Luke Catalano

Consumer Segmentation

K-Modes Clustering for Coffee Customer Survey Data

Project Introduction

Summary

- Clustering is a powerful unsupervised learning technique that can uncover natural patterns in consumer activity and allow us to classify or “segment” them based on historical behaviour

- Data is constantly being collected on consumers, and this is just one of the many ways businesses can utilize this information to inform pricing and capacity planning

Motivation

- Standard clustering algorithms like K-Means rely on Euclidean distance, making them ineffective for categorical data common in real-world transaction logs (e.g., price brackets, frequency labels)

- In competitive retail markets a "one-size-fits-all" marketing strategy is inefficient. High-value consumers require retention incentives, while price-sensitive "casuals" respond better to discounts.

- This project was motivated by the challenge of applying unsupervised learning to strictly categorical consumer data without losing interpretability

Data Origin & Cleaning Pipeline

Source & Scope

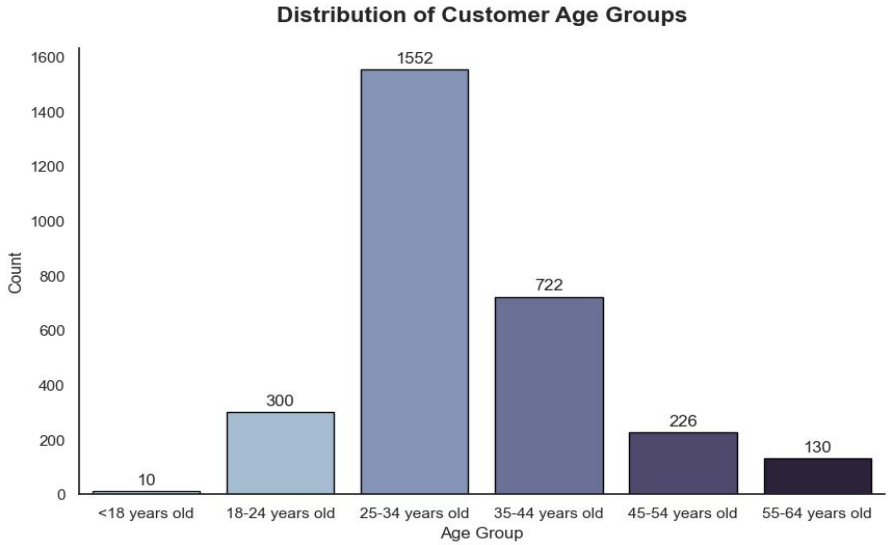
- **Origin:** "Great American Coffee Taste Test" (2023), an open-source survey analyzing consumer preferences and purchasing habits.
- **Initial Volume:** Raw dataset contained **4,042** survey responses sourced via the Tidy Tuesday repository.
- **Features:** The dataset includes a mix of:
    - **Demographic Variables:** Age, Gender, Employment Status.
    - **Behavioral Variables:** Cups per Day, Brewing Method, Favorite Drink, Monthly Spend.
    - **Preference Variables:** Subjective taste test scores and coffee specificities.

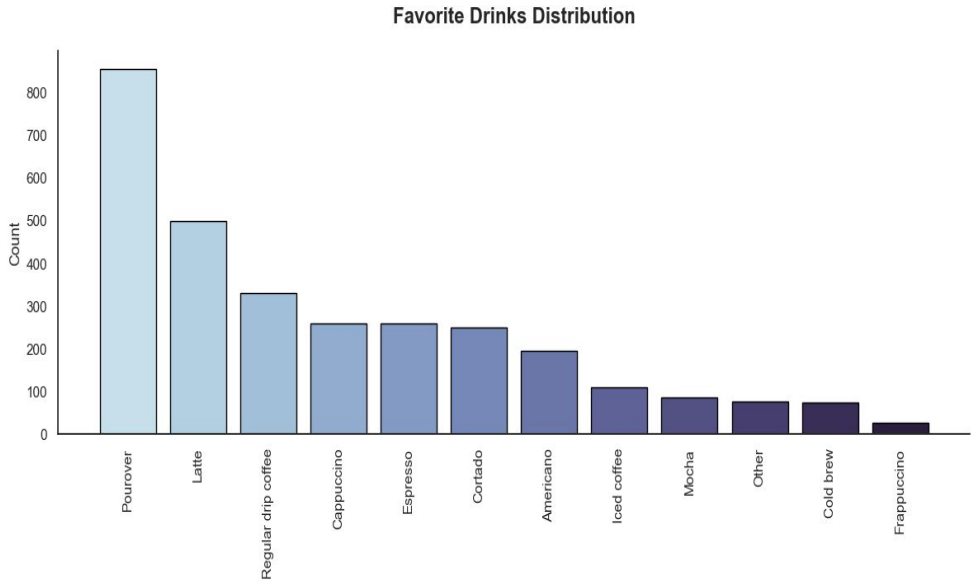
Quality Assurance (QA)

- **Data Integrity Check:** Rigorous filtering was applied to remove incomplete consumer profiles, ensuring that the K-Modes algorithm—which is sensitive to missing values—received a complete feature set.
- **Final Sample:** The cleaning pipeline retained **3,002 high-quality responses** (74% retention rate).
- **Implication:** This strict inclusion criteria eliminated noise and ensured that the resulting clusters represent fully defined consumer archetypes rather than artifacts of missing data.

Exploratory Data Analysis

When individuals of a sample are extremely similar in characteristics, it becomes much more difficult to categorize them based on the observable data, so its best to check key distributions prior to modeling.





N = 3002

K-Modes Clustering

The Challenge: Categorical Constraints

- Standard K-Means clustering relies on Euclidean distance (calculating the mean), which is mathematically impossible for categorical fields like price brackets (e.g., "$50-$100") or purchase frequency labels.
- One-hot encoding would have resulted in high dimensionality and sparse vectors, diluting the cluster quality.

The Solution: K-Modes Algorithm

- **Dissimilarity Metric:** Utilized the **K-Modes** algorithm which calculates distance based on the number of matching categories between data points (Hamming distance) rather than geometric distance.
- **Centroids:** Clusters are defined by **modes** (most frequent values) rather than means, preserving the interpretability of the categorical attributes.
- **Initialization:** Applied 'Huang' initialization to select starting centroids based on frequency, optimizing convergence speed.

Model Optimization & Reproducibility

We can apply a quantitative framework to score our model's ability to efficiently cluster observations as we increase the number of clusters

The Cost Function (Elbow Method)

- The K-Modes Clustering algorithm seeks to minimize a cost function which measures the total "error" or "dissimilarity" in the model.
- As we add more clusters, Cost always goes down because groups get smaller and more specific. The "Elbow Method" involves looking for the specific point where the curve flattens out, where adding another cluster stops providing a significant decrease in the cost function
- This point will serve as our optimal amount of clusters

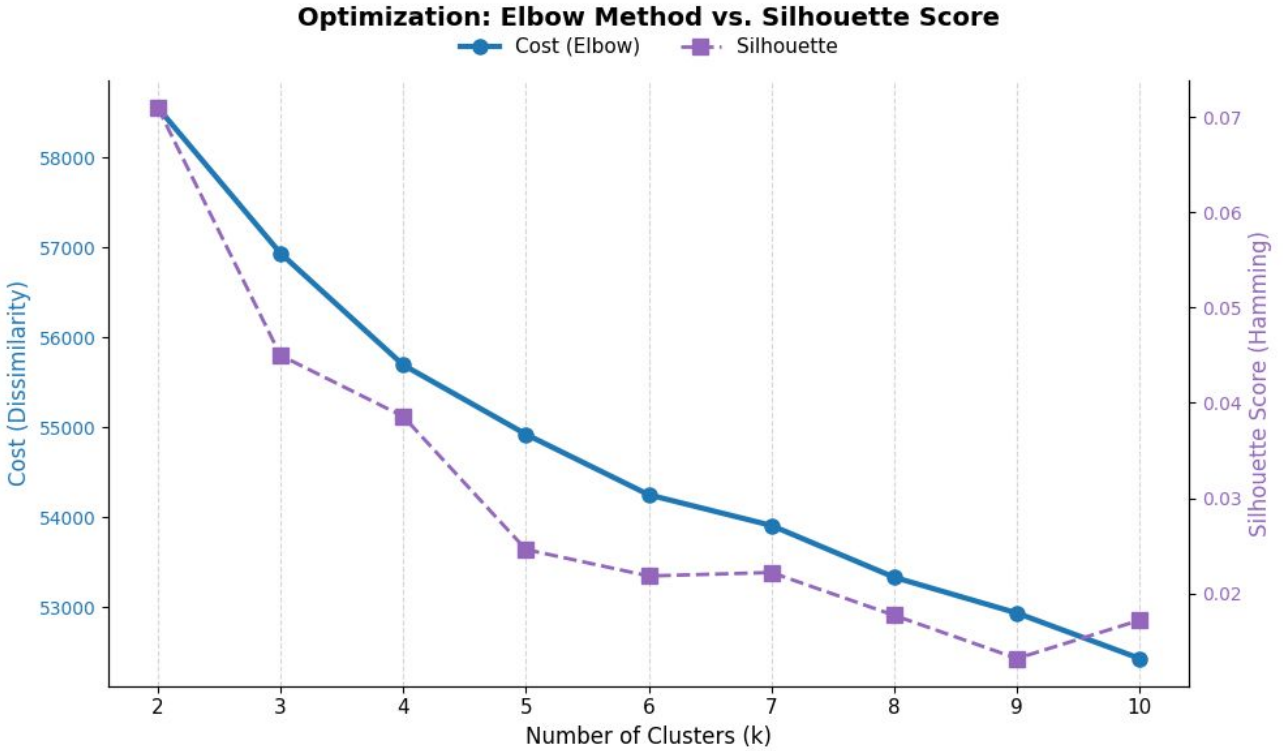
The Silhouette Score

- Compares how similar an observation is to its own cluster (cohesion) versus how similar it is to the nearest other cluster (separation).
- We generally want to maximize this score, however we can pair this metric with intuition to establish the optimal cluster count.
- For example, a cluster count of two will often maximize Silhouette Score, but provide minimal segmentation between observations, binning them into 2 distinct groups.

Model Optimization & Reproducibility

Optimal Cluster Amount:

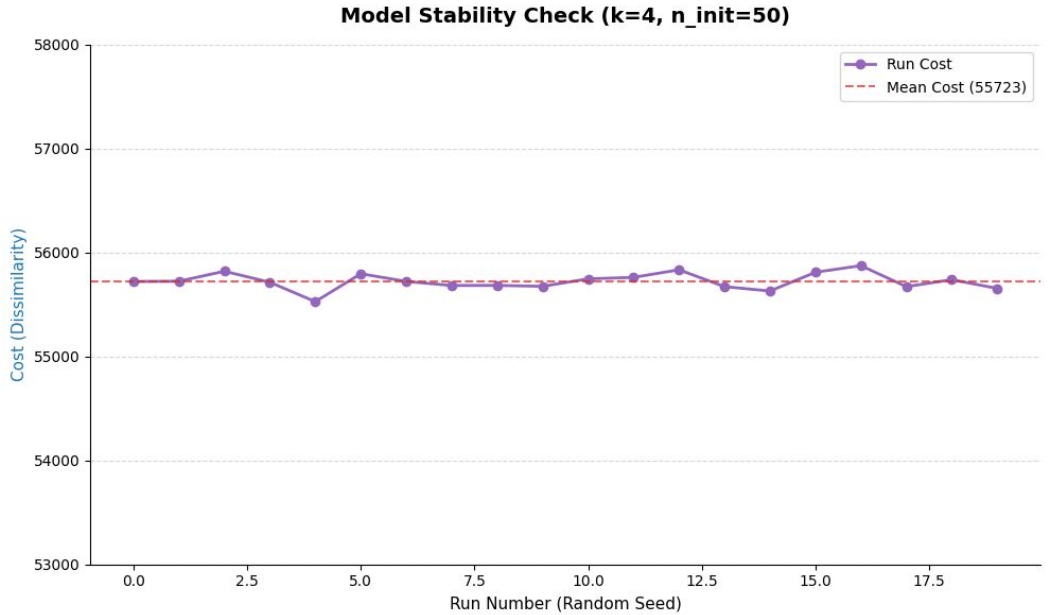
- By combining both the Cost Function and Silhouette Score, we can see that the Cost function sees maller losses after 4 clusters, and this is precisely the cluster count before the Silhouette Score falls drastically
- While the Silhouette Score is higher at 3 clusters, the significant decrease in the cost function and the added interpretability of an extra customer segment outweighs this larger score



Model Optimization & Reproducibility

Model Stability Assessment:

- The K-Modes simulation was ran **20 independent times** with unique random seeds, forcing the algorithm to restart its search pattern from scratch each time.
- **The Result:** The model converged to around the same solution of approx. Cost = 55,723 in every iteration.
- **Conclusion:** The segment definitions are **structurally robust**. If we ran this analysis again next month with new data, the archetypes would likely hold.



Now that we know how many customer segments we have, and that we can reproduce our results, it's time to analyze them.

Feature Engineering for Analysis

To visualize the relative impact of each customer segment, I implemented a **Latent Variable Extraction** pipeline:

- **Parsing:** Deconstructed categorical pricing string bounds into numeric features.
  - *Input:* "$50-$100" ⟶ *Output:* Lower: 50, Upper: 100
- **Aggregation:** Calculated the mean of these extracted bounds for every user within a cluster to establish a "Cluster Centroid."
- **Normalization (Min-Max Scaling):**
  - The centroids were normalized to a **0.0 – 1.0 scale**.
  - **0.0** represents the lowest spending bound in the dataset.
  - **1.0** represents the highest spending bound.
  - Normalization is used only for visualization and comparison, not as an input to clustering.
- These features allow us to construct the relative spending intensity profile for each cluster

Consumer Behavior Analysis

Cluster 0: The "Premium Customers"

- **Dominant Performers:** This group hits the outer boundary (1.0) on every single metric, maximizing both "Total Spend" and "Most Paid." They are price-insensitive and volume-heavy. These are your ideal candidates for premium subscriptions, high-margin upsells, and exclusive "gold tier" rewards.

Cluster 1: The "Core Regulars"

- **The Reliable Baseline:** Sitting comfortably in the upper-middle range (~0.6), this group likely represents the daily habit-formers. They are consistent spenders who value quality but aren't necessarily buying the most extravagant items every visit. Retention strategies should focus on consistency and loyalty perks.

Cluster 3: The "Budget Conscious"

- **Low-Tier Engagement:** Compressed into the lower quartile (~0.3), these customers engage with the brand but are highly constrained by price.They likely stick to entry-level items or visit infrequently. They are the primary target for discount coupons or "value bundles" designed to increase their basket size.

Cluster 2: The "Minimalists"

- **Lowest Value / At-Risk:** Concentrated at the absolute center (0.0), this group has negligible spending footprints. They likely purchase only the cheapest item (pastries, drip coffee) or are at high risk of churning. Marketing spend here yields the lowest ROI.