Capstone Project

Customer Segmentation

Team Consists of:

- Vridhi Parmar
- Hrithik Chourasia

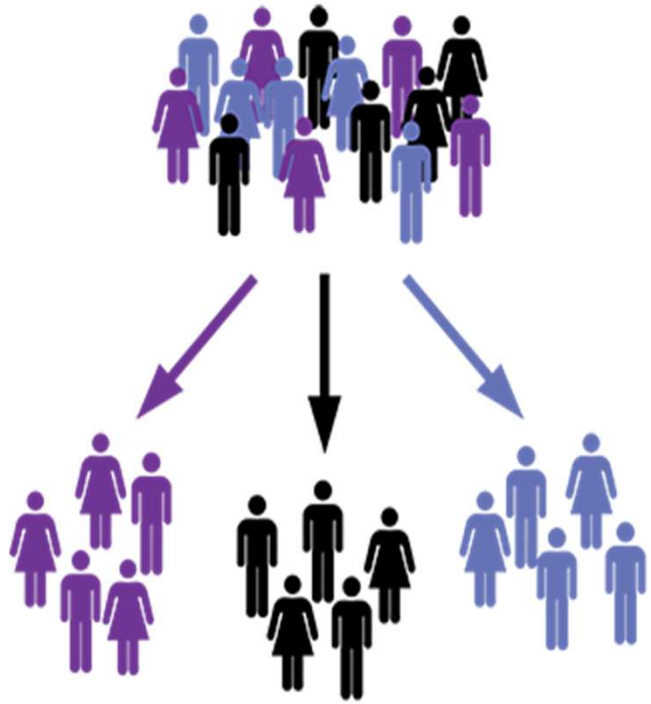
Problem Statement:

- 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.

What is Customer Segmentation?

- Practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

- Allows us to better understand our customers helping us target these customers in a more efficient manner and improve the customer experience.



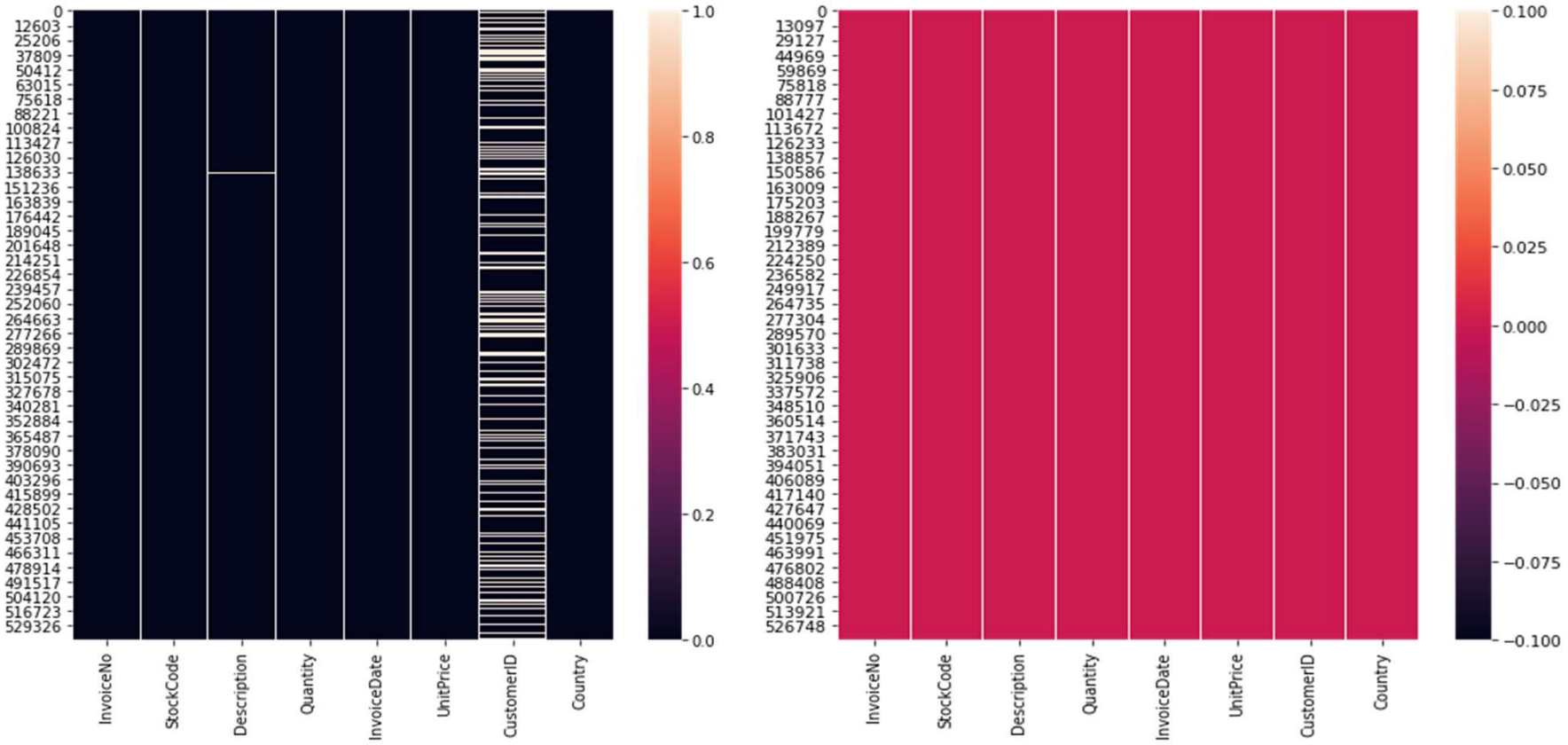
Factors for segmentation for a business to business marketing company:

- Industry
- Number of Employees
- Location
- Market Cap/ Company Size
- Age of Company

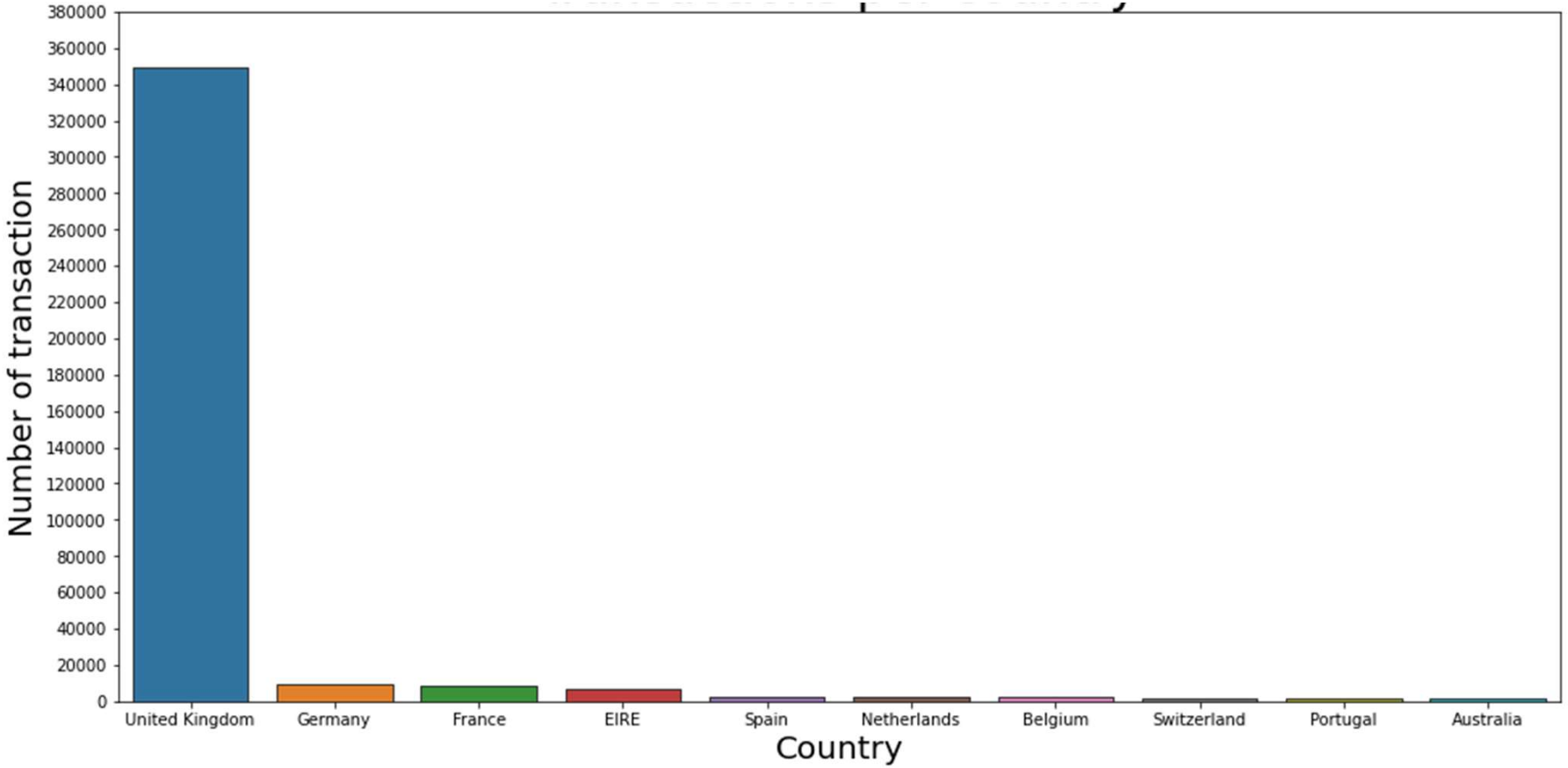
Data description:

- **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:** 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.

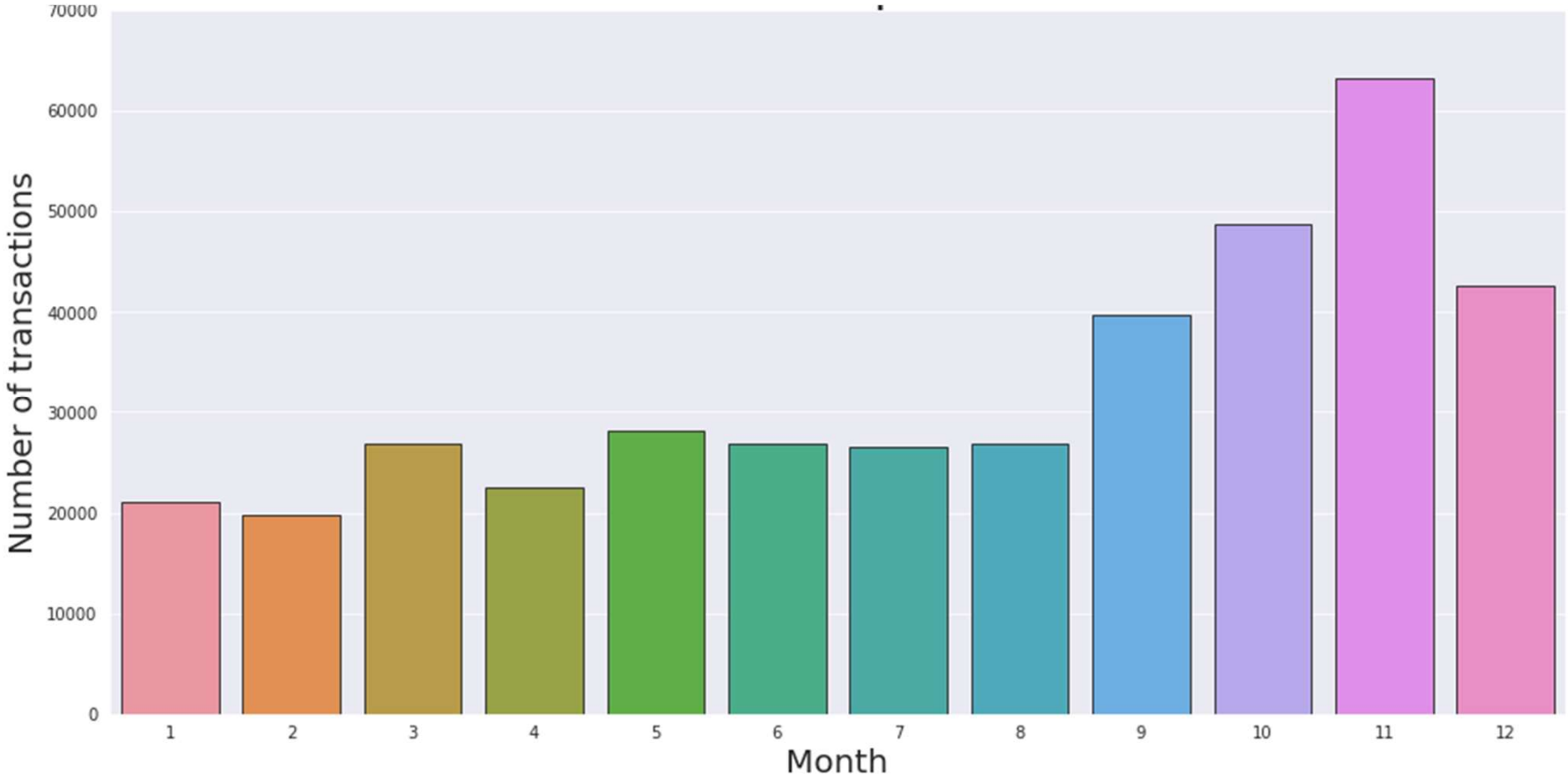
Data Exploration



Transactions per Country



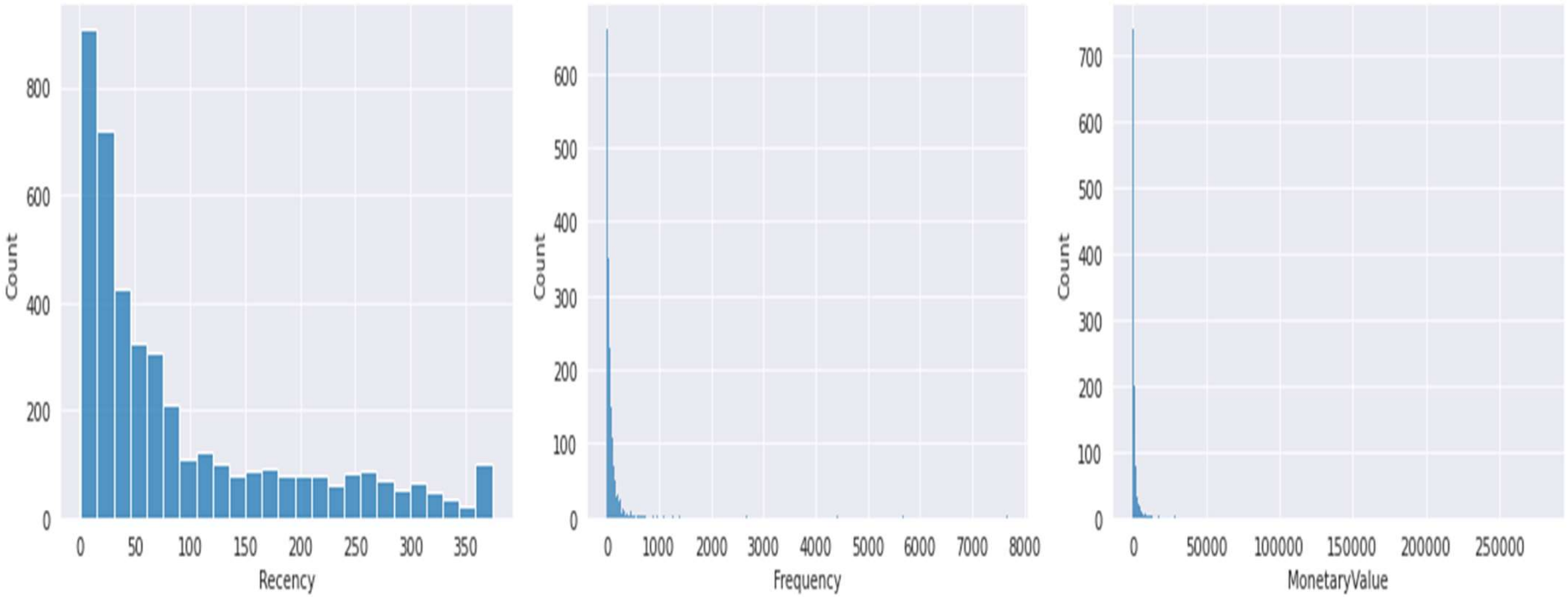
Transactions per month



RFM Segmentation

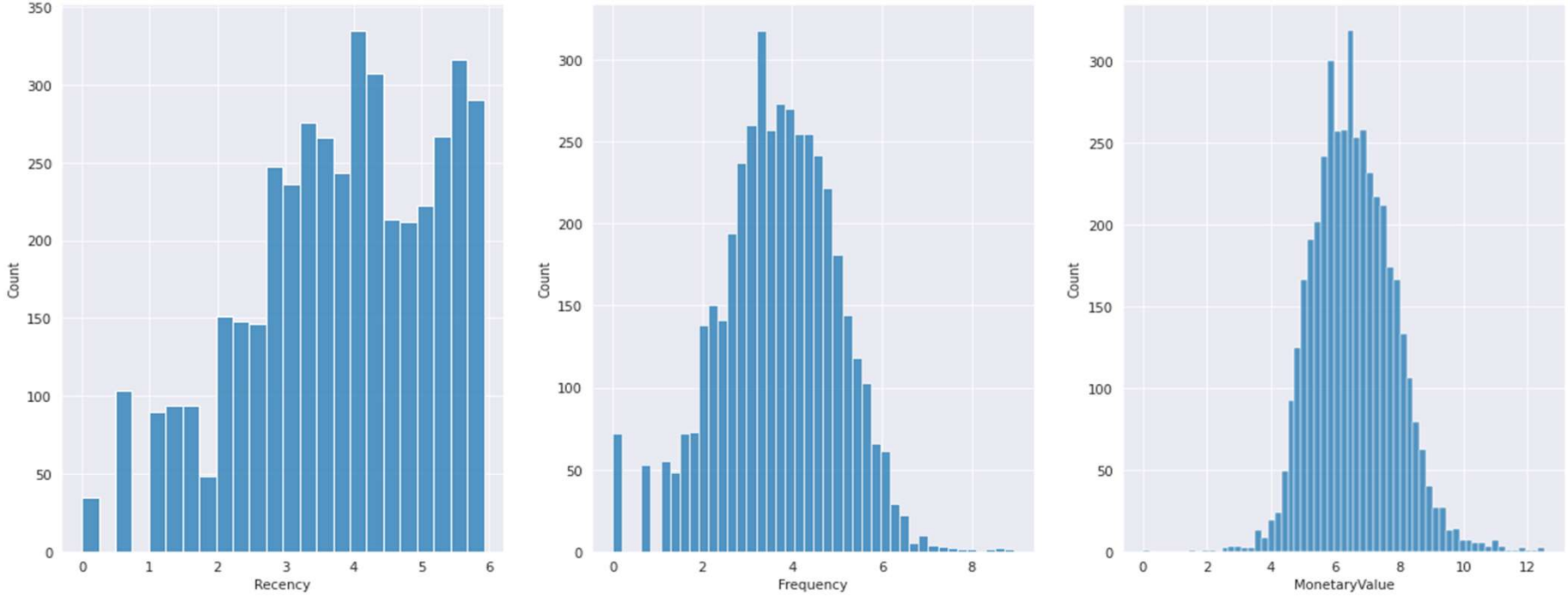
- RFM stands for Recency, Frequency and Monetary.
- RFM analysis is commonly used technique to generate and assign a score to each customer based on:
  - How recent their last transaction was (Recency)
  - How many transactions they have made in the last year (Frequency)
  - What monetary value of their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

Feature Extraction



Our data is right-skewed. We can apply log transformation to remove the skewness.

Data distributions after log transformation

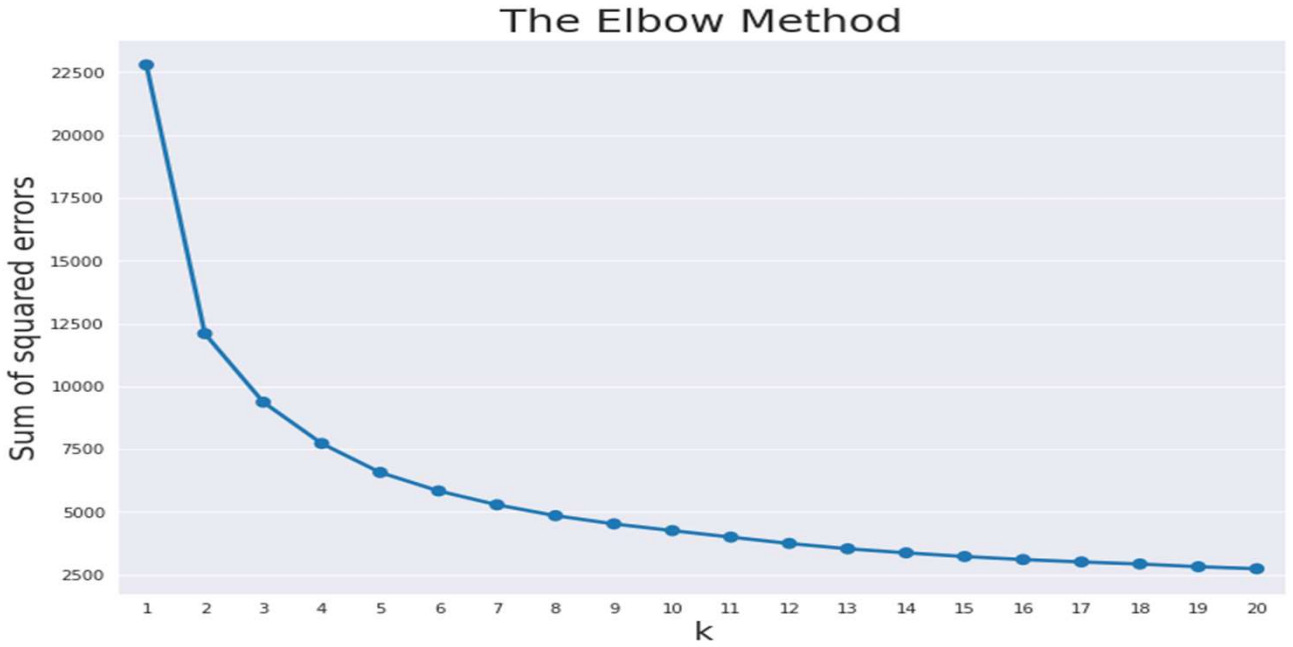


We have removed skewness from data.

Data Visualization

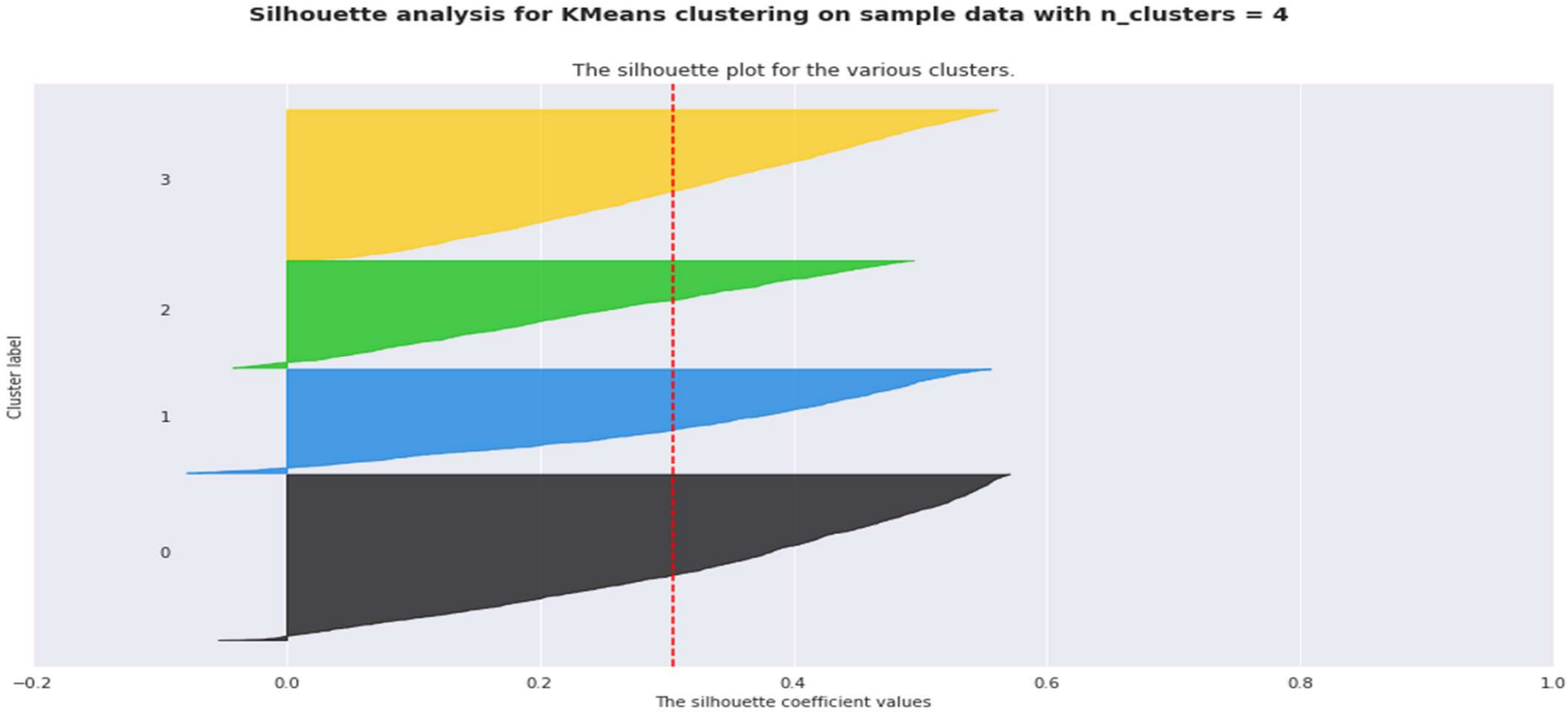
- Each transaction is assigned values based on Recency, Frequency and Monetary.
- Each point in plot represent a Transaction.

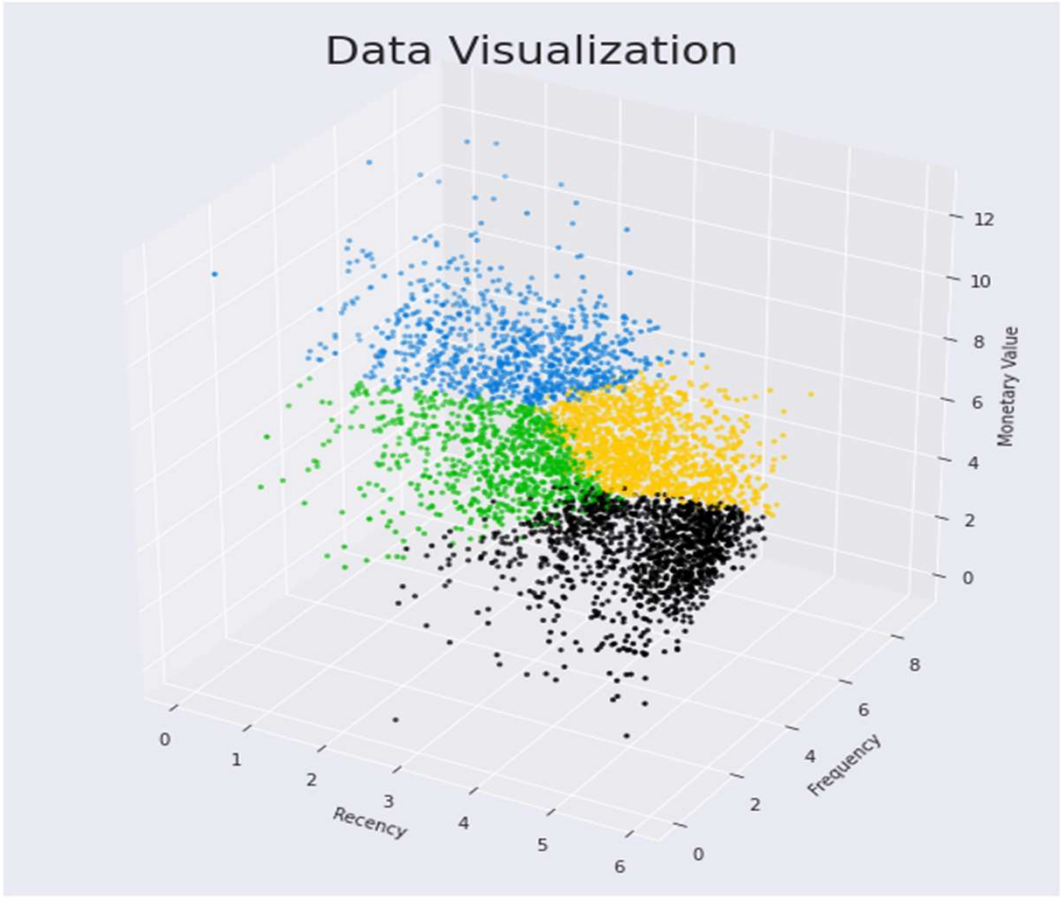
K-Means Clustering



We can see that error is decreasing as K increases. For values of k at 4 or 5 slope of the curve is decreasing very fast. This means errors do not decrease much faster as the increase in number of clusters.

Silhouette analysis for k-means clustering





- 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

CONCLUSTION

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