NYC Taxi Fare Prediction — EDA Summary

Automatidata • NYC Taxi Analytics Project

Project Overview:

This project aims to predict NYC taxi fares before a ride begins using historical trip data. EDA was conducted to validate data quality, identify key fare drivers, and remove anomalous records that could distort predictions.
The cleaned dataset is prepared for regression-based fare estimation.

Key Insights

❖ Data quality issues were identified, including **zero-distance trips with high fares** and **negative fare amounts**, indicating data anomalies.

❖ Correlation analysis showed that **trip distance and trip duration** are the primary drivers of fare amount, while several variables added little predictive value and were removed as noise.

❖ Extreme fare values were further evaluated using **distance–fare scatter plots** and a **fare-per-mile analysis**, which revealed unrealistic pricing patterns.

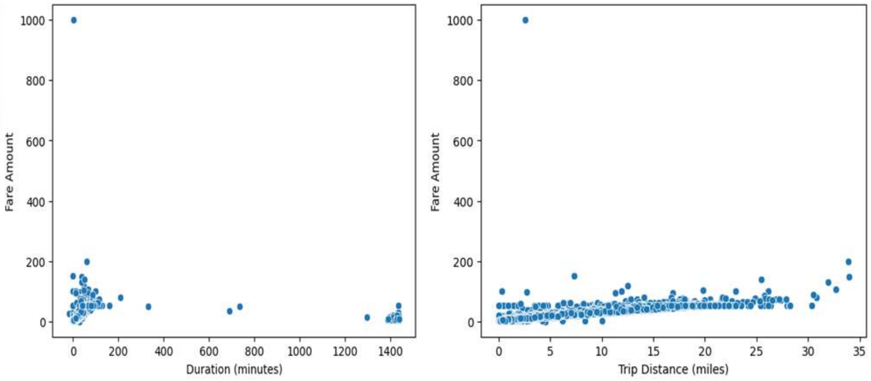
❖ Based on visual trends, pricing logic, and fare-per-mile analysis, anomalous fare records were removed.

❖ Outcome:

❖ The cleaned dataset shows a **stable and interpretable relationship** between distance, duration, and fare

❖ This ensures the predictive model will learn **realistic pricing patterns**

Details



Validation:

❖ Low trip distance with very high fares.
❖ Very short and very long durations paired with implausible fares.
❖ This prompted deeper validation rather than immediate removal.

Scatter plots of:

Trip distance vs. fare amount
Trip duration vs. fare amount

❖ revealed that these extreme fares **did not follow the expected pricing trend**, confirming that they were inconsistent with normal taxi fare behavior.

Next Steps

Conclusion:

❖ The EDA process successfully identified and corrected data quality issues, isolated meaningful predictors, and removed anomalous fare values that violated real-world pricing constraints. The resulting dataset is well-structured, reliable, and suitable for predictive modelling.

Next Step:

❖ **Model Selection and Baseline Regression;**
❖ The next phase will focus on selecting an appropriate regression model to estimate taxi fares using the finalized feature set.