New York City TLC Project Preliminary Data Summary

Executive summary report

Prepared by **Automatidata**

OVERVIEW

The NYC Taxi & Limousine Commission has partnered with Automatidata to create a regression model capable of forecasting taxi fares. In this phase of the project, the Automatidata data team carried out a preliminary assessment of the dataset supplied by the Commission to document key variable characteristics and verify that the data is suitable for generating reliable analytical insights.

PROJECT STATUS

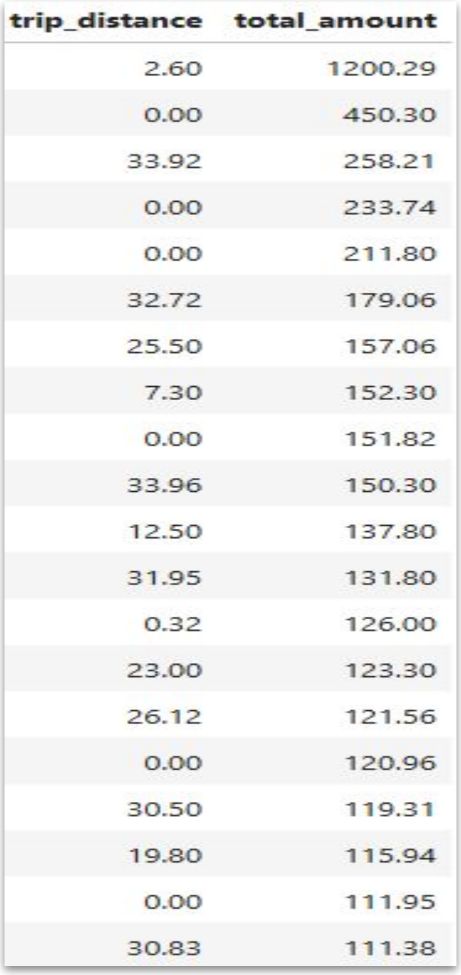
- Explored the dataset to identify any unexpected values.
- Evaluated which variables would be most effective for predictive modeling — specifically **total_amount** and **trip_distance**, as they jointly represent the core elements of a taxi ride.
- Assessed potential interactions between these key variables.
- Determined which parts of the dataset would yield meaningful insights.
- Established the foundation for deeper exploratory analysis, visualizations, and future modeling work.

NEXT STEPS

1. Complete an end-to-end exploratory data analysis.
2. Clean the data and examine any irregular or outlier values.
3. Use descriptive statistics to understand key patterns.
4. Build and run a regression model.

KEY INSIGHTS

- The dataset includes variables useful for fare-prediction modeling of taxi rides.
- The main irregularities observed are short trips with unexpectedly high fares associated as shown in the screenshot below.



trip_distance	total_amount
2.60	1200.29
0.00	450.30
33.92	258.21
0.00	233.74
0.00	211.80
32.72	179.06
25.50	157.06
7.30	152.30
0.00	151.82
33.96	150.30
12.50	137.80
31.95	131.80
0.32	126.00
23.00	123.30
26.12	121.56
0.00	120.96
30.50	119.31
19.80	115.94
0.00	111.95
30.83	111.38