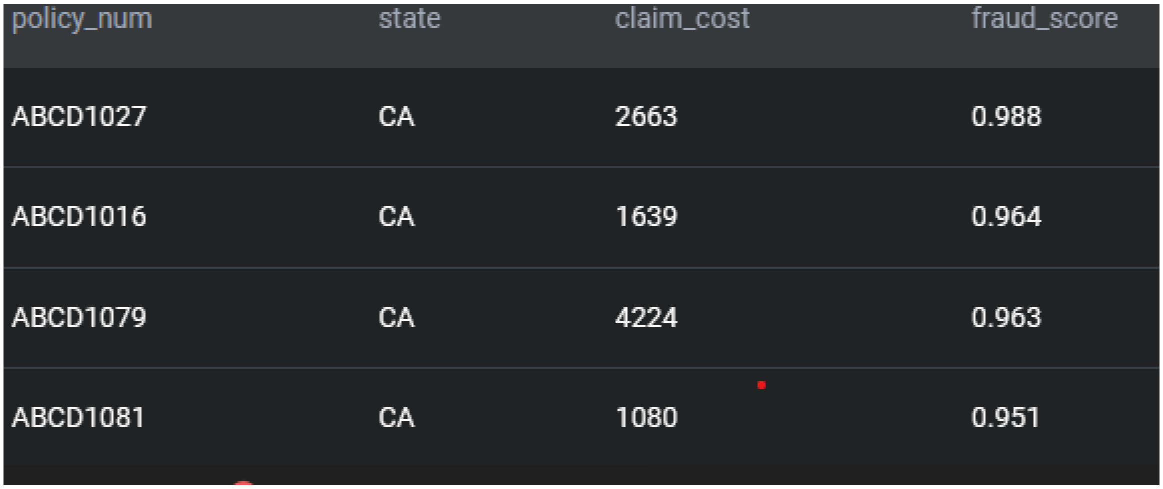
Strata hard questions

ABC Corp is a mid-sized insurer in the US and in the recent past their fraudulent claims have increased significantly for their personal auto insurance portfolio. They have developed a ML based predictive model to identify propensity of fraudulent claims. Now, they assign highly experienced claim adjusters for top 5 percentile of claims identified by the model. Your objective is to identify the top 5 percentile of claims from each state. Your output should be policy number, state, claim cost, and fraud score.

```
-- q1. Top Percentile Fraud(google)
select
policy_num, state, claim_cost, fraud_score
from
(select *
, ntile(100) over (partition by state order by fraud_score DESC) AS percentile
from fraud_score) as t
where percentile <= 5;
```

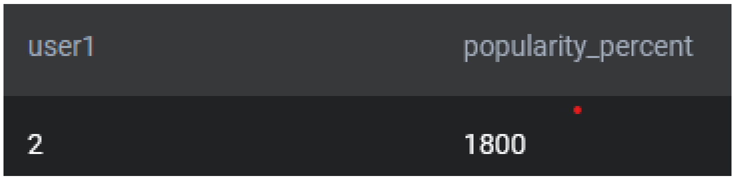


policy_num	state	claim_cost	fraud_score
ABCD1027	CA	2663	0.988
ABCD1016	CA	1639	0.964
ABCD1079	CA	4224	0.963
ABCD1081	CA	1080	0.951

Popularity Percentage

Find the popularity percentage for each user on Meta/Facebook. The popularity percentage is defined as the total number of friends the user has divided by the total number of users on the platform, then converted into a percentage by multiplying by 100. Output each user along with their popularity percentage. Order records in ascending order by user id. The 'user1' and 'user2' column are pairs of friends.

```
WITH cte AS (
    SELECT
        user1,
        COUNT(user2) AS friend
    FROM facebook_friends
    GROUP BY user1
    UNION
    SELECT
        user2,
        COUNT(user1) AS friend
    FROM facebook_friends
    GROUP BY user2
)

SELECT
    user1,
    (total_dost / total_pu) * 100 AS popularity_percent
FROM (
    SELECT
        user1,
        SUM(friend) AS total_dost,
        COUNT(user1) OVER () AS total_pu
    FROM cte
) AS t;
```



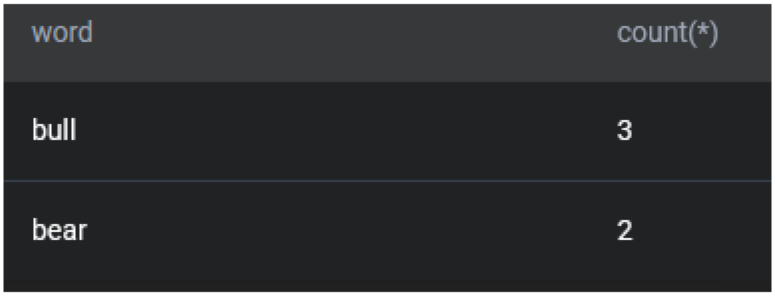
user1	popularity_percent
2	1800

Counting Instances in Text

Find the number of times the words 'bull' and 'bear' occur in the contents. We're counting the number of times the words occur so words like 'bullish' should not be included in our count. Output the word 'bull' and 'bear' along with the corresponding number of occurrences.

draft1.txt	The stock exchange predicts a bull market which would make many investors happy.
draft2.txt	The stock exchange predicts a bull market which would make many investors happy, but analysts warn of possibility of too much optimism and that in fact we are awaiting a bear market.
final.txt	The stock exchange predicts a bull market which would make many investors happy, but analysts warn of possibility of too much optimism and that in fact we are awaiting a bear market. As always predicting the future market is an uncertain game and all investors should follow their instincts and best practices.

```
(select

'bull' as word,

count(*)

from google_file_store

where contents like '%bull%'

group by word)

union all

(select

'bear' as word,

count(*)

from google_file_store

where contents like '%bear%'

group by word);
```



word	count(*)
bull	3
bear	2

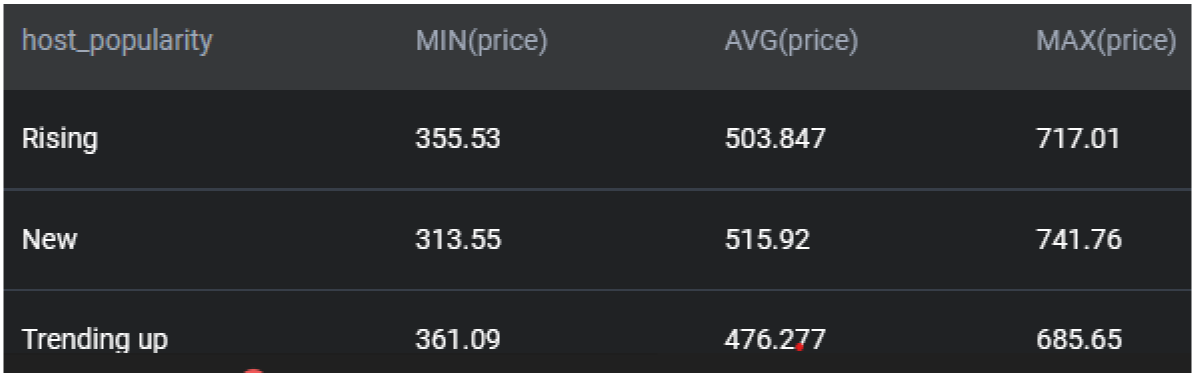
Host Popularity Rental Prices

You’re given a table of rental property searches by users. The table consists of search results and outputs host information for searchers. Find the minimum, average, maximum rental prices for each host’s popularity rating. The host’s popularity rating is defined as below: 0 reviews: New 1 to 5 reviews: Rising 6 to 15 reviews: Trending Up 16 to 40 reviews: Popular more than 40 reviews: Hot

Tip: The `id` column in the table refers to the search ID. You'll need to create your own host_id by concating price, room_type, host_since, zipcode, and number_of_reviews.

Output host popularity rating and their minimum, average and maximum rental prices.

```
-- q4. Host popularity rental prices(airbnb)
WITH cte1 AS (
    SELECT
        CONCAT(price, room_type, host_since, zipcode, number_of_reviews) AS host_id,
        number_of_reviews,
        price
    FROM airbnb_host_searches
),
cte2 AS (
    SELECT DISTINCT
        host_id,
        number_of_reviews,
        price,
        CASE
            WHEN number_of_reviews = 0  THEN 'New'
            WHEN number_of_reviews BETWEEN 1 AND 5 THEN 'Rising'
            WHEN number_of_reviews BETWEEN 6 AND 15 THEN 'Trending
up'
            WHEN number_of_reviews BETWEEN 16 AND 40 THEN
'Popular'
            WHEN number_of_reviews > 40 THEN 'Hot'
        END AS host_popularity
    FROM cte1
)

SELECT
    host_popularity,
    MIN(price),
    AVG(price),
    MAX(price)
FROM cte2
GROUP BY 1;
```



host_popularity	MIN(price)	AVG(price)	MAX(price)
Rising	355.53	503.847	717.01
New	313.55	515.92	741.76
Trending up	361.09	476.277	685.65

Chelsea

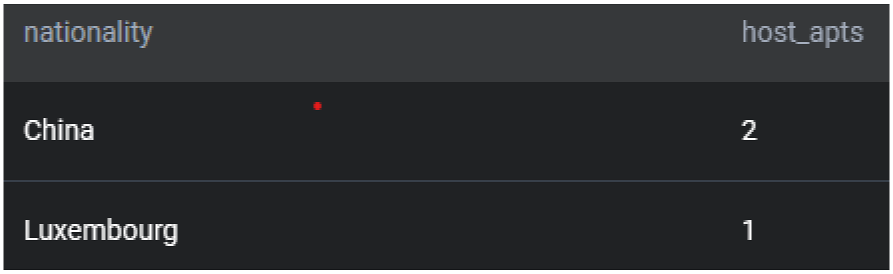
Number Of Units Per Nationality

Find the number of apartments per nationality that are owned by people under 30 years old.

Output the nationality along with the number of apartments.

Sort records by the apartments count in descending order.

```
select * from airbnb_hosts; -- (host_id), nationality, gender, age
select * from airbnb_units; --  (host_id), unit_id, unit_type, n_beds n_bedrooms, country, city
SELECT
h.nationality
, COUNT(DISTINCT unit_id) AS host_apts
FROM airbnb_units u
JOIN airbnb_hosts h
ON u.host_id = h.host_id
WHERE h.age < 30
    AND u.unit_type = 'Penthouse'
GROUP BY h.nationality
order by host_apts DESC;
```



nationality	host_apts
China	2
Luxembourg	1