Starbucks-Customer-Clusters

Cluster analysis to show Starbucks customers fall into 4 distinct categories.

The data set contains simulated data that mimics customer behaviour on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.
Not all users receive the same offer, and that is the challenge to solve with this data set.
The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.
Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. It is evident in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.
The transactional data is given showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

Example:

To give an example, a user could receive a discount offer buy 10 dollars get 2 off on Monday. The offer is valid for 10 days from receipt. If the customer accumulates at least 10 dollars in purchases during the validity period, the customer completes the offer. However, there are a few things to watch out for in this data set. Customers do not opt into the offers that they receive; in other words, a user can receive an offer, never actually view the offer, and still complete the offer. For example, a user might receive the "buy 10 dollars get 2 dollars off offer", but the user never opens the offer during the 10 day validity period. The customer spends 15 dollars during those ten days. There will be an offer completion record in the data set; however, the customer was not influenced by the offer because the customer never viewed the offer.

Data Cleanup:

The Starbucks dataset is split up into three different files: **profile**, which has low level information about customers, **portfolio**, which has information about different promotional offers that can be received, and **transcript**, which has all purchase history and information on when the customer received, viewed, and completed their promotional offers.

```
profile.head()
```

	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN

```
profile.head()
```

	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN

Raw profile data

```
portfolio.head()
```

	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5

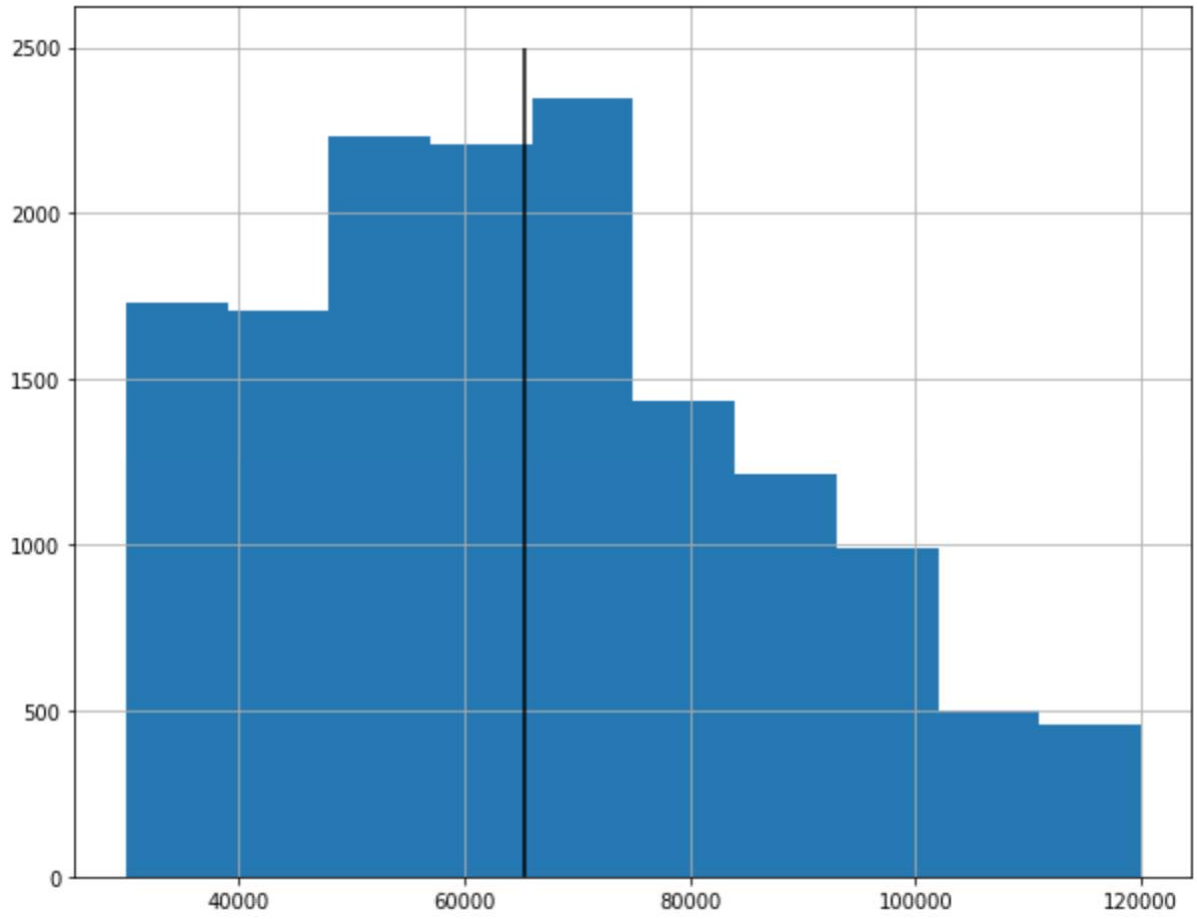
Raw portfolio data

```
transcript.head()
```

	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}
3	offer received	8ec6ce2a7e7949b1bf142def7d0e0586	0	{'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'}
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}

Raw transcript data

PROFILE: For cleaning up the profile data, I needed to figure out how I wanted to handle genders and income that were NULL, and change the become_member_on field to an actual date and not just a string. Looking at the missing genders, those records were responsible for NULL incomes as well. I decided to make missing genders have the value of "U" for unknown instead of deleting those records (they made up about 15% of the data). For figuring out an appropriate selection for income, I looked at the distribution of income for the whole population:



This needs to be cleaned up a little nicer

The black line shows where the mean of the distribution is, and that seemed like a good selection since it doesn't lean too far one way or the other on our income distribution. After creating the member_date field, which was derived from the became_member_on field, I also split up that date into the member_year, member_month, and member_day to see if that

would help give any more additional information. Here is what the final output looked like:

```
profile.head()
```

	age	gender	id	income	member_year	member_month	member_day	member_date
0	118	U	68be06ca386d4c31939f3a4f0e3dd783	65404.991568	2017	2	12	2017-02-12
1	55	F	0610b486422d4921ae7d2bf64640c50b	112000.000000	2017	7	15	2017-07-15
2	118	U	38fe809add3b4fcf9315a9694bb96ff5	65404.991568	2018	7	12	2018-07-12
3	75	F	78afa995795e4d85b5d9ceeca43f5fef	100000.000000	2017	5	9	2017-05-09
4	118	U	a03223e636434f42ac4c3df47e8bac43	65404.991568	2017	8	4	2017-08-04

Cleaned profile information

PORTFOLIO: Each promotion has an array that shows the different ways someone could receive the promotion. It’s good practice to transform this into bit flags.

```
portfolio.head()
```

	difficulty	duration	id	offer_type	reward	email	mobile	social	web
0	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10	1	1	1	0
1	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10	1	1	1	1
2	0	4	3f207df678b143eea3cee63160fa8bed	informational	0	1	1	0	1
3	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5	1	1	0	1
4	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5	1	0	0	1

Cleaned portfolio information

TRANSCRIPT: The value column in the transcript dataframe is a dictionary that has a key and value associated with it. The different key values are offer id and amount. Offer id is associated with the receiving, viewing, and completing of an offer, while amount is just the transactional amount. I changed the value field to show the value of the promotion id or transaction amount, and the value_type field then became the key of the dictionary.

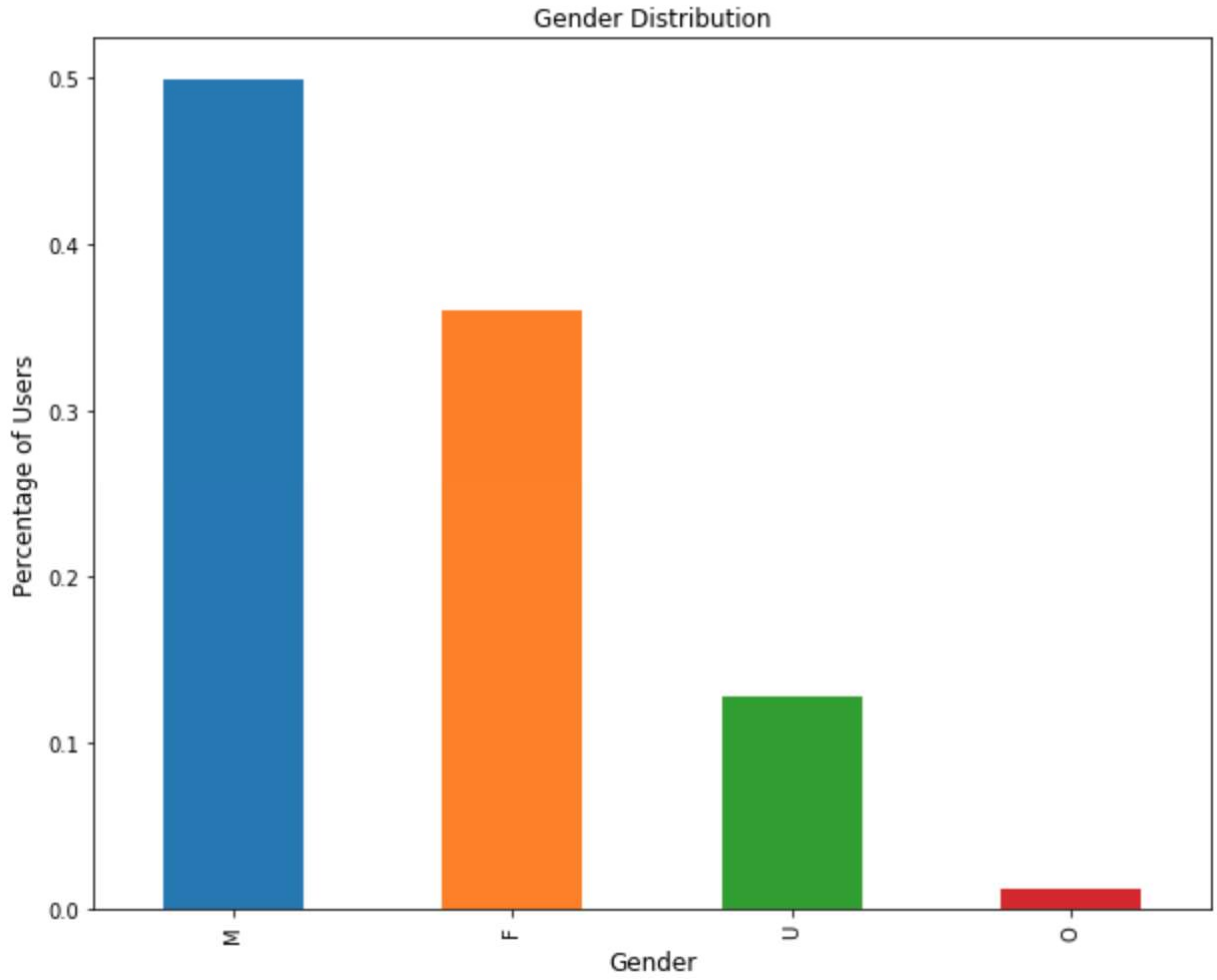
```
transcript.head()
```

	event	person	time	value	value_type
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	9b98b8c7a33c4b65b9aebfe6a799e6d9	offer id
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	0b1e1539f2cc45b7b9fa7c272da2e1d7	offer id
2	offer received	e2127556f4f64592b11af22de27a7932	0	2906b810c7d4411798c6938adc9daaa5	offer id
3	offer received	8ec6ce2a7e7949b1bf142def7d0e0586	0	fafdcd668e3743c1bb461111dcafc2a4	offer id
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	4d5c57ea9a6940dd891ad53e9dbe8da0	offer id

Cleaned transcript information

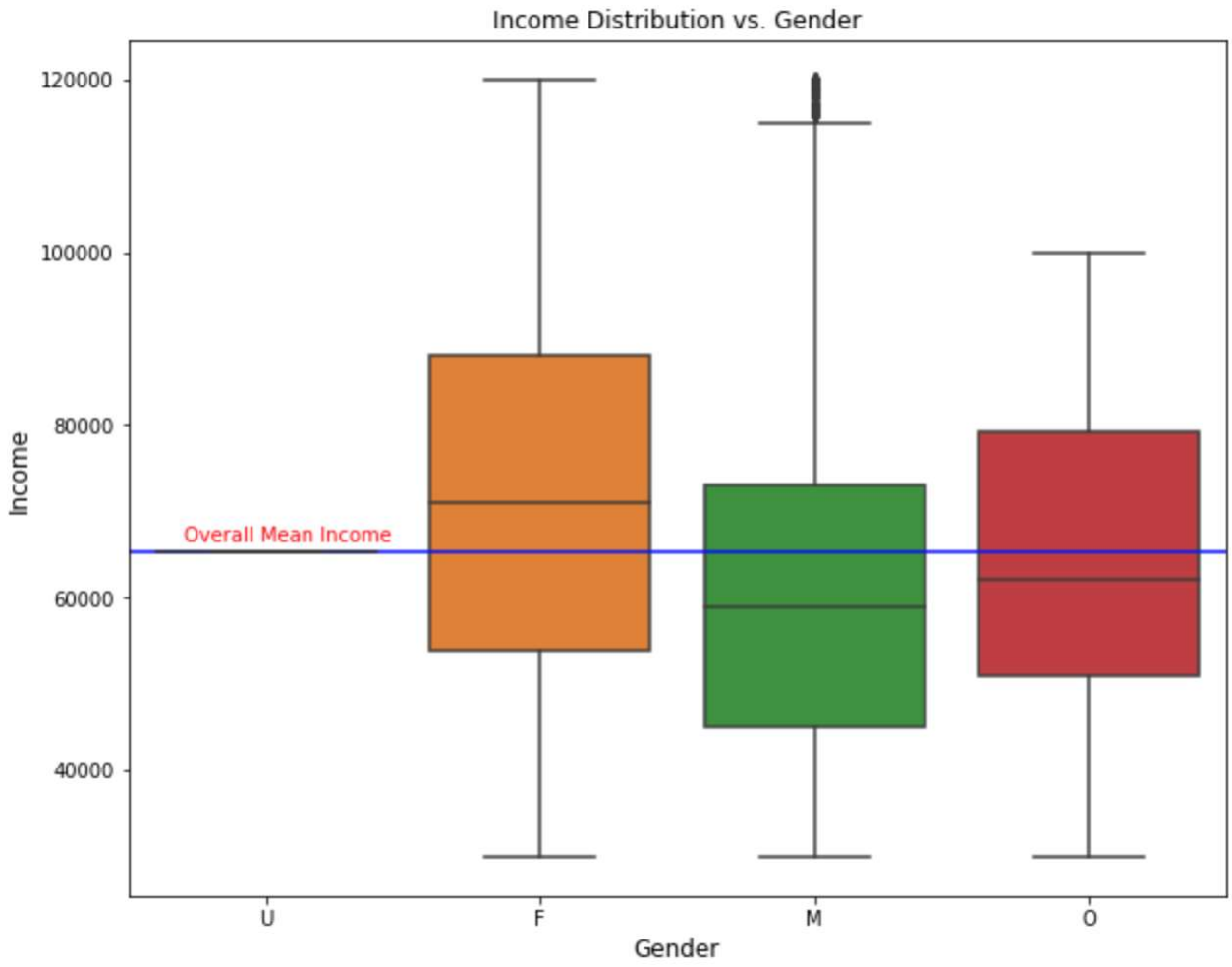
Exploratory Analysis

First thing to look at customer data is to see what type of people make up the customer segments without having to go too in depth. Starbucks customers are roughly 50% male, 35% female, and the rest being other or unknown.



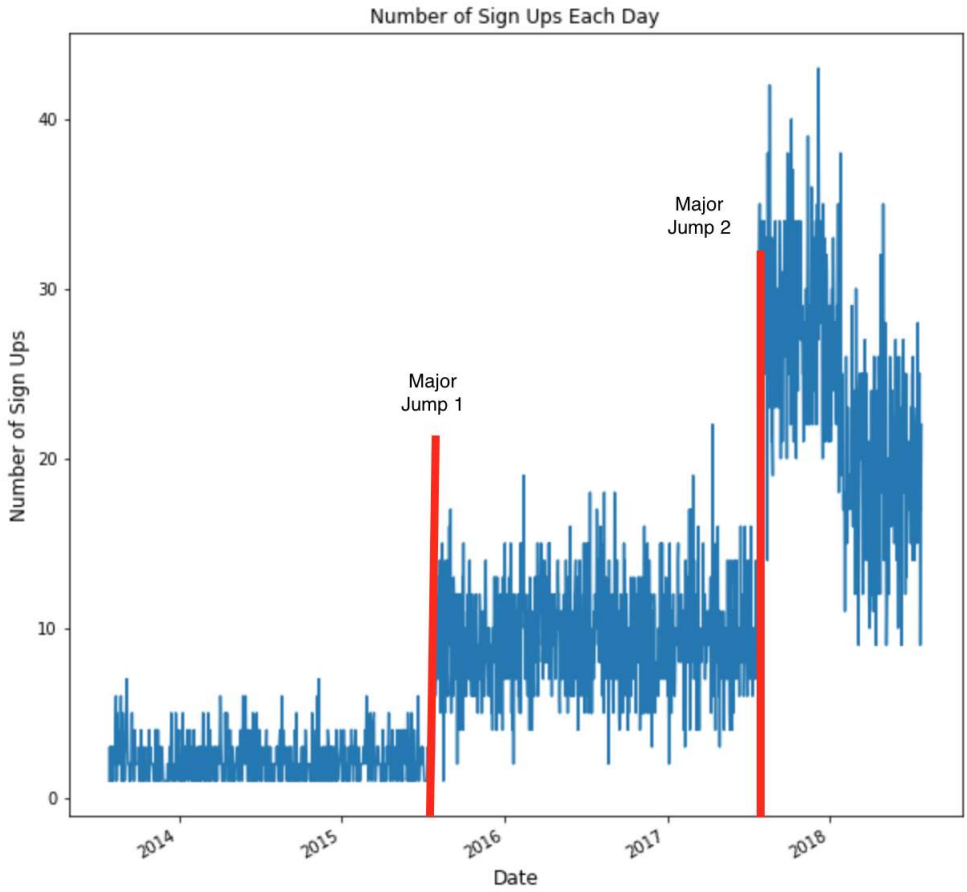
M — Male, F — Female, U — Unknown, O — Other

When looking at the income distribution, females have a wider distribution of incomes and also have higher income than males and other genders. Since we replaced all unknown gender records with the average income of the whole population, there is no distribution to analyze.



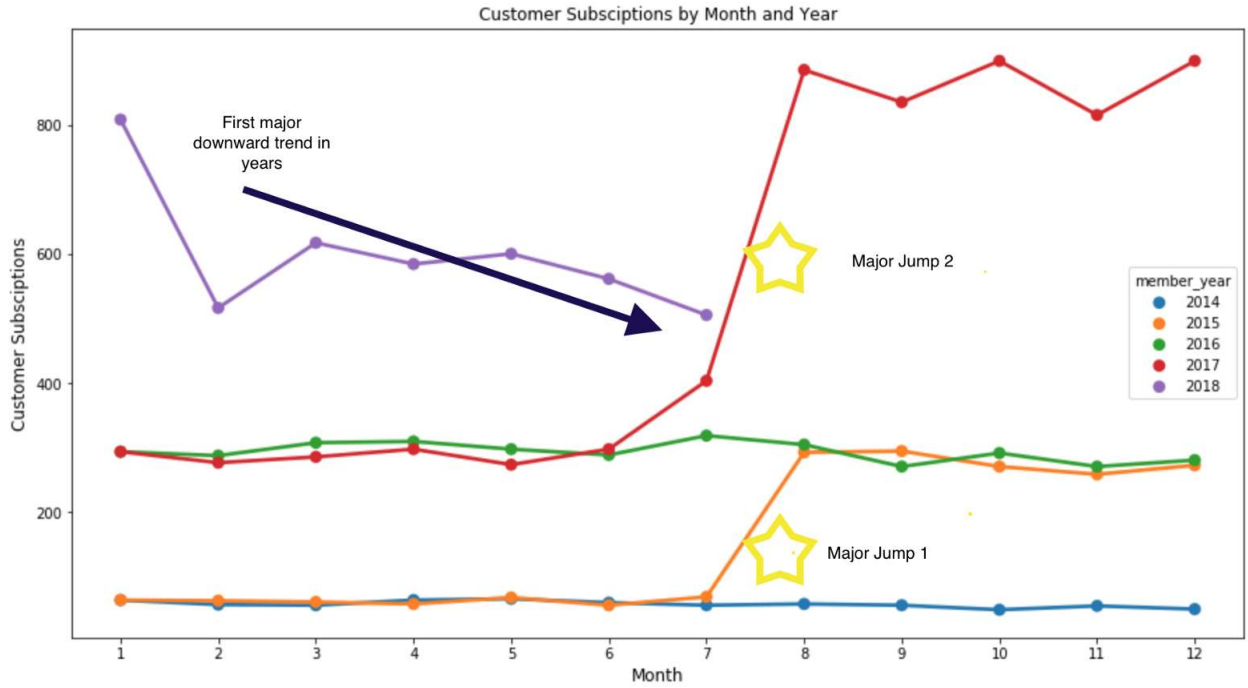
Females are holding up the weight for income distribution and are causing the overall mean income to be higher than both the mean male income and mean other income.

Another piece of interesting information to look at is the trend of when customers became Starbucks members. Looking at the count of customers who signed up over the last few years, there are two large jumps of increased user amounts. I would be more curious to see why these dates are so relevant, whether that be app upgrades or universal wide promotions (download the app and we will give you a free coffee mug, etc).



Number of daily customer sign ups over time

To make things a more clear, I changed up the graph to show customer sign ups at a monthly level with year being the line color. I starred the two major jumps that we saw in the graph above, and they look like they took off at exactly the same time. That has to be due to something done by the marketing department. Starbucks also has a downward trend in monthly subscriptions this year, almost wiping out any progress that they had made from the large surge in customer sign ups from the previous year.



Customer subscriptions per month with year being the color code

Customers seem to have similar patterns of when they sign up for being a Starbucks member and there are different income distributions based on their gender input, but that still doesn't give a good representation of the type of buyer they are and what characteristics they have in common with other buyers. For that, we need to take a deeper look into their purchase history and how responsive they are to different promotions.

Transactional and Promotional Analysis

Looking through the transactional data, it is easy to see when a person receives a promotion, when they view it, and when they complete it. This example is a random person viewing a discount offer that day. The discount is spend $10, receive a $2 credit. You can see that the time is 0 for all of the different values, meaning they acted on this offer on the same day they received it. From this segment, we can try to figure the following information:

1. How many promotions have they received (BOGO, discount, informational)
2. What was the completion percentage of each type of promotion (BOGO and discount, every informational promotion is completed the same day it is issued)
3. How many total transactions have they made since becoming a Starbucks member
4. What is the average transaction amount spent
5. What is the average and median days between purchases for each customer
6. What is the average number of days it takes to complete an offer

	event	person	time	value	value_type
36	offer received	9fa9ae8f57894cc9a3b8a9bbe0fc1b2f	0	2906b810c7d4411798c6938adc9daaa5	offer id
12656	offer viewed	9fa9ae8f57894cc9a3b8a9bbe0fc1b2f	0	2906b810c7d4411798c6938adc9daaa5	offer id
12658	offer completed	9fa9ae8f57894cc9a3b8a9bbe0fc1b2f	0	2906b810c7d4411798c6938adc9daaa5	offer_id

Random customer receiving, viewing, and completing a promotional offer.

When trying to see people's success rates in completing promotional offers, I have to join the transaction dataframe to itself, joining on person and value where the value_type says "offer id". We also need to make sure that the offer is completed after or on the same day as the received time, and that it fits in a window of the duration of the offer. This can be tricky since a customer can receive the same promotion more than once (I found that out the hard way and it was a nightmare to go back and figure that out). This also raised other

questions because there were instances where people completed their offer at a date later than the promotion expiration date. An example:

	person	offer_time	value	CompletedOffer	DaysToCompleteOffer	difficulty	duration	offer_type
48845	db1805ed333844978e8b46ed3e4643ae	576.0	f19421c1d4aa40978ebb69ca19b0e20d	0	NaN	5	5	bogo
48846	eec70ab28af74a22a4aeb889c0317944	576.0	f19421c1d4aa40978ebb69ca19b0e20d	1	18.0	5	5	bogo
48847	31e915c24163436790b97c1d45b545f6	576.0	f19421c1d4aa40978ebb69ca19b0e20d	1	18.0	5	5	bogo
48848	361539b15a6243dc834c6b25e481570b	576.0	f19421c1d4aa40978ebb69ca19b0e20d	0	NaN	5	5	bogo
48849	eb7dd979f4644052a5c401a01f129132	576.0	f19421c1d4aa40978ebb69ca19b0e20d	0	NaN	5	5	bogo

This promotion was offered to these people at day 576 to them, and both people completed the offer 18 days later. This doesn't make sense because the duration was only for 5 days? So maybe I am misunderstanding what duration means, or duration doesn't really matter. I ended up ignoring the duration column.

From here, we can aggregate the data to see at a high level the different promotions, their success rate, and their net reward to the consumer, and the net worth to Starbucks. From an overall perspective, the higher the difficulty (the more money you have to spend), the less likely people are going to complete the offer. Even though the $5 for $5 BOGO and the $10 for $10 BOGO have the same net reward for the consumer, people are more inclined to get the $5 for $5 probably due to convenience. The most successful offer is the $3 for $7 discount, which actually makes Starbucks money in the end. We also see that the $5 for $20 discount is a complete waste of time, only being completed 10% of the time. For the average time it takes to complete the offer, people are more likely to get the $5 for $5 BOGO quicker than any other offer, while the $5 for $20 discount promotion takes the longest to complete, probably due to the larger amount that needs to be spent.

	offer_type	difficulty	reward	CompletedOffers	TotalCompletions	AvgDaysToComplete	NetReward	NetWorth
0	bogo	5	5	0.308293	2606	8.638526	0	-0.000000
1	bogo	10	10	0.217532	2149	9.001396	0	-0.000000
2	discount	7	3	0.383878	1362	8.938326	-4	1.535513
3	discount	10	2	0.279512	2060	8.906796	-8	2.236092
4	discount	20	5	0.090034	393	9.511450	-15	1.350515

Completed Offers is the success percentage of that offer being completed, Total Completions is the number of times that promotion was completed, Avg Days To Complete is the average days it takes for that promotion to be completed, Net Reward is the reward

minus the difficulty, and Net Worth is the Net Reward * Completed offers. This shows how much money Starbucks makes on that promotion.

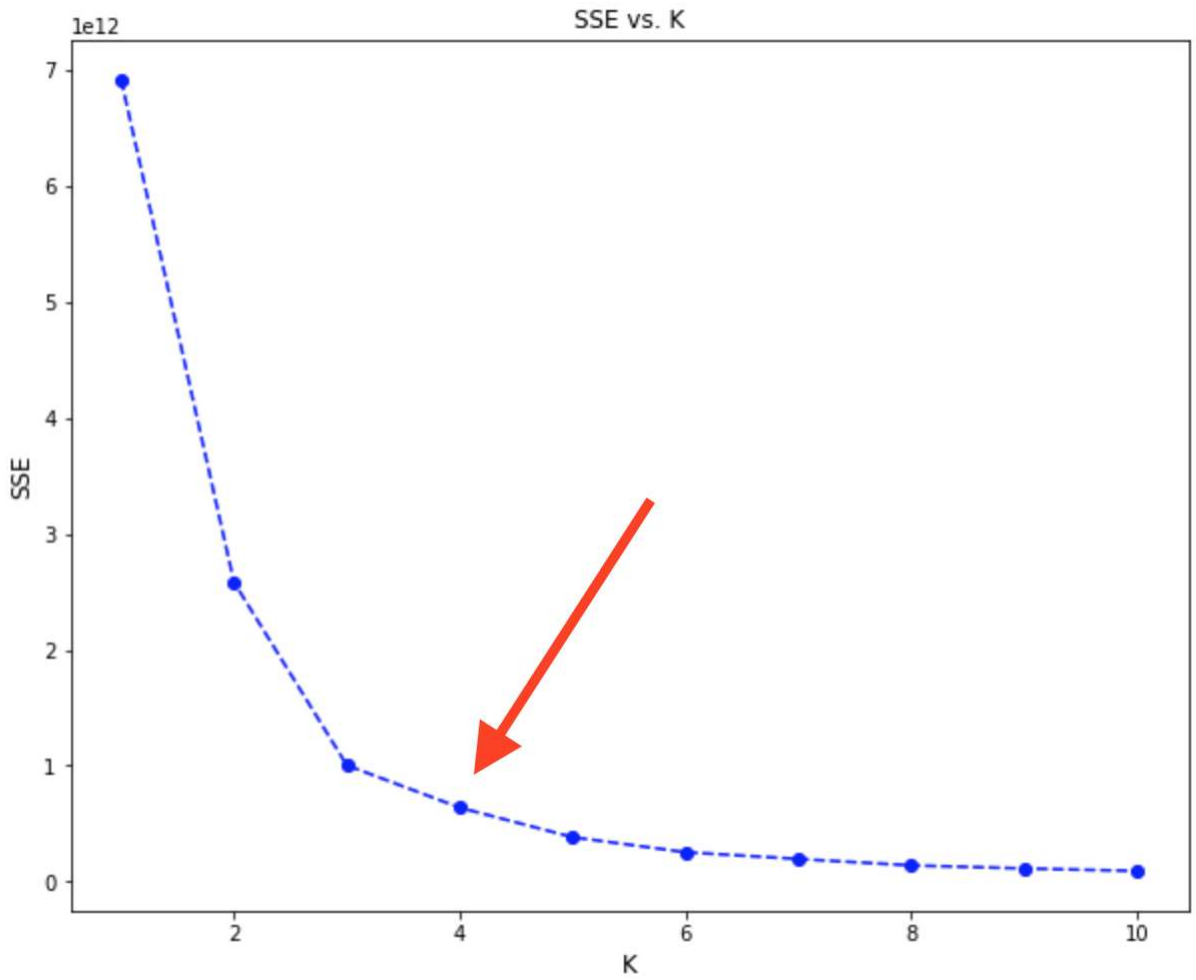
Net worth should be the most important segment to Starbucks. With BOGO's, Starbucks doesn't make any money off of the deal. They're just biting the bullet to get people to get their reward and then hopefully spend more money in the future. So, for both BOGO deals, the net worth will always be $0 any time they send out that promotion; on the other hand, discounts do give some monetary value to Starbucks. The net worth of discount deals are the total value that can be made off of the discount, multiplied by the percentage chance that the promotion is completed. The $5 for $20 promotion might favor Starbucks by $15, but if only 10% of people complete the promotion, they should only expect to receive $1.35 * the number of people who actually receive the promotion. So in terms of net worth, the $2 for $10 discount is the best promotion to give their customers.

The structure of the discount and BOGO promotions should be pushed in such a way that is financially smart for Starbucks. Discount promotions should be given to customers that will come back on a much more constant basis and do not need to be highly incentivized to come back. If they're going to come back anyway, giving them a BOGO doesn't make sense because you're not making money on the deal, but you will make more money back if you give them discounts, it's a win-win for both Starbucks and the customer. People who are not guaranteed to come back consistently should be given BOGO's to reengage their interest and get them to become more likely to spend more money at their stores. The next step is to break out the customers into more appropriate clusters based on their purchasing habits and make better decisions on who gets what promotional deals.

Clustering Segments

Using machine learning with clustering algorithms is one of the more interesting analysis that data scientists encounter due to the fact that there is no correct “answer”. Predicting stock prices or predicting fraudulent bank transactions can be looked at historical yes / no answers, but clustering doesn’t have a label associated to it. It just shows which values are most similar to each other and groups them together.

You can technically pick any number of clusters that you would like to represent your data, but one way to evaluate whether your cluster is performing at a high level is to run the **elbow method**. In the elbow method, we are looking for significant drops in the sum of squared errors (SSE) from each point to it’s affiliated centroid. Since each increase in K (number of clusters) will create more clusters with fewer number of points, there will be an overall decreasing trend of SSE and K increases. Since a lower number of clusters is easier to decipher and analyze, we want K to be the last significant drop in SSE before it starts to flatten out, hence, looking for the “elbow” in the graph.



Elbow method chart looking at the number of clusters that should be analysed.

From the chart, 4 clusters are an appropriate number of clusters to analyse on the data. That fits well with our problem since the 4 clusters can be broken up into these types of customers:

1. Won't react to any promotional deals
2. Will favour BOGO's over discounts
3. Will favour discounts over BOGO's
4. Will respond to body BOGO's and discounts

The input data would be based off a customer matrix that included the following features:
discount_total_offers, discount_completion_pct,
discount_min_completion_days, discount_max_completion_days,
discount_completed_offers,
discount_avg_completion_days, discount_avg_net_reward,
bogo_total_offers, bogo_completion_pct, bogo_completed_offers
bogo_min_completion_days,
bogo_max_completion_days, bogo_avg_completion_days,
bogo_avg_net_reward, informational_promotions, age, gender,
income, total_transactions, min_transaction_day,
max_transaction_day, avg_transaction, total_transaction_amount,
median_days_between_purchases, avg_days_between_purchases

This is a combination of personal attributes (age, gender, income), BOGO and discount attributes (percent completed, average net reward, average number of days to complete, the fastest time it took to complete, the largest time it took to complete), informational promotions (how many they received), and overall transaction trends (number of transactions made, total spent, average and median days between transactions). If any of the values were NULL, they were replaced with a 0 since the customer didn't participate in that specific sector of interest (never made a transaction, never completed a promotional offer, etc). I also had to turn the gender field from a categorical variable to 4 dummy variables since the clustering algorithm only takes numeric values.

There is no testing and training set with clustering due to there being no right or wrong answers and no before and after values. Once the clusters were appended to the customers matrix, I wanted to visualize different plots of distributions that can help identify what types of customers were clustered together.

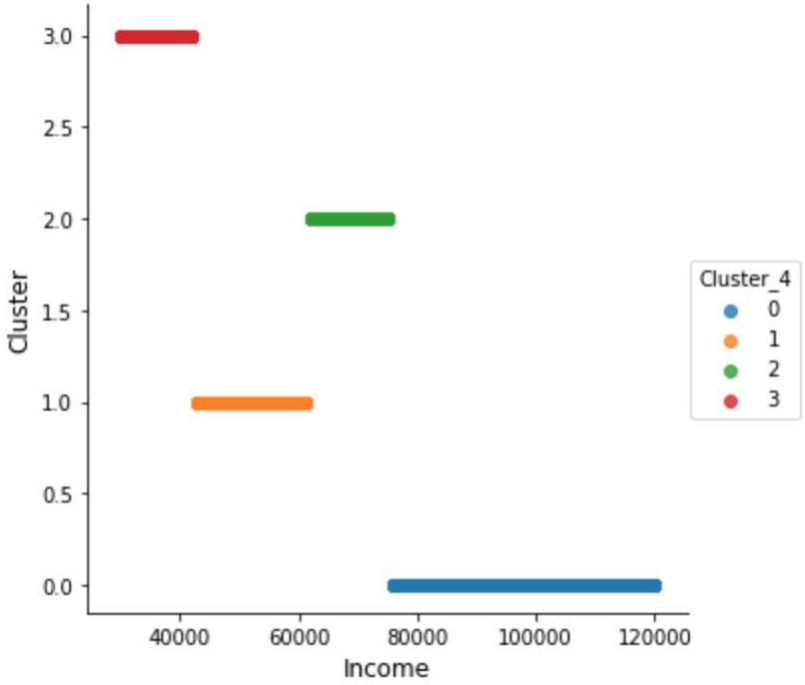
2	0.334941
1	0.260647
0	0.256176
3	0.148235

Cluster 2 had the largest group size at 33% of the total population (17,000 customers which comes out to about 5,694 customers), while cluster 3 was the smallest segment at roughly 15%. I am most interested in cluster 3 because that seems to be a very small niche market compared to the other clusters.

I first started with looking at information that didn't have to do with promotional success values to get an idea of what type of person fits into these clusters. I looked for plots that seemed to be very color segmented with little to no overlap.

Seaborn pair plot that cross analyzes different distributions and color codes them based on the associated cluster.

Looking at the different plots above, the plots most segmented seem to be any plot that has to do with income. It looks like cluster 3 is people who have a small income, while cluster 0 looks like it has people who are making a much larger income. This makes sense why cluster 3 was the smallest group of all the clusters because most Starbucks income lower quartiles were right around $45K, but this group's income tops out at around that amount.

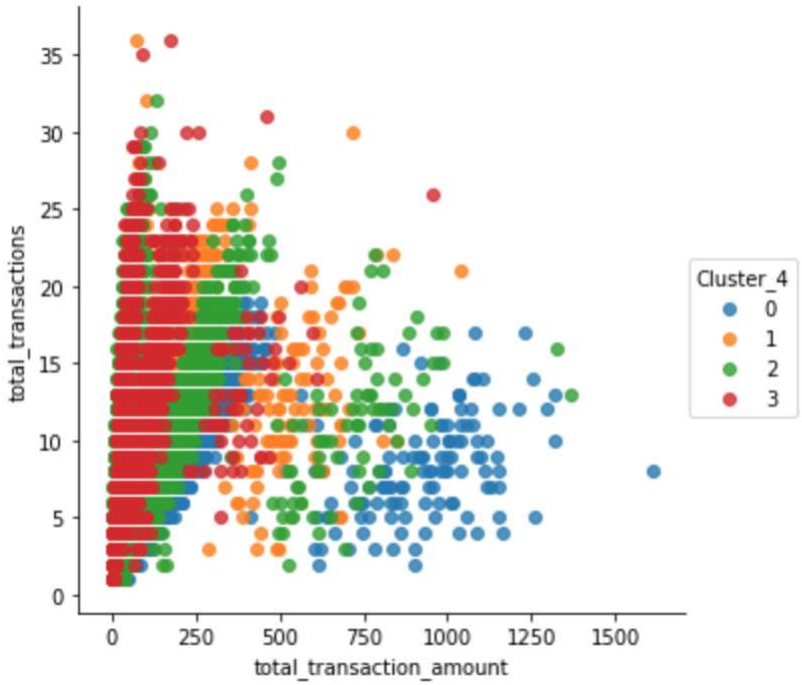


Income distributions separated out by Cluster

Now looking at transactional data, it's much harder to separate out the clusters, but we can still take away some information. If we take a look at completion percentages of BOGO's and discounts, it looks like cluster 3 seems range all over the place with both promotional offers, but cluster 0 seems to stay in the success rate above .5 for both types of promotions.

Seaborn pair plots on transactional information

Looking at total transactions vs total transaction amounts, it looks like cluster 0 is people who spend more money per transaction and don't go as often, while cluster 3 is the opposite. Clusters 1 and 2 are both in the middle, which makes sense, they don't really lean one way or the other since they make up a majority of the population.



Total transactions vs total transaction amount shows how many times customers make transactions and how much they spend total.

These all are interesting facts about the similarities and differences of these clusters, but how would Starbucks make more money than just blindly giving people different promotions throughout the year?

We can establish if our clustering algorithm is helpful in our customer segmentation by the **expected return** of our promotions. The expected return takes into account the total possible value Starbucks makes from the discount multiplied by the probability of the event (in this case someone completing their promotional deal). We want to evaluate the net worth of the overall population compared to our cluster to determine if our cluster is more fit for one type of promotion or the other.

If we compare how each cluster did for each reward compared to the overall population average, it can help leverage which promotion we should focus on. Remember, BOGO's should be to excite the customer to get back to Starbucks and discounts should focus on customers who are consistently coming in to retain some of their

purchases back. If we look at cluster 0, they responded to the $5 for $20 promotional deal at a 34% success rate! The whole population success rate was just 9%, so why in the world would you offer a BOGO deal to them? You can make $5.10 on each customer when you give them that deal as opposed to $1.35 that you would make on average for the whole population. Clusters 2 and 3 didn't respond as well to the discount promotions as clusters 1 and 2, so it would be smarter to give them BOGO promotions to reengage them back into making "free" purchases at Starbucks as opposed to keeping them away by sending them discounted deals.

				ClusterCompletedOffers	AvgCompletedOffers	BOGO_or_Discount
Cluster_4	offer_type	difficulty	reward			
0	bogo	5	5	0.511532	0.308293	discount
		10	10	0.526596	0.217532	discount
	discount	7	3	0.602015	0.383878	discount
		10	2	0.561009	0.279512	discount
		20	5	0.340974	0.090034	discount
1	bogo	5	5	0.329114	0.308293	discount
		10	10	0.201728	0.217532	discount
	discount	7	3	0.401826	0.383878	discount
		10	2	0.262621	0.279512	discount
		20	5	0.054434	0.090034	discount
2	bogo	5	5	0.224138	0.308293	bogo
		10	10	0.149921	0.217532	bogo
	discount	7	3	0.275216	0.383878	bogo
		10	2	0.202934	0.279512	bogo
		20	5	0.048321	0.090034	discount
3	bogo	5	5	0.219118	0.308293	bogo
		10	10	0.087638	0.217532	bogo
	discount	7	3	0.306122	0.383878	bogo
		10	2	0.141766	0.279512	bogo
		20	5	0.013237	0.090034	bogo

Breakdown of how each cluster responded to each of the promotional deals. The BOGO_or_Discount field is based off a function that looks at the cluster completion

percentage and the overall population completion percentage, and if the cluster value is greater than the overall population completion percentage (with a 5% buffer), then you should offer them a discount deal. If not, you should offer them a BOGO deal to get them reengaged.

By focusing on cluster 0 and 1 as being the “discount promotional clusters”, you will have much higher expected returns on your discounts which will help make up for offering clusters 2 and 3 BOGO deals. You’re retaining your everyday customers that will continue to come back no matter what with small rewards that they will appreciate, while you focus on getting clusters 1 and 2 into the door and enjoying your product. Starbucks makes customers happier from their promotions, and they also retain more money in the process.

Reflection and Conclusion

This is a classic problem companies have to deal with, and I thought that the 4 clusters was the best approach to solving this problem. Of the different datasets that I got to work with, I wanted to keep as much untouched information for our process. This raises problems since there were people who had their listed age being 118 years old or people completing their offer after that deal expired. These would be things that I wish I would have more clarity on to have more clean data and better results.

Some improvement that could have worked more in my favor for this project would have been using a better tool like SQL databases to run some of my analysis. Had I used SQL to figure out rolling differences of people's transactions all at once instead of doing it individually would have made the client matrix run much faster. Since I didn't have that ability in python, it took roughly 15 minutes to run that section of the analysis alone. So if I messed up, it was going to take another 15 minutes to see the results again. I also wished I had more factual information to work with, like what what specific product(s) were purchased or rough living locations of where these customers were. Do people's living area affect spending habits? Do people use their BOGO or discount deals for specific products? That information would help with the clustering results and possibly make better promotional decisions.

Utilizing Starbuck's personal, transactional, and promotional data can make huge waves in their customers approval and spending habits at their stores. My clustering methods show that there are people who are low income and low spending customers who should have BOGO promotions, people who respond decent to both types of promotional deals, and high income, frequent spenders at Starbucks that should be offered more discount promotions. When affiliating these like minded people together correctly, Starbucks can make sure the right people are receiving the promotion that will elevate their

status with Starbucks while continuing to reap the benefits of smart business decisions.