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	(id)	
ii.	Hours =	1562	(business_id)	
iii.	Category =	2643	(business_id)	
iv.	Attribute =	1115	(business_id)	
v.	Review =	10000	(id),	8090 (business_id), 9581 (user_id)
vi.	Checkin =	493	(business_id)	
vii.	Photo =	10000	(id),	6493 (business_id)
viii.	Tip =	537	(user_id),	3979 (business_id)
ix.	User =	10000	(id)	
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:

```
select id, name, review_count, yelping_since, useful, funny, cool, fans, average_stars,
                        compliment_hot, compliment_more, compliment_profile, compliment_cute, compliment_list,
                        compliment_note, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos
                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: 1.0 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 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
MontrÃ©al	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

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 as [Star Rating], count(stars) as [Count]
from business b
where city = 'Avon'
group by stars
```

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

Star Rating	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 as [Star Rating], count(stars) as [Count]
from business b
where city = 'Beachwood'
group by stars
```

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

Star Rating	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 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?

- Yes! it seems correlated. Users with more reviews tend to have more fans.

Please explain your findings and interpretation of the results:

-
```
SELECT range AS fans_range,
COUNT(*) AS num_user,
AVG(review_count) AS avg_num_review,
AVG(fans) AS avg_num_fans
FROM (SELECT CASE
```

```
            WHEN fans BETWEEN 0 AND 9 THEN '0 - 9'
            WHEN fans BETWEEN 10 AND 99 THEN '10 - 99'
            ELSE '100-1000' END AS range,
            review_count,
            fans
            FROM user) AS subtable
     GROUP BY subtable.range
```

Result:

fans_range	num_user	avg_num_review	avg_num_fans
0 - 9	9690	15.0085655315	0.447265221878
10 - 99	294	283.326530612	25.5986394558
100-1000	16	891.5	189.75

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

Answer: more reviews with the word "love"

SQL code used to arrive at answer:

```
select (select count(text)
        from review
        where text like "%love%") as  love_text,

       (select count(text)
        from review
        where text like "%hate%") as hate_text
```

Results:

love_text	hate_text
1780	232

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 choose Las Vegas and Shopping category.

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

```
      SELECT CASE WHEN stars > 4.0 THEN '4-5 stars'
            WHEN stars > 3.0 THEN '3-4 stars'
            WHEN stars > 2.0 THEN '2-3 stars'
            ELSE 'below 2' END AS 'STAR',           -- divide the businesses into 4
groups based on their ratings

            COUNT(DISTINCT b.id) AS count,               -- count the distinct
number of businesses from the business inner join of the business and hours table
            COUNT(hours) AS open_days_total,             -- count the number of
entries in the hours table (grouped by the stars category), which happens to be the
total number of days open
            COUNT(hours) / COUNT(DISTINCT b.id)  AS open_days_avg    -- divide the
total number days open by the number of distinct businesses in the hours table
            FROM business b
            INNER JOIN hours h
               ON b.id = h.business_id                   -- creating an inner join
of the business and hours table such that only the business IDs that show up in
both tables are counted
            WHERE city = 'Toronto'                     -- set the city to "Toronto"
            GROUP BY STAR;
```

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

```
SELECT business.city,
       category.category,
       business.name,
       business.stars,
       business.review_count
           ON business.id = hours.business_ID)
       from (business INNER JOIN category ON business.id = category.business_id)
WHERE CITY = 'Las Vegas' AND category.category ='Shopping'
ORDER BY business.stars
```

Result:

city	category	name	stars	review_count
Las Vegas	Shopping	Walgreens	2.5	6
Las Vegas	Shopping	Wooly Wonders	3.5	11
Las Vegas	Shopping	Red Rock Canyon Visitor Center	4.5	32
Las Vegas	Shopping	Desert Medical Equipment	5.0	4

There is different number of review between the two groups; 17 & 36.

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

SQL code used for analysis:

```
       SELECT business.city,
       business.neighborhood,
       category.category,
       business.name,
       business.stars,
       business.review_count,
       hours.hours
       from ((business INNER JOIN category ON business.id = category.business_id)
               INNER JOIN HOURS ON business.id = hours.business_ID)
                 from (business INNER JOIN category ON business.id =
category.business_id)
       WHERE CITY = 'Las Vegas' AND category.category ='Shopping'
       ORDER BY business.stars
```

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: # of business. The number of open-business is bigger than closed

one.

ii. Difference 2: # of review & average of stars. Both of them in the open-business are bigger.

SQL code used for analysis:

```
SELECT is_open,
       count(distinct business.id) num_business,
       count(distinct review.id) num_review,
       avg(review.stars)
FROM business
JOIN review ON business.id =  review.business_id
GROUP BY is_open
```

Result:

is_open	num_business	num_review	avg(review.stars)
0	61	71	3.64788732394
1	446	565	3.7610619469

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:
What is the most successful category of business?

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
Among the categories, I calcalate the average of stars and the proportion of opening on each category. To get statistical reasoning, I only consider the set of category with more than 10 of business.
From the output, we can see that "Local Service", "Health & Medica", "Home Services", "Shopping", and "Beauty & Spas" are successful; they are getting better reviews and higher opening rate. However, "Bars", "Nightlife", and "Restaurants" have lower stars and close frequently.

iii. Output of your finished dataset:

category	num_business	avg_stars	avg_isopen
Local Services	12	4.21	0.83
Health & Medical	17	4.09	0.94
Home Services	16	4.0	0.94
Shopping	30	3.98	0.83
Beauty & Spas	13	3.88	0.92
American (Traditional)	11	3.82	0.73
Food	23	3.78	0.87
Bars	17	3.5	0.65
Nightlife	20	3.48	0.6
Restaurants	71	3.46	0.75

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

```
SELECT category.category,
       count(business.id) num_business,
       round(avg(business.stars),2) avg_stars,
       round(avg(business.is_open),2) avg_isopen
FROM (business INNER JOIN category ON business.id = category.business_id)
GROUP BY category.category
HAVING num_business > 10
ORDER BY avg_stars DESC, avg_isopen DESC
```