
# Starbucks Project Final Report

David Hundley | July 2019
Udacity - ML Engineer Nanodegree

## Introduction

It's no secret that I'm a big fan of Starbucks. I visit Starbucks perhaps two to three times per week. A daddy to two little girls under two years old, I often leverage one of our local Starbucks locations as a getaway means to study or focus on projects… like this one!

Let's introduce what this project is looking to do. As a means to attract and retain customers, Starbucks leverages a rewards program that honors regular customers with special offers not available to the standard customer. For this project, we'll be combing through some fabricated customer and offer data provided by Starbucks / Udacity to understand how Starbucks may choose to alter its rewards program to better suit specific customer segments.

## Domain Background

A desire to glean insights about customer behavior is one of the hottest uses of machine learning today, and rightfully so. It is often difficult to understand necessarily how certain behaviors influence others, so leveraging things like unsupervised predictive learning models go a long way to better understand customer behavior.

Throughout my professional work experiences, I have had experience working directly with customers in things like focus groups, if not utilizing predictive modeling. (At least, not yet.) Fortunately, I am also a student of Udacity's Data Science nanodegree program and recently completed the Arvato project on determining customer segments. Coincidentally, that was another offering I could have selected for this particular capstone, but I wanted to focus my sights on something else entirely. Still, the work from that project will serve me well as I seek

to pivot the unsupervised learning knowledge applied there to this particular project. (Link to Hundley - Arvato Customer Segments GitHub)

## Problem Statement

The problem we are looking to solve here is conceptually easy to understand, albeit difficult to answer. **We are looking to best determine which kind of offer to send to each customer segment based on their purchasing decisions.** We'll touch more on what these offers are and data we'll be utilizing down in the next section. We will leverage traditional evaluation metrics to determine which model is most appropriate for our dataset. These evaluation metrics will be discussed in an upcoming section.

## Datasets and Inputs

For this project, we will be leveraging the data graciously provided to us by Starbucks / Udacity. This is given to us in the form of three JSON files. Before delving into those individual files, let us first 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.)

With that understanding established, let's now look at the three provided JSON files and their respective elements:

**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
Given that we do not have any labels or ground truth that would enable us to leverage supervised learning models, **we will be leveraging unsupervised learning methods amongst our data to determine potential strategies for adjusting the Starbucks Rewards program given our customer insights**. Specifically, we'll be leveraging hierarchical modeling to cluster our data into a few respective customer segments for analysis.

## Evaluation Metrics
Given that we will be leveraging unsupervised clustering models for our project, we will be using some metrics that enable us to validate our clusters without having labelled data. Namely, we will be leveraging the **silhouette coefficient**. Because we 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, we will be leveraging 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 our final algorithm.

## Initial Cleansing

The data in its initial form is decent, but we will need to clean it up some in order to best leverage it for our unsupervised model later on in the project. Specifically, I cleaned up the initial datasets and then later combined them to form a master dataset that we'll be regularly working from in the remainder of the project. We'll discuss about that initial cleansing here and talk more about that latter preprocessing down in another section.

### Portfolio Clean Up

- Changing the column name from 'id' to the more descriptive 'offer_id' since the id column is present in our other datasets
- One hot encoding the 'offer_type' column to work well with our algorithms later
- Separating and one hot encoding the 'channels' column to also work with our algorithms later
- 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

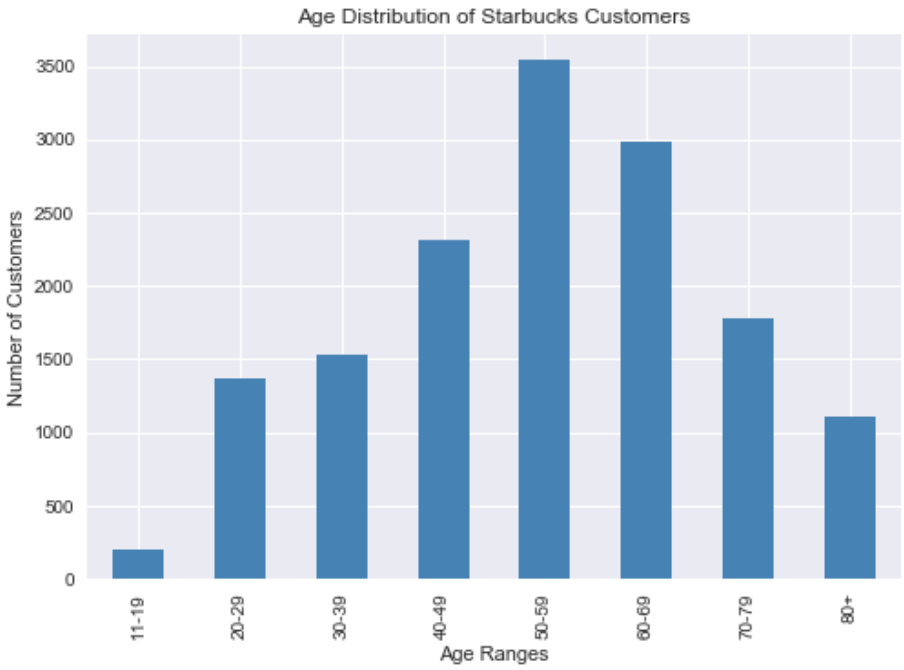
- Changing the name of the 'person' column to 'customer_id'
- Removing the customers that are not reflected in the 'profile' dataset
- One hot encoding the 'event' values
- Changing the 'time' column to 'days' along with appropriate values
- Separating value from key in 'value' dictionary in order to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

## Exploratory Data Analysis

With initial cleansing complete, we can go ahead now and perform an exploratory data analysis on our datasets. We're going to separate this question into four distinct questions (Q#'s), initial thoughts, visual analyses, and reflective summaries (A#'s).

### Q1: What are the general age ranges of our customers?

Given the general hip, young vibe associated with Starbucks, I'm expecting this to be skewed right, meaning that we'll see more customers in those younger age ranges like 20's or 30's. With that thought in mind, let's view what the dataset actually tells us.

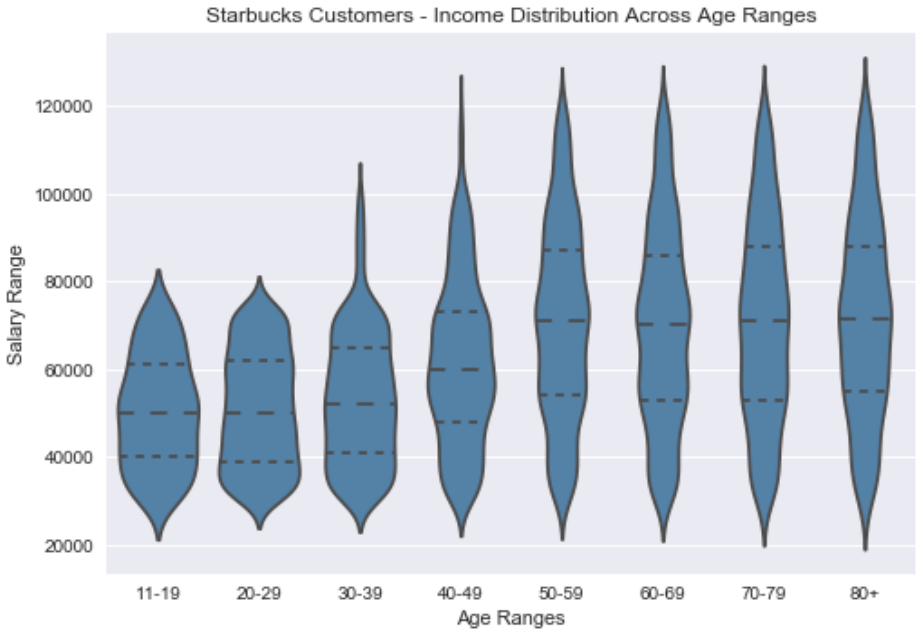


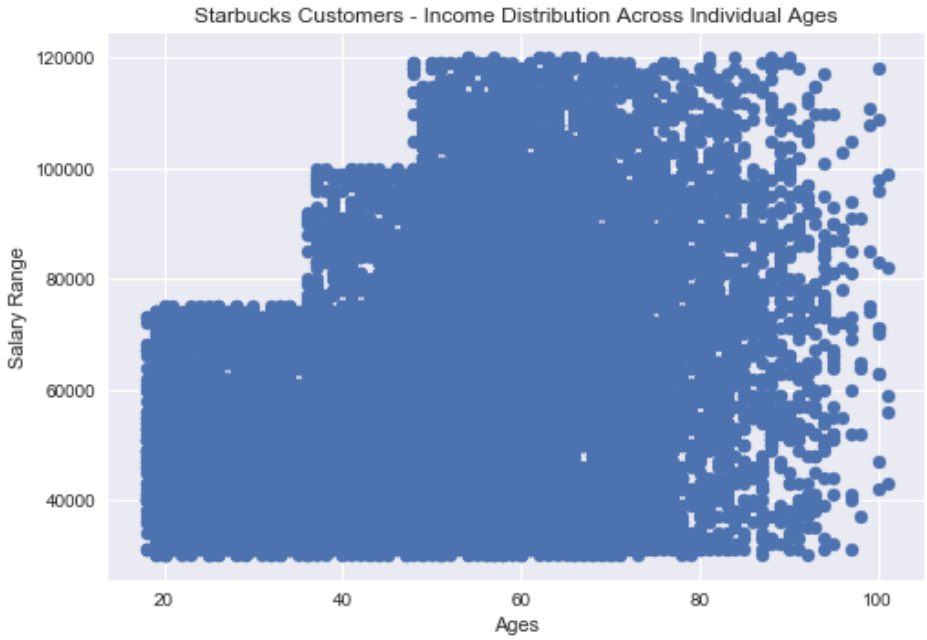

### A1: The actual age distributions of Starbucks customers.

Well, I was certainly wrong with my initial assessment! This is a classic example of why it's important to not make assumptions and let the data speak for itself. I suppose now that I think about it, I do tend to see many folks in those 40-60ish age ranges any time I visit. Perhaps it's because some of them are now retired and enjoy visiting with friends at Starbucks. Perhaps also it is that people in their twenties typically don't have the money to spend on things like Starbucks. I don't know; the data isn't particularly clear on this reasoning. No matter. Let's move on.

**Q2: What are the salary ranges of people across different age groups?**

Similar to our last question, 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. Given our first assessment, I'm going to guess that those 40-60 age ranges have the highest salary ranges given that these people are generally further along in their careers and thus make more money to show for it. Likewise, I definitely expect those younger age ranges to be on the lower end. Let's go ahead and take a look!



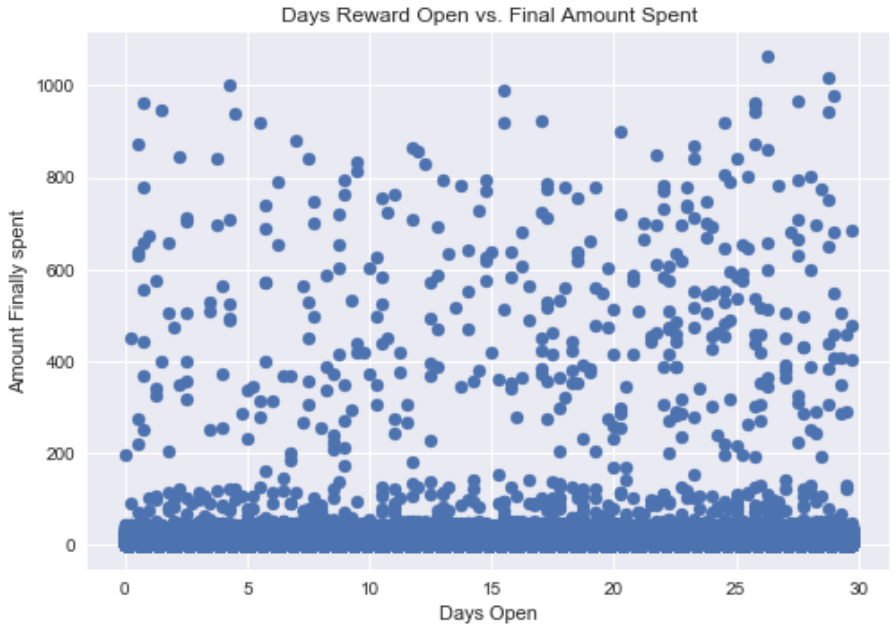



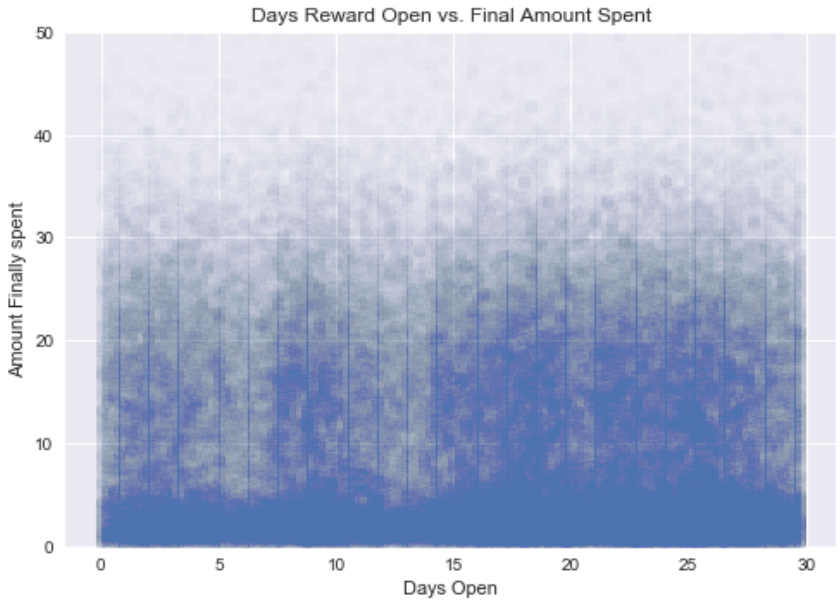
### A2: Analysis of Income Distributions

Okay, when first visualizing the data with the violin plots, my thoughts were affirmed that older customers definitely tend to make more money; however, I noticed something odd in the violin plot: there seemed to be a hard cap on the salary range of younger people. Visualizing the data in a scatter plot using individual ages instead of age ranges, I verified that not only are their caps for younger people, but there are caps on everybody's salaries. Looks like even for older people, the income caps out at $120,000. Theis seems a really odd choice on the data capturer's part. Clearly, there are many people that make more than $120,000 per year, and there are definitely people in their 20's and 30's that make more than that as well. Unfortunately, we're not clued at all into why these caps were put in place, so we're just going to have make due with what we've been given and note that in our finalization as well.

### Q3: What does the correlation between number of days an offer has been open vs. final transaction amount have to tell us?

Honestly, I don't know what to expect from this one, but it still has me curious. Part of me wants to believe that the longer a reward has been available, the less the person is prone to spend. My reasoning for this is that the customer clearly wasn't that excited to run out and redeem the reward right away, so the longer time would be a smaller amount because it's like a "Well, I gotta use it or lose it" kind of thing. My other hypothesis is that we're not going to see any patterns at all here, that there is no correlation at all here. Let's go ahead and let the data tell us which concept is right.






### A3: Analysis of Amount Spent vs. Days Open

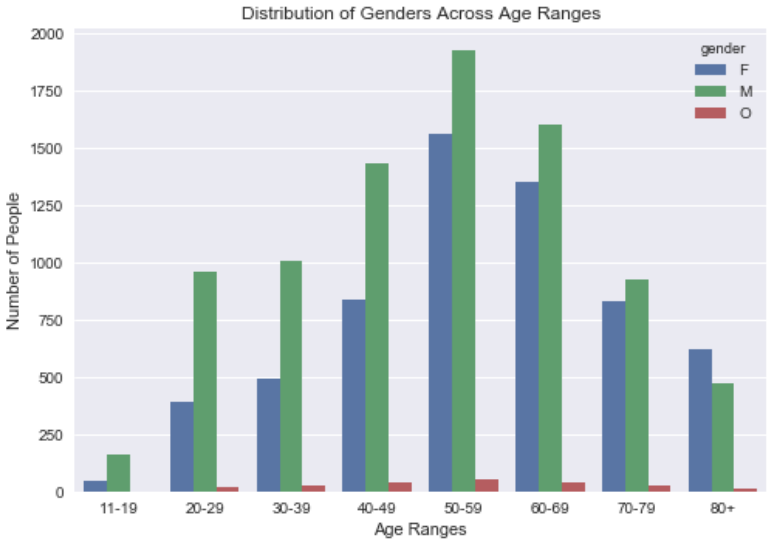
A couple of observations here. First, I didn't realize that rewards only stretched as long as 30 days. This makes my initial curiosity of "use or lose it" tough to measure because I know as a Starbucks customer myself that I sometimes let my rewards sit out there longer than 30 days.

Second, my initial visualization wasn't all that helpful because there were surprisingly a lot of outliers of people upwards of $1000 in a single purchase, which is... sort of beyond me to think about. (Are these people buying the full menu...?) Anyway, this wasn't all that helpful, so I pared down the data in the next visualization. Before moving onto that, it is worth noting that even with these higher dollar amounts, there is no correlation at all between days open and amount spent.

Coming down to our pared down visualization, we again see no definitive ties between how long a reward is open and the dollar amount finally spent. As noted above, I sort of expected this, but I still am glad I got verification on that!

### Q4: Do gender distributions have any major effect on our data here?

For our final EDA piece here, let's take a look at how gender may affect our final analysis. As a reminder, our dataset has indicated three distinct genders: Male, Female, and Other.





### A4: Analysis of gender across our customer data

Several points of interest here. First, we definitely see more males across this dataset than any other gender category. In fact, the only age range we see more women is in the 80+ category, and I'm going to guess that has to do with the fact that women generally tend to live longer than men.

The other really interesting thing here is the salary distribution for women in particular. Where our 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. This

makes me curious that this data has been oddly captured since the general sentiment is that men make more than women. Why is it that this dataset falls counter to that? Unfortunately this is another one of those instances where we truly don't know given solely based on our data.

## Data Preprocessing

With initial analysis / clean up and EDA under our belt, let's move on into formalizing the master dataset I will work from in the remainder of the project. We are going to take several things in mind to engineer some new features that I feel will be helpful when we actually move toward running our unsupervised algorithms. Here are the features we will leverage as part of this master dataset:

- **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

We've come a long way in this project, and now it's time to finally get to what we've been building toward all along: machine learning modeling. As noted in the proposal for this project, we're going to leverage some unsupervised algorithms to cluster data in such a way to find commonalities across customer segments based on a number of features. That said, we're going to take our final dataset (customer_transactions) and drop a few columns just to keep those I think will be relevant to our project.

### Feature Selection

Before moving onto scaling our final data, we'll need to remove some features from our customer_transactions DataFrame. We'll remove the following features for the following reasons:

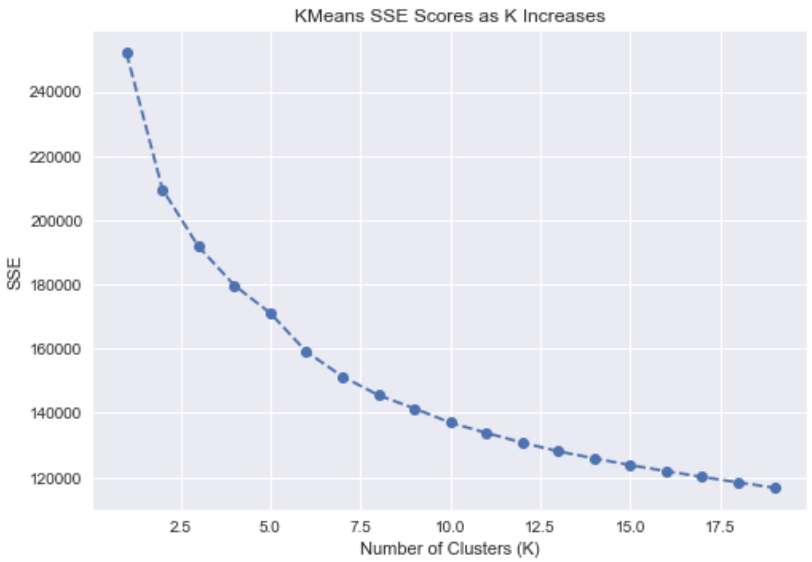
- **customer_id**: proprietary to the row, it is a wholly unique value
- **became_member_on**: a date column that can't be scaled
- **age_range**: a categorical column that can't be scaled

### Feature Scaling

Prior to running our data through the unsupervised model, we will need scale the data so that we are given the best results. To do this, we will leverage scikit-learn's handy tool, StandardScaler.

### Number of Clusters

Now using our final scaled dataset, we're just about ready to run our final clustering algorithm. Before we can do that, however, we first need to determine the ideal number of clusters to place into our algorithm. We'll do that by performing the elbow method and running a series of silhouette scores. When running these, here is the outcome we get:



### Silhouette Scores

```
For n_clusters =  2 The avg silhouette_score is : 0.15698941128814237
For n_clusters =  3 The avg silhouette_score is : 0.13427119202276
For n_clusters =  4 The avg silhouette_score is : 0.13488219332971987
For n_clusters =  5 The avg silhouette_score is : 0.11768868435263667
For n_clusters =  6 The avg silhouette_score is : 0.12344969115506356
For n_clusters =  7 The avg silhouette_score is : 0.12587930376251508
For n_clusters =  8 The avg silhouette_score is : 0.11720782781736412
For n_clusters =  9 The avg silhouette_score is : 0.12062524581312133
For n_clusters =  10 The avg silhouette_score is : 0.10764354206822273
For n_clusters =  11 The avg silhouette_score is : 0.10848069499502569
For n_clusters =  12 The avg silhouette_score is : 0.10597903597420155
For n_clusters =  13 The avg silhouette_score is : 0.10401624479641476
For n_clusters =  14 The avg silhouette_score is : 0.10378410165693669
For n_clusters =  15 The avg silhouette_score is : 0.10332614987699437
For n_clusters =  16 The avg silhouette_score is : 0.09888594464818426
For n_clusters =  17 The avg silhouette_score is : 0.09886660326029735
For n_clusters =  18 The avg silhouette_score is : 0.09656559327175529
For n_clusters =  19 The avg silhouette_score is : 0.09505693204242804
```

Given this information, I've settled on leveraging 4 clusters for our final algorithm. Transparently, I could have gone as high as 9 given the diminishing returns, but I felt 9 clusters would be too many for the purposes of this project. Four clusters is a reasonable amount to comb through in the final evaluation of this project.

### Clustering with a Hierarchical Algorithm

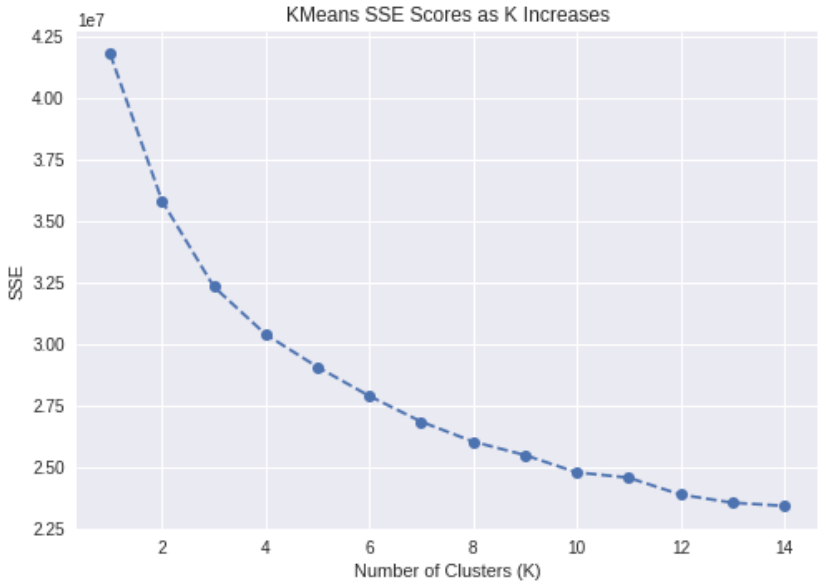
Now that we've determined the number of clusters we'll leverage, we'll go beyond simply using K-Means to leveraging one of the hierarchical-based unsupervised algorithms. This is because K-Means alone tends to be a little too simplistic. Now, we could either leverage DBSCAN or hierarchical-based clustering, but given that DBSCAN focuses more on weeding out noise, I'm not sure that is the best choice for our purposes. That said, we'll use leverage scikit-learn's AgglomerativeClustering and leave it with the default of a ward-based linkage.

## Benchmark Definition & Comparison

Before we move on, I want to discuss the benchmarks and metric evaluation since these came heavily into play in the prior section. Given that there isn't necessarily a labelled right or wrong to the provided dataset, we can't really objectively evaluate how well our unsupervised dataset performed after it has already processed through the data. What we can do, however, is leverage our benchmark and metrics to determine the ideal number of clusters for the final algorithm.

We already explored leveraging the elbow method and silhouette score in the previous section, so I instead want to focus on our benchmark comparison here. For this project, I chose to leverage insights from a similar project I did as part of the Udacity Data Scientist nanodegree program. That project was the Bertelsmann-Arvato customer segmentation project. (Link to Hundley - Arvato Customer Segments GitHub) Given that that project was also very focused on customer segmentation and already reviewed by Udacity, I believe the insights gleaned there will serve well as a benchmark for this project.

Utilizing a very similar elbow method in that project, here is what the results were from there:

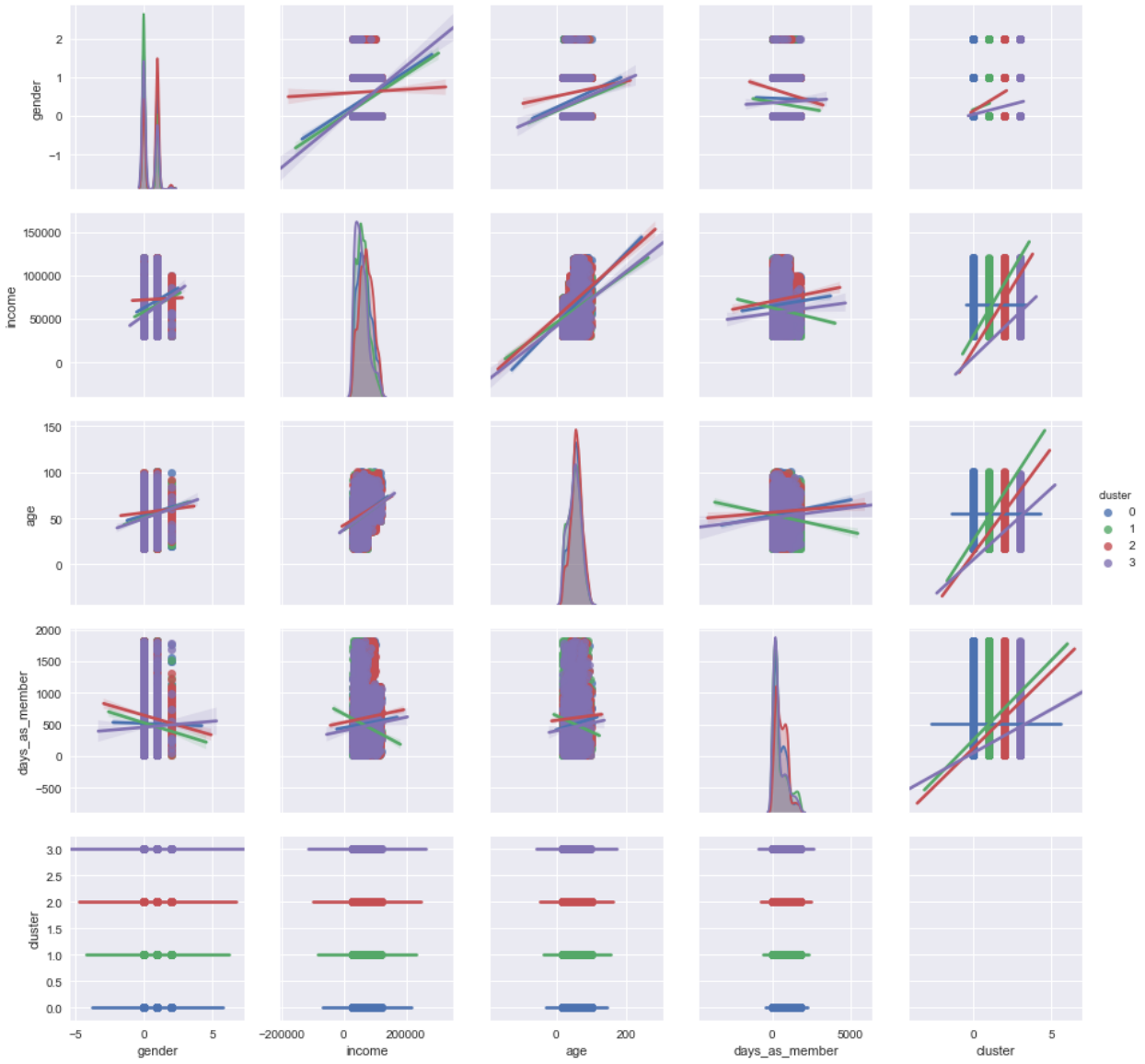


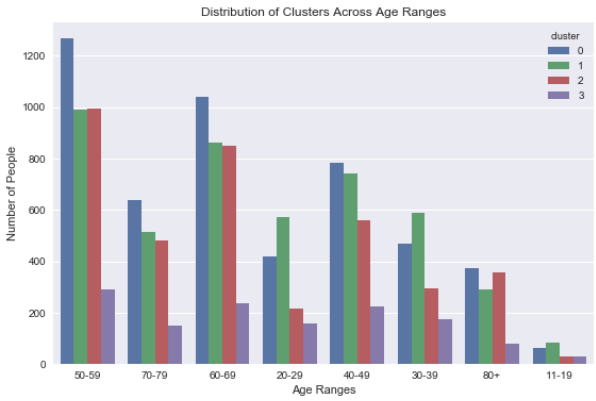

As you can see, the diagram follows a very similar shape to our one from above, including how we might have opted to leverage nine clusters instead of four. In fact, in that project I did end up leveraging nine clusters, but that was partially because there was so much more data in that dataset. Given the relative size of this dataset, I still hold to utilizing four clusters, but this benchmark affirms that our methodology for determining the number of clusters is sound.

## Final Analysis & Evaluation

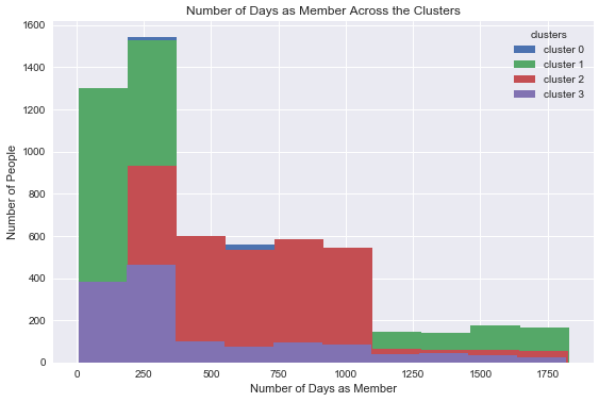
Now that we've gathered our various clusters from our hierarchical algorithm, let's go ahead and visualize the results! Like we did in the Exploratory Data Analysis section, we're going to explore two more high level questions and how these might be utilized by Starbucks to adjust their rewards program.

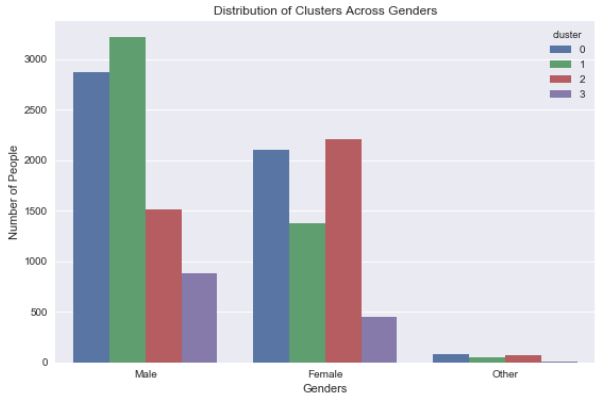
**Q5: What personal attributes of our customers are defined throughout each of our clusters?**
First, let's take a look at the personal attributes of our customers as clustered by our algorithm. We'll visualize this first in a PairPlot and then further down in more easy to read diagrams.










## A5: Analysis of Clustered Personal Attributes

Lots of great insights here! Let's cover each of the respective clusters within the respective sections below.

### Cluster 0

Easily the largest cluster, cluster 0 tended to consist of older people with higher incomes. As evidenced by our countplot with the age ranges, the ages of these folks 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, and I would probably only account the difference as the fact that the original dataset had more men to begin with, anyway. Admittedly, however, 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 distrubution here amongst the 20-40 year old crowd. Interestingly, the gap between males and females is quite large in this cluster. These folks 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

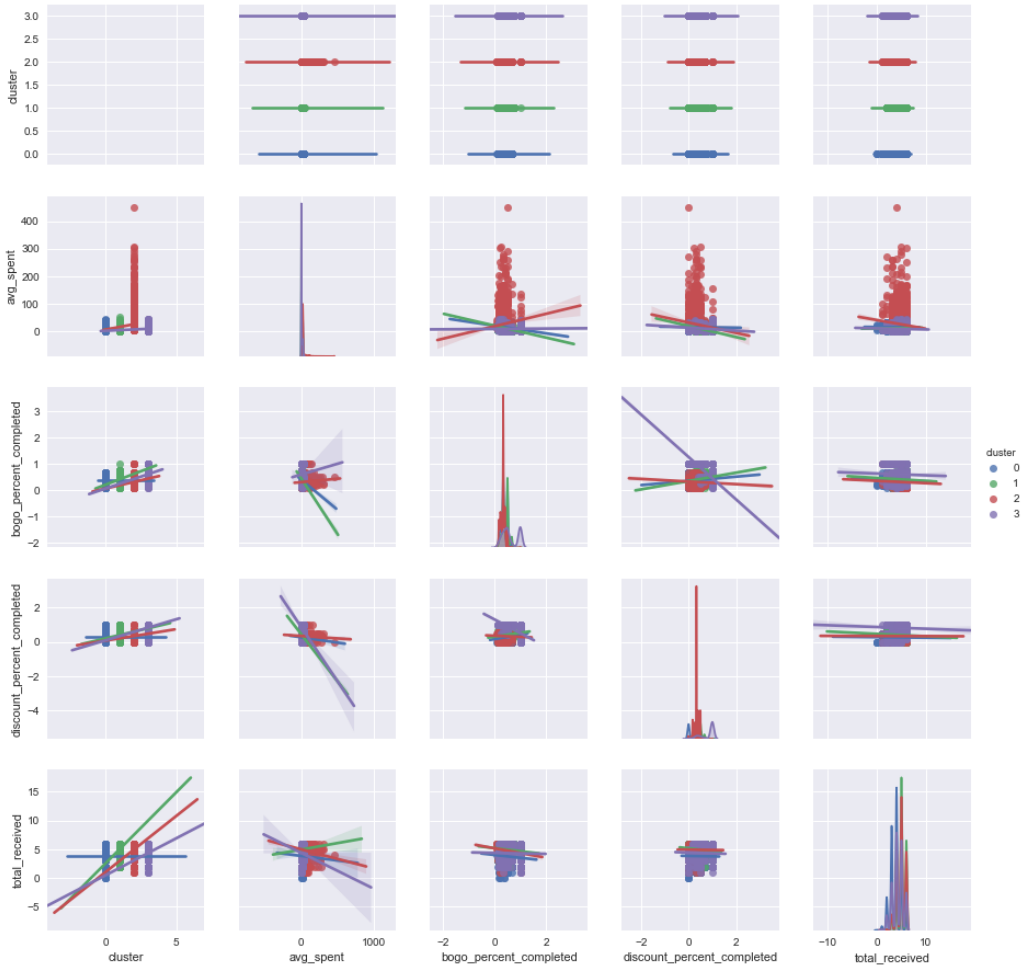
This cluster here has a couple of interesting highlights. First, this is the only customer where there were more males than females, and this cluster also contained the largest distribution of people who have been members of the Starbucks rewards program for some time. Income tended to be higher here which is not surprising given that our EDA showed us that women tended to make more than men.

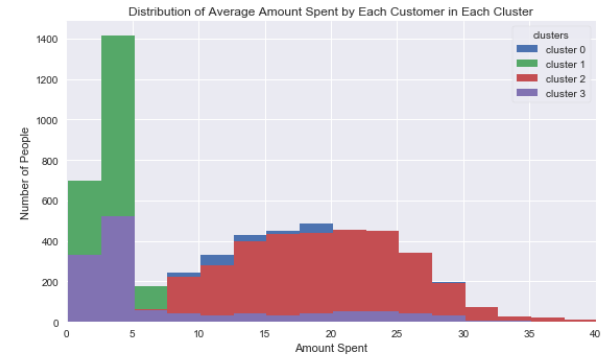
### Cluster 3

Finally, this last cluster is easily the smallest of our four clusters, yet it bears some striking similarities to cluster 0 in a few ways. Namely, the distribution of days as member and gender are fairly similar. Perhaps the exception that separates it from cluster 0 is that there tended to be more people in the younger age range, and I suppose that would make sense given that younger people consisted of a smaller subset of the original data provided.

### Q6: What do the behavioral attributes look like across our clusters?

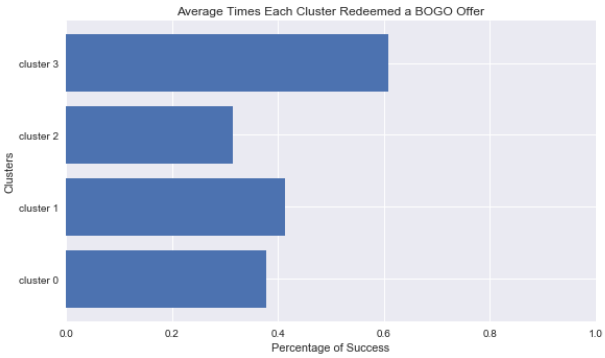
Finally, let's wrap up our analysis by looking at the clusters across some of the more behavioral attributes.

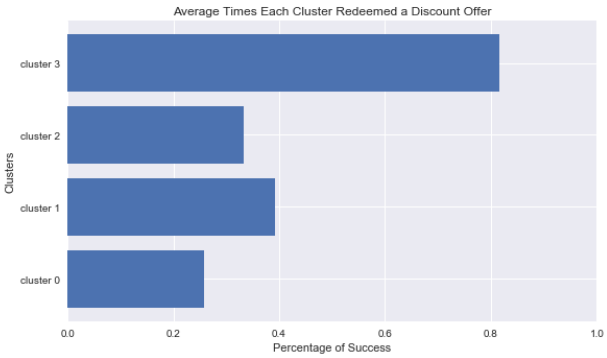











## A6: Analysis of Clustered Behavioral Attributes

Even more fascinating stuff in here. Let's cover each of the respective clusters within the respective sections below.

### Cluster 0

Lots of interesting things to note in here. First, this cluster on average seems to be the biggest spenders, with an average transaction total peaking at ~$19. I noted earlier on in the report that I felt that this amount is really high, but now that I think about it, maybe it's on point.

Remember, this particular cluster contains our older customers, and perhaps these customers are buying for multiple people, like family members or friends. Considering that, perhaps that average number makes more sense.

Interestingly, this cluster has the lowest yield of using discount offers and pretty close to last for BOGO offers. It's also interesting to note that Starbucks almost caps out how many offers they send to these folks. We have a lot of people reporting receiving 2-3 offers, and that number falls off big time after that. Is it because Starbucks is tired of trying to appeal to a group that doesn't take advantage of these offers? Not to try stereotyping older folks too much, but they do tend toward not leveraging technology as well. Is it because these offers are being managed electronically? More on this down in cluster 3.

### Cluster 1

Recalling from above that this cluster tends toward a younger crowd, we see things like the average amount spent spike toward the lower end. Interestingly, we also see that this particular group gets hit hard with offers, spiking big time around 5 offers sent. Taking a closer look, it looks like there's a stronger leaning toward BOGO offers over discount offers. In both cases, the success percentage for both is about equal, hovering around 40%. This number isn't necessarily great, and it's not clear why this is the case. Perhaps the difficulty levels for these rewards are too high for this cluster's general demographics.

### Cluster 2

Our 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. I'm not sure why this is necessarily the case since the cluster doesn't tend to react well to these clusters, noting the lowest BOGO success precentage at about 25%. I think there's definitely a lot of opportunity for improvement here given the demographic information of this group and current low success rating across both offers.

### Cluster 3

This group is the most curious of all. Recall from the previous section that this group's demographics closely aligned to that of cluster 0's with the exception that this group tends a little younger than the cluster 0's folks. I say that this group is curious because of how high of a success rate this group has with offers, especially with discount based offers. Now granted, this cluster is definitely the smallest of the bunch, but that doesn't negate the fact that something is definitely working. Perhaps Starbucks could lean into leaning more about this cluster to discover what is causing the level of success.

## Conclusion

For as much work that I've put into this project, it wouldn't have been unfeasible to take other different approaches. There were certain pieces of data that were intentionally omitted that we just as easily could have used. We could have leveraged deep learning and allow a model like that to determine the features for me. Those are admittedly beyond my skill level at this point, but I still think we ended up with a lot of great insights in the end. I'm glad we took the approach that we did and am looking forward to applying these skills in other projects!