Data Intake Report - Taxi-Cab-Data-Analysis-and-Insights

Name: <Taxi-Cab-Data-Analysis-and-Insights>
Report date: <02/08/2021>
Internship Batch:<LISUM02>
Version:<1.0>
Data intake by:<Igor Azevedo de Queiroz>
Data intake reviewer:< >
Data storage location: <https://github.com/IgorQueiroz23031988/Taxi-Cab-Data-Analysis-and-Insights>

Tabular data details:

Total number of observations	< 359392 rows >
Total number of files	<Number of files received>
Total number of features	< 24 columns >
Base format of the file	<.ipynb, .csv, .txt, .png>
Size of the data	<3.740 in MB>

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

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?
- Which company provided more taxi travels?
- Which company has more customers?

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

- Which area of US the Cab companies provide higher incomes and taxi travels?
- Which season of US the Cab companies provide higher incomes and taxi travels?

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

- Reports recommending which Cab company The firm XYZ should invest.
- A website created by the data analyst showing a table of which company has the higher amount of taxi travels, profit and customers. 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 7 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#

Us-states.json

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

Attributes	Meaning
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 of 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).

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.

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, taxi travels and customers each company made?

1º - Group the variable Company by the variable Profit, 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 taxi travels, it is possible do the same method, however, it is necessary replace the variable Profit by the variable Transaction Id.

3º - For customers, it is possible do the same method, however, it is necessary eliminate the duplications and replace the variable Profit by the variable Customer Id.

- **Second way:** Here it is made a verification of the amount of money and taxi travels by region and season.

- 1º - Group the variable Company by the variable Profit and Region, and sum the amount of price for each company. This method shows a specific perspective about each company in each region. For taxi travels, it is possible do the same method; however, it is necessary replace the variable Profit by the variable Transaction 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.

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 10 Data Insights.

Hypothesis 01: The number of users of the Yellow Cab plus Pink Cab is the same number of the Total Cab Users at every city.

False: There are others Cab companies at each state, which the range of customers in others companies floats between 70% and 99%, depending of the state.

Hypothesis 02: The number of users of the Yellow Cab is higher than the company Pink Cab in each city.

False: There are cities where the Pink cab company has more customers than the Yellow cab company, cities such: San Diego, Nashville, Sacramento and Pitsburgh.

Hypothesis 03: People uses more taxi during New Year's Day than other holidays in all 3 years.

False: In total, the holiday which people uses more taxis is Veterans Day, However, splitting this in years, 2016 and 2018, Thanksgiving Day is the holiday which people uses more taxis, and in 2017 the holiday is Veterans Day.

Hypothesis 04: The number of users for each company increases WoW (Week over Week).

False: The number of users for the company Yellow Cab Company WoW (Week over Week) suffered a decrease of 2.26 percent.

The number of users for the company Pink Cab Company WoW (Week over Week) suffered a decrease of 2.39 percent.

Hypothesis 05: The amount of money made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

False: The amount of money made by the company Yellow Cab Company YoY (Year over Year) suffered a decrease of 1.48 percent.

And the amount of money made by the company Pink Cab Company YoY (Year over Year) suffered a decrease of 4.58 percent.

However, the amount of money made by the company Yellow Cab Company MoM (Month over Month) suffered an increase of 1.81 percent.

And the amount of money made by the company Pink Cab Company MoM (Month over Month) suffered an increase of 3.94 percent.

Hypothesis 06: The number of travels made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

True: The number of travels made by the company Yellow Cab company YoY (Year over Year) suffered a decrease of 7.06 percent.

And the number of travels made by the company Pink Cab company YoY (Year over Year) suffered a decrease of 8.01 percent.

Also, the number of travels made by the company Yellow Cab Company MoM (Month over Month) suffered an increase of 3.35 percent.

And the number of travels of money made by the company Pink Cab Company MoM (Month over Month) suffered an increase of 4.11 percent.

Hypothesis 07: Companies make more profit during second semester of the year.

True: In total and in the second semester, the company Yellow Cab make $3770264.7852 (7.64%) more profit than the first semester.

And in the second semester, the company Pink Cab make $2134648.873 (4.33%) more profit than the first semester.

Hypothesis 08: People with the amount of income higher than the median uses more taxis.

False: People with the amount of income higher than the median uses less taxi than people which has income inferior than median, however this difference is minimal, which is 1 for Pink Cab company and 3 for Yellow Cab company.

Hypothesis 09: Men use more cash than women for payment.

True: In total, the number of men using cash is 20978 (5.84%) more than women.

Hypothesis 10: Young people use more taxis than old people.

True: In total, there are 166884 (46.44%) young people more using taxi than old people.

6. Financial Results.

Yellow Cab Company has that percentage higher than Pink Cab Company:

- Number of taxi trips (52.86%);
- Profit (78.48%);
- Number of customers (35.73%).

7. Conclusion.

In conclusion, Yellow Cab Company has a higher number of taxi trips than Pink Cab Company, as well as a higher profit and higher number of customers.

The data also show that Yellow Cab has superior results in almost all cities, especially in New York, which is the city with the highest profit and higher number of taxi trips.

The same company also stands out from the other company in relation to the season, where the fall or spring (depending which measure is more important) season shows the highest profit and number of taxi trips as well.

Also other insights provided by the hypotheses, inform that Yellow Cab Company present better results than Pink Cab Company, however some of them shows that the number of taxi trips, the profit and the number of customers, for both Companies, are decreasing over the year, giving space for other Cab companies.

Due those all information, it is recommended to invest in Yellow Cab Company.

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 number of taxi trips, the profit or the number of customers for each company.