Data Intake Report

Name: Week 02 Exploratory Data Analysis
Report date: November 14, 2022
Internship Batch: LISUM15
Version: 1.0
Data intake by: Richard Flores
Data intake reviewer: N/A
Data storage location:
https://github.com/DataDaimon/DataGlacier/tree/main/Week%2002/Datasets

Tabular data details: Cab_Data

Total number of observations	359,393
Total number of files	01
Total number of features	07
Base format of the file	CSV
Size of the data	20.1 MB

Tabular data details: City

Total number of observations	21
Total number of files	01
Total number of features	03
Base format of the file	CSV
Size of the data	4.0 KB

Tabular data details: Customer_ID

Total number of observations	49,172
Total number of files	01
Total number of features	04
Base format of the file	CSV
Size of the data	1.0 MB

Tabular data details: Transaction_ID

Total number of observations	440,099
Total number of files	01
Total number of features	03
Base format of the file	CSV
Size of the data	8.58 MB

Tabular data details: weather_data_nyc_centralpark_2016

Total number of observations	367
Total number of files	01
Total number of features	07
Base format of the file	CSV
Size of the data	12.0 KB

Proposed Approach:

- Data deduplication validation was accomplished through data cleaning in Tableau Prep software and verified in the Jupyter Python Exploratory Data Analysis Notebook
- We assume that we will join CSV tables for Cab_Data, Customer_ID, and Transaction_ID through primary keys and append the City and weather_data_nyc CSV files.