NYC Yellow Taxi Trip Data Cleaning and Preprocessing Report

Team Members: Shelan Mohammed Tahir Tyeb, Aya Firas Abdulkareem, Ahmad Bahdin Jalal and Omar Safar Younis

Introduction

This project uses the New York City Yellow Taxi Trip dataset from the official NYC TLC (Taxi and Limousine Commission) data portal. The overall dataset has millions of records on the Taxi frequency infrastructure of their rides; therefore, looking at each month's dataset, it still contains millions of taxi trip records that have pickup and drop-off times, trip distances, fares, passenger counts, and payment types.

For this project, the following sample of data is selected: the data from October, November, and December 2023 were combined to create a dataset of over 10 million rows and 19 columns. The main goal and initial step are to clean and preprocess the dataset for further analysis or prediction tasks, such as predicting trip duration or fare amount.

Data Loading and Combining

The monthly Parquet files were downloaded from the TLC cloud storage using this URL: https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_YYYY-MM.parquet

Python and pandas were used to load each month's data (Oct, Nov, Dec 2023) and merge them into a single DataFrame using pd.concat(). After combining:

- Total rows: 10,238,567
- Total columns: 19

Preprocessing and Data Cleaning Steps

The following preprocessing steps were applied to clean the dataset:

Step 1: Work on a Copy

A copy of the dataset was created to keep the original data safe.

Step 2: Parse Date and Time Columns

The columns 'tpep_pickup_datetime' and 'tpep_dropoff_datetime' were converted to proper datetime format to allow calculations and time-based analysis.

Step 3: Create Trip Duration Feature

A new column 'duration_min' was created to calculate trip duration in minutes using the difference between pickup and drop-off times. This step increased the number of columns from 19 to 20.

Step 4: Apply Basic Filters

In real-world datasets — especially large ones like the NYC Taxi Trip data — some records are incorrect, incomplete, or unrealistic. These can happen because of:

- Sensor or GPS errors
- Manual data entry mistakes
- Corrupted trip meters
- Missing drop-off or pickup readings

Such bad data can cause wrong analysis results or bias in machine-learning models.

That's why we apply basic filters simple logical rules to remove trips that clearly don't make sense.

Filter	Why It Was Applied
Duration: between 1 and 180 minutes	Trips lasting less than 1 minute usually mean the meter was started/stopped by mistake. Trips over 3 hours (180 minutes) are extremely rare for yellow taxis and often caused by missing drop-off times.
Distance: between 0 and 100 miles	A 0-mile trip means no movement was recorded (error). Over 100 miles is unrealistic for city taxis and likely due to faulty GPS or data entry errors.
Fare amount: 0–1000 USD	Fares under 0 or extremely high (over $1000) usually indicate data corruption or misread meters — typical NYC taxi fares are below $200.
Total amount: 0–1500 USD	Includes taxes, tips, and surcharges. Anything above $1500 is clearly invalid and would distort model averages.
Passenger count: between 1 and 6	A count of 0 means no passengers, which makes no sense for a trip. Over 6 is not possible for standard yellow cabs.

Step 5: Handle Missing (NaN) Values

Missing data was handled as follows:
Numerical columns: filled with median values
Categorical columns: filled with the most frequent (mode) value
After cleaning, there were no missing (NaN) values remaining in the dataset.
Before Cleaning average percentage of missing data (only in columns with Nan): 4.57 %

Step 6: Final Clean Dataset

After all preprocessing steps, the final cleaned dataset had:
- Rows: 9,880,959

- Columns: 20

- No missing values (NaN = False)

This dataset is now ready for analysis or machine learning modeling.

Summary of Cleaning Process

- Load and combine data | Combined Oct, Nov, Dec 2023 data
- Fix datetime columns | Parsed pickup and drop-off times
- Create duration | Added new feature in minutes
- Filter data | Removed unrealistic values
- Handle missing values | Used median/mode imputation
- Create mph column | Calculated average trip speed
- Save dataset | Exported as cleaned_taxi_data.csv

Results

- Original dataset: 10,238,567 rows and 19 columns
- Cleaned dataset: 9,880,959 rows and 20 columns
- No missing values remaining
- Data ready for further data mining and predictive modeling tasks.