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

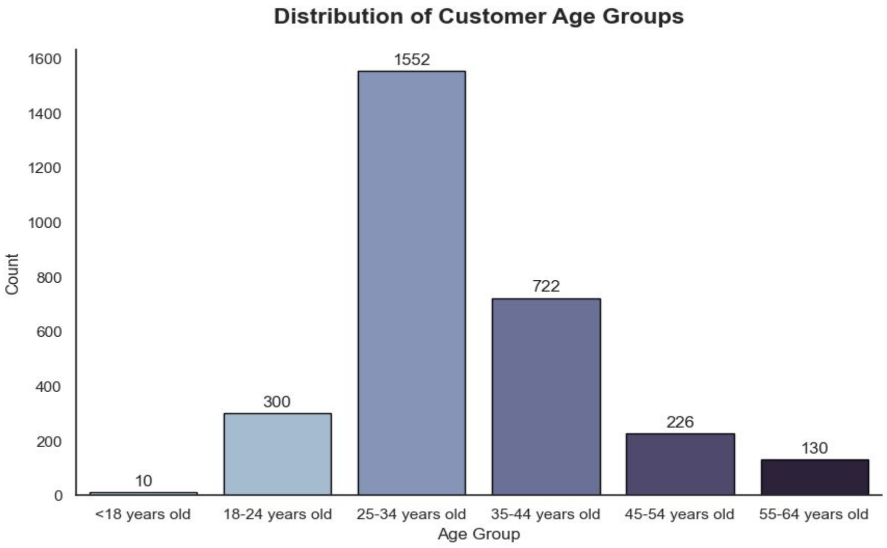
- **Origin:** "Great American Coffee Taste Test" (2023), an open-source survey analyzing consumer preferences and purchasing habits.
- **Initial Volume:** Raw dataset contained **4,042** survey responses sourced via the Tidy Tuesday repository.
- **Features:** The dataset includes a mix of:
  - **Demographic 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.

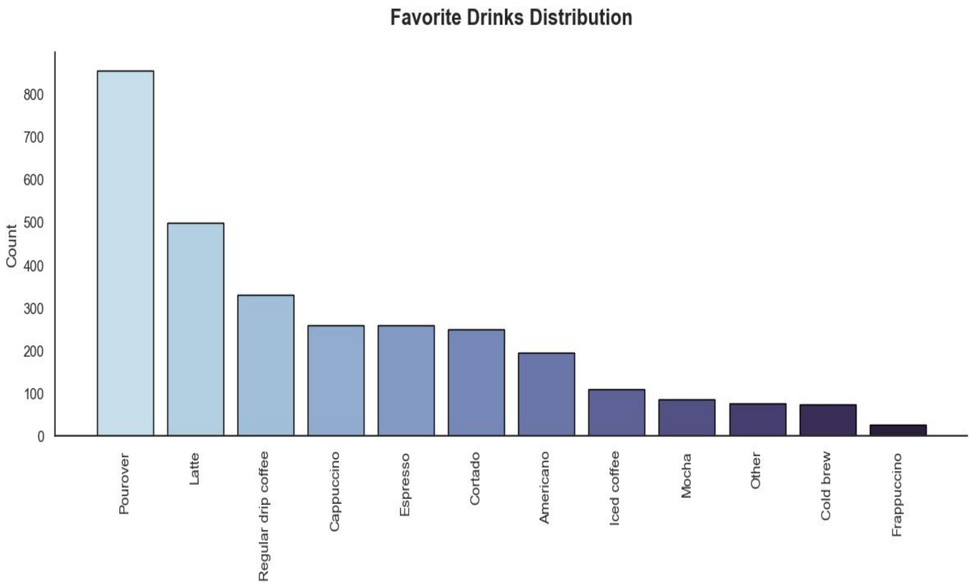
Quality Assurance (QA)

- **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.

Feature Engineering for Analysis

To visualize the relative strength of each cluster, I implemented a **Latent Variable Extraction** pipeline:

- **Parsing:** Deconstructed categorical 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 (Blue): The "Big Ticket" Buyers.** They dominate the single-purchase metric (`max_most_paid`) but have lower total volume, suggesting they buy expensive items (like equipment or bulk) rather than daily consumables.

- **Cluster 1 (Orange): The Core Middle.** These customers sit squarely in the average range for all metrics; they are reliable but price-sensitive, representing the mass-market baseline for revenue.

- **Cluster 2 (Green): Budget Tier.** Concentrated at the center with the lowest scores everywhere, this group spends the absolute minimum and represents the lowest value or highest churn risk.

- **Cluster 3 (Red): The Premium Segment.** The most valuable segment, hitting the maximum bounds on every axis; they consistently spend the most per visit and per month.

- **Cluster 4 (Purple): The Upsell Candidates:** Statistically similar to the Core Middle but with slightly higher spending ceilings, making them the ideal target audience for campaigns designed to move users into the premium tier.