Gabriel Gonzales
SQL for Data Science
Peer-Graded Assignment
7/4/2021

Part 1: Yelp Dataset Profiling and Understanding

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

Code Used:

```
SELECT COUNT (*)
FROM Insert table name ;
```

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. 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.

Code Used:

```
 SELECT COUNT (DISTINCT(insert primary key id))
FROM Insert table name;
```

i. Business = Business = id: 10000
ii. Hours = Hours = business_id: 1562
iii. Category = Category = business_id: 2643
iv. Attribute = business_id: 1115
v. Review = id: 10000, business_id: 8090, user_id: 9581
vi. Checkin = business_id: 493
vii. Photo = id: 1000, business_id: 10000
viii. Tip = business_id: 537, user_id: 537
ix. User = id: 10000
x. Friend = user_id: 11
xi. Elite_years = user_id: 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 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;
```

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

Code used:

```
SELECT (MIN, MAX, AVG)(Insert column name)
FROM Insert table name ;
```

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: 24.995

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:

```
SELECT stars,
SUM(review_count) AS '#_of_reviews'
FROM business
Where City = 'Avon'
GROUP BY stars;
```

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

```
+--------+--------------+
| stars  | #_of_reviews |
+--------+--------------+
|   1.5  |           10 |
|   2.5  |            6 |
|   3.5  |           88 |
|   4.0  |           21 |
|   4.5  |           31 |
|   5.0  |            3 |
+--------+--------------+
```

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars,
SUM(review_count) AS '#_of_reviews'
FROM business
Where City = 'Beachwood'
GROUP BY stars
;
```

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

```
+--------+--------------+
| stars  | #_of_reviews |
+--------+--------------+
|   2.0  |            8 |
|   2.5  |            3 |
|   3.0  |           11 |
|   3.5  |            6 |
|   4.0  |           69 |
|   4.5  |           17 |
|   5.0  |           23 |
+--------+--------------+
```

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
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:

NO, there is no distinct correlation between review_count and fans showing that an increase in reviews results in more fans.

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

```
+------------------------+-----------+--------------+------+
| id                     | name      | review_count | fans |
+------------------------+-----------+--------------+------+
| -9I98YbNQnLdAmcYfb324Q | Amy       |          609 |  503 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi      |          968 |  497 |
| --2vR0DIsmQ6WfcSzKWigw | Harald    |         1153 |  311 |
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald    |         2000 |  253 |
| -0IiMAZI2SsQ7VmyzJjokQ | Christine |          930 |  173 |
| -g3XIcCb2b-BD0QBCcq2Sw | Lisa      |          813 |  159 |
| -9bbDysuiWeo2VShFJJtcw | Cat       |          377 |  133 |
| -FZBTkAZEXoP7CYvRV2ZwQ | William   |         1215 |  126 |
| -9da1xk7zgnnfO1uTVYGkA | Fran      |          862 |  124 |
| -lh59ko3dxChBSZ9U7LfUw | Lissa     |          834 |  120 |
| -B-QEUESGWHPE_889WJaeg | Mark      |          861 |  115 |
| -DmqnhW4Omr3YhmnigaqHg | Tiffany   |          408 |  111 |
| -cv9PPT7IHux7XUc9dOpkg | bernice   |          255 |  105 |
| -DFCC64NXgqrxlO8aLU5rg | Roanna    |         1039 |  104 |
| -IgKkE8JvYNWeGu8ze4P8Q | Angela    |          694 |  101 |
| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon      |         1246 |  101 |
| -4viTt9UC44lWCFJwleMNQ | Ben       |          307 |   96 |
| -3i9bhfvrM3F1wsC9XIB8g | Linda     |          584 |   89 |
| -kLVfaJytOJY2-QdQoCcNQ | Christina |          842 |   85 |
| -ePh4Prox7ZXnEBNGKyUEA | Jessica   |          220 |   84 |
| -4BEUkLvHQntN6qPfKJP2w | Greg      |          408 |   81 |
| -C-l8EHSLXtZZVfUAUhsPA | Nieves    |          178 |   80 |
| -dw8f7FLaUmWR7bfJ_Yf0w | Sui       |          754 |   78 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri      |         1339 |   76 |
| -0zEEaDFIjABtPQni0XlHA | 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: There are more reviews with the word “l`ove” than there are reviews with the word “hate”.

SQL code used to arrive at answer:

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

```
+-----------+
| COUNT(*)  |
+-----------+
|      1780 |
+-----------+
```

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

```
+-----------+
| COUNT(*)  |
+-----------+
|       232 |
+-----------+
```

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

SQL code used to arrive at answer:

```
SELECT id, name, fans
FROM user
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 |
| -9da1xk7zgnnfO1uTVYGkA | 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. Although most are open on the same day, some are open early in the morning, some open midday, while others close in the afternoon and some stay open til midnight.

ii. Do the two groups you chose to analyze have a different number of reviews? Yes. One business has a 5-star rating with only 6 reviews, while another with only 6 reviews has a 2.5 star review. Others have a range from 32 – 168 reviews

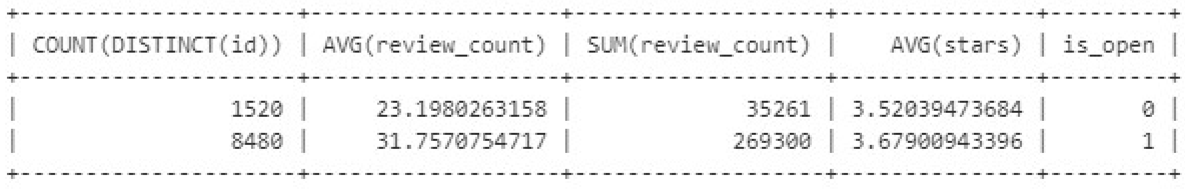
iii. Are you able to infer anything from the location data provided between these two groups? Explain. Unfortunately, no. Several of the values for the "neighborhood" attribute were NULL, so no clear correlation could be established.

SQL code used for analysis:

```
SELECT
b.name,
b.city,
b.neighborhood,
b.postal_code,
b.stars,
b.review_count,
h.hours,
c.category,
CASE
  WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'
  WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'
 END AS star_rating

FROM business AS b
INNER JOIN category AS c ON b.id = c.business_id
INNER JOIN hours AS h ON h.business_id = c.business_id
WHERE (b.city == 'Las Vegas')
--AND c.category LIKE 'Bars')
AND (b.stars BETWEEN 2 AND 3
OR
b.stars BETWEEN 4 AND 5)
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.



```
+---------------------+------------------+------------------+---------------+---------+
| COUNT(DISTINCT(id)) | AVG(review_count) | SUM(review_count) |     AVG(stars) | is_open |
+---------------------+------------------+------------------+---------------+---------+
|                1520 |    23.1980263158 |            35261 | 3.52039473684 |       0 |
|                8480 |    31.7570754717 |           269300 | 3.67900943396 |       1 |
+---------------------+------------------+------------------+---------------+---------+
```

i. Difference 1: Businesses that are open have a higher review count(31.75) on average than those that are closed (23.19)

ii. Difference 2: Business that are open have a significantly hire number of total reviews overall(269,300) than those that are closed (35,261).

SQL code used for analysis:

```
SELECT COUNT(DISTINCT(id)),
AVG(review_count),
SUM(review_count),
AVG(stars),
is_open
FROM business
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:

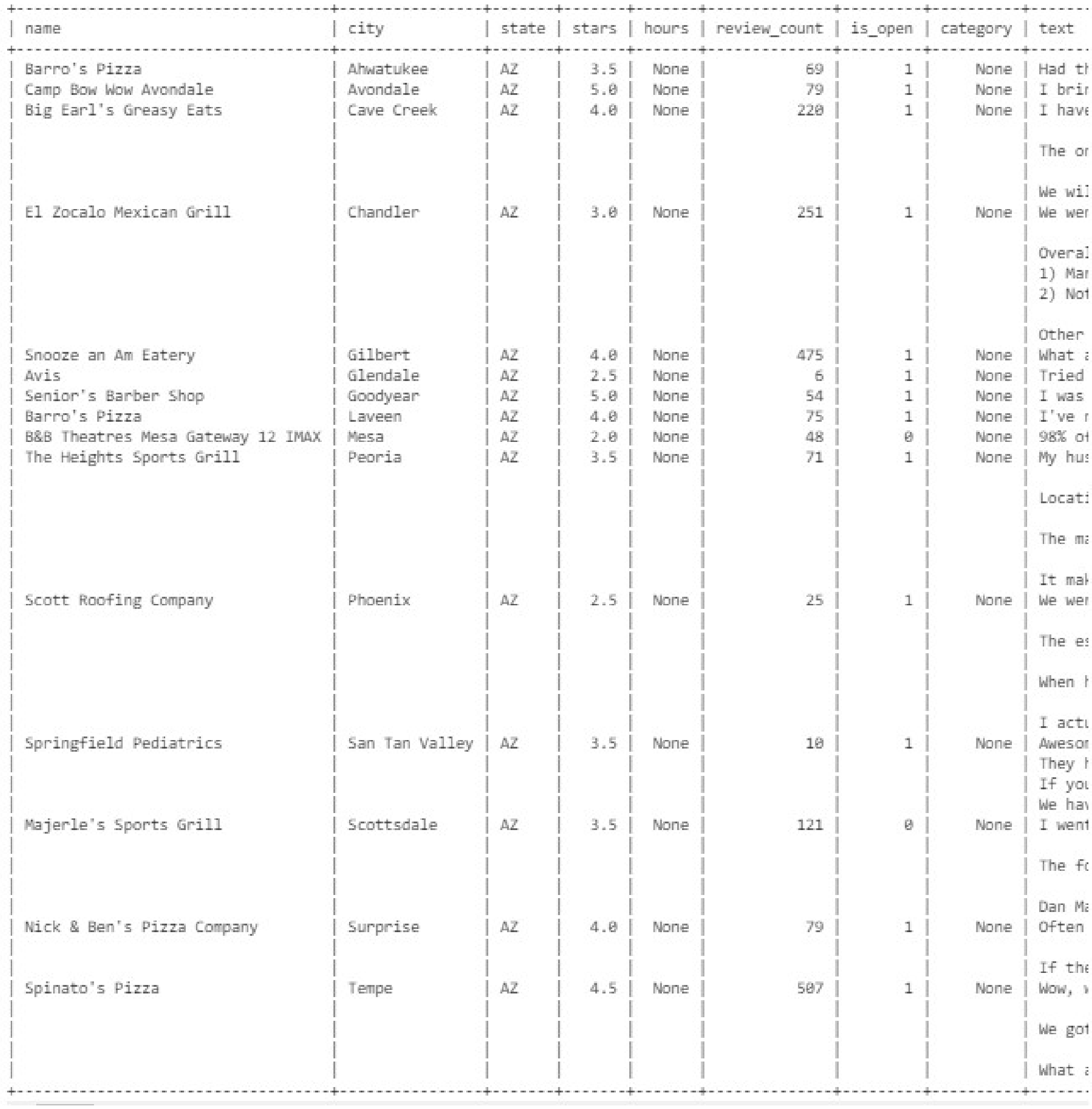
i. Indicate the type of analysis you chose to do:

I would like to prep the data for NLP analysis to determine the most common used words in reviews for restaurants that are open and for those that are closed. We could also look at which words are most common in each star category. This would provide insight to current and future restaurant owners on what and what not to focus on in their business.

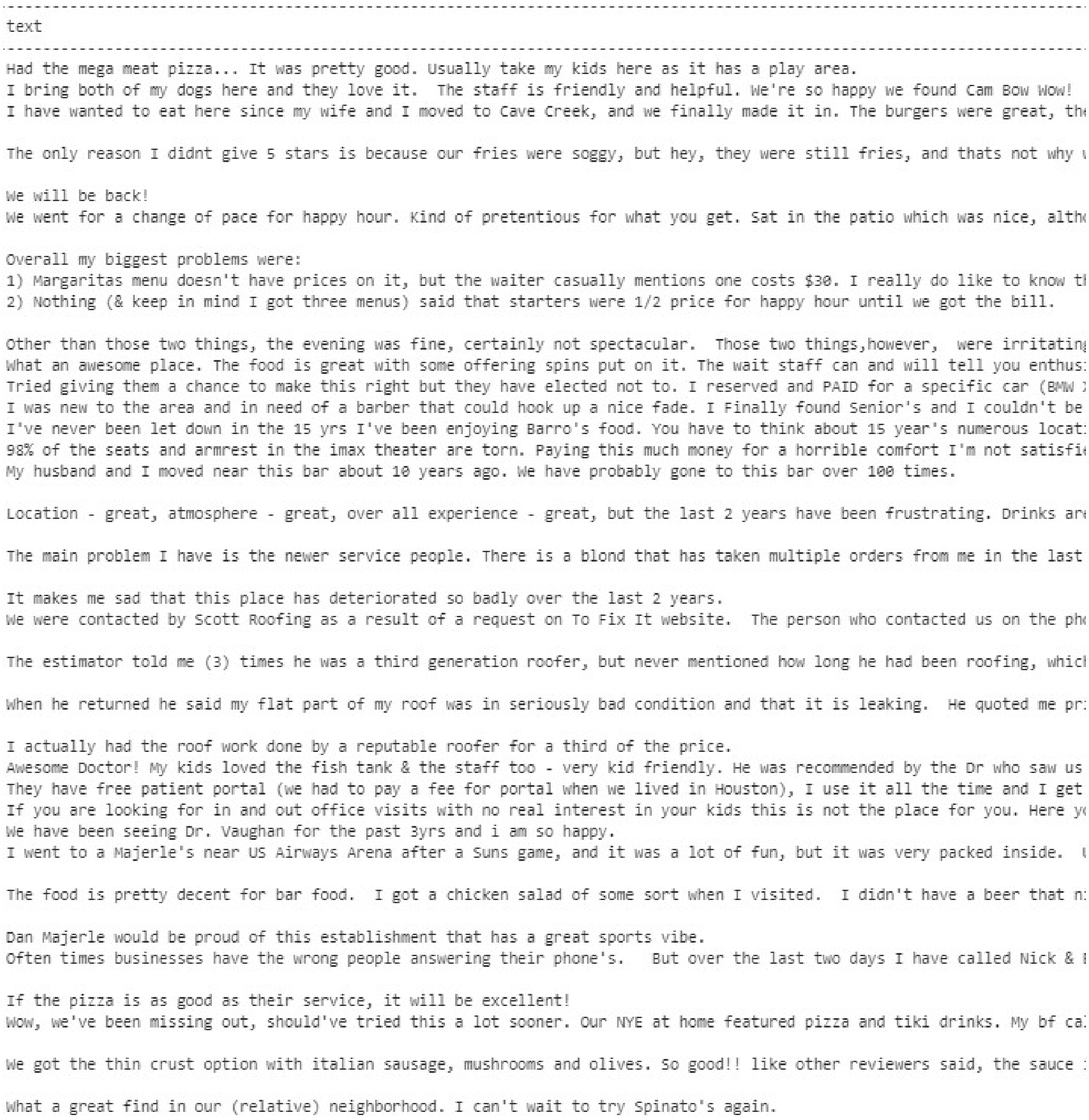
ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

The type of data we will need for this analysis are yelp reviews for all types of food categories, their location, star ratings and open status. From here we can separate the food categories or genre and look at different food types. I chose this data due to the importance of customer reviews and sentiment of those reviews about a restaurant. Yelp reviews are used all over the world and this information can help a business succeed. To start small, the query was just on businesses in the state of Arizona.

iii. Output of your finished dataset:



```
+-----------------------------------+----------------+-------+-------+-------+--------------+---------+----------+------
| name                              | city           | state | stars | hours | review_count | is_open | category | text
+-----------------------------------+----------------+-------+-------+-------+--------------+---------+----------+------
| Barro's Pizza                     | Ahwatukee      | AZ    |   3.5 |  None |           69 |       1 |     None | Had th
| Camp Bow Wow Avondale             | Avondale       | AZ    |   5.0 |  None |           79 |       1 |     None | I brin
| Big Earl's Greasy Eats            | Cave Creek     | AZ    |   4.0 |  None |          220 |       1 |     None | I have
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | The or
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | We wil
| El Zocalo Mexican Grill           | Chandler       | AZ    |   3.0 |  None |          251 |       1 |     None | We wer
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | Overal
|                                   |                |       |       |       |              |         |          | 1) Mar
|                                   |                |       |       |       |              |         |          | 2) Not
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | Other
| Snooze an Am Eatery               | Gilbert        | AZ    |   4.0 |  None |          475 |       1 |     None | What a
| Avis                              | Glendale       | AZ    |   2.5 |  None |            6 |       1 |     None | Tried
| Senior's Barber Shop              | Goodyear       | AZ    |   5.0 |  None |           54 |       1 |     None | I was
| Barro's Pizza                     | Laveen         | AZ    |   4.0 |  None |           75 |       1 |     None | I've r
| B&B Theatres Mesa Gateway 12 IMAX | Mesa           | AZ    |   2.0 |  None |           48 |       0 |     None | 98% of
| The Heights Sports Grill          | Peoria         | AZ    |   3.5 |  None |           71 |       1 |     None | My hus
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | Locati
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | The ma
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | It mak
| Scott Roofing Company             | Phoenix        | AZ    |   2.5 |  None |           25 |       1 |     None | We wer
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | The es
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | When h
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | I actu
| Springfield Pediatrics            | San Tan Valley | AZ    |   3.5 |  None |           10 |       1 |     None | Awesom
|                                   |                |       |       |       |              |         |          | They h
|                                   |                |       |       |       |              |         |          | If you
|                                   |                |       |       |       |              |         |          | We hav
| Majerle's Sports Grill            | Scottsdale     | AZ    |   3.5 |  None |          121 |       0 |     None | I went
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | The fo
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | Dan Ma
| Nick & Ben's Pizza Company        | Surprise       | AZ    |   4.0 |  None |           79 |       1 |     None | Often
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | If the
| Spinato's Pizza                   | Tempe          | AZ    |   4.5 |  None |          507 |       1 |     None | Wow, w
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | We got
|                                   |                |       |       |       |              |         |          |
|                                   |                |       |       |       |              |         |          | What a
+-----------------------------------+----------------+-------+-------+-------+--------------+---------+----------+------
```



```
---------------------------------------------------------------------------------------------------------------------
text
---------------------------------------------------------------------------------------------------------------------
Had the mega meat pizza... It was pretty good. Usually take my kids here as it has a play area.
I bring both of my dogs here and they love it.  The staff is friendly and helpful. We're so happy we found Cam Bow Wow!
I have wanted to eat here since my wife and I moved to Cave Creek, and we finally made it in. The burgers were great, the

The only reason I didnt give 5 stars is because our fries were soggy, but hey, they were still fries, and thats not why w

We will be back!
We went for a change of pace for happy hour. Kind of pretentious for what you get. Sat in the patio which was nice, altho

Overall my biggest problems were:
1) Margaritas menu doesn't have prices on it, but the waiter casually mentions one costs $30. I really do like to know th
2) Nothing (& keep in mind I got three menus) said that starters were 1/2 price for happy hour until we got the bill.

Other than those two things, the evening was fine, certainly not spectacular.  Those two things,however,  were irritating
What an awesome place. The food is great with some offering spins put on it. The wait staff can and will tell you enthusi
Tried giving them a chance to make this right but they have elected not to. I reserved and PAID for a specific car (BMW X
I was new to the area and in need of a barber that could hook up a nice fade. I Finally found Senior's and I couldn't be
I've never been let down in the 15 yrs I've been enjoying Barro's food. You have to think about 15 year's numerous locati
98% of the seats and armrest in the imax theater are torn. Paying this much money for a horrible comfort I'm not satisfie
My husband and I moved near this bar about 10 years ago. We have probably gone to this bar over 100 times.

Location - great, atmosphere - great, over all experience - great, but the last 2 years have been frustrating. Drinks are

The main problem I have is the newer service people. There is a blond that has taken multiple orders from me in the last

It makes me sad that this place has deteriorated so badly over the last 2 years.
We were contacted by Scott Roofing as a result of a request on To Fix It website.  The person who contacted us on the pho

The estimator told me (3) times he was a third generation roofer, but never mentioned how long he had been roofing, which

When he returned he said my flat part of my roof was in seriously bad condition and that it is leaking.  He quoted me pri

I actually had the roof work done by a reputable roofer for a third of the price.
Awesome Doctor! My kids loved the fish tank & the staff too - very kid friendly. He was recommended by the Dr who saw us
They have free patient portal (we had to pay a fee for portal when we lived in Houston), I use it all the time and I get
If you are looking for in and out office visits with no real interest in your kids this is not the place for you. Here yo
We have been seeing Dr. Vaughan for the past 3yrs and i am so happy.
I went to a Majerle's near US Airways Arena after a Suns game, and it was a lot of fun, but it was very packed inside.

The food is pretty decent for bar food.  I got a chicken salad of some sort when I visited.  I didn't have a beer that ni

Dan Majerle would be proud of this establishment that has a great sports vibe.
Often times businesses have the wrong people answering their phone's.   But over the last two days I have called Nick &

If the pizza is as good as their service, it will be excellent!
Wow, we've been missing out, should've tried this a lot sooner. Our NYE at home featured pizza and tiki drinks. My bf ca

We got the thin crust option with italian sausage, mushrooms and olives. So good!! like other reviewers said, the sauce i

What a great find in our (relative) neighborhood. I can't wait to try Spinato's again.
```

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

```
SELECT
b.name,
b.city,
b.state,
b.stars,
h.hours,
b.review_count,
b.is_open,
c.category,
r.text
```

```
FROM business AS b
LEFT JOIN category AS c ON b.id = c.business_id
LEFT JOIN hours AS h ON b.id = h.business_id
LEFT JOIN attribute AS a ON b.id = a.business_id
LEFT JOIN review AS r ON b.id = r.business_id
WHERE (r.text IS NOT NULL
AND
b.state =='AZ')
GROUP BY b.city;
```