GRT INSTITUTE OF ENGINEERING AND TECHNOLOGY, TIRUTTANI - 631209

Approved by AICTE, New Delhi Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CUSTOMER SEGMENTATION USING DATA SCIENCE

PROJECT REPORT

SUBMITTED BY

HARISH PRABU .S

3RD YEAR 5TH SEM

110321104014

harishprabu2003@gmail.com

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **“CUSTOMER SEGMENTATION USING DATA SCIENCE”** is the bonafide work of **“HARISH PRABU. S [110321104014]”** who carried out the project work under my/our supervision.

SIGNATURE
Dr.N. Kamal
head of the department

Department of Computer Science And Engineering
GRT Institute of Engineering and Technology
Tiruttani

SIGNATURE
Mr.T.Vinayagam
supervisor
Assistant professor

Department of Computer Science And Engineering
GRT Institute of Engineering and Technology
Tiruttani

Certified that the candidates were examined in Viva-voce in the Examination Held on______________________

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We thank our Management for providing us all support to complete this project successfully. Our sincere thanks to honorable **Chairman**, **Shri. G. RAJENDRAN and Managing Director, Shri.G.R. RADHAKRISHNAN** for creating a wonderful atmosphere inside the campus.

We are very grateful to **Dr.S. ARUMUGAM, M.E., Ph.D., Principal**, for providing us with consistent guidance and motivation to execute a real time project to learn and experience the project work in an environments to complete our project successfully.

Our sincere thanks to **Dr.N. KAMAL, M.E., Ph.D., Professor and Head, Department of Computer Science and Engineering** for giving me this wonderful opportunity to do the project and providing the require facilities to fulfill our work.

We are highly indebted and thankful to our project Evaluators **Mrs V. Priya M.E**., And **Mrs. Edith Esther M.E., Assistant Professor, Department of Computer Science and Engineering** for his immense support in doing the project.

We are very grateful to our internal guide **Mr. T.A. VINAYAGAM, M.Tech., AssistantProfessor, Department of Computer Science and Engineering** for guiding us with her valuable suggestions to complete our project.

We also dedicate equal and grateful acknowledgement to all the **respectable members of thefaculty and lab in-charges** of the Department of Computer Science and Engineering, friends andour families for their motivation, encouragement and continuous support.

Our sincere thanks **to IBM and Skill Up Team members** for giving me this wonderful opportunity to do the projectand providing the require guidance and valuable online session

CUSTOMER SEGMENTATION USING DATA SCIENCE

ABSTRACT

Customer segmentation is a vital strategy for businesses aiming to enhance their marketing efforts, optimize product offerings, and improve overall customer satisfaction. This abstract provides an overview of the concept of customer segmentation and its application through data science techniques.

Customer segmentation involves dividing a customer base into distinct groups based on shared characteristics and behaviors. These segments can encompass a wide range of variables, including demographics, psychographics, purchase history, and online behavior. The ultimate goal of segmentation is to gain a deeper understanding of customers and tailor marketing strategies and product offerings to meet their specific needs and preferences.

Data science plays a pivotal role in the process of customer segmentation. By leveraging advanced analytical tools, machine learning algorithms, and vast datasets, businesses can uncover hidden patterns and insights within their customer data. These insights empower organizations to create more personalized and targeted marketing campaigns, which, in turn, result in higher conversion rates and customer loyalty.

This abstract explores various data science techniques commonly used in customer segmentation,such as clustering algorithms like k-means and hierarchical clustering, decision trees,and predictive modeling. These techniques help identify distinct customer segments with precision,allowing businesses to craft tailored marketing messages and strategies.

Furthermore, the abstract highlights the benefits of customer segmentation through data science,including improved customer acquisition,reduced churn rates,increased customer lifetime value,and enhanced customer satisfaction.It also underscores the importance of ethical considerations and data privacy when implementing datadriven segmentation strategies. Customer segmentation using data science is a powerful approach for businesses seeking to gain a competitive edge in today's data-driven marketplace.By harnessing the potential of data science techniques, organizations can better understand their customers,boost marketing effectiveness, and foster long-lasting customer relationships.

TABLE OF CONTENTS

CHAPTER No	TITLE	PAGE No
	ABSTRACT	4
1.	**PHASE 1**	
	1.1 INTRODUCTION	8
	1.2 PROBLEM DEFINITION	8
	1.3 DESIGN THINKING	9
	1.4 OBJECTIVES	10
	1.5 SYSTEM DESIGN AND THINKING	11
	1.6 SYSTEM ARCHITECTURE	12
	1.7 E – R DIAGRAM	12
	1.9 USE CASE DIAGRAM	13
	1.10 ARCHITECTURE	14
	1.11 SEQUENCE DIAGRAM	15
2.	**PHASE 2**	
	2.1 SHORT EXPLAINATION ABOUT CUSTOMER SEGMENTATION USING DATA SCIENCE	16

	2.2 WHERE I GOT THE DATASETS AND ITS DETAILS	17
	2.3 DETAILS ABOUT COLUMNS	18
	2.4 DETAILS OF LIBRARIES TO BE USED AND WAY TO DOWNLOAD	19
	2.5 HOW TO TRAIN AND TEST THE DATASET	20
	2.6 REST OF EXPLAINATION	21
	2.7 WHAT METRICS USED FOR THE ACCURACY CHECK	22
3.	**PHASE 3**	
	3.1DATASET AND ITS DETAIL EXPLANATION AND IMPLEMENTATION OF CUSTOMER SEGMENTATION USING DATA SCIENCE	24
	3.2 BEGIN BUILDING THE PROJECT BY LOAD THE DATASET CUSTOMERIDS	24
	3.3 PREPROCESS DATASET	28
	3.4 PERFORMING DIFFERENT ANALYSIS NEEDED	31

4.	**PHASE 4**	
	4.1: IN THIS TECHNOLOGY YOU WILL CONTINUE BUILDING YOUR PROJECT BY PREPROCESSING YOUR DATASET	35
	4.2: IN THIS TECHNOLOGY YOU WILL CONTINUE BUILDING YOUR PROJECT BY PERFORMING FEATURE ENGINEERING	38
	4.3:MODEL TRAINING AND EVALUATION	43
	4.4: PERFORM DIFFERENT ANALYSIS AS NEEDED	47
	REFERENCE	50

CHAPTER 1

PHASE 1

1.1 INTRODUCTION

Customer segmentation is a critical strategy in marketing and business that involves dividing a company's customer base into distinct groups or segments based on shared characteristics, behaviors, and preferences. These segments are created to help businesses better understand their customers and tailor their products, services, and marketing efforts to meet the specific needs and expectations of each group. Data science plays a pivotal role in this process by providing the tools and techniques needed to analyze and categorize customers effectively.

The primary goal of customer segmentation using data science is to enhance customer experiences, increase customer retention, and drive revenue growth. This is achieved by uncovering hidden patterns, relationships, and insights within the vast amounts of customer data that modern businesses accumulate. Data science techniques, including machine learning, statistical analysis, and data mining, are employed to identify meaningful distinctions among customers. These distinctions can include demographic factors (age, gender, location), behavioral data (purchase history, browsing habits), psychographic information (lifestyle, interests), and more.

1.2 PROBLEM DEFINITION

In the modern business landscape, understanding your customers is essential for success. Customer segmentation is a powerful strategy that enables companies to divide their customer base into distinct groups based on shared characteristics. This process allows businesses to tailor their marketing, product development, and customer service efforts to specific customer segments, ultimately leading to improved customer satisfaction, higher retention rates, and increased profitability.

However, the challenge lies in identifying the most effective way to segment customers. Traditional methods often rely on basic demographic data, such as age, gender, or location, which may not provide a comprehensive understanding of customer behavior and preferences. To address this issue, data science offers a sophisticated and data-driven approach to customer segmentation.

Problem statement :

The problem at hand is to develop a data science solution for customer segmentation. The goal is to divide the customer base into meaningful and actionable segments based on their shared characteristics and behaviors. These segments should be used to customize marketing strategies, product offerings, and customer experiences, ultimately maximizing the business's return on investment.

1.3 DESIGN THINKING

Design Thinking for Customer Segmentation Using Data Science Design Thinking is a human-centered approach that can revolutionize the way businesses undertake customer segmentation using data science.

In this iterative process, the primary focus is on understanding, empathizing with, and addressing the needs and preferences of customers.

Here's a concise overview of applying Design Thinking to customer segmentation:

Empathize :

Start by truly understanding your customers. Conduct interviews, surveys, and data analysis to uncover their behaviors, pain points, and aspirations. Develop customer personas to embody these insights.

Define : Clearly define the problem you want to solve through segmentation and establish concrete objectives. For example, it could be optimizing marketing strategies for better engagement or personalizing product recommendations.

Ideate : Gather a cross-functional team of data scientists, marketers, and other stakeholders to brainstorm segmentation ideas. Encourage innovative thinking and creativity to generate diverse solutions.

Prototype : Utilize data preprocessing techniques to clean and consolidate relevant data, including demographic, behavioral, and psychographic information. Apply clustering algorithms or machine learning to create initial customer segments. Visualize these segments to enhance understanding.

Test : Share the proposed segments and insights with stakeholders. Validate the segments through user testing, A/B testing, or other experiments. Gather feedback and iteratively refine the segments.

Implement : Work closely with marketing teams to operationalize the segmentation strategy.Ensure seamless integration of customer segments with marketing tools and databases. Provide training to relevant teams.

Evaluate : Define key performance indicators (KPIs) to assess the success of the segmentation strategy, such as conversion rates or customer retention. Regularly monitor and review performance to make data-driven adjustments.

Iterate and Improve : Recognize that customer behavior evolves. Continuously update andrefine segments based on feedback, changing trends, and new data. Maintain a feedback loop for ongoing improvement.

1.4 OBJECTIVES

Objective 1: Enhance Customer Understanding

- To gain a comprehensive understanding of our customer base by leveraging data science techniques to segment customers based on behavior, preferences, and demographics.

Objective 2: Improve Marketing Effectiveness

- To optimize marketing efforts by tailoring campaigns to specific customer segments, resulting in increased conversion rates, higher engagement, and improved return on investment (ROI).

Objective 3: Enhance Product and Service Personalization

- To deliver more personalized products and services by identifying the unique needs and preferences of different customer segments, ultimately improving customer satisfaction and loyalty.

Objective 4: Increase Customer Retention

- To reduce customer churn and increase customer loyalty through targeted retention strategies designed for each customer segment.

Objective 5: Optimize Resource Allocation

- To allocate resources more efficiently by directing advertising budgets, customer support, and other resources toward the segments that offer the highest potential for growth and profitability.

Objective 6: Drive Data-Driven Decision-Making

- To foster a culture of data-driven decision-making within the organization, where customer segmentation insights are used to inform strategic planning and business operations.

Objective 7: Measure and Monitor Segment Performance

- To establish key performance indicators (KPIs) for each customer segment and continuously monitor their performance to adapt strategies as needed and ensure long-term success.

Objective 8: Enhance Competitive Advantage

- To gain a competitive edge by leveraging data science to create more nuanced and effective customer segments compared to competitors using traditional segmentation methods.

Objective 9: Foster Innovation

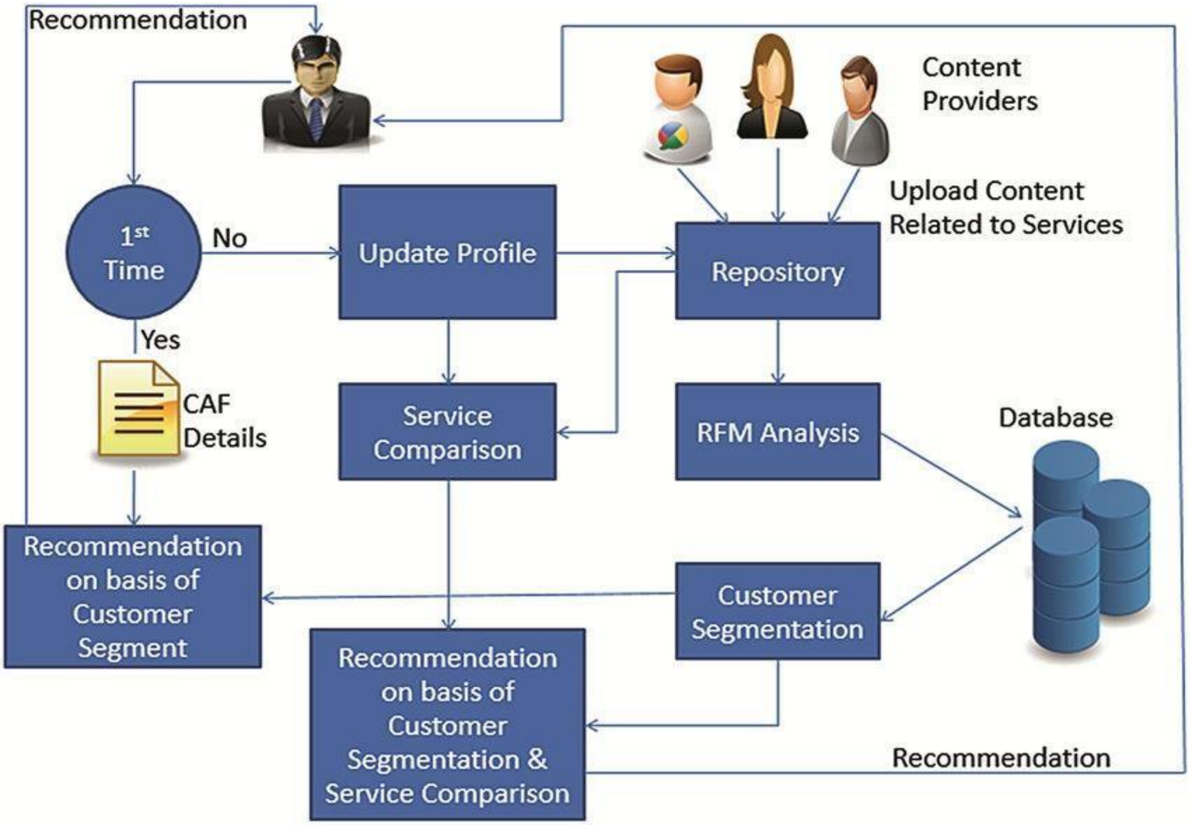
- To encourage innovation by using customer segmentation insights to identify unmet customer needs and develop new products or services tailored to specific segments.

Objective 10: Ensure Ethical Data Handling

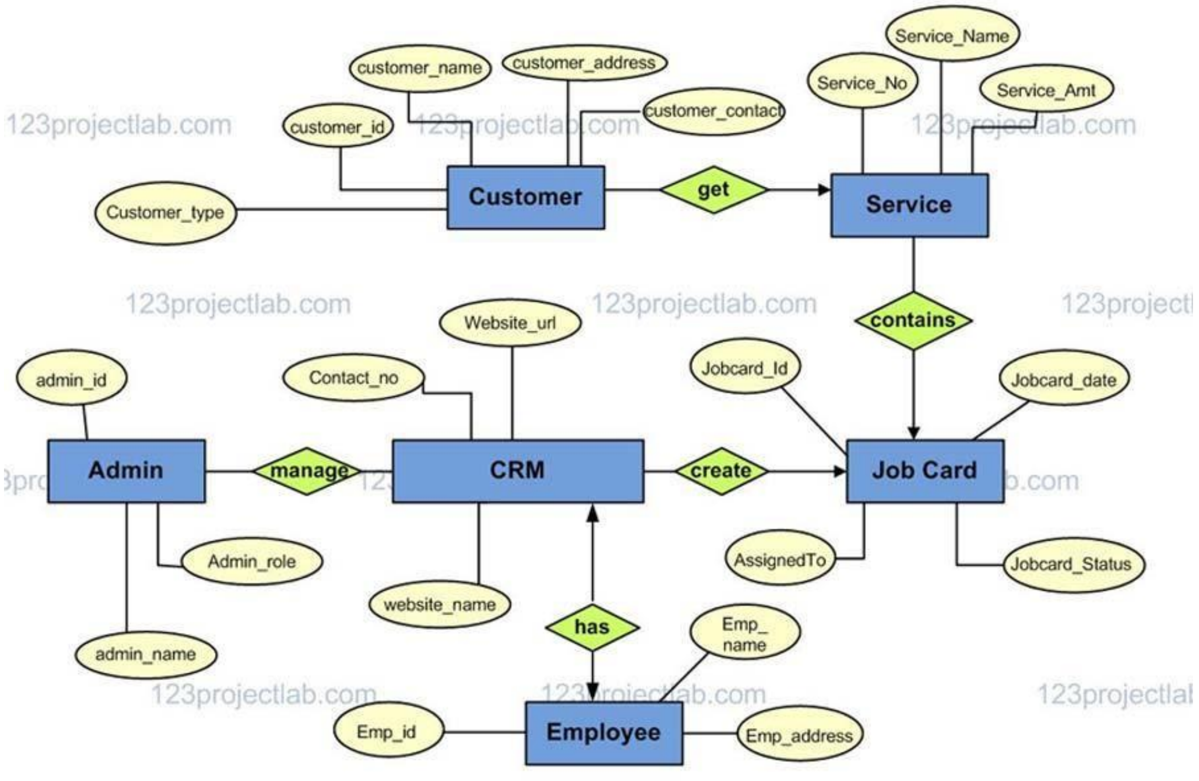
- To prioritize ethical data handling and privacy considerations throughout the segmentation process, ensuring compliance with data protection regulations and maintaining customer trust. These objectives collectively aim to harness the power of data science for customer segmentation, enabling the organization to better understand, target, and serve its diverse customer base while maintaining a commitment to ethical and responsible data practices.

SYSTEM DESIGN AND THINKING

2.1SYSTEM ARCHITECTURE

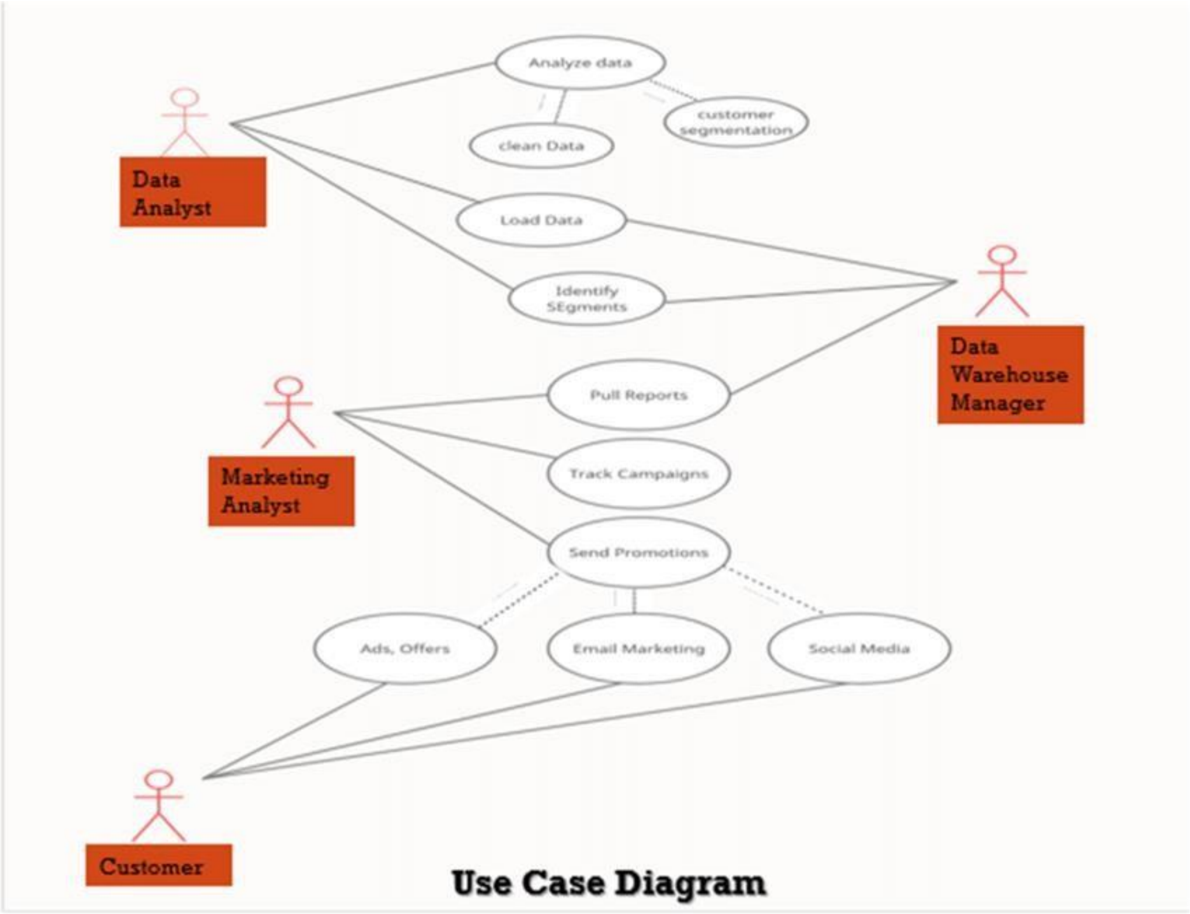


2.1 E – R DIAGRAM

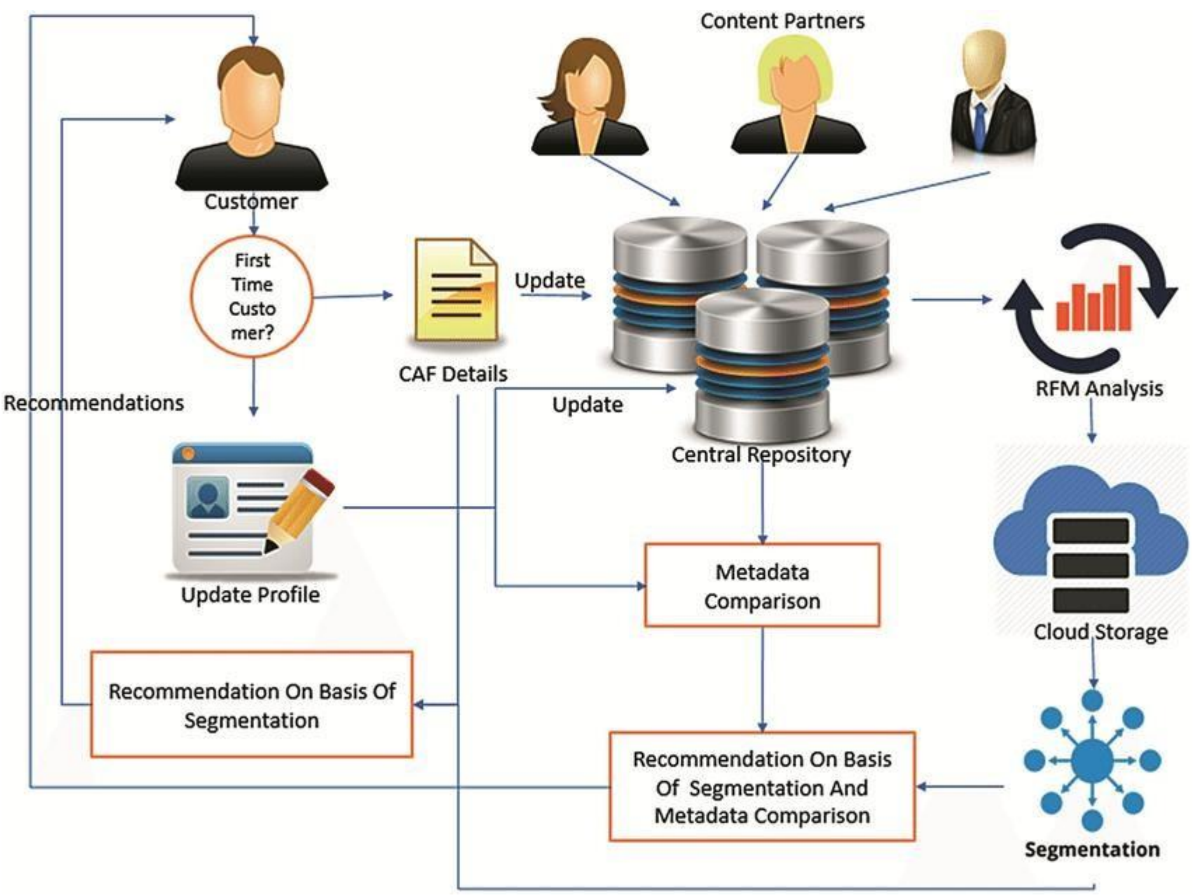


ER-Diagram for Customer Relationship Management System

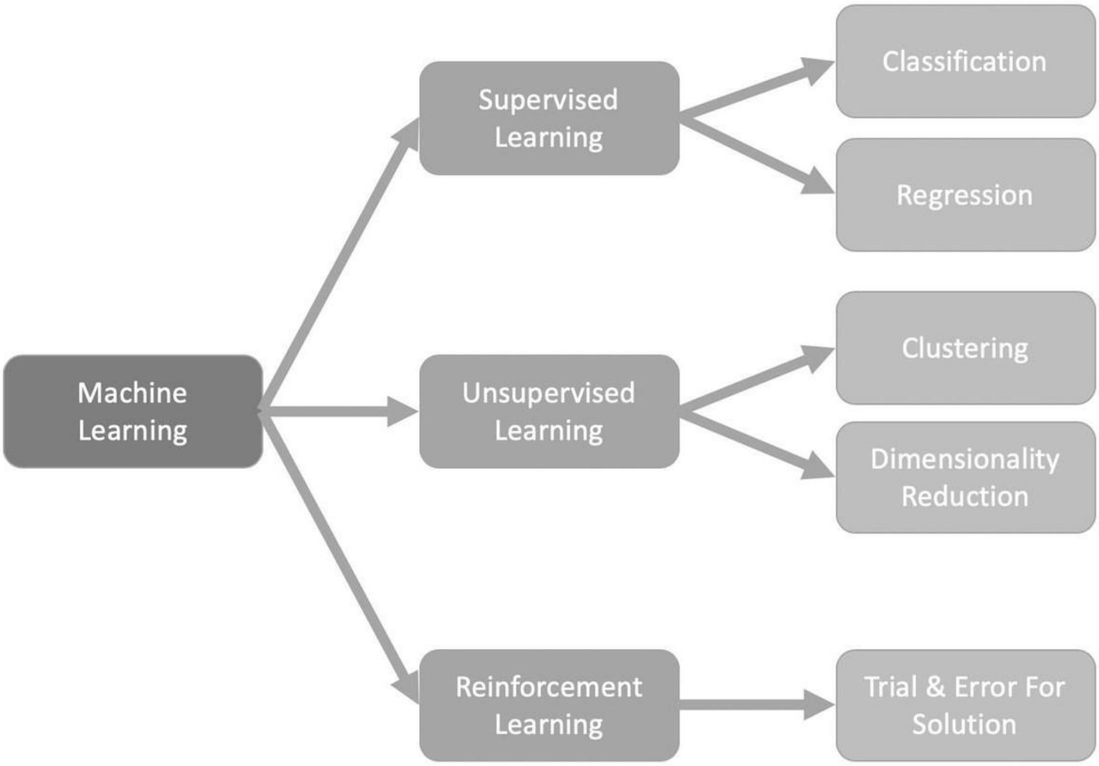
2.2 USE CASE DIAGRAM



2.3 ARCHITECTURE



2.4 SEQUENCE DIAGRAM



CHAPTER 2

PHASE 2

2.1 SHORT EXPLAINATION ABOUT CUSTOMER SEGMENTATION USING DATA SCIENCE

Customer segmentation using data science is the process of dividing a company's customerbase into distinct groups or segments based on common characteristics or behaviors, with the goal of tailoring marketing, product development, and customer service strategies to better meet the needs of each segment. Here's a brief explanation of how it works:

1. Data Collection: Gather relevant data about your customers, which can include demographic information (age, gender, location), psychographic data (lifestyle, interests), transaction history, website behavior, and more.

2. Data Preprocessing: Clean and prepare the data by handling missing values, outliers, and ensuring data consistency. This step is essential for accurate analysis.

3. Segmentation Techniques: Apply data science techniques such as clustering, classification, or dimensionality reduction to group customers with similar characteristics or behaviors together. Common methods include k-means clustering, hierarchical clustering, and decision tree algorithms.

4. Feature Selection: Identify the most important features (variables) that contribute to the segmentation. This helps in focusing on relevant aspects of customer behavior.

5. Model Validation: Evaluate the quality of your segmentation model using metrics like silhouette score, Davies-Bouldin index, or domain-specific criteria to ensure that the segments are meaningful and distinct.

6. Segment Profiling: Once segments are defined, create detailed profiles for each group. Understand their needs, preferences, and pain points to develop tailored marketing strategies and product offerings.

7. Targeted Marketing: Design and implement personalized marketing campaigns and strategies for each segment. This can involve customizing product recommendations, messaging, and advertising channels.

8. Monitoring and Iteration: Continuously monitor customer behavior and segment performance. Adjust your strategies as needed based on changing trends and feedback to ensure the segments remain relevant.

Customer segmentation using data science can lead to more effective marketing, increased customer satisfaction, higher retention rates, and improved overall business performance. It allows companies to treat different customer groups in a way that resonates with their unique preferences and characteristics, ultimately driving better results and stronger customer relationships.

2.2 WHERE I GOT THE DATASETS AND ITS DETAILS

You can find datasets for customer segmentation and various other data science projects from several reputable sources.

KAGGLE : Kaggle is a popular platform for data science competitions and dataset sharing. Ithosts a wide range of datasets on various topics, including customer data. You can browse datasets, read their descriptions, and download them for free. Kaggle also provides a community where you can discuss and collaborate on data science projects.

Website : *https://www.kaggle.com/datasets/akram24/mall-customers*

NAME OF THE DATASET : Mall Customers

DATA DESCRIPTION :

Customer segmentation is a common application of data science in the retail industry, including malls. To perform customer segmentation effectively, you need relevant data about mall customers. Once you have collected and cleaned the relevant data, you can apply variousdata science techniques such as clustering, classification, and regression to segment mall customers effectively. The goal is to identify groups of customers with similar characteristics and preferences to tailor marketing strategies, promotions, and store layouts to meet their needsand maximize the mall's revenue.

2.3 DETAILS ABOUT COLUMNS

CustomerID – in this column fill the ID details belongs to the customer that was givenGender

– Mention the gender of the customer

Age – Mention the age of the customer

Annual Income (k$) – Mentioning the customer annual income for a count purpose and calculating the spending score.

Spending Score (1-100) – In this by using the annual income column the spending score iscalculated for the customer in that specified mall

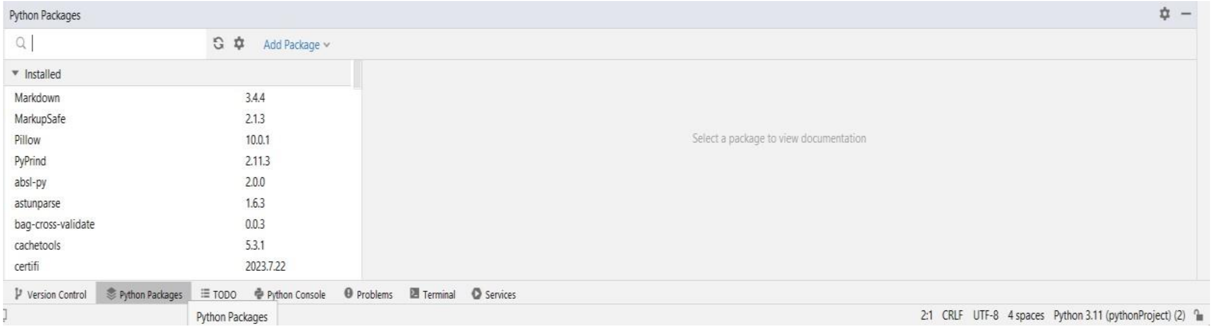
2.4 DETAILS OF LIBRARIES TO BE USED AND WAY TO DOWNLOAD

LIBRARIES TO BE USED

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.cluster import KMeans

WAY TO DOWNLOAD THE LIBRARIES

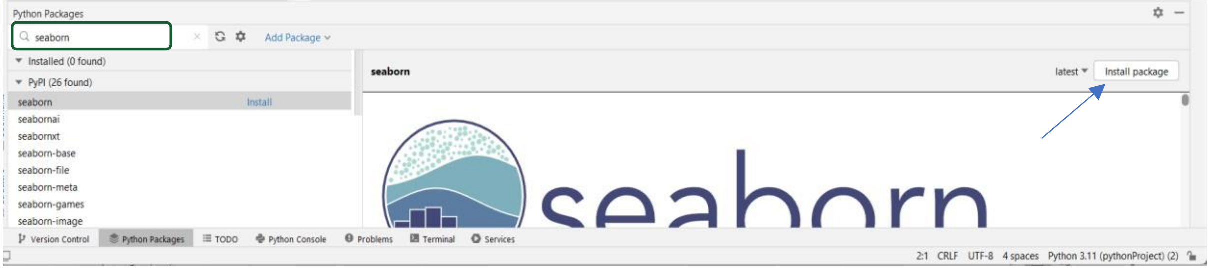
1. Click the python packages in the bottom of your project in pycharm



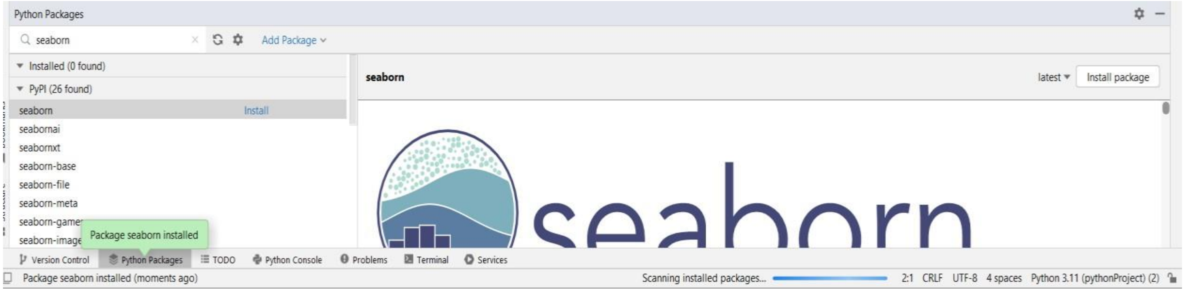
Markdown	3.4.4
MarkupSafe	2.1.3
Pillow	10.0.1
PyPrind	2.11.3
absl-py	2.0.0
astunparse	1.6.3
bag-cross-validate	0.0.3
cachetools	5.3.1
certifi	2023.7.22

Select a package to view documentation

2. Type the required library in the search box and click install package in the right end top of the python packages.



3. After installation process finished it shows the package was installed in the python packages.



Package seaborn installed

Package seaborn installed (moments ago)

Scanning installed packages...

2.5 HOW TO TRAIN AND TEST THE DATASET

To train and test a machine learning model using a dataset of mall customers with the givencolumn names (CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100)), youcan follow these steps:

Data Preprocessing :

Load your dataset into a data analysis or machine learning environment (e.g., Python with libraries like pandas and scikit-learn).

Explore and clean the data to handle any missing values, duplicates, or outliers.

Encode categorical variables like "Gender" into numerical values (e.g., 0 for Male, 1 for Female) if needed.

Splitting the Data :

Divide your dataset into two parts: a training set and a testing set. A common split is 80% fortraining and 20% for testing, but you can adjust this ratio as needed.

Ensure that the split maintains a representative distribution of data, especially if you have imbalanced classes or segments.

Selecting a Machine Learning Model :

Choose an appropriate machine learning model for your task. Since you want to segment customers, unsupervised learning techniques like clustering (e.g., K-Means, hierarchical clustering) are often used.

Feature Selection :

In this case, you can use features like "Age," "Annual Income," and "Spending Score" for customer segmentation. Exclude "CustomerID" and "Gender" if they do not contribute to the segmentation.

Training the Model :

Fit your chosen machine learning model to the training data using the selected features.

For example, if you're using K-Means clustering in Python with scikit-learn, you can use thefollowing code:

```
kmeansmodel = KMeans(n_clusters = 5 , init = "k-means++",random_state=0)

y_kmeans = kmeansmodel.fit_predict(x)
```

Model Evaluation :

Since clustering is unsupervised, traditional evaluation metrics like accuracy do not apply. Instead, you can use within-cluster sum of squares (WCSS) or silhouette score to assess the quality of the clusters.

Testing the Model :

Use the trained model to predict the clusters for the samples in your testing dataset.

```
plt.scatter(x[y_kmeans==0,0],x[y_kmeans==0,1],s=100,c="red",label = "cluster 1")

plt.scatter(x[y_kmeans==1,0],x[y_kmeans==1,1],s=100,c="blue",label = "cluster 2")

plt.scatter(x[y_kmeans==2,0],x[y_kmeans==2,1],s=100,c="green",label = "cluster 3")

plt.scatter(x[y_kmeans==3,0],x[y_kmeans==3,1],s=100,c="cyan",label = "cluster 4")

plt.scatter(x[y_kmeans==4,0],x[y_kmeans==4,1],s=100,c="magenta",label = "cluster 5")

plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=300,c="yellow",label="centroids")

plt.title("clusters of customers")


plt.xlabel("Yıllık gelir")

plt.ylabel=("harcama skoru")

plt.legend()
```

Analysing the Results :

Examine the clustering results to understand how customers are grouped based on their features.

Visualize the clusters, if possible, to gain insights.

Utilizing Segmentation :

Once you have clustered customers, you can tailor marketing strategies, promotions, and storelayouts for each segment based on their characteristics.

2.6 REST OF EXPLAINATION

Certainly! Continuing from where we left off, here is the rest of the explanation on how to work with a mall customer dataset for customer segmentation using data science:

Validation and Fine-Tuning : you may want to fine-tune your model by adjusting hyperparameters or exploring different clustering algorithms. Cross-validation can help you assess the robustness of your model and choose the best configuration.

Interpreting the Segmentation : Interpret the meaning of each customer segment. What are the distinguishing characteristics of each group? For example, do you have a high-income, high-spending segment and a low-income, low-spending segment? Use visualizations like scatter plots or bar charts to illustrate the differences between segments.

Targeted Marketing and Strategy : Develop tailored marketing strategies for each customersegment. For example, create promotions or advertisements that resonate with the unique preferences and behaviours of each group. Optimize store layouts, product placements, and inventory based on the identified segments.

Monitoring and Feedback : Continuously monitor the effectiveness of your strategies and promotions for each segment. Collect feedback from customers in each segment and use it to make data-driven improvements.

Retraining the Model : Over time, as new data becomes available, consider retraining your customer segmentation model. Customer preferences and behaviours can change, and your model should adapt accordingly.

Integration with Customer Relationship Management (CRM) : Integrate the segmentationresults with your CRM system to ensure that customer interactions and communications are personalized and consistent with the identified segments.

Privacy and Compliance : Ensure that you handle customer data with care and in compliance with relevant privacy regulations (e.g., GDPR, CCPA). Anonymize or pseudonymize customerdata as needed to protect privacy.

A/B Testing : Implement A/B testing for marketing campaigns to measure the impact of changes on different customer segments accurately.

Documentation and Reporting : Document your data pre-processing steps, model selection,and results thoroughly. This documentation is essential for future reference and model maintenance.

Scaling and Scalability : Consider how your customer segmentation process can scale as thedataset and business grow. Ensure that your infrastructure and tools can handle larger volumesof data.

2.7 WHAT METRICS USED FOR THE ACCURACY CHECK

When performing customer segmentation using data science, traditional accuracy metrics like classification accuracy are not applicable because customer segmentation is an unsupervised learning task. In unsupervised learning, there are no ground truth labels to compare predictions against. Instead, you use different metrics to evaluate the quality of the segmentation. Here are some commonly used metrics for assessing the accuracy of customer segmentation:

Silhouette Score : The silhouette score measures how similar each data point in one cluster isto the data points in the same cluster compared to the nearest neighboring cluster. A higher silhouette score indicates better-defined clusters.

Davies-Bouldin Index : This index measures the average similarity between each cluster andits most similar cluster. Lower values indicate better clustering, with a lower Davies-Bouldin Index representing more distinct clusters.

CHAPTER 3

PHASE 3

3.1DATASET AND ITS DETAIL EXPLANATION AND IMPLEMENTATION OF CUSTOMER SEGMENTATION USING DATA SCIENCE

- Customer segmentation is a technique used to group customers based on their characteristics. It helps organizations understand their customers better and make strategic decisions regarding product growth and marketing1.
- Machine learning can be employed to automate customer segmentation, which can be a tedious task when done manually. There are different methodologies for customer segmentation, and they depend on four types of parameters: geographic, demographic, behavioral, psychological1.
- There are several datasets available online that can be used for customer segmentation. One such dataset is the e-commerce dataset that contains the annual income of approximately 300 customers and the amount they spend annually on an e-commerce website2.
- To implement customer segmentation using machine learning techniques, you can use ensemble techniques such as Support Vector Machine (SVM), Logistics Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost Classifier, and Gradient Boosting Classifier3. You can also use clustering algorithms such as k-means and hierarchical clustering to derive the optimum number of clusters and understand the underlying customer segments based on the data provided

3.2 BEGIN BUILDING THE PROJECT BY LOAD THE DATASET

CUSTOMERIDS.

- Age.
- Gender.
- Annual Income.
- Spending Score.

- To begin building your customer segmentation project using the Mall Customers dataset from Kaggle, you'll need to load the dataset and start exploring it. You can use Python and popular libraries like Pandas for data manipulation and Matplotlib or Seaborn for data visualization. Here's a step-by-step guide:

1. IMPORT NECESSARY LIBRARIES:

- You'll need to import the necessary Python libraries to work with the dataset.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```

2. LOAD THE DATASET:

- Download the Mall Customers dataset from Kaggle and save it as a CSV file in your working directory. You can then load it into a Pandas DataFrame.

```
# Load the dataset

df = pd.read_csv(' C:\Users\CSE\Downloads\Mall_Customers.csv')
```

3. EXPLORE THE DATASET:

- Now, let's start exploring the dataset to understand its structure and the type of information it contains.

```
# Display the first few rows of the dataset

print(df.head())



# Check the basic statistics of the dataset
```

```
print(df.describe())


# Check for missing values
print(df.isnull().sum())


# Check the data types of each column
print(df.dtypes)
```

This code will give you an overview of the dataset, including the first few rows, summary statistics, missing values, and data types of each column.

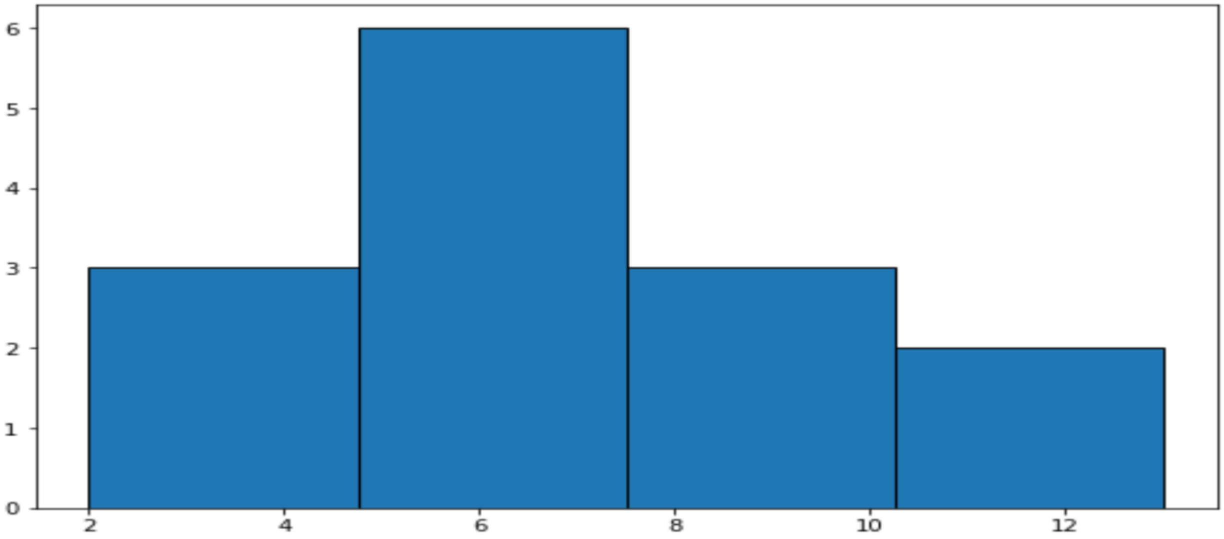
4. DATA VISUALIZATION:

- Data visualization is an important step in understanding the dataset and identifying potential trends and patterns. You can use libraries like Matplotlib and Seaborn for this purpose.

```
# Example: Visualize the distribution of Age
plt.figure(figsize=(8, 6))
sns.histplot(df['Age'], bins=20, kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```

You can create various types of plots and visualizations to gain insights into the dataset.

OUTPUT:



5. DATA PREPROCESSING:

- Depending on your analysis goals and the segmentation method you plan to use, you may need to preprocess the data. This could include handling outliers, scaling features, and encoding categorical variables.

6. CUSTOMER SEGMENTATION:

- Once you've explored and preprocessed the data, you can move on to customer segmentation using one of the techniques mentioned earlier (e.g., clustering using K-Means). You may also want to select the features (columns) that are relevant for your analysis.

7. EVALUATE AND INTERPRET:

- After segmentation, evaluate the quality of the segments and interpret the results to gain insights into customer groups. You can use visualization and statistical methods to do this.

From there, you can implement personalized marketing strategies, product recommendations, or any other actions based on the segments you've identified.

8. MONITOR AND ITERATE:

- Keep monitoring your customer segments over time and adjust your strategies as needed to maintain their relevance and effectiveness.

Remember to handle sensitive customer data responsibly and in accordance with data privacy regulations.This is a basic outline to get you started with your project using the Mall Customers dataset. Your specific analysis and segmentation approach may vary depending on your business goals and the insights you want to gain from the data.

3.3 PREPROCESS DATASET

1. IMPORT LIBRARIES:

- First, import the necessary libraries, including Pandas for data manipulation.

```
import pandas as pd
```

2. LOAD THE DATASET:

- Load the dataset from the CSV file. Make sure to download the dataset from Kaggle and place it in your working directory.

```
# Load the dataset

df = pd.read_csv('Mall_Customers.csv')
```

3. EXPLORE THE DATASET:

- Explore the dataset to understand its structure, check for missing values, and review data types.

```
# Display the first few rows of the dataset

print(df.head())


# Check for missing values

print(df.isnull().sum())


# Check the data types of each column



print(df.dtypes)
```

4. HANDLE MISSING VALUES:

- In this dataset, it's possible that there are no missing values. However, if there were any missing values, you'd need to decide how to handle them. Options include dropping rows with missing values, filling them with a default value, or using more advanced imputation techniques.

5. ENCODE CATEGORICAL DATA (IF ANY):

- The Mall Customers dataset doesn't contain categorical variables that need encoding. However, if your dataset had categorical data (e.g., "Gender"), you'd need to encode it, typically using one-hot encoding or label encoding.

6. FEATURE SELECTION:

- Depending on your analysis goals, you may want to select a subset of features for segmentation. For example, you might choose to focus on "Annual Income" and "Spending Score" for clustering customers.

```
# Select specific columns for analysis

selected_columns = ['Annual Income (k$)', 'Spending Score (1-100)']

df = df[selected_columns]
```

7. STANDARDIZE/NORMALIZE DATA (IF NEEDED):

- If you're using clustering algorithms that rely on distances (e.g., K-Means), it's often a good practice to standardize or normalize the data to bring features to the same scale. This can be done using techniques like Min-Max scaling or Z-score standardization.

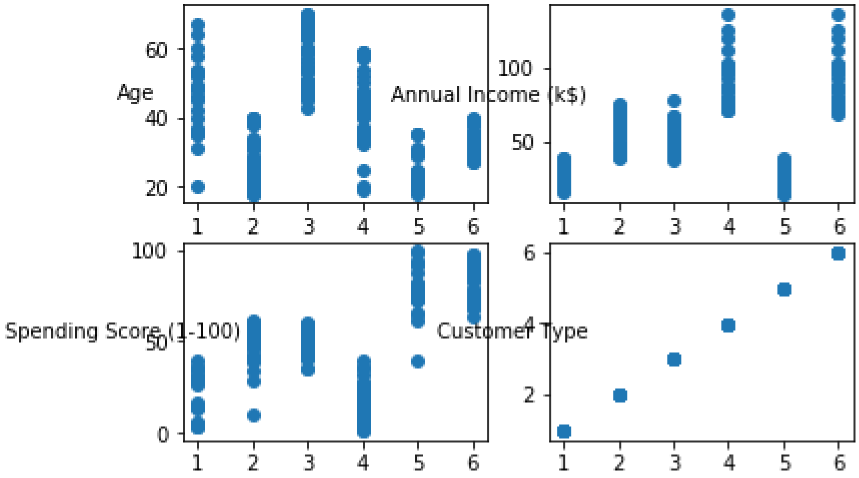
```
from sklearn.preprocessing import StandardScaler



scaler = StandardScaler()

df_scaled = scaler.fit_transform(df)
```

8. SAVE THE PREPROCESSED DATA (OPTIONAL):

- If you want to save the preprocessed data for future use, you can save it to a new CSV file.

```
df_preprocessed = pd.DataFrame(df_scaled, columns=selected_columns)

df_preprocessed.to_csv('mall_customers_preprocessed.csv', index=False)
```

OUTPUT:



The steps mentioned above provide a general outline for preprocessing the Mall Customers dataset. Keep in mind that the specific preprocessing steps can vary based on your analysis goals and the nature of the dataset. Additionally, you can add further preprocessing steps, such as handling outliers or creating new features, depending on your project requirements.

3.4 PERFORMING DIFFERENT ANALYSIS NEEDED

1.DESCRIPTIVE STATISTICS:

- Calculate summary statistics such as mean, median, standard deviation, and percentiles for features like 'Age,' 'Annual Income,' and 'Spending Score.'

2.DATA VISUALIZATION:

- Create various types of plots and visualizations to explore the data. For example:

Histograms to visualize the distribution of features.

Scatter plots to explore the relationship between 'Annual Income' and 'Spending Score.'

Box plots to identify outliers in the data.

Pair plots or correlation matrices to understand the relationships between different features.

3.CUSTOMER SEGMENTATION**:**

- Implement customer segmentation using clustering algorithms like K-Means, Hierarchical Clustering, or DBSCAN to group customers into different segments based on 'Annual Income' and 'Spending Score.'

Visualize the segments to understand their characteristics.

4.EXPLORATORY DATA ANALYSIS (EDA):

- Conduct EDA to uncover patterns or trends in the data. For example, you can explore the distribution of customers by age and gender or identify correlations between variables.

5.CUSTOMER PROFILES:

- Create customer profiles or personas based on common attributes or behaviors. This can help in tailoring marketing strategies.

6.RFM ANALYSIS:

- Perform RFM (Recency, Frequency, Monetary) analysis to segment customers based on their shopping behavior.

Identify high-value customers or those who might need re-engagement.

7.MARKET BASKET ANALYSIS:

- Analyze product associations by examining which products are frequently purchased together. This can help with product placement and recommendations.

8.CUSTOMER CHURN ANALYSIS:

- Analyze customer churn by tracking changes in customer behavior over time. Identify factors that lead to customers leaving and take steps to retain them.

9.PREDICTIVE MODELING:

- Build predictive models to forecast customer behavior, such as predicting future spending based on historical data or predicting customer churn.

10.CUSTOMER SATISFACTION ANALYSIS:

- Collect and analyze customer feedback and satisfaction data to identify areas for improvement.

A/B Testing:

If applicable, conduct A/B tests on different marketing strategies, product offerings, or store layouts to evaluate their impact on customer behavior.

11.CUSTOMER LIFETIME VALUE (CLV) ANALYSIS:

- Calculate the CLV of each customer to understand their long-term value to the business. This can inform marketing and retention strategies.

12.GEOSPATIAL ANALYSIS (IF LOCATION DATA IS AVAILABLE):

- If the dataset contains location information, you can perform geospatial analysis to understand customer distribution and behavior by region.

13.TIME SERIES ANALYSIS (IF APPLICABLE):

- Analyze time-dependent data, such as customer visits or spending, over time to identify trends and seasonality.

14.CUSTOMER RETENTION ANALYSIS:

- Analyze customer retention rates and understand why some customers continue to visit the mall while others don't.

Remember that the choice of analysis depends on your specific business goals and questions you want to answer. Also, consider combining multiple types of analysis to gain a more comprehensive understanding of your customer data and make informed business decisions.

CHAPTER 4

PHASE 4

4.1: IN THIS TECHNOLOGY YOU WILL CONTINUE BUILDING YOUR PROJECT BY PREPROCESSING YOUR DATASET

4.1.1 PREPROCESS DATASET

1. IMPORT LIBRARIES:

- First, import the necessary libraries, including Pandas for data manipulation.

import pandas as pd

2. LOAD THE DATASET:

- Load the dataset from the CSV file. Make sure to download the dataset from Kaggle and place it

in your working directory.

3. EXPLORE THE DATASET:

- Explore the dataset to understand its structure, check for missing values, and review data types.

4. HANDLE MISSING VALUES:

- In this dataset, it's possible that there are no missing values. However, if there were any missing values, you'd need to decide how to handle them. Options include dropping rows with missing values, filling them with a default value, or using more advanced imputation techniques.

5. ENCODE CATEGORICAL DATA (IF ANY):

- The Mall Customers dataset doesn't contain categorical variables that need encoding. However, if

your dataset had categorical data (e.g., "Genre"), you'd need to encode it, typically using one-hot encoding or label encoding.

6. FEATURE SELECTION:

• Depending on your analysis goals, you may want to select a subset of features for segmentation.

7. STANDARDIZE/NORMALIZE DATA :

• If you're using clustering algorithms that rely on distances (e.g., K-Means), it's often a good practice to standardize or normalize the data to bring features to the same scale. This can be done using techniques like Min-Max scaling or Z-score standardization.

8. SAVE THE PREPROCESSED DATA :

• If you want to save the preprocessed data for future use, you can save it to a new CSV file.

PROGRAM:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("C:/Users/lokeshwar k/OneDrive/Documents/naan mudhalavan/Mall_Customers.csv")
print(df.head())
# Check the basic statistics of the dataset
print(df.describe())
# Check for missing values
print(df.isnull().sum())
# Check the data types of each column
print(df.dtypes)
plt.figure(figsize=(8, 6))
sns.histplot(df['Age'], bins=20, kde=True)
plt.title('Distribution of Age')
```

plt.xlabel('Age')

plt.ylabel('Count')

plt.show()

selected_columns = ['Annual Income (k$)', 'Spending Score (1-100)']

df = df[selected_columns]

from sklearn.preprocessing import StandardScaler

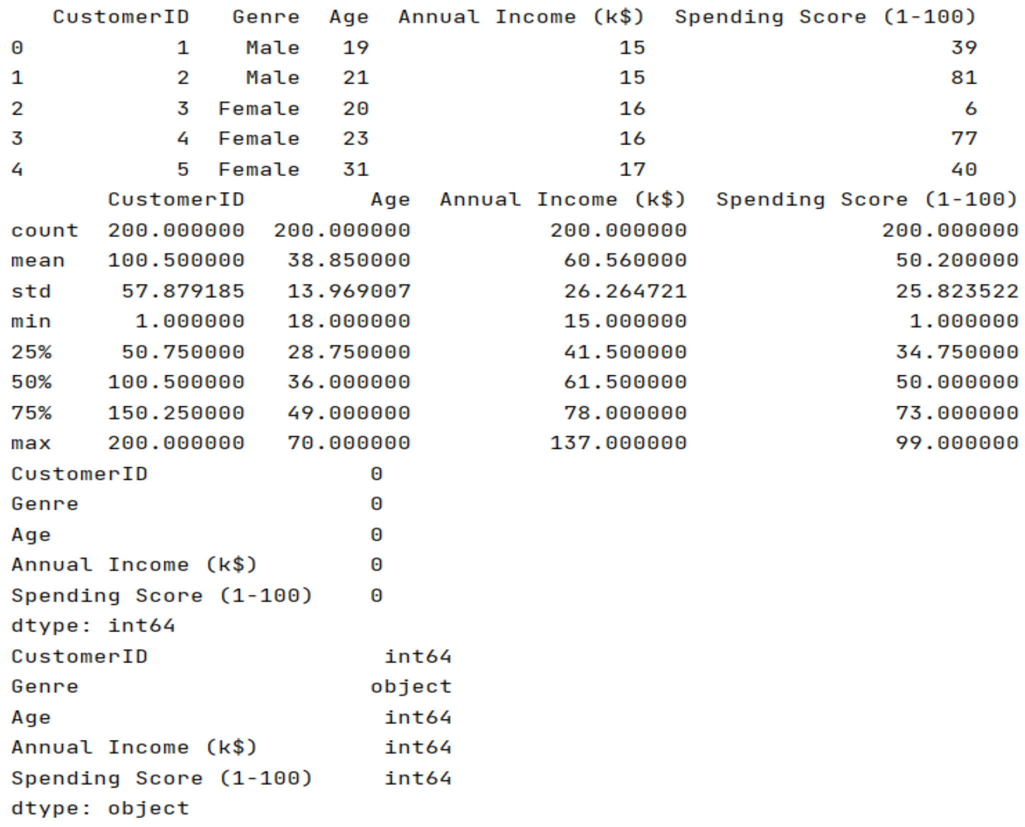
scaler = StandardScaler()

df_scaled = scaler.fit_transform(df)

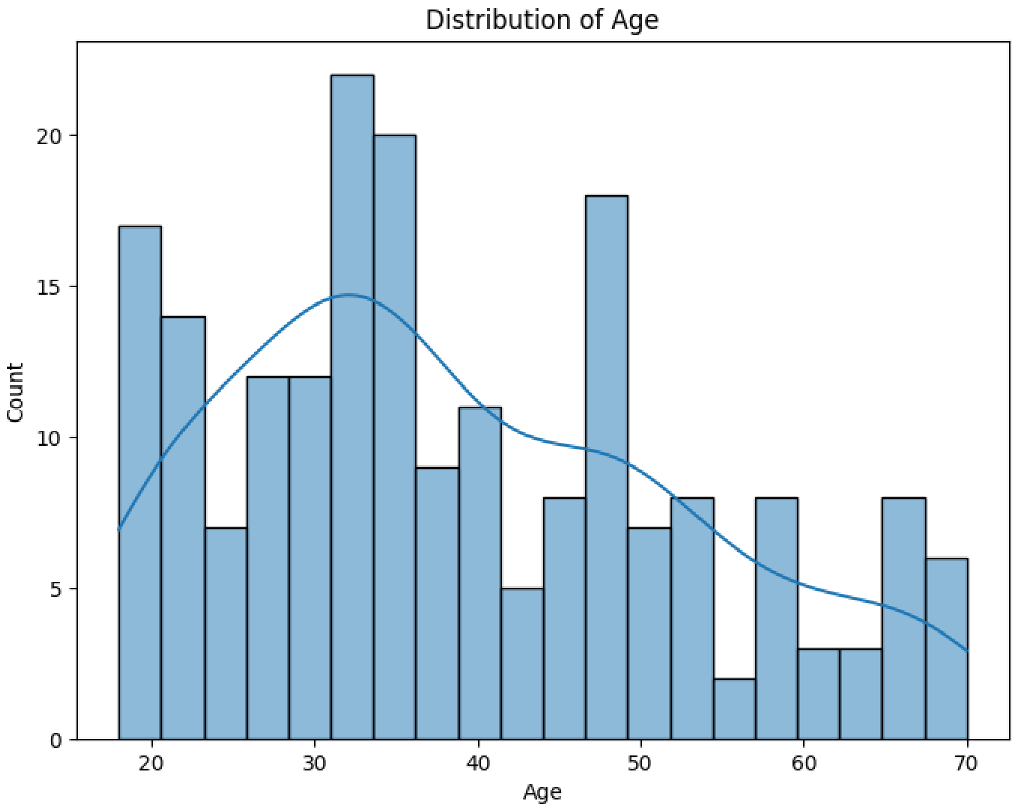
df_preprocessed = pd.DataFrame(df_scaled, columns=selected_columns)

df_preprocessed.to_csv('mall_customers_preprocessed.csv', index=False)

OUTPUT:



```
   CustomerID   Genre  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
CustomerID                0
Genre                     0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
CustomerID                 int64
Genre                     object
Age                        int64
Annual Income (k$)         int64
Spending Score (1-100)     int64
dtype: object
```



4.2: IN THIS TECHNOLOGY YOU WILL CONTINUE

BUILDING YOUR PROJECT BY PERFORMING FEATURE ENGINEERING

DATA COLLECTION:

- Start by collecting relevant data about mall customers.
- This data can come from sources such as customer surveys, transaction records, loyalty programs, and even demographic information.

DATA PREPROCESSING:

- Clean and preprocess the data to ensure its quality and consistency.
- This step includes handling missing values, outliers, and standardizing or normalizing the data.

FEATURE SELECTION/ENGINEERING:

- Identify the relevant features (attributes) that can be used for segmentation.
- These could include age, gender, income, shopping frequency, spending behavior, and more.
- You may also create new features if they are informative, such as customer age groups or spending categories.

EXPLORATORY DATA ANALYSIS (EDA):

- Use data visualization and summary statistics to gain insights into your customer data.
- Explore relationships between features and look for patterns and trends.

CHOOSE SEGMENTATION VARIABLES:

- Select the variables that you will use to segment customers.
- Common variables include:

- Demographics: Age, gender, income, marital status, etc.
- Behavioral: Purchase history, frequency of visits, average spending, etc.
- Psychographics: Lifestyle, preferences, and interests.
- Geographic: Location or proximity to the mall.

SELECT A SEGMENTATION METHOD:

- Choose an appropriate segmentation technique based on your data and objectives.
- Common methods include:

- K-Means Clustering: Groups customers into clusters based on similarity.
- Hierarchical Clustering: Builds a tree-like structure of customer segments.
- DBSCAN: Identifies dense regions of data points as clusters.
- PCA (Principal Component Analysis): Reduces dimensionality for better visualization and interpretation.

- Machine Learning Algorithms: Use supervised learning to predict customer segments.

SEGMENTATION MODELING:

- Implement the chosen segmentation method.
- This process assigns each customer to a specific segment or cluster based on the variables you selected.

EVALUATE AND INTERPRET SEGMENTS:

- Analyze the characteristics of each segment.
- What distinguishes one segment from another?
- Do they have unique preferences, behaviors, or needs?
- Use visualizations and descriptive statistics to understand the segments.

VALIDATION AND REFINEMENT:

- Validate the quality of your segments using techniques like silhouette score (for clustering) or cross-validation (for machine learning models).
- If necessary, refine your segmentation based on validation results.

APPLICATION:

- Apply the customer segments to marketing strategies, product recommendations, store layout, and other aspects of mall management.
- Tailor your approach to the specific needs and preferences of each segment.

CONTINUOUS MONITORING:

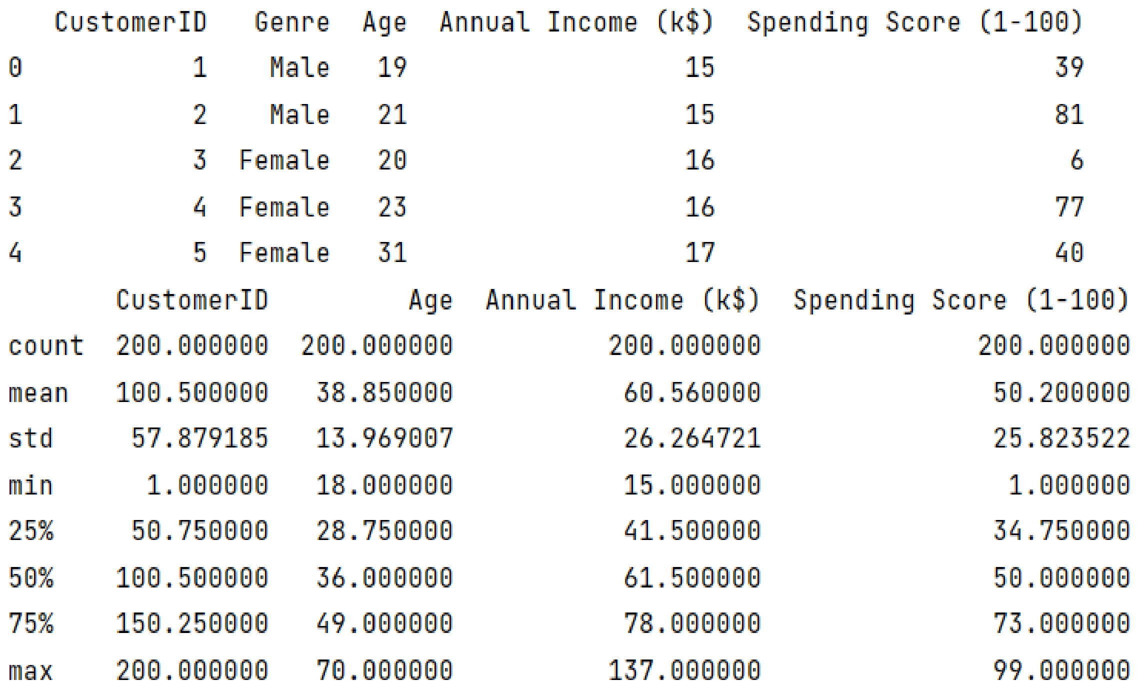
- Customer preferences and behaviors may change over time.
- Continuously monitor and update your segmentation as needed to ensure its relevance.

Customer segmentation in mall management is an iterative process that can yield valuable insights and improve the overall shopping experience for customers while increasing the mall's profitability. Data science and machine learning techniques help make this process more data-driven and effective.

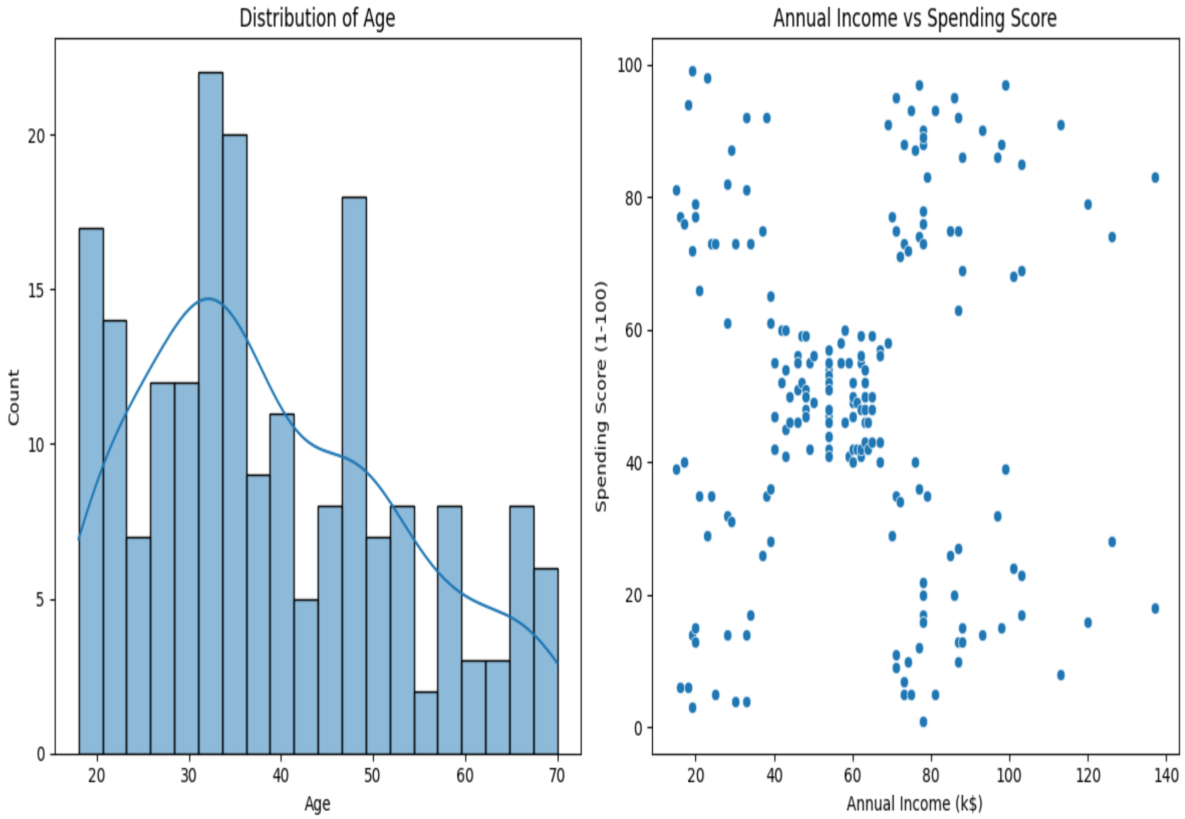
PROGRAM:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

df = pd.read_csv("C:/Users/lokeshwar k/OneDrive/Documents/naan mudhalavan/Mall_Customers.csv")

# Basic data exploration

print(df.head())

print(df.describe())

# Data visualization

plt.figure(figsize=(12, 6))

# Plot a histogram of 'Age' with a density curve

plt.subplot(1, 2, 1)

sns.histplot(df['Age'], bins=20, kde=True)

plt.title('Distribution of Age')

plt.xlabel('Age')

# Plot a scatterplot of 'Annual Income' vs 'Spending Score'

plt.subplot(1, 2, 2)

sns.scatterplot(data=df, x='Annual Income (k$)', y='Spending Score (1-100)')

plt.title('Annual Income vs Spending Score')

plt.tight_layout()

plt.show()
```

OUTPUT:



```
   CustomerID   Genre  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
```



4.3:MODEL TRAINING AND EVALUATION

To build a model for the Mall Customers dataset from Kaggle, you can follow these general steps:

DATA PREPARATION:

- Download the dataset from the Kaggle link you provided.
- Load the dataset using a library like Pandas.
- Explore the dataset to understand its structure, including the columns and data types.

DATA CLEANING:

- Handle missing values if any.
- Convert categorical variables to numerical format through techniques like one-hot encoding.

FEATURE ENGINEERING:

- Create relevant features that can enhance the model's performance, as discussed in the previous response.

DATA SPLITTING:

- Split the dataset into training and testing sets. A common split is 70-30 or 80-20 for training and testing, respectively.

MODEL SELECTION:

- Choose a machine learning model suitable for the problem.
- For a mall customer dataset, you might consider clustering techniques like K-Means, hierarchical clustering, or even regression/classification models depending on the specific problem you want to solve.

MODEL TRAINING:

- Train the selected model on the training dataset.
- Use libraries like Scikit-Learn to implement the model.

MODEL EVALUATION:

- Evaluate the model's performance using appropriate metrics. For clustering models, you can use metrics like Silhouette Score or Davies-Bouldin Index. For regression or classification models, use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), accuracy, precision, recall, F1-score, etc.

HYPERPARAMETER TUNING (IF APPLICABLE):

- Optimize model hyperparameters to improve performance.
- You can use techniques like grid search or random search.

MODEL DEPLOYMENT (IF NEEDED):

- If the model performs well and is ready for production use, deploy it in your desired environment

Remember that the choice of the model and specific steps will depend on the problem you want to solve with this dataset. You might want to do further data analysis and consider different types of models, such as regression or classification, depending on your goals.

To perform different analyses on the Mall Customers dataset from Kaggle, you can use various data analysis techniques and tools. Here's a step-by-step guide for conducting different types of analysis on this dataset:

HERE'S A GENERAL EXAMPLE IN PYTHON FOR K-MEANS CLUSTERING:

Remember that the choice of the model and specific steps will depend on the problem you want to solve with this dataset. You might want to do further data analysis and consider different types of models, such as regression or classification, depending on your goals.

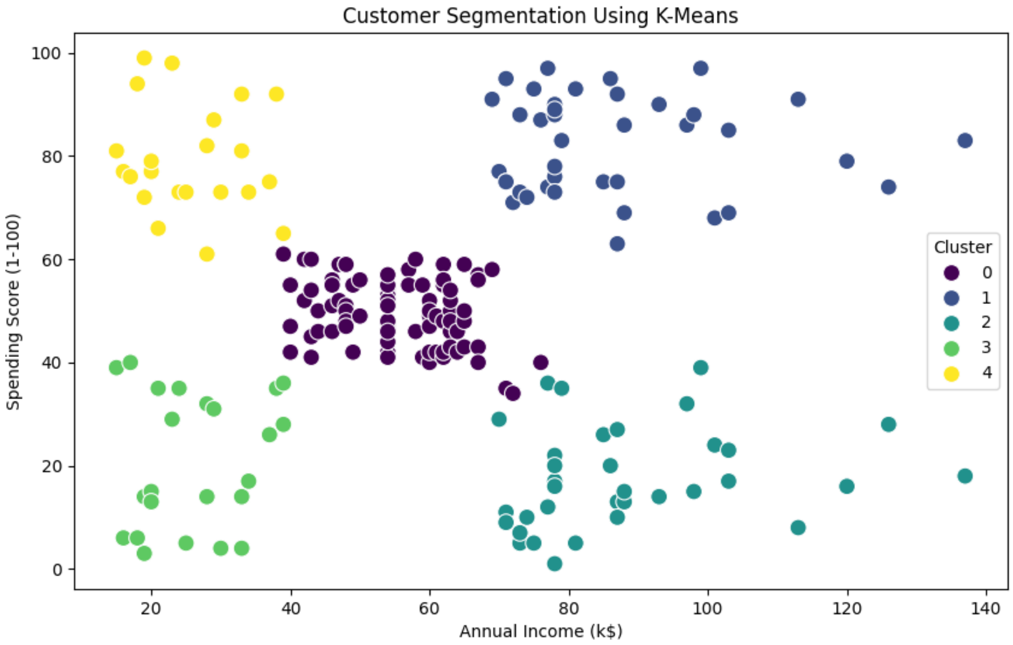
To perform different analyses on the Mall Customers dataset from Kaggle, you can use various data analysis techniques and tools. Here's a step-by-step guide for conducting different types of analysis on this dataset:

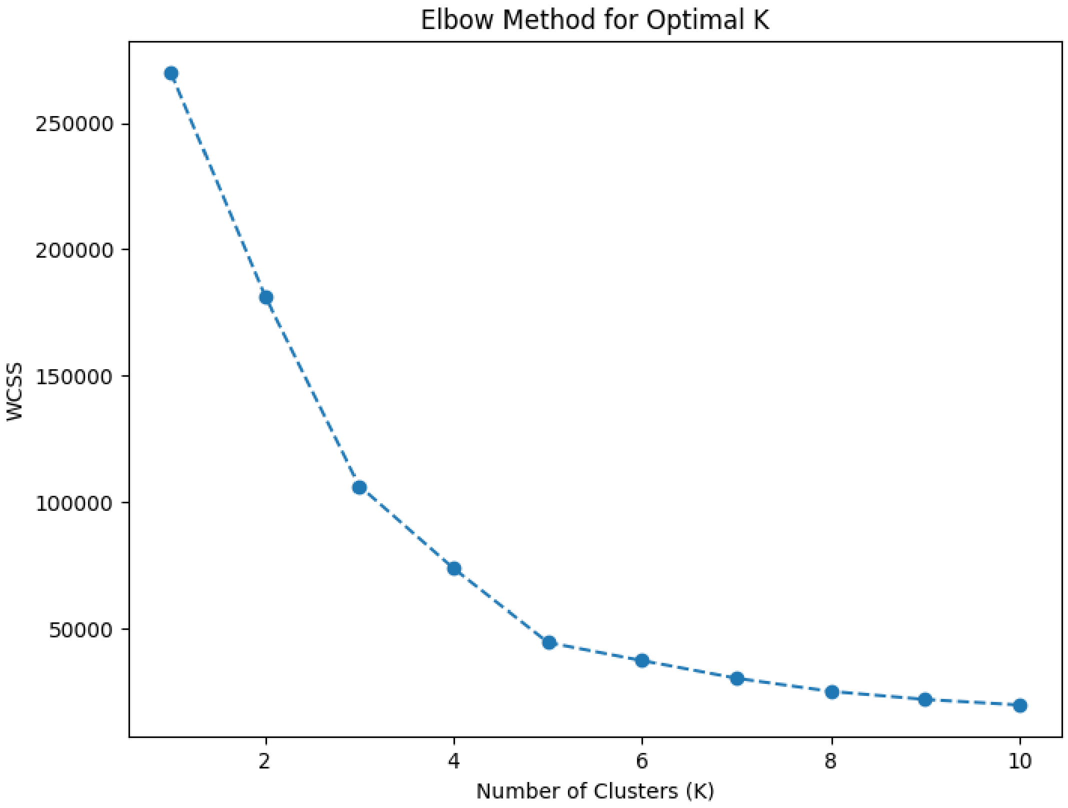
PROGRAM:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
# Load the dataset
df = pd.read_csv("C:/Users/lokeshwar k/OneDrive/Documents/naan mudhalavan/Mall_Customers.csv")
# Select the relevant features for clustering
X = df[['Annual Income (k$)', 'Spending Score (1-100)']]
# Determine the optimal number of clusters (K) using the Elbow Method
wcss = []  # Within-Cluster-Sum-of-Squares
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
# Plot the Elbow Method graph to find the optimal K
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method for Optimal K')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('WCSS')
plt.show()
# Based on the Elbow Method, let's choose K=5
n_clusters = 5
# Train the K-Means model
kmeans = KMeans(n_clusters=n_clusters, random_state=0)
df['Cluster'] = kmeans.fit_predict(X)
```

```
# Visualize the clusters
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Annual Income (k$)', y='Spending Score (1-100)', hue='Cluster',
palette='viridis', s=100)
plt.title('Customer Segmentation Using K-Means')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.show()
# Evaluate the clustering using silhouette score
silhouette_avg = silhouette_score(X, df['Cluster'])
print(f'Silhouette Score: {silhouette_avg:.2f}')
```

OUTPUT:





4.4: PERFORM DIFFERENT ANALYSIS AS NEEDED

DATA EXPLORATION:

- Load the dataset and examine its structure.
- Explore the summary statistics of numeric columns (e.g., age, income, and spending score) using Pandas.
- Visualize the data to gain insights, e.g., use histograms, box plots, or scatter plots to understand the distribution and relationships among variables.

CUSTOMER SEGMENTATION:

- Perform customer segmentation using clustering techniques like K-Means or Hierarchical Clustering to group customers based on similar characteristics.
- Analyze the resulting clusters to understand customer behavior.

DESCRIPTIVE STATISTICS:

- Calculate and analyze descriptive statistics for various customer groups. For example, compute the mean, median, and standard deviation of spending scores for different segments.

CUSTOMER PROFILING:

- Create customer profiles or personas by summarizing the characteristics of each customer group, such as average age, income, and spending score.

DATA VISUALIZATION:

- Use data visualization libraries like Matplotlib or Seaborn to create informative charts and graphs.
- Visualize the distribution of spending scores, income, and other relevant features across customer segments.

CORRELATION ANALYSIS:

- Explore the correlations between different features. Determine whether there are any strong relationships between variables like age, income, and spending score.

HYPOTHESIS TESTING:

- Formulate hypotheses and conduct statistical tests to confirm or reject these hypotheses. For example, you could test whether there's a significant difference in spending scores between male and female customers.

TIME SERIES ANALYSIS (IF APPLICABLE):

- If the dataset contains timestamp data, perform time series analysis to identify patterns and trends in customer behavior over time.

MACHINE LEARNING PREDICTIVE MODELING (IF NEEDED):

- Build regression or classification models to predict customer behavior, such as spending score, based on other variables.
- Evaluate model performance using appropriate metrics like Mean Squared Error (MSE) for regression or accuracy for classification.

CUSTOMER CHURN ANALYSIS (IF APPLICABLE):

- Analyze customer churn by identifying customers who have stopped shopping at the mall.
- Create churn prediction models to identify customers at risk of leaving.

MARKET BASKET ANALYSIS (IF APPLICABLE):

- If you have transaction data, conduct market basket analysis to identify which products or services are frequently purchased together.

CUSTOMER RETENTION STRATEGIES:

- Based on your analyses, develop customer retention strategies to improve customer satisfaction and loyalty.

To conduct these analyses, you can use Python libraries like Pandas, Matplotlib, Seaborn, Scikit-Learn, and statsmodels for data analysis, visualization, and statistical testing. Additionally, you may use Jupyter notebooks to document your analysis steps and findings.

Keep in mind that the specific analyses you perform should align with your business objectives and the questions you aim to answer using the dataset. The results of your analyses can inform marketing strategies, customer targeting, and business decisions for the mall.

REFERENCE

1. DATA COLLECTION:

Gather relevant data on mall customers. This may include demographic information, purchase history, visit frequency, etc.

You can collect data through surveys, loyalty programs, or other sources.

2. DATA PREPROCESSING:

Clean and preprocess the data to handle missing values, outliers, and ensure data consistency.

3. FEATURE SELECTION/ENGINEERING:

Identify and select relevant features for segmentation.

Create new features that can aid in customer segmentation, such as RFM (Recency, Frequency, Monetary) metrics.

4. EXPLORATORY DATA ANALYSIS (EDA):

Perform EDA to gain insights into the data.

Visualize data to understand customer behavior.

5. CUSTOMER SEGMENTATION TECHNIQUES:

Apply various data science techniques for customer segmentation, such as:

K-Means Clustering: Divide customers into groups based on similarities in their features.

Hierarchical Clustering: Create a hierarchy of clusters.

DBSCAN: Identify clusters with varying shapes and densities.

PCA (Principal Component Analysis): Reduce dimensionality for visualization and clustering.

6. MODEL EVALUATION:

Evaluate the quality of the segmentation. Common metrics include silhouette score, Dunn index, or domain-specific metrics.

7. CUSTOMER PROFILING:

Profile each segment to understand the characteristics and behaviors of the customers in each group.

8. BUSINESS INSIGHTS:

Provide actionable insights to mall management based on the segmentation.

Suggest marketing and service strategies for each customer segment.

9. VISUALIZATION:

Create visualizations to communicate the results effectively to non-technical stakeholders.

10. IMPLEMENTATION:

If applicable, implement the recommended strategies and monitor their impact.

11. DOCUMENTATION:

Document the entire project, including data sources, methods, and results.

12. PRESENTATION:

Prepare a presentation to share with stakeholders.

13. FUTURE WORK:

Discuss potential future work, such as using predictive modeling for customer behavior.

For project references, you can search for customer segmentation case studies or projects on platforms like Kaggle, GitHub, or research papers. While I can't provide specific project references beyond my last knowledge update in January 2022, a search on those platforms for"mall customer segmentation" or related keywords should yield relevant projects that you can study for inspiration and guidance.