Customer Segmentation using RFM Analysis on Online Retail II Dataset

Group – Xi

Abstract

E-commerce sales have increased significantly in the retail industry, and retailers are competing fiercely to understand their customers and provide them with the right products and marketing strategies. We are using the Online Retail II dataset. The goal of this study is to use the Recency, Frequency and Monetary model abbreviated as RFM to analyze transactions made by customers of a UK-based online retail firm. We preprocess the data and handle data issues over the features. Using the K-Means, Agglomerative and DBSCAN clustering algorithms, we computed the RFM scores and divided the customers into different groups. The Elbow method along with silhouette score is used to determine the optimal number of clusters, and customers are assigned to one of them based on their RFM scores. We use scatter plots and box plots to visualize the clustered data in order to gain insight. Since our dataset does not have any ground truth values, we are using intrinsic metrics like Silhouette score, Calinski Harabasz score, and Davies Bouldin score to compare the KMeans, Agglomerative and DBSCAN clustering models to determine the better clustering models for customer segmentation.

Introduction

The problem under consideration belongs to the retail domain. The Center for Retail Research of Western Europe said in a statement that the combined e-commerce sales in 2015 totaled £152.20 billion throughout the UK and some EU countries. They totaled £328.91 billion by 2022 (+116.1% rise).[1](Center for Retail Research, (2022))

In recent years, the number of retail firms has increased and there is lots of competition among individual firms. In order to sustain in the market, they should properly understand their customers and serve them with the right kind of products and marketing strategies.

In our problem, we are trying to understand the transactions made by customers, recorded on an Online Transactions Processing system of a UK based online retail firm . In the data set we are making use of a couple of distinctive features such as Invoice Date, Price, Quantity, and Customer ID to build an RFM model. We are preprocessing data by dropping the null values, duplicates, and canceled orders in the dataset, along with that we are creating features like, Recency, Frequency and Monetary to perform our analysis. We have used Silhouette score and Elbow Method to determine the hyper-parameters for KMeans, Agglomerative and DBSCAN clustering algorithms.

Related Work

Apart from the RFM[3](Lourth Hallishma, 2023) model, cohort is another technique to generate groups of people who share similar preferences of purchases during a specific time period. Cohort analysis techniques can be used on transaction data to understand patterns of a group of customers over a year. Cohort analysis is different from the RFM model, Cohort analysis tracks the behavior of the group of customers over a defined time period whereas RFM model focuses on the individual customer based on their recency, frequency and monetary.

Behavior segmentation is another analysis technique where customers behave in areas other than their purchasing patterns, such as their interactions with customer service representatives, websites, and social media. In behavior segmentation we use the RFM values to categorize the purchase patterns and develop special marketing strategies based on these clusters.

Data

The Online Retail II dataset is from the UCI ML repository website(2). There are 814,213 data objects each having a total of 8 attributes. The dataset has different types of attributes varying from nominal, ordinal, categorical, and text data.

S. No	Feature	Attributes
1	Invoice	Nominal
2	StockCode	Nominal
3	Description	Nominal
4	Quantity	Ratio
5	InvoiceDate	Nominal
6	Price	Ratio
7	Customer ID	Nominal
8	Country	Nominal

Table 1 :- Features and Attributes

On the dataset, we performed the following data cleaning and feature engineering tasks.

1. We have dropped Null values from the dataset.
2. The dataset contains data for multiple European countries but 90% of the data is from UK based retailers so we have taken UKs data for the analysis.

S. No	Countries	Percentage
1	United Kingdom	0.914318
2	Germany	0.017521
3	France	0.015792
4	EIRE	0.015124
5	Spain	0.004674

Table 2 :- Top 5 Countries and their percentage of data

3. We only took positive values into consideration after eliminating all the negative values from the price and quantity column.
4. We also removed cancelled invoices from the Invoice column that have character C in the beginning and followed by invoice numbers. (Ex. C123456 is a canceled order)
5. We have deleted all the duplicates in the dataset.
6. For the analysis we have considered only 12/01/2010 to 12/01/2011 data.
7. We have created a TotalPrice feature as Quantity * Price which will be used for generating Monetary Feature.
8. To perform DBSCAN we had to copy and drop the CustomerID in the RFM dataset and perform DBSCAN clustering and then again add the CustomerID after the clusters are added to the RFM data frame.

Creating RFM Data frame:

1. Recency: is created by subtracting 12/01/2011 – latest invoice data for a given CustomerID.
2. Frequency: is created by counting the number of unique invoices for a given CustomerID.
3. Monetary: is the total sum of TotalPrice for a given CustomerID.
4. We performed Log transformation on the RFM as data is not normally distributed.

Methods

Initially we would like to create an RFM model. Where recency feature describes how recently a customer made a purchase, frequency describes how often a customer made a purchase, and monetary describes how much money a customer spends on their purchases. Once we build the RFM model we will build a KMeans, Agglomerative and DBSCAN clustering models to determine the better clustering models for customer segmentation.

In order to introduce the recency, frequency, and monetary values into the dataset for clustering, we calculated the recency value by subtracting the most recent purchase date of the customer with the highest date in the dataset. We calculated the frequency value by calculating the unique invoices generated by each customer and monetary value is calculated by summing up the multiple of the unit price and the quantity.
The RFM scores are used to cluster the customers into diverse groups using the clustering algorithms.

KMeans Clustering:
In KMeans we have used both Elbow Method and Silhouette score over a range 2 - 15 to determine best neighbor values. To determine the best cluster size, we have to consider both elbow point and the maximum silhouette score.

Agglomerative Clustering:
In Agglomerative we have compared Silhouette score over a range 2 - 15 cluster. To determine the best cluster size, we have taken the maximum silhouette score.

DBSCAN Clustering:
In DBSCAN also, we have compared Silhouette score over eps in range 0.3 to 2 in steps of 0.1. To determine the best eps values, we have taken the maximum silhouette score.
And for the minimum number of samples, we have taken twice the number of features in the dataset (R,F,M).

The optimal number of clusters is determined using the above-mentioned methods, and the customers are assigned to one of the clusters based on their RFM scores.

Metrics:
As we are performing an unsupervised machine learning task and our data does not have any ground truth values, we are using Intrinsic metrics.[7](Wong, 2022)
In order to determine the best clustering model, we have considered the following metrics.

1. Silhouette score : Values range between (-1,1) and values close to 1 potentially true classification and -1 is incorrect classification.
2. Calinski Harabasz Score : Higher the score better the classification.
3. Davies Bouldin Index : Values close to Zero are better classified.

We visualize clustered data using 2D/3D scatter plots, box plots and histograms to gain insights into the characteristics of each cluster. The analysis helped us in identifying customer segments with high and low RFM scores, and these results can be used to target specific groups with personalized marketing strategies.

Experiments and Results

Based on the RFM values we applied the KMeans, Agglomerative and DBSCAN clustering algorithms

KMeans: Hyperparameter tuning in order to get k neighbors' values, we have by iterated the k values in the range of 2 to 15 and constructed the elbow plot. Based on the elbow point we found to be the elbow point to be **six**.

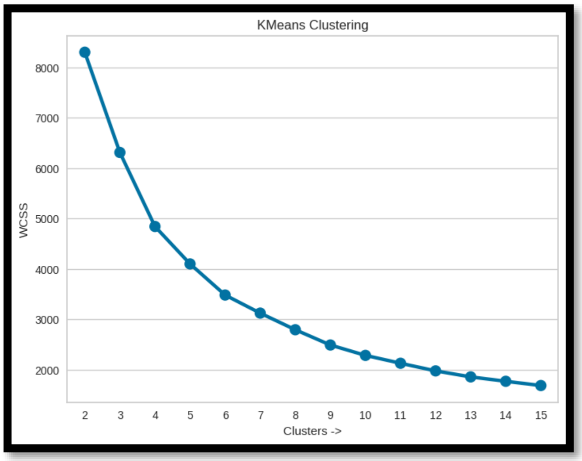

Fig 1: Elbow Plot

Cluster Size	Silhouette Score
2	0.4165
3	0.3323
4	0.3530
5	0.3266
6	0.3279
7	0.3227
8	0.3067

Table 3: Silhouette Score of KMeans

Although the Silhouette Score suggests 4 to be a better K value but by taking the elbow point into consideration and K values of 6 was chosen for KMeans and the model was built.

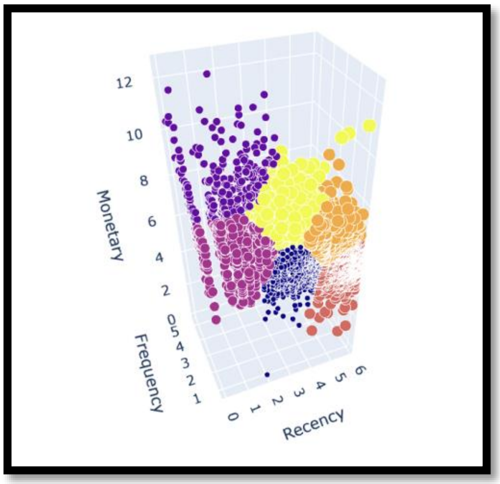

Fig 2: 3D Scatter plot of KMeans

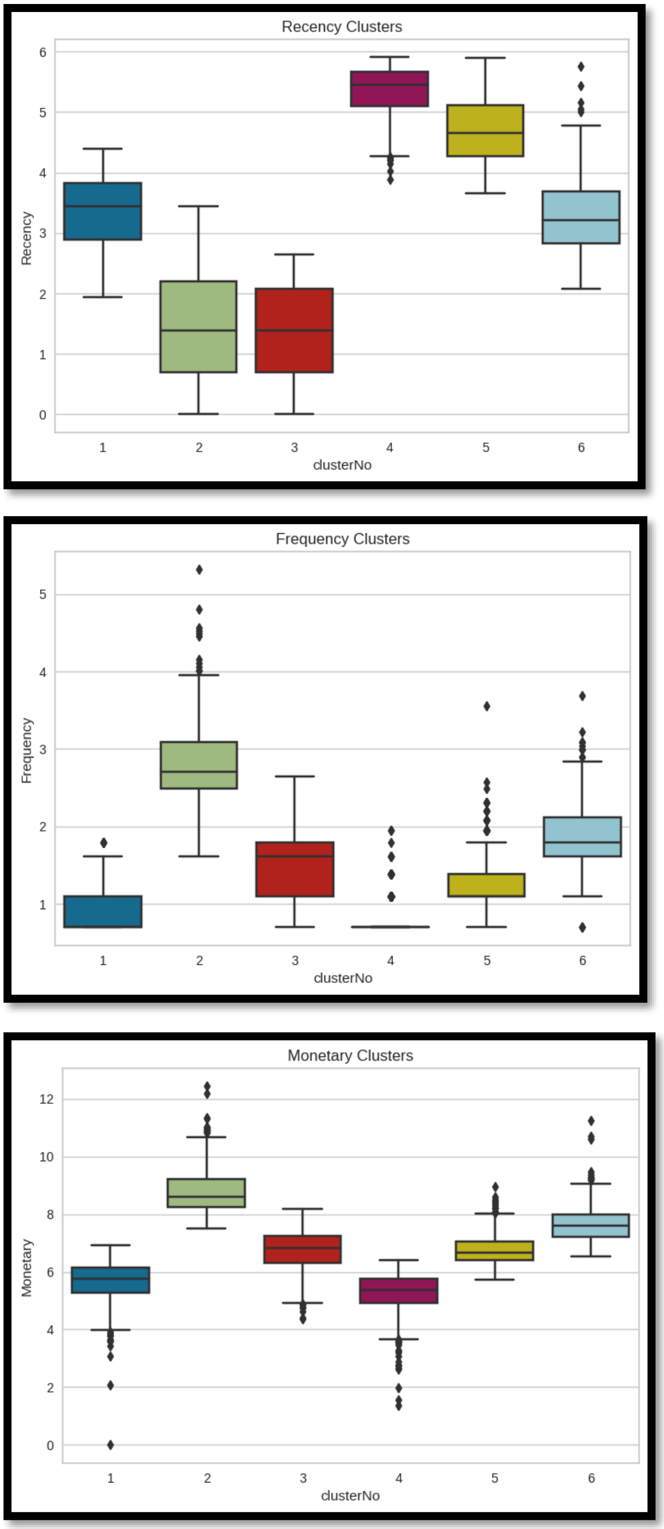

Fig 3: Box plots of KMeans

Based on these plots we can identify cluster of interest:

1. Low Frequency new Customers: 1, 2
2. Loyal Customers: 3, 4
3. High value Customers: 5, 6

Agglomerative Clustering: Hyperparameter tuning we are using complete linkage and for tuning the number of clusters and by iterating the values in the range of 2 to 15 will be applied the Silhouette Score.

Cluster Size	Silhouette Score
2	0.3273
3	0.3588
4	0.4512
5	0.4112
6	0.3675
7	0.4453
8	0.4345
9	0.4089

Table 4: Silhouette Score of Agglomerative

The Silhouette Score suggests 7 to be a better Cluster value.

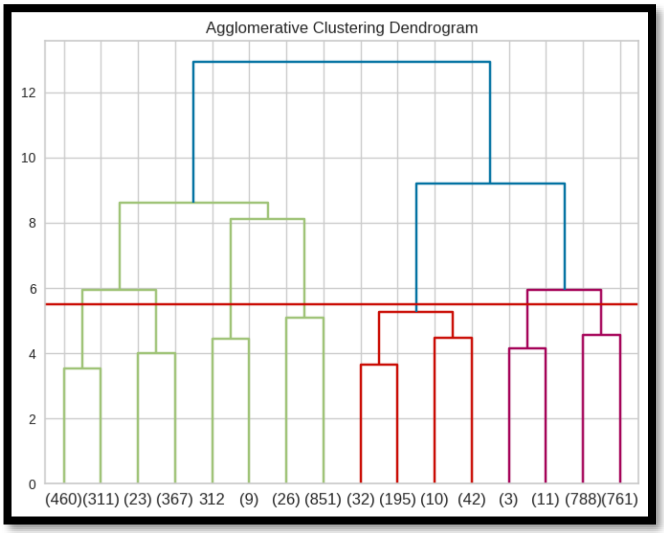

Fig 4: Dendrogram

DBSCAN:

Hyperparameter tuning for finding the eps value we have iterated the values in the range of 0.3 to 2 in steps of 0.1and applied the Silhouette Score.

eps	Silhouette Score
0.6	0.2923
0.7	0.408
0.8	0.421
0.9	0.4272
1	0.4562
1.1	0.4628
1.2	0.4553
1.3	0.4564

Table 5: Silhouette Score of DBSCAN

The Silhouette Score suggests eps of 1.1 and for the minimum number of samples, we have taken 6. (Twice the number of features used in clustering)

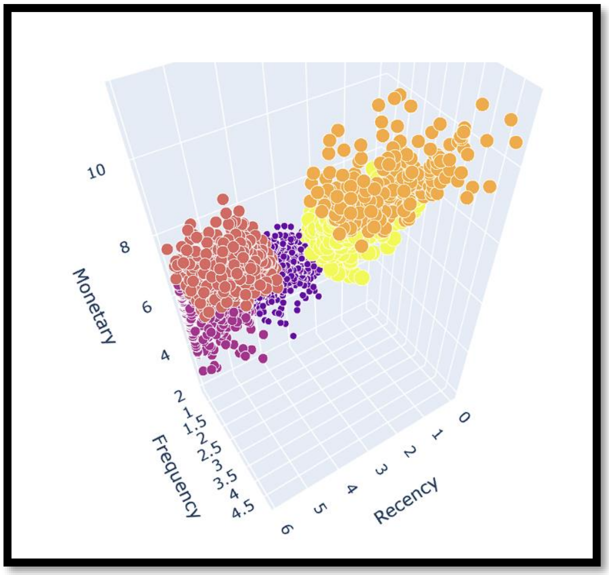

Fig 5: 3D Scatter plot of DBSCAN

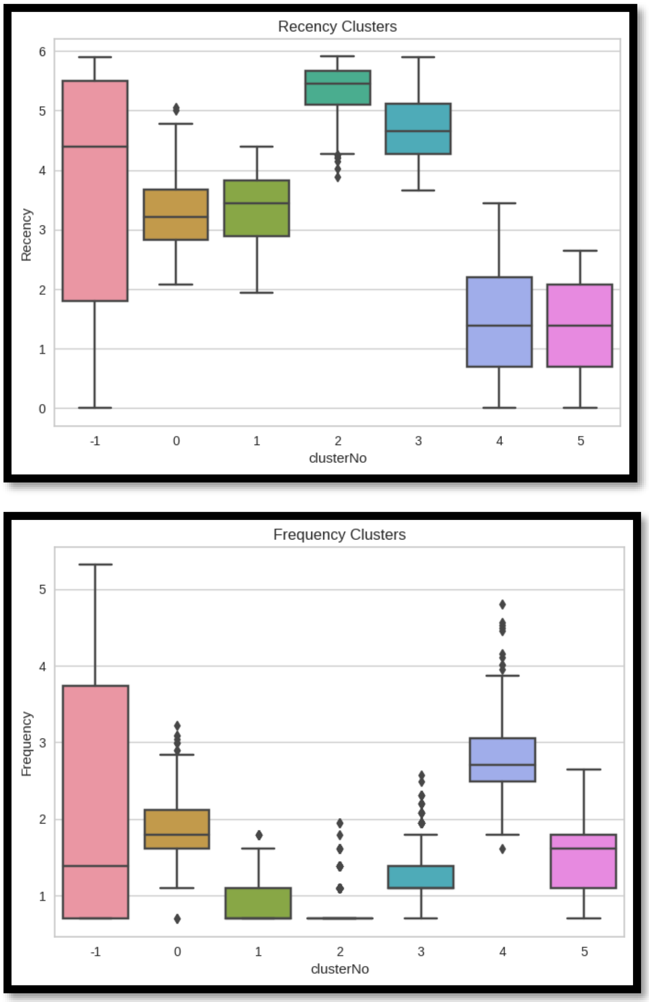

Fig 6: Box plots of DBSCAN

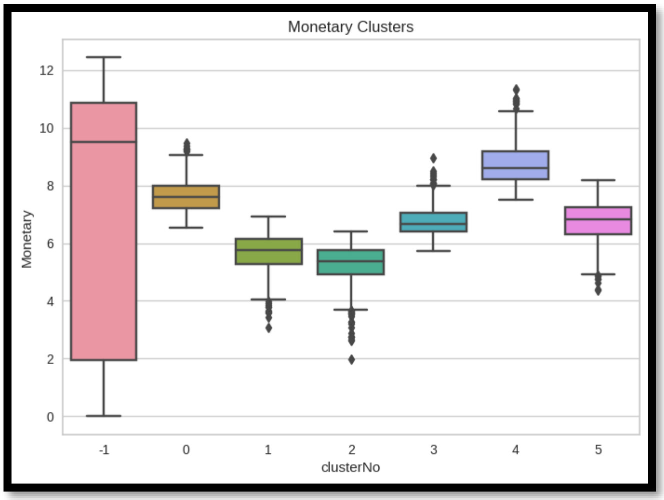

Fig 7: Box plots of DBSCAN contd.

Based on these plots we can identify cluster of interest:

1. '-1' Cluster is noise from the DBSCAN and is ignored in the analysis.
2. Low Frequency new Customers: 1, 2
3. Loyal Customers: 3, 4
4. High value Customers: 5, 6

To compare the KMeans, Agglomerative and DBSCAN clustering models the following metrics are used.

Model	Silhouette Score	Calinski Harabasz Score	Davies Bouldin Score
KMeans	0.50925	5503.43	0.77972
Agglomerative	0.44525	3011.44	0.87019
DBSCAN	0.50918	4294.76	2.46542

Table 6: Metrics comparison between models

After comparing the above metrics, Occam's Razor comes to my mind as the metrics suggest that the simplest KMeans models is the better for this RFM analysis as it has the highest silhouette score close to one, lowest Davies Bouldin score, and highest Calinski Harabasz scores. And surprisingly the DBSCAN and KMeans generated clusters are similar to each other.(after ignoring the noise in DBSCAN)

Conclusions

Currently, we have preprocessed the Data, generated the recency, frequency, and monetary features in the data frame. And built the KMeans, Agglomerative and DBSCAN models based on elbow and silhouette score. After building the model have analyzed them using three different metrics and found that KMeans is the best suited for this RFM analysis. We analyzed the K Means clusters to identify patterns of the high value customer or high frequency new customers and loyal customers.

Customer segmentation based on KMeans for clusters of interest:

1. Clusters 1 and 2 are low frequency new customers for them we could develop retention mechanisms.
2. Clusters 3 and 4 are loyal customers we could help increasing customer satisfaction programs .
3. Cluster 5 and 6 high value customers we could help them in choosing premium products and better incentive plans.

Possible additional analysis routes include customer relationship management and recommender systems for the high value customers.

References

1. Center for Retail Research. (2022). Online Retail : UK, Europe & N. America : The Centre for Retail Research Online Retail : UK, Europe & N. America : The Centre for Retail Research. Retailresearch.org. https://www.retailresearch.org/online-retail.html
2. UCI Machine Learning Repository: Online Retail II Data Set. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
3. Lourth Hallishma. (2023). Customer Segmentation Based on RFM Analysis and Unsupervised Machine Learning Technique. Communications in Computer and Information Science, 46–55. https://doi.org/10.1007/978-3-031-28183-9_4
4. McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of Business Research, 60(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015
5. Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197–208. https://doi.org/10.1057/dbm.2012.17
6. Wong, K. J. (2022, December 9). 7 Evaluation Metrics for Clustering Algorithms. Medium. https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2
7. Mandeng, J. M. (2020, April 14). Data visualization and RFM(Recency, Frequency and Monetary) analysis using Python-Customer…. Analytics Vidhya. https://medium.com/analytics-vidhya/data-visualization-and-rfm-recency-frequency-and-monetary-analysis-using-python-customer-d7e129437aac