Data Intake Report

Name: Tehesuma Mensah Imoro
Report date: 14th February 2024
Internship Batch: LISUM30 (30th Jan 24 – 30th April 24)
Version:1.0
Data intake by: Tehesuma Mensah Imoro
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

1. **Cab_Data.csv:**

Total number of observations	358,547
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20.1+ MB

Approach:

- Dedup validation: After checking for duplicates in the dataset using the nunique() function, there were no duplicates found.
- Here are a few things that were observed about the data quality of the dataset:
  a. Completeness - This dataset has some missing values.
  b. Timeliness - This dataset is not up to date. The maximum year is 2018.
  c. Validity - The date of travel column of this dataset is not in the right data format.
  d. Uniqueness – There were no duplicates found.
  e. Accuracy – The City column of this dataset has an inaccurate entry.

2. **City.csv:**

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	4 + KB

Approach:

- Dedup Validation: After checking for duplicates in the dataset using the nunique() function, there were no duplicates found
- Here are a few things that were observed about the data quality of the dataset:
  a. Completeness – This dataset has no missing values
  b. Uniqueness – There were no duplicates found.
  c. Validity - The 'Users' and 'Population' columns of this dataset are in the wrong format.

3. **Transaction_ID.csv:**

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8.58+ MB

Approach:

- Dedup validation: After checking for duplicates in the dataset using the nunique() function, there were no duplicates found
- Here are a few things that were observed about the data quality of the dataset:
  a. Completeness - This dataset has no missing values.
  b. Uniqueness – There were no duplicates found.

4. **Customer_ID.csv:**

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1+ MB

Approach:

- Dedup Validation: After checking for duplicates in the dataset using the nunique() function, there were no duplicates found
- Here are a few things that were observed about the data quality of the dataset:
  a. Completeness - This dataset has no missing values.
  b. Uniqueness – There were no duplicates found.