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 = 10,000
ii. Business table = 10,000
iii. Category table = 10,000
iv. Checkin table = 10,000
v. elite_years table = 10,000
vi. friend table = 10,000
vii. hours table = 10,000
viii. photo table = 10,000
ix. review table = 10,000
x. tip table = 10,000
xi. user table = 10,000

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 = 10,000
ii. Hours = 1562
iii. Category = 2643
iv. Attribute = 1115
v. Review = 10,000
vi. Checkin = 493
vii. Photo = 10,000
viii. Tip = 537 (Foreign key = user_id)
ix. User = 10,000
x. Friend = 11
xi. Elite_years = 2780

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

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 *
FROM user
WHERE compliment_photos IS NULL;
```

(Changed the column name)

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

ii. Table: Business, Column: Stars

min:1.0 max:5.0 avg: 3.6549

iii. Table: Tip, Column: Likes

min:0 max:2 avg: 0.0144

iv. Table: Checkin, Column: Count

min:1 max:53 avg: 1.9414

v. Table: User, Column: Review_count

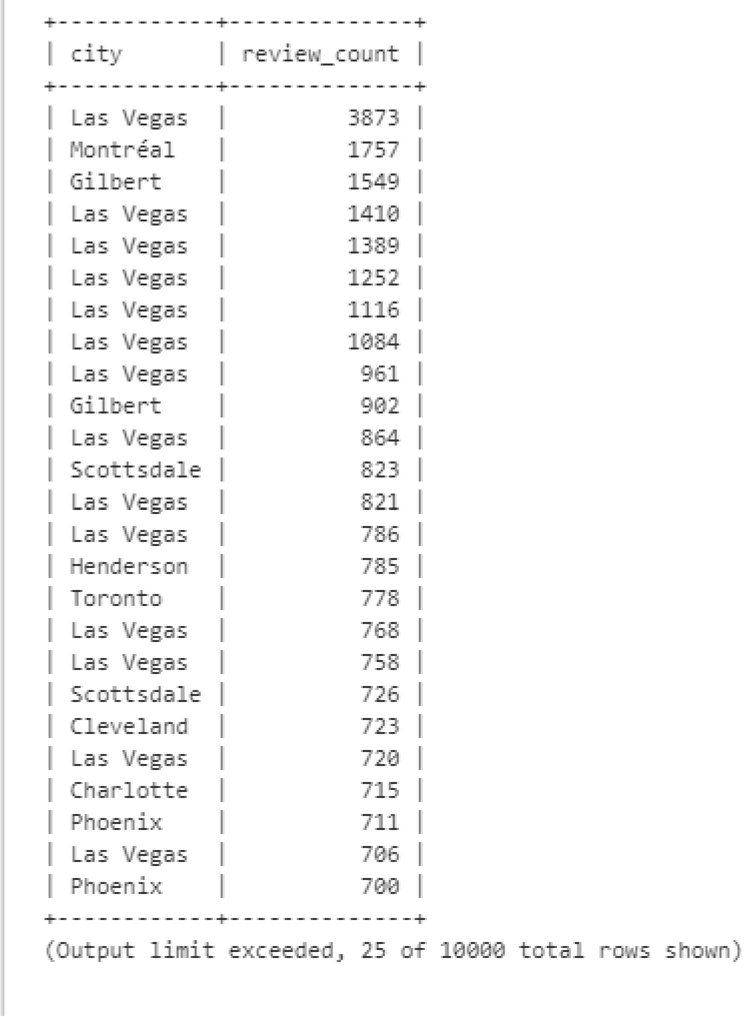
min:0 max:2000 avg: 24.2995

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

SQL code used to arrive at answer:

```
SELECT city
,review_count
FROM business
ORDER BY review_count desc;
```

Copy and Paste the Result Below:



```
+------------+--------------+
| city       | review_count |
+------------+--------------+
| Las Vegas  |         3873 |
| Montréal   |         1757 |
| Gilbert    |         1549 |
| Las Vegas  |         1410 |
| Las Vegas  |         1389 |
| Las Vegas  |         1252 |
| Las Vegas  |         1116 |
| Las Vegas  |         1084 |
| Las Vegas  |          961 |
| Gilbert    |          902 |
| Las Vegas  |          864 |
| Scottsdale |          823 |
| Las Vegas  |          821 |
| Las Vegas  |          786 |
| Henderson  |          785 |
| Toronto    |          778 |
| Las Vegas  |          768 |
| Las Vegas  |          758 |
| Scottsdale |          726 |
| Cleveland  |          723 |
| Las Vegas  |          720 |
| Charlotte  |          715 |
| Phoenix    |          711 |
| Las Vegas  |          706 |
| Phoenix    |          700 |
+------------+--------------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

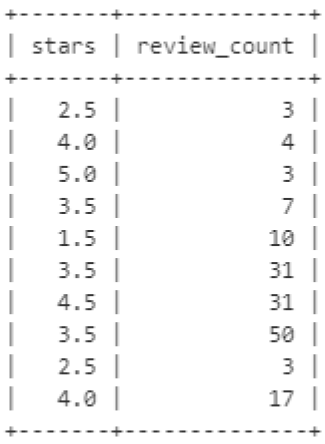
6. Find the distribution of star ratings to the business in the following cities:

i. Avon

SQL code used to arrive at answer:

```
SELECT stars
,review_count
FROM business
WHERE city = 'Avon';
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):



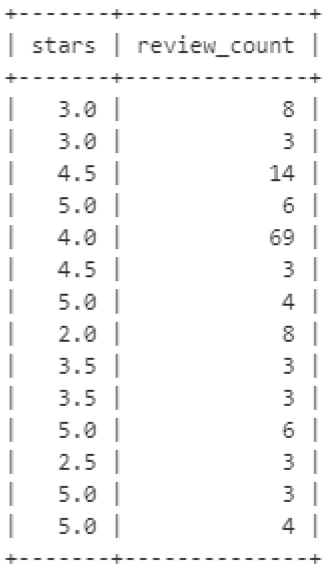
```
+-------+--------------+
| stars | review_count |
+-------+--------------+
|   2.5 |            3 |
|   4.0 |            4 |
|   5.0 |            3 |
|   3.5 |            7 |
|   1.5 |           10 |
|   3.5 |           31 |
|   4.5 |           31 |
|   3.5 |           50 |
|   2.5 |            3 |
|   4.0 |           17 |
+-------+--------------+
```

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars
,review_count
FROM business
WHERE city = 'Beachwood';
```

Copy and Paste the Resulting Table Below (2 columns â€‟ star rating and count):



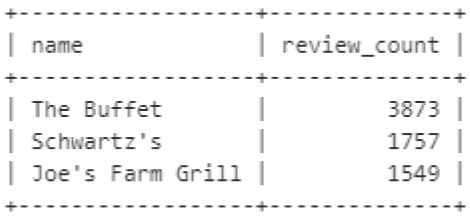
```
+-------+--------------+
| stars | review_count |
+-------+--------------+
|   3.0 |            8 |
|   3.0 |            3 |
|   4.5 |           14 |
|   5.0 |            6 |
|   4.0 |           69 |
|   4.5 |            3 |
|   5.0 |            4 |
|   2.0 |            8 |
|   3.5 |            3 |
|   3.5 |            3 |
|   5.0 |            6 |
|   2.5 |            3 |
|   5.0 |            3 |
|   5.0 |            4 |
+-------+--------------+
```

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

SQL code used to arrive at answer:

```
SELECT name
,review_count
FROM business
ORDER BY review_count desc
LIMIT 3;
```

Copy and Paste the Result Below:



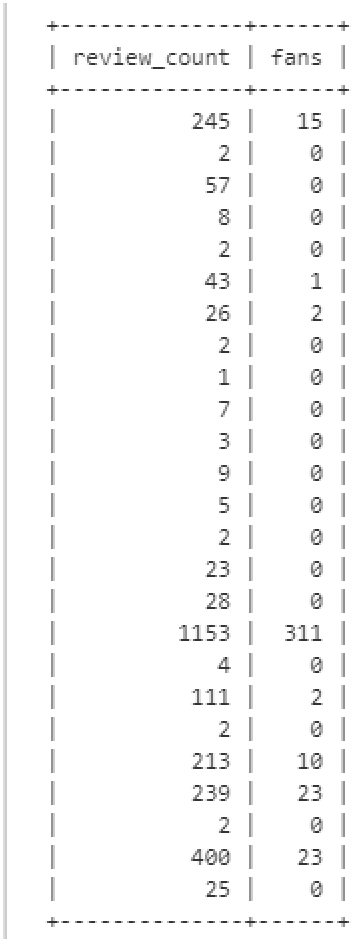
```
+------------------+--------------+
| name             | review_count |
+------------------+--------------+
| The Buffet       |         3873 |
| Schwartz's       |         1757 |
| Joe's Farm Grill |         1549 |
+------------------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

I selected columns review count and fans from table user.

```
My code: SELECT review_count
,fans
FROM user;
Result:
```



```
+--------------+------+
| review_count | fans |
+--------------+------+
|          245 |   15 |
|            2 |    0 |
|           57 |    0 |
|            8 |    0 |
|            2 |    0 |
|           43 |    1 |
|           26 |    2 |
|            2 |    0 |
|            1 |    0 |
|            7 |    0 |
|            3 |    0 |
|            9 |    0 |
|            5 |    0 |
|            2 |    0 |
|           23 |    0 |
|           28 |    0 |
|         1153 |  311 |
|            4 |    0 |
|          111 |    2 |
|            2 |    0 |
|          213 |   10 |
|          239 |   23 |
|            2 |    0 |
|          400 |   23 |
|           25 |    0 |
+--------------+------+
```

By looking at the table, we can nearly state that more no. of reviews more no. of fans.

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

Answer: “love”

SQL code used to arrive at answer:

```
For love : SELECT COUNT(text)
FROM review
WHERE text LIKE '%love%';
FOR hate : SELECT COUNT(text)
FROM review
WHERE text LIKE '%hate%';
```

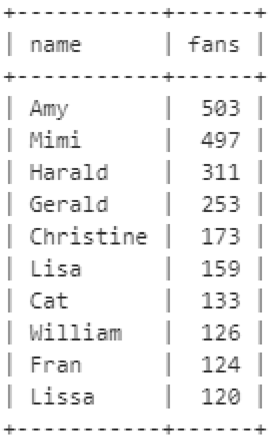
10. Find the top 10 users with the most fans:

SQL code used to arrive at answer:

```
SELECT name
,fans
FROM user
```

```
ORDER BY fans desc
LIMIT 10;
```

Copy and Paste the Result Below:

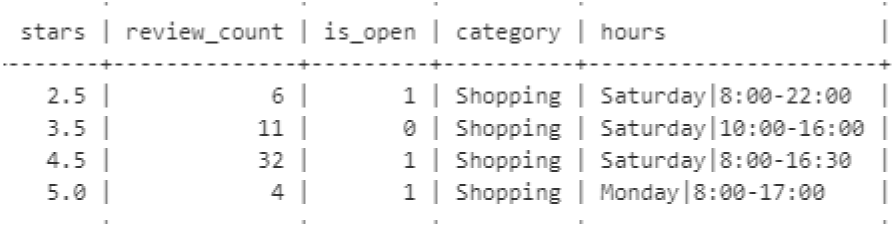


```
+-----------+------+
| name      | fans |
+-----------+------+
| Amy       |  503 |
| Mimi      |  497 |
| Harald    |  311 |
| Gerald    |  253 |
| Christine |  173 |
| Lisa      |  159 |
| Cat       |  133 |
| William   |  126 |
| Fran      |  124 |
| Lissa     |  120 |
+-----------+------+
```

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?
No, 2-3 stars and 4-5 stars tend to open up early.



```
 stars | review_count | is_open | category | hours                 |
-------+--------------+---------+----------+-----------------------+
   2.5 |            6 |       1 | Shopping | Saturday|8:00-22:00   |
   3.5 |           11 |       0 | Shopping | Saturday|10:00-16:00  |
   4.5 |           32 |       1 | Shopping | Saturday|8:00-16:30   |
   5.0 |            4 |       1 | Shopping | Monday|8:00-17:00     |
```

ii. Do the two groups you chose to analyze have a different number of reviews?
Yes, the businesses having average stars i.e. 3.5-4.5 tend to have high no. of reviews than others.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
The businesses with less star lie in the region of Tropicana.

SQL code used for analysis:

```
SELECT b.*
,c.category
,h.hours
```

```
FROM (business b INNER JOIN category c ON b.id =
c.business_id) INNER JOIN hours h ON h.business_id =
b.id
WHERE b.city = 'Las Vegas' AND c.category = 'Shopping'
GROUP BY b.stars;
```

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:Business that are closed have got only one star’4’ in review while that are open have got all the stars.

ii. Difference 2:There is only one business named “Stella's Pizza & Italian Restaurant” which is closed and has got a checkin . Rest all businesses who have got a checkin are opened.

SQL code used for analysis:

```
   i.   SELECT b.is_open
          ,c.*
           FROM business b INNER JOIN review c
          ON b.id = c.id
          WHERE is_open = 0;
   ii.  SELECT b.is_open
        ,b.name
        ,c.*
        FROM business b INNER JOIN checkin c
        ON b.id = c.business_id
        WHERE is_open = 0;
```

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:

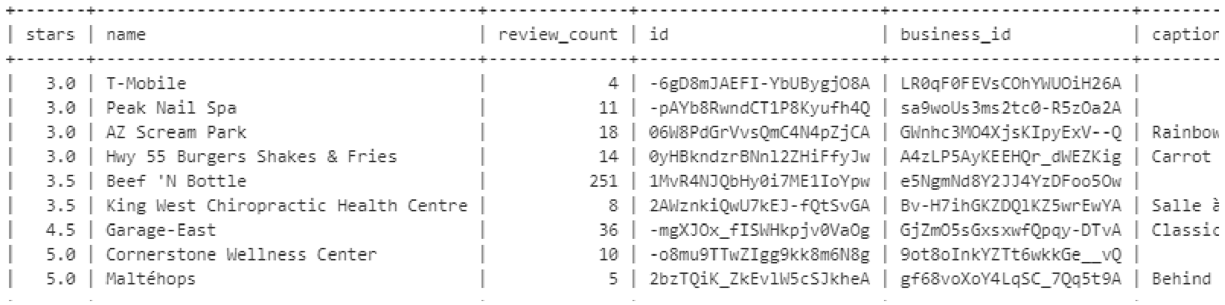
i. Indicate the type of analysis you chose to do:
predicting which businesses are likely to have a photo on the basis of no. of stars they have. Also, analyzing the relation between review count and the photo.

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

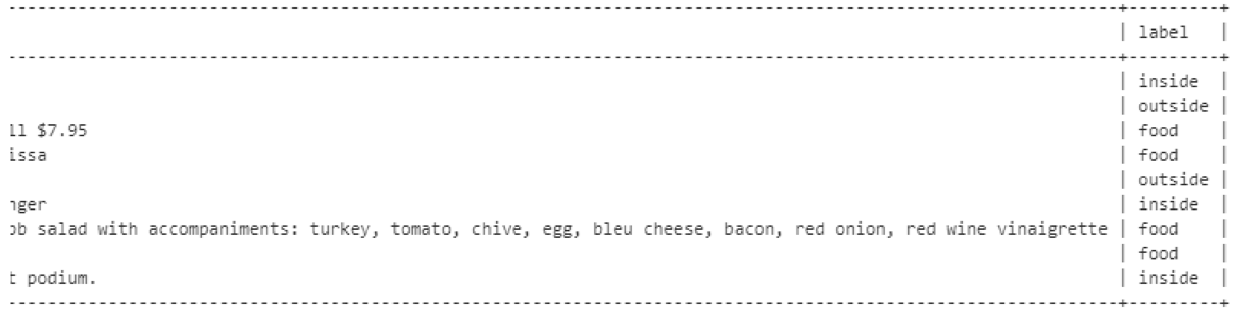
From business table , I picked out the name , stars and no. of reviews it has . Similarly from photo table I picked out id, caption and the label.
I created an inner join by the ids to investigate the relation.

iii. Output of your finished dataset:

I found out that majorly the businesses having stars more than 3.0 tend to have their photo. The label was majorly inside , outside and food . There was no direct relation between the no. of reviews and photo.



```
+-------+------------------------------------+--------------+------------------------+------------------------+--------
| stars | name                               | review_count | id                     | business_id            | captior
+-------+------------------------------------+--------------+------------------------+------------------------+--------
|   3.0 | T-Mobile                           |            4 | -6gD8mJAEFI-YbUBygjO8A | LR0qF0FEVsCOhYWUOiH26A |
|   3.0 | Peak Nail Spa                      |           11 | -pAYb8RwndCT1P8Kyufh4Q | sa9woUs3ms2tc0-R5zOa2A |
|   3.0 | AZ Scream Park                     |           18 | 06W8PdGrVvsQmC4N4pZjCA | GWnhc3MO4XjsKIpyExV--Q | Rainbow
|   3.0 | Hwy 55 Burgers Shakes & Fries      |           14 | 0yHBkndzrBNnl2ZHiFfyJw | A4zLP5AyKEEHQr_dWEZKig | Carrot
|   3.5 | Beef 'N Bottle                     |          251 | 1MvR4NJQbHy0i7ME1IoYpw | e5NgmNd8Y2JJ4YzDFoo5Ow |
|   3.5 | King West Chiropractic Health Centre |          8 | 2AWznkiQwU7kEJ-fQtSvGA | Bv-H7ihGKZDQ1KZ5wrEwYA | Salle à
|   4.5 | Garage-East                        |           36 | -mgXJOx_fISWHkpjv0VaOg | GjZmO5sGxsxwfQpqy-DTvA | Classic
|   5.0 | Cornerstone Wellness Center        |           10 | -o8mu9TTwZIgg9kk8m6N8g | 9ot8oInkYZTt6wkkGe__vQ |
|   5.0 | Maltéhops                          |            5 | 2bzTQiK_ZkEv1W5cSJkheA | gf68voXoY4LqSC_7Qq5t9A | Behind
```



```
----------------------------------------------------------------------------------------------------------+---------+
                                                                                                          | label   |
----------------------------------------------------------------------------------------------------------+---------+
                                                                                                          | inside  |
                                                                                                          | outside |
ll $7.95                                                                                                  | food    |
issa                                                                                                      | food    |
                                                                                                          | outside |
nger                                                                                                      | inside  |
ob salad with accompaniments: turkey, tomato, chive, egg, bleu cheese, bacon, red onion, red wine vinaigrette | food    |
                                                                                                          | food    |
t podium.                                                                                                 | inside  |
----------------------------------------------------------------------------------------------------------+---------+
```

iv. Provide the SQL code you used to create your final dataset:

```
SELECT b.stars

,b.name

,b.review_count

,c.*

FROM business b INNER JOIN photo c

ON b.id = c.id

ORDER BY stars asc;
```