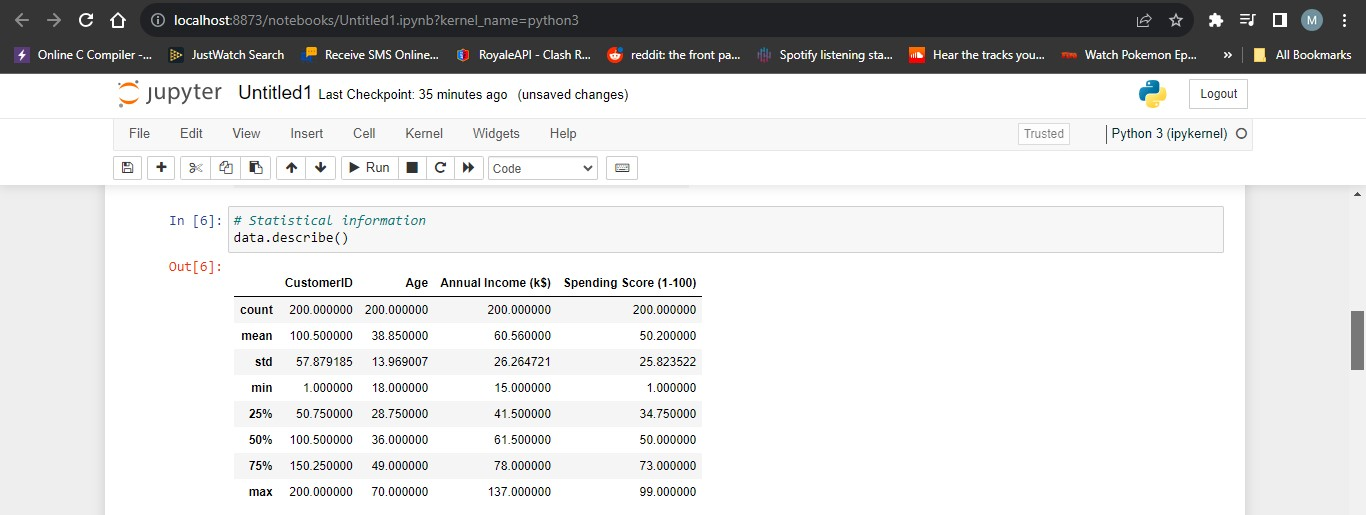
**1. pd.read\_csv ("Mall\_Customers.csv"):** Uses the read\_csv function from the pandas library to read the Mall Customers dataset from a CSV file and create a Data Frame named data.

**2. data.head ( ):** Returns the first 5 rows of the data frame.

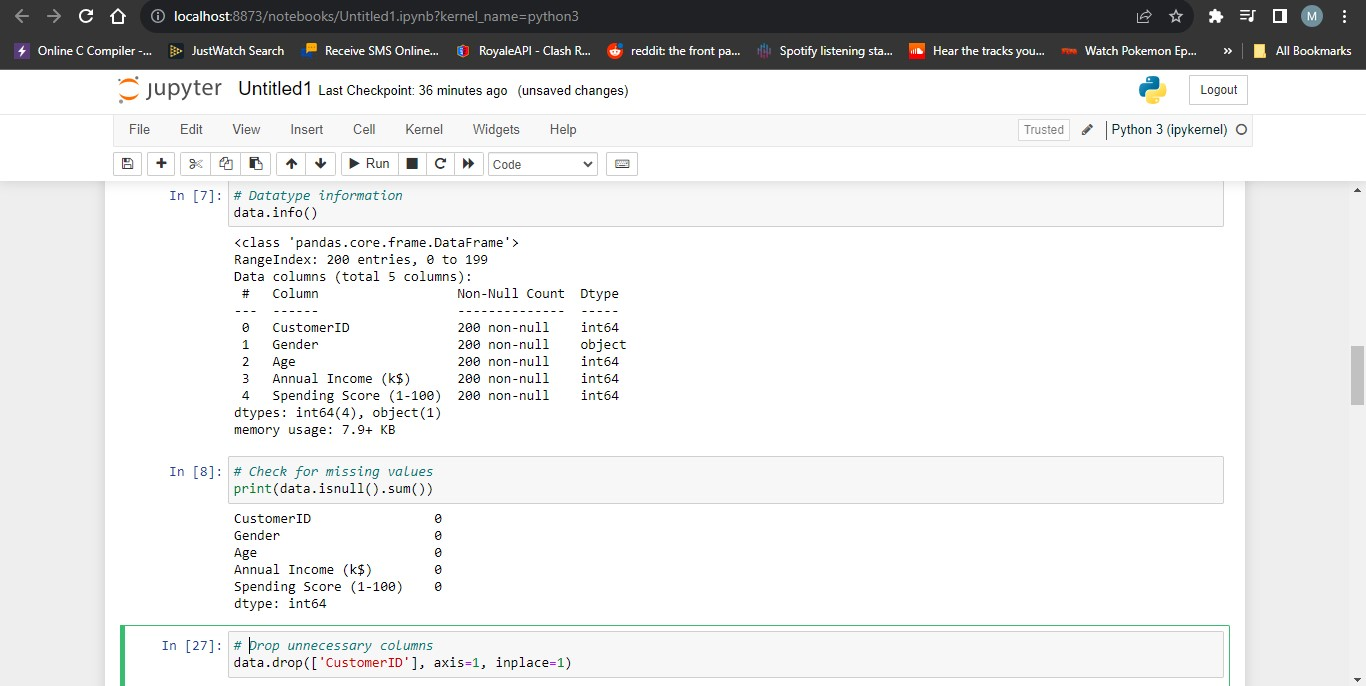
**3. data.shape ( ):** Return a tuple representing the dimensionality of the Data Frame. Tuple of array dimensions.

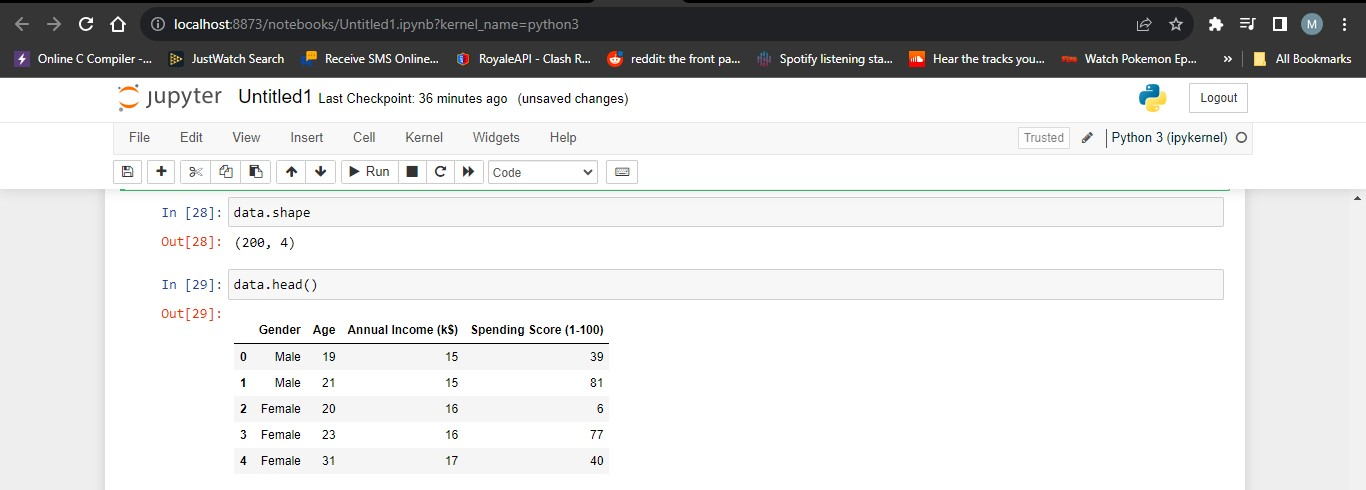
**3) Data Pre-Processing:**



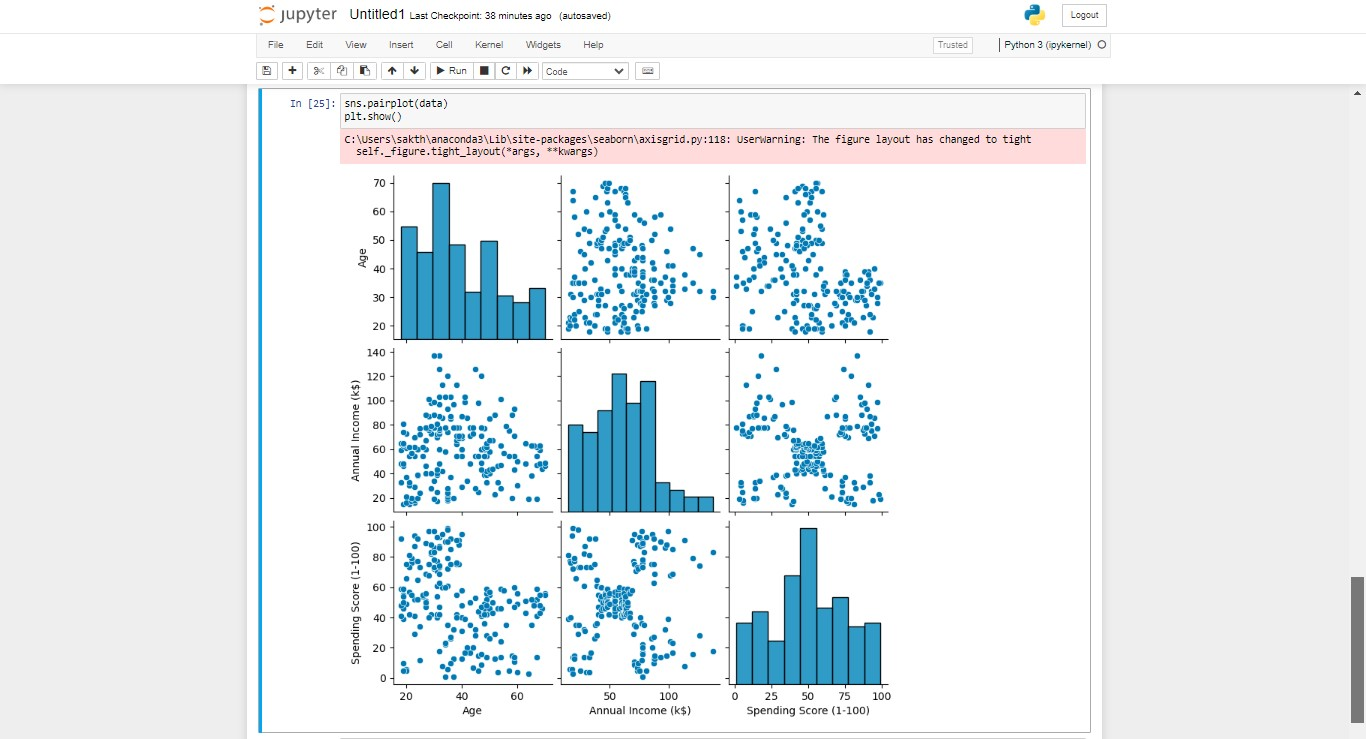


**1. Dropping Unnecessary Columns:** data = data.drop (columns= ['CustomerID']): This line removes the 'CustomerID' column from the dataset as it is not likely to contribute to the clustering process. Dropping unnecessary columns helps simplify the dataset.





**2. Checking for Missing Values:** print (data.isnull ().sum ()) this line prints the number of missing values in each column of the dataset. Checking for missing values is an essential step in data pre-processing to decide how to handle them.



**3. Visualizing Data Distribution:** sns.pairplot (data) This line creates a pair plot using Seaborn, allowing you to visually inspect the relationships between different pairs of features in the dataset.

**4. Feature Scaling:** scaler = StandardScaler (); data\_scaled = scaler.fit\_transform (data) this block of code standardizes the features using the StandardScaler. Standardization is crucial for clustering algorithms like K-Means, as it ensures that all features contribute equally to the distance computations.

After completing these steps, the dataset is ready for further analysis, such as determining the optimal number of clusters and applying the K-Means algorithm for customer segmentation.