Exploratory Data Analysis(EDA) of New York City TLC Data

Executive summary report

Prepared by **Automatidata**

Project Overview

The NYC Taxi & Limousine Commission(TLC) has hired Automatidata to build a machine learning model that can predict ride fares. In this part of the project, the data is analyzed, explored, cleaned and structured appropriately as required for the modelling.

Key Insights

The Problem: Early EDA shows data that look anomalous. There are data entries with negative total fare, 0 trip distance, outliers in fare amount. We need to check the source of these data entries and decide whether to keep them or drop them before our model construction.

Proposed solution: After analysis, we recommend removing data entries with trip distance recorded of 0 and negative total fare amount. The fare outliers with non-zero trip distance will be kept since some factors such as high traffic/demand can cause this.

Keys to success

- Ensuring that the sample provided is an accurate reflection of NYC TLC data as a whole and there are not any biases present in the data.
- Ways for handling the outliers such as low trip distance paired with high costs which we decided to incorporate in our model construction.

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. This scatter plot was created in Tableau to enhance the provided visualization.

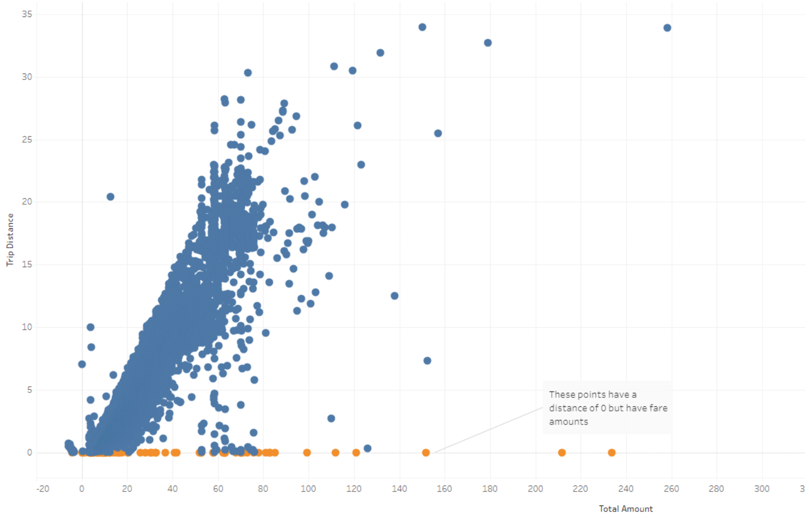

Tableau Viz: New York City TLC data plotting variables for total distance and total amount. Shows clear outliers in trip distance.

Next Steps

- Check for any unusual data left that could cause problems with model training.
- Determine the variables that have the largest impact on trip fares. Plot a correlation heatmap to drop any redundant features.
- From all the available features, select the appropriate ones for running regression and statistical analysis.