Exploratory Data Analysis of New York City TLC Data

Executive summary report

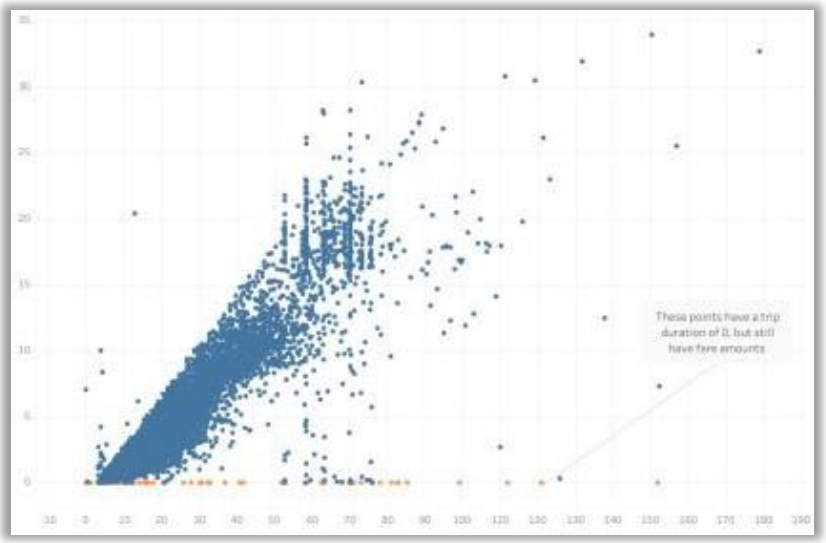
Commission Prepared by Automatidata

Project Overview

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab ride fares. In this part of the project, the data needs to be analyzed, explored, cleaned and structured prior to any modeling.

Details

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.



Alt Text: Graph displaying New York City TLC data plotting variables for total distance (y axis) and total amount (x axis)

Key Insights

The Problem: 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. The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab ride fares. In this part of the project, the data needs to be analyzed, explored, cleaned and structured prior to any modeling.

Keys to success:

● Ensuring with New York City TLC that the sample provided is an accurate reflection of their data as a whole.

● Plan for handling other outliers, such as low trip distance paired with high costs.

Next Steps:

● Determine any unusual data points that could pose a problem for future analysis in predicting trip fares. For example, locations that have longer durations.

● Determine the variables that have the largest impact on trip fares.

● Filter down to consider the most relevant variables for running regression, statistical analysis, and parameter tuning.