**Scenario:**

InteliData, a data consulting firm, partners with clients to transform unused and stored data into actionable insights. They specialize in data-driven solutions such as performance dashboards, customer-facing tools, and strategic business insights, catering to a range of industries by understanding and addressing their unique business needs.

**Client**:

The New York City Taxi and Limousine Commission (TLC), which regulates and licenses taxi cabs and for-hire vehicles, has approached InteliData to develop a machine learning model to estimate taxi fares before rides. With over 200,000 licensees and approximately one million trips made each day, TLC possesses a massive amount of trip data that can be leveraged for this task.

**Problem Statement:**

TLC aims to provide taxi fare estimates to passengers before their rides begin, enhancing customer experience and transparency. InteliData’s goal is to develop a **regression model** using TLC’s vast data repository to accurately predict fare prices based on multiple factors.

**Answer the question given below and upload this file and your code to repository given by us.**

**Dataset overview:**

| **Column name** | **Description** |
| --- | --- |
| ID | Trip identification number |
| VendorID | A code indicating the TPEP provider that provided the record.  **1= Creative Mobile Technologies, LLC;**  **2= VeriFone Inc.** |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| Passenger\_count | The number of passengers in the vehicle.  This is a driver-entered value. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone in which the taximeter was engaged |
| DOLocationID | TLC Taxi Zone in which the taximeter was disengaged |
| RateCodeID | The final rate code in effect at the end of the trip.  **1= Standard rate**  **2=JFK**  **3=Newark**  **4=Nassau or Westchester**  **5=Negotiated fare**  **6=Group ride** |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before being sent to the vendor, aka “store and forward,”  because the vehicle did not have a connection to the server.  **Y= store and forward trip**  **N= not a store and forward trip** |
| Payment\_type | A numeric code signifying how the passenger paid for the trip.  **1= Credit card**  **2= Cash**  **3= No charge**  **4= Dispute**  **5= Unknown**  **6= Voided trip** |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| MTA\_tax | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The  improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in trip. |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |

**Task to be performed:**

1. **Understand the data**

* Create a pandas dataframe for data learning, exploratory data analysis (EDA), and statistical activities.
  + **Question 1:** When reviewing the df.info() output, what do you notice about the different variables? Are there any null values? Are all of the variables numeric? Does anything else stand out?
    - * **Answer: Dataset has 18 columns and 22,699 entries and there are no NULL values in the dataset. Most variable are numeric except object-type columns representing dates and categories.**
  + **Question 2:** When reviewing the df.describe() output, what do you notice about the distributions of each variable? Are there any questionable values?
    - * **Answer: The data contain negative values in fare\_amount, mta\_tax, and total\_amount, 0 in trip\_distance and passenger\_count and extreme outliers in fare\_amount, total\_amount and tip\_amount. RatecodeID max have 99 invalid entries**
* Write a compiled summary information about the data to inform next steps.
  + - * **Answer: 1)Some columns might have null values.**
      * **2) columns like tpep\_pickup\_datetime and tpep\_dropoff\_datetime are non-numeric and need to be converted.**
      * **3) columns such as fare\_amount and Trip\_distance might have outliners**

1. Understand the variables

* Use insights from your examination of the summary data to guide deeper investigation into specific variables.
  + Sort and interpret the data table for two variables: trip\_distance and total\_amount. **Answer the following three questions:**
    - **Question 1:** Sort your first variable (trip\_distance) from maximum to minimum value, do the values seem normal?
      * **Answer: maximum values range 27.97 to 33.96**
    - **Question 2:** Sort by your second variable (total\_amount), are any values unusual?
      * **Answer: when sorting by total\_amount values are within a reasonable range, but there are unusually high values. These outliner are likely errors or special cases**
    - **Question 3:** Are the resulting rows similar for both sorts? Why or why not?
      * **Answer: the resulting rows for sorting by trip\_distance and total\_amount are unlikely to be identical this is because total\_amount is depends on many other factors not just trip\_distance.**

1. Develop a machine learning (regression) model
   * What is the error in prediction?
     + - **Answer: Mean Squared Error (MSE)=0.44**
       - **Mean Absolute Error(MAE)=0.12**
       - **Root mean squared error (RMSE)=0.66**
   * What is the percentage of accuracy in prediction?
     + - **Answer: 1.00**