New York City TLC Project - Exploratory Data Analysis

Executive Summary Report II

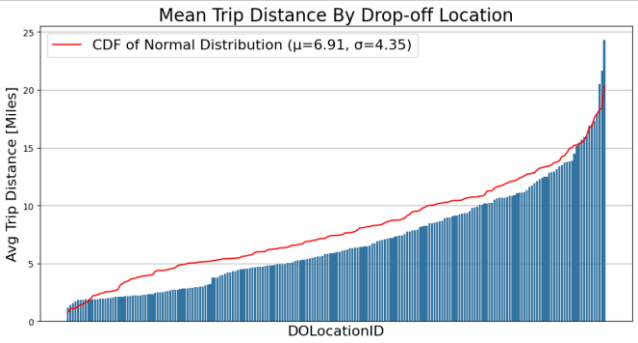
Commission Prepared by **Automatidata**

Project Overview

The NYC Taxi & Limousine Commission contracted Automatidata to develop a regression model for predicting taxi fares. Before modeling can begin, the data must be explored, cleaned and organized. Exploratory data analysis was performed to understand at an even deeper level the key variables and confirm that the data is appropriate for deriving valuable insights.

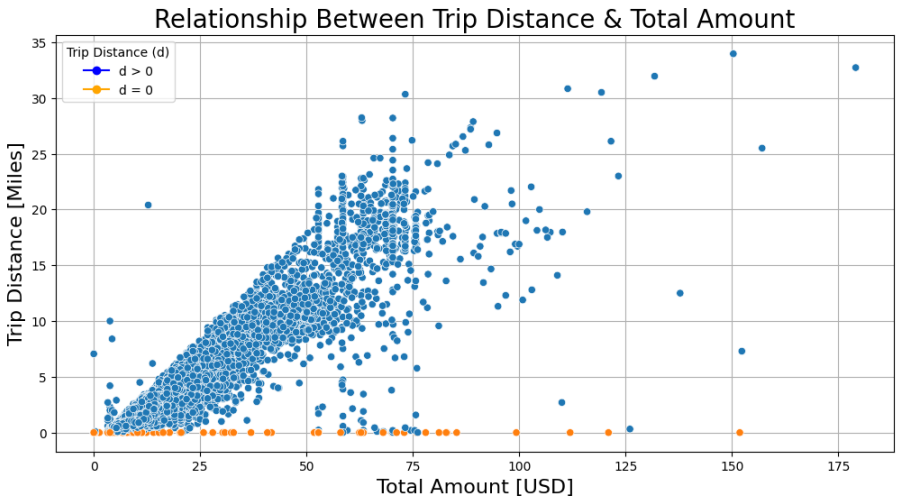
Key Insights

- Initial EDA revealed issues in the dataset that could affect fare prediction accuracy, particularly trips with a total cost but no recorded distance. These appear to be outliers and will need to be accounted for or removed. We decided to remove outliers where the trip distance is recorded as 0.
- The chart for the average distance traveled per trip for each unique destination shows a characteristic curve of a cumulative density function (CDF) for a normal distribution. This indicates that drop-off points are evenly distributed geographically, despite the lack of location data.
- After conducting EDA, the Automatidata team identified the trip distance and the total amount as key variables for modeling taxi rides.



Average distance traveled per trip for each unique destination.

Details



Connection between the distance traveled and the total fare charged.

Next Steps

1. Confirm with NYC TLC that the sample is representative of the broader dataset and plan for handling other outliers, such as trips with low distances but high costs.
2. Determine which data points could cause problems in future fare predictions (e.g. locations with longer trip durations).
3. Identify the variables that most impact trip fares and focus on the most relevant variables for regression and analysis.