Starbucks Customer Data Analysis

Starbucks

- Starbucks is one of the leading coffee brands worldwide with thousands of stores spread across various countries. In the digital era and the development of information technology, companies like Starbucks collect a vast amount of data about their consumer behavior.
- Data analysis for Starbucks customers is crucial to understand consumer preferences and habits, as well as to identify trends and patterns that can provide valuable insights for business decision-making. We will explore how data analysis can be utilized to uncover valuable information from Starbucks customer data.
- Starbucks customer profile data contains 17000 records and 6 columns where each record has data about each person, their age, salary, id, when they became member on(Date)and gender and unnnamed column.

About Starbucks Dataset

	Unnamed: 0	gender	age	id	became_member_on	income
0	0	NaN	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	2	NaN	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	4	NaN	118	a03223e636434f42ac4c3df47e8bac43	20170804	NaN

Starbucks customer profile data contains 17000 records and 6 columns where each record has data about each person, their age, salary, id, when they became member on(Date)and gender and unnnamed column.

Data Wrangling

```
df_customer.isna().sum()
```

```
gender              2175
age                    0
id                     0
became_member_on       0
income              2175
dtype: int64
```

For df_customer, there are 2175 missing values in column gender and the same number of missing values in columnincome. Need to check whether these missing values are from the same observations.

```
missing_gender = df_customer[df_customer['gender'].isna()]
missing_income = df_customer[df_customer['income'].isna()]

np.sum(missing_gender['id'] == missing_income['id'])
```

```
2175
```

All missing values in columns gender and income are from the same observations.

Logically,income is likely to play an important role in customer behaviour, hence observations with missing income will be removed from the dataset. This means 12.8% of the observations will be removed which is not ideal, however the alternative (replacing NaNs with the average or the median) will make our analysis less robust. As there are enough obsevations in our data for analysis, this should not have an important impact on the analysis (unless observations removed share some common behavior, inthat case, removing them will skewer the sample distribution).

```
id_to_remove = missing_income['id']
df_customer_no_na = df_customer[~df_customer['id'].isin(id_to_remove)]
df_customer_no_na = df_customer_no_na.reset_index(drop = True)
df_customer_no_na.info()
```

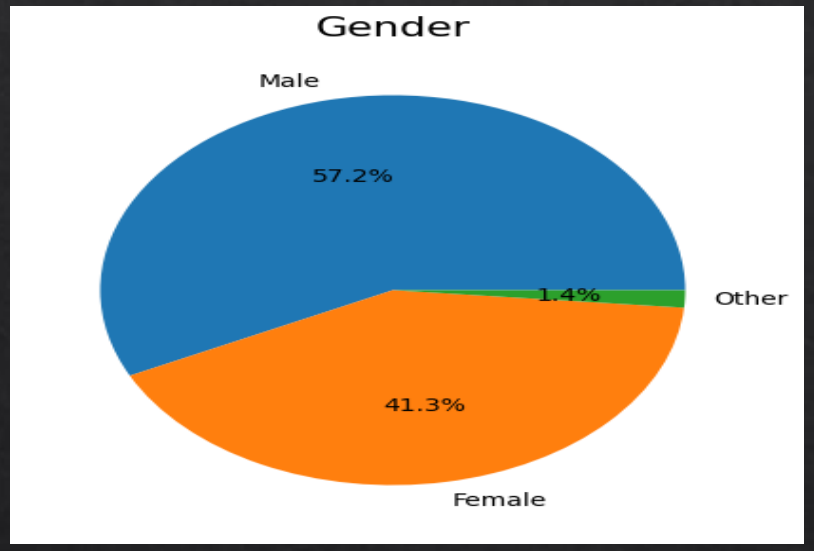
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14825 entries, 0 to 14824
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   gender            14825 non-null  object 
 1   age               14825 non-null  int64  
 2   id                14825 non-null  object 
 3   became_member_on  14825 non-null  int64  
 4   income            14825 non-null  float64
dtypes: float64(1), int64(2), object(2)
memory usage: 579.2+ KB
```

Remove rows with missing values in df_customer.

There are no missing values in df_customer_no_na.

Exploratory Data Analysis (EDA)

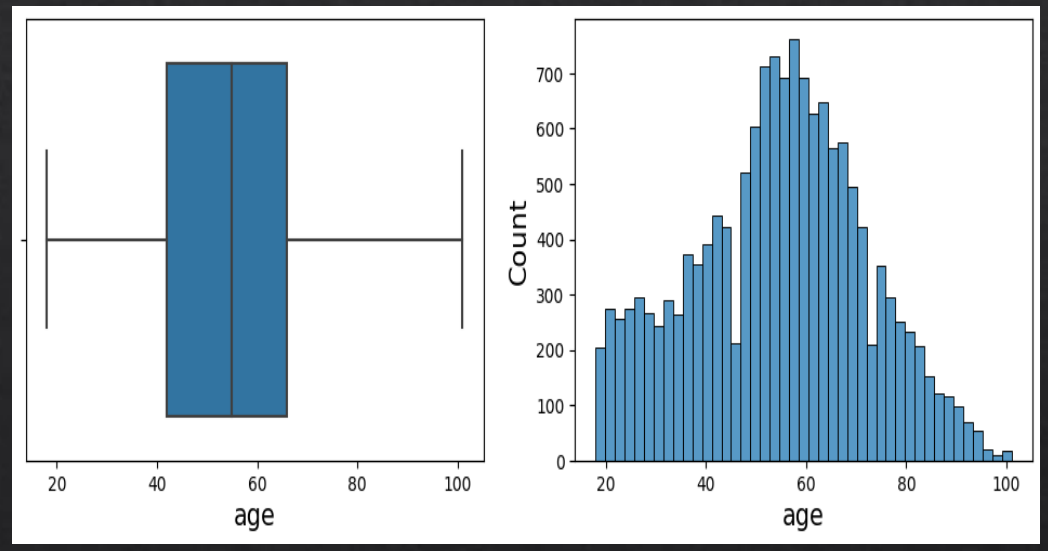
Univariate Analysis - Gender



Observations:

There are more males (57.2%) than females (41.3%) customers, and a small share (1.4%) of customers of other genders.

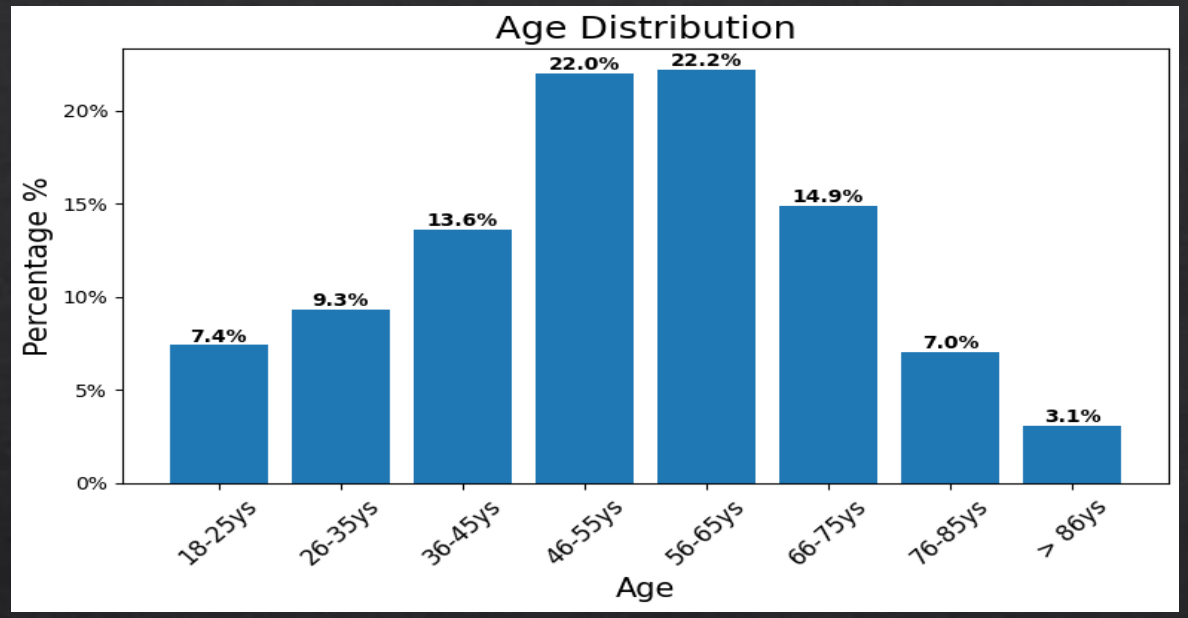
Univariate Analysis - Age



Observations:

- The youngest customer is 18 years old and the oldest is 101 years old.
- The age of the customers roughly follows a normal distribution with the mean and the standard deviation equal to 54 and 17.

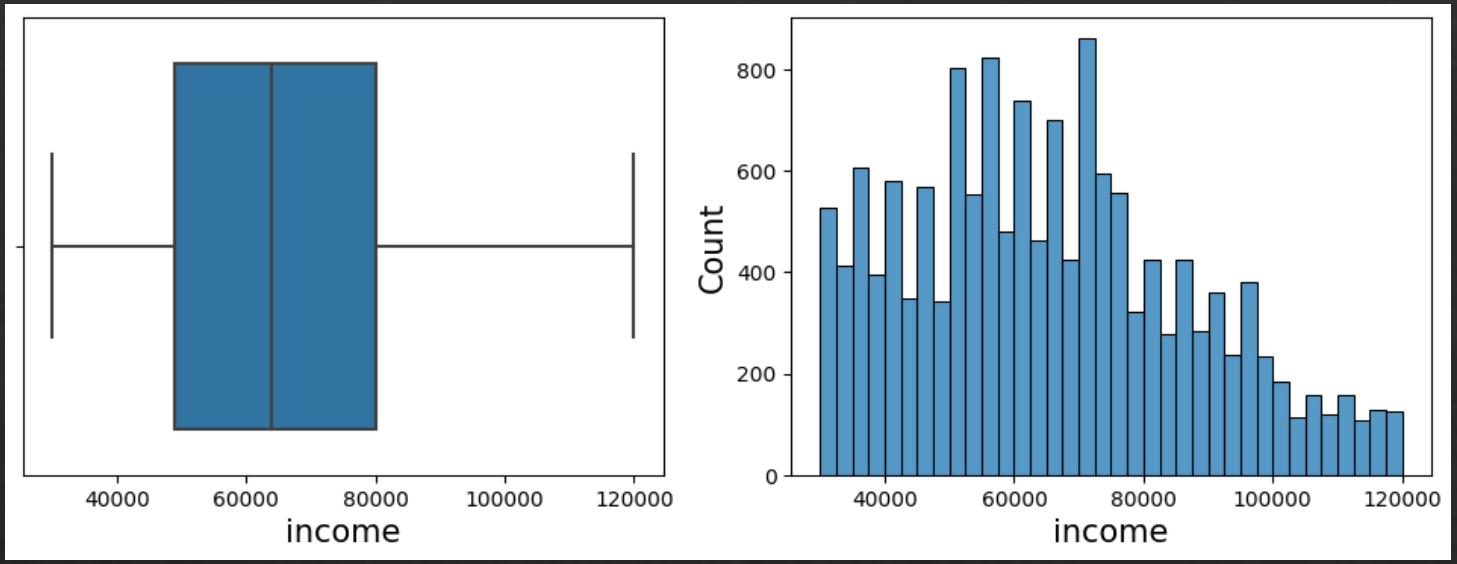
Divide Customers Into Age Groups

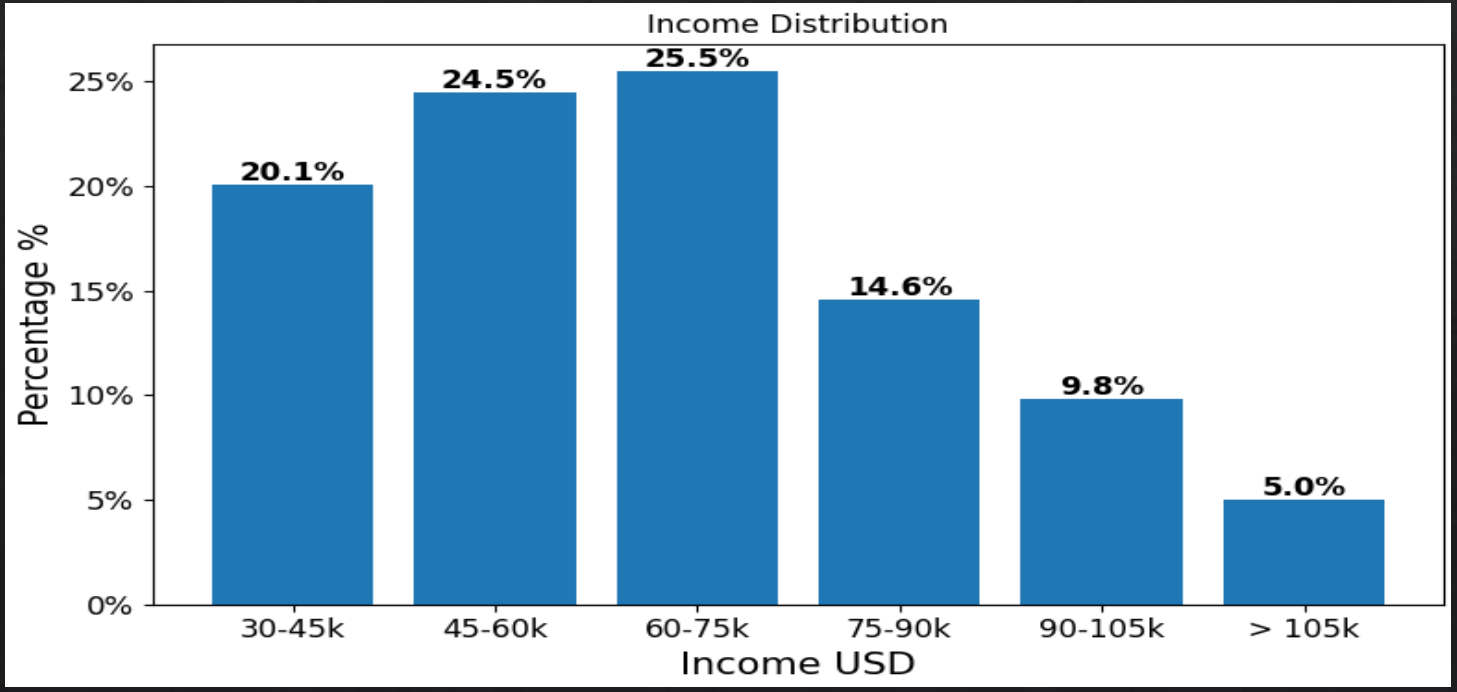


Observations:

- The largest age group of customers is 56-65 years olds, closely followed by age group 46-55 years olds. The third largest age group is 66-75 years olds.
- Top 3 age groups account for approximately 60% of customers.
- This makes sense as people between 46 - 75 years old tend to be less busy than younger people, therefore have more time to visit cafes. They also tend to have better mobility than more senior people (above 75 years old), therefore are more likely to be our customers.

It would be very interesing to investigate how age affects customers response to offers and their purchasing habbits. This will be investigated in a later section.



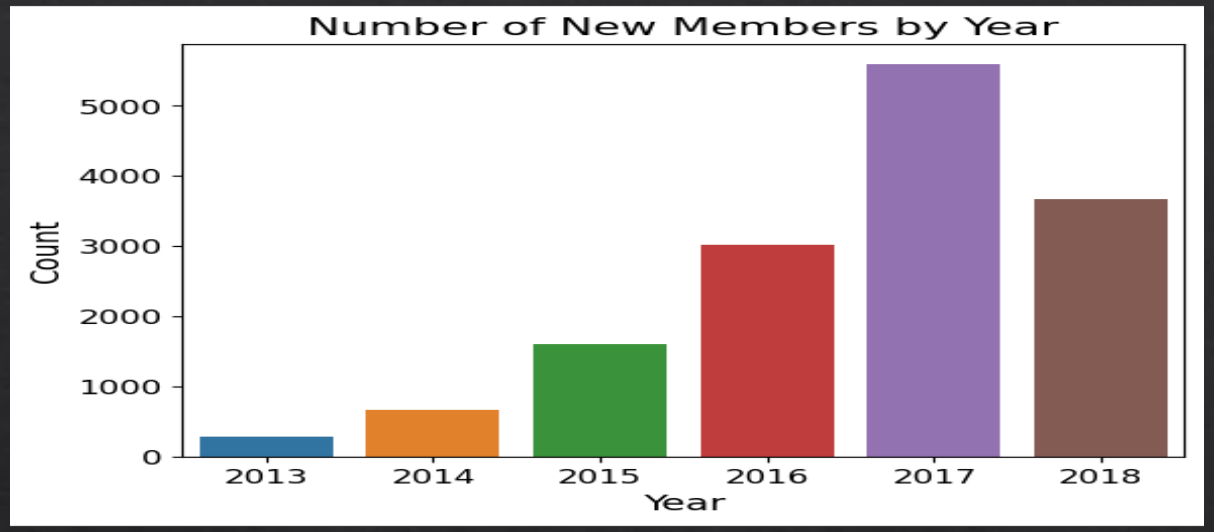


Univariate Analysis - Income

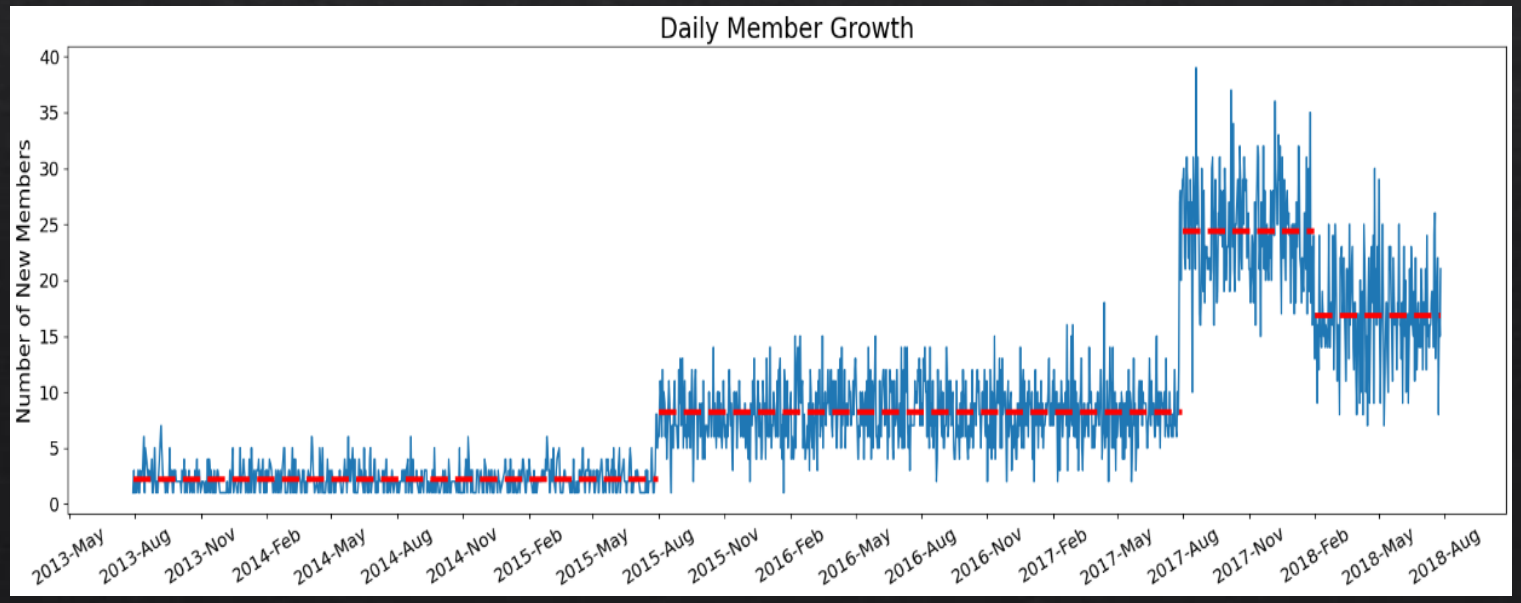
Observations:

- Income of customers has a range of 30k to 120k, with an average of 65.4k.
- Top 2 income groups consist of customers with annaul income of 60-75k and 45-60k. These account for approximately 50% of customers.
- The income distribution among the customers may, to a large extent, in line with that over the whole population, except that people with extremely low income might find it hard to afford regular cafe visits, while people with extremely high income may prefer other ways to enjoy their coffee.

Member Growth by Year



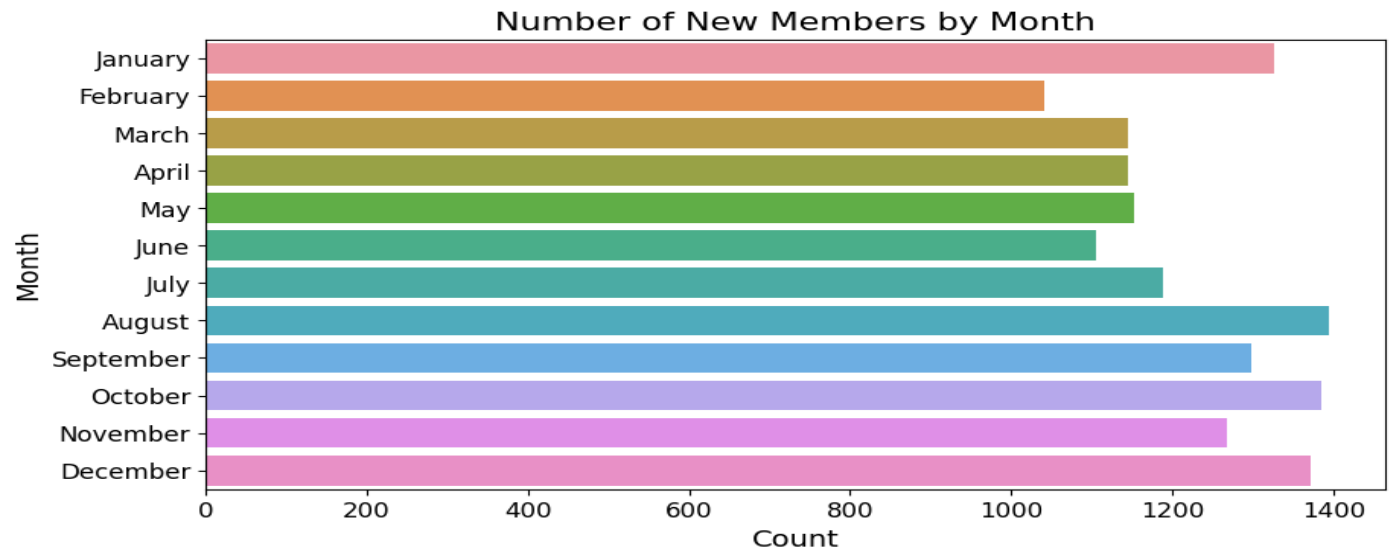
Daily Member Growth



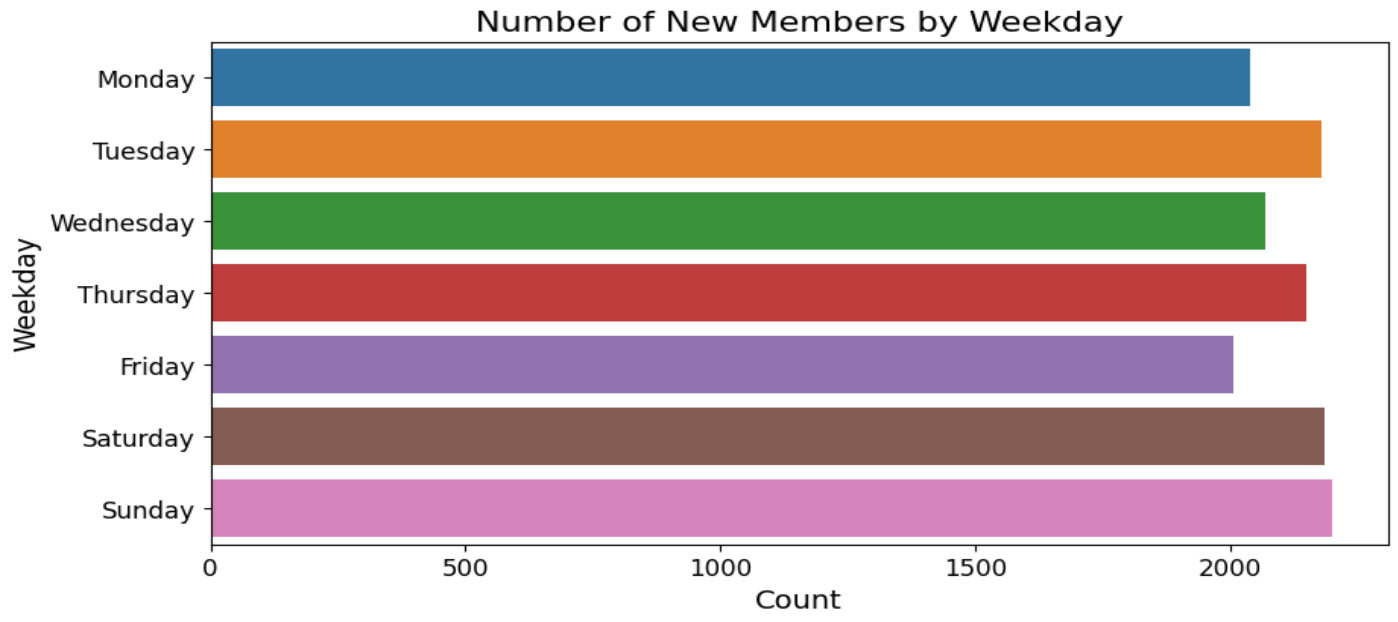
Observations:

- The data contains customers became members from July 2013 to July 2018.
- Very few customers (~2.2) chose to become a member from 2013 to mid-2015.
- The number of new members started to pick up from mid-2015 and really took off from mid-2017. The company must have had some sucessful compaigns around mid of 2015 and mid of 2017 that drastically boosted member growth.
- However, from early 2018 the daily number growth dropped by 31% (from 24.4 to 16.8). Maybe some new strategies were applied then and had a negative impact on the member growth.

Member growth by Month



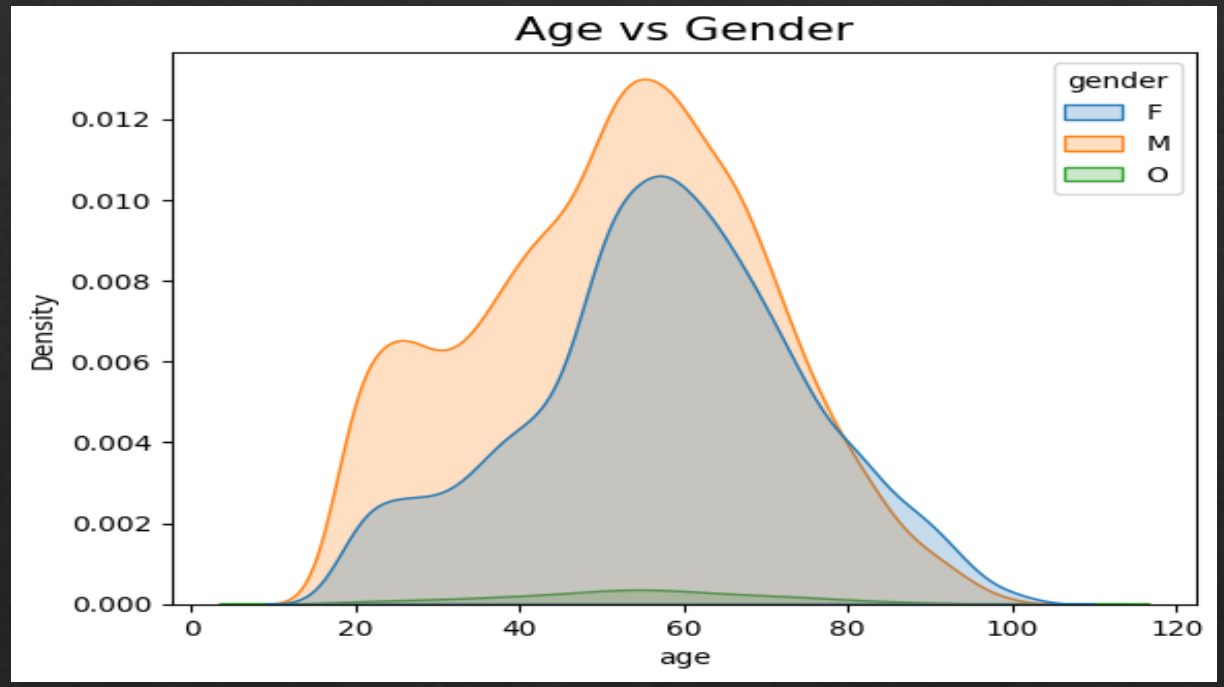
Member growth by weekday



Observations:

- The months saw most customers becoming memebrs was August, followed by October and Decemebr.
- There were least new memebrs in Feburary on average.
- In terms of days in a week, unsuprisingly, weekend saw the higher member growth than weekdays. Tuesday had the highest member growth among weekdays.
- There were least new memebrs on Friday on average.

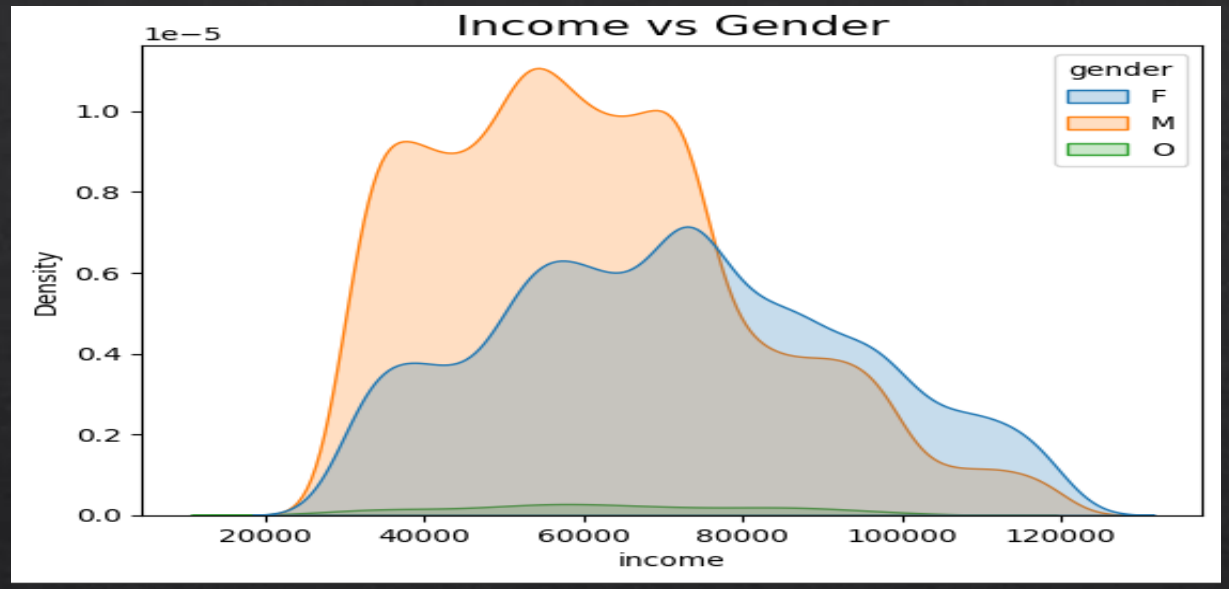
Bivariate Analysis - Age vs Gender



Observations:

- The number of female and male cutomers roughly follows a normal distribution.
- There is a larger proportion of young customers in males than females.
- Female customers, male customers and cutomers of other genders are 57.5, 52.1 and 54.4 year old on average.

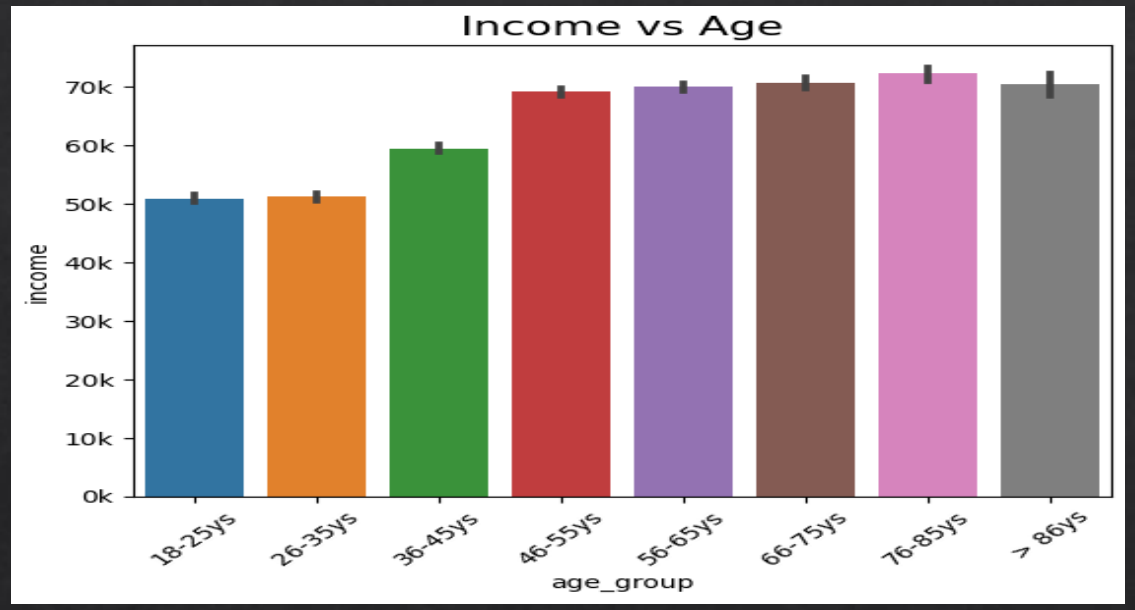
Bivariate Analysis - Income vs Gender



Observations:

- Income of female customers roughly follows a normal distribution. This indicates female customers accross the whole income range enjoys the company's products.
- Income of male customers is skewed to the right. This means among the company's male customers, more people are on the lower half of the income spectrum (among customers, not among US population, as shown below).
- Female customers have a much higher average income than other genders. This maybe becase they are on average older than other gender groups, assumming older customers have higher incomes (it will be checked whether this assumption is true next).
- Average income of female customers, male customers and cutomers of other genders are 71k, 61k and 63k.

Bivariate Analysis - Income vs Age



Observations:

- Customers in the two younger age groups (18-35 years old) has an average annual income of about 51k USD. The middle age groups (36-55 years old) has an average annual income of about 65k. All age groups above 56 years old have very similar average annual income of round 70k. On average, older customers of the company have higher incomes.
- The company's customers in all age groups have a much higher average income than the median income of US citizens (30,119 USD in 2018). I use the median income of 2018 as the benchmark (rather than a more recent year) because in this data set, the latest time a customer became member was July 2018. Profile data, such as income, is most likely collected when the customers signed up to become memebrs and this infomation is usually left unupdated by most people. Hence, it makes more sense to compare customer income with that of the population of the year customers last became members.

Conclusion

- There are more males (57.2%) than females (41.3%) customers, and a small amount (1.4%) of customers of other genders.
- Customer age ranges from 18 to 101, roughly following a normal distribution with the mean and standard deviation being 54 and 17. Customers in top 3 age groups (46-75 years old) account for 60% of all customers.
- Customer income (30 - 120k) is skewed to the right, having a mean of 65.4k and a standard deviation of 21.6k. Customers with a income between 45k-75k account for approximately 50% of all customers.
- In general, customers in younger age groups have an average income lower than those in more senior age groups. Customers in the two younger age groups (18-35 years old) has an average annual income of about 51k. The middle age groups (36-55 years old) has an average annual income of about 65k. All age groups above 56 years old have very similar average annual income of round 70k.
- There is a higher proportion of young people in male customer than in other genders, and (therefore) male customers have a lower average income than other gender groups.

Thank You

- Linkedin
  www.linkedin.com/in/aldilah-ariwibowo

- Github
  https://github.com/aldilahariwibowo