Exploratory Data Analysis of New York City TLC Data

Executive summary report

By Automatidata

ISSUE / PROBLEM

The NYC Taxi & Limousine Commission has partnered with Automatidata to develop a regression model for predicting taxi ride fares. At this stage of the project, the focus is on analyzing, exploring, cleaning, and organizing the data to ensure it is properly prepared for modeling.

IMPACT

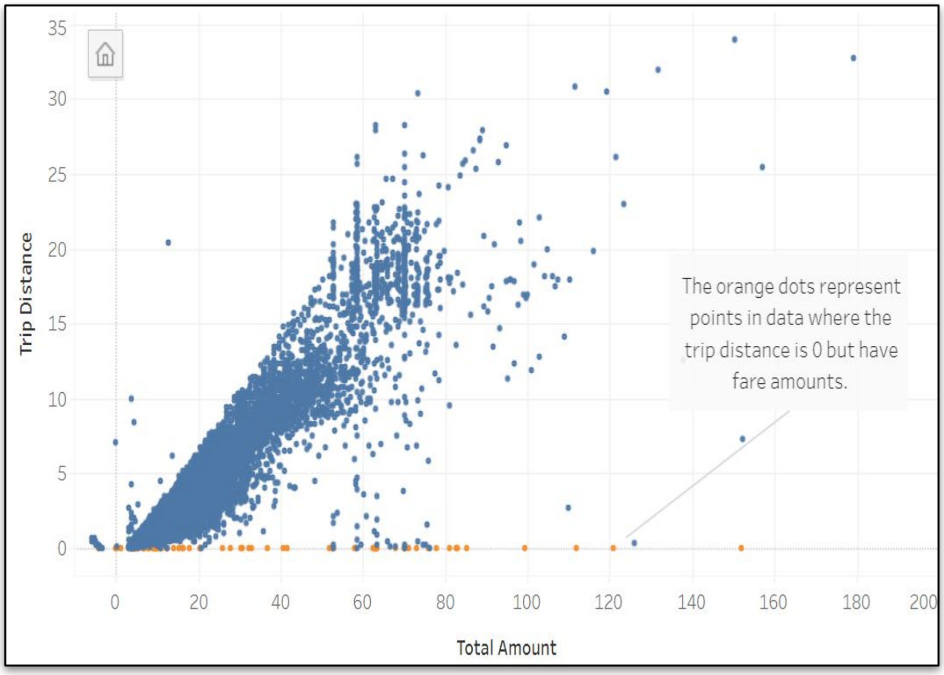
After running initial exploratory data analysis (EDA) on a sample of the data provided by New York City TLC, it is clear that some of the data will prove an obstacle for accurate ride fare prediction. Namely, trips that have a total cost entered, but a total distance of "0." At this point, our analysis indicates these to be anomalies or outliers that need to be factored into the algorithm or removed completely.

Proposed solution: After analysis, we recommend removing outliers with a total distanced recorded of 0.

RESPONSE

- Identify any outliers or anomalies in the data that could interfere with the accuracy of future fare predictions, such as trips with unusually long durations.
- Determine which variables have the greatest influence on trip fares.
- Narrow the dataset to focus on the most relevant variables for regression modeling, statistical analysis, and parameter tuning.

KEY INSIGHTS



The orange dots represent points in data where the trip distance is 0 but have fare amounts.

As a result of the conducted exploratory data analysis, the Automatidata data team considered trip distance and total amount as key variables to depict a taxi cab ride. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau to enhance the provided visualization.