Advancing Customer Segmentation: A Comparative Study of Hierarchical Level Weighted Method and K-Means Algorithms

1st Anmol Malik
Department of AIT-CSE
Chandigarh University
Punjab, India
21BCS5335@cuchd.in

2th Dr. Gurwinder Singh
Department of AIT-CSE
Chandigarh University
Punjab, India
gurwinder.e11253@cumail.in

***Abstract*—This paper explores into the crucial aspect of customer segmentation and its significance in understanding consumer behavior for businesses. Customer segmentation is the process of grouping customers according to their similar characteristics, behavior, and preferences. The study explores the application of a weighted approach and clustering techniques in the customer segmentation process. The supermarket data-set used in this research comprises transactions data from a supermarket, encompassing details such as date, time, products purchased, individual product prices, and the total transaction amount. Leveraging this data-set, data analysis techniques are employed to gain insights into supermarket customers' behavior. For instance, the relationship between the day of the week and total transaction amounts is studied, factors influencing customer spending patterns are analyzed, and trends in customer behavior are identified. This project focuses on customer segmentation utilizing two distinct approaches: K-means clustering and a weighted method. The first approach involves employing K-means clustering to segment customers based on their characteristics. The second approach utilizes a weighted method, assigning varying importance to specific attributes for enhanced segmentation. The project leverages these two approaches on a data-set to gain valuable insights into customer behavior and preferences. By exploring these techniques, businesses can tailor marketing strategies effectively, improve customer engagement, and optimize overall performance for success.**

Keywords: Customer Segmentation, Consumer Behavior, EDA, Clustering Techniques, Hierarchical level weighted method, K-means clustering

1. Introduction

Businesses in a market that is undergoing fast change have to cope with intense competition and shifting consumer demands. Businesses must ensure their clients are satisfied in order for them to flourish and do well. Offering customised services and marketing tactics that are adapted to each customer's tastes and needs is one successful strategy. One strategy that may be used to aid with this is customer segmentation.It entails dividing the customer base of a business into several groups or segments depending on many aspects like demographics, behavior, buying habits, interests, and other similar characteristics. Businesses may learn a lot about the distinctive patterns, tastes, and behaviors within these consumer clusters by grouping consumers with similar traits.Businesses require access to thorough and accurate consumer data in order to conduct client segmentation efficiently. Numerous methods, including online interactions, client feedback, purchase history, surveys, and social media, can be used to get this information. Then, you may use cutting-edge data analytics and machine learning approaches to find significant trends and develop useful consumer groups.

Using customer segmentation, organizations may divide their target market into several groups based on shared characteristics, behaviors, needs, and preferences. The procedure aids companies in developing deeper client insights and effectively modifying their marketing initiatives to cater to the particular requirements of each group. By applying segmentation criteria, businesses can create homogeneous and unique customer groups, allowing for more targeted marketing strategies. Through personalized experiences and focused approaches, customer satisfaction and loyalty can be improved, leading to increased sales, revenue, and a competitive edge in the market. Regular monitoring and refinement ensure that businesses adapt to changing customer preferences and stay ahead in the dynamic business landscape. It is an indispensable tool in modern marketing and customer service strategies. This can help businesses build reliable customer relationships, drive brand loyalty, ultimately achieve sustainability in growth amid the dynamic and competitive market landscape.Customer segmentation can be done on these factors; demographics, Psychographics, Geographic Location, Socioeconomic Status, Buying Frequency, Purchase History. Armed with this valuable insight, businesses can then offer tailored solutions and experiences that resonate with specific customer segments, setting themselves apart from competitors who may take a one-size-fits-all approach. The increase global connectivity and improved supply chains leads to the need for customer segmentation more pronounced. These factors have brought about significant changes in the business landscape, making customer-centric a critical aspect of remaining relevant and successful in today's competitive markets. A customer-centric approach, fueled by data-driven insights from segmentation, enables companies to deliver personalized experiences, stay ahead of market trends, and build enduring customer relationships. Customer segmentation continues to be a key tactic for establishing relevance, success, and long-term growth as organizations adjust to these changing market pressures.

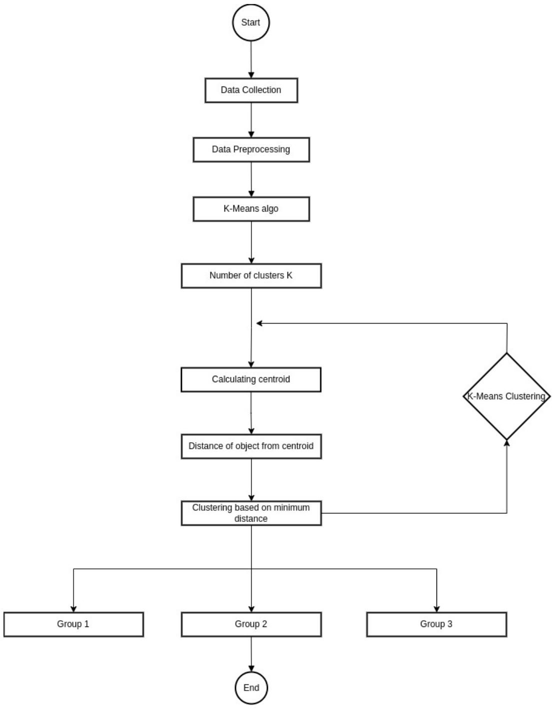

Fig. 1: K-Means Algorithm

There is an explosion of data in the business world today due to the unheard-of volume and diversity of data made available to firms through digital platforms. Text, pictures, videos, transaction records, customer interactions, and more are just a few of the different formats in which this data is available. Effective data management and integration are essential for maximizing the value of this data. Businesses may make data-driven choices and stay competitive in an increasingly digital and data-driven corporate world by gathering, storing, and integrating data from diverse sources. However, it is essential to balance data utilization with data privacy and compliance to build and maintain trust with customers. Leveraging the wealth of available data, customer segmentation has reached new heights as businesses harness demographic, psychographic, and behavioral information to create comprehensive customer profiles. By integrating these diverse data sources, companies gain valuable insights into customer sentiment, preferences, and behaviors. Armed with this deeper understanding, businesses can design highly targeted marketing campaigns that resonate with specific customer segments, delivering personalized recommendations and experiences. The result is enhanced customer satisfaction and loyalty, as customers feel seen and understood by the brand. With data-driven segmentation, businesses can build stronger connections with their clientele, fostering long-term loyalty and sustainable growth in today's competitive market landscape.

There are many algorithms and techniques we can achieve customer segmentation, the most widely used methods is clustering, which includes algorithms like K-means, density-based clustering, and hierarchical clustering. K-means is known for its simplicity and efficiency in dividing customers into distinct clusters based on similarities in their attributes and behaviors. Hierarchical clustering creates hierarchical representation of customer groups, allowing for insights at various levels of granularity. Density-based clustering, identifies clusters based on the density thresholds, accommodating irregularly shaped segments and handling noisy data efficiently. Beyond clustering, other techniques like decision trees, regression analysis, and neural networks are utilized for customer segmentation. Decision trees segment customers based on a series of binary decisions, while regression analysis helps identify significant predictors of customer behavior. Neural networks can capture complex relationships among customer attributes and create accurate segmentation. Each technique has its own strengths and limitations, businesses may combine these methods to gain a comprehensive understanding of their customer's and develop targeted marketing strategies that match to specific customer segments.

By targeting high-value customers and offering personalized solutions, companies can optimize their marketing efforts, reduce acquisition costs, and improve customer retention rates. This, in turn, enhances the company's overall profitability and strengthens its competitive position in the market. Overall, customer segmentation empowers businesses to understand the diverse needs of their customer base, creating differentiation and competitive advantage in a crowded marketplace. By identifying and prioritizing high-value customers, companies can make smarter decisions about resource allocation, ensuring that efforts are focused on nurturing the relationships that matter most to the company's success. Through tailored solutions and personalized experiences, businesses can enhance customer satisfaction and build long-term loyalty, securing their position as market leaders in the face of intense competition.

Businesses may forecast future trends and requests by analyzing previous data and behavior patterns, making it a crucial tool in sectors that are undergoing fast change. Companies may remain ahead of the curve and adjust their strategy proactively to suit changing client wants and market dynamics by utilizing data-driven insights. Analyzing historical data entails looking at previous trends, consumer behavior, and market developments to find patterns and connections. Businesses may obtain important context and insights into the elements that have affected their performance and consumer interactions by knowing what has transpired in the past. The basis for forecasting probable future scenarios is this historical viewpoint. Companies identify new preferences and altering needs by analyzing consumer behaviour, such as purchasing history and interaction with marketing initiatives.

Knowing what will happen next in terms of trends and desires is essential in sectors that are changing quickly. Technological breakthroughs, shifting customer tastes, shifting economic situations, or pressures from competitors can cause market conditions to change swiftly. Businesses may obtain insight into these developments by examining past data and behavior patterns, which enables them to proactively alter their plans and operations. Adapting strategies based on predictive insights helps businesses stay relevant and competitive. For instance, a retail company that notices a growing trend toward online shopping can invest more resources in enhancing its e-commerce platform and digital marketing efforts. Similarly, a technology company that identifies a surge in demand for a particular type of product can focus its research and development efforts to capitalize on the opportunity. Predictive analytics also aids in resource optimization. By anticipating future demands, companies can allocate their resources efficiently, avoiding overproduction of products that may become less popular and ensuring they can meet the needs of a growing market segment.

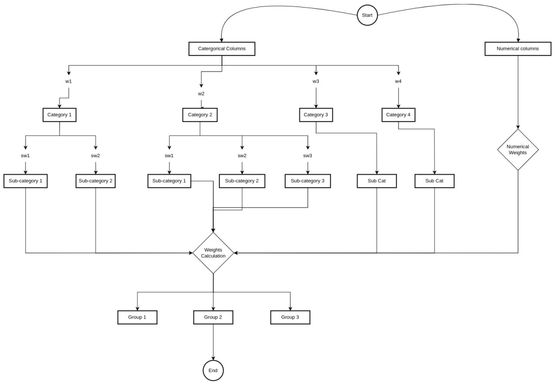

Fig. 2: Hierarchical Level Weighted Method

We begin by thoroughly exploring the dataset through visualizations and summary statistics to gain valuable insights into its underlying patterns and characteristics. Followed by employing two distinct clustering techniques for segmentation purposes. We apply the widely used K-means clustering technique, which partitions the data into distinct clusters based on their similarities. By iteratively optimizing cluster centroids, K-means groups data points into clusters, allowing us to identify different customer segments effectively. Then we provide weighted clustering, a revolutionary methodology. In contrast to conventional clustering, our suggested technique gives weights to certain variables while considering their relative significance in identifying clusters. By using a weighted technique, we may build segments that are more specialized and detailed, which improves the accuracy of our clustering findings. providing a clear and simple comprehension of the consumer segments by visualizing the clusters produced by both methodologies. By making it easier to spot specific patterns and correlations in the data, these visual representations help in further analysis and decision-making. We summarize our results and contrast the K-means and weighted clustering methods' outputs. In order to give readers useful insights into each technique's uses in consumer segmentation, we emphasize its benefits and drawbacks. Our conclusions serve as a basis for data-driven strategies, allowing businesses to adjust their marketing efforts, optimize product offerings, and enhance overall customer experiences.

This study offers significant contributions to the subject of the customer segmentation:

1. The study explores the features of the customer dataset, including demographic information and shopping behaviour. By diving deeper into these characteristics, the study delivers significant insights into customer preferences, behaviours, and trends, establishing the groundwork for better informed retail decision-making.

2. This study introduces a unique experimental hierarchical-level weighted method for grouping. By adding weighted features at multiple hierarchical levels, this approach goes beyond typical clustering algorithms. This approach allows for a more refined and personalized segmentation of customers.

3. To cluster the supermarket customer dataset, the study used the widely utilised K-means technique. The study shows the applicability and efficiency of this well-established approach in custome segmentation by using it. K-means helps identify distinct customer groups based on their shared characteristics, enabling retailers to understand the varying needs and preferences of their clientele.

2. Literature Survey

One significant study [20] used K-means clustering to classify customers of an e-commerce site based on their past purchases and browsing habits. The study was successful in identifying distinct customer groups and making tailored product recommendations, which increased conversion rates and customer retention significantly. After developing personalized product recommendations for each segment based on a better understanding of customer behavior, the research team continued. These personalized recommendations aimed to increase the usefulness and allure of product offerings for particular customers, ultimately resulting in increased client engagement and satisfaction. The results showed a significant increase in conversion rates and client retention, proving the personalized strategy's efficiency in boosting sales and cultivating client retention.

It was suggested in the study [21] to combine the RFM (Recency, Frequency, Monetary) and LTV (Life Time Value) methods to create a novel approach for customer segmentation. In order to analyze the RFM and LTV metrics of the customers, a statistical method was used. Recency, frequency, and monetary value were the three main components of customer behavior that were the focus of the RFM analysis. The LTV analysis sought to determine the overall value of each customer to the company over the course of their entire relationship. The combined RFM and LTV metrics were then subjected to the K-means clustering algorithm. The researchers then used a neural network to improve the segmentation after applying the K-means model. The study aimed to enhance the precision and granularity of the segmentation process by using a neural network. By incorporating a neural network into the segmentation process, the study aimed to improve the accuracy and granularity of the identified customer segments. The study demonstrated the potential for machine learning techniques to optimize customer segmentation further.

Their study [17], presents a novel method for market segmentation in the retailing industry, focusing on a customer's lifestyle as a key factor. The researchers utilized a large transactional database from a European retailing company to extract relevant information about shoppers' purchasing behaviors across various product categories. With the segmentation results in hand, the researchers proposed tailored promotional policies for each customer segment. By customizing marketing strategies based on the specific lifestyle preferences of different segments, the retailing company aimed to foster loyal relationships with its customers and, in turn, boost overall sales.

The researcher investigated the function of product categories in customer segmentation in a related study [19]. Their investigation compared various methods for grouping customer-level sales data. In the end, they used the popular K-means clustering algorithm to divide consumers into groups according to their shopping preferences. Han et al. identified different customer segments using their segmentation approach, including those who bought regular, seasonal, or convenience category items. The researchers were able to identify the various shopping preferences and patterns of various customer groups thanks to this level of granularity. By identifying these customer segments, Han et al. were able to provide valuable insights to the retailing company. With a clear understanding of the distinct customer behaviors and preferences, the company could develop targeted marketing strategies and promotions, catering to each segment's unique needs. This tailored approach aimed to enhance customer satisfaction and loyalty while optimizing sales and revenue.

Clustering

Clustering is a unsupervised machine-learning algorithm in data analysis that involves the grouping of similar data points or objects into distinct clusters. The goal of clustering is to identify inherent patterns, structures, relationships within a data-set, without the need for predefined labels or target variables [18]. By organizing data points into meaningful clusters based on their similarities and dissimilarities, clustering facilitates data exploration, pattern recognition, and the discovery of natural groupings within the data. In clustering, each data point is assigned to a cluster, and points within the same cluster share similar characteristics or properties, while points in different clusters exhibit dissimilar traits. Clustering is driven by mathematical algorithms that try to minimise intra-cluster distance and maximise inter-cluster distance, ensuring that data points within the same cluster are more comparable to each other than those in other clusters.

Clustering finds wide applications in various fields, including customer segmentation, image processing, anomaly detection, and market research. It plays a pivotal role in customer segmentation, where businesses aim to divide their customer base into distinct groups based on shared characteristics, behaviors, or preferences. By leveraging clustering algorithms, businesses can tailor marketing strategies, optimize product offerings, and enhance overall customer experiences based on the unique needs of different customer segments [15]. Clustering is a widely used technique in data analysis and plays a crucial role in customer segmentation. Researchers and practitioners have extensively employed various clustering algorithms to uncover meaningful patterns and groupings within datasets. Clustering methods, such as K-means, density-based clustering, and hierarchical clustering have been applied to diverse domains, including marketing, retail, and e-commerce, to identify distinct customer segments based on shared attributes and behaviors.

K-means, as one of the most popular clustering algorithms, partitions data into clusters by iteratively updating centroids to minimize intra-cluster variance. Its simplicity, efficiency, and ease of implementation make it a go-to choice for many customer segmentation tasks [18]. Hierarchical clustering, on the other hand, creates a hierarchical representation of data, revealing nested clusters at different levels of granularity. This method offers insights into both broad and fine-grained customer groups. Density-based clustering, represented by algorithms like DBSCAN, identifies clusters based on density thresholds, accommodating irregularly shaped clusters and handling noisy data effectively. The flexibility of density-based clustering makes it suitable for various customer segmentation scenarios [16].

Advancements in clustering techniques have also led to hybrid approaches, combining multiple algorithms to capitalize on their strengths and overcome limitations. Genetic algorithms, particle swarm optimization, and neural networks have been integrated with clustering to optimize cluster formation and improve segmentation accuracy.

In customer segmentation the importance of feature selection, data preprocessing, and the evaluation of clustering results using relevant metrics, clustering can significantly impact marketing strategies, personalized recommendations, and customer retention efforts. The insights obtained from clustering analyses aid businesses in developing targeted marketing campaigns, improving customer satisfaction, and fostering long-term customer loyalty.

3. Data

The data-set used in this study comprises historical sales data from a supermarket company recorded over three months in three different branches. The data-set is sourced from Kaggle(https://www.kaggle.com). A collection of data called the supermarket data-set contains details about the purchases made at a supermarket. This data set typically contains details about the transaction's date and time, the items purchased, their prices, the total amount paid for the transaction, and other pertinent information. This supermarket data set is a useful resource for comprehending and forecasting supermarket patrons' behavior as well as for making data-driven decisions that can enhance a supermarket's operational performance. 1000 rows and 17 columns make up the data set, which offers a significant amount of data for analysis. This data set provides plenty of opportunities for identifying hidden trends and patterns that can have a big impact on business decisions thanks to its wide range of variables.

These are the columns:

- Invoice id: unique number
- Branch: 3 branches A, B, and C.
- City: Location
- Customer type: Member and Normal
- Gender: Male and female
- Product line: Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
- Unit price
- Quantity
- Tax
- Total
- Date
- Time
- Payment: Cash, Credit card, and Ewallet
- COGS(Cost of goods sold)
- Gross margin percentage
- Gross income
- Rating: From 1 to 10

4. Exploratory Data Analysis(EDA)

Exploratory Data Analysis (EDA) is a fundamental step in the data analysis process, playing a crucial role in understanding and uncovering insights from raw data. It involves visually exploring, summarizing, and interpreting the dataset to gain valuable insights before applying more advanced modeling or statistical techniques. EDA provides an opportunity to grasp the overall structure of the data, detect patterns, relationships, and potential issues, and validate assumptions. EDA helps become familiar with the dataset's characteristics, including the number of observations, variables, and their respective data types. Understanding the data's scope and limitations allows for informed decision-making throughout the analysis process.[1,2,5]

- Summary Statics: The summary of the dataset provided numerical measures, such as maximum, minimum, quartiles, count, and standard deviation, to grasp its characteristics. These statistics offer insights into data distribution, central tendency, variability, and potential outliers.
- Checking null values: Our data contain no null values
- Dropping columns: During the data preprocessing phase, we performed column removal to enhance the relevance and efficiency of our analysis and clustering process. The dropped columns include Invoice ID, Date, Time, gross margin percentage, Tax 5%, cogs, gross income, Total.

  The invoice ID column did not provide any relevant information about customer behavior or preferences, as it only represented a transactional reference number.

  Date and Time: Our data does not include individual customer identification, these attributes lack relevance for our customer segmentation analysis. Retaining them would not contribute meaningful results to our research objectives.

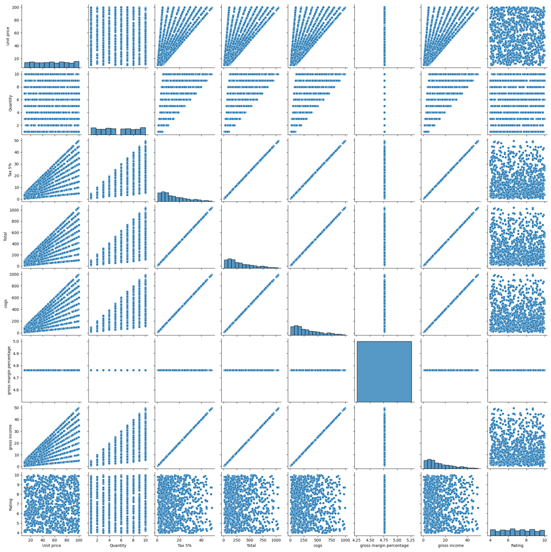

Fig. 3: Scatterplots and Distributions of dataset

  Gross Margin Percentage, Tax 5%, cogs, gross income, and Total are removed are looking at the correlation matrix, we observed a high correlation among these variables. High correlation indicates that these columns share similar information and could potentially lead to multicollinearity issues during the analysis. Multicollinearity can affect the stability and interpretability of our results and may lead to unreliable insights. To avoid redundancy and potential bias in our analysis, we decided to drop these.
- Handling categorical columns: We used both Label Encoder and One-Hot Encoder, for visualization and biased free clustering in our project respectively.

  Label Encoder: It assigns a unique integer value to each category in the column. but this can increase the biases in the column and thus cannot be used in clustering but it is better to use it in visualization.

  One-Hot Encoder: One-Hot Encoder is another approach for dealing with categorical data. Unlike Label Encoder, One-Hot Encoder creates binary columns for each category in the original column. It assigns a value of 1 to the corresponding category and 0 to all other categories.
- Correlation matrix: A correlation matrix is a powerful tool used in data analysis to explore the relationships between multiple variables in a dataset. It provides a comprehensive overview of the pairwise correlations between all pairs of attributes, allowing us to understand how the variables are related to each other. The values in the correlation matrix range from -1 to 1. When three or more variables have a correlation value of 1 in the correlation matrix, it indicates a perfect positive linear relationship between these variables. In the context of our dataset, the variables Total, COGS, Tax 5%, and Gross

Income exhibit a perfect positive correlation with each other.

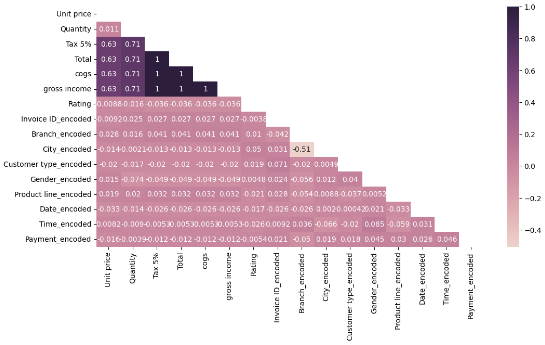

Fig. 4: Correlation Matrix of dataset

5.

5.1 K-means Clustering

K-means clustering is a method that separates observations with dissimilar attributes and groups similar observations together in order to create distinct, non-overlapping clusters. K initial points are chosen as centroids at random or from the dataset itself to start the process. The closest centroid for each data point is then determined by calculating the Euclidean distance between each data point and the discovered centroids. The first clusters are made in this step. After calculating the average of the data points in each cluster and reassigning the data points to the closest centroids, the method updates the centroids repeatedly. Until the centroids stabilize or a predetermined number of iterations have been reached, this method is repeated. The primary purpose of K-means clustering is to minimise the total within-cluster variance, which is the sum of the squared Euclidean distances between the feature values of each observation and its associated centroid. K-means seeks to build clusters with little internal variation by constantly altering the centroids based on the closeness of data points, resulting in clusters that are unique and well-separated from one another.[1,3,23,24,25,27]

The aim of K-means clustering is to minimise total within-cluster variance, which is computed as follows:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

where:

- C_k represents the k-th cluster containing data points x_i.
- μ_k is the centroid of the k-th cluster.
- $(x_i - \mu_k)^2$ denotes the squared Euclidean distance between the data point x_i and the centroid μ_k.

In our research, we use the optimum number of clusters K = 4 based on the Elbow technique described above.
We may further exhibit these groups through plots by fitting the K-means clustering into our data-set.

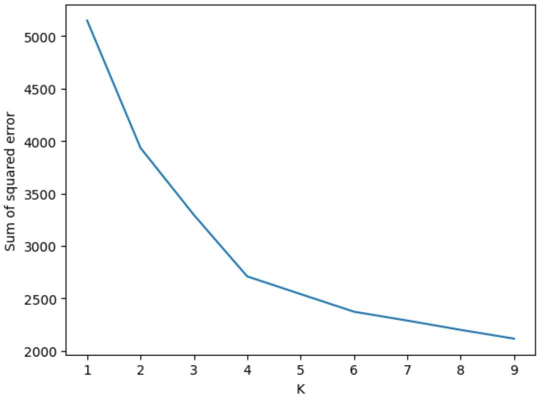

Fig. 5: Elbow Plot

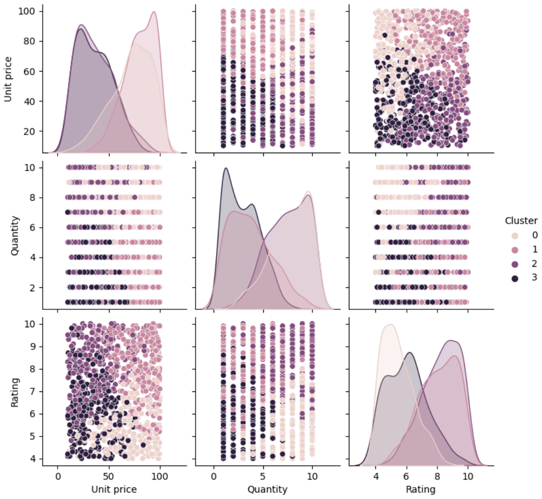

Fig. 6: Clusters resulted by fitting K-means into the data-set

5.2 Validation of K-means Clustering

A. Within-Cluster Sum of Squares (WCSS)

WCSS calculates the sum of squared distances between data points and their corresponding cluster centroids to determine cluster compactness. A lower WCSS indicates better grouping. By charting WCSS versus the number of clusters and looking for the ”elbow” point, which represents the point when adding additional clusters does not significantly diminish the WCSS, the elbow approach may be used to discover the best number of clusters. Within-Cluster Sum of Squares (WCSS) is a key notion in cluster analysis, particularly in the context of k-means clustering, an unsupervised machine learning approach that is frequently employed. WCSS evaluates cluster compactness or tightness, which reflects how near data points are to cluster centroids. The WCSS value serves as a measure of how well the data can be represented by the chosen number of clusters (k). The ideal scenario is to achieve a low WCSS value, indicating that the data points are tightly clustered around their respective centroids, resulting in well-defined and compact clusters.By minimizing the WCSS, k-means aims to

Cluster 1	Cluster 2	Max Dissimilarity	Average Dissimilarity	Isolation
0	1	2.153021	1.571281	0.729803
0	2	2.153021	1.569386	0.728923
0	3	2.153021	1.573823	0.730984
1	2	2.100931	1.554850	0.740077
1	3	2.153021	1.559287	0.724232
2	3	2.153021	1.557392	0.723352

TABLE I: Clusters resulted by fitting K-means into the data-set

create well-defined and tight clusters that effectively capture the underlying structure of the data. The optimal number of clusters can be determined by inspecting the ”Elbow” point on a plot of WCSS values against different values of k.

B. Silhouette Score

The Silhouette Score metric is used to assess the quality of clustering findings. It assesses how well the data has been grouped by measuring the cohesion and separation of data points inside clusters. The Silhouette Score is a value between -1 and +1, with higher values indicating better-defined and well-separated clusters. Two distances are considered when computing the Silhouette Score for a data point: a) the average distance between the data point and all other points within the same cluster (intra-cluster distance), and b) the average distance between the data point and all points in the nearest neighbouring cluster (inter-cluster distance). The Silhouette Score for the data point is then calculated by dividing (b - a) by the maximum of (a, b). It may also be used to identify the ideal number of clusters, as the score is maximum when the data is truly classified into different groups. However, it is critical to remember that the Silhouette Score has limits and should not be used only to assess the quality of clustering findings, particularly when working with high-dimensional or non-linear data.

C. Calinski-Harabasz Index

The Calinski-Harabasz Index, commonly known as the Variance Ratio Criterion, is an unsupervised machine learning assessment statistic used to assess the quality of clustering findings. It aids in selecting the best number of clusters to use for a particular dataset. The index assesses cluster separation as well as cluster compactness inside each cluster, with the goal of achieving a balance that results in well-defined clusters. The Calinski-Harabasz Index is calculated by dividing the variation across clusters (inter-cluster variance) by the variance inside each cluster (intra-cluster variance). The index yields a higher value when clusters are well-separated and more compact, indicating a better clustering solution. Conversely, lower values suggest that either the clusters are too spread out or the data points within the clusters are too dispersed. The formula for the Calinski-Harabasz Index involves the use of the centroids of the clusters, which means it can be sensitive to the shape and size of the clusters. This makes it more effective for evaluating data-sets with relatively spherical clusters. Researchers and practitioners often use it in conjunction with other validation techniques to gain a comprehensive understanding of clustering performance.

D. Davies-Bouldin

The Davies-Bouldin measures the average similarity between each cluster and its most similar cluster while considering the compactness of the clusters. The index aims to find a balance between well-separated and compact clusters, providing a quantitative measure of the clustering performance. To compute the Davies-Bouldin Index, the dissimilarity between each pair of clusters is calculated based on their centroids. The index is then obtained by averaging these dissimilarities across all clusters. A lower Davies-Bouldin Index indicates better clustering performance, with well-separated and compact clusters. Davies-Bouldin Index does not require computing distances between data points, making it computationally less intensive, especially for large datasets. Additionally, the index can be used with any distance metric that defines a meaningful similarity measure between data points. However the Davies-Bouldin Index also has its limitations, it may not perform well with irregularly shaped or non-convex clusters, as it heavily relies on the centroid-based representation of clusters. Consequently, its effectiveness might be reduced when dealing with complex and overlapping cluster structures.

5.3 Hierarchical Level Weighted Method

The ”Hierarchical Level Weighted Method” technique is an approach that offers a comprehensive and nuanced evaluation of categorical and numerical data. The technique involves dividing each unique categorical value into a hierarchical ladder. The method assigns weights to each categorical value based on its percentage representation in the column by breaking each distinct categorical value into a hierarchical ladder. This makes sure that categories that are used more frequently have a bigger impact on the final calculation. Quantity and other non-laddering numerical columns are added together separately. The method uses a step-by-step process of climbing down the ladder to determine the final score for each data entry. At each level, the corresponding weight is applied, and these weighted scores are multiplied together. Each customer’s final score can be calculated after which they can be ranked according to their scores. This technique introduces a thoughtful and data-driven way to cluster information using quartiles, allowing for meaningful insights and decision-making based on the hierarchical levels and weights assigned to different data components. The hierarchical structure and the weight assignments provide a way to prioritize certain categories over others in the final calculation.

The inclusion of numerical columns that are not a part of the hierarchical ladder is an important component of the

S No.	Score	K-Means	HLWM
1.	Silhouette Score	-0.011	-0.051
2.	Davies Bouldin	26.13	3.11
3.	Calinski-Harabasz Index	1.11	416.05

TABLE II: Clustering Results

hierarchical approach that improves the analysis. The method makes sure that all pertinent information is appropriately taken into account in the final scoring process by incorporating these numerical variables. The numerical columns provide useful quantitative data that enhances the categorical columns' qualitative insights. A comprehensive scoring system that takes into account both categorical and numerical factors is produced by adding these numerical values together and multiplying them by the weights given to the categorical columns. Because it captures a wide range of data and relationships between various variables, this integrated approach makes it possible to understand the data set in a more comprehensive way.

Additionally, the choice to cluster data using quartiles provides flexibility and insightful information about the data set. Analysts can cluster data points into meaningful clusters based on their numerical values by dividing the data into quartiles. This method makes it possible to spot specific patterns and trends in the data, which can be very useful for a variety of projects. It becomes easier to discern high or low-performing groups, identify customer segments with distinct preferences, or detect regions with specific characteristics. Additionally, the use of quartiles aids in identifying outlier behaviors, which are data points that significantly deviate from the rest of the data-set. These outliers might represent unique opportunities or potential issues that require special attention in strategic planning and decision-making processes.

The combination of including numerical columns and using quartiles for data clustering in the hierarchical approach empowers analysts to gain deeper insights and make more informed decisions. The approach becomes particularly valuable in strategic planning, marketing campaigns, and resource allocation. Companies can tailor their marketing strategies to specific customer segments based on the identified clusters, thus optimizing the effectiveness of their campaigns. Moreover, resource allocation decisions can be guided by understanding regional differences or outlier behaviors, allowing organizations to allocate their resources more efficiently and capitalize on opportunities. In summary, the incorporation of numerical columns and the use of quartiles enhance the hierarchical approach's analytical capabilities, making it a powerful tool for data-driven decision-making across various domains.

The method effectively gives more weight to categories with a higher prevalence in the data by allocating weights to each category, reflecting their true significance in the final calculation. As a result of taking into account the varying effects of each category on the analysis, this weighting mechanism makes sure that the resulting average or aggregated value is more indicative of the entire dataset. As a result, the hierarchical approach turns into a useful tool for market trend analysis and customer segmentation because it offers a more precise and nuanced understanding of the underlying characteristics of the data. The insightful information it provides into data patterns and trends is one of the main benefits of using quartiles for data clustering. It is possible to locate distinct groups or clusters of data by dividing the dataset into quartiles. These clusters may represent various market or customer segments with distinctive traits and preferences. Detecting outlier behaviors, or data points that significantly deviate from the norm, is another benefit of using quartiles. Outliers can be a sign of exceptional or unusual events that demand consideration during decision-making processes. Making knowledgeable and sensible decisions in a variety of areas, including marketing tactics, product development, and resource allocation, depends on this information. Overall, the hierarchical approach's capacity to offer insightful information and support data-driven decision-making is improved by the combination of giving categories weights and using quartiles for data clustering.

The method's reliance on percentage representation within the dataset can be problematic, particularly in situations where the dataset is imbalanced, meaning some categories have significantly more or fewer data points than others. In imbalanced datasets, categories with higher percentages may be overemphasized during the weighting process, leading to a biased analysis. As a result, the conclusions drawn from the study may be skewed and not accurately reflect the true trends or relationships in the data. Assigning weights might often require subjective judgment, especially in cases where there is a lack of domain knowledge or clear guidelines for assigning weights. Subjective decisions made during this process can introduce inconsistencies in the analysis, potentially leading to less reliable results. Different analysts may assign weights differently, leading to variations in the conclusions.

The hierarchical approach is a powerful data analysis technique with significant benefits for customer segmentation and market trend analysis, particularly in large datasets rich in diverse factors such as demographics, preferences, spending habits, and purchase quantities. When applied appropriately, it can provide nuanced evaluations and valuable insights. However, caution must be exercised when dealing with very small datasets that lack categorical diversity and have few data points, as the hierarchical approach may add unnecessary complexity and fail to yield meaningful results. Moreover, extreme outliers within the data can distort the weighting process, leading to biased outcomes. Therefore, analysts should carefully consider the dataset's characteristics and research objectives before employing the hierarchical approach, as it may not be universally suitable for all contexts. Flexibility in selecting alternative approaches when necessary is essential to ensure sound and reliable data analysis.

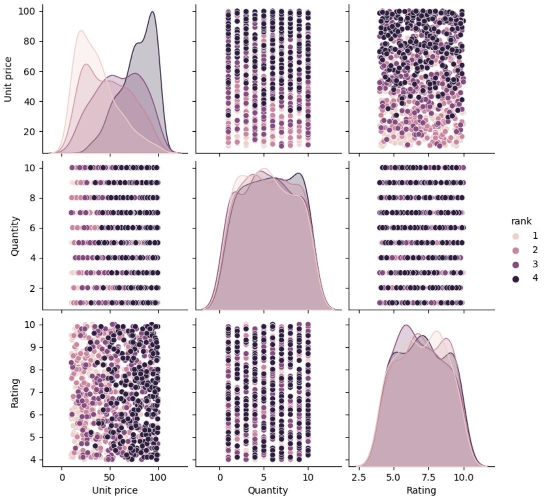

Fig. 7: Clusters resulted by fitting the Hierarchical Level Weighted Method dataset

6. Conclusion

In this study, we explored the domain of clustering analysis and its uses in market segmentation and customer segmentation. Finding loyal consumers and creating marketing plans that were suited to their requirements were the objectives. Our data allowed us to divide the customers into four different groups. Our goal was to develop a deeper understanding of consumer behavior, preferences, and value to help firms better target their marketing efforts and increase client happiness. Two clustering techniques, K-means and Hierarchical Weighted Level Methods, were used in our study.

Through appropriate data cleaning and scaling, we addressed the dataset and the nuances of customer segmentation. We ensured the dataset's quality and made it ready for clustering analysis by dealing with missing values, removing irrelevant columns, and standardizing the numerical characteristics. To apply clustering methods, it was also necessary to encode categorical columns while converting qualitative data to a numerical format. For firms to make data-driven decisions and develop specialized marketing strategies for particular client groups, these preprocessing stages were crucial in creating relevant and precise consumer segments.

We used K-means clustering to segment our customer dataset into distinct groups based on their similarities. Using the K-means algorithm, we selected K initial centroids and iteratively assigned data points to the nearest centroid, forming clusters with minimal within-cluster variation. We were able to categorize customers based on shared traits, tastes, and behaviors, which gave us important information for developing focused marketing campaigns.

To better comprehend the customer categories, we visualized the findings of the clustering study. We produced several visualization graphs using the Matplotlib and seaborn libraries. We were able to spot potential patterns and trends across different customer categories due to these visualizations, which also made it easier to interpret the clustering results.

Furthermore, we evaluated the efficiency of the K-means clustering algorithm by generating and displaying clustering errors for various K values. The ideal number of clusters was determined using the Elbow method. The ”elbow point,” or the point at which including more clusters has little effect on the sum of squares within a cluster, was discovered using the Elbow approach. Based on the errors and validation metrics, we decided on the ideal number of clusters. The ideal number of clusters was determined using the Elbow method. The ”elbow point,” or the distance at which more clusters have no appreciable impact on the total number of squares within a cluster, was discovered using the Elbow technique.

We employed a wide range of quantitative metrics, such as Max Dissimilarity, Average Dissimilarity, and Isolation, to assess the effectiveness and caliber of the K-means clustering results. The segmentation and homogeneity of the customer segments were made possible by these metrics. By calculating the dispersion of data points inside each cluster using Max Dissimilarity and Average Dissimilarity, we were able to better understand the intra-cluster similarity and compactness. We were able to evaluate the distinctiveness and separation across clusters using the Isolation measure in order to further ensure the identification of distinct and non-overlapping consumer categories.

To better segment customers, we developed and introduced the Hierarchical Level Weighted Method. This innovation made a significant contribution to the field of market analysis. We aimed to improve the precision and applicability of the clustering procedure by combining the benefits of hierarchical clustering with the inclusion of weighted features. With our method, we were able to identify more detailed and insightful consumer groupings by giving varying weights to elements based on their importance in describing customer behavior. The use of this methodology enabled the development of focused marketing strategies by giving useful insights into the distinctive traits and preferences of various consumer groups.

Businesses can obtain profound insights into customer behavior and preferences by using the right tools and carefully understanding the data, which will result in more effective marketing campaigns and higher levels of customer satisfaction. Future developments in complex and precise customer segmentation strategies are made possible by the continued investigation of enhanced clustering techniques. It was also found that numerous different clustering algorithms should be used in addition to investigating all accessible data insights. Different clustering types are probably best fitted to take use of different aspects of the data. K-means clustering appeared to show the greatest fit for the data in this project's specific case. Our results demonstrated the potency of client segmentation based on data. By grouping customers with similar behaviors and preferences, businesses can deliver personalized experiences, targeted promotions, and tailored recommendations.

This leads to increased customer loyalty, higher sales, and a competitive edge in the market.

References

[1] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning: With Applications in R. Springer.

[2] Alboukadel Kassambara. 2017. R Graphics Essentials for Great Data Visualization. Sthda.com.

[3] Cory Lesmeister. 2017. Mastering Machine Learning with R: Advanced prediction, algorithms, and learning methods with R 3.x. Packt Publishing Ltd.

[4] Abdulhafedh, A. (2017). A Novel Hybrid Method for Measuring the Spatial Autocorrelation of Vehicular Crashes: Combining Moran's Index and Getis-Ord Gi Statistic. Open Journal of Civil Eng , 7, 208-221.

[5] Aurélien Géron. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly.

[6] 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), 2018, https://doi.org/10.1109/icebe.2018.00027.

[7] Alloghani, Mohamed, et al. "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science." Unsupervised and Semi-Supervised Learning, 2019, pp. 3–21, https://doi.org/10.1007/978-3-030-22475-2_1.

[8] Han, Shui Hua, et al. "Segmentation of Telecom Customers Based on Customer Value by Decision Tree Model." Expert Systems with Applications, 2012, https://doi.org/10.1016/j.eswa.2011.09.034.

[9] "Data Mining Techniques for Segmentation." Data Mining Techniques in CRM, pp. 65–132, https://doi.org/10.1002/9780470685815.ch3.

[10] Bayer, Judy. "Customer Segmentation in the Telecommunications Industry." Journal of Database Marketing Customer Strategy Management, 2010, https://doi.org/10.1057/dbm.2010.21.

[11] Kanchanapoom, Kessara, and Jongsawas Chongwatpol. "Integrated Customer Lifetime Value (CLV) and Customer Migration Model to Improve Customer Segmentation." Journal of Marketing Analytics, 2022, https://doi.org/10.1057/s41270-022-00158-7.

[12] Cheng, Guang Hua. "A Collaborative Filtering Recommendation Algorithm Based on User Clustering in E-Commerce Personalized Systems." Advanced Materials Research, 2011, https://doi.org/10.4028/www.scientific.net/amr.267.789.

[13] Chang, H.H. and Tsay, S.F. (2004), "Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation", JOURNAL OF INFORMATION MANAGEMENT.

[14] Caliński, T. and Harabasz, J. (1974), "A dendrite method for cluster analysis", Communications in Statistics-Theory and Methods, Taylor Francis, Vol. 3 No. 1, pp. 1–27.

[15] Zeying Li. "Research on Customer Segmentation in Retailing Based on Clustering Model." 2011 International Conference on Computer Science and Service System (CSSS), 2011, https://doi.org/10.1109/csss.2011.5974496.

[16] Sharma, Arvind, et al. "Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data." Mathematical Problems in Engineering, vol. 2016, 2016, pp. 1–9, https://doi.org/10.1155/2016/1564516.

[17] Siagian, Romadansyah, et al. "The Implementation of K-Means Dan K-Medoids Algorithm for Customer Segmentation on e-Commerce Data Transactions." SISTEMASI, vol. 11, no. 2, 2022, p. 260, https://doi.org/10.32520/stmsi.v11i2.1337.

[18] Chander, Satish, and P. Vijaya. "Unsupervised Learning Methods for Data Clustering." Artificial Intelligence in Data Mining, 2021, pp. 41–64, https://doi.org/10.1016/b978-0-12-820601-0.00002-1.

[19] Han, Shuihua, et al. "Category Role Aided Market Segmentation Approach to Convenience Store Chain Category Management." Decision Support Systems, vol. 57, 2014, pp. 296–308, https://doi.org/10.1016/j.dss.2013.09.017.

[20] Mohanty, Swaroop, et al. "Study of Factors That Influence Retailers in Product Assortment as per the Customers Preference of Products, Leading to Improved Retailer Performance for Customer Satisfaction and Retention." Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 11-SPECIAL ISSUE, 2019, pp. 57–64, https://doi.org/10.5373/jardcs/v11sp11/20192929.

[21] Namvar, Morteza, et al. "An Approach to Optimised Customer Segmentation and Profiling Using RFM, LTV, and Demographic Features." International Journal of Electronic Customer Relationship Management, vol. 5, no. 3/4, 2011, p. 220, https://doi.org/10.1504/ijecrm.2011.044688.

[22] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, T.-S. Chua, Disease inference from health-related questions via sparse deep learning, IEEE Transactions on knowledge and Data Engineering 27 (8) (2015) 2107–2119.

[23] Michael Colins. 2017. Machine Learning: An Introduction to Supervised and Unsupervised Learning Algorithms.

[24] Chirag Shah. 2020. A Hands-On Introduction to Data Science. Cambridge University Press.

[25] Sunil Kumar Chinnamgari. 2019. R Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5. Packt Publishing Ltd.

[26] A. Abdallah, A. Berendeyev, I.Nuradin, D. Nurseitov, Tncr: Table net detection and classification dataset, Neurocomputing 473 (2022) 79–97.

[27] Kevin Jolly. 2018. Machine Learning with scikit-learn Quick Start Guide: Classification, regression, and clustering techniques in Python. Packt Publishing Ltd