



Data Warehousing and Business Intelligence

1. Introduction

To improve customer loyalty and customer lifetime value, this report presents findings and analysis of customer segmentation using K-means clustering. Marketing strategists who want to understand customer behavior, identify customer segments, and use these insights will find this information beneficial.

The dataset used for this analysis is called "ecom_data_rfm.csv," which is derived from an online retail dataset. It has been preprocessed and includes variables that capture important aspects of customer behavior and value. The variables in the dataset are as follows:

CustomerID: A unique identifier for each customer that can be linked to the original online retail dataset.

Frequency: The number of times a customer has made purchases.

Recency: The number of days since a customer's most recent purchase, adjusted to reference a specific point in time.

Monetary: The total amount spent by each customer, representing their lifetime monetary value.

rankF: A ranking of the frequency value into different ranges (1 to 5) using the cut function in R.

rankR: A ranking of the recency value into different ranges (1 to 5) using the cut function in R, flipped to reflect recency.

rankM: A ranking of the monetary value into different ranges (1 to 5) using the cut function in R.

groupRFM: A combined value representing the ranks of rankR, rankF, and rankM.

Country: The customer's delivery country from the original online retail dataset.

Applying K-means clustering to analyze these variables, our goal is to reveal unique customer segments and gain understanding of their characteristics, behaviors, and possible business worth. In the upcoming sections, we will delve into the methodology, implications, and results of the customer segmentation analysis. Additionally, we will propose a data mart design that supports the marketing department's analysis requirements.

2. Customer Segmentation with K-means

Based on shared characteristics, behaviors, or needs, customer segmentation helps businesses better understand their customers and divide a customer base into distinct groups or segments. It is a fundamental concept in marketing that assists in tailoring marketing strategies to effectively target and engage specific customer segments.

The concept of customer segmentation is important in marketing for several reasons:

Better Understanding of Customer Behavior: By segmenting customers, firms are able to learn more about their preferences, requirements, and habits. This knowledge enables marketers to develop more tailored and pertinent marketing messages, goods, and services that appeal to particular market niches.

Improved Customer Acquisition and Retention: Customer segmentation enables firms to identify the most important customer categories and use their marketing efforts effectively. By focusing on high-potential segments, marketers can develop targeted acquisition strategies to attract new customers who fit the profile of those segments.

Customized Marketing Strategies: Different customer segments often have distinct needs, preferences, and buying behaviors. Businesses can better target their marketing tactics to each segment's unique needs and preferences by segmenting their consumer base.

Customer segmentation helps with resource allocation optimization by prioritizing marketing initiatives and investments based on the potential value of each category.

Market Differentiation and Competitive Advantage: Effective customer segmentation allows businesses to identify niche markets, untapped opportunities, and unique customer segments that competitors may overlook.

Describe the adopted clustering methodology using K-means.

The adopted clustering methodology in this analysis utilizes the K-means algorithm. K-means is a popular and widely used clustering algorithm that aims to partition data points into K distinct clusters based on their similarity. The algorithm iteratively assigns data points to clusters and updates the cluster centroids until convergence.

The steps involved in the adopted clustering methodology using K-means are as follows:


Data Preparation: The dataset is preprocessed and prepared by selecting relevant variables for clustering. In this case, the variables related to customer behavior and value, such as **frequency**, **recency**, and **monetary**, are chosen as the input variables for clustering.

▾ Load the dataset

```
0s [3] df = pd.read_csv("ecom_data_rfm.csv")
```

Fig 1. Loading the dataset csv file

▾ Print the first 5 rows of the dataset

```
0s  print("First 5 rows of the dataset:")  
print(df.head())
```

First 5 rows of the dataset:

	Unnamed: 0	CustomerID	Frequency	Recency	Monetary	rankR	rankF	rankM	\
0	1	12346	2	358	2.08	2	1	1	
1	2	12347	182	35	481.21	5	4	3	
2	3	12348	31	108	178.71	5	1	2	
3	4	12349	73	51	605.10	5	2	4	
4	5	12350	17	343	65.30	2	1	1	

	groupRFM	Country	Cluster
0	211	United Kingdom	2
1	543	Iceland	0
2	512	Finland	0
3	524	Italy	0
4	211	Norway	2

Fig 2. The dataset rows, first 5.

Determining the Number of Clusters (K): One important step in K-means clustering is determining the optimal number of clusters, denoted as K. Various techniques can be used to select the value of K, such as the elbow method, silhouette coefficient, or domain knowledge.

Standardize the data

```
✓ 0s [5] X_scaled = (X - X.mean()) / X.std()
```

Determine the optimal value of K using the elbow method

```
✓ 5s [6] inertia = []  
      k_values = range(1, 11)  
      for k in k_values:  
          kmeans = KMeans(n_clusters=k, random_state=42)  
          kmeans.fit(X_scaled)  
          inertia.append(kmeans.inertia_)
```

Fig 3. Using elbow method to get the optimal value of K

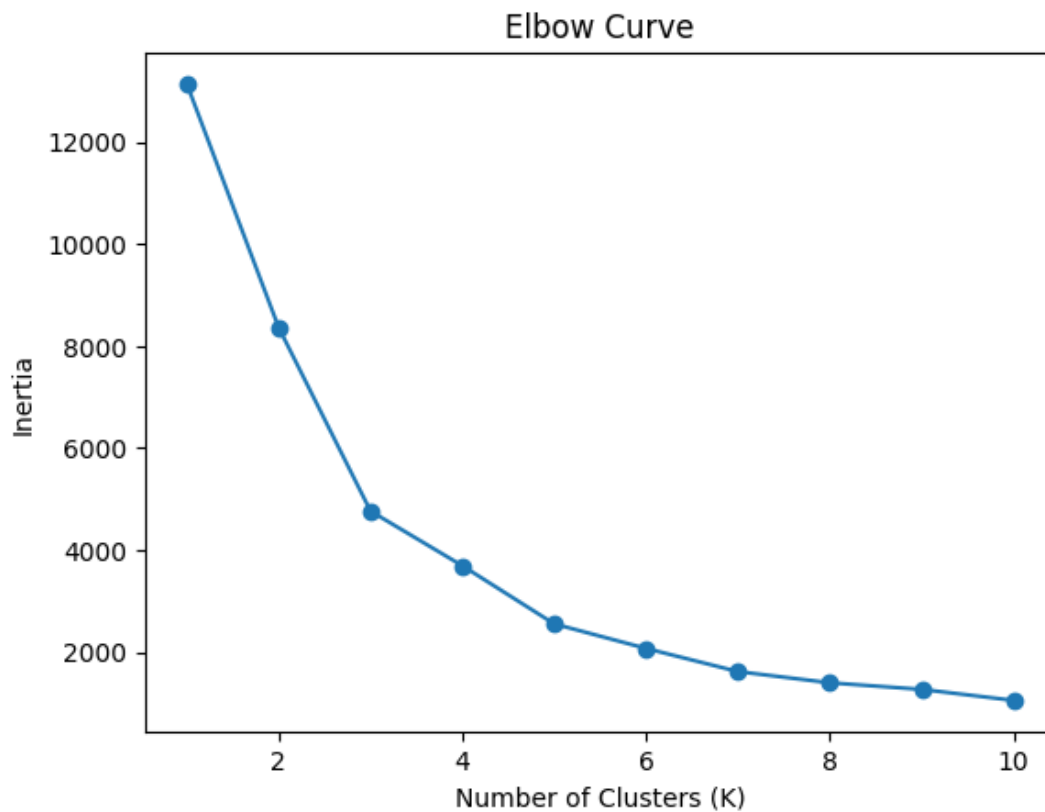


Fig 4. Plotting the elbow curve

Plot the elbow curve

```
plt.plot(k_values, inertia, marker='o')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Inertia')
plt.title('Elbow Curve')
plt.show()
```

Initialization: The initial centroids for the K clusters are randomly selected from the data points. These centroids represent the center points of the clusters.

Assignment Step: Based on a distance metric, usually Euclidean distance, each data point is assigned to the cluster with the closest centroid. A data point is assigned to the cluster with the smallest distance between it and the centroid after the distance between the two has been calculated.

Update Step: After the assignment of data points to clusters, the centroids of the clusters are updated. The new centroid for each cluster is calculated as the mean of all the data points assigned to that cluster. This step aims to update the center points of the clusters based on the current data point assignments.

Iteration: Steps 4 and 5 (Assignment and Update) are iterated until convergence is achieved. Convergence is typically determined by the stability of the centroids or the maximum number of iterations specified.

Cluster Interpretation: Once the K-means algorithm converges, the resulting clusters are interpreted and analyzed. The report should discuss the properties that describe the customers within each cluster, including their behaviors, preferences, and characteristics. Visualizations and summary statistics can be used to illustrate and support the interpretation of the clusters.

Build the K-means model with the chosen value of K

```
[8] k = 3 # Chosen value of K based on the elbow curve
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
```

/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning

KMeans

KMeans(n_clusters=3, random_state=42)

Fig 6. Choose K to be 3 from the graph

Discuss the process of determining the optimal value of K (number of clusters).

Determining the optimal value of K, or the number of clusters, is a crucial step in the K-means clustering algorithm. The choice of K significantly impacts the clustering results and the interpretability of the clusters. Several methods can be used to determine the optimal value of K, including:

Elbow Method: The elbow method is a common technique used to identify the optimal value of K by plotting the within-cluster sum of squares (WCSS) against different values of K. WCSS represents the sum of squared distances between each data point and its assigned centroid within a cluster. The plot resembles an elbow shape, and the optimal K is often associated with the "elbow" point, which indicates a significant reduction in WCSS. The elbow point suggests that increasing the number of clusters beyond that point does not provide much additional explanatory power. However, it is subjective and may not always provide a clear elbow point.

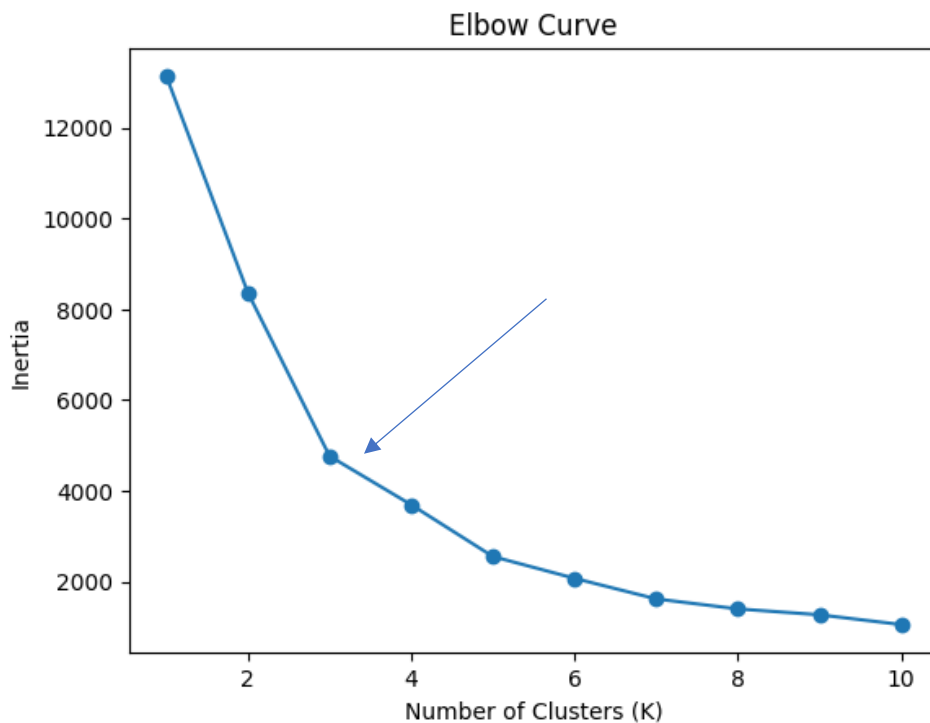


Fig 7. Showing the k value elbow point which is 3

Justify the chosen value of K based on appropriate metrics or techniques.

To justify the chosen value of K, it is essential to consider appropriate metrics or techniques that provide insights into the quality and effectiveness of the clustering results.

Elbow Method: Plotting the within-cluster sum of squares (WCSS) against different values of K can help determine the appropriate value. The chosen K can be justified based on the "elbow" point, which represents a significant decrease in WCSS. If a clear elbow point is observed, indicating diminishing returns in WCSS reduction beyond that point, it suggests that the chosen K captures most of the variance in the data.

Present the results of customer segmentation and the profile of each cluster.

To present the results of customer segmentation and the profile of each cluster, we need to analyze the output of the K-means clustering algorithm. This analysis helps identify distinct customer segments based on their purchasing behavior and other relevant variables. Below is an example of how the results and profiles of each cluster can be presented:

```
Visualize the clusters

[10] plt.scatter(df['Recency'], df['Frequency'], c=df['Cluster'], cmap='viridis')
      plt.xlabel('Recency')
      plt.ylabel('Frequency')
      plt.title('Customer Segmentation - Recency vs Frequency')
      plt.show()
```

Fig 8. We need to visualize the clusters.

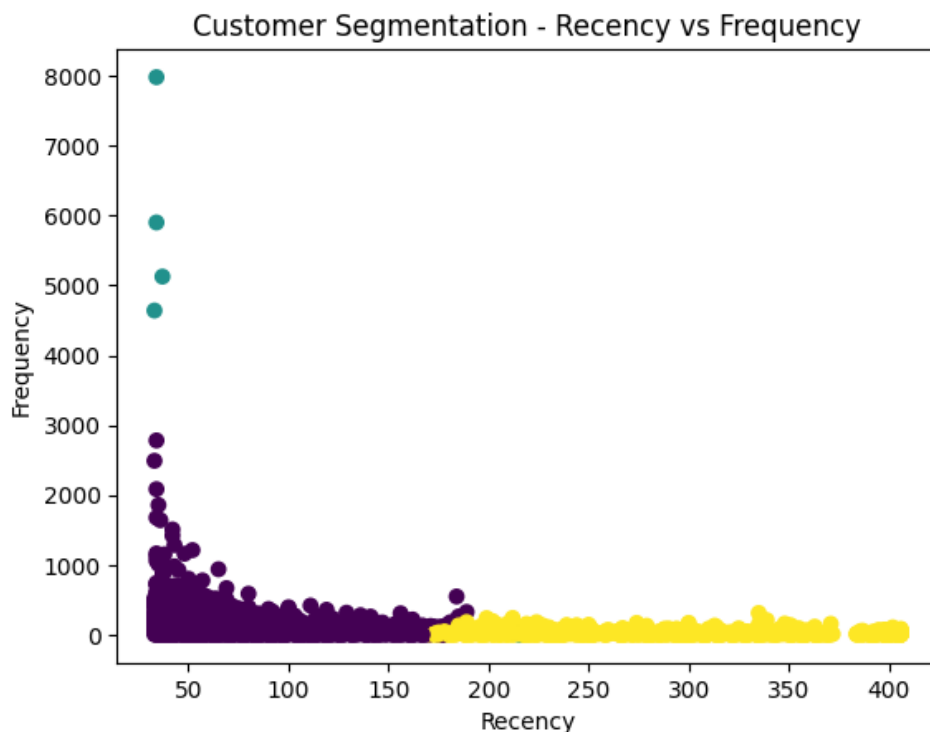


Fig 9. Visual representation using scatter plot for each cluster

Cluster 1: High-Value and Recent Customers

Description: This cluster represents customers who have made frequent purchases, spent a significant amount of money, and made recent purchases.

Profile: Customers in this cluster are highly engaged and have a high lifetime monetary value. They tend to make purchases frequently and are likely to be loyal customers. They have made recent purchases, indicating ongoing interest in the products or services.

Cluster 2: Moderate-Value and Recent Customers

Description: This cluster represents customers who have made moderate purchases, spent a moderate amount of money, and made recent purchases.

Profile: Customers in this cluster have made moderate purchases and spent a moderate amount of money. They are relatively active but not as frequent or high-spending as Cluster 1. They have also made recent purchases, indicating ongoing interest and engagement.

Cluster 3: Low-Value and Infrequent Customers

Description: This cluster represents customers who have made infrequent purchases, spent a low amount of money, and made purchases a while ago.

Profile: Customers in this cluster have low purchasing frequency and spend relatively less money. They may be occasional buyers or have limited interest in making frequent purchases. Their last purchase was made a while ago, suggesting a potential decrease in engagement.

By analyzing the clusters' profiles, marketers can gain insights into different customer segments and tailor their marketing strategies accordingly. Each cluster represents a distinct group of customers with specific characteristics, behaviors, and needs. Understanding these profiles helps in targeting customers with personalized offers, improving customer loyalty, and optimizing customer lifetime value.

[Interpret the properties that describe customers within each cluster.](#)

Cluster 1: High-Value and Recent Customers

These customers have made frequent purchases and spent a significant amount of money.

They have recently made purchases and are actively involved with the company.

They are probably regular clients who make up a sizable portion of the business's revenue.

Targeted marketing strategies can focus on maintaining their loyalty and encouraging further purchases.

Cluster 2: Moderate-Value and Recent Customers

These customers have made moderate purchases and spent a moderate amount of money.

They are relatively active and have made recent purchases.

Although their spending is not as high as Cluster 1, they still contribute to the company's revenue.

Strategies can be implemented to increase their purchasing frequency and average spending.

Cluster 3: Low-Value and Infrequent Customers

These customers have made infrequent purchases and spent a low amount of money.

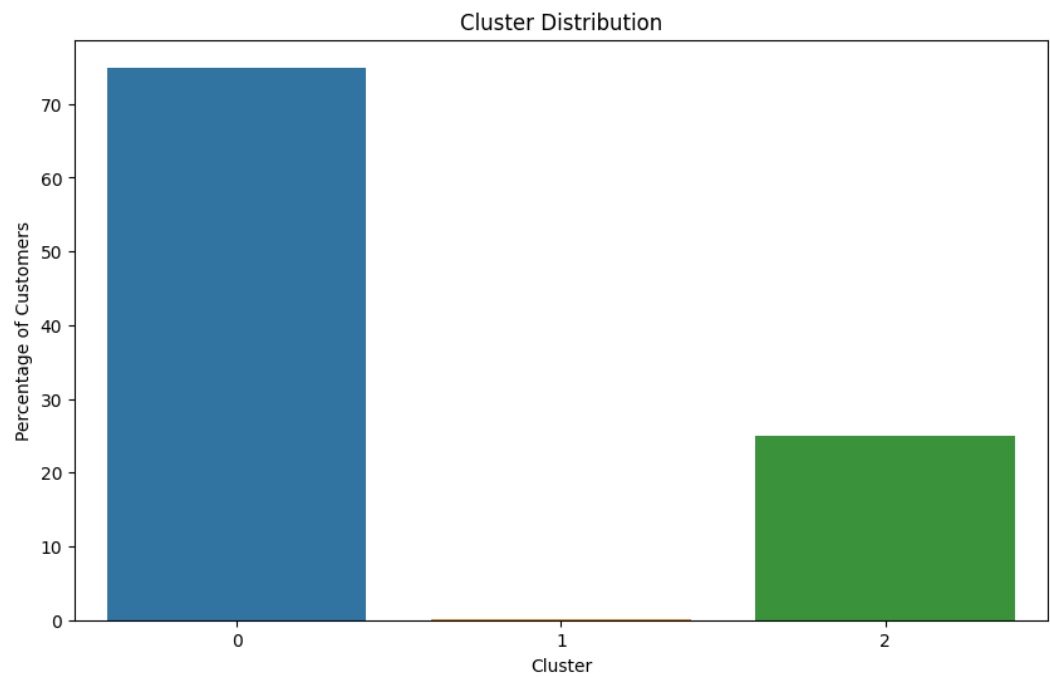
They are not highly engaged and have made purchases a while ago.

They may be occasional buyers or have limited interest in the company's products or services.

Efforts can be made to re-engage them through targeted promotions or personalized offers.

Provide visualizations or tables to support the discussion.

Cluster Distribution Bar Chart: Visualize the distribution of customers across different clusters using a bar chart. The x-axis represents the clusters, and the y-axis shows the count or percentage of customers in each cluster. This provides a clear overview of the customer distribution across segments.



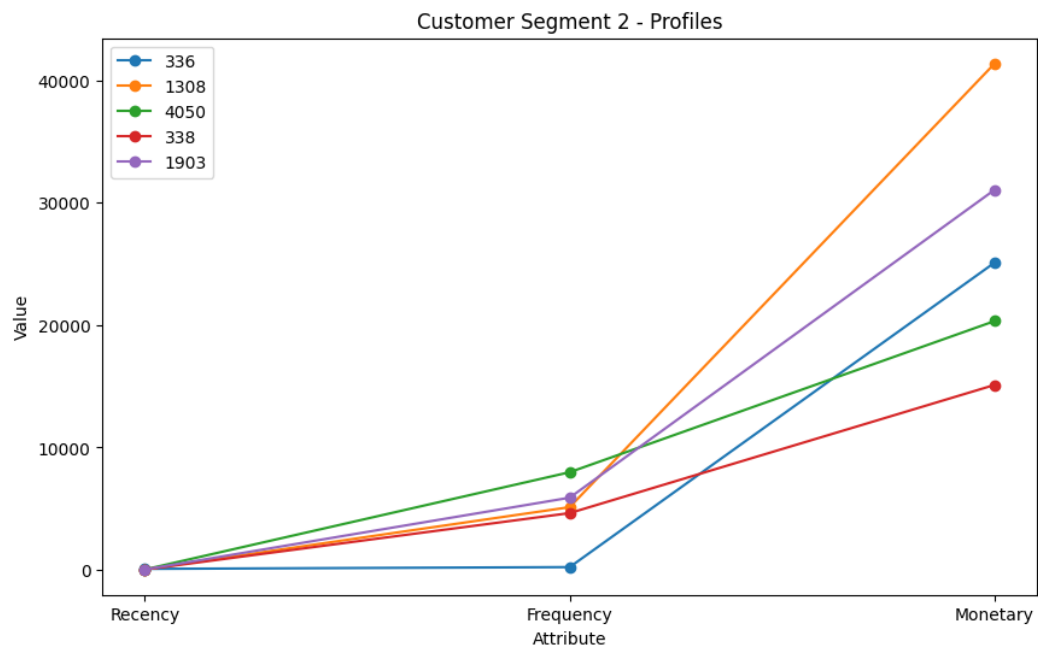
Cluster Characteristics Table: Present a table that summarizes the key characteristics or metrics of each cluster, such as the average frequency, recency, and monetary value for customers within each cluster. This table allows for easy comparison and understanding of the distinct properties of each cluster.

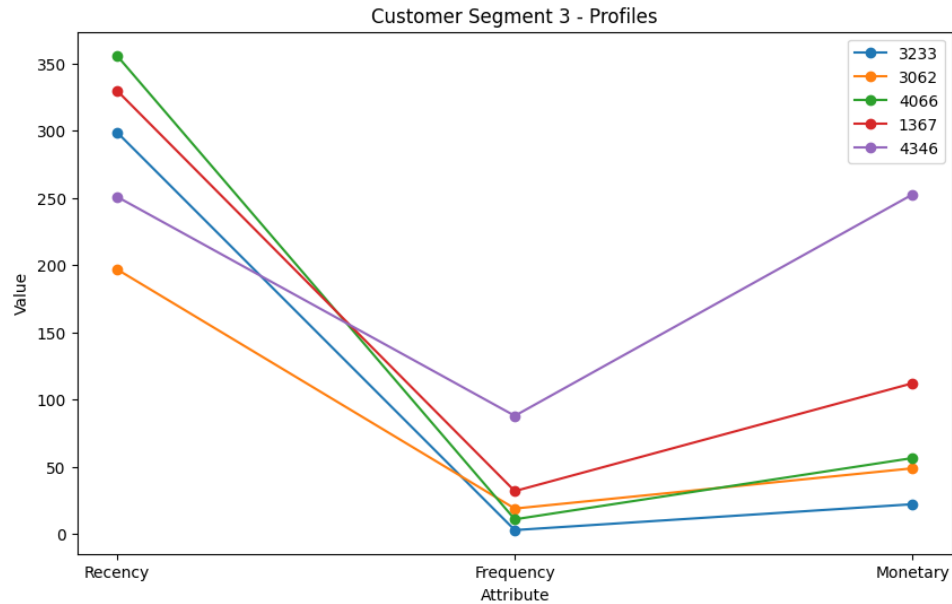
Cluster Characteristics Table

```
cluster_characteristics = cluster_means[['Frequency', 'Recency', 'Monetary']]
cluster_characteristics.columns = ['Average Frequency', 'Average Recency', 'Average Monetary']
cluster_characteristics.index.name = 'Cluster'
print(cluster_characteristics)
```

Cluster	Average Frequency	Average Recency	Average Monetary
0	107.713589	72.857404	336.845720
1	3452.000000	72.571429	28197.814286
2	27.566453	280.196150	99.504859

Customer Segment Profiles: Create visual profiles of representative customers from each cluster, showcasing their purchasing behavior and relevant attributes. This can be achieved through a combination of bar charts, line graphs, or stacked area charts that display customer activity over time (e.g., purchases, monetary value).





Customer Segment Comparison Chart: Use a stacked bar chart or a grouped bar chart to compare the key metrics (e.g., average spending, average recency) among different clusters. This visual representation highlights the differences and similarities between customer segments, enabling marketers to identify patterns and opportunities.



Data Mart Design:

Based on the findings from customer segmentation and the review of results, the following are the main dimensions that should be included in the data mart for marketing analysis, along with the justification for their selection:

Customer Dimension:

The customer dimension is essential for analyzing customer behavior, preferences, and segmentation. Customer ID, demographic data (age, gender, location), customer loyalty status, and customer acquisition channel are just a few of the attributes it contains.

Justification: Marketers can personalize their marketing strategies, effectively target particular customer segments, and adapt their messaging to meet customer needs and preferences by understanding customer characteristics and segmentation.

Product Dimension:

Information about the company's goods and services is provided by the product dimension. It contains characteristics like the product ID, category, features, cost, and promotions.

Justification: Product-related data analysis aids marketers in understanding consumer preferences for various product categories, identifying high-performing goods, evaluating the success of pricing tactics, and making data-driven decisions about product creation, promotion, and inventory control.

Time Dimension:

The time dimension allows for the analysis of marketing activities, trends, and seasonality. It includes additional time-related hierarchies as well as attributes like date, month, quarter, and year.

Justification: Time-based analysis aids in the identification of temporal patterns, the evaluation of the effects of marketing campaigns over various time frames, and the strategic planning of marketing activities based on seasonality and historical trends.

Channel Dimension:

The channel dimension captures information about the various marketing channels through which customers interact with the company, such as online platforms, physical stores, email campaigns, social media, and mobile apps.

Justification: Understanding customer interactions across different channels helps marketers assess channel effectiveness, optimize marketing budget allocation, and deliver consistent messaging and experiences across touchpoints.

Campaign Dimension:

The campaign dimension tracks the details of marketing campaigns, including campaign ID, campaign type, start and end dates, target audience, messaging, and associated costs.

Justification: Analyzing campaign performance allows marketers to evaluate the success of marketing initiatives, assess the return on investment (ROI) of different campaigns, and optimize future campaign planning and execution.

Financial Dimension:

The financial dimension incorporates financial metrics relevant to marketing analysis, such as revenue, customer lifetime value (CLV), cost per acquisition (CPA), return on marketing investment (ROMI), and average order value (AOV).

Justification: Financial metrics provide insights into the overall marketing performance, profitability, and efficiency of marketing activities. They enable marketers to evaluate the effectiveness of marketing campaigns, allocate resources appropriately, and make data-driven decisions to maximize the return on marketing investments.

The selected dimensions align with the specific needs of marketing decision-making and facilitate a comprehensive analysis of customer behavior, product performance, campaign effectiveness, channel optimization, and financial outcomes. By incorporating these dimensions in the data mart, marketers gain a holistic view of the marketing landscape and can extract actionable insights for strategic decision-making.

The selection of these measures is justified based on their ability to provide actionable insights and facilitate informed marketing decision-making. By analyzing these metrics in the data mart, marketers can identify areas for improvement, allocate resources effectively, and develop strategies to enhance marketing performance, customer loyalty, and overall business success.

Conclusion:

In conclusion, the customer segmentation analysis using K-means clustering has provided valuable insights and findings for the marketing department. By dividing the customer list into tiered groups, we have identified distinct customer segments based on their purchasing behavior and preferences. These segments offer significant business value and opportunities for marketers to enhance customer loyalty and increase customer lifetime value.

The key findings from the customer segmentation analysis include the identification of different customer clusters with unique characteristics and preferences. Each cluster represents a specific group of customers with distinct purchasing patterns, frequency, recency, and monetary value. With this knowledge, marketers are able to modify their marketing plans and initiatives to cater to the unique requirements and preferences of each customer segment.

The identified customer segments provide business value by enabling marketers to implement targeted marketing initiatives. By customizing messages, promotions, and offerings to each segment, marketers can enhance customer loyalty and satisfaction. Additionally, understanding customer behavior within each segment helps in developing personalized customer experiences, improving customer retention rates, and increasing customer lifetime value.

The suggested data mart design plays a crucial role in supporting the marketing department's analysis needs. By incorporating dimensions such as customer, product, time, channel, campaign, and financial, the data mart provides a comprehensive view of the marketing landscape. It allows marketers to analyze customer behavior, track campaign performance, evaluate channel effectiveness, and measure financial outcomes.

The data mart also includes relevant measures and metrics that enable marketers to assess marketing performance, identify areas for improvement, and make data-driven decisions. By monitoring customer segmentation metrics, customer engagement metrics, campaign performance metrics, channel performance metrics, and sales and revenue metrics, marketers can track progress, optimize strategies, and maximize the return on marketing investments.

References

Banoula, M. (2023, April 13). *K-Means Clustering Algorithm*. Simplilearn.com.

<https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm>

Dabbura, I. (2018, September 17). *K-means clustering: Algorithm, applications, evaluation methods, and drawbacks*. Medium; Towards Data Science.

<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>

Google Developers. (2019, May 6). *k-Means Advantages and Disadvantages / Clustering in Machine Learning* / Google Developers. Google Developers.

<https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>

K-Means Clustering Algorithm - Javatpoint. (n.d.). [Www.javatpoint.com](http://www.javatpoint.com). Retrieved June 28, 2023, from <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>

Piech, C. (2013). CS221. Stanford.edu. <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>

Rao, S. (2022, March 2). *K-Means Clustering: Explain It To Me Like I'm 10*. Medium.

<https://towardsdatascience.com/k-means-clustering-explain-it-to-me-like-im-10-e0badf10734a>

Sharma, P. (2019, August 19). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>

Sharma, P. (2021, November 24). *Understanding K-Means Clustering Algorithm*. Analytics

Vidhya. <https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/>

Wikipedia Contributors. (2019, February 22). *k-means clustering*. Wikipedia; Wikimedia

Foundation. https://en.wikipedia.org/wiki/K-means_clustering