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.

Key Insights

The Problem: After running an exploratory data analysis on a sample of the data provided by the 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 model or removed completely.

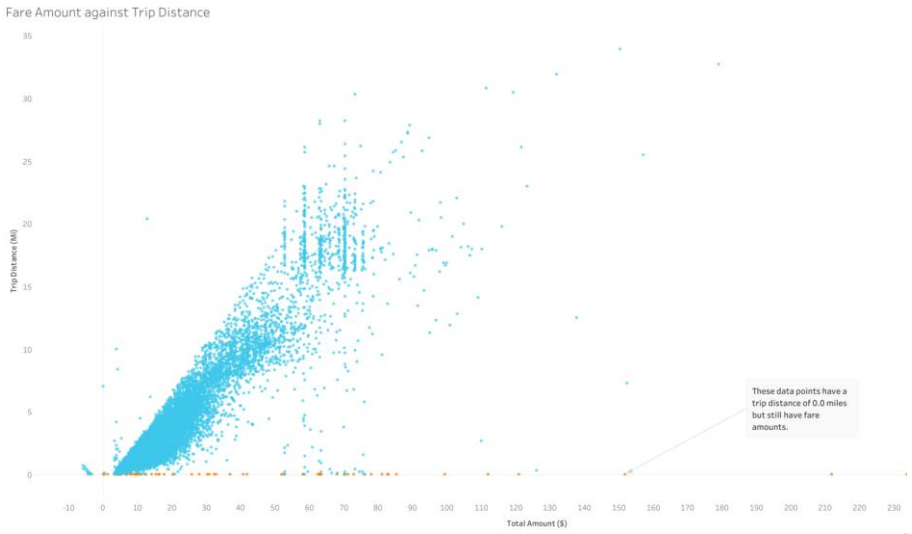
Proposed solution: We recommend removing outliers with a total distance recorded of 0.

Keys to success

- Check with the 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 with high costs, and trip distance with negative costs.

Details

As a result of the exploratory analysis, the Automatidata data team considered **trip distance** and **total amount** as key variables to represent a taxi cab ride. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau.
Link:
https://public.tableau.com/app/profile/kuebiko/viz/GoogleAdvancedDACourse3_3AutomatidataVisualization/Sheet1?publish=yes



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.