FINAL PROJECT REPORT

Customer Segmentation via RFM Analysis and Clustering Using Online Retail Data

Athaudage Sanuja Vihanga Senadeera, Justin Kyle Pedro, Mayra Geraldine Reinoso Varon,

Naveen Karan Krishna, Thi Anh Tram Le

Seneca Polytechnic

BAN230NAA: Applied Data Mining and Modelling

Professor: Roya Barzegar

Due Date: 11th August 2025

Table of Contents

Executive Summary

This project aimed to improve customer engagement strategies for a UK-based online retail company by identifying distinct customer segments based on their purchasing behaviors. Using the publicly available Online Retail dataset from UCI, we conducted an RFM (Recency, Frequency, Monetary) analysis to profile customers and applied KMeans clustering to segment them into actionable groups. This dual approach enables the business to identify loyal customers, high-value spenders, and at-risk groups, thereby maximizing return on marketing investments through personalized engagement campaigns. Key findings show the presence of valuable customer groups like "Champions," "Loyal Customers," and "Big Spenders," as well as vulnerable segments like "At Risk" and "Lost." These insights can inform data-driven strategies such as loyalty rewards, win-back programs, and tailored marketing communication.

Problem Statement

A UK-based online retail company is seeking to enhance its marketing performance by understanding different types of customers and tailoring engagement strategies accordingly. The lack of customer segmentation hinders the company's ability to personalize offers or re-engage inactive clients effectively. The primary objective of this project is to segment customers using their Recency, Frequency, and Monetary values derived from historical transaction data. By applying both RFM scoring and unsupervised clustering techniques, we aim to reveal customer behavior patterns and derive meaningful, business-ready segments that can inform targeted marketing initiatives.

Dataset Overview

The dataset used in this project is the Online Retail dataset from the UCI Machine Learning Repository. It consists of over 540,000 transactions made between December 2010 and December 2011 by customers from various countries, primarily the UK. Key attributes include Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country. For this project, we focused on cleaning and preparing the dataset by removing return transactions (negative quantities), dropping rows with missing Customer ID or Description, and calculating a new TotalPrice feature. The final dataset allowed us to build an RFM table with well-defined customer behavior metrics.

This dataset was created by Daqing Chen from the School of Engineering at London South Bank University and is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license allows for sharing and adapting the dataset for any purpose, including commercial use, provided appropriate credit is given to the original creator (Chen, 2012).

Methodology

We began by calculating three essential customer behavior metrics: Recency (days since last purchase), Frequency (number of unique purchases), and Monetary value (total amount spent). These metrics were computed using the cleaned dataset, grouped by Customer ID, with a reference date set one day after the most recent invoice in the data. After creating the RFM table, we applied RFM scoring by assigning quantile-based scores from 1 to 5 for each metric. Customers were then segmented into categories like "Champions," "Big Spenders," and "At Risk" based on their individual RFM scores using business rules.

For unsupervised clustering, we standardized the RFM values using StandardScaler to normalize the feature scales. We then used the KMeans algorithm to perform clustering, determining the optimal number of clusters via both the Elbow Method and Silhouette Score analysis. The Silhouette Score peaked at k=2 (0.89), but we selected k=3 for interpretability and balance. The resulting clusters were labeled as "Big Spenders," "Loyal Customers," and "Lost," based on their average RFM values.

Findings and Interpretation

The RFM-based segmentation revealed meaningful customer groups. "Champions" had the lowest Recency (5.45), highest Frequency (18.24), and the highest Monetary value (11,221.74), representing the most engaged and profitable customers. "At Risk" customers showed high Recency (228.5), low Frequency (1.04), and low Monetary value (230.3), indicating a high probability of churn. "Big Spenders" stood out with a high Monetary value (4236.60) and above-average Frequency (3.64), making them ideal for loyalty incentives.

Cluster analysis complemented the RFM segmentation. The three clusters showed distinct behavior patterns: Cluster 0 ("Big Spenders") had the highest Monetary average; Cluster 1 ("Loyal Customers") exhibited high Frequency and low Recency, indicating active engagement; Cluster 2 ("Lost") had the highest Recency and lowest Frequency. A heatmap comparison of RFM segments vs. cluster assignments showed good alignment between the two techniques. For instance, most "Champions," "Loyal Customers," and "Recent Customers" fell into Cluster 1, validating the clustering outcome.

Business Recommendations

Based on the combined RFM and clustering analysis, we recommend the following marketing strategies:

- **Champions and Loyal Customers:** Implement loyalty rewards, early access to promotions, and exclusive offers to retain these high-value customers.
- **Big Spenders:** Target with upselling campaigns, premium membership tiers, and personalized product bundles.
- **At Risk and Lost:** Initiate win-back campaigns through limited-time offers, re-engagement emails, and personalized discounts to re-capture their interest.
- **Recent Customers:** Send welcome messages and offer discounts on second purchases to nurture long-term loyalty.

These tailored strategies will help the company focus resources where they yield the highest ROI.

Limitations

This project has several limitations that should be acknowledged. First, the dataset covers only one year of transactions, which may not fully capture long-term customer behavior. Second, external factors like seasonality or marketing promotions are not considered in the RFM metrics. Third, the analysis assumes each customer ID corresponds to a unique customer, but shared IDs or household purchases could skew results. Also, the KMeans algorithm is sensitive to outliers, and while we applied standardization, additional outlier detection or robust scaling could improve clustering accuracy. Lastly, the analysis does not consider product categories or geographic diversity, which may provide additional segmentation dimensions.

Conclusion

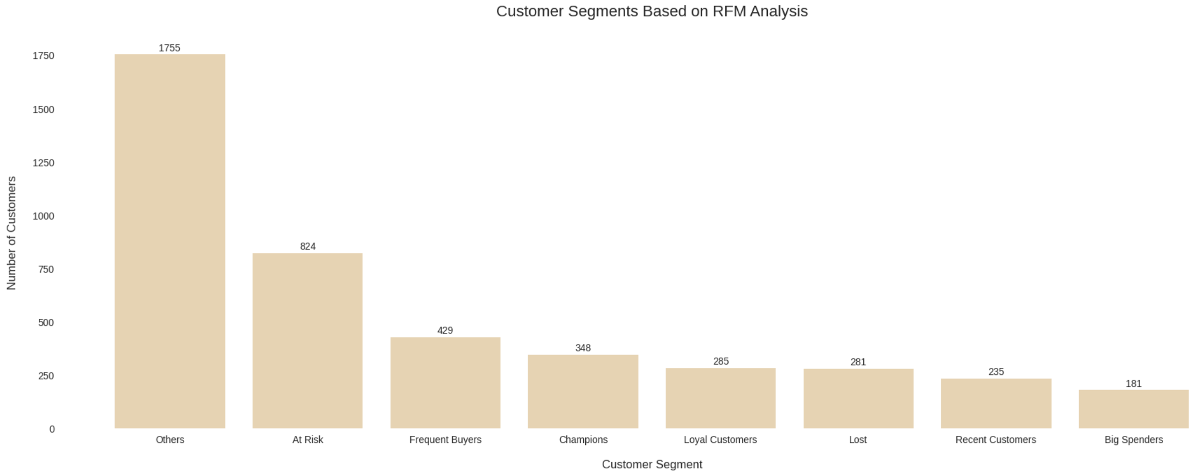
This project successfully demonstrated the value of combining RFM analysis with KMeans clustering to generate actionable customer segments. By cleaning and preparing the Online Retail dataset, building RFM metrics, and applying unsupervised learning, we identified customer groups with distinct behaviors and strategic importance. The synergy between RFM scores and clustering results enabled a robust segmentation model, providing valuable insights into customer loyalty, spending, and risk. With these insights, the retailer can enhance its marketing performance, improve customer retention, and optimize resource allocation. Future work could incorporate more sophisticated algorithms (e.g., DBSCAN, Gaussian Mixture Models), product-level insights, and longer-term data for a more comprehensive view.

References

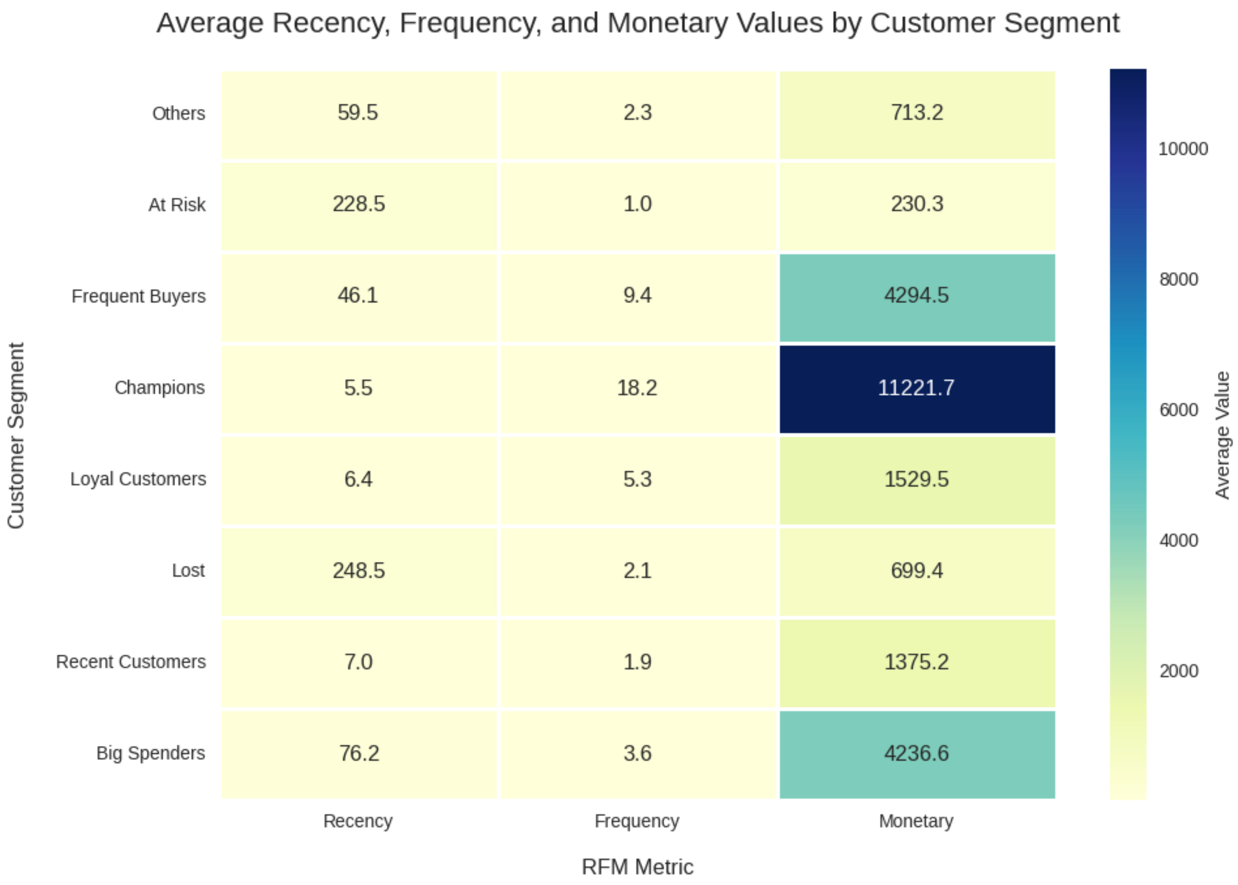
Chen, D. (2012). *Online Retail Dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Online+Retail

Appendices

Appendix A

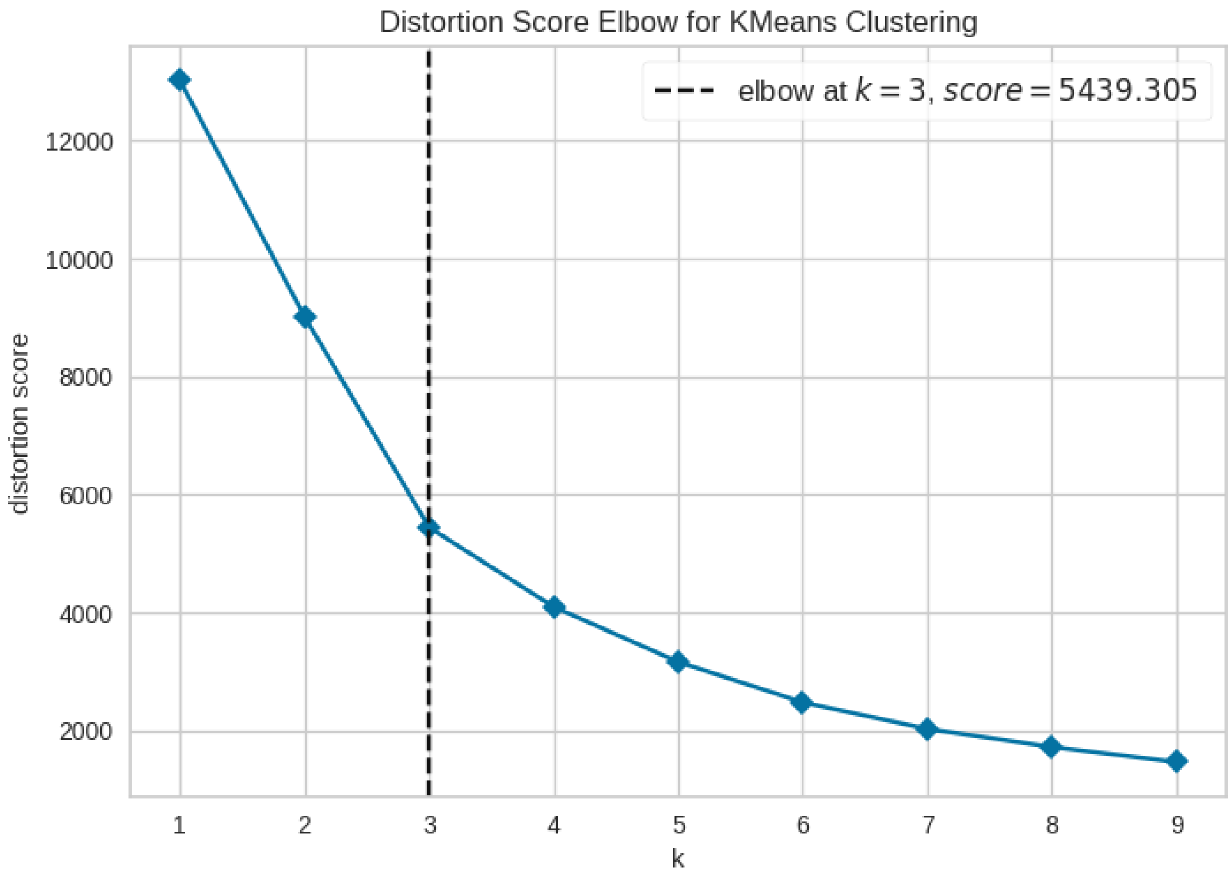


Appendix B

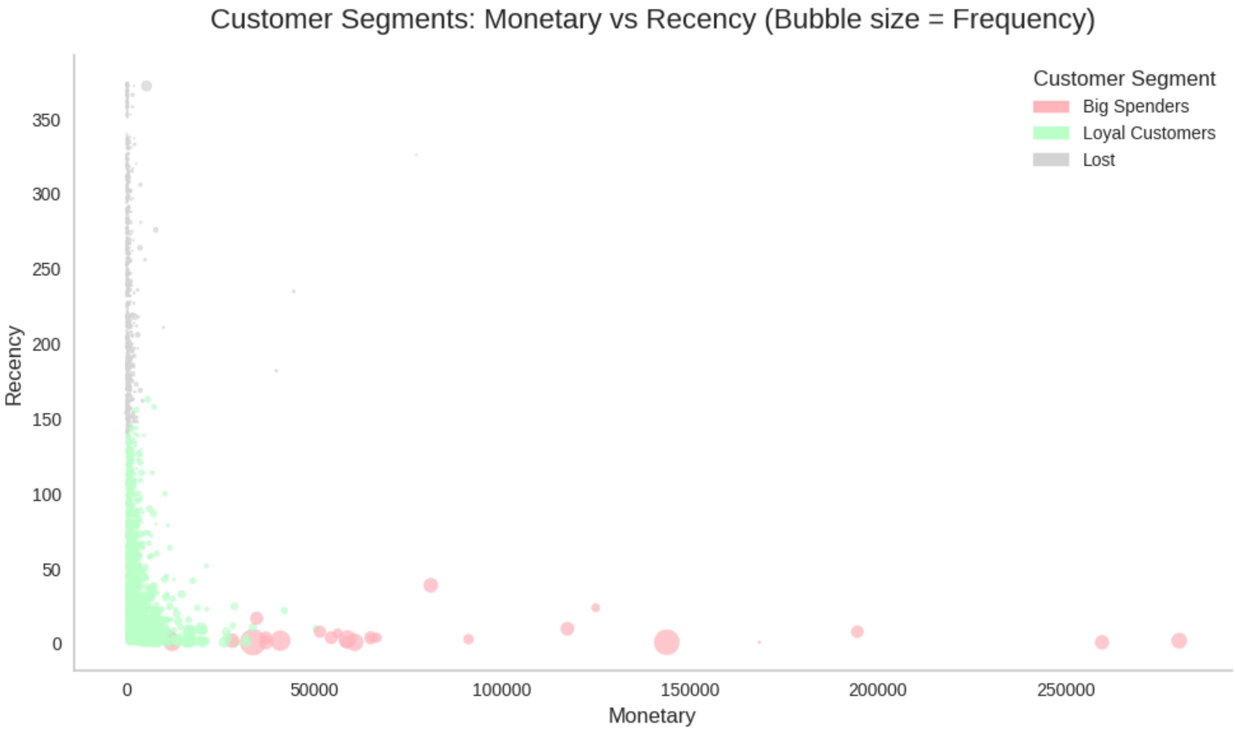


	Recency	Frequency	Monetary
Others	59.5	2.3	713.2
At Risk	228.5	1.0	230.3
Frequent Buyers	46.1	9.4	4294.5
Champions	5.5	18.2	11221.7
Loyal Customers	6.4	5.3	1529.5
Lost	248.5	2.1	699.4
Recent Customers	7.0	1.9	1375.2
Big Spenders	76.2	3.6	4236.6

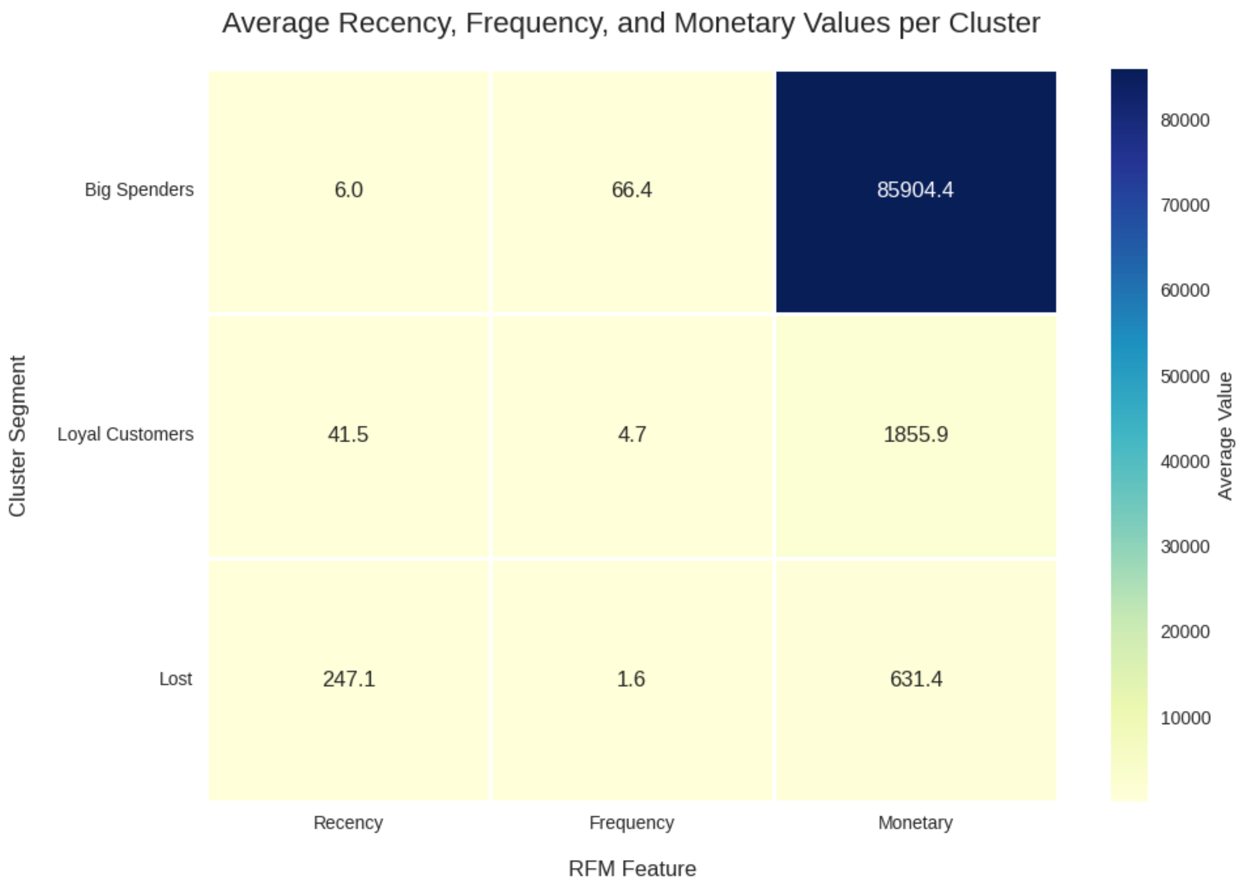
Appendix C


Distortion Score Elbow for KMeans Clustering
elbow at $k = 3$, $score = 5439.305$
distortion score
12000
10000
8000
6000
4000
2000
1
2
3
4
5
6
7
8
9
k

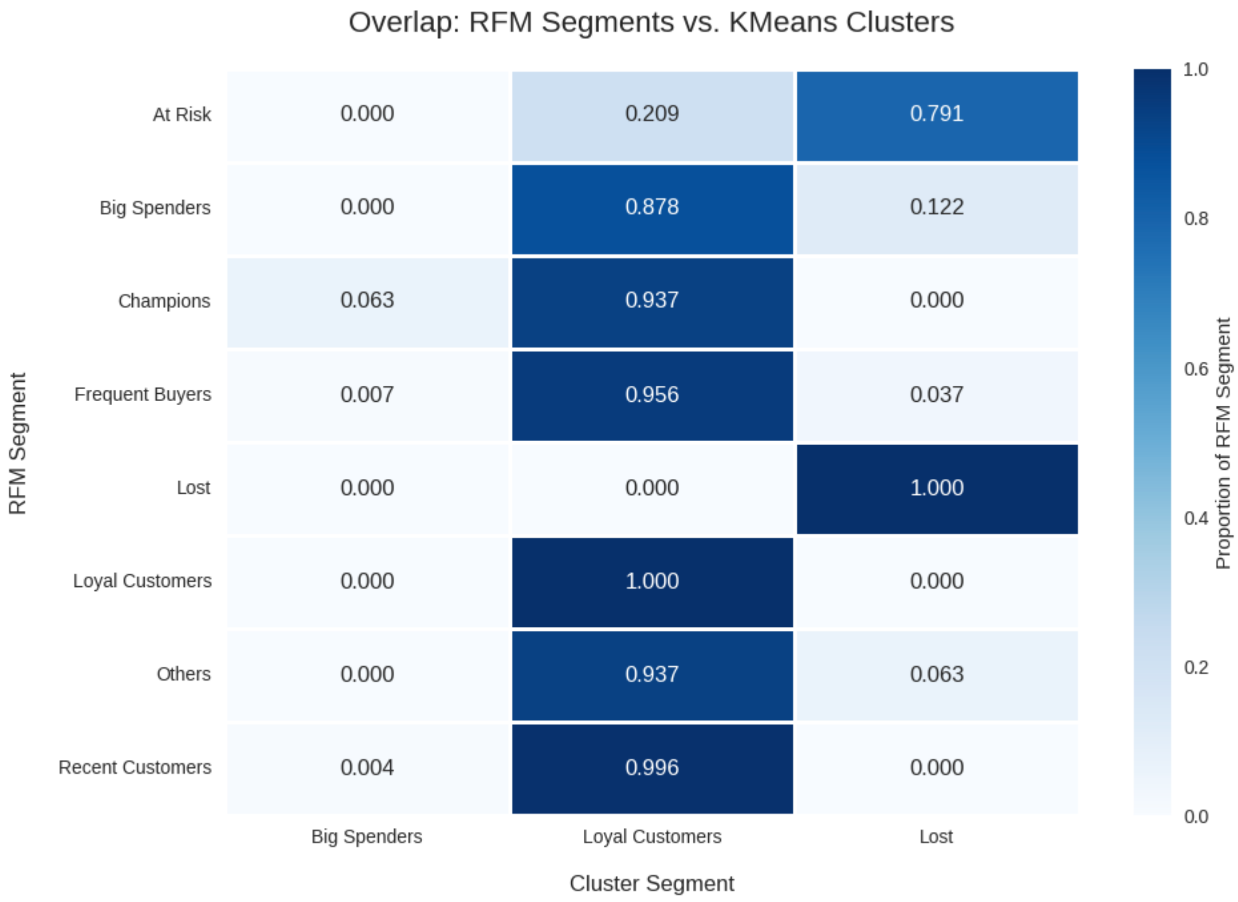
Appendix D


Customer Segments: Monetary vs Recency (Bubble size = Frequency)
Customer Segment
Big Spenders
Loyal Customers
Lost
Recency
350
300
250
200
150
100
50
0
0
50000
100000
150000
200000
250000
Monetary

Appendix E


Average Recency, Frequency, and Monetary Values per Cluster
Cluster Segment
Big Spenders
Loyal Customers
Lost
6.0
66.4
85904.4
41.5
4.7
1855.9
247.1
1.6
631.4
Recency
Frequency
Monetary
RFM Feature
80000
70000
60000
50000
40000
30000
20000
10000
Average Value

Appendix F



RFM Segment	Big Spenders	Loyal Customers	Lost
At Risk	0.000	0.209	0.791
Big Spenders	0.000	0.878	0.122
Champions	0.063	0.937	0.000
Frequent Buyers	0.007	0.956	0.037
Lost	0.000	0.000	1.000
Loyal Customers	0.000	1.000	0.000
Others	0.000	0.937	0.063
Recent Customers	0.004	0.996	0.000