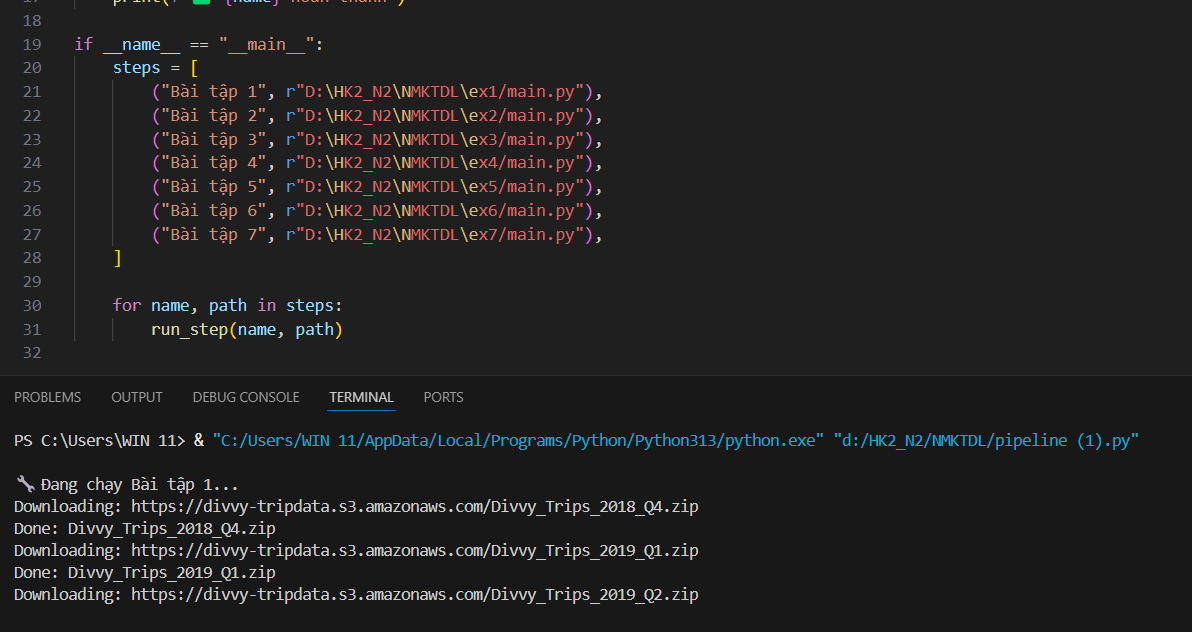
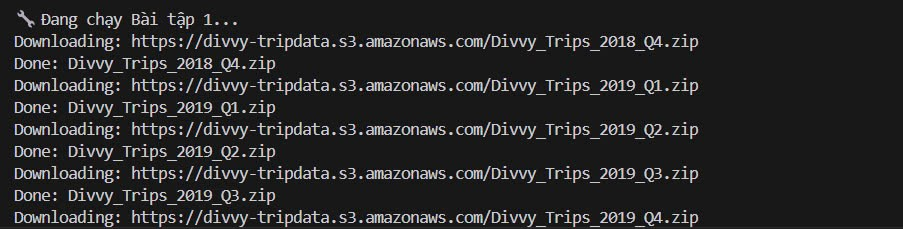
**Báo cáo các bài tập**

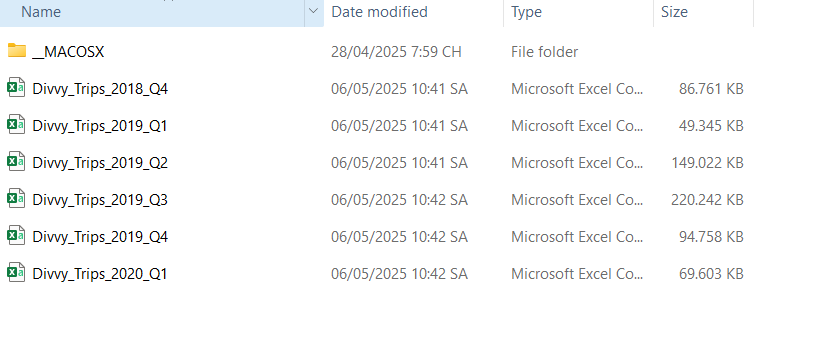
**Pipeline chạy lần lượt các Exercise-1, Exercise-2, Exercise-3, Exercise-4, Exercise-5, Exercise-6, Exercise-7**



**Exercise 1**

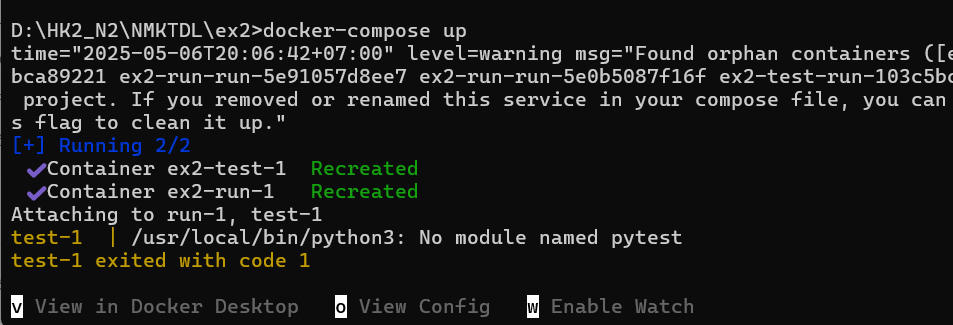
Sau khi thực hiện lệnh docker build --tag=exercise-1 ., thêm mã vào file main.py và chạy docker-compose up, các file zip đã được tải xuống (ngoại trừ một liên kết bị lỗi: [https://divvy](https://divvy/) tripdata.s3.amazonaws.com/Divvy\_Trips\_2220\_Q1.zip). Các file zip được lưu trong thư mục downloads\_url.

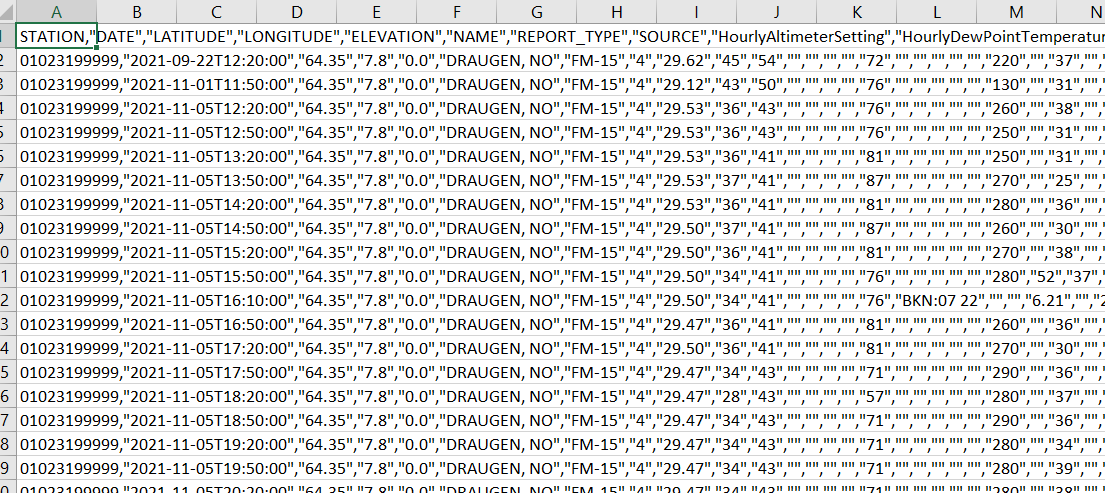




**Exercise 2**

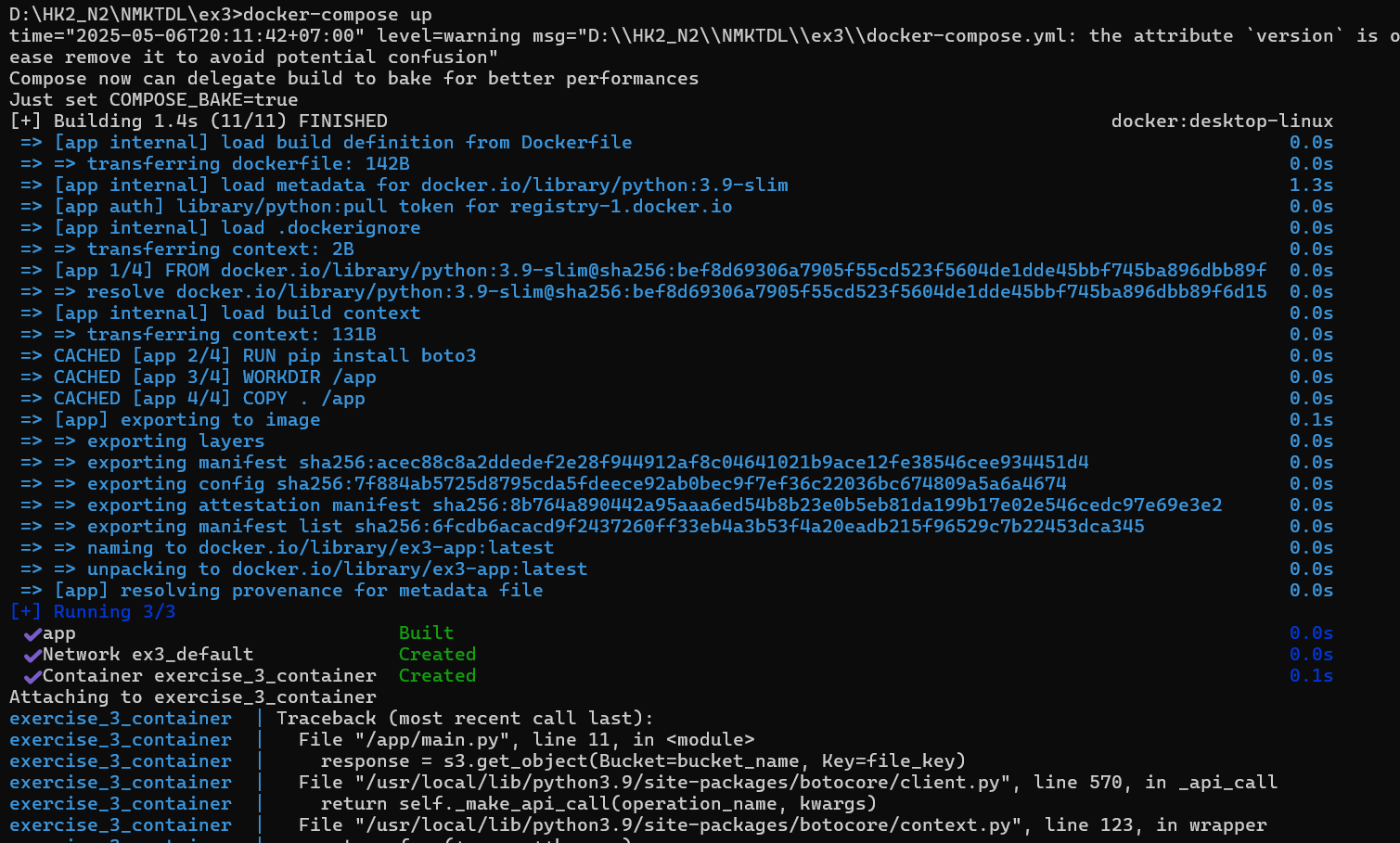
Sau khi docker build --tag=exercise-2 ., thêm code vào main.py và chạy docker-compose up. Cào dữ liệu từ website và lưu vào file CSV.

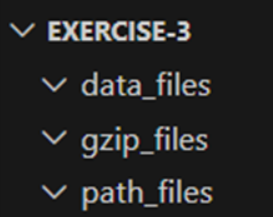




Exercise 3

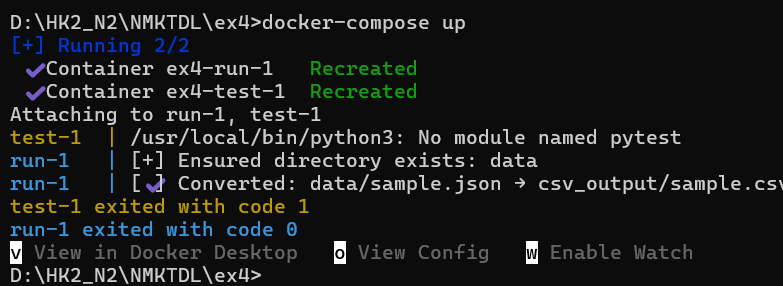
Sau khi chạy lệnh docker build --tag=exercise3 ., bạn thêm mã xử lý vào main.py để tải và giải nén dữ liệu từ Common Crawl. Khi thực thi docker-compose up trong thư mục EXERCISE-3, chương trình sẽ tự động tải file wet.paths.gz từ S3 và tạo ba thư mục: gzip\_files (chứa file .gz tải về), path\_files (chứa file .txt giải nén, liệt kê đường dẫn WET), và data\_files (lưu nội dung .wet đã giải nén từ web).

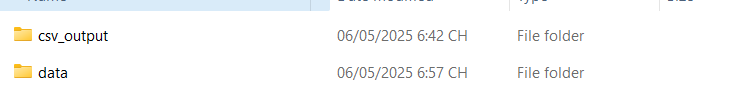


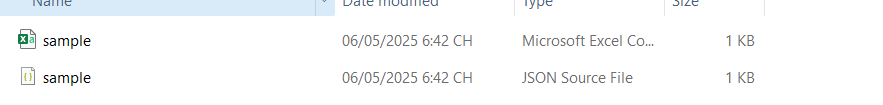


**EXERCISE-4**

Sau khi docker build --tag=exercise4 ., thêm code vào main.py và chạy docker-compose up. Tìm tất cả file JSON trong thư mục data và các thư mục con, sau đó chuyển từng file sang CSV.

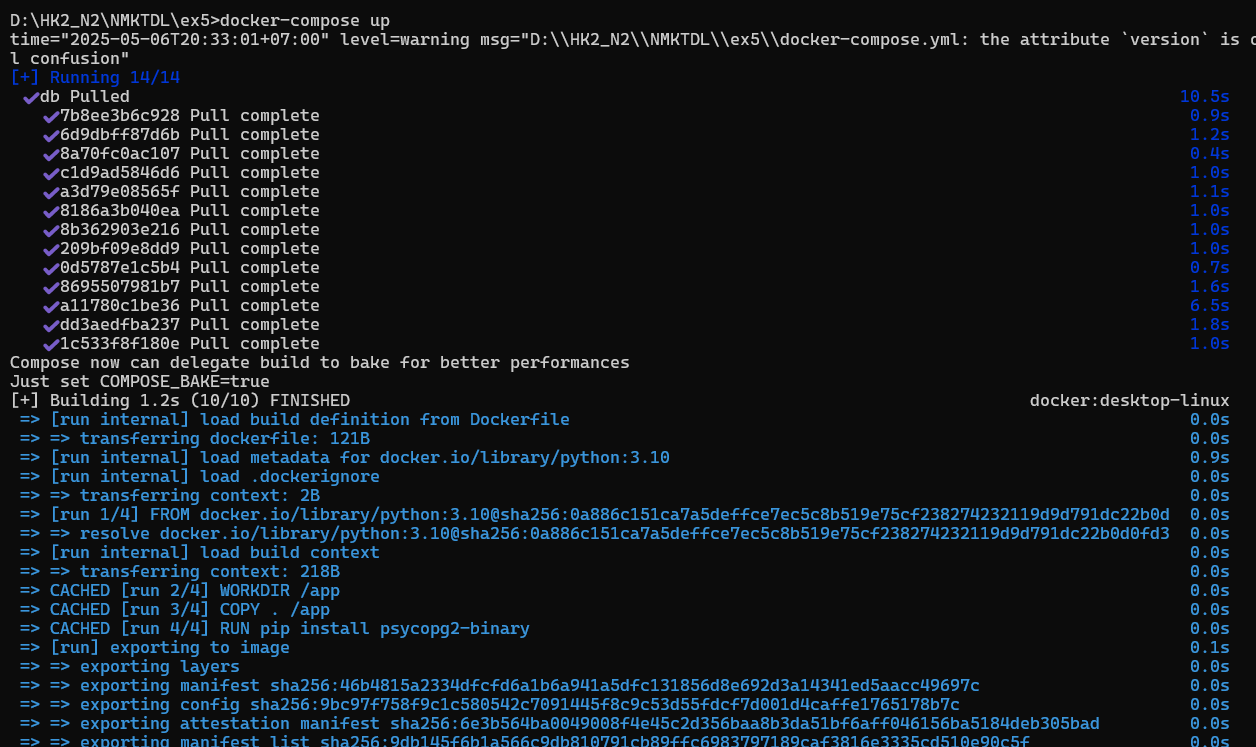


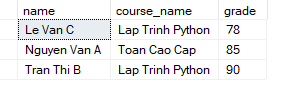


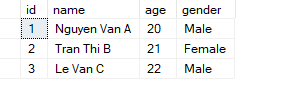


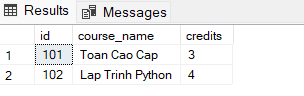
**Exercise 5**

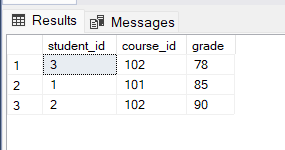
Sau khi docker build --tag=exercise5 ., thêm code vào main.py và chạy docker-compose up. Kết nối PostgreSQL và pgAdmin4, tạo bảng cho 3 file CSV trong data và insert dữ liệu tương ứng.





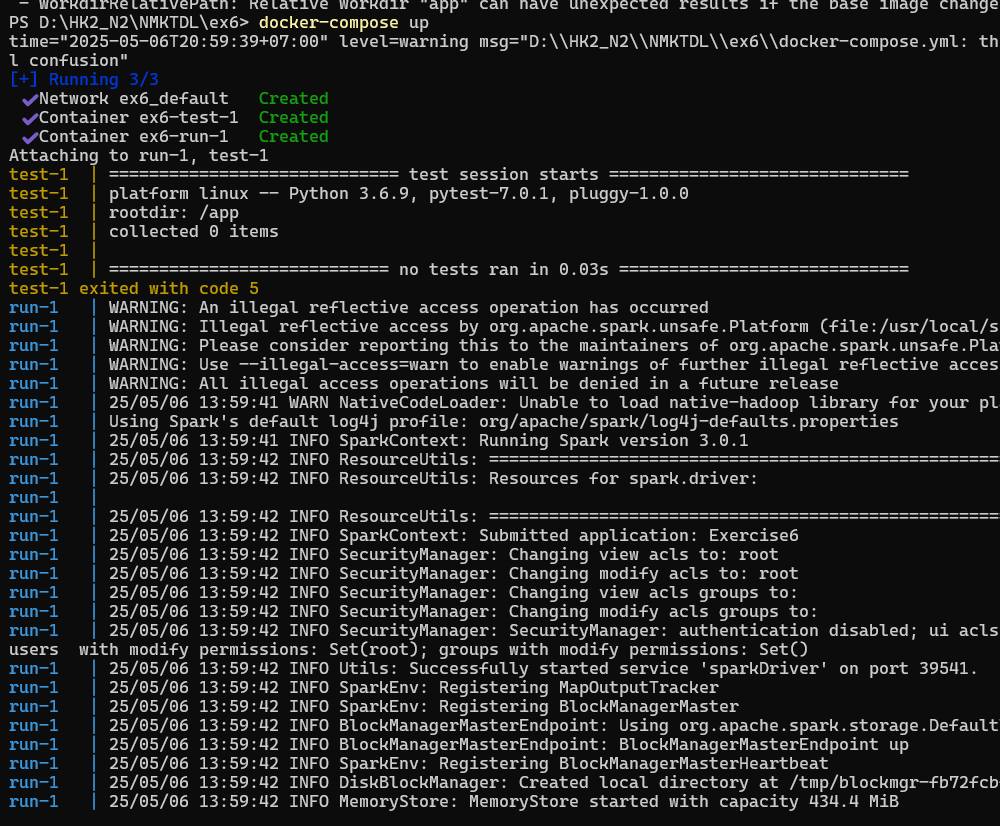


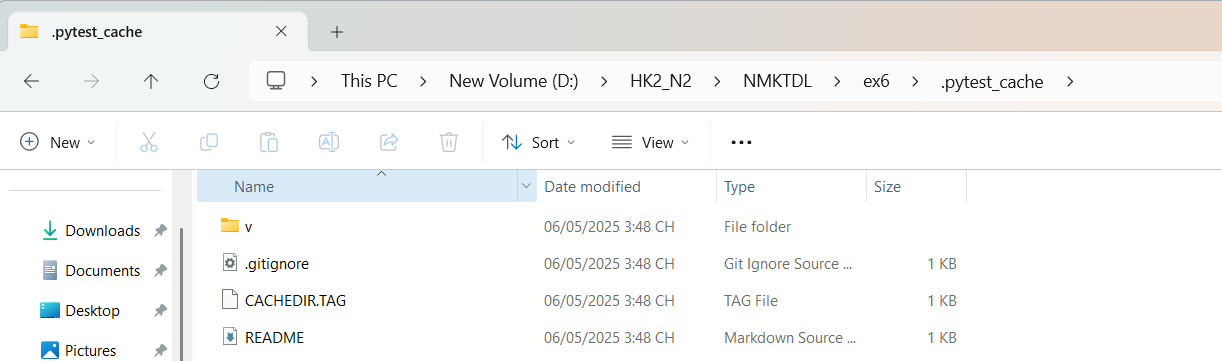




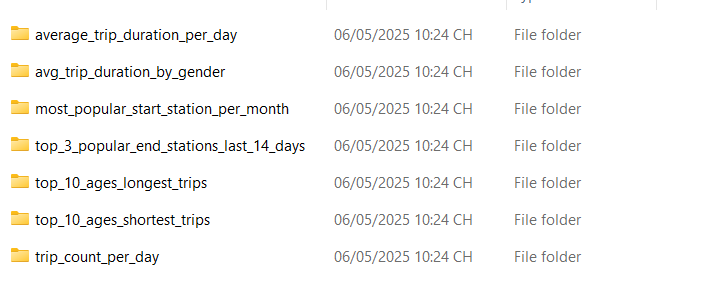
**Exercise 6**

Sau khi docker build --tag=exercise6 ., thêm code vào main.py và chạy docker-compose up.Dữ liệu được lưu vào file csv và viết filecode PySpark. Sinh ra file report , trong file report gồm các file dữ liệu

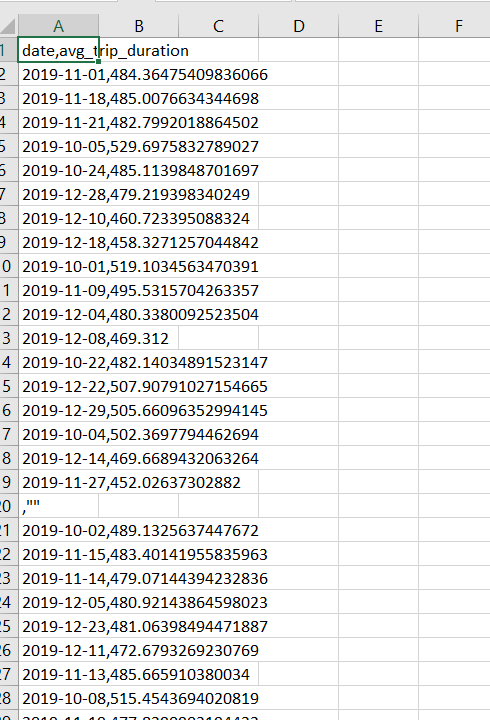
****

****

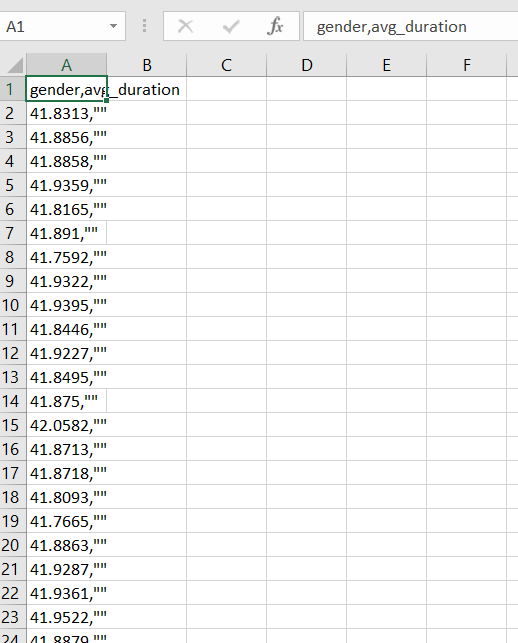
**File dữ liệu trong file report**

****

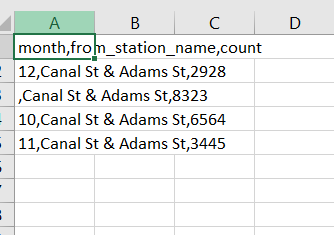
File average\_trip\_duration\_per\_day:

****

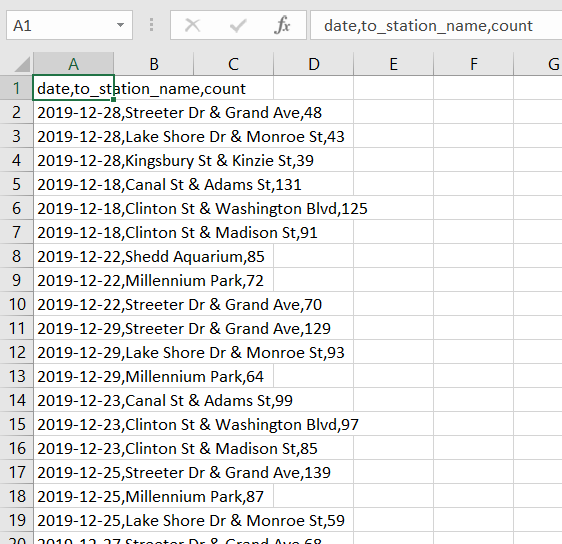
**File avg\_trip\_duration\_by\_gender**

****

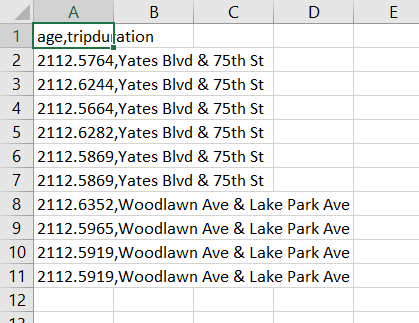
**File most\_popular\_start\_station\_per\_month**

****

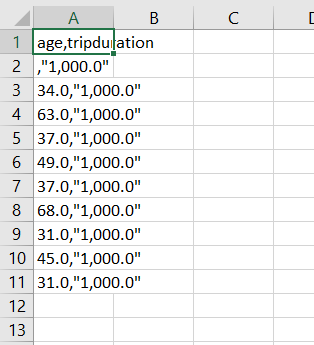
**FILE top\_3\_popular\_end\_stations\_last\_14\_days**

****

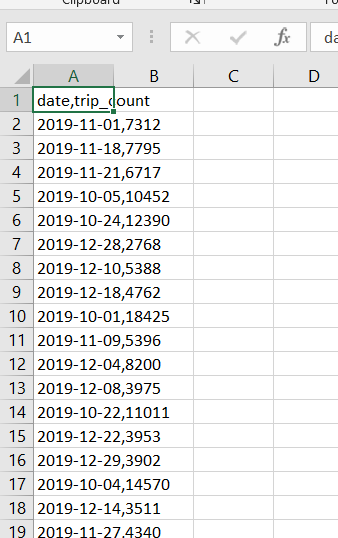
**FILE top\_10\_ages\_longest\_trips**

****

**File top\_10\_ages\_shortest\_trips**

****

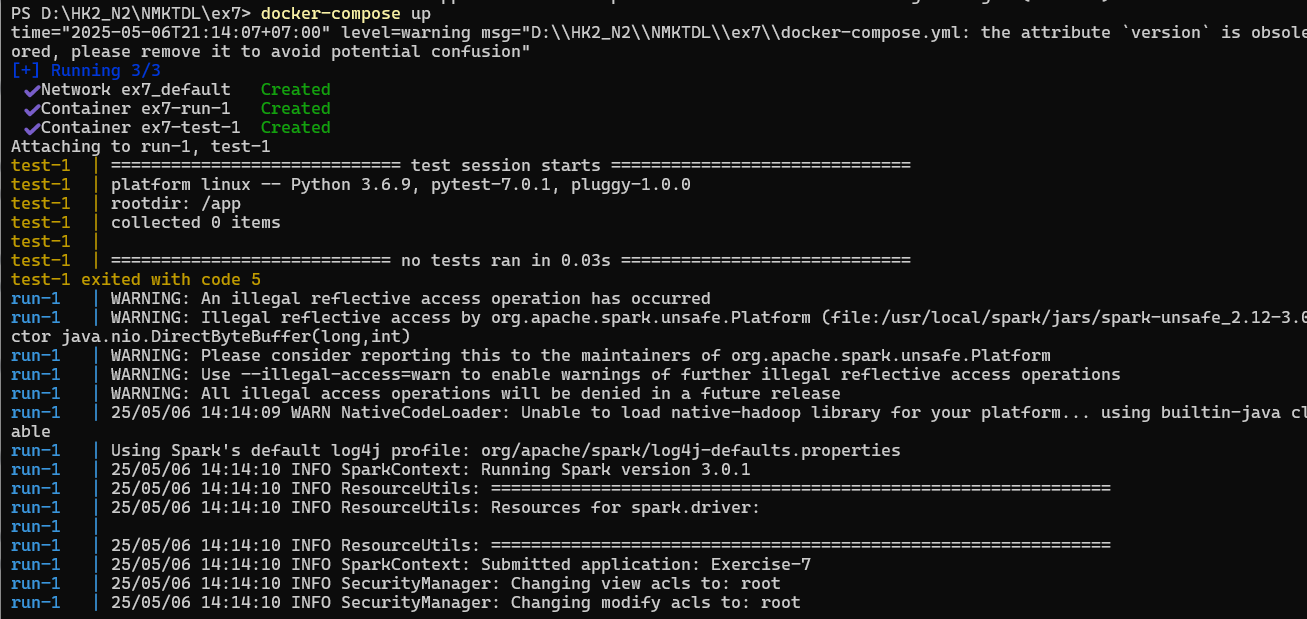
**File trip\_count\_per\_day**

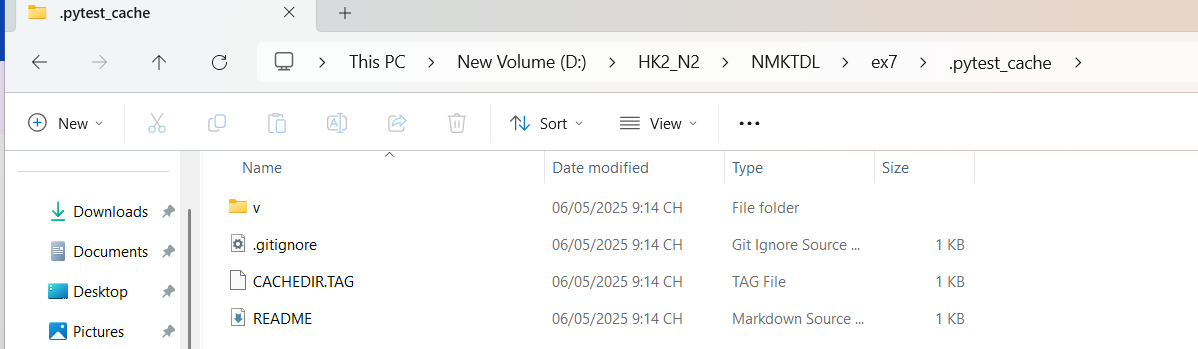
****

**Exercise 7**

Sau khi docker build --tag=exercise7 ., thêm code vào main.py và chạy docker-compose up.

Sinh ra các file

****

****