# Week 2: In-Class Assignment 1

### **Dataset: taxi\_data.csv**

## Data Information:

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| vendorid | BIGINT |
| tpep\_pickup\_datetime | STRING |
| tpep\_dropoff\_datetime | STRING |
| store\_and\_fwd\_flag | STRING |
| ratecodeid | DOUBLE |
| pulocationid | BIGINT |
| dolocationid | BIGINT |
| passenger\_count | BIGINT |
| trip\_distance | DOUBLE |
| fare\_amount | DOUBLE |
| extra | DOUBLE |
| mta\_tax | DOUBLE |
| tip\_amount | DOUBLE |
| tolls\_amount | DOUBLE |
| improvement\_surcharge | DOUBLE |
| total\_amount | DOUBLE |
| payment\_type | BIGINT |

## Queries:

#### **Q.1 Create a database and create a table or external table named taxi\_data.**

* **Task**:
  + Write a query to create a new database.
  + Create a table or an external table named taxi\_data within that database.

#### **Q.2 Load the data from the provided CSV file.**

* **Task**:
  + Load the taxi\_data.csv file into the table created.
  + Ensure that all the data types align with the table schema.
  + Optionally, you can load the file using the Hue portal.

#### **Q.3 What is the total number of trips (equal to the number of rows)?**

* **Task**:
  + Write a query to count the total number of trips recorded in the dataset.

#### **Q.4 What is the total revenue generated by all the trips?**

* **Task**:
  + Calculate the total revenue from all trips using the total\_amount column.

#### **Q.5 What fraction of the total revenue is paid for tolls?**

* **Task**:
  + Calculate the fraction of the total revenue that is allocated to tolls using the tolls\_amount column.

#### **Q.6 What fraction of the total revenue is driver tips?**

* **Task**:
  + Calculate the fraction of the total revenue that comes from driver tips using the tip\_amount column.

#### **Q.7 For each payment type, display the following details:**

* **Task**:
  + a. **Average fare generated**: Calculate the average fare using the fare\_amount column.
  + b. **Average tip**: Calculate the average tip.
  + c. **Average tax**: Calculate the average tax using the mta\_tax column.

#### **Q.8 On average, which hour of the day generates the highest revenue?**

* **Task**:
  + Determine the hour of the day (extracted from tpep\_pickup\_datetime) that, on average, generates the highest revenue.

#### **Q.9 What is the correlation between the number of passengers on any given trip and the tip paid per trip?**

* **Task**:
  + Analyze whether multiple passengers on a trip tend to tip more compared to solo travelers using the CORR(Col\_1, Col\_2) function.

#### **Q.10 Segregate the data into five segments of tip paid: [0-5), [5-10), [10-15), [15-20) and >=20.**

* **Task**:
  + Calculate the percentage share of trips that fall into each of these tip segments.