**Applied Data Science**

**Phase 3: Development Part 1**

**Customer Segmentation using Data Science**

**Project definition:**

* Customer segmentation is the method of distributing a customer base into collections of people based on mutual characteristics so organizations can market to group efficiently and competently individually.
* The purpose of segmenting customers is to determine how to correlate with customers in multiple segments to maximize customer benefits. Perfectly done customer segmentation empowers marketers to interact with every customer in the best efficient approach

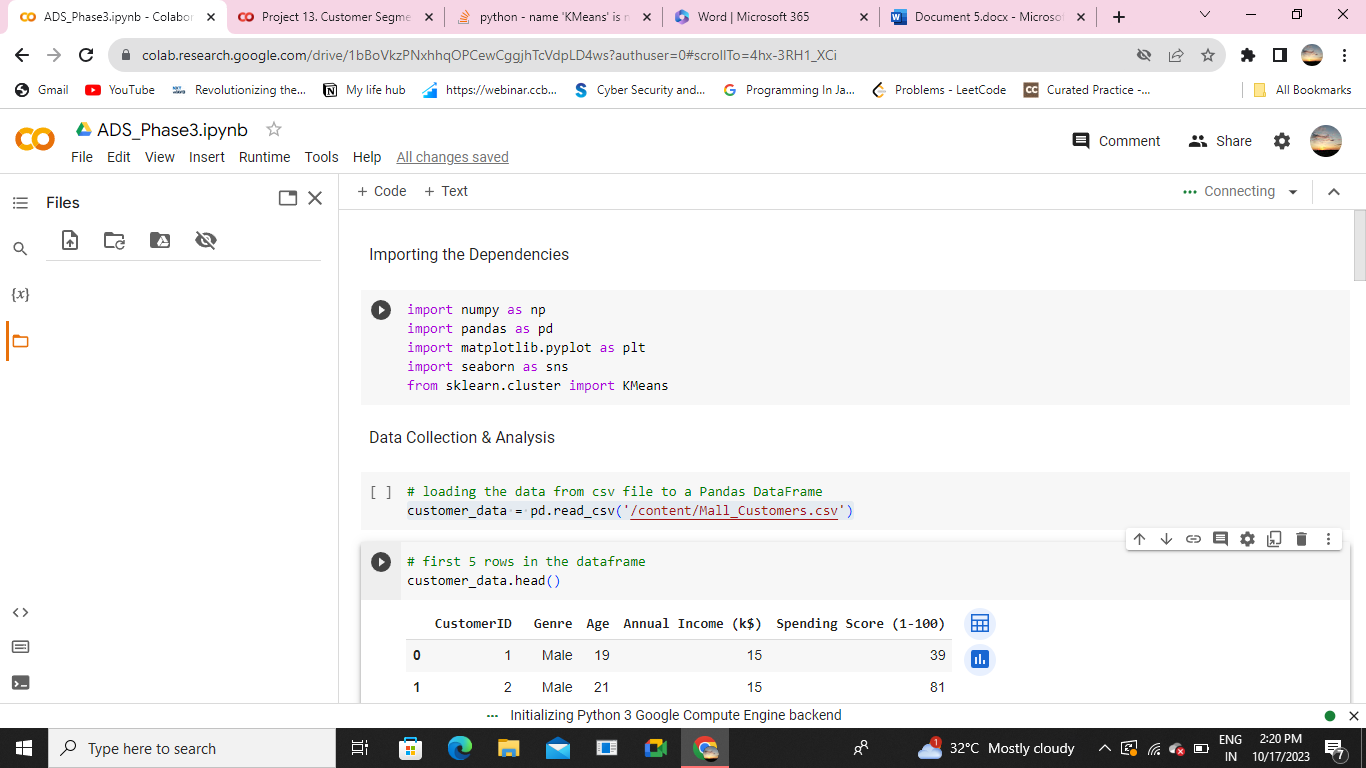
**Dataset explanation:**

The data includes the following features:

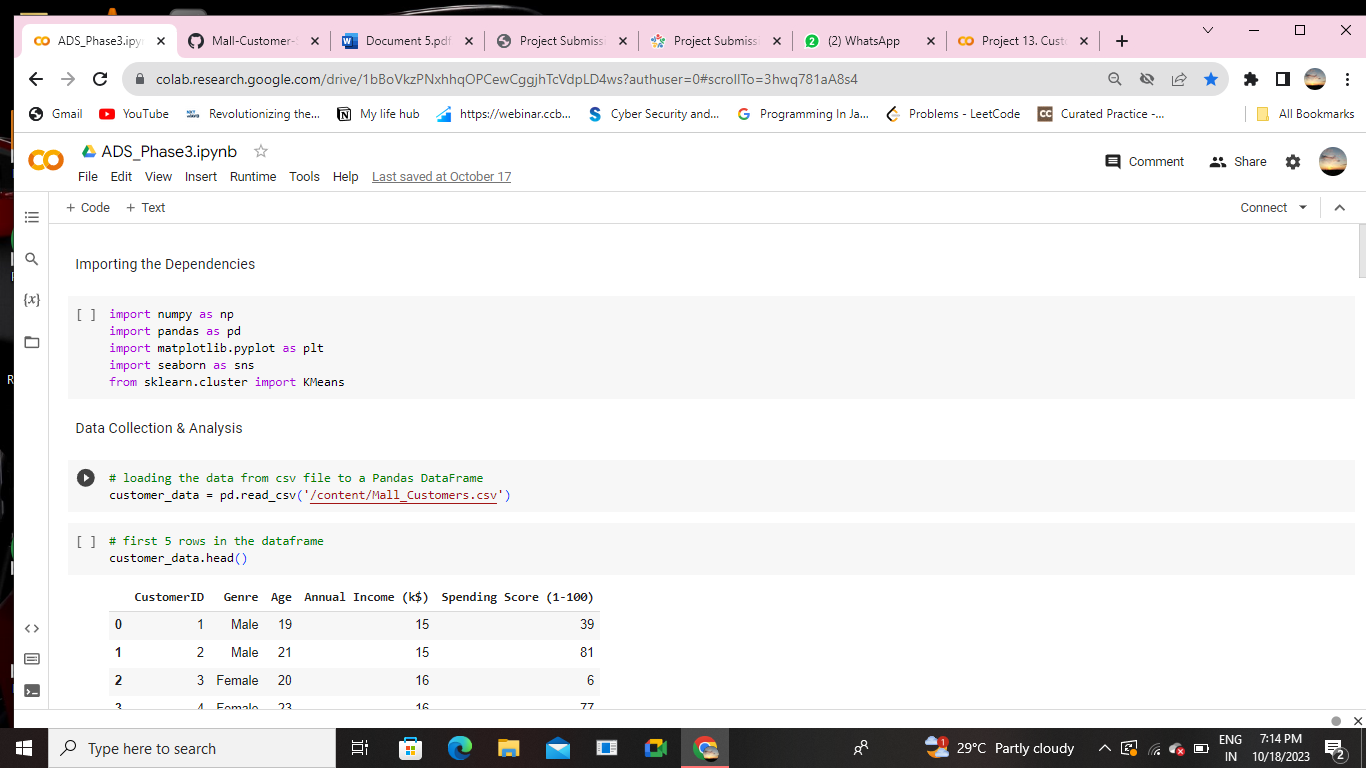
1. Customer ID
2. Customer Gender
3. Customer Age
4. Annual Income of the customer (in Thousand Dollars)
5. Spending score of the customer (based on customer behavior and spending nature)

**Importing the required libraries:**

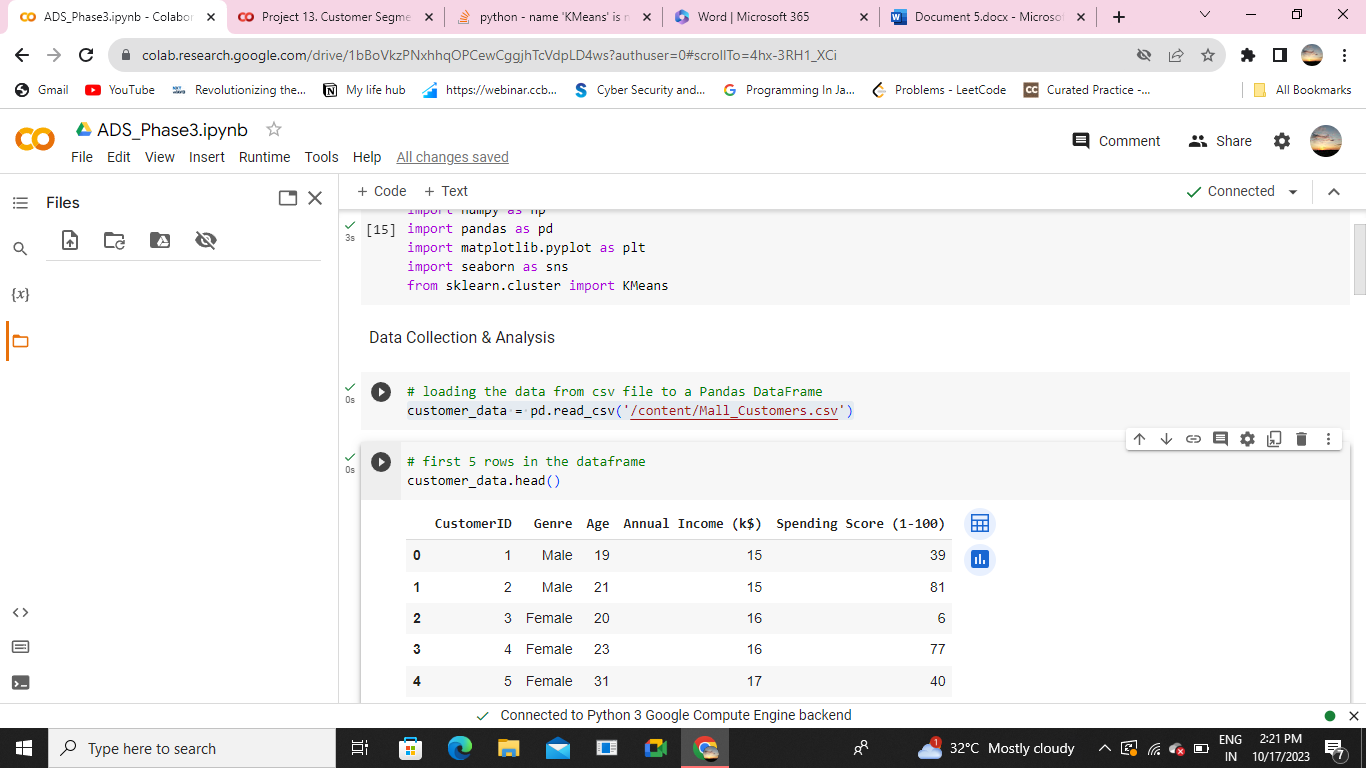
Let’s start the development part of customer segmentation using data science by importing the necessary Python libraries.



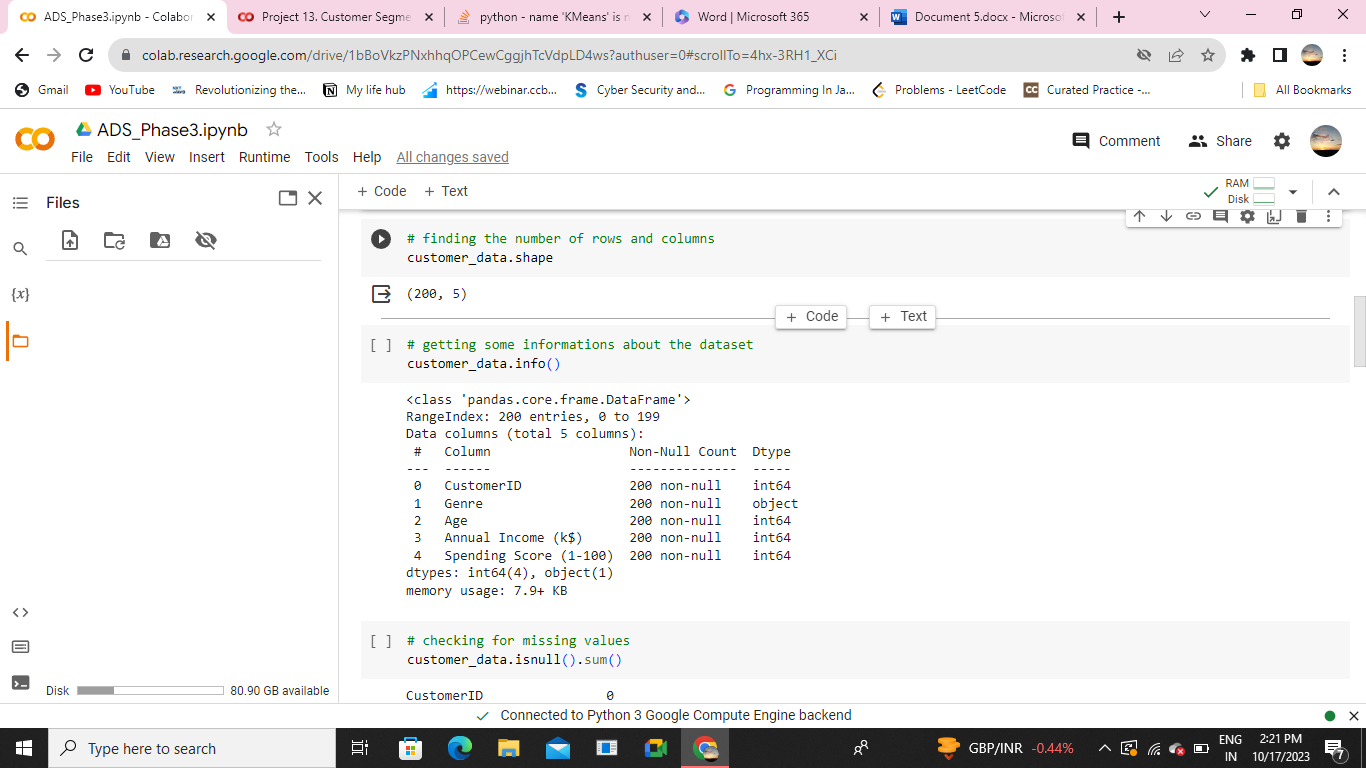
**Importing the dataset:**

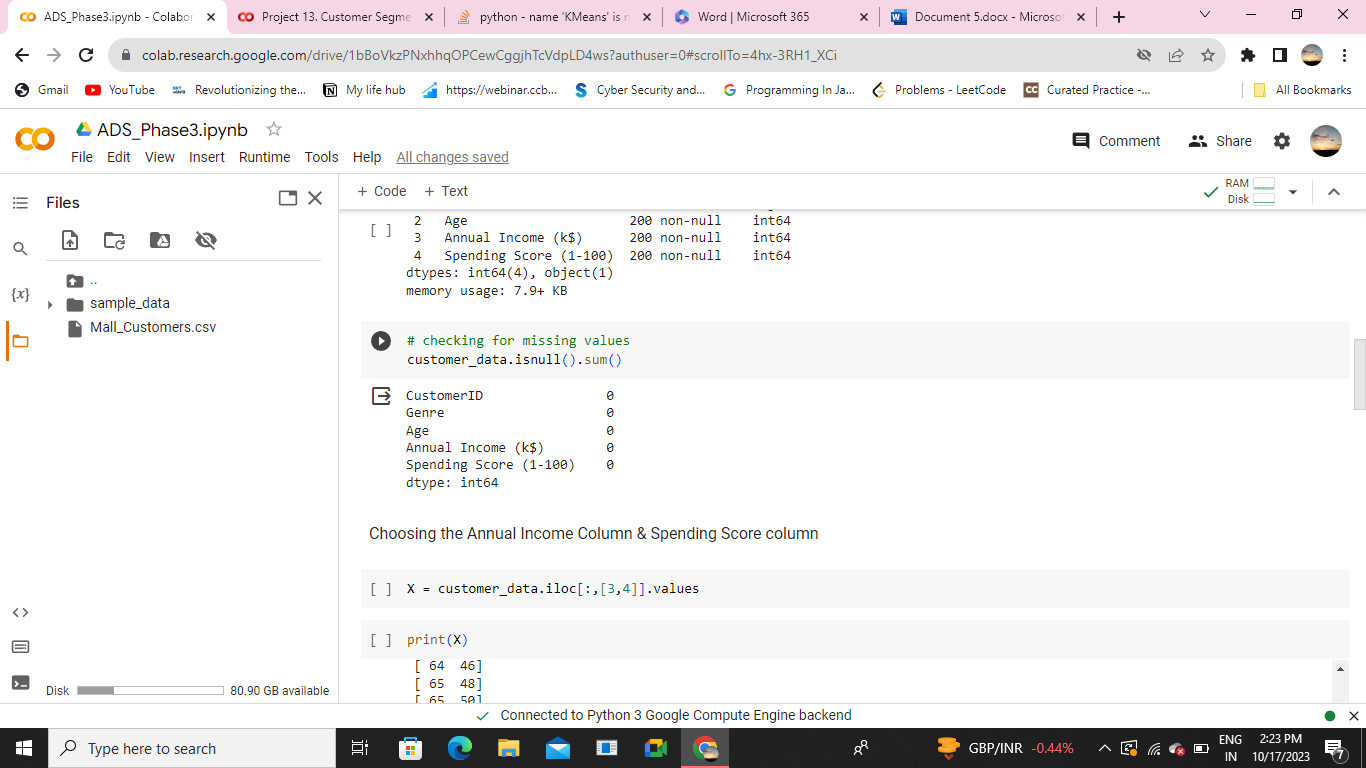


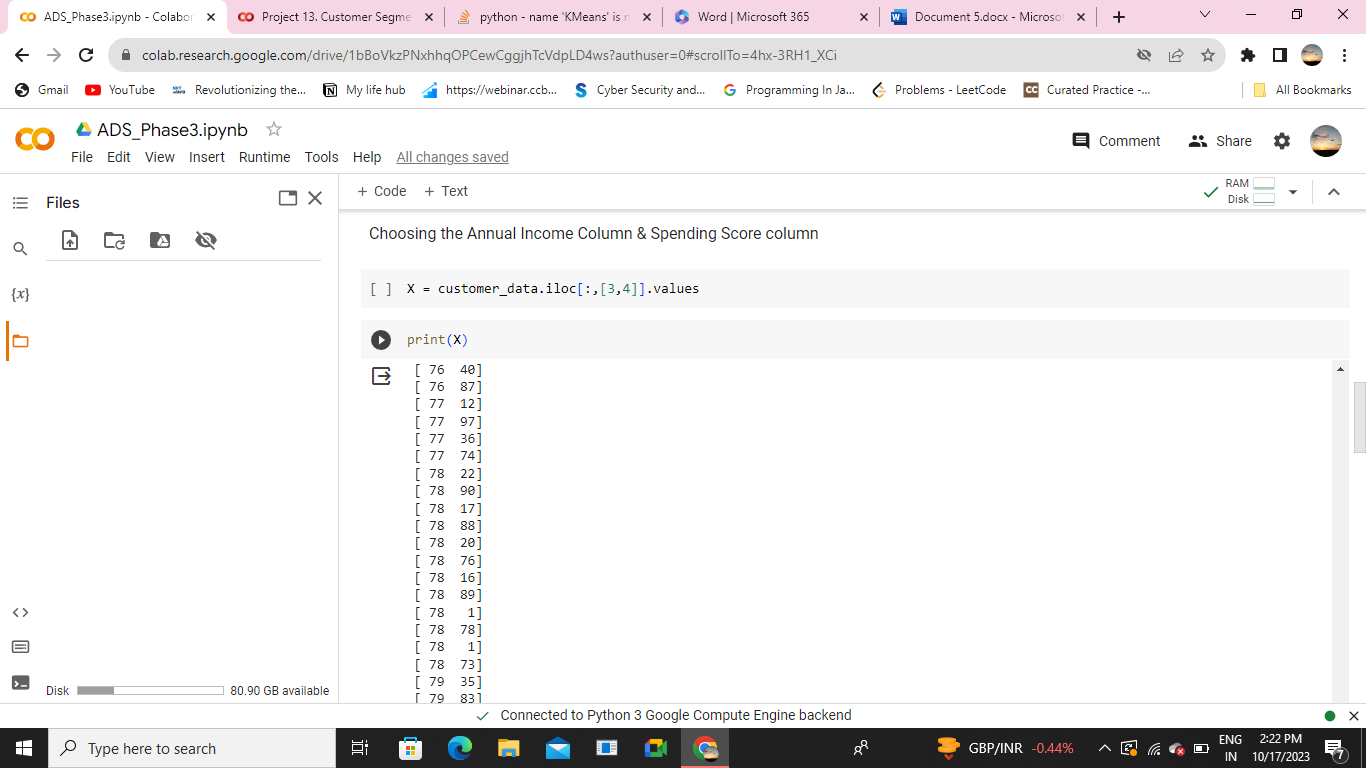
**Printing first five rows of the dataset:**

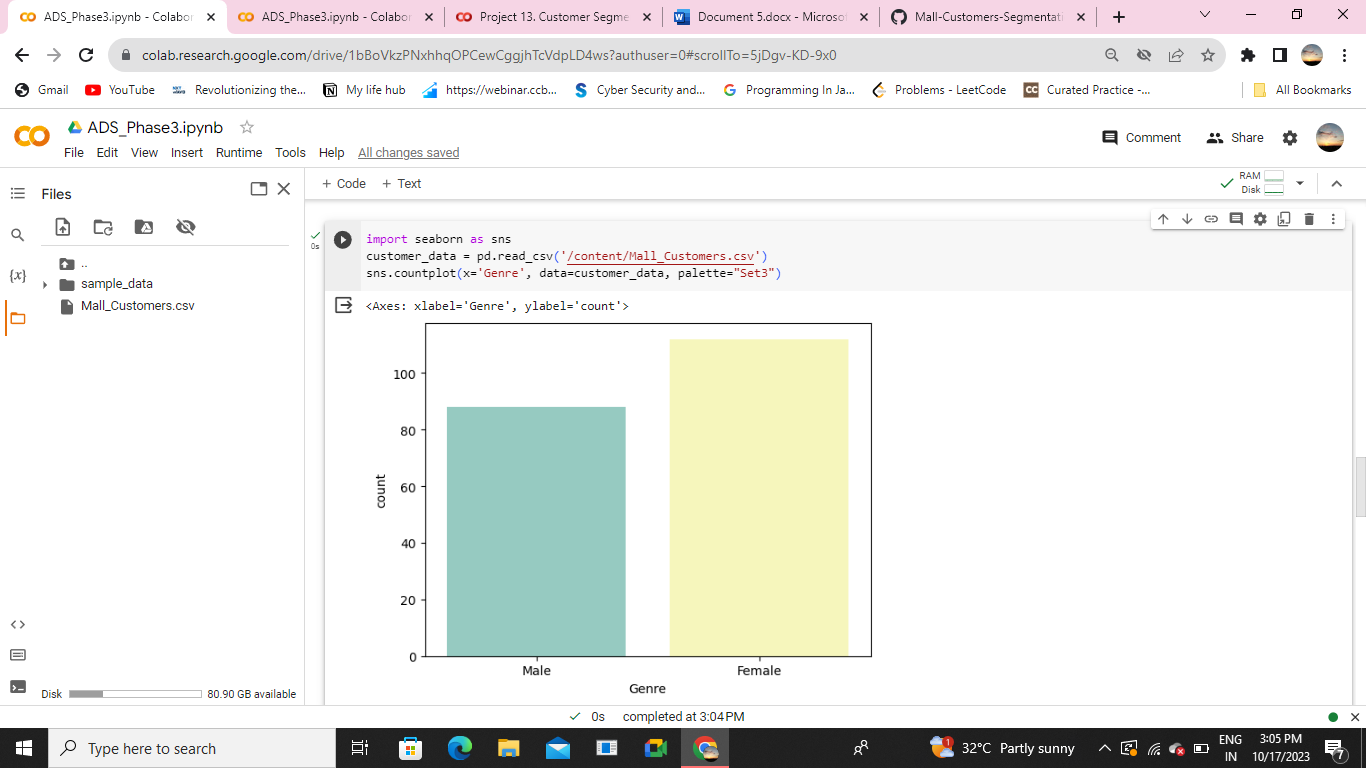


**Analyzing the dataset:**



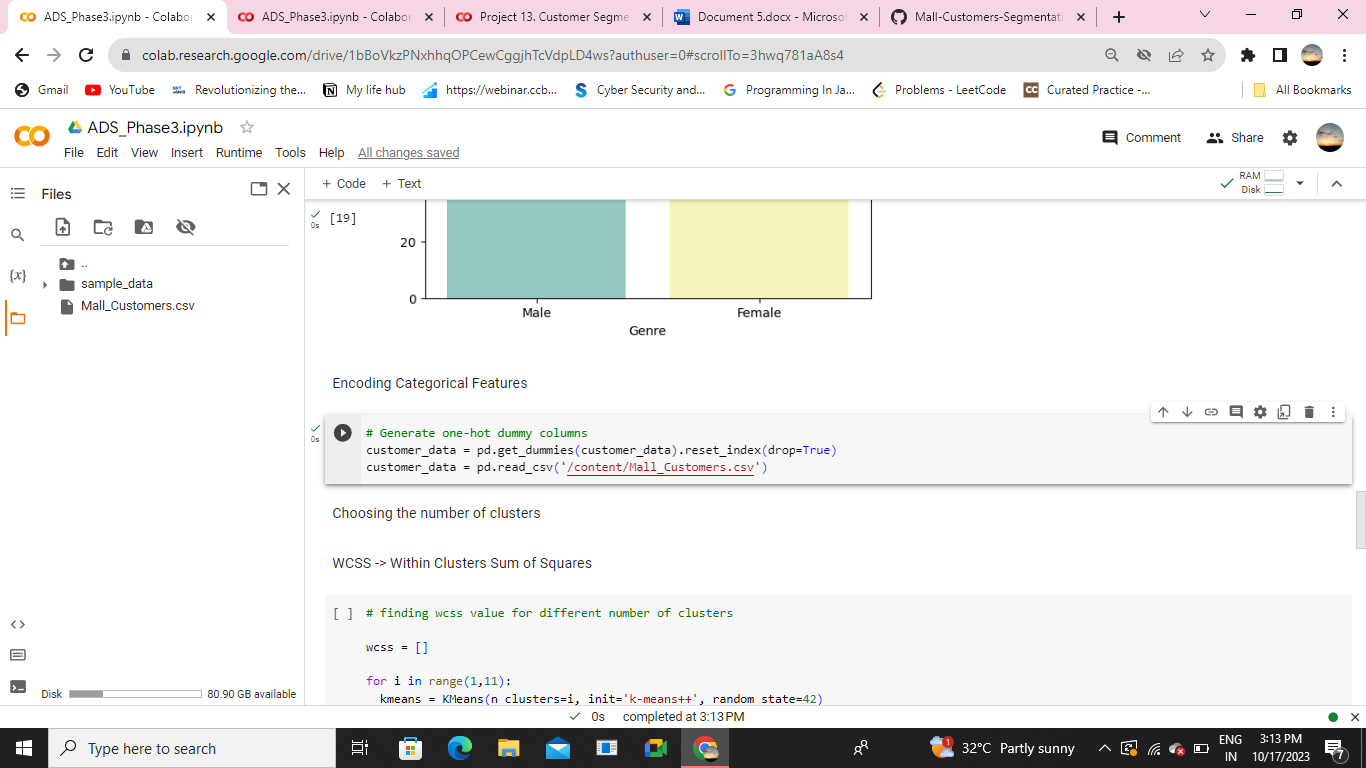






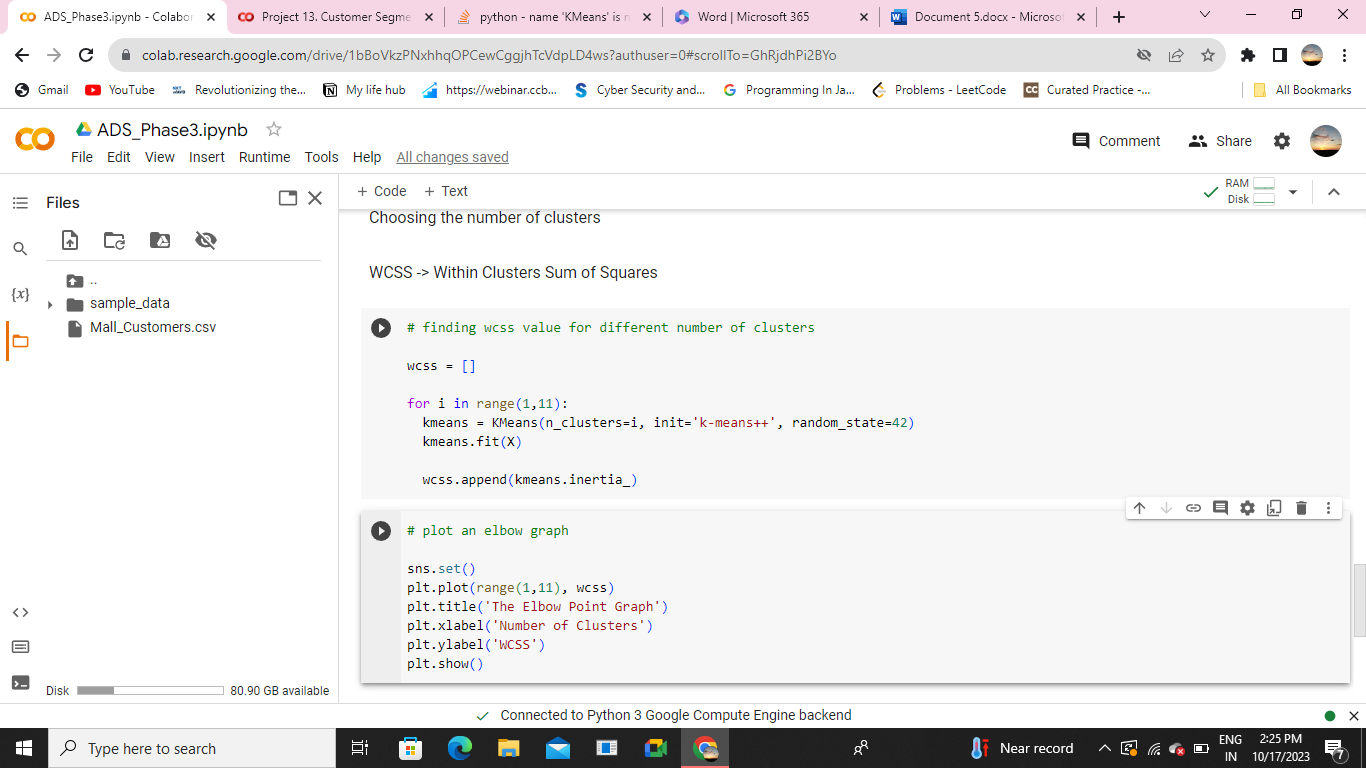
**Encoding the categorical features:**

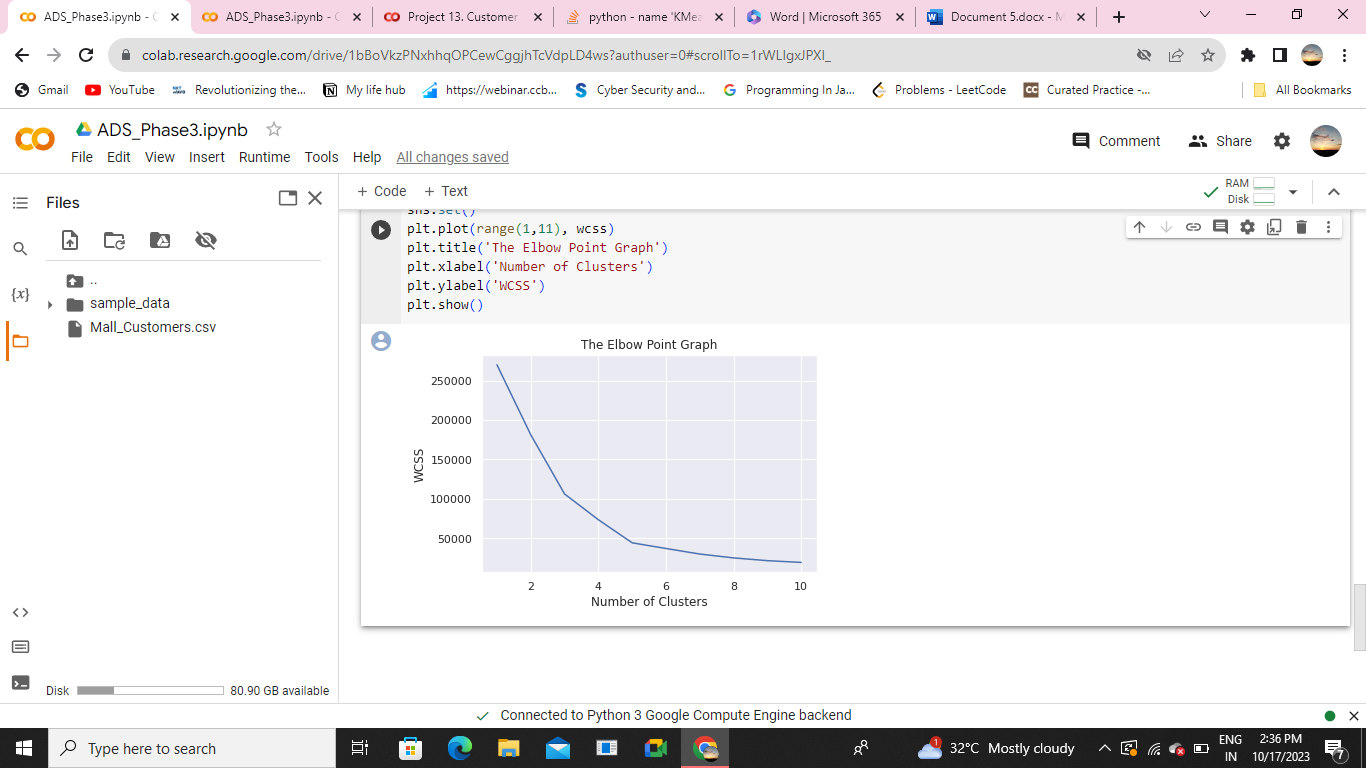
* Encoding categorical data using one- hot encoding (OneHotEncoder) is a process of representing categorical variables in a numerical format that can be used in various machine learning algorithms
* Encoding categorical data using OneHotEncoder is a process in data projects where categorical variables are transformed into a numerical format, specifically binary vectors.
* Each unique category or label within a categorical variable is converted into a separate binary column, and for each observation, the column corresponding to its category is marked with a “1,” while all other columns are set to “O.”
* This method is used to make categorical data compatible with machine learning algorithms that require numerical input.
* OneHotEncoder ensures that the categorical data is represented in a way that doesn’t introduce false ordinal relationships or numerical values, preventing biases in the model’s interpretation of the data



**Choosing the number of clusters:**

K Means Clustering Using the Elbow Method For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow.

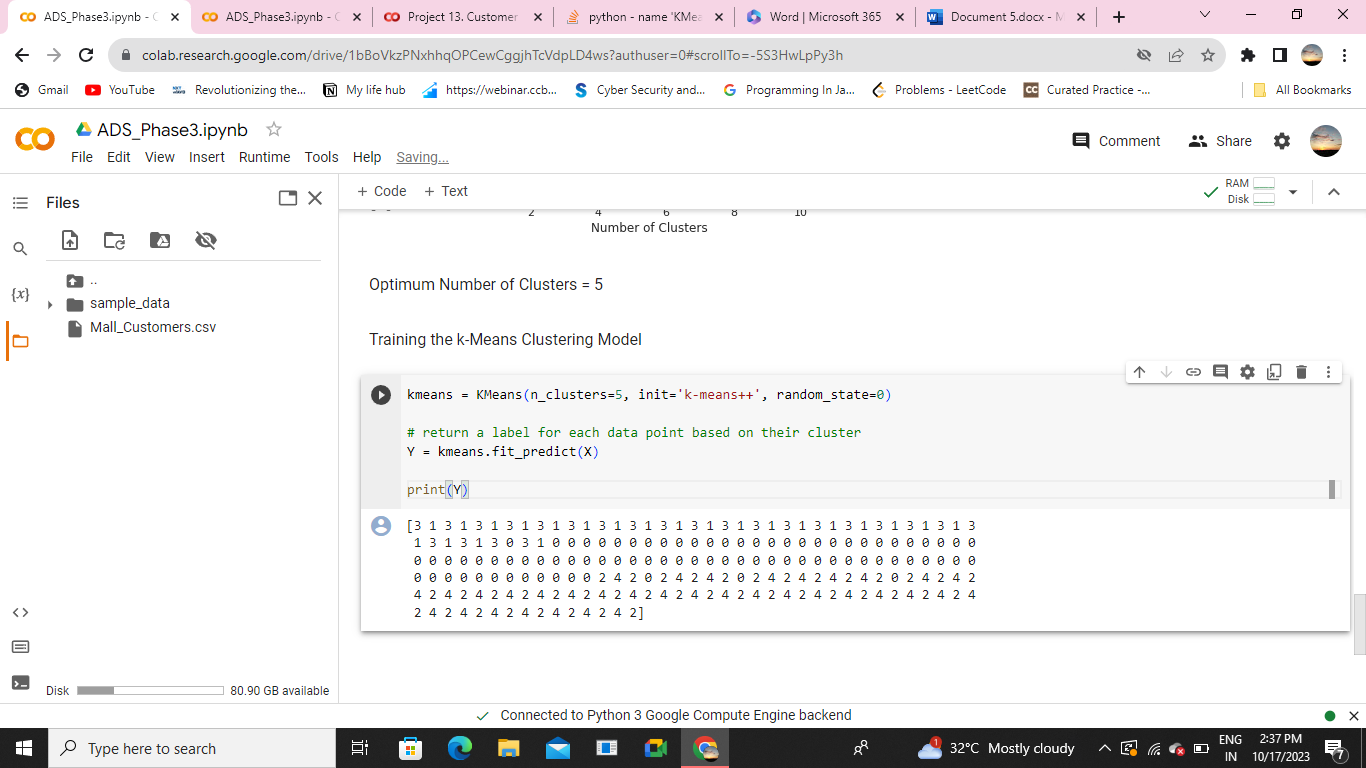




**Training the k-Means algorithm:**

**1.Testing set:** The testing set is a smaller portion of the data, usually around 20-30% of the dataset. It is kept separate and is not used during the model training phase. Instead, it is used to evaluate the model’s performance by making predictions or performing analyses and comparing them to the actual, known outcomes.

**2.Training set**: This subset contains a majority of the data, typically around 70-80% of the dataset. It is used to train machine learning models or perform data analysis tasks. The model learns patterns, relationships, and trends within the data from this set.



* So, this is how we can do Customer Segmentation using data science.