Customer Segmentation Analysis Report

Overview

This project analyzed a dataset of 200 customers with features: Age, Gender, Annual Income (k$), and Spending Score (1-100). The objectives were to explore data distributions, evaluate and mitigate skewness, and segment customers using K-Means clustering. Additional tasks optimized cluster counts and applied alternative transformations. We used Python with pandas, Matplotlib, Seaborn, scikit-learn, and scipy, printing all statistics to the console.

Exploratory Data Analysis (EDA)

- **Distributions**: Age and Annual Income showed mild right-skewness; Spending Score was nearly symmetric.
- **Boxplots**: Annual Income had outliers above 100 k$; Age and Spending Score showed few outliers.
- **Scatter Plots**: Annual Income vs. Spending Score indicated potential customer segments; Gender showed no distinct patterns.
- **Pairplot**: Suggested clustering potential in Income vs. Spending Score, with weak Age-Spending Score correlation.

Skewness Evaluation and Transformations

- **Skewness**:
  - Age: 0.486 (mildly positive).
  - Annual Income: 0.322 (mildly positive).
  - Spending Score: -0.047 (nearly symmetric).
- **Initial Threshold**: |Skewness| > 0.5; no transformations needed, as no features qualified.
- **Revised Threshold (Bonus Task)**: |Skewness| > 0.3, applied to Age and Annual Income.
- **Transformation Results**:
  - **Age**:
    - Original: Skewness = 0.486, Mean = 38.850, Std = 13.969.
    - Log: Skewness = -0.089, Mean = 3.623, Std = 0.357.
    - Square Root: Skewness = 0.195, Mean = 6.133, Std = 1.114.
    - Box-Cox: Skewness = -0.016, Mean = 4.763, Std = 0.624.
  - **Annual Income**:
    - Original: Skewness = 0.322, Mean = 60.560, Std = 26.265.
    - Log: Skewness = -0.736, Mean = 4.012, Std = 0.495.
    - Square Root: Skewness = -0.230, Mean = 7.581, Std = 1.761.
    - Box-Cox: Skewness = -0.066, Mean = 20.126, Std = 6.395.

- **Insight**: Box-Cox most effectively reduced skewness to near zero; Square Root was moderately effective; Log overcorrected, introducing negative skewness.

K-Means Clustering

- **Two-Axis Clustering (Annual Income, Spending Score, k=3)**:
    - Cluster 1: (87.00 k$, 18.63) – High-Income, Low-Spenders.
    - Cluster 2: (86.54 k$, 82.13) – High-Income, High-Spenders.
    - Cluster 3: (44.15 k$, 49.83) – Medium-Income, Medium-Spenders.
- **Three-Axis Clustering (Age, Annual Income, Spending Score, k=5)**:
    - Cluster 0: (46.21, 47.72 k$, 41.80) – Older, Medium-Income, Medium-Spenders.
    - Cluster 1: (32.45, 108.18 k$, 82.73) – Middle-Aged, High-Income, High-Spenders.
    - Cluster 2: (24.69, 29.59 k$, 73.66) – Young, Low-Income, High-Spenders.
    - Cluster 3: (40.39, 87.00 k$, 18.63) – Middle-Aged, High-Income, Low-Spenders.
    - Cluster 4: (31.79, 76.09 k$, 77.76) – Middle-Aged, Middle Income, High-Spenders.
- **Insight**: k=5 with Age refined segments than 3
- **Optimal k**: k=5, where the WCSS curve elbowed, indicating effective variance capture with minimal complexity.