Exploratory Data Analysis

G2M insight for Cab Investment firm

Nov 18, 2022

Agenda

Problem Summary

Approach

EDA

EDA Summary

Conclusions and Recommendations

Case background

Client:
XYZ, a private firm in US

Description:
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 they want to understand the market before taking final decision

Objectives:
- provide actionable insights to help XYZ firm in identifying the right company for making investment
- give recommendations on investments

Data Sources' Content Analysis: Assumptions and Insights

Data Glacier
Your Deep Learning Partner

Data Sources: 4 .csv files which can be linked by either identification fields or string field

Main assumptions:

- "Cost of Trip" feature is total cost summing up waiting time fee, cost of trip distance by counter and other direct costs applied;
- Observed differences are due to companies' price policies and marketing campaigns; no other factors;
- All customers are valid credit card holders (based on age of some customers <21 years)

Insights:

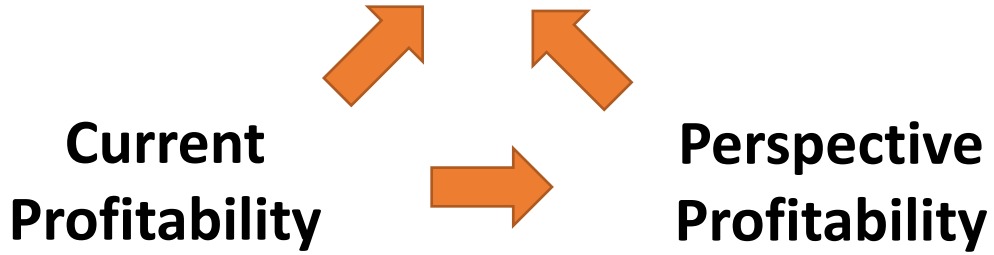
- Number of transactions exceeds number of trips detailed records (**not all the trips info is available**)
- As there are near 80 100 transactions (22% from analyzed transactions quantity) without trip details, we should try to found features that could make it possible to classify transactions as belonging to a particular company

Data discrepancies:

- "Population" feature **values not suite** the official US Census Bureau data (for example, population of Miami, FL in 2016 was 453,579 but the value of 1,339,155 provided, Boston has official population 673,184 but provided value is 248,968)

Approach

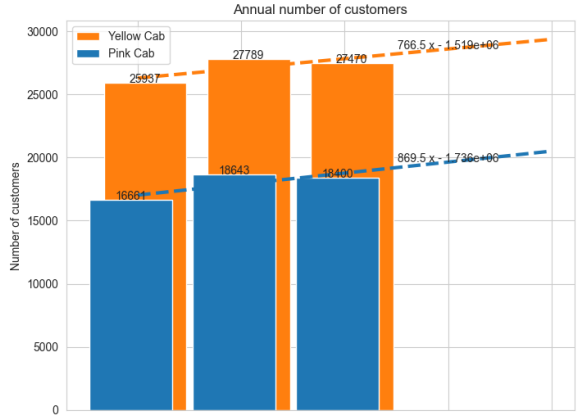
Investment Decision



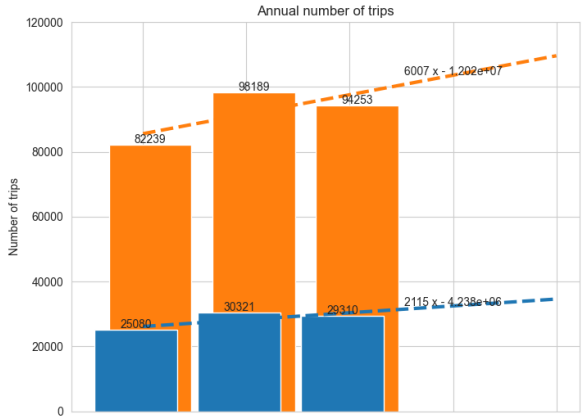
Profitability:
Margin = Total Revenue – Total Costs

Total Revenue = Trips Quantity * Average Trip Distance * Average Revenue per 1 km
Total Costs = Trips Quantity * Average Trip Distance * Average Costs per 1 km

Overall characteristics

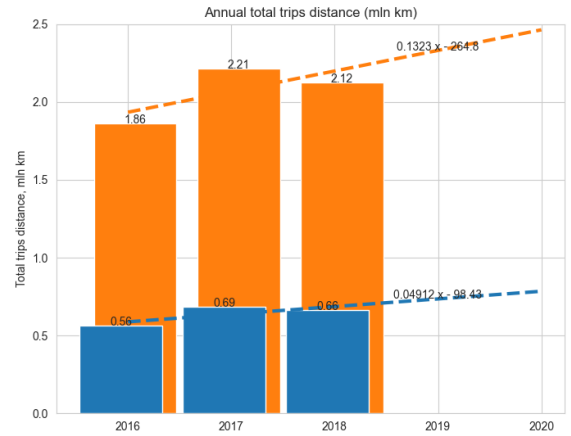


Number of customers grows almost equal



“Pink Cab”’s number of customers grows slightly faster but not enough to influence on investments decision

“Yellow Cab” grows in annual number of trips substantially faster than “Pink Cab” does

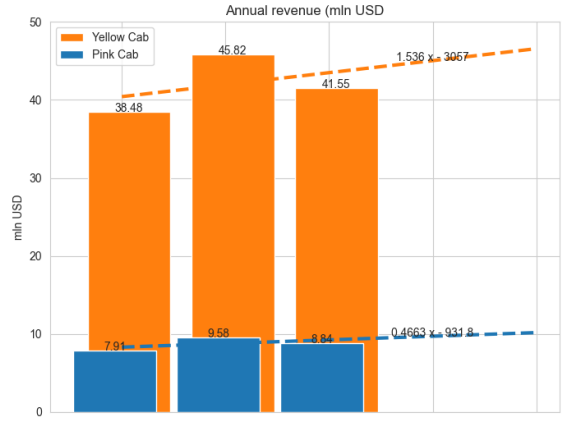


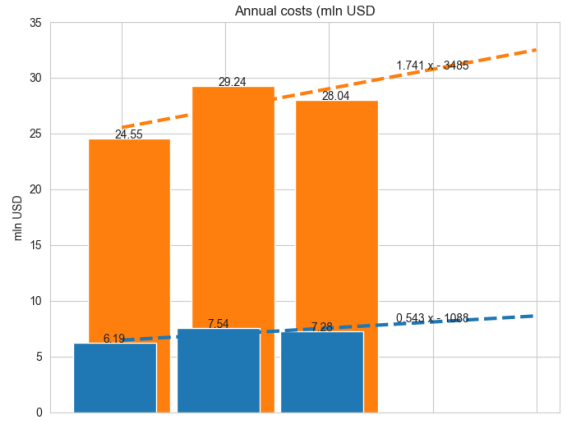
“Yellow Cab” grows in annual trips distance also substantially faster than “Pink Cab” does

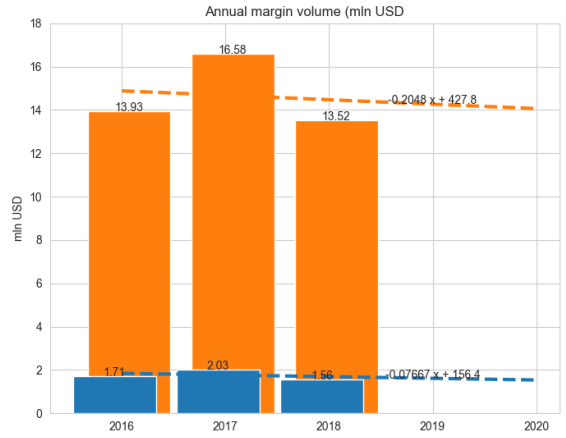
Based on the performance in physical terms, “Yellow Cab” looks better than “Pink Cab”

Data Glacier
Your Deep Learning Partner

Financial indicators







"Yellow Cab" grows in annual revenue substantially faster than "Pink Cab" does

BUT "Yellow Cab"'s annual costs grows substantially faster as well.

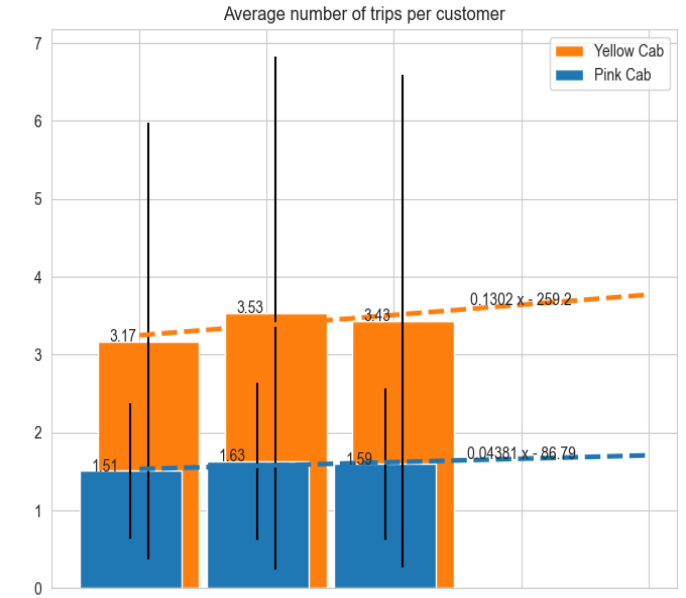
"Yellow Cab"'s margin is around 7x higher than "Pink Cab"'s margin
BUT it is descending faster

Both companies' annual costs grow faster than annual revenues

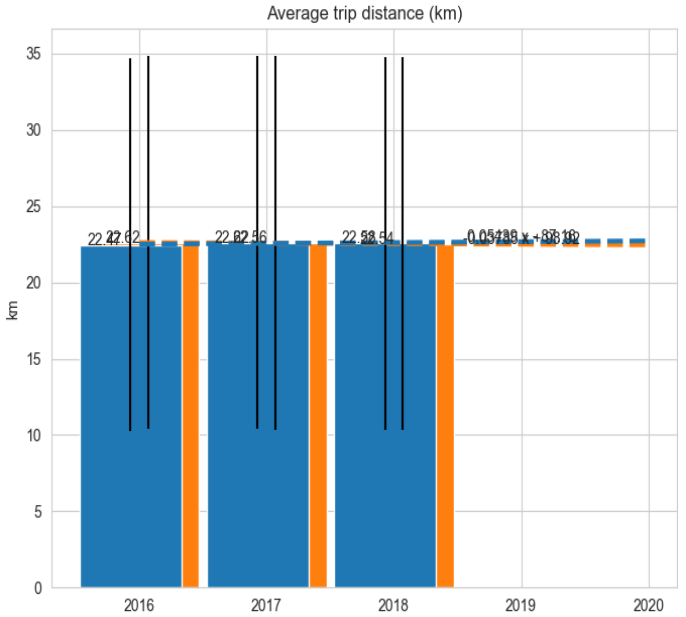
Changes in costs in 2018 looks positive but it is not enough to make a trend

Data Glacier
Your Deep Learning Partner

Specific physical indicators



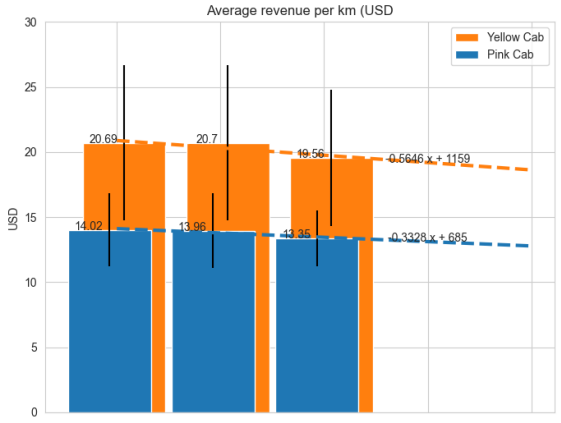
“Yellow Cab” has higher average number of trips per customer than “Pink Cab”.
The slope of the trend of “Yellow Cab” is much steeper also

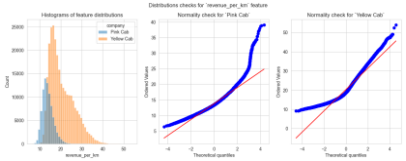


Both companies has average trip distance values near equal.
Standard deviation ranges are also similar

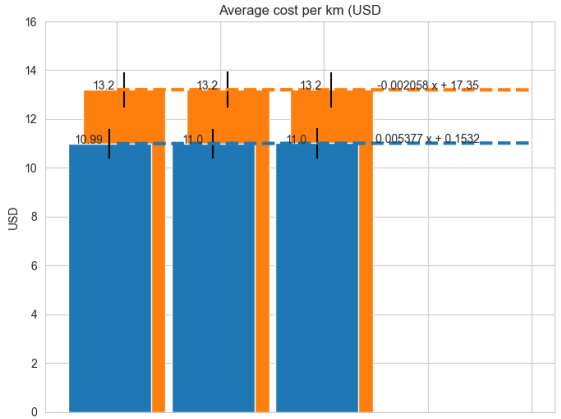
Data Glacier
Your Deep Learning Partner

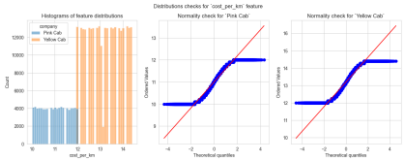
Specific financial indicators



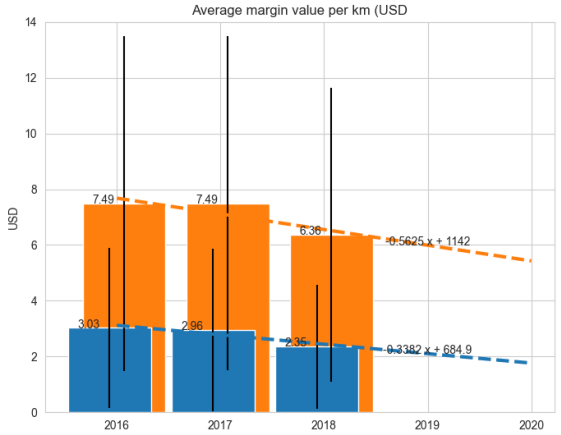


Both companies have their average **revenue per 1 km values decreasing**. "Yellow Cab"'s value is **decreasing faster** than "Pink Cab"





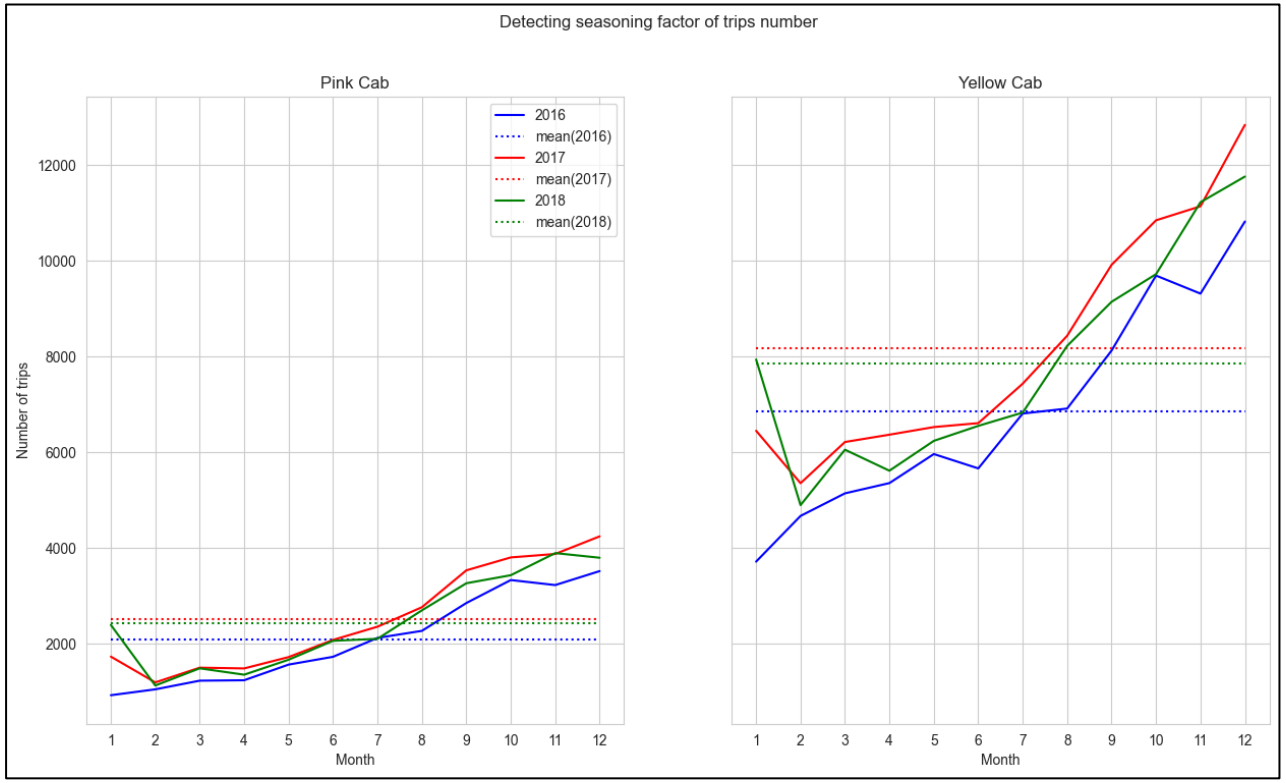
Both companies has **different but stable** average level of cost per 1 km. The **difference is statistically significant**



Both companies have their 1 km **profitability decreasing**
If current trends continue, "Yellow Cab" and "Pink Cab" **will pass the break-even point in 2030 and 2025** respectively

Data Glacier
Your Deep Learning Partner

Seasonality analysis



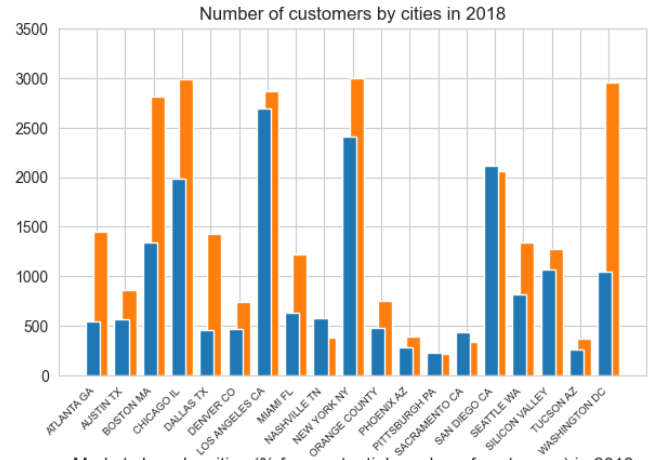
Both companies have **strongly marked seasonality** in their businesses

"Yellow Cab" has **substantially wider seasonal range** in number of trips

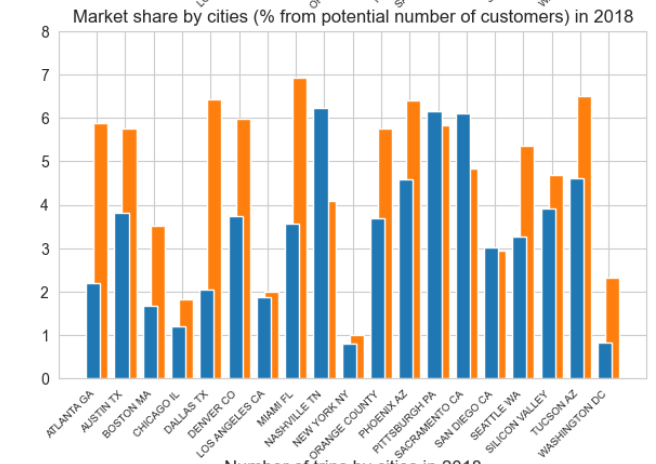
Seasonality was relatively stable in 2016-2018 and had repeating pattern

Data Glacier
Your Deep Learning Partner

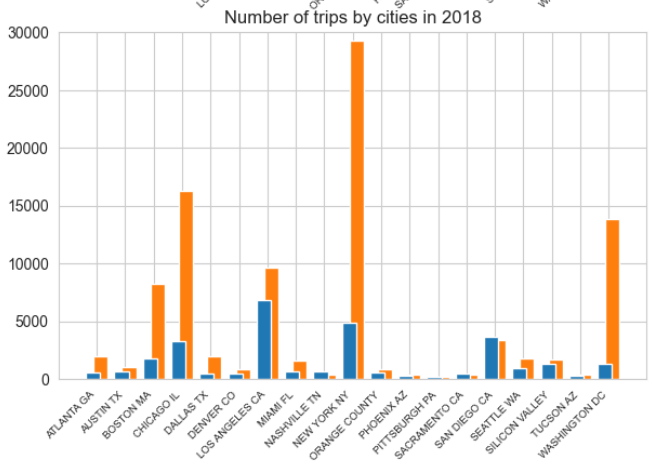
Geographical analysis (Physical indicators)



Both companies have similar number of clients in all cities except of 4 (Atlanta, Boston, Dallas and Washington) in which “Yellow Cab” **significantly outperforms** “Pink Cab”



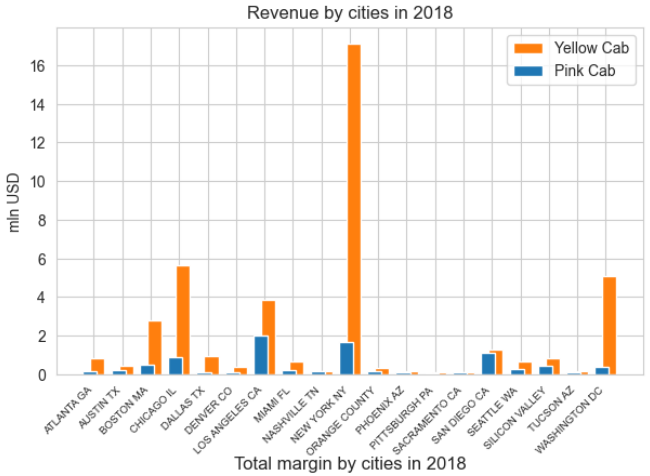
Market share of both companies is not too high but “Yellow Cab” has higher market share in most of cities



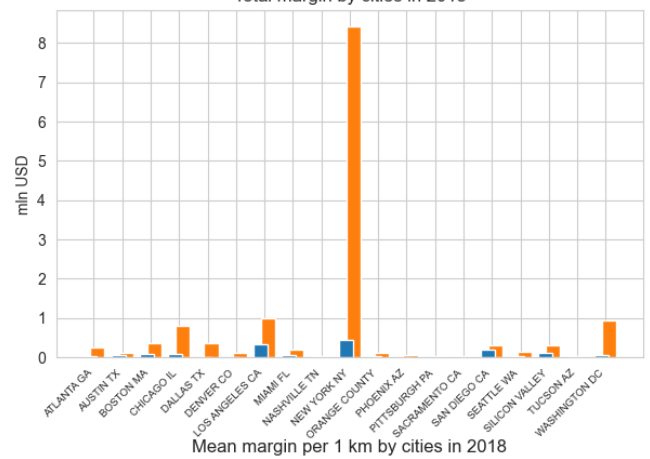
“Yellow Cab” has **abnormally high number of trips** in 4 cities (Boston, Chicago, New York, Washington

Data Glacier
Your Deep Learning Partner

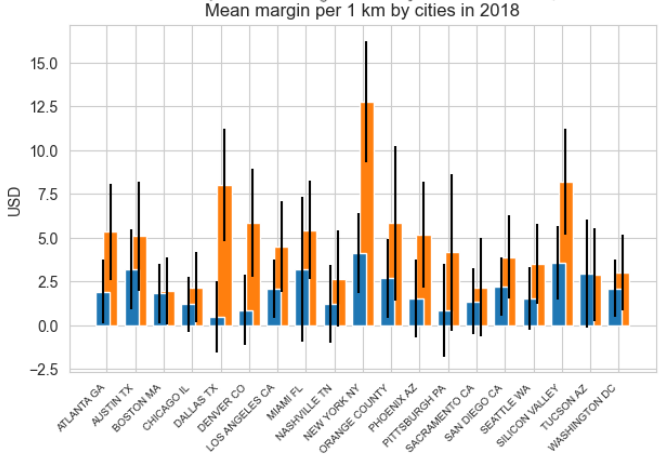
Geographical analysis (Financial indicators)



"Yellow Cab" had **83% of revenue** in 2018 earned in **5 cities** (Boston, Chicago, Los Angeles, New York and Washington) including **41% of revenue in New York alone**.
Revenue structure was **stable in 2016-2018**



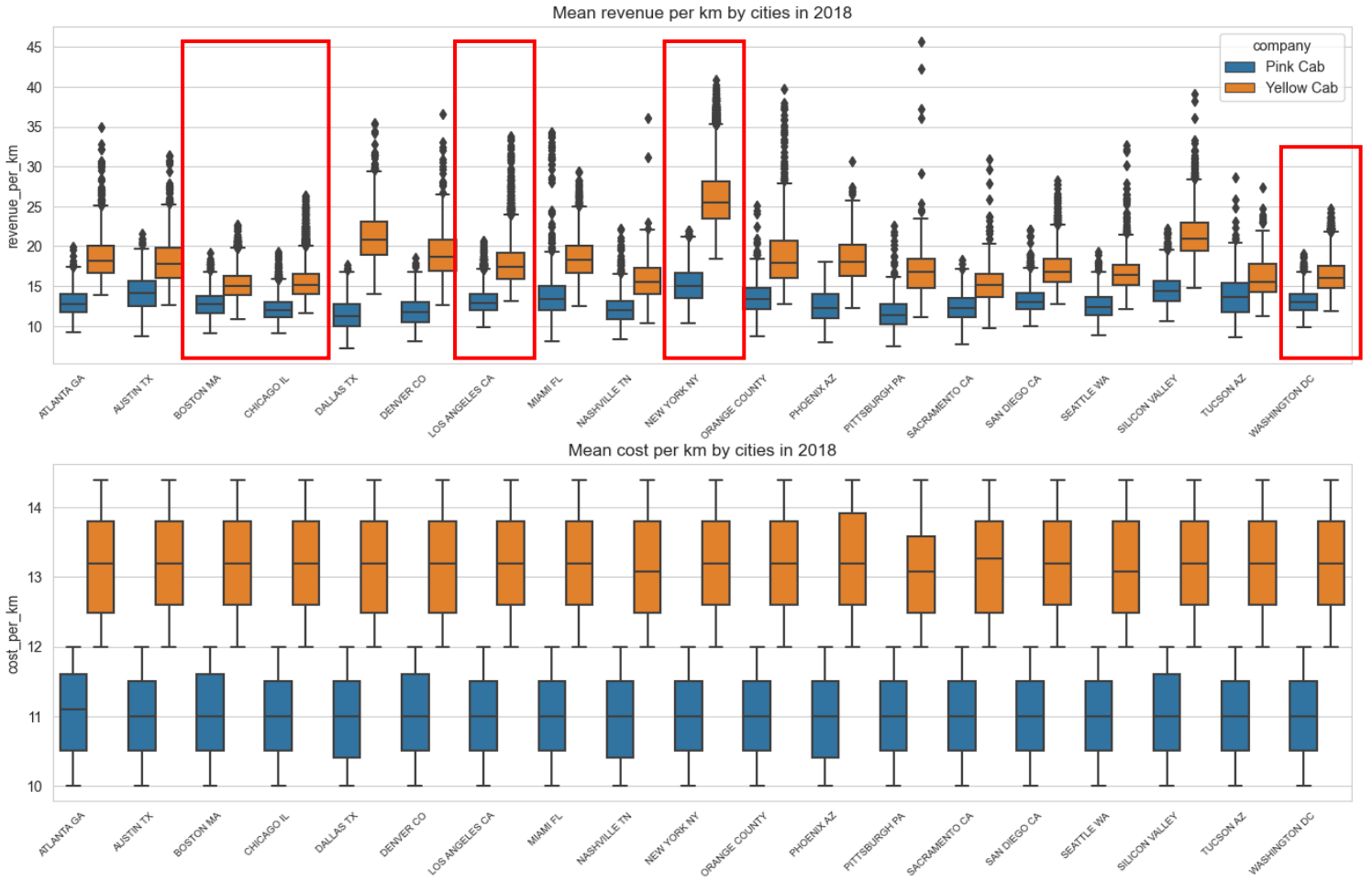
"Yellow Cab" had **85% of margin** in 2018 earned in same **5 cities** including **62% of margin in New York alone**



"Yellow Cab" outperforms "Pink Cab" on margin per 1 km indicator in almost all cities.
In several cities **"Pink Cab"** operates with **near-zero margin**

Data Glacier
Your Deep Learning Partner

Geographical analysis (Exploring margin anomalies)



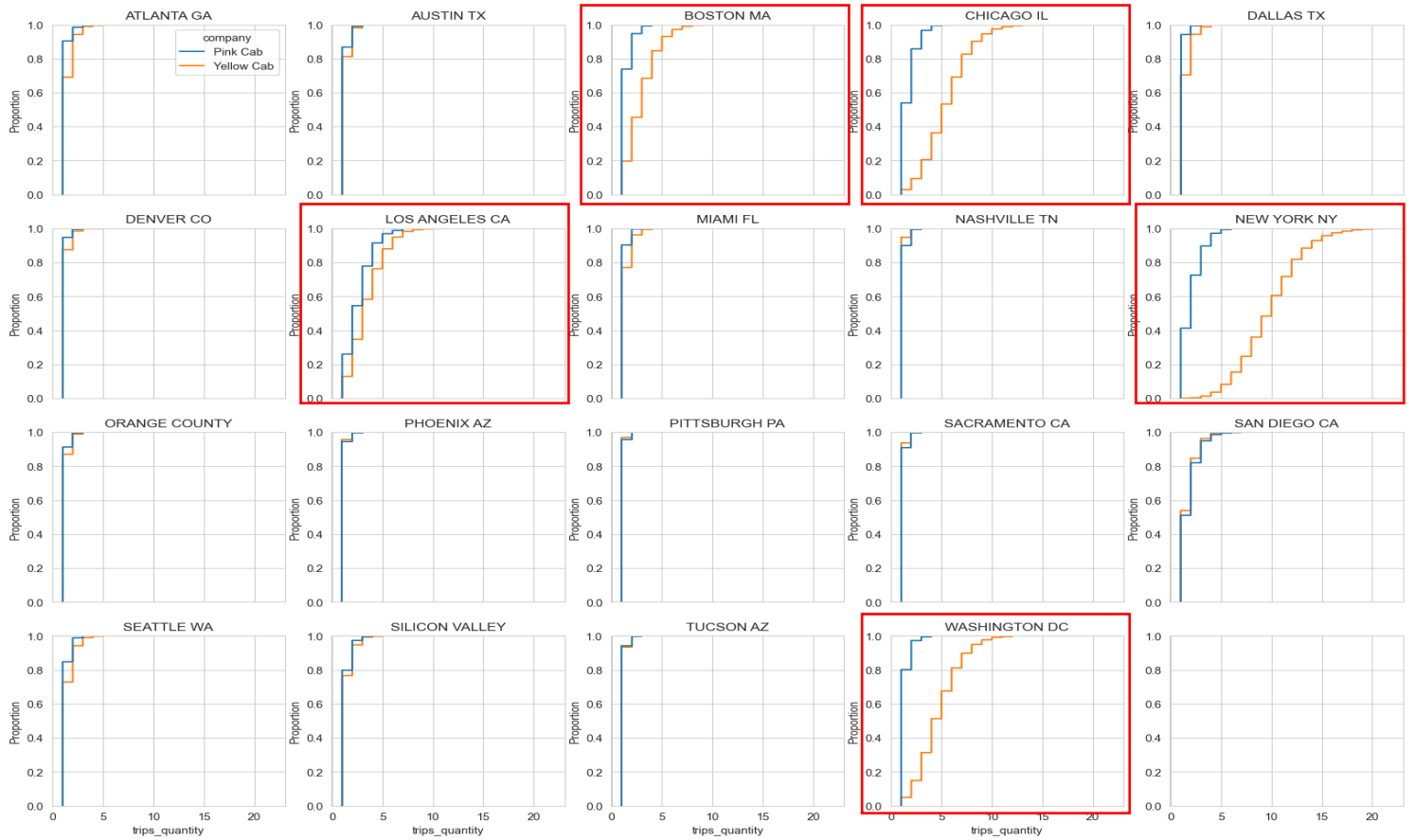
Both companies have the **same average cost per 1 km level across all cities**

In all cities **“Yellow Cab”** demonstrates **higher revenue per 1 km** level. This difference is **statistically significant**

The largest gap between companies’ revenue per 1 km is in New York which is the “most abnormal” city in terms of margin value

Data Glacier
Your Deep Learning Partner

Geographical analysis (Exploring margin anomalies)



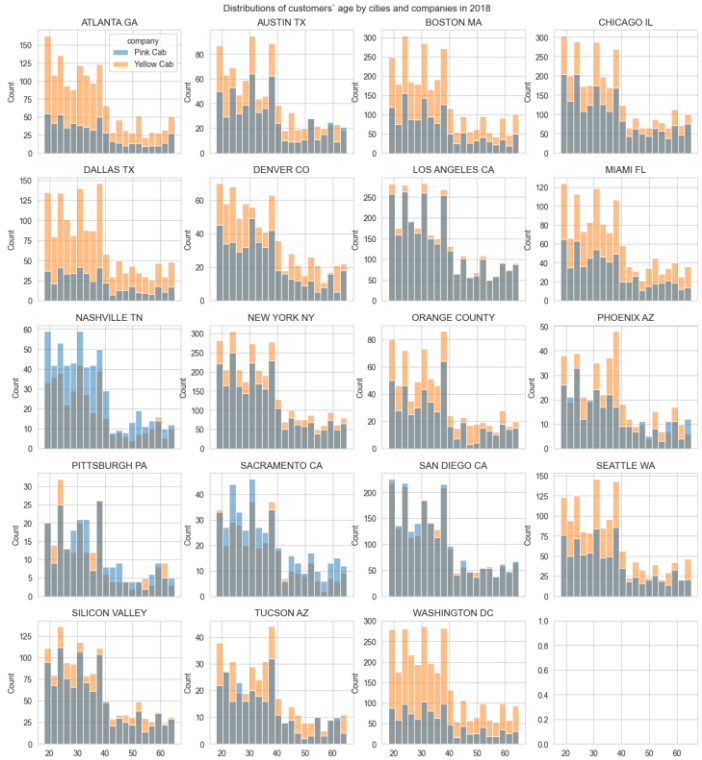
We can clearly see the differences in distributions of the number of trips per customer indicator: as in most cities maximum value didn't exceed 5 trips per customer, in "anomaly" cities it exceeded 10 **up to 20 for "Yellow Cab" in New York**.

Combined with **high margin per 1 km** this gave an outstanding total margin value

This situation demands more exploration as **such a strong dependency on 1 city is risky**

Customers segmentation (Geographical position)

Trying to find dependency between customers' location and income: failed

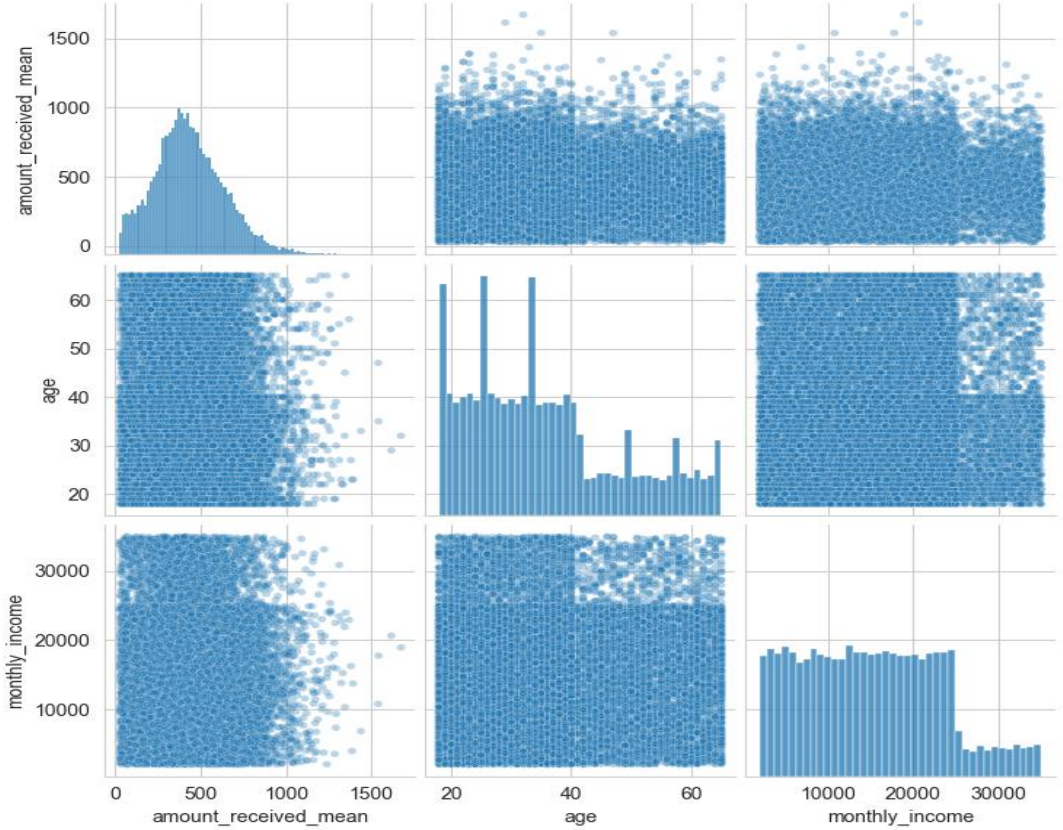


Trying to find dependency between customers' location and age: failed

All distributions are similar and there are no differences in distributions except of frequencies. Thus, we can't use this for classification

Data Glacier
Your Deep Learning Partner

Customers segmentation (other features)



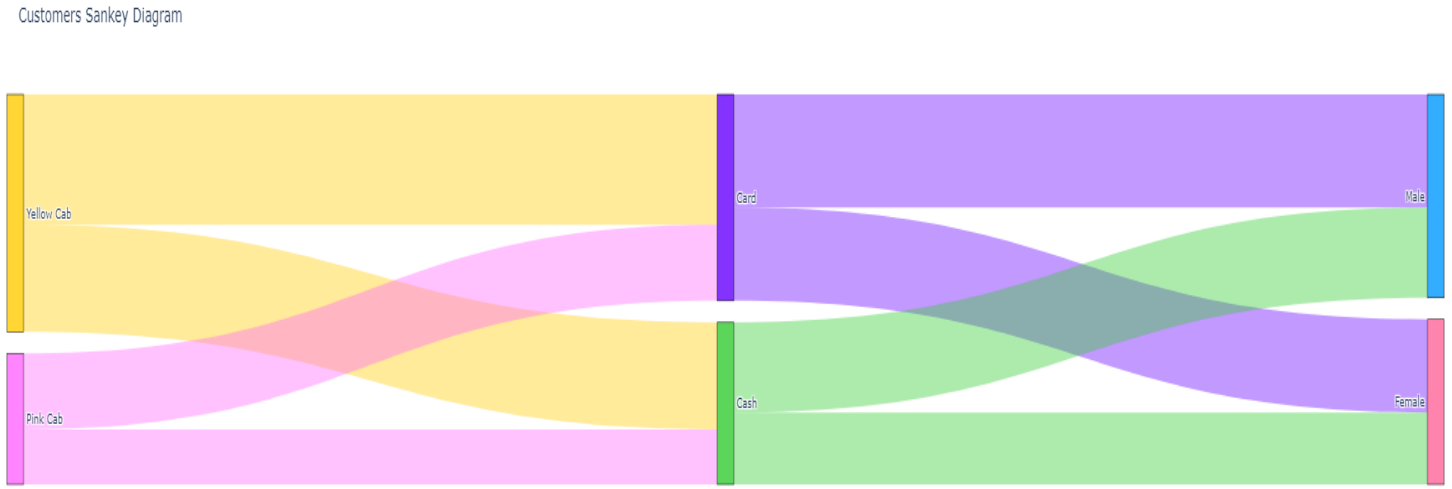
Correlation testing: **failed**

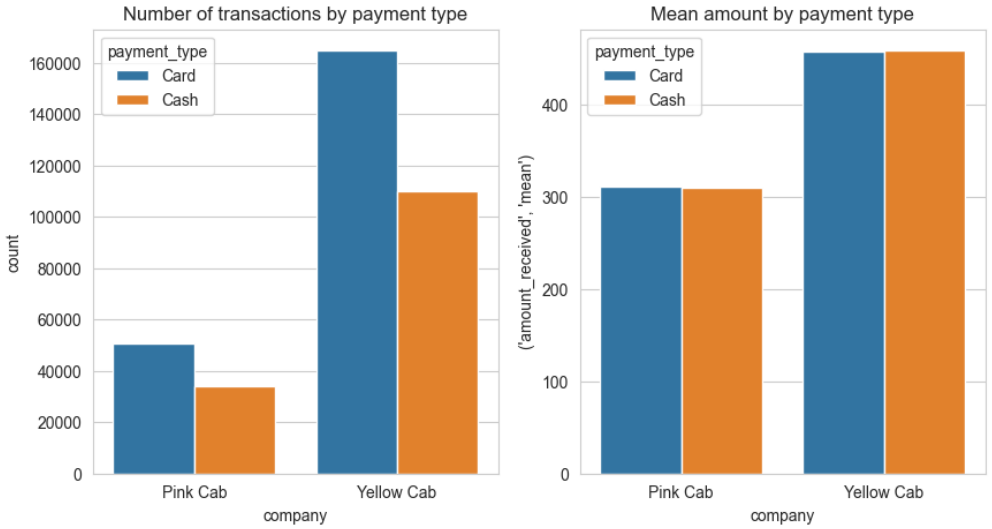
There's **no visible correlation** between age and monthly income features and amount received per trip.

All statistical coefficients are small with respective p-values greater than significance level (0.05) thus displaying **no linear or other correlation between values**

Data Glacier
Your Deep Learning Partner

Customers segmentation (other features)





Trying to find dependencies between company, payment types and customers' gender: **failed**

There's **no visible dependencies** between features being analyzed.

Data Glacier
Your Deep Learning Partner

Conclusions and recommendations

Data Glacier
Your Deep Learning Partner

Let's recap:

- Both companies' annual costs grow faster than annual revenues. Thus, **the margin of both companies is descending**
- "Yellow Cab"'s **margin is around 7x higher** than "Pink Cab"'s margin, but it is **descending faster and is based on 1-city margin**: 62% of margin is in New York alone. This situation demands more exploration as **such a strong dependency on 1 city is risky**
- In physical terms both companies show growth, but **"Yellow Cab" grows faster than "Pink Cab"**
- Both companies have their **1 km profitability decreasing**. If current trends continue, "Yellow Cab" and "Pink Cab" will **pass the break-even point in 2030 and 2025** respectively
- Both companies have **strongly marked seasonality** in their businesses. Seasonality is **relatively stable** in 2016-2018 and had repeating pattern
- Market share of both companies is not too high but **"Yellow Cab" has higher market share in most cities**
- In all cities **"Yellow Cab"** demonstrates **higher revenue per 1 km** level

Recommendations:

Considering descending margin of both companies, low profitability of "Pink Cab" and 1-city dependency of "Yellow Cab", **DO NOT invest in neither "Yellow Cab" nor "Pink Cab"** until full understanding of margin drop reasons and possibilities to rectify the situation

THANK YOU