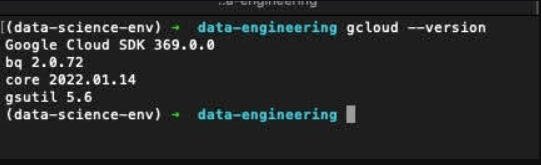
**ANSWER KEY FOR HOMEWORK – WEEK 1 /DATA ENGINEERING**

[**https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/week\_1\_basics\_n\_setup/homework.md**](https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/week_1_basics_n_setup/homework.md)

**Question 1. Google Cloud SDK**

**Install Google Cloud SDK. What's the version you have?**

**To get the version, run gcloud –version**



**Google Cloud account**

Create an account in Google Cloud and create a project.

**Question 2. Terraform**

**Now install terraform and go to the terraform directory (week\_1\_basics\_n\_setup/1\_terraform\_gcp/terraform)**

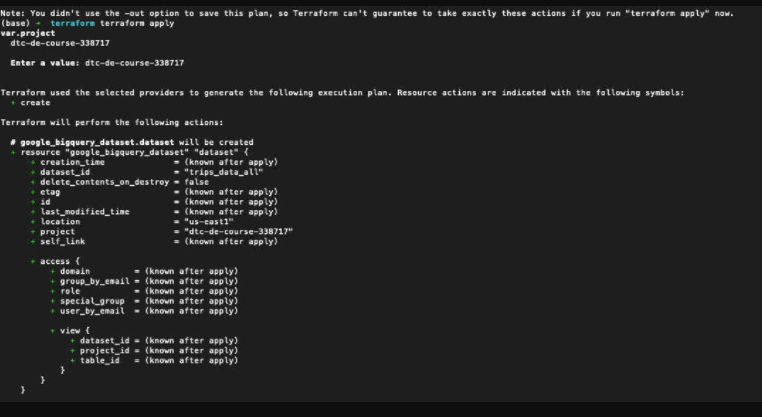
After that, run

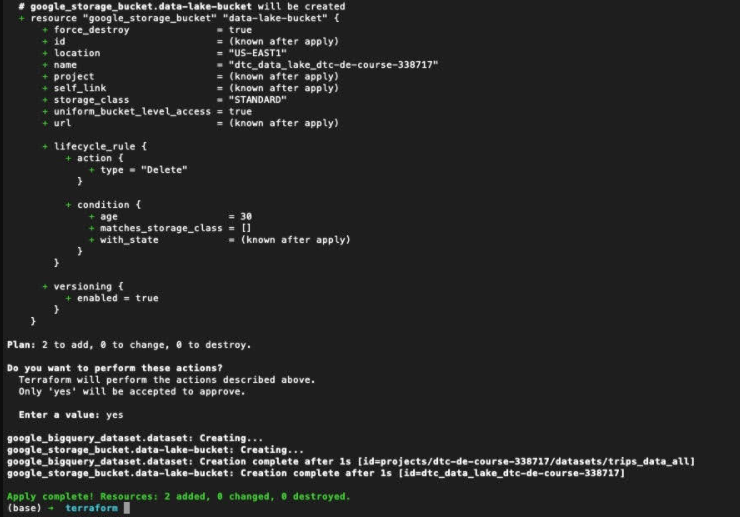
terraform init

terraform plan

terraform apply

Apply the plan and copy the output (after running apply) to the form





**Prepare Postgres**

Run Postgres and load data as shown in the videos

We'll use the yellow taxi trips from January 2021:

wget https://s3.amazonaws.com/nyc-tlc/trip+data/yellow\_tripdata\_2021-01.csv

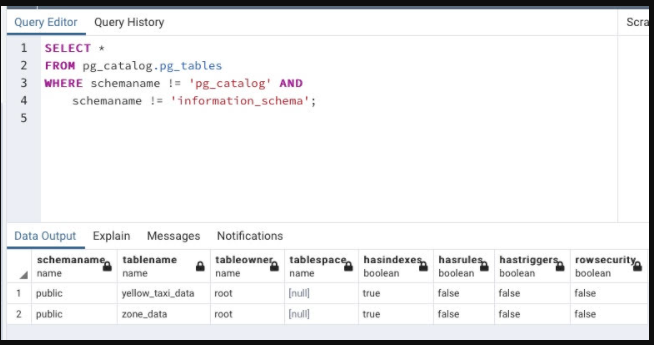
You will also need the dataset with zones:

wget https://s3.amazonaws.com/nyc-tlc/misc/taxi+\_zone\_lookup.csv

Download this data and put it to Postgres

**Execute following commands :**

docker run -it \c                               
e- POSTGRES\_USER="root" \  
e- POSTGRES\_PASSWORD="root"  
: 1642606466:0;docker run -it \  
-e POSTGRES\_USER="root" \  
-e POSTGRES\_PASSWORD="root" \  
-e POSTGRES\_DB="nyc\_taxi" \  
-v /Users/subramanianhariharan/Documents/kumar\_work/data-engineering/nyc\_data:/var/lib/postgresql/data/ \  
-p 5432:5432 \  
--name postgres\_data \  
--network=pg-network \  
postgres:13  
  
docker run -it \  
-e PGADMIN\_DEFAULT\_EMAIL="[admin@admin.com](mailto:admin@admin.com)" \  
-e PGADMIN\_DEFAULT\_PASSWORD="root" \  
-p 8080:80 \  
--name pgadmin \  
--network=pg-network \  
dpage/pgadmin4  
  
# convert ipynb to python script  
jupyter nbconvert --to=script <filename>  
  
#running py script ingest\_data.py with env variables  
URL ="<https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2021-01.csv>"  
python ingest\_data.py \  
--user=root \  
--password=root \  
--db=nyc\_taxi \  
--host=localhost \  
--port=5432 \  
--table\_name=yellow\_taxi\_data \  
--url=$(URL)  
  
#build a Dockerfile for running dtc\_ingest\_data.py  
docker build -t ingest\_data:v1 .  
docker run -it \  
--network=pg-network \  
taxi\_ingest:v1 \  
--user=root \  
--password=root \  
--db=nyc\_taxi \  
--host=postgres\_data \  
--port=5432 \  
--table\_name=yellow\_taxi\_data \  
--url="<https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2021-01.csv>"  
  
#pushing zone data into nyc\_data  
docker build -t ingest\_data:v1 .  
docker run -it \  
--network=pg-network \  
taxi\_ingest:v1 \  
--user=root \  
--password=root \  
--db=nyc\_taxi \  
--host=postgres\_data \  
--port=5432 \  
--table\_name=zone\_data \  
--url="<https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv>"



**Question 3. Count records**

How many taxi trips were there on January 15?

Consider only trips that started on January 15.

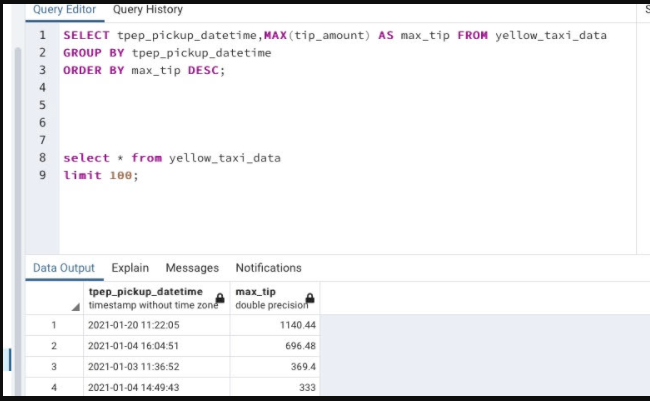
Ans 53024

## Question 4. Average

Find the largest tip for each day. On which day it was the largest tip in January?

Use the pick up time for your calculations.

(note: it's not a typo, it's "tip", not "trip")

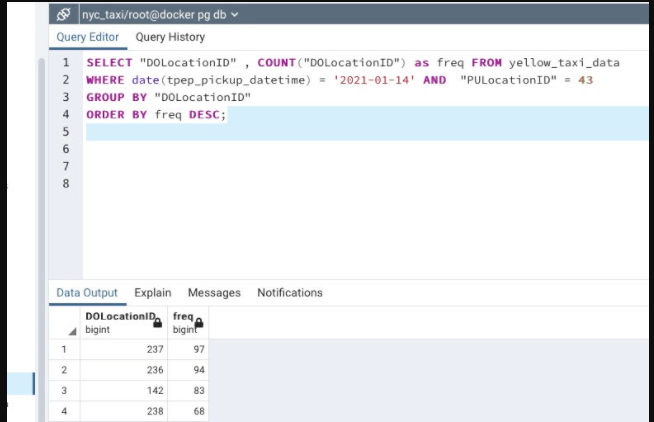


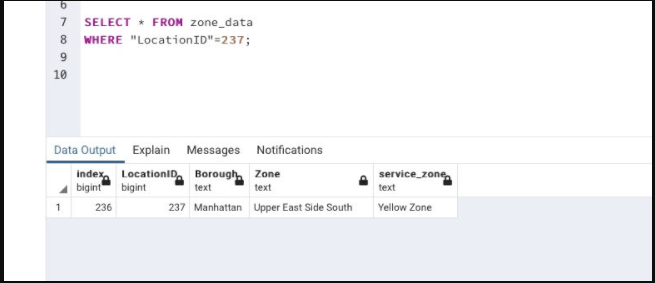
## Question 5. Most popular destination

What was the most popular destination for passengers picked up in central park on January 14?

Use the pick up time for your calculations.

Enter the zone name (not id). If the zone name is unknown (missing), write "Unknown"





## Question 6.

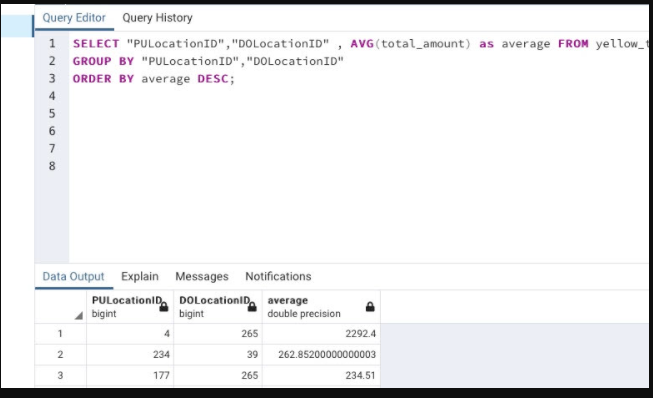
What's the pickup-dropoff pair with the largest average price for a ride (calculated based on total\_amount)?

Enter two zone names separated by a slash

For example:

"Jamaica Bay / Clinton East"

If any of the zone names are unknown (missing), write "Unknown". For example, "Unknown / Clinton East".



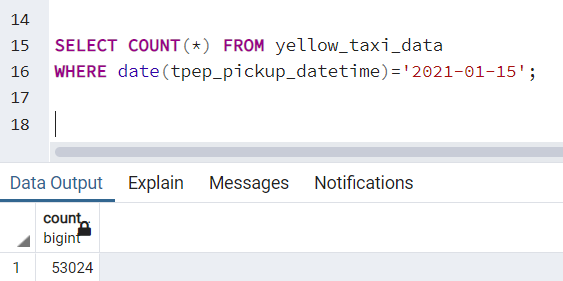


**ATTEMPT 2**

**Question 3. Count records**

**How many taxi trips were there on January 15?**

**Consider only trips that started on January 15.**

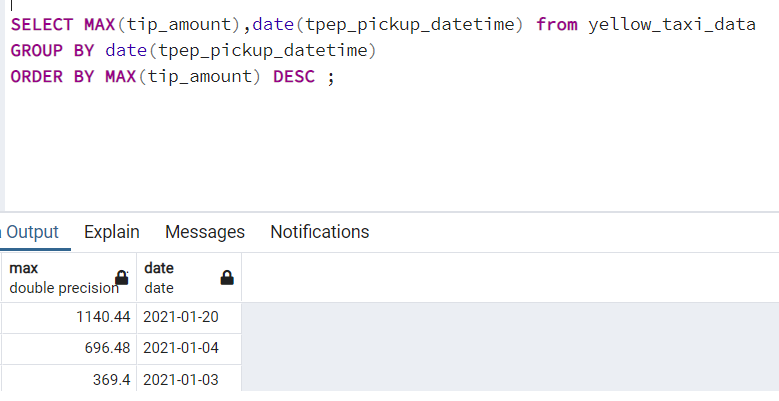


## Question 4.

**Find the largest tip for each day. On which day it was the largest tip in January?**

**Use the pick up time for your calculations.**

**(note: it's not a typo, it's "tip", not "trip")**



**Q No 5 Most popular destination from Central park**

SELECT t.pulocationid,t.dolocationid ,COUNT(dolocationid),zpu.zone\_name as pick\_up\_zone,

zdo.zone\_name as dr\_off\_zone

FROM yellow\_taxi\_data t,

zone\_data zpu,

zone\_data zdo

WHERE t.pulocationid= zpu.locationid AND

t.dolocationid= zdo.locationid AND

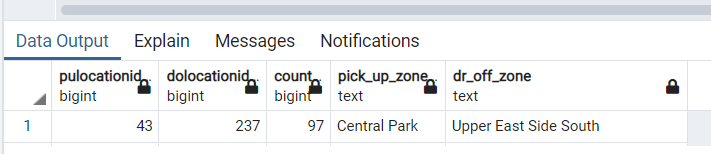
zpu.zone\_name='Central Park' AND

date(tpep\_pickup\_datetime)='2021-01-14'

GROUP BY t.pulocationid,t.dolocationid,zpu.zone\_name,zdo.zone\_name

ORDER BY COUNT(dolocationid) DESC

LIMIT 100;



## Question 6.

What's the pickup-dropoff pair with the largest average price for a ride (calculated based on total\_amount)?

Enter two zone names separated by a slash

For example:

"Jamaica Bay / Clinton East"

If any of the zone names are unknown (missing), write "Unknown". For example, "Unknown / Clinton East".

