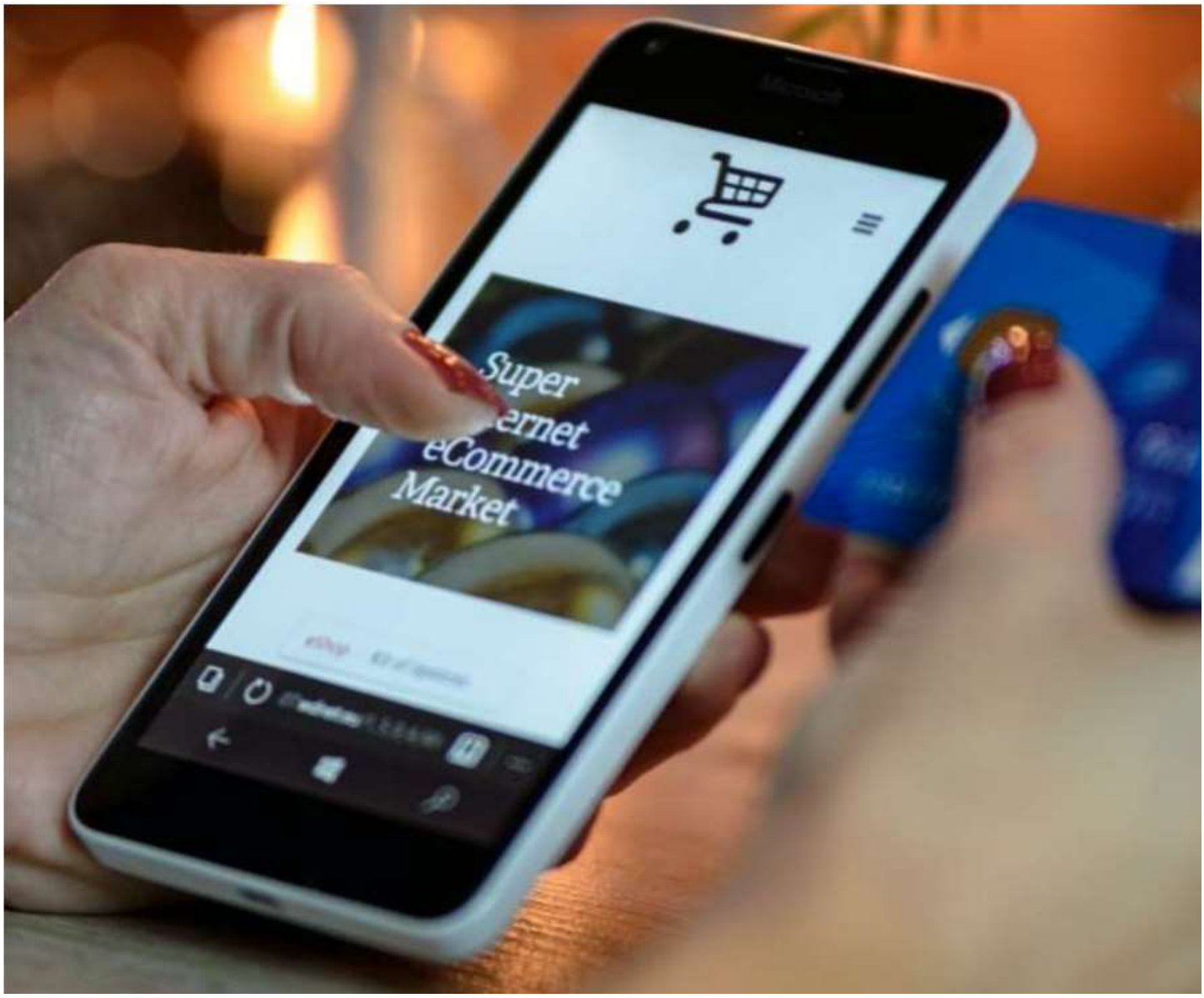


Customer Segmentation Project

Capstone Project 2

Gokul venugopal

8/15/23

DS AI INTELLIPAAT

Table of Contents

Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm. This project undertakes to review sales records with a view to provide useful insights to the company and also segment the customers based on their purchasing behaviour.

Project Objective

An online retail store is trying to understand the various customer purchase patterns for their firm. The task is to come up with useful insights to the company and also to segment the customers based on their purchasing behaviour.

Data Description

The walmart.csv contains 5,41,909 rows and 8 columns.

Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

From the given data set of the company, it is observed that the data consist of 5 Lakh forty-one thousand nine hundred and nine (5,41,909) records with 8 features as follows.

Stores: There are 25900 unique transaction entries with below information.

- InvoiceNo (Unique Number to identify a single transaction)
- StockCode (Unique Number to identify a product)
- Description (Description of Product)
- Quantity (Number of Products ordered)
- InvoiceDate (Date on which order transaction was made)
- UnitPrice (Price of single product)
- CustomerID (Unique Number to identify a customer)
- Country (Country from which order transaction was made)

Data Preprocessing Steps and Inspiration

The Preprocessing steps included the following steps:

Step 1: Load Data
Step 2: Perform Exploratory Data Analysis

a. Check number of records and its distribution
b. Check Data types
c. Check for missing data, invalid entries and duplicates
d. Create new features which can be useful for the model using RFM Analysis.
e. Check for outliers that are known to distort

Step 3: Model Building, two approaches for customer segmentation using unsupervised machine learning model.

a. KMEANS

b. DBSCAN

Step 4: Select appropriate Number of Clusters using Elbow method and Silhouette score.

Step 5: Segment the customers based on the selected Number of Clusters.

Choosing the Algorithm for the Project

Model Selection

Examination of the entries shows unsupervised data and need of feature engineering for creating better model. An unsupervised Machine Learning model (KMeans, DBscan) will be employed for the clustering.

KMeans

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K-center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Assumptions of K-means

- Limited to spherical shaped clusters

As K-means calculates distance from centroid, it forms a spherical shape. Thus, it cannot cluster complicated geometrical shape.

- Size of clusters

This algorithm considers only distance, thus does not account for clusters with different sizes or densities. It assumes that features within a cluster have equal variance.

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.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters:

- minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.
- eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.

Model Technique and Evaluation

Model Design and Analysis

RFM analysis is a customer segmentation technique commonly used in marketing and retail to categorize customers based on their purchasing behaviour. "RFM" stands for Recency, Frequency, and Monetary Value. These three key metrics are used to gain insights into customer behaviour and create distinct segments for targeted marketing and business strategies.

- Recency (R): Recency refers to how recently a customer made a purchase. It's a measure of how long it has been since the customer's last transaction.

- Frequency (F): Frequency measures how often a customer makes purchases within a specific timeframe.

- Monetary Value (M): Monetary value represents the total amount of money a customer has spent over a specific period.

Model Approach

- Extract and preprocess the relevant data for RFM analysis. Calculate recency, frequency, and monetary values for each customer based on their purchasing behavior.
- Since KMeans is distance-based, it's important to scale the features. Standardize the RFM values to have a mean of 0 and a standard deviation of 1. This ensures that no single feature dominates the clustering process.
- Decide how many clusters we want to divide our customers into. We will be using techniques like the "Elbow Method" or "Silhouette Score" to determine the optimal number of clusters.
- Use the scaled RFM values as input to the KMeans algorithm. Initialize KMeans with the chosen number of clusters (K).
- Fit the KMeans model to the scaled RFM data. The algorithm will assign each customer to one of the K clusters.

- Analyze the cluster assignments and the centroids of each cluster. This will give us insights into the characteristics of each segment, allowing us to interpret and label the clusters.
- Create visualizations (like 3d plots) to help us understand the separation between clusters.

Model Evaluation

The Elbow Method and Silhouette Score are both techniques used to evaluate the appropriate number of clusters (K) in unsupervised machine learning, particularly in algorithms like KMeans.

- **Elbow Method**:

The Elbow Method is a graphical approach to determine the optimal number of clusters in a dataset.

It involves plotting the sum of squared distances (inertia) between data points and their cluster's centroid for different values of K.

As K increases, the inertia typically decreases because each data point is closer to its cluster's centroid. However, at some point, adding more clusters may not significantly reduce inertia.

The "elbow point" on the graph is where the reduction in inertia starts to slow down. This point indicates a good balance between reducing intra-cluster distance and avoiding overfitting.

- **Silhouette Score**:

The Silhouette Score is a metric that measures the quality of clustering.

For each data point, it calculates how similar the point is to its own cluster (a) compared to other clusters (b).

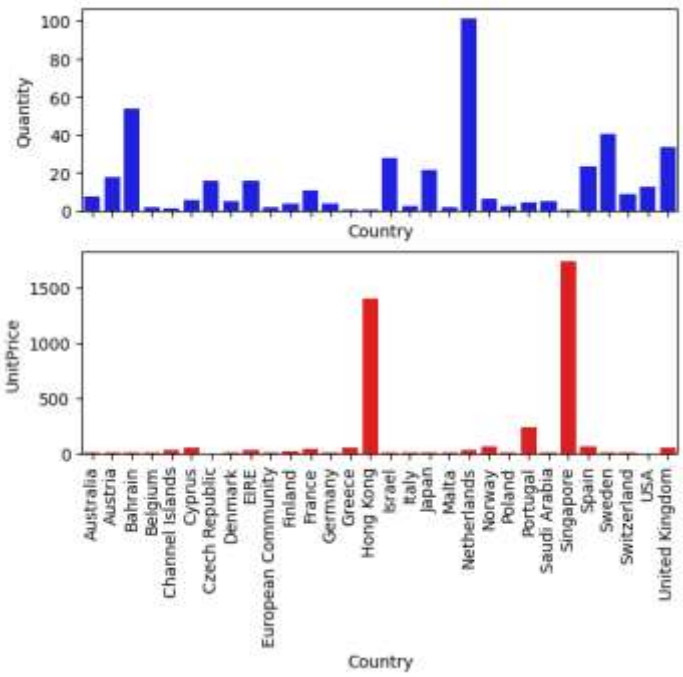
The Silhouette Score ranges from -1 to 1. A high score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

A high average Silhouette Score across all data points suggests that the chosen number of clusters is appropriate.

Inferences from the Project

1. Findings from the Dataset.

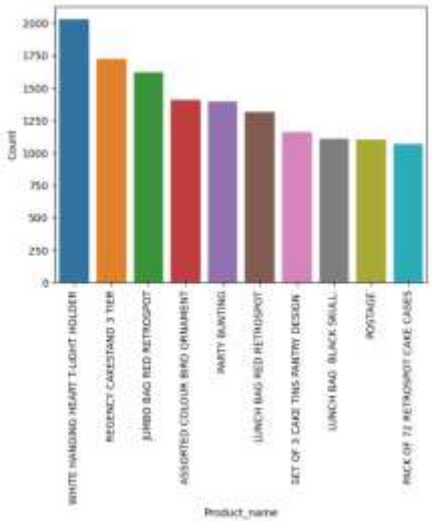
Graph of Unit Price and Quantities of Cancelled Orders



Observation:

- **Netherlands, Bahrain and Sweden have higher cancelled orders in terms of quantities.**
- **Singapore and Hong Kong have costly cancelled orders.**

Top 10 Selling Products

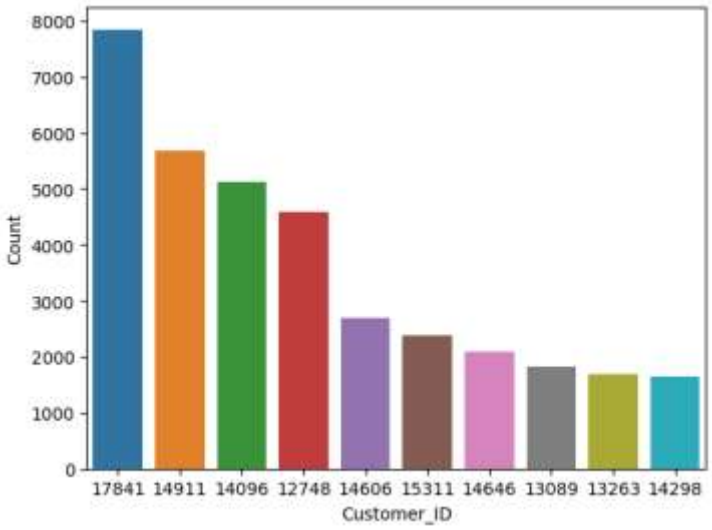


Product_name	Count
WHITE HANGING HEART T-LIGHT HOLDER	2028
REGENCY CAKESTAND 3 TIER	1724
JUMBO BAG RED RETROSPOT	1618
ASSORTED COLOUR BIRD ORNAMENT	1408
PARTY BUNTING	1397
LUNCH BAG RED RETROSPOT	1316
SET OF 3 CAKE TINS PANTRY DESIGN	1159
LUNCH BAG BLACK SKULL.	1105
POSTAGE	1099
PACK OF 72 RETROSPOT CAKE CASES	1068

Observations:

- **WHITE HANGING HEART T-LIGHT HOLDER is the Highest selling product**

Most Frequent Customers

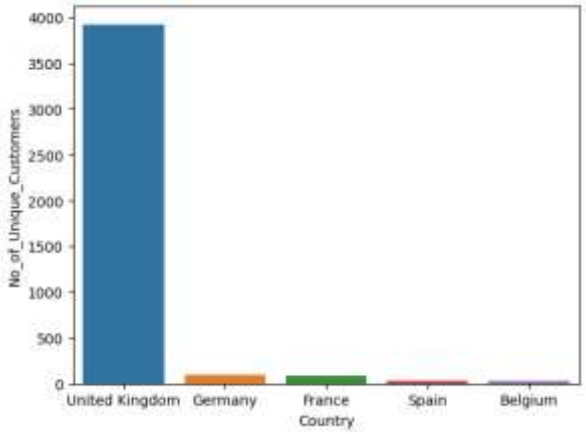


Customer_ID	Count
17841	7847
14911	5677
14096	5111
12748	4596
14606	2700
15311	2379
14646	2080
13089	1818
13263	1677
14298	1637

Observations:

- **Customer 17841 is the Most involved Customer.**

Top 5 countries with count of unique customers

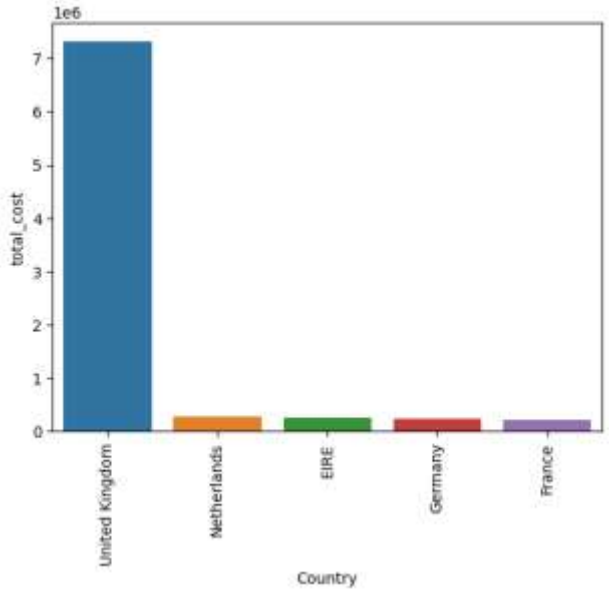


Country	No_of_Unique_Customers
United Kingdom	3921
Germany	94
France	87
Spain	30
Belgium	25

Observation

- **United Kingdom has the highest number of customers**

Top 5 Countries based Total sales

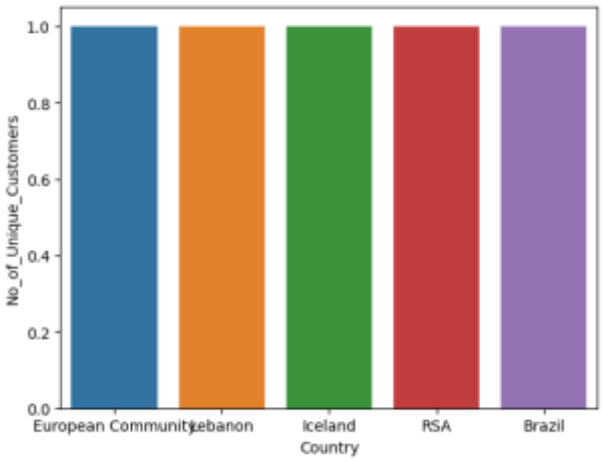


Country	total_cost
United Kingdom	7308391.554
Netherlands	285446.340
EIRE	265545.900
Germany	228867.140
France	209024.050

Observation:

- **UK has the highest total sales**

Top 5 Countries with least number of Customers

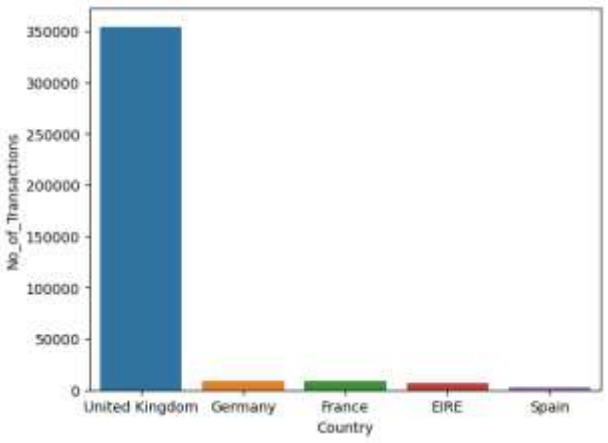


Country	No_of_Unique_Customers
European Community	1
Lebanon	1
Iceland	1
RSA	1
Brazil	1

Observations:

- **European Community, Lebanon, Iceland, RSA, Brazil only have one customer**

Top 5 countries with Highest Number of Transactions

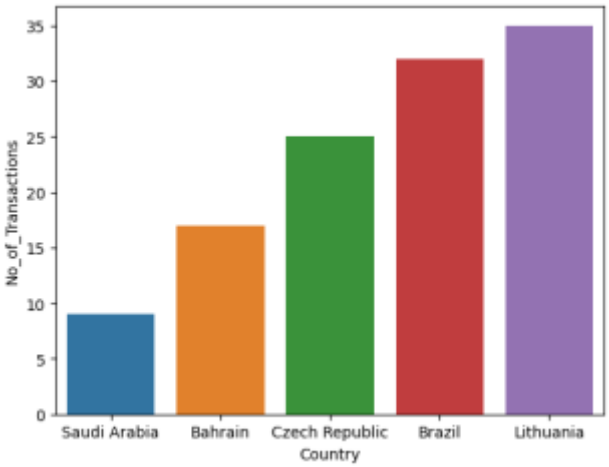


Country	No_of_Transactions
United Kingdom	354345
Germany	9042
France	8342
EIRE	7238
Spain	2485

Observations:

- **UK has the highest number of transactions**

Top 5 countries with lowest number of Transactions

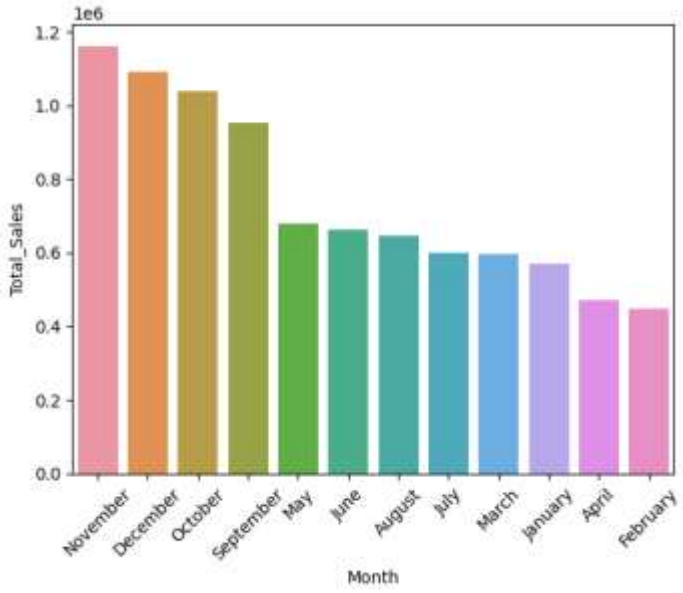


Country	No_of_Transactions
Saudi Arabia	9
Bahrain	17
Czech Republic	25
Brazil	32
Lithuania	35

Observations:

- **Saudi Arabia has least number of transactions.**

Monthly Sales

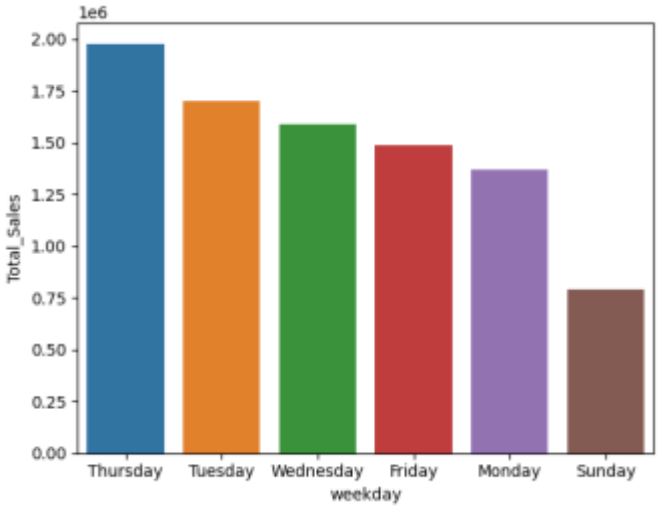


Month	Total_Sales
November	1161817.380
December	1090906.680
October	1039318.790
September	952838.382
May	678594.560
June	661213.690
August	645343.900
July	600091.011
March	595500.760
January	569445.040
April	469200.361
February	447137.350

Observations:

- **Highest Sales observed generally at the end of the year.**

Weekly Sales

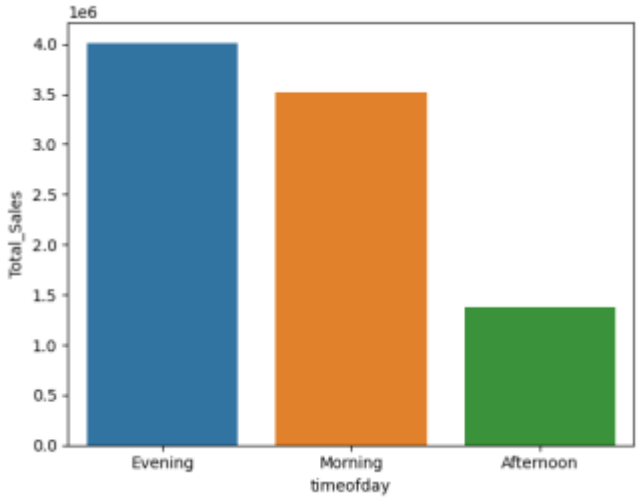


weekday	Total_Sales
Thursday	1976859.070
Tuesday	1700634.631
Wednesday	1588336.170
Friday	1485917.401
Monday	1367146.411
Sunday	792514.221

Observations

- **Thursday has the highest weekly sales.**

Graph of Time of Day Sales



timeofday	Total_Sales
Evening	4011300.842
Morning	3521535.582
Afternoon	1378571.480

Observations:

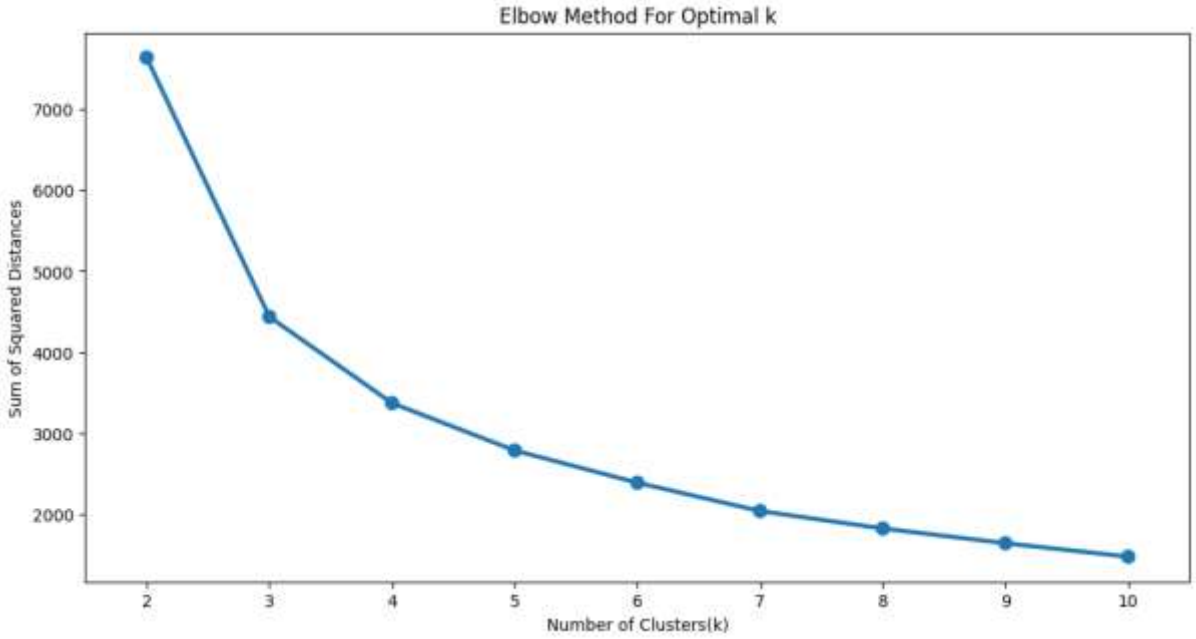
- **We observed high sales during Evening and least sales in afternoon.**

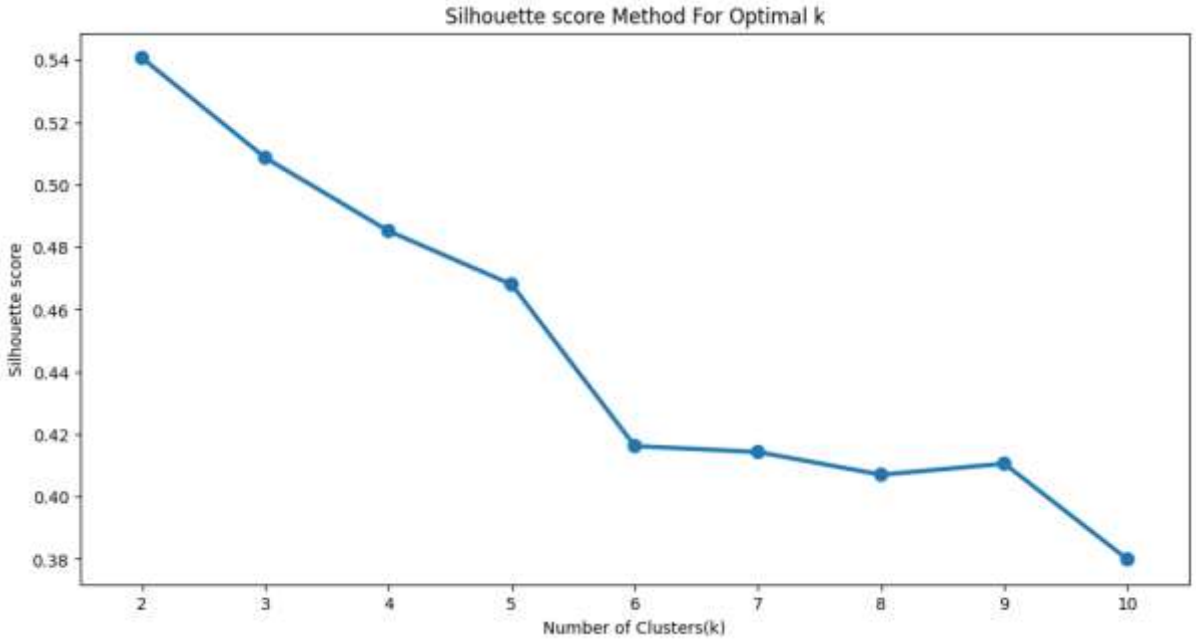
2. KMeans Model:

a. Model Building

Segmentation was performed for 9 different number of clusters [2, 3, 4, 5, 6, 7, 8, 9, 10] using KMeans. The Evaluation metrics are summarized in the Table and Graphs below.

No of Clusters	KMeans Model	
	Sum of Squared Distances	Sillhouette Score
2	7632.92	0.541
3	4437.18	0.509
4	3373.87	0.485
5	2790.04	0.468
6	2392.27	0.416
7	2043.24	0.414
8	1828.11	0.407
9	1646.52	0.411
10	1480.23	0.380





We can see good results in segmentation through both elbow method and Sillhouette score using Number of clusters as 2. As a result, 2 clusters will be formed and types of clusters formed will be evaluated.

3. DBSCAN Model:

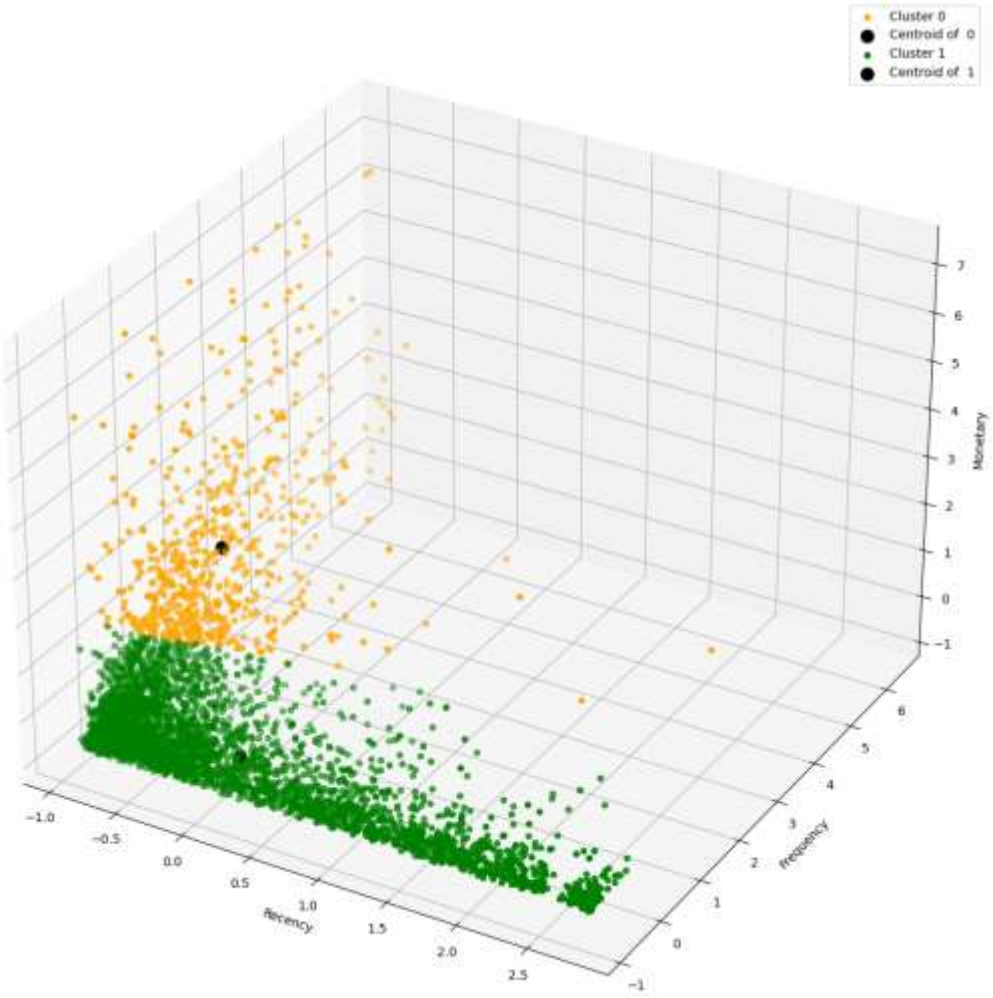
a. Model Building

Segmentation was performed using DBSCAN. The predictions are summarized in the Table and Graphs below

Estimated No of Clusters	DBSCAN Model	
	Estimated Number of Noise Points	Sillhouette Score
2	7632.92	0.414

4. Segmentation

As seen the 2 number of cluster has high Sillhouette score and elbow method also gave the same result. Hence using KMeans and Number of clusters as 2 to segment the customers in the present dataset.



Cluster	Recency			Frequency			Monetary			Inference
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	
0	0.69	-0.93	2.78	1.81	-0.73	6.41	1.78	-0.23	7.20	**Customers who are frequent and whose monetary values are generally high**
1	0.12	-0.93	2.79	0.32	-0.76	1.86	0.31	-0.73	2.00	**Customers who are less frequent, recent and low in monetary value.**

Future Possibilities

The evolution of customer segmentation using machine learning offers a plethora of possibilities for businesses seeking to enhance their understanding of customer behaviour and preferences. This transformation is fuelled by the rapid accumulation of data and the ever-improving capabilities of machine learning algorithms. As this field progresses, several exciting directions emerge, promising more nuanced insights and tailored strategies.

Advanced feature engineering is a significant avenue for growth. Machine learning algorithms have the potential to uncover intricate patterns in customer behaviour. Future developments could involve creating more sophisticated features that capture subtle nuances across various data sources. This integration might encompass social media interactions, browsing habits, and historical purchasing behaviour, allowing businesses to gain a deeper understanding of customer preferences.

Real-time segmentation is on the horizon, thanks to improvements in data processing capabilities. Machine learning models could facilitate dynamic customer segmentation that rapidly adapts to evolving trends. This agility would empower businesses to respond promptly, ensuring their strategies remain relevant and engaging to customers. Moreover, the fusion of various segmentation techniques could be a pivotal advancement. Combining strengths from methods like RFM analysis, clustering, and predictive modelling might yield a more comprehensive understanding of customer behaviour. This holistic perspective could lead to more effective targeting and engagement strategies. As data privacy gains prominence, future segmentation models are likely to incorporate privacy-preserving techniques, aligning with regulations and preserving customer trust.

In summation, the trajectory of customer segmentation using machine learning is marked by remarkable potential. As technology evolves and customer preferences evolve, businesses have the opportunity to harness increasingly sophisticated tools for deciphering and engaging their customer base effectively.

Conclusion

The project undertook a study of online retail company with 25,900 unique transaction and 4070 different products. Some important findings form the report include the following.

1. Customers were segmented into 2 types
    a. One who are more frequent and whose monetary values are generally high.
    b. Another with ones having less frequency, recency and are low in monetary value.
2. To improve the sales, the following steps are recommended:
    a. Ensure the website is easy to navigate, responsive, and user-friendly across all devices. Using high-quality images and clear product descriptions can enhance the shopping experience. Simplify the checkout process to reduce cart abandonment rates.
    b. Leverage customer data to provide personalized product recommendations based on their preferences and behaviour.
    c. Optimize website for search engines (SEO) to increase its visibility in search results. Utilize social media marketing and paid search ads to drive targeted traffic to website.
    d. Use interactive content and engaging visuals on social media platforms to build a loyal online community. Implement email marketing campaigns to share special offers, new product launches, and personalized recommendations.
    e. Offer limited-time discounts, bundle deals, or free shipping to incentivize purchases. Display customer reviews and testimonials to build trust and credibility with potential buyers.
3. Collect feedback from customers is key step to identify pain points and areas for improvement.

References

1. https://www.qualtrics.com/au/experience-management/brand/customer-segmentation/#:~:text=Customer%20segmentation%20is%20the%20process,to%20those%20customers%20more%20effectively
2. https://www.forbes.com/advisor/business/customer-segmentation/
3. https://www.barilliance.com/rfm-analysis/#:~:text=is%20RFM%20Analysis%3F-,A%20definition%20and%20context.,much%20they've%20spent%20overall
4. https://www.analyticsvidhya.com/blog/2021/07/customer-segmentation-using-rfm-analysis/
5. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished
6. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html