Capstone Project -3

Online Retail Customer Segmentation

By

Snehendu Jana

Outline:

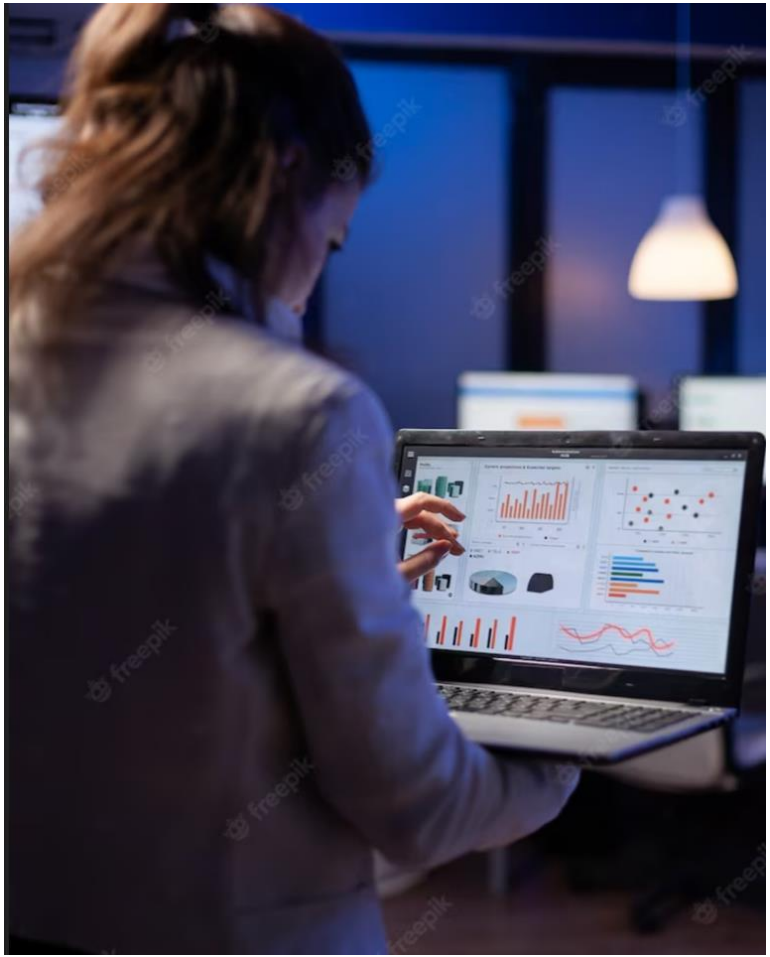
- Introduction to Online Retail Customer Segmentation
- Unsupervised Machine Learning Techniques
- Data Preparation and Feature Engineering
- Case Study: Customer Segmentation in the Online Retail Industry
- Conclusion and Next Steps

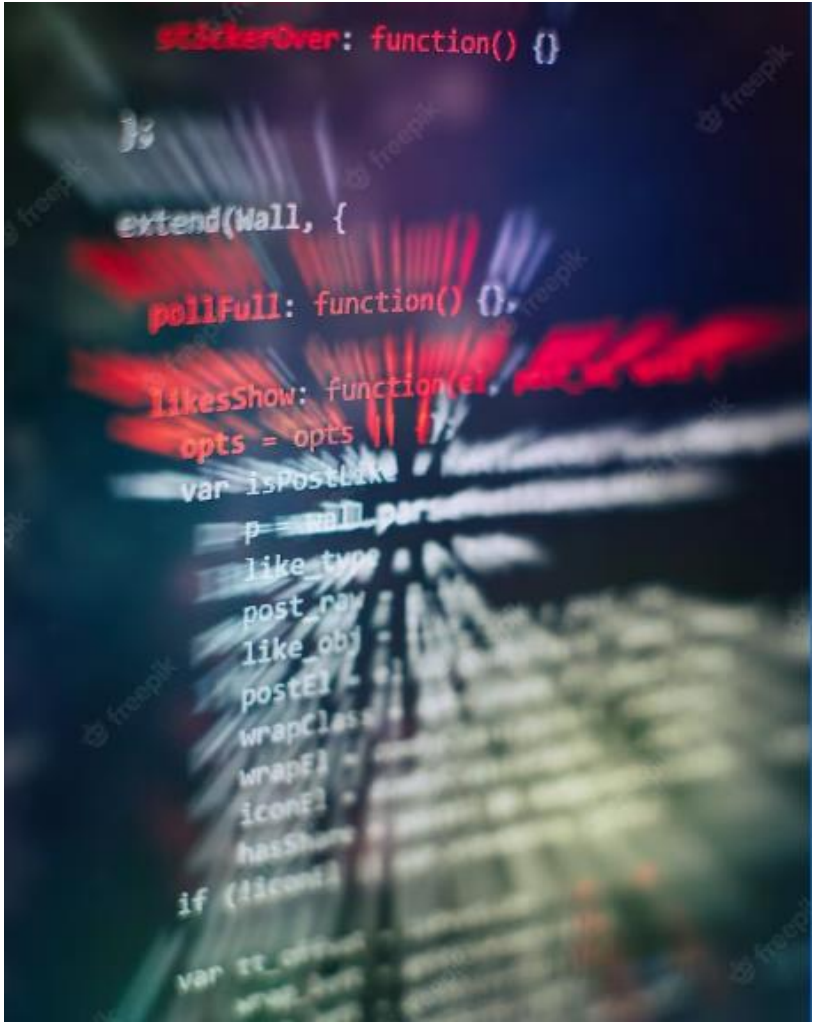
Introduction

This presentation will delve into the outcomes of an online retail customer segmentation project. The invaluable insights extracted from this endeavor will provide businesses with a profound comprehension of their customer base, enabling the refinement of targeted marketing strategies in alignment with customer preferences and behaviors.

Data Collection and Preprocessing

Data Collection and Preprocessing are critical steps in the customer segmentation process. We must gather relevant data from various sources, clean and preprocess it to remove any inconsistencies or errors, and transform it into a format that can be used by machine learning algorithms.





What is Unsupervised ML?

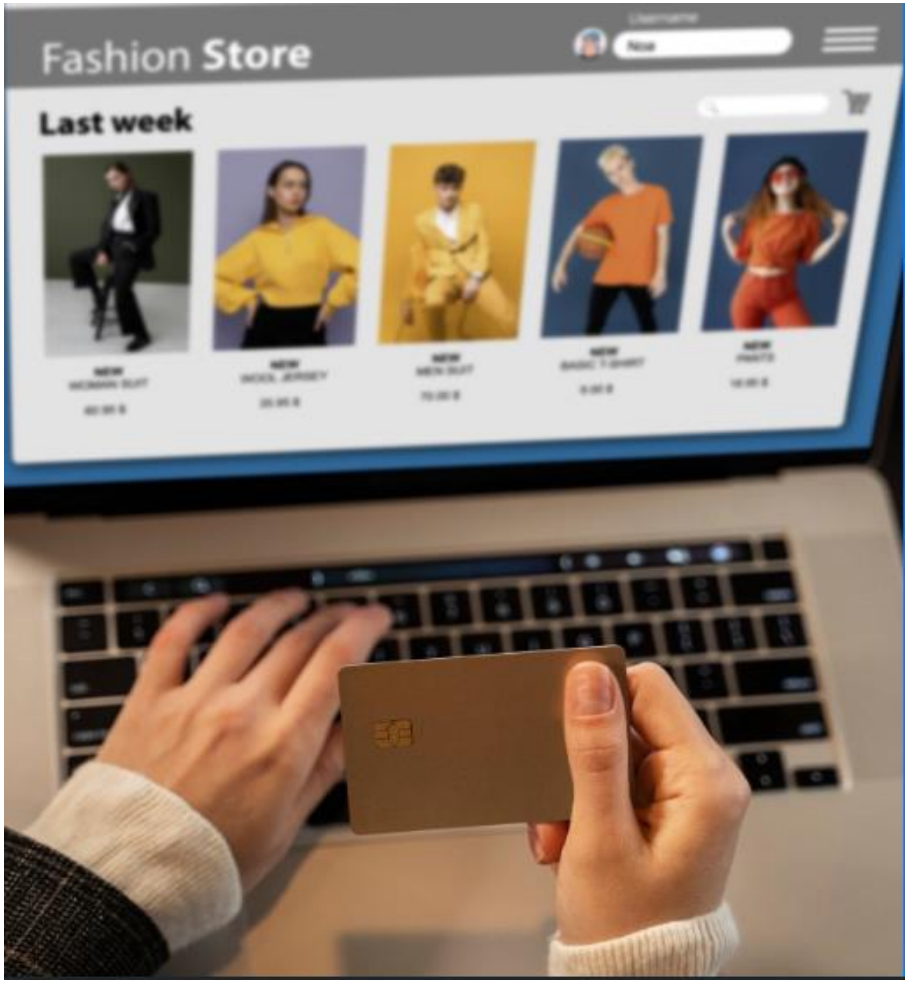
Unsupervised Machine Learning is a technique used to discover hidden patterns or data groupings without the need for labeled data. It is particularly useful for discovering customer segments in online retail.

Why Use Unsupervised ML?

Unsupervised ML allows for the discovery of hidden patterns in large, unstructured datasets. This can help online retailers better understand their customers and tailor their marketing strategies to specific customer segments.

Problem Statement

The challenge addressed in the Online Retail Customer Segmentation project is the absence of a structured approach to harness the power of unsupervised machine learning for categorizing customers. This limits businesses' ability to personalize marketing strategies, resulting in suboptimal customer experiences and inefficient resource allocation. This presentation emphasizes the significance of implementing unsupervised ML techniques to create meaningful customer segments, thereby revolutionizing targeted marketing efforts and elevating customer satisfaction in the online retail landscape.



Case Study

We will present a case study of an online retailer that used Unsupervised ML to segment their customers. We will show how this helped them improve their marketing strategies and increase sales.

Introduction

Online retail customer

- Online retail customer segmentation involves categorizing shoppers into distinct groups based on behavior and characteristics. This aids tailored marketing, personalized experiences, and better decision-making. Unsupervised ML techniques reveal patterns, enhancing business strategies.

Methodology

- Unsupervised Machine Learning Techniques

Database:

- Online retail customer segmentation
- 102673 rows and 8 columns
- Data from last decade

Data Description

- **InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to eachtransaction. 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 wasgenerated.
- **UnitPrice:** Unit price. Numeric, Product price per unit in sterling.
- **CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to eachcustomer.
- **Country:** Country name. Nominal, the name of the country where each customer resides.

Data Summary

The dataset contains 8 columns and 541909 rows.

There also exist some null values in our data:

- Percentage of null values in Description : 30.68%
- Percentage of null values in CustomerID : 9.22%

	DataType	Non-null_Values	Unique_Values	NaN_Values	NaN_Values_Percentage
InvoiceNo	object	541909	25900	0	0.000000
StockCode	object	541909	4070	0	0.000000
Description	object	540455	4223	1454	0.268311
Quantity	int64	541909	722	0	0.000000
InvoiceDate	datetime64[ns]	541909	23260	0	0.000000
UnitPrice	float64	541909	1630	0	0.000000
CustomerID	float64	406829	4372	135080	24.926694
Country	object	541909	38	0	0.000000

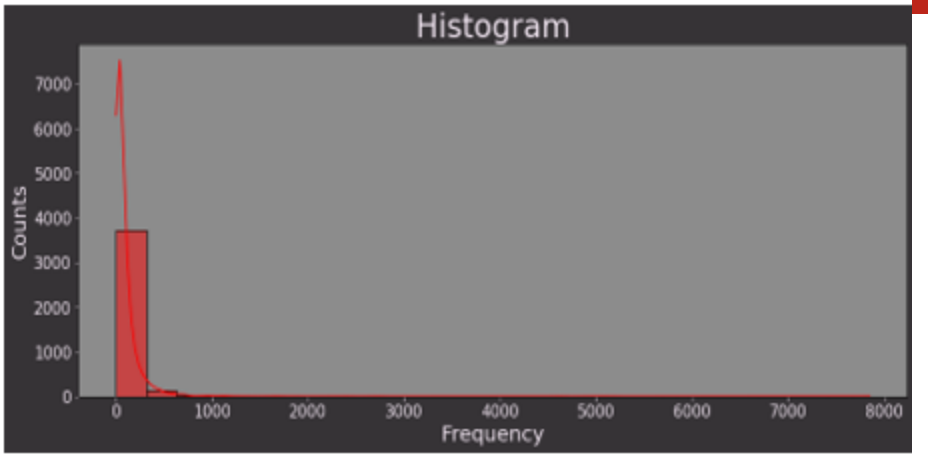
RFM Modelling

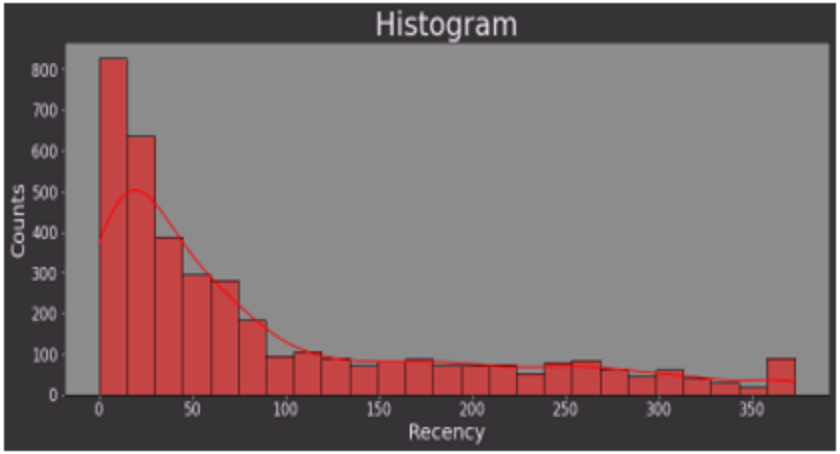
RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).

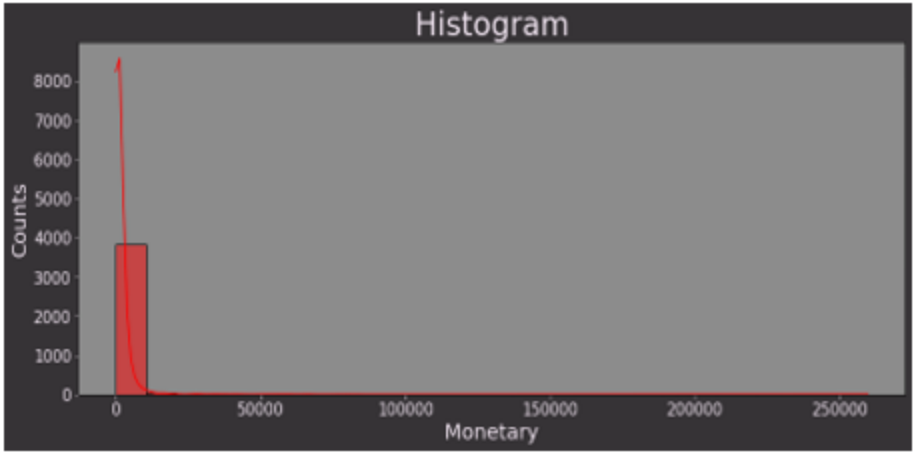
•RFM analysis helps to answer the following questions: Who was our most recent customer? How many times has he purchased items from our shop? And what is the total value of his trade? All this information can be critical to understanding how good or bad a customer is to the company.

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4595	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

- With the help of histogram we can say that Recency is right skewed where as Frequency and Monetary are left skewed



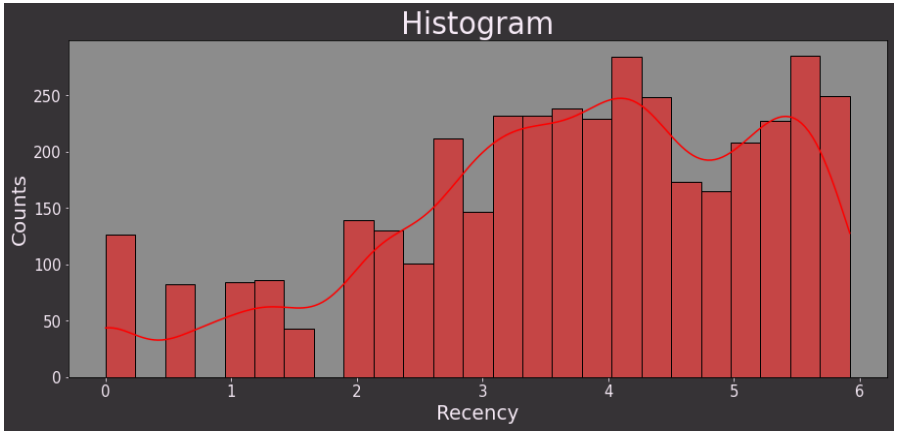


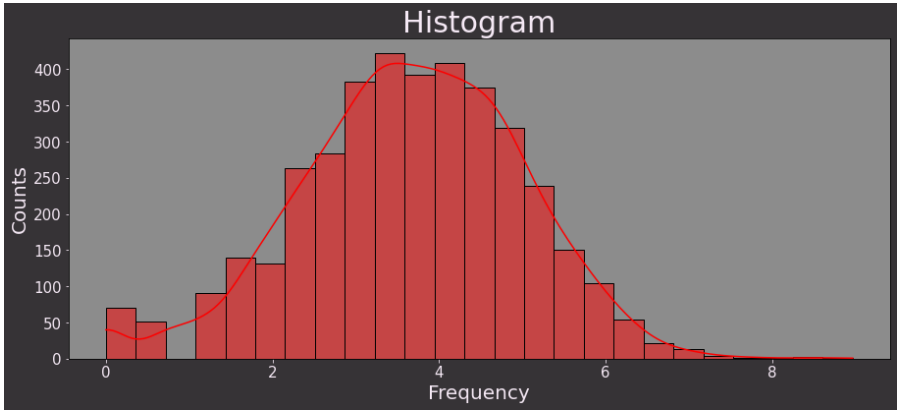


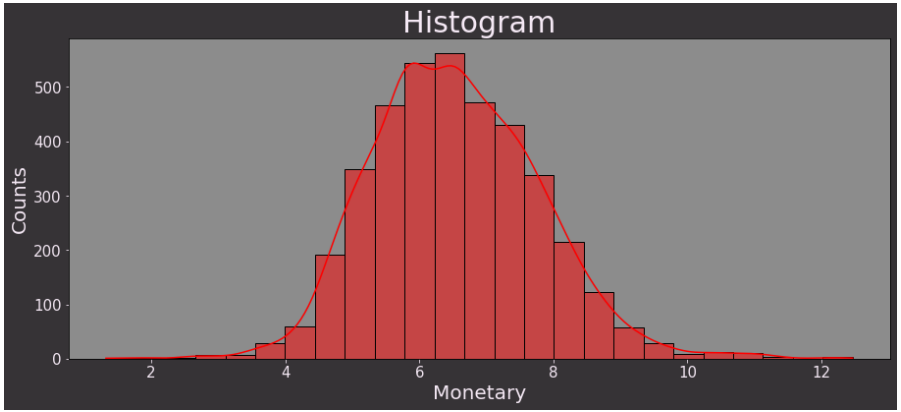
Log Transformation

```
#Perform Log transformation to bring data into
normal or near normal distribution

Log_Data = RFMScores[['Recency', 'Frequency',
'Monetary']].apply(np.log, axis = 1).round(3)
```







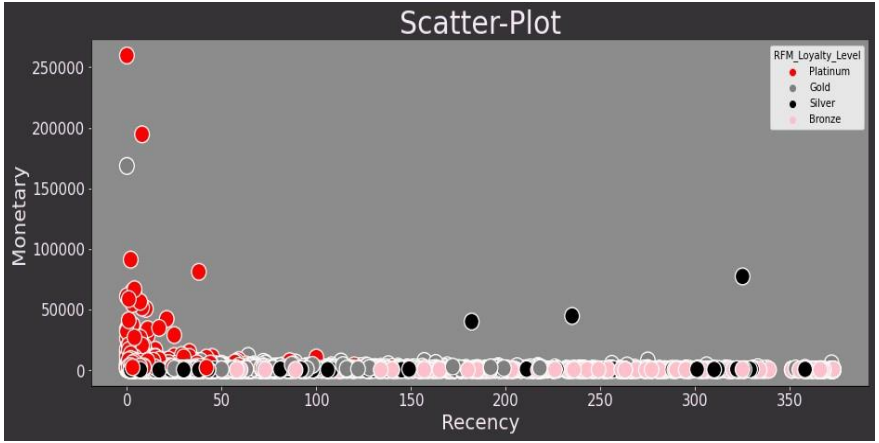
After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assigning the required order. For example, suppose that we divide each metric into 4 cuts. For the recency metric, the highest value, 4, will be assigned to the customers with the least recency value (since they are the most recent customers). For the frequency and monetary metric, the highest value, 4, will be assigned to the customers with the Top 25% frequency and monetary values, respectively. After dividing the metrics into quartiles, we can collate the metrics into a single column (like a string of characters {*like ‘213’*}) to create classes of RFM values for our customers. We can divide the RFM metrics into lesser or more cuts depending on our requirements

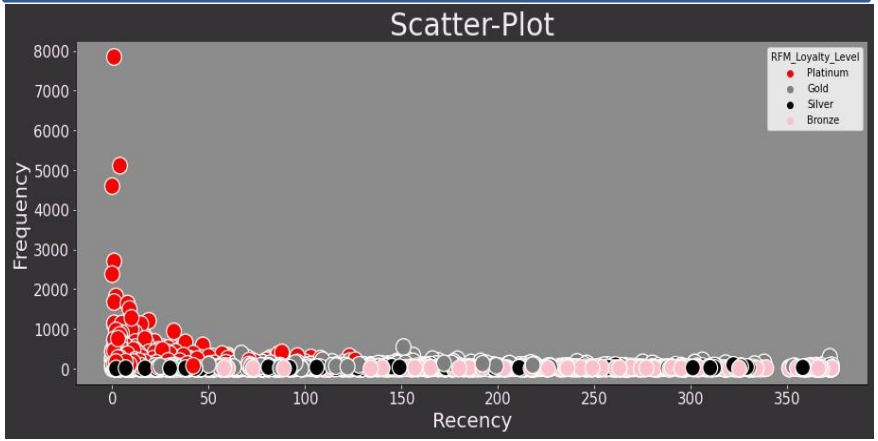
	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4595	33719.73	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum

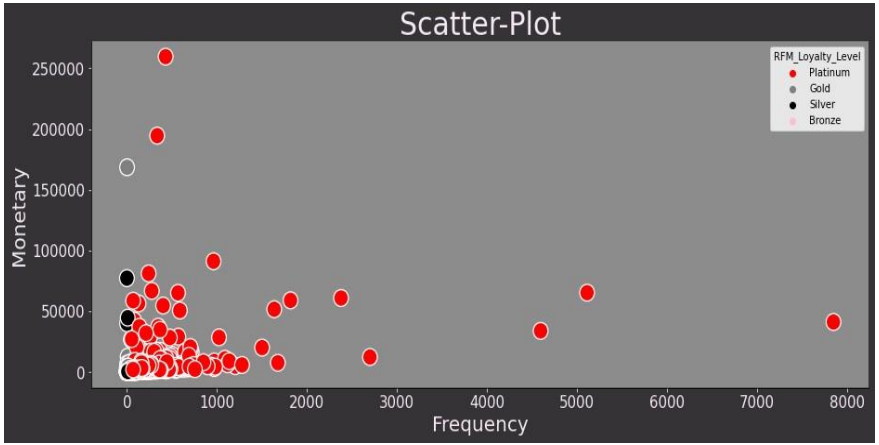
RFM LOYALTY LEVEL

We divided our customers into Four loyalty level which help us to distinguish the customers according to their RFM scores.

A. platinum
B. Gold
C. Silver
D. Bronze







K - Means

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

K-Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group.

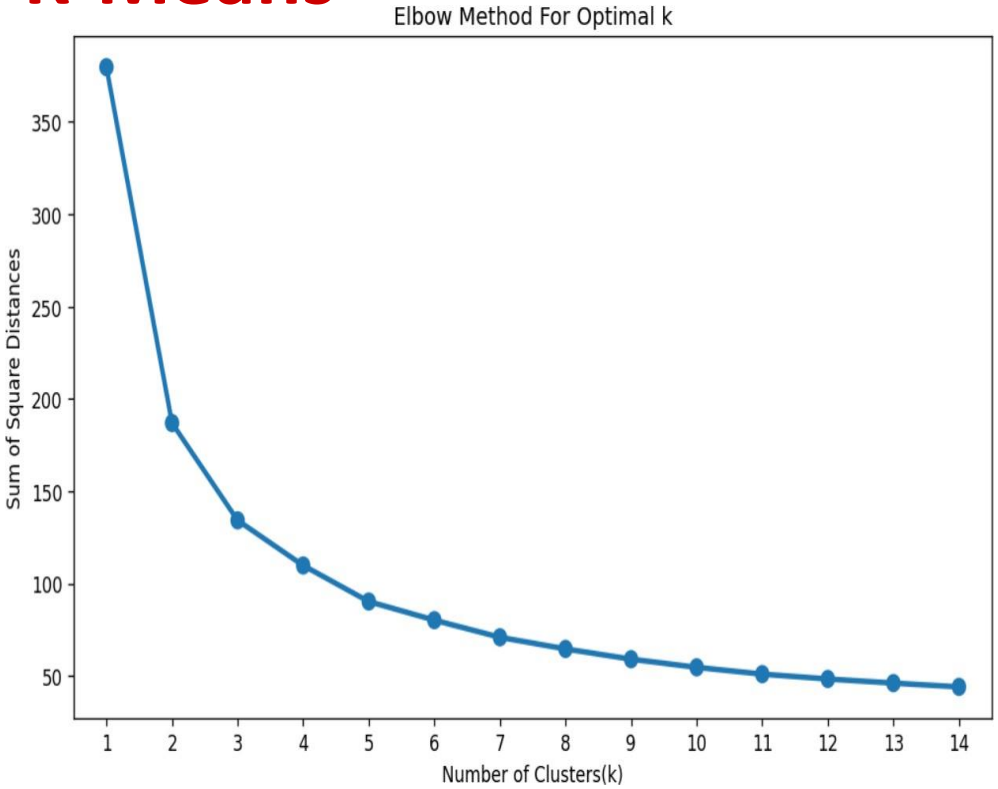
1. Elbow Curve:

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE)to choose an ideal value of k based on the distance between the data points and their assigned clusters.

2. Silhouette score :

- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

K-Means



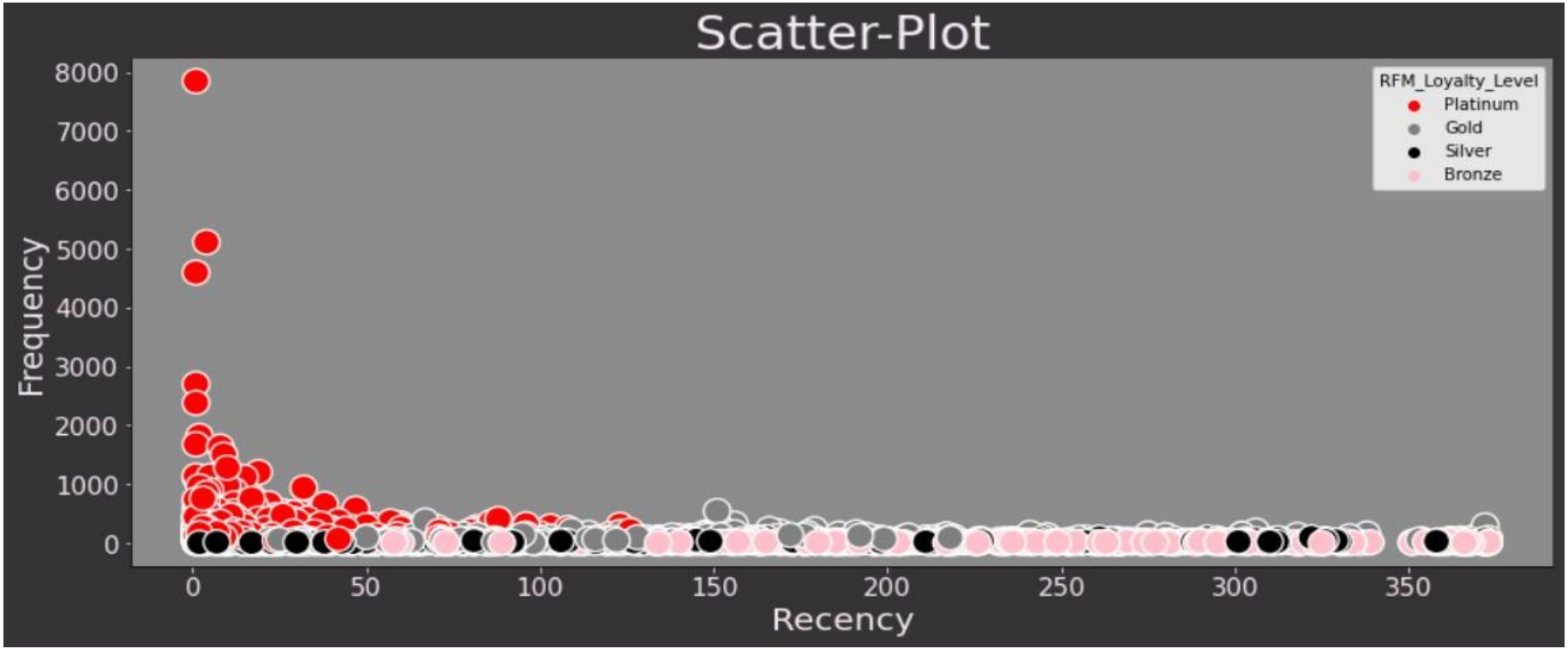
Cluster K3	Recency mean	Frequency mean	Monetary mean	Monetary count
0	192.0	24.0	545.0	1480
1	6.0	241.0	5621.0	718
2	41.0	85.0	1432.0	1722

Cluster K4	Recency mean	Frequency mean	Monetary mean	Monetary count
0	214.0	24.0	556.0	1251
1	38.0	22.0	423.0	843
2	5.0	233.0	5504.0	696
3	50.0	128.0	2146.0	1130

To understand what these 4 clusters mean in a business scenario, we should look back the table comparing the clustering performance of 3 and 4 clusters for the mean values of recency, frequency, and monetary metric. On this basis, let us label the clusters as ‘New customers’, ‘Lost customers’, ‘Best customers’, and ‘At risk customers’.

Cluster	Type of customers	RFM Interpretation	Recommended action
0	New customers	Customers who transacted recently and have lower purchase frequency, with low amount of monetary spending.	Need to handled with care by improving relationships with them. Company should try to enhance their purchasing experience by providing good quality products and services, and customer care services.
1	Lost customers	Customers with the least monetary spending and the least number of transactions. Made their last purchase long ago.	These customers may have already exited from the customer base. The company should try to understand why they left the system so that it does not happen again.
2	Best customers	Most frequent spenders with the highest monetary spending amount and had transacted recently.	Potential to be the target of new products made by a company and can increase company revenue by repeated advertising. Heavy discounts not required.
3	At risk customers	Customers who made their last transaction a while ago and made less frequent and low monetary purchases.	At high risk of churning. Need to be addressed urgently with focussed advertising. May perform well if discounts are provided to them. Company should find out why they are leaving.

Final we make 4 clusters



Conclusion

Customer segmentation is a highly effective strategy for organizations because it allows them to know which customers care about them and understand their needs enough to send a message that ensures brand success.

- we used RFM Modeling to see the relation between Recency,Frequncy and Monetary.
- After RFM model we used this data to perform clustering with the help of k mean clustering Algorithm.
- At the end we make 4 clusters of customers named as.
- **Clustor 0** - New Customer = Customer who transacted recently and have lower purchase frequency, with low amount of monetary spending.
- **Clustor 1** - Lost Customers = Customers with the least Monetary spending and the least number of transaction.
- **Clustor 2** - Best Customers = Most frequent spenders with the highest monetory spending amount and had transacted recently.
- **Clustor 3** - At Risk Customers = Customers who made their last transaction a while ago and made less frequent and low monetary purchases.

Thank you