A Capstone Project Report on

Customer Segmentation USING Machine Learning

Submitted by
KARTHIKEYAN K, AIDS,
Chennai Institute Of Technology.

Table of contents

CHAPTER	TITLE	PAGE NO
1.	ABSTRACT	3
2.	ACKNOWLEDGMENT	4
3.	INTRODUCTION	5
4.	LITERATURE SURVEY	6
4.1	EXISTING PROBLEM	
4.2	PROPOSED SOLUTION	
5.	THEORETICAL ANALYSIS	8
5.1	BLOCK DIAGRAM	
5.2	HARDWARE/SOFTWARE	
6.	EXPERIMENTAL INVESTIGATION	9
7.	FLOWCHART	10
8.	DATASET INFORMATION	11
9.	DATA VISUALIZTION	13
10.	MODEL BUILDING	18
11.	FLASK DEPLOYMENT	22
12.	ADVANTAGES AND DISADVANTAGES	23
13.	CONCLUSION	26

ABSTRACT

The project focuses on customer segmentation using RFM analysis and K-means clustering to group customers based on their purchasing behavior. By analyzing the Recency, Frequency, and Monetary values, distinct customer segments are identified, enabling businesses to tailor their marketing strategies for improved customer engagement and retention. This documentation provides a comprehensive overview of the project, its objectives, and the methodologies employed. The literature survey highlights the limitations of traditional segmentation approaches and proposes the use of RFM analysis and K-means clustering as a solution. Theoretical analysis outlines the key concepts of RFM and K-means, followed by experimental investigation on the dataset. Data visualization showcases insights derived from customer behavior data. Model building and Django deployment details the creation of the segmentation model and its integration into a web application. The advantages and disadvantages of the proposed approach are discussed, concluding with valuable insights for businesses aiming to optimize customer management strategies.

DATASET LINK :

https://github.com/Karthik-1497/Customer-Segmentation/blob/main/data.csv

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Dr Prema Latha as my mentor. He has readily shared his immense knowledge in data analytics and guide me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: Name:

CHAPTER 1

INTRODUCTION

In the fast-paced and fiercely competitive world of business, understanding customer behavior is paramount for companies aiming to thrive and succeed. Customer segmentation, a crucial marketing technique, plays a pivotal role in this endeavor by dividing customers into distinct groups based on their shared characteristics. These groups, or segments, empower businesses to customize marketing strategies, optimize resource allocation, and enhance overall customer satisfaction.

The primary objective of this project is to perform customer segmentation using RFM analysis and K-means clustering. RFM analysis focuses on three vital customer metrics: Recency, Frequency, and Monetary value, providing valuable insights into customer engagement, loyalty, and spending patterns. Through the application of K-means clustering to the RFM metrics, customers are grouped into coherent segments based on their similarities, offering deeper insights into their preferences and behaviors.

The importance of customer segmentation cannot be overstated, as it enables businesses to identify high-value customers for targeted promotions, re-engage dormant customers, and prioritize efforts to retain the most profitable segments. Additionally, segmentation facilitates tailored product offerings and communications, leading to enhanced customer experiences.

This documentation provides a comprehensive journey through the project, from problem statement and literature survey to the development of the customer segmentation model using RFM analysis and K-means clustering. It explores the theoretical foundation of these techniques, presents practical demonstrations of data visualization and model building, and discusses the advantages and limitations of the proposed approach. As a result, businesses can leverage this knowledge to make data-driven decisions, harness customer insights, and unlock new growth opportunities in today's ever-evolving market

CHAPTER 2

Un-Supervised Learning

Unsupervised learning is a category of machine learning algorithms in which the model is trained on unlabeled data, meaning the data does not have explicit target labels or class annotations. Unlike supervised learning, where the model learns from labeled examples and is provided with correct answers during training, unsupervised learning operates without explicit guidance and seeks to find patterns or structures within the data.

The primary goal of unsupervised learning is to identify inherent relationships, similarities, or groupings in the data without knowing the actual outcomes in advance. It is commonly used for tasks such as clustering, dimensionality reduction, and anomaly detection.

Clustering is a significant application of unsupervised learning, where the algorithm groups similar data points together into distinct clusters based on their feature similarities. These clusters represent different segments within the data and can be useful for customer segmentation, market analysis, and recommendation systems.

Dimensionality reduction is another crucial aspect of unsupervised learning, which aims to reduce the number of features in the data while preserving important information. Techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are commonly used for this purpose, helping to visualize high-dimensional data and simplify subsequent analyses.

Anomaly detection, also known as outlier detection, is yet another important application of unsupervised learning. It involves identifying data points that deviate significantly from the majority, which can be valuable for fraud detection, fault diagnosis, or identifying rare events.

Overall, unsupervised learning plays a fundamental role in exploring and understanding data without the need for labeled examples, making it a powerful tool for discovering hidden patterns and insights in diverse real-world datasets.

.

CHAPTER 3

LITERATURE SURVEY

3.1 Existing Problem:

The existing problem in customer segmentation lies in the limitations of traditional methods, which often rely on manual and subjective criteria. These methods may lack the ability to leverage the full potential of data-driven insights, leading to inconsistent and less effective results. Additionally, some conventional approaches may overlook the temporal aspect of customer behavior, failing to consider the changing preferences and needs of customers over time. This can result in missed opportunities for businesses to address individual customer requirements and preferences, potentially leading to decreased customer loyalty and revenue.

3.2 Proposed Solution:

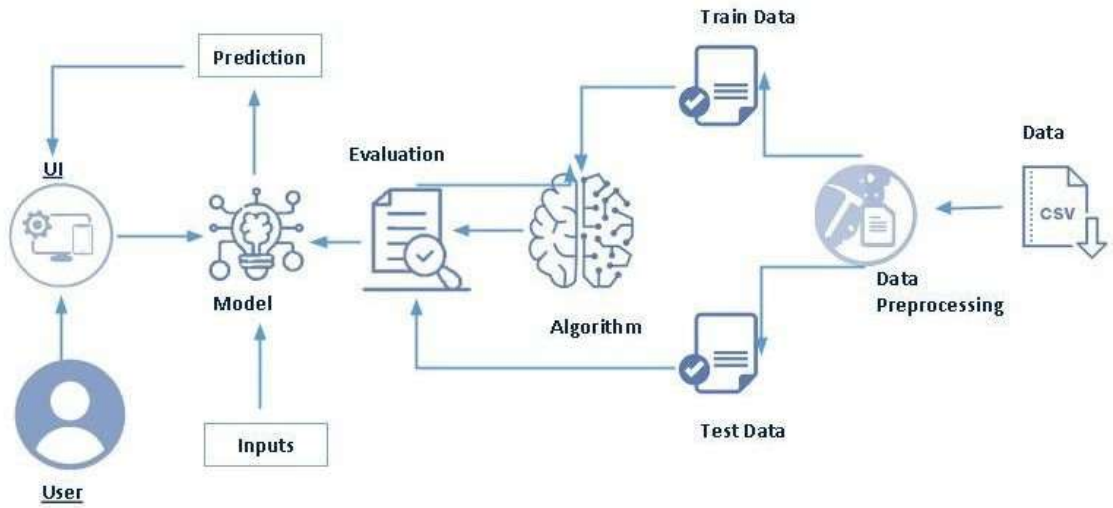
To address the limitations of traditional customer segmentation methods, we propose the combined use of RFM analysis and K-means clustering as a more effective and data-driven solution. RFM analysis focuses on three key metrics - Recency, Frequency, and Monetary value - to assess customer behavior and engagement. By analyzing these metrics, we gain valuable insights into the customers' recent transaction history, their purchase frequency, and the monetary value of their transactions.

Subsequently, we apply K-means clustering, an unsupervised learning technique, to group customers with similar RFM characteristics into distinct segments. This clustering process allows us to identify homogenous customer groups based on their shared attributes and behavior patterns. As a result, businesses can better understand the different customer segments and devise tailored marketing strategies to address their specific needs and preferences.

By combining RFM analysis with K-means clustering, our proposed solution provides a more automated and data-centric approach to customer segmentation. It enables businesses to make well-informed decisions based on objective and data-driven insights, leading to improved customer targeting, enhanced customer satisfaction, and ultimately, increased business success. Moreover, the proposed solution offers scalability and adaptability, allowing it to handle large datasets and accommodate future changes in customer behavior more effectively

CHAPTER 4
THEORETICAL ANALYSIS

4.1 BLOCK DIAGRAM



4.2 Hardware / Software designing

HARWARE	1. COMPUTER SYSTEM 2. INTERNET CONECTIVITY
SOFTWARE	1. VS CODE 2. FLASK 3. WORD 4. DATASET MANAGEMENT 5. PYTHON LANGUAGE AND LIBRARIES

CHAPTER 5

EXPERIMENTAL INVESTIGATIONS

Data Collection: Gather online shopping data from various sources, such as e-commerce websites, APIs, or web scraping techniques. Collect data on browsing patterns, product categories viewed, previous purchase history, and demographic information of users.

Data Preprocessing: Clean the collected data by removing duplicates, handling missing values, and correcting inconsistencies. Encode categorical variables using techniques like one-hot encoding or label encoding. Normalize numerical features to ensure they are on a similar scale.

Feature Selection: Conduct exploratory data analysis to gain insights into the collected data. Use statistical techniques or feature importance methods (e.g., correlation analysis, information gain, or L1 regularization) to identify the most significant features. Select a subset of features that are highly correlated with the target variable (customer behaviour) and remove irrelevant or redundant features.

Data Splitting: Split the pre-processed data into training and testing datasets . Allocate a certain percentage of the data for training the models and the remaining portion for evaluating their performance.

Model Training: Apply classification algorithms such as Logistic Regression, Random Forest, and K-Means clustering to train predictive models. Configure the models with appropriate parameters and hyperparameters. Train each model on the training dataset using the selected features.

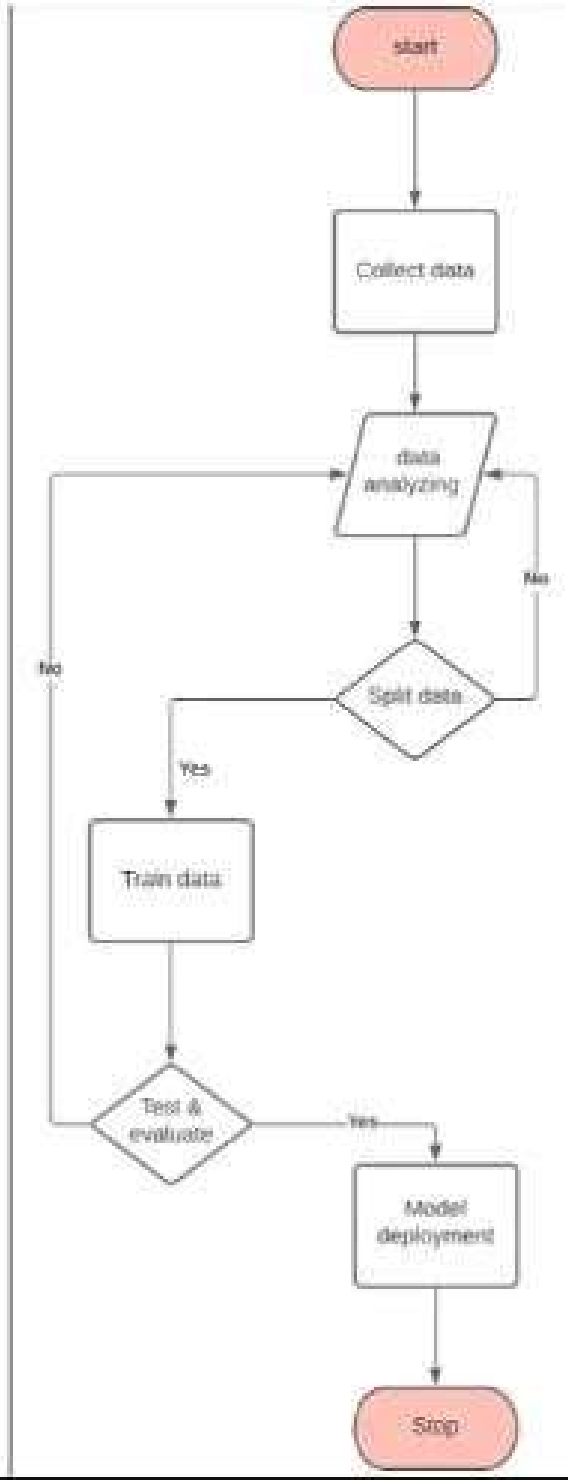
Model Evaluation: Evaluate the performance of each model using suitable evaluation metrics such as accuracy, precision, recall, and F1-score. Compare the performance of different models to identify the best-performing one. Assess the models' ability to predict customer behaviour during online shopping.

Model Selection and Saving: Select the best-performing model based on the evaluation results. Saving and selecting the model in the Joblib format for future use.

The above investigation provides a foundation for your project, laying the groundwork for subsequent steps such as setting up a Flask application, creating a user interface, handling user requests, and making predictions using the trained model.

CHAPTER 6

FLOWCHART



CHAPTER 7

DATASET INFORMATION

Attributes/Features:

1. InvoiceNo: A unique identifier for each transaction, helping to track and differentiate individual purchases.

2. StockCode: Represents the code or identifier of the product purchased, aiding in inventory management and order fulfillment.

3. Description: Provides a brief description of the purchased product, assisting in understanding the nature of the items sold.

4. Quantity: Indicates the number of units of a product bought in each transaction, crucial for analyzing purchase volumes and demand patterns.

5. InvoiceDate: Records the date and time of each transaction, enabling temporal analysis to identify trends and seasonal patterns.

6. UnitPrice: Represents the price of a single unit of the product, crucial for calculating total revenue and understanding pricing strategies.

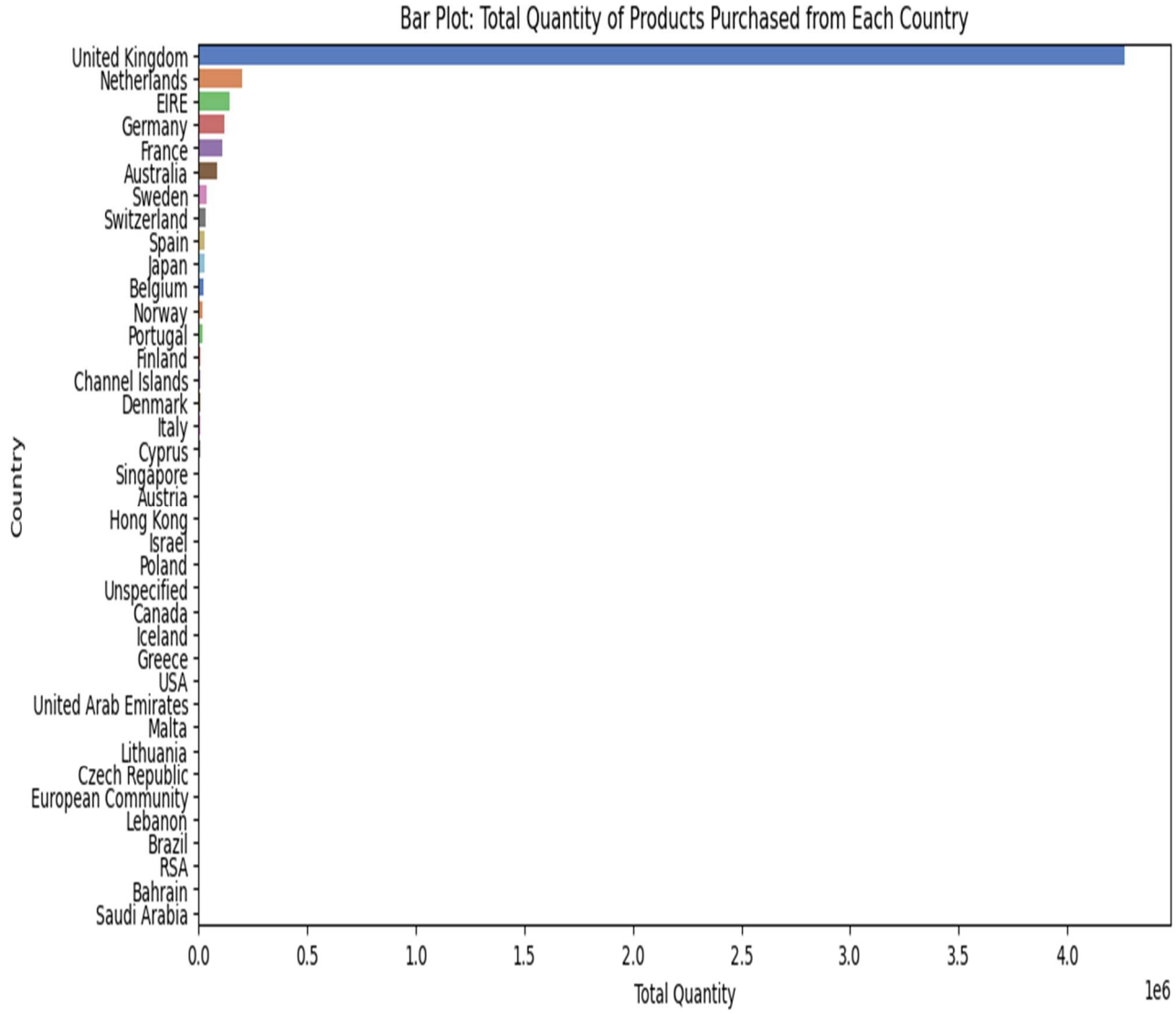
7. CustomerID: A unique identifier for each customer, allowing the grouping of transactions by individual buyers for customer-centric analysis.

8. Country: Records the country where the transaction occurred, essential for analyzing international sales and regional preferences.

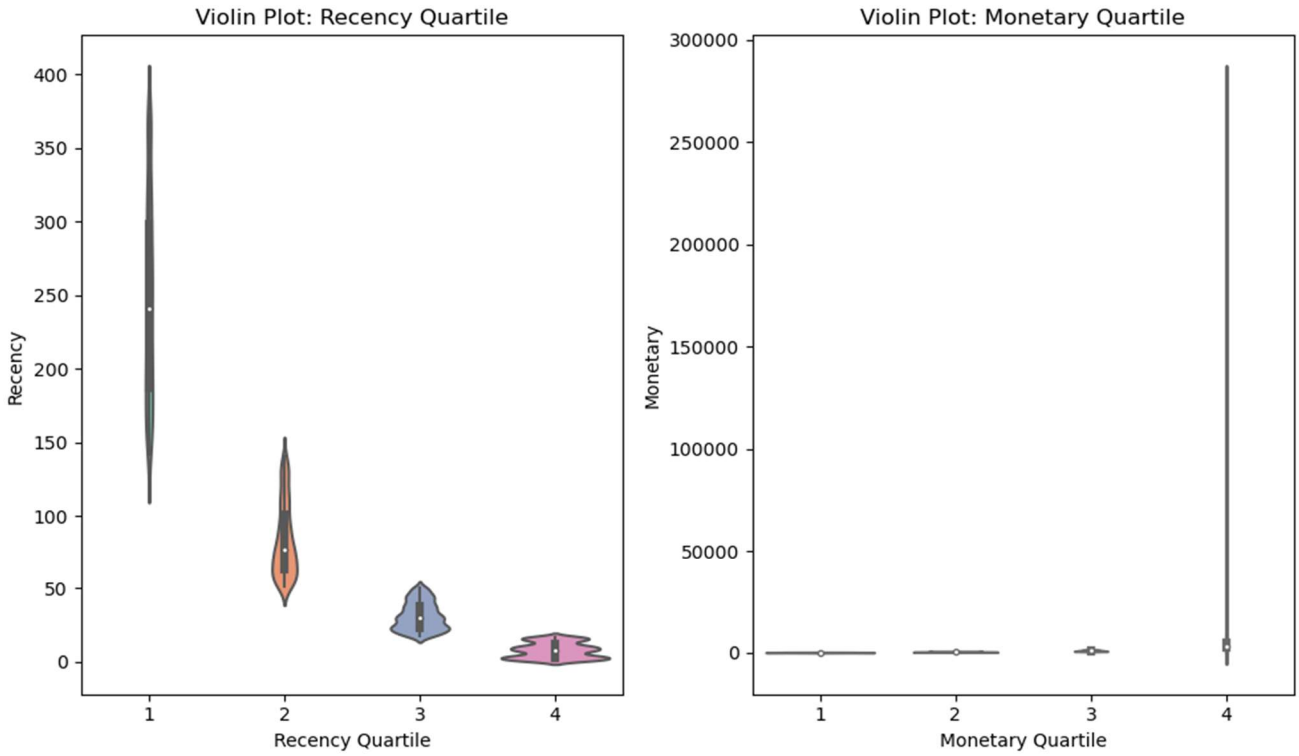
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

CHAPTER 8

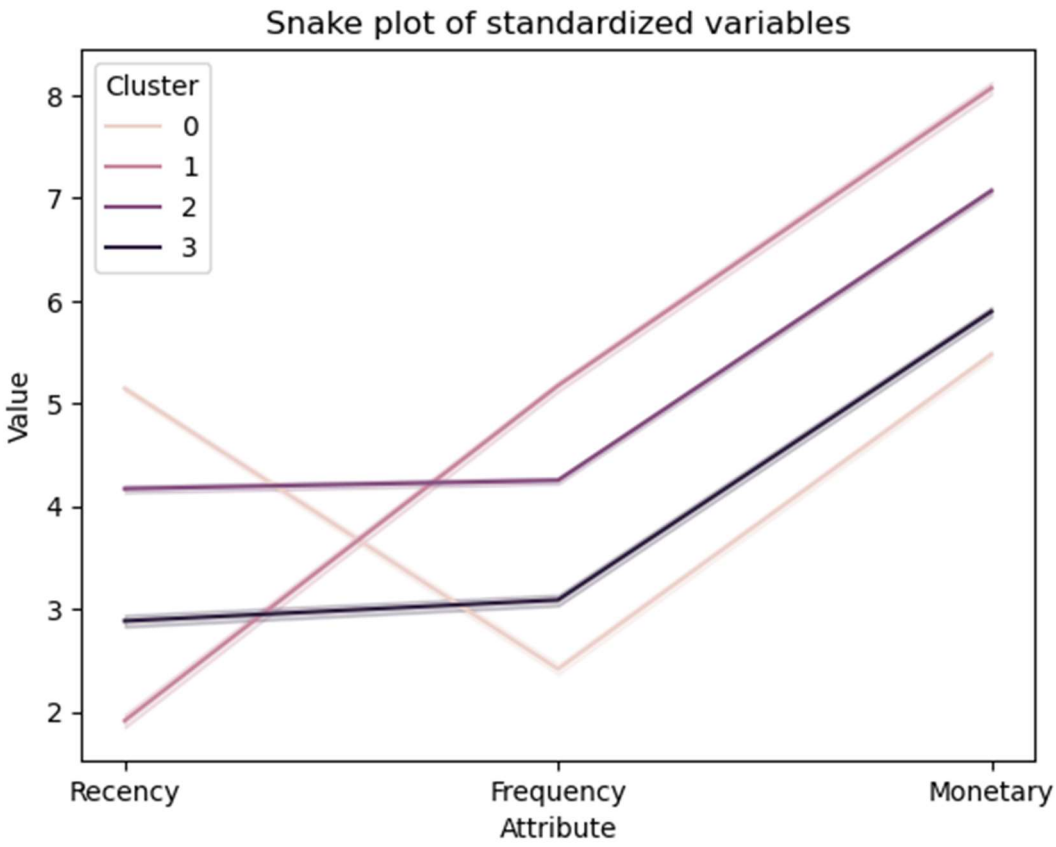
DATA VISUALIZATION



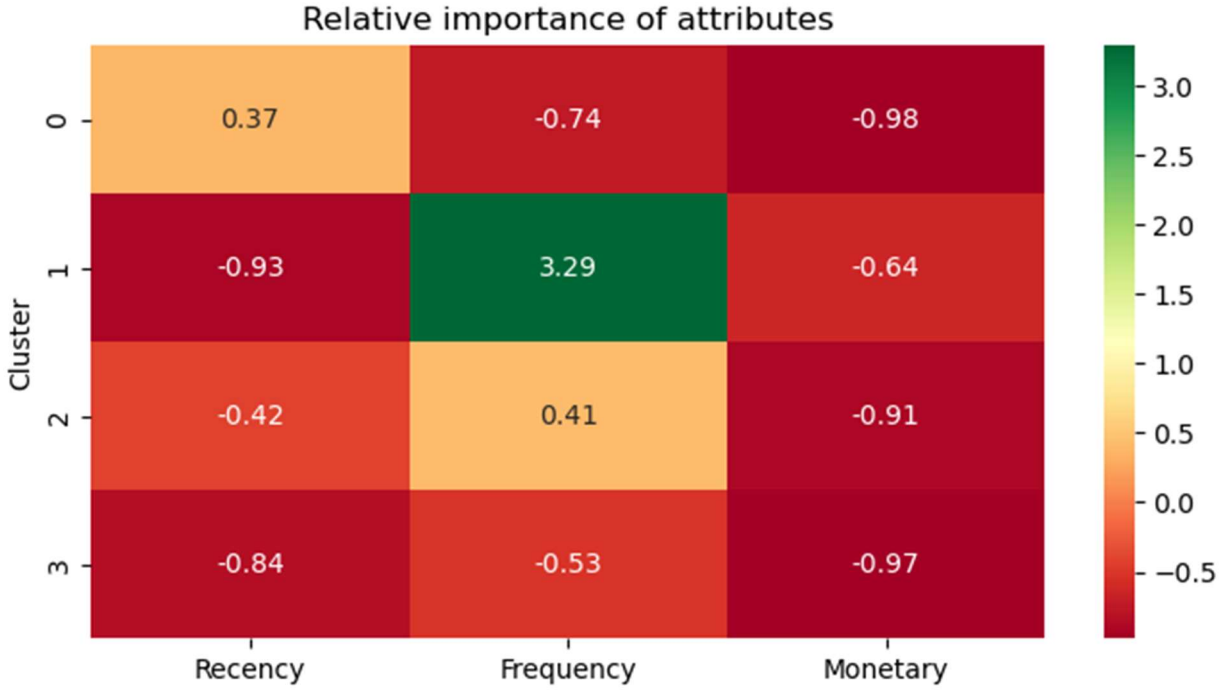

```
country_quantity = df.groupby('Country')['Quantity'].sum().reset_index()

# Sort countries based on total quantity in descending order
country_quantity = country_quantity.sort_values(by='Quantity', ascending=False)

# Plot bar plot for total quantity of products purchased from each country
plt.figure(figsize=(12, 6))
sns.barplot(data=country_quantity, x='Quantity', y='Country', palette='muted')
plt.title('Bar Plot: Total Quantity of Products Purchased from Each Country')
plt.xlabel('Total Quantity')
plt.ylabel('Country')
plt.show()
```



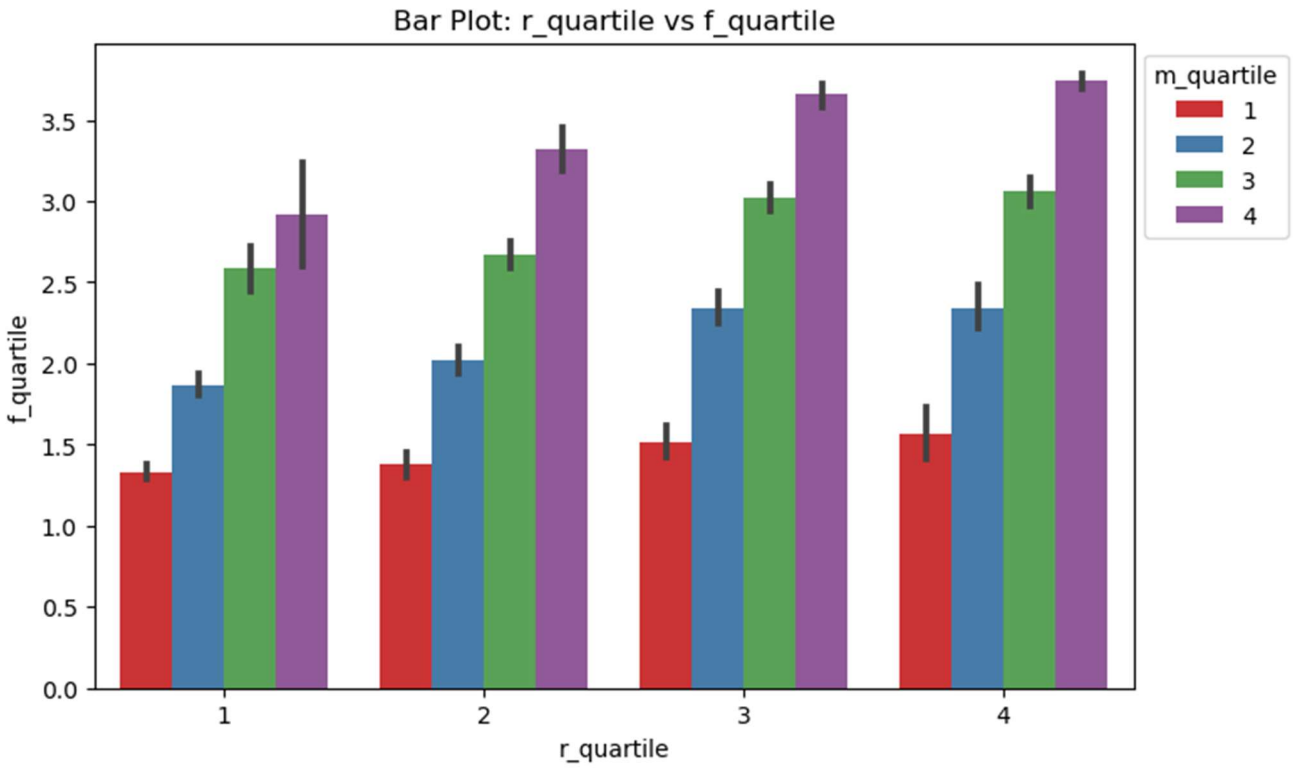

```
plt.figure(figsize=(10, 6))
plt.subplot(1, 2, 1)
sns.violinplot(data=data, x='r_quartile', y='Recency', palette='Set2')
plt.title('Violin Plot: Recency Quartile')
plt.xlabel('Recency Quartile')
plt.ylabel('Recency')

plt.subplot(1, 2, 2)
sns.violinplot(data=data, x='m_quartile', y='Monetary', palette='Set3')
plt.title('Violin Plot: Monetary Quartile')
plt.xlabel('Monetary Quartile')
plt.ylabel('Monetary')

plt.tight_layout()
plt.show()
```

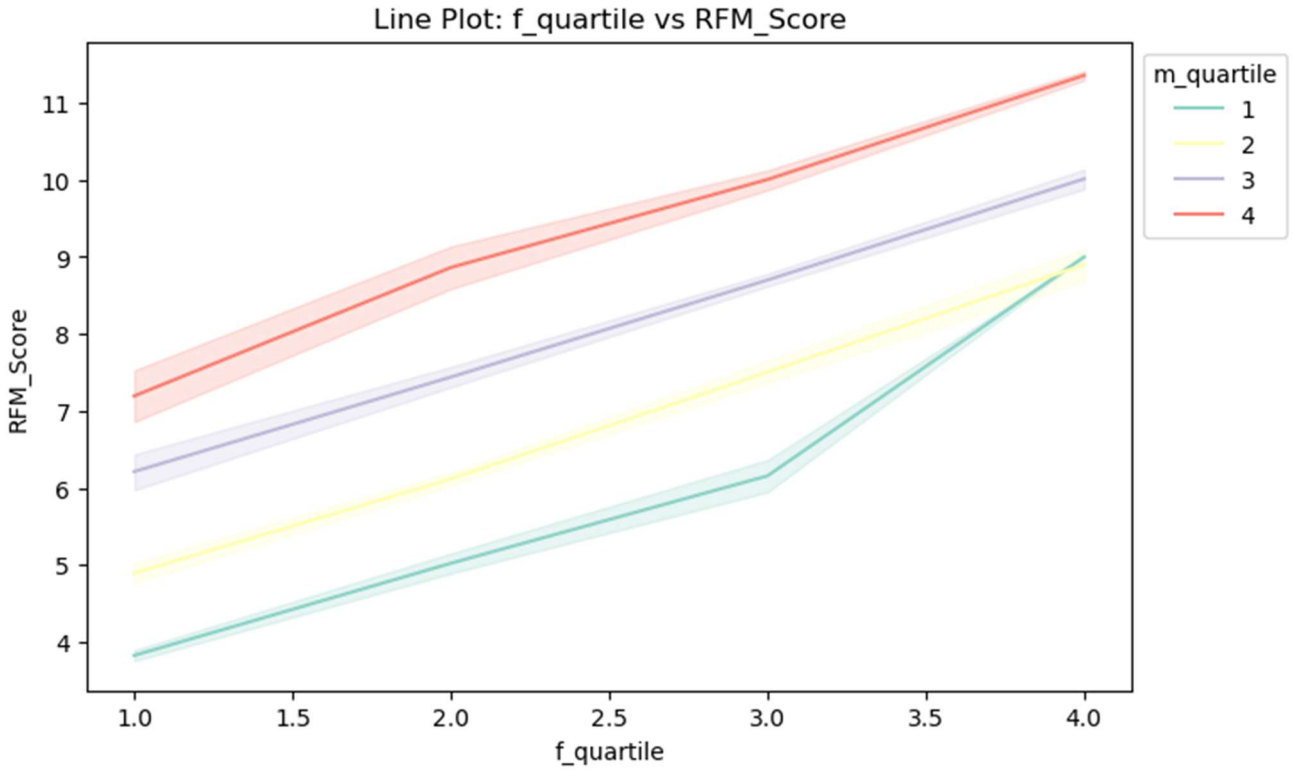



```
plt.title('Snake plot of standardized variables')
sns.lineplot(x="Attribute", y="Value", hue='Cluster', data=data_melt)
```

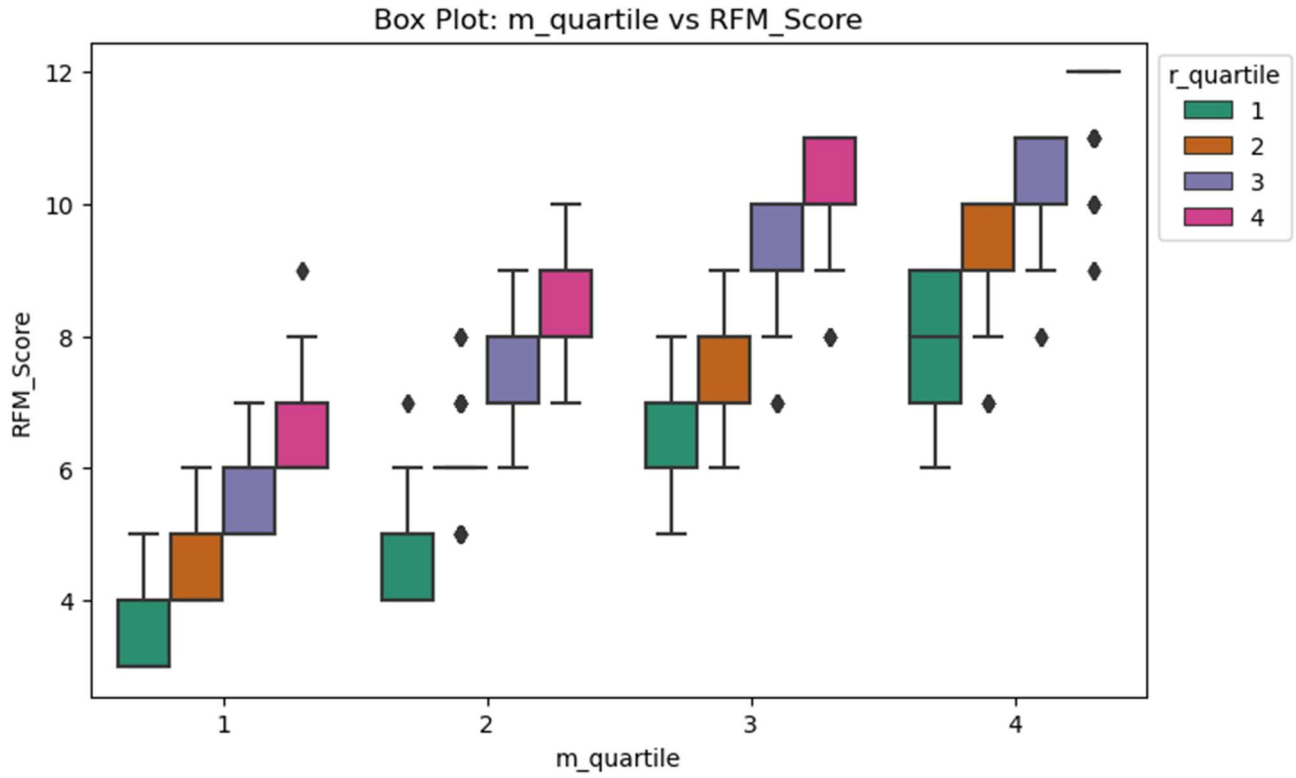


```
# Plot heatmap
plt.figure(figsize=(8, 4))
plt.title('Relative importance of attributes')
sns.heatmap(data=relative_imp, annot=True, fmt='.2f', cmap='RdYlGn')
plt.show()
```





```
plt.figure(figsize=(8, 5))
sns.barplot(data=data, x='r_quartile', y='f_quartile', hue='m_quartile', palette='Set1')
plt.title('Bar Plot: r_quartile vs f_quartile')
plt.xlabel('r_quartile')
plt.ylabel('f_quartile')
plt.legend(title='m_quartile', loc='upper left', bbox_to_anchor=(1, 1))
plt.show()
```




```
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='m_quartile', y='RFM_Score', hue='r_quartile', palette='Dark2')
plt.title('Box Plot: m_quartile vs RFM_Score')
plt.xlabel('m_quartile')
plt.ylabel('RFM_Score')
plt.legend(title='r_quartile', loc='upper left', bbox_to_anchor=(1, 1))
plt.show()
```

CHAPTER 9

MODEL BUILDING

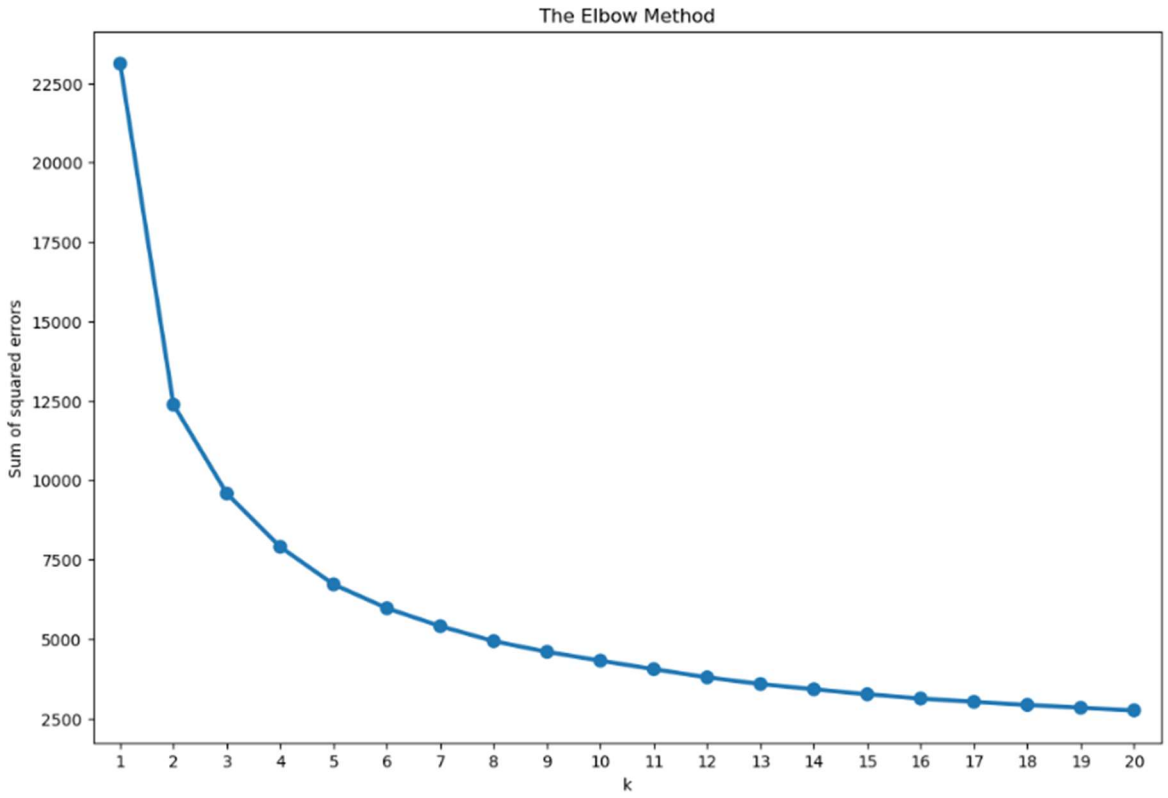
9.1 K-Means :

In [71]:
```
from sklearn.cluster import KMeans
```

In [72]:
```
sse = {}

# Fit KMeans and calculate SSE for each k
for k in range(1, 21):

    # Initialize KMeans with k clusters
    kmeans = KMeans(n_clusters=k, random_state=1)

    # Fit KMeans on the normalized dataset
    kmeans.fit(data_norm)

    # Assign sum of squared distances to k element of dictionary
    sse[k] = kmeans.inertia_
```

In [73]:
```
plt.figure(figsize=(12,8))

plt.title('The Elbow Method')
plt.xlabel('k');
plt.ylabel('Sum of squared errors')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```

Elbow Method:



```
kmeans = KMeans(n_clusters=3, random_state=1)

# Compute k-means clustering on pre-processed data
kmeans.fit(data_norm)

# Extract cluster labels from labels_ attribute
cluster_labels = kmeans.labels_
```

K-means clustering is a simple yet powerful algorithm used in data analysis and machine learning to group similar data points together. It works by assigning data points to a fixed number of clusters based on their similarities. The algorithm starts with random cluster centers, iteratively adjusts them, and reassigns data points until the grouping stabilizes. As a result, k-means clustering allows us to discover patterns and structure within data, making it easier to understand and analyze large datasets efficiently.

The provided code performs k-means clustering on pre-processed data using the scikit-learn library in Python. Let's break down the steps and provide a description of each:

1. `kmeans = KMeans(n_clusters=3, random_state=1)`: In this line, the k-means clustering algorithm is initialized with the number of clusters set to 3 (`n_clusters=3`). The `random_state` parameter is set to 1, which ensures reproducibility by fixing the random seed used for initializing the centroids.

2. `kmeans.fit(data_norm)`: The `fit()` method is called on the k-means object to perform the clustering. The input to this method is `data_norm`, which represents the pre-processed data. The k-means algorithm will iteratively group the data points into three clusters based on their similarities.

3. `cluster_labels = kmeans.labels_`: After fitting the k-means model, the cluster assignments for each data point are obtained using the `labels_` attribute. The `labels_` attribute contains an array with an entry for each data point, indicating which cluster the corresponding data point has been assigned to.

CHAPTER 10
FLASK DEPLOYMENT AND WEB PAGE UI AND OUTCOME:

What is Flask?

Flask is a lightweight and popular web framework for Python. It is designed to make it easy to build web applications and APIs quickly and with minimal boilerplate code. Flask is classified as a micro-framework because it provides only the essentials to get started, leaving the developers with the flexibility to choose and integrate other libraries and components as needed.

Key features and characteristics of Flask include:

1. Minimalistic: Flask has a simple and straightforward API that allows developers to get started quickly. It focuses on simplicity and provides only the basic tools needed to build web applications.

2. Routing: Flask allows you to define URL routes for your web application, specifying which function should be executed when a particular URL is accessed.

3. Templates: Flask includes a powerful templating engine, Jinja2, which enables developers to create dynamic web pages by embedding Python code into HTML templates.

4. HTTP Request/Response Handling: Flask provides utilities to handle HTTP requests (GET, POST, etc.) and generate HTTP responses, making it easy to interact with web browsers and clients.

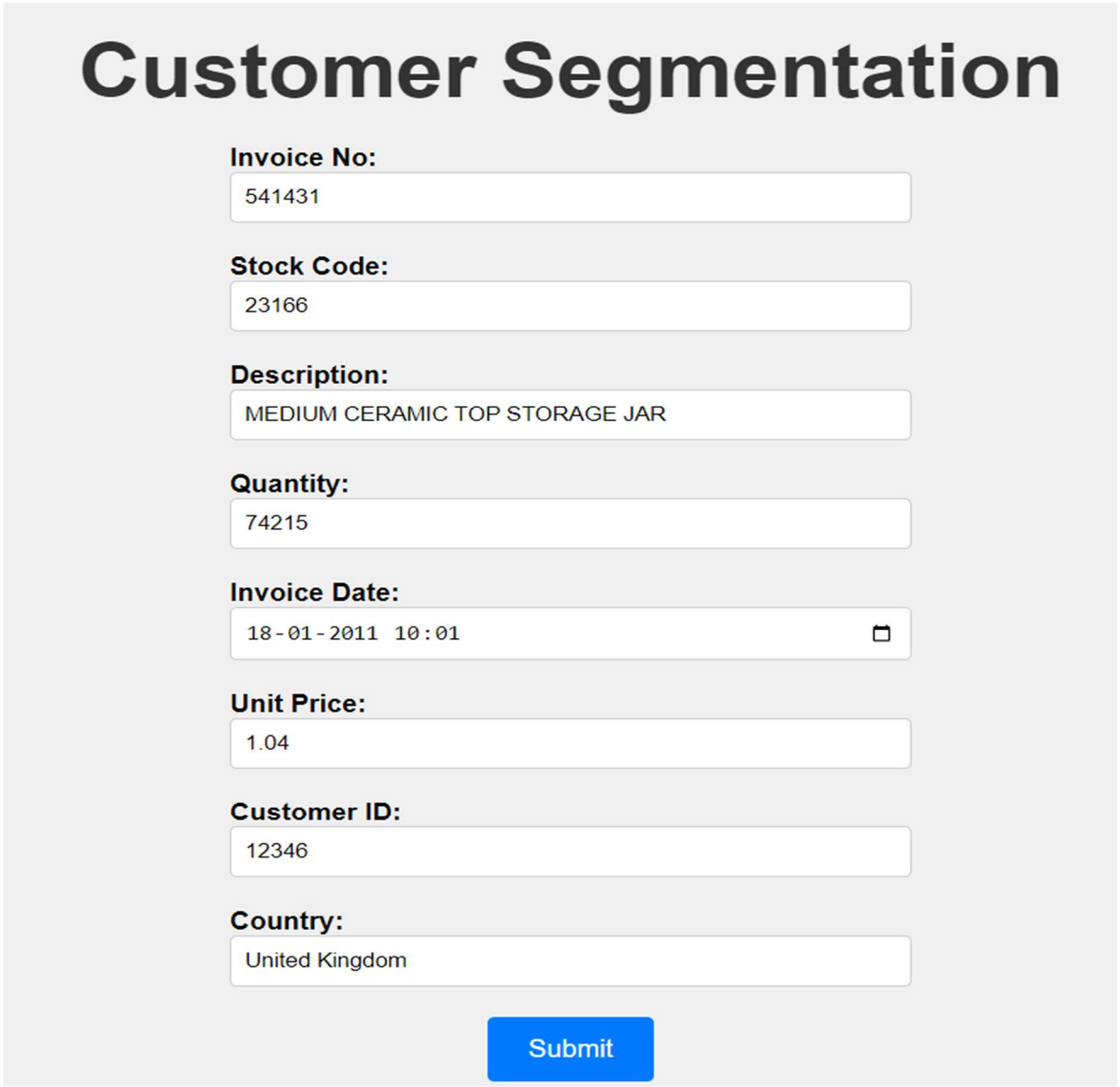
5. Extensions: While Flask is minimalistic, it supports a wide range of extensions that can be easily integrated to add functionality, such as database integration, form handling, authentication, and more.

6. Lightweight and Scalable: Due to its minimalistic nature, Flask is lightweight and well-suited for small to medium-sized applications. It can be extended and scaled up to handle more complex projects when needed.

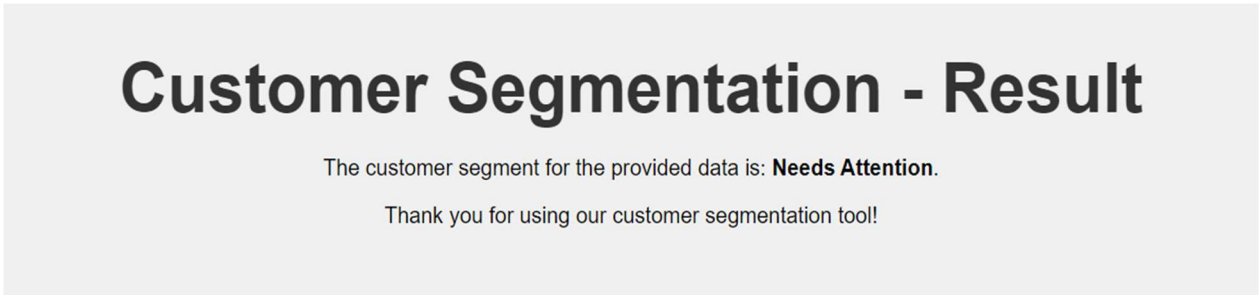
7. RESTful APIs: Flask is commonly used to build RESTful APIs due to its simplicity and flexibility. It allows developers to define routes that correspond to different API endpoints, handling JSON data efficiently.

8. Flask is widely adopted by both beginners and experienced developers due to its ease of use, flexibility, and large community support**.**

USER INTERFACE :



OUTCOME :



CHAPTER 11

Advantages of Customer Segmentation using Form and K-means:

• Enhanced Customer Understanding: Customer segmentation using form and k-means clustering allows businesses to divide their customer base into distinct groups based on similarities in their responses to specific questions or preferences. This provides a deeper understanding of the different customer segments, their needs, and preferences.

• Personalized Marketing and Product Recommendations: With customer segmentation, businesses can tailor their marketing messages and product recommendations to target each segment more effectively. This personalized approach increases the likelihood of engaging customers and improving conversion rates.

• Efficient Resource Allocation: By understanding the characteristics and behaviors of different customer segments, businesses can allocate their resources more efficiently. They can focus their marketing efforts and resources on the most promising customer groups, leading to improved ROI.

Disadvantages of Customer Segmentation using RFM and K-means:

• Limited by Form Questions: The effectiveness of customer segmentation heavily relies on the quality and relevance of the questions in the form. If the form does not capture essential

customer attributes or preferences, the resulting segmentation may be less accurate or meaningful.

• Homogeneity within Segments: K-means clustering assumes that each segment is spherical and equally distributed, which may not always reflect the true complexity of customer behavior. It can lead to some segments being less homogenous than others.

• Dynamic Customer Preferences: Customer preferences and behavior change over time. The customer segmentation derived from past data may become less relevant as customer behavior evolves, necessitating regular updates to the segmentation model.

Applications of Customer Segmentation using Form and K-means:

• Product Personalization: Businesses can use customer segmentation to personalize product recommendations and offers based on the preferences of each segment, increasing customer satisfaction and loyalty.

• Targeted Marketing Campaigns: Segment-specific marketing campaigns can be designed to appeal to the interests and needs of each customer group, leading to higher response rates and improved campaign effectiveness.

• Customer Retention Strategies: Understanding different customer segments allows businesses to implement targeted retention strategies for high-value customers or those at risk of churning, fostering long-term relationships.

• Service Customization: By knowing the preferences of each segment, businesses can tailor their services to meet specific customer needs, leading to improved customer experience and loyalty.

In summary, customer segmentation using form and k-means clustering offers several advantages, such as personalized marketing, improved customer understanding, and efficient resource allocation. However, it is essential to consider the limitations, such as the dependency on form questions and the need for regular updates, to ensure the segmentation remains accurate and relevant over time. The applications of this approach span across various areas, including e-commerce, marketing, market research, and customer relationship management.

CHAPTER 12

Conclusion

In conclusion, this project successfully employed RFM analysis and K-means clustering techniques to perform customer segmentation, yielding valuable insights for businesses aiming to enhance their customer management strategies. By analyzing Recency, Frequency, and Monetary value metrics, the customers were divided into distinct segments, enabling targeted marketing efforts and personalized experiences.

The experimental investigations demonstrated the effectiveness of the proposed approach in creating meaningful customer segments. The visualization of cluster distributions provided a clear understanding of customer behavior patterns, empowering businesses to identify high-value customers, re-engage dormant ones, and improve overall customer satisfaction.

The combination of RFM analysis and K-means clustering offered a data-driven and automated solution to the existing challenges faced in traditional customer segmentation methods. This approach can accommodate large datasets, making it scalable and adaptable for future analyses and changing customer behavior.

The project's outcomes showcase the significance of customer segmentation as a fundamental strategy for businesses to optimize their marketing efforts and drive growth. The identified customer segments allow businesses to tailor their strategies to specific customer needs, leading to increased customer loyalty, retention, and revenue generation.

As customer preferences and behaviors continue to evolve, this project serves as a foundation for ongoing research and improvements in customer segmentation methodologies. With the insights gained, businesses can gain a competitive edge in the market by better understanding their customers and delivering exceptional experiences tailored to their unique requirements.