Udacity Data Science Nanodegree Capstone Project

Author: Rohan Hadvani



Starbucks is an American multinational chain of coffeehouses and roasting reserves headquartered in Seattle, Washington, and operates more than 30,000 locations worldwide in more than 70 countries as of early 2020. Once every few days, Starbucks sends out an offer to users of the Starbucks rewards mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or a buy one get one free.

In this blog we are going to analyze 3 datasets:

1. *portfolio*.json: metadata for each and every offer (10 rows, 6 columns)
2. *profile*.json: data about each customer (17k rows, 5 columns)
3. *transcript*.json: all the records (306k rows, 4 columns)

The process which we will follow: Define our question, understanding the Datasets, Data preprocessing, EDA, model the data, compare various ML algorithms, and finally selecting one model and improving it.



**Question?**

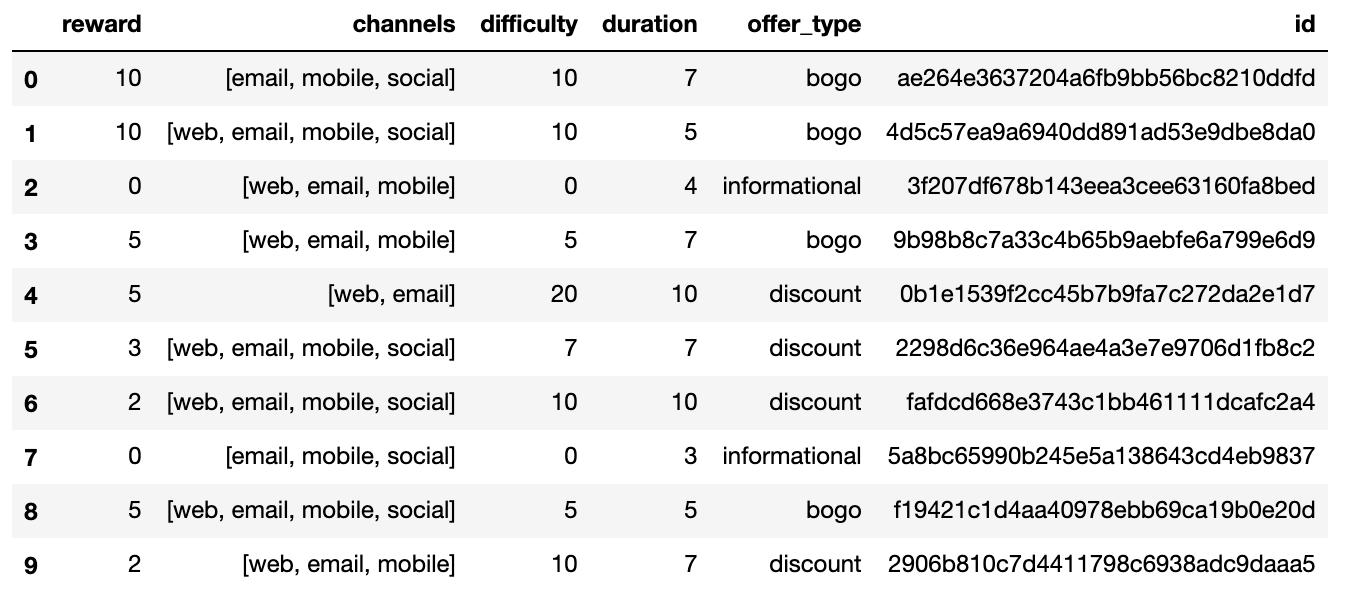
The question that we will answer in this analysis is that how does a customer react when an offer is given to them?

**Datasets**

Let’s look at all the datasets:

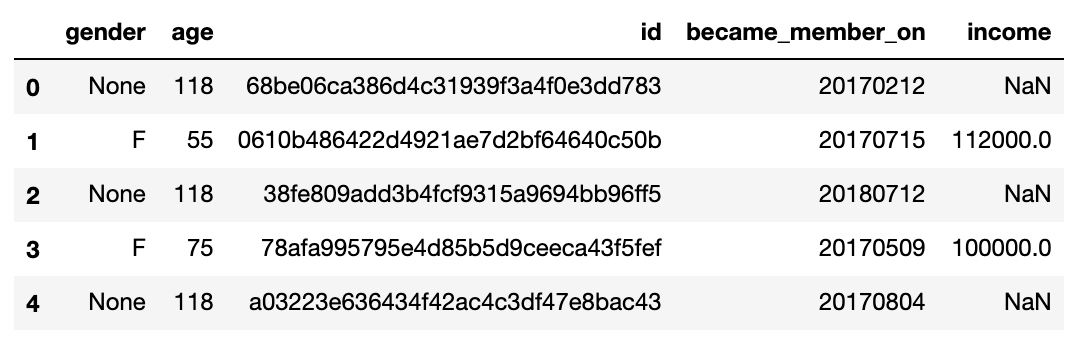
1. **Portfolio.json**

* id (string): offer id.
* offer\_type (string): type of offer i.e. BOGO, discount, informational.
* difficulty (int): minimum required spend to complete an offer.
* reward (int): reward is given for completing an offer.
* duration (int): time for the offer to be open, in days.
* channels (list of strings): channels include mobile, web, email and/or social.

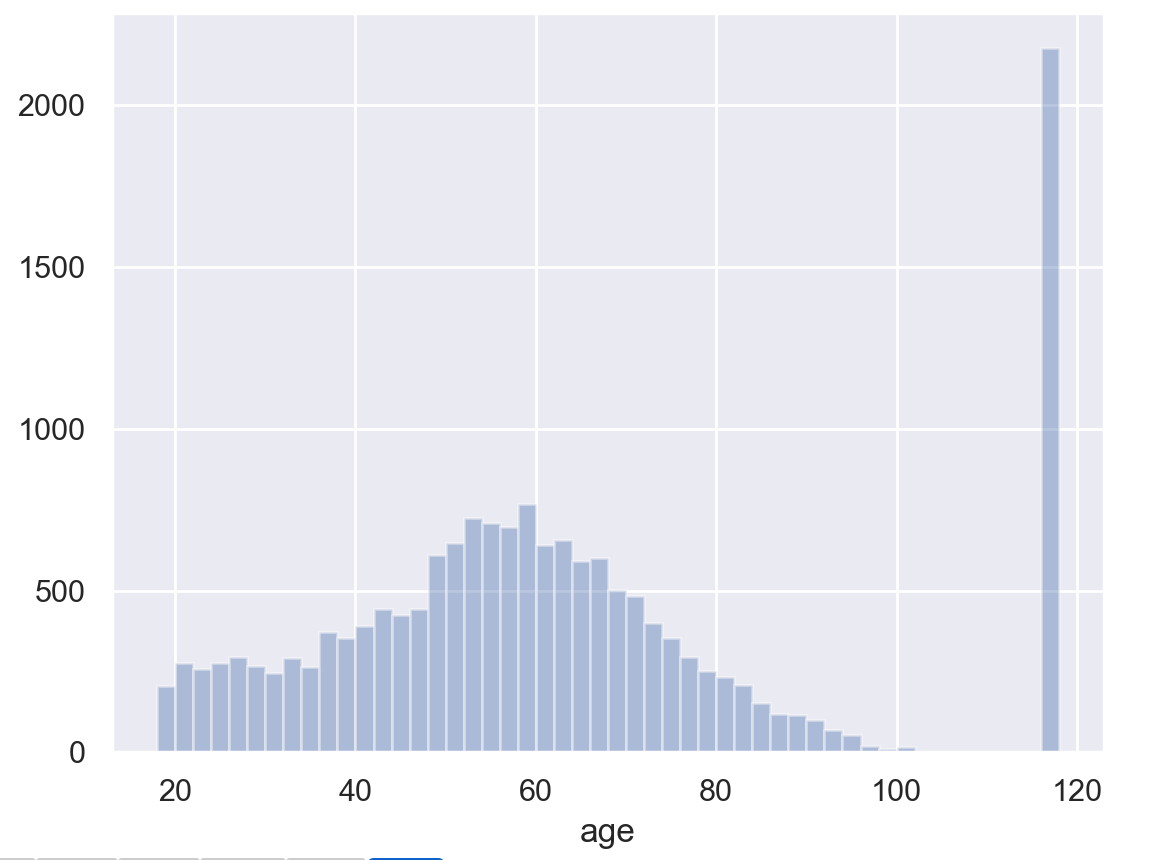


2. **Profile.json** :

* age (int): age of the customer.
* became\_member\_on (int): the date when customer created an app account.
* gender (str): gender of the customer (M: Male, F: Female, O: Others).
* id (str) : customer id.
* income (float) : customer’s income.

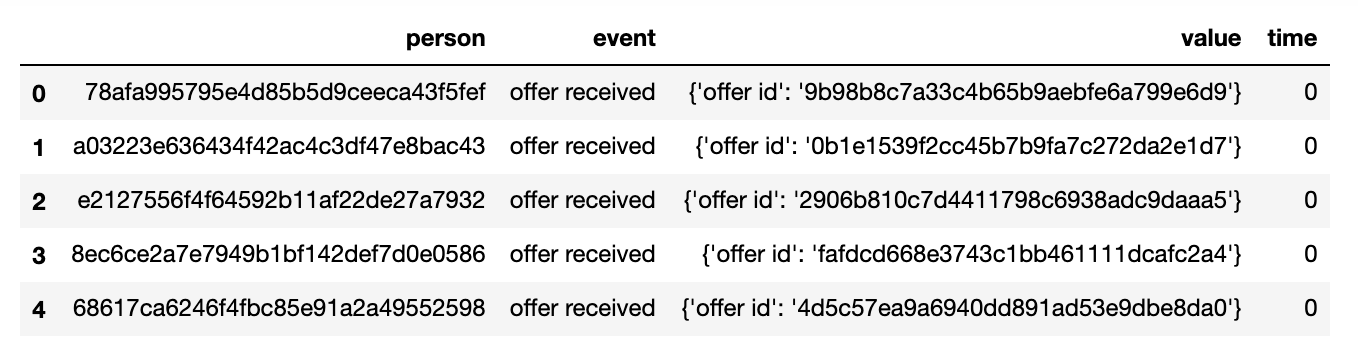


In this dataset, we find that there are a few people with the age of 118. These people are the ones who don’t have any gender or income. My best guess over here is that either they are fake profiles or people who don’t want to share their private details.



3. **Transcript.json**

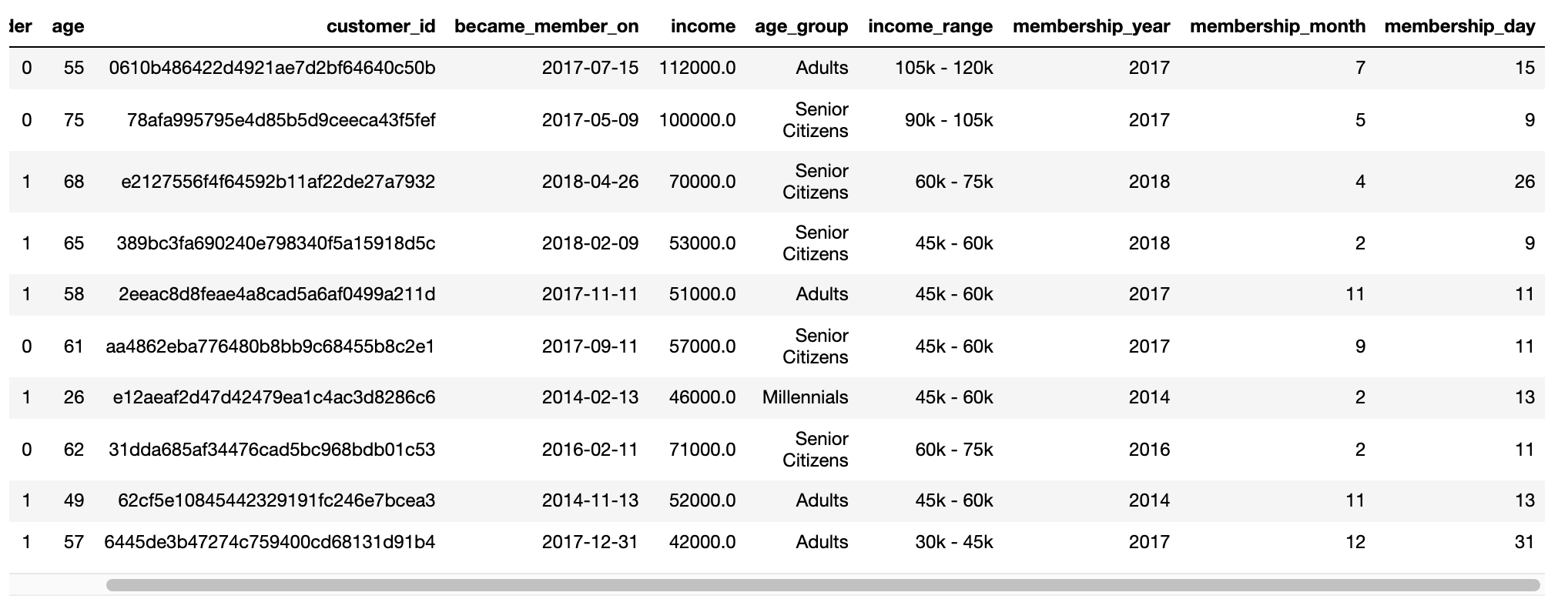
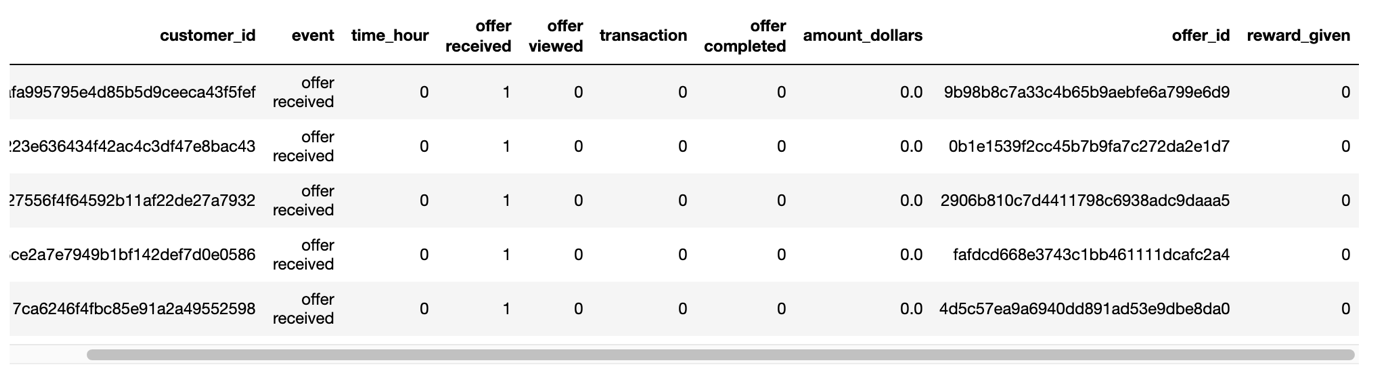
* event (str): record description (i.e. transaction, offer received, offer viewed, etc.)
* person (str) : customer id.
* time (int): time in hours since the start of the test.
* value (dict of strings): it can hold the values of ‘offer id’,’ amount’,’ reward’ and/or ‘difficulty’.



**Data Preprocessing**

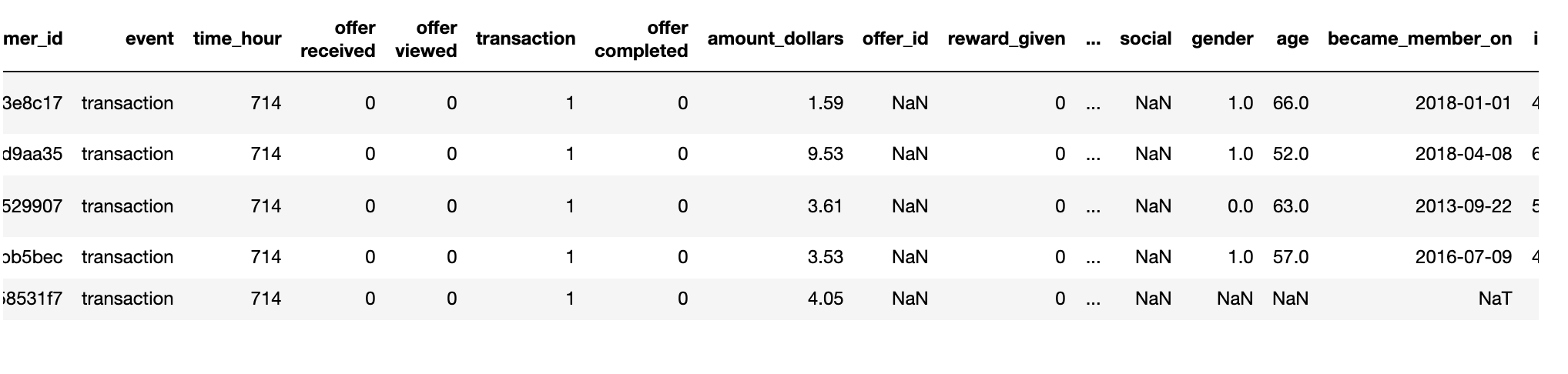
In this section, I clean all the data frame taking out the essential information from the various columns and make a new data frame. I also change the names of a few columns in order for better identification.

All the data frames after cleaning are:

PortfolioProfileTranscript

After cleaning all the data frames we merge them into 1 creating a master data frame which will be used further.

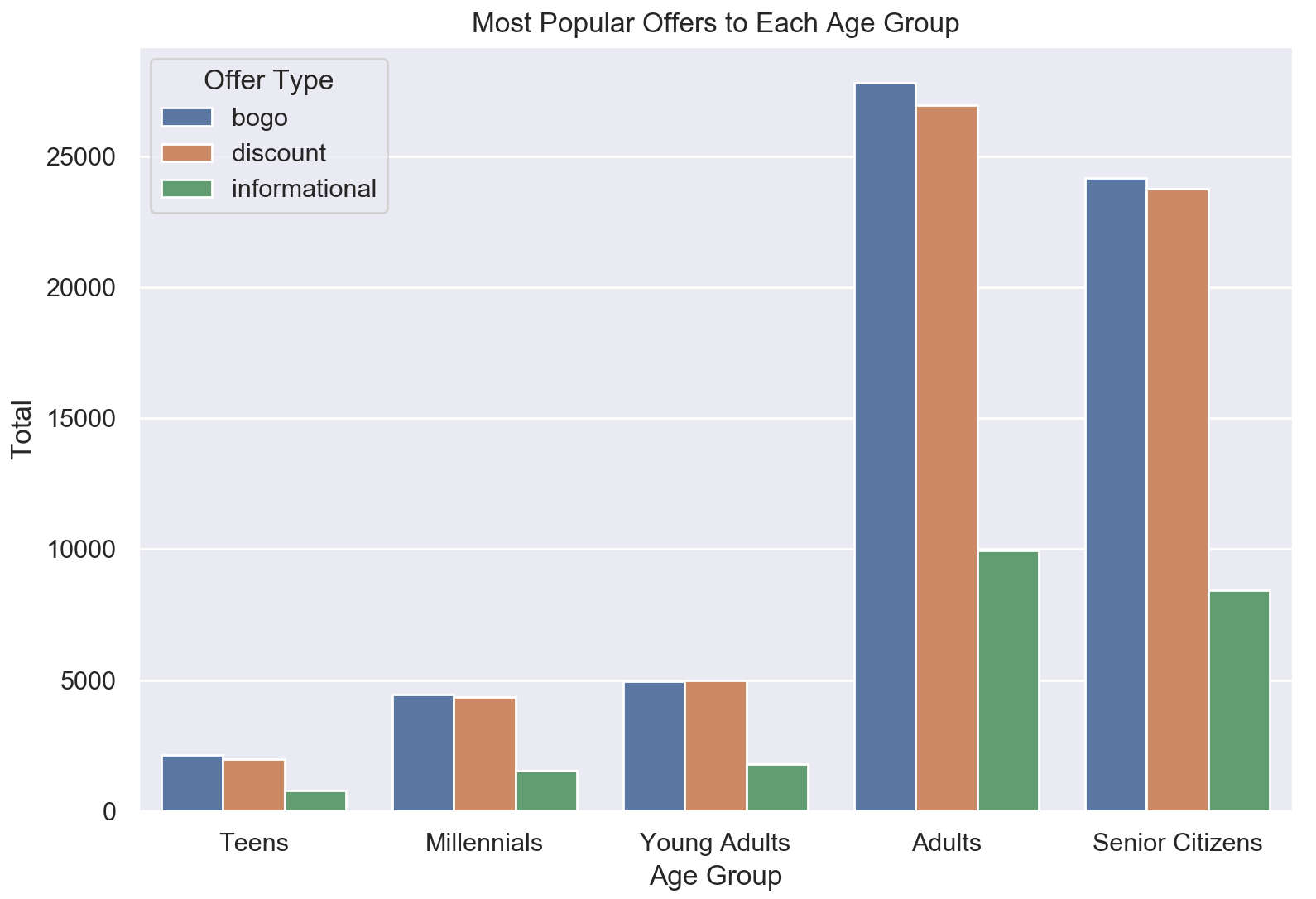
It contains 306k rows and 30 columns.

master\_df

**Exploratory Data Analysis**

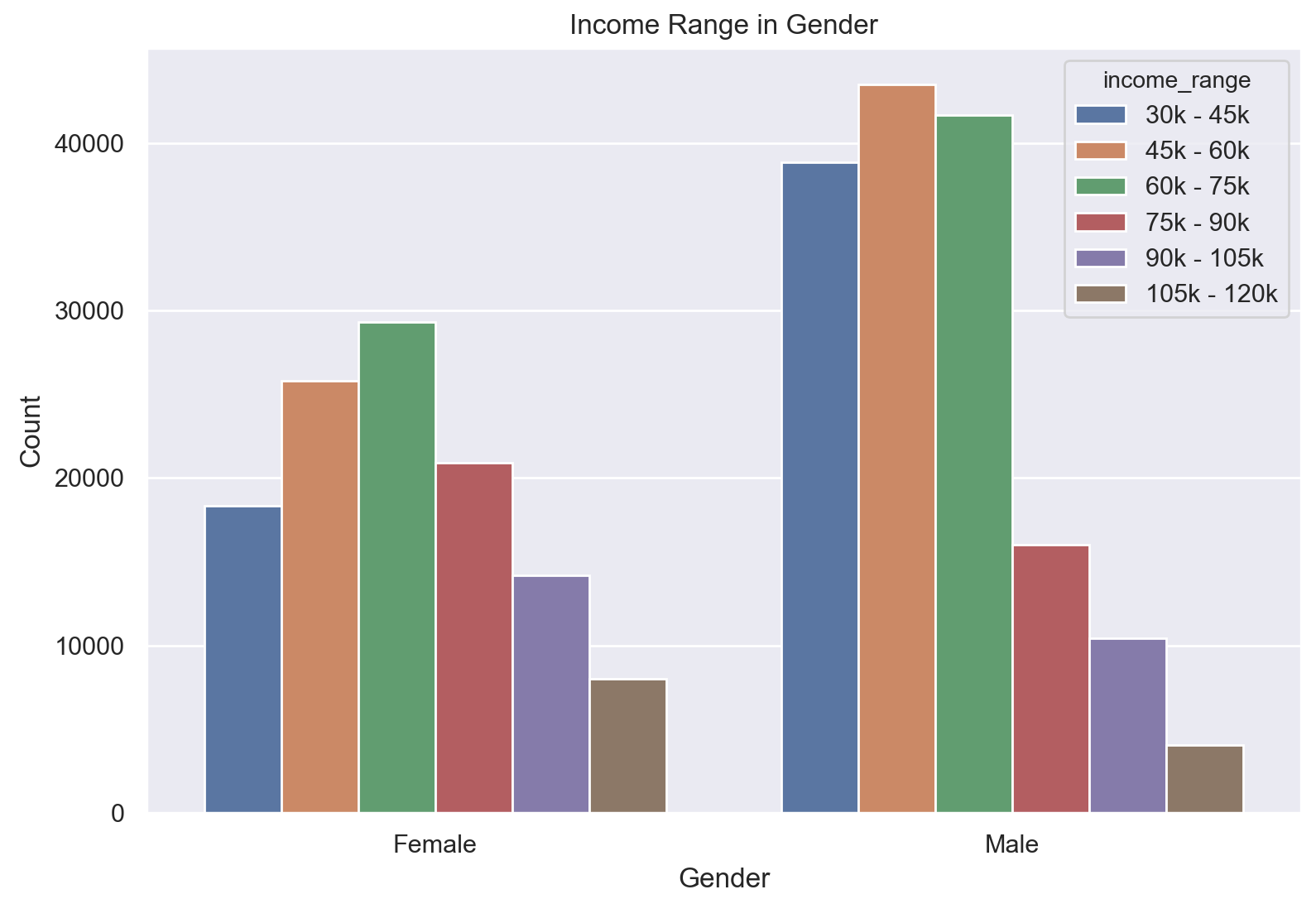
In this section with the help of various charts, we will answer a few more questions. (this is my favourite part)

*Q1. What are the most popular offers in the various age groups?*



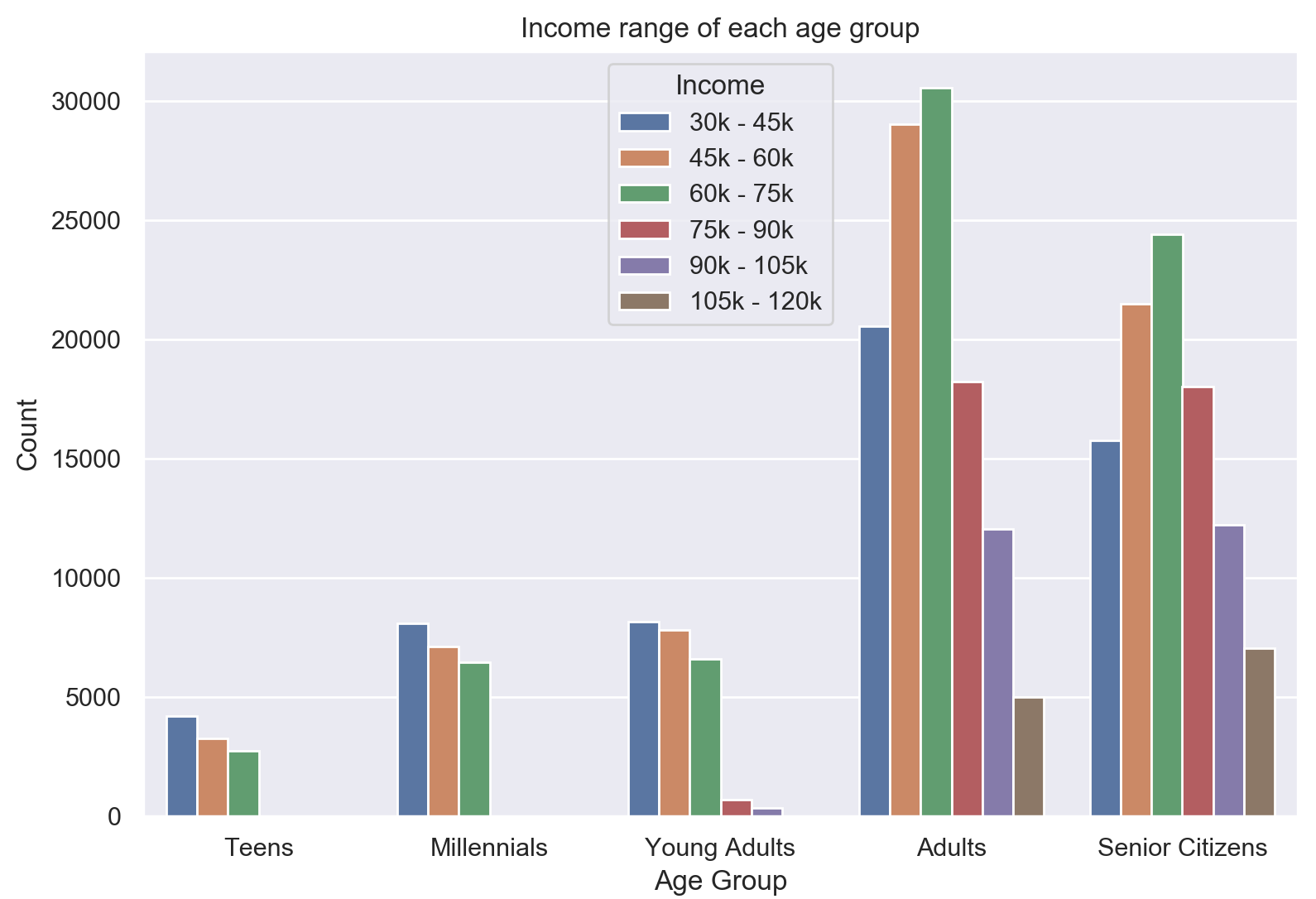
We can infer from the above plot that in all the age groups the BOGO offer & the discount offer is the most popular whereas the people don’t prefer the informational offer.

*Q2. How the income range varies for each gender?*



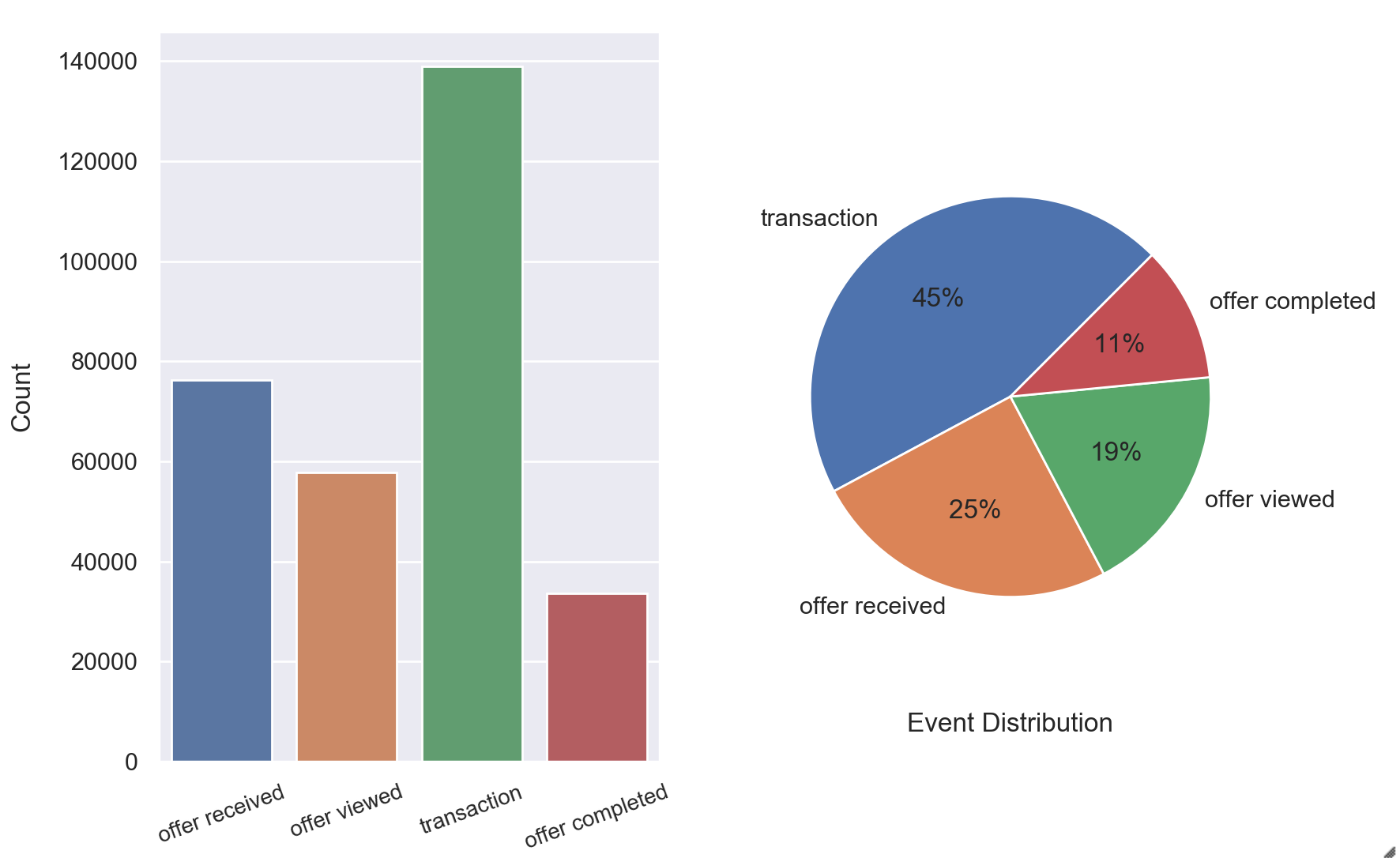
In the lower spectrum(less than 75k) we can see that males earn more than females but in the higher spectrum(75k onwards) females earn more.

*Q3. How the income range varies for each age group?*



We can see that as there are more adults they are most high earners too. Very few Millennials or young adults earn above 75k. This could be also that the majority of the customers are belonging to the adult category.

*Q4. How is the data is divided between different events?*



We can see that the dataset has two types of events:

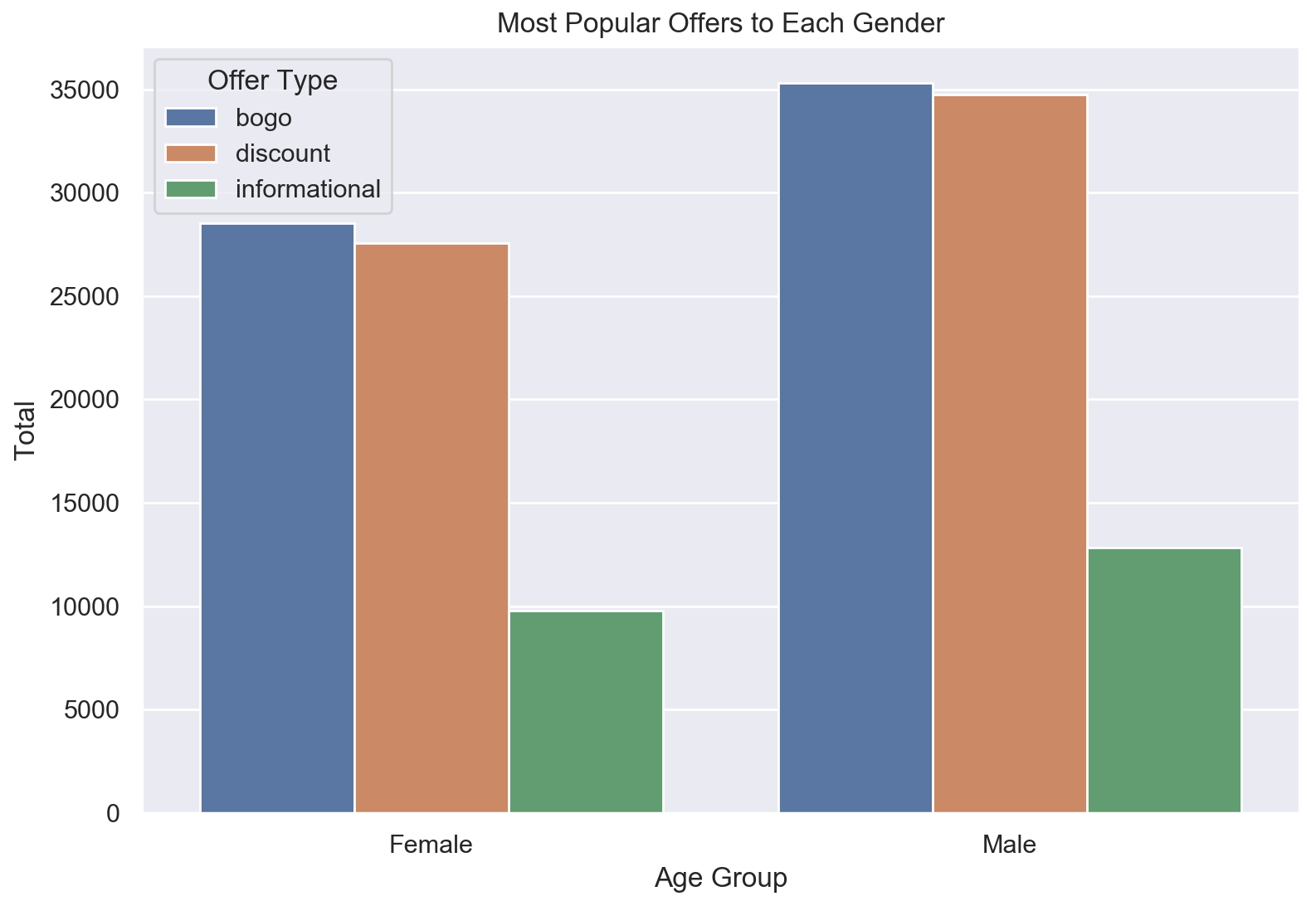
1. Offer-which is further divided into three classes:

* offers received
* offers viewed
* offers completed

2. Transaction-which holds records where a transaction has occurred

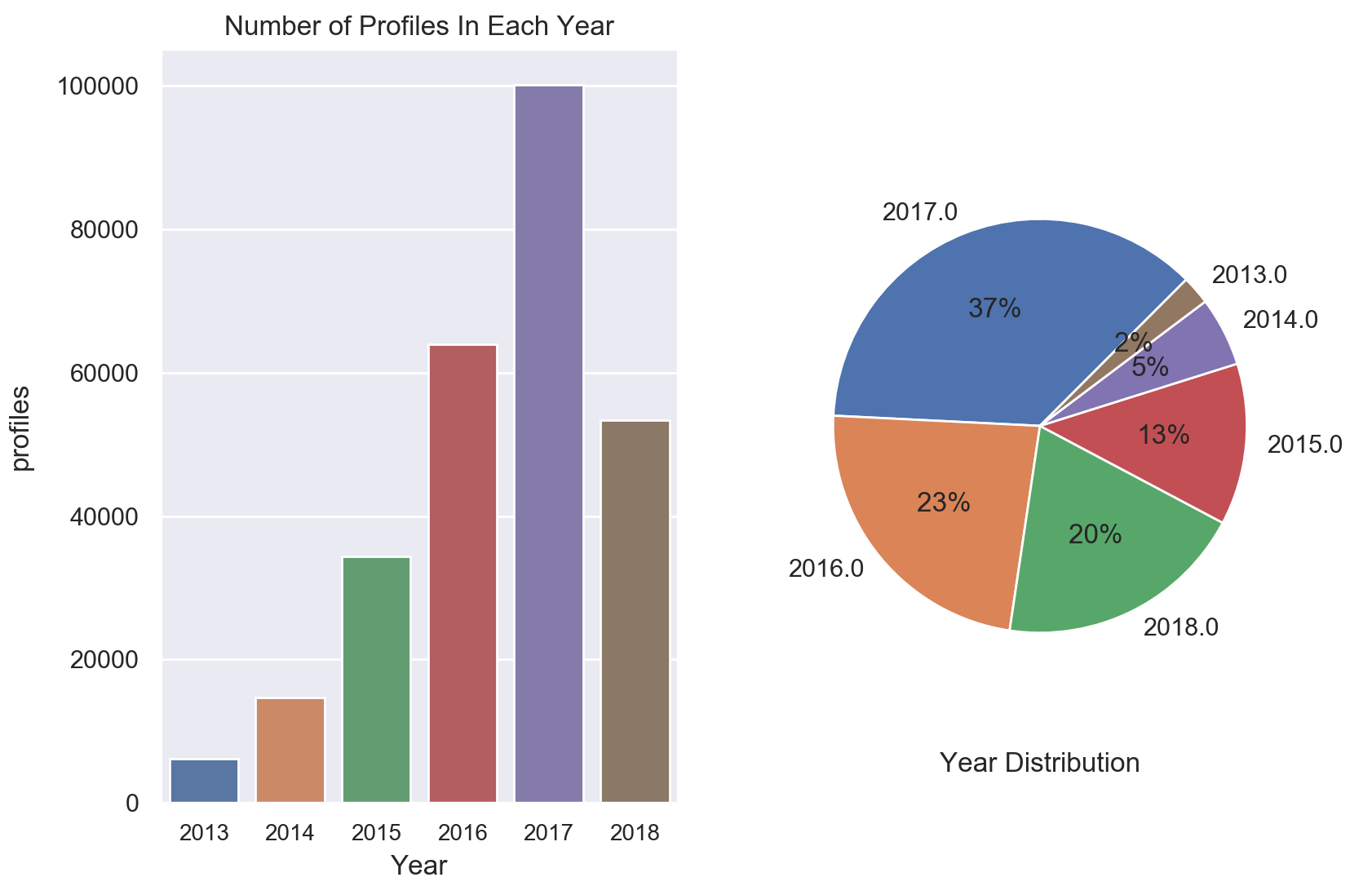
Offers make up 55 per cent of the dataset, leaving 45 per cent for transactions. As expected, it was not viewed by all users who received the offer and not by all users who viewed the offer.

*Q5. What are the most popular offers for each gender?*



The chart we got showed that both genders like BOGO and Discount offer and they have the same reaction toward Informational offers, they both seem to be not interested in it.

*Q6. How many profiles are added each year?*



We can see that the number of profiles increases every year till 2017 whereas they start to fall back again in 2018.

*Q7. Who are the most loyal customers?*



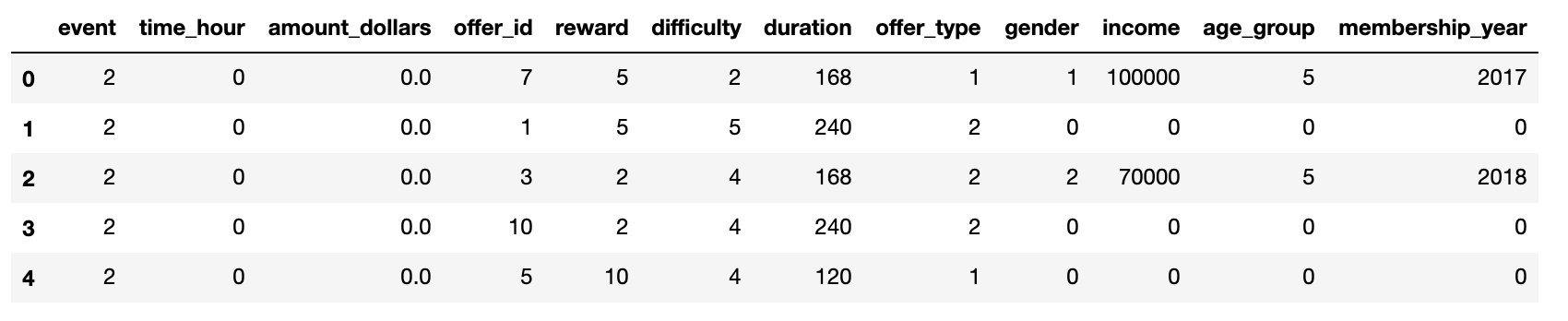
**Modelling**

In this section, we actually make the model and fit it to the dataset. I have selected to predict the ‘event’ column of the data frame.

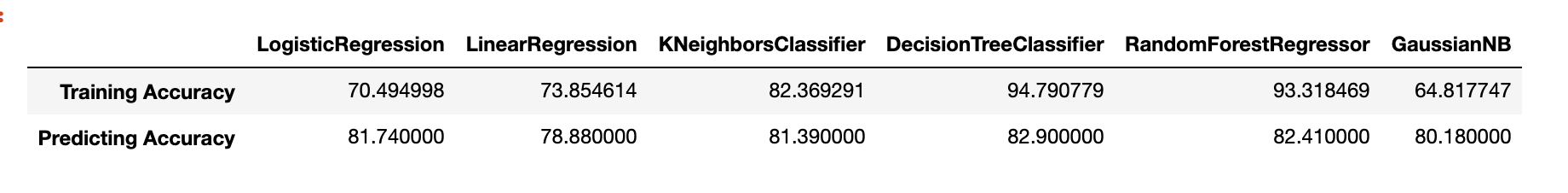
As the master\_df has a lot of columns we will make a separate data frame just for modelling purposes.

It will contain only these columns: *event,time\_hour,amount\_dollars,offer\_id,reward,difficulty,duration,offer\_type,gender,income,age\_group,membership\_year.*

We then perform some more operations on the model\_df data frame like converting some categorical data to numeric, converting float datatype to integer and replacing all the null values with 0. After all these the data frame looks like:



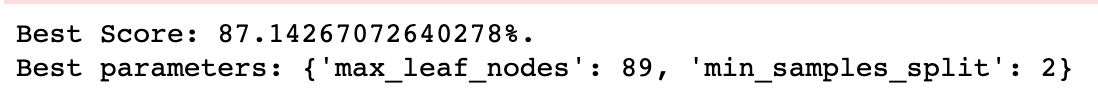
We normalize a few columns and then perform the fitting. The algorithms I choose are *Logistic Regression, Linear Regression, KNN classifier, Naive Bayes, Decision Tree classifier and Random Forest.*



From the above table and analysis, we come to the conclusion that the best method is the decision tree classifier. Is provides the maximum training accuracy and the best predicting accuracy. We will go ahead and perform improvements on this model. It is good here because we have a good amount of data to work with.

**Model Improvements**

After using *Grid Search* with *Decision Tree Classifier* we managed to get better results as shown here:



After this, we compute the training accuracy and predicting the accuracy of the new model with the best parameters.

The results are:

* Training Accuracy: **87.32%**
* Predicting Accuracy: **86.78%**

From the original fitting to using the best parameters for fitting we gain almost 4% in accuracy which is very good. Even though the accuracy on the training model goes down.

**Conclusion**

Throughout this project, I have analysed the dataset given by Starbucks and I have tried to predict what happens when a customer is presented with an offer. Whether the person just views it or completes it or performs a transaction. The dataset given was very huge and extensive. First, I’ve examined every dataset, evaluated its different aspects in order to gain an overall understanding of the data. Then I switched to the pre-processing portion. Pre-processing Data was the job that took much of the time and energy. The dataset is a little complicated and the data included allowed me to use my wrangling/engineering and pre-processing expertise to get the final clean version of the three datasets.

According to me, the preprocessing and making the data ready for analysis was the hardest part because there were three sets to be analysed. I loved doing the exploratory data analysis and plotting all the charts.

**Quick Facts and Interesting analysis**

1. The income of everyone is between 30k and 120k where men earn more in the lower spectrum (less than 75k) and women earn more in the higher spectrum.
2. There are 17k customers in the database with almost 2175 with no age, gender or income. These are the people who don’t want to give their personal details.
3. The most common offer type among all age groups is the BOGO, followed by the Discount Offers. Whereas, the least common offer to be sent is the informational offers. I believe that BOGO offers are more attractive compared to other offers provided by Starbucks.

Finally, I would conclude that I have successfully made a model which tells with 87% accuracy that what happens when a customer is presented with an offer. I really enjoyed doing the whole project. Though it was very extensive it was fun and engaging.



**Improvements**

I believe that there is always scope for improvement. Here we can use deep learning for a much better fitting. Or we could look at a few more features. We could also collect features like the demographics, branch, time of day etc.