Luke Catalano

# Stat-155 Final Presentation

Coffee Consumer Clustering Project Journey

# Introduction

## Summary

- Exploring consumer trends of American coffee enthusiasts

- Attempting to group consumers through an unsupervised clustering algorithm using numerous features from a coffee survey data set

## Motivation

- Worked as a Business Intelligence Intern last summer, and was tasked with finding new ways to group current customers into distinct categories

- Worked as a barista at Starbucks

# Data Source

- 2023 survey of “*Great American Coffee Taste Test*” viewers
- This data is downloaded directly from a web URL every time a project script is ran.

- ***Tidy Tuesday*** provided a csv containing 4042 valid survey responses, and a cleaning script which turned the survey responses into discrete and continuous variables.

- 2970 survey responses had zero NA values, and were viable for modeling and analysis.
- This is a considerable drop in valid responses, potentially harming analysis.

# Exploratory Data Analysis

**Key Features:**

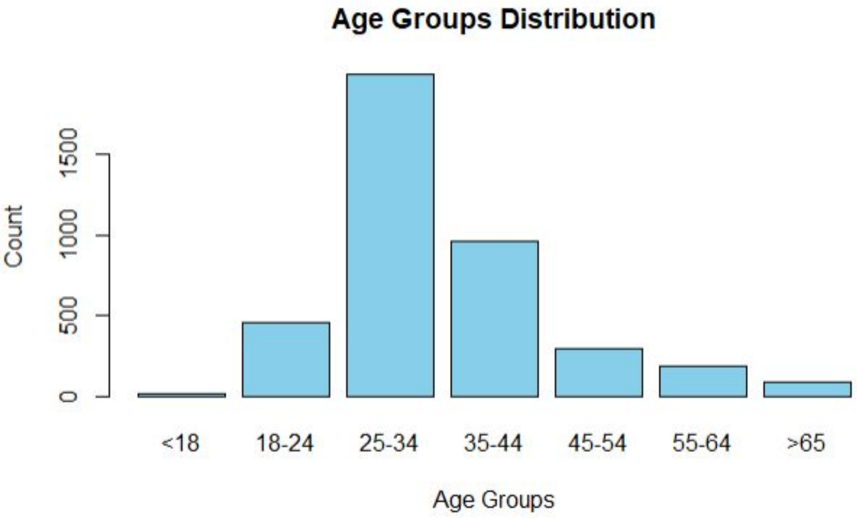
- Favorite Coffee Drink
- Cups of Coffee Per Day
- Favorite Coffee Spot
- Do you Work from Home?
- Age, Education, Gender
- Avg Monthly Coffee Spending
- Do you Brew Coffee At Home?

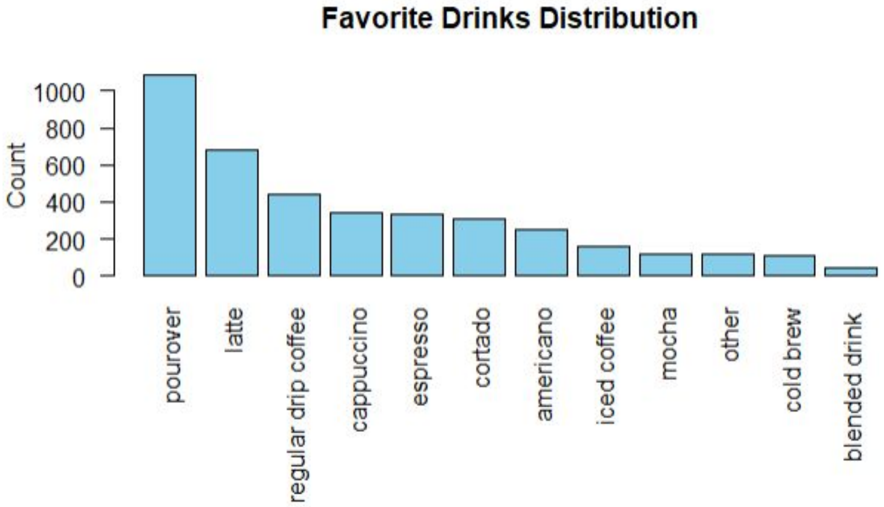
Pourover/Drip Coffees are highly preferred amongst every subgroup based on age.

| Age Group | Favorite Drink | Count |
|---|---|---|
| <18 | Latte | 8 |
| 18-24 | Pourover | 103 |
| 24-34 | Pourover | 566 |
| 35-44 | Pourover | 273 |
| 45-54 | Pourover | 78 |
| 55-64 | Drip Coffee | 40 |
| >65 | Drip Coffee | 32 |

# 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.



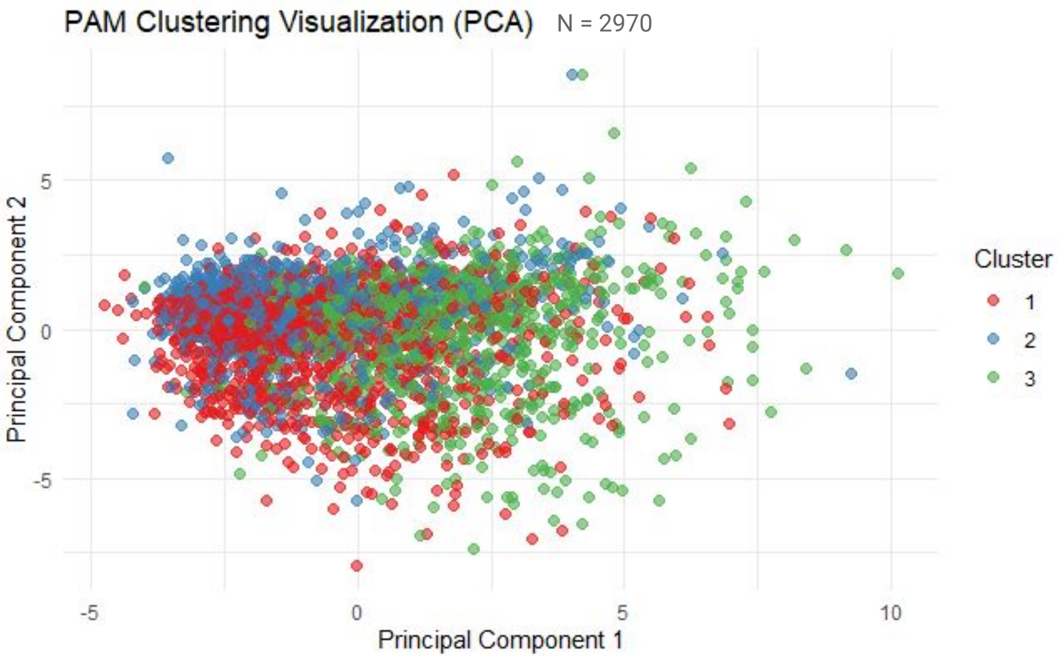




N = 4042

# Modeling - PAM

- Used Grower Distance and PAM (Partition Around Medoids) to assign each consumer to a cluster
- Grower Distance normalize the differences between each pair of features and then averages these differences, resulting in a matrix of differences.
- PAM utilizes these grower distances to find naturally occurring groupings in the consumer dataset, minimizing the average differences between each observation

## Principal Component Plot




**Cluster sizes**

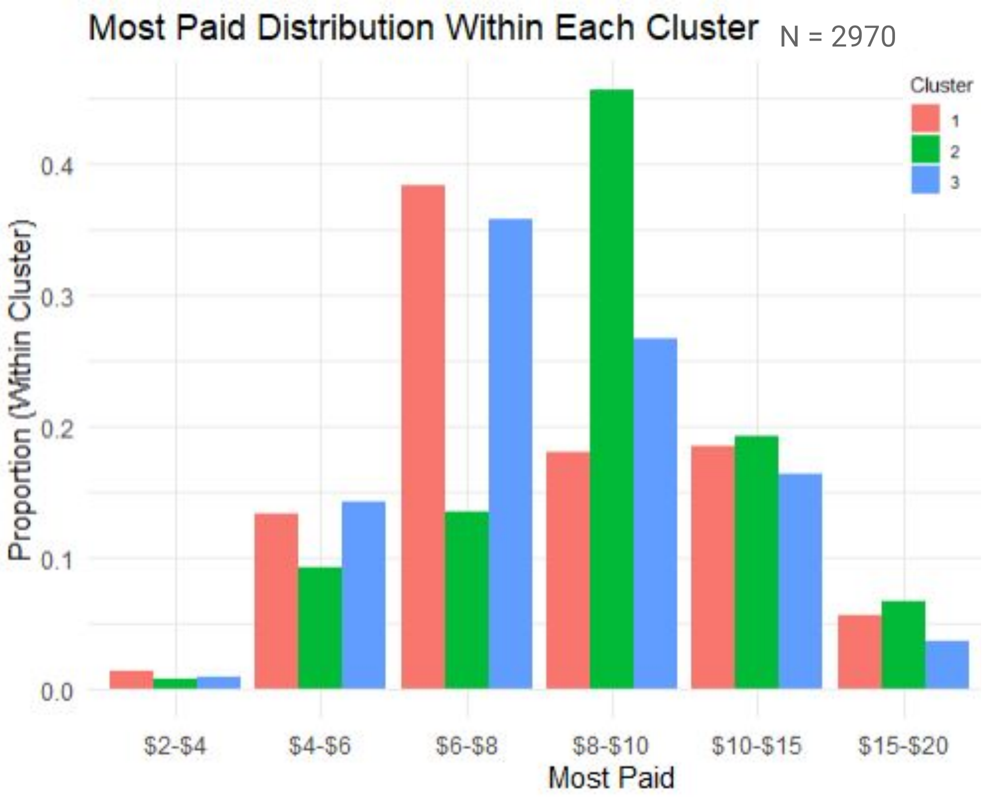
| 1 | 2 | 3 |
|---|---|---|
| 1218 | 827 | 925 |

# Modeling - Results

- Principal Component Analysis flattens the dimensions of the data, which in turn may be causing the clusters to appear relatively close to each other.

- Further Analysis can be done to see the differences in certain key variables between the 3 clusters.

- Here we explore the “Most_Paid” variable, showing the distribution per cluster.



# Monte Carlo Simulation – Proximus

## Experimental Design

### Factors: / Levels:

| Factors: | Levels: |
|---|---|
| Sample Size | - {500, 1500, 2970} |
| 90% Column Sparsity | - {1 Col, 3 Cols, 6 Cols} |

- Data is converted into a logical matrix of true or false questions (Do you value coffee from a cafe?... etc)

- Each Factor-Level will have 10 simulations with randomly generated data, that fits the current distribution of True / False in the current dataset (with some normalized error)

- Each simulation is a proximus model constructed off of the generated data. We will compare the Jaccard Similarity Error between all 90 of the simulations.

## Key Findings

| Obs | Sparse | Mean_Jsim | SD_Jsim |
|---|---|---|---|
| 500 | 1 | 0.942 | 0.0199 |
| 500 | 3 | 0.925 | 0.0158 |
| 500 | 6 | 0.898 | **0.0432** |
| 1500 | 1 | 0.943 | 0.0275 |
| 1500 | 3 | 0.907 | 0.0258 |
| 1500 | 6 | 0.884 | 0.0259 |
| 2970 | 1 | 0.936 | 0.0135 |
| 2970 | 3 | 0.888 | 0.0343 |
| 2970 | 6 | 0.88 | 0.0141 |

*** 10 Reps Each***

We find that as matrix sparsity increases, the Jaccard Similarity measure decreases. Sample size has little effect on the average Jaccard Similarity score

# Summary

## Key Findings

- While it is possible to group coffee consumers using survey data, the population needs to be diverse enough, and the questions revealing enough, to create distinct customer classes.

- Our current sample size is likely too small, and too similar, to construct distinct classifications

## What I've Learned

- Reproducibility is not as simple as providing a link to your data, and there is an extensive process to ensure others can accomplish a valid reproduction of your work.

- RStudio has far more features than I learned in courses like Econ 114 and Econ 124, and the reproducibility features like QMD's and RStudio Projects are extremely useful for keeping track of extensive projects.

# Reflection and Considerations

## Important Considerations for Clustering

- During the EDA stage, ensure your sample is diverse enough for high quality groupings
- Ensure the data has questions that can effectively group individuals

## Project Considerations

- Especially when working with recursive clustering algorithms, always have a virtual backup of your work (Github)
- Regardless of if data is pulled from the web or not, keep a local copy of the data so you can work while offline