Customer Segmentation

Chandrashekhar Awate

Data science Trainee

Almabetter,Banglore

Problem Statement: In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Description

Attribute Information:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Exploratory Data Analysis

Exlporatory Data analysis was done to know the customers from the viewpoint of there purchase in terms of product category, price of product etc.

Null values

There were some null values in Description, CustomerID,but we removed it as we have large dataset and null values are few only.

Fitting different models

1)K means Clustering

2) DBSCAN

3) Hierarchical clustering

<u>RFM model (Recency, Frequency,Monetary value)</u>

Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors: . Frequency: How often a customer makes a purchase. Monetary Value: How much money a customer spends on Performing RFM Segmentation and RFM Analysis, Step by Step The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer. ... The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M), using Excel or another tool.

<u>Calculating RFM scores</u>

The number is typically 3 or 5. If you decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 333 to a low of 111. Generally speaking, the higher the RFM score, the more valuable the customer.

<u>Algorithms</u>

1)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.

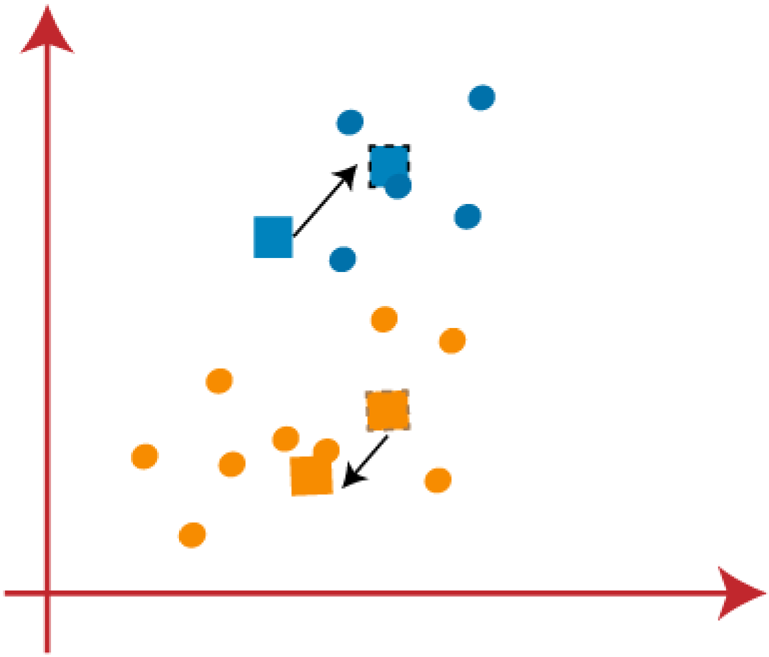
In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

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

K-means implements the Expectation-Maximization strategy to solve the problem. The Expectation-step is used to assign data points to the nearest cluster, and the Maximization-step is used to compute the centroid of each cluster.



2)DBSCAN

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

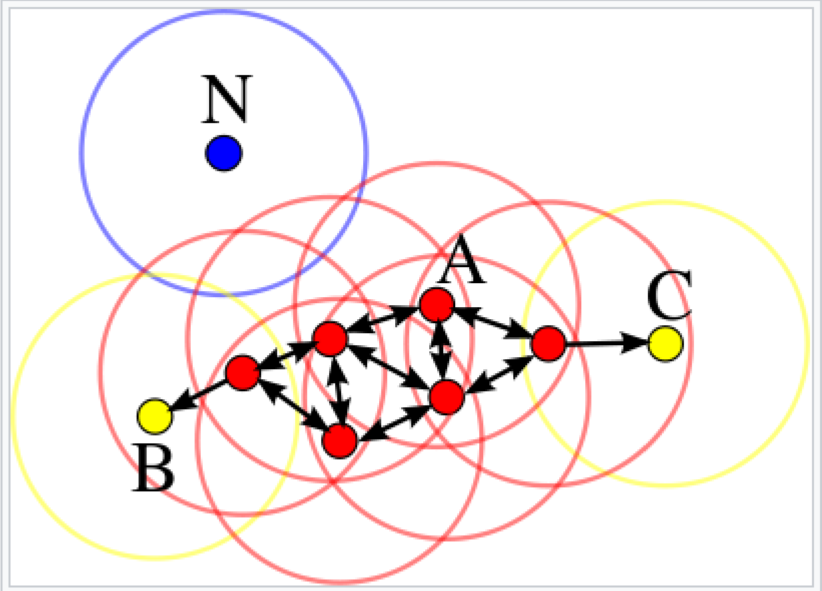
There are two key parameters of DBSCAN:

- **eps: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.**
- **minPts: Minimum number of data points to define a cluster.**

Based on these two parameters, points are classified as core point, border point, or outlier:

- Core point: A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps.
- Border point: A point is a border point if it is reachable from a core point and there are less than minPts number of points within its surrounding area.
- Outlier: A point is an outlier if it is not a core point and not reachable from any core points.

These points may be better explained with visualization



In this case, minPts is 4. Red points are core points because there are at least 4 points within their surrounding area with radius eps. This area is shown with the circles in the figure. The yellow points are border points because they are reachable from a core point and have less than 4 points within their neighborhood. Reachable means being in the surrounding area of

a core point. The points B and C have two points (including the point itself) within their neigborhood (i.e. the surrounding area with a radius of eps). Finally N is an outlier because it is not a core point and cannot be reached from a core point.

We have learned the definitions of parameters and different type points. Now we can talk about how the algoritm works. It is actually quite simple:

- **minPts and eps are determined.**
- **A starting point is selected at random at it's neighborhood area is determined using radius eps. If there are at least minPts number of points in the neighborhood, the point is marked as core point and a cluster formation starts. If not, the point is marked as noise. Once a cluster formation starts (let's say cluster A), all the points within the neighborhood of initial point become a part of cluster A. If these new points are also core points, the points that are in the neighborhood of them are also added to cluster A.**

Note: A point that is marked as noise may be revisited and be part of a cluster.

- **Next step is to randomly choose another point among the points that have not been visited in the previous steps. Then same procedure applies.**
- **This process is finished when all points are visited.**

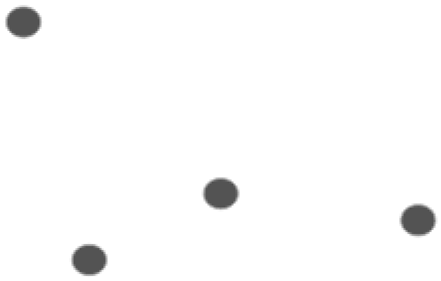
The distance between points is determined using a distance measurement method as in k-means algorithm. The most commonly used method is euclidean distance.

By applying these steps, DBSCAN algorithm is able to find high density regions and separate them from low density regions.

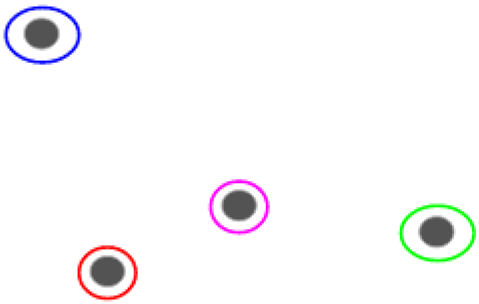
A cluster includes core points that are neighbors (i.e. reachable from one another) and all the border points of these core points. The required condition to form a cluster is to have at least one core point. Although very unlikely, we may have a cluster with only one core point and its border points.

<u>3) Hierarchical Clustering</u>

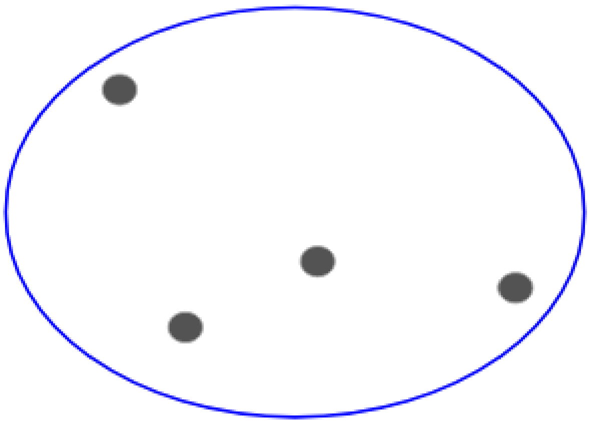
Let's say we have the below points and we want to cluster them into groups:



We can assign each of these points to a separate cluster:



Now, based on the similarity of these clusters, we can combine the most similar clusters together and repeat this process until only a single cluster is left:



We are essentially building a hierarchy of clusters. That's why this algorithm is called hierarchical clustering. I will discuss how to decide the number of clusters in a later section. For now, let's look at the different types of hierarchical clustering.

Types of Hierarchical Clustering

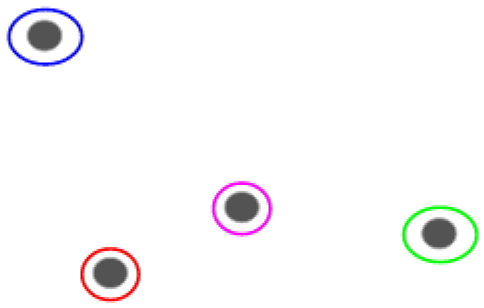
There are mainly two types of hierarchical clustering:

1. Agglomerative hierarchical clustering
2. Divisive Hierarchical clustering

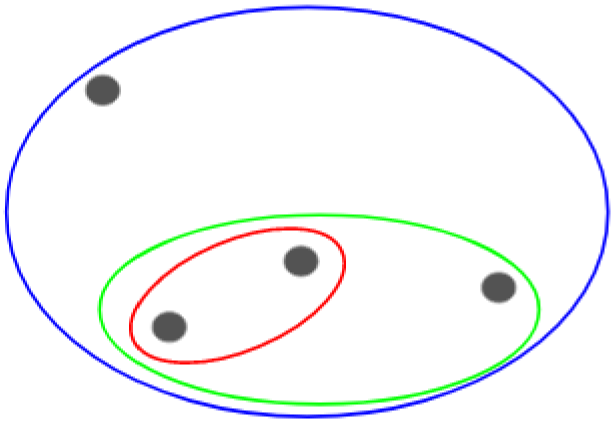
Let's understand each type in detail.

Agglomerative Hierarchical Clustering

We assign each point to an individual cluster in this technique. Suppose there are 4 data points. We will assign each of these points to a cluster and hence will have 4 clusters in the beginning:



Then, at each iteration, we merge the closest pair of clusters and repeat this step until only a single cluster is left:

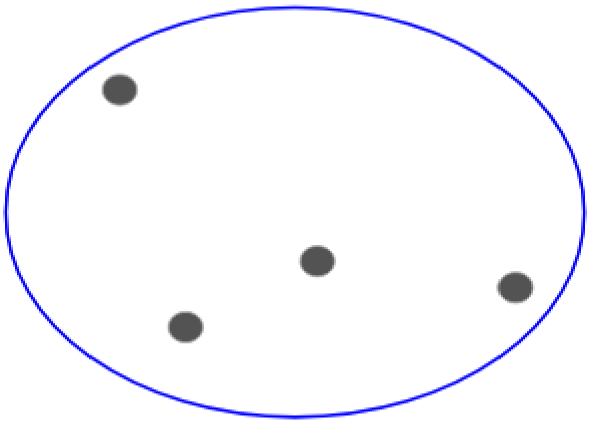


We are merging (or adding) the clusters at each step, right? Hence, this type of clustering is also known as additive hierarchical clustering.

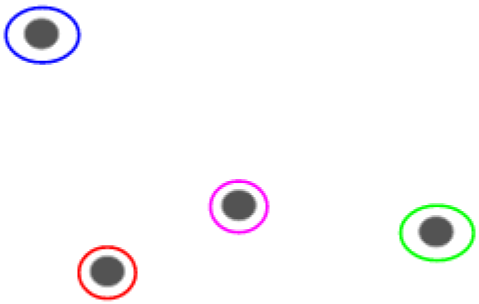
Divisive Hierarchical Clustering

Divisive hierarchical clustering works in the opposite way. Instead of starting with n clusters (in case of n observations), we start with a single cluster and assign all the points to that cluster.

So, it doesn't matter if we have 10 or 1000 data points. All these points will belong to the same cluster at the beginning:



Now, at each iteration, we split the farthest point in the cluster and repeat this process until each cluster only contains a single point:



We are splitting (or dividing) the clusters at each step, hence the name divisive hierarchical clustering.

Agglomerative Clustering is widely used in the industry and that will be the focus in this article. Divisive hierarchical clustering will be a piece of cake once we have a handle on the agglomerative type.

Calculation of Silhouette score

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters: Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance.

The mean distance is denoted by a. Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b. The Silhouette Coefficient for a sample is S=(b−a)max(a,b).

Elbow method

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

Dendogram

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

Conclusions

From Exploratory Data Analysis

1)Top 5 selling products are WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT, ASSORTED COLOUR BIRD ORNAMENT and PARTY BUNTING.

2) Top 5 stock name based on selling

1. **85123A**
2. **22423**
3. **85099B**
4. **84879**
5. **47566**

3) Most of the customers are from unitedkingdom,Germany,france,eire,spain

4) Most of the customer buy things on thursday,Wednesday and Tuesday.

5) most numbers of customers have purches the gifts in the month of November ,October and December September

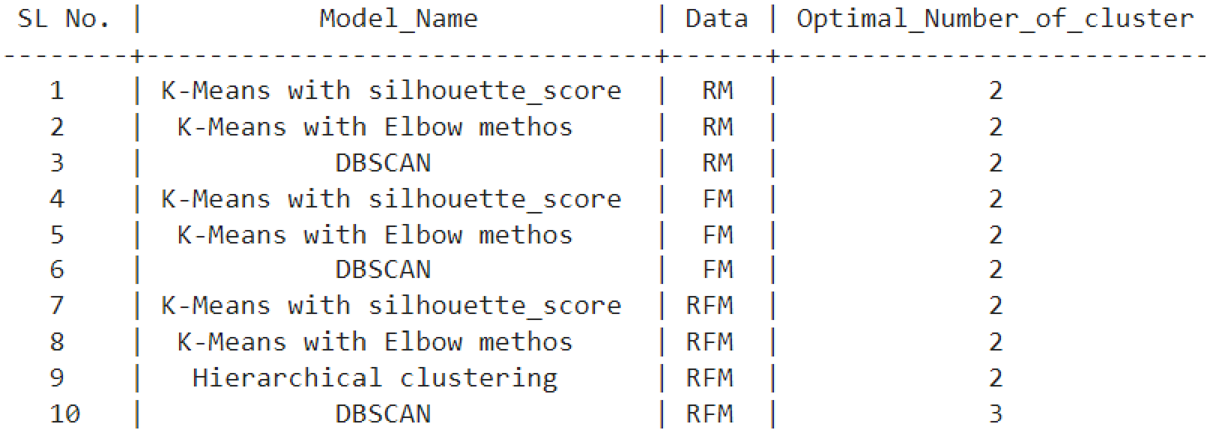
6)Less numbers of customers have purches the gifts in the month of April ,january and February.

7)In AfterNone Time most of the customers have purches the item.

8) Most of the customers have purches the items in Aftrnoon ,moderate numbers of customers have purches the items in Morning and least numbers of customers have purches the items in Evening.

From Models

The Final report after applying RFM



```
 SL No. |            Model_Name            | Data | Optimal_Number_of_cluster
--------+----------------------------------+------+--------------------------
   1    | K-Means with silhouette_score    |  RM  |             2
   2    |  K-Means with Elbow methos       |  RM  |             2
   3    |            DBSCAN                |  RM  |             2
   4    | K-Means with silhouette_score    |  FM  |             2
   5    |  K-Means with Elbow methos       |  FM  |             2
   6    |            DBSCAN                |  FM  |             2
   7    | K-Means with silhouette_score    | RFM  |             2
   8    |  K-Means with Elbow methos       | RFM  |             2
   9    |   Hierarchical clustering        | RFM  |             2
   10   |            DBSCAN                | RFM  |             3
```