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 prim key
ii. Hours = 1562 business id
iii. Category = 2643 business id
iv. Attribute = 1115 business id
v. Review = 10000 prim key
vi. Checkin = 493 business id
vii. Photo = 10000 prim key
viii. Tip = 3979 business_id
ix. User = 10000 prim key
x. Friend = 11 user_id
xi. Elite_years = 2780 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:

```
    -- Counting in how much rows ANY column contains a NULL value
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      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:
-- JOINING reviews with business table to get a table
-- with both reviews and the city of the review. The counting and grouping by city.
-- Lastly ordered by the number of reviews
SELECT
    COUNT(*) AS NumReviews
    , b.city
FROM review AS r
LEFT JOIN business AS b ON r.business_id = b.id
GROUP BY b.city
ORDER BY NumReviews DESC

    Copy and Paste the Result Below:
+------------+-----------------+
| NumReviews |            city |
+------------+-----------------+
```

```
|       9364 |            None |
|        193 |       Las Vegas |
|         65 |         Phoenix |
|         51 |         Toronto |
|         37 |      Scottsdale |
|         30 |       Henderson |
|         28 |           Tempe |
|         23 |      Pittsburgh |
|         22 |        Chandler |
|         21 |       Charlotte |
|         18 |        Montréal |
|         16 |         Madison |
|         13 |         Gilbert |
|         13 |            Mesa |
|         12 |       Cleveland |
|          6 | North Las Vegas |
|          5 |       Edinburgh |
|          5 |        Glendale |
|          5 |        Lakewood |
|          4 |      Cave Creek |
|          4 |       Champaign |
|          4 |         Markham |
|          4 |      North York |
|          3 |     Mississauga |
|          3 |        Surprise |
+------------+-----------------+
(Output limit exceeded, 25 of 68 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  COUNT(*) AS 'Count'
        , stars
FROM business
WHERE city = 'Avon'
GROUP BY stars

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):
+-------+-------+
| Count | stars |
+-------+-------+
|     1 |   1.5 |
|     2 |   2.5 |
|     3 |   3.5 |
|     2 |   4.0 |
|     1 |   4.5 |
|     1 |   5.0 |
+-------+-------+

ii. Beachwood

SQL code used to arrive at answer:
SELECT  COUNT(*) AS 'Count'
        , stars
FROM business
WHERE city = 'Beachwood'
GROUP BY stars

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):
+-------+-------+
| Count | stars |
+-------+-------+
|     1 |   2.0 |
|     1 |   2.5 |
|     2 |   3.0 |
|     2 |   3.5 |
|     1 |   4.0 |
|     2 |   4.5 |
|     5 |   5.0 |
+-------+-------+
```

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

SQL code used to arrive at answer:

```
SELECT  user_id
        , COUNT(*) AS NumOfReviews
FROM review
GROUP BY user_id
ORDER BY NumOfReviews DESC
LIMIT 3
```

Copy and Paste the Result Below:

```
+------------------------+--------------+
| user_id                | NumOfReviews |
+------------------------+--------------+
| CxDOIDnH8gp9KXzpBHJYXw |            7 |
| U4INQZOPSUaj8hMjLlZ3KA |            7 |
| 8teQ4Zc9jpl_ffaPJUn6Ew |            5 |
+------------------------+--------------+
```

8. Does posing more reviews correlate with more fans?

There are only 69 cases of which both the number of fans as well as the number of reviews are known.
On a total of 10000 for both reviews and users, this subset is to small to answer this question. Also. only three of those matches have done more than two reviews!

```
SELECT  r.user_id
        , u.name
        , COUNT(*) AS NumOfReviews
        , u.fans
FROM review AS r
LEFT JOIN user AS u ON r.user_id = u.id
WHERE u.name IS NOT NULL
GROUP BY user_id
ORDER BY NumOfReviews DESC
```

```
+------------------------+-----------+--------------+------+
| user_id                | name      | NumOfReviews | fans |
+------------------------+-----------+--------------+------+
| -0udWcFQEt2M8kM3xcIofw | Kaitlan   |            2 |    3 |
| -Biq3Dt8YhkRJEO_ITrvww | Christina |            2 |   27 |
| -l4iRZ2wYow39RTevk21Dg | Craig     |            2 |    1 |
| --Qh8yKWAvIP4V4K8ZPfHA | Dixie     |            1 |   41 |
| --i0PK1aTXScdV2UkNDkSQ | A'Starra  |            1 |    0 |
| -0DgO-WJ7yBjAihY_PoUpw | Tonia     |            1 |    0 |
| -0WUJsVpizkaAYQp05giUA | Jeffrey   |            1 |    0 |
| -0oUqPRPpbi2MyiK39cCTg | soragamii |            1 |    1 |
| -27BmUSrHjQQqItaFFIwxA | Cassandra |            1 |    0 |
| -3uEQhR9MXaC5QTHJ4lraw | Amy       |            1 |    2 |
| -4ARERfWgDsMfy1Pu7AbLA | Patricia  |            1 |    3 |
| -594af_E7Z9VVjQc9pJK3g | Andrea    |            1 |   10 |
| -5Y3idbK2Yzuz9munIF3tg | Laura     |            1 |    0 |
| -5psN9APmb8RcprBFA6lfw | Frank     |            1 |    0 |
| -7ABF3eAKE98kiufwQ2dng | Rachel    |            1 |    0 |
| -7Mo5iv_3t7u000eYaduGw | Ed        |            1 |    1 |
| -7bM_DeL2Kj2CuYuVDsLNg | Jade      |            1 |    6 |
| -7rFFU0fNQm4w0zn-r_9Xg | Ryan      |            1 |    1 |
| -8EE28ZzxVFxwjRqJuDumg | Meg       |            1 |    0 |
| -8nmj3B-tfY_vFiimtBOsw | Sanaz     |            1 |    3 |
| -973s-koCwNBKWLh2CdSYQ | Lynn      |            1 |    1 |
| -9S_Fh-sQebyB1yhEM5zHw | Bob       |            1 |    3 |
| -9ly39RQ8CvqxPuNZrGEbA | Brenda    |            1 |    0 |
| -ARdx8hOcEWlMDjzwLYZ_g | AJA       |            1 |   16 |
| -AkZkFH_md2-2kaSsvgrkg | Alan      |            1 |    1 |
+------------------------+-----------+--------------+------+
(Output limit exceeded, 25 of 69 total rows shown)
```

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
    (SELECT count(*)
    FROM review
    WHERE text LIKE '%hate%') AS NumOfHate

    ,(SELECT count(*)
    FROM review
    WHERE text LIKE '%love%') AS NumOfLove
```

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

SQL code used to arrive at answer:

```
SELECT  name
        , id
        , fans
FROM user
ORDER BY fans DESC
LIMIT 10
```

Copy and Paste the Result Below:

name	id	fans
Amy	-9I98YbNQnLdAmcYfb324Q	503
Mimi	-8EnCioUmDygAbsYZmTeRQ	497
Harald	--2vR0DIsmQ6WfcSzKWigw	311
Gerald	-G7Zkl1wIWBBmD0KRy_sCw	253
Christine	-0IiMAZI2SsQ7VmyzJjokQ	173
Lisa	-g3XIcCb2b-BD0QBCcq2Sw	159
Cat	-9bbDysuiWeo2VShFJJtcw	133
William	-FZBTkAZEXoP7CYvRV2ZwQ	126
Fran	-9da1xk7zgnnfO1uTVYGkA	124
Lissa	-lh59ko3dxChBSZ9U7LfUw	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, city las vegas and category Coffee & Tea

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

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
One is only from vegas, the other isnt. Dont really know what they want in this section.
SQL code used for analysis:

```
SELECT  COUNT(*)
        , CASE
        WHEN stars>=2.0 AND stars<=3.0 THEN '2-3 stars'
        WHEN stars>=4.0 AND stars <= 5.0 THEN '4-5 stars'
        END starCategory
FROM business AS b

INNER JOIN review AS h ON b.id = h.business_id

WHERE b.city = 'Las Vegas'
    AND starCategory IS NOT NULL
```

```
GROUP BY starCategory, hours
```

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: AVG Stars of open businesses is a bit larger

ii. Difference 2: More businesses are open atm

SQL code used for analysis:

```
SELECT  AVG(r.stars)
        , b.is_open

FROM business AS b
INNER JOIN review r ON b.id = r.business_id
GROUP BY is_open

SELECT  COUNT(*)
        , b.is_open

FROM business AS b
INNER JOIN review r ON b.id = r.business_id
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 want to match the location of the businesses with the number of reviews, stars, etc. and cluster those.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
I need a data set containing, the all information about the businesses with the location and name as indicators.

iii. Output of your finished dataset:

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