Capstone Project

Customer Segmentation

By

Govind Wakure

Outline

- Introduction
- Problem Statement
- What is Customer Segmentation?
- Data Description
- Data Exploration
- RFM Segmentation
- Feature Extraction
- K-Means Clustering
- Conclusion

Problem Statements

- 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.

What is Customer Segmentation?

- Customer segmentation is the process of separating customers into groups based on their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other.
- The main goal is to identify customers that are most profitable and the ones who churned out to prevent further loss of customer by redefining company policies.
- Having large number of customers, each with different needs it is crucial to find which customer are most important for business and target them with appropriate strategy.

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.

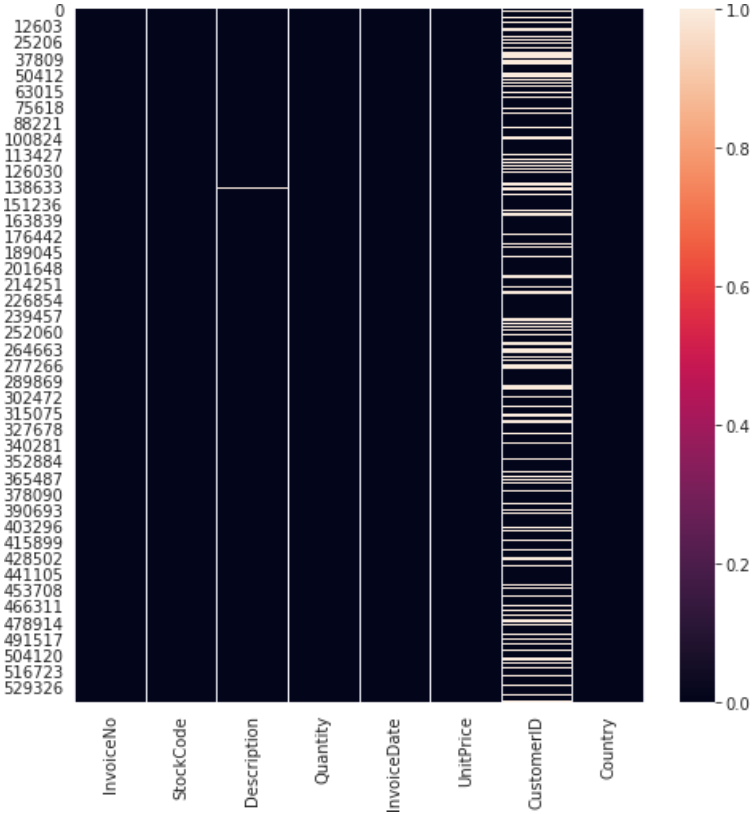
InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

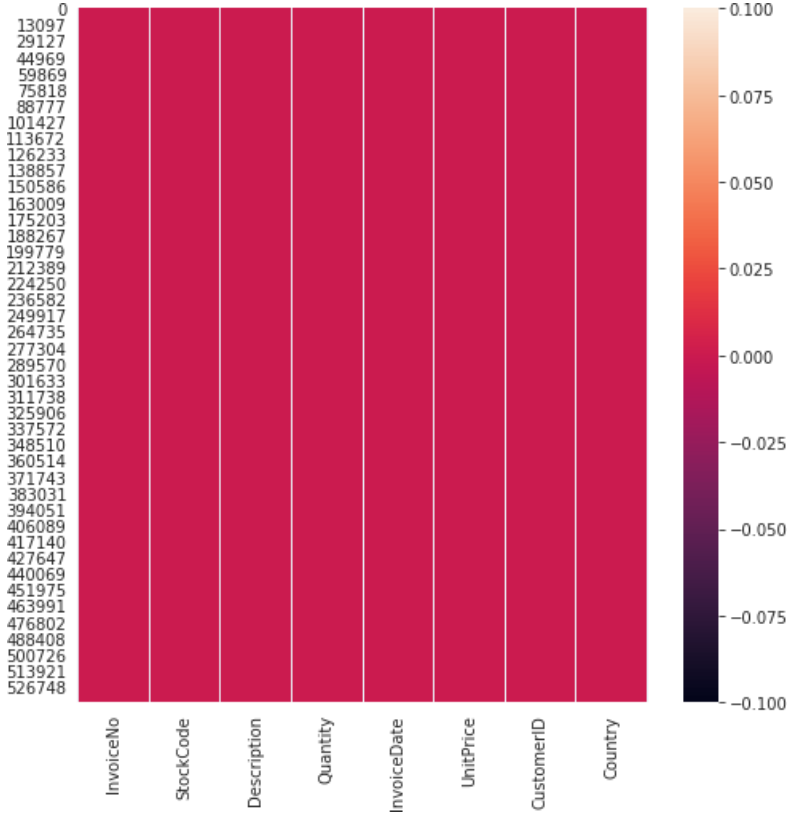
UnitPrice: Unit price. Numeric, Product price per unit in sterling.

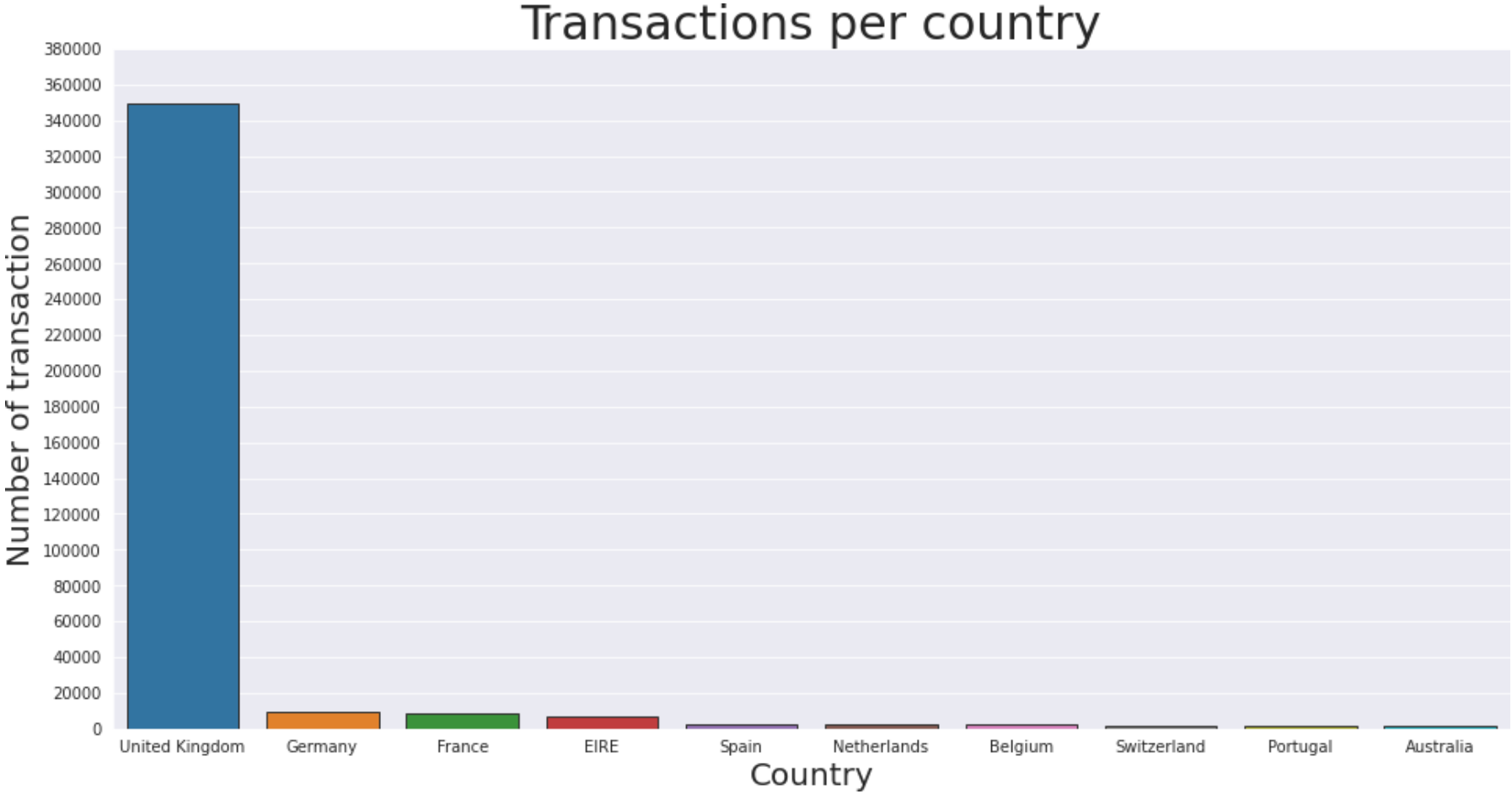
CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

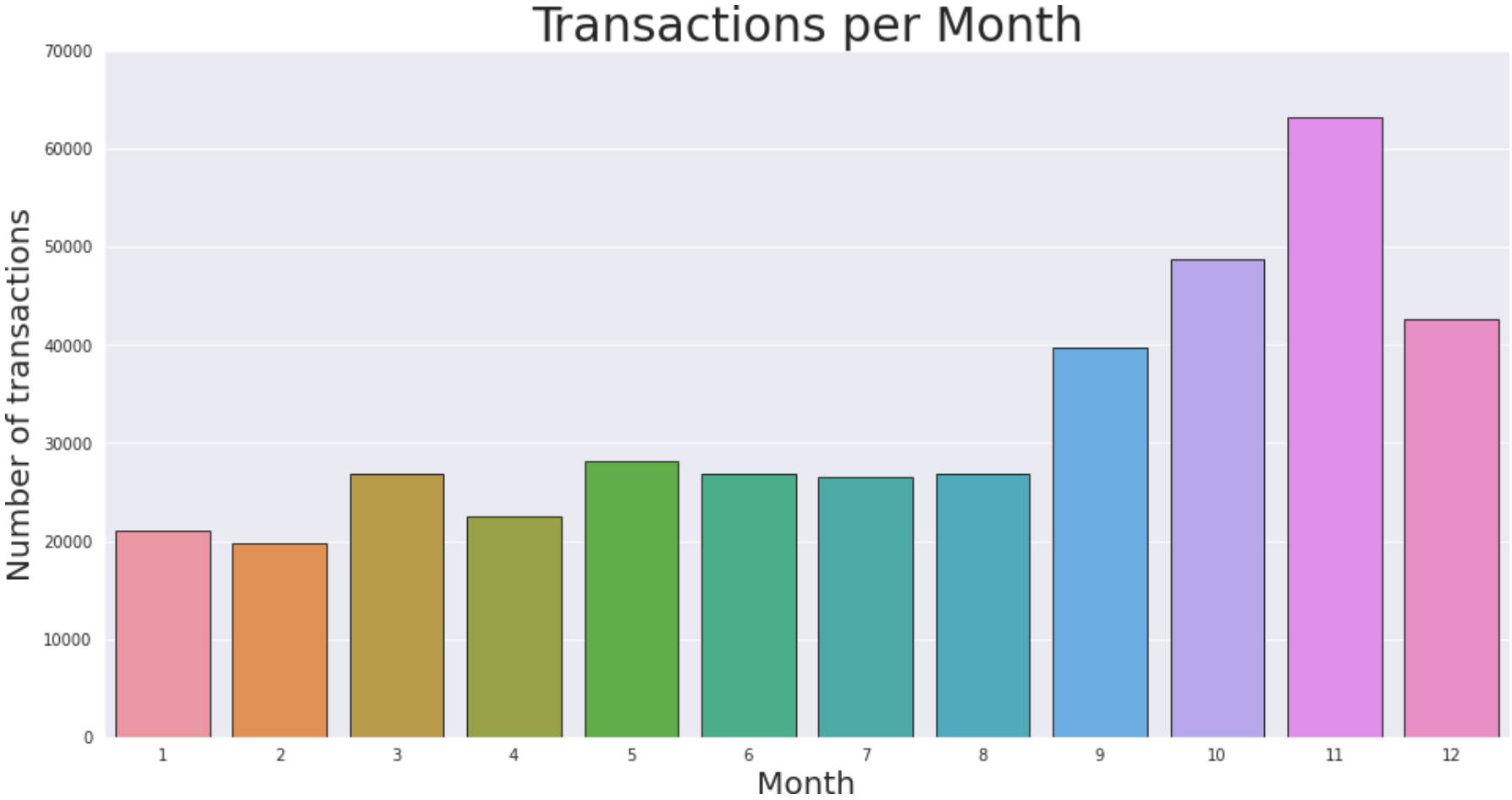
Country: Country name. Nominal, the name of the country where each customer resides.

Data Exploration





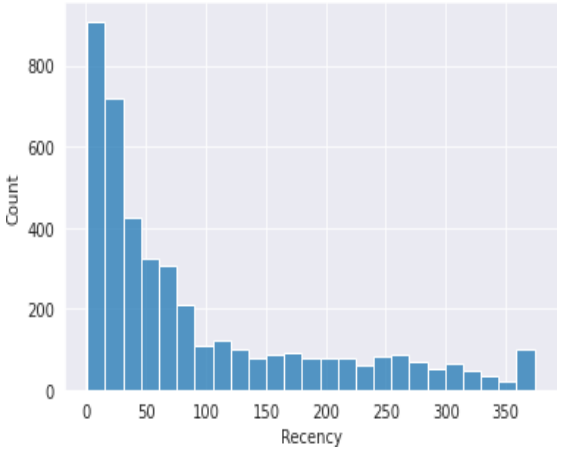

Transactions per country
Number of transaction
380000
360000
340000
320000
300000
280000
260000
240000
220000
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
United Kingdom
Germany
France
EIRE
Spain
Netherlands
Belgium
Switzerland
Portugal
Australia
Country

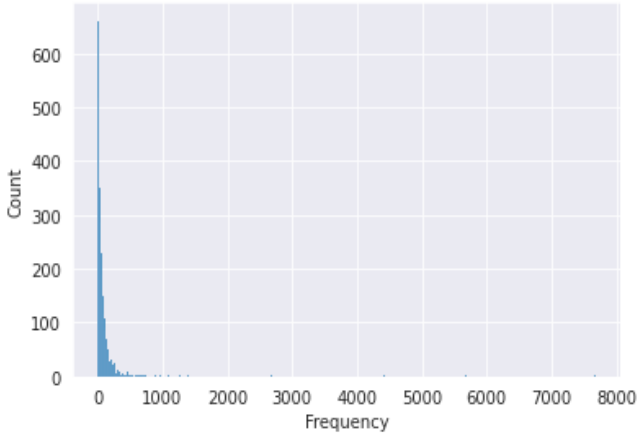

Transactions per Month
Number of transactions
70000
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

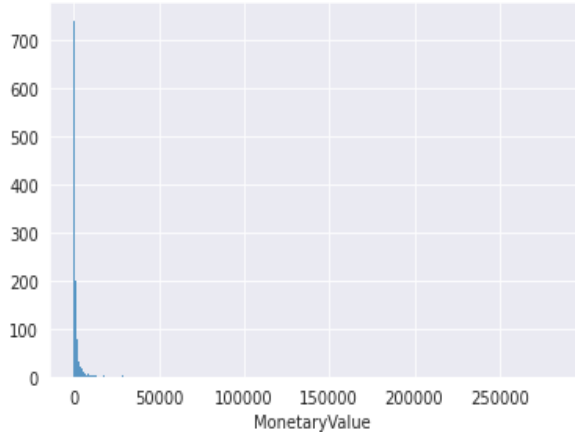
RFM Segmentation

- RFM stands for Recency, Frequency and Monetary
- RFM analysis is commonly used technique to generate and assign a score to each customer based on:
  - How recent their last transaction was (Recency)
  - How many transactions they have made in the last year (Frequency)
  - What monetary value of their transaction was (Monetary)

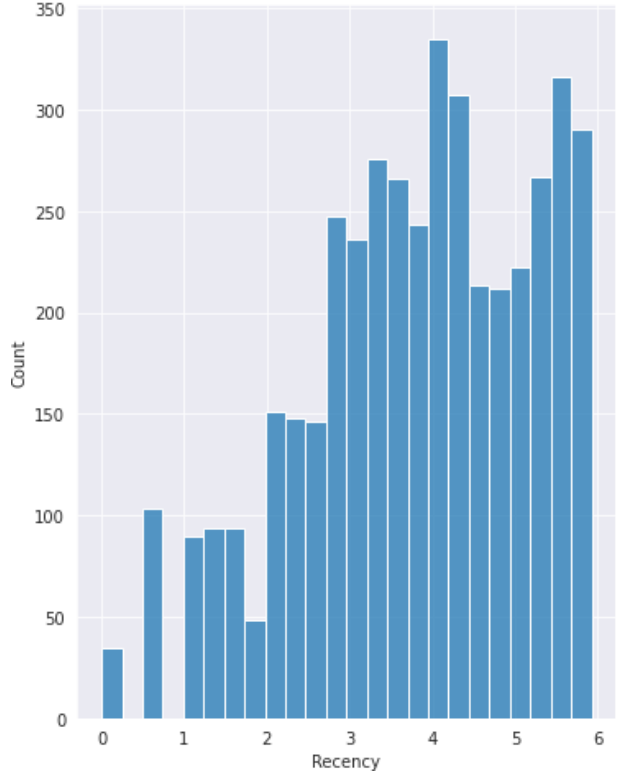
Feature Extraction

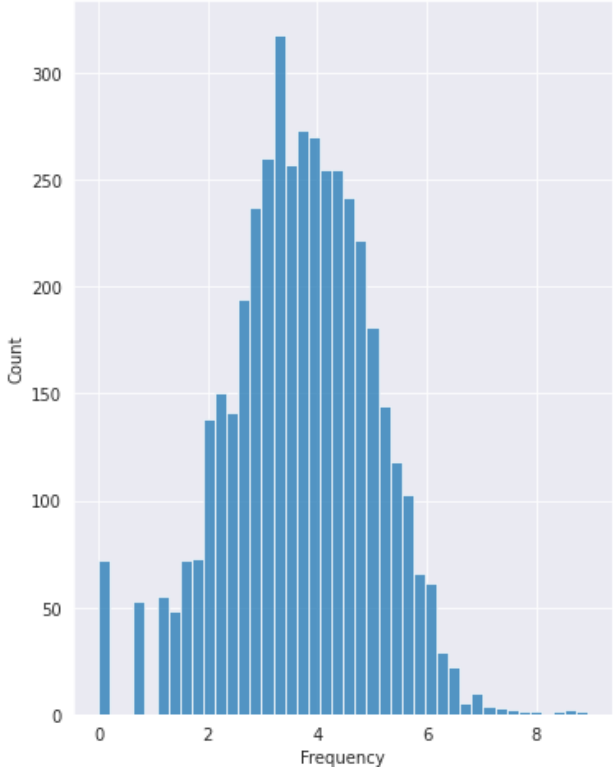


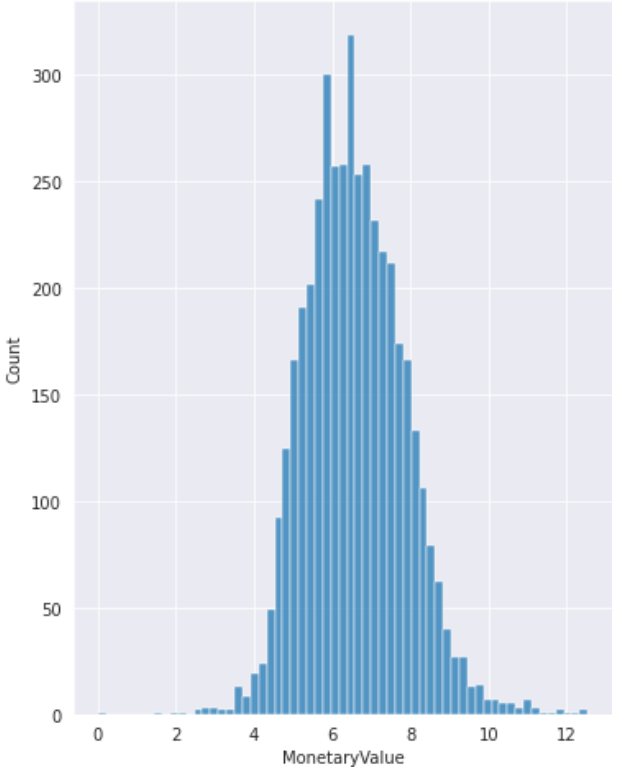




Data distribution after log transform







- Each transaction is assigned values based on Recency, Frequency and Monetary
- Each point in plot represent a Transaction

Data Modeling:

K-means Clustering:

- K Means clustering algorithm is an unsupervised machine learning algorithm that uses multiple iterations to segment the unlabeled data points into K different clusters in a way such that each data point belongs to only a single group that has similar properties

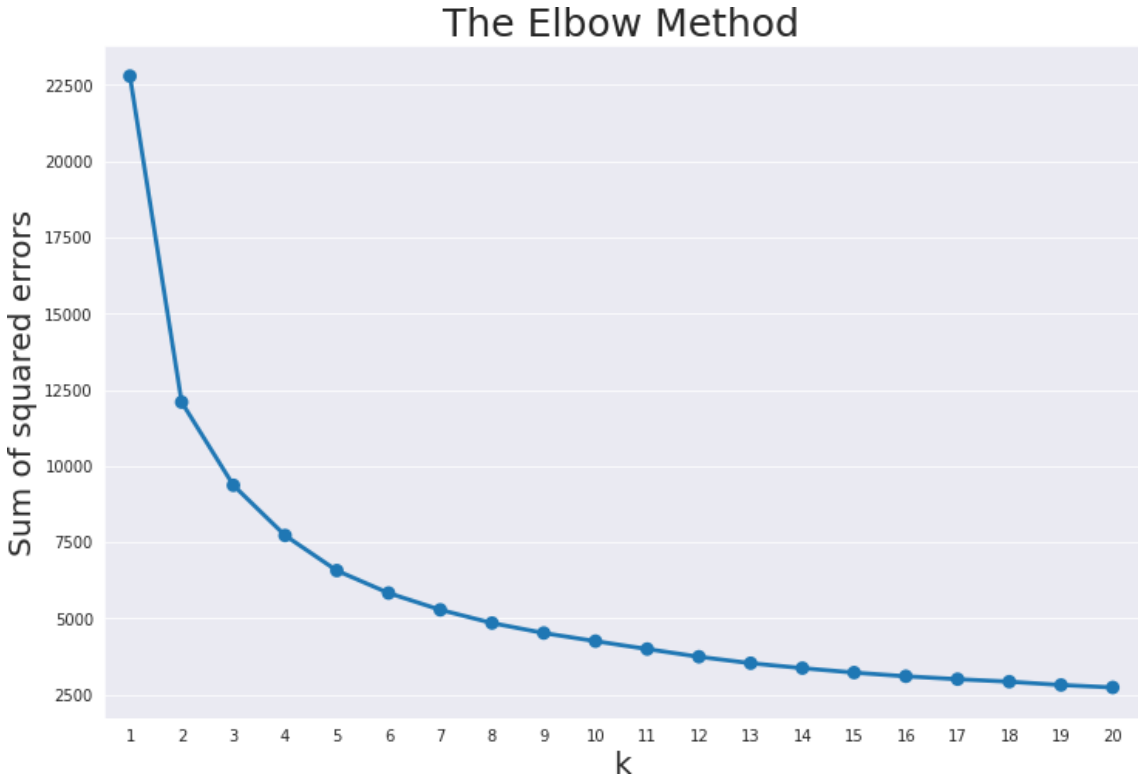
K means gives the best result under the following conditions

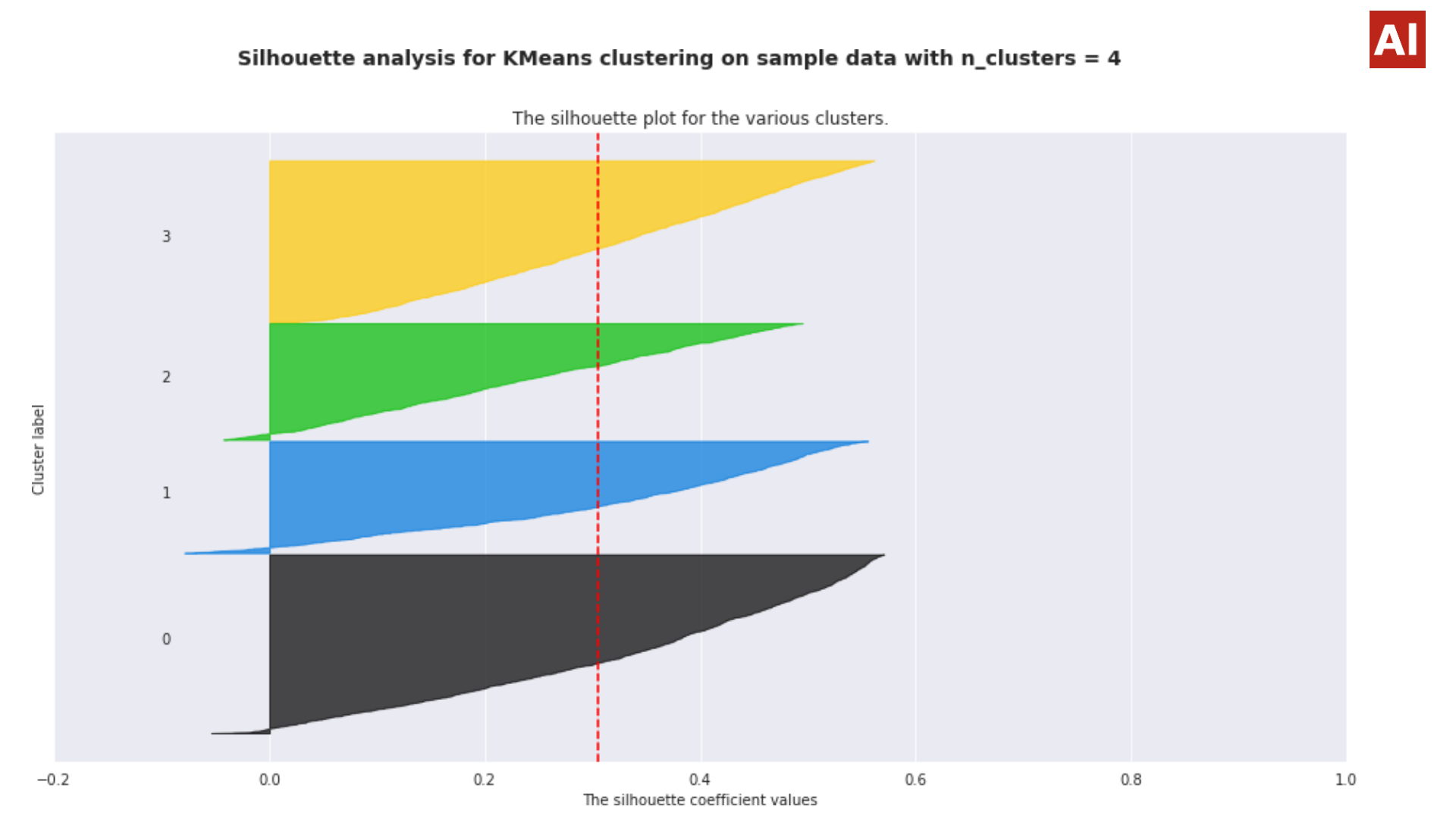
- Data's distribution is not skewed
- Data is standardised
- The data is highly skewed, therefore I will perform log transformations to reduce the skewness of each variable and I standardised the data

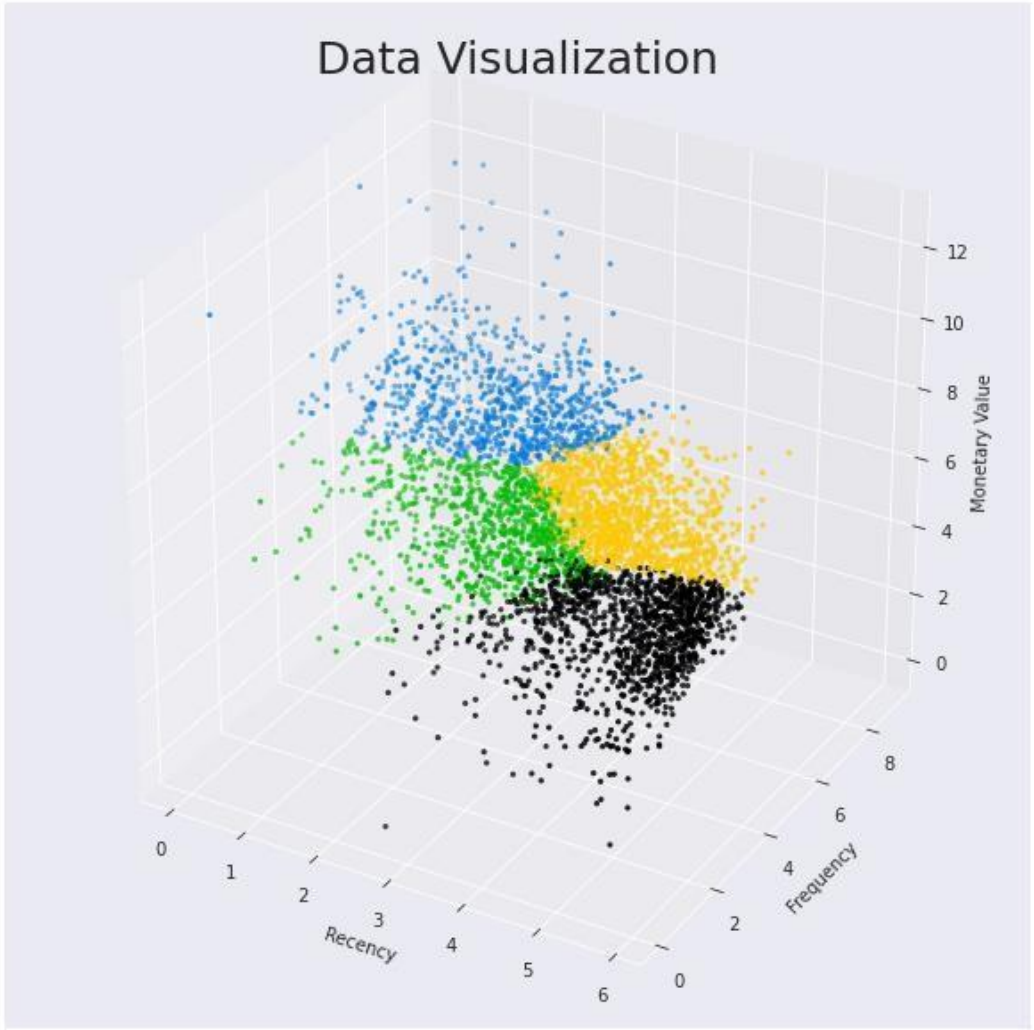
Why k-means?

As our feature variables are numerical and our goal is unsupervised to find out some sort of structure in the customers, I used k-means clustering

K-Means Clustering




Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values



- Each transaction is assigned a cluster based on Recency, Frequency and Monetary
- Optimal number of cluster by silhouette analysis is four
- Each color in plot represent a Cluster

Mean value of each feature

Cluster	Recency	Frequency	Monetary Value
0	150	10	228
1	8	197	3697
2	15	30	486
3	77	65	1123

Conclusion

The customer segments thus deduced can be very useful in targeted marketing, scouting for new customers and ultimately revenue growth. After knowing the types of customers, it depends upon the retailer policy whether to chase the high value customers and offer them better service and discounts or try and encourage low/ medium value customers to shop more freqently or of higher monetary values.

The visualization of clusters in Silhoutte Analysis show some overlapp between the customer segments.However, the dataset does not distinguish between wholesale and retail customers, it is quite likely that high value frequent clients are the wholesale dealers and medium/ low valued ones are individual retail purchasers.

Conclusion

Cluster	RFM Interpretation	Type of Customer
0	Last purchase long ago, Least number of transactions, Least monetary spending	Churned
1	Recent transaction, Most frequent transactions, Highest monetary spending	Best (target)
2	Recent transaction, Low purchase frequency Low monetary spending	New
3	Last purchase while ago, Less frequent transactions Low monetary spending	At Risk

Thank You