# TLC Fare Prediction Project

Executive summary for Automatidata team on data variable assessment for TLC data

## Overview

The recently submitted New York City TLC project proposal has been approved. The Automatidata team now has access to the New York City TLC data to analyze, identify key variables, and prepare for exploratory data analysis.

## Objective

- Load data, explore, and extract the New York City TLC data with Python
- Use custom functions to organize the information within the New York City TLC dataset
- Build a dataframe for the New York City TLC project

## Results

- Used a Jupyter notebook.
- Read data from the provided 2017_Yellow_Taxi_Trip_Data.csv file into a dataframe.
- Conducted exploration of the dataframe.
  - Used methods such as head(), info(), describe() etc.
  - Created dataframe copies based on sorting of specific columns in descending order.
  - Noticed possible anomalies and outliers.
    - For example, minimum fare_amount is -120.
    - 33 rows have passenger_count = 0.
  - Noted the average values for numeric columns.
- Created a new column for the duration of taxi trips.
  - trip_duration calculated subtraction of pickup time from dropoff time.

## Next Steps

- Further investigation on the data may be needed.
  - Outliers, negative values, columns with zero values for specific trips, trip duration anomalies.
- Further data cleaning is needed.