| **Field Name** | **Description** |
| --- | --- |
| VendorID | A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc. |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| Passenger\_count | The number of passengers in the vehicle.  This is a driver-entered value. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone in which the taximeter was engaged. |
| DOLocationID | TLC Taxi Zone in which the taximeter was disengaged. |
| RateCodeID | The final rate code in effect at the end of the trip.  1= Standard rate  2= JFK  3= Newark  4= Nassau or Westchester  5= Negotiated fare  6= Group ride |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server.  Y= store and forward trip  N= not a store and forward trip |
| Payment\_type | A numeric code signifying how the passenger paid for the trip. 1= Credit card  2= Cash  3= No charge  4= Dispute  5= Unknown  6= Voided trip |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| MTA\_tax | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in trip. |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |
| Congestion\_Surcharge | Total amount collected in trip for NYS congestion surcharge. |
| Airport\_fee | $1.25 for pick up only at LaGuardia and John F. Kennedy Airports. |

**New York City Taxi Trip Data Analysis**

**1. Introduction**

This document outlines the methodology for analyzing a dataset of New York City (NYC) yellow taxi trip data. The objective is to reduce a large dataset (approximately 1 million observations for a month) to a more manageable size (720 observations) while preserving the data's time-based characteristics.

**2. Data Source and Description**

The data used originates from trip records provided to the NYC Taxi and Limousine Commission (TLC) by authorized technology providers under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The data captures information on individual taxi trips, including:

* Pick-up and drop-off dates/times
* Pick-up and drop-off locations
* Trip distances
* Fares and fare types
* Payment methods
* Driver-reported passenger counts

**3. Data Reduction Methodology**

The data reduction process will prioritize preserving the dataset's temporal characteristics. Here's the planned approach:

**3.1 Data Conversion**

* The initial data format Parquet will be converted to a CSV file using pandas in Python. This conversion promotes wider accessibility and facilitates analysis using various tools.

**3.2 Data Preprocessing**

* The focus will be on extracting and manipulating time-related features. It contains one trip for each hour for the month of September, 2024.

**4. Data Analysis**

Following the data reduction and preprocessing steps, a comprehensive analysis will be conducted:

**4.1 Exploratory Data Analysis (EDA)**

* Descriptive statistics will be generated to summarize the data's key characteristics.

**4.2 Visualization**

* Visualizations such as box plots and scatter plots will be used to explore relationships between variables.

**4.4 Statistical Analysis**

* **One-way ANOVA tests:** These tests will be used to analyze the influence of individual factors (e.g., pickup hour) on a dependent variable (e.g., total fare amount).
* **Two-way ANOVA tests:** These tests will be conducted to examine the combined effects of two factors on a dependent variable.

This data dictionary describes yellow taxi trip data. For a dictionary describing green taxi data, or a map of the TLC Taxi Zones, please visit

https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page