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

Executive summary report

OVERVIEW

The New York City TLC has contracted with us 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.

PROJECT STATUS

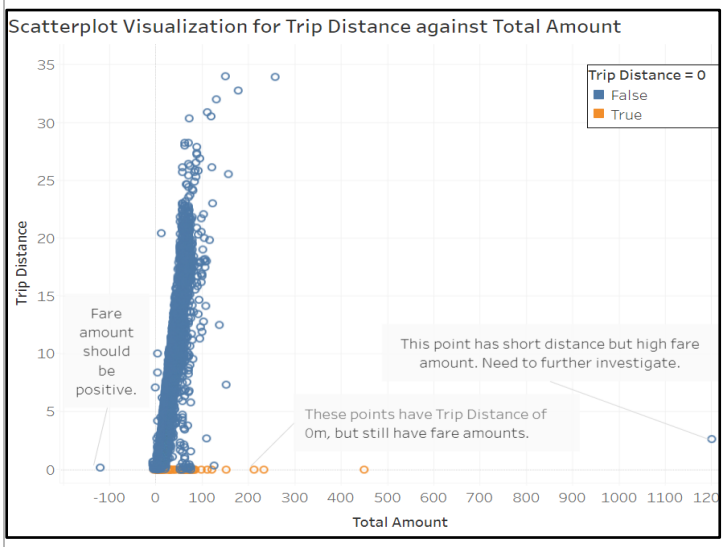
The Problem: After running initial EDA on a sample of the data provided, it is clear that some of the data will prove an obstacle for accurate ride fare prediction. For examples, trips that have a total cost entered, but a total distance of "0" and total cost that are overly charged or negatively charged.

Proposed solution: We recommend removing these outliers as these data are human error during recording.

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.

KEY INSIGHTS



As a part of EDA using Tableau for visualization, the data team considered trip distance and total amount as key variables to represent a taxi cab ride. The scatterplot above shows the relationship between the two variables.

There are some obvious outliers to be factored in as this will affect the regression gradient for analysis on the modelling stage.

In overall, the total amount is directly proportional to the trip distance based on the observation.