

Exploratory Data Analysis of New York City TLC Data

Summary Report - Project 3 Automatidata - NYC TLC 2017

Project Overview

In this part of the project, the data needs to be analyzed, explored, cleaned and structured prior to any modeling. Meaningful insights and scapulatory data filtering is prepared for future modeling.

Key Insights

The Problem: Exploratory analysis identified significant data anomalies where high fares were recorded for trips with zero distance. These discrepancies and skewed ride durations pose a risk of producing biased results if the raw data is used directly for predictive modeling.

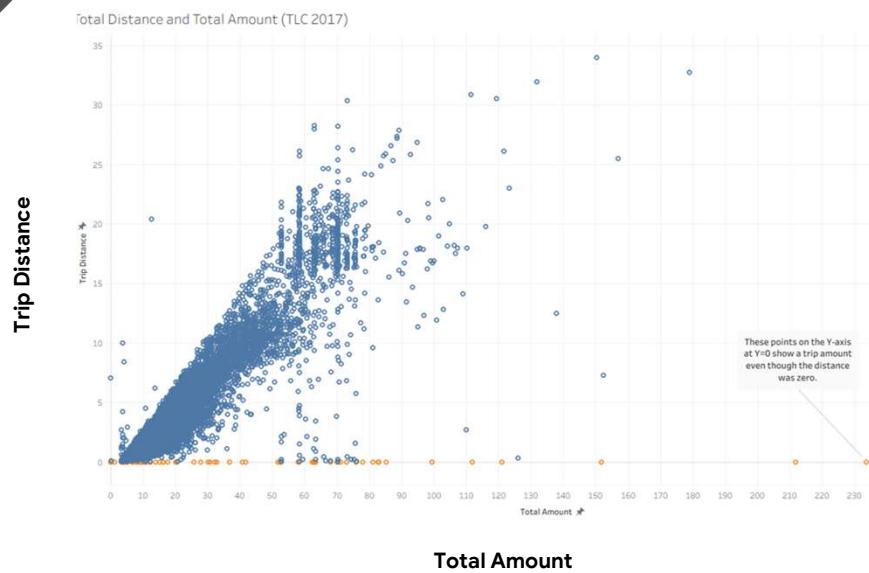
Proposed solution: By applying rigorous cleaning and visualizing these outliers could be prepared for accurate revenue forecasting and business reporting.

Keys to success

- Make sure the data source is verifiable and accurate.
- Plan for determining which data is useful, and which data is in error or is an outlier that may skew results.

Details

The scatter plot was created in Tableau to enhance the provided visualization. The data visually presented shows a meaningful, linear trend between Total Amount and Trip Distance. This relationship that should further studied to allow for insights into achieving better profit margins.



Alt Text: New York City TLC data scatter plot graph showing Trip Distance against Total Amount.

Next Steps

- Determine data points that may be in error or outliers.
- Determine what affects profitability.
- Determine which variables need to be further analyzed to present regression modeling and future forecasting.