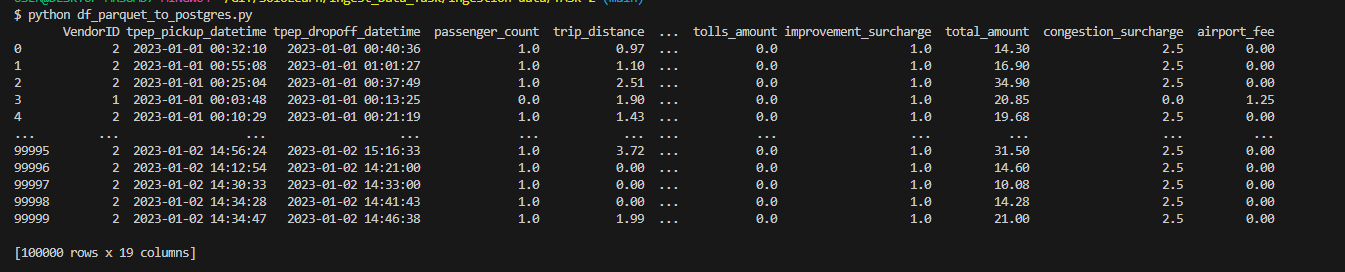
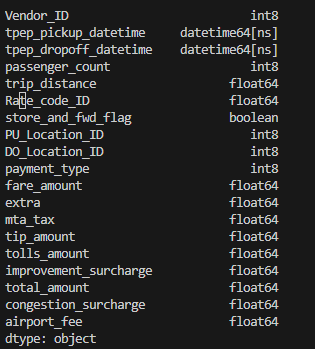
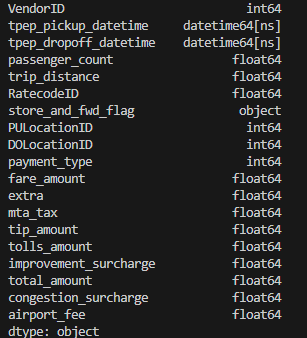
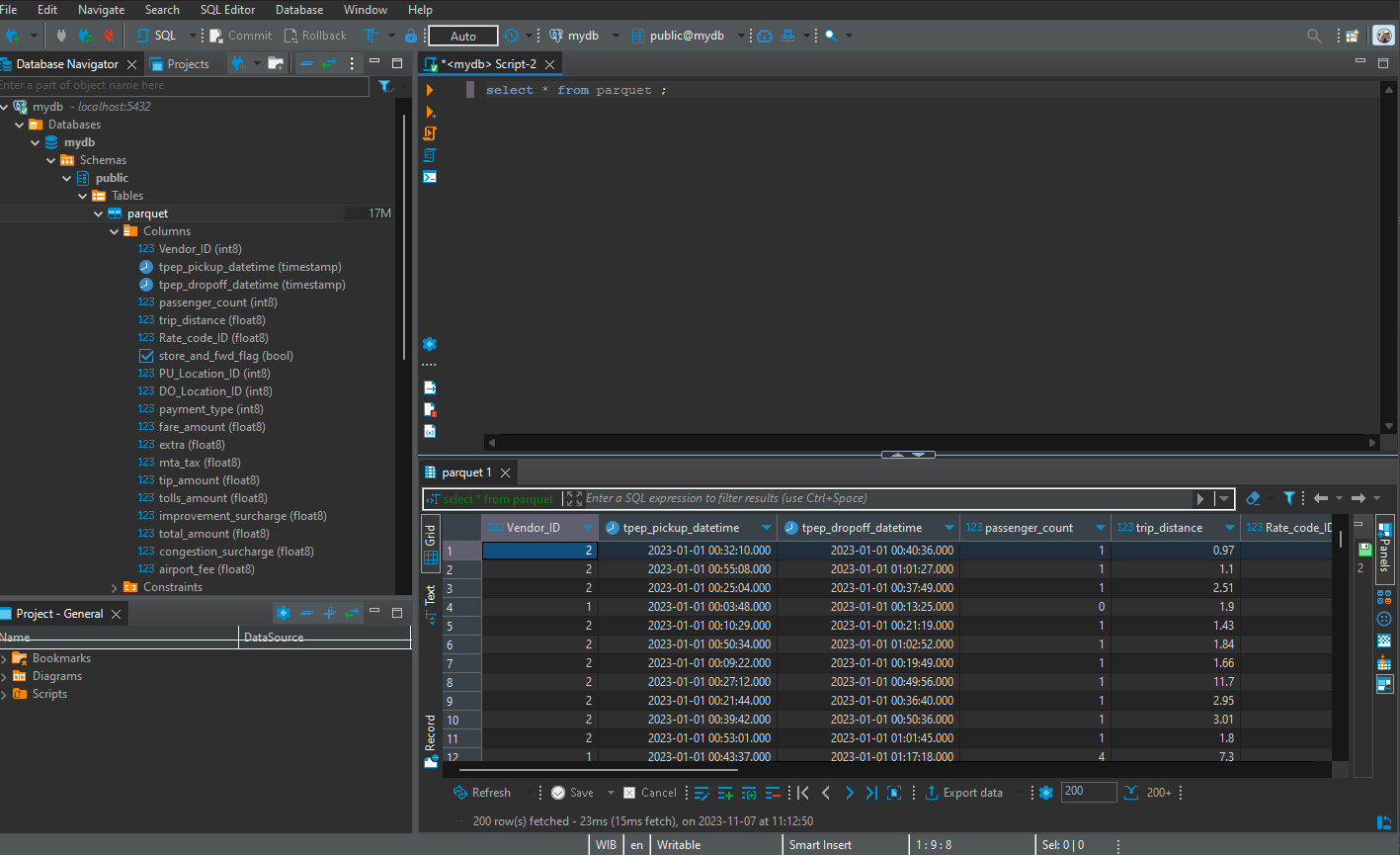
Nurul Taufik Kurrahman

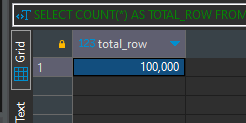
**Task 2**

1. We are going to create a DataFrame from a parquet file on our datasets. (file yellow\_tripdata\_2023-01.parquet) (mirip dengan file test.py tapi ganti format ke parquet dan tambah/import dengan library sesuai nomor selanjutnya (nomor 1-5))
2. [](https://github.com/ALTA-DE1-CHARIS-14101996/ingestion-data/blob/main/TASK-2/image-1.png)Load the parquet file to a DataFrame with fastparquet library.
3. Clean the Yellow Trip dataset.

[](https://github.com/ALTA-DE1-CHARIS-14101996/ingestion-data/blob/main/TASK-2/image-4.png)[](https://github.com/ALTA-DE1-CHARIS-14101996/ingestion-data/blob/main/TASK-2/image-2.png)\*Before Cleaning Dataset \*After Cleaning Dataset

1. Define the data type schema when using to\_sql method.
2. [](https://github.com/ALTA-DE1-CHARIS-14101996/ingestion-data/blob/main/TASK-2/image-14.png)Ingest the Yellow Trip dataset to PostgreSQL
3. Count how many rows are ingested. (masuk ke dbeaver use count(SQL))

SELECT COUNT(\*) AS TOTAL\_ROW FROM parquet ;

[](https://github.com/ALTA-DE1-CHARIS-14101996/ingestion-data/blob/main/TASK-2/image-15.png)