Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 04.07.2022
Internship Batch:LISUM 10:30
Version: 1.0
Data intake by:Prajesh Tejani
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://www.kaggle.com/datasets/gopalkalpnde/nyc-tlc-data

Tabular data details:

Total number of observations	7821717
Total number of files	1
Total number of features	19
Base format of the file	yellow_tripdata_2015-01.csv
Size of the data	2.0 GB

Proposed Approach:

- Firstly I Read data using different method such as dask,modin and pandas
- Validate column(Remove White_space and convert name in small Letters)
- Validation with YAML file
- convert data into .txt.gz format
- Location:https://drive.google.com/file/d/1bgywre2DHTTrVyPDhN28oSIVKDOqcKH5/view?usp=sharing