Online Retail Customer Segmentation

Sapana Pawar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Customer segmentation plays a key role in making business decisions. In the competitive field of e-commerce, it is very important to satisfy the customer needs and to identify the potential customer and these things should be done at the right time in the right manner. In this paper, various segments of customer segmentation are discussed and different techniques in customer segmentation are presented. Among them, clustering is best and by comparing the techniques of clustering we analyze that K-Means algorithm is the most efficient and it is very simple to use.

Keywords:machine learning,Customer segmentation,clustering

1.Problem Statement

In this project, 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.

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.

2. Introduction

The online retail industry has changed the way customers shop as everything is available online. In order to build a loyal customer base, a company needs to deploy various marketing strategies focused on the diverse nature of its customers.
A possible solution is to segment customers and make targeted marketing strategies for which historical data of customers is required.
RFM (Recency, Frequency, Monetary) analysis is a technique that helps in extracting insights from the records and can be used for segmentation of customers as well.

3. Customer Segmentation

3.1 What is Customer Segmentation?

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

3.2 Why Segment Customers?

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

- Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).
- Select the best communication channel for the segment, which might be email, social media posts, radio advertising, or another approach, depending on the segment.
- Identify ways to improve products or new product or service opportunities.
- Establish better customer relationships.
- Test pricing options.
- Focus on the most profitable customers.
- Improve customer service

4. Steps involved:

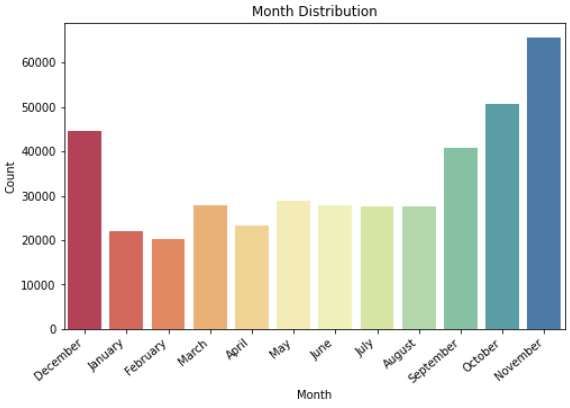
Data Cleaning and Visualization

After loading the dataset we cleaned our data i.e. removed null values,
changing the datatype of Customer Id as per Business understanding, and dropped the duplicates. And Converting ‘InvoiceDate’ column to datetime to proper datatype.

Exploratory Data Analysis and Visualization

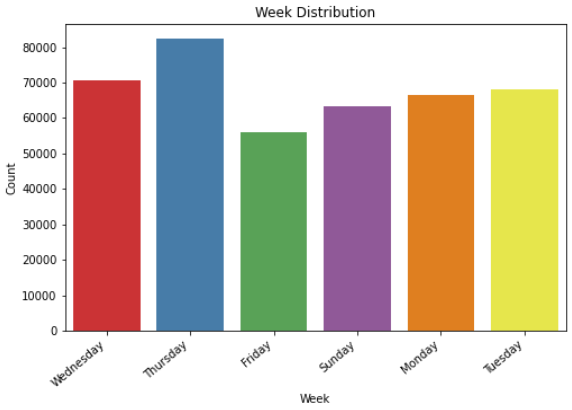
In EDA and visualization, I have done univariate and bivariate analysis using different plots. Got some meaningful trends and insights from the data.

1.Orders per months



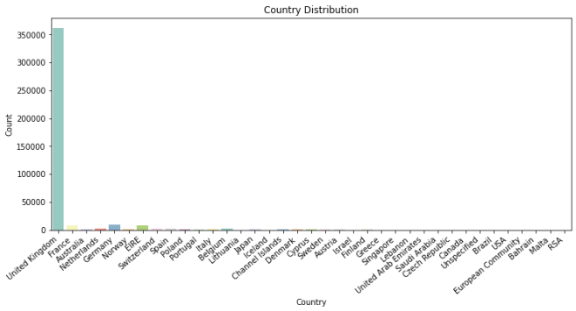
- We can see that the months with higher sales were oct, nov and dec.

2. Orders per weeks



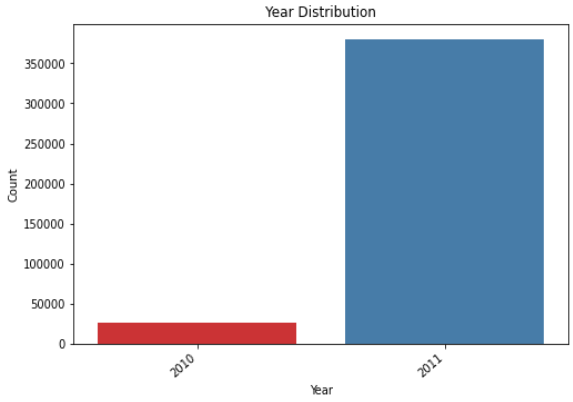
- As we can see Thursday have shown more transactions and Fridays the lowest. No transactions on Saturdays.

3. The total sales per country



- As expected, the company receives the highest number of orders in the UK (since it is a UK based company).

4. Transaction per year



- As we can see more transaction is seen in 2011 as compared to 2010.

Data Preparation

In this step, we have to prepare or process the data before feeding it to the model.
I am going to segment the data based on three factors: Recency, Frequency and Monetary.
Recency can be evaluated by subtracting the transaction date from the most recent date of transaction from the dataset.
Frequency is nothing but the number of transactions made by the customer. And Monetary can be calculated by multiplying the number of units and price of the product.

Outliers

I used the Inter Quartile Range method to remove outliers from the new data frame containing Recency, frequency and Monetary columns.

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 algorithm to it.
The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.
I have used standard scalar to scale the data.

Fitting model

For modeling I had tried the K-means clustering algorithm and hierarchical clustering.

5.1. Algorithms:

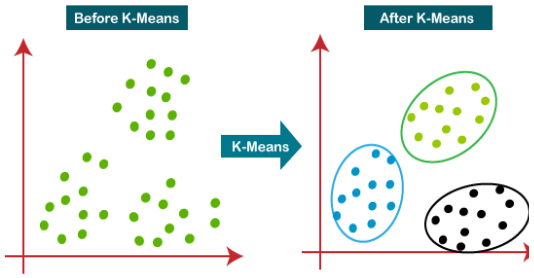
1. K-Means Clustering:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

- First we initialize k points, called means, randomly.
- We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
- We repeat the process for a given number of iterations and at the end,

we have our clusters.



5.2. Model Performance

Model can be evaluated by metric such as:

Silhouette Score

To study the separation distance between the clusters formed by the algorithm silhouette analysis could be used. The distance between the cluster can be calculated by different types of distance metrics (Euclidean, Manhattan, Minkowski, Hamming). Silhouette score returns the average silhouette coefficient applied on all the samples.

The Silhouette Coefficient is calculated by using the mean of the distance of the intra-cluster and nearest cluster for all the samples. The Silhouette Coefficient ranges from [-1,1]. The higher the Silhouette Coefficients (the closer to +1), the more is the separation between clusters. If the value is 0 it indicates that the sample is on or very close to the decision boundary between two neighboring clusters whereas a negative value indicates that those samples might have been assigned to the wrong cluster.

2. 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:

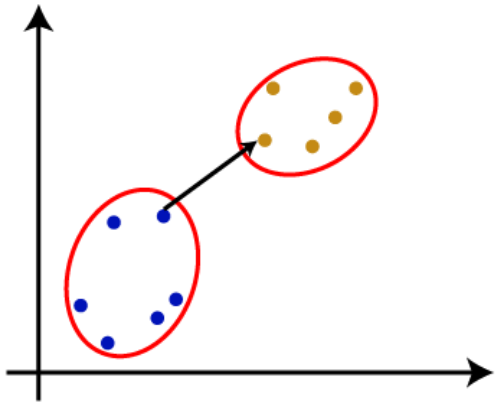
1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

Measure for the distance between two clusters

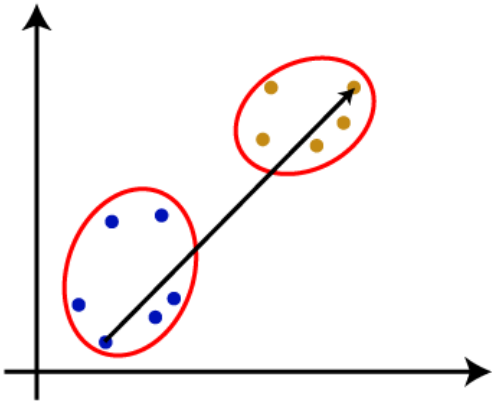
The **closest distance** between the two clusters is crucial for the hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for

clustering. These measures are called **Linkage methods**. Some of the popular linkage methods are given below:

Single Linkage: It is the Shortest Distance between the closest points of the clusters. Consider the below image:



Complete Linkage: It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.



Average Linkage: It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

6. Conclusion:

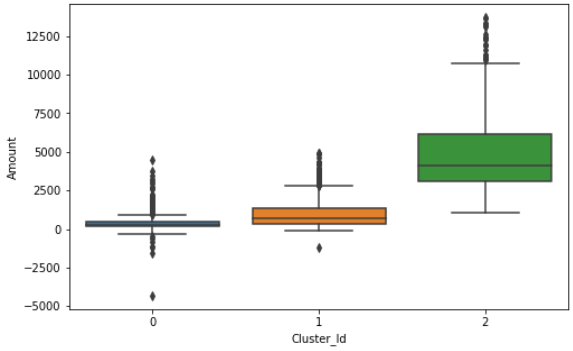
That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA and Visualization, null values treatment, feature engineering, data preprocessing and then model building.

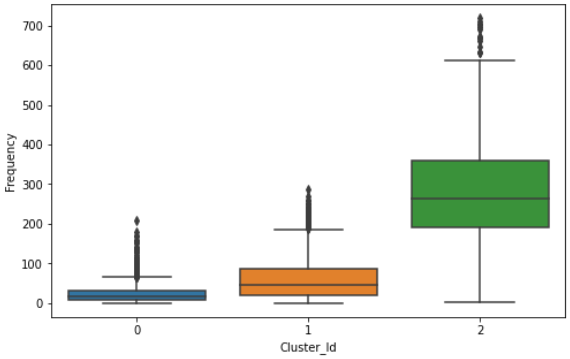
By observing all the silhouette scores for different values of k, I got the best results for k=3.
So the silhouette score of our best model is 0.50 which can be said to be good for this type of dataset.

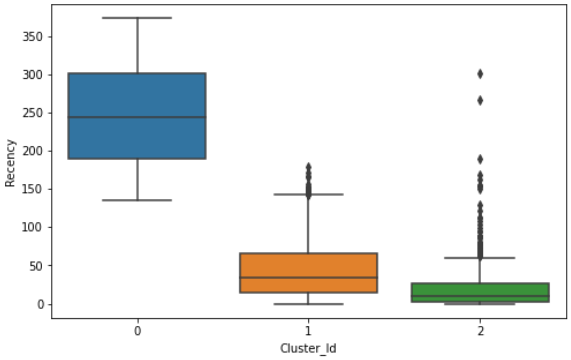
Observations:

Inference:

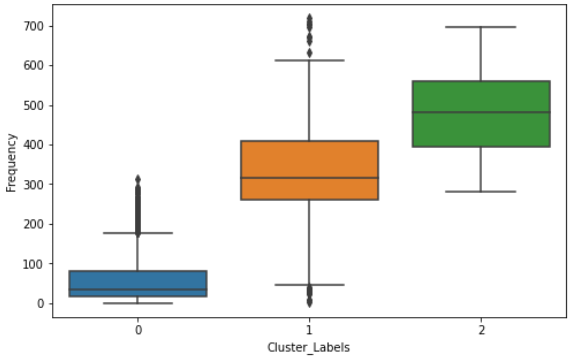
K-Means Clustering with 3 Cluster Ids:

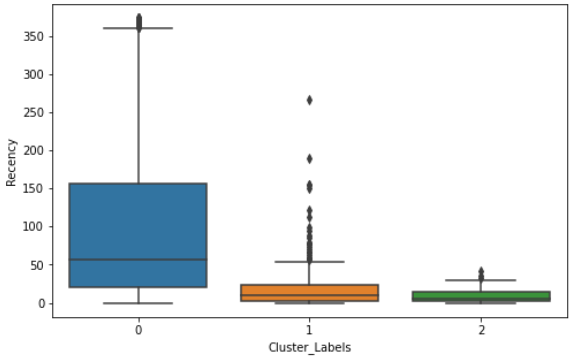






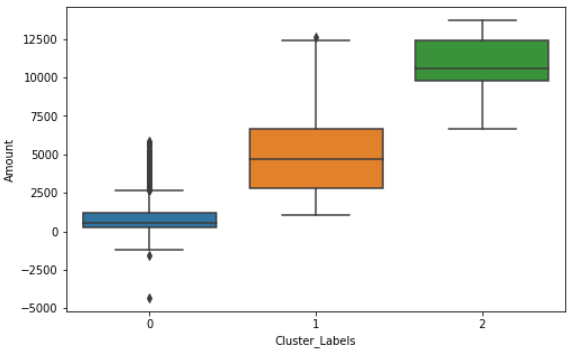
- Customers with Cluster Id 2 are the customers with a high amount of transactions as compared to other customers.
- Customers with Cluster Id 2 are frequent buyers.
- Customers with Cluster Id 1 are not recent buyers and hence least of importance from a business point of view.





- Customers with Cluster Labels 2 are the customers with high amount of transactions as compared to other customers.
- Customers with Cluster Labels 2 are frequent buyers.
- Customers with Cluster Labels 0 are not recent buyers and hence least of importance from business point of view.

Hierarchical Clustering with 3 Cluster La bels:



References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Javatpoint