**ONLINE RETAIL CUSTOMER SEGMENTATION**

**(A project on unsupervised Machine Learning)**



**Submitted by:**

**Shreesha K, Data Science Trainee, Almabetter**

**E-mail:** [**shreeshasrisha@gmail.com**](mailto:shreeshasrisha@gmail.com)

**Github Link:** [**https://github.com/shreesha2304/Online-Retail-Customer-Segmentation**](https://github.com/shreesha2304/Online-Retail-Customer-Segmentation)

**ABSTRACT**

In this era of a highly competitive business environment, a company must make effective marketing policies. A company might have several groups of customers. When a company is small, these customer groups are easily identifiable, and each customer is easily targetable. As the company grows, it becomes difficult to analyze customer behavior, and companies cannot give individual attention to different customer requirements. In such a situation, humans cannot make a proper strategy to segregate customers, and a data-driven approach is much needed to make judgments. For a medium or large company, the focus is not only acquiring new customers but also retaining existing customers. Many companies survive only by certain groups of customers loyal to the company. These customers generate high revenues, make promotions, and help the company to set up newer benchmarks. When the resources are limited, a company cannot invest too much in customers and it becomes crucial to identify those loyal customers. Moreover, companies must focus on preventing customer churn by resolving the issue of these dissatisfied customers. Hence, companies perform customer segmentation.

***Keywords: Clusters, Segmentation, Unsupervised ML Model.***

1. **PROBLEM STATEMENT**

In this project, our 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.

Below there are the description of the attributes that will be used in our model for better understanding of the data:

* 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.

1. **I****NTRODUCTION**

The objective of the project is to build an ML model that can segment the customers of the company based on certain features.

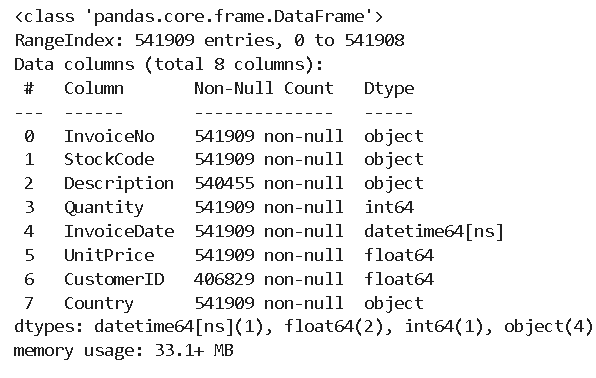
First, we briefly checked the data for null and missing values. Then we explored the data to find the countries with the maximum number of customers. We observed that the UK was the country with the maximum number of customers. We also checked the highest-selling products and their details. Then we checked the year with the highest sales. We also found the monthly, daily, date-wise, and hour-wise distribution of the sales. We calculated the net transaction amount for each order. Then we could identify the countries which bought the highest revenue for the company. We found the month, day, and hour for which the transacted amount was maximum.

To identify the best customer groups, we performed RFM (Recency, Frequency, Monetary) Analysis. Recency indicates the last date when a customer has placed an order. If a customer has placed an order recently, chances are higher that the customer will buy the same product again. Similarly, a customer who placed an order once might place the same order once again. We can even find out how much revenue each customer generates and whether they are more likely to follow the same trend in the coming days.

We used KMeans clustering to perform segmentation. We used the Elbow method and Silhouette score to find the optimum number of clusters. After finding the number of clusters, we analyzed the behavior of customers in each cluster and concluded the results.

1. **DATA HANDLING AND FEATURE ENGINEERING**
   1. **Data loading and Checking**

As discussed earlier, this dataset contains information about the customer-wise orders and transacted amounts for each order.



We loaded the data from the drive using pandas.read\_csv function. Our dataset had had 541909 rows and 8 Columns.

This dataset contains the data of datatype string, int, float, and datetime.

* 1. **Checking for Null values and Duplicates**

Null or missing values are the direct results of errors or inefficiency in data recording. If they are left untreated, our ML models will result in an error. Hence, they need to be treated in the initial stages.

Our dataset had around 1454 Null Values in the 'Description' column and 135080 missing values in the 'CustomerID' column. Since these columns cannot be filled by any other values, the only option was to remove those rows.

Duplicate entries are redundant and they unnecessarily increase the complexity. It is important to remove such observations.

In our dataset, we had around 5225 duplicate entries. We removed those entries by using the 'dropna' command.

* 1. **EDA**

As the first step, we checked the countries with the highest number of customers. We found that most customers were from the United Kingdom. Then we found the most- sold item was a T-Shirt Holder. 2011 was the year with the highest number of sales. After analyzing the monthly sales, we found that November was the month with the highest sales and most sales occurred in the ending months of the year. We noted that there were no sales recorded on Saturdays, indicating that it might be the weekly off day. Most sales happened at the beginning of a month. In the afternoon hours, maximum sales were recorded and there were no sales before 6 AM and after 8 PM. The Netherlands and Australia were the countries that generated the highest revenue for the company. On Sundays, the amount transacted was the least. The highest revenue was generated at the beginning of the month.

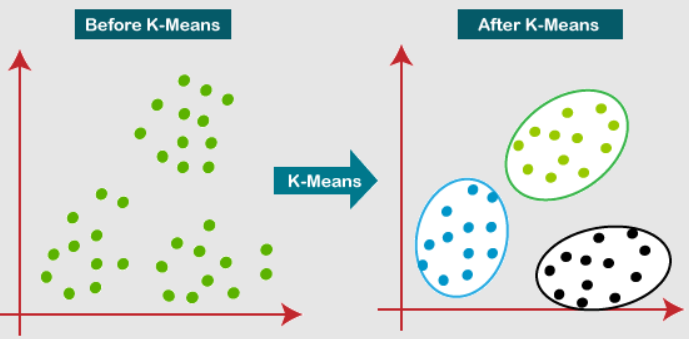
For our further analysis, we generated some new columns such as frequency, total transaction, and purchase recency. We checked the distribution of individual columns and by using the quantile method, we removed the outliers. Since the distribution was not normal, we performed a square root operation on each column to get a nearly normal distribution.

* 1. **Fitting the Model**

The model was built using KMeans Clustering. The optimum number of clusters was found by the Elbow method and Silhouette Score.

1. **ALGORITHMS**
   1. **KMeans Clustering**

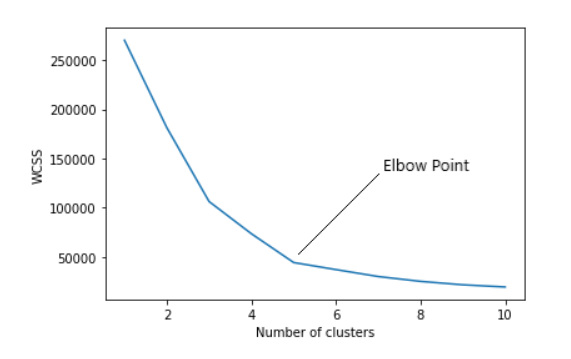
KMeans clustering is an unsupervised ML algorithm that groups the unlabeled dataset into different clusters. The user must specify the number of clusters to be formed. It is a centroid-based algorithm, which indicates that each cluster is associated with a centroid. The algorithm minimizes the sum of squared distances (SSD) between a given data point and its corresponding centroid.



One of the limitations of KMeans Clustering is that the user needs to specify the numberof clusters to be formed. It is not easy to choose the number of clusters intuitively and we can use two ways to find the number of clusters.

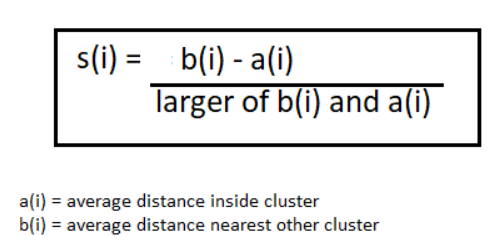
* + 1. **The Elbow Method**

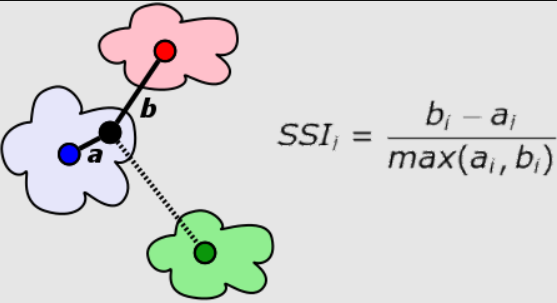
In the Elbow method, we first set several cluster sizes (K). For each value of K, we will fit the model and find the value of WCSS Score (Within Cluster Sum of Squares, also known as SSD). If we plot the values of K with their corresponding WCSS Score, the curve looks like an elbow. As we increase the value of K, we observe that the WCSS value decreases. The WCSS value will rapidly decrease till a particular value of K, then after the curve becomes almost parallel to the X-axis. The K value corresponding to this transition point is the optimum value for K and further analysis is carried out.



* + 1. **The Silhouette Score**

Silhouette Score is a performance parameter used to find to what extent the clustering model has performed. It is a score ranging between -1 to 1. A silhouette score of 1 means the clusters are well apart from one other and distinguished clearly, while -1 indicates the clustering was performed in the wrong way.





1. **CONCLUSION**

We have loaded the data, treated the NULL values, performed feature engineering, performed exploratory data analysis, and built ML models. After segregating the customers into different categories, we analyzed the behavior of customers in each category and concluded the following conclusions.

* 'Cluster 0' Customers- Their recent purchase date is too old. They do not purchase very frequently and the amount they spend is not high. We can classify them as Low Valued Customers.
* 'Cluster 1' Customers- These customers have placed an order more recently, but the frequency with which they order is not that great. They spend comparatively more than cluster 0 customers. We classify them as Medium Valued Customers.
* 'Cluster 2' Customers- These customers are the most loyal to the company. They place the orders more frequently and recency is also low. The average amount they spend on an order is also high. They are 'High Valued Customers'.
* We can reward Cluster 2 customers. They can adapt to new products. Moreover, they help to promote the company.
* For cluster 1 customers, we can offer membership or loyalty programs. We can recommend them related products to upsell them. This initiative helps to make them loyal customers of the company.
* In cluster 0, we have some customers who spent heavily but not recently. We need to identify them and initiate a reactivation campaign. We can also offer them promotions and conduct surveys to know what went wrong. In this way, we

can avoid losing them to competitors.

1. **REFERENCES**

* AnalyticsVidya
* Javapoint
* TowardsDataScience