Online Retail Customer Segmentation

Almabetter, Bangalore

Neha Pasi

**The need of customer segmentation:-**

The differences in customers' behaviour, demographics, geographies, etc. help in classifying them in groups. Learning about different groups in the customer can help with following:

Target Marketing Client understanding optimal product placement searching for new customers Revenue growth.

## **Problem Description**

### 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.

**Introduction to EDA**:

Exploratory Data Analysis is investigating data and drawing out insights from it to study its main characteristics. EDA can be done using statistical and visualization techniques.

Exploring and analysing the data is important to see how features are contributing to the target variable, identifying anomalies and outliers to treat them lest they affect our model, to study the nature of the features, and be able to perform data cleaning so that our model building process is as efficient as possible.

If we don’t perform exploratory data analysis, we won’t be able to find inconsistent or incomplete data that may pose trends incorrectly to our model.

This step also serves as the basis for answering our business questions.

**Introduction to Unsupervised Machine learning**:

Unsupervised learning, also known as [unsupervised machine learning](https://www.ibm.com/cloud/learn/machine-learning), uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

# Recency-Frequency-Monetary (RFM) model to determine customer value:-

The RFM model is quite useful model in retail customer segmentation where only the data of customer transaction is available. RFM stands for the three dimensions:

Recency – How recently did the customer purchase? Frequency – How often do they purchase? Monetary Value – How much do they spend? A combination of these three attributes can be defined to assign a quantitative value to customers. E.g. A customer who recently bought high value products and transacts regularly is a high value customer

**Data Summary**

We have a data shape of shape (541909, 8)

We have the following column provided to us in the dataset:

### 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.

**Steps involved:**

**Null Values Treatment:** There is some null value present in our dataset like in Customer ID and Description. So, we have dropped them. We have to drop some Invoice No which are starts with 'c' because 'c', it indicates a cancellation

**Exploratory Data Analysis:** After loading the dataset we performed this method by comparing our target variable that is Trip duration with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

**Feature Engineering:** We have Converted InvoiceDate columns into date time format

And created a new features Day from Invoice date.

**Model Fitting:** Fitted different models

1. K- means with silhouette and Elbow method Recency, Monetary

2. k- Means with silhouette and Elbow method Frequency, Monetary

3. K-means with Elbow method on Recency, Frequency, Monetary

4. Hierarchical method

**Algorithms:**

**K- Means Clustering**

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by ‘K’ in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

## Working of K-Means Algorithm

We can understand the working of K-Means clustering algorithm with the help of following steps −

* Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.
* Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.
* Step 3 − Now it will compute the cluster centroids.
* Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more −

First, the sum of squared distance between data points and centroids would be computed.

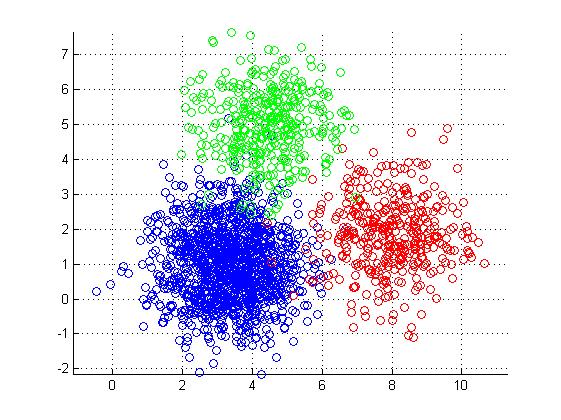
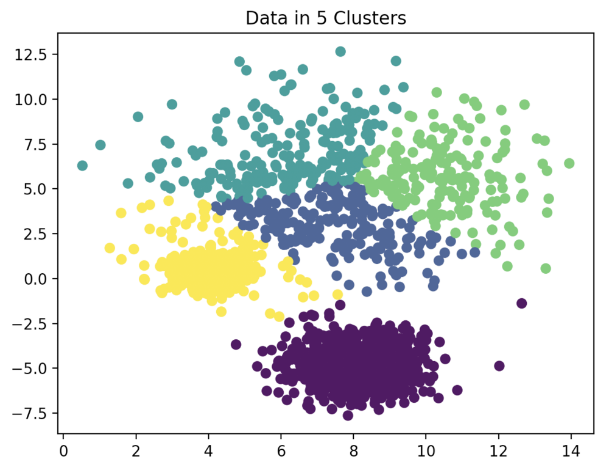
Now, we have to assign each data point to the cluster that is closer than other cluster (centroid).

At last compute the centroids for the clusters by taking the average of all data points of that cluster.

K-means follows Expectation-Maximization approach to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster.

While working with K-means algorithm we need to take care of the following things −

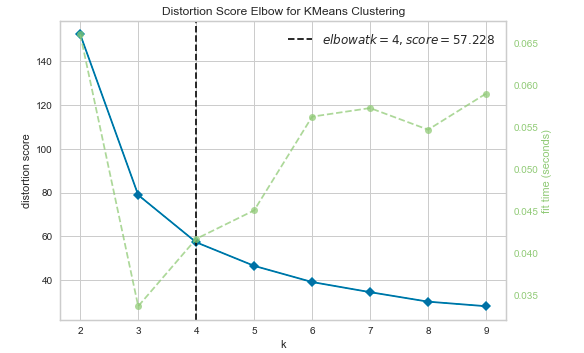
* While working with clustering algorithms including K-Means, it is recommended to standardize the data because such algorithms use distance-based measurement to determine the similarity between data points.
* Due to the iterative nature of K-Means and random initialization of centroids, K-Means may stick in a local optimum and may not converge to global optimum. That is why it is recommended to use different initializations of centroids.

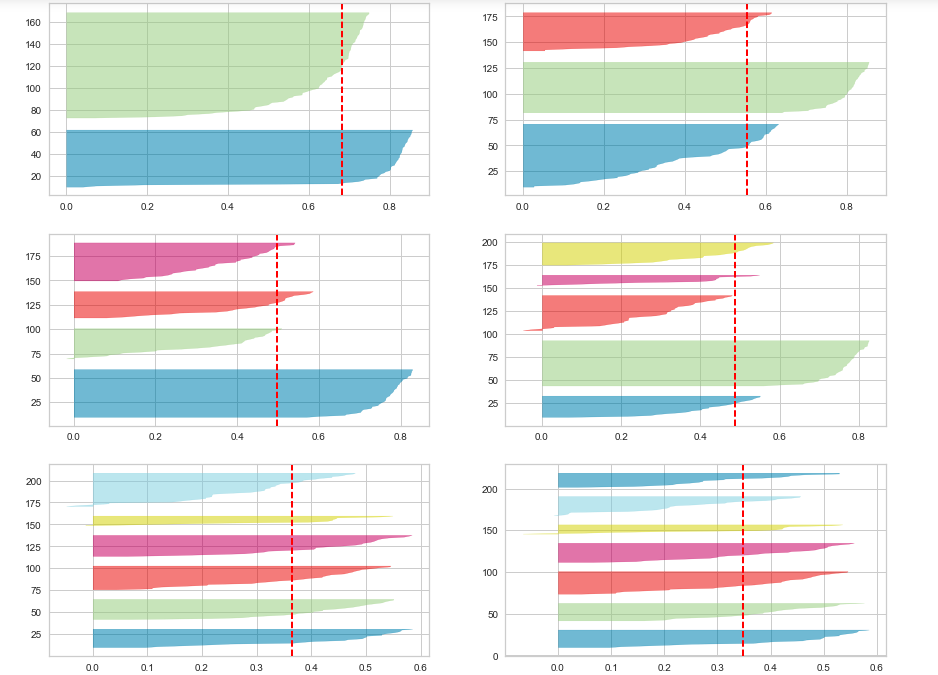
# Elbow Method and Silhouette Score

# In **K-means clustering**, **elbow method** and **silhouette analysis or score** techniques are used to find the number of clusters in a dataset. The elbow method is used to find the “elbow” point, where adding additional data samples does not change cluster membership much. Silhouette score determines whether there are large gaps between each sample and all other samples within the same cluster or across different clusters. In this post, you will learn about these **two different methods to use for finding optimal number of clusters**in **K-means clustering.**Selecting optimal number of clusters is key to applying clustering algorithm to the dataset.

**The Elbow method**is used to find the elbow in the elbow plot. The elbow is found when the dataset becomes flat or linear after applying the cluster analysis algorithm. The elbow plot shows the elbow at the point where the number of clusters starts increasing.



**The silhouette score** of a point measures how close that point lies to its nearest neighbour points, across all clusters. It provides information about clustering quality which can be used to determine whether further refinement by clustering should be performed on the current clustering.



# ****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).

**Hierarchical Clustering Algorithm**

Also called **Hierarchical cluster analysis** or **HCA**is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

For e.g.: All files and folders on our hard disk are organized in a hierarchy.

The algorithm group’s similar objects into groups called **clusters**. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

This clustering technique is divided into two types:

1. Agglomerative Hierarchical Clustering
2. Divisive Hierarchical Clustering

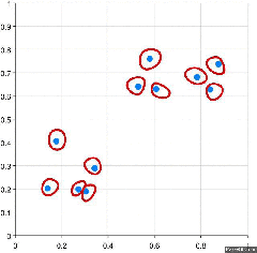
### **Agglomerative Hierarchical Clustering**

The Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting). It's a “[bottom-up](https://en.wikipedia.org/wiki/Top-down_and_bottom-up_design)” approach: **each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.**

**How does it work?**

1. Make each data point a single-point cluster → forms N clusters
2. Take the two closest data points and make them one cluster → forms N-1 clusters
3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
4. Repeat step-3 until you are left with only one cluster.

Have a look at the visual representation of Agglomerative Hierarchical Clustering for better understanding:



[Agglomerative Hierarchical Clustering](https://gfycat.com/somelonelycaterpillar)

There are several ways to measure the distance between clusters in order to decide the rules for clustering, and they are often called Linkage Methods. Some of the common linkage methods are:

* **Complete-linkage**: the distance between two clusters is defined as the longest distance between two points in each cluster.
* **Single-linkage**: the distance between two clusters is defined as the shortest distance between two points in each cluster. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.
* **Average-linkage**: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
* **Centroid-linkage:** finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

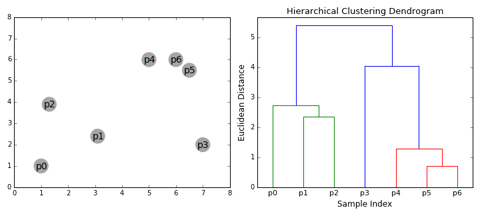
The choice of linkage method entirely depends on you and there is no hard and fast method that will always give you good results. Different linkage methods lead to different clusters.

The point of doing all this is to demonstrate the way hierarchical clustering works, it maintains a memory of how we went through this process and that memory is stored in **Dendrogram**.

**Dendrogram:**

A Dendrogram is a type of tree diagram showing hierarchical relationships between different sets of data.

As already said a Dendrogram contains the memory of hierarchical clustering algorithm, so just by looking at the Dendrgram you can tell how the cluster is formed.



**Conclusion**:

Most numbers of customers have purches in the month of November, October and December September and less numbers of customers have purches in the month of April, January and February

Afternoon Time most of the customers have purchase the item.

Most of the customers have purches the items in Afternoon, moderate numbers of customers have purches the items in Morning and least numbers of customers have purches the items in Evening

By applying different clustering algorithem to our dataset .we get the optimal number of cluster is equal to 2

|  |  |  |  |
| --- | --- | --- | --- |
| SL. No. | Model Name | Data | Optimal Number Of Cluster |
| 1. | K- means with silhouette and Elbow method | RM | 2 |
| 2. | K- means with silhouette and Elbow method | FM | 2 |
| 3. | K- means with Elbow method | RFM | 2 |
| 4. | Hierarchical Method | RFM | 2 |
| 5. | Hierarchical Method | RFM | 2 |