Capstone Project 4

Online Retail Customer Segmentation

Team project by

Sibani choudhury

Sangamesh chandankera

Content

- Problem statement
- Data Summary
- What is customer segmentation?
- Customer segmentation with RFM analysis
- Exploratory Data Analysis
- K means clustering
- Silhouette score
- Principal component analysis
- Challenges
- Conclusion

Problem statement

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Summary

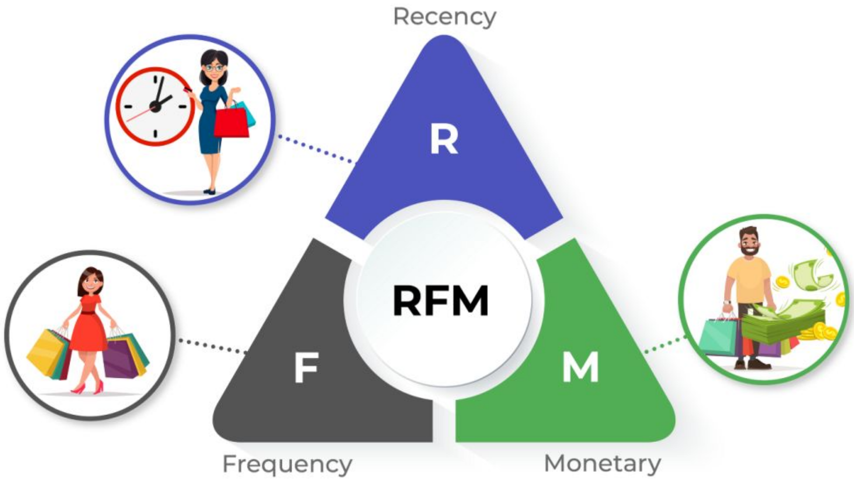
- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invice 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.

Customer Segmentation

Customer segmentation is the process of dividing your customer base into groups based on common characteristics. The customer segmentation process helps you understand who your target audience is, refine your customer experience and reduce churn.

Customer Segmentation with RFM Analysis

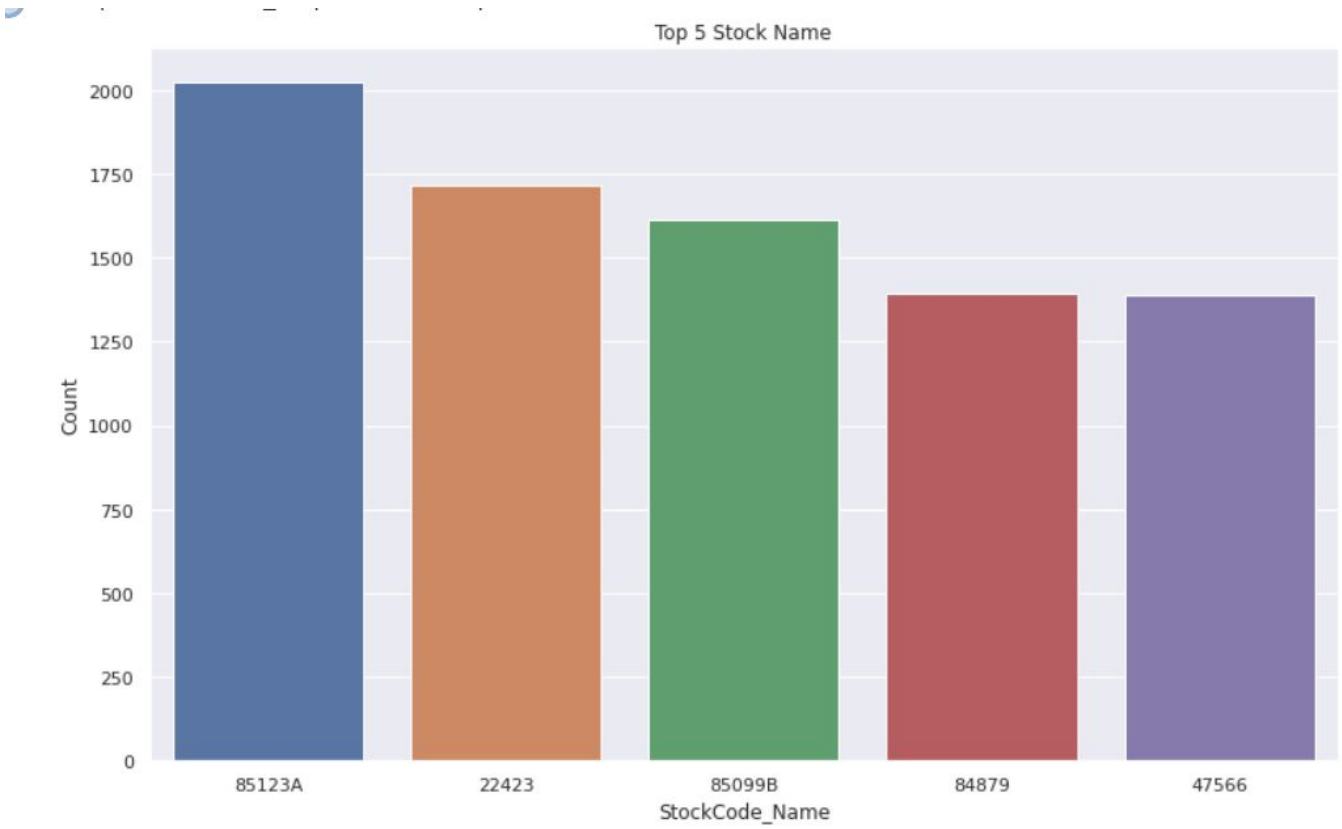
- RFM represents a method used for measuring customer value. An RFM analysis can show you who are the most valuable customers for your business. The ones who buy most frequently, most often, and spend the most. First of all, the metrics you have seen are calculated.



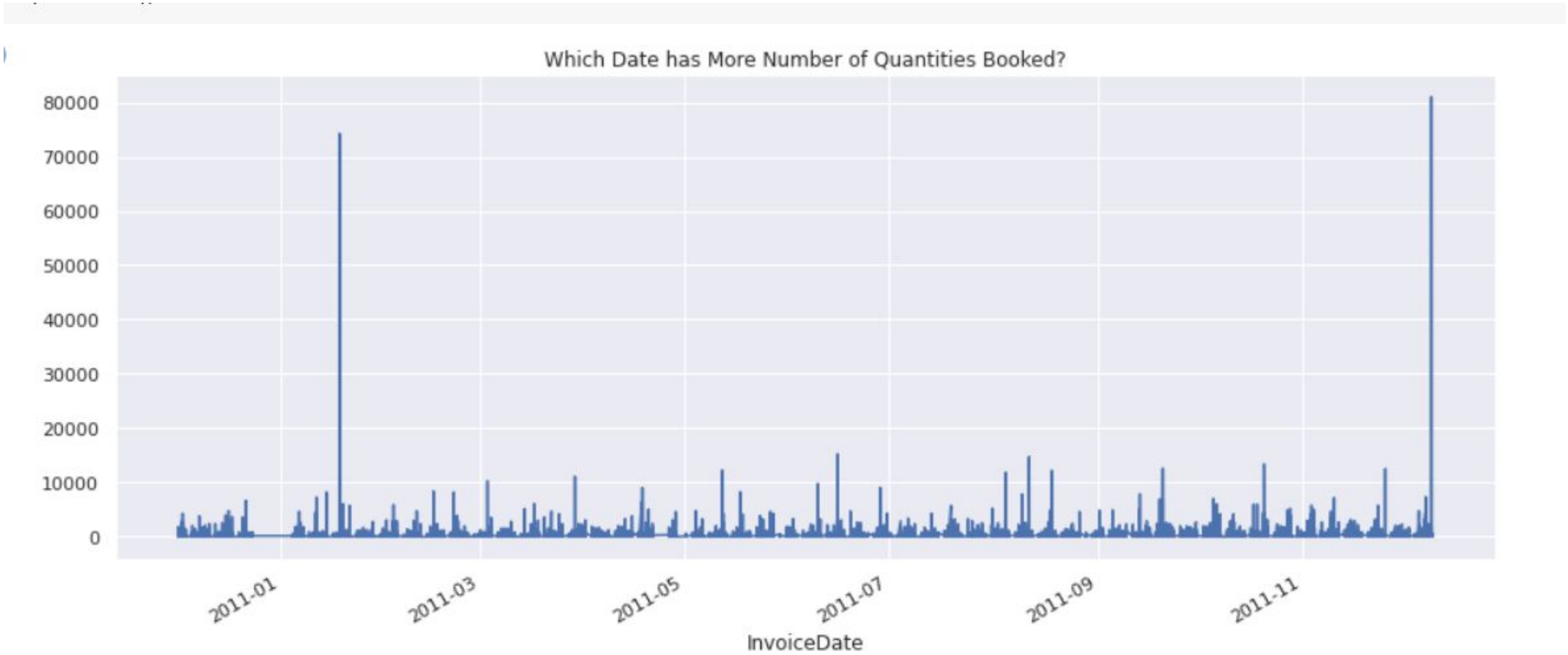
RFM

- Recency : How much time has elapsed since a customer's last activity or transaction with the brand?
- Frequency : How often has a customer transacted or interacted with the brand during a particular period of time?
- Monetary : How much a customer has spent with the brand during a particular period of time?

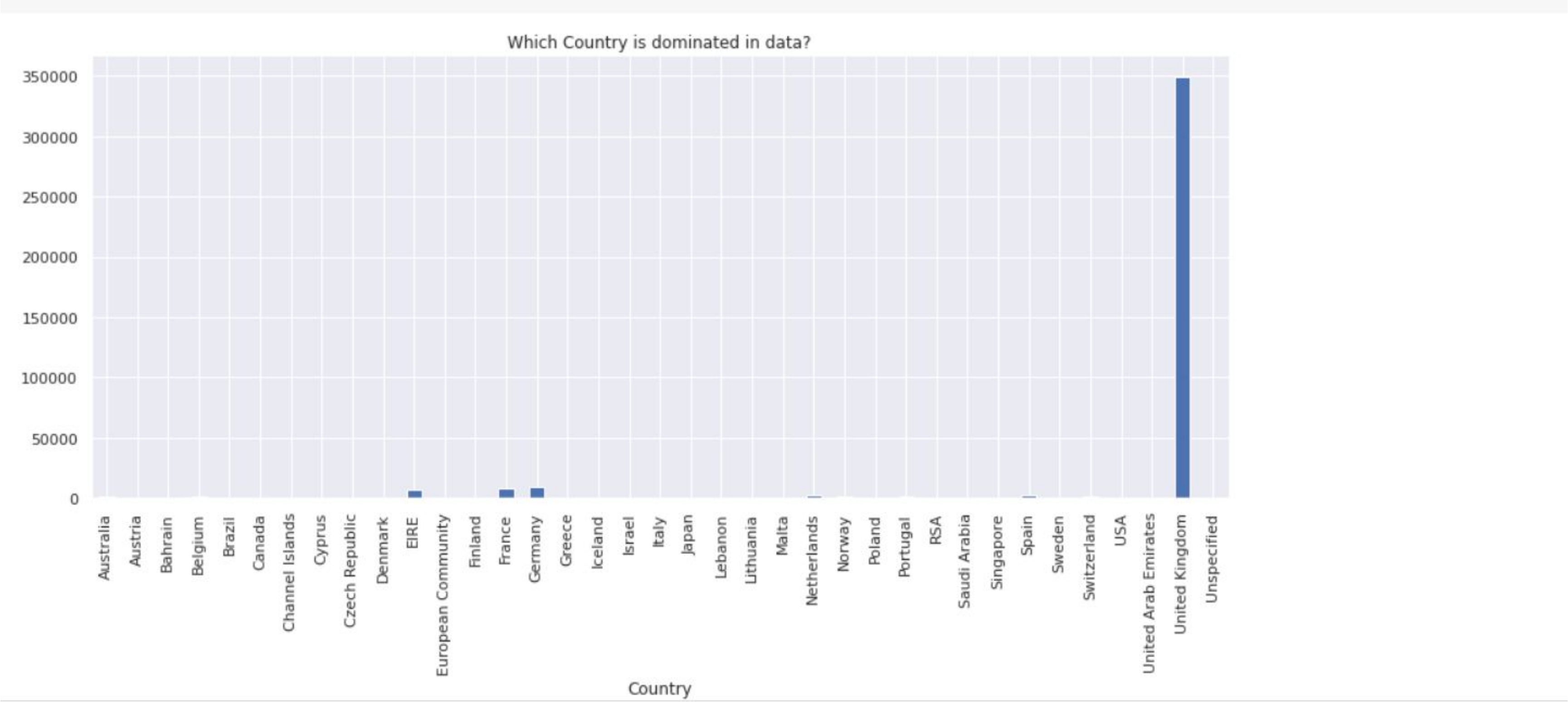
Exploratory Data Analysis



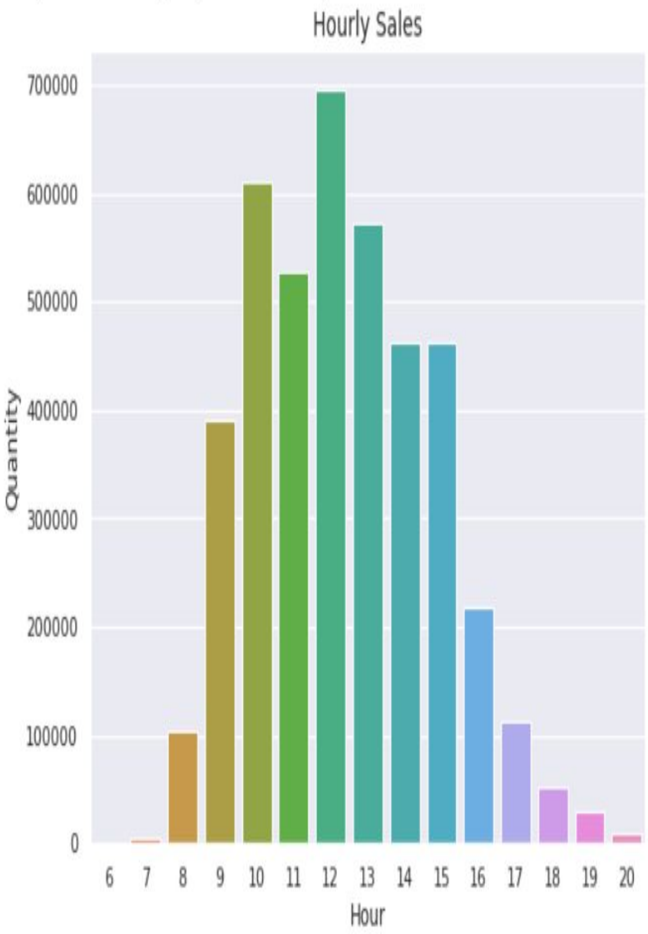
EDA

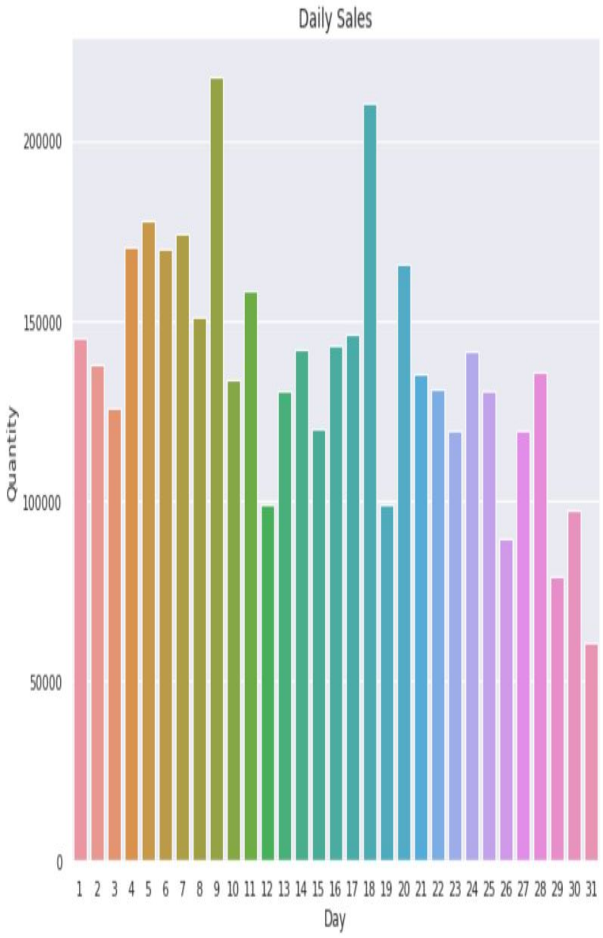

Which Date has More Number of Quantities Booked?
80000
70000
60000
50000
40000
30000
20000
10000
0
2011-01
2011-03
2011-05
2011-07
2011-09
2011-11
InvoiceDate

EDA

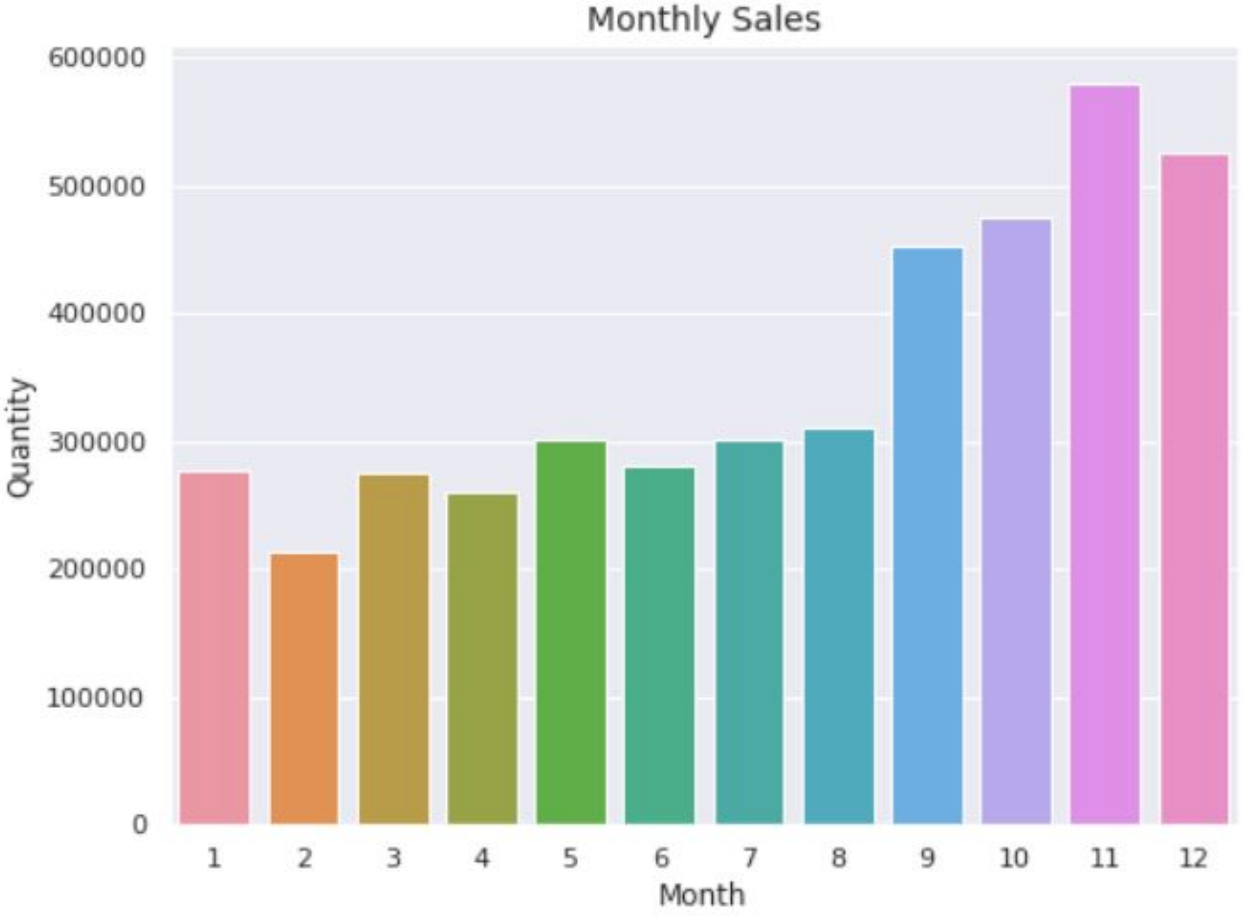

Which Country is dominated in data?
350000
300000
250000
200000
150000
100000
50000
0
Australia
Austria
Bahrain
Belgium
Brazil
Canada
Channel Islands
Cyprus
Czech Republic
Denmark
EIRE
European Community
Finland
France
Germany
Greece
Iceland
Israel
Italy
Japan
Lebanon
Lithuania
Malta
Netherlands
Norway
Poland
Portugal
RSA
Saudi Arabia
Singapore
Spain
Sweden
Switzerland
USA
United Arab Emirates
United Kingdom
Unspecified
Country

EDA

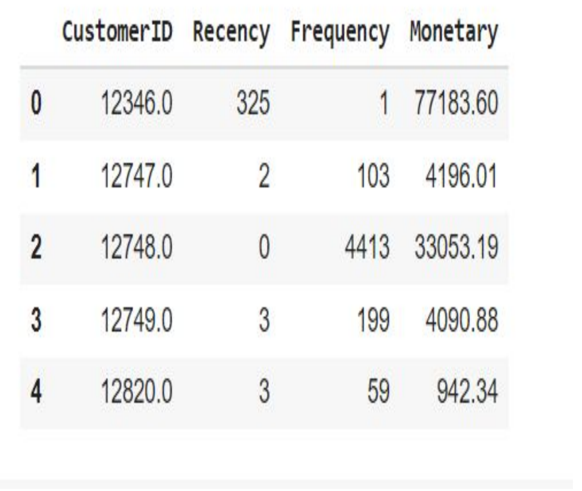

Hourly Sales
Quantity
700000
600000
500000
400000
300000
200000
100000
0
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Hour


Daily Sales
Quantity
200000
150000
100000
50000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day

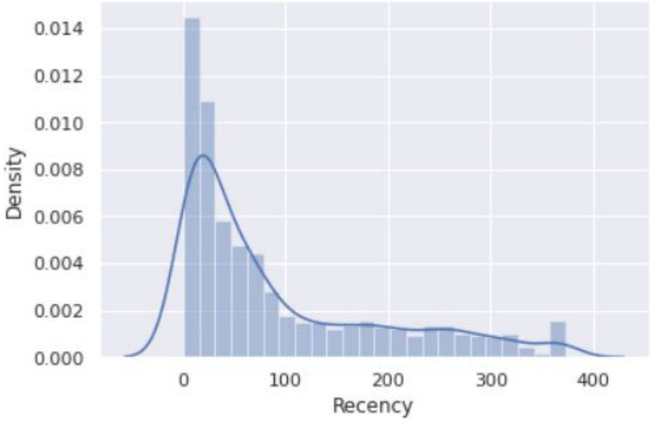
EDA

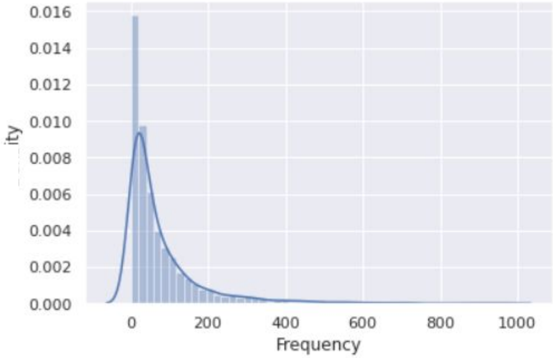

Monthly Sales
600000
500000
400000
300000
200000
100000
0
Quantity
1
2
3
4
5
6
7
8
9
10
11
12
Month

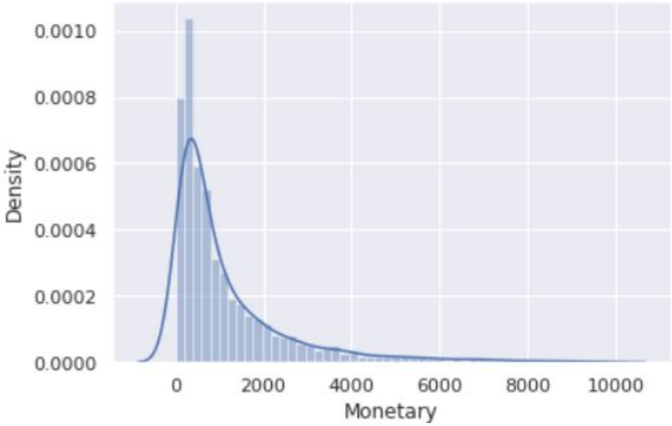
RFM



	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4413	33053.19
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

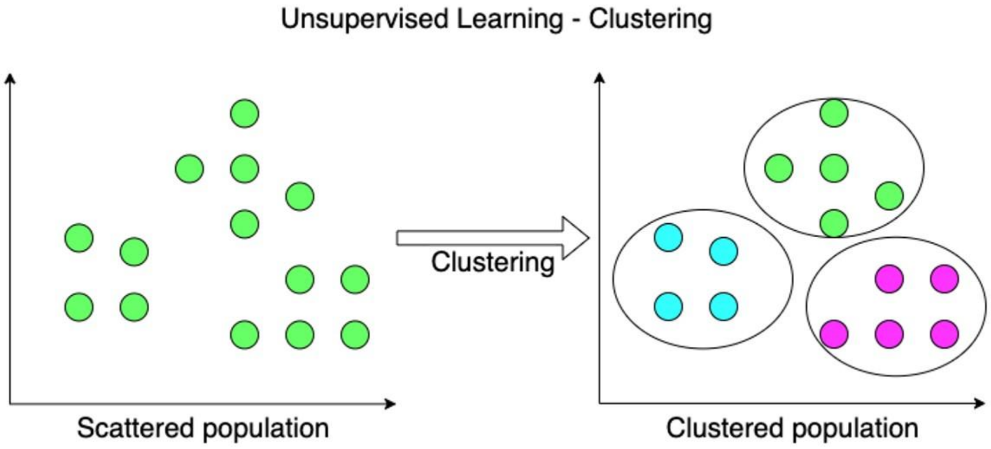






K-Means Clustering

- K-Means clustering is an unsupervised machine learning algorithm that divides the given data into the given number of clusters. Here, the “K” is the given number of predefined clusters, that need to be created.
- It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid.
- The algorithm takes raw unlabelled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.

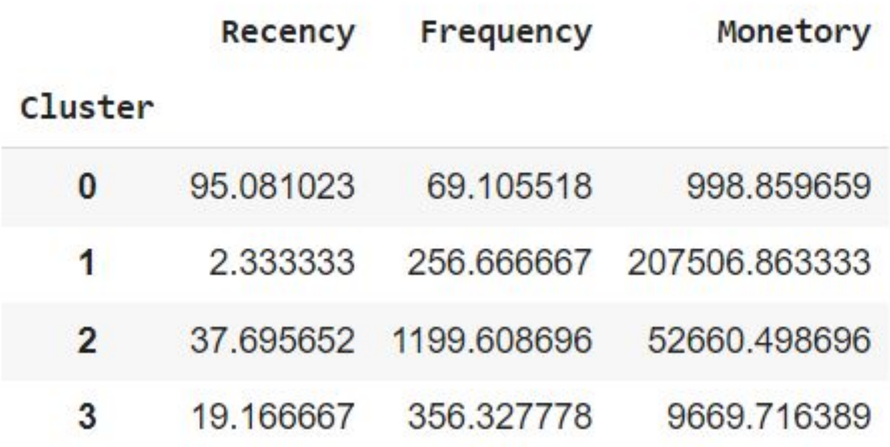


Silhouette score

- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.
- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.
- The Silhouette Coefficient for a sample is $S=(b-a)/max(a,b)$.

Silhouette score

- Cluster 0 contains group of customers with low value of Recency, Frequency and Monetory
- Cluster 1 contains group of customers with high Monetory value
- Cluster 2 contains group of customers with high Frequnecy and monetary value
- Cluster 3 contains group of customers with moderate value of Recency, Frequnecy and Monetory

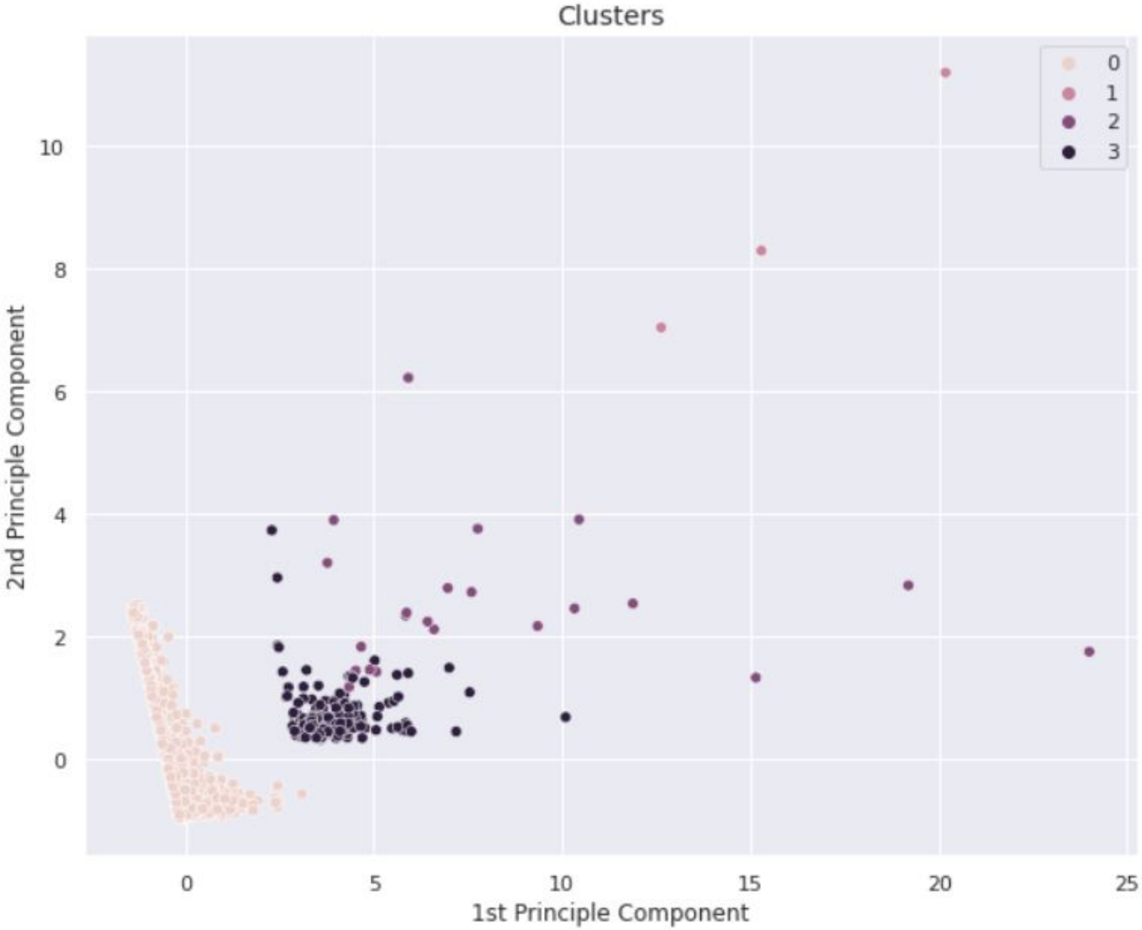


	Recency	Frequency	Monetory
Cluster			
0	95.081023	69.105518	998.859659
1	2.333333	256.666667	207506.863333
2	37.695652	1199.608696	52660.498696
3	19.166667	356.327778	9669.716389

Principal component analysis

- Principal component analysis, or PCA, is a dimensionality reduction method that is used to diminish the dimensionality of large datasets by converting a huge quantity of variables into a small one and keeping most of the information preserved. PCA is such a technique that groups the different variables in a way that we can drop the least important feature

PCA


Clusters
2nd Principle Component
1st Principle Component
10
8
6
4
2
0
0
5
10
15
20
25
0
1
2
3

Challenges

- Whole data Consists of duplicated data initally there was more than 5lakh data after grouped Based Customer ID it was left with 3.5 thousand Customers left.
- Groupping them based on certain assumptions and there was many negative values.

Conclusion

- In this case, we have compared RFM Analysis with Kmeans clustering. How much best cluster making in modeling with Kmeans. Firsty,this data set is better with scaling and centering data, robust scaler should be using in this dataset because so many outliers in our data. After that, we have find out best n_cluster for this data to build how much segmentation should be given and show best silhoutte score for each n_cluster. And in the final we have done PCA(Principal component analysis) to find best components.

AI

Thank you