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

Sample code (including NULL values):

```
select count(*) as

total_records

from attribute;
```

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

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 537

ix. User = 10000

x. Friend = 11

xi. Elite_years = 2780

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

Sample code:

```
select count(distinct name)  + count(distinct business_id)

+ count(distinct value)

as

total_records

from attribute;
```

```
+---------------+

| total_records |

+---------------+

|          1115 |

+---------------+
```

3. Are there any columns with null values in the Users table? Indicate "yes," or "no."

Answer: Zero rows in the answer shows that there is no null values in the User table

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 = NULL or name = NULL or review_count = NULL or yelping_since = NULL or useful = NULL or
funny = NULL or cool = NULL or fans= NULL or average_stars= NULL or compliment_hot= NULL or
compliment_more= NULL or compliment_profile= NULL or compliment_cute= NULL or compliment_list=
NULL or compliment_note= NULL or compliment_plain = NULL or compliment_cool= NULL or
compliment_funny= NULL or compliment_writer= NULL or compliment_photos= 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

Sample code:

```
select min(stars)
,max(stars)
,avg(stars)
from review
```

5. List the cities with the most reviews in descending order:

SQL code used to arrive at answer:

```
select
city
, count(review_count) as total_review
from business
group by city
order by total_review desc
```

Copy and Paste the Result Below:

```
+-----------------+--------------+
| city            | total_review |
+-----------------+--------------+
| Las Vegas       |         1561 |
| Phoenix         |         1001 |
| Toronto         |          985 |
| Scottsdale      |          497 |
| Charlotte       |          468 |
| Pittsburgh      |          353 |
| Montréal        |          337 |
| Mesa            |          304 |
| Henderson       |          274 |
| Tempe           |          261 |
| Edinburgh       |          239 |
| Chandler        |          232 |
| Cleveland       |          189 |
| Gilbert         |          188 |
| Glendale        |          188 |
| Madison         |          176 |
| Mississauga     |          150 |
| Stuttgart       |          141 |
| Peoria          |          105 |
| Markham         |           80 |
| Champaign       |           71 |
| North Las Vegas |           70 |
| North York      |           64 |
| Surprise        |           60 |
| Richmond Hill   |           54 |
+-----------------+--------------+
(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
name
, stars
, review_count
from business
where city = 'Avon'
```

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

StarRating	Count
0	0
1	0
1.5	1
2	0
2.5	2
3	1
3.5	2
4	2
4.5	1
5	1

ii. Beachwood

SQL code used to arrive at answer:

```
select
name
, stars
, review_count
from business
where city = 'Beachwood';
```

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

```
+--------------------------------+-------+--------------+
| name                           | stars | review_count |
+--------------------------------+-------+--------------+
| Maltz Museum of Jewish Heritage |   3.0 |            8 |
| Charley's Grilled Subs         |   3.0 |            3 |
| Sixth & Pine                   |   4.5 |           14 |
| Beechmont Country Club         |   5.0 |            6 |
| Hyde Park Prime Steakhouse     |   4.0 |           69 |
| Origins                        |   4.5 |            3 |
| Fyodor Bridal Atelier          |   5.0 |            4 |
| College Planning Network       |   2.0 |            8 |
| Lucky Brand Jeans              |   3.5 |            3 |
| American Eagle Outfitters      |   3.5 |            3 |
| Shaker Women's Wellness        |   5.0 |            6 |
| Avis Rent A Car                |   2.5 |            3 |
| Cleveland Acupuncture          |   5.0 |            3 |
| Studio Mz                      |   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
, id
, review_count
from user
order by review_count desc
```

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:

As table below shows, posting more reviews does not necessarily correlate with more fans. For example, although, Gerald has posed the most reviews, he has fewer fans in comparison with Mimi. Therefore, sorting the users in descending order based on their total number of reviews does not sort the fans in the same order, meaning that there is not a correlation between the total number of reviews and number of fans.

```
select
name
, id
, review_count
, fans
from user
order by review_count desc
+-----------+------------------------+--------------+------+
| 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 |
+-----------+------------------------+--------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

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

Answer:

As the tables below show there are more reviews with the word "love" in them compared to the word "hate".

SQL code used to arrive at answer:

```
select
count (*)
from review
where text like '%love%';
```

```
+-----------+
| count (*) |
+-----------+
|      1780 |
+-----------+
```

```
select
count (*)
from review
where text like '%hate%';
```

```
+-----------+
| count (*) |
+-----------+
|       232 |
+-----------+
```

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
```

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

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.

Based on the results, we can see that there seems to be a correlation between the location of the business and its rating. The business that are probably located in the same neighbor have close rating.

Also, they have similar working hours. In addition, the business that have longer working hours usually have higher rating.

SQL code used for analysis:

```
select
business.name
, business.city
, category.category
, business.stars
, hours.hours
, business.review_count
, business.postal_code
from (business inner join category on business.id = category.business_id) inner join hours on hours.business_id = category.business_id
where business.city = 'Mesa'
 group 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:

The business that are still open have higher rating.

ii. Difference 2:

The business that are still open have more reviews.

SQL code used for analysis:

```
select

business.name

, business.is_open

, category.category

, business.stars

, hours.hours

, business.review_count

, business.postal_code

from (business inner join category on business.id = category.business_id) inner join hours on hours.business_id = category.business_id

where business.city = 'Mesa'

 group by business.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:

Finding correlation between the likes with the given rates and using "like" in the reviews.

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 five tables from two different sources. First, I join these two tables together based on business and users. Then I put them in order by grade to see if there is a link between the number of stars and the number of likes.

I chose this analysis and the data sets because psychologists have shown that what people think about something can change a lot, even after just a few minutes. They also believe that what people think right after an event is a better indicator of how good that event was than what they say about it after thinking about it. Since the tip table is about the event itself (shopping) and the review is written hours or even days later, comparing these two tables can help us figure out if what psychologists say is true. There is a weak but noticeable link between the number of likes and stars, as shown by the result. So it looks like what scientists say is mostly true.

iii. Output of your finished dataset:

```
+-------+-------+
| stars | likes |
+-------+-------+
|     3 |     2 |
|     5 |     2 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     5 |     1 |
|     3 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
+-------+-------+
(Output limit exceeded, 25 of 1227 total rows shown)
```

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

```
select
  review.stars
, tip.likes
from review inner join tip on review.user_id = tip.user_id
order by tip.likes desc
```