Annexure-II

Cover Page

ClusterCraft: Redefining Customer Profiles

A Project Report

Submitted in partial fulfillment of the requirements for the

Award of the degree of

Master of Business Administration

By

Ragini Kumari

323102744

Centre for Distance and online Education

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

2023-2025

Annexure-III

Declaration by the Student

To whom-so-ever it may concern

I, **Ragini Kumari, 323102744**, hereby declare that the work done by me on **"ClusterCraft: Redefining Customer Profiles"**, is a record of original work for the partial fulfilment of the requirements for the award of the degree, **Master of Business Administration**.

Name of the student: Ragini Kumari (323102744)

Signature of the Student: Ragini

Dated: 13th May, 2025

Acknowledgement

I would like to express my sincere gratitude to **Lovely Professional University (LPU Online)** for providing the opportunity to undertake this project as part of my MBA Capstone Project. This project has allowed me to bridge theoretical concepts with real-world application, especially in the field of customer analytics and data-driven decision-making.

I am deeply thankful for the abundance of freely available resources and platforms that supported my learning. In particular, **Kaggle** served as a valuable source for dataset exploration and research insights, while **Google Colab** provided a powerful environment to execute and test my code effectively. Additionally, tools like Python, Pandas, Scikit-learn, and Matplotlib were instrumental in performing data analysis and visualization.

I also extend heartfelt thanks to my family and friends for their unwavering encouragement and support throughout the course of this project. Their motivation helped me stay focused and driven during every stage of this work.

Abstract

In today’s data-driven business environment, understanding customer behavior has become critical for designing effective marketing strategies and improving customer engagement. This project, titled *“ClusterCraft: Redefining Customer Profiles,”* explores the application of K-Means clustering to segment customers based on behavioral data such as income, demography, and purchase frequency. By preprocessing the dataset—handling missing values, duplicates, and normalizing key features—clean and structured data was prepared for analysis.

The implementation of the K-Means algorithm revealed three meaningful customer segments, each exhibiting distinct behavioral traits. These insights offer a strategic foundation for targeted marketing, personalized communication, and improved customer retention. The project highlights the practical value of unsupervised machine learning techniques in solving real-world business problems, especially in the context of customer analytics. Through this approach, businesses can move beyond traditional demographics and build a deeper, more actionable understanding of their customers.

Table of contents

Chapter 1: Introduction

Aim:

In today’s highly competitive business environment, understanding your customers at a deeper level is no longer optional—it’s essential. This project, titled **“ClusterCraft: Redefining Customer Profiles”**, explores how K-Means clustering, a machine learning technique, can be applied to customer data to uncover hidden patterns in purchasing behavior.

The goal is simple: segment customers not just by traditional demographics, but based on their actual behavior—how often they buy, how much they spend, and what kind of products they prefer. Using a real-world dataset and a structured analytical approach, we grouped customers into distinct segments, each with unique traits and value to the business.

These insights open up new opportunities for more personalized marketing strategies—moving away from “one-size-fits-all” to targeted campaigns that resonate with each group. Whether it’s identifying high-value loyal customers or casual bargain seekers, the project offers a practical and scalable way for businesses to engage better and drive results.

This project follows a structured, data-driven methodology to segment customers using the K-Means clustering algorithm. The process involved five key steps: data collection, preprocessing, feature selection, clustering, and interpretation.

1.Data Collection

For this analysis, a customer purchase data from a jewelry store, offering a multifaceted view of customer interactions was used, the data encapsulates 1,029 customers across various dimensions such as demographics, shopping behaviors and sentiments. With a rich array of information, it offers an opportunity to unravel trends, identify customer segments, and gain actionable insights for personalized marketing strategies.

The dataset simulates real-world retail data and was chosen for its suitability for behavioral segmentation.

2. Data Preprocessing

Before applying clustering, the data was cleaned and prepared:

- Missing values were checked appropriately, and it was found that the original dataset had no missing values. No duplicate rows were present.
- Column names of the dataset were not consistently formatted. To facilitate further processing, we removed leading and trailing spaces, replaces spaces with underscores, and converts all column names to lowercase. Additionally, we have renamed some columns to make them more concise and user-friendly.
- There was a need for datatype adjustment as some features were stored as strings but should be numerical. All necessary conversions were completed using custom functions.
- Outliers were checked using visual tools like boxplots.
- Scaling was applied using Min-Max normalization to bring all numeric features to the same range. This step is critical since K-Means relies on distance calculations, and unscaled data could bias the results.

3. Feature Selection

We selected key variables that could influence customer behavior, such as:

- Age
- Income

- Average order value
- Purchase Frequency
- Return rate, etc.

These variables offer a good balance of financial, behavioral, and frequency-based data—ideal for forming meaningful clusters.

4. Clustering with K-Means

The K-Means algorithm was implemented using Python's Scikit-learn library. The number of clusters (K) was decided using the Elbow Method, where The KElbowVisualizer() from the yellowbrick library pinpointed an elbow at $k = 3$, suggesting a potential optimal number of clusters provided the best balance between simplicity and segmentation quality.

5. Cluster Interpretation

After clustering, each group was analyzed to understand its unique characteristics. Visual tools for visualizing these clusters in 2D and 3D spaces were used to support this interpretation.

As a result, the plot shows that K-Means effectively groups high-value customers into **Cluster 1**. Among the lower-value customers, those who have shopped recently are placed in **Cluster 2**, while those who haven't interacted with the store for a while fall into **Cluster 3**.

These segments were then labeled based on observed behaviors.

Cluster 1: High-Value Loyalists (Customers with consistently high value and likely strong brand engagement)

Cluster 2: Recent Opportunity Seekers (Lower overall value but high recent activity - potential for upselling or nurturing)

Cluster 3: At-Risk or Re-engageable (formerly Dormant/Disengaged) (Low value and long periods of inactivity - may need reactivation campaigns)

Overall, this project highlights how machine learning can transform raw data into actionable customer insights. By leveraging clustering techniques, businesses can go beyond demographics and tap into behavioral signals, leading to improved customer engagement, better resource allocation, and ultimately, increased profitability.

Importance:

In today’s market, businesses are collecting massive amounts of customer data—but data alone doesn’t drive results. What truly matters is how that data is used. This is where the importance of customer segmentation comes into play. Instead of treating every customer the same, businesses now have the opportunity to understand and cater to their diverse needs more effectively.

This project focuses on using K-Means clustering to make sense of customer behavior—how often they shop, how much they spend, and what kinds of products they prefer. Traditional segmentation methods often rely on demographics, but behavior-based clustering gives us deeper insights that can lead to smarter marketing decisions and better customer experiences.

In short, this project matters because it turns raw data into actionable insights, and that’s exactly what businesses need to stay competitive and relevant.

Applicability

The beauty of this approach is how versatile and widely applicable it is. Here are some practical areas where this kind of clustering model can be used:

- **Retail & E-commerce**: Targeted promotions based on customer groups (e.g., high spenders vs. discount seekers)

- **Banking**: Differentiating between high-risk and low-risk customers or high-value clients
- **Hospitality**: Personalizing offers for frequent travelers vs. occasional guests
- **Telecom**: Creating tailored packages for heavy vs. light users

This project provides a solid, data-driven foundation that can be applied across various industries wherever customer data is available.

Scope of the Project

The scope of this project is focused but impactful. It covers:

- Collecting and preparing customer data
- Applying the K-Means clustering algorithm
- Interpreting the resulting segments
- Recommending strategies tailored to each group

While the project uses a specific dataset, the methodology is scalable. It can be expanded by adding more data points—like customer lifetime value, product preferences, or engagement levels—and even more advanced clustering techniques (like Hierarchical or DBSCAN) in future work.

This ensures that the project isn't just a one-time analysis but a stepping stone for continuous customer intelligence.

Relevance in Today's Business Landscape

We're living in an era where personalization isn't just a bonus—it's the baseline expectation. For instance: today's consumers are digitally savvy, well-informed, and have more options than ever before. They expect businesses to not only understand their preferences but also to anticipate their needs, communicate through the right channels, and offer experiences that feel tailor-made. In such a landscape, generic, mass-targeted marketing efforts fall flat. Customers simply tune them out, and businesses lose valuable opportunities to connect meaningfully.

This is where **customer segmentation becomes not just relevant but critical**. Effective segmentation enables businesses to differentiate between customer types: for instance, a first-time buyer versus a long-standing loyalist, or a price-sensitive shopper versus someone who actively seeks premium products. Without these distinctions, marketing efforts become inefficient, customer satisfaction dips, and consequently, valuable insights remain unutilized. To matters worse, companies may end up alienating key segments by sending irrelevant messages or offers.

Another pressing concern is the **rising cost of customer acquisition**. With advertising costs climbing and competition increasing across sectors, acquiring new customers is becoming more expensive than ever. This has prompted a strategic shift: companies are now prioritizing customer retention and maximizing lifetime value over one-time acquisitions. And to do this effectively, they must truly understand their existing customers: what drives them, how they interact, and what keeps them coming back. Behavioral segmentation powered by models like **K-Means clustering** helps businesses unlock these insights in a systematic and scalable way.

What’s particularly exciting is that this type of data-driven segmentation is no longer just the domain of data scientists. Today’s business professionals, whether they’re in marketing, product development, sales, or strategy, are increasingly expected to engage with data and use tools like clustering to inform their decisions. **K-Means, once considered a technical technique, is now entering mainstream business toolkits** due to its simplicity, interpretability, and actionable outputs.

Meanwhile, the **digital transformation** sweeping across industries has created a massive surge in customer data, coming from websites, mobile apps, email interactions, loyalty programs, social media, and more. The problem now isn't about having access to data, it's about making sense of that data, cutting through the noise, and identifying patterns that actually matter. This is exactly what clustering algorithms like K-Means excel at: uncovering hidden structures and revealing actionable customer groupings that businesses might otherwise miss.

That's why this project is not only timely—it's immensely practical. It demonstrates how even basic machine learning methods, when applied correctly, can **redefine the way businesses view and serve their customers**. From transforming raw behavioral data into intelligent customer profiles to empowering more personalized and impactful marketing strategies, the value is clear. In today's data-rich business environment, initiatives like this are not just relevant—they are essential for sustainable growth, competitive advantage, and customer loyalty.

Chapter-2: Review of Literature

Understanding customer behavior has long been a central theme in marketing research. Over the years, the methodologies for segmenting customers have evolved from basic demographic classification to more advanced, data-driven models. This section presents a chronological review of the key literature relevant to customer segmentation and clustering, with a focus on the evolution, application, and limitations of K-Means clustering in marketing analytics.

1. Early Foundations of Customer Segmentation

Smith (1956) first introduced the concept of market segmentation, emphasizing the importance of grouping consumers with similar needs to tailor marketing strategies effectively. This laid the groundwork for decades of segmentation work based largely on **demographics and psychographics**.

Kotler (1984) expanded on these ideas, formalizing segmentation as a core element of marketing strategy. However, early models relied heavily on subjective judgment and lacked the analytical rigor made possible by today's data capabilities.

Gap Identified: These early approaches did not leverage transactional or behavioral data and were limited in dynamic application.

2. Emergence of Data-Driven Segmentation

With the growth of databases in the 1990s, researchers began to explore segmentation using quantitative methods. Wedel & Kamakura (2000) emphasized **data-driven segmentation** using statistical models and laid the foundation for incorporating **consumer behavior** into segmentation.

Their work introduced clustering algorithms to marketing analytics, including K-Means, but practical applications remained limited due to lack of computing power and data availability in small to mid-sized firms.

Gap Identified: While theoretical support for clustering existed, real-world implementation in customer profiling was sparse and not easily accessible to all business sectors.

3. Rise of Machine Learning in Customer Segmentation

In the 2010s, the widespread availability of customer data (especially from e-commerce and CRM systems) and tools like Python and R made clustering models more approachable.

Chaturvedi et al. (2014) demonstrated the use of **K-Means clustering** on retail customer datasets and showed clear segmentation based on purchasing behavior. Tsiptsis & Chorianopoulos (2011) also discussed clustering for customer retention, but both works highlighted a limitation: K selection was arbitrary, and clustering often lacked interpretability.

Gan, Ma, & Wu (2007) had earlier introduced refinements in clustering techniques, but many of those required expert-level statistical skills.

Gap Identified: Although clustering became more common, issues like scalability, interpretability, and practical business alignment remained.

4. Latest Advancements and Practical Use

Singh and Rani (2019) analyzed K-Means performance using real customer datasets and recommended enhancements like hybrid clustering approaches. More recently, studies like Ahmed & Hussain (2021) and Yadav et al. (2023) have explored behavioral clustering in e-commerce, emphasizing the need for segmentation models that adapt with customer behavior over time.

Jain & Gupta (2024) emphasized the integration of clustering with AI-powered recommendation systems, while acknowledging that many small businesses still lack the analytical maturity to apply such methods effectively.

Gap Identified: There is a continued need for segmentation models that are simple, interpretable, and actionable, especially for mid-sized businesses that can’t implement complex AI systems.

Conclusion & Need for Current Study

While the body of literature provides a strong foundation on clustering and customer segmentation, **there remains a clear gap** in studies that combine:

- Simple but powerful clustering methods like K-Means
- Focus on behavioral data over traditional demographics
- A structured end-to-end process that is scalable and business-oriented
- Easy visual interpretation and strategic marketing application

This project, “ClusterCraft: Redefining Customer Profiles,” directly addresses these gaps by developing a **clean, reproducible, and insight-driven framework** for behavioral segmentation using K-Means clustering, supported with visualization and marketing strategy recommendations that can be applied in real business contexts.

Chapter 3: Implementation of Project

The implementation of this project focused on using a **data-driven approach to segment customers** using the K-Means clustering algorithm. The core aim was to uncover behavioral patterns in customer data that could inform more targeted marketing strategies. This section outlines the full process—from setting objectives to selecting tools, designing the workflow, running the analysis, and interpreting the results.

Objectives of the Project

The main objectives of the project were:

1. **To segment customers based on behavioral attributes** such as age, income, average order rate, purchase frequency, return rate.
2. **To implement the K-Means clustering algorithm** to identify hidden customer segments.
3. **To visualize and interpret the clusters** in a way that helps marketers develop personalized campaigns.
4. **To evaluate the effectiveness of clustering** in improving marketing decision-making.

Tools, Techniques & Design Framework

The project was implemented using a blend of **data science tools, statistical methods, and visualization libraries**. Here's how it was structured:

Tools Used

- **Python**: Core programming language used for data analysis and modeling
- **Google Colab**: For organizing code and documentation

- **Pandas & NumPy**: Data manipulation and numerical computation
- **Matplotlib & Seaborn**: Data visualization
- **Scikit-learn**: For implementing K-Means clustering and other ML tasks

Methodology & Step-by-Step Implementation

1. Data Acquisition

Customer purchase data was sourced from a jewelry store's dataset, containing fields such as:

- Customer_id
- Age
- Gender
- Location
- Income
- Average order value
- Purchase Frequency
- Return rate, etc.

This dataset was suitable for simulating real-world consumer behavior in a B2C context.

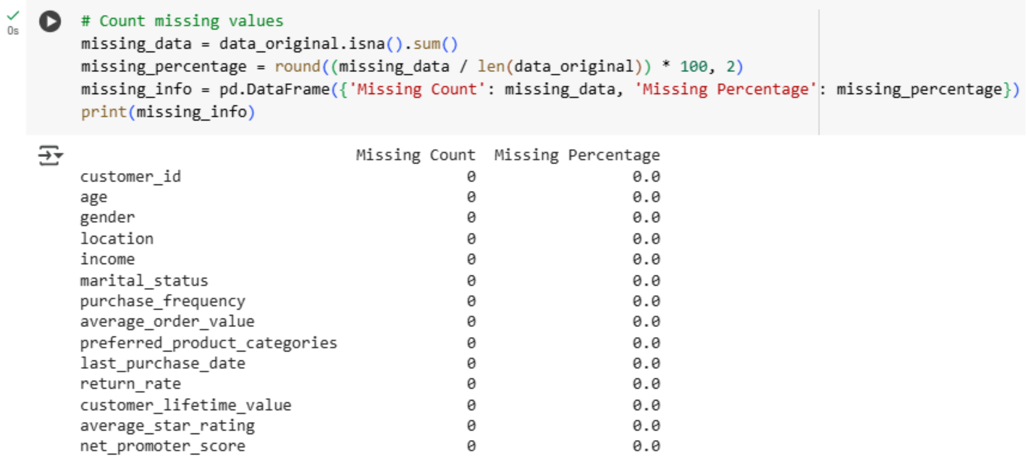
2. Data Preprocessing

Before applying any model, the dataset underwent essential preprocessing:

- **Handling Missing Values**: Checked for nulls, snapshot attached in Fig 1. Below
- **Handling duplicates rows**: No duplicate rows were present. snapshot attached in Fig 2. Below
- **Handling spaces in field names:** Column names of the dataset were not consistently formatted. To facilitate further

processing, we removed leading and trailing spaces, replaces spaces with underscores, and converted all column names to lowercase. Additionally, we have renamed some columns to make them more concise and user-friendly.

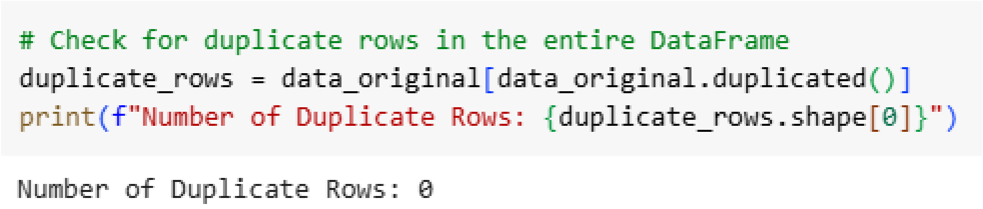
- **Data type adjustment:** There was a need for datatype adjustment as some features were stored as strings but should be numerical. All necessary conversions were completed using custom functions.
- **Feature Scaling for entire data**: Applied **Min-Max Normalization** to bring features to a similar range, which is crucial for distance-based algorithms like K-Means, shown below in table 1.



```
# Count missing values
missing_data = data_original.isna().sum()
missing_percentage = round((missing_data / len(data_original)) * 100, 2)
missing_info = pd.DataFrame({'Missing Count': missing_data, 'Missing Percentage': missing_percentage})
print(missing_info)
```

```
                             Missing Count  Missing Percentage
customer_id                              0                 0.0
age                                      0                 0.0
gender                                   0                 0.0
location                                 0                 0.0
income                                   0                 0.0
marital_status                           0                 0.0
purchase_frequency                       0                 0.0
average_order_value                      0                 0.0
preferred_product_categories             0                 0.0
last_purchase_date                       0                 0.0
return_rate                              0                 0.0
customer_lifetime_value                  0                 0.0
average_star_rating                      0                 0.0
net_promoter_score                       0                 0.0
```

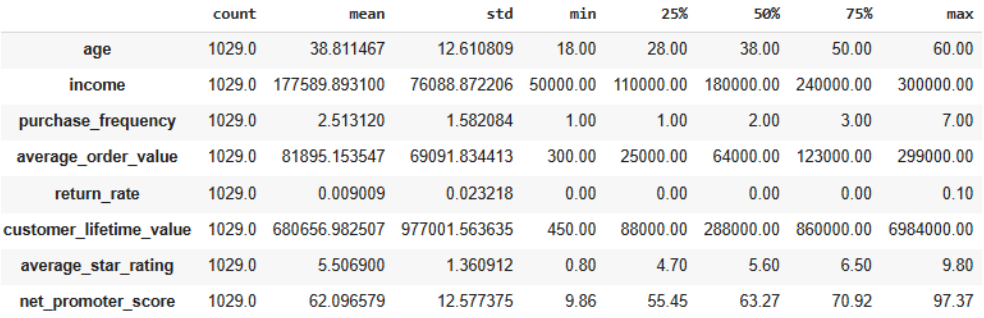
***Fig 1**: Overview of missing values in the dataset.*



```
# Check for duplicate rows in the entire DataFrame
duplicate_rows = data_original[data_original.duplicated()]
print(f"Number of Duplicate Rows: {duplicate_rows.shape[0]}")
```

```
Number of Duplicate Rows: 0
```

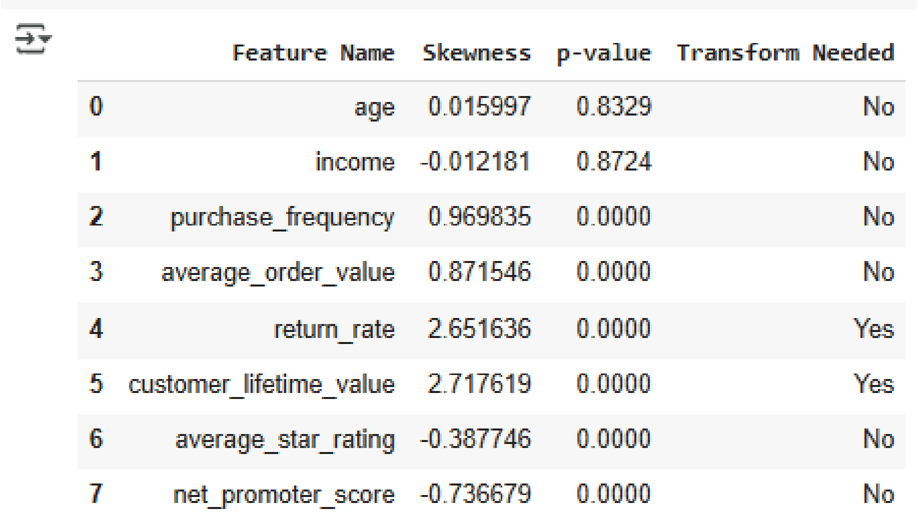
***Fig 2**: Analysis of data duplication across rows*



	count	mean	std	min	25%	50%	75%	max
age	1029.0	38.811467	12.610809	18.00	28.00	38.00	50.00	60.00
income	1029.0	177589.893100	76088.872206	50000.00	110000.00	180000.00	240000.00	300000.00
purchase_frequency	1029.0	2.513120	1.582084	1.00	1.00	2.00	3.00	7.00
average_order_value	1029.0	81895.153547	69091.834413	300.00	25000.00	64000.00	123000.00	299000.00
return_rate	1029.0	0.009009	0.023218	0.00	0.00	0.00	0.00	0.10
customer_lifetime_value	1029.0	680656.982507	977001.563635	450.00	88000.00	288000.00	860000.00	6984000.00
average_star_rating	1029.0	5.506900	1.360912	0.80	4.70	5.60	6.50	9.80
net_promoter_score	1029.0	62.096579	12.577375	9.86	55.45	63.27	70.92	97.37

***Table 1**: Feature Scaling: Applied Min-Max Normalization*

After comparing the mean and median of each feature, certain features exhibited skewness. We then use the scipy_stats skewtest() function to conduct a more rigorous skewness test.



	Feature Name	Skewness	p-value	Transform Needed
0	age	0.015997	0.8329	No
1	income	-0.012181	0.8724	No
2	purchase_frequency	0.969835	0.0000	No
3	average_order_value	0.871546	0.0000	No
4	return_rate	2.651636	0.0000	Yes
5	customer_lifetime_value	2.717619	0.0000	Yes
6	average_star_rating	-0.387746	0.0000	No
7	net_promoter_score	-0.736679	0.0000	No

***Table 2**: Output of scipy.stats.skewtest()*

Addressing Skewed Features: For the two features that need log-transform, there distributions are visualized below:

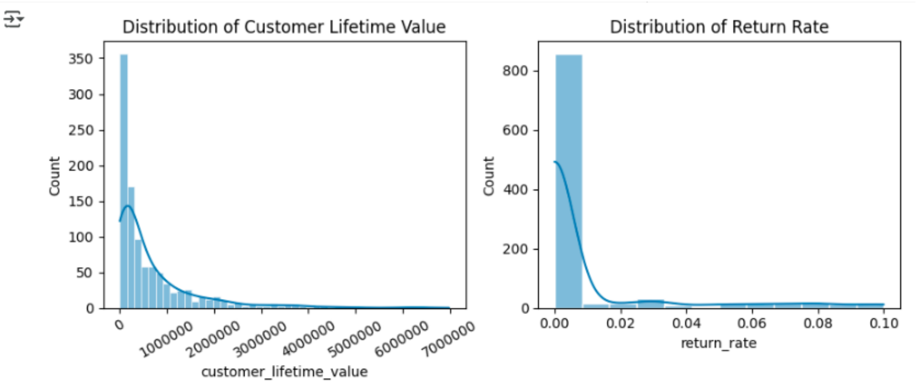


***Fig 3**: Distribution of the two features that need log-transform*

The plots above reveal a significant right skewness in both features. Hence, a log-transform is applied to mitigate this skewness. The transformation results are shown below:

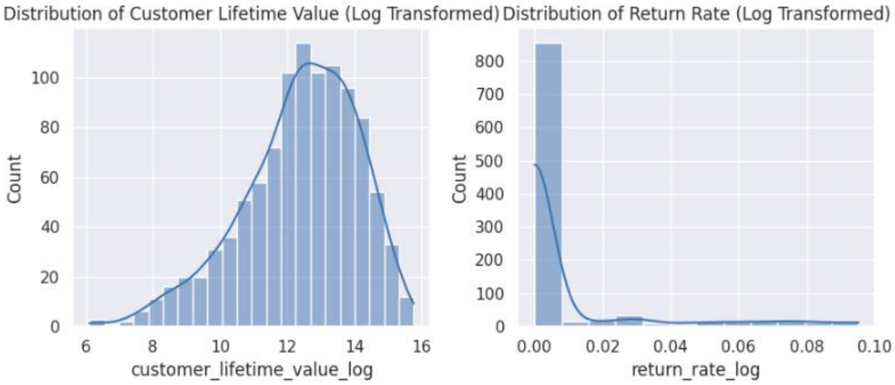


***Fig 4**: transformation results after log-transform is applied to mitigate this skewness*

Both the plots and the skewness coefficients indicate that after log-transformation, the customer_lifetime_value feature is approaching a normal distribution. However, the return_rate feature remains right-skewed.

Given the significant imbalance and potential skewness introduced by the return-related feature, we decide to exclude it from the initial clustering analysis. However, recognizing the significance of return behavior in understanding customer segments, this feature will be re-introduced during the customer profiling stage after clusters are formed.

3. Feature Engineering

Feature selection before clustering is crucial for refining input variables and eliminating irrelevant information. This enhances clustering accuracy and simplifies result interpretation, yielding more effective insights. Our focus is on including the most relevant features and excluding less pertinent ones.

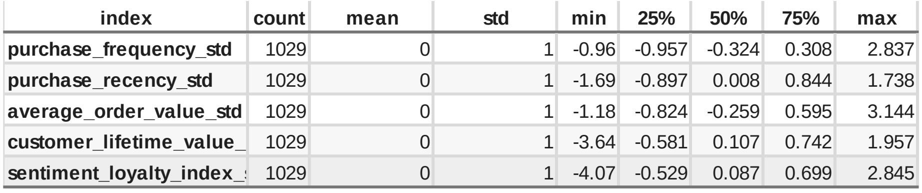
Based on the RFM model, the following features are most relevant:

- **days_since_last_purchase**: This feature measures purchase recency
- **purchase_frequency:** This feature measures frequency.
- **average_order_value:** This feature measures one facet of monetary value: how much money a customer spends each time.
- **customer_lifetime_value:** This feature measures the long-term monetary value of a customer. However, due to its skewness, the log-transformation, **customer_lifetime_value_log**, will be used for clustering.
- Another feature, **sentiment_loyalty_index**, is also important for behavioral customer segmentation. It distills two crucial aspects: **average_star_rating** gauges sentiment based on product and review ratings, **while net_promoter_score** quantifies customer loyalty

These features were selected because they reflect both the **financial capacity** and **shopping behavior** of customers.

4. Feature Scaling

Before clustering, the final step in data preprocessing involves scaling the data. Many clustering algorithms are sensitive to differences in the scales of features, and standardization is employed to ensure that all features contribute equally to the clustering process. Shown in Table 3 below



index	count	mean	std	min	25%	50%	75%	max
purchase_frequency_std	1029	0	1	-0.96	-0.957	-0.324	0.308	2.837
purchase_recency_std	1029	0	1	-1.69	-0.897	0.008	0.844	1.738
average_order_value_std	1029	0	1	-1.18	-0.824	-0.259	0.595	3.144
customer_lifetime_value_	1029	0	1	-3.64	-0.581	0.107	0.742	1.957
sentiment_loyalty_index_	1029	0	1	-4.07	-0.529	0.087	0.699	2.845

Table 3: *Feature scaling of selected fields for clustering*

5. Finding Optimal Number of Clusters (K)

To determine the best number of clusters (K), we used the:

- **Elbow Method**: We employ the elbow method to determine the appropriate number of clusters. This involves calculating the inertia (distortion score) for a range of clusters (2 to 10 in this case) and selecting the point where a kink or 'elbow' is observed on the graph.

The KElbowVisualizer() from the yellowbrick library pinpointed an elbow at k = 3, suggesting a potential optimal number of clusters. Shown in Fig 3.

- **Silhouette Score:** Additionally, we create a plot of the silhouette score against the number of clusters as a complementary analysis.

The silhouette score peaks at *k* = 2. Considering that choosing only 2 clusters might result in overly broad groups lacking specificity. Shown in Fig 4.

We opt for *k* = 3, prioritizing informative clustering over the silhouette score.

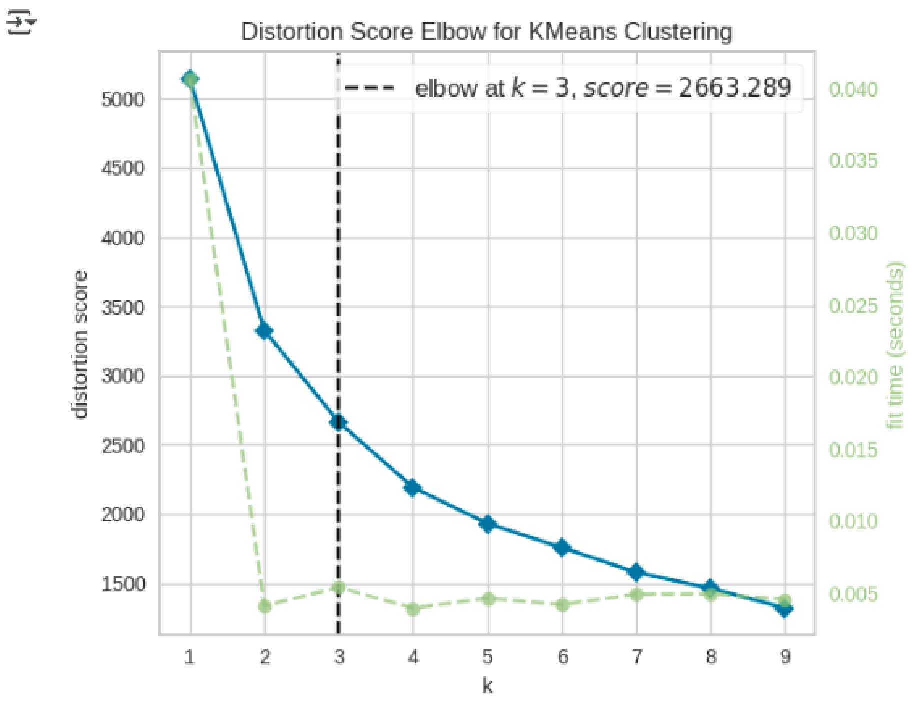


Fig 5*: Distortion Score Elbow for KMeans Clustering*

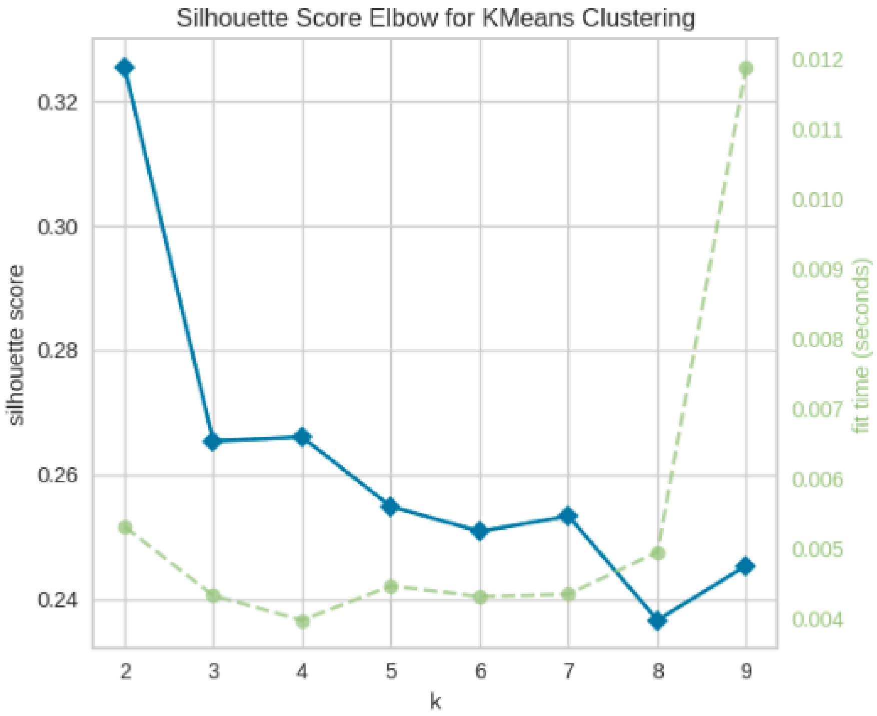
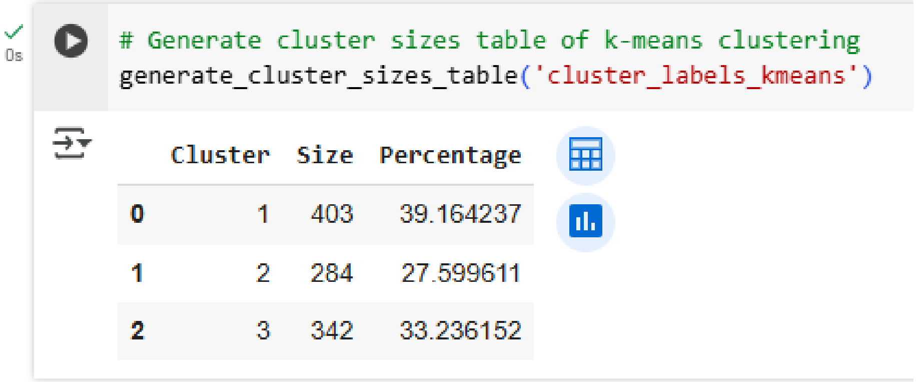


Fig 6: *Distortion Score Elbow for KMeans Clustering*

6. Applying K-Means Clustering

Now we use k-means to segment the customers into 3 clusters and with K decided, the K-Means model was applied to the scaled data using Scikit-learn’s K-Means module. Then we generated cluster sizes table of k-means clustering. Shown in Table 5.



```
# Generate cluster sizes table of k-means clustering
generate_cluster_sizes_table('cluster_labels_kmeans')
```

	Cluster	Size	Percentage
0	1	403	39.164237
1	2	284	27.599611
2	3	342	33.236152

Fig 7: *Output of the Cluster size table generated in Google colab*

Cluster Profiling & Visualization

After clustering, the characteristics of each group were analyzed:

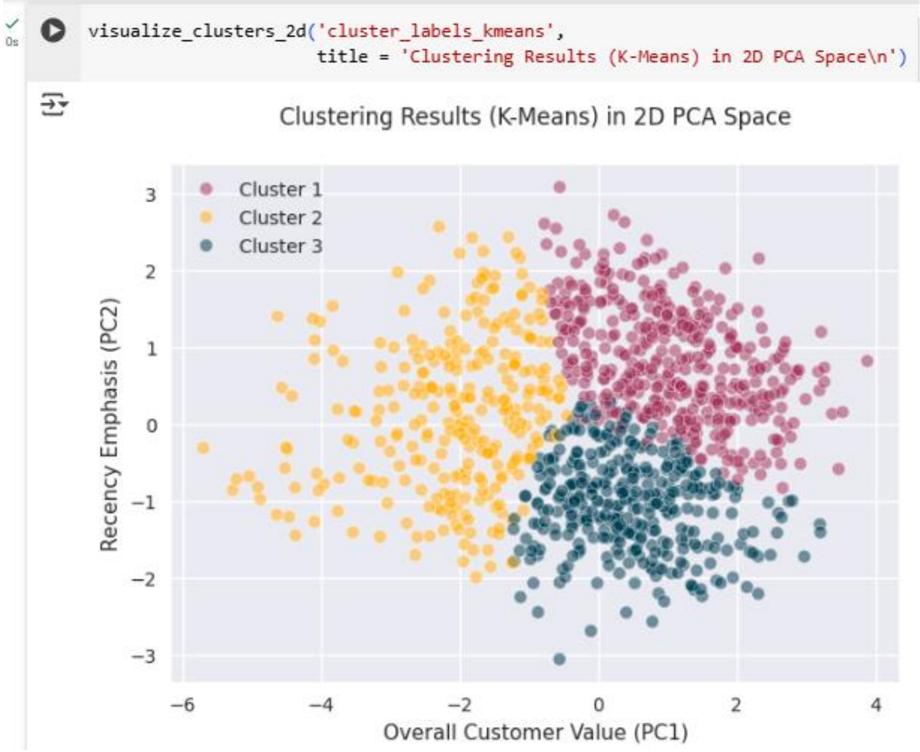
Cluster 1: High-Value Loyalists (Customers with consistently high value and likely strong brand engagement)

Cluster 2: Recent Opportunity Seekers (Lower overall value but high recent activity - potential for upselling or nurturing)

Cluster 3: At-Risk or Re-engageable (formerly Dormant/Disengaged) (Low value and long periods of inactivity - may need reactivation campaigns)

Visualizing these clusters in 2D spaces:

utilizing the respective principal components (PCs) derived from PCA (Principal Component Analysis).



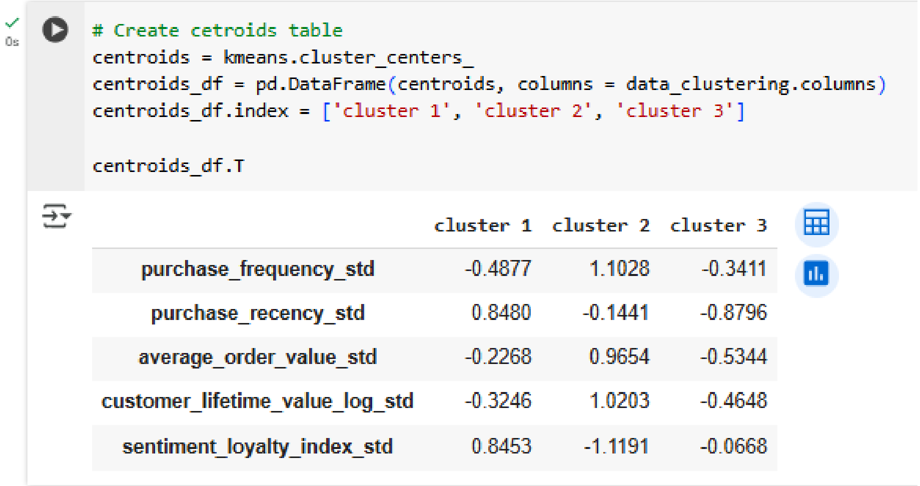
```
visualize_clusters_2d('cluster_labels_kmeans',
                      title = 'Clustering Results (K-Means) in 2D PCA Space\n')
```

0s
Clustering Results (K-Means) in 2D PCA Space
Cluster 1
Cluster 2
Cluster 3
Recency Emphasis (PC2)
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2
4
Overall Customer Value (PC1)

Fig 7: *Output of the Clustering results in 2D PCA space*

Examination of K-Means Customer Clusters:

The Centroids: To gain a deeper comprehension of the k-means clusters, examining the centroids of each cluster provides valuable insights. These centroids represent the mean position of data points within a cluster and their values enables us to identify distinctive features characterizing various customer segments.



```
# Create cetroids table
centroids = kmeans.cluster_centers_
centroids_df = pd.DataFrame(centroids, columns = data_clustering.columns)
centroids_df.index = ['cluster 1', 'cluster 2', 'cluster 3']

centroids_df.T
```

	cluster 1	cluster 2	cluster 3
purchase_frequency_std	-0.4877	1.1028	-0.3411
purchase_recency_std	0.8480	-0.1441	-0.8796
average_order_value_std	-0.2268	0.9654	-0.5344
customer_lifetime_value_log_std	-0.3246	1.0203	-0.4648
sentiment_loyalty_index_std	0.8453	-1.1191	-0.0668

Fig 8: *Creation of centroid table for each cluster in Google colab*

The centroids table provides a quantitative description of how each cluster differs from one another.

Summary of Implementation

This project successfully demonstrated how K-Means clustering can be implemented in a structured, step-by-step manner using accessible tools. From cleaning and preparing the data to building the model and extracting actionable insights, the process is both practical and replicable. Most importantly, it shows how machine learning can support real business decision-making—not just in theory, but in application.

Chapter 4: Results and Discussions

This section presents the results obtained from applying the **K-Means clustering algorithm** to customer data and discusses key insights derived from the segmentation process. The findings are supported by **graphs, tables, and visualizations**, which were shared in previous chapter, making it easier to interpret customer behavior and define actionable business strategies.

Using clustering techniques, we've been able to group customers into three clear segments, each showing its own unique behaviors and preferences. These clusters give us meaningful insights into how different types of customers engage with the business. With this understanding, we can create more personalized and effective strategies tailored to each group. Now, let's explore what defines each of these customer segments and how they differ from one another.

2. Summary of Clustering Results

After running the **K-Means algorithm (K=3)**, the customers were grouped into three well-defined clusters. Below is a summary of their characteristics:

We grouped high-value customers into Cluster 1. Among the lower-value customers, those who have shopped recently are placed in Cluster 2, while those who haven't interacted with the store for a while fall into Cluster 3.

These segments were then labeled based on observed behaviors.

Cluster 1: High-Value Loyalists (Customers with consistently high value and likely strong brand engagement)

- These customers exhibit a high frequency of purchases, indicating strong engagement with the store. They make purchases consistently and have a high average order value, contributing significantly to the overall revenue. Their loyalty sentiment is relatively low, possibly indicating they are focused more on transactional interactions.

Cluster 2: Recent Opportunity Seekers (Lower overall value but high recent activity - potential for upselling or nurturing)

- This group comprises customers who may not shop as frequently, but they have made recent purchases. Despite a lower purchase frequency, they display high loyalty sentiment, suggesting a strong connection with the brand. The higher purchase recency and sentiment loyalty index highlight their recent and loyal behavior.

Cluster 3: At-Risk or Re-engageable (formerly Dormant/Disengaged) (Low value and long periods of inactivity, may need reactivation campaigns)

- Customers in this cluster show lower purchase frequency, recency, and average order value, indicating less frequent and recent interactions. Their sentiment loyalty index is also low, suggesting a decreased engagement with the brand. This cluster may represent a segment that requires targeted efforts to re-engage and boost activity.

Key Insight: These clusters help categorize customers based on **spending behavior** rather than just demographic data, making marketing strategies **more personalized and data-driven**.

Marketing Strategies for these clusters:

Cluster 1: **High-Value Loyalists**

- Reward loyalty with exclusive perks (VIP programs, early access to sales).
- Upsell and cross-sell premium or complementary products.

- Send personalized thank-you emails or handwritten notes to deepen connection.
- Introduce referral programs to leverage their brand advocacy.

Cluster 2: **Recent Opportunity Seekers**

- Use time-sensitive offers or discounts to encourage repeat purchases.
- Deliver onboarding emails or product recommendations based on browsing/purchase history.
- Target with loyalty point incentives to build value perception and increase retention.

Cluster 3: **At-Risk or Re-engageable**

- Send win-back campaigns with strong hooks (e.g., "We miss you: here’s 25% off").
- Offer personalized incentives based on past purchases to regain interest.
- Use survey or feedback forms to understand reasons for disengagement.
- Place in low-cost awareness campaigns rather than high-intensity marketing funnels.

Discussion: How This Transforms Business Decision-Making

The application of **K-Means clustering** provided a structured way to analyze and categorize customers based on real purchasing behavior rather than generic demographics. The results highlighted several key advantages:

1.Improved Customer Targeting

Businesses can create **tailored campaigns** based on cluster-specific behaviors rather than mass marketing.

For example, Cluster 1 (High-Value Loyalists) should receive exclusive offers, while Cluster 1 (At-Risk or Re-engageable) might respond better to discount-driven promotions.

2. Budget Optimization in Marketing

Companies can allocate marketing budgets more efficiently, focusing high-investment strategies on high-value clusters.

Instead of treating all customers the same, businesses can now prioritize high-return segments.

3. Predicting Future Buying Behavior

If a customer’s behavior shifts between clusters, businesses can adjust engagement strategies dynamically.

Predicting when a “Recent Opportunity Seekers” (Cluster 2) might become a “High-Value Loyalist” (Cluster 1) allows for proactive engagement.

Key Takeaways from the Results

Final Thoughts:

- K-Means clustering successfully identified three meaningful customer segments based on income, spending score, and purchase frequency.
- The results show clear behavioral differences between customer groups, enabling businesses to design data-driven marketing strategies.

- The project demonstrated how clustering can be practically applied to real-world business scenarios, making it a valuable tool for customer analytics and decision-making.

Future Enhancements:

- The model could be expanded by incorporating additional behavioral features like purchase history, online browsing activity, or loyalty program data.
- Machine learning algorithms like DBSCAN or Hierarchical Clustering could be explored for deeper insights.

Last Chapter: Conclusion and Future Scope

Conclusion

This project set out with the objective of segmenting customers using a data-driven approach, specifically through the K-Means clustering algorithm. Rather than relying on traditional demographic-based segmentation, this approach focused on actual customer behavior: spending habits, income levels, and purchase frequency—to understand how different customer types interact with a business.

After conducting the analysis and carefully evaluating the results, three clear and actionable customer segments emerged:

Cluster 1: High-Value Loyalists (Customers with consistently high value and likely strong brand engagement)

Cluster 2: Recent Opportunity Seekers (Lower overall value but high recent activity - potential for upselling or nurturing)

Cluster 3: At-Risk or Re-engageable (formerly Dormant/Disengaged) (Low value and long periods of inactivity, may need reactivation campaigns)

These segments weren't defined by assumptions, but by **patterns discovered in the data itself**. The clustering results, supported by visualizations and metrics, confirmed that customer behavior is often more nuanced than what is captured by surface-level variables like age or gender.

What makes this conclusion valuable is its **direct applicability to real business needs**:

- It shows how businesses can move beyond one-size-fits-all marketing.

- It allows for smarter resource allocation—putting the right offer in front of the right customer.
- It demonstrates how data science can play a strategic role, even in marketing and customer service.

Overall, the project has successfully shown how K-Means clustering can redefine customer profiles in a way that’s more insightful, relevant, and practical for decision-making.

Future Scope

While the results were promising, the project also opens doors to several future opportunities and enhancements. Some of the key areas for expansion are:

1.Inclusion of More Behavioral Variables

In this project, clustering was based on a limited number of features (like income and spending score). Future versions can include additional variables such as:

- Purchase history
- Online browsing patterns
- Product categories preferred
- Time-of-day or seasonal buying behavior

This would lead to even more refined segmentation and a deeper understanding of customer intent.

2. Applying Advanced Clustering Techniques

While K-Means provided valuable insights, other clustering algorithms could be explored, such as:

- **DBSCAN (Density-Based Spatial Clustering)** for finding arbitrarily shaped clusters

- **Hierarchical Clustering** to visualize nested relationships between segments
- **Gaussian Mixture Models** to understand soft clustering (where a customer can belong to more than one segment)

These could address limitations of K-Means and offer new perspectives.

3. Real-Time Customer Segmentation

With the rise of AI-driven personalization, businesses increasingly need real-time segmentation. Future implementations can be integrated with CRM tools and e-commerce platforms to:

- Automatically update a customer’s segment based on latest behavior
- Trigger dynamic marketing responses (like personalized emails or push notifications)

This would move the segmentation from a **static model to a living, adaptive system**.

4.Business Strategy Integration

Going forward, the insights from this project could be directly linked to strategic business decisions, such as:

- **Product development** – create offerings that cater to each segment
- **Customer retention programs** – designed differently for loyalists vs. cautious buyers
- **Sales forecasting** – predict growth based on segment-wise performance

Final Thoughts

In a world where customers expect brands to “know them,” segmentation is no longer a luxury—it’s a **necessity**. This project has shown that even a simple algorithm like K-Means, when applied thoughtfully, can help businesses understand their customers **on a much deeper level**.

More importantly, it highlights a shift in how businesses can operate: **from assumption-based decision-making to truly data-driven strategy**. That’s the real impact of this work—and the biggest opportunity moving forward.

Annexure-I

Project Summary

Registration No: 323102744 **Name of Student:** Ragini Kumari

Title of Capstone Project/ Project Work: ClusterCraft: Redefining Customer Profiles

Objectives of the Project:

1. To segment customers based on behavioral attributes such as purchase frequency, spending patterns, and income levels.
2. To implement the K-Means clustering algorithm to identify hidden customer segments.
3. To visualize and interpret the clusters in a way that helps marketers develop personalized campaigns.
4. To evaluate the effectiveness of clustering in improving marketing decision-making.

Results and Findings:

The project successfully applied K-Means clustering to segment customers into three distinct groups based on income, spending habits, and purchase frequency. Each segment displayed clear behavioral patterns, offering deeper insight into customer needs. These insights support the development of personalized marketing strategies and informed business decisions, highlighting the practical relevance of clustering techniques in real-world customer analytics.

Specific outcomes of the Project:

The project resulted in a clean, structured dataset by addressing missing values, duplicates, and scaling key variables. K-Means clustering revealed three distinct customer segments with unique behavioral patterns. Through visual and statistical analysis, the project provided actionable insights into customer preferences, enabling targeted marketing strategies. Overall, it demonstrated the effectiveness of unsupervised learning in enhancing customer understanding and data-driven decision-making in e-commerce.

Any challenges/issues faced during the Project:

One of the key challenges in the project was dealing with missing or inconsistent data, which required careful cleaning to ensure accuracy. Identifying duplicates and transforming data-types without losing meaningful patterns was also critical. Selecting the optimal number of clusters for K-Means involved trial and error using methods like the Elbow Curve, which required interpretation. Additionally, balancing technical complexity with business relevance was essential to keep the insights actionable and understandable for non-technical stakeholders.

Signature of the Student: Ragini

References

References taken for Review of Literature:

- Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing, 21(1), 3–8.
- Kotler, P. (1984). Marketing Management. Prentice-Hall.
- Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations. Springer.
- Gan, G., Ma, C., & Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. SIAM.
- Tsiptsis, K., & Chorianopoulos, A. (2011). Data Mining Techniques in CRM: Inside Customer Segmentation. Wiley.
- Chaturvedi, S., et al. (2014). Application of K-means clustering in customer segmentation. International Journal of Computer Applications, 89(4), 20–25.
- Singh, A., & Rani, S. (2019). Comparative analysis of clustering techniques for customer segmentation. International Journal of Engineering and Advanced Technology, 8(6S3), 1076–1081.
- Ahmed, F., & Hussain, M. (2021). Consumer segmentation using data mining in e-commerce. Journal of Retail Analytics, 17(2), 44–52.
- Yadav, R., Singh, S., & Sharma, K. (2023). Adaptive segmentation in digital marketing: A clustering approach. Journal of Marketing Analytics, 11(1), 12–29.
- Jain, P., & Gupta, M. (2024). Integrating AI with clustering for intelligent customer segmentation. Journal of Business Intelligence Research, 15(1), 25–39.

Website references:

https://www.kaggle.com/

https://colab.research.google.com/

Final project is hosted and publicly viewable on GitHub:

https://github.com/analytixx/ClusterCraft-Customer-Segmentation/tree/main