Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Investment Case Study

19 June 2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

Data Glacier
Your Deep Learning Partner

Background:

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.

Problem Statement

Conduct a data exploration to determine which cab company the XYZ firm should make an investment in based on the datasets

I conducted the analysis with a set of steps:

- **Importing and Understanding Data**
- **Altering Datatypes**
- **Merging Data**
- **Analysis**
- **Hypotheses**

Data Glacier
Your Deep Learning Partner

Approach

1. Importing and Understanding Data
2. Altering Datatypes
3. Merging Data and Adding Info
4. Analysis & Visualizations
5. Hypotheses & Investigation

Data Glacier
Your Deep Learning Partner

Given Data

1. Cab_Data.csv
   - Includes information about details of the cab companies, such as Price Charged, Company, and Cost

2. Customer_ID.csv
   - Includes information about customer demographics, such as Gender, Age, and Income

3. Transaction_ID.csv
   - Includes information about customer mapping and payment methods

4. City.csv
   - Includes information about population, cities, and number of cab users

EDA

Importing and Understanding Data

Data Glacier
Your Deep Learning Partner

Cab_Data.csv

	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
2	10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.776

City.csv

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468
2	LOS ANGELES CA	1,595,037	144,132
3	MIAMI FL	1,339,155	17,675
4	SILICON VALLEY	1,177,609	27,247

Transaction_ID.csv

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card

Customer_ID.csv

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536

EDA

Altering Datatypes

Data Glacier
Your Deep Learning Partner

1. Changing Date of Travel in Cab Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 359392 entries, 0 to 359391
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Transaction ID  359392 non-null  int64
 1   Date of Travel  359392 non-null  datetime64[ns]
 2   Company         359392 non-null  object
 3   City            359392 non-null  object
 4   KM Travelled    359392 non-null  float64
 5   Price Charged   359392 non-null  float64
 6   Cost of Trip    359392 non-null  float64
 7   Profit          359392 non-null  float64
dtypes: datetime64[ns](1), float64(4), int64(1), object(2)
memory usage: 21.9+ MB
```

2. Changing Population and Users in City Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   City        20 non-null     object
 1   Population  20 non-null     float64
 2   Users       20 non-null     float64
dtypes: float64(2), object(1)
memory usage: 608.0+ bytes
```

EDA

Merging Data and Adding Info

Data Glacier
Your Deep Learning Partner

1. Add a Profit Section to Cab Dataset

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Profit
0	10000011	2016-01-10	Pink Cab	ATLANTA GA	30.45	370.95	313.635	57.315
1	10000012	2016-01-08	Pink Cab	ATLANTA GA	28.62	358.52	334.854	23.666
2	10000013	2016-01-04	Pink Cab	ATLANTA GA	9.04	125.20	97.632	27.568
3	10000014	2016-01-09	Pink Cab	ATLANTA GA	33.17	377.40	351.602	25.798
4	10000015	2016-01-05	Pink Cab	ATLANTA GA	8.73	114.62	97.776	16.844

2. Merging Data into a Master Dataset with all Data Included

EDA

Analysis and Visualizations -

Correlation Table

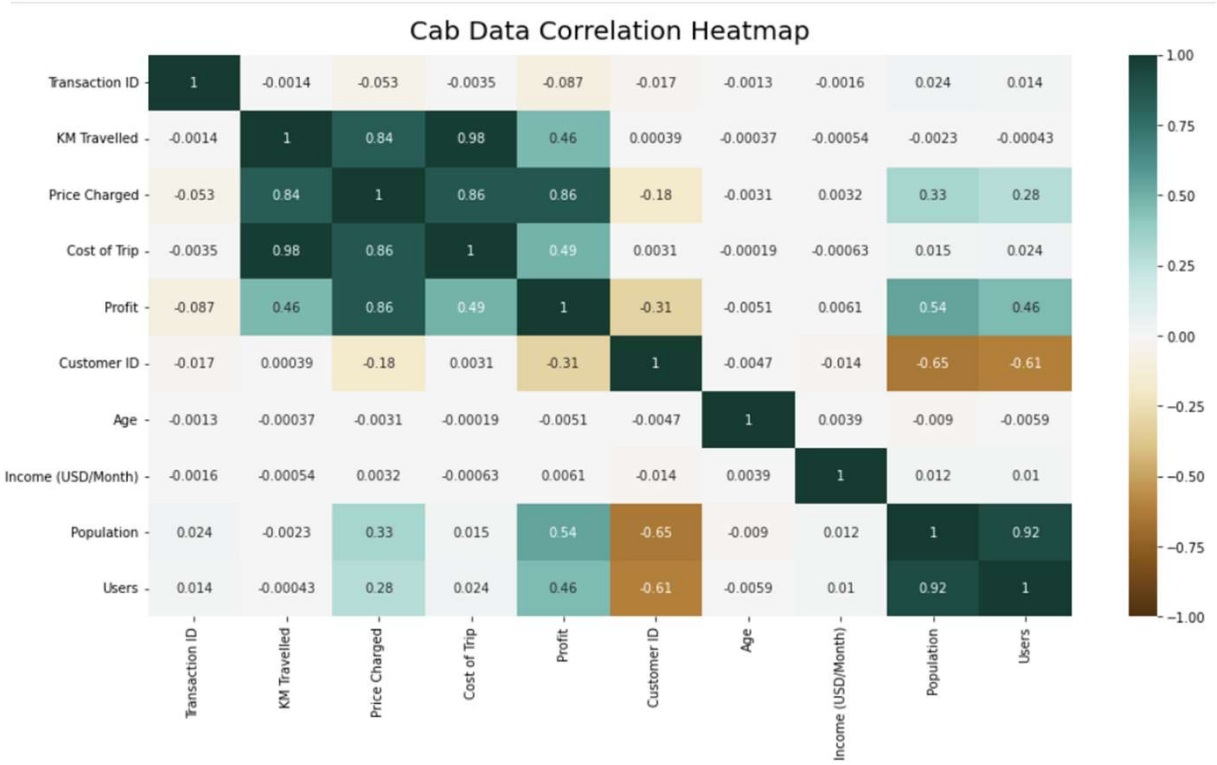
Data Glacier
Your Deep Learning Partner

	Transaction ID	KM Travelled	Price Charged	Cost of Trip	Profit	Customer ID	Age	Income (USD/Month)	Population	Users
Transaction ID	1.000000	-0.001429	-0.052902	-0.003462	-0.087130	-0.016912	-0.001267	-0.001570	0.023868	0.013526
KM Travelled	-0.001429	1.000000	0.835753	0.981848	0.462768	0.000389	-0.000369	-0.000544	-0.002311	-0.000428
Price Charged	-0.052902	0.835753	1.000000	0.859812	0.864154	-0.177324	-0.003084	0.003228	0.326589	0.281061
Cost of Trip	-0.003462	0.981848	0.859812	1.000000	0.486056	0.003077	-0.000189	-0.000633	0.015108	0.023628
Profit	-0.087130	0.462768	0.864154	0.486056	1.000000	-0.306527	-0.005093	0.006148	0.544079	0.457758
Customer ID	-0.016912	0.000389	-0.177324	0.003077	-0.306527	1.000000	-0.004735	-0.013608	-0.647052	-0.610742
Age	-0.001267	-0.000369	-0.003084	-0.000189	-0.005093	-0.004735	1.000000	0.003907	-0.009002	-0.005906
Income (USD/Month)	-0.001570	-0.000544	0.003228	-0.000633	0.006148	-0.013608	0.003907	1.000000	0.011868	0.010464
Population	0.023868	-0.002311	0.326589	0.015108	0.544079	-0.647052	-0.009002	0.011868	1.000000	0.915490
Users	0.013526	-0.000428	0.281061	0.023628	0.457758	-0.610742	-0.005906	0.010464	0.915490	1.000000

EDA

Analysis and Visualizations -

Correlation Heatmap

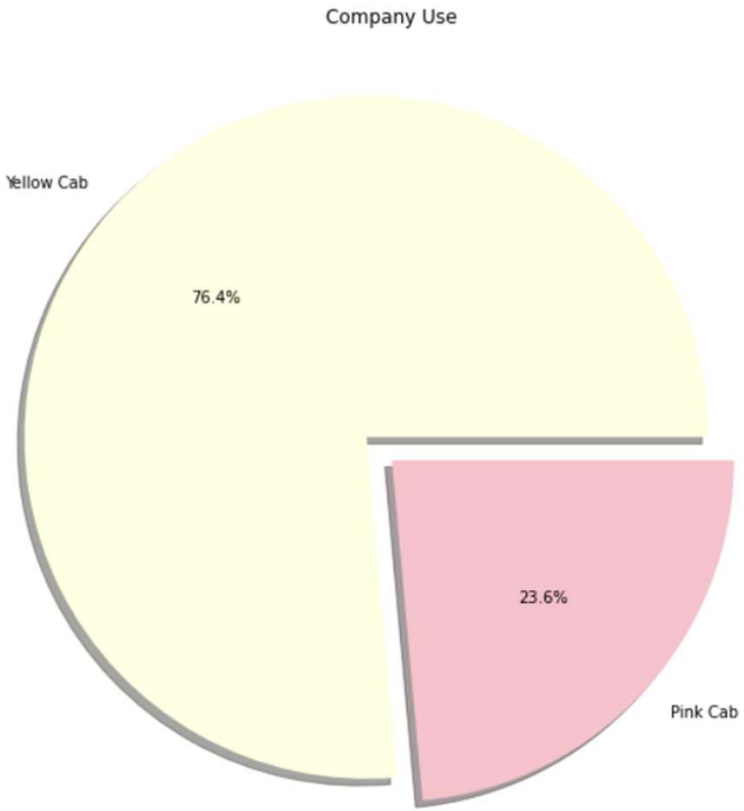


Data Glacier

Your Deep Learning Partner

EDA

Analysis and Visualizations - Company Distribution



We can see that the Yellow Cab company is used more often than the Pink Cab company

Data Glacier
Your Deep Learning Partner

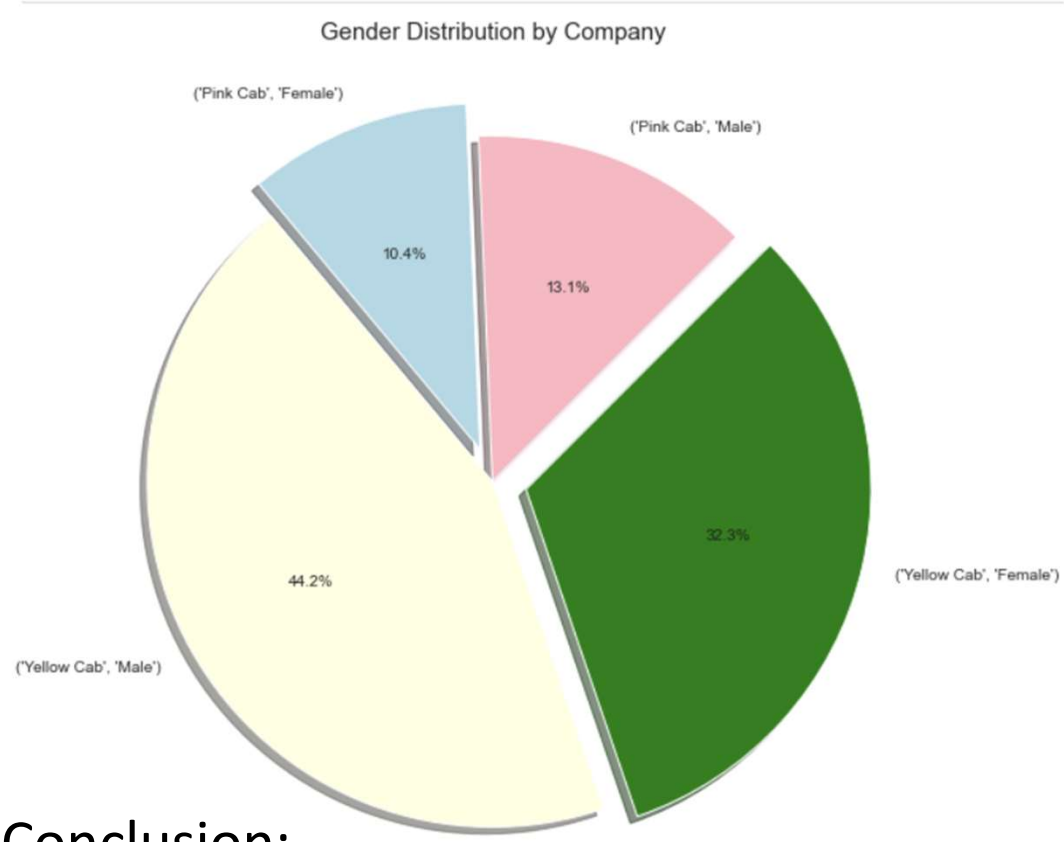
EDA

Hypotheses and Investigation

Data Glacier
Your Deep Learning Partner

Hypothesis 1:

Females are more likely to use the Yellow Cab than the Pink Cab



Conclusion:

Females use the Yellow Cab 32.3% of the time while they only use the Pink Cab 10.4% of the time

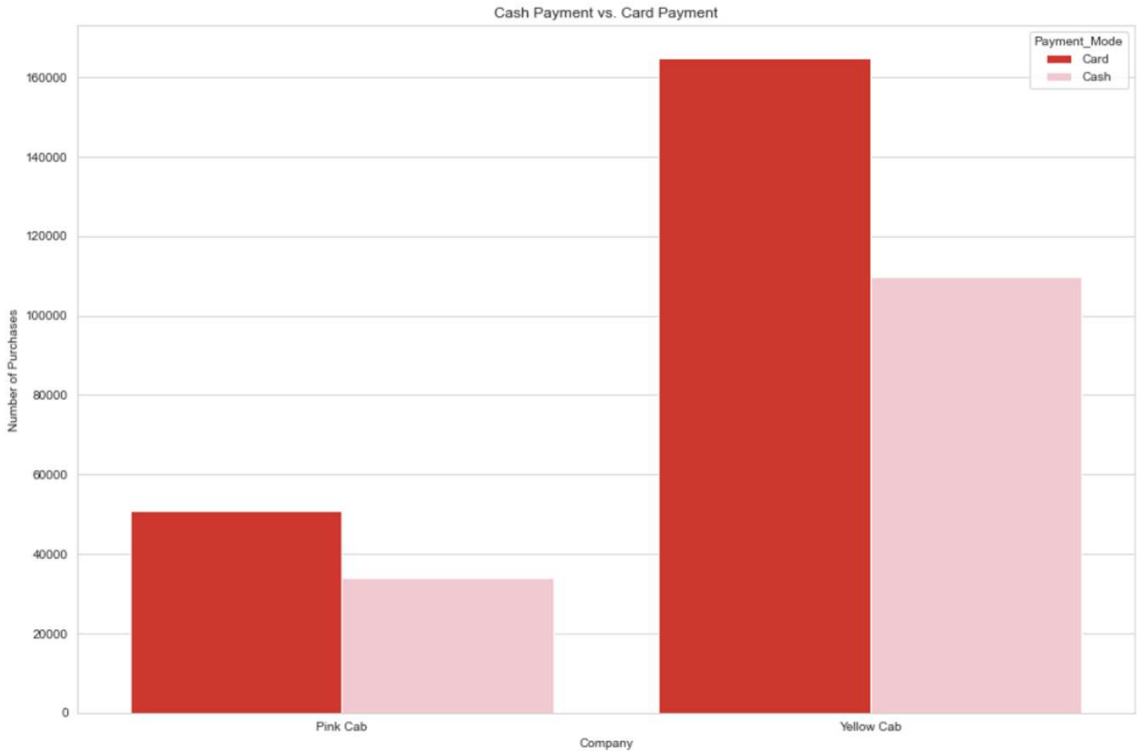
EDA

Hypotheses and Investigation

Data Glacier
Your Deep Learning Partner

Hypothesis 2:

All riders prefer to pay with card than cash



Conclusion:

The majority of riders pay with card rather than cash

EDA

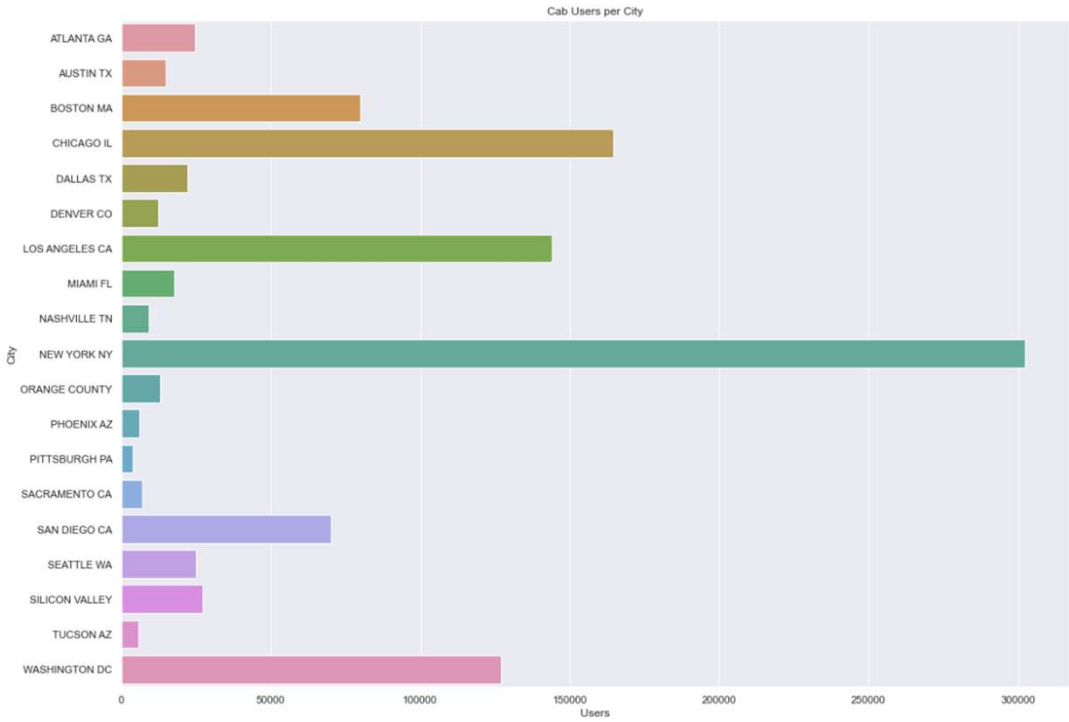
Hypotheses and Investigation

Data Glacier
Your Deep Learning Partner

Hypothesis 3:

Those who tend to live in bigger cities have greater cab service use

City	
NEW YORK NY	99885
CHICAGO IL	56625
LOS ANGELES CA	48033
WASHINGTON DC	43737
BOSTON MA	29692
SAN DIEGO CA	20488
SILICON VALLEY	8519
SEATTLE WA	7997
ATLANTA GA	7557
DALLAS TX	7017
MIAMI FL	6454
AUSTIN TX	4896
ORANGE COUNTY	3982
DENVER CO	3825
NASHVILLE TN	3010
SACRAMENTO CA	2367
PHOENIX AZ	2064
TUCSON AZ	1931
PITTSBURGH PA	1313



Conclusion:

Those who live in bigger cities have greater cab service use

EDA

Hypotheses and Investigation

Data Glacier
Your Deep Learning Partner

Hypothesis 4:

The amount of profit increased as the KM travelled increased

Conclusion:

There does not seem to be much of a relationship between profit and the KM travelled between any of the cab companies

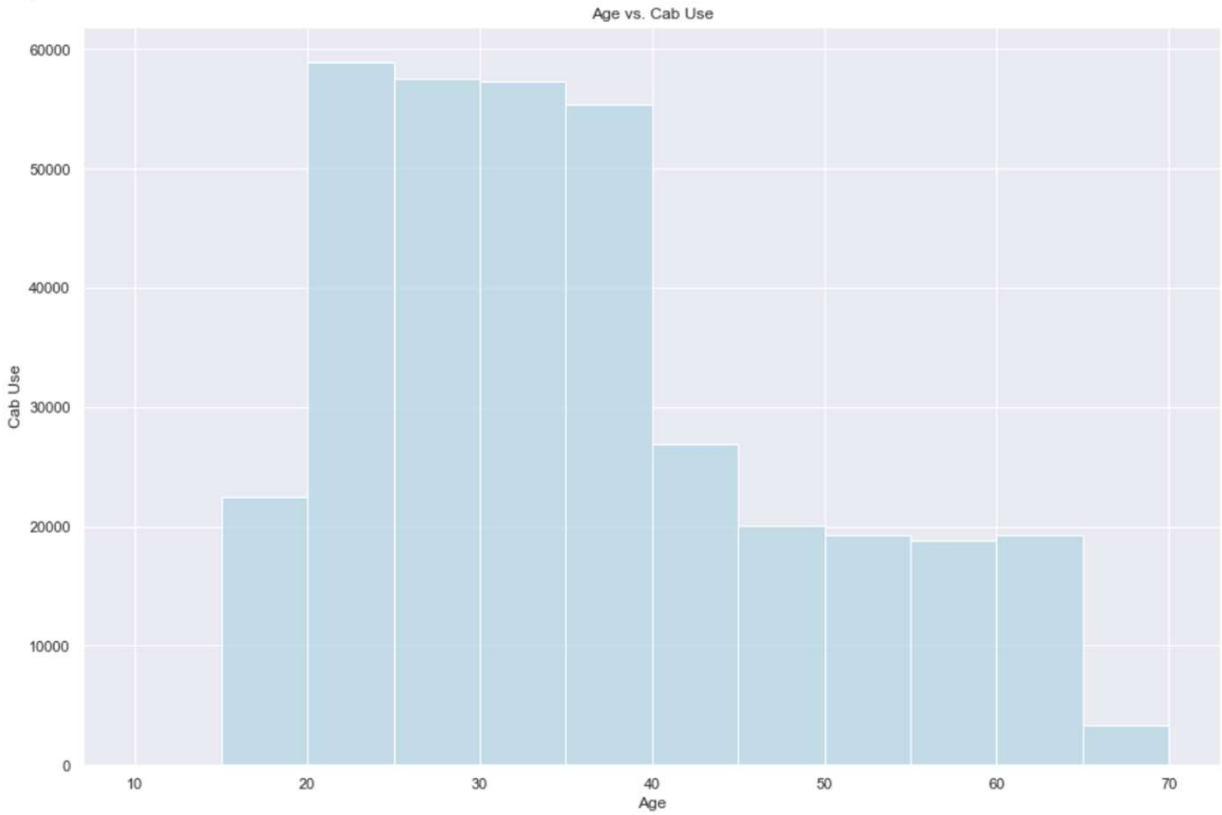
EDA

Hypotheses and Investigation

Data Glacier
Your Deep Learning Partner

Hypothesis 5:

Age plays a role in cab use



Conclusion:

Those who are of ages 20-40 have a much larger cab use than those who are older (45-70)

EDA Summary

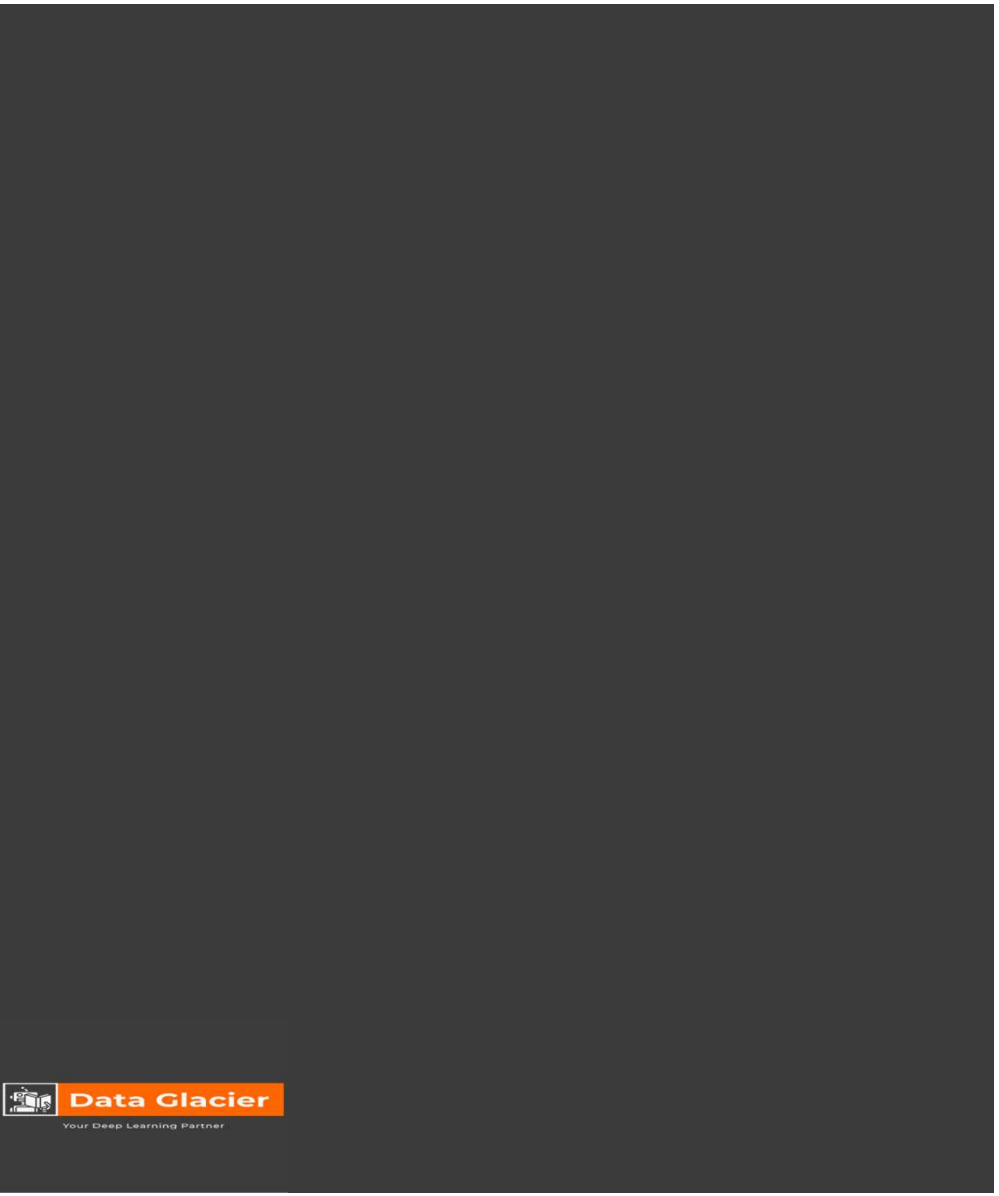
Data Glacier
Your Deep Learning Partner

The analysis shows:

1. A very high correlation between cost of trip and km travelled & population and users
2. There is NO duplicated data or NA values in any of the sets
3. More females prefer the yellow cab (32.3%) then the pink cab (10.4%)
4. The overall popularity for any cab ride purchase is through card
5. Cities with a larger population tend to use the cab services more often than cities with a smaller population
6. Those who are of ages 20-40 have a much larger cab use than those who are older (45-70)

Overall Recommendations

Through the data analysis conducted, investing in the **Yellow Cab** would be more beneficial to the XYZ firm.



Thank You,
Maria Contractor