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



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

T. Mohith Shakthi (411621104033)

Mohanrangan (411621104032)

Shaik Sharukh Ahmed (411621104048)

Krishnaraj (411621104027)

K. Salman Khan (411621104047)

**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.

**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.

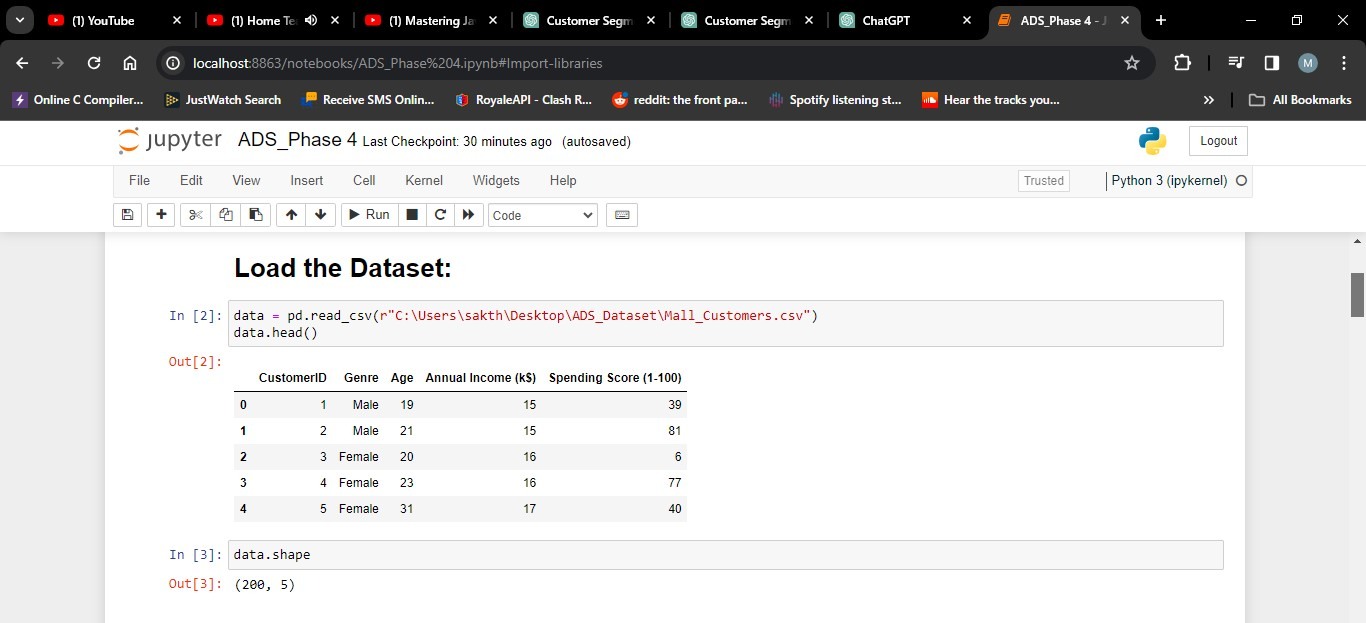
**3. Matplotlib.pyplot:** Imports the matplotlib.pyplot module and assigns it the alias plt. This library is used for creating visualizations such as plots and charts.

**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.

**Loading Dataset:**

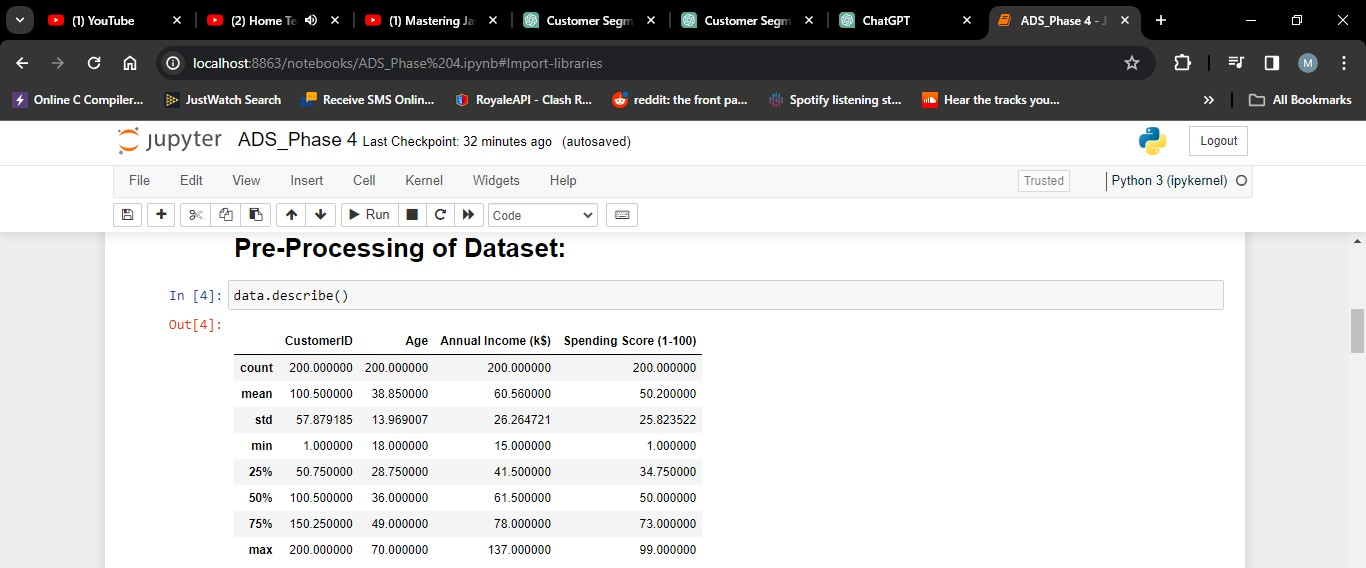


**1. pd.read\_csv ("Mall\_Customers.csv"):** Uses the read\_csv function from the pandas library to read the Mall Customers dataset from a CSV file and create a Data Frame named data.

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

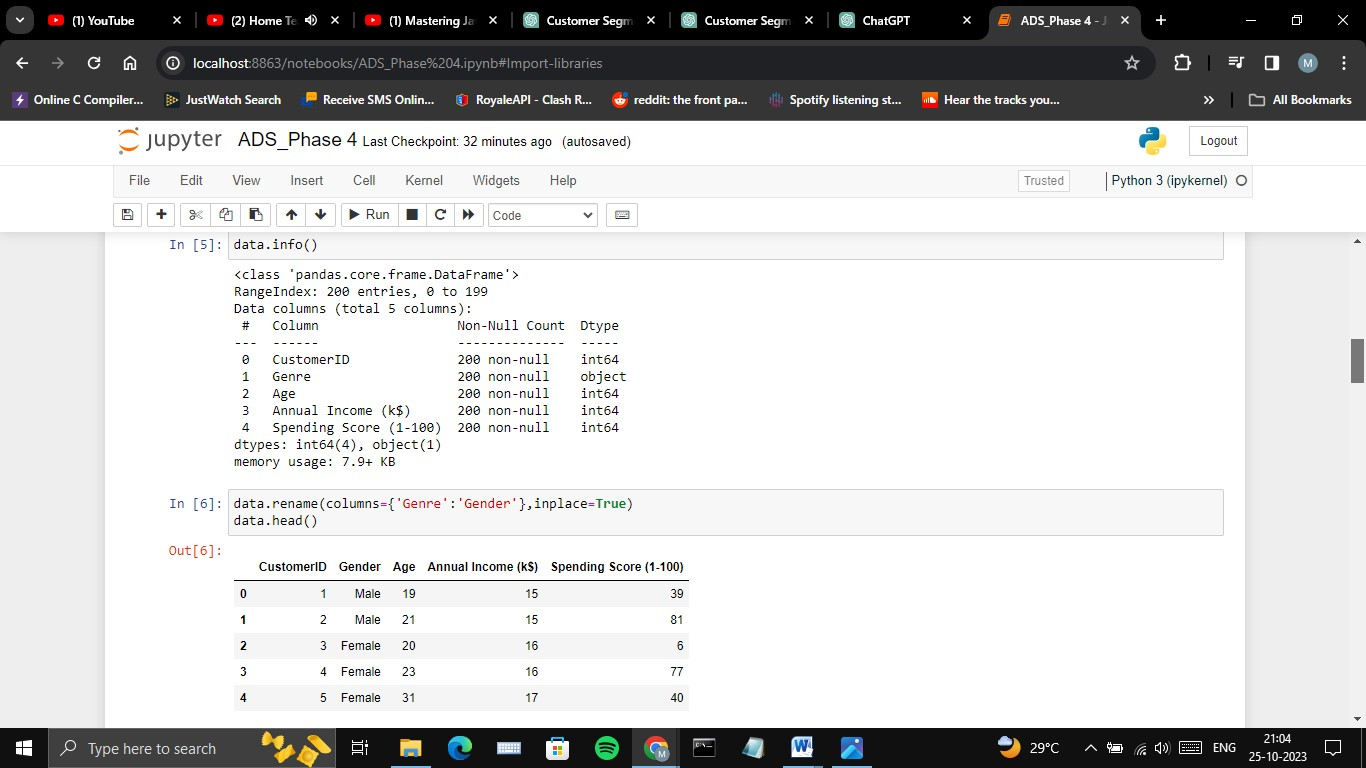
**3. Data.shape ( ):** Return a tuple representing the dimensionality of the Data Frame. Tuple of array dimensions.

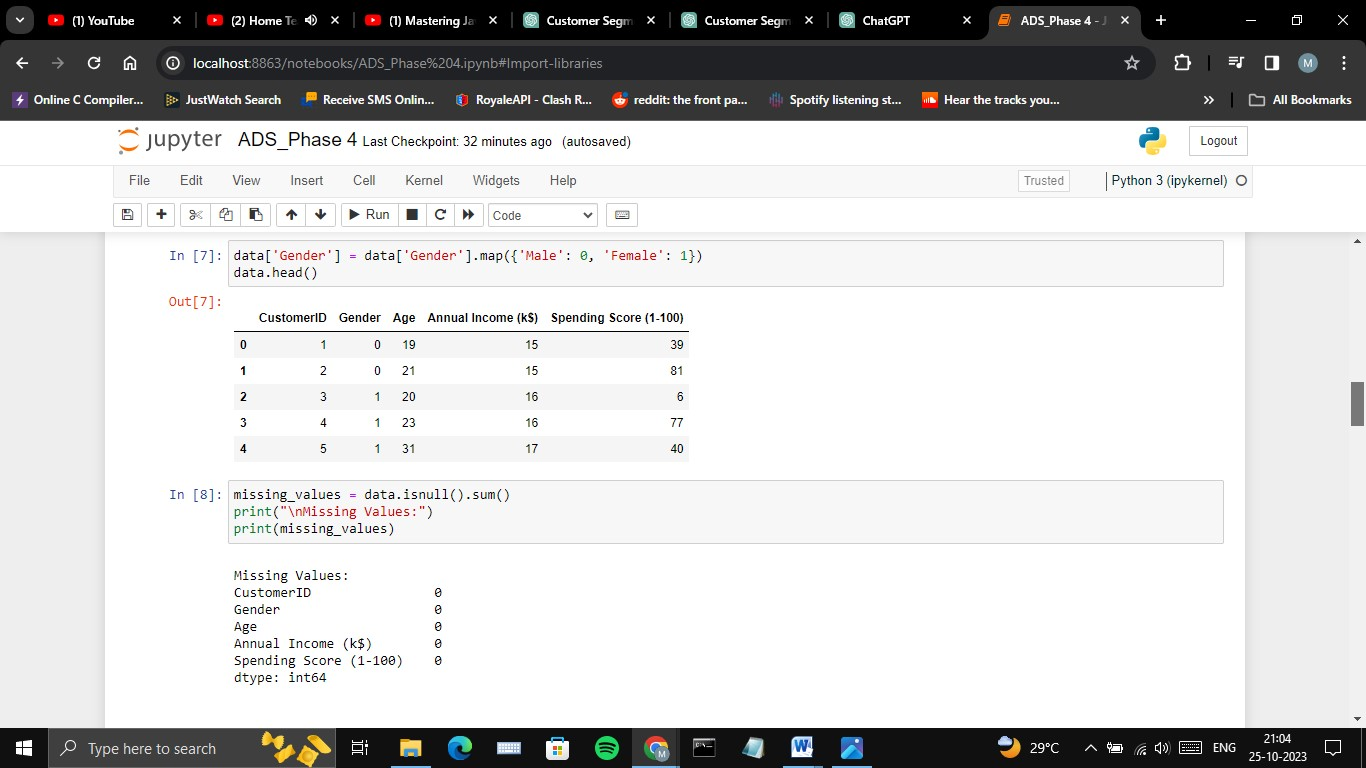
**Data Pre-Processing:**



Data pre-processing is a crucial step in building a machine learning model. In this step, you handle missing data, select relevant features, and split the data into training and testing sets

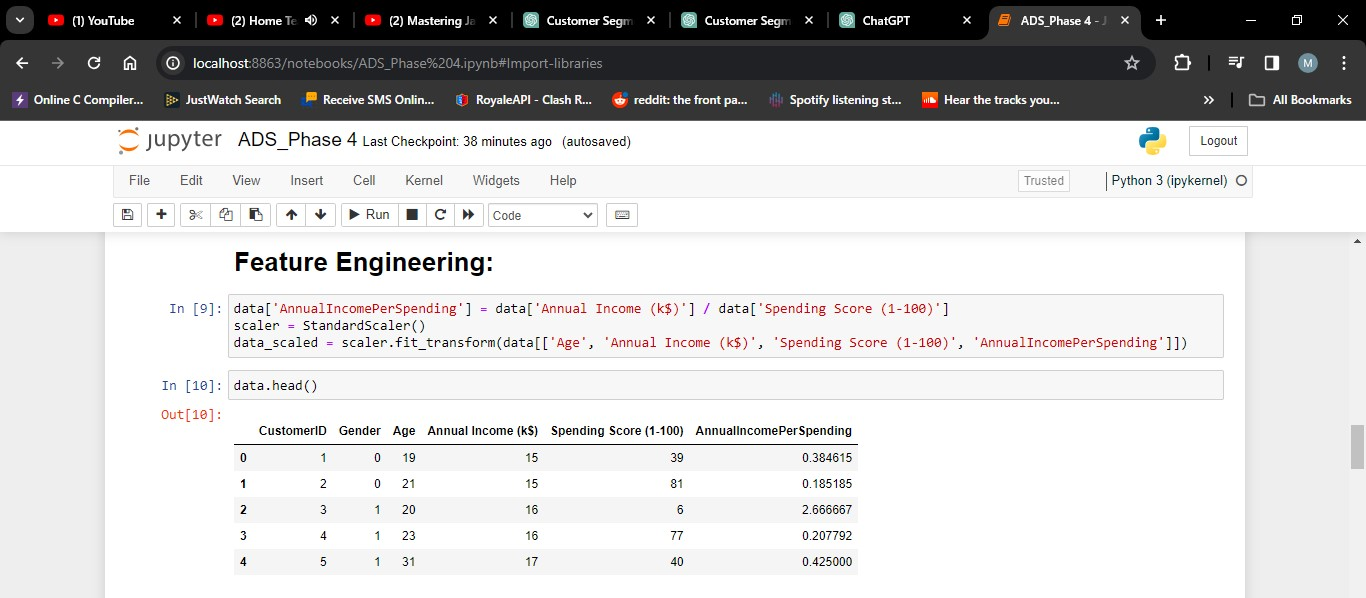
.In this code, the 'Product ID' column is dropped as it's considered irrelevant. If there are missing values, you can choose to drop rows or fill in missing values based on your dataset and domain knowledge.





Now, you have a pre-processed dataset ready for building and evaluating your machine learning model. The next steps involve selecting a model, training it, making predictions, and evaluating the model's performance, as shown in the previous response.

**Feature Engineering:**



Feature engineering is the process of creating, selecting, or transforming input features (variables) in a way that enhances a machine learning model's performance.

In summary, feature engineering is the art of preparing and transforming the dataset to provide the most informative and suitable input to your machine learning model. It can significantly impact the model's predictive performance and is a critical step in the model development process.

**Applying clustering algorithms:**

K-Means is a clustering algorithm used in the model to group customers. It works as follows:

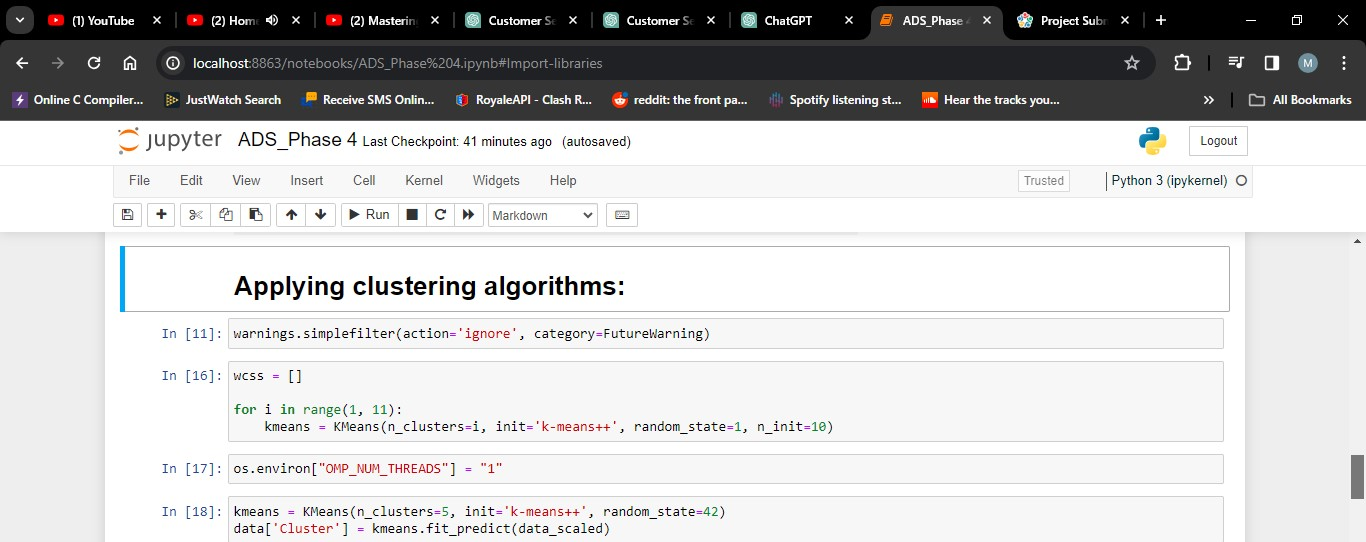
**1. Initialization:** Start by selecting K initial cluster centers randomly or strategically.

**2. Assigning Data Points:** Assign each data point (customer) to the nearest cluster center based on a distance metric, typically Euclidean distance.

**3. Updating Cluster Centres:** Recalculate the cluster center as the mean of the data points assigned to each cluster.

**4. Repeat:** Iteratively repeat steps 2 and 3 until convergence (when cluster assignments no longer change significantly).

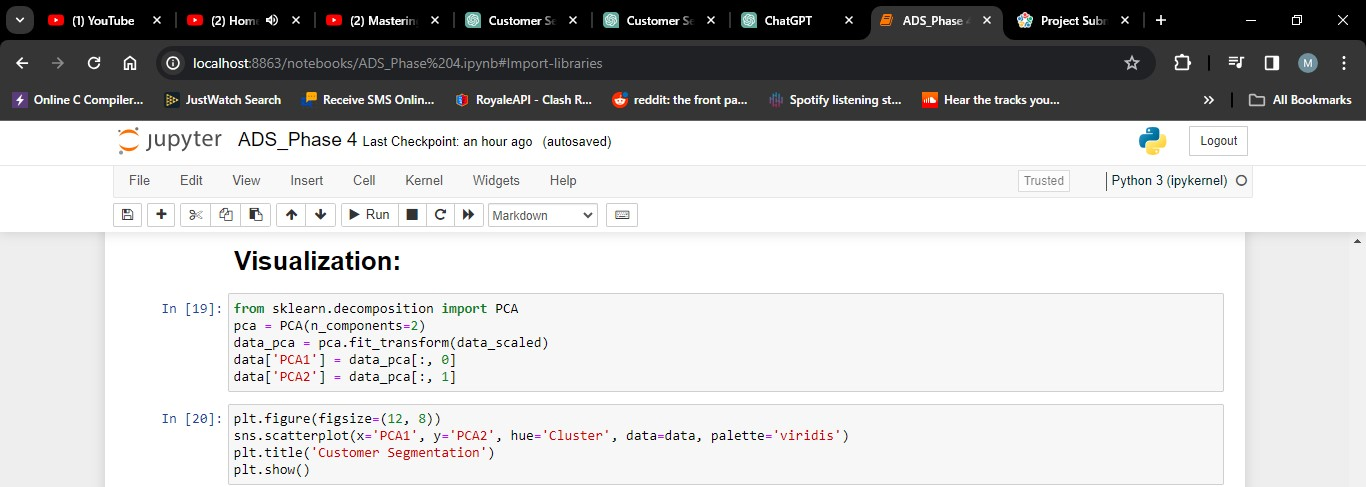
**Result:** The final cluster centers represent the cluster's characteristics, and the data points are grouped into K distinct clusters based on their similarity to these centers.

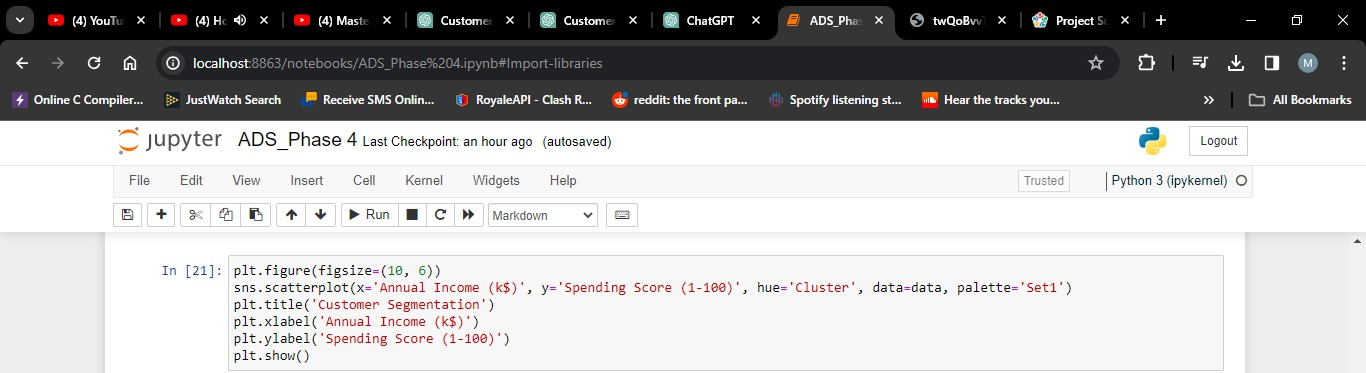


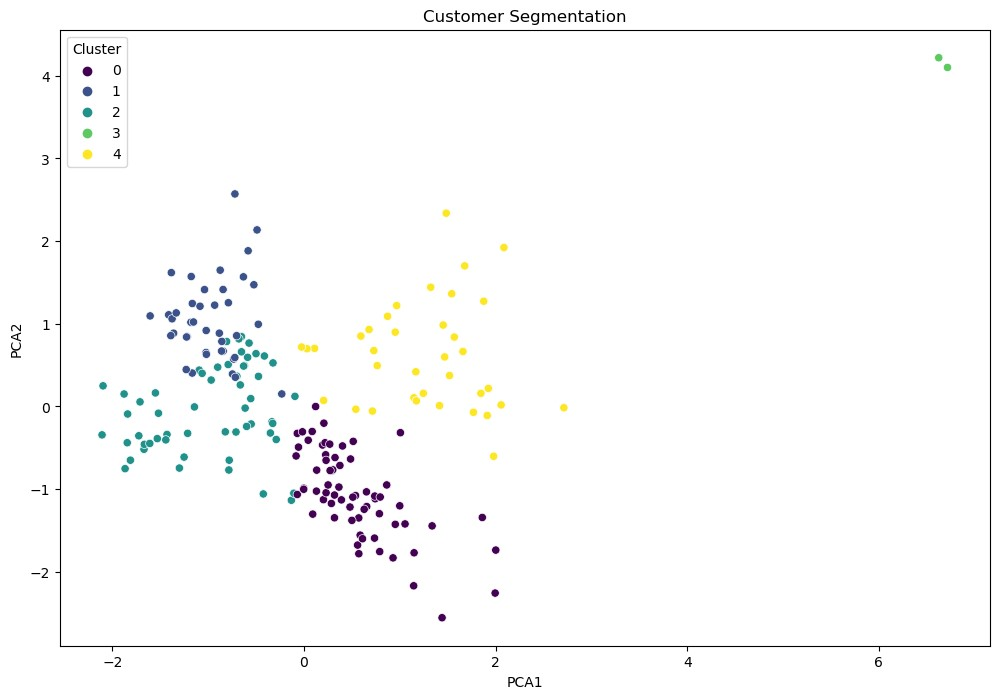
**Visualization:**

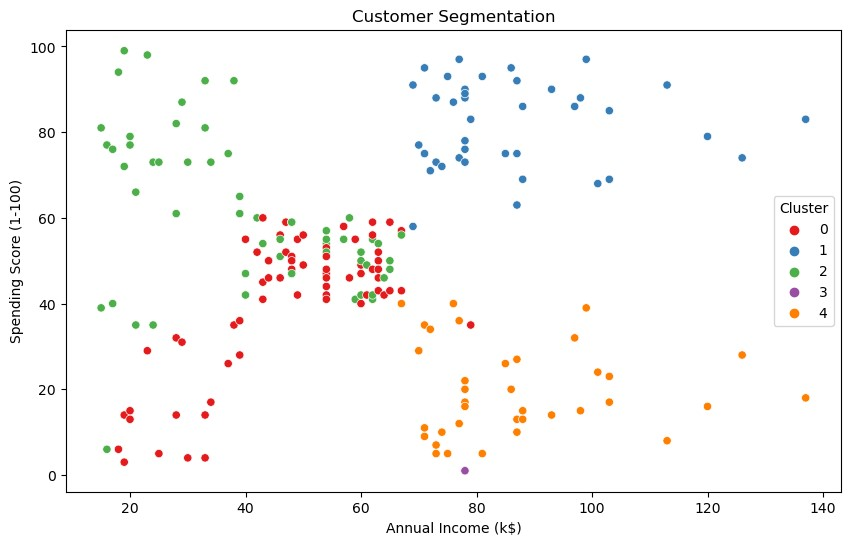
Visualizing the Clusters: We'll visualize the clusters using scatter plots to see how customers are grouped based on their spending behaviour and annual income.

This code will create a scatter plot where each point represents a customer. The points will be coloured according to their cluster membership. This visualization will help you identify distinct customer groups.









**Interpretation:**

After visualizing the clusters, you should interpret and understand the characteristics of each cluster. You can do this by analysing the cluster means and profiles. The cluster means represent the average values of each feature within each cluster.

**1. Cluster 0:** Customers with moderate income and moderate spending.

**2. Cluster 1:** High-income customers with high spending.

**3. Cluster 2:** Low-income customers with low spending.

**4. Cluster 3:** High-income customers with low spending.

**5. Cluster 4:** Customers with low income and high spending.

You can further analyse and interpret these clusters to create customer personas and tailor marketing strategies accordingly. For example, Cluster 1 represents potential high-value customers, while Cluster 2 represents customers who may need targeted promotions to increase their spending.

This interpretation will help you make informed decisions to better serve and market to each customer segment.

