New York City TLC Project Preliminary Data Summary

Executive summary report
Commision Prepared by Automatidata

Overview

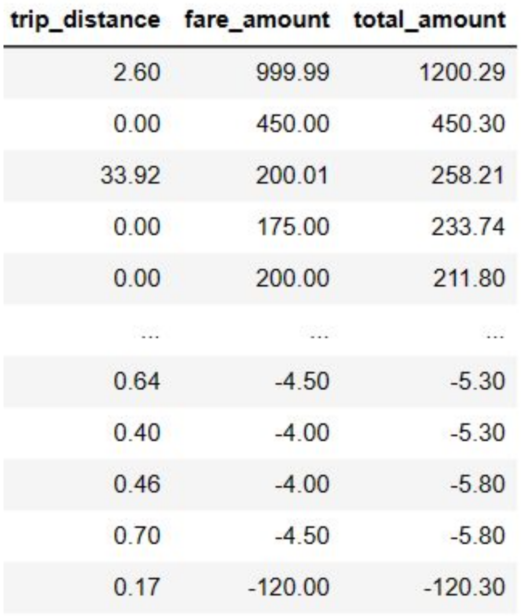
The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab fares. In this part of the project, the Automatidata data team performed a preliminary inspection of the data supplied by the NYC Taxi & Limousine Commission in order to inform the team of key data variable descriptions, and ensure the information provided is suitable for generating clear and meaningful insights.

Objective

- Explore dataset to find any unusual values.
- Consider which variables are most useful to build predictive models.
- Consider potential interactions between the two chosen variables.
- Examine which components of the provided data will provide relevant insights.
- Build the groundwork for future exploratory data analysis, visualization, and models.

Results

- This dataset includes variables that should be helpful for building prediction models on taxi cab ride fares.
- The identified unusual values are some negative values for fare_amount and total_amount, as shown on the screenshot on the right:
- Another unusual values are short distance trips that have high charges associated with them, as shown in the total_amount variable. Reference screenshot on the right:



trip_distance	fare_amount	total_amount
2.60	999.99	1200.29
0.00	450.00	450.30
33.92	200.01	258.21
0.00	175.00	233.74
0.00	200.00	211.80
...	...	...
0.64	-4.50	-5.30
0.40	-4.00	-5.30
0.46	-4.00	-5.80
0.70	-4.50	-5.80
0.17	-120.00	-120.30

Next Steps

1. Conduct a complete exploratory data analysis.
2. Perform any data cleaning and data analysis steps to understand unusual variables.
3. Use descriptive statistics to learn more about the data.
4. Create and run a regression model.