Profiling Customer Types: A Segmentation Analysis

Alyssa Juarez

University of Texas at Arlington

DATA 3421

Profiling Customer Types: A Segmentation Analysis

For this project, I chose customer segmentation, a critical analysis area for businesses to understand their customer base. Effective segmentation enables companies to cater to specific customer needs, optimize inventory, and enhance customer retention. For example, identifying young families as frequent shoppers allows stores to stock baby-related items, reducing the number of shopping trips required and ensuring customers spend more within the store.

Preparation

Data Cleaning

The dataset for this analysis was sourced from Kaggle and contained 2,240 entries with 29 features. During cleaning, I removed null values and handled outliers. For instance, I addressed anomalies like a 120-year-old shopper and an income exceeding $600,000, which were unrealistic and could skew the analysis. After these adjustments, the dataset comprised 2,205 entries.

Preprocessing

I performed feature engineering to enhance the dataset's usability, creating new features such as:

- **Customer_For**: Duration of customer loyalty
- **Age**: Customer's age
- **Spent**: Total expenditure
- **Household_Size**: Combined number of adults and children

Categorical variables like "Education" were converted into numerical categories (undergrad, grad, and postgrad). Due to the high dimensionality of the dataset, I applied Principal Component Analysis (PCA) to reduce complexity, ensuring smoother model processing. This step prepared the data for exploratory analysis and descriptive statistics.

Model Selection and Explanation

To analyze the dataset, three clustering models were selected based on their suitability for unsupervised learning and customer segmentation tasks:

1. **K-Means Clustering.** K-Means was chosen due to its efficiency and ability to group data into distinct, non-overlapping clusters. This algorithm is widely used for customer segmentation because it identifies clear, actionable groupings in the data (Lloyd, 1982).
2. **Agglomerative Hierarchical Clustering.** This model was selected to explore hierarchical relationships within the dataset. Agglomerative clustering builds clusters iteratively, offering insights into the grouping process through dendrograms (Rokach & Maimon, 2005). This approach was expected to provide more granular insights into the structure of customer segments.
3. **K-Modes Clustering.** K-Modes was included as an alternative for handling categorical data effectively. As an adaptation of K-Means, K-Modes optimizes clustering for datasets with categorical variables, making it particularly relevant for customer segmentation tasks with mixed data types (Huang, 1998).

These models were chosen to ensure comprehensive analysis, balancing interpretability, computational efficiency, and compatibility with the dataset.

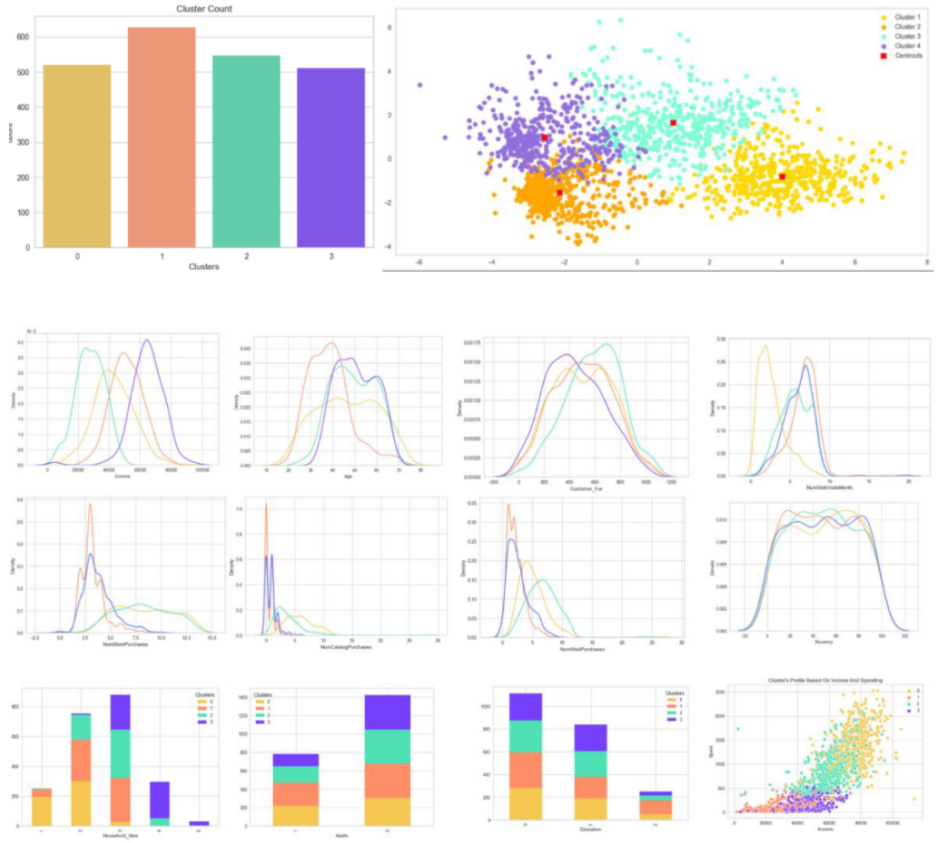
Results and Insights

K-Means Clustering

The K-Means algorithm, configured with four clusters, yielded the most balanced and interpretable segmentation results. The clusters were defined as follows:

- **Group 1.** Moderate income (20–65k), high spenders, primarily without children, luxury-oriented, coupon-driven, and forming the largest segment of shoppers.
- **Group 2.** Low to moderate income (20–80k), small families, low spending habits, and limited luxury purchases.
- **Group 3.** Low income (10–50k), single or mature small families, occasional luxury purchases.
- **Group 4.** High income (30–100k), large families, low luxury expenditure, frequent coupon usage.

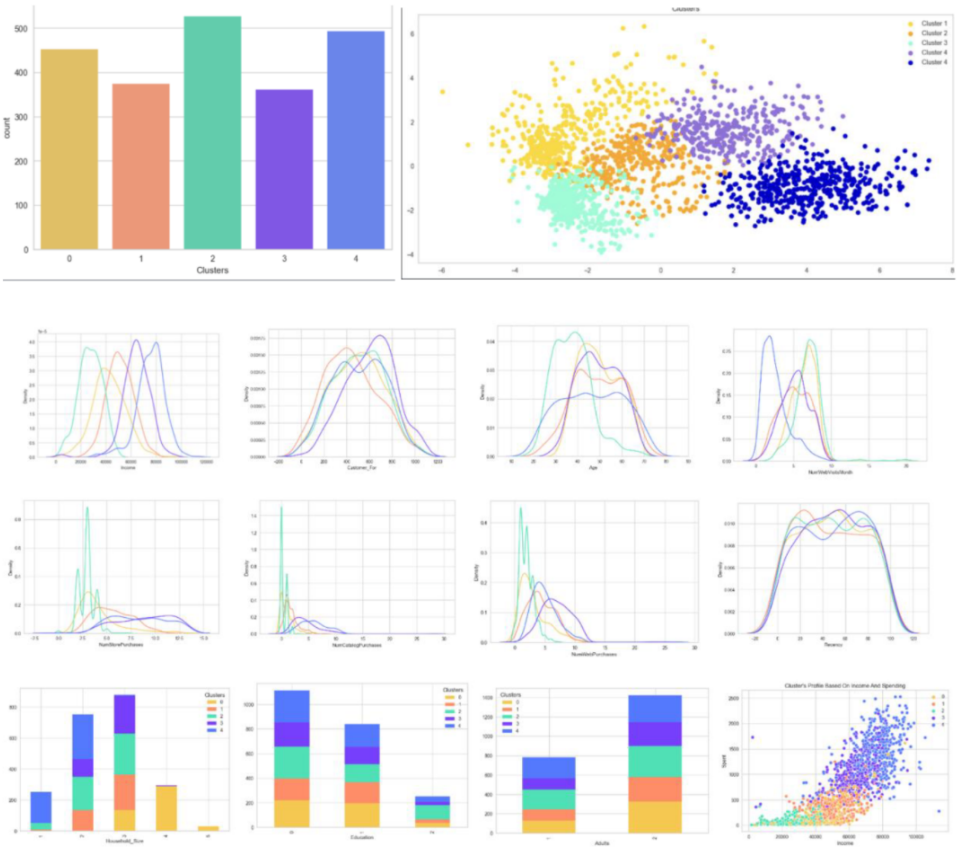
This model provided actionable and well-balanced customer segments, making it the most effective tool for this dataset.



Agglomerative Hierarchical Clustering

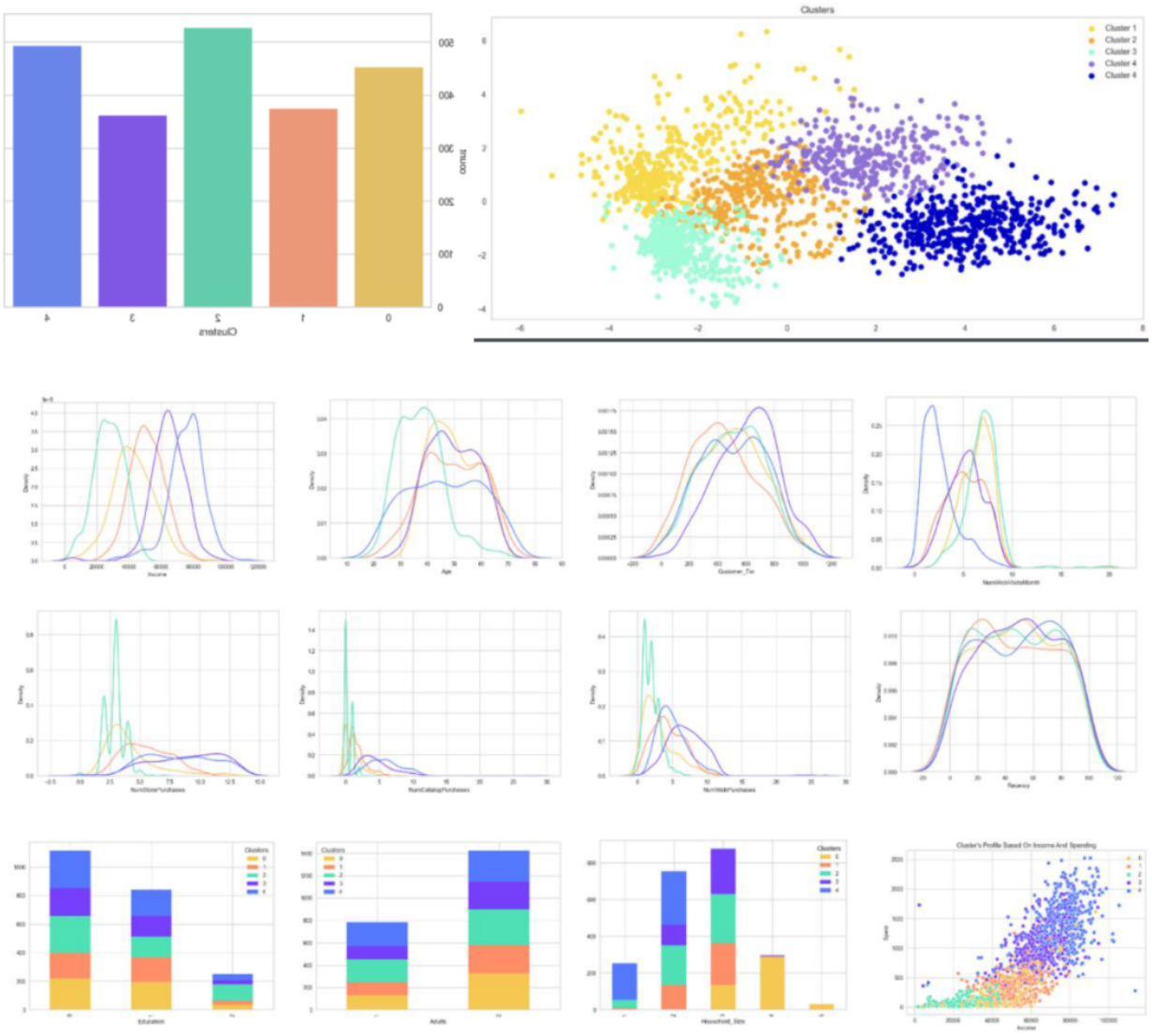
The hierarchical model produced five clusters. While this initially appeared to offer more detailed segmentation, the results were less balanced and harder to interpret effectively:

- **Group 1.** Moderate income (0–80k), large families, moderate luxury purchases.
- **Group 2.** Middle income (20–90k), older families, moderate luxury purchases, highly coupon-motivated.
- **Group 3.** Low income (0–50k), young families, balanced spending on groceries and luxury items.
- **Group 4.** Higher income (30–100k), older families, primarily luxury buyers.
- **Group 5.** Highest income (35–120k), adults without children, grocery-focused shoppers with significant disposable income. Despite the additional cluster, the insights were less actionable compared to those derived from K-Means clustering.



K-Modes Clustering

The K-Modes clustering results closely mirrored those of the agglomerative model. This redundancy indicated limited additional insights for this dataset, suggesting that K-Modes did not add significant value in this specific context.



References

- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283-304. https://doi.org/10.1023/A:1009769707641
- Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489
- Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer. https://doi.org/10.1007/0-387-25465-X_15