Marketing Automation Customers Segmentation

Hamad Murad
hkm2358@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Master's Project is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
RIT
Marketing Automation Customers Segmentation

by

Hamad Murad

A Capstone Submitted in Partial Fulfilment of the Requirements for the
Degree of Master of Science in Professional Studies:
Data Analytics
Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

December 19, 2021
RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Hamad Murad

Graduate Capstone Title: Marketing Automation Customers Segmentation

Graduate Capstone Committee:

Name: Dr. Sanjay Modak

Chair of committee

Name: Dr. Hammou Messatfa
Member of committee
Acknowledgments ............................................................................................................................. 3

Abstract ............................................................................................................................................. 4

Chapter 1 ......................................................................................................................................... 5
  1.1 Background ................................................................................................................................. 5
  1.2 Statement of problem .................................................................................................................... 6
  1.3 Objectives and Deliverables .......................................................................................................... 7
  1.4 Methodology ................................................................................................................................. 7
  1.5 Limitations of the Study .................................................................................................................. 7

Chapter 2 – Capabilities Used ............................................................................................................. 8
  2.1 Python .......................................................................................................................................... 8
  2.2 RFM .............................................................................................................................................. 8
  2.3 K-means ....................................................................................................................................... 9
  2.4 Decision Tree ............................................................................................................................... 9

Chapter 3 – Literature Review ............................................................................................................ 9

Chapter 4- Project Description .......................................................................................................... 13
Chapter 5 - Project Analysis ....................................................................................................................

5.1 The Data Source ............................................................................................................................

5.2 The Socio demographic Data ........................................................................................................

5.3 Data discovery ................................................................................................................................

5.4 RFM Analysis ...................................................................................................................................

5.5 RFM score calculation ....................................................................................................................... 

5.6 Segmentation based on RFM Score ............................................................................................... 

5.7 Segmentation description using decision tree ............................................................................... 

5.8 K-MEANS SEGMENTATION .......................................................................................................... 

5.9 Python K-means program ............................................................................................................. 

5.10 Data ............................................................................................................................................. 

5.11 Result of K-mean ..........................................................................................................................

Chapter 6 Conclusion .............................................................................................................................

6.1 Conclusion ....................................................................................................................................

6.2 Recommendations ........................................................................................................................

References .............................................................................................................................................
Figure 1... Euclidean distance formal ........................................................................................................ 11
Figure 2 ... Cosine correlation ................................................................................................................ 11
Figure 3 ... Project Description. .................................................................................................................. 14
Figure 4... Data Source .................................................................................................................................. 15
Figure 5... The Socio demographic Data ..................................................................................................... 16
Figure 6... Sample Histogram of the dataset fields before applying the RFM ................................................. 17
Figure 7... Simple Histogram Count Of Gender ............................................................................................ 17
Figure 8 ... Simple Histogram Count of Country .......................................................................................... 18
Figure 9... Simple Histogram Count of Income ............................................................................................ 19
Figure 10... RFM data .................................................................................................................................. 20
Figure 11 ... RFM Score .................................................................................................................................. 21
Figure 13... segments .................................................................................................................................. 23
Figure 14... Data Exploration 1. .................................................................................................................... 38
Figure 15... Data Exploration 2 ..................................................................................................................... 23
Figure 16... Data Exploration 3 ..................................................................................................................... 38
Figure 17... Data Exploration 4 ..................................................................................................................... 23
Figure 18... Relationship Map ....................................................................................................................... 39
Acknowledgments

I would like to express my sincere gratitude to several individuals and RIT for approving me throughout my Capstone. First, I wish to express my sincere gratitude to my supervisor, Dr. Hammou Messatfa for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and writing of this thesis.

His immense knowledge, profound experience and professional expertise in Data Quality Control has enabled me to complete this research successfully. Without his support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

Secondly, I would also like to thank the Chair Committee: Dr. Sanjay Modak who helped me in finalizing this project within the limited time frame.

Finally, last but by no means least: Also wish to express my sincere thanks to the RIT University and my parents who helped me a lot and supported me from the beginning.

Thanks for all your encouragement!
Abstract.
Understanding the various facets of a customer’s data is very important for businesses. This insight can help them identify patterns and opportunities to improve their operations. RFM values are often used to identify which customers are valuable for a company. They are then used to identify which promotional activities are most appropriate for them. K-means clustering is an unsupervised learning technique that works when you have unlabeled data. It allows the identification of new data points and groups.

This paper proposes a novel approach that combines data cleaning and analyzing customer data to divide a broad market into various consumer groups, and then designing and implementing marketing campaigns that effectively target these consumers. This study aims to identify profitable segments based on historical data (such as purchased items and the associative monetary expenses). The proposed model is formulated using a decision tree and the RFM model to put the customers in three segments (Gold or Silver or Bronze).

The decision tree is a type of supervised algorithm that can interpret clustering. It will use labels to classify the data and then plot the clusters.

Keywords: RFM, K-means clustering, Decision-Tree.
Chapter 1

1.1 Background.
Customer segmentation is a process that involves grouping different types of customers in order to improve the understanding and effectiveness of marketing and customer relationship management. The characteristics of each segment indicate that the same purchasing behavior exists within the same market, hence, establishing a segment for different services can help you achieve competitive advantage and enhance your business performance and the maximization of a customer’s value. RFM analysis is a technique used to identify the best customers and develop targeted marketing campaigns. It uses recency, frequency, and monetary totals to identify the most profitable customers.

RFM analysis ranks each customer on the following factors:

1. Recency is the amount of time that a customer has been using or buying a product after making a purchase. This is the reason why businesses keep track of it.
2. Frequency is a good measure to evaluate how often a given customer made a purchase in a given period. It can also be used to identify repeat customers.
3. Monetary is referring to how much money did a customer spend during a given period, this is a good indicator of how much money the customer will spend in the future.

RFM analysis scores customers on three main factors: profitability, customer loyalty, and overall satisfaction. A score of 1 to 5 is given, with 5 being the highest. The RFM cell is a collection of three values that identify a customer. It is used to identify the most valuable customers. A dealership may not realize that a high-frequency customer is likely to buy multiple cars in the near future. But, in order to attract this type of customer, the dealership may use the frequency score to evaluate the potential buyer. RFM analysis can also be used by non-profit organizations to find the best donors and potential customers. It can also be used by charities and nonprofits to find out which individuals are most likely to give again in the future.

RFM only collects historical data about customers, and it cannot predict the future behavior of those customers. Also, it uses historical data to collect data about non-existent customers.
1.2 **Statement of problem.**

Nowadays with the market competitors increasing, it is not an easy task to maintain good relationships with the customers, specially in the retails sector. And to stay in competition, we need to realize the customers’ needs properly. Ignoring customer's needs can decrease the customer's engagement. The firms will not reasonably predict which customers are likely to purchase their products again, how much revenue comes from new (versus repeat) clients, and how to turn occasional buyers into habitual ones.

1.3 **Objectives and Deliverables.**

This study aims to establish the link between marketing campaign and the customer segmentation, by grouping customers based on their buying behavior, in terms of how recently they bought, how often did they buy, and how much monetary value they contributed to the online store using RFM and k-means clustering techniques.

This model demonstrates the benefits of using RFM for analyzing customers’ purchasing data in retail sector. Analyses included two phases. Firstly, K-Means clustering was included, where the customers were clustered according to their RFM. Then, with demographic data, each cluster was partitioned again into new clusters.

1.4 **Methodology.**

The proposed research methodology consists of four phases which are Data Cleaning, RFM analysis, K-means and Decision Tree. The first phase involves preparing the data cleaning, then, the second phase is the analysis of customer RFM which initiates by grouping customers based on their buying behavior, in terms of how recently they bought, how often did they buy, and how much monetary value they contributed to the store.

In RFM Analysis, three parameters are analyzed, each denoted by the letters R, F, and M. To satisfy the need of knowing true customer value, analysis of just one parameter will give an inaccurate report of the customer base, so the customer’s lifetime value can’t be reliable. This is why at least three parameters of customer’s purchase behavior are analyzed; with the freedom to add other analytical parameters too. The third phase is the k-means cluster analysis to be performed, which result in groups of customers with more homogeneous characteristics, then the last phase is to generate the decision tree. RFM scores are computed to determine the purchase behavior of a
customer. They are based on a simple rating system that shows the most profitable and most loyal customers.

1.5 Limitations of the Study

The findings of this study have several limitations that indicate the direction of future research. The study only focused on online orders and its results were limited to that data. Also, this study is limited to RFM, K-means and decision Tree. This study would be limited to a medium portion of the database used and a limited number of selective variables.

Chapter 2 – Capabilities Used.

2.1 Python.

Python is a programming language that’s commonly used to build websites and software. It’s also used to perform various tasks, such as analyzing data. Python is a programming language that is built on top of an object-oriented design philosophy. Its code readability is ensured by its use of significant indentation, and its object-oriented approach simplifies programming.

2.2 RFM.

RFM analysis is a technique used to identify and target the best customers based on their recency, frequency, and monetary totals. It is a marketing technique that uses numerical scores to rank and group customers. RFM analysis scores customers on a scale from 1 to 5, with 5 being the highest
score. The RFM cell collects three values for each customer. These values are then averaged together to find the most valuable customers.

2.3 K-means.
K-means clustering is a method that combines the search for groups that have not been explicitly defined in the data with the search for unknown groups. This allows the developers to identify nonlabeled groups in the data. K-Means clustering is a method that enables groups to be separated into k clusters. These k groups are defined as the nearest mean for each object.

2.4 Decision Tree

A decision tree is an algorithm that works for supervised learning problems that involve classification or regression. It has an input feature that is labeled with a value, and the arcs coming from it are indicated with the possible values of that feature.

A tree can be learned by splitting the source set into multiple sub-sets, which are then recursively partitioned. The first set of sub-sets is then recursively partitioned as well.

Chapter 3 – Literature Review

The RFM model is a widely used segmentation technique that consists of three measures: recency, frequency, and monetary. It is typically used for estimating product characteristics. For instance, among individuals with higher frequency, those with lower future purchasing potential were more likely to have lower future purchasing rates.

Different RFM quintiles have different response rates. For recency, customers are sorted by purchase dates. Recency is commonly defined by the number of periods since the last purchase, which measures the interval between the most recent transaction time and the analyzing time (days or months), that is, the lower the number of days, the higher the score of recencies.

A customer having a high score of recency implies that he or she is more likely to make a repeat purchase. The top 20% segment is coded as 5, while the next 20% segment is coded as 4 and so forth. Finally, the recency for each customer in the database is denoted by a number from 5 to 1.
For frequency, the database is sorted by purchase frequency (the number of purchases) made in a certain time period. The definition of frequency is often simplified to consider two states, including single and repeated purchases. The top quintile is assigned a value of 5 and the others are given the values of 4, 3, 2 and 1. However, higher frequency score indicates greater customer loyalty. A customer having a high score of frequency implies that he or she has a great demand for the product and is more likely to purchase the products repeatedly. For monetary, customers are coded by the total amount of money spent during a specified period of time. The definition of monetary is defined by the dollar value that the customer spent in this time period or by the average dollar amount per purchase or all purchases to date (Wei, 2010).

An important marketing strategy that is widely used by businesses is customer segmentation. The point of customer segmentation is to split the user-base into smaller groups that can be targeted with specialized content. The produced customer groups are drawn from user behavior which gives the business a deeper understanding of the types of customers that exists in the system. The benefit of customer segmentation is twofold (Khare, 2016).
Firstly, a better knowledge about the types of users in a system can lead to better business and marketing strategies. Secondly, a user is likely to use an application more often if he/she always receives relevant content.

Another essential point is that if a customer is pleased, he/she is more likely to recommend the business to other people which helps in the expansion of a company. This type of marketing technique is a subset of a company’s Business Intelligence to be able to create a set of similar customer groups, an extensive analysis of the available data combined with research and evaluation of clustering algorithms is needed. The available data is the most vital part of any clustering algorithm. The most important aspects are the quality and amount of the available data. To run some sort of similarity function to cluster items or customers in a system, the data needs to be arranged into feature vectors with a set of feature values. To achieve the best results, a large amount of data is needed and more importantly the absence of data points needs to be minimal. Another important aspect in customer segmentation is to understand the available data. In a system where items are rated using some sort of scale, e.g., a rating from zero to five, it is easy to interpret a user’s preferences. However, in systems where the set of items is not predefined, as in an Emarketplace where users upload items which are removed when sold, it is much harder to determine a user’s preference (Hidayat, Rismayati, Tajuddin and Merawati, 2020).

Clustering algorithms are used to assign users into groups so that users belonging to the same group are more similar than users in another group.

The goal of this division is to find meaningful underlying patterns within the data space. User similarity is determined by a distance measure. This section will introduce the most common similarity measures and clustering algorithms (Matkowski, 2020).

Clustering is highly dependent on defining a relevant similarity or distance measure. The simplest and most common distance measure is Euclidean distance:

\[ d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \]

*Figure 1... Euclidean distance formal*
Where \( n \) is the number of features in the data objects \( x \) and \( y \), and \( x_k \) and \( y_k \) are the \( k \)th attribute of the feature data objects \( x \) and \( y \) respectively. Cosine correlation is widely used in the field of recommender systems, mainly in collaborative filtering. The main idea behind cosine correlation is to compute the cosine value of the angle that two \( n \)-dimensional feature vectors form (Wei, 2010). This is possible using the following equation, where \( n \) is the number of features in the data objects \( x \) and \( y \), \( \cdot \) indicates the vector dot product and \( k \times k \) is the norm of vector \( x \):

\[
\cos(x, y) = \frac{(x \cdot y)}{\| x \| \| y \|}
\]

**Figure 2 ... Cosine correlation**

The final distance measure that will be covered in this report is Pearson correlation. This distance measure is also widely used in recommender systems. Pearson correlation computes the linear relationship between two feature vectors, in other words two feature vectors are similar if a best fitting straight line is close to all data points in both vectors (Heldt, 2021).

K-means clustering is widely used in the field of cluster analysis and customer segmentation. K-means is an algorithm designed to group a set of items into \( K \) subgroup or clusters. The algorithm is dependent on a manually set value for \( K \). The \( K \) centroids are initialized to random observations in the dataset. K-means is then tasked with iteratively moving these centroids to minimize the cluster variance using two steps:

1. For each centroid \( c \) identify the subset of items that are closer to \( c \) than any other centroid using some similarity measure.
2. Calculate a new centroid each cluster after every iteration which is equal to the mean vector of all the vectors in the cluster.

This two-step process is repeated until convergence is reached. The standard implementation of K-means uses Euclidian distance measure described in the section above to find the subset of items that corresponds to each cluster. This is done by calculating mean squared error, which in this case is equivalent with the Euclidian distance, of each item’s feature vector with the \( K \) centroid and choosing the closest result. However, other distance measures can be used instead of Euclidian.
distance. Aggarwal et al. Claim that for high dimensional data, the choice of distance measure used in clustering is vital for its success. (Li and Yan, 2019).

By choosing a cluster, we can then evaluate its characteristics using a decision tree. After identifying the appropriate cluster, we can then train a CART model to highlight its characteristics. To train a decision tree, first train it by using the labeled dataset. Then, structure the problem as a binary classification problem, where \( y \) is equal to 1 if the point is in the chosen cluster and 0 otherwise (Hwang, 2021).

Chapter 4- Project Description.

This project consists of four phases which are Data Cleaning, RFM analysis, K-means and decision Tree. The first phase involves preparing the data cleaning, then the second phase is the analysis of customer RFM which initiates by grouping customers based on their buying behavior, in terms of how recently they bought, how often did they bought, and how much monetary value they contributed to the online store.

The preparing data is the initial stage of this project, the online retail data has been collected and addressed to be stable for the next step of the description analysis, after some digging, we found that there is a need to clean up the data.

This tutorial will help you get the most out of your data, this group filters out transactions where the price is zero or less. We only selected transactions that have clear transactions and are not complex, we have created a sum column where we can calculate the revenue for each order-line.
In this study, we will show the data listings for the RFM model.

Step #1 Calculating the RFM values.

Creating an RFM model from the data collected during a transaction. This step will help make the model easier to understand.

The RFM score is computed by taking the data points for each customer. These are the points that determine the frequency and monetary value of each transaction.

Frequency: The number of purchases within the customers life span

Monetary value: The mean monetary value for the customers transactions.

Recency is the age of the customer at the time of the last transaction, this is different from RFM where recency is computed as the number of days since the last purchase.

Each RFM score is preceded by a segment.

Figure 3 ... Project Description.
Chapter 5 - Project Analysis

We are going to use a publicly available online retail transaction dataset from Kaggle, which includes the transaction information of each customer from all over the world. It includes information such as invoice number, invoice date, customer id, description of the product, purchased quantity, and country where the customer lives. In this study we started with data exploration and data preparation steps, and then start with a clean dataset. In this research, the raw data is a table of transaction records with the following fields:

5.1 The Data Source.
<table>
<thead>
<tr>
<th>InvoiceNo</th>
<th>StockCode</th>
<th>Description</th>
<th>Quantity</th>
<th>InvoiceDate</th>
<th>UnitPrice</th>
<th>CustomerID</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>554065</td>
<td>SMALL PURPLE BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>.850</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>2</td>
<td>554065</td>
<td>SMALL RED BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>.850</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>3</td>
<td>554065</td>
<td>SMALL YELLOW BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>.850</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>4</td>
<td>554065</td>
<td>LARGE YELLOW BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>5</td>
<td>554065</td>
<td>LARGE PURPLE BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>6</td>
<td>554065</td>
<td>LARGE RED BABUSHKA NOTEBOOK</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>7</td>
<td>554065</td>
<td>SET 7 BABUSHKA NESTING BOXES</td>
<td>4</td>
<td>5/22/2011</td>
<td>8.500</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>8</td>
<td>554065</td>
<td>PACK OF 6 HANDBAG GIFT BOXES</td>
<td>6</td>
<td>5/22/2011</td>
<td>2.550</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>9</td>
<td>554065</td>
<td>PACK OF 6 SWEETIE GIFT BOXES</td>
<td>6</td>
<td>5/22/2011</td>
<td>2.550</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>10</td>
<td>554065</td>
<td>PINK GINGHAM CAT WITH SCARF</td>
<td>6</td>
<td>5/22/2011</td>
<td>2.550</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>11</td>
<td>554065</td>
<td>BLUE CIRCLES DESIGN MONKEY DOLL</td>
<td>6</td>
<td>5/22/2011</td>
<td>2.550</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>12</td>
<td>554065</td>
<td>PINK FLOWER FABRIC PONY</td>
<td>4</td>
<td>5/22/2011</td>
<td>3.750</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>13</td>
<td>554065</td>
<td>SET OF 6 T-LIGHTS TOADSTOOLS</td>
<td>6</td>
<td>5/22/2011</td>
<td>2.950</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>14</td>
<td>554065</td>
<td>ASSORTED COLOUR LIZARD SUCTION HOOK</td>
<td>24</td>
<td>5/22/2011</td>
<td>420</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>15</td>
<td>554065</td>
<td>APPLE BATH SPONGE</td>
<td>30</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>16</td>
<td>554065</td>
<td>STRAWBERRY BATH SPONGE</td>
<td>30</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>17</td>
<td>554065</td>
<td>ICE CREAM SUNDAE LIPGLOSS</td>
<td>24</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>18</td>
<td>554065</td>
<td>ICE CREAM PEN LIPGLOSS</td>
<td>24</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>19</td>
<td>554065</td>
<td>PINK DOUGHNUT TRINKET POT</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.650</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>20</td>
<td>554065</td>
<td>STRAWBERRY CERAMIC TRINKET BOX</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.250</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>21</td>
<td>554065</td>
<td>CERAMIC CHERRY CAKE MONEY BANK</td>
<td>12</td>
<td>5/22/2011</td>
<td>1.450</td>
<td>18287</td>
<td>United Kingdom</td>
</tr>
</tbody>
</table>

Figure 4... Data Source

5.2 The Socio demographic Data.
5.3 Data discovery.

Data discovery is a process that involves gathering and analyzing vast amounts of data. This process is usually carried out to identify patterns and uncover hidden gems in the data.

Histogram of the dataset fields before apply the RFM
Figure 6… Sample Histogram of the dataset fields before applying the RFM.
**Figure 7... Simple Histogram Count of Gender**

![Histogram Count of Gender](image-url)
Figure 8  ... Simple Histogram Count of Country
5.4 RFM Analysis.

The first thing we’ll calculate is the three key factors of RFM Analysis (recency, frequency, and monetary).

Recency: How recently customers made their purchase.

Frequency: For simplicity, we’ll count the number of times each customer made a purchase.

Monetary: How much money they spent in total.

We are going to calculate these three key factors by grouping them by customers and taking “2020/12/30” as our reference end date since this is the last transaction date listed in our dataset.

# Frequency = count of invoice no. Of transaction(s)
# Monetary = Sum of Total amount for each customer
# Recency  = Overall latest invoice date - individual customer's last invoice date

# Set 2020/12/30 as the overall last invoice date. This is to calculate recency in days.

Latest_date = F.to_date(F.lit("2020/12/30"), 'yyyy/MM/dd')

# Calculate recency, frequency, and monetary numbers for each customer

Rfm_scores = (rtl_data.groupby("customerid")
               .agg((F.datediff(latest_date, F.max(F.col("invoicedate")))).alias("Recency"),
                    F.count("*").alias("Frequency"),
                    F.sum(F.col("totalamount")).alias("Monetary")).sort("customerid"))

Now explore the RFM values and see how they compare to one another.
5.5 RFM score calculation

Then, we’ll divide the customer data into three segments: recency, frequency, and economic value. We then get a combined segmentation score. We’re going to cluster the RFM score in 3 equal Categories:

<table>
<thead>
<tr>
<th>RFM_Score</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>3.00</td>
<td>34</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>4.00</td>
<td>102</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>5.00</td>
<td>210</td>
<td>4.8</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>350</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>7.00</td>
<td>526</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>8.00</td>
<td>631</td>
<td>14.4</td>
<td>14.4</td>
</tr>
<tr>
<td></td>
<td>9.00</td>
<td>669</td>
<td>15.3</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>10.00</td>
<td>630</td>
<td>14.4</td>
<td>14.4</td>
</tr>
<tr>
<td></td>
<td>11.00</td>
<td>524</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>12.00</td>
<td>350</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>13.00</td>
<td>209</td>
<td>4.8</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>14.00</td>
<td>105</td>
<td>2.4</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>15.00</td>
<td>32</td>
<td>.7</td>
<td>.7</td>
</tr>
<tr>
<td>Total</td>
<td>4372</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Segments = [0.33, 0.66]

Quantiles = rfm_numbers.approxquantile(["Recency","Frequency","Monetary"], Segments, 0)
#### Step 1

# Assign individual R,F,M scores to the three segments of customers

```
Rfm_scores = (rfm_numbers
    .withcolumn("R_Score",
        F.when(F.col("Recency") < quantiles[0][0], F.lit(1))
            .when(F.col("Recency") < quantiles[0][1], F.lit(2))
            .otherwise(F.lit(3)))
    .withcolumn("F_Score",
        F.when(F.col("Frequency") < quantiles[1][0], F.lit(3))
            .when(F.col("Frequency") < quantiles[1][1], F.lit(2))
            .otherwise(F.lit(1)))
    .withcolumn("M_Score",
        F.when(F.col("Monetary") < quantiles[2][0], F.lit(3))
            .when(F.col("Monetary") < quantiles[2][1], F.lit(2))
            .otherwise(F.lit(1))))
```

#### Step 2

# Calculate the combined RFM_Score

```
Rfm_agg_scores = (rfm_scores
    .withcolumn("RFM_Score", F.col("R_Score") + F.col("F_Score") + F.col("M_Score"))
)
5.6 Segmentation based on RFM Score

We have computed the customer’s various recency and frequency scores, as well as their aggregated RFM-score. For simplicity, we will divide our customer base into three segments: Platinum, Gold, and Silver. We will then assign a loyalty badge (Silver, Platinum, and Gold).

In the end, we created the following segmentation:

If the Score <= 8 then segmentation = Bronze
If the Score > 8 and <= 10 then segmentation = Silver  If the Score >10 then segmentation = Gold
Figure 13... segments

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bronze</td>
<td>1853</td>
<td>42.4</td>
</tr>
<tr>
<td>Gold</td>
<td>1220</td>
<td>27.9</td>
</tr>
<tr>
<td>Silver</td>
<td>1299</td>
<td>29.7</td>
</tr>
<tr>
<td>Total</td>
<td>4372</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Figure 14... Data Exploration

5.7 Segmentation description using decision tree.
We used the CHAID method for the decision tree algorithm, CHAID (Chi-Square Automatic Interaction Detector) analysis is a marketing segmentation technique. It is useful for identifying the relationships between categorical response variables. CHAID looks at how they might respond to a marketing campaign. It analyzes according to the attributes of each group.

### Model Summary

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Growing Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>CHAID</td>
<td>Segmentation</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>Recency, Gender, Country, Income, Age, Cluster Number of Case, Number of transactions</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Maximum Tree Depth</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Minimum Cases in Parent Node</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Minimum Cases in Child Node</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>Independent Variables Included</td>
<td>Age, Cluster Number of Case, Gender, Number of transactions, Income</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Number of Terminal Nodes</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Depth</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td></td>
</tr>
</tbody>
</table>
Figure 15... Decision tree
/* Node 1 */.
IF (Age IS MISSING OR (Age <= 38))
THEN
Node = 1
Prediction = 'Bronze'
Probability = 1.000000

/* Node 7 */.
IF (Age NOT MISSING AND (Age > 38 AND Age <= 43)) AND (Income NOT MISSIN
G AND (Income <= 60602))
THEN
Node = 7
Prediction = 'Bronze'
Probability = 0.933333

/* Node 8 */.
IF (Age NOT MISSING AND (Age > 38 AND Age <= 43)) AND (Income NOT MISSIN
G AND (Income > 60602 AND Income <= 74277))
THEN
Node = 8
Prediction = 'Silver'
Probability = 0.924138

/* Node 22 */.
IF (Age NOT MISSING AND (Age > 38 AND Age <= 43)) AND (Income IS MISSING OR (Income > 74277)) AND (Gender = "Male")
THEN
Node = 22
Prediction = 'Silver'
Probability = 0.942857

/* Node 23 */
IF (Age NOT MISSING AND (Age > 38 AND Age <= 43)) AND (Income IS MISSING OR (Income > 74277)) AND (Gender != "Male")
THEN
Node = 23
Prediction = 'Silver'
Probability = 0.989637

/* Node 10 */
IF (Age NOT MISSING AND (Age > 43 AND Age <= 48)) AND (Number of transactions NOT MISSING AND (Number of transactions <= 41))
THEN
Node = 10
Prediction = 'Silver'
Probability = 0.903061

/* Node 11 */
IF (Age NOT MISSING AND (Age > 43 AND Age <= 48)) AND (Number of transactions IS MISSING OR (Number of transactions > 41))
THEN
Node = 11
Prediction = 'Silver'
Probability = 0.990868
/* Node 12 */.
IF (Age NOT MISSING AND (Age > 48 AND Age <= 50)) AND (Number of transactions NOT MISSING AND (Number of transactions <= 41))
THEN
Node = 12
Prediction = 'Silver'
Probability = 0.852941

/* Node 13 */.
IF (Age NOT MISSING AND (Age > 48 AND Age <= 50)) AND (Number of transactions IS MISSING OR (Number of transactions > 41 AND Number of transactions <= 123))
THEN
Node = 13
Prediction = 'Gold'
Probability = 0.600000

/* Node 14 */.
IF (Age NOT MISSING AND (Age > 48 AND Age <= 50)) AND (Number of transactions NOT MISSING AND (Number of transactions > 123 AND Number of transactions <= 212))
THEN
Node = 14
Prediction = 'Gold'
Probability = 0.787879
/* Node 15 */.
IF (Age NOT MISSING AND (Age > 48 AND Age <= 50)) AND (Number of transactions NOT MISSING AND (Number of transactions > 212))
THEN
Node = 15
Prediction = 'Gold'
Probability = 1.000000

/* Node 16 */.
IF (Age NOT MISSING AND (Age > 50 AND Age <= 55)) AND (Income NOT MISSING AND (Income <= 68232))
THEN
Node = 16
Prediction = 'Silver'
Probability = 0.950000

/* Node 24 */.
IF (Age NOT MISSING AND (Age > 50 AND Age <= 55)) AND (Income IS MISSING OR (Income > 68232)) AND (Number of transactions NOT MISSING AND (Number of transactions <= 41))
THEN
Node = 24
Prediction = 'Silver'
Probability = 0.614679

/* Node 25 */.
IF (Age NOT MISSING AND (Age > 50 AND Age <= 55)) AND (Income IS MISSING OR (Income > 68232)) AND (Number of transactions IS MISSING OR (Number of transactions > 41 AND Number of transactions <= 123))
THEN
Node = 25
Prediction = 'Gold'
Probability = 0.758824

/* Node 26 */.
IF (Age NOT MISSING AND (Age > 50 AND Age <= 55)) AND (Income IS MISSING OR (Income > 68232)) AND (Number of transactions NOT MISSING AND (Number of transactions > 123 AND Number of transactions <= 212))
THEN
Node = 26
Prediction = 'Gold'
Probability = 0.904762

/* Node 27 */.
IF (Age NOT MISSING AND (Age > 50 AND Age <= 55)) AND (Income IS MISSING OR (Income > 68232)) AND (Number of transactions NOT MISSING AND (Number of transactions > 212))
THEN
Node = 27
Prediction = 'Gold'
Probability = 0.996016

/* Node 18 */.
IF (Age NOT MISSING AND (Age > 55)) AND (Number of transactions NOT MISSING AND (Number of transactions <= 41))
THEN
Node = 18
Prediction = 'Silver'
Probability = 0.887324

/* Node 19 */.
IF (Age NOT MISSING AND (Age > 55)) AND (Number of transactions IS MISSING OR (Number of transactions > 41 AND Number of transactions <= 123))
THEN
Node = 19
Prediction = 'Gold'
Probability = 0.576687

/* Node 20 */.
IF (Age NOT MISSING AND (Age > 55)) AND (Number of transactions NOT MISSING AND (Number of transactions > 123 AND Number of transactions <= 212))
THEN
Node = 20
Prediction = 'Gold'
Probability = 0.757143

/* Node 21 */.
IF (Age NOT MISSING AND (Age > 55)) AND (Number of transactions NOT MISSING AND (Number of transactions > 212))
THEN
Node = 21
Prediction = 'Gold'
Probability = 0.983193

5.8 K-MEANS SEGMENTATION

We did a K-means using the following variables and we get 5 segments, and below the table showing the assignment of the client to each segment: Online retail data is a great dataset to store and analyze customer data. It allows you to divide the customers into groups.

The data includes the following attributes:
1. Customer ID based

2. Customer Gender

3. Customer Age

4. Income of the customer (in Thousand Dollars)

5. Spending score of the customer (based on customer behavior and spending nature)

### 5.9 Python K-means program

Here we will use the libraries, Matplotlib to create visualizations to perform calculation.

1. Converts the customers’ data frame into K-means data. The model and features are dropped so the customer columns are all on the left. The data frame is transposed to have the customers as rows, and models as columns. The K-means function requires this format.

2. Performs the K-means function to cluster the customer segments. We set min cluster = 4 and max cluster = 8. From our hypothesis, we expect there to be at least four and at most six groups of customers.

3. Uses the silhouette function to obtain silhouette widths. Silhouette is a technique in clustering that validates the best cluster groups. The silhouette function from the cluster package allows us to get the average width of silhouettes, which will be used to programmatically determine the optimal cluster size.

Steps 1 and 2 of K-means were about choosing the number of clusters (k) and selecting random centroids for each cluster. We will pick 3 clusters and then select random observations from the data as the centroids:
# Step 1 and 2 - Choose the number of clusters (k) and select random centroid for each cluster

```python
#number of clusters
K=3

# Select random observation as centroids
Centroids = (X.sample(n=K))
```

```python
Plt.scatter(X['annalincome'],X['totaltrnxamount'],c='black')
Plt.scatter(Centroids['annalincome'],Centroids['totaltrnxamount'],c='red')
Plt.xlabel('annualincome')
Plt.ylabel('Income Amount (In Thousands)')
Plt.show()
```

# Step 3 - Assign all the points to the closest cluster centroid

# Step 4 - Recompute centroids of newly formed clusters

# Step 5 - Repeat step 3 and 4

```python
Diff = 1
J=0
```
While\( (\text{diff} \neq 0) \):

\[ XD = X \]
I=1

For index1,row_c in Centroids.iterrows():

    ED=[]

    For index2,row_d in XD.iterrows():

        D1=(row_c["atincome"]-row_d["aincome"])**2
        D2=(row_c["trxamount"]-row_d["trxamount"])**2
D=np.sqrt(d1+d2)

ED.append(d)

X[i]=ED

I=i+1

C=[]

For index,row in X.iterrows():
    Min_dist=row[1]
    Pos=1

    For i in range(K):
        If row[i+1] < min_dist:
            Min_dist = row[i+1]
            Pos=i+1
            C.append(pos)

X["Cluster"] = C

Centroids_new = X.groupby("Cluster").mean()["trxamount","aincome"]

If j == 0:
    Diff=1
    J=j+1
Else:
\[
\text{Diff} = (\text{Centroids\_new['trxamount']} - \text{Centroids['trxamount']}).\text{sum()}
\]

\[
(\text{Centroids\_new['trxamount']} - \text{Centroids['aincome']}).\text{sum()}
\]

\[+\]
Print(diff.sum())
Centroids = X.groupby(['Cluster']).mean()['trxamount','aincome']

5.10 Data
<table>
<thead>
<tr>
<th>CustomerID</th>
<th>Transaction_count</th>
<th>Amount</th>
<th>Gender</th>
<th>Country</th>
<th>Income</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18287</td>
<td>70</td>
<td>104.55</td>
<td>1.00</td>
<td>5.00</td>
<td>$36,781</td>
</tr>
<tr>
<td>2</td>
<td>18283</td>
<td>756</td>
<td>1220.93</td>
<td>1.00</td>
<td>7.00</td>
<td>$94,639</td>
</tr>
<tr>
<td>3</td>
<td>18282</td>
<td>13</td>
<td>62.68</td>
<td>.00</td>
<td>5.00</td>
<td>$101,768</td>
</tr>
<tr>
<td>4</td>
<td>18281</td>
<td>7</td>
<td>39.36</td>
<td>1.00</td>
<td>5.00</td>
<td>$92,304</td>
</tr>
<tr>
<td>5</td>
<td>18280</td>
<td>10</td>
<td>47.65</td>
<td>.00</td>
<td>9.00</td>
<td>$59,012</td>
</tr>
<tr>
<td>6</td>
<td>18278</td>
<td>9</td>
<td>29.55</td>
<td>1.00</td>
<td>4.00</td>
<td>$30,288</td>
</tr>
<tr>
<td>7</td>
<td>18277</td>
<td>9</td>
<td>37.88</td>
<td>1.00</td>
<td>5.00</td>
<td>$31,716</td>
</tr>
<tr>
<td>8</td>
<td>18276</td>
<td>16</td>
<td>47.25</td>
<td>1.00</td>
<td>9.00</td>
<td>$86,975</td>
</tr>
<tr>
<td>9</td>
<td>18274</td>
<td>22</td>
<td>80.78</td>
<td>1.00</td>
<td>1.00</td>
<td>$71,798</td>
</tr>
<tr>
<td>10</td>
<td>18273</td>
<td>3</td>
<td>7.65</td>
<td>1.00</td>
<td>5.00</td>
<td>$50,035</td>
</tr>
<tr>
<td>11</td>
<td>18272</td>
<td>170</td>
<td>391.16</td>
<td>1.00</td>
<td>4.00</td>
<td>$81,815</td>
</tr>
<tr>
<td>12</td>
<td>18270</td>
<td>13</td>
<td>59.10</td>
<td>1.00</td>
<td>5.00</td>
<td>$91,037</td>
</tr>
<tr>
<td>13</td>
<td>18269</td>
<td>8</td>
<td>27.10</td>
<td>1.00</td>
<td>10.00</td>
<td>$58,527</td>
</tr>
<tr>
<td>14</td>
<td>18268</td>
<td>2</td>
<td>25.50</td>
<td>1.00</td>
<td>6.00</td>
<td>$33,868</td>
</tr>
<tr>
<td>15</td>
<td>18265</td>
<td>46</td>
<td>241.34</td>
<td>1.00</td>
<td>3.00</td>
<td>$92,924</td>
</tr>
<tr>
<td>16</td>
<td>18263</td>
<td>62</td>
<td>60.61</td>
<td>1.00</td>
<td>4.00</td>
<td>$33,555</td>
</tr>
<tr>
<td>17</td>
<td>18262</td>
<td>13</td>
<td>14.60</td>
<td>1.00</td>
<td>5.00</td>
<td>$39,639</td>
</tr>
<tr>
<td>18</td>
<td>18261</td>
<td>21</td>
<td>64.74</td>
<td>1.00</td>
<td>6.00</td>
<td>$44,386</td>
</tr>
</tbody>
</table>

Figure 16... Data

5.11 Result of K-means
<table>
<thead>
<tr>
<th>CustomerID</th>
<th>KmeansSegmentation</th>
<th>Distancetoeachcluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>18287</td>
<td>4</td>
<td>4631.70260</td>
</tr>
<tr>
<td>18283</td>
<td>3</td>
<td>5060.95603</td>
</tr>
<tr>
<td>18282</td>
<td>3</td>
<td>12093.68665</td>
</tr>
<tr>
<td>18281</td>
<td>3</td>
<td>2654.83178</td>
</tr>
<tr>
<td>18280</td>
<td>5</td>
<td>9652.72310</td>
</tr>
<tr>
<td>18278</td>
<td>4</td>
<td>11124.32863</td>
</tr>
<tr>
<td>18277</td>
<td>4</td>
<td>9696.31002</td>
</tr>
<tr>
<td>18276</td>
<td>3</td>
<td>2733.97934</td>
</tr>
<tr>
<td>18274</td>
<td>5</td>
<td>3139.59740</td>
</tr>
<tr>
<td>18273</td>
<td>4</td>
<td>8622.93290</td>
</tr>
<tr>
<td>18272</td>
<td>3</td>
<td>7865.48056</td>
</tr>
<tr>
<td>18270</td>
<td>3</td>
<td>1410.06453</td>
</tr>
<tr>
<td>18269</td>
<td>5</td>
<td>10138.03995</td>
</tr>
<tr>
<td>18268</td>
<td>4</td>
<td>7544.39358</td>
</tr>
<tr>
<td>18265</td>
<td>3</td>
<td>3249.98823</td>
</tr>
<tr>
<td>18263</td>
<td>4</td>
<td>5857.40653</td>
</tr>
<tr>
<td>18262</td>
<td>4</td>
<td>1773.96825</td>
</tr>
<tr>
<td>18261</td>
<td>4</td>
<td>2973.73608</td>
</tr>
<tr>
<td>18260</td>
<td>3</td>
<td>12250.02989</td>
</tr>
<tr>
<td>18259</td>
<td>5</td>
<td>5835.66928</td>
</tr>
<tr>
<td>18257</td>
<td>5</td>
<td>3982.30601</td>
</tr>
<tr>
<td>18256</td>
<td>5</td>
<td>5871.72119</td>
</tr>
</tbody>
</table>

*Figure 17... Result of K-mean*
Figure 18... Data Exploration 1.

<table>
<thead>
<tr>
<th>Cluster Number of Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Valid 1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Figure 19 ... Data Exploration 2
The chart shows that the Platinum customers are more likely to shop more frequently than the Silver customers. It also shows that the Gold customers are more active than the Silver customers and the findings indicate that we need to think about ways to improve the satisfaction of our Platinum customers while at the same time, increasing the engagement of our Gold customers.

Figure 20... Relationship Map
Chapter 6 Conclusion.

6.1 Conclusion.

In order to attract and keep customers; organizations need to develop effective marketing strategies that cater to the varying needs of their customers. This process is known as segmentation. This paper aims to provide a comprehensive review on the RFM model and its various applications. It first, talks about the model's definition and scoring scheme, then discusses the various advantages and disadvantages of RFM model, and finally, it talks about its extended RFM model. The review on RFM model is very useful for researchers and decision makers. It can provide them with valuable insight into the various facets of the RFM model and its effectiveness in different industries.

RFM is a model that helps decision makers identify valuable customers. It can also be used to develop a strategy and implement it in their organization. Through the review of the RFM model, decision makers would be able to identify the areas of their operation where they can improve their efficiency and develop a strategy to satisfy their customers' needs.

6.2 Recommendations.

It is generally recommended that marketers carry out market segmentation based on the combination of various factors, such as customer characteristics, operational characteristics, and personality characteristics. RFM model is a widely used marketing strategy that can help organizations identify and develop effective marketing strategies for different industries. It can also help them identify and develop valuable customers. Through the review of the RFM model, decision-makers would be able to identify the various factors that influence the operation of RFM and develop effective strategies to address these issues.
References


