For Hire Vehicle Taxi Data Quality Report Overview

Overview

This report will outline the initial data quality findings on the TLC For Hire Vehicle (FHV) trip record data taken from the NYC.gov TLC trip record data page. It will include an overview of the dataset, and a review of the continuous and categorical features, including histograms and bar charts. On initial review, this dataset is missing a lot of pickup and drop off data and contains some logical inconsistencies. This dataset is recommended for use in calculating busyness, but will require significant cleaning.

Summary

This dataset consists of FHV fare data in NYC between Jan 2021 and Mar 2024, including pickup and dropoff time, pickup and dropoff taxi zones (as outlined in the taxi maps section on the data page). Fare data and trip distance are also included, however these columns have been dropped as they are not relevant to our task and the dataset was too large to handle otherwise. The dataset consists of 49M rows. There are 1.8M (3%) duplicate lines, as well as some outliers and inconsistencies. Distribution of the data is overall reasonable.

Review Logical Integrity

Test 1: Dropoff time can't be before pickup time.

- 1 instance.

Test 2: Pickup date cannot be before 2021.

- 0 instances.

Test 3: Dropoff date cannot be after 31st Mar 2024

- 717 instances from 1st April 2024 to 1st May 2031.

Test 4: The duration of a taxi ride shouldn't be greater than 5hrs.

- 10.6k instances, ranging from 5hrs to 3061 days.

Test 5: Pickup location ID not in taxi zone lookup table

- 0 instances.

Test 6: Dropoff location ID not in taxi zone lookup table

- 0 instances.

Review Continuous Features

Descriptive Statistics

There are 3 continuous features in this dataset:

- Pickup datetime
    - This has a reasonable distribution between 2021 and 2024, with some outliers that can be seen. The mean datetime is 28th Aug 2022, which is roughly halfway between the start and end of time collection. There are no missing values. There appear to be outliers at both the high end.
- Dropoff datetime
    - This has a near identical distribution to pickup datetime, as is expected. The mean datetime is 28th Aug 2022, which is roughly halfway between the start and end of time collection. There are no missing values. There appear to be outliers at both the low and high end.

Histograms

All Histograms can be found in the appendix as a summary sheet. All features show a plausible distribution.

Review Categorical Features

There are 2 categorical features in this dataset:

- Pickup location ID
    - This has 264 unique values. The most frequent is zone 206 with 264k pickups. 78% of data is missing.
- Dropoff location ID
    - This has 265 unique values. The most frequent is zone 265. 16% of data is missing.

Actions to Take

- Action 1: Drop rows where dropoff time is before pickup time.
- Action 2: Add "passengers" column with 1 as passenger number (median of yellow and green taxi passenger count).
- Action 3: Drop rows where pickup date is before 2021.
- Action 4: Drop rows where dropoff date is after 31st Mar 2024.
- Action 5: Drop rows where the taxi ride duration is greater than 5hrs.

Appendix

Continuous Features

Descriptive Statistics

Feature	count	mean	min	25%	50%	75%	max	%missing	card
pickup_datetime	49111129	49:03.6	01/01/2021 00:00	01/11/2021 20:46	31/08/2022 07:58	03/07/2023 19:13	31/03/2024 23:59	0	33227791
dropOff_datetime	49111129	24:30.5	28:24.3	01/11/2021 21:12	31/08/2022 08:27	03/07/2023 20:10	01/05/2031 16:00	0	33507864

Histograms

Histograms were created from a random sample of 1M rows.
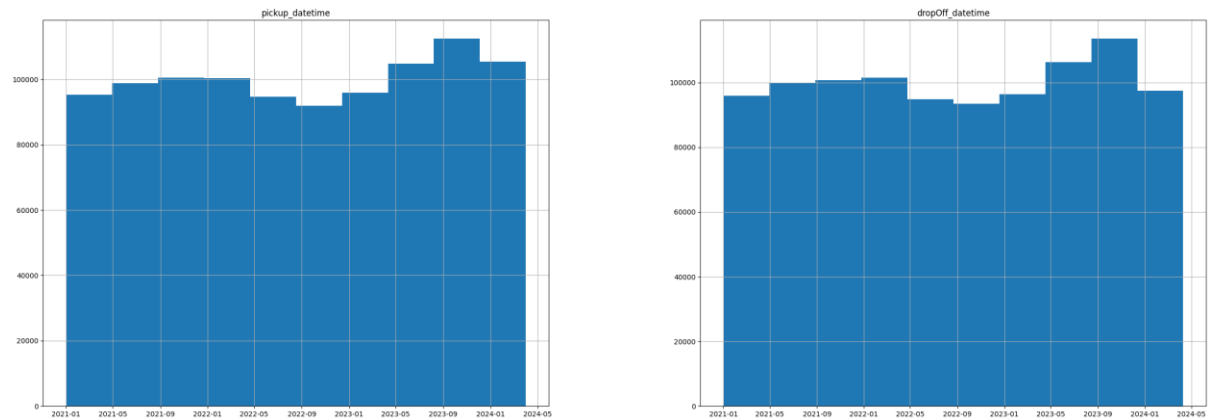
Categorical Features`

Descriptive Statistics

Feature	count	unique	top	freq	second	second_freq	%missing	card
PUlocationID	10480257	264	206	264094	221	236968	78.66011795411994	264
DOlocationID	41267339	265	265	1176798	76	1073384	15.971512281869959	265

Box Plots

Box plots were created from a random sample of 1M rows.

PUlocationID

DOlocationID

DOlocationID