Customer Segmentation through Clustering

Abed Al Kader Al Bakkour
School of Informatics
University of Skövde
Skövde, Sweden
a22abeal@student.his.se

Muhammad Usman
School of Informatics
University of Skövde
Skövde, Sweden
a22muhus@student.his.se

***Abstract*—Customer segmentation is a popular task in data analysis, it enables organizations to target particular customer groups with customized marketing efforts. The Online Retail II data set from the ML Repository, which contains transaction data of an online retailer from 2010 to 2011, was utilized in this study to apply several clustering approaches. K-means, Hierarchical clustering and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) were the techniques that were utilized. So, to find the most valuable customers, we used RFM analysis (Recency, Frequency, Monetary) in addition to clustering. The RFM analysis allowed us to identify the customers with the highest potential value, which can be targeted with personalized marketing strategies. Moreover, we utilized the Silhouette analysis, Davies-Bouldin and Calinski-Harabasz index to evaluate the effectiveness of the various clustering technique.**

Keywords—Customer Segmentation, Unsupervised machine learning algorithms, K-mean clustering, Hierarchical algorithm, DBSCAN algorithm, RFM model, Silhouette analysis, Davies-Bouldin index, Calinski-Harabasz index.

I. INTRODUCTION

The process of splitting a customer base into groups of people who are similar in particular aspects significant to marketing, such as age, gender, interests, and expenditure patterns, is known as customer segmentation. Due to being better able to target their marketing efforts and resources, businesses are able to boost both client retention and profitability. Customer segmentation can be especially helpful in the context of online shopping for locating and targeting the most valued customers as well as for locating chances for increasing sales and cross-selling [1].

RFM (Recency, Frequency, Monetary) analysis is a commonly used method for identifying the most valuable customers in a business. It involves ranking customers based on their purchase recency, frequency, and monetary value, with the assumption that the more recent, frequent, and high value the purchases, the more likely the customer is to make future purchases [4]. RFM analysis can be combined with customer segmentation to target marketing efforts more effectively and to optimize the allocation of resources.

The improvement and comprehension of consumer communications as well as methods that businesses can improve their goods and services are greatly aided by segmentation. The significance of customer segmentation can be attributed to the fact that it aids in understanding various customer trends, such as which products are most in demand. This aids in preserving relationships with customers and supports the maintenance of both supply and demand for the services that are most in demand by customers [5]. This aids in recognizing consumer defection, identifying the clients most inclined to do so, and identifying other market trends and plans. The most important points are those that are anticipated to address consumer relations regarding the company's operations and problem-solving. In order to create client categories with similar values, wants, expectations, and needs, standards are set.

A. Research questions/hypothesis

- **Data quality and availability:** One potential problem is that the customer data may be incomplete, incorrect, or outdated, which can affect the accuracy of the segmentation results. A research question might be: How can we ensure that the customer data used for the segmentation is of high quality and sufficient for the machine learning algorithms to be effective?
- **Algorithm selection and performance**: There are many different methods and algorithms that can be used for customer segmentation, and it is important to choose the most appropriate approach for the specific needs and goals of the business. A research question might be: How can we select the best method or algorithm for customer segmentation, and how can we evaluate its performance?
- **Segmentation accuracy and effectiveness**: The goal of customer segmentation is to identify distinct groups of customers who will respond positively to tailored marketing campaigns and product offerings. A research question might be: How accurate and effective are the customer segments identified through different methods of customer segmentation, and how do they compare to one another?

II. BACKGROUND

In addition to other generic predictive and descriptive clustering methods like Cluster Regression CHAID [6] usage of Logistic Regression for Supervised Classification. Data mining and machine learning algorithms have been utilized by [7] to segment customers and advise on marketing strategy.

The investigation of how customers use various media and channels in contemporary retail [9]. They employ latent-class cluster analysis as their methodology. Data from 2595 Japanese single source panelists were analyzed by the authors, who discovered seven categories, including characteristics of multichannel aficionados and research consumers.

Customers can be targeted with specialized marketing campaigns thanks to the approach of customer segmentation, which is popular in marketing and data analysis for organizations. This may lead to more profitability and consumer loyalty [10]. Customer segmentation can be especially helpful in the context of online shopping for locating and targeting the most valued customers as well as for locating chances for upselling and cross-selling.

Customer segmentation can be done using a variety of strategies, such as clustering techniques, which separate

the data into groups based on similarities. K-means, Hierarchical clustering, and BIRCH are a few popular clustering techniques. Numerous evaluation measures, such the silhouette comparison and the Calinski-Harabasz index, can be used to compare the effectiveness of various clustering techniques [11].

RFM (Recency, Frequency, Monetary) analysis is another often employed technique for determining a company's most valuable clients. Customers are ranked according to how recently, frequently, and how much they spent on their purchases, with the idea being that the more recently, frequently, and highly-priced the transactions, the more probable it is that the consumer would make more purchases in the future. Customer segmentation and RFM analysis can be coupled to more effectively target marketing campaigns and allocate resources [5]. In conclusion, customer segmentation and RFM analysis are useful tools for companies looking to better understand and target their clientele, especially in the context of online shopping. Businesses can increase the effectiveness of their marketing initiatives and use money more wisely by combining these strategies.

Using unsupervised learning approaches, the authors describe a unique method for client segmentation in online shopping. Using a real-world online retail data set, they use their method to prove how successful it is and compare its performance to that of existing supervised and unsupervised techniques [2].

He & Li [8] suggested a three-dimensional strategy for improving customer lifetime value, customer satisfaction, and customer behavior. The authors believe that each customer is unique and has different needs and expectations, and that segmenting customers to understand their needs and expectations can lead to better service.

Cho & Moon [3] proposed a personalized recommendation system using weighted frequent pattern mining. The system uses customer profiling based on the RFM concept to identify potential customers and assigns different weights to each transaction in order to generate weighted association rules through mining. This model aims to increase profits by providing more accurate recommendations to customers.

Zahrotun [12] used Customer Relationship Management to identify the best customers using online customer data. The author applied the CRM concept to online purchasing and segmented potential customers to identify them, which helped increase corporate revenues. The Fuzzy C-Means Clustering Method was used to accurately segment customers and target them with marketing, allowing customers to receive specialized services across various categories using appropriate labeling techniques.

Shah & Singh [13] proposed a novel clustering algorithm that functions similarly to the K-means and K-medoids algorithms, which both use a partitioning strategy. While the suggested technique may not always be the best option, it does reduce the cluster error criterion. According to Saurabh, the novel method performs faster than traditional methods as the number of clusters increases.

III. Methodology

A. Data

The Online Retail II dataset from Kaggle contains data on transactions occurring for a UK-based online retail company between 01-12/2010 and 12-12/2011. The data includes information on each transaction, such as the date of the transaction, the customer ID, the product ID, the quantity, and unit price of the products purchased, and the country where the customer is located.

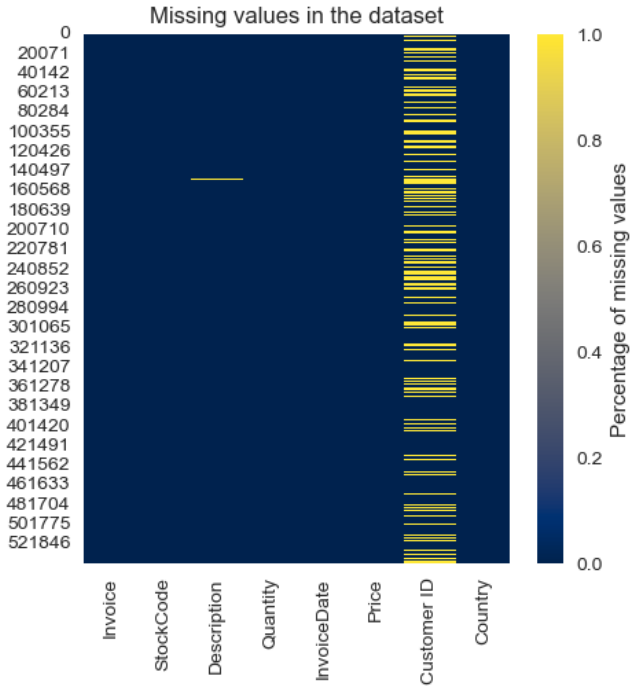
The following is a list of the columns included in the dataset, along with a brief description of each:

- **Invoice**: A unique identifier for the transaction.
- **StockCode**: A unique identifier for the product.
- **Description**: A description of the product.
- **Quantity**: The quantity of the product purchased by the customer.
- **InvoiceDate**: The date and time of the transaction.
- **Price**: The unit price of the product.
- **CustomerID**: A unique identifier for the customer.
- **Country**: The country where the customer is located.
- The data includes 16,487 transactions and 4,372 unique customers. The dataset does not include any information about the customer's demographics or purchase history beyond what is included in the individual transactions.

B. Data Cleaning and Pre-processing

1) Negative values: Some transactions have negative ‘Quantities’ and ‘Price’ values, so they must be eliminated.

2) Missing Values: Transactions with missing ‘CustomerID’ and ‘Description’ values must be eliminated.



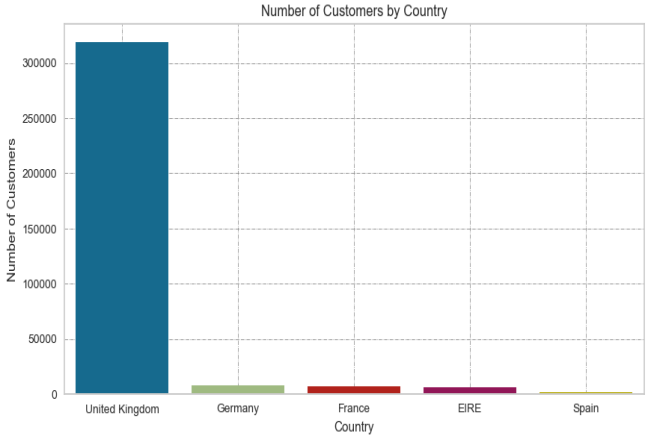
3) Uniqueness: Unique values of ‘Description’ and ‘StockCode’ should be equal since each ‘StockCode’

represents a ‘Description’. So, we eliminate items that have several ‘StockCode’ and ‘StockCode’ that belongs to several items.

4) Add ‘Total Amount’ column: Multiply ‘Price’ by ‘Quantity’

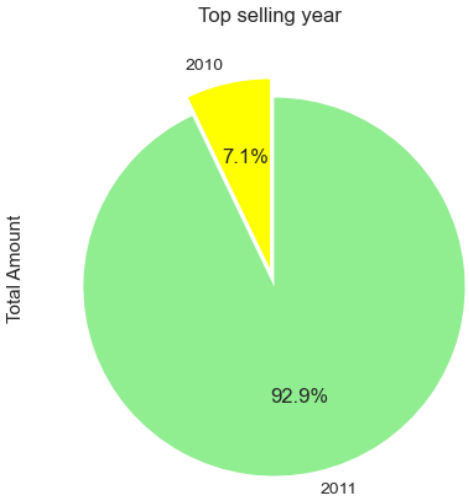
C. *Data Analysis*

1) Graph below shows that the United Kingdom has the highest ratio of customers. So, we decided to analyze everything related to this country.



2) Analysis says that (PAPER CRAFT, LITTLE BIRDIE) are the most popular items based on quantity sold.

3) The pie chart shows that the year 2011 has the most total amount of sales.



4) Table and chart shows that December 2010 has the highest total amount of sales.

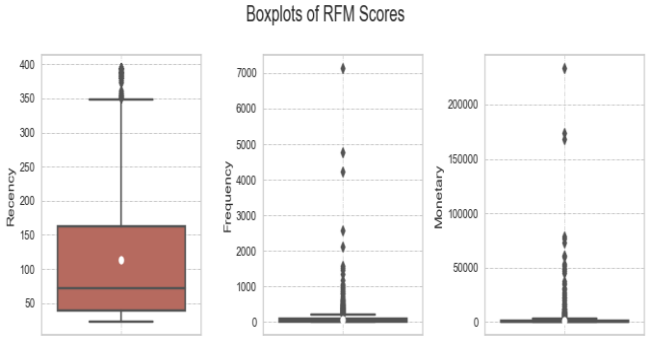
	Year_Month	Total Amount
0	2010-12	455770.400
1	2011-01	396712.670
2	2011-02	325781.830
3	2011-03	431189.440
4	2011-04	368477.131

D. *Data Preparation*

1) RFM (Recency, Frequency, Monetary) is a customer segmentation technique that is used to segment customers based on their historical purchasing behavior. The technique is based on three key metrics:

- **Recency**: The recency is calculated by taking the maximum ‘InvoiceDate’ for each customer and subtracting it from a fixed date in the future (2012-01-01 in this case because the maximum date in the dataset is in 2011).
- **Frequency**: The frequency value is calculated by counting the number of ‘Invoice’ for each customer.
- **Monetary**: The monetary value is calculated by summing the ‘Total Amount’ for each customer.

2) **Outliers**: Figure shows that there is a lot of outliers in the Frequency and Monetary features which should be eliminated.



3) **Scaling and Normalization**:
A data preprocessing method called normalization scales the values of numerical columns in a dataset to a similar scale without distorting variations in their ranges or distributions. For some statistical methods or machine learning algorithms that rely on the assumption that the input variables are on a same scale [7], this can be crucial information.
In the context of our project, normalization might be useful if we are using machine learning algorithms to perform customer segmentation or if we are using statistical techniques that assume that the variables are on a similar scale.
There are many different techniques for normalizing data, and the appropriate technique will depend on the characteristics of the data and the goals of the analysis. The technique that is utilized in this project for normalizing is:

- Min-max scaling: Scales the values of the columns between 0 and 1.

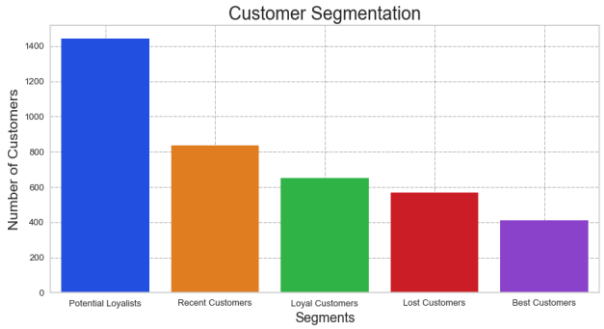
It is important to note that normalization is not always necessary, and in some cases it can even be counterproductive. Whether or not to normalize the data is a decision that should be made based on the characteristics of the data and the requirements of the analysis.

4) **RFM Quantiles**:
In the RFM quantile part of the data preparation process, we calculate the RFM quantiles for each customer based on their recency, frequency, and monetary values. RFM quantiles scores are centred between 1 and 4. 1 means strong and 4 means weak. Once we have calculated the quantiles, we can use these scores to calculate the RFM scores.

$$RFM_{Score} = Recency_{Quartile} + Frequency_{Quartile} + Monetary\_Quartile$$

Customer ID	Recency	Frequency	Monetary	Recency_Quartile	Frequency_Quartile	Monetary_Quartile	RFM_Score
12346.0	163	91	1401.32	3	2	2	7
12747.0	163	91	1401.32	3	2	2	7
12748.0	163	91	1401.32	3	2	2	7
12749.0	163	91	1401.32	3	2	2	7
12820.0	25	47	734.22	1	2	2	5
12821.0	236	4	62.60	4	4	4	12
12822.0	92	42	884.18	2	2	2	6
12823.0	96	5	1759.50	2	4	1	7
12824.0	81	20	324.84	2	3	3	8
12826.0	24	84	1345.02	1	2	2	5

5) **RFM Segmentation**:
In this part we have assigned a RFM group based on the RFM score thresholds. RFM scores that falls between 3 and 4 are called *Best Customers*, between 5 and 6 are called *Loyal Customers*, between 7 and 8 are called *Potential Loyalists*, between 9 and 10 are called *Recent Customers* and lastly between 11 and 12 are called *Lost Customers*.
This way of segmentation will show us the type of customers that we have.

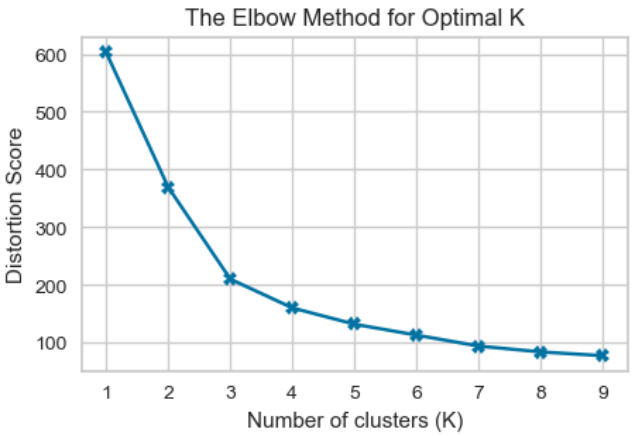


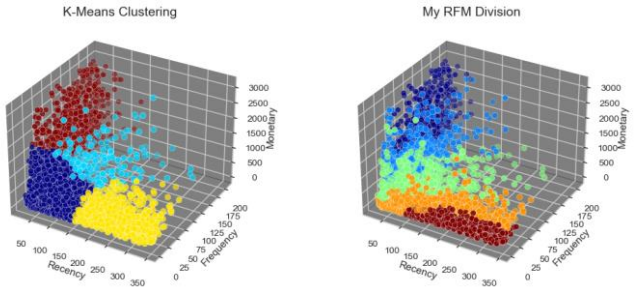
IV. MODELS

Clustering models has been constructed then compared to the RFM segmentation that we have divided, to evaluate how the clustering models are performing. We can compare the results according to the figures below.

A. K-means

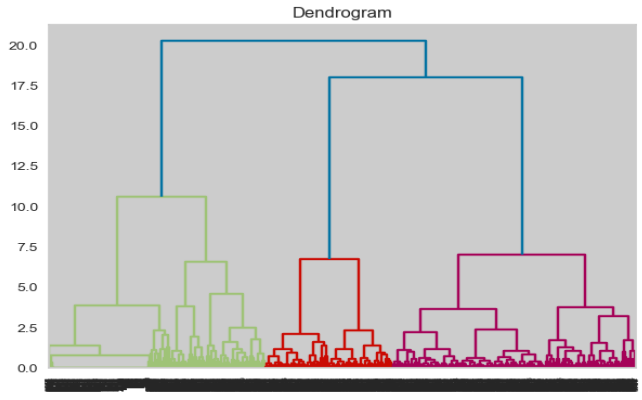
K-means is a well-liked unsupervised learning technique for grouping related data together. K-means was employed in this study to divide the users in the Online Retail II dataset into various groups according to their purchase habits [14]. K-means analysis revealed that the purchase patterns of the clusters detected by this method were clearly different, with certain clusters being more valuable to the business than others.

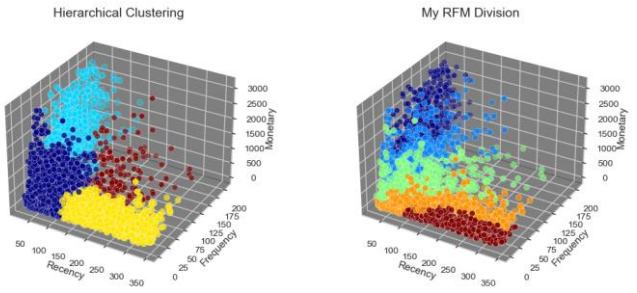




B. Hierarchical

Agglomerative Hierarchical Clustering using the ward linkage method is a type of hierarchical clustering algorithm that starts with each data point as its own individual cluster. It iteratively merges the closest clusters together until a stopping criterion is met, similar to the standard Agglomerative Hierarchical Clustering. The difference is that the ward linkage method uses variance-minimizing criteria to decide which clusters to merge. When merging two clusters, the method chooses the merge that leads to the minimum increase in total within-cluster variance. This linkage method tends to produce more balanced clusters and is particularly useful when the clusters have similar sizes. The result of the algorithm is a tree-based representation of the data called a dendrogram which can be used to determine the number of clusters and the structure of the clusters.



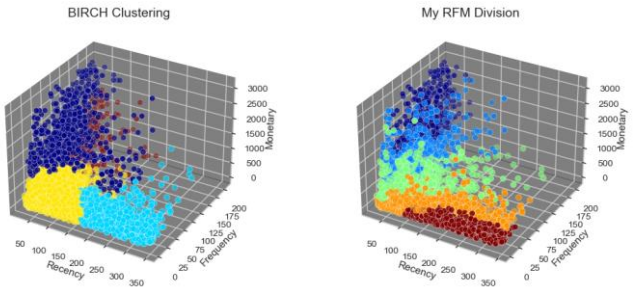


C. *BIRCH*

BIRCH is a data clustering algorithm that was developed to address the scalability issues of traditional clustering methods, such as k-means, on large data sets. It works by constructing a tree-like structure called a BIRCH tree, which organizes the data points into clusters.

The BIRCH tree is constructed in two phases:

In the first phase, the data points are processed in small chunks, called "micro-clusters", which are stored in the tree as leaf nodes. In the second phase, the micro-clusters are merged into larger clusters, which are represented as non-leaf nodes in the tree. One of the key benefits of the BIRCH algorithm is that it allows for incremental cluster construction, meaning that new data points can be easily incorporated into the existing tree structure, without the need to rebuild the entire model [9]. This makes it well-suited for use in online and streaming data applications.



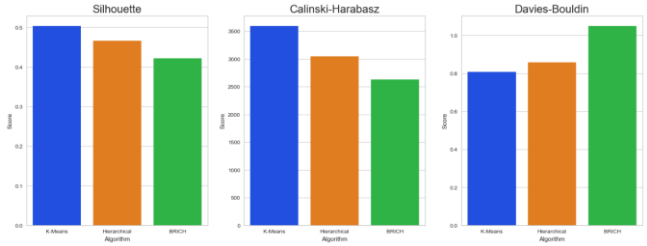
V. Evaluation

The Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index are all evaluation metrics that can be used to compare the performance of different clustering methods.

Silhouette score measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1. A score closer to 1 indicates that the object is well matched to its own cluster, while a score closer to -1 indicates that the object may have been assigned to the wrong cluster. According to this technique we can observe that K-means conducted the best among other algorithms on the dataset in this figure.

Calinski-Harabasz index is also known as the Variance Ratio Criterion. It is a ratio of the between-cluster variance to the within-cluster variance. A higher value indicates a better clustering, because it means that the clusters are more distinct from one another. According to this technique we can observe that K-means conducted the best among other algorithms on the dataset in this figure.

Davies-Bouldin index is a measure of the similarity between each cluster and its most similar cluster. The lower the value the better, because it means the clusters are more separated from one another. According to this technique we can observe that BIRCH conducted the best among other algorithms on the dataset in this figure.



	Metrics	Algorithm	Score
0	Silhouette	K-Means	0.503787
1	Silhouette	Hierarchical	0.466544
2	Silhouette	BRICH	0.421790
3	Davies-Bouldin	K-Means	0.806848
4	Davies-Bouldin	Hierarchical	0.857124
5	Davies-Bouldin	BRICH	1.049782
6	Calinski-Harabasz	K-Means	3595.681664
7	Calinski-Harabasz	Hierarchical	3048.837488
8	Calinski-Harabasz	BRICH	2633.278598

VI. Result

The results of our customer segmentation analysis using clustering techniques (K-means, Hierarchical, and BIRCH) in combination with the RFM technique show that both K-means and BIRCH perform well in identifying different customer segments. Our RFM segmentation, which divided customers into "Best," "Loyal," "Potential Loyalists," "Recent," and "Lost" groups, was used as a benchmark to compare the performance of the clustering models. Two of the techniques we used found that K-means performed best, while the third technique found that BIRCH performed well. Overall, our analysis suggests that both K-means and BIRCH can be useful in identifying and understanding customer segments for businesses looking to improve their marketing strategies. Figure below shows for which cluster number do the customer belongs to.

Customer ID	Recency	Frequency	Monetary	Kmeans_segments	Hierarchical_segments	BIRCH_segments
12346.0	163	91	1401.32	3	3	0
12747.0	163	91	1401.32	3	3	0
12748.0	163	91	1401.32	3	3	0
12749.0	163	91	1401.32	3	3	0
12820.0	25	47	734.22	1	0	2
12821.0	236	4	62.60	2	2	1
12822.0	92	42	884.18	1	0	2
12823.0	96	5	1759.50	1	0	0
12824.0	81	20	324.84	1	0	2
12826.0	24	84	1345.02	0	1	0

A. *Limitations and Challenges*

1) **Limited data**: The dataset only includes data on transactions occurring between December 1, 2010, and December 12, 2011, and does not include any information about the customer's demographics or

purchase history beyond what is included in the individual transactions. This could limit the scope of the analysis and make it difficult to draw conclusions about the customers and their purchasing behavior over time.

2) **Data quality**: The quality of the data can have a big impact on the accuracy and reliability of the analysis. It is important to carefully check the data for errors or inconsistencies and to address any issues that are identified.
3) **Choosing the right approach**: There are many different techniques that can be used for customer segmentation and RFM analysis and choosing the right approach can be challenging. It is important to consider the characteristics of the data and the goals of the analysis when selecting an approach.
4) **Evaluating the effectiveness of the approach**: Once an approach has been selected, it is important to evaluate its effectiveness in terms of the quality of the segments or clusters that are produced and the usefulness of the resulting insights for the business. This can be challenging, as there is often no "right" or "wrong" way to segment customers, and the effectiveness of the approach will depend on the specific context of the business.
5) **Dealing with large and complex datasets**: Working with large and complex datasets can be time-consuming and resource-intensive, and it can be challenging to extract useful insights from the data. It is important to have the right tools and resources to handle the data effectively.
6) **Ensuring data privacy and security**: When working with customer data, it is important to ensure that the data is handled in a way that protects the privacy and security of the customers. This may require taking steps such as de-identifying the data or implementing appropriate security measures to prevent unauthorized access to the data.

VII. Conclusion

In conclusion, customer segmentation through clustering techniques such as K-means, Hierarchical, and BIRCH, combined with the RFM technique, is an effective method for identifying and understanding different groups of customers. By cleaning, scaling, and normalizing the data, and visualizing and analyzing it, we were able to prepare the data for the clustering models. After applying the models and comparing the results using different techniques, we were able to identify the most appropriate model for our data. This approach can provide valuable insights for businesses to improve their marketing strategies and better target their customers.

References

[1] S. Bano and N. Khan, "A Survey of Data Clustering Methods," *International Journal of Advanced Science and Technology*, vol. 113, April 2018.

[2] D. Chen, S. Sain and K. Guo, "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, September 2012.

[3] Y. Cho and S. C. Moon, "Weighted Mining Frequent Pattern based Customer's RFM Score for Personalized u-Commerce Recommendation System," *J. Converg.*, vol. 4, pp. 36-40, January 2013.

[4] A. J. Christy, A. Umamakeswari, L. Priyatharsini and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, pp. 1251-1257, 2021.

[5] P. Sarvari, A. Ustundag and H. Takci, "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, vol. 45, pp. 1129-1157, August 2016.

[6] B. Cooil, L. Aksoy and T. Keiningham, "Approaches to Customer Segmentation," *Journal of Relationship Marketing*, vol. 6, pp. 9-39, January 2007.

[7] C. Dullaghan and E. Rozaki, "Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, p. 13–24, January 2017.

[8] X. He and C. Li, "The Research and Application of Customer Segmentation on E-Commerce Websites," *2016 6th International Conference on Digital Home (ICDH)*, pp. 203-208, 2016.

[9] S. Nakano and F. N. Kondo, "Customer segmentation with purchase channels and media touchpoints using single source panel data," *Journal of Retailing and Consumer Services*, vol. 41, pp. 142-152, 2018.

[10] R. Pradhan, "Customer Segmentation Using Clustering Approach Based on RFM Analysis," in *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, 2021.

[11] V. R. Maddumala, H. Chaikam, J. S. Velanati, R. Ponnaganti and B. Enuguri, "Customer Segmentation using Machine Learning in Python," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, June 2022.

[12] L. Zahrotun, "Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering," *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 299-303, 2017.

[13] S. S. Shah and M. Singh, "Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm," *2012 International Conference on Communication Systems and Network Technologies*, pp. 435-437, 2012.

[14] M. R. K. Ibrahim and R. Tyasnurita, "LRFM Model Analysis for Customer Segmentation Using K-Means Clustering," *2022 International Conference on Electrical and Information Technology (IEIT)*, September 2022.