Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 2021-03-07
Internship Batch: LISPO1
Version: 1.0
Data intake by: Freddy F. Tapia C.
Data intake reviewer: -
Data storage location: https://github.com/Fr3ddy1/Week2_EDA

Tabular data details:

1) **Cab_mod.csv:**

Total number of observations	359,392
Total number of features	7
Base format of the file	csv
Size of the data	24.5 MB

2) **City.csv:**

Total number of observations	20
Total number of features	3
Base format of the file	csv
Size of the data	1 KB

3) **Customer_ID.csv:**

Total number of observations	49,171
Total number of features	4
Base format of the file	csv
Size of the data	1.02 MB

4) **Transaction_ID.csv:**

Total number of observations	440,098
Total number of features	3
Base format of the file	csv
Size of the data	8.7 MB

5) **Final_data.csv**

Total number of observations	359,392
Total number of files	4 (obtained by merging inicial files Cab, Customer, transaction)
Total number of features	19
Base format of the file	csv
Size of the data	64.98 MB

6) **plot_data.RDS:**

Total number of observations	workspace with different variables
Total number of features	workspace with different variables
Base format of the file	RDS
Size of the data	42.2 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

My approach in the process of finding the duplicates was to analyze the different levels of each variable that can be used to merge or combine the four data sets provided,

1) Cab_mod.csv: for this data i analyzed the "Transaction.ID" variable and i found that there are no duplicates in this data. There are 359,392 unique transactions.
2) City.csv: for this data i analyzed the "City" variable and i found that there are no duplicates in this data. There are 19 unique cities.
3) Customer_ID.csv: for this data i analyzed the "Customer.ID" variable and i found that there are no duplicates in this data. There are 49,171 unique customers.
4) Transaction_ID.csv: for this data i analyzed two variables "Transaction.ID" and "Customer.ID" and i found that for the first variable there are no duplicates in this data. There are 440,098 unique transactions. On the other hand there are duplicates for the second variable "Customer.ID", there are only 49,171 uniques values.

I used the variable "Transaction.ID" as a key variable to join or merge the data 3 and 4. After that I merge the new data that i obtained with the data 1, and finally obtained the definitive data.

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

During the analysis carried out to detect the outliers, note that the variable "Price.Charged" presents these values, due to the lack of information regarding the distance traveled ("KM.Travelled"), this variable will not be treated.

Another important issue to consider is the "profit" variable, which is calculated from the subtraction of the price and cost variables. I assume for this analysis that there may be a negative profit for some rides.