New York City TLC Project EDA Summary

Executive Summary Report

Project Overview

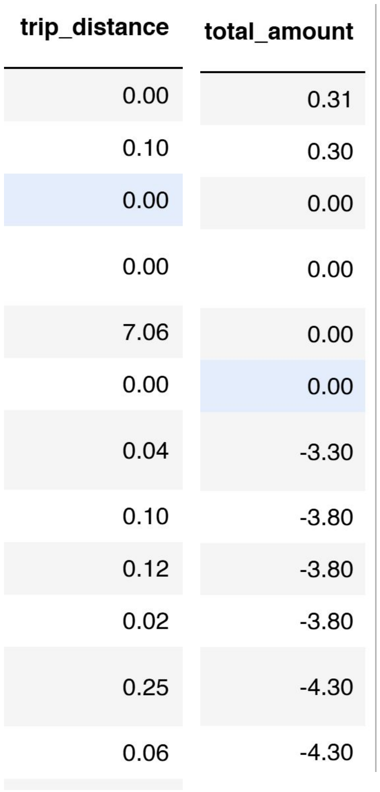
In this part of this project, the Automatidata Team initiated an exploratory data analysis on the data provided by the TLC. The goal is to inform the TLC of summary statistics of some key variables.

Key Insights

- No null values are detected in the data.
- The two most helpful columns of this data would be total_amount (the total amount of money paid by the customer) and trip_distance (the distance of each trip).
- These two variables, along with related variables, are heavily left-skewed, with zero and negative values in cost.
- Some short-distance trips have unusually high costs.

Details

These are some screenshots from the EDA step. Some values are negative or zero, which suggests a sort of imputation mistake. The related trip_distance values on its left, and some distances have zero, which again is unusual, especially a free 7-mile ride.



trip_distance	total_amount
0.00	0.31
0.10	0.30
0.00	0.00
0.00	0.00
7.06	0.00
0.00	0.00
0.04	-3.30
0.10	-3.80
0.12	-3.80
0.02	-3.80
0.25	-4.30
0.06	-4.30

Next Steps

Next steps in this project would include:

1. Do data cleaning and more analysis with unusual values.
2. Perform a more-detailed exploratory data analysis, with visualizations.
3. Do a descriptive statistical analysis on the variables of interest.
4. Conduct hypothesis testing.
5. Establish the final regression model.