Machine Learning Engineer Nanodegree Program

Report

Starbucks Customer Behavior Prediction with Machine Learning Agglomerative Clustering

Carlos André Antunes

2021

Table of Contents

Introduction

FIrst of all, I’m a big fan of Starbucks since I was visiting some places around the word, walking like 20km a day and needed some place that give me some comfort. For a developer nothing more comforting than a cup of coffee, electrical energy and wifi.

To attract and retain customers, Starbucks use their rewards program that honors regular customers with special offers not available to the standard customer. In this project, we'll be combing through some fabricated customer and offer data provided by Udacity to understand how Starbucks may choose what rewards program to better fit specific customer segments, trying to extract for this data some behavioral insights.

Problem Statement

The problem we are looking to solve here is best determine which kind of offer to send to each customer segment based on their purchasing decisions.

Datasets and Inputs

For this project, we will use the data graciously provided by Udacity in JSON format.We need to understand the three types of offers that Starbucks is looking to potentially send its customers:

- Buy-One-Get-One (BOGO): In this particular offer, a customer is given a reward that enables them to receive an extra, equal product at no cost. The customer must spend a certain threshold in order to make this reward available.
- Discount: With this offer, a customer is given a reward that knocks a certain percentage off the original cost of the product they are choosing to purchase, subject to limitations.
- Informational: With this final offer, there isn’t necessarily a reward but rather an opportunity for a customer to purchase a certain object given a requisite amount of money. (This might be something like letting customers know that Pumpkin Spice Latte is coming available again toward the beginning of autumn.)

Was provided 3 JSON files:

1. profile.json

This file contains dummy information about Rewards program users. This will serve as the basis for basic customer information.
(17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string / hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

2. portfolio.json
This file contains offers sent during a 30-day test period. This will serve as the basis to understand our customers' purchasing patterns.
(10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) bogo, discount, informational
- id: (string / hash)

3. transcript.json
This file contains event log information. Complementing the file above, this file will serve as a more granular look into customer behavior.
(306648 events x 4 fields)

- person: (string / hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on the event type
  - offer id : (string / hash) not associated with any "transaction"
  - amount: (numeric) money spent in "transaction"
  - reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

Solution Statement

Analyzing the datasets I realized that I don't have any data that would enable me to use supervised learning models, so I will be using unsupervised learning methods to determine potential strategies for adjusting the Starbucks Rewards program given customer insights. Specifically hierarchical modeling, to cluster our data into a few respective customer segments for analysis.

Evaluation Metrics

Would be use silhouette coefficient. Because I don't have labelled data, the silhouette coefficient is appropriate since it produces a score between the range of -1 and 1 based on internal indices. It also happens to be easy to calculate with help from sci-kit learn.

Additionally, I will use the elbow method of determining k-means clusters through a simple function that will iterate through a number of K-Means clusters and displaying the Sum of Squared Errors (SSE) in visual form. This in conjunction with the silhouette coefficient will idealize the number of clusters for my final algorithm.

Algorithms and Methodologies

In this section I'll talk about the algorithms and methodologies used throuth the execution of project.

K-Means Clustering

The first algorithm used is K-Means. K-Means works by initializing a number of centroids that serve as sort of magnets to attracting nearby clusters of data around them. These centroids are determined by user. For example, in scikit-learn's "KMeans" algorithm, I initialize how many centroids there will be by passing in an "n_clusters" parameter.

Each time the algorithm steps through an iteration, each respective centroid moves closer and closer toward the "center" of the clustered data. I've visualized this with some very basic data in the three screenshots below. Notice that the initialized centroids do a pretty poor job at clustering the data in the first step, then do a slightly better job in the second step, and then finally do a great job in the third step.

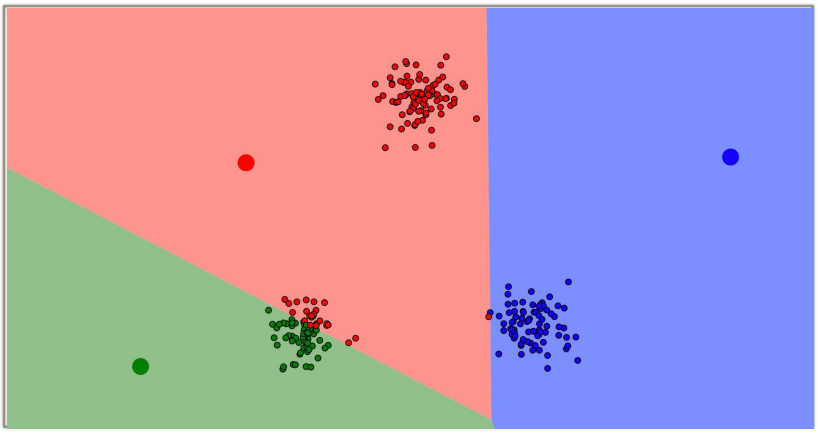


Figure 1 - K Means First Step

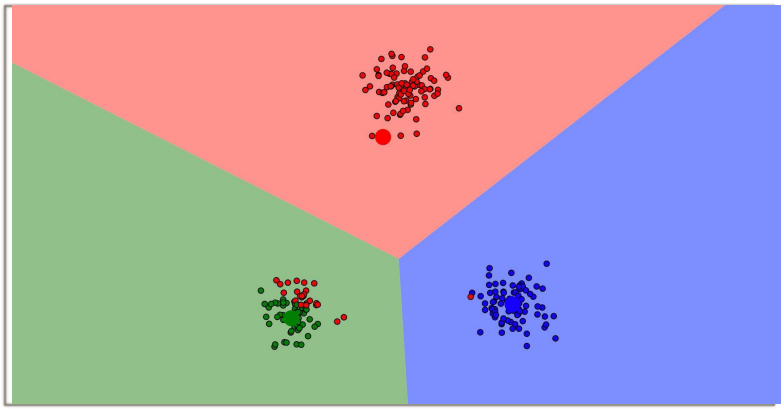


Figure 2 - K Means Second Step

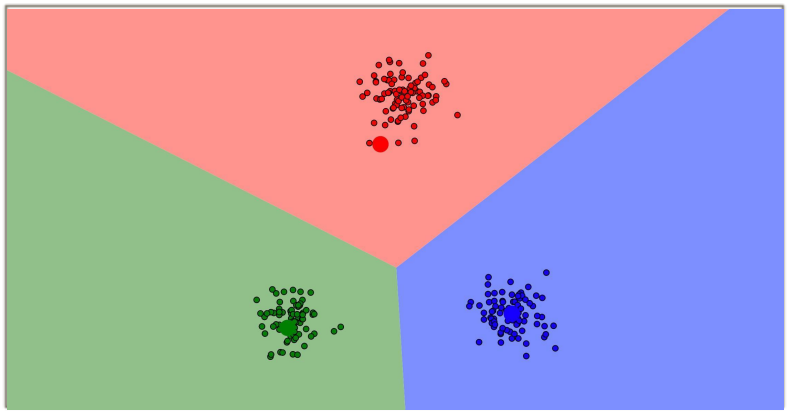


Figure 3 - K Means Third Step

Hierarchical Algorithms

Hierarchical clustering is used to aggregate data by their proximity. The algotithm begins with the underlying assumption that every single data point is already its own cluster. From there, the hierarchical clustering use different forms of linkage to determine how to best cluster the data, as shown below.

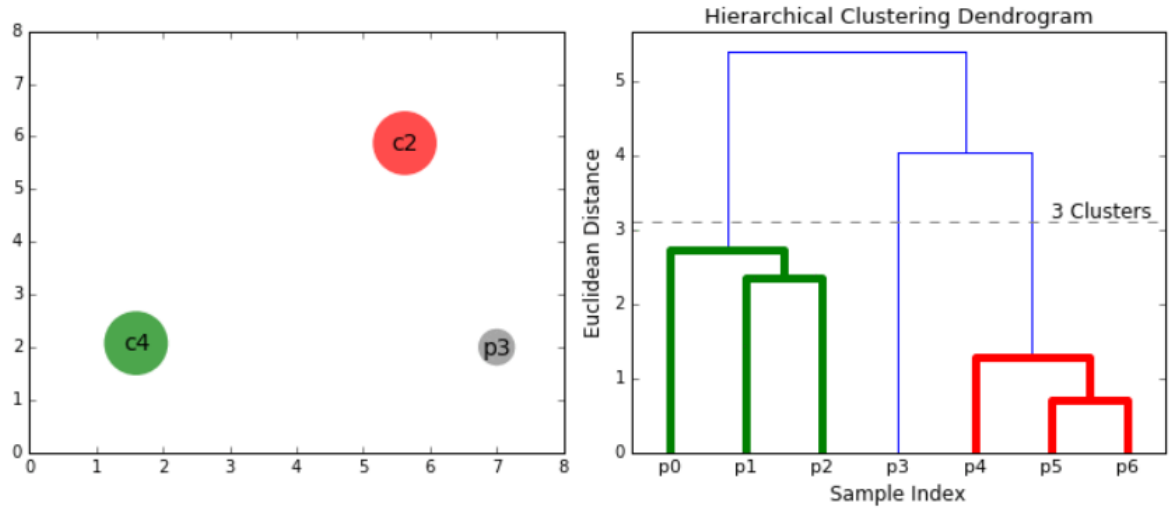


Figure 4 - Hierarchical Clustering.
Source: https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html

This kind of algorithm starts assuming that each data is one cluster, and by proximity this will agglomerating points until aglomerate all points in a unique cluster. Of course, one million of small clusters is irrelevant, one big cluster too. Instead, I want just the right amount of clusters, and that is defined by the hyperparameters of the model.

Initial Cleansing

The initial data need to be cleaned in order to best fit for unsupervised model used later on in the project. Specifically, I cleaned up the initial datasets and then later combined them to form a master dataset that will be regularly used the project.

Portfolio Clean Up

- Changing the column name from ‘id’ to the more descriptive ‘offer_id’
- One hot encoding the ‘offer_type’ column.
- Separating and one hot encoding the ‘channels’ column
- Dropping the ‘offer_type’ and ‘channels’ columns now that they are one hot encoded in other columns

Profile Clean Up

- Dropping rows with null information
- Changing ‘id’ column to ‘customer_id’ name

- Changing the ‘became_member_on’ column to a date object type
- Calculating number of days that a person has been a member as a new ‘days_as_member’ column (as of August 1, 2018)
- Creating new ‘age_range’ column based off ‘age’ column

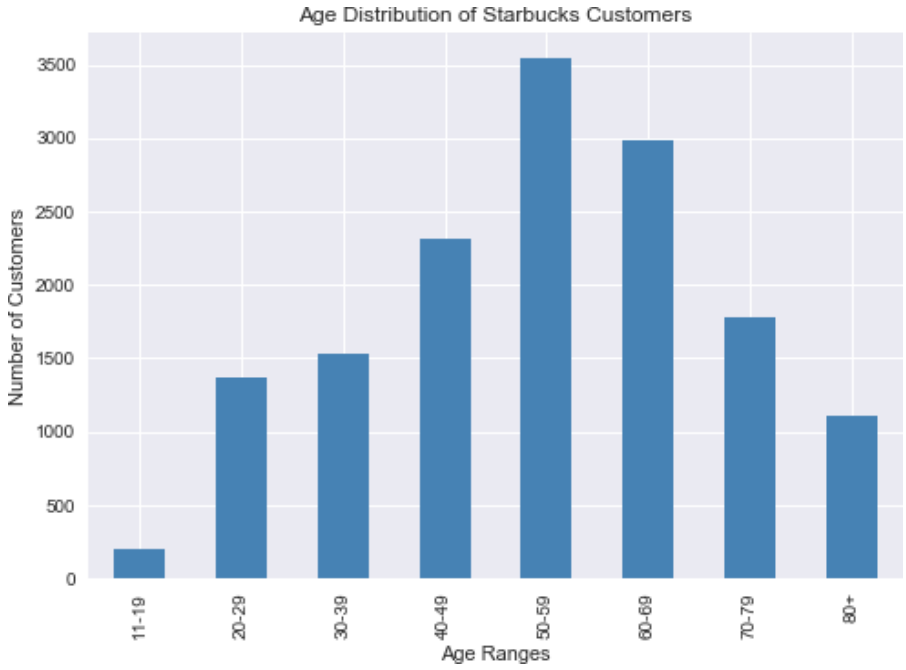
Transcript Clean Up

- Changing the name of the ‘person’ column to ‘customer_id’
- Removing the customers that are not reflected in the ‘profile’ dataset
- One hot encoding the ‘event’ values
- Changing the ‘time’ column to ‘days’ along with appropriate values
- Separating value from key in ‘value’ dictionary in order to form two wholly separate datasets: transcript_offer and transcript_amount

Exploratory Data Analysis

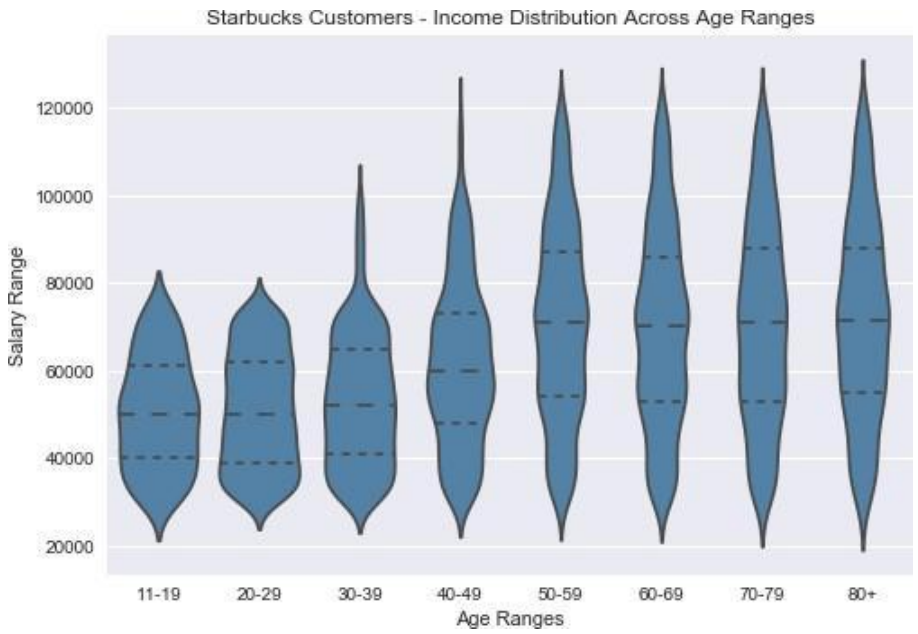
Once the cleansing complete, I can start an exploratory data analysis on the datasets. I'm going to make a serie of questions and reflective summaries.

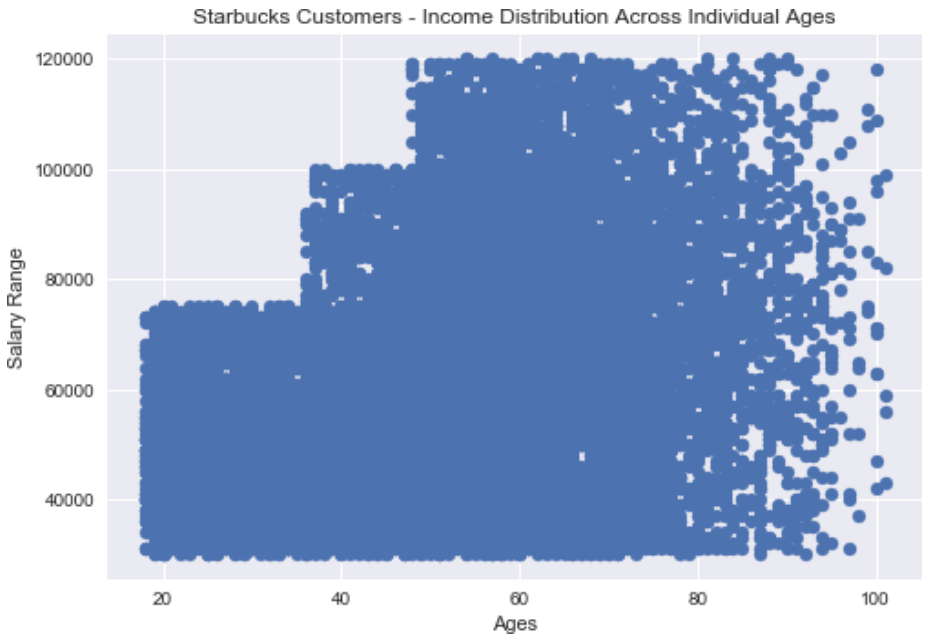
Question 1: What are the general age ranges of our customers?
When I think about Starbucks, I imagine young people using their devices and drinking coffe. So I'm expecting that we'll see more customers in those younger age ranges like 20's or 30's..



Answer 1: The actual age distributions of Starbucks customers
I was wrong with my initial assessment. This is example of why it's important to not make assumptions. The 50-59 and 60-69 has more customers, maybe cause people in this age have a higher wage than the young people. Ok, move on.

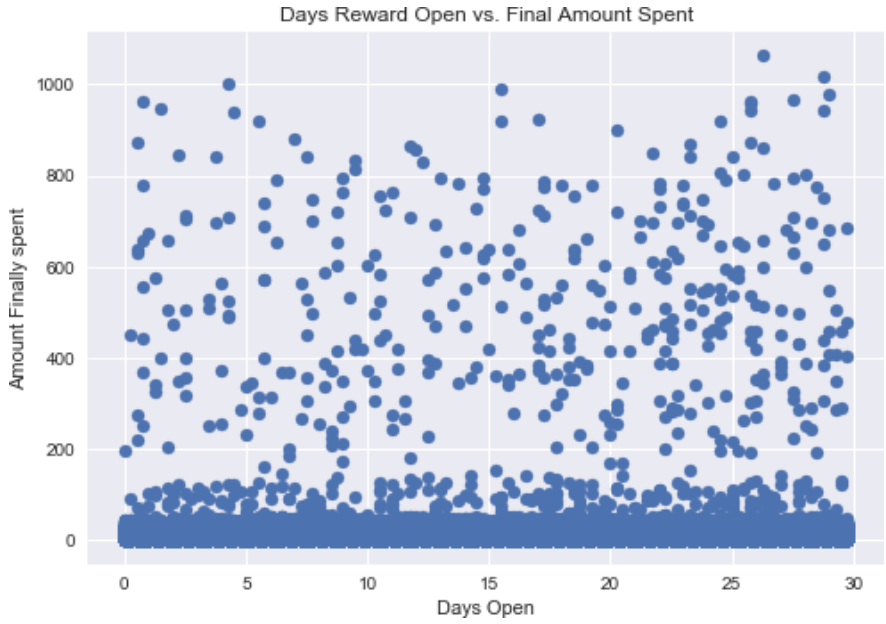
Question 2: What are the salary ranges of people across different age groups?
Now I'm curious to see how the salary ranges of these various age groups might affect how often a person visits Starbucks and utilizes their rewards program. I'm going to guess that those 40-60 age ranges have the highest salary ranges and make more money to show for it.
I'll use **violin plot** and **scatter plot** to investigate it.

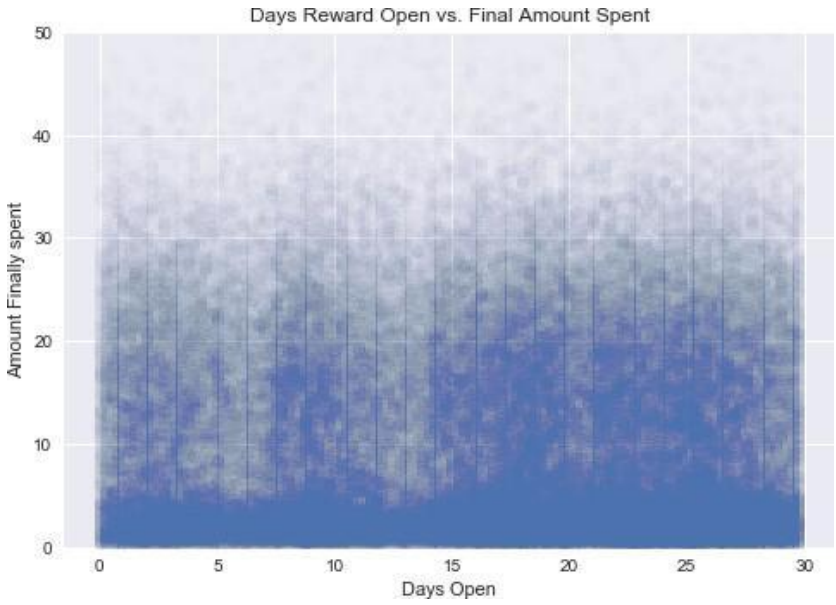




Answer 2: Analysis of Income Distributions
The violin plots shows that older customers tend to make more money. However, seems that to be a hard cap on the salary range of younger people. Scatter plot using individual ages shows that there are caps on everybody's salaries.

Question 3: What we see about correlation between number of days an offer has been open vs. final transaction amount?
Here I expect to see if a offer deadline cause a final transaction. Something like "Well, I gotta use it or lose it". But maybe we couldn't see correlations here.



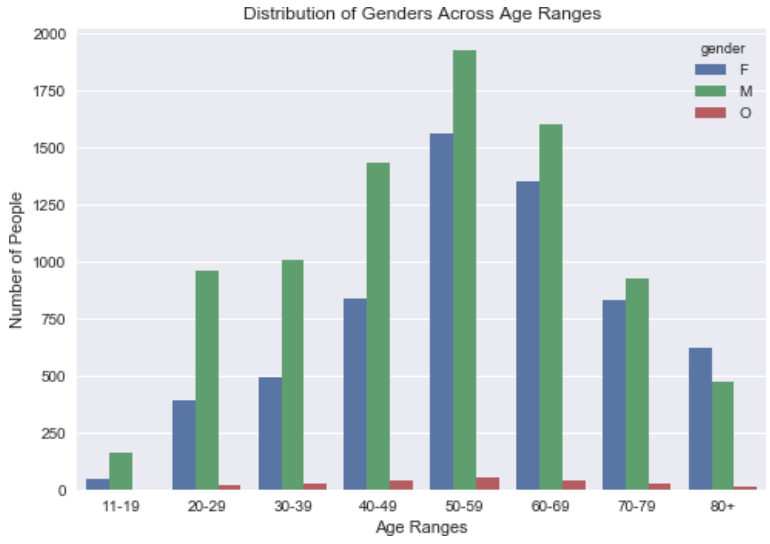


Answer 3: Analysis of Amount Spent vs. Days Open

We can see that when the offer are close to 30 days deadline, more people tend to spend more money in final transaction, but it's a little irrelevant so there is no correlation at all between days open and amount spent.

Question 4: Do gender distributions have any major effect on our data here?

Finally, how gender may affect the final analysis. The label are F: Female, M: Male, O: Other.



Answer 4: Analysis of gender across customer data

First, we see more males across this dataset than any other gender category. The only age range we see more women is in the 80+ category. My guess is the fact that women generally tend to live longer than men.

Something really interesting here is the salary distribution for women in particular. Where the dataset indicates that there are more male customers than female (or other) customers, females in this dataset generally tend to have a higher income than men. And across both the primary genders, it's not as if there's a crazy disparity between the average salary, too.

Data Preprocessing

We're now going to create a function that builds a 'customer_transactions' dataframe. I'll to engineer some new features that I feel will be helpful when we actually move toward building ML models. Each row will represent an individual customer and contain the following columns / features.

Features

- customer_id: The unique customer identifier
- age: The age of the customer
- age_range: The age range the customer falls into
- gender: The gender of the customer, either male (M), female (F), or other (O)
- income: How much money the customer makes each year
- became_member_on: The date that the customer because a Starbucks Rewards member
- days_as_member: How many days that the customer has been a Starbucks Rewards member
- total_completed: The total number of offers actually completed by the customer
- total_received: The total number of offers that Starbucks sent to the customer
- total_viewed: The total number of offers that the customer viewed
- percent_completed: The ratio of offers that the customer completed as compared to how many offers Starbucks sent to the customer
- total_spent: The total amount of money spent by the customer across all transactions
- avg_spent: The mean average amount of money spent by the customer across all transactions
- num_transactions: The total amount of individual monetary transactions performed by the customer
- completed_bogo: The number of completed BOGO offers by the customer
- num_bogos: The total number of BOGO offers sent to the customer by Starbucks
- bogo_percent_completed: The ratio of how many BOGO offers were actually completed by the customer as compared to how many Starbucks sent them
- completed_discount: The number of completed discount offers by the customer
- num_discounts: The number of discount offers sent to the customer by Starbucks
- discount_percent_completed: The ratio of how many discount offers were actually

completed by the customer as compared to how many Starbucks sent them.

Machine Learning Modeling

As noted in the proposal for this project, I'll leverage some unsupervised algorithms to cluster data to find commonalities across customer segments based on a number of features.

Feature Selection

At now we need to remove some columns that can't scale or column ids:

- customer_id: proprietary to the row, it is a wholly unique value
- became_member_on: a date column that can't be scaled
- age_range: a categorical column that can't be scaled

Hierarchical Clustering - Feature Tunning

Let's create a single-link based hierarchical model utilizing our unscaled dataset with a choice of six clusters. This links are build by the three methods below.

Feature Scaling

In the first step of hierarchical model, the model will run on the data without any scaling of the data. This was an intentionally huge couse the numbers across different columns are not created equally. For example, ‘income’ is measured in the thousands of dollars whereas something like ‘average amount spent’ typically hovers pretty low, somewhere around like $10 to $20. Scaling the data sets all the columns on an equal playing field to ensure that no single feature is weighing down the rest of the dataset. To do this I’ll used StandardScaler of Scikit Learn Toolkit.

Linkage

Linkage is the methodology by which data is clustered, literally determining the best link between defined clusters. In the first step, I utilized single-link clustering. Here’s a quick explanation on that clustering along with the other various kinds of clustering.

Single link clustering looks to cluster different sets of data by linking the clusters together at the point at which the two closest points meet. This is often a poor choice of linkage because the clusters could have “longer tails” that don’t at all reflect a proper linkage. This is best illustrated in the image below.

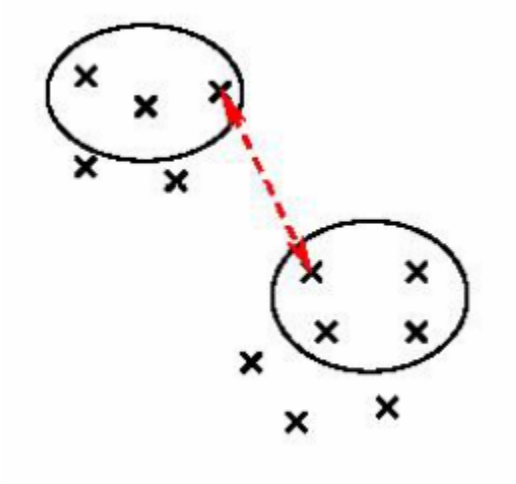


Figure 5 - Single Link
Source: https://medium.datadriveninvestor.com/hierarchical-clustering-514b9d1aa2c1

Complete Link: Where single link clustering took the approach of clustering based on the points in a cluster being closest to one another, complete link takes the opposite approach in linking clusters together by the points that are furthest away in a group. Again, this is also prone to the same issue with single link and “long tail” clusters, so this also isn’t an ideal choice for the final model.

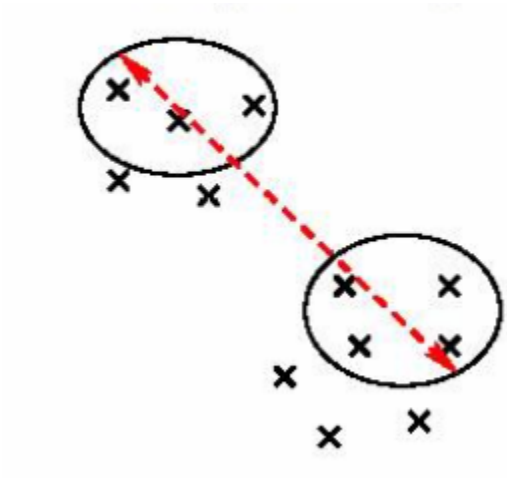


Figure 6 - Complete Link
Source: https://medium.datadriveninvestor.com/hierarchical-clustering-514b9d1aa2c1

Average Link: Whereas the two clustering methods above took a more narrow approach at defining linkage via a single point, average linking takes into account instead the average of the whole cluster and links based on that instead. We’ll see how this differs from Ward’s linkage down below.

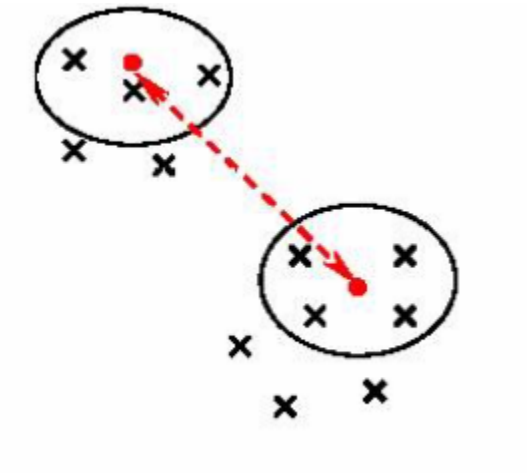


Figure 6 - Avarage Link
Source: https://medium.datadriveninvestor.com/hierarchical-clustering-514b9d1aa2c

Ward's Link: Ward's is similar to average in the fact that it looks at all data points in a cluster before making a decision, except it looks to minimize variance by averaging to a single point between the clusters instead of an average between the clusters.. The screenshot below definitely helps to illustrate the idea nicely.

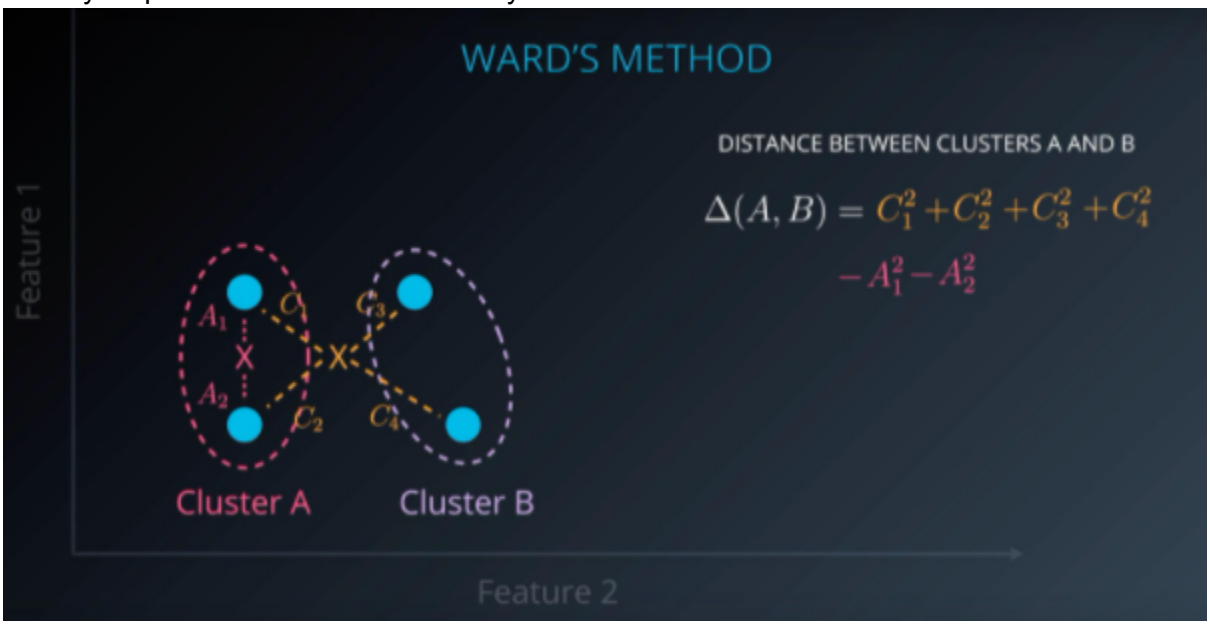


Figure 7 -Ward´s Method
Source: Udacity Data Science Nanodegree

Number of Clusters

Let's determine the number of ideal clusters by viewing KMeans SSE scores as K increases, known as the elbow method, as well as leveraging the silhouette score.

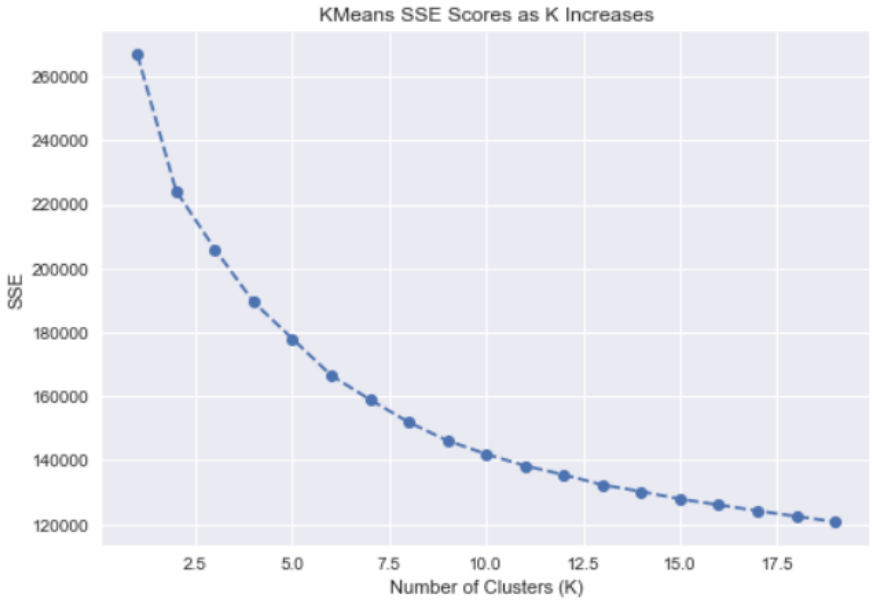


Figure 8 - Elbow Method
Source: Author.

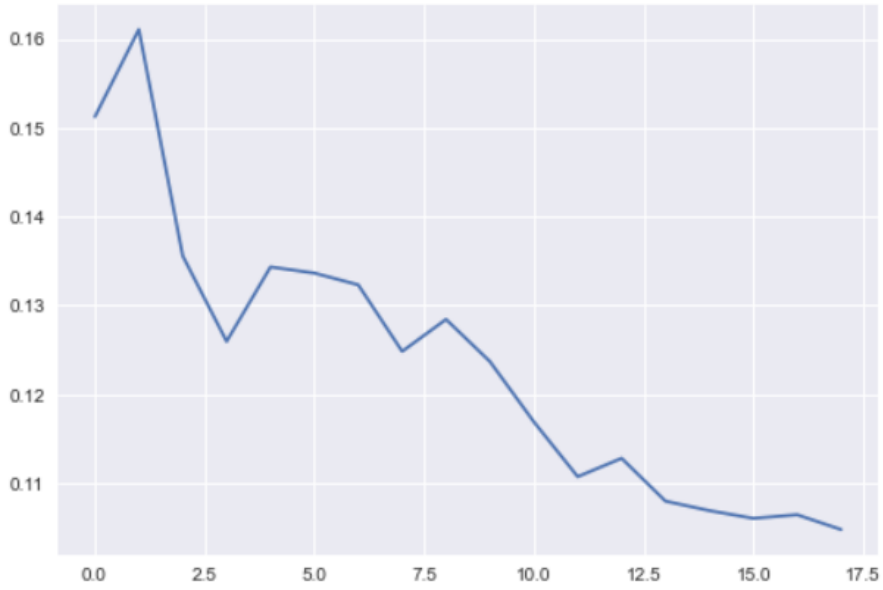


Figure 9 - Sillouete Score
Source: Author.

Given both plots here, we'll use 4 clusters. Both the sihouette score and elbow method started showing diminishing returns following 4 clusters.

Machine Learning Modeling - Model Tunning

Now we've determined how to properly refine the model to optimize the results and extract some insights.

Benchmark Definition and Comparison

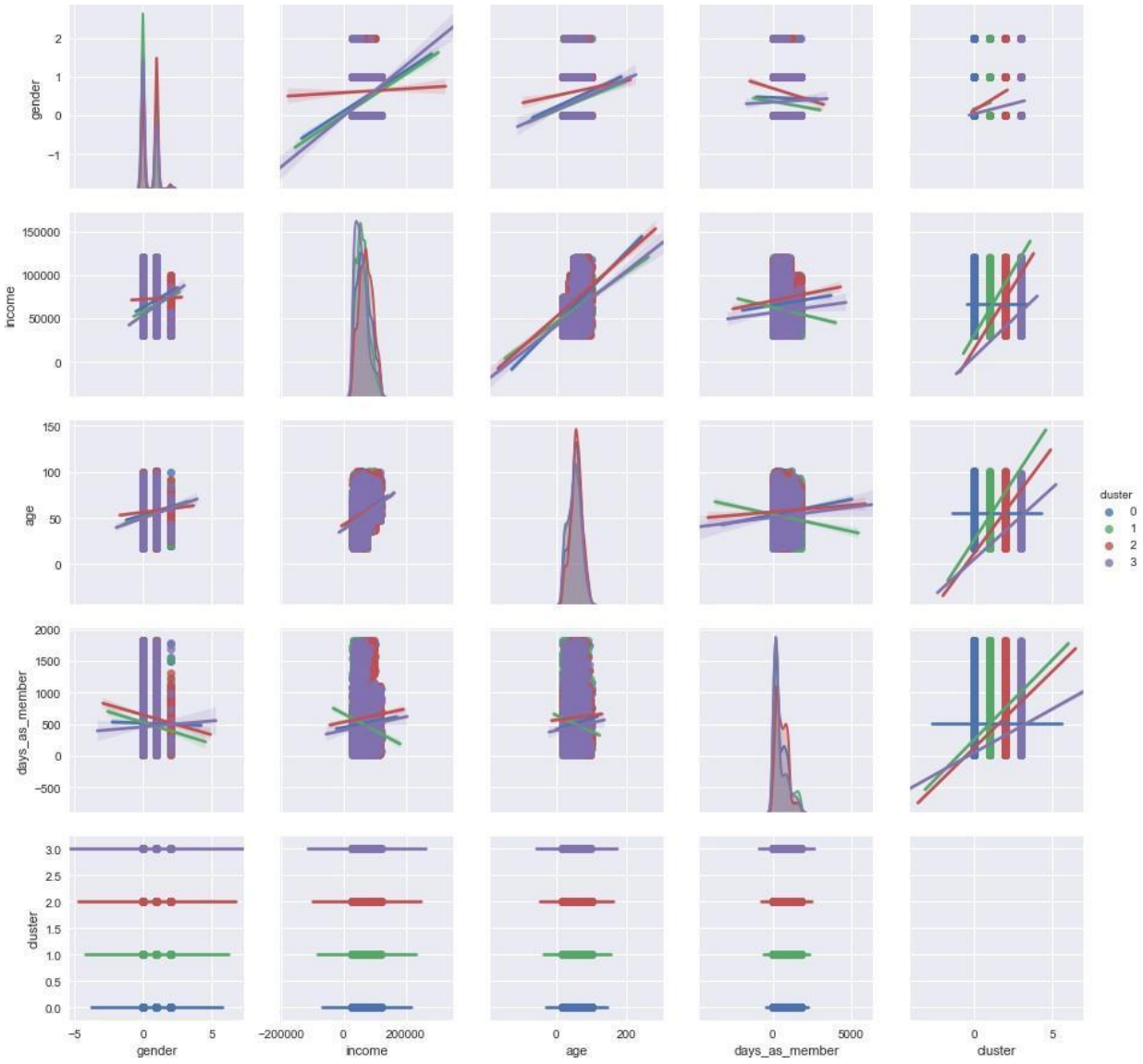
Given that there isn't necessarily a labelled right or wrong to the provided dataset, I can't really objectively evaluate how well the unsupervised dataset perform. What I can do, however, is leverage one benchmark and metrics to determine the ideal number of clusters for the final algorithm. I already explored the elbow method and silhouette score in the previous section.

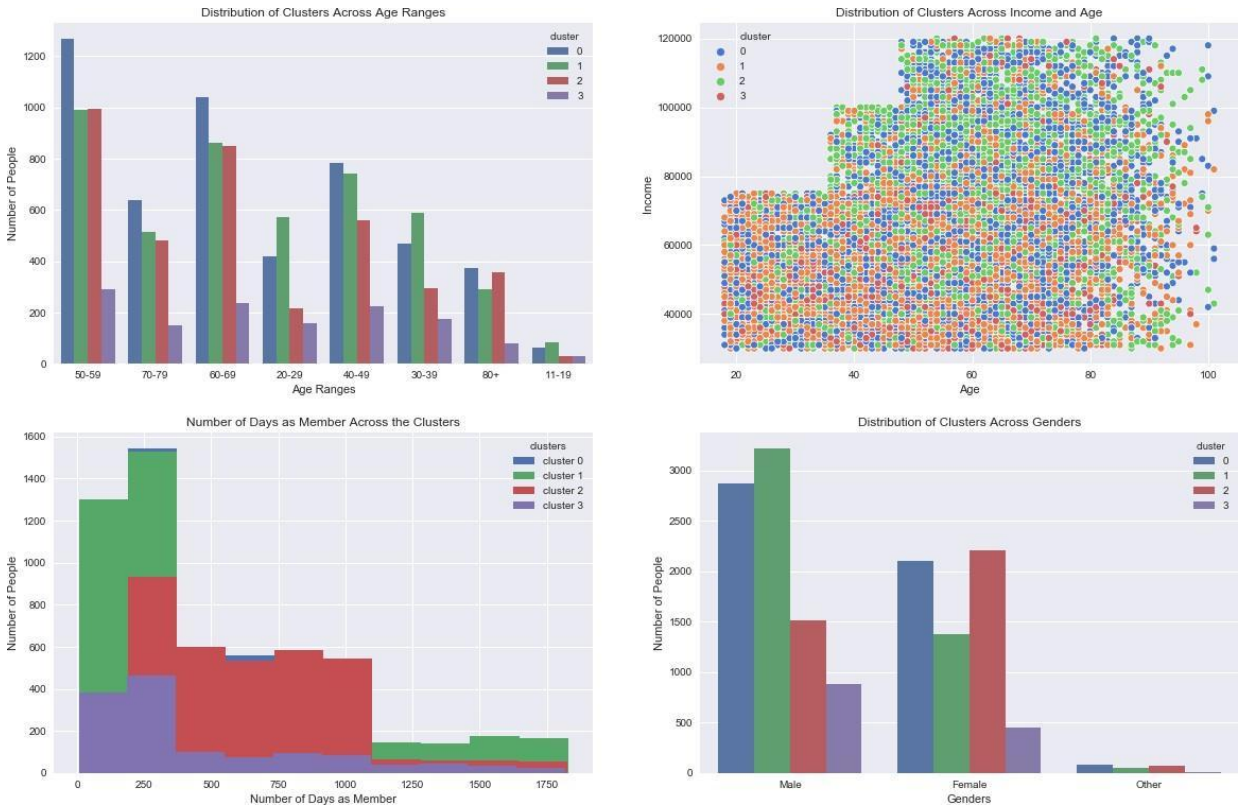
Final Analysis and Evaluation

Now that I've gathered various clusters from hierarchical algorithm, let's visualize the results. I'm going to explore two more high level questions and how these might be utilized by Starbucks to adjust their rewards program.

Question 5: What personal attributes of our customers are defined throughout each of our clusters?

First, let's take a look at the personal attributes of customers as clustered by the algorithm.





Answer 5: Analysis of Clustered Personal Attributes

Now we let's cover each of the respective clusters within the respective sections below.

Cluster 0

This is largest cluster, cluster 0 tended to consist of older people with higher incomes. As evidenced by our countplot with the age ranges, the ages typcally fell into that 50 to 80 year old range. Additionally, it was somewhat common to see these folks have some of the higher income ranges. Gender was somewhat evenly split between males and females.

Because this cluster was the biggest, it had a lot of discrepancies (particularly with income) that make it questionable how much to rely upon it for future inference.

Cluster 1

This cluster seems to be the young person's cluster.

Looking at the age range distribution, we see the strongest distribution here amongst the 20-40 year old crowd. The gap between males and females is quite large in this cluster.

These people also tend to fall toward the lower end when it comes to income.

And as far as number of days as member goes, this cluster's distribution is very similar to that of cluster 0.

Cluster 2

This is the only customer where there were more males than females. This cluster contain the largest distribution of people who have been members of the Starbucks rewards program for some time.

Income tended to be higher here which is not surprising given that EDA showed us that women tended to make more than men.

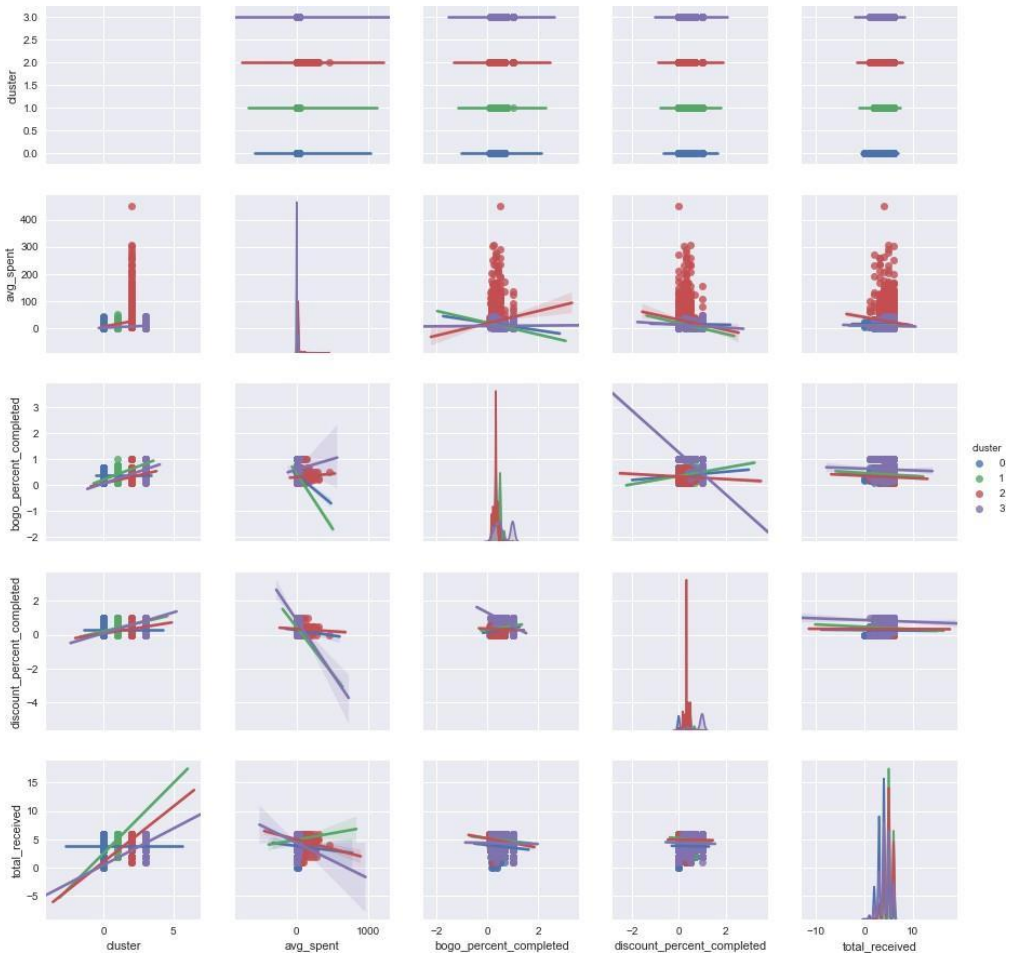
Cluster 3

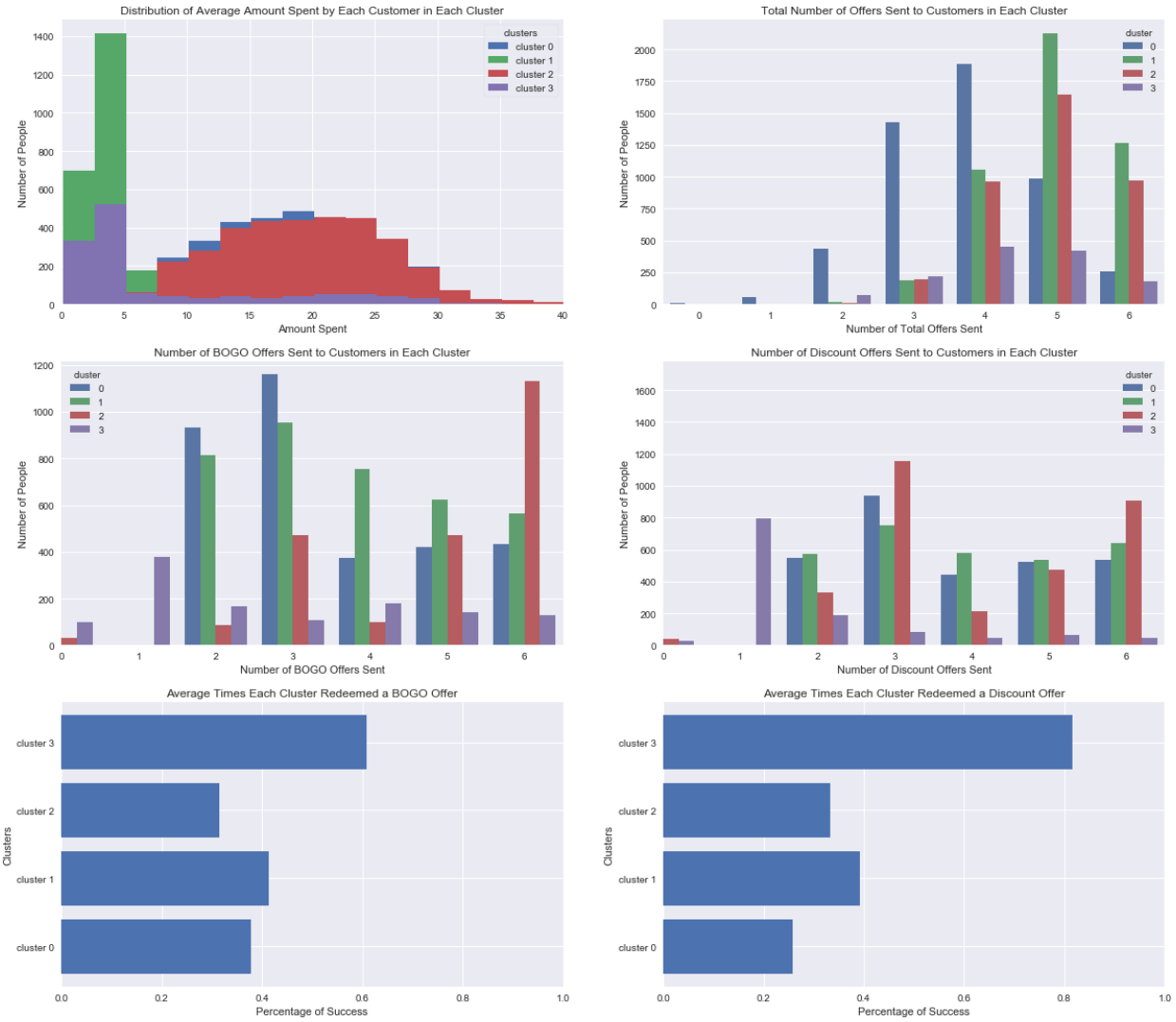
The smallest of four clusters, yet it bears some striking similarities to cluster 0 in a few ways.

Namely, the distribution of days as members and gender are fairly similar. The exception that separates it from cluster 0 is that there tended to be more people in the younger age range.

Question 6: What do the behavioral attributes look like across our clusters?

Finally, let's look at the clusters across some of the more behavioral attributes.





Answer 6: Analysis of Clustered Behavioral Attributes

Let's cover each of the respective clusters within the respective sections below.

Cluster 0

This cluster on average seems to be the biggest spenders. This cluster contains older customers, that maybe are buying for multiple people, like family members. This cluster has the lowest yield of using discount offers and pretty close to last for BOGO offers.

Cluster 1

This cluster tends toward a younger crowd, is possible see things like the average amount spent spike toward the lower end. This particular group gets hit hard with offers, spiking big time around 5 offers sent. It looks like there's a stronger leaning toward BOGO offers over discount offers.

Cluster 2

This is predominantly female cluster, this group also has a high spend amount, much akin to cluster 0. This cluster is hit pretty strongly with a lot of offers, especially BOGO offers.

Cluster 3

This cluster group tends a little younger than the cluster 0's. This group has a high success rate with offers, especially with discount based offers. This cluster is the smallest of the bunch.

Conclusion

For as much work that I've put into this project, it could be concluded by the other kind of approaches. I choose build this project this way because to me it is more interesting to try to understand what is the meaning of unsupervised learning instead use some more basic approaches with algorithms that make all the work. I'm glad we took the approach that we did and am looking forward to applying these skills in other projects. I love this course. Thanks Udacity.