CUSTOMER SEGMENTATION

AASHA SURESH

Problem Statement:

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

Objective:

For large retail stores, the customers must be analysed by dividing into groups. Segmentation of customers can help the stores in building strategies to improve their sales. It can help them in finding their best customers who can be targeted during discount sales campaigns. Also, these kinds of segmentation can help the stores find 'at risk' or churning customers and get them back by offering various discounts. In most of the businesses 20% of customers contribute 80% share of the total revenue of a company. That's why finding this set of people is important and so is customer segmentation. We will look into how customers are segmented and various insights from the same in this project.

Data Description:

• 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

• Price: 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 Pre-processing and EDA:

As a first step in the pre-processing of data, we remove the duplicate and null values from the dataset. The data contained 5268 duplicate entries (about ~1%).

As per the data, InvoiceNo starting with c implicates cancelled orders. There are 8905 such records which is also removed.

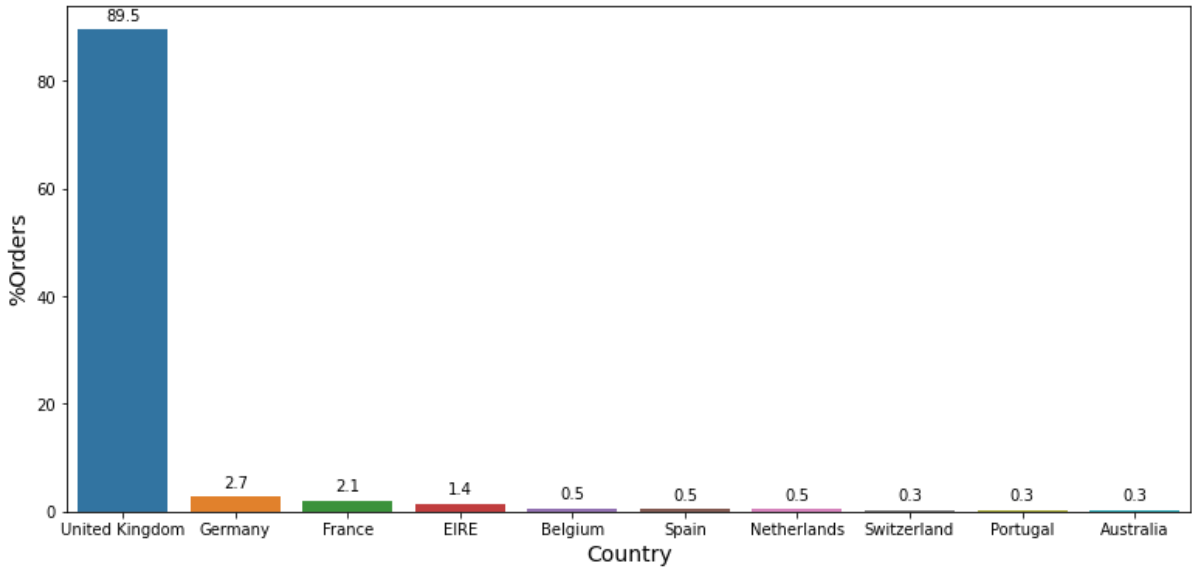
Then, neglect the data where amount is negative i.e. the item is cancelled or returned

Remove transactions with missing customer ID. Amongst 541909 records, 135080 records (about ~25%) are missing Customer ID values

Remove transactions with 0 Monetary value. 40 records are available as such.

EDA:

From the bar plot, it is observed that nearly 90% of the customers are from UK.



Total number of products available: 3684

Number of Transactions occurred per Invoice: 22190

Total number of customers: 4372

	products	transactions	customers
quantity	3684	22190	4372

RFM:

We use the RFM technique to perform customer segementation:

Recency- how recently the purchase is done by any customer. It is obtained by taking the most recent date as reference and subtracting the InvoiceDate column from the same.

Frequency- how many times the customer has made purchase. It is obtained with by the number of Invoices available for each customer ID.

Monetary- for how much amount is the purchase made by the customer. Amount for each transaction is obtained by multiplying Unit Price and Quantity and then it is grouped by CutomerID column for each customer.

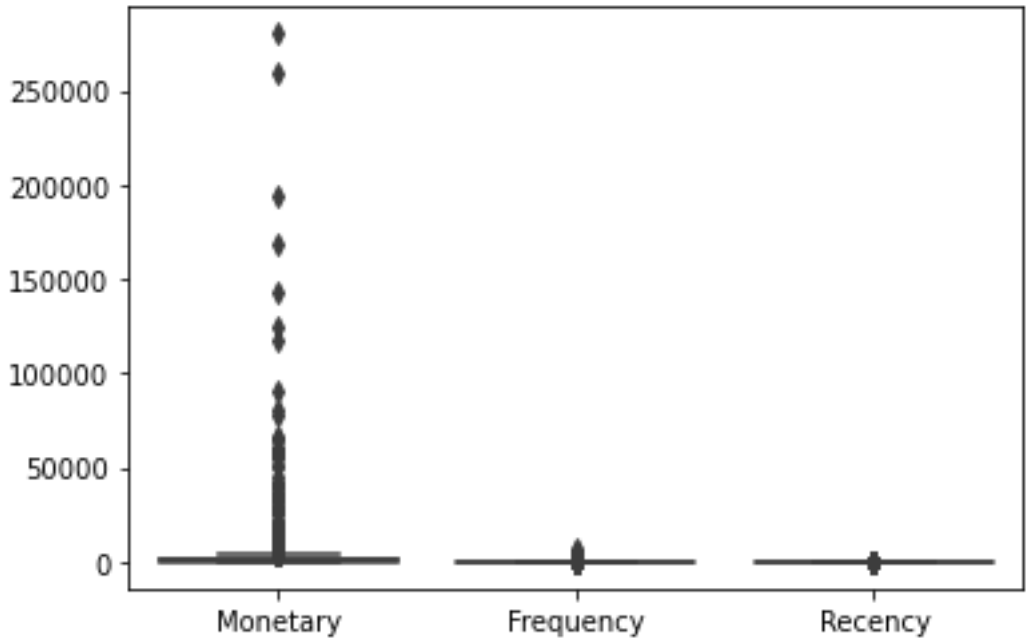
After obtaining the Monetary value for each transaction, we use the .describe() method for further analysis.

```
df['Amount'].describe()
```

```
count    392732.000000
mean         22.629195
std         311.083465
min           0.000000
25%           4.950000
50%          12.390000
75%          19.800000
max      168469.600000
Name: Amount, dtype: float64
```

This shows that a total of 392732 transactions were made.
Average amount per transaction is `22.629195`
Min amount of transaction= 0
Max amount of transaction= 168469.600000

Outlier Analysis:
We use box plots for Monetary, Recency and Frequency to vizualise the outliers. The below plot sho ws that Monetary has a lot of outliers. We use InterQuartile ranges to remove the outliers from all 3 variables.



Algorithm:

I have chosen K-Means clustering for this project. 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 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.

Assumptions:

Key assumptions of the algorithm are:

1. The variables should be distributed symmetrically
2. Variables should have similar average values
3. Variables should have similar standard deviation values

To satisfy these assumptions, we use StandardScaler function is used to normalize the data. It remov es the mean and scales each feature/variable to unit variance

Model Evaluation:

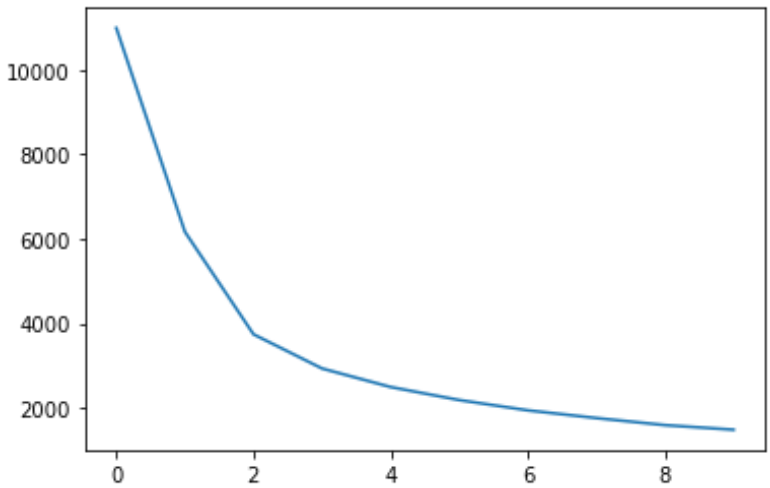
The techniques used for evaluation are:

- Elbow method
- Silhouette method

Elbow Method

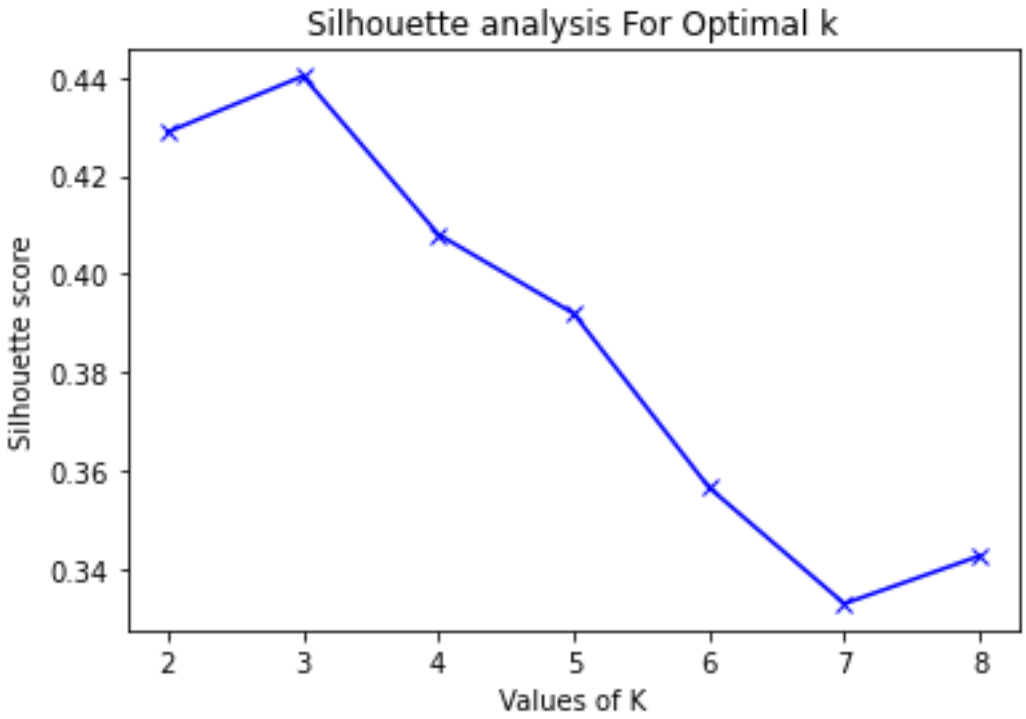
Elbow method identifies the optimal number of clusters with the help of WSS[total within-cluster sum of squares]. We find the best cluster which has the minimum WSS value. It is obtained by plotting each number of clusters with their respective error value to get an elbow curve. After a certain point in the elbow curve, we find the point where the average distance from the centroid falls suddenly ("Elbow").

The optimal number of clusters which is 3 from the elbow method in our case.



Silhouette Method:

The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). To plot the silhouette curve, choose a range of values of k and then plot the silhouette coefficient for each value of k. By doing so, we get the below curve,



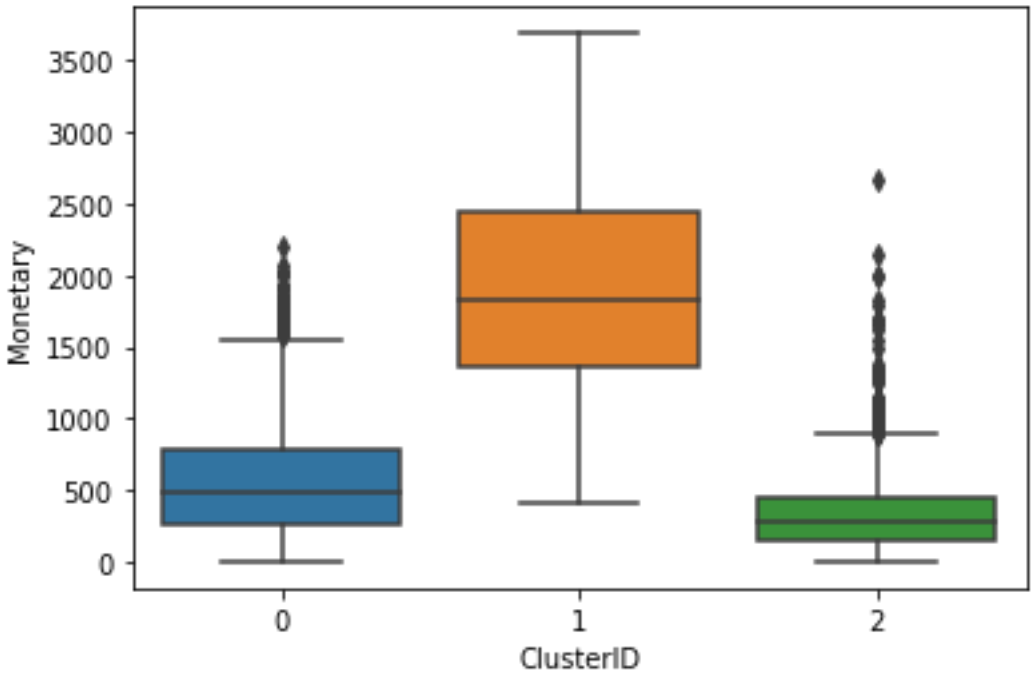
Silhouette scores lie between -1 to +1:

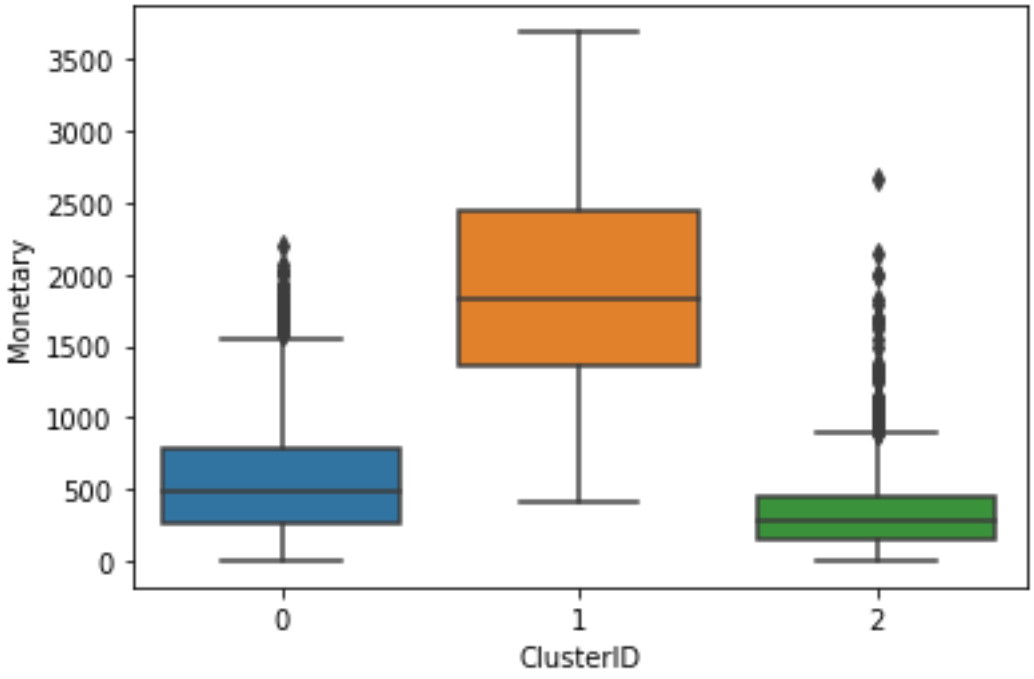
- A score of 1 denotes the best, meaning that the data point is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1.
- Values near 0 denote overlapping clusters.

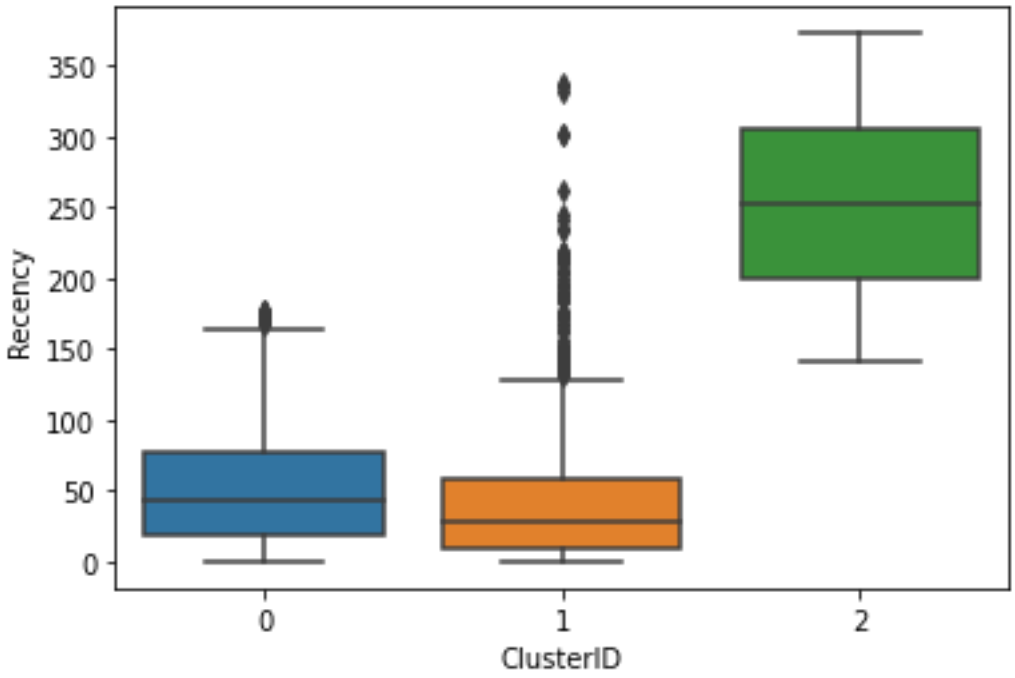
So, we can say that the ideal number of clusters in our case is 3 which confirms the number of clusters obtained from elbow method. The silhouette Method is used in combination with the Elbow Method for a more confident decision

Inferences:

We can extract this information now for each customer that can be used to map the customer with their relative importance by the company from the above process:


3500
3000
2500
2000
1500
1000
500
0
Monetary
0
1
2
ClusterID


3500
3000
2500
2000
1500
1000
500
0
Monetary
0
1
2
ClusterID



Based on the clustering and visualization from the above box plots, we can say that:

Cluster2:

Customers of cluster 2 made recent purchases but they are less frequent. They usually make lost cost purchases

Cluster 1:

Customers of cluster 1 made no recent purchases. But they were very frequent and made high cost purchases in the past.

Cluster 0:

Customers of cluster 0 are consistent buyers based on needs.

Future Possibilities:

Since Cluster 2 consists of new customers. we can focus on more promotional ads for them. So as to sustain them. Also, we should improve the customer care service for them

Cluster 1 members made purchases in the past; they are lost customers. We shoud understand why they left so as not to happen again.

Cluster 0 members are best customers. There is no need to focus on them for any promotions or discounts.

Conclusion:

We have used K-Means clustering to understand customer data and divided the customers into 3 clusters. Based on the above understanding the store has focus and improve the discounts, promotional ads and customer service for the respective group of customers. This can help them with increased sales.