Part 1: Yelp Dataset Profiling and Understanding

1. Profile the data by finding the total number of records for each of the tables below:

i. Attribute table = 10000

```
SELECT count(*) FROM Attribute
```

ii. Business table =10000

```
SELECT count(*) FROM Business
```

iii. Category table =10000

```
SELECT count(*) FROM Category
```

iv. Checkin table =10000

```
SELECT count(*) FROM Checkin
```

v. elite_years table =10000

```
SELECT count(*) FROM elite_years
```

vi. friend table = 10000

```
SELECT count(*) FROM friend
```

vii. hours table =10000

```
SELECT count(*) FROM hours
```

viii. photo table = 10000

```
SELECT count(*) FROM photo
```

ix. review table = 10000

```
SELECT count(*) FROM review
```

x. tip table = 10000

```
SELECT count(*) FROM tip
```

xi. user table =10000

```
SELECT count(*) FROM user
```

2. Find the total distinct records by either the foreign key or primary key for each table. If two foreign keys are listed in the table, please specify which foreign key.

i. Business = id, 10000

```
SELECT COUNT(DISTINCT id) FROM Business
```

ii. Hours = Business_id, 1532

```
SELECT COUNT(DISTINCT Business_id) FROM hours
```

iii. Category = business_id, 2643

```
SELECT COUNT(DISTINCT Business_id) FROM Category
```

iv. Attribute = business_id, 1115

```
SELECT COUNT(DISTINCT Business_id) FROM Attribute
```

v. Review = business_id, 8090

```
SELECT COUNT(DISTINCT business_id) FROM Review
```

vi. Checkin = business_id 493

```
SELECT COUNT(DISTINCT business_id) FROM Checkin
```

vii. Photo = business_id   6493

```
SELECT COUNT(DISTINCT business_id) FROM Photo
```

viii. Tip = user_id 537

```
SELECT COUNT(DISTINCT user_id) FROM Tip
```

ix. User = id 10000

```
SELECT COUNT(DISTINCT id) FROM User
```

x. Friend = user_id 11

```
SELECT COUNT(DISTINCT user_id) FROM Friend
```

xi. Elite_years = user_id 2780

```
SELECT COUNT(DISTINCT user_id) FROM Elite_years
```

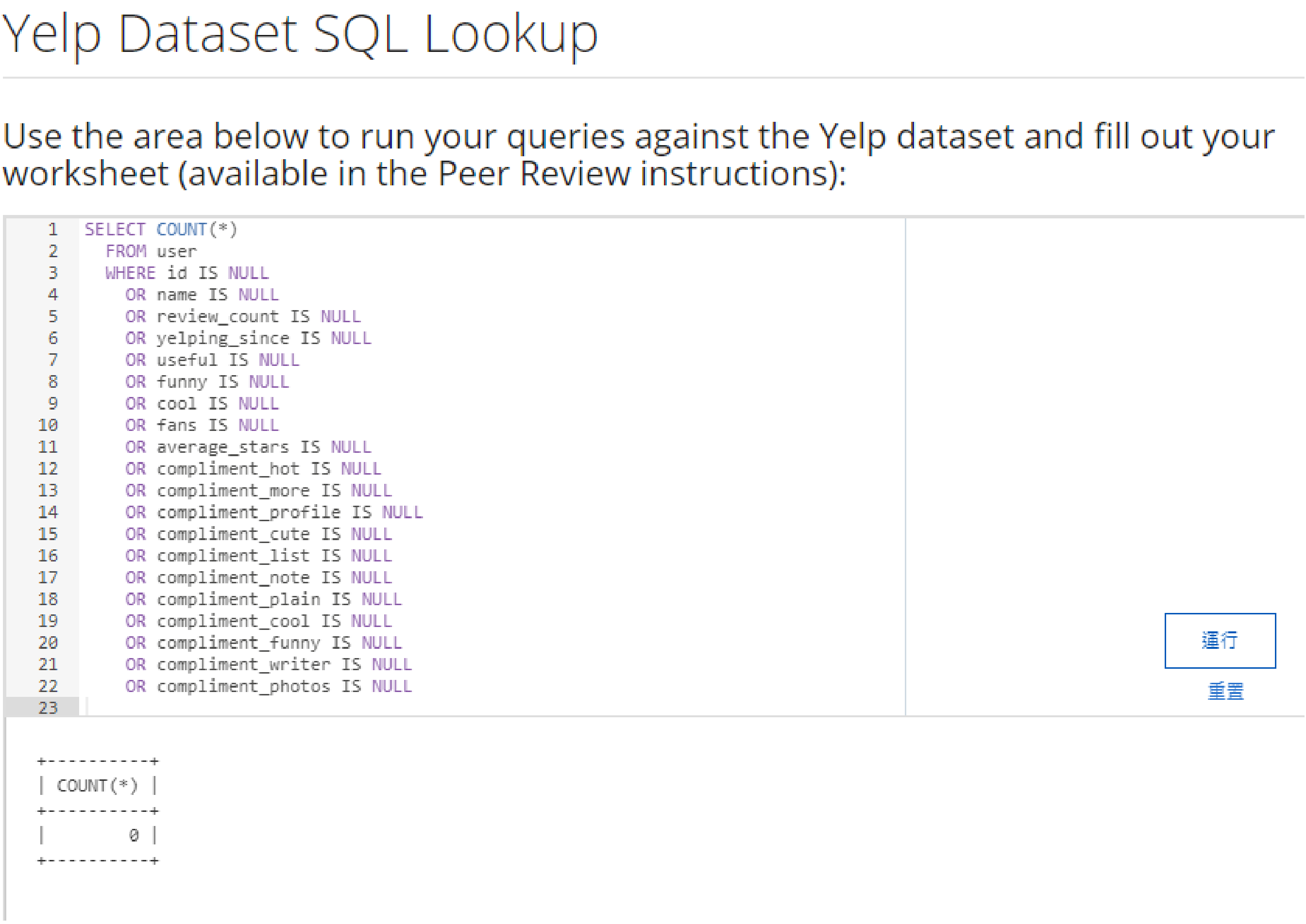
3. Are there any columns with null values in the Users table? Indicate "yes," or "no."

Answer:

no

SQL code used to arrive at answer:

```
SELECT COUNT(*)
FROM user
WHERE id IS NULL
        OR name IS NULL
        OR review_count IS NULL
        OR yelping_since IS NULL
        OR useful IS NULL
        OR funny IS NULL
        OR cool IS NULL
        OR fans IS NULL
        OR average_stars IS NULL
        OR compliment_hot IS NULL
        OR compliment_more IS NULL
        OR compliment_profile IS NULL
        OR compliment_cute IS NULL
        OR compliment_list IS NULL
        OR compliment_note IS NULL
        OR compliment_plain IS NULL
        OR compliment_cool IS NULL
        OR compliment_funny IS NULL
        OR compliment_writer IS NULL
        OR compliment_photos IS NULL
```



Yelp Dataset SQL Lookup

Use the area below to run your queries against the Yelp dataset and fill out your worksheet (available in the Peer Review instructions):

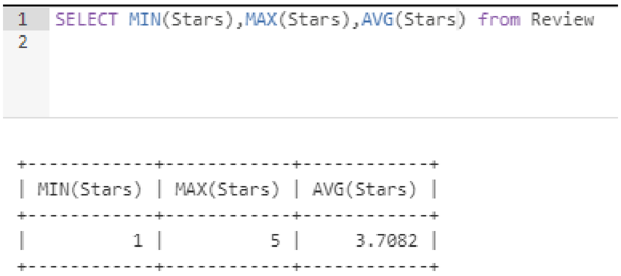
```
SELECT COUNT(*)
  FROM user
  WHERE id IS NULL
    OR name IS NULL
    OR review_count IS NULL
    OR yelping_since IS NULL
    OR useful IS NULL
    OR funny IS NULL
    OR cool IS NULL
    OR fans IS NULL
    OR average_stars IS NULL
    OR compliment_hot IS NULL
    OR compliment_more IS NULL
    OR compliment_profile IS NULL
    OR compliment_cute IS NULL
    OR compliment_list IS NULL
    OR compliment_note IS NULL
    OR compliment_plain IS NULL
    OR compliment_cool IS NULL
    OR compliment_funny IS NULL
    OR compliment_writer IS NULL
    OR compliment_photos IS NULL
```

```
+----------+
| COUNT(*) |
+----------+
|        0 |
+----------+
```

4. For each table and column listed below, display the smallest (minimum), largest (maximum), and average (mean) value for the following fields:

i. Table: Review, Column: Stars

min: 1 max:5 avg: 3.7082

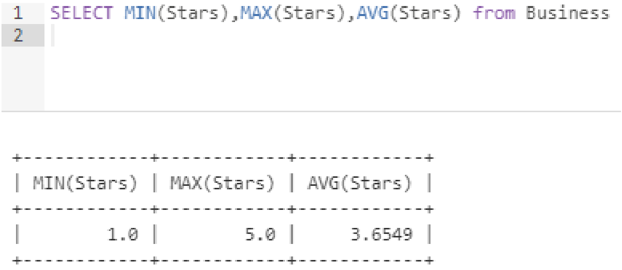


```
SELECT MIN(Stars),MAX(Stars),AVG(Stars) from Review
```

```
+------------+------------+------------+
| MIN(Stars) | MAX(Stars) | AVG(Stars) |
+------------+------------+------------+
|          1 |          5 |     3.7082 |
+------------+------------+------------+
```

ii. Table: Business, Column: Stars

min: 1 max:5 avg: 3.6549

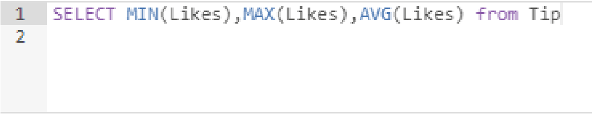


```
SELECT MIN(Stars),MAX(Stars),AVG(Stars) from Business
```

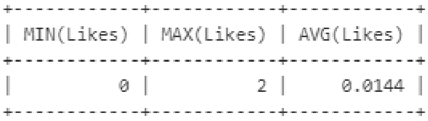
```
+------------+------------+------------+
| MIN(Stars) | MAX(Stars) | AVG(Stars) |
+------------+------------+------------+
|        1.0 |        5.0 |     3.6549 |
+------------+------------+------------+
```

iii. Table: Tip, Column: Likes

min:0 max:2 avg: 0.0144



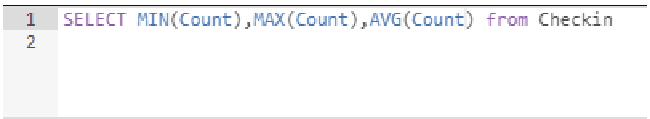
```
SELECT MIN(Likes),MAX(Likes),AVG(Likes) from Tip
```



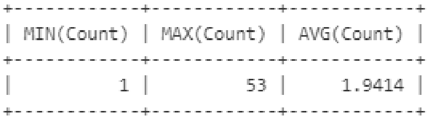
```
+------------+------------+------------+
| MIN(Likes) | MAX(Likes) | AVG(Likes) |
+------------+------------+------------+
|          0 |          2 |     0.0144 |
+------------+------------+------------+
```

iv. Table: Checkin, Column: Count

min: 1 max: 53 avg: 1.9414



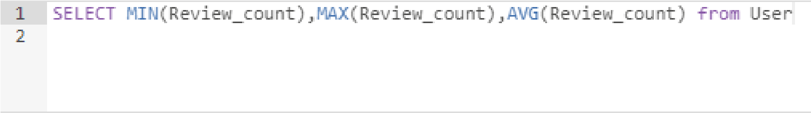
```
SELECT MIN(Count),MAX(Count),AVG(Count) from Checkin
```



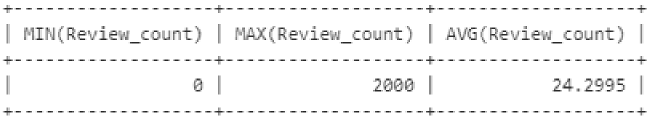
```
+------------+------------+------------+
| MIN(Count) | MAX(Count) | AVG(Count) |
+------------+------------+------------+
|          1 |         53 |     1.9414 |
+------------+------------+------------+
```

v. Table: User, Column: Review_count

min:0 max: 2000 avg: 24.2995



```
SELECT MIN(Review_count),MAX(Review_count),AVG(Review_count) from User
```



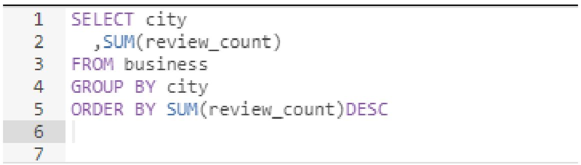
```
+-------------------+-------------------+-------------------+
| MIN(Review_count) | MAX(Review_count) | AVG(Review_count) |
+-------------------+-------------------+-------------------+
|                 0 |              2000 |           24.2995 |
+-------------------+-------------------+-------------------+
```

5. List the cities with the most reviews in descending order:

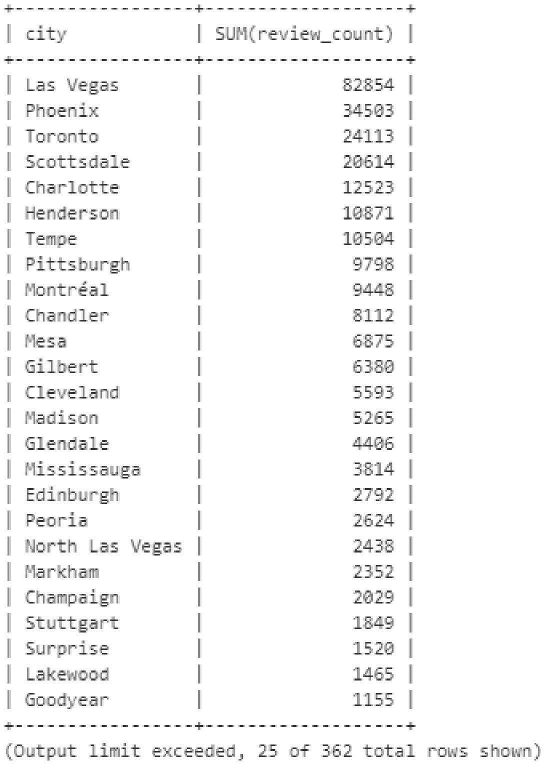
SQL code used to arrive at answer:

```
SELECT city
  ,SUM(review_count)
FROM business
GROUP BY city
ORDER BY SUM(review_count)DESC
```

Copy and Paste the Result Below:



```
SELECT city
  ,SUM(review_count)
FROM business
GROUP BY city
ORDER BY SUM(review_count)DESC
```



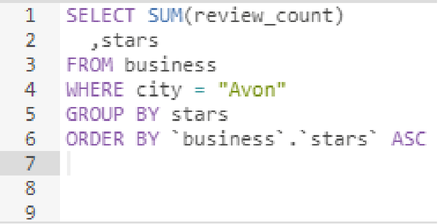
```
+-----------------+-------------------+
| city            | SUM(review_count) |
+-----------------+-------------------+
| Las Vegas       |             82854 |
| Phoenix         |             34503 |
| Toronto         |             24113 |
| Scottsdale      |             20614 |
| Charlotte       |             12523 |
| Henderson       |             10871 |
| Tempe           |             10504 |
| Pittsburgh      |              9798 |
| Montréal        |              9448 |
| Chandler        |              8112 |
| Mesa            |              6875 |
| Gilbert         |              6380 |
| Cleveland       |              5593 |
| Madison         |              5265 |
| Glendale        |              4406 |
| Mississauga     |              3814 |
| Edinburgh       |              2792 |
| Peoria          |              2624 |
| North Las Vegas |              2438 |
| Markham         |              2352 |
| Champaign       |              2029 |
| Stuttgart       |              1849 |
| Surprise        |              1520 |
| Lakewood        |              1465 |
| Goodyear        |              1155 |
+-----------------+-------------------+
(Output limit exceeded, 25 of 362 total rows shown)
```

6. Find the distribution of star ratings to the business in the following cities:

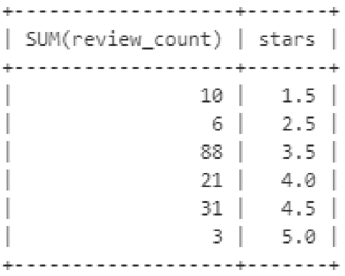
i. Avon

```
SELECT SUM(review_count)

    ,stars

FROM business

WHERE city = "Avon"

GROUP BY stars

ORDER BY `business`.`stars` ASC
```

SQL code used to arrive at answer:



```
SELECT SUM(review_count)
  ,stars
FROM business
WHERE city = "Avon"
GROUP BY stars
ORDER BY `business`.`stars` ASC
```



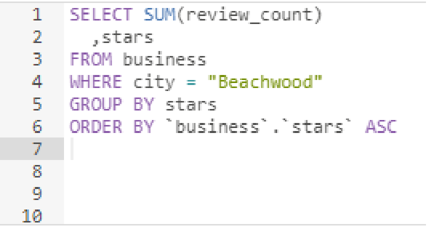
```
+-------------------+-------+
| SUM(review_count) | stars |
+-------------------+-------+
|                10 |   1.5 |
|                 6 |   2.5 |
|                88 |   3.5 |
|                21 |   4.0 |
|                31 |   4.5 |
|                 3 |   5.0 |
+-------------------+-------+
```

Copy and Paste the Resulting Table Below (2 columns �� star rating and count):

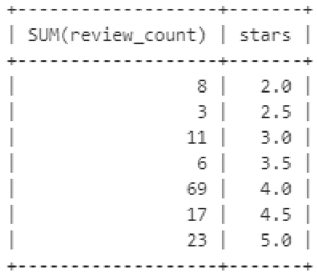
ii. Beachwood

```
SELECT SUM(review_count)
    ,stars
FROM business
WHERE city = "Beachwood"
GROUP BY stars
ORDER BY `business`.`stars` ASC
```

SQL code used to arrive at answer:



```
SELECT SUM(review_count)
  ,stars
FROM business
WHERE city = "Beachwood"
GROUP BY stars
ORDER BY `business`.`stars` ASC
```



```
+-------------------+-------+
| SUM(review_count) | stars |
+-------------------+-------+
|                 8 |   2.0 |
|                 3 |   2.5 |
|                11 |   3.0 |
|                 6 |   3.5 |
|                69 |   4.0 |
|                17 |   4.5 |
|                23 |   5.0 |
+-------------------+-------+
```

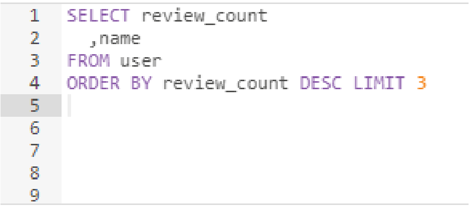
Copy and Paste the Resulting Table Below (2 columns �� star rating and count):

7. Find the top 3 users based on their total number of reviews:

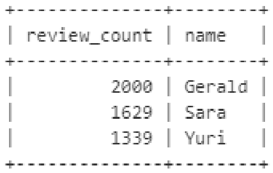
SQL code used to arrive at answer:

```
SELECT review_count
    ,name
FROM user
ORDER BY review_count DESC LIMIT 3
```

Copy and Paste the Result Below:



```
SELECT review_count
  ,name
FROM user
ORDER BY review_count DESC LIMIT 3
```

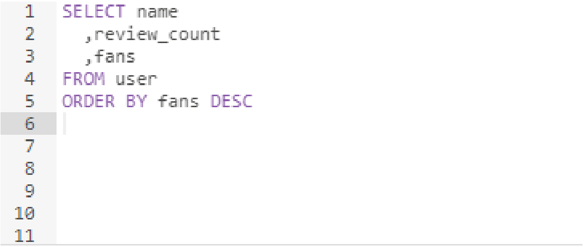


```
+--------------+--------+
| review_count | name   |
+--------------+--------+
|         2000 | Gerald |
|         1629 | Sara   |
|         1339 | Yuri   |
+--------------+--------+
```

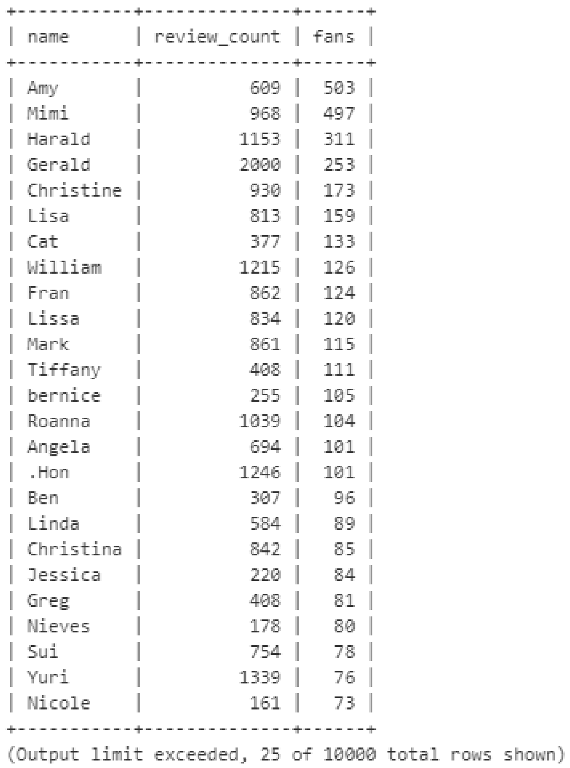
8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

```
SELECT name
    ,review_count
    ,fans
FROM user
ORDER BY fans DESC
```



```
SELECT name
  ,review_count
  ,fans
FROM user
ORDER BY fans DESC
```



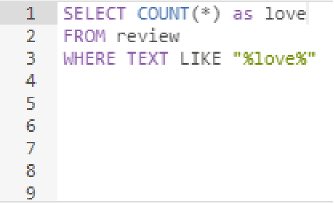
```
+-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Amy       |          609 |  503 |
| Mimi      |          968 |  497 |
| Harald    |         1153 |  311 |
| Gerald    |         2000 |  253 |
| Christine |          930 |  173 |
| Lisa      |          813 |  159 |
| Cat       |          377 |  133 |
| William   |         1215 |  126 |
| Fran      |          862 |  124 |
| Lissa     |          834 |  120 |
| Mark      |          861 |  115 |
| Tiffany   |          408 |  111 |
| bernice   |          255 |  105 |
| Roanna    |         1039 |  104 |
| Angela    |          694 |  101 |
| .Hon      |         1246 |  101 |
| Ben       |          307 |   96 |
| Linda     |          584 |   89 |
| Christina |          842 |   85 |
| Jessica   |          220 |   84 |
| Greg      |          408 |   81 |
| Nieves    |          178 |   80 |
| Sui       |          754 |   78 |
| Yuri      |         1339 |   76 |
| Nicole    |          161 |   73 |
+-----------+--------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

9. Are there more reviews with the word "love" or with the word "hate" in them?

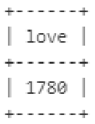
Answer:

```
SELECT COUNT(*)
FROM review
WHERE TEXT LIKE "%love%"

SELECT COUNT(*)
FROM review
WHERE TEXT LIKE "%hate%"
```

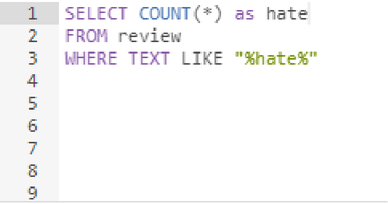
SQL code used to arrive at answer:



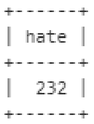
```
SELECT COUNT(*) as love
FROM review
WHERE TEXT LIKE "%love%"
```



```
+------+
| love |
+------+
| 1780 |
+------+
```



```
SELECT COUNT(*) as hate
FROM review
WHERE TEXT LIKE "%hate%"
```



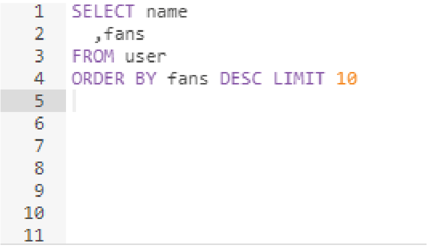
```
+------+
| hate |
+------+
|  232 |
+------+
```

10. Find the top 10 users with the most fans:

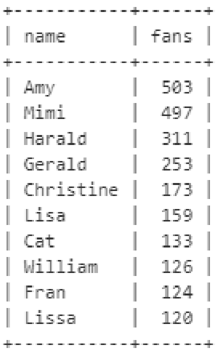
SQL code used to arrive at answer:

```
SELECT name
	,fans
FROM user
ORDER BY fans DESC LIMIT 10
```

Copy and Paste the Result Below:



```
SELECT name
  ,fans
FROM user
ORDER BY fans DESC LIMIT 10
```



```
+-----------+------+
| name      | fans |
+-----------+------+
| Amy       |  503 |
| Mimi      |  497 |
| Harald    |  311 |
| Gerald    |  253 |
| Christine |  173 |
| Lisa      |  159 |
| Cat       |  133 |
| William   |  126 |
| Fran      |  124 |
| Lissa     |  120 |
+-----------+------+
```

Part 2: Inferences and Analysis

1. Pick one city and category of your choice and group the businesses in that city or category by their overall star rating. Compare the businesses with 2-3 stars to the businesses with 4-5 stars and answer the following questions. Include your code.

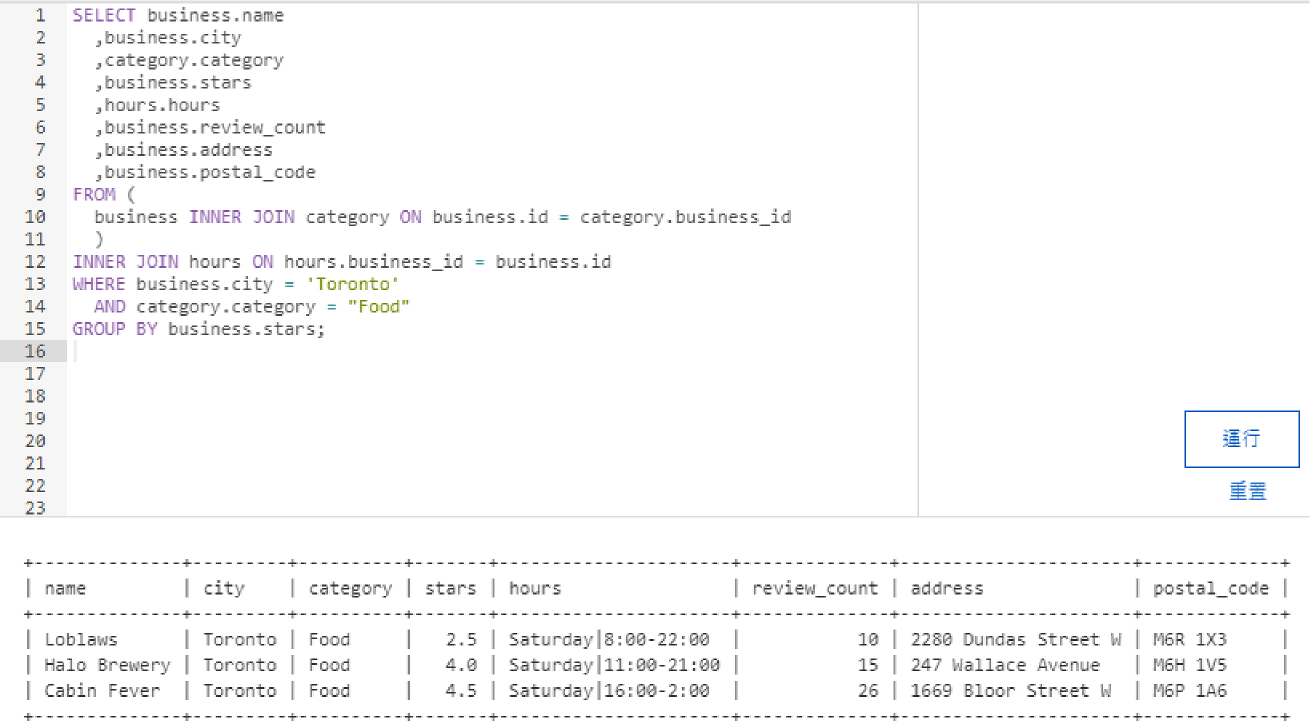
i. Do the two groups you chose to analyze have a different distribution of hours?
Yes ,they do have different distribution hours.   for the restaurant category , the one with 2-3 star ratings operates for longer hoursthan the one with 4-5 star ratings.

ii. Do the two groups you chose to analyze have a different number of reviews?
They have different number of reviews.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
Two groups had different zipcodes, I wasn't able to infer anything

SQL code used for analysis:

```
SELECT business.name
        ,business.city
        ,category.category
        ,business.stars
        ,hours.hours
        ,business.review_count
        ,business.address
        ,business.postal_code
FROM (
        business INNER JOIN category ON business.id = category.business_id
        )
INNER JOIN hours ON hours.business_id = business.id
WHERE business.city = 'Toronto'
        AND category.category = "Food"
GROUP BY business.stars;
```



```
SELECT business.name
  ,business.city
  ,category.category
  ,business.stars
  ,hours.hours
  ,business.review_count
  ,business.address
  ,business.postal_code
FROM (
  business INNER JOIN category ON business.id = category.business_id
  )
INNER JOIN hours ON hours.business_id = business.id
WHERE business.city = 'Toronto'
  AND category.category = "Food"
GROUP BY business.stars;
```

```
+--------------+---------+----------+-------+----------------------+--------------+----------------------+-------------+
| name         | city    | category | stars | hours                | review_count | address              | postal_code |
+--------------+---------+----------+-------+----------------------+--------------+----------------------+-------------+
| Loblaws      | Toronto | Food     |   2.5 | Saturday|8:00-22:00  |           10 | 2280 Dundas Street W | M6R 1X3     |
| Halo Brewery | Toronto | Food     |   4.0 | Saturday|11:00-21:00 |           15 | 247 Wallace Avenue   | M6H 1V5     |
| Cabin Fever  | Toronto | Food     |   4.5 | Saturday|16:00-2:00  |           26 | 1669 Bloor Street W  | M6P 1A6     |
+--------------+---------+----------+-------+----------------------+--------------+----------------------+-------------+
```

2. Group business based on the ones that are open and the ones that are closed. What differences can you find between the ones that are still open and the ones that are closed? List at least two differences and the SQL code you used to arrive at your answer.

i. Difference 1:

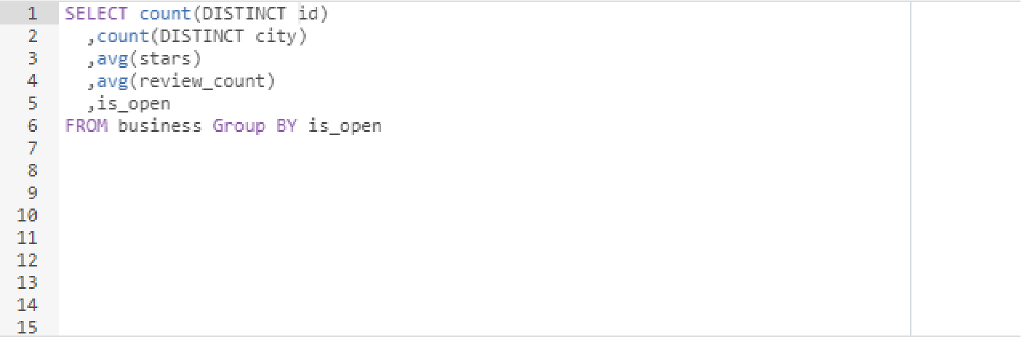
The average review count was 9 points more for business that are open

ii. Difference 2:

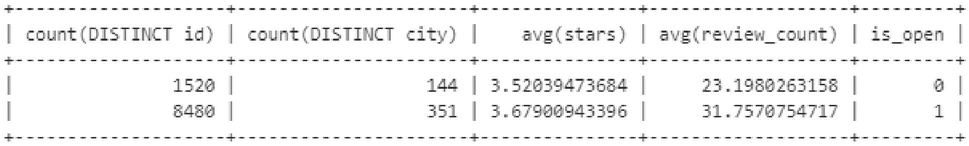
The number of distinct business_id for the one that are open times more than the business that are closed and hence the average review count is higher for the business that are open

SQL code used for analysis:

```
SELECT count(DISTINCT business_id)
        ,count(DISTINCT city)
        ,avg(stars)
        ,avg(review_count)
        ,is_open
FROM business Group BY is_open
```



```
SELECT count(DISTINCT id)
  ,count(DISTINCT city)
  ,avg(stars)
  ,avg(review_count)
  ,is_open
FROM business Group BY is_open
```



```
+--------------------+----------------------+---------------+-------------------+---------+
| count(DISTINCT id) | count(DISTINCT city) |    avg(stars) | avg(review_count) | is_open |
+--------------------+----------------------+---------------+-------------------+---------+
|               1520 |                  144 | 3.52039473684 |     23.1980263158 |       0 |
|               8480 |                  351 | 3.67900943396 |     31.7570754717 |       1 |
+--------------------+----------------------+---------------+-------------------+---------+
```

3. For this last part of your analysis, you are going to choose the type of analysis you want to conduct on the Yelp dataset and are going to prepare the data for analysis.

Ideas for analysis include: Parsing out keywords and business attributes for sentiment analysis, clustering businesses to find commonalities or anomalies between them, predicting the overall star rating for a business, predicting the number of fans a user will have, and so on. These are just a few examples to get you started, so feel free to be creative and come up with your own problem you want to solve. Provide answers, in-line, to all of the following:

i. Indicate the type of analysis you chose to do:
The businesses like restaurants having the arrtibutes like 'goodforkids' ,'alcohol' and 'free wifi' anyway relate to the number of stars or the review counts has more number of restaurants, has the review counts ranging from the least to the highest and the ratings from 2 to 4.5 stars ffor my analysis.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

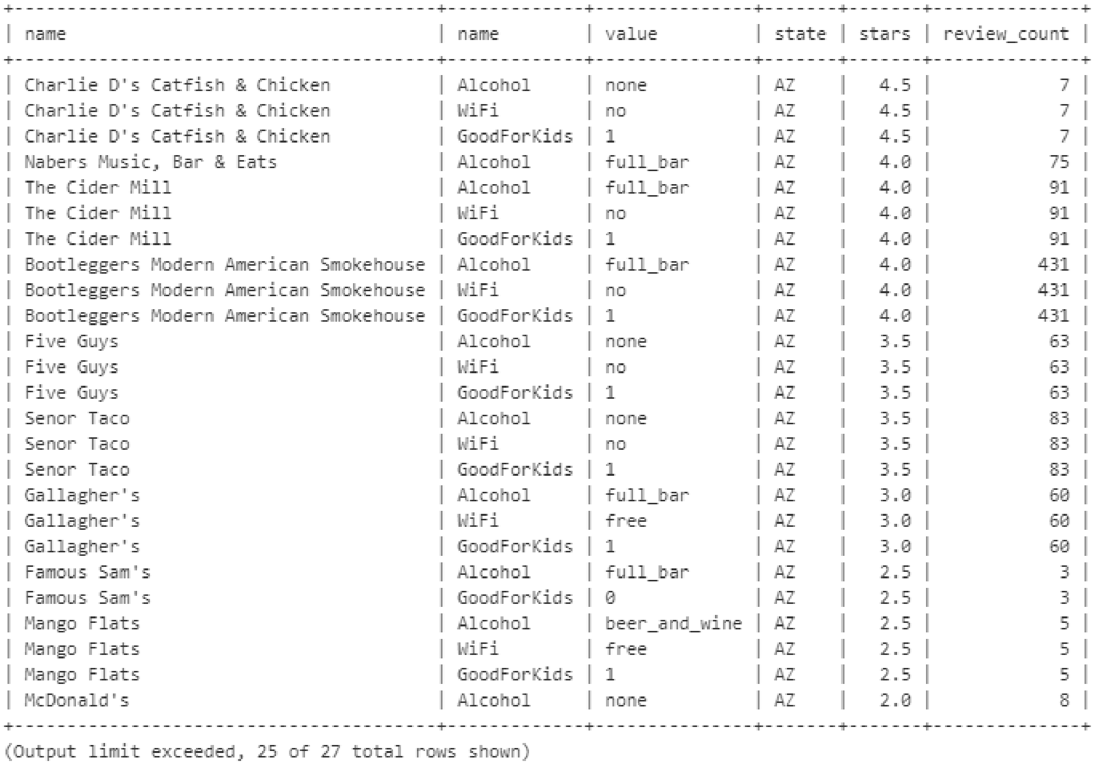
Use 3 tables like business, category and attribute for analysis.
the business name ,their catagory , the state in which they are run , the attributes they have, their ratings and their count of reviews. I took varibles like

1)name, state, stars , review_count from the table
2)category from the table
3)name ,value from the attribute table.
To connect all the 3 tables

Having a free wifi or restaurents good for kids or having full-bar or having any combination of 2 or all the attributes contributes to good rating or having more reviews in particular.

iii. Output of your finished dataset:

```
SELECT business.name
        ,attribute.name
        ,attribute.value
        ,business.STATE
        ,business.stars
        ,business.review_count
FROM business
INNER JOIN category ON category.business_id = business.id
INNER JOIN attribute ON attribute.business_id = business.id
WHERE (
                attribute.name LIKE 'alcohol'
                OR attribute.name LIKE 'wifi'
                OR attribute.name LIKE 'goodforkids'
                )
        AND category = 'Restaurants'
        AND business.STATE = 'AZ'
ORDER BY stars DESC
        ,review_count
```



```
+----------------------------------------+------------+-----------------+-------+-------+--------------+
| name                                   | name       | value           | state | stars | review_count |
+----------------------------------------+------------+-----------------+-------+-------+--------------+
| Charlie D's Catfish & Chicken          | Alcohol    | none            | AZ    |   4.5 |            7 |
| Charlie D's Catfish & Chicken          | WiFi       | no              | AZ    |   4.5 |            7 |
| Charlie D's Catfish & Chicken          | GoodForKids | 1              | AZ    |   4.5 |            7 |
| Nabers Music, Bar & Eats               | Alcohol    | full_bar        | AZ    |   4.0 |           75 |
| The Cider Mill                         | Alcohol    | full_bar        | AZ    |   4.0 |           91 |
| The Cider Mill                         | WiFi       | no              | AZ    |   4.0 |           91 |
| The Cider Mill                         | GoodForKids | 1              | AZ    |   4.0 |           91 |
| Bootleggers Modern American Smokehouse | Alcohol    | full_bar        | AZ    |   4.0 |          431 |
| Bootleggers Modern American Smokehouse | WiFi       | no              | AZ    |   4.0 |          431 |
| Bootleggers Modern American Smokehouse | GoodForKids | 1              | AZ    |   4.0 |          431 |
| Five Guys                              | Alcohol    | none            | AZ    |   3.5 |           63 |
| Five Guys                              | WiFi       | no              | AZ    |   3.5 |           63 |
| Five Guys                              | GoodForKids | 1              | AZ    |   3.5 |           63 |
| Senor Taco                             | Alcohol    | none            | AZ    |   3.5 |           83 |
| Senor Taco                             | WiFi       | no              | AZ    |   3.5 |           83 |
| Senor Taco                             | GoodForKids | 1              | AZ    |   3.5 |           83 |
| Gallagher's                            | Alcohol    | full_bar        | AZ    |   3.0 |           60 |
| Gallagher's                            | WiFi       | free            | AZ    |   3.0 |           60 |
| Gallagher's                            | GoodForKids | 1              | AZ    |   3.0 |           60 |
| Famous Sam's                           | Alcohol    | full_bar        | AZ    |   2.5 |            3 |
| Famous Sam's                           | GoodForKids | 0              | AZ    |   2.5 |            3 |
| Mango Flats                            | Alcohol    | beer_and_wine   | AZ    |   2.5 |            5 |
| Mango Flats                            | WiFi       | free            | AZ    |   2.5 |            5 |
| Mango Flats                            | GoodForKids | 1              | AZ    |   2.5 |            5 |
| McDonald's                             | Alcohol    | none            | AZ    |   2.0 |            8 |
+----------------------------------------+------------+-----------------+-------+-------+--------------+
(Output limit exceeded, 25 of 27 total rows shown)
```