NYC Taxi Dataset Preprocessing Report

1. Objective

The goal of this preprocessing step is to clean and prepare the NYC Taxi dataset by handling missing values, removing outliers, converting datetime columns, and dropping irrelevant features. This prepares the dataset for accurate analysis and modeling.

---

2. Dataset Overview

- **Source:** `/content/train.csv`
- **Initial Shape:** *(rows, columns)* = (7624, 11)
- **Columns:**
    - `id`, `vendor_id`, `pickup_datetime`, `dropoff_datetime`, `passenger_count`
    - `pickup_longitude`, `pickup_latitude`, `dropoff_longitude`, `dropoff_latitude`
    - `store_and_fwd_flag`, `trip_duration`

---

3. Missing Values Handling

Detected Missing Values:

```
dropoff_datetime      1
passenger_count       1
pickup_longitude      1
pickup_latitude       1
dropoff_longitude     1
dropoff_latitude      1
store_and_fwd_flag    1
trip_duration         1
```

- **Strategy:** Drop rows with any missing values (as `fill_missing=False` was used).
- **Rows Dropped:** 1 row

---

4. Outlier Detection and Removal

- **Criteria:** Trip durations that are:
    - Less than or equal to **60 seconds**
    - Greater than or equal to **10,000 seconds**
- **Outliers Detected:** 59 rows
- **Action Taken:** Removed outlier rows
- **Rows After Cleaning:** 7624 - 1 (missing) - 59 (outliers) = **7565 rows**

---

5. Datetime Conversion

- Converted the following columns to `datetime64[ns]`:
    - `pickup_datetime`
    - `dropoff_datetime`

---

6. Dropping Irrelevant Columns

- **Dropped Column:** `id`
- **Reason:** Not useful for modeling or analysis

---

7. Final Data Summary

Metric	Value
Final Row Count	7,565
Final Column Count	10
No. of Null Values	0
Date Range (Pickup)	2016-01-01 to 2016-06-30
Trip Duration (secs)	61 to 7,440 (max ~2 hrs)

---

8. Data Types Overview (after cleaning)

```
vendor_id              int64
pickup_datetime        datetime64[ns]
dropoff_datetime       datetime64[ns]
passenger_count        float64
pickup_longitude       float64
pickup_latitude        float64
dropoff_longitude      float64
dropoff_latitude       float64
store_and_fwd_flag     object
trip_duration          float64
```