

Exploratory Data Analysis of New York TLC Data

Executive summary report

Commission Prepared by **Automatidata**

OVERVIEW

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxicab ride fares. In this part of the project, the data needs to be analysed, explored, cleaned and structured prior to any modelling.

PROJECT STATUS

- Project Proposal has been approved.
- Data is collected from New York TLC.
- Data is inspected and organized.

NEXT STEPS

- Handle outliers in data.
- Eliminate the errors that could pose a problem for future analysis in predicting trip fares.
- Determine the variables that have the largest impact on trip fares.
- Build regression model based on relationships found.

KEY INSIGHTS

- Total amount is related to distance and time duration both.
- For more than 6 passengers the total amount takes a sudden spike.
- Some rides have a large total amount even though the trip distance is 0.
- The dataset contains a lot of outliers and error values like negative total amount and trip duration.
- Sundays record the most no. of rides and thus highest total revenue and Wednesday the least.
- The mean trip distance is normally distributed for the Location IDs.
- Also, some locations are highly popular while some are rarely visited by riders.

