The Battle of Neighborhoods

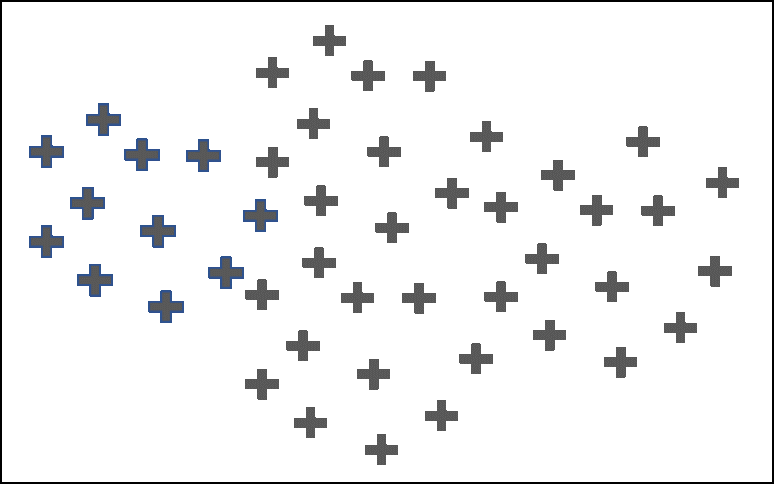
<http://go.microsoft.com/fwlink/p/?LinkId=255141>

**1. Abstract**

This project deals with the customers data of mall. This article demonstrates the concept of segmentation of a customer [data set](https://github.com/sowmyacr/kmeans_cluster/blob/master/CLV.csv) from an e-commerce site using K\_means clustering in python. The data set contains the **annual income** of ~300 customers and their **annual spend** on an e-commerce site. We will use the K\_means clustering algorithm to derive the optimum number of clusters and understand the underlying customer segments based on the data provided. In today’s competitive world, it is crucial to understand customer behavior and categorize customers based on their demography and buying behavior. This is a critical aspect of customer segmentation that allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies.

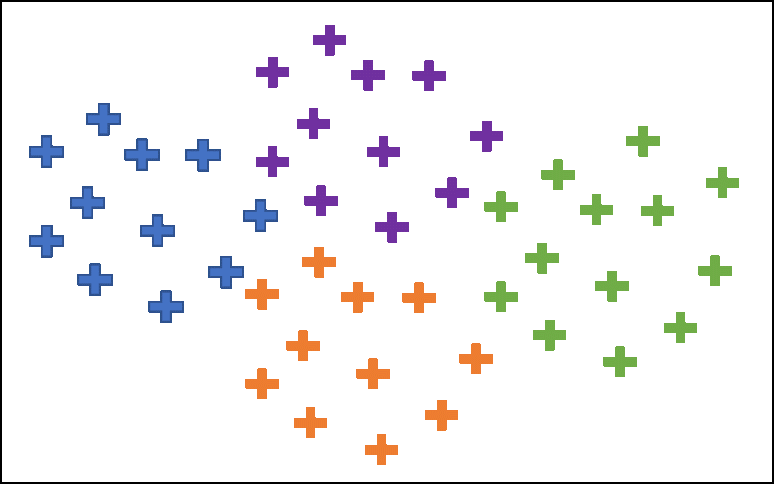
**2. Intorduction**

Clustering is an unsupervised machine learning technique, where there are no defined dependent and independent variables. The patterns in the data are used to identify / group similar observations.



Original Dataset

Fig-1



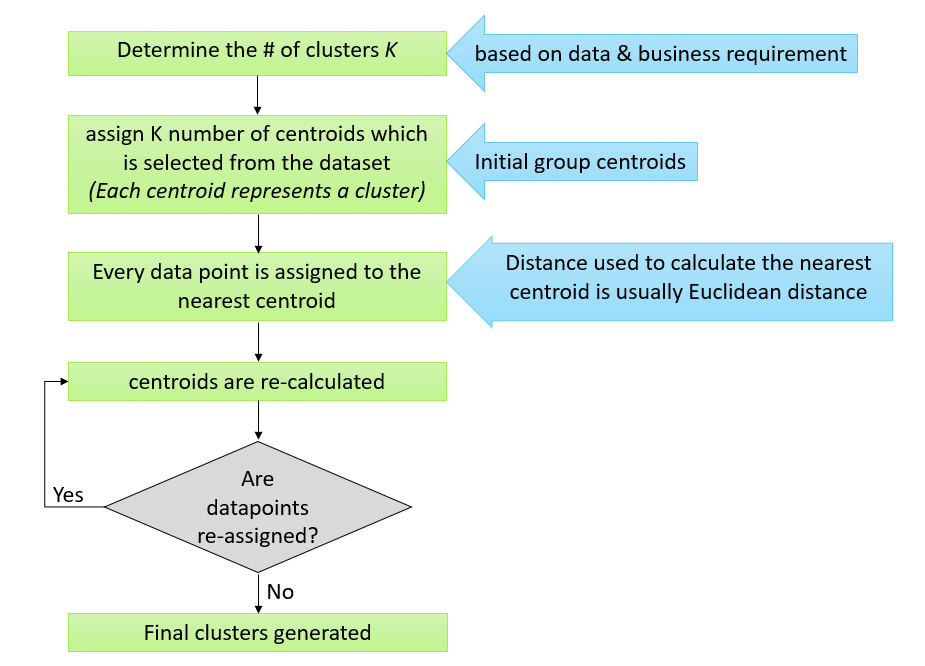
After Clustering

Fig-2

The objective of any clustering algorithm is to ensure that the distance between datapoints in a cluster is very low compared to the distance between 2 clusters. In other words, members of a group are very similar, and members of different groups are extremely dissimilar.

**3. K-Means clustering**

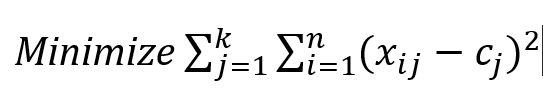
K-means clustering is an iterative clustering algorithm where the number of clusters K is predetermined and the algorithm iteratively assigns each data point to one of the K clusters based on the feature similarity.

**Broad steps of the k-means algorithm-** 

**Fig-3**

**4. Methadology of clustering**

**-** The mathematics behind clustering, in very simple terms involves minimizing the sum of square of distances between the cluster centroid and its associated data points:



* *K* = number of clusters
* *N*= number of data points
* *C*=centroid of cluster j
* (*xij — cj*)– Distance between data point and centroid to which it is assigned

**5.Deciding on the optimum number of clusters ‘K’**

The main input for k-means clustering is the number of clusters. This is derived using the concept of **minimizing within cluster sum of square (WCSS)**. A scree plot is created which plots the number of clusters in the X axis and the WCSS for each cluster number in the y-axis

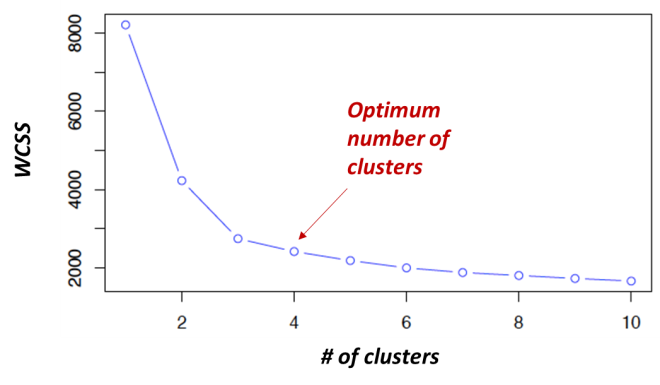


Fig-4

As the number of clusters increase, the WCSS keeps decreasing. The decrease of WCSS is initially steep and then the rate of decrease slows down resulting in an elbow plot. The number of clusters at the elbow formation usually gives an indication on the optimum number of clusters. This combined with specific knowledge of the business requirement should be used to decide on the optimum number of clusters.

**6.Implementation**

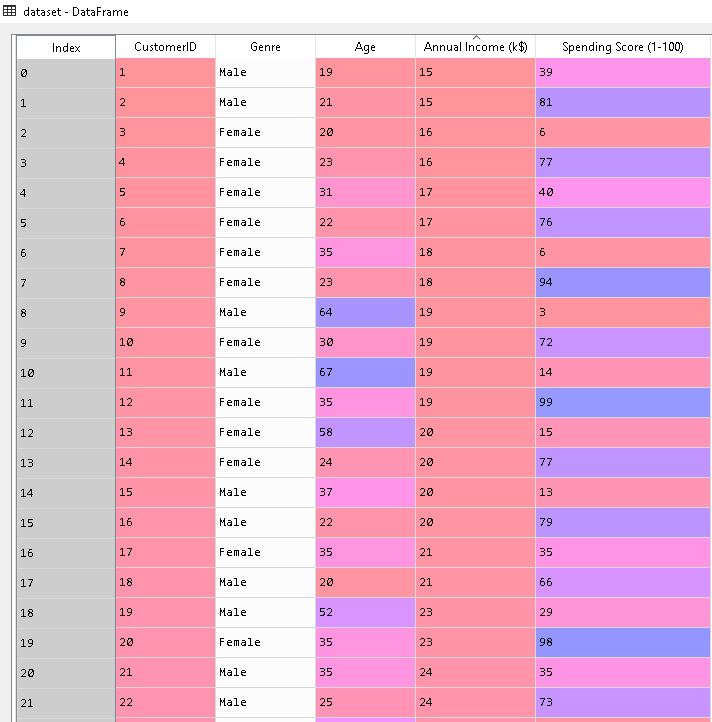
Dataset-

Fig-5

**Obtained Elbow plot(using Wcss Method)**

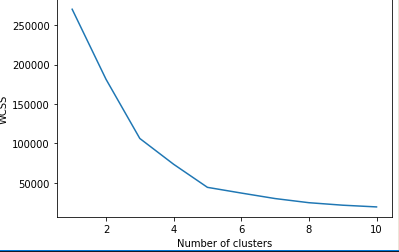


Fig-6(Elbow plot)

We clearly obtained from the elbow plot we can have 5 clusters.i.e k=5

**Obtained graph for cluster:-**

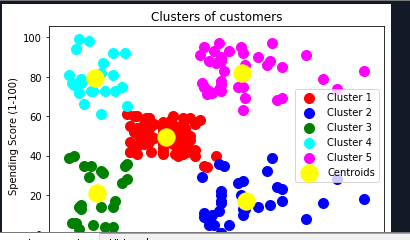


Fig-7

Setting the number of clusters to 5 seems to provide a more meaningful customer segmentation.

1. Cluster 1: Medium income, low annual spend
2. Cluster 2: Low income, low annual spend
3. Cluster 3: High income, high annual spend
4. Cluster 4: Low income, high annual spend
5. Cluster 5: Medium income, low annual spend

**Conclusion-** We have thus seen, how we could arrive at meaningful insights and recommendations by using clustering algorithms to generate customer segments. For the sake of simplicity, the dataset used only 2 variables — income and spend. In a typical business scenario, there could be several variables which could possibly generate much more realistic and business-specific insights.