Customer Segmentation Project Report

Introduction

This project focuses on segmenting customers of a mall using the Mall Customer Segmentation Dataset. The goal is to group customers based on their Age, Annual Income (k$), and Spending Score (1-100) to help the mall develop targeted marketing strategies. I performed this analysis in Python using Google Colab, applying K-Means clustering and exploring various improvements to optimize the results.

My Approach

Here's a step-by-step breakdown of how I approached the project:

1. **Loading the Data**
   I loaded the dataset (Mall_Customers.csv) using pandas. It contains 200 rows and 5 columns: CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1-100). I confirmed there were no missing values and reviewed basic statistics:
   - Age: Mean ~38.85, Range 18–70
   - Annual Income (k$): Mean ~60.56, Range 15–137
   - Spending Score (1-100): Mean ~50.2, Range 1–99
2. **Exploratory Data Analysis (EDA)**
   I explored the data to understand its features:
   - **Histograms**: Showed distributions of Age (slightly right-skewed), Annual Income (multi-modal), and Spending Score (fairly uniform).
   - **Pair Plot**: Visualized relationships between Age, Income, and Spending Score, hinting at potential clusters.
   - **Correlation Heatmap**: Low correlations (e.g., Age and Spending Score ~ -0.33), indicating all features could be used for clustering.
3. **Preparing the Data**
   I selected Age, Annual Income (k$), and Spending Score (1-100) as features for clustering. Since these have different scales, I standardized them using StandardScaler to ensure equal weighting in the K-Means algorithm.
4. **Clustering with K-Means**
   - **Elbow Method**: Plotted inertia for k=1 to 10. The curve suggested k=4 or k=6 as potential optimal points.
   - **Silhouette Scores**: Calculated for k=2 to 10. The highest score was **0.43** at k=6, indicating reasonable cluster separation.
   - Trained K-Means with k=6 on the scaled data and assigned cluster labels to the dataset.
5. **Visualizing the Clusters**
   - **2D Scatter Plots**: Plotted Age vs Spending Score and Annual Income vs Spending Score, colored by cluster, showing distinct groups.
   - **3D Plot**: Visualized all three features (Age, Income, Spending Score) in a 3D scatter plot, reinforcing cluster separation.

6. **Understanding the Groups**
   I calculated the mean values of each feature per cluster to interpret them:
   - **Cluster 0**: Older (56), moderate income (54k), moderate spending (49) – "Mature Moderates"
   - **Cluster 1**: Young (32), high income (86k), high spending (82) – "Young Affluent Spenders"
   - **Cluster 2**: Young (25), low income (26k), high spending (76) – "Young Impulsive Buyers"
   - **Cluster 3**: Young (26), moderate income (59k), moderate spending (44) – "Balanced Youngsters"
   - **Cluster 4**: Middle-aged (44), high income (90k), low spending (18) – "Affluent Savers"
   - **Cluster 5**: Middle-aged (45), low income (26k), low spending (19) – "Budget-Conscious"
7. **Evaluating and Improving**
   - **Initial Silhouette Score**: 0.43 with k=6 (reasonable but below 0.5).
   - **Improvements Tried**:
     - **2D Clustering**: Used only Income and Spending Score, best score ~0.44 with k=5.
     - **With Gender**: Encoded Gender (Male=0, Female=1), best score 0.42 with k=10.
     - **PCA**: Reduced to 2 components, score ~0.39 with k=5.
     - The original setup (k=6, three features) remained the best at 0.43.
   - **Alternative Algorithm**: Tried Hierarchical Clustering on 2D features, but it didn't outperform K-Means.

Challenges I Faced

- **Choosing k**: The elbow plot was ambiguous, so I relied on Silhouette Scores, settling on k=6.
- **Silhouette Score**: The best score (0.43) was below 0.5, suggesting some cluster overlap. Improvements didn't significantly boost it.
- **Interpretation**: Naming clusters required careful analysis of mean values, but the patterns were clear enough to define meaningful groups.

How Well It Worked

- **Silhouette Score**: 0.43 with k=6, indicating decent but not perfect separation.
- **Clusters**: Six distinct customer types emerged, actionable for marketing (e.g., targeting "Young Impulsive Buyers" with promotions).
- **Visualizations**: Scatter plots confirmed the clusters were interpretable and separated.

Improvements

While the score of 0.43 is acceptable for this dataset, it could be enhanced by:

- Adding new features (e.g., spending-to-income ratio).
- Trying other algorithms like DBSCAN for non-spherical clusters.
- Collecting more data for better-defined segments.

Conclusion

I successfully segmented mall customers into six groups using K-Means clustering, achieving a Silhouette Score of 0.43 with Age, Annual Income, and Spending Score. The clusters, such as "Young Affluent Spenders" and "Affluent Savers," provide valuable insights for the mall's marketing strategies. This project deepened my understanding of data preprocessing, clustering, and evaluation techniques—a rewarding experience!