New York City TLC Project Preliminary Data Summary

Executive summary report

Prepared by **Automatidata**

OVERVIEW

The New York City taxi and Limousine Commission (TLC) has approached Automatidata to develop an app that enables TLC rides to estimate ride fares in advance. The project task is to develop an ML model that can do this prediction. In this step, the Automatidata data analyst team performed a preliminary inspection of the data supplied by the the TLC in order to get basic insights into dataset provided. The team looks for anamolies in the dataset and check for the features that would be useful.

PROJECT STATUS

- Explored dataset to find any anomalies and errors.
- Considered which variables are most useful to build predictive models. (here, we can choose number of passengers, trip distance and total_fare).
- Considered potential relationship between the chosen variables.
- Prepared for further exploratory data analysis of the data.

NEXT STEPS

1. Perform a full exploratory data analysis.
2. Remove any unusual data rows such as negative total fare, Zero trip distance and extremely high fare for very small trip.
3. Perform a descriptive statistics on data to learn more about the data.
4. Choose the important features and run a regression model.

KEY INSIGHTS

- The dataset contains columns with useful information such as date and time of trip, trip distance, Location, number of passengers and ride fare.
- This data has some anomalies. We see negative total_fare amount, 0 trip_distance and also very high fare amount for 0 or small trip distance.

total_amount
-120.3
-5.8
-5.8
-5.3
-5.3
-4.8
-4.8

trip_distance	fare_amount
2.60	999.99
0.00	450.00
33.92	200.01
0.00	175.00
0.00	200.00
32.72	107.00
25.50	140.00

These outlier and error values need to removed from the dataset before moving forward.