🧠 Customer Segmentation using K-Means Clustering

This project applies the K-Means clustering algorithm to segment customers based on their annual income and spending score. The goal is to identify distinct customer groups for targeted marketing or business strategies.

📌 1. Import Libraries

In this section, we import all the Python libraries we'll use:

- `pandas` for data handling
- `matplotlib` and `seaborn` for visualization
- `sklearn` for clustering with **KMeans**

These tools allow us to load data, explore it, build the model, and visualize the results.

📂 2. Load Dataset

Here, we load the Mall Customer dataset using `pd.read_csv()`.

We preview the data using:

- `.head()` to see the first few rows
- `.columns` to view column names
- `.describe()` to get summary statistics

This helps us understand the structure and basic distribution of the data.

🔍 3. Exploratory Data Analysis (EDA)

We examine the summary statistics of customer data, including Age, Income, and Spending Score.

We perform EDA to explore patterns and detect any issues like missing data.

This includes:

- Summary statistics (mean, min, max)
- Understanding the distribution of features like `Age`, `Annual Income`, and `Spending Score`

EDA helps us decide which features are most useful for clustering.

📊 4. Visualize Customer Distribution

We use a **scatter plot** to plot `Annual Income` vs. `Spending Score`.

This gives a visual idea of how customers are distributed and whether natural groupings (clusters) seem to exist.

It's a **pre-model insight** step.

📉 5. Elbow Method to Find Optimal K

The elbow method helps us determine the best number of clusters.

We:

- Fit KMeans with different values of **K** (from 1 to 10)
- Calculate **inertia** (within-cluster sum of squares)
- Plot the inertia vs. K

The “elbow” point on the curve indicates the ideal number of clusters to use.

🤖 6. Apply K-Means Clustering

Now that we know the optimal K (for example, 5), we:

- Initialize and fit the KMeans model
- Predict cluster labels for each customer

This step segments our customers into distinct clusters based on their income and spending habits.

🗒️ 7. Add Cluster Labels to Dataset

We add the predicted **cluster number** as a new column to our original DataFrame.

Now every customer is labeled as belonging to Cluster 0, 1, 2, etc.

This enriched dataset is useful for reporting, dashboards, or further analysis.

📈 8. Visualize Final Clusters

We use a colored scatter plot to visualize clusters clearly:

- Each color represents one cluster
- Cluster centroids are shown as large black X marks

This visualization helps explain the segmentation logic and validate that clusters are well-separated.

💾 9. Export Clustered Data (optional)

We export the updated dataset to a `.csv` file with cluster labels:

```
df.to_csv("Mall_Customers_Clustered.csv", index=False)
```

This file can be used in Power BI, Tableau, or Excel for further analysis and dashboard creation.

🧩 10. Insights and Conclusion

In the final step, we summarize what we found:

- The data naturally grouped into K clusters (e.g., 5)
- Each cluster shows a different customer behavior pattern:
    - High income, high spending
    - Low income, low spending
    - High income, low spending, etc.

These insights help businesses:

- Target marketing campaigns
- Design loyalty programs

- Improve customer service