Data Glacier

Your Deep Learning Partner

Virtual Internship

G2M INSIGHT FOR CAB INVESTMENT FIRM

DATA ANALYSIS: AN COMPREHENSIVE REPORT

Presented by Alison March

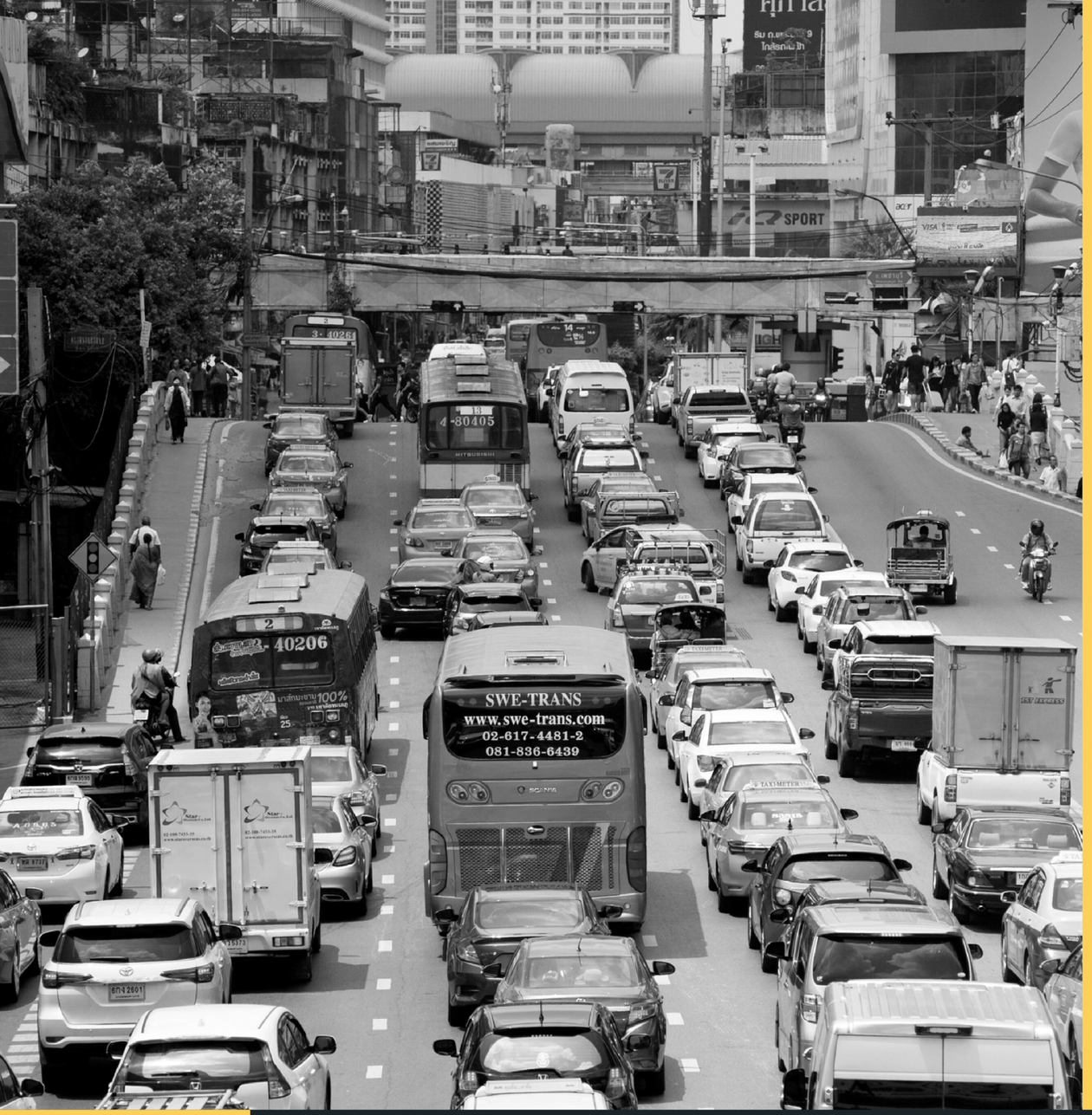


Table of Contents

Topics to be Discussed

- Business Problem & Objectives
- About the Data Source & Data Structure
- Data Intake & Preliminary Analysis
- Compare Companies from data perspectives
- Proposed Recommendations
- Conclusion

Presented by Alison March

Business Problem

Introduction

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.

Presented by Alison March

Project Objective:

TO GENERATE INSIGHTS TO HELP XYZ IDENTIFY THE RIGHT COMPANY TO MAKE THEIR INVESTMENT.

Presented by Alison March

Cab Investment:

Data Exploration & Analysis:

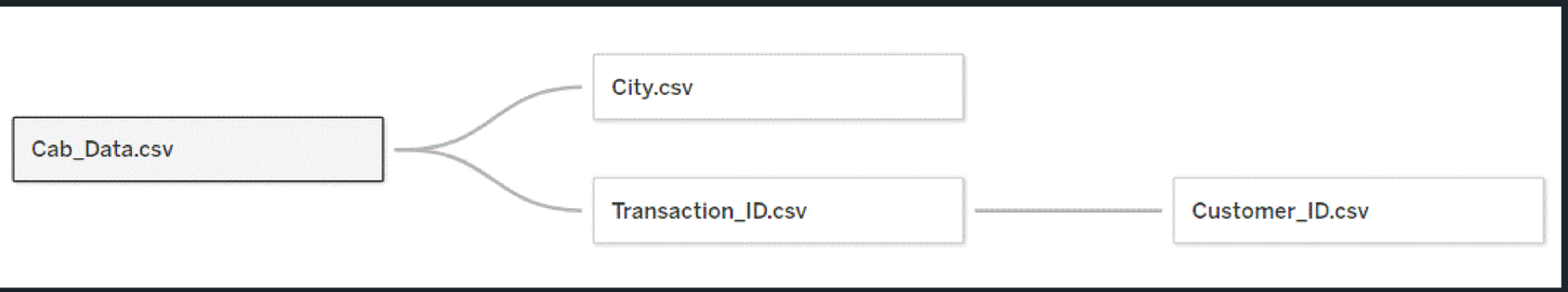
- In next sections we will analyze data in the following process:
  a. **Data Intake and Preliminary Processing**
  b. **Compare Yellow Cab and Pink Cab from different aspects**
  c. **Investment Recommendation**

Presented by Alison March

I. Data Intake
&
Preliminary Analysis

Data Files and Structure

- Cab_Data
- City
- Customer_ID
- Transaction_ID



Cab Data

Transaction ID* | Date of Travel | Company | City* | KM Traveled | Price Charged | Cost of Trip

Transaction_ID

Transaction ID* | Customer ID* | Payment_Mode

Customer_ID

Customer ID* | Gender | Age | Income (USD/Month)

City

City* | Population | Users

First, we imported all 4 datasets, and check any null values. Then we concatenate datasets by the unique identifiers:

- Join Cab_Data.csv with City.csv by "City" = df1 (dataframe 1)
- Join Transaction_ID.csv with Customer_ID.csv by "Customer ID" = df2 (dataframe 2)
- Join df1 with df2 by "Transaction ID"

Final Dataframe(df) **359,392** Row by **14** Columns

	Transaction ID	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Population	Users
0	10000011	29290	Card	Male	28	10813	1/8/2016	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	814,885	24,701
1	10351127	29290	Cash	Male	28	10813	7/21/2018	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	814,885	24,701
2	10412921	29290	Card	Male	28	10813	11/23/2018	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	814,885	24,701
3	10000012	27703	Card	Male	27	9237	1/6/2016	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	814,885	24,701
4	10320494	27703	Card	Male	27	9237	4/21/2018	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	814,885	24,701
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
359387	10439790	38520	Card	Female	42	19417	1/7/2018	Yellow Cab	SEATTLE WA	16.66	261.18	213.9144	671,238	25,063
359388	10439799	12490	Cash	Male	33	18713	1/3/2018	Yellow Cab	SILICON VALLEY	13.72	277.97	172.8720	1,177,609	27,247
359389	10439838	41414	Card	Male	38	3960	1/4/2018	Yellow Cab	TUCSON AZ	19.00	303.77	232.5600	631,442	5,712
359390	10439840	41677	Cash	Male	23	19454	1/6/2018	Yellow Cab	TUCSON AZ	5.60	92.42	70.5600	631,442	5,712
359391	10439846	39761	Card	Female	32	10128	1/4/2018	Yellow Cab	TUCSON AZ	13.30	244.65	180.3480	631,442	5,712

359392 rows × 14 columns

Presented by Alison March

Data Type Conversion:

In [26]:

```
# Convert Data types in the dataframe df
df['Population'] = df['Population'].str.replace(',','')
df['Users'] = df['Users'].str.replace(',','')

#Convert objects to integer datatypes
df['Population'] = df['Population'].astype(int)
df['Users'] = df['Users'].astype(int)
```

Data Types:

```
df.dtypes
```

```
Transaction ID          int64
Customer ID             int64
Payment_Mode           object
Gender                 object
Age                     int64
Income (USD/Month)      int64
Date of Travel         object
Company                object
City                   object
KM Travelled          float64
Price Charged         float64
Cost of Trip          float64
Population              int32
Users                   int32
dtype: object
```

Overall Glance:

In [30]:

```
df.describe()
```

Out[30]:

	Transaction ID	Customer ID	Age	Income (USD/Month)	KM Travelled	Price Charged	Cost of Trip	Population	Users
count	3.593920e+05	359392.000000	359392.000000	359392.000000	359392.000000	359392.000000	359392.000000	3.593920e+05	359392.000000
mean	1.022076e+07	19191.652115	35.336705	15048.822937	22.567254	423.443311	286.190113	3.132198e+06	158365.582267
std	1.268058e+05	21012.412463	12.594234	7969.409482	12.233526	274.378911	157.993661	3.315194e+06	100850.051020
min	1.000001e+07	1.000000	18.000000	2000.000000	1.900000	15.600000	19.000000	2.489680e+05	3643.000000
25%	1.011081e+07	2705.000000	25.000000	8424.000000	12.000000	206.437500	151.200000	6.712380e+05	80021.000000
50%	1.022104e+07	7459.000000	33.000000	14685.000000	22.440000	386.360000	282.480000	1.595037e+06	144132.000000
75%	1.033094e+07	36078.000000	42.000000	21035.000000	32.960000	583.660000	413.683200	8.405837e+06	302149.000000
max	1.044011e+07	60000.000000	65.000000	35000.000000	48.000000	2048.030000	691.200000	8.405837e+06	302149.000000

Check missing values:

In [27]:

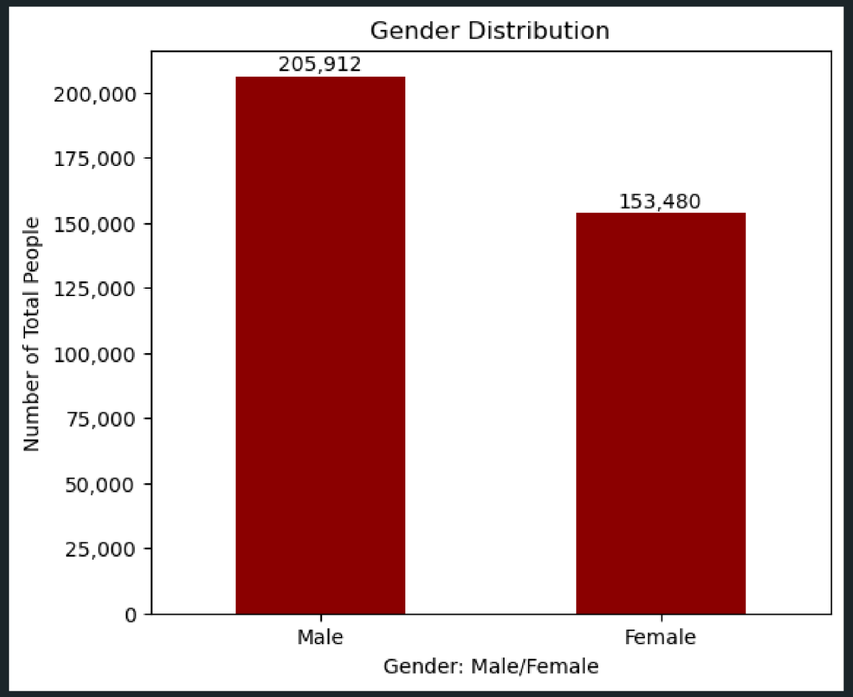
```
# Check any missing values
df.isnull().sum().sum()
```

Out[27]: 0

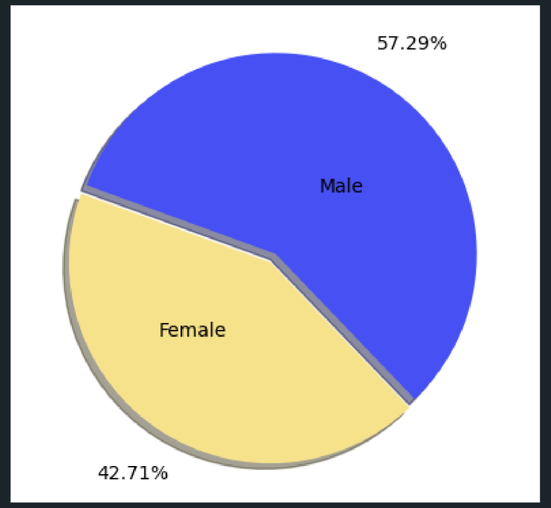
Presented by Alison March

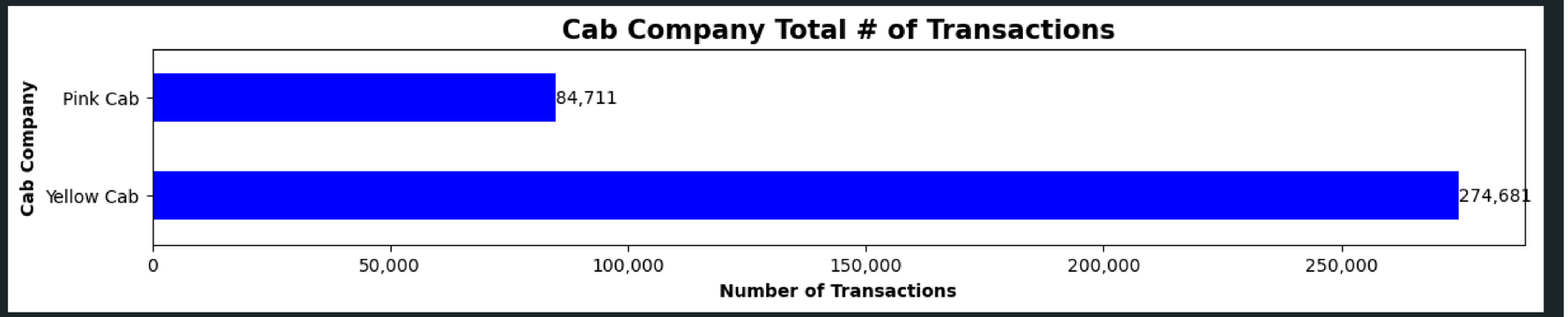
II. Compare Yellow Cab and Pink Cab from different aspects

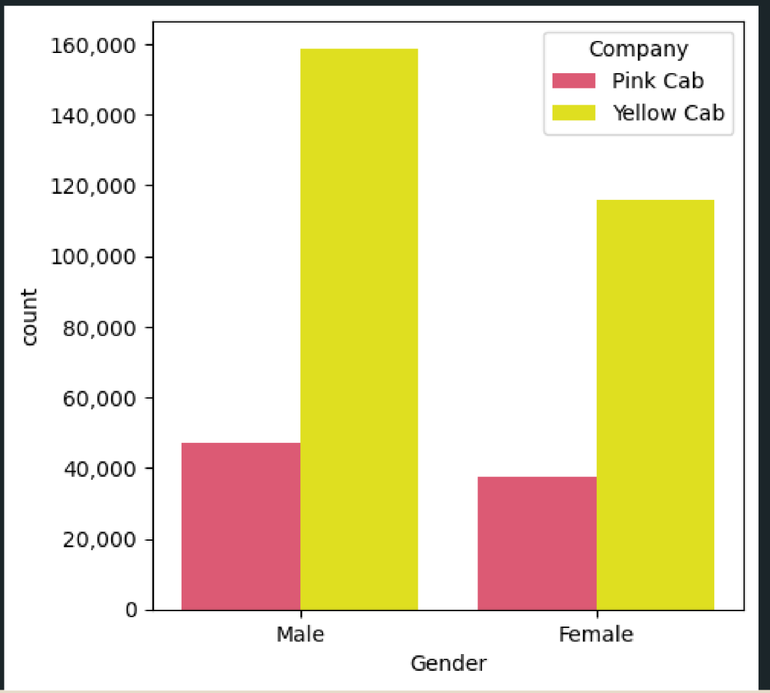
General statistics:



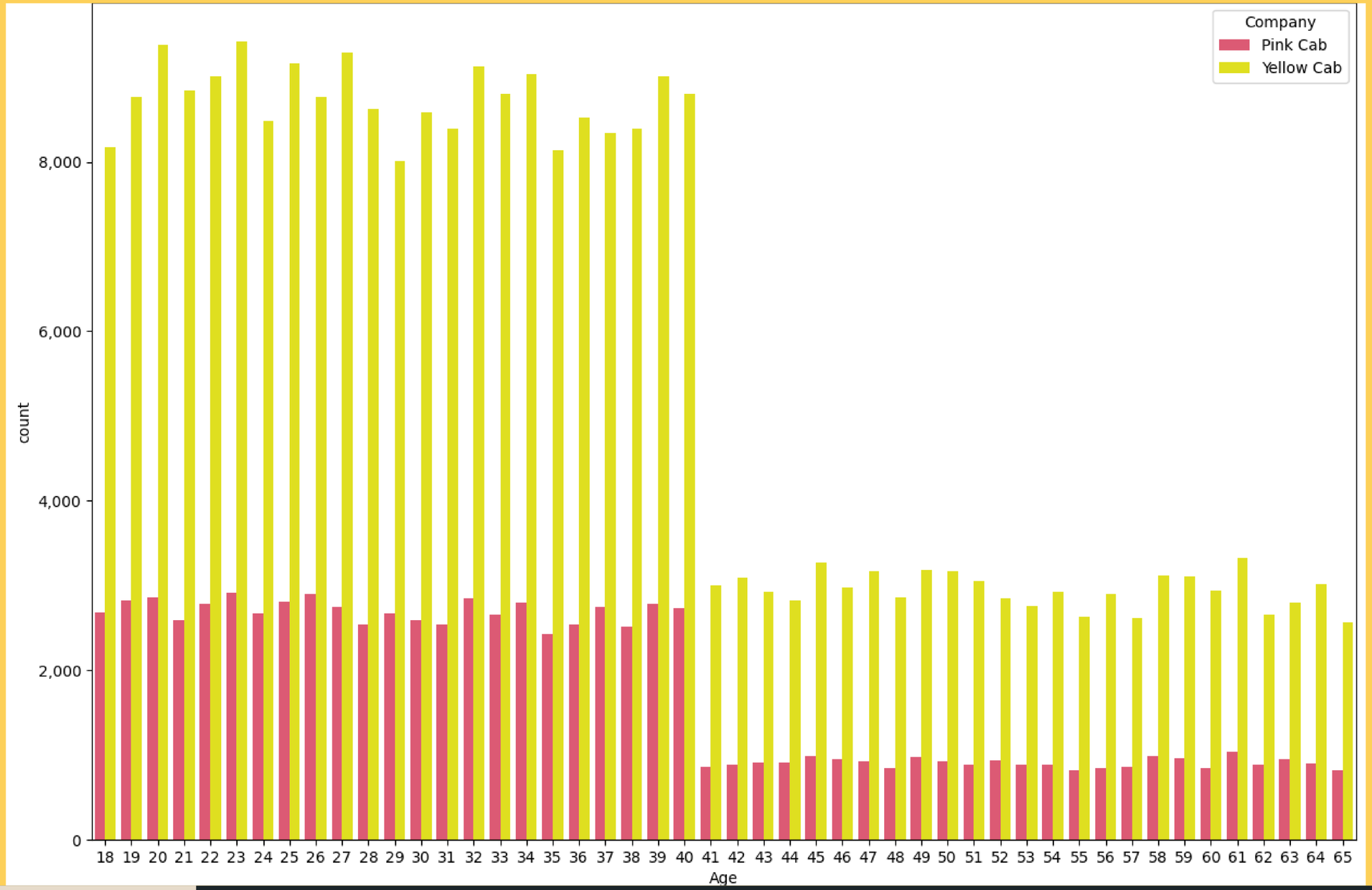
- There are a total of 359,392 Transactions
- Pink Cab has 84,711 Transactions (23.57%)
- Yellow Cab has 274,681 Transactions(76.43%)
- 205,912 passengers are males, 153,480 are female
- The ratio of Male to Female is 57.3% vs. 42.7%







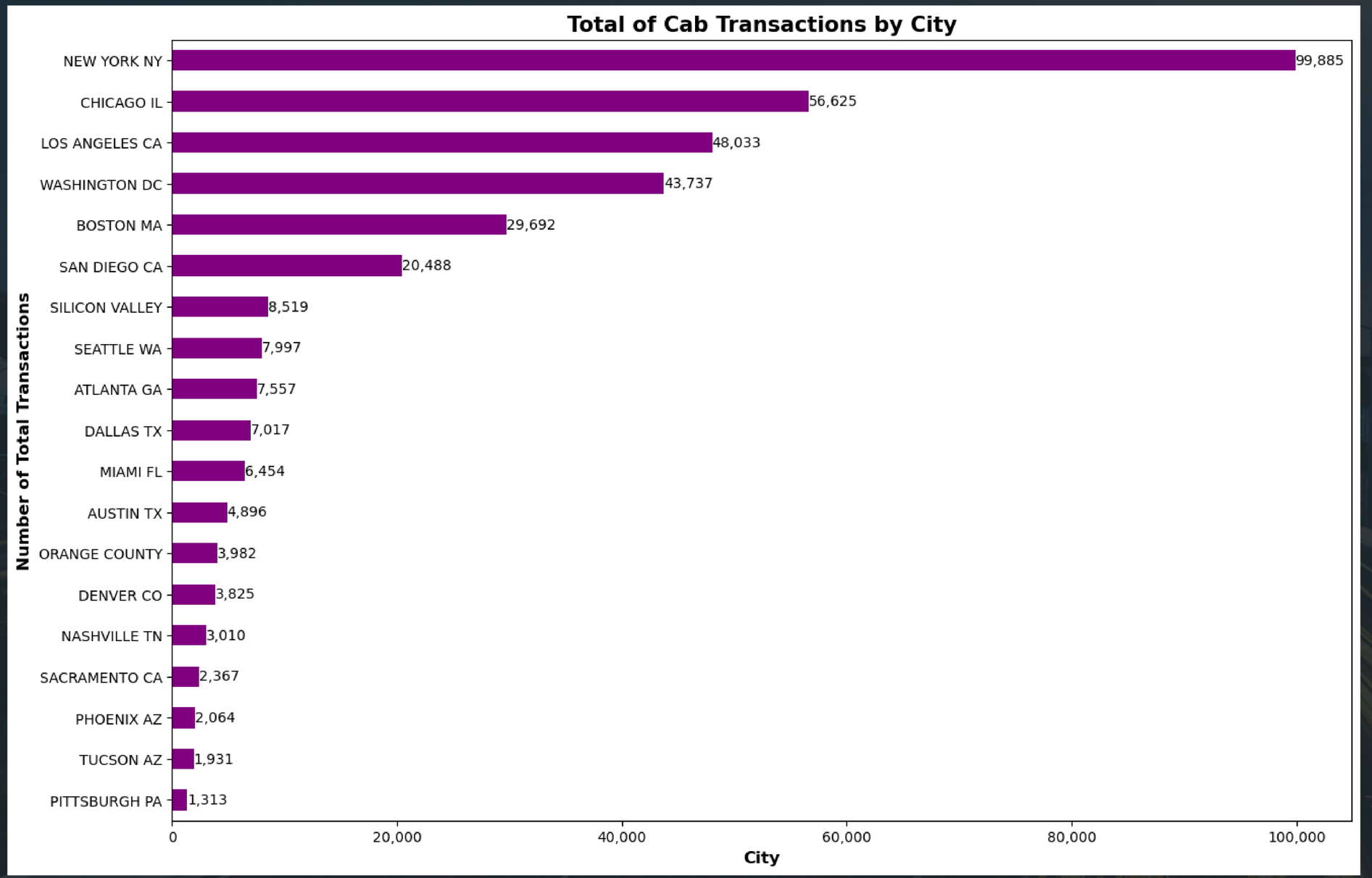
Presented by Alison March



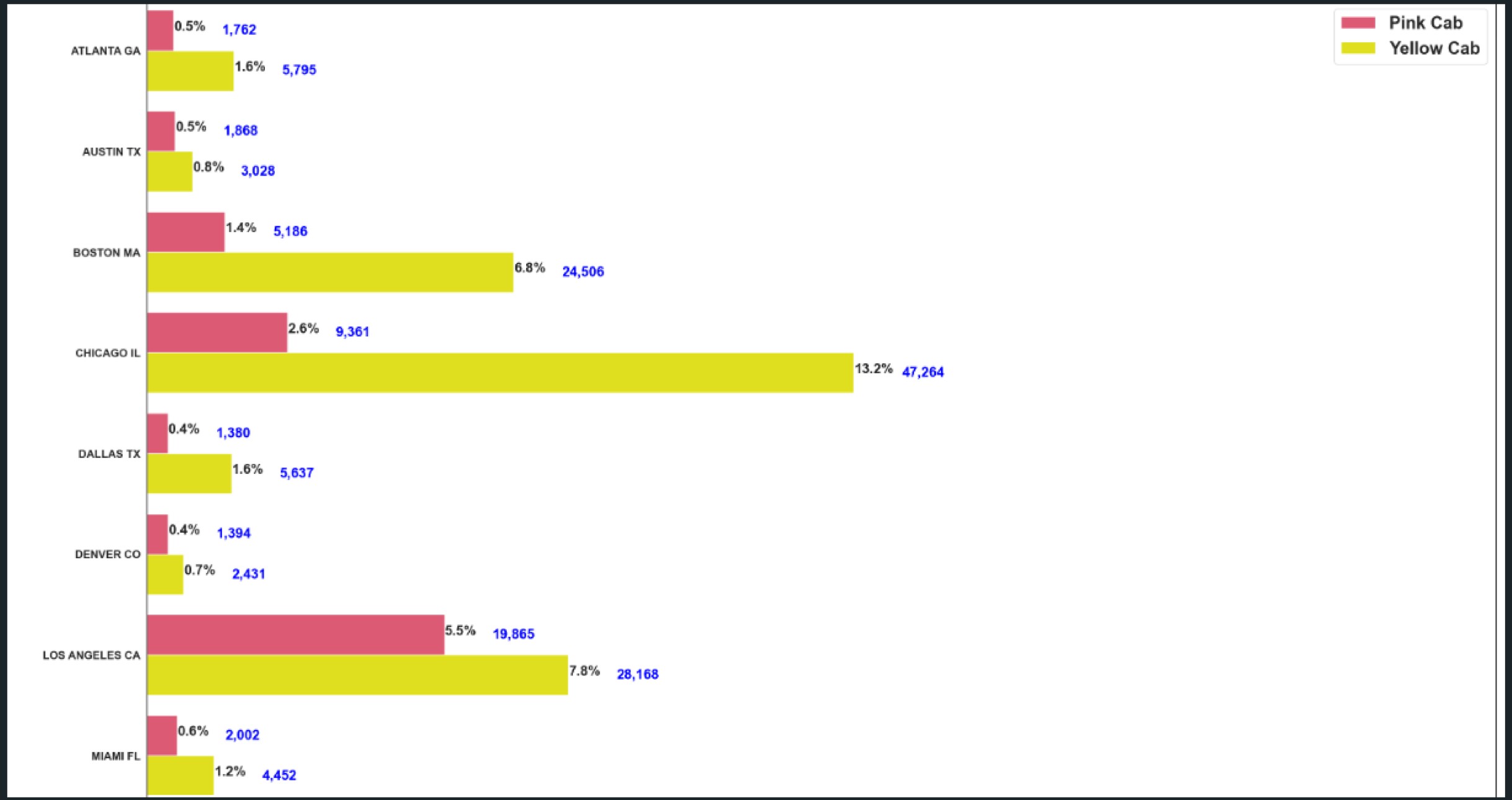
Age Distribution:

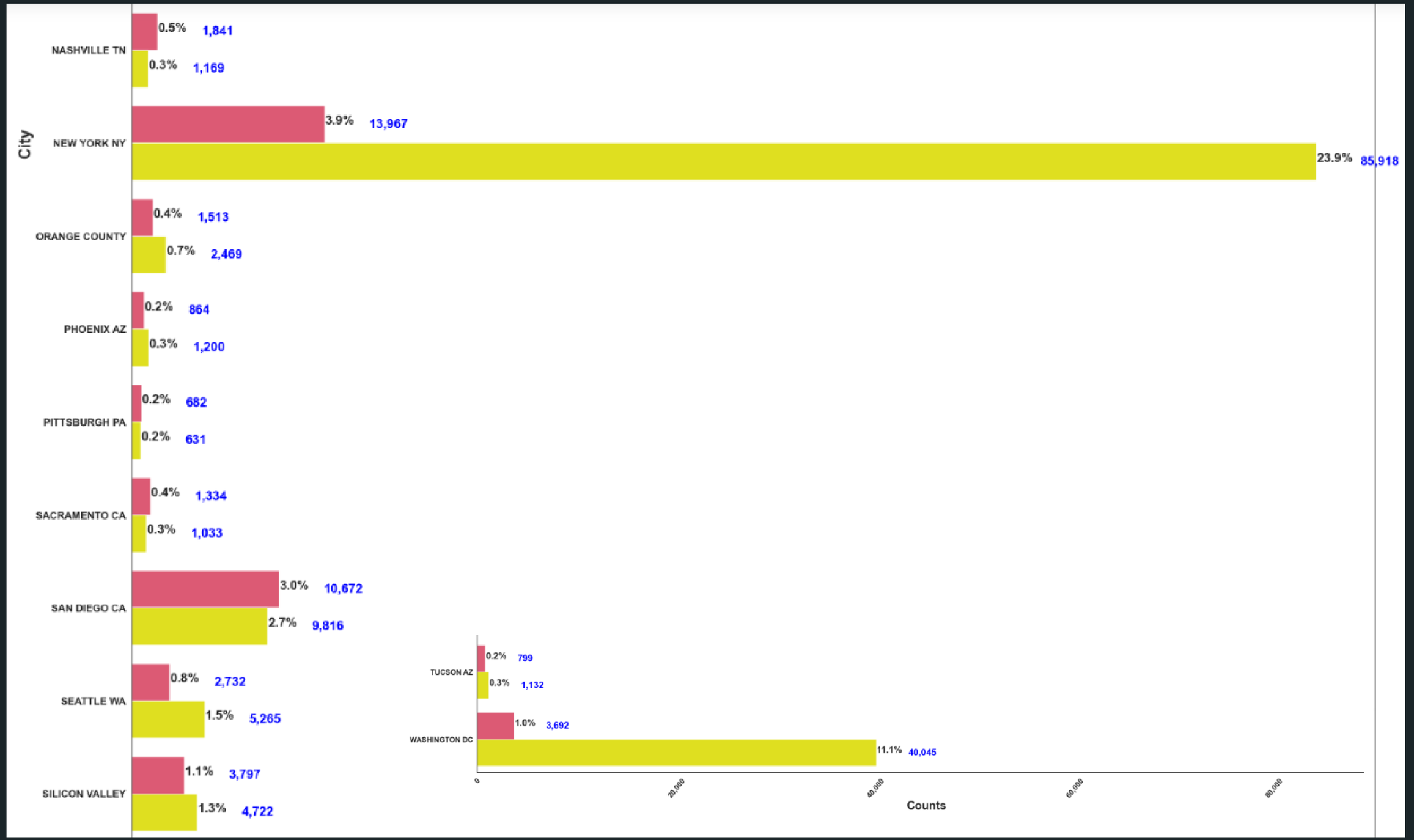
- The age range is between 18-65
- Majority of the transactions occur between the age of 18 and 40
- Notably Age 20 and 24 have the most counts
- Age 65 has the fewest counts
- Yellow Cab has significantly much higher counts than the Pink Cab

Cab Transactions by City Distribution:



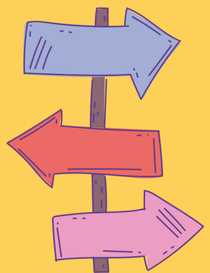
- Top 5 cities with the most transactions: New York City, Chicago, L.A, Washington D.C, & Boston.
- Bottom 5 cities with the least transactions: Pittsburgh, Tuscon, Phoenix, Sacramento and Nashville.
- New York City constitutes of 27.8% of transactions, close to 1/3 of the total rides.
- Top 5 cities combined (278,072) constitutes 77% of the total business transactions.
- The actual % distribution is in the next two slides

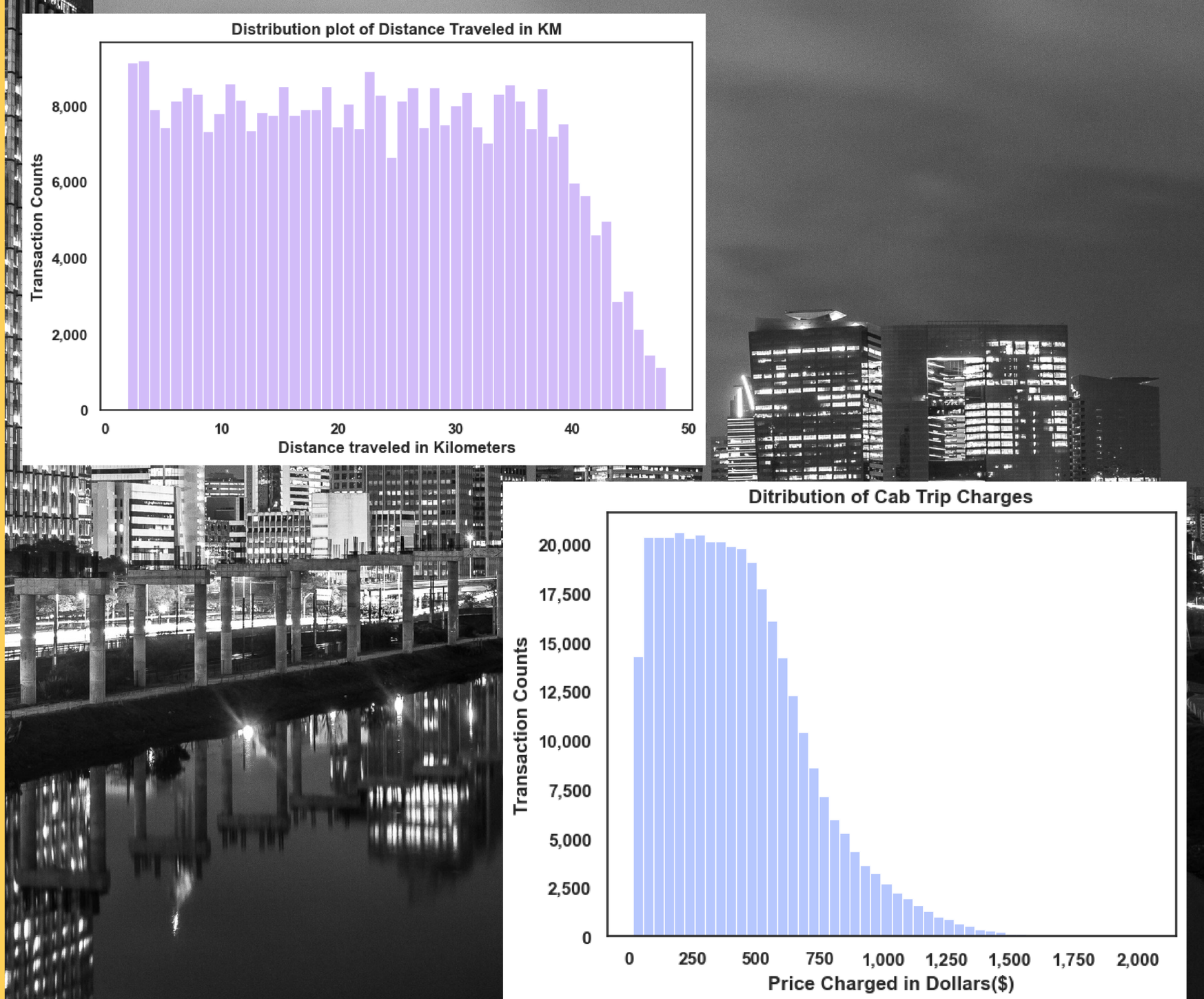

Pink Cab
Yellow Cab
ATLANTA GA
0.5%
1,762
1.6%
5,795
AUSTIN TX
0.5%
1,868
0.8%
3,028
BOSTON MA
1.4%
5,186
6.8%
24,506
CHICAGO IL
2.6%
9,361
13.2%
47,264
DALLAS TX
0.4%
1,380
1.6%
5,637
DENVER CO
0.4%
1,394
0.7%
2,431
LOS ANGELES CA
5.5%
19,865
7.8%
28,168
MIAMI FL
0.6%
2,002
1.2%
4,452


City
NASHVILLE TN
0.5% 1,841
0.3% 1,169
NEW YORK NY
3.9% 13,967
23.9% 85,918
ORANGE COUNTY
0.4% 1,513
0.7% 2,469
PHOENIX AZ
0.2% 864
0.3% 1,200
PITTSBURGH PA
0.2% 682
0.2% 631
SACRAMENTO CA
0.4% 1,334
0.3% 1,033
SAN DIEGO CA
3.0% 10,672
2.7% 9,816
SEATTLE WA
0.8% 2,732
1.5% 5,265
SILICON VALLEY
1.1% 3,797
1.3% 4,722
TUCSON AZ
0.2% 799
0.3% 1,132
WASHINGTON DC
1.0% 3,692
11.1% 40,045
0
20,000
40,000
60,000
80,000
Counts

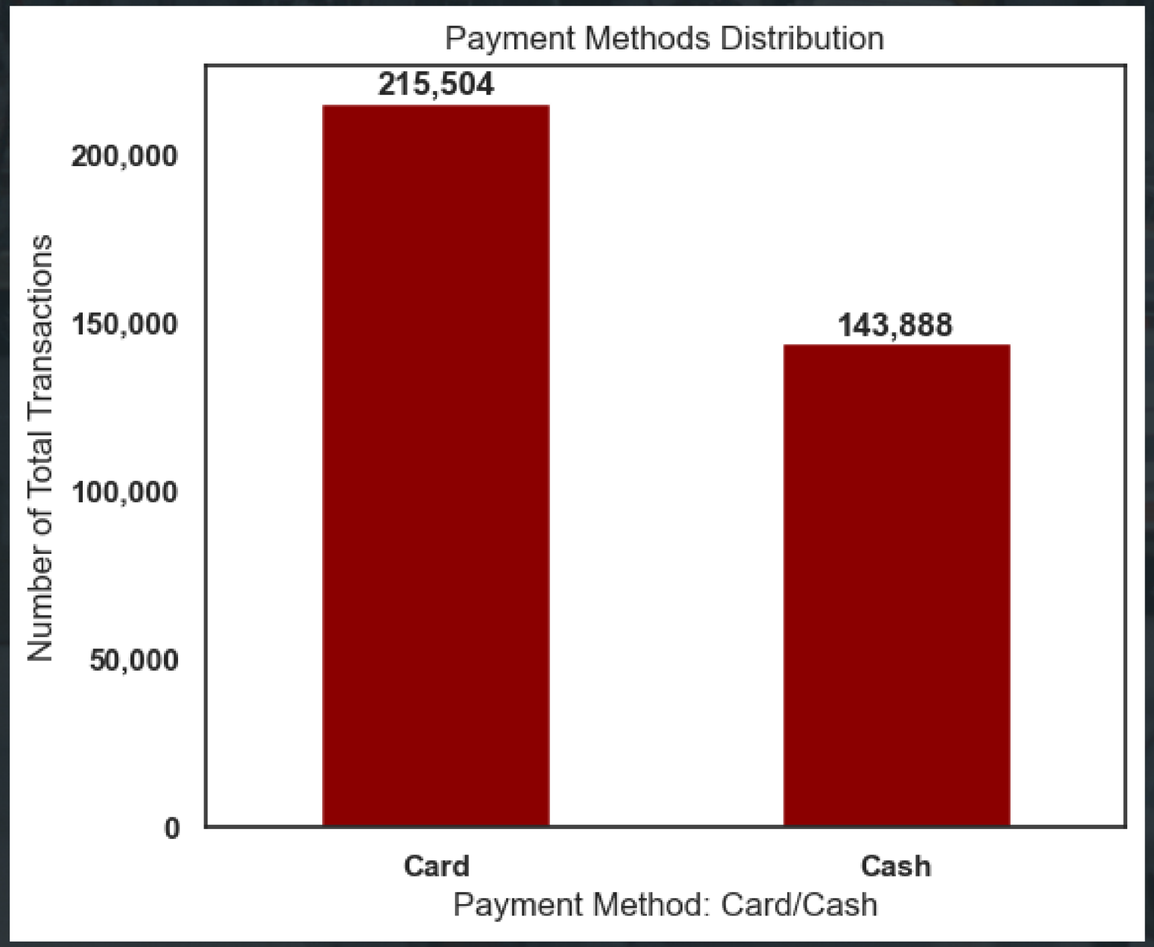
Distance and Trip Charges:

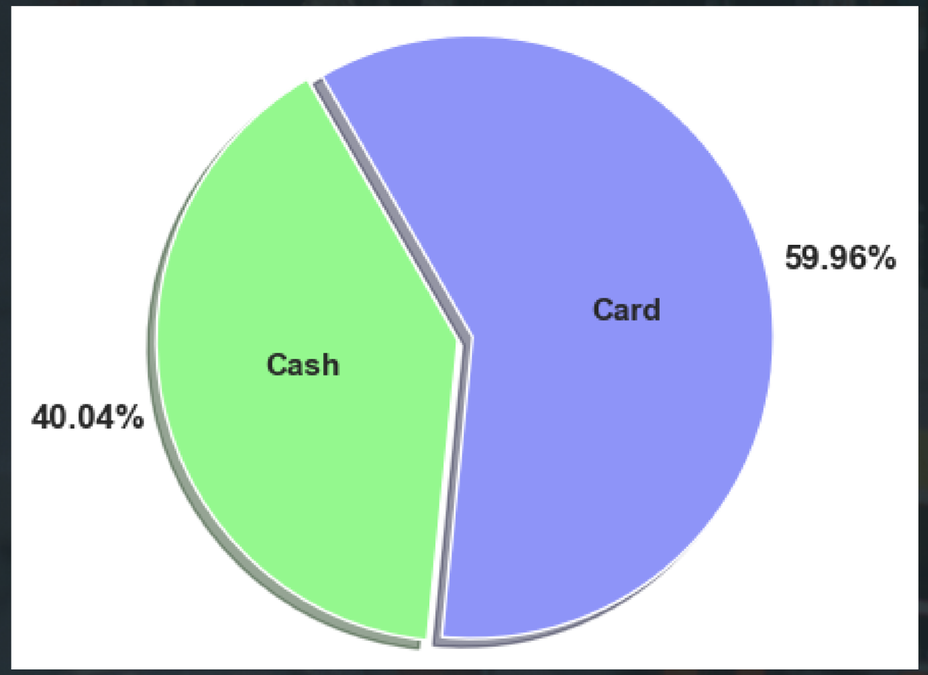
- The most rides are between 1-3 kilometers, reaching 9,000 rides
- The average distance is 22.57KM
- The minimum distance is 1.9KM and the maximium is 48KM
- The average cab ride costs $423.44.
- The minimum charge is $15.6 and the maximum is $2,048



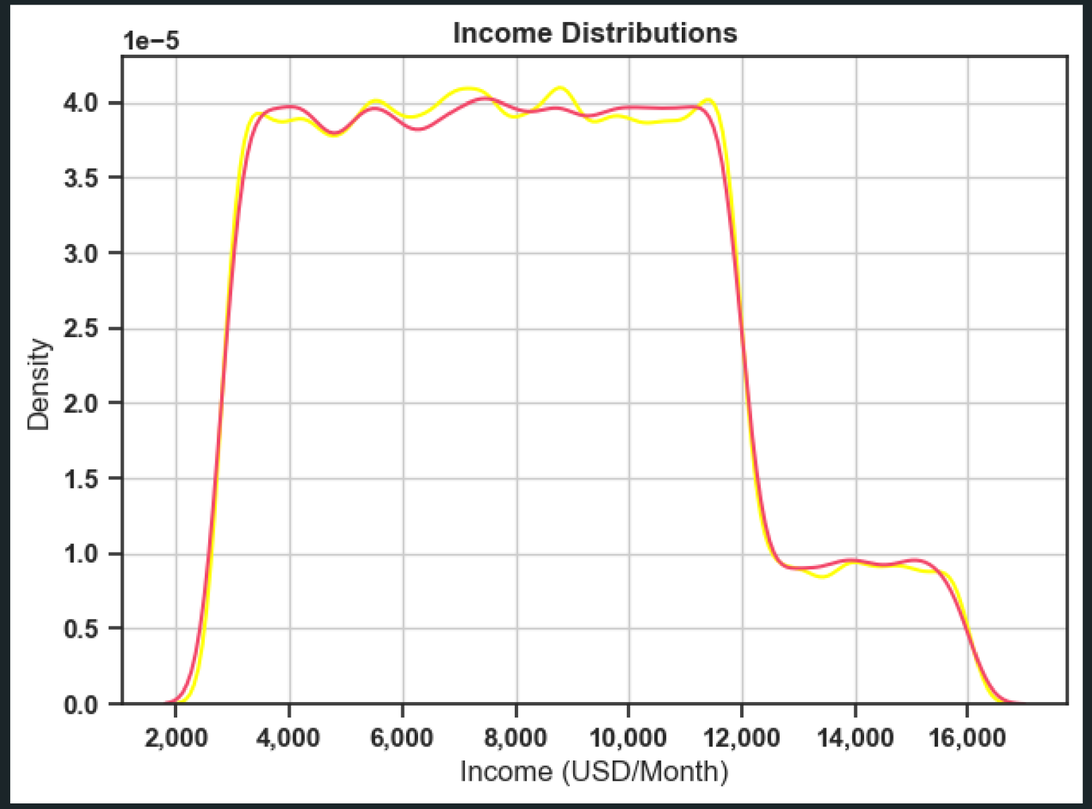


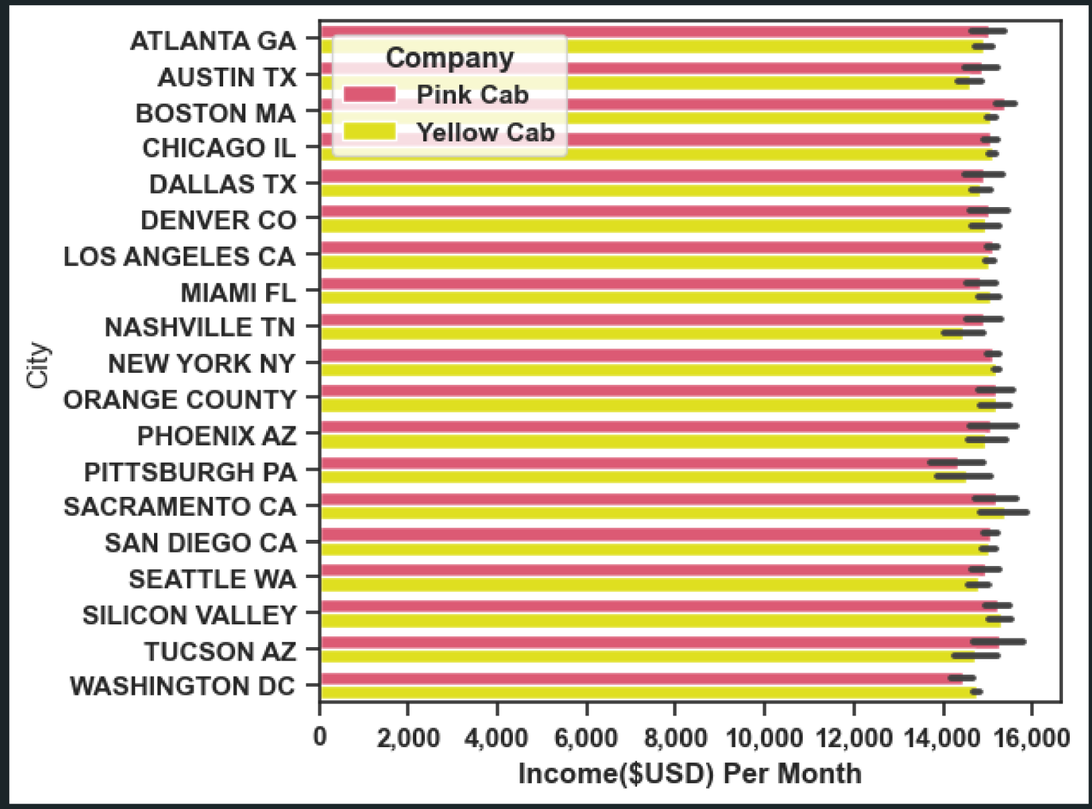
Payment Method Distribution:





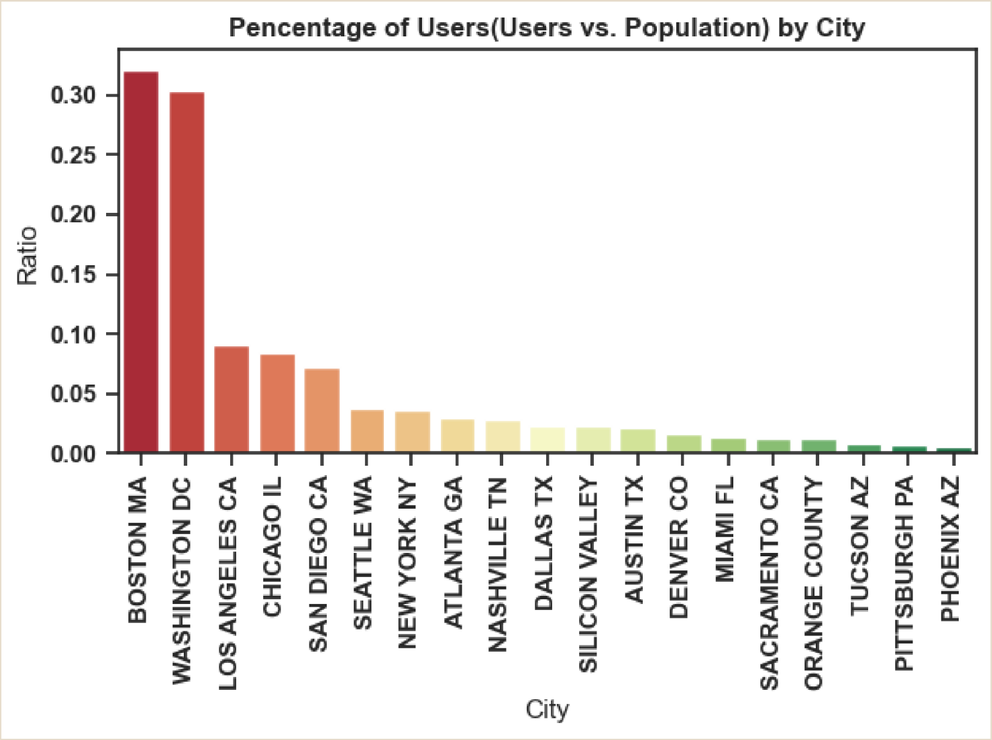
Income Distribution





- Income seems to have no effect on Cab rides and profits(distribution is uniform)
- Majority income range with higher Taxi usage frequency is between $3000-$11,000

Presented by Alison March



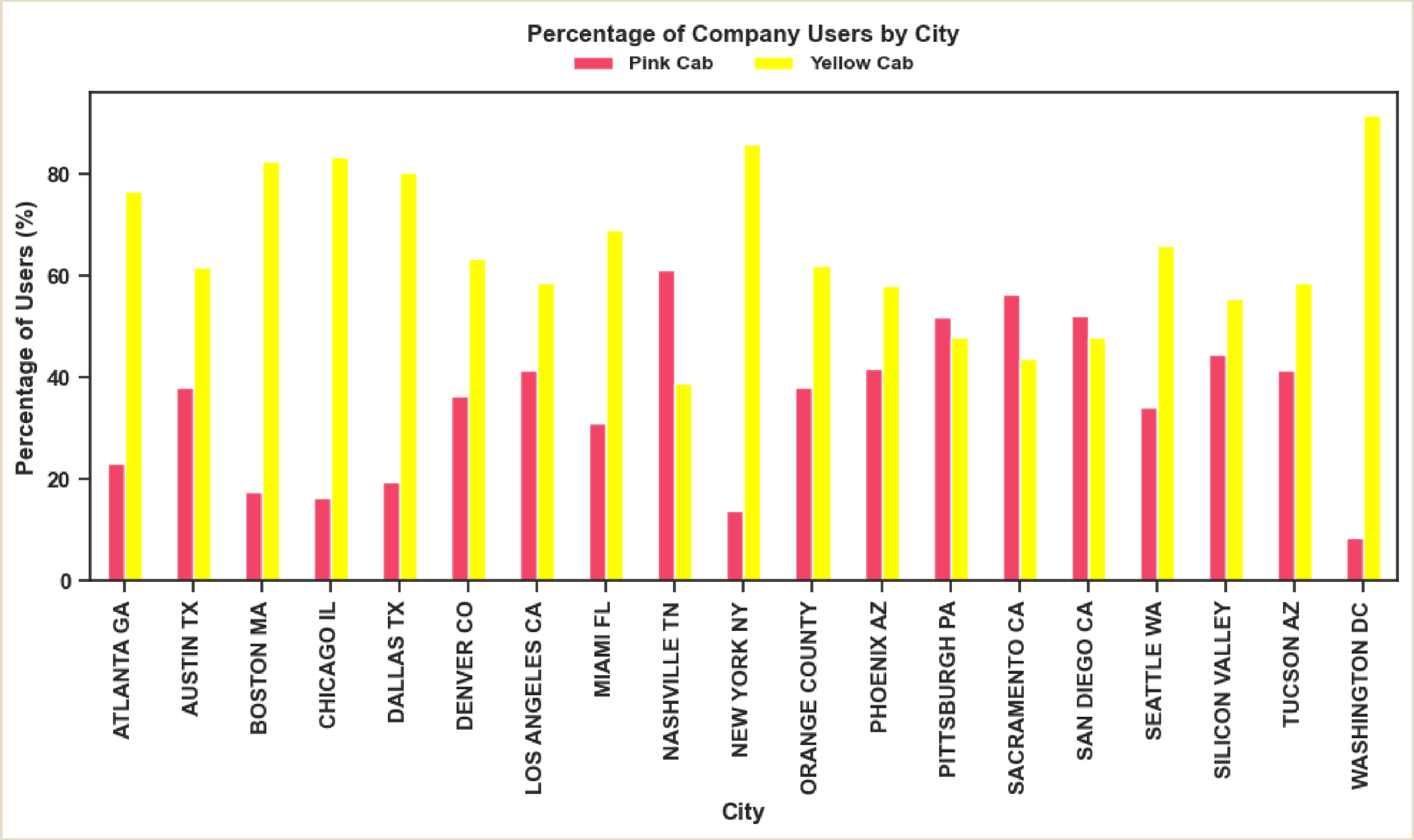
- **Cities with the least users are Tucson, Pittsburgh and Phoenix.**

- **Average users are 158,365 counts**

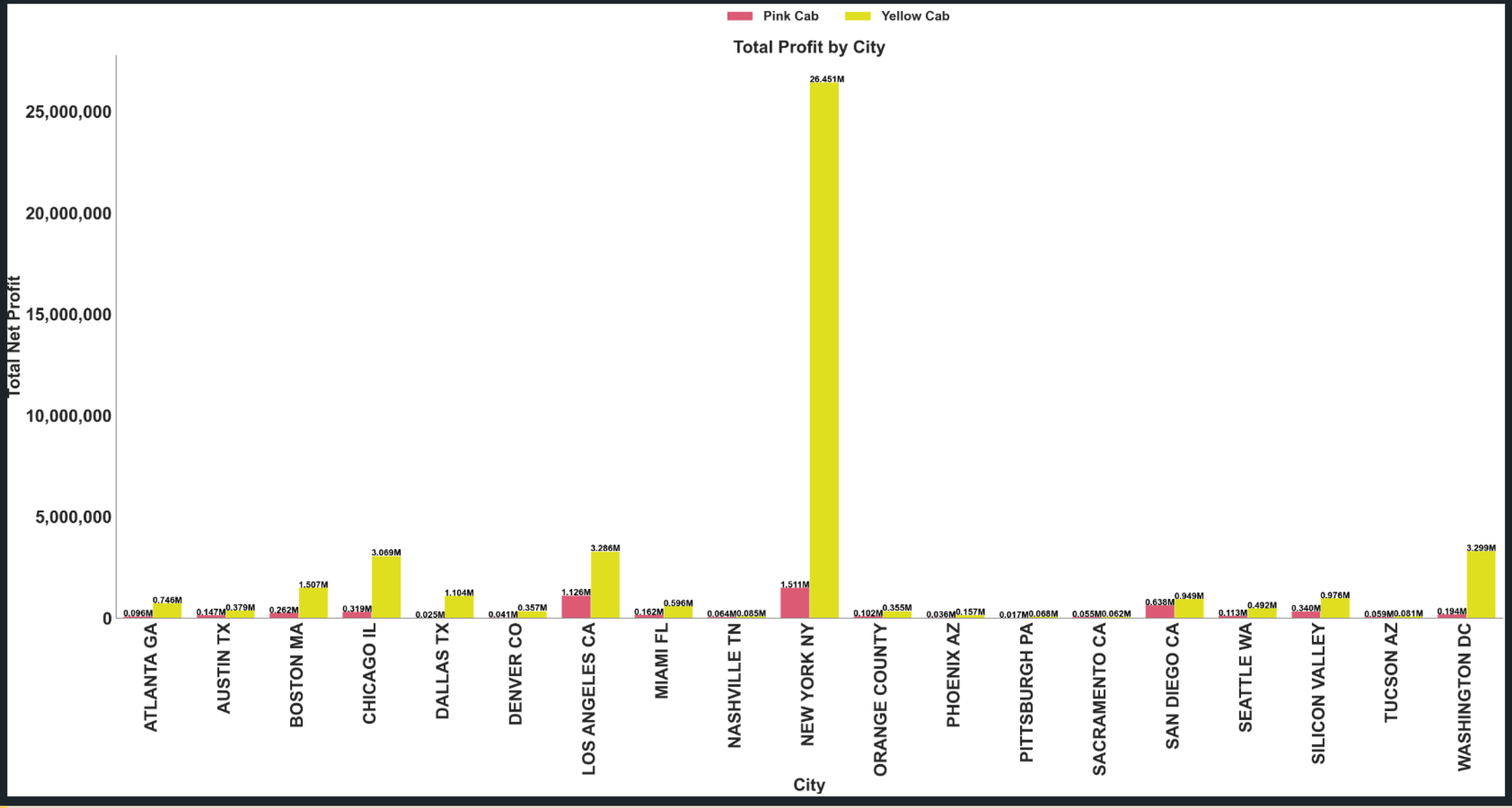
Cab User Distribution:

- **Top 3 cities with the most users are Boston, Washington DC and Los Angeles.**

- **Yellow Cab has more users than the Pink Cab EXCEPT Pittsburgh, Sacramento and San Diego**



Profit Distribution:



Presented by Alison March

Costs and Profits Summary:

Company	Total Costs:	Total Rides:	Total Distance Traveled(KM)	Average Costs Per Ride:	Average Costs Per KM:
Pink Cab	$21,020,923	84,711	1,911,073	$248.149	$11.000
Yellow Cab	$81,833,514	274,681	6,199,417	$297.922	$13.200

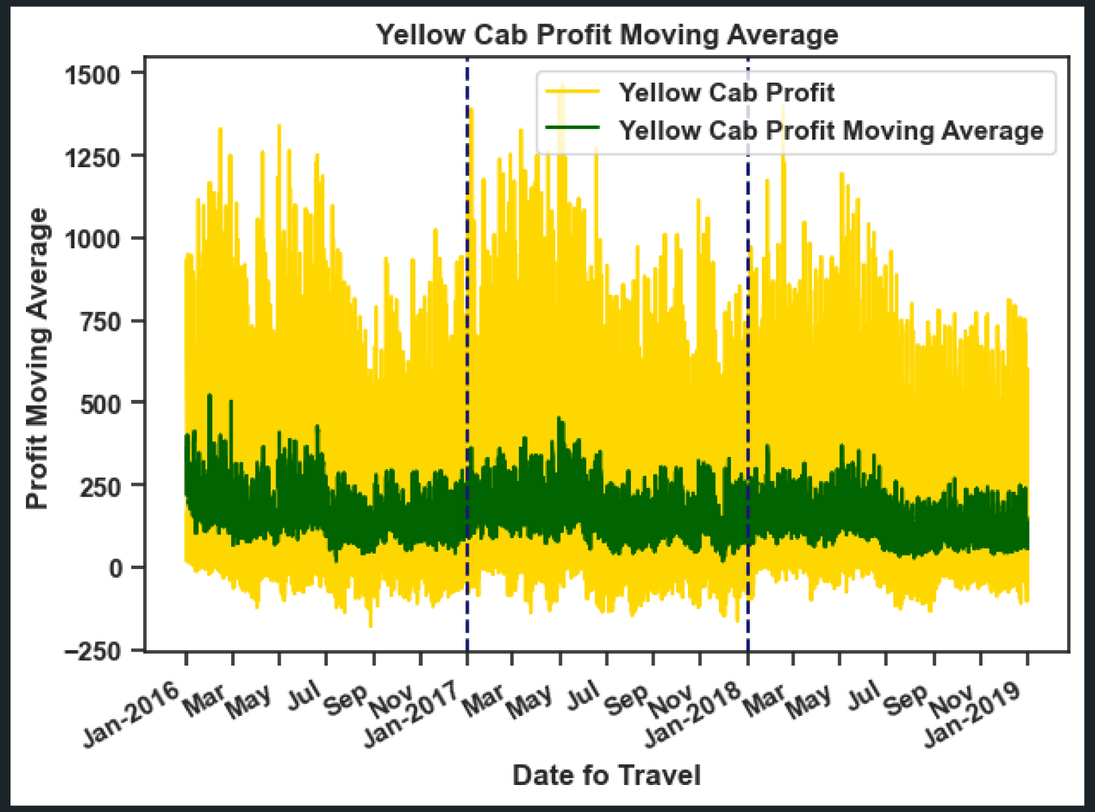
Company	Total Profits:	Total Rides:	Total Distance Traveled(KM)	Average Profits Per Ride:	Average Profits Per KM:
Pink Cab	$5,307,328	84,711	1,911,073	$62.652	$2.777
Yellow Cab	$44,020,373	274,681	6,199,417	$160.260	$7.101

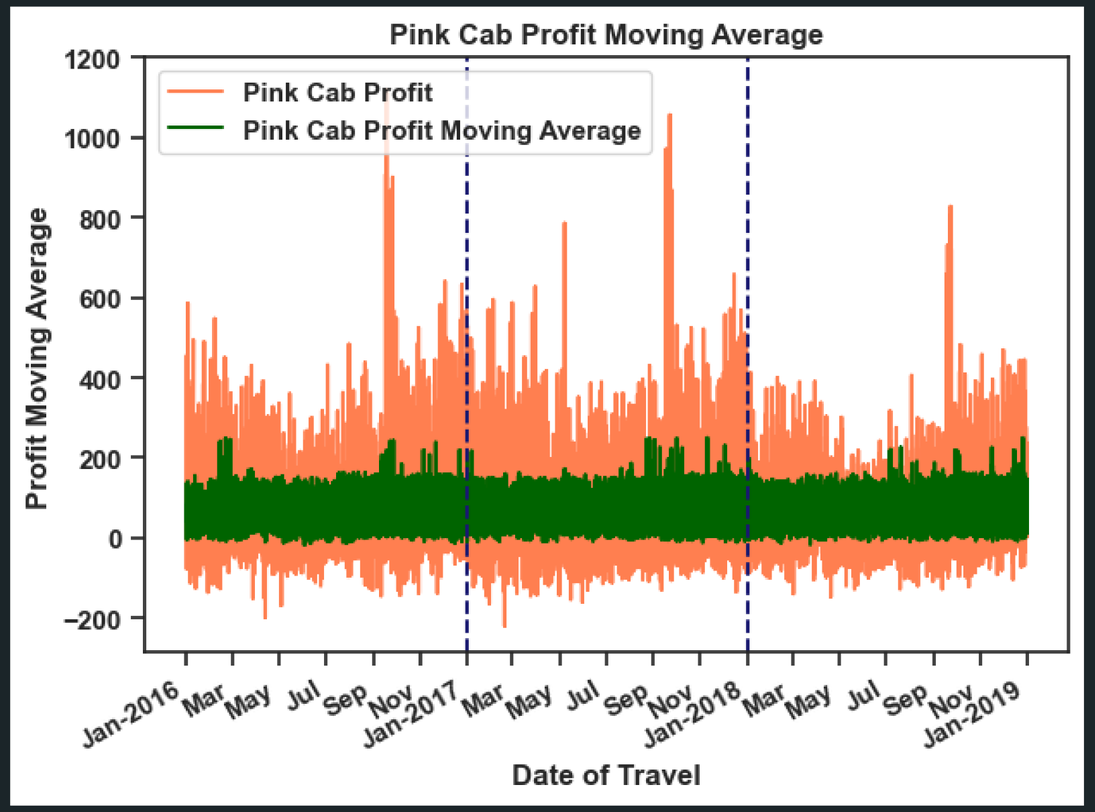
Year	Company	Profit
2016	Pink Cab	$1,713,511.224
	Yellow Cab	$13,926,995.4316
2017	Pink Cab	$2,033,654.908
	Yellow Cab	$16,575,977.968
2018	Pink Cab	$1,560,162.189
	Yellow Cab	$13,517,399.7712

- Yellow Cab travels more than 3 times the distances than the Pink Cab
- The average cost per ride for Pink Cab is $248 and $298 for Yellow Cab
- The average cost per kilometer for Pink Cab is $11 and $13 for Yellow Cab
- Yellow Cab's total profits is 8.8 times higher of Pink Cab's total profits.
- The average profit for Pink Cab is $62.65 and $160.26 for Yellow Cab.
- The average profit per Kilometer for Pink Cab is $2.78 and $7.1 for Yellow Cab

Presented by Alison March

Profit Moving Average:





- Yellow Cab: The profit peaks around beginning of each year until May, then dipped down then rises up again in January thru May of the following year, There is cyclic pattern hence there is seasonality existed. The profit moving average stays between $150-$350.
- Pink Cab: The profit usually peaks after September(October-November period) then dips back down until the next window after September. May appears to be the month with biggest losses. There is also a cyclic pattern hence there is seasonality existed. The profit moving average stays between $0-$150.

Proposed Recommendations:

- We recommend investing in Yellow Cab since it generates 8 times more in annual profits than the Pink Cab between 2016-2018.

- Pink Cab dominates the market over Yellow Cab in the following 3 cities: Pittsburgh, Sacramento and San Diego. In order to generate more profits in those cities, Yellow Cab should expand on incentives, marketing and advertisement to increase the users.

- Yellow Cab should also put emphasis on the second half of the year between June- December with incentives, loyalty program and advertisement to increase more business sales and attract patrons.

- Yellow Cab performs better than the Pink Cab in terms of moving average profits, and there is an estimate of 2.33 times higher in moving average in Yellow Cab compared to Pink Cab.

- In order for Pink Cab to improve profits, it needs to expand market in other cities except Pittsburgh, Sacramento and San Diego.

- Pink Cab also needs to focuses on months with the losses like May. Rolling out incentives, discounts and promotions could help mitigate losses.

Summary:

- There are 359,392 Rows by 14 Columns in the Cab dataset
- There are more male cab users than female users for the cab services.
- Yellow Cab has more than three times the transactions than the Pink Cab business
- Majority of cab users age between 18-40, notably early 20- late 20s are most frequent.
- Top 5 cities with the most transactions: NYC, Chicago, L.A, Washington D.C, and Boston. Top 5 cities combined constitutes 77% of the total business transactions.
- Bottom 5 cities with the least transactions: Pittsburgh, Tucson, Phoenix, Sacramento, and Nashville.
- Yellow Cab has more users than the Yellow Cab except Pittsburgh, Sacramento and San Diego.
- Top 3 cities with the most cab users are Boston, Washington D.C and Los Angeles.
- Bottom 3 cities with the least cab users are Tuscon, Pittsburgh and Phoenix.
- Income does not contribute to the cab services and profits. Majority cab users with higher Taxi frequency is between 3,000 - 11,000.
- Yellow Cab: The profit peaks around beginning of each year until May, then dipped down then rises up again in January thru May of the following year, There is cyclic pattern hence there is seasonality existed. The profit moving average stays between $150-$350.
- Pink Cab: The profit usually peaks after September(October-November period) then dips back down until the next window after September. May appears to be the month with biggest losses. There is also a cyclic pattern hence there is seasonality existed. The profit moving average stays between $0-$150.

Presented by Alison March