Assignment -1
Openrefine
Dataset used - Taxi Trajectory data from Kaggle.

- Step 1
  - The dataset downloaded from Kaggle is around 1.8GB, so divided the dataset is split into files of each 190MB and loaded into open refine
- Step 2
  - As a first step after loading the dataset verified the unique values in the columns “DATA_TYPE”, “ORIGIN_STAND”, “MISSING_DATA”, and “CALL_TYPE”.
- Step 3
  - The column “DATA_TYPE” has only one unique value for the whole dataset.
  - So, this column doesn’t help much in predicting the trip time.
  - The column “DATA_TYPE” can be dropped from dataset.
- Step 4
  - The column “ORIGIN_STAND” is a unique identifier, identifies based on the column “CALL_TYPE”.
  - The column “ORIGIN_STAND” has blank values and the column “CALL_TYPE” is connected.
  - So, dropping the column “ORIGIN_STAND” is dropping from the dataset.
- Step 5
  - The column “MISSING DATA” will be false if there is no missing data and true if there is any.
  - It has few columns with the value “true”, indicating there is data missing in those rows
  - So dropping those rows with column value “True”.
- Step 6
  - The column “CALL_TYPE” has three unique values.
- Step 7
  - The columns “TRIP ID”, “TAXI ID” is just unique identifiers for a trip.
  - Dropping these columns doesn’t show any effect in predicting the trip time.

Screenshots of the above steps followed in open refine