

Data Engineering Zoomcamp Week 1 Homework

Kubilay Çıtak

Question 1. Knowing docker tags

Q: Run the command to get information on Docker

```
``docker --help``
```

Now run the command to get help on the "docker build" command

Which tag has the following text? - *Write the image ID to the file*

- `--imageid string`
- **`--iidfile string`**
- `--idimage string`
- `--idfile string`

A: For this question I run the command "docker build --help" and saw that `--iidfile string`` tag has the text "Write the image ID to the file"

```
--force-rm          Always remove intermediate conta
--iidfile string     Write the image ID to the file
--isolation string   Container isolation technology
```

Question 2. Understanding docker first run

Q: Run docker with the python:3.9 image in an interactive mode and the entrypoint of bash.

Now check the python modules that are installed (use pip list).

How many python packages/modules are installed?

- 1
- 6
- **3**
- 7

A: I run the command "docker run -it --entrypoint=bash python:3.9" which allowed me to run my next command, which is "pip list". With this, I saw that there are 3 modules that are installed, as below;

```
$ docker run -it --entrypoint=bash python:3.9
root@4a6dfc54c37d:/# pip list
Package    Version
-----
pip        22.0.4
setuptools 58.1.0
wheel      0.38.4
```

Preparing Postgres

- First of all I run this command;

```
docker run -it -e POSTGRES_USER="root" -e POSTGRES_PASSWORD="root" -e POSTGRES_DB="ny_taxi" -v $(pwd)/ny_taxi_postgres_data:/var/lib/postgresql/data -p 5431:5432 postgres:13
```

- After it run successfully, I opened another terminal and used this command to connect;

```
pgcli -h localhost -p 5431 -u root -d ny_taxi
```

- In another terminal, I used these commands to get the data to the main folder;

```
wget https://github.com/DataTalksClub/nyc-tlc-data/releases/download/green/green_tripdata_2019-01.csv.gz
```

```
wget https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv
```

- Then I opened a Jupyter Notebook, used pandas and sqlalchemy to read and upload the data to PostgreSQL. You can use this notebook in the homework files, as "upload-data.ipynb"
- For the next questions, I used pgcli to write SQL codes and get the correct answers.

Question 3. Count records

Q: How many taxi trips were totally made on January 15?

Tip: started and finished on 2019-01-15.

Remember that `lpep_pickup_datetime` and `lpep_dropoff_datetime` columns are in the format timestamp (date and hour+min+sec) and not in date.

- 20689

- 20530

- 17630

- 21090

A: 20530

```
root@localhost:ny_taxi> select count(1) from taxi_data td
where td.lpep_pickup_datetime >= '2019-01-15 00:00:00' and
td.lpep_dropoff_datetime < '2019-01-16 00:00:00'
+-----+
| count |
+-----+
| 20530 |
+-----+
SELECT 1
Time: 0.058s
```

Question 4. Largest trip for each day

Q: Which was the day with the largest trip distance
Use the pick up time for your calculations.

- 2019-01-18
- 2019-01-28
- 2019-01-15
- 2019-01-10

A: 2019-01-15

```
root@localhost:ny_taxi> select lpep_pickup_datetime, max(trip_distance) as max_distance
from taxi_data group by lpep_pickup_datetime order by max_distance desc limit 1
+-----+-----+
| lpep_pickup_datetime | max_distance |
+-----+-----+
| 2019-01-15 19:27:58   | 117.99       |
+-----+-----+
SELECT 1
Time: 0.722s
```

Question 5. The number of passengers

Q: In 2019-01-01 how many trips had 2 and 3 passengers?

- 2: 1282 ; 3: 266
- 2: 1532 ; 3: 126
- 2: 1282 ; 3: 254
- 2: 1282 ; 3: 274

A: 2: 1282 ; 3: 254

```
root@localhost:ny_taxi> select passenger_count, count(1) from taxi_data where passenger_count
in (2, 3) and (lpep_pickup_datetime between '2019-01-01 00:00:00' and '2019-01-02 00:00:00')
group by passenger_count
+-----+-----+
| passenger_count | count |
+-----+-----+
| 2               | 1282  |
| 3               | 254   |
+-----+-----+
SELECT 2
Time: 0.049s
```

Question 6. Largest tip

Q: For the passengers picked up in the Astoria Zone which was the drop off zone that had the largest tip? We want the name of the zone, not the id.

Note: it's not a typo, it's `tip`, not `trip`

- Central Park
- Jamaica
- South Ozone Park
- Long Island City/Queens Plaza

A: Long Island City/Queens Plaza

```
root@localhost:ny_taxi> select lu."Zone", max(td."tip_amount") as max_tip from taxi_data td inner join
taxi_data_lu lu on td."DOLocationID" = lu."LocationID" where td."PULocationID" = (select distinct
td."PULocationID" from taxi_data td inner join taxi_data_lu lu on td."PULocationID" = lu."LocationID"
where lu."Zone" = 'Astoria') group by lu."Zone" order by max_tip desc limit 1
+-----+-----+
| Zone | max_tip |
+-----+-----+
| Long Island City/Queens Plaza | 88.0 |
+-----+-----+
SELECT 1
Time: 0.088s
```

Submitting the solutions

- * Form for submitting: [form](https://forms.gle/EjphSkR1b3nsdojv7)
- * You can submit your homework multiple times. In this case, only the last submission will be used.

Deadline: 26 January (Thursday), 22:00 CET

Solution

We will publish the solution here