Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

Intern: Samuel Cueva Lozano

26/06/2021

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Data Glacier

Your Deep Learning Partner

Executive 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.

Problem Statement

There are two cab companies which XYZ wants to invest in, so they require actionable insights to help them identify the right company to make their investment.

Cab Companies:

- Yellow Cab
- Pink Cab

Analysis:

EDA

- Descriptive Analysis.
- Correlation Analysis.
- Contextual Analysis.
- Recommendations.

Approach and Assumptions

Duplicate values

A record in the data set will be considered duplicate if there is a transaction with the same kilometers traveled, the same city, the same company and the same day.

Join tables

The table Cab_Data.csv will be joined with the table Customer_ID.csv by the table Transaction_ID.csv since the latter has the Transaction ID (from Cab_Data.csv) and Customer ID (from Customer_ID.csv) attributes, then the table City.csv will also join them.

Null values

The Empty values, NaN values and values like “?” will be treated as Null values

Fields with too many null values would be removed or filled according to the following:

- Fields like Age, Gender, Income, Company and Users would be filled because they have missing values that depends on their hypothetical values or depend on other attributes.
- The rest of fields would be removed because they have missing values due to bad configuration, issues with data collection, or untraceable random reasons.

After cleaning at the field level, rows with null values would be removed.

Approach and Assumptions

Outlier detection

Outliers will be removed from numerical fields so that they don’t negatively affect the analysis.

A fixed threshold would be used for the Age, Population and Users attribute to avoid inconsistent data, then the IQR Score will be used to filter out the outliers in all attributes.

Data Transformation

The field Date of Travel from the table Cab_Data.csv could be transformed to a more readable format for better understanding.

Exploratory Data Analysis

Descriptive Analysis

Correlation Analysis

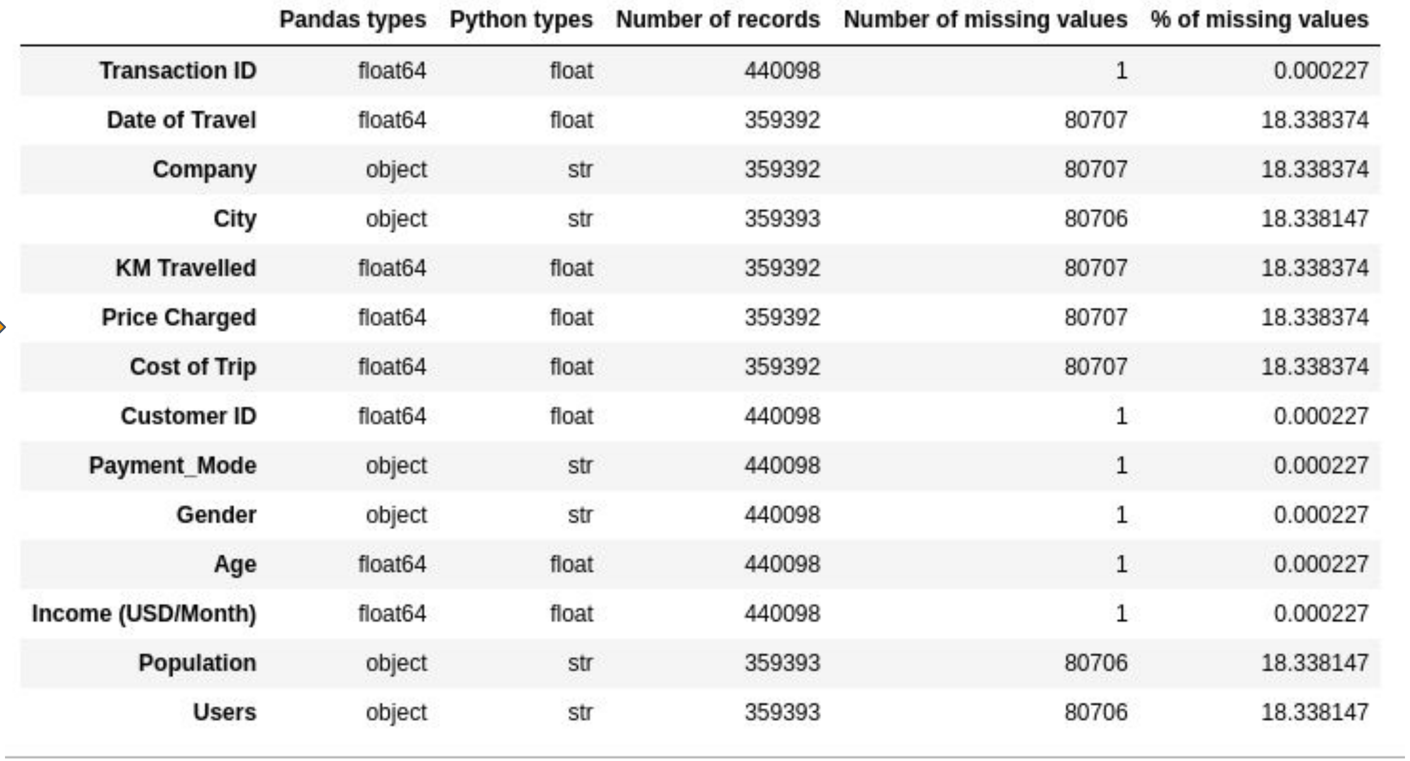
Contextual Analysis

Data

There are four datasets :

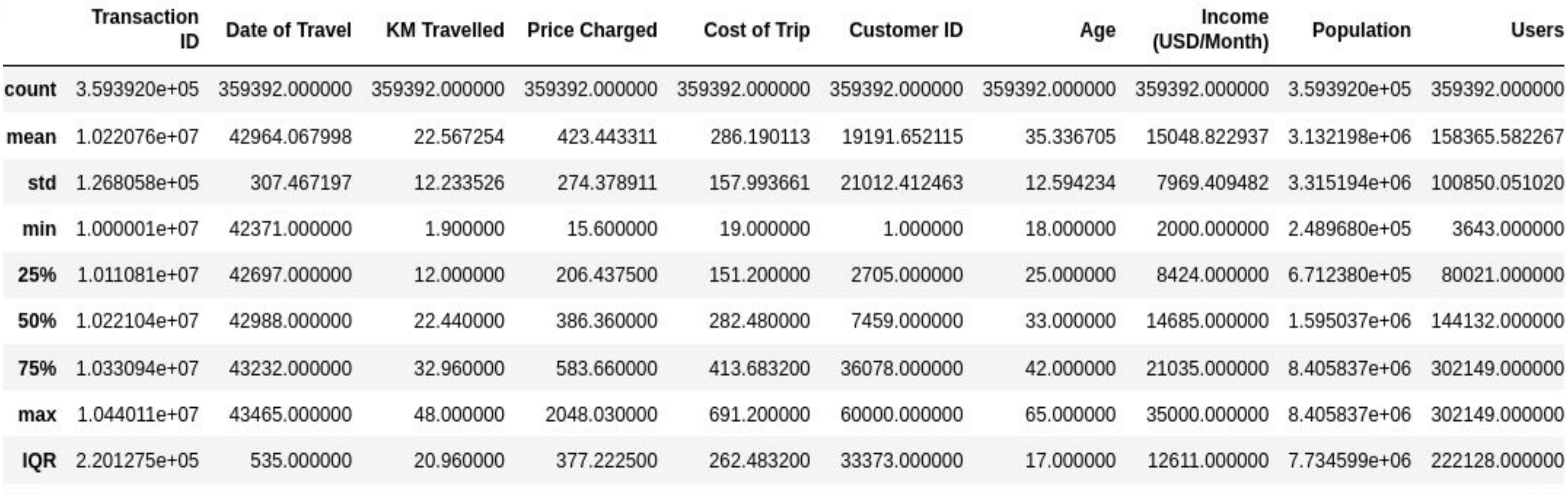
- **Cab_Data.csv –** this file includes details of transaction for 2 cab companies
- **Customer_ID.csv** – this is a mapping table that contains a unique identifier which links the customer’s demographic details
- **Transaction_ID.csv –** this is a mapping table that contains transaction to customer mapping and payment mode
- **City.csv –** this file contains list of US cities, their population and number of cab users

Location:
https://github.com/DataGlacier/DataSets



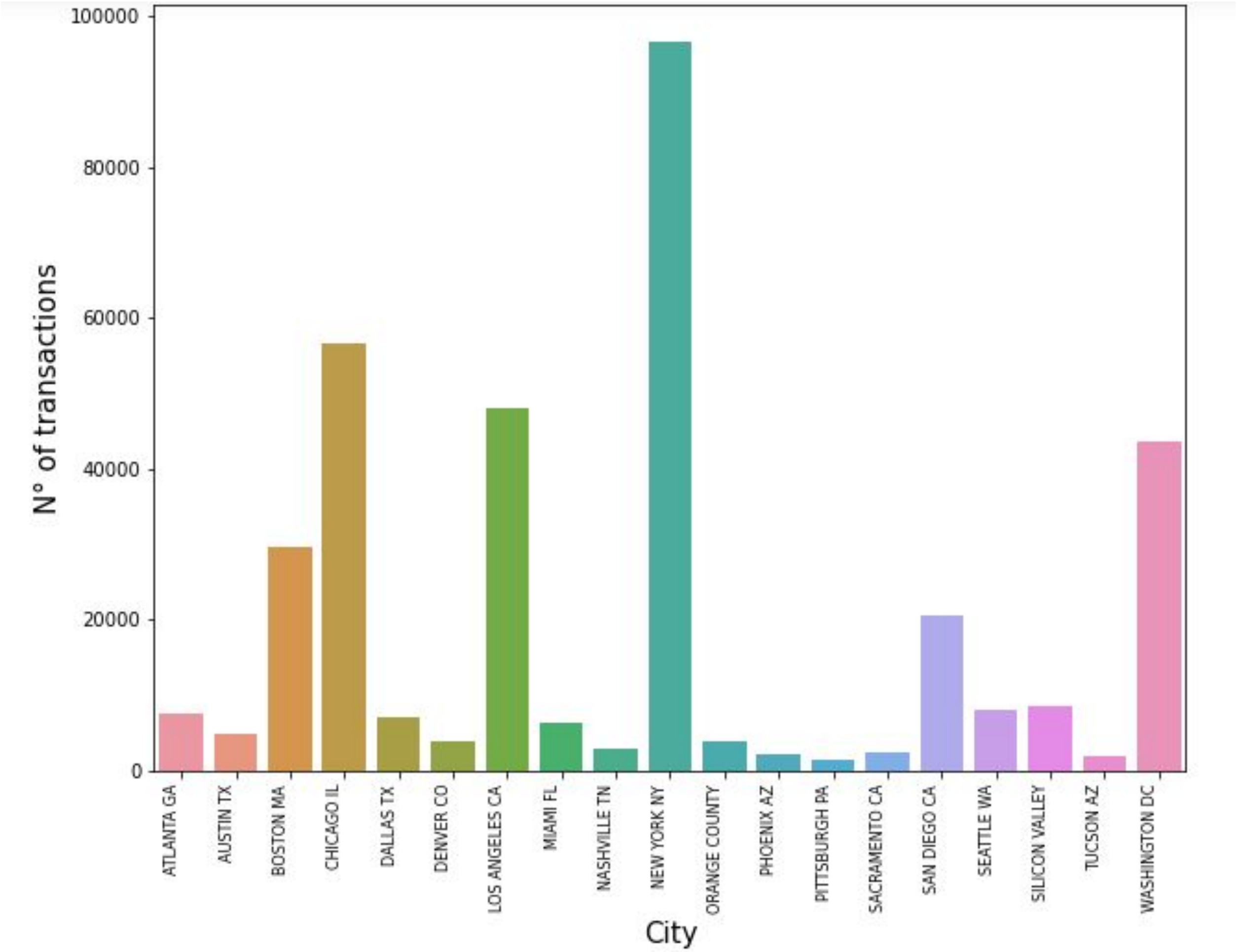
	Pandas types	Python types	Number of records	Number of missing values	% of missing values
Transaction ID	float64	float	440098	1	0.000227
Date of Travel	float64	float	359392	80707	18.338374
Company	object	str	359392	80707	18.338374
City	object	str	359393	80706	18.338147
KM Travelled	float64	float	359392	80707	18.338374
Price Charged	float64	float	359392	80707	18.338374
Cost of Trip	float64	float	359392	80707	18.338374
Customer ID	float64	float	440098	1	0.000227
Payment_Mode	object	str	440098	1	0.000227
Gender	object	str	440098	1	0.000227
Age	float64	float	440098	1	0.000227
Income (USD/Month)	float64	float	440098	1	0.000227
Population	object	str	359393	80706	18.338147
Users	object	str	359393	80706	18.338147

Descriptive Analysis

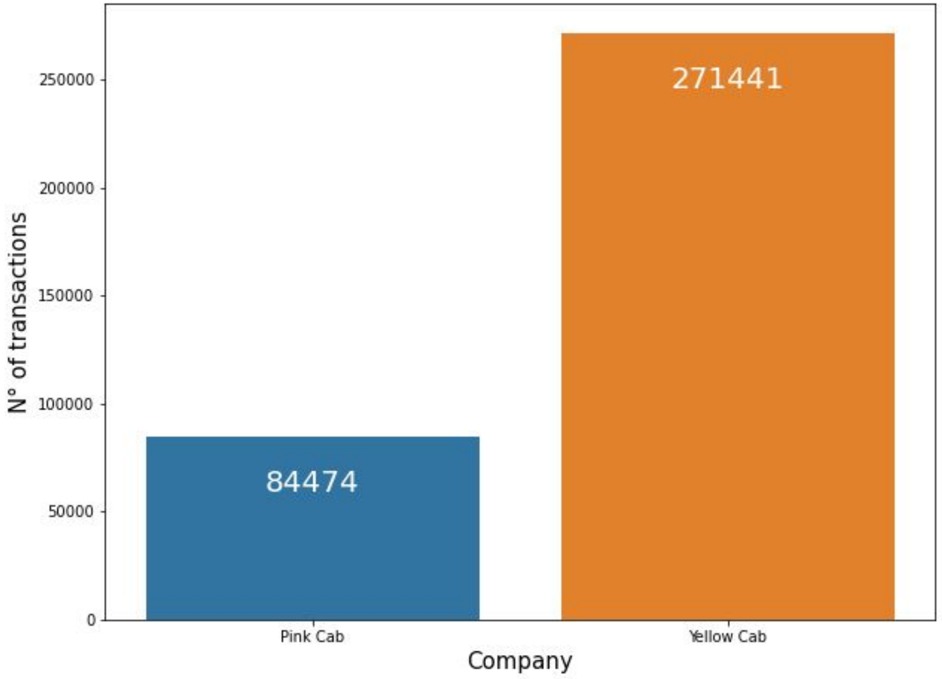


	Transaction ID	Date of Travel	KM Travelled	Price Charged	Cost of Trip	Customer ID	Age	Income (USD/Month)	Population	Users
count	3.593920e+05	359392.000000	359392.000000	359392.000000	359392.000000	359392.000000	359392.000000	359392.000000	3.593920e+05	359392.000000
mean	1.022076e+07	42964.067998	22.567254	423.443311	286.190113	19191.652115	35.336705	15048.822937	3.132198e+06	158365.582267
std	1.268058e+05	307.467197	12.233526	274.378911	157.993661	21012.412463	12.594234	7969.409482	3.315194e+06	100850.051020
min	1.000001e+07	42371.000000	1.900000	15.600000	19.000000	1.000000	18.000000	2000.000000	2.489680e+05	3643.000000
25%	1.011081e+07	42697.000000	12.000000	206.437500	151.200000	2705.000000	25.000000	8424.000000	6.712380e+05	80021.000000
50%	1.022104e+07	42988.000000	22.440000	386.360000	282.480000	7459.000000	33.000000	14685.000000	1.595037e+06	144132.000000
75%	1.033094e+07	43232.000000	32.960000	583.660000	413.683200	36078.000000	42.000000	21035.000000	8.405837e+06	302149.000000
max	1.044011e+07	43465.000000	48.000000	2048.030000	691.200000	60000.000000	65.000000	35000.000000	8.405837e+06	302149.000000
IQR	2.201275e+05	535.000000	20.960000	377.222500	262.483200	33373.000000	17.000000	12611.000000	7.734599e+06	222128.000000

Descriptive Analysis



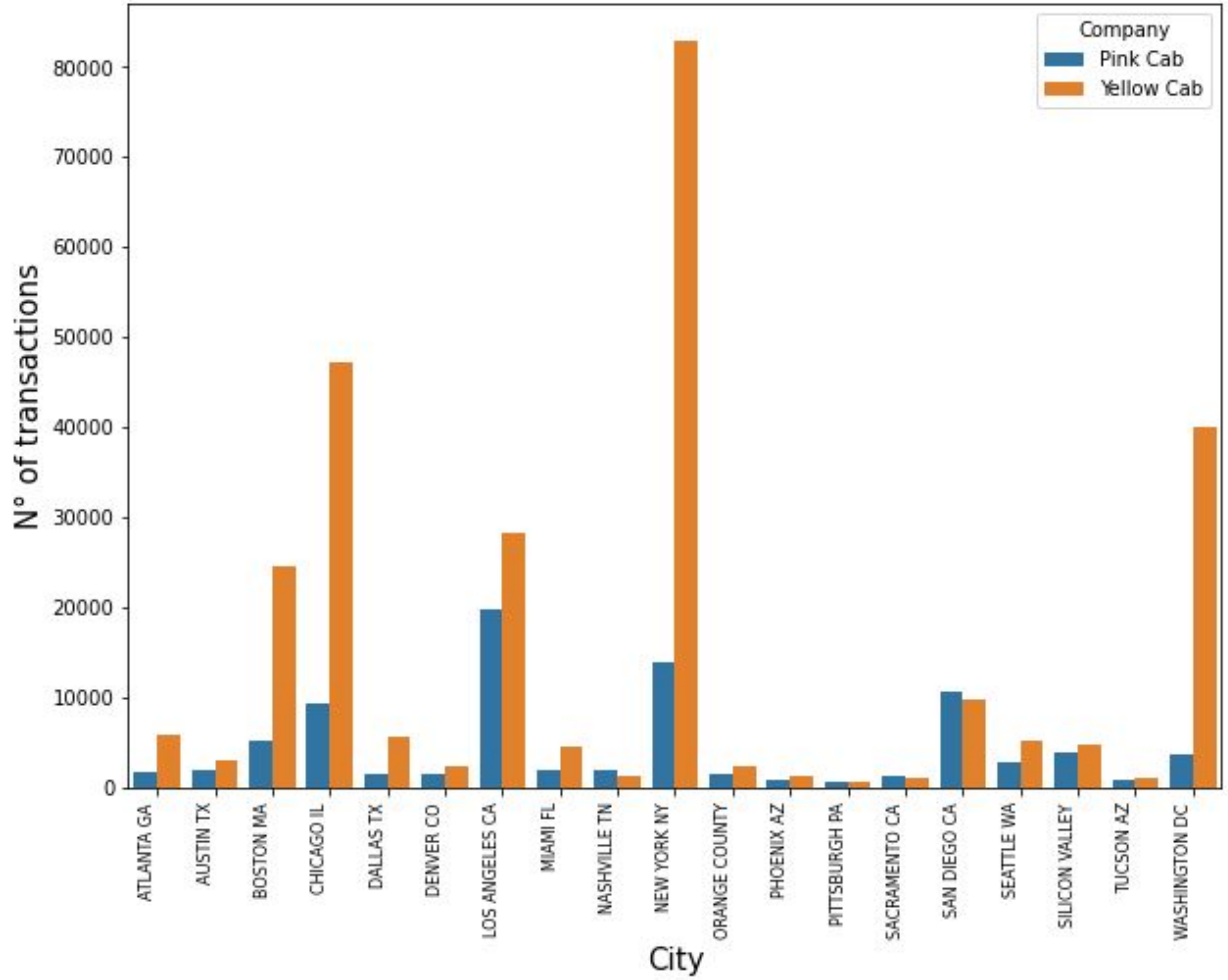
- It can be seen that the number of total transactions in Yellow Cab is much larger than Pink Cab
- There is a greater number of transactions in cities like New York, Chicago, Los Angeles, Washington DC and Boston.

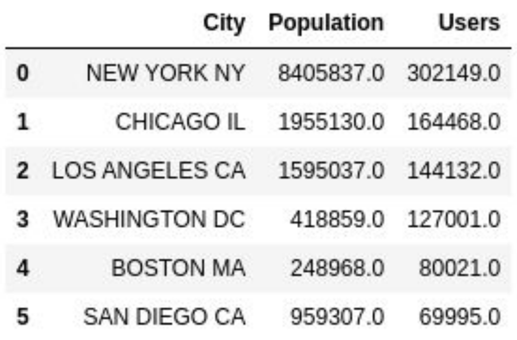


Data Glacier

Your Deep Learning Partner

Correlation Analysis (categorical vs categorical)

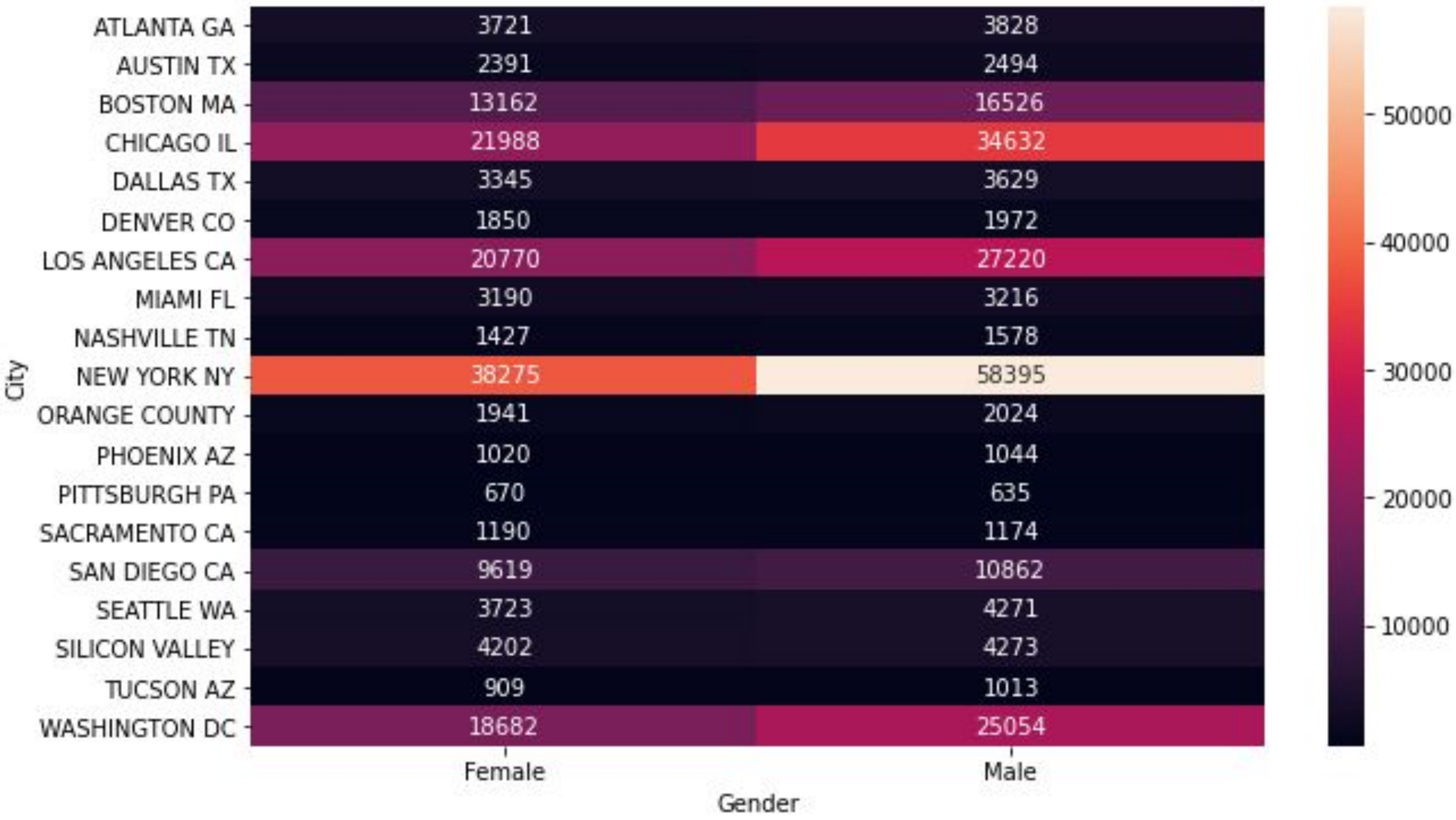




	City	Population	Users
0	NEW YORK NY	8405837.0	302149.0
1	CHICAGO IL	1955130.0	164468.0
2	LOS ANGELES CA	1595037.0	144132.0
3	WASHINGTON DC	418859.0	127001.0
4	BOSTON MA	248968.0	80021.0
5	SAN DIEGO CA	959307.0	69995.0

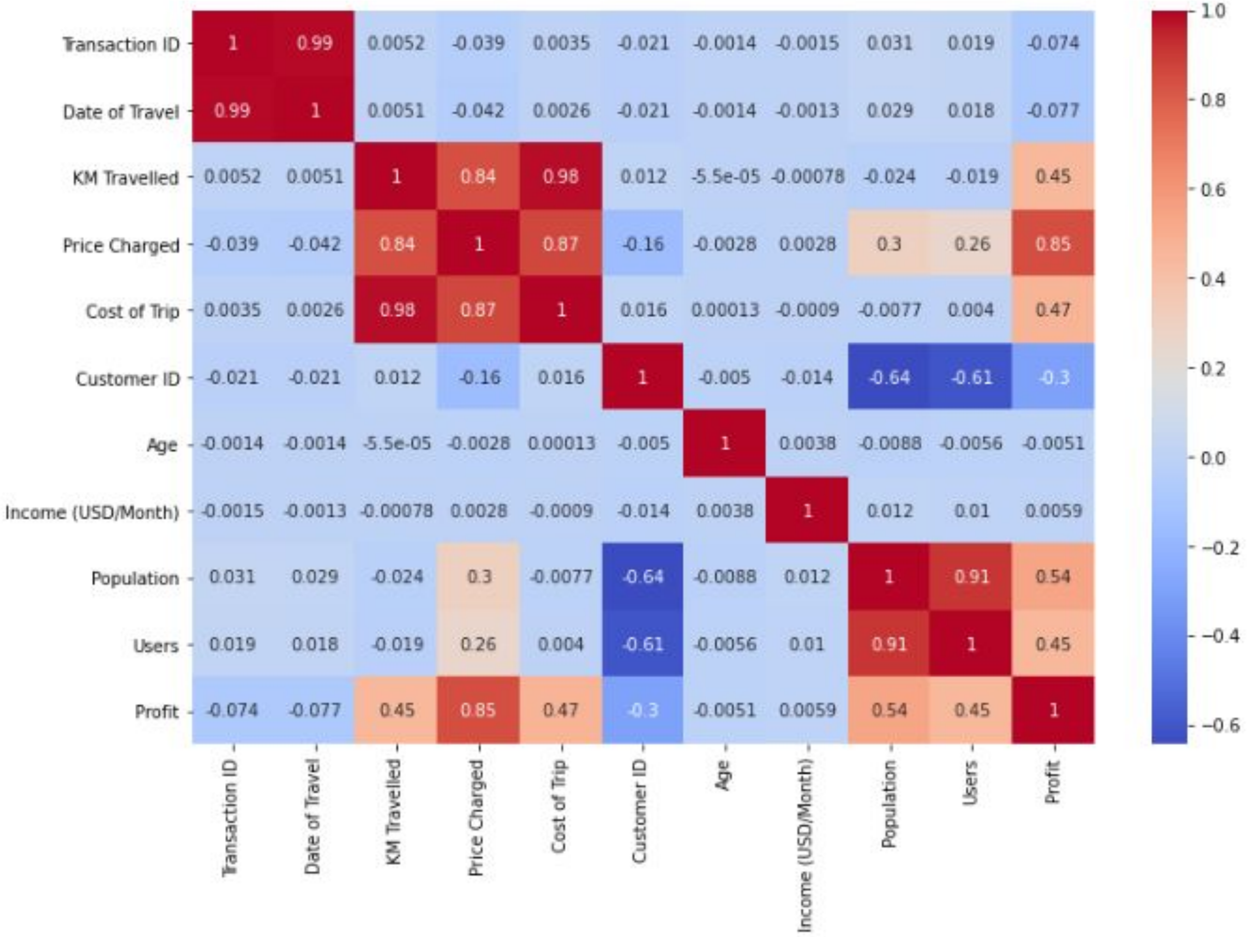
- The Yellow Cab company has a higher number of transactions in the 5 cities with the most population and the most taxi users.
- The Pink Cab company competes with company Yellow Cab in number of trips (transactions) in cities such as: Los Angeles and San Diego and only has a higher number of transactions in San Diego.

Correlation Analysis (categorical vs categorical)



It shows that there are more transactions carried out by men compared to women in cities with more users and with more population(New York,Chicago,Los Angeles,Washington and Boston)

Correlation Analysis (Numerical vs Numerical)



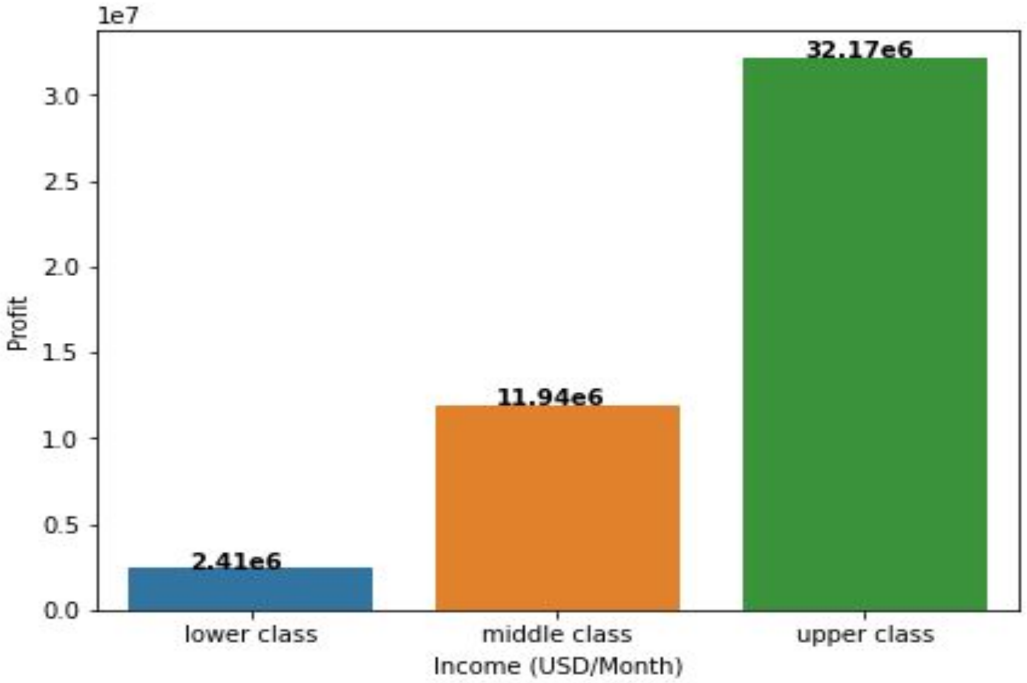
The variable **Profit** will be added to improve the analysis, this variable these would be used if an ML model is built.
Using the Heat map above, the following can be observed:

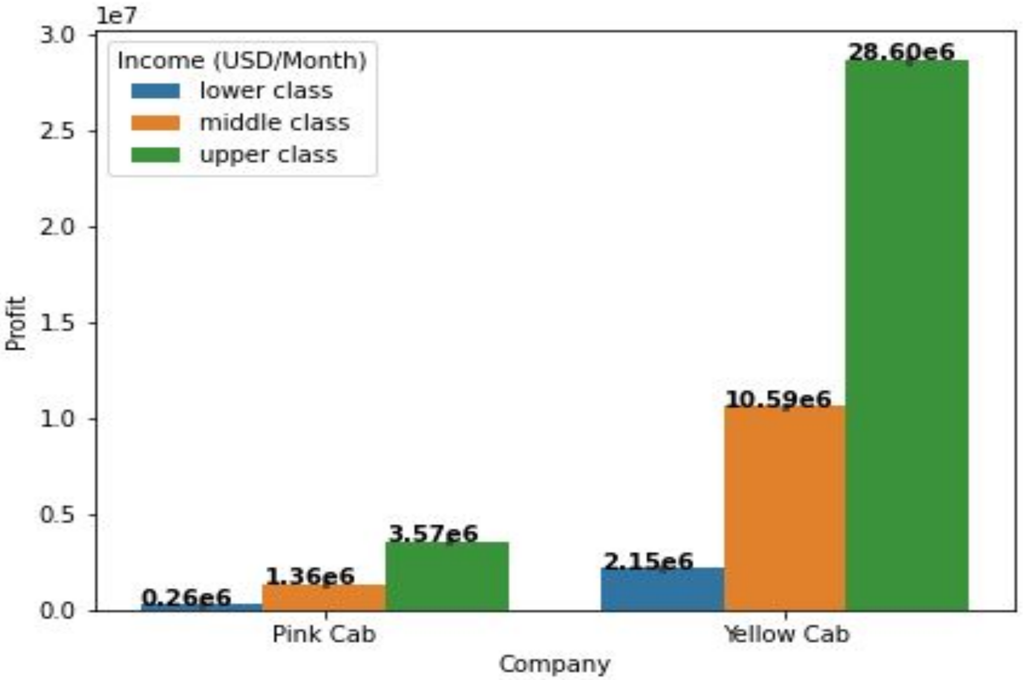
- The variables **Population** and **Users** are strongly correlated, so **Population** would be discarded to build the model.
- The new variable **Profit** is strongly correlated with the variables **Price Charged** as expected.
- The new variable **Profit** has a moderate positive correlation with **KM Travelled, Cost of Trip**.
- **Transaction ID** and **Customer ID** are unique identifiers, they would not be used when building the ML model,but they can be used in Contextual Analysis.
- the variable **Income** and **Age** have no correlation with the objective, so these would be discarded or transformed to build the model, they can be used in Customer Analysis.
- **KM Travelled** and **Cost of Trip** are strongly correlated, so **KM Travelled** would not used when building the ML model, however if a Neural Network is used as algorithm this feature could be considered.

Data Glacier
Your Deep Learning Partner

Correlation Analysis(Categorical vs Numerical)

Income class-wise profit analysis

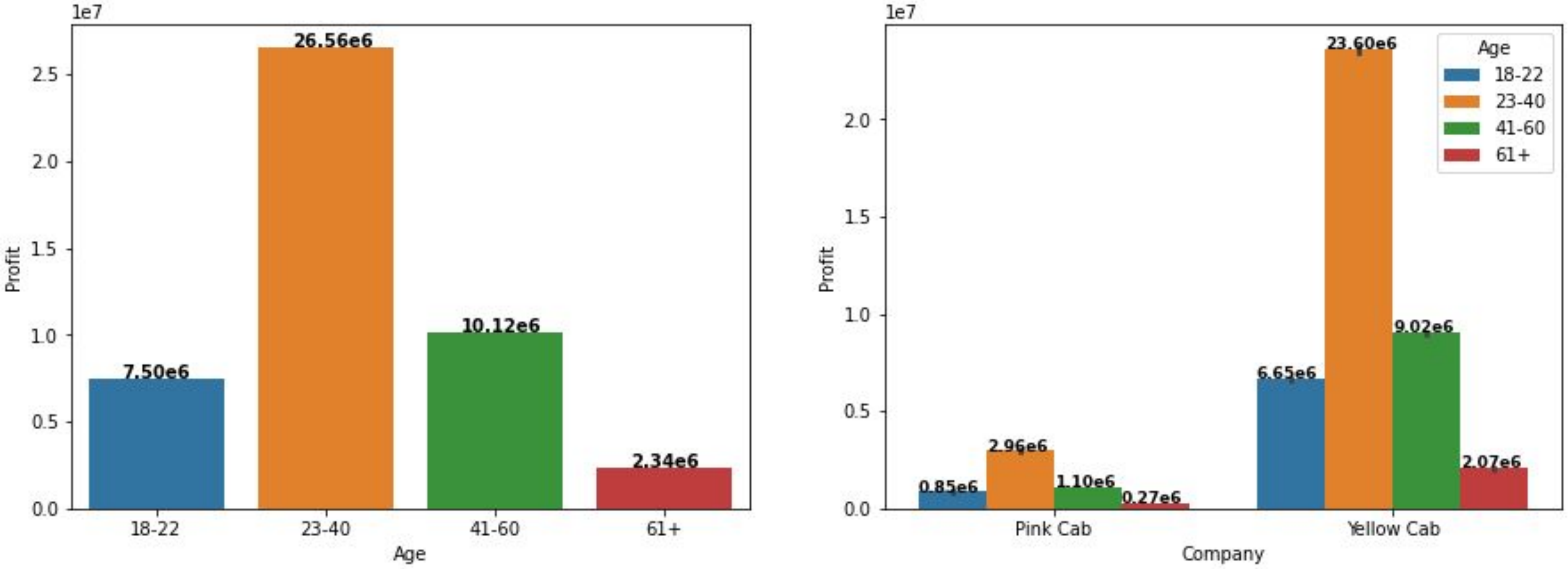




It can be seen that the Upper class produce much more profit, followed by the Middle class.

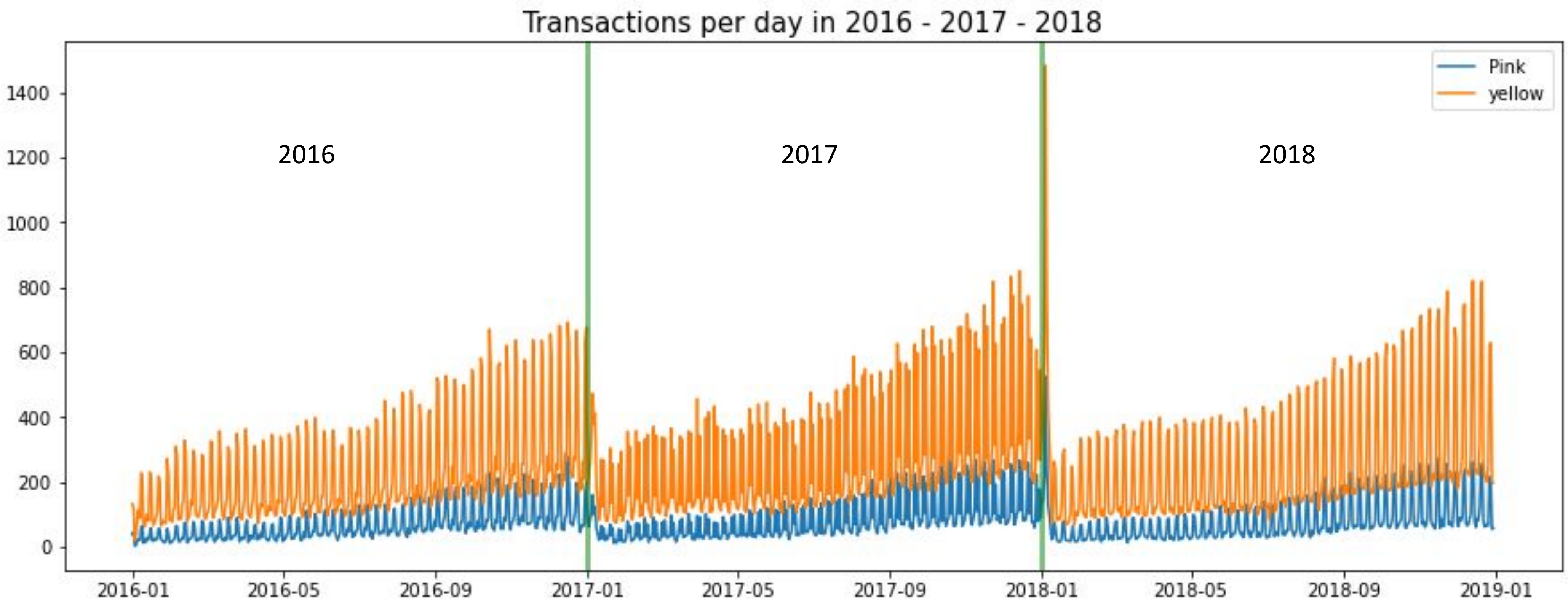
Correlation Analysis(Categorical vs Numerical)

Profit analysis by age



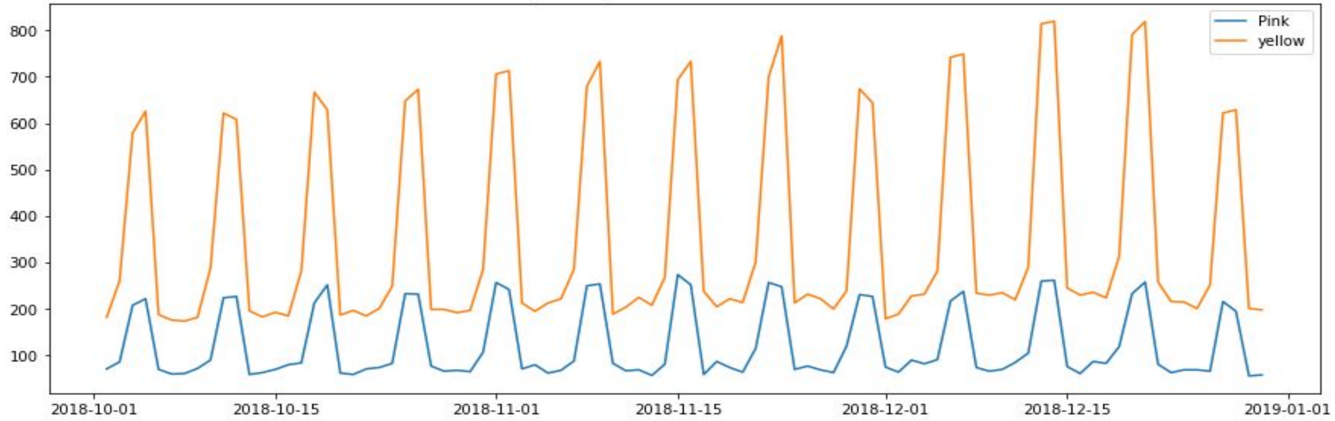
The segment of customers from 23 to 40 years is the one that generates the most profit in both companies

Contextual Analysis

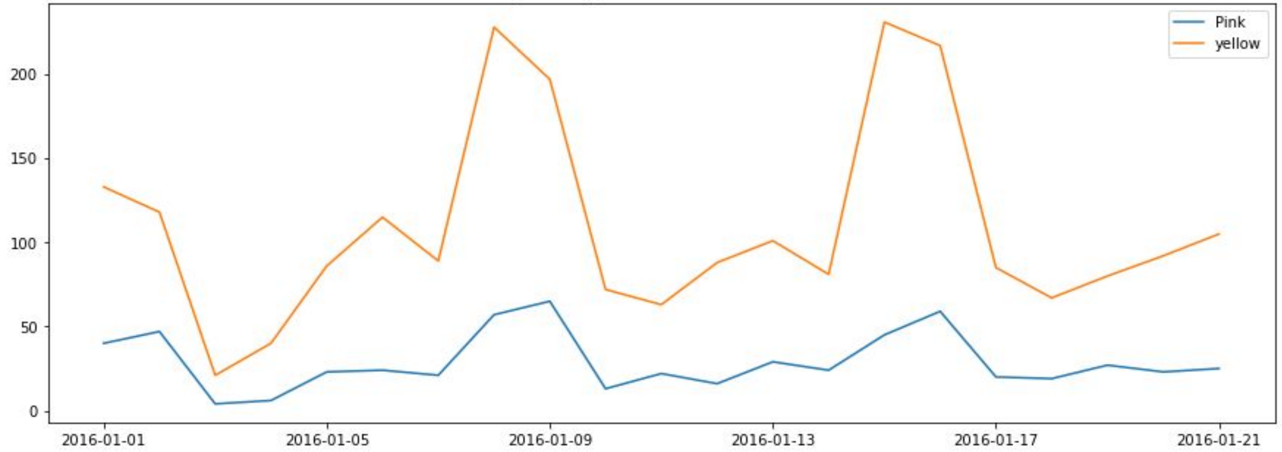


- It can be seen that the higher consumption of the service in both companies increases as the year progresses, with the month of December being stronger.
- Company Yellow has more consumption of its service at each stage of the year.

Contextual Analysis



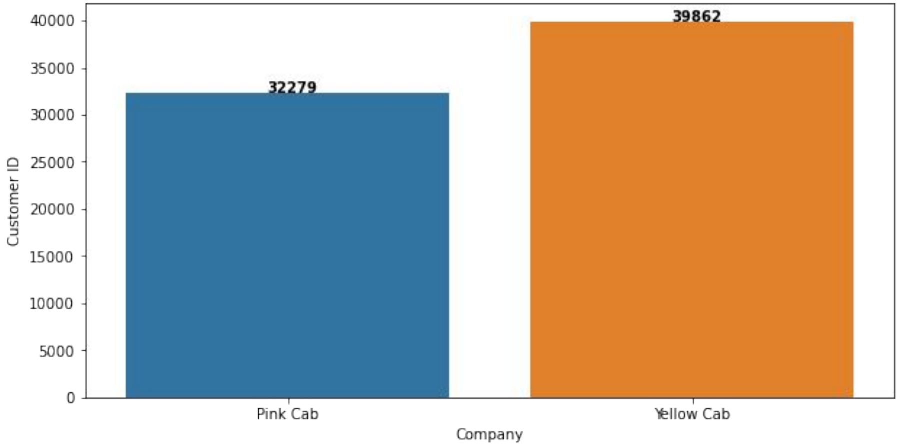
Transactions per day in the last three months in 2018



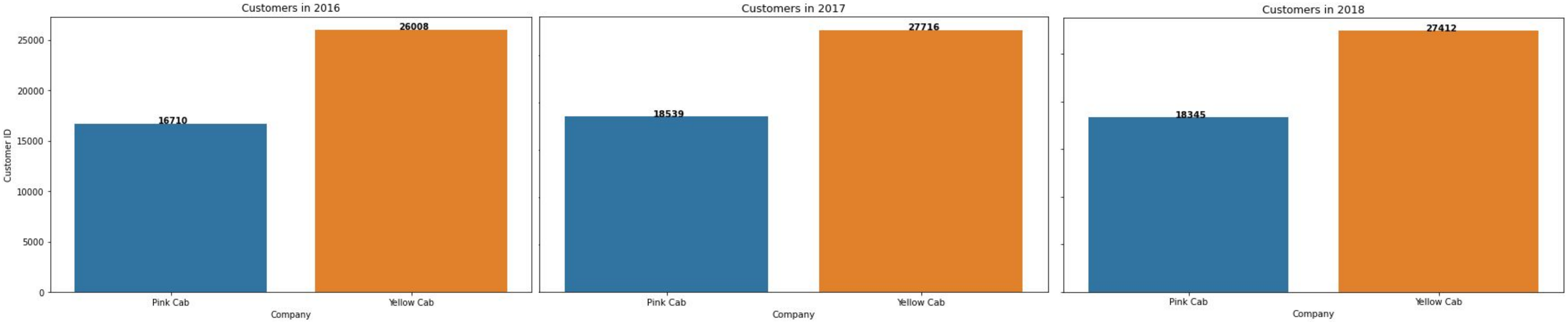
Transactions per day in the first three weeks in 2016

It can be seen that the highest consumption of the service occurs on weekends (especially Friday and Saturday) and that on weekdays consumption is minimal, this pattern is repeated every week throughout the year and both companies

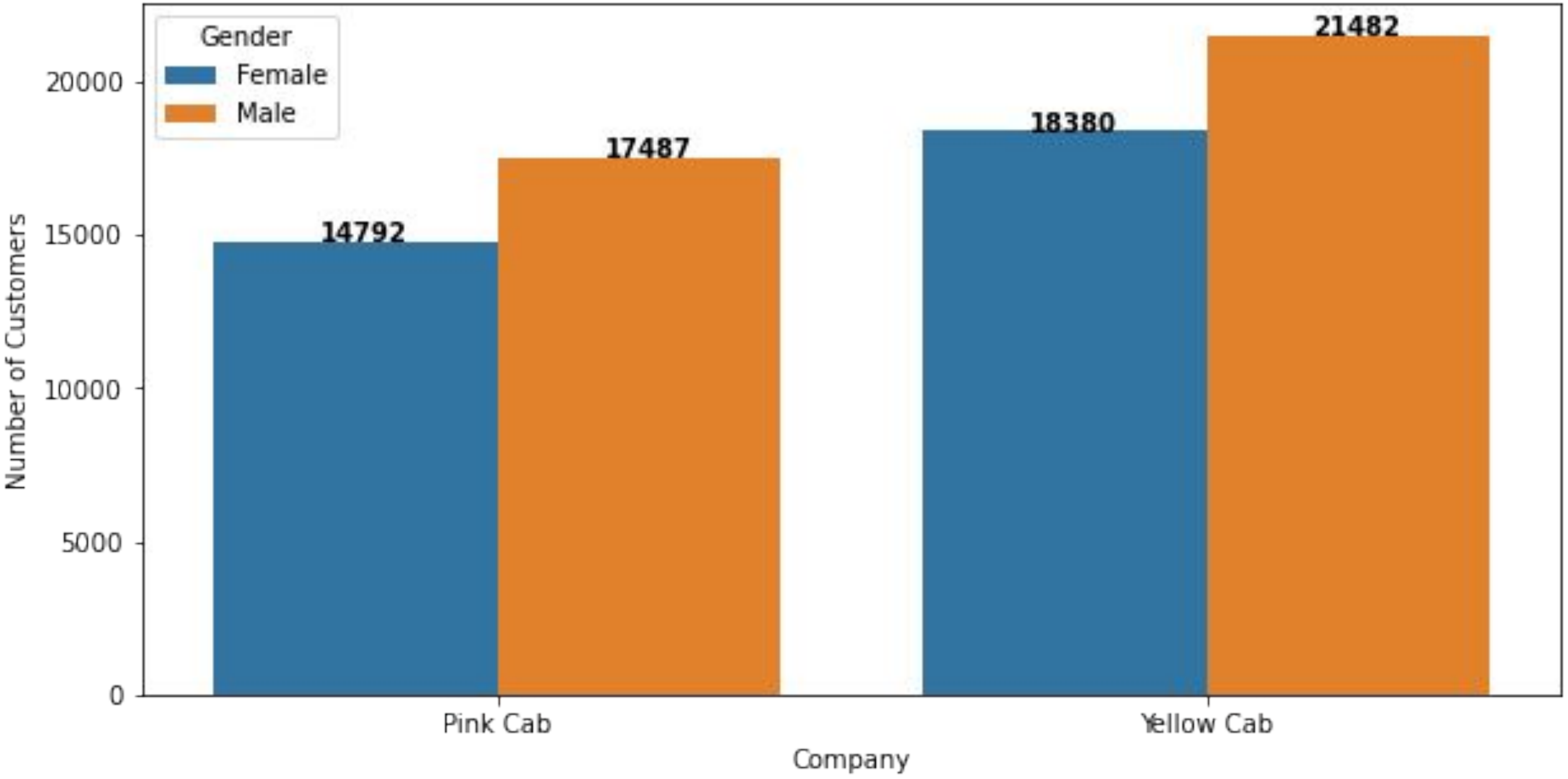
Contextual Analysis



- Total number of customers : 46104
- The total number of clients of company Yellow is slightly higher than that of company Pink, which means that many of the users of company Yellow also use company Pink.
- The number of Yellow Cab customers is greater than the number of Pink Cab customers in each year.

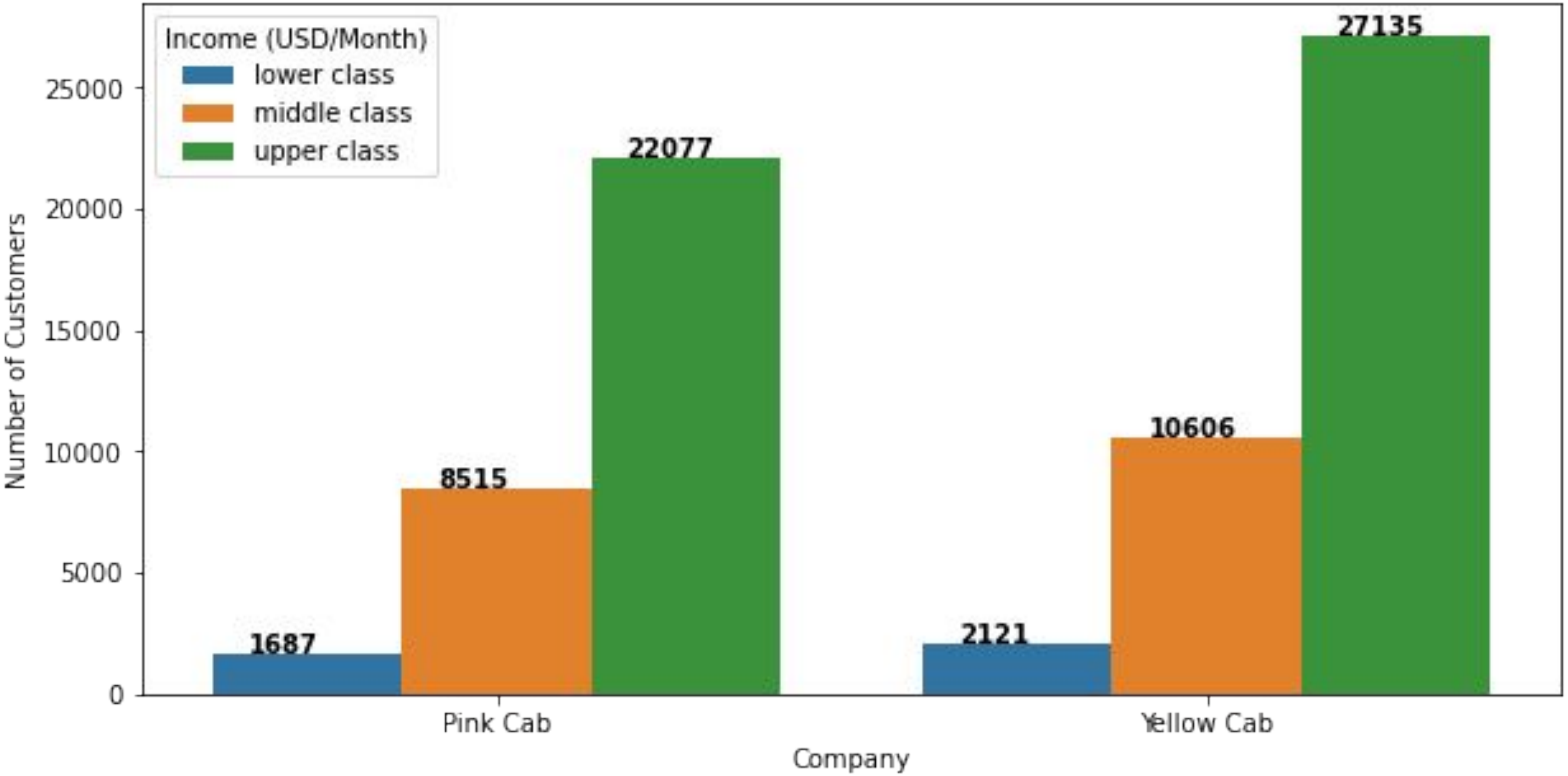


Contextual Analysis



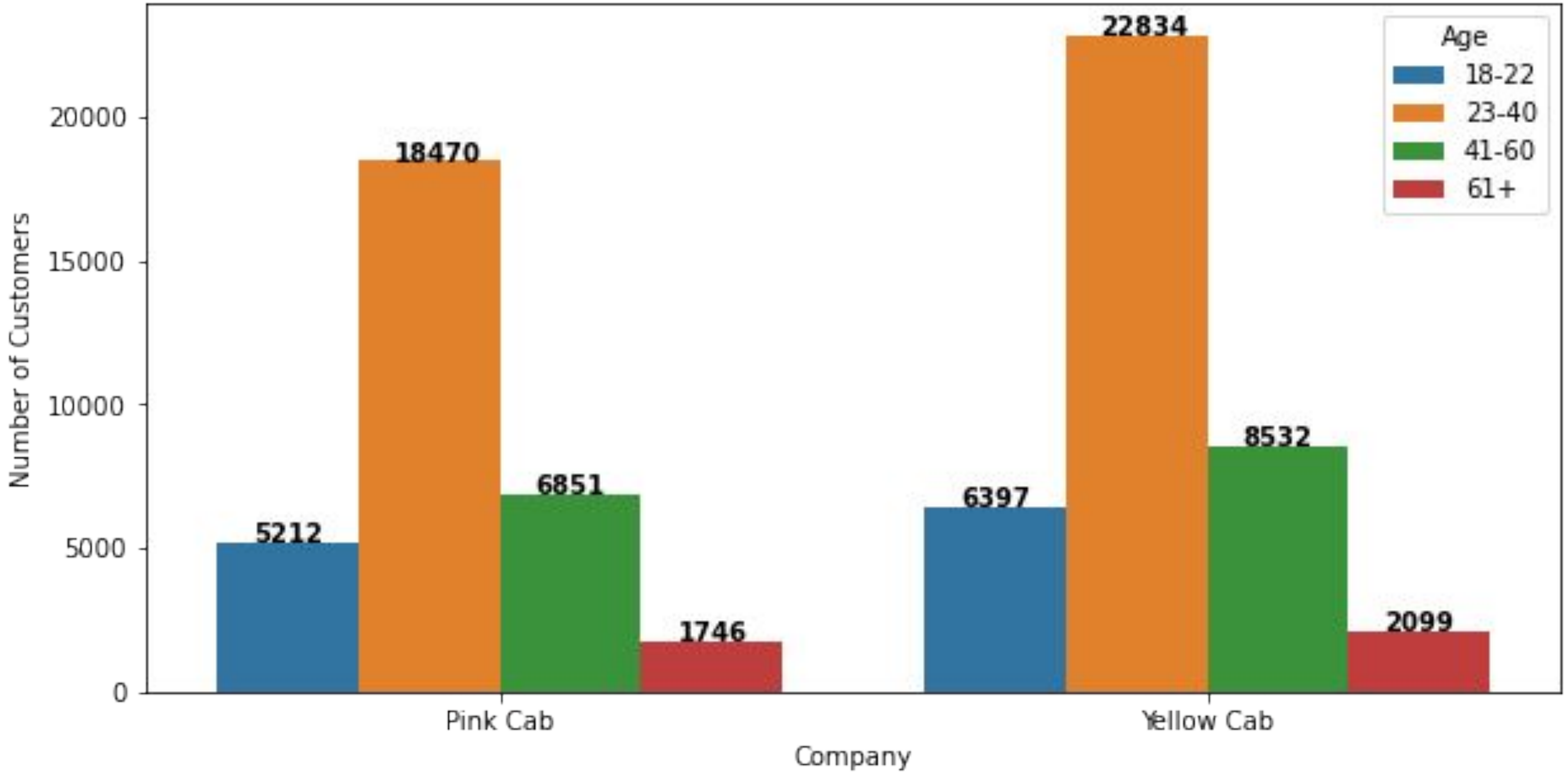
The number of male customers is greater than the number of female clients, but this occurs in both companies

Contextual Analysis



Upper-class clients are the ones that consume the most of the service and both companies have a greater number of clients in this sector.

Contextual Analysis



The group of clients from 23 to 40 years old is the one that most consumes the service and both companies have their largest number of clients in that group.

EDA Summary and Recommendations

- **Presence in big cities** : Yellow Cab has a greater market share than Pink Cab in the largest cities in the United States(New York,Chicago,Los Angeles,etc).Therefore, it is already well positioned in these cities and has a large number of potential clients since these cities have a large population.
- **Demand throughout the year** :Yellow Cab has a greater demand for its service than Pink Cab at each time of the year and in each year(2016,2017,2018).
- **Number of customers**: Yellow Cab has a greater number of clients in general and year after year that has not changed much.
- **Profit according to customer income**: Both companies cover well the segment of clients with high income, which are the ones that generate the greatest profit, even so, Yellow Cab has more clients in that segment.
- **Profit according to the age of the client**: The segment of customers aged 23 to 40 years are the ones that generate the most profits, and specifically in this segment is where there is the most difference in favor of company Yellow Cab.

Based on all the findings mentioned above, the recommendation is to invest in the Yellow Cab company.

Thank You

Data Glacier

Your Deep Learning Partner