Analyzing Customer Behavior on the Starbucks Rewards Mobile App

Report on Udacity Data Scientist Capstone Project

1. Project Definition

Project Overview

This project focuses on analyzing simulated customer behavior data from the Starbucks rewards mobile app. The dataset reflects how users interact with various offers sent by Starbucks, which can range from advertisements to actual discounts or BOGO (buy one get one free) offers. Understanding customer responses to these offers is crucial for optimizing marketing strategies and enhancing customer engagement. The data is structured into three main files: `portfolio.json`, `profile.json`, and `transcript.json`, which contain information about offers, customer demographics, and transaction records, respectively.

Problem Statement

The primary problem to be solved is to determine which demographic groups respond best to different types of offers on the Starbucks app. By analyzing customer interactions with offers, the goal is to identify patterns that can inform targeted marketing strategies, ultimately leading to increased customer satisfaction and sales.

Metrics

To measure the effectiveness of the offers and the success of the analysis, I calculated the completion rate, meaning the percentage of offers completed by users, which indicates the effectiveness of the offers. Furthermore, I assessed how different demographic groups (age, gender, income) respond to various offer types. This helps to understand customer behavior and optimizing marketing strategies.

2. Analysis

Data Exploration

The data is contained in three files:

1. portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
    - 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)

2. profile.json - demographic data for each customer

    - 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

3. transcript.json - records for transactions, offers received, offers viewed, and offers completed

    - 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

The data analysis was done with python via jupyter notebook. The exploring process began with data gathering, where the dataset was imported for exploration. During the data exploration phase, various techniques were employed, including checking the dataset's head and shape, as well as identifying any missing values and duplicates.

A function for bar charts was created in order to be able to visualize data more quickly without repeating the code on and on.

```
def create_bar_chart(x_values,y_values,x_label,y_label,title):
    plt.bar(x_values, y_values)
    plt.xlabel(x_label)
    plt.ylabel(y_label)
    plt.title(title)
```

Fig.:1 Function for creating bar charts

The first bar chart shows the distribution of offer types, revealing 4 BOGO offers, 4 discount offers, and 2 informational offers (using data from the portfolio dataset):

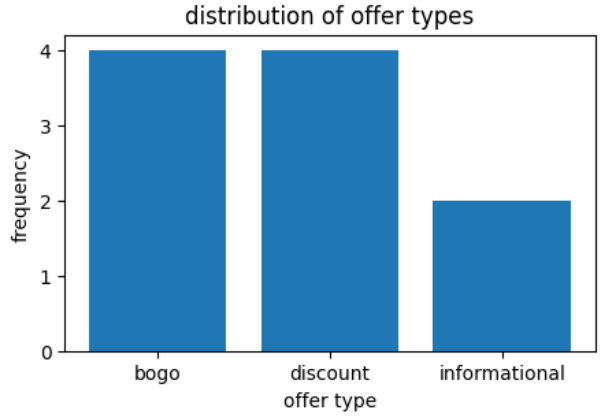

Fig.:2 Distribution of offer types

Additionally, another bar chart from the profile dataset illustrated the gender distribution, showing 50% male, 36% female, and 1% other genders, with 13% of users not providing gender information:

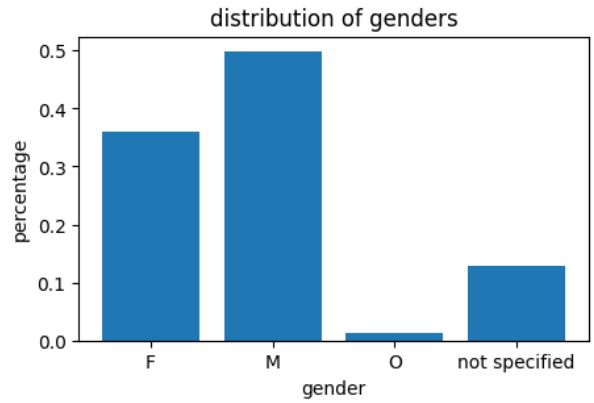

Fig.:3 Distribution of genders

Further analysis included several boxplots and histograms why I created a some more functions:

```
def create_boxplot(values,title,x_ticks,x_label,y_ticks,y_label):
    plt.boxplot(values, patch_artist=True)
    plt.title(title)
    plt.ylabel(y_label)
    plt.xticks(x_ticks, x_label)

def create_histogram(values,bins,title,x_label,y_label):
    plt.hist(values, bins, edgecolor='black')
    plt.title(title)
    plt.xlabel(x_label)
    plt.ylabel(y_label)
```

Fig.:4 Function for creating histograms and boxplots

The examination of the age distribution shows that customer ages range from 18 to 118 years, with a mean age of 62. The graphics below show that there is an outlier in the data (118 years). As only very few people can get this old, the outlier should be looked at in more detail. As there are over 2000 user ids with age of 118 and the difference to the last age before that (101) is quite high, I suppose that 118 is not the real age of these users. Probably this is a default value for those users who did not give information about their age. Furthermore, it is the same number of "missing values" as in the gender column (2175 values). I will further address this issue in the data cleaning section below (chapter 3).

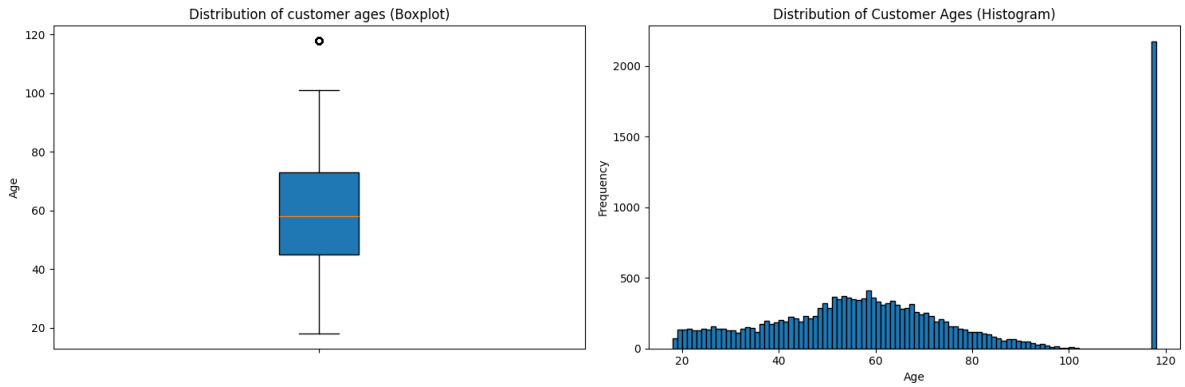

Fig.:5 Distribution of ages

The profile dataset also revealed an income range from $12,000 to $30,000, with a mean income of $65,405, indicating that most users earn between $50,000 and $60,000.

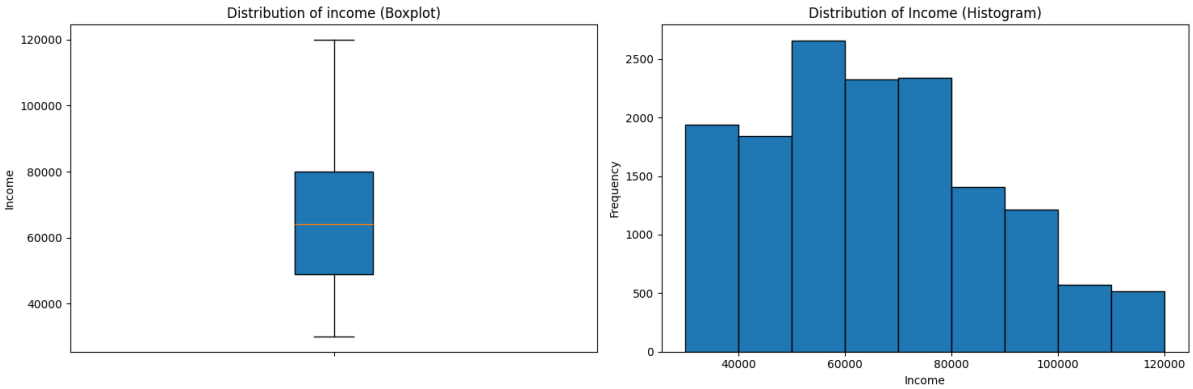

Fig.:6 Distribution of income

3. Methodology

Data Preprocessing

Data preprocessing involved several steps to clean and prepare the data for analysis.

In the portfolio dataset I rename the id column into “offer_id”, calculated reward percentages, and converted the categorical variables “channels” and “offer type” into dummy variables (binary columns with values 0 and 1).

In the profile dataset I renamed the id column into “customer_id”, converted the categorical variable “gender” into dummy variables, handled missing values for gender by imputing “not_specified”, and treated the age outlier of 118 year as a missing value by filling in the mean age.

In the transcript dataset I renamed the “person” column into “customer_id” to make it compatible with the profile dataset. Furthermore I split the column “value” by the ":" and created new columns for offer_id and amount with the respecting value. Then I split the dataset

into two new datasets offers and transactions. Afterwards I converted categorical variable “event” in the new offers dataset into dummy variables (with values of 0 and 1). As I would like to see in one line if an offer was seen, received and viewed by a specific user, I created an identifier for the user and offer combination and then aggregated the dataframe. Also, I aggregating the transactions dataset to see which user spent how much in total.

Then, I merged offers and transactions with each other again and added the information from profile and portfolio to get one clean new dataset “df”. Furthermore, as there is no effect of an offer if it was not viewed, I reduced the df dataset and took only the rows with “offer_viewed != 0” into account.

Implementation

The implementation phase involved creating a correlation matrix to identify relationships between variables in the first step. This was done with the pearson correlation coefficient:

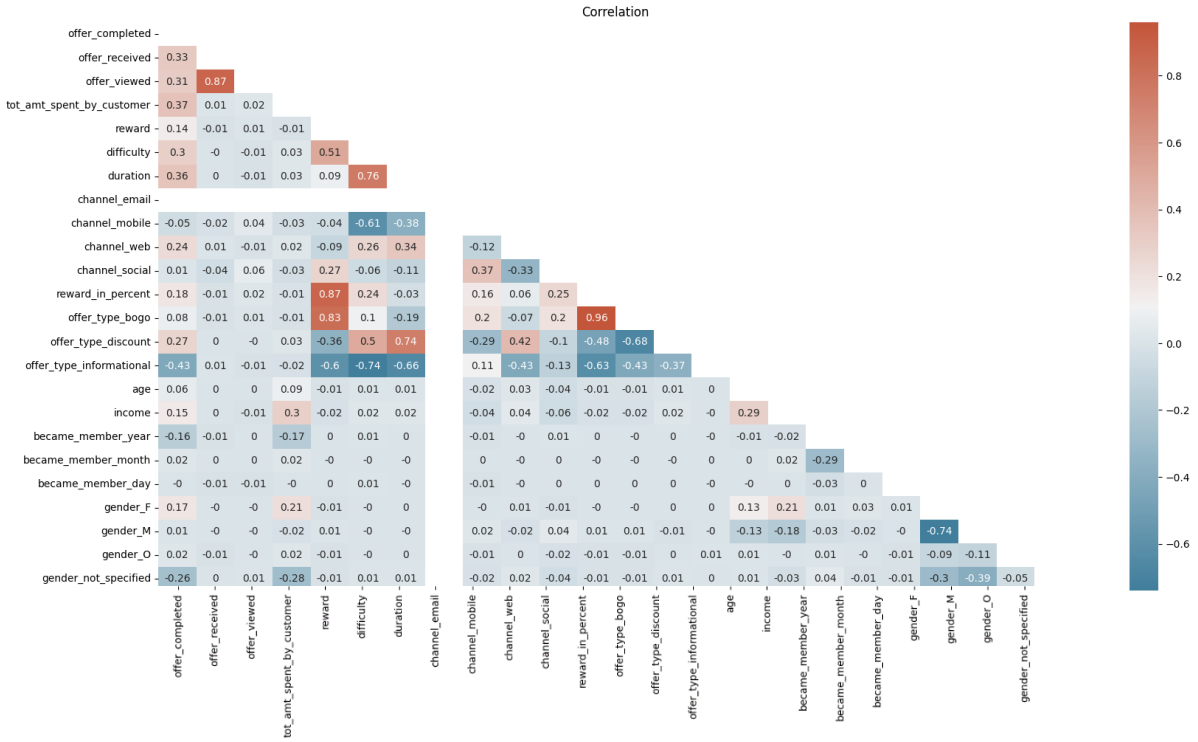

Fig.:7 Correlation matrix

However, the correlation matrix did not reveal significant correlations that could lead to conclusions about demographic groups. Consequently, heuristics and data visualization techniques were employed to further analyze the data.

I created bar charts to look at the interesting categorical variables, which should show which offer types, channels and genders have best completion rate. The completion rate was calculated by the number of completed offers divided by the number of viewed offers.

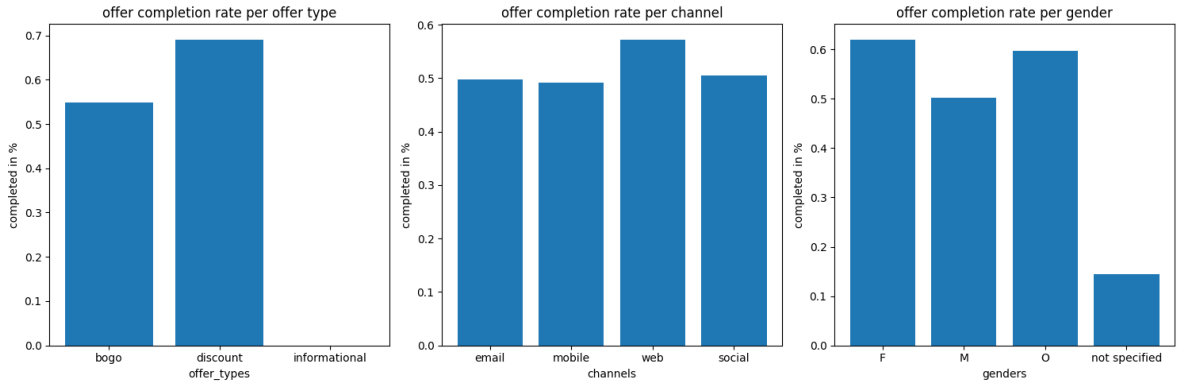

Fig.:8 Effect of categorical variables on completion rate

Over all, discount offers have the highest completion rate (69%). Offers that were sent by the web have a slightly higher completion rate than the other channels (57% vs. about 50%). Female customers have the highest completion rates (62%). However, other genders are following straight behind (60%).

Additional bar charts were generated to explore completion rates by offer types and channels for the different genders.

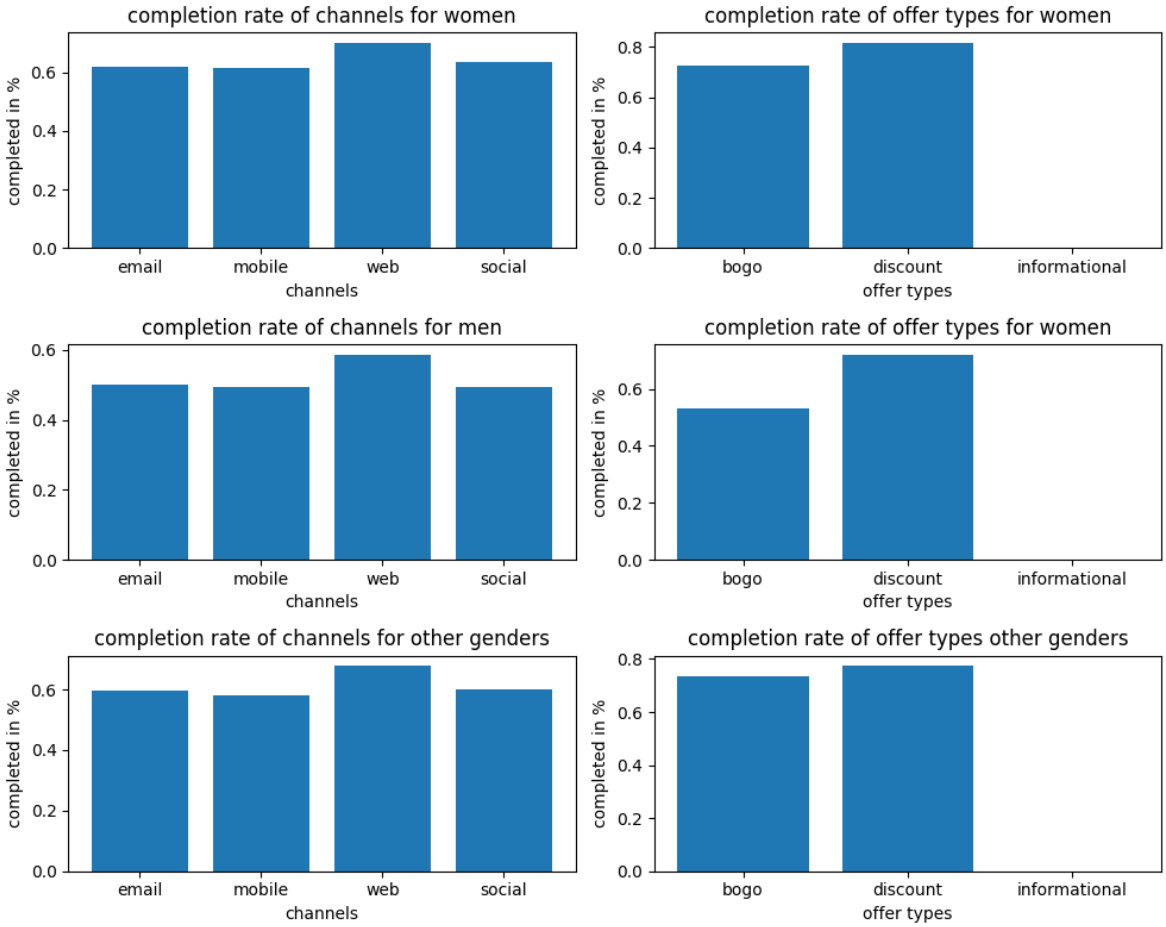

Fig.:9 Effect of categorical variables on completion rate segmented by gender

The findings showed no significant differences in completion rates across channels when segmented by gender, although the difference in completion rates between BOGO and discount offers was most pronounced among male customers.

To look at the numeric variables I created several functions for plotting line graphs (see jupyter notebook). I analyzed which reward percentages, ages, and incomes correlate with the completion rate of offers.

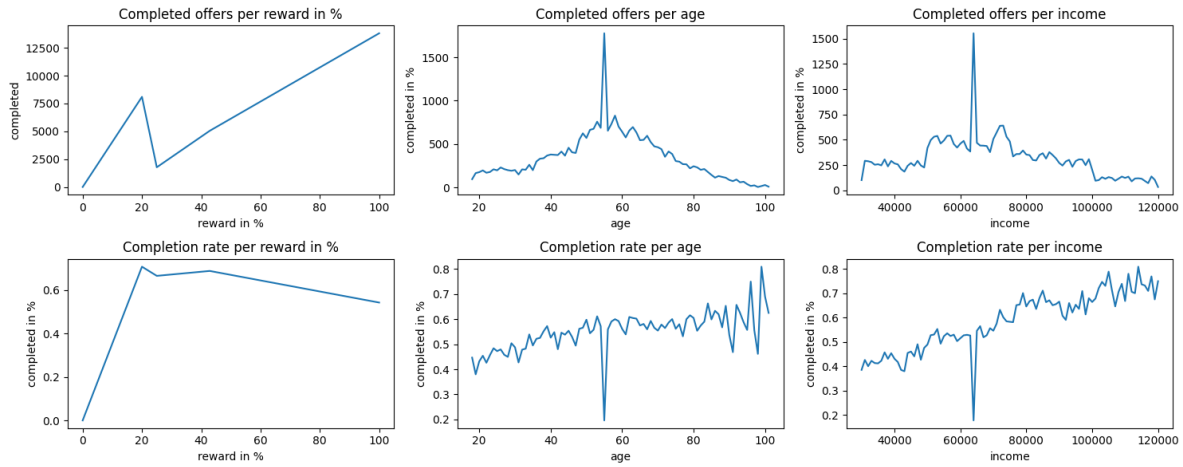

Fig.:10 Effect of numeric variables reward in %, age and income on completion rate

A strong relationship was observed between the total number of completed offers and reward percentages, although completion rates slightly decreased with higher rewards. Completion rates increased with both age and income, but the total number of completed offers decreased as income and age rose due to a declining customer base.

The same analyses were conducted separately by gender, revealing that the increase in completion rates with income was more pronounced for men than for women, while no trend was observed for other genders (graphs see jupyter notebook). When examining the data by offer type, it was found that the increase in completion rates with income and age was stronger for BOGO offers compared to discount offers.

After looking at all offers in general, I want to look at the different offer ids in detail. A bar chart was created to provide an overview of which offers had the best completion rates, indicating that the two informational offers had no completion events recorded.

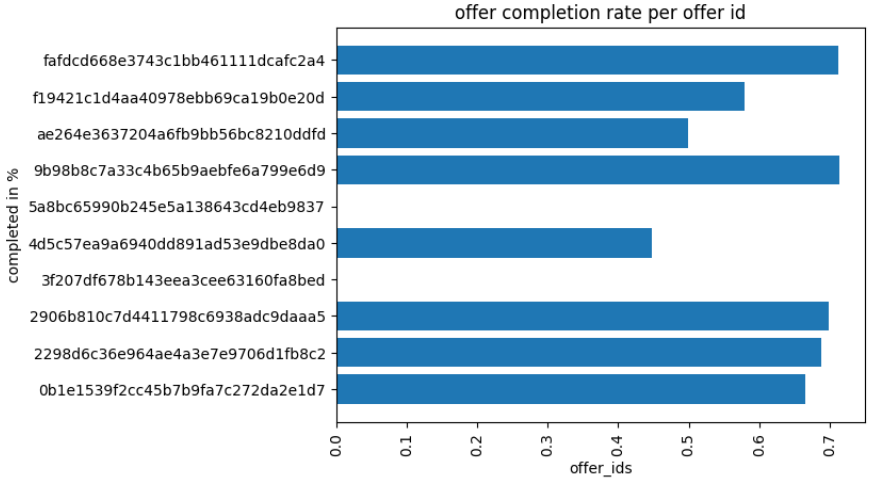

Fig.:11 offer completion by offer id

Finally, bar and line charts were generated to analyze the completion rates by gender, age, and income for each offer separately. This is an example for the first offer table (I simplified the offer ids, see translation table Fig.:14):

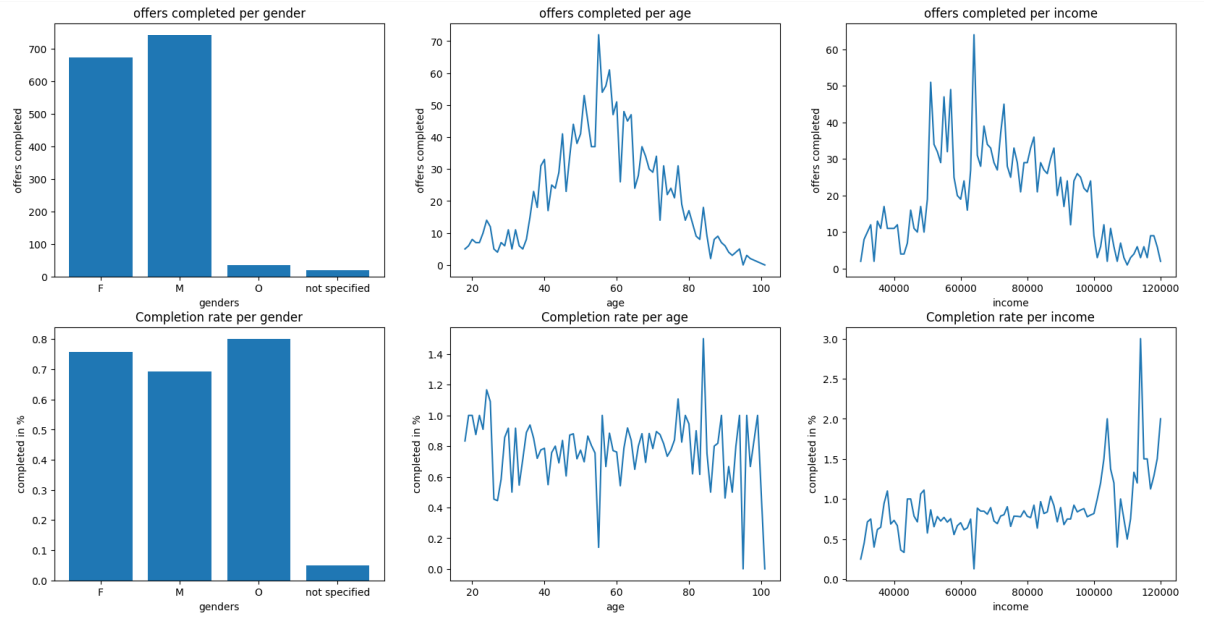

Fig.:12 completion rates by gender, age, and income

I summarized the results for the offers in this table, evaluating the overall trends found before:

offer	gender trend	age trend	income trend
1	Other genders are the best customers for this offers when it comes to completion rate. However, the number of other genders is quite low.	The effect of increasing completion rate with increasing age cannot be observed here.	The graph shows completion rates over 100% which means that some customers completed offers more often than they viewed them. This distorts the result. When looking only at the completion rates below one, there is no trend recognizable.
2	Female customers are the best customers for this offer when it comes to completion rate.	The effect of increasing completion rate with increasing age cannot be observed here.	The graphs show an increase of completion rate with increasing income.
3	Female customers and other genders are the best customers for this offer when it comes to completion rate. However, again other genders have a quite low overall number of completed offers.	The effect of increasing completion rate for increasing age can only be observed for very high ages.	Concerning income, the graphs show a slight increase of completion rate with increasing income.

4	Female customers and other genders are the best customers for this offer when it comes to completion rate. Here the difference to mal customers is very high (about 25 percentage points). Again the overall number of other genders is very low.	The effect of increasing completion rate with increasing age and income can be observed here, while the trend is much stronger for income.	The effect of increasing completion rate with increasing age and income can be observed here, while the trend is much stronger for income.
5	Female customers are the best customers for this offer when it comes to completion rate	The effect of increasing completion rate with increasing age cannot be observed here.	Concerning income, the graphs show an increase of completion rate with increasing income.
6	Female customers and other genders are the best customers for this offer when it comes to completion rate. Here the difference to male customers is very high (about 25 percentage points). Again the overall number of other genders is very low.	The effect of increasing completion rate with increasing age and income can be observed here, while the turned is much stronger for income.	The effect of increasing completion rate with increasing age and income can be observed here, while the turned is much stronger for income.
7	Female customers and other genders are the best customers for this offer when it comes to completion rate. Here the difference to male customers is relatively high (about 15 percentage points). Again the overall number of other genders is very low.	The effect of increasing completion rate with increasing age and income can be observed here.	The effect of increasing completion rate with increasing age and income can be observed here.
8	Female customers are the best customers for this offer when it comes to completion rate.	The effect of increasing completion rate with increasing age and income can be observed here.	The effect of increasing completion rate with increasing age and income can be observed here.

Fig.:13 completion rates by gender, age, and income

offer id	simplification
0b1e1539f2cc45b7b9fa7c272da2e1d7	Offer 1
2298d6c36e964ae4a3e7e9706d1fb8c2	Offer 2
2906b810c7d4411798c6938adc9daaa5	Offer 3
3f207df678b143eea3cee63160fa8bed	No completion rate (as only informational offer)
4d5c57ea9a6940dd891ad53e9dbe8da0	Offer 4
5a8bc65990b245e5a138643cd4eb9837	No completion rate (as only informational offer)
9b98b8c7a33c4b65b9aebfe6a799e6d9	Offer 5
ae264e3637204a6fb9bb56bc8210ddfd	Offer 6
f19421c1d4aa40978ebb69ca19b0e20d	Offer 7
fafdcd668e3743c1bb461111dcafc2a4	Offer 8

Fig.:14 translation table for offer ids

Refinement

The graphics were continuously adjusted, refined and summarized to give the best possible overview.

4. Results

Model Evaluation and Validation

The analysis revealed that discount offers had the highest completion rates, offers sent by the web got the highest completion rates and female customers showing the highest engagement. Furthermore, the completion rates increase with age and income, particularly for BOGO offers. When looking at every offer in details these trends might not be revealed for every offer.

Justification

The final results indicated the correlation matrix (with pearson correlation coefficient) could not help to answer the main question of this project. That's why data visualizations were used. The result captured from that were showed by several graphics and were compared by tables.

5. Conclusion

Reflection

The project successfully identified key insights into customer behavior on the Starbucks rewards app, demonstrating the importance of tailoring offers to specific demographic groups. One interesting aspect was the significant impact of income on offer completion rates, which was more pronounced for male customers compared to female customers.

Improvement

Future research could focus on incorporating more complex models using machine learning and real-world data to further refine the analysis. Additionally, exploring the impact of external factors, such as seasonal promotions or economic conditions, could provide deeper insights into customer behavior and preferences.