Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset Coursera Worksheet

This is a 2-part assignment. In the first part, you are asked a series of questions that will help you profile and understand the data just like a data scientist would. For this first part of the assignment, you will be assessed both on the correctness of your findings, as well as the code you used to arrive at your answer. You will be graded on how easy your code is to read, so remember to use proper formatting and comments where necessary.

In the second part of the assignment, you are asked to come up with your own inferences and analysis of the data for a particular research question you want to answer. You will be required to prepare the dataset for the analysis you choose to do. As with the first part, you will be graded, in part, on how easy your code is to read, so use proper formatting and comments to illustrate and communicate your intent as required.

For both parts of this assignment, use this "worksheet." It provides all the questions you are being asked, and your job will be to transfer your answers and SQL coding where indicated into this worksheet so that your peers can review your work. You should be able to use any Text Editor (Windows Notepad, Apple TextEdit, Notepad ++, Sublime Text, etc.) to copy and paste your answers. If you are going to use Word or some other page layout application, just be careful to make sure your answers and code are lined appropriately.
In this case, you may want to save as a PDF to ensure your formatting remains intact for you reviewer.

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
ii. Business table = 10000
iii. Category table = 10000
iv. Checkin table = 10000
v. elite_years table = 10000
vi. friend table = 10000
vii. hours table = 10000
viii. photo table = 10000
ix. review table = 10000
x. tip table = 10000
xi. user table = 10000

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 = 10000(id)
ii. Hours = 1562(business_id)
iii. Category = 2643(business_id)
iv. Attribute = 1115(business_id)
v. Review = 10000(id)/8090(business_id)/9581(user_id)
vi. Checkin = 493(business_id)
vii. Photo = 10000(id)/6493(business_id)
viii. Tip = 537(user_id)/3979(business_id)

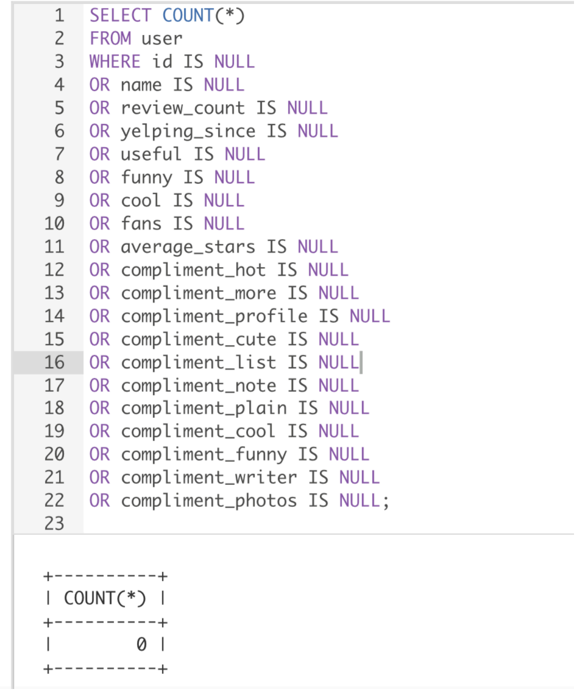
ix. User = 10000(id)
x. Friend = 11(user_id)
xi. Elite_years = 2780(user_id)

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 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

ii. Table: Business, Column: Stars

min:1   max:5   avg:3.6549

iii. Table: Tip, Column: Likes

min:0   max:2   avg:0.0144

iv. Table: Checkin, Column: Count

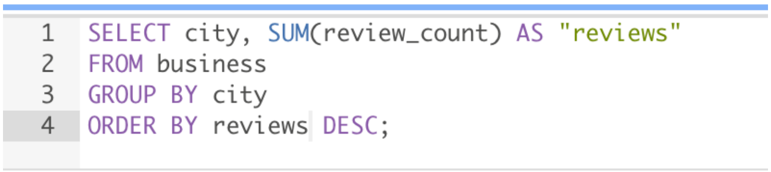
min:1 max:53 avg:1.9414

v. Table: User, Column: Review_count

min:0 max:2000 avg:24.2995

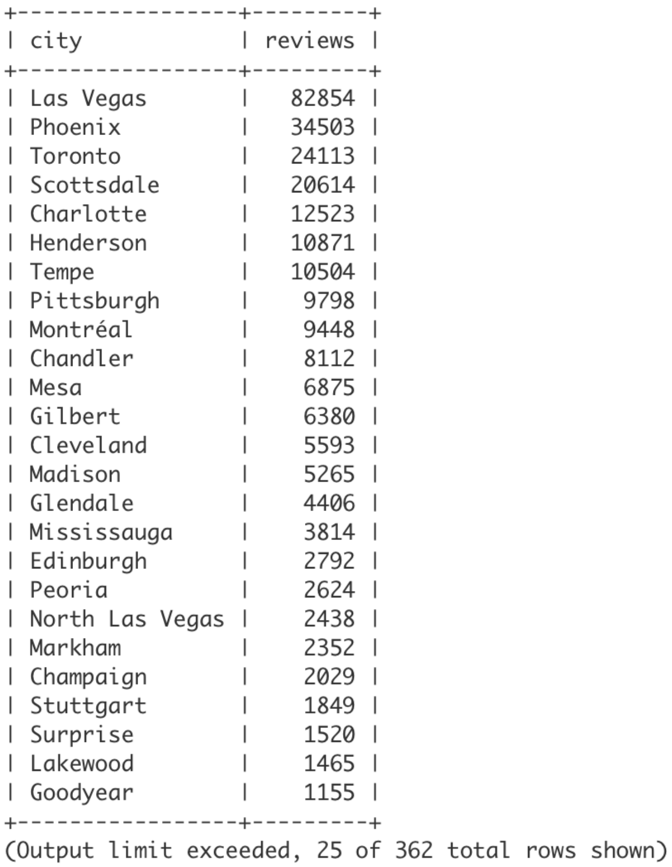
5. List the cities with the most reviews in descending order:

SQL code used to arrive at answer:



```
SELECT city, SUM(review_count) AS "reviews"
FROM business
GROUP BY city
ORDER BY reviews DESC;
```

Copy and Paste the Result Below:

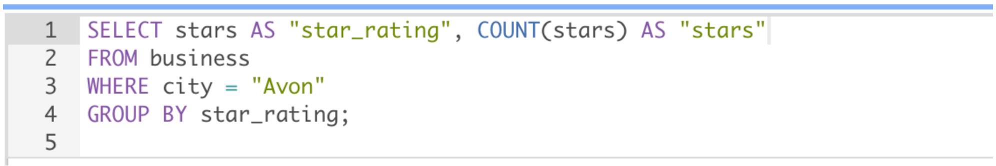


```
+-----------------+---------+
| city            | reviews |
+-----------------+---------+
| Las Vegas       |   82854 |
| Phoenix         |   34503 |
| Toronto         |   24113 |
| Scottsdale      |   20614 |
| Charlotte       |   12523 |
| Henderson       |   10871 |
| Tempe           |   10504 |
| 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:

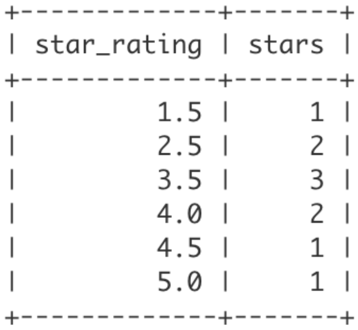
i. Avon

SQL code used to arrive at answer:



```sql
SELECT stars AS "star_rating", COUNT(stars) AS "stars"
FROM business
WHERE city = "Avon"
GROUP BY star_rating;
```

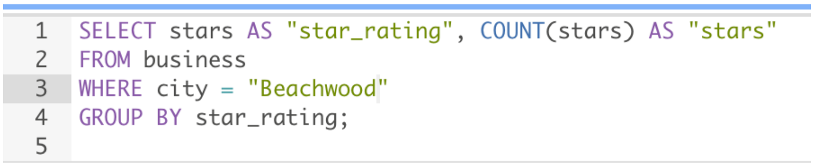
Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):



```
+-------------+-------+
| star_rating | stars |
+-------------+-------+
|         1.5 |     1 |
|         2.5 |     2 |
|         3.5 |     3 |
|         4.0 |     2 |
|         4.5 |     1 |
|         5.0 |     1 |
+-------------+-------+
```

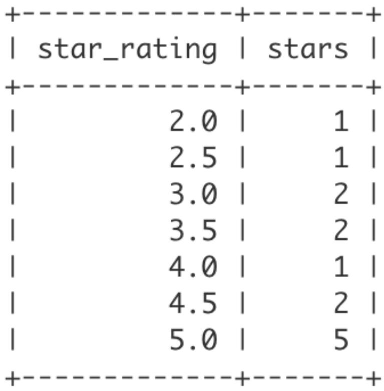
ii. Beachwood

SQL code used to arrive at answer:



```sql
SELECT stars AS "star_rating", COUNT(stars) AS "stars"
FROM business
WHERE city = "Beachwood"
GROUP BY star_rating;
```

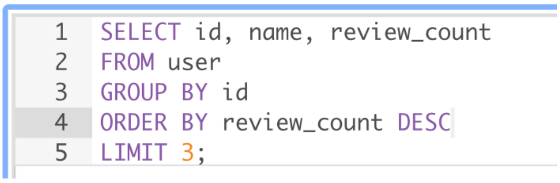
Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):



```
+-------------+-------+
| star_rating | stars |
+-------------+-------+
|         2.0 |     1 |
|         2.5 |     1 |
|         3.0 |     2 |
|         3.5 |     2 |
|         4.0 |     1 |
|         4.5 |     2 |
|         5.0 |     5 |
+-------------+-------+
```

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

SQL code used to arrive at answer:



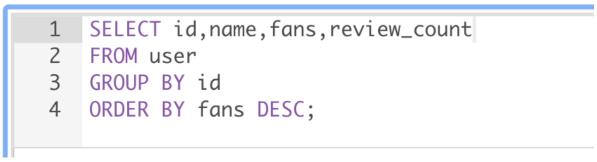
```
SELECT id, name, review_count
FROM user
GROUP BY id
ORDER BY review_count DESC
LIMIT 3;
```

Copy and Paste the Result Below:

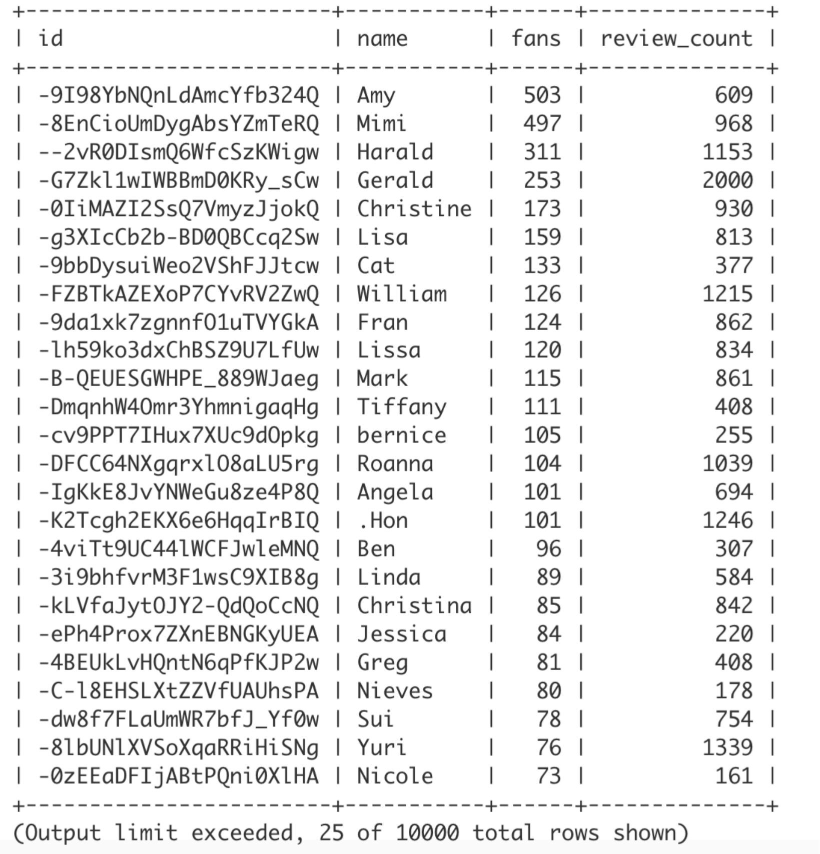
```
+------------------------+---------+--------------+
| id                     | name    | review_count |
+------------------------+---------+--------------+
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald  |         2000 |
| -3s52C4zL_DHRK0ULG6qtg | Sara    |         1629 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri    |         1339 |
+------------------------+---------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:



```
SELECT id,name,fans,review_count
FROM user
GROUP BY id
ORDER BY fans DESC;
```

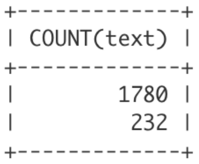


```
+------------------------+-----------+------+--------------+
| id                     | name      | fans | review_count |
+------------------------+-----------+------+--------------+
| -9I98YbNQnLdAmcYfb324Q | Amy       |  503 |          609 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi      |  497 |          968 |
| --2vR0DIsmQ6WfcSzKWigw | Harald    |  311 |         1153 |
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald    |  253 |         2000 |
| -0IiMAZI2SsQ7VmyzJjokQ | Christine |  173 |          930 |
| -g3XIcCb2b-BD0QBCcq2Sw | Lisa      |  159 |          813 |
| -9bbDysuiWeo2VShFJJtcw | Cat       |  133 |          377 |
| -FZBTkAZEXoP7CYvRV2ZwQ | William   |  126 |         1215 |
| -9da1xk7zgnnf01uTVYGkA | Fran      |  124 |          862 |
| -lh59ko3dxChBSZ9U7LfUw | Lissa     |  120 |          834 |
| -B-QEUESGWHPE_889WJaeg | Mark      |  115 |          861 |
| -DmqnhW40mr3YhmnigaqHg | Tiffany   |  111 |          408 |
| -cv9PPT7IHux7XUc9dOpkg | bernice   |  105 |          255 |
| -DFCC64NXgqrxl08aLU5rg | Roanna    |  104 |         1039 |
| -IgKkE8JvYNWeGu8ze4P8Q | Angela    |  101 |          694 |
| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon      |  101 |         1246 |
| -4viTt9UC44lWCFJwleMNQ | Ben       |   96 |          307 |
| -3i9bhfvrM3F1wsC9XIB8g | Linda     |   89 |          584 |
| -kLVfaJyt0JY2-QdQoCcNQ | Christina |   85 |          842 |
| -ePh4Prox7ZXnEBNGKyUEA | Jessica   |   84 |          220 |
| -4BEUkLvHQntN6qPfKJP2w | Greg      |   81 |          408 |
| -C-l8EHSLXtZZVfUAUhsPA | Nieves    |   80 |          178 |
| -dw8f7FLaUmWR7bfJ_Yf0w | Sui       |   78 |          754 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri      |   76 |         1339 |
| -0zEEaDFIjABtPQni0XlHA | Nicole    |   73 |          161 |
+------------------------+-----------+------+--------------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

From the query result, it shows that there is no positive relation between reviews and fans. Although the user Amy has the most number of fans, she doesn’t has the most number of reviews. Also other users like Christina posts quite a lot of reviews, but she has only 85 fans.

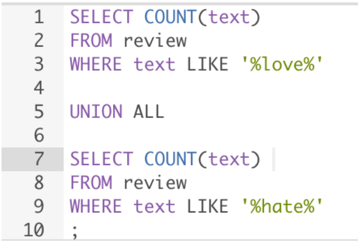
9. Are there more reviews with the word "love" or with the word "hate" in them?

Answer: Yes. There are 1780 “love” and 232 “hate”.



```
+-------------+
| COUNT(text) |
+-------------+
|        1780 |
|         232 |
+-------------+
```

SQL code used to arrive at answer:



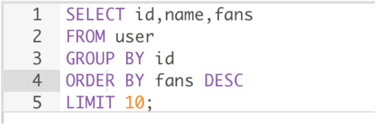
```
SELECT COUNT(text)
FROM review
WHERE text LIKE '%love%'

UNION ALL

SELECT COUNT(text)
FROM review
WHERE text LIKE '%hate%'
;
```

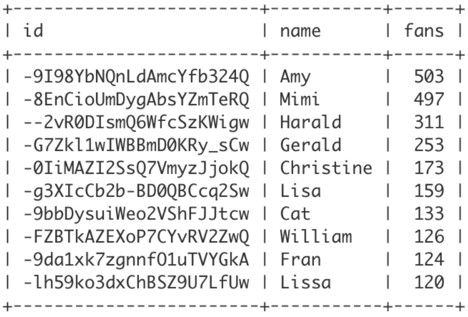
10. Find the top 10 users with the most fans:

SQL code used to arrive at answer:



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

Copy and Paste the Result Below:



```
+------------------------+-----------+------+
| id                     | name      | fans |
+------------------------+-----------+------+
| -9I98YbNQnLdAmcYfb324Q | Amy       |  503 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi      |  497 |
| --2vR0DIsmQ6WfcSzKWigw | Harald    |  311 |
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald    |  253 |
| -0IiMAZI2SsQ7VmyzJjokQ | Christine |  173 |
| -g3XIcCb2b-BD0QBCcq2Sw | Lisa      |  159 |
| -9bbDysuiWeo2VShFJJtcw | Cat       |  133 |
| -FZBTkAZEXoP7CYvRV2ZwQ | William   |  126 |
| -9da1xk7zgnnf01uTVYGkA | Fran      |  124 |
| -lh59ko3dxChBSZ9U7LfUw | 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?

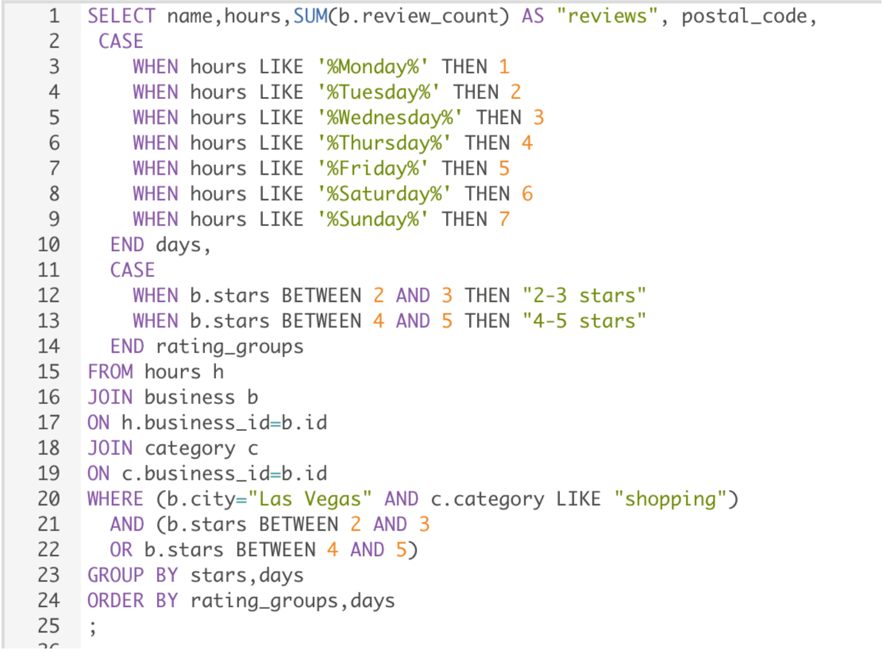
Yes, the 2-3 stars group open longer than the 4-5 stars group.

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

Yes/No. One of the 4-5 stars group has much more reviews, but another 4-5 stars group has similar number of reviews as that of the 2-3 stars group.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
No. All of the groups are in different locations.

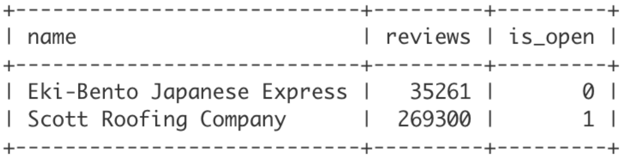
SQL code used for analysis:



```
SELECT name,hours,SUM(b.review_count) AS "reviews", postal_code,
 CASE
    WHEN hours LIKE '%Monday%' THEN 1
    WHEN hours LIKE '%Tuesday%' THEN 2
    WHEN hours LIKE '%Wednesday%' THEN 3
    WHEN hours LIKE '%Thursday%' THEN 4
    WHEN hours LIKE '%Friday%' THEN 5
    WHEN hours LIKE '%Saturday%' THEN 6
    WHEN hours LIKE '%Sunday%' THEN 7
  END days,
  CASE
    WHEN b.stars BETWEEN 2 AND 3 THEN "2-3 stars"
    WHEN b.stars BETWEEN 4 AND 5 THEN "4-5 stars"
  END rating_groups
FROM hours h
JOIN business b
ON h.business_id=b.id
JOIN category c
ON c.business_id=b.id
WHERE (b.city="Las Vegas" AND c.category LIKE "shopping")
  AND (b.stars BETWEEN 2 AND 3
  OR b.stars BETWEEN 4 AND 5)
GROUP BY stars,days
ORDER BY rating_groups,days
;
```

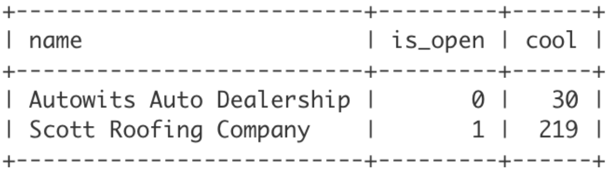
2. Group business based on the ones that are open and the ones that are closed. What differences can you find between the ones that are still open and the ones that are closed? List at least two differences and the SQL code you used to arrive at your answer.

i. Difference 1:
Business that are open have more reviews than that are closed.



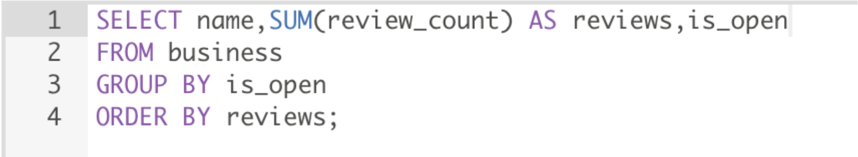
```
+---------------------------+---------+---------+
| name                      | reviews | is_open |
+---------------------------+---------+---------+
| Eki-Bento Japanese Express |   35261 |       0 |
| Scott Roofing Company     |  269300 |       1 |
+---------------------------+---------+---------+
```

ii. Difference 2:
Business that are open have many more “cool” in reviews than that are closed.



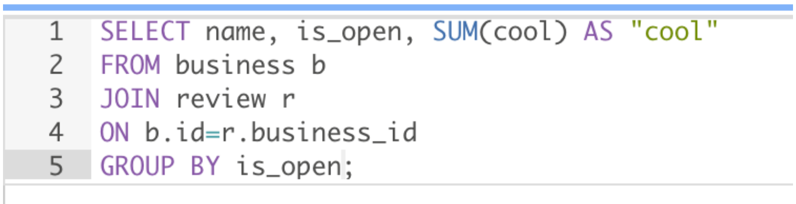
```
+--------------------------+---------+------+
| name                     | is_open | cool |
+--------------------------+---------+------+
| Autowits Auto Dealership |       0 |   30 |
| Scott Roofing Company    |       1 |  219 |
+--------------------------+---------+------+
```

SQL code used for analysis:
Diff 1:



```
SELECT name,SUM(review_count) AS reviews,is_open
FROM business
GROUP BY is_open
ORDER BY reviews;
```

Diff 2:



```
SELECT name, is_open, SUM(cool) AS "cool"
FROM business b
JOIN review r
ON b.id=r.business_id
GROUP BY is_open;
```

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:

I want to predict the real rating of business by sentiment analysis of the review text. Since customers normally are just being polite, the rating given by stars may not show the truth of the business.

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

To figure out the real comment of business for both owners and customers, some data may be important; number of reviews for each business, review text; category will be used to better distinguish between different types of business.

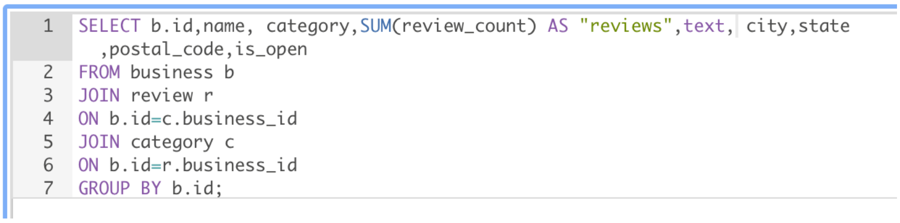
Once I extract all the reviews for classification and prediction, I will generate more specific information for each business for visualization purpose, so that business owners can learn more easily regarding “who is doing better than me?” ,and then take actions based on the comparison result. Such data includes city, state, poste code, is_open.

iii. Output of your finished dataset:

```
     | id                    | name                                    | category
| reviews | text
| city       | state | postal_code | is_open |
+------------------------+--------------------------------------+--------------+-
--------+-----------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------+------------+-------+-------------+------
---+
| -9lOQ0Lfm8wiu8eSdUXS8A | Moondogs Pub                           | Nightlife    |
21 | Mr. Esser...Banning Buddy Guy from future Blues events in the city of
Pittsburgh is the most assinine thing i've ever heard anyone do. It's 2015. It's a
Blues Fest in Pittsburgh, not Dade Cunty Florida. Have you ever been to a Steeler
game? You cheat all Buddy fans and Blues fans alike. Screw ever coming into your
understaffed club again. Im sure Buddy can work around you and entertain Blues fans
before this Legend is gone for good. You've lost the plot for sure.
| Pittsburgh | PA    | 15238       |       1 |
| -Eu04UHRqmGGyvYRDY8-tg | West Side Market                       | Meat Shops   |
14460 | I love this market, crowded, fresh and cheap veggies! Nice collection of
bread, cheese and butter.
| Cleveland  | OH    | 44113       |       1 |
|                        |                                         |              |
| The only thing about it is the close early!
|            |       |             |         |
| -Za5mjo-CYYUMsd1r8GC7Q | Ashbridges Bay Park                    | Parks        |
87 | As the other reviewers have said, this is as close as you can get to a real
beach in Toronto. It has great places for volleyball, a nice boardwalk, and-- this
is key-- it's great for people watching. Hello, extremely large man! I am sorry for
staring, but your belly just jiggles so much!
| Toronto    | ON    | M4L 3W6     |       1 |
|                        |                                         |              |
|
|            |       |             |         |
|                        |                                         |              |
| However, I do have to say that, in Vancouver, water is *blue*. Not green. Your
algae level was a bit excessive-- I didn't feel like going into the water. And
admittedly, it is a bit crowded, and the washrooms are nasty, but that's true at
most beaches I've visited in my life.
|            |       |             |         |
|                        |                                         |              |
|
|            |       |             |         |
|                        |                                         |              |
| But at most beaches, I don't have to wonder whether the lake is cleaner than the
bathroom. I ask for too much out of Lake Ontario. Sorry, nature.
|            |       |             |         |
| 0NDbUCHi9YsRwgG3iZO8Kg | Cafe Tandoor                           | Restaurants  |
96 | The restaurant has a nice atmosphere with great food, and good service.
| Aurora     | OH    | 44202       |       1 |
|                        |                                         |              |
|
|            |       |             |         |
|                        |                                         |              |
| I ordered a lamb biryani. There was a mistake in my order, so I told the manager
about it. He apologized for the mistake and offered me another order for free as a
compensation, which I gladly accepted.
|            |       |             |         |
|                        |                                         |              |
| The manager handle the situation professionally, which I appreciate.
|            |       |             |         |
| 1ZnVfS-qP19upP_fwOhZsA | Big Wong Restaurant                    | Asian Fusion |
10752 | This restaurant has great service. Their foods are very delicious and not
expensive. Thank you!!
| Las Vegas  | NV    | 89146       |       1 |
| 1veVZUawy7IhIc5oDpRRQA | Slyman's Restaurant                    | Restaurants  |
1083 | One of the best places I've ever gone for breakfast. Unassuming and
positively delightful. Old school breakfast of two eggs over easy with hash browns.
| Cleveland  | OH    | 44114       |       1 |
|                        |                                        |              |
| Couldn't have been better. I would never go anywhere else in Cleveland.
|            |       |             |         |
| 20ib4z2Yo2wlfARFMcFwSQ | Vanilla Pastry Studio                  | Food         |
144 | On a recent visit to Pgh I was happily surprised to find that numerous
cupcake shops have sprung up since my last visit. Being the cupcake fanatic that I
am, I tried Dozen, Coco's and Vanilla Pastry Studio. Although I think all three
cupcake shops could have come up with better names, the VPS had the best products
hands down. Over my 2 month visit, I tried their vanilla/vanilla, vanilla caramel,
carrot, mocha and choc peanut butter. They were all incredible with the vanilla
caramel being my favorite. And the atmosphere inside the shop is adorable, it made
me want to skip around like I won the Willy Wonka golden ticket, and after tasting
one of their cupcakes, into another world I went. | Pittsburgh | PA    | 15206
|       0 |
| 24Td_CQH1bonWKff1rt2vg | Matt's Big Breakfast                   | Restaurants  |
752 | I thought the bacon couldn't be topped until I had the jelly that came with
my toast. This place is the sh*t. Need to go back when I am hungover to get the
best experience. My new favorite breakfast spot.
| Phoenix    | AZ    | 85016       |       1 |
| 2skQeu3C36VCiB653MIfrw | Bootleggers Modern American Smokehouse | Barbeque     |
9051 | This restaurant has the absolute best atmosphere. It starts from the moment
you get out of your car and smell the aroma from the smokehouse. From there, it is
dark inside which creates a calm homey feeling. The decor is perfect. The food is
out of this world! Apple pie moonshine? Just say yes!
| Phoenix    | AZ    | 85028       |       1 |
+------------------------+----------------------------------------+--------------+-
--------+-----------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-----------------------------------------+------------+-------+-------------+------
---+
```

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



```
SELECT b.id,name, category,SUM(review_count) AS "reviews",text, city,state
  ,postal_code,is_open
FROM business b
JOIN review r
ON b.id=c.business_id
JOIN category c
ON b.id=r.business_id
GROUP BY b.id;
```