Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

May 21st 2023

Agenda

Executive Summary

Problem Statement

Data Details

Exploratory Data Analysis Approach

Exploratory Data Analysis Summary

Hypothesis

Recommendations

References

Executive Summary

Company: XYZ

Interested in: Cab investment

Strategy: Go to Market (G2M)

Location: US

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 Details

There are 4 datasets.

Time period of data is from 31/01/2016 to 31/12/2018.

Below are the list of datasets which are provided for the analysis:

- Cab_Data.csv – this file includes details of transaction for 2 cab companies (Pink Cab and Yellow Cab)
- Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details
- Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode
- City.csv – this file contains list of US cities, their population and number of cab users

Note: US Holidays dataset has been downloaded to analyse the days off during the year. US Holiday Dates (2004-2021).csv - this file contains list of holidays, with their date, weekday, day, month and year [1].

Exploratory Data Analysis Approach

1. Understanding the Data
2. Data Cleaning
3. Master Data
4. Outlier Detection
5. Data Visualization

Understandi-ng Data

The Cab_Data dataset has 359392 records and 7 columns as shown below.

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.854

The City dataset has 20 records and 3 columns as shown below.

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468

The Customer_ID dataset has 49171 records and 4 columns as shown below.

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237

The Transaction_ID dataset has 440098 records and 3 columns as shown below.

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card

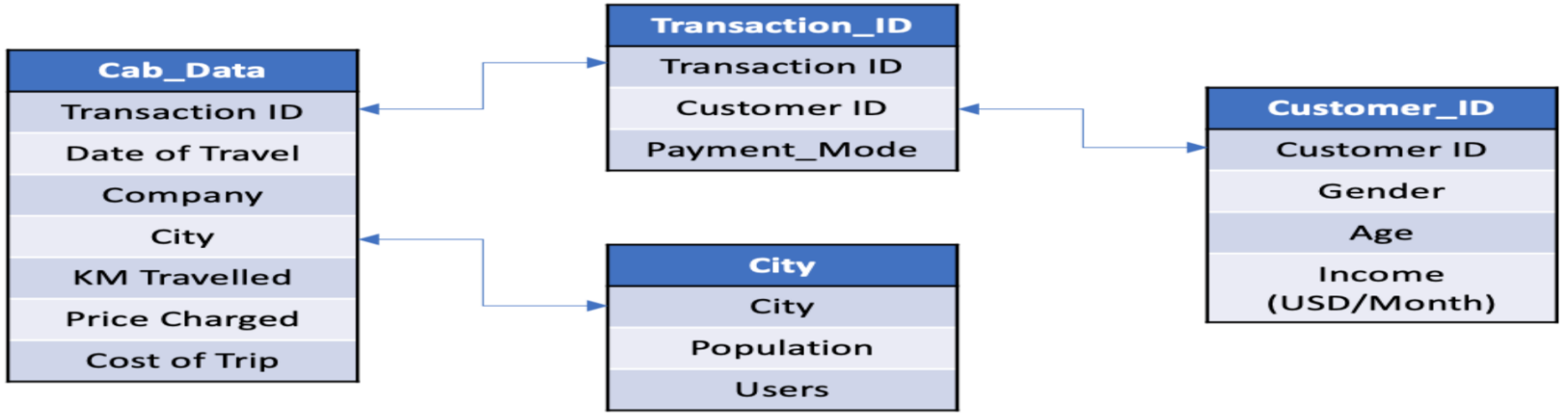
Note: US Holiday data was downloaded to understand the days off during the year 2016 to 2018. The Holidays dataset has 57 records and 6 columns as shown below.

	Date	Holiday	WeekDay	Month	Day	Year
12	2016-07-04	4th of July	Monday	7	4	2016
13	2017-07-04	4th of July	Tuesday	7	4	2017

Data Cleaning

- There are no missing values in any of the datasets.
- Date of Travel variable was converted from int64 to datetime datatype in Cab_Data dataset.
- Population and Users variables were converted to int64 datatype in City dataset.
- Datatypes of the variables on Customer_ID and Transaction_ID datasets looks fine.
- There were no duplicate data in any of the datasets.
- Year and Month were extracted from Date of Travel variable in Cab_Data dataset

Master Data



1. Transaction ID variable is the common key on Cab_Data and Transaction_ID datasets. Hence, Cab_Data and Transaction_ID datasets can be combined by using Transaction ID variable.

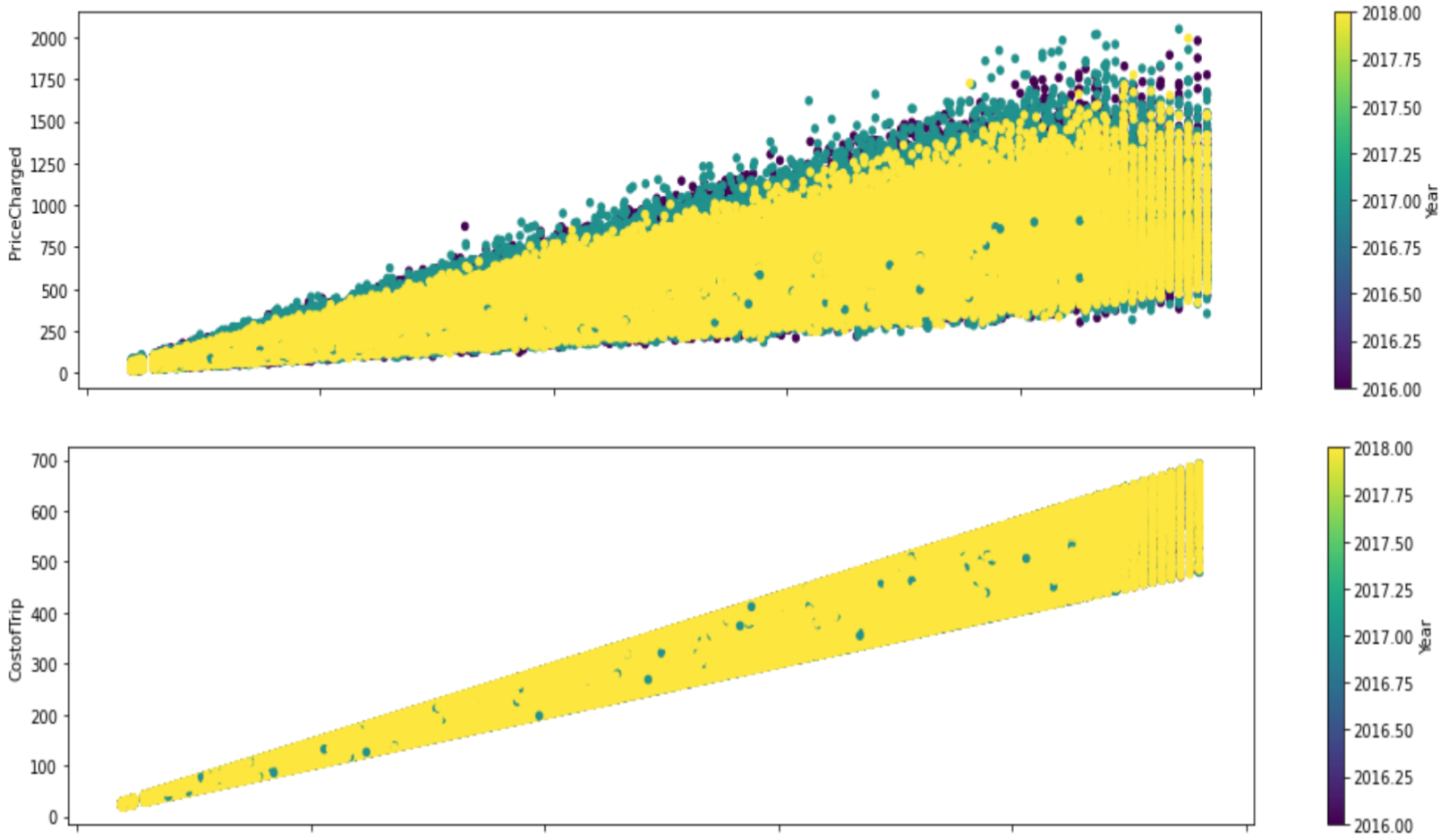
2. Customer ID variable is the common key on Transaction_ID and Customer_ID datasets. Hence, Transaction_ID and Customer_ID datasets can be combined by using Customer ID variable.

3. City variable is the common key on Cab_Data and City datasets. Hence, Cab_Data and City datasets can be combined using City variable.

Master dataset is shown below.

TransactionID	CustomerID	City	DateofTravel	Company	KMTravelled	PriceCharged	CostofTrip	Year	Month	PaymentMode	Gender	Age	Income(USD/Month)	Population	Users
10000011	29290	ATLANTA GA	2016-01-07	Pink Cab	30.45	370.95	313.635	2016	1	Card	Male	28	10813	814885	24701
10000012	27703	ATLANTA GA	2016-01-05	Pink Cab	28.62	358.52	334.854	2016	1	Card	Male	27	9237	814885	24701
10000013	28712	ATLANTA GA	2016-01-01	Pink Cab	9.04	125.20	97.632	2016	1	Cash	Male	53	11242	814885	24701
10000014	28020	ATLANTA GA	2016-01-06	Pink Cab	33.17	377.40	351.602	2016	1	Cash	Male	23	23327	814885	24701
10000015	27182	ATLANTA GA	2016-01-02	Pink Cab	8.73	114.62	97.776	2016	1	Card	Male	33	8536	814885	24701

Outlier Detection



There seems to be some outliers in Price Charged variable. However, keeping these outliers as the distance travelled varies and it might be important to get the Profit Margin.

Data Visualization

1. Correlation between data variables

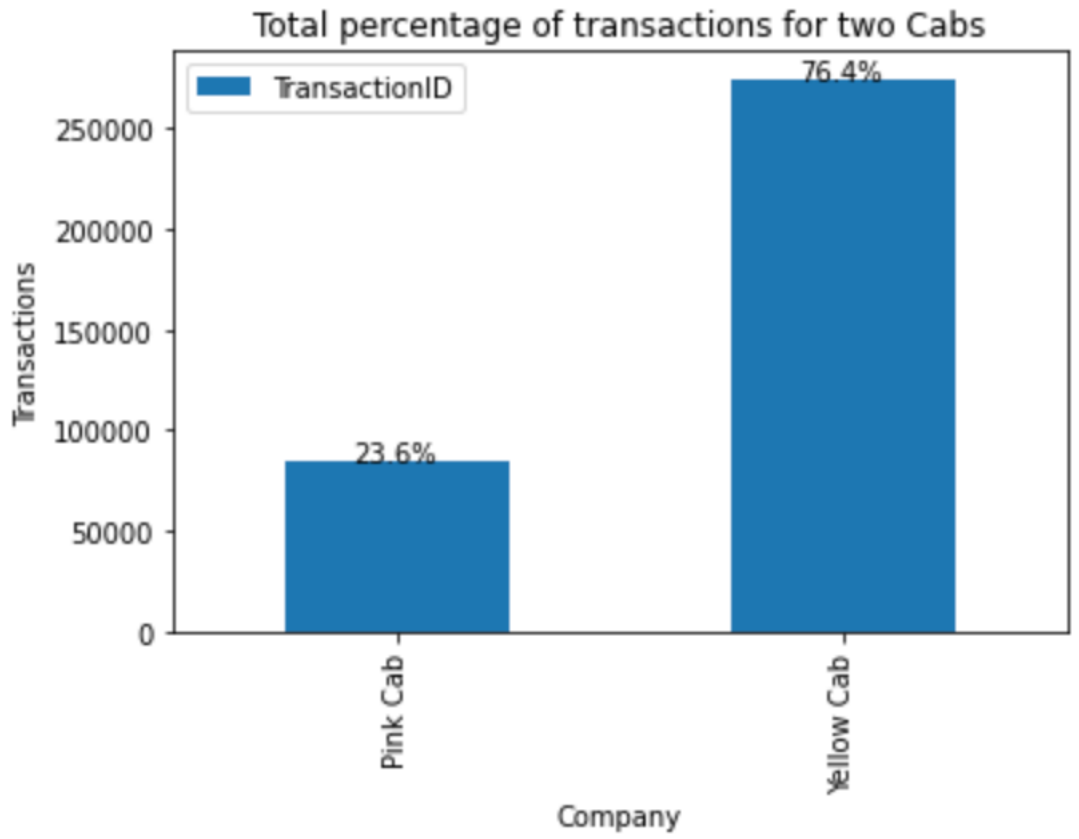
	TransactionID	CustomerID	KMTravelled	PriceCharged	CostofTrip	Year	Month	Age	Income(USD/Month)	Population	Users	Profit
TransactionID	1.000000	-0.016912	-0.001429	-0.052902	-0.003462	0.941475	0.284724	-0.001267	-0.001570	0.023868	0.013526	-0.087130
CustomerID	-0.016912	1.000000	0.000389	-0.177324	0.003077	-0.002480	-0.045030	-0.004735	-0.013608	-0.647052	-0.610742	-0.306527
KMTravelled	-0.001429	0.000389	1.000000	0.835753	0.981848	-0.001094	-0.001773	-0.000369	-0.000544	-0.002311	-0.000428	0.462768
PriceCharged	-0.052902	-0.177324	0.835753	1.000000	0.859812	-0.036903	-0.059639	-0.003084	0.003228	0.326589	0.281061	0.864154
CostofTrip	-0.003462	0.003077	0.981848	0.859812	1.000000	-0.001766	-0.008309	-0.000189	-0.000633	0.015108	0.023628	0.486056
Year	0.941475	-0.002480	-0.001094	-0.036903	-0.001766	1.000000	-0.033169	-0.000497	-0.001679	0.000061	-0.000556	-0.061420
Month	0.284724	-0.045030	-0.001773	-0.059639	-0.008309	-0.033169	1.000000	-0.002376	0.000585	0.064827	0.036285	-0.093886
Age	-0.001267	-0.004735	-0.000369	-0.003084	-0.000189	-0.000497	-0.002376	1.000000	0.003907	-0.009002	-0.005906	-0.005093
Income(USD/Month)	-0.001570	-0.013608	-0.000544	0.003228	-0.000633	-0.001679	0.000585	0.003907	1.000000	0.011868	0.010464	0.006148
Population	0.023868	-0.647052	-0.002311	0.326589	0.015108	0.000061	0.064827	-0.009002	0.011868	1.000000	0.915490	0.544079
Users	0.013526	-0.610742	-0.000428	0.281061	0.023628	-0.000556	0.036285	-0.005906	0.010464	0.915490	1.000000	0.457758
Profit	-0.087130	-0.306527	0.462768	0.864154	0.486056	-0.061420	-0.093886	-0.005093	0.006148	0.544079	0.457758	1.000000

There seems to be a correlation between KMTravelled with Price Charged/Cost of Trip. Also, Price Charged and Cost of Trip are correlated.

There seems to be a correlation between Price Charged with Population and Users.

Data Visualization - cont

2. Analyzing the percentage of Pink and Yellow Cab

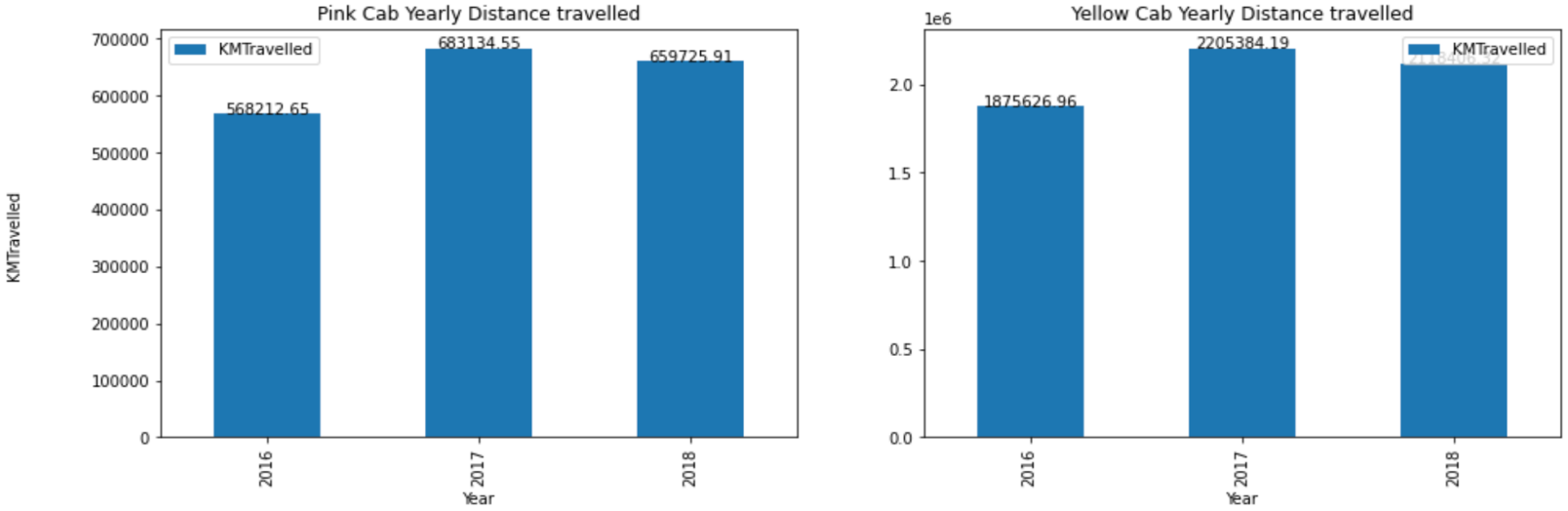


There seems to be a huge difference in the total number of transactions between Pink and Yellow cabs. Pink Cab is just 23.6% and Yellow has been 76.4%.

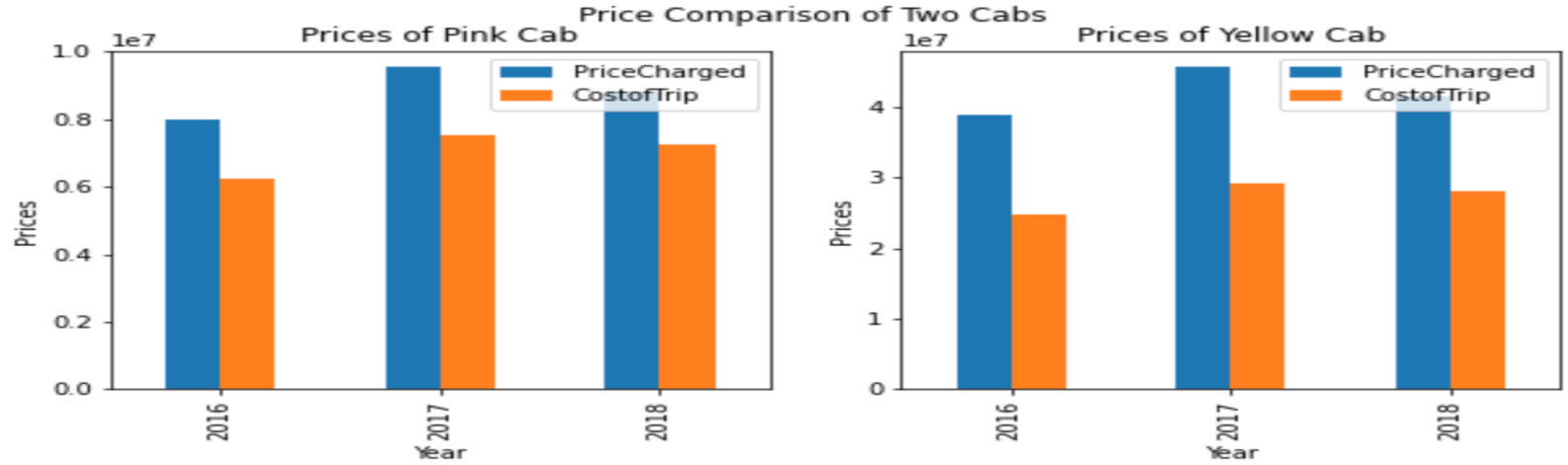
Data Visualization - cont

Data Glacier
Your Deep Learning Partner

3. Yearly Analysis of Pink and Yellow cabs by KMTravelled and Prices



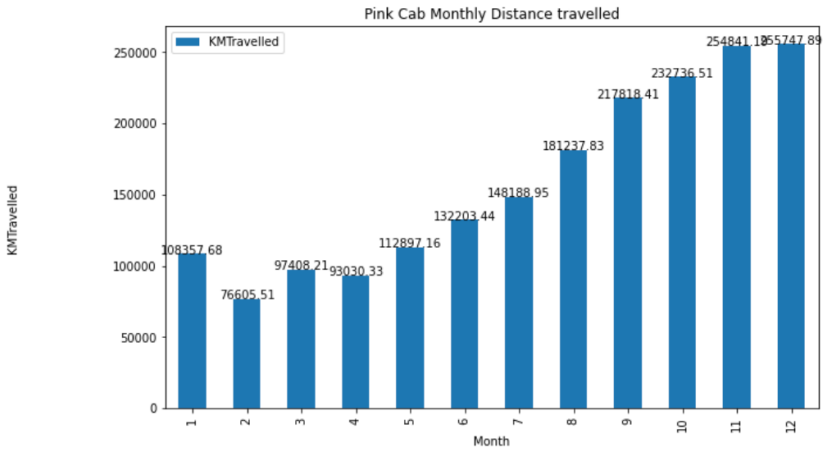
Distance travelled by Yellow cab per year is more than Pink cab, however the percent of travel in each year by each cab is not much different. Year 2017 has the highest percentage of travel compared to 2016 and 2018 for both the cabs.

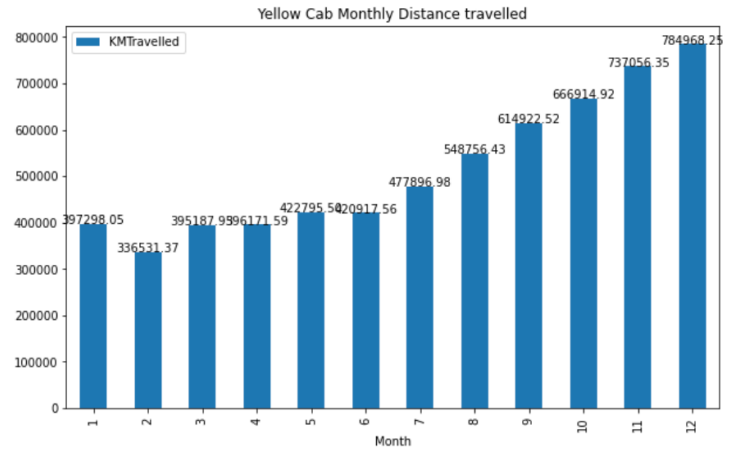


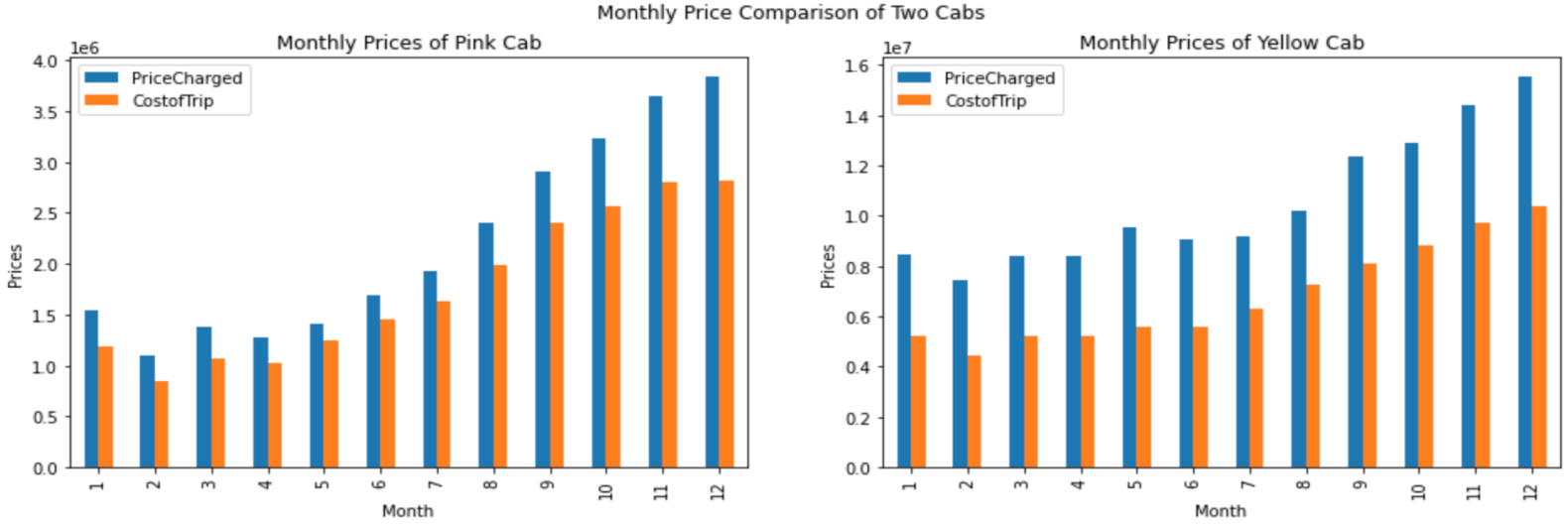
Profit margin by Yellow cab is more than Pink cab, as the difference between Price charged and Cost of Trip depicted in the graph above is more for Yellow cab than Pink cab in each year

Data Visualization - cont

4. Monthly Analysis of Pink and Yellow cabs by KMTravelled and Prices





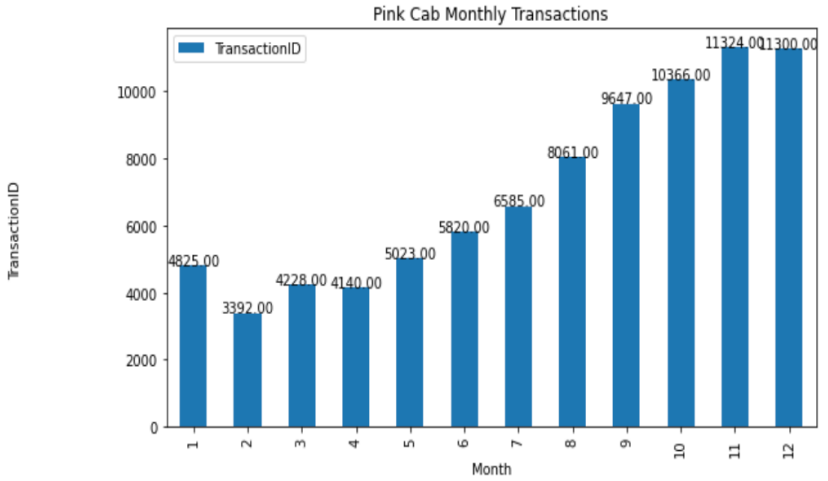


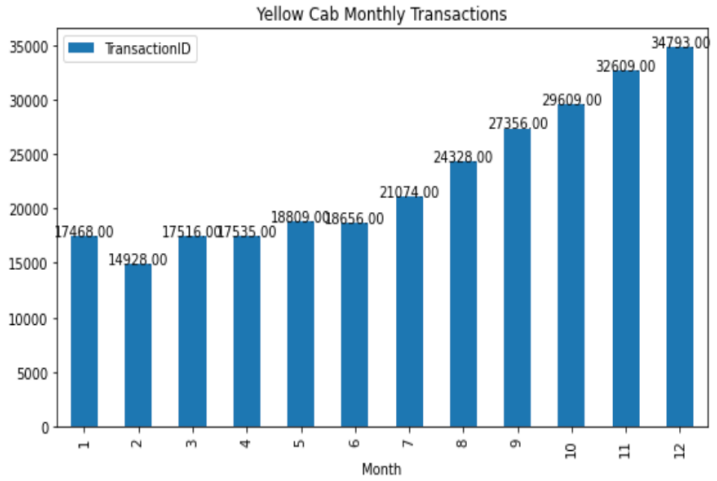
From above 4 graphs, its clear that the distance and profit margin of Yellow cab is more than Pink cab and is maximum in the month of December

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

5. Monthly Analysis of Pink and Yellow cabs by TransactionID





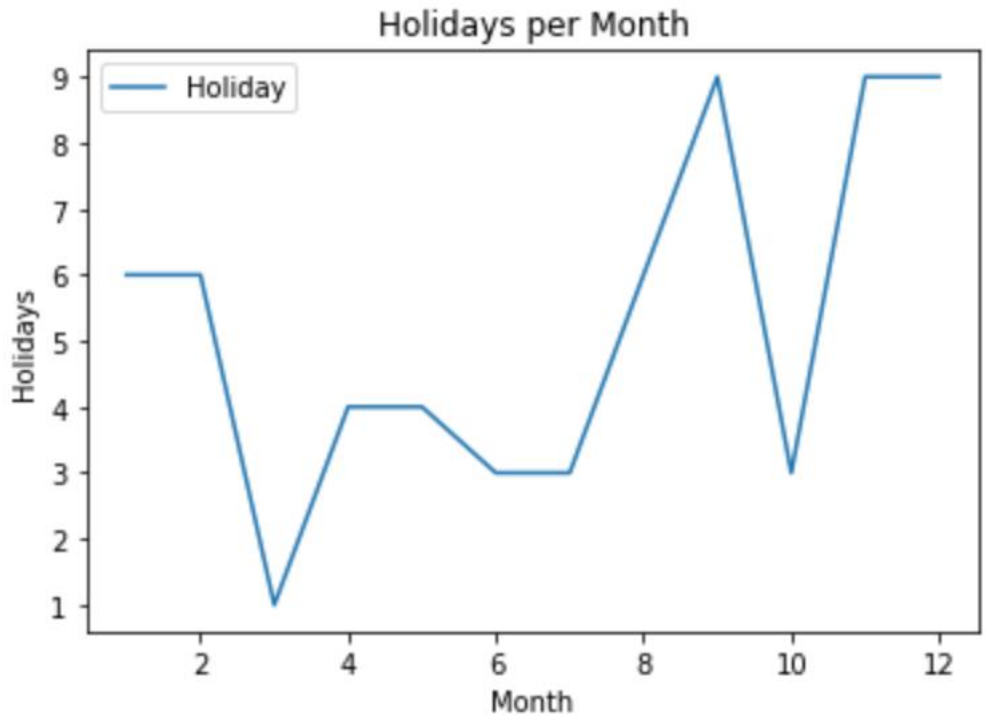
It's clear that the number of transactions in Dec months is more for both cabs compared to other months.

The number of transaction by Yellow cab in the month of Dec is between 30000 and 35000 where as, by Pink cab its between 10000 and 12000.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

6. Analysis of Holidays dataset

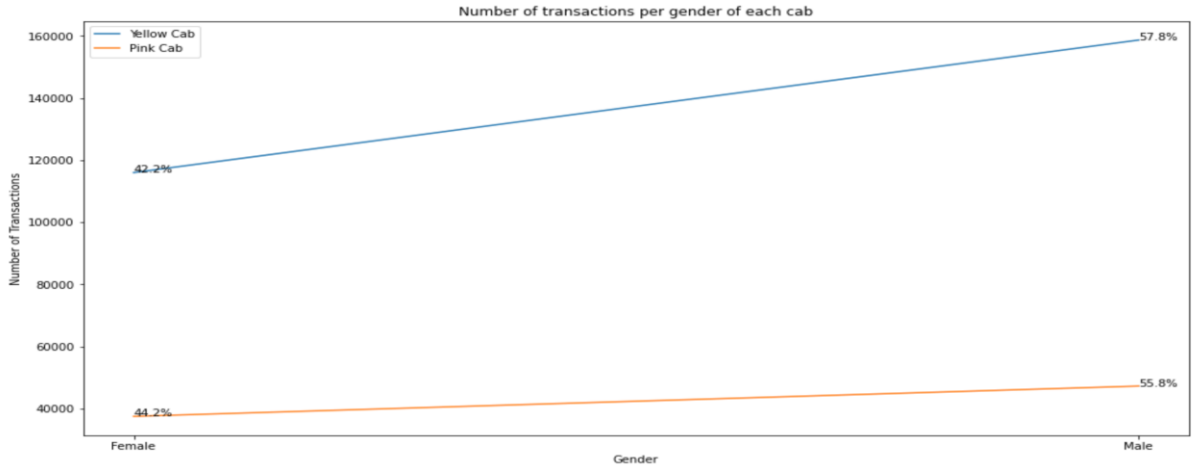


Number of holidays in the month of Sep, Nov and Dec are higher than rest of the months. This depicts that due to holidays in the month of Dec, the travel transactions are higher, which also leads to be the most profitable months

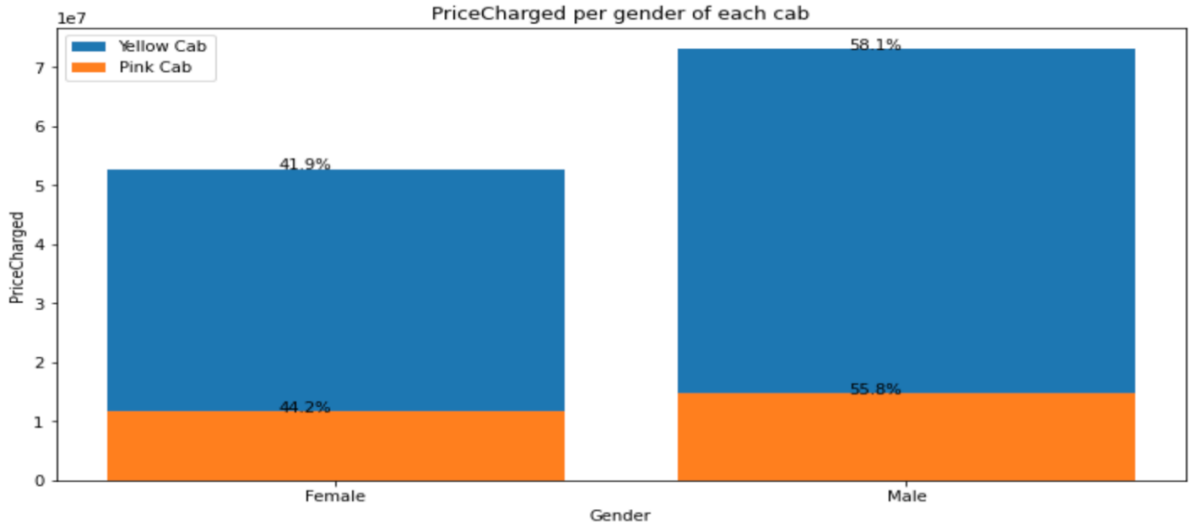
Data Visualization - cont

Data Glacier
Your Deep Learning Partner

7. Analysis of Pink and Yellow cabs by Gender



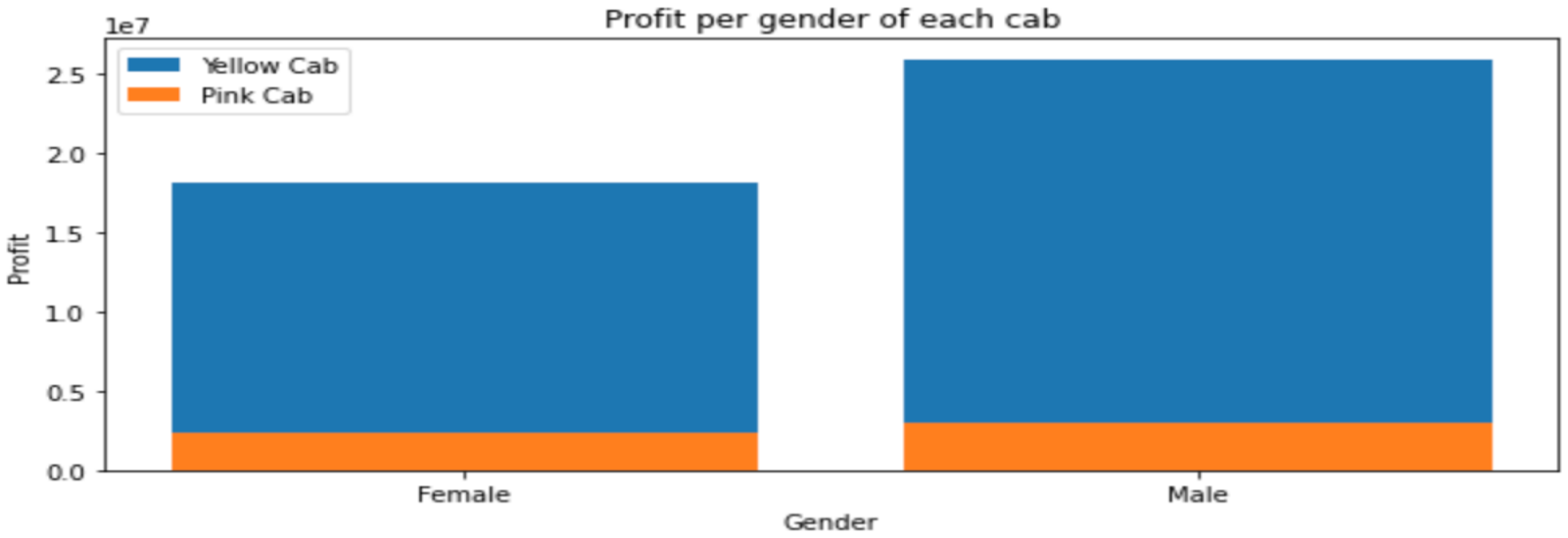
In Yellow cab, the numbers of transactions by Male customers are nearly 15% more than Female customers. Whereas in Pink Cab, the number of Male are 10% more than the Female users.



Pink cab charged nearly same price for Male and Female users.

Yellow cab charged higher Price for Male users than Female users.

Data Visualization - cont

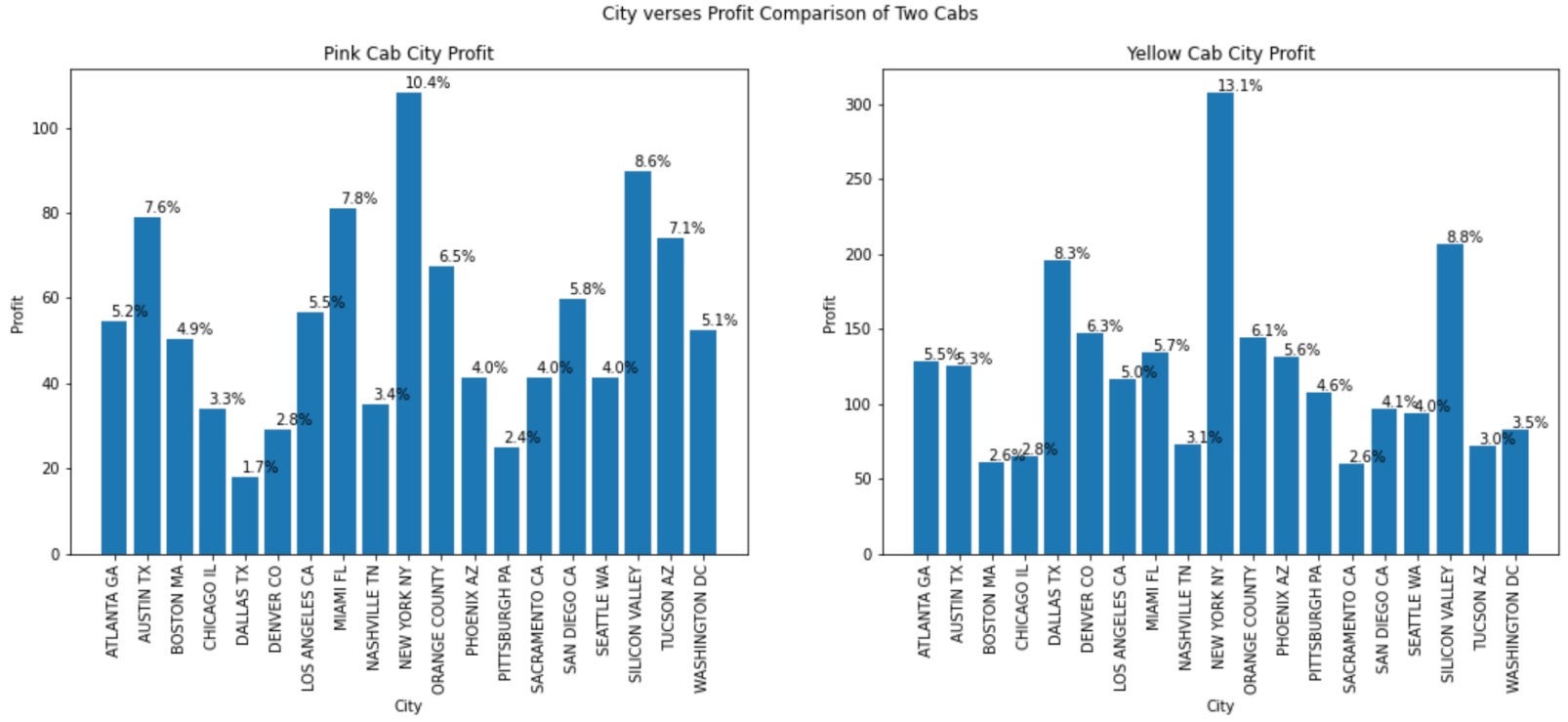


Profit of Pink cab is very less for both Male and Female users than Yellow cab.

Profit of Yellow cab is relatively high for Male users than Female users.

Data Visualization - cont

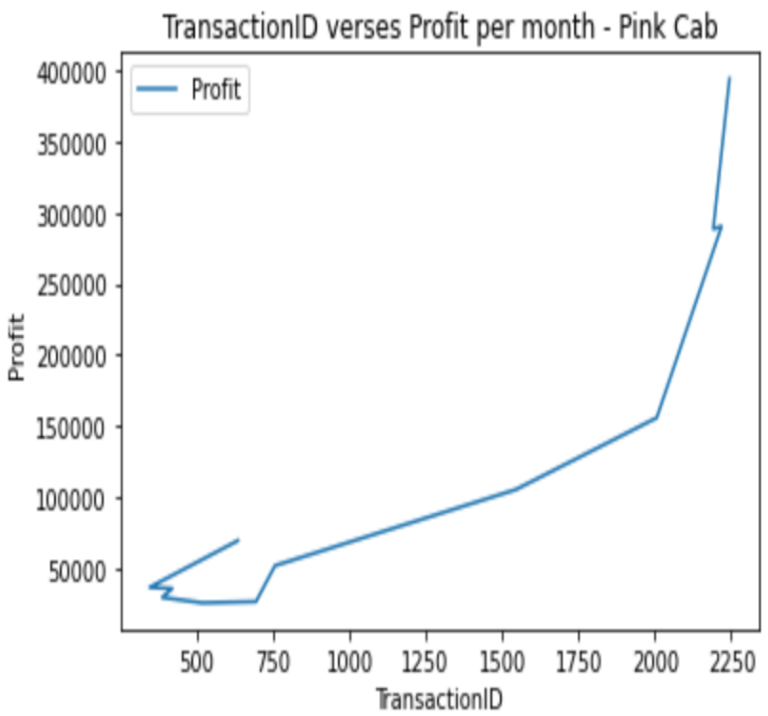
8. Analysis of Pink and Yellow cabs by City making Profit

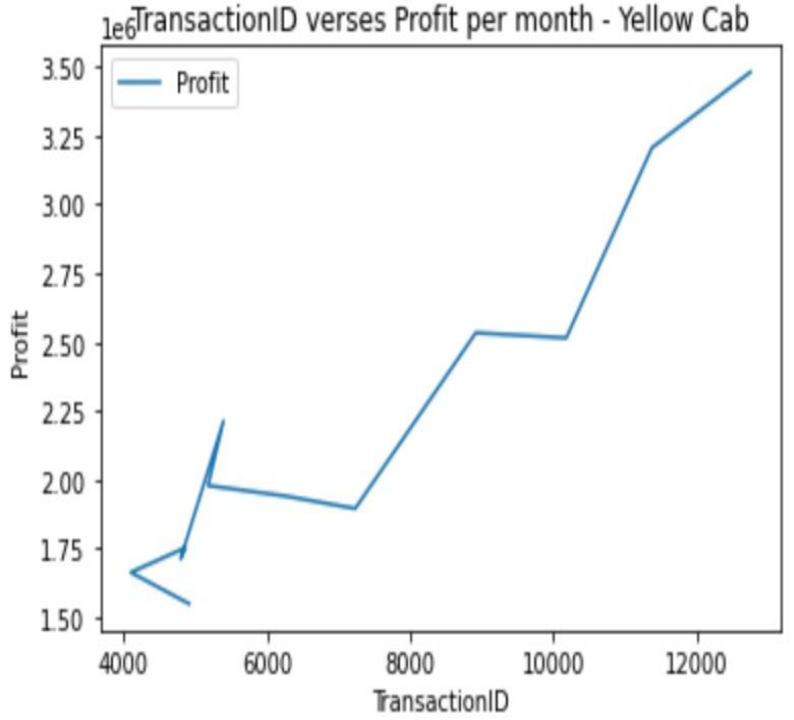


Profit gained by New York City is the highest than any other City by both the cabs. Yellow cab has 13.1% of Profit from New York City and Pink Cab has 10.4% of Profit from New York City.

Data Visualization - cont

9. Analysis the monthly transaction vs Price of New York City for Pink and Yellow cab



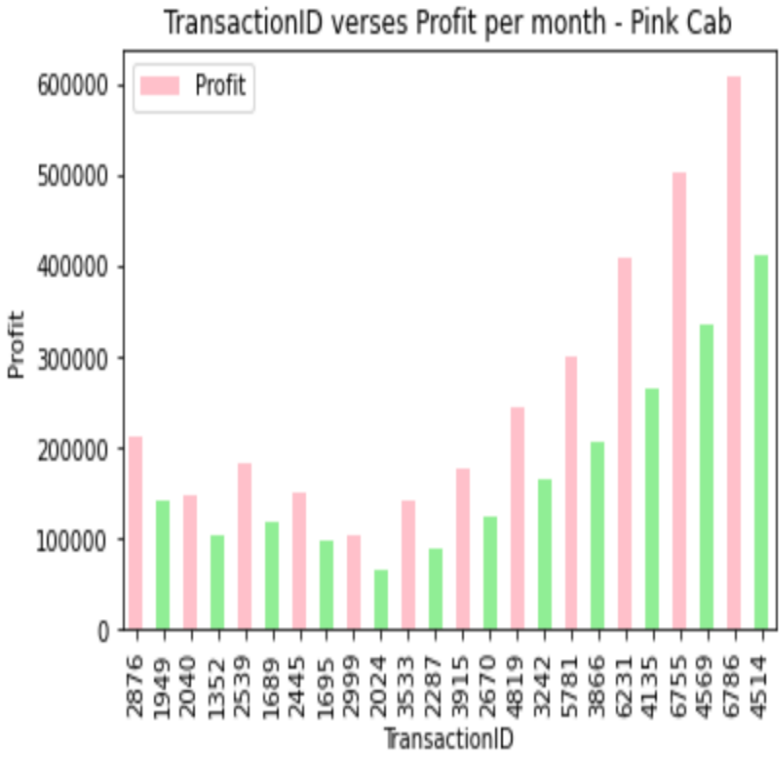


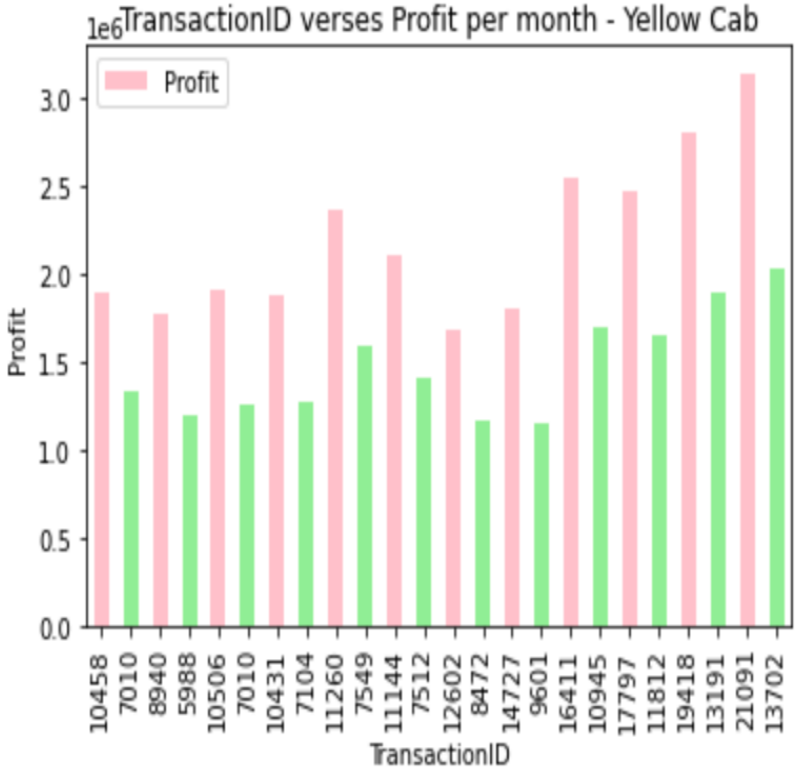
For the above 2 graphs, it is depicted that Profit increased with an increase in number of transactions for both the cabs

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

10. Analysis the monthly transaction vs Price for Pink and Yellow cabs





For the above 2 graphs, it is depicted that Card payment mode was higher for both the cabs compared to Cash on all months.

Data Glacier
Your Deep Learning Partner

EDA Summary

PINK CAB	YELLOW CAB
Number of transactions are 23.6%	Number of transactions are 76.4%
Distance travelled is less than Yellow cab	Distance travelled is more than Pink cab
Profit increased with an increase in transaction	Profit increased with an increase in transaction
Card payment mode is most used	Card payment mode is most used
Profit margin is less than Yellow cab	Profit margin is more than Pink cab
Price charged for Male and Female customers are same	Price charged for Male customers is nearly 15% more than Female customers
Profit from Male and Female customers are nearly same	Profit from Male customers are huge than Female customers
Profit from highest transaction city New York is 10.4%	Profit from highest transaction city New York is 13.1%

Hypothesis

Data Glacier

Your Deep Learning Partner

```
p1 = df_master[df_master['Company']=='Pink Cab'].groupby('Users')['Users'].count()
y1 = df_master[df_master['Company']=='Yellow Cab'].groupby('Users')['Users'].count()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,y1)
```

```
We accept null hypothesis
P value is  0.06451194993747668
```

Hypothesis 1

1. **Are the number of users different for both cabs**
   - Null Hypothesis: The number of users are same for both cabs.
   - Alternate Hypothesis: The number of users are different for both cabs.

   From the above hypothesis test, its proved that the number of users for both cabs are same.

```
p1 = df_pinkcab['Gender'].value_counts()
y1 = df_yellowcab['Gender'].value_counts()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,y1)
```

```
We accept alternate hypothesis
P value is  0.04922373545914366
```

Hypothesis 2

2. **Are the number of male and female customers different for both cabs**
   - Null Hypothesis: The number of male and female customers are not different for both cabs.
   - Alternate Hypothesis: The number of male and female customers are different for both cabs.

   From the above hypothesis test, its proved that the number of male and female customers for both cabs are different.

Hypothesis

Data Glacier

Your Deep Learning Partner

```
# Pink Cab
p1 = df_pinkcab[df_pinkcab['Gender']=='Male'].groupby('TransactionID')['PriceCharged'].mean()
p2 = df_pinkcab[df_pinkcab['Gender']=='Female'].groupby('TransactionID')['PriceCharged'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,p2)
```

```
We accept null hypothesis
P value is  0.8019871421072007
```

```
# Yellow Cab
y1 = df_yellowcab[df_yellowcab['Gender']=='Male'].groupby('TransactionID')['PriceCharged'].mean()
y2 = df_yellowcab[df_yellowcab['Gender']=='Female'].groupby('TransactionID')['PriceCharged'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(y1,y2)
```

```
We accept alternate hypothesis
P value is  2.0207950578635145e-08
```

Hypothesis 3

3. **Is there any difference in Price Charged between the Male and Female customers for both cabs**
   - Null Hypothesis: There is no difference in Price Charged for Male and Female customers.
   - Alternate Hypothesis: There is a difference in Price Charged for Male and Female customers.

From the above hypothesis test, its proved that Pink cab charges nearly same Price for male and female customers. Whereas, Yellow cabs charges different price for male and female customers.

Hypothesis

Data Glacier

Your Deep Learning Partner

```
# Pink Cab
p1 = df_pinkcab[df_pinkcab['Gender']=='Male'].groupby('TransactionID')['Profit'].mean()
p2 = df_pinkcab[df_pinkcab['Gender']=='Female'].groupby('TransactionID')['Profit'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,p2)
```

```
We accept null hypothesis
P value is  0.11515305900425798
```

```
# Yellow Cab
y1 = df_yellowcab[df_yellowcab['Gender']=='Male'].groupby('TransactionID')['Profit'].mean()
y2 = df_yellowcab[df_yellowcab['Gender']=='Female'].groupby('TransactionID')['Profit'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(y1,y2)
```

```
We accept alternate hypothesis
P value is  6.060473042494144e-25
```

Hypothesis 4

4. **Is there any difference in Profit margin between the Male and Female customers for both cabs**
   - Null Hypothesis: There is no difference in Profit for Male and Female customers.
   - Alternate Hypothesis:There is a difference in Profit for Male and Female customers.

From the above hypothesis test, its proved that Pink cab has negligible Profit for male and female customers. Whereas, Yellow cabs different profit for male and female customers.

Hypothesis

Data Glacier

Your Deep Learning Partner

```
# Pink Cab
p1 = df_pinkcab[df_pinkcab['Age']>60].groupby('TransactionID')['PriceCharged'].mean()
p2 = df_pinkcab[df_pinkcab['Age']<=60].groupby('TransactionID')['PriceCharged'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,p2)
```

```
We accept null hypothesis
P value is  0.15163688511281687
```

```
# Yellow Cab
y1 = df_yellowcab[df_yellowcab['Age']>60].groupby('TransactionID')['PriceCharged'].mean()
y2 = df_yellowcab[df_yellowcab['Age']<=60].groupby('TransactionID')['PriceCharged'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(y1,y2)
```

```
We accept alternate hypothesis
P value is  0.00037639135035337577
```

Hypothesis 5

5. **Is there any difference on Price charged for Age > 60 for both cabs**
   - Null Hypothesis: There is no difference on the Price Charged for the users above 60 years of Age.
   - Alternate Hypothesis: There is a difference on the Price Charged for the users above 60 years of Age.

From the above hypothesis test, its proved that Pink cab charges nearly same Price for customers above 60 years of age. Whereas, Yellow cabs charges less for the customers above the age of 60.

Hypothesis

Data Glacier

Your Deep Learning Partner

```
# Pink Cab
p1 = df_pinkcab[df_pinkcab['PaymentMode']=='Card'].groupby('TransactionID')['Profit'].mean()
p2 = df_pinkcab[df_pinkcab['PaymentMode']=='Cash'].groupby('TransactionID')['Profit'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,p2)
```

```
We accept null hypothesis
P value is  0.7900465828793288
```

```
# Yellow Cab
y1 = df_yellowcab[df_yellowcab['PaymentMode']=='Card'].groupby('TransactionID')['Profit'].mean()
y2 = df_yellowcab[df_yellowcab['PaymentMode']=='Cash'].groupby('TransactionID')['Profit'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(y1,y2)
```

```
We accept null hypothesis
P value is  0.2933060638298729
```

Hypothesis 6

6. **Is there any difference on Price charged for Age > 60 for both cabs**
   - Null Hypothesis: There is no difference in the Profit gained by Card and Cash Payment Mode.
   - Alternate Hypothesis: There is a difference in the Profit gained by Card and Cash Payment Mode.

From the above hypothesis test, its proved that there is no difference in the Profit gained by Payment Modes for both Pink cab and Yellow cabs.

Data Glacier

Your Deep Learning Partner

Hypothesis

```
# Pink and Yellow Cab
p1 = df_pinkcab[df_pinkcab['City']=='NEW YORK NY'].groupby('TransactionID')['Profit'].mean()
y1 = df_yellowcab[df_yellowcab['City']=='NEW YORK NY'].groupby('TransactionID')['Profit'].mean()
# Calling TestHypothesis to test the hypothesis
TestHypothesis(p1,y1)
```

```
We accept alternate hypothesis
P value is  0.0
```

Hypothesis 7

7. **Is there any difference in Profit gained from New York City for both cabs**
   - Null Hypothesis: There is no difference in the Profit gained for New York City.
   - Alternate Hypothesis: There is a difference in the Profit gained for New York City.

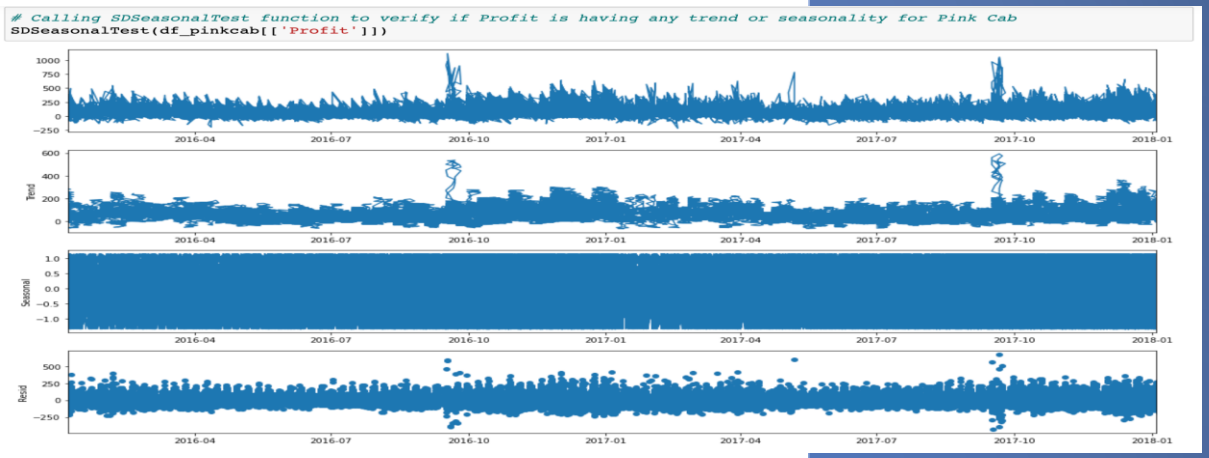
From the above hypothesis test, its proved that there is difference in the Profit gained from New York City for both Pink cab and Yellow cabs.

Data Glacier

Your Deep Learning Partner

Hypothesis

```
# Calling SDSeasonalTest function to verify if Profit is having any trend or seasonality for Pink Cab
SDSeasonalTest(df_pinkcab[['Profit']])
```

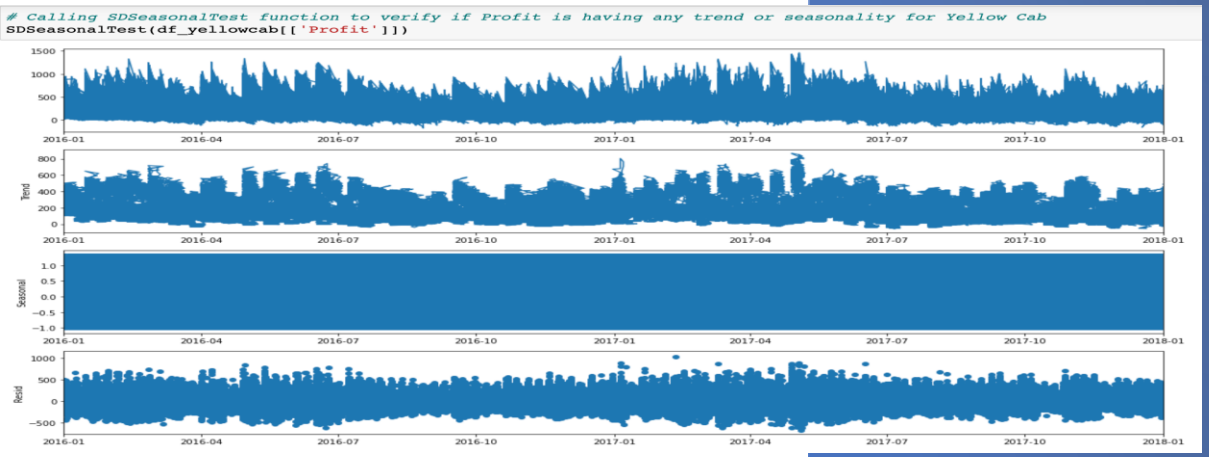


8. Is there any seasonality for Profit for both cabs

- Null Hypothesis: There is no trend and seasonality in the Profit gained for both cabs.
- Alternate Hypothesis: There is a trend and seasonality in the Profit gained for both cabs.

From the above graphs and the p_value (0.000000), its proved that Profit variable is stationary and does not have any trend or seasonality for both Pink and Yellow cabs.

```
# Calling SDSeasonalTest function to verify if Profit is having any trend or seasonality for Yellow Cab
SDSeasonalTest(df_yellowcab[['Profit']])
```




```
# Calling ADFStationarityTest function to verify if
# Profit is stationarity using dickey-fuller-test for Pink Cab
ADFStationarityTest(df_pinkcab['Profit'])
```

```
Augmented Dickey-Fuller Test Results:
ADF Test Statistics             -21.968746
p-value                           0.000000
#Lag Used                        50.000000
Number of Observations Used    84660.000000
Critical Value (1%)              -3.430427
Critical Value (5%)              -2.861574
Critical Value (10%)             -2.566788
dtype: float64
The time series data has no unit roots and hence it is stationary
```

```
# Calling ADFStationarityTest function to verify if
# Profit is stationarity using dickey-fuller-test for Yellow Cab
ADFStationarityTest(df_yellowcab['Profit'])
```

```
Augmented Dickey-Fuller Test Results:
ADF Test Statistics             -20.268509
p-value                           0.000000
#Lag Used                        76.000000
Number of Observations Used    274604.000000
Critical Value (1%)              -3.430374
Critical Value (5%)              -2.861551
Critical Value (10%)             -2.566776
dtype: float64
The time series data has no unit roots and hence it is stationary
```

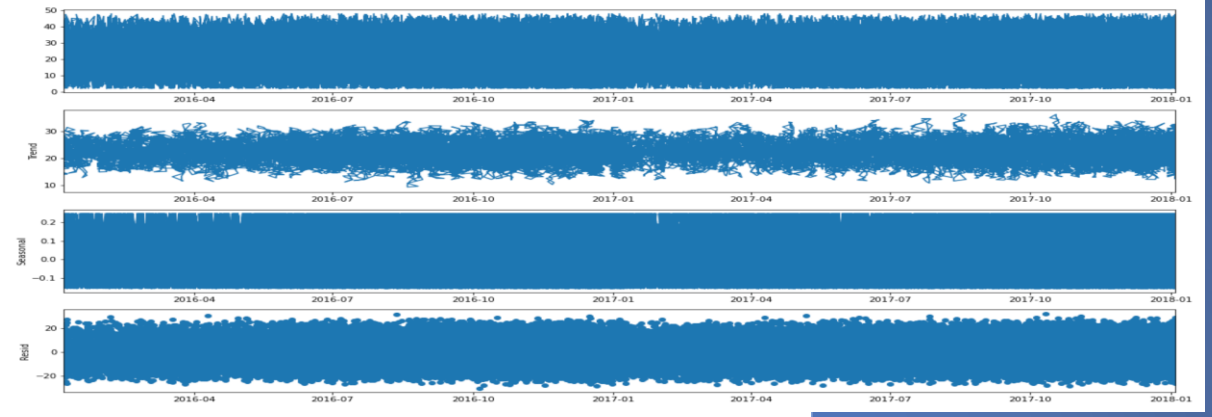
Hypothesis 8

Data Glacier

Your Deep Learning Partner

Hypothesis

```
# Calling SDSeasonalTest function to verify if the KMTravelled is having any trend or seasonality for Pink Cab
SDSeasonalTest(df_pinkcab[['KMTravelled']])
```

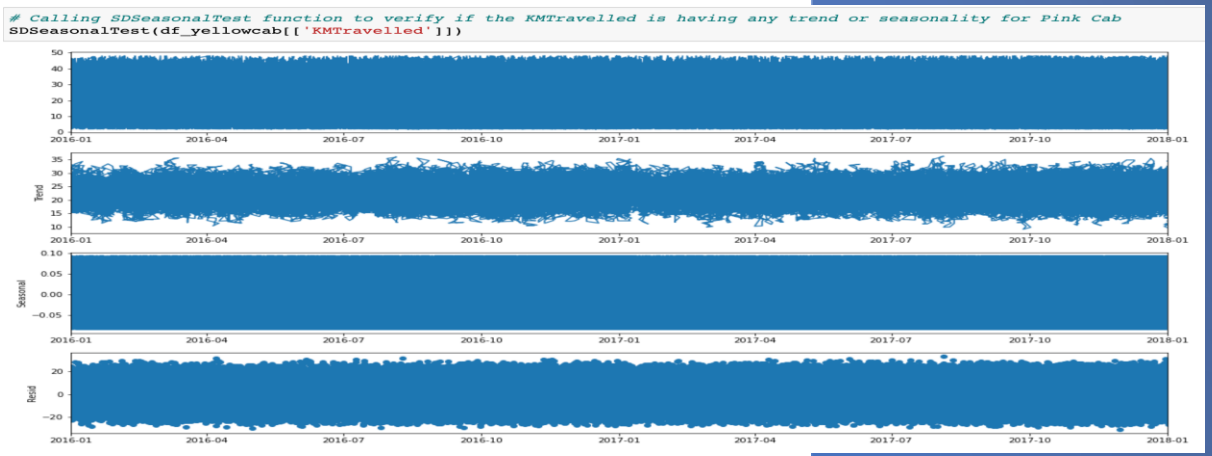


9. Is there any seasonality in KM Travelled for both cabs

- Null Hypothesis: There is no trend and seasonality in the KMTravelled for both cabs.
- Alternate Hypothesis: There is a trend and seasonality in the KMTravelled for both cabs

From the above graphs and the p_value (0.000000), its proved that KMTravelled variable is stationary and does not have any trend and seasonality for both Pink and Yellow cabs.

```
# Calling SDSeasonalTest function to verify if the KMTravelled is having any trend or seasonality for Pink Cab
SDSeasonalTest(df_yellowcab[['KMTravelled']])
```




```
# Calling ADFStationarityTest function to verify if
# KMTravelled is stationarity using dickey-fuller-test for Pink Cab
ADFStationarityTest(df_pinkcab['KMTravelled'])
```

```
Augmented Dickey-Fuller Test Results:
ADF Test Statistics            -292.262786
p-value                           0.000000
#Lag Used                         0.000000
Number of Observations Used    84710.000000
Critical Value (1%)              -3.430427
Critical Value (5%)              -2.861574
Critical Value (10%)             -2.566788
dtype: float64
The time series data has no unit roots and hence it is stationary
```

```
# Calling ADFStationarityTest function to verify if
# KMTravelled is stationarity using dickey-fuller-test for Yellow Cab
ADFStationarityTest(df_yellowcab['KMTravelled'])
```

```
Augmented Dickey-Fuller Test Results:
ADF Test Statistics            -523.786491
p-value                           0.000000
#Lag Used                         0.000000
Number of Observations Used    274680.000000
Critical Value (1%)              -3.430374
Critical Value (5%)              -2.861551
Critical Value (10%)             -2.566776
dtype: float64
The time series data has no unit roots and hence it is stationary
```

Hypothesis 9

Recommendations

Looking at the results of Pinks cab and Yellow cab, the number of transactions and the profit margin for Yellow cab is much more than Pink cab. Hence the recommendation would be to invest in Yellow cab.

References

[1]***https://www.kaggle.com/datasets/donnetew/us-holiday-dates-2004-2021***

Thank You