G2M insight for Cab Investment firm

Joseph Nnodim

20th August 2022

Agenda

1. Introduction
2. Data understanding
3. Data Exploration and Data Quality Check
4. Data Cleaning
5. Statistical Analysis
6. Interpretation

Introduction

Business Analysis

Problem definition

XYZ, a private firm in US needs to gain insights into the cab industry and leverage opportunities in the growing market, as per their Go-to-Market(G2M) strategy. This will inform its decision on the right investment.

Objectives

- To investigate and understand the dataset in terms of schema, structure and quality
- To handle existing data quality issues
- To carry out exploratory through visualization and analytical approaches for the two companies

-To recommend the better company for XYZ's investment that will drive the most value and attain the highest profit.

Solution Requirements

Explore, transform, analysis and generate insight from data using statistical techniques.

DATA INTAKE REPORT

Tabular data details:

Transaction_ID data: The file containing this data is in csv format

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788kb

Cab_Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663

Customer_ID data: The file containing this data is in csv format

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027kb

City data: The file containing this data is in csv format

Total number of observations	19
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1kb

Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 13th August 2022
Internship Batch: LISUM12: 30 July - 30 October 2022
Version: 1.0
Data intake by: Joseph Nnodim
Data intake reviewer:
Data storage location:

DATA INTAKE REPORT

Tabular data details:

Transaction_ID data: The file containing this data is in csv format

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788kb

Cab_Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663

Customer_ID data: The file containing this data is in csv format

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027kb

City data: The file containing this data is in csv format

Total number of observations	19
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1kb

Data Exploration- Understanding the data

Schema

The schema was checked to know the **data types** for each of the tables, which will inform the type of analysis that will be carried out on each column. The schema analysis shows that:

In the Customer demographic table,

-Name, gender, job title, job industry category, wealth segment, deseaced indicator, owns_car, address, state and country are text data types

-Customer id, past_3_years_bike_related_purchases and tenure are integer data types

-DOB is date data types

```
CHECKING DATA TYPES


For the Transaction data table
Transaction ID      int64
Customer ID         int64
Payment_Mode       object
dtype: object
----------------------------------------
For the Customer ID data table
Customer ID              int64
Gender                  object
Age                      int64
Income (USD/Month)       int64
dtype: object
----------------------------------------
For the City data table
City          object
Population    object
Users         object
dtype: object
----------------------------------------
For the Cab_data table
Transaction ID       int64
Date of Travel       int64
Company             object
City                object
KM Travelled       float64
Price Charged      float64
Cost of Trip       float64
dtype: object
```

Data Exploration- Checking for consistency amongst tables

Customer ID occur in both the Transaction data table and the customer id table, so some checks was done to ensure the number of values for the customer IDs column in both tables

```
Ensuring equal number of data for primary keys in tables


Transaction ID    440098
Customer ID        49171
Payment_Mode           2
dtype: int64


Customer ID          49171
Gender                   2
Age                     48
Income (USD/Month)   23341
dtype: int64
```

Data Exploration- Data Quality Check continues…

Missing Values

A comprehensive exploration of the dataset was done to check for quality issues and gain a deep understanding of the properties, qualities and relationship between features in the data.

The data was checked for missing values

```
Checking for missing values


The number of missing values in each column of the Transaction table are:
Transaction ID    0
Customer ID       0
Payment_Mode      0
dtype: int64
--------------------------------------------------------------------------------
The number of missing values in each column of the Customer ID table are:
Customer ID           0
Gender                0
Age                   0
Income (USD/Month)    0
dtype: int64
--------------------------------------------------------------------------------
The number of missing values in each column of the Cab table are:
Transaction ID    0
Date of Travel    0
Company           0
City              0
KM Travelled      0
Price Charged     0
Cost of Trip      0
dtype: int64
--------------------------------------------------------------------------------
The number of missing values in each column of the City_data table are:

City          0
Population    0
Users         0
dtype: int64
```

Data Exploration- Data Quality Check continues…

Duplicate rows

The data was checked for duplicate rows

CHECKING FOR DUPLICATE ROWS

There are 0 duplicates rows in the Transaction data table
There are 0 duplicates row in the City_data table
There are 0 duplicates rows in the Customer_id table
There are 0 duplicates rows in the Cab_data table

Data Exploration- Data Quality Check continues…

Duplicate in individual columns

Individual columns were checked for duplicates

```
Checking individual columns for duplicate


For the Transaction table,
There are 0 duplicates in the Transaction ID Column
There are 390927 duplicates in the Customer ID Column
There are 440096 duplicates in the Payment_Mode Column


For the Cab data Table,
There are 0 duplicates in the 'Transaction ID' Column
There are 358297 duplicates in the 'Date of Travel' Column
There are 359390 duplicates in the 'Company' Column
There are 359373 duplicates in the 'City' Column
There are 358518 duplicates in the 'KM Travelled'Column
There are 260216 duplicates in the 'Price Charged' Column
There are 343101 duplicates in the 'Cost of Trip' Column


For the City table
There are 0 duplicates in the 'City' Column
There are 0 duplicates in the 'Population' Column
There are 0 duplicates in the 'Users' Column


For the Customer ID Table,
There are 0 duplicates in the 'Customer ID' Column
There are 49169 duplicates in the 'Gender'Column
There are 49123 duplicates in the 'Age' Column
There are 25830 duplicates in the 'Income (USD/Month)' Column
```

Data Exploration- Data Quality Check continues…

Outliers

The data was checked for the presence of outliers.

No Outlier detected

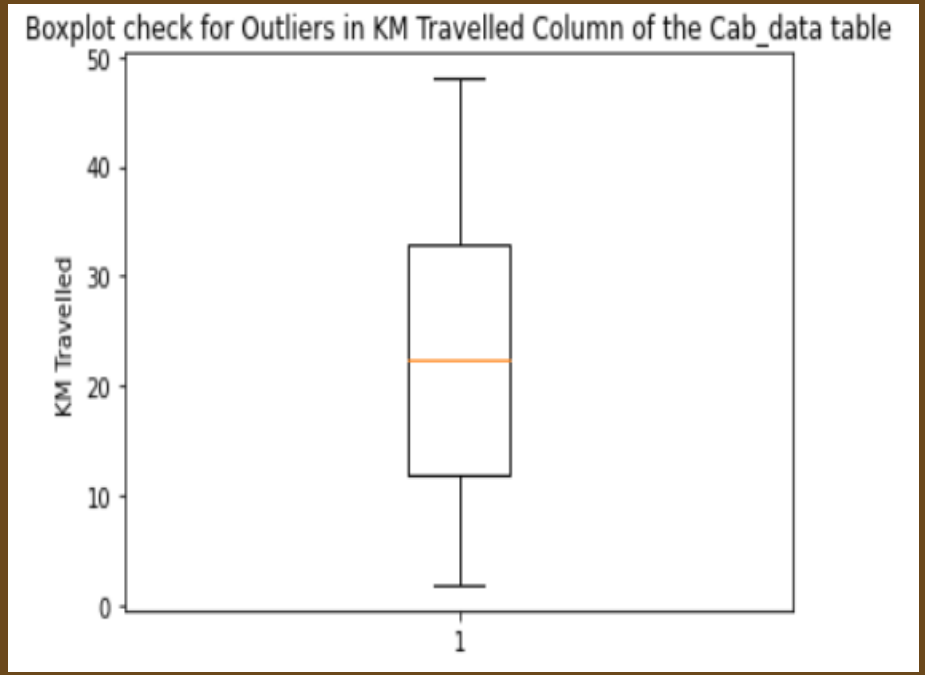
```
CHECKING FOR OUTLIERS


Checking for outliers in the 'KM Travelled' column
step 1: Calculating the first and third quartile
12.0 32.96
-----------------------------------------------------------------------------------------------------------
step 2: Calculating the Interquartile range
20.96
-----------------------------------------------------------------------------------------------------------
step 3: Calculating the lower and upper bounds
Lower bound
-19.44


Upper bound
64.4
Any number outside the range (-19.44 to 64.4) will be considered an outlier. Lets see the minimum and maximum values
-----------------------------------------------------------------------------------------------------------
Step 4: Minimum and maximum values
Minimum distance
1.9
Maximum distance
48.0
```

Data Exploration- Data Quality Check continues…

Checking Outliers for 'KM Travelled' table

The data was visualised for the presence of outliers.

No Observed extreme identified for the KM Travelled column



Data Exploration- Data Quality Check continues…

Checking Outliers for 'Price Charged' table …

The 'Price Charged' was visualized for the presence of outliers.

Extreme identified for the Price Charged

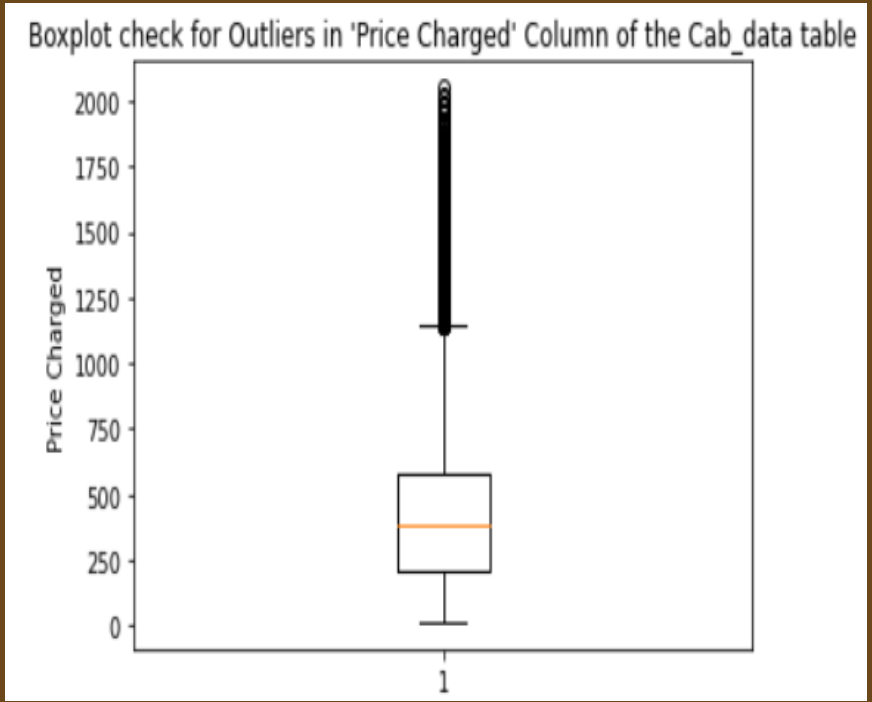
```
CHECKING FOR OUTLIERS FOR THE 'PRICE CHARGE' TABLE


Checking for outliers in the 'Price Charged' column
-----------------------------------------------------
step 1: Calculating the first and third quartile
12.0 32.96
-----------------------------------------------------------
step 2: Calculating the Interquartile range
20.96
-----------------------------------------------------
step 3: Calculating the lower and upper bounds
Lower bound
-19.44


Upper bound
64.4
-----------------------------------------------------------
Any number outside the range (-359.4 to 1149.5) will be considered an outlier. Lets see the minimum and maximum values


Step 4: Minimum and maximum values
Minimum Price
15.6


Maximum Price
2048.03
```

Data Exploration- Data Quality Check continues…

Outliers …

The ‘Price Charged’ was visualized for the presence of outliers.

Extreme identified for the Price Charged

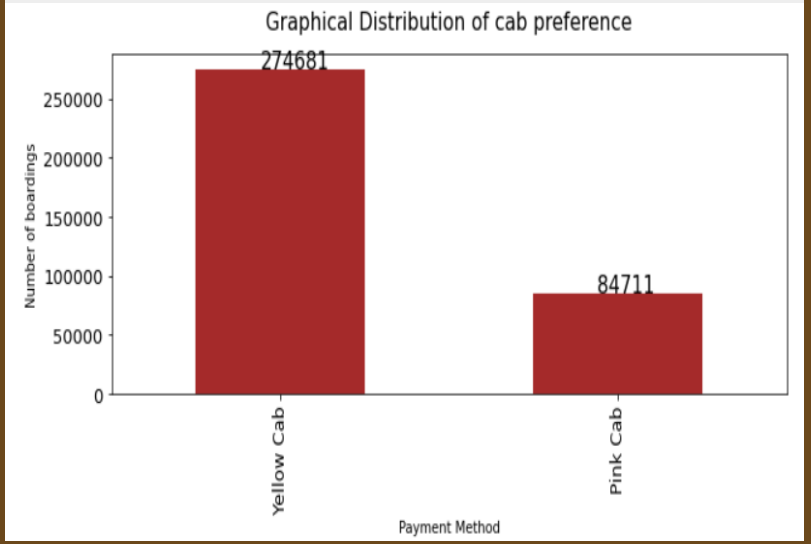


Analysis- Statistical Analysis

Comparison based on customer preference/patronage

The two companies were compared based on customer patronage

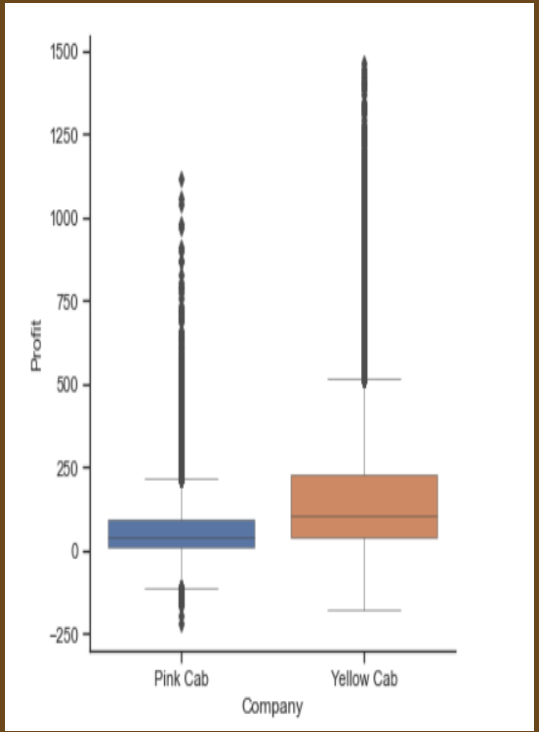
```
Yellow Cab     0.764294
Pink Cab       0.235706
Name: Company, dtype: float64
```

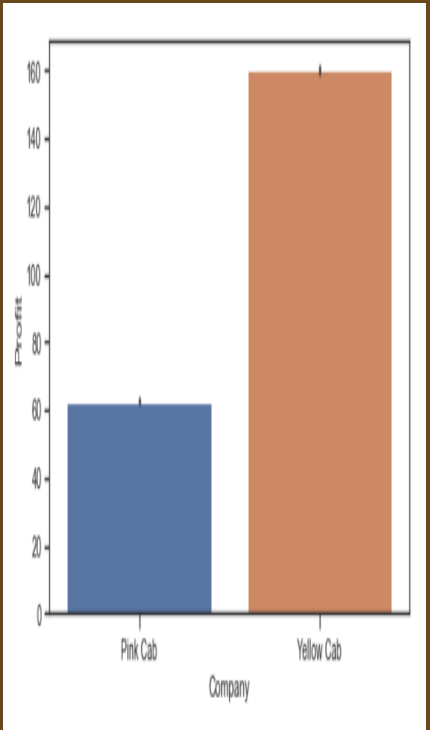


Discriptive statistics and Analysis and Interpretation

Comparison based on Mean Profit

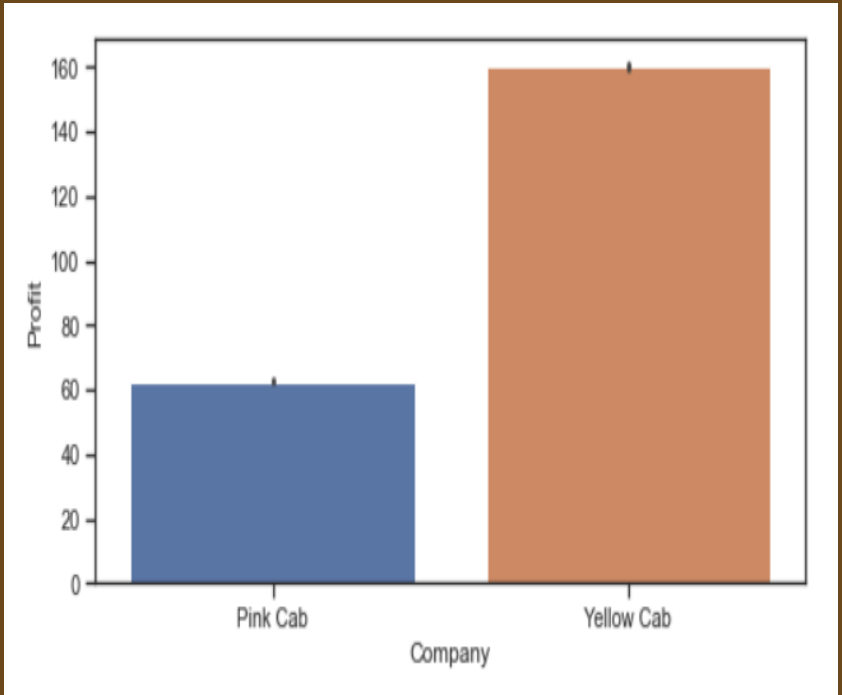
Company with yellow cab make higher mean profit (approximately 160.2) compared to company with pink cab (with profit of approximately 62.7)





Analysis and Interpretation

Based on the current analysis and the number of people who boarded the cab, the yellow cab is recommended for investment



Analysis and Interpretation

Comparison based on Mean Profit

The yellow cab company was patronized in more of the cities except in San Diego, Nashville, Sacramento and Pittsburgh (see red arrows)

Distribution of counts for cities that used Yellow cab

City	
NEW YORK NY	85918
CHICAGO IL	47264
WASHINGTON DC	40045
LOS ANGELES CA	28168
BOSTON MA	24506
SAN DIEGO CA	9816
ATLANTA GA	5795
DALLAS TX	5637
SEATTLE WA	5265
SILICON VALLEY	4722
MIAMI FL	4452
AUSTIN TX	3028
ORANGE COUNTY	2469
DENVER CO	2431
PHOENIX AZ	1200
NASHVILLE TN	1169
TUCSON AZ	1132
SACRAMENTO CA	1033
PITTSBURGH PA	631

Name: City, dtype: int64

Distribution of counts for cities that used pink cab

City	
LOS ANGELES CA	19865
NEW YORK NY	13967
SAN DIEGO CA	10672
CHICAGO IL	9361
BOSTON MA	5186
SILICON VALLEY	3797
WASHINGTON DC	3692
SEATTLE WA	2732
MIAMI FL	2002
AUSTIN TX	1868
NASHVILLE TN	1841
ATLANTA GA	1762
ORANGE COUNTY	1513
DENVER CO	1394
DALLAS TX	1380
SACRAMENTO CA	1334
PHOENIX AZ	864
TUCSON AZ	799
PITTSBURGH PA	682

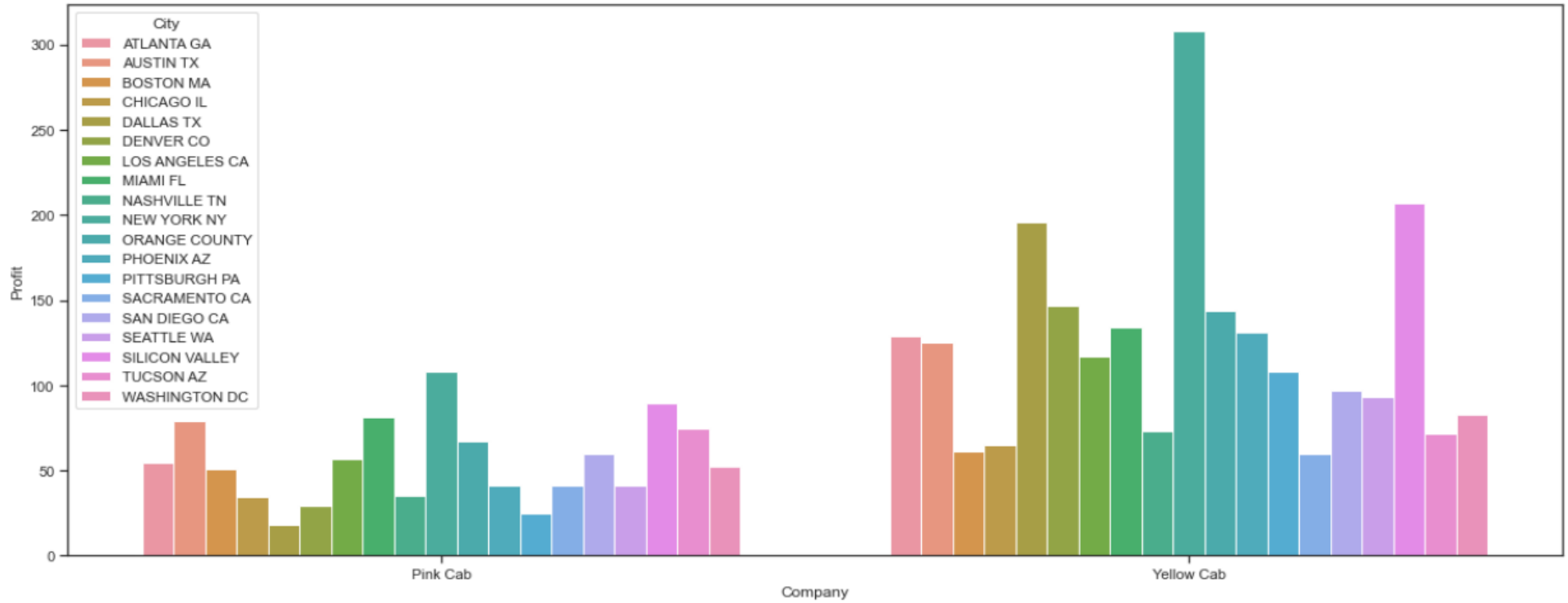
Name: City, dtype: int64

Comparison based on Mean Profit

The yellow cab company was patronized in more of the cities except in San Diego, Nashville, Sacramento and Pittsburgh (see red arrows)

Bar plt showing profit by city for each Company

AxesSubplot(0.125,0.125;0.775x0.755)



Joining Tables

-Customer data and Transaction data have the Customer ID in common.
-Transaction data and Cab data have the transaction ID in common
-Cab data and City data have the 'City' in common

	Customer ID	Gender	Age	Income (USD/Month)	Transaction ID	Payment_Mode	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Profit	Population
0	1	Male	36	16359	10000011	Card	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	57.3150	814,885
1	1	Male	36	16359	10000012	Card	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	23.6660	814,885
2	1	Male	36	16359	10000013	Cash	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.6320	27.5680	814,885
3	1	Male	36	16359	10000014	Cash	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.6020	25.7980	814,885
4	1	Male	36	16359	10000015	Card	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.7760	16.8440	814,885
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
359387	60000	Female	27	20303	10440101	Cash	43108	Yellow Cab	WASHINGTON DC	4.80	69.24	63.3600	5.8800	418,859
359388	60000	Female	27	20303	10440104	Cash	43104	Yellow Cab	WASHINGTON DC	8.40	113.75	106.8480	6.9020	418,859
359389	60000	Female	27	20303	10440105	Cash	43105	Yellow Cab	WASHINGTON DC	27.75	437.07	349.6500	87.4200	418,859
359390	60000	Female	27	20303	10440106	Card	43105	Yellow Cab	WASHINGTON DC	8.80	146.19	114.0480	32.1420	418,859
359391	60000	Female	27	20303	10440107	Card	43102	Yellow Cab	WASHINGTON DC	12.76	191.58	177.6192	13.9608	418,859

359392 rows × 15 columns

Summary of first set of results

Distribution of counts for cities that used Yellow cab

City	
NEW YORK NY	85918
CHICAGO IL	47264
WASHINGTON DC	40045
LOS ANGELES CA	28168
BOSTON MA	24506
SAN DIEGO CA	9816
ATLANTA GA	5795
DALLAS TX	5637
SEATTLE WA	5265
SILICON VALLEY	4722
MIAMI FL	4452
AUSTIN TX	3028
ORANGE COUNTY	2469
DENVER CO	2431
PHOENIX AZ	1200
NASHVILLE TN	1169
TUCSON AZ	1132
SACRAMENTO CA	1033
PITTSBURGH PA	631

Name: City, dtype: int64

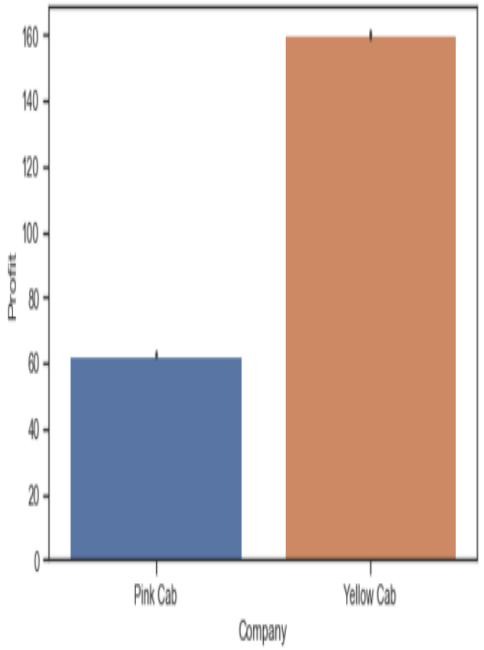
Distribution of counts for cities that used pink cab

City	
LOS ANGELES CA	19865
NEW YORK NY	13967
SAN DIEGO CA	10672
CHICAGO IL	9361
BOSTON MA	5186
SILICON VALLEY	3797
WASHINGTON DC	3692
SEATTLE WA	2732
MIAMI FL	2002
AUSTIN TX	1868
NASHVILLE TN	1841
ATLANTA GA	1762
ORANGE COUNTY	1513
DENVER CO	1394
DALLAS TX	1380
SACRAMENTO CA	1334
PHOENIX AZ	864
TUCSON AZ	799
PITTSBURGH PA	682

Name: City, dtype: int64

Mean profits for both companies

AxesSubplot(0.125,0.125;0.775x0.755)

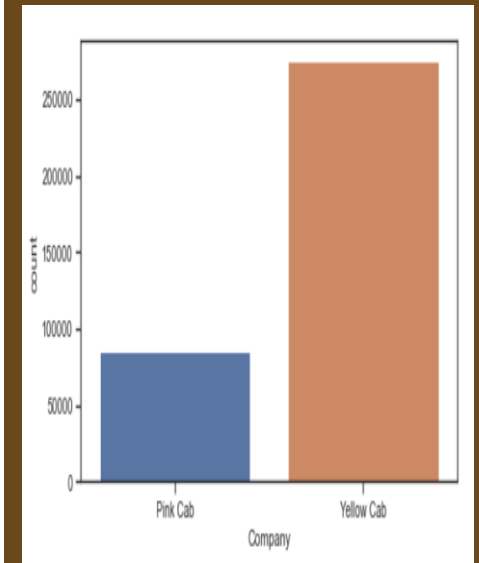


Comparing both cabs colours based on booking counts

Yellow Cab 274681
Pink Cab 84711
Name: Company, dtype: int64

Comparing both cabs colours based on PERCENTAGE of transactions

Yellow Cab 0.764294
Pink Cab 0.235706
Name: Company, dtype: float64



Hypothesis testing

Step 1: Select the appropriate statistic
We would be comparing the mean profits from the cities for the different companies so we would be using Independent samples ttest

Step 2: State the Null hypothesis
As we want to know if the mean profit generated by the yellow cab company is greater that the mean profit generated by the pink cab company, we are doing a one directional ttest; so,

Null hypothesis = Mean Profit(yellow) > Mean Profit(pink)
Alternative hypothesis = Mean Profit(yellow) ≤ Mean Profit(pink)

```
from scipy.stats import ttest_ind
ttest_ind(yellowcab_meanprofit, pinkcab_meanprofit)
```

```
Ttest_indResult(statistic=4.573525705863568, pvalue=5.4866370522558754e-05)
```

Step 3: Select a level of significance
#Two tailed test; a=0.05, df=36; i.e

Step 4: Calculate the statistics
We assume that the 2 samples have approximately equal variances

Step 5: Make a decision
Since **pvalue<0.05**, the test id **significant**
Conclusion: Company with yellow cab generates higher mean profit than

Hypothesis testing

Result of ttest

```
from scipy.stats import ttest_ind
ttest_ind(yellowcab_meanprofit, pinkcab_meanprofit)
```

```
Ttest_indResult(statistic=4.573525705863568, pvalue=5.4866370522558754e-05)
```

Recommendation

Since p value<0.05, we reject the null hypothesis and conclude that the mean profit from the yellow cab company in different states is significantly different from those from the pink cab company in different states

Based on the result, the company with the yellow cab is recommeded

Assumptions

Remember, null hypothesis states that there is no statistical difference between the Profits for different states

Since p value<0.05, we reject the null hypothesis and conclude that the mean profit from the yellow cab

company in different states is significantly different from those from the pink cab company in different states

Assumptions

-Profit is only dependent on cost of trip and price charged

- The data in each subgroup for each of the categories are free normally distributes

- Each company operates in at least 30 states. This is one of the assumptions of independent sample ttest