**Online Retail Customer Segmentation**

**Pradeep Gupta**

**Data science trainee,**

**Almabetter, Bangalore**

**Abstract:**

A company want to understand the major customer segments on the basis of transactions made by the customers. We were provided with data which contains the details of the transactions made by customers like invoice number, invoice date, customer id and other important features.

Our experiment can help in understanding the customers. On the basis of customers recency, frequency and monetary company can decide the offers, suggestions and other important things to retain the customers.

***Keywords: Recency, Frequency, Monetary***

**1.Problem Statement**

A UK based and registered non-store online retail company mainly sells unique all occasions gifts. Many customers of the company are wholesalers. In this project, our task is to identify major customer segments on a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011. We were provided with data contains details of the transactions

* 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.
* Quantity – The quantities of each product 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**

Companies decides there offers, product suggestions and other import things on the basis of the customers precious orders. A UK based registered non store online retail company mainly sells unique all occasions gifts. Company identify major customer segments on a transactional dataset which contains all the transactions of the customers. Transactional dataset contains many features like invoice number, product name, quantity, stock code invoice date, unit price, customer id and country of the customers. Our goal is to understand the recency, frequency and monetary of the customers, this information can help company to attract the customers to make another transaction and also companies uses this data to reduce the recency of the customers.

**3. Steps involved:**

* **Null and Duplicated Values**

Dataset contains 5268 duplicated values and two features contains null values so I removed those duplicated and null values.

* **Data Cleaning**

In feature Invoice number if the code starts with ‘C’, it indicates a cancellation. I removed the cancelled items from the dataset as thew were only approximate 2% of the data.

* **Feature Engineering**

I performed feature engineering and extracted day name, month name, year, hour and minute from the Invoice date feature. I also created a new feature ‘TotalAmount’ by multiplying unit price and quantity.

* **Exploratory Data Analysis**

After handling null values, data cleaning and feature engineering I performed exploratory data analysis on the features and generated some insights. This helped me in finding the top selling products and worst selling products, also in finding the percentage of transactions in different days and also in different months. I find our maximum customer lives in United Kingdom.

* **Standard Scaler**

Standard scaler performs the task of standardization. Usually, a dataset contains variables that are different in scale. For example an employee dataset will contain age column with values on scale 20-70 and salary column with values on scale 10000-80000. As these two columns are different in scale, they are standardized to have common scale while building machine learning model.

* **Fitting different models**

For modelling I tried various clustring algorithms like:

1. **Kmeans Clustering**
2. **Hierarchical Clustering**

**4.1. Clustering Algorithms:**

1. **Kmeans Clustering:**

K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The below diagram explains the working of the K-means Clustering Algorithm:



.

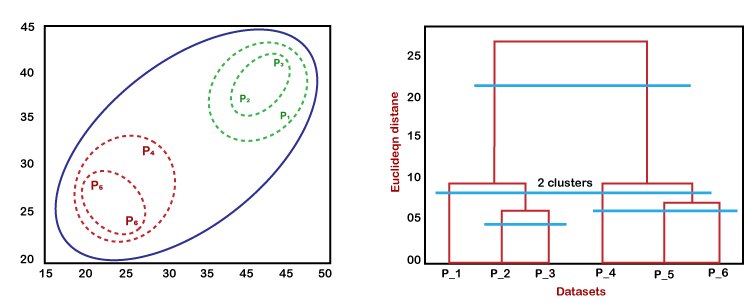
1. **Hierarchical Clustering**

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram.** Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

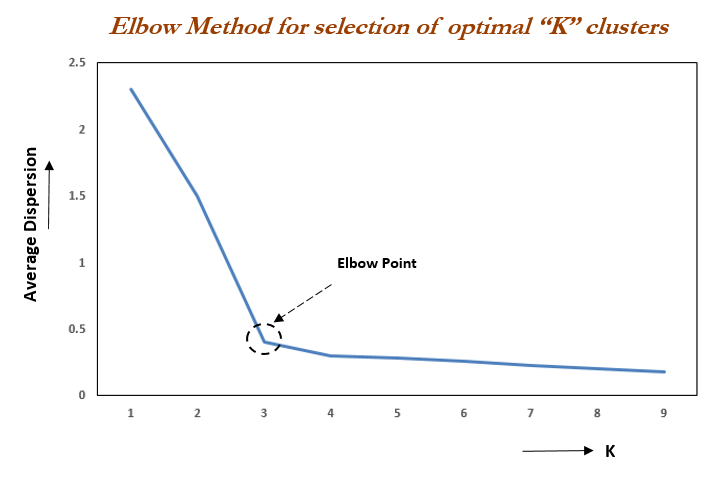
1. **Agglomerative -**Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive -** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

The working of the dendrogram can be explained using the below diagram:



**5. Elbow Plot:**

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The **Elbow Method** is one of the most popular methods to determine this optimal value of k. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



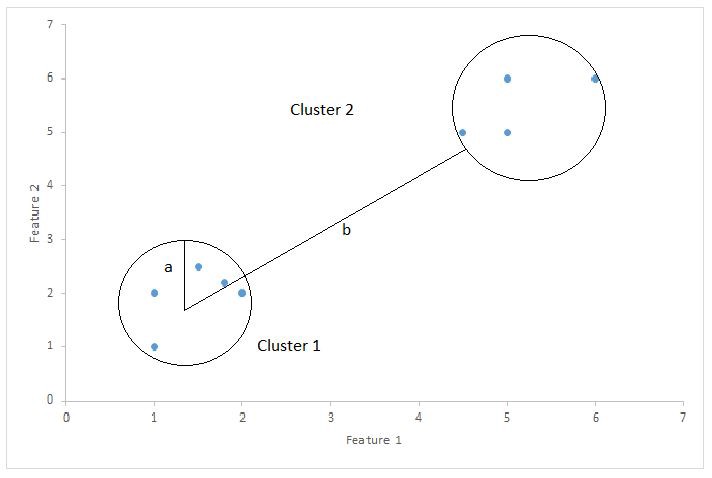
**6. Silhouette Score**

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.



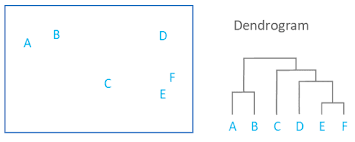
Silhouette Score = (b-a)/max(a,b)

Where, a=average intra-cluster distance i.e. the average distance between each point within a cluster.

b= average inter-cluster distance i.e. the average distance between all clusters.

**7. Dendrogram**

A dendrogram is a tree or branch diagram that visually shows the relationship between similar objects. Each of the branches of the tree represents a category of class, while the entire tree diagram shows the hierarchy relationship between all the classes or branches. The objects in a certain category of class share similar features or characteristics, and are referred to as clusters. A dendrogram is a visual representation that uses different branches to show the relationship and the order of how similar or dissimilar objects are to each other. When looking at a dendrogram, objects that share similar characteristics will be located along the same branch. While objects that do not share similar characteristics will be located on a separate branch. Dendrograms do not have a required number of branches, and it may not be predetermined how many branches you will need. The dendrogram may have a few branches or it may have many branches based on the variety of characteristics that the objects contain.



**8. Conclusion:**

Started with loading the data and then handled duplicated and null values then did data cleaning. Then I proceeded and performed feature engineering and made a new feature that is total amount. After that I performed EDA on different features. Then I calculated Recency, frequency and Monetary of the customers and give score to each customer on basis of there recency, frequency and monetary. I checked the distribution of these three and found all three were positively skewed so applied log transformation to make the distribution normal. I also plotted scatter plot for recency & monetary, frequency & monetary. Frequency and monetary shows positive relation, if the frequency of the items increases then monetary also increases and vice versa. I plotted Elbow plot and dendrogram to find the number of clusters, I also calculated silhouette score to find the number of clusters then I performed two clustering algorithms that are Kmeans clustering and Agglomerative clustering, I have taken the number of clusters as three in both of these clustering methods. Customers are well separated by these clusters.

**References-**

1. Geeksforgeeks
2. Scikit-learn
3. Towardsdatascience
4. Oreilly