**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

**Department of Electronics and Telecommunication Engineering**

**Subject: Machine Learning Program: B.Tech/MBA.Tech**

**Sem: III/V ACAY: 2020-21**

**MINOR PROJECT**

|  |  |
| --- | --- |
| Roll No.:  N035  N041  N049  N063 | Name:  Trisha Sarkar  Swapnil Singh  Tarun Tanmay  Anushka Khare |
| Program: MBA Tech CE (3rd Year) | Semester: V |
| Date of Submission: 31/10/2020 | Grades: |

**Aim:**

To study and implement the K-Medoids clustering algorithm on the chosen dataset.

**Dataset Used:**

The dataset used is Mall Customer Segmentation Data, which has basic information about the customers visiting the mall through membership cards. The dataset has the following attributes are:

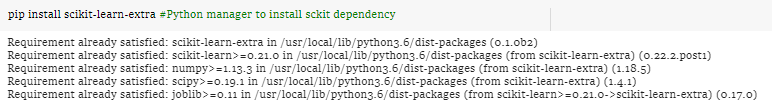
1. CustomerID, which is a unique ID assigned to the customer.
2. Gender, which tells us about the gender of the customer.
3. Age, which is the age of the customer.
4. Annual Income, which is the annual income of the customer.
5. Spending Score is a score assigned by the mall, based on customer behavior and spending nature.

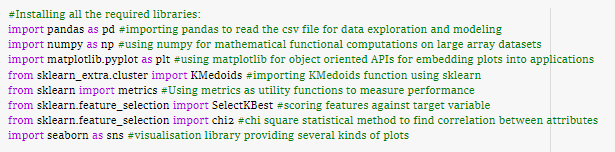
**Problem Statement:**

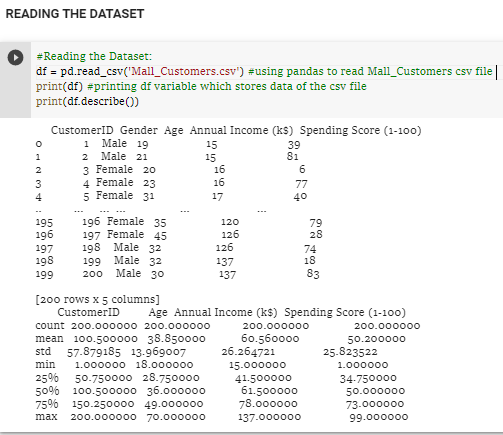
The owner of the mall wants to understand the customers, who can easily converge into target groups so that sense can be given to the marketing team and plan marketing strategies accordingly.

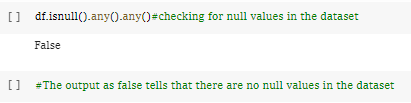
**Code:**

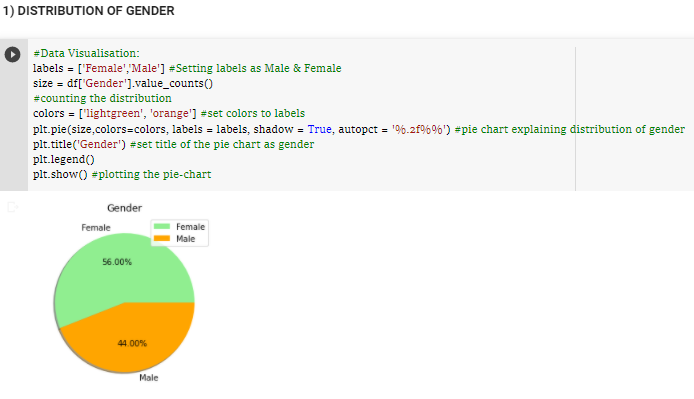
Google Colab File Link: <https://colab.research.google.com/drive/1d7ABkvXBzPAnQw3WiZ4yHHnyG2m8HHEL?usp=sharing&authuser=1#scrollTo=FBHlOl-H_OmN>

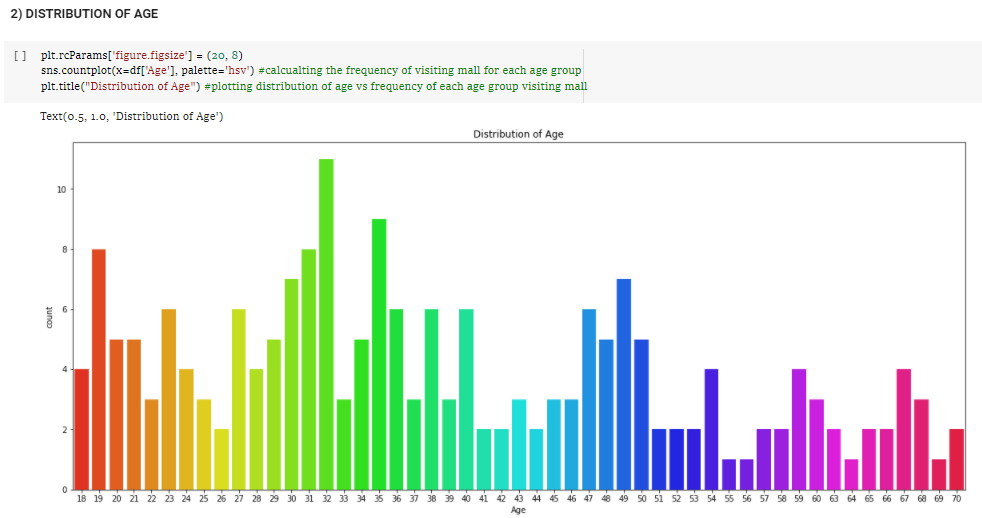
****

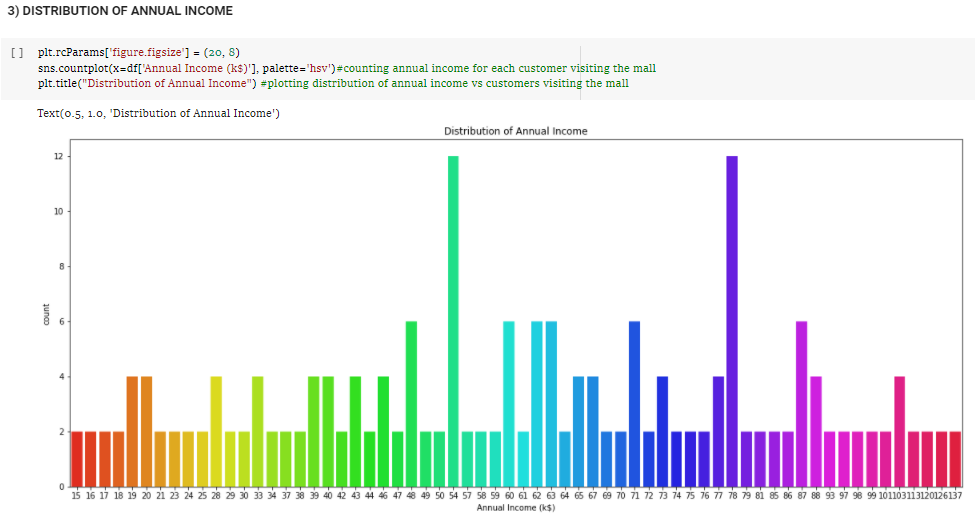
****

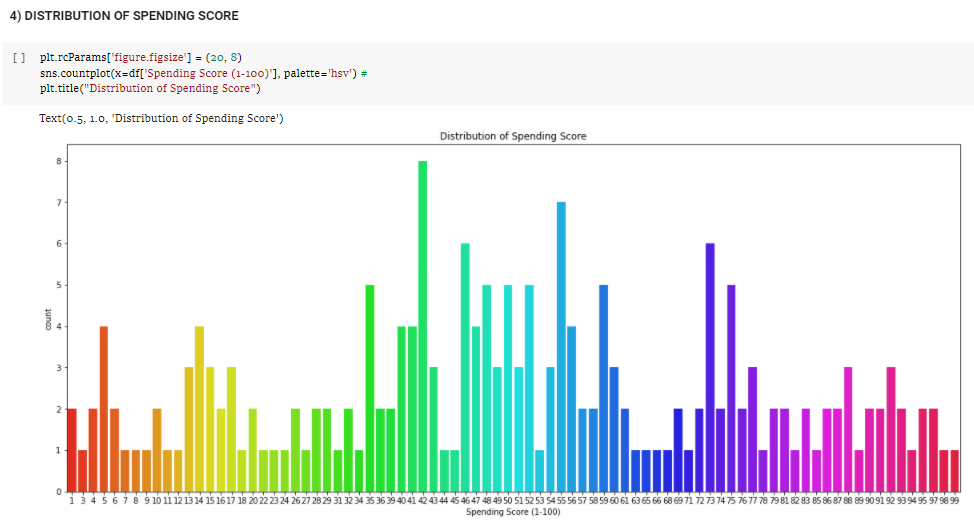
****

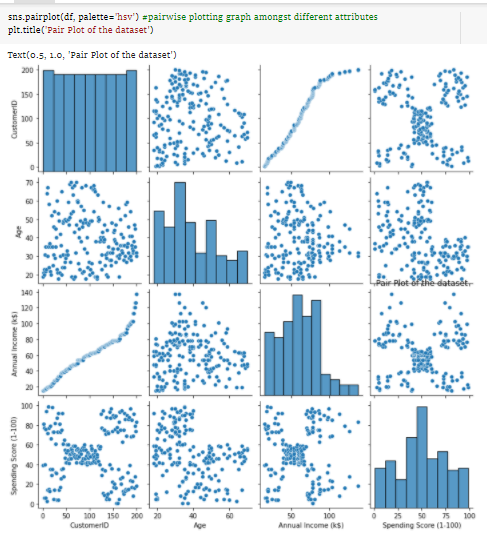
****

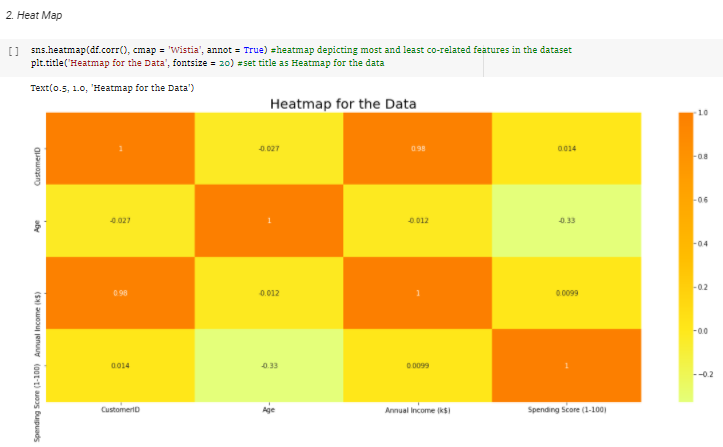
****

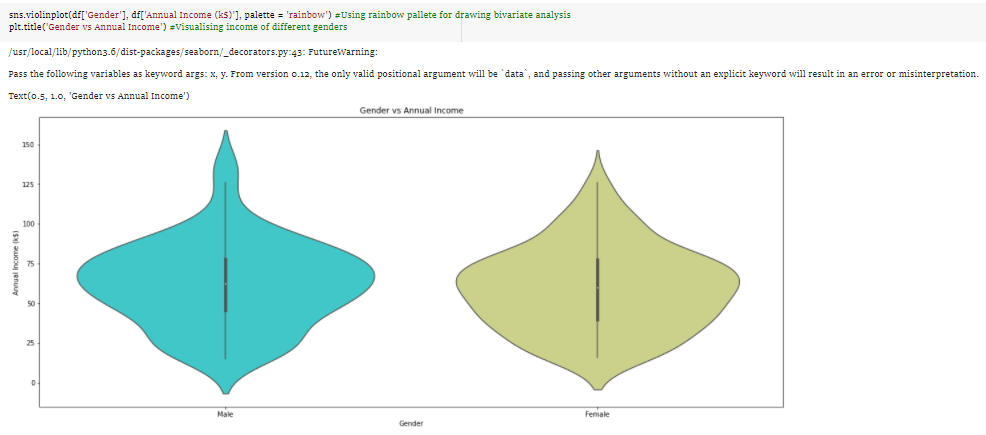
****

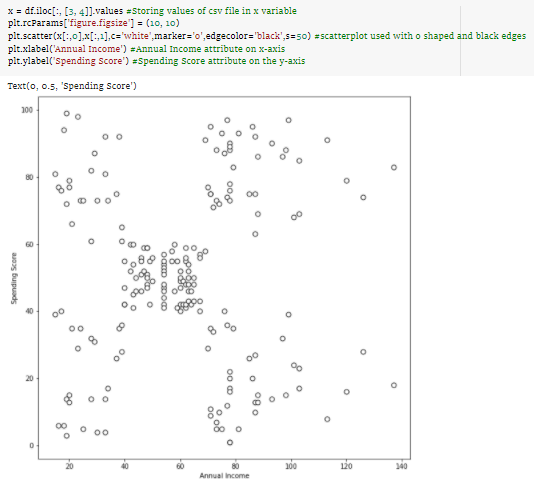
****

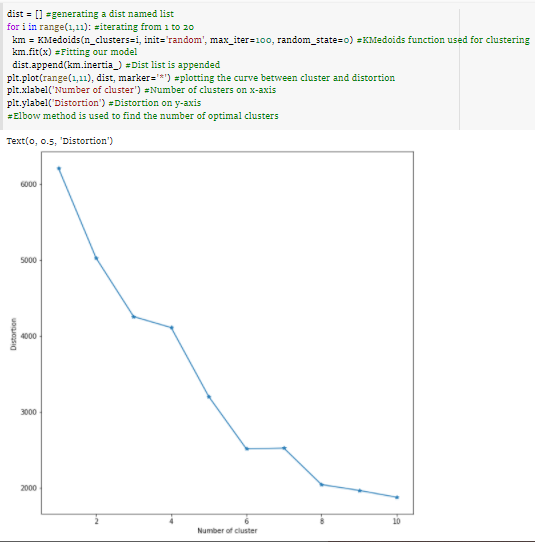
****

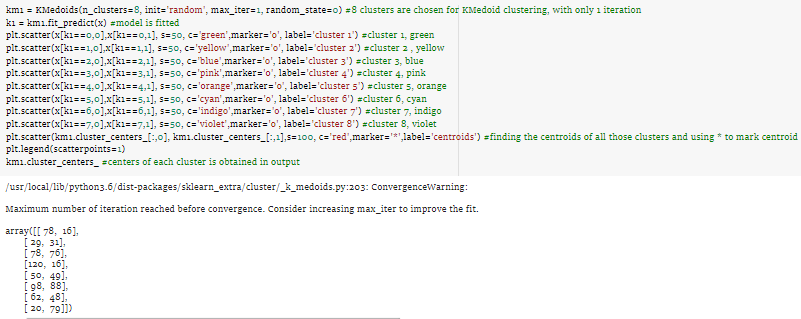
****

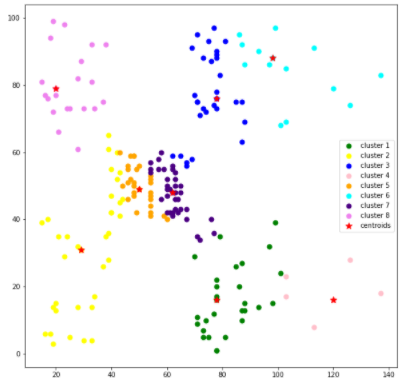
****

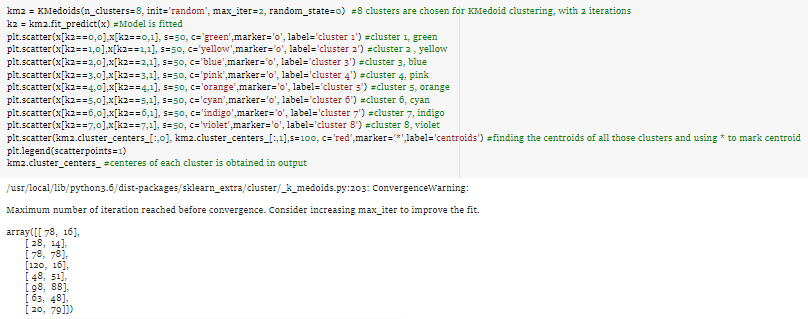
****

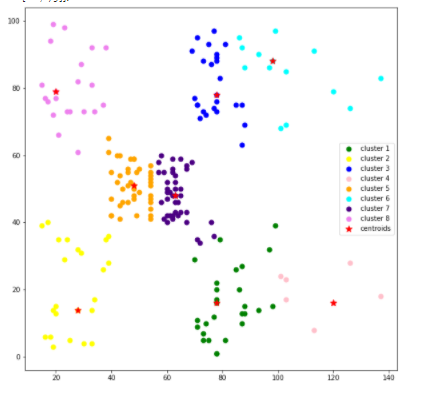
****

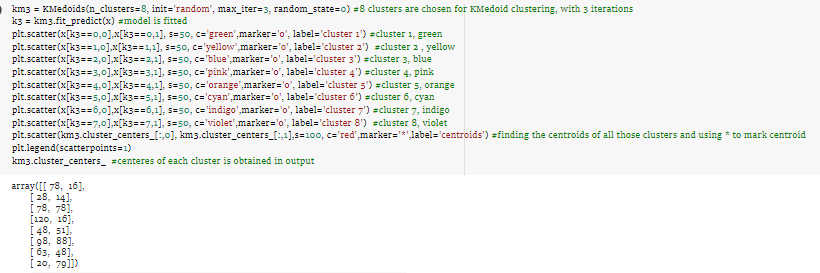
****

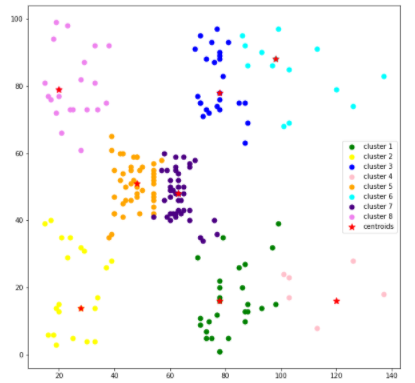
****

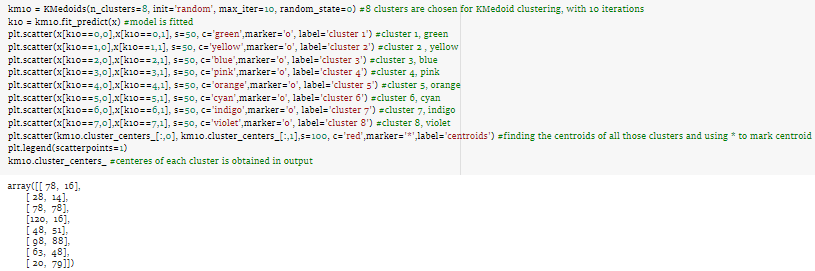
****

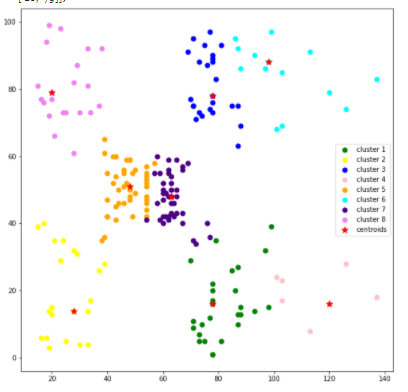
****

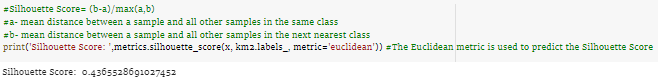
****

****

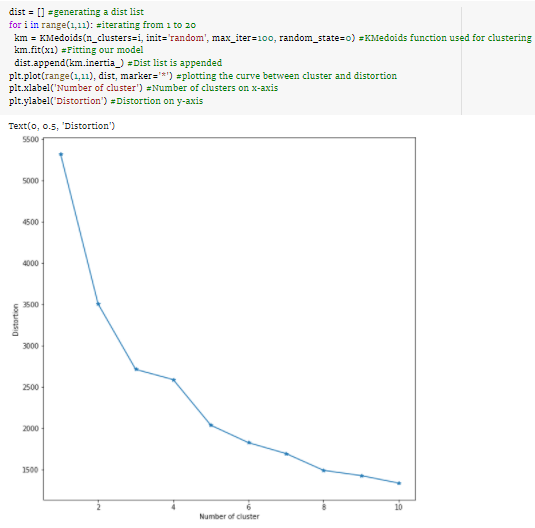
****

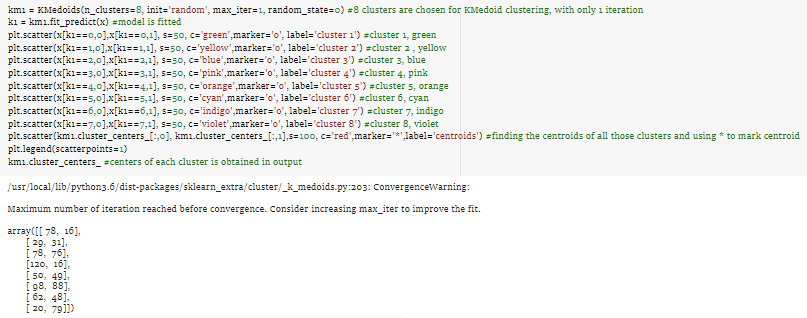
****

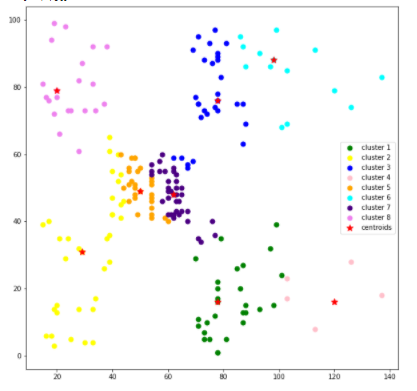
****

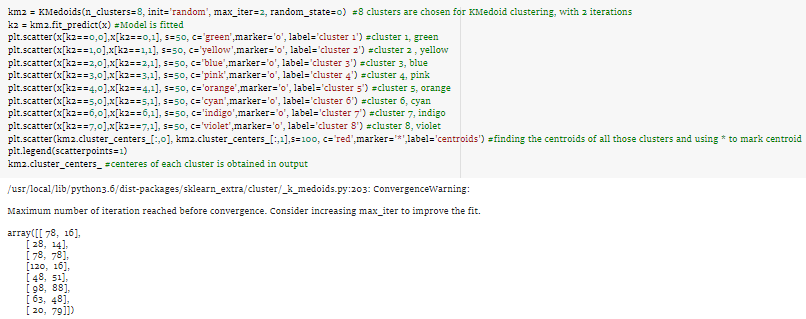
****

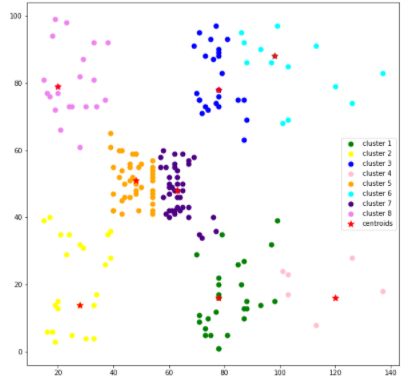
****

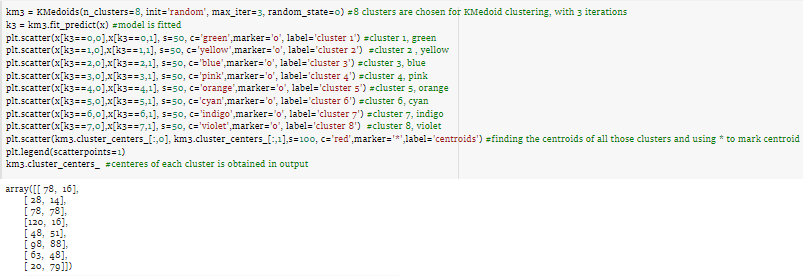
****

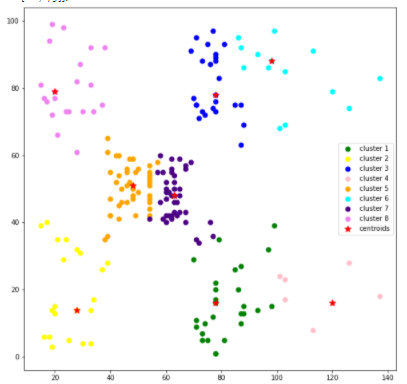
****

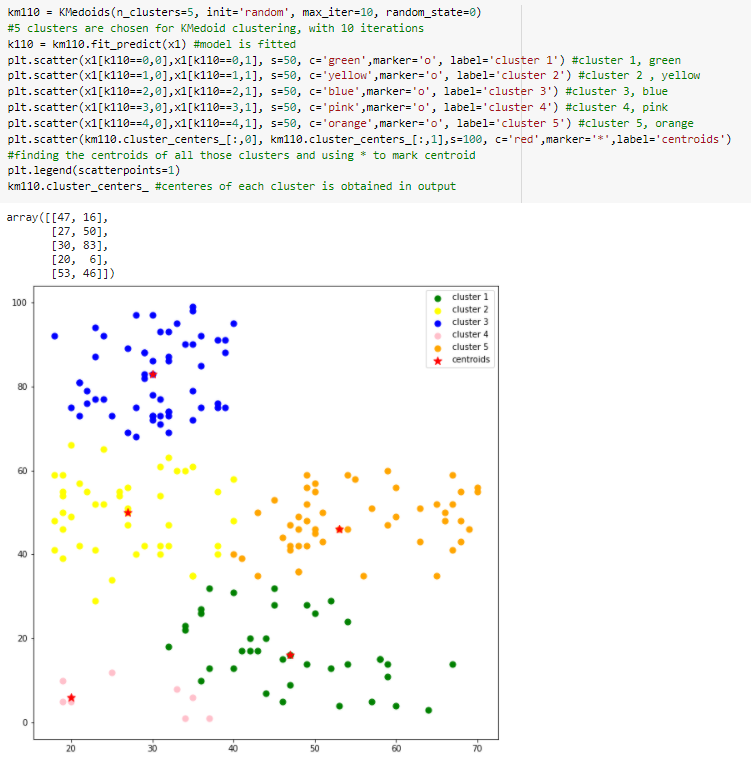
****

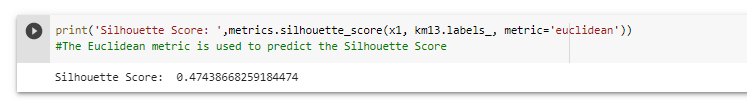
****

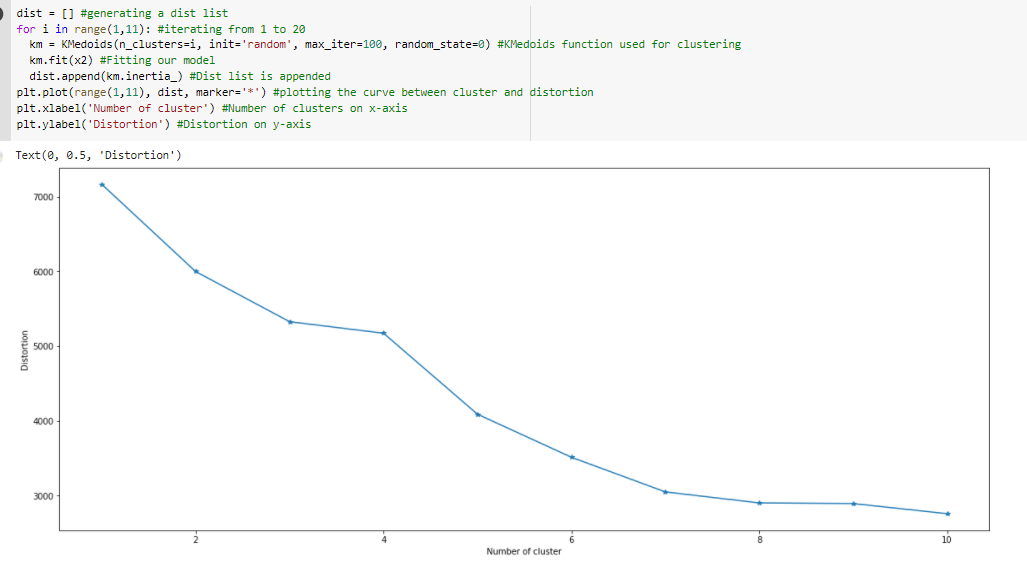
****

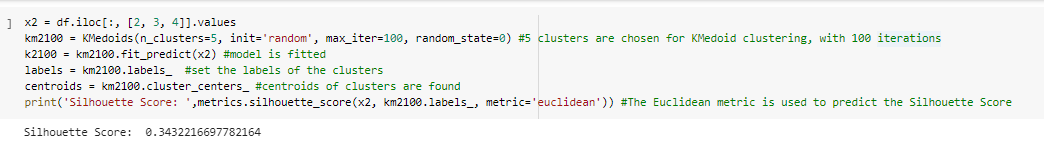
****

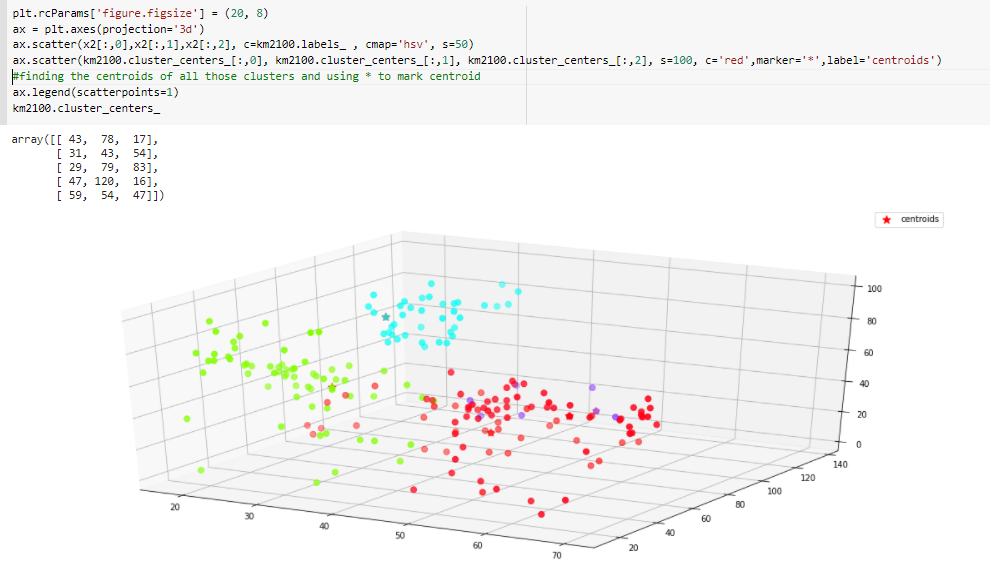
****

****

****

****

****

****

**Innovative Idea:**

The innovative idea used was visualizing all the data to make the inference of the data easier. The data set and its variables have been explored by visualizing each attribute using a pie chart and count plots. Count plots can be thought of as a histogram across a categorical variable, instead of a quantitative variable. The correlation between the attributes has been visualized using a pair plot, a correlation heatmap, and bivariate analysis, which helped decide the attributes to be chosen for clustering. For clustering, the data points and the elbow curve was plotted to see the distribution of data points and decide the elbow point. Then, the clusters were visualized by changing the max iterations to see the change in the centroid values. Lastly, a 3D scatter plot was also made to visualize the clustering after taking three attributes- Age, Income, and Spending Score.

**Conclusion:**

* A Mall Customer Segregation dataset was chosen to understand the buying patterns in customers on a lower level of abstraction. The dataset was cleaned and pre-processed to remove the null values.
* We have tried to venture into an Exploratory Data Analysis theme (EDA) by using data visualization to understand the data set at hand better. Pie plots, count plots, violin plots, 2-D scatter plots, 3-D scatter plots, Scatter plot matrix in the form of pair plots, and Heat maps.
* Choosing an optimal k value using elbow technique, implementation of the K-medoids algorithm by choosing appropriate parameters.
* 3-D scatter plot visualization of the clusters plotted using 3 attributes.
* The accuracy score of the clustering algorithm has been evaluated using the silhouette score. Silhouette Score is in the range of -1 to 1, a value that is towards 1 indicates a high-density cluster, a value towards -1 indicates incorrect clustering.