

**TABLE OF CONTENTS: -**

**Customer Segmentation Based On Purchase History**

***By-SOUVIK SIKDER***

**1. Problem Statement**

**2. Project Objective**

**3. Data Description**

**4. Data Pre-processing Steps and Inspiration**

**5.RFM Analysis**

**6. Choosing the Algorithm for the Project**

**7. Motivation for Choosing the Algorithm**

**8. Assumptions**

**9. Model Evaluation and Techniques**

**10. Inferences from the Same**

**11. Future Possibilities of the Project**

**12. Conclusion**

**13. Reference**

**PROBLEM STATEMENT: -**

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

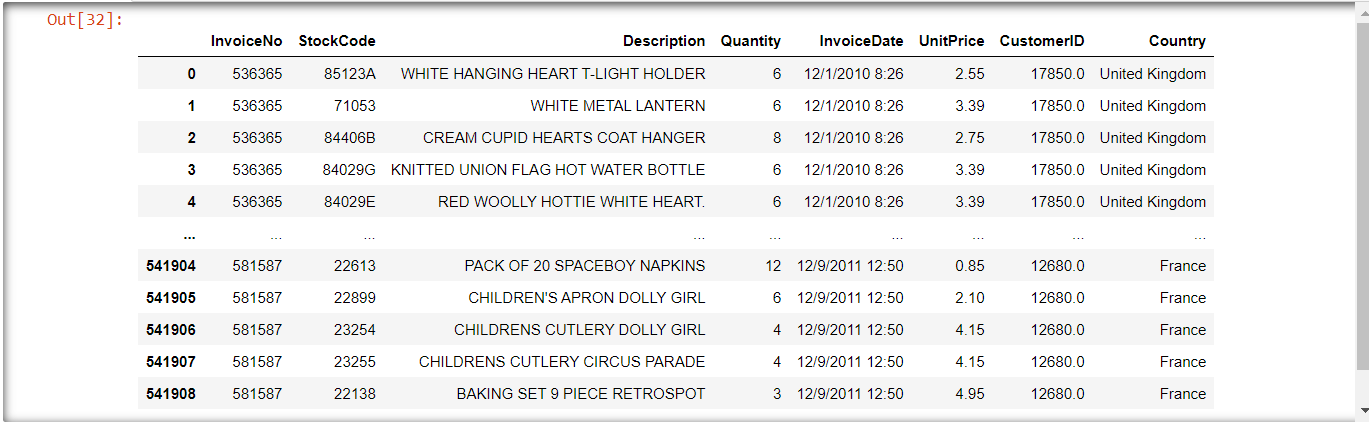
1. Using the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behaviour.

**Objective: -**

The main objective of the project is to segment customer based on purchase history using different customer segmentation algorithms. The project aims to look at purchasing patterns for customers. The overarching objective of the project would be to maximize customer satisfaction, retention, and revenue by implementing targeted strategies based on insights gained from analysing the customer purchasing history. This includes personalized marketing, tailored promotions, and an improved understanding of customer segments to enhance overall business performance.

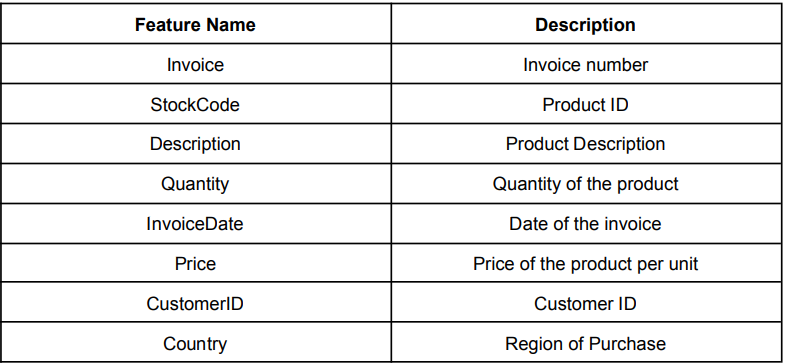
**DATA LOADING:-**

Once all python libraries are imported dataset is loaded for analysis of the dataset and generate important insights about the dataset. Following are the important columns in the dataset.



**DATA DESCRIPTION: -**

Dataset has 541909 rows and 8 columns. Following table describe each and every column in use.



**DATA PREPROCESSING STEPS AND INSPIRATION:** -

The preprocessing steps involved are as follows: -

* Getting information about the dataset and the datatype of each column.
* Check for null values and duplicate values and dropping them.
* Convert Invoice Date column to datetime format.
* Extract Month from the Invoice Date column and generating another column named Month and finally dropping Invoice Date column.

**DATA VISUALIZATION: -**

* Plot bar graph to find total sales country-wise. Found out United States to be the top most selling product.
* Use pie chart to find the most selling product among all the product sold in the countries. Found out White Hanging Heart T-light Holder to be the most selling product.
* Use bar graph to find out the month when maximum sale happened.
* Next for customer segmentation we find out recency, monetary and frequency.

**RFM ANALYSIS:-**

1. **Recency (R):**
   * **Definition:** Recency represents how recently a customer has made a purchase or interacted with the business.
   * **Calculation:** It is often measured as the time elapsed since the customer's last purchase or interaction.
   * **Purpose:** Customers who have interacted recently are often considered more engaged and are potentially more valuable.
2. **Monetary (M):**
   * **Definition:** Monetary represents the total monetary value of a customer's transactions or spending with the business.
   * **Calculation:** It is often calculated as the sum of all the monetary values (e.g., purchase amounts) for a particular customer.
   * **Purpose:** Customers with higher monetary values are typically more valuable to the business, and targeting them may yield higher returns.
3. **Frequency (F):**
   * **Definition:** Frequency represents how often a customer makes a purchase or interacts with the business.
   * **Calculation:** It is often measured as the count of transactions or interactions within a specific period.
   * **Purpose:** Customers with higher frequency may be more loyal or engaged with the business, and they might be more responsive to marketing efforts.

* **Recency ('Recency'):** The code calculates recency by finding the number of days since the customer's last purchase, relative to the maximum date in the 'Invoice Date' column.
* **Monetary ('Monetary'):** The code calculates monetary value by summing up the unit prices for each customer, representing the total spending.
* **Frequency ('Frequency'):** The code counts the number of invoices for each customer, representing how often they have made a purchase.

These three metrics are commonly used in combination for customer segmentation in marketing and business analytics, helping businesses identify and target specific customer segments based on their behaviour.

**CHOOSING ALGORITHM FOR THE PROJECT: -**

We used to customer segmentation or clustering algorithm to segment the customers based on purchase history. We briefly discussed those below.

**K-Means Algorithm: -**

K-Means is a popular clustering algorithm used in machine learning and data analysis. It's an unsupervised algorithm that partitions a dataset into K clusters based on similarities in the data. Here's a brief explanation of the K-Means algorithm:

1. **Initialization:**
   * Choose the number of clusters, K.
   * Randomly initialize K cluster centroids. Each centroid represents the centre of a cluster.
2. **Assignment:**
   * Assign each data point to the nearest centroid. This is usually based on the Euclidean distance between the data point and the centroids.
3. **Update Centroids:**
   * Recalculate the centroids based on the mean of all the data points assigned to each cluster.
4. **Repeat:**
   * Repeat the assignment and centroid update steps iteratively until convergence. Convergence occurs when the centroids no longer change significantly or a set number of iterations is reached.

**Hierarchical Clustering: -**

Hierarchical clustering is another popular technique for clustering data points, and it creates a tree-like structure of clusters. There are two main types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering is more common, so I'll focus on that.

Here are the key steps of agglomerative hierarchical clustering:

1. **Initialization:**
   * Treat each data point as a single-point cluster.
2. **Merge:**
   * Identify the two closest clusters and merge them into a new cluster. This process is repeated until only one cluster remains.
3. **Dendrogram:**
   * Create a dendrogram (tree diagram) that shows the hierarchical relationships between clusters at each stage of the process.
4. **Cutting the Dendrogram:**
   * Decide the number of clusters you want by setting a threshold or cutting the dendrogram at a certain height. The vertical line where you cut the dendrogram represents the clusters.
5. **Assigning Labels:**
   * Assign each data point to a cluster based on the cut made in the dendrogram.

Applying both these algorithms we found out that hierarchical algorithm is well suited for the dataset used where we can easily classify the customers based on purchase history.

Then, we used Apriori Algorithm where we were keen to see association of each product with other product in our dataset.

Association Rules or Market Basket Analysis gives us the idea of association of items. It is type of recommendation system used in retail scenario. It is different from collaborative filtering system. Collaborative filtering is user-centric while association rules are product based. Association rules has algorithms like Apriori Algorithm and FP growth. There are usually 3 metrics used in association rules to set rules for association of items. These are Support, Lift, Confidence.  
Support-Among all the transactions how frequently an item is brought.  
Confidence-If an item is brought how likely another item is brought is given by confidence.  
Lift-Lift indicates how much more likely items are brought together compared to their individual probabilities. Lift greater than 1 indicates positive association that items are more likely to be purchased together. Lift of 1 implies independence while less than 1 suggest negative association.

Association between products can be seen by the table provided below: -



**ASSUMPTIONS OF THE ALGORITHM USED: -**

**Hierarchical Clustering:**

1. **Euclidean Distance:**
   * Assumes that the data can be effectively measured using Euclidean distance. This distance metric is often used to measure the dissimilarity between clusters.
2. **Single Linkage:**
   * The results can be sensitive to the choice of linkage method. Single linkage, complete linkage, and average linkage are common methods, and each makes different assumptions about the relationships between clusters.
3. **Dendrogram Interpretability:**
   * Assumes that the dendrogram structure is meaningful. The interpretation of clusters may depend on visual inspection of the dendrogram, and decisions about the number of clusters may be subjective.

**K-Means:**

1. **Cluster Shape:**
   * Assumes that clusters are spherical and equally sized. K-Means may struggle with clusters of different shapes, sizes, or densities.
2. **Initial Centroid Sensitivity:**
   * The results can be sensitive to the initial placement of centroids. Different initializations may lead to different final clusters.
3. **Fixed Number of Clusters:**
   * Requires specifying the number of clusters (�*K*) in advance. Determining the optimal �*K* value may require experimentation or domain knowledge.

**Apriori Algorithm:**

1. **Binary Data:**
   * Primarily designed for binary datasets where items are either present or absent. If the data is not in a binary format, it needs to be discretized or transformed.
2. **Frequent Itemset Assumption:**
   * Assumes that interesting patterns are frequent item sets. Frequent item sets are sets of items that appear together in a significant number of transactions.
3. **Support Threshold:**
   * Requires the setting of a support threshold, which determines the minimum frequency required for an itemset to be considered "frequent." The choice of this threshold may impact the discovered rules.
4. **Rule Confidence:**
   * Assumes that rule confidence is a meaningful measure of rule quality. Confidence is the conditional probability that a rule is true given that the antecedent is true.
5. **Association Rule Interest:**
   * The algorithm assumes that the goal is to find interesting association rules. What is considered "interesting" may depend on the application and the interpretation of the rules.

**Model Evaluation and Techniques Used: -**

#### **Silhouette Score:**

* Measures how well-separated clusters are. Ranges from -1 (incorrect clustering) to +1 (highly dense and well-separated clusters).
* Higher values indicate better-defined clusters.

**Future Possibilities of the Project: -**

1. **Personalized Marketing Campaigns:**
   * Use the customer segments to tailor marketing campaigns. Develop personalized promotions, recommendations, and content based on the specific needs and preferences of each segment.
2. **Dynamic Segmentation:**
   * Implement dynamic segmentation that adapts in real-time as customer behaviour evolves. This could involve using machine learning models to continuously update and refine segments based on the latest data.
3. **Predictive Analytics:**
   * Integrate predictive analytics to forecast future customer behavior within each segment. This could include predicting future purchase patterns, customer churn, or the likelihood of responding to a particular marketing campaign.
4. **Customer Journey Analysis:**
   * Extend the segmentation analysis to include a comprehensive understanding of the customer journey. Identify touchpoints, challenges, and opportunities at each stage of the customer lifecycle.

**CONCLUSION: -**

In conclusion, the customer segmentation project has provided valuable insights into the purchasing behaviour of our customer base, enabling us to create meaningful clusters that group customers with similar characteristics. Here are the key takeaways and conclusions from the project:

1. **Identification of Distinct Customer Segments:**
   * Through the application of RFM analysis and clustering algorithms such as K-means and hierarchical clustering, we identified distinct customer segments based on recency, frequency, and monetary values. These segments represent groups of customers with similar purchasing patterns.
2. **Understanding Customer Behaviour:**
   * The segmentation analysis has deepened our understanding of customer behaviour by revealing patterns, preferences, and trends within different segments. We gained insights into how recently customers have made purchases, how frequently they buy, and their overall spending patterns.
3. **Targeted Marketing Opportunities:**
   * With the identified customer segments, we now have the opportunity to implement targeted marketing strategies. This includes personalized promotions, product recommendations, and communication strategies tailored to the unique characteristics of each segment.
4. **Improved Customer Engagement:**
   * Tailoring our approach to each customer segment allows for more effective and personalized communication. This can lead to improved customer engagement, satisfaction, and loyalty as we address the specific needs and preferences of different customer groups.

**REFERENCES: -**

* [Apriori Algorithm - GeeksforGeeks](https://www.geeksforgeeks.org/apriori-algorithm/)
* [Clustering in Machine Learning - GeeksforGeeks](https://www.geeksforgeeks.org/clustering-in-machine-learning/)
* [User guide and tutorial — seaborn 0.12.2 documentation (pydata.org)](https://seaborn.pydata.org/tutorial.html)
* [Data Science Courses Online | Learn Fundamentals to Advanced | Intellipaat](https://intellipaat.com/course-cat/data-science/)
* [K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0020025522014633)