Titanic Data Cleaning & SQL Analysis

1. Data Cleaning in Python

- Load the Titanic dataset using pandas.
- Fill missing 'Age' values with the median age.
- Fill missing 'Embarked' values with the most common port.
- Drop the 'Cabin' column (too many missing values).
- Create 'FamilySize' = SibSp + Parch + 1.
- Convert 'Sex' into numerical (0 = Female, 1 = Male).
- Convert 'Embarked' into numerical (C=0, Q=1, S=2).
- Save the cleaned dataset as 'titanic_cleaned.csv'.

2. Storing Data in MySQL

- Create a new MySQL database (e.g., 'titanic_db').
- Establish a connection using SQLAlchemy.
- Insert the cleaned data into a MySQL table.
- Verify the data insertion with basic SQL queries.

3. SQL Queries for Analysis

Check Total Rows

SELECT COUNT(*) FROM passengers;

View First 5 Records

SELECT * FROM passengers LIMIT 5;

Check Column Names & Data Types

DESCRIBE passengers;

Total Number of Survivors & Non-Survivors

SELECT Survived, COUNT(*) FROM passengers GROUP BY Survived;

Average Age of Survivors vs. Non-Survivors

SELECT Survived, AVG(Age) FROM passengers GROUP BY Survived;

Most Common Embarkation Port

SELECT Embarked, COUNT(*) FROM passengers GROUP BY Embarked ORDER BY COUNT(*) DESC;

Top 5 Passengers Who Paid the Highest Fare

```
SELECT Name, Fare FROM passengers ORDER BY Fare DESC LIMIT 5;
```