Title   Investigate, Understand, and Organize New York TLC Dataset Project

Overview

This project's purpose is to investigate and understand the data provided from the New York TLC construct by our Automatidata team. Having the general and concise understanding of the dataset will allow us to continue on further steps in developing the regression model.

Objective

The goal of the project is to inspect the dataframe from the given dataset by using Python, analyze and interpret the variables, and present the key insights/information.

Results

- There is a consideration of the appearance of **negative** values, which come from variables (fare amount, tip amount, total amount). These negatives represent for the minimum values.
- Furthermore, the maximum values for these variables and others tend to be **higher** than the normal range. For example, the maximum trip distance is ~33 miles, compared to the range of 25% of 1 mile and 75% of 3 miles. Specifically, the majority trip distances aren't too long, but there are some longer ones, as well as the majority rides' amount aren't much (below $20), but there are amounts way higher than the normal.
- **Credit card** and **cash** are the most two payment types (15265 and 7267 respectively), which credit card accounts more than 2 times. The average credit card tip is ~$2.73, and $0.00 is for cash tip.
- **One passenger** accounts for the majority credit card payment, which accounts for ~11,000. Average tip that one passenger gave is ~$2.71, which ranges the same either for two, three, or four passengers.
- **Verifone Inc.** accounts for more than 2,000 vendor's appearance than the **Creative Mobile Technologies, LLC** (12626 and 10073 respectively).
- The **trip_distance** and **total_amount** might be most helpful variables for building the regression model in future steps. We might account for other variables if necessary.

Next Steps

- Our Automatidata team needs to further analyze on the outliers and anomalies values appeared in the dataset.
- Also further investigate on variables to deeply understand and fine-tune the dataframe.
- Prepare for Exploratory Data Analysis (EDA) process.
- Perform hypothesis testing and statistical modeling.