UNIVERSITY OF WESTMINSTER

MSc BUSINESS INTELLIGENCE ANALYTICS

MODULE CODE: 7BUIS010W

MODULE TITLE: DATA WAREHOUSING AND BUSINESS INTELLIGENCE

ASSIGNMENT: COURSEWORK 2

STUDENT NAME: POONAM H RANGPARIYA

STUDENT ID: W1878136

SUBMISSION DATE: 04 JULY 2023

TABLE OF CONTENTS

1. **Data Understanding**
   - **1.1 Checking of Missing Value and data cleaning**
   - **1.2 Distribution Analysis**
   - **1.3 Statistical Exploration**
   - **1.4 Correlation Analysis**

2. **Customer Segmentation with K-Means**
   - **2.1 Building of K-means model in python**
   - **2.2 Justification of K value**
   - **2.3 Testing of K means model in Python**

3. **Review of Results**
   - **3.1 Identification of business value customer segments**
   - **3.2 Justification of Business Value from the formed Clusters**

4. **Datamart Design**
   - **4.1 Identification of Dimensions**
   - **4.2 Justification of Selected Dimensions**
   - **4.3 Identification of Measures**
   - 4**.4 Justification of Selected Measures**

1. Data Understanding

The dataset contains columns like CustomerID, Frequency, Recency, Monetary, rankF, rankR, rankM, groupRFM, Country.

1.1 Checking for Missing Value and data cleaning

The most important part before performing any data analysis is to check if there are any missing value in the data or any NaN value in the dataset, because it can lead to wrong or biased analysis, hence it is utmost necessary to remove them.

```
data.isna().sum()
```

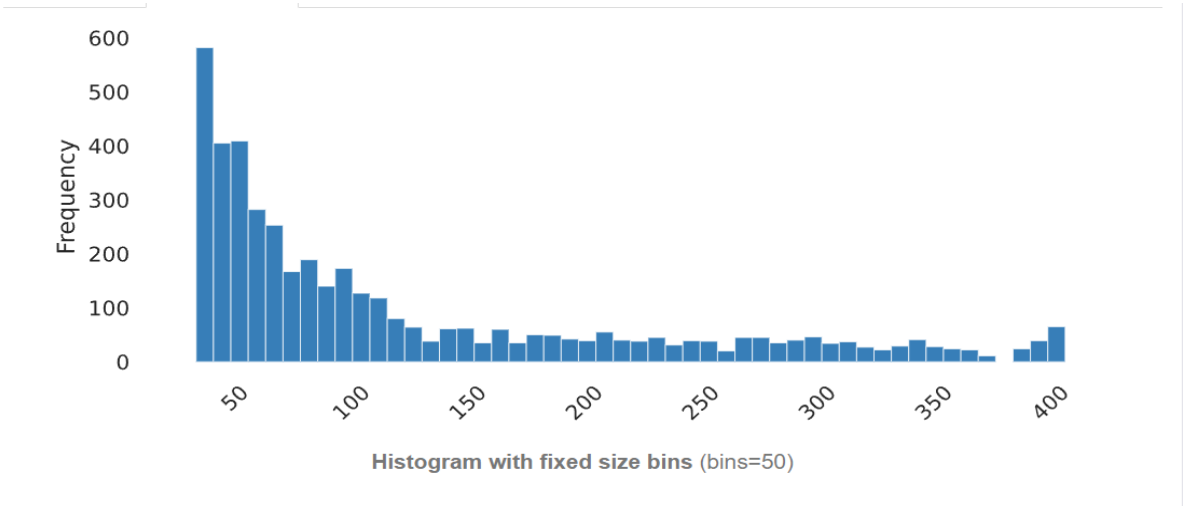
```
CustomerID     0
Frequency      0
Recency        0
Monetary       0
rankR          0
rankF          0
rankM          0
groupRFM       0
Country        0
dtype: int64
```

From the above figure we can see that there are no missing value in the dataset and so we can progress further with the analysis

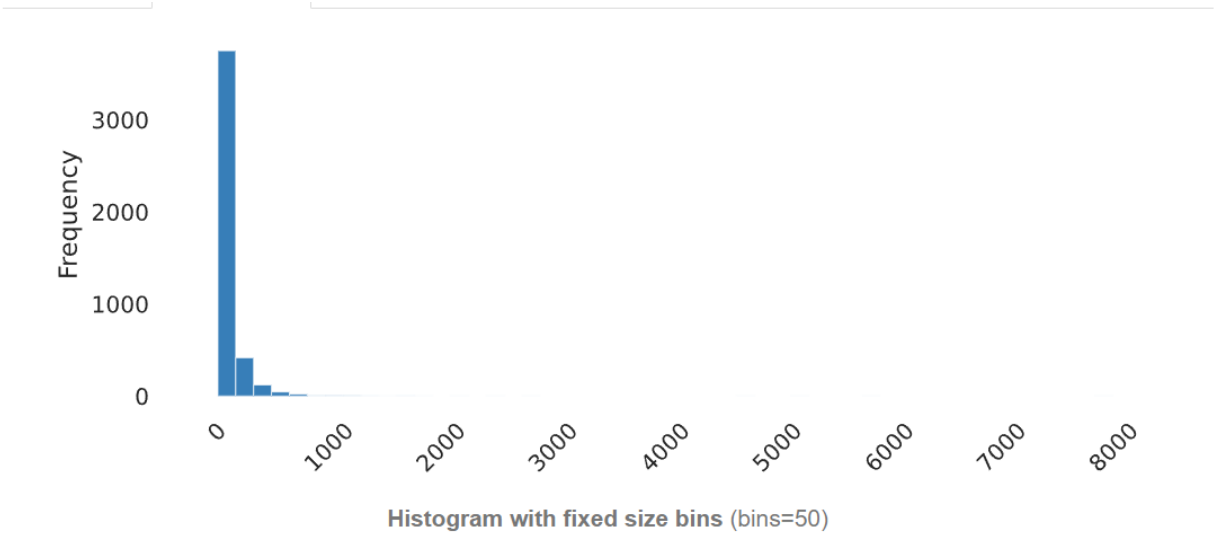
1.2 Distribution Analysis

Distribution analysis is necessary to determine the data is distributed or see the range of the data.

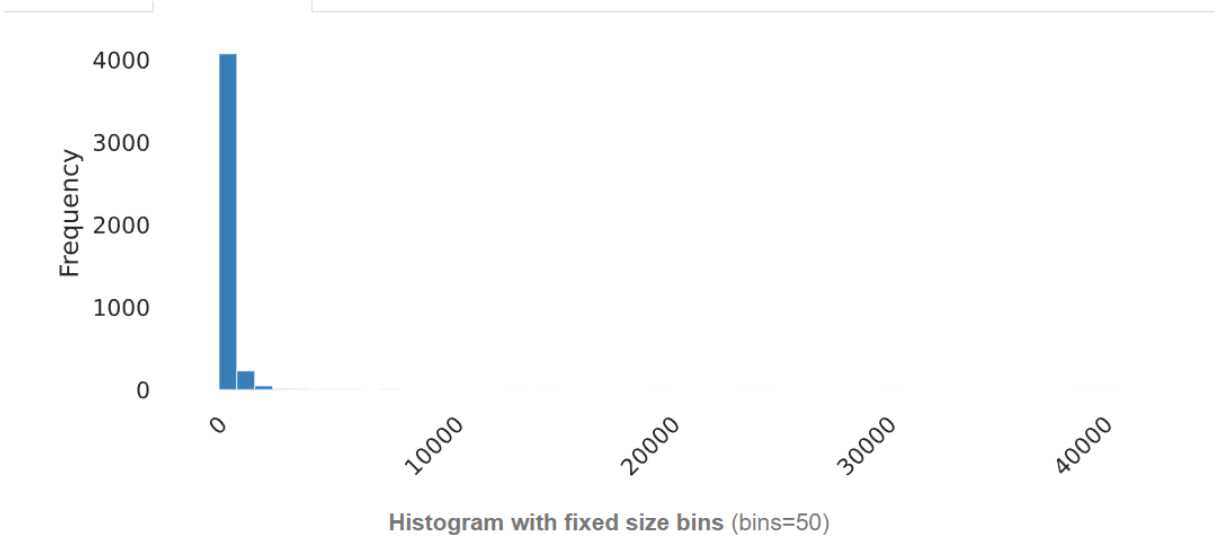
Here we can see the distribution of variables such as Frequency, Monetary, Recency, rankR, rankF, rankM as they are numerical data.



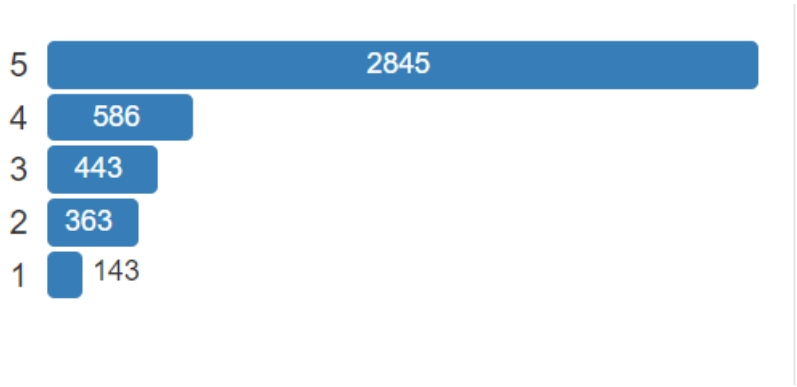
Above is the Histogram of Recency, we can see that the data is right skewed and the most of the values lies in range 500-150.



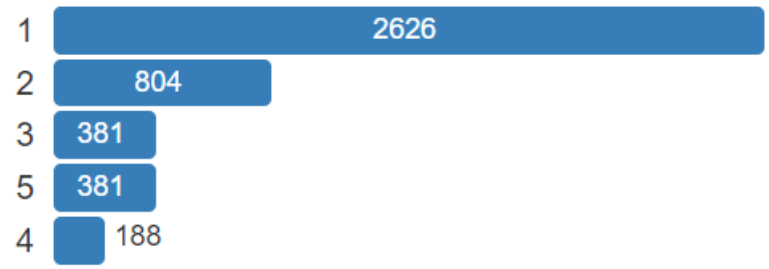
Above is the Histogram of Frequency, we can see that the data is right skewed and the most of the values lies in the range in 0-1000



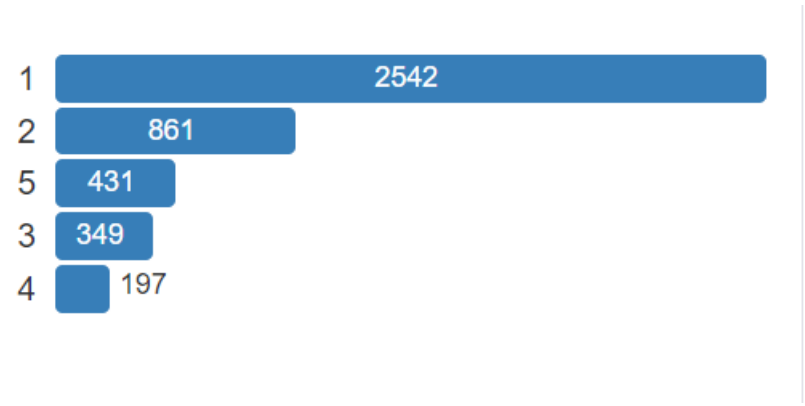
Above is the Histogram of Monetary, we can see that the data is rightly skewed and the most of the value lies in the range of 0-5000.



Above is the image of rankR, from this we can see that highest of customer falls in the rank 5 that the dataset contains more of the most recent customers.



The above image is the bar chart of rankF. The most of the customers comes into rank1 that means the dataset contains most of the customer that are not very frequent (very low visits).



The above image is the bar chart of rankM. The most of the customers comes into the rank 1 that means the dataset contains most of the customer that are low spender.

1.3 Statistical Exploration

Statistical exploration is used to better understand the data such as mean, median, mode, min, max, skewness etc.

Quantile statistics

Minimum	1
5-th percentile	4
Q1	17
median	42
Q3	102
95-th percentile	317.1
Maximum	7983
Range	7982
Interquartile range (IQR)	85

Descriptive statistics

Standard deviation	232.28177
Coefficient of variation (CV)	2.4951115
Kurtosis	486.9167
Mean	93.094749
Median Absolute Deviation (MAD)	31
Skewness	18.16456
Sum	407755
Variance	53954.823
Monotonicity	Not

The above image is the statistics of Frequency and we can see that the minimum value is 1 and the max value is 7983 that means the least time the customer visited is 1 and the highest time is 7983.

Quantile statistics

Minimum	0
5-th percentile	11.1685
Q1	53.0175
median	130.265
Q3	303.4525
95-th percentile	1018.294
Maximum	41376.33
Range	41376.33
Interquartile range (IQR)	250.435

Descriptive statistics

Standard deviation	1283.6995
Coefficient of variation (CV)	3.9835032
Kurtosis	606.1896
Mean	322.25392
Median Absolute Deviation (MAD)	96.08
Skewness	22.3904
Sum	1411472.2
Variance	1647884.4
Monotonicity	Not monotonic

The above is the statistics of Monetary, we can see that the minimum is 0 and the maximum is 41376.33 that means the lowest amount spent by the customer is 0 and the highest amount spent by the customer is 41376.33

Quantile statistics

Minimum	33
5-th percentile	35
Q1	49
median	83
Q3	175
95-th percentile	345
Maximum	406
Range	373
Interquartile range (IQR)	126

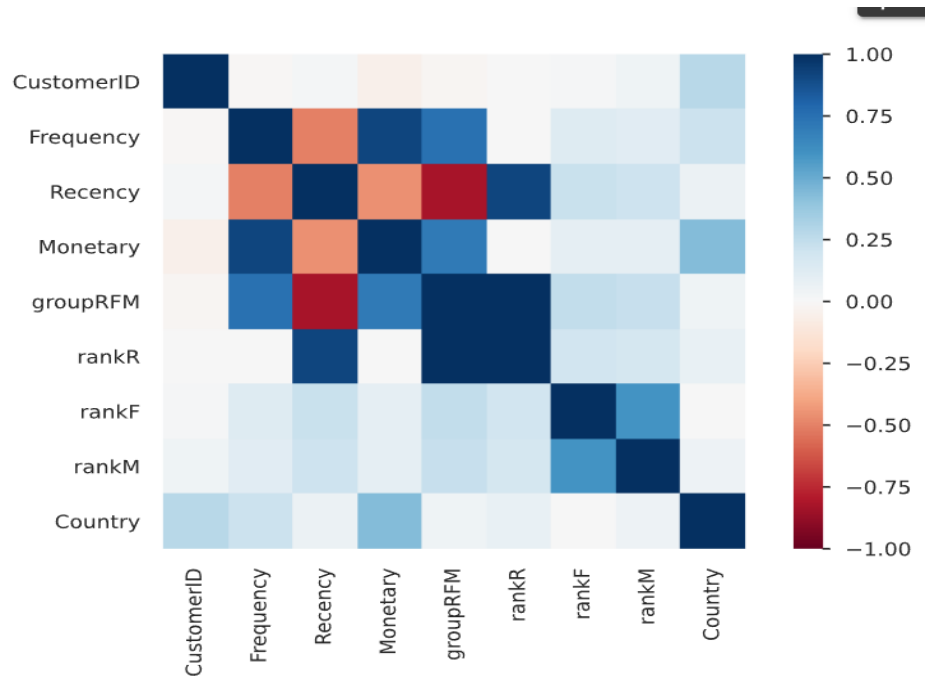
Descriptive statistics

Standard deviation	100.70475
Coefficient of variation (CV)	0.8088587
Kurtosis	0.43720486
Mean	124.50228
Median Absolute Deviation (MAD)	41
Skewness	1.2515912
Sum	545320
Variance	10141.448
Monotonicity	Not monotonic

The above is the statistics of the Recency, we can see that the minimum is 33 and the maximum is 406.

1.4 Correlation Analysis

Correlation between two variables indicated how linearly two variables are related to each other.



From the above heatmap, we can see that Recency is highly correlated with groupRFM

And Recency, Frequency, Monetary are highly correlated with groupRFM.

2. Customer Segmentation with K-Means

The need for Kmeans

All the customers segmented based on the above rfm value doesn't provide much information about the behaviour of customers and how they are represented on the three-dimension plot. Because it labels the customer's recency, frequency, and monetary based on only one column that is itself. It doesn't consider the segmentation based on all the three combined. While, the Kmeans considers all the three and group_rfm as an input and clusters the customers based on it.

2.1 Building of K-means model in python

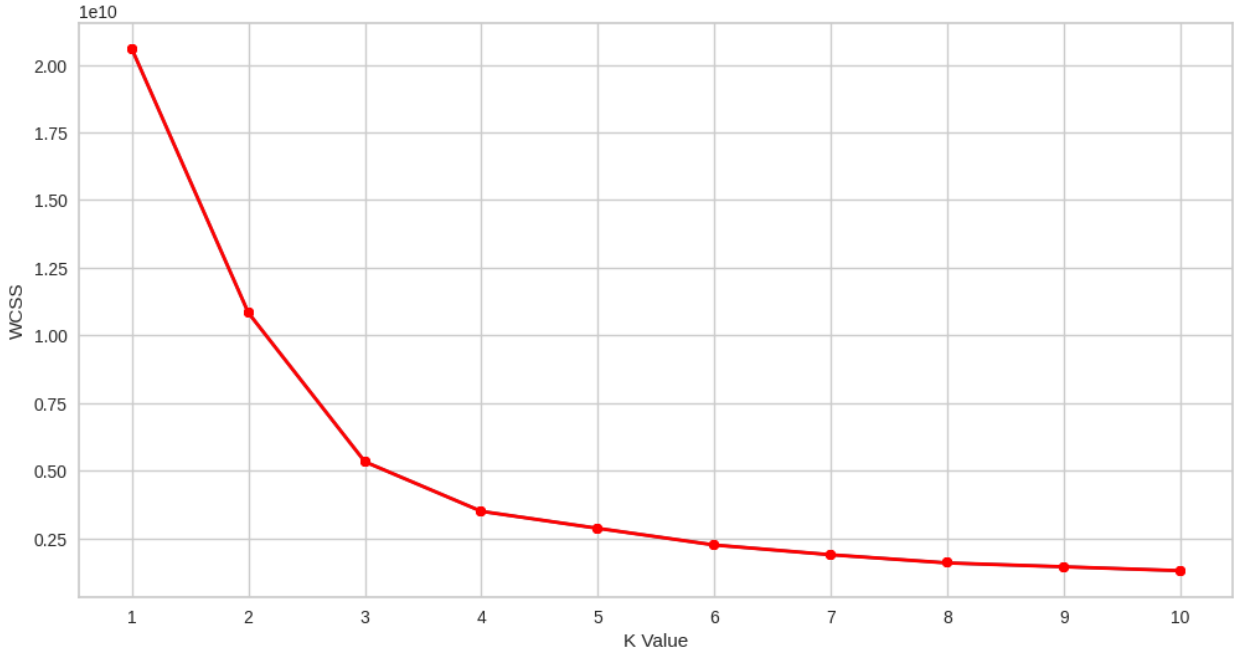
For the Kmeans algorithm, we need to figure out our inputs. So here we used monetary, frequency, recency, group_rfm

The reason to avoid rankF, rankR, rankM as they are correlated with group_rfm.

Now, we calculate the within-cluster sum of squared errors (WCSS) for different values of k. Next, we choose the k for which Wcss first starts to diminish. The Code for the same is given below. The Elbow method is a heuristic used in determining the number of clusters.

```
[ ]  #Importing KMeans from sklearn
     from sklearn.cluster import KMeans
     wcss=[]
     for i in range(1,11):
         km=KMeans(n_clusters=i)
         km.fit(X)
         wcss.append(km.inertia_)
     #The elbow curve
     plt.figure(figsize=(12,6))
     plt.plot(range(1,11),wcss)
     plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
     plt.xlabel("K Value")
     plt.xticks(np.arange(1,11,1))
     plt.ylabel("WCSS")
     plt.show()
```

The output of the elbow graph is as below.



From the above graph we can see that the elbow can be found at k=3. So, the optimal number clusters that can be formed is 3.

2.2 Justification of K value

The next important step is the justification of K value shown by above graph. Here we are using silhouette score. It is a measure used to evaluate the quality of clustering results. It provides a way to assess how well the data points within each cluster are separated from each other and how well they are assigned to their respective clusters. The range for the score is -1 to 1, where a higher value indicates better clustering results. A score close to 1 suggests that the data points are well separated and appropriately assigned to their clusters, while a score close to -1 implies that the data points may have been assigned to incorrect clusters.

```
best_score = -1
best_clusters = -1

# Iterate over different cluster numbers and calculate silhouette scores
for n_clusters in range(min_clusters, max_clusters + 1):
    # Fit K-means model to the data
    kmeans = KMeans(n_clusters=n_clusters)
    cluster_labels = kmeans.fit_predict(X)

    # Calculate the silhouette score for the current cluster number
    score = silhouette_score(X, cluster_labels)

    # Check if the current score is better than the previous best score
    if score > best_score:
        best_score = score
        best_clusters = n_clusters

# Print the best silhouette score and corresponding cluster number
print("Best Silhouette Score:", best_score)
print("Number of Clusters:", best_clusters)
```

Here the Kmean model is evaluate for each value of k (1 to 11) and the best score is updated each time when silhouette_score is generated and if the new score is better than the previous score.

Here for our case below is the output for silhouette score.

```
Best Silhouette Score: 0.5739944873205852
Number of Clusters: 3
```

So, we can know justify that the value of k that we analyzed from the Elbow method is the same given by the silhouette metrics.

2.3 Testing of K means model in Python

To test the Kmeans model we have predicted the labels for each customer that is from 0 to 2

```
km2 = KMeans(n_clusters=3)
y2 = km2.fit_predict(X)
data["label"] = y2
#The data with labels
data.head()
```

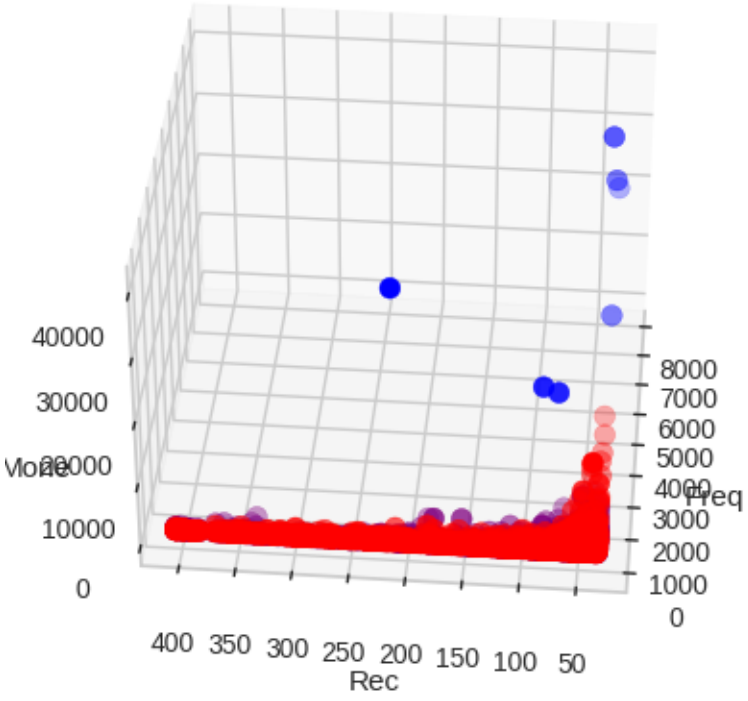
The number of customers in each cluster is

Label 0: 2189

Label 1: 2184

Label 2: 7

The output of the formed cluster in 3D is shown below.



Cluster Interpretation

There are three clusters formed in colour purple, red and blue.

The X -axis is Recency, Y-axis is Frequency and Z -axis is Monetary.

The Kmeans algorithm has divided the customers into various clusters based on the recency value, frequency value and monetary value.

Here the data is not linear so the clusters are not visible and the Kmeans can't distinguish between the clusters. If we increase the data points, we might be able to distinguish each type of customers. Below is our interpretation for each clusters formed.

Label 0 (Color: Purple) – It contains all the customers that has a recency rank of 1.

Label 1 (Color: Red) – It contains all the customers that has a recency value above 1.

Label 2 (Color: Blue) – It contains only the customers that has high value of all recency, frequency and monetary.

3.Review of Results

3.1 Identification of business value customer segments

From the cluster plot we can see that the cluster of color purple has more data points as compare to cluster of color red and the cluster of color blue has very few elements.

Here the features frequency and recency with label 1 are given more importance and the clusters are formed based on them.

The Customer with Label 0:

CustomerID	Frequency	Recency	Monetary	rankR	rankF	rankM	groupRFM	Country	label
15297	41	43	113.54	5	1	1	511	United Kingdom	0
15300	28	97	80.14	5	1	1	511	United Kingdom	0
15303	31	323	119.06	2	1	1	211	United Kingdom	0
15304	26	93	182.9	5	1	2	512	United Kingdom	0
15306	58	97	153.76	5	1	1	511	United Kingdom	0
15307	12	128	80.95	4	1	1	411	United Kingdom	0

The cluster with label 0 has all the lower ranked recency customers and lower ranked monetary customers.

The Customer with Label 1:

CustomerID	Frequency	Recency	Monetary	rankR	rankF	rankM	groupRFM	Country	label
12346	2	358	2.08	2	1	1	211	United Kingdom	1
12347	182	35	481.21	5	4	3	543	Iceland	1
12348	31	108	178.71	5	1	2	512	Finland	1
12349	73	51	605.1	5	2	4	524	Italy	1
12350	17	343	65.3	2	1	1	211	Norway	1
12352	95	69	2211.1	5	2	5	525	Norway	1

The cluster with label 1 has few lower ranked recencies that is 1 and rest are the higher ranked (or above rank 1 recency customers), Also it contains medium ranked monetary customers.

The Customer with Label 2:

CustomerI	Frequency	Recency	Monetary	rankR	rankF	rankM	groupRFM	Country	label
12744	229	84	25108.89	5	4	5	545	Singapore	2
12748	4642	33	15115.6	5	5	5	555	United Kin	2
14096	5128	37	41376.33	5	5	5	555	United Kin	2
14911	5903	34	31060.66	5	5	5	555	EIRE	2
15098	5	215	40278.9	3	1	5	315	United Kin	2
16029	274	71	24111.14	5	5	5	555	United Kin	2
17841	7983	34	20333.18	5	5	5	555	United Kin	2

The cluster with label 2 contains only high ranked recency, frequency and monetary value customers.

Based on the above properties of the clusters we can target each set of customers differently and can obtain business from them.

3.2 Justification of Business Value from the formed Clusters:

The customer labeled as 1 requires special attention as they may be at risk of leaving or have already left. It's important to show them targeted advertisements to retain their interest and improve their recency score. Gathering feedback from them will help address their concerns effectively. By providing personalized assistance, we can enhance their experience and increase the chances of retaining them.

The customer labeled as 0 represents a group with low recency scores. It's not necessary to allocate significant resources to this segment as they already show limited engagement. However, maintaining basic communication is essential to prevent further decline or potential churn.

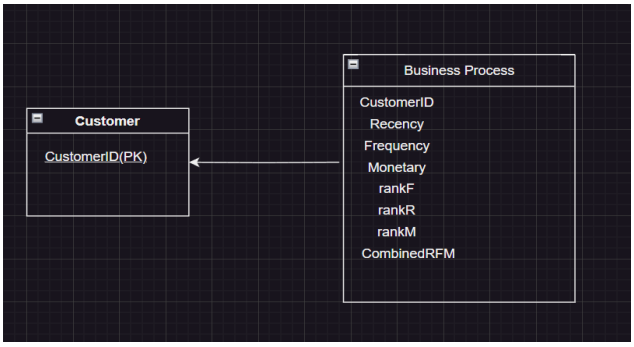
Customers labeled as 2 have high recency, frequency, and monetary value, indicating recent and valuable purchases. Allocating resources, such as targeted ads, special offers, and personalized communication, to this segment can yield positive results. Acknowledging their recent engagement and providing incentives can reinforce their loyalty and encourage repeat transactions.

In summary, the proposed approach focuses on targeted actions and resource allocation based on customer segmentation, with the aim of retaining at-risk customers, optimizing resource usage, and driving business growth.

4. Data Mart Design

A fact table represents the central table in a star schema or snowflake schema. It contains quantitative measures or metrics that are typically numerical and can be aggregated. Dimensions, on the other hand, are descriptive attributes that provide context to the measures in the fact table.

Below is the Star Schema for our model



Assumptions:

The fact table should ideally have additional dimensions (such as Invoice, Product, or Location) to provide a more comprehensive analysis. However, based on the variables provided, the CustomerID serves as the primary dimension in the fact table.

4.1 Identification of Dimensions.

There is only one dimensions in our solution:

Customer (CustomerID)

CustomerID: This variable represents a unique identifier for each customer. It could be considered as a dimension that provides individual customer context.

The rest of the dimensions that can be thought of to provide comprehensive analysis are:

Invoice (Invoice No, InvoiceDate, Quantity)

Product (StockID, UnitPrice)

4.2 Justification of Selected Dimensions

Customer Dimension: This dimension table can provide additional attributes and details about each customer.

The Invoice Dimension: This dimension provides information about Invoice Date and Quantity that are necessary to generate measures like recency (Invoice Date) and Quantity (in addition to price helps to generate monetary value of each customer)

The Product Dimension: This dimension provides information about Unitprice (in addition to quantity) is useful to generate measure like monetary value.

4.3 Identification of Measures

Measures involves numerical attributes and they are part of the fact table. It holds the data to be analyzed.

Monetary: The monetary value associated with each customer.

Frequency: The number of transactions made by each customer.

Recency: The time since the last transaction made by each customer.

RankM: The rank of each customer based on monetary value.

RankF: The rank of each customer based on frequency.

RankR: The rank of each customer based on recency.

4.4 Justification of Selected Measures

Monetary: The "Monetary" variable can indeed be considered a measure as it represents the monetary value associated with each customer. It provides quantitative information about customer spending habits and contributes to revenue-related analysis.

Frequency: The "Frequency" variable can be considered a measure as well. It represents the number of transactions made by each customer and provides quantitative insights into customer engagement and activity.

Recency: The "Recency" variable can also be categorized as a measure. It represents the time since the last transaction made by each customer and provides a quantitative measure of customer engagement recency.

RankM, RankF, RankR: These variables, representing the ranks of customers based on monetary value, frequency, and recency, respectively, can be categorized as measures. They provide quantitative rankings and allow for comparative analysis between customers based on these metrics.

Combined_rfm: It is the combination of all RFM so it is a derived attribute.

Therefore, the revised categorization includes CustomerID as a dimension, while Monetary, Frequency, Recency, RankM, RankF, and RankR are measures.