# NYC TLC Fare Prediction Project

Multiple Linear Regression Phase

## Overview

Goal of the project is to provide the NYC TLC with a reliable way to predict the fare of a trip before it happens, using data collected from past trips. The regression model constructed in this phase supplies our client with the main deliverable: a multiple linear regression (MLR) model.

## Objective

This primary objective of this phase of the project was to build, train, test and evaluate a multiple linear regression model, using a set of features from the dataset to predict the target variable, fare_amount.

## Results

- Of the variables from the dataset and from those created for analysis, ride distance has the highest correlation with fare amount, where for every 2 miles traversed fare amount increases by $7.13.
- Using this model, the coefficient of determination was 0.87 for the testing data, meaning the model can account for 87 percent of the variation in the target variable fare amount.
- Evaluation metrics for test data
  - Mean Absolute Error: 2.12
  - Mean Squared Error: 14.25
  - Root Mean Squared Error: 3.77

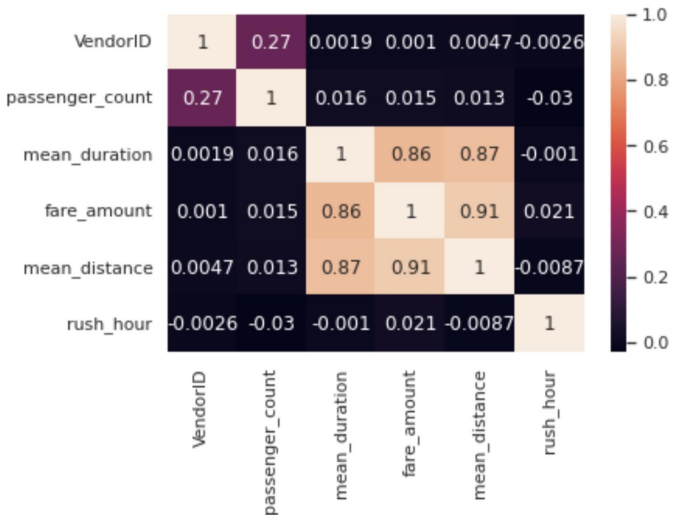


Fig. 1: Heatmap of outcome variable and feature correlations

## Next Steps

Consider implementation of the model into a mobile application that forecasts fare amount when ride distance is calculated based on user provided destination.