uOttawa

CUSTOMER SEGMENTATION

Prepared for

Prof: Bisi Runsewe

Faculty of Engineering, University of uOttawa
Fundamentals/Applied Data Science

Prepared by

Ahmed Ibraheem Ali Badwy
Anas Ibrahim Ali Elbatra
Esmael Alahmady Ebrahim Ezz
Yousef Abd Al Haleem Ahmed Shindy

Jul 28,2023

Contents

1 Abstract:

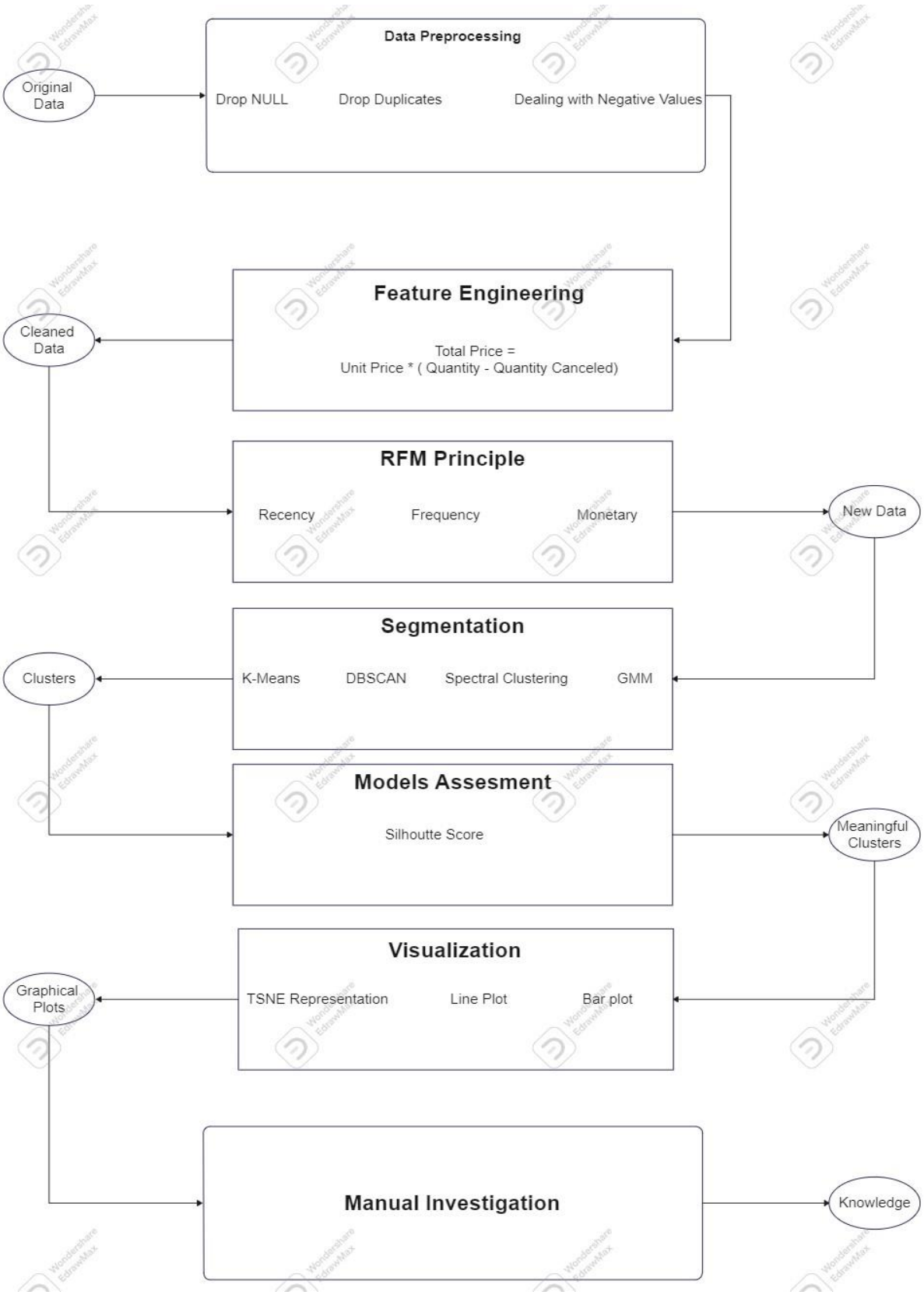
Customer segmentation is a critical part of modern business strategy. It allows companies to understand and meet the diverse needs of their customers effectively. This project will explore various customer segmentation techniques to divide the customer population into distinct groups based on shared characteristics, preferences, and behaviors. By using advanced data analysis and machine learning algorithms, we hope to uncover valuable insights that can drive personalized marketing strategies, enhance customer satisfaction, and optimize resource allocation.

2 Introduction:

The primary goal of this project is to identify distinct customer segments within the dataset based on their transactional behavior and characteristics. The dataset went through various steps of preprocessing and cleaning to make sure it was ready for analysis. We explore the use of RFM (Recency, Frequency, Monetary) technique combined with various clustering algorithms.

The data preprocessing stage involved several critical steps to enhance data quality and prepare it for analysis. Missing values were handled appropriately, and any inconsistent or erroneous entries were carefully addressed. The RFM technique focuses on three fundamental aspects of customer behavior: Recency, Frequency, and Monetary value. Recency refers to the time since a customer's last purchase, Frequency indicates the number of transactions made, and Monetary represents the total amount spent by the customer. To effectively identify customer segments, we employed three different clustering algorithms: K-means, DBSCAN, Spectral Clustering and Gaussian Mixture Model (GMM) then compared their performance to get the best approach.

3 System Architecture



4 Data

4.1 Data Description

The data is a transactional dataset from an e-commerce platform. Each row in the dataset represents a specific transaction with details about the products purchased and the associated customer information.

4.2 Data Understanding

1. The data contains 541909 rows with 8 features:

InvoiceNo: A unique identifier for each transaction or invoice.

StockCode: A code or identifier for the product or item purchased.

Description: A description for the product or item purchased.

Quantity: The quantity of the product purchased in that particular transaction.

InvoiceDate: The date and time of the transaction.

UnitPrice: The price of a single unit of the product.

CustomerID: A unique identifier for each customer, indicating the customer who made the purchase.

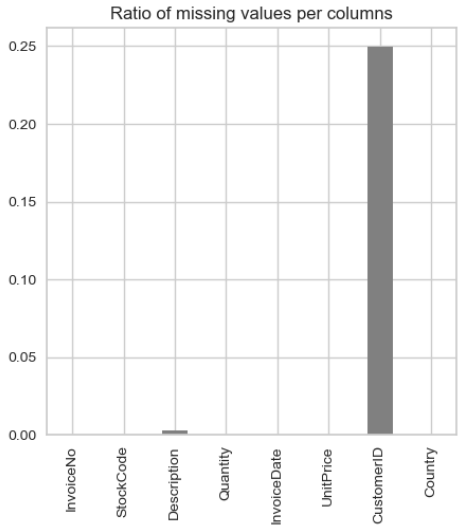
Country: The country where the transaction took place or the customer is located.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

4.3 Data Preprocessing

1. During the data understanding phase, an examination of the dataset revealed that approximately 25 percent of the data contains missing values. These missing values need to be addressed to ensure the data's integrity and reliability for further analysis and modeling.



2. The dataset contains over 5000 duplicated values. Identifying and handling these duplicates will be necessary to avoid potential biases and ensure the accuracy of analyses and models based on the dataset.

```
Dupplicate entries: 5225
```

3. The dataset includes customers from 37 different countries worldwide. Notably, the majority of customers are located in the United Kingdom, with a significant difference in the number of customers compared to other countries.

	Country	CustomerID
35	United Kingdom	3950
14	Germany	95
13	France	87
30	Spain	31
3	Belgium	25

4. The dataset contains negative values in the "Quantity" column, with the minimum quantity of products purchased being negative.

	Quantity	UnitPrice	CustomerID
count	401604.000000	401604.000000	401604.000000
mean	12.183273	3.474064	15281.160818
std	250.283037	69.764035	1714.006089
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13939.000000
50%	5.000000	1.950000	15145.000000
75%	12.000000	3.750000	16784.000000
max	80995.000000	38970.000000	18287.000000

5. After analyzing the data, we discovered that approximately 16% of the Quantity column contains negative values, With the presence of a "C" before some Invoice numbers. These negative quantities indicate canceled orders. However, not all canceled transactions have an exact counterpart in the database, especially when there are discounts involved.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	12/1/2010 9:41	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	12/1/2010 9:49	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	12/1/2010 10:24	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	12/1/2010 10:24	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	12/1/2010 10:24	0.29	17548.0	United Kingdom

6. To gain a deeper understanding, we decided to exclude the discount values and reevaluate whether each canceled transaction has a corresponding positive quantity counterpart. During this process, we observed that customers can also cancel only a portion of their original transactions, which is a reasonable occurrence.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
14498	C537597	D	Discount	-1	12/7/2010 12:34	281.00	15498.0	United Kingdom
17612	537771	21232	STRAWBERRY CERAMIC TRINKET BOX	24	12/8/2010 12:29	1.25	15498.0	United Kingdom
17613	537771	22174	PHOTO CUBE	48	12/8/2010 12:29	1.48	15498.0	United Kingdom
17614	537771	21524	DOORMAT SPOTTY HOME SWEET HOME	20	12/8/2010 12:29	6.75	15498.0	United Kingdom
17615	537771	48173C	DOORMAT BLACK FLOCK	10	12/8/2010 12:29	6.75	15498.0	United Kingdom

7. Based on this new piece of information, we will remove all counterpart rows and create a new column named "QuantityCanceled" to store the negative values from the Quantity column in a positive format. This step allows us to maintain a cleaner dataset that properly reflects canceled orders and their corresponding quantities.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
884	536488	22960	JAM MAKING SET WITH JARS	8	12/1/2010 12:31	4.25	17897.0	United Kingdom
939	C536506	22960	JAM MAKING SET WITH JARS	-6	12/1/2010 12:38	4.25	17897.0	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	QuantityCanceled
884	536488	22960	JAM MAKING SET WITH JARS	8	12/1/2010 12:31	4.25	17897.0	United Kingdom	6

5 Feature Engineering

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	QuantityCanceled	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	35	0	15.30
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	35	0	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	35	0	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	35	0	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	35	0	20.34

During the feature engineering stage, we created a new column called "Total Price" using the following equation:

Total Price = Unit Price * (Quantity - Quantity Canceled)

This equation calculates the total price of each transaction by multiplying the unit price with the difference between the original quantity and the quantity that was canceled. The "Quantity Canceled" represents the positive counterparts of the canceled quantities, which we derived earlier. By applying this equation, we were able to compute the total price for each transaction, taking into account the canceled quantities and their respective unit prices. This new column provides valuable information for analyzing and understanding the financial aspect of the data.

6 RFM

RFM (Recency, Frequency, Monetary) analysis is a widely used technique in marketing and customer segmentation to evaluate the behavior of customers based on their historical transaction data. The purpose of this analysis is to categorize customers into segments, allowing businesses to target them more effectively and tailor their marketing strategies accordingly.

1. Recency (R):

   Recency refers to the time since the customer's last transaction. We calculated the difference between the most recent transaction in the dataset and the most recent transaction the customer made. Customers with more recent transactions are considered more engaged and are assigned a higher recency score. On the other hand, customers with longer intervals since their last purchase receive lower recency scores.

2. Frequency (F):

   Frequency represents the number of transactions made by each customer within a specific period. By counting the total number of transactions for each customer, we identified the most active and loyal customers who frequently engage with our business. Customers with a higher frequency score are those who make more purchases, indicating their loyalty and interest in our products or services.

3. Monetary Value (M):

   Monetary value relates to the total amount of money spent by each customer during their transactions. We computed this by summing the total monetary value of all purchases made by each customer. Customers with a higher monetary value score are those who have made significant purchases, contributing significantly to our revenue.

4. RFM Segmentation:

   To create meaningful customer segments, we combined the individual R, F, and M scores into a single RFM score for each customer. The RFM score is usually represented as a three-digit number, where each digit corresponds to the respective R, F, and M scores.

5. Interpretation of Segments:

   The RFM segments provide valuable insights into customer behavior and can be interpreted as follows:

   High-Value Customers: Customers with high RFM scores, indicating recent, frequent, and high-spending behavior. These are the most valuable customers and should be prioritized for special offers and personalized engagement.

   Loyal Customers: Customers with high recency and frequency scores but relatively lower monetary value. These customers are engaged and frequent buyers, but their spending potential can be increased through targeted marketing efforts.

   Recent Customers: Customers with high recency scores but lower frequency and monetary scores. These are customers who have made recent purchases and might need incentives to become more loyal.

At-Risk Customers: Customers with low recency and frequency scores but significant monetary value. These customers were once active and high-spending but have recently decreased their engagement. Retention strategies should be applied to prevent their churn.

Inactive Customers: Customers with low RFM scores across all factors. These customers have not engaged recently, making them less likely to return. Re-engagement campaigns may be initiated to win them back.

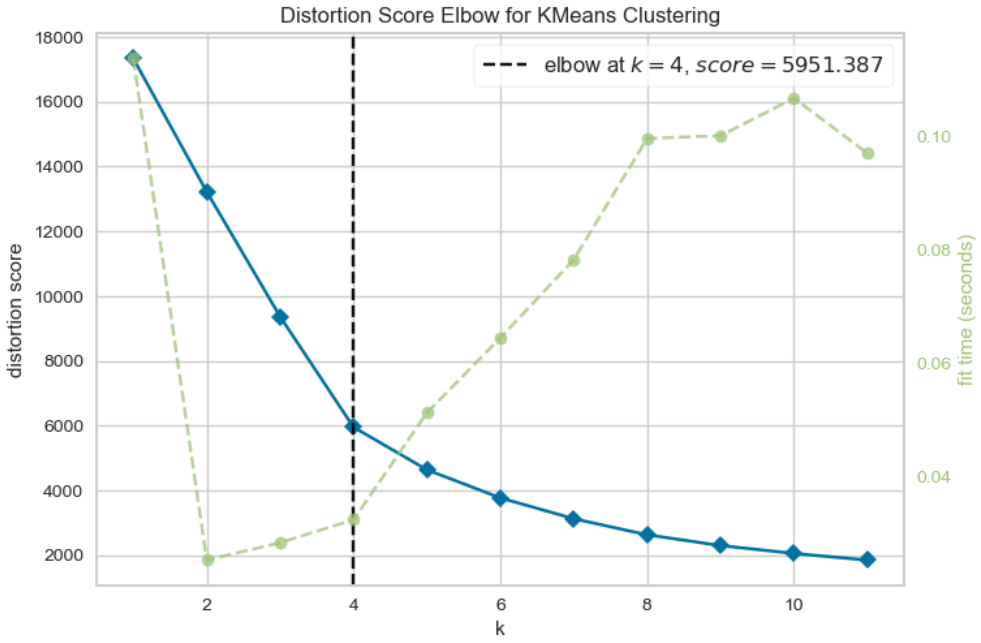
CustomerID	min_recency	max_recency	frequency	monetary_value
17850	372.0	373.0	34	5322.84
13047	56.0	373.0	9	3105.70
12583	2.0	373.0	15	6690.18
13748	95.0	373.0	5	948.25
15100	333.0	373.0	3	843.15
...	...	...	...	...
13436	1.0	1.0	1	196.89
15520	1.0	1.0	1	343.50
13298	1.0	1.0	1	360.00
14569	1.0	1.0	1	227.39
12713	0.0	0.0	1	794.55

Segment	RFM	Description	Marketing
Best Customers	111	Bought most recently and most often, and spend the most	No price incentives, new products, and loyalty programs
Loyal Customers	X1X	Buy most frequently	Use R and M to further segment
Big Spenders	XX1	Spend the most	Market your most expensive products
Almost Lost	311	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Customers	411	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Cheap Customers	444	Last purchased long ago, purchased few, and spent little	Don't spend too much trying to re-acquire

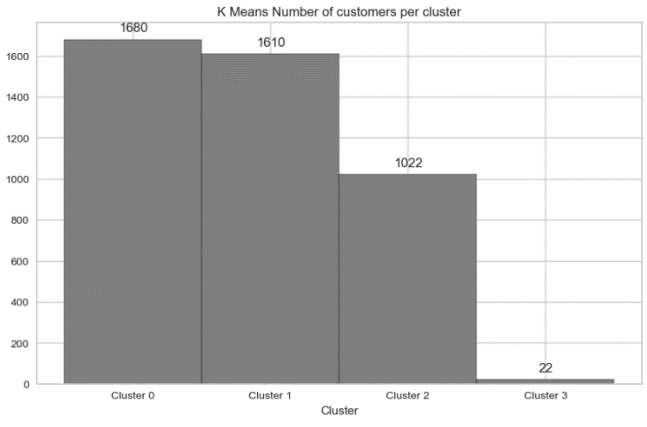
7 Modeling

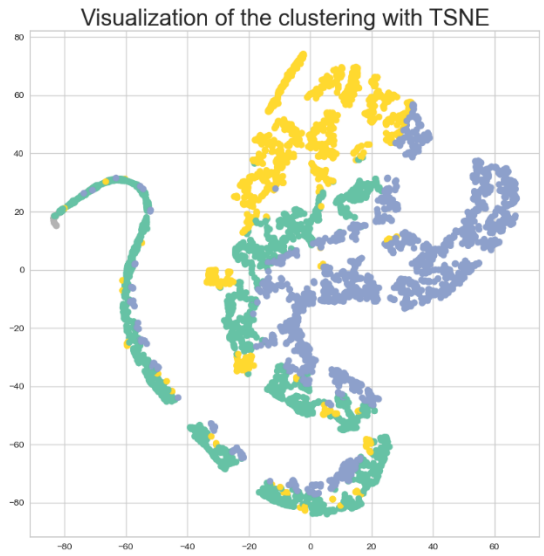
7.1 Kmeans

We used Kmeans and elbow method to determine the best K value for clustering our customers. From the figure we get the best K equal 4



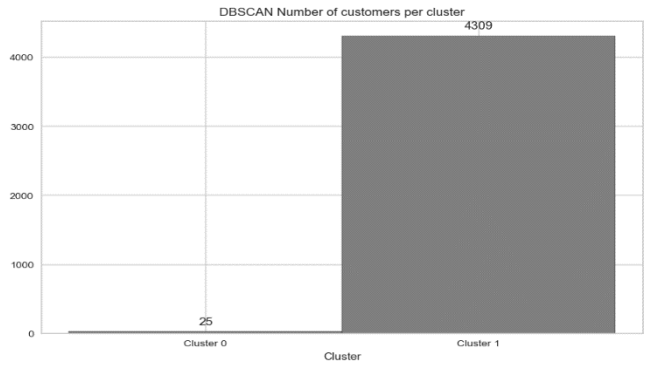
Then we fit the kmeans over the data to get the clusters and plot the TSNE and bar graph representations.

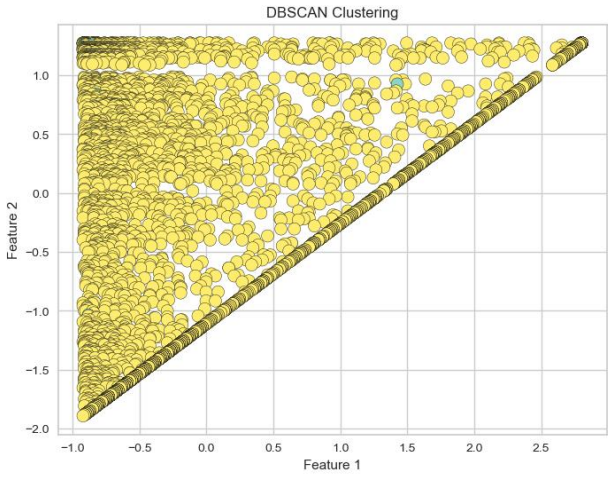




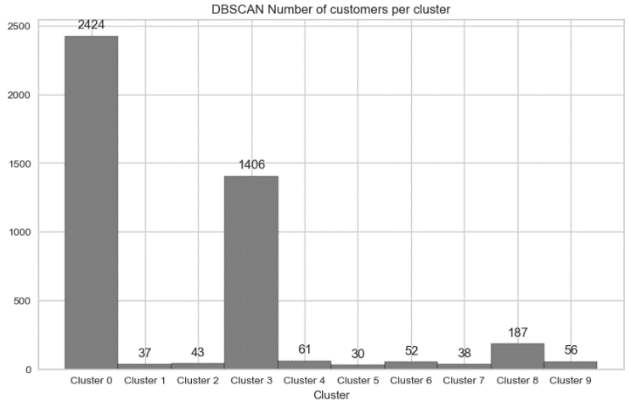
7.2 DBSCAN

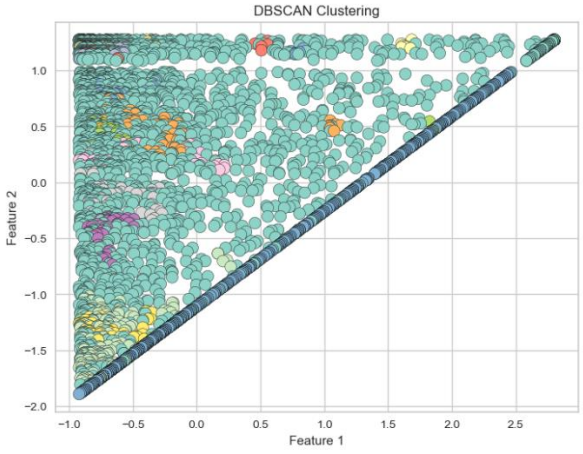
Then we used DBSCAN and fit the data and let It decide the best number of clusters for a given parameters. It clusters almost all the data as noise.





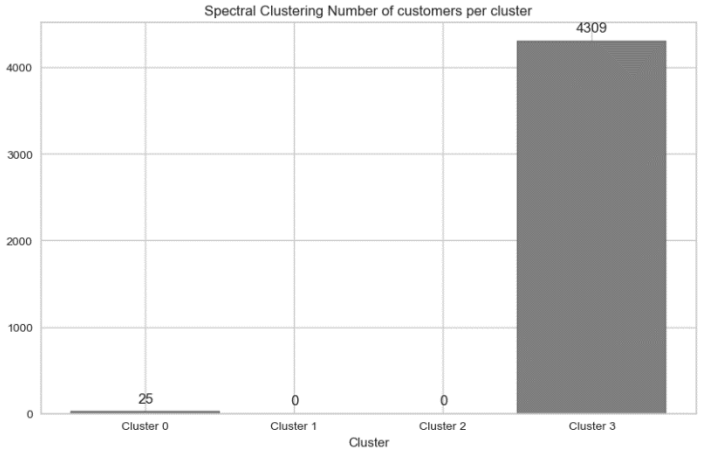
Then we fine tune the parameters to get better results, we get more clusters.

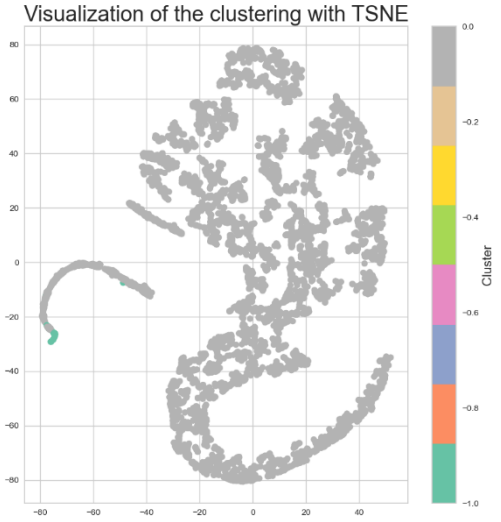




7.3 Spectral Clustering

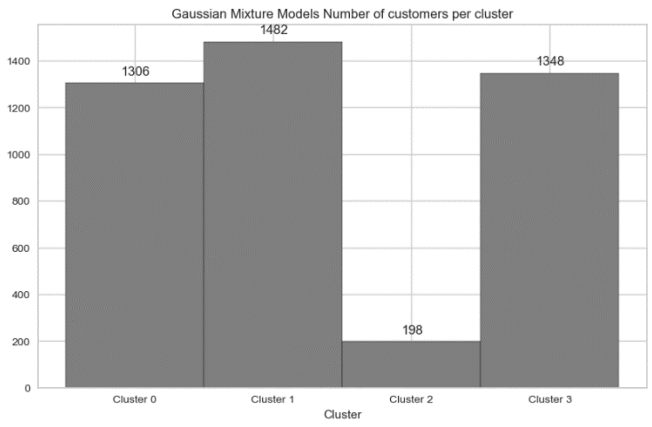
We used Spectral clustering to try to get better performance. We did get high silhoutte accuracy but in representing the TSNE plot it appears not very result

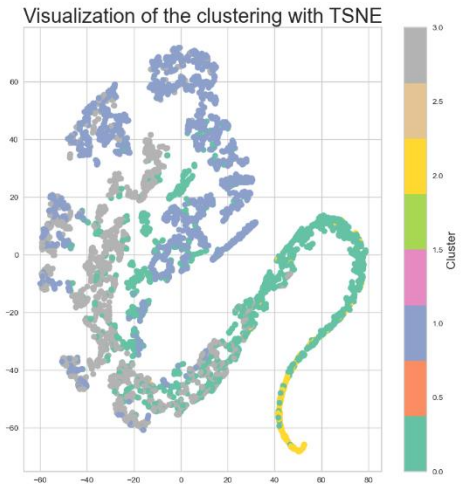




7.4 Gaussian Mixture Model

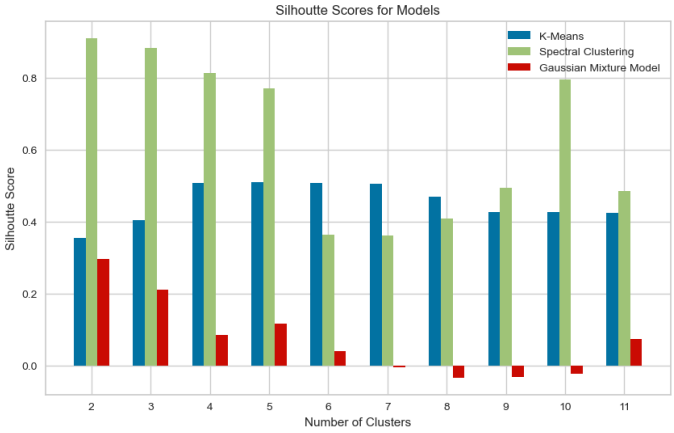
The last model we use is GMM, this gives us not the best silhouette score but it represents well in the TSNE visualization





8 Performance Evaluation

We plot the silhouette score for each K from 2 to 11 to visualize the best score



We found that the spectral clustering has the highest silhouette score In K equal 4 followed by the K-Means then Gaussian Mixture Model. However, despite its high silhouette score, the TSNE plots and resulting clusters generated by spectral clustering did not show satisfactory performance or clear separations between clusters.

9 Summary and Conclusion

In this customer segmentation project, we conducted a thorough analysis of our transaction data to gain valuable insights into customer behavior and preferences. By employing various clustering techniques, including K-Means, DBSCAN, Spectral Clustering, and GMM, we successfully divided our customer base into distinct segments based on their Recency, Frequency, and Monetary (RFM) values. The results of this analysis offer critical insights for our business strategy and marketing efforts.

10 References

[1] https://www.kaggle.com/datasets/carrie1/ecommerce-data

[2] https://www.putler.com/rfm