Data Glacier

Your Deep Learning Partner

G2M Case Study

Data Science Virtual Internship

20-July-2024

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary with problem statement

Due to remarkable growth in the cab industry over the last few years and the presence of multiple key players in the market, XYZ Corporation is planning to invest in the cab industry as part of their Go-to-Market (G2M) strategy. The goal is to evaluate and recommend the best investment option between the two cab industries.

Problem Statement:
XYZ Corporation needs to decide between two cab companies, each with differing strengths and challenges. A comprehensive analysis, including Exploratory Data Analysis (EDA), will be conducted to derive insights. Based on these insights, a recommendation will be made on which company to invest in.

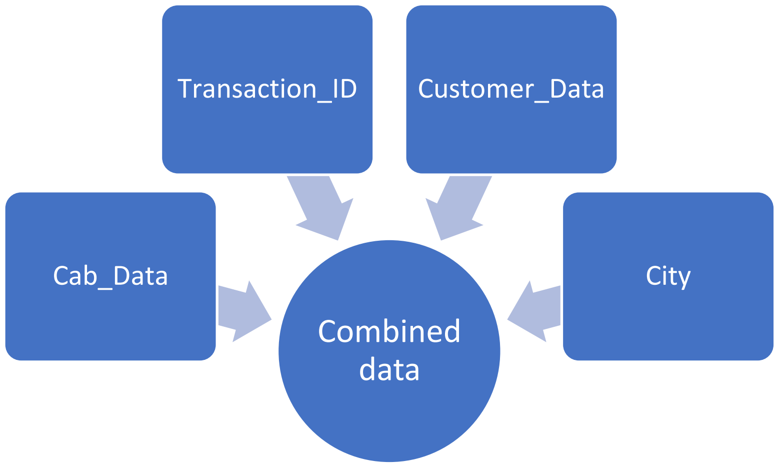
Approach

1. **Data Preparation**
   - Clean data by checking for null values, missing data, data types, and duplicates.
2. **Data Integration**
   - Merge multiple datasets using appropriate foreign keys (e.g., city, customer ID, transaction ID).
3. **Feature Engineering**
   - Enhance the dataset by adding relevant columns (features) to facilitate visualizations and trend analysis.
4. **Exploratory Data Analysis**
   - Look closely at the data: what it shows and how things relate.
5. **Hypothesis Testing**
   - Formulate hypotheses based on insights from EDA and Check our ideas with numbers to make sure they’re right.
6. **Insights and recommendations**
   - Decide on the best option for investing in cab industries.

EDA

(Understanding the Data)

Before diving into the analysis, let's understand the dataset. There are four datasets provided, collecting data between **2016** and **2018**, with some common columns among them.



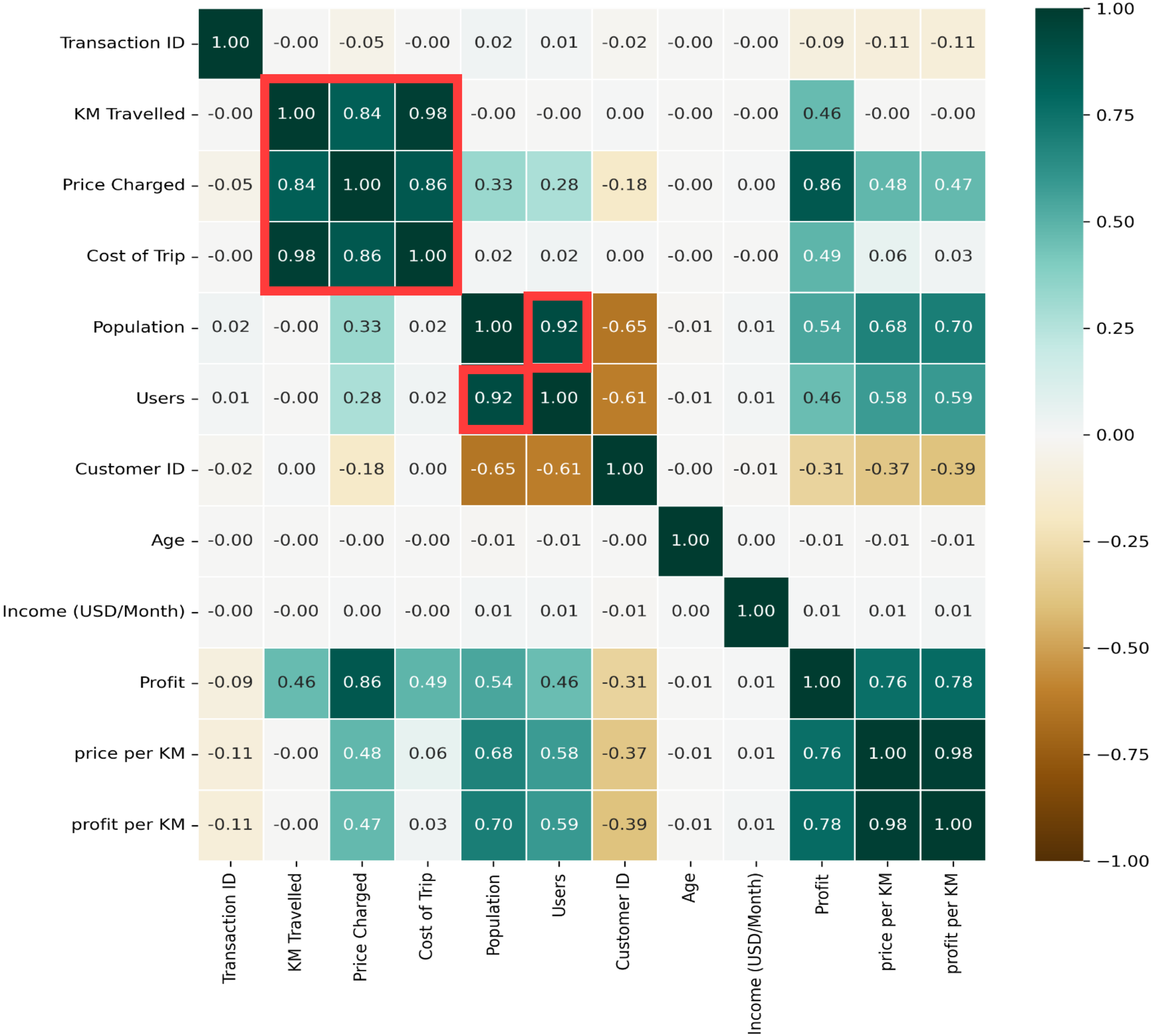
- After merging these four datasets, we have conducted checks for null values, missing data, data types, and duplicates using their shared columns. The combined dataset now consists of **359,352** total rows and **19** columns (**5** of which were added specifically for analysis).

EDA

(Correlation)

- A correlation heatmap shows how closely different things in a group are connected. It helps to quickly spot patterns and decide which variables might be important for further Analysis.
- From the heatmap we can see that there is a strong correlation between:
  - Population and Users
  - Cost/price and distance.

Note:- correlation does not imply causation means that just because two things correlate does not necessarily mean that one causes the other



	Transaction ID	KM Travelled	Price Charged	Cost of Trip	Population	Users	Customer ID	Age	Income (USD/Month)	Profit	price per KM	profit per KM
Transaction ID	1.00	-0.00	-0.05	-0.00	0.02	0.01	-0.02	-0.00	-0.00	-0.09	-0.11	-0.11
KM Travelled	-0.00	1.00	0.84	0.98	-0.00	-0.00	0.00	-0.00	-0.00	0.46	-0.00	-0.00
Price Charged	-0.05	0.84	1.00	0.86	0.33	0.28	-0.18	-0.00	0.00	0.86	0.48	0.47
Cost of Trip	-0.00	0.98	0.86	1.00	0.02	0.02	0.00	-0.00	-0.00	0.49	0.06	0.03
Population	0.02	-0.00	0.33	0.02	1.00	0.92	-0.65	-0.01	0.01	0.54	0.68	0.70
Users	0.01	-0.00	0.28	0.02	0.92	1.00	-0.61	-0.01	0.01	0.46	0.58	0.59
Customer ID	-0.02	0.00	-0.18	0.00	-0.65	-0.61	1.00	-0.00	-0.01	-0.31	-0.37	-0.39
Age	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	1.00	0.00	-0.01	-0.01	-0.01
Income (USD/Month)	-0.00	-0.00	0.00	-0.00	0.01	0.01	-0.01	0.00	1.00	0.01	0.01	0.01
Profit	-0.09	0.46	0.86	0.49	0.54	0.46	-0.31	-0.01	0.01	1.00	0.76	0.78
price per KM	-0.11	-0.00	0.48	0.06	0.68	0.58	-0.37	-0.01	0.01	0.76	1.00	0.98
profit per KM	-0.11	-0.00	0.47	0.03	0.70	0.59	-0.39	-0.01	0.01	0.78	0.98	1.00

EDA
(Distribution of Cab Companies and Users)

- From pie chart (1), we can see that Yellow Cab occupies nearly 80% of the market share compared to Pink Cab.
- From pie chart (2), we can see that Yellow Cab has more average users compared to Pink Cab.

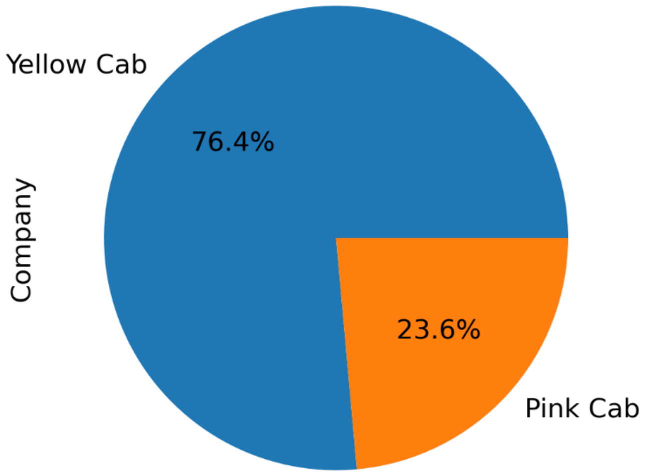


Fig1:- Market Share

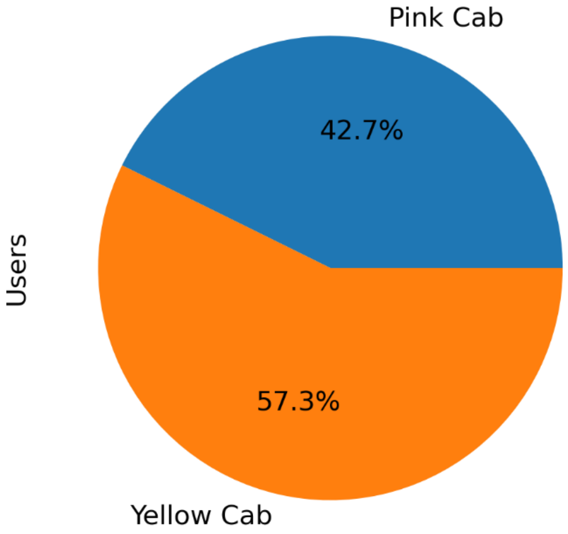
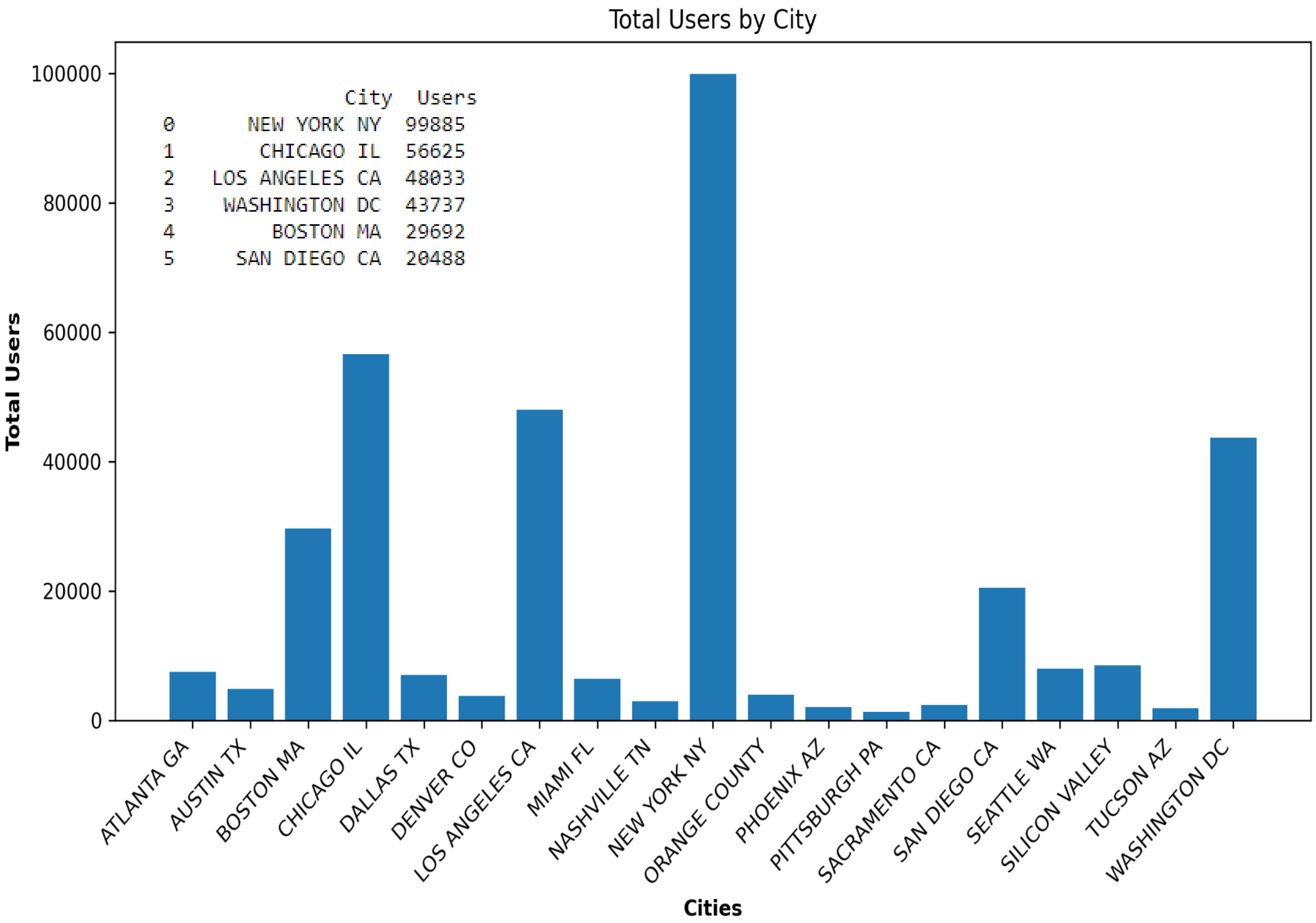


Fig2:- Average users per company

EDA

(Users by City)

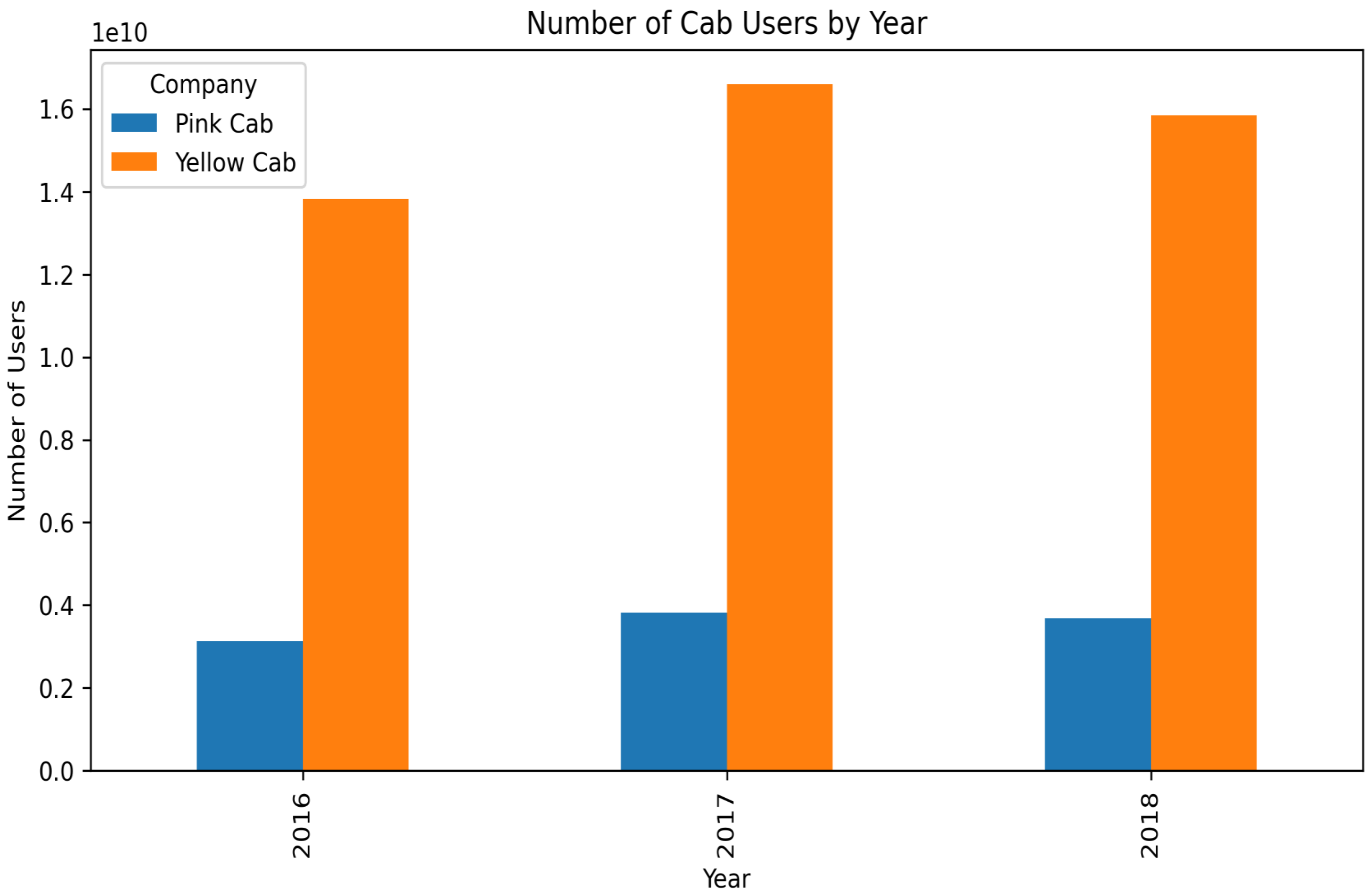
- **Cities such as New York, NY and Chicago, IL have significantly higher numbers of cab service users compared to other cities.**



```
             City  Users
0     NEW YORK NY  99885
1      CHICAGO IL  56625
2  LOS ANGELES CA  48033
3   WASHINGTON DC  43737
4       BOSTON MA  29692
5    SAN DIEGO CA  20488
```

EDA
(Booking Trends-As per Year)

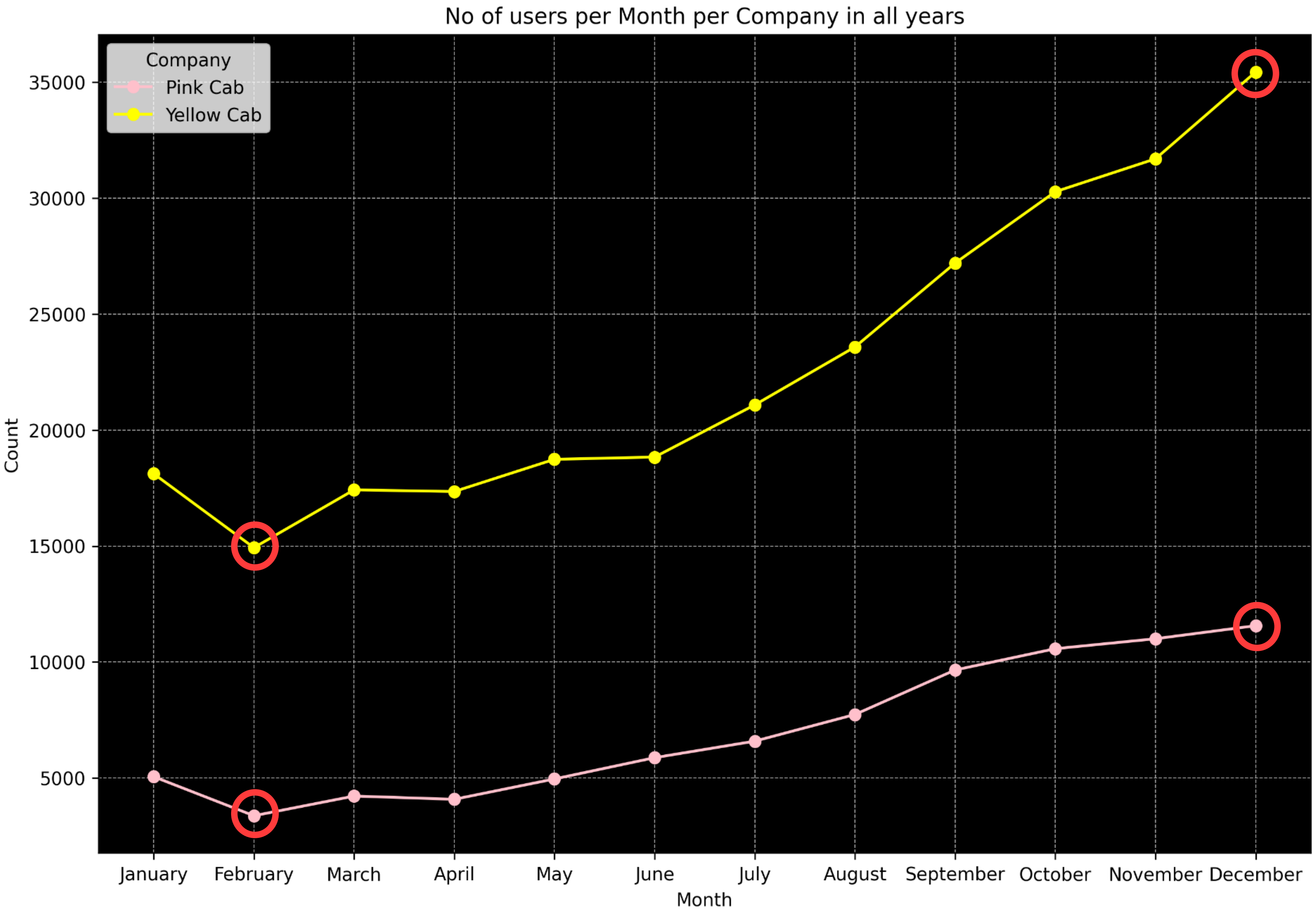
- From the bar chart, we can see that each year the number of users for Yellow Cab is higher compared to Pink Cab, meaning that yellow cab has more bookings.



EDA

(Booking Trends-As per month)

- From the line graph, we can see that the number of bookings for both cab companies increased each month
- Additionally, February has fewer bookings, while December has more bookings compared to the other months.



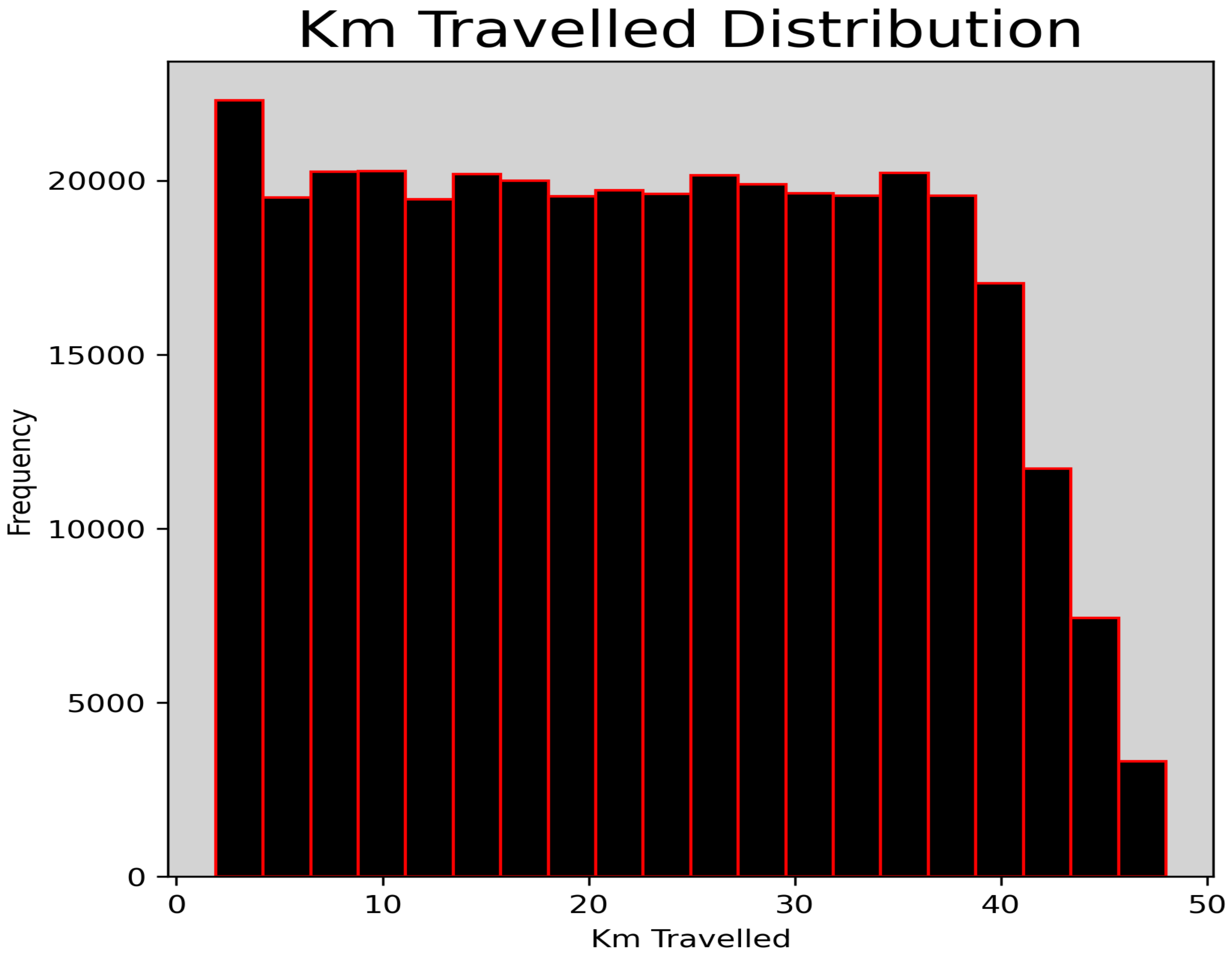
Data Glacier

Your Deep Learning Partner

EDA

(Kilometers Travelled)

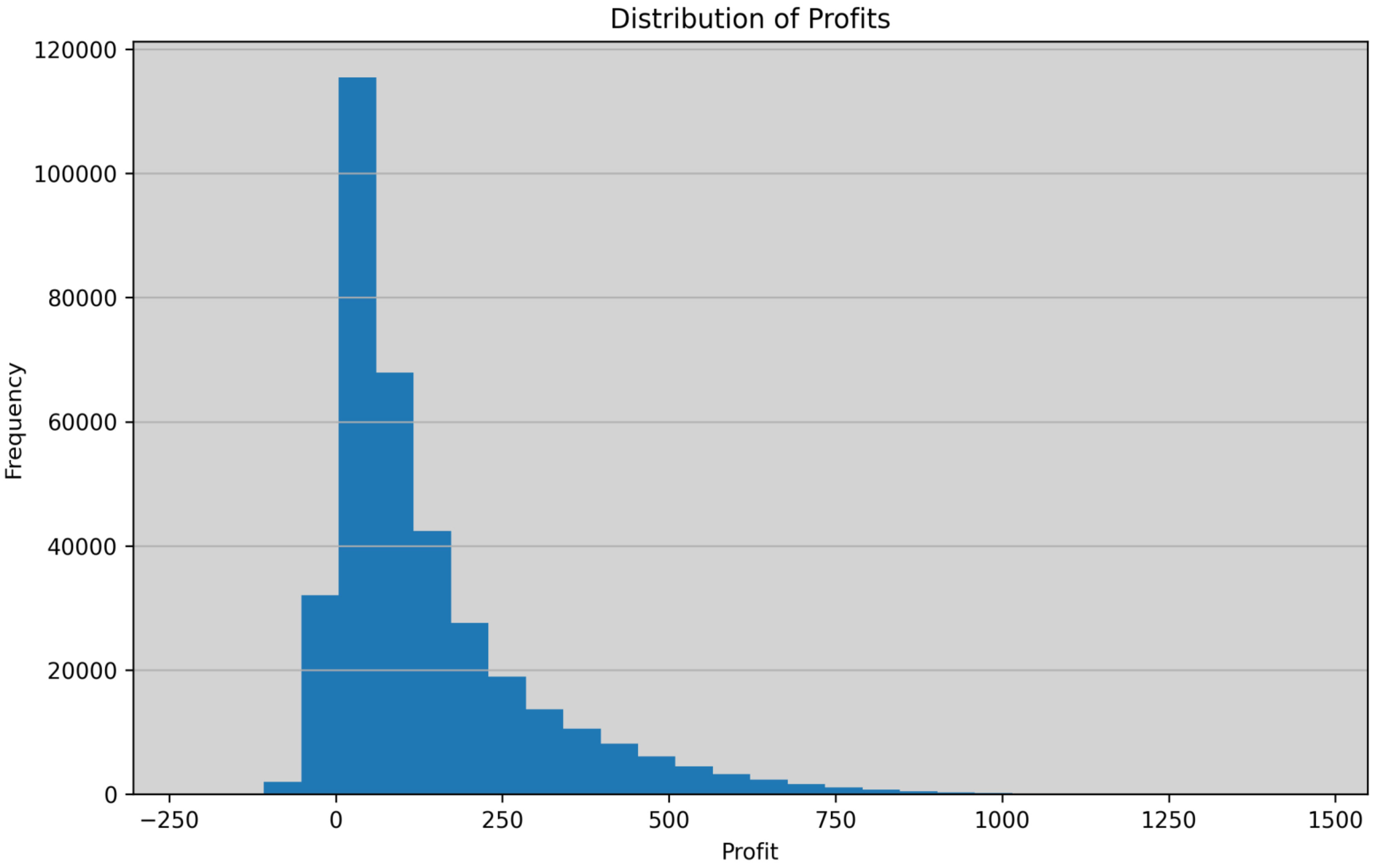
- The majority of cab rides range approximately from 2 km to 48 km.



EDA

(Profit Margins-Distribution of profits)

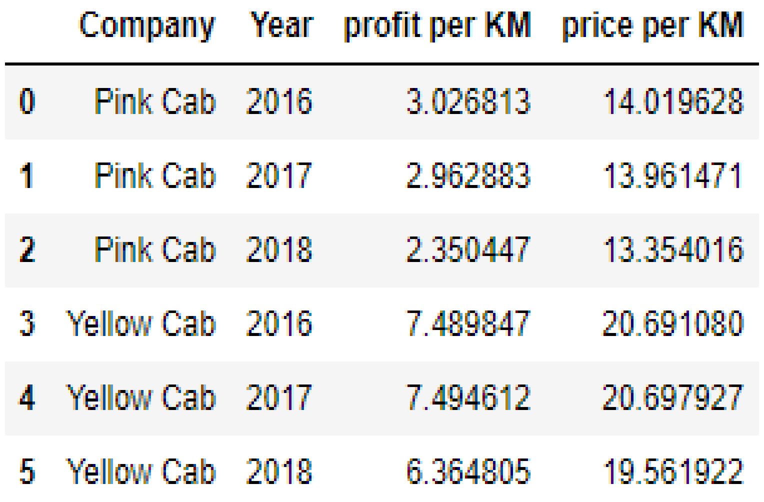
- The majority of profits fall within the range of approximately 10 to 140. Although there are some losses, most of the results are profitable



EDA

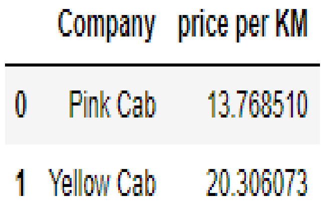
(Profit Margins-Pricing Analysis)

- On average, Yellow Cab charges higher fares and earns more profits compared to Pink Cab. However, Pink Cab offers cheaper fares in comparison to Yellow Cab.

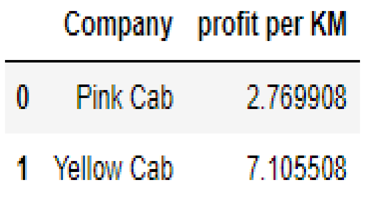


	Company	Year	profit per KM	price per KM
0	Pink Cab	2016	3.026813	14.019628
1	Pink Cab	2017	2.962883	13.961471
2	Pink Cab	2018	2.350447	13.354016
3	Yellow Cab	2016	7.489847	20.691080
4	Yellow Cab	2017	7.494612	20.697927
5	Yellow Cab	2018	6.364805	19.561922

:- Average price and profit per year

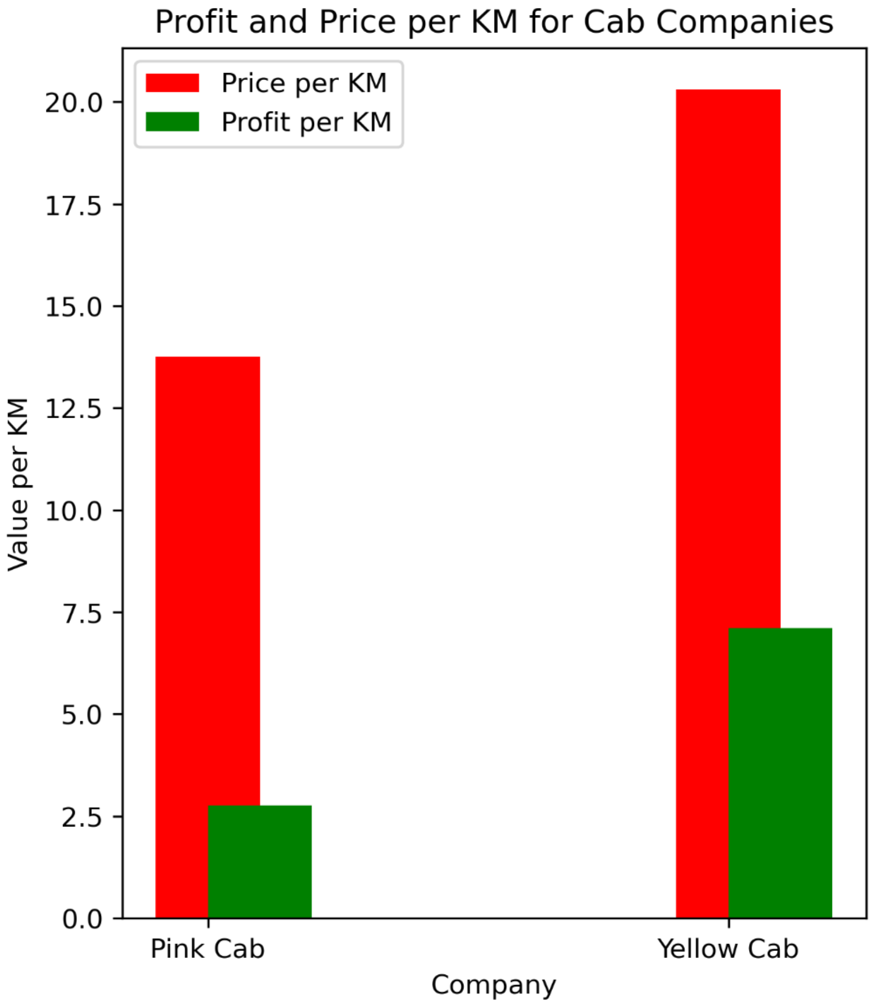


	Company	price per KM
0	Pink Cab	13.768510
1	Yellow Cab	20.306073



	Company	profit per KM
0	Pink Cab	2.769908
1	Yellow Cab	7.105508

Fig:-Average prices and profit

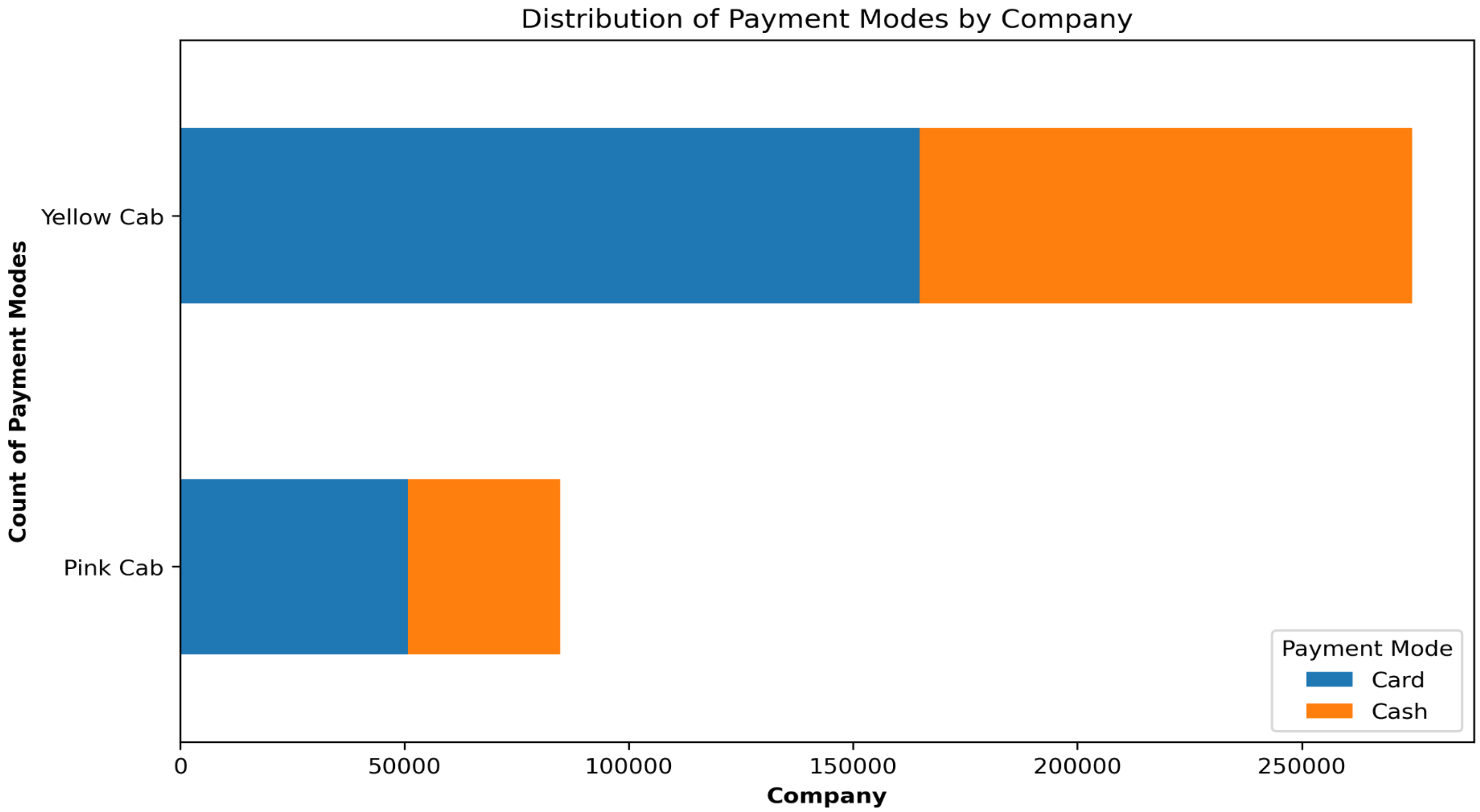


Profit and Price per KM for Cab Companies

EDA

(Payment Preferences)

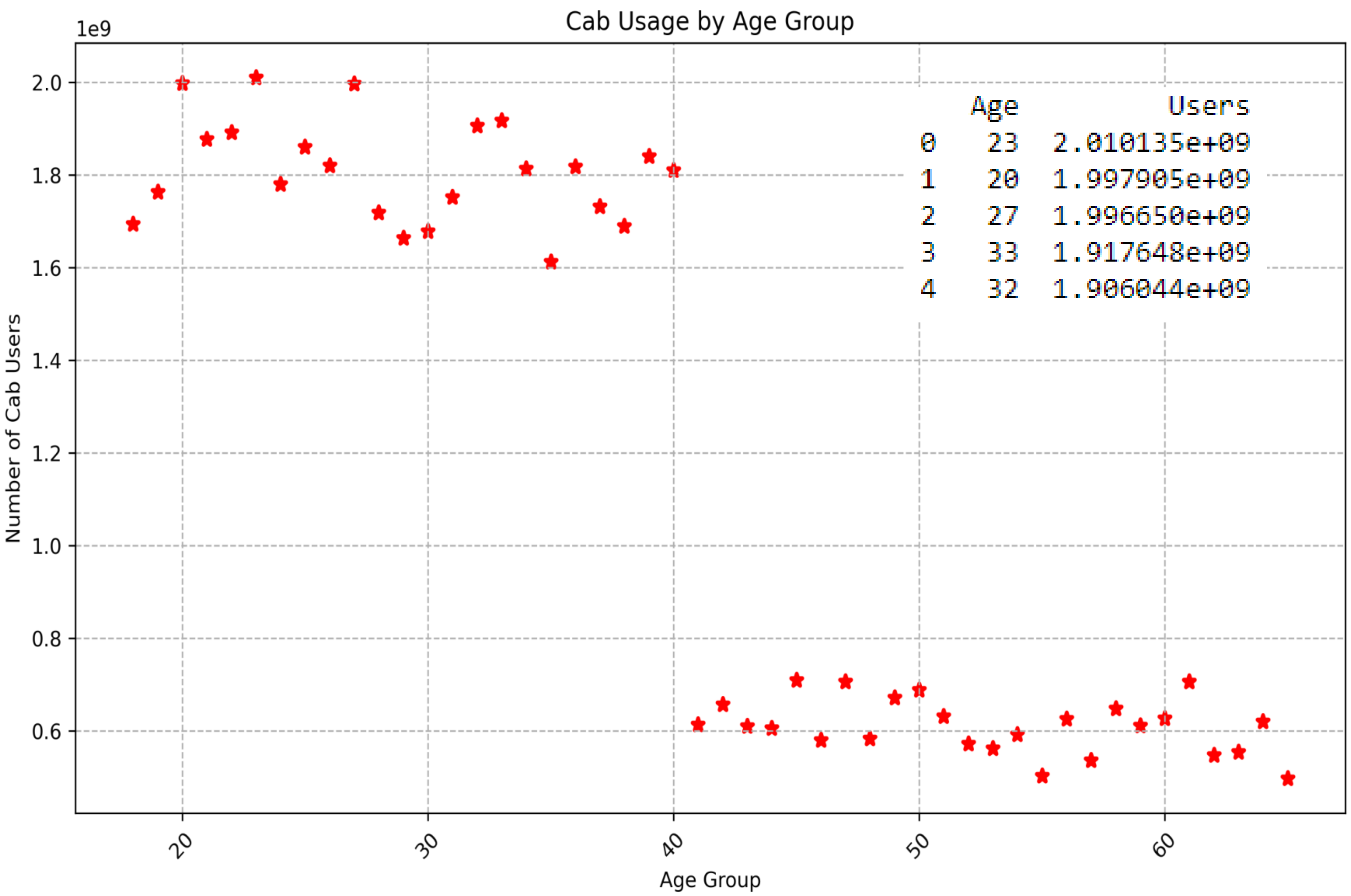
- The stacked bar graph shows that most payments are made by card in both companies.



EDA

(Age Groups)

- Users aged between 20 and 40 years old show higher cab usage compared to those aged 40 and above with most users being 23 years old.



```
   Age         Users
0   23  2.010135e+09
1   20  1.997905e+09
2   27  1.996650e+09
3   33  1.917648e+09
4   32  1.906044e+09
```

Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or inferences about a population based on sample data.

Steps:-

1. Define the following:

- Null Hypothesis (H0) : statement asserting that there is no effect or no difference.
- Alternative Hypothesis (H1) : statement suggesting that there is an effect or a difference.

2. Find the p-value

- Since the population standard deviation(sigma) is unknown, we use the t-distribution.
- Calculate the test statistic and corresponding p-value using the sample data.

3. Reject or Fail to Reject the Hypothesis

- Compare the p-value with the significance level (alpha), typically set at 0.05.

  Decision Rule:

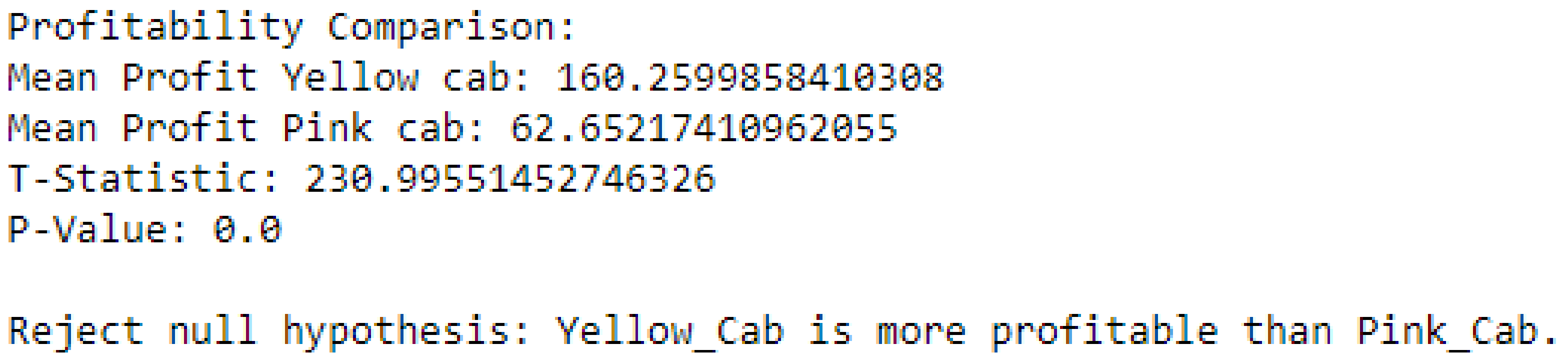
  - If (p)-value < (alpha): Reject the null hypothesis (H0).
  - If (p)-value > (alpha): Fail to reject the null hypothesis (H0).

Conclusion: Based on the decision rule, determine if there is enough evidence to conclude that there is a significant difference.

1. Profitability Comparison

Hypothesis Testing to Determine if Yellow Cab is More Profitable.

- Null Hypothesis (Ho): Yellow Cab is not more profitable than Pink Cab.
- Alternative Hypothesis (H1):Yellow Cab is more profitable than Pink Cab.



```
Profitability Comparison:
Mean Profit Yellow cab: 160.2599858410308
Mean Profit Pink cab: 62.6521741096205
T-Statistic: 230.99551452746326
P-Value: 0.0

Reject null hypothesis: Yellow_Cab is more profitable than Pink_Cab.
```

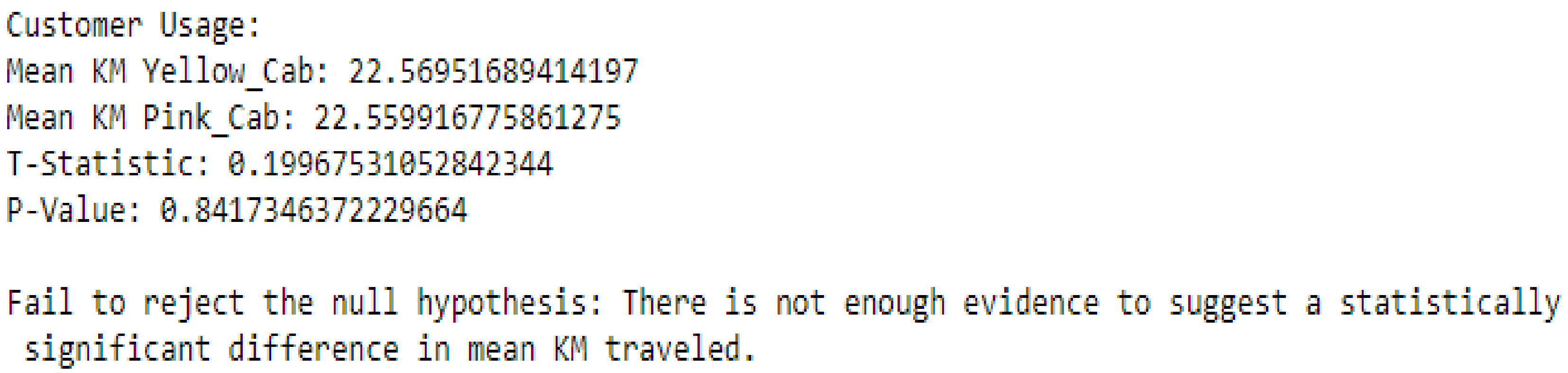
Conclusions:

Yellow Cab has a much higher mean profit of 160 compared to Pink Cab, which has a mean profit of 62. This suggests that, on average, Yellow Cab is significantly more profitable. Additionally, the p-value is 0, so we reject the null hypothesis, meaning Yellow Cab is more profitable than Pink Cab. This evidence indicates not only a significant difference in mean profits but also a practical and substantial difference in profitability between the two cab companies.

2. Customer Usage

Hypothesis Testing to see if the Customers of yellow cab use the service more frequently (km travelled) than those of pink cab.

- Null Hypothesis (Ho): There is no difference in the mean kilometres travelled by customers of yellow cab and Pink cab..
- Alternative Hypothesis (H1):There is a difference in the mean kilometres travelled by customers of yellow cab and Pink cab.



```
Customer Usage:
Mean KM Yellow_Cab: 22.56951689414197
Mean KM Pink_Cab: 22.559916775861275
T-Statistic: 0.19967531052842344
P-Value: 0.8417346372229664

Fail to reject the null hypothesis: There is not enough evidence to suggest a statistically
 significant difference in mean KM traveled.
```

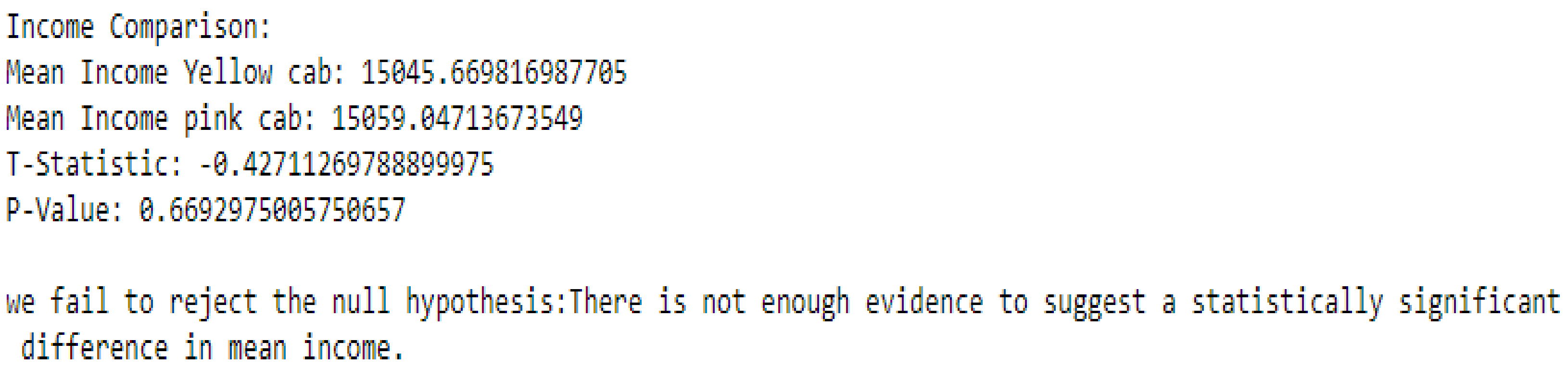
Conclusions:

The difference in the mean between the two companies is very small. Additionally, the p-value is 0.84 (high), so we fail to reject the null hypothesis, meaning there is no significant difference in the kilometres travelled by both companies. This indicates that the small difference in mean kilometres travelled (0.01 kilometres) is likely due to random variability rather than a true difference in usage patterns.

3. Income

Hypothesis Testing to see if Income effect the choice of cab company?

- Null Hypothesis (Ho): The mean income of users of yellow cab is equal to the mean income of users of Pink cab.
- Alternative Hypothesis (H1):The mean income of users of yellow cab is different from the mean income of users of Pink cab.



```
Income Comparison:
Mean Income Yellow cab: 15045.669816987705
Mean Income pink cab: 15059.04713673549
T-Statistic: -0.42711269788899975
P-Value: 0.6692975005750657

we fail to reject the null hypothesis:There is not enough evidence to suggest a statistically significant
 difference in mean income.
```

Conclusions:

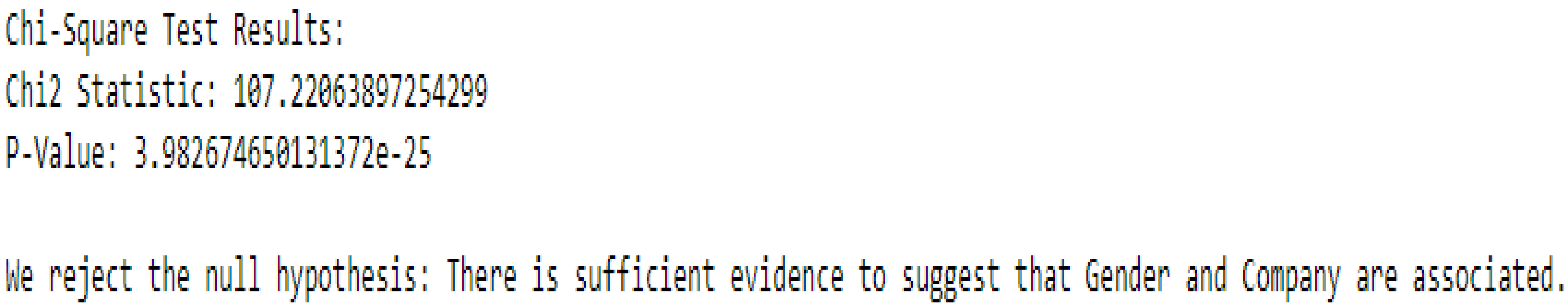
The difference in mean income between the two companies is very small. Additionally, the p-value is 0.66 (high), so we fail to reject the null hypothesis, meaning that income does not affect the choice of cab. This indicates that the small difference in mean income is likely due to random variability rather than a true difference in income levels.

4. Gender

Hypothesis Testing to see if Gender effect the choice of cab company?

- Null Hypothesis (Ho): There is no association between Gender and Company.
- Alternative Hypothesis (H1):There is an association between Gender and Company.

Note: We use the chi-square test to see if there's a connection between two types of categories. In this case, the two categorical variables are gender (e.g., male, female) and cab company preference (e.g., yellow, Pink). The chi-square test helps determine whether the distribution of one categorical variable differs depending on the level of the other categorical variable.



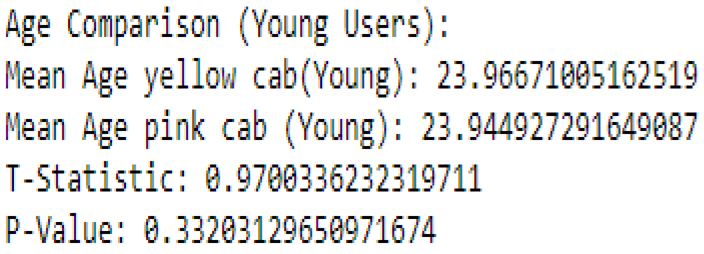
```
Chi-Square Test Results:
Chi2 Statistic: 107.22063897254299
P-Value: 3.982674650131372e-25

We reject the null hypothesis: There is sufficient evidence to suggest that Gender and Company are associated.
```

Conclusions:
Based on the very small p-value (much less than 0.05), we reject the null hypothesis. Therefore, we conclude that there is a statistically significant association between 'Gender' and 'Company'.

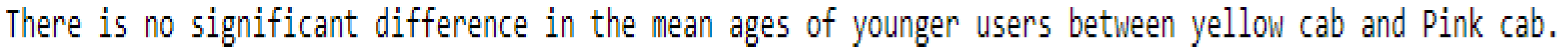
5. Demographic Preference

Hypothesis Testing to see Is Yellow cab is preferred by younger users compared to Pink cab ?

- Null Hypothesis (Ho): There is no significant difference in the mean ages of younger users (defined as ages 10-30) between yellow cab and Pink cab.
- Alternative Hypothesis (H1):There is a significant difference in the mean ages of younger users between yellow cab and Pink cab.



```
Age Comparison (Young Users):
Mean Age yellow cab(Young): 23.96671005162519
Mean Age pink cab (Young): 23.944927291649087
T-Statistic: 0.9700336232319711
P-Value: 0.33203129650971674
```



```
There is no significant difference in the mean ages of younger users between yellow cab and Pink cab.
```

Conclusions:
Based on the p-value (0.33), we fail to reject the null hypothesis. There is no significant difference in the mean ages of younger users between Yellow Cab and Pink Cab Company B. Both companies appear to have similar mean ages for their younger user base.

EDA Summary

- **Distribution of Cab Companies and Users:**
  - Yellow Cab: 80% market share, more average users.
- **Users by City:**
  - New York, NY and Chicago, IL: highest number of users.
- **Booking Trends :**
  - Increasing bookings each month.
  - February: fewer bookings.
  - December: more bookings.
- **Kilometres Travelled:**
  - Majority of rides: 2 km to 48 km.
- **Profit Margins:**
  - Profits range: 10 to 140.
  - Yellow Cab: higher fares, more profits.
  - Pink Cab: cheaper fares.
- **Payment Preferences:**
  - Most payments by card for both companies.
- **Age Groups:**
  - Higher usage: ages 20-40, peak at 23 years old.

Recommendations

Investment Decision:

- **Invest in Yellow Cab** due to its dominant market share and higher average profits.
- **Monitor Pink Cab** for potential growth opportunities in emerging markets due to its competitive pricing strategy.

Rationale:

- **Market Dominance**: Yellow Cab's 80% market share indicates a strong position and brand recognition.
- **Profitability**: Yellow Cab has higher fares and profit margins, making it a more profitable investment.
- **Customer Base**: Yellow Cab is preferred by more users in major cities, indicating reliability and customer loyalty.

Thank You

Data Glacier

Your Deep Learning Partner