



MGT7215: Marketing Analytics

Title : Strategic Customer Segmentation
and Predictive Analysis in Marketing

Name: Dhanush Mathighatta Shobhan
Babu

Student ID: 40412492

Word Count: 2191

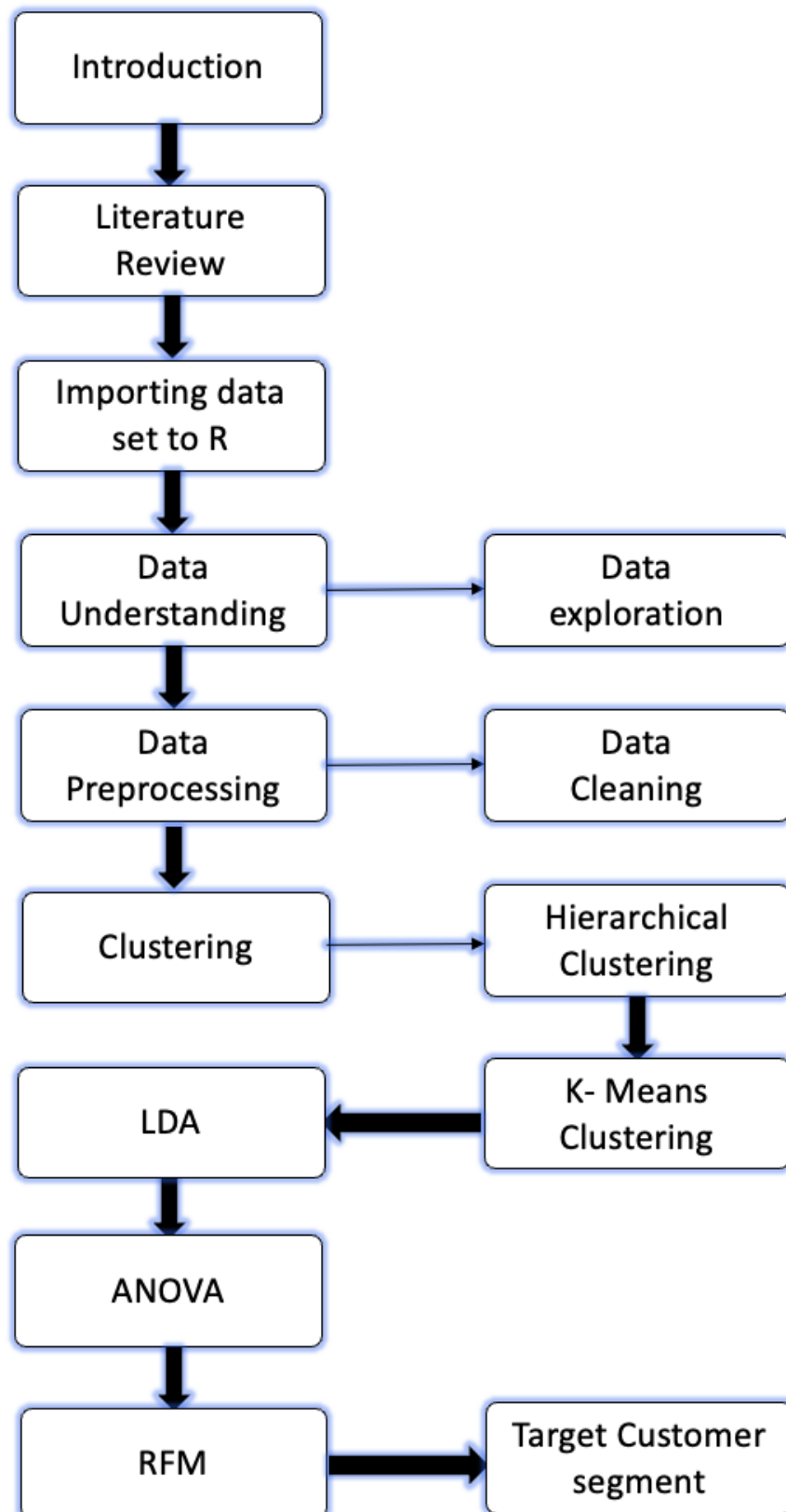
Table Of Content

Sl. No	Content	Page No
1.	1.0 Introduction and Background	1
2.	2.0 Literature review	2-6
3.	Methodology	7-20
4.	Analysis of Segments Using Tableau	21-22
5.	Conclusion	23
6.	References	24-26
7.	Appendix	27-33

Table Of Figures

Sl. No	Content	Page No
1.	Fig 3.0	7
2.	Fig 3.3	13
3.	Fig 3.4	14
4.	Fig 3.5	15
5.	Fig 3.5.1	16
6.	Fig 3.6	17
7.	Fig 3.7.1	18
8.	Fig 3..7.2	19
9.	Fig 3..7.3	19
10.	Fig 3..7.4	19
11.	Fig 4.0	21
12.	Fig 4.1.1	22

Infographic Representation of the report



1. Introduction and Background

1.1 Introduction :

In the dynamic realm of e-commerce, enterprises must employ data analytics to predict and comprehend consumer behaviour, so enhancing their marketing tactics and augmenting client pleasure. This study investigates the functioning of a renowned e-commerce corporation located in the United Kingdom. The company is renowned for providing a wide range of gifts for various occasions and has a substantial customer base. This retailer exclusively operates in the online marketplace and confronts the obstacle of precisely defining its customer segments to boost its marketing techniques and expand its product offerings.

1.2 Background to the Problem:

E-commerce has supplanted the direct interpersonal engagement that brick-and-mortar retailers once had with clients. Consequently, businesses currently depend on customers' online activities, like their purchase history and website navigation, to collect vital data. A comprehensive examination of the sales data, along with the demographic and psychographic profiles, is essential for the shop in question to fully comprehend its consumer segments. The analysis should cover the timeframe from December 1, 2020, to November 24, 2021. Understanding this principle is crucial for staying competitive and achieving long-term success. (Cai Qiuru et al., 2012)

1.3 Objective of Analysis

The primary objective of this study is to utilise sophisticated analytical techniques to precisely categorise the varied customer base of the shop into clearly defined segments. This analysis will offer vital insights into the unique requirements and behavioural tendencies of different segments, notably differentiating between individual consumers and bulk buyers. The acquired knowledge will be utilised to create tailored marketing strategies and goods. The goal is to enhance consumer engagement and loyalty by delivering communication that establishes a deep connection with the specific preferences of each segment. The reference (Tsai & Chiu, 2004) is given

2.0 Literature review

Title of the Paper	Year of Publication	Author	Conclusion
Visualization method for customer targeting using customer map	2005	Ji Young Woo Sung Min Bae Sang Chan Park	The paper introduces the "customer map," a novel visualization method for customer targeting in the service industry. By integrating and analysing diverse customer data, the map displays value distribution across customer needs and characteristics, aiding in strategic decision-making. Applied to a credit card company, it demonstrated enhanced targeting efficiency and strategy development for retaining profitable customers.
Buyer Targeting Optimization: A Unified Customer Segmentation Perspective	2016	Jingyuan Yang, Chuanren Liu, Mingfei Teng, Hui Xiong, March Liao	This paper introduces a novel approach by integrating customer segmentation and buyer targeting into a unified optimization framework, significantly enhancing marketing analytics. The developed K-Classifiers Segmentation algorithm simultaneously optimizes customer segmentation and targeting, offering substantial improvements in targeting accuracy and meaningful segmentation based on customers' buying preferences. Tested on synthetic and real-world B2B datasets, the method outperforms traditional techniques, providing actionable insights for tailored marketing strategies. The inclusion of a Profile-Consistent Algorithm addresses profile inconsistency, ensuring segmentation aligns with customer profiles and decision

			preferences. This work presents a significant step forward in marketing optimization, offering practical tools for more effective buyer targeting and customer understanding.
Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment	2012	You-Shyang Chen Ching-Hsue Cheng Chien-Jung Lai Cheng-Yi Hsu Han-Jhou Syu	The study emphasizes the necessity for hospitals to identify patient contributions effectively to allocate healthcare resources properly and enhance healthcare quality. It introduces a hybrid clustering-classification approach to solve classification challenges in hospital management, demonstrating through empirical results superior accuracy over other methods. The approach also identifies specific types of patients contributing more to hospital revenue and discovers significant diagnostic items and patient characteristics. This model aids in focusing on high-yield patients, fostering better relationships, and enhancing consumer satisfaction.
Customer Segmentation Using Clustering and Data Mining Techniques	2013	Kishan R . Kashwan,C. M. Velu	The cluster analysis applied to a sample of respondents revealed significant insights into potential market segments. Using a non-hierarchical k-means clustering algorithm, initial cluster centres were defined, and final stable cluster centres were determined through iterative refinement. This process ensured the identification of distinct, stable clusters, which were further analysed for stability by splitting the sample data. This approach not only provided a method for effective market segmentation but also

			facilitated intelligent, automated decision-making for managers. The future work aims to further automate market forecasting and planning processes
Segmenting and Targeting Customers Through Clusters Selection & Analysis	2015	Ilung Pranata, Geoff Skinner	Utilizing k-means clustering and validation techniques such as the Elbow Method, Davies-Bouldin Index, and Silhouette Width, the study segments customers of a wholesale distributor into three categories based on their annual spending across six product categories. It highlights high spenders with an average expenditure above \$90,000, primarily on groceries, alongside middle and low spenders. This approach facilitates targeted marketing efforts aimed at optimizing customer retention and enhancing competitive advantage within the business landscape.
Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce	2021	Ritu Punhani, V.P.S Arora, Sai Sabitha, Vinod Kumar Shukla	The study utilized data mining techniques, specifically the K-means clustering algorithm, to uncover hidden patterns and behaviors among e-commerce customers. It analyzed a dataset from a small online store to identify which products had the highest sales and which payment method (credit or PayPal) was more popular. The findings revealed that products in categories 503-505 were the most sold, and payment by credit was more commonly used than PayPal. The research highlights the importance of personalized marketing strategies in engaging customers effectively and

			adapting to changes in the virtual market.
Customer Segmentation using K-means Clustering	2018	Tushar Kansal, Suraj Bahuguna , Vishal Singh, Tanupriya Choudhury	The paper delves into customer segmentation using the K-means clustering method, focusing on internal cluster validation to select the most appropriate algorithm for unlabelled datasets. This approach facilitates the precise grouping of customers based on their attributes, enabling businesses to tailor their marketing and service strategies more effectively. The study underscores the significance of internal validation for accurate clustering without external labels, which is pivotal for analyzing unlabelled data sets. Through this methodology, businesses can achieve targeted customer engagement and improve service delivery, highlighting the practical benefits of customer segmentation in a data-driven business environment.
GPHC: A heuristic clustering method to customer segmentation	2021	Zhao-Hui Sun , Tian-Yu Zuo , Di Liang , Xinguo Ming , Zhihua Chen , Siqi Qiu	This paper presents the Gaussian Peak Heuristic Clustering (GPHC) method for customer segmentation, effectively handling real, ambiguous customer requirement data. By using a standardized Gaussian distribution for data modeling and combining niching genetic algorithms with hierarchical clustering, GPHC accurately segments customers based on preference patterns. Demonstrated through a case study and numerical experiments, GPHC outperforms traditional

			clustering methods, offering clear, actionable segmentation insights for businesses.
Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm	2021	Yue Li, Xiaoquan Chu, Dong Tian, Jianying Feng, and Weisong Mu.	The study introduces a novel customer segmentation method integrating an improved K-means algorithm with an Adaptive Particle Swarm Optimization (PSO) algorithm to enhance optimization accuracy. The Adaptive Learning PSO (ALPSO) improves PSO's optimization by redesigning inertia weight, learning factors, and updating methods. It uses ALPSO for optimizing K-means cluster centers (KM-ALPSO) and introduces an improved KM-ALPSO (IKM-ALPSO) for customer segmentation, demonstrating superior performance over existing models on UCI datasets and a real-world grape-customer dataset, achieving higher accuracy in customer segmentation.
A Two Phase Clustering Method for Intelligent Customer Segmentation	2010	Morteza Namvar, Mohammad R. Gholamian, Sahand KhakAbi	This paper introduces a novel two-phase clustering method for customer segmentation, combining RFM, demographic, and LTV data through K-means clustering. Applied to an Iranian bank's dataset, the method effectively segments customers into actionable groups, enabling targeted management strategies. The approach demonstrates significant utility in enhancing customer relationship management practices.

3.0 Methodology

CRISP-DM, an industry-neutral data mining process model, consists of six iterative phases spanning from understanding business needs to deploying solutions (Schröder et al., 2021). To address our business challenge, we'll tailor CRISP-DM and begin with understanding the data, preparing it, and then pre-processing it before applying unsupervised learning techniques. This methodology offers a structured approach to solve business problems, encompassing model generation. Having completed the initial phase of understanding the business context, we'll now move on to the subsequent steps. The following flowchart illustrates the specific actions that will be taken to address this business challenge.

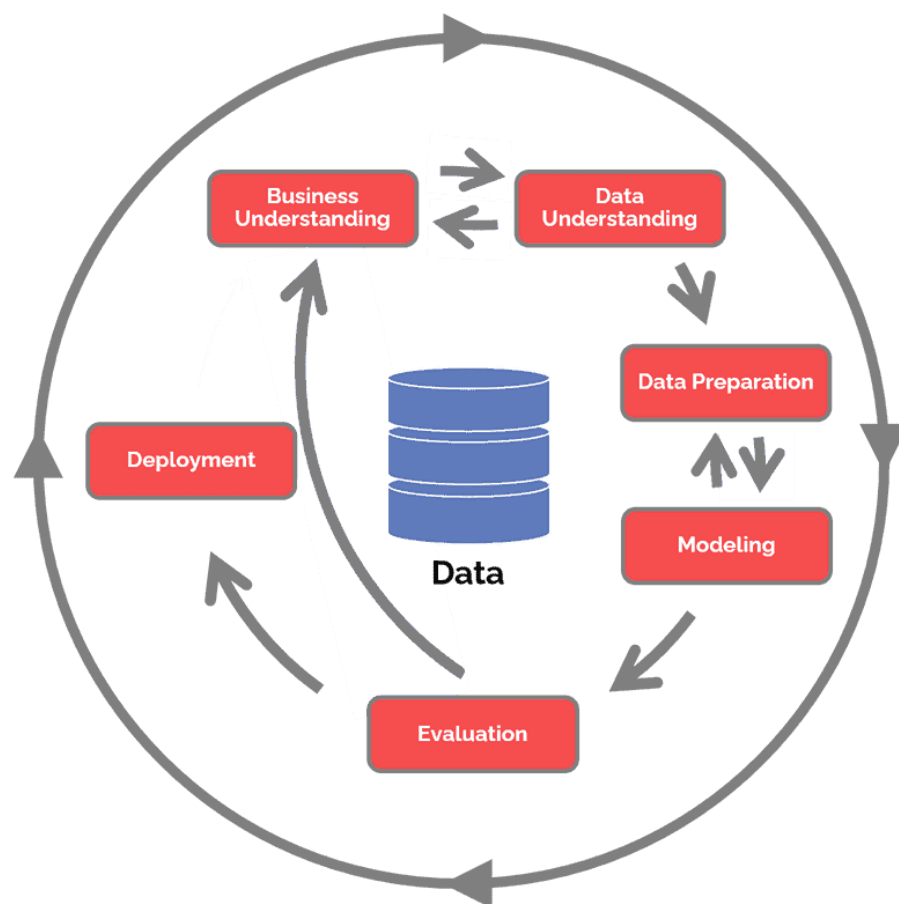
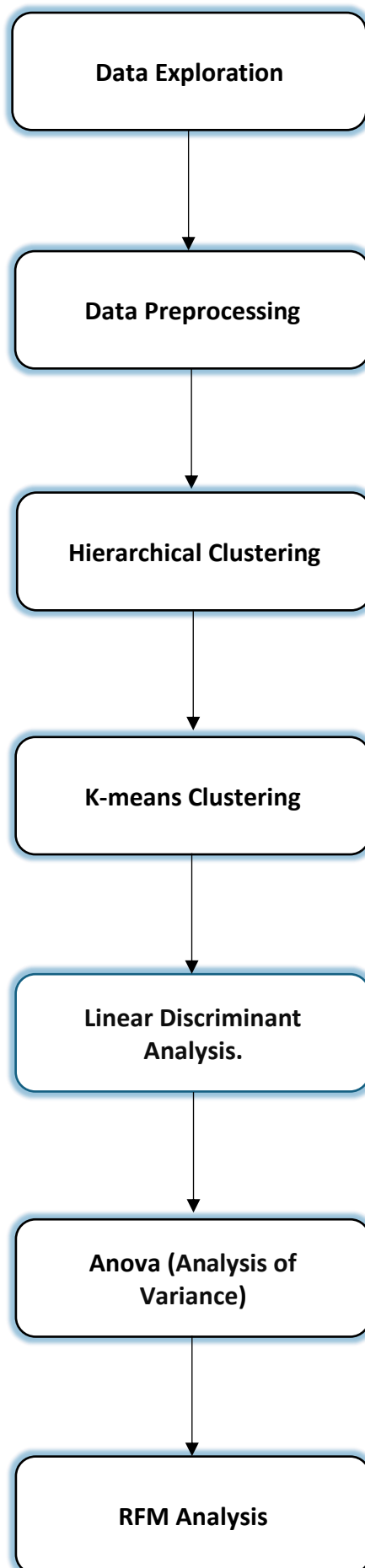


Fig 3.0: CRISP – DM Methodology



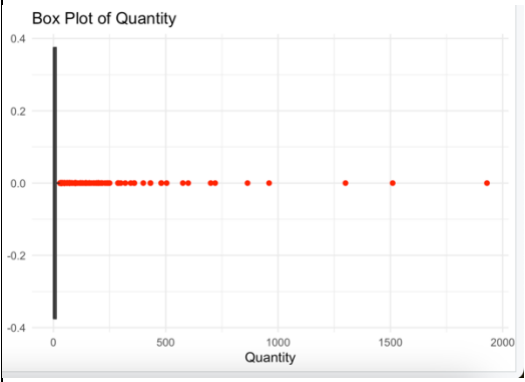
Flow chart of the workflow

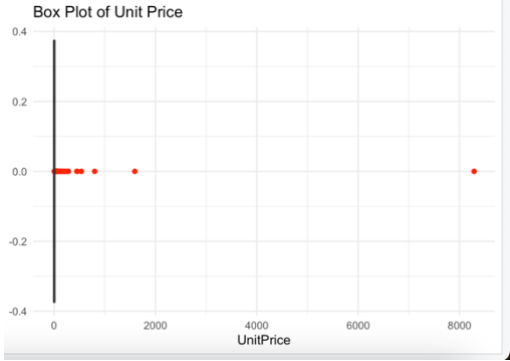

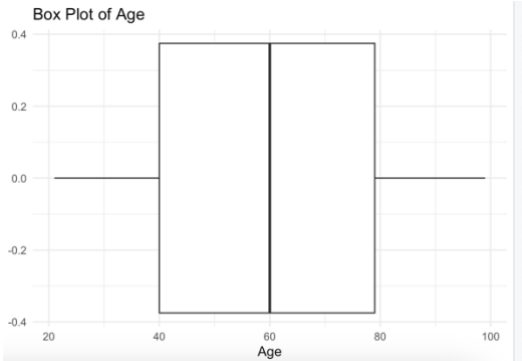
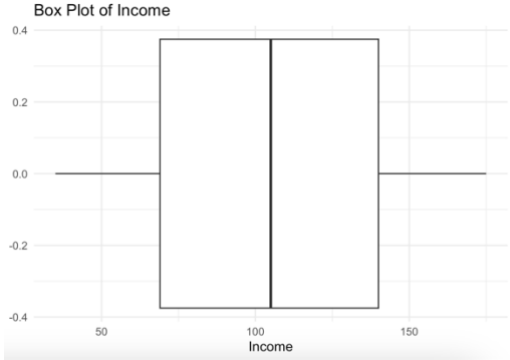
3.1 Data exploration

The dataset comprises 10,000 rows and 14 columns.

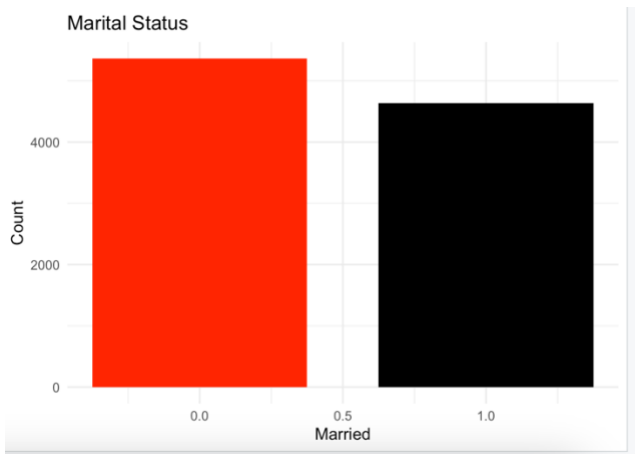
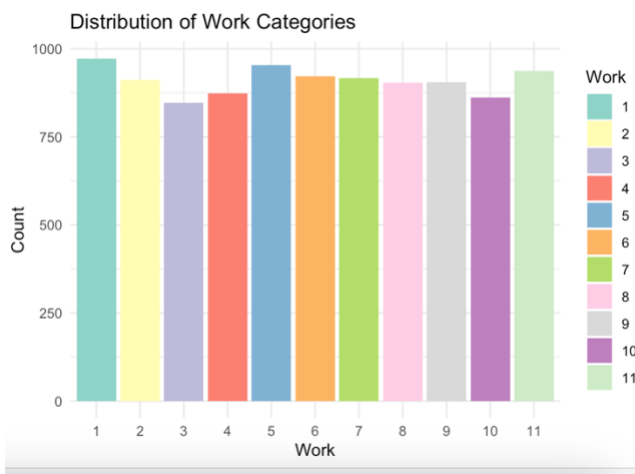
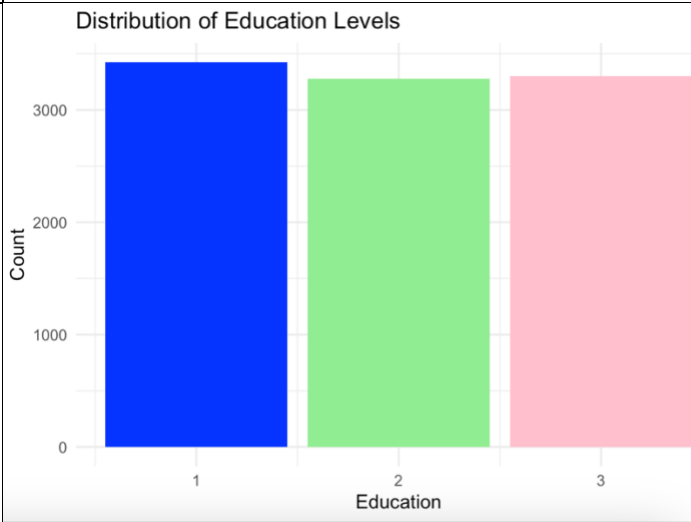
Data Variable Name	Type of data	Description
Invoice no	Nominal	The 6-digit integer number that is exclusively assigned to each individual transaction. If the code begins with the letter 'c', it signifies a cancellation.
Stock code	Nominal	The 5-digit integral number that is exclusively assigned to each individual product.
Description	Nominal	Product name
Customer id	Nominal	The 5-digit integral number that is exclusively assigned to each customer.
Zip code	Nominal	The customer's residential zip code, consisting of five digits
Invoice date	date	The specific date and time at which each transaction was created

Numerical data:

Data Variable Name	Min	Median	Mean	Max	Visualization
Quantity	1	3	10.66	1930	 A box plot titled 'Box Plot of Quantity' showing the distribution of the 'Quantity' variable. The x-axis is labeled 'Quantity' and ranges from 0 to 2000 with major ticks every 500 units. The y-axis ranges from -0.4 to 0.4 with major ticks every 0.2 units. The plot shows a very narrow distribution with a median line at approximately 3, a mean line at approximately 10.66, and a maximum value at 1930. The data points are represented by red dots along the x-axis.

Unit Price	0.04 0	2.080	4.802	8286.22 0	 <p>Box Plot of Unit Price</p> <p>This box plot shows the distribution of Unit Price. The y-axis ranges from -0.4 to 0.4, and the x-axis ranges from 0 to 8000. The plot shows a very narrow distribution with a median near 0 and a few outliers at higher prices.</p>
Return Rate	0	0.1947	0.1767 1	6.90909	 <p>Box Plot of Return Rate</p> <p>This box plot shows the distribution of Return Rate. The y-axis ranges from -0.4 to 0.4, and the x-axis ranges from 0 to 6. The plot shows a very narrow distribution with a median near 0 and a few outliers at higher return rates.</p>
Age	21	60	59.95	99	 <p>Box Plot of Age</p> <p>This box plot shows the distribution of Age. The y-axis ranges from -0.4 to 0.4, and the x-axis ranges from 20 to 100. The plot shows a very narrow distribution with a median near 60 and a few outliers at higher ages.</p>
Income	35	105.0	104.8	175	 <p>Box Plot of Income</p> <p>This box plot shows the distribution of Income. The y-axis ranges from -0.4 to 0.4, and the x-axis ranges from 50 to 150. The plot shows a very narrow distribution with a median near 105 and a few outliers at higher incomes.</p>

Categorical data:

Data Variable Name	Distinct Value	Visualization																								
Married	1 or 0	 <table border="1"><caption>Marital Status Data</caption><thead><tr><th>Married</th><th>Count</th></tr></thead><tbody><tr><td>0.0</td><td>~4500</td></tr><tr><td>1.0</td><td>~4000</td></tr></tbody></table>	Married	Count	0.0	~4500	1.0	~4000																		
Married	Count																									
0.0	~4500																									
1.0	~4000																									
Work	1,2,3,4,5,6,7,8,9,10,11	 <table border="1"><caption>Distribution of Work Categories</caption><thead><tr><th>Work</th><th>Count</th></tr></thead><tbody><tr><td>1</td><td>~950</td></tr><tr><td>2</td><td>~900</td></tr><tr><td>3</td><td>~850</td></tr><tr><td>4</td><td>~850</td></tr><tr><td>5</td><td>~950</td></tr><tr><td>6</td><td>~900</td></tr><tr><td>7</td><td>~900</td></tr><tr><td>8</td><td>~900</td></tr><tr><td>9</td><td>~900</td></tr><tr><td>10</td><td>~850</td></tr><tr><td>11</td><td>~900</td></tr></tbody></table>	Work	Count	1	~950	2	~900	3	~850	4	~850	5	~950	6	~900	7	~900	8	~900	9	~900	10	~850	11	~900
Work	Count																									
1	~950																									
2	~900																									
3	~850																									
4	~850																									
5	~950																									
6	~900																									
7	~900																									
8	~900																									
9	~900																									
10	~850																									
11	~900																									
Education	1,2,3	 <table border="1"><caption>Distribution of Education Levels</caption><thead><tr><th>Education</th><th>Count</th></tr></thead><tbody><tr><td>1</td><td>~3500</td></tr><tr><td>2</td><td>~3200</td></tr><tr><td>3</td><td>~3200</td></tr></tbody></table>	Education	Count	1	~3500	2	~3200	3	~3200																
Education	Count																									
1	~3500																									
2	~3200																									
3	~3200																									

3.2 Data Pre-Processing

3.2.1 Data quality issues

Variable Name	Data Quality Issues
Return Rate	There are more than 150 data items, exceeding the value of 1 (which is a representation of 100%). The return rate can logically reach a maximum of 100%. Hence, any value more than 1 cannot be deemed acceptable.
Description	There are 26 missing row values in which 3 of them are indicated as '?' that need to be excluded.
Customer ID	Entirety there are 2491 row values which are recorded as NAs, and these values cannot be considered for further analysis.
Work, Married, Education, ZipCode	There are unfactored values that need to be factored in order to avoid complications in categorization and analysis.

3.2.2 Addressing data quality issues

Resolve the issue of inadequate data quality through the process of data cleansing. Create derived characteristics based on the selected model from the first phase. The optimal approach for these processes is contingent upon the model utilised (Schröer et al., 2021).

Variable Name	Addressing the issue
Return Rate	The values which were more than 1, were removed using the filter() function.
Description	The values with incomplete information were filtered out
Customer ID	These missing or NAs were omitted from the data by using omit() function
Work, Married, Education, ZipCode	The non-factorised values were factorised by using the as.factor() function.

3.3 Hierarchical clustering

Clustering, often known as cluster analysis, is a significant topic in data mining (Tripathi et al., 2018). The dataset is divided into several groupings, known as clusters, where the data points within each cluster exhibit a higher degree of similarity to one another compared to those in different clusters. Hierarchical clustering is a cluster analysis technique that constructs a hierarchy of data points as they are assigned to or removed from a cluster (Tripathi et al., 2018).

After performing Hierarchical clustering, we plot the elbow plot shown below.

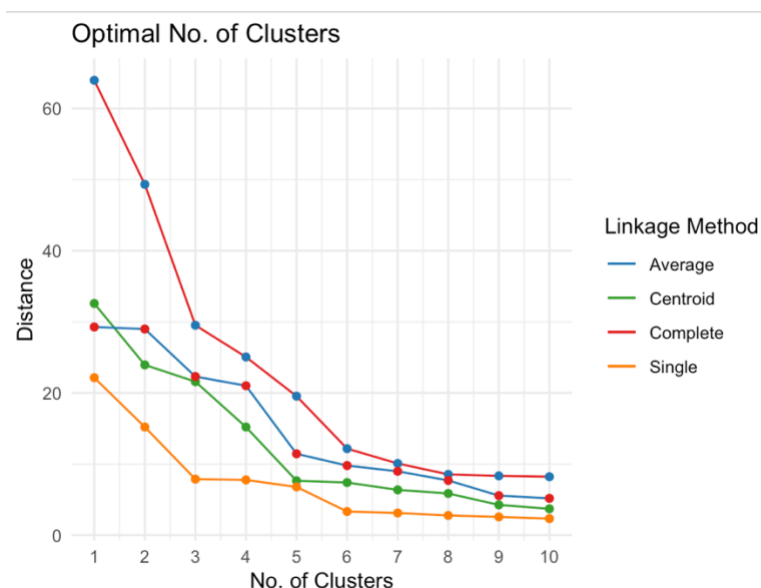


Fig 3.3: Elbow Plot

The elbow plot is a graphical instrument used to estimate the ideal number of clusters for segmenting data. The x-axis is labelled "No. of Clusters" and ranges from 1 to 10, indicating several possible cluster solutions. The y-axis, labelled as "Distance," represents the within-cluster sum of squares, ranging from 0 to approximately 60. The hierarchical clustering linkage methods are represented by four lines, each having a distinct colour: Average (blue), Centroid (green), Complete (red), and Single (orange). The 'elbow' refers to the point on each curve where the reduction in distance becomes smallest as the number of clusters increases. This indicates that the best number of clusters is 3.

3.4 K-means clustering

K-means is an iterative algorithm that aims to partition the data into k separate groups (Shirole et al., 2021).

Steps to take when utilising the K-means clustering algorithm:

- Determine the number of clusters K in advance.

- Randomly select K data points to initialise the Centroid.
- Calculate the distance between the upcoming data points and all centroids.
- Allocate the data point to the closest cluster. • Iterate this process until all data points converge to a cluster.

Formula for Centroid Determination:

$$C_i = 1/M \sum_{j=1}^m X_j$$

Formula for Euclidean Distance

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

The K-means clustering analysis was performed on the dataset after normalisation. The dataset was partitioned into three clusters as the desired outcome. The analysis began with a predetermined starting point to ensure consistency. It used 1000 random starting points and a maximum iteration limit of 500 to optimise the cluster centroids. After running the programme, we noted that the sizes of the resulting clusters were 1589, 17, and 5621 for clusters one, two, and three accordingly. The 2D visualisation, possibly obtained using principal component analysis, showed that 35.8% of the variability was explained by the first principal component (Dim1), whereas 33.6% was explained by the second principal component (Dim2). This visualisation clarified the structure of the data, highlighting cluster one (shown in blue) and cluster three (in grey) which exhibited large concentrations of data points, indicating significant homogeneous segments. In contrast, cluster two (highlighted in yellow) exhibited a scattered arrangement of data points, possibly indicating a subset of abnormalities or a separate subgroup within the sample. The clusters were accentuated by using ellipses to depict the confidence zones, and the core points of each cluster were indicated with stars to enhance identification.

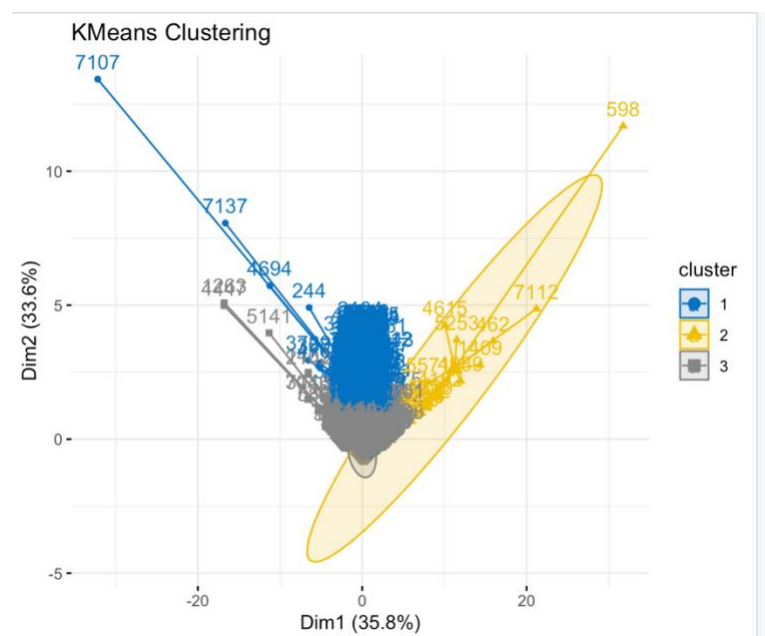


Fig 3.4: K-means plot

3.5 Linear Discriminant analysis

LDA is a supervised learning technique that reduces the dimensionality of a sample set by projecting it onto a lower-dimensional sample space. It then determines the optimal linear combination of features (Yang et al., 2011). This linear combination minimises the variation within each group and maximises the divergence between the two groups, resulting in the highest level of separability among the samples.

```
Call:
lda(cluster ~ Married + Age + Income + Work + Education, data = data_clustered)

Prior probabilities of groups:
      1      2      3
0.21986993 0.00235229 0.77777778

Group means:
      Married1      Age      Income      Work10      Work11      Work2      Work3      Work4      Work5      Work6
1 0.3228446 60.53996 106.4305 0.05663940 0.03901825 0.1334172 0.1227187 0.08118313 0.09376967 0.06544997
2 0.5294118 59.23529 108.4706 0.11764706 0.00000000 0.2352941 0.00000000 0.05882353 0.11764706 0.11764706
3 0.3180929 59.93809 104.8312 0.05959794 0.05034691 0.1442804 0.1225761 0.07863370 0.09179861 0.06635830
      Work7      Work8      Work9      Education2      Education3
1 0.05789805 0.07866583 0.08873505 0.3064821 0.2693518
2 0.00000000 0.05882353 0.05882353 0.5882353 0.2941176
3 0.06671411 0.06618039 0.07578723 0.3047500 0.2691692

Coefficients of linear discriminants:
      LD1      LD2
Married1 -0.76776857 -0.50839981
Age      -0.02062944 0.02300125
Income   -0.01944559 0.01328375
Work10    -0.05046040 -1.14928273
Work11    2.34196725 -0.25036799
Work2     0.45880104 -0.79211046
Work3     0.91062042 0.93057110
Work4     0.31649256 0.46438721
Work5     0.06823691 -0.04098228
Work6     0.03712925 -0.49614918
Work7     1.56816062 0.09731632
Work8     -0.53942313 1.23797545
Work9     -0.52661265 1.31584668
Education2 -1.04188710 -1.24654234
Education3 -0.34312840 -0.84568514

Proportion of trace:
      LD1      LD2
0.5771 0.4229
```

Fig 3.5: LDA Analysis

In a combined K-means and Linear Discriminant Analysis (LDA) approach, the dataset was segmented into three clusters, revealing distinct groupings that highlight homogeneity and potential outliers. LDA further elucidated the discriminatory power of specific predictors across the identified clusters. The analysis pinpointed work-related variables ('Work10', 'Work3', 'Work7', 'Work9', and 'Work8') and an educational variable ('Education2') as significant discriminants. In particular, 'Work10' displayed a pronounced negative correlation with the first discriminant function, LD1, whereas 'Work3', 'Work7', and 'Work9' exhibited

positive associations. Conversely, 'Work2' and 'Education2' were strongly inversely related to the second discriminant function, LD2.

The discriminant functions explained a considerable portion of the variance between the groups, with LD1 accounting for 57.71% and LD2 for 42.29%. This analytical fusion offered an in-depth perspective on the underlying structure of the data. By leveraging the strengths of both K-means clustering to identify inherent data groupings and LDA to comprehend the directional influence of predictors on these groupings, a robust narrative emerged. This narrative not only identifies but also characterizes the nature of the clusters, enriching the analysis with insights into the defining attributes of each cluster.

```
> #confusion Matrix
> confusionMatrix(fit.predict, as.factor(data_clustered$cluster))
Confusion Matrix and Statistics
```

	Reference		
Prediction	1	2	3
1	0	0	0
2	0	0	0
3	1589	17	5621

```
Overall Statistics

                Accuracy : 0.7778
                95% CI : (0.768, 0.7873)
    No Information Rate : 0.7778
    P-Value [Acc > NIR] : 0.5067

                Kappa : 0

McNemar's Test P-Value : NA

Statistics by Class:
```

	Class: 1	Class: 2	Class: 3
Sensitivity	0.0000	0.000000	1.0000
Specificity	1.0000	1.000000	0.0000
Pos Pred Value	NaN	NaN	0.7778
Neg Pred Value	0.7801	0.997648	NaN
Prevalence	0.2199	0.002352	0.7778
Detection Rate	0.0000	0.000000	0.7778
Detection Prevalence	0.0000	0.000000	1.0000
Balanced Accuracy	0.5000	0.500000	0.5000

```
> |
```

Fig3.5.1: Confusion Matrix of LDA Analysis

The confusion matrix for the LDA classifier indicates an overall accuracy of 77.78%, which coincides with the No Information Rate, signifying that the model's predictive capability is no

better than random guessing. The Kappa statistic is 0, reinforcing the model's lack of predictive improvement over chance. The model exhibits complete bias towards predicting every instance as belonging to class 3, as evidenced by the sensitivity for classes 1 and 2 being 0, and specificity being 1 due to no positive predictions for these classes. The Positive Predictive Value is not applicable for classes 1 and 2, and the Negative Predictive Value for class 2 is high but trivial. Detection rates for classes 1 and 2 are 0, while class 3 is 77.78%, matching the model's accuracy. Balanced accuracy across classes is 0.5, reflecting performance equivalent to random chance.

3.6 Analysis of Variance (ANOVA)

ANOVA, or Analysis of Variance, is used for a range of analytical and data processing tasks, such as clustering and data mining.(Kashwan & Velu, 2013). The ANOVA test provides appropriate inference while dynamically splitting and merging groups.

```
> #ANOVA
> anova(lm(ld[,1] ~ data_clustered$cluster ))
Analysis of Variance Table

Response: ld[, 1]
              Df Sum Sq Mean Sq F value    Pr(>F)
data_clustered$cluster    1   14.9  14.9025   14.885 0.0001153 ***
Residuals              7225 7233.6   1.0012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm(ld[,2] ~ data_clustered$cluster ))
Analysis of Variance Table

Response: ld[, 2]
              Df Sum Sq Mean Sq F value    Pr(>F)
data_clustered$cluster    1    7.1   7.0602   7.0505 0.007941 **
Residuals              7225 7234.9   1.0014
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Fig 3.6 Anova Output

An Analysis of Variance (ANOVA) was utilised to assess the relevance of the discriminant functions obtained from Linear Discriminant Analysis (LDA) in explaining the variance of the clusters. The analysis of variance (ANOVA) results for the first discriminant function (LD1) revealed a substantial F-value of 14.885, accompanied by a p-value of about 0.0001. This indicates a highly significant association between LD1 and the cluster assignments. The importance is shown by three asterisks, which signify a p-value below 0.001. The second discriminant function (LD2) showed statistical significance, as indicated by an F-value of 7.0505 and a p-value of about 0.0079. The p-value is denoted by two asterisks, which signifies that it is less than 0.01. The results confirm that the discriminant scores are relevant predictors for the clusters, with LD1 being a more powerful predictor than LD2.

3.7 RFM

RFM analysis, a well-known method, is utilised to assess customers' purchasing patterns, taking into account recency, frequency, and monetary factors (Christy et al., 2021). A scoring methodology has been devised to assess the scores of Recency, Frequency, and Monetary. Ultimately, the scores of the three variables are combined into a single metric called RFM. This metric is utilised to forecast future patterns by examining the customer's present and prior histories (Christy et al., 2021).

Score Category	1	2	3	4	5
Recency Score	459	472	455	449	494
Frequency Score	976	-	493	451	409
Monetary Score	509	497	411	457	455

The RFM study conducted on consumer transactional data provided valuable insights into purchase behaviours, categorised based on Recency, Frequency, and Monetary value. Customers were evaluated using a rating system ranging from 1 to 5 for each RFM category. Quintiles were employed to ensure a detailed and precise categorization of customers. The output indicated a uniform distribution in Recency, indicating a diverse range of time intervals since the last purchases within the client base. The frequency and monetary scores exhibited a moderate concentration in the central quintiles, suggesting that customers likely to have average transaction frequencies and expenditure levels. The RFM scores play a crucial role for organisations in understanding customer value, developing focused marketing strategies, and improving customer relationship management. The distribution of scores across quintiles helps to illustrate the range of consumer engagement and spending patterns. This segmentation enables customised client outreach, with the goal of enhancing customer loyalty and optimising long-term corporate revenue.

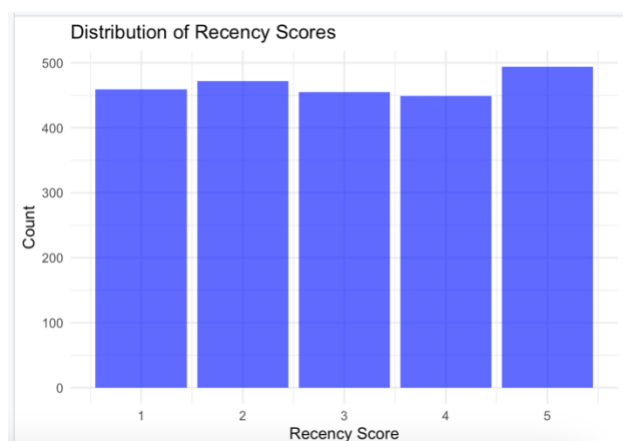


Fig 3.7.1: Recency Score Graph

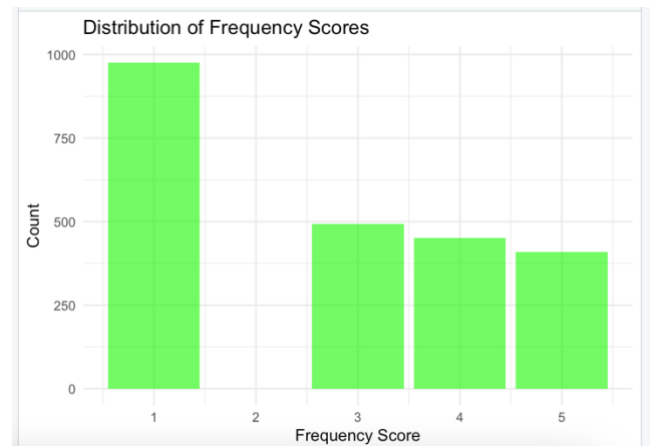


Fig 3.7.2 Frequency Score Graph

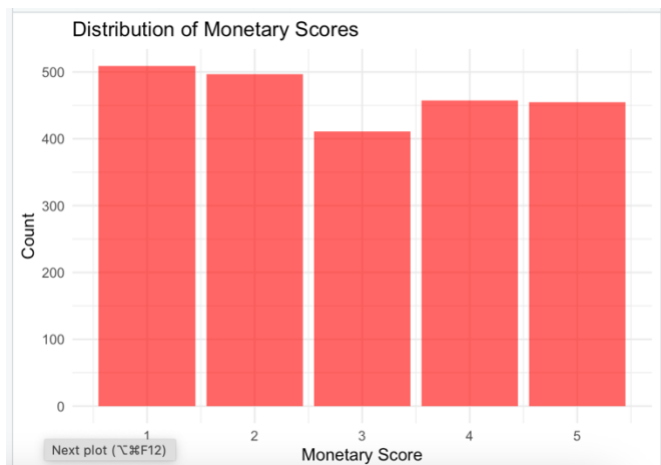


Fig 3.7.3 Monetary Score Graph



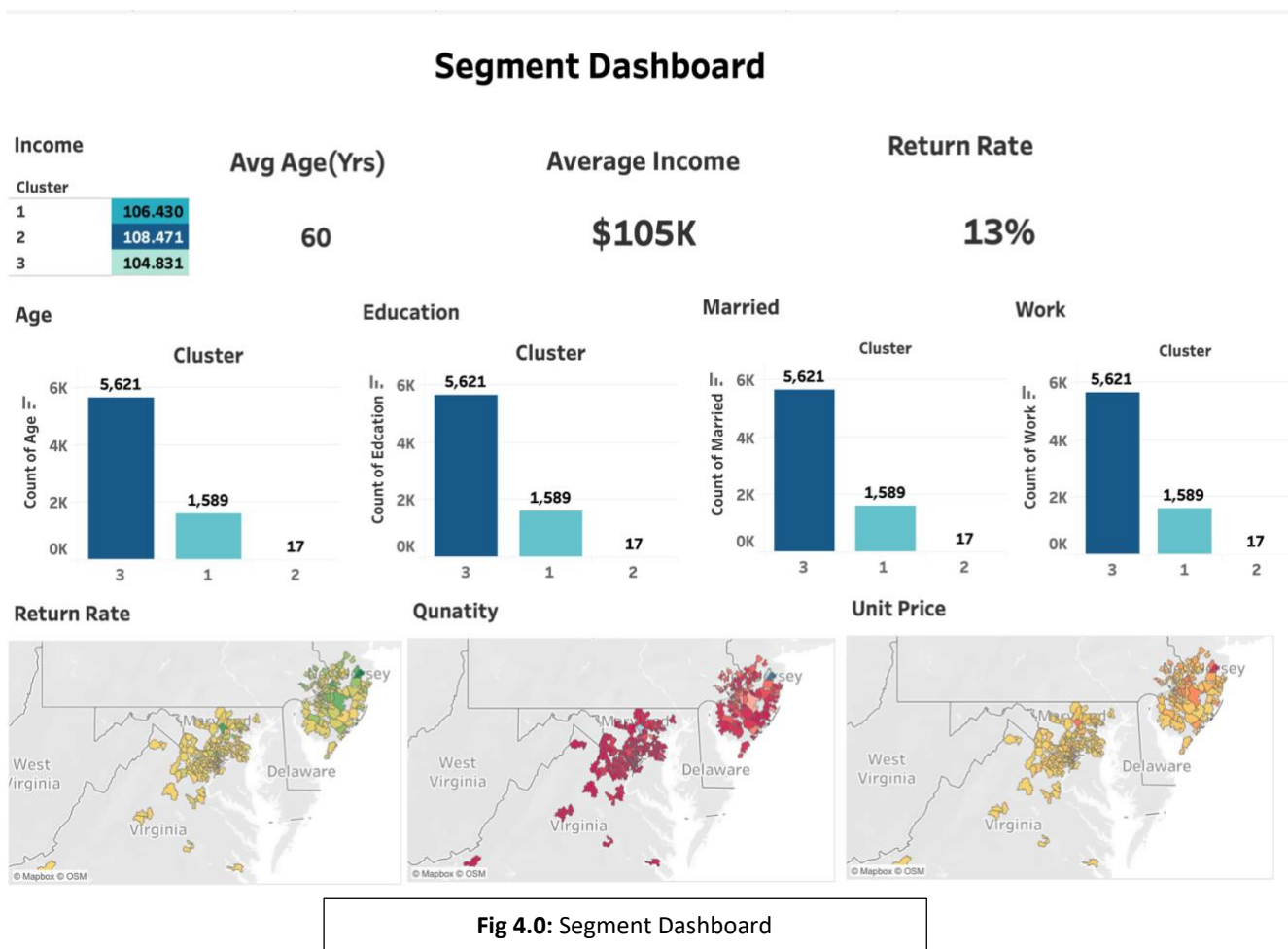
Fig 3.7.4 Frequency vs Monetary by Recency Score

3.8 Customer Targeting Based on Revenue per cluster

Cluster	Total Customers	Total Profits	Revenue/Segment
1	1589	\$31361	\$19.7
2	17	\$15465	\$910
3	5621	\$116500	\$20.7

Cluster 2, the smallest with only 17 customers, is the most lucrative, averaging 910 units in revenue per customer, indicating high-value, possibly premium market consumers. In contrast, Cluster 3's vast 5,621 customer base contributes the most overall profit but has the lowest spend per person at 20.7 units, suggesting a volume-driven segment. Strategic focus on Cluster 2 could yield higher per-customer profitability, whereas Cluster 3 offers broad market coverage opportunities.

4.0 Analysis of segments using Tableau



The dashboard provides a concise overview of client segmentation, presenting crucial facts across three distinct groupings. Cluster 1, albeit smaller in size, demonstrates substantial financial capacity, with an average income of approximately \$106,400. Cluster 2 outperforms this with a significantly higher mean income of around 108.5K, suggesting the possibility of targeting luxury or premium products. In contrast, Cluster 3, although it is the largest group, has a somewhat lower average income of 104.8K.

The clusters are expected to reflect a mature customer base with a respectable level of brand loyalty or product satisfaction, as indicated by their average age of 60 years across all segments and a return rate of 13%. The distribution of demographics, with a strong emphasis on Cluster 3 due to criteria like marital status and job, indicates a wide audience for mainstream marketing initiatives. From a geographical perspective, analysing the distribution of indicators such as return rates and unit pricing across different regions can provide valuable insights for developing strategies tailored to certain regions. Based on revenue levels and consumer distribution, focused initiatives in Cluster 2 have the potential to generate significant profits, whilst methods that prioritise volume could be advantageous for Cluster 3. The lesser size of Cluster 1 should not detract from its significant spending power.

Segmenatation / Clusters Results			
	Cluster		
	1	2	3
Count of seg.csv	1,589	17	5,621
Avg. Age	61	59	60
% of Total Count of Edcation along Table (Across)	21.99%	0.24%	77.78%
Avg. Income	106	108	105
% of Total Count of Married along Table (Across)	21.99%	0.24%	77.78%
% of Total Avg. Return Rate along Table (Across)	337.16%	91.89%	32.98%
% of Total Count of Work along Table (Across)	21.99%	0.24%	77.78%

Fig 4.1.1: Segment Values

This visualization presents a customer segmentation analysis. Cluster 1 consists of 1,589 customers with an average age of 61 and average income of 106 units, highly educated and married. Cluster 2, with just 17 customers, has the youngest and highest earning individuals. Cluster 3 is the largest group with 5,621 customers, mostly educated, of average age 60, with a moderate income level of 105 units. The data suggests targeting Cluster 2 for high-value engagements, while Clusters 1 and 3 offer volume and market breadth.

5.0 Conclusion

In this exploratory analysis, hierarchical and K-means clustering were applied to segment the customer base into three distinct clusters. The resultant Linear Discriminant Analysis (LDA), while indicating a high level of accuracy in cluster assignment, suggested that the discriminative strength of the chosen variables was no better than chance. ANOVA testing corroborated this by demonstrating significant discrimination for only the primary discriminant function. Nevertheless, the clusters were well-defined: Cluster 1, with the smallest customer count, presented the highest average revenue per customer, identifying it as a segment with high-value clients ideal for premium marketing initiatives. In contrast, Cluster 3's substantial customer volume but lower average revenue per customer suggests a focus on mass-market strategies might be beneficial.

6.0 References

- Cai Qiuru et al. (2012) 'Telecom Customer Segmentation based on Cluster Analysis', 2012 International Conference on Computer Science and Information Processing (CSIP) [Preprint]. doi:10.1109/csip.2012.6309069 Available at : https://ieeexplore.ieee.org/abstract/document/6309069?casa_token=olD0GX0bE-EAAAAA:KcVvKH63weCd6aj-8-K3hvPJbDLuY-dhaKBmN-IXm43lwRRQg_dPrVVE32-cQlGWu1EONPB (Accessed on 25th Feb 2023).
- Chen, Y.-S. et al. (2012) 'Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment', Computers in Biology and Medicine, 42(2), pp. 213–221. doi:Available at : https://www.sciencedirect.com/science/article/pii/S0010482511002290?ref=pdf_download&fr=RR-2&rr=85f47749ca3f35b9 (Accessed on march 2nd 2024).
- Christy, A.J. et al. (2021) 'RFM ranking – an effective approach to customer segmentation', Journal of King Saud University - Computer and Information Sciences, 33(10), pp. 1251–1257. doi:Available at : <https://www.sciencedirect.com/science/article/pii/S1319157818304178> (Accessed on march 2nd 2024).
- Kashwan, K.R. and Velu, C.M. (2013) 'Customer segmentation using clustering and data mining techniques', International Journal of Computer Theory and Engineering, pp. 856–861. doi:Available at : https://www.researchgate.net/profile/K-R-Kashwan/publication/271302240_Customer_Segmentation_Using_Clustering_and_Data_Mining_Techniques/links/57093e7908ae2eb9421e2d86/Customer-Segmentation-Using-Clustering-and-Data-Mining-Techniques.pdf (Accessed on march 2nd 2024).
- Kansal, T. et al. (2018) 'Customer segmentation using K-means clustering', 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) [Preprint]. doi:Available at : <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8769171> (Accessed on march 2nd 2024).
- Li, Y. et al. (2021) 'Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm', Applied Soft Computing, 113, p. 107924. doi:Available at : https://www.sciencedirect.com/science/article/pii/S1568494621008462?casa_token=oW6bOM9t2gQAAAAA:FEqw-HYlgR7Vuhj-rAnuJIWdVhTwtSOMxtoJJwHJvnpniKBfScPLtr6bXIYOrQTUAGZBQfQkg [Accessed on march 2nd 2024].
- Namvar, M., Gholamian, M.R., and KhakAbi, S. (2010) 'A two phase clustering method for intelligent customer segmentation', 2010 International Conference on Intelligent Systems, Modelling and Simulation [Preprint]. doi:Available at :

<https://ieeexplore.ieee.org/abstract/document/5416093> (Accessed on march 2nd 2024).

- Punhani, R. et al. (2021) 'Application of clustering algorithm for effective customer segmentation in e-commerce', 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) [Preprint]. doi:Available at : <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9410713> (Accessed on march 2nd 2024).
- Schröer, C., Kruse, F., and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', *Procedia Computer Science*, 181, pp. 526–534. doi: Available at: <https://www.sciencedirect.com/science/article/pii/S1877050921002416> (Accessed on: 8th march 2023).
- Shirole, R., Salokhe, L., and Jadhav, S. (2021) 'Customer segmentation using RFM model and K-means clustering', *International Journal of Scientific Research in Science and Technology*, pp. 591–597. doi:Available at : <https://ijsrst.com/paper/8152.pdf> (Accessed on march 2nd 2024).
- Sun, Z.-H. et al. (2021) 'GPHC: A heuristic clustering method to customer segmentation', *Applied Soft Computing*, 111, p. 107677. doi:Available at : https://www.sciencedirect.com/science/article/pii/S1568494621005986?casa_token=kBUZFW6ZYvsAAAAA:WZFqVE0H7WCqpte0qQwgje6Dv0EcacMsiMiHDU0wQM2Aa3Wyh_qjaknC_9rQEalHy60bRbRsQ (Accessed on march 2nd 2024).
- Tripathi, S., Bhardwaj, A., and E, P. (2018) 'Approaches to clustering in customer segmentation', *International Journal of Engineering & Technology*, 7(3.12), p. 802. doi:Available at : https://www.researchgate.net/publication/326706602_Approaches_to_Clustering_in_Customer_Segmentation/link/5b602d6b458515c4b2548801/download?tp=eyJjb250ZXh0Ijp7ImZpcnNOUGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19 (Accessed on march 2nd 2024).
- Tsai, C.-Y. and Chiu, C.-C. (2004) 'A purchase-based market segmentation methodology', *Expert Systems with Applications*, 27(2), pp. 265–276. doi:10.1016/j.eswa.2004.02.005 Available at: https://www.sciencedirect.com/science/article/pii/S0957417404000132?casa_token=L0N2n3lnO84AAAAA:cgs02PBLZjcYJpY7HG8F1oxZaM_vzp6r78Muz6gRvVkvVqKIGzWvtipRo9SXnMul1miQOyjY (Accessed on 2nd March 2024).
- WOO, J., BAE, S., and PARK, S. (2005) 'Visualization method for customer targeting using customer map', *Expert Systems with Applications*, 28(4), pp. 763–772. doi:Available at : https://www.sciencedirect.com/science/article/pii/S0957417404001800?casa_token=ZB8XX2xKCDcAAAAA:0KQhIO9-

EDOOxM9x xv40JWVrZ 059FTn9FPO Z8eKaLZN0VQIHUfvN7rRTYFGxrWEc8AGCarw
(Accessed on 1st March 2024).

- Yang, J. et al. (2016) 'Buyer targeting optimization: A unified customer segmentation perspective', 2016 IEEE International Conference on Big Data (Big Data) [Preprint]. doi:Available at : <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7840730> (Accessed on March 2nd 2024).
- Yang, Z., Cao, S., and Yan, B. (2011) 'Using linear discriminant analysis and data mining approaches to identify e-commerce anomaly', 2011 Seventh International Conference on Natural Computation [Preprint]. doi:Available at : <https://ieeexplore.ieee.org/document/6022591> (Accessed on march 2nd 2024).

7.0 Appendix

```
#data explorartion
```

```
library(tidyverse)
```

```
library(ggplot2)
```

```
getwd()
```

```
setwd('/Users/dhanush/Desktop/Bussiness analytics /Sem 2/Marketing Analytics/Ass_1')
```

```
data <- read_csv('/Users/dhanush/Desktop/Bussiness analytics /Sem 2/Marketing  
Analytics/Ass_1/Assignment1-data (1).csv')
```

```
names(data)
```

```
attach(data)
```

```
summary(InvoiceNo)
```

```
library(ggplot2)
```

```
#Quantity
```

```
library(ggplot2)
```

```
ggplot(data, aes(x = Quantity)) +  
  geom_boxplot(outlier.color = "red") +  
  labs(title = "Box Plot of Quantity") +  
  theme_minimal()
```

```
#Unit Price
```

```
ggplot(data, aes(x = `UnitPrice`)) +  
  geom_boxplot(outlier.color = "red") +  
  labs(title = "Box Plot of Unit Price") +  
  theme_minimal()
```

```
#Return Rate
```

```
ggplot(data, aes(x = `ReturnRate`)) +  
  geom_boxplot(outlier.color = "red") +  
  labs(title = "Box Plot of Return Rate") +  
  theme_minimal()
```

```
#Age
```

```
ggplot(data, aes(x = Age)) +  
  geom_boxplot(outlier.color = "red") +  
  labs(title = "Box Plot of Age") +  
  theme_minimal()
```

```
#Income
```

```
ggplot(data, aes(x = Income)) +
  geom_boxplot(outlier.color = "red") +
  labs(title = "Box Plot of Income") +
  theme_minimal()
```

```
#marriage
ggplot(data, aes(x = Married, fill = factor(Married))) +
  geom_bar(width = 0.75, fill = c("red", "black")) +
  labs(title = "Marital Status", x = "Married", y = "Count") +
  theme_minimal()
```

```
# Convert 'Work' variable to factor
n_data<- data
n_data$Work <- factor(n_data$Work)
```

```
ggplot(n_data, aes(x = Work, fill = Work)) +
  geom_bar() +
  scale_fill_brewer(palette = "Set3") +
  labs(title = "Distribution of Work Categories", x = "Work", y = "Count") +
  theme_minimal()
```

```
#Education
ggplot(data, aes(x = Edcation, fill = factor(Edcation))) +
  geom_bar(fill = c("blue", "lightgreen", "pink")) +
  labs(title = "Distribution of Education Levels", x = "Education", y = "Count") +
  theme_minimal()
```

```
install.packages("factoextra")
install.packages("NbClust")
install.packages("fpc")
library(tidyverse, warn.conflicts = FALSE, quietly = TRUE)
library(lubridate)
library(factoextra)
library(NbClust)
```

```
library(fpc)
library(cluster)
library(SmartEDA)
library(psych)
library(ggplot2)
library(dplyr, warn.conflicts = FALSE, quietly = TRUE)
```



```

# Reading the data
data <- read_csv(file.choose())
View(data)

#####
#####
get_mode <- function(x) {
  tab <- table(x)
  mode <- names(tab)[which.max(tab)]
  return(mode)
}

# Imputating mean for the numeric data and mode for the categorical data.
data <- data %>%
  group_by(CustomerID) %>%
  mutate(Age = round(mean(Age)),
         Income = round(mean(Income)),
         Married = get_mode(Married),
         Work = get_mode(Work),
         Education = get_mode(Education),
         ZipCode = get_mode(ZipCode)) %>%
  ungroup()

# A) Data Preparation and understanding.
# A.1) Omitted the CustomerID.

Missing_values_Columns <- colSums(is.na(data))
barplot(Missing_values_Columns)
data <- na.omit(data)

# A.2) Return Rate with greater than 1 will be omitted.
data <- data %>% filter(ReturnRate < 1)

# A.3) change the data type of the characters.
data$Work <- as.factor(data$Work)
data$Married <- as.factor(data$Married)
data$Education <- as.factor(data$Education)
data$ZipCode <- as.factor(data$ZipCode)

# A.3.1) Created the variables year and month to include it in the base variables.
data$Year <- year(data$InvoiceDate)
data$Month <- month(data$InvoiceDate)

# B) Descriptive Statistics.
str(data)

```

```
#####
#####
# C) Cluster Analysis.
# C.1) Finding the Optimal Number of cluster using Base Descriptor.

base_variables <- data %>% dplyr::select(Quantity, ReturnRate, UnitPrice)
base_var_scale <- scale(base_variables)

### Hclust
hir_complete <- hclust(dist(base_var_scale, method = "euclidean"), method = "complete")
hir_centroid <- hclust(dist(base_var_scale, method = "euclidean"), method = "centroid")
hir_avg <- hclust(dist(base_var_scale, method = "euclidean"), method = "average")
hir_single <- hclust(dist(base_var_scale, method = "euclidean"), method = "single")

y_complete <- sort(hir_complete$height, decreasing = TRUE)[1:10]
y_centroid <- sort(hir_centroid$height, decreasing = TRUE)[1:10]
y_avg <- sort(hir_avg$height, decreasing = TRUE)[1:10]
y_single <- sort(hir_single$height, decreasing = TRUE)[1:10]
x <- c(1:10)

# Create the data frame
df <- data.frame(x = x,
  y_complete = y_complete,
  y_centroid = y_centroid,
  y_avg = y_avg,
  y_single = y_single)

# Visualising the results
# Define aesthetically pleasing colors
my_colors <- c("#1f78b4", "#33a02c", "#e31a1c", "#ff7f00")

# Plot with updated colors
ggplot(df) +
  geom_line(aes(x = x, y = y_complete, color = "Complete")) +
  geom_line(aes(x = x, y = y_centroid, color = "Centroid")) +
  geom_line(aes(x = x, y = y_avg, color = "Average")) +
  geom_line(aes(x = x, y = y_single, color = "Single")) +
  geom_point(aes(x = x, y = y_complete), color = "#1f78b4") +
  geom_point(aes(x = x, y = y_centroid), color = "#33a02c") +
  geom_point(aes(x = x, y = y_avg), color = "#e31a1c") +
  geom_point(aes(x = x, y = y_single), color = "#ff7f00") +
  scale_color_manual(values = my_colors) +
  labs(color = "Linkage Method", title = "Optimal No. of Clusters",
```

```

    x = "No. of Clusters", y = "Distance") +
scale_x_continuous(breaks = 1:10)+
theme_minimal()

```

```

plot(hir_complete)

```

D) After the thorough consideration we have come to a conclusion that the optimal number of clusters for our data set is 6.

Hence the optimal no. of cluster is 3.

D.1) KMeans algorithm.

```

set.seed(40412492)

```

```

km.res <- kmeans(x = base_var_scale, centers = 3, iter.max = 500, nstart = 1000)

```

```

table(km.res$cluster)

```

```

km.res_cluster <- fviz_cluster(km.res,
                               base_var_scale,
                               ellipse = TRUE,
                               star.plot = TRUE,
                               ggtheme = theme_minimal(),
                               palette = "jco",
                               ellipse.type = "norm") +
labs(title = "KMeans Clustering")

```

```

km.res_cluster

```

```

#####
#####

```

cluster is added as the column in the data and the new data is named as data_clustered.

```

data_clustered <- cbind(data, cluster = km.res$cluster)

```

Assuming 'data_clustered' is your dataframe

```

write.csv(data_clustered, file = file.choose(new = TRUE), row.names = FALSE)

```

E) Discriminant Analysis

```

library(MASS)

```

To check which discriminant functions are significant.

We are running the classification algorithm LDA to predict the classes using the Descriptor variables.

```

fit <- lda(cluster ~ Married + Age + Income + Work + Education, data = data_clustered)
fit

ldaPred <- predict(fit, data_clustered)
ld <- ldaPred$x

fit.predict <- ldaPred$class

#confusion Matrix
confusionMatrix(fit.predict, as.factor(data_clustered$cluster))

#ANOVA
anova(lm(ld[,1] ~ data_clustered$cluster ))
anova(lm(ld[,2] ~ data_clustered$cluster ))

summary_revenue <- data_clustered %>% group_by(cluster) %>%
  summarise(total_Customers = n(),total_profits = sum(Quantity*UnitPrice)) %>%
  mutate("Revenue/Segment" = total_profits/total_Customers)

summary_revenue

## Check Discriminant Model Fit
pred.seg <- predict(fit)$class
tseg <- table(data_clustered$cluster, pred.seg)
tseg # print table
sum(diag(tseg))/nrow(data_clustered) # print percent correct

#####
#####
# RFM

library(rfm)
analysis_date <- lubridate::as_date("2022-01-31")

rfm <- data %>% group_by(CustomerID) %>%
  summarise(total_trans = n(),
            total_revenue = sum(UnitPrice*Quantity),
            recency_days = min(analysis_date - date(InvoiceDate)))

rfm$recency_days <- as.numeric(rfm$recency_days)
colnames(rfm)[1] <- "customer_id"

rfm_table <- rfm_table_customer(data = rfm,
                                customer_id = customer_id,
                                n_transactions = total_trans,

```

```

        recency = recency_days,
        total_revenue = total_revenue,
        analysis_date = analysis_date,
        recency_bins = 5,
        frequency_bins = 5,
        monetary_bins = 5)

table(rfm_table$rfm$recency_score)
table(rfm_table$rfm$frequency_score)
table(rfm_table$rfm$monetary_score)

# Write the RFM table to a CSV file
write.csv(rfm_table$rfm, "rfm_analysis_results.csv", row.names = FALSE)

plot(rfm)

rfm_data <- read_csv('rfm_analysis_results.csv')

library(ggplot2)

# Recency Score Distribution
ggplot(rfm_table$rfm, aes(x = recency_score)) +
  geom_bar(fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Recency Scores", x = "Recency Score", y = "Count")

# Frequency Score Distribution
ggplot(rfm_table$rfm, aes(x = frequency_score)) +
  geom_bar(fill = "green", alpha = 0.7) +
  labs(title = "Distribution of Frequency Scores", x = "Frequency Score", y = "Count")

# Monetary Score Distribution
ggplot(rfm_table$rfm, aes(x = monetary_score)) +
  geom_bar(fill = "red", alpha = 0.7) +
  labs(title = "Distribution of Monetary Scores", x = "Monetary Score", y = "Count")
#####
# Scatter plot of Frequency vs Monetary
ggplot(rfm_table$rfm, aes(x = frequency_score, y = monetary_score)) +
  geom_point(aes(color = recency_score), alpha = 0.6) +
  scale_color_gradient(low = "yellow", high = "red") +
  labs(title = "Frequency vs. Monetary by Recency Score",
        x = "Frequency Score",
        y = "Monetary Score",
        color = "Recency Score")
#####

```