# MALL-CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

SUJITH NAARAYAN HIRENDRA BABU(002190143)
2nd Semester, MS Data Science, Khoury College of Computer Sciences,
hirendrababu.s@northeastern.edu

## 1. ABSTRACT :

In today's world, all businesses are run based on the needs of the Customer. There are different types of customers with different needs and it is very difficult to find out who is the potential customer. Customer Segmentation is the process by which customers are separated into groups based on common characteristics that reflect the similarity of each individual in the group. Customer segmentation is performed to determine how to relate customers in each segment in order to maximize the value of each customer to the enterprise. To find out the target customer segment, we use the customer segmentation method by applying Clustering Technique.

## 2. INTRODUCTION

Clustering is one of the most common ways to explore data and to get a clear understanding of its structure. This can be characterized as a task of finding subgroups in a complete dataset. Similar data will be grouped into the same subgroup. A cluster refers to a collection of aggregated data points because of some similarities. Clustering is used in market basket analysis to segment customers based on customer behaviour.

K means clustering algorithm is applied which partitions the dataset into 'k' different number of clusters that are non overlapping and each data point belongs to only one sub group which has similar properties. It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid.

The algorithm takes raw unlabelled data as input and divides the dataset into clusters and the process is repeated until the best clusters are found. We use PCA , a linear dimensionality reduction technique to reduce the date dimensionality by maximizing the variance. t-SNE, a non-linear technique is used for exploration.

## 3. DATA SET :

The dataset (Mall_customers.csv) has been downloaded from Kaggle repository. The dataset contains data about customers from a Mall. The dataset contains 200 customers and five variables which includes : 1) Customer ID, 2) Customer Gender 3) Customer Age 4) Annual Income of the customer (in Thousand Dollars) 5) Spending score of the customer (1 – 10)

## 4. METHODOLOGY :

### 4.1 DATA EXPLORATION:

We analyse the data by calling the *describe* method, which gives us the min, max, mean, median, mode and standard deviation of the dataset columns.

Table 1,: Table showing the Min,Mean, Median, Std Dvn, Max

| | Count | Mean | Std.Dvn | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Age | 200.000000 | 38.850000 | 13.969007 | 18.000000 | 28.750000 | 36.000000 | 49.000000 | 70.000000 |
| Annual Income (K$) | 200.000000 | 60.560000 | 26.264721 | 15.000000 | 41.500000 | 61.500000 | 78.000000 | 137.000000 |
| Spending Score (1-100) | 200.000000 | 50.200000 | 25.823522 | 1.000000 | 34.750000 | 50.000000 | 73.000000 | 99.000000 |

We see that the mean of age, annual income and spending score of the customers is 38.85, 60.56 and 50.20.

**4.2 VARIABLE CORRELATIONS** :The relationship among the different variables age, annual income and spending scores in the dataset is examined by generating the pair plots using the pairplot seaborn function.

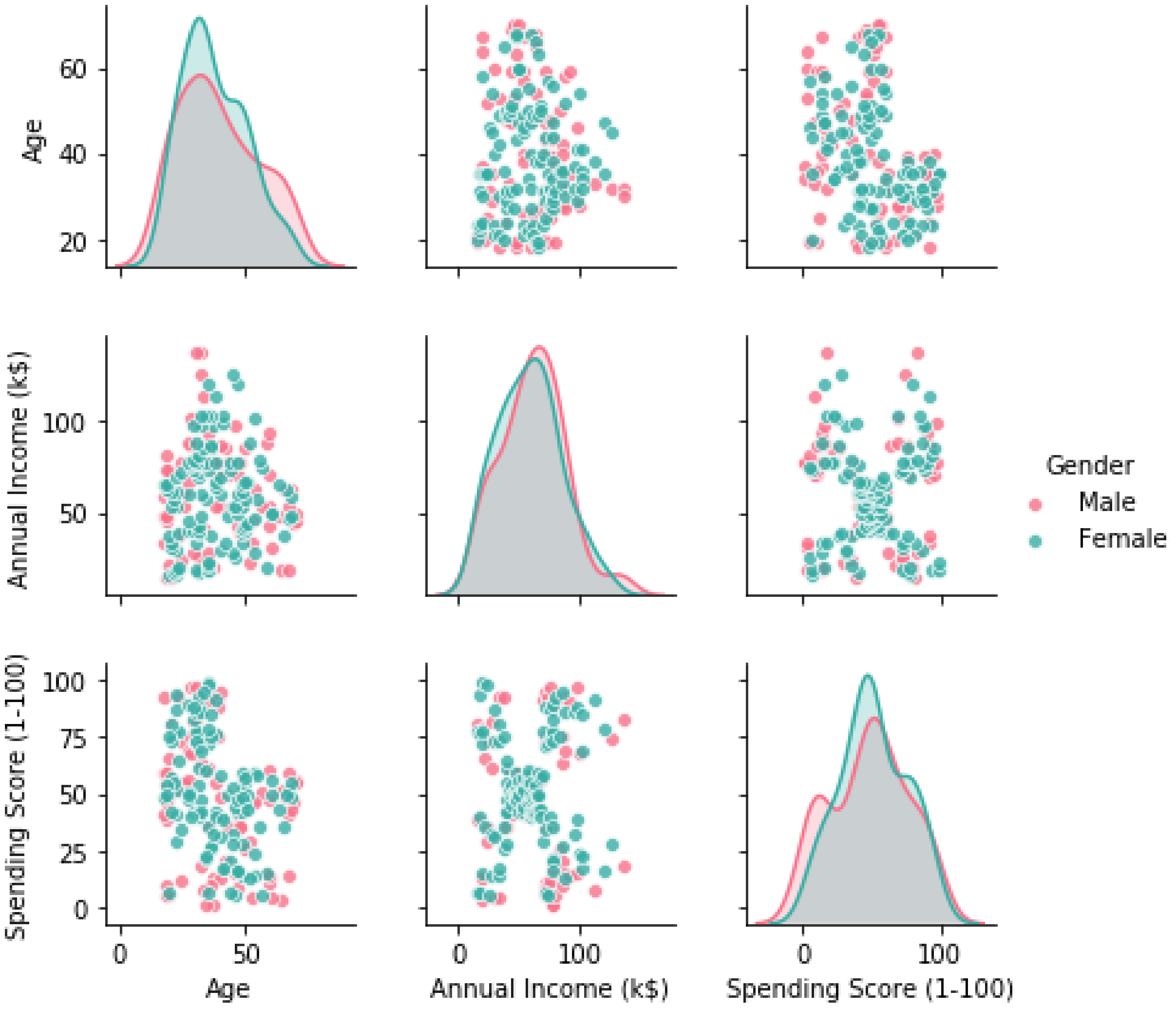


Figure 1 Pair Plots for variable correlations

From the above plot, it is observed that females play an important role in shopping rather than their male counterparts with respect to age, annual income and spending score i.e. 56% of our customers are female and the rest are male.

### 4.3 DATA PRE-PROCESSING :

The variable '*Gender*' is categorical, hence it is replaced with Numerical values. We dropped '*Gender_female'*, because, if '*Gender_male*' is 0, then it is inferred that '*Gender_female'* is 1

**Standardization :**
All the variables in the dataset are standardized in order to have them around the same scale and the output generated.

### 4.4 K-MEANS CLUSTERING :

We first need to decide on the number of clusters we want to create in our dataset. We do so by using the Silhouette and Elbow methods.

We select k centers for each cluster. We do it randomly, pass certain points that we believe are the center or place them in a smart way (e.g. as far away from each other as possible). Then, we calculate the Euclidean distance between each point and the cluster centers. We assign the points to the cluster center where the distance is minimum. After that, we recalculate the new cluster center. We select the point that is in the middle of each cluster as the new center. And we start again, calculate distance, assign to cluster, calculate new centers till the centers do not move anymore.

#### 4.4.1 ELBOW METHOD :

The Elbow method is used to find the optimum number of clusters. The value of the cost function of different values of 'k' is plotted in the elbow method. This is done by ranging k from 1 to 10 clusters. We use the WCSS (Within Cluster Sum of Squares ) as our cost function. The formula for WCSS is given below.

$$WCSS = \sum_{C_k}^{C_n} \left( \sum_{d_1 \, in \, C_1}^{d_m} distance(d_i, C_k)^2 \right)$$

where C is the Cluster Centroid and d is the data point in each cluster.

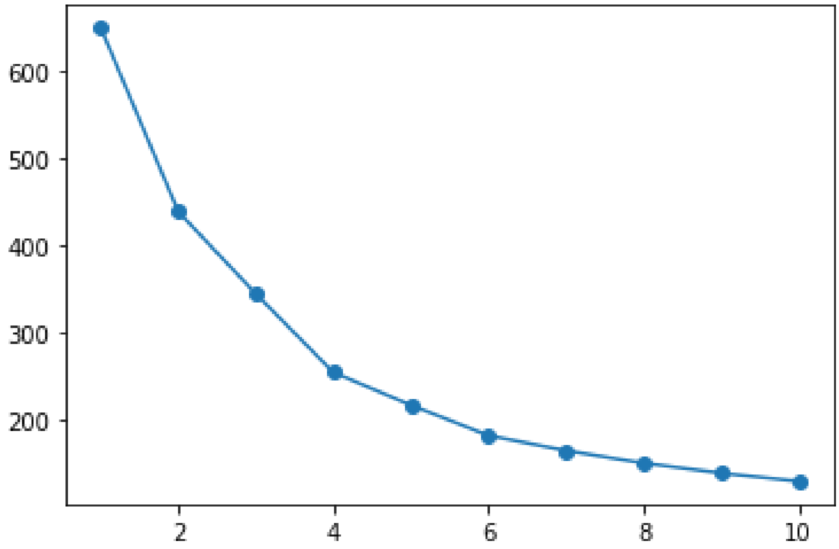


Figure 2 Elbow Plot

The plot denotes the optimum number of clusters required in our model, which is 5. The optimum clusters can be found from the graph where there is a bend in the graph.

**4.4.2 Silhoutte Analysis :** The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The silhouette score is an evaluation metric which is used evaluate the quality of clusters created and to find the optimal number of clusters. A silhouette score of +1 indicates good clustering and -1 indicates poor clustering.

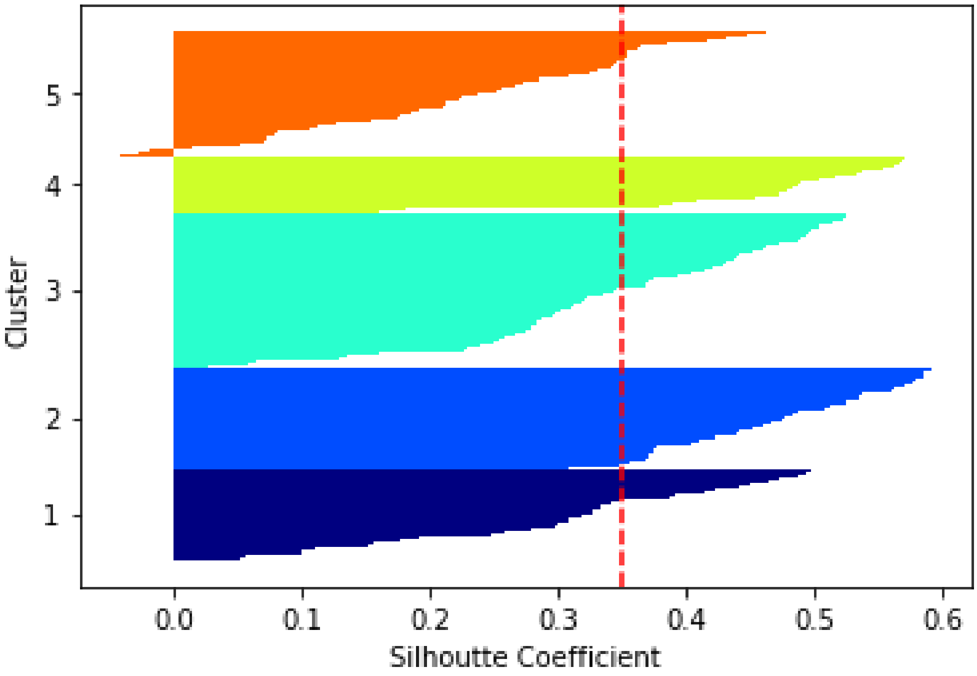
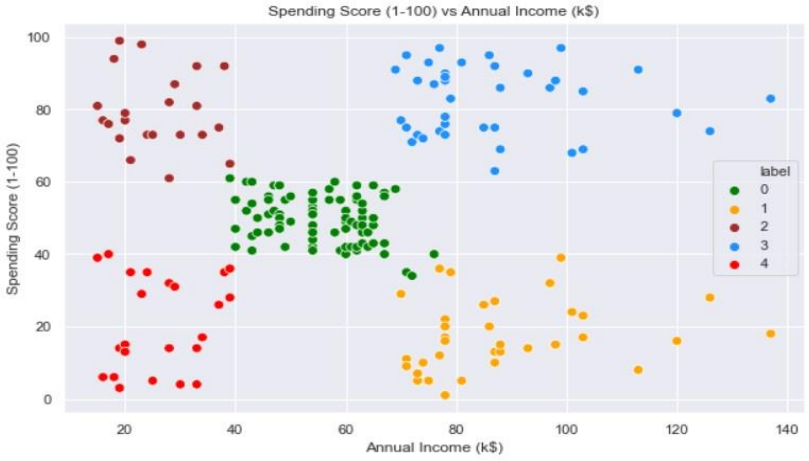


Figure 3 Silhoutte Plot

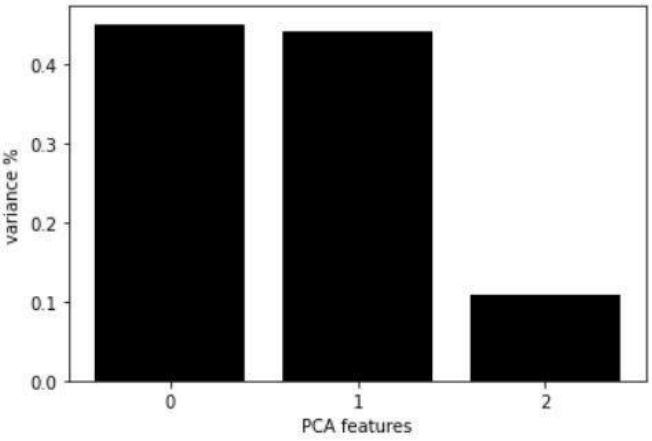
From the above, we know that the optimal number of clusters is 5 and the silhouette score is 0.35.

Now we perform k means clustering with 5 clusters. A new dataframe is created with original features and a new column with the assigned clusters is added and we get the following plot.



Cluster 0 Green Colour) represents average in terms of earning and also spending
Cluster 1 (cyan Colour) represents high earning but spending less
Cluster 2(magenta Colour) represents earning less, but spending high
Cluster 3(Blue Colour) represents high earning and spending is also more(Target Customers)
Cluster 4 (Red Colour) represents less earning and spending less

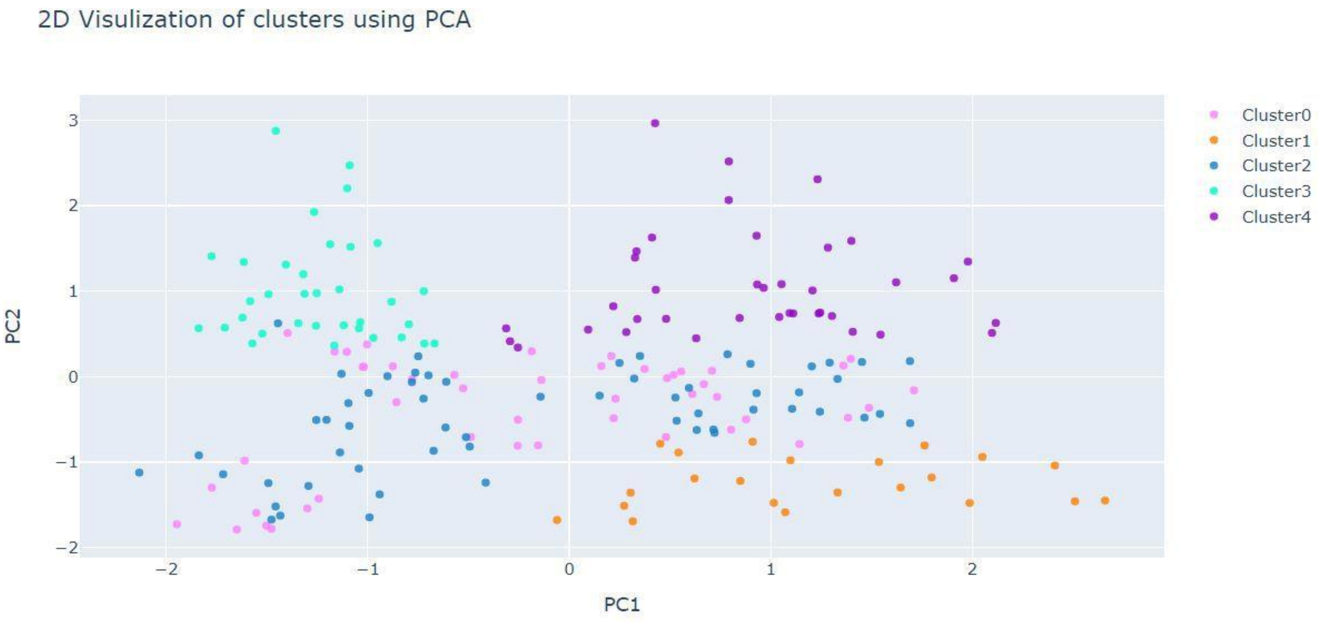
**4.5 PRINCIPAL COMPONENT ANALYSIS** :
We use the Dimensionality Reduction Technique PCA to see which dimensions best maximize the variance of features involved. When PCA is run on a dataframe new components are created and these components explain the maximum variance in the model

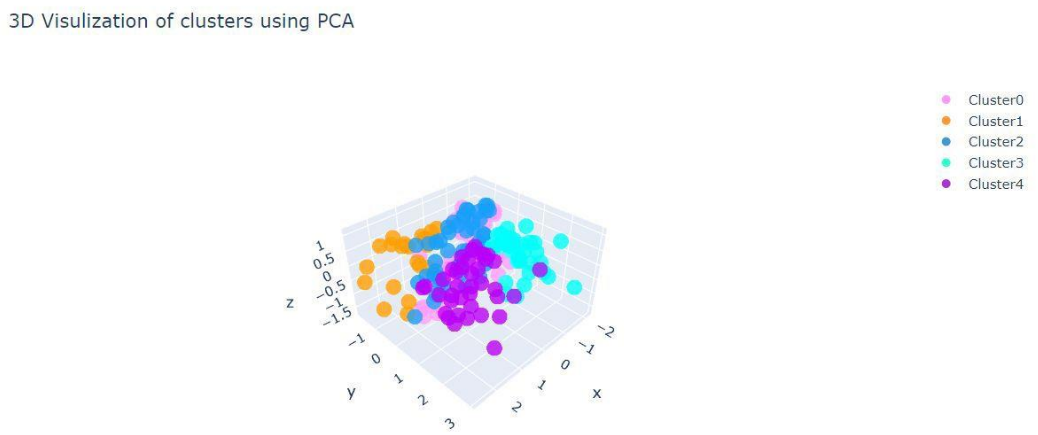


The above chart shows the PCA components along with its variance. In the PCA space, the variance is maximized along PC1 and PC2. The first three Principal components explains about 90% of the dataset variance. Hence, we have fed these three components and built the model and the output is given below :

**Table 2 : Table with values of three PCs**

| | Scaled_Age | Scaled_AnnualIncome | Scaled_SpendingScore | Gender_Male | Cluster | PC1 | PC2 | PC3 |
|---|---|---|---|---|---|---|---|---|
| 1 | -1.424569 | -1.738999 | -0.434801 | 1 | 2 | -1.649886 | -1.789124 | 0.067789 |
| 2 | -1.281035 | -1.738999 | 1.195704 | 1 | 0 | 0.315381 | -1.692772 | 2.179818 |
| 3 | -1.352802 | -1.700830 | -1.715913 | 0 | 2 | -1.477989 | -1.781317 | 0.078518 |
| 4 | -1.137502 | -1.700830 | 1.040418 | 0 | 0 | -0.060545 | -1.678691 | 0.688083 |
| 5 | -0.563369 | -1.662660 | -0.395980 | 0 | 2 | -1.502961 | -1.743807 | 0.156516 |





### 4.6 t-SNE :

We have also used t-SNE for data visualization and plotted the 2D and 3D Plots.

## 5. RESULT :

In the above Mall customer segmentation Project, we found that the most important features are the  Annual Income and Spending score followed by the Age. We segmented our customers into 5 groups. The first segment, Segment 0 has customers whose income is middle and spend in the same range. Segment 1 has customers  whose earnings are high but they spend less. Segment 2 is where we have customers whose income is low range but spend high. Segment 3 has  customers whose income is very high and their spending is also equal. Finally, we have Segment 4 customers whose earnings are little and spend also in the same range.  From these segments, the customers could be understood better and used to increase the revenue of the company.

The K means clustering solution was able to distinguish the segments. Using PCA we can reduce the number of variables and the three PC's contributed for 90% of the variance. In order to analyze in depth we can also use other dimensionality reduction techniques.  t-SNE is used for visualization using 2D and 3D plots.

**LIMITATIONS:** Though K means is the one of the best methods for customer Segmentation, but it has its disadvantages.  We have to specify the number of clusters in the beginning and it can deal only with numerical data.

## 6. FUTURE WORK :

Use a hybrid approach to study and analyze customer segmentation and clustering in depth using different algorithms like GMM, DBSCAN, Hierarchical Clustering and Mini Batch KMeans.

## 7. REFERENCES:

1)D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation” Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp. 245-249, 2009.
2) T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.