



# Executive Summary

## Title: New York City Taxi Fare Prediction


### Project Overview

The New York City Taxi and Limousine Commission (TLC) is leveraging data collected from the NYC area to predict taxi fare amounts accurately. To achieve this, the agency has partnered with Automatidata to develop a robust regression model that estimates taxi fares prior to the ride. As a preliminary step, we are constructing a comprehensive DataFrame that will serve as the foundation for this predictive model.

### Data Quality Check

- **Correctness and Coherence:** The dataset, created for pedagogical purposes, effectively represents real-world entities and events. It includes key attributes such as Vendor ID, pick-up and drop-off times, distance traveled, fare amount, payment mode, and tip given. The data types are appropriately aligned with each attribute, and the dataset is logically structured, with each column building on the others to provide a coherent view of the taxi trips.
- **Completeness and Accountability:** For the task of predicting taxi fare amounts, the dataset is comprehensive, containing essential transactional information like fare amount, trip distance, time, and payment details, including tips. The data was generated by Google to serve as a foundational resource for the Google Advanced Analytics Certificate, ensuring both completeness and accountability.

### Understanding the Dataset

1. **Trip Distance:** The values for trip distance appear reasonable, with a maximum of 33.96 miles and a minimum of 0 miles. Considering that the longest distance across New York City is approximately 30 miles, these values are consistent with expected travel distances within the city.
  2. **Total Fare Amount:** While most values are plausible, there are some anomalies, such as negative and zero values, which may indicate errors or trips that were terminated prematurely. Additionally, the top two fare amounts are significantly higher than the rest, suggesting potential outliers.
  3. **Long Trips vs. High Fares:** There is no direct correlation between trip length and fare amount. The most expensive trips are not necessarily the longest, suggesting that factors other than distance, such as time of day or traffic conditions, may influence fare amounts.
- 



## Distribution Analysis

- **Vendors:** The dataset records trips from two vendors: Creative Mobile Technologies LLC (44.4%) and VeriFone Inc. (55.6%), indicating a near-even distribution.
- **Payment Modes:** Among the six recorded payment modes, credit card is the most prevalent, accounting for approximately 67% of transactions, followed by cash at 32%.
- **Tips:** No tips were recorded for cash payments, while the average tip for credit card transactions is \$2.07, suggesting that credit card users tend to be more generous.

## Summary and Next Steps

Initial analysis indicates that **total\_amount** and **trip\_distance** are the primary variables influencing taxi fare predictions, as they encapsulate the key characteristics of a taxi ride. The next step will involve conducting a comprehensive Exploratory Data Analysis (EDA) to uncover deeper insights and refine the predictive model.

