Data Intake Report

Name: Week 2
Report date: 11/19/2022
Internship Batch: LISUM15
Version:<1.0>
Data intake by: John Wu
Data intake reviewer:<intern who reviewed the report>
Data storage location:
https://github.com/DataGlacier/DataSets
https://www.kaggle.com/datasets/donnetew/us-holiday-dates-2004-2021

Tabular data details:

Total (Cab_Data, City, Customer_ID, Transaction_ID):

Total number of observations	848681
Total number of files	4
Total number of features	14
Base format of the file	.csv
Size of the data	30 MB

US Holiday Dates

Total number of observations	57
Total number of files	1
Total number of features	6
Base format of the file	.csv
Size of the data	3 KB

Proposed Approach:

- I checked for duplicate transaction IDs and null values.
- I assumed that the data for 'City.csv' is at the end of the 3-year period and that users in 'City.csv' are the total users (who may have inactive in the 3-year period), as the unique number of users from 'Cab_Data.csv' do not match city data.
- Joining was done in the .ipynb file with 'Cab_Data' joined with "City" and"Transaction_ID" with "Transaction_ID" joined with "Customer_ID".