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 distinct id's, 8090 business_id's, 9581 user_id's
vi. Checkin = 493 distinct business_id
vii. Photo = 6493 distinct business_id, 10000 id
viii. Tip = 537 distinct user_id, 3979 business_id
ix. User = 10000 id
x. Friend = 11 distinct user_id
xi. Elite_years = 2780 distinct 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
```

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

Copy and Paste the Result Below:

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

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)
FROM business
WHERE city = 'Avon'
GROUP By stars
ORDER By stars DESC
```

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

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

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars, SUM(review_count)
FROM business
WHERE city = 'Beachwood'
GROUP By stars
ORDER By stars DESC
```

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

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

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

SQL code used to arrive at answer:

```
SELECT *
FROM user
ORDER BY review_count DESC
LIMIT 3
```

Copy and Paste the Result Below:

id	name	review_count	yelping_since	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cute	compliment_list	compliment_note	compliment_plain	compliment_cool	compliment_funny	compliment_writer	compliment_photos
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	2000	2012-12-16 00:00:00	17524	2324	15008	253	3.6	206	58	19	1	0	352	667	704	704	430	220

```
        | -3s52C4zL_DHRK0ULG6qtg | Sara   |         1629 | 2010-05-16 00:00:00 |
25 |    10 |     2 |   50 |          3.42 |             11 |                6 |
           1 |               2 |                0 |              19 |
44 |              16 |               16 |                11 |                 2 |
        | -8lbUNlXVSoXqaRRiHiSNg | Yuri   |         1339 | 2008-01-03 00:00:00 |
1166 |   220 |   561 |   76 |          4.11 |             109 |                8 |
              6 |               3 |               14 |              34 |
79 |              91 |               91 |                41 |                47 |

+------------------------+--------+-------------+--------------------+--------+---
----+-------+------+---------------+----------------+-----------------+------------
-------+-----------------+-----------------+-----------------+------------------+---
-------------+------------------+------------------+------------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:
Posting more reviews does not correlate with more fans. If there was a correlation, we'd either see that there would be a decrease or increase in the decrease of fans, but the change in review_count is independent of the change in fans

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

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

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

Answer:
There are more reviews with the word love in them. There is a total of 1780 words with "love" in them while there are only 232 words with "hate" in them

SQL code used to arrive at answer:

```
SELECT COUNT(text)
FROM review
WHERE text LIKE "%hate%"
+-------------+
| COUNT(text) |
+-------------+
|         232 |
+-------------+
```

```
SELECT COUNT(text)
FROM review
WHERE text LIKE "%love%"
+-------------+
| COUNT(text) |
+-------------+
|        1780 |
+-------------+
```

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?
Yes. The hours spent at 4-5 star restaurants in Las Vegas either started earlier or had less hours if starting at the same time

ii. Do the two groups you chose to analyze have a different number of reviews?
Yes. There were more reviews for the 4-5 star restaurants

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
These locations had different very different zipcodes so I can not make any inferences as to the location data

SQL code used for analysis:

```
SELECT bus.name
        ,cat.category
        ,bus.city
        , bus.postal_code
        ,hours ,
CASE
        WHEN stars BETWEEN 2 AND 3
        THEN '2-3 stars'
        WHEN stars BETWEEN 4 AND 5
        THEN '4-5 stars'
        END AS rating,
                bus.review_count
FROM business bus
INNER JOIN hours h ON bus.id = h.business_id
INNER JOIN category cat ON cat.business_id = bus.id
WHERE city = 'Las Vegas'
        AND category = 'Restaurants'
        AND (rating = '2-3 stars' OR rating = '4-5 stars')
GROUP BY name
```

```
ORDER BY stars DESC
```

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:
There was almost an average of 9 more reviews for the businesses that are open than closed

ii. Difference 2:
There is a higher average rating for these restaurants by approximately .16 stars

SQL code used for analysis:

```
SELECT AVG(stars) as average_stars
    , AVG(review_count) as average_reviews
        , 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.

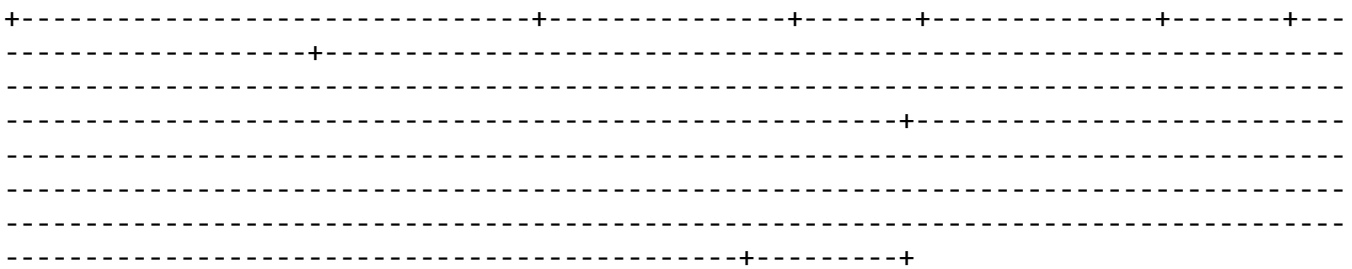
Ideas for analysis include: Parsing out keywords and business attributes for sentiment analysis, clustering businesses to find commonalities or anomalies between them, predicting the overall star rating for a business, predicting the number of fans a user will have, and so on. These are just a few examples 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 decided to predict the overall star rating for a business based on the hours, reviews, location, and type of business

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
This data helps businesses observe how other businesses in similar categories have higher or lower ratings showing them what they should do. For the data, I will need to have the review table and the business table essentially in order to grab not only the places that the businesses are in but also the actual reviews towards the businesses and the stars. Having something on the categories and hours would be helpful in finding patterns or correlations as to how businesses can receive more

stars

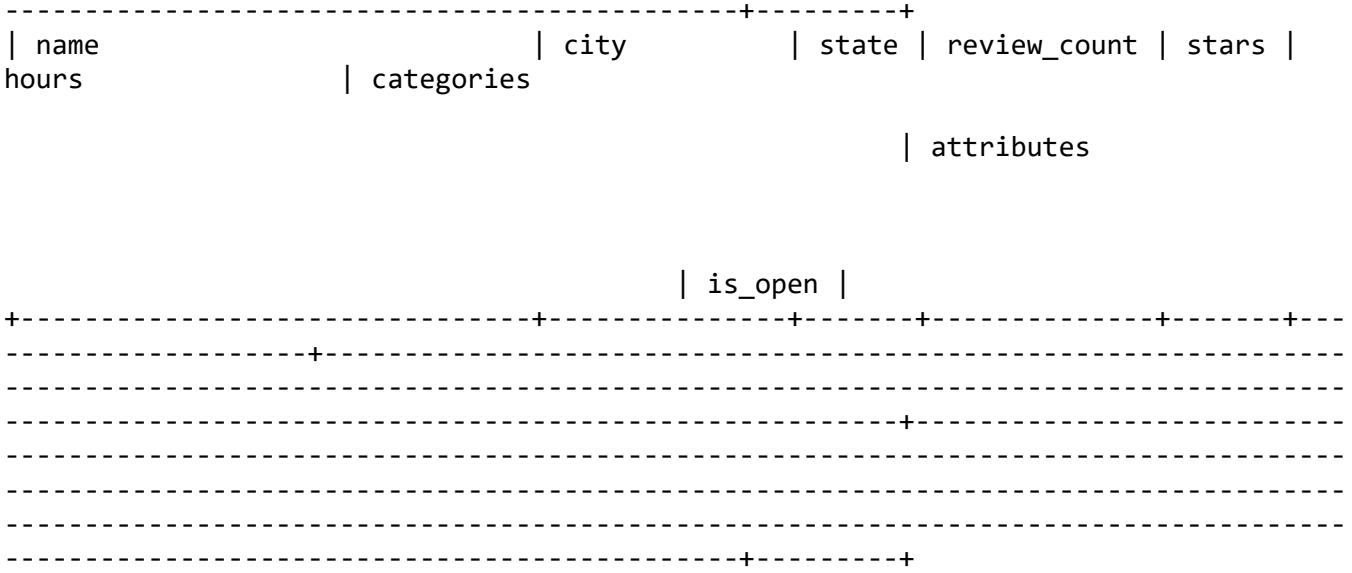
iii. Output of your finished dataset:



```
| name                              | city          | state | review_count | stars |
hours                | categories

                                                        | attributes


                                          | is_open |
```



```
| Flaming Kitchen                   | Markham       | ON    |           25 |   3.0 |
Saturday|12:00-23:00 | Asian Fusion,Restaurants

                                                       |
RestaurantsTableService,GoodForMeal,Alcohol,Caters,HasTV,RestaurantsGoodForGroups,No
iseLevel,WiFi,RestaurantsAttire,RestaurantsReservations,OutdoorSeating,RestaurantsPr
iceRange2,BikeParking,RestaurantsDelivery,Ambience,RestaurantsTakeOut,GoodForKids,Bu
sinessParking                                                          |       1 |
| Freeman's Car Stereo              | Charlotte     | NC    |            8 |   3.5 |
Saturday|9:00-17:00  | Electronics,Shopping,Automotive,Car Stereo Installation

                                                       |
BusinessAcceptsCreditCards,RestaurantsPriceRange2,BusinessParking,WheelchairAccessib
le

                                                                       |       1 |
| Motors & More                     | Las Vegas     | NV    |            7 |   5.0 |
Saturday|8:00-12:00  | Home Services,Solar Installation,Heating & Air
Conditioning/HVAC
                                                                        |
BusinessAcceptsCreditCards,BusinessAcceptsBitcoin,ByAppointmentOnly
```

```
                                                                    |       1 |
| Baby Cakes                    | Willoughby    | OH    |            5 |   3.5 |
Saturday|10:00-17:00 | Bakeries,Food

                                                          |
BusinessAcceptsCreditCards,RestaurantsTakeOut,WheelchairAccessible,RestaurantsDelive
ry

                                                                     |       1 |
| Snip-its Rocky River          | Rocky River   | OH    |           18 |   2.5 |
Saturday|9:00-17:30  | Beauty & Spas,Hair Salons

                                                          |
BusinessAcceptsCreditCards,RestaurantsPriceRange2,GoodForKids,BusinessParking,ByAppo
intmentOnly

                                                                     |       1 |
| Standard Restaurant Supply    | Phoenix       | AZ    |           15 |   3.5 |
Saturday|9:00-17:00  | Shopping,Wholesalers,Restaurant Supplies,Professional
Services,Wholesale Stores
                                                                |
BusinessAcceptsCreditCards,RestaurantsPriceRange2,BusinessParking,BikeParking,Wheelc
hairAccessible

                                                                     |       1 |
| What A Bagel                  | York          | ON    |            8 |   3.0 |
Saturday|6:00-15:30  | Restaurants,Bagels,Breakfast & Brunch,Food

                                                          |
NoiseLevel,RestaurantsAttire,RestaurantsTableService,OutdoorSeating


                                                                    |       1 |
| Pinnacle Fencing Solutions    | Phoenix       | AZ    |           13 |   4.0 |
Monday|8:00-16:00    | Home Services,Contractors,Fences & Gates

                                                          |
BusinessAcceptsCreditCards,ByAppointmentOnly


                                                                    |       1 |
| Alterations Express           | Strongsville  | OH    |            3 |   4.0 |
Saturday|8:00-18:00  | Shopping,Bridal,Dry Cleaning & Laundry,Local Services,Sewing
& Alterations
                                                          |
BusinessParking,BusinessAcceptsCreditCards,RestaurantsPriceRange2,BusinessAcceptsBit
coin,BikeParking,ByAppointmentOnly,WheelchairAccessible

                                                                     |       1 |
| Extra Space Storage           | Chandler      | AZ    |            5 |   4.0 |
```

```
Saturday|8:00-17:30  | Home Services,Self Storage,Movers,Shopping,Local
Services,Home Decor,Home & Garden
                                                                    |
BusinessAcceptsCreditCards


                                                                    |       1 |
| Gussied Up                      | Toronto       | ON    |            6 |   4.5 |
Saturday|11:00-17:00 | Women's Clothing,Shopping,Fashion

                                                        |
BusinessAcceptsCreditCards,RestaurantsPriceRange2,BusinessParking,BikeParking


                                                                    |       1 |
| Buddy's Muffler & Exhaust       | Gastonia      | NC    |            4 |   5.0 |
Saturday|9:00-15:00  | Automotive,Auto Repair

                                                       | BusinessAcceptsCreditCards



                                          |       1 |
| Five Guys                       | Phoenix       | AZ    |           63 |   3.5 |
Saturday|10:00-22:00 | American (New),Burgers,Fast Food,Restaurants

                                                        |
RestaurantsTableService,GoodForMeal,Alcohol,Caters,HasTV,RestaurantsGoodForGroups,No
iseLevel,WiFi,RestaurantsAttire,RestaurantsReservations,OutdoorSeating,BusinessAccep
tsCreditCards,RestaurantsPriceRange2,BikeParking,RestaurantsDelivery,Ambience,Restau
rantsTakeOut,GoodForKids,DriveThru,BusinessParking                      |       1 |
| All Storage - Anthem            | Henderson     | NV    |            3 |   3.5 |
Saturday|9:00-16:30  | Truck Rental,Local Services,Self Storage,Parking,Automotive

                                                        |
BusinessAcceptsCreditCards,BusinessAcceptsBitcoin


                                                                    |       1 |
| Mood                            | Edinburgh     | EDH   |           11 |   2.0 |
Thursday|22:30-3:00  | Dance Clubs,Nightlife

                                                        |
Alcohol,OutdoorSeating,BusinessAcceptsCreditCards,RestaurantsPriceRange2,AgesAllowed
,Music,Smoking,RestaurantsGoodForGroups,WheelchairAccessible

                                                                     |       0 |
| Starbucks                       | Phoenix       | AZ    |           52 |   3.0 |
Saturday|5:00-20:00  | Coffee & Tea,Food
```

```
                                                                |
BusinessParking,Caters,WiFi,OutdoorSeating,BusinessAcceptsCreditCards,RestaurantsPri
ceRange2,BikeParking,RestaurantsTakeOut

                                                                 |       1 |
| Big Smoke Burger               | Toronto       | ON    |           47 |   3.0 |
Saturday|10:30-21:00 | Poutineries,Burgers,Restaurants

                                                                |
RestaurantsTableService,GoodForMeal,Alcohol,Caters,HasTV,RestaurantsGoodForGroups,No
iseLevel,WiFi,RestaurantsAttire,RestaurantsReservations,OutdoorSeating,BusinessAccep
tsCreditCards,RestaurantsPriceRange2,WheelchairAccessible,BikeParking,RestaurantsDel
ivery,Ambience,RestaurantsTakeOut,GoodForKids,DriveThru,BusinessParking |       1 |
| Subway                         | Charlotte     | NC    |            7 |   3.5 |
Saturday|10:00-21:00 | Fast Food,Restaurants,Sandwiches

                                                                |
Ambience,RestaurantsPriceRange2,GoodForKids


                                                                 |       1 |
| Red Rock Canyon Visitor Center | Las Vegas     | NV    |           32 |   4.5 |
Saturday|8:00-16:30  | Education,Visitor Centers,Professional Services,Special
Education,Local Services,Community Service/Non-Profit,Hotels & Travel,Travel
Services,Gift Shops,Shopping,Parks,Hiking,Flowers & Gifts,Active Life |
BusinessAcceptsCreditCards,GoodForKids


                                                                 |       1 |
| Scent From Above Company       | Scottsdale    | AZ    |           14 |   4.5 |
Monday|6:00-16:00    | Home Cleaning,Local Services,Professional Services,Carpet
Cleaning,Home Services,Office Cleaning,Window Washing
                                                          |
BusinessAcceptsCreditCards,ByAppointmentOnly


                                                                 |       1 |
| The Charlotte Room             | Toronto       | ON    |           10 |   3.5 |
Saturday|18:00-2:00  | Event Planning & Services,Bars,Nightlife,Lounges,Pool
Halls,Venues & Event Spaces
                                                              |
BusinessParking,HasTV,CoatCheck,NoiseLevel,OutdoorSeating,BusinessAcceptsCreditCards
,RestaurantsPriceRange2,Music,WheelchairAccessible,Smoking,Ambience,BestNights,Resta
urantsGoodForGroups,HappyHour,GoodForDancing,Alcohol
                                                                   |       0 |
| PC Savants                     | Sun City      | AZ    |           11 |   5.0 |
Saturday|11:00-18:00 | IT Services & Computer Repair,Electronics Repair,Local
Services,Mobile Phone Repair
                                                             |
BusinessAcceptsCreditCards,BusinessAcceptsBitcoin
```

```
                                                                    |       1 |
| Sweet Ruby Jane Confections    | Las Vegas     | NV    |           30 |   4.0 |
Saturday|10:00-19:00 | Food,Chocolatiers & Shops,Bakeries,Specialty Food,Desserts

                                                    |
BusinessAcceptsCreditCards,RestaurantsPriceRange2,BusinessParking,WheelchairAccessib
le

                                                                     |       0 |
| Oinky's Pork Chop Heaven       | North Randall | OH    |            3 |   3.0 |
Saturday|6:00-23:00  | Soul Food,Restaurants

                                                    |
RestaurantsAttire,RestaurantsGoodForGroups,GoodForKids,RestaurantsReservations,Resta
urantsTakeOut

                                                                     |       1 |
| Sushi Osaka                    | Toronto       | ON    |            8 |   4.5 |
Saturday|11:00-23:00 | Sushi Bars,Restaurants,Japanese,Korean

                                                    |
RestaurantsTakeOut,WiFi,RestaurantsGoodForGroups,RestaurantsReservations


                                                                    |       1 |
+-------------------------------+--------------+-------+-------------+-------+---
------------------+-----------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------+---------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-----------------------------------------------+---------+
```

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

```
                SELECT B.name,
                           B.city,
                           B.state,
                           B.review_count,
                           B.stars,
                           H.hours,
                           GROUP_CONCAT(DISTINCT(C.category)) AS categories,
                           GROUP_CONCAT(DISTINCT(A.name)) AS attributes,
                           B.is_open
                FROM business B
                INNER JOIN hours H
                ON B.id = H.business_id
                INNER JOIN category C
```

```
ON B.id = C.business_id
INNER JOIN attribute A
ON B.id = A.business_id
GROUP BY B.id
```