Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 29/08/2021
Internship Batch: LISUM03
Version:<1.0>
Data intake by: KABORE Rimbesougri
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/Rim-B/Cab_Investment

Tabular data details:

File name	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	20.1 MB

File name	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	1 KB

File name	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

File name	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

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

Proposed Approach:

- Mention approach of dedup validation (identification) :
  To detect duplicated datas, we can use this python function :
  datasets[x][datasets[x].duplicated()].count()

- Mention your assumptions (if you assume any other thing for data quality analysis) :
  Some datas need to be reformatted - especially the ‘Date of Travel’(Cab_Data.csv) and the ‘Population’ and ’Users’ (City.csv)