Alyssa Juarez

Dr. Rostami

DATA 3421

April 26, 2023

Customer Segmentation

Introduction

For this project, I choose Customer Segmentation. It is crucial that stores know who is shopping so that they are able to cater to the needs of their customer base. So that if we find young families shop at a store, we can make sure to have baby-food, formula and other such items on hand simultaneously reduce the number of stores the customer needs to go in while making sure they spend the most at our store.

Preparation

Cleaning

In this project, I found a dataset that had 2240 features with 29 features from Kaggle. During the cleaning process, I removed any and all null values and trimmed outliers. Though most of my categories did have outliers, I thought it only relevant to only eliminate the outliers in the columns of age and income since the other outliers are crucial to my analysis. There was definitely a typo or rare anomaly in the age, as I had a 120-year-old shopper! I also had an individual who had an income of over 600K! Removing these outliers was crucial to the dataset and left me with 2205 instances.

Preprocessing

To preprocess, I began with feature engineering. From the current features, I created Customer_For (how long they have been a customer), Age (customer age), Spent (how much customer spent), Adult (number of adults), Children (number of children), Household_Size(total number of individuals in household). This allowed further insight into my dataset and overall aided me in making a clear distinction between groups. I also had to turn my categorical variable, "Education" into numerical. I decided to divide it into undergrad, grad and postgrad and assigned numerical value to each category. Finally, due to the number of variables and the correlation, I had to scale and perform PCA to reduce the dimensionality. Doing so makes the data easier for my models to process. After this, I was able to do my exploratory data analysis and collect the descriptive statistics for my dataset.

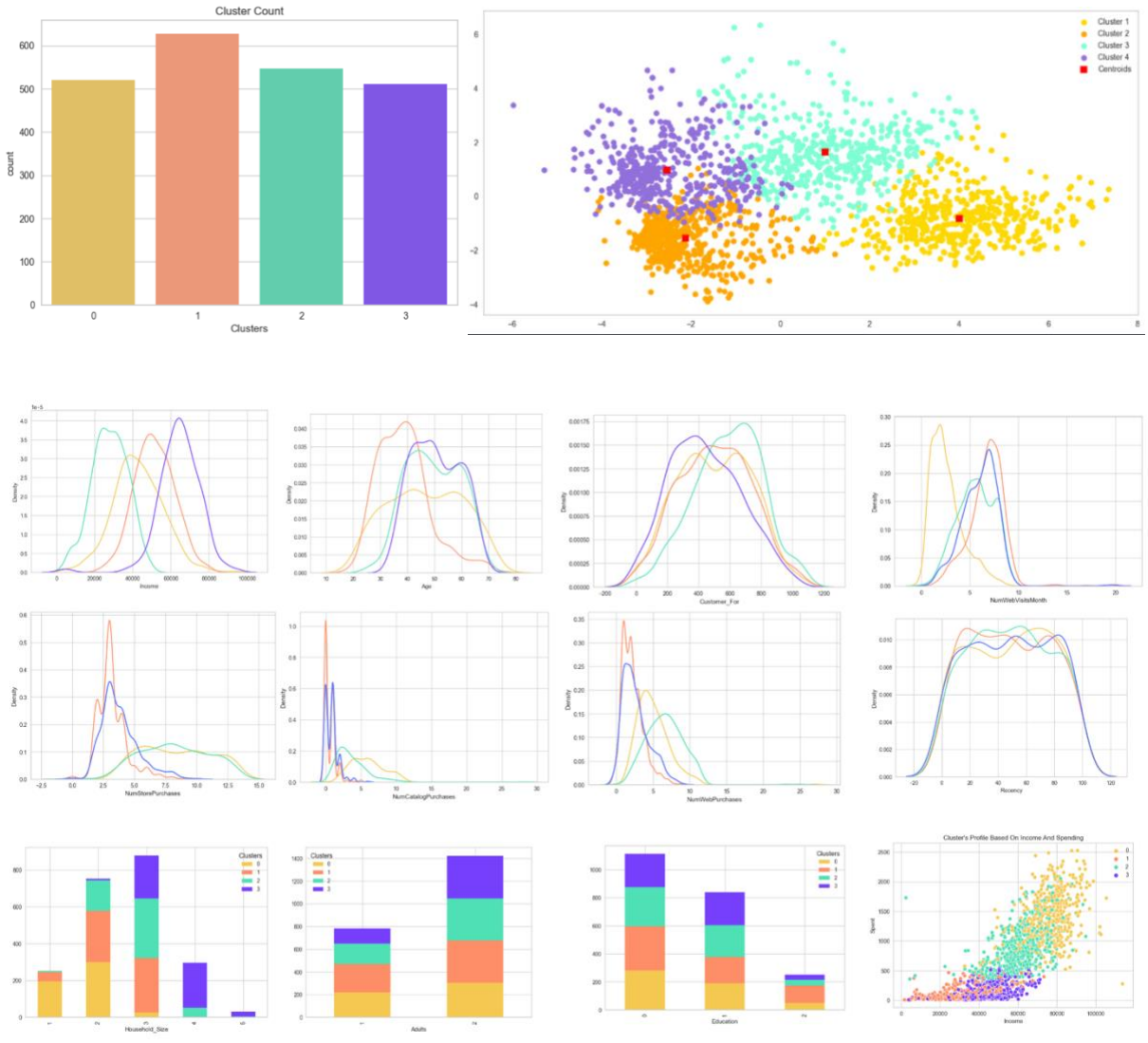
Modeling

Model Selection

When it came to model selection, I knew I needed a model familiar with unsupervised datasets and was able to group the dataset into smaller sections. Thus, I choose K-Means, K-Modes and a hierarchal model (Agglomerative).

K-Means

K-Means required that I use 4 clusters. Overall, it performed efficiently, and I felt had the best distribution. I also felt like the groups were better organized and easier to read and utilize. Below are the graphs I used.

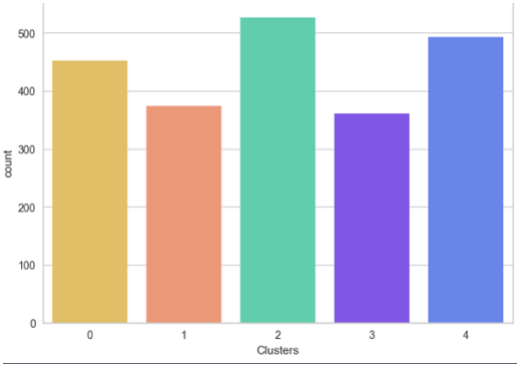


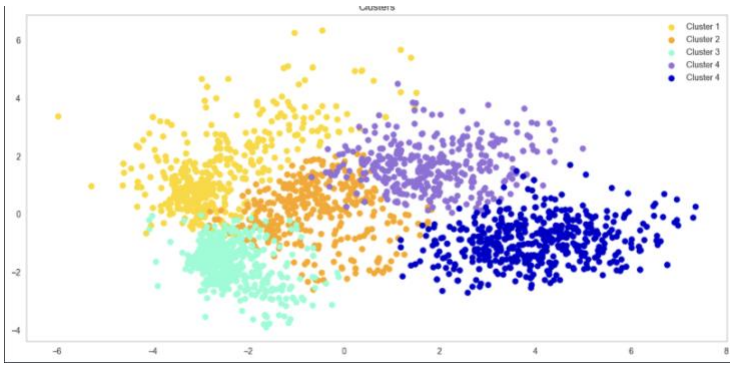
The results were as follows:

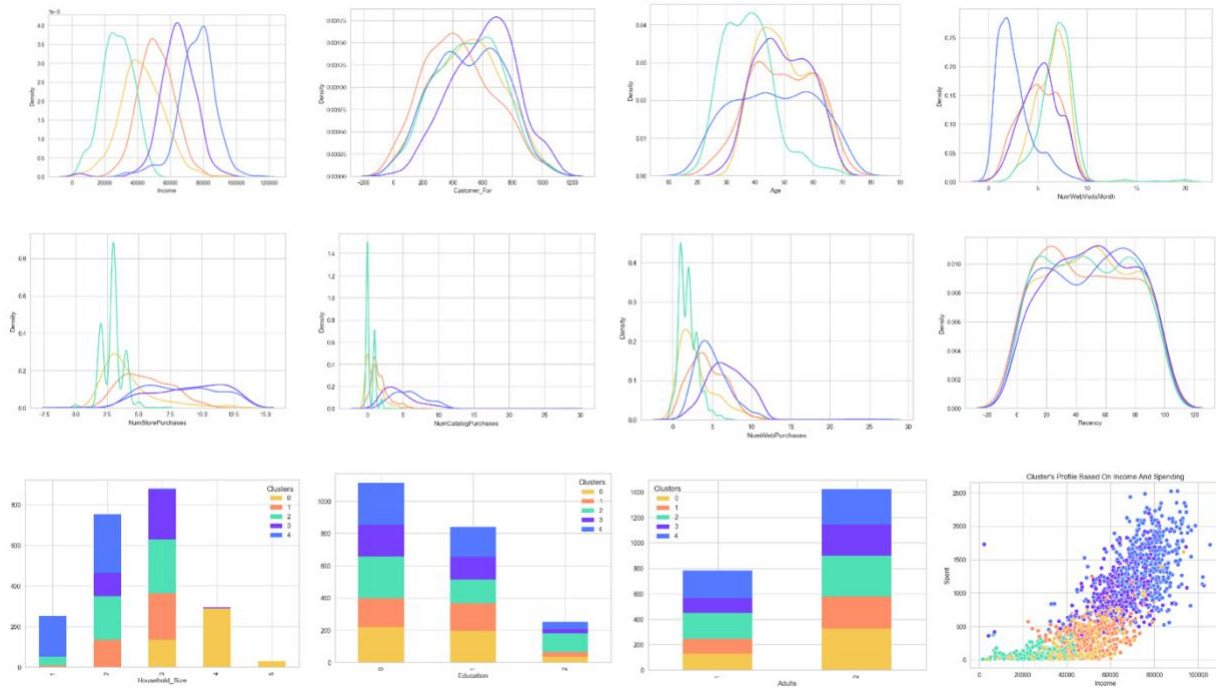
- Group 1
    - 20-65k income
    - Highest spenders
    - mostly doesn’t have kids.
    - Buys the most luxury items.
    - Is the majority of shoppers?
    - Coupon motivated.
- Group 2
    - 20-80k income
    - 20-50 years old
    - Spend the least.
    - is a small, new family.
    - Buys the least luxury items.
- Group 3
    - Lowest income (10-50k)
    - Is likely a single child matured family.
    - Buys luxury.
- Group 4
    - Largest income (0-100k)
    - is a large family.
    - Doesn’t buy much luxury.
    - Spends less on grocery.
    - Coupon motivated & uses the most coupons per vis

Agglomerative

The agglomerative model insisted on 5 clusters. My initial thought was that it would be more detailed but instead they were just unnecessary in my opinion. The distribution wasn't even, and the groups did not seem organized to the best for this dataset.





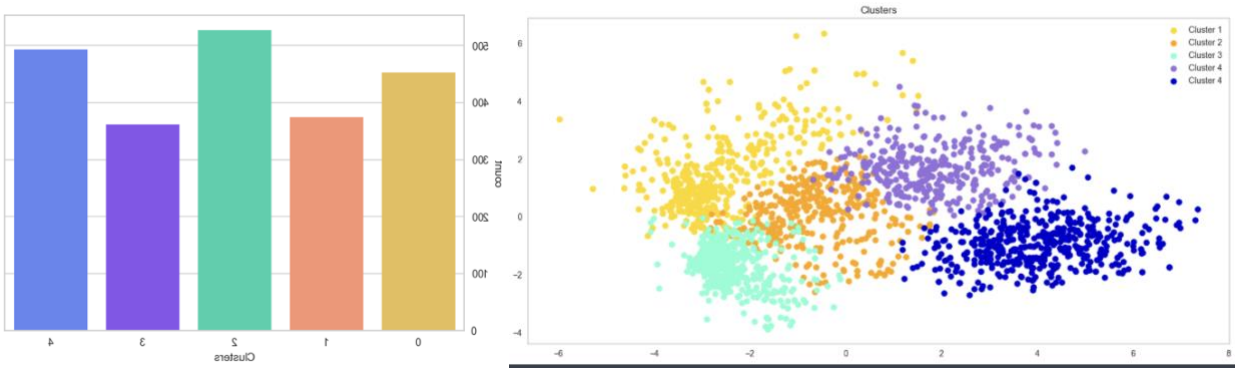


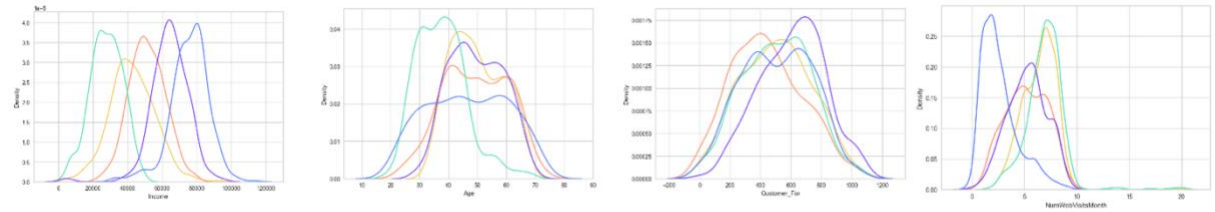
The results are as follows:

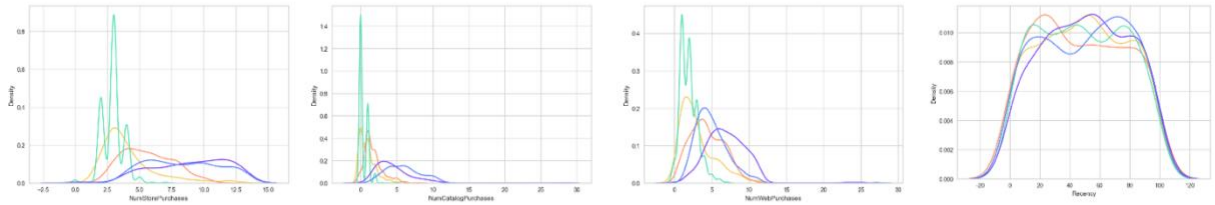
- Group 1
  - 0-80k income
  - Spend 2nd most.
  - Large Family
  - Purchase some luxury.
- Group 2
  - Middlemost group
  - 20-90k income
  - Older family
  - Buys more luxury.
  - Possibly buys better cuts of meat.
  - Coupon motivated.
- Group 3
  - 0-50k income (lowest)
  - mostly young families
  - Spends equally on grocery and luxury but doesn't purchase much.
  - Possibly buys cheapest meats.
- Group 4
  - 30-100k income
  - Older family
  - seems to mainly purchase luxury items.
- Group 5
  - 35-120k income (highest)
  - Adults without kids
  - seems to do all their shopping at the grocery store.
  - Most excess income
  - Prefers to spend on groceries than general luxury.

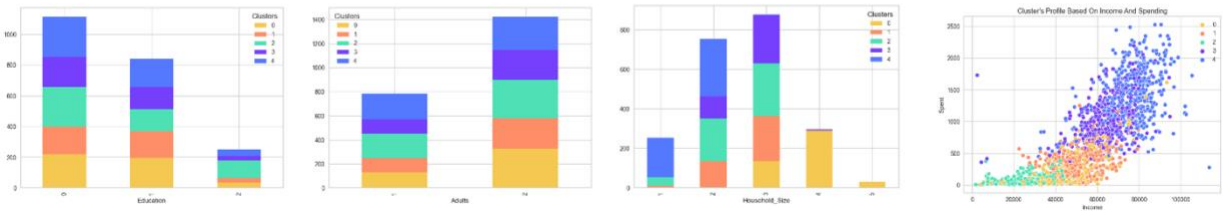
K-Modes

For some reason, My K-Modes results were identical to my agglomerative results.









The results are as follows:

- Group 1
  - 0-80k income
  - Spend 2nd most.
  - Large Family
  - Purchase some luxury.
- Group 2
  - Middlemost group
  - 20-90k income
  - Older family
  - Buys more luxury.
  - Possibly buys better cuts of meat.
  - Coupon motivated.
- Group 3
  - 0-50k income (lowest)
  - mostly young families
  - Spends equally on grocery and luxury but doesn't purchase much.
  - Possibly buys cheapest meats.
- Group 4
  - 30-100k income
  - Older family
  - seems to mainly purchase luxury items.
- Group 5
  - 35-120k income (highest)
  - Adults without kids
  - seems to do all their shopping at the grocery store.
  - Most excess income
  - Prefers to spend on groceries than general luxury.

Conclusion

I feel that my dataset was limited and with a larger dataset I would be able to get a better understanding of the customer. Overall, I appreciated the K-Means model the most. It gave me the least trouble and gave the best insight into the shopping needs of each of the clusters. My least favorite model was K-Modes, as it was temperamental and was not instant, unlike the other sets.

Citations

Patel, Akash. “Customer Personality Analysis.” *Kaggle*, 22 Aug. 2021, https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.