CUSTOMER SEGMENTATION WITH CLUSTERING AND RFMT MODEL

Hatice K. Erdogan
American University, Department of Math & Statistics

ABSTRACT

In today's competitive market, understanding customer data in depth is crucial for fostering customer-company engagement. Identifying similarities and differences among customers, predicting their behaviors, and providing personalized offers have become critical for enhancing customer satisfaction and loyalty. Customer segmentation plays a pivotal role in this process. The RFMT (Recency, Frequency, Monetary, Time) model has long been used to identify high-value customers; target disengaged segments and guide promotional efforts. This paper explores the application of machine learning techniques with a focus on clustering to segment customers effectively. Clustering helps organizations group customers and other entities based on shared attributes, enabling data-driven decision-making and more targeted marketing strategies.

***Keywords*—** clustering, customer segmentation, RFMT model, K-means, Non-negative Factorization, Gaussian Mixture models

1. INTRODUCTION

Decision-makers rely on various factors to segment customers effectively. Demographic variables—such as age, gender, family structure, education level, and income—are among the most commonly used due to their simplicity and accessibility. In addition to demographics, other key segmentation criteria include socio-cultural, geographic, psychographic, and behavioral variables, providing deeper insights into customer preferences and behaviors [1].

2. RELATED WORK

Customer segmentation is the process of dividing customers into distinct groups based on shared characteristics, such as demographics, interests, behaviors, or location. This approach enables businesses to focus their marketing efforts and allocate resources more effectively, targeting high-value and loyal customers to drive better outcomes and achieve strategic goals.

2.1. Recency, Frequency and Monetary Model for Customer Segmentation

E-commerce provides businesses with powerful tools to track and analyze customer behavior through internal analytics or external web crawling software. The availability of data on browsing habits and purchasing patterns has made behavioral segmentation one of the most popular approaches for identifying customer groups (Massimino, 2016; Velotio Technologies, 2019, as cited in [2]). A widely adopted framework for behavioral segmentation is the Recency, Frequency, and Monetary (RFM) model, which has been extensively used to assess customer value and behavior over time.

The RFM model evaluates customer behavior along three key dimensions:

Recency: How recently a customer made a purchase.
Frequency: How often a customer makes purchases.
Monetary: The total amount a customer has spent during a specified period.

In practice, businesses generate RFM data from historical purchase records and categorize customers into distinct groups through RFM scoring. Two common methods for scoring are customer quintile scoring and behavior quintile scoring. Customer quintile scoring sorts customers by Frequency and Monetary values in descending order and Recency in ascending order, dividing them into five equal quintiles. In contrast, behavior quintile scoring assigns customers to quintiles based on behavioral patterns, with the highest scores given to those with the most recent, frequent, and high-value purchases.

Clustering techniques are often applied to RFM data to further refine customer segmentation. K-means clustering, Expectation-Maximization (EM), and hierarchical clustering are commonly used methods for this purpose. These algorithms help group customers with similar RFM scores into meaningful clusters, which can then be labeled to create customer profiles (Anitha & Patil, 2029; Yoseph & Heikkila, 2018, as cited in [2]). For example, businesses may label clusters as "platinum," "gold," and "iron" to reflect varying levels of customer value.

While the RFM model offers valuable insights, it has limitations. Specifically, it does not capture customer loyalty or long-term behavioral changes. As a result, several extensions have been proposed. For example, researchers have added dimensions such as "time since first purchase" and "churn probability" to address dynamic changes in behavior. Other variations, such as the LRFM model,

incorporate the time span between the first and last purchases to account for customer lifetime patterns (Yeh et al., 2008; Alvandi et al., 2012, as cited in [2]). Despite these enhancements, the RFM model remains a fundamental tool for segmenting customers based on purchasing behavior.
The model's ability to identify and profile key customer groups makes it a valuable asset for businesses seeking to enhance customer engagement and drive growth.

3. DATASET

The data I'll be using for the project is the customer segmentation – online retail dataset, provided by Kaggle. It contains all the transactions recorded between December 1, 2010, and December 9, 2011, for a UK-based, non-store online retail company specializing in unique all-occasion gifts. Many customers of the company are wholesalers, contributing to the dynamics of the dataset. The dataset consists of 541,909 rows and 8 features: Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID and Country. There are 4372 distinct customer IDs in the dataset. For each customer, I will create the three key variables: Recency, Frequency, and Monetary (RFM). These variables will serve as the foundation for customer segmentation. Additionally, the Interpurchase Time (T) will be incorporated to extend the analysis.

4. METHODOLOGY

4.1. Interpurchase Time (T) Calculation

The Interpurchase Time (T) measures the average time gap between consecutive shopping sprees. If the dates of the first purchase (T_1) and last purchase (Tn) for a customer are known, the holistic shopping cycle (L) can be approximated as the months between T_1 and Tn. F accounts for the count of unique invoices per customer. The T value can then be calculated as:

$$T = \frac{L}{F - 1} = \frac{T_n - T_1}{F - 1}$$

This formula ensures that only customers with at least two purchases ($F \geq 2$) are included in the calculation. This step ensures that meaningful insights about shopping frequency are captured [2].

4.2. RFMT Scoring

To align the variables for clustering analysis, I will convert the RFMT values into a five-quintile scale. As each variable—Recency, Frequency, Monetary, and Interpurchase Time—has different units and ranges, normalization is performed using Min Max Scaler by scikit-learn. Following Miglautsch's (2000) scoring rules (as cited in [2]) the quintile ranges for Recency, Frequency, and Monetary values will be:

Fig. 1. Customers segmented using RFMT scores

Recency Quintiles:
1 = 9+ months
2 = 5–9 months
3 = 2–5 months
4 = 1–2 months
5 = Within the last month

Frequency Quintiles: Higher frequency scores correspond to more frequent purchases. Customers with many transactions receive higher scores (e.g., 100+ purchases).

Monetary Quintiles: Higher spending results in a higher score.

Time (T) Quintiles:
1 = 7+ months
2 = 4–7months
3 = 2–4 months
4 = 1–2 months
5 = Within a month

These quintiles allow for a consistent scale across all dimensions. Each customer is assigned a score between 1 and 5 in each category, with higher scores reflecting more favorable behavior (e.g., recent, frequent, or high-value purchases). Then all the scores are aggregated for each customer to be assigned for one of the 6 classes as illustrated in Figure 1. "Champions" are the customers with the highest engagement across all metrics. "Loyal customers" are regular buyers with a high Frequency score ($F \geq 4$), "big spenders" are customers who spend significantly, indicated by a high Monetary score ($M \geq 4$) regardless of recency or spending. "Recent customers" are the ones who made purchases recently ($R \geq 4$). "At-risk" customers are with low recency scores ($R \leq 2$), signaling a decline in engagement. Lastly, "lost customers" do not fall into the above categories, indicating inactivity or disengagement. Figure 1 illustrates the distribution of the customers based on this scoring logic. "Champions", "at-risk", "recent" and "loyal" customers all account for representing more than 800 customers each.

4.3. Clustering Analysis

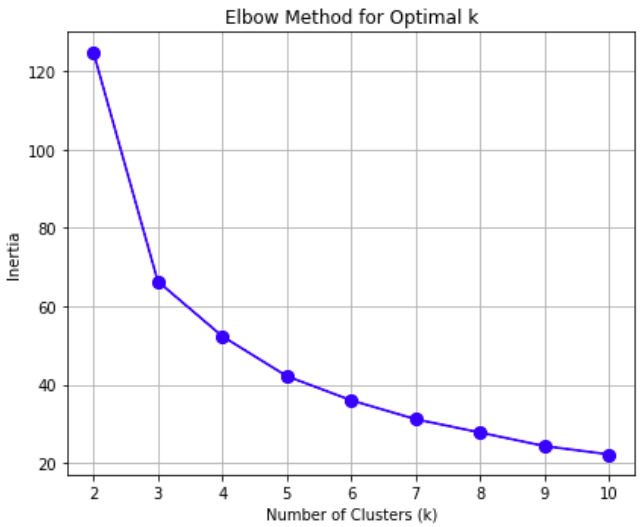

Fig. 2. Elbow plot for the optimal k

The RFM and RFMT scores will be used to cluster customers using algorithms such as K-means, agglomerative clustering, Non-negative Factorization and Gaussian Mixture Models. This segmentation identifies groups like high-value customers (recent, frequent, and high-spending buyers) to support targeted strategies for customer retention and engagement [2] and potentially bring back lost or at-risk customers.

K -Means clustering is a widely used method for identifying distinct customer segments based on purchasing behaviors defined by the RFMT variables. By partitioning customers into K clusters, K-Means enables the identification of groups with similar purchasing patterns [4]. Figure 2 illustrates the implementation of the elbow method, using inertia as a reference metric. The "elbow" in the graph indicates that the optimal number of clusters is either 4 or 5. I decided to try out both 4 and 5 clusters for each algorithm and decide which one performed better in terms of silhouette score and visual analysis. Therefore, while most of the algorithms performed best with 5 clusters, some performed better with 4 clusters.

Figures 3 and 4 present the implementation of the K-Means algorithm with 5 clusters using both the RFM and RFMT models. Principal Component Analysis (PCA) was applied to reduce dimensionality and visualize the clustering results in a 2-dimensional feature space. The clusters are well separated, and the results reveal notable differences between the RFM and RFMT models, particularly in the distribution of data points across clusters.

Non-Negative Matrix Factorization (NMF) is a valuable technique for reducing the dimensionality of data while maintaining non-negativity. Unlike PCA and vector quantization, NMF stands out because it is part-based and applies non-negativity constraints, allowing only additive combinations rather than subtractive ones [5]. It is

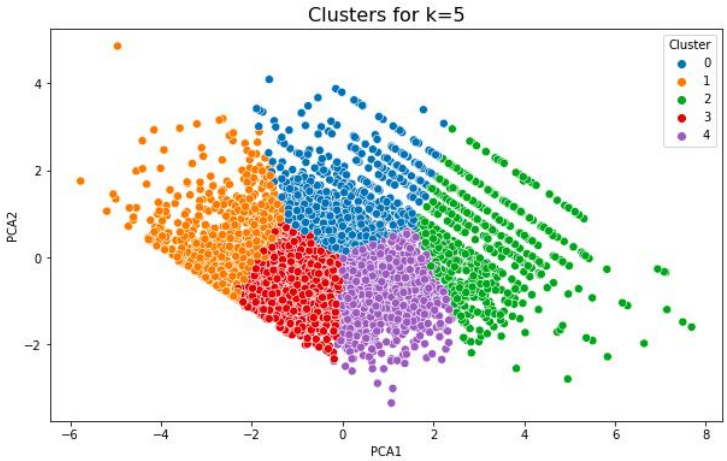

Fig. 3. K-Means RFM Model (5 clusters)

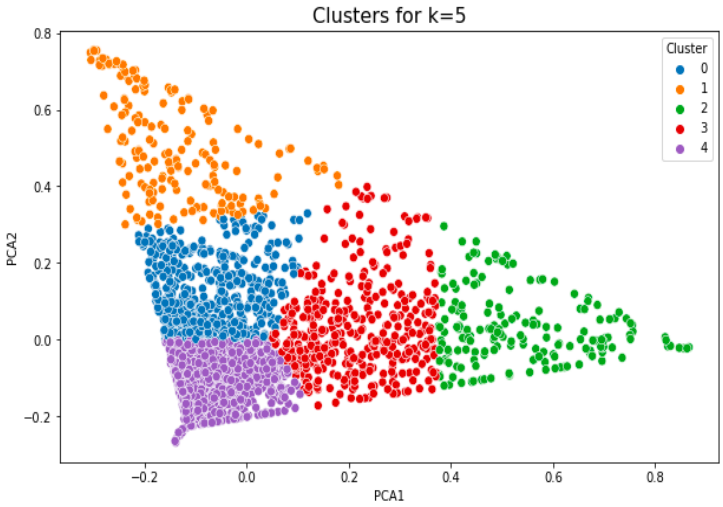

Fig. 4. K-Means RFMT Model (5 clusters)

particularly suitable for count data such as purchase frequency and monetary value which are two of the key features of interest in this study. By decomposing the data into non-negative components, NMF uncovers latent features in customer behavior that are not directly observable. This provides a more interpretable representation of customer segments.

To enhance the clustering process, the NMF-generated components were incorporated as features in the K-Means algorithm.

Figures 5 and 6 illustrate the implementation of K-Means with 4 clusters using the RFM and RFMT models, plotted in a 2-dimensional space with NMF Component 1 on the x-axis and NMF Component 3 on the y-axis. In the RFM model, the clusters are arranged vertically, with significant overlap observed between Cluster 1 and Cluster 3, indicating less distinct segmentation, while nice distinction between Cluster 0 and 2. In contrast, the RFMT model, which incorporates the "Time" variable, displays clusters that are better separated

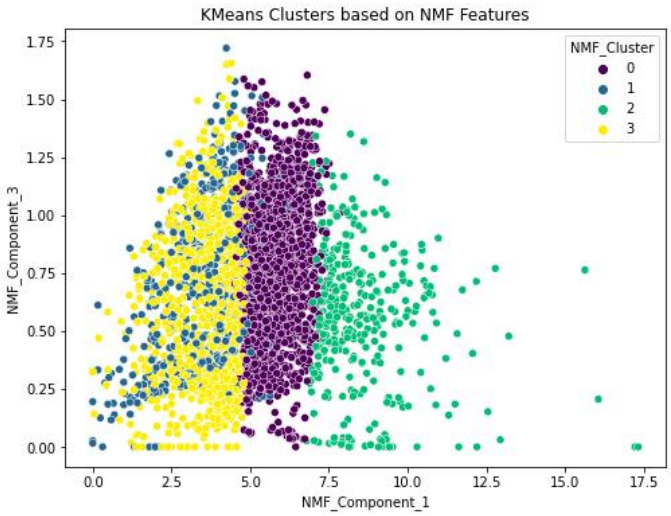

Fig. 5. K-Means with NMF features on RFM Model (4 clusters)

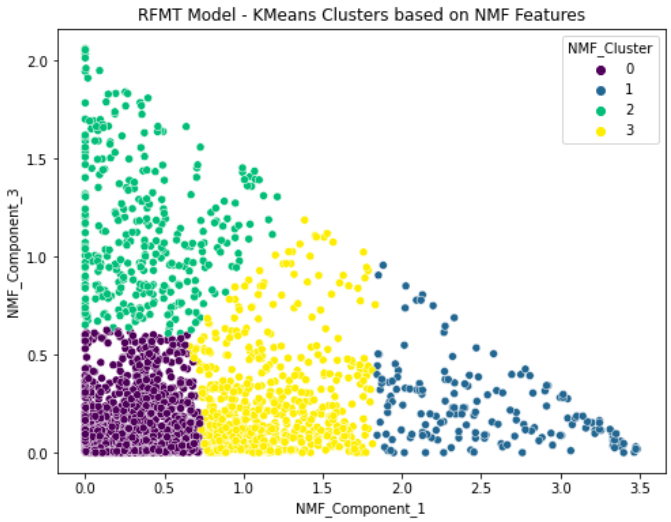

Fig. 6. K-Means with NMF features on RFMT Model (4 clusters)

and more regionally defined, highlighting the added value of including the "Time" variable in the analysis.

Hierarchical clustering decomposes data into a tree-like structure based on group similarities, using either a bottom-up (agglomerative) or top-down (divisive) approach.

Agglomerative clustering, in particular, starts with each data point as its own cluster and iteratively merges the most similar clusters, creating a hierarchy of nested groups. The algorithm captures non-linear relationships in the data and reveals valuable insights into the structure and connections among customer groups [2].

Figures 7 and 8 illustrate agglomerative clustering applied to the RFM and RFMT models, each with 5 clusters. Compared to k-means, the resulting clusters exhibit a wider spread, with

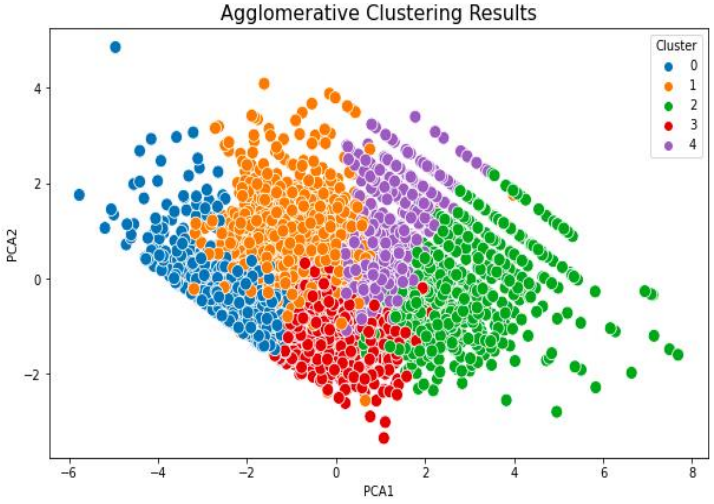

Fig. 7. Agglomerative Clustering RFM Model (5 clusters)

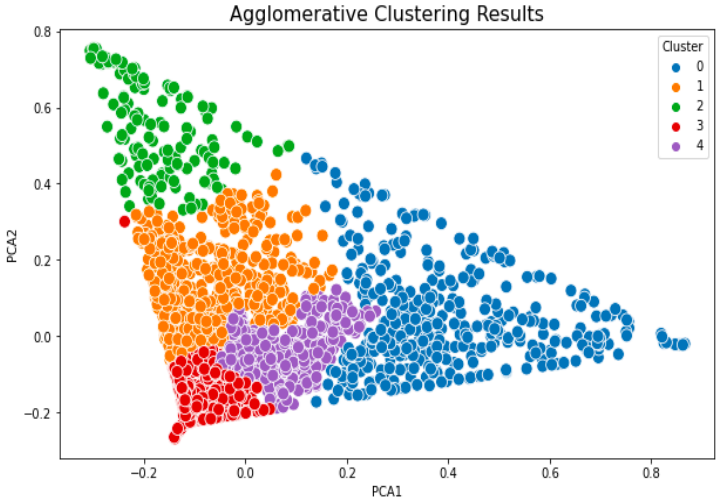

Fig. 8. Agglomerative Clustering RFMT Model **(**5 clusters)

the RFM model showing more out-of-cluster points assigned to different clusters, especially through the middle of the graph. The RFMT model demonstrates better-defined clusters which stresses the importance of incorporating the "time" variable for improved segmentation.

Gaussian Mixture Models (GMM) extend traditional clustering methods by modeling data as a mixture of multiple gaussian distributions which allows for clusters with varying shapes and overlapping characteristics. This flexibility is particularly useful for customer segmentation, where purchasing behaviors often overlap between groups. GMM provides a probabilistic framework, enabling the estimation of cluster membership probabilities, which can capture uncertainty and refine segmentation insights [6].

Figure 9 showcases the GMM applied to the RFM model with 4 clusters. The clusters are distinctly separated with the

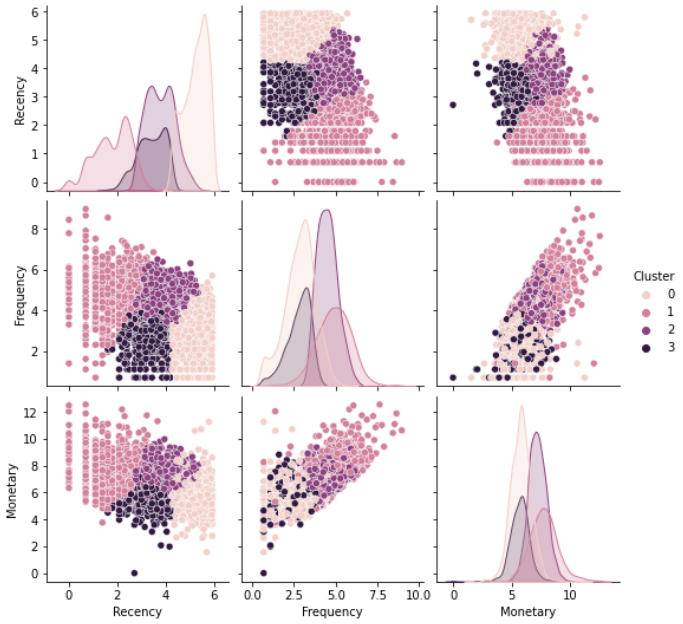

Fig. 9. GMM RFM Model (4 clusters)

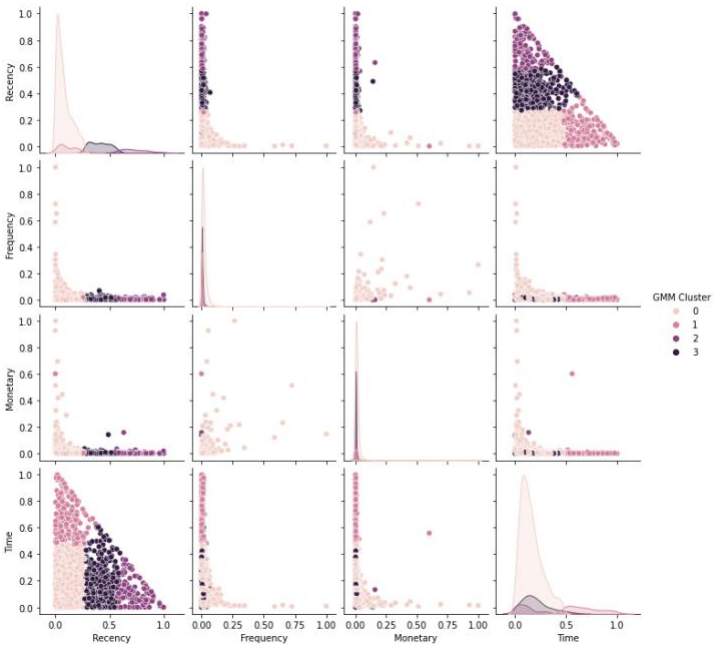

Fig. 10. GMM RFMT Model (4 clusters)

exception of the frequency-monetary plot, where overlap is more pronounced. When incorporating the "time" variable into the RFMT model on GMM, the clustering results improve significantly. As illustrated in Figure 10, the clusters become more distinct and better distributed across the feature space especially for the recency-time plot, while monetary-frequency plot show poor results. Table 1 summarizes the performance of various clustering algorithms applied to the customer segmentation dataset, using both RFM and RFMT models. Silhouette Score is used to compare the models and evaluate the quality of the clusters. Bayesian Information Criterion (BIC) is added as an additional metric to evaluate the GMMs performance.

Among the clustering algorithms, Gaussian Mixture Models (GMM) with the RFMT model achieved the highest Silhouette Score of 0.50 and the lowest BIC value of -28,374.66, indicating well-defined and distinct clusters with strong model performance. The incorporation of the "Time" variable in the RFMT model appears to enhance cluster separation and provide a more accurate representation of customer behaviors. In contrast, GMM with the RFM model resulted in a lower Silhouette Score of 0.33 and a higher BIC value of -15,137.75, demonstrating the added value of temporal dimensions in improving clustering outcomes.

The K-Means algorithm also demonstrated notable results. Using RFMT variables, K-Means achieved a Silhouette Score of 0.46, outperforming its performance with the RFM model (Silhouette Score: 0.33). Notably, when Non-Negative Matrix Factorization (NMF) was applied as a dimensionality reduction technique, the RFMT model with K-Means produced a higher Silhouette Score of 0.48, further highlighting the utility of incorporating time and latent features in customer segmentation.

Agglomerative clustering, while offering a hierarchical approach, performed less effectively compared to K-Means and GMM. The RFMT model with Agglomerative clustering produced a Silhouette Score of 0.36, while the RFM model achieved a lower score of 0.21, suggesting this method may not capture customer behavior relationships as well as GMM or K-Means.

Overall, the results indicate that the inclusion of the "Time" variable in the RFMT model consistently enhances cluster quality across all algorithms. Additionally, the combination of dimensionality reduction techniques, such as NMF, with clustering algorithms like K-Means, provides further improvement in segmentation accuracy. Gaussian Mixture Models with the RFMT model stand out as the most effective approach for identifying well-separated and meaningful customer clusters. It is worthy to note that this dataset contained customer data over a one-year period, additional data across further timeframe might improve model performance.

Clustering Algorithm	Number of Clusters	Silhouette Score	BIC
K-Means RFM w/out PCA	5	0.27	
K-Means RFMT w/out PCA	5	0.44	
K-Means RFM	5	0.33	
K-Means RFMT	5	0.46	
Agglomerative RFM	5	0.21	
Agglomerative RFMT	5	0.36	
K-Means w/ NMF RFM	4	0.30	
K-Means w/ NMF RFMT	4	**0.48**	
GMM RFM	4	0.33	-15137.75
GMM RFMT	5	**0.50**	-28374.66

Table 1. Clustering algorithm results

5. REFERENCES

[1] Doğan, Onur, Ejder Ayçin, and Zeki Bulut. "Customer segmentation by using RFM model and clustering methods: a case study in retail industry." *International Journal of Contemporary Economics and Administrative Sciences* 8 (2018).

[2] Zhou, Jinfeng, Jinliang Wei, and Bugao Xu. "Customer segmentation by web content mining." *Journal of Retailing and Consumer Services* 61 (2021): 102588.

[3] Tavakoli, Mohammadreza, et al. "Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study." *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. IEEE, 2018

[4] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171

[5] Jiajia, Wang, et al. "Clustering product features of online reviews based on nonnegative matrix tri-factorizations." *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2016.

[6] Oloo, John Mark. "Examining Gaussian Mixture Models using clustering algorithms."