

# **UNSUPERVISED MACHINE LEARNING CAPSTONE PROJECT**

## **CUSTOMER SEGMENTATION**

Presented by-

**Prayag Raj Dubey**

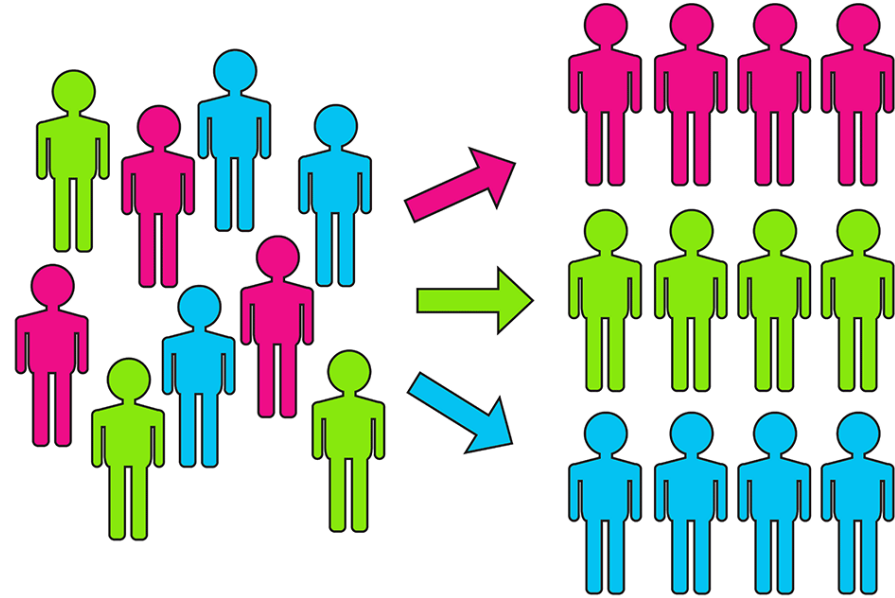
# Contents:

- Introduction
- Problem Statement
- Data Analysis Steps
- Data Preview
- Data Summary
- RFM Modeling
- MinMaxScalar
- Clustering
- Conclusion



# Introduction:

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable. Insights from customer segmentation are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning. A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the company philosophy and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy.



# Problem Statement:

Customer segmentation has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers. For example, a company can make a high-end product, a budget product, or a cheap alternative product, depending upon whether the product is intended for its most high yield customers, frequent purchasers or for the low-value customer segment

Finally, this technique can also be used by companies to test the pricing of their different products, improve customer service, and upsell and cross-sell other products or services.



# Data Analysis Steps:

## Imported Libraries

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

## Descriptive Statistics

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this told mean, median, standard deviation

## Missing Value Imputation

We will now check for missing values in our dataset. after checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

## Graphical Representation

We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models helps us predict the Risk.

# Dataset Preview:

- **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.

# 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%

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

# RFM Modeling

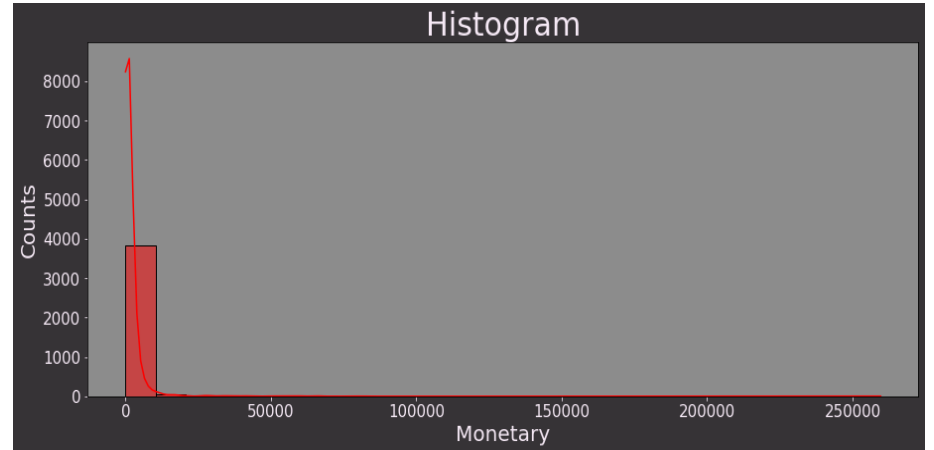
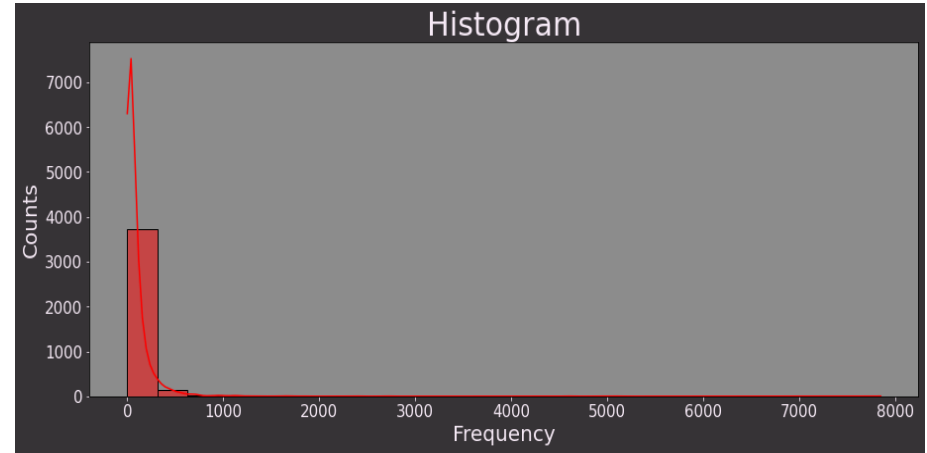
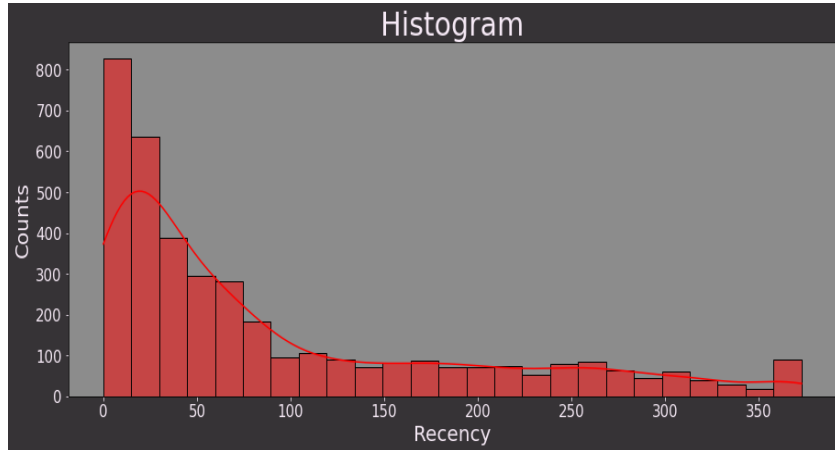
RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their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



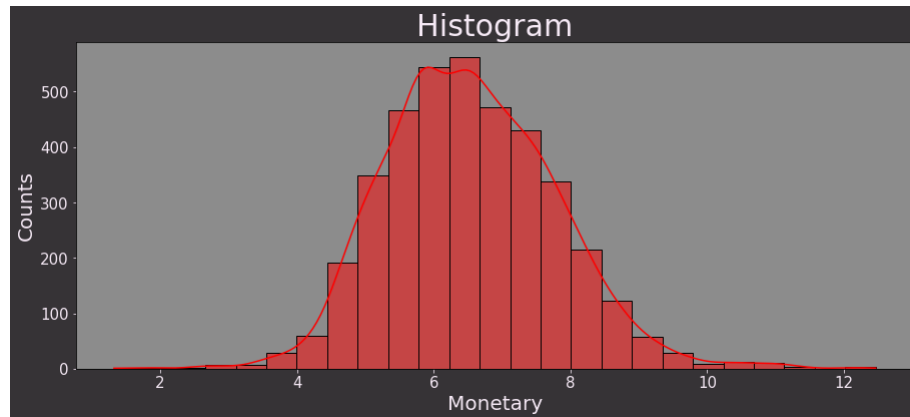
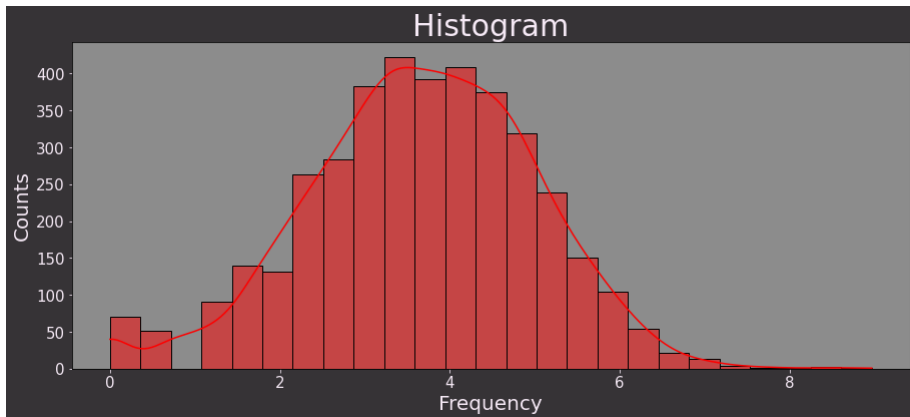
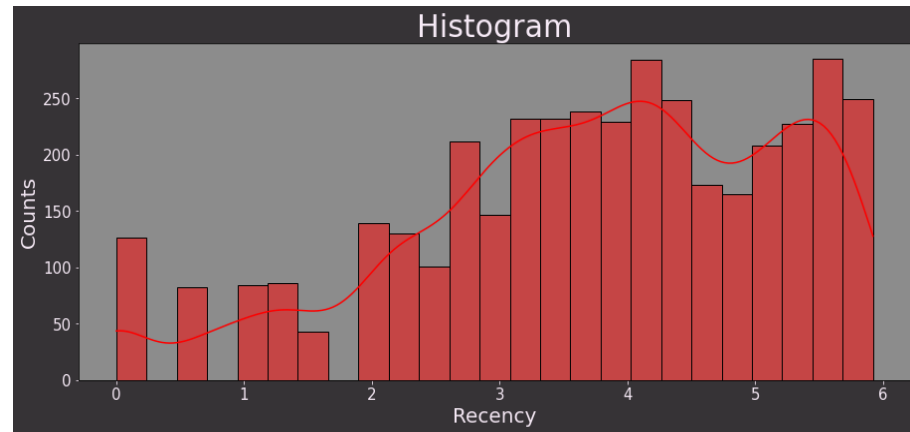
- With the help of histogram we can say that Recency is right skewed where as Frequency and Monetary are left skewed



# 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)
```



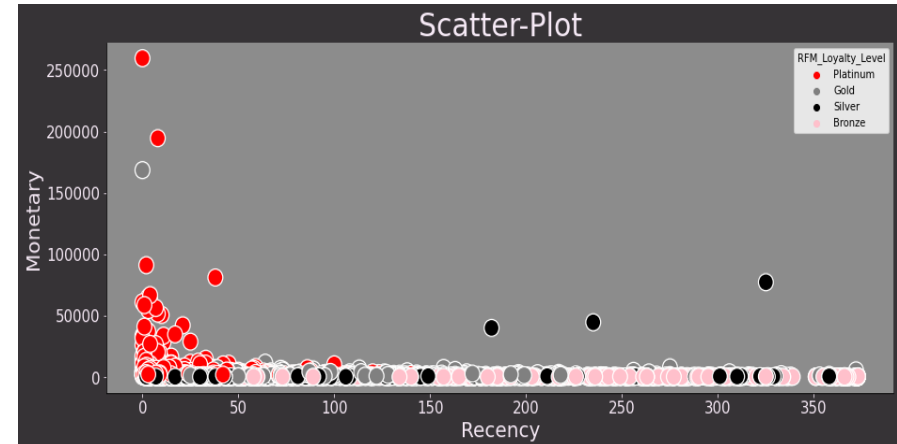
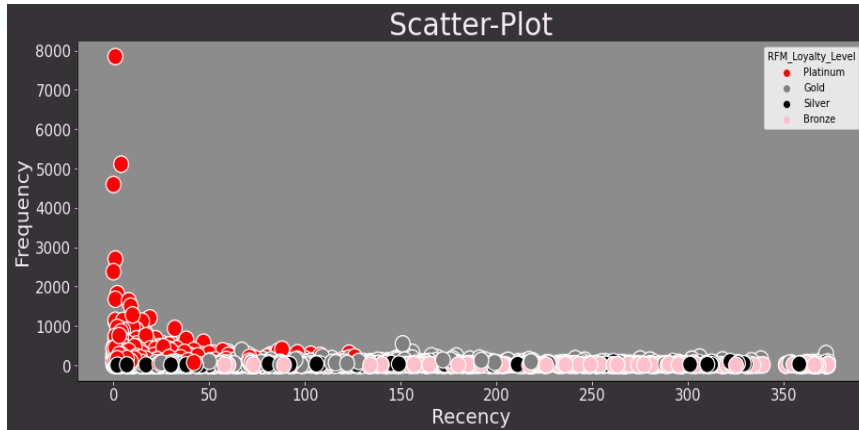
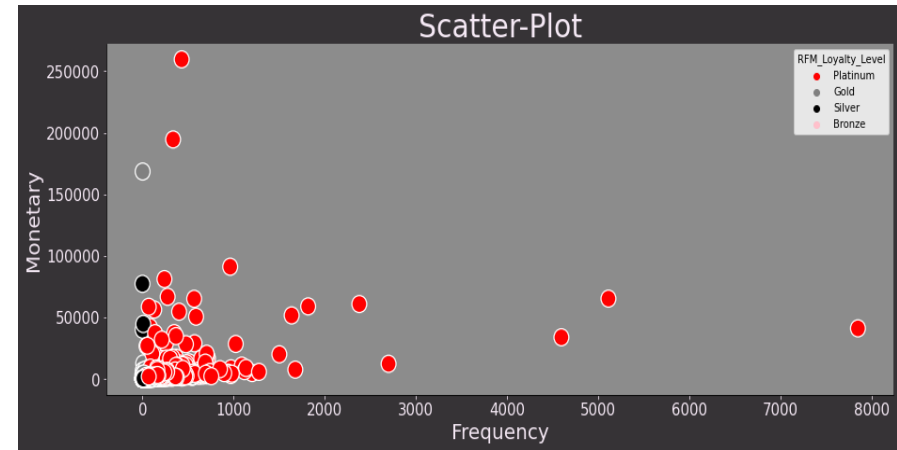
After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assigning the required order. For example, suppose that we divide each metric into 4 cuts. For the recency metric, the highest value, 4, will be assigned to the customers with the least recency value (since 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.

1. platinum
2. Gold
3. Silver
4. Bronze



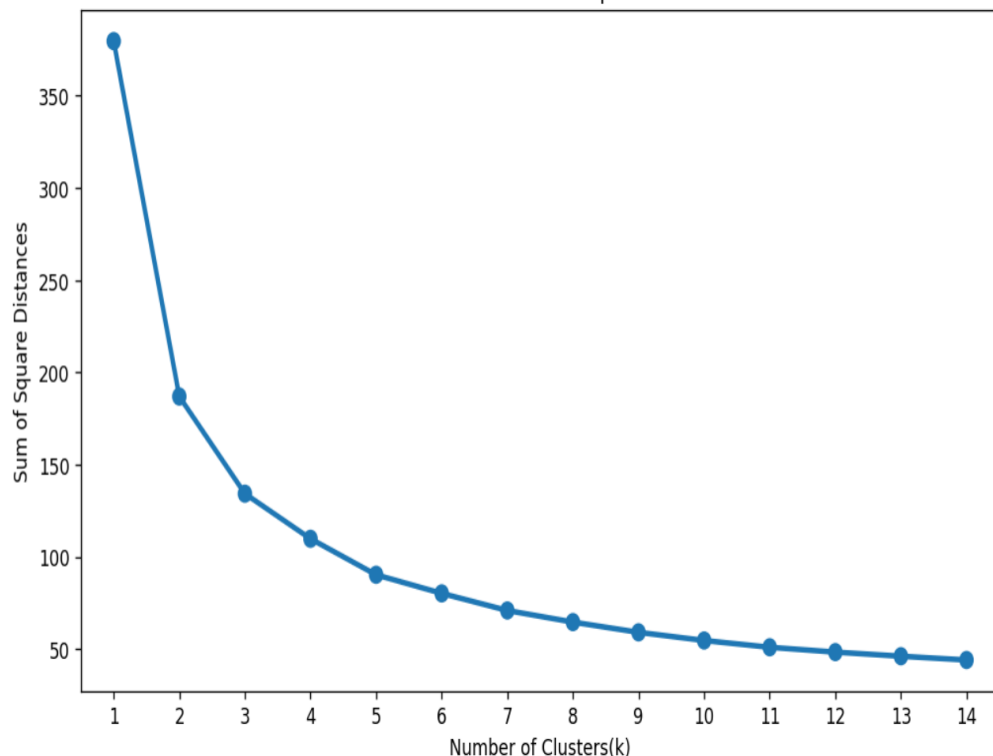
# K-Mean Clustering

- Here we perform K-mean to make clusters.

```
from sklearn.cluster import KMeans
plt.figure(figsize=(10,6), dpi=120)
inertias = {}
for k in range(1,15):
    model = KMeans(n_clusters= k, init= 'k-means++', max_iter= 1000,random_state=42)
    model = model.fit(Scaled_Data)
    inertias [k] = model.inertia_

#Plot the graph for the sum of square distance values and Number of Clusters
sns.pointplot(x = list(inertias.keys()), y = list(inertias.values()))
plt.xlabel('Number of Clusters(k)')
plt.ylabel('Sum of Square Distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

Elbow Method For Optimal k



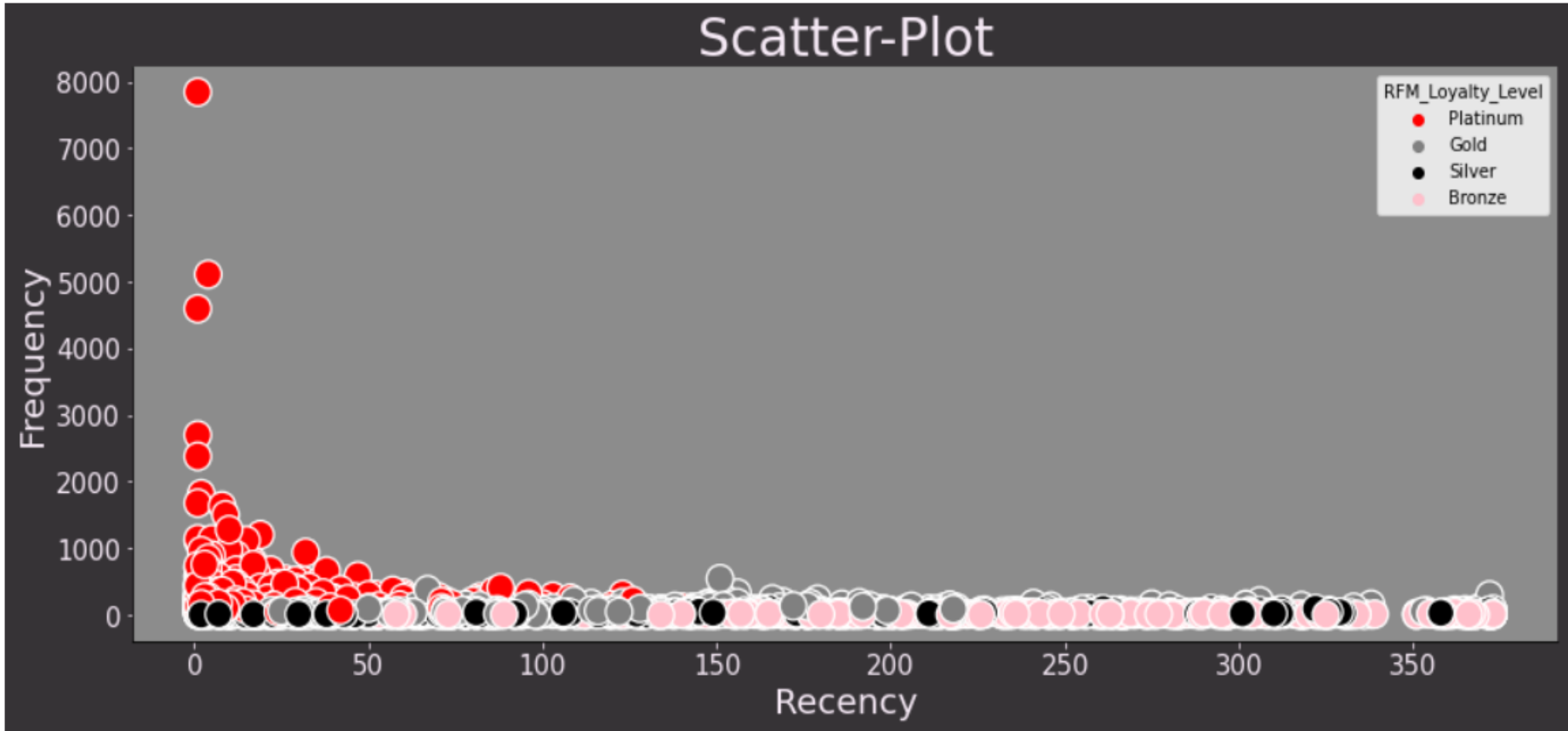
	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster K3				
0	192.0	24.0	545.0	1480
1	6.0	241.0	5621.0	718
2	41.0	85.0	1432.0	1722
	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster K4				
0	214.0	24.0	556.0	1251
1	38.0	22.0	423.0	843
2	5.0	233.0	5504.0	696
3	50.0	128.0	2146.0	1130

Using the **Elbow** Method we select the optimal number of **clusters** to be **3 or 4** , From the above analysis, we can see that there should be 4 clusters in our data.

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, Frequency 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.

- Cluster 0** - New Customer = Customer who transacted recently and have lower purchase frequency, with low amount of monetary spending.

- Cluster 1** - Lost Customers = Customers with the least Monetary spending and the least number of transaction.

- Cluster 2** - Best Customers = Most frequent spenders with the highest monetary spending amount and had transacted recently.

- Cluster 3** - At Risk Customers = Customers who made their last transaction a while ago and made less frequent and low monetary purchases.

**THANK YOU**