New York City TLC Project Preliminary Data Summary

Executive summary report
Commission Prepared by **Automatidata**

OVERVIEW

Our Customer, the NYC Taxi & Limousine Commission has hired our company, Automatidata to build a regression model to predict taxi cab fares. We have completed the preliminary inspection of the data supplied by the customer. We have detected the key variables and their description. In this phase we confirm that the information provided is sufficient to proceed with the model and create the necessary insights..

PROJECT STATUS

- Data exploration, finding unusual values.
- Defining the important variables to build predictive models (**total_amount** and **trip_distance**, which work together to depict a taxi cab ride).
- Considered interdependencies between the two chosen variables.
- Examined which data components will provide meaningful insight
- Built the groundwork for next steps, future exploratory data analysis, visualizations, and models.

NEXT STEPS

1. Conduct Exploratory data analysis.
2. Perform necessary data cleaning and data analysis steps to understand unusual variables (e.g. outliers).
3. Apply descriptive statistics to learn more about the data.
4. Create and test the regression model

KEY INSIGHTS

- We have identified the variables that should be helpful for building prediction model(s) on taxi cab ride fares.
- We also identified unusual values on these variables, trips that are a short distance but have high charges associated with them, as shown in the total_amount variable:

Total_amount variable

```
df_sort = df.sort_values(by=['fare_amount'],ascending=False)
df_selected = df_sort[['fare_amount', 'trip_distance', 'total_amount']]
df_selected.head(10)
```

	fare_amount	trip_distance	total_amount
8476	999.99	2.60	1200.29
20312	450.00	0.00	450.30
13861	200.01	33.92	258.21
15474	200.00	0.00	211.80
12511	175.00	0.00	233.74
3582	152.00	7.30	152.30
9280	150.00	33.96	150.30
16379	140.00	25.50	157.06
10291	131.00	31.95	131.80
11269	120.00	0.00	151.82

The total_amount variable indicates the necessity of further analyzing outlier variables.