Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Case Study

Virtual Internship

21/09/24

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Data Glacier
Your Deep Learning Partner

Executive Summary

XYZ is a private equity 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.

Objective : Provide actionable insights to help XYZ firm in identifying the right company for making investment.

The analysis has been divided into four parts:

Data Understanding

Profit Calculation and number of rides for each cab type in 2016,2017,2018

Finding the most profitable Cab company

Recommendations for investment

Data Glacier

Your Deep Learning Partner

Problem Statement

- 19 Features(including 2 derived features)
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points :3,59,393

Assumptions:

- Outliers are present in Price_Charged feature but due to unavailability of trip duration details ,It's not considered as outlier.

- Profit of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit in Microsoft Excel.

- Users feature of city dataset is treated as number of cab users in the city. It's been assumed that this can be other cab users as well(including Yellow and Pink cab)
- The attributes for customer segments considered for analysis are Age, Gender, Income level, Geographic attributes city, and behavioral Attribute i.e., Purchasing mode

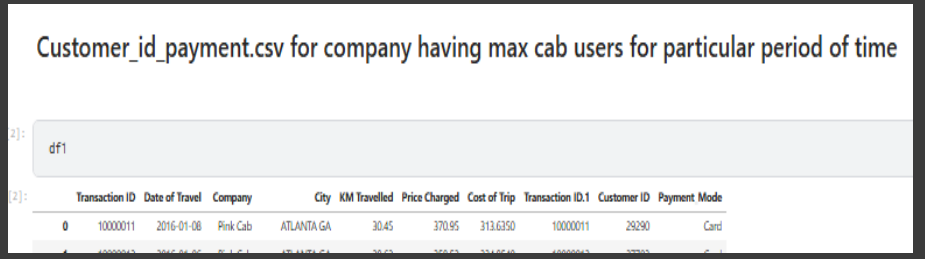
Data Glacier
Your Deep Learning Partner

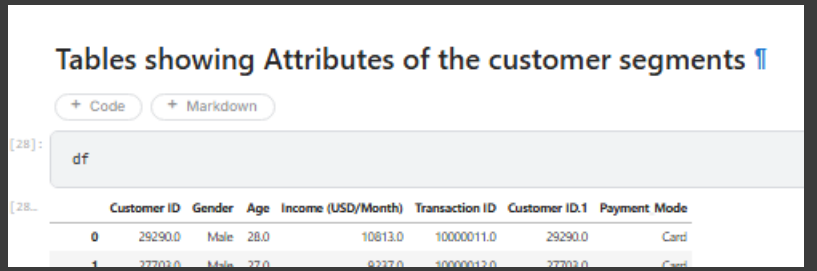
Approach

Two consolidated tables embedded in Python notebook

1.Customer_id_payment.csv for company having max cab users for particular period of time

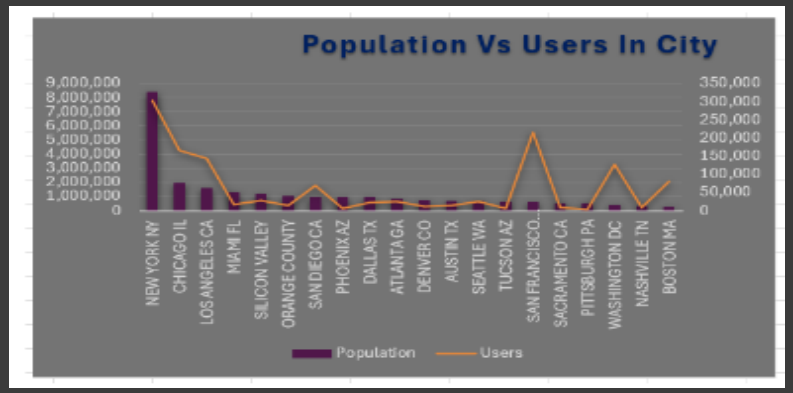
2.Tables showing Attributes of the customer segments¶

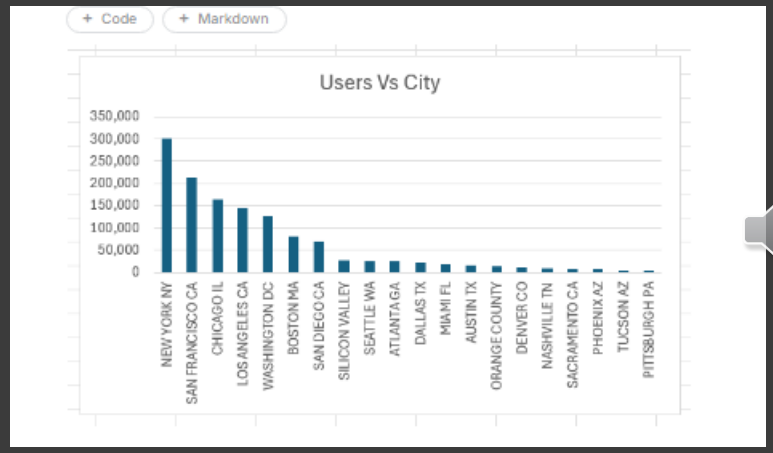




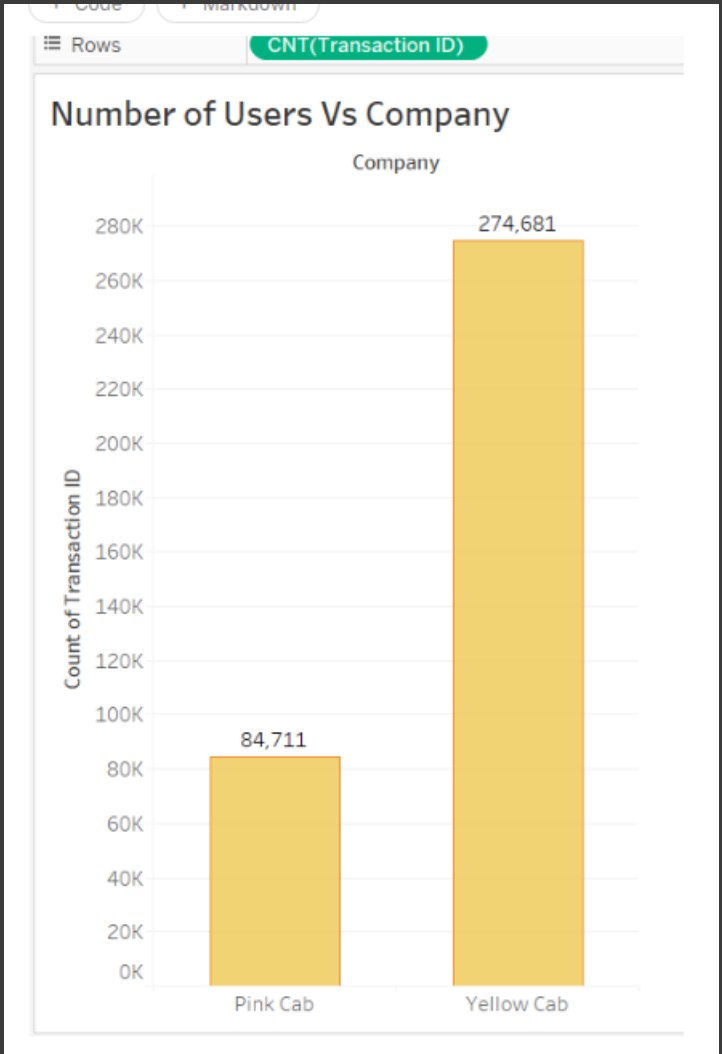
Univariante Analysis to determine Customers in City Newyork have more Cab users

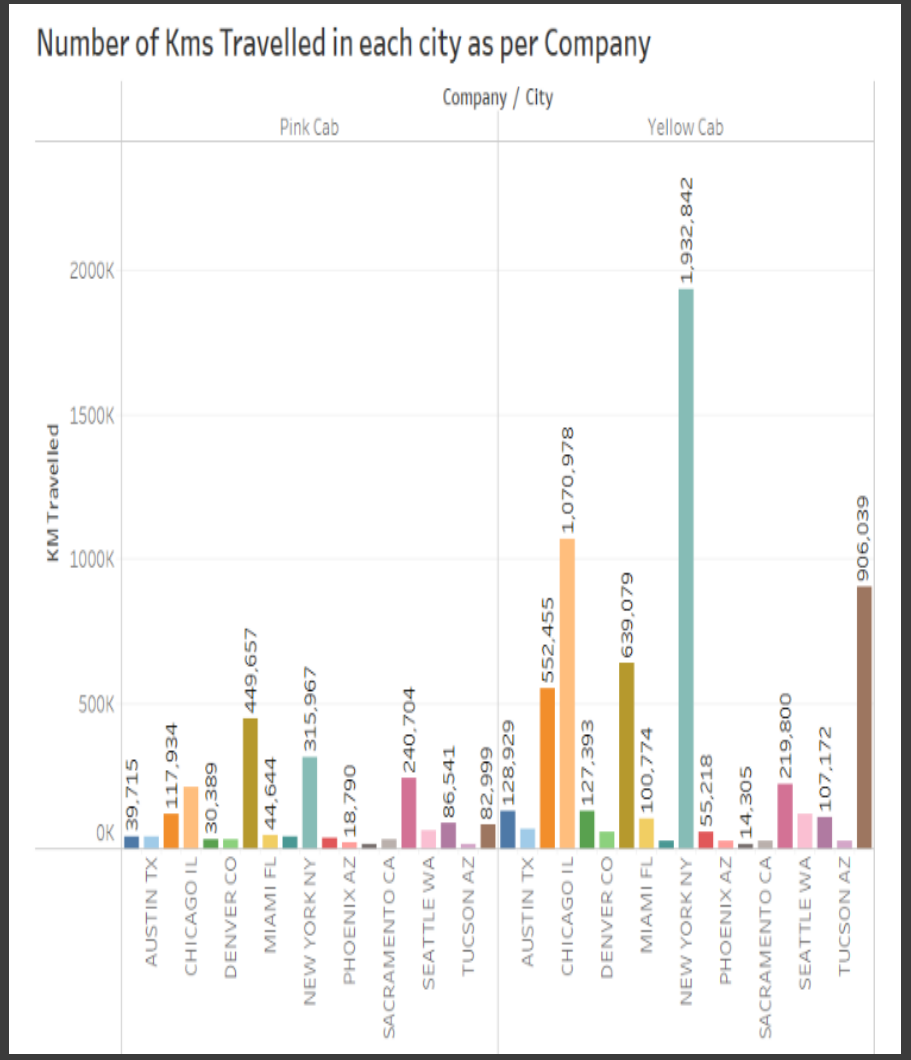
Bivariante Analysis to determine Population Vs Cab users in City





Exploratory Data Analysis
Cab Company Analysis

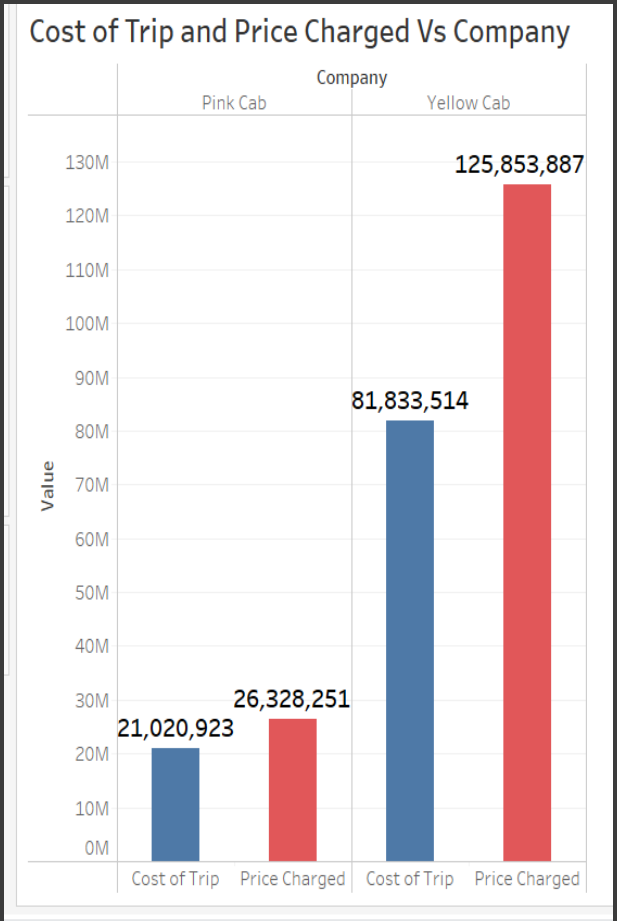




- Yellow Cab have a greater number of users than Pink Cab
- Yellow Cab users – 274,681
- Pink Cab users -84711

- Customer Base in each city for Yellow Cab Company and Pink Cab Company

Exploratory Data Analysis
Profit Analysis



	F	G	H	I	J	K	L
	Price Charged	Cost of Trip	Transaction ID	Customer	Payment_Mode	Profit	
30.45	370.95	313.635	10000011	29290	Card	57.315	
28.62	358.52	334.854	10000012	27703	Card	23.666	
9.04	125.2	97.632	10000013	28712	Cash	27.568	
33.17	377.4	351.602	10000014	28020	Cash	25.798	
8.73	114.62	97.776	10000015	27182	Card	16.844	
6.06	72.43	63.024	10000016	27318	Cash	9.406	
44	576.15	475.2	10000017	33788	Card	100.95	
35.65	466.1	377.89	10000018	34106	Card	88.21	
14.4	191.61	146.88	10000019	59799	Cash	44.73	
10.89	156.98	113.256	10000020	57982	Cash	43.724	
39.6	570.83	475.2	10000021	58774	Cash	95.63	
21.8	317.27	220.18	10000022	58627	Card	97.09	
12	158.01	134.4	10000023	59007	Card	23.61	
32.67	470.94	392.04	10000024	58215	Cash	78.9	
25.52	360.79	298.584	10000025	59372	Cash	62.206	
15.54	234.78	169.386	10000026	57950	Cash	65.394	
6.65	101.39	76.475	10000027	59533	Card	24.915	
34.22	498.02	407.218	10000028	58346	Cash	90.802	
21.34	324.21	226.204	10000029	58925	Card	98.006	
41.3	646.06	454.3	10000030	58551	Card	191.76	
1.96	27.71	20.776	10000031	59804	Card	6.934	
13.44	193.74	145.152	10000032	57200	Card	48.588	
23.4	295.15	259.74	10000033	57931	Cash	35.41	
23.2	396.73	278.4	10000034	3077	Cash	118.33	
4.48	55.27	51.52	10000035	4734	Card	3.75	
37.76	541.76	407.808	10000036	4004	Card	133.952	

- Profit is calculated from the difference between Cost of Trip and Price charged
- Profit Margin Proportionally increase with increasing number of Customers

Customer base Analysis Gender wise

df

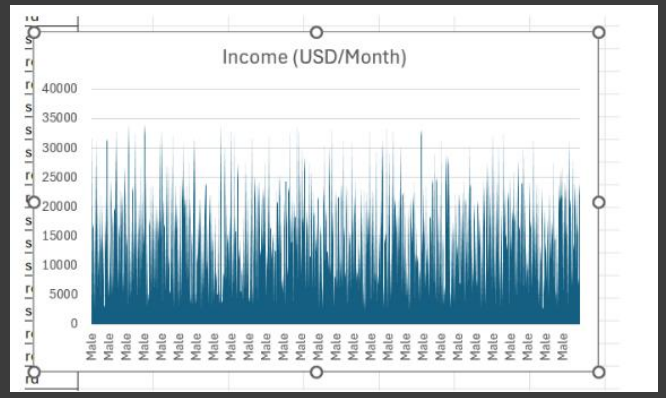
	Customer ID	Gender	Age	Income (USD/Month)	Transaction ID	Customer ID.1	Payment_Mode
0	29290.0	Male	28.0	10813.0	10000011.0	29290.0	Card
1	27703.0	Male	27.0	9237.0	10000012.0	27703.0	Card
2	28712.0	Male	53.0	11242.0	10000013.0	28712.0	Cash
3	28020.0	Male	23.0	23327.0	10000014.0	28020.0	Cash
4	27182.0	Male	33.0	8536.0	10000015.0	27182.0	Card
...	...	...	...	...	...	...	...
440094	NaN	NaN	NaN	NaN	10440104.0	53286.0	Cash
440095	NaN	NaN	NaN	NaN	10440105.0	52265.0	Cash
440096	NaN	NaN	NaN	NaN	10440106.0	52175.0	Card
440097	NaN	NaN	NaN	NaN	10440107.0	52917.0	Card
440098	NaN	NaN	NaN	NaN	10440108.0	51587.0	Card

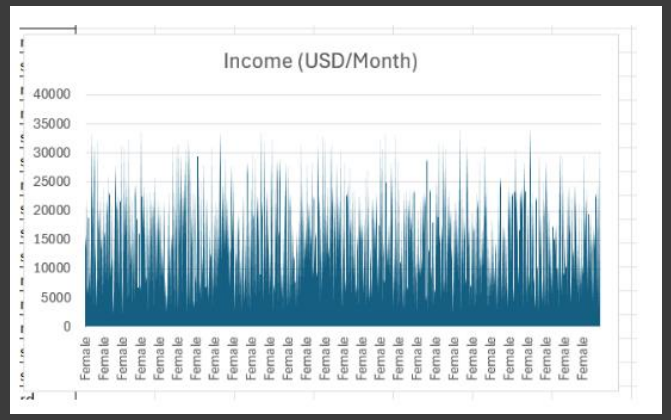
440099 rows × 7 columns

+ Code + Markdown

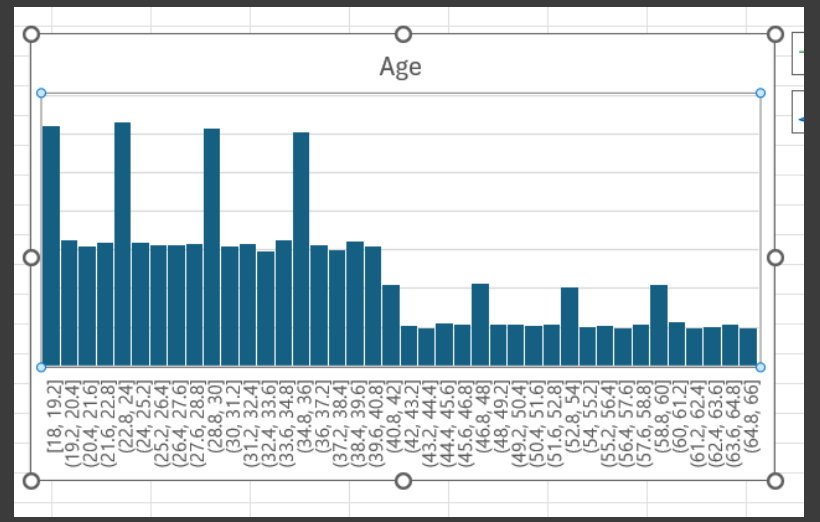
```
count_A = df['Gender'].value_counts().get('Male', 0)
count_B = df['Gender'].value_counts().get('Female', 0)

print(f"Count of 'Male': {count_A}")
print(f"Count of 'Female': {count_B}")
```

```
Count of 'Male': 26562
Count of 'Female': 22609
```



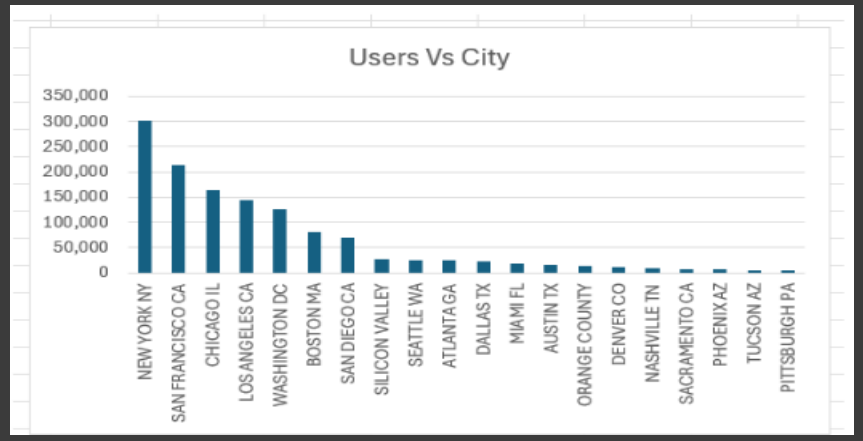


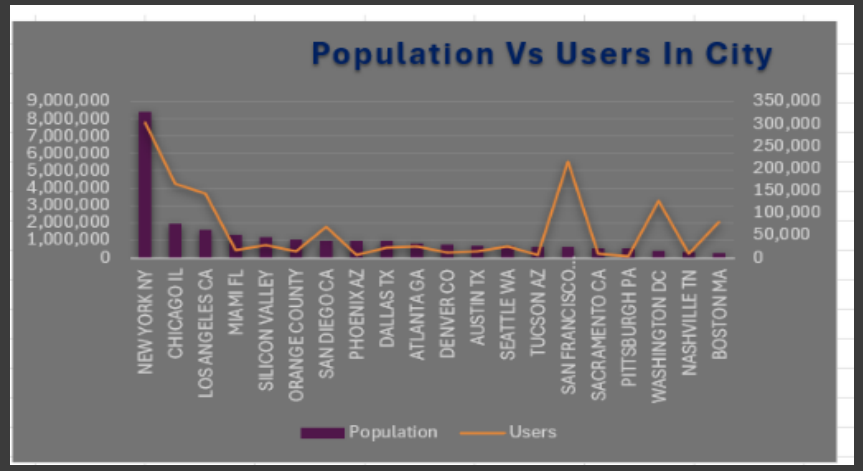
Count of Payment_Mode	Column Labels			
Row Labels	Card	Cash	(blank)	Grand Total
Female	13388	9221		22609
Male	15944	10618		26562
(blank)	234659	156268		390927
Grand Total	**263991**	**176107**		**440098**



Citywide customer base Analysis

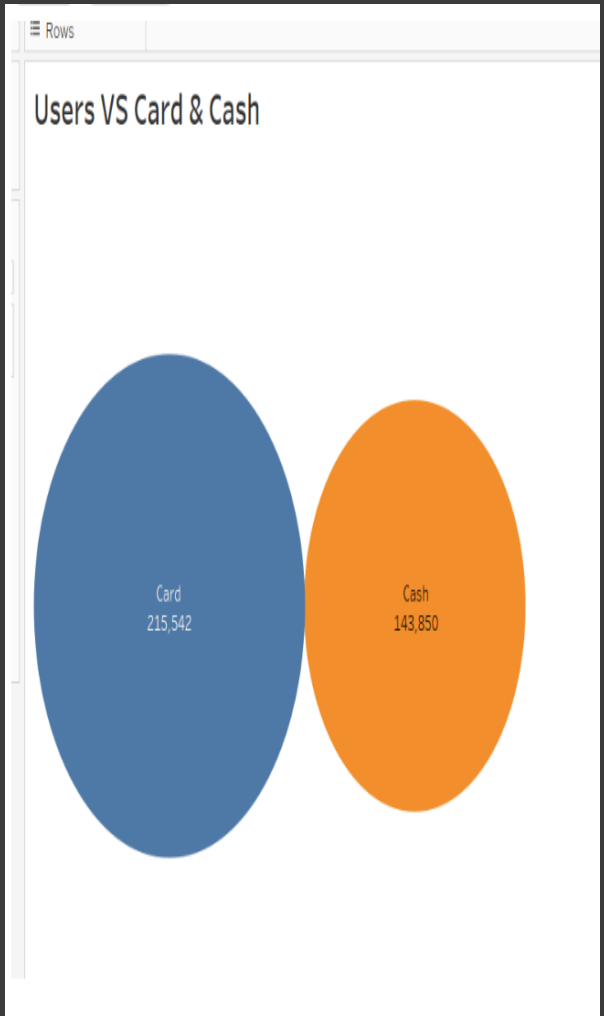
City	Population	Users
NEW YORK NY	8,405,837	302,149
CHICAGO IL	1,955,130	164,468
LOS ANGELES CA	1,595,037	144,132
MIAMI FL	1,339,155	17,675
SILICON VALLEY	1,177,609	27,247
ORANGE COUNTY	1,030,185	12,994
SAN DIEGO CA	959,307	69,995
PHOENIX AZ	943,999	6,133
DALLAS TX	942,908	22,157
ATLANTA GA	814,885	24,701
DENVER CO	754,233	12,421
AUSTIN TX	698,371	14,978
SEATTLE WA	671,238	25,063
TUCSON AZ	631,442	5,712
SAN FRANCISCO CA	629,591	213,609
SACRAMENTO CA	545,776	7,044
PITTSBURGH PA	542,085	3,643
WASHINGTON DC	418,859	127,001
NASHVILLE TN	327,225	9,270
BOSTON MA	248,968	80,021

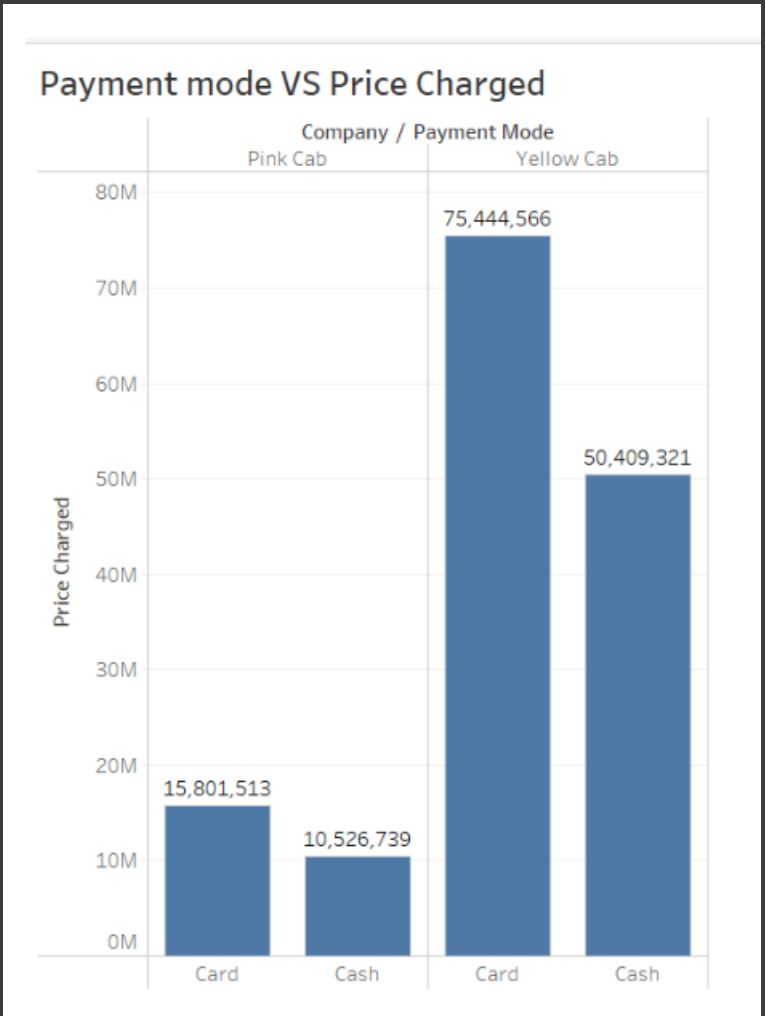




- Visualization to analyse Cab Users in City
- Also to analyse the ratio of Population against Cab users in City

Customer Base in terms of Payment Mode





- Cab users Payment mode shows users of Card is greater than that of Cash payers
- Also, Yellow Cab Company has Greater margin of Card USERS.

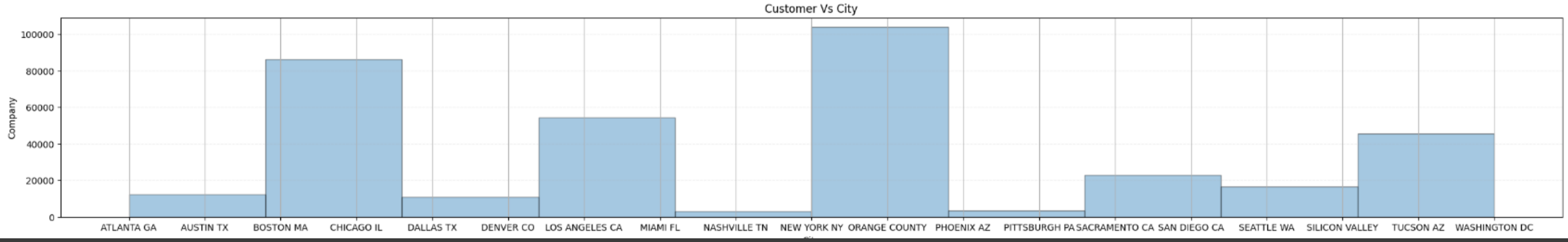
City Wise Cab Users Covered By Company

- Number of Kms Travelled for Each Cab Company have increased from 2016 to 2017.
- A slighter increase in the trend was found in 2018.
- Yellow cab have greater number of Kms Travelled

Customer Presence of cab city wise

Newyork have more Cab users

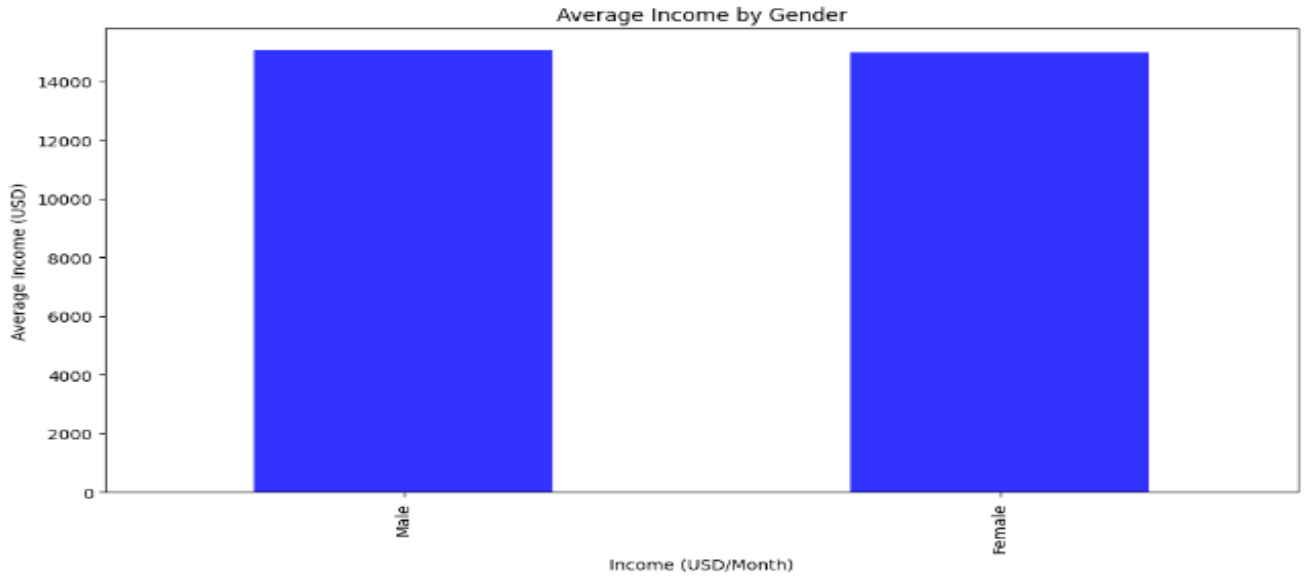
```
import pandas as pd
import matplotlib.pyplot as plt

# Assuming 'df1' is your DataFrame
plt.figure(figsize=(30, 4))  # Optional: Set the figure size
df1['City'].hist(bins=10, edgecolor='black', alpha=0.4)
plt.xlabel('City')
plt.ylabel('Company')
plt.title('Customer Vs City')
plt.grid(axis='y', linestyle='--', alpha=0.2)
plt.show()
```



Univariant Analysis to determine Average Income by Gender

Univariante Analysis to determine Average income by gender

```
plt.figure(figsize=(12, 6))
avg_income_by_gender = df.groupby('Gender')['Income (USD/Month)'].mean().sort_values(ascending=False)[:10]
avg_income_by_gender.plot(kind='bar', color='blue', alpha=0.8)
plt.title('Average Income by Gender')
plt.xlabel('Income (USD/Month)')
plt.ylabel('Average Income (USD)')
plt.show()
```

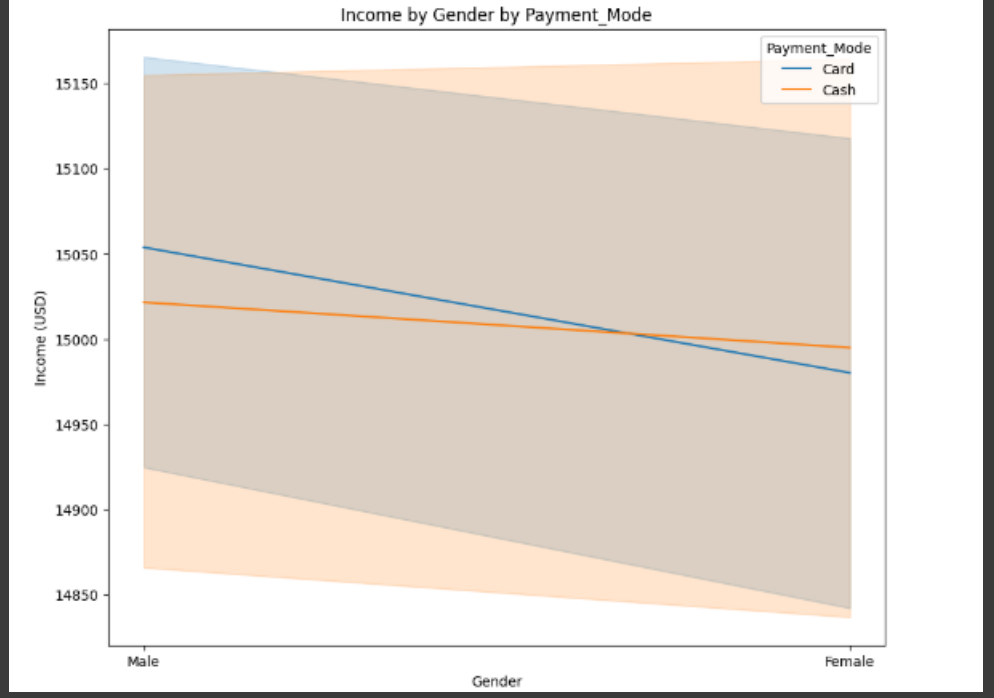


- Analysis performed for the Gender Column to calculate average income in Male And Female that's been observed from Groupby function

User Covered by Company and Customer base Year wise

```
plt.figure(figsize = (10, 8))
sns.lineplot(x = "Gender", y = "Income (USD/Month)", data = df, hue = "Payment_Mode")

plt.title("Income by Gender by Payment_Mode")
plt.xlabel("Gender")
plt.ylabel("Income (USD)")
plt.show()
```

```
/opt/conda/lib/python3.10/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will
  with pd.option_context('mode.use_inf_as_na', True):
/opt/conda/lib/python3.10/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will
  with pd.option_context('mode.use_inf_as_na', True):
/opt/conda/lib/python3.10/site-packages/seaborn/_oldcore.py:1075: FutureWarning: When grouping with a length-1 list-like, yo
  data_subset = grouped_data.get_group(pd_key)
/opt/conda/lib/python3.10/site-packages/seaborn/_oldcore.py:1075: FutureWarning: When grouping with a length-1 list-like, yo
  data_subset = grouped_data.get_group(pd_key)
```



- The Analysis in Python code details the payment behaviour observed among the Male and Female of the Gender Column.
- Also, it details Income range for the Card and Cash Payment mode Card Users

Day Frequency in Rides Vs Year

- The visualisation details the average Day frequency rides Vs Year
- The Number of customers undertaking rides is been increased in 2017&2018.
- There is decline in trend during the last days of months in each year
- Most Importantly there is increase in rides during 5th day of the month

Summary

The Client

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.

From the above hypothesis the Answers for Questions are found to be

Which company has maximum cab users at a particular time period?

Yellow Cab have more number of users than Pink Cab
Yellow Cab users - 274682
Pink Cab users - 84712

Does margin proportionally increase with increase in number of customers?

Yes Margin Proportionally increase with increase in number of customers.
Yellow Cab shows Marginal increase proportionally in more number of customers

What are the attributes of these customer segments?

The below attributes for customer segments are considered

for analysis

Demographic Attributes:

Age

Gender

Income level

Geographic Attributes:

Location (country, region, city)

Behavioral Attributes:

Purchasing behavior

Recommendations

I have evaluated both the cab companies on following points and found Yellow cab better than Pink cab:

- **Customer Reach :** Yellow cab has higher customer reach in 25 cities while Pink cab has higher customer reach in 4 cities. We have also observed that Yellow cab is doing good in covering other cab users as compared to Pink cab.
- **Customer Retention:** Yellow Cab has higher customer reach**.**
- **Age wise Reach :** Cab Company has customer in all age group and it's been observed that it's even popular in 60+ age group as equally as its in 18-35 age group.
- **Average Profit per KM:** Yellow cab's average profit per KM is almost three times the average profit per KM of the Pink cab.
- **Income wise Reach :**Both the cabs are very popular in high and medium income class but here also Yellow cab is performing better than Pink cab in offering their services to all the three income class group (low, medium and high)

On the basis of above point , I will recommend Yellow cab for investment.

Thank You