Data Glacier

Your Deep Learning Partner

G2M Insight for Cab Investment Firm

Data Science Virtual Internship

Submitted by: Elissa Kuon

Date: January 14, 2023

Outline

INTRODUCTION

DATA INFORMATION

EXPLORATORY DATA ANALYSIS

MODEL BUILDING

FINAL STATEMENT

INTRODUCTION

Problem Statement

- XYZ is a private equity firm in the US. Due to remarkable growth in the Cab Industry in the last few years and multiple key players in the market, it is planning for an investment in the Cab Industry.
- Objective: Provide actionable insights to help XYZ firm in identifying the right company for making an investment

- The analysis was divided into three parts:
  - Data Understanding
  - Visualization & Model Building
  - Recommendations for Investment

DATA INFORMATION

Data Background

Investigate and analyze four datasets to achieve insights within the cab industry:

- Cab_Data.csv – includes details of transactions between the cab companies
- Customer_ID.csv – a mapping table containing a unique identifier linking the customer's demographic details
- Transaction_ID.csv – a mapping table containing transaction to customer mapping and payment mode
- City.csv – contains a list of US cities, along with their respective population, and the number of cab users

We proceed to combine all four datasets into one master data to work with

Data Background

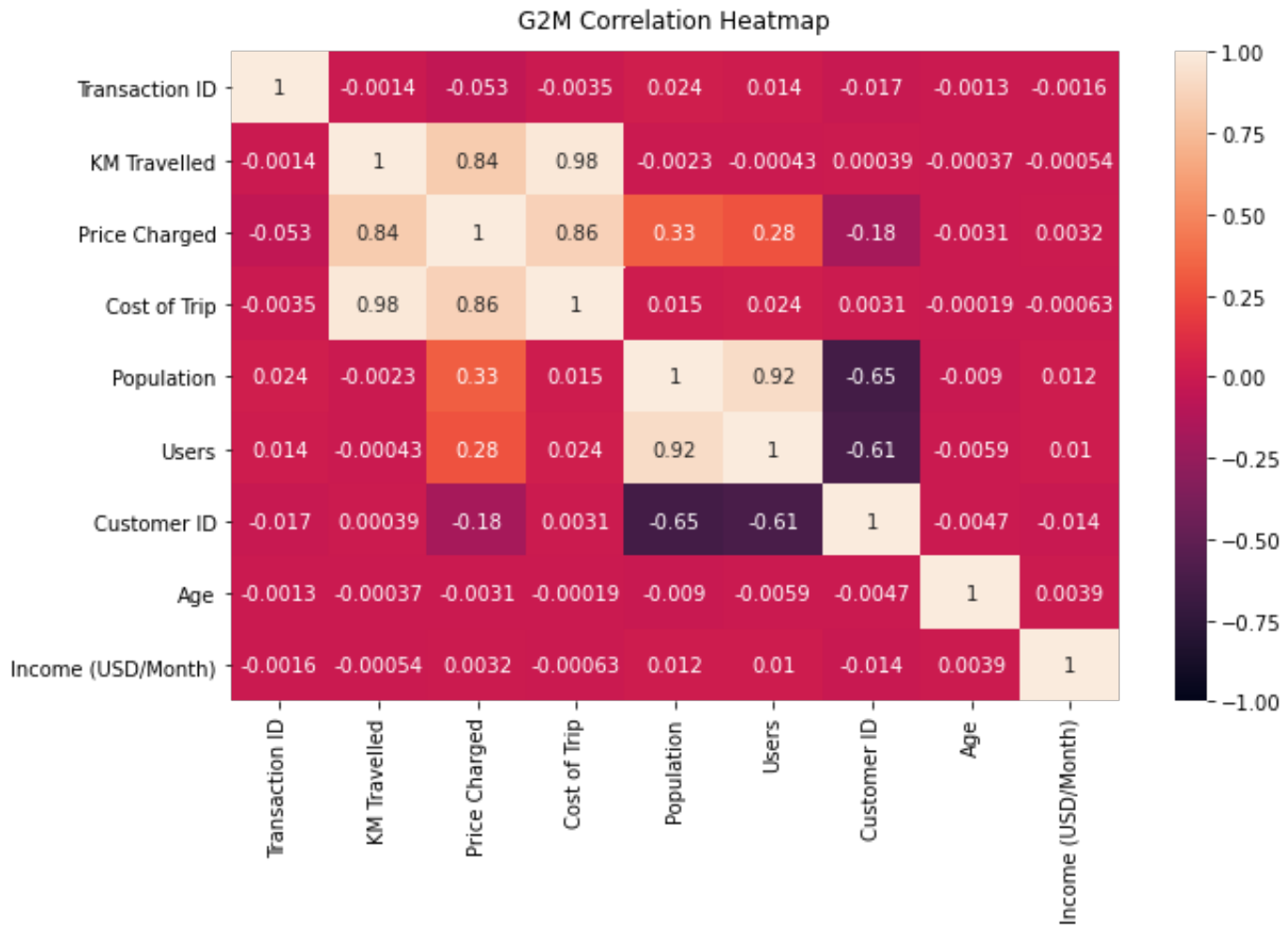
- The timeframe of our dataset is from January 1st, 2016, till December 31st, 2018
- 18 features with 359,392 observations

- **Assumptions:**
  - ❖ Users are treated as the number of cab users in the city
  - ❖ Profit Margin is calculated by subtracting Cost_of_Trip from Price_Charged
  - ❖ Take into consideration the assumptions that are followed with model building and hypothesis testing
    - ➢ May affect the results we've obtained and may need further investigation, however, can be used as a baseline for understanding the Cab Industry

EXPLORATORY DATA ANALYSIS

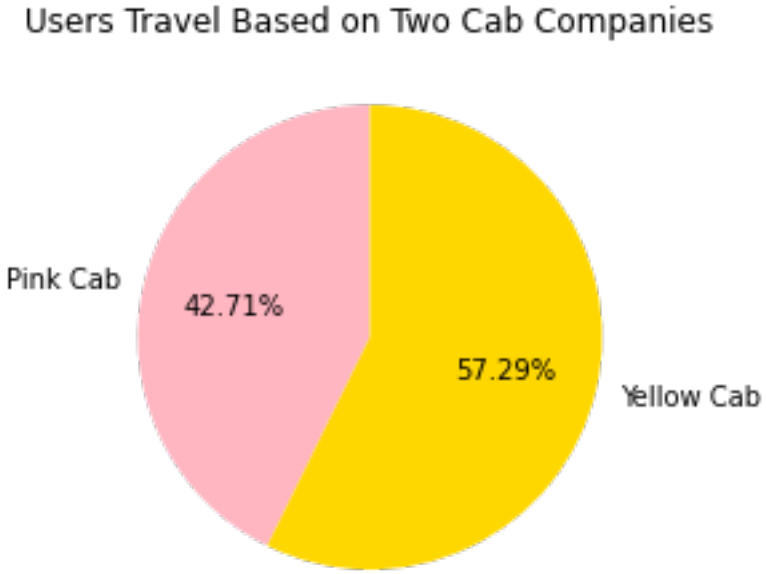
Correlation Between Variables

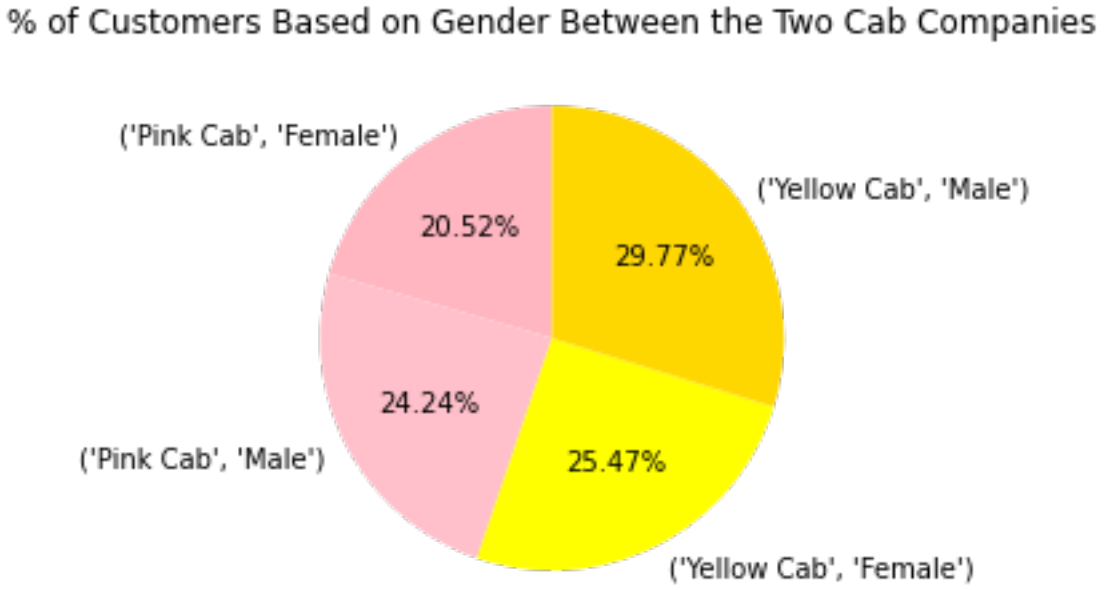
- There is a strong correlation between the variables:
  - Population and Users
  - Price_Charged and KM Travelled
  - KM Travelled and Cost_of_Trip



% of Cab Users and Gender Distribution

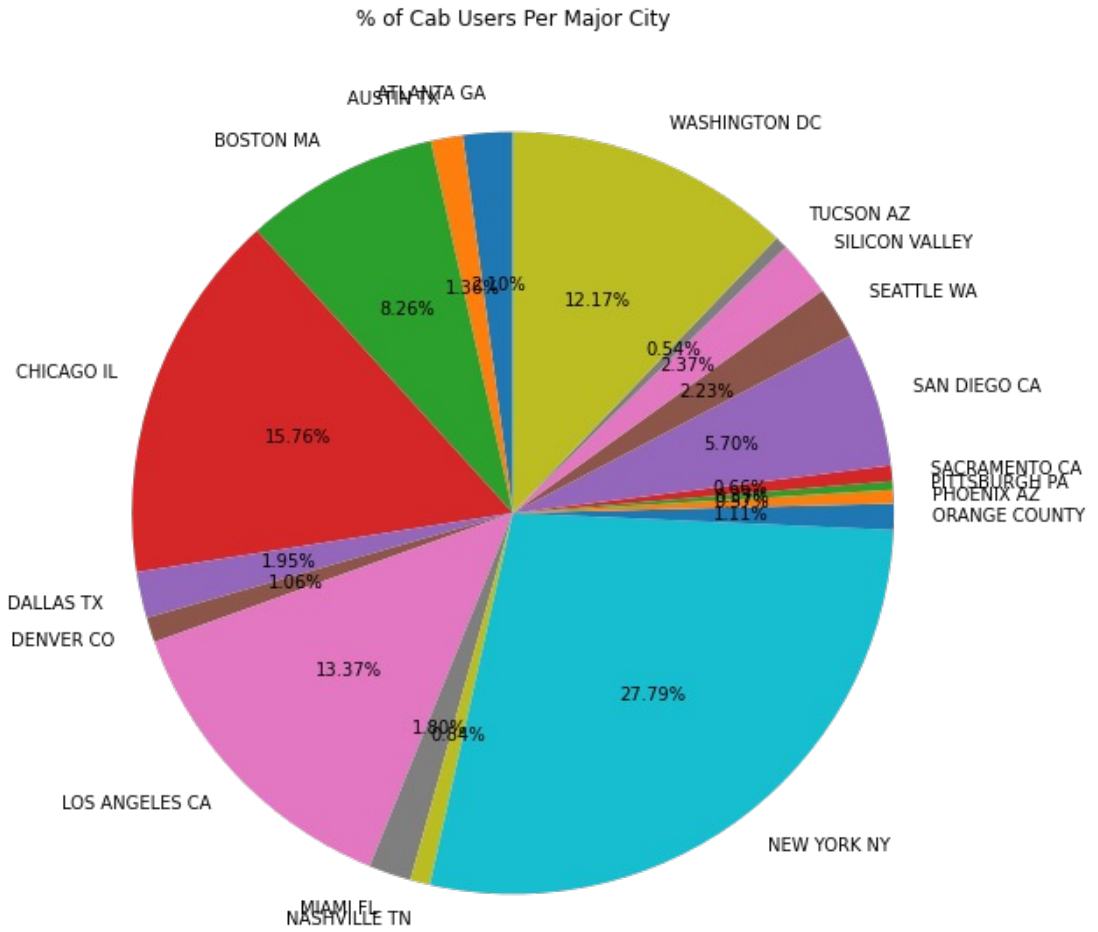
- Users prefer to take Yellow Cab than Pink Cab, with a difference of 14.58%
- Pretty even gender distribution overall, with slightly more male users by 0.29%





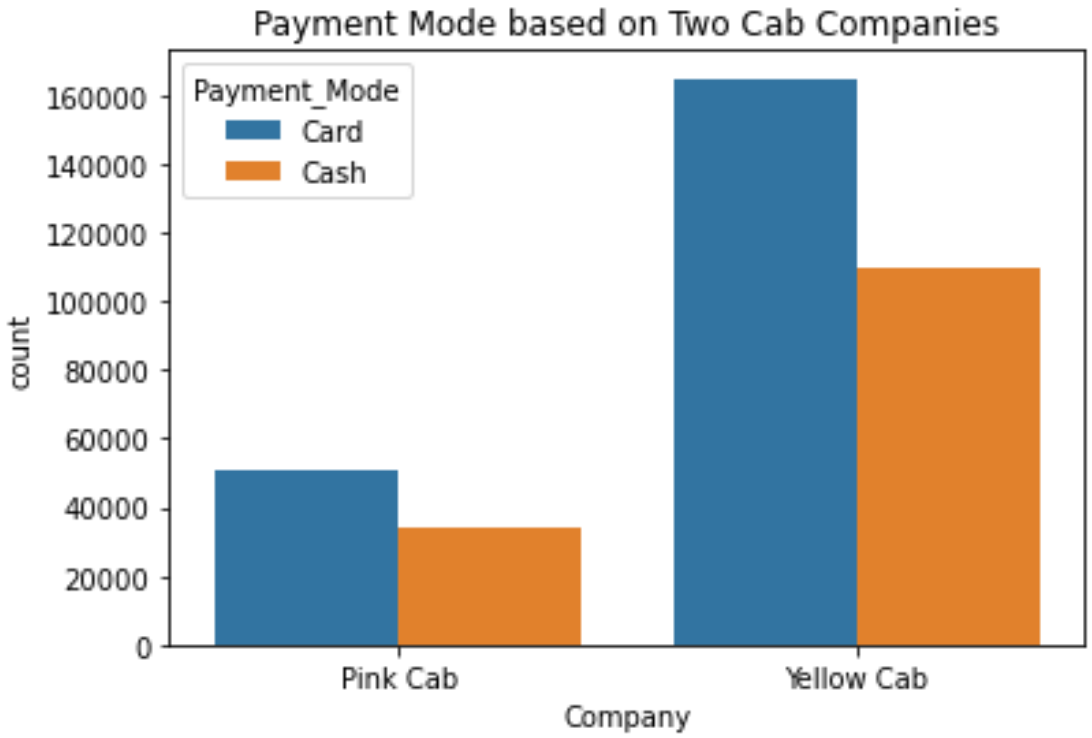
% of Cab Users Within US Major Cities

- Most cab users reside in New York City with 27.79% of all cab users, followed by Chicago and Los Angeles



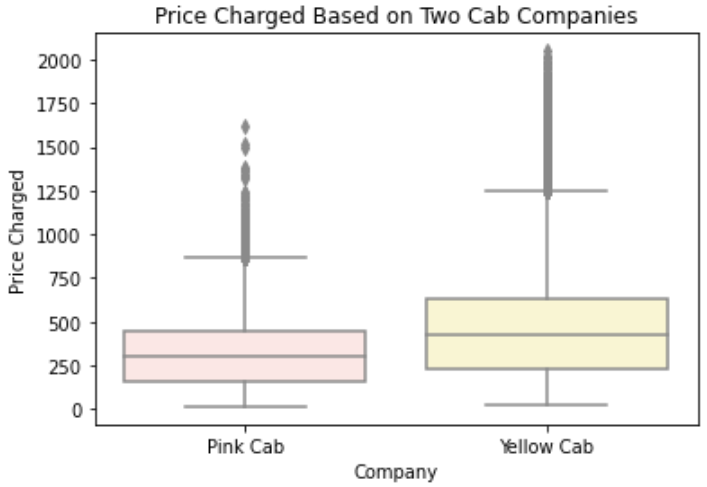
Payment Mode Distribution

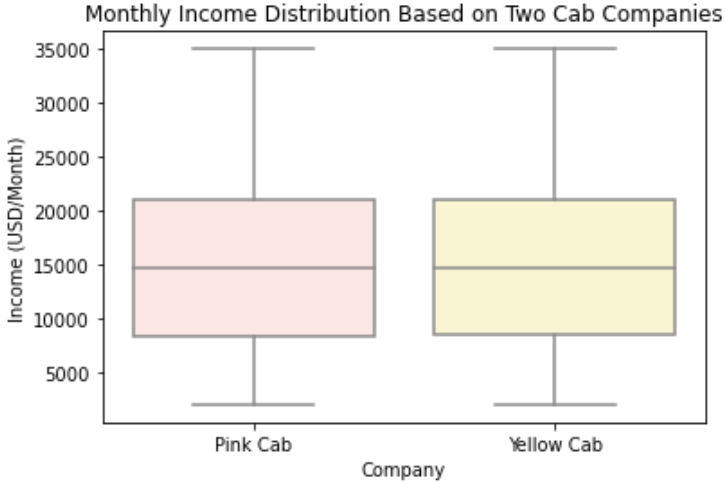
- Most cab users prefer to pay their rides through card payments than with cash for both companies

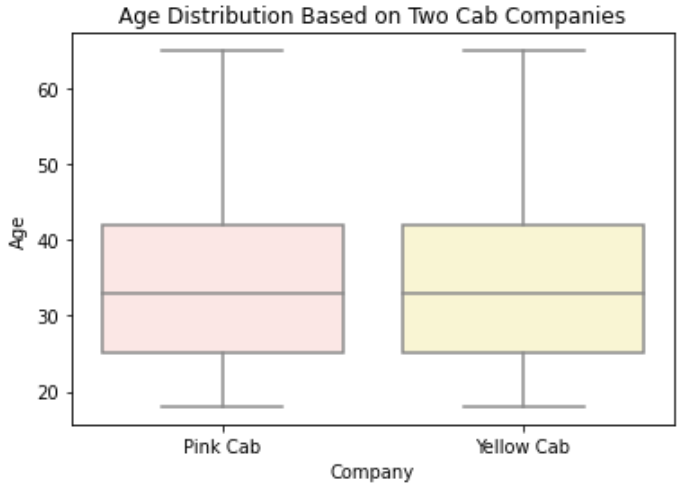


Price Charged, Income, and Age Distribution

- Yellow Cab charged slightly higher for their cab services
- For both companies, the average age of cab users is around 33 years old with an average monthly income of 15K USD

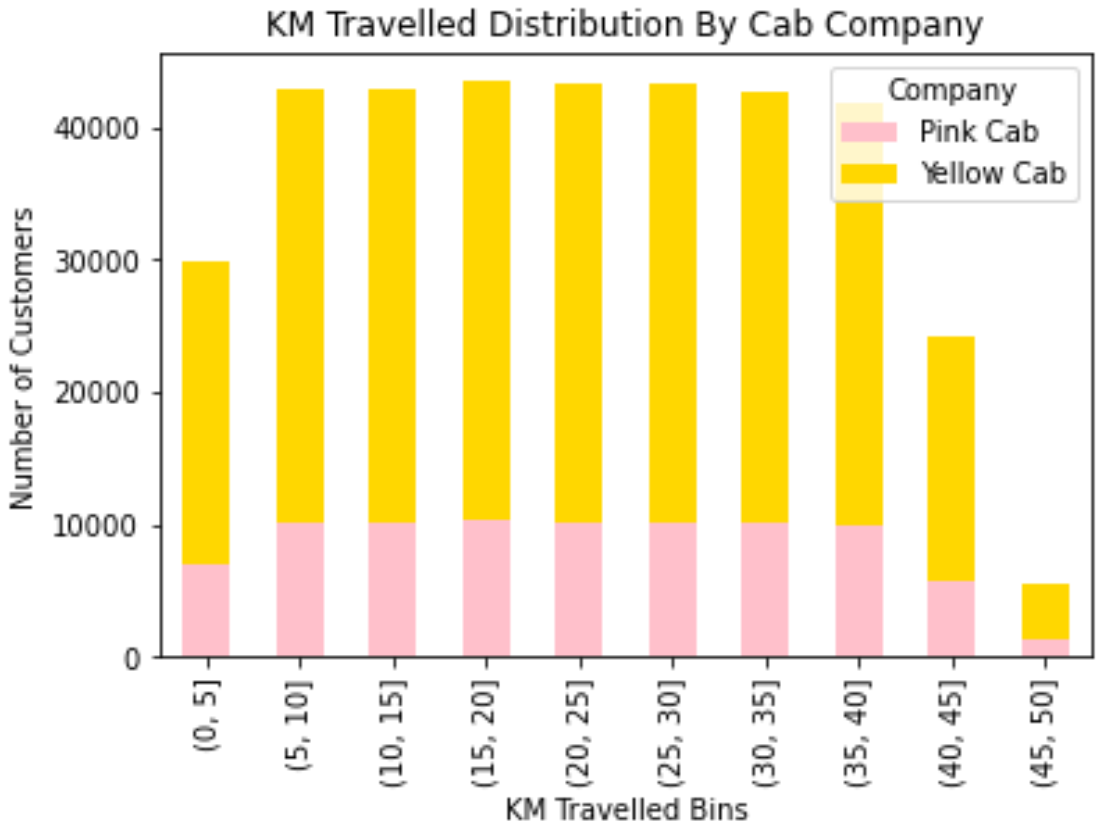






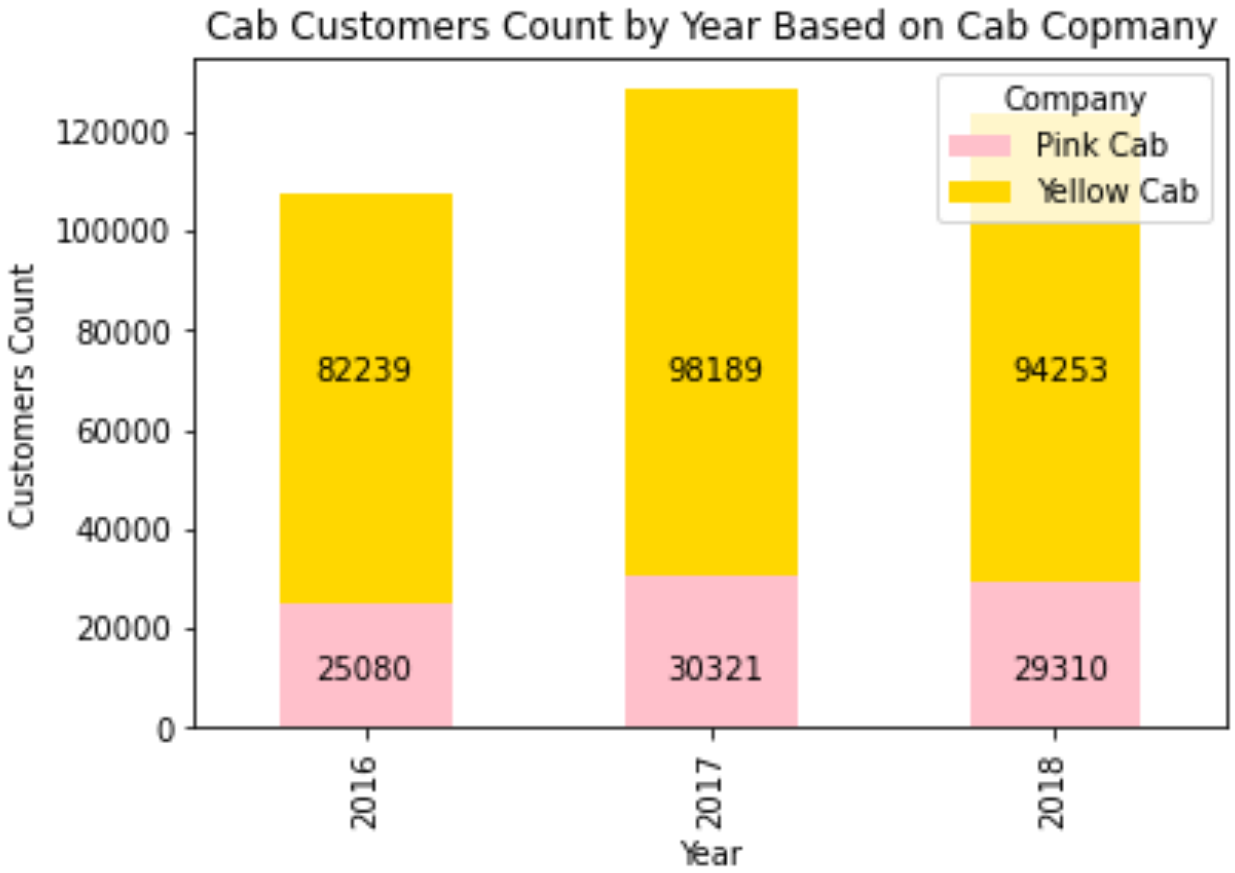
KM Travelled Distribution

- Many cab users mostly travel between 5 – 40 km
- Yellow Cab has a uniform shape in terms of distance traveled compared to Pink Cab
  - **Note**: A big difference between the two companies in short and long distances, where Yellow Cab has an overwhelmingly better coverage



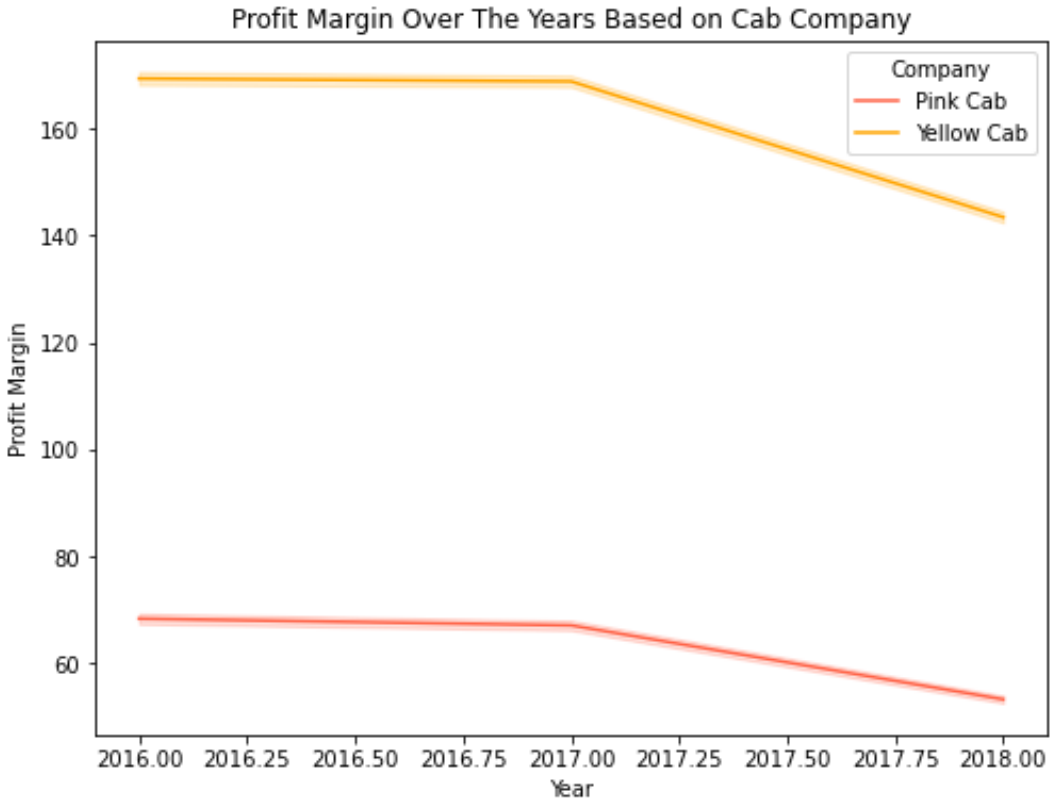
Users Count by Year

- Yellow Cab has a higher customer count compared to Pink Cab
- In the transition from 2017 to 2018, Yellow Cab suffered a 4.18% decrease in customers, while Pink Cab suffered a 3.45% decrease in customers



Profit Margin

- Yellow Cab has a much higher profit margin in comparison to Pink Cab
- Profit Margin for both cab companies decrease over the span of 2 years and may continue to decrease looking at the trend of the graph



MODEL BUILDING

Model Building Objective

- Profit Margin will be treated as our targeted feature
- Not interested in wanting to predict profit margin due to the complexity of the formulas of the models
- More interested in the relationship between profit margin and the other features

Linear Regression

- Important features in inferencing Profit Margin is Company, KM Travelled, Price Charged, Cost of Trip, Population, and Users
- Dropped Gender, Payment Mode, Age, and Income as they were deemed to not have any influence on Profit Margin

OLS Regression Results

Dep. Variable:	PM	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	1.222e+32
Date:	Tue, 17 Jan 2023	Prob (F-statistic):	0.00
Time:	18:19:48	Log-Likelihood:	8.9651e+06
No. Observations:	359392	AIC:	-1.793e+07
Df Residuals:	359385	BIC:	-1.793e+07
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.354e-12	2.26e-14	-236.394	0.000	-5.4e-12	-5.31e-12
Company_Yellow Cab	6.487e-12	2.04e-14	317.724	0.000	6.45e-12	6.53e-12
KM Travelled	9.726e-14	3.59e-15	27.125	0.000	9.02e-14	1.04e-13
Price Charged	1.0000	5.6e-17	1.79e+16	0.000	1.000	1.000
Cost of Trip	-1.0000	2.82e-16	-3.55e+15	0.000	-1.000	-1.000
Population	6.81e-19	4.98e-21	136.711	0.000	6.71e-19	6.91e-19
Users	-2.864e-17	1.51e-19	-190.160	0.000	-2.89e-17	-2.83e-17

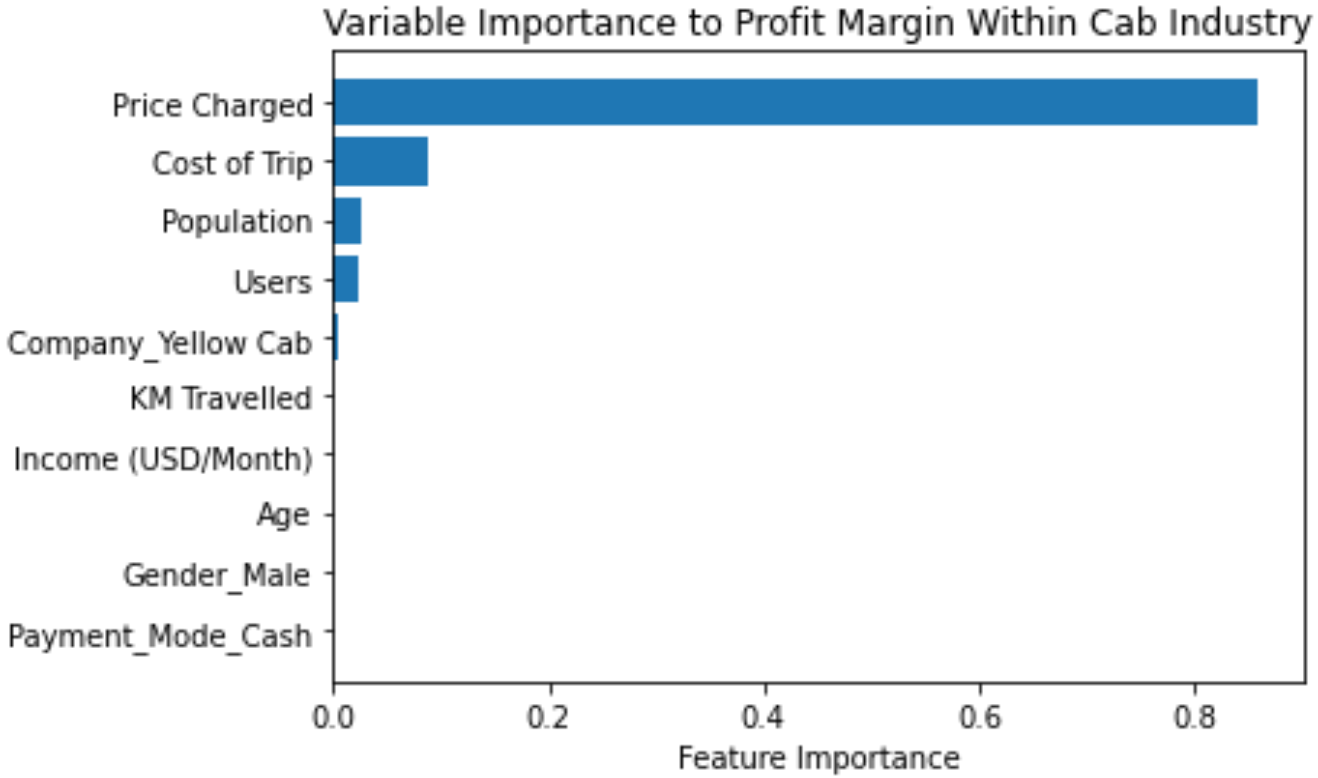
Omnibus:	28554.639	Durbin-Watson:	0.914
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36041.455
Skew:	0.776	Prob(JB):	0.00
Kurtosis:	2.995	Cond. No.	2.14e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.14e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Random Forest

- Price Charged is the most important feature in influencing Profit Margin
- Cost of Trip, Users, Population, and Company have also some or little importance
- Age, Gender, Income, and Payment Mode have no importance to Profit Margin

NOTE: This is in line with the variable importance of the linear regression model and hypothesis testing



FINAL STATEMENT

Recommendations

We have evaluated both cab companies and conclude Yellow Cab is better than Pink Cab:

- Customer Reachability: Users mostly travel between 5 to 40 km. Overall, Yellow Cab has much better coverage covering short and long distances compared to Pink Cab. Yellow Cab customers consist of about three times more customers than Pink Cab customers
- Expected Profit Margin and Users Count: Through statistical testing, there is strong evidence that the expected profit margin of Pink Cab is different than the expected profit margin of Yellow Cab. From the visualization, Yellow Cab has a much higher profit margin in comparison to Pink Cab. Based on the trend of the graphs between the user counts and profit margins within the span of 2 years, we are expected to see a steady decline. To get a better forecast of profit margin, we may investigate the feature importance from model building.
- Income-Wise: Through statistical testing, most Pink Cab and Yellow Cab users are of the mid-income class with the low-income class being the least number of cab users. Most of the users utilize Yellow Cab by a great amount in the low and high-income classes. Yellow Cab company has the highest number of users from the mid-income class compared to Pink Cab, which greatly covers all three income classes.

Based on the above points, we recommend Yellow Cab for investment. However, we must consider the assumptions we have made initially and may need to do further analysis for a more accurate insight.

THANK YOU