DS 21

FINAL PROJECT

ONLINE RETAIL CUSTOMER SEGMENTATION (EDA, RFM, K MEANS)

BY YUNITA APRILIA

OVERVIEW

- Data Background
- Data PreProcessing
- Exploratory Data Analyst
- Modeling
- Conclusion & Recomendation

DATA BACKGROUND

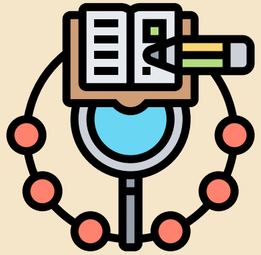
BUSINESS UNDERSTANDING

E-commerce has become a new channel to support businesses development by providing cheaper and more efficient distribution channels for their products or services. But, there are more trade competitors of the retail industry. The company must be recognize to understand its customer segmentation and marketing strategies accordingly. So, what efforts should be made the company?

BUSINESS PROBLEM

How to build good relation with customer based on the customer behavior ?

How to implement marketing strategies ?



What kind of machine learning model that suitable to predict default of a client?

DS 21

ABOUT DATA

SOURCE : *https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset*

DS 21

- This project uses the “**Online Retail**” dataset which contains all the transactions occurring between 10/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.
- There are 4.372 user data & 23.260 transaction data recorded only on the web store

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3.39	17850	United Kingdom

541909 rows

GOALS :

1. Increase sales/revenue
2. Improve marketing
3. Increase customer retention

OBJECTIVES

DS 21

MODELING

Perform customer segmentation (clustering) through the customer dataset.

RFM ——— K - MEANS CLUSTERING

ANALYSIS

Analysis of the characteristics of each cluster resulting from segmentation.

RECOMMENDATION

Provide business insight related to the analysis results.

DATA PREPROCESSING

DS 21

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

MISSING VALUE

	feature	missing_value	percentage
0	CustomerID	135080	24.93
1	Description	1454	0.27

Need to Drop !

DUPLICATED DATA

```
df.duplicated().sum()

5225
```

1,28% are duplicated

Need to Drop !

DATA PREPROCESSING

DS 21

```
df.describe()
```

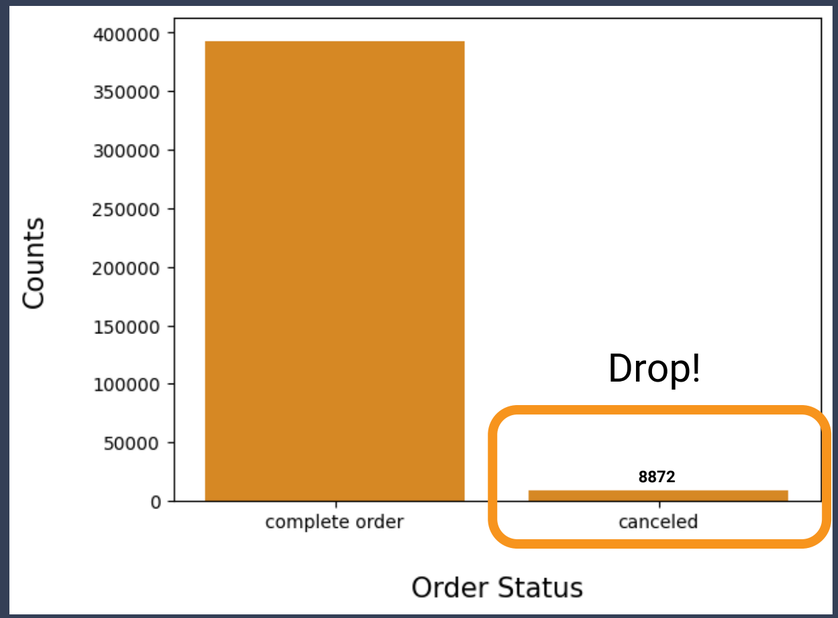
	Quantity	UnitPrice	CustomerID
count	401604.000000	401604.000000	401604.000000
mean	12.183273	3.474064	15281.160818
std	250.283037	69.764035	1714.006089
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13939.000000
50%	5.000000	1.950000	15145.000000
75%	12.000000	3.750000	16784.000000
max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

STATISTICAL SUMMARY

Observation :

- The min value for Quantity is 80995, this could represent cancelled or returned orders.
- The UnitPrice also have few negative values which is uncommon, these transactions could represent cancelled orders by customers or bad-debt incurred by the business.

Need to Clean!




```
len(df[df['UnitPrice']==0])
```

```
40
```

Final Data

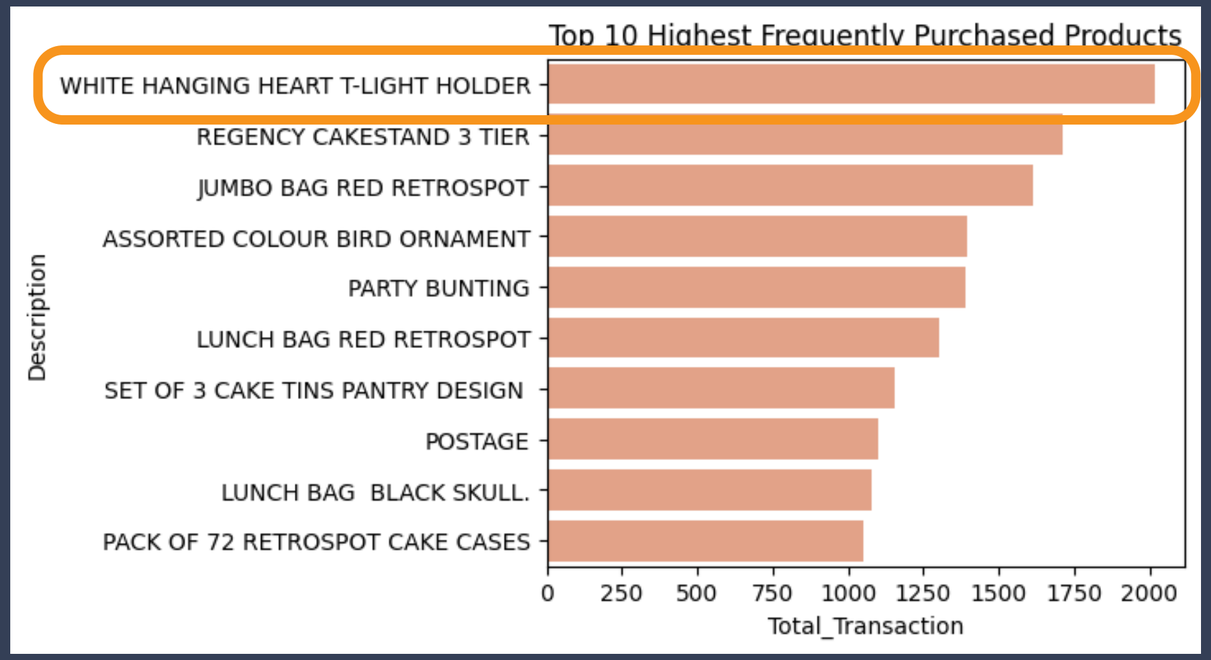
```
df.shape
```

```
(392692, 11)
```

EXPLORATORY DATA ANALYST ____

DS 21

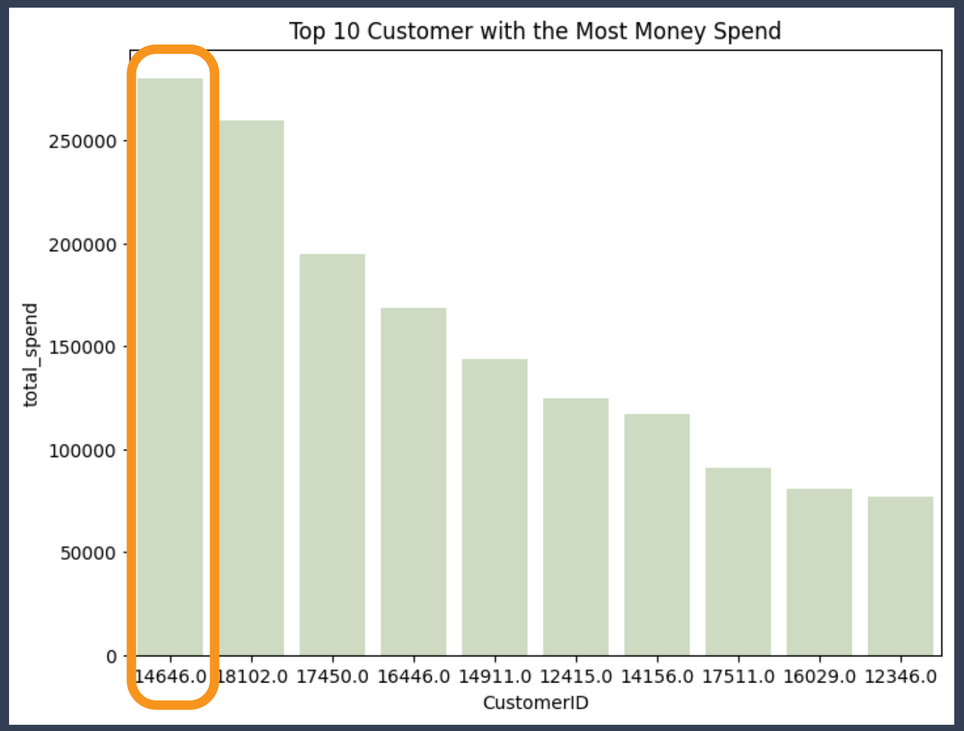
WHAT ARE THE MOST FREQUENTLY PURCHASED PRODUCTS ?



Observation :

- The most frequently purchased products is the WHITE HANGING HEART T-LIGHT HOLDER, with a total quantity of 2016 units purchased.

WHO ARE THE TOP 10 CUSTOMERS IN MONEY SPEND ?



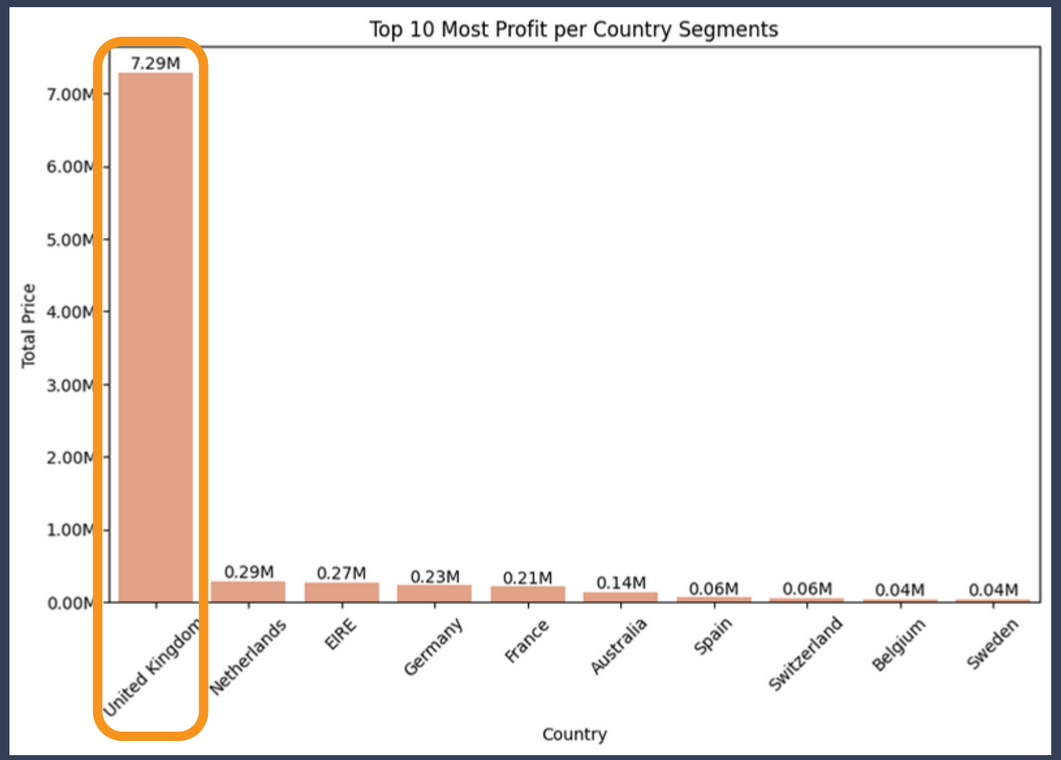
Observation :

- CustomerID 14646 is the top customer with a total money spend 280206.02 per unit in sterling

EXPLORATORY DATA ANALYST

DS 21

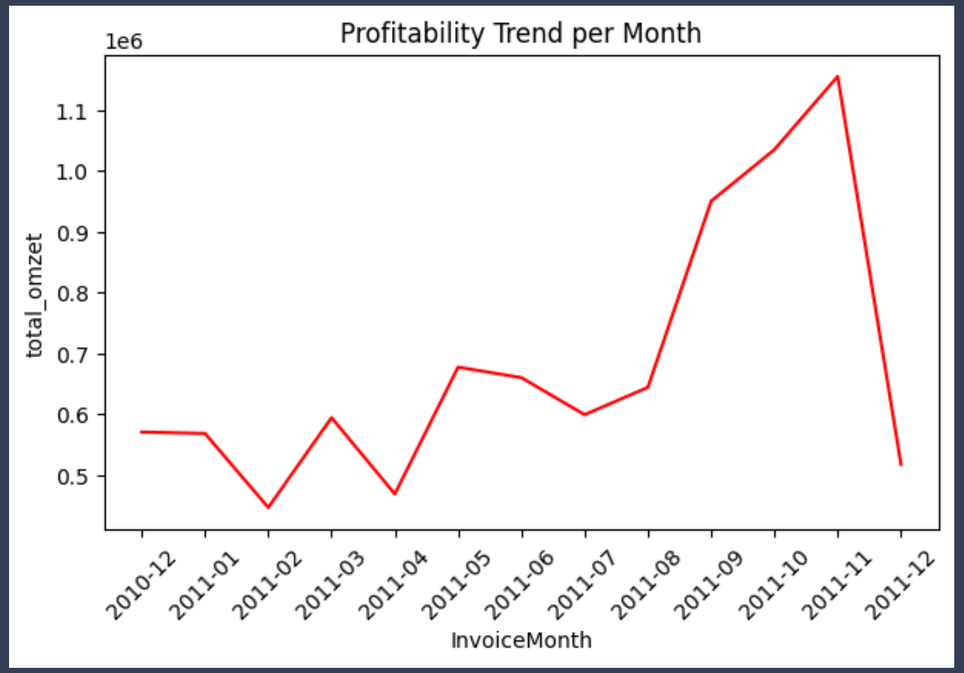
WHAT ARE THE MOST PROFITABLE SEGMENT CUSTOMERS ?



Observation :

- The United Kingdom stands out as the top-performing country in terms of profitability, generating over 7 million pounds sterling in profit.

HOW ABOUT MONTHLY TIME SERIES OF TOTAL OMZET ?



Observation :

- The trend of total omzet for several months from the end of 2010 showed instability until August 2011, followed by a consistent and steady increase from September culminating in the highest total buyers in November, with a total omzet 1.156.205.

EXPLORATORY DATA ANALYST _____

DS 21

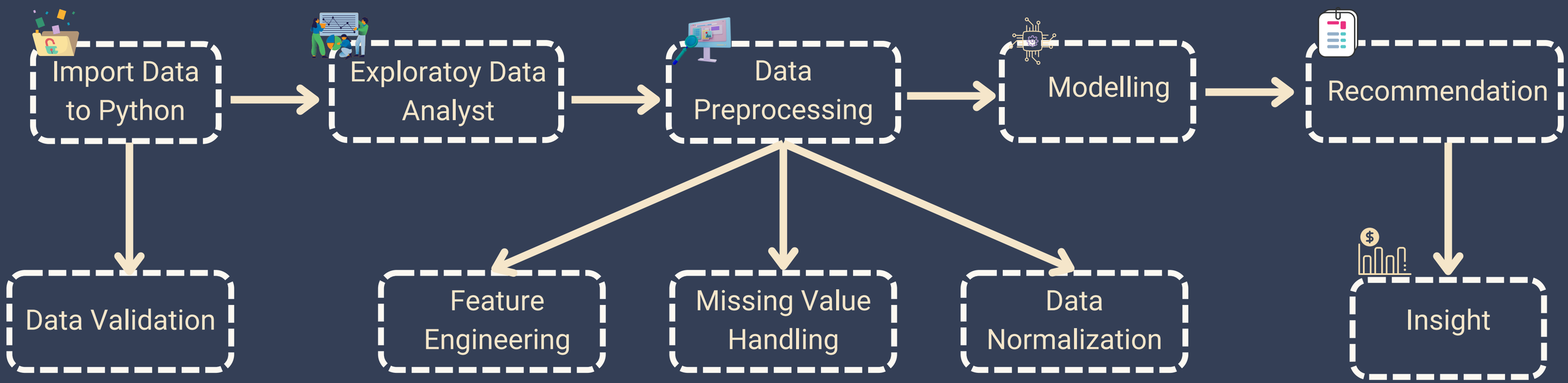
RECOMMENDATION FOR EDA

1. Analyzing the reasons behind purchase cancellations and look for strategies to reduce them and improve product descriptions or packaging if necessary.
2. Focus on top product sales: provide promotions on products with the highest demand "WHITE HANGING HEART T-LIGHT HOLDER" to boost sales.
3. Provide promotions in months when sales are slow to maintain customer interest.
4. Implement new marketing strategies on products with low demand.

MODELING

DS 21

CUSTOMER SEGMENTATION WORKFLOW



RFM ANALYSIS

DS 21

- **RECENCY :** Day since last purchase
- **FREQUENCY :** Total number of purchases
- **MONETARY :** Total amount spent

	CustomerID	Recency	Frequency	Monetary	Rscore	Fscore	Mscore	RFM_Group	RFM_score	RFM_Loyalty_customer
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12347.0	2	182	4310.00	1	1	1	111	3	Platinum
2	12348.0	75	31	1797.24	3	3	1	331	7	Gold
3	12349.0	18	73	1757.55	2	2	1	221	5	Platinum
4	12350.0	310	17	334.40	4	4	3	443	11	Bronz

RFM SCORE

CUSTOMER LOYALITY LEVEL

RFM ANALYSIS

DS 21

1

	CustomerID	Recency	Frequency	Monetary	Rscore	Fscore	Mscore	RFM_Group	RFM_score	RFM_Loyalty_customer
0	12346.0	325	1	77183.60	4	4		441	9	Silver
1	12347.0	2	182	4310.00	1	1		111	3	Platinum
2	12348.0	75	31	1797.24	3	3		331	7	Gold
3	12349.0	18	73	1757.55	2	2		221	5	Platinum
4	12350.0	310	17	334.40	4	4		443	11	Bronz

3. RFM SCORE

4

CUSTOMER LOYALITY LEVEL

5

2

```
quantiles = rfm_df.quantile(q = [0.25, 0.50, 0.75])
quantiles
```

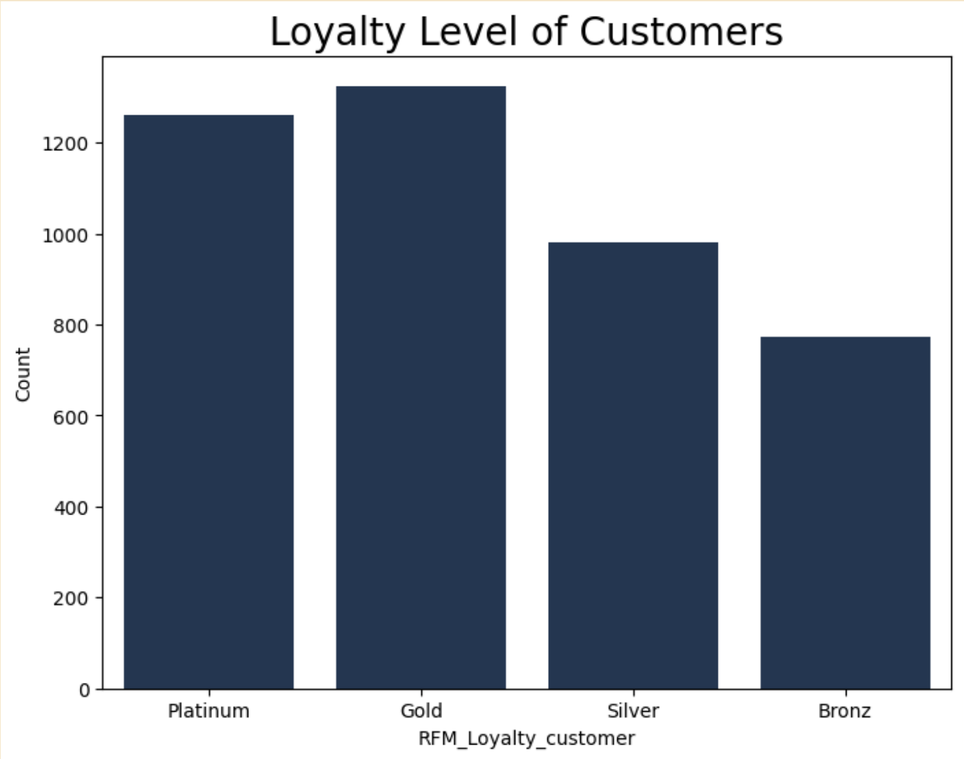
```
def RScoring(x,p,d):
    if x <= d[p][0.25]:
        return 1
    elif x <= d[p][0.50]:
        return 2
    elif x <= d[p][0.75]:
        return 3
    else:
        return 4
```

```
def FnMScoring(x,p,d):
    if x <= d[p][0.25]:
        return 4
    elif x <= d[p][0.50]:
        return 3
    elif x <= d[p][0.75]:
        return 2
    else:
        return 1
```

RFM ANALYSIS

DS 21

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
RFM_Loyalty_customer										
Platinum	19.550357	0	140	225.884219	20	7676	5253.556788	316.25	280206.02	1261
Gold	63.433962	0	372	57.012075	1	521	1162.987662	114.34	168472.50	1325
Silver	125.178571	1	373	24.337755	1	98	579.452461	6.90	77183.60	980
Bronz	217.585492	51	373	10.958549	1	39	199.030725	3.75	660.00	772

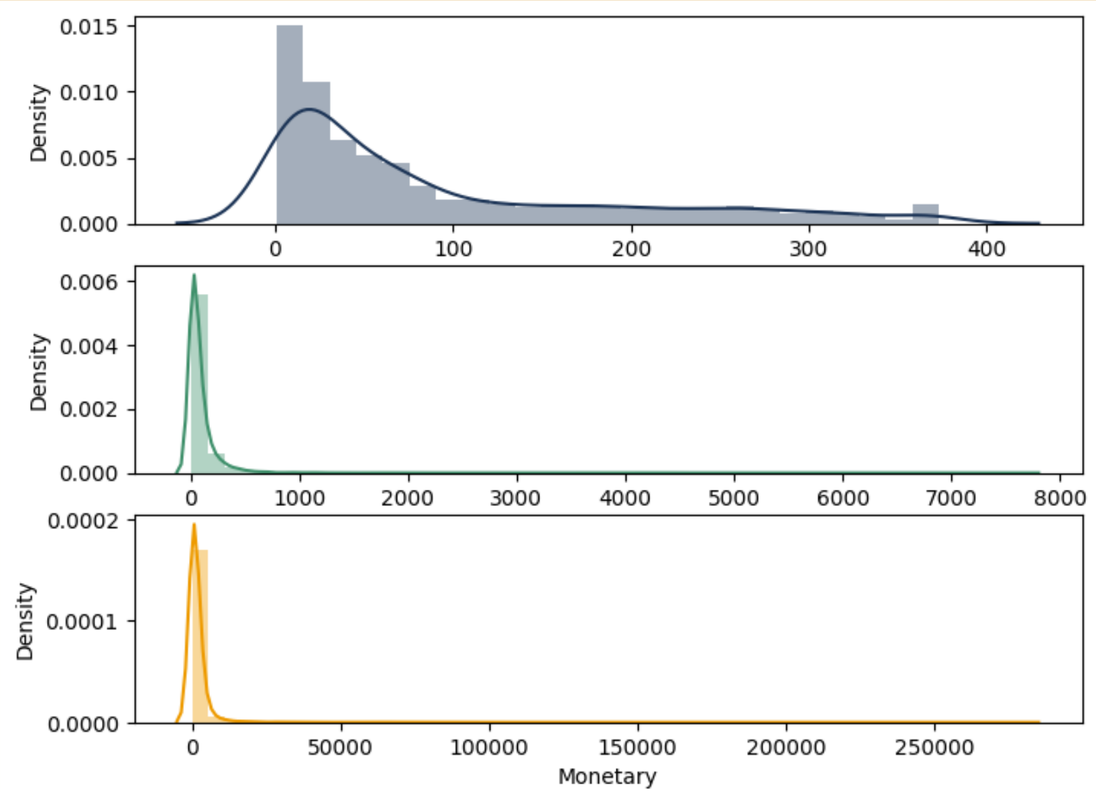


- **Platinum customers** = 1261 (less recency but high frequency and heavy spendings)
- **Gold customers** = 1325 (good recency,frequncy and moentary)
- **Silver customers** = 980 (high recency, low frequency and low spendings)
- **Bronz customers** = 772 (very high recency but very less frequency and spendings)

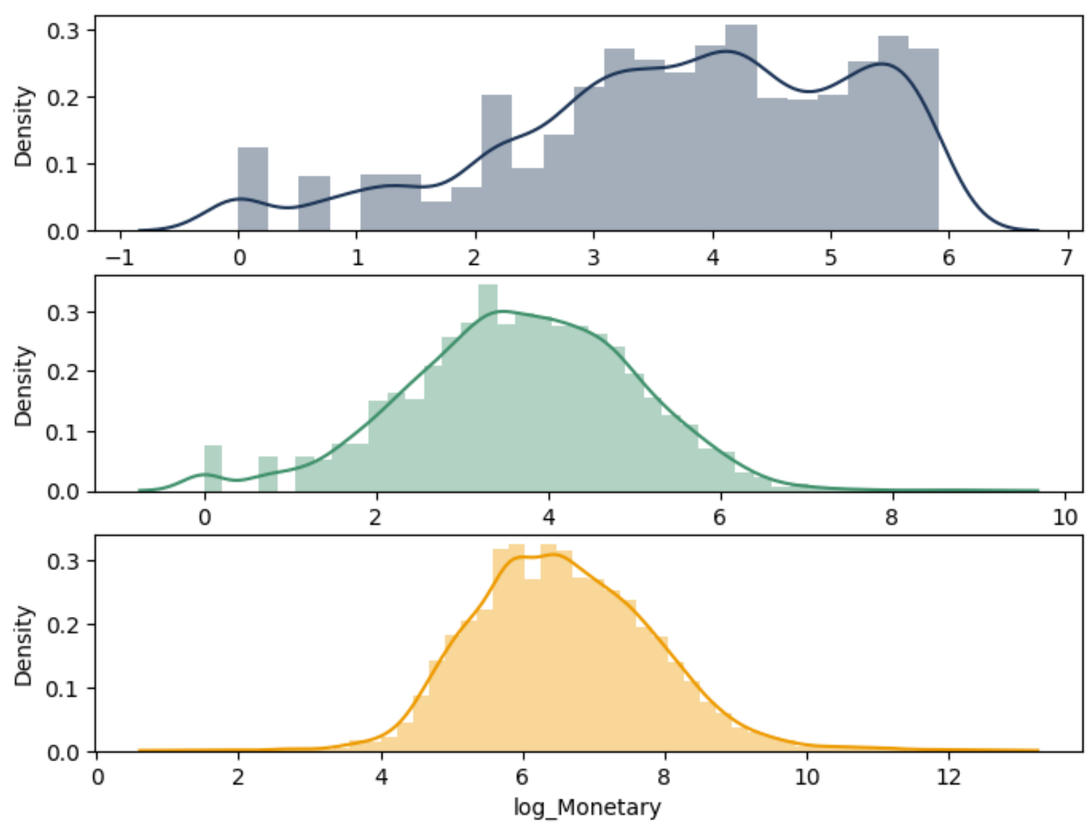
NUMERIC DISTRIBUTION

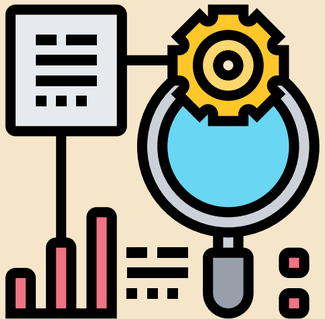
DS 21

ORIGINAL DATA RFM



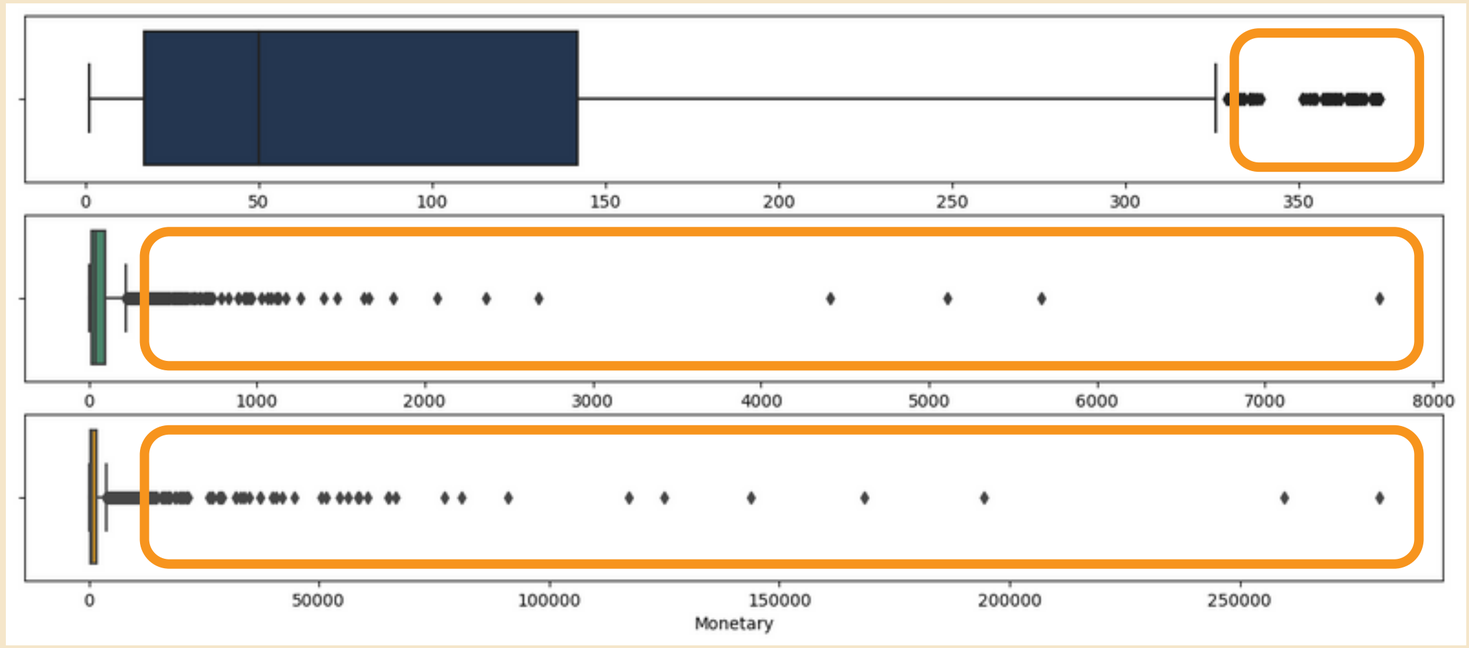
LOG TRANSFORMATION DATA RFM





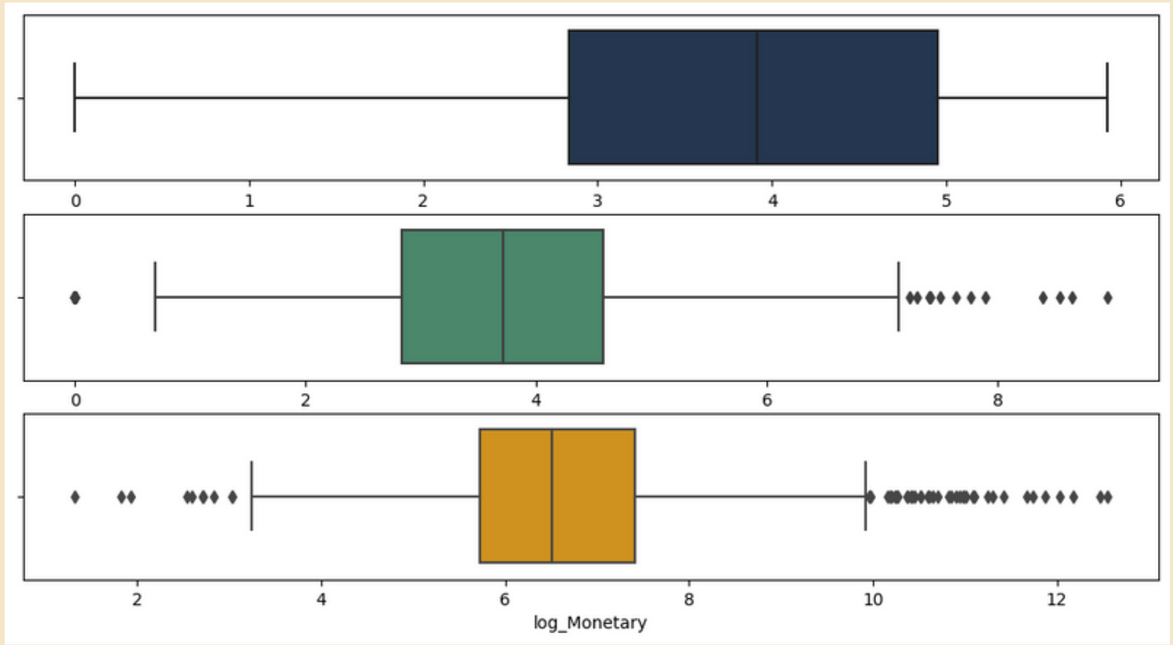
OUTLIER HANDLING

DS 21



ORIGINAL DATA RFM

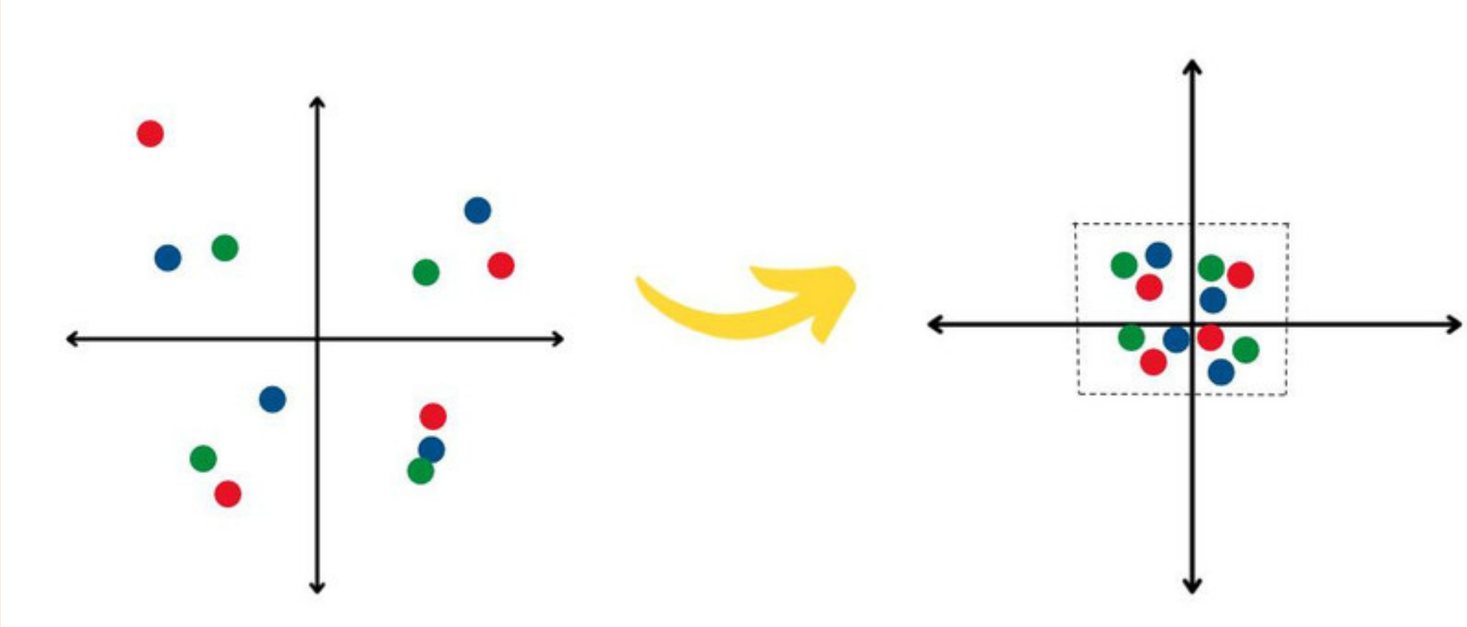
LOG TRANSFORMATION DATA RFM



K MEANS CLUSTERING

DATA SCALING

StandardScaler to normalize the data so that the data used does not have large deviations.



ACTUAL DATA

AFTER STANDARDIZATION

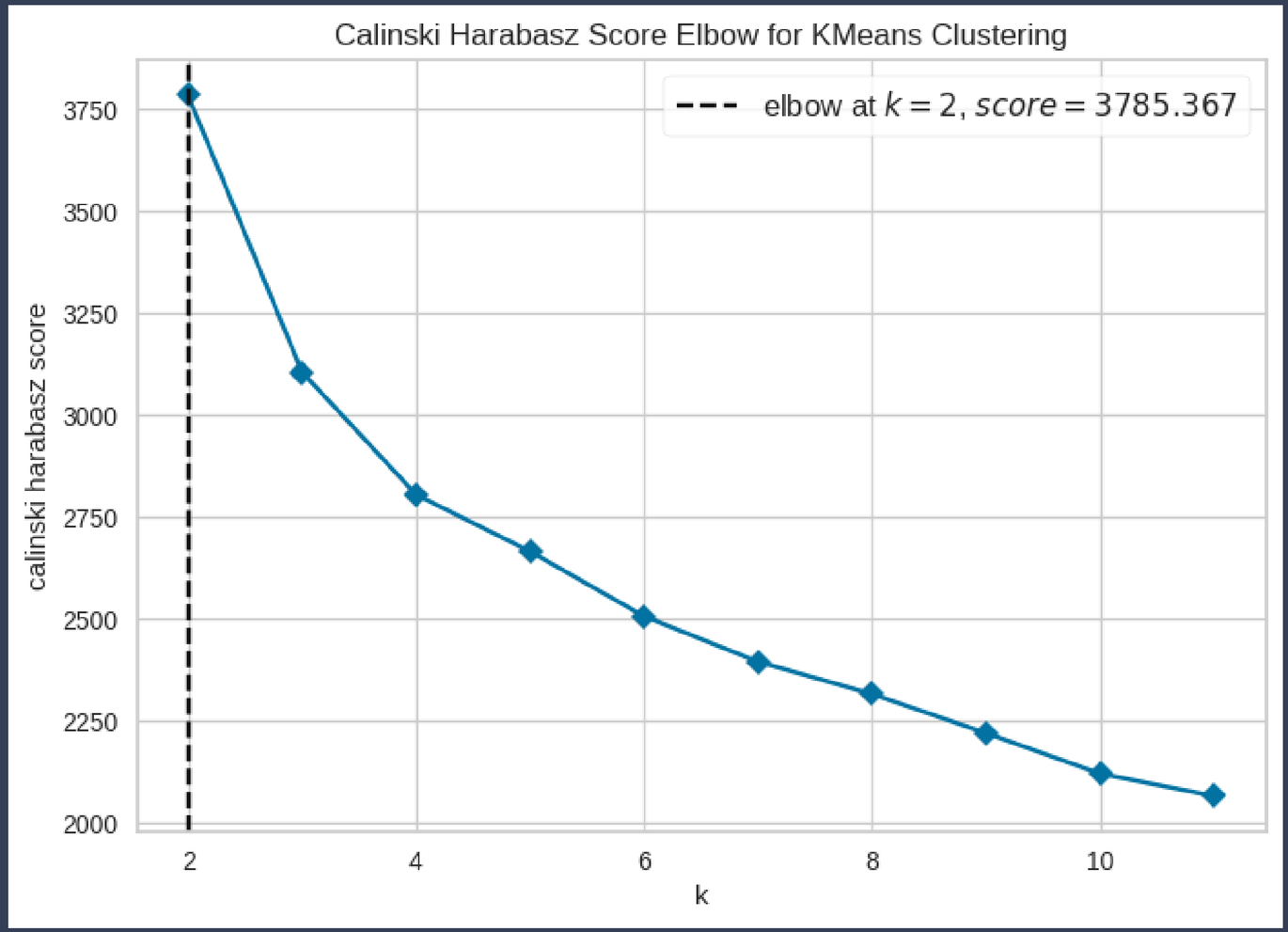
```
# taking only values of recency and monetory in X.
X = rfm_df_log.values

# standardising the data
scaler = StandardScaler()
X_std = scaler.fit_transform(X)
```

DS 21

ELBOW METHODE

Before doing clustering, it would be better to determine the best and right number of clusters first.

According to the graphic (Elbow method), the angle change starts to occur at point 2, then the correct K value for **K-Means Clustering is K = 2**.



VISUALISASI CLUSTERING

From the results of this customer clustering and visualized with a scatterplot as shown below.
This diagram shows the distribution of customer data which is divided into clusters according to the K-Means Clustering algorithm.

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_based_on_freq_mon_rec										
0	25.081465	1	372	177.574631	1	7676	4124.197333	120.03	280206.02	1829
1	140.898764	1	373	27.065763	1	232	535.692297	3.75	77183.60	2509

- **Cluster 0 has low recency but very high frequency and monetary**. Cluster 0 conatins 1829 customers. **Thus generates more revnue to the retail business**
- **Cluster 1 has high recency but they are frequent buyers and spends very low money than other customers as mean monetary value is very low**.

ANALYSIS

DS 21

RFM ANALYSIS

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
RFM_Loyalty_customer										
Platinum	19.550357	0	140	225.884219	20	7676	5253.556788	316.25	280206.02	1261
Gold	63.433962	0	372	57.012075	1	521	1162.987662	114.34	168472.50	1325
Silver	125.178571	1	373	24.337755	1	98	579.452461	6.90	77183.60	980
Bronz	217.585492	51	373	10.958549	1	39	199.030725	3.75	660.00	772

Based on RFM analysis. We had 4 clusters/Segmentation of customers.

K MEANS CLUSTERING

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_based_on_freq_mon_rec										
0	25.081465	1	372	177.574631	1	7676	4124.197333	120.03	280206.02	1829
1	140.898764	1	373	27.065763	1	232	535.692297	3.75	77183.60	2509

Based on K MEANS Clustering. We had 2 clusters/Segmentation of customers based on K value for K-Means Clustering

Observations:

Cluster 1 is heterogenous in nature. It comprises Platinum and Gold Customers.

Cluster 0 is heteregenuous in nature. It comprises Silver and Bronz Customers.

BUSINESS _____ RECOMMENDATION

Platinum Customers: High revenue generating and frequent buyers.

- Offers VIP members with features providing the latest information about new products etc.
- give promo buy 2 get 3.

Gols Customer: customer whose purchases are fairly frequent and generate moderate revenue.

- Give vouchers (max 5%) to attract users' attention with a maximum redemption time limit of 1 week.
- Provide a big sale event voucher (remaining warehouse stock) to impress many discounted items at affordable prices.

BUSINESS _____ RECOMMENDATION

Silver Customers: customer who are less active and are not bery frequent buyers and generate low revenue.

- Provide special discount promos for new users to make transactions.
- Provide free shipping for the first 3 transactions, etc.

Bronz Customer: customers generating very low revenue and are occasional buyers

- Do a reminder via email or WhatsApp at least once a week.
- Provide recommendations for the best-selling items to remind users of retail products.
- Provide information about developments and changes in e-commerce applications such as the ease of online transactions now, the existence of one-day shipping services, better application display, etc.

DS 21

THANK YOU

BY YUNITA APRILIA

yunitaaprilia054@gmail.com
linkedin.com/in/yunitaapril/
github.com/yunitaapril