Understanding customers purchase behavior at Starbucks

Definition:

Marketeers can achieve the best results by planning their campaigns with more insights about their potential customers. These insights can be obtained from the data present in their site, current campaigns and social media. Data science projects help marketers target the right customers and thereby enable profit maximization.

This Udacity capstone project is one such project where on analyzing the dataset, we get insights that help the marketing team to perform well. Thereby achieving business objectives. In this project, we take Starbucks which gives various promotions to its customers. There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational.

- In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount.
- In a discount, a user gains a reward equal to a fraction of the amount spent.
- In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend.

These promotional offers use multiple channels and they are e-mail, social media, on the web, or via the Starbucks's app. One of the goals of every marketing campaign is to bring in more profit, that is, the profit generated must be higher than marketing costs. The campaign aims to attract the customers that would eventually buy the product. Targeting people who are not likely to buy Starbuck drinks is not ideal. We need to find people who possess a high probability to buy Starbucks products by using promotions.

In this project, with the data provided, we analyze and find patterns between various features and find out which offer is appropriate to give to which kind of customers. That way, the offer leads that customer to make a purchase at Starbucks. To find a solution to the above stated problem, in this project we apply machine learning techniques to understand customers' behavior by analyzing their previous transactions with Starbucks. To find out which offer to send to a specific kind of customer, we perform Exploratory Data Analysis and find information such as which offer the customers are most interested in, demographics details of those customers that make the purchase using the offer, and others. To find out the appropriate response of a customer to an offer, we will use models such as Logistic regression, Decision Tree classifier and Random Forest classifier to determine the data that best represents our data. We use accuracy in this project as an evaluation metric. As for our benchmark model, a quick and fairly accurate model can be considered as a benchmark. we use the

KNeighborsClassifier to build the benchmark, as it is a fast and standard method for binary classification machine learning problems and evaluate the model result using accuracy as the evaluation metric. Also, we use accuracy since it is one of the common evaluation metrics in classification problems, that is the total number of correct predictions divided by the total number of predictions made for a dataset.

Analysis:
About the datasets:

The data set that is going to be used in this project is provided by Udacity and Starbucks as part of the Machine Learning Engineer Nanodegree program. It contains simulated data that mimics customer behavior on the Starbucks mobile app. The program used to create the data simulates how people make purchasing decisions and how those decisions are influenced by promotional offers. Each person in the simulation has some hidden traits that influence their purchasing patterns and are associated with their observable traits. People produce various events, including receiving offers, opening offers, and making purchases. Only the amounts of each transaction or offer are recorded.

This dataset contains simulated data that mimics customer behavior on the Starbucks rewarding system in their mobile application. Once every few days, Starbucks sends out an offer to users of the mobile app. The message can be 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. We are going to analyze three file:

- portfolio: containing offer ids and meta data about each offer (duration, type, etc.). 10 rows, 6 columns.
- profile: demographic data for each customer. 17000 rows, 5 columns.
- transcript: records for transactions, offers received, offers viewed, and offers completed. 306534 rows, 4 columns.

The process of our analysis will be by the following step: Define our Business question, understanding the Datasets, Data preparation and wrangling, analyze the data, model the data, compare model performance, and finally selecting one model and improving it.

To get an overview of the three dataframes, here are some snippets of those dataframes:

1. <u>Portfolio dataframe:</u>

	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

2. Profile dataframe:

	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

3. Transcript dataframe:

	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'}

From this we understand that offers can be delivered via multiple channels: email, social media, on the web, via the Starbucks's app. Each offer has a validity period (duration) before the offer expires. We see that informational offers have a validity period even though these ads are merely providing information about a product. Here, the duration is the assumed period in which the customer is feeling the influence of the offer after receiving the advertisement. As we can see, offer_type and channels are presented as a categorical values, which must be converted to columns using one hot encoding.

Data Exploration:

In the portfolio dataframe, the channels columns is packed with too much information. We need to dissect the columns into many columns. This makes it easy for us to derive insights into the information present in them. We use one hot encoding and get the following result. We then delete the channel column as its redundant now.

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

Moving on to the next dataframe 'profile', we check for null values in 'gender' and 'income' columns. We replace 'None' with NA in the 'gender' column and replace the NaN values in income column with the mean of the income column.

	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

Finally, we move on to the transcript dataframe.

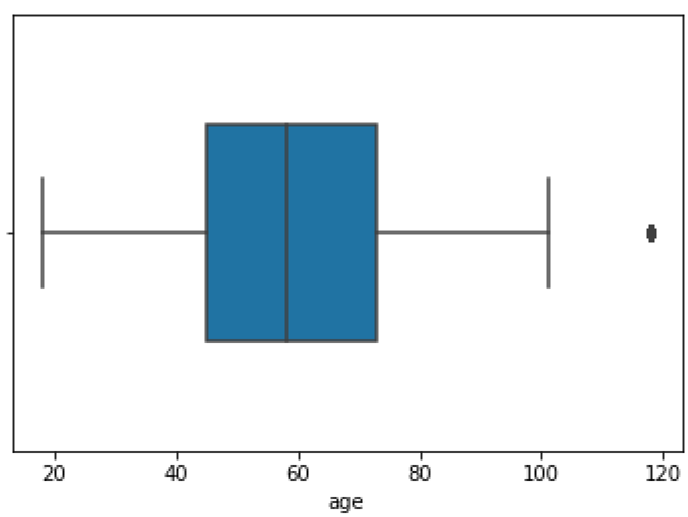
	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}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

The ‘value’ column contains a dictionary that means we have to separate each value and drop the ‘value’ column as it is no longer needed. To see what value it holds, we use a for-loop and find the keys. After iterating through the ‘value’ column we can find that we have the following keys:[‘offer id’, ‘amount’, ‘offer_id’, ‘reward’]. Our next step is to iterate over the transcript table, check the value column and update it, put each key in a separate column, and finally delete the ‘value’ column. After applying what I’ve discussed, the table will look like this:

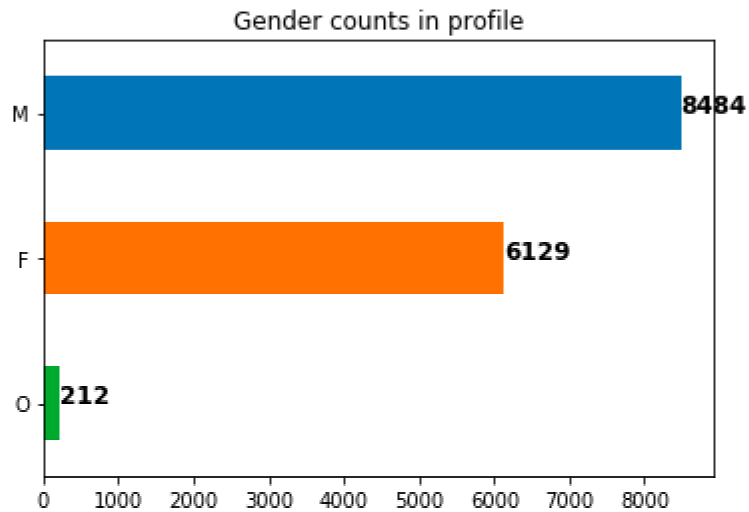
	event	person	time	offer_id	amount	reward
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	9b98b8c7a33c4b65b9aebfe6a799e6d9	0	0
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	0b1e1539f2cc45b7b9fa7c272da2e1d7	0	0
2	offer received	e2127556f4f64592b11af22de27a7932	0	2906b810c7d4411798c6938adc9daaa5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

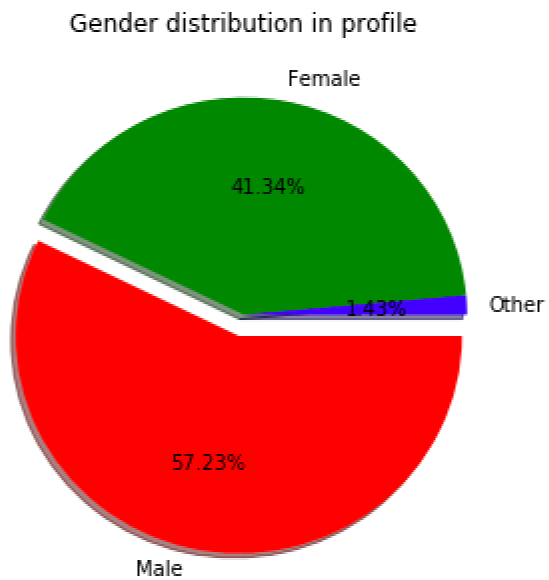
Exploratory Visualization:

We can perform various visualizations to understand the given data in depth. Following are some of them.
This is the boxplot of the Age column, we can see that the median age of customers is around 60.

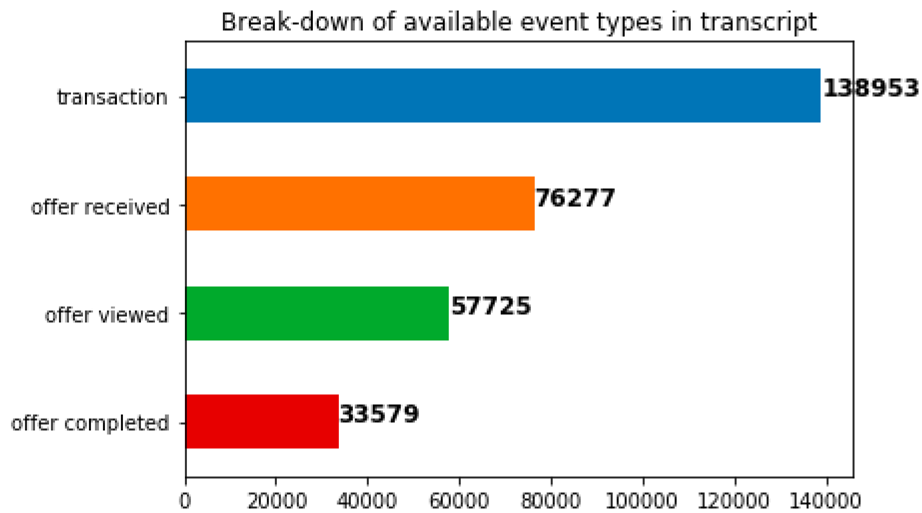


From this viz, we can assess the count of various genders as given in the dataset. The number of male customers is clearly high.





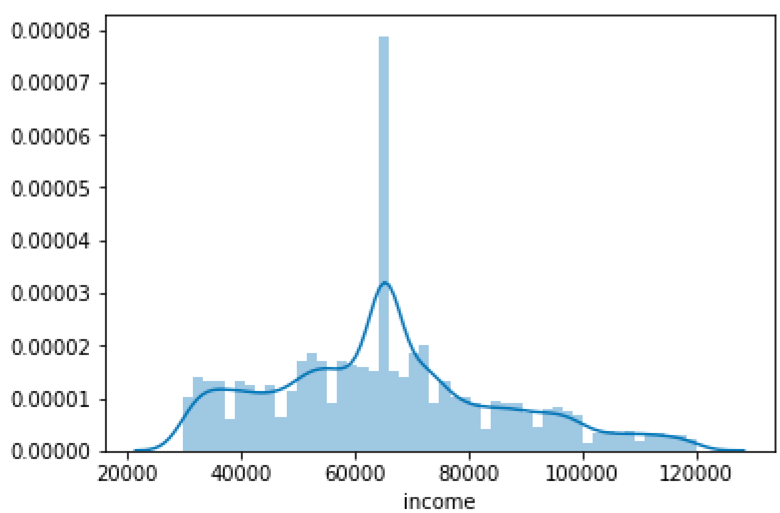
Here, we can see the breakdown of the event types in the transcript dataframe. The transactions are high meaning, the number of people who used are high as well but this also includes regular customers who bought products without any offers.



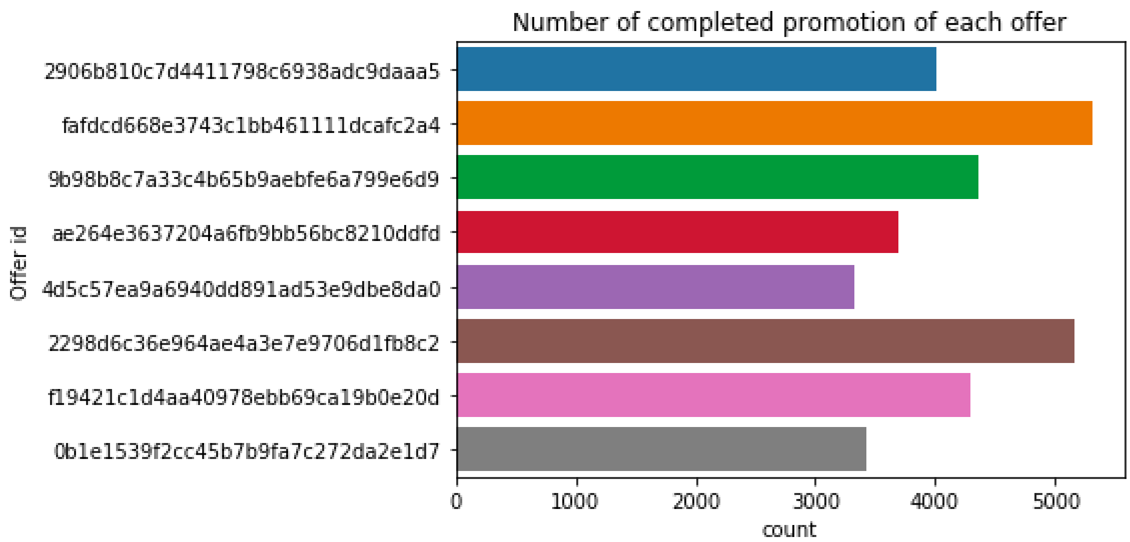
We find that the the average income for Starbucks customers by simply using the .mean() and the average income is:

```
65404.991568296799
```

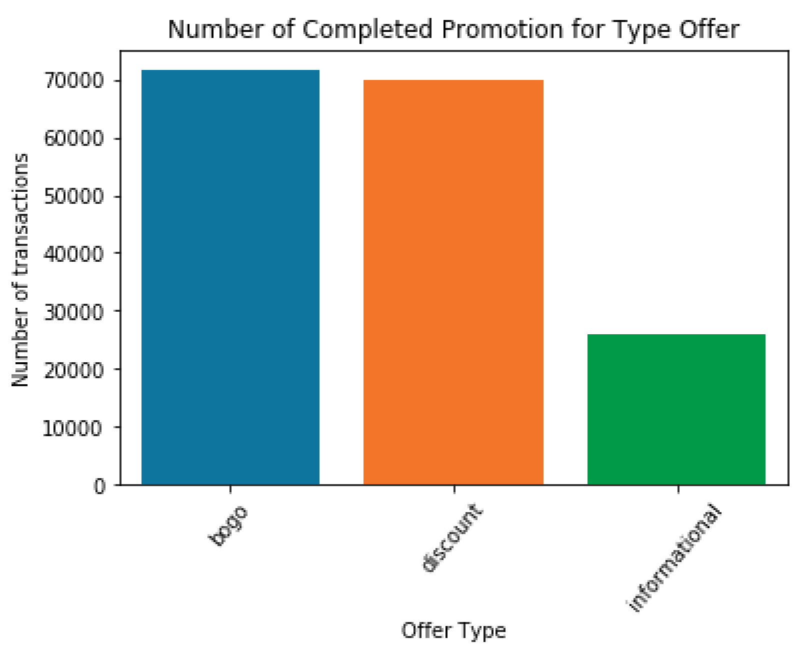
The distribution of the income column can be found in this graph:



This graph displays the number of transactions that has occurred for its respective offer.



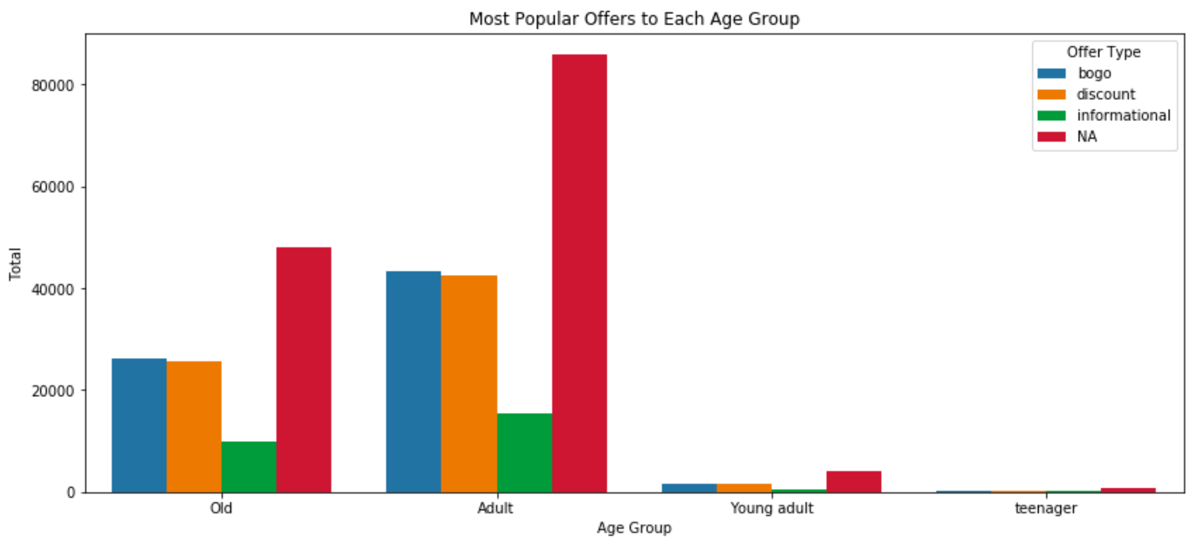
From the graph displayed below, we can find the most common promotion. We can assess that Bogo and Discount seem the most and they are close to each other with bogo been slightly higher.



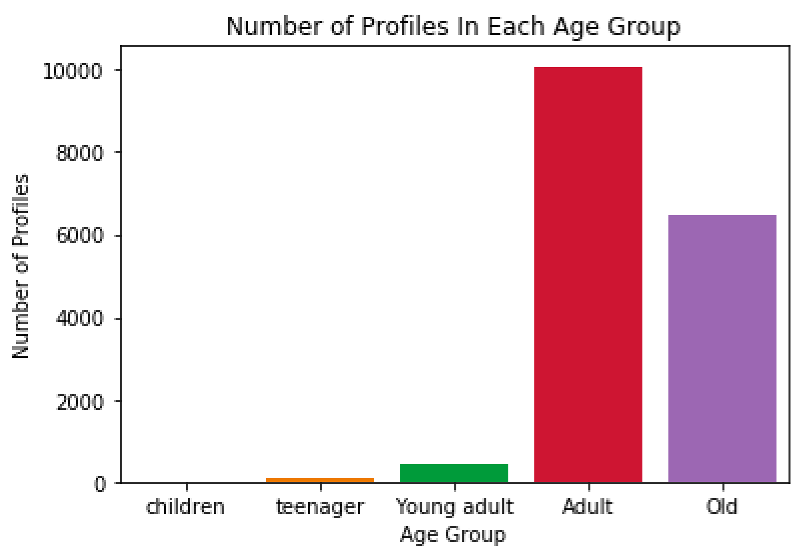
We can also calculate who are the most loyal customers with most transactions. Here we filter top 10 loyal customers. In this image, we can see the customer's unique profile ID, number of completed offers, and the amount. From this data, we can reward their loyalty by giving them extra and unique promotions.

```
Profile ID: 3c8d541112a74af99e88abbd0692f00e
Number of Completed Offers:5
Amount:$1606
Profile ID: f1d65ae63f174b8f80fa063adcaa63b7
Number of Completed Offers:6
Amount:$1360
Profile ID: ae6f43089b674728a50b8727252d3305
Number of Completed Offers:3
Amount:$1320
Profile ID: 626df8678e2a4953b9098246418c9cfa
Number of Completed Offers:4
Amount:$1314
Profile ID: 73afdeca19e349b98f09e928644610f8
Number of Completed Offers:5
Amount:$1314
Profile ID: 52959f19113e4241a8cb3bef486c6412
Number of Completed Offers:5
Amount:$1285
Profile ID: ad1f0a409ae642bc9a43f31f56c130fc
Number of Completed Offers:3
Amount:$1256
Profile ID: d240308de0ee4cf8bb6072816268582b
Number of Completed Offers:5
Amount:$1244
Profile ID: 946fc0d3ecc4492aa4cc06cf6b1492c3
Number of Completed Offers:4
Amount:$1224
Profile ID: 6406abad8e2c4b8584e4f68003de148d
Number of Completed Offers:3
Amount:$1206
```

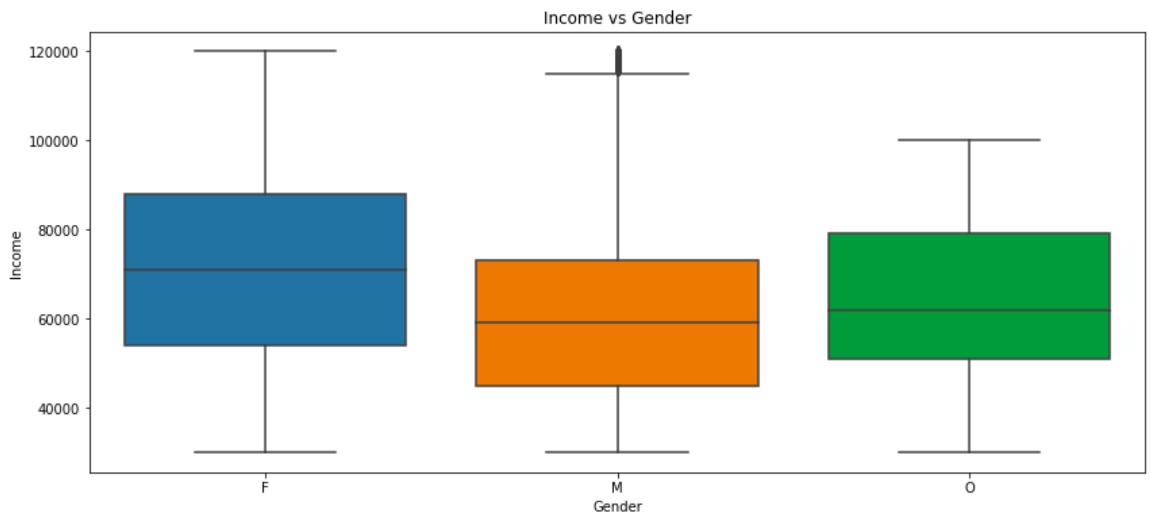
From this, we can answer which is the most popular promotion among various age groups.

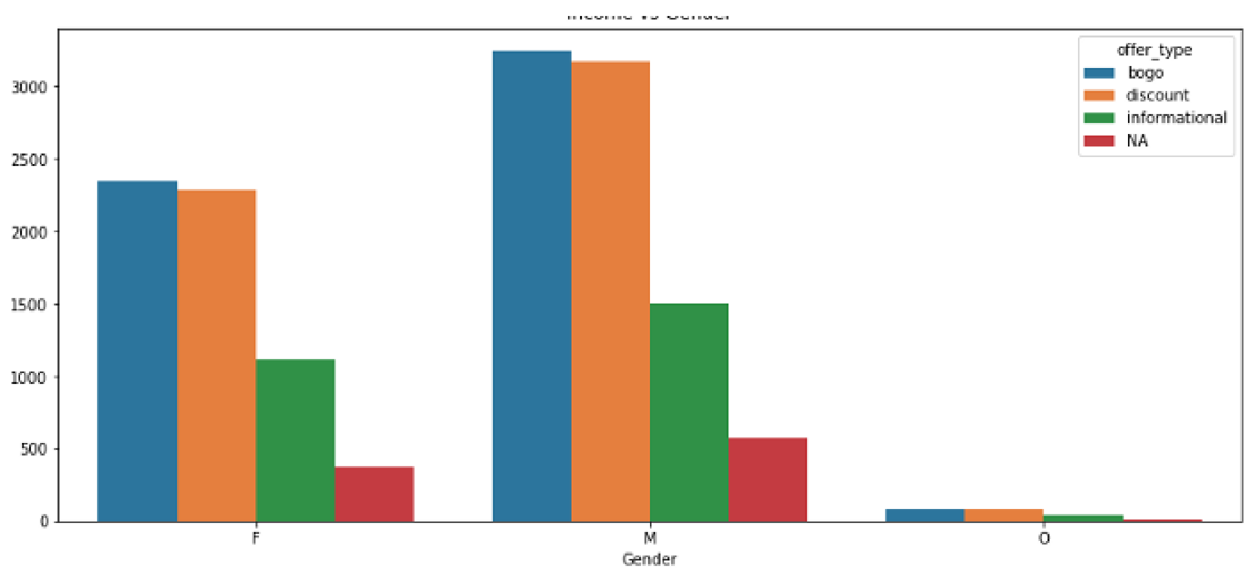


We can see that all age groups show similar pattern, they all favour BOGO after ofcourse the ones that purchase without any offers which is the red bar in the graph. From this graph, we can also say that the target customers are adults and Old people.



We use boxplots and plot income and gender. The graph shows that income median (the white dot) for females (around 70k) is higher than males (around 60k) we can also see that for females the income spreads from 40k to 100k. For males most of them around 40k to 70k which close to median.





Once again we focus on the gender to conclude that the most customers who use offers are male and that all genders prefer BOGO. Now, we will focus and machine learning and applying different models.

Modeling the Data

We build a model that can find the offers that we can present to a customer. Our model will guess the offer_type. Therefore, only consider ones with offer ids, and we ignore the ones without offer ids. This is a classification problem, hence we use accuracy as our evaluation metrics. We would like to see how well our model by seeing the number of correct predictions vs total number of predictions. Let us take the time to take about accuracy here, To define accuracy, it is the ratio of the correctly labeled subjects to the whole pool of subjects. Also, accuracy answers questions like: How many students did we correctly label out of all the students? It’s similar to our situation right? because we want to see how many customers use Starbucks offers. Furthermore, Accuracy = (TP+TN)/(TP+FP+FN+TN). Not to forget, that this is a simple classification problem, so this is my opinion and reasoning on why to use the easiest (accuracy).

The features we use now are Event, Time, offer_id, Amount, Reward, Age_gorup, Gender, and Income. Some are the features that are categorical will be changed to numerical and others will be normalized. The target variable is offer type
The models that I have used are: Logistic Regression, K-Nearest Neighbors, Decision Tree, and Support Vector Machine.

Compare model performance:
Now that we have trained the data, it's time to evaluate their performance based on accuracy.

	LogisticRegression	KNeighborsClassifier	DecisionTreeClassifier	SVC
Training Accuracy	80.522836	100.0	100.0	100.0
Predicting Accuracy	92.800000	100.0	100.0	100.0

Eventhough it can be observed that we obtained 100 percent accuracy on Decision tree and SCV training and test datasets, we will choose logistic regression since it got good results 80.5% on training and 92.8% on testing datasets, and as it means our model will not suffer overfitting. Also, in this scenario with the datasets we are using binomial outcomes must be favoured.

Conclusion:

In this project, we tried to analyze and make a model to predict the best offer to give a Starbucks customer. First we explored the data and see what I have to change before starting the analysis. Then I did some exploratory analysis on the data after cleaning. In conclusion, the company should give more offers to Females than Males since they have more completed offers. And they should focus more on BOGO and Discount offers since they are the one that tend to make customers buy more.