# Executive Summary of Regression Model Building for New York TLC Data

Commission Prepared by Automatidata

## Project Overview

The NYC Taxi & Limousine Commission has consulted with Automatidata to build a regression model to predict taxi cab fares. The Automatidata team built a multiple linear regression model with a 87% accuracy score on the test data.

## Key Insights

EDA

- Outliers were imputed
- Values for the JFK airport had a set fare amount for the year the data was collected

Preprocessing

- Irrelevant features were removed
- New continuous features added

Modeling

- The model predict 87% of the variance of the target variable for the test set
- The absolute mean error for the test set was $2, inferring that the predicted values were off by $2 from the actual values.
- Mean_distance had the greatest effect of fare_amount

## Details

EDA and Preprocessing

- The dataset was cleaned with missing values removed and outliers imputed.
- New columns were added (mean_distance, mean_duration, rush_hour) from transformation applied the dataset.
- Irrelevant columns and columns that would cause data leakage were removed.

Modeling

- A train and test set were created were 80% of the dataset was used for training
- Categorical variables were encoded and continuous values were standardized.
- Visualizations of data relationships were realized.

## Next Steps

- Encourage taxi cab drivers to prioritize longer trips to maximize earnings.