E-commerce Customer Segmentation and Recommendation

Fathima Sharsoom | Manasi Jaykisan Dudhane | Siddarth Goud Gundrathi | Siddharth Miteshkumar Haveliwala | Swarupa Rani Danthuluri | Saivomanesh Chekuri

*Abstract*—This project focuses on the strategic implementation of customer segmentation within the realm of e-commerce, employing advanced methodologies such as the K-means clustering algorithm and RFM analysis. Leveraging clustering techniques, including the Elbow, Silhouette methods and gap statistics method we determine the optimal number of clusters for our dataset, unraveling distinct patterns and preferences within our diverse customer base. The K-means clustering visualization provides a graphical representation of customer clusters, aiding in the identification of diverse customer segments with unique characteristics. The application of the Pareto Principle further refines customer segmentation by pinpointing the top revenue generating customers. The nuanced segmentation allows for tailored strategies to address the unique needs of each customer group, facilitating effective approaches to customer engagement and retention.

# Introduction

In the rapidly evolving business environment, with unheard-of technological breakthroughs and dynamic market trend comprehending consumer behaviour has become essential for enterprises looking to establish and maintain a competitive edge.

Comparable to assembling a vast puzzle, customer segmentation involves the intricate process of categorizing individuals into distinct groups based on shared traits, behaviours, or preferences. Each segmented group represents a unique set of customers with similar tastes, buying habits, or interests. This strategic organizational approach empowers businesses to tailor their product offerings and marketing strategies to meet the specific needs of each segment, fostering a deeper connection with their diverse clientele.

The fundamental principle underlying customer segmentation lies in the ability to understand and respond to the nuanced demands of various consumer groups. For instance, a business identifying a segment with a keen interest in technology may curate specialized items or exclusive offers on the latest devices, thus aligning with the preferences of that particular customer subset. By comprehensively comprehending these diverse groupings, companies not only enhance the potential for customer satisfaction and loyalty but also optimize their communication strategies for maximum efficiency.

Looking towards the future, the landscape of customer segmentation is poised for significant transformation, driven by the integration of cutting-edge technologies. The advent of artificial intelligence and machine learning holds the promise of forecasting future customer preferences based on their current behaviours. This predictive strategy empowers businesses to stay ahead of evolving consumer demands, ensuring they can adapt their product and service offerings to suit changing trends. Moreover, the future of customer segmentation envisions seamless, real-time interactions across various platforms, allowing companies to engage with customers whether they are online or offline.

In essence, customer segmentation represents a proactive and dynamic approach that enables companies to continually adjust and provide personalized experiences in an ever-evolving market. By leveraging advanced technologies and staying attuned to shifting consumer behaviours, businesses can not only meet but exceed customer expectations, fostering lasting relationships in the competitive landscape of the business world.

# Problem Definition & Questions

Despite the abundance of data-driven tools and analytics, businesses are finding it difficult to interpret the nuances of consumer behavior in real-time, which is hindering their ability to make informed strategic decisions. The fundamental problem lies in complexity and dynamic nature of consumer preferences, which are influenced by a variety of factors including changing market dynamics, technical advancements, and socioeconomic considerations. Traditional market research methods often fall short in capturing the fluidity of customer behavior, leading to an information gap that hampers businesses from effectively tailoring their products, services, and marketing strategies to meet the ever-changing needs and expectations of their target audience.

Furthermore, the advent of digital platforms, social media, and e-commerce has exponentially increased the volume and velocity of customer data generated daily. Although this situation offers businesses a chance to extract valuable information, they encounter the difficulty of sorting through the abundance of data to pinpoint significant trends that can lead to actionable insights. The absence of a thorough grasp of customer behavior not only obstructs the creation of personalized customer experiences but also obstructs the enhancement of supply chain management, inventory, and resource allocation. This problem is exacerbated by the fact that consumer expectations are constantly evolving, influenced by global events, cultural shifts, and emerging technologies. Companies that do not align their strategies with these changes jeopardize their competitive advantage in the market. In light of these challenges, there is an urgent need for businesses to invest in sophisticated data analytics to decode the intricate web of customer behavior.

**Questions:**

1. How can businesses establish a feedback loop using customer behaviour data to continuously improve overall customer satisfaction?
2. What are the most important patterns and trends in the defined client segments, and how may these guide targeted marketing campaigns?
3. How can each identified segment's sales and customer engagement be improved through the use of a recommendation engine?
4. How might an in-depth analysis of customer behaviour influence the development and modification of services to better cater to specific customer preferences?
5. How can businesses leverage customer behaviour data to create more personalized and engaging experiences across various touchpoints?
6. How might dynamic pricing strategies be implemented based on real-time customer behaviour data to optimize pricing for different segments and maximize revenue?

# DATA DESCRIPTION

The dataset used in this analysis was obtained from the UC Irvine Machine Learning Repository and specifically originates from the Online Retail dataset. The dataset includes every transaction made by a UK-based, registered online retailer without a physical store between December 1, 2010, and December 12, 2011. The dataset under consideration exhibits distinctive features that characterize its nature, including being sequential, multivariate, and time-series in structure. It is centered around the topic area of "Company," suggesting a focus on business-related aspects. The dataset encompasses various types of features, namely characters and integers, reflecting a diverse range of data representations. With a substantial amount of information, the dataset comprises 541,909 observations, and it is organized across six variables, indicating the richness and complexity of the data captured within this analytical framework.

## Description of the variables in the dataset:

The “Invoice No” serves as a distinctive identifier for each transaction, denoted by a six-digit nominal number. Notably, if the initial letter is 'c,' it indicates a cancellation code. “Stock Code”, on the other hand, functions as the code assigned to each product, represented by a five-digit integral nominal number unique to each item. The “Description” field contains the name of the product and is considered notional. “Quantity” refers to the quantitative aspect, representing the number of units of a product involved in a particular transaction. The “Date of Invoice” encompasses both the time and date of the transaction, providing information on its generation. “Unit Price” denotes the cost per unit in sterling for the respective product. “Customer ID” is a unique five-digit nominal number assigned to each customer, playing a crucial role in customer identification. Lastly, “Nation” captures the name of the nation to which each client belongs, represented as a nominal field.

## Data Observations

The dataset comprises quantities with both positive and negative values, ranging from a minimum of -80,995 to a maximum of 80,995. Negative quantities may signify refunds or cancellations. Unit prices in the dataset exhibit a minimum of -11,062.06 and a maximum of 38,970.00, with negative unit costs potentially necessitating further investigation. The Customer ID variable contains 135,080 missing values (NAs), indicating instances where client data is absent, possibly suggesting transactions without associated client information. The dataset encompasses transactions from various countries. The Invoice Date variable is of character data type, and for effective time-related analysis, it is recommended to convert it to a datetime type.

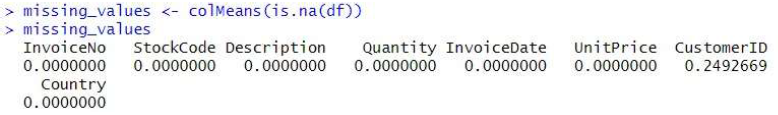
Time stamps for every transaction are contained in the characters that make up the Invoice Date variable. Negative unit prices and quantities in particular call for careful examination since they can be signs of inaccuracies or anomalies in the data. Before performing in-depth analyses or modelling on this sales dataset, more investigation and data pretreatment is advised, such as addressing missing values and converting Invoice Date to a datetime format.

# Data Preparation

## A. Data Screening

Data collection and pre-processing is a vital step of any data-driven study. Once data is collected, pre-processing becomes imperative to enhance its usability for analysis. This involves cleaning and organizing the raw data, addressing issues such as missing values, outliers, and inconsistencies. Properly formatted and cleaned data not only ensures the accuracy of analytical results but also facilitates the application of various statistical and machine learning techniques. Additionally, pre-processing may involve transforming variables, scaling features, or encoding categorical data to make it suitable for modeling.

## Missing Data

Fig. 1 Missing Values in DataSet

Upon examining the dataset, we identified missing data, quantifying the absence of values for each variable. Notably, the "Customer ID" variable exhibits a relatively high rate of missing values at 24.93%, signifying that approximately a quarter of the records lack customer identifying information. Addressing these missing values requires a thoughtful approach, considering factors such as the data's nature and specific analytic objectives. Potential strategies involve imputation, where missing values are replaced with estimated values, or the removal of records containing missing values. In this particular case, our chosen method involves excluding records with missing values from the dataset.

## Outliers Analysis

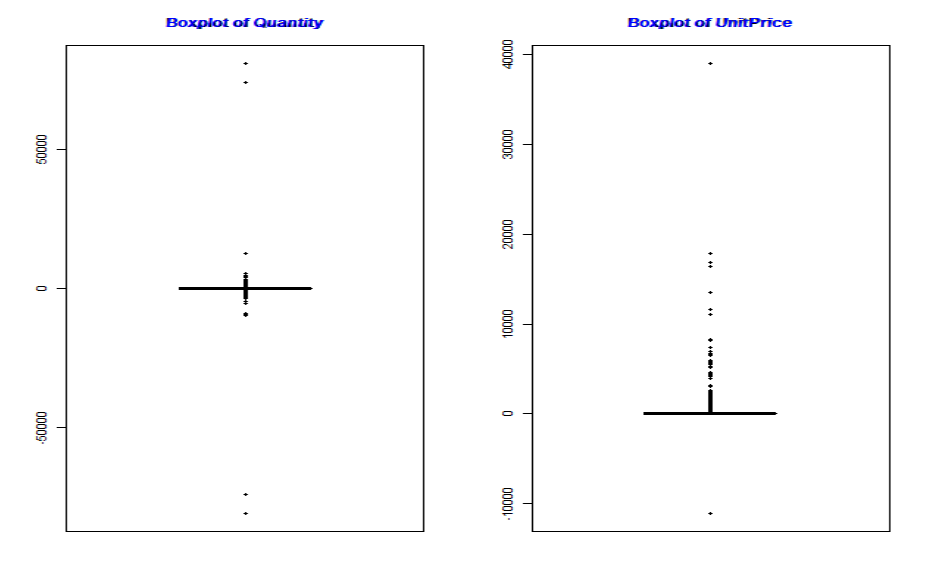


Fig. 2 Boxplot for Quantity & Unit Price

The analysis of the dataset reveals a substantial presence of outliers in both the "Quantity" and "UnitPrice" fields, indicating potential anomalies in purchasing behavior and item costs. The outliers in the "Quantity" variable, representing the number of items bought in each transaction, may signify unusually high or low buying amounts. Similarly, the "UnitPrice" variable, denoting the cost of items per unit, exhibits several outliers that could indicate atypical pricing for specific items. Given the critical nature of these variables, it becomes imperative to delve into the nature of these outliers to discern whether they reflect legitimate data points or stem from potential input errors.

Investigating the characteristics of these outliers is essential for distinguishing between genuine data points and recording errors. For instance, the presence of large negative quantities might suggest cancellations or returns, emphasizing the need for a nuanced examination. To ensure data integrity, a meticulous review of the outliers in both the "Quantity" and "Unit Price" variables is conducted, leading to their removal from the dataset. This proactive approach not only safeguards the accuracy of subsequent analyses but also fosters a more reliable foundation for deriving meaningful insights from the data.

## Linearity & Normality Analysis

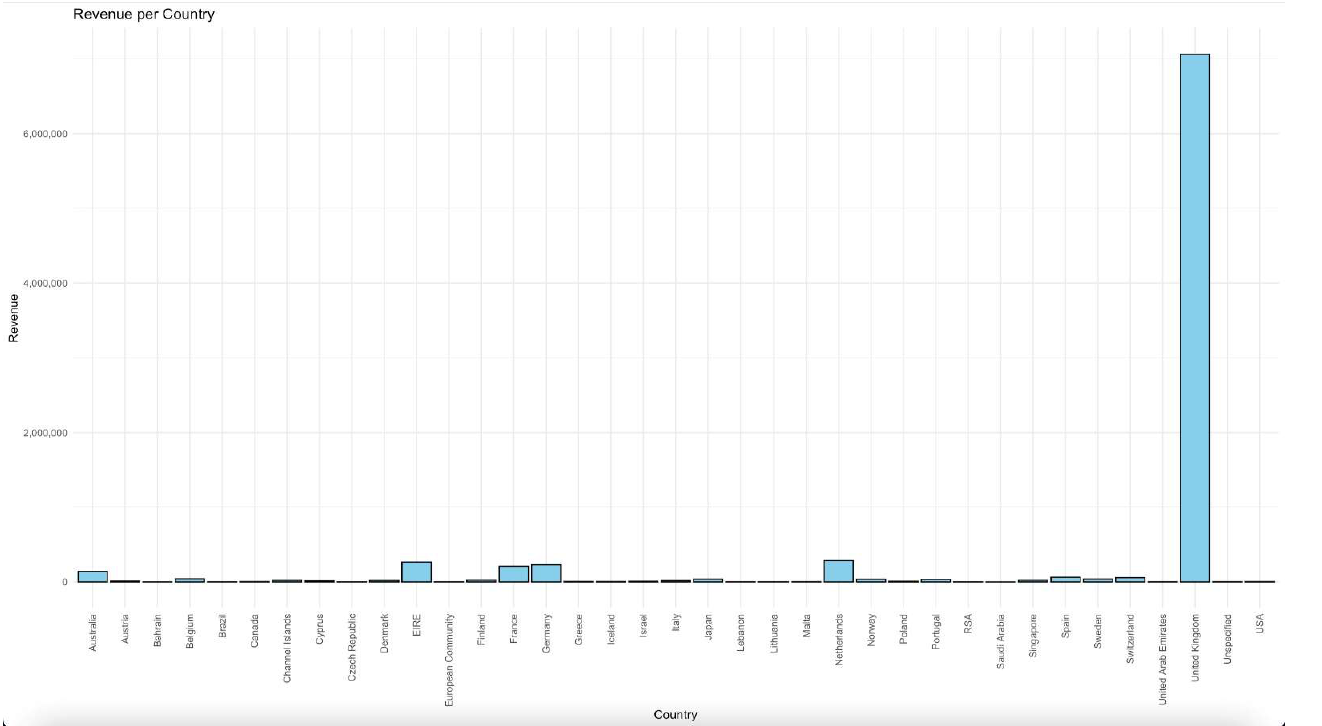


Fig. 3 Histogram of revenue per country

In conducting a normality analysis, the distribution of revenue across different nations was visually depicted through histograms. Each bar in these histograms represents the frequency or count of observations within a specific range of revenue numbers, providing a comprehensive view of the dataset's distribution. The height of each bar indicates the frequency of observations falling within corresponding revenue ranges.

The observation is particularly noteworthy in the context of the United Kingdom, as the histogram underscores its prominence among other nations. The larger number of observations for the United Kingdom implies that transactions originating from this country constitute a substantial portion of the dataset, contributing significantly to the overall income depicted in the analysis.

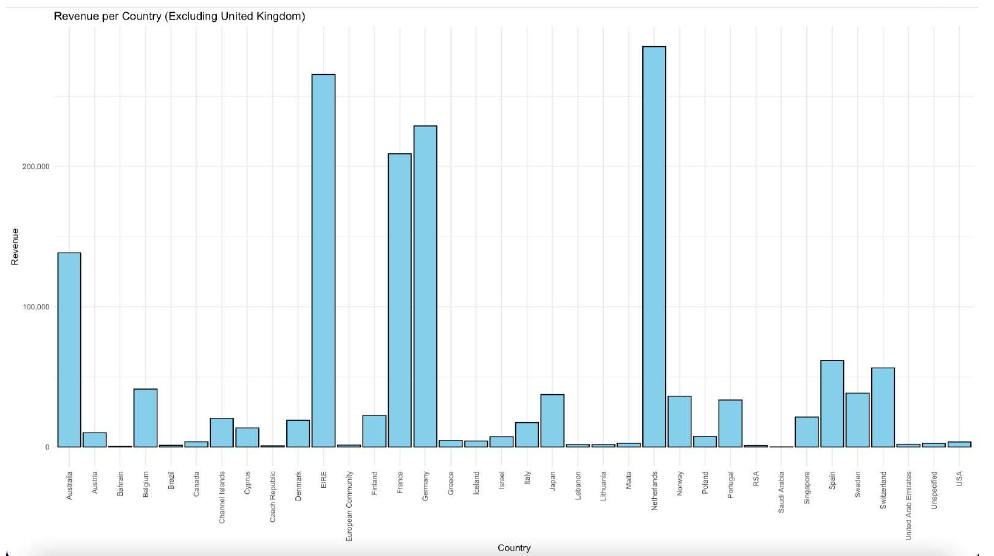


Fig. 4 Histogram of revenue per country (Excluding UK)

To gain a more granular understanding of revenue distribution across various countries, a strategic decision was made to exclude the United Kingdom from the analysis. By doing so, the focus shifts specifically to the revenue contributions of other nations, allowing for a clearer and more nuanced examination of their individual impacts on the dataset. This exclusion of the United Kingdom ensures that the revenue patterns and trends of other countries are more distinctly visible, facilitating a comprehensive analysis of their respective roles in the overall revenue landscape

## Data Reduction

In this particular stage of the process, the primary objective was to address the issue of missing values within the dataset, a common challenge encountered in statistical studies and modeling. Typically denoted by terms such as NA or NULL, missing values can introduce complications in data analyses. To mitigate this concern, a strategy of data reduction was employed, specifically targeting observations with missing values. By eliminating these instances, the dataset underwent simplification, rendering it more suitable for subsequent thorough examination.

The focus of the data reduction phase was to enhance completeness by systematically removing observations (rows) containing missing values. This strategic removal was particularly emphasized in the context of the "Customer ID" variable to improve overall dataset quality. While this process contributes to the precision of subsequent studies and modeling efforts, it concurrently prompts important considerations regarding potential biases introduced by the removal procedure. Striking a balance between data completeness and the potential impact of removal procedures becomes pivotal in ensuring the reliability and accuracy of future analyses and modeling endeavors.

## B. Validity and reliability analysis

To enhance validity, we rigorously applied data preprocessing techniques, including data cleaning, ensuring that the information collected is a faithful representation of the underlying e-commerce transactions. This attention to validity is paramount as it safeguards against drawing incorrect or misleading conclusions about customer segmentation and recommendation strategies.

In terms of reliability, we adopted a meticulous approach to guarantee the consistency and stability of our results. Our decision to use an authentic source for the dataset, along with a thorough data preprocessing phase, aimed to mitigate sources of variation, noise, and measurement errors. Despite starting with over 500,000 data points, we carefully cleaned and processed the data, resulting in a reliable dataset of 430,000 samples. This reliability is essential for the subsequent stages of analysis and segmentation, ensuring that our findings are consistent and reproducible. The reliability of our data is instrumental in building trust in the outcomes of our research, allowing for robust customer segmentation and recommendation strategies.

To further strengthen the validity and reliability of our research, we employed diverse data collection and analysis methods, including data triangulation and cross-validation. The use of established instruments, such as the dataset from UCI ML Repository, contributed to the overall robustness of our research. Additionally, we welcomed peer review and feedback to continually assess and improve the validity and reliability of our methods. Through a combination of careful sourcing, preprocessing, and validation techniques, we aim to provide accurate, trustworthy, and meaningful insights into e-commerce customer segmentation and recommendation.

# Data Analysis

A central aspect of our analysis revolves around solving the clustering problem, which entails identifying homogeneous groupings of data points within our dataset. This is essential for understanding patterns and trends among customers. For this purpose, we leverage the widely recognized K-means clustering algorithm, a method extensively researched and proven effective in partitioning data points into distinct groups. The primary objective of K-means clustering is to minimize the distance between data points within a cluster while maximizing the distance between different clusters, ensuring that individuals within a group are similar, and those in separate groups are distinctly dissimilar.

## A. Recency, Frequency & Monetary Value

To further enhance our customer segmentation and identification, we integrate the RFM (Recency, Frequency, Monetary Value) analysis, a powerful tool in database marketing, particularly vital in the retail industry. RFM assigns scores to customers based on three key elements: Recency, which measures the time since their last purchase; Frequency, indicating the interval between subsequent purchases; and Monetary, reflecting the amount spent during a specific period. These scores provide valuable insights into customer behavior and preferences.

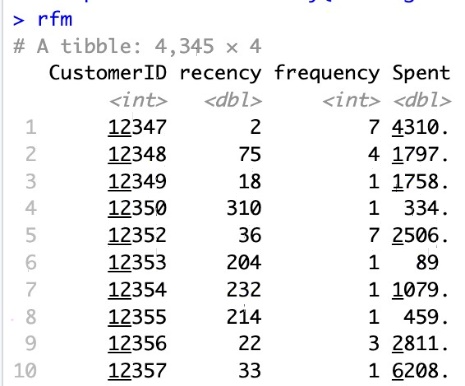


Fig. 5 RFM Scores of each customer

Having previously engineered the total revenue per customer through the incorporation of the "Spent" feature, we streamline our datagram by retaining only the relevant columns, dropping extraneous information. With this refined dataset, encompassing the engineered RFM values and total revenue per customer, we prepare a comprehensive data frame that serves as the foundation for our RFM clustering analysis. The integration of these three crucial dimensions—Recency, Frequency, and Monetary—into a unified data frame sets the stage for a meaningful exploration of customer segments, laying the groundwork for our subsequent analysis in the realm of "E-commerce Customer Segmentation and Recommendation." The K-means clustering algorithm is employed iteratively to partition customers based on their RFM values, utilizing the Euclidean distance metric. The determination of the optimal number of clusters, denoted by K, is a critical step in our analysis.

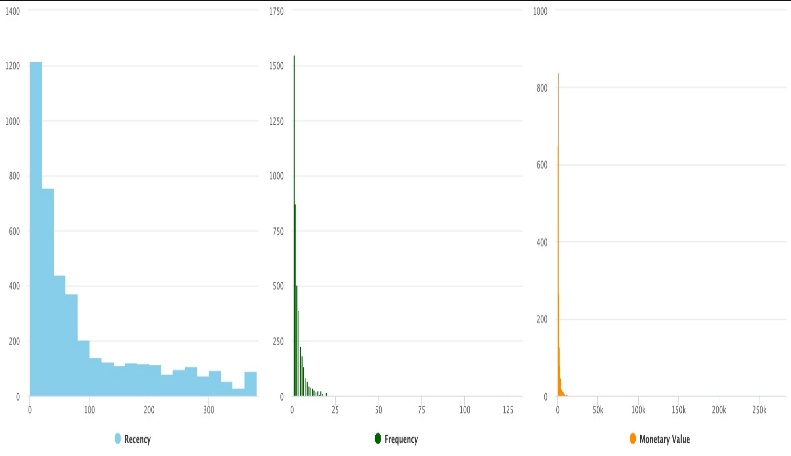


Fig. 6 Graphical Representation of RFM

In analyzing the RFM metrics for our customer base, we observe that the average time since the last purchase, or recency, stands at 91.97 days, with a median recency of 50.00 days. This suggests that, on average, customers made a purchase approximately 92 days ago, and half of our customers made a purchase within the last 50 days. Turning to frequency, the mean number of purchases is 3.869, indicating that, on average, customers make approximately 3.87 purchases. The median frequency is 2.000, illustrating that half of our customers have made two purchases. In terms of monetary value, the average spending per customer is $1994.4, while the median spending is $671.8. Notably, a significant portion of our customer base, 1216 customers to be precise, has a recency of 0 to 20 days, signifying recent engagement. Similarly, the frequency distribution indicates that 1548 customers fall within the range of 1 to 1.5 purchases, highlighting a substantial group making one or one-and-a-half purchases. This analysis provides valuable insights into customer behavior, enabling targeted marketing strategies and customer segmentation for more personalized and effective engagement.

Normalizing the RFM data is essential because it ensures that all variables (Recency, Frequency, Monetary) are on a consistent scale. This process is crucial for accurate analysis, preventing variables with larger magnitudes from dominating and biasing the results. By standardizing the data, we enable fair comparisons between variables and create a foundation for unbiased insights, especially in algorithms sensitive to variable scales. Our methodological approach, combining K-means clustering with RFM analysis and the careful determination of the number of clusters, ensures a robust and insightful exploration of customer patterns. Prior to embarking on the clustering process for our dataset of wholesale customers, it is crucial to identify the optimal number of clusters (k).

## B. Methods used to identify optimal number of clusters (k):

*I. Elbow Method.*

The Elbow Method involves plotting the number of clusters against the corresponding within-cluster sum of squares (WCSS) and observing the chart for an "elbow" point, where the rate of decrease in WCSS slows.

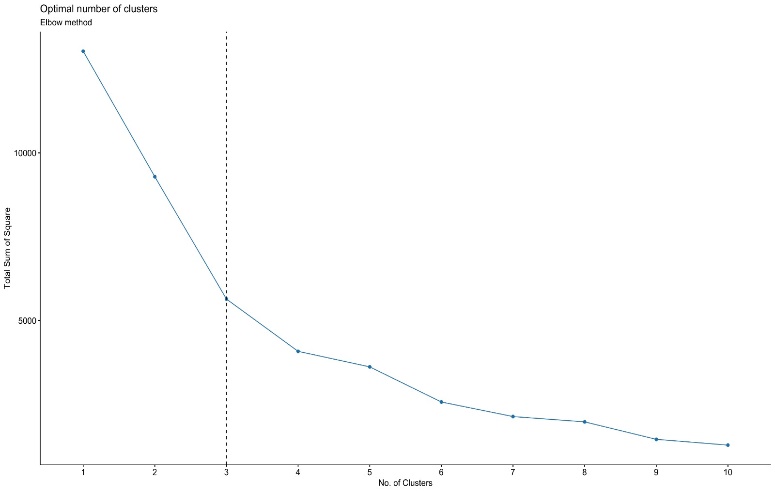


Fig. 7 Elbow Method

In our case, the chart suggests potential cluster values of 3, 4, 5, or 6, and the choice among them can be somewhat subjective based on the slope of the curve as it approaches zero.

*II. Silhouette Method.*

The Silhouette Method evaluates the compactness and separation of clusters. By calculating silhouette scores for different cluster numbers, we identify the number of clusters that maximizes the average silhouette score. A higher silhouette score indicates better-defined and well-separated clusters.

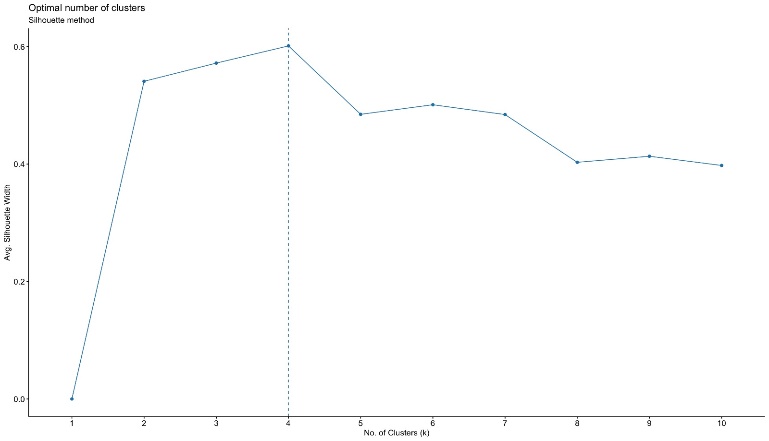


Fig. 7 Silhouette Method

The silhouette plot shows the silhouette co efficient over values of k ranging from 1 to 10. This plot shows the highest average silhouette co-efficient occurring when k = 4.

*III. Gap Statistic Method.*

The Gap Statistic Method involves comparing the WCSS of the actual clustering with the WCSS of a reference random clustering. The optimal number of clusters is determined by identifying the point where the gap between the two WCSS values is maximized.

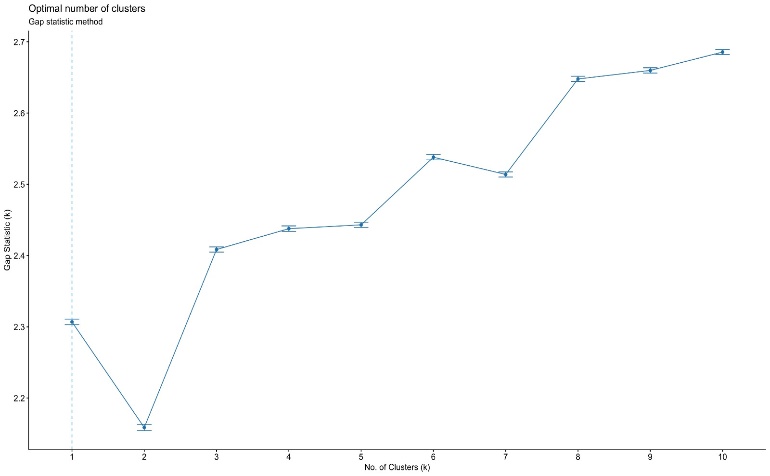


Fig. 7 Gap Statistic Method

As we can see from the graph the rate of increase begins to slow down on k = 3, as such gap statistics proposes that the optimal number of clusters k is 3.

## C. K- Means Clustering Visualization

While these methods provide valuable insights into potential cluster numbers, the choice of the optimal k can still involve a degree of subjectivity, and selecting the most appropriate method depends on the specific characteristics of the dataset and the goals of the analysis. In selecting the number of clusters for our analysis, we considered multiple methods, including the elbow method, silhouette analysis, and gap statistics. In determining the optimal number of clusters for our analysis, we employed various methods, including the elbow method, silhouette analysis, and gap statistics. These methods suggested potential cluster counts of 3, 4, 5, 6, and 10. The evaluation of these options revealed that with k=3, approximately 56.72% of the variance was explained.

**Selection of Optimal Number of Clusters & K-Means Algorithm:**

Choosing k=6 substantially increased the explanatory power to 80.27%. While k=10 offered an even higher explanatory percentage of 90%, the decision was made to opt for k=6 to avoid the risk of overfitting. This choice strikes a balance between capturing the underlying data structure and maintaining a model that is less complex and more interpretable.

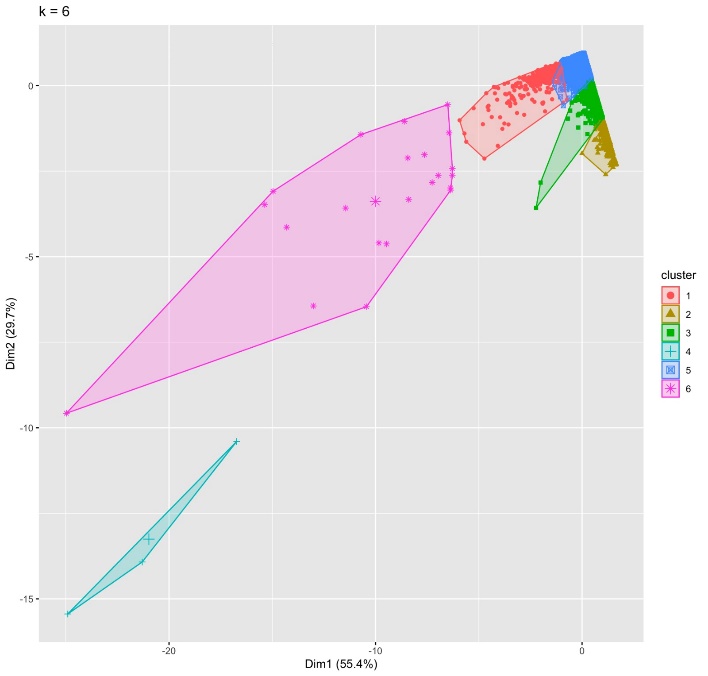


Fig. 8 Visualization of the clustering algorithm results (k= 6)

The cluster graph depicts the distribution of six clusters within a two-dimensional space. The first dimension, Dim1, accounts for 55.4% of the data's variance, while the second dimension, Dim2, accounts for 29.7%. The clusters form a triangular arrangement, with cluster 1 at the top, cluster 6 at the bottom, and the others in between. Additionally, the clusters appear to be separated by gaps, indicating that they are well-defined and distinct from one another. This is further supported by the observation that the maximum distance within each cluster is relatively small compared to the average distance between clusters

In the cluster plot generated from k-means clustering with six clusters, each group exhibits distinct characteristics based on recency, frequency, and spending patterns. Cluster 1, with a centroid of (-0.79, 1.87, 0.65), represents customers with recent and frequent transactions and moderate spending. Cluster 2, characterized by (2.03, -0.44, -0.19), comprises customers with high recency, low frequency, and low spending. Cluster 3, with a centroid of (0.64, -0.32, -0.14), reflects customers with a balanced profile of moderate recency, slightly low frequency, and slightly low spending. Cluster 4, marked by (-0.89, 4.84, 28.42), includes highly engaged customers with strongly negative recency, very high frequency, and exceptionally high spending. Cluster 5, with a centroid of (-0.57, -0.11, -0.08), captures customers with recent transactions but at a lower frequency and moderate spending. Finally, Cluster 6, characterized by (-0.86, 8.28, 6.69), consists of actively engaged customers with negative recency, extremely high frequency, and substantial spending. These numerical interpretations provide a quantitative perspective on the diverse customer segments identified through clustering, facilitating targeted marketing strategies tailored to the specific behaviors of each cluster.

## D. Customer Segmentation using Pareto Principle

The Pareto Principle, the principle implies that a significant portion of a company's revenue is generated by a small percentage of its customers. In our analysis, the Pareto Principle is applied to the RFM scores to identify the top 20% of customers who contribute the most to the company's revenue. These high-value customers exhibit recent, frequent, and substantial transaction behavior, making them crucial for the business's success. By focusing on and retaining this vital segment, companies can strategically allocate resources and tailor marketing efforts to maximize the impact of this select customer group on overall business outcomes.

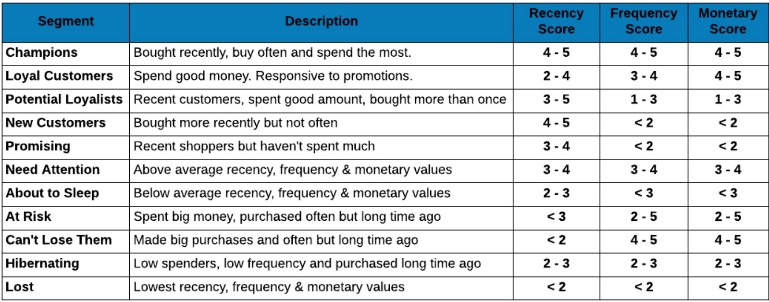


Fig. 9 Customer Segments based on different RFM range

The above table explains about different customer segments based on RFM Score.



Fig. 10 Visual Representation Different Segments

In the process of segmenting customers using the RFM, distinct categories have been established, each defined by specific score ranges. For instance, the "Champion" segment comprises customers with scores equal to or greater than 4 in recency, frequency, and monetary aspects. Customers falling into the "Loyal Customer" category exhibit scores of 2 or more in recency, and 3 or more in both frequency and monetary dimensions. The "Potential Loyalist" group includes customers with a recency score of 3 or higher but with frequency and monetary scores of 3 or less. These delineations continue for various segments, such as "New Customer," "Promising," "Need Attention," "About to Sleep," "At Risk," "Can't lose them," "Hibernating," and "Lost," each defined by specific score combinations within the RFM parameters. This nuanced segmentation allows businesses to tailor their strategies to the unique behaviors and needs of each customer group, aiming for more personalized and effective approaches to customer engagement and retention.

# Results

Our results show compelling insights into customer behavior through the segmentation based on RFM scores. Notably, the "Champion" segment emerges as a powerhouse with a median recency of 10 days, an impressive transaction count of 1590, and a substantial median amount of $2639, attaining a perfect recency score of 5. This indicates a highly engaged and valuable customer group. Conversely, the "At Risk" segment displays a median recency of 236 days, signaling potential churn risks. The "Loyal Customer" segment, characterized by a median recency of 53 days and a median amount of $986, represents a consistently engaged customer base. These findings provide nuanced perspectives on customer behaviors and preferences, offering valuable guidance for businesses to tailor their strategies for customer retention, re-engagement, and overall satisfaction across distinct segments.

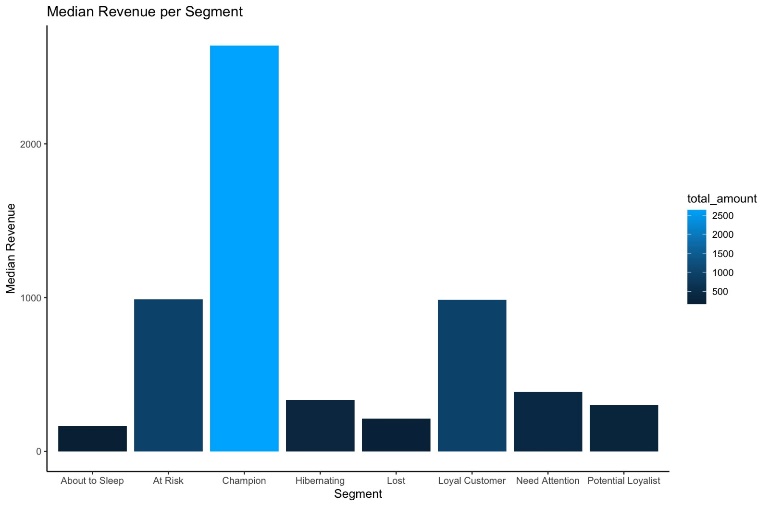


Fig. 11 Median Monetary Value by Segment

# Managerial discussion

By analyzing the historical purchasing behavior of our customers, we aimed to derive actionable insights for targeted marketing, personalized engagement, and improved customer satisfaction.

**Key Findings:**

1. *Segmentation Results*

The RFM analysis successfully segmented our customer base into distinct groups based on their recency of purchase, frequency of transactions, and monetary value.

Identified segments include high-value loyal customers, at-risk customers, and potential high spenders.

*2. Customer Behavior Patterns*

Discovered patterns in customer behavior, such as a decline in engagement for certain customer segments and an increase in spending for others. Recognized the importance of recency in predicting future customer engagement and monetary value.

*3. Marketing Channel Effectiveness*

Evaluated the effectiveness of different marketing channels in engaging customers from various segments. Found opportunities to optimize marketing spend by tailoring strategies to specific customer segments.

**Insights and Implications:**

*1. Tailored Marketing Strategies*

Personalized marketing strategies for each segment are crucial. This involves targeted promotions and communication to re-engage at-risk customers and cultivate potential high spenders.

*2. Resource Allocation*

Recommending a reallocation of marketing resources based on identified segments. By focusing efforts on high-value segments, we can ensure a more efficient and cost-effective approach

*3.Product Recommendations*

Implementing personalized product recommendations based on historical purchase patterns and preferences. Cross-selling and upselling strategies can enhance the overall shopping experience for each segment.

##### References

1. 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.
2. E. Y. L. Nandapala and K. P. N. Jayasena, "The practical approach in Customers segmentation by using the K-Means Algorithm," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 344-349, doi: 10.1109/ICIIS51140.2020.9342639.
3. M. Namvar, M. R. Gholamian and S. KhakAbi, "A Two Phase Clustering Method for Intelligent Customer Segmentation," 2010 International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, UK, 2010, pp. 215-219, doi: 10.1109/ISMS.2010.48. K. Elissa, “Title of paper if known,” unpublished.
4. Chinedu Pascal Ezenkwu, Simeon Ozuomba and Constance kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services” International Journal of Advanced Research in Artificial Intelligence(IJARAI), 4(10), 2015.
5. M. Namvar, M. R. Gholamian and S. KhakAbi, "A Two Phase Clustering Method for Intelligent Customer Segmentation," 2010 International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, UK, 2010, pp. 215-219, doi: 10.1109/ISMS.2010.48.
6. D. Jutla, J. Craig and P. Bodorik, "Enabling and measuring electronic customer relationship management readiness", proceedings of the 34th annual hawaii international conference on system sciences organizational systems and technologies track, pp. 1-10, 2001
7. T. Nelson Gnanaraj, K. Ramesh Kumar and N. Monica, "Survey on mining clusters using new k-mean algorithm from structured and unstructured data", IJACST, 2014.
8. Jayant Tikmani, Sudhanshu Tiwari and Sujata Khedkar, "Telecom customer segmentation based on cluster analysis An Approach to Customer Classification using k-means", IJIRCCE, 2015
9. T. Nelson Gnanaraj, K. Ramesh Kumar and N. Monica, "Survey on mining clusters using new k-mean algorithm from structured and unstructured data", IJACST, 2014.
10. Omar Kettani, Faycal Ramdani and Benaissa Tadili, "An Agglomerative Clustering Method for Large Data Sets", IJCA, 2014.
11. K. Keerthi, G. Lakshmi Thirupathamma, N.Vijayalakshmi, D. Aparna, U. Vineela, "Customer Segmentation Analysis and Visualization", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 280-284, January-February 2021.
12. H.V. Shashidhar, V. Subramanian, Customer segmentation of bank based on data mining security value based heuristic approach as a replacement to k-means algorithm,International Journal Computer Application, 19 (2011), 13-18.
13. Yun Chen, Guozheng Zhang, Dengfeng Hu, Shanshan Wang School of Public Economy Administration of Shanghai University of finance & economics, Shanghai, 20043, CUSTOMER SEGMENTATION IN CUSTOMER RELATIONSHIP MANAGEMENT BASED ON DATA MINING
14. Sulekha Goyat. The basis of market segmentation: a critical review of literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online).Vol 3, No.9, 2011