

Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 17/06/2022

Internship Batch: LISUM10 : 30

Version: 1.0

Data intake by: Amr El-Bana

Data intake reviewer: Data-Glacier

Data storage location: <https://github.com/DataGlacier/DataSets>

Cab_Data.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	20.2 MB

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	759 bytes

Customer_ID.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_ID.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:

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

I assume the data is messy.