**XYZ Investments in Cab Industry**

**1. Business Problem.**

Data Analysis to create insights informing which cab industry the XYZ firm should invest.

**2. Business Understanding.**

XYZ it is a private firm located is 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.

The object of this project is providing answer of the main questions made by the company’s CEO, which are:

* Which Cab companies XYZ firm should invest, in general?
* What is maximum amount of money made by each company through this period of time?
* Which company has more users?

Also there are secondary questions that could help the CEO decision, which are?

* Which area of US the Cab companies provide higher incomes?
* Which period of time people use taxi more frequently, seasons?
* Which season through the year the profit is higher?
* Which company provides more income in in holidays?

The answer for those questions is presented in two different methods:

* Reports recommending which Cab company The XYZ should invest.
* A website created by the data analyst showing a table of which company has the higher amount of users and profit. Also a collection of hypotheses and a dashboard is presented in order to auxiliary the CEO decisions.

The tools used for this project are: Python 3.8, Pycharm, Jupyter Notebook, Streamlit and Heroku.

**2. Data Understanding.**

There are 6 different datasets provided:

https://www.kaggle.com/nishantdhingra/cabs-fare-data

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.

States – this file contains list of US cities, regions and divisions. – provided by website: https://www.kaggle.com/omer2040/usa-states-to-region

US-federal-holidays-2011-2020 – this file contains list of US holidays from 2011 to 2020. – provided by website: https://data.world/sudipta/us-federal-holidays-2011-2020#

| **Attributes** | **Meaning** |
| --- | --- |
| Transaction ID | Unique ID for each cab service |
| Date of Travel | Date of travel by the user on cab |
| Company | Name of the cab company |
| City | City of the cab travel |
| KM Travelled | KM Travelled made in each travel |
| Price Charged | Price charged plus tax – Total charged |
| Cost of Trip | Price Charged for the travel |
| Customer ID | Customer ID |
| Gender | Customer gender |
|  |
| Age | Customer Age |
| Income (USD/Month) | Customer Income |
| Payment\_Mode | Payment method |
| Population | Each city population |
| Users | Cab users by city |
| State | US State |
| State Code | US state code |
| Region | US region |
| Division | US division |
| Date | Date oh holiday |
| Holiday | Holiday name |

**3. Business Assumptions.**

For further analysis, it is necessary to know the period of time of each season on US, those information will be included in a variable called Season. According to the website: <https://www.timeanddate.com/calendar/aboutseasons.html>, each season is specified on the followed days:

* spring runs from March 1 to May 31;
* summer runs from June 1 to August 31;
* fall (autumn) runs from September 1 to November 30; and
* winter runs from December 1 to February 28 (February 29 in a [leap year](https://www.timeanddate.com/date/leapyear.html)).

Due the dataset City.csv does not inform the State of two cities: SILICON VALLEY and ORANGE COUNTY, it is necessary specify the States of them.

According to the website: <https://www.britannica.com/place/Silicon-Valley-region-California>, The SILICON VALLEY State is California, CA.

Also according to the website: https://en.wikipedia.org/wiki/Orange\_County,\_California, The ORANGE COUNTY State is California, CA.

OBS: The States will be added on excel format.

**4. Solution Strategy.**

The answer for which Cab company XYZ firm should invest can be answered as 2 different ways:

* **First way:** How much money and users each company made through those years?

1º - Group the variable Company by the variable Price Charge, and sum the amount of price for each company. This method shows a general perspective about each company in all cities and years.

2º - For users, it is possible do the same method, however, it is replaced the variable Price Charge by the variable Customer Id.

* **Second way:** Here it is made a verification of the amount of money and users by region, season and holidays.
* 1º - Group the variable Company by the variable Price Charge and Region, and sum the amount of price for each company. This method shows a specific perspective about each company in each region. For users, it is possible do the same method, however, it is replaced the variable Price Charge by the variable Customer Id.
* 2º - Same method is used than first one, however replace the variable Region by Season. This method shows a specific perspective about each company in each season over the years.
* 3º - Same method is used than first one and second one, however replace the variable Region by Holidays. This method shows a specific perspective about each company in the most important days of the year over the years.

The general and specific overview about price and clients will help the decision about which company XYZ should invest.

Furthermore hypotheses will be created to influence that decision as well.

**5. Top 08 Data Insights.**

**Hypothesis 01:** Houses which has water view are 20% more expensive, in general.

**False:** Houses with water view are 212.57668803323867 percent more expensive.

**Hypothesis 02:** Houses that was built before 1955 are 50% cheaper, in general.

**False:** Houses that was built before 1955 are -0.7757205525248732 percent cheaper.

**Hypothesis 03:** Houses without basement are 40% bigger them house with basement, related to total area (sqft\_lot).

**False:** Houses without basement are 22.483151526642544 percent bigger them houses with basement.

**Hypothesis 04:** The growth of house prices YoY (Year over Year) (May 2014 compared to May 2015) is 10%, in general.

**False:** The total houses price YoY (Year over Year) suffered a decrease of -62.79177358882806 percent.

**Hypothesis 05:** Houses with 3 bathrooms have a growth MoM (month over Month) of 15%.

**False:** The total houses price MoM (month over Month) suffered a decrease of -9.953899240174858 percent.

**Hypothesis 06:** Houses with number of bedrooms above 8 have a number of bathrooms 40% higher than houses with number of bedrooms between 5 and 8, and 94% higher than houses with number of bedrooms between 1 and, 4 on average.

**True:** Houses with number of bedrooms above 8 have a number of bathrooms 39.9514563106796 percent higher than houses with number of bedrooms between 5 and 8, and 94.48676155875182 higher than houses with number of bedrooms between 1 and 4.

**Hypothesis 07:** Houses with 7 bedrooms has the total area (sqft\_lot) bigger between 132 to 320 percent than houses with 8 to 11 bedrooms, on average.

**True:** Houses with 7 bedrooms has the total area (sqft\_lot) bigger between 132.29431644290653 and 320.17243208828523 percent than houses with 8 to 11 bedrooms.

**Hypothesis 08:** Renovated Houses have living rooms 12% bigger than houses not renovated, on average.

**True:** Renovated Houses have living rooms 12.132344286788795 percent bigger than houses not renovated, on average.

**6. Financial Results.**

House Rocket Company would have a profit of almost 19 percent, which are more than $771 million, if applies this data analytics method.

**7. Conclusion.**

In conclusion, it is possible to identify that the application of data analytics project at dataset from House Rocket Company was very successful, providing a huge profit opportunity based on which houses to buy and when to sell.

**8. Next Steps.**

Other project that can be made with this dataset is the exploration data analyses, which identify the best’s attributes in order to apply machine learning algorithms, with the objective to predict the price of futures houses to buy.