High Volume For Hire Vehicle Taxi Data Quality Report

Overview

This report will outline the initial data quality findings on the TLC For Hire Vehicle High Volume (FHVHV) 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 FHVHV 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 680M rows. Due to the large size of this dataset, a sample of 138M rows was chosen to evaluate the dataset. Of these rows, there are 1200 (>>0.1%) 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.

- 7558 instance.

Test 2: Pickup date cannot be before 2021.

- 0 instances.

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

- 0 instances.

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

- 1606 instances ranging from 5hrs to 1 day 17hrs.

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 23rd Sept 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 23rd Sept 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 262 unique values. The most frequent is zone 206 with 264k pickups. There are no missing values.
- Dropoff location ID
    - This has 263 unique values. The most frequent is zone 265. There are no missing values.

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

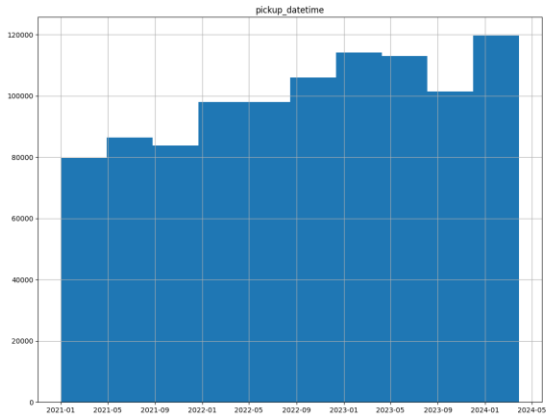
Continuous Features

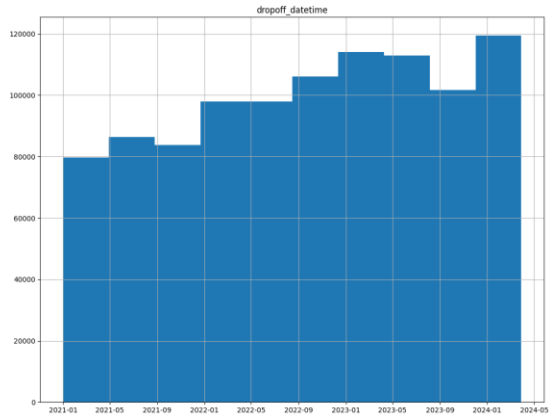
Descriptive Statistics

Feature	count	mean	min	25%	50%	75%	max	%missing	card
pickup_datetime	1.39E+08	05:12.1	01/01/2021 00:00	23/12/2021 16:05	16/10/2022 10:56	10/07/2023 09:16	28/03/2024 16:59	0	20622121
dropoff_datetime	1.39E+08	24:20.9	01/01/2021 00:02	22:25.5	16/10/2022 11:13	10/07/2023 09:36	28/03/2024 21:08	0	20797430

Histograms

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





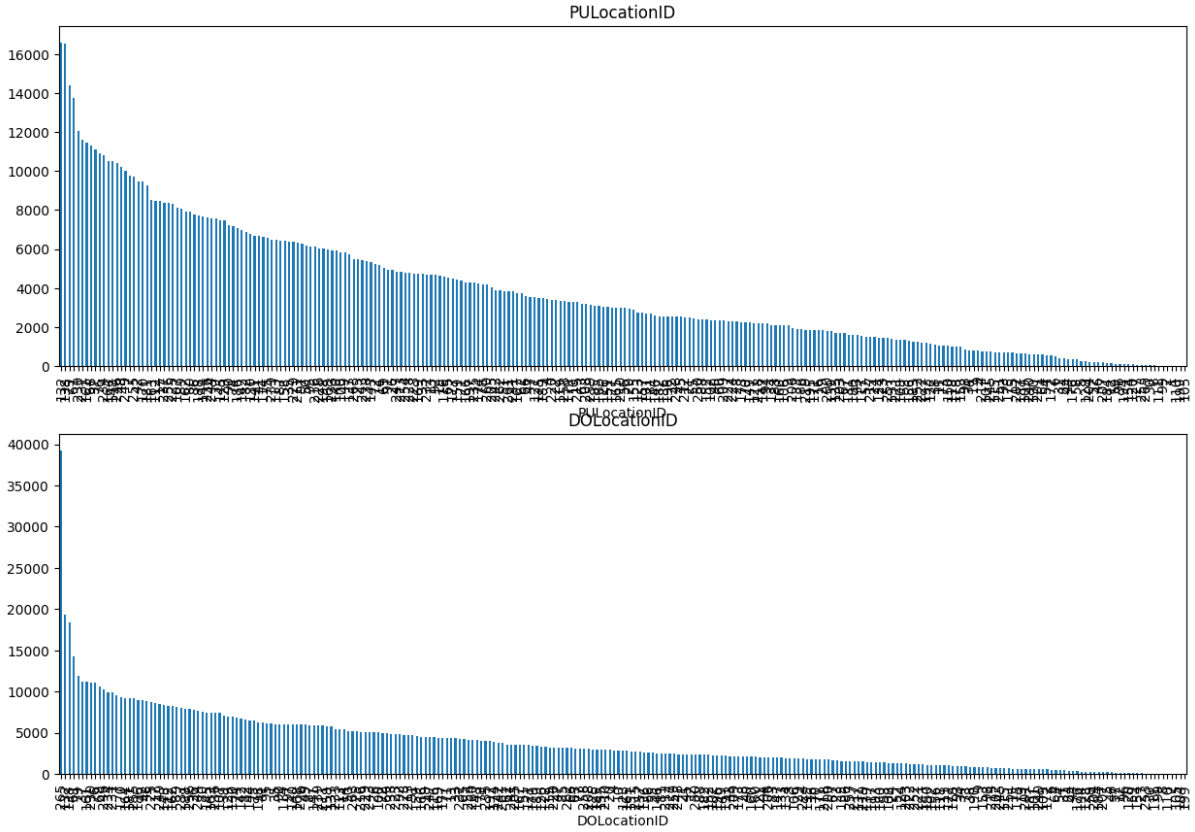
Categorical Features

Descriptive Statistics

Feature	count	unique	top	freq	second	second_freq	%missing	card
PULocatio	1.39E+08	262	138	2292361	132	2291029	0	262
DOLocatic	1.39E+08	263	265	5471778	132	2693935	0	263

Box Plots

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


PULocationID
16000
14000
12000
10000
8000
6000
4000
2000
0
PULocationID

DOLocationID
40000
35000
30000
25000
20000
15000
10000
5000
0
DOLocationID