**Executive Summary**

**Name of the Project:**

**NYC Taxi Fare Data Preparation and Analysis**

**Project Overview**

As part of the Google Advanced Data Analytics Certificate program, I contributed to a fictional consulting firm, Automatidata, in preparing a dataset from the New York City Taxi and Limousine Commission (TLC). The goal was to inspect, clean, and prepare the data for future predictive modeling, specifically to estimate taxi fares before a trip begins. This required understanding the dataset structure, identifying data issues, and recommending appropriate next steps for analysis.

**Details**

The dataset contains **22,699 taxi ride records** with **18 variables**, including pickup/dropoff timestamps, trip distances, fare amounts, tips, and payment methods. Using Python (Pandas and NumPy), I performed the following key tasks:

* Imported and explored the dataset using .head(), .info(), and .describe().
* Verified there were no missing values across any columns.
* Identified and sorted anomalous records with zero-distance rides and high or negative fares.
* Investigated variable distributions and outliers, particularly in fare\_amount, tip\_amount, and total\_amount.
* Grouped and analyzed payment types and tip behavior.
* Filtered the dataset to assess credit card transactions and their relation to passenger count.

**Key Insights**

* ✅ **No Missing Data**: Every record in the dataset is complete with no null values.
* 🚩 **Anomalies Detected**:
  + Negative and extreme values in fare\_amount (e.g., –$120) and total\_amount (up to $1200).
  + Several trips reported **0.00 miles** yet had high total charges.
* 💳 **Tip Recording Bias**: Tips are only recorded for **credit card** transactions, so tip amounts are always $0 for cash trips.
* ❌ **Invalid RatecodeID Values**: Codes like **99** appear, even though valid codes are 1–6.
* ⏱️ **Datetime Fields**: Stored as strings (object) and must be converted to datetime type for time-based insights.
* 🔍 **Top Predictors for Modeling**:
  + trip\_distance: Strong positive relationship with fare.
  + payment\_type: Affects tip recording and total amount patterns.

**Next Steps**

* Clean and remove or flag negative or unrealistic fare and total values.
* Convert tpep\_pickup\_datetime and tpep\_dropoff\_datetime into datetime format.
* Correct or exclude invalid RatecodeID values.
* Engineer new features such as **hour of day**, **day of week**, or **trip duration** for future modeling.
* Consider integrating **external data sources** such as traffic, weather, or location zoning for improved prediction accuracy.