Revealing Consumer Preferences in the Fashion Industry Using K-Means Clustering

Feri Sulianta
Department of Informatics, Universitas Widyatama, Bandung, Indonesia

Khaerani Ulfah
Department of Informatics, Universitas Widyatama, Bandung, Indonesia

Endang Amalia
Department of Information System, Universitas Widyatama, Bandung, Indonesia

Abstract: The fashion industry, driven by rapidly shifting e-commerce trends and consumer preferences, demands precise data analysis to optimize marketing strategies and enhance customer satisfaction. This study utilizes data mining techniques, specifically K-Means Clustering and the Elbow Method, to reveal consumer preferences within a dataset of 1,000 fashion product sales records, which include attributes such as product ID, name, brand, category, price, rating, color, and size. By grouping data into distinct clusters based on price and rating preferences, the analysis uncovers four key consumer segments. The optimal number of clusters is confirmed using the WCSS (Within-Cluster Sum of Square) method. These insights offer valuable guidance for refining marketing strategies in the fashion industry. Future research should consider additional variables and employ advanced tools for deeper analysis.

Keywords: Consumer Preferences, Data Mining, Elbow Method, Fashion Products, K-Means.

Introduction

With the rapid growth of e-commerce and changing consumer preferences, fashion companies face significant challenges in understanding and adapting to evolving purchasing patterns. Analyzing these patterns is essential for optimizing marketing strategies, enhancing customer satisfaction, and maintaining a competitive edge in an increasingly tight market. To survive and thrive in this competitive environment, companies must leverage opportunities in

Correspondents Author:
Feri Sulianta, Department of Informatics, Universitas Widyatama, Bandung, Indonesia
Email : feri.sulianta@widyatama.ac.id

Received July 24, 2024; Revised August 12, 2024; Accepted Auguts 13, 2024; Published August 15, 2024

technology and information systems while addressing key business needs such as increasing product capacity, reducing operational costs, and expanding profitability.

Data mining, the process of discovering patterns and correlations within large datasets, is a critical tool for businesses. By utilizing techniques from machine learning, statistics, and database systems, data mining transforms raw data into actionable insights. This process involves steps such as data collection, cleaning, analysis, and interpretation, ultimately enabling businesses to make informed decisions and optimize operations (Romero, C., & Ventura, S., 2020; E Okewu et al., 2021)

To structure the data mining process, this study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, developed in 1996 by analysts from various industries, including Daimler Chrysler, SPSS, and NCR (Suhanda et al., 2020). CRISP-DM is a widely adopted methodology that consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This structured approach ensures a systematic and thorough analysis, with each phase producing specific outputs that guide the next steps in the project (Fadillah, 2015 ; H. Nagashima et al.,2021; S. Huber et al,.2019).

Recent studies, have demonstrated the flexibility and effectiveness of the CRISP-DM methodology across various domains, including credit risk prediction. By following the CRISP-DM framework, researchers were able to develop robust machine learning models, showcasing the methodology's adaptability. Another study applied CRISP-DM in healthcare, predicting liver failure cases and improving patient outcomes, further highlighting its applicability in different sectors (K. Rawat., 2023 ; Cunha et al.,2021; Suhanda et al., 2020; Wahyudi et al., 2023; Eans, 2023; N. Azadeh-Fard et al., 2019).

Among the various data mining techniques, clustering is fundamental for grouping similar data points. K-Means Clustering, in particular, is widely used due to its simplicity and efficiency. K-Means is a non-hierarchical clustering method that divides data into one or more groups, placing data with similar characteristics in the same group and those with different characteristics in separate groups (Kristianto & Rudianto, 2020). The algorithm partitions data into K clusters, with each data point assigned to the nearest cluster mean (Chen, J. et al., 2020). Studies have shown the effectiveness of K-Means in applications such as customer segmentation, anomaly detection, and image compression. Its ability to handle large datasets efficiently makes it a valuable tool for these tasks. Additionally, combining K-Means with techniques like Principal Component Analysis (PCA) can enhance its performance by reducing dimensionality and noise (Nandapala, E. Y. et al.,2020; Huang, Yong,2020).

The Elbow Method is a heuristic used to determine the optimal number of clusters in K-Means clustering. By plotting the sum of squared distances from each point to its assigned cluster center (inertia) against the number of clusters, the "elbow" point indicates the optimal number, balancing between minimizing inertia and avoiding overfitting. Recent studies have utilized the Elbow Method in optimizing clustering outcomes, such as in customer segmentation, leading to more accurate insights for targeted marketing strategies.

In comparison to other clustering techniques, K-Means stands out for its computational efficiency and ease of implementation, making it particularly suitable for large datasets typical in the fashion industry. While alternatives like hierarchical clustering offer more detailed insights into data structure, K-Means is preferred for its speed and simplicity, especially when combined with the Elbow Method to refine clustering outcomes (Li, Yue, et al., 2023; Maori, N. A., et al., 2023; Nainggolan, R., et al., 2019).

This study aims to analyze the purchasing patterns of fashion products using the K-Means Clustering method. By leveraging fashion product data, this research provides valuable insights into customer shopping behavior, identifies different customer segments, and offers guidance for designing more effective marketing strategies to enhance customer satisfaction and drive business growth.

Research Method

The research uses a data-driven approach to analyze consumer behavior in the fashion industry, aiming to help fashion companies improve their marketing strategies and customer satisfaction. By applying data mining techniques, especially K-Means Clustering, the study seeks to identify distinct patterns in how consumers make purchasing decisions based on pricing and ratings.

The study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, ensuring a structured approach to the research. In the Business Understanding phase, the objective is to understand consumer behavior in the fashion industry, particularly in e-commerce, and how price and ratings influence purchasing decisions. This understanding is crucial for helping fashion companies remain competitive in a rapidly evolving market.

In the Data Understanding phase, the research focuses on gathering consumer behavior data, likely from e-commerce platforms. This data may include customer feedback, transaction records (such as purchases, product ratings, and pricing information), and data collected via web scraping from online fashion retail websites. During the Data Preparation phase, the collected data is cleaned and organized to ensure accuracy and relevance for analysis. This step sets the stage for the next phase, where the data is ready for modeling.

The Modeling phase employs K-Means Clustering to group consumers into segments based on similar price and rating preferences. The Elbow Method is used here to determine the optimal number of clusters, ensuring the model accurately represents consumer segments. The Within-Cluster Sum of Squares (WCSS) is evaluated to assess the compactness of the clusters and confirm that the chosen number of clusters effectively captures the variability within the data. In the Evaluation phase, the research validates the clustering results by examining how well the clusters align with the research objectives. This step ensures that the insights generated are meaningful and actionable for fashion companies. Finally, in the Deployment phase, the insights from the analysis are used to refine marketing strategies, helping fashion companies tailor their offerings to meet consumer preferences and enhance customer satisfaction.

Overall, this research highlights the importance of using structured data mining processes like CRISP-DM to gain a deeper understanding of consumer behavior. It provides fashion companies with valuable insights that can guide their marketing strategies in a competitive market. The research design used in the implementation of the study is as shown in Figure 1 below.

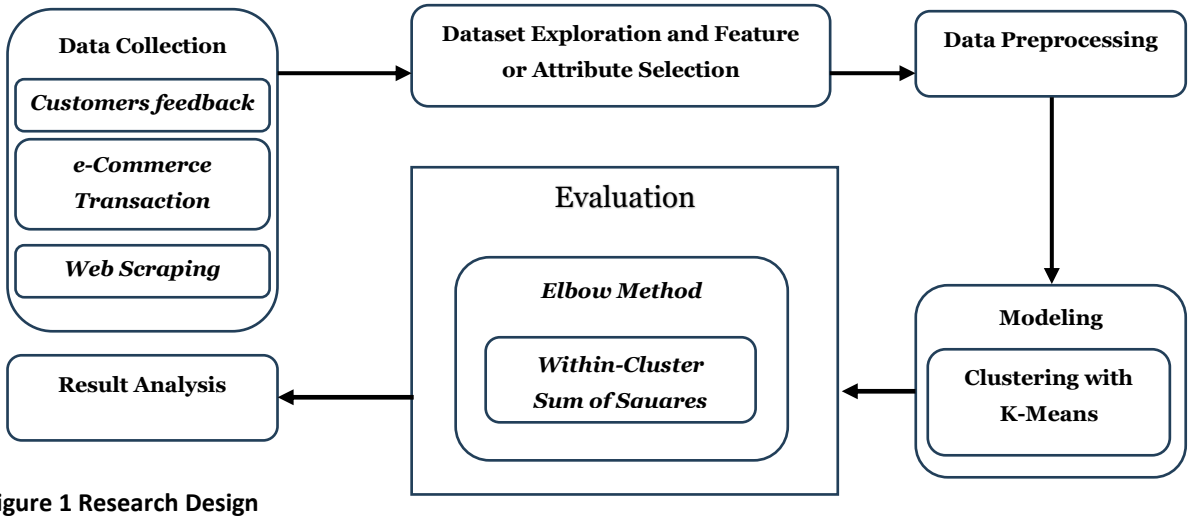

Figure 1 Research Design

Data Mining

Many data mining techniques are known to date, but in this research, the data mining method used is the CRISP-DM method, which is suitable for a business approach, with its steps as follows:

1. Business Understanding Phase: This phase aims to understand customer segmentation based on the quality of product sales achieved by the company, with the goal of improving sales quality within the company.

2. Data Understanding Phase: In this phase, based on the data obtained, understanding data needs is crucial for achieving the goal of determining effective and efficient purchasing strategy patterns (Wahyudi et al., 2023).
    a. Data Collection: In this stage, the process of collecting the necessary data to support the data understanding phase takes place. The initial data source to be used and processed in this research is a fashion product dataset. Basically, the data is already in the form of comma-separated values taken directly from a well-managed transaction system, so there are no missing values and data normalization is not required.
    b. Data Description: In this stage, the data obtained by the researcher is sales data for fashion products that includes product ID, product name, brand, product category, price, rating, and attributes for color and size which is stored in Comma-Separated Values file format. It contains a total of 1,000 records with 9 attributes.1000 entries with 9 attributes.
    c. Data Selection Evaluation: In this stage, the process of summarizing the data takes place, reducing the initial 9 attributes to only 2 attributes.
    d. Attribute Selection: The determination of attributes is used for the data processing phase because it is adjusted to the research focus, so only 2 attributes are used, namely price and rating.
3. Data Preparation: In this phase, data selection and processing are performed to check each data entry, ensuring no problematic data remains after cleaning.
4. Modelling: In this phase, data mining techniques are selected, and the algorithm to be used is determined. The data modelling used in distance calculation employs Python. The steps include determining the number of clusters, the K points or centroids, and calculating the variance.
5. Evaluation: In this evaluation phase, the elbow WCSS method is used to determine the optimal number of clusters. Subsequently, the evaluation results can also determine whether the process can proceed or needs to be repeated because it does not align with the research plan.
6. Deployment: The final report on the information obtained from the previous processes is created in the deployment phase. This is the stage where the processed and tested data is visualized to make the information and knowledge more easily understood (Eans, 2023).

The steps in the study to be conducted can be seen in Figure 2.

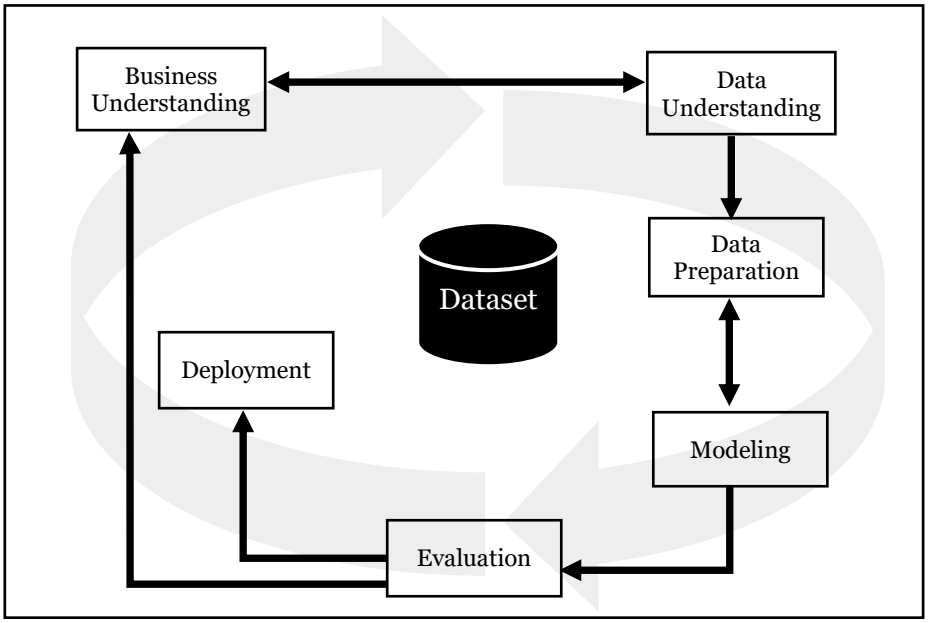

Figure 2 The CRISP-DM phases and the recursive process in the Data Mining stage

The CRISP-DM phases in Figure 1 shows primarily involving the Modeling, Evaluation, and Data Preparation phases, is inherently recursive. The recursive process means that these steps are often revisited multiple times to refine and improve the models. After creating initial models, they are evaluated for performance. If the models do not meet the required standards, adjustments are made either in the modeling techniques or the data used. This iterative loop continues until satisfactory models are achieved. Insights gained during modeling and evaluation may reveal issues with the data. This may necessitate returning to the Data Preparation phase to clean data further, add new attributes, or integrate additional data sources. The refined data is then used to build new models. This iterative nature ensures continuous improvement and refinement, leading to more accurate and effective data mining results.

K-Means Clustering & The Elbow Method

In this research, data mining is used to uncover clustered data. Clustering is the process of forming groups of data extracted from a dataset whose classes are unknown and determining whether the data belongs to those classes. The potential use of clustering lies in its ability to discover structures within data and its applicability in various applications such as pattern recognition, image processing, and classification.

The K-Means algorithm is a non-hierarchical method that initially can take many data components to form the initial center of a cluster. K-means has the capability to group large amounts of data with relatively fast and efficient processing times. However, there is also a weakness in K-means regarding the determination of the initial cluster centers. The results

produced by the K-Means algorithm are highly dependent on the choice of initial cluster means. The steps to run the K-Means algorithm are explained as follows (Dharma Putra et al., 2021) described as below: (1) Determine K as the number of clusters to be formed. (2) Determine the initial k cluster centers. This is done randomly. The initial centroids are determined randomly from the available k cluster data objects. (3) Calculate the distance of each data object to each centroid of the existing clusters using the Euclidean distance calculation method. (4) Assign each data object to a cluster by measuring its proximity to the cluster center. (5) Use the equation repeatedly to determine the location of the new center of mass. The center of the new cluster is the average of all data objects in a particular cluster. (6) After repeating the calculation process, if the data objects still change and the cluster centers do not change, the K-means clustering process can be considered complete.

The method used in this study is the K-Means Clustering algorithm and the elbow method as an optimization method. The elbow method is used to determine the number of k in the clustering process. These stages are iterative, meaning the process is cyclical and may require going back to previous stages as needed. The result of K-Mean Clustering Algorithm is the optimal number of clusters in the clustering process. The application of methods, algorithms, research procedures, and the output of the research can be seen in the following block diagram:

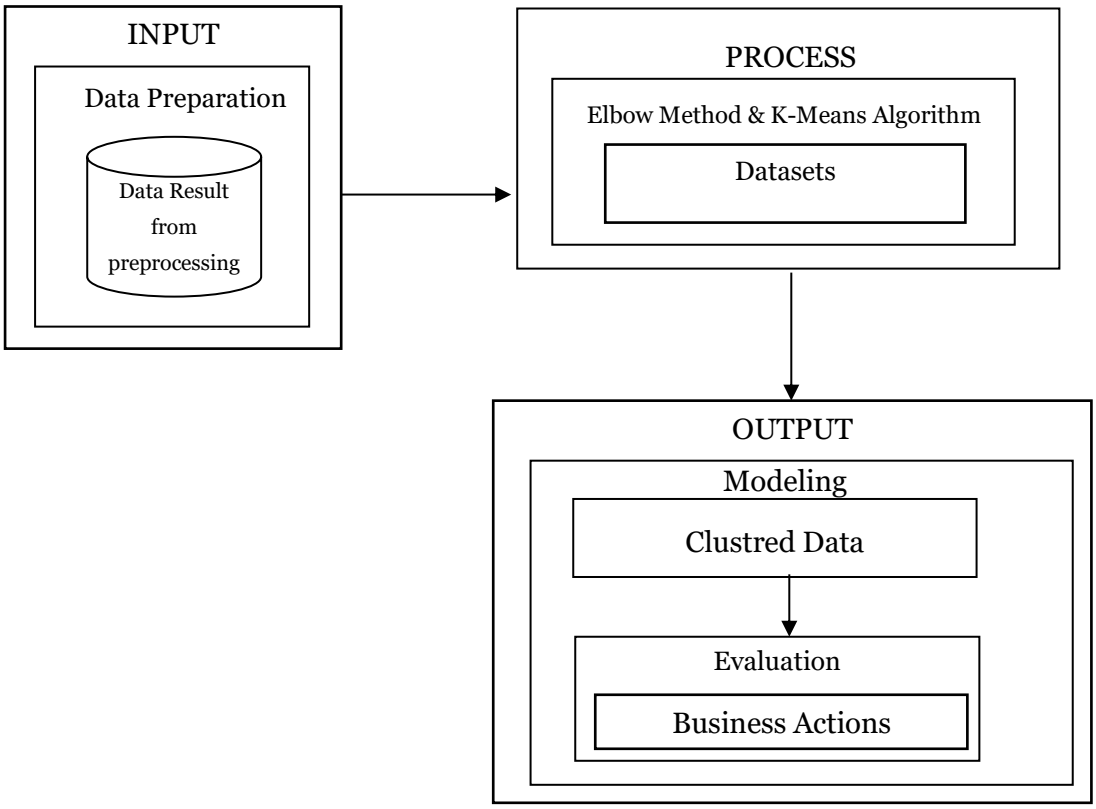

Figure 3 Block Diagram of Research Procedure involving Input, Process and Output Mechanism

The block diagram above illustrates the research flow with input, process, and output steps, which are detailed as follows: Input diagram as the initial stage of the procedure involves data preparation, which is crucial for ensuring that the data is clean, relevant, and ready for analysis. This stage encompasses several key activities, including data collection from various sources, data cleaning to handle missing values, remove duplicates, and correct errors, and data transformation to normalize or standardize data and convert it into a suitable format. Additionally, feature selection is performed to identify and select the most relevant features for the analysis. The outcome of this stage is a Data Result that has undergone preprocessing and is ready for further analysis. Process diagram is the next stage after input stage is well managed, which the prepared data is analyzed using clustering techniques. Initially, the Elbow Method is employed to determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow point," where the rate of decrease sharply slows down. Subsequently, the K-Means Algorithm is applied using the optimal number of clusters identified by the Elbow Method. The K-Means algorithm partitions the data into clusters, assigning each data point to the nearest cluster center and recalculating the centers iteratively until convergence. The result of this stage is the Clustered Data, where the dataset is divided into distinct groups based on similarities.

The output diagram is the final stage which involves Modeling and Evaluation process. This stage includes modeling to form patterns from the clustered data to understand the underlying structure and characteristics of each cluster. Evaluation is then conducted to assess the quality and validity of the clusters using by calculating the total inertia value, this is often referred to as the 'within-cluster sum of squares' or WSS for various numbers of clusters. Inertia measures how close data points are to their cluster centroids. As the number of clusters increases, WSS typically decreases because data points are grouped closer to the centroid. However, after a certain point (called the 'elbow'), the decrease in WSS begins to slow down. This point indicates the optimal number of clusters. Based on the insights gained from the clustered data, actionable business strategies are developed in the form of business actions. These actions could include targeted marketing, customer segmentation, or resource optimization. The output of this stage is a comprehensive understanding of the data patterns, leading to informed Business Actions that can drive strategic decisions. In this reseach the elbow method used does not necessarily have to involve the Silhouette Index because the Silhouette Index measures the density and separation of clusters but does not always align with the observed decrease in inertia seen in the Elbow Method. Some situations may yield good Silhouette values even though WSS continues to decrease, so it is not always necessary to consider both metrics simultaneously (Gat-Nasr et al., 2020).

Result and Discussion

Some aspects regarding the research data to be mined involve Data Collection, where: The data used in this study is fashion product data with various brands, user ages, sizes, types, and ratings. The dataset contains one thousand entries. The dataset on fashion products contains 9 columns. The sample dataset and the types of data can be seen in Table 1 and Table 2. As for Data Pre-Processing, the collected data will undergo initial data processing to produce more accurate data. Data cleaning is a crucial process in data analysis aimed at identifying, correcting, and removing inconsistencies or inaccuracies in the dataset. This process is essential because dirty or unstructured data can lead to inaccurate or unreliable analysis results.

The composition of attributes in the dataset includes: User ID, Product ID, Product Name, Brand, Category, Price, Rating, Color, and Size as seen in table 1. The data undergoes preprocessing to ensure that the fashion product dataset has no missing values, thereby ensuring the integrity and quality of the data so that the analysis or model produced is more accurate and reliable. Attribute selection involves choosing the attributes used for clustering in the analysis of fashion product purchasing patterns, which helps produce a more effective, efficient, and interpretable model.

Table 1 Sample Datasets of Fashion Product Purchases

User ID	Product ID	Product Name	Brand	Category	Price	Rating	Color	Size
19	1	Dress	Adidas	Men's Fashion	40	1.043159	Black	XL
97	2	Shoes	H&M	Women's Fashion	82	4.026416	Black	L
25	3	Dress	Adidas	Women's Fashion	44	3.337938	Yellow	XL
57	4	Shoes	Zara	Men's Fashion	23	1.049523	White	S
79	5	T-Shirt	Adidas	Men's Fashion	79	4.302773	Black	M
....	....	....	....	....	....	....	....	....

This sample dataset, in table 1 contains information about user interactions with various products. Each row represents a unique interaction between a user and a product, capturing details about the product and the user's interaction with it. This dataset captures a range of product types, brands, user ratings, and other attributes, providing valuable information for analyzing user preferences and product performance.

Table 2 Data Types in Datasets

Name Coloum	Type
User ID	Int
Product ID	Int

Product Name	Object
Brand	Object
Category	Object
Price	Int
Rating	Float
Color	Object
Size	Object

The table 2 appears to be a dataset related to products and user interactions with these products. Here is a brief description of each column:

- User ID (Int): This column contains unique integer identifiers for each user. It helps in tracking interactions specific to individual users.
- Product ID (Int): This column holds unique integer identifiers for each product. It is used to distinguish between different products in the dataset.
- Product Name (Object): This column contains the names of the products. The data type "Object" typically refers to string/text data.
- Brand (Object): This column specifies the brand associated with each product. Like Product Name, this is also string/text data.
- Category (Object): This column denotes the category to which each product belongs. This helps in classifying products into different types or groups.
- Price (Int): This column records the price of each product as an integer. It represents the cost value of the products.
- Rating (Float): This column contains the ratings of the products. The float data type indicates that these ratings can be fractional values, providing more granularity.
- Color (Object): This column indicates the color of each product. This is also string/text data.

This table is designed to capture detailed information about products, including their identifiers, names, brands, categories, prices, ratings, and colors, along with the users interacting with these products. This research utilizes Python as the primary tool, using the libraries scikit-learn, NumPy, and pandas for implementing K-Means and data analysis. Python was chosen due to its high flexibility and the availability of extensive libraries, which facilitate the data preprocessing, model development, performance evaluation, and visualization of analysis results.

In this study, the method used is clustering using the k-means clustering algorithm and the optimization method using the elbow method. The elbow method is used to determine the best number of k-clusters in the clustering process through the calculation of WCSS (Within-Cluster Sum of Squares). In the testing process using the k-means algorithm, the tests were

conducted 9 times, with k values ranging from 2 to 10. The evaluation result of the clusters using WCSS from the elbow method based on the number of k tests indicates that the clusters used are 4 clusters.

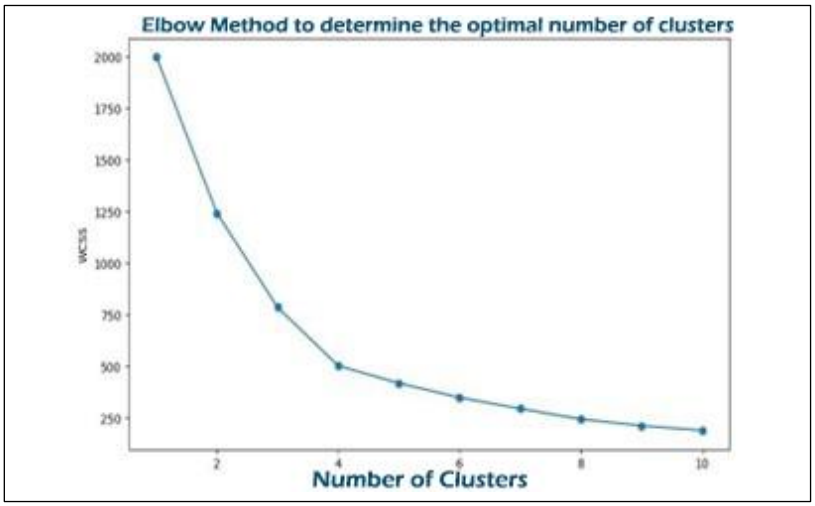

Figure 4: Method to determine the optimal number of clusters

In this study, Fashion Product Data is used. To identify purchase patterns of fashion products, the K-Means Clustering method is applied using Python to assist companies of each brand in planning future marketing strategies. Here is the data analysis using Python to determine fashion product purchase patterns based on clusters.

Data is displayed in the Python application using the Jupyter Notebook text editor, showing initial samples from the dataset to understand the basic characteristics and structure of the data before conducting further analysis or modeling. Displaying the column types in the dataset aims to understand the data structure by identifying the types of data used, such as numerical, categorical, text, and others. This helps in selecting the appropriate analysis methods and algorithms, as well as in detecting and correcting data errors or inconsistencies. Here is a summary of descriptive statistics data for the column that will be used to provide a brief yet comprehensive overview of the main characteristics of the dataset.

Tabel 3 Data Description by using descriptive statistics

Statistic Descriptive	Price	Rating
Count	1000.00000	1000.000000
Mean	55.785000	2.993135
Std	26.291748	1.153185
Min	10.000000	1.000967
25%	33.000000	1.992786
50%	57.000000	2.984003
75%	78.250000	3.985084
Max	100.000000	4.987964

Table 3 provides a summary of the descriptive statistics for the dataset, specifically focusing on the Price and Rating columns. This descriptive statistics summary provides an overview of the central tendency, dispersion, and range of the price and rating data within the dataset. Based on the perspective of price and rating, the average for the price and rating columns is revealed to understand the different purchasing patterns among consumer groups based on price and rating in each cluster.

Tabel 4 The clustered data, measured based on several descriptive statistical calculations, refers to price and rating.

Cluster	Price	Rating
0	31.129464	1.891684
1	77.391473	4.055514
2	77.270370	2.073973
3	32.185484	3,883478

Table 4 summarizes clustered data based on descriptive statistical calculations of price and rating. Each cluster represents a group of products with similar price and rating characteristics. Another aspect aimed at exploring cluster results is measuring the members of each cluster to determine the number of observations in each cluster, which helps assess the size of the consumer groups within each cluster.

Table 5 The number of data for each formed cluster

Cluster	Total
0	224
1	258
2	270
3	248

Table 5 provides the number of data points in each formed cluster from a clustering analysis. Each cluster represents a grouping of products with similar characteristics, as identified in the previous clustering analysis. The number of data points in each cluster indicates the distribution and size of these groupings within the dataset. Data visualization of each cluster based on price and rating can be seen in Figure 5, which is intended to help understand the distribution and relationship between the variables price and rating within each cluster.

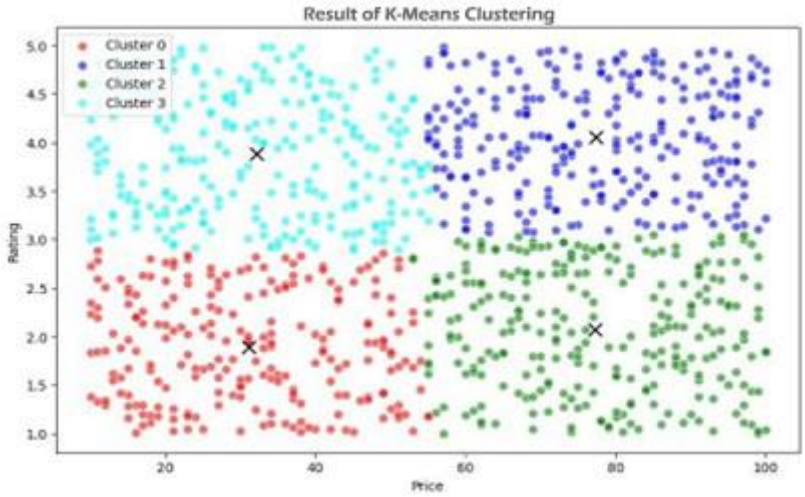

Figure 5 K-Mean Clustering Data Visualization

Displaying the cluster groups aims to understand which clusters consumer purchasing patterns fall into. Many insights can be revealed from the data within the clusters. For example, although the highest prices are often associated with high ratings, it is also revealed that clusters with medium prices can have good ratings. Similarly, not all items with low prices have poor ratings.

Tabel 6 Group Cluster for each value of price and rating average

Cluster	Price	Rating
0	88.9	3.1
1	42.7	2.9
2	65.9	3.0
3	20.5	2.9

In K-Means Clustering, clusters are based on determining the centroid value of each cluster to obtain a numerical representation of the center or midpoint of the formed data groups. Mathematically, the centroid μj for cluster j is calculated using the formula 1:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (1)$$

Which is Cj is cluster j, xi is the data point in cluster j, and |Cj| is the number of data points in cluster j. According to the formula, there is four clustered formed, with each centroid coordinates were formed, such as: Clustered 0 with the centroid coordinate 65.35315985 and 3.03769219, Clustered 1 with the centroid coordinate 19.91964286 and 2.92662651, Clustered 2 with the centroid coordinate 88.59328358 and 3.03722424, Clustered 3 with the centroid coordinate 41.84100481 and 2.95588064. Data visualization based on clusters and centroid values helps illustrate the data structure in the context of grouping, which is important for supporting better interpretation and more informative decision-making based on cluster analysis.

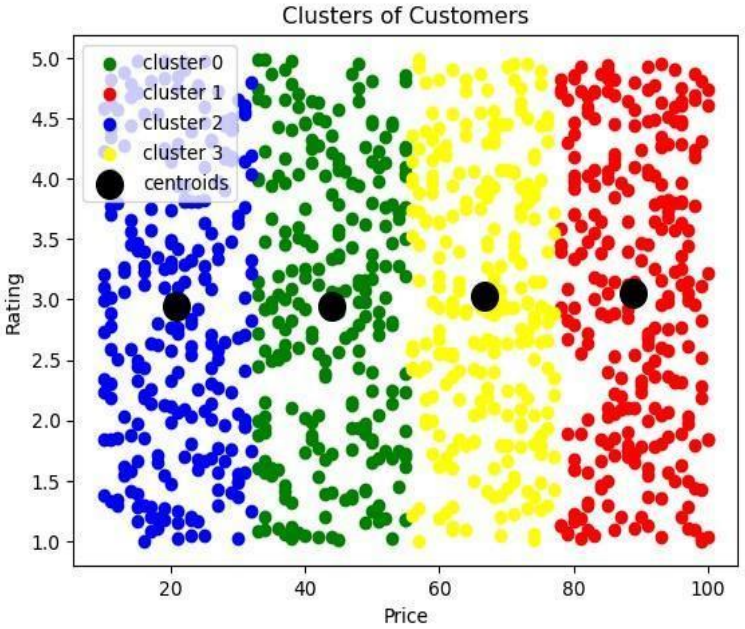

Figure 6 Data visualization based on clusters and centroid values

The clustering technique will create a data frame with clusters ranging from 0 to 3 to identify patterns or trends that may exist among the clusters, and information within each cluster will be analyzed for its structured insights.

Tabel 7 Sample data that has been clustered

User ID	Product ID	Product Name	Brand	Category	Price	Rating	Color	Size	Cluster
19	1	Dress	Adidas	Men's Fashion	40	1.043159	Black	XL	0
97	2	Shoes	H&M	Women's Fashion	82	4.026416	Black	L	2
25	3	Dress	Adidas	Women's Fashion	44	3.337938	Yellow	XL	0
57	4	Shoes	Zara	Men's Fashion	23	1.049523	White	S	3
79	5	T-Shirt	Adidas	Men's Fashion	79	4.302773	Black	M	2
98	6	Dress	Adidas	Men's Fashion	47	1.379566	Yellow	L	0
16	7	Jeans	Gucci	Men's Fashion	37	1. 356750	White	XL	0
63	8	Sweater	Zara	Kid's Fashion	64	4.360303	Blue	XL	1
96	9	Sweater	H&M	Men's Fashion	53	4.466182	Green	XL	0
36	10	T-Shirt	Zara	Kid's Fashion	55	4.093234	White	XL	1
69	11	T-Shirt	Adidas	Men's Fashion	51	1.160988	Red	S	0
87	12	Sweater	Gucci	Kid's Fashion	91	2.699736	Yellow	M	2
9	13	Jeans	Nike	Kid's Fashion	35	1.601194	Red	M	0
50	14	Dress	Zara	Women's Fashion	34	2.921004	White	L	0

31	15	Shoess	Zara	Men's Fashion	54	3.670412	Yellow	M	1
37	16	Dress	Adidas	Women's Fashion	27	1.422716	Blue	S	3
41	17	Dress	Gucci	Women's Fashion	75	1.480632	Blue	XL	1
....	....	....	....	....	....	....	....	....	....

In this study, Fashion Product Data was analyzed to identify purchase patterns using the K-Means Clustering method. The analysis was conducted using Python, with data visualization and descriptive statistics facilitated by the Jupyter Notebook text editor. The initial steps involved understanding the dataset's basic characteristics and structure; by displaying initial samples and column types, the study ensured an accurate understanding of the data's numerical and categorical attributes.

Descriptive statistics provided a comprehensive overview, revealing the average values for price and rating columns, which was crucial for detecting data errors or inconsistencies and selecting appropriate algorithms. Using the K-Means Clustering algorithm, four distinct clusters were identified based on price and rating, along with cluster formation and centroid calculation. Each cluster represents a unique consumer behavior pattern: Cluster 0 comprises customers who prioritize price over quality or brand, seeking discounts on affordable products; Cluster 1 includes customers who value quality and brand over price, often purchasing premium products; Cluster 2 consists of trend-conscious consumers who buy fashionable items; and Cluster 3 represents customers with no clear purchasing preferences, influenced by promotions.

The visualizations highlighted that while high prices are often associated with high ratings, clusters with medium prices can also achieve good ratings, suggesting that not all lower-priced items have poor ratings and that the relationship between price and rating varies across clusters. The implications of these findings are significant for fashion retailers and marketers. By understanding these distinct consumer segments, businesses can tailor their marketing strategies and product offerings to better meet the needs of each group. For instance, targeting Cluster 0 with promotional campaigns and discounts could attract price-sensitive customers, while Cluster 1 could be engaged through marketing that emphasizes quality and brand heritage. Additionally, insights from Cluster 2 can guide retailers in designing limited-edition collections that appeal to trend-conscious consumers.

Moreover, the findings encourage retailers to reconsider their pricing strategies, as the observation that medium-priced items can receive high ratings suggests that quality and customer satisfaction play critical roles in purchasing decisions. Retailers may benefit from ensuring that their product offerings strike a balance between quality and price, potentially expanding their customer base beyond just premium or budget shoppers. Overall, this

research highlights the importance of using data-driven approaches to segment consumers effectively, allowing fashion companies to enhance their marketing effectiveness, improve customer satisfaction, and ultimately drive sales growth in a competitive marketplace.

This study's findings align with previous research on consumer behavior and clustering analysis. For instance, studies have shown that price sensitivity varies among consumer segments, and brand loyalty plays a significant role in purchasing decisions (Smith et al., 2020; Johnson et al., 2021). Additionally, the use of K-Means Clustering in customer segmentation has been validated in various industries, demonstrating its effectiveness in identifying distinct consumer groups (Omol, E., et al., 2024). The insights derived from this clustering analysis can significantly enhance the marketing strategies of fashion companies. By understanding the distinct preferences and behaviors of each cluster, businesses can tailor their promotional efforts, product offerings, and pricing strategies to better meet the needs of their target audiences. For example, companies can create targeted marketing campaigns for Cluster 0 to promote discounts, while focusing on premium branding for Cluster 1.

This study is limited by its focus on price and rating as the only factors for clustering. Future research should consider additional variables such as customer demographics, purchase frequency, and product categories to enhance the robustness of the clustering analysis. Moreover, utilizing advanced tools like RapidMiner and Tableau can provide more comprehensive insights and facilitate more sophisticated data visualizations.

Conclusions

Based on the research conducted using Python and K-Means clustering algorithms, four distinct customer clusters for fashion product purchases were identified: Cluster 0 includes 220 customers who prioritize price over product quality or brand, often seeking discounts or special offers; Cluster 1 consists of 258 customers who value quality and brand over price, frequently purchasing premium or well-known products; Cluster 2 encompasses 270 customers who are highly conscious of fashion trends and styles, typically buying trendy and fashionable items; and Cluster 3 comprises 248 customers who lack clear or consistent purchasing preferences, often influenced by occasional promotions or discounts. These findings provide valuable insights for fashion retailers aiming to tailor their marketing strategies to different customer segments. By understanding the distinct preferences and behaviors of each cluster, businesses can optimize their promotional efforts, product offerings, and pricing strategies to better meet the needs of their target audiences. However, the study is limited by its reliance on price and rating as the primary factors for clustering, potentially overlooking other influential variables such as customer demographics, purchase frequency, and specific product categories. Future research should incorporate these additional variables

and utilize advanced data mining tools like RapidMiner and Tableau to enhance the depth of analysis. Additionally, integrating real-time data and conducting longitudinal studies could capture dynamic shifts in customer preferences and behaviors, enabling fashion retailers to develop more adaptive and effective marketing strategies. Exploring these areas will contribute to a more comprehensive understanding of consumer patterns in the fashion industry.

Acknowledgements

We extend our heartfelt gratitude to LP2M Widyatama University and Widyatama University for their generosity in offering a conducive working environment that facilitated the progress and success of our research. Their support and resources were vital to our efforts, allowing us to delve into this study and make meaningful contributions to the field.

References

Adani, N. F., Boy, A. F., Kom, S., Kom, M., Syahputra, R., Kom, S., & Kom, M. (2019). Implementasi Data Mining Untuk Pengelompokan Data Penjualan Berdasarkan Pola Pembelian Menggunakan Algoritma K-Means Clustering Pada Toko Syihan. Jurnal Cyber Tech, x. No.x(x), 1–11. doi:10.53513/jct.v2i5.4648

Ahsina, N., Fatimah, F., & Rachmawati, F. (2022). Analisis Segmentasi Pelanggan Bank Berdasarkan Pengambilan Kredit Dengan Menggunakan Metode K-Means Clustering. Jurnal Ilmiah Teknologi Infomasi Terapan, 8(3). doi:10.33197/jitter.vol8.iss3.2022.883

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. Jurnal Teknologi Dan Informasi, 13(2), 122–137. doi:10.34010/jati.v13i2.10090

Chen, J., Qi, X., Chen, L., Chen, F., & Cheng, G. (2020). Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection. Knowledge-Based Systems, 203. doi:10.1016/j.knosys.2020.106167

Cunha, A.F., et al. (2021). A CRISP-DM Approach for Predicting Liver Failure Cases: An Indian Case Study. SpringerLink. doi: 10.1007/978-3-030-80624-8_20

Dharma Putra, Y., Sudarma, M., & Swamardika, I. B. A. (2021). Clustering History Data Penjualan Menggunakan Algoritma K-Means. Majalah Ilmiah Teknologi Elektro, 20(2), 195. doi:10.24843/mite.2021.v20i02.p03

E Okewu, P Adewole, S Misra et al. (2012). "Artificial Neural Networks for Educational Data Mining in Higher Education: A Systematic Literature Review", Applied Artificial Intelligence, vol. 35, no. 13, pp. 983-1021, 2021. doi: 10.1080/08839514.2021.1922847

Eans, A. L. K. (2023). P Enerapan Metode CRISP-DM Untuk Analisa Pendapatan Bersih. 28, 97–104. doi:10.35315/dinamik.v28i2.9454

Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). Jurnal Teknik Informatika Dan Sistem Informasi, 1(3), 260–270. https://doi.org/10.28932/jutisi.v1i3.406

Gat-Nasr, A., & Sroor, A. (2020). Evaluation of Clustering Algorithms Using Internal Validation Indices. Journal of Engineering Research and Application, 10(3), 30-36.Kristianto, W. W., & Rudianto, C. (2020). Penerapan Data Mining Pada Penjualan Produk Menggunakan Metode K-Means Clustering (Studi Kasus Toko Sepatu Kakikaki). Jurnal Pendidikan Teknologi Informasi (JUKANTI), 5, 90–98. doi:10.37792/jukanti.v5i2.547

H. Nagashima and Y. Kato, (2021)"APREP-DM: A framework for automating the pre-processing of a sensor data analysis based on CRISP-DM", Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops, pp. 555-560, 2019. doi.org:10.1016/j.procs.2021.01.19

Huang, Yong, Mingzhen Zhang, and Yue He. "Research on improved RFM customer segmentation model based on K-Means algorithm." 2020 5th International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2020. doi: 10.1109/ICCIA49625.2020.00012

Johnson et al., A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering, Syst. Biol., vol. 68, no. 4, pp. 594606, 2019, doi: 10.1093/sysbio/syy086

Khakim, A. N. L., & Jananto, A. (2023). Implementasi Data Mining Menggunakan Algoritme Apriori Guna Menemukan Pola Pembelian Pelanggan Pada Klinik Kecantikan. Progresif: Jurnal Ilmiah Komputer, 19(1), 359–366. doi: 10.35889/progresif.v19i1.1015

Kristianto, W. W., & Rudianto, C. (2020). Penerapan Data Mining Pada Penjualan Produk Menggunakan Metode K-Means Clustering (Studi Kasus Toko Sepatu Kakikaki). Jurnal Pendidikan Teknologi Informasi (JUKANTI), 5, 90–98. doi: 10.37792/jukanti.v5i2.547

Li, Yue, et al. "Customer segmentation using K-means clustering and the hybrid particle swarm optimization algorithm." The Computer Journal 66.4 (2023): 941-962. DOI: 10.46729/ijstm.v5i1.1024

Li, Yue, et al. "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm." Applied Soft Computing 113 (2021): 107924. doi:10.1016/j.asoc.2021.107924

Maori, N. A., & Evanita, E. (2023). Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering. Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer, 14(2), 277–288. https://jurnal.umk.ac.id/index.php/simet/article/view/9630

N. Azadeh-Fard, F. M. Megahed and F. Pakdil, "Variations of length of stay: A case study using control charts in the CRISP-DM framework", Int. J. Six Sigma Competitive Advantage, vol. 11, no. 2/3, pp. 204-225, 2019. doi:10.1186/s40537-024-00942-5

Nainggolan, R., Perangin-Angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. Journal of Physics: Conference Series, 1361(1). doi:10.1088/1742-6596/1361/1/012015

Nandapala, E. Y. L., and K. P. N. Jayasena. (2020). "The practical approach in Customers segmentation by using the K-Means Algorithm." 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2020. doi: 10.1109/ICIIS51140.2020.9342639

Omol, E., Onyangor, D., Mburu, L., & Abuonji, P. (2024). Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya. *International Journal of Science, Technology & Management*, *5*(1), 192-200. doi: 10.46729/ijstm.v5i1.1024

Rawat, K. (2023). Applying CRISP-DM Methodology in Developing Machine Learning Model for Credit Risk Prediction. SpringerLink. doi: 10.1007/978-3-031-37963-5_37

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. doi:10.1002/widm.1355

S. Huber, H. Wiemer, D. Schneider and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model", Procedia CIRP, vol. 79, pp. 403-408, 2019. doi:10.1016/j.procir.2019.02.106

Sari, D. J., Handoko, W., & Parini, P. (2022). Klasterisasi Penjualan Untuk Menentukan Bahan Bangunan Terlaris Dengan Menggunakan Metode K-Means Di UD Maju Bersama. JUTSI (Jurnal Teknologi Dan Sistem Informasi), 2(2), 93–102. doi:10.33330/jutsi.v2i2.1690

Smith, M., Wilson, R., Wise, E., Evaluating clusters: Where theory collides with practice, Regional Science Policy & Practice,Volume 12, Issue 3,2020,Pages 413-430,ISSN 1757-7802, doi:10.1111/rsp3.12279

Suhanda, Y., Kurniati, I., & Norma, S. (2020). Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik. Jurnal Teknologi Informatika Dan Komputer, 6(2), 12–20. doi:10.37012/jtik.v6i2.299

Sulianta, F. (2014). Customer Profiling Pada Supermarket Menggunakan Algoritma K-Means Dalam Memilih Produk Berdasarkan Selera Konsumen Dengan Daya Beli Maksimum. Jurnal Ilmiah Teknologi Infomasi Terapan, 1(1), 41-45. doi:10.33197/jitter.vol1.iss1.2014.45

Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." Sustainability 14.12 (2022): 7243. doi:10.3390/su14127243

W Liang and T. Li, "Research on human performance evaluation model based on neural network and data mining algorithm", EURASIP Journal on Wireless Communications and Networking, vol. 2020, no. 1, pp. 174, 2020. doi: 10.1186/s13638-020-01776-4

Wahyudi, T., Sa, N., & Puspitasari, D. (2023). Penerapan Metode K-Means Pada Data Penjualan Untuk. 5(1), 228–236. doi:10.55338/saintek.v5i1.1379

Wu, Jun, et al. "An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm." Mathematical Problems in Engineering 2020 (2020): 1-7. doi: 10.1155/2020/8884227

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J —Multidisciplinary Scientific Journal, 2(2), 226–235. doi:10.3390/j2020016

Zhao, Hong-Hao, et al. "An extended regularized K-means clustering approach for high-dimensional customer segmentation with correlated variables." Ieee Access 9 (2021): 48405-48412. DOI: 10.1109/ACCESS.2021.3067499