Starbucks Capstone Challenge¶

Introduction

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely 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.

Not all users receive the same offer, and that is the challenge to solve with this data set.

Your task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

You'll be given transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

Keep in mind as well that someone using the app might make a purchase through the app without having received an offer or seen an offers

The following steps I have made to data set

importing Libraries

starting by uploading necessary data.

Reading Files

and then reading 3 files of data we have.

- portfolio
- profile
- transcript

***Cleaning up of this portfolio dataset*¶**

1- changing the name of id to offer_id as we need to make it as primamry key and we will use use it to join other dataset.

2- making a one hot encoded with channel and sperate every channel .

3- making one hot encoding to offer_type and drop offer type

Cleaning up of this profile dataset :

1- removing outlier from age .

2- change "became_member_on" datatype to be datetime .

3- changing column id to customer_id .

4- check null information and drop column with null .

5- create age_range from age

6- create days_as_member from become a member

Cleaning up of this transcript dataset :

1- change a person to Customer_id .

2- change value be "offer id " and remove str bepfre it .

3- make one hot code for event value.

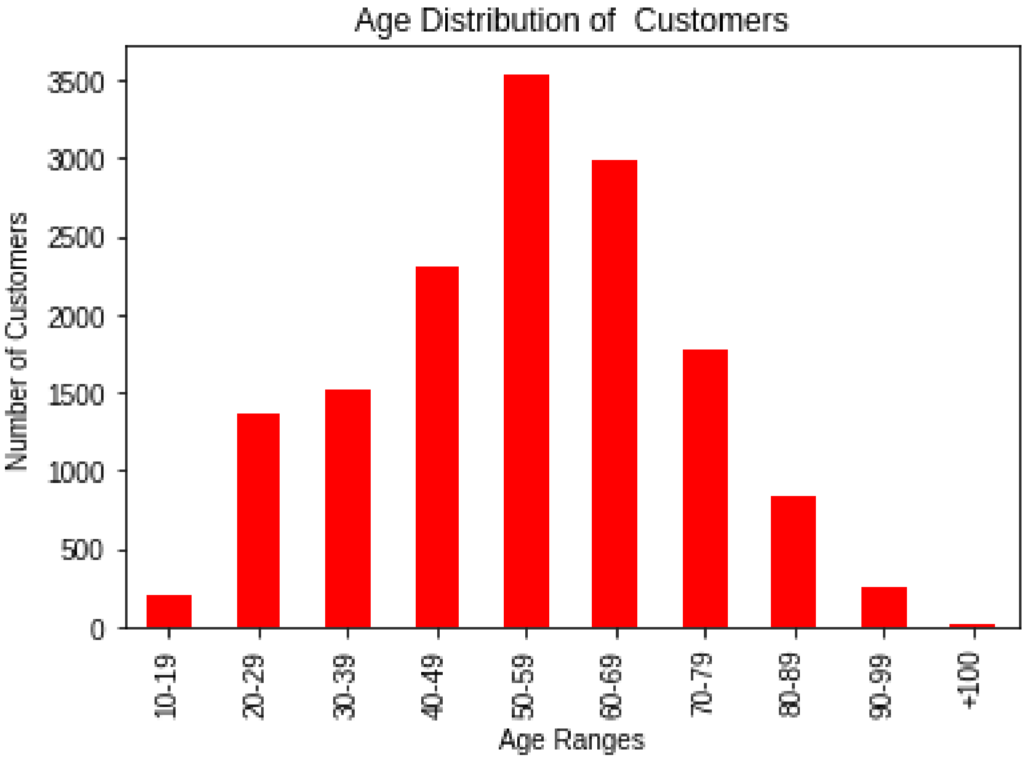
4- change time to days

Data Exploration

Trying to generate question and its answer from data point of view

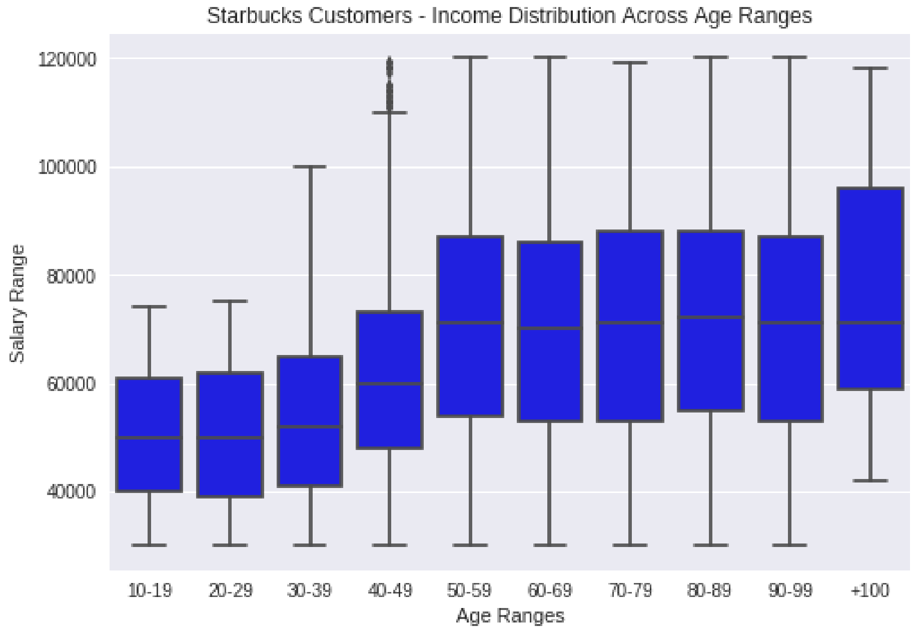
what is distribution of customer range of age ?

customer age take the bell curve for the age range what is called normal distribution mean = median = mode

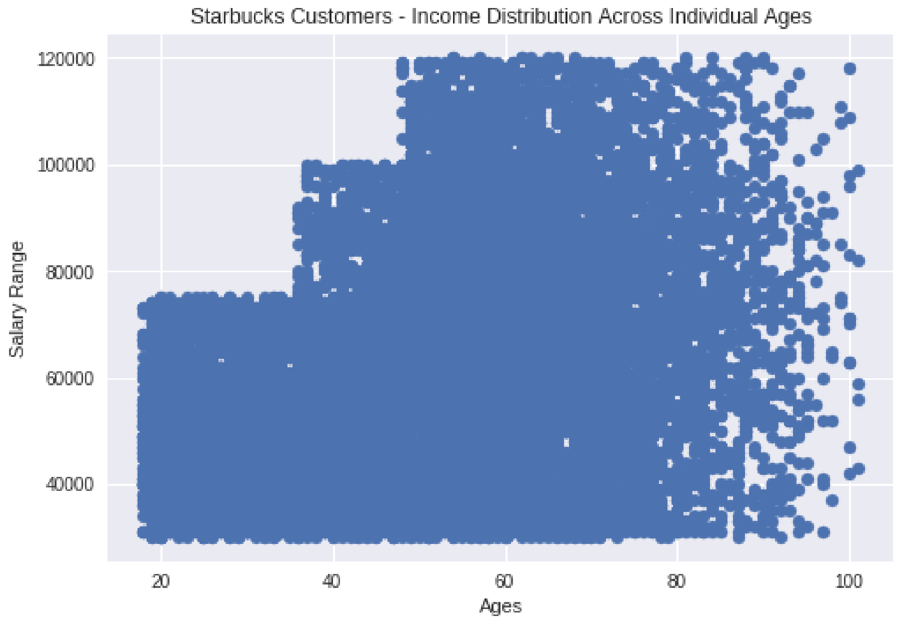


what the distribution of income with age_range ?

it shows that : for young age range income is low and range of icome not hight and thus understood , also income increase by age_range increae also range of icome increase untill reach to the age range more than 100 years old



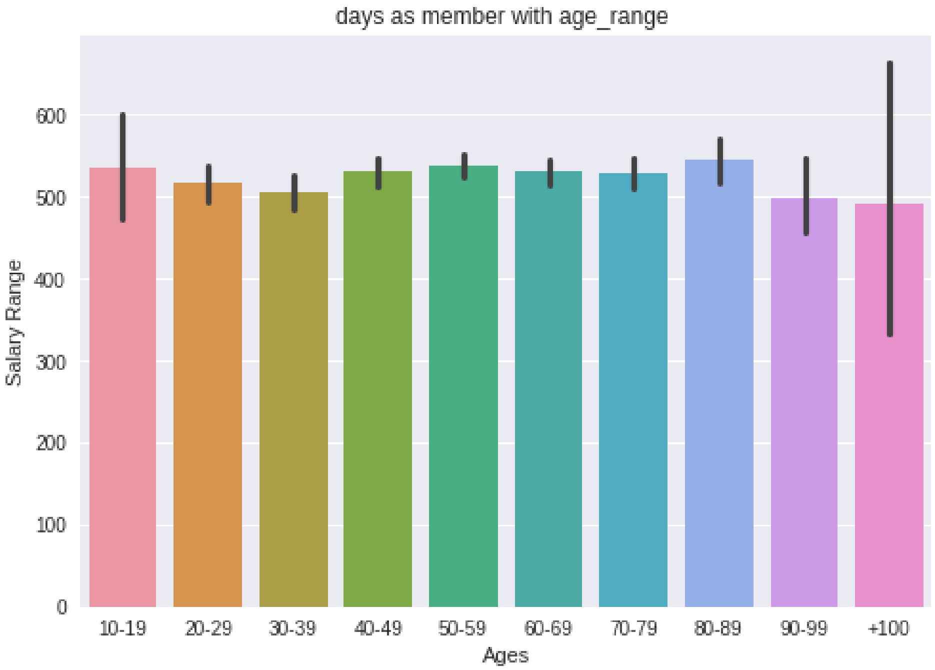
Starbucks Customers - Income Distribution Across Age Ranges

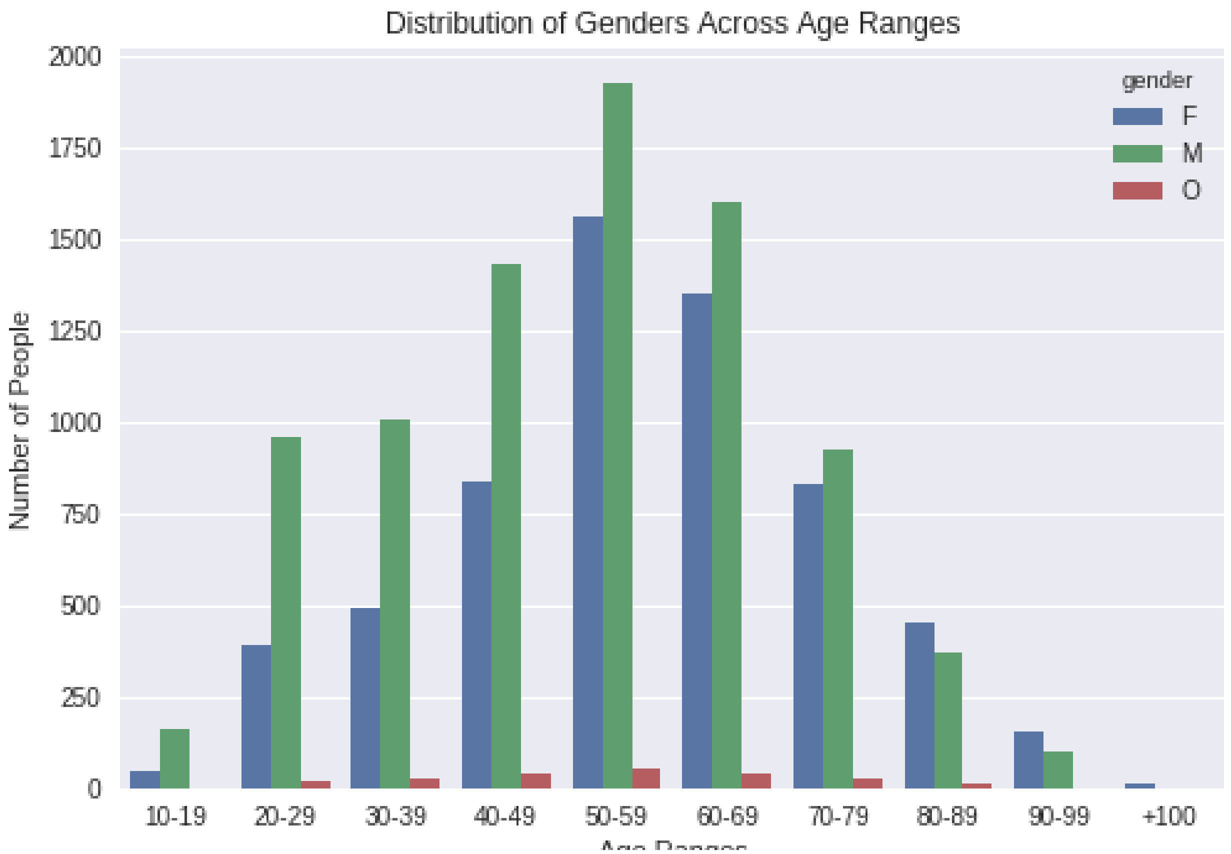


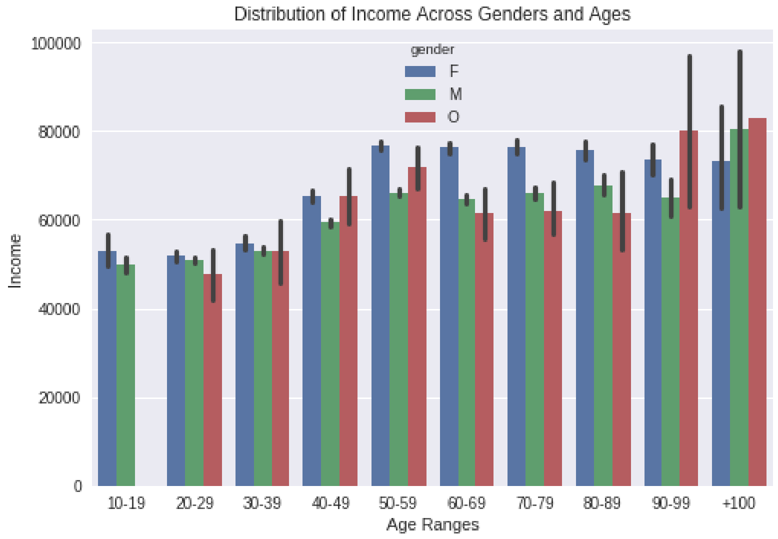
Starbucks Customers - Income Distribution Across Individual Ages

at what age range they will keen to keep their member ship ?

we found approximatly all is the same

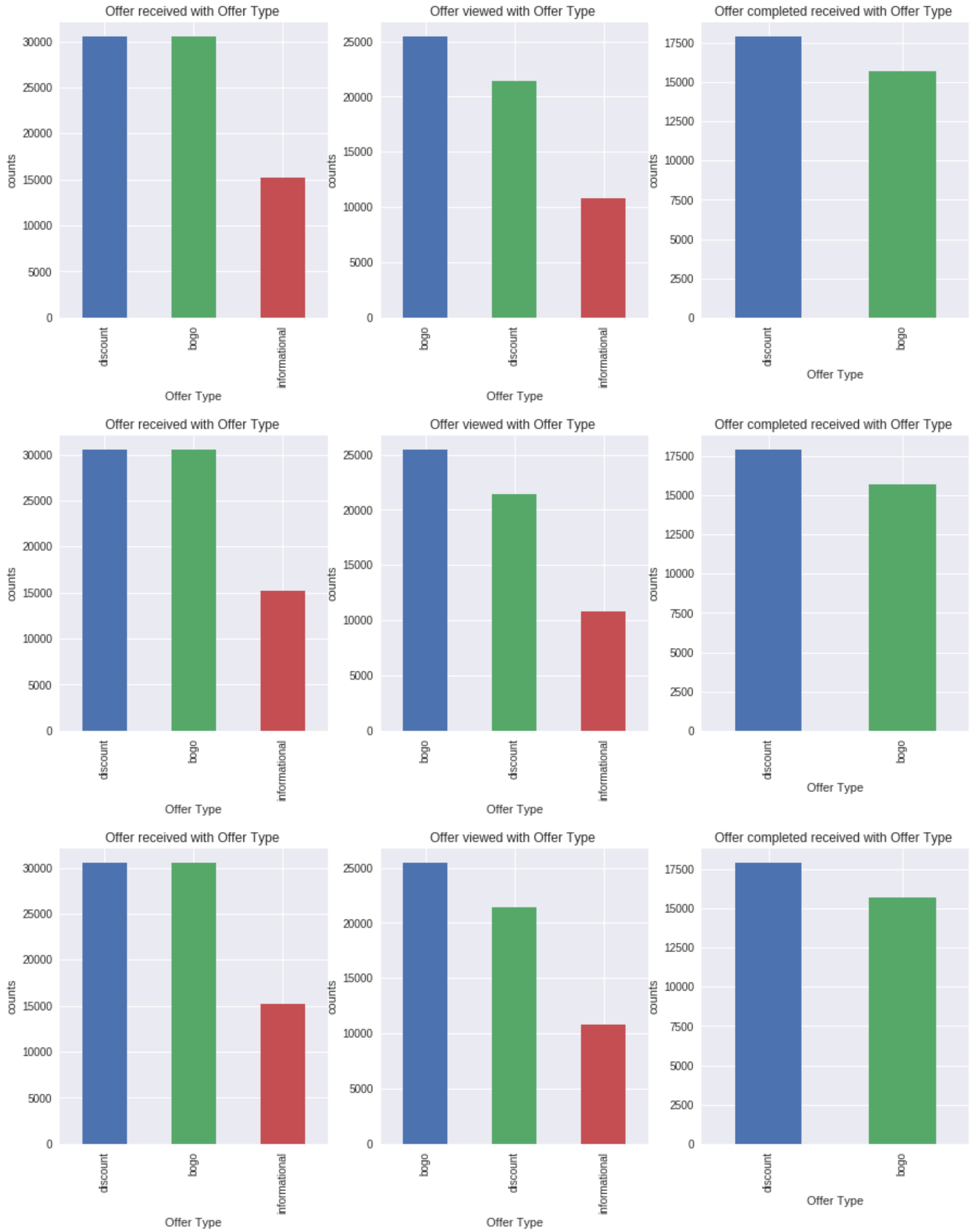

days as member with age_range
Salary Range
0
100
200
300
400
500
600
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99
+100
Ages





Also we can see that the distribution male and female of gender and come female have the highest income until age range 80 and mela is more than age range

We have distribution of offer statues regarded to type of offer



Prepare date for model

Fill null for income by forward fill

And use maxmin scaler to normalize numerical data

Make data label and feature

Use unsupervised classification three model

- KNeighborsClassifier
- RandomForestClassifier
- DecisionTreeClassifier

Final Discussion

i have used final 3 model to classify the result we see that random forest classifier and decision tree approximatly the same and KNeighborsClassifier is the least