Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

06/20/2024

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Data Glacier

Your Deep Learning Partner

Executive Summary

The Client: XYZ

- Private firm in the US
- Interest: Investing in the Cab Industry due to recent growth
- Objective: Understand the market for informed investment decision

Project Delivery

- Data Sets Provided: Information on two cab companies
- Outcome: Presentation to XYZ's Executive team

Evaluation Criteria:

- Visuals
- Quality of analysis
- Value of recommendations and insights

Data Sets

- Cab_Data.csv – Transaction details for two cab companies
- Customer_ID.csv – Customer demographic details
- Transaction_ID.csv – Transaction to customer mapping and payment mode
- City.csv – US cities, population, and cab users

Cab_Data.csv

- Observations: 359,392
- Features: 7 (Transaction ID, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip)
- Size: 20.1 MB

Customer_ID.csv

- Observations: 49171
- Features: 4
- Size: 1.3 MB

Transaction_ID.csv

- Observations: 440098
- Features: 3
- Size: 8.7 MB

City.csv

- Observations: 20
- Features: 3
- Size: 1 KB

Problem Statement

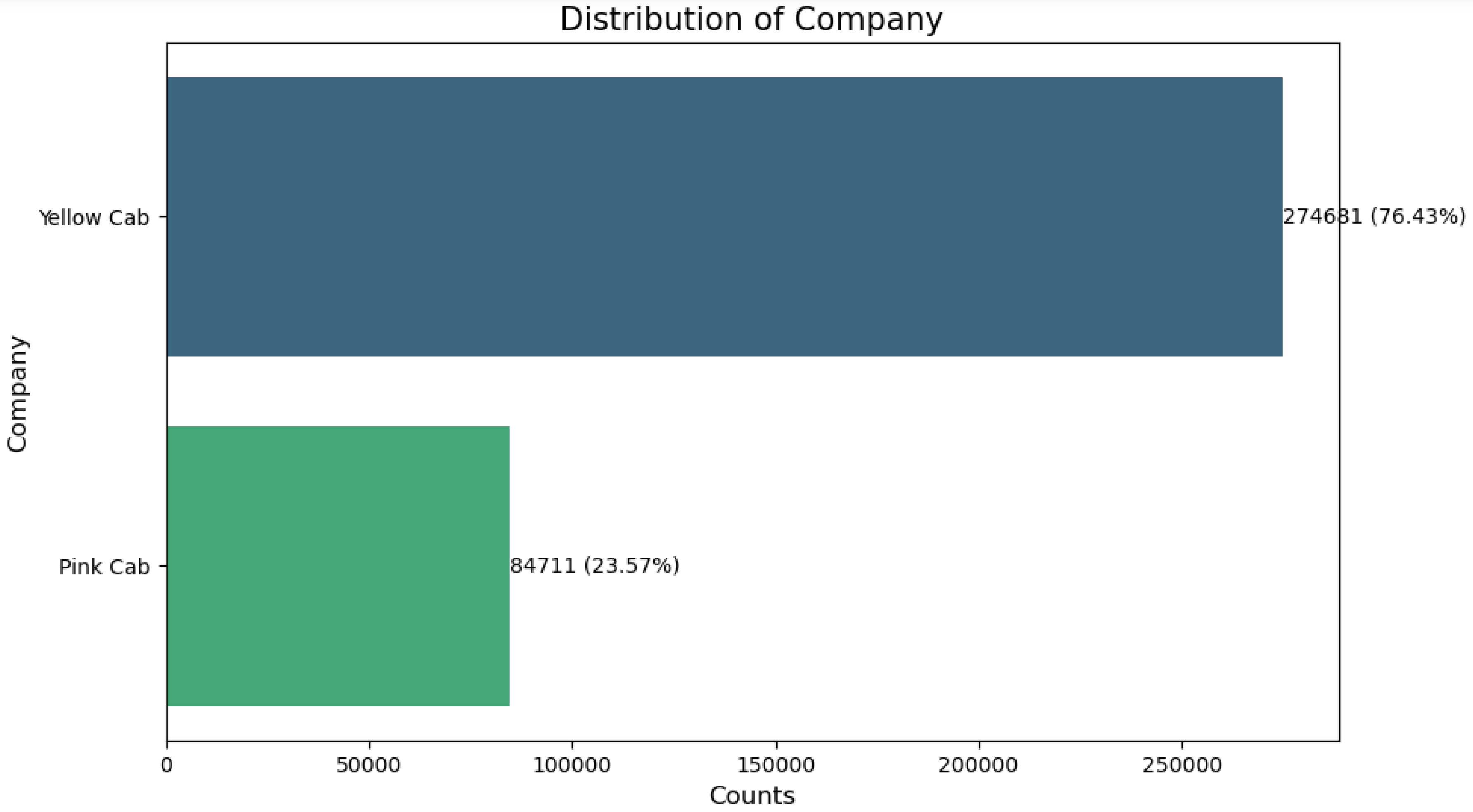
- Analyze customer profiles, transaction details, and city-specific information
- Preprocess and merge datasets for comprehensive analysis
- Perform exploratory and hypothesis-driven analyses

Approach

- Reading the CSV files
- Ensuring column names are consistent
- Renaming Columns
- Converting relevant columns to string
- Inspecting Columns
- Merging the DataFrames
- Displaying the final DataFrame
- Exploratory data analysis on final dataset
- Performed the hypothesis test

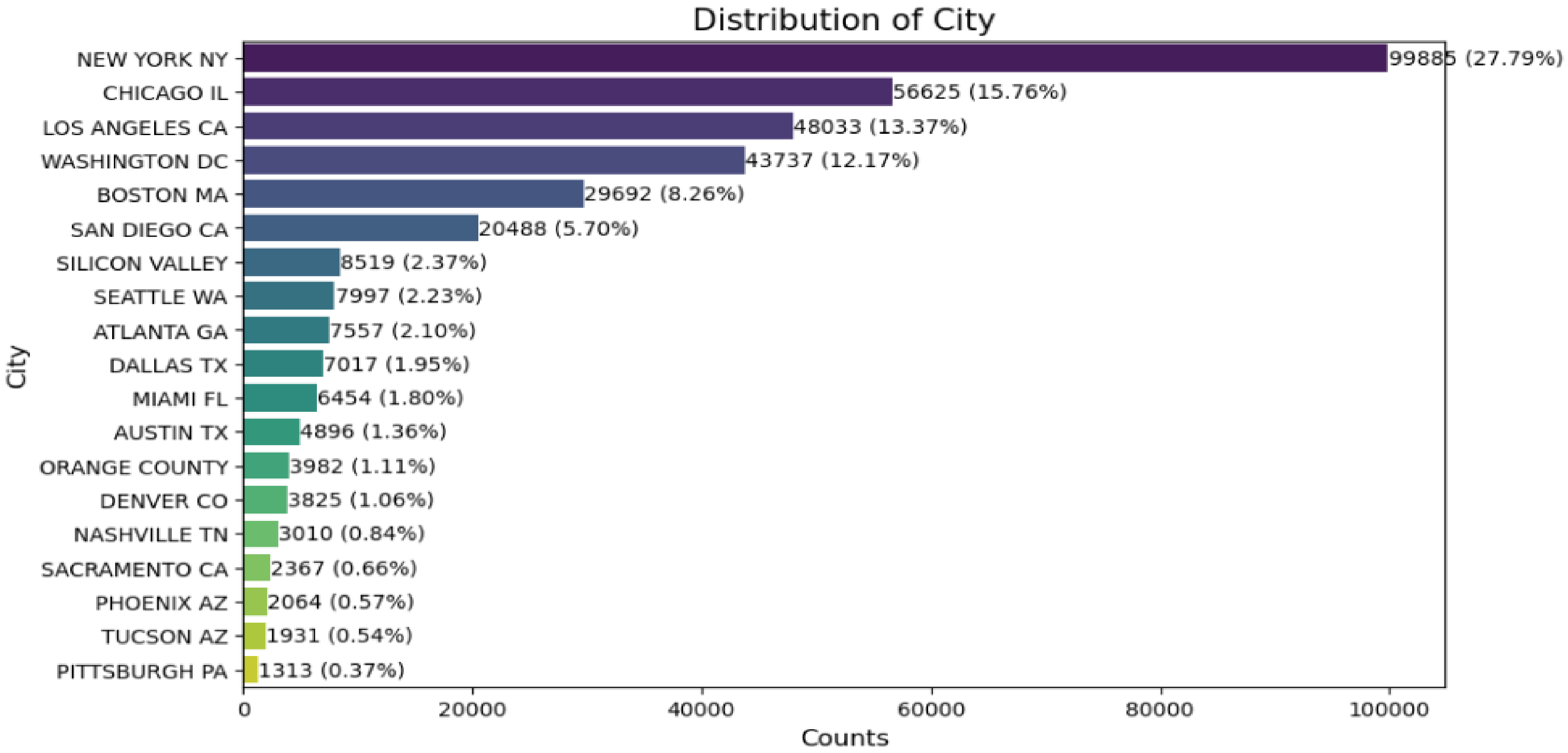
EDA(Cab_Data)

Distribution of Company

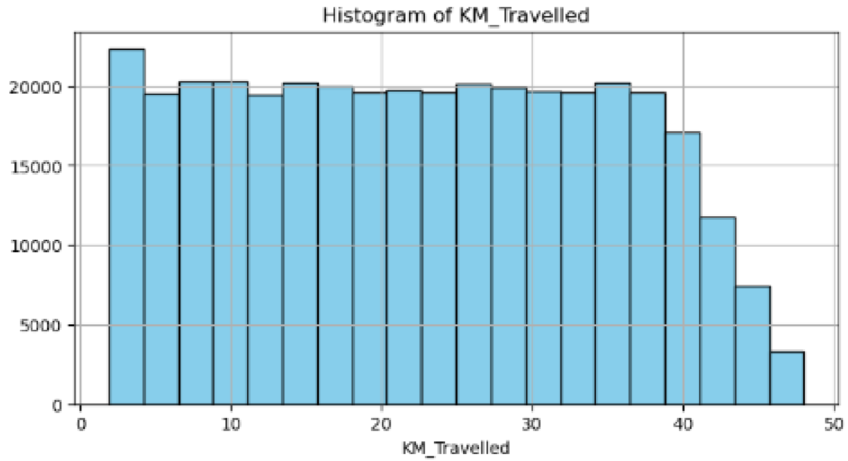


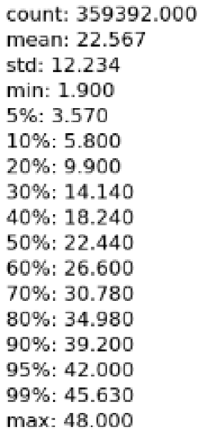
- The Yellow Cab company leads with 76.43% of the operations, while the Pink Cab company represents 23.57%.This clearly shows Yellow Cab's significant advantage.

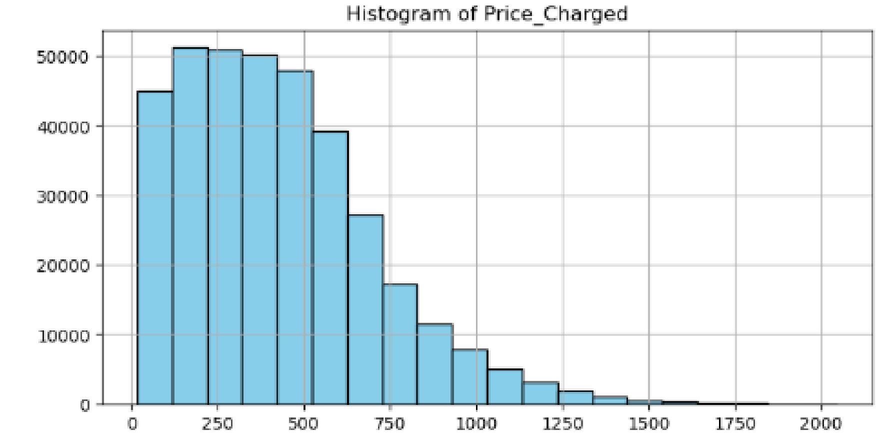
Data Glacier
Your Deep Learning Partner

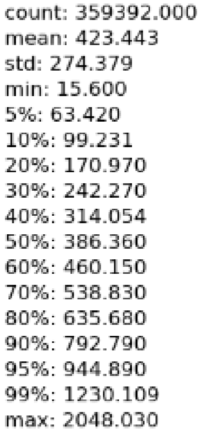


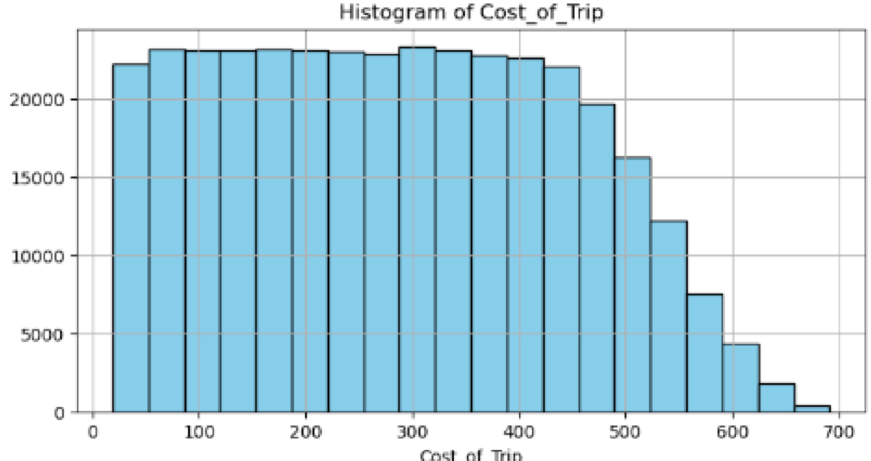
- In terms of cities, NEW YORK, CHICAGO, and LOS ANGELES have the highest number of transactions.

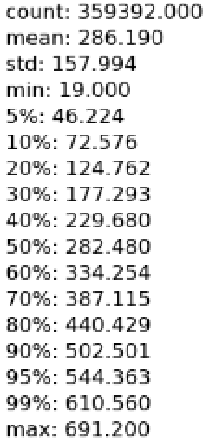

Histogram of KM_Travelled


```
count: 359392.000
mean: 22.567
std: 12.234
min: 1.900
5%: 3.570
10%: 5.800
20%: 9.900
30%: 14.140
40%: 18.240
50%: 22.440
60%: 26.600
70%: 30.780
80%: 34.980
90%: 39.200
95%: 42.000
99%: 45.630
max: 48.000
```


Histogram of Price_Charged


```
count: 359392.000
mean: 423.443
std: 274.379
min: 15.600
5%: 63.420
10%: 99.231
20%: 170.970
30%: 242.270
40%: 314.054
50%: 386.360
60%: 460.150
70%: 538.830
80%: 635.680
90%: 792.790
95%: 944.890
99%: 1230.109
max: 2048.030
```


Histogram of Cost_of_Trip

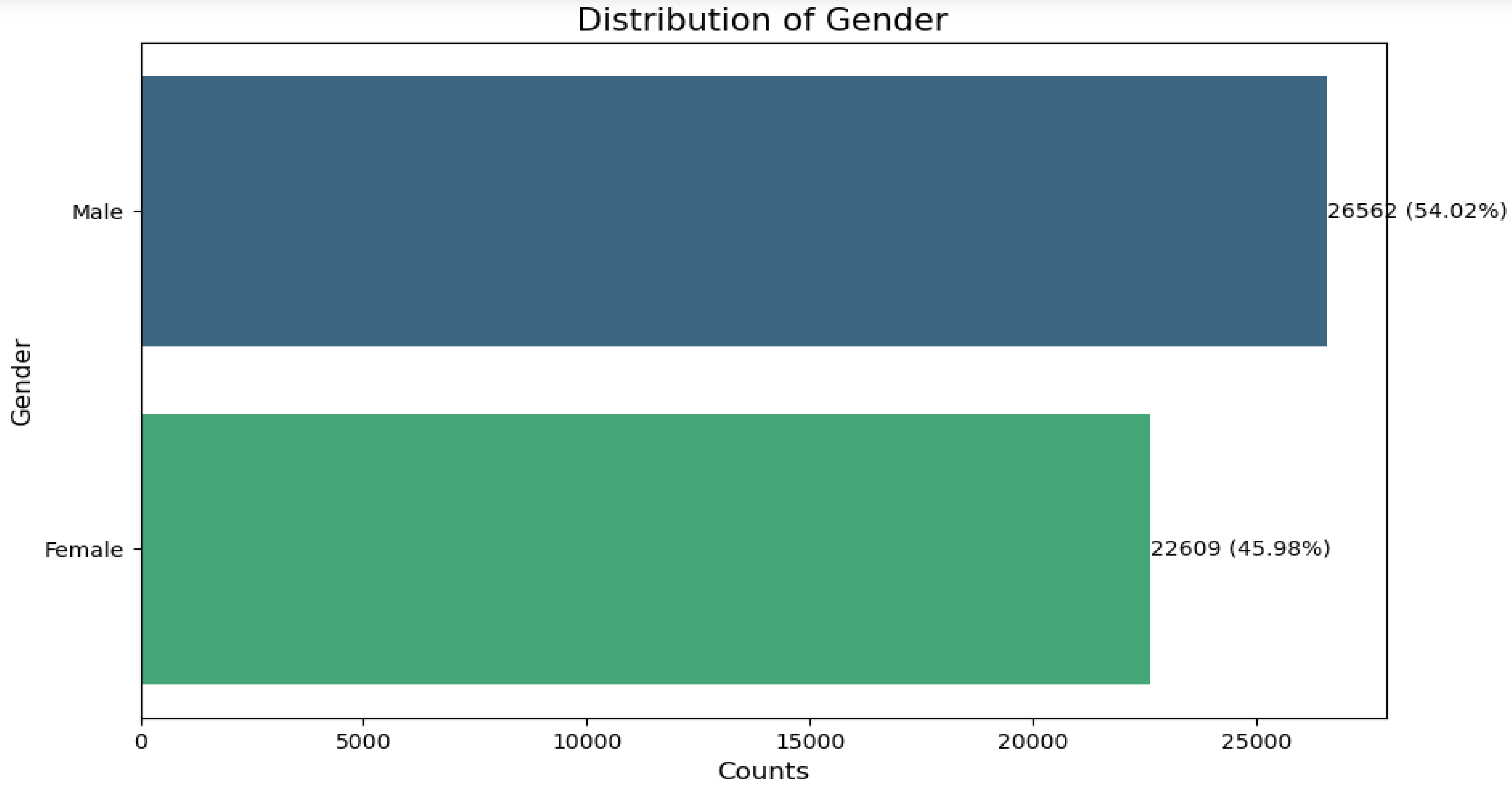

```
count: 359392.000
mean: 286.190
std: 157.994
min: 19.000
5%: 46.224
10%: 72.576
20%: 124.762
30%: 177.293
40%: 229.680
50%: 282.480
60%: 334.254
70%: 387.115
80%: 440.429
90%: 502.501
95%: 544.363
99%: 610.560
max: 691.200
```

- The graph shows that the shortest distance traveled is 1.9 km, and the longest is 48 km.

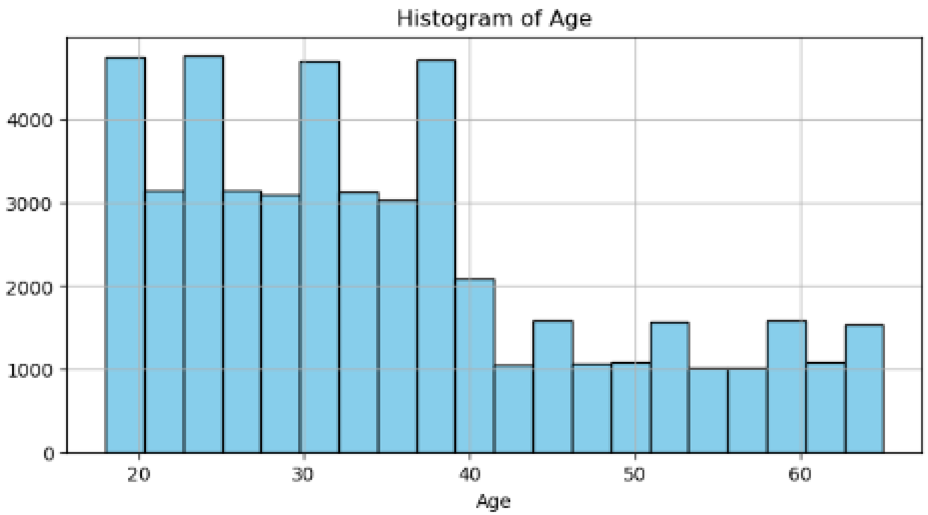
- Examining the Price_Charged variable reveals a lack of normal distribution, likely due to outliers, with a maximum value of around 2048 and a median value of about 386.

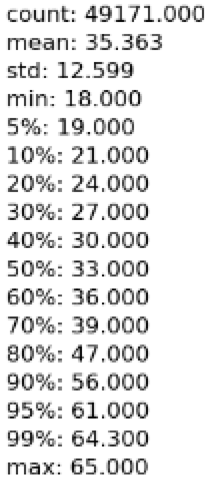
- The minimum trip cost is 19 dollars, while the maximum is approximately 691 dollars.

EDA (Customer_Data)

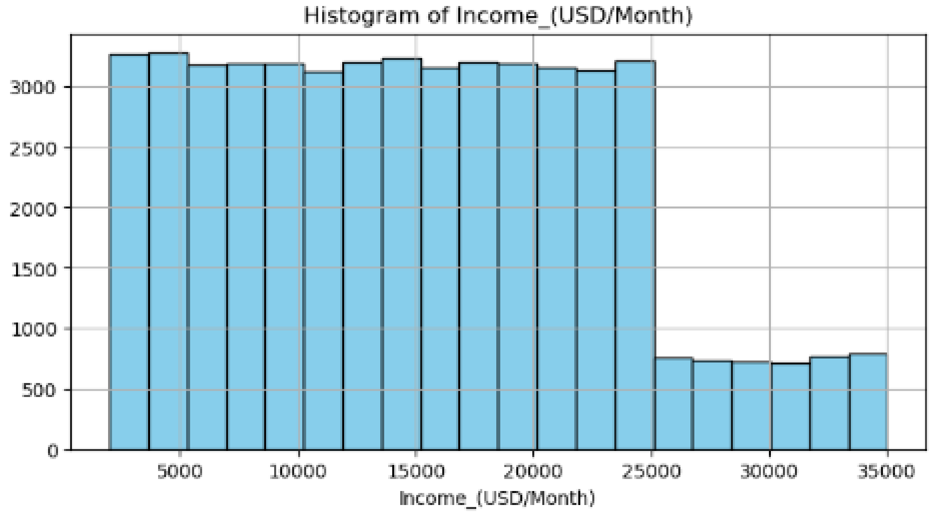


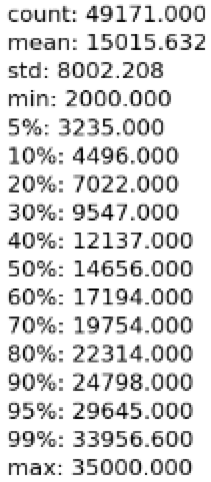
- Upon examining the gender distribution of customers, it is noted that male customers constitute approximately 54%, whereas female customers represent about 46%.





```
count: 49171.000
mean: 35.363
std: 12.599
min: 18.000
5%: 19.000
10%: 21.000
20%: 24.000
30%: 27.000
40%: 30.000
50%: 33.000
60%: 36.000
70%: 39.000
80%: 47.000
90%: 56.000
95%: 61.000
99%: 64.300
max: 65.000
```

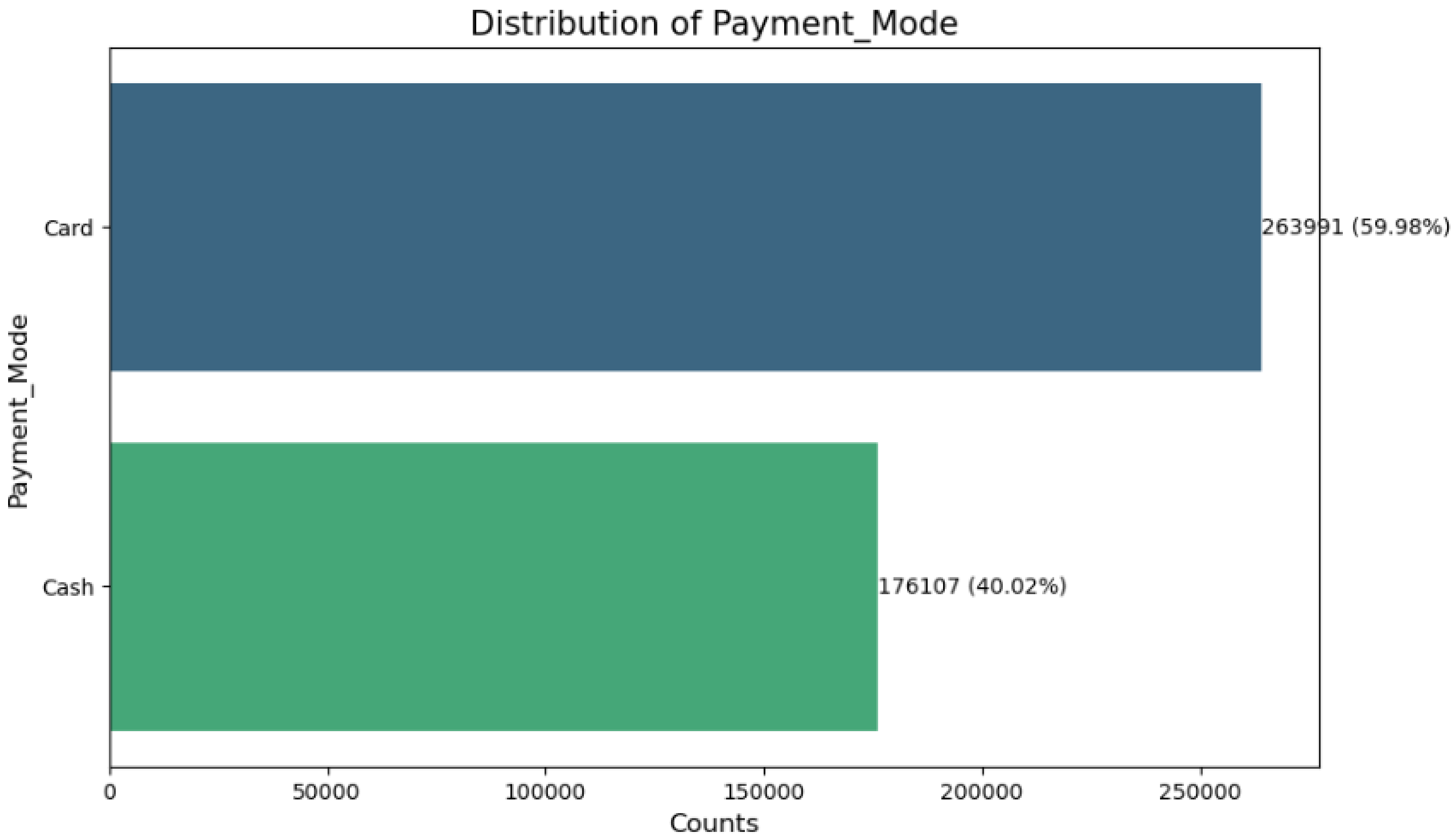




```
count: 49171.000
mean: 15015.632
std: 8002.208
min: 2000.000
5%: 3235.000
10%: 4496.000
20%: 7022.000
30%: 9547.000
40%: 12137.000
50%: 14656.000
60%: 17194.000
70%: 19754.000
80%: 22314.000
90%: 24798.000
95%: 29645.000
99%: 33956.600
max: 35000.000
```

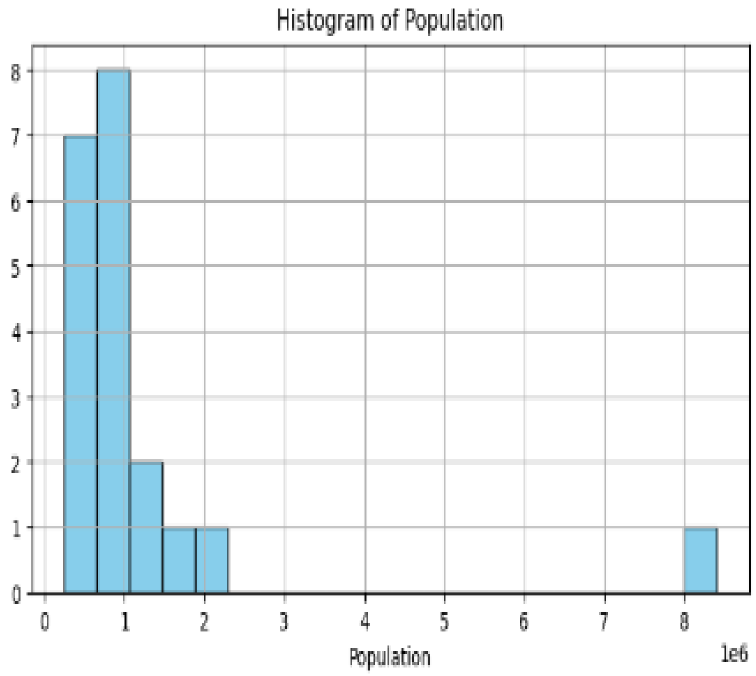
- More than 70% of customers are below the age of 40.
- Around 10% of customers have a monthly income exceeding $25,000.

EDA (Tramsaction_Data)

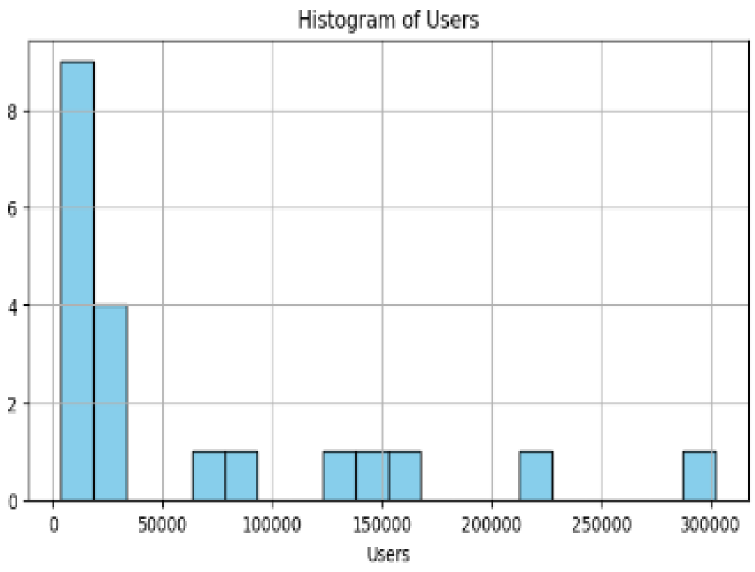


- When we examine the graph, we observe that nearly 60% of the transactions have been made using a card.

EDA (City_Data)


Histogram of Population

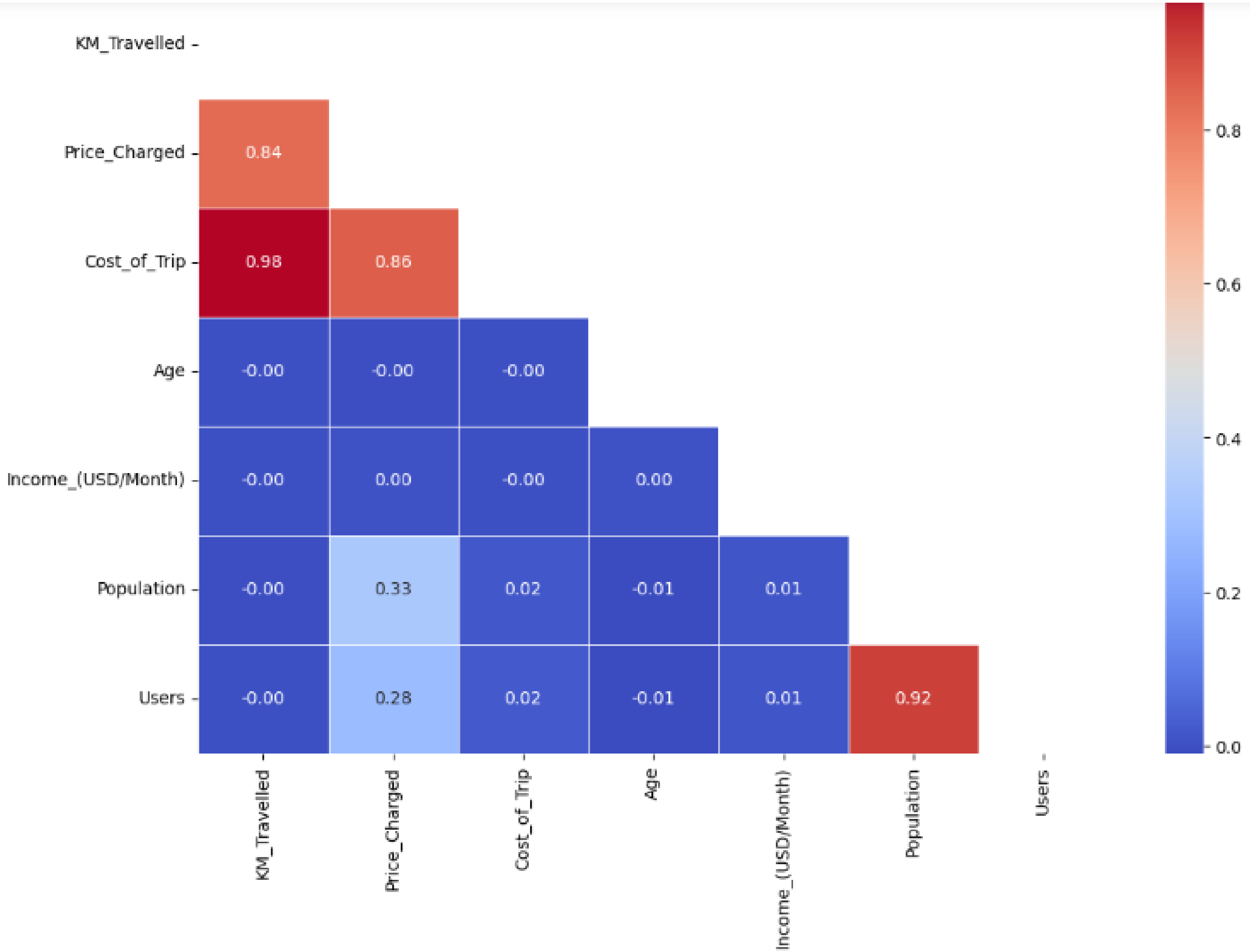
```
count: 20.000
mean: 1231592.000
std: 1740126.700
min: 248968.000
5%: 323312.150
10%: 409695.600
20%: 545037.800
30%: 630886.700
40%: 687517.800
50%: 784559.000
60%: 943344.400
70%: 980570.400
80%: 1209918.200
90%: 1631046.300
95%: 2277665.350
99%: 7180202.670
max: 8405837.000
```


Histogram of Users

```
count: 20.000
mean: 64520.650
std: 83499.375
min: 3643.000
5%: 5608.550
10%: 6090.900
20%: 8824.800
30%: 12822.100
40%: 16596.200
50%: 23429.000
60%: 25936.600
70%: 73002.800
80%: 130427.200
90%: 169382.100
95%: 218036.000
99%: 285326.400
max: 302149.000
```

- Observing the graph reveals that the city with the smallest population has about 250,000 residents, while the city with the largest population totals around 8.5 million.
- The number of users across cities where taxi companies operate ranges from a minimum of 3,643 to a maximum of 302,149.
- This dataset includes outlier values in the "Population" data, but I chose not to handle these outliers.

Correlation Analysis

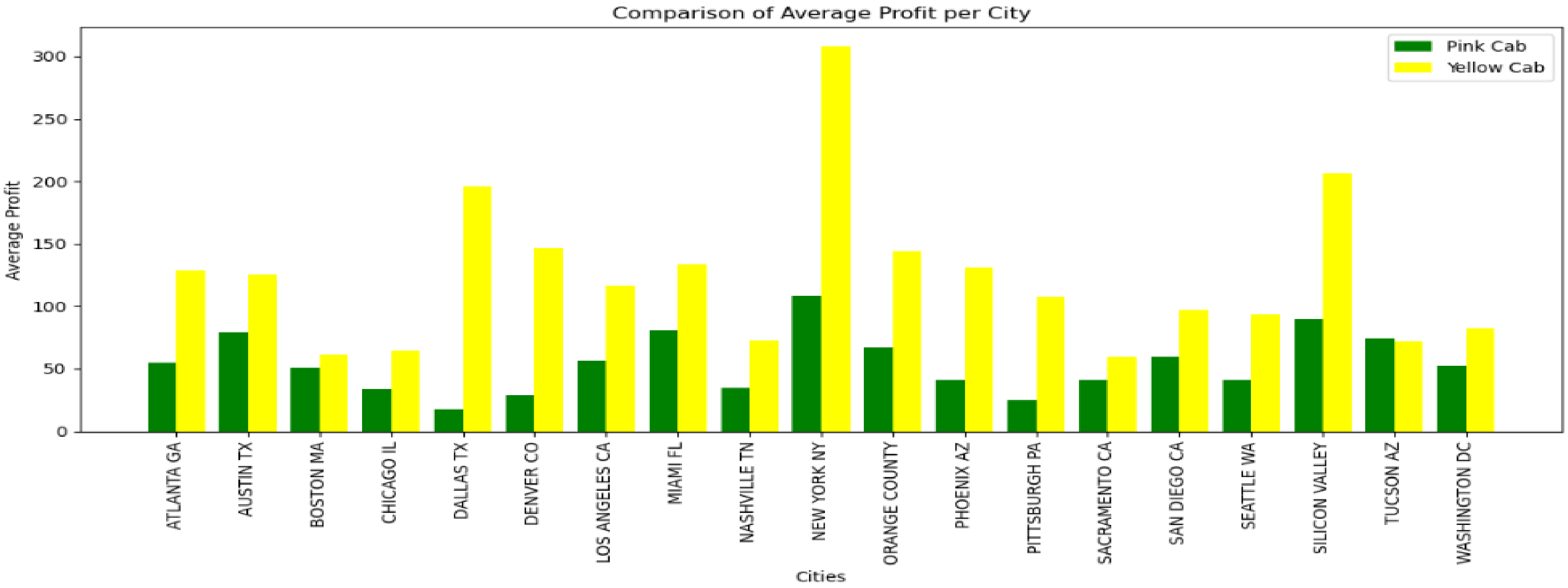


- There is a strong correlation between Population and Users based on observations.

- Moreover, there is a significant correlation among Price_Charged, Cost_of_Trip, and KM_Travelled.

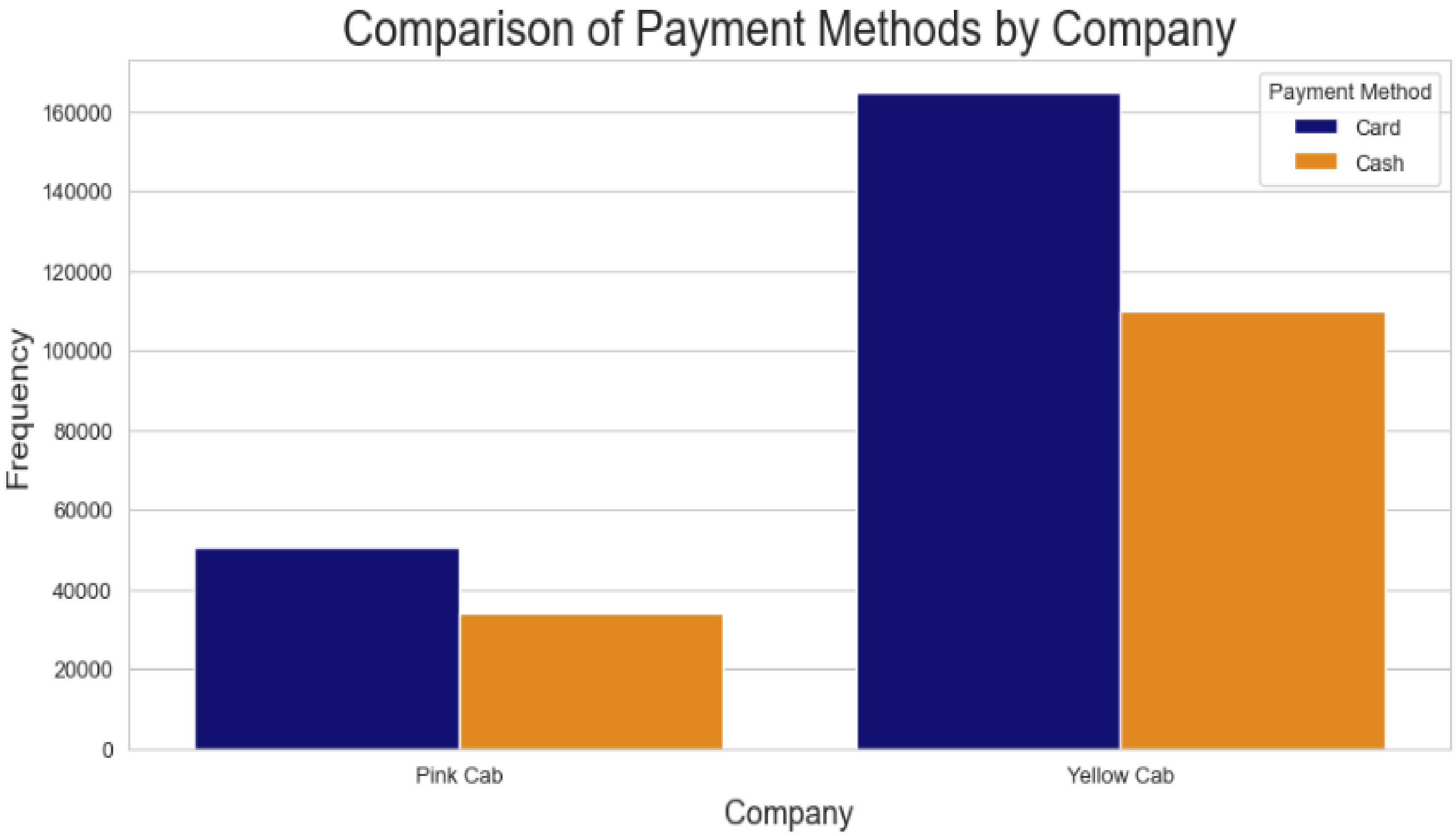
Data Glacier

Overall Analysis



- In every city except Tucson, AZ, Yellow Cab's average profit surpasses that of Pink Cab.

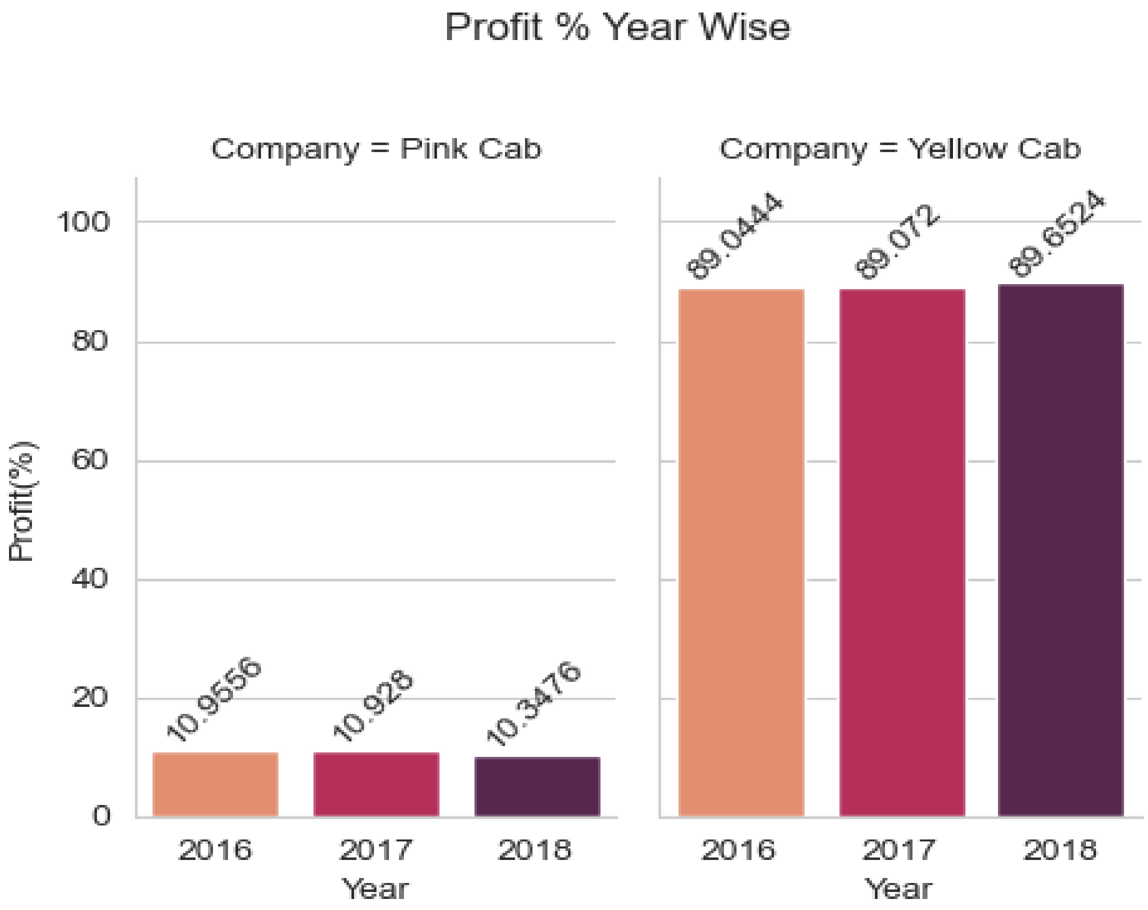
Payment Mode Comparison


Comparison of Payment Methods by Company

- Yellow Cab shows a larger overall number of transactions compared to Pink Cab, for both payment methods.

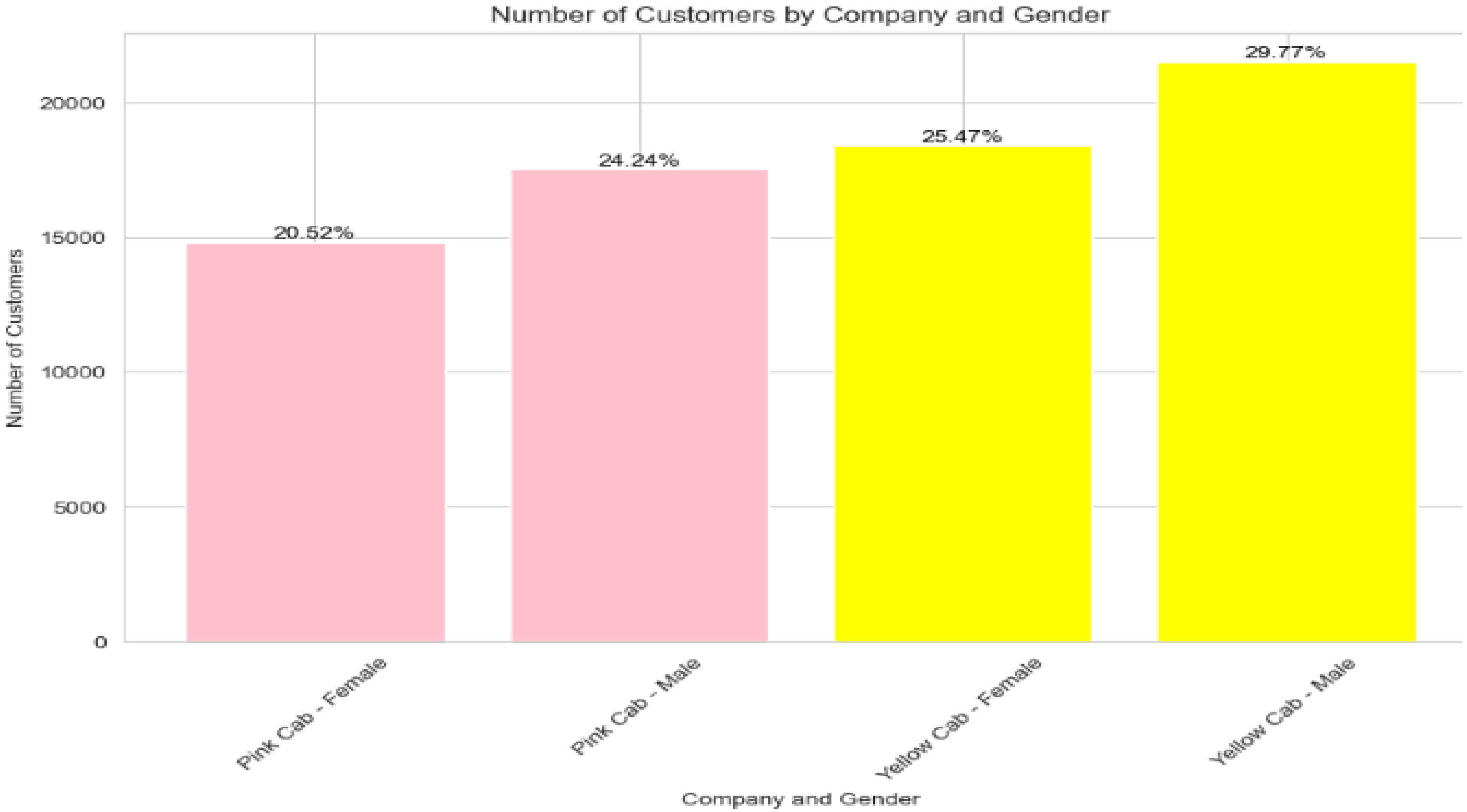
- The frequency of Cash payments is higher for Yellow Cab than Pink Cab, but the difference is less pronounced compared to Card payments.

Overall Profit by year



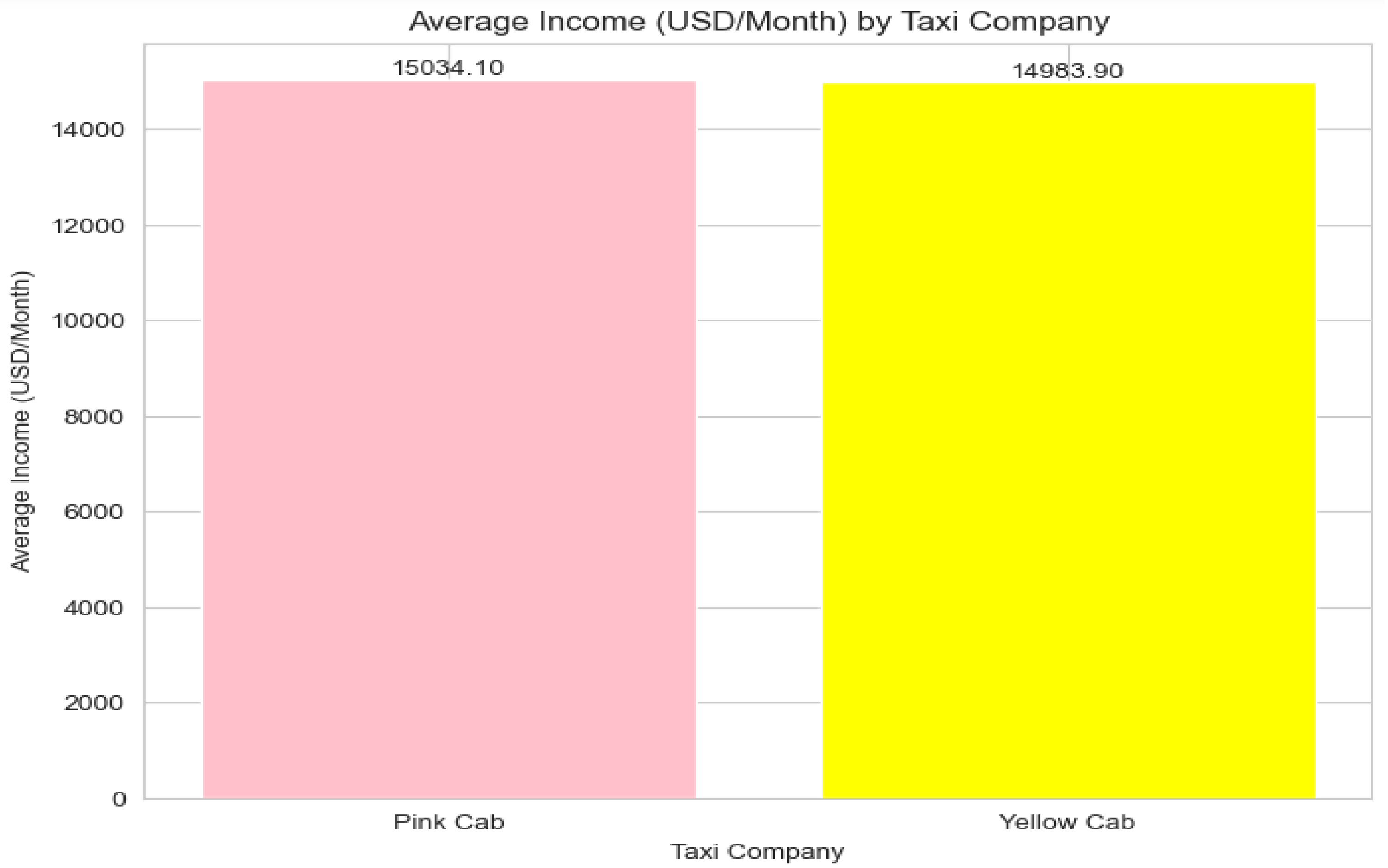
- Pink Cab shows a small decline in profit percentage over the three years.
- Yellow Cab maintains high profit percentages with a slight increase in 2018.

Hypothesis Testing



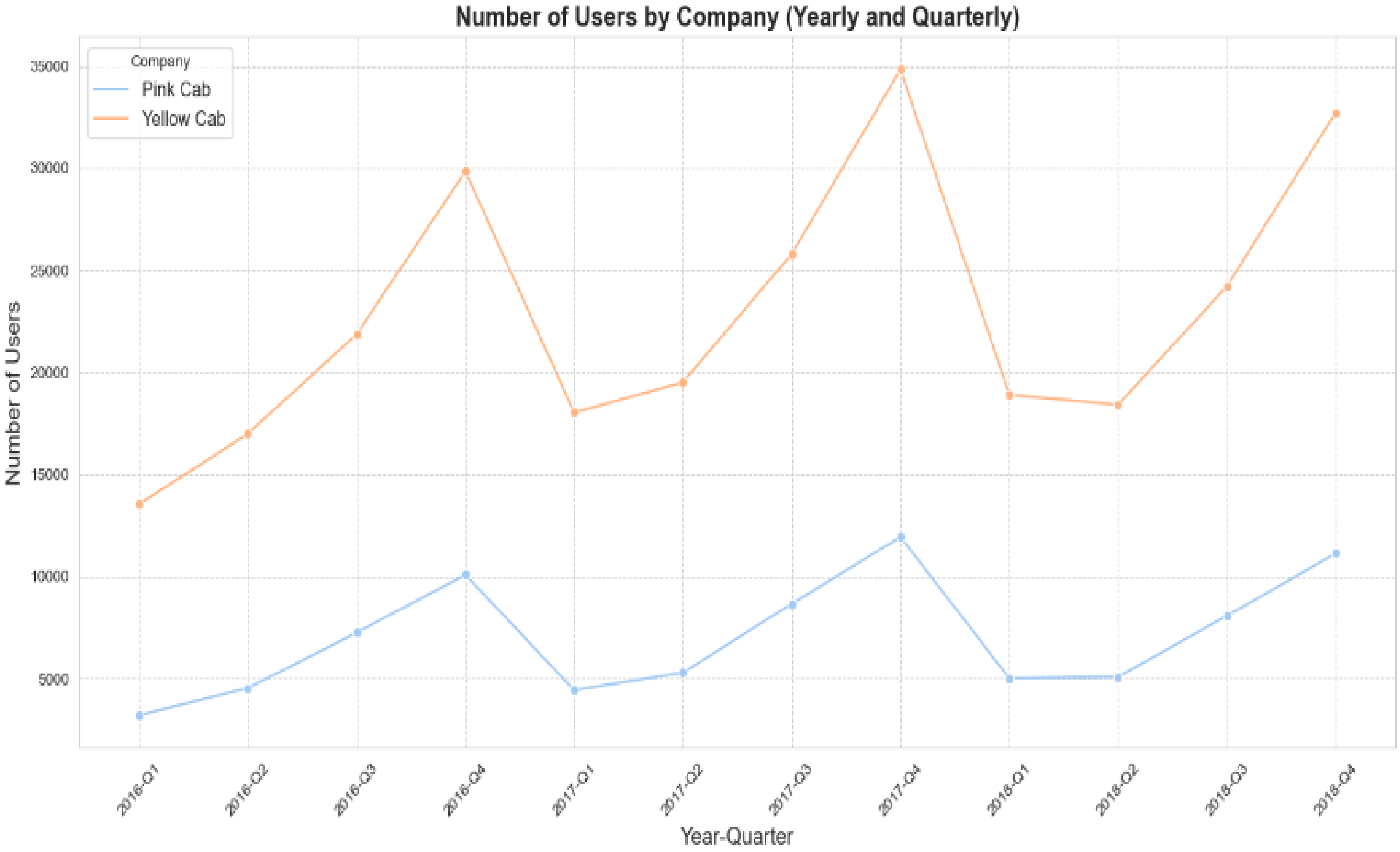
- When comparing the two cab companies, it is evident that both have a higher number of male users.

Avg Income of Cab customers



- The average monthly incomes of Pink Cab and Yellow Cab are very similar, with Pink Cab having a marginally higher income. This indicates that both companies have comparable revenue performance on a monthly basis.

Number of Users by Company (Yearly and Quarterly

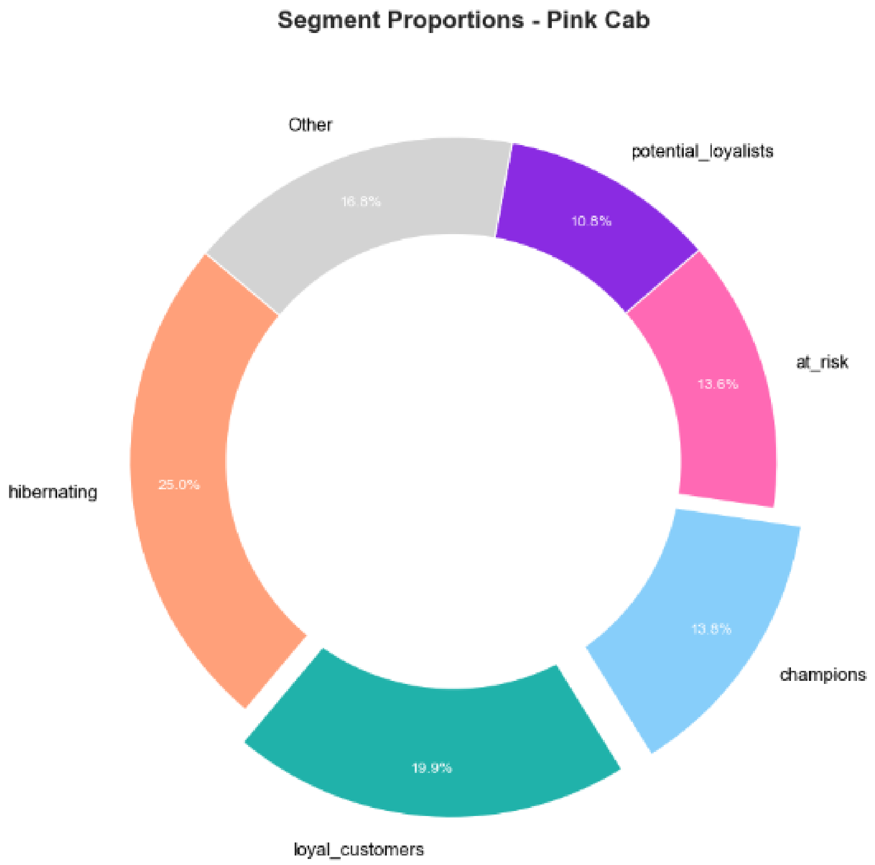


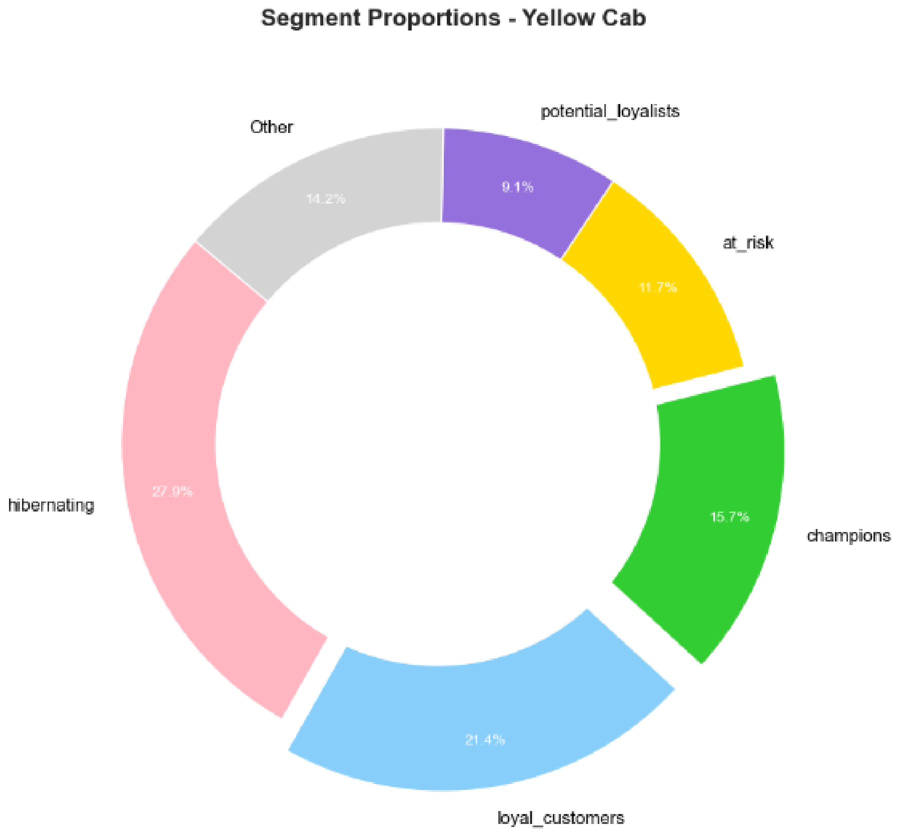
- Upon examining the graph, there is a noticeable rise in customer demand for cab usage from the first quarter to the fourth quarter in each of the years 2016, 2017, and 2018.

- The increasing demand for cab usage from the first quarter to the fourth quarter in each of the years 2016, 2017, and 2018 can be influenced by several factors, such as seasonal variations, economic factors, special events, holidays, urban population growth, and a preference for cabs during inclement weather conditions.

Data Glacier

The Impact of Loyal Customers and Champions on Profit in Yellow Cab and Pink Cab Companies


Segment Proportions - Pink Cab


Segment Proportions - Yellow Cab

- The hypothesis tests revealed that Yellow Cab's loyal customers and top users generate more profit compared to those of Pink Cab.

Recommendations

- Number of Users: Yellow Cab clearly has a larger user base compared to Pink Cab.
- Average Profit per City: Yellow Cab's average profit surpasses that of Pink Cab in all cities except Tucson, AZ.
- Loyal Customers: The higher number of loyal customers for Yellow Cab suggests a potential for more stable future revenue.
- Gender Preference: Yellow Cab's popularity among male users could play a crucial role in shaping its marketing strategies.

Thank You

Data Glacier

Your Deep Learning Partner