Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Insight for Cab Investment Firm

Actionable insights for potential investment in the cab industry

Data Science Virtual Internship - LISUM30
Submitted by: Konstantinos Soufleros
Date: 18/02/2024

Contents

Data Glacier
Your Deep Learning Partner

Introduction

1

OVERVIEW OF THE CAB INDUSTRY IN THE US

The cab industry in the United States is a large and growing market, with an estimated value of over $100 billion.

The industry is dominated by a few large players, such as Yellow Cab and Pink Cab.

2

XYZ's INVESTMENT STRATEGY

XYZ is a private equity firm that invests in a variety of industries, including the transportation sector.

XYZ is looking to invest in a cab company that has a strong market position and the potential for growth.

3

PURPOSE OF THIS ANALYSIS

The purpose of this analysis is to provide XYZ with insights into the cab industry in the United States and to identify potential investment opportunities.

This analysis will compare the performance of two cab companies, Yellow Cab and Pink Cab, and will assess their market position and growth potential.

Data sources

Cab_Data.csv: This dataset contains transaction details for two cab companies, Yellow Cab and Pink Cab. The data includes information such as the transaction ID, date of travel, company, city, kilometers traveled, price charged, and cost of trip.

City.csv: This dataset contains information on various US cities.The data includes information such as the city name, population, and number of cab users.

Transaction_ID.csv: This dataset links transaction IDs to customer IDs and payment modes.This data allows us to integrate the transaction data with the customer demographic data.

Customer_ID.csv: This dataset contains demographic details of cab users.The data includes information such as the customer ID, gender, age, and income.

Additional data sources from Kaggle:

US Holiday Dates (2004-2021).csv: This dataset contains information on official US holidays from 2004 to 2021.

WeatherEvents_Jan2016-Dec2022.csv: This dataset contains information on weather events in the United States from January 2016 to December 2022.

These two additional datasets were used to investigate an extreme spike in cab usage that was observed on January 7, 2018. The US Holiday Dates dataset was used to determine if there was a holiday on that day, and the WeatherEvents dataset was used to determine if there was any unusual weather activity that could have caused the spike in cab usage.

Methodology of the analysis

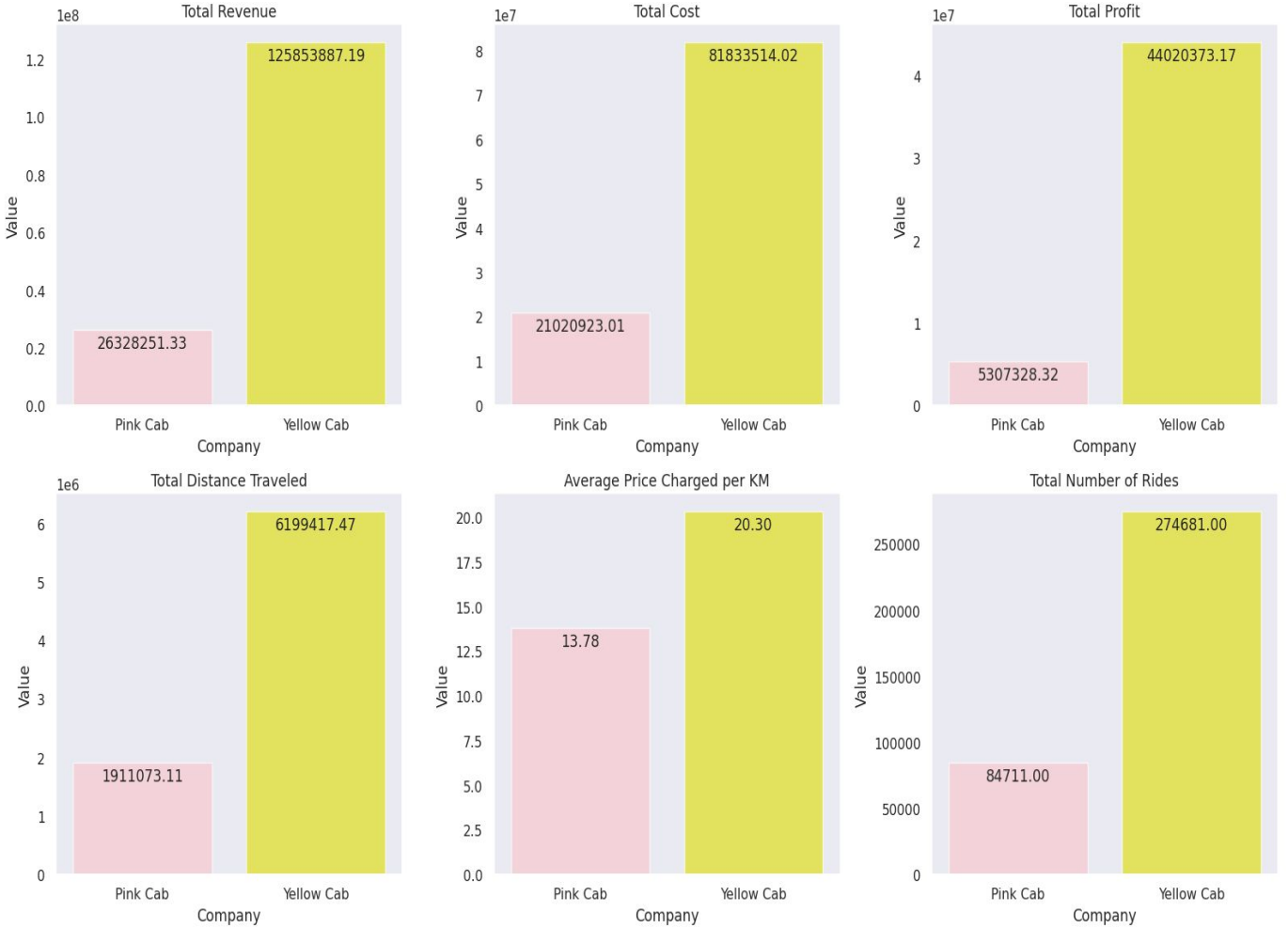
Our analysis approach involved a multi-step process:

- ✓ Data Understanding and Preparation: We thoroughly reviewed each dataset, understanding its structure, field names, and data types. Data cleaning and preprocessing steps were undertaken to handle missing values, outliers, duplicates, and ensure consistency across datasets.
- ✓ Exploratory Data Analysis (EDA): Each dataset was explored separately, employing descriptive statistics and visualizations to uncover patterns, trends, and relationships. Hypotheses were formulated later, based on the insights gained from EDA.
- ✓ Data Integration and Further Analysis: Datasets were merged as necessary to create a comprehensive dataset for further analysis. Additional analysis, such as customer segmentation and external data integration, were conducted to enhance insights.
- ✓ Hypothesis Testing: Formulated hypotheses were tested using appropriate statistical tests, allowing us to assess the validity of assumptions and draw actionable conclusions.
- ✓ Presentation Preparation: Findings and recommendations were organized into a coherent narrative for presentation to XYZ's executive team. Visual aids and documentation were provided to facilitate understanding and decision-making.

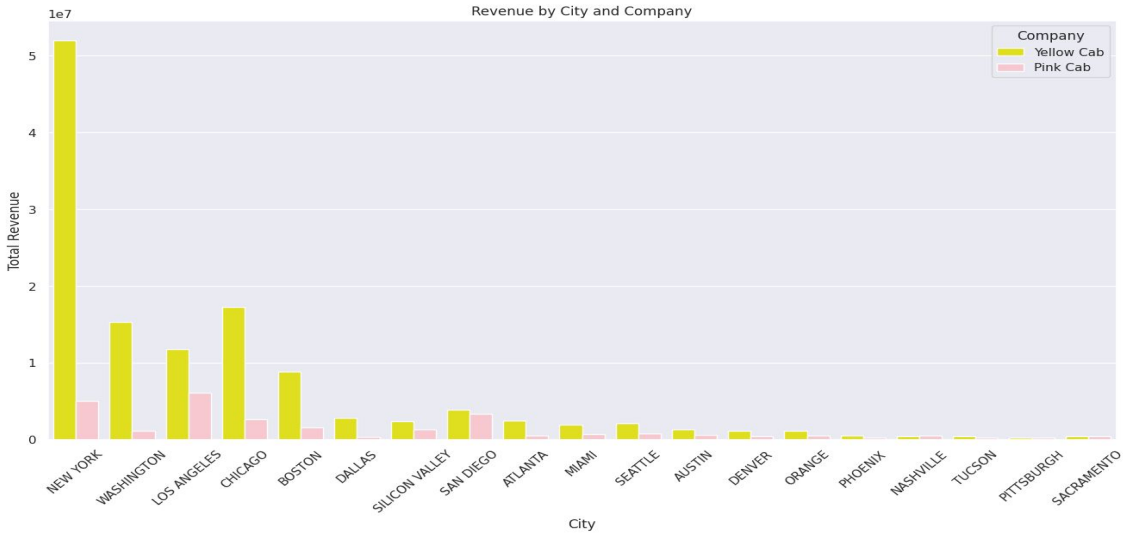
Key Findings - Descriptive Statistics

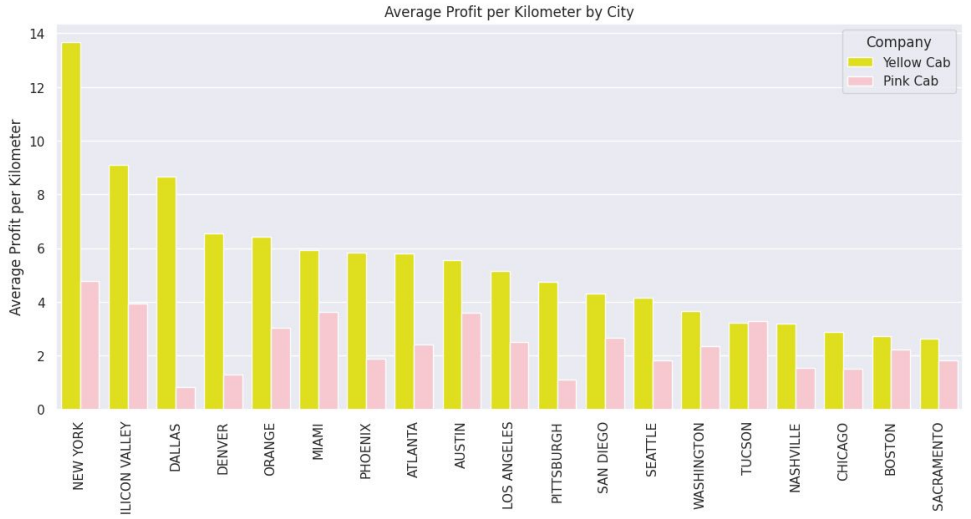
Metric	Yellow Cab	Pink Cab
Average KM traveled	22.57 km	22.56 km
Average price charged	$458.18	$310.80
Average cost of trip	$297.92	$248.15
Profit for Average KM traveled	$160.26	$62.65

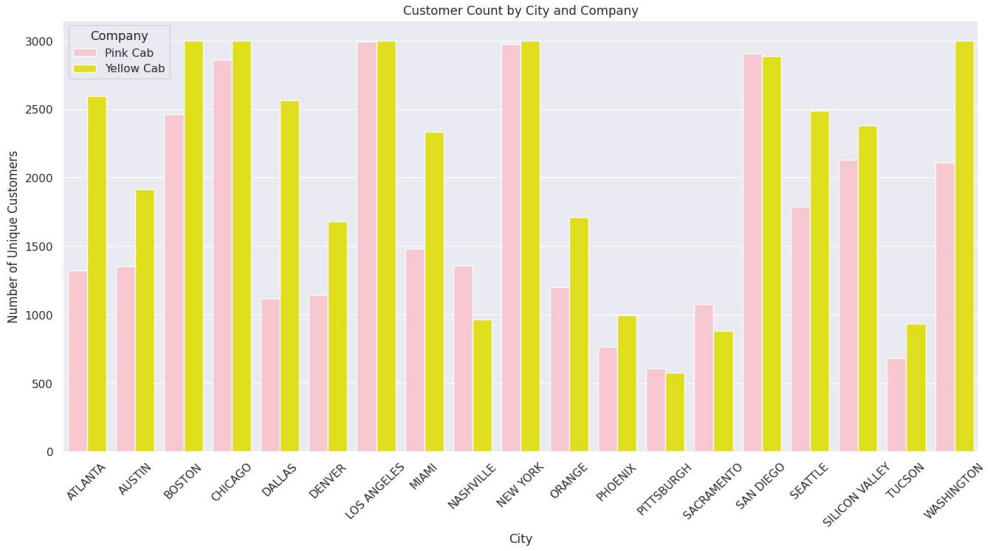
From the table above, we can see clearly that for the same average distance of 22.56 km, Yellow Cab charges 458.18 dollars with a cost of trip of 297.92 dollars, leaving a profit of 160.26 dollars. On the other hand, Pink Cab charges 310.80 dollars for a cost of trip of 248.15 dollars, leaving a profit for the company of only 62.65 dollars, that is almost 2.56 times less than Yellow Cab.



Key Findings - Market Share Analysis



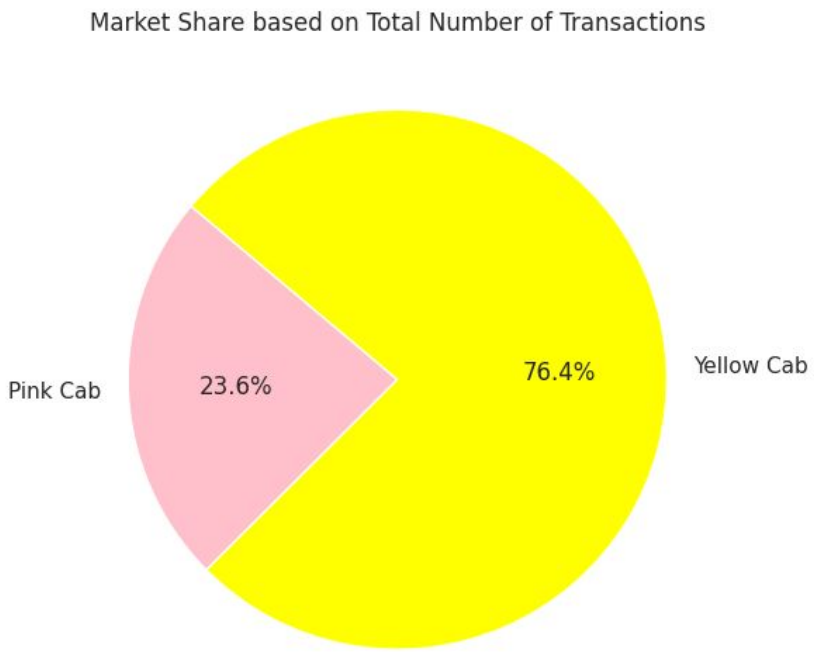


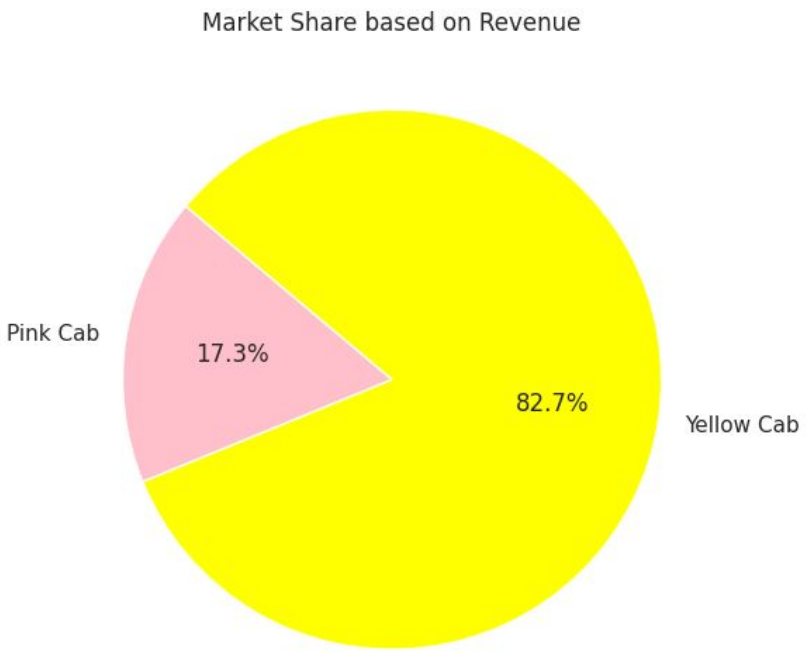


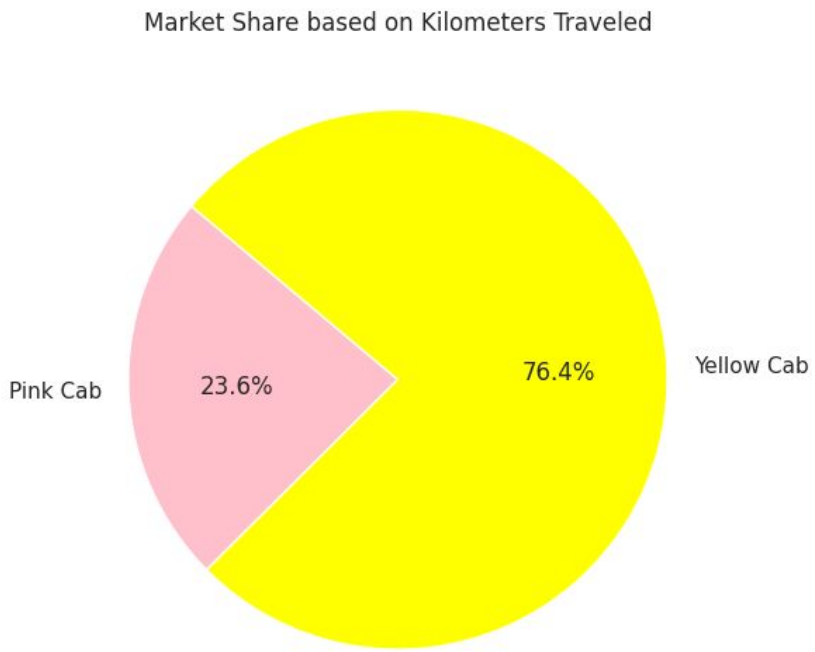
Market share based on Revenue, Average Profit/km and Customer Count by City

Yellow Cab has a dominant market position in most cities compared to Pink Cab. However, Pink Cab has a stronger market position in a few key cities, like Tuckson where it has better average profit per km, San Diego, Sacramento and Nashville, where it has more customers than than Yellow Cab.

Key Findings - Market Share Analysis (continued)



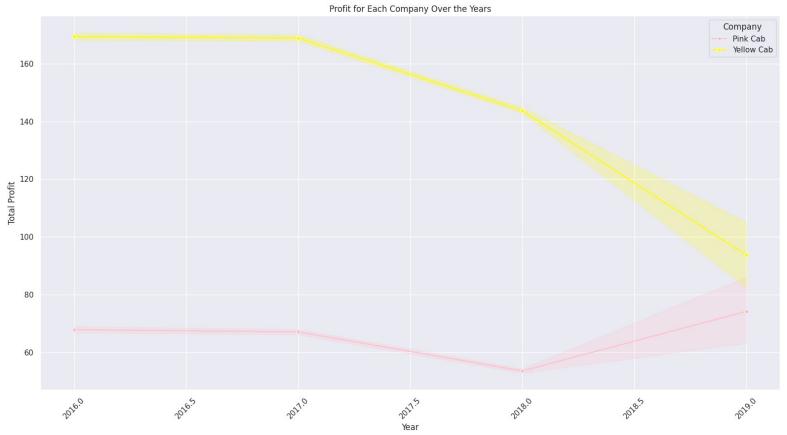


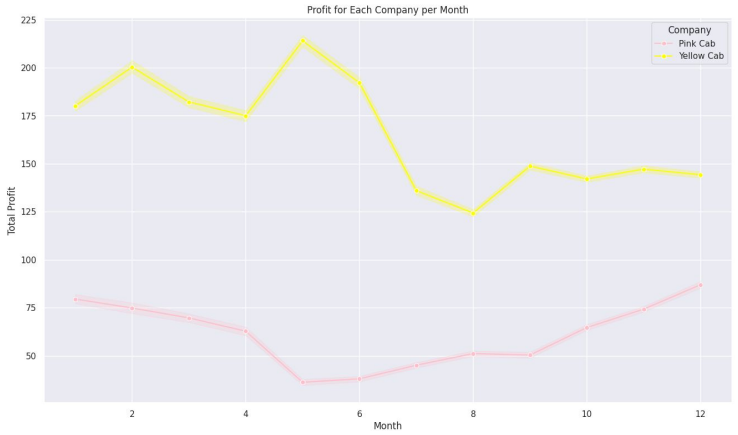


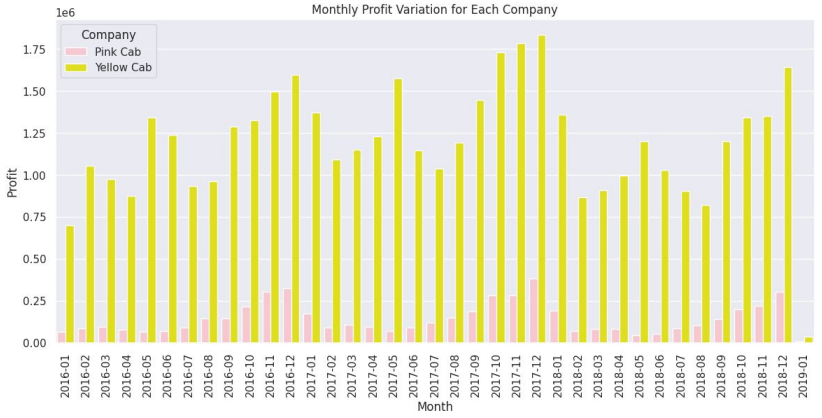
Market share based on total number of transactions, Revenue and Km travelled

Yellow Cab has a significantly higher market share than Pink Cab based on the total number of transactions. This is consistent with the findings from the previous slide, which showed that Yellow Cab has a higher market share in most cities.

Key findings - Profit Analysis

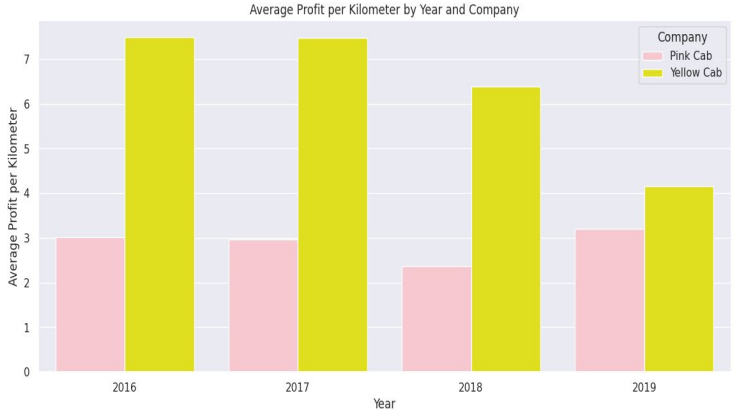




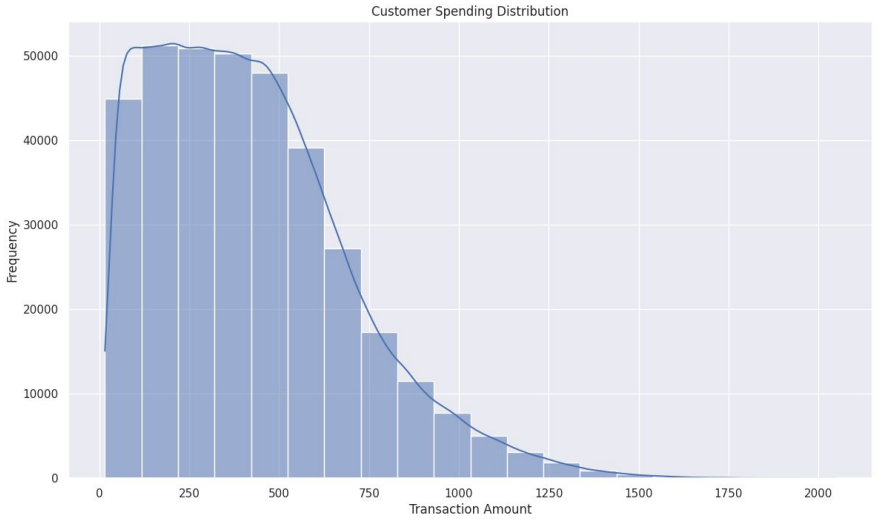


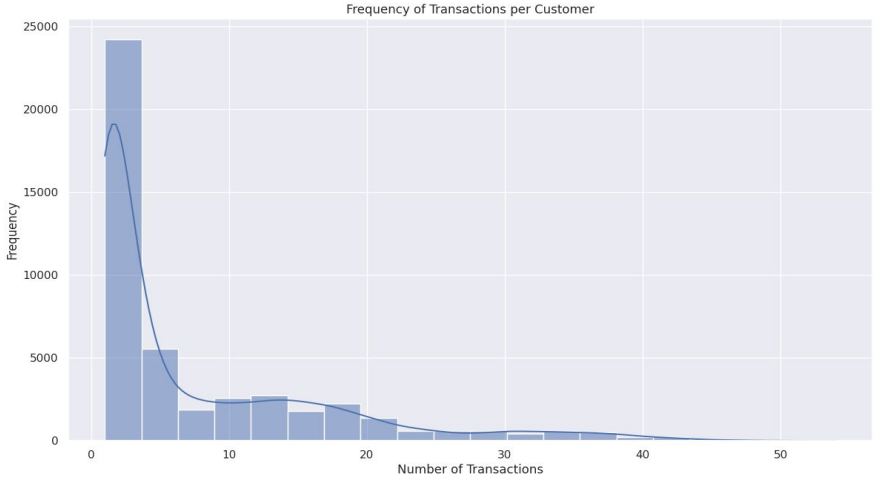
The profit analysis shows that Yellow Cab is more profitable than Pink Cab. This is due to the fact that Yellow Cab charges a higher price per kilometer.

- Yellow Cab initially had better profitability compared to Pink Cab, as indicated by the higher profit levels in the earlier years. However, Yellow Cab's profitability shows a declining trend over time. On the other hand, Pink Cab's profitability seems to be stable and gradually increasing over the years.
- The monthly profitability trends also suggest a similar pattern, with Yellow Cab's profit declining over the months and Pink Cab's profit showing a slight upward trend. Most profitable month for Yellow Cab is May, and for Pink Cab is December.
- The Yellow Cab line appears consistently above the Pink Cab line, indicating that, on average, Yellow Cab tends to generate higher profits per kilometer traveled compared to Pink Cab.



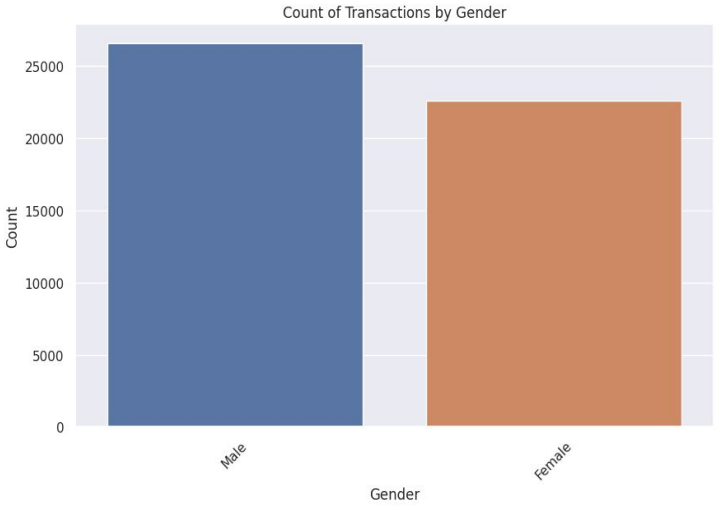
Key findings - Customer Analysis

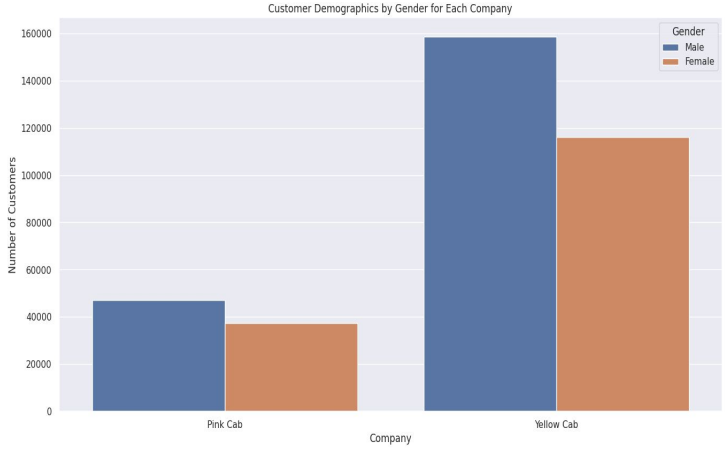


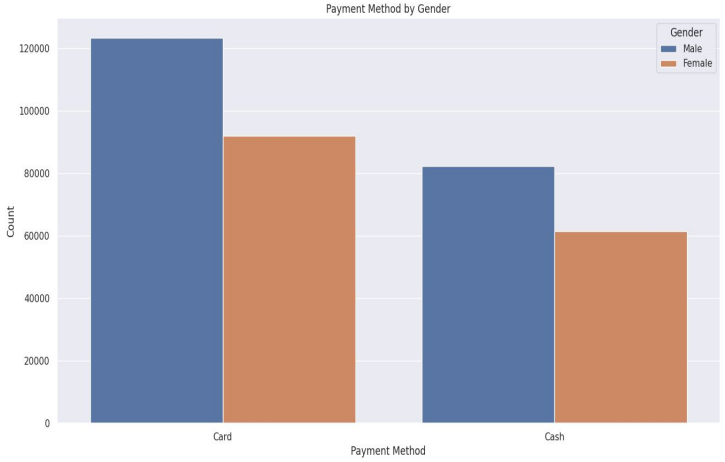


- The number of customers for both companies is higher in the autumn months and lowest in the winter months with December being the best month for both companies.
- The majority of customers for both companies spend between $0 and $500 on taxi rides. However, there is a small percentage of customers who spend over $1,000 on taxi rides for the specific timeframe.
- Most of the customers for both companies have only a few transactions per year. However, there is a small percentage of customers who have over 30 transactions for the specific timeframe.

Key findings - Customer Analysis (continued)

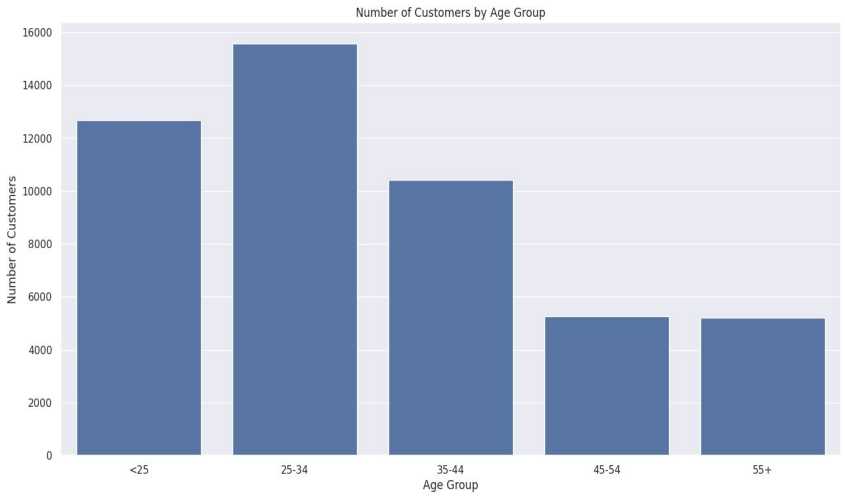


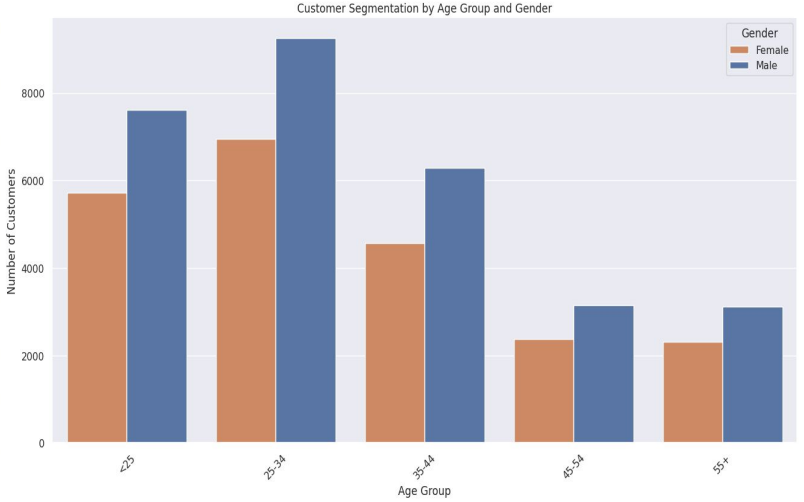


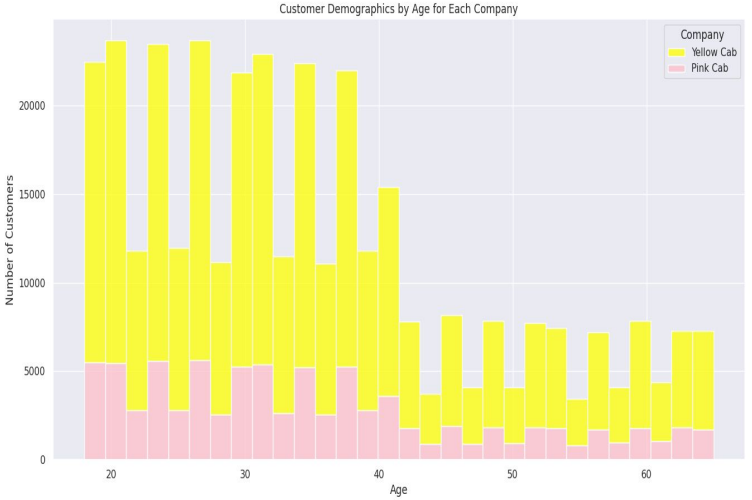


- The majority of transactions for both companies are made by male customers.
- Most of the transactions for both genders are made using card for both companies.

Key findings - Customer Analysis (continued)

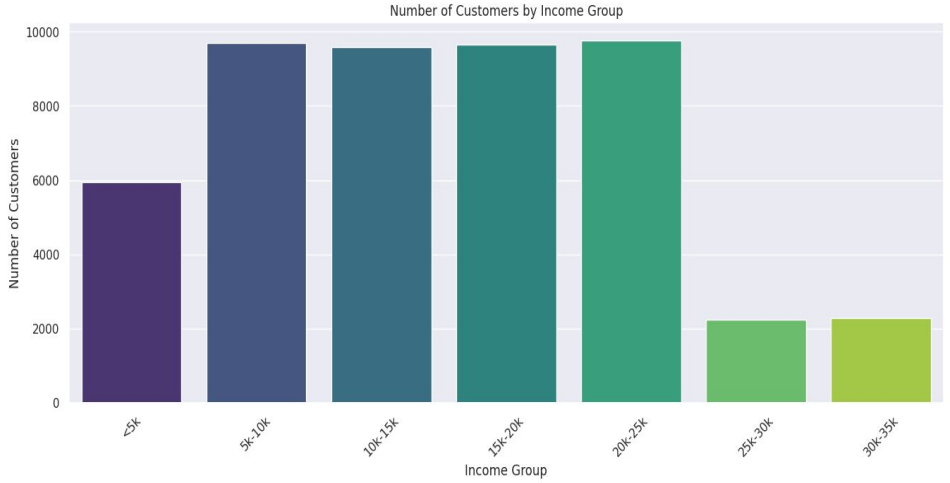


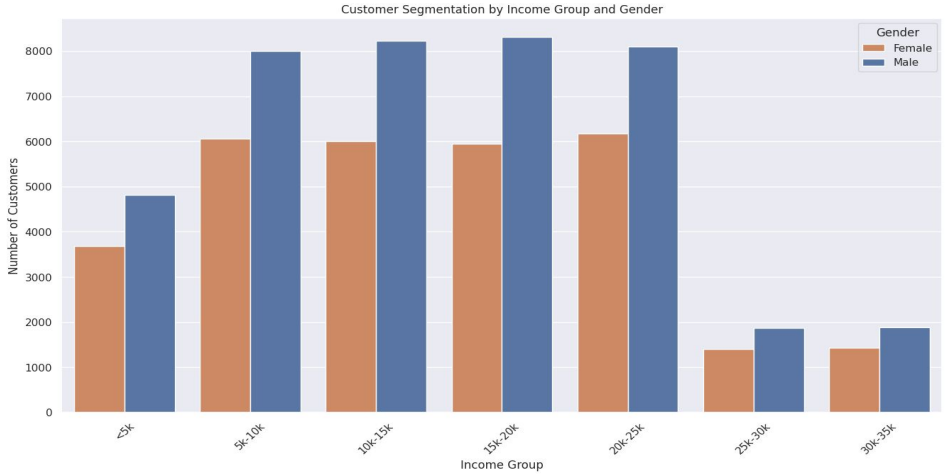


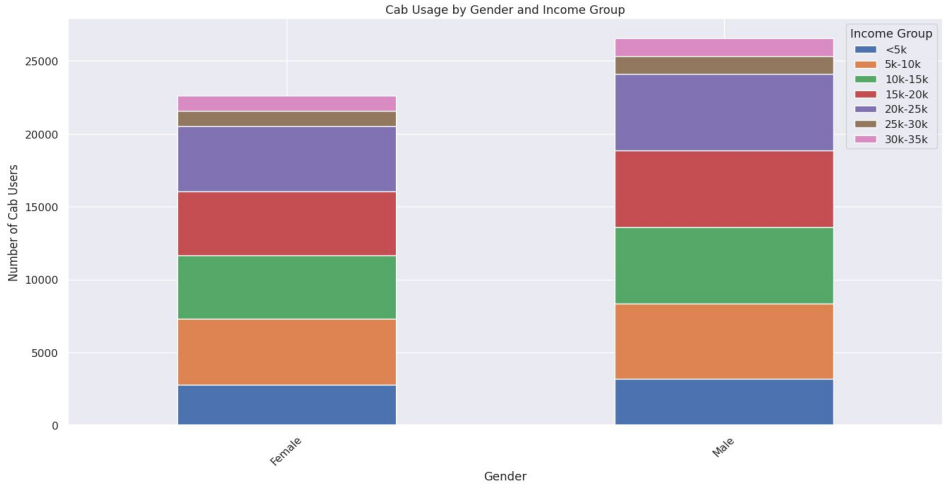


- The largest age group for both companies is the "25-34" age group. The "25-34" age group is followed by the "<25" age group and the "35-44" age group. The "55+" age group has the fewest customers for both companies.
- Additionally, Yellow Cab has a slightly higher percentage of customers in the younger age groups compared to Pink Cab.

Key findings - Customer Analysis (continued)

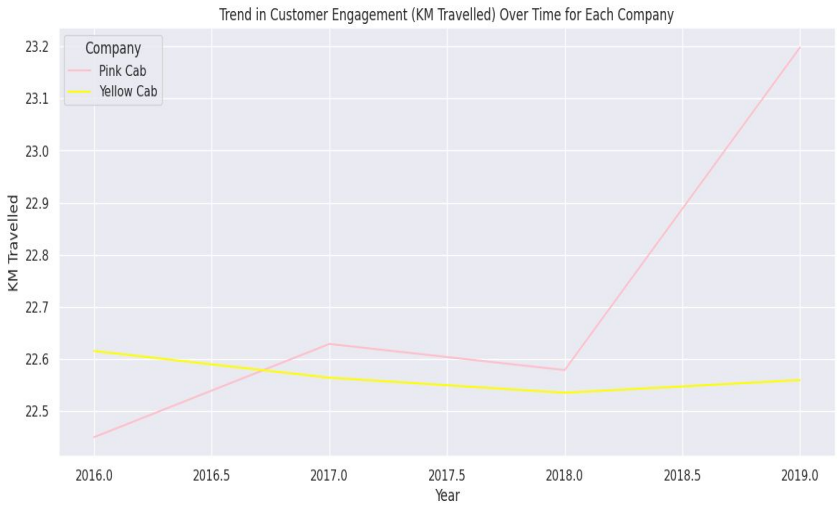


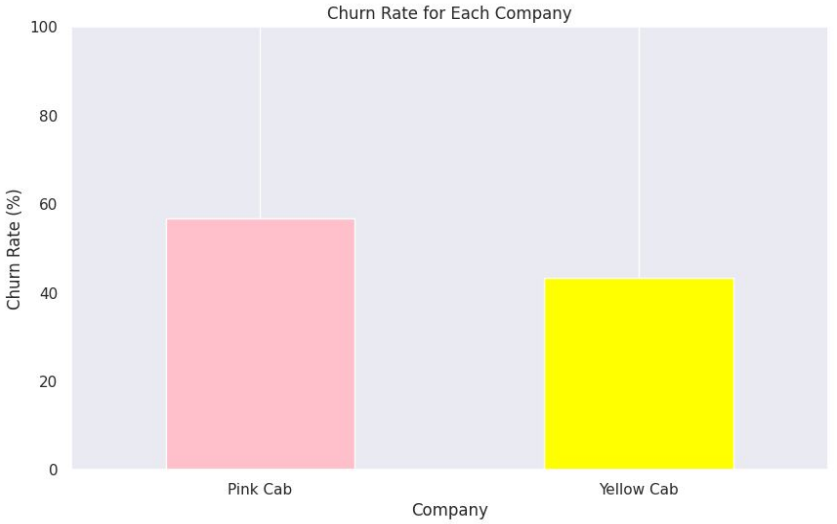


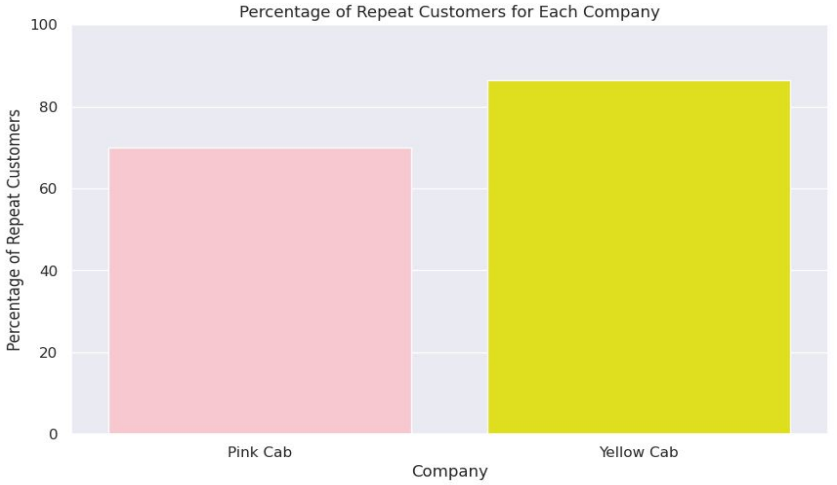


- The majority of customers for both companies fall into lower income brackets (i.e., below $20,000 annually). The highest count of customers is observed in the "$20k-25k" income bracket for both companies.
- There are fewer customers in the higher income brackets, such as "25k-30k" and "30k-35k," indicating that the service may not be as popular among higher-income individuals.

Key findings - Customer Analysis (continued)





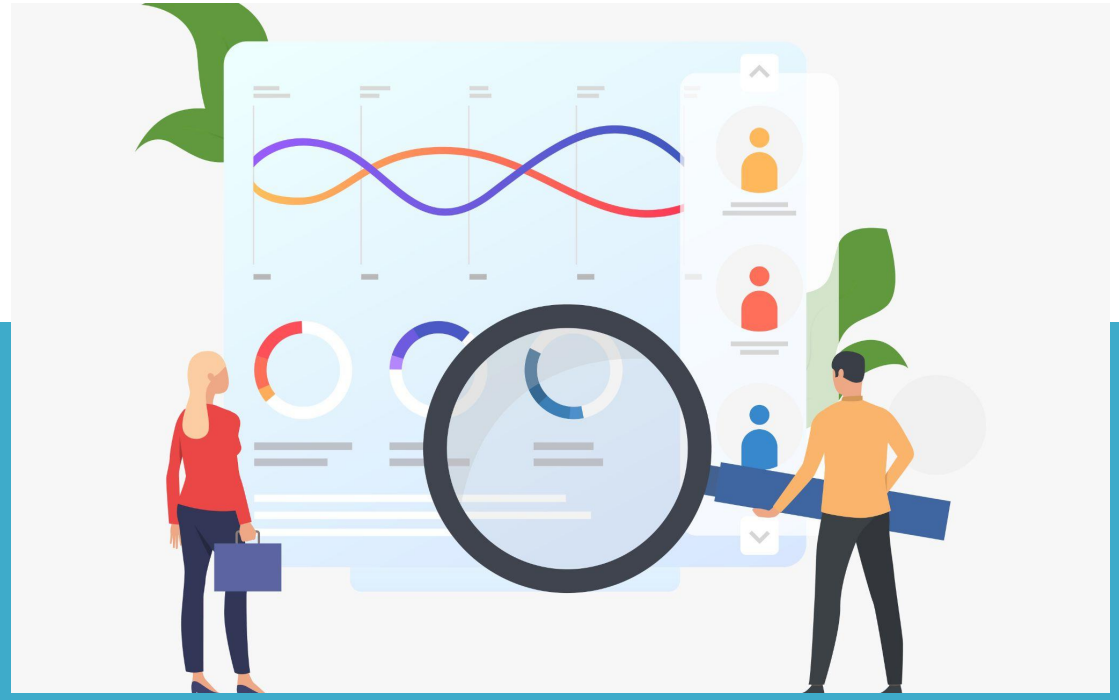


- Pink Cab has a increasing trend in customer engagement over time. However, Pink Cab also has a higher churn rate than Yellow Cab.
- Yellow Cab, on the other hand, has a relatively stable trend in customer engagement over time and a lower churn rate.
- Additionally, Yellow Cab has a higher percentage of repeat customers than Pink Cab.

Hypothesis 1

Seasonality Hypothesis

- **Null Hypothesis (H0): There is no seasonality in the number of customers using cab services.**
- **Alternative Hypothesis (H1): The number of customers using cab services exhibits seasonality, with peak demand during certain months or seasons.**



1

Seasonal_Mann_Kendall_Test(trend='increasing', h=True, p=7.923130374010157e-09, z=5.770085540729481, Tau=0.9090909090909091, s=120.0, var_s=425.3333333333333, slope=1.0, intercept=9196.25)

2

The cab service usage shows a significant increasing trend with strong evidence of seasonality (p-value=7.92e-09). The z-score of 5.77 confirms the trend's significance. Tau of 0.909 indicates a strong positive association between time and customer numbers.

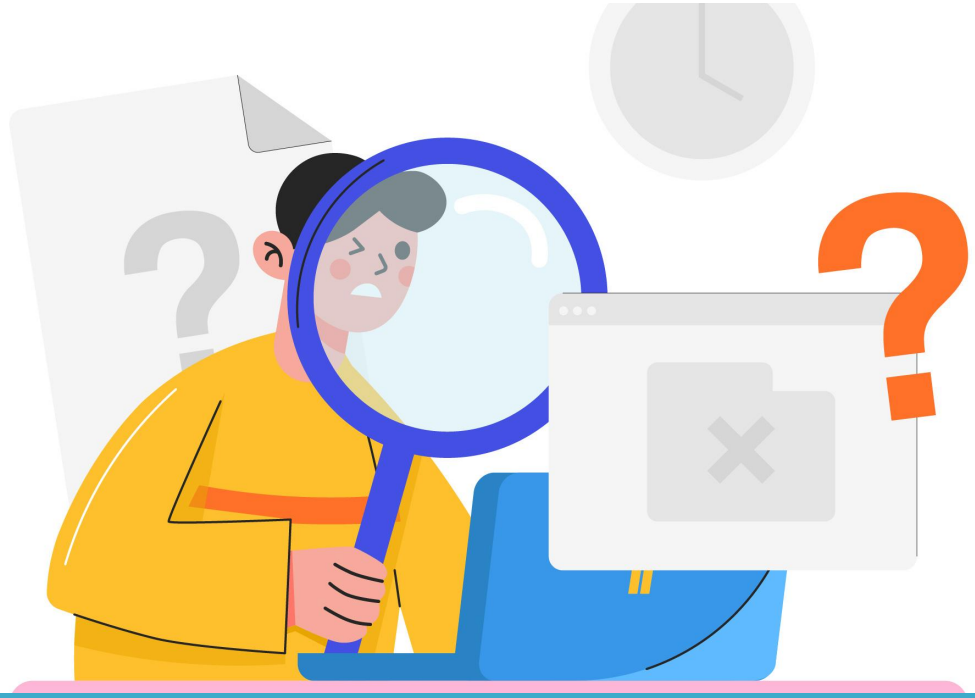
3

The test results provide strong evidence to support **the alternative hypothesis** that the number of customers using cab services exhibits seasonality, with a statistically significant increasing trend over the months.

Hypothesis 2

Company Performance Hypothesis

- **Null Hypothesis (H0): There is no significant difference in the number of cab users between Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): One of the cab companies has a significantly higher number of cab users compared to the other.**



1

Two-Sample T-Test Results:

T-Statistic: -12.187627722674286

P-Value: 3.6747050468286167e-34

2

Reject the null hypothesis. There is a significant difference in the number of customers between Pink Cab and Yellow Cab.

3

We can conclude that there is indeed a significant difference in the number of customers between Pink Cab and Yellow Cab.

Hypothesis 3

Payment Mode Preference Hypothesis

- **Null Hypothesis (H0): There is no association between the payment mode used (e.g., cash, card) and the frequency of cab usage for Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): Customers using specific payment modes (e.g., card) tend to use Pink Cab or Yellow Cab services more frequently compared to those using other payment modes.**

1 Chi2-Statistic: 0.3733235887859897

P-Value: 0.5411981778304723

2 Fail to reject the null hypothesis. There is no significant association between payment mode and cab company

3 The choice of payment mode does not appear to be related to the preference for a particular cab company (Pink Cab or Yellow Cab).

Hypothesis 4

City Population Influence Hypothesis

- **Null Hypothesis (H0): There is no relationship between the population of a city and the number of cab users in that city for Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): Cities with larger populations have a higher number of Pink Cab or Yellow Cab users compared to cities with smaller populations.**

1 For Pink Cab:

Correlation Coefficient: 0.8857457277364514, P-Value: 0.0. Reject the null hypothesis. There is a significant relationship between city population and number of Pink Cab users.

2 For Yellow Cab:

Correlation Coefficient: 0.9237028641158833, P-Value: 0.0. Reject the null hypothesis. There is a significant relationship between city population and number of Yellow Cab users.

3 Cities with larger populations tend to have a higher number of cab users for both Pink Cab and Yellow Cab services.

Hypothesis 5

Income Influence Hypothesis

- **Null Hypothesis (H0): There is no relationship between customers' income levels and their frequency of cab usage for Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): Customers with higher incomes are more likely to use Pink Cab or Yellow Cab services frequently compared to those with lower incomes.**

1

For Pink Cab:

Correlation Coefficient: Income (USD/Month) 0.003725, dtype: float64, P-Value: 0.5029910596599918. Fail to reject the null hypothesis. There is no significant relationship between customers' income levels and their frequency of cab usage for Pink Cab.

2

For Yellow Cab:

Correlation Coefficient: Income (USD/Month) 0.006502, dtype: float64, P-Value: 0.19406771887211996. Fail to reject the null hypothesis. There is no significant relationship between customers' income levels and their frequency of cab usage for Yellow Cab.

3

There is no significant relationship between customers' income levels and their frequency of cab usage for Pink Cab and Yellow Cab.

Hypothesis 6

Profitability between Age Groups Hypothesis

- **Null Hypothesis (HO): There is no significant difference in profitability between age groups for Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): There is a significant difference in profitability between age groups for Pink Cab and Yellow Cab.**

1 ANOVA Test Result:

F-statistic: 9304.004822497693

p-value: 1.490028414300129e-13

2 Reject the null hypothesis. There is a significant difference in profitability between age groups.

3 There is a significant difference in profitability between age groups. This suggests that age has a notable impact on profitability within the studied context for Pink Cab and Yellow Cab.

Hypothesis 7

Weekend vs. Weekday Usage Hypothesis

- **Null Hypothesis (H0): There is no difference in cab usage between weekends and weekdays for Pink Cab and Yellow Cab.**
- **Alternative Hypothesis (H1): Cab usage is higher during weekends compared to weekdays due to factors such as leisure activities and reduced availability of public transportation for Pink Cab and Yellow Cab.**

1

Pink Cab:

Weekend Cab Trips: 28798, Weekday Cab Trips: 55913, T-Statistic: 39.777073402672784, P-Value: 0.0. Reject the null hypothesis. There is a significant difference in cab usage between weekends and weekdays.

2

Yellow Cab:

Weekend Cab Trips: 92351, Weekday Cab Trips: 182330, T-Statistic: 71.08295050186481, P-Value: 0.0. Reject the null hypothesis. There is a significant difference in cab usage between weekends and weekdays.

3

There is a significant difference in cab usage between weekends and weekdays. This suggests that both Pink Cab and Yellow Cab experience substantially different levels of demand on weekends compared to weekdays.

Recommendations

Based on the extensive exploratory data analysis (EDA) and hypothesis testing conducted, here are the key recommendations for our client, XYZ:

Company Selection:

Based on the profitability analysis, it's evident that Yellow Cab has consistently higher profitability compared to Pink Cab. Therefore, XYZ may consider Yellow Cab as the preferable option for investment due to its better financial performance.

More specifically:

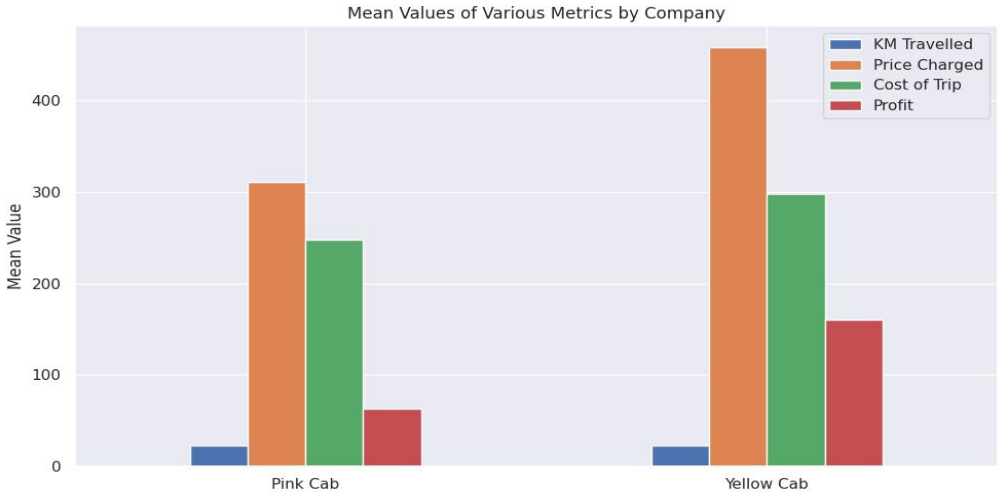
- ✓ Yellow Cab demonstrates a higher number of customers compared to Pink Cab, as indicated by statistical tests such as the Two-Sample T-Test. Profitability trends, including factors such as payment modes, city populations, and seasonal variations, should be carefully evaluated to understand Yellow Cab's financial performance and growth potential. Yellow Cab's market presence and customer base may offer opportunities for market expansion and growth.
- ✓ While Pink Cab may have fewer customers compared to Yellow Cab, there are still significant opportunities for growth and market penetration, especially in segments where it demonstrates a competitive advantage. Market segmentation insights reveal that Pink Cab may appeal to specific demographic segments, which can be leveraged to tailor marketing strategies and enhance customer experience.

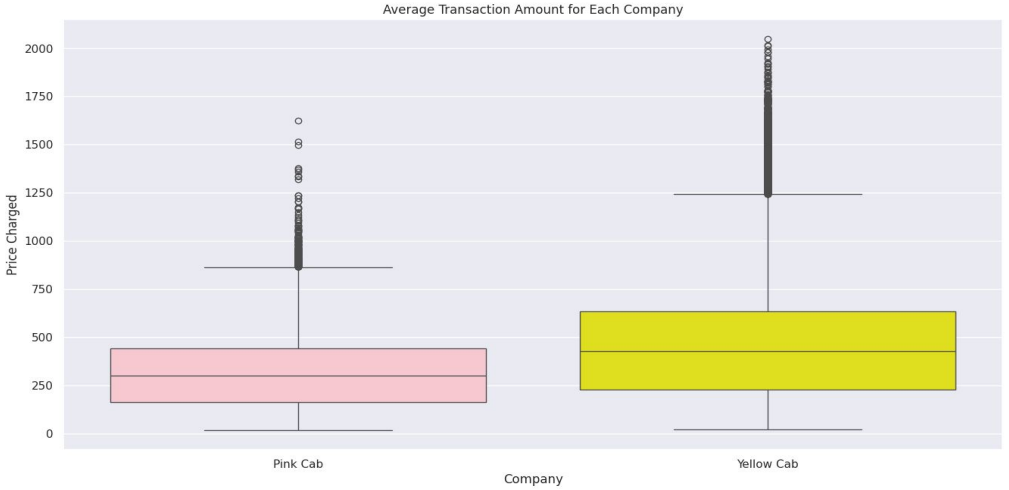
Conclusion

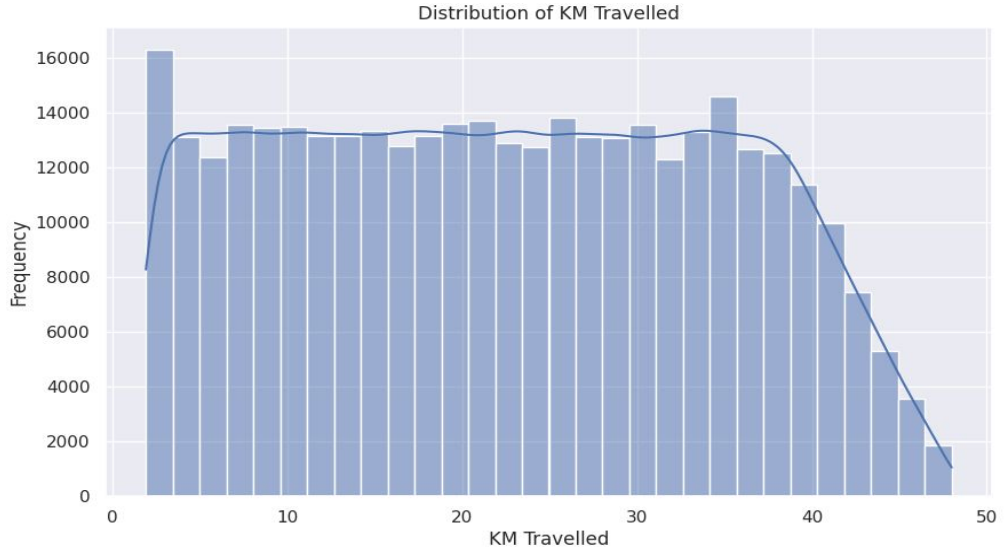
This analysis provides valuable insights into the cab industry in the United States and the performance of Yellow Cab and Pink Cab.

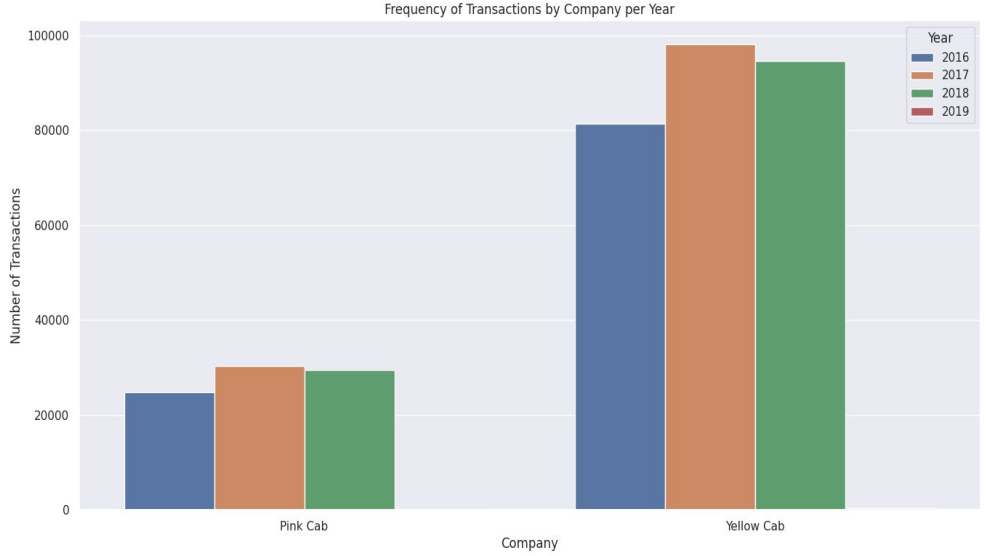
- ✓ Yellow Cab has a higher market share than Pink Cab in most cities.
- ✓ Yellow Cab is more profitable than Pink Cab.
- ✓ The majority of customers for both companies fall into the lower income brackets (i.e., below $20,000 annually).
- ✓ The majority of cab users for both companies are male.
- ✓ Pink Cab has a increasing trend in customer engagement over time.
- ✓ Yellow Cab has a relatively stable trend in customer engagement over time.
- ✓ Pink Cab has a higher churn rate than Yellow Cab.
- ✓ Yellow Cab has a higher percentage of repeat customers than Pink Cab.

APPENDIX 1


Mean Values of Various Metrics by Company
KM Travelled
Price Charged
Cost of Trip
Profit
Mean Value
0
100
200
300
400
Pink Cab
Yellow Cab


Average Transaction Amount for Each Company
Price Charged
0
250
500
750
1000
1250
1500
1750
2000
Pink Cab
Yellow Cab
Company


Distribution of KM Travelled
Frequency
0
2000
4000
6000
8000
10000
12000
14000
16000
0
10
20
30
40
50
KM Travelled


Frequency of Transactions by Company per Year
Year
2016
2017
2018
2019
Number of Transactions
0
20000
40000
60000
80000
100000
Pink Cab
Yellow Cab
Company

APPENDIX 2

January 7, 2018:

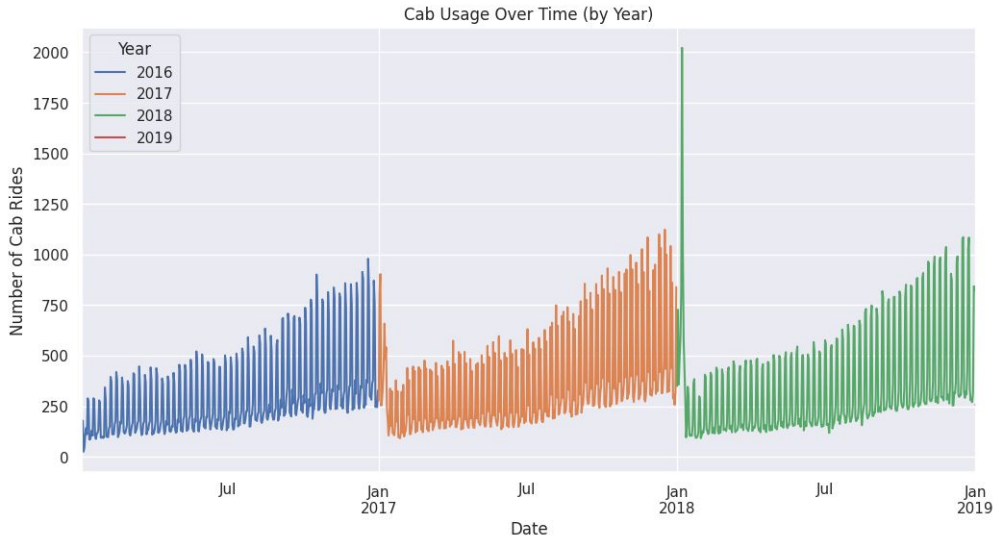
Highest cab usage observed in dataset (2022 rides).

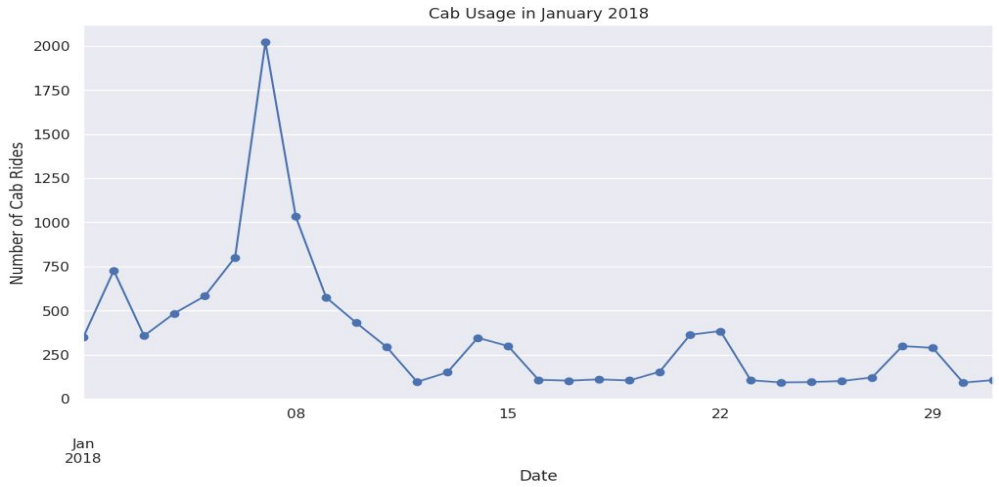
Significant spike in time series plot.

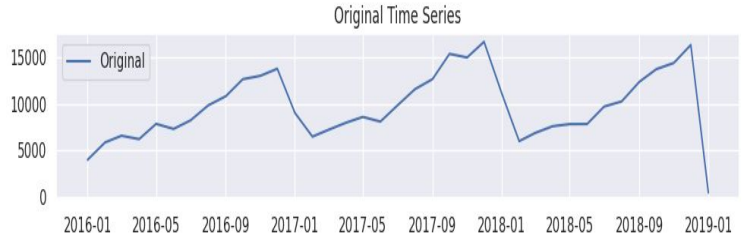
Moderate to severe weather events reported across majority of cities in dataset.

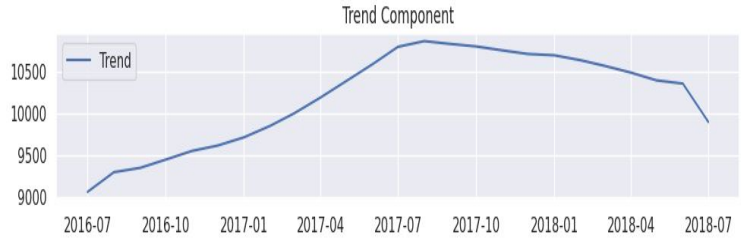
Identified extreme weather phenomenon: Winter Storm Grayson, Blizzard of 2018, or Storm Brody.

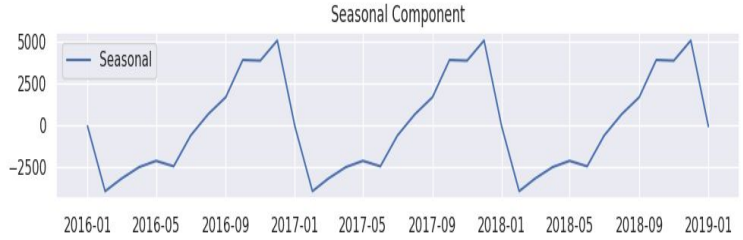
Extreme weather likely contributed to increased demand for cab services.

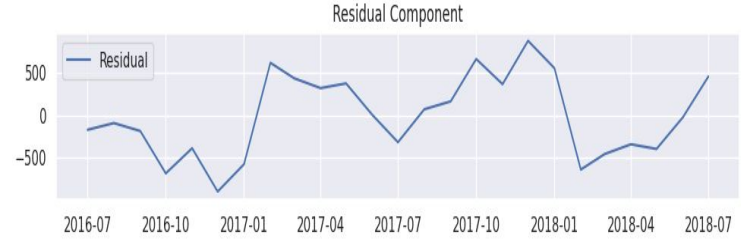




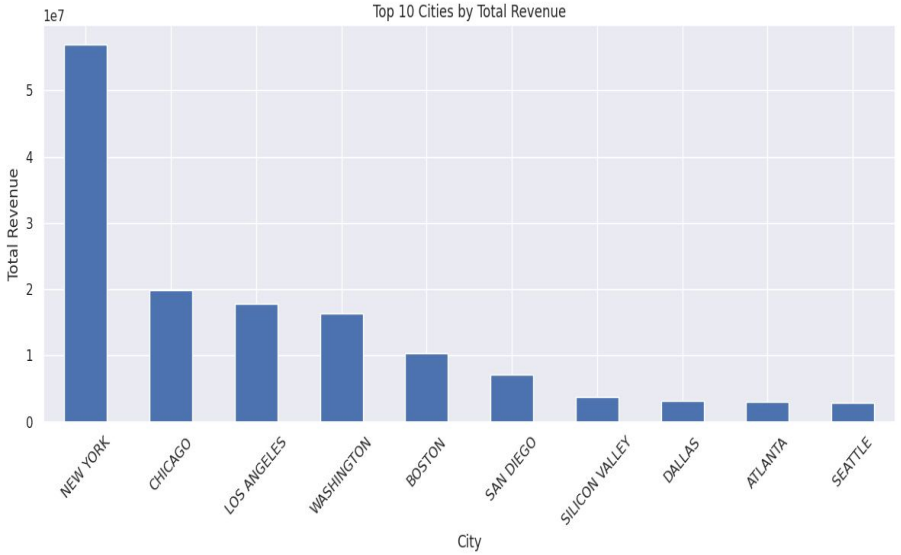


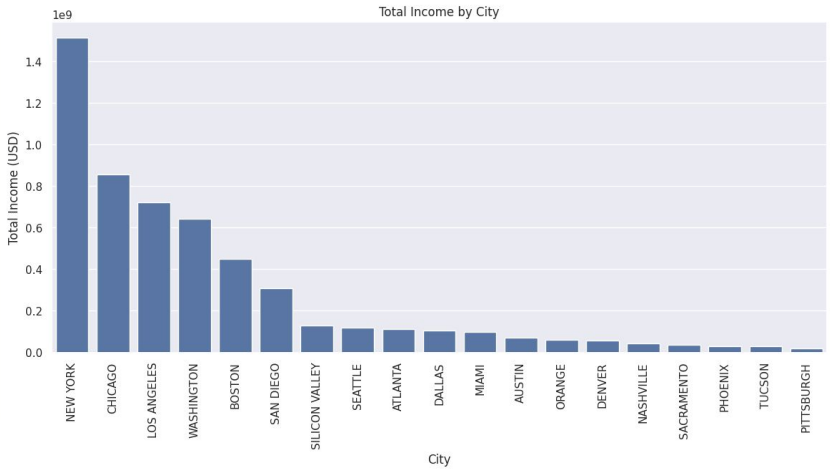


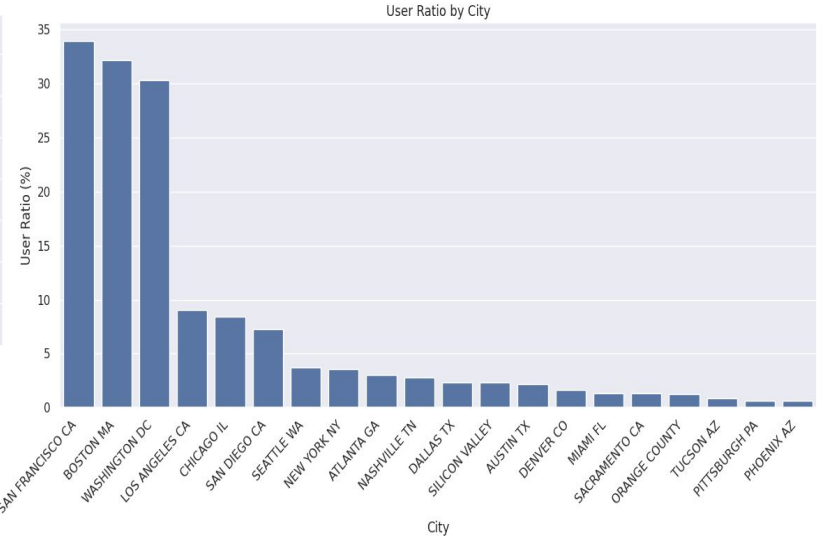


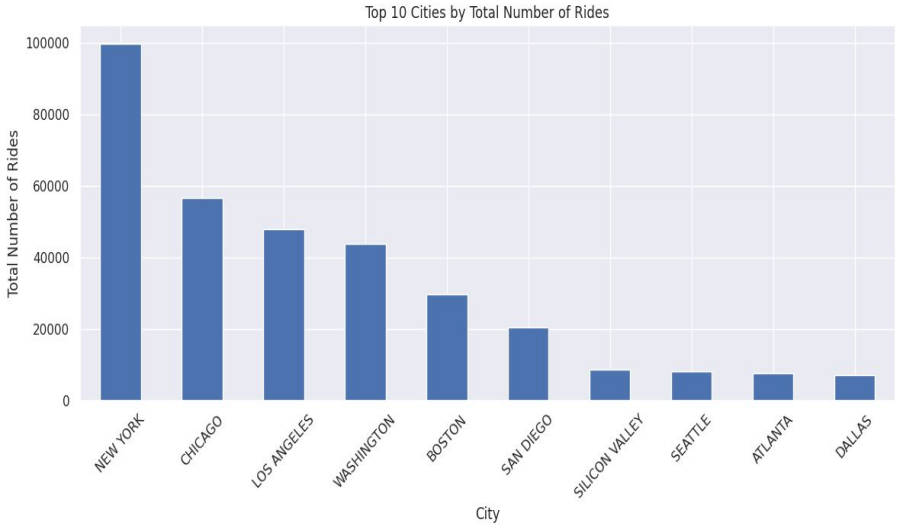


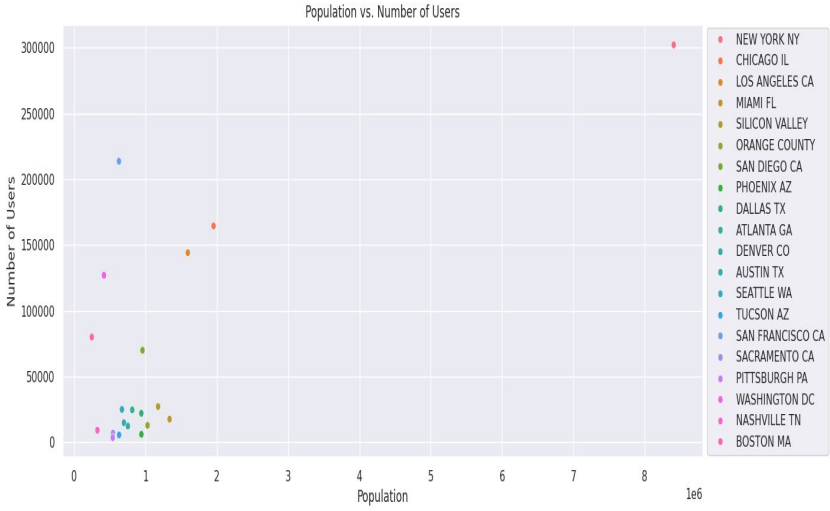
APPENDIX 3

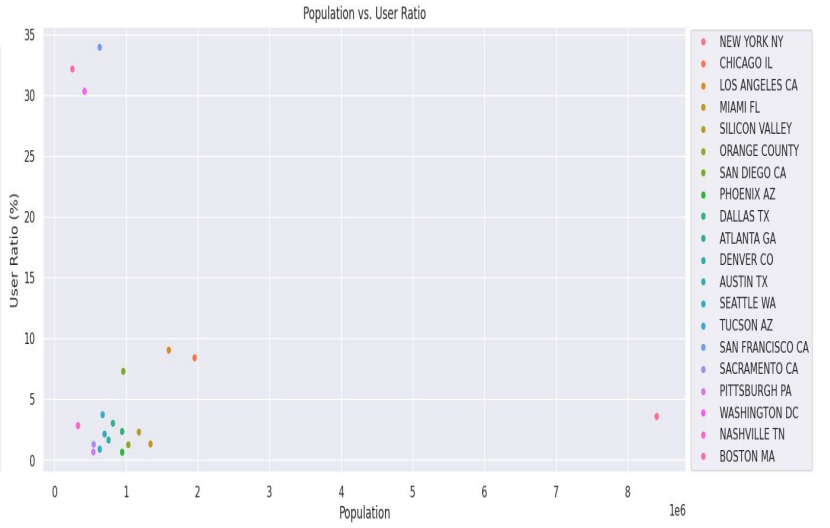

Top 10 Cities by Total Revenue
1e7
Total Revenue
City
NEW YORK
CHICAGO
LOS ANGELES
WASHINGTON
BOSTON
SAN DIEGO
SILICON VALLEY
DALLAS
ATLANTA
SEATTLE


Total Income by City
1e9
Total Income (USD)
City
NEW YORK
CHICAGO
LOS ANGELES
WASHINGTON
BOSTON
SAN DIEGO
SILICON VALLEY
SEATTLE
ATLANTA
DALLAS
MIAMI
AUSTIN
ORANGE
DENVER
NASHVILLE
SACRAMENTO
PHOENIX
TUCSON
PITTSBURGH

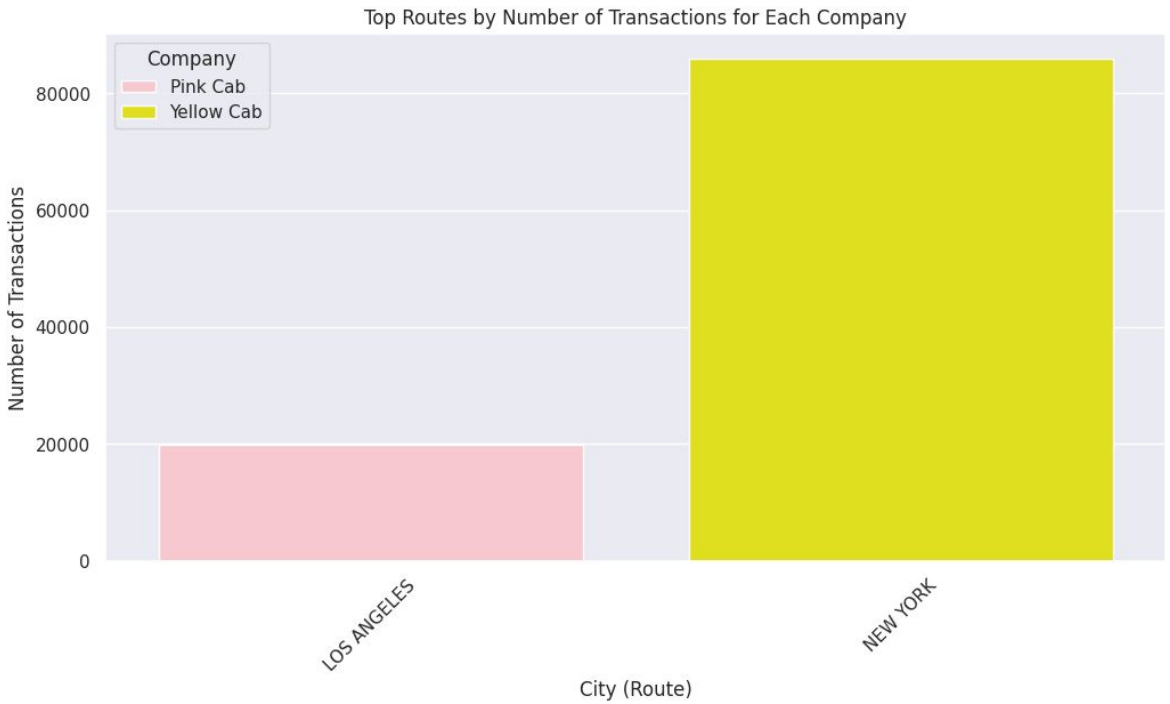

User Ratio by City
User Ratio (%)
City
SAN FRANCISCO CA
BOSTON MA
WASHINGTON DC
LOS ANGELES CA
CHICAGO IL
SAN DIEGO CA
SEATTLE WA
NEW YORK NY
ATLANTA GA
NASHVILLE TN
DALLAS TX
SILICON VALLEY
AUSTIN TX
DENVER CO
MIAMI FL
SACRAMENTO CA
ORANGE COUNTY
TUCSON AZ
PITTSBURGH PA
PHOENIX AZ

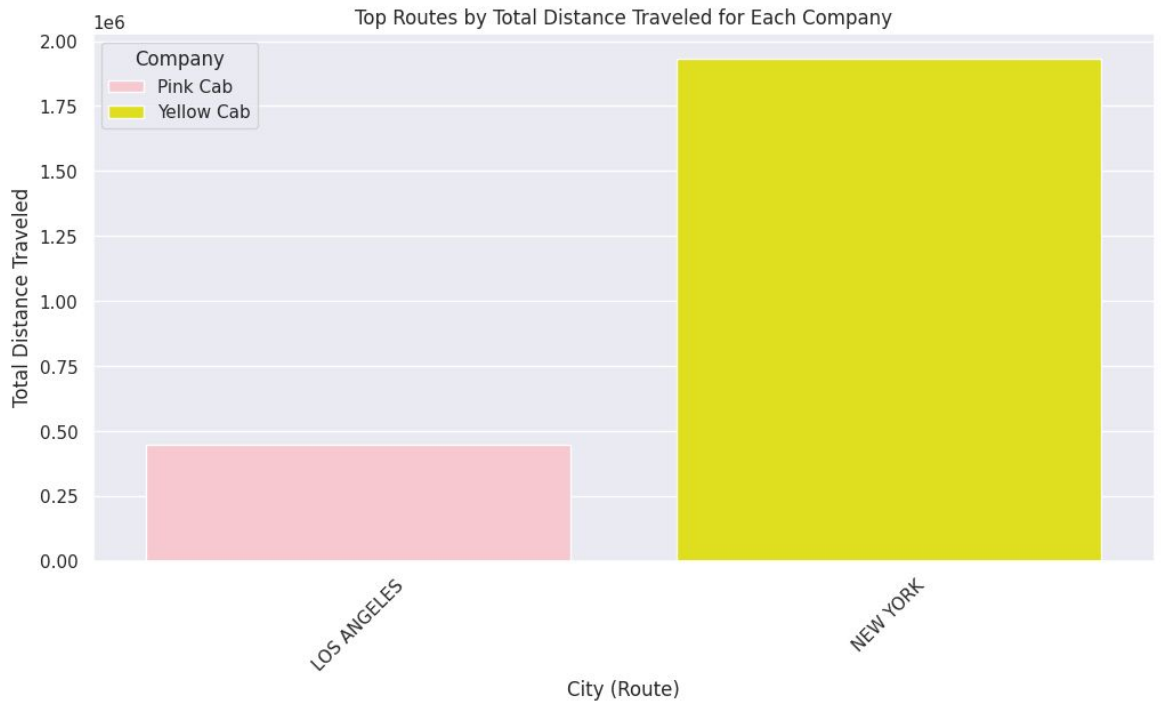

Top 10 Cities by Total Number of Rides
Total Number of Rides
City
NEW YORK
CHICAGO
LOS ANGELES
WASHINGTON
BOSTON
SAN DIEGO
SILICON VALLEY
SEATTLE
ATLANTA
DALLAS


Population vs. Number of Users
Number of Users
Population
1e6
NEW YORK NY
CHICAGO IL
LOS ANGELES CA
MIAMI FL
SILICON VALLEY
ORANGE COUNTY
SAN DIEGO CA
PHOENIX AZ
DALLAS TX
ATLANTA GA
DENVER CO
AUSTIN TX
SEATTLE WA
TUCSON AZ
SAN FRANCISCO CA
SACRAMENTO CA
PITTSBURGH PA
WASHINGTON DC
NASHVILLE TN
BOSTON MA


Population vs. User Ratio
User Ratio (%)
Population
1e6
NEW YORK NY
CHICAGO IL
LOS ANGELES CA
MIAMI FL
SILICON VALLEY
ORANGE COUNTY
SAN DIEGO CA
PHOENIX AZ
DALLAS TX
ATLANTA GA
DENVER CO
AUSTIN TX
SEATTLE WA
TUCSON AZ
SAN FRANCISCO CA
SACRAMENTO CA
PITTSBURGH PA
WASHINGTON DC
NASHVILLE TN
BOSTON MA

APPENDIX 4


Top Routes by Number of Transactions for Each Company
Company
Pink Cab
Yellow Cab
Number of Transactions
0
20000
40000
60000
80000
LOS ANGELES
NEW YORK
City (Route)


Top Routes by Total Distance Traveled for Each Company
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Company
Pink Cab
Yellow Cab
Total Distance Traveled
LOS ANGELES
NEW YORK
City (Route)

APPENDIX 5

We observed the presence of data points with negative profit values for both companies.

Pink Cab Profit Analysis:

13.14% of Pink Cab trips result in negative profitability.

Negative profit instances:

2016: 3232 (13.01%)

2017: 4362 (14.40%)

2018: 3535 (12.00%)

2019: No data available.

Possible reasons for negative profits: high operational costs, inefficient pricing, and unforeseen circumstances.

Concerning proportion of trips yielding negative profits.

Yellow Cab Profit Analysis:

4.98% of Yellow Cab trips result in negative profitability.

Negative profit instances:

2016: 3879 (4.77%)

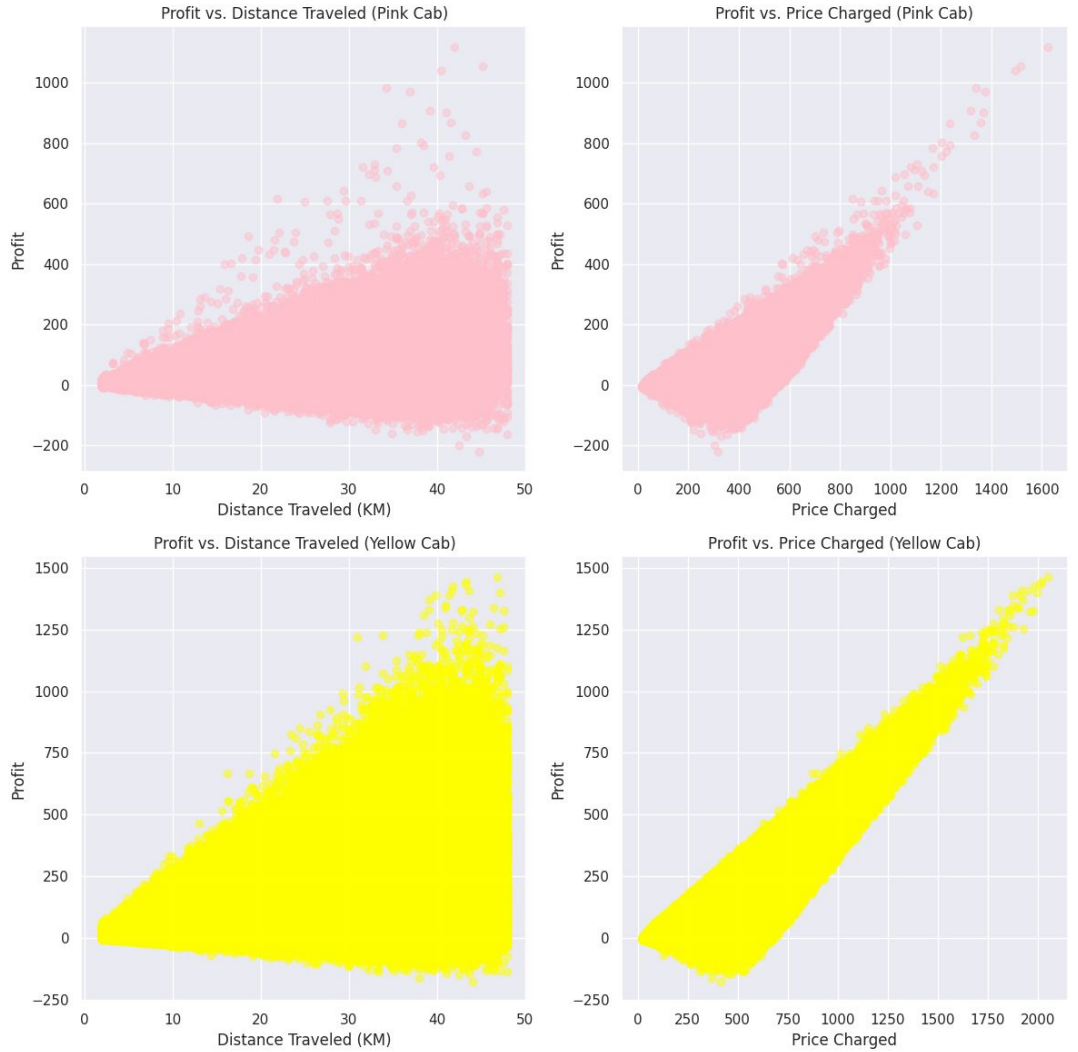
2017: 5071 (5.16%)

2018: 4704 (4.97%)

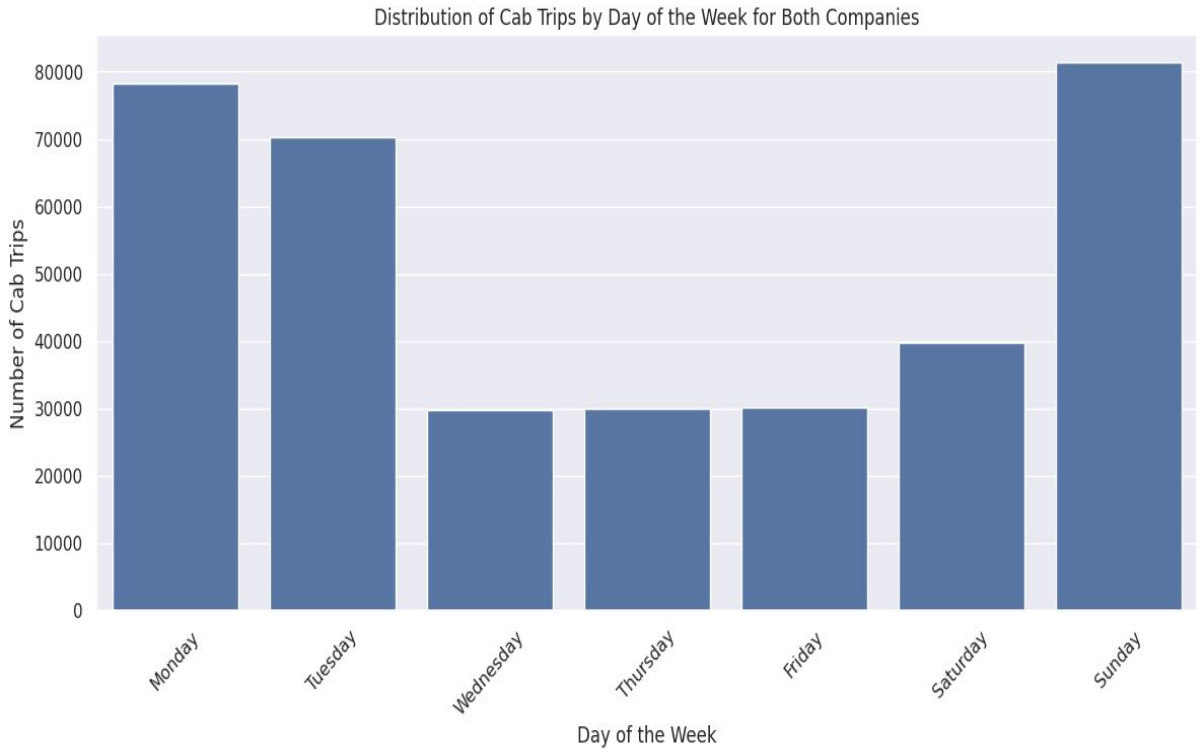
2019: 36 (9.02%)

Lower percentage of negative profit instances compared to Pink Cab.

Suggests better cost management or pricing strategies.



APPENDIX 6


Distribution of Cab Trips by Day of the Week for Both Companies
Number of Cab Trips
80000
70000
60000
50000
40000
30000
20000
10000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week


Pink Cab - Distribution of Cab Trips
50000
40000
30000
20000
10000
0
Yellow Cab - Distribution of Cab Trips
Number of Cab Trips
175000
150000
125000
100000
75000
50000
25000
0
Weekend
Weekday
Day Type

APPENDIX 7

Price Charged and Profit: These two variables exhibit a strong positive correlation (0.86). It suggests that as the price charged for a cab ride increases, the profit tends to increase as well.

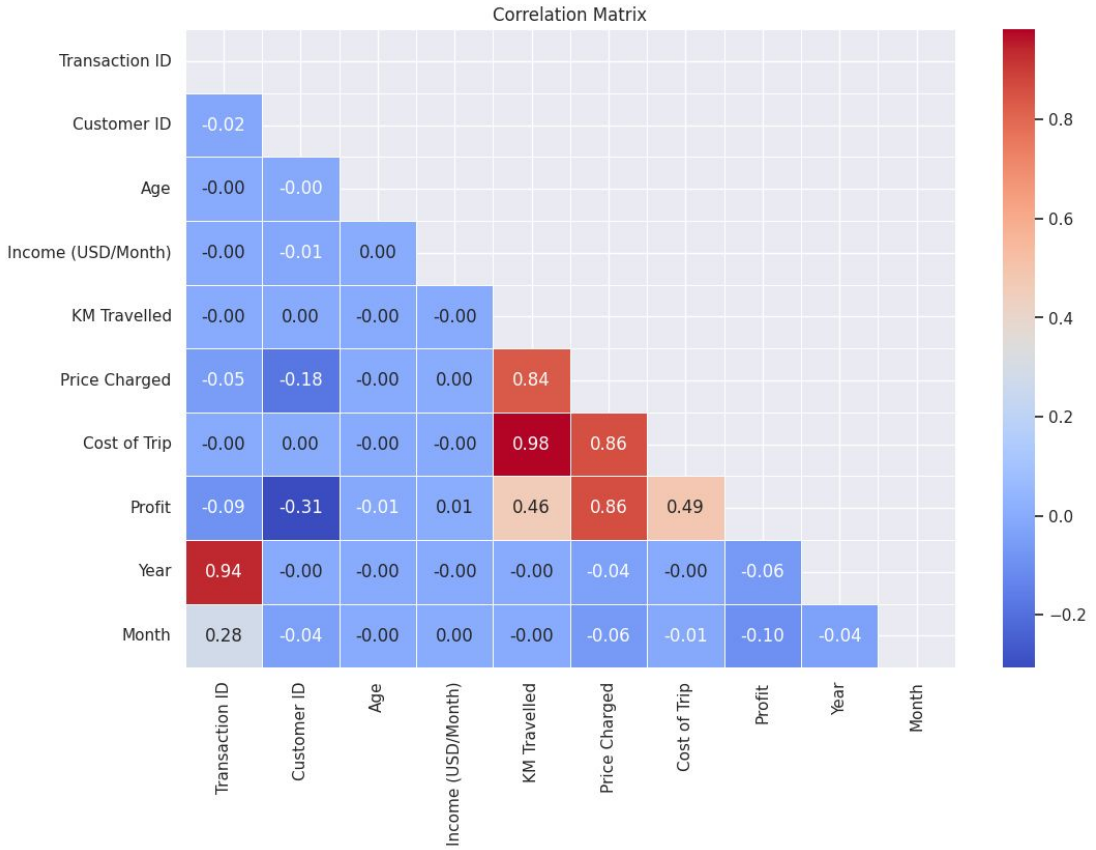
Cost of Trip and Profit: Similarly, there is a strong positive correlation (0.49) between the cost of the trip and profit. This indicates that as the cost of the trip increases, the profit tends to increase too.

Price Charged and Cost of Trip: There is a strong positive correlation between the price charged for a cab ride and the cost of the trip. This suggests that as the cost of providing the service increases, the price charged to the customer also tends to increase, which is expected in a profitable business model.

Cost of Trip and KM Travelled: There is an extremely strong positive correlation between the cost of the trip and the distance traveled (in kilometers). This indicates that the cost of providing the service is directly proportional to the distance traveled, which is intuitive as longer distances typically incur higher costs.

Price Charged and KM Travelled: There is a strong positive correlation between the price charged for a cab ride and the distance traveled (in kilometers). This suggests that customers are charged more for longer-distance rides, which aligns with common pricing models in the transportation industry.

Year and Transaction ID: There is a very strong positive correlation between the year and the transaction ID. This indicates a consistent increase in the number of transactions over the years, which could be attributed to the growth and expansion of the cab service business over time.



soufleros.kostas@gmail.com

Data Glacier

Your Deep Learning Partner

Thank You