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

2. Find the total distinct records by either the foreign key or primary key for each table. If two foreign keys are listed in the table, please specify which foreign key.

i. Business = primary: 10,000
ii. Hours = business_id: 1,562
iii. Category = business_id: 2643
iv. Attribute = business_id: 1115
v. Review = primary: 10,000, business_id: 8090, user_id: 9581
vi. Checkin = business_id: 493
vii. Photo = primary: 10,000, business_id: 6493
viii. Tip = user_id: 537, business_id: 3979
ix. User = primary: 10,000
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 *
FROM user
WHERE id is null or name is null or ...
```

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 review_count,
		city
	FROM business
	ORDER BY review_count DESC
```

Copy and Paste the Result Below:

```
+--------------+------------+
| review_count | city       |
+--------------+------------+
|         3873 | Las Vegas  |
|         1757 | MontrÃ©al   |
|         1549 | Gilbert    |
|         1410 | Las Vegas  |
|         1389 | Las Vegas  |
|         1252 | Las Vegas  |
|         1116 | Las Vegas  |
|          1084 | Las Vegas  |
|           961 | Las Vegas  |
|           902 | Gilbert    |
|           864 | Las Vegas  |
|           823 | Scottsdale |
|           821 | Las Vegas  |
|           786 | Las Vegas  |
|           785 | Henderson  |
|           778 | Toronto    |
|           768 | Las Vegas  |
|           758 | Las Vegas  |
|           726 | Scottsdale |
|           723 | Cleveland  |
|           720 | Las Vegas  |
|           715 | Charlotte  |
|           711 | Phoenix    |
|           706 | Las Vegas  |
|           700 | Phoenix    |
+--------------+------------+
```

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

i. Avon

SQL code used to arrive at answer:

```
SELECT review_count,
        stars
FROM business
WHERE city = "Avon"
ORDER BY review_count DESC
```

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

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

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT review_count,
        stars
FROM business
WHERE city = "Beachwood"
ORDER BY review_count DESC
```

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

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

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

        SQL code used to arrive at answer:

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

        Copy and Paste the Result Below:

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

8. Does posing more reviews correlate with more fans?

        Please explain your findings and interpretation of the results:

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

result:

```
+-----------+------------------------+--------------+------+
| name      | id                     | review_count | fans |
+-----------+------------------------+--------------+------+
| Gerald    | -G7Zkl1wIWBBmD0KRy_sCw |         2000 |  253 |
| Sara      | -3s52C4zL_DHRK0ULG6qtg |         1629 |   50 |
| Yuri      | -8lbUNlXVSoXqaRRiHiSNg |         1339 |   76 |
| .Hon      | -K2Tcgh2EKX6e6HqqIrBIQ |         1246 |  101 |
| William   | -FZBTkAZEXoP7CYvRV2ZwQ |         1215 |  126 |
| Harald    | --2vR0DIsmQ6WfcSzKWigw |         1153 |  311 |
| eric      | -gokwePdbXjfS0iF7NsUGA |         1116 |   16 |
| Roanna    | -DFCC64NXgqrxlO8aLU5rg |         1039 |  104 |
| Mimi      | -8EnCioUmDygAbsYZmTeRQ |          968 |  497 |
| Christine | -0IiMAZI2SsQ7VmyzJjokQ |          930 |  173 |
| Ed        | -fUARDNuXAfrOn4WLSZLgA |          904 |   38 |
| Nicole    | -hKniZN2OdshWLHYuj21jQ |          864 |   43 |
| Fran      | -9da1xk7zgnnfO1uTVYGkA |          862 |  124 |
| Mark      | -B-QEUESGWHPE_889WJaeg |          861 |  115 |
| Christina | -kLVfaJytOJY2-QdQoCcNQ |          842 |   85 |
| Dominic   | -kO6984fXByyZm3_6z2JYg |          836 |   37 |
| Lissa     | -lh59ko3dxChBSZ9U7LfUw |          834 |  120 |
| Lisa      | -g3XIcCb2b-BD0QBCcq2Sw |          813 |  159 |
| Alison    | -l9giG8TSDBG1jnUBUXp5w |          775 |   61 |
| Sui       | -dw8f7FLaUmWR7bfJ_Yf0w |          754 |   78 |
| Tim       | -AaBjWJYiQxXkCMDlXfPGw |          702 |   35 |
| L         | -jt1ACMiZljnBFvS6RRvnA |          696 |   10 |
| Angela    | -IgKkE8JvYNWeGu8ze4P8Q |          694 |  101 |
| Crissy    | -hxUwfo3cMnLTv-CAaP69A |          676 |   25 |
| Lyn       | -H6cTbVxeIRYR-atxdielQ |          675 |   45 |
+-----------+------------------------+--------------+------+
```

As shown from the result, as the review_count decreases, there's no decreasing pattern shown in the fans as expected. Therefore, there's no any positive correlationship between the two variables.

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

Answer: love

SQL code used to arrive at answer:

```
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 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?
I chose Toronto, with the category of Restaurants. They do have. One of the 4-5 stars restaurants have hours in the night, whereas the rest restaurants operate during noon.

ii. Do the two groups you chose to analyze have a different number of reviews?
Restaurants with high star rating tend to have more number of reviews.

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

SQL code used for analysis:

```
SELECT name,
    postal_code,
    longitude,
    stars,
    review_count,
    hours,
    c.category
FROM business b
INNER JOIN hours h ON b.id = h.business_id
INNER JOIN category c ON b.id = c.business_id
WHERE b.city = "Toronto"
AND c.category = "Restaurants"
GROUP BY stars
ORDER BY stars ASC
```

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:
The ones that are open tend to have more stars than those not.

ii. Difference 2:
The ones that are open tend to have more reviews than those not.

SQL code used for analysis:

```
SELECT is_open,
    AVG(review_count),
    AVG(stars)
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:

Which category is more popular and successful for a business.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
I will basically group categories and calculate the average star rating and review counts of each of them, and rank them in a descending order. For further analysis, we may also include the number of sample size into account. For example, some categories may have very high average of star ratings, but actually there are only 4 samples in this category, leading to an unfair comparison.

iii. Output of your finished dataset:

```
+----------------------------+----------------+---------------+--------------------+
| category                   | average_review |  average_star | number_of_business |
+----------------------------+----------------+---------------+--------------------+
| Restaurants                |  63.4366197183 | 3.45774647887 |                 71 |
| Shopping                   |  32.5666666667 | 3.98333333333 |                 30 |
| Food                       |  77.4347826087 | 3.78260869565 |                 23 |
| Nightlife                  |          67.55 |         3.475 |                 20 |
| Bars                       |  77.7647058824 |           3.5 |                 17 |
| Health & Medical           |  11.9411764706 | 4.08823529412 |                 17 |
| Home Services              |          5.875 |           4.0 |                 16 |
| Beauty & Spas              |  9.15384615385 | 3.88461538462 |                 13 |
| Local Services             |  8.33333333333 | 4.20833333333 |                 12 |
| American (Traditional)     |  102.545454545 | 3.81818181818 |                 11 |
| Active Life                |           13.1 |          4.15 |                 10 |
| Automotive                 |           22.0 |           4.5 |                  9 |
| Hotels & Travel            |  42.3333333333 | 3.22222222222 |                  9 |
| Burgers                    |         37.125 |         3.125 |                  8 |
| Sandwiches                 |         121.75 |        3.9375 |                  8 |
| Arts & Entertainment       |  55.4285714286 |           4.0 |                  7 |
| Fast Food                  |  26.4285714286 | 3.21428571429 |                  7 |
| Mexican                    |  46.7142857143 |           3.5 |                  7 |
| American (New)             |  80.1666666667 | 3.33333333333 |                  6 |
| Event Planning & Services  |  19.6666666667 |          3.75 |                  6 |
| Hair Salons                |  10.8333333333 | 4.08333333333 |                  6 |
| Bakeries                   |           47.8 |           4.1 |                  5 |
| Doctors                    |           11.0 |           4.2 |                  5 |
| Indian                     |           12.6 |           3.6 |                  5 |
| Japanese                   |           30.4 |           3.8 |                  5 |
+----------------------------+----------------+---------------+--------------------+
```

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

```
SELECT c.category,
    AVG(b.review_count) AS average_review,
    AVG(b.stars) AS average_star,
    COUNT(c.business_id) AS number_of_business
FROM business b
INNER JOIN category c ON b.id = c.business_id
GROUP BY c.category
ORDER BY number_of_business DESC
```