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



**Phase -5:  Project Documentation & Submission**

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.

**Problem Definition:**

In a dynamic landscape of business, understanding customers is vital for effective marketing strategies. Our project aims to employ data science techniques for customer segmentation, focusing on behaviour, preferences, and demographic attributes. Traditional one-size-fits-all approaches often fail to resonate with individual customers' unique characteristics, prompting the need for a data-driven strategy. The objective of our project is to overcome the limitations of generic marketing strategies by dissecting the diverse customer base into distinct groups. Our goal is to identify patterns within customer data and formulate personalised marketing strategies to enhance overall customer satisfaction.

**Design Thinking:**

The design thinking process outlines a structured approach to solving the problem of customer segmentation for personalized marketing. Each step is iterative, allowing for continuous learning and improvement. By understanding the needs of customer, defining data requirements, ideating creative solutions, and prototyping and testing various components, using this process we aim to deliver a robust and effective customer segmentation solution.

**Project Approach:**

**Data Collection:**

We have collected relevant data on customer behaviour, preferences, and demographic attributes from various sources, such as customer databases, online interactions, surveys, and third-party data providers.

This may include:

**Behavioural Data:** Purchase history, website interactions, and responses to marketing campaigns.

**Preference Data:** Surveys, feedback, reviews.

**Demographic Data:** Age, gender, location, income.

We have Ensured data privacy and compliance with regulations like GDPR. We anonymize and aggregate data to protect customer privacy.

**Data Pre-processing:**

We clean and pre-process the data collected to ensure its quality and usability. Steps may include:

**Handling missing data:** Impute or remove missing values.

**Removing duplicates:** Ensureeach customer is represented once.

**Data transformation:** Normalise numerical features, encode categorical variables.

**Feature Engineering:**

We can create new features or transform existing ones to extract more meaningful information. For example, you might create features like:

**RFM (recency, Frequency, and Monetary) features:** Calculate these metrics based on customer transaction history.

**Demographic features:** Create categories or bins for age, income, or other demographic attributes.

**Preference features:** Incorporate sentiment analysis scores from customer reviews or feedback.

**Clustering Algorithms:**

We must conduct EDA (Exploratory Data Analysis) to understand the data distribution, identify outliers, and visualise relationships between variables. Use summary statistics, histograms, scatter plots, and correlation matrices.

We are applying clustering algorithms to group or segment customers based on their attributes. Common clustering methods include:

**K-Means Clustering:** Divide customers into K clusters based on similarities.

**Hierarchical Clustering:** Build a hierarchical tree of clusters.

**DBSCAN:** Identify dense regions of data points as clusters.

**Model Evaluation:**

Evaluate the performance of the clustering models using appropriate metrics. Adjust the number of clusters (K) if needed.

**Visualisation:**

We are visualising the clusters to provide a clear representation of customer segments. Examples are:

**Scatter Plots:** Visualise clusters using scatter plots to understand the distribution of customers.

**Heat-maps:** Use heat-maps to showcase the relationships between different features.

**Dimensionality Reduction:** Apply techniques like PCA or t-SNE for a concise representation of clusters.

**Interpretation of Results:**

We analyse the characteristics of each cluster to understand what distinguishes one segment from another. We identify common behaviours, preferences, and demographic profiles within each cluster. We identify actionable insights for personalised marketing strategies.

**Project/Model Building:**

**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.

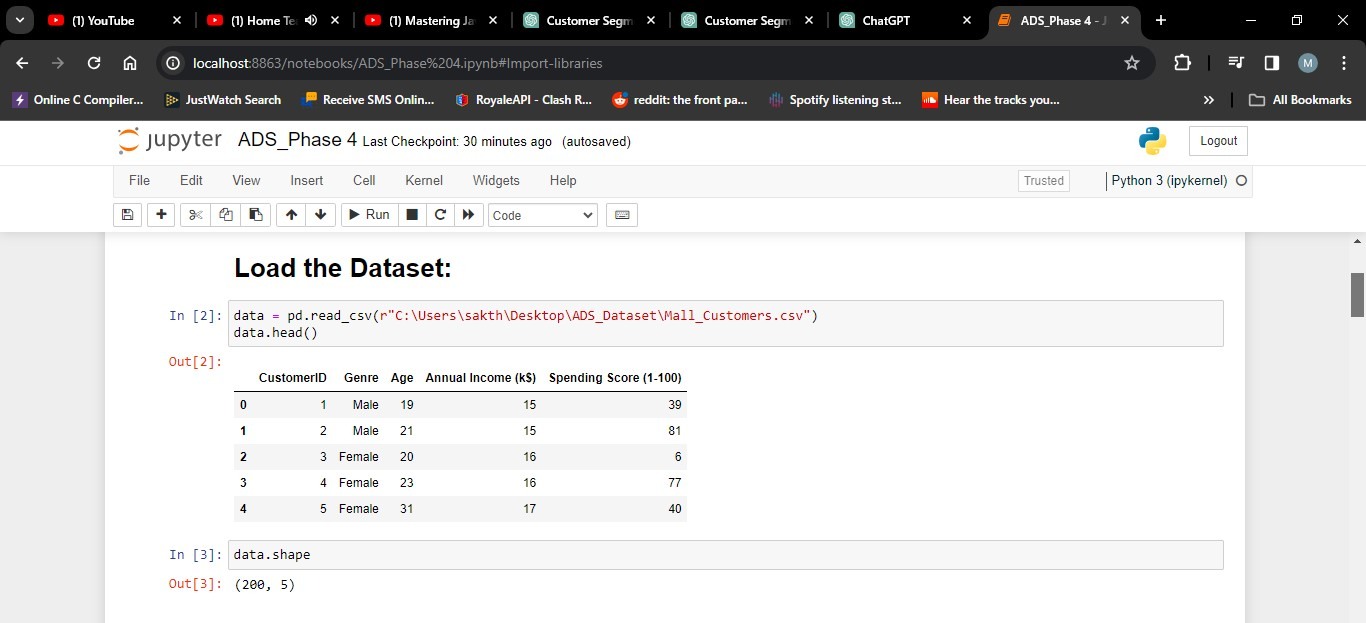
**3. Matplotlib.pyplot:** Imports the matplotlib.pyplot module and assigns it the alias plt. This library is used for creating visualizations such as plots and charts.

**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:**

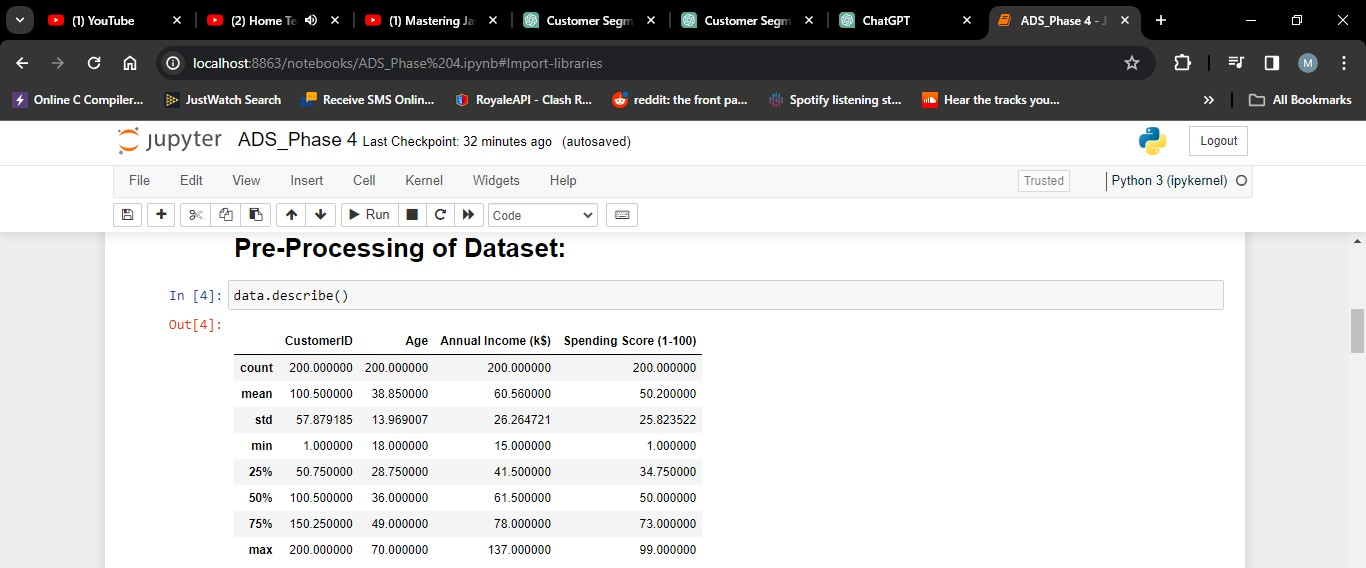


**1. pd.read\_csv ("Mall\_Customers.csv"):** Uses the read\_csv function from the pandas library to read the Mall Customers dataset from a CSV file and create a Data Frame named data.

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

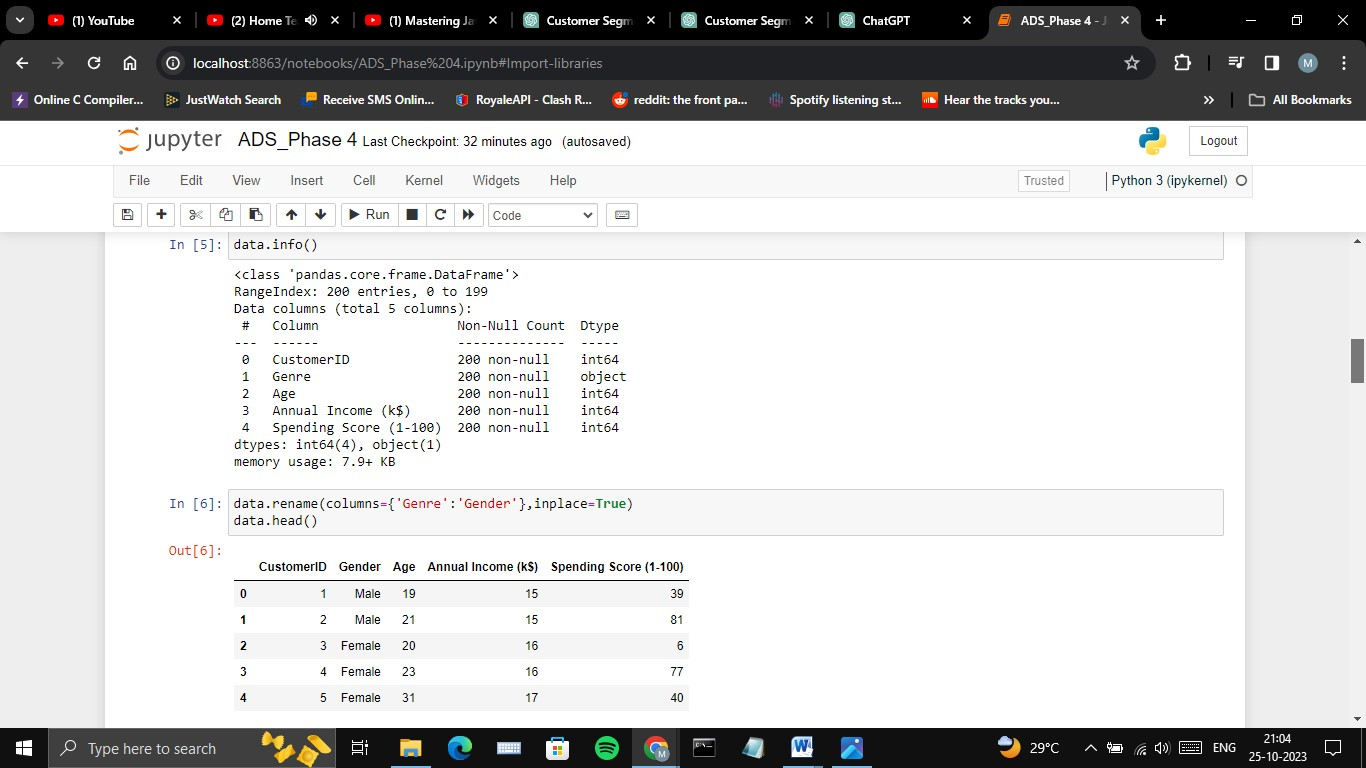
**3. Data.shape ( ):** Return a tuple representing the dimensionality of the Data Frame. Tuple of array dimensions.

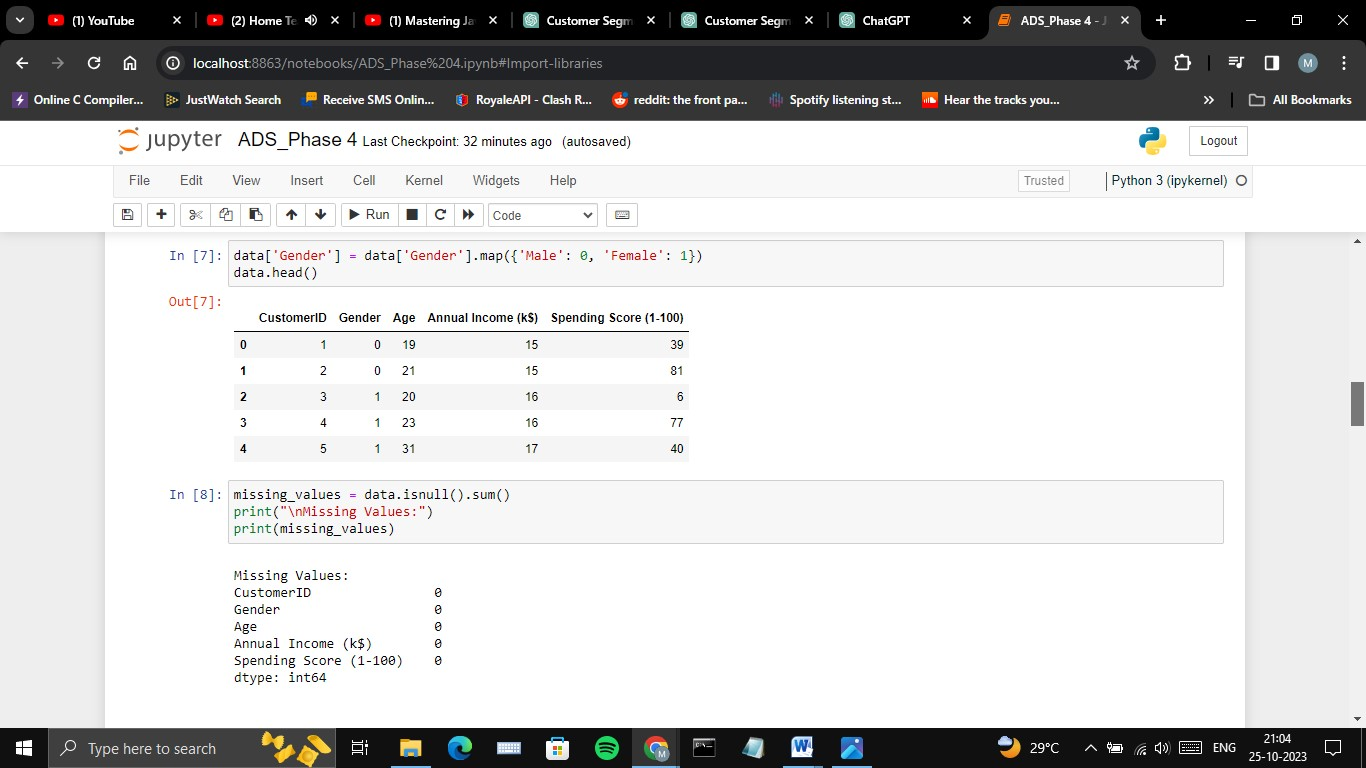
**Data Pre-Processing:**



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

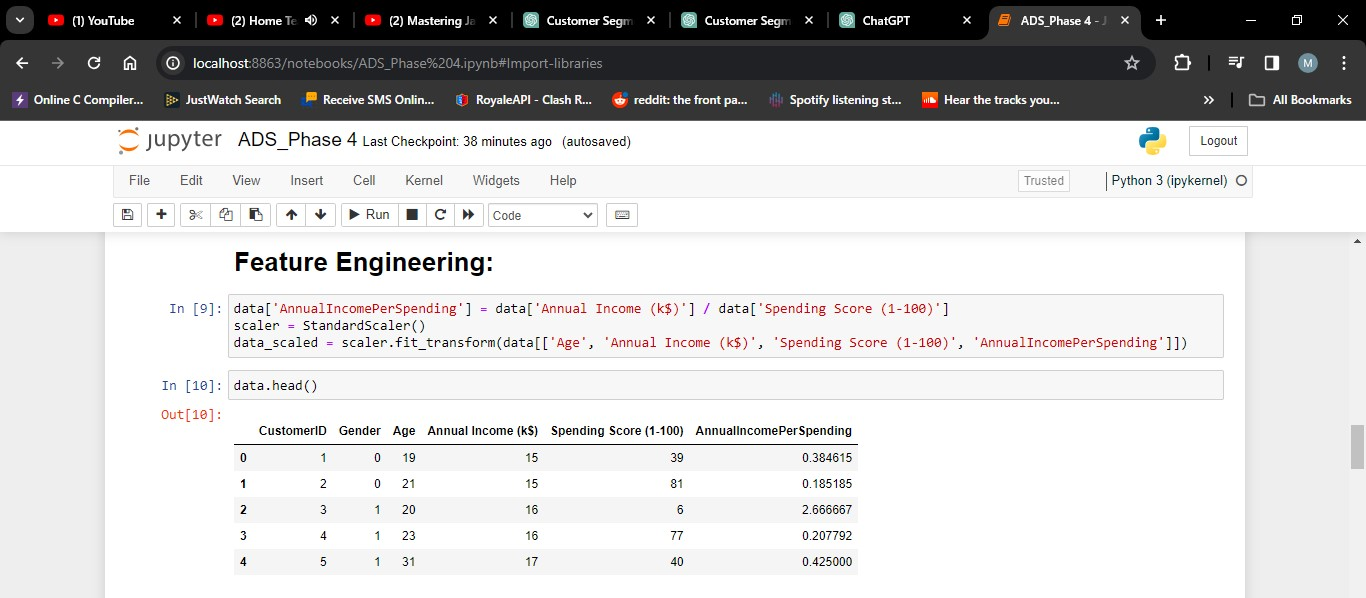
.In this code, the 'Product ID' column is dropped as it's considered irrelevant. If there are missing values, you can choose to drop rows or fill in missing values based on your dataset and domain knowledge.





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:

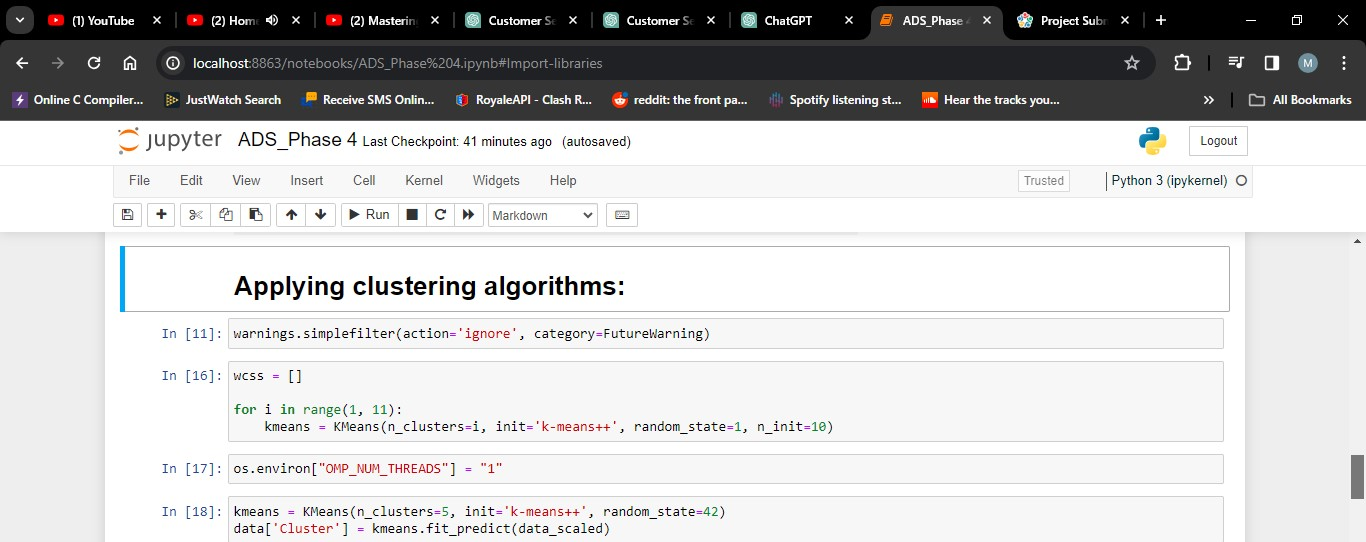
**1. Initialization:** Start by selecting K initial cluster centers randomly or strategically.

**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).

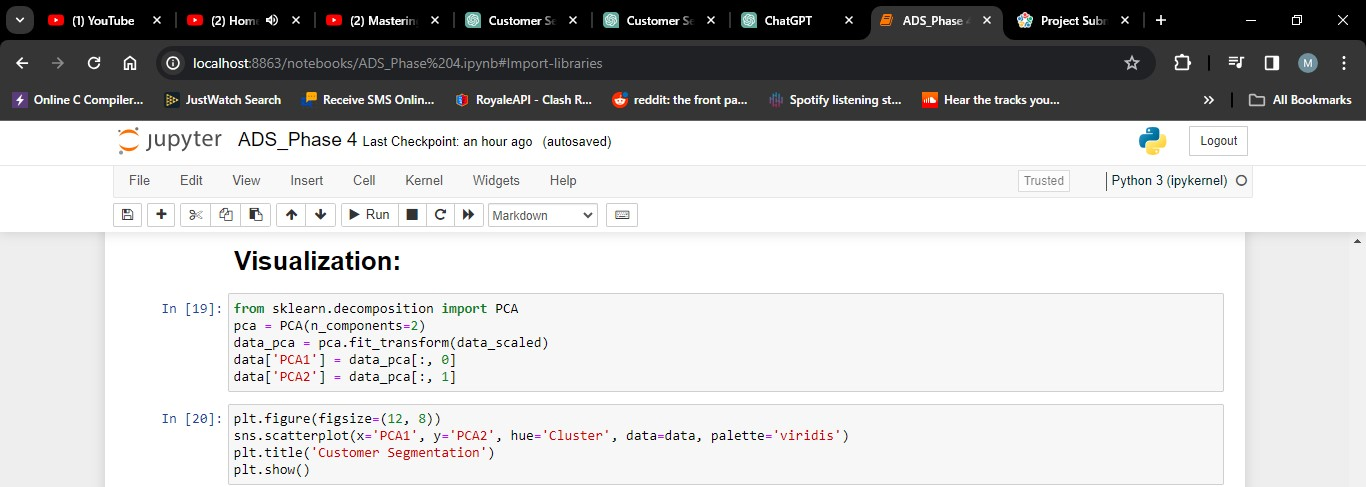
**Result:** The final cluster centers represent the cluster's characteristics, and the data points are grouped into K distinct clusters based on their similarity to these centers.

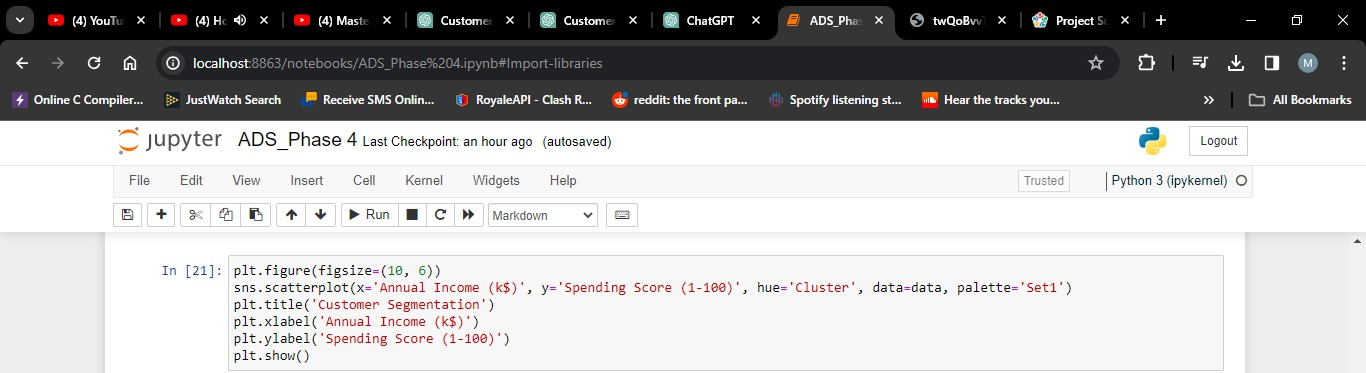


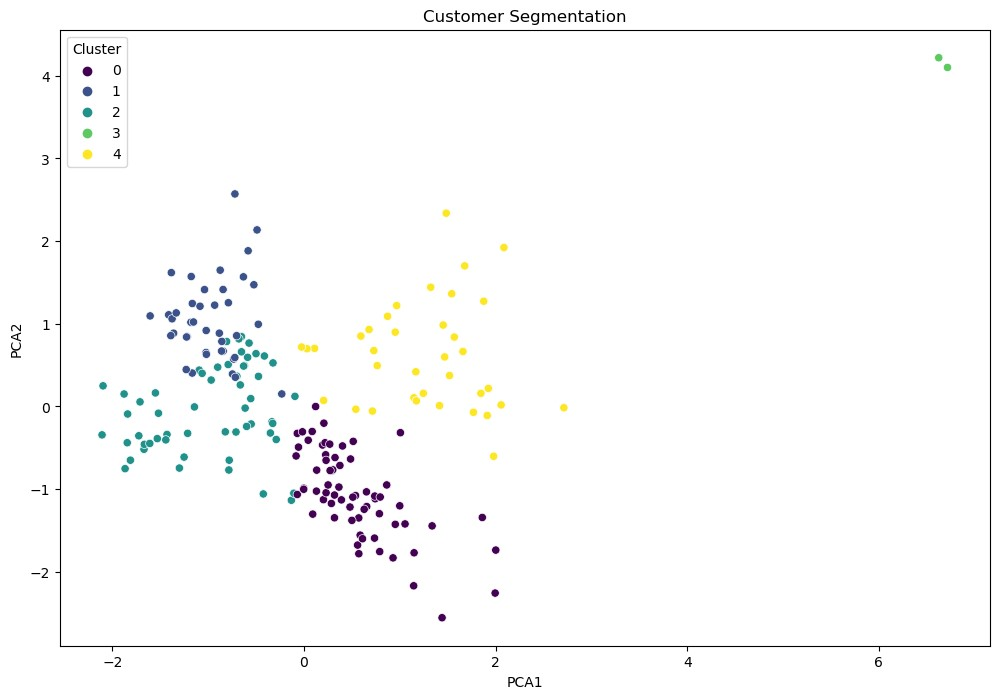
**Visualization:**

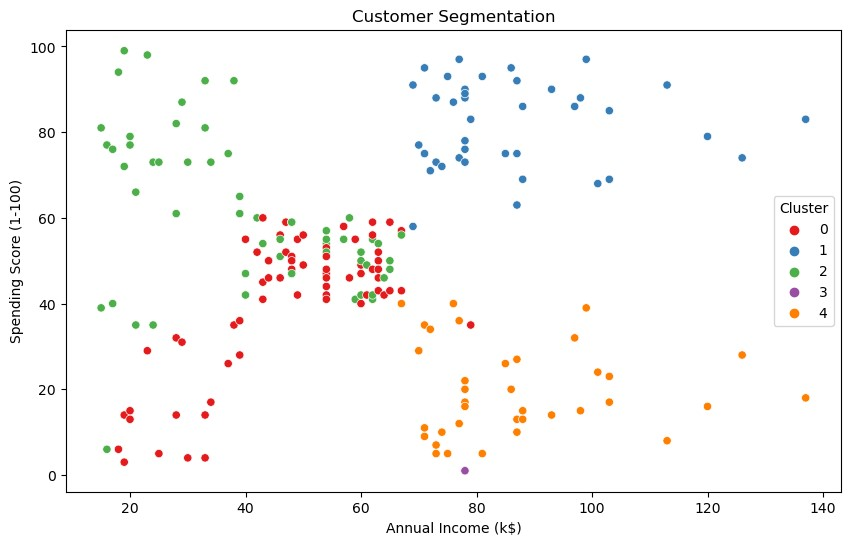
**Visualizing the Clusters:** We'll visualize the clusters using scatter plots to see how customers are grouped based on their spending behaviour and annual income.

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.

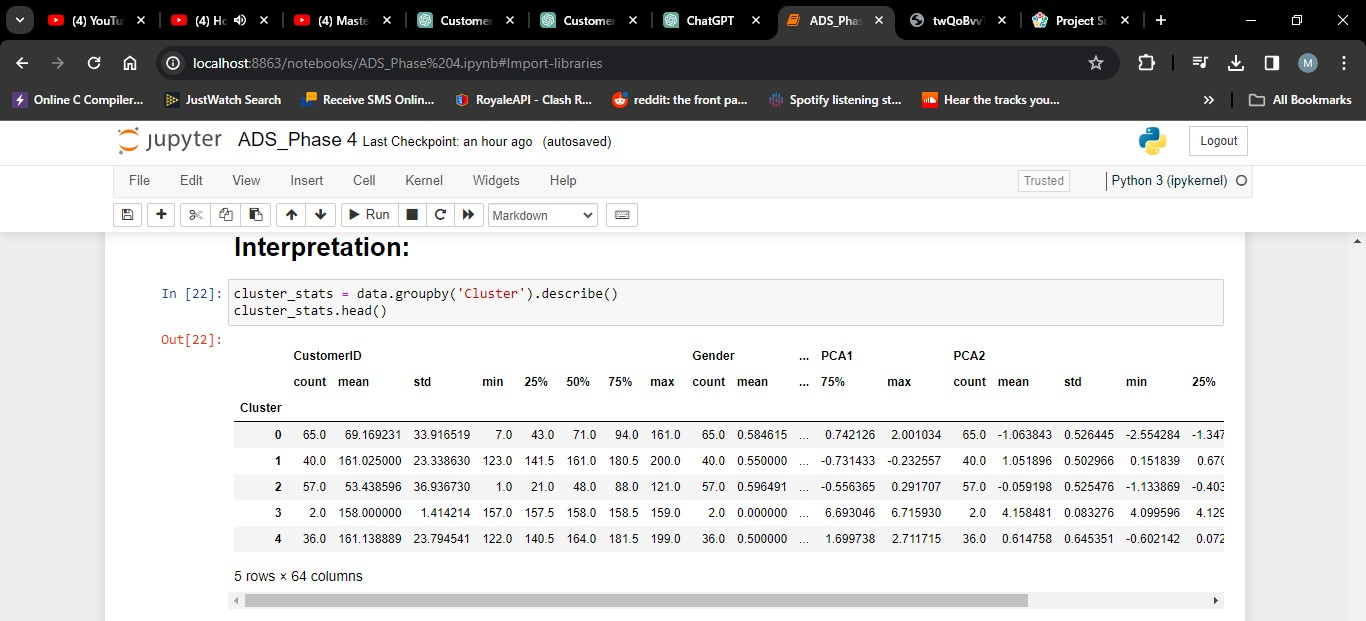
**3. Cluster 2:** Low-income customers with low spending.

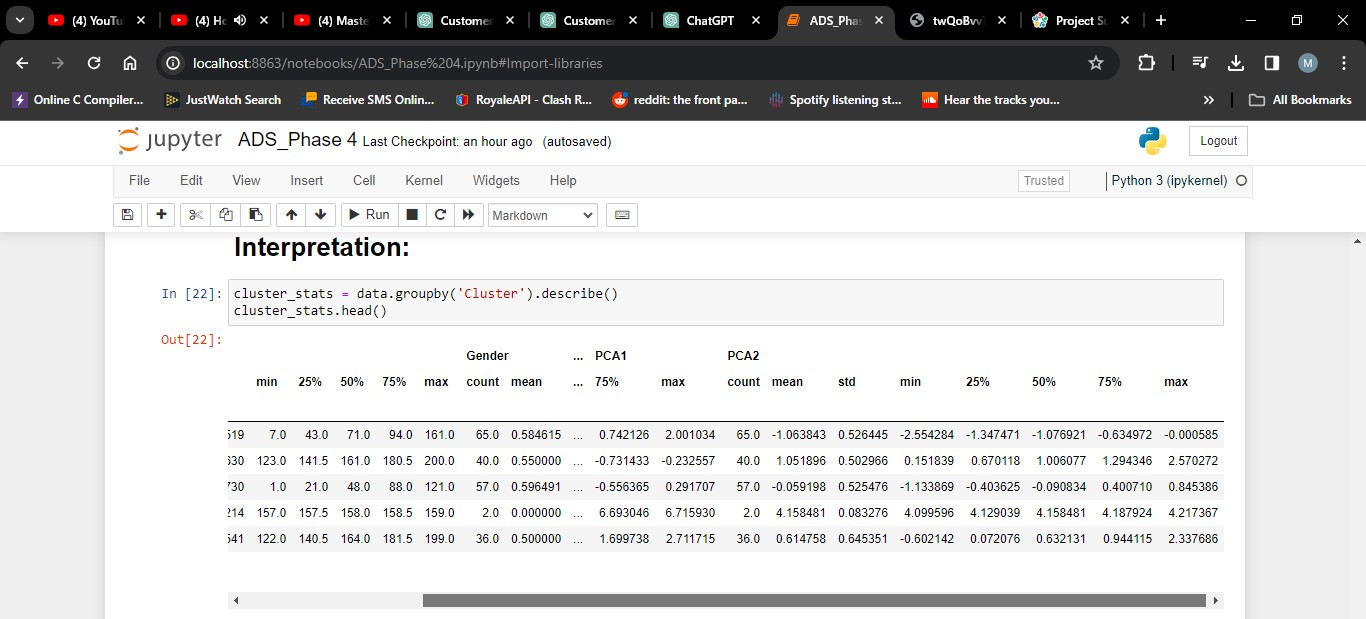
**4. Cluster 3:** High-income customers with low spending.

**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.







**Conclusion:**

In this design thinking process we have outlined a structured approach for solving the problem of customer segmentation for personalised marketing. Each step is iterative, allowing for continuous learning and improvement. By understanding the needs of customers, defining data requirements, ideating creative solutions, and prototyping and testing various components, in this process we aim to deliver a robust and effective customer segmentation solution.