**Customer Segmentation Analysis**

**Gottipati GopiChand,**

**Turab Ahmed.**

**Alma Better.**

**Abstract:**

Customer segmentation is a method of dividing customers into groups or clusters on the basis of common characteristics*.* The market researcher can segment customers into the B2C model using various customer's **demographic** characteristics such as occupation, gender, age, location, and marital status. **Psychographic** characteristics such as social class, lifestyle and personality characteristics and behavioral characteristics such as spending, consumption habits, product/service usage, and previously purchased products. In the B2B model using various company's characteristics such as the size of the company, type of industry, and location.

**1.Problem Statement**

In the Retail sector, the various chain of hypermarkets generating an exceptionally large amount of data. This data is generated daily across the stores. This extensive database of customers transactions needs to analyze for designing profitable strategies.

All customers have different-different kind of needs. With the increase in customer base and transaction, it is not easy to understand the requirement of each customer. Identifying potential customers can improve the marketing campaign, which ultimately increases the sales. Segmentation can play a better role in grouping those customers into various segments. The datasets contain the below fields,

* **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**: Invoice 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**.**

**2. Introduction**

Customer segmentation is a method of dividing customers into groups or clusters on the basis of common characteristics. why do we need customer segmentation?

* It will help in identifying the most potential customers.
* It will help managers to easily communicate with a targeted group of the audience.
* It will help managers to design special offers for targeted customers, to encourage them to buy more products.
* It also helps in identifying new products that customers could be interested in.

*Types of segmentation:*

* Demographic(eg. gender,age,occupation etc.)
* Geographic(eg. location,region,urban/rural etc.)
* Behavioral(eg. spending,consumption habits,previously purchased product etc.)
* psychographic(eg. social status,lifestyle,personality characterstics etc.)

**3. Steps involved**

* ***Exploratory Data Analysis:***

After loading the dataset, we performed EDA to help the retail market to analyze the customer strategy in terms of which product is sold mostly, sales based on monthly, weekly and hourly strategy. By analyzing this strategies, this helps the retail market to give priority during the traffic periods.

* ***Outliers and Null values Treatment:***

Our dataset contains outliers and Null values which might tend to disturb the model in terms of resulting in an error. To overcome from this, we have dropped them.

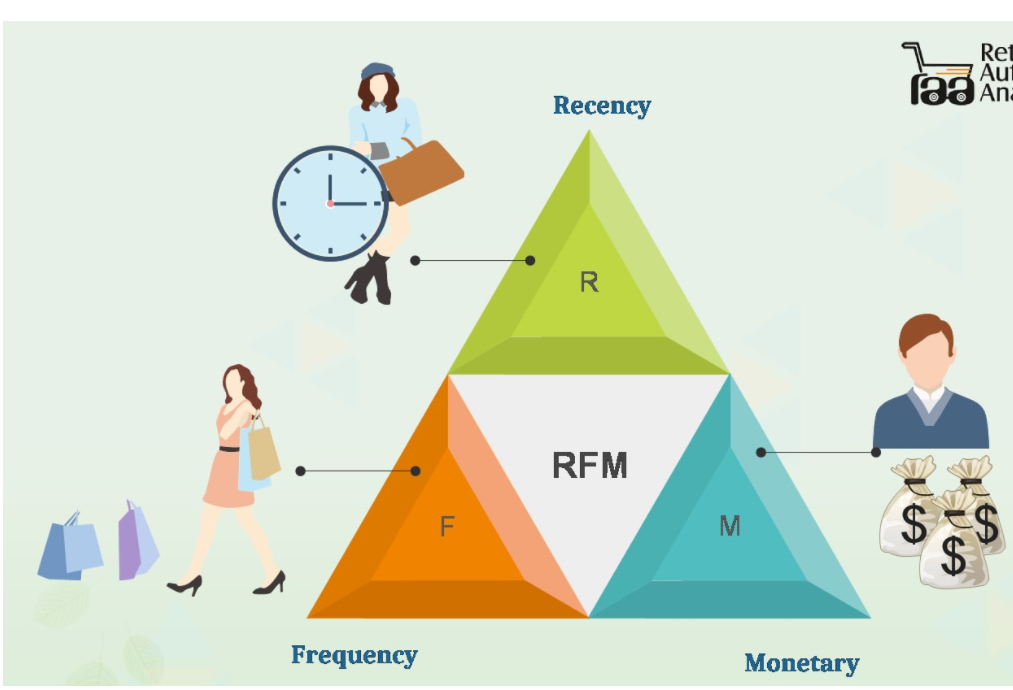
* ***RFM(Recency, Frequency, Monetary):***

We extracted Date from InvoiceDate column which helps us to analyze the RFM metrics.

**Recency:-** This refers to how recently the customer bought our product. This is calculated by subtracting current date with last purchased date by the customer.

**Frequency:-** This refers to how frequently the customers have purchased our products. This is calculated by number of orders purchased by the customer

**Monetary:-** This refers to how much revenue generated through the customer. This is calculated by sum of purchase price for each customer.



* ***Feature Selection:***

Based on RFM values, we have segregated the customers from values 1 to 4. Where 1 represents best whereas 4 represents worst. As the system understands numerical data, this values are converted as numeric datatype

* ***Log Transformation:***

Our dataset are mostly rightly skewed(Positively skewed) when seen through distribution plot. The model requires symmetrical shape to give us a good result, so we have applied ‘Logarithmic’ transformation to our feature dataset which resulted in symmetric shape.

* ***Standardization of features:***

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way, while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

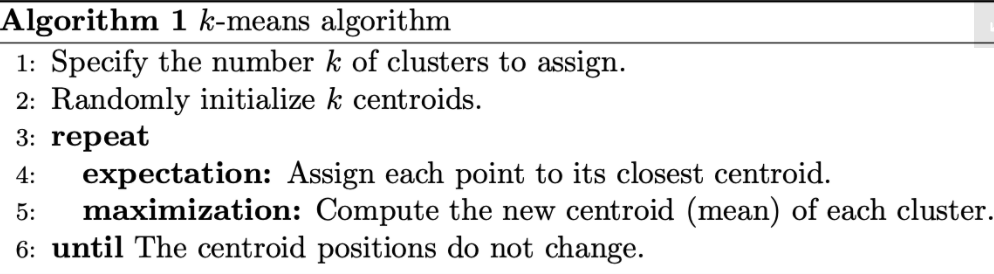
* ***Applying Clusters:***

We make use of Silhoutte score and Elbow method to figure out the optimal number of clusters required to segment the customers.

**4. Algorithms:**

* **K-Means Clustering**

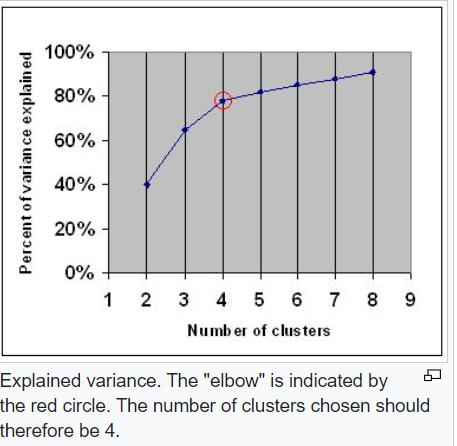
k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition ‘n’ observations into ‘k’ clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster Sentimental Analysis of the user.

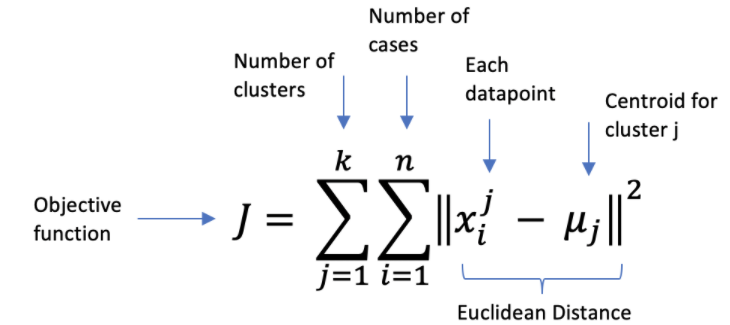


**5.Evaluating the quality of clusters:**

**5.1 Elbow Method:**

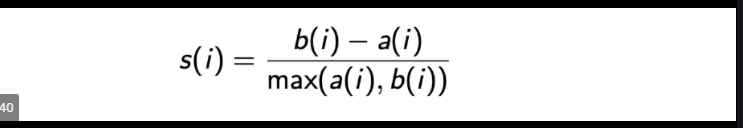
In [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis), the elbow method is a [heuristic](https://en.wikipedia.org/wiki/Heuristic) used in [determining the number of clusters in a data set](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set). The method consists of plotting the [explained variation](https://en.wikipedia.org/wiki/Explained_variation) as a function of the number of clusters, and picking the [elbow of the curve](https://en.wikipedia.org/wiki/Elbow_of_the_curve) as the number of clusters to use.





**5.2 Silhouette Analysis:**

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].



**6. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , Outliers treatment, dropping Null values, feature selection, preprocessing the data, scaler transformation and then model building.

To determine the optimal number of clusters, we applied Elbow method and Silhouette analysis which results in k=4 for optimal number of clusters.

**7.References:**

1.Wikipedia

2.Machine Learning Mastery

3.Analytics Vidya

4.Towards Data Science