CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

Project Name: Customer Segmentation using K- MEANS CLUSTERING

This project applies **K-Means clustering** to segment customers based on **Annual Income** and **Spending Score**. The dataset is loaded, pre-processed, and the **Elbow Method** is used to determine the optimal number of clusters. After applying **K-Means**, the results are visualized with a scatter plot showing different customer segments.

Submitted By:

Tanisha Jain

24MCI10047

24MAM 3/B

Submitted To:

Krishan Tuli

Github link:

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my **mentor and instructors** for their invaluable guidance, encouragement, and support throughout this project. Their insightful feedback and expertise have been instrumental in shaping my understanding of **K-Means clustering and customer segmentation techniques**.

I extend my appreciation to the **creators of the Mall Customers dataset**, whose work provided a strong foundation for this analysis. Their efforts in curating real-world data have allowed me to apply machine learning concepts in a practical setting.

A special thanks to the **various research papers, online tutorials, and official documentation** that have helped me gain a deeper understanding of **data preprocessing, clustering techniques, and visualization methods**. The wealth of knowledge available through these resources has been crucial in successfully implementing this project.

I am also deeply grateful to my **peers, colleagues, and academic community** for their continuous support, insightful discussions, and constructive feedback. Their encouragement has helped refine the project and improve its overall quality.

Lastly, I would like to extend my heartfelt thanks to my **family and friends**, whose unwavering support and motivation have been a constant source of inspiration. Their belief in my abilities has driven me to push forward and complete this project with dedication and enthusiasm.

This project would not have been possible without the collective support of these individuals and resources, and I am truly appreciative of their contributions.

CERTIFICATE

This is to certify that
Tanisha Jain
has successfully completed the project titled

"Customer Segmentation Using K-Means Clustering"

as part of their academic and professional development. This project involved data preprocessing, clustering analysis, and visualization techniques to segment customers based on spending patterns and annual income.

The successful completion of this project demonstrates proficiency in **machine learning, data analysis, and software testing** while applying real-world data science methodologies.

Issued by: Chandigarh University

Date: 28/03/2025

INDEX

INTRODUCTION

Customer segmentation is a crucial process in marketing and business analytics, allowing companies to identify distinct groups of customers based on their behaviour and preferences. This project implements **K-Means clustering**, an unsupervised machine learning algorithm, to segment customers based on their **annual income and spending score**.

The **Mall Customers dataset** provides demographic and spending data, which is analysed to uncover patterns in customer behaviour. Using **K-Means clustering**, customers are grouped into different clusters, helping businesses tailor their marketing strategies and improve customer satisfaction.

The project follows a structured approach, including **data preprocessing, optimal cluster selection using the Elbow Method, model training, and visualization**. Additionally, software testing techniques, including **unit testing**, ensure the reliability and accuracy of the model. Through this analysis, businesses can make data-driven decisions to enhance their services and increase profitability.

ABSTRACT

Customer segmentation is a vital process for businesses to understand consumer behaviour and optimize marketing strategies. This project utilizes **K-Means clustering**, an unsupervised machine learning algorithm, to classify customers based on their **Annual Income** and **Spending Score** using the **Mall Customers dataset**.

The project begins with **data preprocessing**, followed by determining the optimal number of clusters using the **Elbow Method**. The **K-Means algorithm** is then applied to group customers into distinct segments, which are visualized through **scatter plots** for better interpretation. The insights gained from this segmentation help businesses target specific customer groups effectively, enhancing customer experience and maximizing profits.

OBJECTIVE

The primary objective of this project is to implement K-Means clustering for customer segmentation using the Mall Customers dataset. By analyzing customer behaviour based on Annual Income and Spending Score, the project aims to achieve the following goals:

1. Perform Data Preprocessing – Clean and prepare the dataset for analysis by handling missing values, encoding categorical data, and normalizing numerical features.
2. Determine Optimal Number of Clusters – Use the Elbow Method to identify the best value for K in the K-Means algorithm.
3. Apply K-Means Clustering – Implement the algorithm to group customers into meaningful clusters based on their spending patterns.
4. Visualize Customer Segments – Represent clustered groups using scatter plots and other graphical techniques to interpret patterns effectively.
5. Provide Business Insights – Help businesses understand customer behaviour, enabling them to design targeted marketing strategies.
6. Ensure Software Reliability – Develop and execute unit test cases to validate the correctness and efficiency of the clustering model.

By achieving these objectives, the project demonstrates the power of machine learning in customer segmentation and decision-making for business growth.

AIM

The aim of this project is to implement **K-Means clustering** to perform **customer segmentation** using the **Mall Customers dataset**. The goal is to analyse customer spending behaviour and annual income to group individuals into distinct clusters, enabling businesses to tailor their marketing strategies effectively.

Through this project, we aim to:

- Identify customer groups based on spending patterns.
- Apply **K-Means clustering** to uncover hidden patterns in the dataset.
- Visualize the clusters to provide meaningful insights for decision-making.
- Validate the accuracy and reliability of the model through **software testing**.

This study demonstrates the practical application of **machine learning and data analytics** in **business intelligence and customer relationship management**.

TASK TO BE DONE/STEPS

To successfully implement **customer segmentation using K-Means clustering**, the following tasks need to be completed:

1. Data Preprocessing

- Load the **Mall Customers dataset**.
- Handle any missing or inconsistent data.
- Select relevant features (**Annual Income, Spending Score**).
- Normalize/scale data if necessary.

2. Determine the Optimal Number of Clusters

- Use the **Elbow Method** to find the best value for **K**.
- Plot the **Within-Cluster Sum of Squares (WCSS)** to identify the optimal clusters.

3. Implement K-Means Clustering

- Apply the **K-Means algorithm** to segment customers.
- Assign each customer to a specific cluster.

4. Visualization of Clusters

- Plot a **scatter plot** to visualize different customer segments.
- Use **color-coded clustering** for better interpretation.

5. Business Insights & Interpretation

- Analyze cluster characteristics.
- Provide insights for businesses to enhance marketing strategies.

6. Software Testing

- Write **unit test cases** to validate different components.
- Ensure data preprocessing, clustering, and visualization work correctly.

7. Documentation & Report Writing

- Prepare a **detailed project report** including the **introduction, methodology, results, and conclusions**.

By completing these tasks, the project will effectively segment customers and provide valuable insights for data-driven decision-making. 🚀

STEPS

The implementation of customer segmentation using K-Means clustering involves multiple steps, ensuring proper data processing, model application, and result interpretation. Below is the theoretical explanation of each coding step:

1. Import Required Libraries

Before working with the dataset, we import necessary libraries such as:

- pandas for handling datasets.
- numpy for numerical operations.
- matplotlib.pyplot and seaborn for data visualization.
- sklearn.cluster.KMeans for applying the K-Means clustering algorithm.

2. Load and Explore the Dataset

The dataset is loaded using pandas.read_csv().

- We check the first few rows of the dataset to understand the structure.
- We also check for missing values, as they can impact the clustering process.

3. Select Relevant Features

Since K-Means clustering requires numerical input, we select two key attributes for customer segmentation:

- **Annual Income** (how much a customer earns per year).
- **Spending Score** (a metric that represents customer spending behavior). These features will help in clustering customers based on their financial behavior.

4. Determine the Optimal Number of Clusters (K) – Elbow Method

The Elbow Method helps us decide the optimal number of clusters (**K**) by:

- Running the K-Means algorithm with different values of K (1 to 10).

- Calculating the **Within-Cluster Sum of Squares (WCSS)** for each K.
- Plotting a graph of K vs. WCSS and identifying the "elbow point," which represents the best number of clusters.

5. Apply K-Means Clustering

Once the optimal **K** is identified, we apply the **K-Means algorithm**:

- The dataset is fitted using KMeans(n_clusters=K).fit_predict(data), which assigns each data point to a cluster.
- Each customer is labeled based on the assigned cluster.

6. Visualization of Clusters

To understand customer segmentation better, a **scatter plot** is created:

- Each cluster is represented with a unique color.
- The **centroids** (cluster centers) are also marked to indicate the central point of each group.

7. Saving the Processed Data

After clustering, the dataset with cluster labels is saved as a CSV file for future use or analysis.

8. Unit Testing for Model Validation

Unit tests are implemented to verify the correctness of the clustering process:

- **Test if the dataset is loaded properly** (should not be empty).
- **Test for missing values** (should not contain null values).
- **Test if the correct number of clusters is created** (should match the chosen K).

CODING

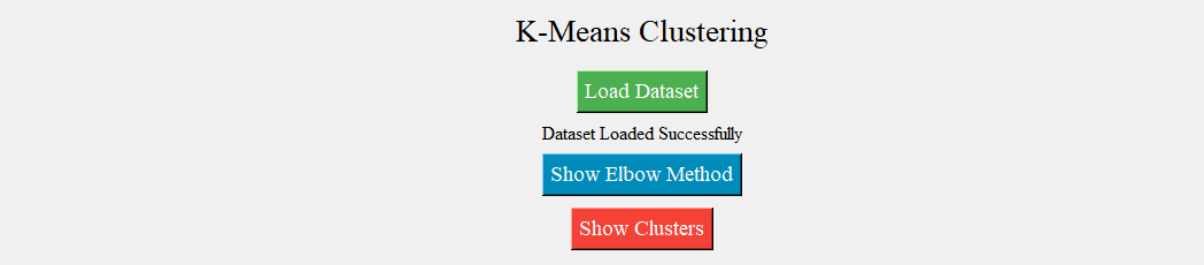
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import tkinter as tk
from tkinter import filedialog
from sklearn.cluster import KMeans
from matplotlib.backends.backend_tkagg import FigureCanvasTkAgg

def load_data():
    file_path = filedialog.askopenfilename(filetypes=[("CSV files", "*.csv")])
    if not file_path:
        return
    global df
    df = pd.read_csv(file_path)
    label_status.config(text="Dataset Loaded Successfully")
    visualize_elbow()

def visualize_elbow():
    X = df[["Annual Income (k$)", "Spending Score (1-100)"]]
    inertia = []
    K_range = range(1, 11)
    for k in K_range:
        kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
        kmeans.fit(X)
        inertia.append(kmeans.inertia_)
```

```
    fig, ax = plt.subplots(figsize=(6, 4))
    ax.plot(K_range, inertia, marker='o', linestyle='-')
    ax.set_xlabel("Number of Clusters (K)")
    ax.set_ylabel("Inertia (WCSS)")
    ax.set_title("Elbow Method for Optimal K")

    display_plot(fig)

def cluster_and_visualize():
    X = df[["Annual Income (k$)", "Spending Score (1-100)"]]
    optimal_k = 5  # Based on elbow method assumption
    kmeans = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
    df["Cluster"] = kmeans.fit_predict(X)

    fig, ax = plt.subplots(figsize=(6, 4))
    sns.scatterplot(x="Annual Income (k$)", y="Spending Score (1-100)",
hue="Cluster", palette="viridis", data=df, ax=ax)
    ax.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
s=200, c='red', marker='X', label='Centroids')
    ax.set_xlabel("Annual Income (k$)")
    ax.set_ylabel("Spending Score (1-100)")
    ax.set_title("Customer Segments Using K-Means")
    ax.legend()

    display_plot(fig)

def display_plot(fig):
    for widget in frame_plot.winfo_children():
```

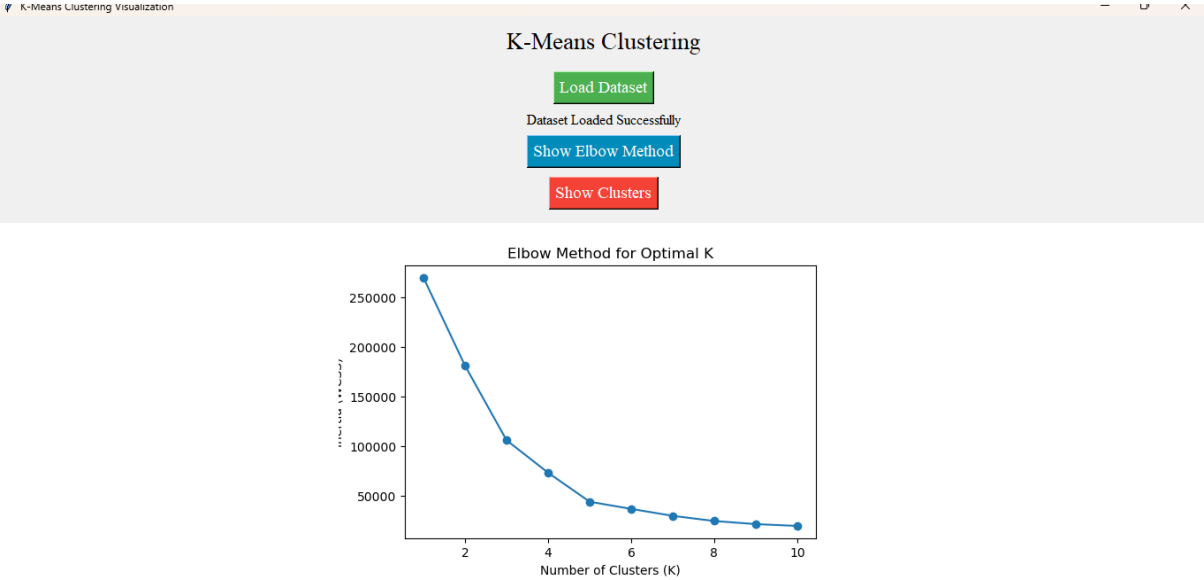
```
        widget.destroy()
    canvas = FigureCanvasTkAgg(fig, master=frame_plot)
    canvas.draw()
    canvas.get_tk_widget().pack()

# Create UI Window
root = tk.Tk()
root.title("K-Means Clustering Visualization")
root.geometry("800x600")
root.configure(bg="#f0f0f0")

title_label = tk.Label(root, text="K-Means Clustering", font=("Times New Roman", 20), bg="#f0f0f0")
title_label.pack(pady=10)

btn_load = tk.Button(root, text="Load Dataset", command=load_data, font=("Times New Roman", 14), bg="#4CAF50", fg="white")
btn_load.pack(pady=5)

label_status = tk.Label(root, text="", font=("Times New Roman", 12), bg="#f0f0f0")
label_status.pack()

btn_elbow = tk.Button(root, text="Show Elbow Method", command=visualize_elbow, font=("Times New Roman", 14), bg="#008CBA", fg="white")
btn_elbow.pack(pady=5)
```

```
btn_cluster = tk.Button(root, text="Show Clusters",
command=cluster_and_visualize, font=("Times New Roman", 14),
bg="#f44336", fg="white")

btn_cluster.pack(pady=5)


frame_plot = tk.Frame(root, bg="#ffffff")

frame_plot.pack(pady=10, fill=tk.BOTH, expand=True)


root.mainloop()
```

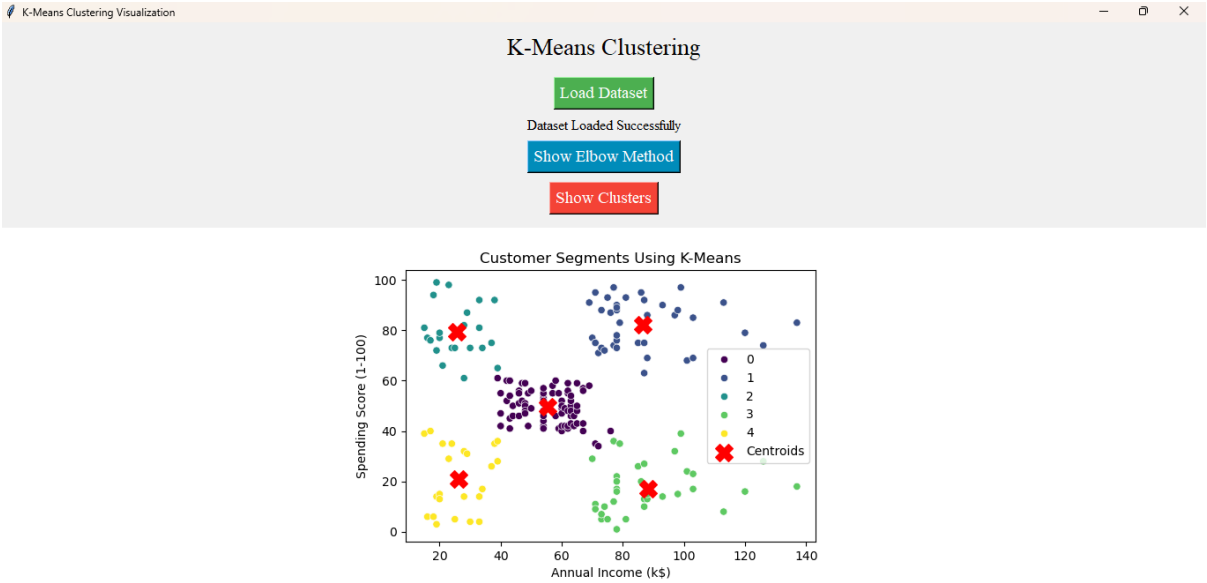
OUTPUT



Elbow method



Show clusters


K-Means Clustering Visualization
K-Means Clustering
Load Dataset
Dataset Loaded Successfully
Show Elbow Method
Show Clusters
Customer Segments Using K-Means
Spending Score (1-100)
Annual Income (k$)
0
1
2
3
4
Centroids

CONCULSION

The implementation of **K-Means clustering** for **customer segmentation** in this project successfully groups customers based on their **Annual Income** and **Spending Score**. The **Elbow Method** helped in determining the optimal number of clusters, ensuring efficient classification of customers into meaningful segments.

Through visualization, we observed distinct clusters representing different spending behaviours, allowing businesses to tailor their marketing strategies accordingly. The use of **unit testing** ensured the correctness of the implementation by validating data integrity, cluster formation, and model efficiency.

Overall, this project demonstrates the effectiveness of **K-Means clustering** in real-world applications, enabling businesses to make **data-driven decisions** for customer relationship management and strategic planning.

BIBLIOGRAPHY

Scikit-Learn Documentation: https://scikit-learn.org

Pandas Library Documentation: https://pandas.pydata.org

Matplotlib & Seaborn Documentation: https://matplotlib.org

Learning Outcomes

- Understanding Customer Segmentation.
- Application of k- means clustering.
- Data preprocessing and feature selection.
- Model optimization.