Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 16.10.2021
Internship Batch: LISUM04
Version: 1
Data intake by: A.M. Nimasha Chathurangani Attanayake
Data intake reviewer: J.M. Tharindu Chinthaka Jayaweera
Data storage location: https://github.com/DataGlacier/DataSets
https://www.kaggle.com/donnetew/us-holiday-dates-2004-2021

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	20.1 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	1 KB

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	1 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	8.58 MB

holiday_data

Total number of observations	342
Total number of files	1
Total number of features	6
Base format of the file	csv
Size of the data	15.3 KB

Proposed Approach:

- Mention approach of dedup validation (identification)

  Since the transaction IDs are unique to each and every trip, duplications can be checked based on transaction ID.

- Mention your assumptions (if you assume any other thing for data quality analysis)

  According to the plotted graphs, there are outliers in the Price Charged feature. But since we do not have enough information on the components that made the Price Charged, it is not appropriate to treat it as an outlier.