Data Glacier

Your Deep Learning Partner

Project : G2M insight for Cab Investment firm

Data Science Virtual Internship Program

Name: Harold Wilson

Date : 21st March 2023

Agenda

- Executive Summary
- Business Understanding
- Data Understanding
- Data Preparation
- Conclusion &
- Recommendation

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.

This project analysis will be carried out using the Crisp-DM methodology to help in the work flow of this project.

CRISPS-DM Methodology:
The cross-industry standard practice for data mining process has been extensively used to carry out most data analysis or mining project. The CRISP-DM data mining methodology published in 2000 which outline the steps the needs to be adhered to when performing a data mining process to achieve the outmost results.

- Business understanding
- Data understanding
- Data Preparation
- Data modelling
- Evaluation
- Deployment

This presentation will look at the business understanding, data understanding and preparation and finally draw conclusion and recommendation from the results.

Business Understanding

This outlines understanding the business problem, investigating the business objectives and requirements of the business plan. Also looks at understanding which data needs to be collected to meets these business objectives.

According to ibis world statistics, the taxi market was valued at USD 213.14 Billion in 2021 and it is projected to reach USD 356.47 Billion by 2027, registering a CAGR of 8.95% during the forecast period 2022 - 2027

The main business objective of this project is to help the client XYZ to make a decision based on the quality of the analysis and the value of recommendations and insights on which cab firm to invest in. This is based on the provision of four data sets over time period of about three years ranging from 31/01/2016 to 31/12/2018 that will be used in the analysis.

Data Understanding

This is the most complex and tedious stage of the process where the data collected needs to be prepared in line with the business plan.

In this stage the main objective includes cleaning the data, dealing with missing and unknown values, reducing data dimensionality, transforming data values, and sometimes reformatting the data to suite the desired mining solution. Other operations performed under this stage includes data aggregation, normalisation, and attribute creation i.e., making new variables to tackle specific business queries.

Data Understanding- Background

Cab_Data.csv

Cab_Data.csv – This file includes details of transaction for 2 cab companies. There are 7 features in this data set which has two data types as objects and five data types as numeric with 359392 observations

City.csv

City.csv – This file contains list of US cities, their population and number of cab users. There are 3 features in this data set which all have data types as objects with 20 observations

Customer_ID.csv

Customer_ID.csv – This is a mapping table that contains a unique identifier which links the customer's demographic details. There are 4 features in this data set which has one data types as objects and three data types as numeric with 49171 observations

Transaction_ID.csv

Transaction_ID.csv – This is a mapping table that contains transaction to customer mapping and payment mode. There are 3 features in this data set which has two data types as numeric and one data types as object with 440098 observations

Data Preparation-Transformation

Cab_Data.csv

The dates in the original cab_Data set was represented in excel serial date format, this was changed into datetime format to aid in the analysis.
The column Date of Travel was changed and replaced by Date for simplicity
Also the date was converted and split into year and month for data aggregation and analysis.

City.csv

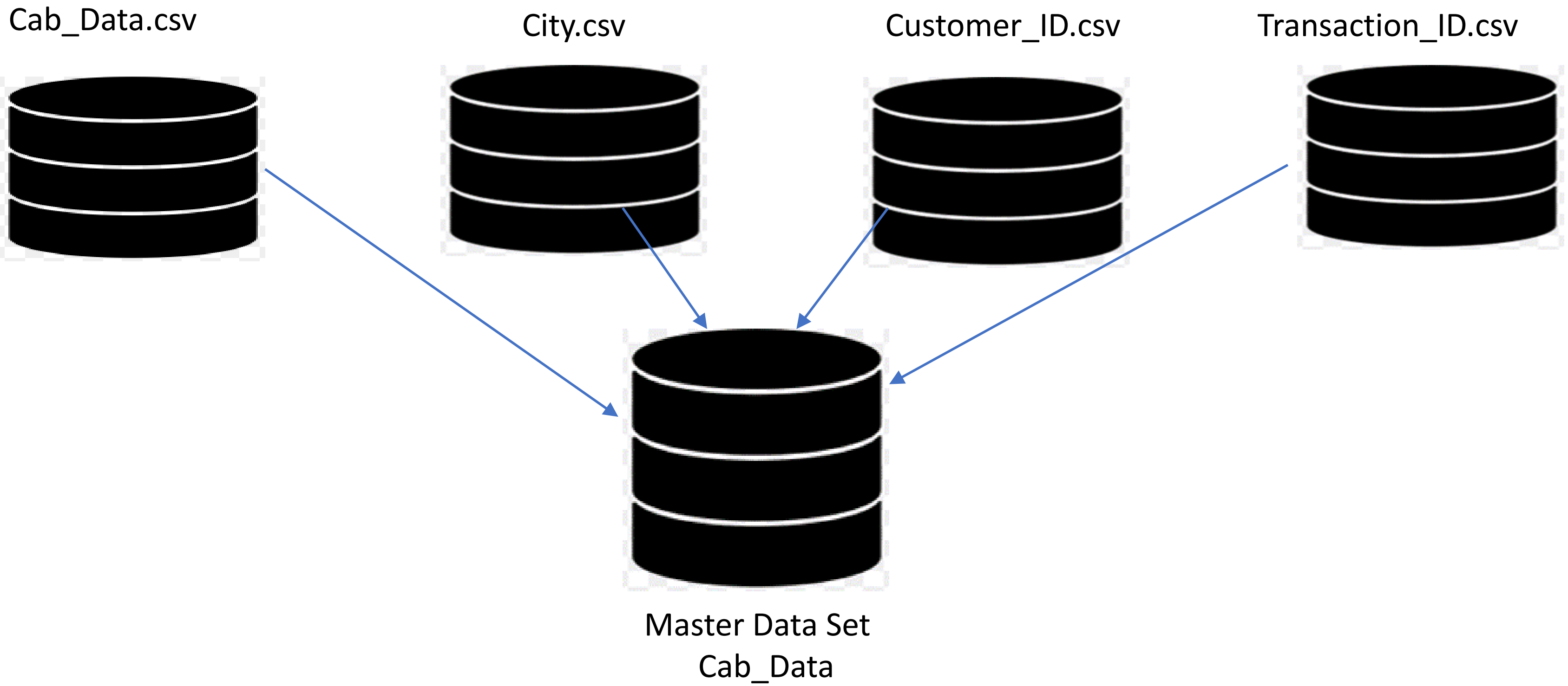
The columns in the City data set i.e. population and users had commas and spaces, theses were removed to present them as just numbers for the sake data manipulation and transformation.
The data types of column population and users was converted to numeric to help in analysis.

Customer_ID.csv

The column Income(USD/Month) was changed and replaced by Income to simplicity.

In general there were no missing and duplicate data recorded in all these data sets

Data Preparation- Merging



Cab_Data.csv

City.csv

Customer_ID.csv

Transaction_ID.csv

Master Data Set
Cab_Data

Merging these data sets into one master data set for the analysis

Cab_Data and City Data were merged by a common column called City

Customer_ID and Transaction_ID were merged by a common column called Customer_ID

Then finally these two merges were merged again by common column called Transaction_ID

Data Preparation- Feature Engineering

Feature Engineering : Is the process of transforming raw data into some interesting features to be used as KPI (key performance indicators) for analysis depending on the business plan or objective.

- Profit_margin
- Profit_Rate
- User_Ratio
- Profit_per_KM
- Market size

Feature variables and how they were derived.

Profit_margin = Price charged – Cost of Trip

Profit_Rate   = ((Price charged – Cost of Trip)/Cost of Trip) * 100

User_Ratio   = Users/Population *100

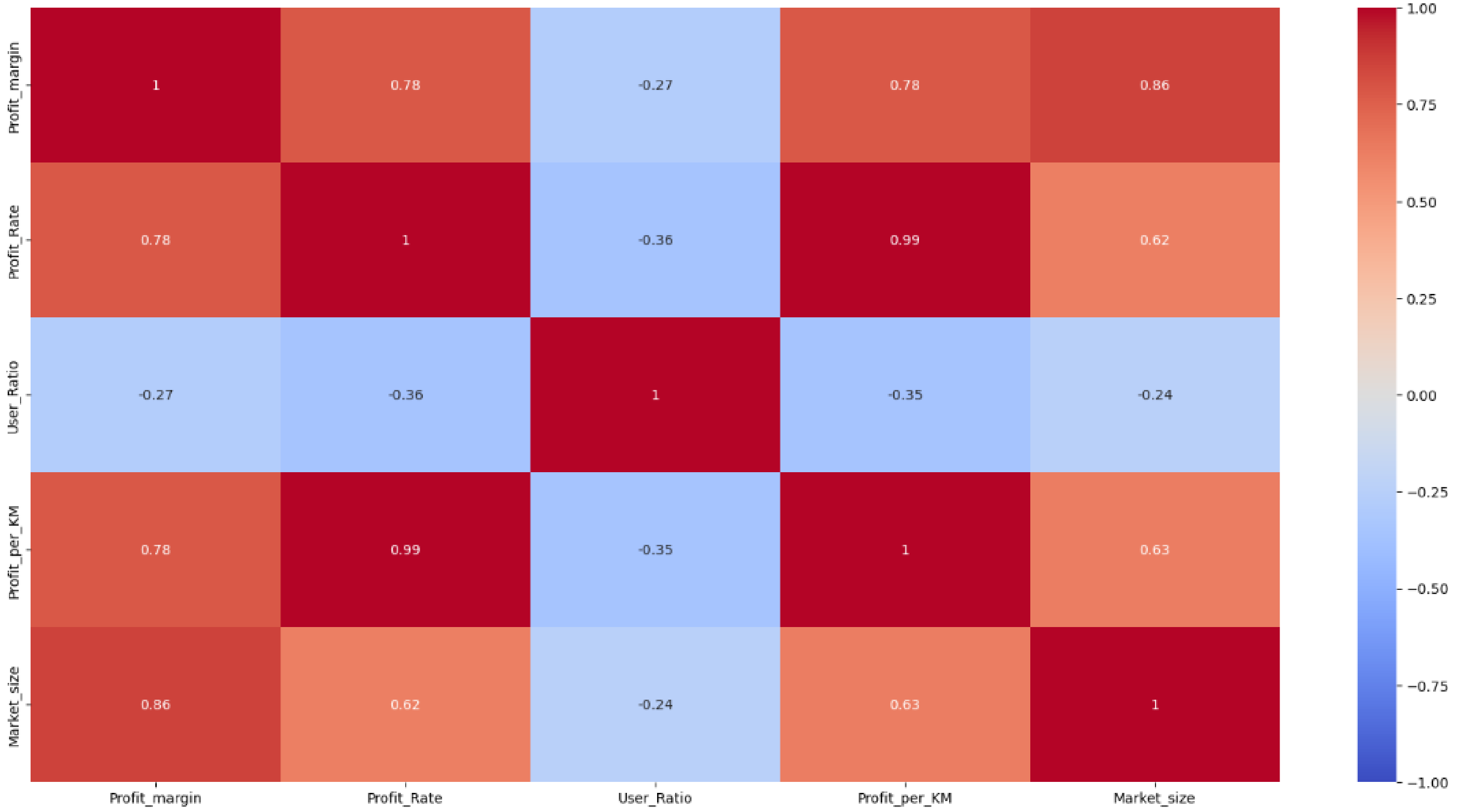
Profit_per_KM = Profit_margin / KM Travelled

Market size   = Price charged * Users

Data Glacier

Your Deep Learning Partner

Data Preparation- Correlationships amongst Features



The above heat map,
shows a 99% strong and positive relationship between

- Profit_margin and Market_size
- Profit_Rate and Profit_per_KM

Also there is a 87% strong and positive relationship between

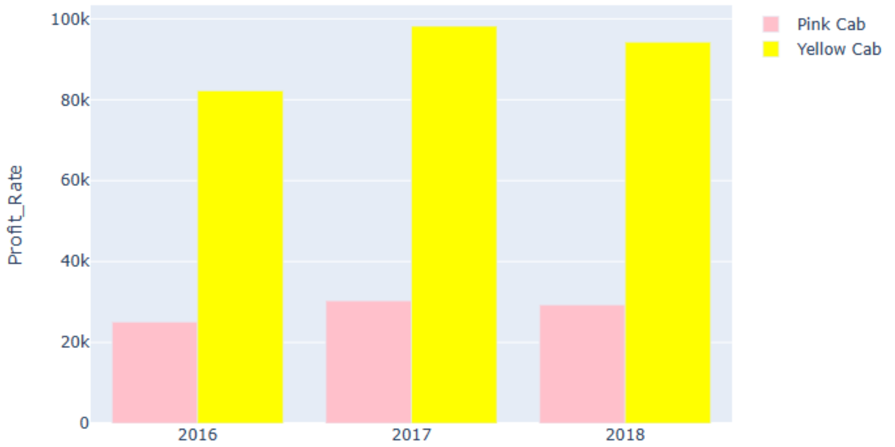
- Profit_margin and Profit_Rate
- Profit_margin and Profit_per_KM

Data Glacier

Your Deep Learning Partner

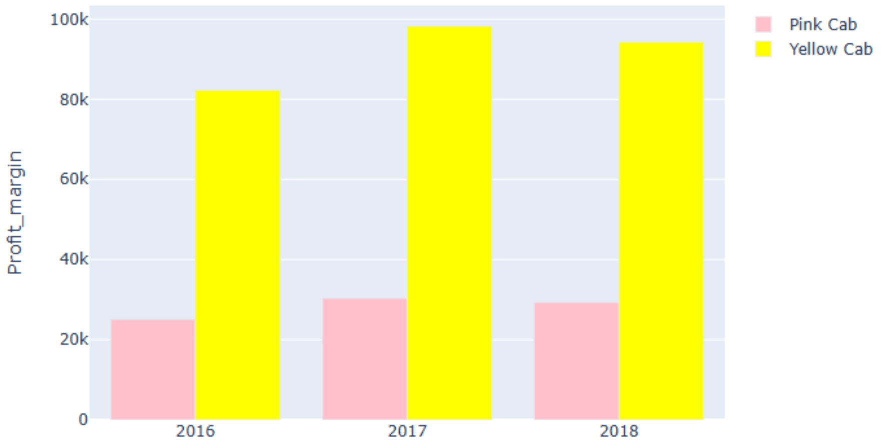
Data Visualisation

Pink & Yellow Cab Firm Profit_Rate Distribution by the Year



The yellow and pink cab firms showed a high profit rate in 2017 with the yellow cab firm showing more profit rate (about four times) than the Pink cab firm over the years.

Pink & Yellow Cab Firm Profit_margin Distribution by the Year



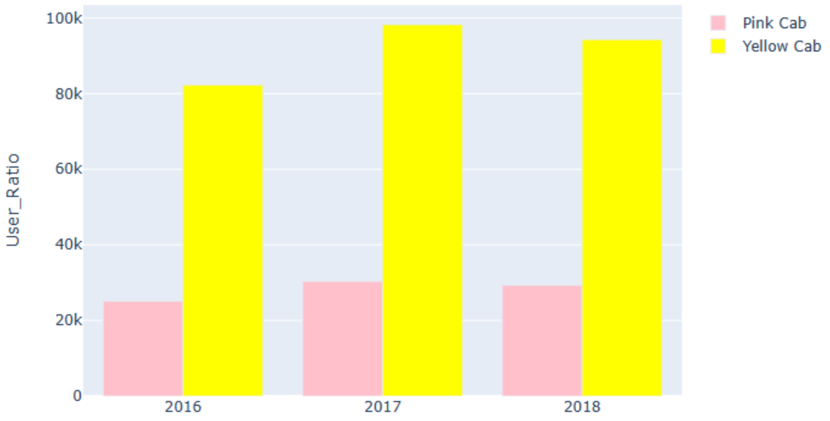
The Yellow cab firm made the most profit margin (about four times) over the years than Pink cab firm Also there was a high profit margin recorded in 2017

Data Glacier

Your Deep Learning Partner

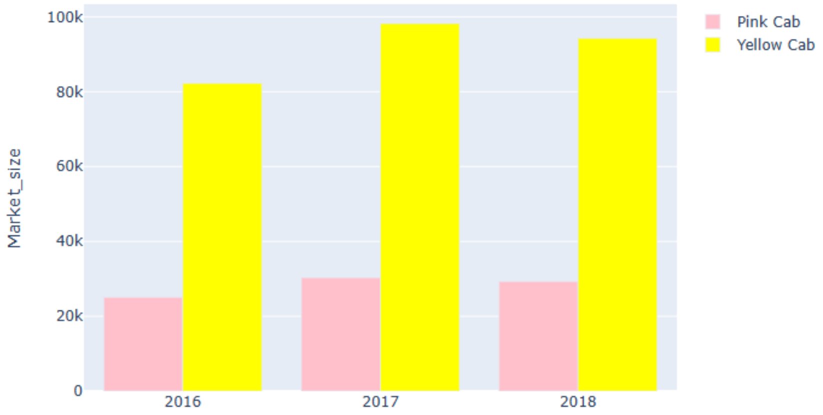
Data Visualisation

Pink & Yellow Cab Firm User_Ratio Distribution by the Year



The user ratio for Yellow cab firm was about four times more than the Pink cab firm. 2017 recorded more yellow cab users than pink cab firm.

Pink & Yellow Cab Firm Market_size by the Year

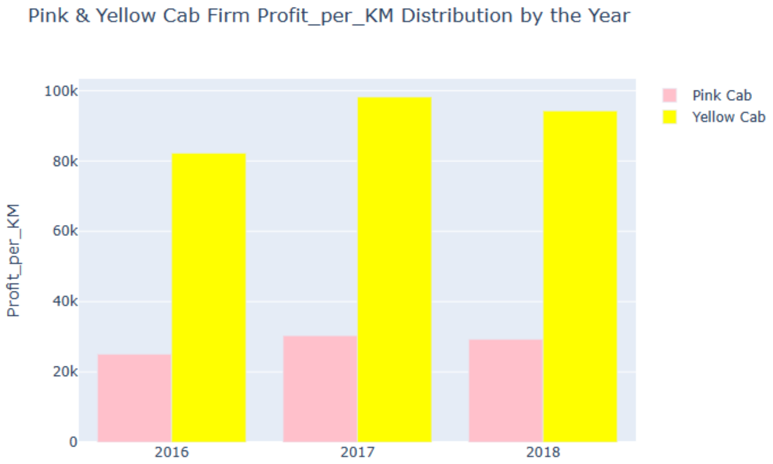


The Yellow cab firm had the bigger market size (about four times) over the years than Pink cab firm. Also there was a high market size recorded in 2017 for both cab firms.

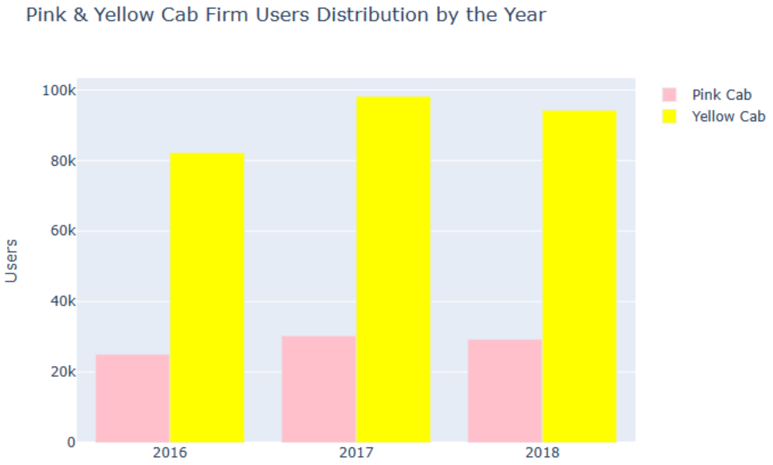
Data Glacier

Your Deep Learning Partner

Data Visualisation



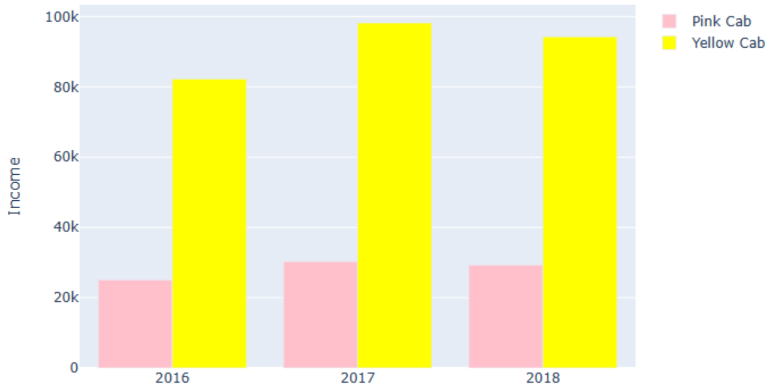
The yellow cab firm had a higher profit per KM (about four times) than the Pink cab firm. Also 2017 recorded the highest profit per KM for both Yellow and Pink cab firms.



There were more Yellow cab users (about four times) over the years than Pink cab users. In 2017 more people used Yellow cab than Pink cab

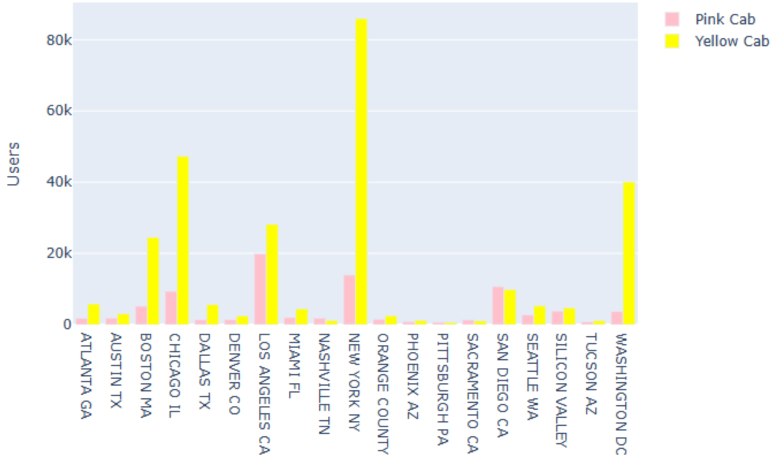
Data Visualisation

Pink & Yellow Cab Firm Income Distribution by the Year



High income earners used the Yellow cab than the Pink cab
2017 registered the highest income records for both Yellow and Pink cab firms.

Pink & Yellow Cab Firm Users Distribution Over City



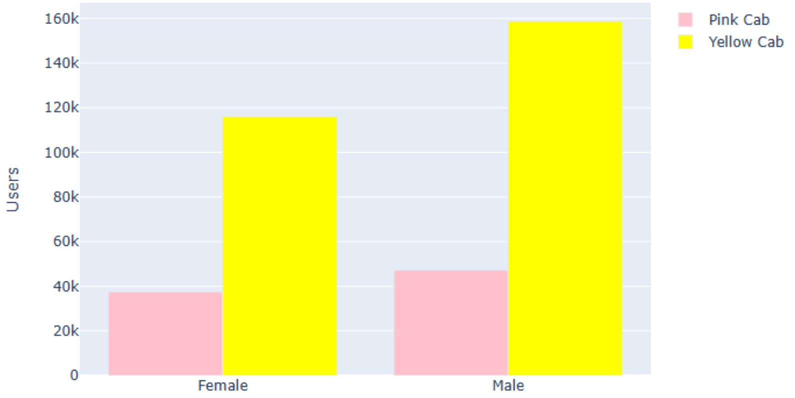
From the Bar Chart, the Yellow Cab firm, records showed the highest number of users on a city basis are in New York, Washington and Chicago, while for the Pink Cab Company, the most are in Los Angeles, New York and San Diego

Data Glacier

Your Deep Learning Partner

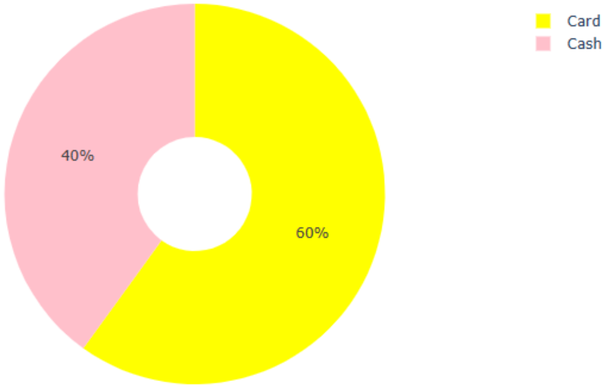
Data Visualisation

Pink & Yellow Cab Firm Gender users Distribution Over City



More males used both the Yellow and the Pink cab firm were than females. However about three times more males used the Yellow cab firm than the females.

Total Users Overview by Payment Method



Yellow cab firm mostly used card payment than cash as method of payment than the Pink cab firm.

Conclusion & Recommendations

In conclusion the analysis showed that the business operations on the Yellow cab firm achieved more than the Pink cab firm in that it recorded appreciably more and high value (about four times) in the following KPIs

- Profit_margin
- Profit_Rate
- User_Ratio
- Profit_per_KM
- Market_size

Hence to advise the XYZ investment about which company to invest in will be the Yellow cab firm since it surpass it's competitors in all grounds, hence we will recommend the Yellow cab firm for investment.

Data Glacier

Your Deep Learning Partner

Thank You