CUSTOMER SEGMENTATION USING K-MEANS

Final Project Report

Subject: - ISM6136.006F22 Data Mining

Submitted by:

G. SIVA KRISHNA REDDY	**U55391402**
P. DINESH DATTA	**U96239813**
P. SANTHI MADHURYA	**U93636652**
M. AKHILASRI REDDY	**U82181510**
V. SRI VAISHNAV	**U29421544**

Major in

BUSINESS ANALYTICS AND INFORMATION SYSTEMS

Under the guidance of

DR. MOHAMMADREZA EBRAHIMI

MUMA COLLEGE OF BUSINESS

UNIVERSITY OF SOUTH FLORIDA

INDEX Page. No.

1. INTRODUCTION TO CUSTOMER SEGMENTATION

1.1. INTRODUCTION

Customer segmentation is a tool for the businesses to strict to their strategy and tactics with and target their customers in a better way. Every customer is unique, and their journey is different so single approach is not going to work for everyone. In this case, customer segmentation comes into picture and plays a vital role.

1.2.WHAT IS CUSTOMER SEGMENTATION?

Customer segmentation is a process by which we divide customers into segments or groups based on demographics or behaviors, so we can market to those customers more effectively. Customer segmentation will help us to understand the customer in a better way and meet their unique needs. By enabling customer segmentation:

- The retention rate can be maximized
- Customer experience can be improved
- Marketing costs can be reduced.
- Marketing campaigns (Email marketing, SMS, Facebook ads)

1.3. CUSTOMER SEGMENTATION FINDINGS

Insight into these factors make it possible for us to identify the right actions and target the execution

- **Customer needs:** Needs are the customers motivation on to buy any product or service
- **Customer Behavior:** Behavior, like purchase frequency and transaction size, helps us understand the market dynamics and how far we have come
- **Customer Demographics:** Demographics and socio-demographics help us describe the target customer group(s) in general terms

1.4. CUSTOMER SEGMENTATION MODELS

- Demographic Segmentation
- Geographic Segmentation
- Psychographic Segmentation
- Technographic Segmentation
- Behavioral Segmentation
- Needs-based Segmentation
- Value-based Segmentation

2. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics. Below are the observations captured using exploratory data analysis:

This data frame contains 8 variables that correspond to:

- Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- Stock Code: 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.
- Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- Unit Price: Unit price. Numeric, Product price per unit in sterling.
- Customer: 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.
- This dataset is an unlabeled set and contains 500000+ Rows with 8 variables. It contains transactions made by 37 unique countries.

- The dataset consists of the majority of Customers from the UK region (Fig 1.1).

	Country	CustomerID
35	United Kingdom	3950
14	Germany	95
13	France	87
30	Spain	31
3	Belgium	25
32	Switzerland	21
26	Portugal	19
18	Italy	15
12	Finland	12
1	Austria	11

Fig 1.1 Number of Customers based on the Countries

- Based on the transactions, 70 percent of the revenue is from the UK region followed by the Netherlands (Fig 1.2).

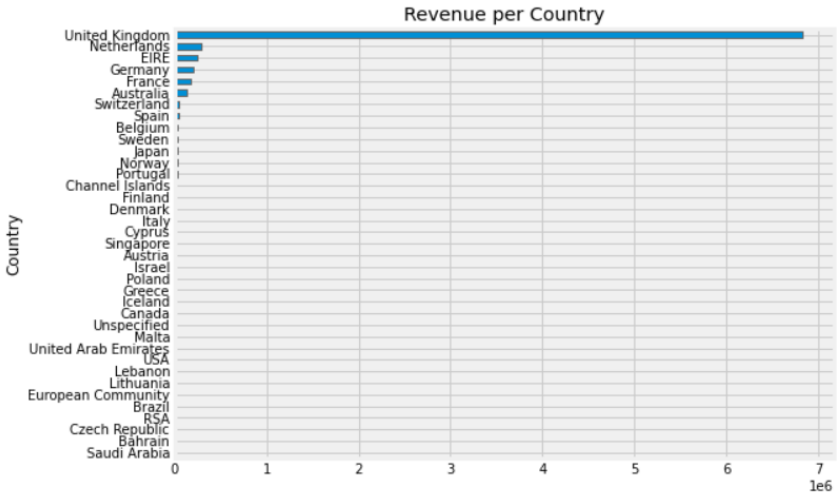


Fig 1.2 Revenue per country

- In some cases, stock code starts with D and they are offered under discount.
- There are few invoices starts with “C”, those transactions were cancelled.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	12/1/2010 9:41	27.50	14527	United Kingdom
8963	537159	22112	CHOCOLATE HOT WATER BOTTLE	6	12/5/2010 13:17	4.95	14527	United Kingdom
8964	537159	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/5/2010 13:17	4.95	14527	United Kingdom
8965	537159	21479	WHITE SKULL HOT WATER BOTTLE	1	12/5/2010 13:17	3.75	14527	United Kingdom
8966	537159	22114	HOT WATER BOTTLE TEA AND SYMPATHY	6	12/5/2010 13:17	3.95	14527	United Kingdom

- 16.46% of orders were cancelled in the dataset.

```
nb_products_per_basket.InvoiceNo = nb_products_per_basket.InvoiceNo.astype(str)
nb_products_per_basket['order_canceled'] = nb_products_per_basket['InvoiceNo'].apply(lambda x:int('C' in x))
len(nb_products_per_basket[nb_products_per_basket['order_canceled']==1])/len(nb_products_per_basket)*100
```

```
16.466876971608833
```

16.46% of transactions were cancelled.

3. DATA PRE-PROCESSING:

Data preprocessing can refer to manipulation or dropping of data before it is used to ensure or enhance performance and is an important step in the data mining process. Below are the insights captured for E-commerce dataset:

- Dataset contains 541909 entries with 8 columns.

```
df_data.shape
```

```
(541909, 8)
```

- Null values exist in the Customer ID column. Those Null values doesn't add much value to the dataset and so drop those ID's (Fig 1.4)

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
622	536414	22139	NaN	56	12/1/2010 11:52	0.00	NaN	United Kingdom
1443	536544	21773	DECORATIVE ROSE BATHROOM BOTTLE	1	12/1/2010 14:32	2.51	NaN	United Kingdom
1444	536544	21774	DECORATIVE CATS BATHROOM BOTTLE	2	12/1/2010 14:32	2.51	NaN	United Kingdom
1445	536544	21786	POLKADOT RAIN HAT	4	12/1/2010 14:32	0.85	NaN	United Kingdom
1446	536544	21787	RAIN PONCHO RETROSPOT	2	12/1/2010 14:32	1.66	NaN	United Kingdom

Fig 1.4 Null Values in Data set

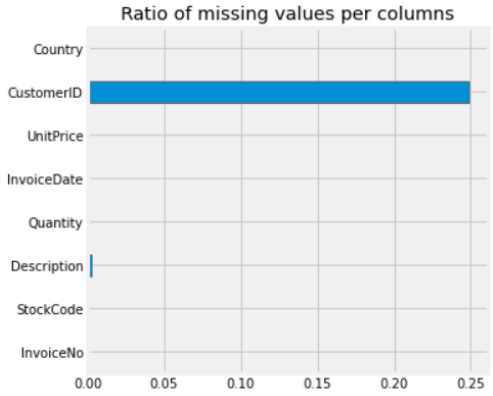


Fig 1.5 Percentage of null values in customer ID column.

- Below are the histograms for key metrics: Recency, Frequency and Monetary

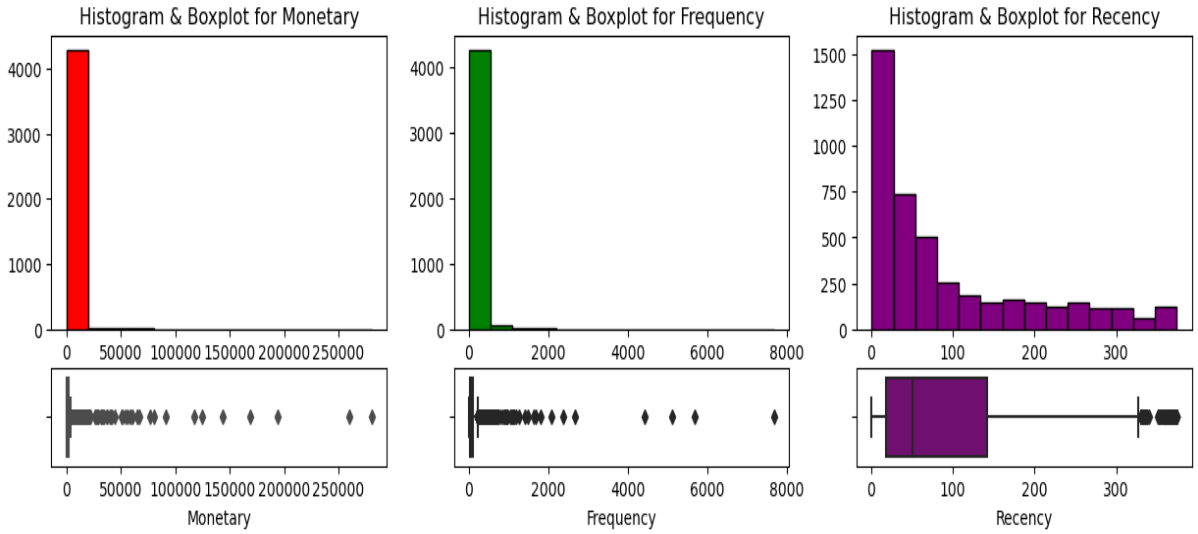


Fig 1.6 RFM metrics of the dataset.

1. Recency: How recently a customer has made a purchase.
2. Frequency: How often a customer makes a purchase.
3. Monetary Value: How much money a customer spends on purchases.

- Below is the histogram for Monetary dataset after removing outliers in the Dataset (Fig 1.7)

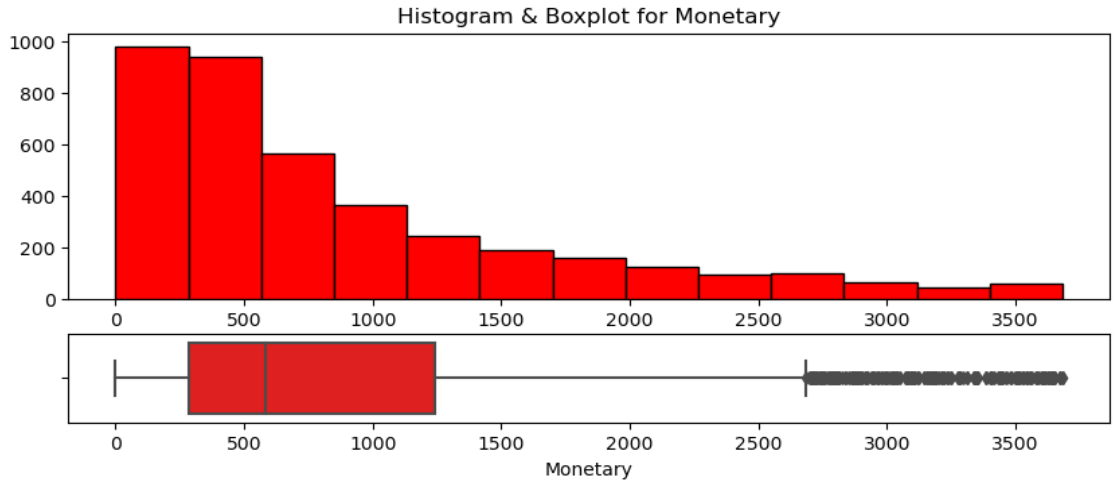
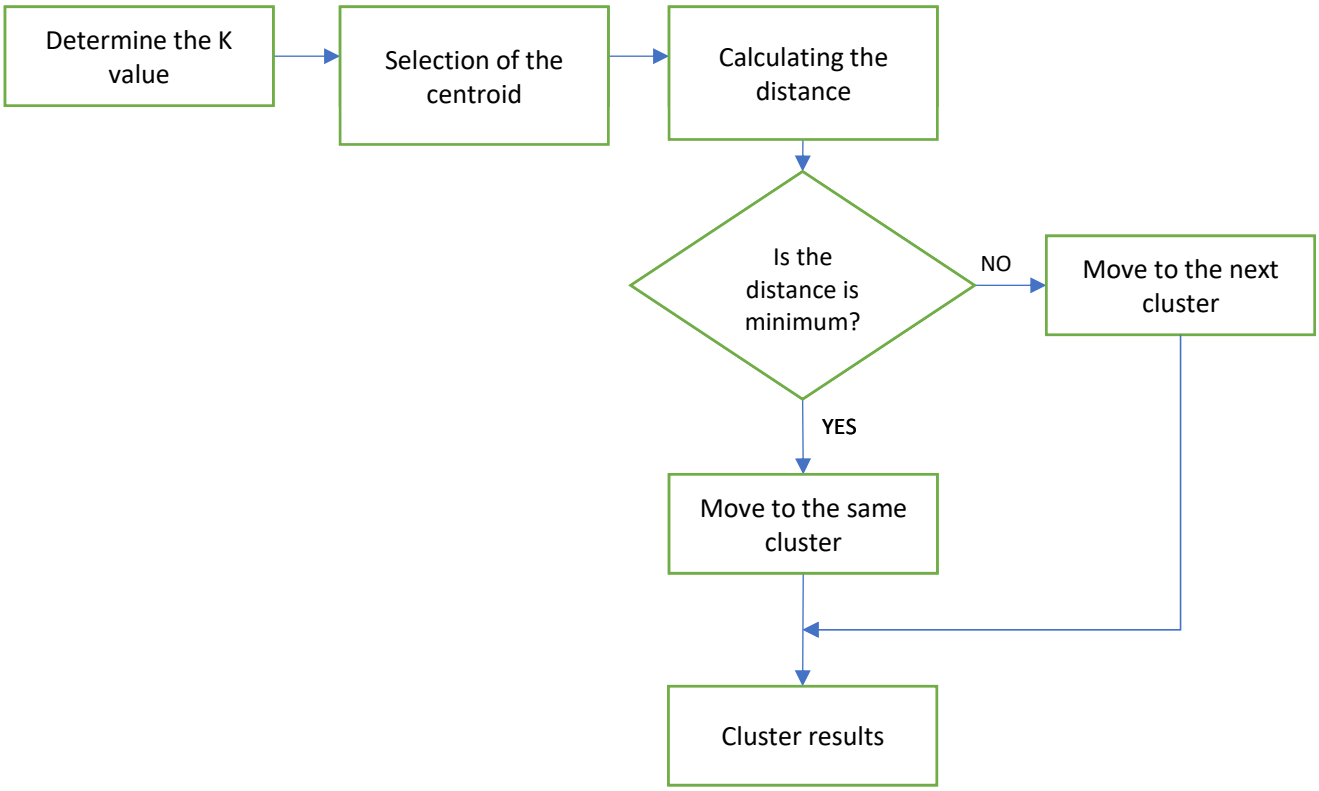


Fig 1.7 Monetary after removing outliers.

4. K-MEANS ALGORITHM AND CLUSTERING

K-means clustering is a centroid-based algorithm, where we calculate the distances to assign a point to cluster. In K-means, each cluster is associated with a centroid. It is an iterative algorithm where we divide un-labelled dataset into K different clusters.



In such a way, that each dataset belongs to one group which has similar properties. optimal value of k is determined by elbow method. Based on the elbow method the optimal value of K is 4. Therefore, we have formed 4 clusters. Below is the flow chart for K-means clustering algorithm based on e-commerce dataset:

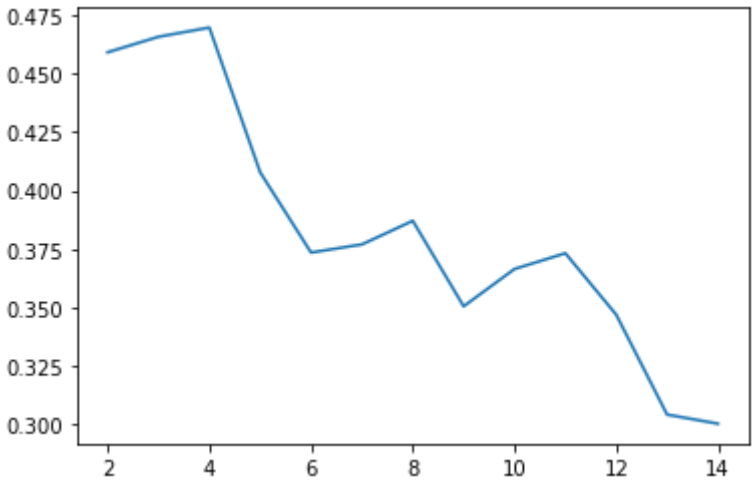


Fig 1.8 Elbow Method to find the optimal K value

- Feature Scaling

Scaling the RFM features using the standard scaling technique to converge to global centroids faster using the K means clustering algorithm.

```
# standardise all parameters
from sklearn.preprocessing import StandardScaler

standard_scaler = StandardScaler()
norm_new = standard_scaler.fit_transform(new_df)
norm_new_df = pd.DataFrame(norm_new)
norm_new_df.columns = new_df.columns
norm_new_df
```

	Monetary	Frequency	Recency
0	1.103611	-0.390975	-0.245052
1	1.055663	0.161481	-0.796234
2	-0.663580	-0.575127	2.067942
3	1.959880	0.319325	-0.628911
4	-0.960036	-0.746126	1.024634
...	...	...	...
3909	-0.849378	-0.667203	1.752981
3910	-0.969918	-0.706665	0.798256
3911	-0.852459	-0.640896	-0.904502
3912	1.403559	8.685090	-0.943872
3913	1.151981	0.122020	-0.560013

3914 rows × 3 columns

5. RESULTS

After applying K-means clustering algorithm to e-commerce data set. Data is segmented into 4 cluster, and they are cluster0, cluster 1, cluster 2 and cluster 3. Based on the Rag status of the graph:

- Cluster 3- Frequency is the one of the key metrics which needs to be focused. The first-time user who has placed order from the company has never come back to place the order again. To increase the frequency, we need to reach out to customers in various ways and also include various product categories to maximize frequency.
- Customers who are in cluster 2 are revenue generators and they add value to the company.
- Customers in cluster 1 are Revenue generators and made purchases from the company recently.
- Customers in cluster 0 are made purchases from the company (~350 customers).

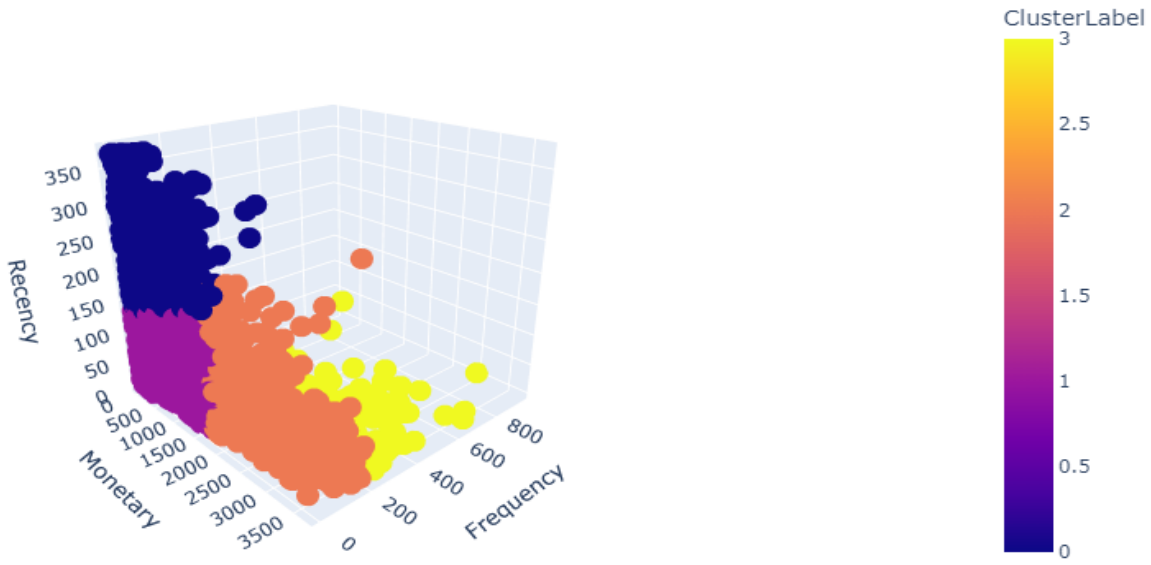


Fig 1.8 Resultant cluster

6. CONCLUSION

Customer segmentation improves customer experience and boosts company revenue. That's why segmentation is a must if you want to surpass your competitors and get more customers. Based on above results, we can conclude that:

- Marketing team can further target the potential customers.
- Based on the past purchases by customers, companies can re-stock the products.
- Attract the loyal customers who are in cluster 0 with best offers.

7. REFERENCES

- Kaggle data source: https://www.kaggle.com/code/fabiendaniel/customer-segmentation
- Related videos: https://www.youtube.com/watch?v=1XqG0kaJVHY&t=5s
- K-means explanation: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
- K-means related pages: https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932