Initial dataset stats

- 6,88GB
- 16.477.365 samples
- 23 columns

Pre-cleaning with Miller

Drop unnecessary columns.
Trip ID is irrelevant, Payment/ operator information other than the fare are not needed
mlr -I --csv cut -x -f "Trip ID,Tips,Tolls,Extras,Trip Total,Payment Type,Company" taxi_main.csv

Test for- and drop more unncessary columns

If filter returns 0 rows -- we can drop them, because we don't have to map them to get missing values of 'Centroid Location

1. Filter samples that have truthy Pickup Latitude and Longitude Values but NA's in Pickup Centroid Location column
mlr --csv filter -S '$["Pickup Centroid Longitude"] != "" && $["Pickup Centroid Latitude"] != "" && $["Pickup Centroid Location"] == "" taxi_main.csv

2. Filter samples that have truthy Dropoff Latitude and Longitude Values but NA's in Dropoff Centroid Location column
mlr --csv filter -S '$["Dropoff Centroid Longitude"] != "" && $["Dropoff Centroid Latitude"] != "" && $["Dropoff Centroid Location"] == "" taxi_main.csv

Since both filters returned nothing, we can drop the 4 columns.
mlr -I --csv cut -x -f "Pickup Centroid Longitude,Pickup Centroid Latitude,Dropoff Centroid Longitude,Dropoff Centroid Latitude" taxi_main.csv

3. Filter samples that have truthy Pickup Community Area Values but NA's in Pickup Centroid Location column. Community Areas provide less accurate geospatial data than the census tracts or GPS data. Thus, if they are not needed to map NA's of the other two, they are nood needed.
mlr --csv filter -S '$["Pickup Community Area"] != "" && $["Pickup Centroid Location"] == "" taxi_main.csv
4. Filter samples that have truthy Dropoff Community Area Values but NA's in Dropoff Centroid Location column
mlr --csv filter -S '$["Dropoff Community Area"] != "" && $["Dropoff Centroid Location"] == "" taxi_main.csv
Since both filters returned nothing, we can drop the 2 columns.
mlr -I --csv cut -x -f "Pickup Community Area,Dropoff Community Area" taxi_main.csv

Drop invalid rides

1. check whether there are missing timestamps:
mlr --csv filter -S '$["Trip Start Timestamp"] == "" && $["Trip End Timestamp"] == “”’ taxi_main.csv

As this is not the case, we donot need to drop samples here, but there are some samples where either timestamp is missing, but then there is still information about the trip duration for these samples.

2. check whether there are missing geo data. That is, there are no valid values for the Pickup Census Tract, the Pickup Community Area or the Pickup Centroid Location
mlr --csv filter -S '$["Pickup Census Tract"] == "" || $["Pickup Community Area"] == "" || $["Pickup Centroid Location"] == """ taxi_main.csv
As this is the case, we filter out the invalid rides.
mlr -I --csv filter -S '$["Pickup Census Tract"] != "" || $["Pickup Community Area"] != "" || $["Pickup Centroid Location"] != """ taxi_main.csv

3. we also drop the rides without a valid taxi ID, because these rides cannot be matched to a taxi:
mlr -I --csv filter -S '$["Taxi ID"] != """ taxi_main.csv

Adjusted dataset stats:

- 4,25 GB
- 15.247.964 samples
- 10 columns