Online Retail Customer Segmentation

1. Introduction

Customer segmentation is one of the most basic and important steps for winning in the market. RFM analysis seeks to segment customers based on when their last purchase was (Recency), how often they have purchased in the past (Frequency), and how much they spend (Monetary Value). These three measures have proven to be effective predictors of a customer's willingness to engage in marketing messages and offers. Cohort analysis can provide answers about marketing strategy that are essential for small and medium-sized enterprises as well as large corporations.

This project uses the Online Retail dataset from Kaggle, which contains all the transactions made from 2010-12-01 to 2011-12-09 for a UK-based online-only retailer. I conducted a cohort analysis, RFM analysis, and built customer segments by running the K-Means clustering model.

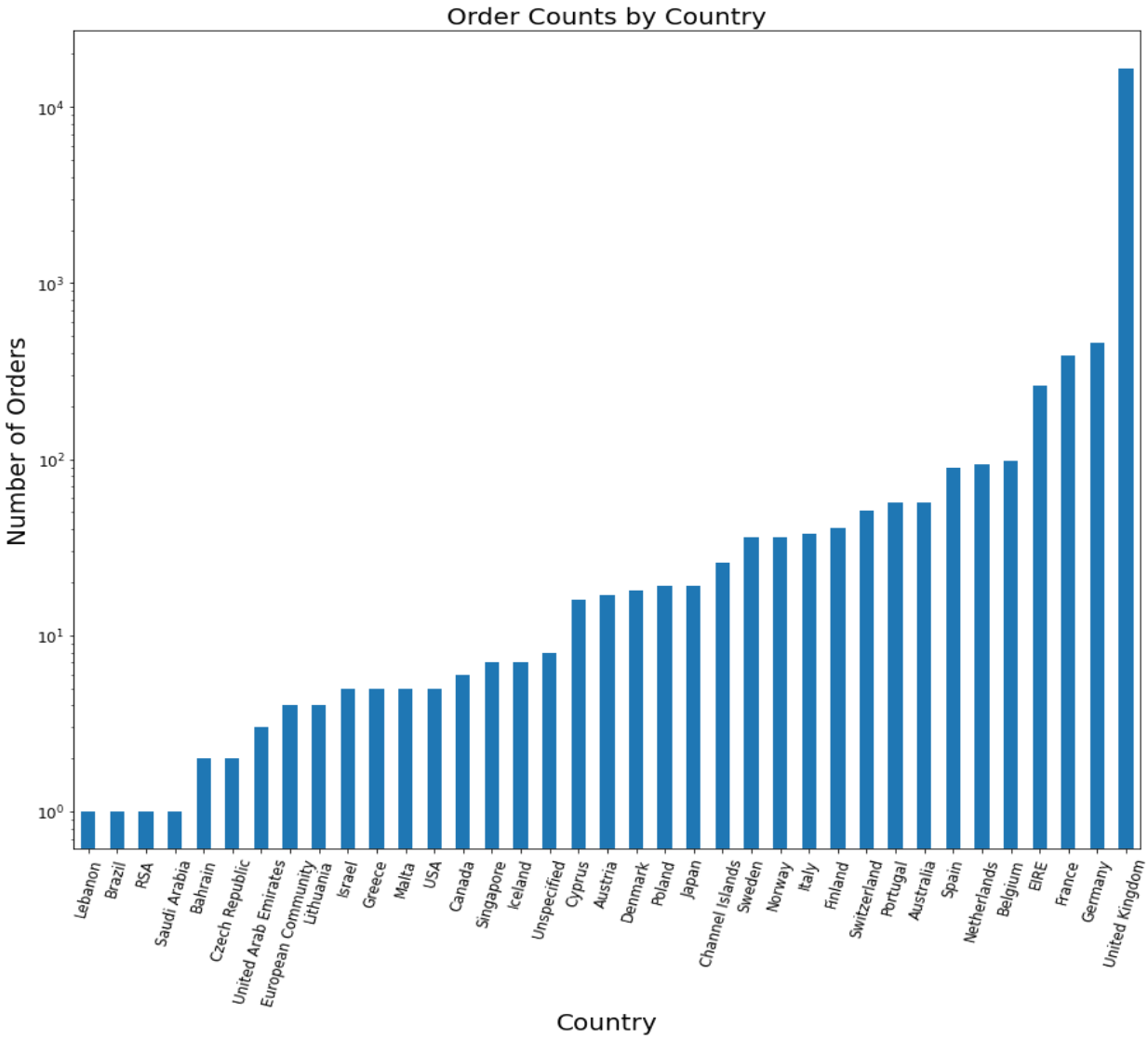
2. Data Wrangling

There are 541,909 rows and 8 columns in the Online Retail dataset. The columns are as follows:

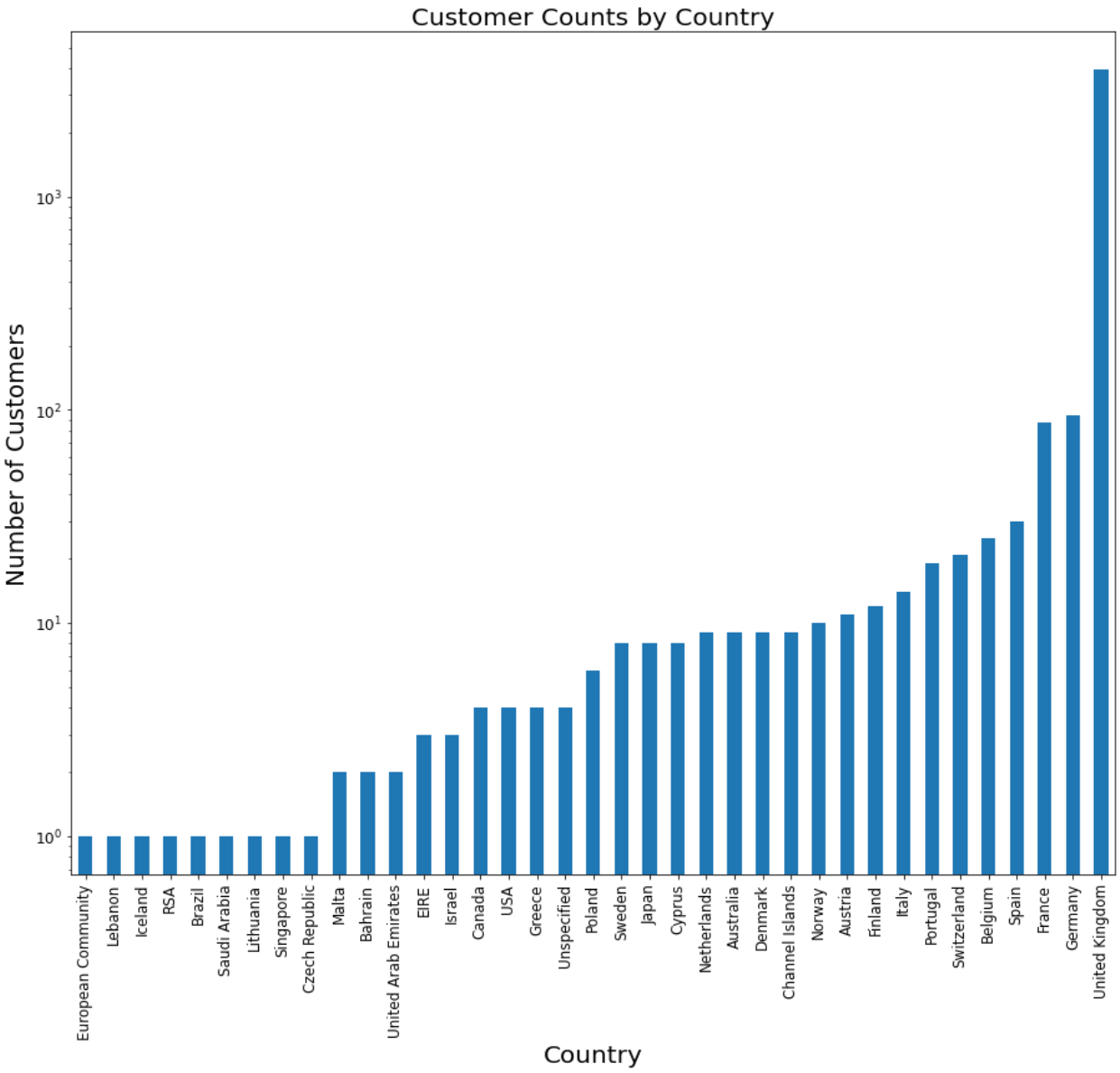
InvoiceNo: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. Invoice numbers starting with the letter 'C' indicate a cancellation.
StockCode: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
InvoiceDate: Invoice date and time. Numeric. The day and time when a transaction occurred.
UnitPrice: Unit price. Numeric.
CustomerID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal. The name of the country where a customer resides.
There is one invoice number per order (associated with all the items purchased in that order). Customer ID serves as identification for each customer (associated with all the orders placed by that customer).
There were 135,080 rows without Customer ID information in the data. Since Customer ID was very essential for this analysis and customer segmentation, I decided to remove all these 135,080 rows.
There were negative quantities in the data. I found that wherever the negative quantities occurred, the InvoiceNo always started with 'C' and vice versa, so it made sense that these were cancelled orders. However, I could not find record of the corresponding order placed before the cancelation happened, so I decided to remove all the rows with negative quantities.
I also identified 40 rows with unit price 0. This could be a manual error. I decided to remove these rows.
There were 1,445 rows with null values in the Description column. However, my analysis and customer segmentation did not use this feature, so I decided to completely remove the Description column when cleaning the data. After data wrangling, 397,884 out of 541,909 rows were retained.

3. Exploratory Data Analysis

The first finding tabulates order counts by country. From the following plot, we can see that there are 16,646 orders from within the UK, which is far more than the number of orders from any other country. The UK is followed by Germany with 457 orders, France with 389 orders, Eire with 260 orders and Belgium with 98 orders.



The second finding shows the customer counts by country. There are 3,920 customers in the UK, 94 customers in Germany, 87 customers in France, 30 customers in Spain, and 25 customers in Belgium.



The customers with the top 5 most orders in the UK are as follows:

Customer ID	12748	17841	13089	14606	15311
Number of Orders	209	124	97	93	91

The UK customers with the top 5 highest total spending amounts are presented in the following table:

Customer ID	18102	17450	16446	17511	16029
Total Spending Amount	$259,657	$194,550	$168,472	$91,062	$81,024

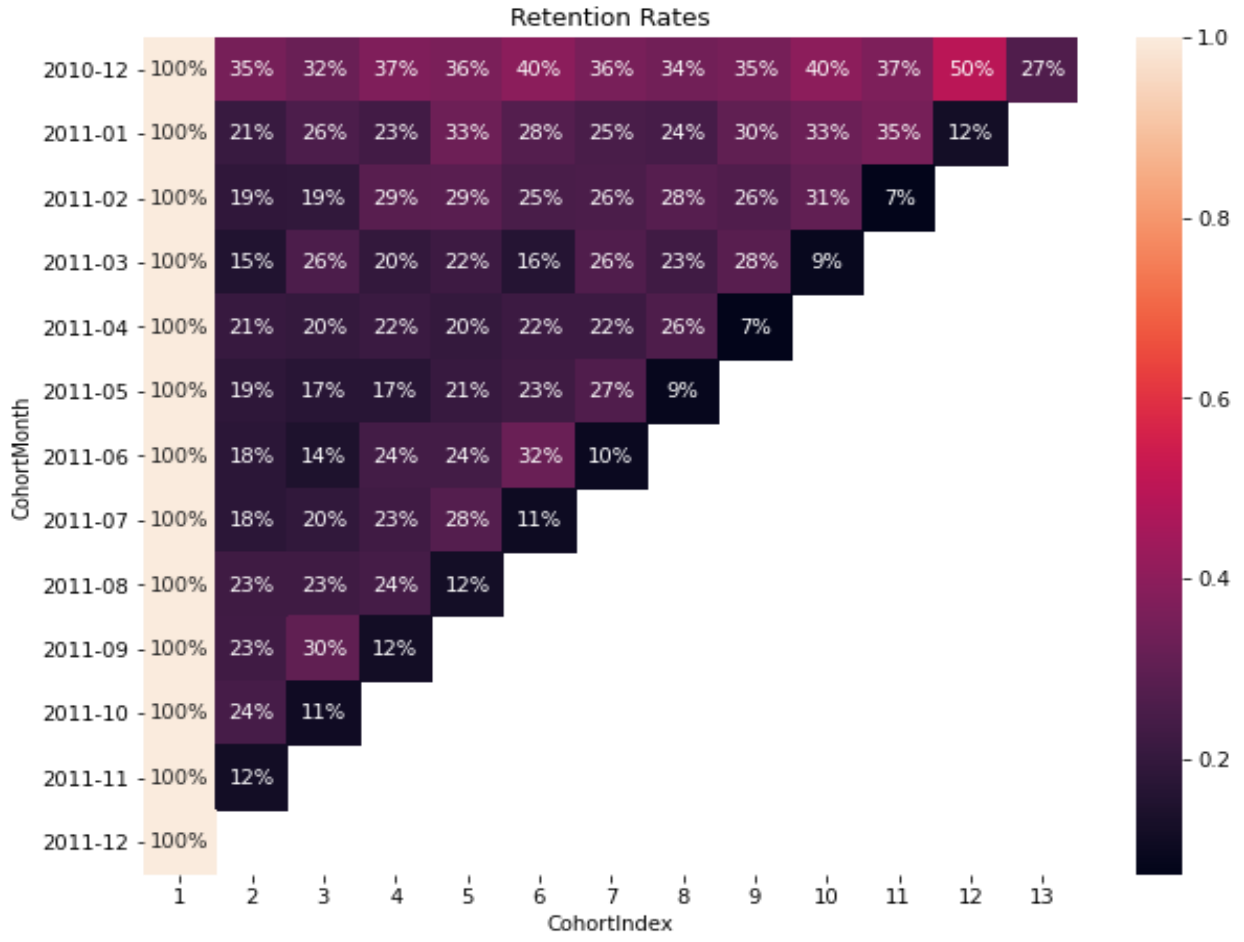
The top 5 dates with the greatest number of orders in the UK are shown in the following table:

Date	2010-12-02	2011-12-22	2011-11-23	2101-12-01	2011-12-29
Total Number of Orders	135	117	116	115	115

The top 5 days with the greatest transaction amounts in the UK are as follows:

Date	2011-12-09	2011-09-20	2011-01-18	2101-09-15	2011-10-03
Total Amount	$179,562	$100,475	$84,038	$67,891	$61,917

We also did cohort analysis on customers in the UK and calculated their retention rates. Below, I have included a heatmap corresponding to the retention rates of customers in the UK.

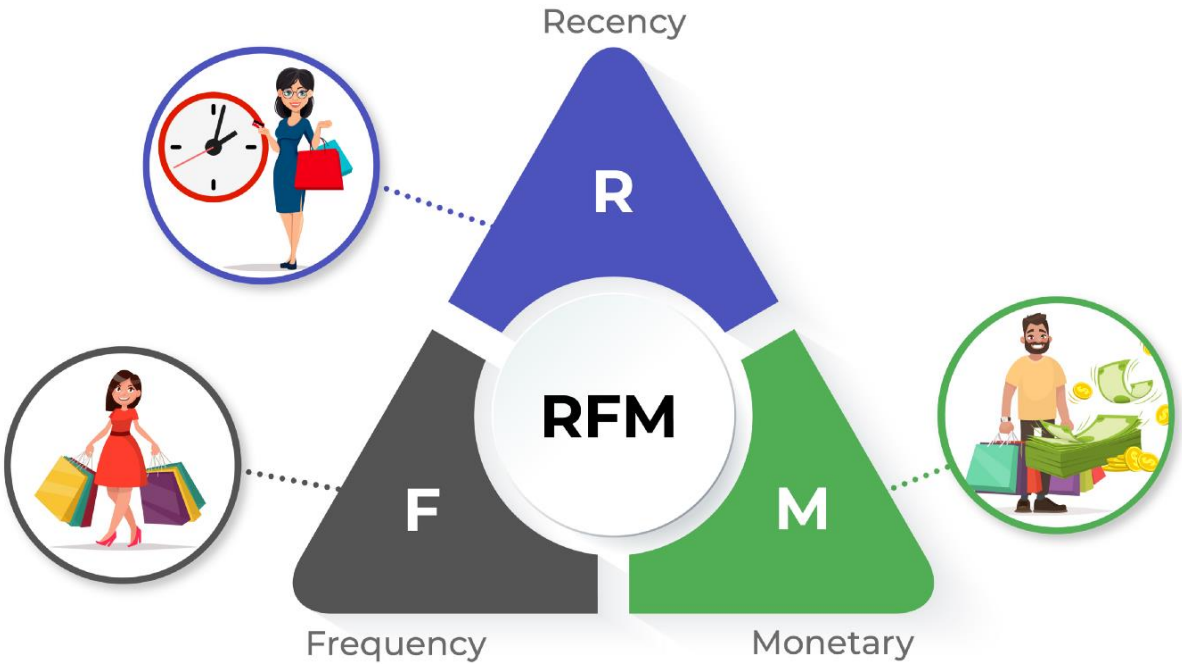


4. Pre-Processing

Since the number of orders and customers in the UK are far greater than those in other countries, I decided to focus on the UK for my RFM analysis and customer segmentation. In this step, I constructed a Recency, Frequency, and Monetary Value (RFM) table in preparation for conducting RFM analysis and building RFM customer segments. I performed log transformation on RFM data to reduce the skewness and used sklearn StandardScaler to standardize RFM data to the same average values and standard deviation; this is necessary for running the K-Means clustering model in the next step.

5. Modeling

5.1 RFM Customer Segmentation



Based on the RFM values, I assigned a score between 1 and 4 for each customer's Recency, Frequency and Monetary Value by quartiles. 4 is the best score and 1 is the worst score in each category; for example, customers who made purchases most recently, most often, and spent the most money would have a Recency score, Frequency score and Monetary Value score of 4-4-4. By combining each score, I built RFM segments and then calculated RFM score by adding them up.

CostomerID	Recency	Frequency	Monetary Value	Recency Score	Frequency Score	Monetary Value Score	RFM segment	RFM Score
12346	326	1	$77,183	1	1	4	114	6
12747	3	11	$4,196	4	4	4	444	12
12748	1	211	$34,345	4	4	4	444	12
12749	4	5	$4,090	4	3	4	434	11
12820	4	4	$942	4	3	3	433	10

Based on RFM segmentation, the top 10 largest segments are:

RFM Segment	444	111	112	211	333	344	433	233	212	311
Size	423	396	210	187	187	167	159	139	137	136

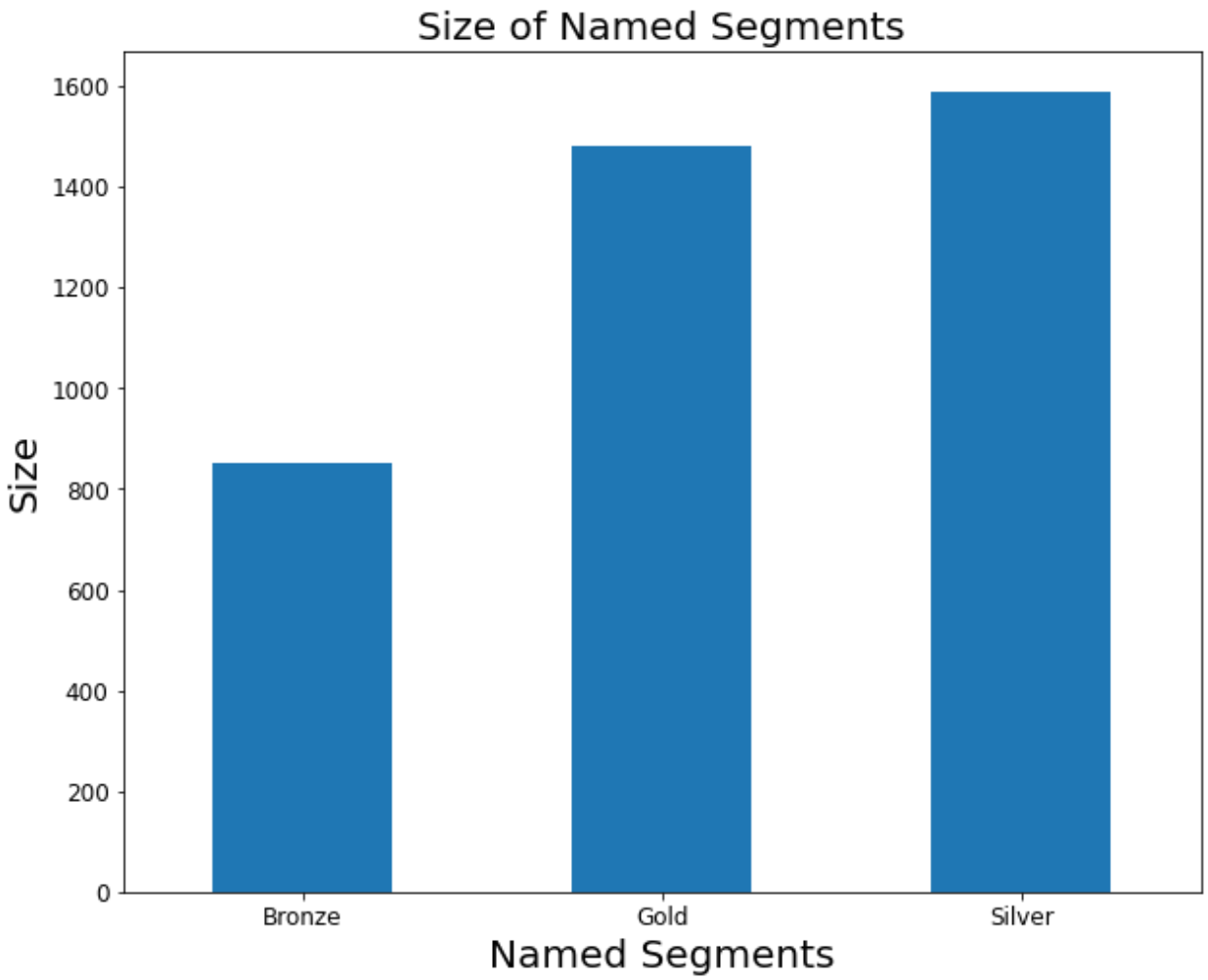
The bottom 10 smallest segments are:

RFM Segment	141	242	414	441	314	142	442	413	143	424
Size	1	1	1	1	2	2	3	3	4	4

The summary metrics based on the RFM scores are the following:

RFM Score	Count	Recency Mean	Frequency Mean	Monetary Value Mean
3	396	265.5	1.0	158.6
4	454	184.7	1.1	280.7
5	443	110.5	1.3	363.6
6	407	90.3	1.7	704.1
7	374	76.6	2.3	698.8
8	366	58.7	3.0	1126.1
9	409	45.8	4.0	1401.8
10	347	30.0	5.2	2337.6
11	301	21.1	8.0	3476.3
12	423	7.7	15.8	8469.8

Based on customers' RFM Scores, we also group customers into named segments: Gold (RFM Score > = 9), Silver (9>= RFM Score >= 5), and Bronze (RFM Score <= 4).

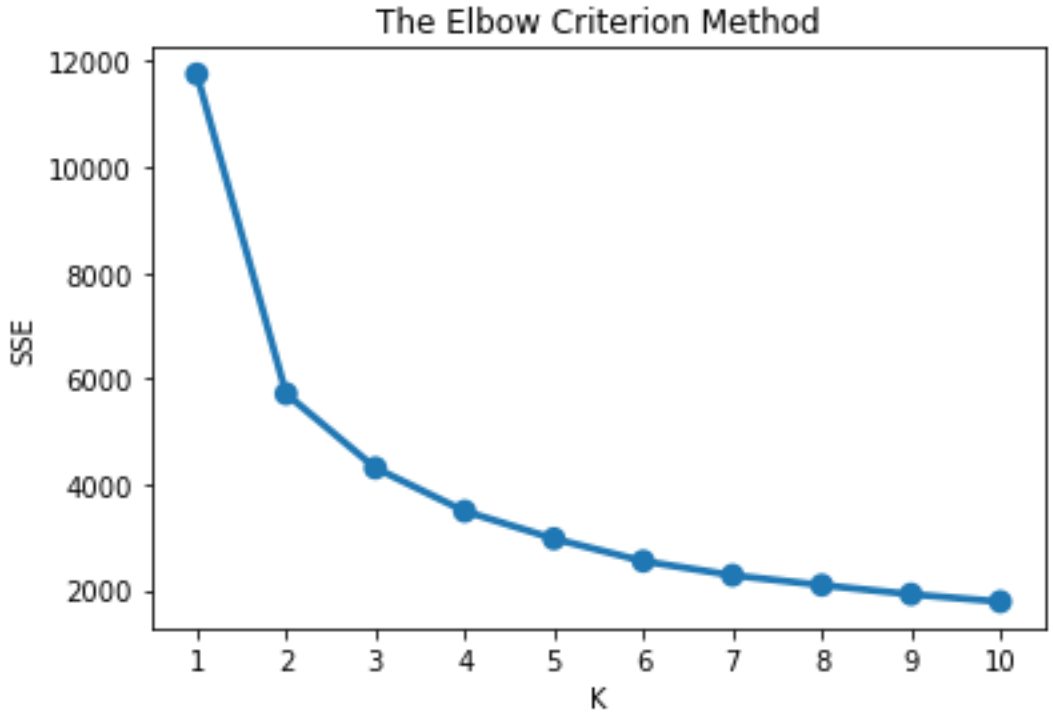


The summary metrics for each named segment is in the following table:

Named Segment	Count	Recency Mean	Frequency Mean	Monetary Value Mean
Gold	1480	26.2	8.5	$4,063
Silver	1590	85.4	2.0	$705
Bronze	850	222.4	1.1	$223

5.2 K-Means Clustering Model

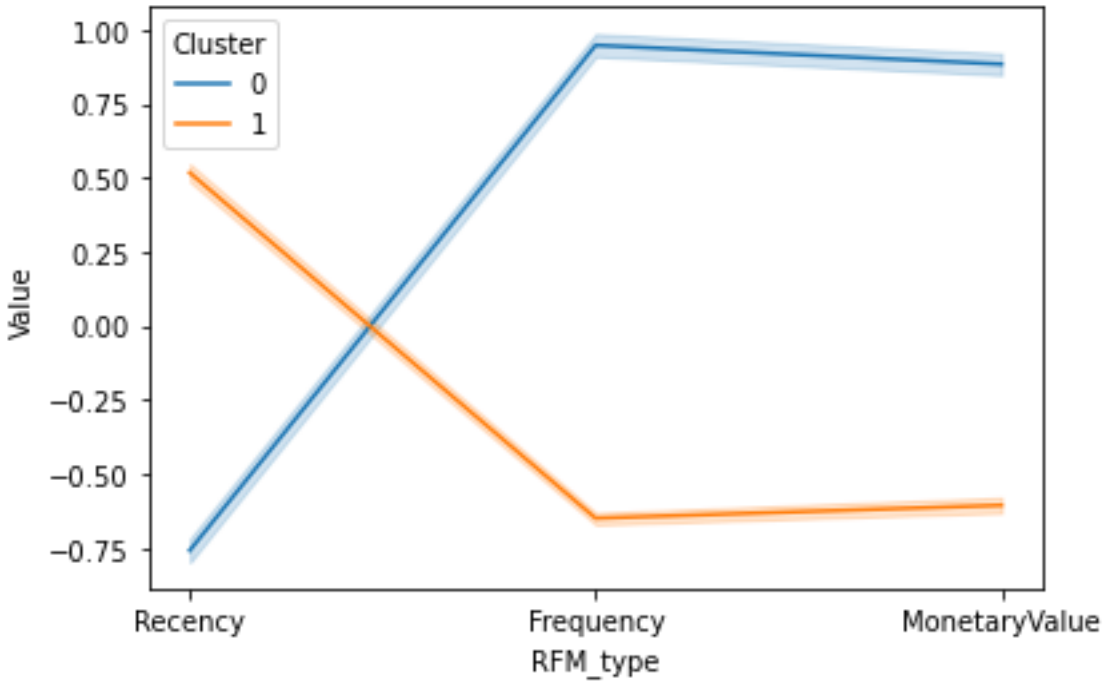
We used the elbow criterion method to choose number of clusters.



The best was K = 2. Running the K-Means clustering model with K=2 over the normalized RFM data, the summary metrics for each cluster is in the following table:

Cluster	Count	Recency Mean	Frequency Mean	Monetary Value Mean
0	1592	28.7	8.1	$3,957
1	2328	136.5	1.6	$440

The plot of standardized RFM is the following:



From this plot, we can clearly see the average Recency, average Frequency and average Monetary Value of customers in cluster 0 are much better than those in cluster 1 (smaller values are better for recency while higher values are better for frequency and monetary value).

6. Conclusion and Future Work

After conducting RFM analysis and constructing the K-means clustering model, I have concluded that the Gold, Silver, and Bronze segmentation method garners the most marketing insight. I recommend that the retailer target the Silver segment because this is the largest group, and these customers hold the most potential for the company. They are involved enough to be persuaded, but there is still plenty of room to improve through marketing. For example, the company might offer deals on Silver customers' next purchase before a certain date to boost recency and frequency scores. They could also try bundle deals to encourage higher monetary value purchases. I also suggest that the retailer take steps to maintain the Gold segment because this group includes many of their best customers and most profitable relationships. One possible strategy is to implement a point system with exclusive benefits; this is a great way to reward continued customer involvement.

Additionally, there are many opportunities for further research that could benefit this retailer. First of all, we could delineate between recency, frequency, and monetary value in terms of impact on RFM score. This could help us understand customer behavior within each segment. For example, two customers may both belong to the Silver segment, but one might be a regular customer who makes smaller purchases frequently while another might have made one large purchase a while ago. These two customers may respond to different kinds of marketing. Further analysis could provide these details so that the company can cater to these differences. We could also pinpoint the products that have been purchased most frequently and most recently and adjust prices accordingly. It might be helpful to find out when people spend the most money as well. This might provide some insight on seasonality and time-specific deals.