Customer Segmentation Analysis

Ibrahim Alghrabi - 201724510

2025-05-07

1. Experimental Setup

Dataset Overview

The dataset consists of 2000 observations of customer purchasing behavior with 8 features. The data includes demographic information such as sex, marital status, age, education level, income, occupation, and settlement size. The data contains no missing values.

Initial exploratory analysis revealed: - Slightly more male than female customers - Equal distribution of marital status - Right-skewed age distribution (mean: 35 years, std: ~12 years) - Most customers have high school education - Right-skewed income distribution (mean: $120,000, std: $38,000) - Most customers are employed or self-employed - Most customers live in small cities

Preprocessing Steps

The preprocessing pipeline included:

1. Selecting relevant features for clustering: Income, Age, and Education
2. Applying log transformation to the Income variable to address its right-skewed distribution
3. Standardizing numerical features (Income and Age) using StandardScaler to ensure equal contribution to the clustering algorithm

Clustering Methodology

I implemented K-means clustering with the following approach:

1. Used K-means++ initialization method to improve convergence and avoid poor local optima
2. Determined the optimal number of clusters using both WCSS (Within-Cluster Sum of Squares) and Silhouette Coefficient
3. Trained the final model with the selected optimal K value
4. Predicted cluster assignments for each customer

2. Visualization and Interpretation

Optimal Number of Clusters

I evaluated K values ranging from 2 to 15 using both WCSS and Silhouette scores:

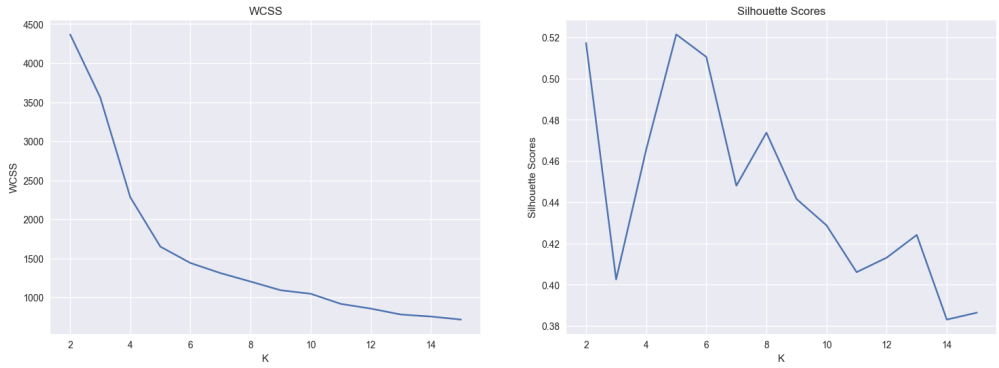

Figure 1: WCSS and Silhouette Scores

Based on the elbow method, the rate of WCSS decrease slows down between K=4 and K=5. 2. A highest score is around $K = 5$. Note that different run could result in differnt graphs, but generaly the performance is better around $K = 5$. . Considering both metrics and the need for practical, interpretable customer segments, I selected K=5 as the optimal number of clusters.

Clustering Results

Since education was a categorical feature, using scatter plot will result in overlapping of point across a horizontal line making it harder to interpret each cluster. Thus, I opted to use swarm plot as shown in Figure 3

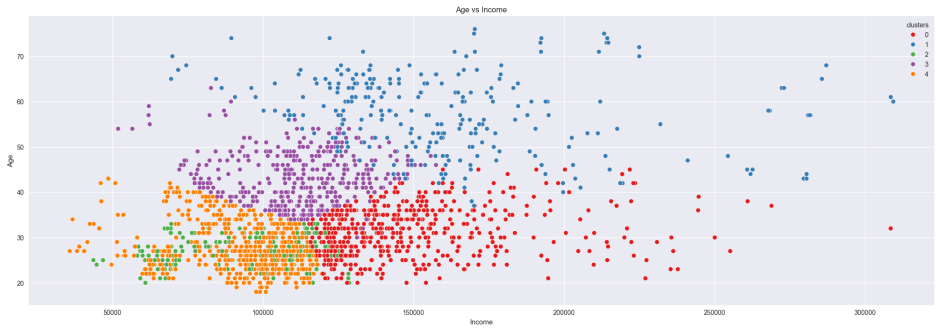

Figure 2: Age vs Income Scatter Plot

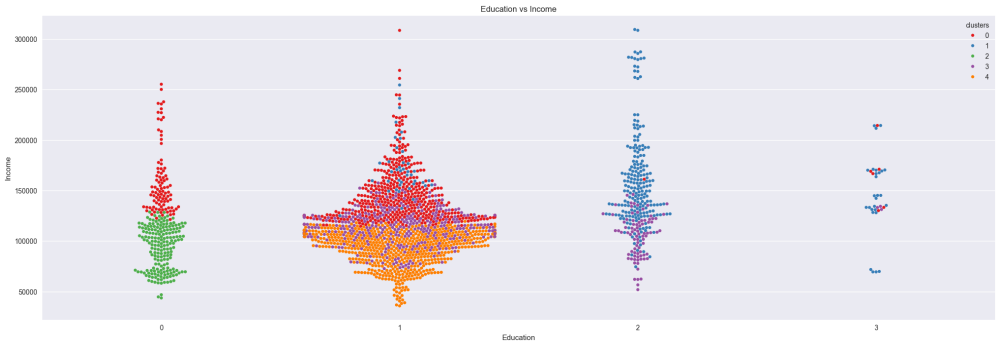

Figure 3: Income vs Education

Cluster Interpretation

Based on the analysis, I identified five distinct customer segments:

Cluster	Count	Income (Mean)	Age (Mean)	Education (Most Common)
0	293	Lowest (70k)	Young 20-40 (30)	Mostly high school (~200)
1	334	Highest (170k)	Young 20-40 (33)	Mostly high school (~250)
2	234	High (150k)	High >40 (60)	Mostly university (~180)
3	442	Mid (110k)	Mid >30 (42)	Mostly high school (~350)
4	697	Mid (110k)	Young <30	Mostly high school (~550)

Marketing Strategies

Cluster 0: Budget-Conscious Young Adults

- **Profile:** Young adults with low income and mostly high school education
- **Strategy:** Offer affordable essential products with clear value propositions. Focus on mobile and social media marketing channels that emphasize cost-effectiveness and practical benefits.

Cluster 1: Affluent Young Professionals

- **Profile:** Young customers with the highest income level, mostly high school graduates
- **Strategy:** Target with premium, trendy products and experiential offerings. Leverage digital influencer marketing and platforms frequented by affluent young consumers.

Cluster 2: Established Professionals

- **Profile:** Older customers with high income and university education
- **Strategy:** Offer high-quality, premium products or services. Emphasize comfort, reliability, and long-term benefits through professional channels and publications.

Cluster 3: Middle-Income Mature Segment

- **Profile:** Middle-aged customers with moderate income, mostly high school educated
- **Strategy:** Market practical family or lifestyle improvement products at mid-range price points. Highlight value and dependability with relatable messaging.

Cluster 4: Young Middle-Income Group

- **Profile:** Young customers with moderate income, predominantly high school educated
- **Strategy:** Attract with promotions for popular items and affordable lifestyle enhancements. Focus on social media engagement to build community and brand loyalty.