Starbucks

Introduction

the given data set contains simulated data that mimics customer behaviors on the Starbucks rewards mobile app

In this project, I aim to address the following two questions:

1. To whom should Starbucks send offers?
2. How much someone will spend based on demographics and offer type? Will they use the offer? If so, how much would they spend?

The data provided consists 3 datasets:

- **portfolio**—containing offer ids and meta data about each offer (duration, type, etc.)
- **profile**—demographic data for each customer (age, income, gender, etc)

- **transcript**—records for events, i.e. transactions, offers received, offers viewed, and offers completed.

The project is carried out in the following steps:

1. **Data Preprocessing**
2. **Data Preparation**
3. **Data Exploration**
4. **Model Implementation**
5. **Conclusion**

. . .

Data Preprocessing

portfolio

- *id* (string)—offer id
- *offer_type* (string)—type of offer, ie bogo (buy one get one free), discount and informational offer
- *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)—methods that the offer been sent

This is a very small dataset containing 10 offers. There is no missing values and only the channels column needs preprocessing:

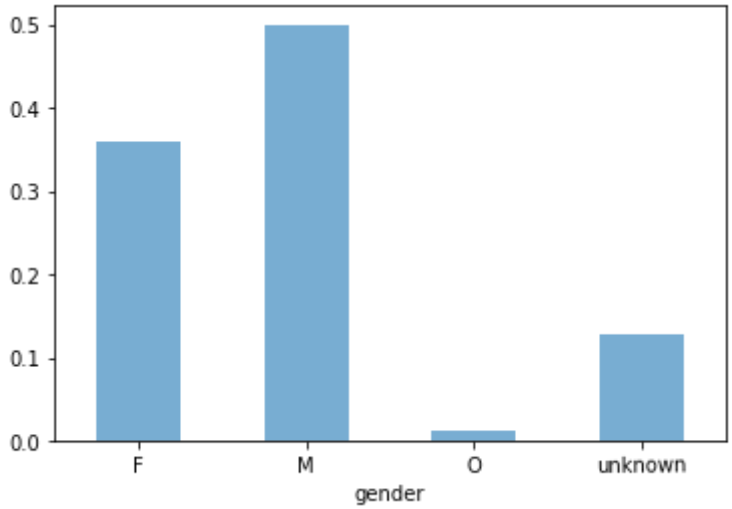
- *channels* is expanded into four categorical variables i.e. *email*, *mobile*, *social* and *web*. Each of them holds 0s and 1s.
- A new feature is engineered as the number of methods the offer been set.

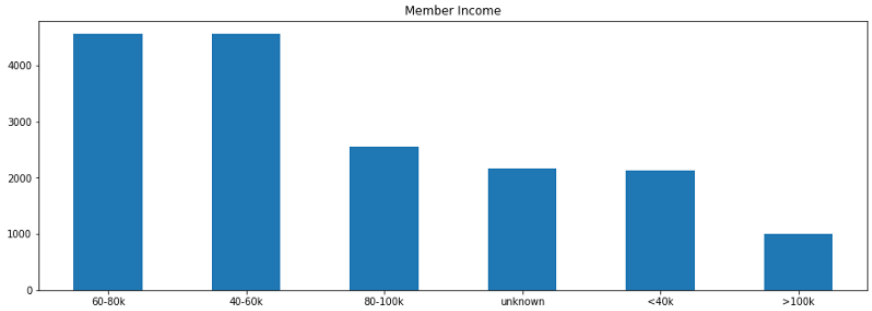
profile

- *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

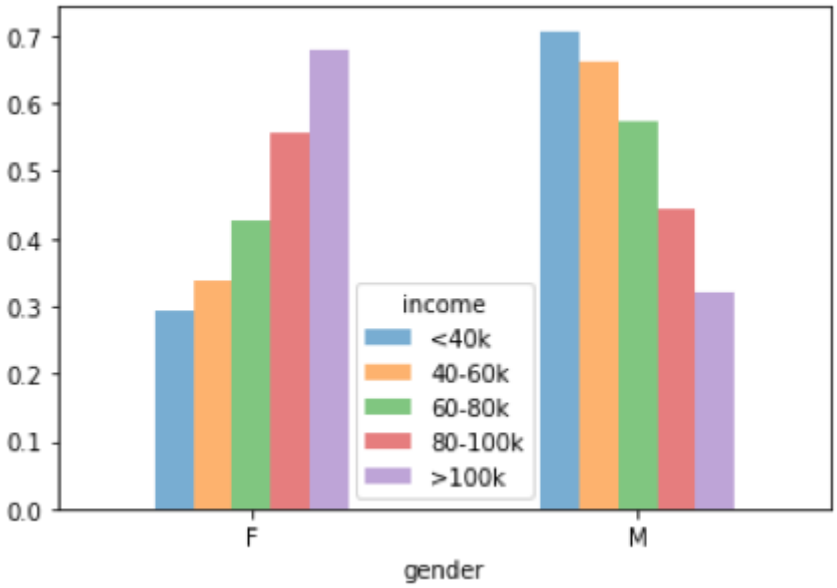
After taking a closer look at the dataset, there are several issues need to be addressed:

- There are about 12.79% missing values for *gender* and *income*. NULL in *gender* is filled with “unknown”. Since 12.79% is a relatively large percentage to be filled with median or mean of the income, so it makes more sense to turn *income* into categorical data. i*ncome* is grouped into 5 bins (‘<40k’, ‘40–60k’ ,’60–80k’, ‘80–100k’,’>100k’). Missing incomes are filled as “unknown”.

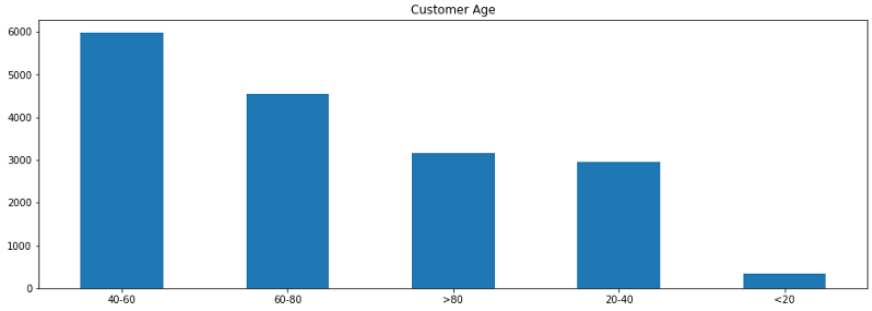




customer incomes and counts



- There is not any missing values in *age*, but 12.79% of the values are 118. These values are replaced with “unknown” with the rest numerical data been grouped into 5 bins (‘<20’, ‘20–40’, ‘40–60’ ,’60–80', ‘>80’).



customer ages and their counts

- Feature engineering to turn *became_member_on* into *membership_days* indicating number of days a customer has been a membership.

transcript

- *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 or reward amount depending on the record

Preprocessing of this dataset includes:

- expanding *value* column into three categorical columns: *offer id*, *amount* and *reward*. Fill missing values in *offer id* with "unknown" and fill missing values in *amount* and *reward* with 0.
- Associate transactions with an offer id: upon initial investigation, `transaction` events do not have *offer id* associated with them. So an assumption was made that if a `transaction` follows an `offer viewed` event, it is considered to be potentially influenced by the offer. So *offer id* of the `offer viewed` event is assigned to *offer id* of the `transaction` event.

Moreover, initial investigation shows that BOGO and discount offers have an `offer completed` event while informational offers do not have this event.

. . .

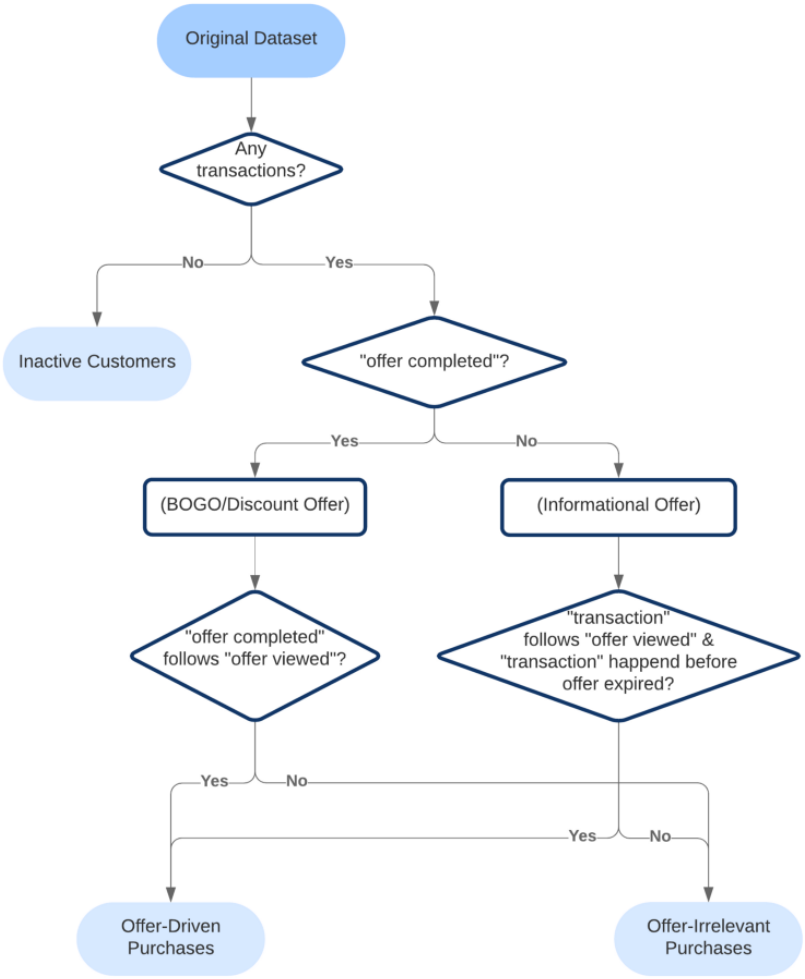
At this point, the three dataset can be joined together to get a complete dataset holding all transaction informations including transaction details, offer information and customer information. The combined dataset is used from this point on.

Dataset Preparation

As shown in the flow chart below, three features are engineered in this section:

- *inactive customer*: if a customer does not have any transaction event associated with him/her.
- *offer_driven*: purchases that are inspired or driven by an offer received
- *offer_irrelevant*: purchases that are not inspired or driven by an offer received.

Finding all *offer_driven* and *offer_irrelevant* events are the trickiest part in dataset preparation.



work flow for engineered features

offer_driven

- For BOGO and discount offers, if an `offer completed` happens after an `offer viewed` event, then it is considered to be an offer driven purchase.
- If a `transaction` happens after an `offer viewed` event and the `transaction` happens within the window of an offer, then it is considered to be an offer driven purchase. More specially, this includes two types of scenarios: 1) For informational offer, meeting this requirement means that the offer is completed; 2) for BOGO/discount offer, as long as the transaction happens after offer view and within offer valid window, it is considered to be offer driven in this study, even though the transaction does not meet transaction amount requirement, i.e. no `offer completed` event follows `transaction` .

offer_irrelevant

- If an `offer completed` does not follows an `offer viewed` event (for BOGO and discount offers), it indicates that a customer made a purchase without knowing the offer. So even though the offer is completed, it is still categorized as an offer irrelevant purchase.
- If a `transaction` does not happen after an `offer viewed` event, then the transaction is considered to be a standalone transaction which means that it has nothing to do with offer. This type of transaction does not have an *offer id* associated with it.
- If a `transaction` happens after an `offer viewed` event, but the `transaction` happens after an offer expires, then it is considered to be an offer irrelevant purchase since the customer does not feel the influence of the offer at the time of purchase. This is true for all three types of offers.

Dataset Cleaning

One complete event sequence may be associated several rows, for example:

- there could be an `offer viewed` alone indicating the offer is never been viewed or completed.
- an offer can be received and viewed, but never be redeemed. In this case, a complete event sequence is `offer received` → `offer viewed`
- also there could be a standalone `transaction`
- there could be `offer received` → `offer viewed` → `transaction` or `offer received` → `offer viewed` → `transaction` → `offer completed`
- there could also be cases like `offer received` → `transaction` or `offer received` → `transaction` → `offer completed`
- etc…

The purpose of cleaning is to integrate all information within a complete event sequence into one record without losing any information. More specifically, the cleaning is carried out in the following steps:

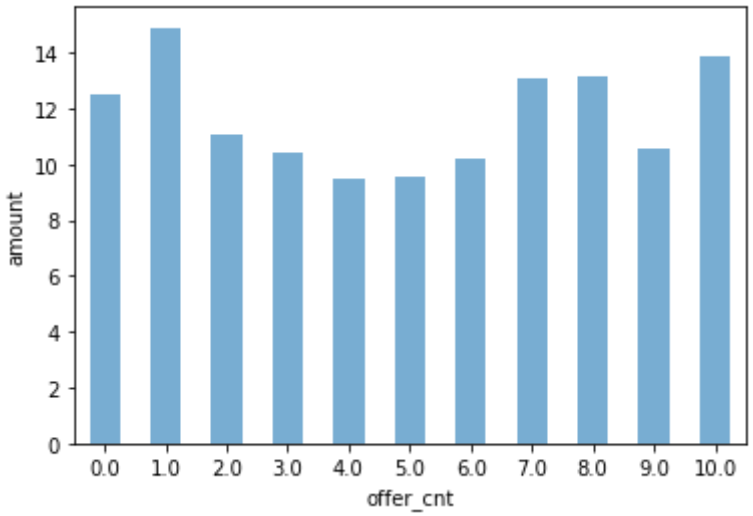
- Each customer may received multiple offers. No matter it is redeemed or not, it is a history of `offer received`. So all `offer received` event should be kept.
- There may be standalone `transaction` without any offers. All of them don't have an *offer id* associated with them. They will need to be kept.
- Information held in `transaction` and `offer received` will be copied to their corresponding `offer received` event and only the `offer received` will be kept.

More details about dataset cleaning can be found in this git repo.

. . .

Data Exploration

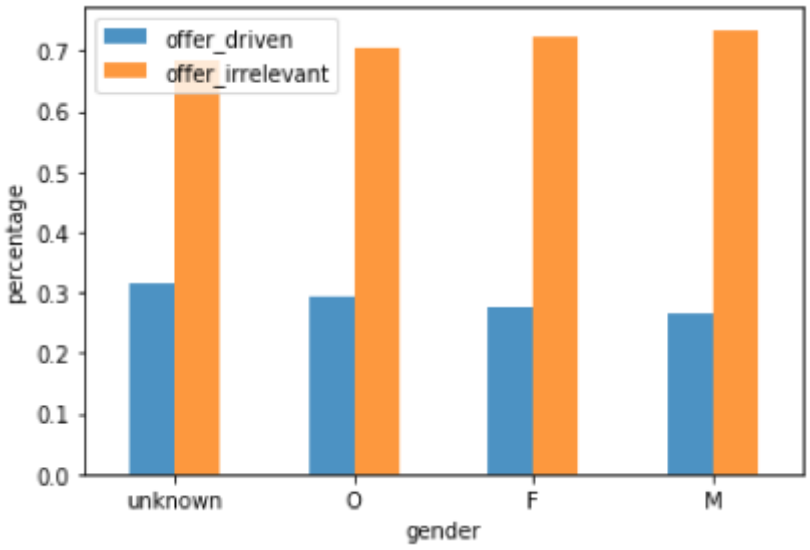
Q: Does the average amount of transaction influenced by the number of offers received?



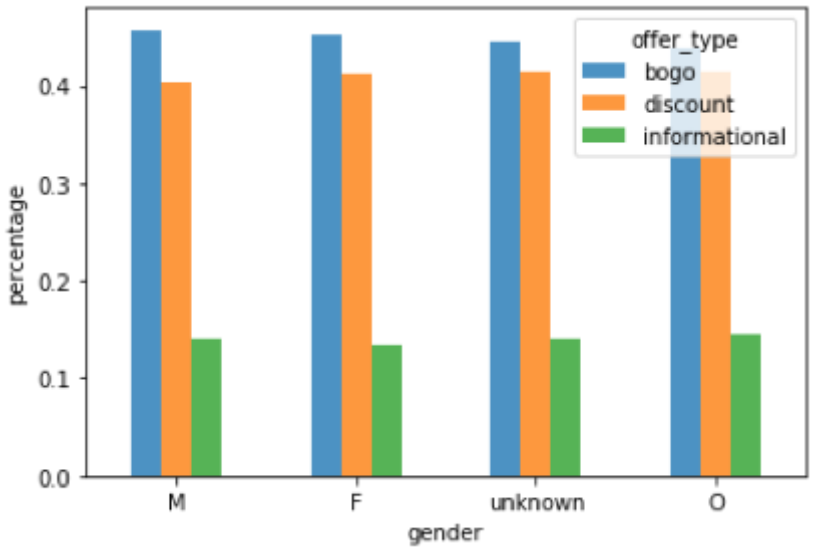
average transaction amount V.S. number of offer received

Average amount of transaction seems to be highest for customers receiving only 1 offer. Customers receiving 4–5 offers have the lowest average transaction amount.

Q: How each gender response to offers?

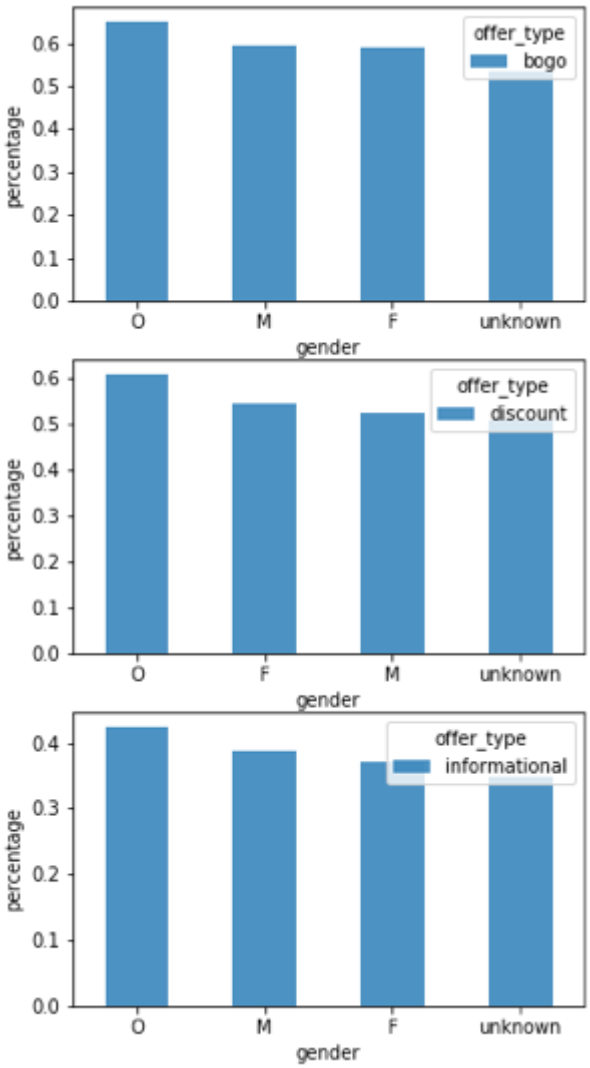


percentage of offer driven and offer irrelevant purchases within each gender

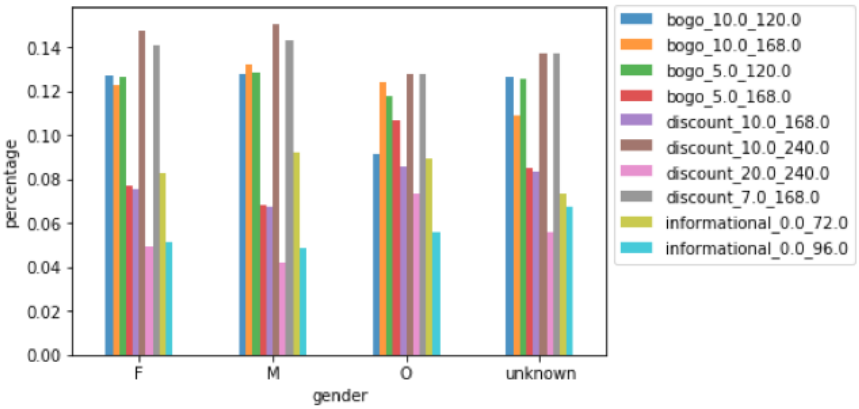


For all offer-driven purchases, what's the response rate of different types of offers?

Figures above show that different gender groups have similar purchasing behaviors and similar responses to the three types offers. It can be seen that only around 10% of offer driven purchases are based on informational offers. Does that mean informational offer is not as effective as bogo and discount offer? *For all offer-driven purchases, how many percentages of offer sent leads to an effective purchases?*

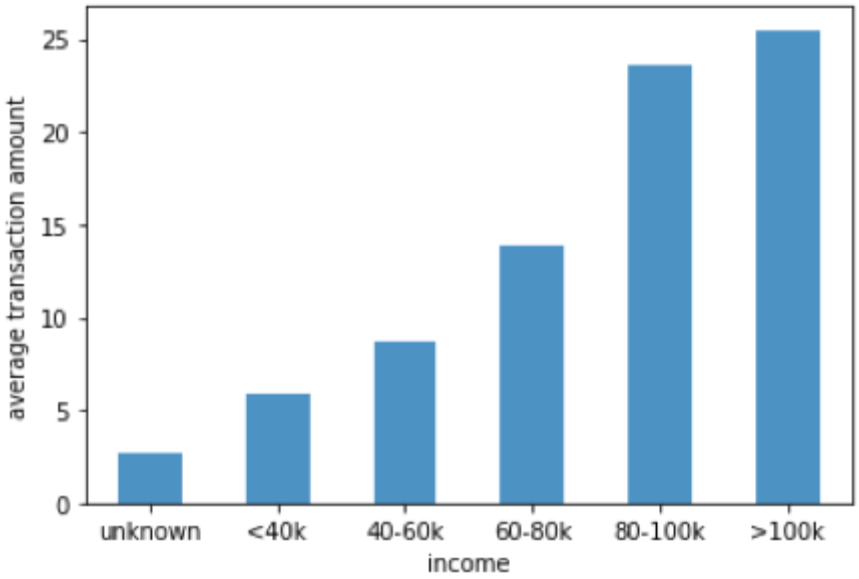


For the figure above, it can be seen that for all genders, the response rate for bogo offer is higher than that for discount offer while the response rate for informational offer is the lowest. There are only slightly differences among each gender. To take a closer look, figure below presents the response rate for each specific offer. It can be seen that gender O is not very responsive for “bogo_10.0-_120.0” while it is most responsive to offer “bogo_5.0-_168.0”. More information can be obtained from this plot.



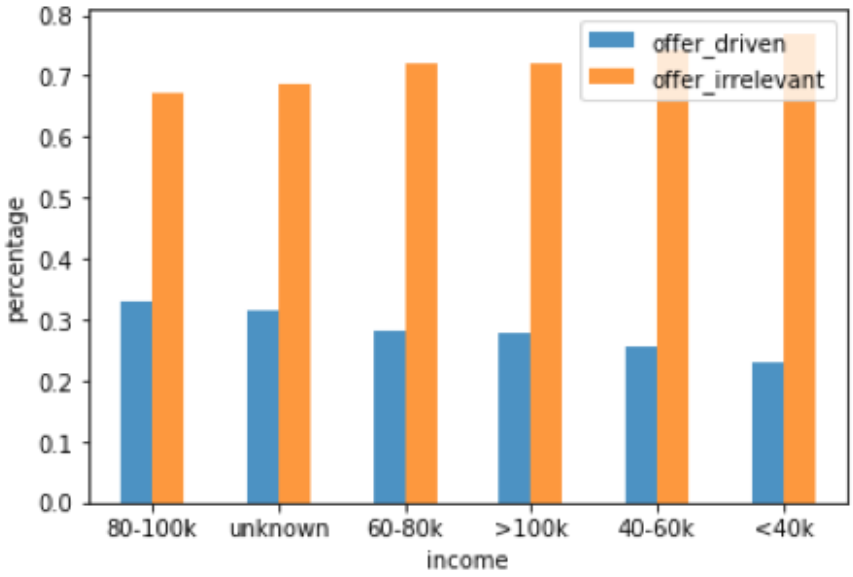
Q: How does income influence purchasing behaviors?

Figure below shows that the higher the income is, the higher the average transaction amount is.



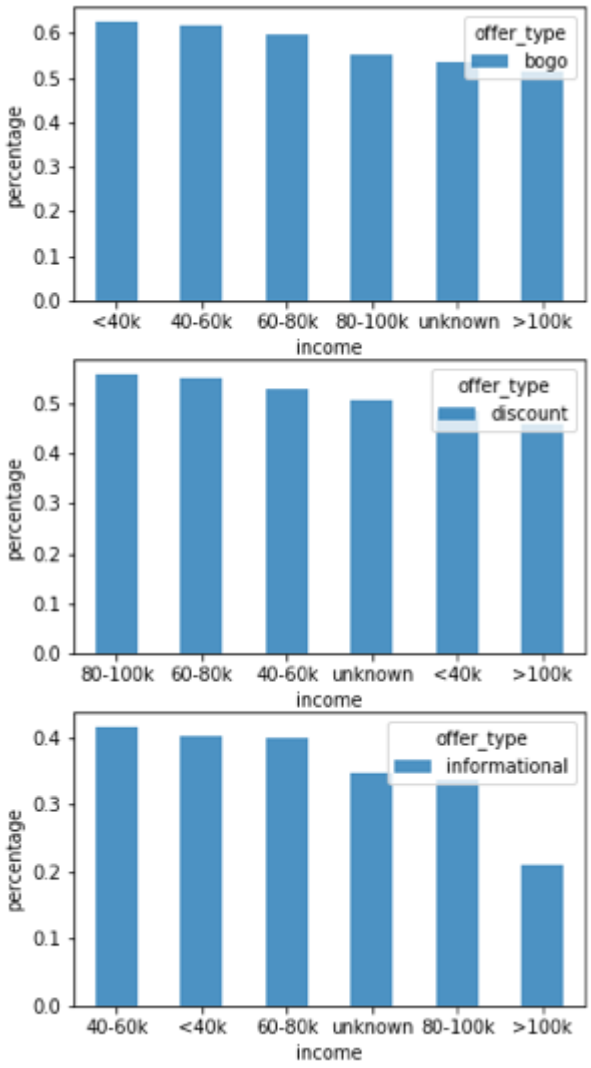
How many percentages of purchases are driven by offers, for each income range? Which income level is most responsive to offers? And what type of offers?

Figure below shows that customers with income between 80–100k is most responsive to offers as they have the highest percentage of purchases driven by offer. Customers with salaries between 40k are cares about offers the least.

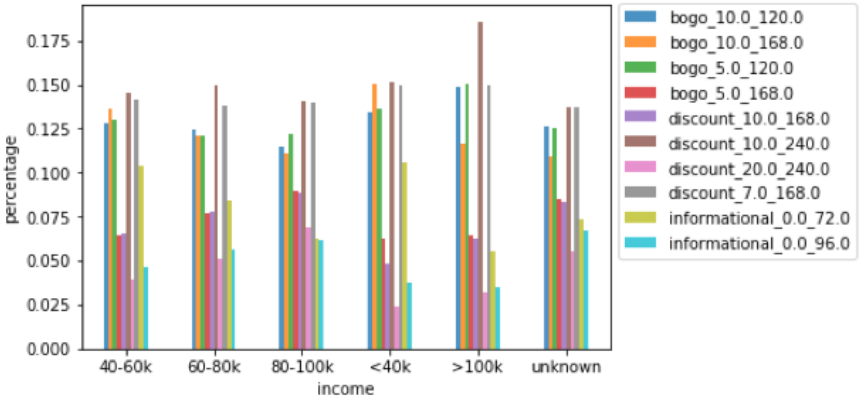


For each offer type, how many percentages of offer sent leads to an effective purchases?

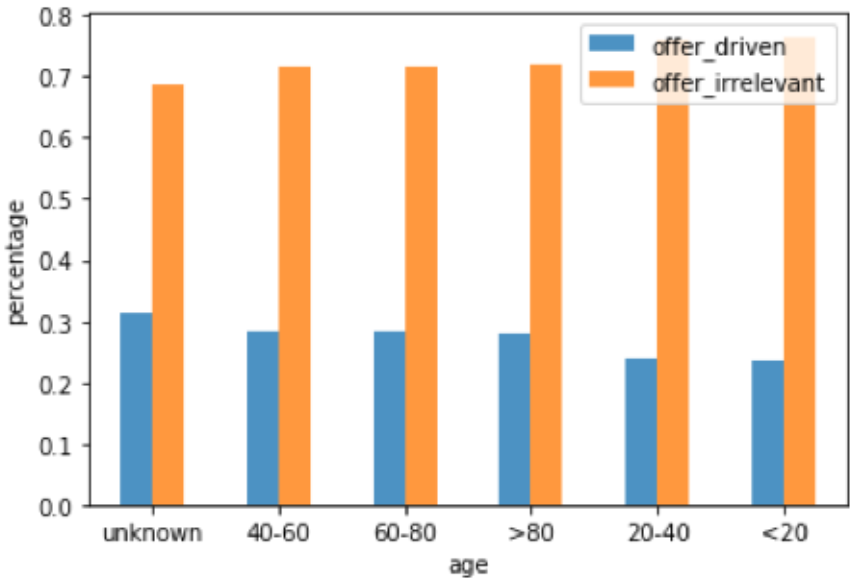
From figures below, it can be seen that customers with salary less than 40k is most responsive to bogo offers. Customers with salary greater than 80k does not like informational offers.

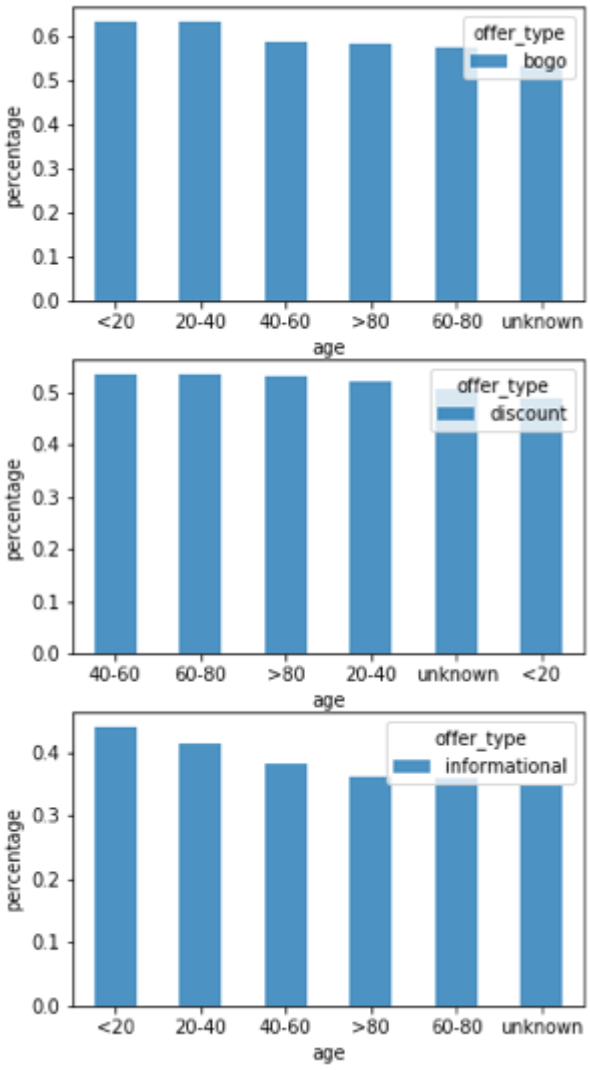


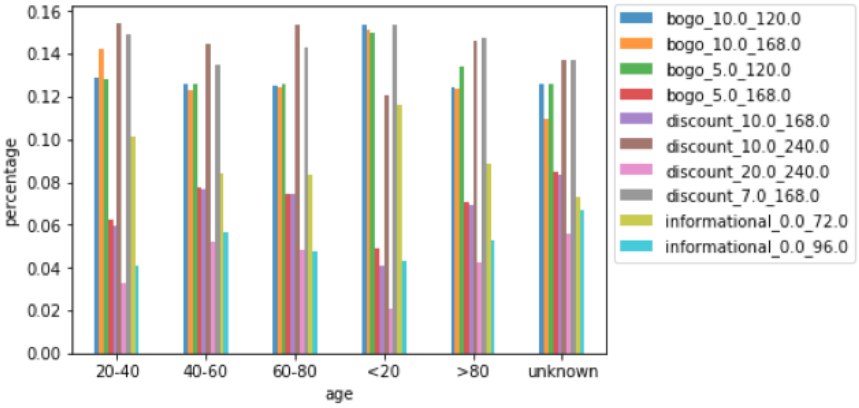
To take a closer look, response rates to each offer is plotted for each income level. We can see that offer “discount_10.0_240” and “discount_7.0_168” are very well received for all income levels.



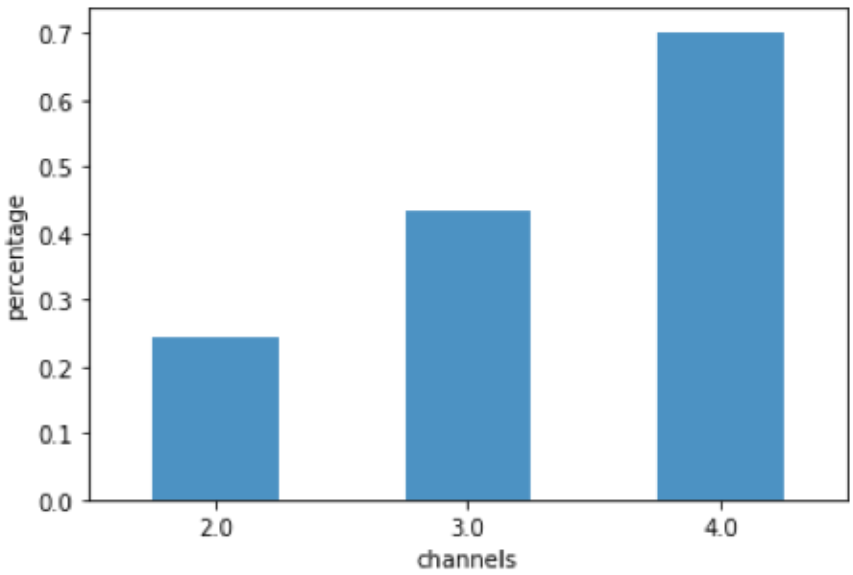
Q: How does age influence purchasing behaviors?

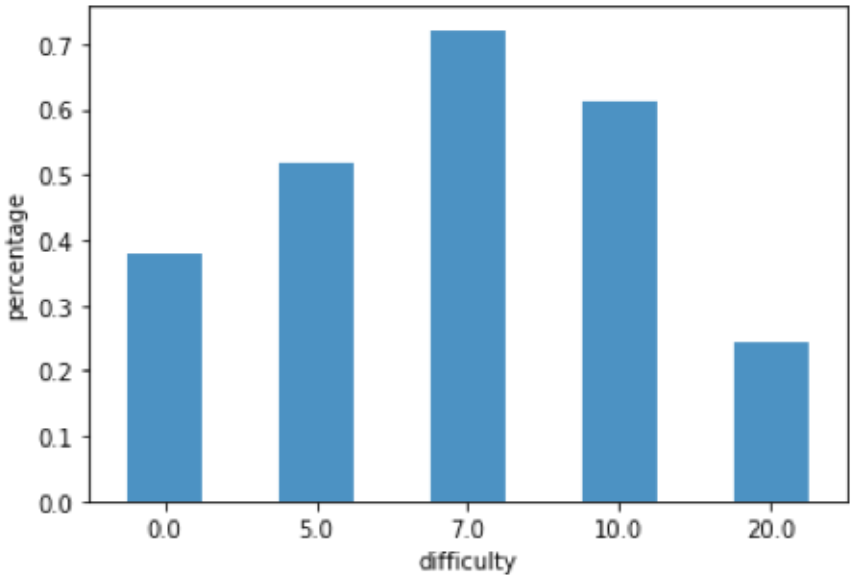


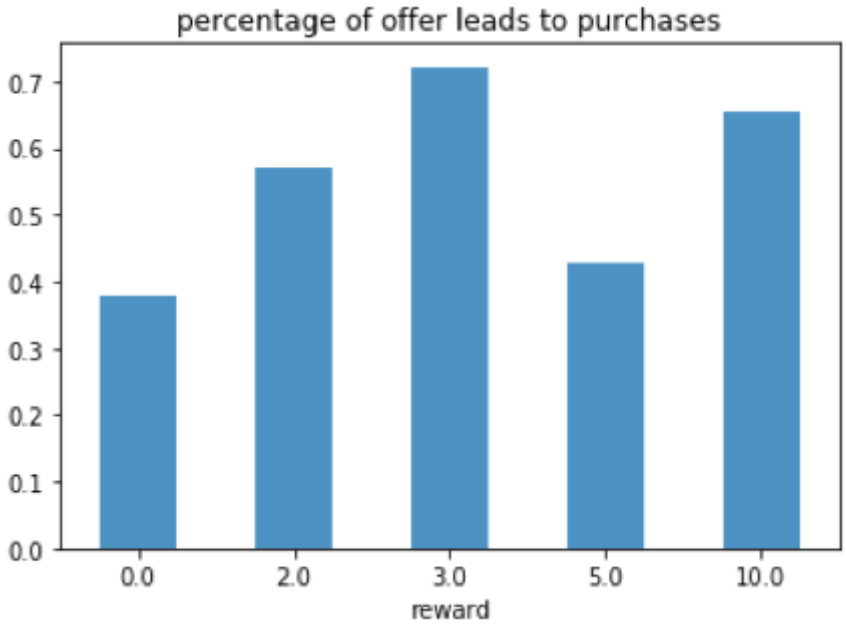




Q: How offer could influence customers responses to offers.


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
percentage
2.0
3.0
4.0
channels


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
percentage
0.0
5.0
7.0
10.0
20.0
difficulty


percentage of offer leads to purchases
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.0
3.0
5.0
10.0
reward