

Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 9/13/2023

Internship Batch: LISUM25: 30 Aug - 30 Nov 2023

Version: 1.0

Data intake by:

Data intake reviewer:

Data storage location:

Tabular data details:

Cab_Data.csv:

| | |
|-------------------------------------|--------|
| Total number of observations | 359393 |
| Total number of files | |
| Total number of features | 7 |
| Base format of the file | .csv |
| Size of the data | 20.1MB |

City.csv:

| | |
|-------------------------------------|------|
| Total number of observations | 21 |
| Total number of files | |
| Total number of features | 3 |
| Base format of the file | .csv |
| Size of the data | 4KB |

Customer_ID.csv:

| | |
|-------------------------------------|-------|
| Total number of observations | 49172 |
| Total number of files | |
| Total number of features | 4 |
| Base format of the file | .csv |
| Size of the data | 1MB |

Transaction_ID.csv:

| | |
|-------------------------------------|--------|
| Total number of observations | 440099 |
| Total number of files | |
| Total number of features | 3 |
| Base format of the file | .csv |

| | |
|------------------|--------|
| Size of the data | 8.58MB |
|------------------|--------|

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)
 - Mention your assumptions (if you assume any other thing for data quality analysis)
-
- Assuming users with income < 3500 as 'low income class', 3500 <= income <= 15000 as 'middle income class', and > 15000 as 'high income class'
 - Assuming 'Price Charged' - 'Cost of Trip' = 'Tips', When 'Tips' is under 0, assuming the user used a discount.