# Definition

## Project Overview

Starbucks is an American multinational chain of coffeehouses. Starbucks stores serve both hot and cold drinks and bakery. However, there are many other local stores also serve nice coffee and drinks in each city. Starbucks have to use marketing strategies to stimulate their sales. They will send the promotions to consumers who have registered as Starbucks members by different kinds of channel. Sometimes, the consumers will respond to the promotions, sometimes, the consumers won't respond. Starbucks want to find a way to increase effective promotions to which consumers have higher possibilities to respond. Thus, in this project, I will analyze the data and to see which offer should send to certain customers.

## Problem Statement

The goal of this project is to improve the promotion results by sending the appropriate offers based on customers' attributes. I will use the following steps to answer the questions and then provide a model in the end which can predict if the customers will respond to these types of offer.

Step 1: Download json data and read it as dataframe in python.
Step 2: Access data and clean the data
Step 3: Visualize and check the distribution of data
Step 4: Use the results to answer the below questions.

Question:

1. The distribution of the ages of member.
2. The distribution of events between different genders.
3. The distribution of events through time.
4. The distribution of events for different offers.
5. The response rate between people of different ages.
6. The distribution of events for different offers within different age interval.
7. The percentage of events of two types(discount/bogo) for different age class.

Step 4: Build model by machine learning to predict the response of customers.

## Metrics

In this project, Starbucks want to send the promotions as long as the members will respond to it. They don't want to miss any opportunity that the members might respond to the promotions. Thus, I use recall rate to measure the effects of the models. The more true positive results are caught, the better the model is. Recall rate is a trade-off between precision, but catching false positive doesn't hurt much to Starbucks. They just send the promotions to people who don't respond, but won't miss the members who will respond.

$$Recall\ rate = \frac{True\ positives}{True\ positives + False\ positives}$$

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

| | | Predicted | |
|---|---|---|---|
| | | 0 (Negative) | 1 (Positive) |
| Actual | 0 (Negative) | True Negative | False Positive |
| | 1 (Positive) | False Negative | True Positive |

Starbucks will like the False Negative as small as possible and True Positive as large as possible.

False Negative might be caused the delay between the offers sent to the members and the members actually respond to it. Thus, we have to also consider the time - time in hours since start of test.

## Data Exploration

### Data Access

Portfolio

- id - offer id
- offer type - type of offer, BOGO(buy one get on free), discount, informational
- difficulty - minimum required spend to complete an offer
- reward - reward given for completing an offer
- duration - time for offer to be open, in days
- channels – method release the offers: email, mobile, social, web

Profile

- age  - age of the customer
- became member on - date when customer created an app account
- gender - gender of the customer: F, M , O (others)
- id - customer id
- income - customer's income

Transcript

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

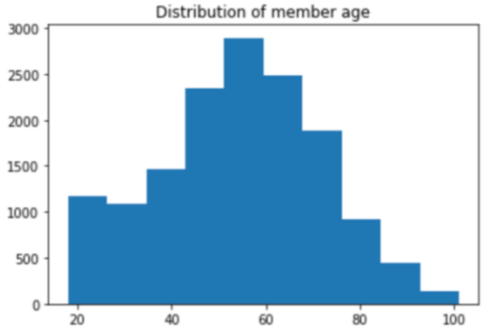
## Data Clean

Portfolio

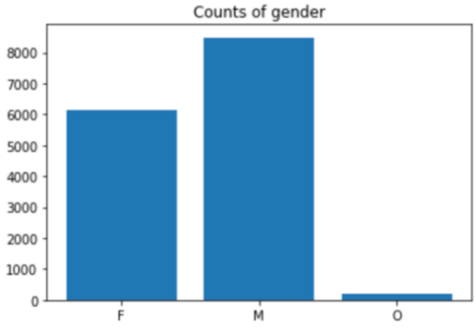
- simplify the id to be 1,2,3…10
- make offer type and channel to be dummy variables.

Profile

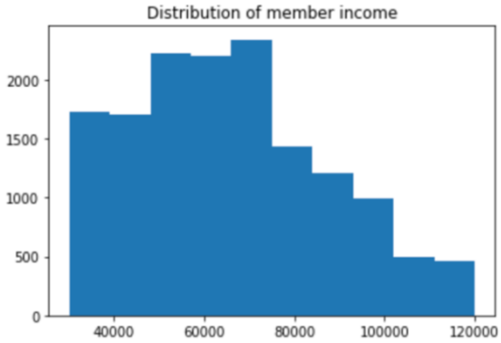
- transfer became_member_on to be datetime type and then create a new variable which calculate the day counts from the day became a member to today
- make gender dummy variables
- missing value in age/gender/income when make data showing as 118/NaN/NaN. The distribution is right skew. The min age is 18 and max is 101, and the mean for non-missing value is 54.39, mode is 58, medium is 55. The distribution is like the below figure. I decide to use the mean to replace missing value.



- Secondly, I use bootstrapping to replace the missing value for gender. I randomly choose a member from the data and use the gender to fill in the missing gender and repeat it for each missing gender because the distribution of gender in the dataset is like the below table. Gender is not evenly distributed in the data, so bootstrapping is a appropriate to fill in the missing value.



- The distribution of income is as the below figure. The min income is 30,000 and max is 120,000, and the mean for non-missing value is 65,404, mode is 73,000, and medium is 64,000.

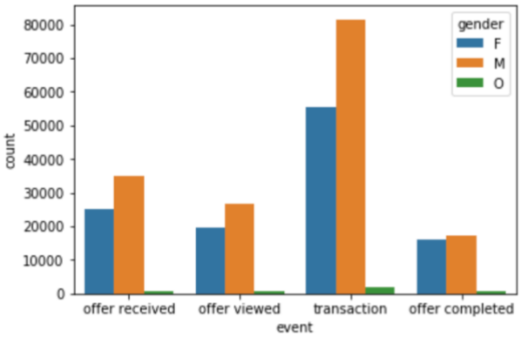


Transcript

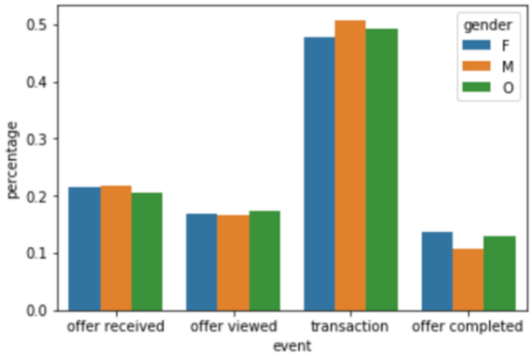
- event - record description (ie transaction, offer received, offer viewed, etc.)
- person - customer id
- time - time in hours since start of test. The data begins at time t=0
- value - either an offer id or transaction amount depending on the record

## Visualization

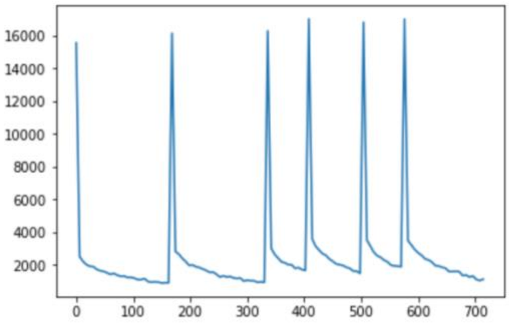
1. The distribution of events between different genders.

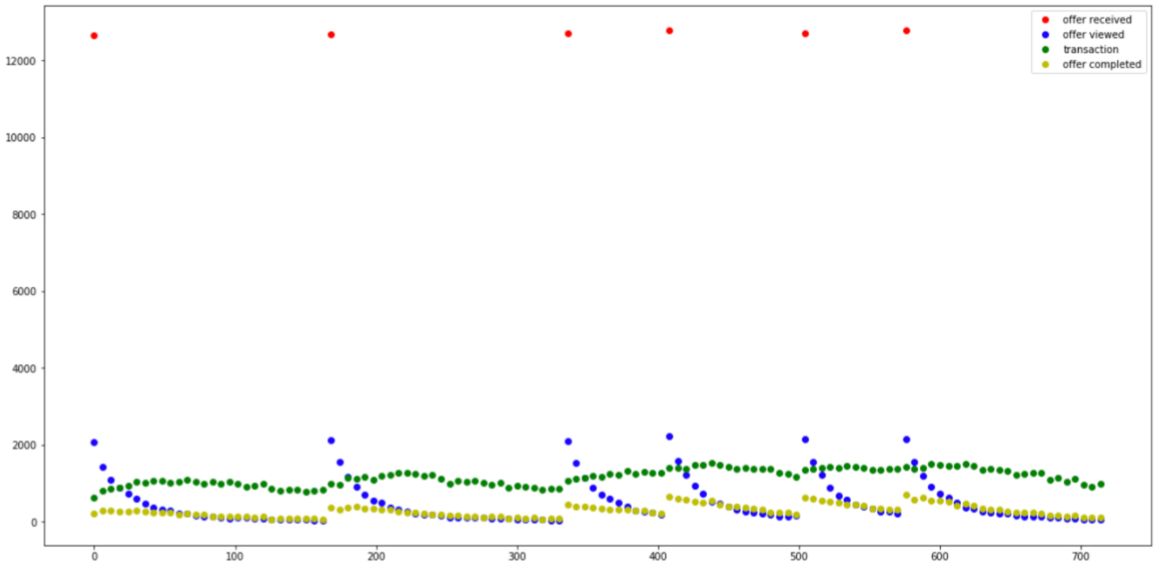


The above figure shows that more males have transaction and offer completed. However, from the counts of gender, we can see that the data have more male. Thus, I will use the percentage to see if certain gender has higher transaction or offer completed rate. We can see from the below figure that male has highest numbers in transaction but lowest in offer completed. Furthermore, Starbucks should care response rate more than offer complete rate so I plus the percentage of transaction and offer completed. We can see male and female have similar response rate.



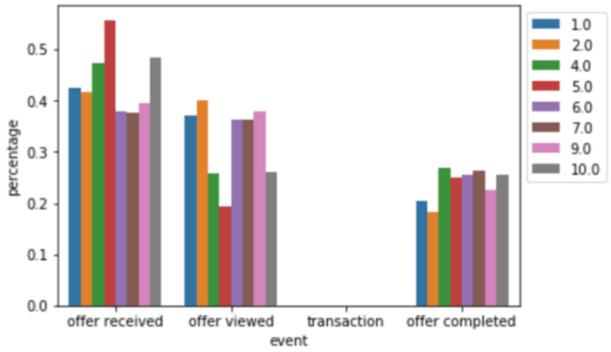
2. The distribution of events through time.





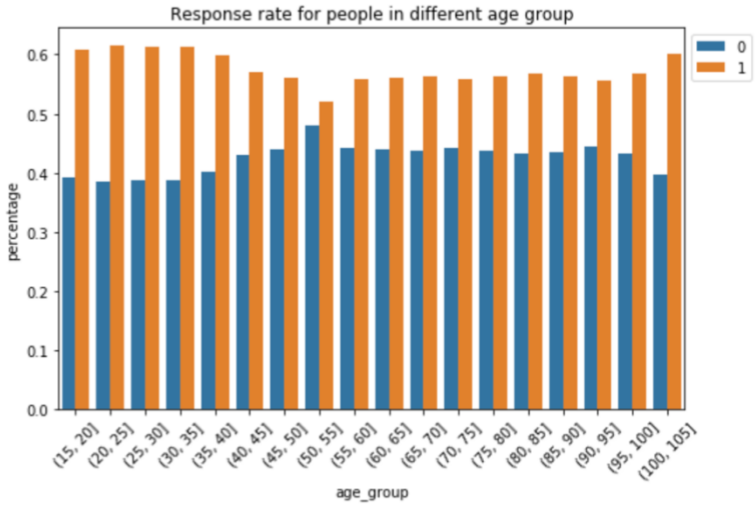
Offers are released periodically. The phenomenon is indicated by the red points in the graph. Offer reviewed decrease throughout the time as it becomes transaction and offer completed. Therefore, I conclude that time variable is correlated to the consumer's action.

3. The distribution of events for different offers.



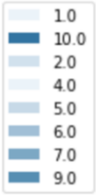
Offer 1,2,4,9 are bogo and offer 5,6,7,10 are discount. I exclude informational offers as they only have offer received and offer viewed status. The bar chart indicates that discount offers have higher completed rate overall. Therefore I suggest Starbucks to promote more discount offers.

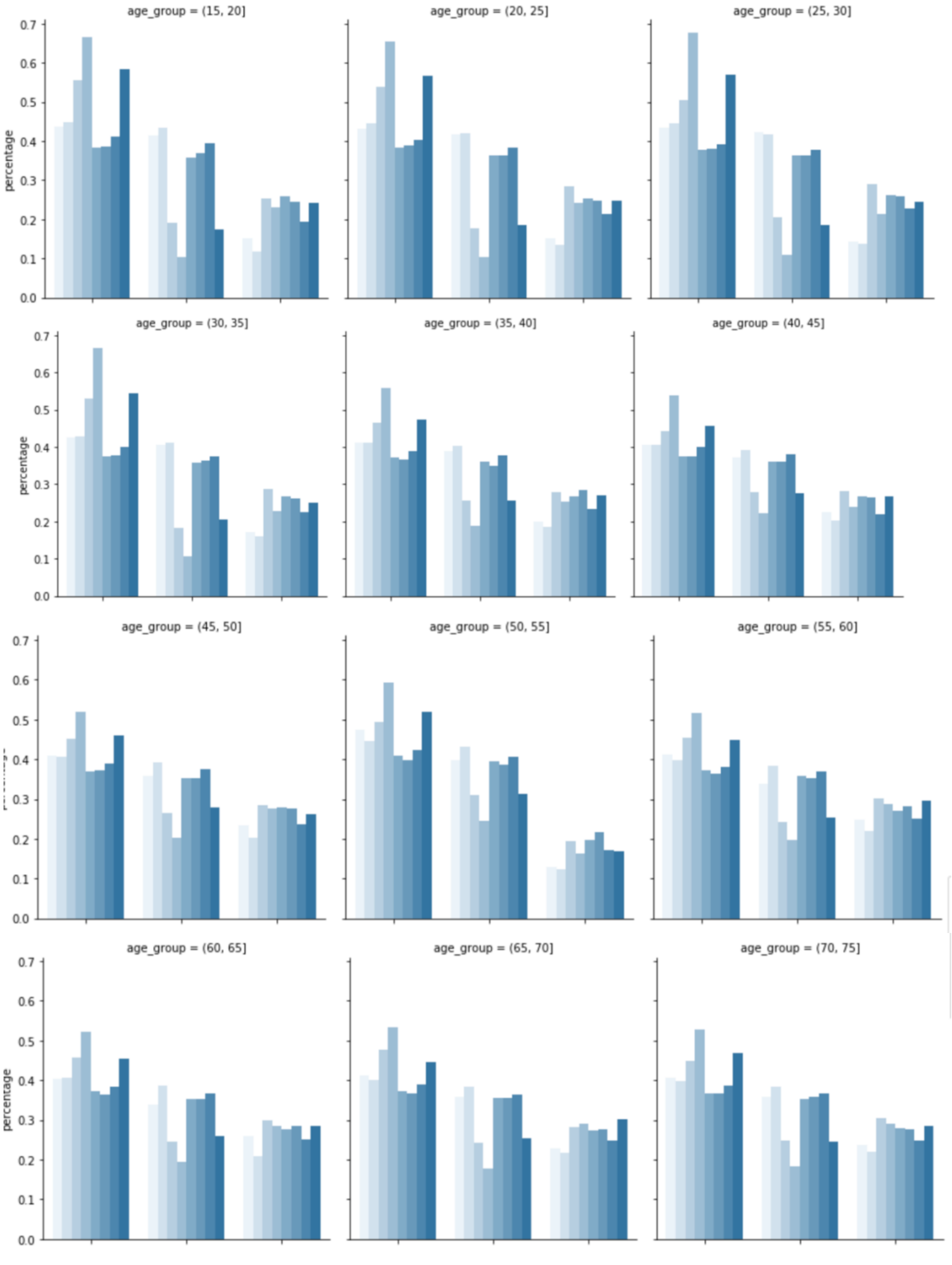
4. The response rate between people of different ages.

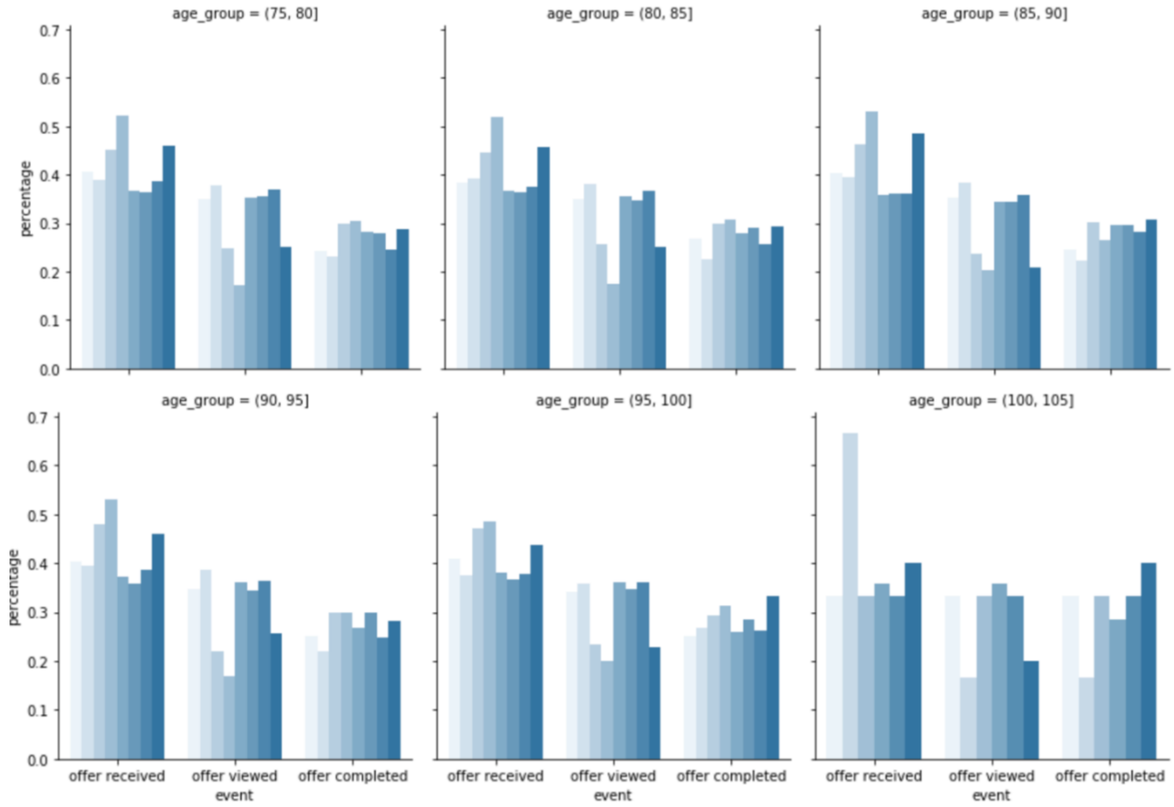


Responded rate decline from 35 to 55. However, it increases and stays stable after 55. One possible reason might be that people within age 35-55 are too busy to respond to the offers. For younger and older people, they might have time to respond to the offers. The other possible reason is that people in middle age are not sensitive to discount/promotion. People under 35 have more than 60% to the offers. People over 55 have second response rate and the people aged between 35-55 have the lowest response rate.

5. The distribution of events for different offers within different age interval.

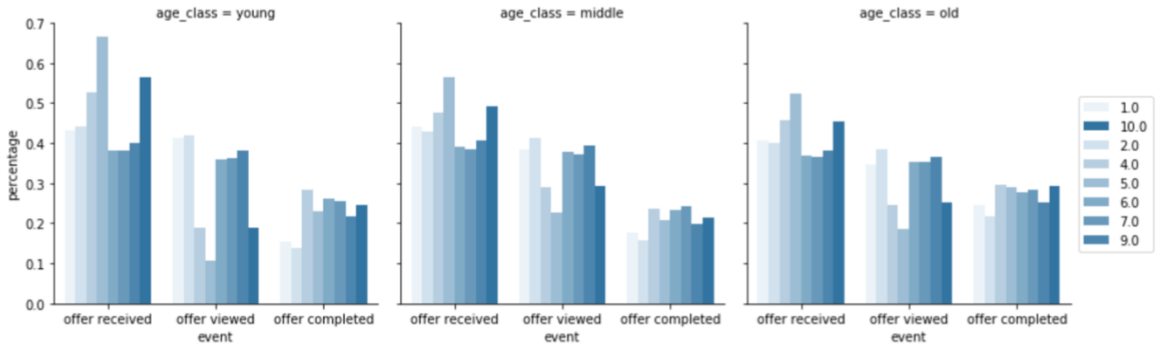


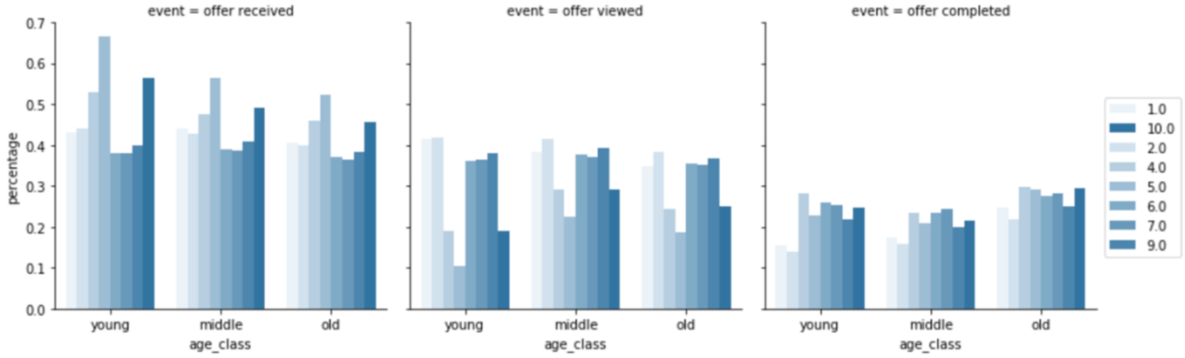

age_group = (15, 20]
age_group = (20, 25]
age_group = (25, 30]
age_group = (30, 35]
age_group = (35, 40]
age_group = (40, 45]
age_group = (45, 50]
age_group = (50, 55]
age_group = (55, 60]
age_group = (60, 65]
age_group = (65, 70]
age_group = (70, 75]
percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7



From the above figures, we see that the distribution of event within different age intervals doesn't change much from the distribution of question 3. Thus, I will say the age doesn't make the offer preference change. However, some age group shows obvious preference for some offer. For example, age group 25-30 prefer offer 4 much more than offer 1 and 2. Age 50-55 obviously less prefer to offer 9 and 10. Therefore, Starbucks should send more offer 4 to people in age 25-30, and decrease offer 9 and 10 to people in age 50-55.

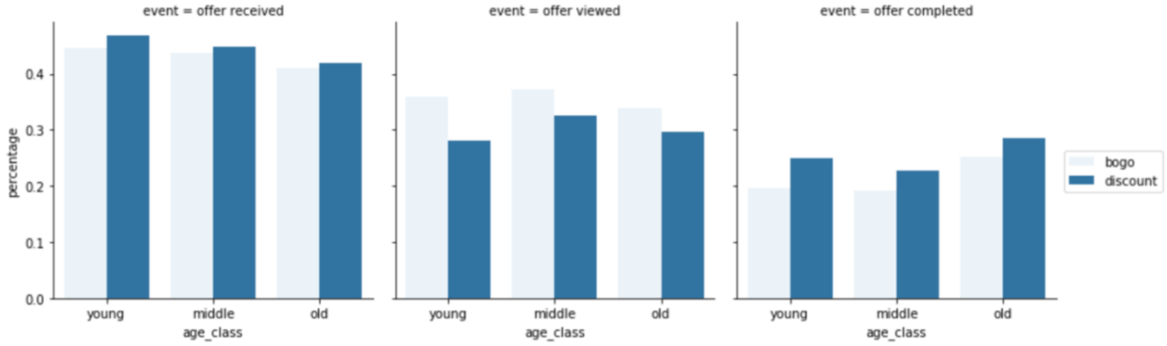
6. The distribution of events for different offers within different age class.





For young group, they don't complete much to offer 1 and 2. Young people prefer offer 4 most. The reason might be that the difficulty, 5, is the least and duration, 7, is longer compared offer 9 which is also difficulty 5 but duration only 5. People in all age class less prefer offer 1 and 2, but for middle age and old age, there preference is distributed more evenly through offer 4-10.

7. The percentage of events of two types(discount/bogo) for different age class.



People in middle age have higher offer viewed rate but lower offer completed rate. However, young and old people have lower offer viewed rate but higher offer completed, which means if they reviewed the offers they have higher possibility to complete the offers than people in middle age do.

People in all age have higher view rate to bogo offers but, interestingly, lower completed rate to bogo offers. However, we can not get the offer_id for those transaction events. Therefore, we cannot certainly sure if discount offers have higher completed rate unless we can get the offer_id for transaction events.

8. The percentage of events of two types(discount/bogo) for different age class

```
reward            -0.877836
bogo              -0.615480
social            -0.526682
mobile            -0.142961
difficulty        -0.033187
web                0.440926
simple_id          0.540107
duration           0.592372
discount           0.615480
completed rate     1.000000
email                   NaN
informational           NaN
```

email/ informational is Nan is because all offer_id have same values for these two variables.
Discount is positively correlated to completed rate, but bogo is negatively correlated to completed rate. Therefore, Starbucks should use discount as promotion more than bogo.

## Data Preprocessing steps

I combine transcript and profile to create a dataframe to build machine learning model. I want to use these model to predict if the member will react to the offer, but if the event is in transaction, then I am unable to know the offer type. Therefore, we only use offer received, offer viewed and offer completed to build the model. We will miss the transaction data but this model can predict if members will completed it or not.

In this model, I use these variables : time, offer_id, age, income, gender, membership days, difficulty, duration, reward, channel, offer type, completed rate. I transfer some variables, including time, age, income, membership days, difficulty, duration, reward, to standard distribution to optimize the model accuracy. Originally, I use MinMaxScaler, but the accuracy is worse. The reason is because that the variables are continuous variables and they are not sparse. Thus, StandardScaler has better results.

Later, I will also use all events to build another data but, in this model, we can only predict the response given the time, which is the time in hours since start of test, and the member information. This model cannot suggest Starbucks which offer type should send to the certain people but can use to decide if the member will respond to the offer through time pass.

## Implementation

I try both RandomForestClassifier and GradientBoostingClassifier and it turns out that GradientBoostingClassifier is much better. The accuracy is higher than 0.8

RandomForestClassifier

```
cross_val_score = 0.716
recall_score = 0.16
accuracy_score = 0.71
```

| | | Predicted | |
|---|---|---|---|
| | | 0 (Negative) | 1 (Positive) |
| Actual | 0 (Negative) | 28513 | 3822 |
| | 1 (Positive) | 8481 | 1639 |

GradientBoostingClassifier()

```
cross_val_score = 0.815
recall_score = 0.45
accuracy_score = 0.814
```

| | | Predicted | |
|---|---|---|---|
| | | 0 (Negative) | 1 (Positive) |
| Actual | 0 (Negative) | 30023 | 2312 |
| | 1 (Positive) | 5574 | 4546 |

## Refinement

The most important part of this model is to duplicate the responded samples because I want to increase the recall rate. If the number of non-responded sample is much more than the number of respond samples, the weight of responded will be too small to increase the recall rate. Therefore, duplicate the responded data can increase the weight and also increase the recall rate of the model.

```
cross_val_score = 0.858
recall_score = 0.943
accuracy_score = 0.646
```

| | | Predicted | |
|---|---|---|---|
| | | 0 (Negative) | 1 (Positive) |
| Actual | 0 (Negative) | 17900 | 14435 |
| | 1 (Positive) | 573 | 9547 |

## Results

We can use the analysis results to find the preference for people from different demography. I suggest Starbucks to send more offers to people under age 35. They will respond actively.

Besides, young people prefer longer duration and easy to complete offers. Secondly, Starbucks can send offers to old people, they respond secondly actively. Last but not least, if Starbucks have to send offers to people in middle age, they should send offer 4,6,7 which have better complete rate.

## Reflection

In this project, the first step is to access the data and to see if there are missing values. Then, fill in the missing values by the structure of the features.

Secondly, I visualize the data and see if there are correlation between features and the reaction of the offers.

Last step, I use the cleaned and transformed data to build a model to predict if the members will responded to the certain offer, based on the time and both offer and member features.

## Future Improvements

If I can get the offer type for the transaction events then I can improve the accuracy. The reason is that I can take more samples into consideration. Starbucks care much about if members respond, even though they might not complete the offers. The other method is use the current data to predict which type the offers are, such as using amount to predict. If the amount is large, then it might probably be the difficulty 20 offer type, which means the members have spent much but still cannot complete the offer.