Data Intake Report

Name: File ingestion and schema validation
Report date: 11/07/2022
Internship Batch: LISUM10:30
Version: 1.0
Data intake by: Jigar Pravinbhai Borad
Data intake reviewer:
Data storage location:
https://drive.google.com/file/d/1D0eAFUXs-OQJWDrfeNWmiNx5UUagMEZH/view?usp=sharing

Tabular data details:

Total number of observations	20692840
Total number of files	
Total number of features	9
Base format of the file	.csv
Size of the data	2.3Gb

Proposed Approach:

- I have read this csv data with different methods such as pandas, modin, rays and dask. Here I found that with dask it is fastest and rays is slowest. Modin is an extended version of pandas so it is faster than pandas but slower than dask.
- Then I created a test file to validate the dataset columns name, but they were already correct so i do not need to correct it.
- Then I ingested data , where I created a yaml file with some column names from the dataset. Then I validate those columns with dataset columns and write the difference between the yaml file and dataset.
- Then I converted the dataset into a pipe separated text file with gz format.