Starbucks Capstone challenge report

-Vas Mandalapu, September 2019

Introduction

The project is based on the simulated data provided by Starbucks that mimics customer behavior on the Starbucks rewards mobile app. Periodically, 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). However, not all users receive the same offer, and that is the challenge to solve with this data set.

The simulated data has 17000 customers demographic data, details related to 10 different offers and transcript data that covers 3 offer related events and transactions over 29 days.

Problem Statement

The objective of this capstone project is to discover customer segments that share common traits and find which customer segment is going to respond best to which offer type. To solve this problem need to

- gain an understanding of what types of customer characteristics and demographics are there.
- how different customer segments respond to different kinds of offers

Model

To discover the customer segments K-Means algorithm is used to cluster the customers based on similarity. K-Means use distance-based measurements to determine the similarity between data points.

Benchmark Model

Hierarchical clustering (with ward linkage) is used as benchmark model to evaluate how the similar the groups. For hierarchical cluster, the silhouette score is used to evaluate the clustering.

Evaluation Metric

Unlike supervised learning, where we have ground truth to evaluate the models performance, clustering analysis doesn't have a solid evaluation metric. Instead we rely on intuition and how well the cluster are separated.

The metrics that will be considered for this project are

- Elbow method
- Silhouette analysis

Elbow Method

Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow.

Silhouette Analysis

Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:

Compute the average distance from all data points in the same cluster (ai).

Compute the average distance from all data points in the closest cluster (bi).

Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

The coefficient take values in the interval [-1, 1].
If it is 0 –> the sample is very close to the neighboring clusters.
If it is 1 –> the sample is far away from the neighboring clusters.
If it is -1 –> the sample is assigned to the wrong clusters.
Therefore, we want the coefficients to be as big as possible and close to 1 to have a good cluster.

Solution Approach

To solve the problem statement, the solution is split into following 5 steps

1. Data exploration and cleaning
2. Data preparation and feature engineering
3. Dimension reduction and Clustering
4. Cluster analysis to find distinct traits of each cluster
5. Conclusions

Data Exploration

The data is contained in three files:

- portfolio.json — containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json — demographic data for each customer
- transcript.json — records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

Profile

Let's start data exploration with customer demographic data in profile dataset. This dataset has 17000 unique customers. Here is summary statistics

	count	mean	std	min	25%	50%	75%	max
age	14825	54.39	17.38	18	42.0	55.0	66.0	101.0
income	14825	65405	21598	30000.	49000	64000	80000	120000
days_as_ member	17000	517.44	411.22	0.0	208.0	358.0	791.0	1823.0

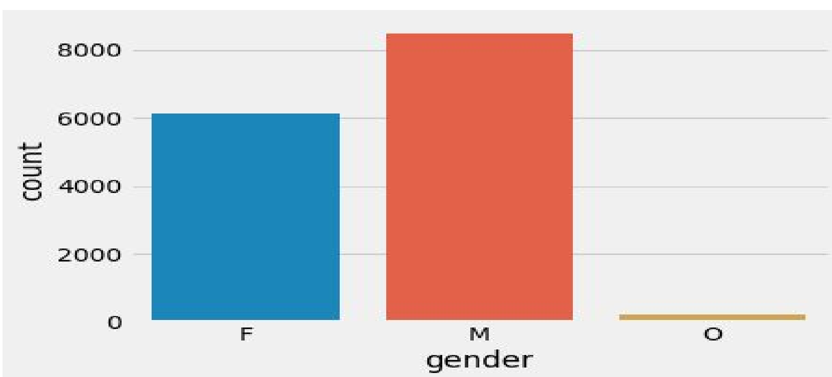
This dataset has 2175 customers with missing income, gender and age set to 118.

	column_name	missing_count	missing_ratio
	gender	2175	0.127941
	income	2175	0.127941

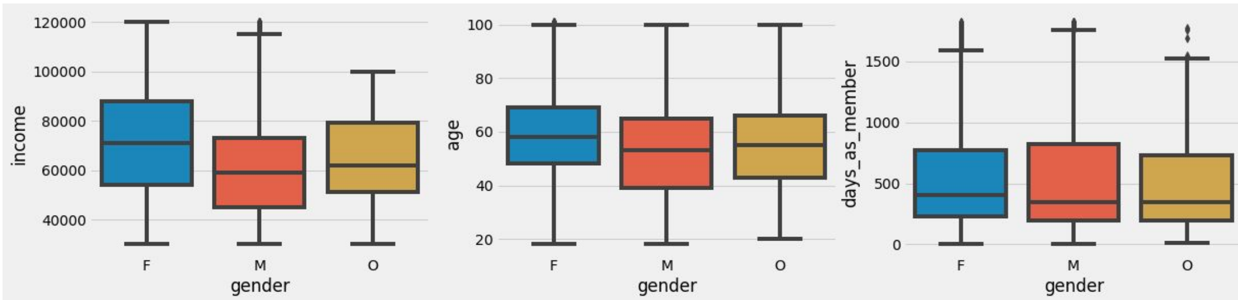
As part of data cleaning , for missing customers the age is set as null and became_member_on column is converted to number of days as member.

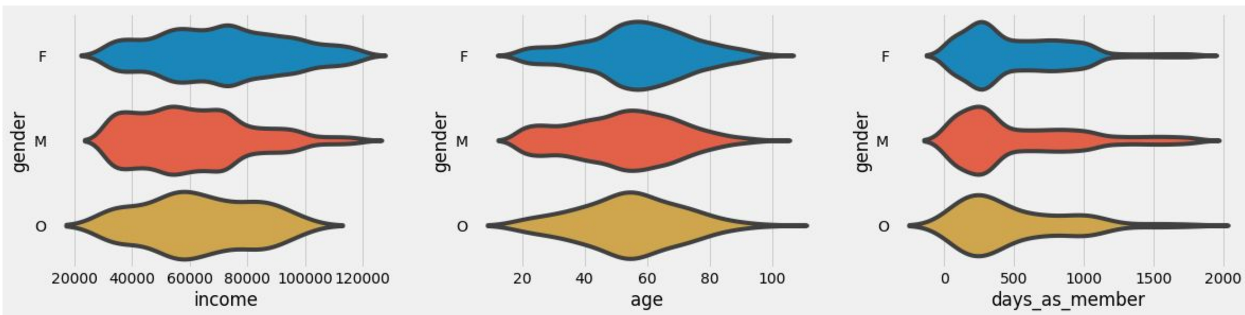
Here are summary statistics of customer demographics by gender

Histogram of gender , with female, male and Other



Customer demographics are analyzed by gender using box plots and violin plots to understand the distribution



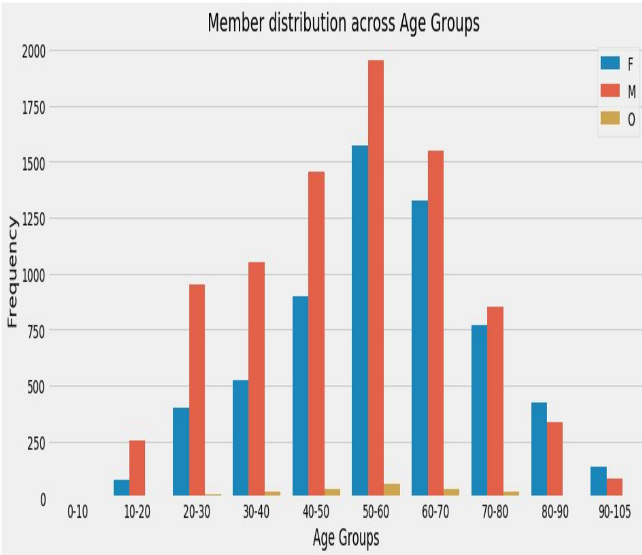


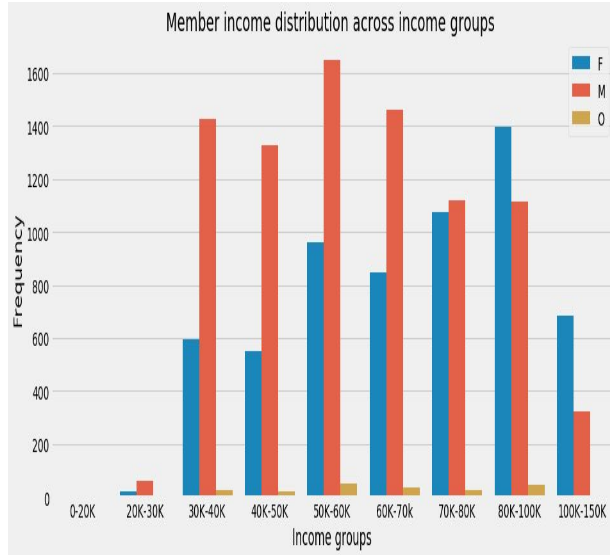
From above two plots we can conclude that

- Females have higher median income than males and also the median age of female is older than median age of male.
- The income distribution is more uniform for females , for males income distribution skewed to left
- The age distribution is more in center for females and for males the age is skewed to left.
- The days as member distribution is same across different genders.

For the following, we will decompose 'age' and 'income' in ranges as follow:

- ten categories for the age with 10 year intervals
- Nine categories for the income: [0-20K], [20K-30K], [30K-40K], [40K-50K], [50K-60k], [60K-70K], [70K-80K] , [80K-100K]and [100K-150K].





From above plots we see that females are better represented in income bins over 70K and age bins over 60.

Portfolio

Portfolio dataset contains 10 offer related information. As part of preprocessing channels columns is one hot encoded , additional column called duration hours (duration in days is converted to hours) , relative difficulty(difficulty/duration) are added.

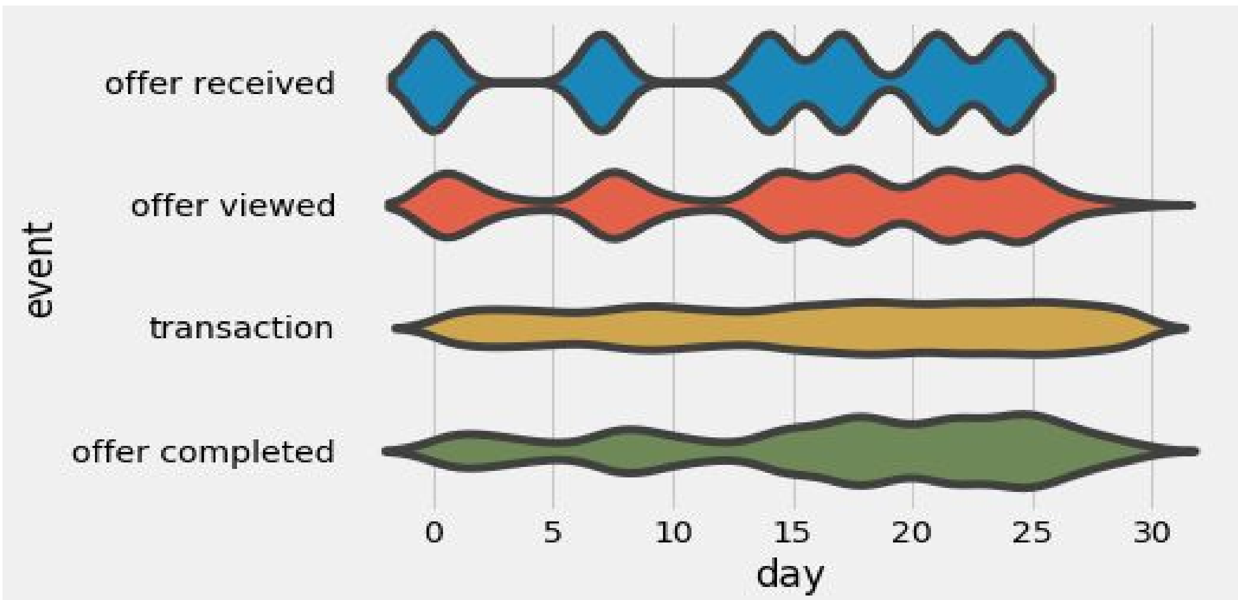
	difficulty	duration	id	offer_type	reward	web	email	mobile	social	duration_hours	relative_difficulty
0	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10	0	1	1	1	168	1.428571
1	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10	1	1	1	1	120	2.000000
3	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5	1	1	1	0	168	0.714286
8	5	5	f19421c1d4aa40978ebb69ca19b0e20d	bogo	5	1	1	1	1	120	1.000000
4	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5	1	1	0	0	240	2.000000
5	7	7	2298d6c36e964ae4a3e7e9706d1fb8c2	discount	3	1	1	1	1	168	1.000000
6	10	10	fafdcd668e3743c1bb461111dcafc2a4	discount	2	1	1	1	1	240	1.000000
9	10	7	2906b810c7d4411798c6938adc9daaa5	discount	2	1	1	1	0	168	1.428571
2	0	4	3f207df678b143eea3cee63160fa8bed	informational	0	1	1	1	0	96	0.000000
7	0	3	5a8bc65990b245e5a138643cd4eb9837	informational	0	0	1	1	1	72	0.000000

Transcript

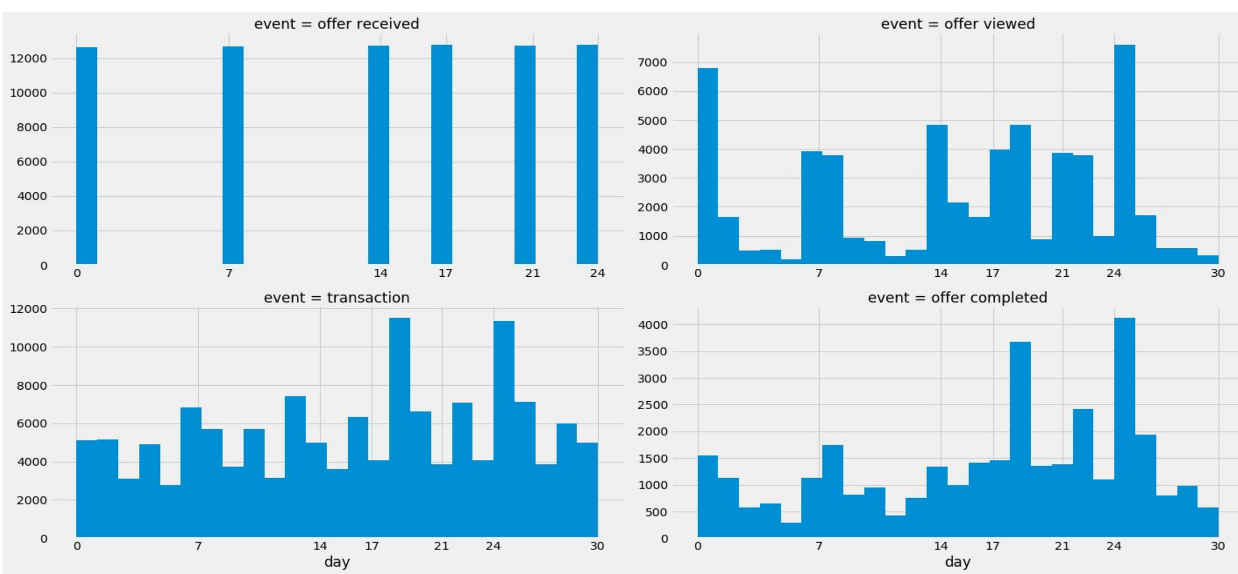
Transcript dataset is that needs extensive cleaning and elaborative preparation steps.

Let's explore the dataset before starting cleanse and preparation steps.

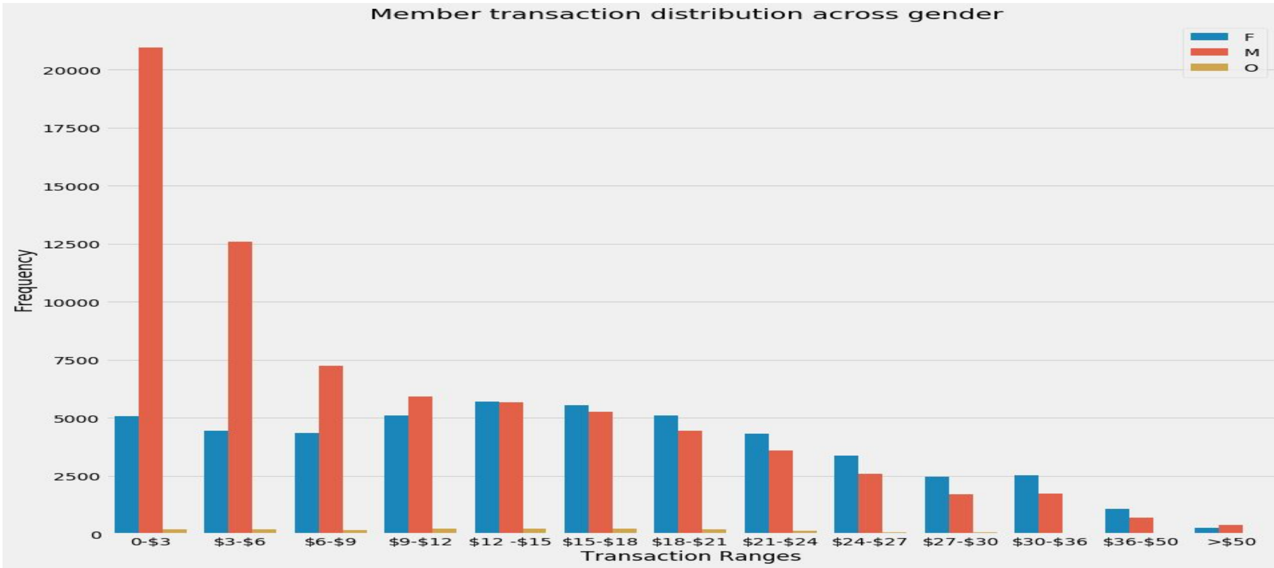
The following plot, showing the distribution of roughly 317,000 'events' over the 30 trial day period, shows that, release of offers is followed by offer viewings , a raise in transactions and eventually, if a certain spending threshold is met, with an offer completion that leads to a reward



Another representation of events over 30 days



Also will decompose transaction amounts into ranges and observe spending pattern by gender



Data Cleaning and Feature Engineering

The challenge with transcript data is that the transactions a customer makes are not linked to any offers he/she has received. And even when an offer is sent to a customer, this doesn't mean he/she has viewed it and hence is influenced by it. So we have to ensure that we only count the viewed offer completions which occur within the defined period of validity and result in a reward.

Transcript dataset is provided in below format

	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
15561	offer viewed	78afa995795e4d85b5d9ceeca43f5fef	6	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
47582	transaction	78afa995795e4d85b5d9ceeca43f5fef	132	{'amount': 19.89}
47583	offer completed	78afa995795e4d85b5d9ceeca43f5fef	132	{'offer_id': '9b98b8c7a33c4b65b9aebfe6a799e6d9...
49502	transaction	78afa995795e4d85b5d9ceeca43f5fef	144	{'amount': 17.78}
53176	offer received	78afa995795e4d85b5d9ceeca43f5fef	168	{'offer id': '5a8bc65990b245e5a138643cd4eb9837'}
85291	offer viewed	78afa995795e4d85b5d9ceeca43f5fef	216	{'offer id': '5a8bc65990b245e5a138643cd4eb9837'}
87134	transaction	78afa995795e4d85b5d9ceeca43f5fef	222	{'amount': 19.67}
92104	transaction	78afa995795e4d85b5d9ceeca43f5fef	240	{'amount': 29.72}
141566	transaction	78afa995795e4d85b5d9ceeca43f5fef	378	{'amount': 23.93}

Here are the data preparation and cleaning steps

1. The value field is decoded to populate offer id and amount depending on whether it's offer event or transaction event. The columns after this steps will be ['event', 'person', 'time', 'amount','offer_id']
2. As we are interested in customer response to offer, we will consider only customer that received offers. So we merge transcript data set with Profile and Portfolio datasets

3. From merged dataset, we pick all event of each customer and iterate thru below steps for each customer
    a. A customer event dataset is split into 4 based on event. This data sets are offer_received , offer_viewed , offer_completed and transaction
    b. Offer received dataset is compared against offer completed dataset of that customer to mark offers that are completed and any reward earned
    c. Now above data is compared against offer viewed to mark whether offer is viewed or not. Also whether an offer is success (viewed and completed) or not
    d. Dataset from above step is now compared against transaction data to check if the transaction is within offer duration period, then that transaction spend is tied to that offer.

    Note: Iterative approach to process ~17000 customers took more than an hour on a personal laptop

Features Engineering

Following features are derived to better understand the customer response to offer

Total_Spent : This total amount spent by customer during the 30 day period

Total_tranx : This total number of transaction by customer during the 30 day period

Total_spent_between_promos: This total amount spent by customer between promotions

Total_tranx_between_promos: This total number of transactions by customer between promotions

Total_participated_duration: Number of hours between viewed to completed (for completed offers). For uncompleted offers , number of hours between view to end of offer duration

Tranx_from_view_to_complete: Number of transactions between viewed to completed (for completed offers). For uncompleted offers, number of transactions between viewed to end of offer duration

Spent_from_view_to_complete: Amount spent between viewed to completed (for completed offers). For uncompleted offers, amount spent between viewed to end of offer duration

np_tranx_in_duration: Number of transactions while not participating in completing an offer. Number of transactions before viewed and number of transaction after completing the offer within the duration

np_spent_in_duration: Amount spent while not participating in completing an offer. Amount spent before viewed and amount spent after completing the offer within the duration

spend_per_tranx_in_promo: This provides average amount spent per transaction during promotion (view to complete). =*spent_from_view_to_complete/ tranx_from_view_to_complete*

spend_per_tranx_in_np: This provides average amount spent per transaction during non-promotion period

= *np_spent_in_duration/np_tranx_in_duration*

rate_of_completion : This indicates how quickly someone has completed the offer. For uncompleted offer this is 0.

For completed offers , the formula is =

(offer_duration – participation_hours)/offer_duration

perc_spend_in_promo: This indicates what percentage of spend happened within promotion period (from viewed to completed state) of overall spend during offer duration

perc_tranx_in_promo: This indicates percentage of transactions that happened while completing offer compared to overall transactions during offer duration

per_hour_spend_in_promo : This indicates rate of spend during promotion

spent_from_viewed_to_complete/participation_duration

per_hour_spend_in_np : This indicates rate of spend during non promotion i.e before offer is viewed and after offer is completed

np_spent_in_duration/(offer_duration-participation_duration)

per_hour_tranx_in_promo : This indicates rate of transactions during promotion

tranx_from_viewed_to_complete/participation_duration

per_hour_tranx_in_np : This indicates rate of transactions while not participating in completing an offer i.e before offer is viewed and after offer is completed

np_tranx_in_duration/(offer_duration-participation_duration)

Data Insights

From simulated data, it appears social channel is the one of the effective way to reach customers, when social channel is not used the view rate dropped below 47%

portfolio_id	email_view_rate	web_view_rate	social_view_rate	mobile_view_rate
bogo-10.0-5.0-10.0	0.894245	0.894245	0.894245	0.894245
bogo-10.0-7.0-10.0	0.795900	0.000000	0.795900	0.795900
bogo-5.0-5.0-5.0	0.874917	0.874917	0.874917	0.874917
bogo-5.0-7.0-5.0	0.465156	0.465156	0.000000	0.465156
discount-10.0-10.0-2.0	0.899171	0.899171	0.899171	0.899171
discount-10.0-7.0-2.0	0.469078	0.469078	0.000000	0.469078
discount-20.0-10.0-5.0	0.321857	0.321857	0.000000	0.000000
discount-7.0-7.0-3.0	0.883076	0.883076	0.883076	0.883076
informational-0.0-3.0-0.0	0.807561	0.000000	0.807561	0.807561
informational-0.0-4.0-0.0	0.500459	0.500459	0.000000	0.500459

For below figure, we can conclude that BOGO offer completion depends more on difficulty and discount offer completion depends on relative difficulty i.e. depends on duration as well as difficulty.

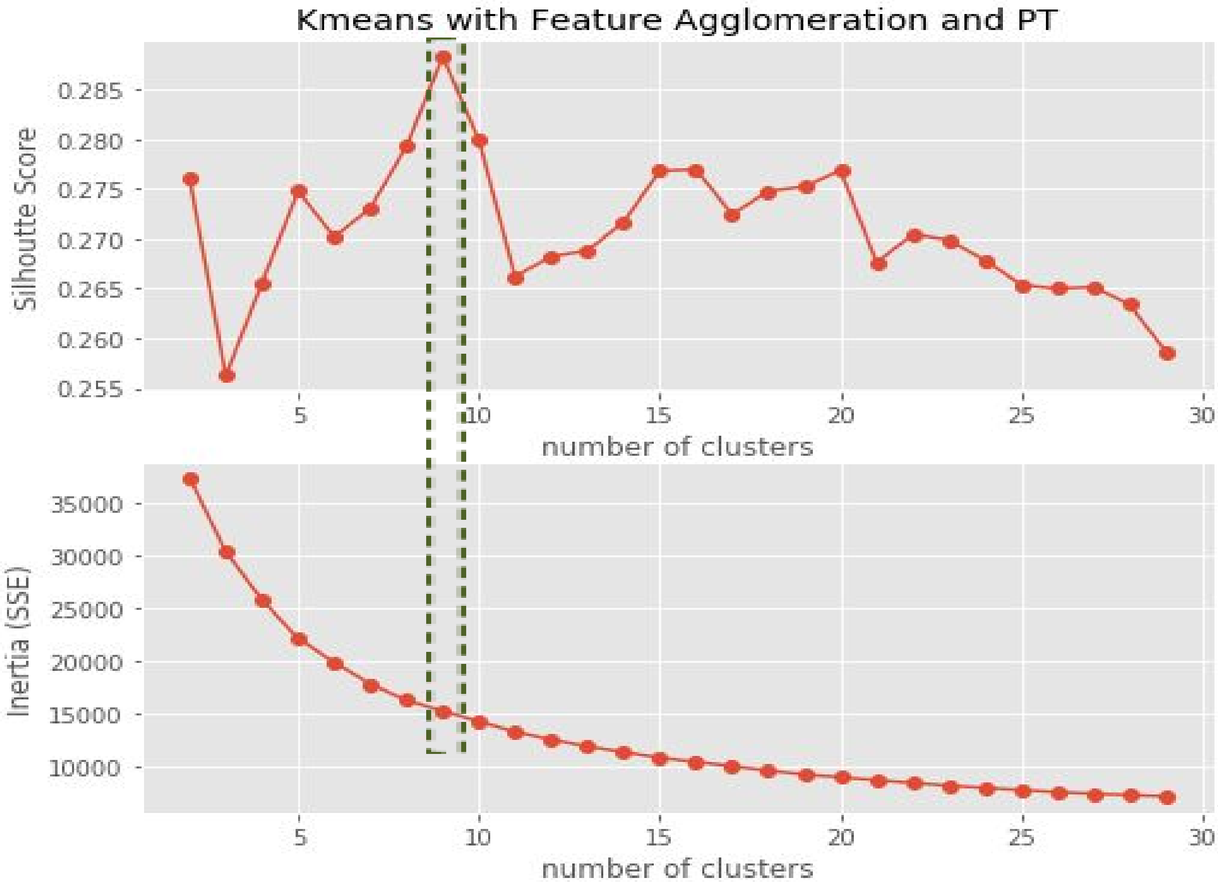
portfolio_id	duration	difficulty	relative_difficulty	recieved	viewed	completed	success	invalid_complete
bogo-10.0-5.0-10.0	5.0	10.0	2.000000	7593	6790	3331	2769	562.0
bogo-10.0-7.0-10.0	7.0	10.0	1.428571	7658	6095	3688	2630	1058.0
bogo-5.0-5.0-5.0	5.0	5.0	1.000000	7571	6624	4296	3546	750.0
bogo-5.0-7.0-5.0	7.0	5.0	0.714286	7677	3571	4354	2135	2219.0
discount-10.0-10.0-2.0	10.0	10.0	1.000000	7597	6831	5332	4678	654.0
discount-10.0-7.0-2.0	7.0	10.0	1.428571	7632	3580	4025	2149	1876.0
discount-20.0-10.0-5.0	10.0	20.0	2.000000	7668	2468	3440	1360	2080.0
discount-7.0-7.0-3.0	7.0	7.0	1.000000	7646	6752	5165	4410	755.0
informational-0.0-3.0-0.0	3.0	0.0	0.000000	7618	6152	0	0	0.0
informational-0.0-4.0-0.0	4.0	0.0	0.000000	7617	3812	0	0	0.0

There is bounce in spending, when offer has good view rate like BOGO (bogo-10-7-10) offer with 79% view rate, average spend during 7 day offer duration is $21.44, while BOGO (bogo-5-7-5) offer and discount (discount-10-7-2) offer with 46% view rate , the average spending during the 7 day promotion period is $12.52 and $13.18 respectively

portfolio_id	viewed	spent_from_view_to_complete	spent_in_duration	np_spent_in_duration
bogo-10.0-5.0-10.0	0.894245	10.425849	18.783910	3.345712
bogo-10.0-7.0-10.0	0.795900	10.628515	21.445226	7.083261
bogo-5.0-5.0-5.0	0.874917	9.129274	18.050930	3.859864
bogo-5.0-7.0-5.0	0.465156	5.301847	12.521081	10.866613
discount-10.0-10.0-2.0	0.899171	12.470186	36.351974	6.134776
discount-10.0-7.0-2.0	0.469078	5.844696	13.186424	10.616779
discount-20.0-10.0-5.0	0.321857	5.722139	12.236493	17.591352
discount-7.0-7.0-3.0	0.883076	10.281477	25.215086	5.380970
informational-0.0-3.0-0.0	0.807561	9.119734	10.695876	1.618048
informational-0.0-4.0-0.0	0.500459	7.290394	8.741132	4.407607

Dimension Reduction and Clustering

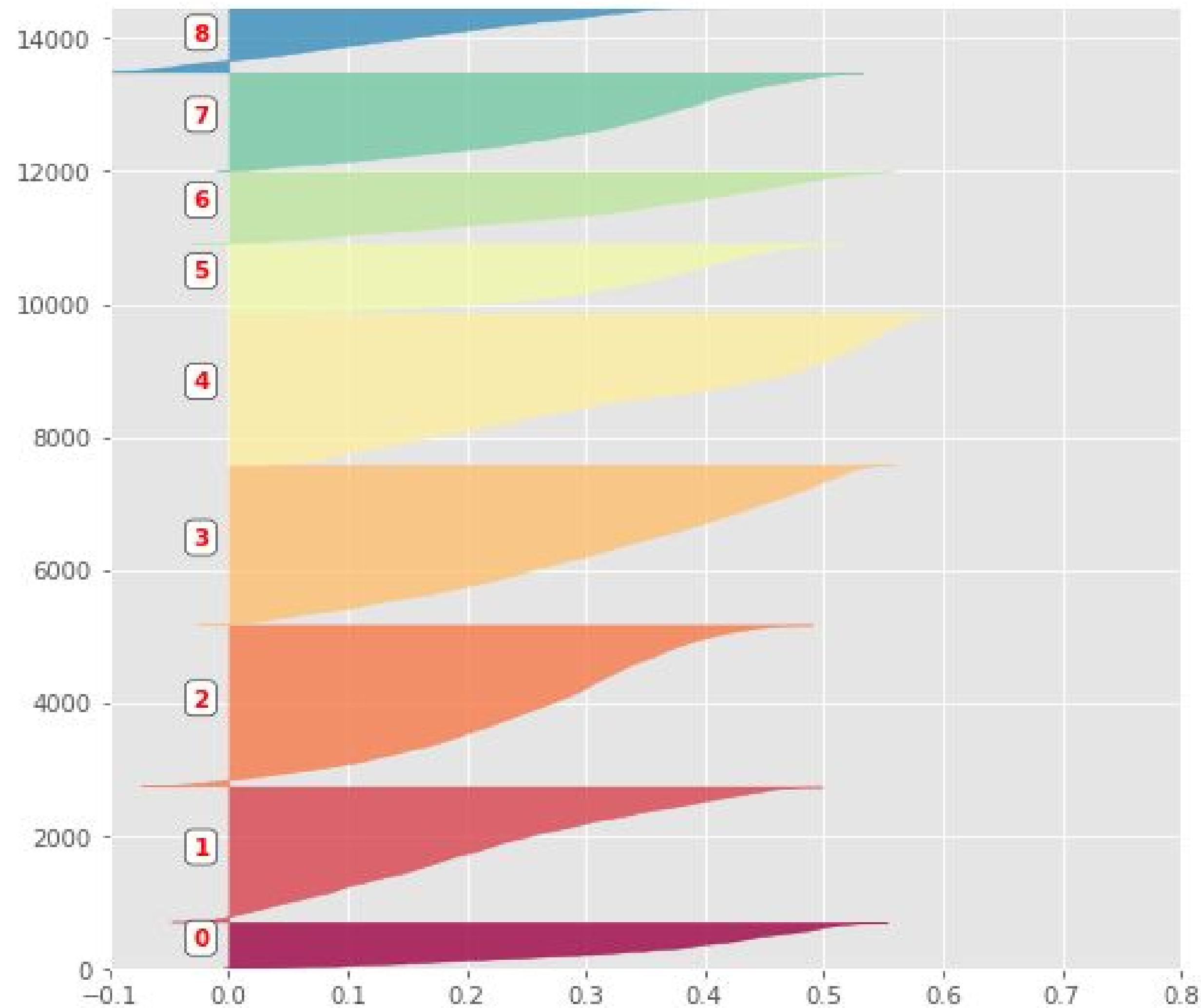
Normalization and Scaling: Prepared dataset is skewed, so explored scaling techniques like standardization, normalization and power transformer with Yeo-Johnson method. Followed by Principal Component Analysis (pca) and Feature Agglomeration dimension reduction techniques applied on scaled data. Out of all combinations, power transformer with agglomeration method yielded best silhouette score when K-Means Clustering is applied



For choosing the model, considered the silhouette score and inertia, out of all models considered the chosen model has best silhouette score (~0.29) and lowest inertia (15000). For benchmarking, created hierarchical model with "ward" linkage, the best silhouette score that we got is 0.23 for 9 clusters

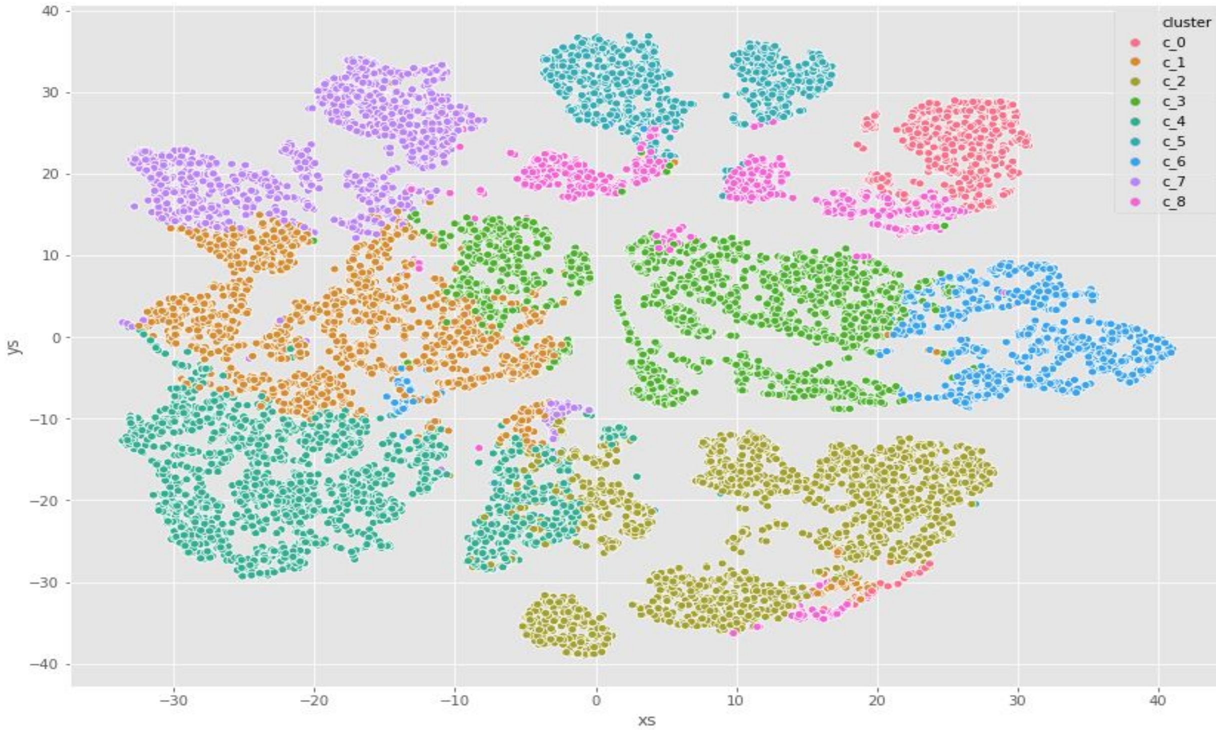
Silhouette intra-cluster score

In order to have an insight on the quality of the intra cluster classification, the silhouette scores of each element of the different clusters are represented below



From above plot, we can conclude cluster 0,4, 5 and 6 have best intra-cluster silhouette score. And cluster 8 has worst intra cluster score.

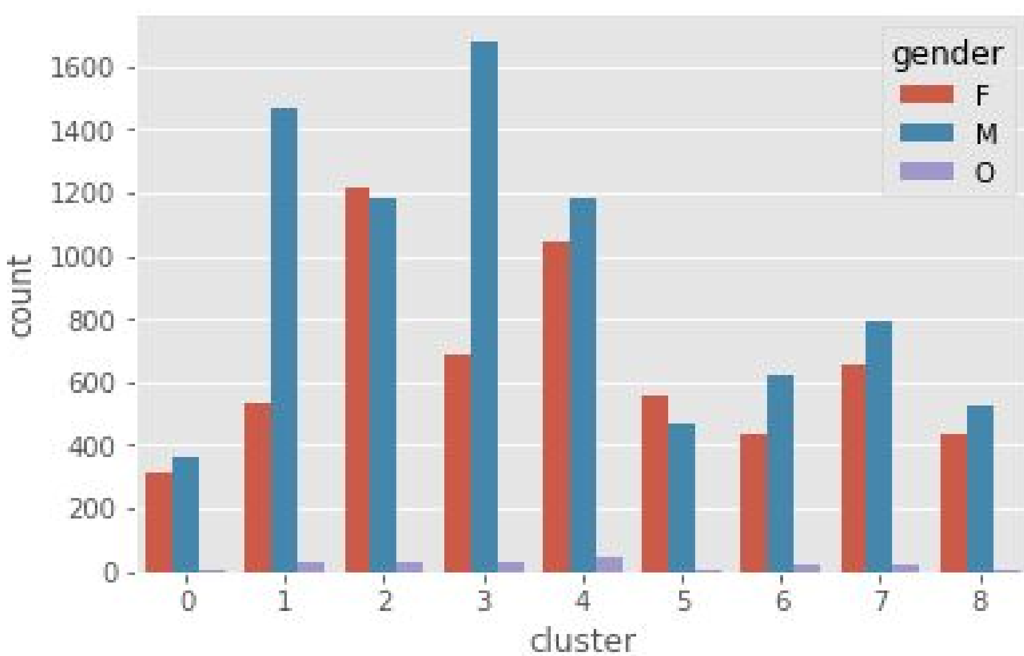
The agglomeration reduced dataset is projected on 2 dimensional TSNE map and the clusters visualized using scatter plot, to see cluster separation



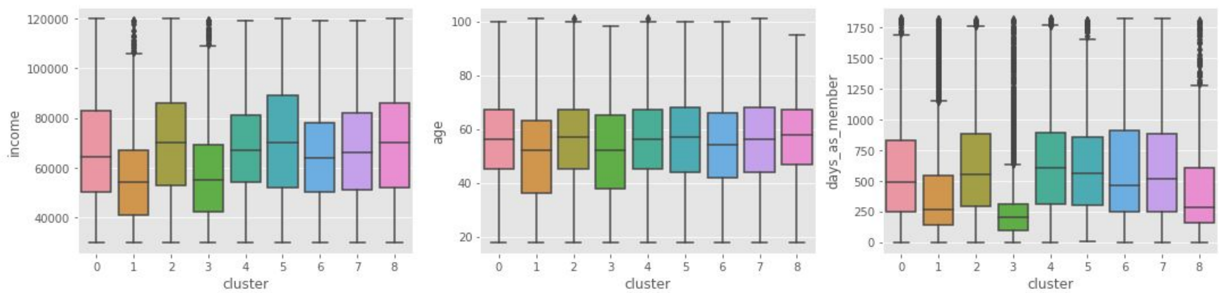
Above plot shows cluster 0,2,3,6 and 7 are more concentrated, while cluster 4 and 8 are more scattered.

Segment Analysis

Gender distribution across the cluster segments is shown below; cluster 1 and cluster 3 are disproportionately represented by Male. While cluster 2 and 5 have more Female representation.



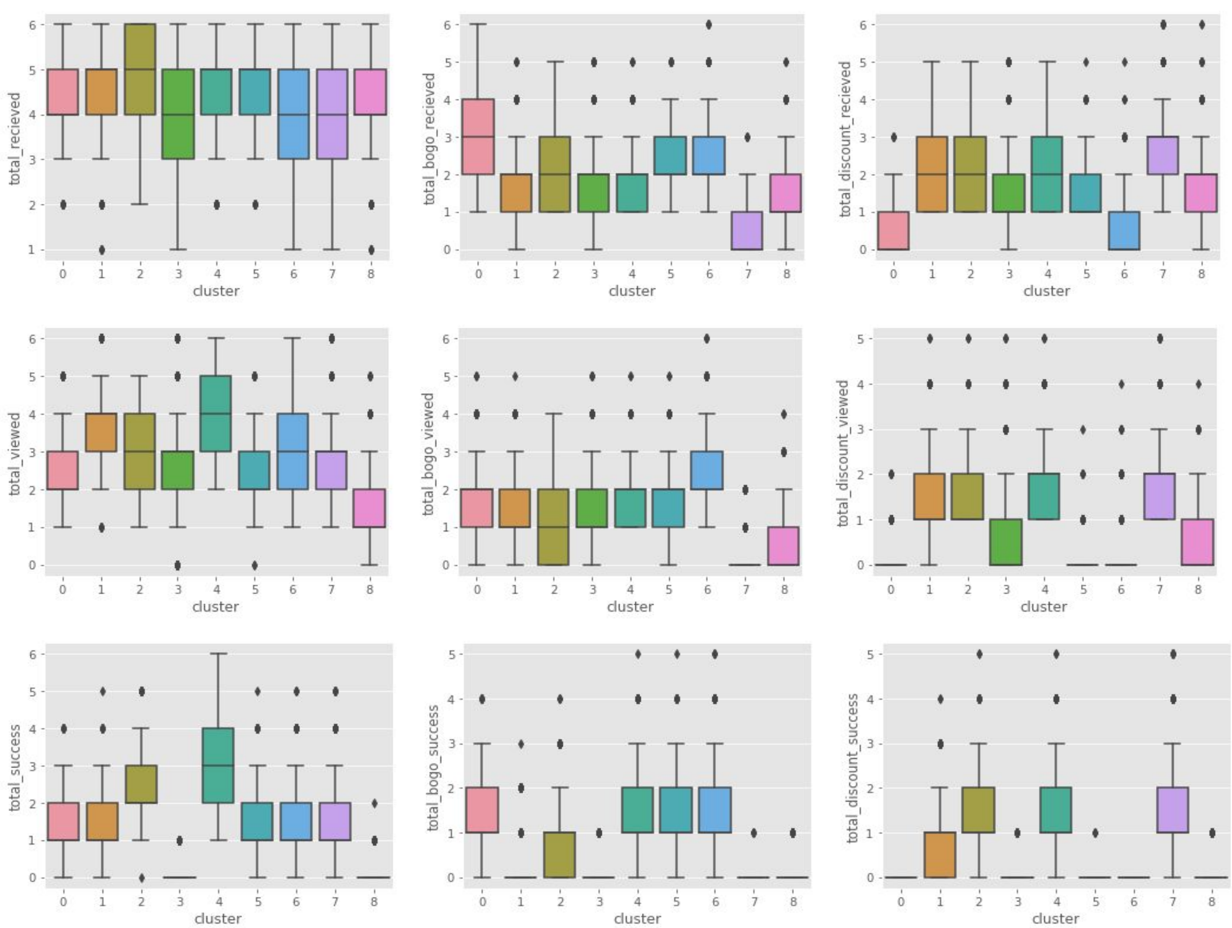
In below plot, the IQR of demographic features are represented across the segments



We observe following properties from the above inter quartile ranges

- Cluster 2 and 5 have highest median age and income.
- Cluster 1 and 3 has lowest median income and age.
- Cluster 3 is heavily represented by newest members

In below plots, the IQR of offers received, viewed and offer success across cluster segments are represented

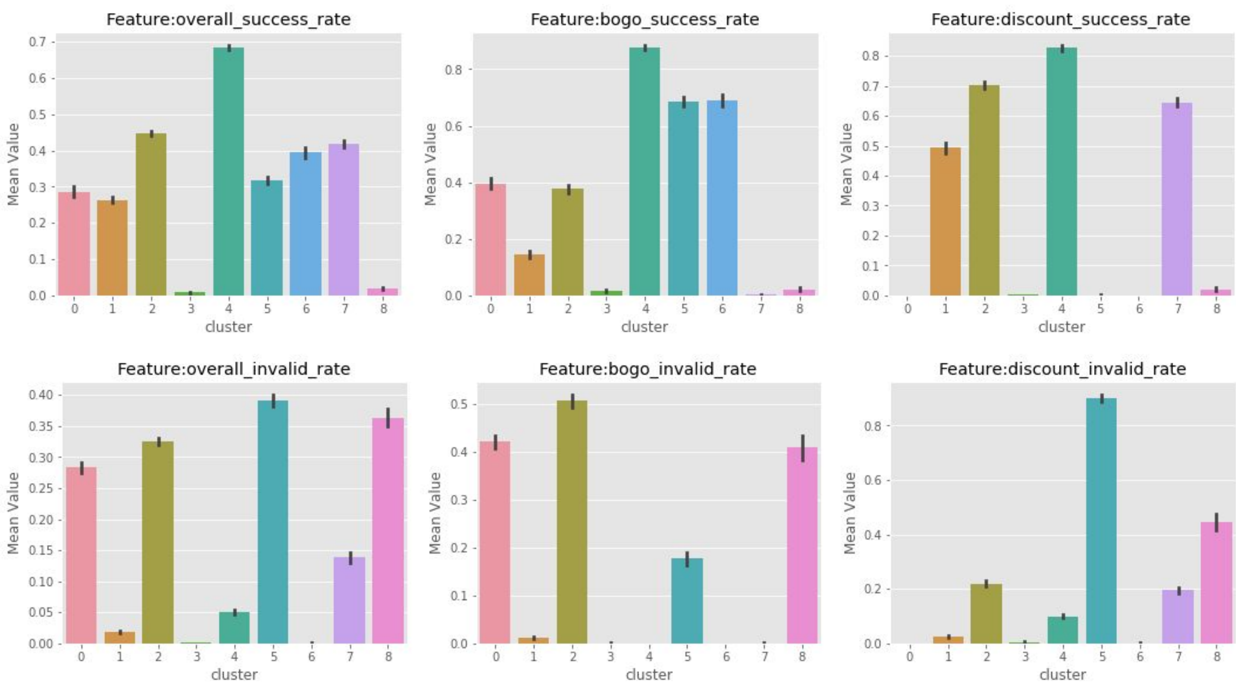


We observe from above IQR representation

- Clusters 0 and 6 has more BOGO success , partly as this group got more BOGO offers and almost no discount offers.

- Cluster 7 is reverse of cluster 0 and 6 i.e., this group has more success with discount offers as this group received more discount offers and almost no BOGO offer.
- Cluster 3 and 8 have worst success rate. Cluster 8 also has worst view rate
- Cluster 1 has prefers discount offers, as this group equal view rate for both offers.
- Cluster 2 received healthy balance of BOGO and discount offers. But cluster 2 has better success with discount.
- Cluster 5 received good mix of BOGO and discounts offers, responded to BOGO better and almost zero successful discount offers partly because of this group's worst view rate for discount offers.
- Cluster 4 has equal success across both offer types and also most engaged group

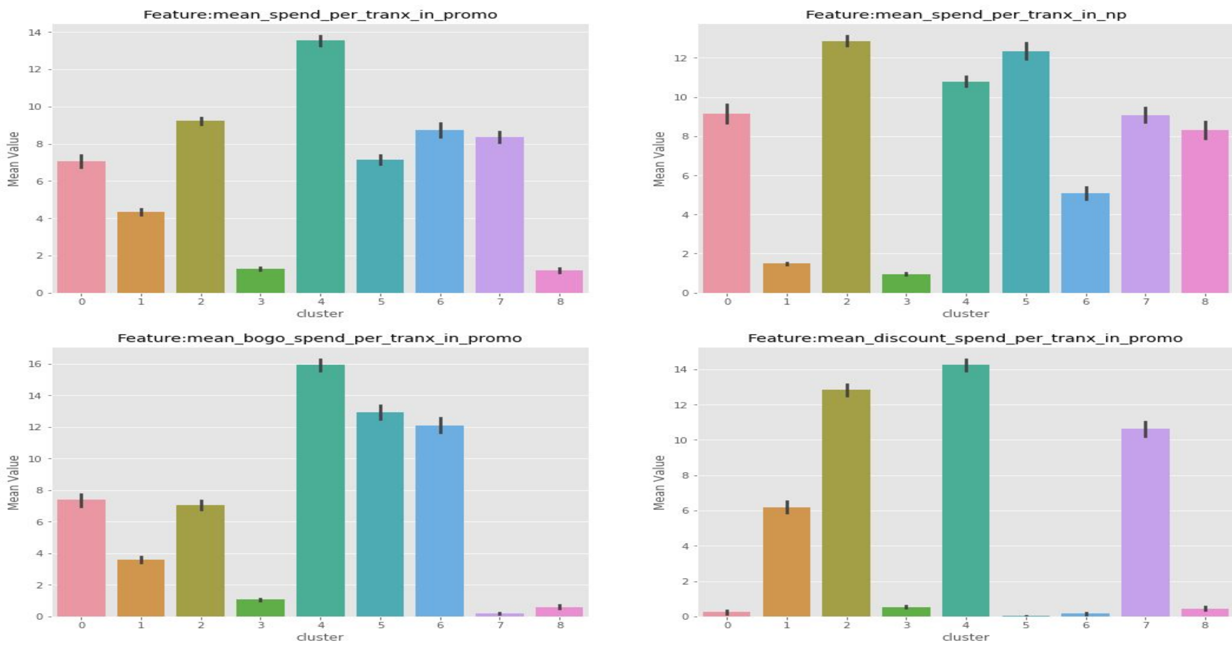
In below plot we represent mean success rate (number of successful offers/number of offers received) and invalid completed rate (number of invalid complete/number of offers received) across the cluster segments



Following observations can be made from above plot

- Cluster 0 and 2 has almost equal percent of successfully completed offers versus invalid completed offers.
- Cluster 1 responds to discount offers positively (mean conversion rate of 50%) and very low rate of invalid offer completions
- Cluster 3 has lowest success rate as well as lowest invalid rate. This group most inactive group.
- Cluster 4 is most engaged customer who respond well to both types of offers and very low rate of invalid offer completions
- Cluster 5 has most invalid offer completed rate and most of the invalid completed offer are discounts.
- Cluster 6 responds well to BOGO offers and has zero invalid offer completion rate.
- Cluster 8 seems like regular coffee buyer who don't view the offers and because of their spending they earned the rewards.

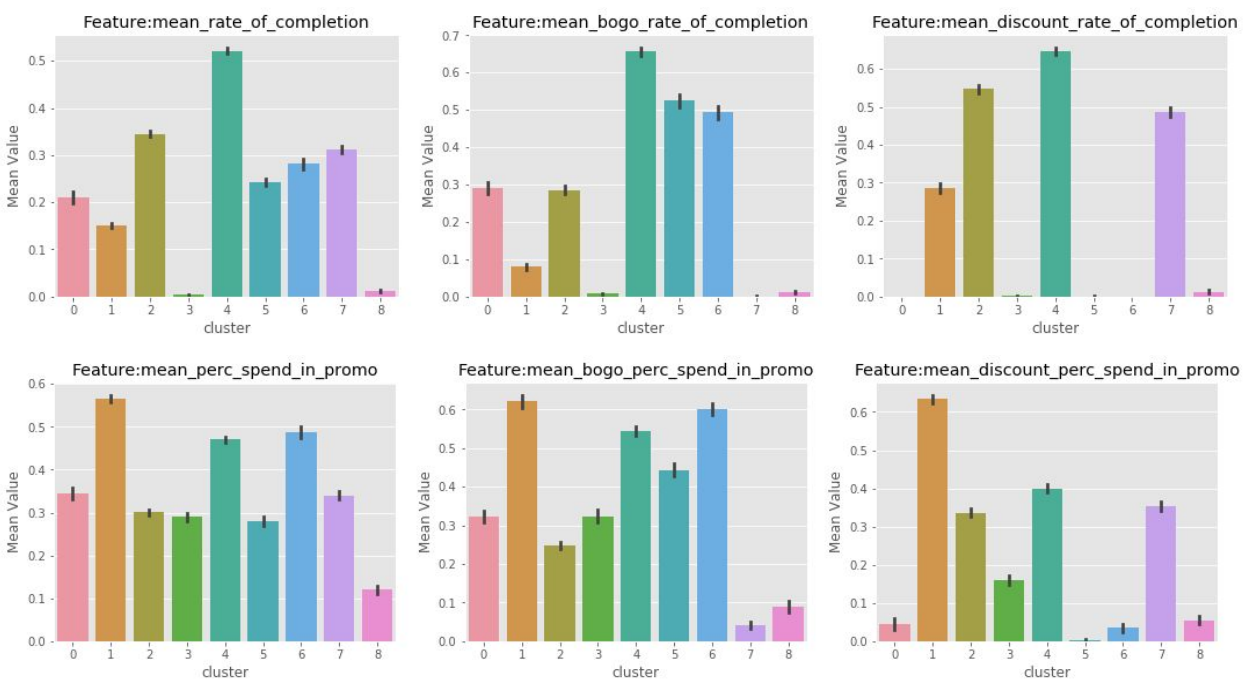
Below we examine mean amount spent per transaction during valid promotion completion effort and number of transactions that are not influenced by the promotion.

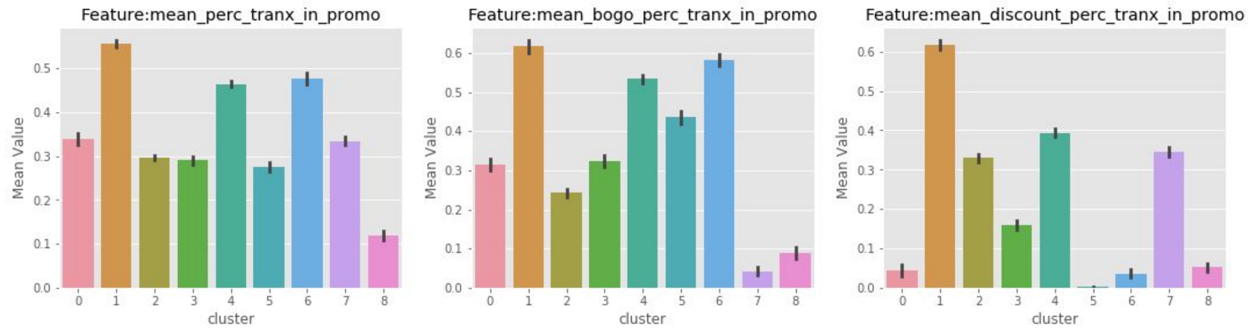


Following observations can be made from above bar chart

- For cluster 0, 2 ,3,5 and 8 there is no influence of promotion on amount spent per transaction.
- For cluster 1,4,6 and 7 there is respond well to promotions with increase in mean amount spent per transaction during promotion period

Let's examine mean rate of completion, mean percentage of spending and number of transactions during promotional period thru below plots

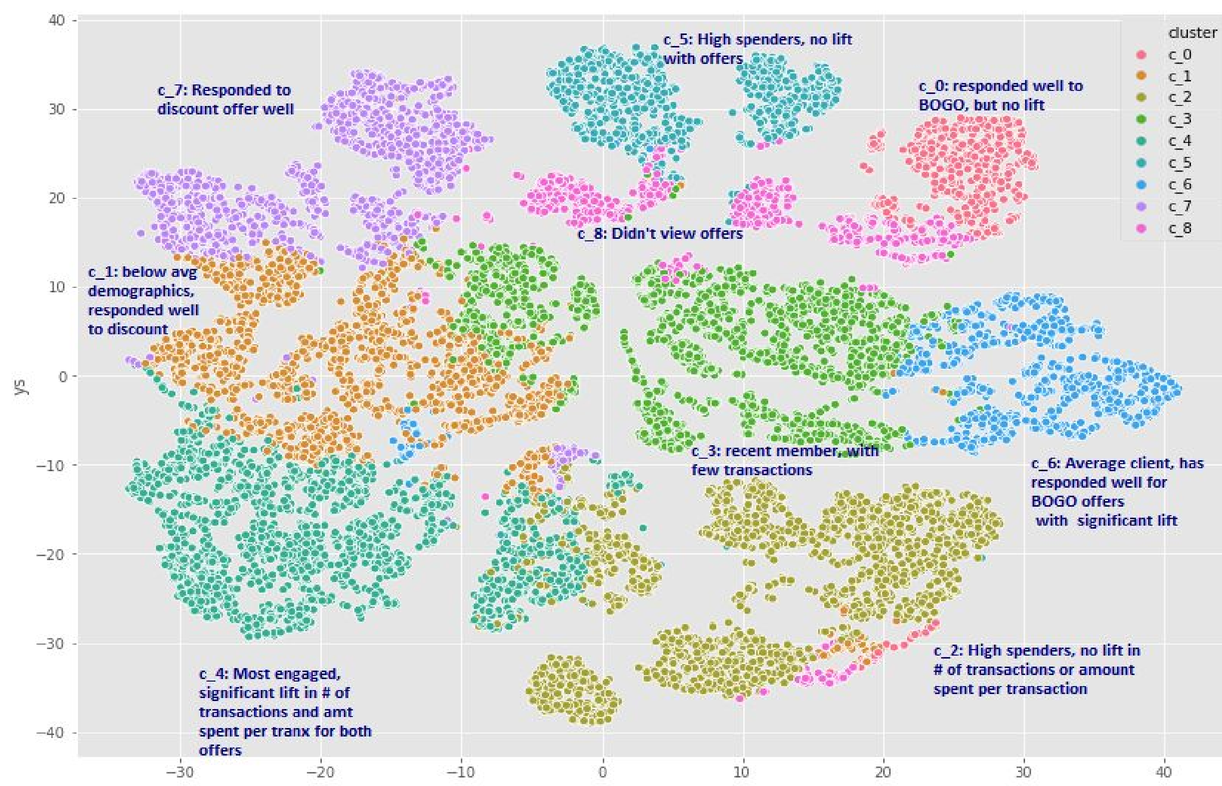




From above three plots, following observations can be made

- Cluster 4 has best mean rate of completion that means this group can complete takes very less time from viewing to completing an offer.
- Cluster 1 and 6 has most of the spending as well as most of transaction during promotional period.

Summary of Customer cluster segments



Cluster Segment 0 (4.8% of population with demographic data): This group mostly received BOGO offers and successfully completed 40% of those offers. This group also 40% invalid completed offers. An offer has no significant influence on client's spending in this group..

Cluster Segment 1 (14.1%): The core demographics of this group are young customers with below mean income and mostly male. This group has most number of transactions with least amount per transaction. This group responds more favorably to discount offer with increase in amount spent per transaction and number of transactions.

Cluster Segment 2 (16.8%): This group has more number of females than male, higher income than mean income of the population and also very active customers with high per transaction spend. And mostly likely view their offer thru social media and most of the offers are not challenging enough for them as this group has significant invalid offers completed. An offer has no significant influence on their spending.

Cluster Segment 3 (16.7%) : The core demographics of this group are slightly younger , mostly like a male , lower income than mean income of population, recently became a member , occasional coffee buyer and didn't respond to offers much.

Cluster Segment 4(15.8%): This group responds to promotional offers well and there is a boost in spending and number of transaction during promotional period. This group is also much more engaged and viewed most of the offers received.

Cluster Segment 5 (7.2%): This group is similar to cluster segment 2, major difference is this group didn't view the discount offers they received but got invalid rewards for discount offers. Like cluster 3 this group spending is not significantly influenced by offers.

Cluster Segment 6 (7.5%): This group's mean income and mean age are close to the mean income and age of the population. The client in this group also more engaged as they have second highest offer viewing rate. The clients in this group mostly received BOGO offers and responded to BOGO offers well with increased amount spent per transaction and number of transactions while participating in completing the offer. The clients in this group didn't have any invalid rewards.

Cluster Segment 7 (10.3%): This group's mean income and mean age are slightly more than mean income and age of the population. This group mostly received discount offers and responded to discount offers with slight increase in amount spent per transaction while participating in completing the offer. The clients in this group have some invalid rewards.

Cluster Segment 8 (6.8%): The clients in this group have mean income and mean age above the mean income and age of the population. One of defining traits is the clients in this group have very low viewing rate and the mean amount spent per transaction for this group is approximately $8.

Recommendations

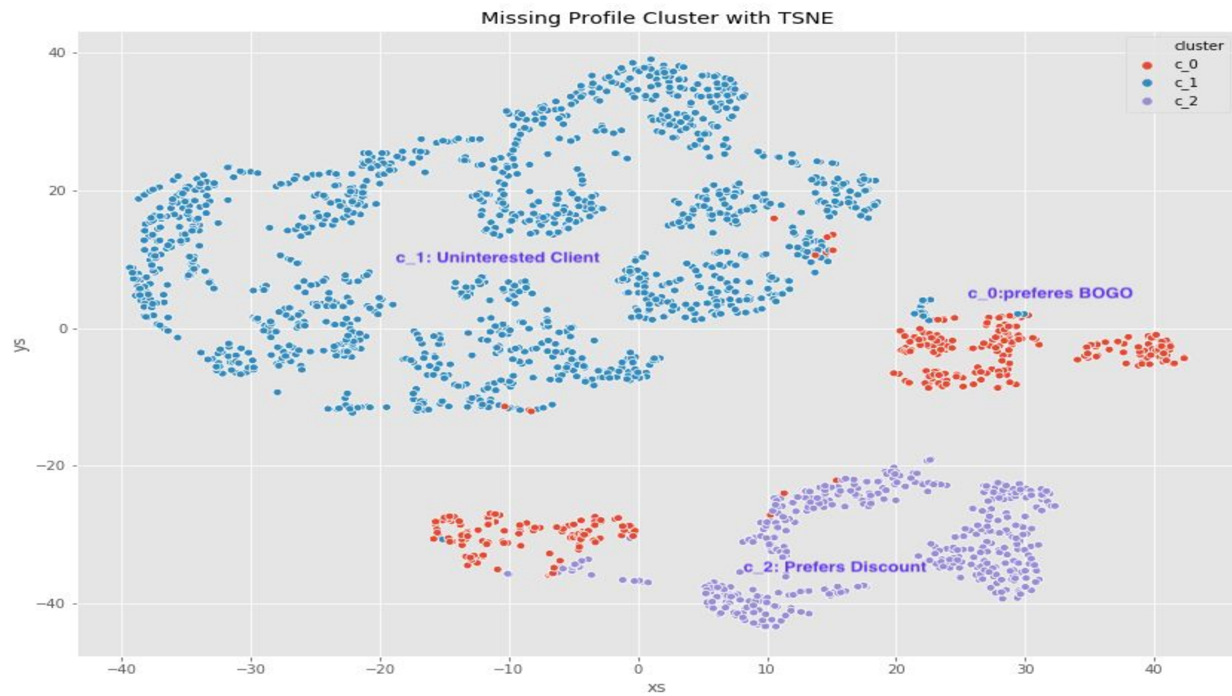
For Starbucks rewards program to be profitable, the program should incremental increase in number transactions or amount spent per transaction during promotional offer than when there are no promotional offers. Based on the cluster analysis, we offer following recommendations to personalize the targeting the offers for different clusters segments.

- For Cluster 0, 2 and 5, the clients in this group should be targeted with offers that are more difficult and with less offer duration. This clients should be targeted more for discounted offers and informational offers as the cost to company is less compared to BOGO offer.
- For Cluster 1, Starbucks should design offers to increase the amount spent per transaction. As this group got most transactions in promotion period, an increase in amount spent per transaction will be significant.
- For cluster 3, as this group have most of recent members with least number of transactions. Starbucks need to find new kind of offers to increase number of transactions.
- For cluster segment 6 and 7, A/B testing need to be done to find if this clusters have any preference to one type offer, in the simulated data this two groups are exposed mostly to one of the offer type and they responded well to those offers. Based on this it is hasty to label this groups have preference for BOGO or discount
- For cluster segment 8, Starbucks need to find new channels (like printing promo on receipt etc) to reach out to these clients and engage them as they have above average spending.

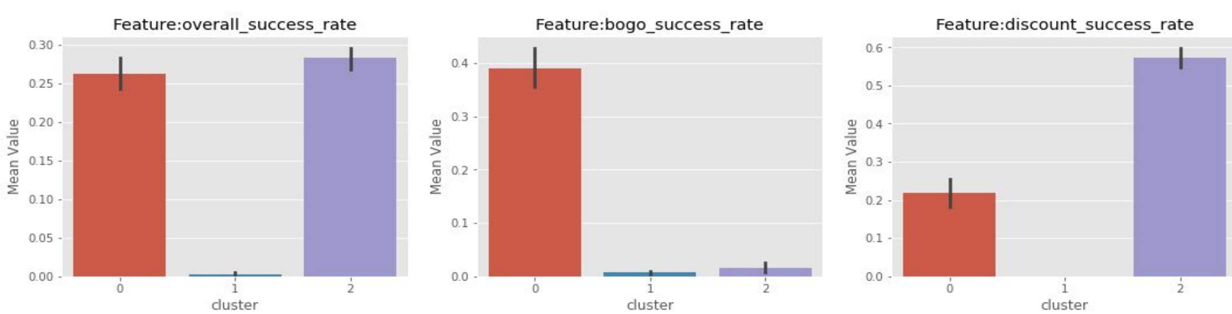
Bonus

Clients with missing demographics data is analyzed using kmeans clustering , 3 clusters were chosen based on silhouette score (0.32) and inertia (5500). Please refer to notebook for details.

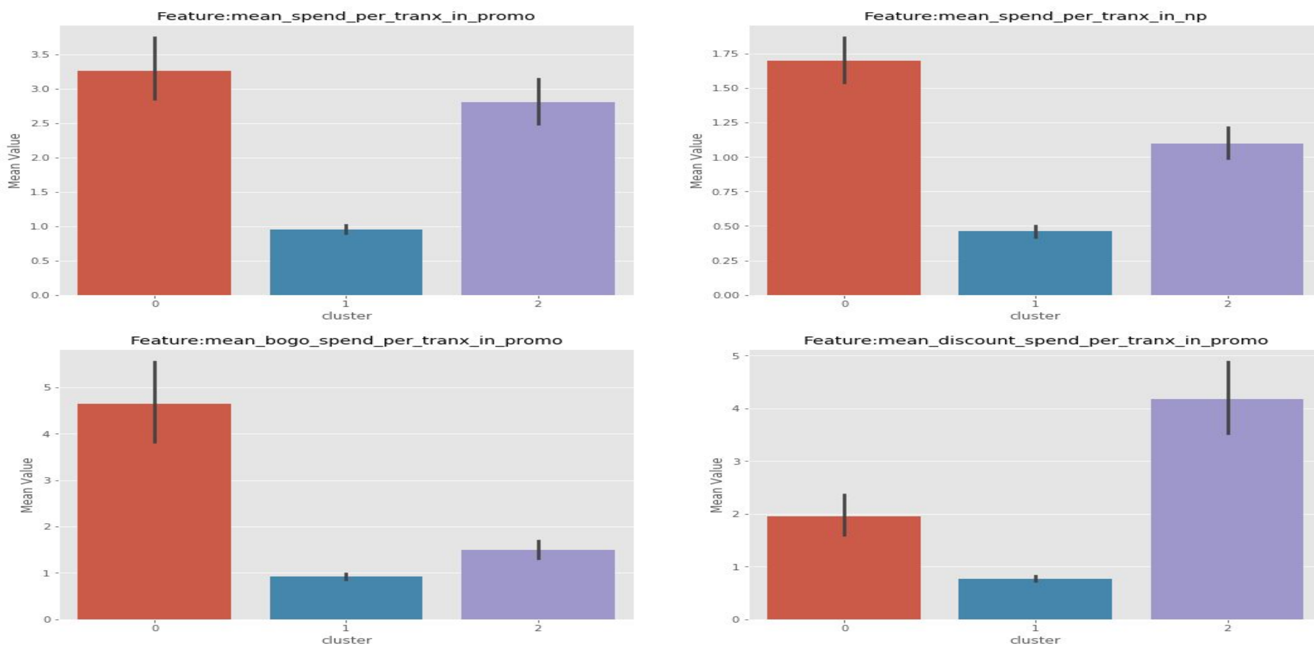
Here are few observations about 3 clusters



Let's look at the rate of success across three clusters



And also analyze mean amount spent per transaction during different phases of offers.



Summary of cluster analysis:

When compared to clients with demographic information, the clients with missing demographic information have spent less and fewer transactions over the observed period.

- Cluster 0 (16% of Missing profile population): The clients in this cluster responded well to BOGO offers with bounce in amount spent per transaction while participating in completing the BOGO offers.
- Cluster 1 (51.3% of Missing Profile population): The clients in this cluster mostly didn’t complete any offers, the clients in this cluster average around 5 transactions over observed period of 30 days and average amount spent per transaction is under $1.
- Cluster 2 (32.6%): The clients in this cluster responded well to discount offers with bounce in amount spent per transaction while participating in completing the discount offers.