Data Intake Report

Name: G2M insight for Cab Investment
Report date: 25/02/2021
Internship Batch: LISP01
Version: 1.0
Data intake by: Camilo Arrieta
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Cab_Data

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	<52.82 MB>

Transaction_ID

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	<21.56 MB>

Customer_ID

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	<2.85 MB>

City

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	<2.32 KB>

Proposed Approach:

- Approach of dedup validation (identification): to explore existing duplicates I will first merge all data (Inner Join) in a pandas DataFrame, and then use the pandas DF method duplicated() along with the sum() function to check the total amount of duplicated rows.
- Assumptions (if you assume any other thing for data quality analysis): to begin, I will assume there is a chance that some of the columns have Null and/or extreme values which will have to be corrected. Also, I assume the data was correctly collected, which allows to make accurate predictions and deductions.