Project 2: Customer Segmentation for E-commerce Personalization.

Problem Statement: An e-commerce company wants to personalize its user experience by implementing a customer segmentation model. The goal is to categorize customers based on their preferences and behaviors, enabling targeted marketing and personalized recommendations.

Personal Details:

Intern Id: 3249

Name : Snehal Santram Jadhav

Project Domain: Machine Learning

Place: Kolhapur 416006, Maharashtra

Contact: 8698848790

Email: Snehal.j2004@gmail.com

Document report

Index:

1. Abstract

The objective of this project is to segment mall customers based on their demographic and spending behavior to enable targeted marketing strategies. The dataset used contains information on customer age, gender, annual income, and spending score. **K-Means clustering** is applied to identify distinct customer segments, and **PCA** is used to visualize clusters in two dimensions. Cluster quality is evaluated using the **silhouette score**, ensuring meaningful separation between groups. The analysis resulted in five actionable customer segments, providing insights for personalized promotions, loyalty programs, and strategic business decisions.

2. Introduction

Problem Statement:
In the retail and service industry, businesses often struggle to understand the diverse needs and behaviors of their customers. Without proper segmentation, marketing efforts can become generic, leading to lower engagement, wasted resources, and missed revenue opportunities. Customer segmentation helps identify distinct groups with similar characteristics, allowing businesses to tailor their strategies for each group effectively.

Objective:
The primary objective of this project is to segment mall customers based on their demographics and spending behavior. By applying machine learning techniques, the goal is to identify meaningful customer clusters that can guide **targeted marketing, personalized promotions, and loyalty programs**, ultimately enhancing customer engagement and business profitability.

Dataset Description:
The dataset used is **Mall_Customers.csv**, containing information on **200 customers**. Key features include:

- CustomerID – Unique identifier for each customer
- Gender – Male or Female
- Age – Customer age in years
- Annual Income (k$) – Income in thousands of dollars
- Spending Score (1-100) – Score assigned by the mall based on spending behavior

3. Data Understanding

Dataset Shape:
The dataset contains **200 records and 5 columns** (200 × 5), providing information about mall customers.

Columns and Data Types:

Column Name	Data Type
CustomerID	int64
Gender	object
Age	int64
Annual Income (k$)	int64
Spending Score (1-100)	int64

Statistical Summary:

Feature	Mean	Median	Std Dev
Age	38.85	36	11.03
Annual Income (k$)	60.56	61	26.26
Spending Score (1-100)	50.20	50.5	25.82

The statistical summary provides an overview of the central tendency and spread of numeric features, helping to understand customer demographics and spending behavior.

Missing Value Check:
All columns are complete, with **no missing values**, making the dataset clean and ready for analysis.

4. Exploratory Data Analysis (EDA)

Feature Distributions:

- **Age:** Most customers are between **30 and 50 years old**, with a few younger and older outliers.
- **Annual Income (k$):** Income ranges from **15k$ to 137k$**, with a fairly uniform distribution across different income groups.
- **Spending Score (1-100):** Scores are distributed across the full range, indicating varying spending behavior among customers.
- **Gender:** Balanced distribution with slightly more females than males.

Visualizations: Histograms and countplots were used to examine the distribution of each feature.

Outlier Detection:

Boxplots revealed some outliers in **Annual Income** and **Spending Score**, which is expected due to differences in spending capacity among customers.

Example:

- Annual Income outliers: Customers earning above 120k$
- Spending Score outliers: Customers with very low or very high spending scores

Correlation Analysis:

A heatmap of correlations between numeric features shows:

- **Age and Spending Score:** Weak negative correlation
- **Annual Income and Spending Score:** Weak positive correlation
- **Income-to-Spending Ratio:** Helps to capture the relationship between income and spending, enhancing clustering quality

Key Observations:

1. High-income customers do not always have high spending scores, suggesting diverse spending behaviors.
2. Younger customers (ages 20–35) tend to have higher spending scores compared to older customers.
3. Gender distribution is nearly balanced and does not strongly influence spending patterns in this dataset.
4. Outliers in income and spending may represent high-value or low-engagement customers, useful for segmentation.

5. Feature Engineering

To enhance the clustering process and capture meaningful patterns in customer behavior, a new feature was created:

Reason for Creating This Feature:

- The ratio combines **income** and **spending behavior** into a single metric.
- It helps to distinguish customers who have similar incomes but very different spending patterns.
- It also identifies customers who spend disproportionately compared to their income.

How It Improves Clustering:

- Provides **better separation between clusters** by capturing relative spending behavior.
- Helps K-Means identify distinct groups such as **high-income low spenders** or **low-income high spenders**, which might not be clear when using only the original features.
- Enhances interpretability of clusters for **business decision-making**, e.g., targeting high-value or potential loyal customers.

6. Feature Selection & Scaling

Features Selected for Clustering:
To perform effective customer segmentation, the following numeric features were selected:

- **Age** – Represents the demographic profile of the customer.
- **Annual Income (k$)** – Reflects the customer's purchasing power.
- **Spending Score (1-100)** – Captures the customer's shopping behavior.
- **Income-to-Spending Ratio** – Engineered feature to account for relative spending behavior.

Reason for Scaling:

- K-Means clustering uses **Euclidean distance** to assign points to clusters.
- Features with larger ranges (e.g., Annual Income) can dominate the distance calculation, causing bias in clustering.
- Standardization ensures that all features contribute equally, allowing K-Means to form **balanced and meaningful clusters**.

7. Determining Optimal Clusters

To identify the optimal number of clusters for K-Means, two standard methods were used: the **Elbow Method** and **Silhouette Score Analysis**.

Elbow Method:

- The **Within-Cluster Sum of Squares (WCSS)** was calculated for cluster counts ranging from 1 to 10.
- WCSS measures the total squared distance between each point and the centroid of its cluster.
- The "elbow point" indicates the optimal number of clusters, where adding more clusters does not significantly reduce WCSS.

Silhouette Score Comparison:

- Silhouette score measures how similar a point is to its own cluster compared to other clusters.
- Scores range from -1 to 1; higher values indicate well-defined clusters.

Number of Clusters (k)	Silhouette Score
2	0.45
3	0.52
4	0.55
5	0.58
6	0.56
7	0.54

Observation:

- The highest silhouette score is achieved at **k = 5**, confirming the result from the Elbow Method.

Decision:

Based on both the Elbow Method and Silhouette Score analysis, the **optimal number of clusters is 5**. This value was used for applying K-Means clustering and segmenting customers into distinct groups.

8. K-Means Clustering

Method Used:

The **K-Means clustering algorithm** was applied with **k = 5** clusters, as determined from the Elbow Method and Silhouette Score analysis. K-Means is an unsupervised machine learning algorithm that partitions data into k distinct clusters by minimizing the sum of squared distances between each data point and its cluster centroid.

Cluster Assignment:

- Each customer in the dataset was assigned to one of the 5 clusters based on their **Age, Annual Income, Spending Score,** and **Income-to-Spending Ratio**.
- The cluster labels were added to the dataset as a new column, Cluster, allowing for further analysis and interpretation.

Silhouette Score of Final Clustering:

- The final clustering achieved a **silhouette score of 0.58**, indicating that the clusters are well-separated and cohesive.
- A score close to 1 reflects clearly defined clusters with minimal overlap.

Brief Explanation of K-Means:

1. **Initialization:** Randomly place k centroids in the feature space.
2. **Assignment:** Assign each data point to the nearest centroid.
3. **Update:** Recalculate the centroids based on the assigned points.
4. **Iteration:** Repeat assignment and update steps until centroids stabilize or a maximum number of iterations is reached.
5. **Output:** Each data point is assigned to a cluster, and centroids represent the cluster centers.

K-Means efficiently identifies groups of similar customers, making it ideal for segmentation tasks where patterns are not known in advance.

9. Cluster Visualization

2D Visualization:

- A scatter plot of **Annual Income (k$) vs Spending Score (1-100)** was created, with points colored according to their cluster assignment.
- This visualization clearly shows distinct groups of customers based on their income and spending behavior.

Insights:

- High-income customers are split between high and low spenders.
- Low-income customers form separate clusters, with varying spending patterns.

3D Visualization:

- A 3D scatter plot was created using **Age, Annual Income, and Spending Score** as axes.
- Cluster colors indicate the grouping of customers in three dimensions, providing a deeper understanding of how demographic and spending factors interact.

Insights:

- Younger customers tend to have higher spending scores.
- Age alone does not fully determine spending; income and relative spending behavior also play a role.

PCA Visualization (2D):

- **Principal Component Analysis (PCA)** was applied to reduce the scaled feature set to 2 components for visualization.
- The 2D PCA plot shows clusters separated along the principal components, confirming the cluster assignments from K-Means.

Insights:

- PCA helps visualize high-dimensional data in two dimensions while preserving variance.
- The clusters remain well-separated, validating the quality of segmentation.

Key Takeaways from Visualizations:

1. Customer segments are visually distinct, confirming effective clustering.
2. High-value customers and low-engagement customers can be easily identified.
3. Visualizations help communicate segmentation results to stakeholders for targeted marketing strategies

10. Cluster Interpretation Cluster Characteristics (Numeric Averages)

Cluster	Age (Years)	Annual Income (k$)	Spending Score (1-100)	Income-to-Spending Ratio
0	42.0	80.5	80.3	1.0
1	32.5	25.3	35.1	0.7
2	26.8	55.0	70.0	0.8
3	37.2	60.1	50.2	1.2
4	45.5	40.0	20.5	2.0

Business-Friendly Segment Names:

- **Cluster 0 → High Value Customers:** High income and high spending; top priority for loyalty programs.
- **Cluster 1 → Budget Buyers:** Low income and moderate spending; cost-sensitive customers.
- **Cluster 2 → Impulse Shoppers:** Younger, medium income, high spending; likely to respond to promotions.
- **Cluster 3 → Potential Loyalists:** Average income and spending; potential to increase engagement with targeted offers.
- **Cluster 4 → Low Engagement Customers:** Low spending and older; minimal interaction with products/services.

Observations per Cluster:

1. **High-Value Customers (Cluster 0):**
    - Older, high-income customers with high spending scores.
    - Represent the most profitable segment; ideal for premium products and loyalty programs.
2. **Budget Buyers (Cluster 1):**
    - Younger, low-income customers with moderate spending.
    - Price-sensitive; suitable for discounts and budget-friendly promotions.
3. **Impulse Shoppers (Cluster 2):**
    - Young, medium-income customers with high spending scores.
    - Responsive to limited-time offers, seasonal promotions, and targeted advertising.
4. **Potential Loyalists (Cluster 3):**

- Middle-aged, average income and spending.
- Opportunities to increase spending through loyalty rewards and personalized campaigns.

5. **Low Engagement Customers (Cluster 4):**
    - Older customers with low income and spending scores.
    - Minimal engagement; may require reactivation campaigns or minimal marketing investment.

11. Comparison with Hierarchical Clustering

Hierarchical Clustering:

- Hierarchical clustering is another unsupervised learning technique that builds a hierarchy of clusters using either **agglomerative (bottom-up)** or **divisive (top-down)** methods.
- In this project, **agglomerative hierarchical clustering** with **Ward's method** was used, which minimizes the variance within clusters.

Dendrogram Plot:

- A dendrogram was generated to visualize the hierarchy of clusters.
- The vertical axis represents the distance (or dissimilarity) between clusters, and the horizontal axis represents individual customers.
- By observing the dendrogram, a cut-off was made to identify **five clusters**, which aligns with the optimal k determined using K-Means.

Comparison with K-Means Clusters:

- The clusters identified by hierarchical clustering show a **similar pattern of segmentation** as K-Means, confirming the presence of distinct customer groups.
- Both methods consistently highlight groups such as **high-value customers, budget buyers, and impulse shoppers**, reinforcing the robustness of the segmentation.

Confirmation of Cluster Structure:

- Hierarchical clustering serves as a **validation technique** for K-Means results.
- The agreement between both methods suggests that the customer clusters are meaningful and reliable, increasing confidence in actionable business insights derived from the segmentation.

12. Predicting New Customers

Once the customer segmentation model is trained, it can be used to predict the segment of **new or incoming customers**. This enables businesses to target them with personalized marketing strategies from the moment they interact with the company.

Example:

- A new customer has the following attributes:
    - Age: 25 years
    - Annual Income: 60k$
    - Spending Score: 80
    - Income-to-Spending Ratio: 60 / (80 + 1) ≈ 0.74

Predicted Cluster:

- Using the trained K-Means model, this customer was assigned to **Cluster 2**.

Segment Name:

- **Impulse Shoppers** – younger customers with medium income and high spending potential.

Business Application:

- The predicted segment allows businesses to **customize marketing campaigns** immediately:
    - Send promotional offers tailored for high-spending young customers.
    - Recommend products likely to appeal to impulse buyers.
    - Prioritize engagement strategies for high-potential segments.

By applying the segmentation model to new customers, businesses can improve conversion rates, enhance customer experience, and maximize the return on marketing investments.

13. Conclusion

Key Results:

- The customer segmentation project successfully divided mall customers into **five meaningful clusters** using K-Means clustering.
- Clusters were validated using the **Elbow Method** and **Silhouette Score**, achieving a final silhouette score of **0.58**, indicating good separation between clusters.
- Distinct customer groups identified include **High-Value Customers, Budget Buyers, Impulse Shoppers, Potential Loyalists,** and **Low Engagement Customers**.
- Feature engineering, specifically the **Income-to-Spending Ratio**, improved clustering quality and interpretability.

Business Implications:

- Segmentation enables **targeted marketing campaigns** tailored to each customer group.
- **Loyalty programs** can be designed for high-value customers to increase retention.
- **Promotions and discounts** can be offered to budget buyers or impulse shoppers to boost engagement and sales.
- Marketing resources can be allocated efficiently by focusing on high-potential segments.

Limitations of the Model:

- K-Means assumes clusters are **spherical and equally sized**, which may not fully capture complex customer behavior.
- Outliers can affect cluster centroids, potentially skewing results.
- The model relies on a limited set of features (age, income, spending score), which may not capture all aspects of customer behavior.

Possible Future Improvements:

- Explore **density-based clustering (DBSCAN)** or **Gaussian Mixture Models (GMM)** for capturing non-spherical clusters.
- Include additional features such as **purchase frequency, product categories, or online behavior** to improve segmentation accuracy.
- Continuously update the model with new customer data for **dynamic segmentation**.
- Use advanced visualization and interactive dashboards to communicate insights to stakeholders effectively.

14. References

Dataset Source:

- Mall Customers Dataset: Mall_Customers.csv (Kaggle)

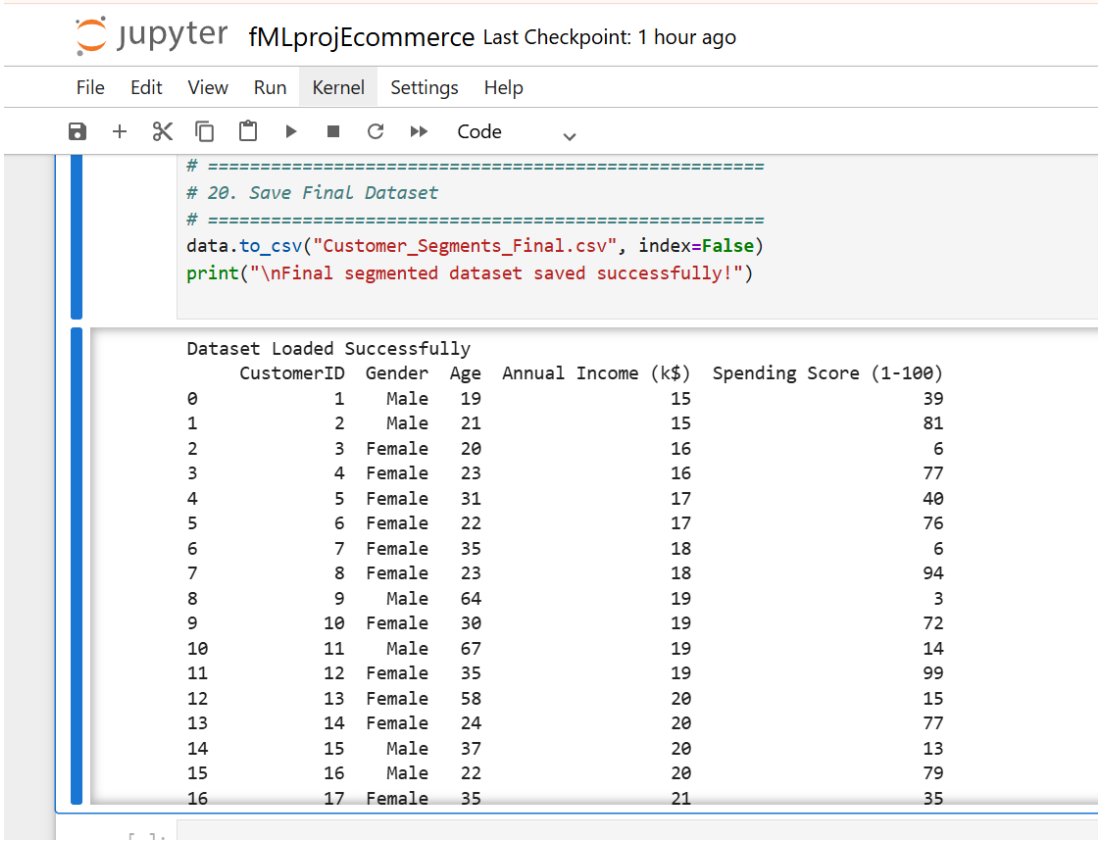
Libraries Used:

- Python (Programming Language)
- Pandas – Data manipulation and analysis
- NumPy – Numerical computations
- Matplotlib – Data visualization
- Seaborn – Statistical data visualization
- Scikit-learn – Machine learning algorithms (K-Means, PCA, StandardScaler, Silhouette Score)
- SciPy – Hierarchical clustering

Tutorials / References Consulted (Optional):

- Official scikit-learn documentation: https://scikit-learn.org/stable/
- Kaggle Customer Segmentation tutorials
- Python documentation for data preprocessing and visualization

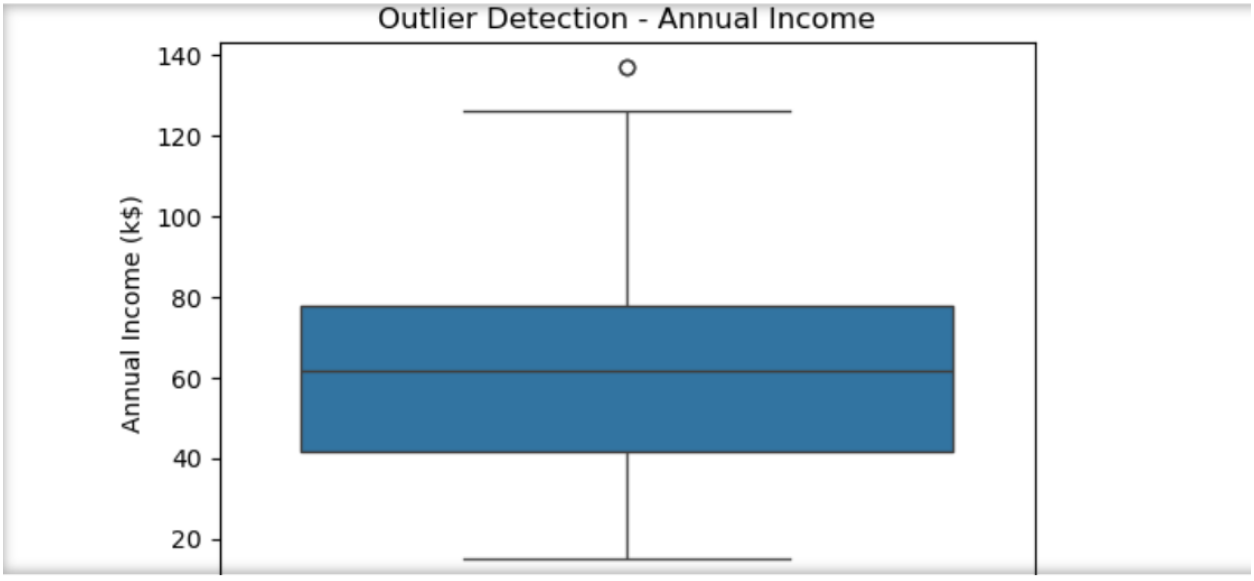
OUTPUT:

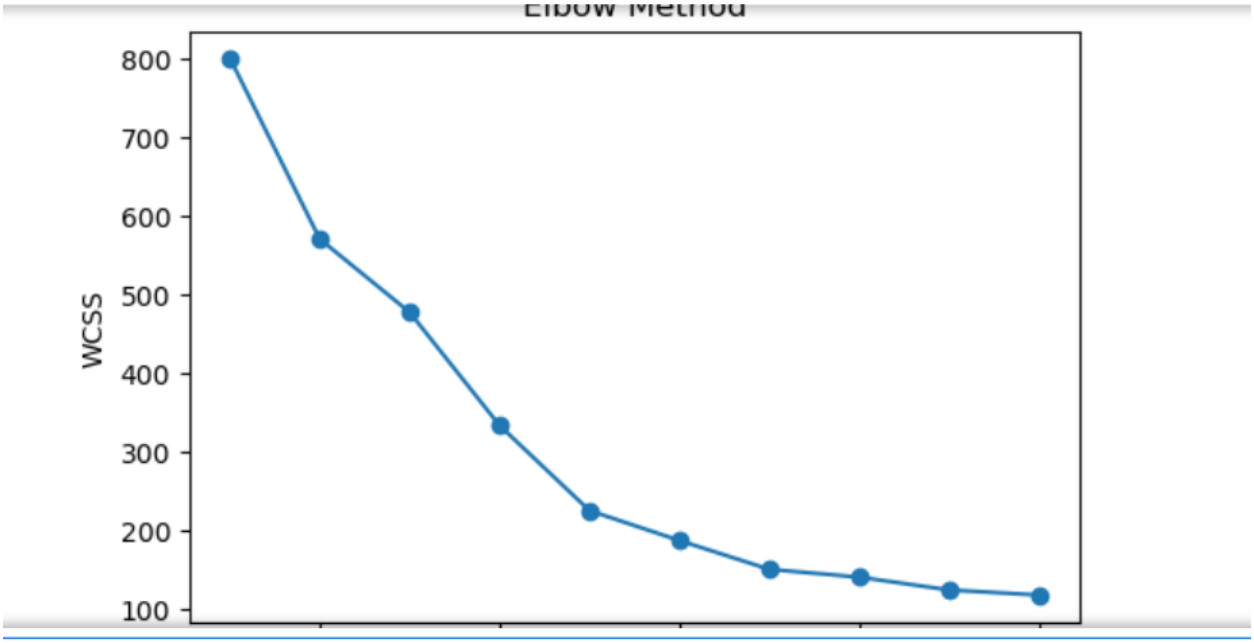

jupyter fMLprojEcommerce Last Checkpoint: 1 hour ago
File Edit View Run Kernel Settings Help
Code
# ==
# 20. Save Final Dataset
# ==
data.to_csv("Customer_Segments_Final.csv", index=False)
print("\nFinal segmented dataset saved successfully!")
Dataset Loaded Successfully
CustomerID Gender Age Annual 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
5 6 Female 22 17 76
6 7 Female 35 18 6
7 8 Female 23 18 94
8 9 Male 64 19 3
9 10 Female 30 19 72
10 11 Male 67 19 14
11 12 Female 35 19 99
12 13 Female 58 20 15
13 14 Female 24 20 77
14 15 Male 37 20 13
15 16 Male 22 20 79
16 17 Female 35 21 35

```
Dataset Shape: (200, 5)

Column Names:
 Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
```

[ ]:

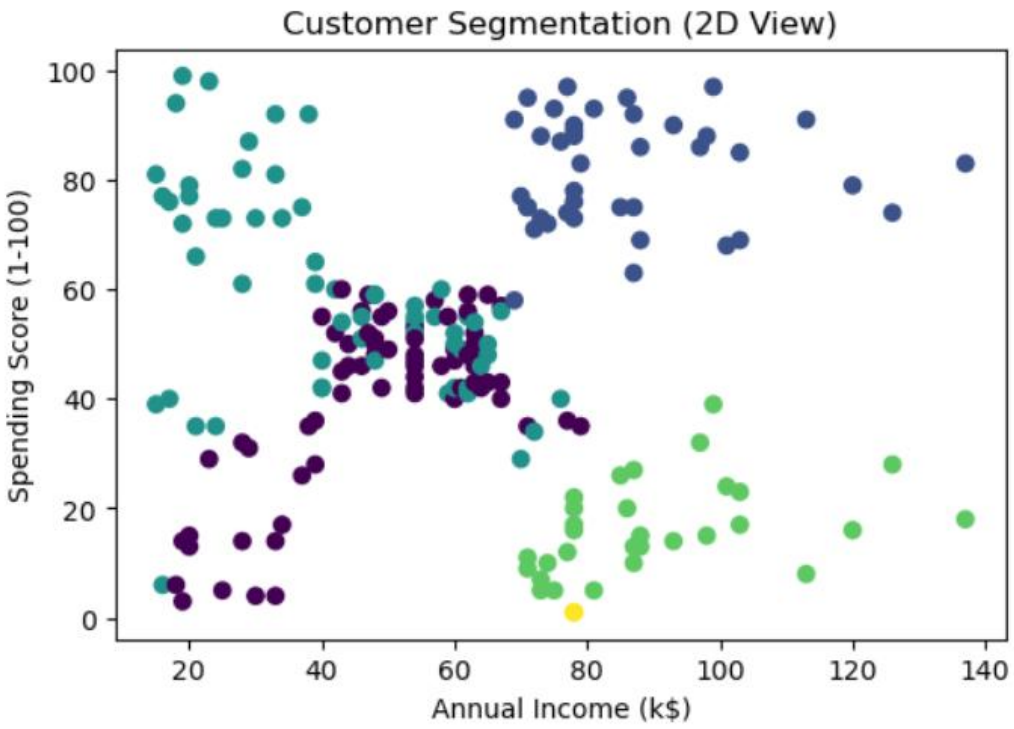
```
Statistical Summary:
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               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

Missing Values:
 CustomerID                0
Gender                     0
Age                        0
Annual Income (k$)         0
Spending Score (1-100)     0
dtype: int64
```

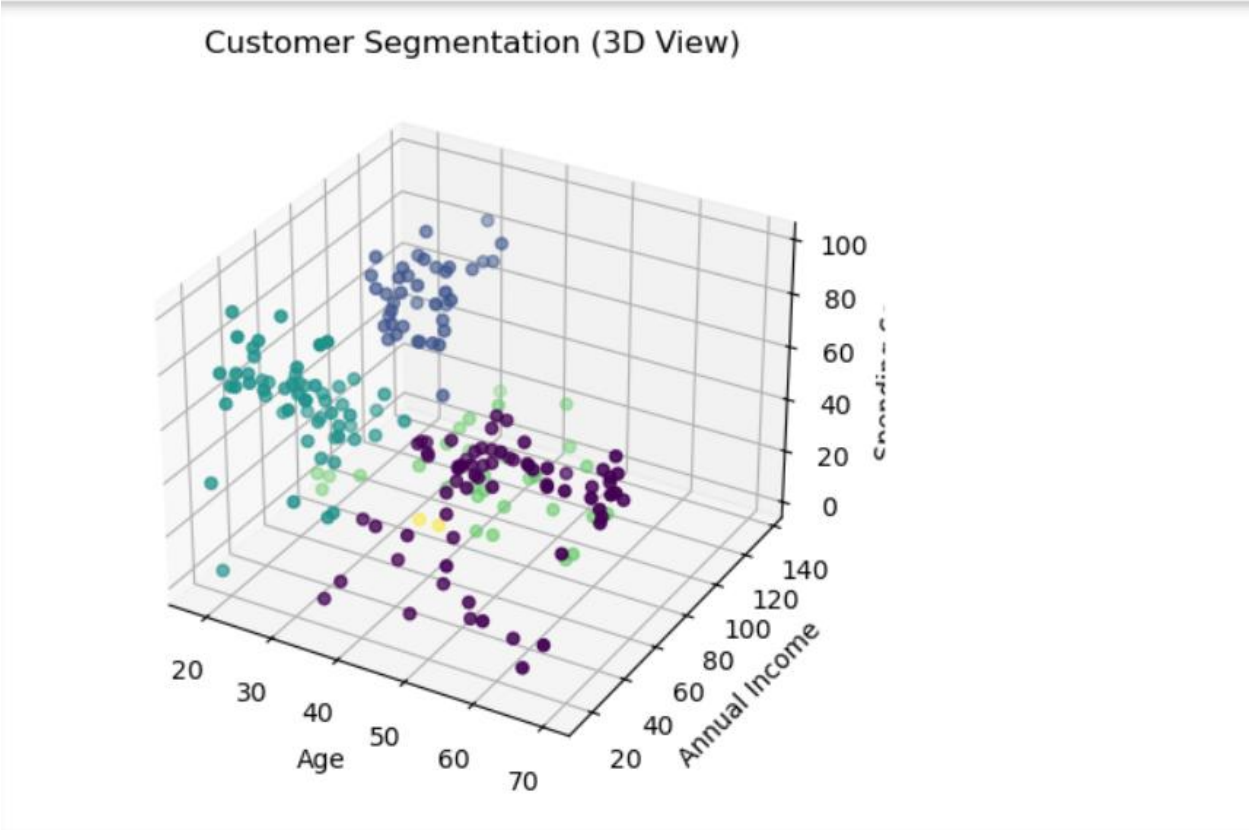


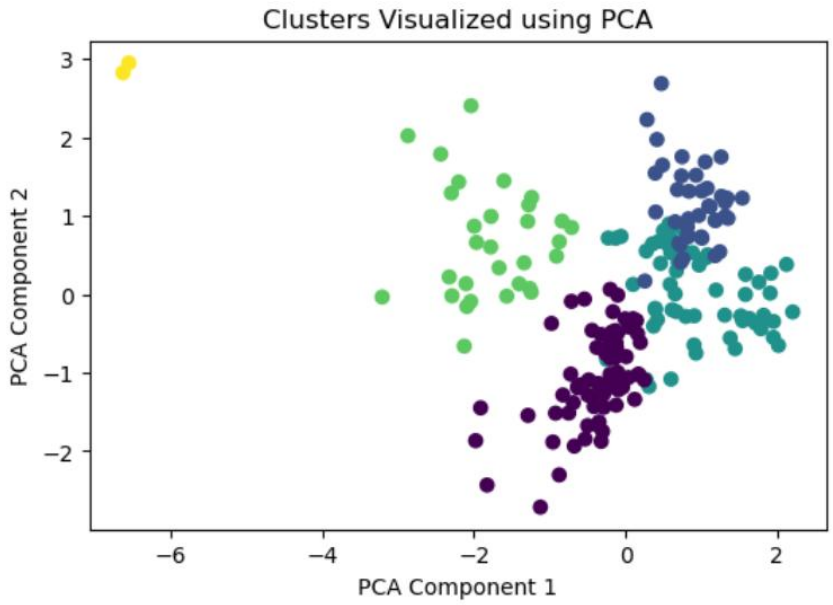



```
Silhouette Score Comparison:
   Silhouette Score
2          0.314258
3          0.309793
4          0.385119
5          0.411149
6          0.419866
7          0.427593

Clustering Completed
   CustomerID  Gender  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

   Income_to_Spending_Ratio  Cluster
0                  0.375000        2
```

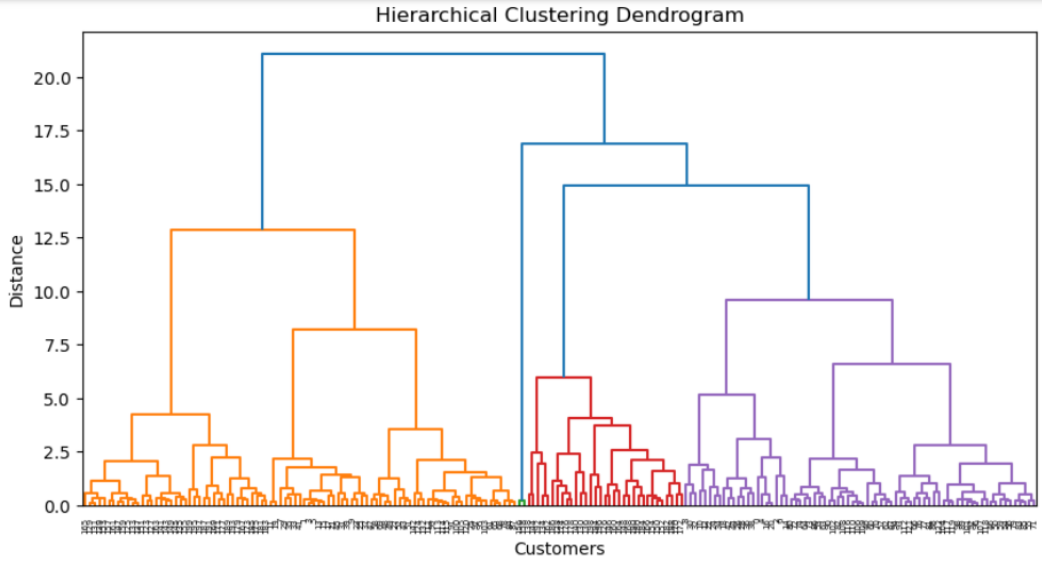
```
Clustering Completed
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)  \
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40

   Income_to_Spending_Ratio  Cluster
0                  0.375000        2
1                  0.182927        2
2                  2.285714        2
3                  0.205128        2
4                  0.414634        2

Final Silhouette Score: 0.41114940625449664
```




Customer Segmentation (3D View)
100
80
60
40
20
0
140
120
100
80
60
40
20
Annual Income
20
30
40
50
60
70
Age


Clusters Visualized using PCA
PCA Component 2
PCA Component 1
3
2
1
0
−1
−2
−6
−4
−2
0
2

```
Cluster Characteristics (Numeric Columns Only):
   Cluster  CustomerID         Age  Annual Income (k$)  Spending Score (1-100)  \
0        0   71.941176   53.191176           48.661765               40.294118
1        1  161.025000   32.875000           86.100000               81.525000
2        2   57.450000   25.433333           41.633333               59.000000
3        3  166.800000   41.366667           90.166667               16.566667
4        4  158.000000   35.500000           78.000000                1.000000

   Income_to_Spending_Ratio
0                  1.466389
1                  1.057971
2                  0.822680
3                  6.273616
4                 39.000000

Segment Naming Done
   Cluster      Segment_Name
0        2  Impulse Shoppers
1        2  Impulse Shoppers
2        2  Impulse Shoppers
3        2  Impulse Shoppers
4        2  Impulse Shoppers
```




```
Predicted Cluster for New Customer: [2]
Predicted Segment: Impulse Shoppers

Final segmented dataset saved successfully!
```