# Yellow Taxi Data Quality Report Overview

## Overview

This report will outline the initial data quality findings on the TLC yellow 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 yellow 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 118M rows. There are 784k (0.7%) 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.

- 54k instances (.04%).

Test 2: Passenger count can't be negative.

- 0 instances.

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

- 1084 instances, ranging from 7 to 112.

Test 4: Pickup date cannot be before 2021.

- 923 instances. Some appear to be taxis where pickup was before the new year and dropoff was after the new year.

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

- 621 instances randing from 1st April 2024 to 11th Sept 2098.

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

- 142k instances, ranging from 5hrs to 7 days 21hrs.

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 16th Dec 2021, 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 16th Dec 2021, 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, centered around 1. Roughly 4% of values are missing. These should be treated as 0 passengers. The mean number of passengers is 1.39, with a max of 112 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 263 unique values. The most frequent is zone 237 with 5.6M pickups. There is no missing data.
- Dropoff location ID

- This has 262 unique values. Taxi zone 199 is missing from this dataset, which corresponds to Riker's island. This is a jail. Only 52 pickups have happened from this location, it is reasonable to think that prisoners would be picked up from this location but would be dropped off by family members. 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

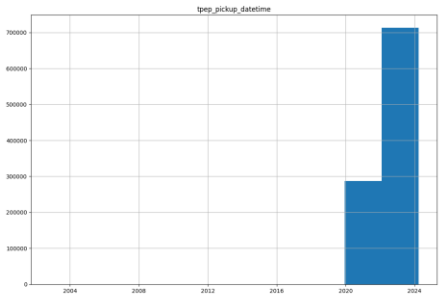
### Continuous Features

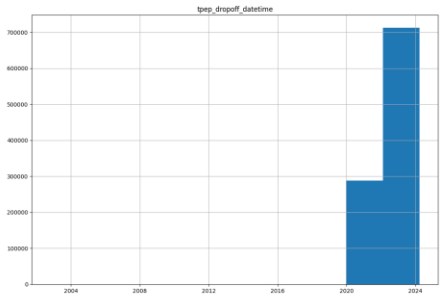
#### *Descriptive Statistics*

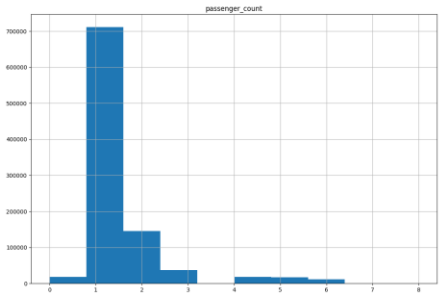
| Feature | count | mean | min | 25% | 50% | 75% | max | std | %missing | card |
|---|---|---|---|---|---|---|---|---|---|---|
| tpep_pickup_datetime | 118425410 | 25:15.9 | 01/01/2001 00:03 | 09:11.2 | 22/09/2022 06:41 | 50:32.7 | 11/09/2098 02:23 | | 0 | 61065577 |
| tpep_dropoff_datetime | 118425410 | 54:13.0 | 20/01/1970 10:16 | 16/12/2021 14:28 | 22/09/2022 06:56 | 07:41.7 | 11/09/2098 02:52 | | 0 | 61029171 |
| passenger_count | 113517094 | 1.3927819 | 0 | 1 | 1 | 1 | 112 | 0.9512203385240618 | 4.144647673164062 | 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 | 1.18E+08 | 263 | 237 | 5643457 | 132 | 5362488 | 0 | 263 |
| DOLocationID | 1.18E+08 | 262 | 236 | 5240099 | 237 | 4974312 | 0 | 262 |

## *Box Plots*

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