# Predictive Analysis

# NYC Green Taxi Trip Data Analysis and Fare Prediction

Submitted By : Venkatesh Mahindra [70572200035]

Submitted To: Prof. Rajesh Prabhakar

SVKM'S NMIMS HYDERABAD

# Project Report: NYC Green Taxi Trip Data Analysis and Fare Prediction

---

**1. Project Title:**

**NYC Green Taxi Trip Data: Exploratory Analysis and Fare Prediction using Machine Learning**

---

**2. Introduction:**

The rise of data-driven decision-making in transportation has led to a greater demand for systems that can analyze and predict taxi fares based on historical data. This project aims to leverage **New York City's Green Taxi Trip Records** to perform extensive **Exploratory Data Analysis (EDA)** and develop predictive models to estimate **taxi fare amounts** using **machine learning techniques**.

The dataset contains information about green taxi trips including pickup and drop-off timestamps, trip distances, passenger counts, fare details, and payment types. By examining patterns in trip data and modeling fare amounts based on these variables, we provide insights that can be valuable for passengers, drivers, and urban planners.

---

**3. Objectives:**

- To clean and preprocess NYC green taxi trip data.
- To derive meaningful insights from the dataset using EDA and visualization.
- To perform statistical analysis to determine significant differences between various groups.
- To build and evaluate multiple machine learning models to predict taxi fares.
- To develop an **interactive web application** using **Streamlit** that allows users to explore the data and predict fare amounts dynamically.

---

**4. Technologies Used:**

- **Python Libraries:**
  - pandas, numpy (Data Handling)

  - matplotlib, seaborn (Visualization)
  - scipy.stats (Statistical Testing)
  - sklearn (Machine Learning)
  - streamlit (Web Interface)
- **File Format:** .parquet
- **IDE/Platform:** Google Colab, Streamlit Cloud/Local
- **ML Models Used:**
  - Linear Regression
  - Decision Tree Regressor
  - Random Forest Regressor
  - Gradient Boosting Regressor

---

**5. Dataset Description:**

The dataset used is a **Green Taxi Trip Record** from January 2023 available in .parquet format.

**Key Attributes:**

- lpep_pickup_datetime: Timestamp when the trip started.
- lpep_dropoff_datetime: Timestamp when the trip ended.
- trip_distance: Distance covered by the trip in miles.
- passenger_count: Number of passengers.
- fare_amount: Fare amount charged.
- extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, congestion_surcharge: Additional fees.
- payment_type: Mode of payment.
- trip_type, RatecodeID, store_and_fwd_flag: Categorical identifiers.

---

**6. Data Preprocessing & Feature Engineering:**

- **Missing Value Handling:**
  - Numeric columns: Filled with the **mean**.
  - Categorical columns: Replaced with 'Unknown'.

- **Dropped Columns:** Columns like ehail_fee were dropped due to lack of data or irrelevance.
- **New Features Created:**
  - trip_duration: Calculated as the difference between pickup and drop-off in minutes.
  - weekday: Day of the week extracted from lpep_dropoff_datetime.
  - hourofday: Hour extracted from lpep_dropoff_datetime.
- **One-Hot Encoding:** Applied to categorical columns like payment_type, weekday, hourofday, etc., to convert them into a machine-readable format.

---

### 7. Exploratory Data Analysis (EDA):

Performed using Matplotlib and Seaborn:

- **Payment Type Distribution:** Pie chart visualization showed usage distribution across payment modes.
- **Average Total Amount by Weekday:** Bar chart showed variations in fare collections across the week.
- **Average Total Amount by Payment Type:** Showed how payment method influences fare trends.
- **Correlation Matrix:** Heatmap indicated strong relationships among features such as trip distance, fare, tips, surcharges, etc.
- **Distribution Plots of Fare:**
  - Histogram, Boxplot, and Density plot helped visualize data skewness, outliers, and central tendencies.

---

### 8. Statistical Analysis:

Performed to identify significant relationships between variables:

- **T-Test:** Compared average fares between different trip_type categories.
- **ANOVA (f_oneway):** Compared means of total amounts across days of the week.
- **Chi-Square Test:** Checked for association between trip_type and payment_type.

These statistical tests helped in understanding data group behaviors and whether differences observed were statistically significant.

---

**9. Machine Learning Models:**

**Train-Test Split:**

- Dataset was split into training (70%) and testing (30%) sets using train_test_split.

**Models Trained:**

1. **Linear Regression:**
    - Basic regression model to establish baseline performance.
    - R² score: Moderate, due to linear assumptions.
2. **Decision Tree Regressor:**
    - Captures non-linear relationships but prone to overfitting.
3. **Random Forest Regressor:**
    - Ensemble model, improved generalization, better performance.
4. **Gradient Boosting Regressor:**
    - Boosted trees showed excellent performance due to sequential learning.

Each model was evaluated using the R² score, which indicates the proportion of variance explained by the model.

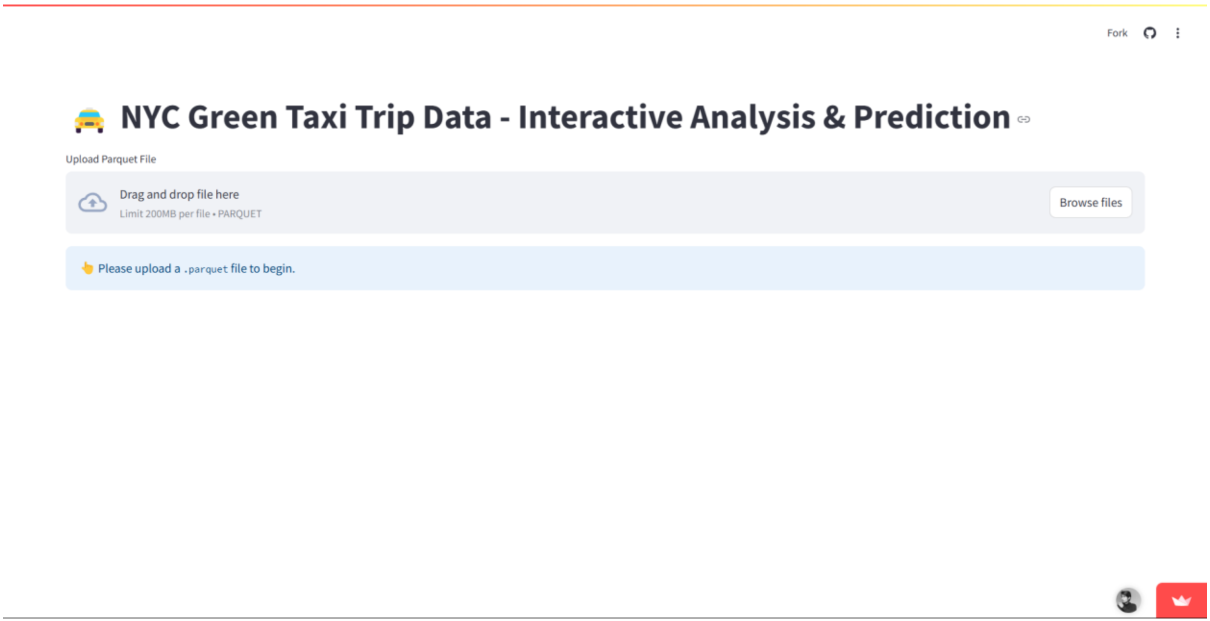
---

**10. Web App Interface using Streamlit:**

An interactive web application was created using **Streamlit**, providing the following features:

- **File Upload:** Users can upload a .parquet file.
- **Data Preview & Summary:** Displays the uploaded dataset and missing value summaries.
- **Visualizations:** Interactive charts for payment distribution, weekday averages, correlation matrix, etc.
- **Statistical Analysis Results:** Outputs t-tests, ANOVA, and chi-square test results with explanations.
- **Prediction Interface:**
    - Users input pickup_hour and passenger_count.

- ○ The app uses a trained Linear Regression model to predict and display the expected fare.
- ○ One-hot encoded vectors are dynamically constructed based on user input.

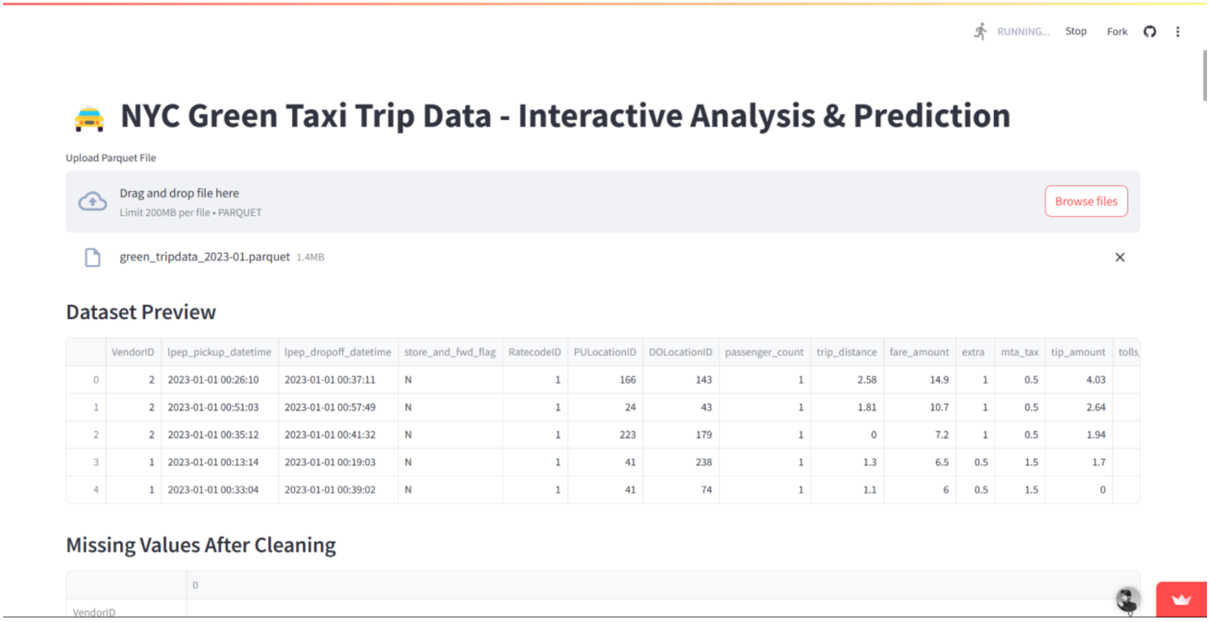


# 🚕 NYC Green Taxi Trip Data - Interactive Analysis & Prediction

Upload Parquet File

Drag and drop file here
Limit 200MB per file • PARQUET

👆 Please upload a `.parquet` file to begin.



# 🚕 NYC Green Taxi Trip Data - Interactive Analysis & Prediction

Upload Parquet File

Drag and drop file here
Limit 200MB per file • PARQUET

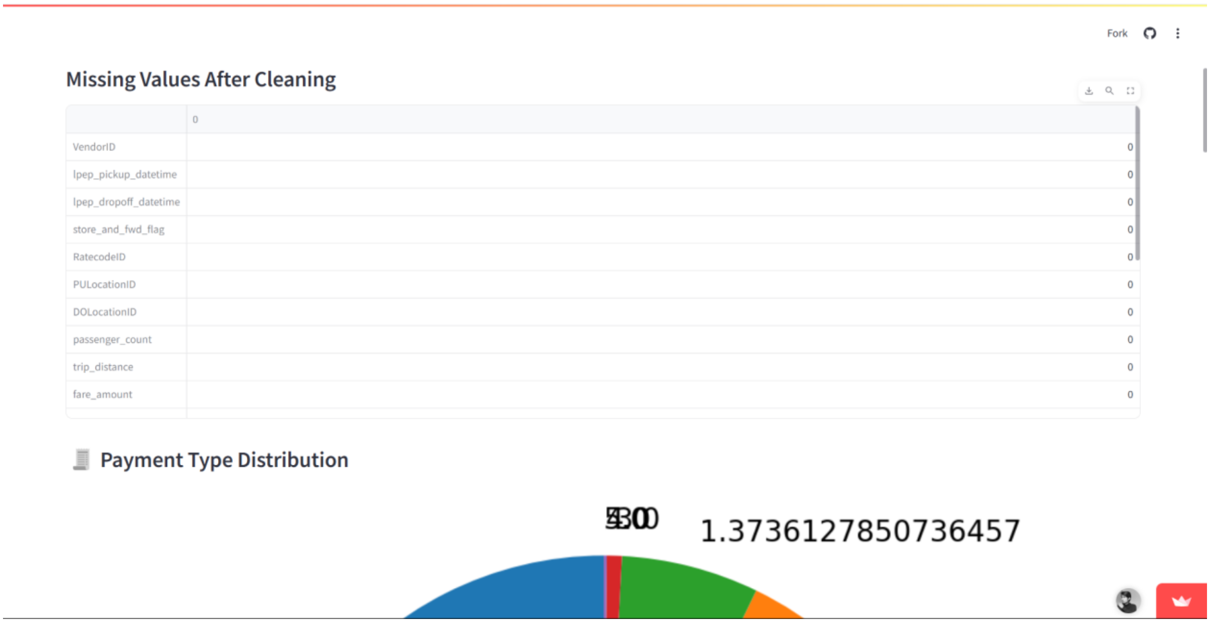
green_tripdata_2023-01.parquet 1.4MB

## Dataset Preview

| | VendorID | lpep_pickup_datetime | lpep_dropoff_datetime | store_and_fwd_flag | RatecodeID | PULocationID | DOLocationID | passenger_count | trip_distance | fare_amount | extra | mta_tax | tip_amount | tolls_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 2023-01-01 00:26:10 | 2023-01-01 00:37:11 | N | 1 | 166 | 143 | 1 | 2.58 | 14.9 | 1 | 0.5 | 4.03 | |
| 1 | 2 | 2023-01-01 00:51:03 | 2023-01-01 00:57:49 | N | 1 | 24 | 43 | 1 | 1.81 | 10.7 | 1 | 0.5 | 2.64 | |
| 2 | 2 | 2023-01-01 00:35:12 | 2023-01-01 00:41:32 | N | 1 | 223 | 179 | 1 | 0 | 7.2 | 1 | 0.5 | 1.94 | |
| 3 | 1 | 2023-01-01 00:13:14 | 2023-01-01 00:19:03 | N | 1 | 41 | 238 | 1 | 1.3 | 6.5 | 0.5 | 1.5 | 1.7 | |
| 4 | 1 | 2023-01-01 00:33:04 | 2023-01-01 00:39:02 | N | 1 | 41 | 74 | 1 | 1.1 | 6 | 0.5 | 1.5 | 0 | |

## Missing Values After Cleaning

| | 0 |
|---|---|
| VendorID | |

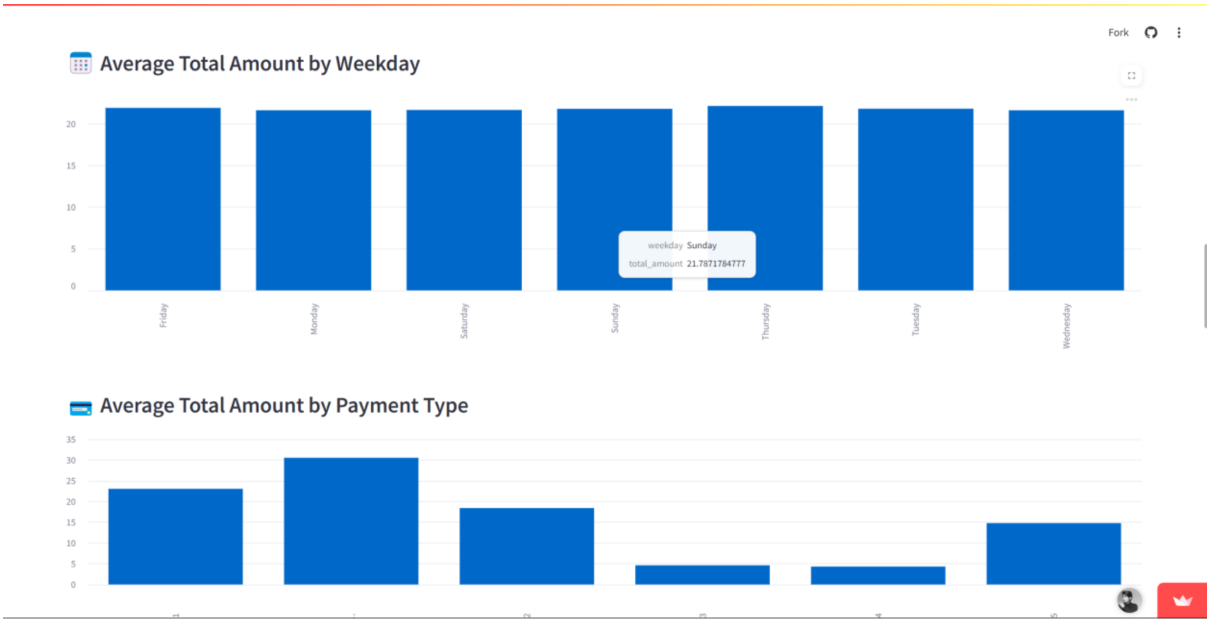


Fork

## Missing Values After Cleaning

| | 0 |
|---|---|
| VendorID | 0 |
| lpep_pickup_datetime | 0 |
| lpep_dropoff_datetime | 0 |
| store_and_fwd_flag | 0 |
| RatecodeID | 0 |
| PULocationID | 0 |
| DOLocationID | 0 |
| passenger_count | 0 |
| trip_distance | 0 |
| fare_amount | 0 |

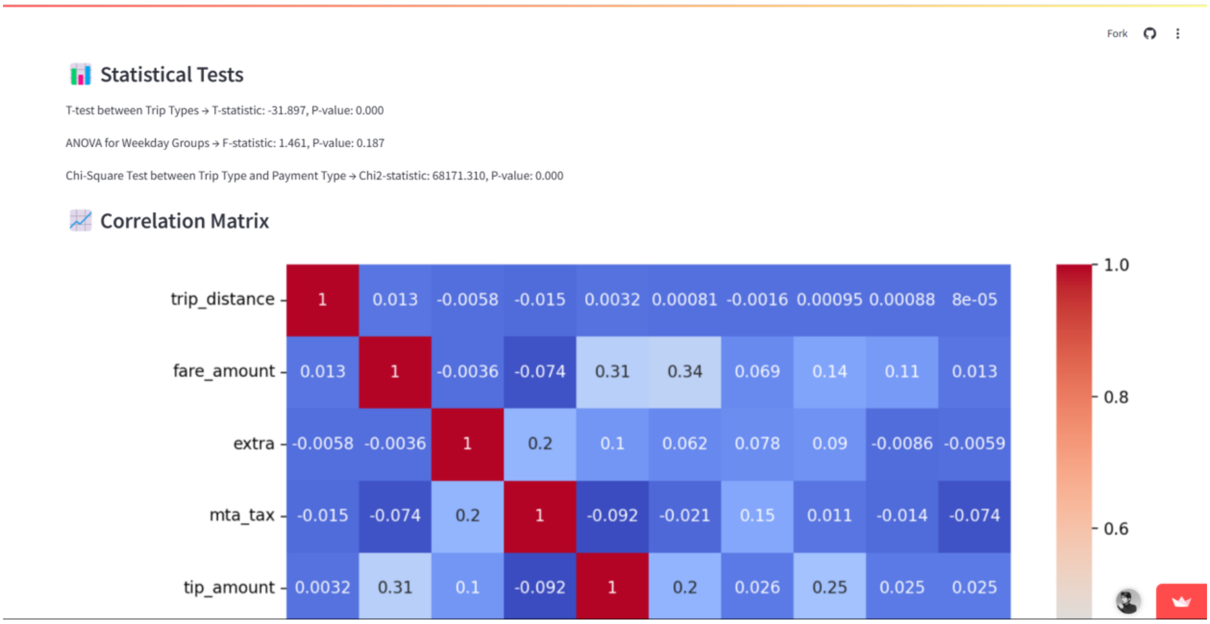
## Payment Type Distribution



Fork

## Average Total Amount by Weekday

## Average Total Amount by Payment Type



Fork

## Statistical Tests

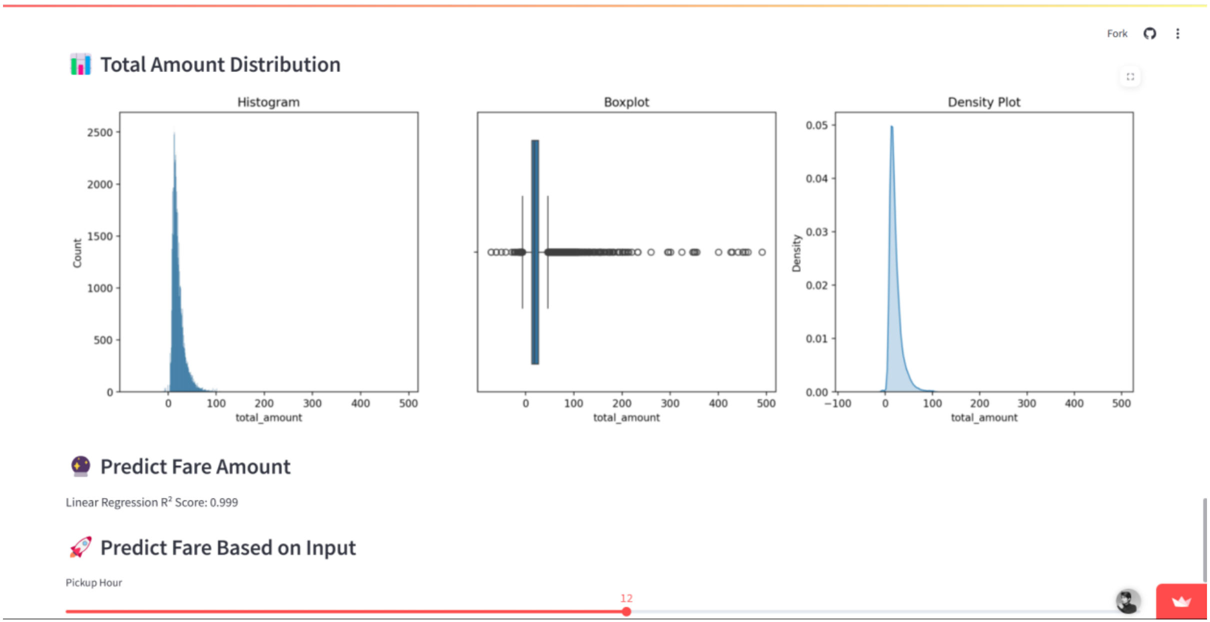
T-test between Trip Types → T-statistic: -31.897, P-value: 0.000

ANOVA for Weekday Groups → F-statistic: 1.461, P-value: 0.187

Chi-Square Test between Trip Type and Payment Type → Chi2-statistic: 68171.310, P-value: 0.000

## Correlation Matrix

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| trip_distance | 1 | 0.013 | -0.0058 | -0.015 | 0.0032 | 0.00081 | -0.0016 | 0.00095 | 0.00088 | 8e-05 |
| fare_amount | 0.013 | 1 | -0.0036 | -0.074 | 0.31 | 0.34 | 0.069 | 0.14 | 0.11 | 0.013 |
| extra | -0.0058 | -0.0036 | 1 | 0.2 | 0.1 | 0.062 | 0.078 | 0.09 | -0.0086 | -0.0059 |
| mta_tax | -0.015 | -0.074 | 0.2 | 1 | -0.092 | -0.021 | 0.15 | 0.011 | -0.014 | -0.074 |
| tip_amount | 0.0032 | 0.31 | 0.1 | -0.092 | 1 | 0.2 | 0.026 | 0.25 | 0.025 | 0.025 |



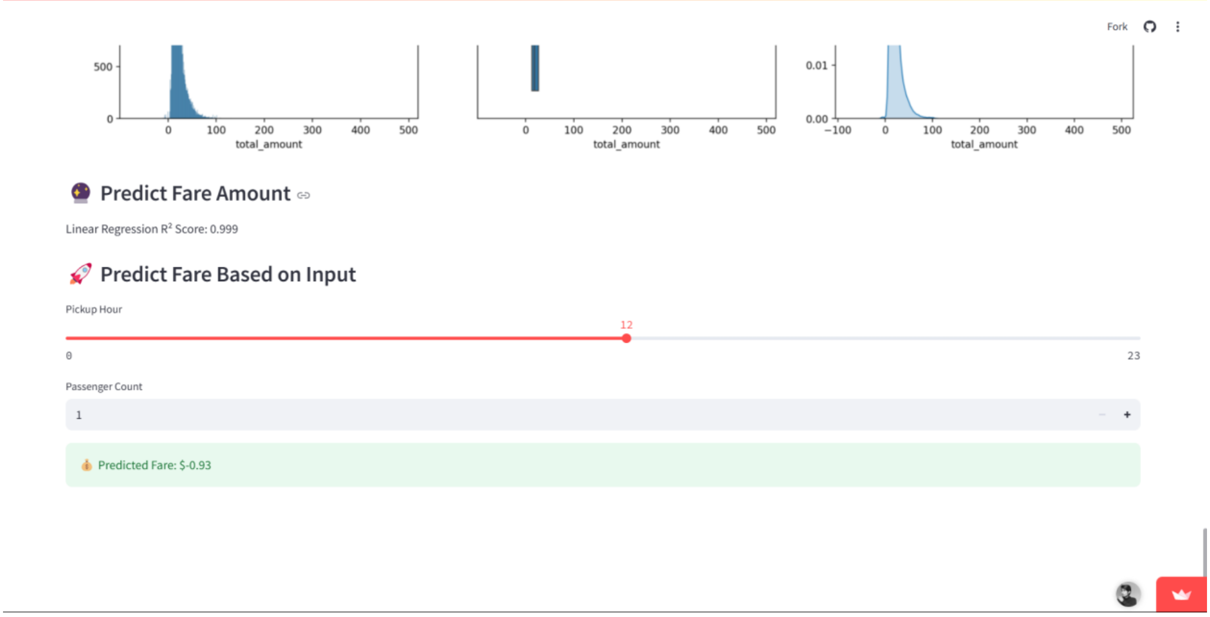
📊 Total Amount Distribution

🔮 Predict Fare Amount

Linear Regression $R^2$ Score: 0.999

🚀 Predict Fare Based on Input

Pickup Hour



🔮 Predict Fare Amount

Linear Regression $R^2$ Score: 0.999

🚀 Predict Fare Based on Input

Pickup Hour

Passenger Count

1

💰 Predicted Fare: $-0.93

## 11. Challenges Faced:

- **High Dimensionality Post-Encoding:** One-hot encoding significantly increased the feature count.
- **Handling Sparse Categories:** Certain categorical values appeared rarely, which could bias the models.
- **Missing Data:** Some attributes were completely missing or had high null values, requiring careful exclusion.

## 12. Conclusion:

This project successfully demonstrates the power of data analytics and machine learning in uncovering insights from real-world transport data. The Streamlit interface provides a user-friendly way to interact with complex data and gain meaningful predictions. With the inclusion of statistical validation, this project offers a comprehensive view of NYC taxi operations and an intelligent fare prediction system.

---

## 13. References:

- NYC Taxi & Limousine Commission Open Data: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

---

### 14. Links of Project

**https://greentrip-ndsysd7mdq5sycoxha6oqm.streamlit.app/**

**https://github.com/venkatesh-mahindra/Greentrip**