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 = 3003246173
ii. Business table = 526028173
iii. Category table = 2406449491
iv. Checkin table = 11762
v. elite_years table = 171764
vi. friend table = 10000
vii. hours table = 3605696075
viii. photo table = 10098471
ix. review table = 10191474
x. tip table = 59561
xi. user table = 20128005

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 = 31704530
ii. Hours = 600814040
iii. Category = 600816466
iv. Attribute = 600803295
v. Review = 10186608
vi. Checkin = 924
vii. Photo = 10086787
viii. Tip = 2744
ix. User = 6463373
x. Friend = 11
xi. Elite_years = 42167

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

Copy and Paste the Result Below:

Woodmere Village, Mount Lebanon, Charlotte, McMurray, and North York

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

i. Avon

SQL code used to arrive at answer:

```
SELECT city,
stars
FROM business
WHERE city = 'avon';
```

Copy and Paste the Resulting Table Below (2 columns – star rating and count):

```
+------+-------+
| city | stars |
+------+-------+
| Avon |   2.5 |
| Avon |   4.0 |
| Avon |   5.0 |
| Avon |   3.5 |
| Avon |   1.5 |
| Avon |   3.5 |
| Avon |   4.5 |
| Avon |   3.5 |
| Avon |   2.5 |
| Avon |   4.0 |
+------+-------+
```

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars,
review_count
FROM business
WHERE city = 'Beachwood';
```

Copy and Paste the Resulting Table Below (2 columns – star rating and count):

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

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

SQL code used to arrive at answer:

```
SELECT name,
review_count
FROM user
ORDER BY review_count desc
LIMIT 3;
```

Copy and Paste the Result Below:

```
+--------+--------------+
| name   | review_count |
+--------+--------------+
| Gerald |         2000 |
| Sara   |         1629 |
| Yuri   |         1339 |
+--------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results: There is no correlation between number of review and fans as seen in below table. There are more reviews, but less fans and vice versa.

```
+-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Gerald    |         2000 |  253 |
| Sara      |         1629 |   50 |
| Yuri      |         1339 |   76 |
| .Hon      |         1246 |  101 |
| William   |         1215 |  126 |
| Harald    |         1153 |  311 |
| eric      |         1116 |   16 |
| Roanna    |         1039 |  104 |
| Mimi      |          968 |  497 |
| Christine |          930 |  173 |
+-----------+--------------+------+
```

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

Answer: There are more reviews with word love than hate which are 1780 and 232 reviews respectivley.

SQL code used to arrive at answer:

```
SELECT *
FROM review
WHERE text like '%love%';
```

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? There is almost no hours information in cities that selected.

ii. Do the two groups you chose to analyze have a different number of reviews? Yes, they have different number of reviews; in catagory 2 and 3 stars, the number of review is 413, while its 1465 in stars rating 4 and 5.

iii. Are you able to infer anything from the location data provided between these two groups? Explain. I am looking into datas from Edinbugh city. The businesses are evely located in all neighborhoods based on their ratings.

SQL code used for analysis:

```
SELECT business.name, business.neighborhood, business.city, business.stars, business.review_count, hours.hours
FROM business
LEFT JOIN hours ON business.id = hours.business_id
WHERE city = 'Edinburgh' AND stars IN (2, 3, 4, 5)
ORDER BY review_count, 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:

Average Star rating for closed business: 3.5
Average Star rating for open business: 3.6

ii. Difference 2:

Average review counts for closed business: 23
Average review counts for open business: 31

SQL code used for analysis:

```
SELECT is_open, AVG(stars), AVG(review_count)
FROM business
WHERE is_open = 1
UNION
SELECT is_open, AVG(stars), AVG(review_count)
FROM business
```

```
WHERE is_open = 0;
```