## **1. Introduction**

This report presents an analysis of the NYC Yellow Cab dataset for the year 2024, specifically for the months of February, March, April, and May. The dataset was obtained from the [NYC Taxi & Limousine Commission (TLC)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page) and initially stored in Parquet format. Due to the large size of the dataset, only four months of data were combined and converted into CSV format for processing in Jupyter Notebook. The primary objectives of this analysis were data cleaning, outlier detection, feature engineering, and exploratory data analysis.

## **2. Data Collection & Processing**

* The dataset was collected from official sources (NYC TLC website).
* The original data was stored in Parquet files, which were converted to CSV for easier handling.
* Four months of data from 2024 (February - May) were combined for analysis.

## **3. Data Cleaning & Transformation**

* Converted relevant columns to appropriate data types (e.g., timestamps to datetime format).
* Removed missing values and handled inconsistencies.
* Eliminated records with negative fares, trip distances, and unrealistic trip durations.
* Identified and removed outliers using the Interquartile Range (IQR) method.
* Ensured all necessary columns were retained for analysis.

## **4. Feature Engineering**

New features were engineered to enhance the dataset:

* **Trip Duration:** Calculated as the difference between pickup and dropoff times (in minutes).
* **Hour of Day:** Extracted from pickup time to analyze time-based trends.
* **Day of the Week:** Extracted from pickup time to study ride distribution across days.
* **Is Weekend:** A binary variable indicating whether a trip occurred on a weekend.

## **5. Exploratory Data Analysis & Findings**

### **Revenue Analysis by Borough**

* The highest total fare revenue was observed in Manhattan, followed by Brooklyn and Queens.
* Trips originating from JFK and LaGuardia airports contributed significantly to revenue.

### **Trip Duration & Distance**

* The average trip duration was around 12-15 minutes, with a few long-duration trips (>60 minutes).
* Outlier detection helped remove unrealistic trip durations (e.g., trips lasting several hours).
* The majority of trips covered short distances (1-5 miles), with some long-distance rides primarily to airports.

### **Peak Ride Hours & Weekday Trends**

* The highest number of rides occurred between **8 AM - 10 AM** and **5 PM - 7 PM**, indicating rush-hour peaks.
* Weekends had a slightly different distribution, with peak ride times extending into the late evening.

### **Payment Trends**

* **Credit card payments** were the most common payment method.
* **Cash payments** were more frequent for shorter trips.

### **Correlation Analysis**

* A moderate correlation was observed between trip distance and fare amount.
* Congestion surcharge and total fare had a strong positive correlation, indicating that higher congestion charges significantly impact total fare.

## **6. Model Implementation**

To identify patterns and predict trip fare, a **Random Forest Regression model** was implemented using the cleaned dataset:

* **Target Variable:** Total fare amount.
* **Features:** Trip duration, trip distance, hour of day, passenger count, and congestion surcharge.
* **Model Performance:**
  + Achieved a R² **score of 0.85**, indicating a good fit.
  + The most influential factors in fare prediction were trip distance and congestion surcharge.

## **7. Conclusion**

* The analysis provided key insights into ride patterns, revenue distribution, and fare determinants.
* Feature engineering and outlier removal improved data quality for accurate modeling.
* The Random Forest model successfully predicted fare amounts with high accuracy.
* The findings can be useful for **taxi operators, city planners, and ride-sharing companies** to optimize pricing strategies and improve service efficiency.

## **Note:**

* Only four months (Feb-May 2024) were analyzed due to computational limitations in the Jupyter Notebook.
* Data was initially stored in Parquet format and later converted to CSV through the notebook itself and combined with a csv file.