Green Taxi Data Quality Report Overview

Overview

This report will outline the initial data quality findings on the TLC green cab 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 mostly clean and well documented, however has some missing data and logical inconsistencies. This dataset is recommended for use in calculating busyness.

Summary

This dataset consists of green cab taxi 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 2.8M rows. There are 6.9k (0.2%) 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.

- 3 instances.

Test 2: Passenger count can't be negative.

- 0 instances.

Test 3: Passenger count can't be greater than 8 for green cab taxis. This is the highest passenger count that could be found for a green cab taxi online.

- 396 instances, ranging from 7 to 48.

Test 4: Pickup date cannot be before 2021.

- 44 instances.

Test 5: dropoff date cannot be after 31st Mar 2024

- 5 instances all on 1st April 2024.

Test 6: The duration of a taxi ride shouldn't be greater than 5hrs.

- 10.6k instances, ranging from 24hrs.

Test 7: Pickup location ID not in taxi zone lookup table

- 0 instances.

Test 8: 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 3rd Aug 2022, which is roughly halfway between the start and end of time collection. There are no missing values.
- Dropoff datetime
    - This has a near identical distribution to pickup datetime, as is expected. The mean datetime is 3rd Aug 2022, which is roughly halfway between the start and end of time collection. There are no missing values.
- Passenger count
    - This appears to have a left-skewed distribution, centred around 1. Roughly 4% of values are missing. These should be treated as 0 passengers. The mean number of passengers is 1.29, with a max of 48 and a min of 0.

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 260 unique values. The most frequent is zone 237 with 5.6M pickups. There is no missing data. Taxi zones 110 and 199 appear not to show up, which correspond to great kills park and Riker's island respectively. These would be expected to be low traffic zones.
- Dropoff location ID
    - This has 262 unique values. The most frequent is zone 236 and there is no missing data.

Actions to Take

- Action 1: Drop rows where dropoff time is before pickup time.
- Action 2: Drop rows where passenger count is greater than 6.
- 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.
- Action 6: Fill missing passenger values with 0.

Appendix

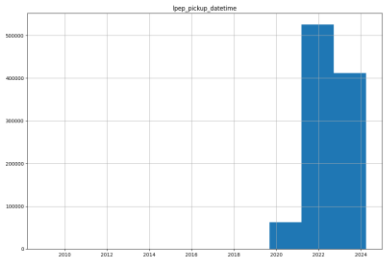
Continuous Features

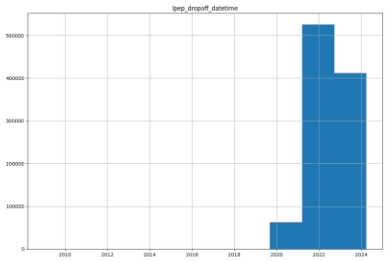
Descriptive Statistics

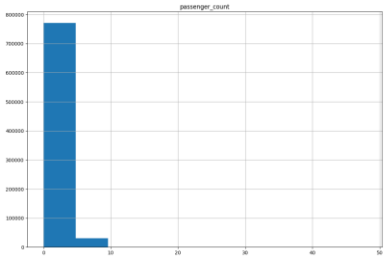
Feature	count	mean	min	25%	50%	75%	max	std	%missing	card
lpep_pickup_datetime	2863802	50:43.1	31/12/2008 17:04	21/09/2021 11:22	41:02.5	50:47.5	01/04/2024 00:01		0	2617298
lpep_dropoff_datetime	2863802	12:06.5	31/12/2008 17:55	45:42.8	31/05/2022 19:59	09:48.7	01/04/2024 16:11		0	2617820
passenger_count	2296975	1.2863030	0	1	1	1	48	0.9217956	19.7928138	12

Histograms

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







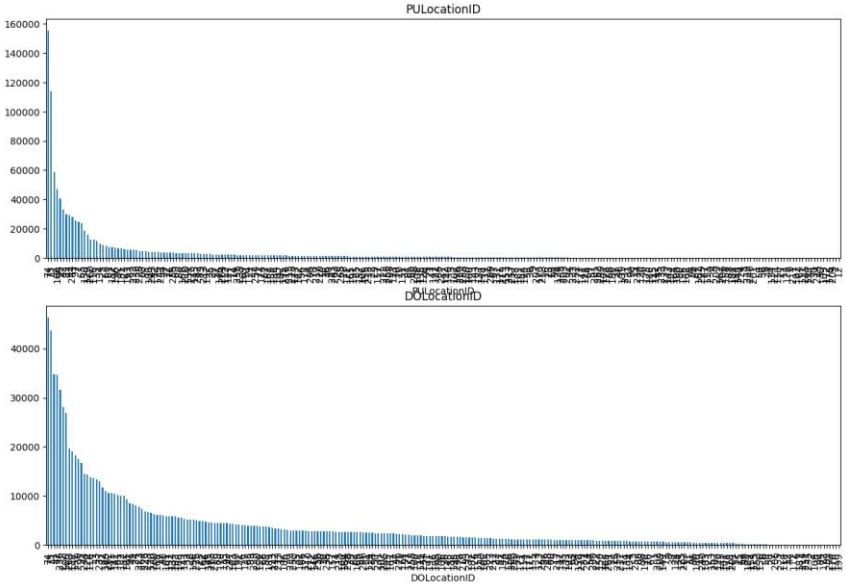
Categorical Features`

Descriptive Statistics

Feature	count	unique	top	freq	second	second_freq	%missing	card
PULocationID	2863802	260	74	445552	75	325502	0	260
DOLocationID	2863802	262	74	132873	75	125275	0	262

Box Plots

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


PULocationID
160000
140000
120000
100000
80000
60000
40000
20000
0
PULocationID

DOLocationID
40000
30000
20000
10000
0
DOLocationID