Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Taxi-Cab Market Exploration Project

January 31, 2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Data Glacier

Your Deep Learning Partner

Executive Summary

A private equity firm is seeking actionable insights into the businesses of two multi-city American cab companies, Pink Cab and Yellow Cab.

Problem Statement

This analysis seeks to understand the profits of these two companies during a period spanning 2016 to 2018 using four related data files. The goal of this study is to determine which company would be the better investment, Yellow Cab or Pink Cab.

Approach

The data files were processed into a master data file then analyzed in response to the following hypotheses …

Hypotheses

•

1. The cab company with more rides overall will have on average greater profitability per ride

•

2. Some cities will be more profitable than others and the more profitable company will dominate those cities

•

3. The differences in average profitability for rides will not vary much by gender of customer

•

4. The trends in profitability over time will be similar between Yellow and Pink Cab

•

5. There will be one clearly more profitable company and it will be possible to make a recommendation on this basis for the contemplated investment.

Exploratory Data Analysis

1. The provided data files were explored
2. Outliers and duplicates were removed as required
3. Data was merged into one master file
4. Data was manipulated to create visualizations utilizing key features for the purpose of testing the hypotheses

Provided Data Utilized in this case study: 4 .csv files

Cab_Data.csv includes details of transactions for two companies across 20 cities in America, with 359392 observations (rides) and 15 additional data columns associated with each ride (including Transaction ID, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip, Customer ID, Payment Mode, Gender, Age, Income)

Customer_ID.csv provides customer specific information

There are 49171 observations (customer IDs) and 3 additional data columns for each (Gender, Age and Income)

Transaction_ID.csv data set correlates transactions to customers

There are 440098 observations (transactions) and 2 additional columns for each (Customer ID and Payment Mode)

City.csv data set provides information about cab use by city

There are 20 observations (cities) with 2 additional columns of data associated with each (Population and Users)

Data Manipulation

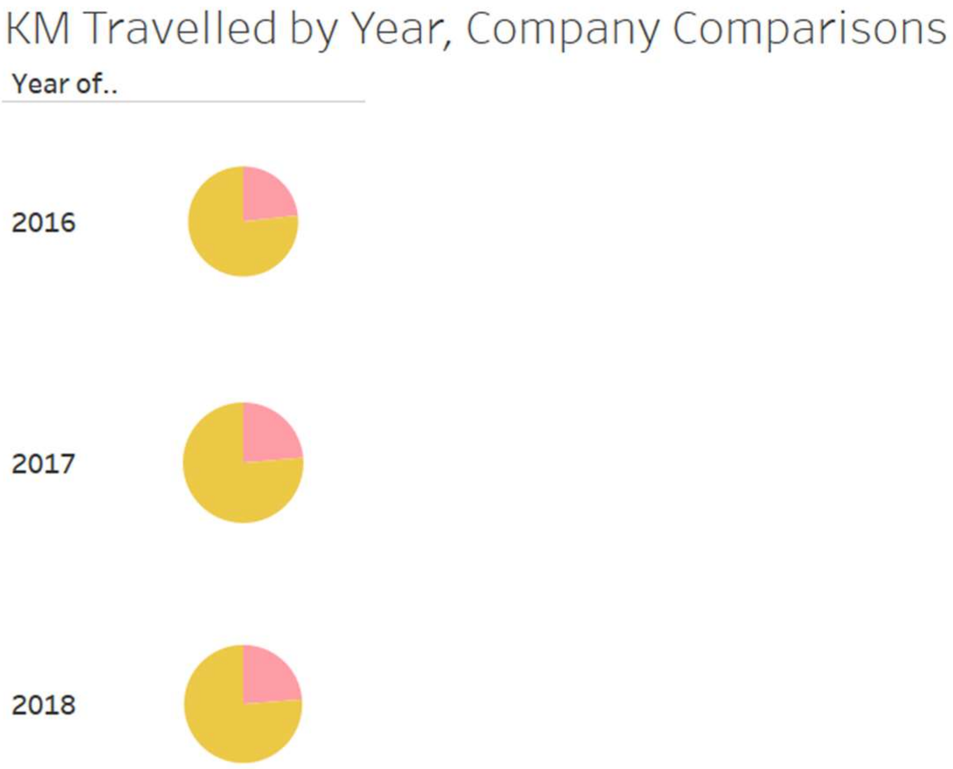
- A created master dataframe incorporating the data from all four .csv data files provided information in an optimal format, with each row representing a single trip/transaction and columns including all necessary data for the analysis.
- Profit per ride data was calculated as well as profit per km (by ride) and these important results were added as new columns to this master dataset.
- There were no extreme outliers in the profit related columns such as would distort the results (in terms of mean profits by company); therefore, no transactions were removed before producing the master dataframe.

EDA Results Summary

1. KM driven per year by each company
2. Profit Trends by Company
3. City Profits and Dominance by Company
4. Age and Gender of Customers
5. Review of Hypotheses in Light of Results

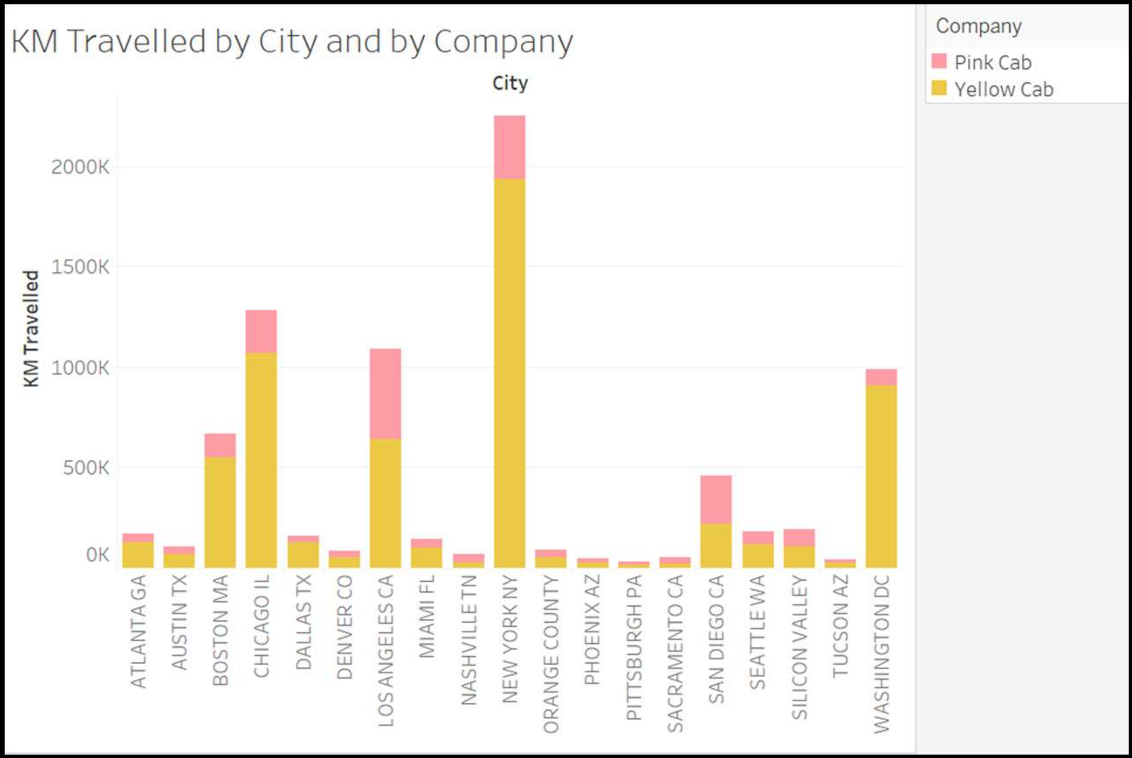
Two Cab Companies, Pink and Yellow

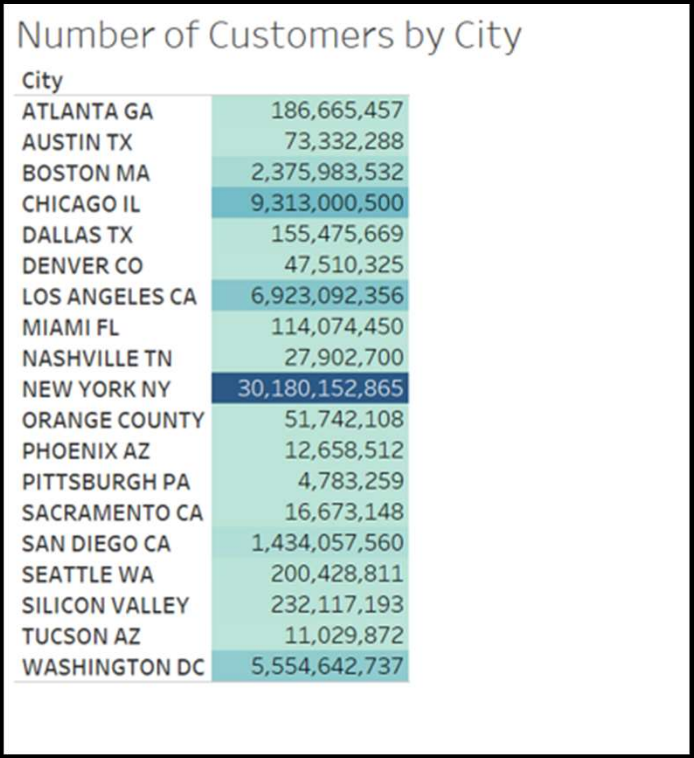
- The Yellow Cab Company accounts for more KM driven all three years



KM Travelled by Year, Company Comparisons

This relationship holds across most cities studied.

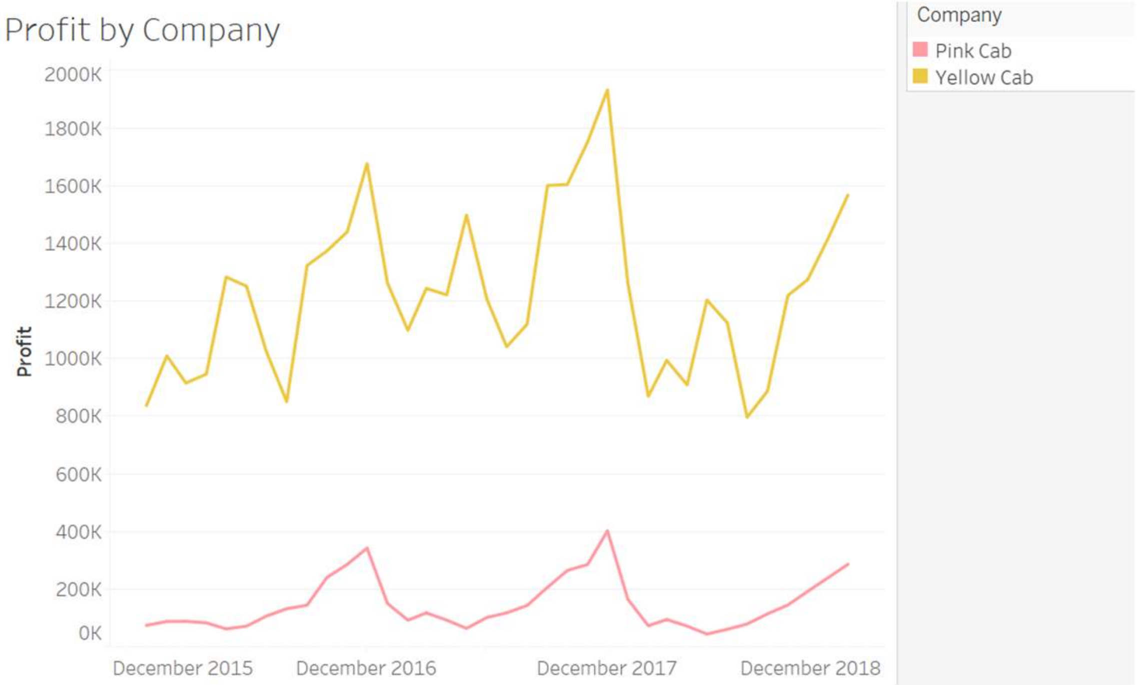




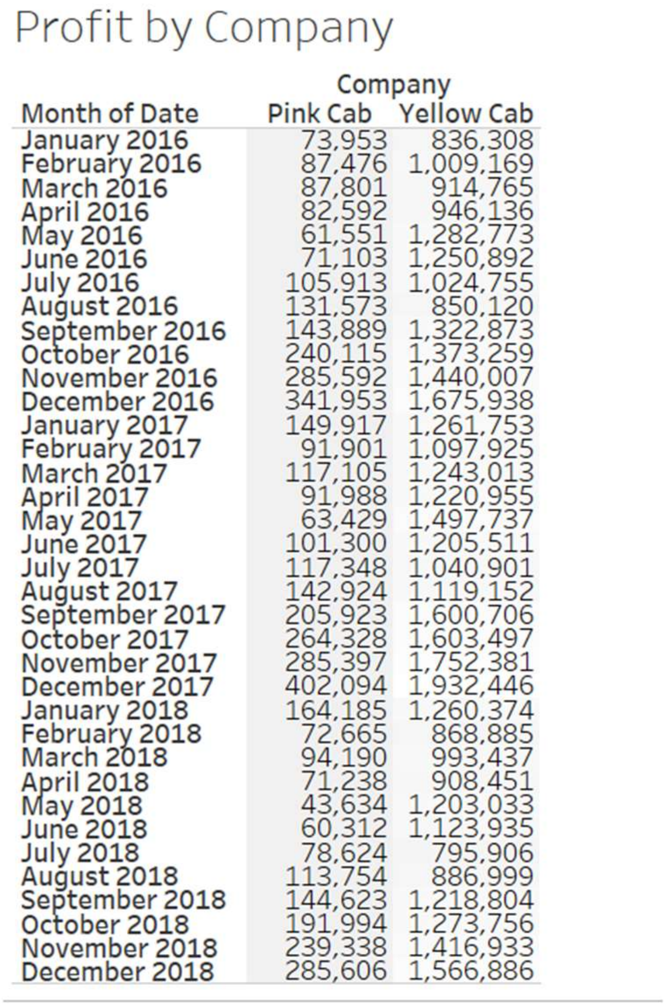
Number of Customers by City

City	
ATLANTA GA	186,665,457
AUSTIN TX	73,332,288
BOSTON MA	2,375,983,532
CHICAGO IL	9,313,000,500
DALLAS TX	155,475,669
DENVER CO	47,510,325
LOS ANGELES CA	6,923,092,356
MIAMI FL	114,074,450
NASHVILLE TN	27,902,700
NEW YORK NY	30,180,152,865
ORANGE COUNTY	51,742,108
PHOENIX AZ	12,658,512
PITTSBURGH PA	4,783,259
SACRAMENTO CA	16,673,148
SAN DIEGO CA	1,434,057,560
SEATTLE WA	200,428,811
SILICON VALLEY	232,117,193
TUCSON AZ	11,029,872
WASHINGTON DC	5,554,642,737

Not surprisingly, Yellow Cab has greater profits over the period studied, but the trends are similar.



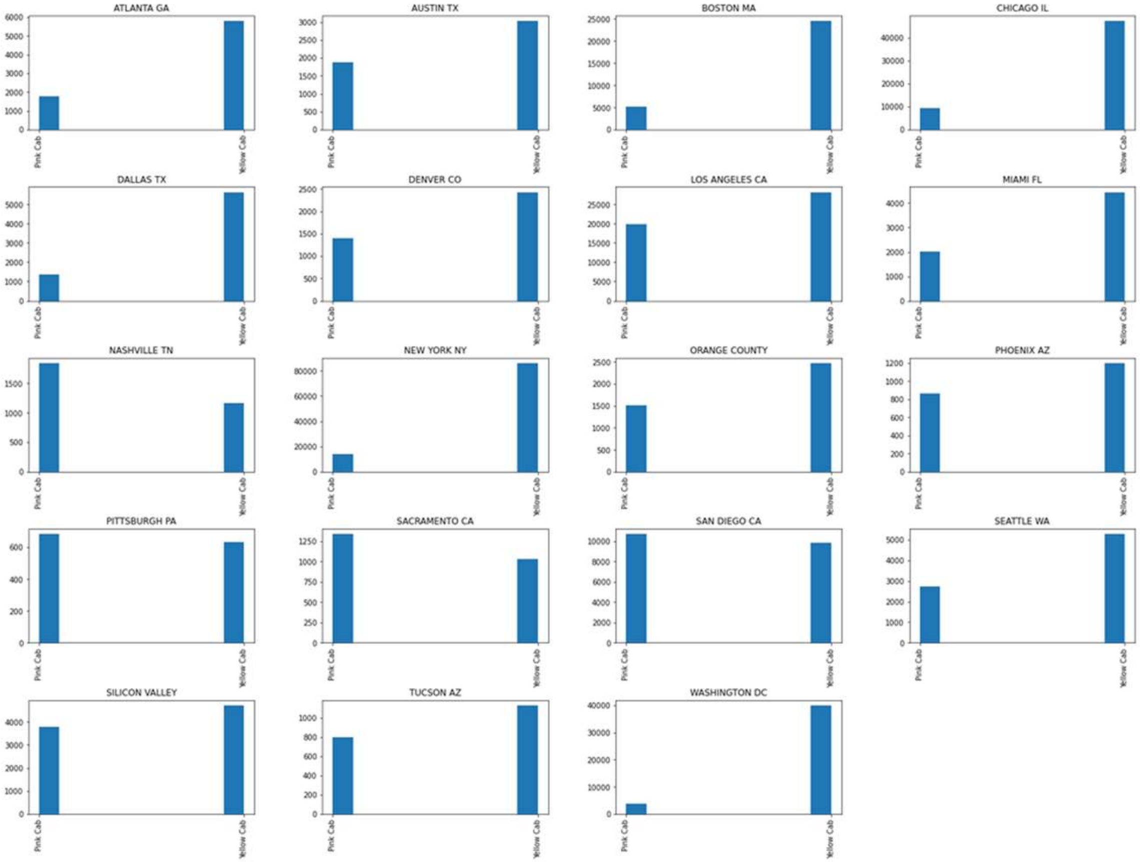
Profit by Month and Company



Profit by Company

	Company	
Month of Date	Pink Cab	Yellow Cab
January 2016	73,953	836,308
February 2016	87,476	1,009,169
March 2016	87,801	914,765
April 2016	82,592	946,136
May 2016	61,551	1,282,773
June 2016	71,103	1,250,892
July 2016	105,913	1,024,755
August 2016	131,573	850,120
September 2016	143,889	1,322,873
October 2016	240,115	1,373,259
November 2016	285,592	1,440,007
December 2016	341,953	1,675,938
January 2017	149,917	1,261,753
February 2017	91,901	1,097,925
March 2017	117,105	1,243,013
April 2017	91,988	1,220,955
May 2017	63,429	1,497,737
June 2017	101,300	1,205,511
July 2017	117,348	1,040,901
August 2017	142,924	1,119,152
September 2017	205,923	1,600,706
October 2017	264,328	1,603,497
November 2017	285,397	1,752,381
December 2017	402,094	1,932,446
January 2018	164,185	1,260,374
February 2018	72,665	868,885
March 2018	94,190	993,437
April 2018	71,238	908,451
May 2018	43,634	1,203,033
June 2018	60,312	1,123,935
July 2018	78,624	795,906
August 2018	113,754	886,999
September 2018	144,623	1,218,804
October 2018	191,994	1,273,756
November 2018	239,338	1,416,933
December 2018	285,606	1,566,886

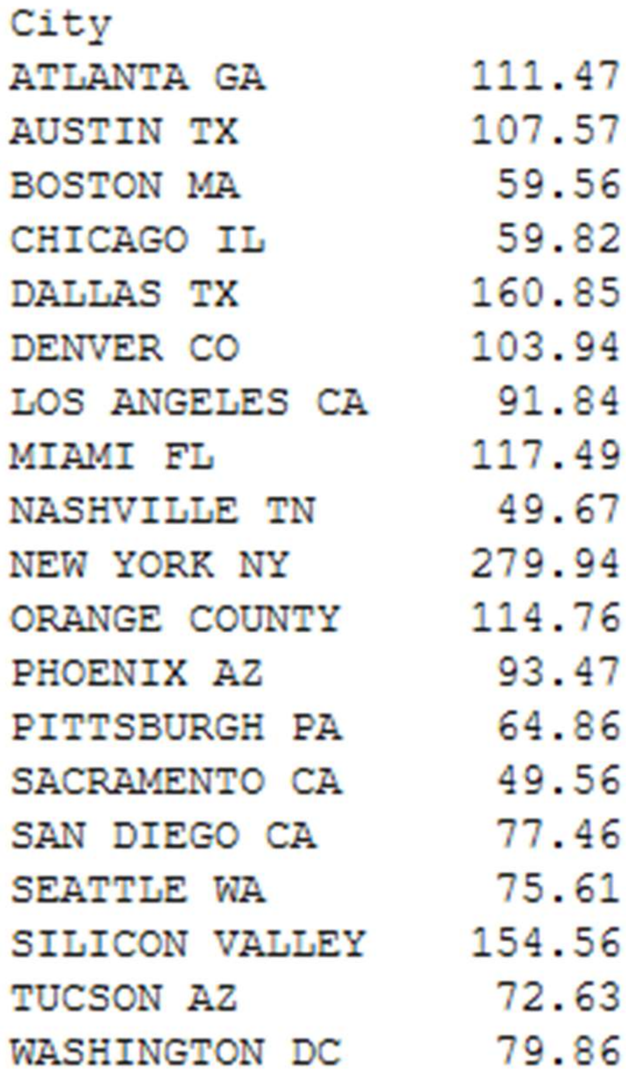
Dominance of companies by city



Pink Cab dominates in Pittsburgh, Sacramento, San Diego, and Nashville.

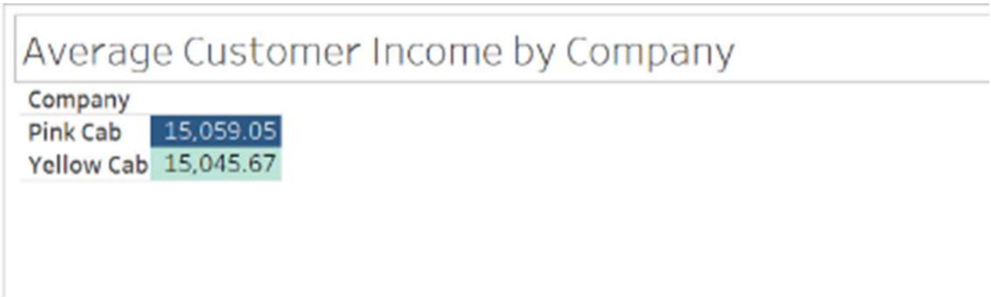
Yellow Cab is larger but is especially dominant in Atlanta, Boston, Chicago, NYC, and Washinton, DC.

This distribution does not appear to give a special advantage to either company: Yellow Cab dominates in the most profitable NYC market, but it also dominates in several of the least profitable markets such as Boston and Chicago.



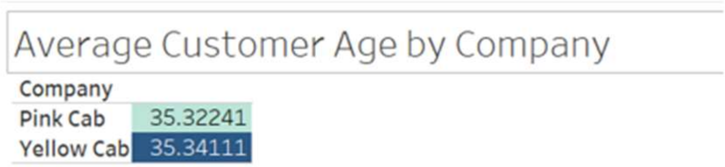
```
City
ATLANTA GA          111.47
AUSTIN TX           107.57
BOSTON MA            59.56
CHICAGO IL           59.82
DALLAS TX           160.85
DENVER CO           103.94
LOS ANGELES CA       91.84
MIAMI FL            117.49
NASHVILLE TN         49.67
NEW YORK NY         279.94
ORANGE COUNTY       114.76
PHOENIX AZ           93.47
PITTSBURGH PA        64.86
SACRAMENTO CA        49.56
SAN DIEGO CA         77.46
SEATTLE WA           75.61
SILICON VALLEY      154.56
TUCSON AZ            72.63
WASHINGTON DC        79.86
```

Customer Age and Income are Similar across the two companies



Average Customer Income by Company

Company	
Pink Cab	15,059.05
Yellow Cab	15,045.67

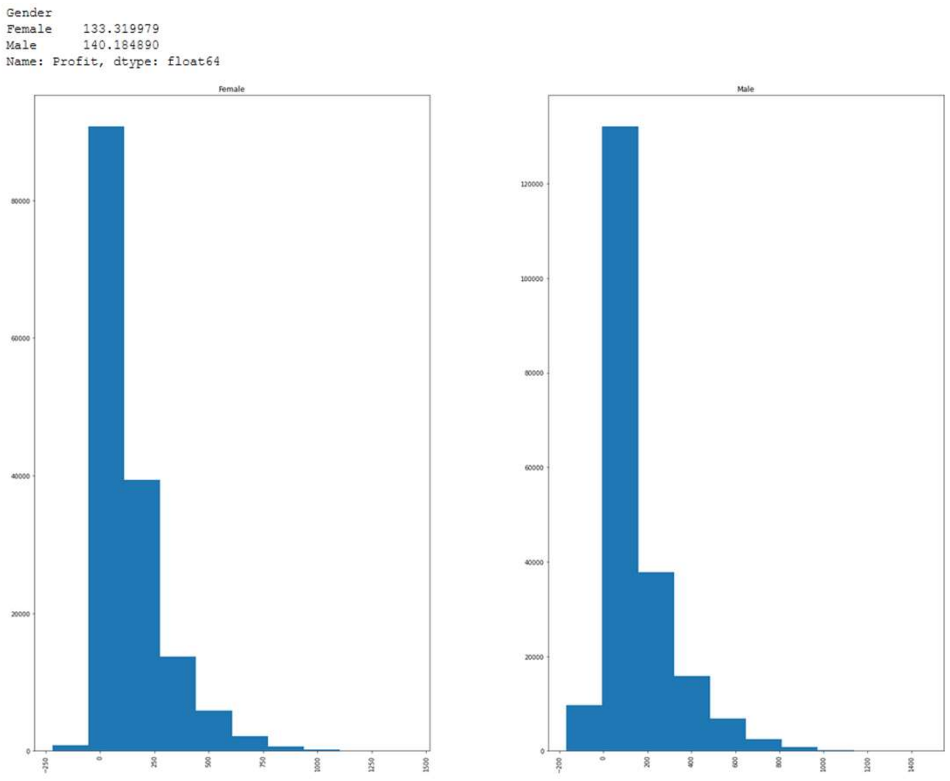


Average Customer Age by Company

Company	
Pink Cab	35.32241
Yellow Cab	35.34111

Cab Company Profits by Gender of Customer

The average profit histograms by gender indicate there is not much difference, so it probably is not worthwhile advertising for gender or considering potential gender dominance for either company.

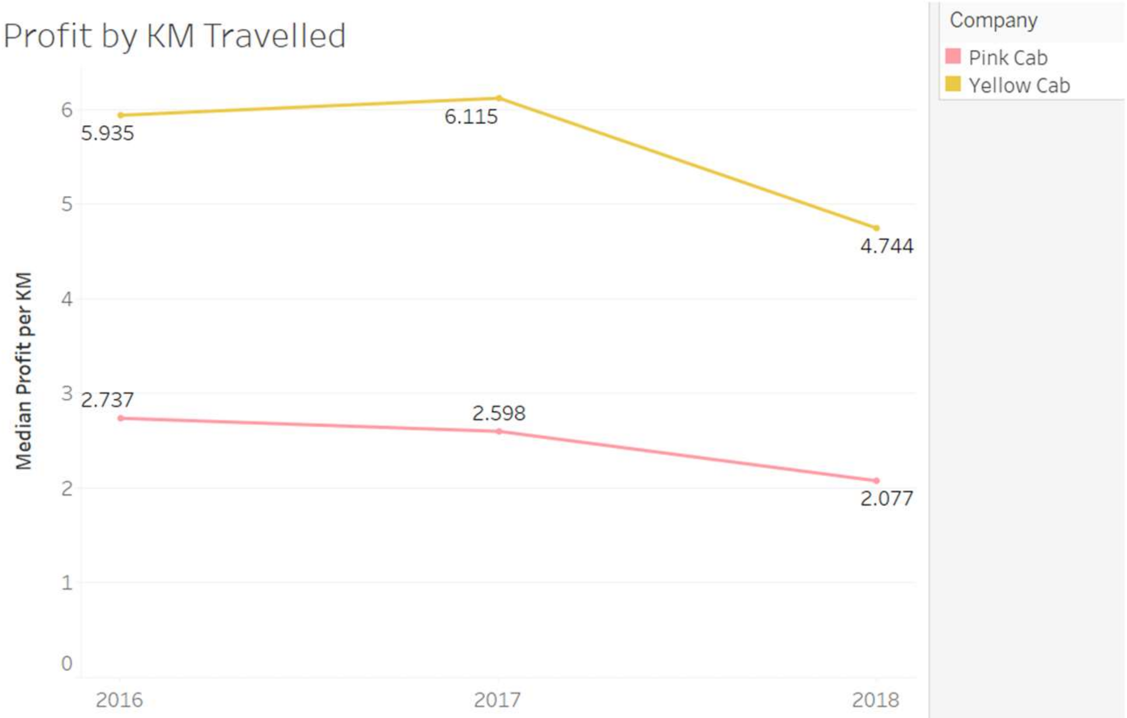


```
Gender
Female    133.319979
Male      140.184890
Name: Profit, dtype: float64
```

Is the Greater Profit by the Yellow Cab Company Related to the Company Size?

More trips travelled and more KM driven accounting for the larger profit?

Not Necessarily: The Yellow Cab Company is also More Profitable by KM Travelled



Yellow Cab Vs. Pink Cab

- More KM are travelled each year by Yellow Cab
- Yellow Cab dominates Pink Cab in most cities in terms of profit
- The customer base for both cab companies is similar in terms of average age and income.

An important finding:

Profit by KM is declining for both companies, but Yellow Cab has substantially higher profit by KM than Pink Cab every year studied indicating Yellow Cab is more profitable in general.

Hypotheses Revisited

1. The cab company with more rides overall will have on average greater profitability per ride as well. TRUE

2. Some cities will be more profitable than others and the more profitable company will dominate in those cities. FALSE

3. The differences in profitability will not vary much by gender of the passenger. TRUE

4. The trends in profitability over time will be similar between Yellow and Pink Cab. TRUE

5. There will be one clearly more profitable company overall, and it will be possible to make a recommendation on this basis for which company is the better investment. TRUE

Recommendations:

The better financial investment would be in the Yellow Cab Company.
Profits are superior overall as well as by distance travelled; the Yellow Cab Company already dominates in most cities.

Final Recommendation: Invest in the Yellow Cab Company

Thank You

Data Glacier

Your Deep Learning Partner