Data Intake Report

Name:
Report date: 14/7/2023
Internship Batch: LISUM23
Version:1.0
Data intake by: Shiby Kurian
Data intake reviewer:
Data storage location:

Tabular data details: Cab Dataset

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	

Tabular data details: Customer Dataset

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	

Tabular data details: City Dataset

Total number of observations	20
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	

Tabular data details: Transaction Dataset

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	

Proposed Approach:

Approach for Deduplication Validation (Identification):

- Identify the fields or combination of fields that should be unique within the dataset.
- Sort the dataset based on the identified unique field(s) in ascending or descending order.
- Iterate through the dataset and compare consecutive records to identify any duplicates.

Assumptions for Data Quality Analysis:

- The provided datasets are relatively clean and have undergone basic data cleaning processes, such as removing obvious errors or inconsistencies.
- The datasets do not have any external data dependencies or references that could affect data quality analysis.
- The data within each dataset is consistent, meaning the field formats and data types are accurate and aligned.
- The data does not contain any hidden or subtle duplicates that may require more complex deduplication techniques.
- The datasets do not contain missing values that would require specific handling techniques such as imputation or deletion. However, if missing values are present, appropriate handling methods should be applied during the analysis.