

## Question 1. Understanding docker first run

Run docker with the python:3.12.8 image in an interactive mode, use the entrypoint bash.

What's the version of pip in the image?

- ☒ 24.3.1
- ☐ 24.2.1
- ☐ 23.3.1
- ☐ 23.2.1

Solution

```
cd c:/users/zhaoy/data-engineering-zoomcamp/2_docker_sql
```

```
docker run -it --entrypoint=bash python 3.12.8
```

```
zhaoy@ZYX MINGW64 ~
$ cd c:/users/zhaoy/data-engineering-zoomcamp/2_docker_sql

zhaoy@ZYX MINGW64 /c/users/zhaoy/data-engineering-zoomcamp/2_docker_sql (master)
$ docker run -it --entrypoint=bash python 3.12.8
Unable to find image 'python:latest' locally
latest: Pulling from library/python
10a1c0fc7d27: Download complete
98a3d7a4f991: Download complete
1db49f4673ea: Download complete
Digest: sha256:6ee79759eb6c6843f7aec973df1d3ae60f7199822669deaf77fba16a7b27d1db
Status: Downloaded newer image for python:latest
bash: 3.12.8: No such file or directory

zhaoy@ZYX MINGW64 /c/users/zhaoy/data-engineering-zoomcamp/2_docker_sql (master)
$ python
  pip.__version__
    ^
SyntaxError: invalid syntax
>>> pip.__version__
'24.3.1'
>>> |
```

## Question 2. Understanding Docker networking and docker-compose

Given the following docker-compose.yaml, what is the hostname and port that pgadmin should use to connect to the postgres database?

services:

db:

container\_name: postgres

image: postgres:17-alpine

environment:

POSTGRES\_USER: 'postgres'

POSTGRES\_PASSWORD: 'postgres'

POSTGRES\_DB: 'ny\_taxi'

ports:

- '5433:5432'

volumes:

```
- vol-pgdata:/var/lib/postgresql/data
```

```
pgadmin:
```

```
  container_name: pgadmin
```

```
  image: dpage/pgadmin4:latest
```

```
  environment:
```

```
    PGADMIN_DEFAULT_EMAIL: "pgadmin@pgadmin.com"
```

```
    PGADMIN_DEFAULT_PASSWORD: "pgadmin"
```

```
  ports:
```

```
    - "8080:80"
```

```
  volumes:
```

```
    - vol-pgadmin_data:/var/lib/pgadmin
```

```
volumes:
```

```
  vol-pgdata:
```

```
    name: vol-pgdata
```

```
  vol-pgadmin_data:
```

```
    name: vol-pgadmin_data
```

☐ postgres:5433

☐ localhost:5432

☐ db:5433

☐ postgres:5432

☒ db:5432

### Question 3. Trip Segmentation Count

During the period of October 1st 2019 (inclusive) and November 1st 2019 (exclusive), how many trips, respectively, happened:

Up to 1 mile

In between 1 (exclusive) and 3 miles (inclusive),

In between 3 (exclusive) and 7 miles (inclusive),

In between 7 (exclusive) and 10 miles (inclusive),

Over 10 miles

Answers:

☐ 104,802; 197,670; 110,612; 27,831; 35,281

☐ 104,802; 198,924; 109,603; 27,678; 35,189

☐ 104,793; 201,407; 110,612; 27,831; 35,281

☐ 104,793; 202,661; 109,603; 27,678; 35,189

☒ 104,838; 199,013; 109,645; 27,688; 35,202

Solution



Query

Query History

6

SELECT

7

DATE(lpep\_pickup\_datetime) AS pickup\_date,

8

trip\_distance,

9

ROW\_NUMBER() OVER (PARTITION BY DATE(lpep\_pickup\_datetime) ORDER BY tri

10

FROM

11

green\_taxi\_data

12

)

13

SELECT

14

pickup\_date,

15

trip\_distance

16

FROM

Data Output

Messages

Notifications

+

📄

▼

📋

▼

🗑️

🗄️

⬇️

📈

SQL

Showing rows: 1 to 1

✎

Page No:

	pickup_date date	trip_distance double precision
1	2019-10-31	515.89

### Question 5. Three biggest pickup zones

Which were the top pickup locations with over 13,000 in total\_amount (across all trips) for 2019-10-18?

Consider only lpep\_pickup\_datetime when filtering by date.

- ☒ East Harlem North, East Harlem South, Morningside Heights
- ☐ East Harlem North, Morningside Heights
- ☐ Morningside Heights, Astoria Park, East Harlem South
- ☐ Bedford, East Harlem North, Astoria Park

QueryQuery History

```

1 SELECT zs."Zone", yt."PULocationID" AS pickupID, SUM(yt.total_amount) AS total_am
2 FROM green_taxi_data yt
3 JOIN zones zs ON yt."PULocationID" = zs."LocationID"
4 WHERE yt.lpep_pickup_datetime BETWEEN '2019-10-18' AND '2019-10-19'
5 GROUP BY zs."Zone", yt."PULocationID"
6 HAVING SUM(yt.total_amount) > 13000
7 ORDER BY total_amount_sum DESC;

```

Data OutputMessagesNotifications

+

📄

▼

📋

▼

🗑️

🗄️

⬇️

📈

SQL

Showing rows: 1 to 3

✎

Page No

	Zone text	pickupid bigint	total_amount_sum double precision
1	East Harlem North	74	18686.680000000008
2	East Harlem South	75	16797.2600000000057
3	Morningside Heights	166	13029.790000000003

### Question 6. Largest tip

For the passengers picked up in October 2019 in the zone named "East Harlem North" which was the drop off zone that had the largest tip?

Note: it's tip , not trip

We need the name of the zone, not the ID.

- ☐ Yorkville West
- ☐ JFK Airport
- ☒ East Harlem North
- ☐ East Harlem South

Query
Query History

```

8
9 SELECT
10     zpu."Zone" AS pickup_zone,
11     zdo."Zone" AS dropoff_zone,
12     MAX(gtd.tip_amount) AS max_tip
13 FROM
14     green_taxi_data gtd
15 JOIN
16     zones zpu ON gtd."PULocationID" = zpu."LocationID"
17 JOIN
18     zones zdo ON gtd."DOLocationID" = zdo."LocationID"

```

Data Output
Messages
Notifications

+

📄

▼

📋

▼


🗑️

🗄️

⬇️

📈

SQL

Showing rows: 1 to 1  Page No:

	pickup_zone text	dropoff_zone text	max_tip double precision
1	East Harlem North	East Harlem North	87.3

Which of the following sequences, respectively, describes the workflow for:

1. Downloading the provider plugins and setting up backend,
2. Generating proposed changes and auto-executing the plan
3. Remove all resources managed by terraform`

Answers:

- ☐ terraform import, terraform apply -y, terraform destroy
- ☐ teraform init, terraform plan -auto-apply, terraform rm
- ☐ terraform init, terraform run -auto-approve, terraform destroy
- ☒ terraform init, terraform apply -auto-approve, terraform destroy
- ☐ terraform import, terraform apply -y, terraform rm