CUSTOMER SEGMENTATION OF A MALL

Author : Rahul Mallah

University Name : National Institute Of Technology Silchar

Mail ID : rahulmallah785671@gmail.com

INTRODUCTION

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively. Use K-means clustering and hierarchical clustering and also visualize the gender and age distributions. Then analyze their annual incomes and spending scores.

To make predictions and find the clusters of potential customers of the mall and thus find appropriate measures to increase the revenue of the mall is one of the prevailing methods of unsupervised learning. For example, a group of customers have high income but their spending score (amount spent in the mall) is low so from the analysis we can convert such types of customers into potential customers (whose spending score is high) by using strategies like better advertising, accepting feedback and improving the quality of products. To identify such customers, this project analyses and forms clusters based on different criteria which are discussed in the further sections.

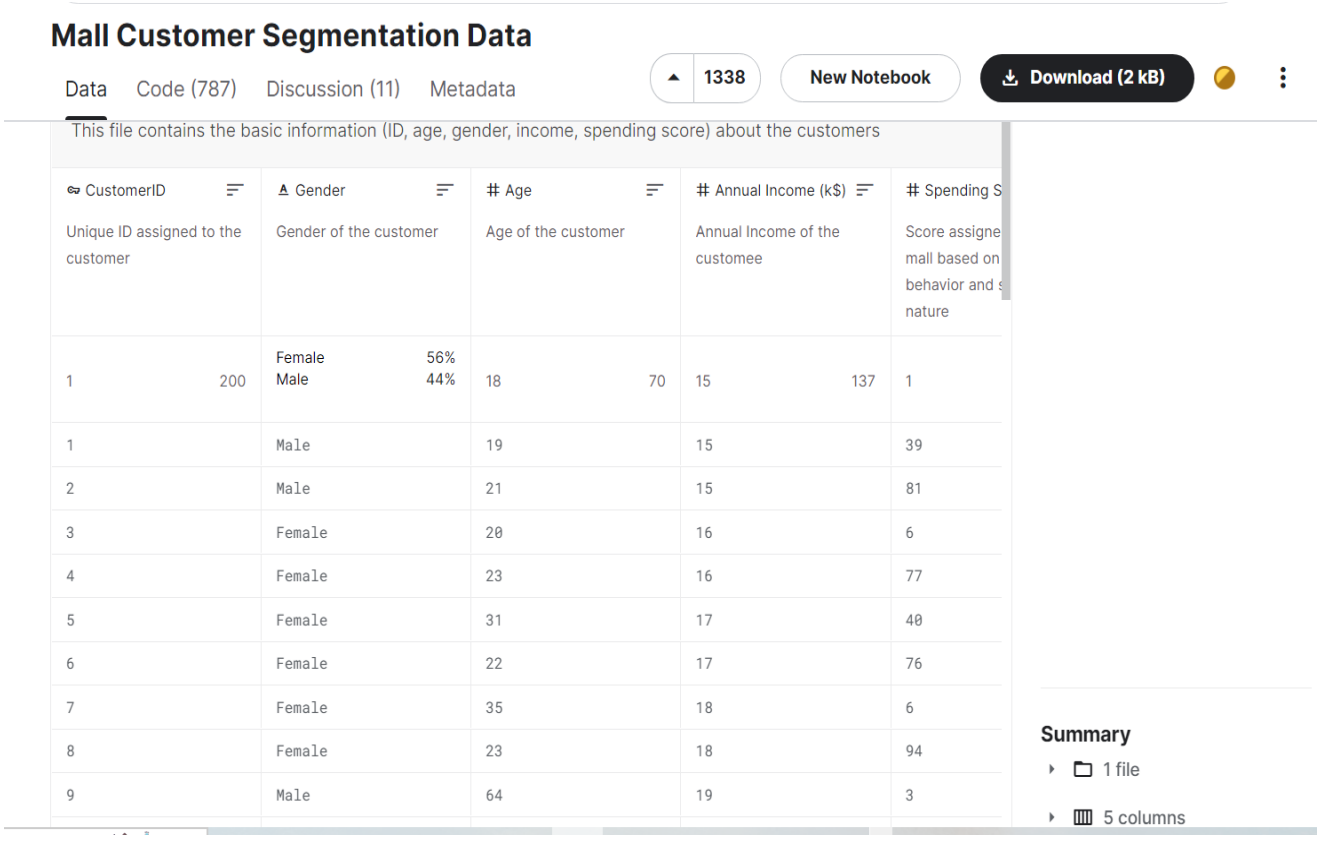
You own the mall and want to understand the customers like who can easily converge [Target Customers] so that the sense can be given to the marketing team and plan the strategy accordingly.

DATASET

The dataset name is Mall_Customers.csv consists of 5 columns which are CustomerID, Gender,Age, Annual Income (k$), Spending Score (1-100) where Gender is a categorical value and rest all features are numeric.

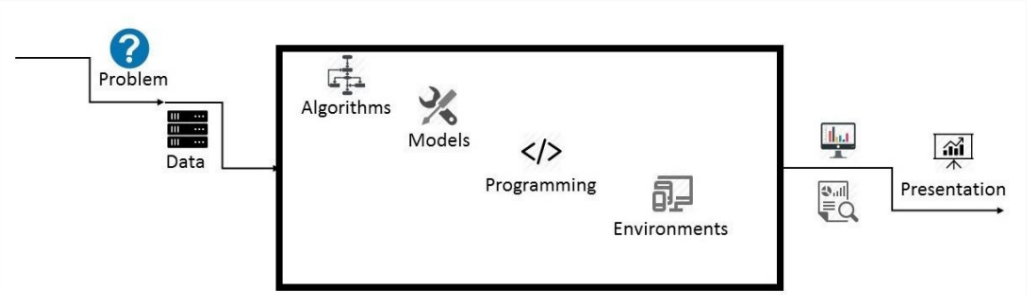
Dataset link →

https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python



METHODOLOGY

ARCHITECTURE OVERVIEW :



3.1 Data Science Project Architecture

PROJECT ARCHITECTURE :

Data : The size of the dataset is (200, 5) which is 200 rows and 5 columns. Also on dataset does not contain any NULL or NaN values.

Algorithms : K-means algorithm is used in this project to analyze and form clusters of customers based on their income and spending score features.

Model : K-means model is used and is hyper tuned parameters like n_clusters=5 using elbow method to find the optimal number of clusters also init='k-means++' to avoid random initialization trap.

Technologies and Libraries : Python 3.7, Numpy, Pandas, Matplotlib, Seaborn, Jupyter Notebook.

ABOUT K-MEANS CLUSTERING :

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

Working of K-Means Algorithm :

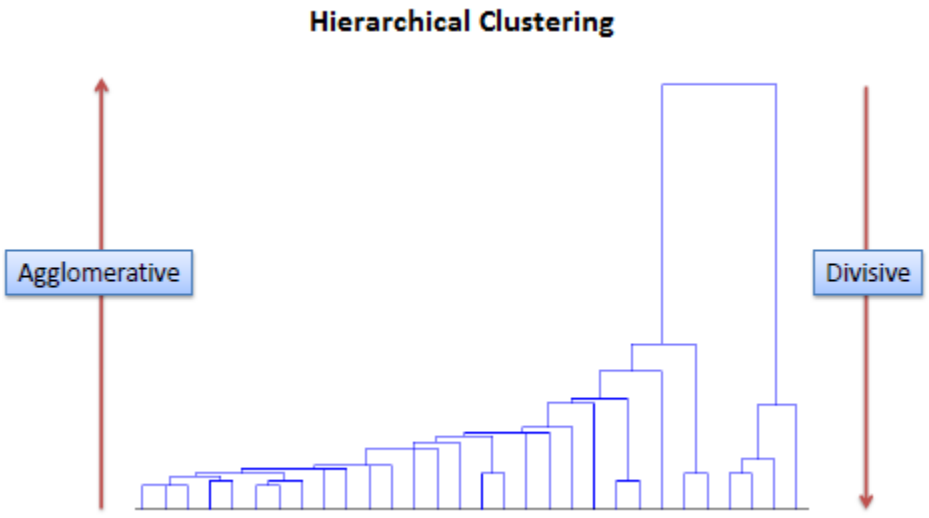
The following stages will help us understand how the K-Means clustering technique works-

- *Step 1:* First, we need to provide the number of clusters, K, that need to be generated by this algorithm.
- *Step 2:* Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
- *Step 3:* The cluster centroids will now be computed.
- *Step 4:* Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.

- 4.1 The sum of squared distances between data points and centroids would be calculated first.
- 4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
- 4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.

ABOUT HIERARCHICAL CLUSTERING :

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, *Divisive* and *Agglomerative*.



Divisive method

In the divisive or *top-down clustering* method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

Agglomerative method

In *agglomerative* or *bottom-up clustering* methods we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left. The related algorithm is shown below. Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between each cluster is measured.

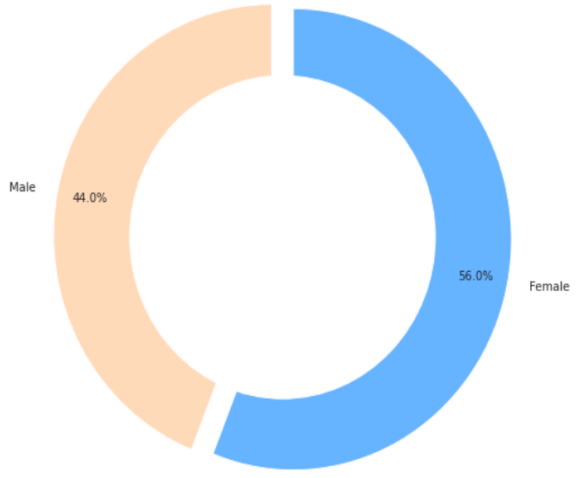
BLUEPRINT OF SOLVING THE PROBLEM STATEMENT :

- Creating an approach to solve the given problem statement
- Exploring the dataset and obtaining useful insight from the same
- Cleaning the dataset by handling nan values, removing duplicate records, etc.
- Data Visualization used to obtain important information from the data

● Data Preprocessing is performed to make the data ready to fit the model this includes feature scaling, splitting the dataset into features and labels, etc.

● Model Building

IMPLEMENTATION



Gender Plot Analysis :

From the above plot, it is shown that the number of Female customers is more than the total number of Male customers.

```
plt.figure(figsize=(20,7))
gender = ['Male', 'Female']
for i in gender:
    plt.scatter(x='Annual Income (k$)',y='Spending Score (1-100)', data=data[data['Gender']==i],s = 200 , alpha = 0.5 , label
= i)
plt.legend()
plt.xlabel("Annual Income (k$)")
plt.ylabel("Spending Score (1-100)")
plt.title("Spending Score according to Annual Income")
plt.show()
```

Annual Income VS Spending Score

Scatter Plot Analysis :

We observe that there are 5 clusters and can be categorized as:

a. High Income, High Spending Score (Top Right Cluster)

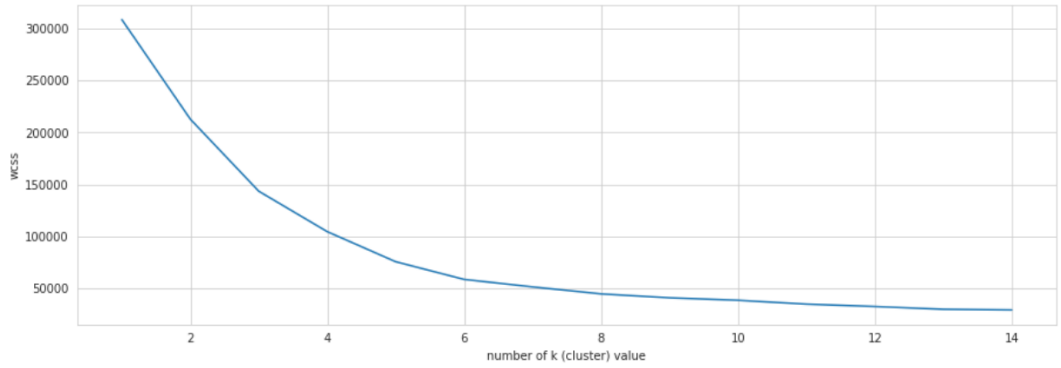
b. High Income, Low Spending Score (Bottom Right Cluster)

c. Average Income, Average Spending Score (Center Cluster)

d. Low Income, High Spending Score (Top Left Cluster)

e. Low Income, Low Spending Score (Bottom Left Cluster)

Elbow method :



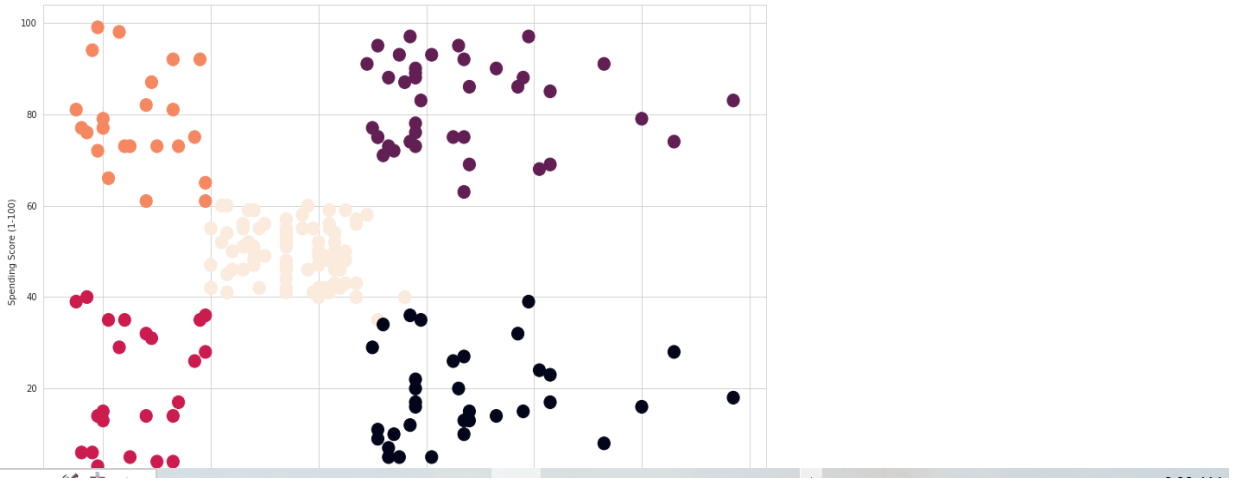
From the elbow method it is clearly shown that the value of k is 5.

CLUSTERING ANALYSIS

K-MEANS CLUSTERING ALGORITHM :

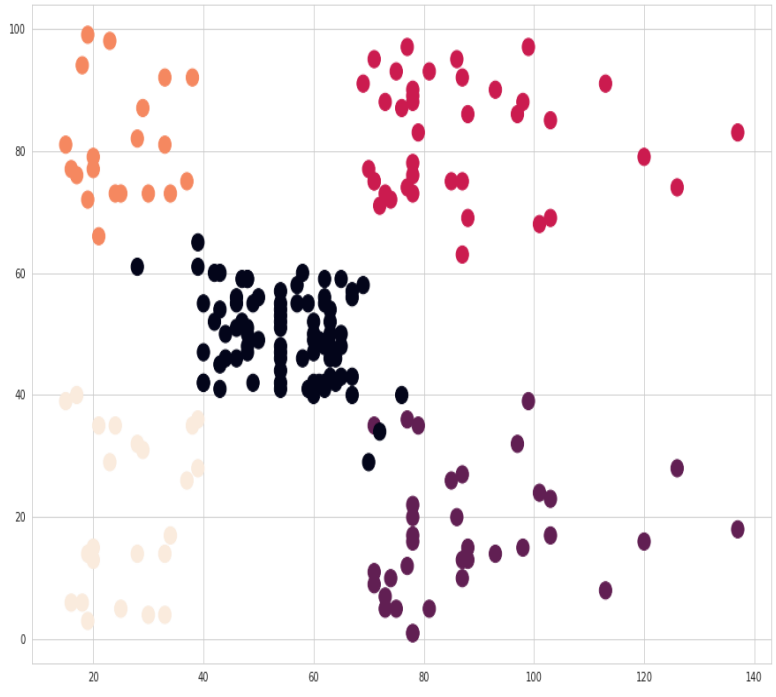
```
#create model
kmeans = KMeans(n_clusters=5)
data_predict = kmeans.fit_predict(data_model)

plt.figure(figsize=(15,10))
plt.scatter( x = 'Annual Income (k$)' ,y = 'Spending Score (1-100)' , data = data_model , c = data_predict , s = 200 )
plt.xlabel("Annual Income (k$)")
plt.ylabel("Spending Score (1-100)")
plt.show()
```



HIERARCHICAL CLUSTERING ALGORITHM :

```
#create model
hiyerartical_cluster = AgglomerativeClustering(n_clusters = 5,affinity= "euclidean",linkage = "ward")
data_predict = hiyerartical_cluster.fit_predict(data_model)
plt.figure(figsize=(15,10))
plt.scatter( x = 'Annual Income (k$)' ,y = 'Spending Score (1-100)' , data = data_model , c = data_predict , s = 200 )
plt.show()
```



From both the algorithms, We can clearly see that there are 5 different categories of customers based on their demography.

CONCLUSION

In this machine learning project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means and hierarchical clustering. We analyzed and visualized the data and then proceeded to implement our algorithm.

REFERENCES

1. https://data-flair.training/blogs/r-data-science-project-customer-segmentation/
2. https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
3. https://scikit-learn.org/stable/
4. https://ieeexplore.ieee.org/document/9752447