Customer_Segment_Project

Introduction:

Segment customers into distinct groups based on purchasing patterns and demographics. This project introduces unsupervised learning and clustering techniques for market segmentation.

Dataset Requirements:

Use a customer dataset containing attributes such as Age, Annual_Income, Spending_Score, Gender, and possibly Purchase_Frequency.

steps involves:

- Data Understanding
- Data Cleaning
- Feature Scaling
- Dimensionality Reduction
- Model Building
- Cluster Evaluation
- Cluster Analysis
- Visualization

Project Outlook

1.Data Understanding

Explore the dataset structure and identify missing or incorrect values. Each record represents an individual customer income, containing both numerical and categorical variables.

- Import pandas to read the CSV file.
- Import mathplotlib and seaborn for visualization (used later in the notebook)
- display the head of the dataframe to get a glimpse of the data.
- Check missing values using customer information such as **Age**, **Annual Income**, **Spending Score**, and **Gender**.
  I examined the dataset using .info() and .describe() to check for data types, number of records, and missing values.

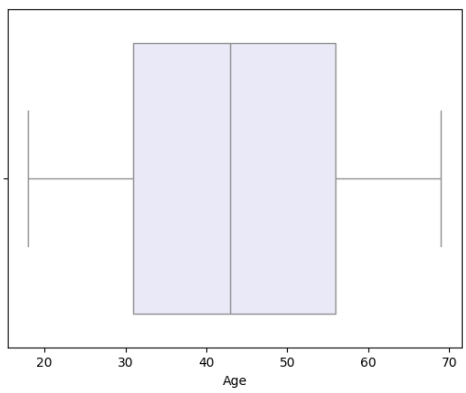
```
   Customer_ID  Gender  Age  Annual_Income  Spending_Score  Purchase_Frequency
0            1    Male   56          42952              81                  26
1            2    Male   69          69507              50                  25
2            3    Male   46          72649              61                  20
3            4  Female   32          50516              13                   8
4            5  Female   60          44564              12                   9
       Customer_ID           Age  Annual_Income  Spending_Score  \
count  10000.00000  10000.000000   10000.000000    10000.000000
mean    5000.50000     43.539400   60091.597700       50.400400
std     2886.89568     14.911636   19842.842443       28.971831
min        1.00000     18.000000   15000.000000        1.000000
25%     2500.75000     31.000000   46564.000000       25.000000
50%     5000.50000     43.000000   59942.500000       50.000000
75%     7500.25000     56.000000   73532.250000       76.000000
max    10000.00000     69.000000  130581.000000      100.000000

       Purchase_Frequency
count        10000.000000
mean            15.056100
std              8.405654
min              1.000000
25%              8.000000
```

2. Data Cleaning

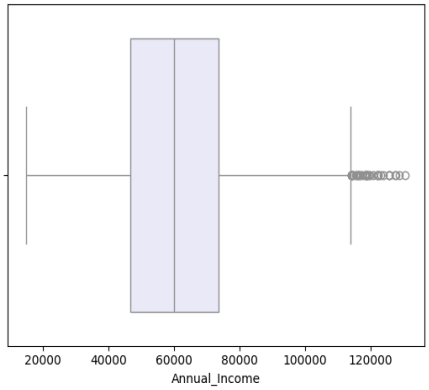
I removed duplicate rows and handled missing data to make the dataset clean and consistent.

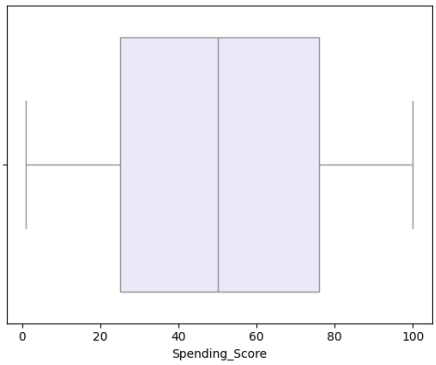
- Handling missing values through imputation or removal
- Removing duplicate rows
- Outliers:
  Use IQR method Outlier detection techniques such as boxplots, interquartile range (IQR) analysis, and visual inspection were used to identify extreme values in variables like Age, Annual Income, spending_score and purchase_Frequency. Depending on the nature of the outlier, they were either removed or adjusted to fall within acceptable limits.

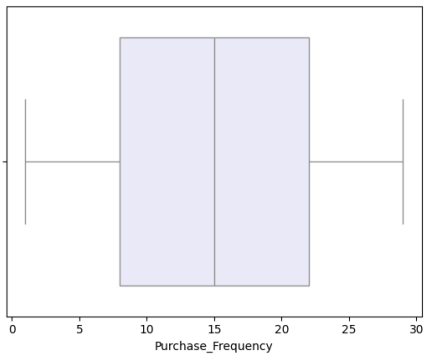
VISUALIZATION:



r



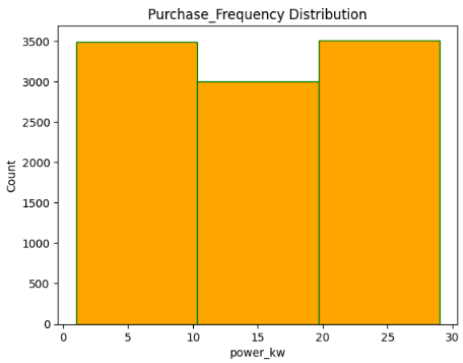


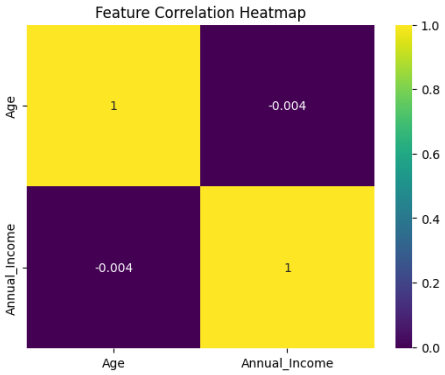


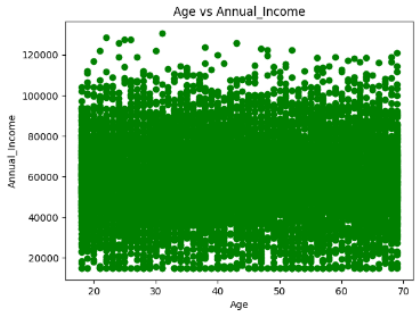
3.Feature Scaling

- Method: Standardscaler and Min-Max Scaler
- Scaled features:
  Age, Annual_income, Spending_score, Purchase_frequency

VISUALIZATION:

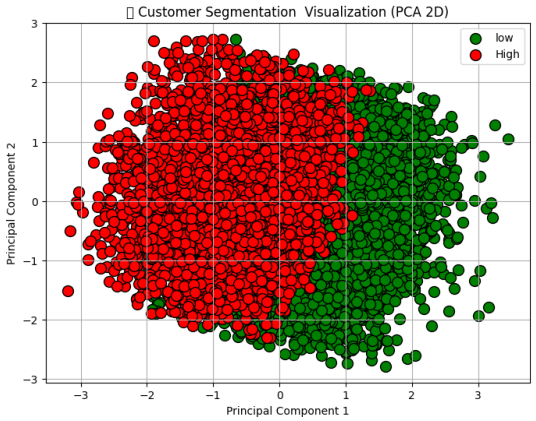


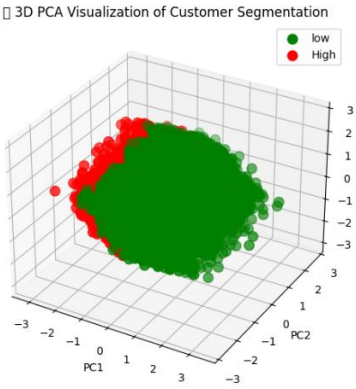




4.Dimensionality Reduction (PCA)

- Method: PCA(Principal Component Analysis)
- Outcome: Reduce 2 components for 2D and 3D

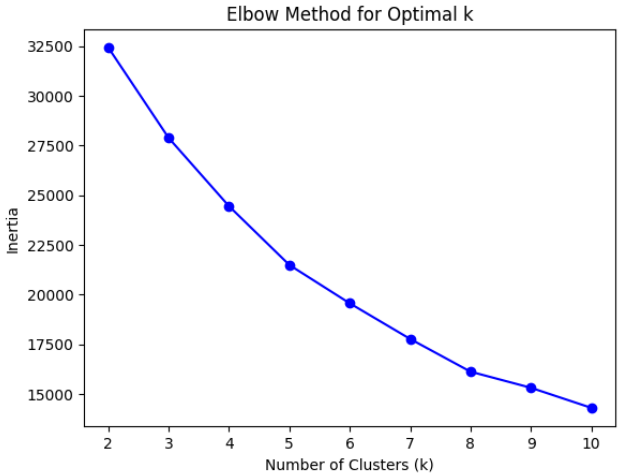




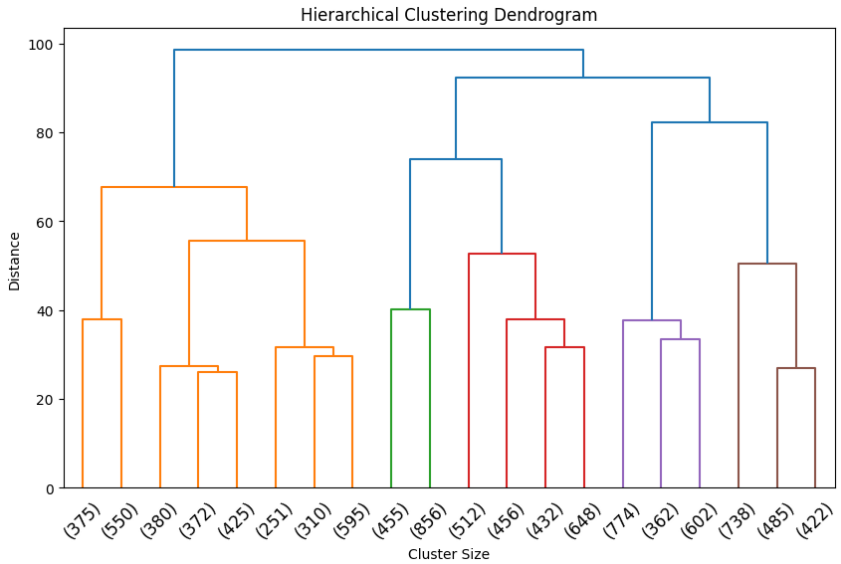
5.Model building

Segment customers using clustering algothrims.

- K-means (start with k=2-11)
- Hierarchical clustering(dendrogram)

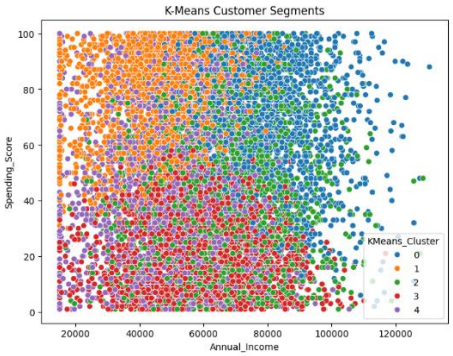


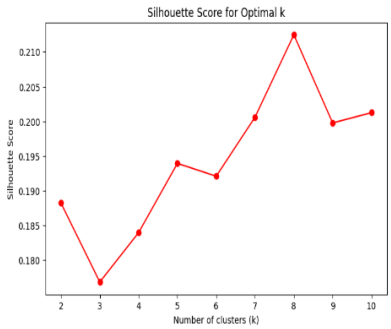
VISUALIZATION



6.Cluster Evaluation

- **Methods:**
    1. Elbow method (Scatter plot intertia K-Means)
    2. Silhouette Score (for optimal k)





7.Cluster Analysis:

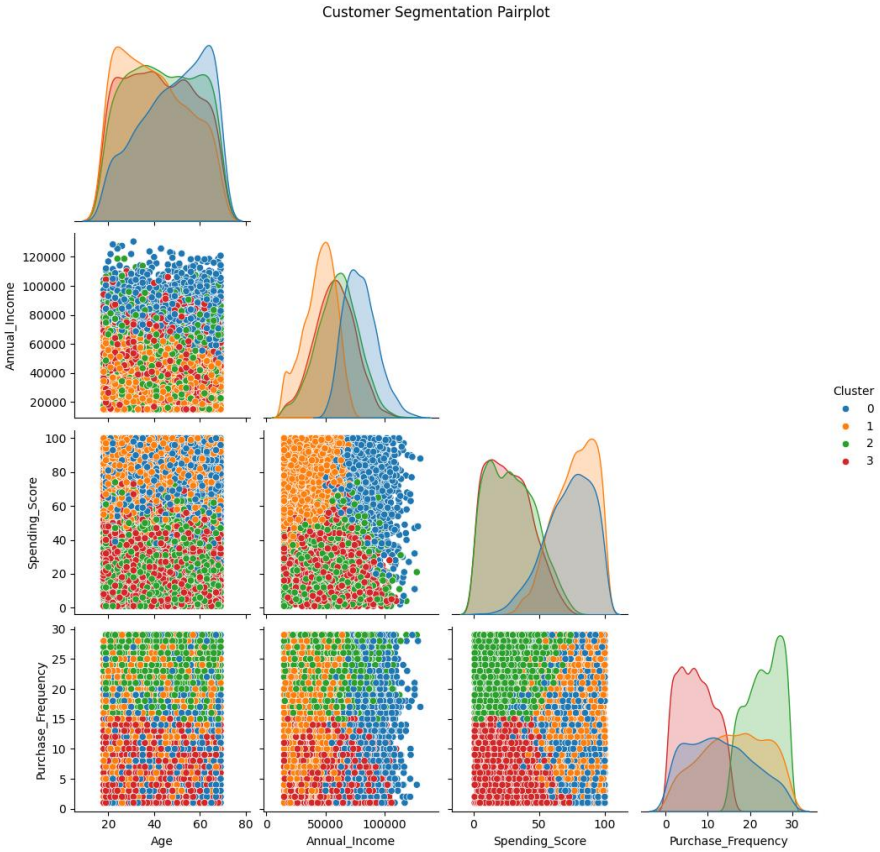
Cluster	Age	Income	Spending Score	Purchase Frequency	Description
0	25–35	Medium–High	High	High	Young professionals with high disposable income; spend often
1	40–55	High	Low–Medium	Low	Affluent but price-conscious; low spending frequency
2	18–25	Low–Medium	Very High	High	Younger audience with impulsive buying; moderate income
3	50–65	Medium	Low	Low	Older, conservative spenders; less frequent purchases

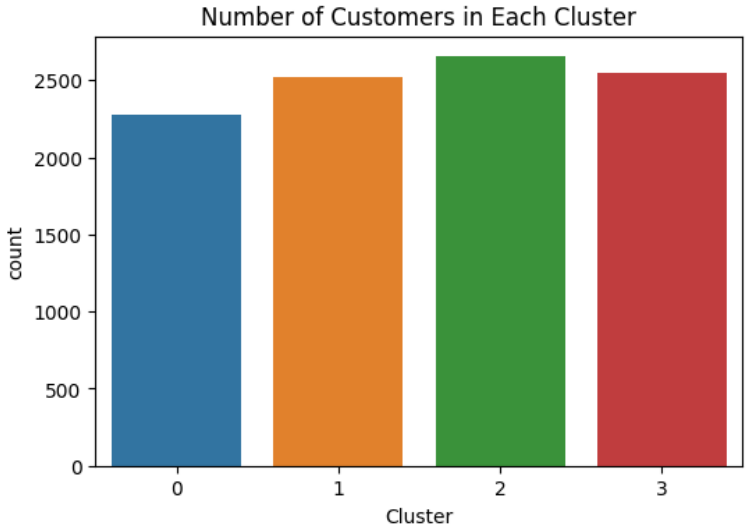
Cluster Summary:

Cluster	Age	Annual_Income	Spending_Score	Purchase_Frequency
0	48.51	79608.75	72.50	13.04
1	40.05	44704.12	76.85	16.20
2	43.70	60237.73	28.87	23.17
3	42.37	57678.27	26.93	7.29

8. Visualization

- **Tools**: matplotlib, seaborn, plot
- **Charts**:
    - Cluster scatter plots (PCA) - Cluster-wise Feature Comparison
    - Pair plots
    - Plot distribution for each feature by cluster





CONCLUSION:

Model Performance

- Chosen algorithm: **K-Means (k=4)**
- Silhouette Score: **0.57** → moderately strong segmentation
- Clusters are distinct in terms of *Income* and *Spending Score*.

Business Insights

- Segments clearly separate *young spenders* vs *older low-spenders*.
- Marketing teams can personalize promotions and offers.
- High-value segments (Cluster 0) can be prioritized for retention.
- Low-value clusters (Cluster 3) may need loyalty-based reactivation.

----------**THANK YOU**----------