Data Intake Report

Name: G2M insight for Cab Investment firm (Must for all Specialization)
Report date: 06/14/2022
Internship Batch: LISUM10: 30
Version:<1.0>
Data intake by: Ruizhe Zhang
Data intake reviewer: Ruizhe Zhang
Data storage location: https://github.com/richardzzhang/Cab

Tabular data details:

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

Name of file	City
Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	1KB

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

Name of file	Holiday
Total number of observations	1095
Total number of features	2
Base format of the file	.csv
Size of the data	15KB

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

Proposed Approach:

- Approach of dedup validation:
  Functions in Pandas such as merge(), groupby(), drop_duplicates(),etc.

- Assumptions (if you assume any other thing for data quality analysis)
  Rows with NaN values are not useful and should be deleted.