Exploratory Data Analysis of New York TLC Data

Executive Summary
Commision prepared by **Automatidata**

Project Overview

New York City Taxi and Limousine Commission (New York City TLC), has hired the Automatidata team to create a regression model to predict the taxi cab fares. In this part of the project, we analyzed, cleaned, and structured the data before creating the modelling.

Key Insights

Problem: We performed Exploratory Data Analysis (EDA) on the data provided by New York City TLC. When performing the analysis to relate the trip distance and total amount, we notice in the trip distance values with 0. Once finished our analysis, we determine that the te values mentioned earlier are outliers and it could interfere in the creation of the regression model.

Proposed solution: We recommend to eliminate the 0 values from total distance column.

Key Values:

- Ensure that the data provided by New York TLC is representative.
- Remove the outliers, such as trip distance.

Details

As a result of the EDA, we conclude that the trip distance and total amount are the important variables for the creation of the regression model. The Scatter Plot was created in Tableau

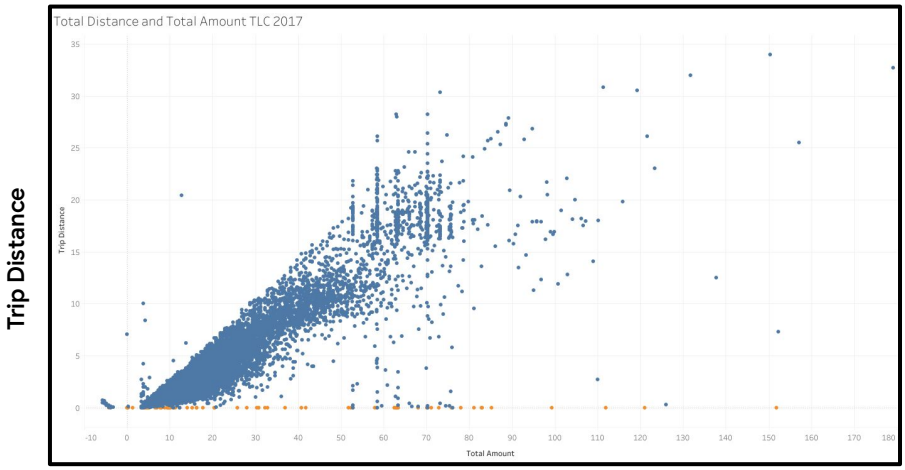

Figure 1. Graph displaying New York City TLC data plotting variables for total distance and total amount.

Next Steps

- Check the data for the outliers
- Look for other variables that may have an impact on the trip fares.
- Filter and organized the data to find the most important variables for the regression model.