**ONLINE RETAIL CUSTOMER SEGMENTATION**

**Shashank Someshwar Nanhe**

**Data science trainee,**

**AlmaBetter**

**Abstract:**

For a medium to large size retail store, it is also imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their ‘best’ or high-valued customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, companies use the technique of customer segmentation.

My experiment can help you understand what could be the best performing model for our project.

1. **PROBLEM STATEMENT:**

To identify major customer segments on a transnational data set.Data set contains all the transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based and registered non-store online retail.

The company mainly sells unique all-occasion gifts. Many customers of thecompany are wholesalers.

**Data description:**

* **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.
* I**nvoiceDate**: 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.

1. **INTRODUCTION:**

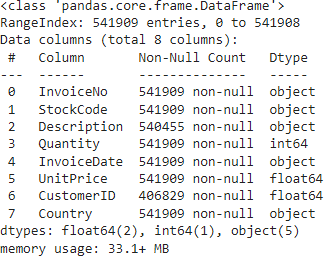
Exploratory Data Analysis (EDA) is for understanding and analyzing the datasets given to get better insights into the data and can be used for taking important business decisions. We can summarize the main characteristics of the datasets and can plot the data visually by using Python libraries. We can store our data in data frames by using the Pandas library from Python.

1. **STEPS INVOLVED:**
   1. **Data Exploration:**

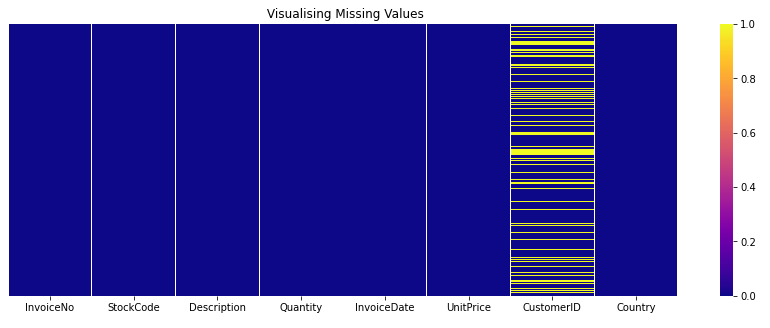
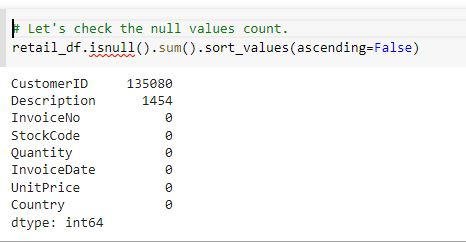
There are data sets which were given Online Retail Customer Segmentation csv dataset in which one dataset having Total Rows= 541909 Total features=8

**Data Wrangling:**

* 1. **Information of the data.**



* Invoice date to datetime.
* If Invoice No starts with C means it's a cancellation.
* Shape of data after dropping entries=397884
* **Null Values :**



* 1. **Data Analysis:**

It is a systematic way of describing and illustrating, condensing, and evaluating data by systematically using statistical or logical methods.

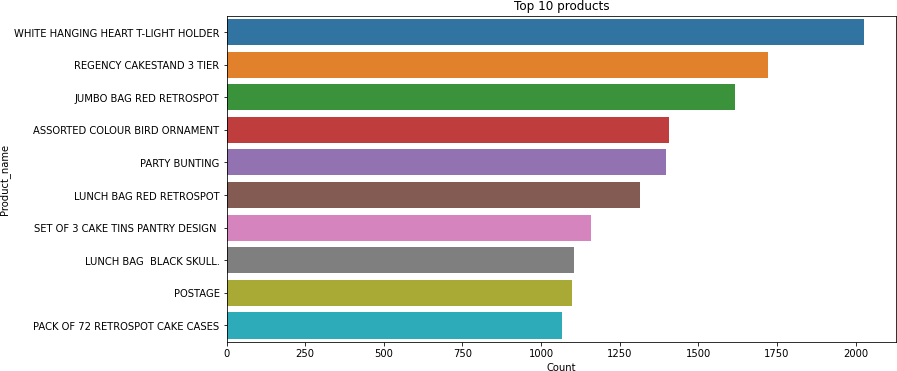
* 1. **Data Visualization:**

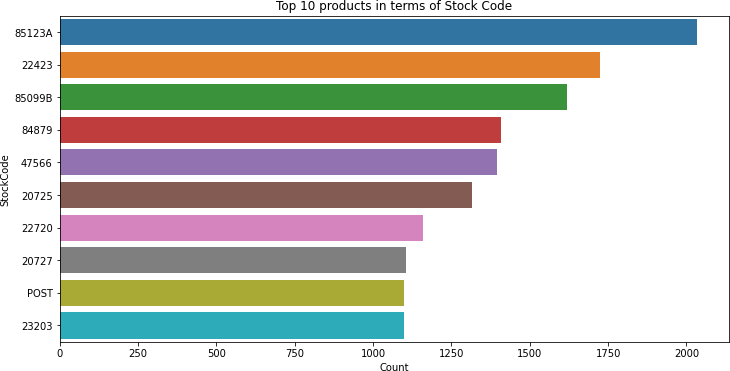
Python's libraries provide plenty of features with which users can create highly customized, elegant, and interactive plots, making data visualization with Python one of the most widely used features in today's data science environment.

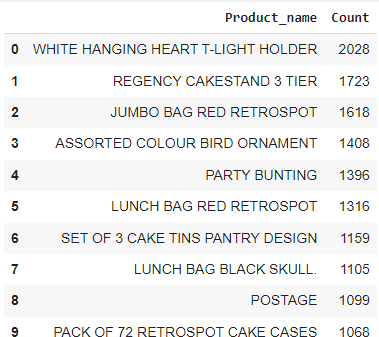
Data Visualization libraries in Python are:

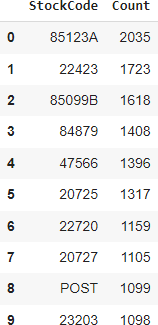
* **Matplotlib:** With a Python library, you can visualize arrays in 2D. NumPy is used to create Matplotlib, which is written in Python. There is a wide variety of plots in Matplotlib, including line, bar, scatter, histogram, etc. that can help us understand trends, patterns, correlations.
* **Seaborn:** A Python library for representing statistics with datasets, Seaborn is a dataset-oriented library. This library is built on matplotlib and creates a variety of visual representations. Pandas data structures are incorporated into it. To create informative visuals, the library performs mapping and aggregation internally

1. **Exploratory Data Analysis:**



****





Top 10 Products(Description wise)

* WHITE HANGING HEART T- LIGHT HOLDER is the highest selling product almost 2018 units were sold.
* REGENCY CAKESTAND 3 TIER is the 2nd highest selling product almost 1723 units were sold.

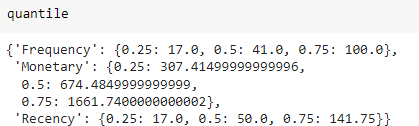
Top 10 products(Stock Code wise)

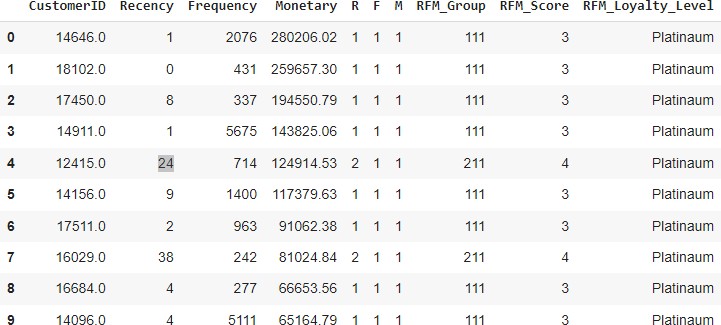
* StockCode-85123A is the first highest selling product.
* StockCode-22423 is the 2nd highest selling product.

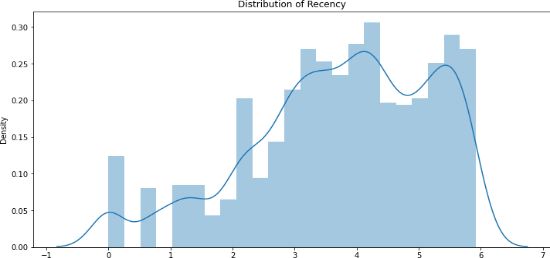
**Model Building:**

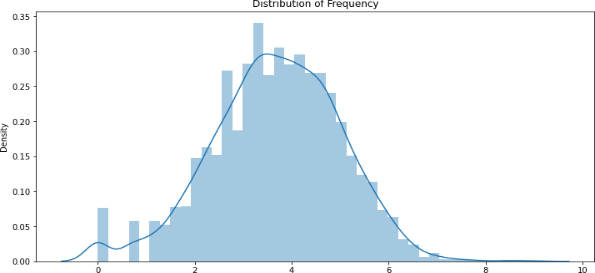
* 1. **RFM Model Analysis:**

Recency = Latest Date - Last Invoice Data.

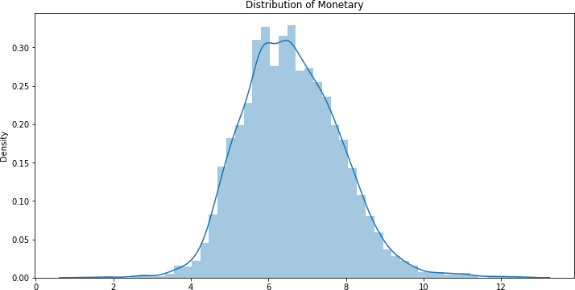
* + Frequency = Count of invoice no. of transaction(s).
  + Monetary = Sum of Total Amount for each customer.



*  Log transformation on Frequency, Recency and Monetary.

 Recency

Frequency



Monetary

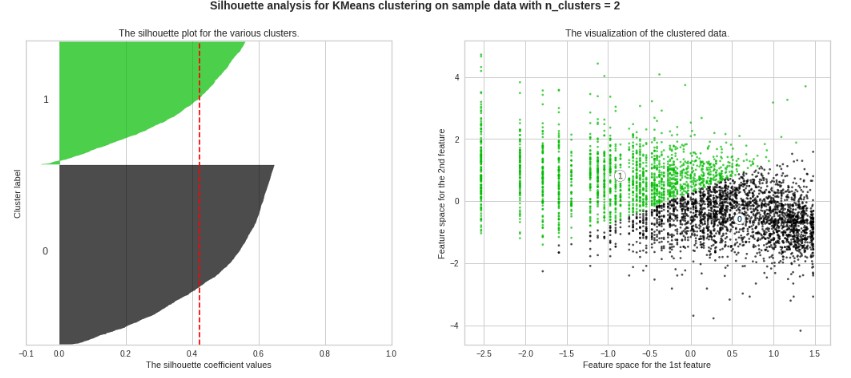
**What is RFM:**

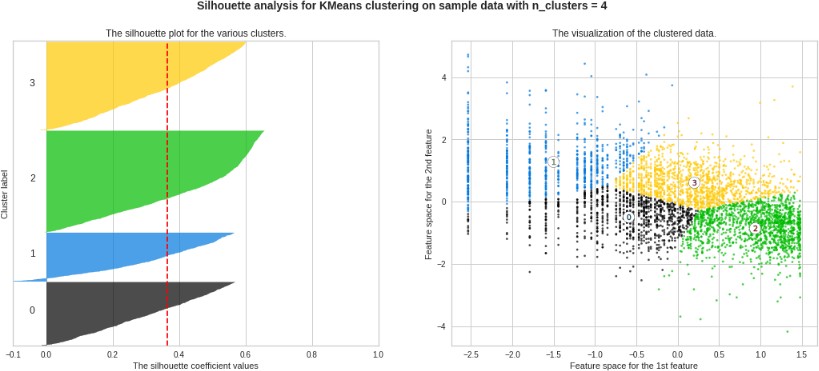
* + **RFM** is a method used to analyze customer value. RFM stands for RECENCY, Frequency, and Monetary.
  + **RECENCY**: How recently did the customer visit our website or how recently did a customer purchase?
  + **Frequency**: How often do they visit or how often do they purchase?
  + **Monetary**: How much revenue we get from their visit or how much do they spend when they purchase?

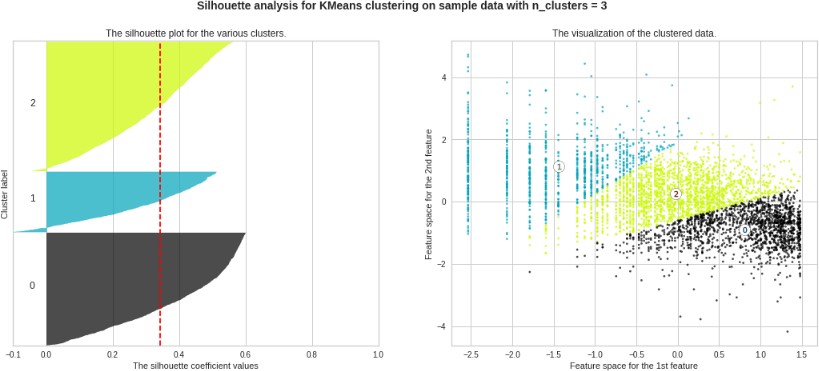
**Why it is Needed?**

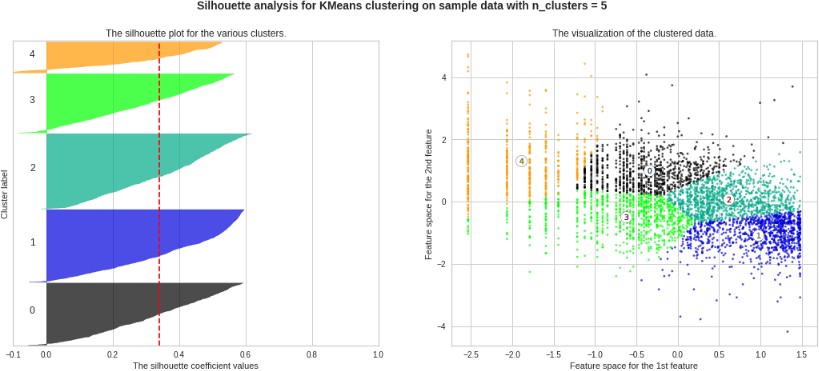
**RFM Analysis is a marketing framework that is used to understand and analyze customer behavior based on the above three factors RECENCY, Frequency, and Monetary.**

**The RFM Analysis will help the businesses to segment their customer base into different homogenous groups so that they can engage with each group with different targeted marketing strategies.**

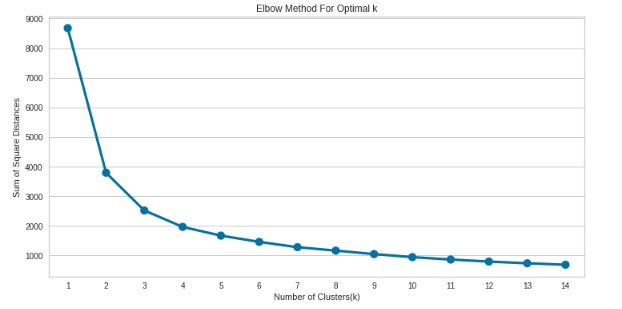
**K-Means Custering : (Recency and Monetary)**

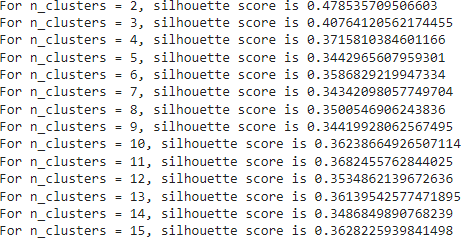


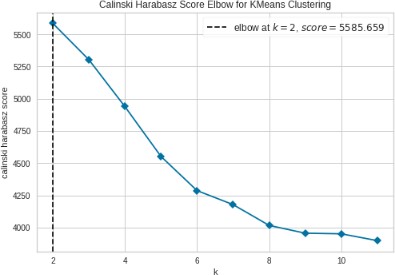




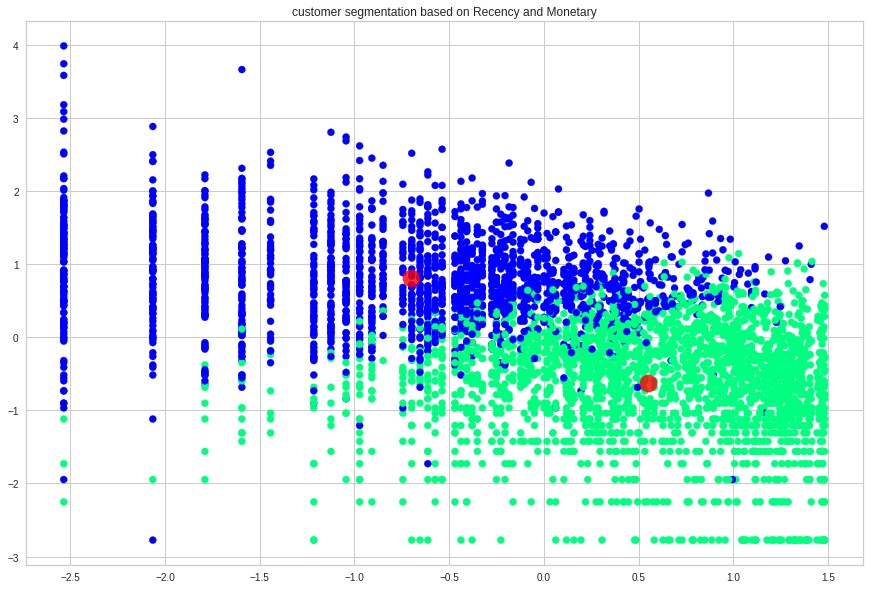
* + Finding the Optimal value of cluster using Elbow method and Silhouette Score.



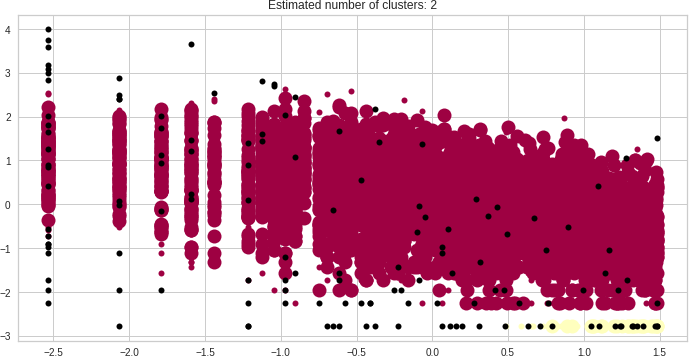


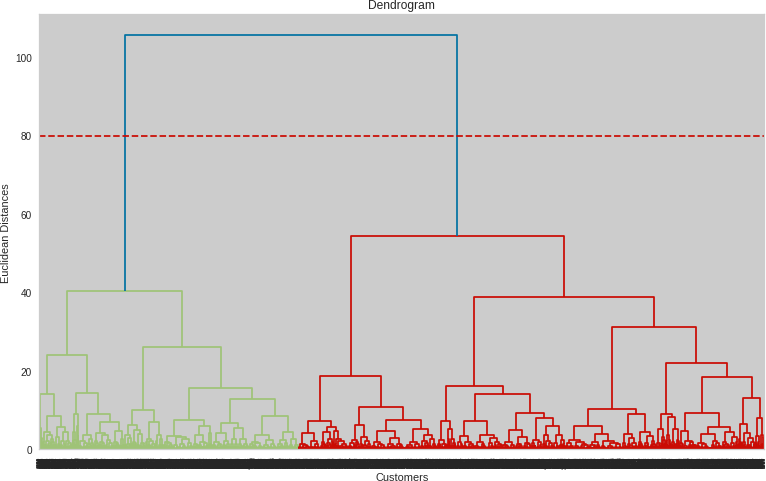


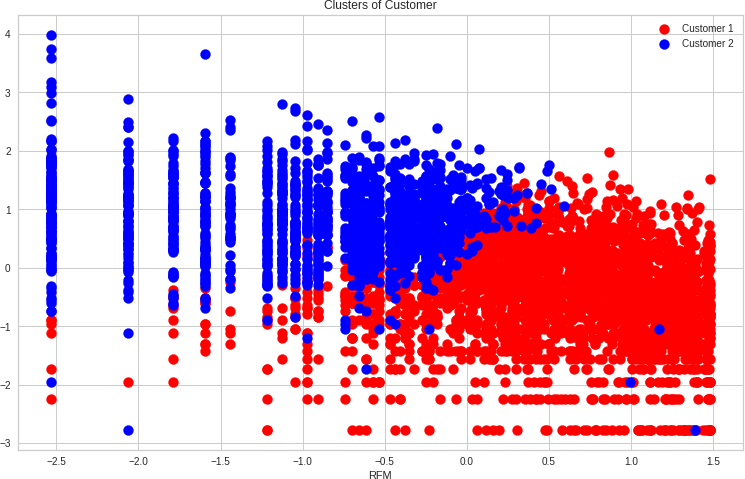
**K-means Clustering: ( Recency,Frequency and Monetary)**



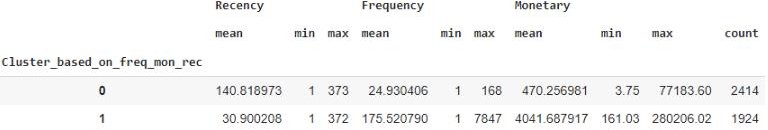
**DBSCAN Algorithm ( Recency and Monetary)**

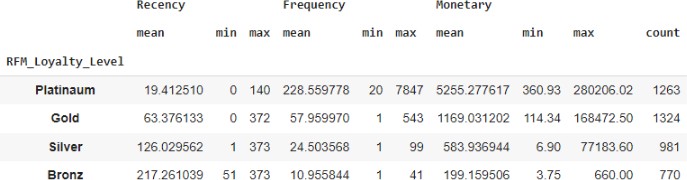


**Hierarchical Clustering(Recency, Frequency and Monetary)**



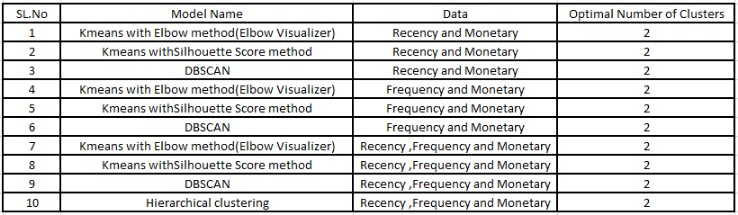
**Optimal Number of clusters using Dendogram.(Optimal Clusters=2)**

**Summary And Conclusion:**

Firstly we did clustering based on RFM analysis. We had 4 clusters/Segmentation of customers based on RFM score.

* **Platinum customers=1263 ( less recency but high frequency and heavy spending)**
* **Gold customers=1324 (good recency, frequency and monetary)**
* **Silver customers=981(high recency, low frequency and low spending)**
* **Bronze customers=770 (very high recency but very less frequency and spending)**

**Later we implemented the machine learning algorithms to cluster the customers.**



* **Above clustering is done with recency, frequency and monetary data(Kmeans Clustering) as all 3 together will provide more information.**
* **Cluster 0 has high recency rate but very low frequency and monetary. Cluster 0 contains 2414 customers.**
* **Cluster 1 has low recency rate but they are frequent buyers and spends very high money than other customers as mean monetary value is very high. Thus generates more revenue to the retail business.**
* **With this, we are done. Also, we can use more robust analysis for the clustering, using not only RFM but other metrics such as demographics or product features.**

**References:**

* https://www.analyticsvidhya.com
* https://www.kaggle.com
* https://www.geeksforgeeks.org/overview-of-data-science/