Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

20th June 2023

BY
SUKURAT SALAM

Agenda

Problem Statement

Data Processing

EDA

Hypothesis Testing

Summary

Recommendations

Data Glacier

Your Deep Learning Partner

INTRODUCTION

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.

Data Source

Link: https://github.com/DataGlacier/DataSets

The dataset contain 4 individual dataset for the period of 31/01/2016 to 31/12/2018.

- Cab data
- Transaction _ID data
- Customer_ID data
- City data

Objective of Study

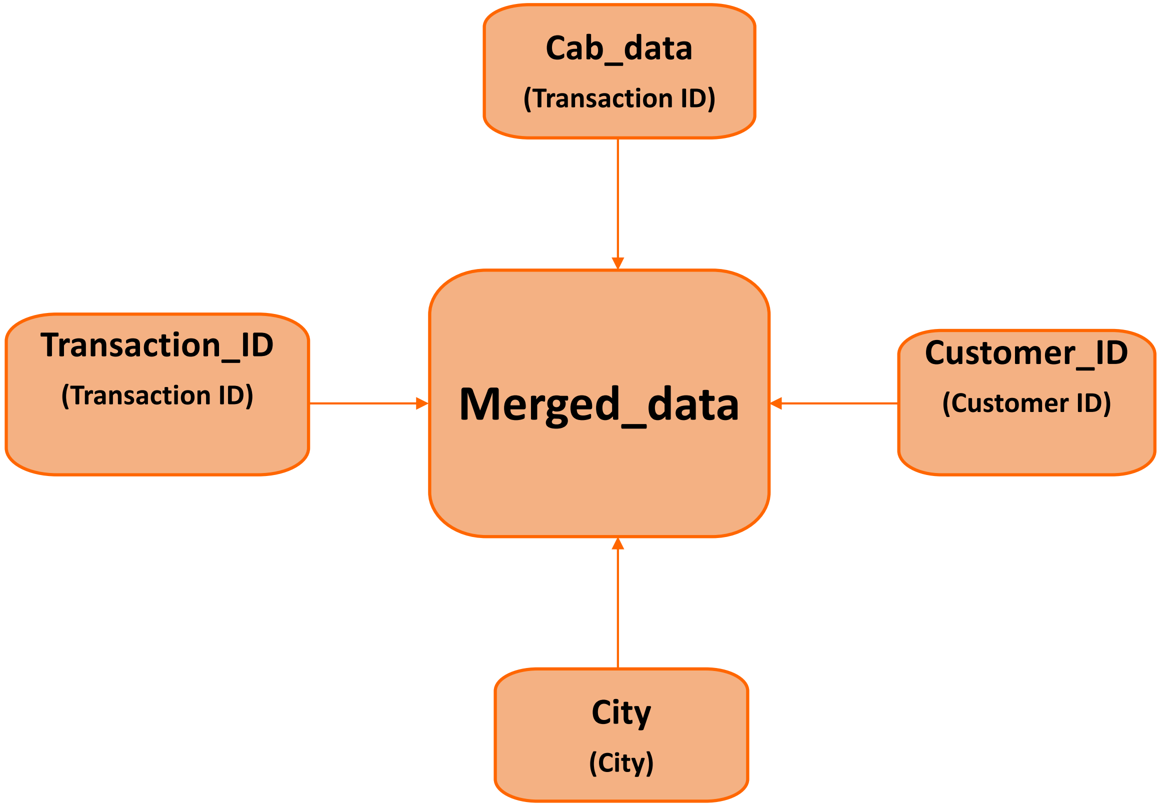
The objective of the study is to use actionable insights to help XYZ identify the right company to make their investment.

This will be achieved with the following hypothesis:

- What are the contributing factors to the profit made over the years?
- Is there any significant difference between the profit made by the two companies?
- Is there any significant difference between methods of payment by revenue generated?
- Is there any relationship between cities' population and the profit made?
- Is there any relationship between the kilometres covered by the two companies?
- Is there any significant difference in the price charged by cities?
- What is the yearly trend of the profit margin?

DATA PROCESSING

- The four datasets was merged to form a dataset that contains all the variables.
- The new column was calculated for the profit margin using the price charged and the cost of the trip variable.
- The age of customers was also categorized into 4 groups.
- The travelled date column was converted into dateTime format and split into day, month, and year.



EXPLORATORY DATA ANALYSIS WITH VISUALS

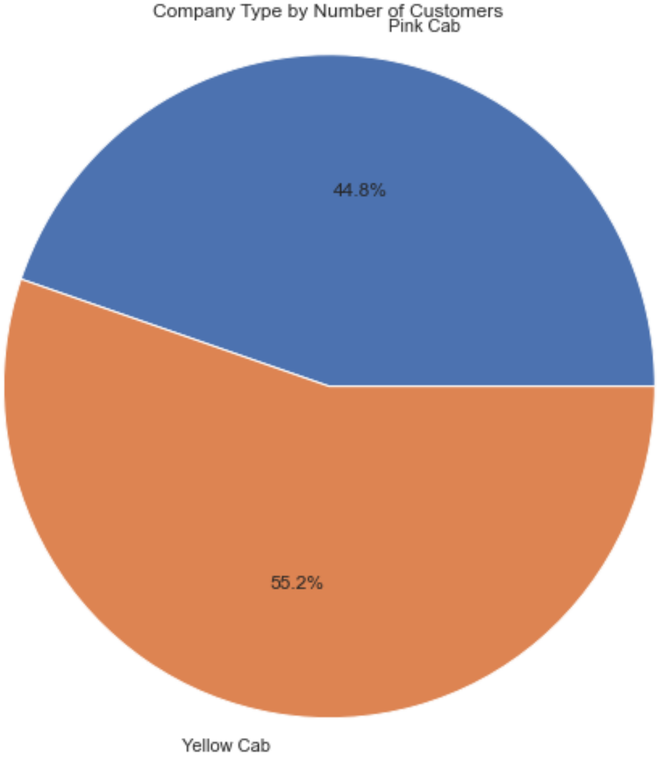


Fig 1

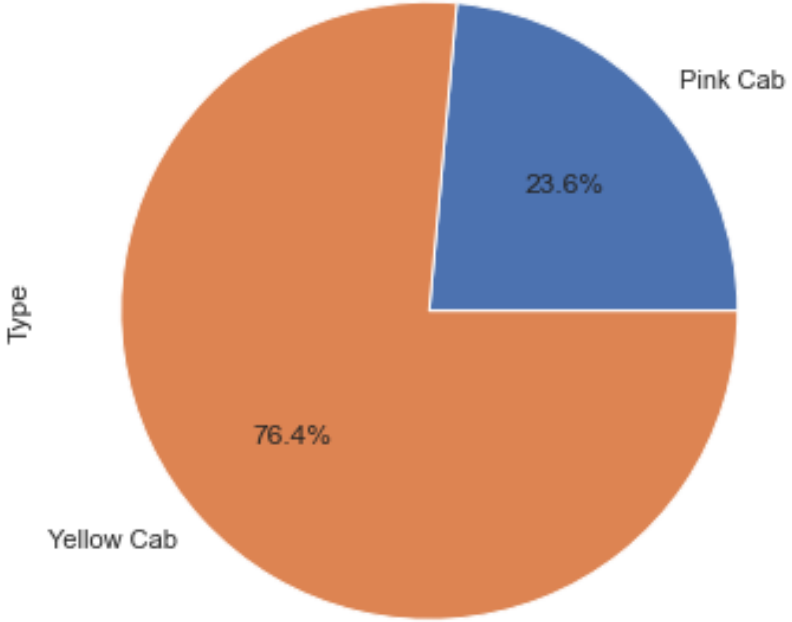


Fig 2

- Fig 1 shows that 55.2% of the customers uses yellow cab and 44.8% of the customer uses pink cab
- Fig 2 shows that 76.4% of the transaction goes to yellow cab and 23.6% goes to pink cab company

Gender Distribution of Customers

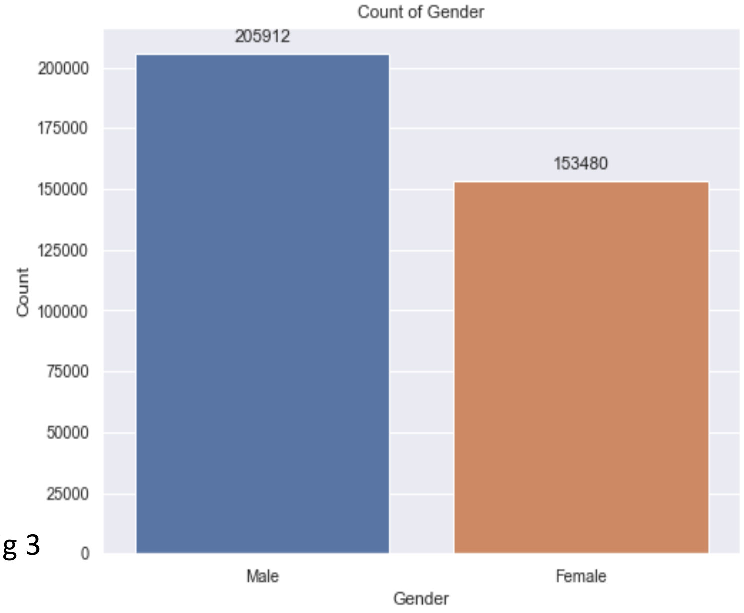


Fig 3

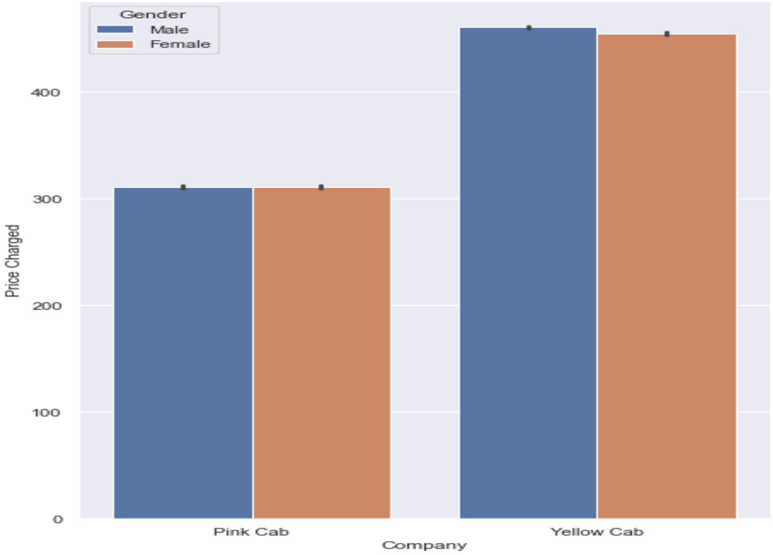


Fig 4

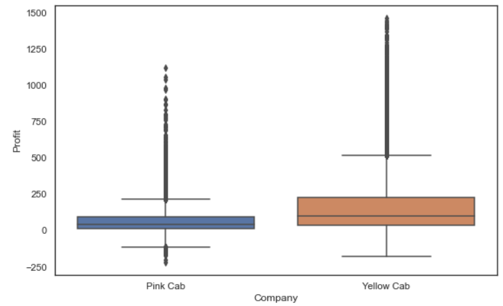
Fig 3: This shows that more transactions was recorded from male compare to female.

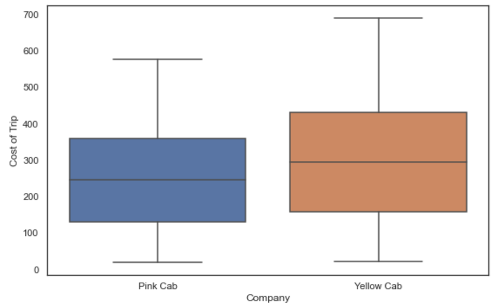
Fig 4: This shows that pink cab charged the same price for both gender and yellow cab charges maie a bit higher than male

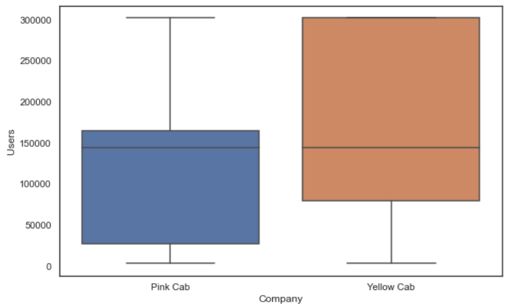
Company Distribution by Prices

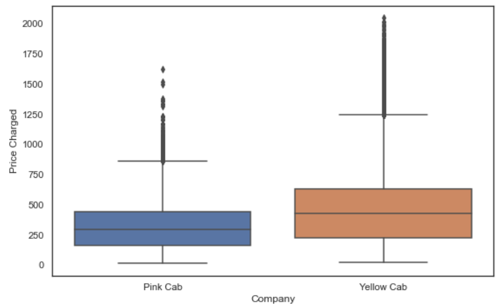
The box plots show that:

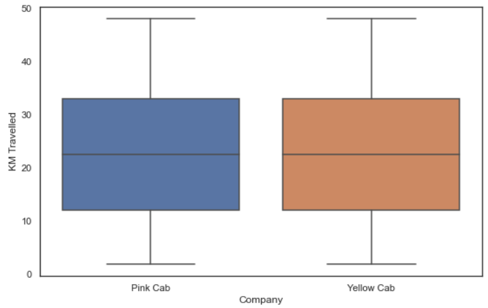
- Yellow cab company makes more profit than the blue cab
- The cost of the trip for the yellow cab is more than the pink cab
- Yellow cab company has a higher number of users compared to the pink cab.
- Yellow cabs charged higher than pink cabs in terms of price
- The two companies both cover the same KM

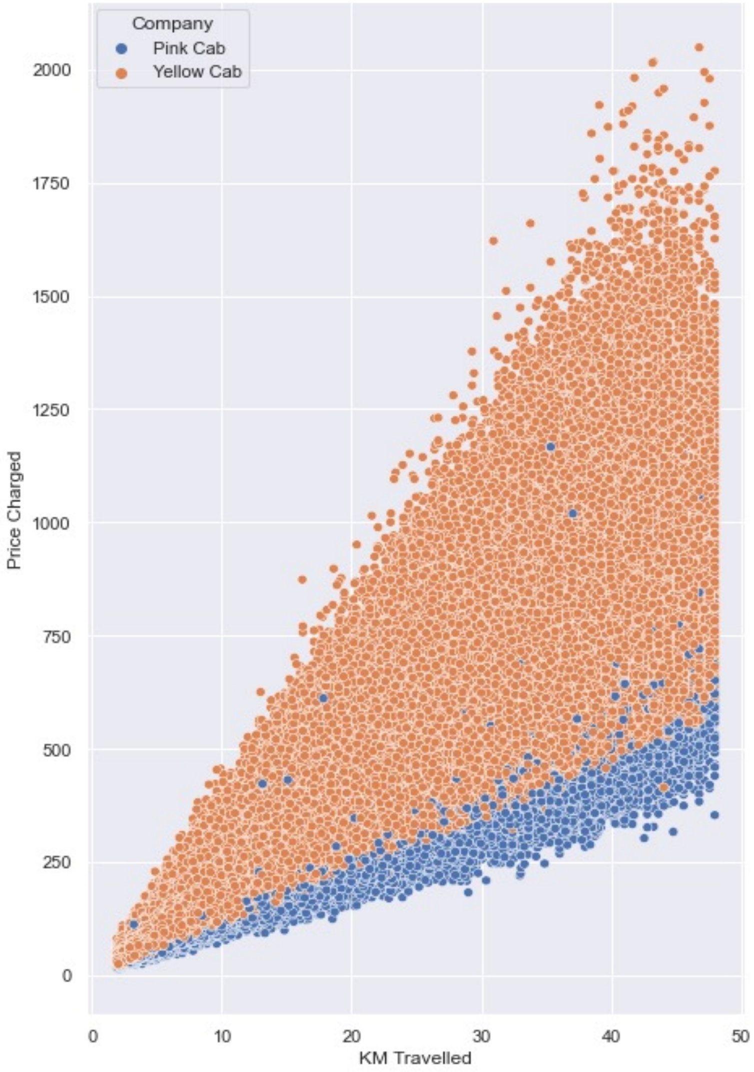


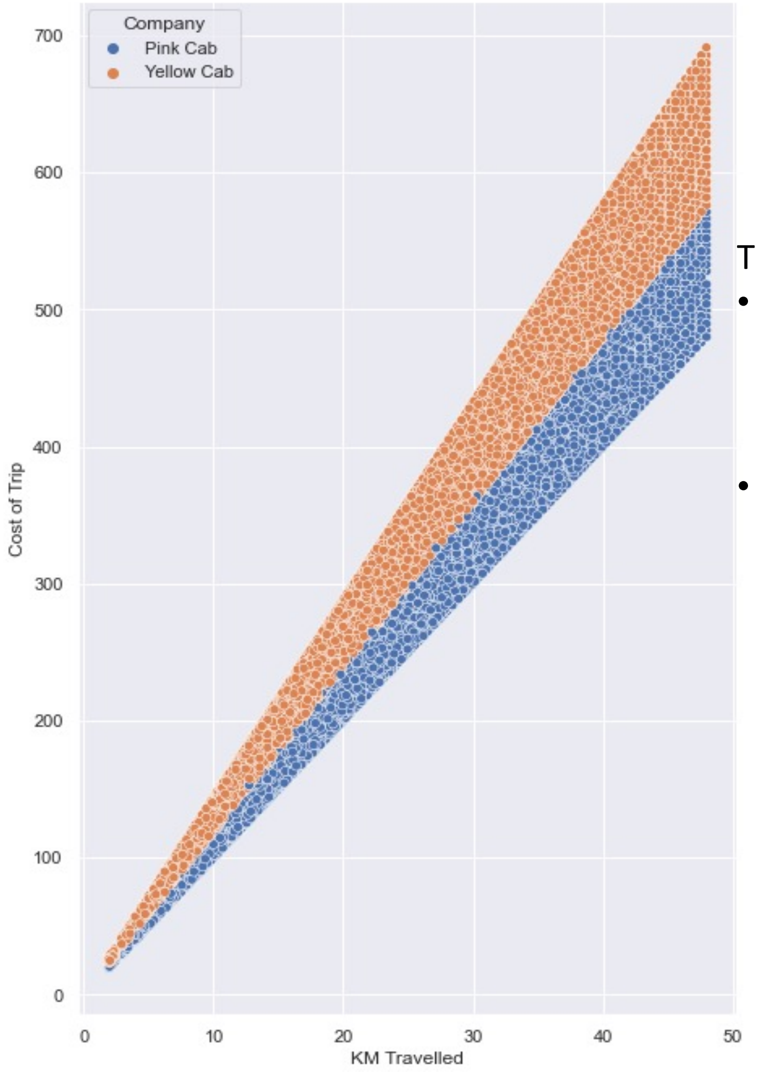






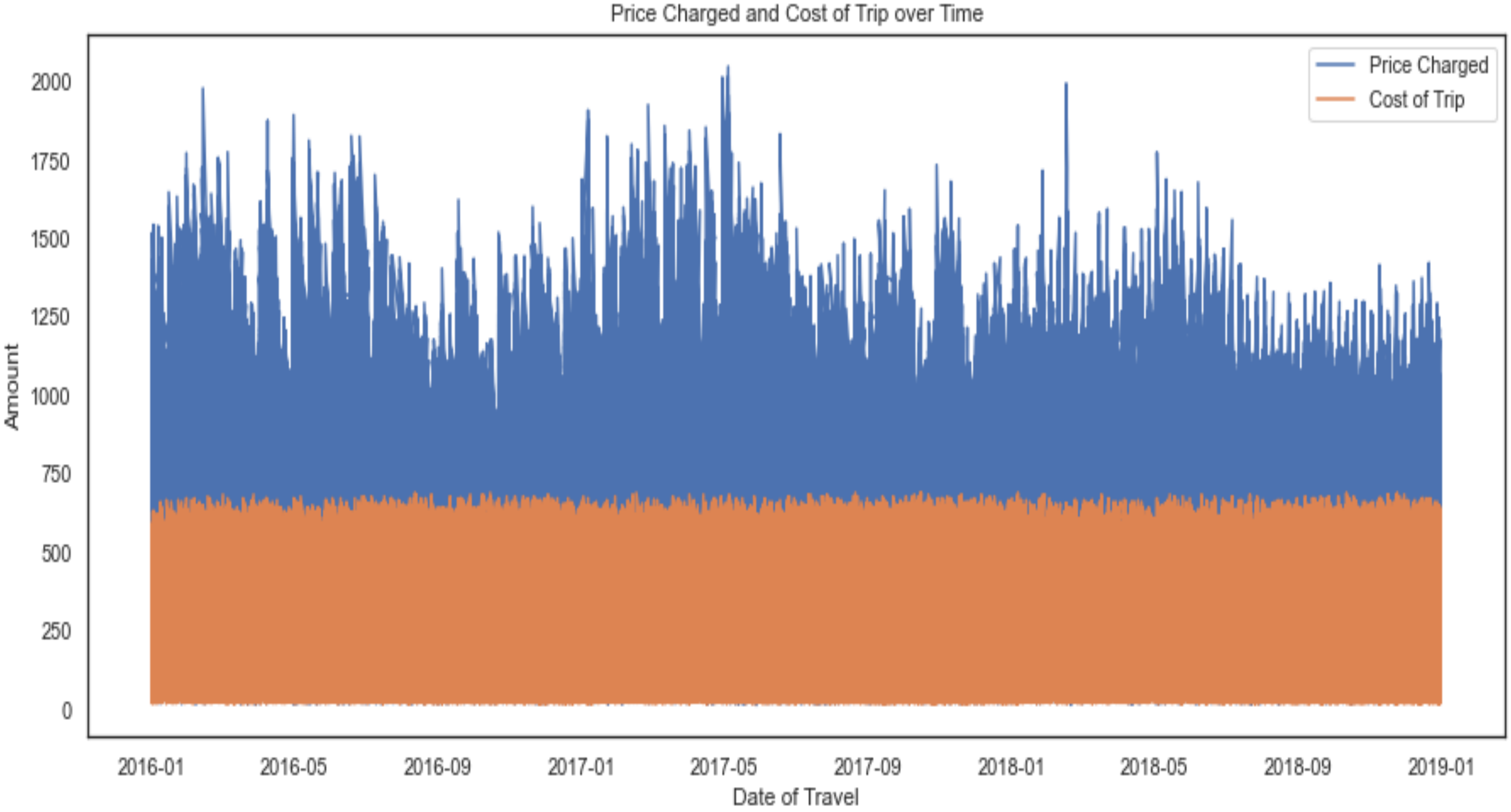




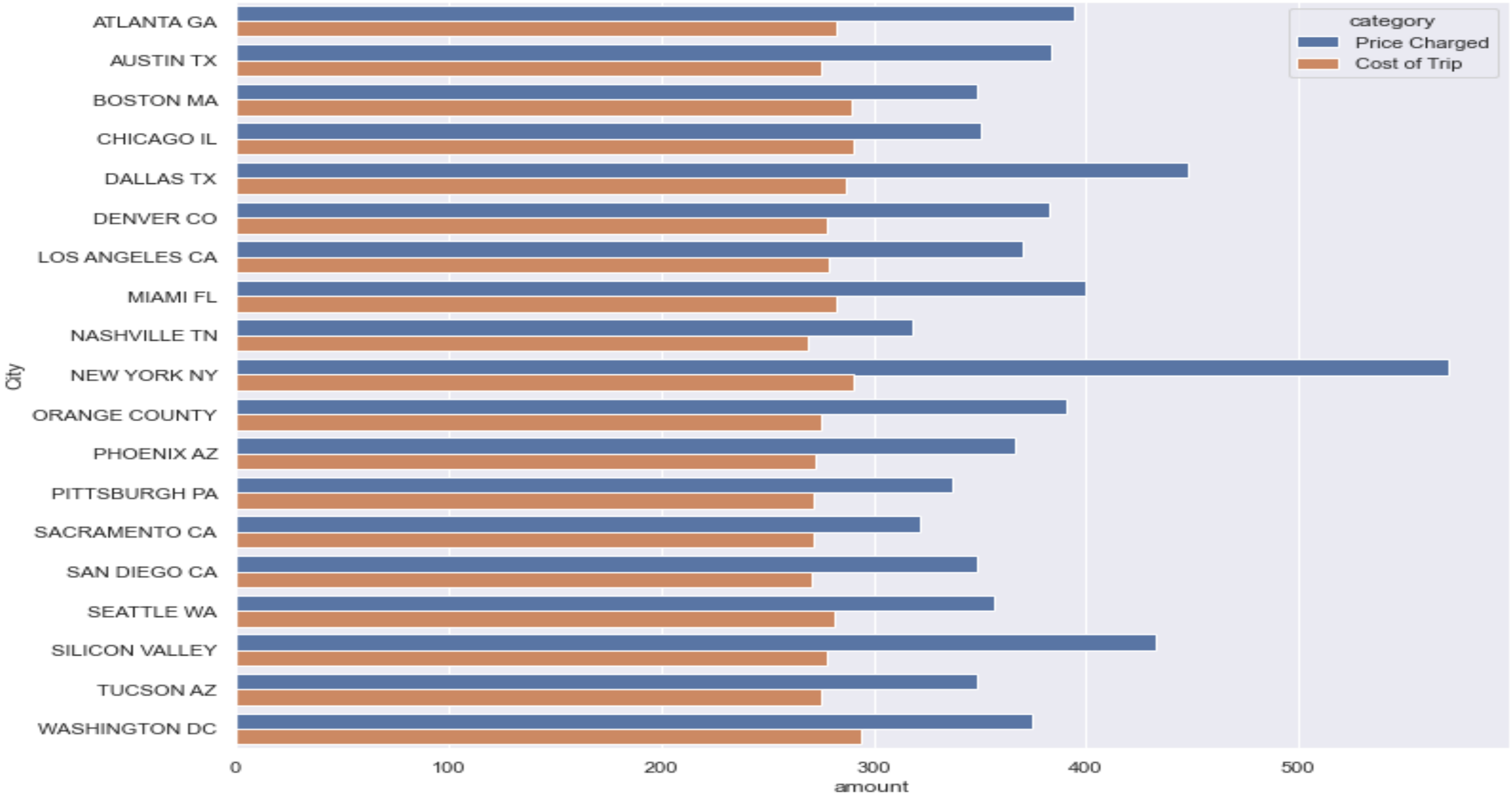


The scater plot show that:

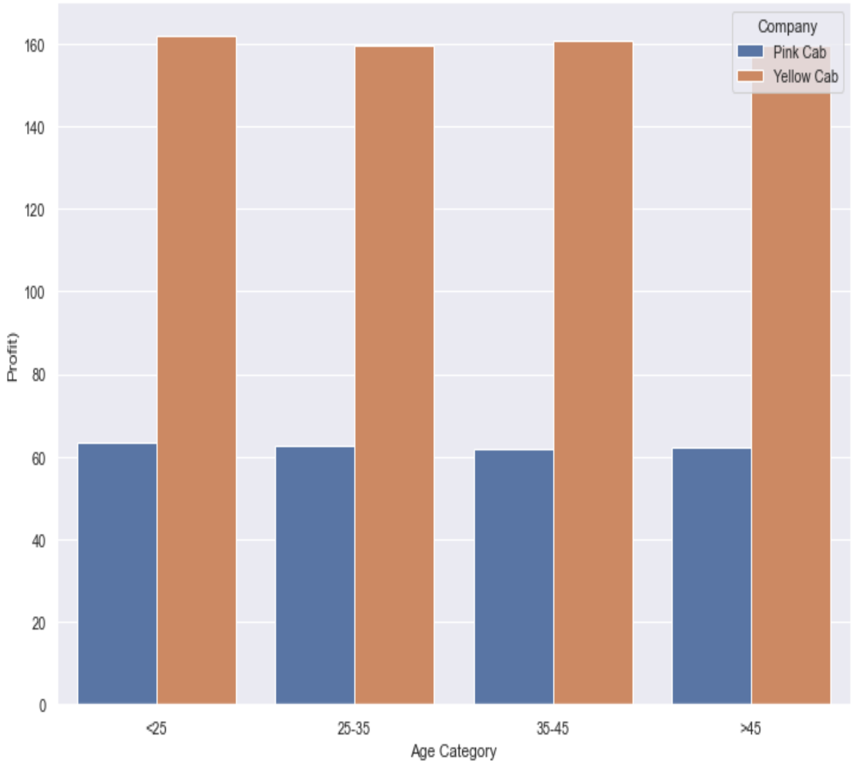
- The more the Km travelled the more the cost of trip and the price charged.
- Also the cost of trip and price charged by the yellow cab is mot than that of the pink cab.


Price Charged and Cost of Trip over Time
Price Charged
Cost of Trip
Amount
2000
1750
1500
1250
1000
750
500
250
0
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date of Travel

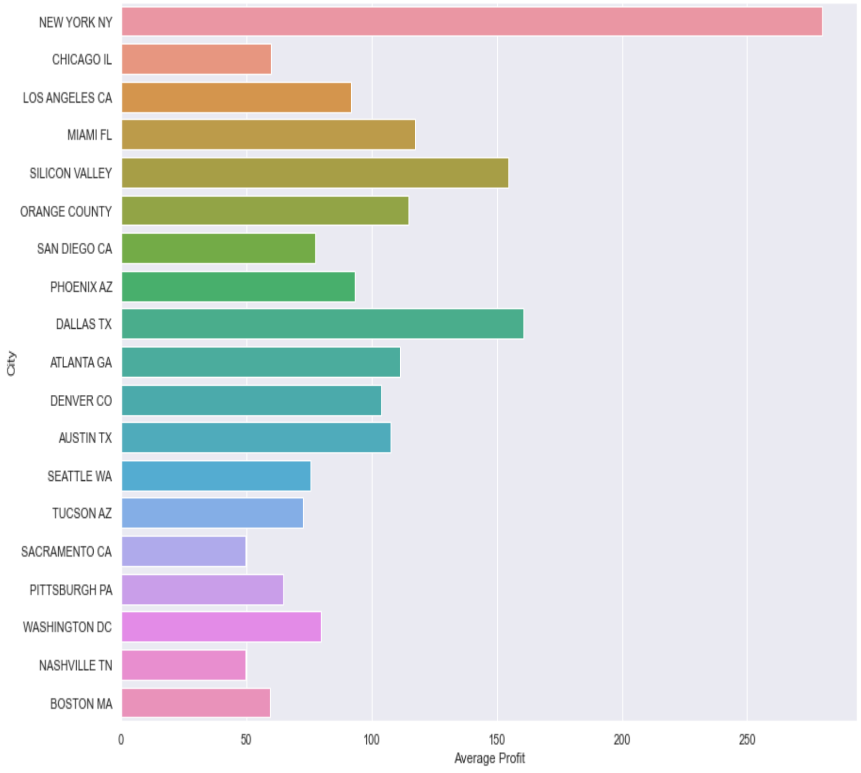
Distribution of cities by the price charged and cost of trip



Profit Analysis by Visualiziation

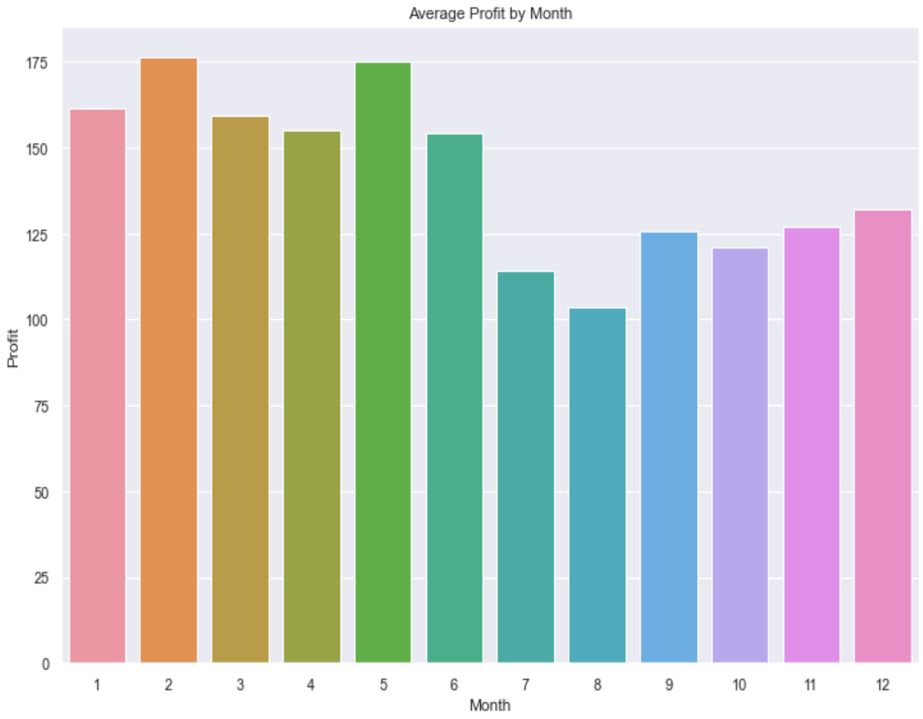


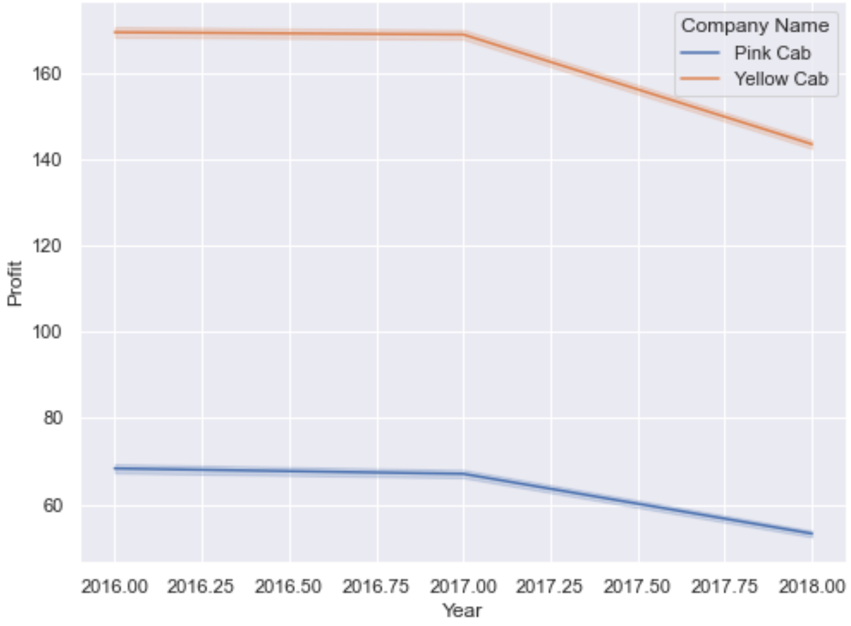
Profit by Age Category (Separated by Company Type)



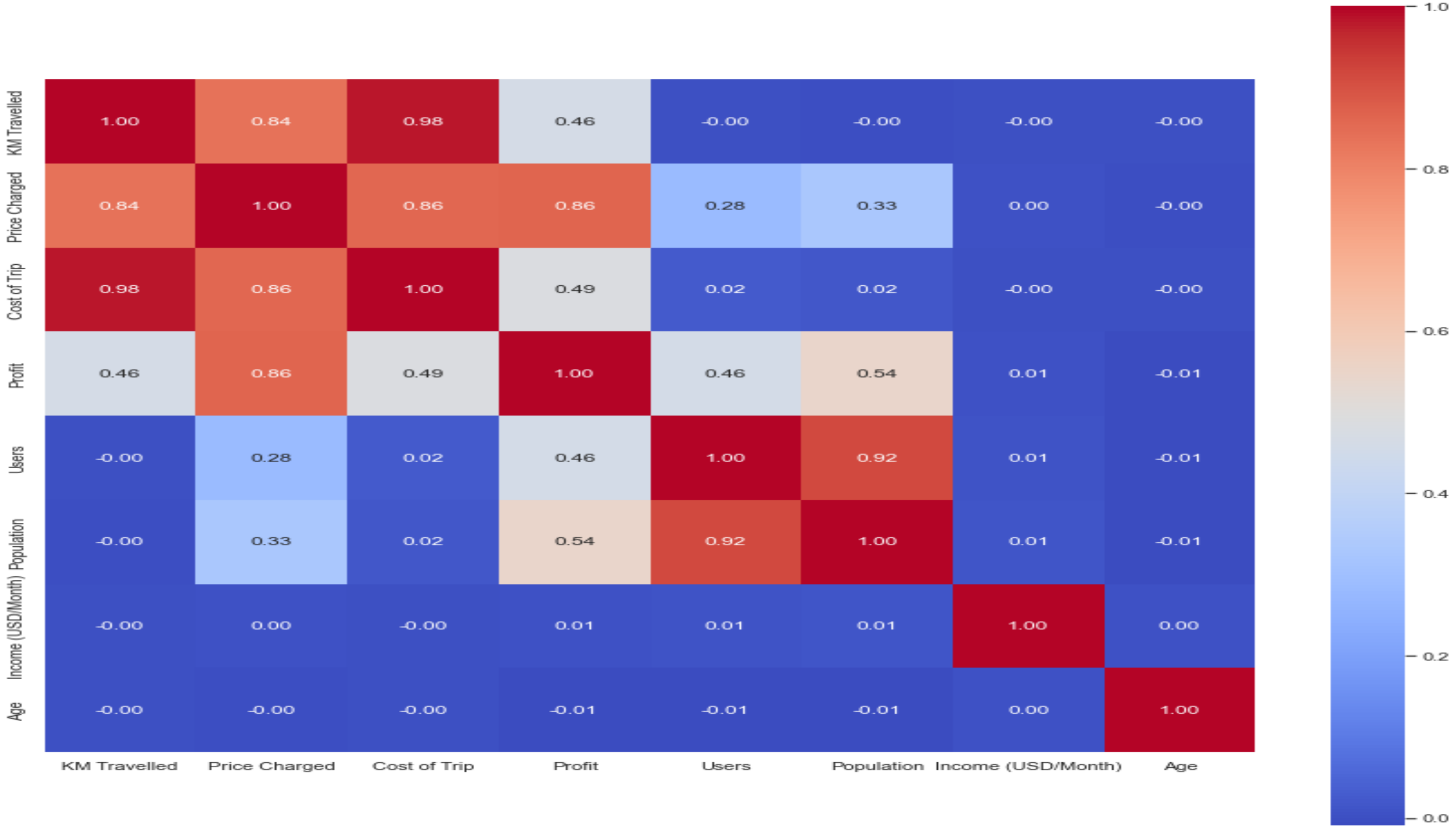
Average profit by city sorted by population in descending order

Average profit made by month and year





Pairwise correlation of the numeric variables



	KM Travelled	Price Charged	Cost of Trip	Profit	Users	Population	Income (USD/Month)	Age
KM Travelled	1.00	0.84	0.98	0.46	-0.00	-0.00	-0.00	-0.00
Price Charged	0.84	1.00	0.86	0.86	0.28	0.33	0.00	-0.00
Cost of Trip	0.98	0.86	1.00	0.49	0.02	0.02	-0.00	-0.00
Profit	0.46	0.86	0.49	1.00	0.46	0.54	0.01	-0.01
Users	-0.00	0.28	0.02	0.46	1.00	0.92	0.01	-0.01
Population	-0.00	0.33	0.02	0.54	0.92	1.00	0.01	-0.01
Income (USD/Month)	-0.00	0.00	-0.00	0.01	0.01	0.01	1.00	0.00
Age	-0.00	-0.00	-0.00	-0.01	-0.01	-0.01	0.00	1.00

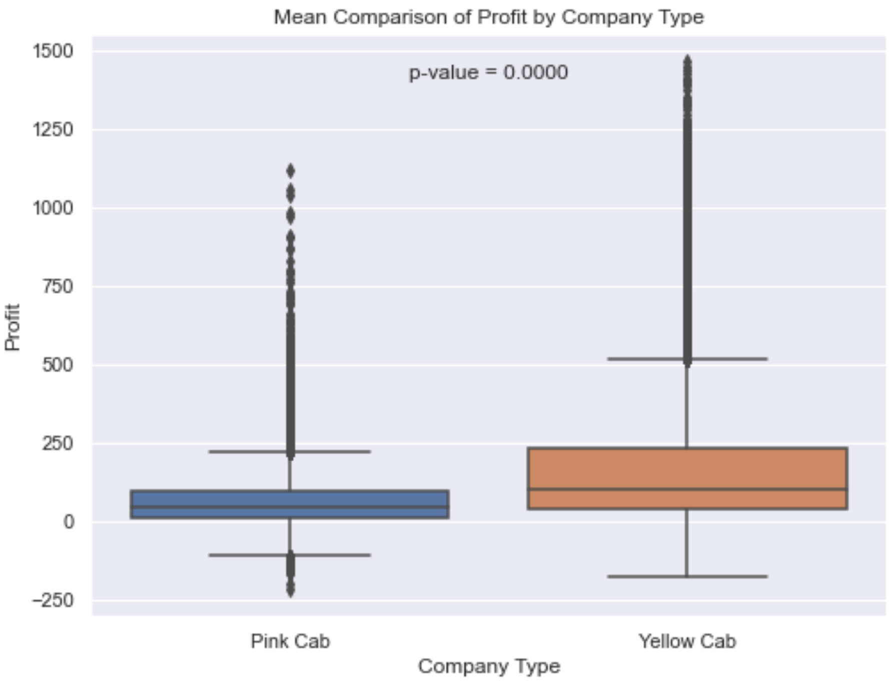
Insight from the heatmap

The above heatmap shows the relationship between each of the numerical variables in the overall merged dataset.

- ❖Checking for the variables that are highly correlated with the profit margin,
- ❖It was observed that the price charged is highly positively correlated with the value 0.86, which implies that the higher the price charged the more the profit.
- ❖Also the population of a city has a significant relationship with the profit of 0.54, which indicates that the higher the population of a city the more the profit .
- ❖The Users, Cost of trip, and KM travelled also have a positive relationship with the profit margin.
- ❖ In general the orange colour represents a significant positive relationship while the darker shade of blue indicates an insignificant relationship.

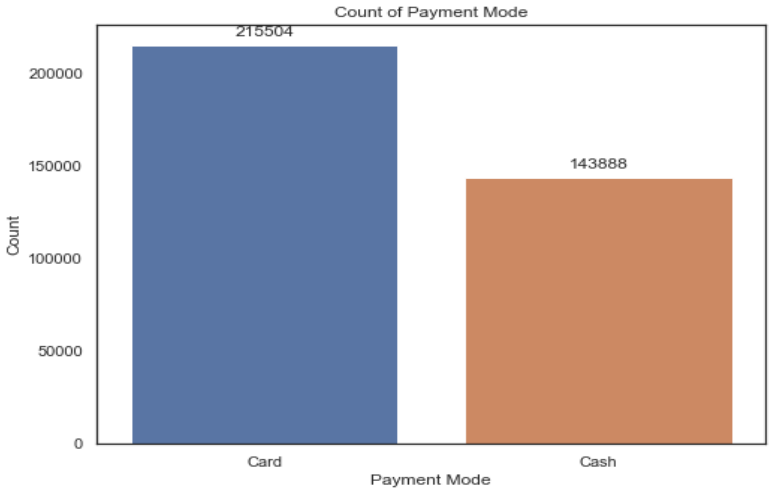
Hypothesis 1: Is there any significant difference between the profit made by the two companies?

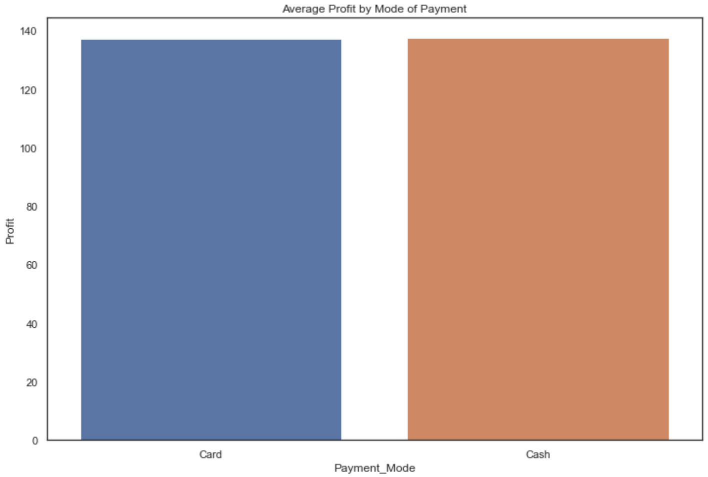
- The t-test statistic was used to compare the average profit of the two companies.
- The Test statistics is -160.3715175947807 and the p-value is 0.0 which is less than 0.05
- This shows there is a significant difference between the profit made by pink cab and yellow cab.
- The yellow cab makes more profit than the pink cab company.



Hypothesis 2: Is there any significant difference between methods of payment by revenue generated?

- T-test was also adopted to compare the revenue generated from the two method of payment.
- The Test statistics is -0.1358 and the p value is 0.8920 which is greater than 0.05.
- There is no significant difference between the revenue generated by card and cash method of payment.
- This was obvious in the bar chart of average profit by mode of payment





Hypotesis 3: Is there any relationship between cities' population and the profit made?

- Since it involves two numeric variables population and profit made in each city, a correlation analysis approach was adopted (Pearson coefficient).
- The correlation coefficient is 0.5440785836878613 and the p-value is 0.0 which is less than 0.05.
- Therefore, there is a significant relationship between the cities' population and the profit made.
- The correlation is positive, indicating a direct relationship between the two variables, this implies that a unit increase in population will lead to 0.544 increase in the profit.

Hypotesis 4: Is there any relationship between the kilometres covered by the two companies?

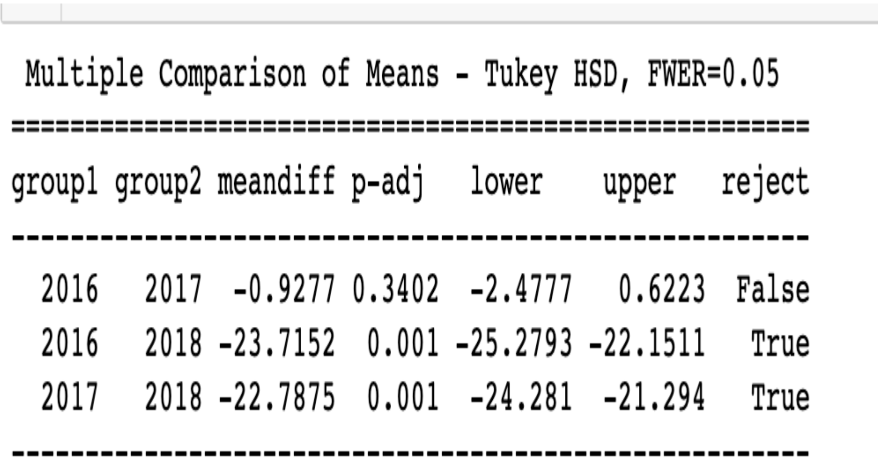
- The Test statistics is -0.19967531052842344 and the p value is 0.8417346372229664 which is greater than 0.05.
- There is no significant difference between KM travelled by pink cab and yellow cab.

Hypotesis 5: Is there any significant difference in the price charged by cities

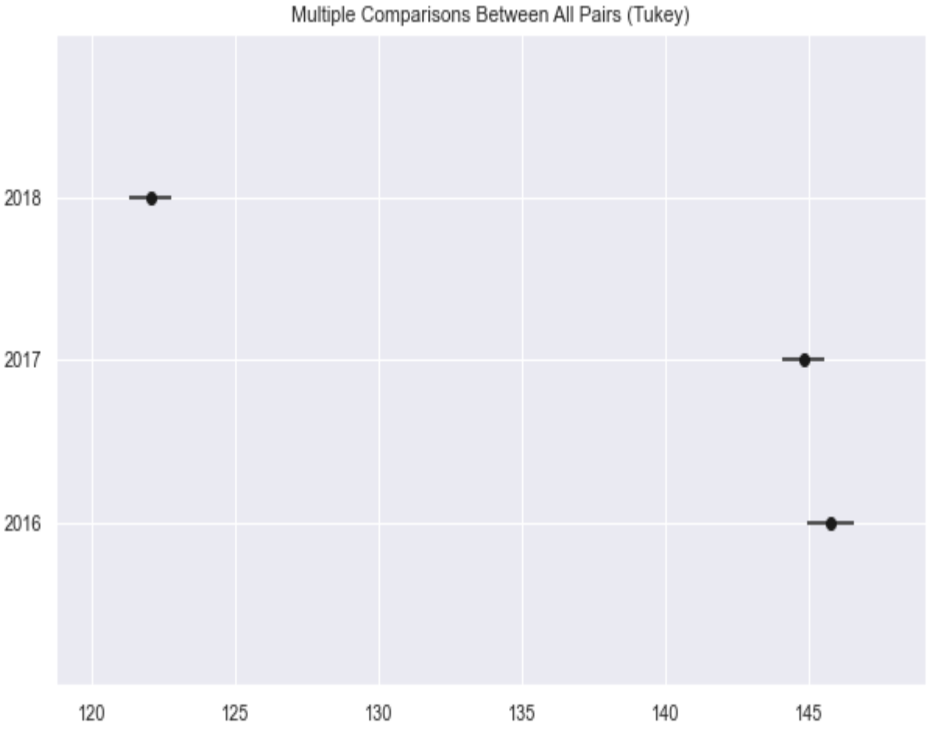
- ANOVA test was used as we have 19 groups to compare
- The Test statistics is 2610.2928727897047 and the p value is 0.0 which is less than 0.05
- There is a significant difference in the price charged by the 19 cities.

Hypotesis 6: Is there any significant difference in the profit generated across year?

- The Test statistics is 49413541.66574912 and the p value is 0.0 < 0.05
- Therefore, there is a significant difference in the profit generated across years.



```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
======================================================
group1 group2 meandiff p-adj  lower    upper   reject
------------------------------------------------------
  2016   2017  -0.9277 0.3402 -2.4777   0.6223  False
  2016   2018 -23.7152  0.001 -25.2793 -22.1511   True
  2017   2018 -22.7875  0.001  -24.281  -21.294   True
------------------------------------------------------
```



CONCLUSION

From the above visualization and hypothesis testing, the following conclusion was drawn:

Yellow cab has a higher number of customers and transactions compared to the pink cab

The distribution by Gender shows that male uses cab than female and most customers pay by card.

The profit made is mostly determined by the price charge and the yellow cab charged higher and make more profit than the pink cab

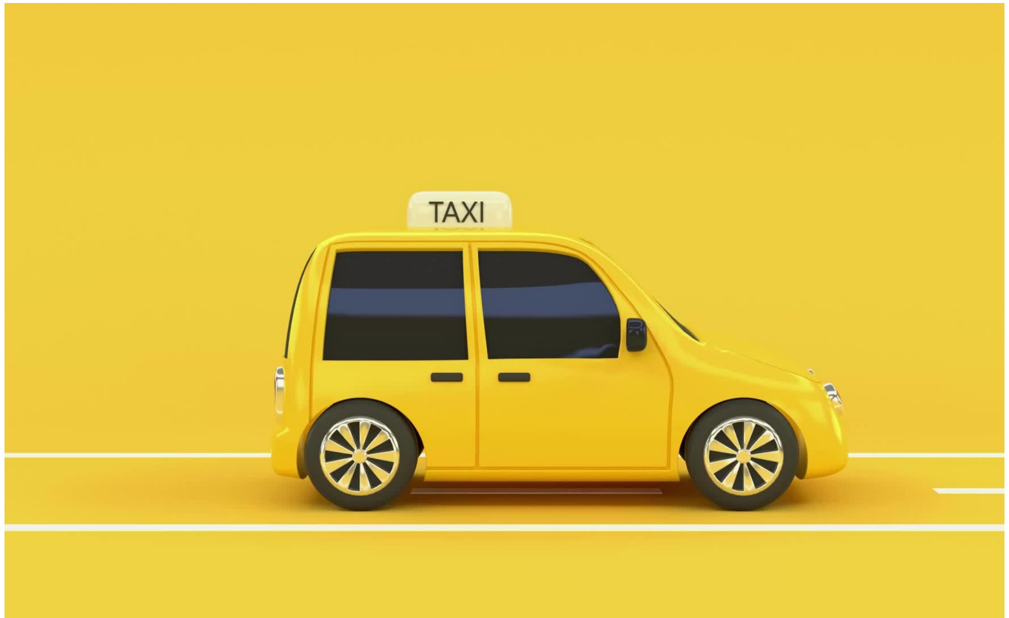
Most sales and profit was made in the first half of the year, the weather might be a contributing factor to this

There is a positive relationship between the profit made and the population of the cities, the higher the population the more the profit made.

The price charged by the two cab companies is determined by the location(cities)

There is a significant difference in the profit and revenue generated across the year, there was a great reduction in 2018 compared to 2016 and 2017.

RECOMENDATION



Based on the observations from the data visualization and hypothesis testing above , it is recommended that company XYZ invested in the yellow cab company.

Data Glacier

Your Deep Learning Partner

Thank You