Data Intake Report

Project name: **Taxi_prediction_newyork**

Report date: 12th July 2022

Internship Batch: LISUM10

Version:1.0
Data intake by: Mohini Kalbandhe

Data intake reviewer:

Data storage location: https://github.com/amohini099/File-ingestion-and-schema-validation

Tabular data details:

Name	yellow_tripdata_2015
Total number of observations	1985964692
Total number of files	1
Total number of features	19
Base format of the file	.csv
Size of the data	2 GB

Proposed Approach:

1. Given a region and a particular time interval, predict the no of pickups as accurately as possible in that region and nearby regions.Based on the data, machine learning model predicts the pickup demand of cabs in 10 minutes time frame. In this python notebook different machine learning model have been trained and accuracy is tested.
2. Given a location and current time of a taxi driver, as a taxi driver, he/she excepts to get the predicted pickups in his/her region and the adjoining regions in few seconds. Hence, there is a medium latency requirement.
3. As long as taxi driver gets good prediction result, he/she is not be much interested in the interpretability of the result. He/she is not much interested in why he/she is getting this result. Hence, there is a no interpretability required.
4. It is inferred from the source https://www.flickr.com/places/info/2459115 that New York is bounded by the location coordinates (lat,long) - (40.5774, -74.15) & (40.9176,-73.7004). So, any coordinates not within these coordinates are not considered by us as we are only concerned with pickups which originate within New York.

5. According to NYC Taxi & Limousine Commision Regulations the maximum allowed trip duration in a 24 hour interval is 12 hours. So we remove the points where the duration is >12 hr
6. We found that the 99.9th percentile value of speed is 45.31 mph. So, we consider only the data points where the computed speed is <45.31mph. We also observed that the avg speed in New York is 12.45miles/hr, i.e. a cab driver can travel 2 miles per 10min on avg
7. The 99.9th percentile value of the distance covered in a ride is ~23 miles. So we remove rows with large trip distances
8. From percentile and graphical analysis, we set the upper limit of total fare to be 1000$ and consider only the data points which satisfies the limit
9. Creadted " testutility.py" and "store.yaml" file and read file with dask and pandas framework, however dask took less time to read file. Validated number of columns and column name of ingested file with YAML, Wrote the file in pipe separated text file (|) in gz format.