Exploratory Data Analysis

G2M insight for Cab Investment Firm

Author: Chooladeva Piyasiri

Agenda

Problem Statement

Datasets Exploration

EDA

Hypothesis Investigation

Summary & Recommendations

Problem Statement

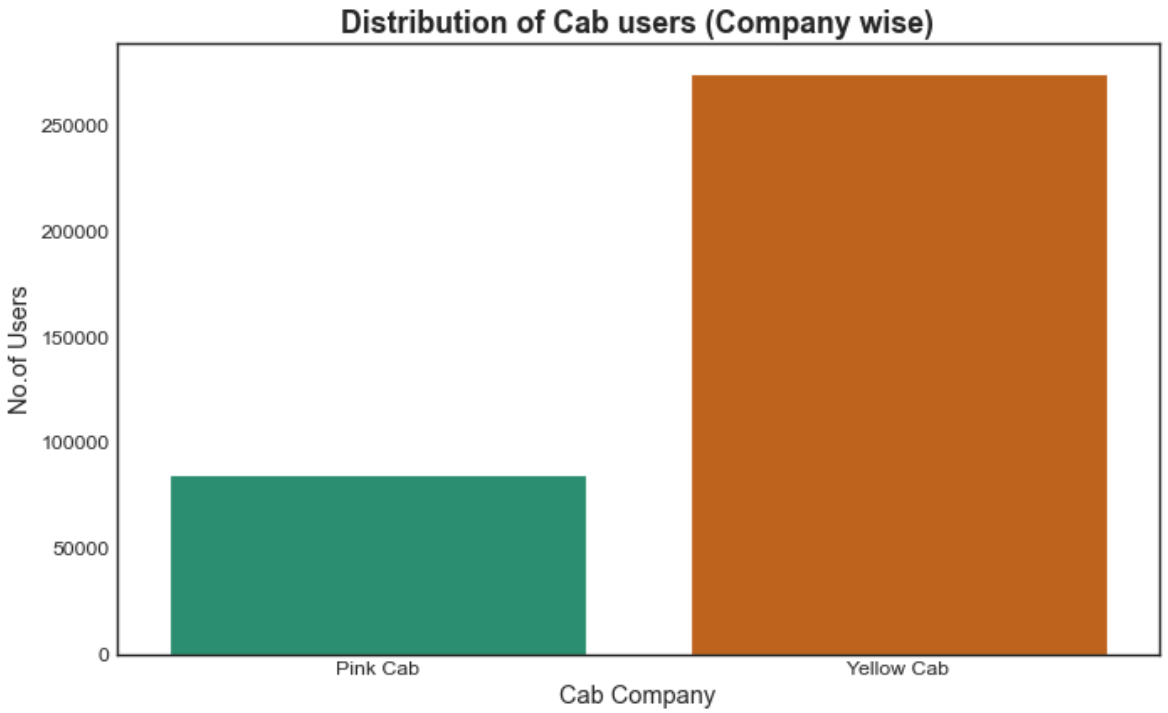
- XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.
- Their problem is their inability to identify the right company for making investment.

Datasets Exploration

- Below are the list of datasets which are provided for the analysis:
  - **Cab_Data.csv** – this file includes details of transaction for 2 cab companies
  - **Customer_ID.csv** – this is a mapping table that contains a unique identifier which links the customer's demographic details
  - **Transaction_ID.csv** – this is a mapping table that contains transaction to customer mapping and payment mode
  - **City.csv** – this file contains list of US cities, their population and number of cab users.
- Timeframe of the data: 2016-01-01 to 2018-12-31

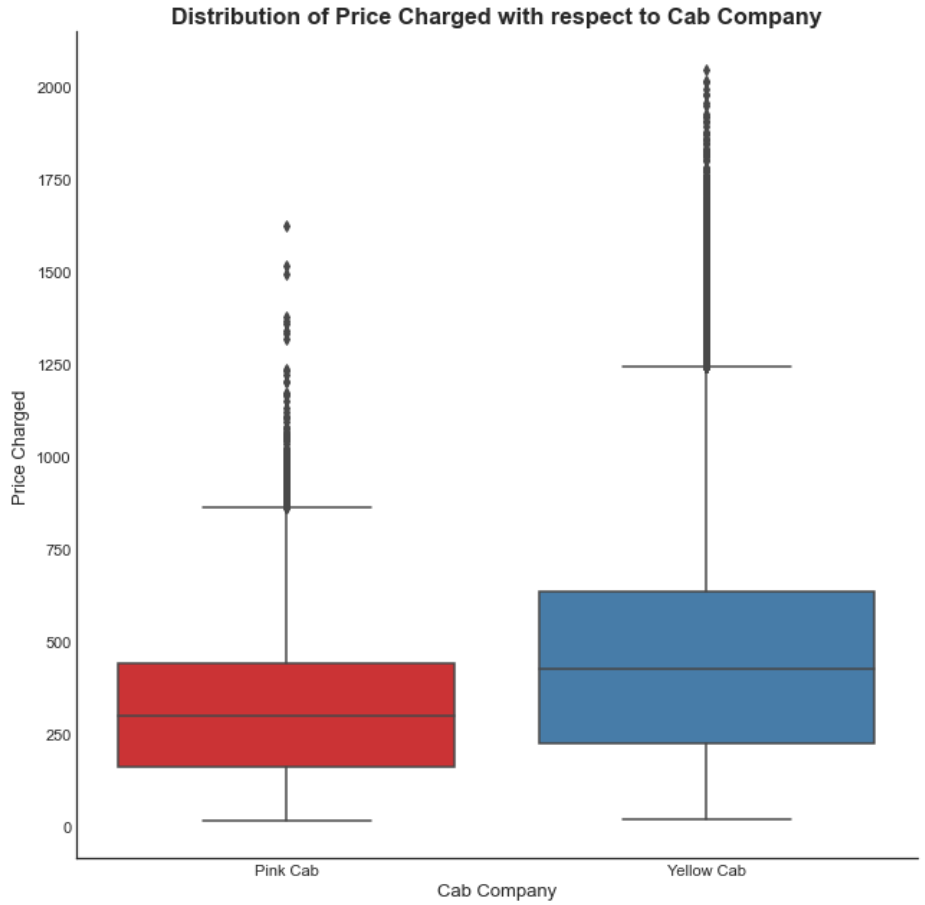
EDA

Figure 01: Distribution of Cab users (Company wise)



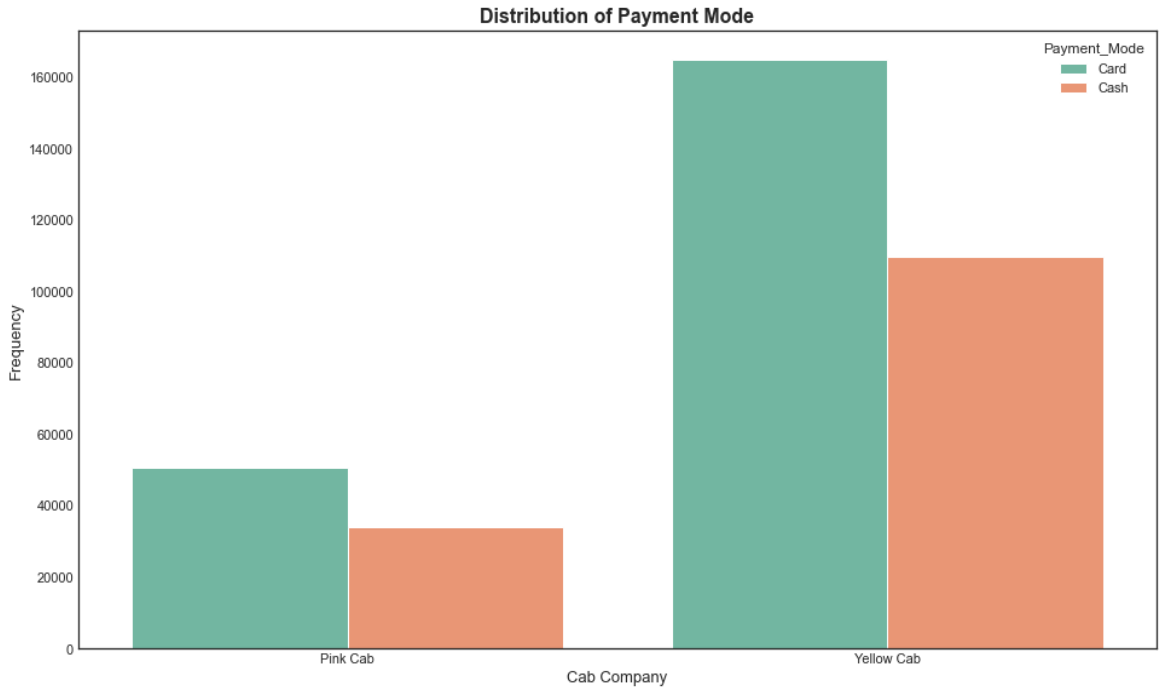
- Users prefer to ride in the 'Yellow Cab' over the 'Pink Cab.'

Figure 02: Distribution of Price Charged with respect to Cab Company



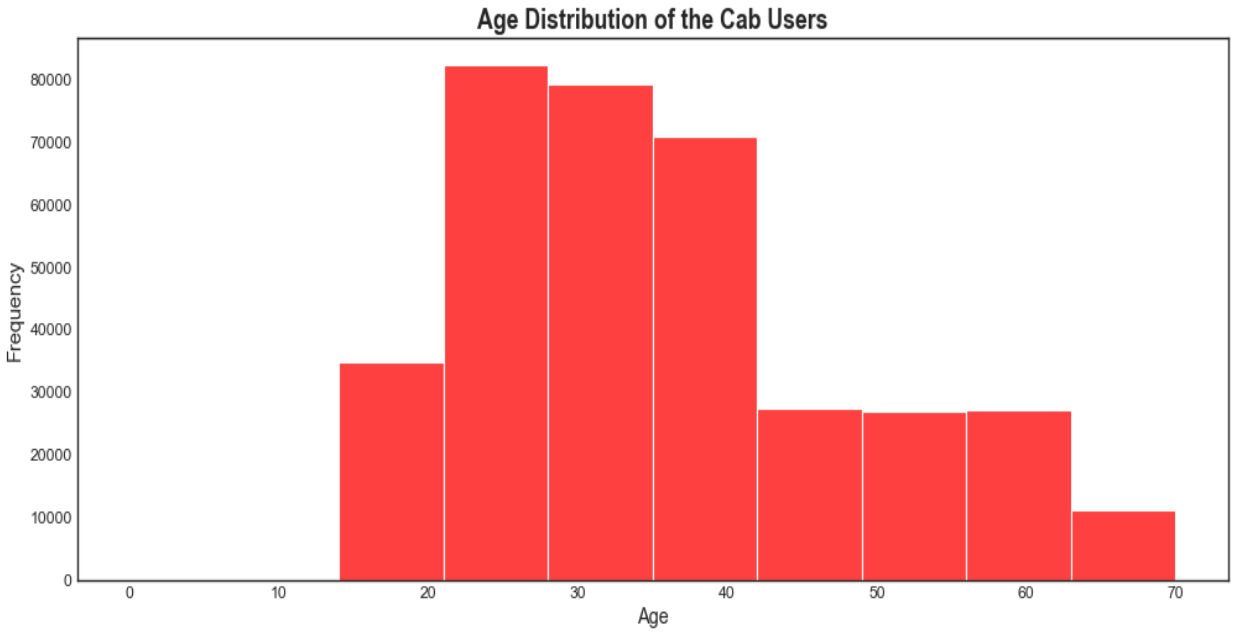
- When compared to the 'Pink Cab,' the 'Yellow Cab' charges the most.
- And also, the 'Yellow cab' has a bigger pricing range than the 'Pink cab’.
- These outliers can be caused by the use of luxury cars, the weather, or the holiday season.

Figure 03: Distribution of Payment Mode



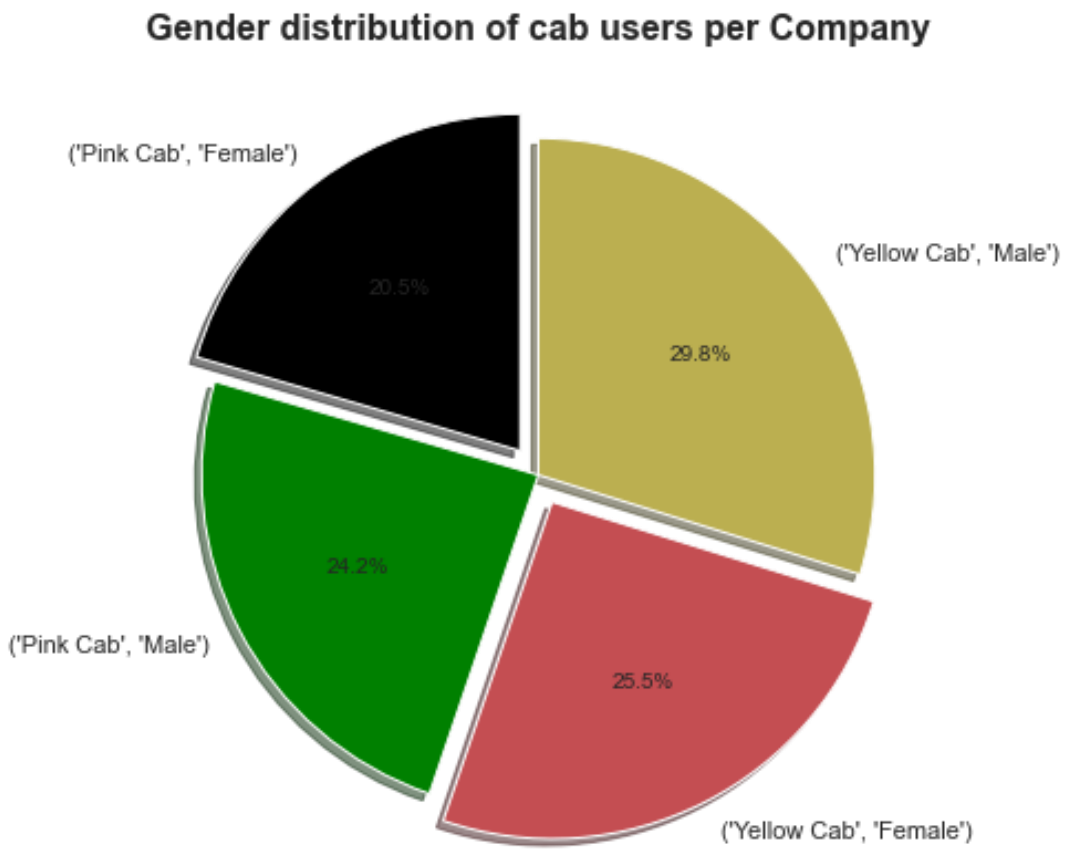
- Users prefer to pay with their cards rather than cash.

Figure 04: Age Distribution of the Cab Users



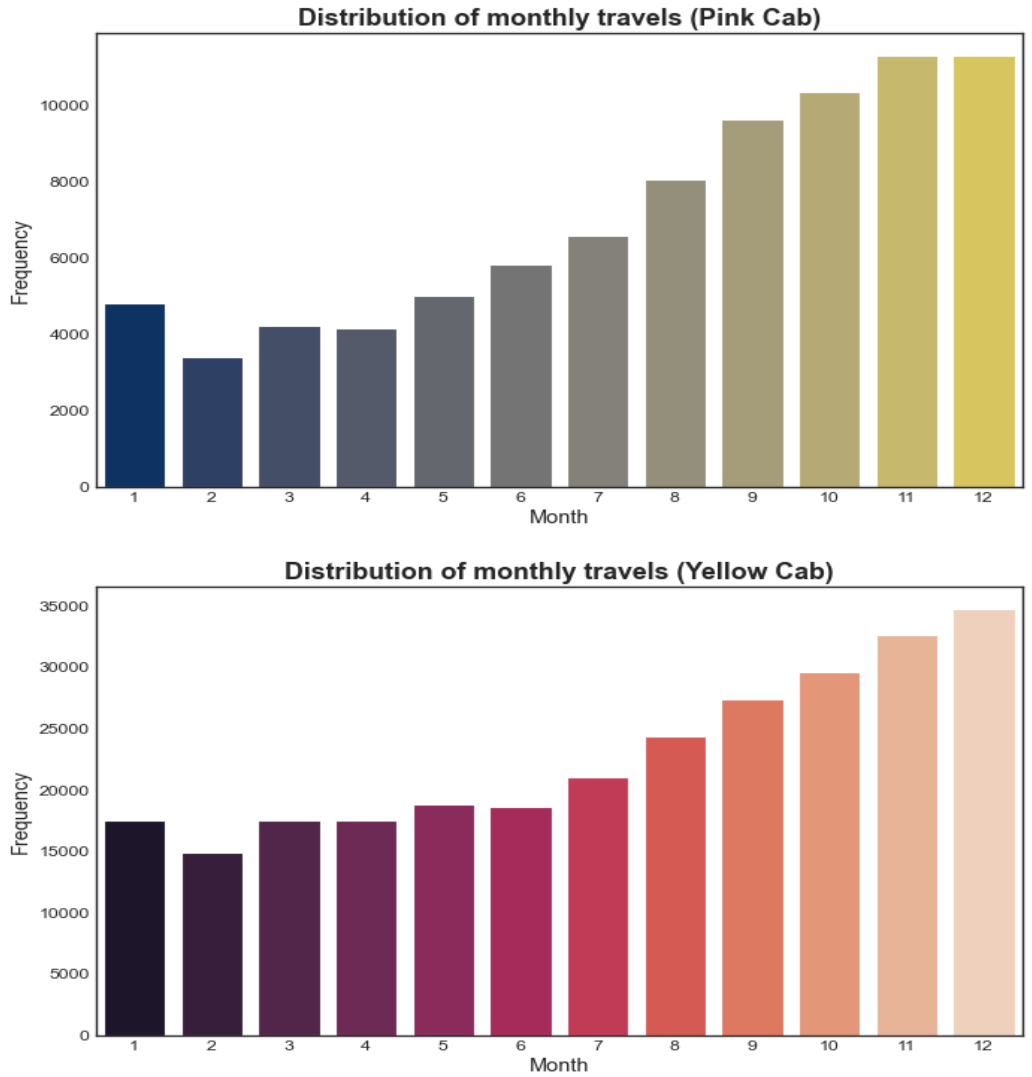
- The majority of cab users are between the ages of 20 and 40.

Figure 05: Gender distribution of unique cab users per Company



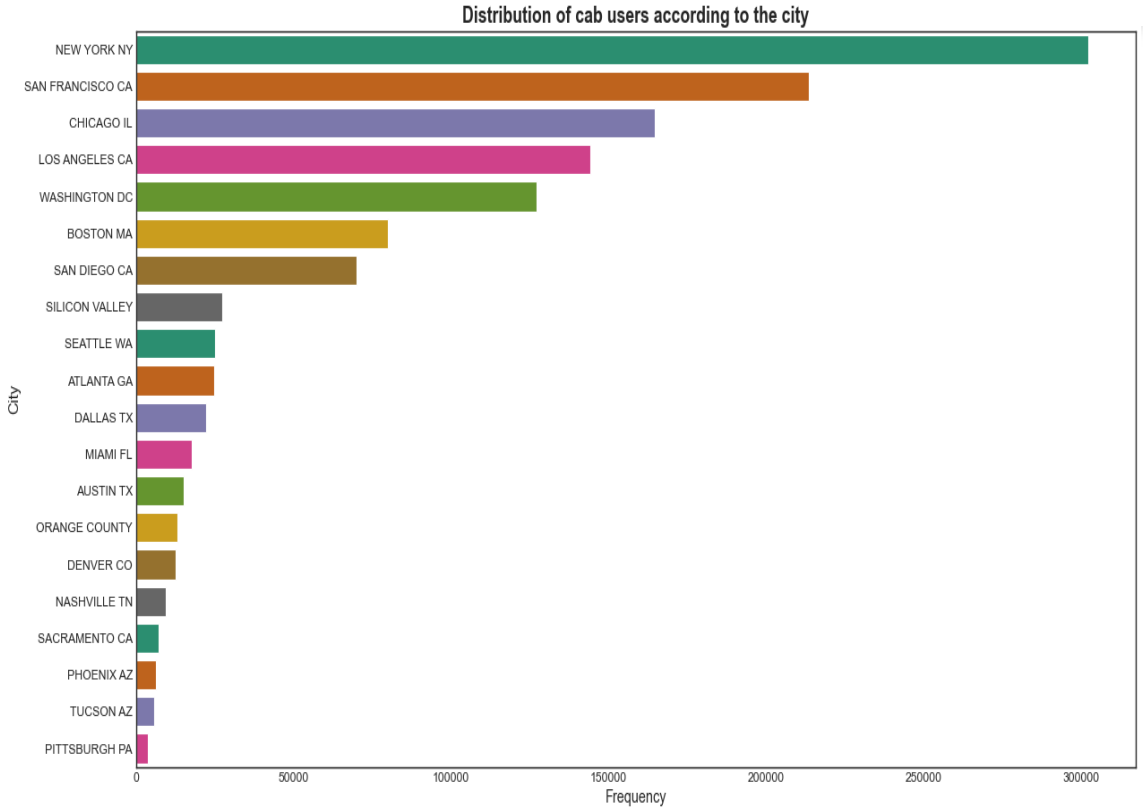
- Males prefer cabs over females for both companies.

Figure 06: Seasonal distribution of User transactions



- We can see a gradual increase in monthly travels for both cab companies beginning in the middle of the year and continuing until the end of the year.

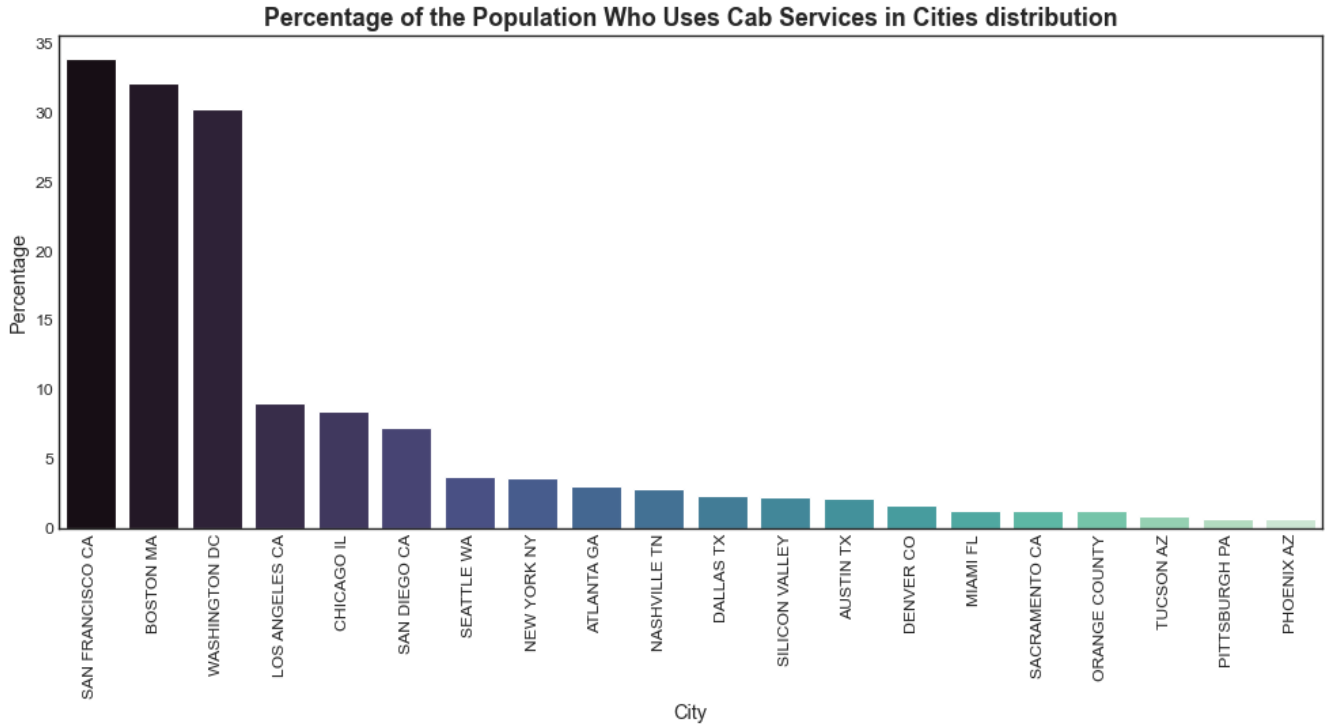
Figure 07: Distribution of cab users according to the city



	City	Users
0	NEW YORK NY	302149
1	SAN FRANCISCO CA	213609
2	CHICAGO IL	164468
3	LOS ANGELES CA	144132
4	WASHINGTON DC	127001
5	BOSTON MA	80021
6	SAN DIEGO CA	69995
7	SILICON VALLEY	27247
8	SEATTLE WA	25063
9	ATLANTA GA	24701
10	DALLAS TX	22157
11	MIAMI FL	17675
12	AUSTIN TX	14978
13	ORANGE COUNTY	12994
14	DENVER CO	12421
15	NASHVILLE TN	9270
16	SACRAMENTO CA	7044
17	PHOENIX AZ	6133
18	TUCSON AZ	5712
19	PITTSBURGH PA	3643

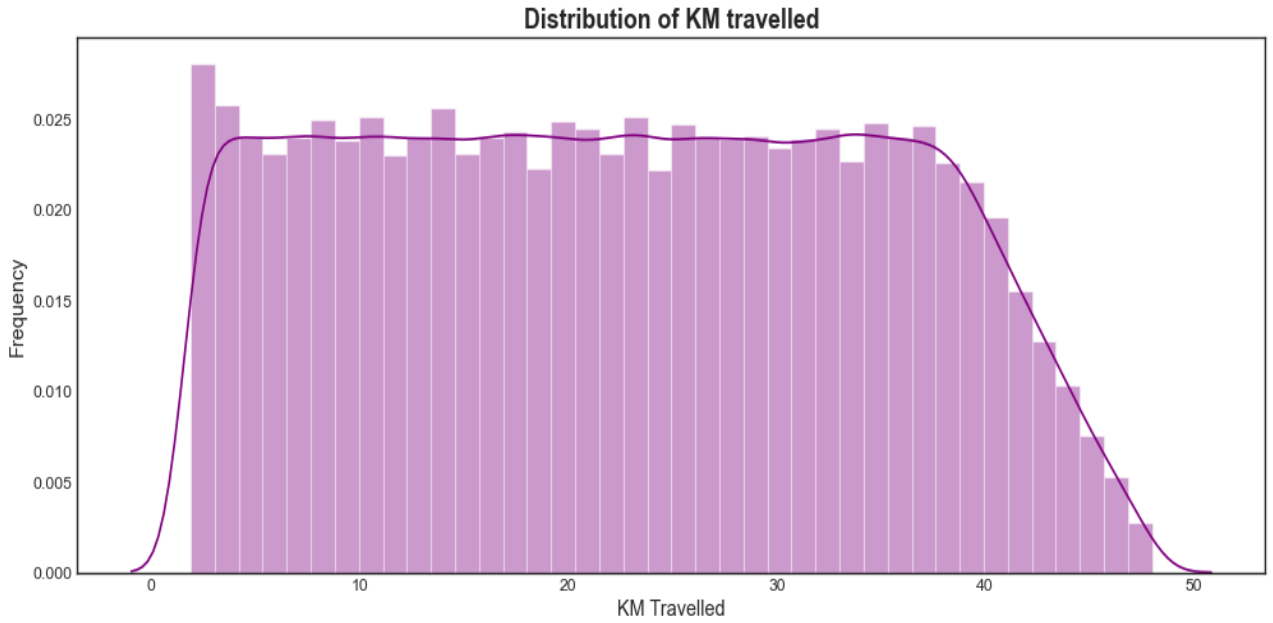
- New York has the most cab user transactions, followed by San Francisco, Chicago, and Los Angeles.

Figure 08: Percentage of the Population Who Uses Cab Services in Cities



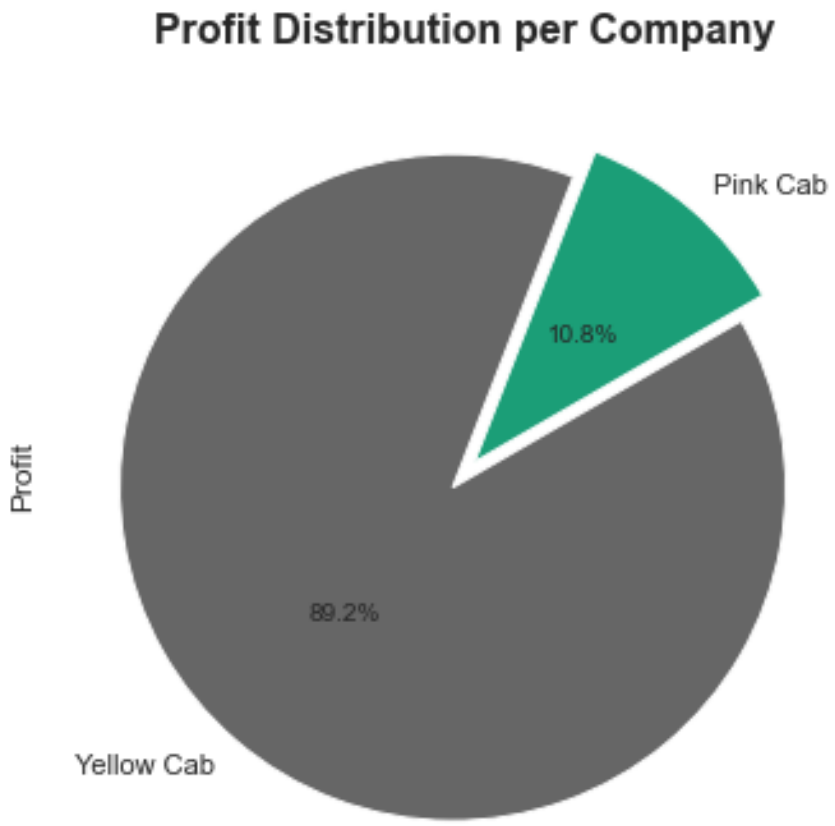
- In San Francisco, Washington, and Boston, cabs are used by more than 30% of the population.

Figure 09: Distribution of KM travelled



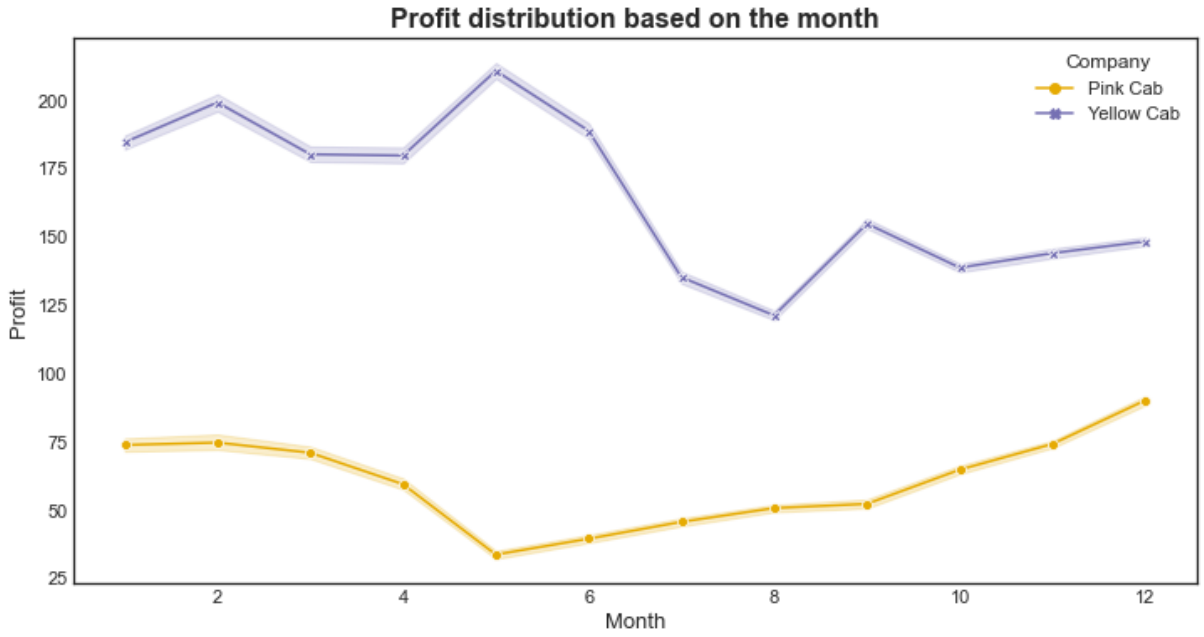
- We can see that the majority of the rides are between 2 and 48 kilometers long.

Figure 10: Profit Distribution per Company



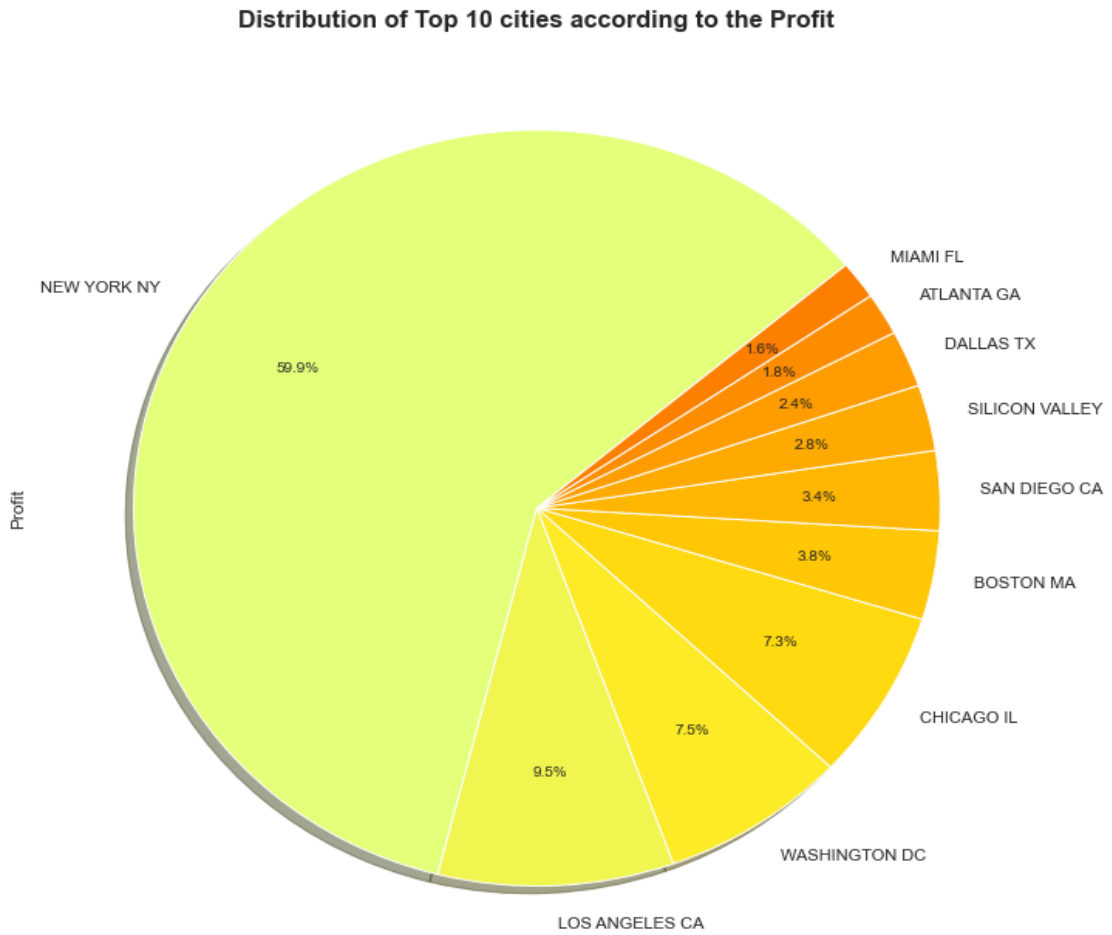
- The 'Yellow Cab' Company received 89.2% of the total profit.

Figure 11: Profit distribution based on the month



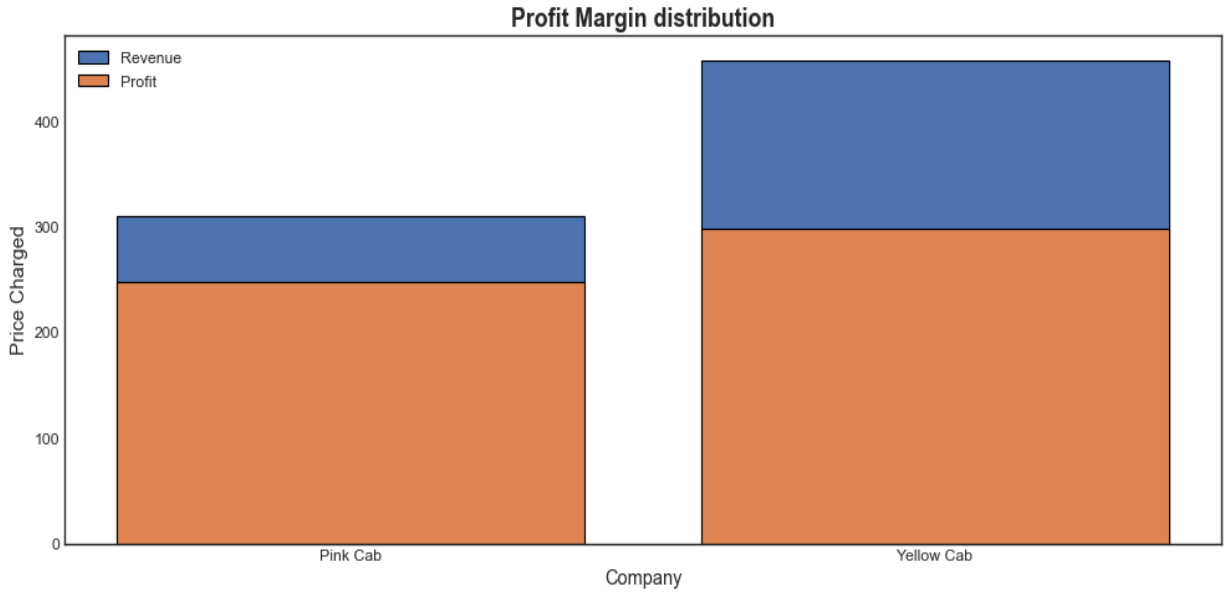
- The most profitable month for Pink Cabs is December, followed by November and January.
- May is the most profitable month for Yellow Cab, followed by February and January.
- When compared to other quarters, Yellow Cabs' profit drops significantly near the end of the year.

Figure 12: Distribution of Top 10 cities according to the Profit



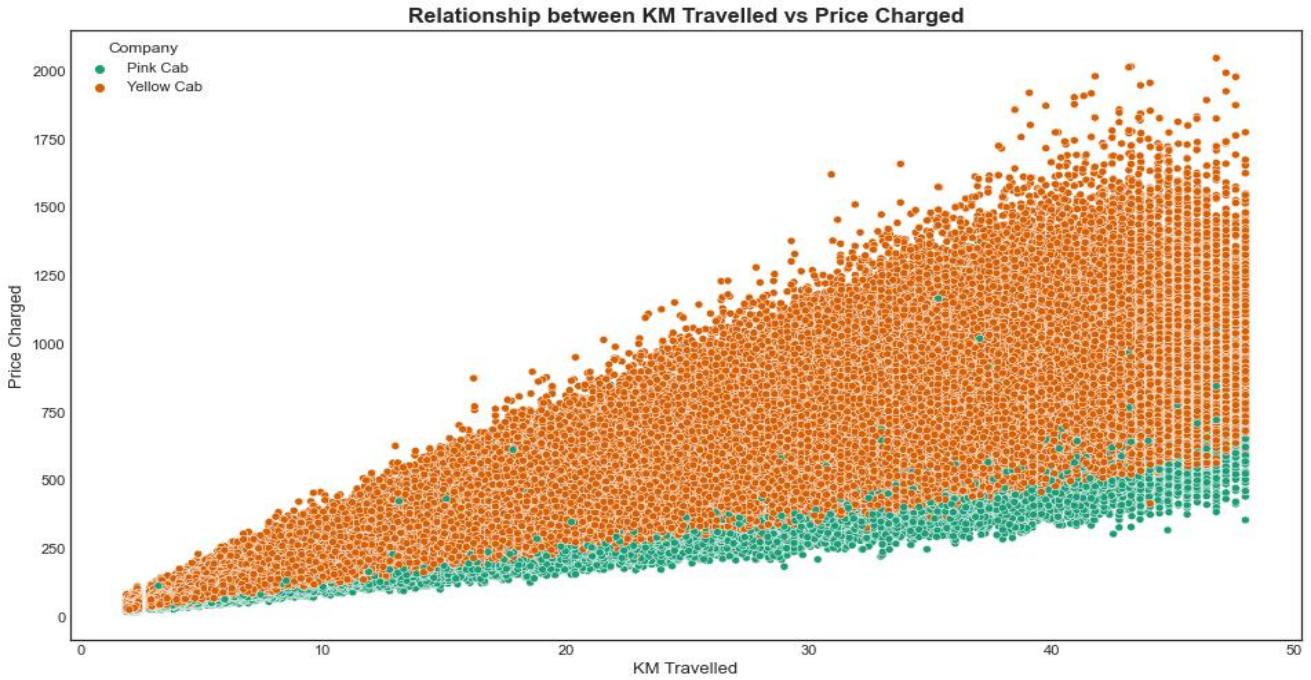
- New York City generates 60% of the profits for both companies and has the highest number of users.

Figure 13: Profit Margin distribution



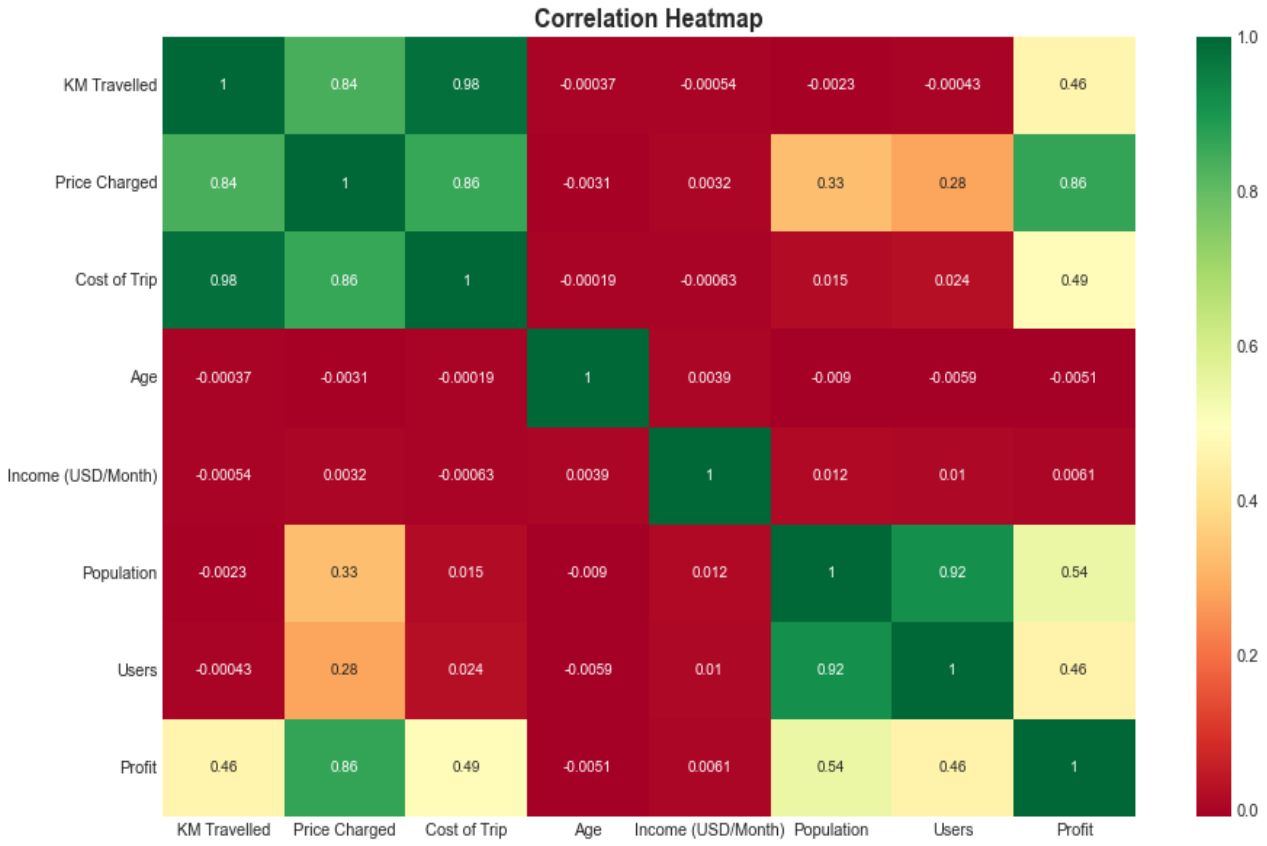
- When compared to the Pink cab, the Yellow cab has a significantly higher profit margin.

Figure 14: Relationship between KM Travelled vs Price Charged



- As expected, we can observe that there is a positive correlation between 'KM Traveled' and 'Price Charged' for both the Pink and Yellow cabs.

Figure 15: Correlation Matrix



- Age and user income have no effect on any of the other attributes.
- There is a moderate correlation between Profit and Population, as well as Profit and Users.

Hypothesis Investigation

- **Hypothesis 1: Does the number of kilometers traveled vary according to the user's income level?**

  - No, According to the correlation matrix, a score of 1.0 is perfectly correlated. But the "KM Traveled" and "Income (USD/Month)" obtained a score of -0.00054. which indicates that they're weakly negatively correlated. Check out the Figure 15.

- **Hypothesis 2: Is there a seasonal variation in the number of customers using the cab service?**

  - Yes, We can see a gradual increase in monthly travels for both cab companies beginning in the middle of the year and continuing until the end of the year. Check out the Figure 06.

- **Hypothesis 3: Is the city with the highest number of cab services and the city with the most cab users relative to their population the same?**

  - No,
  - Among the cities, New York has the most cab users. However, only 3.59% of their population uses cabs. But cities like San Francisco, Boston, and Washington have the most cab users relative to their populations. More than 30% of the population in these cities uses cabs. Check out the Figure 07 and Figure 08.

- **Hypothesis 4: Do older people (Above 50) use cabs more often than younger people?**

  - No,
  - The majority of cab users are between the ages of 20 and 40. The least number of cab users are over 70 years old. Cab users aged between 40 and 60 are evenly distributed. Check out the Figure 04.

- **Hypothesis 5: Do females use cabs more often than males?**

  - No,
  - Males prefer cabs over females for both companies. Check out the Figure 05.

- **Hypothesis 6: Do people prefer to pay with cards rather than cash?**

  - Yes,
  - For both companies, the majority of cab users prefer to pay with their cards rather than cash. Check out the Figure 03.

Summary & Recommendations

- This report presents the Exploratory Data Analysis (EDA) of two cab companies in the United States: Pink Cab and Yellow Cab. All datasets were merged and cleansed before being used to create data visualizations for insight. We discovered that the master data frame contains no null values and duplicate values. Then I went through the EDA process and discovered that,

  - The ‘Yellow Cab’ appears to make more profits than the 'Pink Cab, owning 89% of both companies' total profit.
  - Most users prefer to travel in a 'Yellow Cab' rather than a 'Pink Cab’. (may be due to Offers, Advertising and PR)
  - The ‘Yellow cab’ costs more than the ‘Pink cab.’ This could be due to the facilities, time management, and reputation of the Yellow cab Company.
  - The 'Yellow Cab' has the highest cab utilization in the United States and a significantly higher profit margin than the 'Pink Cab.’ Means that the business do well at managing their sales costs.

- In conclusion, I recommend **Yellow Cab** would be the **better option** to invest in.
- Other overall cab investment recommendations are mentioned below.

- Cab companies should improve their card payment options. Since Users prefer to pay with their cards rather than cash. (like Zettle by PayPal, PayaTaxi)
- Cabs should target more younger male users aged 20 to 40 and improve their services accordingly.
- Companies should increase the number of cabs near the end of the year as the monthly travel usage gradually increases.
- Cab companies should increase their operations in cities such as New York, San Francisco, Chicago, Washington, Boston, and Los Angeles.

Thank You