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 primary key - business_id
ii. Hours = 1562 foreign key - business_id
iii. Category = 2643 foreign key - business_id
iv. Attribute = 1115 foreign key - business_id

v. Review = 10000 primary key - id
8090 foreign key – business_id
9581 foreign key – user_id
vi. Checkin = 493 foreign key – business_id
vii. Photo = 10000 primary key - id
6493 foreign key – business_id
viii. Tip = 3979 foreign key – business_id
537 foreign key – user_id
ix. User = 10000 primary key - id
x. Friend = 11 foreign key – user_id
xi. Elite_years = 2780 foreign key – 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 *
FROM User
WHERE 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 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 SUM_RC
FROM business
GROUP BY city
ORDER BY SUM_RC DESC
```

Copy and Paste the Result Below:

```
+-----------------+--------+
| city            | SUM_RC |
+-----------------+--------+
| 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
       ,count(stars) AS Stars_Count
FROM business
WHERE city = 'Avon'
GROUP BY stars
```

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

```
+-------+-------------+
| stars | Stars_Count |
+-------+-------------+
|   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:

```
SELECT stars
        ,count(stars) AS Stars_Count
FROM business
WHERE city = 'Beachwood'
GROUP BY stars
```

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

```
+-------+-------------+
| stars | Stars_Count |
+-------+-------------+
|   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
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
        ,review_count
        ,fans
FROM user
ORDER BY fans DESC
        ,review_count DESC
```

```
               LIMIT 5
```

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

Posting more reviews does correlate with more fans, but that’s not the only factor. As you can see, Gerald’s review count is 2000, but Amy has more fans than Gerald though her review Count is lesser.

```
               SELECT id
                       ,name
                       ,review_count
                       ,fans
               FROM user
               ORDER BY fans ASC
                       ,review_count ASC LIMIT 5
```

```
+------------------------+---------------+--------------+------+
| id                     | name          | review_count | fans |
+------------------------+---------------+--------------+------+
| -61V4ZkRsKUChYFZtdZDvQ | Sonnenschein1 |            0 |    0 |
| -9TyYbKtEz-pxeZyLICOgA | svenher       |            0 |    0 |
| -arJ-0bq2eycINnHrm0LFA | Schweinefe    |            0 |    0 |
| -d8nnc-pp6qj_6qnp4IN-g | Luke          |            0 |    0 |
| -Dhxu5B36bkm65ciME0vxg | Limon-Du      |            0 |    0 |
```

At the same time, if reviews don’t exist, there are no fans at all, as shown above. So, there is a positive correlation between review counts and fans, but other factors also affect these values.

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

Answer: number of reviews with the word “love” is 1780. The number of reviews with the word “hate” is 232. So, there are more reviews with the word “love”.

SQL code used to arrive at answer:

```
select count(*) from review

where text like '%love%'

select count(*) 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
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 have picked the city as Toronto and category as Restaurant.

Businesses with 2-3 star ratings:

```
+------------------+-------+-----------------------+------------------------+------+
| name             | stars | hours                 | neighborhood           | r_ct |
+------------------+-------+-----------------------+------------------------+------+
| Big Smoke Burger |   3.0 | Monday|10:30-21:00    | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Tuesday|10:30-21:00   | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Friday|10:30-21:00    | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Wednesday|10:30-21:00 | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Thursday|10:30-21:00  | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Sunday|11:00-19:00    | Downtown Core          |   47 |
| Big Smoke Burger |   3.0 | Saturday|10:30-21:00  | Downtown Core          |   47 |
| Pizzaiolo        |   3.0 | Monday|9:00-23:00     | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Tuesday|9:00-23:00    | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Friday|9:00-4:00      | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Wednesday|9:00-23:00  | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Thursday|9:00-23:00   | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Sunday|10:00-23:00    | Entertainment District |   34 |
| Pizzaiolo        |   3.0 | Saturday|10:00-4:00   | Entertainment District |   34 |
| 99 Cent Sushi    |   2.0 | Monday|11:00-23:00    | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Tuesday|11:00-23:00   | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Friday|11:00-23:00    | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Wednesday|11:00-23:00 | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Thursday|11:00-23:00  | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Sunday|11:00-23:00    | Downtown Core          |    5 |
| 99 Cent Sushi    |   2.0 | Saturday|11:00-23:00  | Downtown Core          |    5 |
+------------------+-------+-----------------------+------------------------+------+
```

```
Businesses with 4-5 star ratings:
+-------------+-------+-----------------------+--------------+------+
| name        | stars | hours                 | neighborhood | r_ct |
+-------------+-------+-----------------------+--------------+------+
| Cabin Fever |   4.5 | Monday|16:00-2:00     | High Park    |   26 |
| Cabin Fever |   4.5 | Tuesday|18:00-2:00    | High Park    |   26 |
| Cabin Fever |   4.5 | Friday|18:00-2:00     | High Park    |   26 |
| Cabin Fever |   4.5 | Wednesday|18:00-2:00  | High Park    |   26 |
| Cabin Fever |   4.5 | Thursday|18:00-2:00   | High Park    |   26 |
| Cabin Fever |   4.5 | Sunday|16:00-2:00     | High Park    |   26 |
| Cabin Fever |   4.5 | Saturday|16:00-2:00   | High Park    |   26 |
| Sushi Osaka |   4.5 | Monday|11:00-23:00    | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Tuesday|11:00-23:00   | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Friday|11:00-23:00    | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Wednesday|11:00-23:00 | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Thursday|11:00-23:00  | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Sunday|14:00-23:00    | Etobicoke    |    8 |
| Sushi Osaka |   4.5 | Saturday|11:00-23:00  | Etobicoke    |    8 |
| Edulis      |   4.0 | Sunday|12:00-16:00    | Niagara      |   89 |
| Edulis      |   4.0 | Friday|18:00-23:00    | Niagara      |   89 |
| Edulis      |   4.0 | Wednesday|18:00-23:00 | Niagara      |   89 |
| Edulis      |   4.0 | Thursday|18:00-23:00  | Niagara      |   89 |
| Edulis      |   4.0 | Saturday|18:00-23:00  | Niagara      |   89 |
+-------------+-------+-----------------------+--------------+------+
```

i. Do the two groups you chose to analyze have a different distribution of hours?

Yes, the distribution of hours is different. Cabin Fever Restaurant which has the highest rating of 4.5 is open in the late hours 2.00 which definitely seems like an advantage. On the other hand, restaurants with low ratings esp. 99 cent Sushi is open only till 23.00.

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

Yes, the two groups have different reviews, but nothing can be inferred much from the review count.   For example, Sushi Osaka has a high rating of 4.5, but the review count is only 8.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.

All the neighborhoods seem like important places. HighPark, Etobicoke, and Niagara are all major neighborhoods, and the businesses located there have a 4-5 rating. At the same time, Downtown Core and Entertainment District are busy places, but the businesses there have only 2-3 ratings. So, no inference can be made using location.

SQL code used for analysis:

Businesses in 'Toronto' with category 'Restaurants':

Rating: 2-3:

```
SELECT b.name
        ,b.stars
        ,h.hours
        ,b.neighborhood
        ,b.review_count AS r_ct
FROM business b
JOIN category c ON b.id = c.business_id
JOIN hours h ON b.id = h.business_id
WHERE city = 'Toronto'
        AND category = 'Restaurants'
        AND stars BETWEEN 2
                AND 3
ORDER BY stars DESC
```

Rating 4-5:

```
SELECT b.name
        ,b.stars
        ,h.hours
        ,b.neighborhood
        ,b.review_count AS r_ct
FROM business b
JOIN category c ON b.id = c.business_id
JOIN hours h ON b.id = h.business_id
WHERE city = 'Toronto'
        AND category = 'Restaurants'
        AND stars BETWEEN 4
                AND 5
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.

```
+---------+----------------+---------------+-------------------+
| is_open | count(is_open) |    avg(stars) | avg(review_Count) |
+---------+----------------+---------------+-------------------+
|       0 |           1520 | 3.52039473684 |     23.1980263158 |
|       1 |           8480 | 3.67900943396 |     31.7570754717 |
+---------+----------------+---------------+-------------------+
```

i. Difference 1:

The average star for open businesses is 3.68, while the average star for closed businesses is 3.52. This means the businesses that are open have a slightly higher rating.

ii. Difference 2:

The average review count for businesses that are open is 31.76, and the average review count for businesses that are closed is 23.20. This means the businesses that are open have a higher review count.

SQL code used for analysis:

```
SELECT is_open
        ,count(is_open)
        ,avg(stars)
        ,avg(review_Count)
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:
The type of analysis I like to do is : Creating a Recommendation Engine for the user.

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

Recommendation Engine for the user:
The user has given ratings for some of the businesses. Based on the reviews given by the user, we can find similar businesses the user may like, and we can recommend that to the user.

For example, in the below dataset, the second row, the user gave a 5-star rating to '808 Sushi', which is in the 'Southwest' neighborhood in 'Las Vegas.' Based on this, we can recommend other Japanese restaurants in the same neighborhood with a good rating to the user.

iii. Output of your finished dataset:

```
+------------------------+-------------+-----------------------------------------+---------------
--+------------+----------------+
| user_id                | user_rating | business_name                           |
business_rating | city       | neighborhood   |
+------------------------+-------------+-----------------------------------------+---------------
--+------------+----------------+
| zh9vXKAaUAErsqxY-0mHWw |           5 | Spinato's Pizza                         |
4.5 | Tempe      |                |
| QqF3cU-IkgmNNRxHwKIZ2w |           5 | 808 Sushi                               |
3.5 | Las Vegas  | Southwest      |
| 0tpJmUYvbQSRbNa6DI0WVA |           5 | Kimberfire                              |
5.0 | Toronto    | Downtown Core  |
| sip_xNt4-6y70S6MVxDACA |           5 | Vanity Nails & Spa                      |
3.5 | Las Vegas  | Southeast      |
| n1h8zhEt2x1nGH8hPcbEmw |           5 | Ocean Blue Caribbean Restaurant and Bar |
3.5 | Chandler   |                |
| c95xNHRgG_pGmZCZQEwoHw |           5 | D & D Discount Motorcycles              |
5.0 | Tempe      |                |
| nSU3-MtoodU0EQDoY4nBPQ |           5 | El Fish Taco                            |
4.5 | Las Vegas  | Southeast      |
| _GX0dMS_5sJoaKmDfY8SwA |           5 | Michael Mina                            |
4.0 | Las Vegas  | The Strip      |
| XaWdI5CnfNLAp0ERO1Xl_A |           5 | Food Palace Gelato                      |
4.0 | Toronto    | Alexandra Park |
| G4-nOvLBU4nZWxpASiASRg |           5 | Pizza Taglio                            |
4.0 | Pittsburgh | East Liberty   |
| EU0Vma7jgzDN2ax6f3keJw |           5 | Nandini Indian Cuisine                  |
4.5 | Tempe      |                |
| fRclDad6qMwgW_l3jtRqig |           5 | Tortilla Fish                           |
4.5 | Phoenix    |                |
| HDO6J5DrptQMjC0iccV1Ig |           5 | Greens and Proteins                     |
4.0 | Las Vegas  | Spring Valley  |
| 4cBTUgitY98C-y8rW1-crw |           5 | Pam's Caribbean Kitchen                 |
4.0 | Toronto    | Dovercourt     |
| SCWJXT-8faRzx_2L3lqDDg |           5 | Kimberfire                              |
5.0 | Toronto    | Downtown Core  |
| PTf6pH-zCMshuocmMpNlwA |           5 | Oakmont Bakery                          |
4.5 | Oakmont    |                |
| Km0uiEr3PABtRLlHo4I_zw |           5 | Pier W                                  |
4.0 | Lakewood   |                |
| KgmNR7n0H9wbBIlY_CEh5Q |           5 | The Art Theater                         |
5.0 | Champaign  |                |
| MClheGeTfikaRVNTkxW5mA |           5 | Hertz Rent A Car                        |
2.5 | Las Vegas  | Southeast      |
| FwkBQO-lA-EgkA8t_EHqfw |           5 | Woodlot Restaurant                      |
4.0 | Toronto    | Little Italy   |
```

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

```
SELECT r.user_id
        ,r.stars AS user_rating
        ,b.name AS business_name
        ,b.stars AS business_rating
        ,b.city
        ,b.neighborhood
FROM review r
JOIN business b ON r.business_id = b.id
ORDER BY user_rating DESC limit 20
```