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?

- [x] 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

(I know the process, but maybe there are some issues with my dataset. So the answer is based on Excel output, attach is the screenshot of PgAdmin)

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

```sql
SELECT COUNT(1)
FROM green_taxi_data
WHERE trip_distance <= 1
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT COUNT(*)
FROM green_taxi_data
WHERE trip_distance <= 3 AND trip_distance > 1
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT
```

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page

	count bigint
1	79242

```sql
SELECT COUNT(*)
FROM green_taxi_data
WHERE trip_distance <= 3 AND trip_distance > 1
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT
```

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page

	count bigint
1	151599

```sql
SELECT COUNT(*)
FROM green_taxi_data
WHERE trip_distance <= 7 AND trip_distance > 3
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT
```

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page

	count bigint
1	89550

```sql
SELECT COUNT(*)
FROM green_taxi_data
WHERE trip_distance <= 10 AND trip_distance > 7
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT
```

Data Output | Messages | Notifications

Showing rows: 1 to 1 | Page

	count bigint
1	23825

```sql
SELECT COUNT(*)
FROM green_taxi_data
WHERE trip_distance > 10
AND lpep_dropoff_datetime BETWEEN '2019-10-01' AND '2019-11-01';

SELECT
```

Data Output | Messages | Notifications

Showing rows: 1 to

	count bigint
1	32002

Question 4. Longest trip for each day

Which was the pick up day with the longest trip distance? Use the pick up time for your calculations.

Tip: For every day, we only care about one single trip with the longest distance.

- [ ] 2019-10-11
- [ ] 2019-10-24
- [ ] 2019-10-26
- [x] 2019-10-31

```sql
SELECT
    DATE(lpep_dropoff_datetime) AS pickup_date,
    SUM(trip_distance) AS total_distance
FROM
    green_taxi_data
GROUP BY
    DATE(lpep_dropoff_datetime)
ORDER BY
    total_distance DESC
LIMIT 1;
```

Data Output | Messages | Notifications

	pickup_date date	total_distance double precision
1	2019-10-18	62243.95999999994

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

```
21  ORDER BY total_amount_sum DESC;
22
23  SELECT zs."Zone", yt."PULocationID", SUM(yt.total_amount) AS total_amount_sum
24  FROM yellow_taxi_data yt
25  JOIN zones zs ON yt."PULocationID" = zs."LocationID"
26  WHERE yt.tpep_pickup_datetime BETWEEN '2019-10-18' AND '2019-10-19'
27  GROUP BY zs."Zone", yt."PULocationID"
28  HAVING SUM(yt.total_amount) > 13000
29  ORDER BY total_amount_sum DESC;
30
31
```

Data Output | Messages | Notifications

Showing rows: 1 to 53 | Page No: 1

	Zone text	PULocationID bigint	total_amount_sum double precision
1	JFK Airport	132	520934.96000000346
2	LaGuardia Airport	138	342691.36999999976
3	Midtown Center	161	207698.1300000061
4	Times Sq/Theatre District	230	186122.85000000425
5	Upper East Side South	237	185788.4100000058
6	Midtown East	162	174646.12000000355
7	Penn Station/Madison Sq West	186	170863.66000000358

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
- [x] East Harlem South

```sql
SELECT zs2."Zone" AS dropoff_zone, MAX(yt.tip_amount) AS max_tip
FROM yellow_taxi_data yt
JOIN zones zs1 ON yt."PULocationID" = zs1."LocationID"
JOIN zones zs2 ON yt."DOLocationID" = zs2."LocationID"
WHERE zs1."Zone" = 'East Harlem North'
  AND yt.tpep_pickup_datetime BETWEEN '2019-10-01' AND '2019-10-31'
GROUP BY zs2."Zone"
ORDER BY max_tip DESC
LIMIT 10;
```

Data Output | Messages | Notifications

Showing rows: 1 to 10 | Page N

	dropoff_zone text	max_tip double precision
1	West Concourse	150
2	East Harlem North	98
3	Upper East Side North	81
4	Outside of NYC	45.58
5	LaGuardia Airport	40
6	East Harlem South	31.23

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