**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:**
      * **Data types are : float64(8), int64(7), object(3)**
      * **No there is no Null Values**
      * **The dataset contains 22,699 records, with all the variables populated, as there are no null values in the data.**
      * **The Unnamed: 0 column seems to be an index or identifier, which should likely be dropped as it doesn't provide any useful information for analysis.**
      * **The VendorID, RatecodeID, PULocationID, DOLocationID, and Payment\_type columns are categorical in nature, with numerical encoding, while others like Trip\_distance, Fare\_amount, Tip\_amount, Tolls\_amount, etc., are continuous numerical variables.**
  + **Question 2:** When reviewing the df.describe() output, what do you notice about the distributions of each variable? Are there any questionable values?
    - * **Answer:**
      * **I notice that trip distance min is 0 which is likely invalid**
      * **I fare amount also the min is around -120 which is in negative which detect to anomaly**
      * **The tip values show a lot of variation, with an average of 1.83, indicating some extreme outliers in tip amounts.**
      * **There is some extra charge in negative which is also a thing because it should in positive**
      * **The total fare reaches 999.99, which seems unusually high and could be an outlier or error in the data.**
* Write a compiled summary information about the data to inform next steps.
  + - * **Answer:**
      * **There are 22699 records**
      * **18 columns of different categories**
      * **1. Numeric data : like trip distance , fare amount and total amount**
      * **2. Integer data : like number of passenger ,payment type and some location IDs**
      * **3. Text data : Like pick and drop time and flags**
      * **There are no missing values**
      * **There is problem in Unnamed Column , Negative value in Fare amount , Zero trip Distance , Outliers in Total Amount**

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:**
      * **After sorting trip distance from max to min value seems normal**
      * **This head(10) values :**
      * trip distance
      * 9280 33.96
      * 13861 33.92
      * 6064 32.72
      * 10291 31.95
      * 29 30.83
      * 18130 30.50
      * 5792 30.33
      * 15350 28.23
      * 10302 28.20
      * 2592 27.97
    - **But Some Min trip distance is like 0.0 could indicate erroneous records or trips where no actual movement occurred, possibly due to technical issues or cancelled rides.**
    - This is tail()

trip\_distance

20042 0.0

4910 0.0

3865 0.0

5630 0.0

* + - **Question 2:** Sort by your second variable (total\_amount), are any values unusual?
      * **Answer:**
      * **Yes, are unusual data in total amount which Negative amounts which leads to anomaly**
      * total amount
      * 8476 1200.29
      * 20312 450.30
      * 13861 258.21
      * 12511 233.74
      * 15474 211.80
      * ... ...
      * 11204 -5.30
      * 14714 -5.30
      * 20698 -5.80
      * 17602 -5.80
      * 12944 -120.30
    - Here we can see negative amount goes up to -120
    - **Question 3:** Are the resulting rows similar for both sorts? Why or why not?
      * **Answer:** **The sorted rows for trip distance and total amount may not be directly comparable because the total fare is influenced by multiple factors such as distance, tolls, time, and rate codes, whereas the trip distance is just the physical distance of the trip. Thus, even if a trip has a high distance, the fare may not be correspondingly high if other factors (like rate codes or payment type) are different.**

1. Develop a machine learning (regression) model
   * What is the error in prediction?
     + - **Answer:**
       - **Error Metrics:**
       - **Mean Absolute Error (MAE): 0.013949490923875861**
       - **Mean Squared Error (MSE): 0.027297496227734582**
       - **Root Mean Squared Error (RMSE): 0.16521953948529994**
       - **There is error but very least**
   * What is the percentage of accuracy in prediction?
     + - **Answer: Accuracy Percentage: 99.99435882430187%**