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 = id - 10000
ii. Hours = business_id - 1562
iii. Category = business_id - 2643
iv. Attribute = business_id - 1115
v. Review = business_id - 8090, user_id - 9581, id - 10000
vi. Checkin = business_id - 493
vii. Photo = id - 10000, business_id - 6493
viii. Tip = business_id - 3979, user_id - 573
ix. User = id - 10000
x. Friend = user_id - 11
xi. Elite_years = user_id - 2780

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 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
R 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) AS Rating FROM business
GROUP BY city
ORDER BY Rating DESC
```

Copy and Paste the Result Below:

```
 +----------------+--------+
| city           | Rating |
+----------------+--------+
| 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 sum(review_count), city, stars
from business
where city = 'Avon'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

```
+------------------+------+-------+
| sum(review_count) | city | stars |
+------------------+------+-------+
|               10 | Avon |   1.5 |
|                6 | Avon |   2.5 |
|               88 | Avon |   3.5 |
|               21 | Avon |   4.0 |
|               31 | Avon |   4.5 |
|                3 | Avon |   5.0 |
+------------------+------+-------+
```

ii. Beachwood

SQL code used to arrive at answer:

```
select sum(review_count), city, stars
from business
where city = 'Beachwood'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

```
+------------------+-----------+-------+
| sum(review_count) | city      | stars |
+------------------+-----------+-------+
|                8 | Beachwood |   2.0 |
|                3 | Beachwood |   2.5 |
```

```
|              11 | Beachwood |   3.0 |
|               6 | Beachwood |   3.5 |
|              69 | Beachwood |   4.0 |
|              17 | Beachwood |   4.5 |
|              23 | Beachwood |   5.0 |
+------------------+-----------+-------+
```

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

SQL code used to arrive at answer:

```
select review_count, name
from user
order BY review_count DESC
limit 3
```

Copy and Paste the Result Below:

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

No it does not because the people with the most reviews do not have the highest number of fans. After reviewing the data from the table, there seems to only be a small correlation between number of review and number of fans. Some people have few fans and a lot of reviews, while other have a lot of fans and few reviews. My guess would be that the about of fans is more related to the type of content the reviewer is posting.

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

Answer:

Love - 1780
Hate - 232

SQL code used to arrive at answer:

```
SELECT COUNT(*)
FROM review
WHERE text LIKE "%hate%"

SELECT COUNT(*)
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, review_count
from user
order BY fans DESC
limit 10
```

Copy and Paste the Result Below:

```
+-----------+------+--------------+
| name      | fans | review_count |
+-----------+------+--------------+
| Amy       |  503 |          609 |
| Mimi      |  497 |          968 |
| Harald    |  311 |         1153 |
| Gerald    |  253 |         2000 |
| Christine |  173 |          930 |
| Lisa      |  159 |          813 |
| Cat       |  133 |          377 |
| William   |  126 |         1215 |
| Fran      |  124 |          862 |
```

```
| Lissa     |  120 |          834 |
+-----------+------+--------------+
```

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?

More of the places with 3.5 stars or lower were open for fewer hours overall. They also closed earlier, such as 3 or 4.

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

While places with 2-3 stars had a higher number of reviews overall, there were more places that had 4-5 stars.

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

As someone who is familiar with this location, the restaurants with higher stars were in the areas that are more affluent and generally have higher income levels and housing expenses.

SQL code used for analysis:

```
select
business.city, business.neighborhood, business.name
,business.stars, business.review_count
,hours.hours
,category.category
from (business inner join hours on business.id =
hours.business_id) inner join category on
category.business_id = category.business_id
where business.city = 'Charlotte' and category = 'Coffee &
```

```
    Tea'
    group by neighborhood
```

```
+----------+----------------+------------------------------
+-------+-------------+---------------------+-------------+
| city      | neighborhood    | name                          |
stars | review_count | hours               | category     |
+----------+----------------+------------------------------
+-------+-------------+---------------------+-------------+
| Charlotte |                 | Big City Grill                |
5.0 |            4 | Saturday|11:00-20:00 | Coffee & Tea |
| Charlotte | Arboretum       | Journey's Dry Carpet Cleaning |
5.0 |            3 | Saturday|8:00-20:00  | Coffee & Tea |
| Charlotte | Myers Park      | Gorgeous Glo                  |
3.5 |           10 | Saturday|11:00-16:00 | Coffee & Tea |
| Charlotte | South End       | Dilworth Custom Framing       |
3.5 |            6 | Saturday|10:00-15:00 | Coffee & Tea |
| Charlotte | South Park      | Camden Fairview               |
5.0 |            6 | Saturday|10:00-17:00 | Coffee & Tea|
| Charlotte | University City | Highlife North Tryon          |
4.0 |            5 | Saturday|12:00-22:00 | Coffee & Tea |
```

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:

Out of the two Coffee & Tea places that were closed, both were rated 5 stars.

ii. Difference 2:

The two places with the lease number of views were also the two that were closed.

SQL code used for analysis:

```
    select
```

```
      business.city, business.neighborhood, business.name
      ,business.stars, business.review_count, business.is_open
      ,hours.hours
      ,category.category
      from (business inner join hours on business.id =
hours.business_id) inner join category on category.business_id =
                          category.business_id
      where business.city = 'Charlotte' and category = 'Food'
      group by name
```

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:

For this question, I chose to determine if users that had been using yelp the longest had more fans and higher star averages.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

In order to predict this question, I needed to have the users names, their number of fans, the average number of stars they had, and the year they started yelping in. I chose this data as I thought it would be the most valuable in helping to determine if ones length of time on yelp meant they would have higher star averages or higher numbers of fans. Though my query, I discovered that there was no direct correlation between the number of fans or average number of stars one had. The user who had been on yelp the longest, had one of the least amounts of fans, although he did have the second highest star average. The user who had been on yelp for the second shortest about of time had one of the highest star averages, but she also had one of the smallest number of reviews. The user that

had been on yelp for the second longest time, had the highest number of fans, but I am thinking that they category of their reviews may be more correlated with the number of fans than the length of time on the website.

iii. Output of your finished dataset:

```
+---------+------+--------------+--------------------
+-----------------------+-----------------------+
| name    | fans | average_stars | yelping_since       | id
| user_id               |
+---------+------+--------------+--------------------
+-----------------------+-----------------------+
| Matt    |   14 |          3.77 | 2006-10-11 00:00:00 |
-aAgfEUH4UoFDRXZCfJSUA | -aAgfEUH4UoFDRXZCfJSUA |
| Lissa   |  120 |          3.68 | 2007-08-14 00:00:00 |
-lh59ko3dxChBSZ9U7LfUw | -lh59ko3dxChBSZ9U7LfUw |
| Ed      |   38 |           3.6 | 2009-08-10 00:00:00 |
-fUARDNuXAfrOn4WLSZLgA | -fUARDNuXAfrOn4WLSZLgA |
| Elaine  |   18 |          3.26 | 2010-04-21 00:00:00 |
-a0LRFr94D9ohyBJCKVvXQ | -a0LRFr94D9ohyBJCKVvXQ |
| Dixie   |   41 |          3.19 | 2011-01-19 00:00:00 | --
Qh8yKWAvIP4V4K8ZPfHA | --Qh8yKWAvIP4V4K8ZPfHA |
| Dominic |   37 |          3.47 | 2011-02-06 00:00:00 |
-kO6984fXByyZm3_6z2JYg | -kO6984fXByyZm3_6z2JYg |
| Justin  |   13 |          3.51 | 2012-10-07 00:00:00 |
-lh59ko3dxChBSZ9U7LfUw | -lh59ko3dxChBSZ9U7LfUw |
| Nieves  |   80 |          3.64 | 2013-07-08 00:00:00 |
-fUARDNuXAfrOn4WLSZLgA | -fUARDNuXAfrOn4WLSZLgA |
| Lalena  |   25 |          3.94 | 2014-02-20 00:00:00 |
-a0LRFr94D9ohyBJCKVvXQ | -a0LRFr94D9ohyBJCKVvXQ |
| Kristen |   15 |          3.32 | 2015-12-23 00:00:00 | --
Qh8yKWAvIP4V4K8ZPfHA | --Qh8yKWAvIP4V4K8ZPfHA |
```

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

```
     Select u.name, u.fans, u.average_stars, u.yelping_since,
u.id
     ,ey.user_id, ey.year
```

```
From user u INNER JOIN elite_years ey ON
u.id = ey.user_id
where fans > 10
group by name
order by yelping_since asc
```