

# Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 06/27/2024

Internship Batch: LISUM33

Version: 1.0

Data intake by: Balamurugan

Purushothaman

Data intake reviewer:

Data storage location:

## Tabular data details:

### Cab\_Data.csv

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	20663 KB

### City.csv

Total number of observations	20
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	4 KB

### Customer\_ID.csv

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1027 KB

### Transaction\_ID.csv

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8788 KB

**Proposed Approach:**

- Mention approach of dedup validation (identification):  
Merged all the datasets with the help of pd.merge.  
I checked for duplicate in the merged dataset (non found)
- Mention your assumptions (if you assume any other thing for data quality analysis):  
In Cab\_Data.csv, for each transaction the profit for that transaction is (Price Charged – Cost of the Trip).  
Date of travel is number of days starting from January 1, 1900.  
In the City.csv, the number of Users is the total number of cab users in the city including Pink cab and Yellow cab as well as other cab companies.  
A ride can have multiple transactions. The cost is adjusted to only count the cost once for over lapping routes for multiple people. The charge varies.