**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”

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**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;

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**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

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**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

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**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

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**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

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

## 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