Exploratory Data Analysis of New York City TLC Data

Executive summary report

Commission Prepared by **Akshaj Piri**

Project Overview

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a machine learning model that predicts taxi/limousine ride durations.

Key Insights

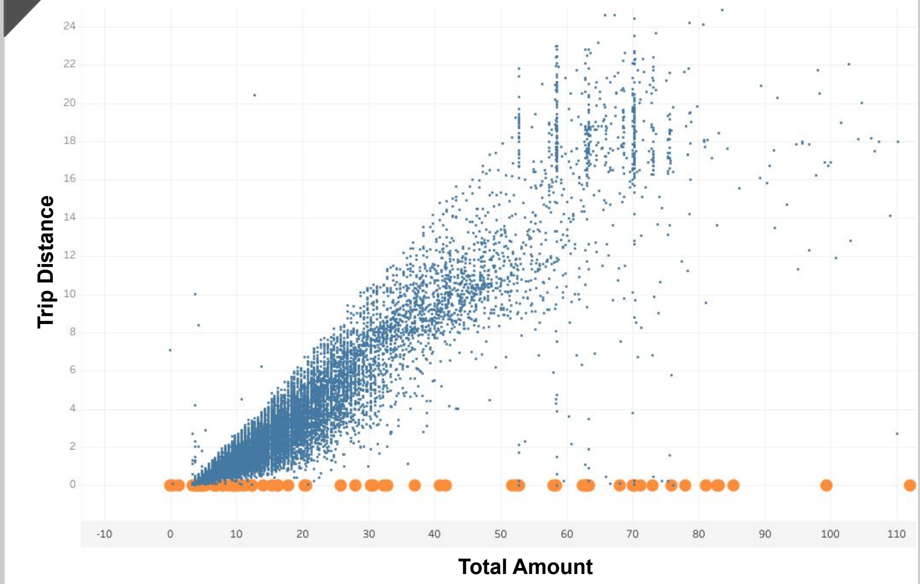
The Problem: After running initial exploratory data analysis (EDA) on a sample of the data provided by New York City TLC, it is clear that some of the data will prove an obstacle for accurate ride duration prediction. Namely, trips that have a total cost entered, but a total distance of "0." At this 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.

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 high costs.

Details



Next Steps

- Determine "problem areas" for predicting trip duration.
  - For example, locations that have longer durations.
- Determine the variables that have the largest impact on trip duration.
- Pare down data to the most relevant variables for running regression, statistical analysis, and parameter tuning.