Customer Segmentation using RFM Analysis and K-Means Clustering

By

Chandana Rajanna

Table of Contents

List of Figures

List of Tables

Abstract

This report analyzes customers' transaction data of an online retail store to gain insights about customers' purchase behavior, in order to help the business, make effective decisions about inventory management, monitoring dynamic shopping trends and channel marketing resources in the correct direction. RFM (Recency, Frequency, and Monetary values) analysis was used for initial analysis and aggregation. Using K-Means unsupervised learning algorithm, four clusters, each with about 20-30% of the total number of customers were identified with distinct purchase behavior and marketing needs.

Introduction

The E-Commerce's customer database contains transactional data over a period of one year. This transaction data mainly contains unique all-occasion gifts. Each individual has different needs and desires, and it is very important to able to identify and satisfy the needs of different customer groups for a business to be profitable. Clustering customers based on their past purchase behavior using RFM analysis (Recency, Frequency, and Monetary values) where customer data were segmented based on three metrics i.e. Recency, Frequency, and Monetary values into different groups such as best customers, loyal customers, big spenders, and lost customers. Later, using R, F, M values K-Means unsupervised learning algorithm was implemented to understand underlaying patterns in purchase behavior.

Overview

To analyze purchase behavior of customers during online shopping and determine valuable customers to the business. In order to extract information, and gain insights about shoppers' behavior, analysis is done with Recency, Frequency, Monetary (RFM) methodology and then an unsupervised learning algorithm K-Means is applied to create different customer segments.

Solution Methodology

The analysis approach followed to create customer segments using the online shopping data is summarized in the flow chart below.

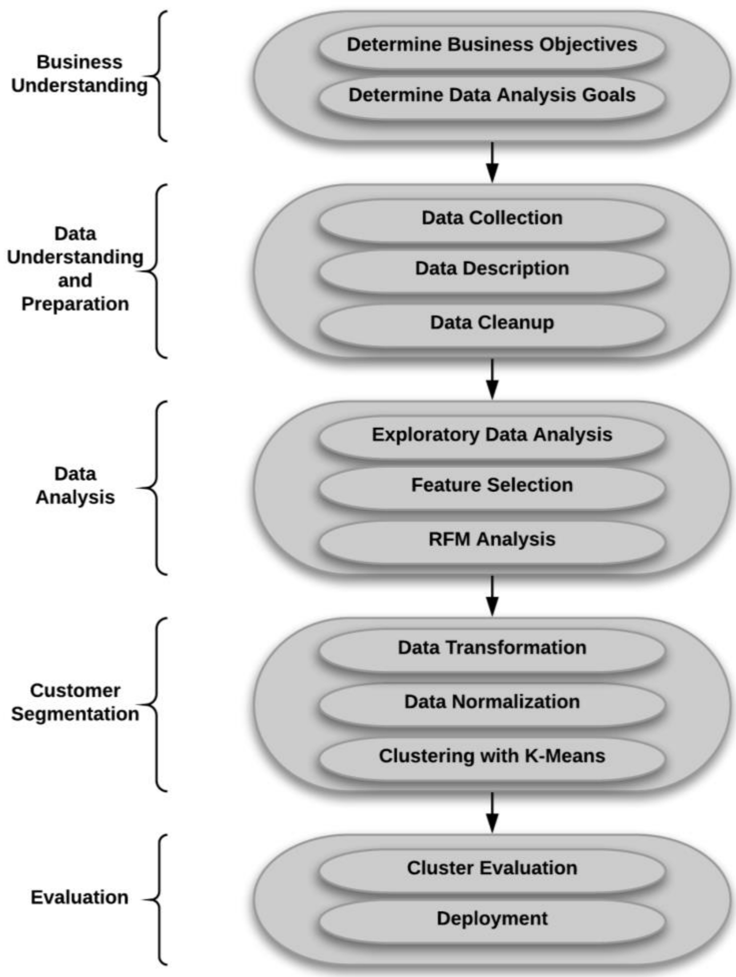

Figure 1: Solution Methodology flow

The first step is to understand the business requirements which in this case is to create customer segmentation to help direct marketing resources and maximize customer growth and retention. The goal was to perform data analysis, explore different data mining options and chose the best suited method to perform customer segmentation.

The next step is to collect a large amount of shopping data with records such as transaction ID, type, units sold, cost per unit, date of transaction etc. all associated with a unique customer ID since the overall goal is to perform customer segmentation using their past shopping behavior. The shopping data which satisfies these parameters was obtained through Kaggle (Carrie. 2017, August 17). This data is described in *Table 1*. Further, the data was cleaned up to eliminate illegal values which might have been a result of bad data entry.

With the clean data, a number of exploratory data analysis was performed to better understand the data, its spread and dimensionality. The appropriate features (columns) of the data were selected based on their usefulness to the analysis and the features which do not have sufficient variance were discarded for the analysis. After considering a number of different data analysis algorithms, RFM (Recency, Frequency, Monetary) analysis (RFM Analysis, Model, Marketing & Software. (n.d.). was chosen since it was the best suited to fully utilize the available data dimensions.

With the generated RFM data per customer, an unsupervised learning algorithm called K-Means was applied to create 4 different customer segments. The RFM data obtained earlier was further transformed and normalized to make it suitable for K-Means clustering K, D. (2020, January 29). The optimum number of clusters was identified using Elbow Method K, D. (2020, January 29).

The customer segments were further analyzed to draw insights into the shopping behavior of the customers who fall into the different categories and suitable marketing recommendations are suggested to maximize customer growth and retention.

Data Description (Metadata)

The database chosen for analysis has 541909 rows and 8 columns. The data set contains five numerical columns namely InvoiceNo, StockCode, Quantity and UnitPrice, two categorical columns namely description and country, and one timestamp column of invoice.

Data Column	**Type**	**Description**
InvoiceNo	object	A 6-digit unique number given to each transaction. Letter ‘c’ is the first letter if it’s a cancelled order
StockCode	object	A 5-digit unique number to identify distinct product
Description	object	Gives product name
Quantity	int64	The quantities of each product per transaction
InvoiceDate	datetime64[ns]	The date and time of each transaction
UnitPrice	float64	Product’s price per unit
CustomerID	object	A 5-digit unique number assigned to each customer
Country	object	customers’ residency and order placed location

Table 1: Data Description

Data Cleaning / Data Preparation

Data cleaning and its preparation is the initial and essential step to perform before starting to analyze or apply data mining methodologies. The cleaning process starts with observation on data spread, checking for data types and converting variables to appropriate data types. *(Refer Appendix A)*

- The first step of data preparation is to check for missing values or NAs in the data frame. A total of 135080 missing values in "Description" and "CustomerID" columns were found. On further investigation, it was found that all rows where "Description" was missing also had a null CustomerID. Since "CustomerID" is essential to our customer segment analysis, it would not be meaningful to keep missing values. The percentage of rows with null values is small compared to whole dataset. So, it is prudent to delete these rows for the purpose of this project's data mining methods. *(Refer Appendix A)*
- Next step is to check for duplicate entries into the database if any. Looking at the duplicate rows, it seems that unique ids are repeated which can be attributed to error in data entry. So, only the first copy was preserved and the duplicates were eliminated from data frame. *(Refer Appendix A)*
- During our initial data description, observed some cancelled orders so on further investigation on cancelled orders. Found 8872 cancelled invoice and 8872 negative quantity. Further, checked for two scenarios 1. a cancel order exists without counterpart and 2. There's at least one counterpart with the exact same quantity.
  The index of the corresponding cancel order is respectively kept and doubtful entries were removed. *(Refer Appendix A)*
- Adding column "AmountSpent" and parsing "InvoiceDate" column to day, yearmonth, and date columns for the purpose of segmentation. *(Refer Appendix A)*

Exploratory Data Analysis

Exploratory data analysis is important process in initial investigations on data to discover patterns, identify anomalies or to check assumptions with summary statistics and visualizations Patil, P. (2018, May 23). *(Refer Appendix A)*

Summary

	Quantity	UnitPrice	AmountSpent
Count	392732	392732	392732
Mean	13.153718	3.125596	22.629195
Std	181.58842	22.240725	311.083465
min	1	0	0
25%	2	1.25	4.95
50%	6	1.95	12.39
75%	12	3.75	19.8
max	80995	8142.75	168469.6

Table 2: Data summary

Summary table depicts the overall picture of the dataset which indicates minimum value and maximum value in the table shows the low and high values in that column. From summary table, for example we see that there is a wide spread in unit price with standard deviation of 22.24, the 1.95 is the median, and mean of 3.12. *(Refer Appendix A)*

Correlation

Correlation measures the strength of the variables, pointing the linear relationship between the variables. As per the correlation graph shown below. Pink color represents positive correlation and black color represents the negative correlation between variables. The correlation analyses the data with respect to p values. Strong/Positive correlation is found between variables Quantity and AmountSpent column. Negative correlation is high with all other variables. *(Refer Appendix A)*

Figure 2: Correlation Matrix

Geographic

Plotting customer data with country, it is evident that United Kingdom has highest number of customers and. Number of sales. (*Refer Appendix A*)

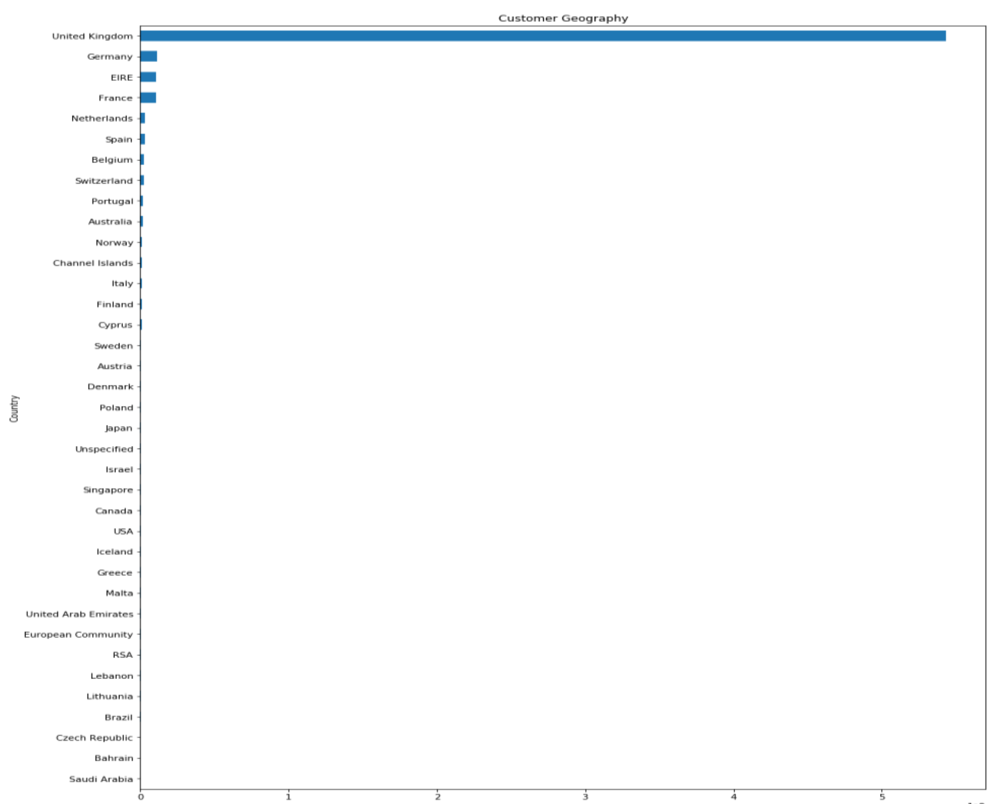

Figure 3: Customer geography

Customers and products

The dataframe contains around 400,000 entries. The unique number of transactions, cutomerIds, and products are as seen below

	Products	Customers	Transactions
Count	3665	4339	18536

Table 3: Unique counts

It can be seen that data is of 4339 users and that they bought 3665 unique items. The total number of transactions carried out is of the order of 18536 during a period of one year. *(Refer Appendix A)*

Orders and Month

The plot of orders vs month shows that the orders are evenly spread across the months except for holiday season around October and November which is generally expected. *(Refer Appendix A)*

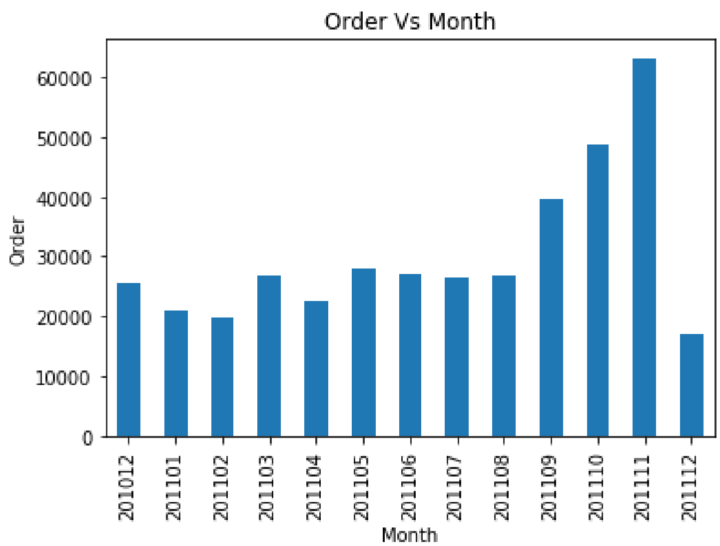

Figure 4: Order Vs Invoice Month

Revenue and Month

The plot of revenue vs month follows the orders plot closely which is expected since revenue generated is a factor of orders received. From the plot it is seen that Q4 has high revenue comparatively. *(Refer Appendix A)*

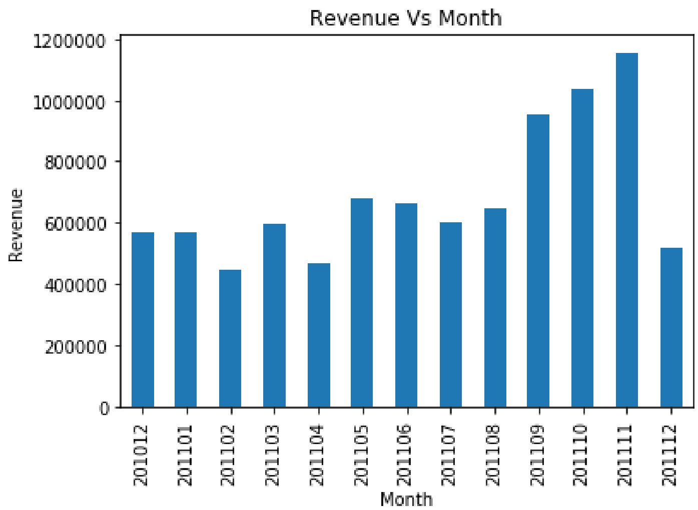

Figure 5: Revenue Vs Invoice Month

Product sales exploration

Top items ordered on the online retail

Analyzing the sales of products and its frequency, we can see the most frequently ordered items here are T-light holder, Cake stand, and Jumbo bag red retrospot, as seen in Table 4. Also, table 5 shows the top items ordered which generated the most revenue. It can be seen that there are a few similar items between the two tables, such as T-light holder, Cake stand and Jumbo bag red retrospot. This information can be used in effective inventory management. *(Refer Appendix A)*

	StockCode	Description	UnitPrice	Quantity
0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	2.95	1640
1	22423	REGENCY CAKESTAND 3 TIER	12.75	1392
2	84879	ASSORTED COLOUR BIRD ORNAMENT	1.69	1341
3	20725	LUNCH BAG RED RETROSPOT	1.65	1229
4	47566	PARTY BUNTING	4.95	1199
5	22720	SET OF 3 CAKE TINS PANTRY DESIGN	4.95	1082
6	85099B	JUMBO BAG RED RETROSPOT	2.08	1075
7	20727	LUNCH BAG BLACK SKULL.	1.65	1044
8	23298	SPOTTY BUNTING	4.95	972

Table 4: Top ordered items by quantity

	StockCode	Description	Revenue
0	23843	PAPER CRAFT , LITTLE BIRDIE	168469.60
1	22423	REGENCY CAKESTAND 3 TIER	142264.75
2	85123A	WHITE HANGING HEART T-LIGHT HOLDER	100392.10
3	85099B	JUMBO BAG RED RETROSPOT	85040.54
4	23166	MEDIUM CERAMIC TOP STORAGE JAR	81416.73
5	POST	POSTAGE	77803.96
6	47566	PARTY BUNTING	68785.23
7	84879	ASSORTED COLOUR BIRD ORNAMENT	56413.03
8	M	Manual	53419.93

Table 5: Top ordered items by revenue

Top items customers reordered

After learning about most sold products and high revenue generating items by the online retail, further analysis was done to understand what items are being reordered by a customer. This indicates that these items are well liked by customers and have come back to buy again. The table below also correlates well with above two tables, and no anomalies are observed. *(Refer Appendix A)*

Description	reorder
WHITE HANGING HEART T-LIGHT HOLDER	1160
JUMBO BAG RED RETROSPOT	980
REGENCY CAKESTAND 3 TIER	833
LUNCH BAG RED RETROSPOT	772
POSTAGE	768
ASSORTED COLOUR BIRD ORNAMENT	717
PARTY BUNTING	682
LUNCH BAG BLACK SKULL.	620
LUNCH BAG SUKI DESIGN	603
SET OF 3 CAKE TINS PANTRY DESIGN	512

Table 6: Top reordered items

First buy and reorder comparison

It is useful to know to if there any underlying shopping patterns between first buy and reorder based on the month ordered. It is seen that in the beginning of the year, there are far more first orders than reorders probably because the data is collected only from Dec 2010. It can be observed that the ratio between first order and reorders almost evens out in the later months. The box plot below represents the dispersion of the first buy and reorder data. There is a difference between first buy's mean and reorder's mean and an outlier is seen for reorder record during November of 2011, where not only is the re-order revenue greater than the first order revenue but is also greater than the re-order revenue of other months. *(Refer Appendix A)*

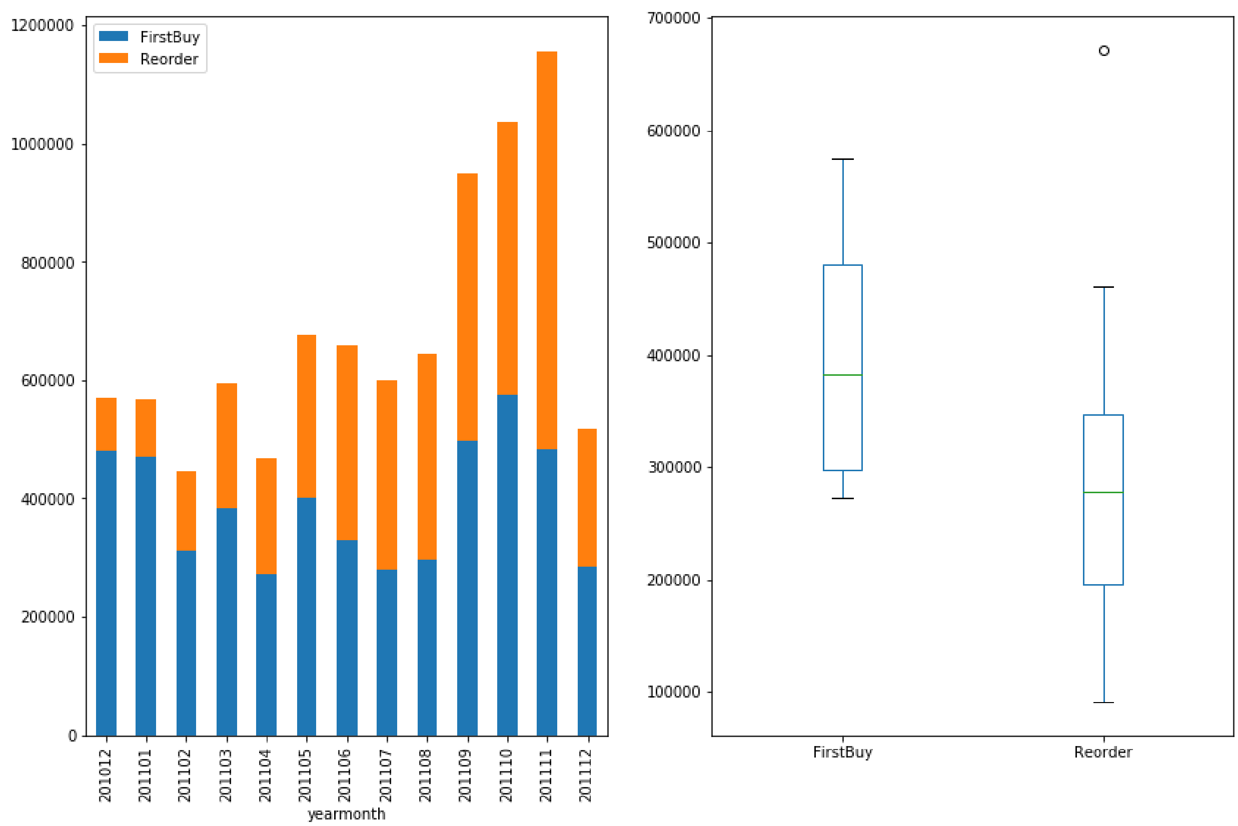

Table 7: First buy and Reorder comparison

Frequency of products

This plot represents the wordcloud of products in the data set. It is a visualization of word frequency that gives greater importance to word that appear more frequently under product column of the database. From the plot below, it can be observed that the items such as Jumbo bag red retrospot, T-light holder, and Lunch bag are most frequently appeared in the database, these items were also evidenced in the top ordered and reordered items above which confirms that they were ordered most frequently. (Henderson, S.,et T. 2015, March 3) *(Refer Appendix A)*

Figure 6: Wordcloud of Products sold

Clustering Methodologies

After exploring the data and having learned the dimensionality and variance of the dataset, the next step was to identify suitable classification algorithm which can handle all the datatypes constituting the online retail database.

After researching few clustering algorithms such as Linear classifiers, k-nearest neighbor, and Random forests, it was observed that none work well with categorical data such as StockCode and CustomerId. Since the categorical variable are essential in the classification of customers based on their purchase behavior, RFM analysis was found to be most appropriate to apply on the dataset.

RFM analysis

RFM analysis is a method of classifying the customers based on three key metrics namely Recency, Frequency, and Monetary. This classification helps to identify the top 20% or most valuable customers to the business.

Recency (R): Time between now and last purchase

Frequency (F): Number of purchases

Monetary Value (M): Total amount spent

Procedure

- In the First step, customers divided into different groups according to the distribution of values for Recency, Frequency, and Monetary values *(Refer Appendix B)*
    - Recency: To calculate the time between now and last purchase
        - A date point is chosen from which days are counted to find the customer's last purchase.
        - A new dataframe is formed with the original dataframe grouped by CustomerID and a column is created to contain the date of their last purchase.
        - A Recency column was added by calculated time since the last purchase of a customer and the chosen date point.

- Frequency: To calculate the total number of purchases by a customer
    - A new dataframe is formed with the original dataframe grouped by CustomerID and Invoice count is created to contain the total number of orders.
    - A Frequency column was added by counting total number of orders placed by a customer

- Monetary: To calculate the total spending by customers
    - A new dataframe is formed with the original dataframe grouped by CustomerID and Amount spent is created to contain the total amount spent by customers
    - A Monetary column was added by counting total amount spent by a customer

- Created a new table containing RFM values, by merging Recency, Frequency, and Monetary values. *(Refer Appendix B)*

- Next step, is to split metrices into segments by using quartiles (0.25, 0.5, 0.75). A score from 1 to 4 is assigned to Recency, Frequency and Monetary. Four is the best/highest value, and one is the lowest/worst value. Created two segmentation classes since, high recency is bad, while high frequency and monetary value is good. *(Refer Appendix B)*

- Finally, A final RFM score is calculated simply by combining individual RFM score numbers as seen in Table 8. *(Refer Appendix B)*
  Best customers are the ones having RFM score of 444.
  Best Recency score = 4: most recently purchase.
  Best Frequency score = 4: most quantity purchase.
  Best Monetary score = 4: spent the most.

	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore
1690	14646.0	2	2080	280206.02	4	4	4	444
4202	18102.0	1	431	259657.30	4	4	4	444
3729	17450.0	9	336	194390.79	4	4	4	444
1880	14911.0	2	5672	143711.17	4	4	4	444
1334	14156.0	10	1395	117210.08	4	4	4	444
3772	17511.0	3	963	91062.38	4	4	4	444
3177	16684.0	5	277	66653.56	4	4	4	444
1290	14096.0	5	5111	65164.79	4	4	4	444
997	13694.0	4	568	65039.62	4	4	4	444
2177	15311.0	1	2366	60632.75	4	4	4	444

Table 8: Top 10 champions (best customers)

Customers in each segment

Here is the visualization of all the six types of customers from RFM analysis. From this figure, it is seen that Loyal customers who buy most frequently and Big spenders who spend the most are high in numbers. The segments which contains customers who are almost lost or lost to the business are lesser comparatively. The customers in lost cheap segment are the ones who purchased few items, spent little, and whose last purchase was very long ago, and this is the segment that the company should not spend much resources to reacquire. *(Refer Appendix B)*

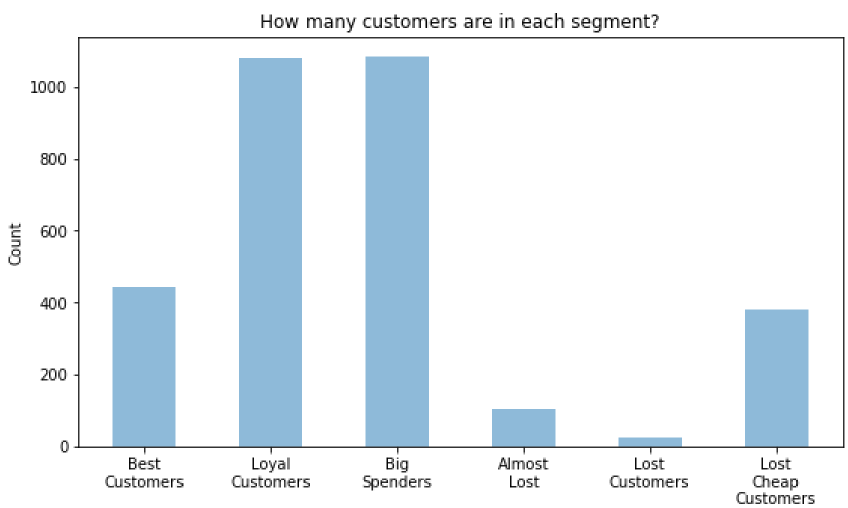

Figure 7: Number of customers in each segment

Customers distribution on RFM score counts

This Venn diagram represents the number of customers with high quartile scores in each metric of Recency, Frequency, and Monetary value. Intersection of all three circles of RFM shows the count of customers who are having high score of 4 in all three metrics i.e. Recency, Frequency, and Monetary value. It is seen from the Venn diagram that the number of customers is high at Recency only, intersection of Recency and Frequency, and at the intersection of Recency, Frequency and Monetary values. *(Refer Appendix B)*

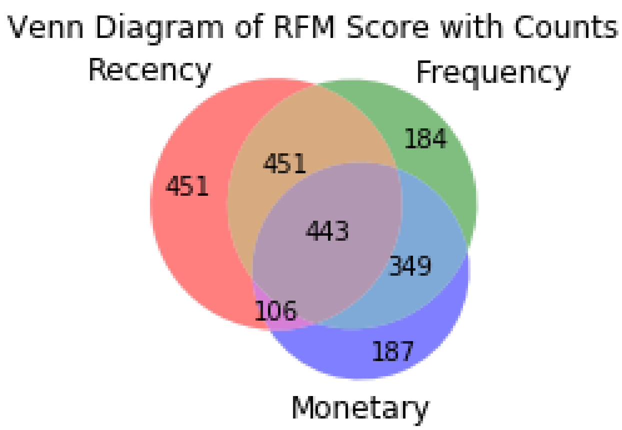

Figure 8: Venn diagram of RFM score

Customer Segmentation using unsupervised learning

To discover unknown patterns and find more insights about customers purchase behavior, an unsupervised learning approach was followed. After exploring few unsupervised clustering algorithms such as Hierarchical clustering, Density-based spatial clustering of applications with noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS), and K-Means clustering, it was found that they don't work well with categorical or hybrid (combination of categorical and numerical) data set. Since the RFM data per customer is fully numerical data and is also a good representation of the customer's shopping behavior, the RFM data was used to perform unsupervised learning for customer segmentation. Among the various algorithms considered here, K-Means clustering was chosen as it was the best suited for the available dataset since it is simple, fast and works well of large datasets.

K-Means clustering

K-Means clustering algorithm works iteratively to create a pre-defined (k) number of clusters within the given dataset with the following objectives (Pallu, P. (n.d.)):

- Each data-point belongs to only one cluster.
- Each data-point within a cluster is as similar to the other data-points within the cluster as possible.
- Each data-point in one cluster is as dissimilar to a data-point from another cluster as possible.

The algorithm works iteratively using the following steps: (Aggelis, et. (2005)).

1. Select k points randomly in the data set. These are the initial centroids of the k clusters.
2. Assign every other data point to one of these clusters based on their shortest Euclidian distance to the centroids.
3. Re-calculate the centroids for the clusters formed in step 2 by calculating the average of all the data points belonging to a particular cluster.
4. If the cluster centroids have changed, then repeat steps 2 and 3. If the centroids have not changed, then the clustering is complete and each data-point has been assigned to its final cluster. (Refer Appendix C)

Since K-Means clustering works based on distances between various data points, it is essential to transform the data if necessary, to get a symmetrical distribution across the range of values each variable can take and to normalize the data, i.e. the mean should be 0 and standard deviation should be 1. (Pallu, P. (n.d.)).

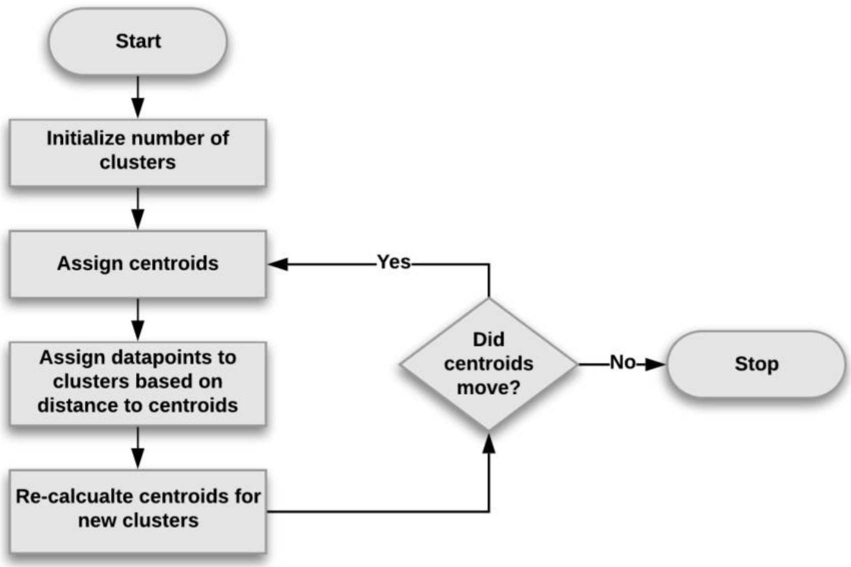

Figure 9:K-Means Generic Algorithm (Pallu, P. (n.d.))

Procedure

Data preprocessing

Data processing is a technique essential for transforming the raw data into useful and efficient format. Since Data cleaning was already performed in the initial analysis above. Here, the focus is on Data transformation, checking for normalization, data reduction, and dimensionality reduction.

There are few ways to remove or reduce skewness from the data.

1. Logarithmic transformation: Useful only with positive values.

2. Z-Transformation: A widely used method which works well on the mixed (negative and positive) values.

Identify the type of data and statistics of the RFM table:

	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile
count	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000
mean	15299.936852	93.041484	90.512100	2048.215924	2.506107	2.487670	2.499885
std	1721.889758	100.007757	225.515328	8984.248352	1.122159	1.122724	1.118266
min	12346.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000
25%	13812.500000	18.000000	17.000000	306.455000	1.500000	1.000000	1.500000
50%	15299.000000	51.000000	41.000000	668.560000	3.000000	2.000000	2.000000
75%	16778.500000	142.500000	98.000000	1660.315000	4.000000	3.000000	3.500000
max	18287.000000	374.000000	7676.000000	280206.020000	4.000000	4.000000	4.000000

Table 9: Output summary of RFM

Next step is to identify skewness in the data set. Skewness is a measure of the asymmetry of the probability distribution of data around its mean. The distribution is symmetric if it is evenly spread on left side and right side from center point. (Skewness. (2020, April 14). (Refer Appendix C)

The distribution plot is one of the effective graphical technique to represent skewness. So, distribution plot of Recency, Frequency, and, Monetary is plotted. ((n.d.). Retrieved) (Refer Appendix C)

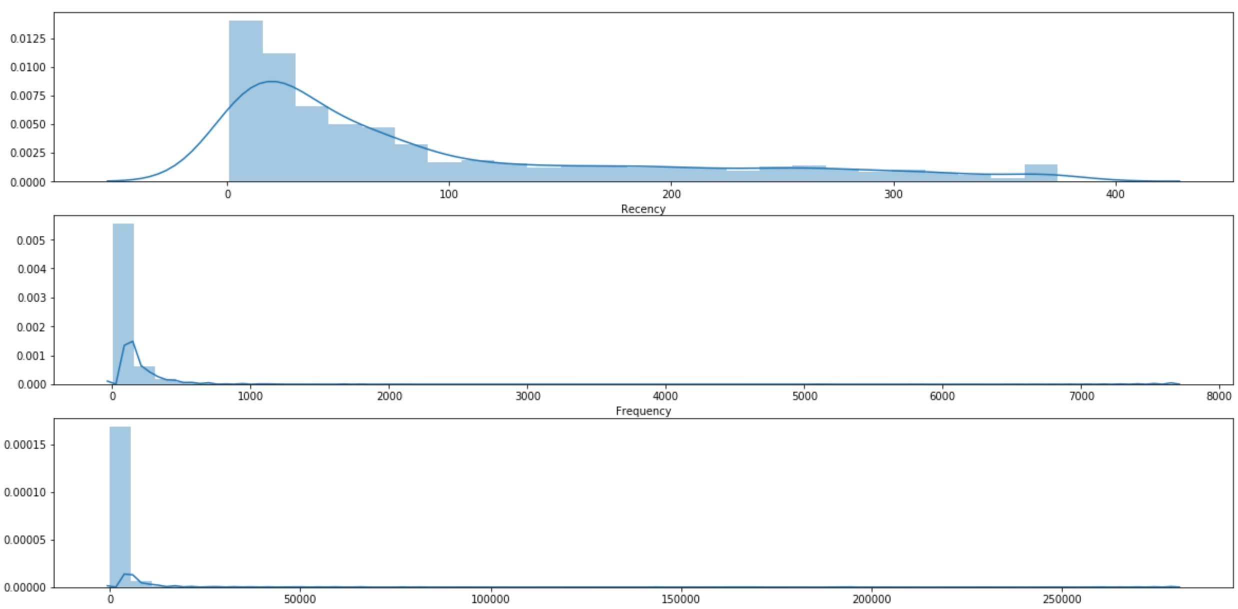

Figure 10: Distribution plot of R, F, M values before transformation

From the distribution plot above, it can be seen that all metrics in RFM data are right skewed. To reduce the skewness in the data, Log Transformation or Z-Transformation can be used, but for this data Log Transformation is performed because there are no negative values. (Refer Appendix C)

After Log Transformation distribution plot was plotted.

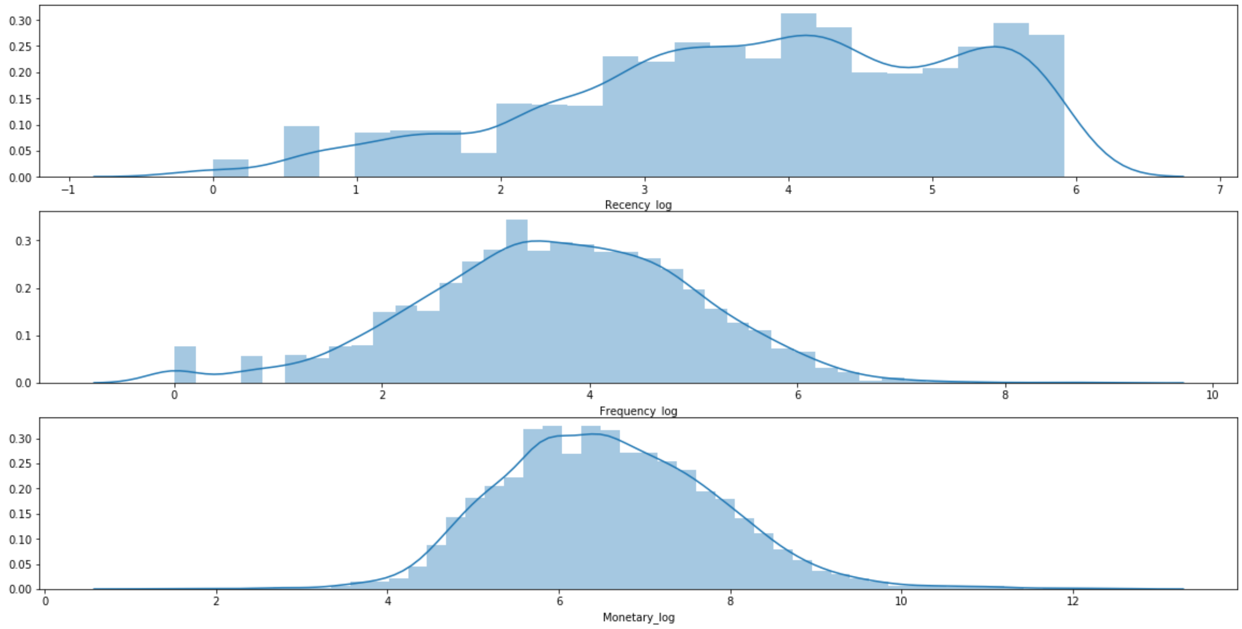

Figure 11: Distribution of R, F, M values after Log transformation.

From the plot above, it can be seen that Recency log transformed data doesn't seem to have ideal normal shape of distribution. So, trying out StandardScaler method from sklearn on log transformed data.

The distribution plot after StandardScaler method (Refer Appendix C)

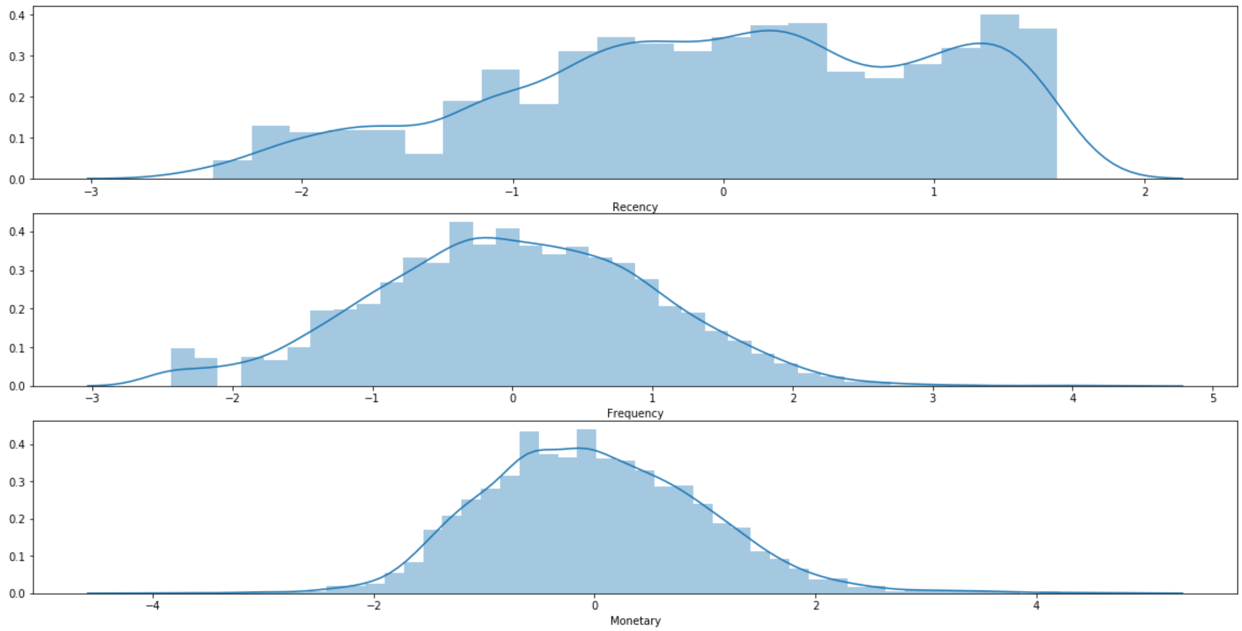

Figure 12: Distribution of R, F, M values after Standard Scalar transformation.

Form the plot, still the Recency's distribution doesn't look normalized.

Checking the statistic values of the distribution of the normalized data (Refer Appendix C)

```
Mean value of the data:

Recency      0.000
Frequency    0.000
Monetary     0.000
Cluster      1.438
dtype: float64
```

Table 10: Mean values

```
Standard Deviation value of the data:

Recency      1.000
Frequency    1.000
Monetary     1.000
Cluster      1.194
dtype: float64
```

Table 11: Standard Deviation

This data shows that the normalized RFM values are normally distributed from statistical perceptive.

Output of K-Means Clustering

After applying k- means clustering algorithm, each customer is assigned a cluster number (0 to 4). *Table 12* shows the average R, F, M values for the customers in each cluster and the number of customers who were grouped together in each cluster. (Refer Appendix C)

cluster	Recency	Frequency	Monetary	Count
0	187.330645	14.866569	296.959737	1364
1	20.957778	38.481111	608.359279	900
2	13.793884	276.388448	6961.902820	883
3	98.346767	78.731318	1500.648423	1191

Table 12: R, F, M averages and count per cluster

Visualization of clustering

A 3D plot was generated with data points for each customer using R, F, M values on the axes. As seen in the plot the clusters are color coded and show a clear distinction in 3D space. The centroids of clusters formed by K-Means are also shows in the plot as black dots. (Refer Appendix C)

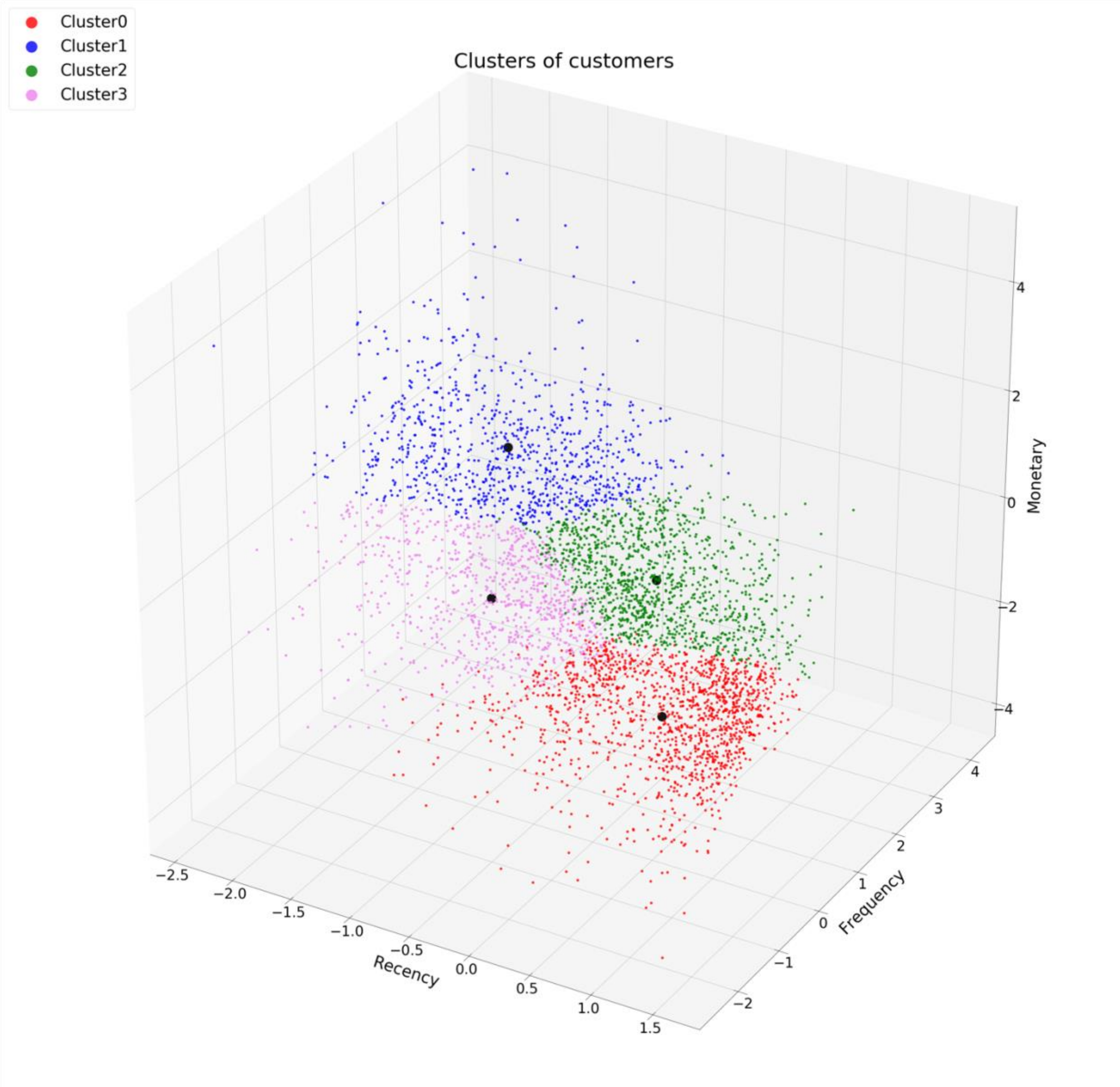

Figure 13: 3D visualization of clusters

Clustering Validation

Validation of the clustering was done using the Elbow method, which is a common heuristic in mathematical optimization to choose the point of diminishing returns. As the number of clusters requested increases, the computational intensity of K-Means increases. So, a balance has to be struck to obtain the highest number of clusters possible without adding unnecessary computational overhead.

To find the optimum number of clusters to be used in K-means clustering, a range of values of k are chosen and the Sum of Squared Errors (SSE) are calculated each time. *Figure 14* shows the plot of SSE vs k values. Elbow method dictates that the point at which the curve bends is the optimum value of k. Looking at the curve, k=4 is the optimum number of clusters to be used. This validates the number of clusters used in the K-Means algorithm. (Refer Appendix C)

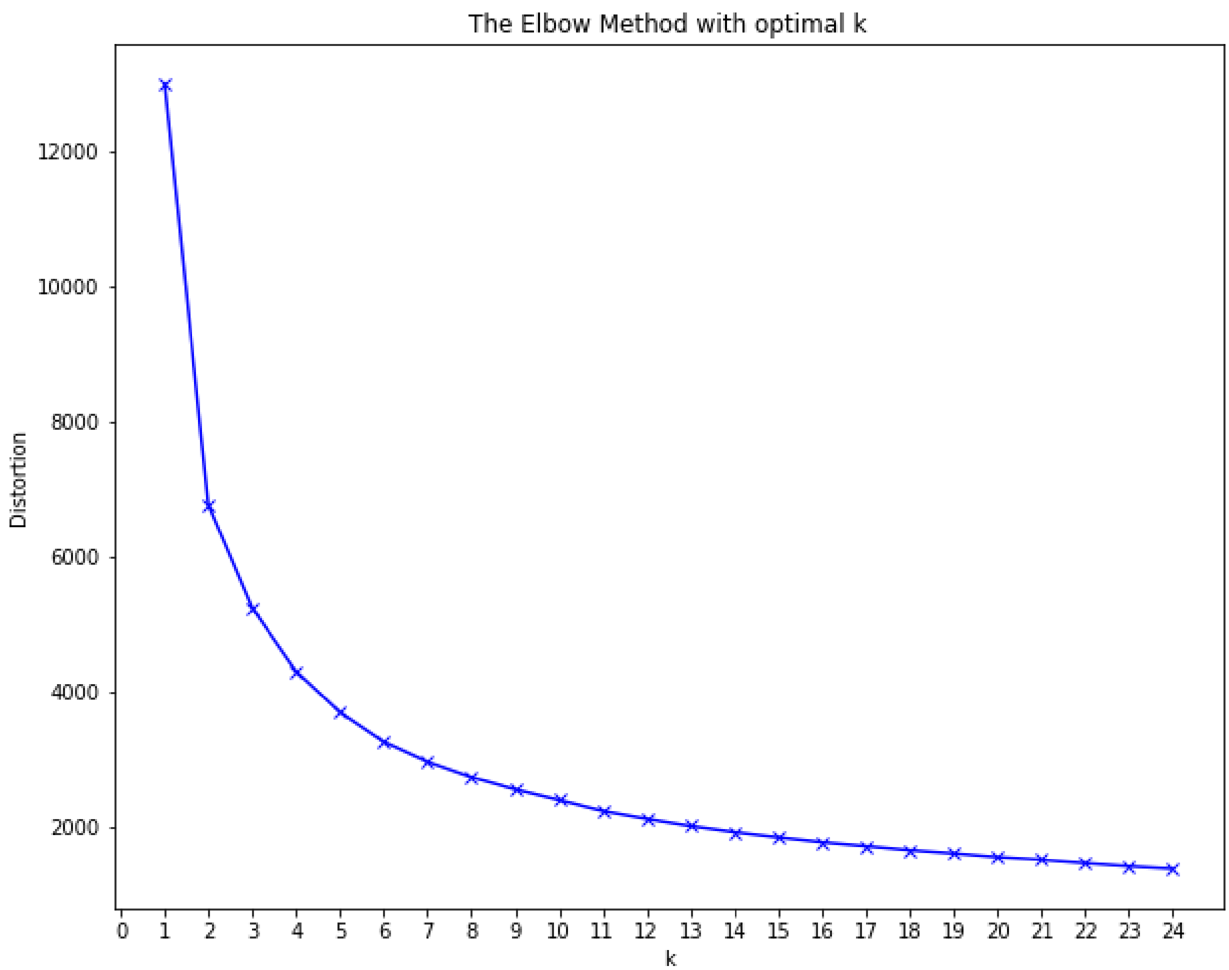

Figure 14: Elbow Method to find optimum clusters

Analysis and Results

RFM analysis assigns a score of 1-4 for R, F and M attributes of customers, with 1 being the least desirable and 4 being the most desirable. Using the RFM scores, customer shopping behavior can be inferred and marketing recommendations are suggested for different combinations of R, F and M values as seen in *Table 13*. (RFM Analysis Boosts Sales. (2019, December 31)).
(Refer Appendix C)

Segment	RFM	Description	Marketing
Best Customers	444	Bought most recently and most often, and spend the most	No price incentives, new products, and loyalty programs
Loyal Customers	X4X	Buy most frequently	Use R and M to further segment
Big Spenders	XX4	Spend the most	Market your most expensive products
Almost Lost	244	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Customers	144	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Cheap Customers	111	Last purchased long ago, purchased few, and spent little	Don't spend too much trying to re-acquire

Table 13: Key RFM segments

Customer segmentation using K-Means algorithm grouped the customers into 4 clusters based on unsupervised learning. The average R, F and M values for customers in different clusters are normalized and color-graded in *figure 15* to show the relative importance of each field among clusters. (Refer Appendix C)

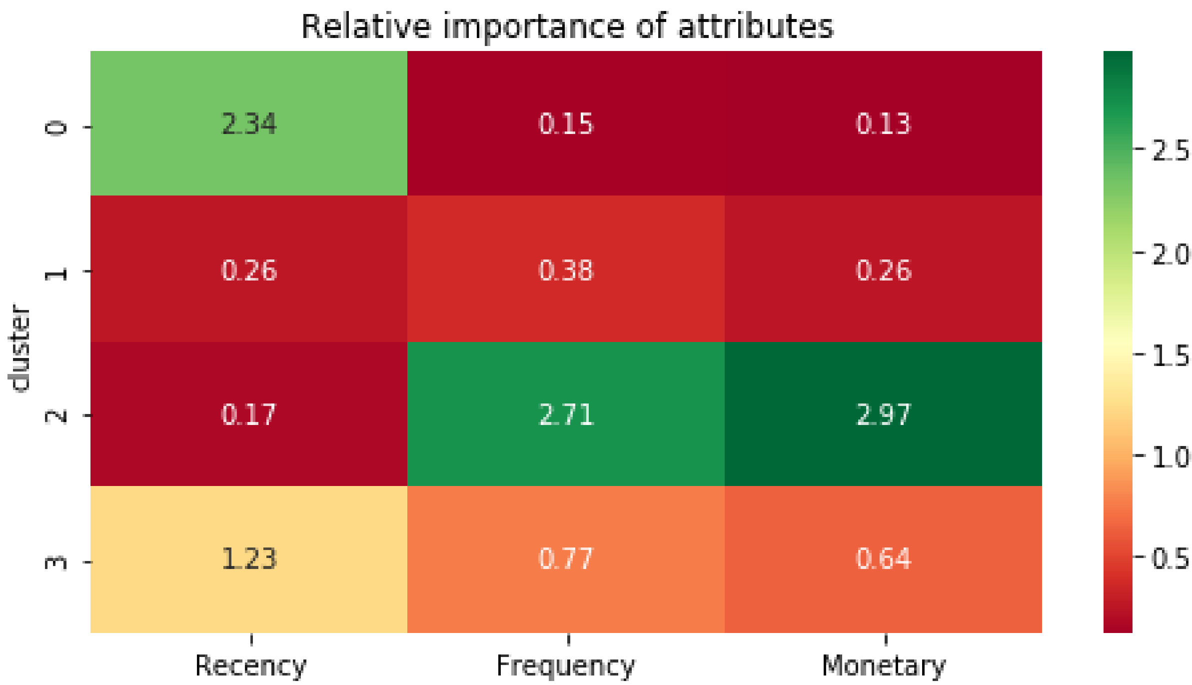

Figure 15: Heat Map of Relative Importance of Attributes

This gives valuable insights into the shopping behavior of the customers in each cluster:

- Cluster 0 (31.44%): Very high R, coupled with very low F and M shows that these are recent new customers who are just trying out the products since they are spending small amounts and are not buying repeatedly yet. This segment may be a result of some new product introduced in the website or a result or marketing campaigns to attract new customers. It is interesting to see that this is the largest of the four segments, so marketing resources should focus on earning the trust and repeated business of this segment for future growth.
- Cluster 1 (20.74%): Low values of R, F and M show the customers spent very little a long time ago and did not return to buy anything else. This is the segment of customers who are

almost lost and very not very profitable to begin with. Marketing resources should not be spent on this segment.

- Cluster 2 (20.35%): Very high values of F and M, with low value of R means shows that this segment of customers very frequent buyers on the e-commerce website and also very big spenders. However, they have not made very recent transactions. This segment, although the smallest of the four segments, should be the highest priority to focus on for marketing because the website seems to be losing highly profitable customers.
- Cluster 3 (27.45%): Average values all across R, F and M shows that these are regular customers who shop frequently and spend moderately and continue to shop at the website. This segment is loyal to the e-commerce website and the target should be to get more customers to this segment.

Conclusions

The shopping data of an e-commerce website was used and extensively analyzed manually using RFM analysis. An unsupervised learning classification algorithm called K-Means was also used to create four customer segments. It was seen that the customer segmentation created by K-Means was of high quality in the sense that it created four distinct categories with clearly distinct shopping behaviors. The segments are fairly sized (between 20-30% of total customers each) making each of the segments significant. It would be very time consuming to come up with these customer segments manually using RFM analysis. The customer segments were further analyzed and marketing strategies were suggested for each category.

Recommendations for future

There are a few possibilities for future work in this area:

- Exploring some newer and more complicated classification algorithms which can handle hybrid data (categorical and numerical) well so that analysis and segmentation can be done on the original dataset itself. Few examples are K-nearest neighbor and K-Means using Gowers distance.
- The R, F, M values of customers were found to be very skewed towards lower values. Resampling data points can be tried to add minority instances or delete few of the majority instances to reduce skew.
- The customer segmentation identified here can be used to build recommendation system to suggest users new products based on their shopping history and correlation with other customers with similar preferences.

Reference

Carrie. (2017, August 17). E-Commerce Data. Retrieved from https://www.kaggle.com/carrie1/ecommerce-data

RFM Segmentation: RFM Analysis, Model, Marketing & Software. (n.d.). Retrieved from https://www.optimove.com/resources/learning-center/rfm-segmentation

K, D. (2020, January 29). K-means clustering using sklearn and Python. Retrieved from https://heartbeat.fritz.ai/k-means-clustering-using-sklearn-and-python-4a054d67b187

Patil, P. (2018, May 23). What is Exploratory Data Analysis? Retrieved from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Henderson, S., Evergreen, S., Jarosewich, T., Henderson, S., Evergreen, S., & Jarosewich, T. (2015, March 31). Word Cloud. Retrieved from https://www.betterevaluation.org/en/evaluation-options/wordcloud

Pallu, P. (n.d.). A Systematic Review on K-Means Clustering Techniques. Retrieved from https://www.academia.edu/35874865/A_Systematic_Review_on_K-Means_Clustering_Techniques

Aggelis, Vasilis & Piraeus, Winbank & Greece, Athens & Gr, Aggelisv@winbank. (2005). Customer clustering using RFM analysis.

How RFM Analysis Boosts Sales. (2019, December 31). Retrieved from https://www.blastanalytics.com/blog/rfm-analysis-boosts-sales

Skewness. (2020, April 14). Retrieved from https://en.wikipedia.org/wiki/Skewness

(n.d.). Retrieved from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm