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 for id
ii. Hours = 1526 for business_id
iii. Category = 2643 for business_id
iv. Attribute = 1115 for business_id
v. Review = 10000 for id, 8090 for business_id, 9581 for user_id
vi. Checkin = 493 for business_id
vii. Photo = 10000 for id, 6493 for business_id
viii. Tip = 537 for user_id, 3979 for business_id
ix. User = 10000 for user
x. Friend = 11 for user_id
xi. Elite_years = 2780 for 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 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)
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 |
+-----------------+--------------------+
(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)
FROM business
WHERE city = 'Avon'
GROUP BY stars
```

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

stars	COUNT(stars)
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)
FROM business
WHERE city = 'Beachwood'
GROUP BY stars
```

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

stars	COUNT(stars)
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 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:
Apparently not, it is observed that people with more fans do not have more reviews.

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

Answer: Yes, there are more reviews with the word “love” (1780) than the word “hate” (232).

SQL code used to arrive at answer:

```
SELECT COUNT(text)
FROM review
WHERE text LIKE '%love%'

SELECT COUNT(text)
FROM review
WHERE text LIKE '%rate%'
```

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, for the city ‘Chandler’ and category ‘Nightlife’, businesses with 2-3 stars have longer opening hours on all days of the week than businesses with 4-5 stars.

ii. Do the two groups you chose to analyze have a different number of reviews?
Yes, businesses with 2-3 stars have more reviews than businesses with 4-5 stars.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
Just that the businesses have different addresses

SQL code used for analysis:

```
SELECT b.city, c.category, b.stars, h.hours, b.review_count, b.address
FROM business b INNER JOIN category c
ON b.id = c.business_id
INNER JOIN hours h ON b.id = h.business_id
WHERE b.city = 'Chandler' AND category = 'Nightlife'
ORDER BY stars DESC
```

```
+----------+-----------+-------+----------------------+--------------+----
----------------+
| city     | category  | stars | hours                | review_count |
address            |
+----------+-----------+-------+----------------------+--------------+----
----------------+
| Chandler | Nightlife |   4.0 | Monday|11:00-0:00    |           75 | 825
N 54th St      |
| Chandler | Nightlife |   4.0 | Tuesday|11:00-0:00   |           75 | 825
N 54th St      |
```

```
| Chandler | Nightlife |   4.0 | Friday|11:00-2:00    |           75 | 825
N 54th St      |
| Chandler | Nightlife |   4.0 | Wednesday|11:00-0:00 |           75 | 825
N 54th St      |
| Chandler | Nightlife |   4.0 | Thursday|11:00-0:00  |           75 | 825
N 54th St      |
| Chandler | Nightlife |   4.0 | Sunday|11:00-0:00    |           75 | 825
N 54th St      |
| Chandler | Nightlife |   4.0 | Saturday|11:00-2:00  |           75 | 825
N 54th St      |
| Chandler | Nightlife |   3.0 | Monday|11:00-0:30    |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Tuesday|11:00-0:30   |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Friday|11:00-2:30    |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Wednesday|11:00-0:30 |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Thursday|11:00-0:30  |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Sunday|9:00-0:30     |          141 | 58
S San Marcos Pl |
| Chandler | Nightlife |   3.0 | Saturday|9:00-2:30   |          141 | 58
S San Marcos Pl |
+----------+-----------+-------+----------------------+--------------+----
----------------+
```

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 average of the reviews counted on closed businesses (23.2) is lower than the average of the reviews counted on open businesses (31.72).

ii. Difference 2:
The average of stars on closed businesses (3.52) is lower than the average of stars on open businesses(3.68).

SQL code used for analysis:

```
SELECT COUNT(id), is_open, ROUND(AVG(stars),2), ROUND(AVG(review_count),2)
FROM business
GROUP BY is_open
```

```
+-----------+---------+---------------------+----------------------------+
| COUNT(id) | is_open | ROUND(AVG(stars),2) | ROUND(AVG(review_count),2) |
+-----------+---------+---------------------+----------------------------+
|      1520 |       0 |                3.52 |                       23.2 |
|      8480 |       1 |                3.68 |                      31.76 |
+-----------+---------+---------------------+----------------------------+
```

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 chose to analyze the nightlife in different cities.

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 need the information of business name, city, stars and review count to study the differences between businesses in the nightlife category according to the city of the business.

Star rating and number of reviews are useful to get insights on how satisfaction is for

clients about the nightlife according to city and business.

iii. Output of your finished dataset:

name	city	category	stars	review_count
Bootleggers Modern American Smokehouse	Phoenix	Nightlife	4.0	431
Irish Republic	Chandler	Nightlife	3.0	141
Eklectic Pie - Mesa	Mesa	Nightlife	4.0	129
Hi Scores - Blue Diamond	Las Vegas	Nightlife	3.5	105
Cabin Club	Westlake	Nightlife	4.0	105
TWIISTED Burgers & Sushi	Medina	Nightlife	4.0	94
Nabers Music, Bar & Eats	Chandler	Nightlife	4.0	75
Gallagher's	Phoenix	Nightlife	3.0	60
The Wine Mill	Peninsula	Nightlife	4.5	42
The Fox & Fiddle	Toronto	Nightlife	2.5	35
The Erin Mills Pump & Patio	Mississauga	Nightlife	3.0	27
Cabin Fever	Toronto	Nightlife	4.5	26
Restaurant Rosalie	Montréal	Nightlife	3.0	19
Halo Brewery	Toronto	Nightlife	4.0	15
Mood	Edinburgh	Nightlife	2.0	11
Innovative Vapors	Tempe	Nightlife	4.5	11
The Charlotte Room	Toronto	Nightlife	3.5	10
Moondogs Pub	Pittsburgh	Nightlife	3.5	

```
|            7 |
| Brubaker's Pub                          | Hudson      | Nightlife |   3.0
|            5 |
| Iron City Grille                        | Coraopolis  | Nightlife |   2.0
|            3 |
+-----------------------------------------+-------------+-----------+------
-+--------------+
```

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

```
SELECT b.name, b.city, c.category, b.stars, b.review_count
FROM business b INNER JOIN category c
ON b.id = c.business_id
WHERE category = 'Nightlife'
ORDER BY review_count DESC
```