|  |
| --- |
| **A Report on**  MALL CUSTOMER SEGMENTATION  Submitted for partial fulfillment of award of  **DEGREE**  **OF**  **BACHELOR OF COMPUTER APPLICATIONS**  Submitted By  RUDRAKSH KAUSHIK  Under supervision  Dr. Rohit Kumar  BCA Dept., ITS  C:\Users\aa\AppData\Local\Temp\samagra-2017.png    **INSTITUTE OF TECHNOLOGY & SCIENCE**  **MOHAN NAGAR, GHAZIABAD**  **Batch: 2021-2024**  **CERTIFICATE** This is to certify that RUDRAKSH KAUSHIK has carried out the project work presented in this report entitled “Mall Customer segmentation” for the award of Bachelor Of Computer Applications from Institute of Technology & Science, Mohan Nagar, Ghaziabad, under my supervision. The report embodies result of original work and studies carried out by Student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else. Date : Dr. Rohit Kumar  ( BCA Dept., ITS) |

#### ACKNOWLEDGEMENT

I extend my heartfelt appreciation to all the individuals and organizations who have been a part of this college project and have contributed to its successful completion. I would like to express my gratitude to my project supervisor, trainer Dr. Rohit kumar for their guidance, expertise, and constant support throughout the project duration

.Furthermore, I would like to acknowledge the support received from ITS for granting permission to conduct the research and providing necessary resources. Lastly, I would like to thank my friends and family for their unwavering support and encouragement during this journey.

Name - RUDRAKSH KAUSHIK

Roll No - 210934106262

Signature:-

ABSTRACT

A lot of customers buy products from the mall and to generate more revenue for the mall, the authorities need to attract these customers and for this large amount of capital is required. After the advertisement, the output is only around 30-40%. Hence customer segmentation comes into

the picture.

Customer Segmentation is a popular application of unsupervised learning and by using this technique we'll only focus on the potential customers (customers whose probability of buying the product is very high). With this technique, the output will drastically increase to 90-95%.

# 

# TABLE OF CONTENTS

* **Chapter 1: Introduction …………………………………… 8**

• 1.1 Overview…………………………………………………….. **9**

• 1.2 Problem statement…………………………………………… **9**

• 1.3 Scope………………………………………………………… **10**

• 1.4 Purpose………………………………………………………. **10**

• 1.5 Tools and methodology……………………………………… **12**

* **Chapter 2: System analysis………………………..……….. 16**

• 2.1 Identification…………………………………………………

• 2.2 Preliminary investigation…………………………………….

* **Chapter 3: Means of Project…………………….………....**

• 3.1 Hardware require…………………………………………….

• 3.2 Software requirement………………………………………..

• 3.3 Product Perspective…………………………………………

* **Chapter 4: Overall description………………….…………**

• 4.1 Product Perspective………………………………………….

* **Chapter 5: Specific Diagrams……………………..………**

• 5.1 Dataset………………………………………………………

• 5.2 Implementation and analysis………………………………..

• 5.3 Age plot analysis……………………………………………

• 5.4 Age vs spending score analysis……………………………..

• 5.5 Annual income vs spending score analysis………………....

* **Chapter 6: Product and project features……………..……**

• 6.1 Advantages of customer segmentation………………………

• 6.2 Why use customer segmentation?...........................................

• 6.3 Why use k-means clustering for customer segmentation?......

* **Chapter 7: Code and Output………………………...………**

• 7.1 Code…………………………………………………………...

• 7.2 Output........................................................................................

* **Chapter 8: Test cases and hypothesis problem…………....**

• 8.1 How to find optimal number of clusters?................................

* **Chapters 9: Conclusion……………………………………..**

• 9.1 Clustering analysis……………………………………………

* **Chapter 10: Future Scope of the Project……..…………….**

* **Chapter 11: Bibliography……………………………….….**

**LIST OF TABLE**

5.1 Snapshot of dataset

8.1 Customer data

**LIST OF IMAGE**

5.2 Gender plot

5.3 Age plot

5.4 Age vs spending score

5.5 Annual income vs spending score

8.3 Aptimal cluster

9.1 Annual income vs spending score after clustering

# CHAPTER 1: INTRODUCTION

* 1. **Overview and Problem statement**

Mall customer segmentation is the process of dividing mall customers into groups based on their similarities and differences. Clustering algorithms are a type of unsupervised machine learning algorithm that can be used to segment customers into distinct groups. These algorithms work by grouping customers based on their features, such as age, annual income, and spending score.

**Problem Statement :**

Problem Statement: You own a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income, and spending score. The problem is to understand the customers like who can be easily converged (Target Customers) so that the sense can be given to the marketing team and plan the strategy accordingly.

In other words, the goal is to identify the key customer groups that present the greatest potential for driving sales growth by leveraging the available data and utilizing a spending score metric. This will empower the mall's marketing department to tailor their strategies effectively, enhancing customer engagement and optimizing revenue generation.

**1.2 Purpose**

Data analysis helps businesses acquire relevant, accurate information, suitable for developing future marketing strategies, business plans, and realigning the company’s vision or mission.This project aims to perform customer segmentation on a Mall customer dataset using the K-Means clustering algorithm. The goal of this project is to cluster the customers based on their purchasing behavior and demographic characteristics.

**1.3 Scope**

The scope of a system in mall customer segmentation using clustering involves identifying and understanding the target customer groups based on their characteristics, behavior, and preferences. The system aims to:

**1. Customer Segmentation:** Divide the customer base into distinct groups based on demographic, geographic, psychographic, and behavioral characteristics.

**2. Clustering Analysis**: Apply clustering algorithms, such as K-Means, to identify patterns and relationships within the customer data.

**3. Target Customer Identification:** Identify the most profitable and valuable customer segments to target with marketing strategies.

**4. Marketing Strategy Development:** Develop customized marketing campaigns, product offerings, and pricing strategies tailored to each target customer segment.

**5. Customer Retention and Loyalty**: Implement strategies to retain and increase customer loyalty within each target segment.

**6. Data Analysis and Visualization:** Analyze and visualize customer data to gain insights into their behavior, preferences, and needs.

**7. System Integration**: Integrate the customer segmentation system with existing systems, such as CRM, ERP, and marketing automation platforms.

The system will analyze customer data to identify distinct customer segments, enabling mall management to tailor marketing strategies, optimize store placements, and enhance customer experience. The scope includes data collection, preprocessing, segmentation analysis, and generating actionable insights.

**1.4 Tools used:**

**Purpose:** The primary language used for scripting and implementing the entire project. It is chosen for its rich ecosystem of libraries and ease of use for data analysis and machine learning tasks.

Jupyter Notebook:

**Purpose:** An interactive development environment that allows for writing and running Python code in a cell-based format. It is ideal for iterative data analysis, visualization, and documenting the workflow.

Pandas:

**Purpose:** A powerful library for data manipulation and analysis. It provides data structures like DataFrames to handle structured data efficiently.

**Usage:** Loading the dataset, cleaning and preprocessing data, exploratory data analysis (EDA), and feature engineering.

NumPy:

**Purpose:** A library for numerical computing in Python. It provides support for arrays and matrices, along with mathematical functions to operate on these data structures.

**Usage:** Numerical operations and handling of large datasets during preprocessing and feature engineering.

Scikit-learn:

**Purpose:** A machine learning library that provides simple and efficient tools for data mining and data analysis.

**Usage:** Implementing clustering algorithms such as K-Means, Hierarchical Clustering, and DBSCAN. It also includes utilities for evaluating the performance of clustering methods (e.g., Elbow Method, Silhouette Analysis).

Matplotlib:

**Purpose:** A plotting library for creating static, interactive, and animated visualizations in Python.

**Usage:** Creating basic visualizations like histograms, scatter plots, and line graphs to explore data and visualize clustering results.

Seaborn:

**Purpose:** A Python visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.

**Usage:** Enhanced visualizations for exploratory data analysis and cluster visualization, such as pair plots, box plots, and heatmaps.

Plotly:

**Purpose**: A graphing library that makes interactive, publication-quality graphs online.

**Usage:** Creating interactive visualizations to explore data and present clustering results in a more engaging manner.

SciPy:

**Purpose**: A library used for scientific and technical computing in Python.

**Usage:** Performing hierarchical clustering and other advanced mathematical operations.

Yellowbrick:

**Purpose:** A mac hine learning visualization library that extends the Scikit-learn API with visual analysis and diagnostic tools.

**Usage:** Visual diagnostics for clustering evaluation, such as Silhouette Visualizer and Elbow Method.

Workflow and Application of Tools:

Data Loading and Preprocessing:

Use Pandas to load the dataset.

Handle missing values and data cleaning with Pandas and NumPy.

Exploratory Data Analysis (EDA):

Use Pandas for summary statistics and initial exploration.

Create visualizations with Matplotlib and Seaborn to understand data distributions and relationships.

Feature Engineering:

Use Pandas and NumPy to create new features or transform existing ones.

Normalize or standardize data using Scikit-learn preprocessing tools.

Clustering:

Implement clustering algorithms like K-Means, Hierarchical Clustering, and DBSCAN using Scikit-learn and SciPy.

Determine the optimal number of clusters with the Elbow Method and Silhouette Analysis using Scikit-learn and Yellowbrick.

Visualization of Clusters:

Create static plots with Matplotlib and Seaborn to visualize clusters.

Use Plotly for interactive visualizations to explore and present the clustering results.

Reporting:

Compile findings and visualizations into a comprehensive report using Jupyter Notebook markdown cells and visual output.

**1.5 Methodology used:**

Methodology for Mall Customer Segmentation Using Python

The methodology for a mall customer segmentation project involves several systematic steps, from data collection and preprocessing to clustering and evaluation. Here’s a detailed breakdown of each step in the methodology:

**1. Data Collection**

* Objective: Gather relevant data for customer segmentation.
* Data Sources: Typically, the dataset includes demographic information and purchasing behavior, such as:
* CustomerID: Unique identifier for each customer.
* Gender: Gender of the customer.
* Age: Age of the customer.
* Annual Income: Annual income of the customer.
* Spending Score: A score assigned by the mall based on customer behavior and spending patterns.

**2. Data Exploration and Preprocessing**

* Objective: Understand the dataset and prepare it for analysis.
* Steps:
* Load the Data: Use Pandas to load the dataset into a DataFrame.
* Explore the Data: Perform initial exploration to understand data structure, types, and basic statistics.
* Handle Missing Values: Identify and handle missing values through imputation or removal.
* Da ta Cleaning: Correct or remove inconsistent, duplicate, or irrelevant data.
* Feature Engineering: Create new features if needed and transform existing ones.
* Normalization/Standardization: Scale the data to ensure all features contribute equally to the clustering algorithm.

**3. Exploratory Data Analysis (EDA)**

* Objective: Gain insights into the dataset and visualize relationships between variables.
* Summary Statistics: Use Pandas to get summary statistics of the data.
* Visualization: Use Matplotlib and Seaborn to create visualizations like histograms, scatter plots, box plots, and pair plots to understand distributions and correlations.
* python

***In this project, the prescribed sequence is:***

* + Creating an approach to solve the given problem statement
  + Exploring the dataset and obtaining useful insight from the same
  + Cleaning the dataset by handling nan values, remove duplicate records, etc.
  + Data Visualization used to obtain important information from the data
  + Data Preprocessing is performed to make the data ready to fit the model this includes
  + feature scaling, splitting the dataset into features and labels, etc.
  + Model Building

**1.6 Technology used:**

To implement a mall customer segmentation project using clustering techniques in Python, several technologies and libraries are used to handle data processing, analysis, clustering, and visualization. Below is a detailed list of the technologies and their roles:

1. **Python Programming Language**

* **Purpose:** The primary language used for scripting and implementing the entire project. Python is chosen for its simplicity, extensive library support, and active community.
* **Features:** Versatility, ease of use, and a rich ecosystem of data science and machine learning libraries.

1. **Jupyter Notebook**

* **Purpose:** An interactive development environment that allows for writing and running Python code in a cell-based format. Ideal for iterative data analysis, visualization, and documentation.
* **Features:** Interactive coding, inline visualization, and easy-to-share notebooks.

1. **Data Manipulation and Analysis**

* **Pandas:**
* **Purpose:** A powerful library for data manipulation and analysis. It provides data structures like DataFrames to handle structured data efficiently.
* **Usage:** Loading the dataset, cleaning and preprocessing data, exploratory data analysis (EDA), and feature engineering.

**Installation:** **pip install pandas**

* **NumPy:**
* **Purpose:** A library for numerical computing in Python. It provides support for arrays and matrices, along with mathematical functions to operate on these data structures.
* **Usage:** Numerical operations and handling large datasets during preprocessing and feature engineering.

**Installation:** **pip install numpy**

1. **Data Visualization**

* **Matplotlib:**
* **Purpose:** A plotting library for creating static, interactive, and animated visualizations in Python.
* **Usage:** Creating basic visualizations like histograms, scatter plots, and line graphs to explore data and visualize clustering results.

**Installation:** **pip install matplotlib**

* **Seaborn:**
* **Purpose:** A Python visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.
* **Usage:** Enhanced visualizations for exploratory data analysis and cluster visualization, such as pair plots, box plots, and heatmaps.

**Installation:** **pip install seaborn**

* **Plotly:**
* **Purpose:** A graphing library that makes interactive, publication-quality graphs online.
* **Usage:** Creating interactive visualizations to explore data and present clustering results in a more engaging manner.

1. **Machine Learning and Clustering**

* **Scikit-learn:**
* **Purpose:** A machine learning library that provides simple and efficient tools for data mining and data analysis.
* **Usage:** Implementing clustering algorithms such as K-Means, Hierarchical Clustering, and DBSCAN. Includes utilities for evaluating the performance of clustering methods (e.g., Elbow Method, Silhouette Analysis).

**Installation:** **pip install scikit-learn**

**CHAPTER 2: SYSTEM ANALYSIS**

**2.1 Identification of need**

***About Python :***

Python is a dynamic, interpreted (bytecode-compiled) language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code short and flexible, and you lose the compile-time type checking of the source code. Python tracks the types of all values at runtime and flags code that does not make sense as it runs..

Identify details like age, geographical location, relationship status, and income for B2C customers. For B2B customers, identify details like company revenue, industry, products and services, and the technologies they use.

**2.2 Preliminary investigation**

The basic purpose behind Preliminary Investigation is to first clarify, understand and evaluate the Project Request. Our preliminary investigation started off in two ways i.e observing the visting patterns of customers that happened on site. And by asking people about their monthly income and expenses to understand them better.

Preliminary Investigation basically refers to the collection of information that guides the management of an organization to evaluate the merits and demerits of the project request and make an informed judgment about the feasibility of the proposed system.

***On site Observation***

Here a detailed study was carried out, checking the existing visitors to understand the patterns effectively.This information helped us to understand how the project should operate. But after interviewing the persons, we got more details that further explain the project and shown whether assistance is merited economically, operationally and technically.

***Conducting Interviews***

This method of investigation conducted by us involved questioning the concerned personnel to get the user’s (client) view about the project and the features they desired it to have. Some of the questions we asked are:

a) The number of customers regularly visiting their stores

b) The growth of Crowd with changing times and conditions

c) The level of money expenses made in a month acculumated sums up to the profit.

**CHAPTER 3: MEANS OF PROJECT**

**3.1 Hardware Requirements**

For the mall customer segmentation project using Python clustering, the hardware requirements are relatively minimal. However, having a sufficiently powerful machine can help in processing large datasets and running clustering algorithms efficiently. The typical hardware requirements include:

* **RAM:** 8 GB or higher
* **Reason:** Adequate memory is necessary to load and process data in memory, especially for larger datasets.
* **Storage:** 256 GB SSD or higher
* **Reason:** Fast storage improves the performance of data read/write operations. An SSD is preferred over an HDD for quicker data access and improved overall system responsiveness.

**3.2 Software Requirements**

The software requirements include the operating system, development environment, and specific libraries needed for data analysis, visualization, and clustering.

***Operating System***

* **Windows 10/11, macOS, or Linux (Ubuntu preferred)**
* **Reason:** Python and its associated libraries are cross-platform and can run on any major operating system. Ubuntu is preferred for its compatibility with many data science tools and ease of setup for development environments.

***Python***

* **Version:** Python 3.8 or higher
* **Reason:** Python 3.8+ ensures compatibility with the latest libraries and features in data science and machine learning.

***Integrated Development Environment (IDE)***

* **Jupyter Notebook:**
* **Reason:** An interactive development environment that allows for writing and running Python code in a cell-based format. Ideal for data analysis, visualization, and documenting the workflow.

***Installation:******pip install notebook***

* **Alternative IDEs:** PyCharm, VS Code
* **Reason:** Advanced IDEs provide robust features for coding, debugging, and project management.

***Python Libraries***

* + **Pandas:**
    - **Purpose:** Data manipulation and analysis
    - **Installation:** **pip install pandas**
  + **NumPy:**
    - **Purpose:** Numerical computing
    - **Installation:** **pip install numpy**
  + **Matplotlib:**
    - **Purpose:** Data visualization
    - **Installation:** **pip install matplotlib**
  + **Seaborn:**
    - **Purpose:** Statistical data visualization
    - **Installation:** **pip install seaborn**
  + **Plotly:**
    - **Purpose:** Interactive data visualization
    - **Installation:** **pip install plotly**
  + **Scikit-learn:**
    - **Purpose:** Machine learning, including clustering algorithms
    - **Installation:** **pip install scikit-learn**

**3.3 Product Perspective**

The mall customer segmentation project aims to provide insights into customer behavior and preferences by segmenting customers into distinct groups based on their demographic and purchasing patterns. The project leverages clustering algorithms to identify these segments, which can be utilized by the mall management for targeted marketing and improved customer service. Here is an overview of the product perspective:

* **Input:**
  + Customer demographic data (e.g., age, gender, annual income)
  + Customer purchasing behavior data (e.g., spending score)
* **Processing:**
  + Data preprocessing: Cleaning, normalization, and transformation of input data
  + Clustering: Application of clustering algorithms (e.g., K-Means, Hierarchical Clustering) to segment customers
* **Output:**
  + Identified customer segments with distinct characteristics
  + Visualizations of customer segments (e.g., scatter plots, 3D plots)
  + Insights and recommendations for targeted marketing strategies
* **Benefits:**

**For Mall Management:**

* + - Improved understanding of customer demographics and behaviors
    - Ability to design targeted marketing campaigns and personalized offers
    - Enhanced customer experience and satisfaction

**For Customers:**

* + - More personalized shopping experiences
    - Relevant promotions and offers based on their preferences
* **Deployment:**
  + The final product, including the code, visualizations, and reports, can be deployed on a local machine or a cloud-based environment for easy access and sharing with stakeholders.
* **Scalability:**
  + The methodology and code can be easily scaled to handle larger datasets and more complex clustering scenarios as the mall expands and collects more customer data.

By following this structured approach, the mall customer segmentation project can effectively deliver valuable insights to mall management, helping them optimize their marketing strategies and improve overall customer satisfaction.

Top of Form

Bottom of Form

**CHAPTER 4: OVERALL DESCRIPTION**

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support.

According to Clustering techniques consider data tuples as objects. They partition the data objects into groups or clusters, 2 so that objects within a cluster are similar to one another and dissimilar to objects in other clusters. Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

**4.1 Product Perspective :**

The overall product that the study deals with in the project report is the concept of people spending their resources. The existing method is storing customer data through paperwork and computer software (digital data) is increasing day by day. At end of the day they will analyse their data as how many things are sold or actual customer count etc. By analysing the collected data they got to know who is beneficial to their business and increase their sales. It requires more time and more paperwork. Also, it is not much effective solution to find the desired customers data.

Customer segmentation simply means grouping your customers according to various characteristics (for example grouping customers by age).

It’s a way for organizations to understand their customers. Knowing the differences between customer groups, it’s easier to make strategic decisions regarding product growth and marketing. The opportunities to segment are endless and depend mainly on how much customer data you have at your use. Starting from the basic criteria, like gender, hobby, or age, it goes all the way to things like “time spent of website X” or “time since user opened our app”.

**CHAPTER 5: SPECIFIC DIAGRAMS AND IMPLEMENTATION**

**5.1 Dataset**

The dataset name is ‘Mall\_Customers.csv’ consists of 5 columns which are CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100) where Gender is a categorical value and rest all features are numeric.

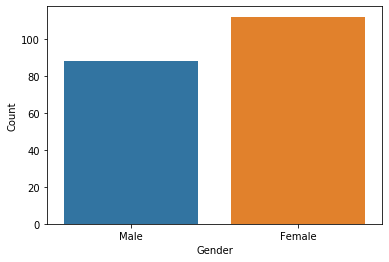


***5.1 Snapshot of Dataset***

**5.2 Implementation and analysis**

On performing data visualization on the dataset, a lot of insights were obtained such as:

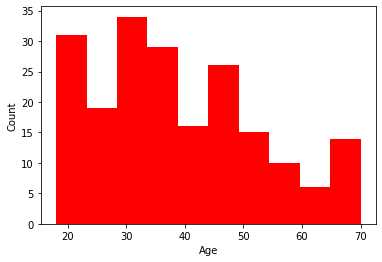
***Gender Plot Analysis :***



***5.2 Gender Plot***

From the Count plot, it is observed that the number of Female customers is more than the total number of Male customers.

**5.3 Age Plot Analysis**

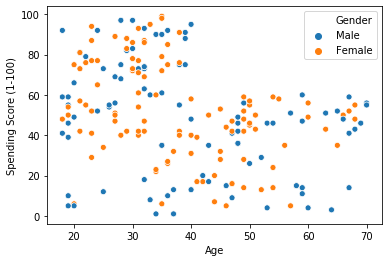


***5.3 Age Plot Analysis***

From the Histogram it is evident that there are 3 age groups that are more frequently shop at the mall, they are: 15-22 years, 30-40 years, and 45-50 years.

**5.4 Age vs Spending score analysis**

* + From the Age Vs Spending Score plot we observe that customers whose spending score is more than 65 have their Age in the range of 15-42 years. Also from the Scatter plot it is observed that customers whose spending score is more than 65 consists of more Females than Males.
  + The customers having average spending score ie: in the range of 40-60 consists of the age group of the range 15-75 years and the count of males and females in this age group is also



***5.4 Age Vs Spending Sore***

**5.5 Annual Income Vs Spending Score Analysis**

We observe that there are 5 clusters and can be categorized as:

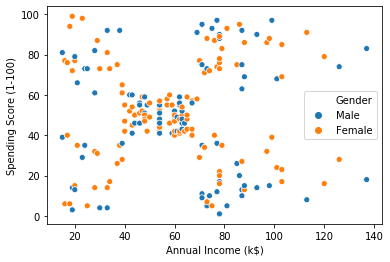
**a**) High Income, High Spending Score (Top Right Cluster)

**b**) High Income, Low Spending Score (Bottom Right Cluster)

**c**) Average Income, Average Spending Score (Center Cluster)

**d**) Low Income, High Spending Score (Top Left Cluster)

**e**) Low Income, Low Spending Score (Bottom Left Cluster)

****

***5.5 Annual Income Vs Spending score***

**Chapter 6: PRODUCT AND PROJECT FEATURES**

**6.1 Advantages of customer segmentation**

Implementing customer segmentation leads to plenty of new business opportunities. You can do a lot of optimization in:

**• Budgeting**

Nobody likes to invest in campaigns that don’t generate any new customers. Most companies don’t have huge marketing budgets, so that money has to be spent right. Segmentation enables you to target customers with the highest potential value first, so you get the most out of your marketing budget.

• **Product design**

Customer segmentation helps you understand what your users need. You can identify the most active users/customers, and optimize your application/offer towards their needs.

**• Promotion**

Properly implemented customer segmentation helps you plan special offers and deals. Frequent deals have become a staple of e-commerce and commercial software in the past few years. If you reach a customer with just the right offer, at the right time, there’s a huge chance they’re going to buy. Customer segmentation will help you tailor your special offers perfectly.

**• Marketing**

The marketing strategy can be directly improved with segmentation because you can plan personalized marketing campaigns for different customer segments, using the channels that they use the most.

**• Customer satisfaction**

By studying different customer groups, you learn what they value the most about your company. This information will help you create personalized products and services that perfectly fit your customers’ preferences.

**6.2 Why use customer segmentation??**

• **More time**

Manual customer segmentation is time-consuming. It takes months, even years to analyze piles of data and find patterns manually. Also if done heuristically, it may not have the accuracy to be useful as expected.

Customer segmentation used to be done manually and wasn’t too precise. You’d manually create and populating different data tables, and analyze the data like a detective with a looking glass. Now, it’s much better (and relatively easy thanks to rapid progress in ML) to just use machine learning, which can free up your time to focus on more demanding problems that require creativity to solve.

**• Ease of retraining**

Customer Segmentation is not a “develop once and use forever” type of project. Data is ever-changing, trends oscillate, everything keeps changing after your model is deployed. Usually, more labeled data becomes available after development, and it’s a great resource for improving the overall performance of your model.

There are many ways to update customer segmentation models, but here are the two main approaches:

✓ Use the old model as the starting point and retrain it.

✓ Keep the existing model and combine its output with a new model.

• **Better scaling**

Machine learning models deployed in production support scalability, thanks to cloud infrastructure. These models are quite flexible for future changes and feedback. For example, consider a company that has 10000 customers today, and they’ve implemented a customer segmentation model. After a year, if the company has 1 million customers, then ideally we don’t need to create a separate project to handle this increased data. Machine Learning models have the inherent capability to handle more data and scale in production.

**• Higher accuracy**

The value of an optimal number of clusters for given customer data is easy to find using machine learning methods like the elbow method. Not only the optimal number of clusters but also the performance of the model is far better when we use machine learning.

**6.3 Why use K-means clustering for customer segmentation?**

K-means clustering is a popular and widely used technique for customer segmentation due to its simplicity, scalability, and effectiveness. Here are some reasons why K-means clustering is a good choice for customer segmentation:

1. **Ease of Use**: K-means clustering is a simple and intuitive algorithm that is easy to implement and interpret. It requires minimal input parameters and can be applied to large datasets with ease.

2. **Scalability**: K-means clustering can handle large datasets with thousands or even millions of data points. This makes it a suitable choice for customer segmentation in large retail organizations with extensive customer databases.

3. **Fast Computation**: K-means clustering is a fast algorithm that can converge quickly, even for large datasets. This is important for customer segmentation, as it allows retailers to quickly identify customer segments and respond to changing market conditions.

4. **Effectiveness**: K-means clustering has been shown to be effective in identifying meaningful customer segments in various industries, including retail, finance, and healthcare. It can help retailers identify patterns in customer behavior, preferences, and demographics, which can inform marketing strategies and improve customer engagement.

5**. Interpretability**: K-means clustering produces clusters that are easy to interpret and visualize. This allows retailers to gain insights into customer behavior and preferences, and to develop targeted marketing campaigns that resonate with each segment.

However, it's important to note that K-means clustering has some limitations, such as sensitivity to initial conditions, assumption of spherical clusters, and inability to handle categorical variables. Therefore, it's important to carefully evaluate the suitability of K-means clustering for a given customer segmentation problem and consider alternative clustering algorithms if necessary.

**CHAPTER 7: CODE & RESTULS**

**7.1 CODE**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from mpl\_toolkits.mplot3d import Axes3D

%matplotlib inline

**# Reading the csv file**

data = pd.read\_csv("Mall\_Customers.csv")

**# Number of customers we have**

print("Number of customers we have data for-", len(data))

data.head()

data.corr

**# Distribution of Annual Income**

plt.figure(figsize=(10, 6))

sns.set(style='whitegrid')

sns.distplot(data['Annual Income (k$)'])

plt.title('Distribution of Annual Income (k$)', fontsize=20)plt.xlabel('Range of Annual Income (k$)')

plt.ylabel('Count')

**# Distribution of age**

plt.figure(figsize=(10, 6))

sns.set(style='whitegrid')

sns.distplot(data['Age'])

plt.title('Distribution of Age', fontsize=20)

plt.xlabel('Range of Age')

plt.ylabel('Count')

**# Distribution of spending score**

plt.figure(figsize=(10, 6))

sns.set(style='whitegrid')

sns.distplot(data['Spending Score (1-100)'])

plt.title('Distribution of Spending Score (1-100)', fontsize=20)

plt.xlabel('Range of Spending Score (1-100)')

plt.ylabel('Count')

**# Gender Analysis**

genders = data.Gender.value\_counts()

sns.set\_style("darkgrid")

plt.figure(figsize=(10, 4))

sns.barplot(x=genders.index, y=genders.values)

plt.show()

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Age', y='Annual Income (k$)', hue="Gender", data=data, s=60)

plt.xlabel('Age')

plt.ylabel('Annual Income (k$)')

plt.title('Age vs Annual Income w.r.t Gender')

plt.legend()

plt.show()

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Annual Income (k$)', y='Spending Score (1-100)', hue="Gender", data=data, s=60)

plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.title('Spending Score (1-100) vs Annual Income (k$)')

plt.legend()

plt.show()

**# Age Buckets**

age18\_25 = data.Age[(data.Age <= 25) & (data.Age >= 18)]

age26\_35 = data.Age[(data.Age <= 35) & (data.Age >= 26)]

age36\_45 = data.Age[(data.Age <= 45) & (data.Age >= 36)]

age46\_55 = data.Age[(data.Age <= 55) & (data.Age >= 46)]

age55above = data.Age[data.Age >= 56]

x = ["18-25", "26-35", "36-45", "46-55", "55+"]

y = [len(age18\_25.values), len(age26\_35.values), len(age36\_45.values), len(age46\_55.values), len(age55above.values)]

plt.figure(figsize=(10, 6))

sns.barplot(x=x, y=y)

plt.title("Customer and Ages Barplot")

plt.xlabel("Age")

plt.ylabel("Number of Customers")

plt.show()

**# Spending Score Buckets**

ss1\_20 = data["Spending Score (1-100)"][(data["Spending Score (1-100)"] >= 1) & (data["Spending Score (1-100)"] <= 20)]

ss21\_40 = data["Spending Score (1-100)"][(data["Spending Score (1-100)"] >= 21) & (data["Spending Score (1-100)"] <= 40)]

ss41\_60 = data["Spending Score (1-100)"][(data["Spending Score (1-100)"] >= 41) & (data["Spending Score (1-100)"] <= 60)]

ss61\_80 = data["Spending Score (1-100)"][(data["Spending Score (1-100)"] >= 61) & (data["Spending Score (1-100)"] <= 80)]

ss81\_100 = data["Spending Score (1-100)"][(data["Spending Score (1-100)"] >= 81) & (data["Spending Score (1-100)"] <= 100)]

score\_x = ["1-20", "21-40", "41-60", "61-80", "81-100"]

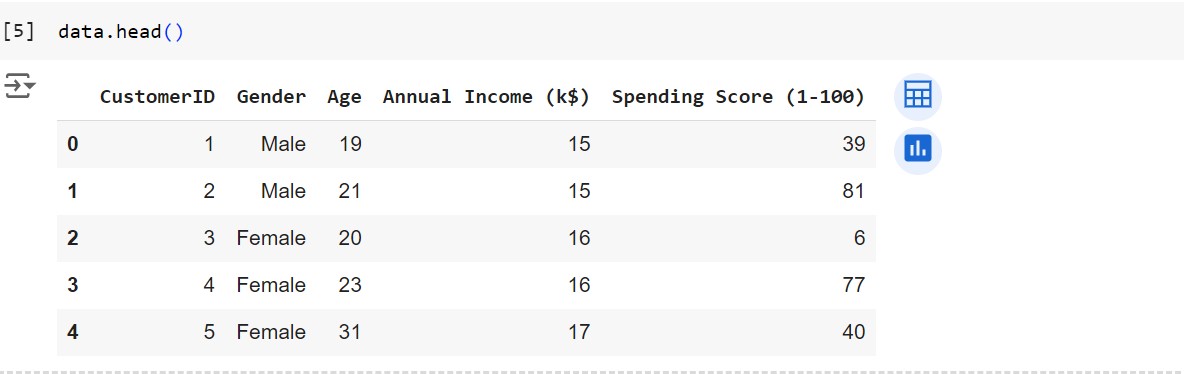
score\_y = [len(ss1\_20.values), len(ss21\_40.values), len(ss41\_60.values), len(ss61\_80.values), len(ss81\_100.values)]

plt.figure(figsize=(10, 6))

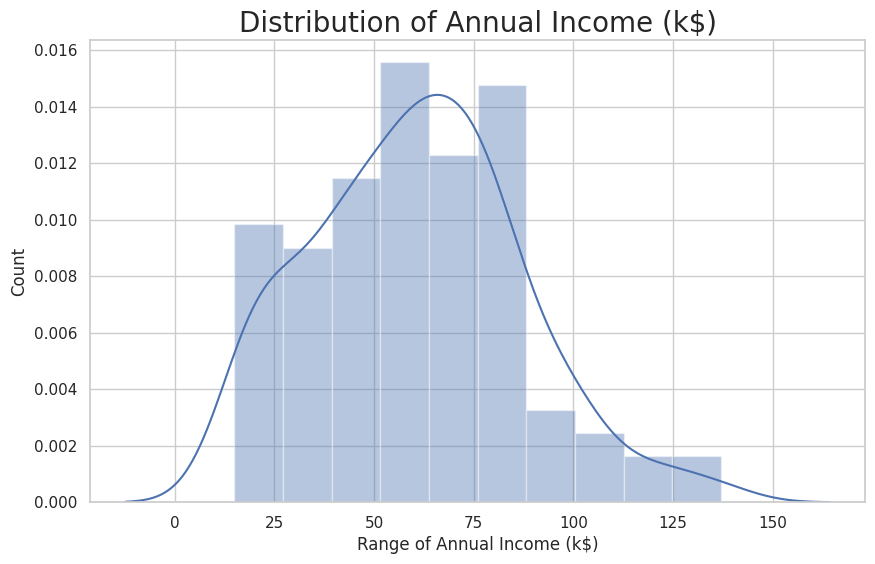
sns.barplot(x=score\_x, y=score\_y, palette

**7.2 OUTPUT**

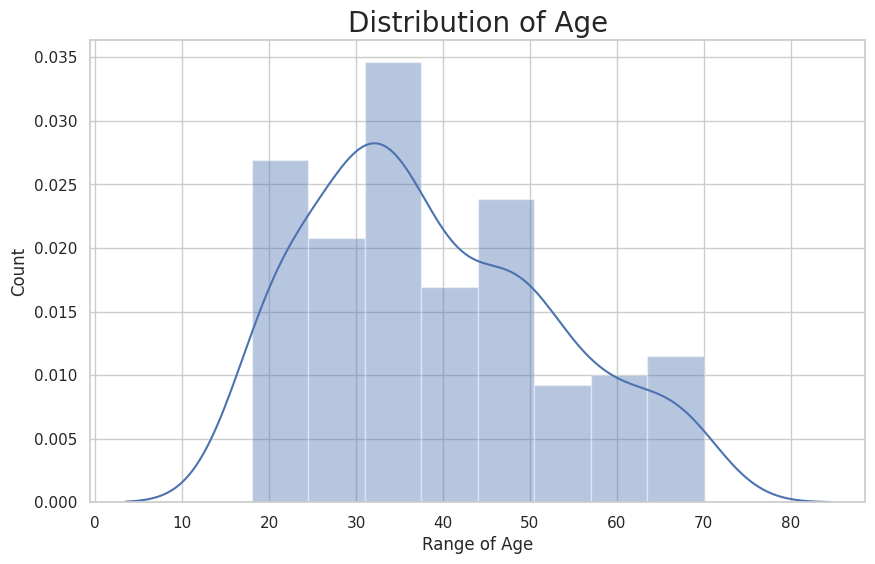
***7.2.1 Dataset***



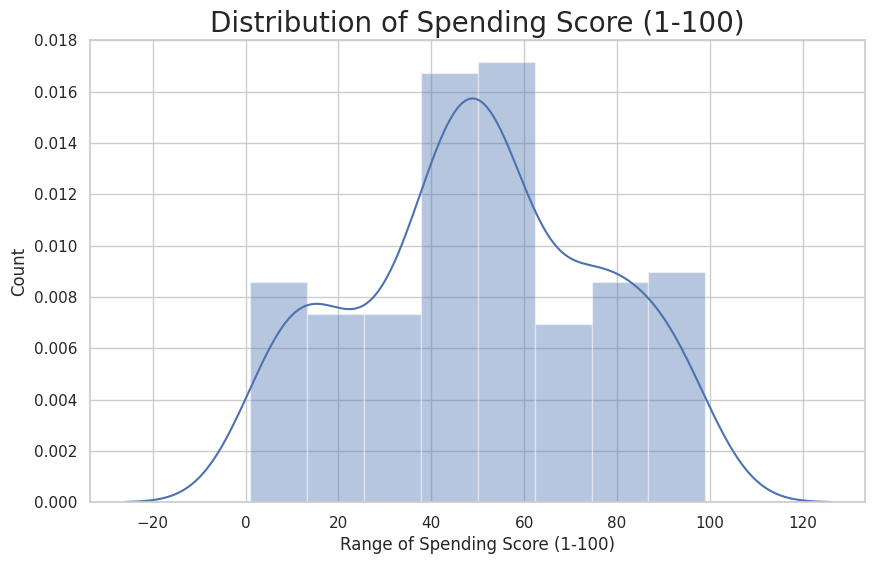
***7.2.2 Distribution of Spending Score***



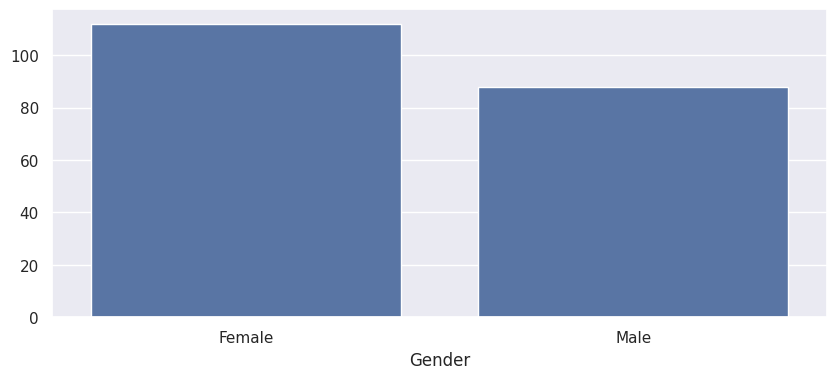
***7.2.3 Distribution of Age***



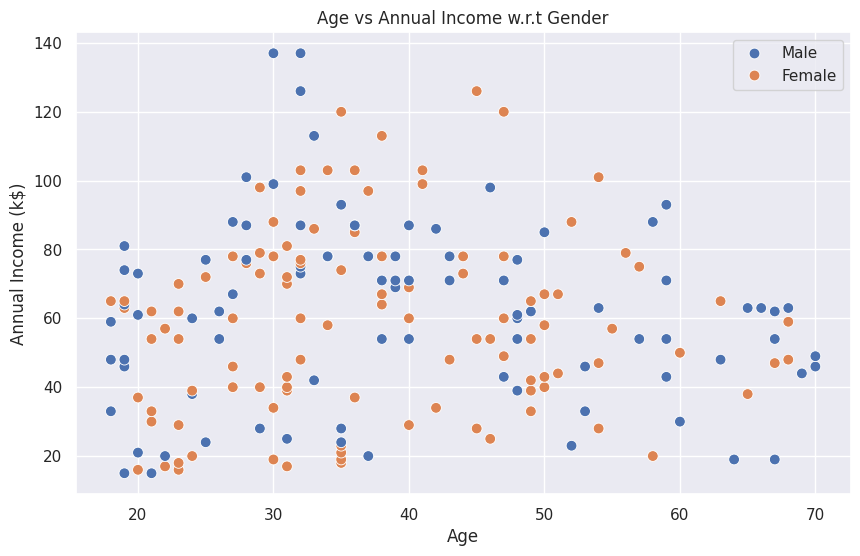
***7.2.4 Distribution of Spending Score***

******

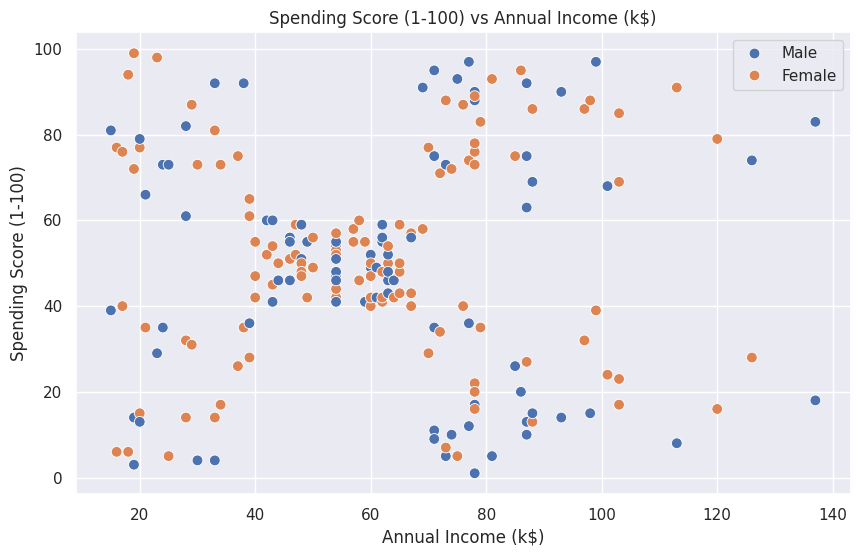
***7.2.5 Gender Analysis***



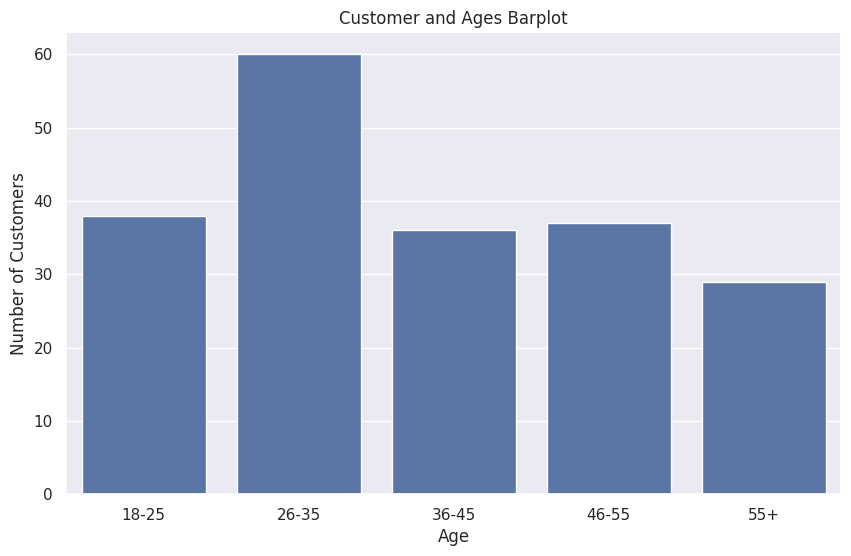
***7.2.6 Age Vs Annual Income w.r.t(Gender)***



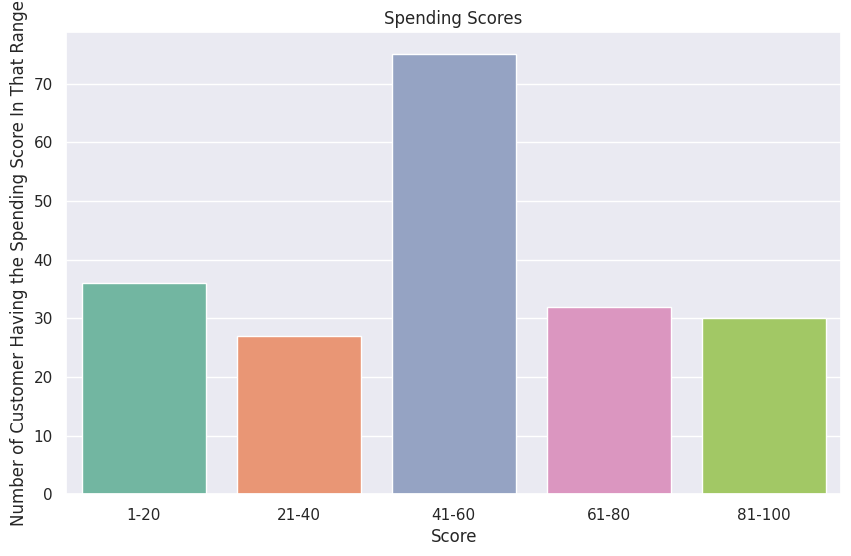
***7.2.7 Spending Score (1-100) Vs Annual Income (k$)***



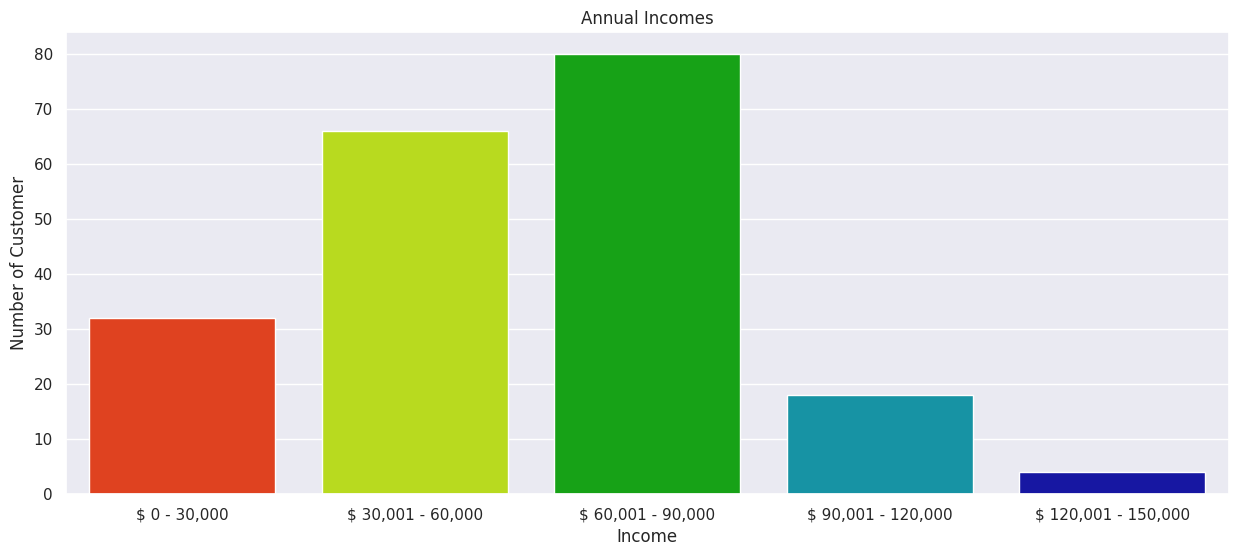
***7.2.8 Age Buckets***



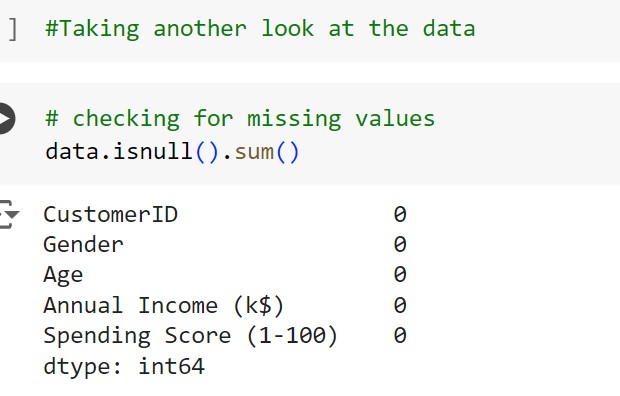
***7.2.9 Spending Score Buckets***



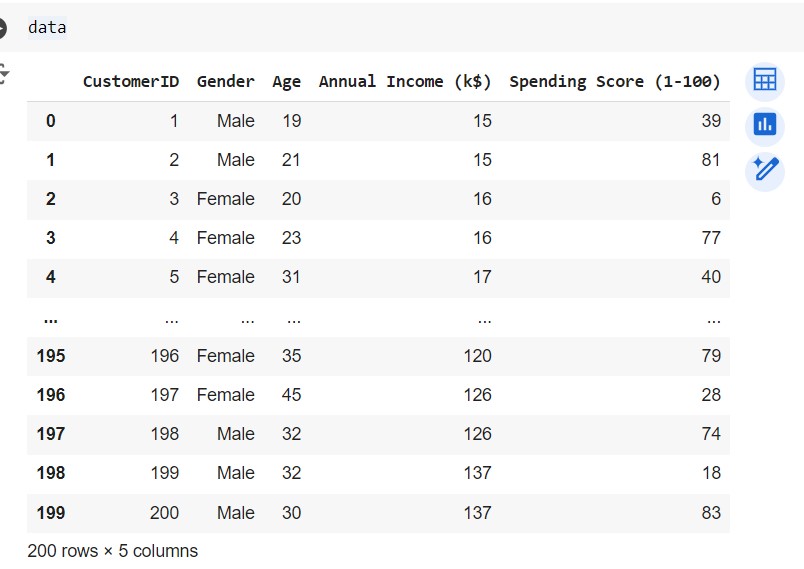
***7.2.10 Annual Income (1000USD)***



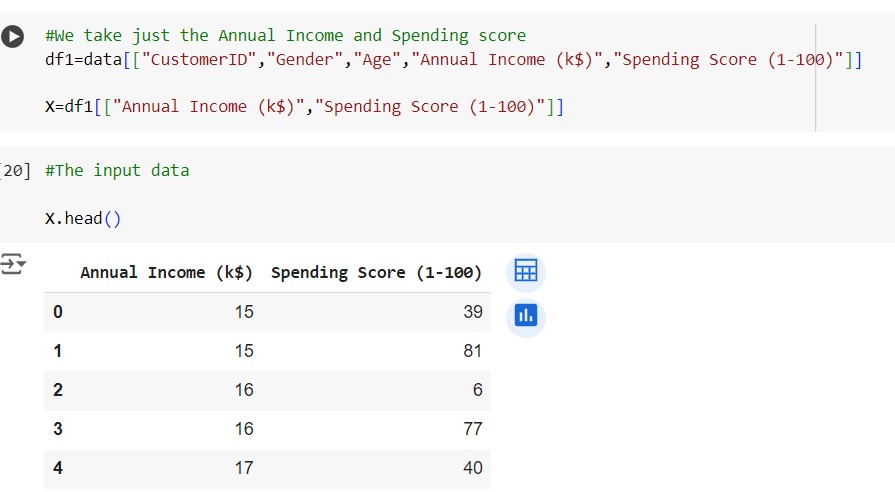
***7.2.11 Checking For Missing Value***



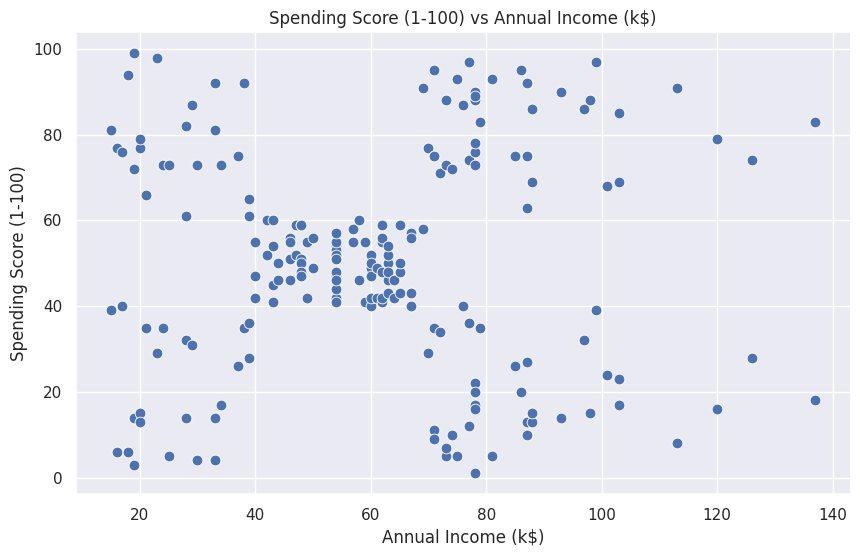
***7.2.12 Data***



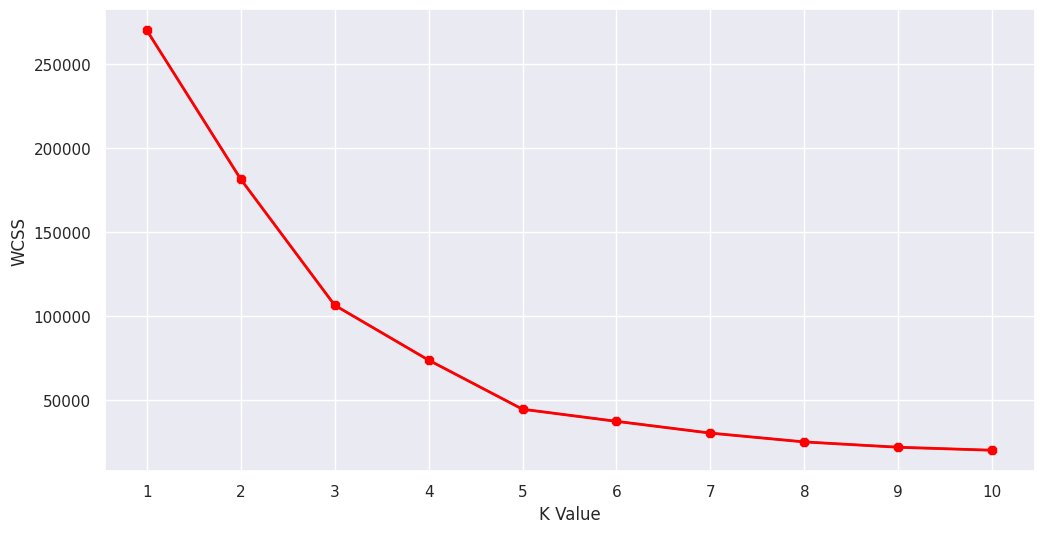
***7.2.13 Annual Income And Spending Score***



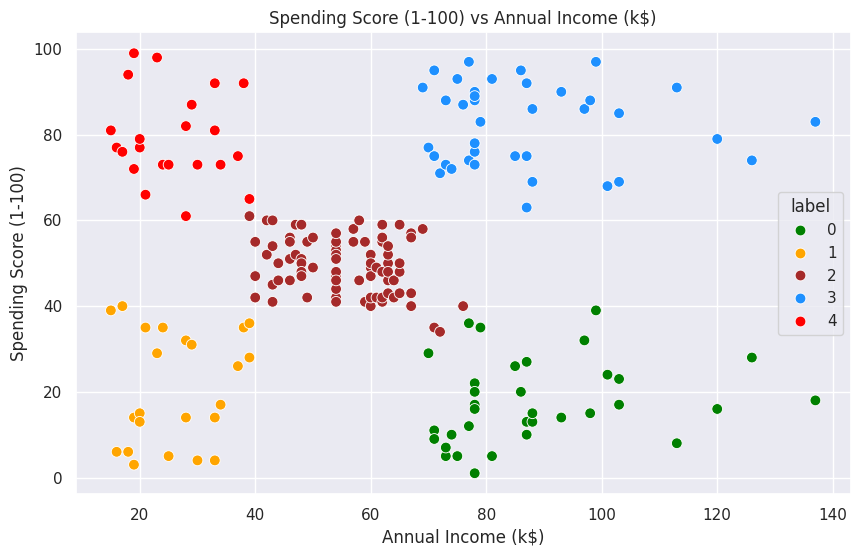
***7.2.14 Scatterplot of The Input Data***



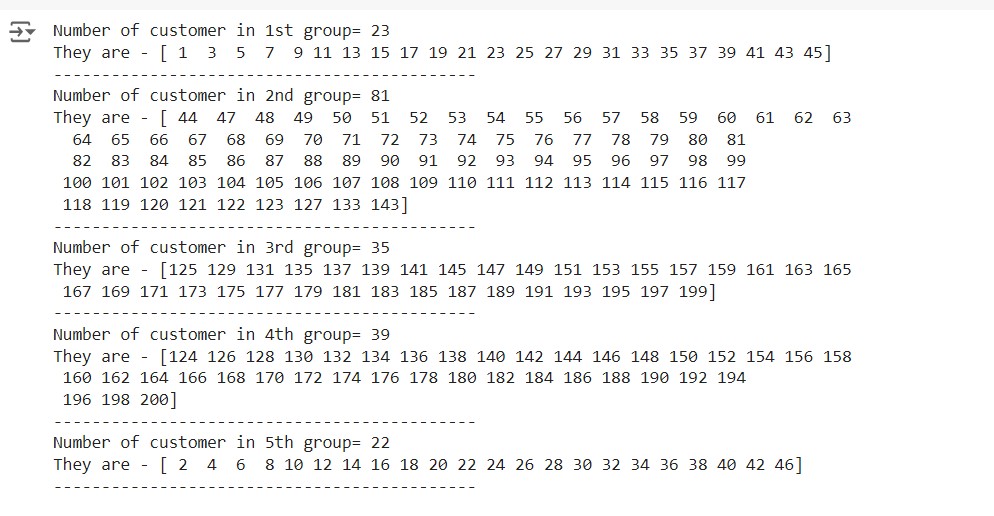
***7.2.15 Elbow Curve***



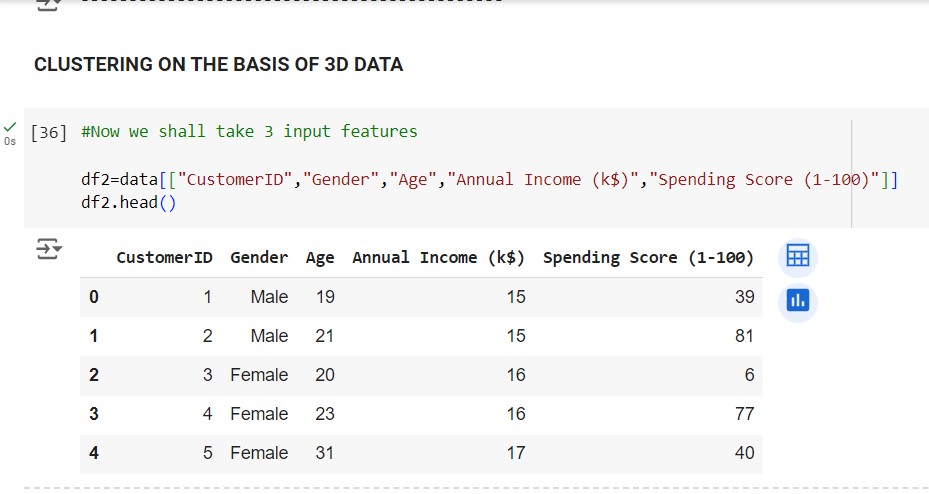
***7.2.16 Scatterplot of The Clusters***



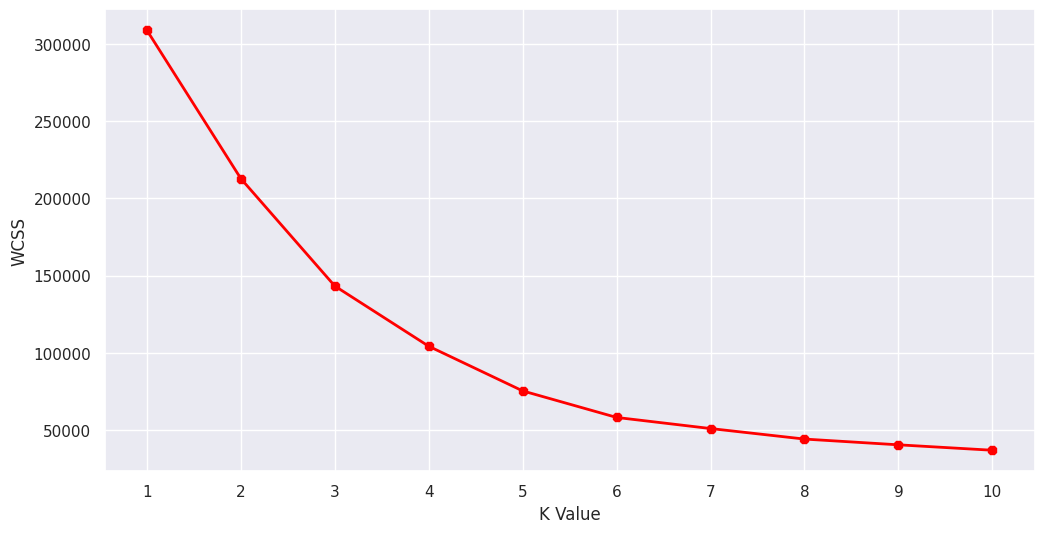
***7.2.17 Now Printing The Customer ID According To The Groups***



***7.2.18 Clustering on The Basis of 3D Data***



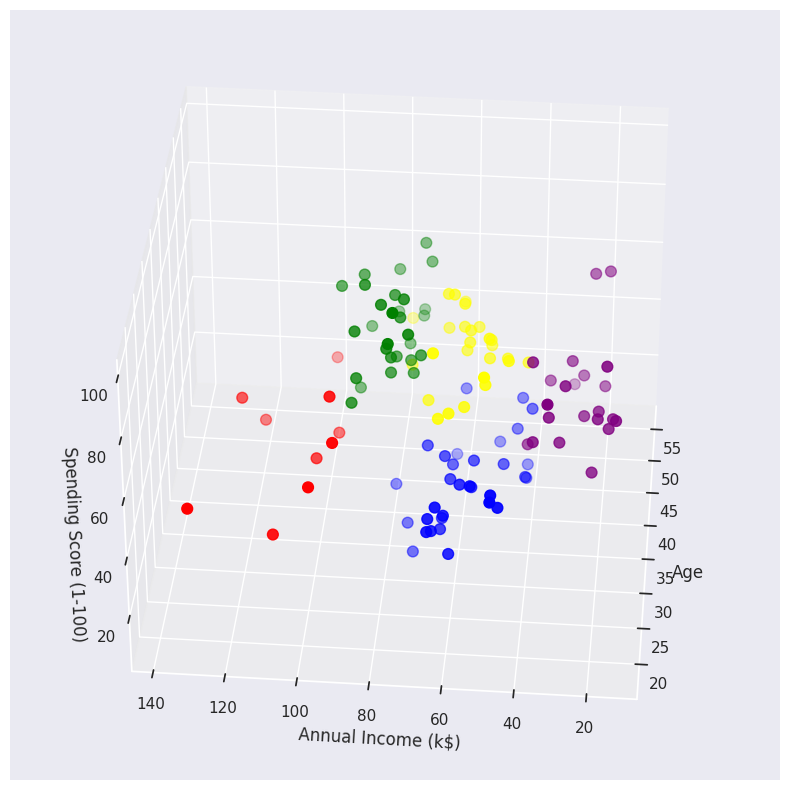
***7.2.20 Elbow curve Based on 3D Data***



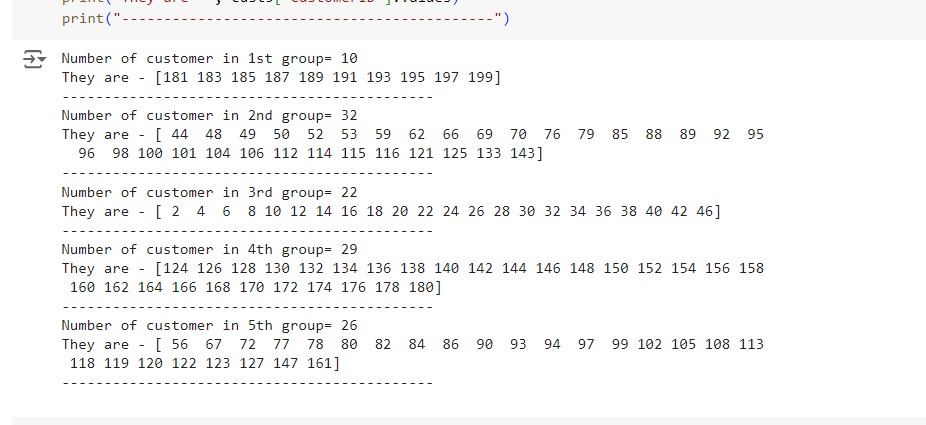
***7.2.20 New Data***



***7.2.21 3D Plot***



***7.2.22 Now Printing The Customer ID According To The Groups***



**CHAPTER 8: TEST CASES AND HYPOTHESIS PROBLEM**

**1. Define the Test Case Objectives**

The primary objectives of this test case are:

* To verify the correct implementation of clustering algorithms.
* To ensure the clustering results are meaningful and interpretable.
* To validate the model's performance using appropriate metrics.

**2. Prepare the Dataset**

We will use a synthetic dataset that mimics typical mall customer data. The dataset will include:

* Customer ID
* Gender
* Age
* Annual Income (k$)
* Spending Score (1-100)

**3. Implement the Clustering Algorithm**

We will use K-Means clustering as an example. Other clustering algorithms like DBSCAN or hierarchical clustering can also be tested similarly.

**4. Validate the Results**

We will validate the clustering results using metrics such as silhouette score and visualizations.

**5. Test Case Implementation**

***Step 1: Import Libraries python***

import numpy as np

import pandas as pd

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

import matplotlib.pyplot as plt

import seaborn as sns

*Setting random seed for reproducibility*

np.random.seed(42)

***Step 2: Create Synthetic Dataset python***

*Create a synthetic dataset*

data = {

'CustomerID': range(1, 101),

'Gender': np.random.choice(['Male', 'Female'], 100),

'Age': np.random.randint(18, 70, 100),

'Annual Income (k$)': np.random.randint(15, 137, 100),

'Spending Score (1-100)': np.random.randint(1, 101, 100)

}

df = pd.DataFrame(data)

**Step 3: Preprocess the Data python**

*Encode Gender*

df['Gender'] = df['Gender'].map({'Male': 0, 'Female': 1})

*Select features for clustering*

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

```

**Step 4: Apply K-Means Clustering python**

*Apply K-Means clustering*

kmeans = KMeans(n\_clusters=5, random\_state=42)

df['Cluster'] = kmeans.fit\_predict(X)

**Step 5: Evaluate the Clustering python**

*Calculate silhouette score*

silhouette\_avg = silhouette\_score(X, df['Cluster'])

print(f'Silhouette Score: {silhouette\_avg}')

**Step 6: Visualize the Clusters python**

*Visualize the clusters*

plt.figure(figsize=(10, 6))

sns.scatterplot(data=df, x='Annual Income (k$)', y='Spending Score (1-100)', hue='Cluster', palette='Set1')

plt.title('Customer Segments')

plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.legend(title='Cluster')

plt.show()

**6. Test Case Execution and Validation**

Running the above script should:

1. Create a synthetic dataset.

2. Preprocess the data and apply K-Means clustering.

3. Print the silhouette score to evaluate clustering performance.

4. Visualize the clusters to ensure they are well-separated and interpretable.

**7. Expected Outcomes**

- A silhouette score typically between 0.5 and 1 indicates well-defined clusters.

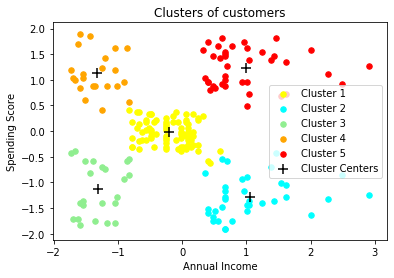
- The visualization should show distinct clusters based on annual income and spending score.

**Conclusion**

This test case provides a comprehensive framework to validate mall customer segmentation using K-Means clustering. By following these steps, you can ensure your clustering implementation is accurate and produces meaningful segments. Adjust parameters and clustering algorithms as needed to fit specific project requirements.

**CHAPTER 9: CONCLUSION**

For this project, the K-means algorithm is used and performs the best (with n\_clusters = 5 and init = ‘kmeans++’). After the clustering algorithm is applied to the dataset, this is the output.



***9.1 Annual Income Vs Spending Score after Clustering***

**9.1Clustering Analysis**

a) High Income, High Spending Score (Cluster 5) - Target these customers by sending new product alerts which would lead to an increase in the revenue collected by the mall as they are loyal customers.

b) High Income, Low Spending Score (Cluster 2) - Target these customers by asking the feedback and advertising the product in a better way to convert them into Cluster 5 customers.

c) Average Income, Average Spending Score (Cluster 1) - May or may not target these groups of customers based on the policy of the mall.

d) Low Income, High Spending Score (Cluster 4) - Can target these set of customers by providing them with Low-cost EMI's, etc.

e) Low Income, Low Spending Score (Cluster 3) - Don't target these customers since they have less income and need to save money.

**CHAPTER 10: Future Scope of the Project**

The future scope of mall customer segmentation using clustering in the retail industry is vast and promising. With the increasing use of data analytics and machine learning, retailers can gain valuable insights into customer behavior, preferences, and shopping patterns. Here are some potential future directions:

1. **Personalized Marketing**: By segmenting customers based on their characteristics and behavior, retailers can create targeted marketing campaigns that resonate with each group, increasing the likelihood of conversion.

2. Dynamic Pricing: Clustering can help retailers identify price-sensitive customers and adjust prices accordingly, maximizing revenue and profitability.

3. **Enhanced Customer Experience**: By understanding customer preferences and behavior, retailers can create a more personalized and engaging shopping experience, leading to increased customer loyalty and retention.

4. **Supply Chain Optimization**: Clustering can help retailers optimize their supply chain by identifying patterns in customer demand, reducing inventory costs, and improving logistics.

5. **Competitive Advantage**: Retailers that adopt customer segmentation using clustering can gain a competitive advantage over their rivals, as they can respond more effectively to changing customer needs and preferences.

**CHAPTER 11: BIBLIOGRAPHY**

**Books**

1. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

* Comprehensive guide to data mining techniques, including clustering and segmentation.

1. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press

* Detailed explanation of machine learning algorithms and probabilistic models.

1. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.

* Practical guide to data analysis with Python, focusing on the pandas library.

1. Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

* Practical introduction to machine learning with Python, including clustering techniques.

1. Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

* Advanced methods for text mining and unstructured data analysis.

**Online Articles and Tutorials**

1. Machine Learning Mastery (2020). A Gentle Introduction to k-means Clustering.

* Available at: https://machinelearningmastery.com/k-means-clustering
* Practical introduction to k-means clustering with examples in Python.

1. Towards Data Science (2021). Customer Segmentation using K-Means Clustering in Python.

* Available at: https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-in-python-9a3e9146ac4b
* Step-by-step guide to customer segmentation using k-means clustering in Python.

1. Scikit-learn Documentation. (2023). Clustering.

* Available at: https://scikit-learn.org/stable/modules/clustering.html
* Official documentation for clustering algorithms in scikit-learn.

**Websites and Blogs**

1. Kaggle (2023). Customer Segmentation Dataset and Notebooks.

* Available at: https://www.kaggle.com/datasets
* Various datasets and notebooks related to customer segmentation.

1. Analytics Vidhya (2021). Customer Segmentation Using Machine Learning.

* Available at: https://www.analyticsvidhya.com/blog/2021/06/customer-segmentation-using-machine-learning/
* Articles and tutorials on customer segmentation using different machine learning techniques.

1. Real Python (2023). Python Data Science Tutorials.

* Available at: https://realpython.com/tutorials/data-science/

Tutorials on data science with Python, including cluster