Capstone Project

Online Retail Customer Segmentation

Submitted by:

Avanish Dixit

Flow of the Presentation

- Introduction
- Data Preparation
- EDA
- Feature Engineering
- Different clustering Techniques
- Conclusion

Introduction

In recent years, there has been a massive increase in the competition among firms in sustaining in the online field. The profits of the company can be improved by a customer segmentation model. Customer retention is more important than the acquisition of new customers..

Objective: Basically, we have to segregate the customers into different clusters very effectively. using suitable methods.

Methodology: Machine Learning (ML) Clustering Algorithms (Unsupervised ML Model)

Dataset Summary:

- The data were collected from the 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.
- Initially, It has 541909 rows and 8 columns.
- The dataset also contains some canceled orders.
- Dataset contains null values.

Data Preparation

The database has the following features:

The dataset contains total of 8 features. The names are given below.

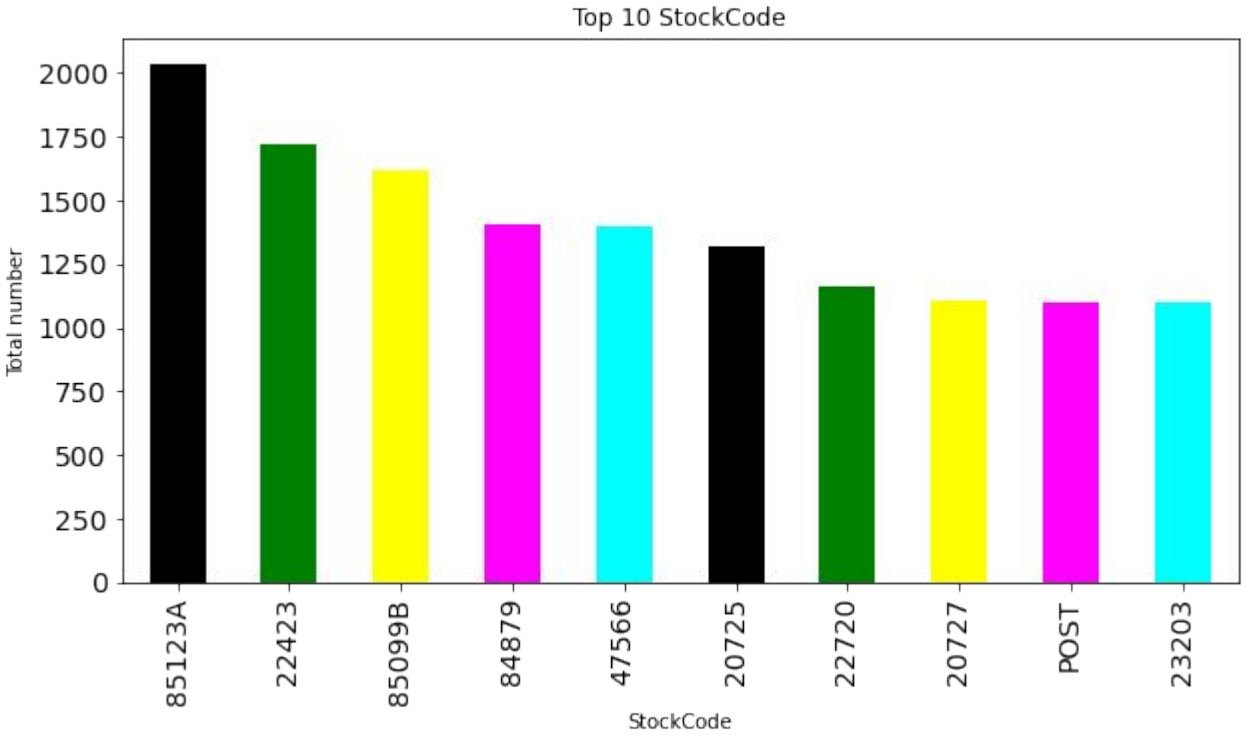
- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description**: 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.

Overview of Dataset

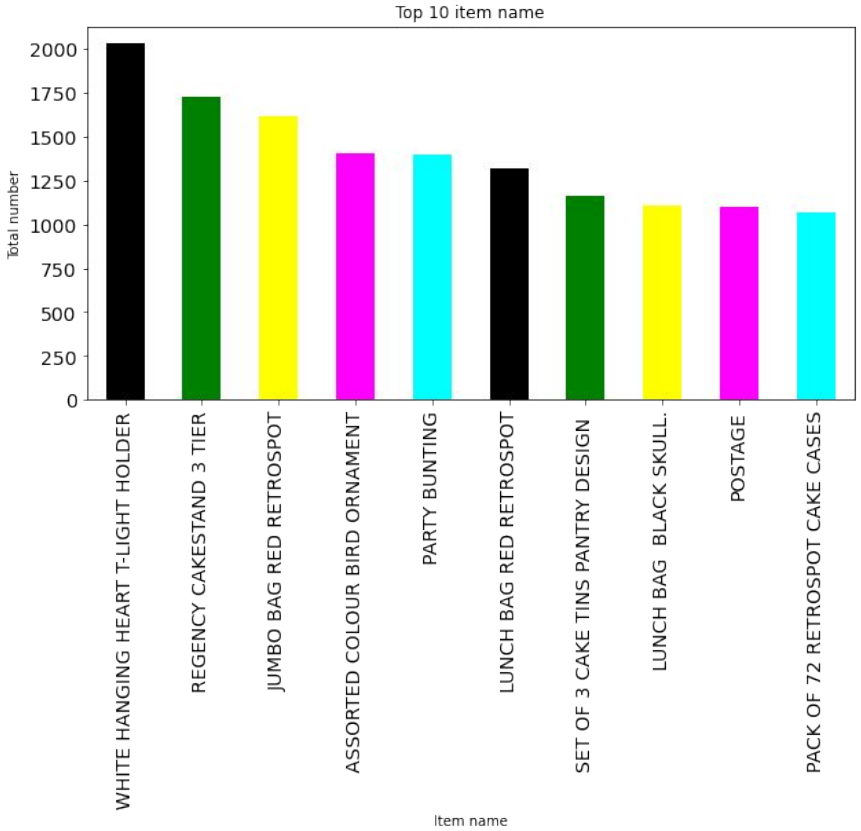
- It has ‘NaN’ or ‘Null’ values in ‘CustomerID’ and in ‘Description’ features.
- It is a problem of unsupervised algorithms so it does not have any dependent variable.
- Data does not contain Duplicate values.
- InvoiceDate contains date as well as time also..
- Price and Quantity have skewed histogram.
- InvoiceNo. Feature contains canceled order also.

EDA

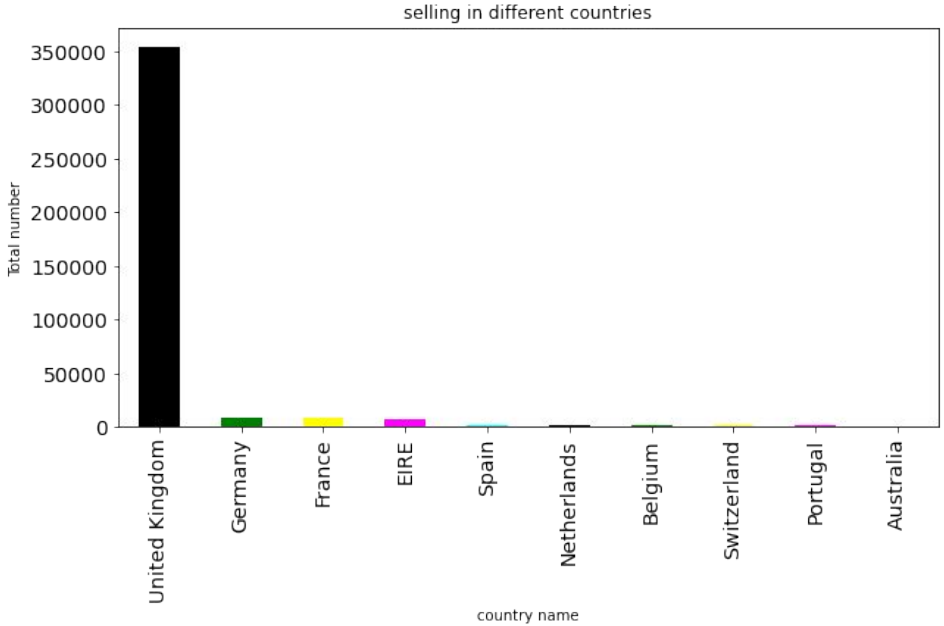
Plot of the Distribution of Top 10 Stocks



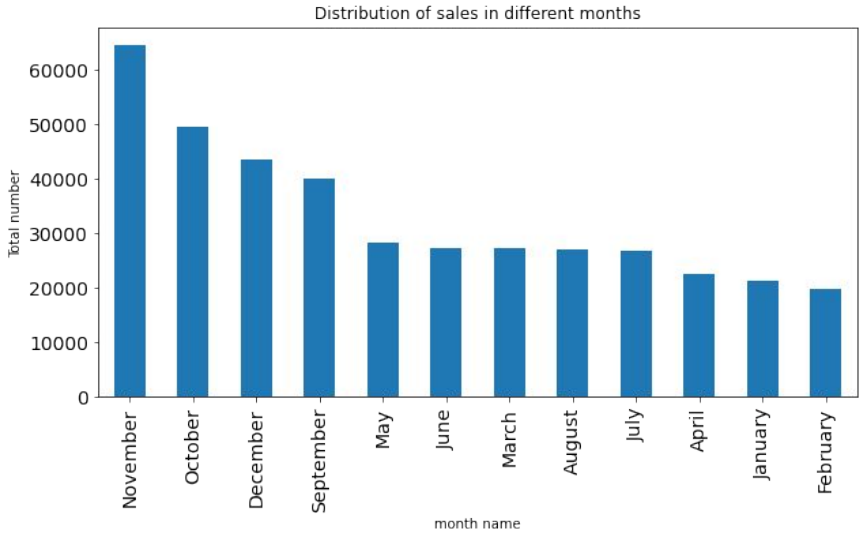
Plot of Top 10 items description

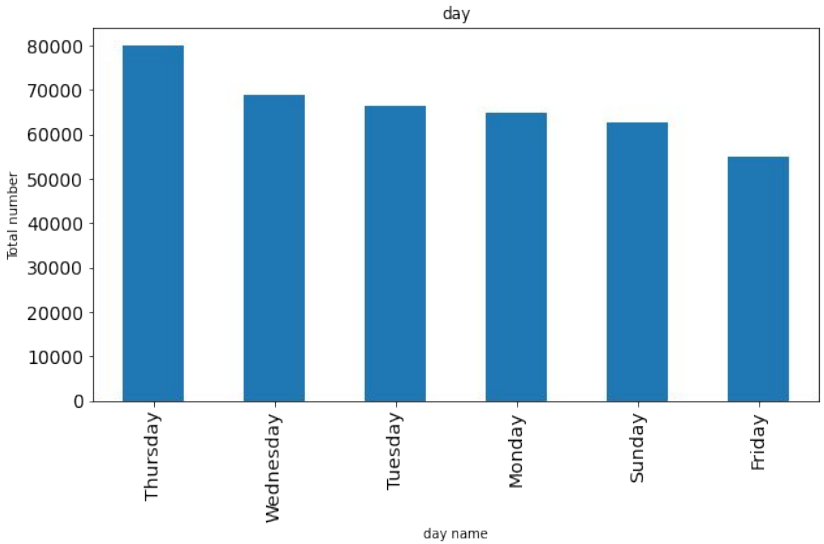


Customers from Different countries

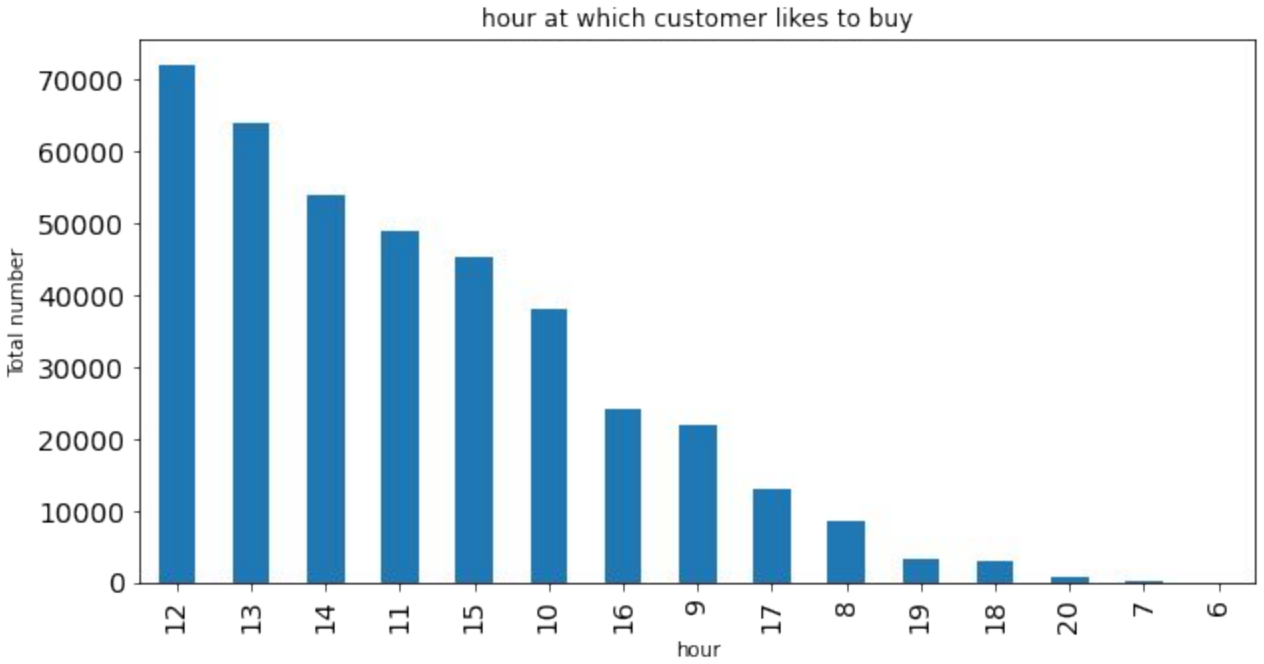


Sales in Months and days

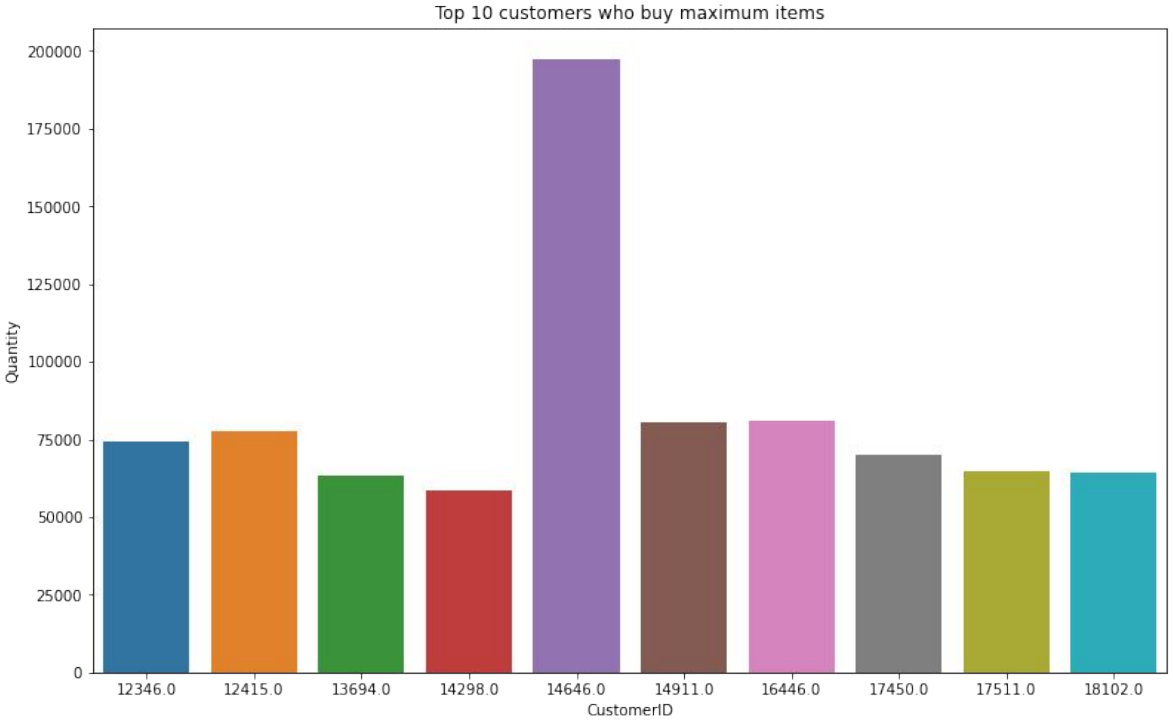




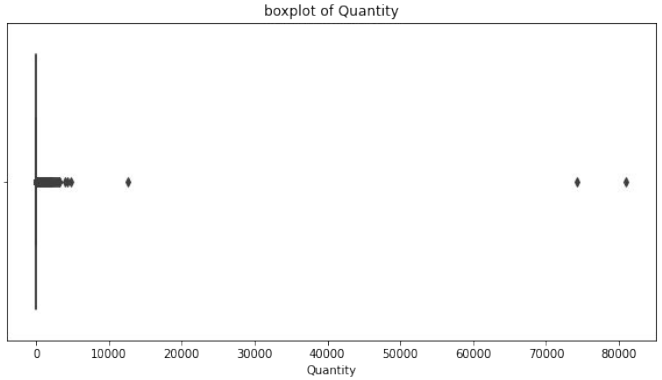
Maximum order placed

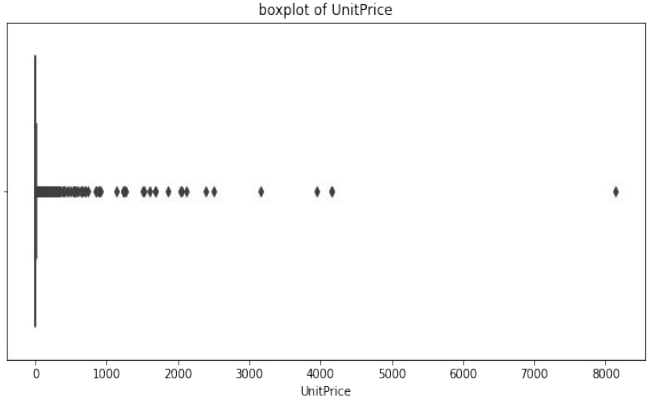


Customers (Bought the maximum quantity)



Outlier in Unit price and Quantity





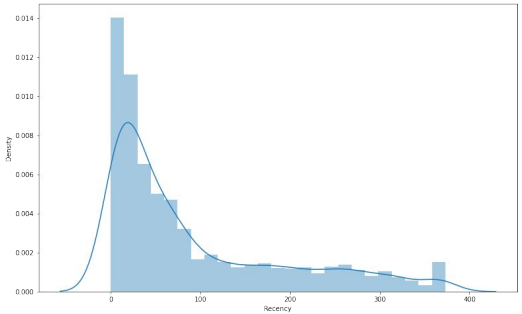
Both contain outliers.

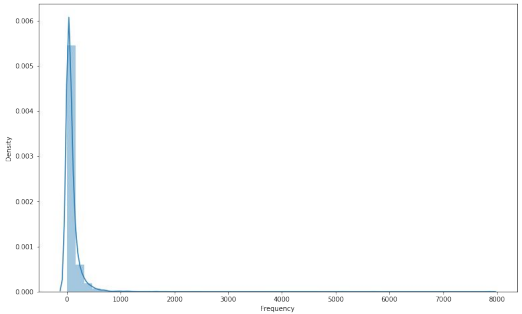
Algorithms Implementation

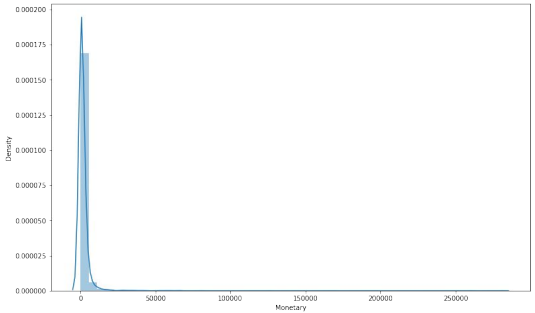
- RFM
- K-means clustering
- DBSCAN(density based spatial clustering of applications with noise)
- Hierarchical clustering

RFM(Recency, Frequency, Monetary)

1. **Recency** : It means when was the last time the customer made a purchase?
2. **Frequency** : It means how many times a customer bought a product.
3. **Monetary** : It is self explanatory of how much money did a customer spend.

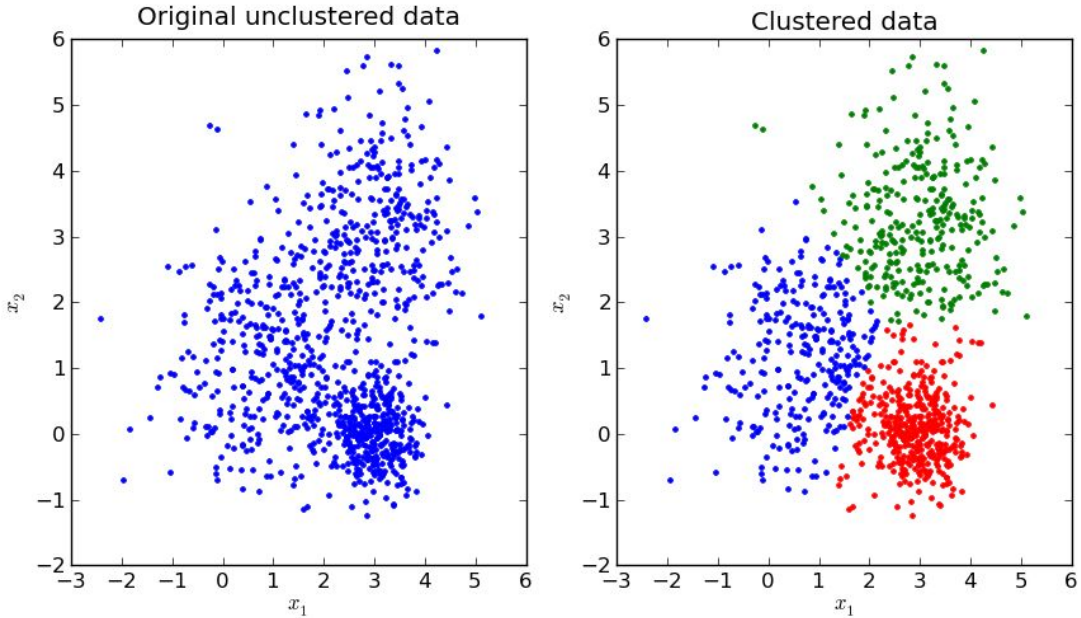


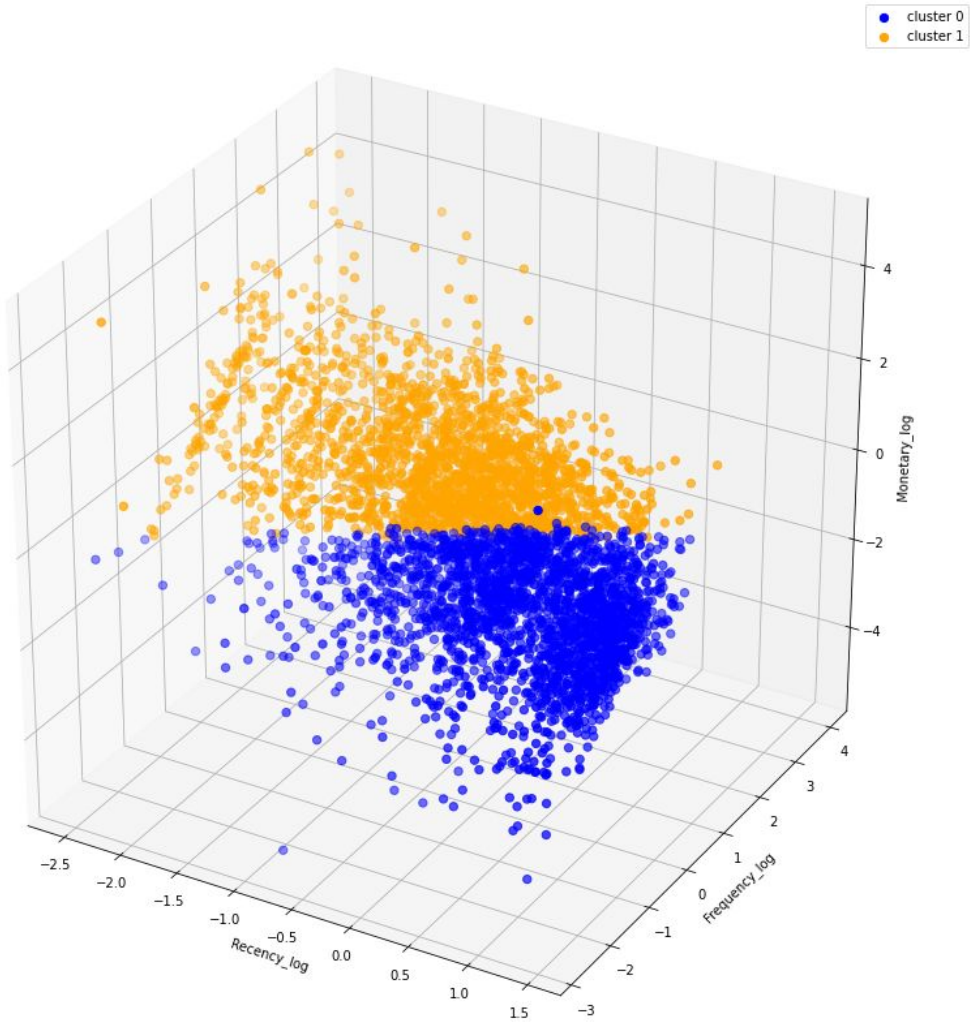




K-means Clustering

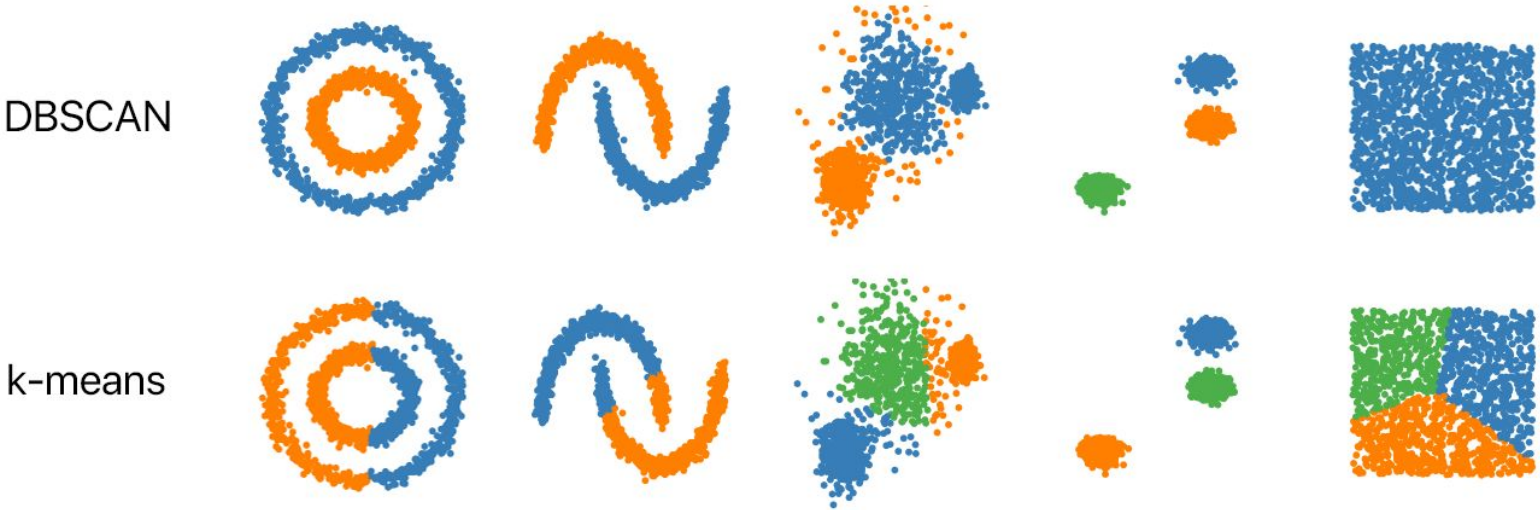
K-Means is a standard algorithm which takes the parameters and the number of clusters as inputs and partitions the data into the defined number of clusters such that the intra-cluster similarity is high.

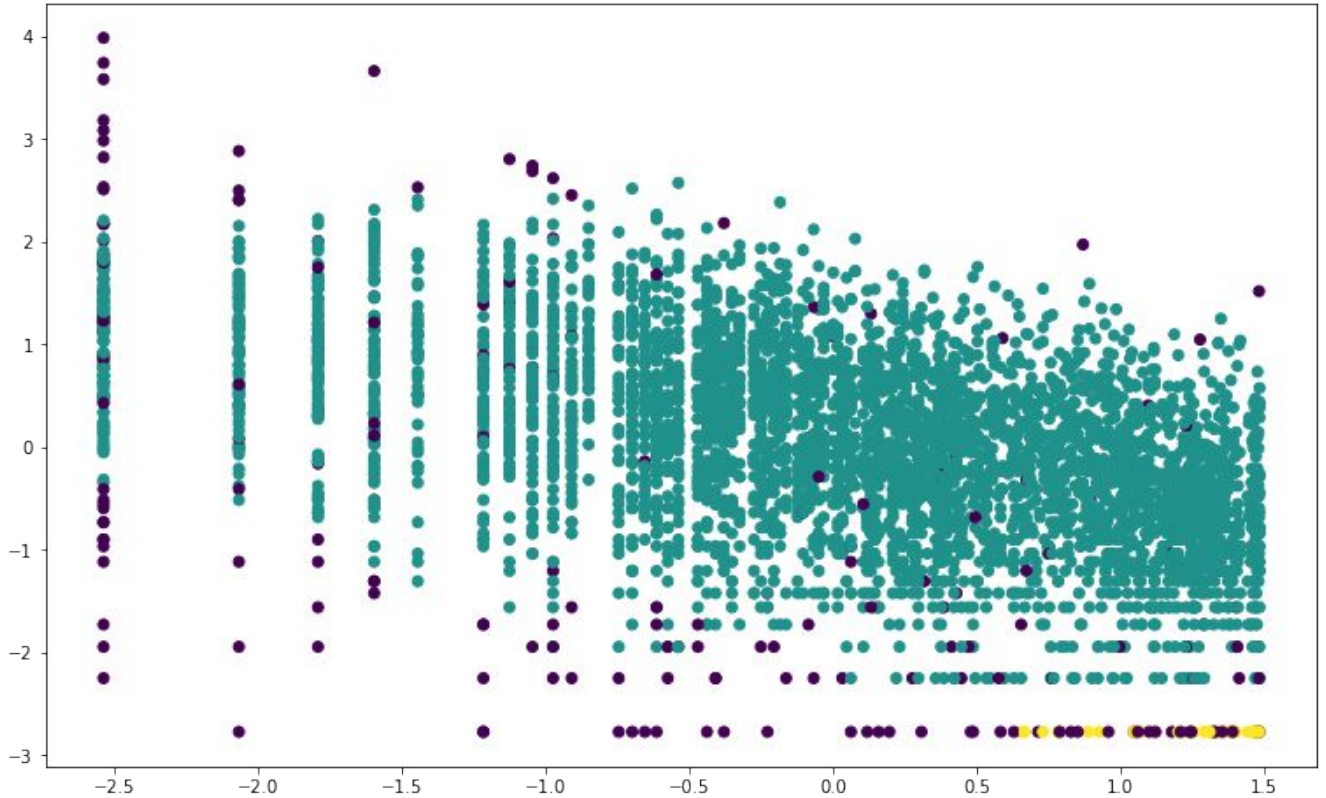



cluster 0
cluster 1
Recency_log
Frequency_log
Monetary_log

DBSCAN

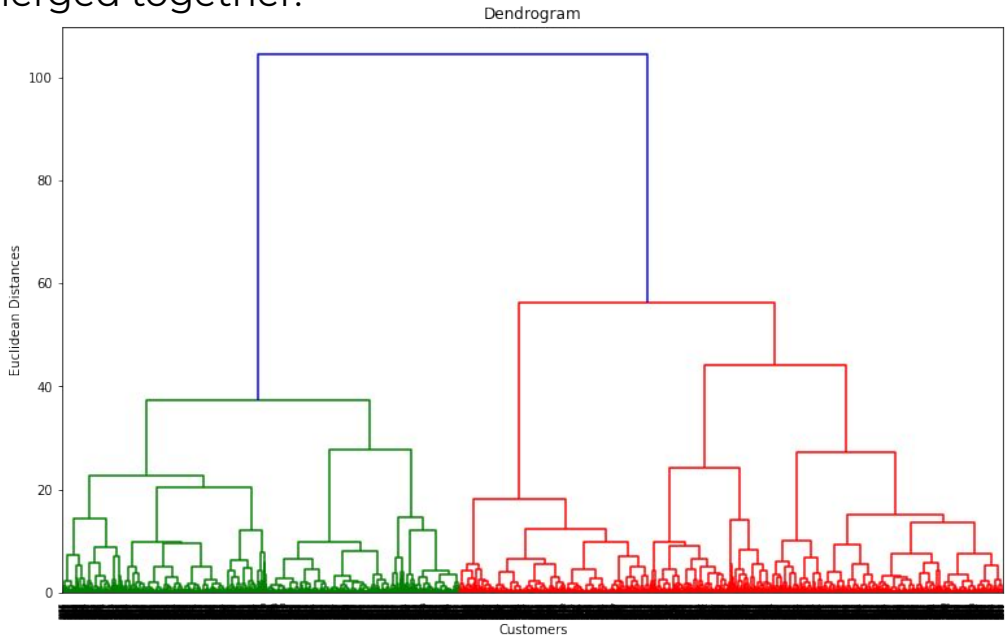
Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.




4
3
2
1
0
−1
−2
−3
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Hierarchical clustering

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together.



Challenges

- The Dataset is large.
- The project requires some domain knowledge.
- Dataset contains null values.
- Dataset also contains some canceled orders.
- Numerical features also skewed.
- Different algorithms were giving different performance.

Conclusion

After the treatment of th null values, duplicate values I did EDA on the given dataset.I applied the RFM because it is one of the best and common practice to do for the customer segmentation problems .then I apply different Clustering models on dataset like K-means clustering, DBSCAN, Hierarchical clustering. I got the following result from different models.

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with silhouette_score	RFM	2
2	K-Means with silhouette_score	RM	2
3	K-Means with silhouette_score	FM	2
4	K-Means with silhouette_score	RF	2
5	DBSCAN	RFM	3
6	DBSCAN	RM	2
7	DBSCAN	FM	3
8	DBSCAN	RF	3
9	Hierarchical clustering	RFM	2

Thank You