Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 08/31/23
Internship Batch: LISUM25
Version: 1.0
Data intake by: Faris Chaudhry
Data intake reviewer: N/A
Data storage location: https://github.com/farischaudhry/DataGalcier/tree/main/Week%202

Tabular data details:

City.csv

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

Cab.csv

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

Customer.csv

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Transaction.csv

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Proposed Approach:

Dedup validation done by using df.drop_duplicates(inplace = True). Furthermore, by explicitly dropping any duplicates of primary keys such as TransationID and CustomerID. This should be sufficient given the data sets are quite small. Null checks are done by using df.info().

Assumptions:

- We assume that only the cities listed in the transactions are where the businesses operate (more cities would change the number of profit/trips and therefore the whole strategy)
- We assume that the number of users listed in ‘City.csv’ includes other cab services apart from Pink and Yellow.
- We assume that profit is the price charged minus the cost of the trip (without calculating any extra overheads such as fuel cost, wages, depreciation of the car); these factors might be accounted for within the cost of the trip.
- We assume the data is otherwise entered correctly.
- We cannot remove outliers from prices charged because there is no data on the duration of the trip, i.e., the price is totally dependent on the distance travelled where in reality there is often a flat fee for time spent waiting.