Homework 1

New Attempt

Due Oct 10 by 11:59pm **Points** 60 **Submitting** a file upload

In the Datasets Module you will find two schema and two CSV files corresponding to the Airbnb data shown in class.

1. Create and populate the Airbnb_listings and Airbnb_Reviews tables in your local PostgreSQL DBMS. Verify that the schema of your DBMS matches the schema given to you. (10 points)

2. For each question given below, write a corresponding SQL query and the corresponding query (as a hand-drawn picture). You can submit these in a Word File. State any assumptions that you make. (10 points each)

(a) Get the address and price of listings where the place is described as "quiet".

(b) Get the address and weekly price of listings for 1 bedroom apartments in Washington DC.

(c) How many bed and breakfast places are there in each city and what are the median prices? (You will have to look up how medians are computed from PostgreSQL documentation)

(d) In which cities are houses cheaper than townhouses on an average? Note that the answer can be "None".

3. Inspect the schema for Airbnb listings. Suggest at least 3 improvements of the schema. (10 points)

DSC-202: ASSIGNMENT-01

1) **Create and populate the Airbnb_listings and Airbnb_Reviews tables in your local PostgreSQL DBMS. Verify that the schema of your DBMS matches the schema given to you.**

Table Airbnb_listings :

```
create table "Airbnb_listings"
(
id                              integer,
listing_url                     text,
scrape_id                       integer,
last_scraped                    date,
name                            varchar(256) not null,
summary                         text,
space                           text,
description                     text,
experiences_offered             text,
neighborhood_overview           text,
notes                           text,
transit                         text,
thumbnail_url                   text,
medium_url                      text,
picture_url                     text,
xl_picture_url                  text,
host_id                         integer,
host_url                        text,
host_name                       varchar(128),
host_since                      text,
host_location                   varchar(256),
host_about                      text,
host_response_time              text,
host_response_rate              text,
host_acceptance_rate            text,
host_is_superhost               boolean,
host_thumbnail_url              text,
host_picture_url                text,
host_neighbourhood              text,
host_listings_count             integer,
host_total_listings_count       integer,
host_verifications              text[],
host_has_profile_pic            boolean,
host_identity_verified          boolean,
street                          text         not null,
neighborhood                    text,
neighborhood_cleansed           text,
neighborhood_group_cleansed     text,
city                            text,
state                           char(2)      not null,
zipcode                         integer,
market                          text,
smart_location                  text,
country_code                    varchar(5),
country                         text,
latitude                        numeric,
longitude                       numeric,
is_located_exact                boolean,
property_type                   text,
room_type                       text,
accommodates                    integer,
```

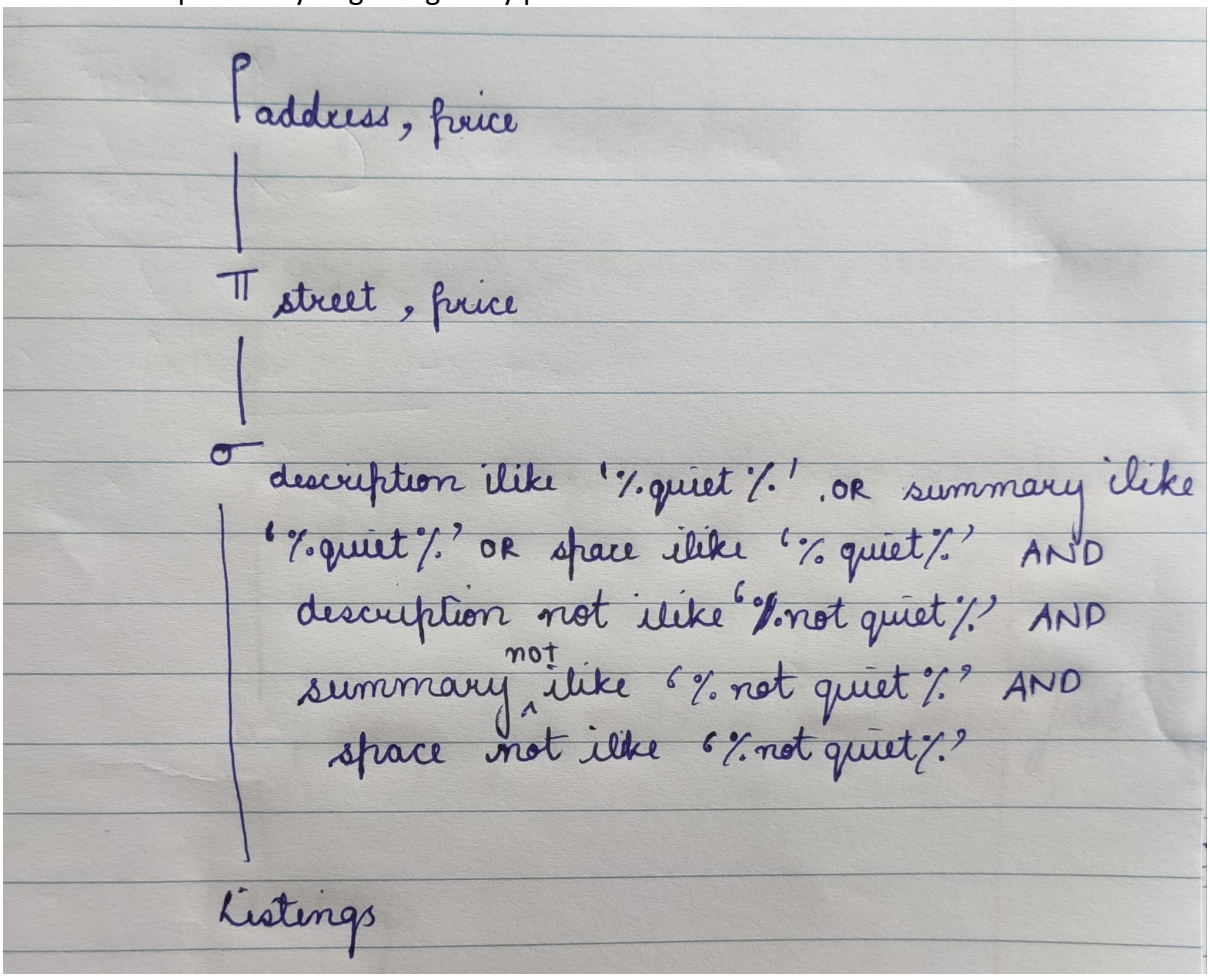
```
bathrooms                          numeric,
bedrooms                           integer,
beds                               integer,
bed_type                           text,
amenities                          text[],
square_feet                        text,
price                              text,
weekly_price                       text,
monthly_price                      text,
security_deposit                   text,
cleaning_fee                       text,
guests_included                    integer,
extra_people                       text,
minimum_nights                     integer,
maximum_nights                     integer,
calendar_updated                   text,
has_availability                   boolean,
availability_30                    integer,
availability_60                    integer,
availability_90                    integer,
availability_365                   integer,
calendar_last_scraped              date,
number_of_reviews                  integer,
first_review                       date,
last_review                        date,
review_scores_rating               numeric,
review_scores_accuracy             numeric,
review_score_cleanliness           numeric,
review_scores_checkin              numeric,
review_scores_communication        numeric,
review_scores_location             numeric,
review_scores_value                numeric,
requires_license                   boolean,
license                            text,
jurisdiction_names                 text,
instant_bookable                   boolean,
cancellation_policy                text,
require_guest_profile_picture      boolean,
require_guest_phone_verification   boolean,
calculated_host_listings_count     integer,
reviews_per_month                  numeric
);
```

We created a new Database Schema as above, put the first row as header and imported the saved CSV File to Datagrip in order to create a table on which we will run queries.

2) a) **Get the address and price of listings where the place is described as "quiet".**

```
a)  select street as address, price
from "Airbnb"."Listings"
where description ilike '%quiet%'
   or summary ilike '%quiet%'
   or space ilike '%quiet%' and description not ilike '%not quiet%' and
      summary not ilike '%not quiet%' and space not ilike '%not quiet%';
```

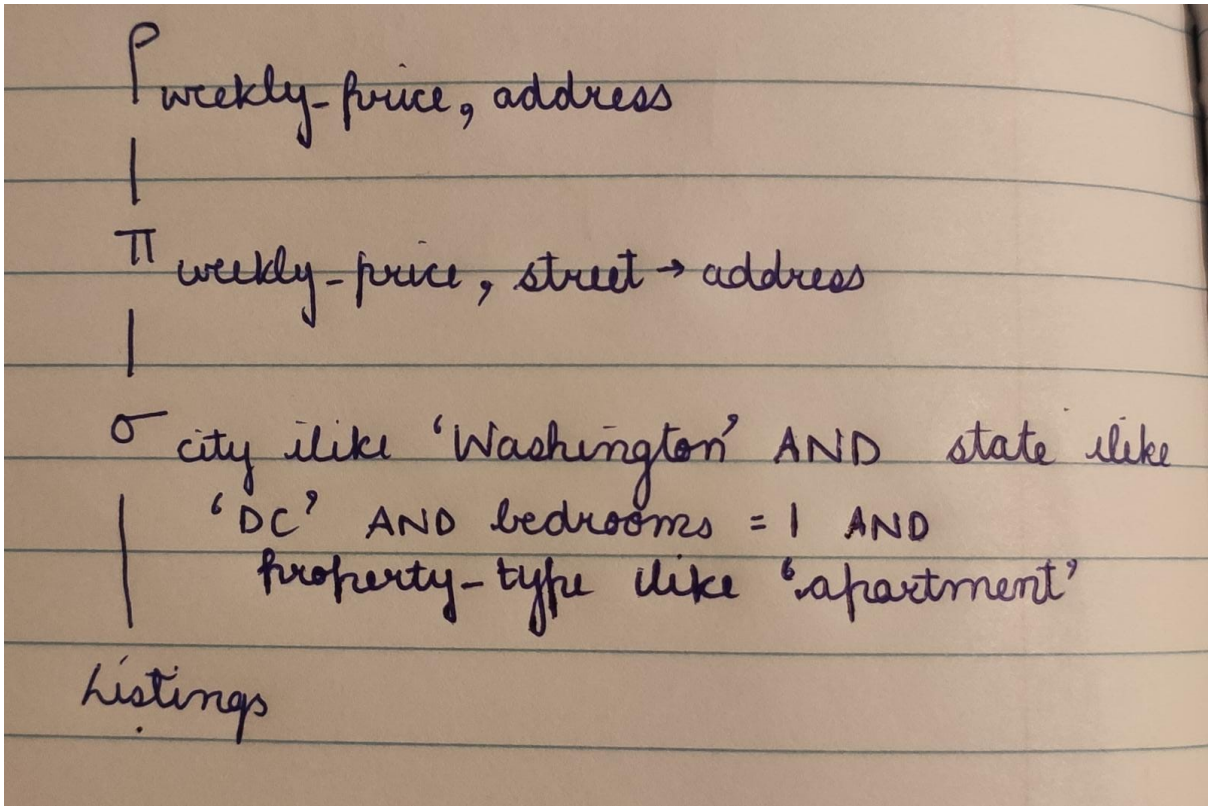
In a) we assume that on removing the situation where ‘not quiet’ occurs , we have removed the possibility of getting noisy places.



b) **Get the address and weekly price of listings for 1 bedroom apartments in Washington DC.**

```
b)  select street as address, weekly_price from "Airbnb"."Listings" where
property_type ilike 'apartment' and bedrooms = 1 and city ilike
'Washington' and state ilike 'DC'
```

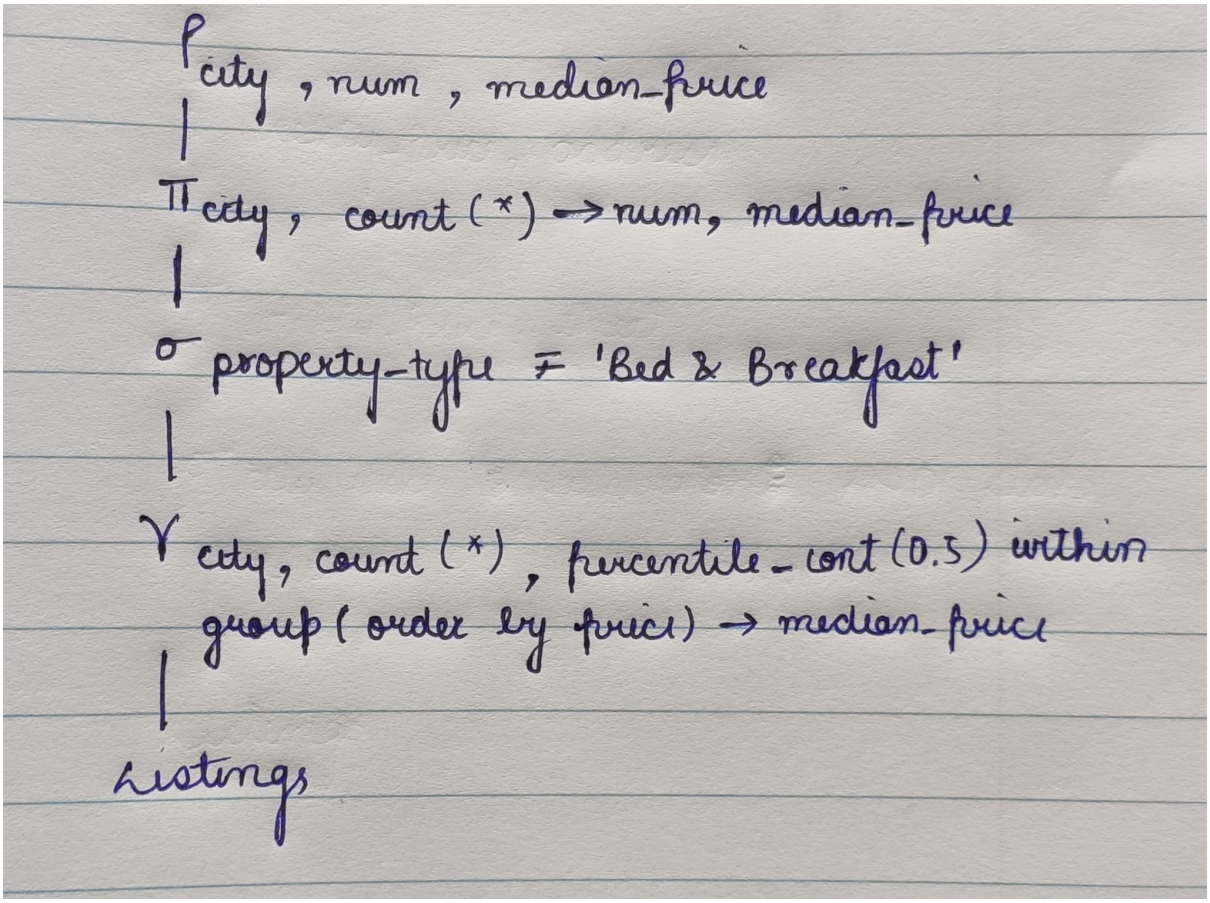
Assuming that City + State covers all the ‘Washington DC’ specific apartment and 1BHK means 1 bedroom and weekly_price is not NULL.



c) **How many bed and breakfast places are there in each city and what are the median prices?**

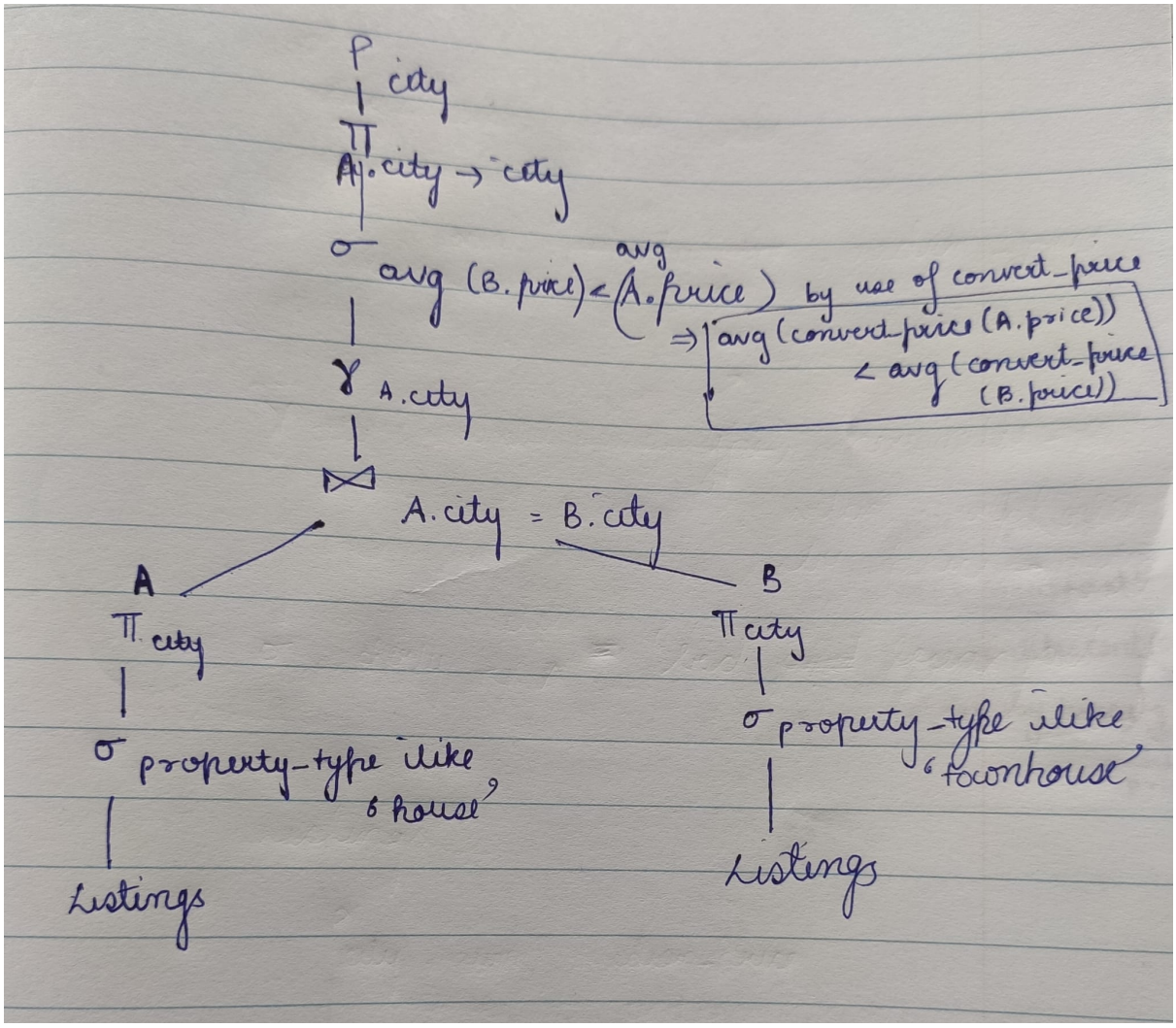
```
c)  select city, count(*) as num, PERCENTILE_CONT(0.5) WITHIN GROUP(ORDER
BY ltrim(regexp_replace(price,',',' '), '$'):: numeric ) as median_price
from "Airbnb"."Listings" group by city, property_type having property_type
= 'Bed & Breakfast'
```

Assumption: Here we take the values of cities which don't have null values for 'Bed & Breakfast' places



d) **In which cities are houses cheaper than townhouses on an average? Note that the answer can be "None".**

```
 CREATE FUNCTION convert_price(text) RETURNS numeric
    AS 'SELECT ltrim(replace($1, '',','''''), ''$'')::numeric;'
    LANGUAGE sql
    IMMUTABLE
    RETURNS NULL ON NULL INPUT ;
select A.city from "Airbnb"."Listings" A, "Airbnb"."Listings" B where
A.property_type ilike 'Townhouse' AND B.property_type ilike 'House' AND
A.city = B.city group by A.city HAVING avg(convert_price(B.price) )<
avg(convert_price(A.price));
```



3) Inspect the schema for Airbnb listings. Suggest at least 3 improvements of the schema.

1) Usage of proper data types
2) Eliminate unnecessary attributes like multiple columns containing addresses of the same place
3) There are certain columns which are like city and state which are included in column called street, so it's a repeat