Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: July 14th, 2022
Internship Batch: LISUM11: 30
Version:<1.0>
Data intake by: Yining Liu
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/DataGlacier/DataSets>

Tabular data details:

Customer_ID

Total number of observations	16384
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

City

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 B

Cab_Data

Total number of observations	16384
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

Transaction_ID

Total number of observations	16384
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

In this assignment, I used a python library called “Seaborn” to plot the graphs to better represents the tendencies and features in this data set. From many different kinds of plots like histogram and boxplot, we found how the variables change relying or not relying on other variables.

- Mention your assumptions (if you assume any other thing for data quality analysis)

Assumption 1: Will the consumers choose different cab companies on different dates of travel?

From the plot above we can see that the median, upper/lower interquartile are pretty similar. And the boundary of maximum and minimum value are at the same level. So there is no relation between the date of travel and the cab companies chosen, which means that consumers take taxi randomly without much special preference on the companies.

Assumption 2: Whether the PRICE people are willing to pay for a cab ride is related to the GENDER and AGE of the passenger.

From the plot above we can observe that generally male are more likely to pay a higher price on each taxi ride no matter the age group. And there is also a slight trend that elder people would not like to spend more on a cab ride than the younger people.

Assumption 3: The unit price per kilometer for a cab drive is dependent on then economic development. The higher the development level, the higher the unit price would be.

From the boxplot above, it is obvious that in all cities the price per kilometer of Yellow Cab is much higher than that in Pink Cab. In addition, unit price of cab rides in New York city is much higher than the prices in other cities, which makes sense since it is one of the most developed city in this world so there's a higher traffic demand and the cost of everything is New York is just higher than other cities.
And we can also see that large cities like dallas silicon valley have higher unit price for the same reason above. So our assumption is right. Higher development level means that it's more expensive for a taxi ride.

Assumption 4: The number of customers in different cities of the two companies in relation to the local economic situation

From the histogram above we can easily point out that in most of the cities, the costumer volume of Yellow Cab is much greater than that of the Pink Cab.
But there are also 2 exceptions: Nashiville, Sacramento. In San Diego, 2 companies have the similar customer volume.

Assumption 5: In different states, is there a difference in price acceptance between the different genders of passengers ?

From the plot above we can see that in different states, there shows no obvious tendencies that passenger in different genders have different acceptability on the price of each taxi ride. Gender is not a factor deciding the price a passenger would pay for a taxi drive.

Assumption 6: The 2 cab companies have different profitability rates on different payment method.

From the boxplot above we can observe that generally payment method does NOT affect the profit rate of each company.
But generally speaking Yellow Cab makes more money than Pink Cab, whcih mean that the Yellow Cab has a higher profit rate.

Assumption 7: Profit rate of each company does NOT varies state-wisely.

Obvious that the patterns have no difference. And in each state, profit rate of both Yellow Cab and Pink Cab are similar to each other.