MERCER UNIVERSITY
SCHOOL OF BUSINESS

PROFILING HOTEL CUSTOMERS: AN APPLICATION OF RFM ANALYSIS AND CLUSTERING METHOD

Nhan Nguyen

I. INTRODUCTION

In a competitive industry, understanding customer behaviors and providing services to fit individual needs to retain old customers has become critical for long-term success in the continuously changing hotel sector landscape. The business must apply useful strategies to maximize customer needs and satisfaction which will contribute to business success in terms of profit. Therefore, Customer Relationship Management (CRM) is introduced to help marketers build the relationship between business and customer. One way to enhance CRM strategies, businesses should apply data mining techniques to understand patterns of customer buying behavior by profiling customer segmentation. In this research, RFM analysis and clustering method will be applied to profiling customer segmentation. By applying these techniques, businesses will understand each group of customers' purchasing behavior and their spending patterns. From that, businesses will customize more effective marketing strategies for each targeted group of customers such as loyal, potential, and risk customers.

II. LITERATURE REVIEW

Customer segmentation is essential for hotel to customize their marketing strategies for each targeted group customer. According to Frochot & Morrison (2000), customer segmentation is the division of market into different groups of consumers based on demographics, socioeconomic, and geographic which are frequently used variables to segment market. And the method widely used to segment customers based on buying behavior is RFM analyis. According to Christy et al. (2021), RFM analysis is used to divide customer into distinct groups based on three dimensions: rencency (R), frequency (F) , monetary (M). Recency is the last time a customer made a purchase, Frequency is the number of purchases that a customer made in a specific period and Monetary is total amount of spending by a customer. Several studies have shown the effectiveness of combination of RFM analysis and clustering method for customer segmentation. By applying RFM analysis and clustering method, Cengizci, A. D., & Caber, M. (2016) identified eight distinct groups of customers based on rencency, frequency, and monetary values. Additionally, Doğan and Ayçin (2018) proposed two separate clustering models utilizing RFM values, highlighting the limitations of solely relying on spending data for customer segmentation. These findings suggests that combination of RFM analysis and clustering method is an valuable approach to segment customers effectively. This allows businesses can develop personalized marketing strategies, improve customer churn and loyalty.

III. DATA

3.1 Data Source

The dataset is taken from "A Hotel's customers personal, behavioral, geographic dataset from Lisbon, Portugal (2015 – 2018) ". This data collected customer personal, behavioral, demographic, and geographical information of customers in three years.

3.2 Data dictionary:

The dataset consists of 31 variables with 83,590 observations. However, in this research, only 4 variables out 31 are chosen for the purpose of doing RFM analysis and clustering method. That is "DaysSinceLastStay", "BookingsCheckedIn" , "LodingRevenue", and "OtherRevenue". The description of each variable will be displayed in Figure 1.

3.3 Data Preprocessing

In order to ensure the precision of model, a comprehensive data preprocessing phase was undertaken. Firstly, mode imputation was employed to address the missing value of "Age" variable, because the "Age" variable is likely to be skewed, and the mode represents the most frequently occurring value in the dataset. By replacing missing values with mode method, it will preserve the central tendency of Age distribution. Besides, there is a large number of observations where "DaySinceLastStay" has value -1 which indicates that these customers never stayed in the hotel, in order to avoid potential outliers in the dataset, all of rows containing "DaySinceLastStay" -1 will be removed from the dataset. Furthermore, there is a combination of two columns "LodgingRevenue" and "OtherRevenue" into a new variable "TotalRevenue", it will facilitate for advantageous dataset for model development.

3.4. Exploratory Data Analysis

- The box plot of variable "Age" which displayed in Figure 2 show the outliers such as negative values which does not make sense for "Age". So, in order to remove those outliers, Interquartile Range (IQR) method was deployed. And Figure 3 shows the new box plot of Age without outliers with the range around from 1 to 88.

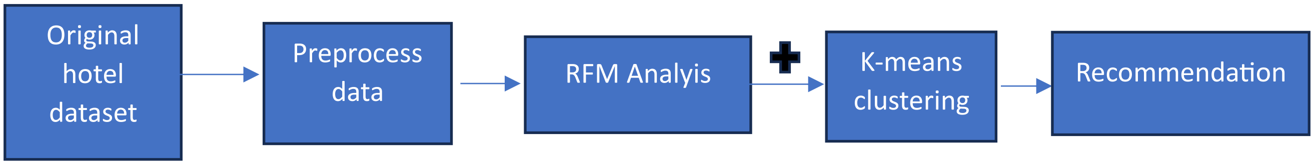
- The scatter plot of number of bookings and total revenue which displayed in Figure 4 shows that there is no obvious relationship between these two variables.

- Figure 5 which shows the bar plot of distribution channel. It can be observed that the most of customers book for the hotel from agent or operator. And it was followed by direct booking was accounted for approximately 10000 customers.

3.5. Data partition

The data will be partitioned into 70% for training dataset with 44569 observations and 30% for testing dataset with 19101 observations. Data partition is a very crucial step to ensure model perfomance, prevent overfitting, and provide a reliable evaluation of model's ability to generalize to new and unseeen data.

IV. METHODOLOGY



4.1. RFM Analysis

The RFM analysis was introduced by Bult and Wansbeek (1995) which proved its effectiveness in marketing databases. The RFM analysis is a powerful tool that helps business improve customer segmentation by dividing customers into different groups for better identifying which customers group will be potential. From that, business can have better marketing strategies for target customers. The analysis will be scored by three factors: Recency(R) - Time since last purchase, Frequency(F) - Total number of purchases and Monetary (M) - Total monetary value. The smaller recency value, the more a customer recently visited to hotel. While the higher value of frequency, the higher number of bookings that a customer makes. Similarly, the higher monetary value, the more revenue the hotel earned.

The segmentation process initiates by focusing on recency, evaluating the time since the last purchase. Customers are sorted from the most recent to the least recent, and then divided into quintiles. The top 20% receive a recency score of 5, the next 20% a score of 4, and so forth. This same approach is then applied to frequency, with customers ranked from most to least frequent. The top 20% in terms of frequency are assigned a score of 5, and the remaining quintiles receive scores of 4, 3, 2, and 1. The process is repeated for monetary value. Subsequently, all customers are ranked by concatenating their recency, frequency, and monetary values. Based on the quintile system, the segmentation will be assigned by 125 scores (5x5x5), from the highest as segment 555 to the lowest 111. Each segmentation will be explained by each score of R-F-M, for example, the segment 555 which would be a group customer with the most recently purchase, most frequently, and spend the most amount of money on goods and services.

The RFM result table (figure 6) and distribution chart (figure 7) reveals distinct customer segments, with Segment 311 standing out as the largest group, encompassing 15,154 customers. This segment is characterized by the medium recent stays, lowest frequent visits, and lowest monetary spending. Following closely is Segment 511, comprising customers who have most recent stays, the lowest visits, and lowest amount of money spent. Meanwhile, Segment 111, with 10,231 customers, represents the group with the lowest RFM scores, indicating lowest recent stays, infrequent visits, and low monetary value. Notably, other segments such as 531 and 551 exhibit negligible customer numbers. Overall, these findings provide valuable insights into customer behavior, allowing for targeted strategies to engage and retain customers across different segments.

4.2. K-means clustering

K-means is an algorithm where inputs as specified parameters and the number of clusters so that maximize the similarity within each cluster. K-means operates through an iterative method where centroids are computed prior to each iteration. During each iteration, data points are reassigned to different clusters based on the centroids calculated at that point. This iterative process continues until no result in decrease in overall sum (Christy et al., 2021).

Before running the model, the dataset with recency, frequency, and monetary values will be standardized with Z-score standardization method. Standardization is necessary preprocessing steps to ensure all values of different variables in the same scale. Next, the Elbow method is applied to determine the optimal number of cluster (k). The method will run clustering algorithms on the dataset for a range of k (starting from k = 1 to kmax) and for each value of k, the algorithm will calculate the within sum of squared (WSS) errors. The elbow method will plot number of clusters (k) against WWS values. It suggest that the optimal point where the rate of decrease in WWS slows down drastically. As a result, it can be observed from figure 8, the elbow method suggest that the optimal numer of clusers is where k = 4. So, the k-means algorithms will perfom k-means with optimal value for k = 4. After running k-means algorithms, each segment will be assigned to each cluster based on recency, frequency, monetary values. In the 3D scatter plot which showed in figure 9, cluster 1 and 4 have high Recency and cluster 1 have highest the monetary value. While cluster 2, 3 have low Recency, and cluster 3 have highest frequency value.

From the table result in figure 10 , the cluster 1 indicates the highest Recency value , which means that the last time of this group customer stayed in the hotel average 777 days ago, not recently. And this group moderate Frequency, with average 29 number of check-in bookings. Remarkably, this group generates the highest total revenue for the hotel. Based on these buying behavior characteristics, this group customer can be classified as "*can't lose them*" segment.

-Cluster 2: this group customer shares the similarity in terms of lowest recency, which means that the time they stayed at the hotel recently, low frequent visit average 9 numbers of bookings and low total revenue generated. These customers can be segmented as "*potential group*" of customers.

-Cluster 3: this group customer recently active with low recency value, and highest frequent visit with average 54 bookings, and moderate value of revenue. These customers could be considered as "*loyal customers*" group.

- Cluster 4: this group customer with high Recency which indicates the last time they stayed at the hotel was really long time ago, and lowest frequent visit and low monetary spending, average $4000. This group is "*at risk/ lost*"" customers, further actions needed to revive their interest.

V. MARKETING STRATEGIES RECOMMENDATION

Through the purchasing behavior pattern of customers thorugh RFM analysis, we can accurately tailor marketing strategies to each targeted group customers:

Group customer 1 with moderate frequent visit brings the highest revenue for the hotel, however, they have not returned for long time. So, the hotel should give them promotion, discount program to encourage them to return. Moreover, tailoring their preferences and past booking history to enhance their loyalty.

Group customer 2 recently visited to the hotel, however low frequent visit and lowest spending. So, the hotel should introduce attractive packages with promotion to encourage them return and enhance their spending. Besides that, we should collect feedback from them, so we can understand their needs and preferences, so we can adjust accordingly.

Group customer 3 recently active, with highest frequent visit to the hotel and moderate spending. So, this group with loyal customers should be treated like VIP by introducing exclusive services, give them rewards through loyalty program.

Group customer 4 is the most inactive group with lowest frequent visit and low monetary spending. So, the hotel should understand their disengagement through the survey, feedback. Besides, the promotion or exclusive incentives for this group should be introduced to encourage their return.

VI. CONCLUSION

In conclusion, customer segmentation is a necessary method to profile customers into distinct groups. In this paper, RFM analysis and k-means cluster method was applied to identify four distinct groups with recency, frequency, and monetary values. By leverage these techniques, we can identify the most and the least profitable customer segments .Consequently, these findings can provide insights for crafting effective marketing strategies tailored to each identified group, as hightlighted in the paper's recommendation . However, it is essential to acknowledge that the limitation of this paper, which is relied on RFM values for segmentation. Therefore, futher investigations should aim to other factors to have more comprehensive understanding of customer behavior and preferences.

REFRENCES

1. Frochot, I., & Morrison, A. M. (2000). Benefit Segmentation: A review of its applications to travel and tourism research. *Journal of Travel & Tourism Marketing*, *9*(4), 21–45. https://doi.org/10.1300/j073v09n04_02

2. Cengizci, A. D., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, *18*, 153–160. https://doi.org/10.1016/j.tmp.2016.03.001

3. Doğan, O., Ayçın, E., & Bulut, Z. A. (2018). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: a CASE STUDY IN RETAIL INDUSTRY. *International Journal of Contemporary Economics and Administrative Sciences*, *8*(1),19

4. Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science, 14(4), 378–394. https://doi.org/10.1287/mksc.14.4.378

5. Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021b). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

APPENDIX

Variables	**Type**	**Description**
BookingsCheckedIn	Numeric	Number of bookings that customer made
DaysSinceLastStay	Numeric	The number of days elapsed between the last dat of the extraction and customer's last arrival data
OtherRevenue	Numeric	Total amount spent on other expense

LodgingRevenue	Numeric	Total amount spent on lodging expenses

Figure 1: Data dictionary

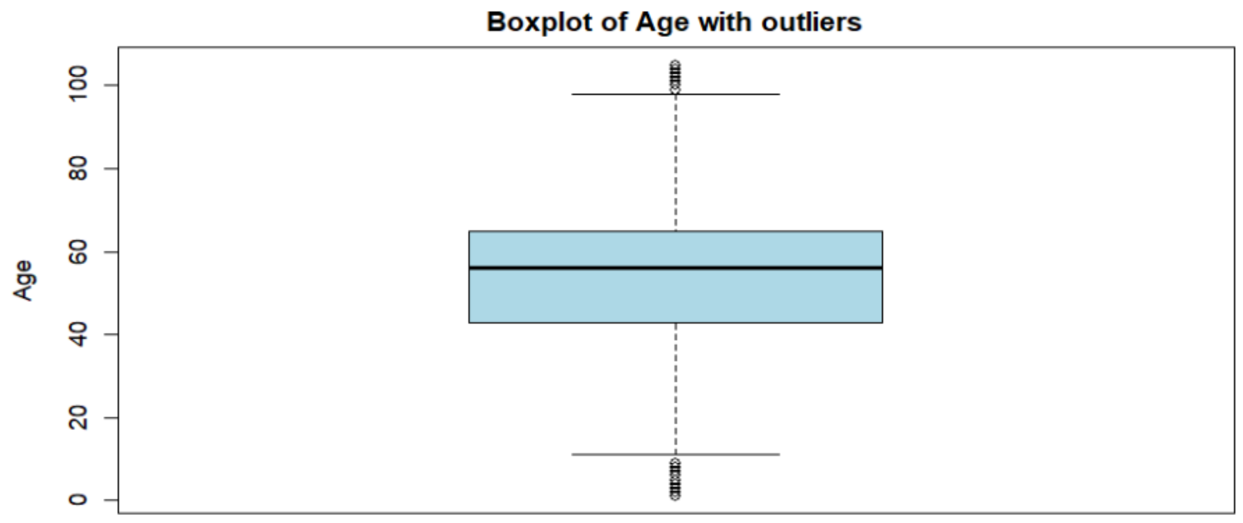


Figure 2: The box plot of Age with outliers

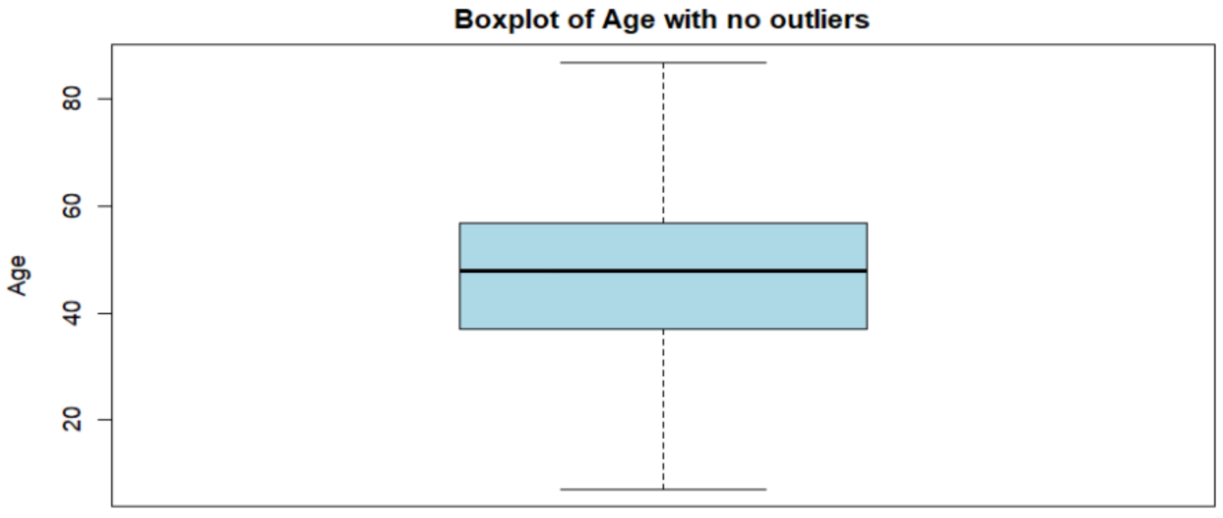


Figure 3: The box plot of with no outliers

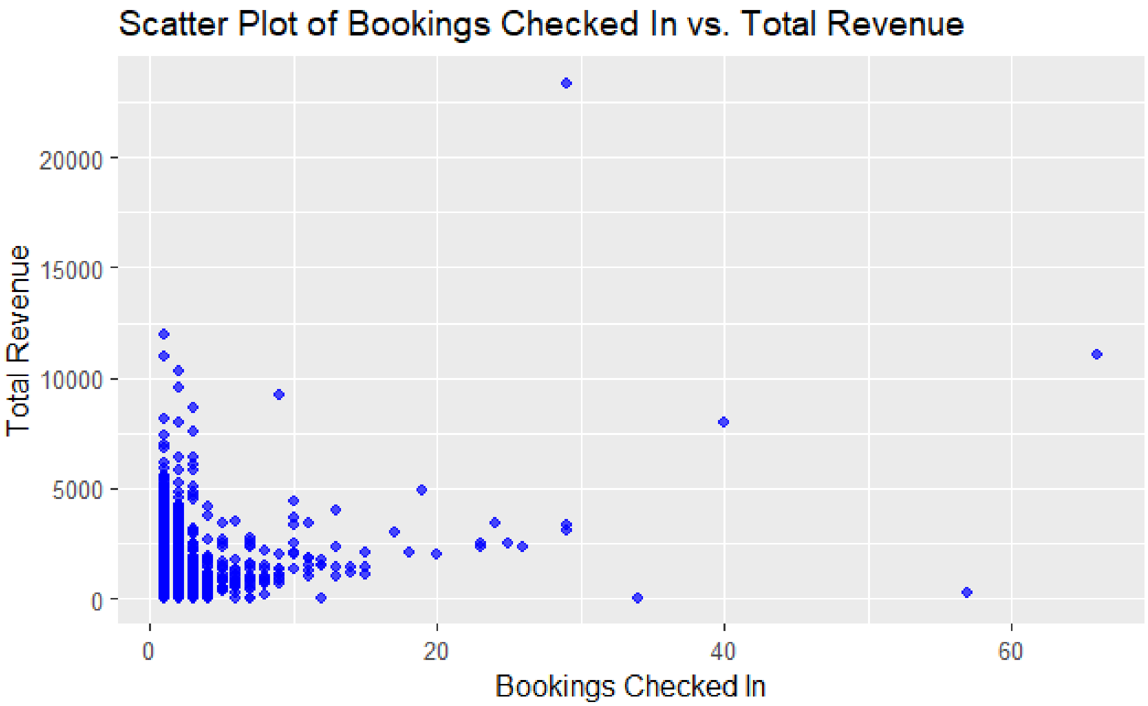


Figure 4: Scatterplot between bookings and total revenue

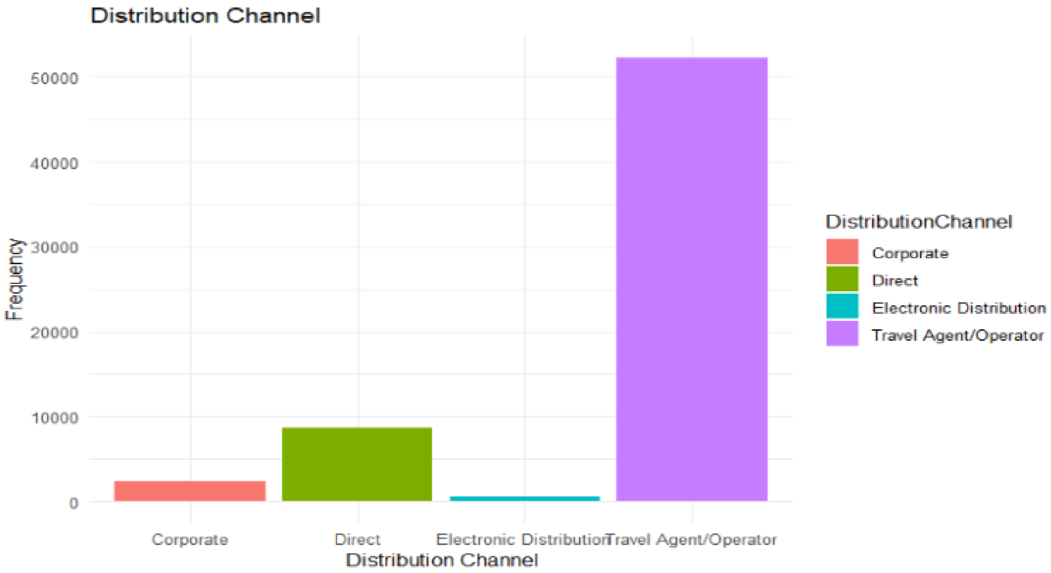
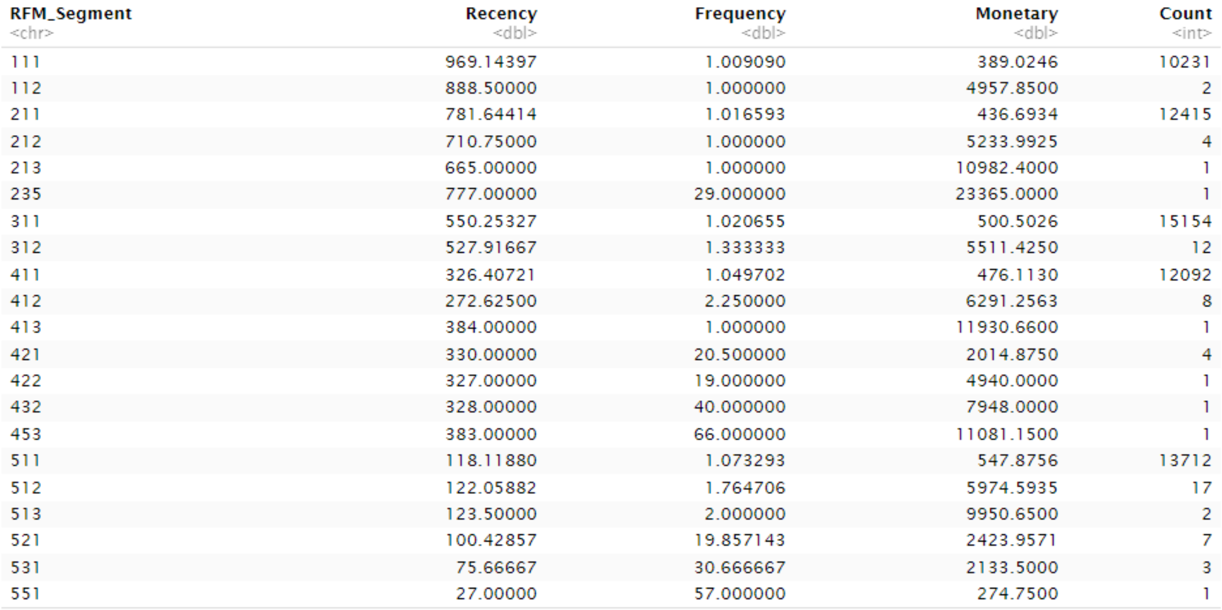


Figure 5: Distribution channels



RFM_Segment	Recency	Frequency	Monetary	Count
<chr>	<dbl>	<dbl>	<dbl>	<int>
111	969.14397	1.009090	389.0246	10231
112	888.50000	1.000000	4957.8500	2
211	781.64414	1.016593	436.6934	12415
212	710.75000	1.000000	5233.9925	4
213	665.00000	1.000000	10982.4000	1
235	777.00000	29.000000	23365.0000	1
311	550.25327	1.020655	500.5026	15154
312	527.91667	1.333333	5511.4250	12
411	326.40721	1.049702	476.1130	12092
412	272.62500	2.250000	6291.2563	8
413	384.00000	1.000000	11930.6600	1
421	330.00000	20.500000	2014.8750	4
422	327.00000	19.000000	4940.0000	1
432	328.00000	40.000000	7948.0000	1
453	383.00000	66.000000	11081.1500	1
511	118.11880	1.073293	547.8756	13712
512	122.05882	1.764706	5974.5935	17
513	123.50000	2.000000	9950.6500	2
521	100.42857	19.857143	2423.9571	7
531	75.66667	30.666667	2133.5000	3
551	27.00000	57.000000	274.7500	1

Figure 6: Distribution of RFM segments table

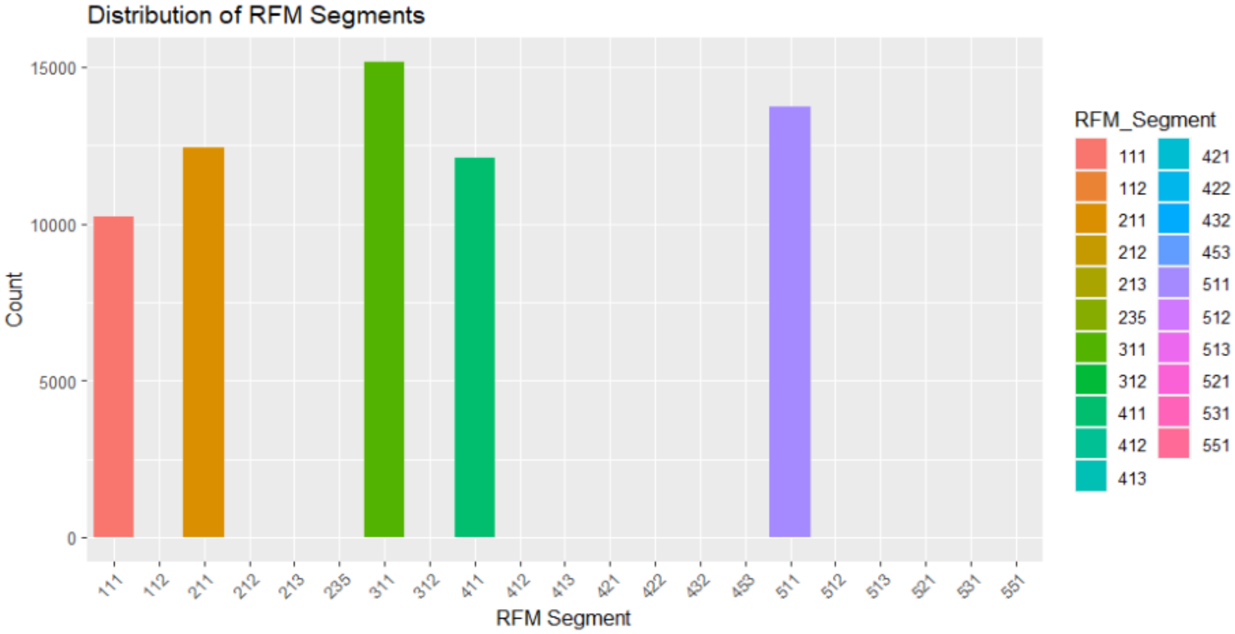


Figure 7: Distribution chart of RFM segments

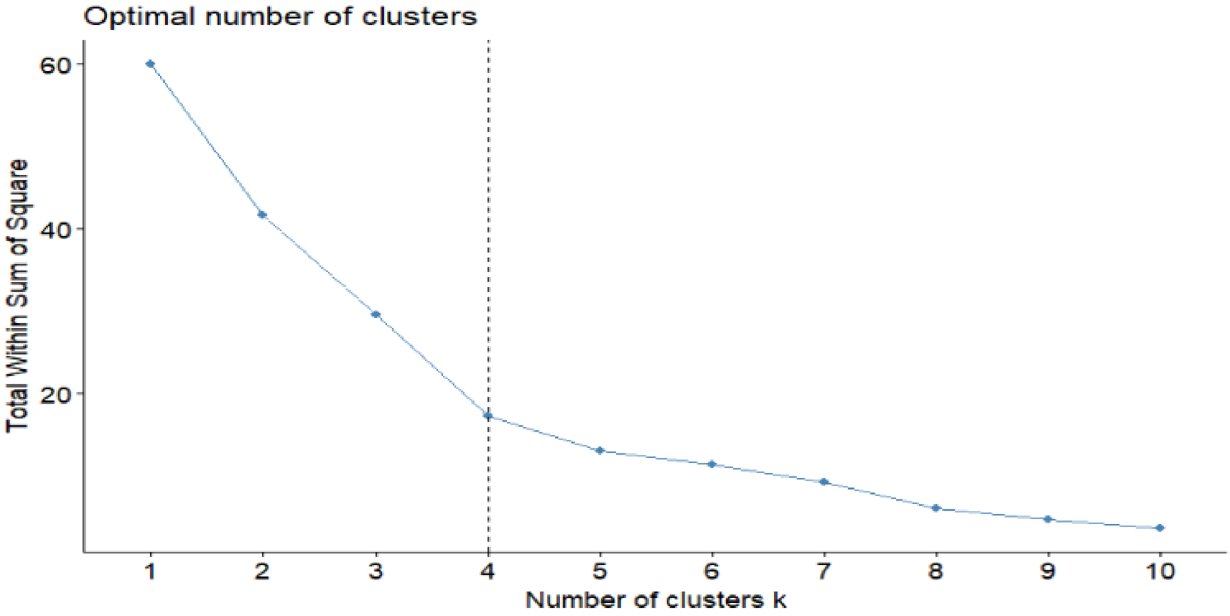


Figure 8: Elbow graph

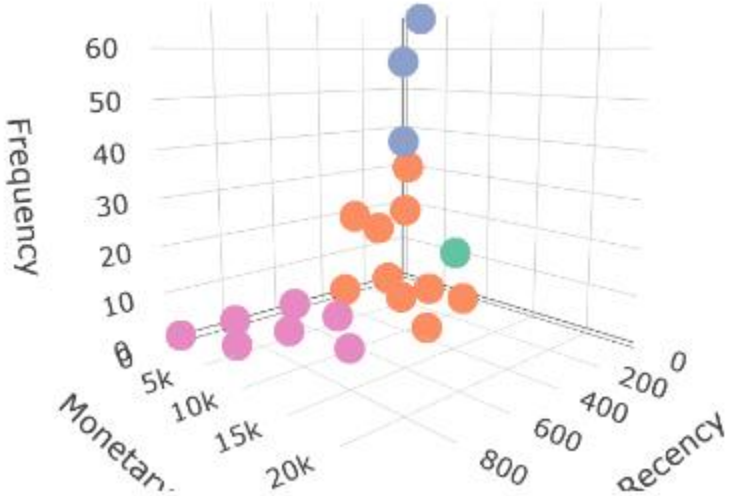
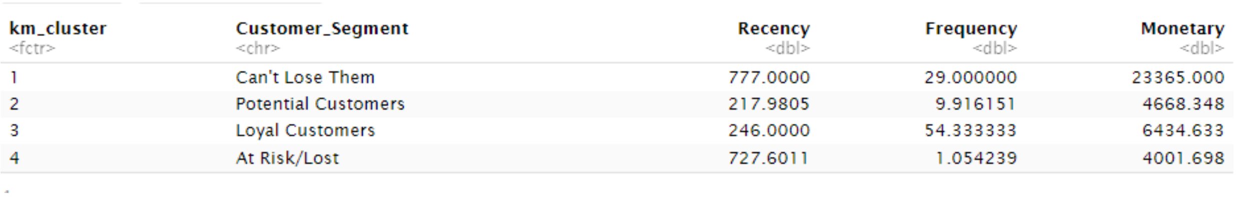


Figure 9: 3D scatter plot between frequency, monetary, recency



km_cluster <fctr>	Customer_Segment <chr>	Recency <dbl>	Frequency <dbl>	Monetary <dbl>
1	Can't Lose Them	777.0000	29.000000	23365.000
2	Potential Customers	217.9805	9.916151	4668.348
3	Loyal Customers	246.0000	54.333333	6434.633
4	At Risk/Lost	727.6011	1.054239	4001.698

Figure 10: K-means clustering result table