Exploratory Data Analysis

G2M Insights for Cab Investment Firm

By Bisma Azeem (Cohort-LISUM32)

21-04-2024

Agenda

Executive Summary

Problem Statement

EDA for XYZ

Comparative Analysis

The Master Dataset

Hypotheses Testing

Recommendations

Executive Summary

Introduction:

XYZ, a US-based private firm, seeks to capitalize on the recent growth surge in the cab industry. This analysis aims to identify the optimal investment opportunity through a data-driven evaluation of two key players.

- Four datasets encompassing customer transactions, demographics, payment methods, and city details were provided.
- Comprehensive data exploration of both potential candidates, involved understanding field names, identifying relationships, and performing transformations.
- A master dataset was created by joining relevant tables and addressing duplicates, missing values, and outliers.
- Several hypotheses were formulated and investigated to gain deeper understanding.
- Based on the analysis of usage patterns, profitability, and market trends, a recommendation for the company presenting the more attractive investment opportunity for XYZ will be provided.

Benefits for XYZ:

This data-driven analysis offers valuable insights into the competitive landscape of the cab industry, enabling XYZ to:

- Make informed investment decisions based on comprehensive data analysis.
- Identify the company with greater customer base, profitability potential, and market fit.
- Gain a competitive advantage by understanding customer behavior and trends.

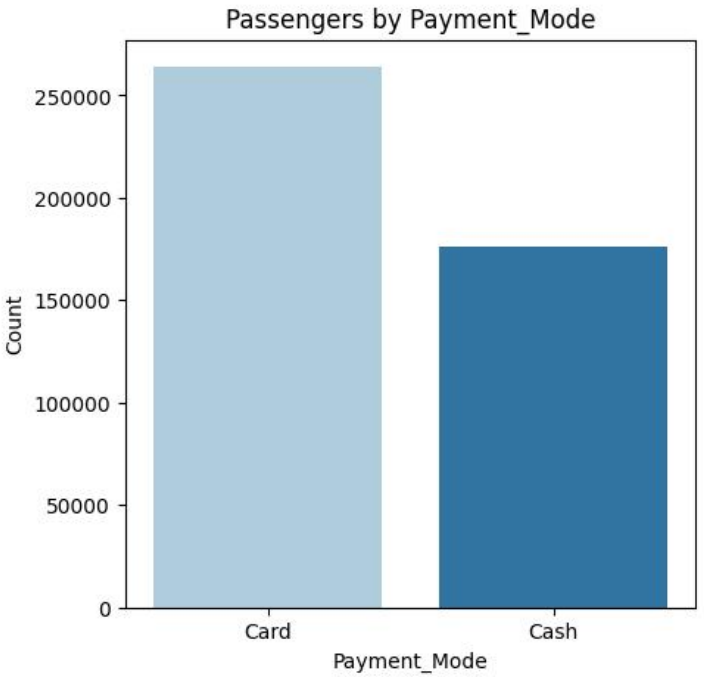
Problem Statement:

XYZ, a US private firm, is seeking to invest in the growing cab industry. However, with multiple key players in the market, XYZ needs to make a strategic decision to maximize their return on investment. To achieve this, they require a data-driven analysis to identify the cab company that offers the most promising investment opportunity based on factors like customer base, profitability, market positioning, and alignment with XYZ's Go-to-Market (G2M) strategy.

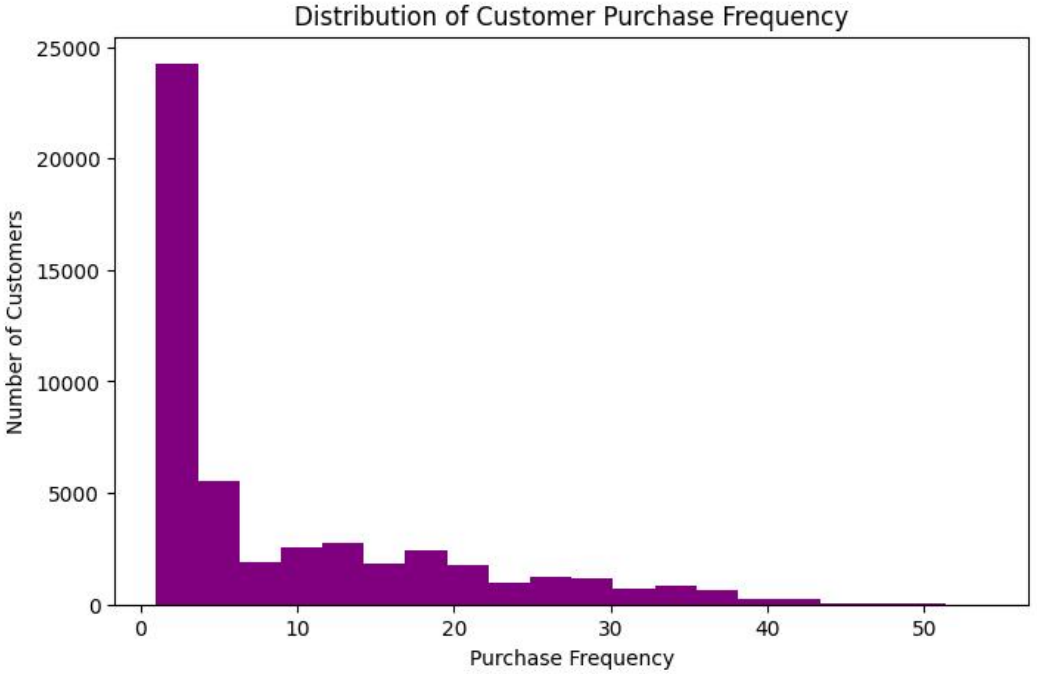
Unveiling Investment Potential: A data-driven Analysis for XYZ

A General Overview

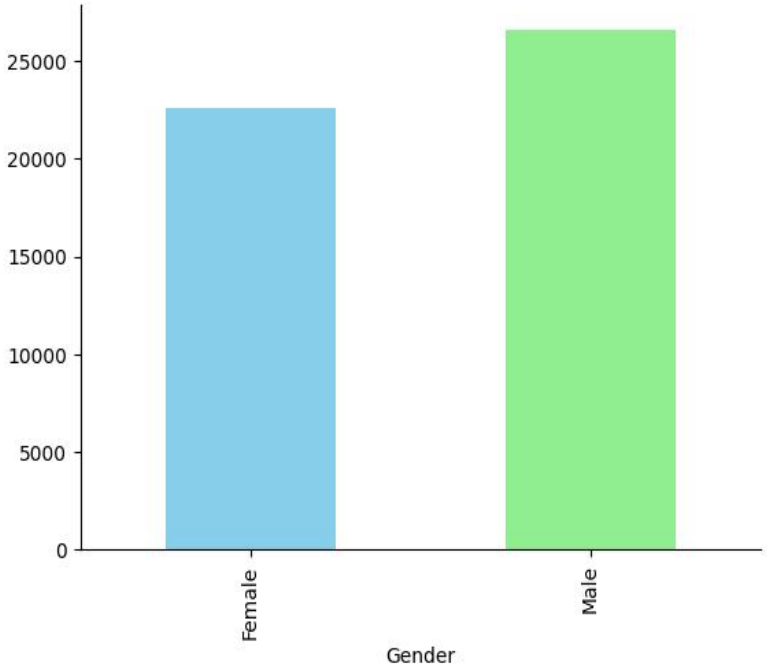
"Card" is the preferred mode of Payment by overall Customers



The majority of customers (around 60%) purchase between 1 and 10 times overall. There is a significant drop-off in the number of customers who purchase between 11 and 20 times.



The customer base appears to be relatively balanced between genders.

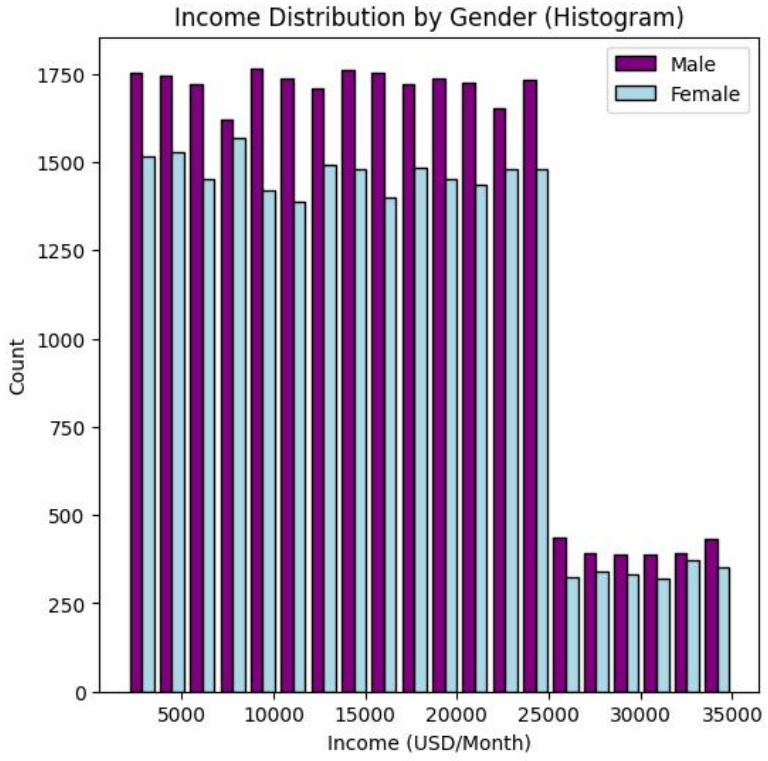


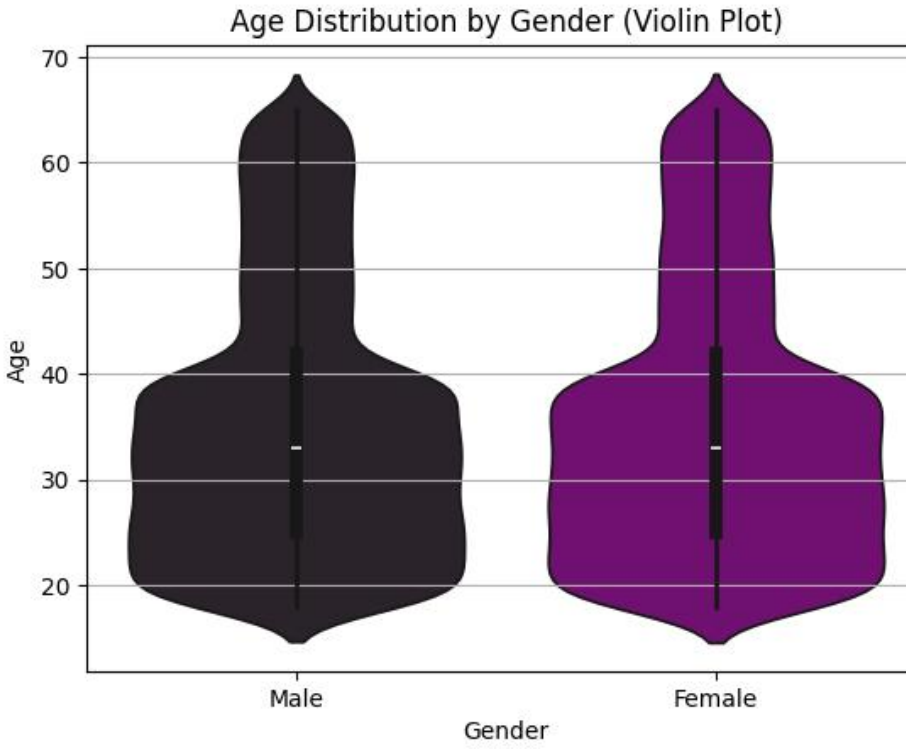
Data Glacier

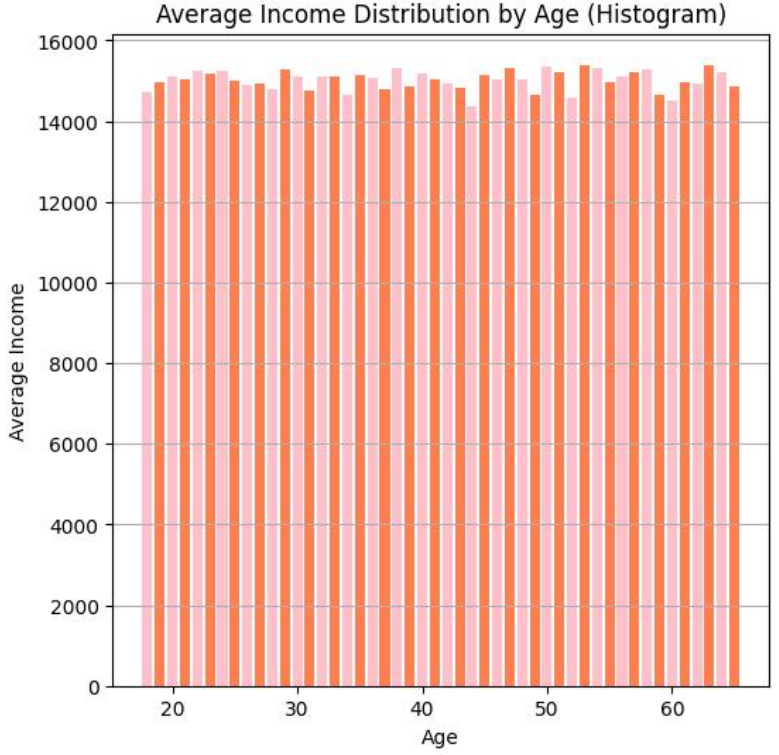
Male customers appear to have higher earnings compared to females, as evidenced by the income distribution.

The data indicates a relatively balanced age distribution across genders when it comes to customer base.

The graph suggests that, on average, males and females within the customer base have comparable income levels.

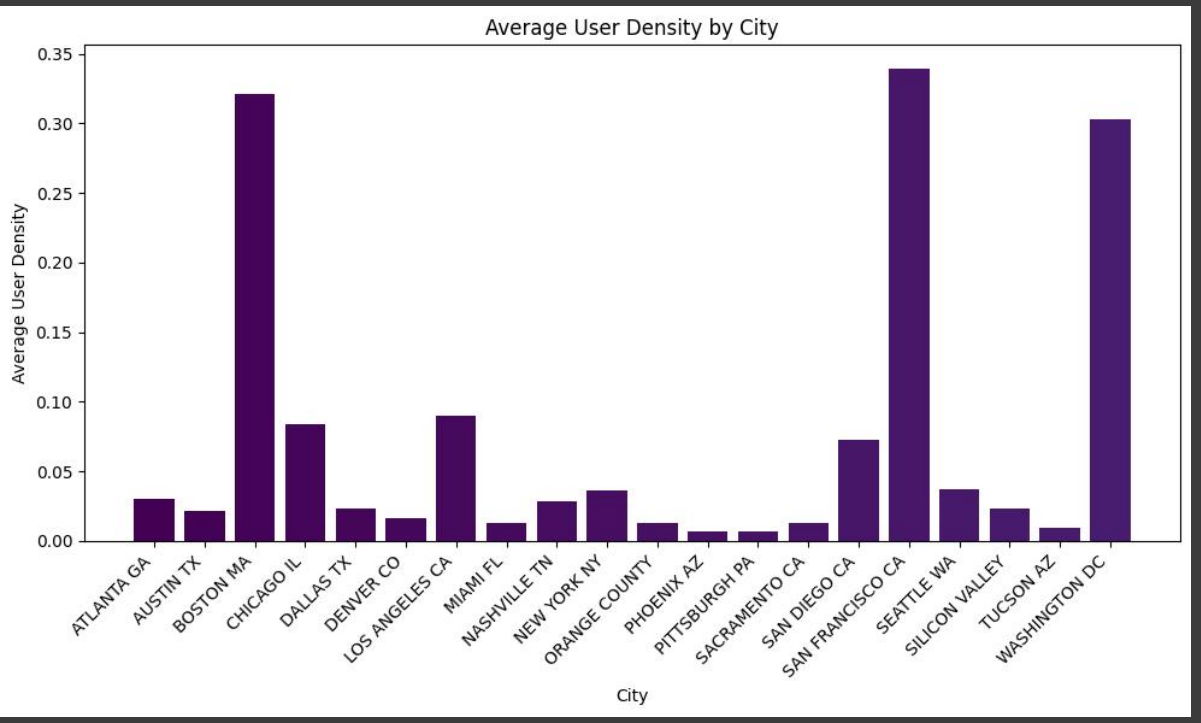


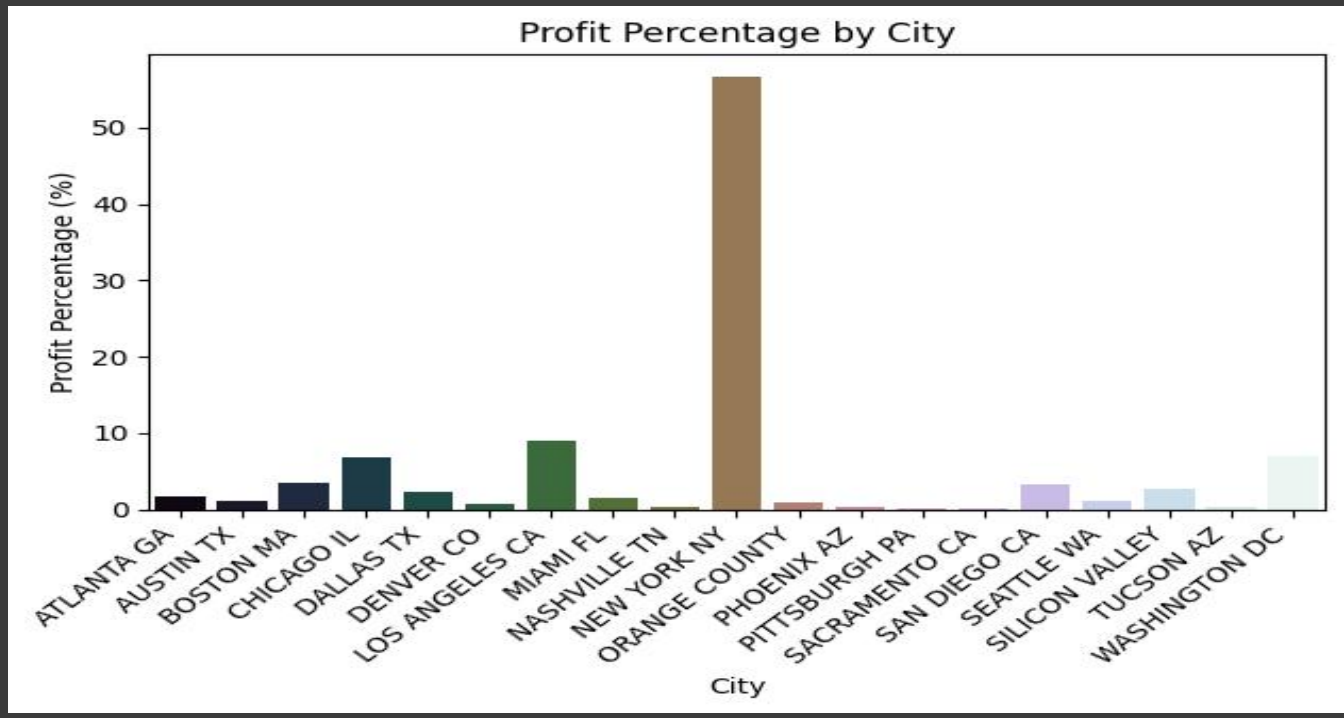




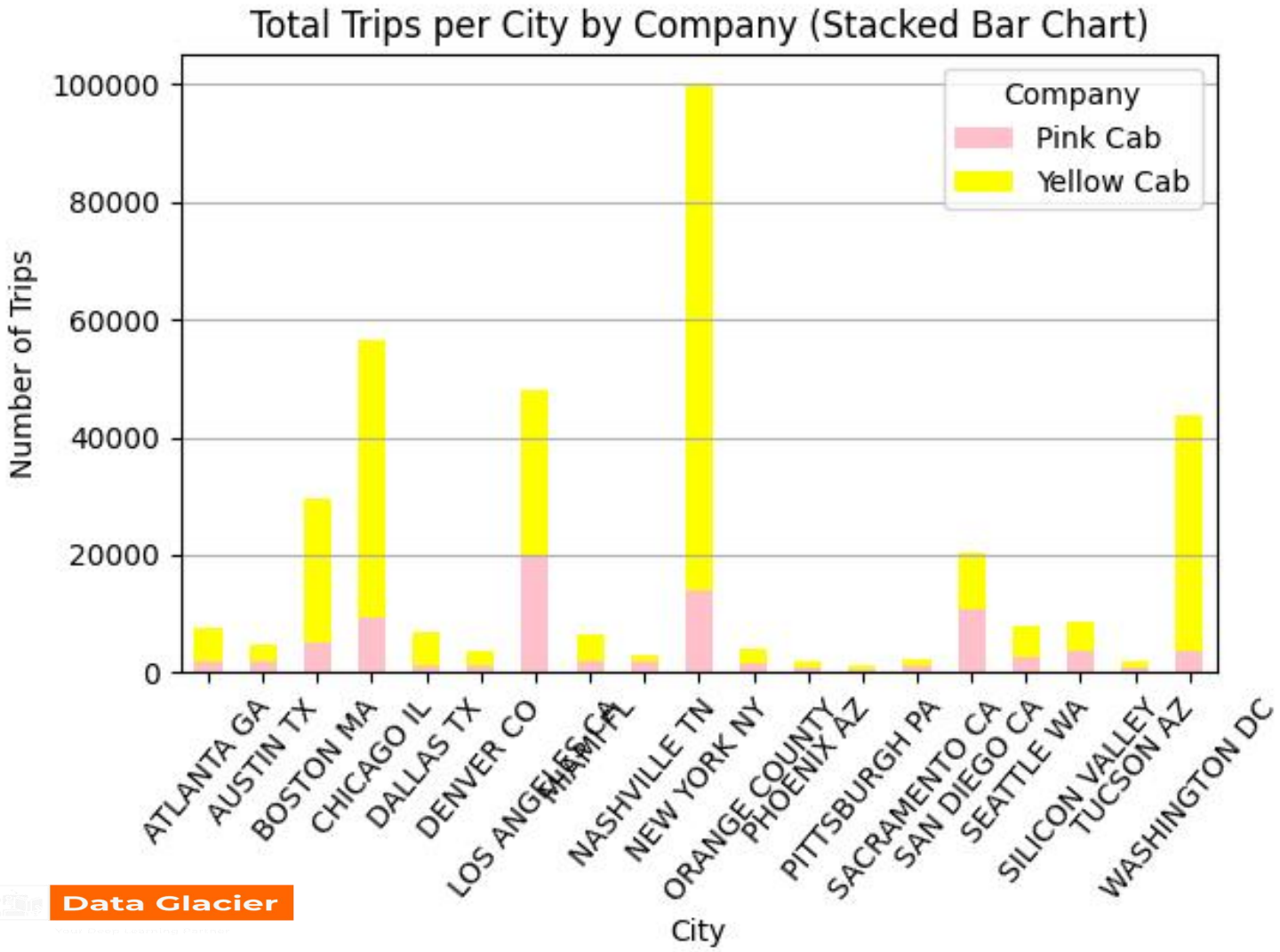
Data Glacier

Although the average user density of NYC is quite low but It generates the most profit percentage among all cities.





Comparative Analysis:
Unveiling Both Companies' Performance

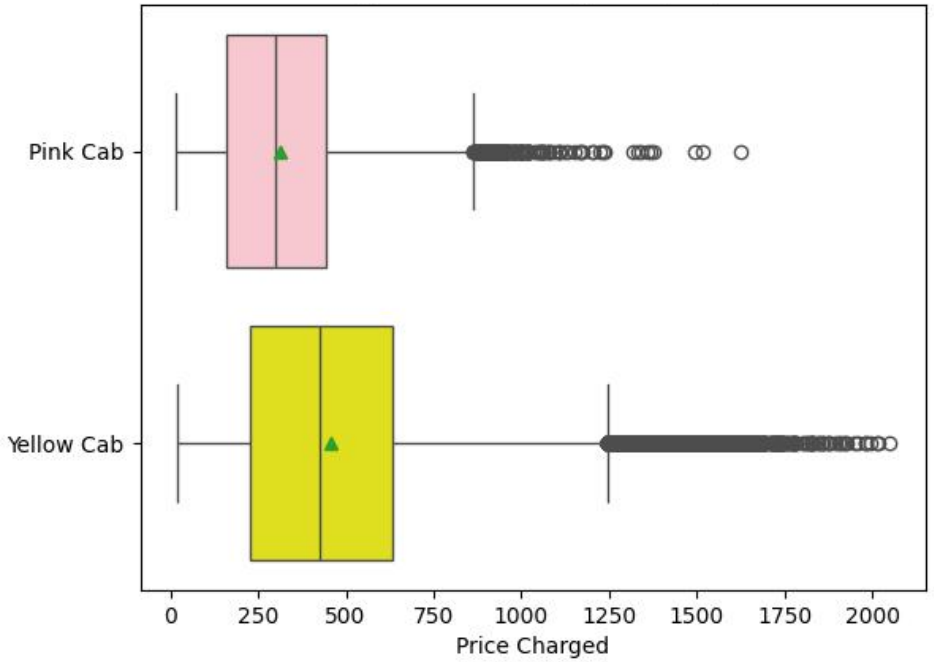


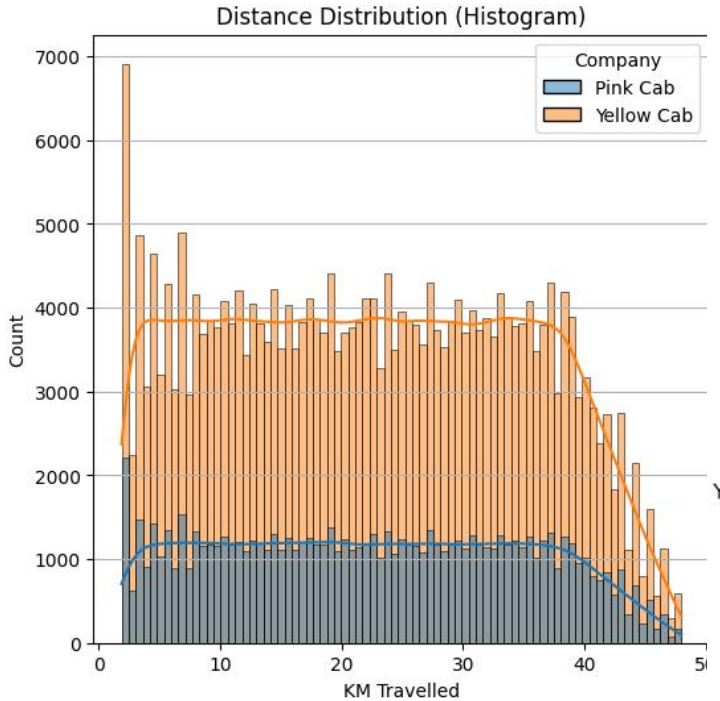
The stacked bar chart indicates that Yellow Cab appears to have a larger share of total rides compared to Pink Cab in almost all cities across the US. The size of the yellow section within each stacked bar suggests this potential dominance.

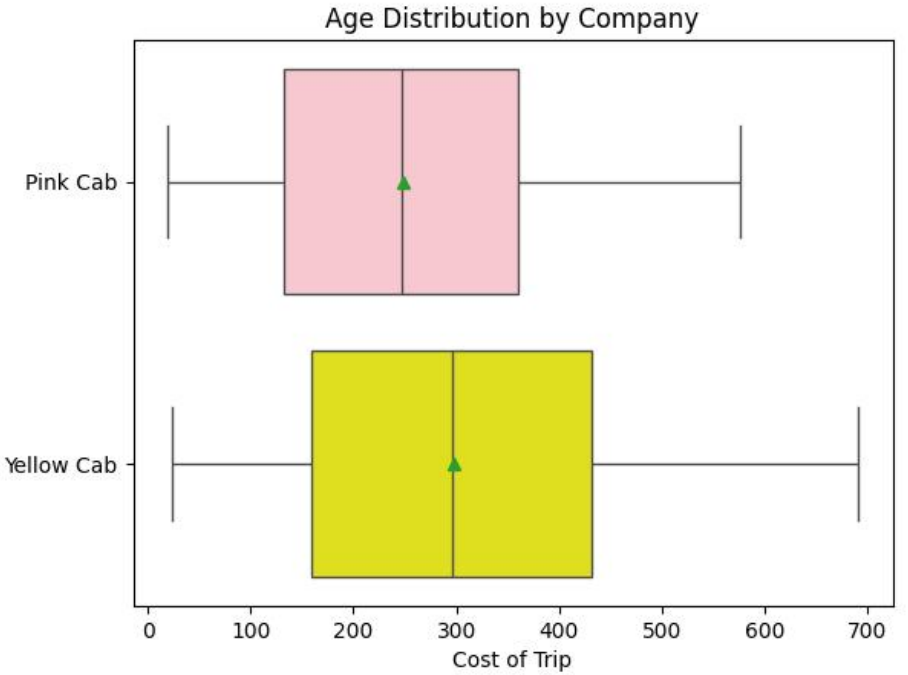
From the visuals below, It is evident that “Yellow Cab” performs relatively much better than the “Pink Cab”. It also seem prefered by customers too.

Data Glacier
Your Deep Learning Partner

Price Charged Distribution by Company

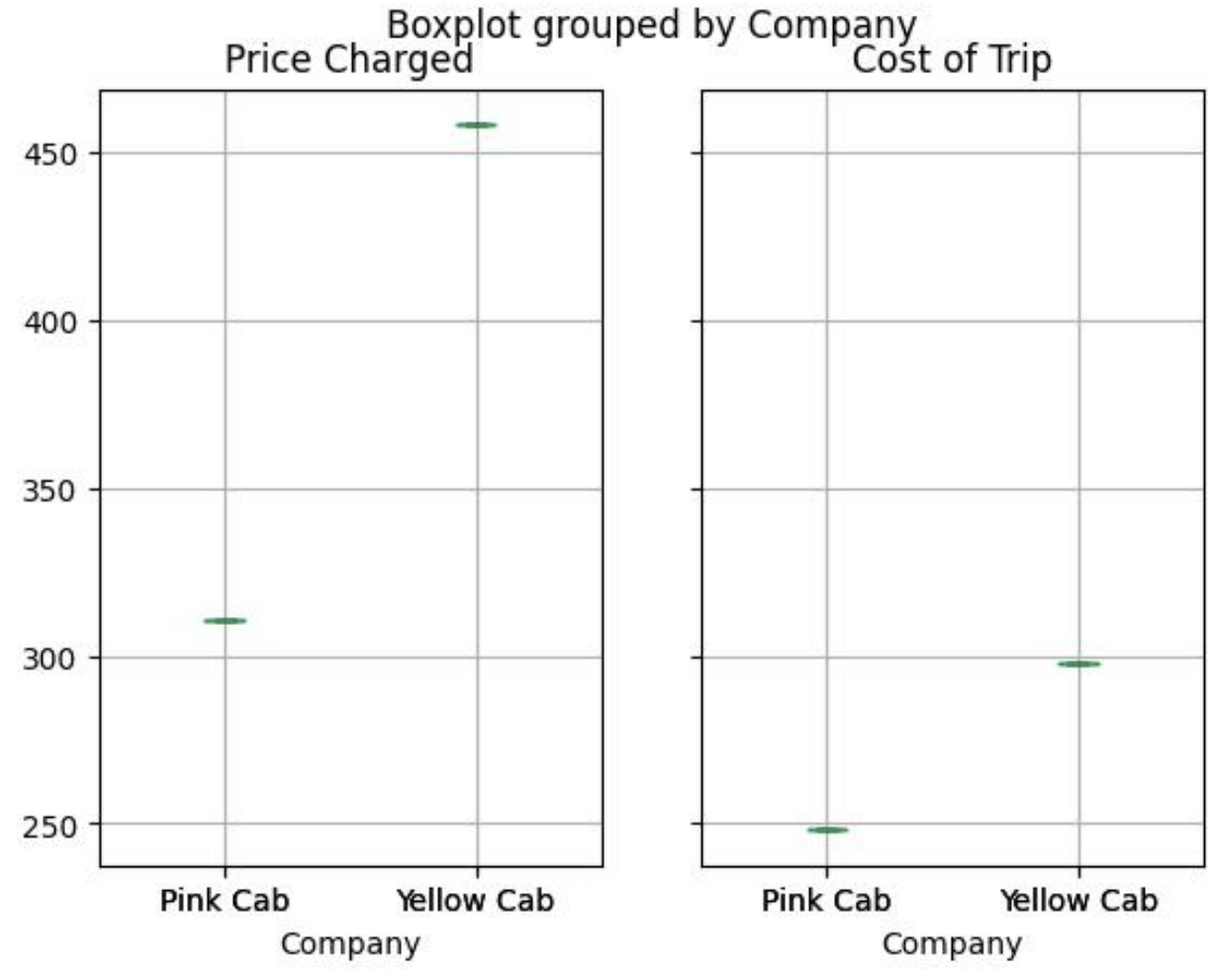


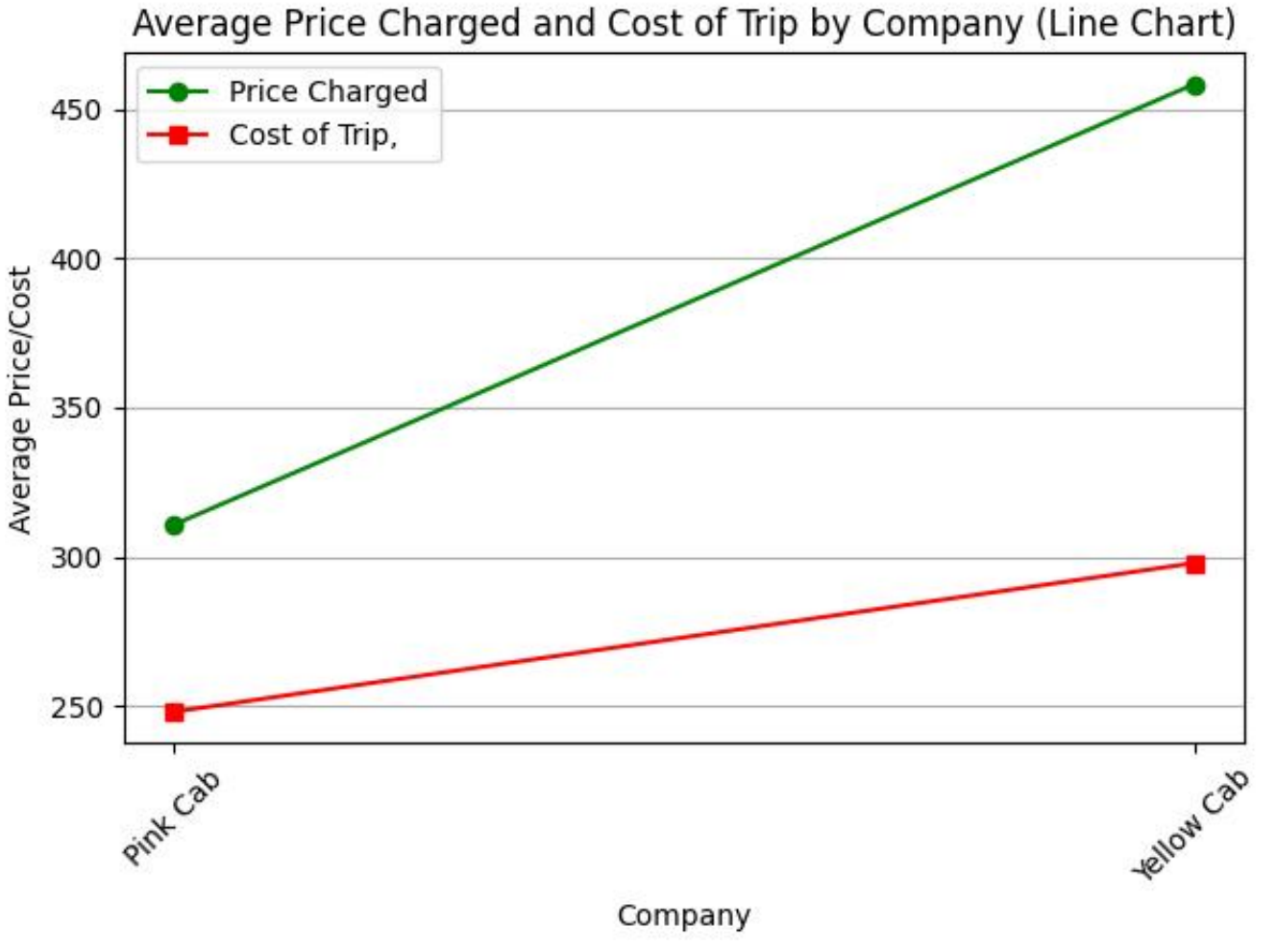




However, There is a huge difference between Cost of Trip and Price Charged for both comapnies, especially for “Yellow Cab”, which can be unfavourable for the company.

Data Glacier
Your Deep Learning Partner





The Master Dataset:

In individual analysis, “Yellow Cab” seems to be doing much better,
so I created a “Master Dataset” by combining all Important attributes,
Derived a few attributes and also merged them all based on “US Holiday” dataset
to further analyze the companies’ performances in Holiday Seasons.

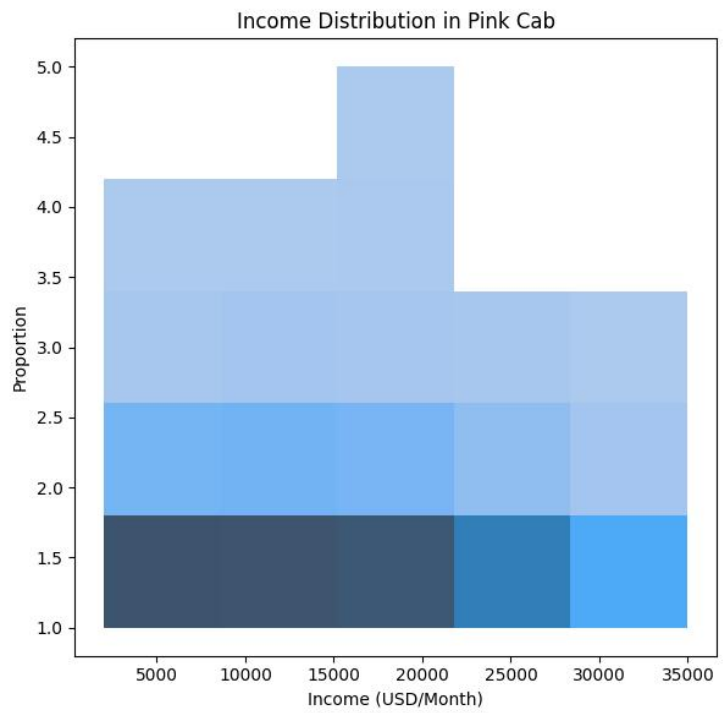
```
master_data.columns
```

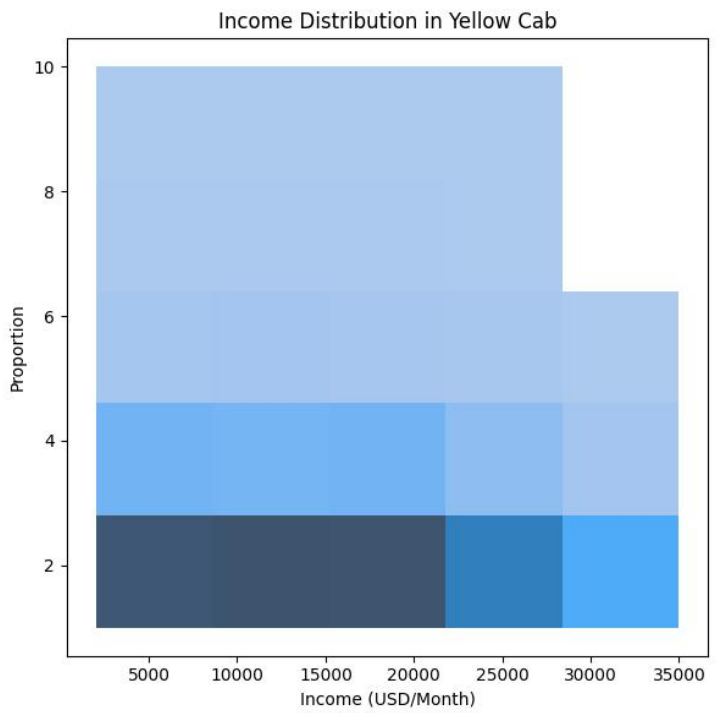
```
Index(['Transaction ID', 'Customer ID', 'Payment_Mode', 'Gender', 'Age',
       'Income (USD/Month)', 'Date', 'Company', 'City', 'KM Travelled',
       'Price Charged', 'Cost of Trip', 'Holiday', 'WeekDay', 'Month', 'Day',
       'Year', 'profit', 'profit_per_Km', 'profit_pctg', 'profit_pctg_per_km'],
      dtype='object')
```

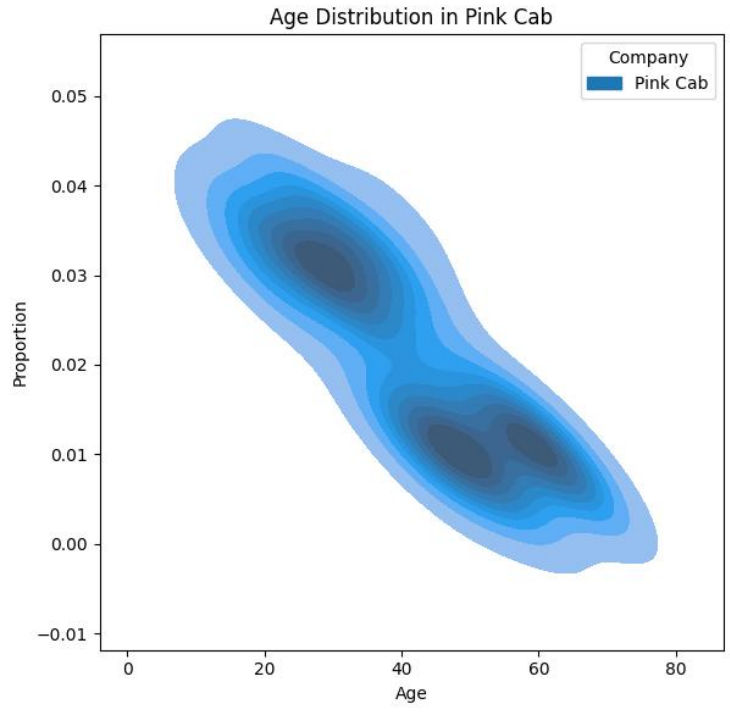
Data Glacier
Your Deep Learning Partner

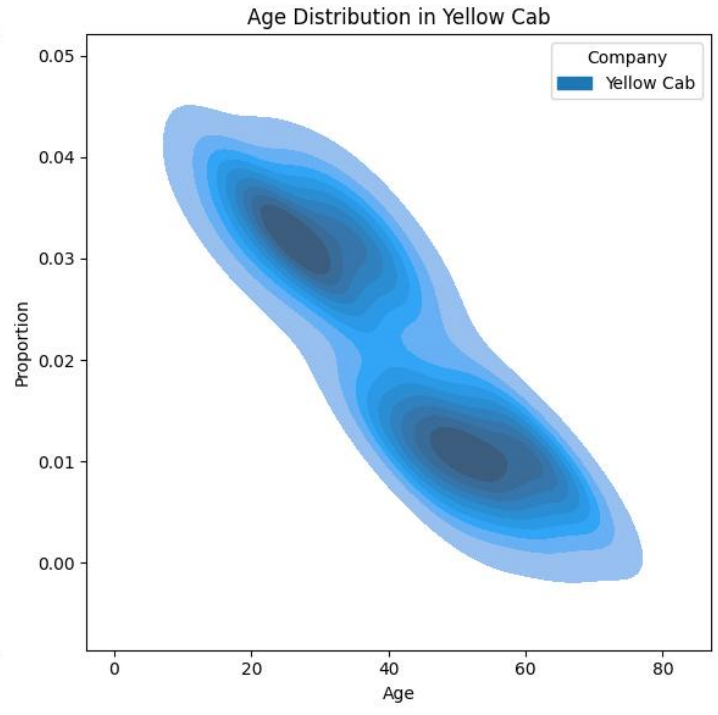
	Transaction ID	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	Date	Company	City	KM Travelled	Price Charged	Cost of Trip	Holiday	WeekDay	Month	Day	Year
19105	10004222	52362	Card	Male	50	10662	2016-01-18	Yellow Cab	WASHINGTON DC	30.74	524.14	442.6560	Martin Luther King, Jr. Day	Monday	1	18	2016
19028	10001874	4548	Card	Male	40	21588	2016-01-18	Yellow Cab	CHICAGO IL	43.32	952.07	519.8400	Martin Luther King, Jr. Day	Monday	1	18	2016
19029	10001915	4731	Cash	Male	28	2770	2016-01-18	Yellow Cab	CHICAGO IL	7.84	167.33	111.9552	Martin Luther King, Jr. Day	Monday	1	18	2016
19030	10001920	5778	Card	Male	19	7032	2016-01-18	Yellow Cab	CHICAGO IL	30.74	643.11	438.9672	Martin Luther King, Jr. Day	Monday	1	18	2016
19031	10001921	4281	Card	Male	65	6853	2016-01-18	Yellow Cab	CHICAGO IL	3.80	73.88	53.3520	Martin Luther King, Jr. Day	Monday	1	18	2016
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11526	10434162	58133	Card	Female	22	12560	2018-12-31	Yellow Cab	BOSTON MA	36.86	496.29	451.1664	New Year's Eve	Monday	12	31	2018
11527	10436755	4416	Cash	Male	59	24183	2018-12-31	Pink Cab	CHICAGO IL	39.24	523.33	443.4120	New Year's Eve	Monday	12	31	2018
11528	10434356	4173	Card	Female	30	18923	2018-12-31	Yellow Cab	CHICAGO IL	8.32	109.88	112.8192	New Year's Eve	Monday	12	31	2018
11521	10434702	25417	Card	Female	33	9597	2018-12-31	Yellow Cab	DALLAS TX	22.88	375.32	280.0512	New Year's Eve	Monday	12	31	2018
11491	10436463	53058	Card	Male	21	21151	2018-12-31	Yellow Cab	WASHINGTON DC	38.76	552.48	553.4928	New Year's Eve	Monday	12	31	2018

19106 rows × 17 columns







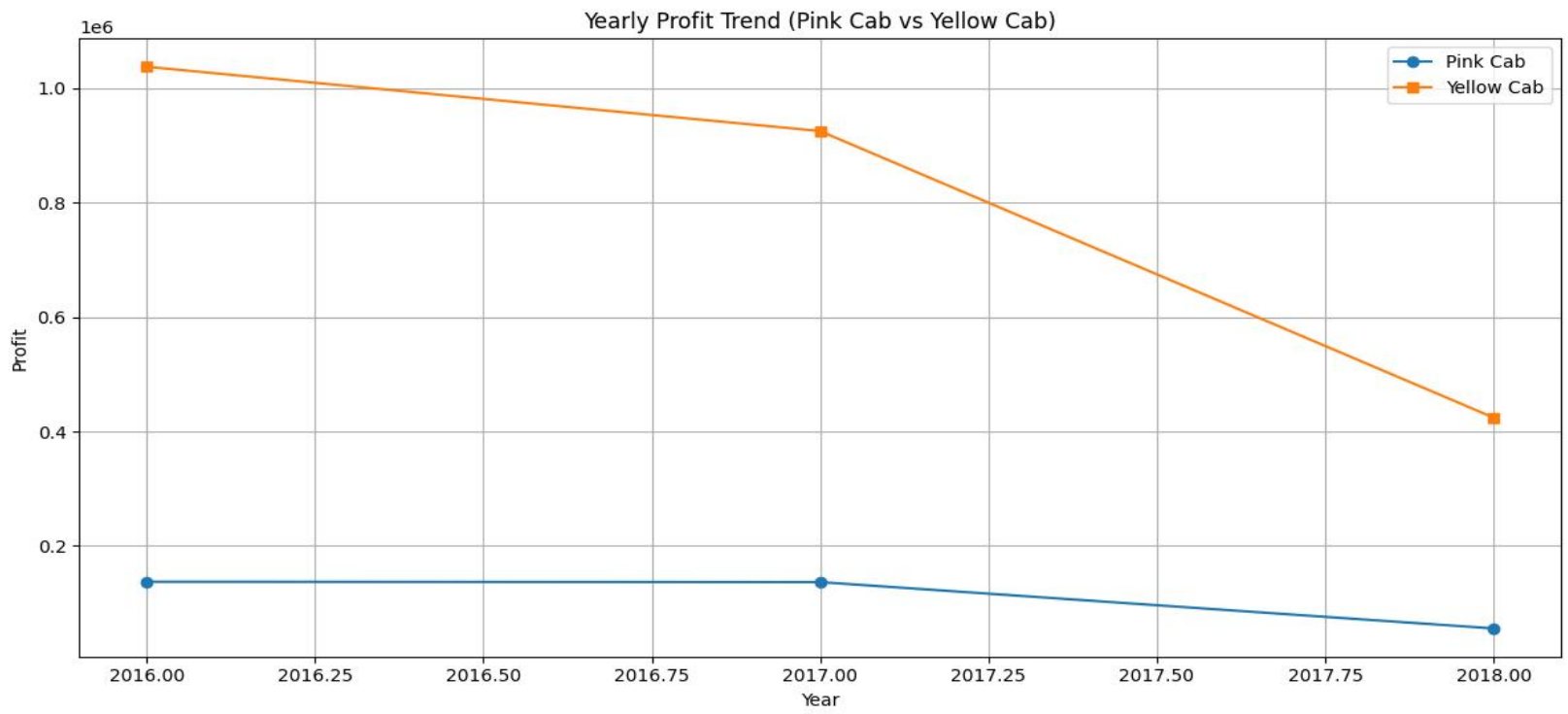


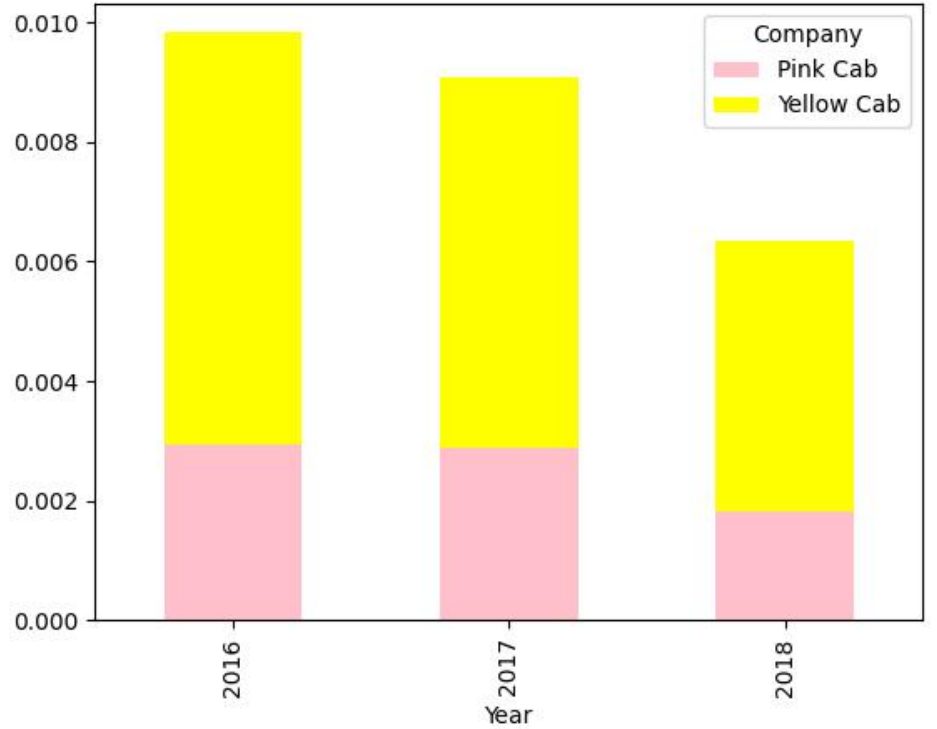
The distribution for Income for both comapnies is comparatively same. However, The Age distribution chart depicts that elderly people (Age 60-70) seems to prefer “Pink Cab” over “Yellow Cab”.

Data Glacier
Your Deep Learning Partner

Hypotheses Testing:

1. “Yellow Cab” Company is more likely to increase profit next year.





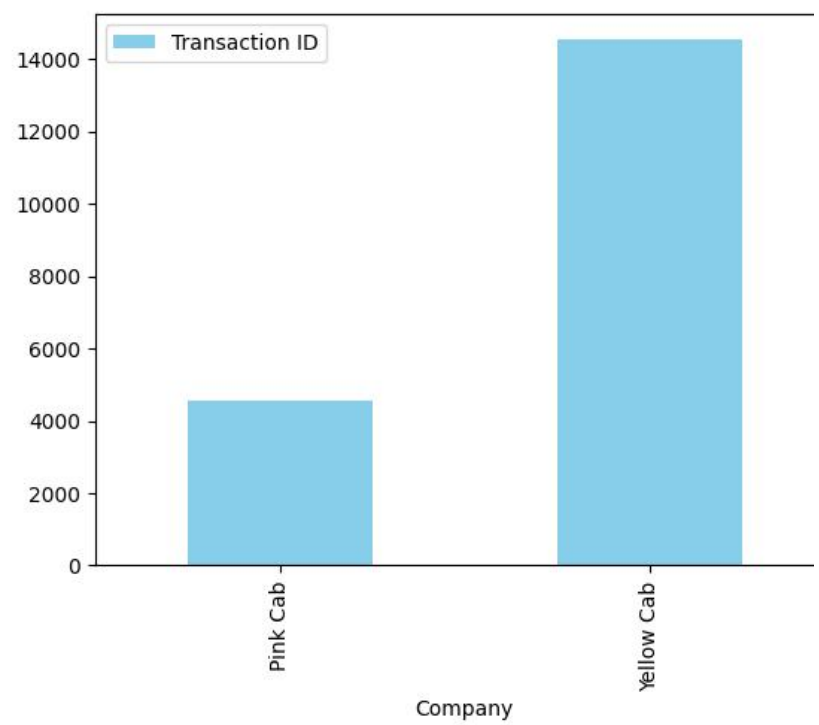
Although Yellow Cab currently leads in ridership, a concerning downward trend in their trips compared to Pink Cab might indicate a shift in the market landscape.

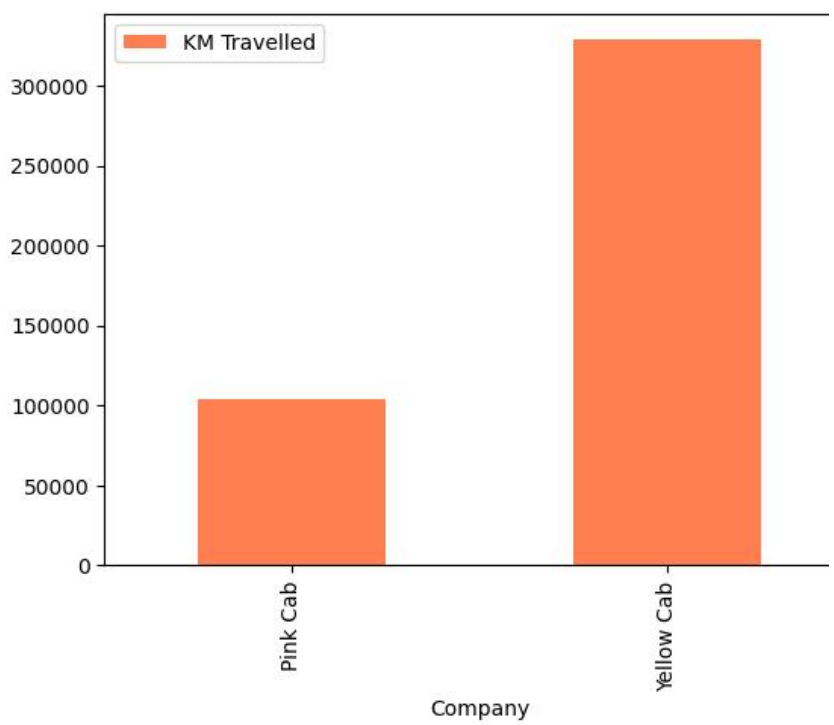
So, Hypothesis # 1 is rejected.

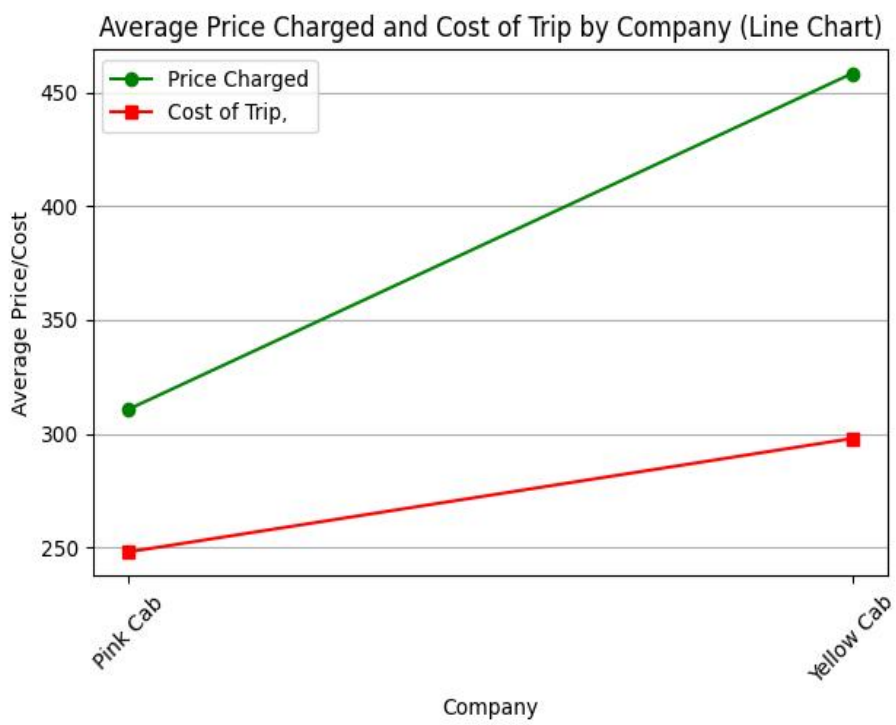
Data Glacier
Your Deep Learning Partner

Hypotheses Testing:

2. “Yellow Cab” has the highest numbers in total # of trips, total_distance, price_charged and cost_of_trip







As evident in the graphs, “Yellow Cab” does have the highest numbers in total # of trips, total_distance, price_charged and cost_of_trip

Hypothesis # 2 accepted.

Hypotheses Testing:

3. “Yellow Cab” company has higher churn rate


```
Estimated Churn Rate for Pink Cab (based on 60-day inactivity): 89.16%

Retention Rate for Pink Cab (2016-01-31 to 2018-12-31): 100.00%

Estimated Churn Rate for Yellow Cab (based on 60-day inactivity): 87.23%

Retention Rate for Yellow Cab (2016-01-31 to 2018-12-31): 100.00%
```

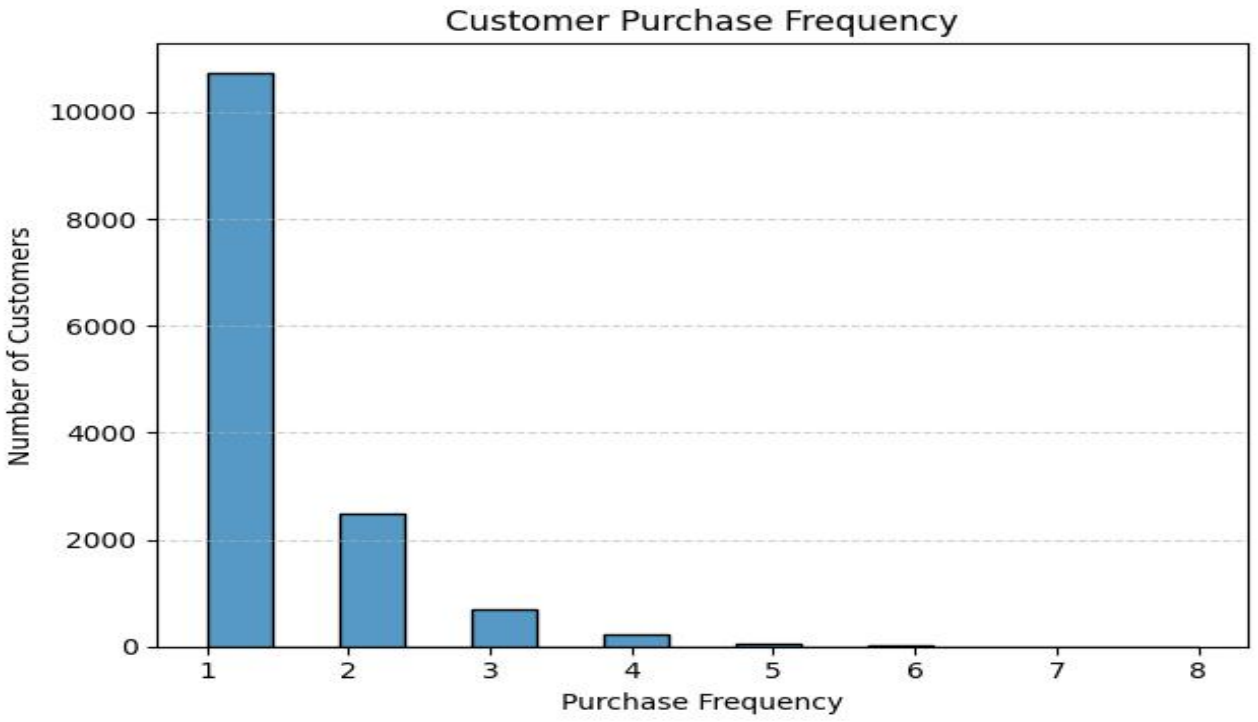
Although the trend shows downfall in the profits but still the estimated churn rate is lower than “Pink Cab”

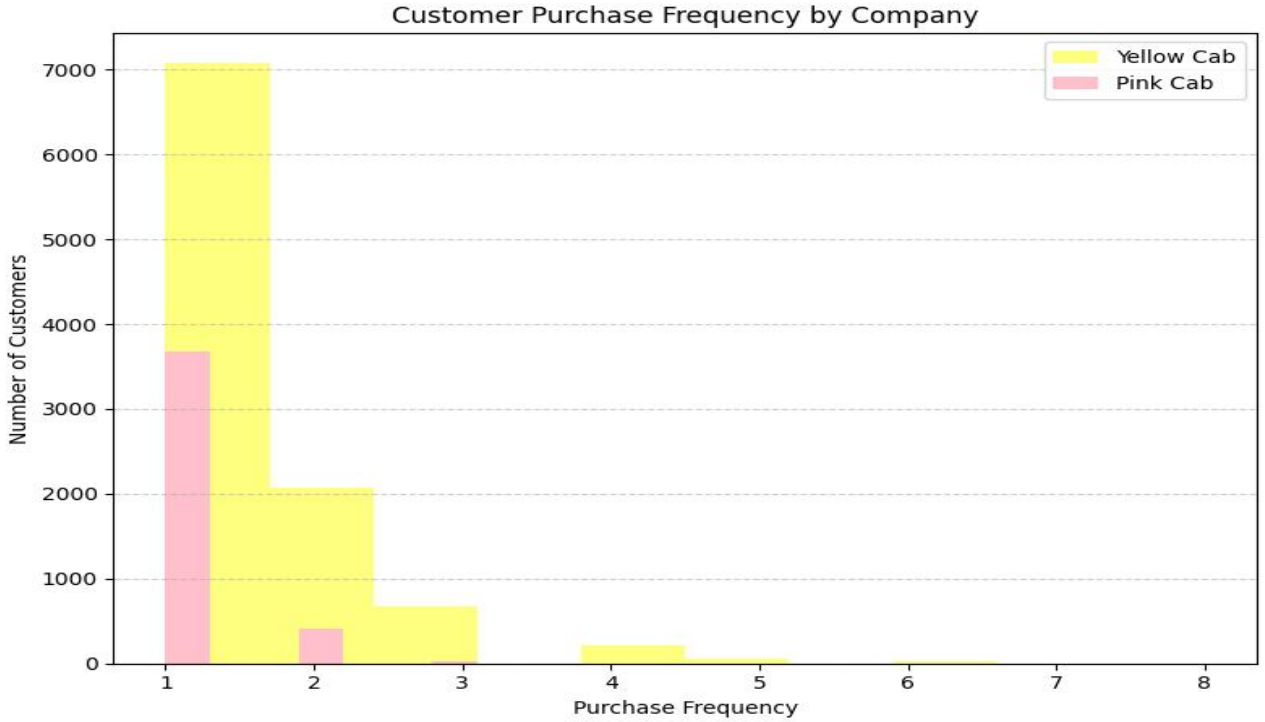
Hypothesis # 3 rejected.

Data Glacier
Your Deep Learning Partner

Hypotheses Testing:

4. “Pink Cab” has higher purchase frequency





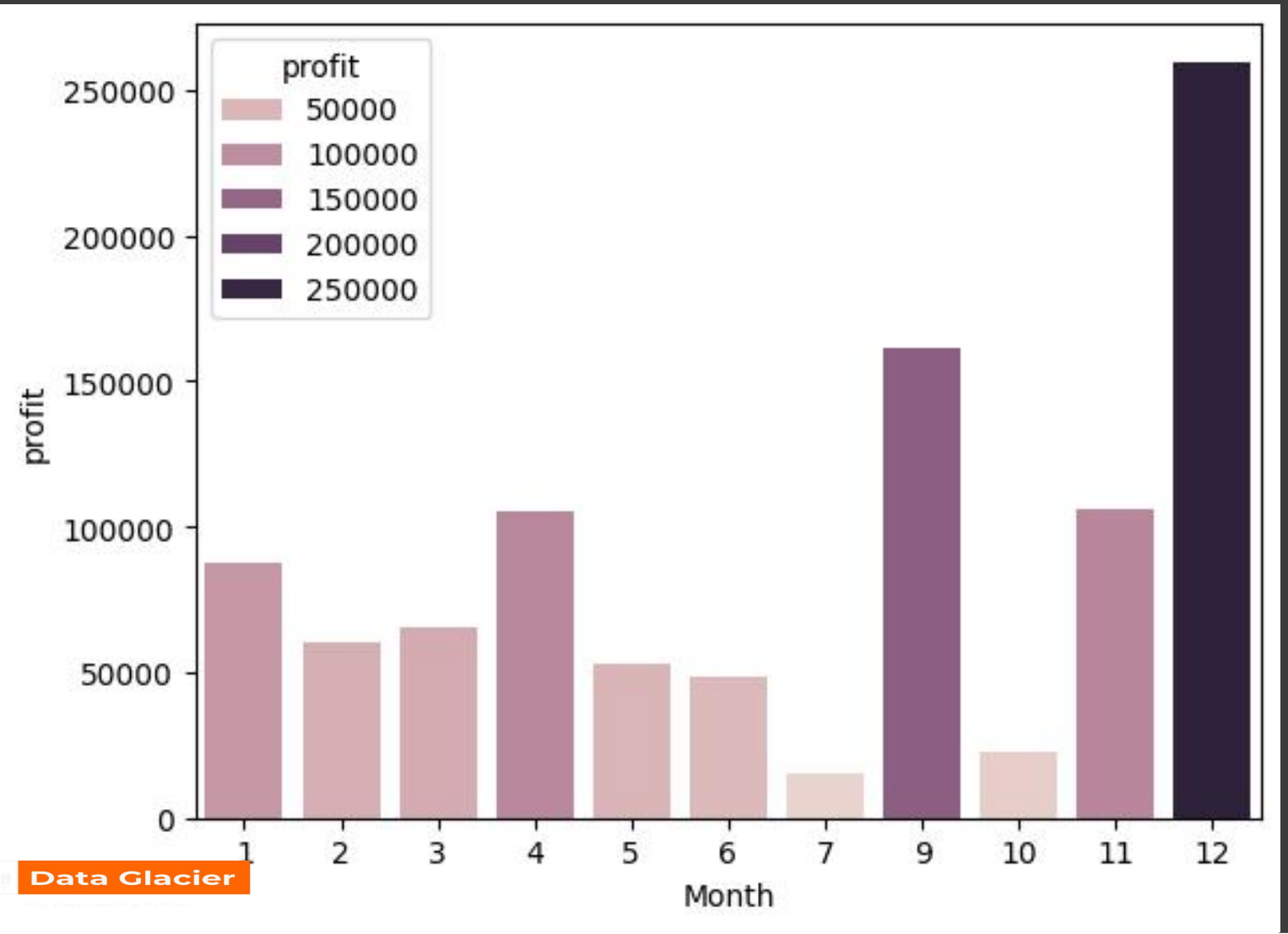
The first graph shows over all CPF and The second Graph clearly shows that “Yellow Cab” have much higher CPF as compared to “Pink Cab”.

Hypothesis # 4 rejected.

Data Glacier
Your Deep Learning Partner

Hypotheses Testing:

5. There’s seasonality in profit.

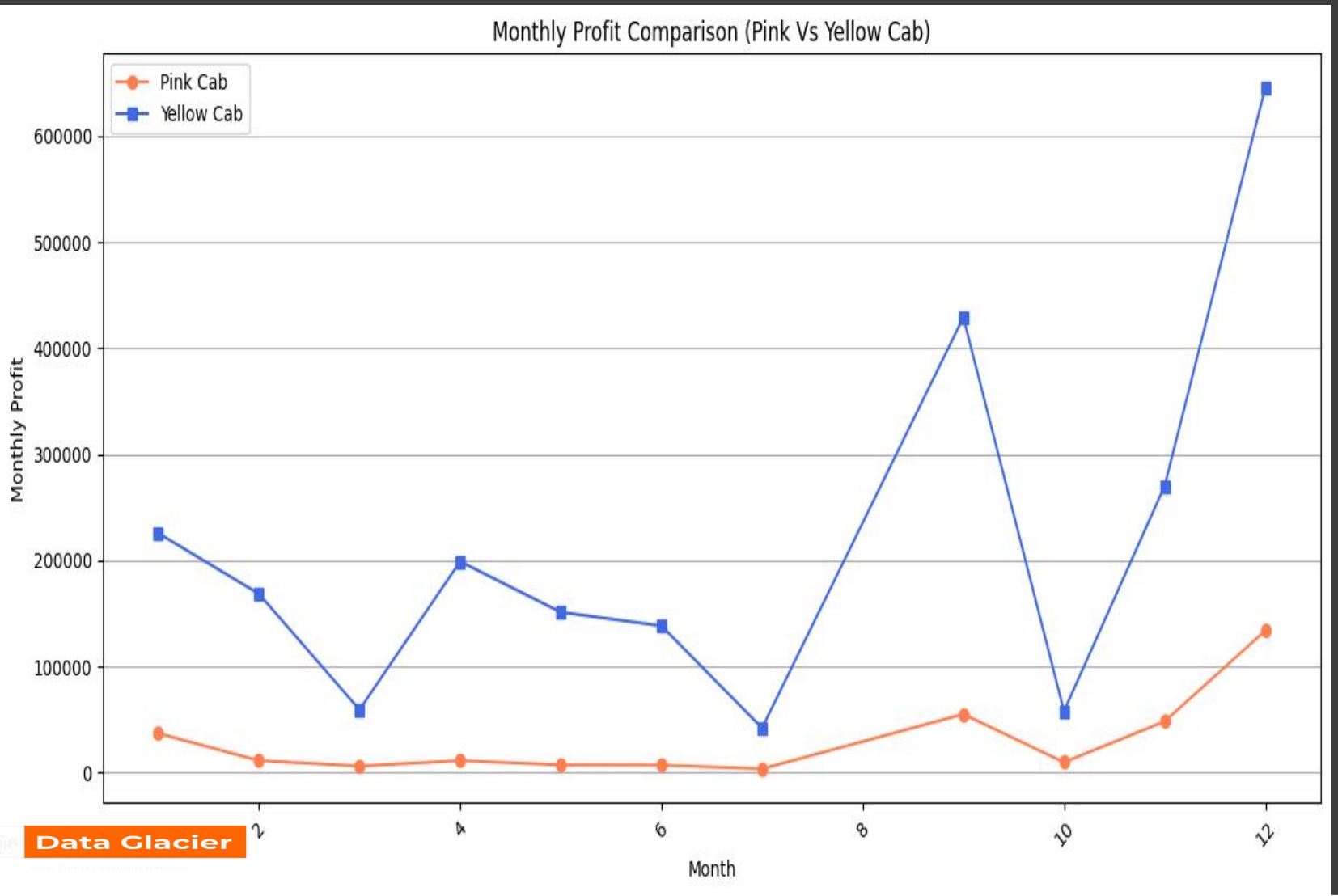


Yes,It is evident that there is seasonality in profit, especially by the end of year.

Hypothesis # 5 accepted.

Hypotheses Testing:

6. “Yellow Cab” has the most profits in peak-seasons



The “Yellow Cab” does have comparatively a higher profit in the peak-seasons.

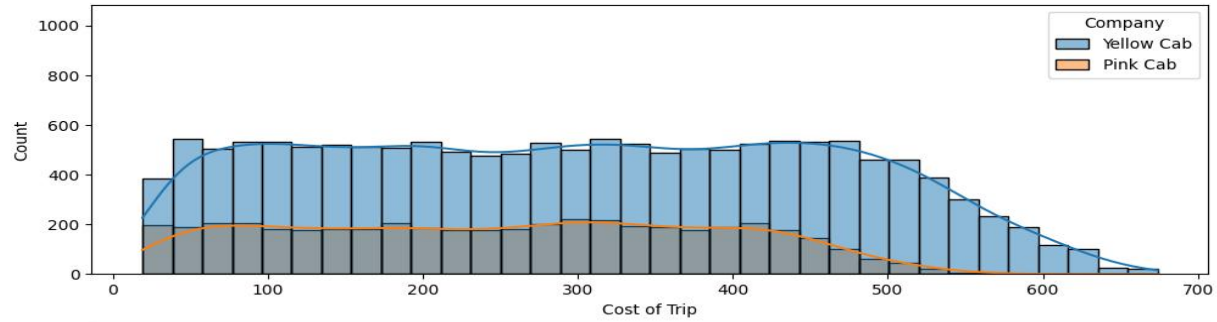
Hypothesis # 6 accepted.

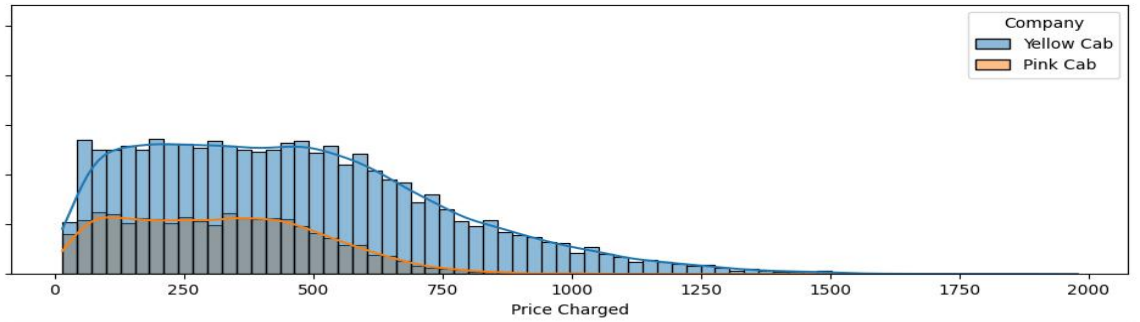
Hypotheses Testing:

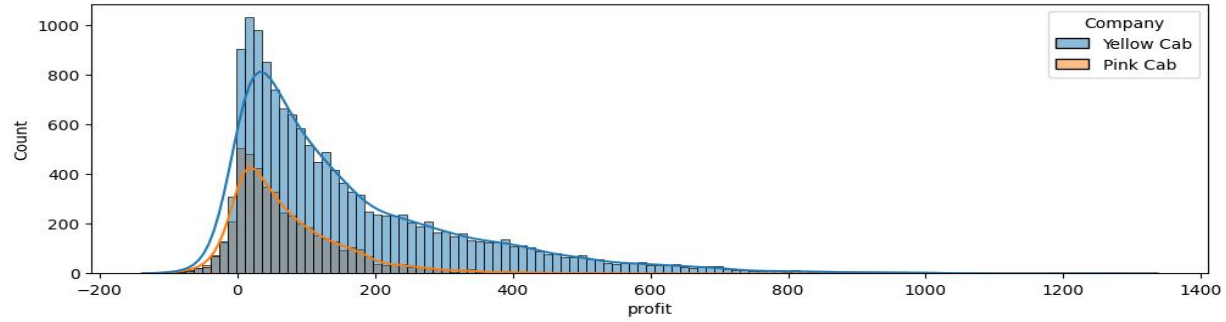
7. “Yellow Cab” is leading in every aspect

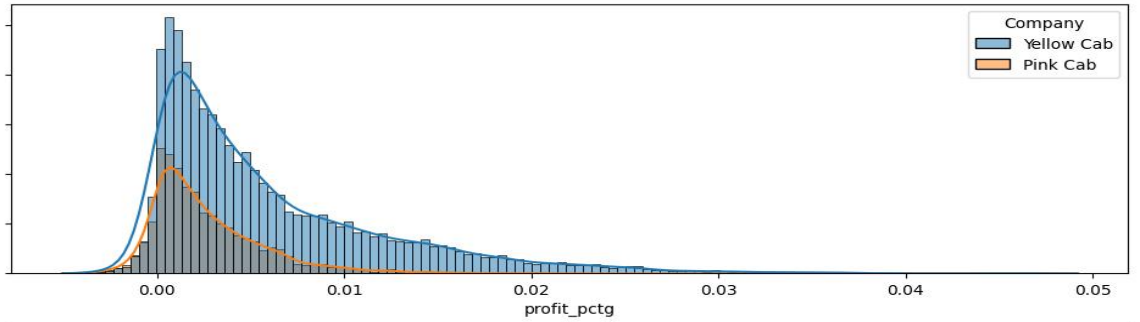
Data Glacier
Your Deep Learning Partner

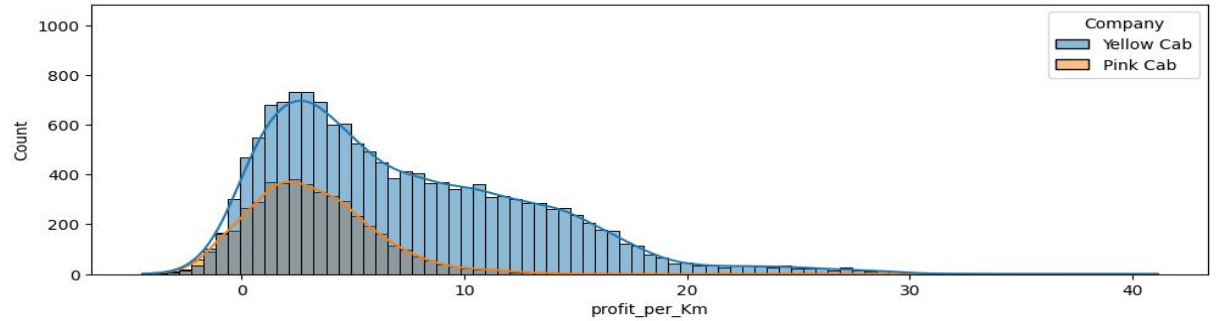
Distributions of Target Variables

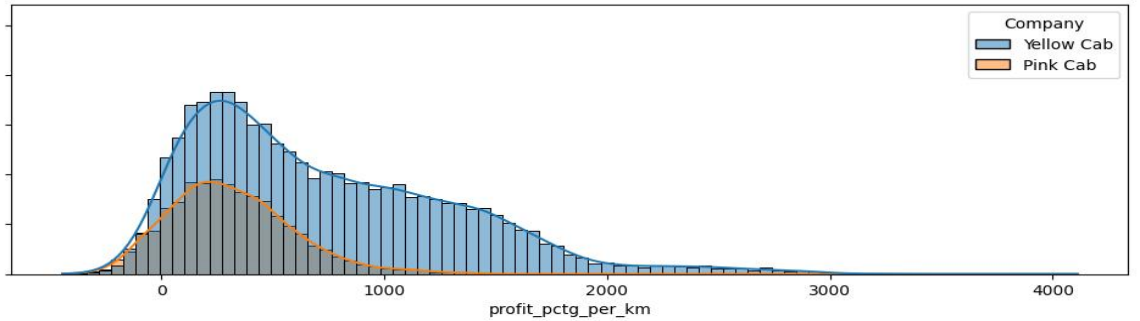






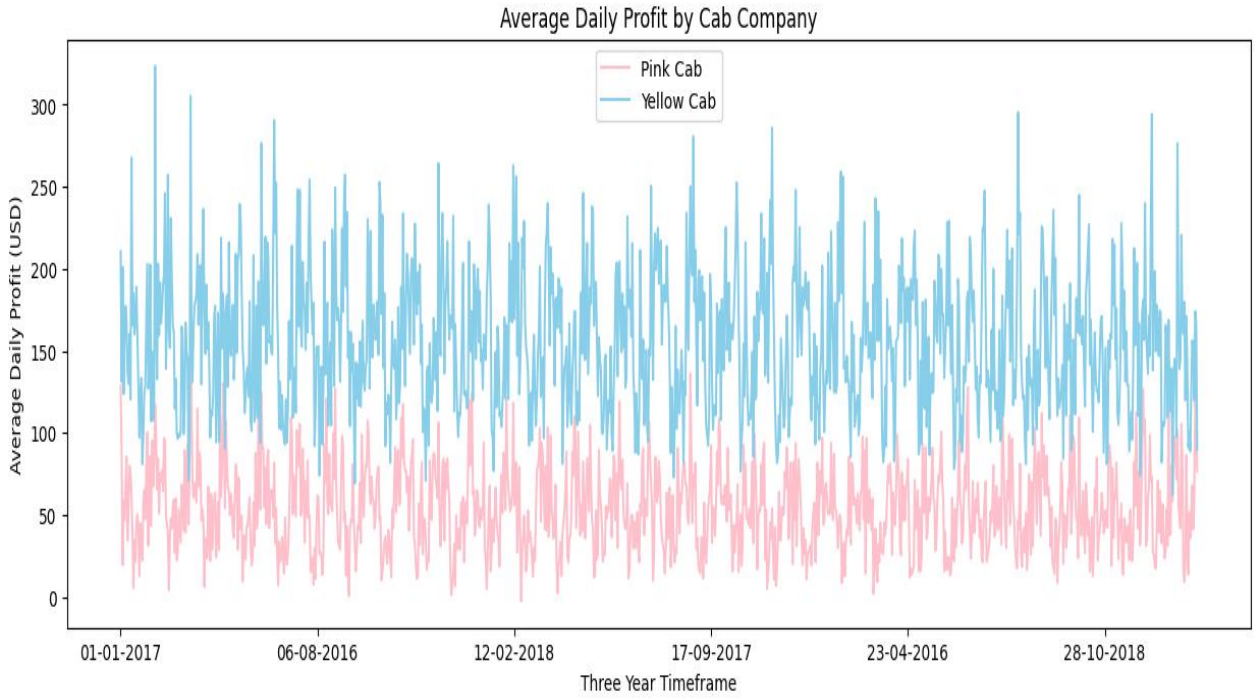


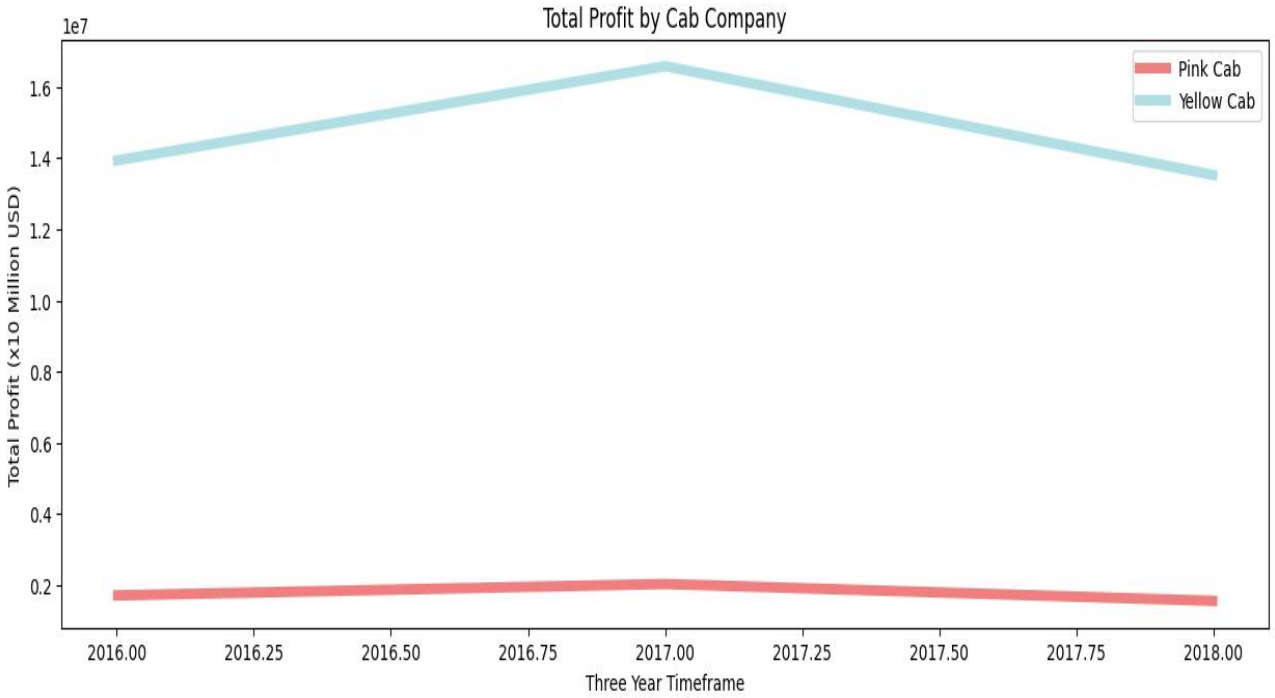




Hypotheses Testing:

7. “Yellow Cab” is leading in every aspect





As evident from the graphs. “Yellow Cab” does seem to lead in every aspect.

Hypothesis # 7 accepted.

Data Glacier
Your Deep Learning Partner

Recommendations

Based on the analysis, Here are a few recommendations:

- **Market Leader:** While facing profitability challenges, Yellow Cab currently boasts a higher customer frequency rate and lower churn rate compared to Pink Cab. This indicates a strong existing customer base.
- **NYC as a Starting Point:** If a single-city investment is chosen, NYC appears most lucrative due to its high overall profit rate.
- **High-Density User Cities:** For user base expansion, consider Boston and San Francisco, which show the highest user density.
- **Holiday Promotions:** Target peak profit seasons, particularly year-end holidays, with special promotions to capitalize on increased ridership.
- **Cost vs. Price Disparity:** A significant gap between Yellow Cab's "Cost of Trip" and "Price Charged" seems to be impacting profitability. The company should prioritize strategies to address this cost-price gap.

Bisma Azeem

Thank You