Customer Segmentation: Clustering and RFM Analysis

MSiA421 - Final Project
Group 9: Ziqiao Ao, Xin Shu, Ke Xu, Yuwen Meng

Abstract

Through the analysis of customer personalities and the identification of distinct customer segments, businesses can gain a deeper understanding of their customer base, allowing them to customize their products to meet the specific requirements, behaviors, and concerns of each segment. This study utilizes a publicly available customer behavior dataset from Kaggle to perform customer segmentation and provide valuable insights into the business side. Our approach involves two key methods: 1) dimensionality reduction and general KMeans clustering of the provided features, and 2) RFM analysis to cluster customers based on their purchasing behavior. The clustering results indicate that customers could be grouped into three clusters regarding their demographic and behavioral characteristics. RFM analysis grouped customers into Diamond, Gold, Silver, and Bronze groups. Each group also represented specific demographic and behavioral characteristics. Marketing strategies are recommended accordingly for each segment within both methods.

1. Literature Review

Customer segmentation is a popular approach to help businesses make better decisions by understanding their customers better and generating tailored strategies accordingly. Numerous researchers applied various segmentation techniques across different industries. Most of them leveraged features that are relevant to customer values to create clusters, such as the RFM features. Others extended the feature selection to age, gender, income, education level, and other demographic information.

Conducting customer segmentation using RFM and other features related to customer values is widely applied. The RFM model directly connects to short-term and long-term profitability by incorporating both the spending and lifetime of the customers. Dogan et al. (2018) used RFM values to segment customers in the sports retail industry. He took two approaches, including a 2-step model using log-likelihood to measure distance and BIC as clustering criteria and a k-means clustering method on RFM features. The resulting clusters improved the company's membership structure which was solely built on spending and facilitated marketing, pricing, and other crucial business strategies. Similar research was conducted in the online retail business. Chen et al. (2012) performed k-means on RFM and enhanced the clustering analysis phase by creating nested segments within the cluster using the decision tree. The authors highlighted the importance of performing detailed analysis on resulting clusters, such as studying popular products purchased in each cluster or identifying customers who have the potential to evolve into the best customers. In 2006, Kim et al. published a paper that proposed a customer LTV (lifetime value) model and tested its effect on a wireless communication company. The authors listed three approaches, including considering only

LTV scores, all components of LTV scores, and both LTV scores and other features. The LTV score was composed of current value, potential value, and customer loyalty. Again, the decision tree was used to visualize how customers' portraits were associated with their LTV scores, which facilitated the following strategy establishment.

Other research branched out to the incorporation of other relevant information, such as demographic features and geographic locations. To resolve the curse of dimensionality, dimensional reduction techniques, such as PCA are integrated into the analysis. D. Zakrzewska and J. Murlewski (2006) applied k-means, two-phase clustering (combining k-means and agglomerative hierarchical clustering), and DBSCAN to the bank business. They tested each algorithm's performance in high dimensionality, outlier detection, and scalability. K-means is proven as efficient for the dataset with large dimensions. Azad Abdulhafedh (2021) applied k-means and agglomerative hierarchical clustering to a credit card company with the PCA technique. The dataset has 18 features including RFM related features and other credibility related features. PCA could help both algorithms to detect more clusters or patterns hidden in the data, and thus improve the clustering performance.

2. Model Approach

The data science pipeline of this project is given in Figure 1. A general clustering algorithm on given variables and RFM analysis will be utilized to perform customer segmentation. Before clustering, we will conduct dimensionality reduction first and compare the performance of different methods (PCA, Tsne, and Umap). Recency, frequency, and monetary features are generated using the provided features for RFM analysis, a more detailed process will be introduced in the following sections.

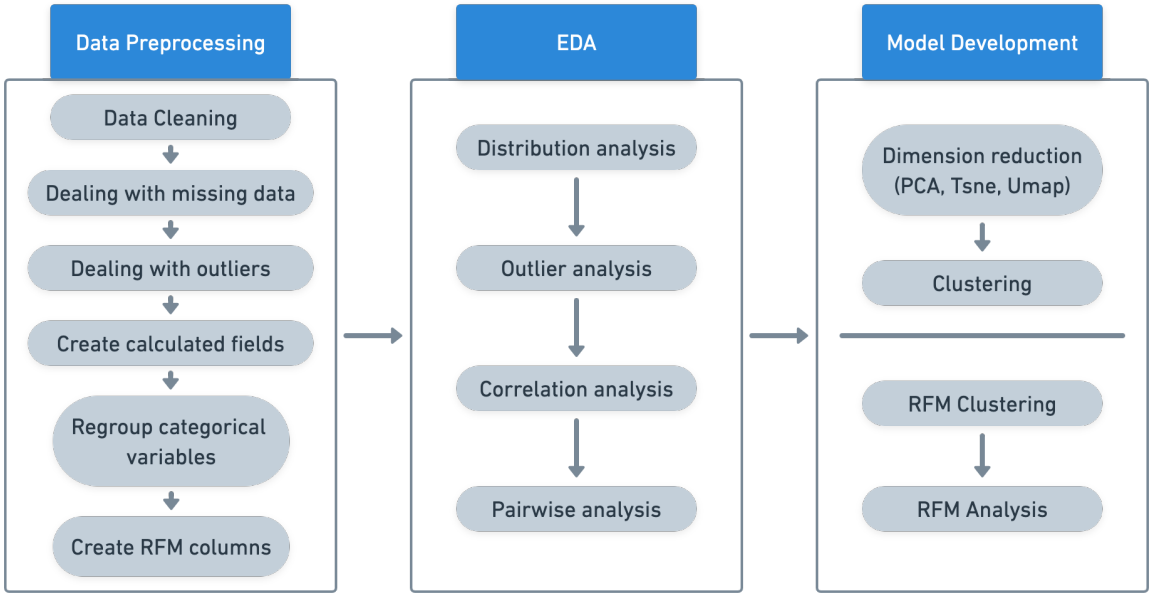


Figure 1. Data Science pipeline

3. Data Preprocessing and EDA

The data has 29 columns and 2240 rows in total. The dataset includes demographic features (such as income, education, and marital status), RFM features (recency, spending, and orders on various products), and customers' responses to campaigns. Before modeling, we dropped the 24 missing data in the income column and the outliers in income, year_birth, and MntMeatProducts that may have huge influences on the date. After cleaning, we have 2204 rows in total. Further, we created calculated variables age, number of accepted campaigns, and RFM columns using the existing variables. For categorical variables marital_status and education, we regrouped them into more clear categories for further EDA and modeling.

EDA is conducted to examine the correlation between variables, variable distribution, and pairwise relationships to understand, gain insights from the provided data and prepare for clustering modeling.

Distribution Analysis

To analyze the distribution of variables, we drew histograms of the features grouped by education level (Figure 2). As is illustrated, the distributions are largely right-skewed in the purchase monetary, indicating that most of the customers purchase small amounts in the data given. The income and age have normal distribution while recency gets a relatively uniform distribution. Besides, customers with higher education levels tend to have higher monetary, frequency, tenure, and income.

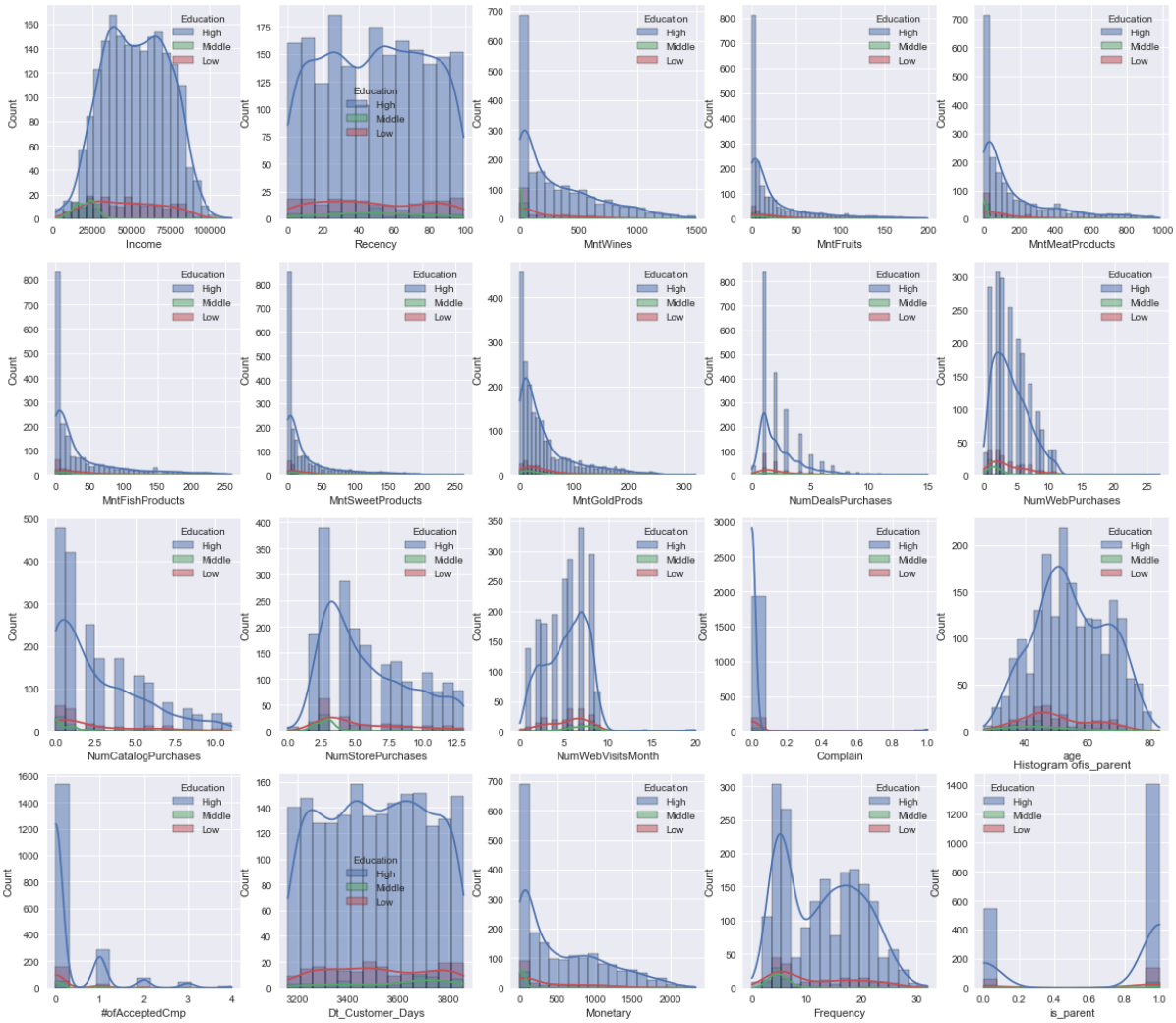


Figure 2. Histograms of the variables

Correlation Analysis

We plotted the correlation between variables, and the correlation coefficient quantifies the strength and direction of the relationship in a numerical value between -1 and 1.

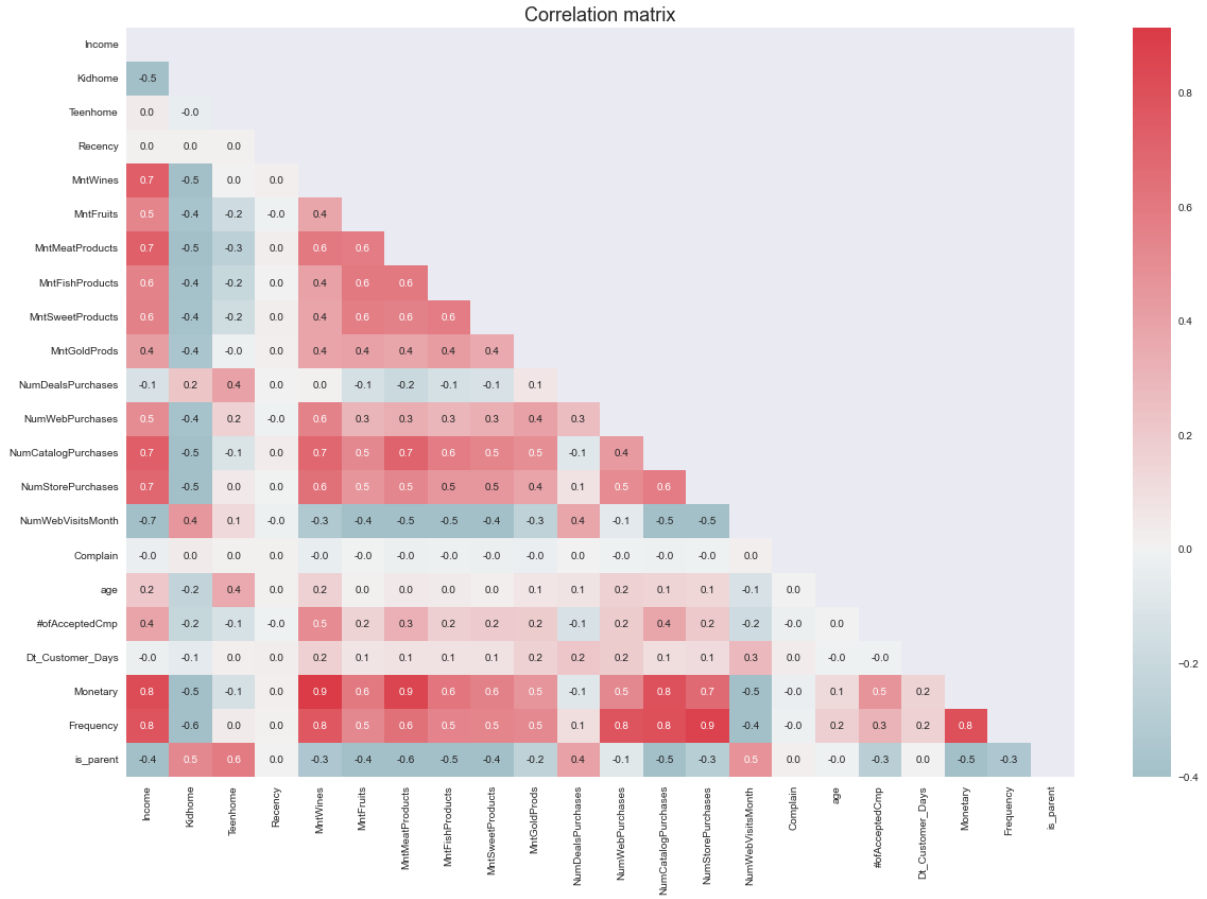


Figure 3. Correlation heatmap

Overall, most of the variables are positively related. However, there are some exceptions. The number of kids and teenagers at home, and whether the customer is a parent is negatively correlated with most of the other variables.

Pairwise Relationship Analysis

To further investigate the relationship between variables, pairs plots (also known as the scatter plot matrix) is drawn, which allows seeing both the distribution of single variables and relationships between two variables in a dataset. We plotted the pairs plot of variables, grouped by education levels (Appendix), and marital status (Appendix), respectively.

Based on the scatter plots, as income increases, the monetary and frequency increase in general, this clear positive relationship also exists between monetary and frequency. Besides, customers with higher education levels have relatively higher incomes and money spent. In terms of marital status, single customers' money spent varies largely (a larger range compared with married customers).

4. Model Development

4.1 PCA + Clustering

Our study aimed to compare the performance of two clustering algorithms, KMeans, and DBSCAN, in combination with or without four dimensionality reduction methods. The analysis was carried out on a dataset of customers that included their purchasing behaviors and demographic information. The goal was to determine the best clustering algorithm and dimensionality reduction method combination based on the histogram of each cluster regarding each variable and the silhouette score to measure the clustering quality.

The preprocessed dataset underwent dimensionality reduction using four methods: Principal Component Analysis (PCA), Kernel PCA, t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). The reduced features obtained from each method were used as input to the KMeans and DBSCAN clustering algorithms, resulting in a total of 10 algorithm combinations. The performance of each combination was evaluated based on the histogram of each cluster regarding each variable, as well as the silhouette score, which measures the clustering quality by assessing the similarity of data points within clusters and dissimilarity between clusters.

In order to select the appropriate number of components for PCA and Kernel PCA, we leveraged a scree plot to determine the number of components required to explain 95% of the variance in the data. This resulted in 17 and 18 components, respectively. As for t-SNE and UMAP, we selected 3 as the number of components based on prior experience and knowledge of the dataset.

To determine the optimal value of k in KMeans, we generated both elbow plots and silhouette plots. Elbow plots were used to identify the value of k where the decrease in the within-cluster sum of squares becomes less significant. Silhouette plots were then used to assess the quality of the resulting clusters. Additionally, 2D and 3D plots were generated to visually evaluate the performance of each clustering algorithm.

To optimize the epsilon and min sample parameters of the DBSCAN algorithm, we experimented with various values of epsilon ranging from 0.1 to 10 with a step size of 0.1, and min sample ranging from 2 to 20 with a step size of 1, using silhouette score as the metric to evaluate clustering performance. Despite achieving the highest silhouette score in combination with each dimensionality reduction method when compared to KMeans, DBSCAN was not chosen as the optimal clustering algorithm due to its tendency to produce imbalanced clusters and the excessive number of clusters generated.

Upon analyzing the clustering results through visualization of the clusters and evaluating the silhouette score (plots were shown below), it was concluded that KMeans with K=3 after PCA with 18 components was the most effective clustering algorithm due to its ability to effectively separate customers and achieve a high silhouette score.

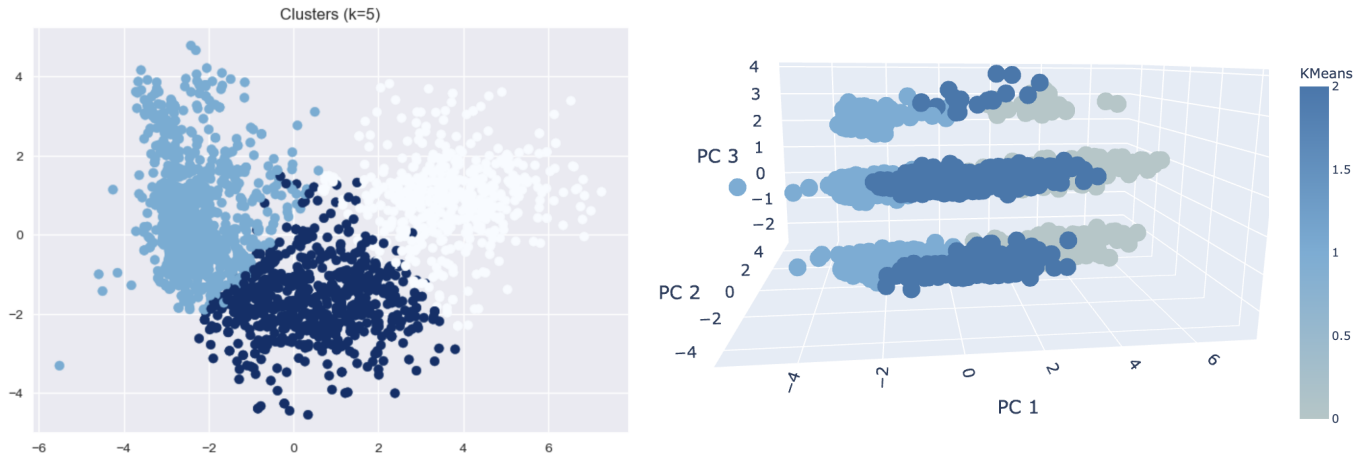


Figure 4. Clustering plots

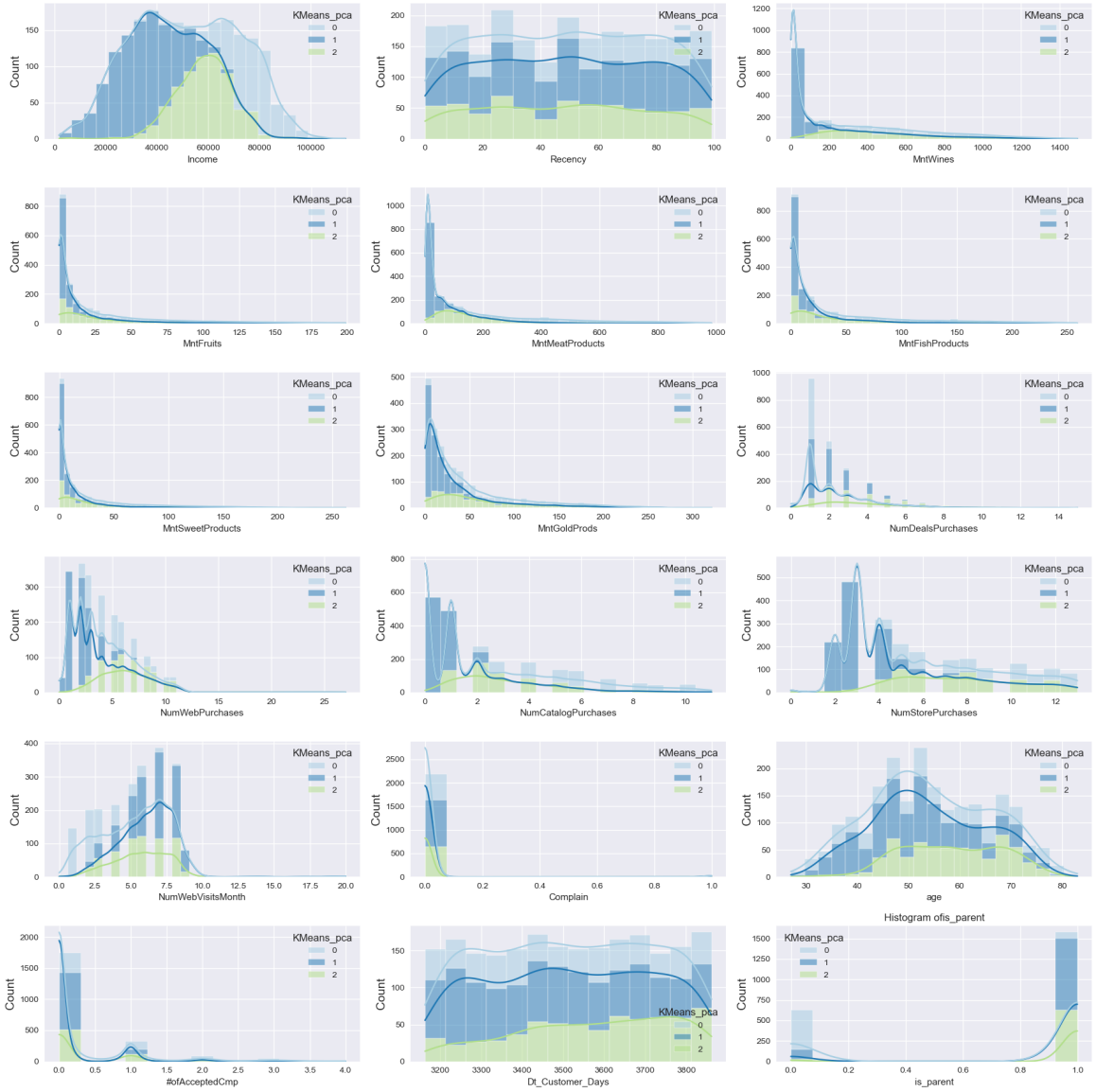


Figure 5. Histograms of variables by cluster

The clusters were defined as follows:

Cluster 0:

This cluster was composed of customers with high income, an average age of 55, and a middle education level (90% of people have high education). Most people in this cluster did not have children. Additionally, they had spent the most in each product category in the last 2 years, had the highest rate of accepting the company's campaign, and had the lowest complaint rate in the last year.

Cluster 1:

This cluster was composed of customers with low income, an average age of 50, and the lowest education level among the three clusters (83% of people have high education). Most people in this cluster had kids at home and tended to spend the least in each product category in the last 2 years. They also visited the company's website most often in the last month but had the lowest rate of accepting the company's campaign.

Cluster 2:

This cluster comprised customers with a middle income, an average age of 58, and the highest education level (96% of people have high education). Most people in this cluster had teenagers at home and tended to spend in the middle in each product category in the last 2 years. They also had the highest single rate among the three clusters.

4.2 RFM Analysis

4.2.1 Model

In addition to PCA+clustering, our study has also featured RFM analysis which is a popular customer segmentation technique that aims to identify and target specific groups of customers based on their Recency, Frequency, and Monetary Value.

For feature engineering, Recency was provided in the initial dataset from Kaggle. We generated Monetary by taking the sum of MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProds. Frequency was generated by adding NumWebPurchases, NumCatalogPurchases, and NumStorePurchases.

As we chose the KMeans algorithm for clustering based on RFM value, standardization of features was required because the algorithm is sensitive to the scale of input data. Our study rescaled RFM value in the range of 1-5. Based on the Elbow method and Silhouette plot, K=4 was chosen as the optimal.

Based on the result of KMeans clustering, we assigned four names for the new clusters based on the cluster average of RFM. New cluster names were Diamond, Gold, Silver, and Bronze in the descending rank of customer's overall value to the business.

4.2.2 Findings

Features of four RFM groups of customers were:

- Diamond: High Monetary, High Frequency, Low Recency
- Gold: High Monetary, High Frequency, High Recency
- Silver: Low Monetary, Low Frequency, Low Recency
- Bronze: Low Monetary, Low Frequency, High Recency

To evaluate the quality of RFM clusters, our study created a 3D plot of RFM value colored by the RFM cluster as shown in Figure 6. Customers were well segmented into four groups without a clear overlap of points in the graph.

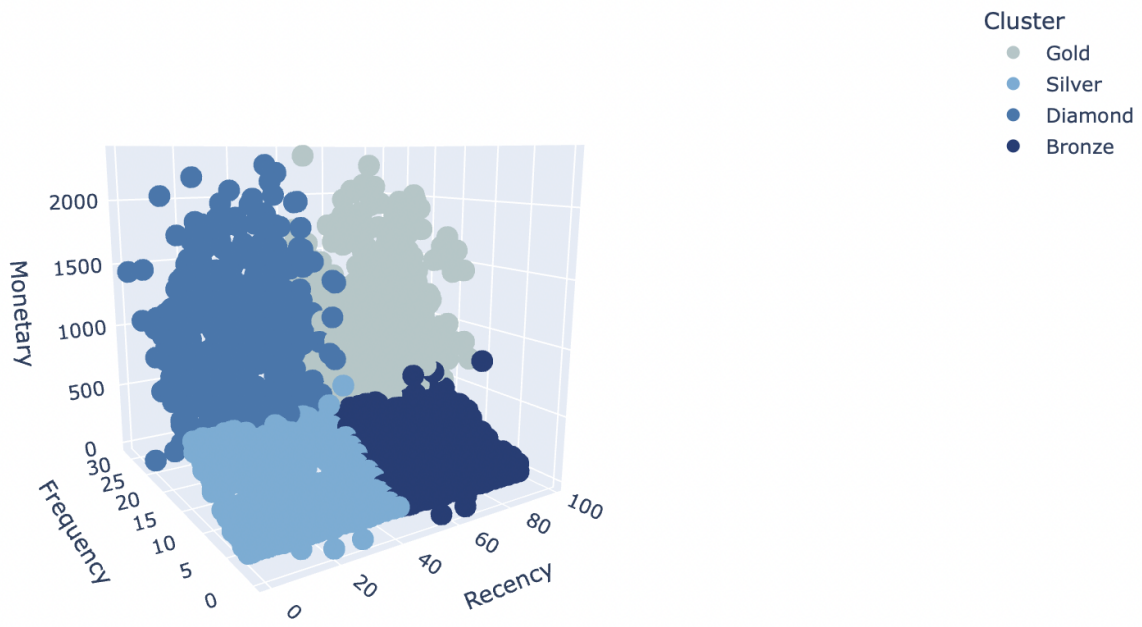


Figure 6. 3D Plot of RFM value colored by Cluster

In terms of other numerical variables in our dataset, their distributions in the RFM clusters were found to be different. Diamond and Gold customers had a higher median income of ~$70k and a higher median age of ~56, while Silver and Bronze customers had a lower median income of ~$38k and a lower median age of ~52.

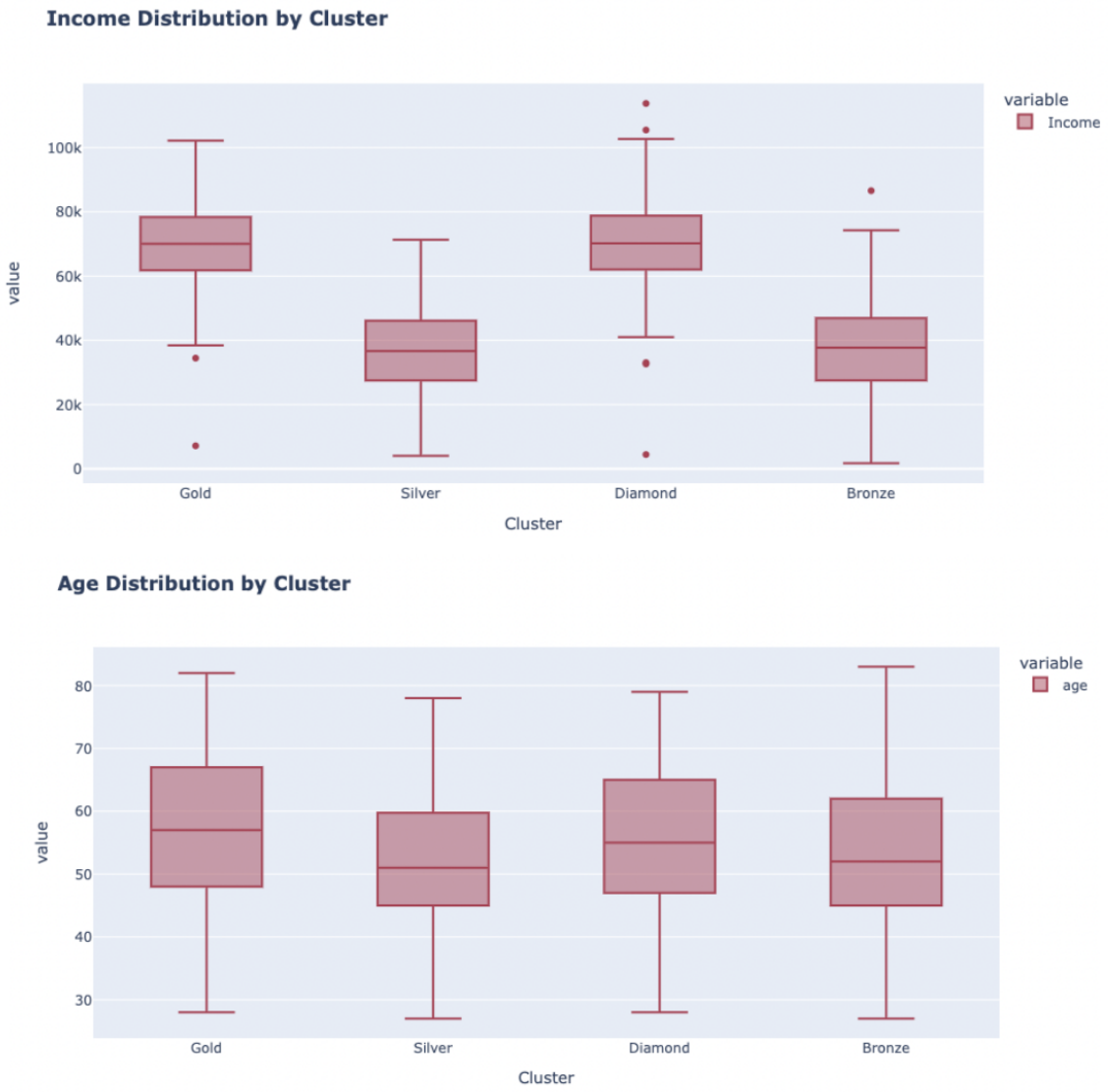


Figure 7. Numerical Variable Distributions by Cluster

In terms of other categorical variables in our dataset, findings for customers in different RFM groups were explained below:

- Silver and Bronze customers had much higher parent rates than Diamond and Gold customers.
- Diamond customers complained the least in the last 2 years.
- Middle-level education customers only appeared in the Silver and Bronze groups.
- Gold customers had the highest alone marital status rate among all groups.

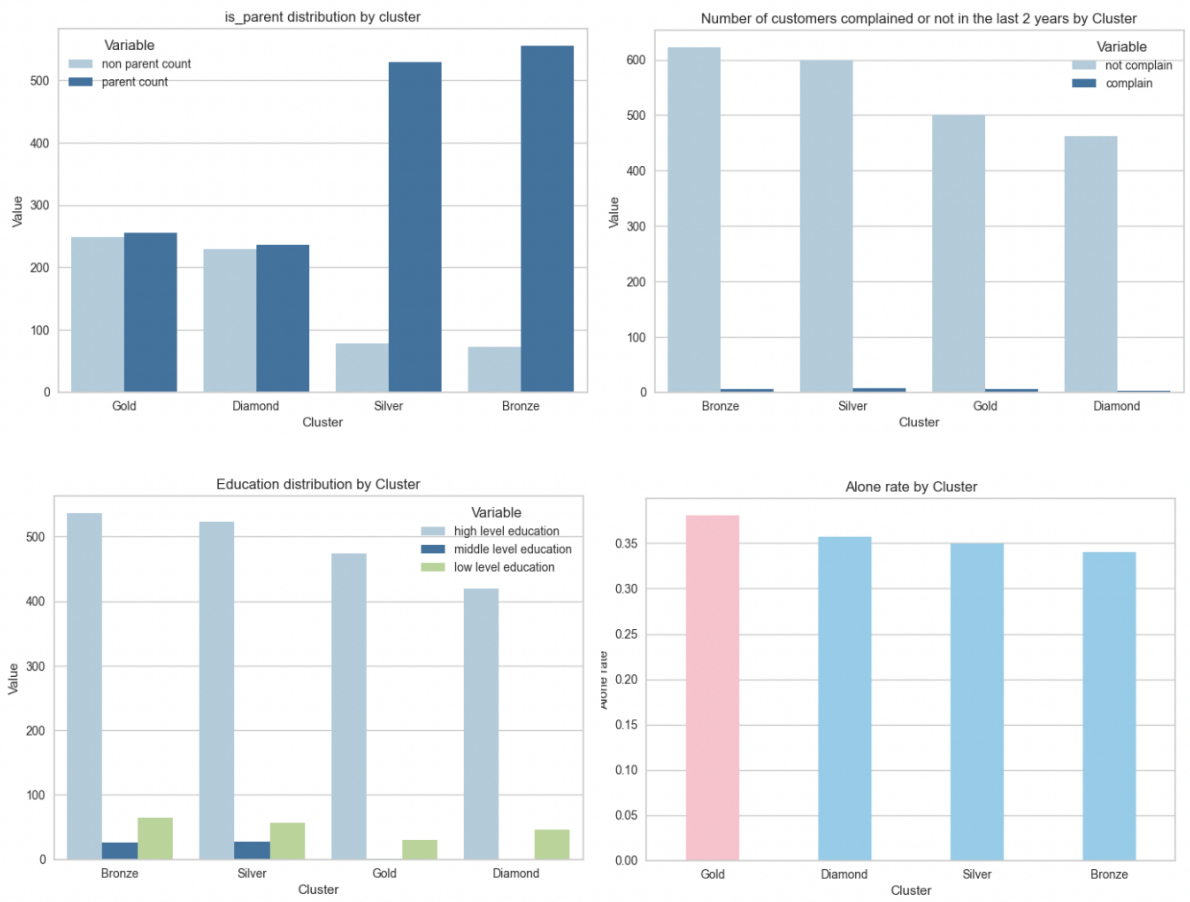


Figure 8. Categorical Variable Distributions by Cluster

Based on the analysis, the following is the suggested marketing strategy for each group:

1) Diamond customers are high-value, loyal customers who are unlikely to complain. To retain them, consider offering personalized loyalty rewards or exclusive perks such as early access to new products or services. The company could also send personalized emails or messages to show appreciation for their loyalty and make them feel valued.
2) Gold customers are high-value customers who may need a little extra attention to keep them engaged with the brand. To keep them coming back, consider creating targeted marketing campaigns that highlight new products or services that align with their interests. You could also offer incentives for repeat purchases, such as a discount or free shipping. Gold customers also had the highest alone marital status rate among all groups, which suggests that this segment may have unique preferences or behaviors that could be leveraged in marketing efforts. Consider tailoring marketing messages to highlight products or services that appeal to this demographic, such as solo travel packages or high-end gadgets that cater to their interests.
3) Silver and Bronze customers are at risk of churning, so it's important to target them with retention strategies. Consider offering them personalized discounts or promotions to encourage them to make repeat purchases. You could also create

targeted email or social media campaigns that highlight the benefits of staying loyal to your brand.

4) The fact that middle-level education customers only appeared in the Silver and Bronze groups suggests that there may be an opportunity to target this segment with relevant marketing messages or promotions. Consider creating content or campaigns that appeal to this demographic, such as educational resources or promotions that align with their interests or needs.

5. Conclusions & Implications

The KMeans clustering algorithm with K=3 after PCA with 18 components produced the best model for this analysis. The resulting clusters differ in various demographic and behavioral factors, such as income, age, education level, parental status, purchasing habits, and engagement with the company's campaign and website. Based on these findings, customized marketing strategies could be implemented for each cluster to improve customer satisfaction and maximize profits.

Cluster 0, representing the high-income group with high education and no children, who spent the most in each product category and had the lowest complaint rate, could be targeted with higher-end products and personalized marketing campaigns that cater to their spending habits. For Cluster 1, which represents the low-income group with kids at home and frequent website visits but the least spending, marketing strategies could focus on affordability and family-friendly products, along with incentives to encourage campaign acceptance. Cluster 2, representing the middle-income group with the, who spent in the middle range, had teenagers at home, and had the highest single rate among the three clusters, could be targeted with tech-savvy or cultural products that could increase their spending. By implementing these customized marketing strategies, the company can better meet the needs and preferences of each customer cluster, ultimately leading to increased customer satisfaction and revenue.

In addition to KMeans clustering based on PCA, this study also constructed RFM clusters using KMeans. Customers were segmented into Diamond, Gold, Silver, and Bronze with Diamond customers representing low recency, high frequency, and high Monetary. High-value groups including Diamond and Gold also had higher median age and income, fewer complaints in the past 2 years, higher alone rate, less likely to be a parent, and excluded people with a middle level of education. Related marketing strategies are proposed for each group of customers.

Our study would be significant for companies to leverage their abundant data regarding customer behavior and the strategies proposed might be generalized in more business settings. A further improvement would be combining multiple segmentation techniques as K-means clustering and RFM analysis both have their limitations. By combining multiple segmentation techniques, such as demographic segmentation or behavioral segmentation, it is likely to gain a more complete picture of customers' preferences, needs, and behaviors. In addition, although this study has identified customer segments, it's important to validate the

results to ensure that they are meaningful and actionable. Further steps might include testing the effectiveness of targeted marketing campaigns or conducting customer surveys to get feedback on the segmentation approach.

References

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. Journal of City and Development, 3(1), 12–30. https://doi.org/10.12691/jcd-3-1-3 .

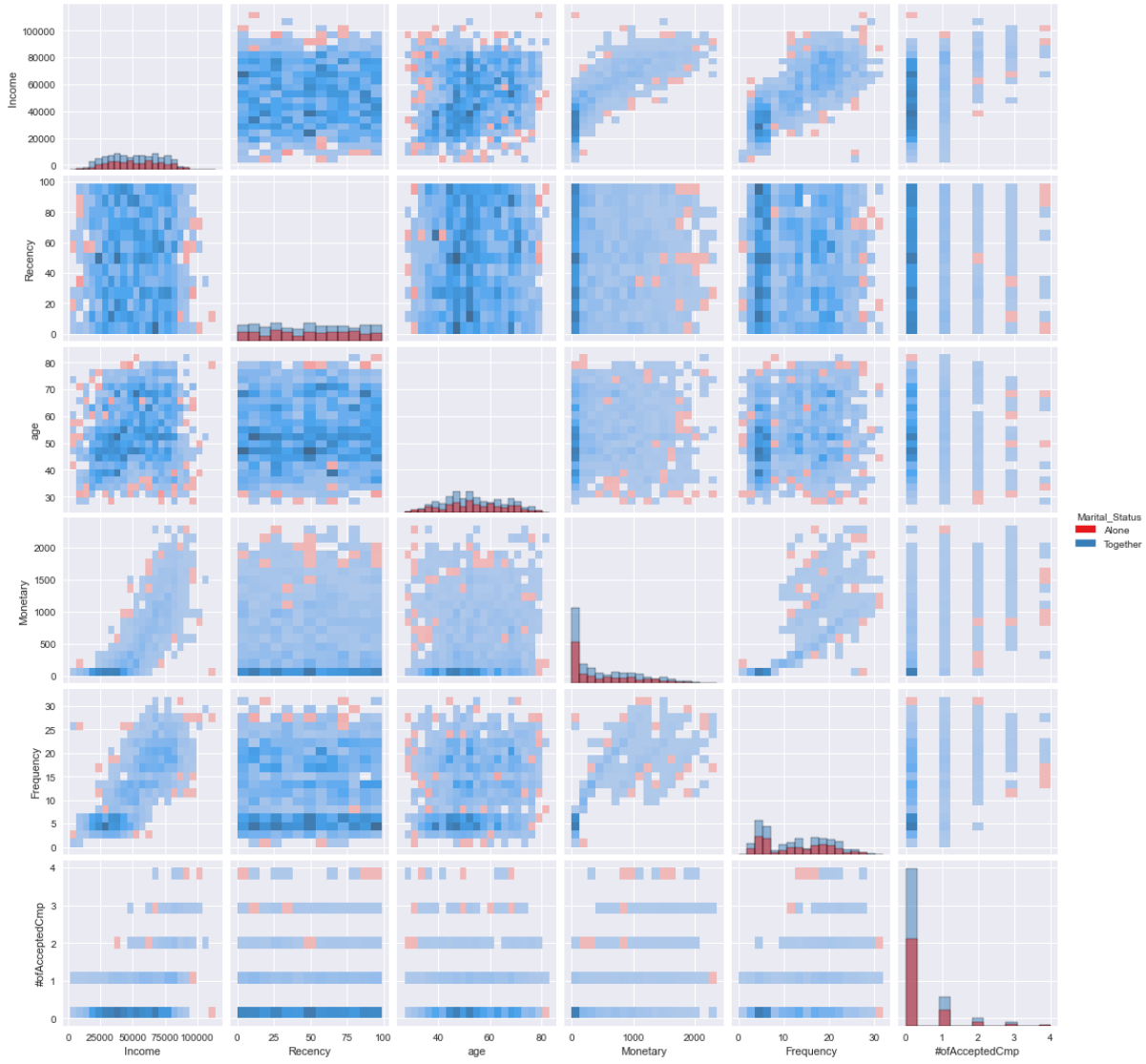
Chen, Daqing, et al. "Data Mining for the Online Retail Industry: A Case Study of RFM Model-Based Customer Segmentation Using Data Mining." *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 3, 27 Aug. 2012, pp. 197–208, https://doi.org/10.1057/dbm.2012.17.

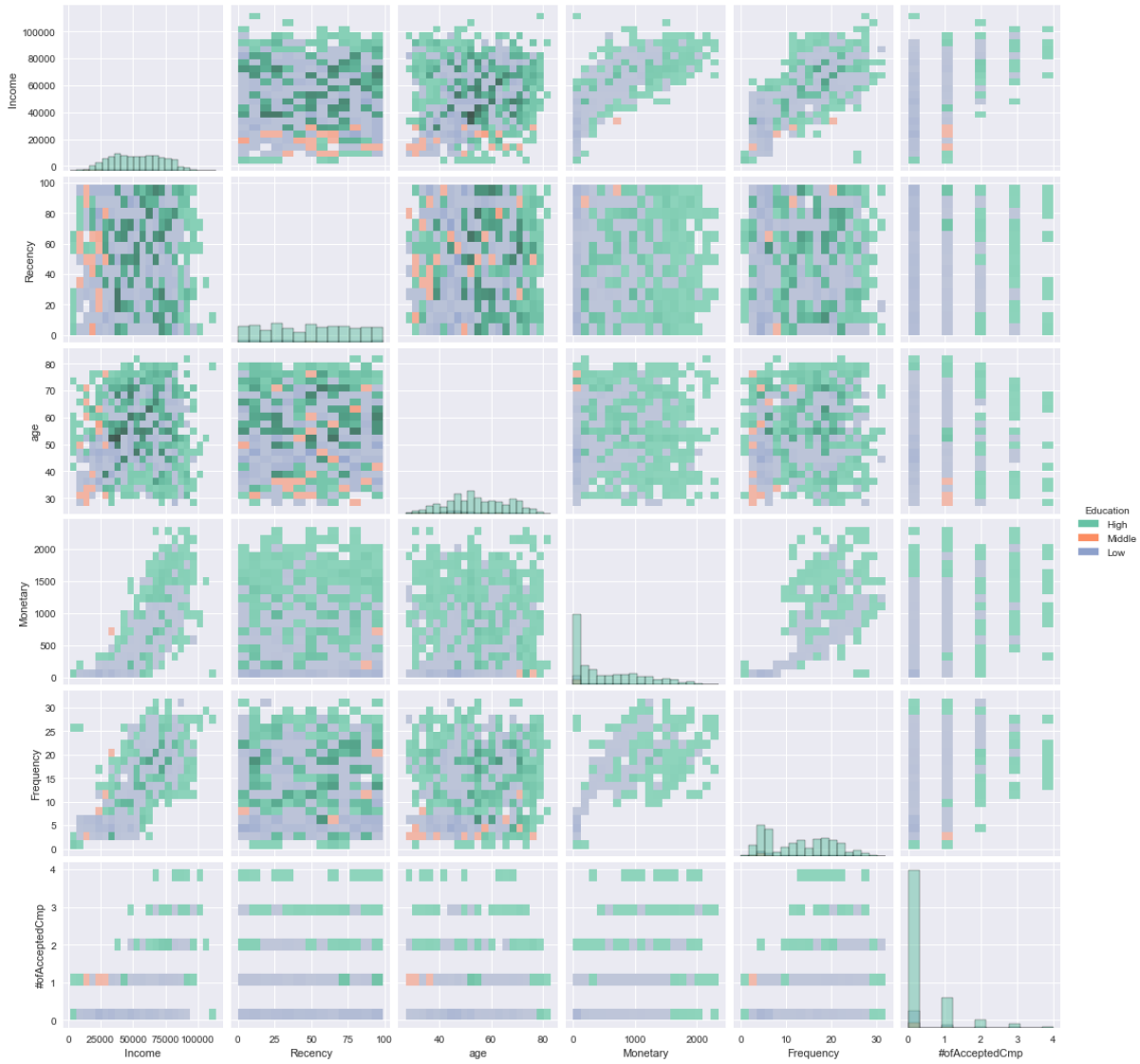
Dogan, Onur, et al. "CUSTOMER SEGMENTATION by USING RFM MODEL and CLUSTERING METHODS: A CASE STUDY in RETAIL INDUSTRY." *INTERNATIONAL JOURNAL of CONTEMPORARY ECONOMICS and ADMINISTRATIVE SCIENCES*, vol. 8, 2018, avesis.kocaeli.edu.tr/yayin/894de1af-d068-4e33-ad18-d0d727c24fbe/customer-segmentation-by-using-rfm-model-and-clustering-methods-a-case-study-in-retail-industry. Accessed 6 Mar. 2023.

Kim, Su-Yeon, et al. "Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A Case Study." *Expert Systems with Applications*, vol. 31, no. 1, July 2006, pp. 101–107, https://doi.org/10.1016/j.eswa.2005.09.004.

Zakrzewska, D., and J. Murlewski. "Clustering Algorithms for Bank Customer Segmentation." *IEEE Xplore*, 1 Sept. 2005, ieeexplore.ieee.org/abstract/document/1578784.

Appendix


Income
Recency
age
Monetary
Frequency
#ofAcceptedCmp
Marital_Status
Alone
Together


Income
Recency
age
Monetary
Frequency
#ofAcceptedCmp
Education
High
Middle
Low