# Data Intake Report

Name: File ingestion and schema validation
Report date: 10.01.2024
Internship Batch: LISUM28
Version: 1.0
Data intake by: Parthkumar Sutariya
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