Regression Model for New York City TLC Project

Executive Summary Report

OVERVIEW

The purpose of this project is to create a multiple linear regression model with several different variables that seem to be related with taxi ride fare. This model will be used to explain which variables are most closely related to ride fare, so that future ride fares can be predicted.

PROJECT STATUS

Exploratory data analysis was performed on NYC taxi data.

- The variables that may impact ride fare amounts the most seem to be average ride distance, average ride duration, passenger count, rush hour, and vendor.
- Outlier data points were examined and removed to prepare the data for modeling.
- The above variables were included in the regression model to ensure accurate prediction of ride fare.

NEXT STEPS

Based on the results of the model from this notebook, it would be recommended to:

- Incentivize customers to ride taxis for longer distances.
- Further research ride distance and ride duration to see if their relationships with fare amount remain consistent based on other variables, such as pick-up and drop-off location.

KEY INSIGHTS

The multiple linear regression model indicates that the variables included in the model explain approximately 86.83% of the variation of taxi fare amount.

Additionally, the model indicates that:

- The strongest predictor of taxi fare amount is the average ride distance,with every 1 mile increase being associated with a fare amount increase of approximately $7.13.
- The other strong predictor of taxi fare amount is the average ride duration (with every 1 minute increase being associated with a fare amount increase of approximately 2.81 dollars.)
- The other variables (passenger count, rush hour, and vendors) were not strong predictors of taxi fare amount.