# Exploratory Data Analysis of New York City TLC Data

**Executive summary report**
Commission Prepared by **Automatidata**

## Project Overview

The New York City Taxi and Limousine Commission (TLC) is responsible for licensing and regulating taxi cabs in the city. The TLC partnered with Automatidata to create a regression model that predicts taxi fares before a trip. For this project, the dataset needed to be analyzed, explored, cleaned, and structured before starting any modeling process.

## Key Insights

**The Problem:** The initial exploratory data analysis (EDA) on a sample of TLC data showed that some values could reduce the accuracy of fare predictions. In particular, some trips have a total_amount entered but show a trip distance of 0. These values need to be considered in the algorithm or removed.

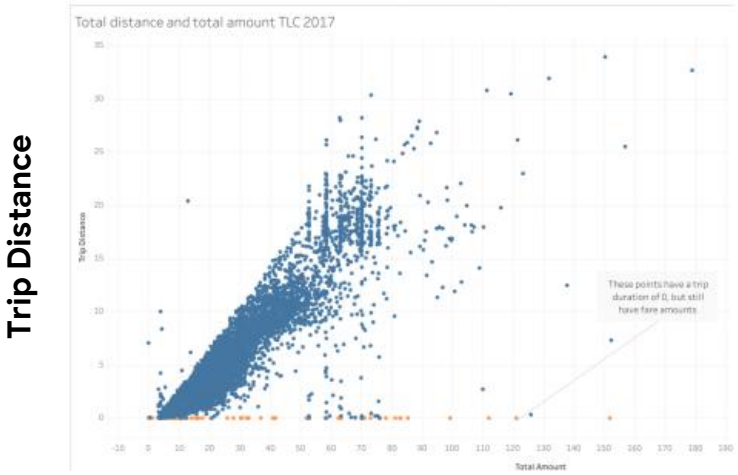
**Proposed Solution:** After analysis, we recommend removing outliers where the trip distance is 0.

**Keys to Success:**

- Ensure that the sample provided by TLC accurately reflects the full dataset.
- Create a plan for handling other outliers, such as very short trips with unusually high fares.

## Details

After completing the exploratory data analysis, the Automatidata team found that trip distance and total amount are key variables for understanding taxi cab rides. A scatter plot was created in Tableau to show the relationship between these two variables and to improve visualization.




Alt Text: Scatter plot showing New York City TLC data comparing trip_distance and total_amount.

## Next Steps

- Identify and handle outliers that could cause issues when predicting trip fares.
  - Such as trips with long durations or unusual values.
- Determine which variables have the strongest impact on taxi fares.
- Filter the data to select the most relevant variables for regression, statistical analysis, and parameter tuning.