**Online Retail Customer Segmentation**

**Sharath Diwakar**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

# Abstract:

Customer segmentation is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. This proposes an integrated approach for determining target customers using predictive model and also discover their associative buying patterns using K-Means algorithm and other methods. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

***Keywords: Online Retail, Clustering, Customer Segmentation, K-Means, Silhouette Score, Elbow Method.***

# Problem Statement:

The 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**

Online retail in the United Kingdom has been gaining ground in the past decade. With the onset of the coronavirus (COVID-19) crisis, the value of online retail sales in the UK is estimated to reach just below 100 billion British pounds in 2020. In the previous year, around 19 percent of retails sales was attributed to online in the UK. According to some survey websites, Statistics gives information on the online share of retail

sales in the United Kingdom in 2011-2019, with a forecast to 2020. Online sales are expected to account for 26.2 percent of the online retail industry in 2020 in the United Kingdom but it depends on several factors which type product you are selling what are your strategies and planning against other retail competitors.

The goal here is to explore the data and apply clustering algorithms to find useful insights from the data and find out customer segmentation based on the products purchased.

# Data Summary:

This dataset has around 541909 observations in it with 8 columns and it is a mix of categorical and numeric values It contains different countries, different products and the behavior of customers. Exploring them will definitely help us have a very good understanding of the online retail trends.

**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:** 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.

# Steps involved:

## Exploratory Data Analysis

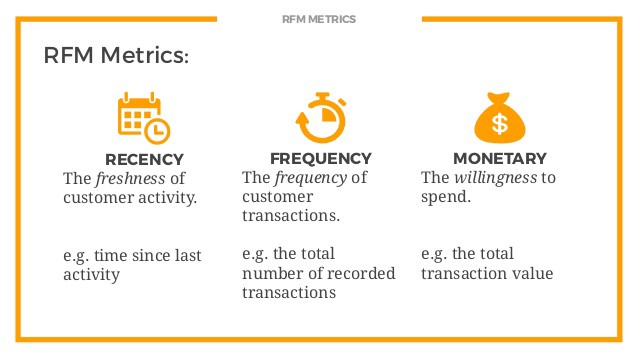
After loading the dataset, we compared our target variable with other independent variables. This process helped us figure out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

## Null values Treatment

The given dataset had large number of null values and those null values were dropped as there was no way to fill those with any other values.

## Data Transformation

In this section, a **Recency, Frequency and Monetary (RFM) analysis** about the data is done. Recency signifies the days since order, frequency signifies the number of times the customer is been billed and monetary signifies the sales each customer has provided.



## Modeling

For Cluster Modeling,

* K-Means

For Optimal Clusters Modeling,

* K-Means with Silhouette Analysis
* K-Means with Elbow Method

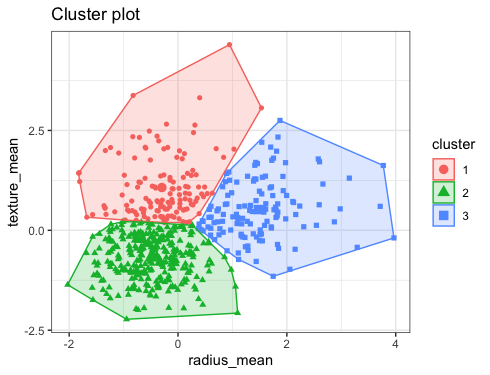
# Algorithms:

## K-Means Clustering

K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

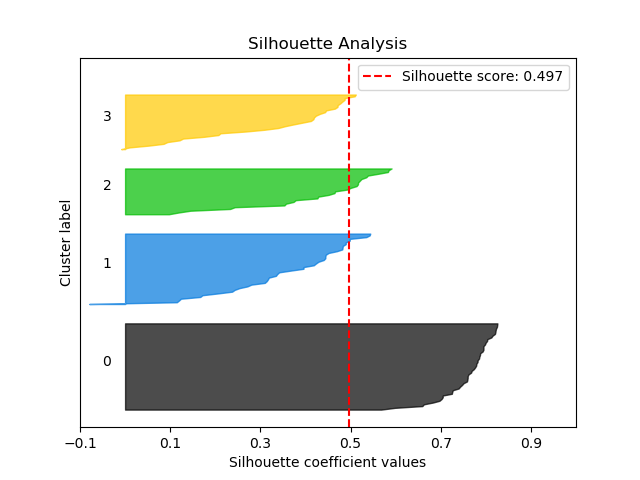
We then proceeded to perform K-means Clustering which will create different clusters to group similar spending activity based on their age and annual income. K Means Clustering selects random values from the data and forms clusters assigned. The closest values from the center of each cluster were taken to update the cluster and reshape the plot (just like k- NN). The closest values are based on Euclidean Distance.



1. **K-Means with Silhouette Analysis**

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

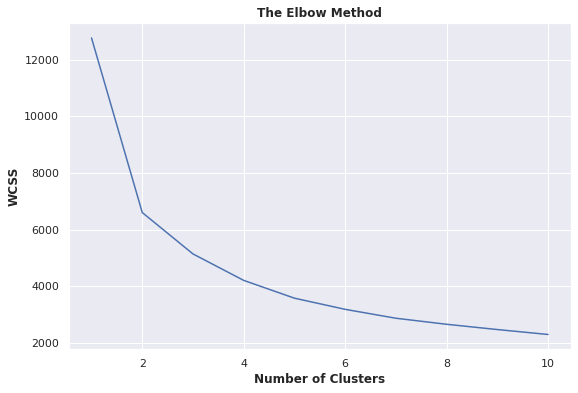
****

1. **K-Means with Elbow Method**

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k. The “arm” can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

We use the Elbow Method which uses Within Cluster Sum of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value.



**Conclusion:**

* Missing and duplicate values were found in the given dataset.
* Most of the purchases are from United Kingdom followed by Germany, France, Ireland and Spain.
* Most of the customers purchased items on Thursday, Wednesday, Tuesday.
* Most of the customers purchased items in the month of November, October, December, and the least number of purchases in April, January, February.
* Most of the customers start their purchase from 10:00 A.M till 2:00 P.M & the 12th hour of the day is the peak for purchasing. After 2:00 P.M the purchasing frequency gradually reduces.
* Top Five purchased products on the basis of their frequency:

1. WHITE HANGING HEART T-LIGHT HOLDER
2. REGENCY CAKESTAND 3 TIER
3. JUMBO BAG RED RETROSPOT
4. ASSORTED COLOUR BIRD ORNAMENT
5. PARTY BUNTING

* **RFM (Recency, Frequency and Monetary) analysis** of the data was carried out to understand the behavior of the customer and the market, as it makes easy to recommend and display newly launched products to the customers and understand their choices.
* Applied Clustering Algorithms:
  1. **K-Means Clustering** - Optimal number of Clusters (3)
  2. **K-Means with Silhouette Analysis** - Optimal number of Clusters (2)
  3. **K-Means with Elbow Method** - Optimal number of Clusters (4)

**References:**

* 1. Stack Overflow
  2. Medium
  3. GeeksforGeeks