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 distinct records by primary key ‘id’ of Business table
ii. Hours = 1562 distinct records by foreign key ‘business_id’
iii. Category = 2643 distinct records by foreign key ‘business_id’
iv. Attribute = 1115 distinct records by foreign key ‘id’
v. Review = 10000 distinct records by primary key ‘id’, 9581 user_id foreign key, 8090 business_id foreign key
vi. Checkin = 493 distinct records by foreign key ‘business_id’
vii. Photo = 10000 distinct records for primary key ‘id’, 6493 business_id foreign key
viii. Tip = 537 distinct records for foreign key ‘user_id’, 3979 business_id foreign key
ix. User = 10000 distinct records by primary key ‘id’
x. Friend = 11 distinct records by foreign key ‘user_id’
xi. Elite_years = 2780 distinct records by foreign key ‘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 *
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:

```
Select city,
Sum(review_count) As reviews
From business
Group By city
Group By review DESC
```

Copy and Paste the Result Below:

```
+-----------------+---------+
| city            | reviews |
+-----------------+---------+
| 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) AS count
   From business
   Where city = ‘Avon’
   Group By stars
```

Copy and Paste the Resulting Table Below (2 columns â€

“ star rating and count):

```
+-------+-----+
| stars |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, count(stars) AS count
From business
Where city = ‘Beachwood’
Group By stars
```

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

```
+-------+-----+
| stars |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 id, name, review_count
     From user
     Order By review_count DESC
     Limit 3
```

     Copy and Paste the Result Below:

```
+------------------------+--------+--------------+
| id                     | name   | review_count |
+------------------------+--------+--------------+
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald |         2000 |
| -3s52C4zL_DHRK0ULG6qtg | Sara   |         1629 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri   |         1339 |
+------------------------+--------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:
As seen in the table below, having a high review_count doesn’t have a correlation to more fans. This can be seen by the difference in Sara and William. Sara has a greater number of reviews but she has less than half the number of fans William has.

```
+-----------+--------------+------+---------------------
+------+
| name      | review_count | fans | yelping_since       |
text |
+-----------+--------------+------+---------------------
+------+
| Gerald    |         2000 |  253 | 2012-12-16 00:00:00 |
None |
| Sara      |         1629 |   50 | 2010-05-16 00:00:00 |
None |
| Yuri      |         1339 |   76 | 2008-01-03 00:00:00 |
None |
| .Hon      |         1246 |  101 | 2006-07-19 00:00:00 |
None |
| William   |         1215 |  126 | 2015-02-19 00:00:00 |
None |
| Harald    |          1153 |  311 | 2012-11-27 00:00:00 |
None |
| eric      |          1116 |   16 | 2007-05-27 00:00:00 |
None |
| Roanna    |          1039 |  104 | 2006-03-28 00:00:00 |
None |
| Mimi      |           968 |  497 | 2011-03-30 00:0:00 |
None |
| Christine |           930 |  173 | 2009-07-08 00:00:00 |
None |
+-----------+-------------+------+---------------------
+------+
```

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

```
    Answer:  -Love is contained in 1780 reviews.
            -Hate is contained in only 232 reviews.
```

SQL code used to arrive at answer:

```
    Select Count(text)
    From review
    Where Lower(text) like ‘%love%’
    Select Count(text)
    From review
    Where Lower(text) like ‘%hate%’
```

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

SQL code used to arrive at answer:

```
    Select user.id, name, fans
    From user
    Order by fans DESC
    Limit 10
```

Copy and Paste the Result Below:

```
+------------------------+-----------+------+
| id                     | name      | fans |
+------------------------+-----------+------+
| -9I98YbNQnLdAmcYfb324Q | Amy       |  503 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi      |  497 |
| --2vR0DIsmQ6WfcSzKWigw | Harald    |  311 |
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald    |  253 |
| -0IiMAZI2SsQ7VmyzJjokQ | Christine |  173 |
| -g3XIcCb2b-BD0QBCcq2Sw | Lisa      |  159 |
| -9bbDysuiWeo2VShFJJtcw | Cat       |  133 |
| -FZBTkAZEXoP7CYvRV2ZwQ | William   |  126 |
| -9da1xk7zgnnfO1uTVYGkA | Fran      |  124 |
| -lh59ko3dxChBSZ9U7LfUw | 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 have chosen the city of Mesa. It had 129 reviews. Subsequently, I chose Shopping as there were 4 businesses which was the highest among all the categories along with Health & Medical and Restaurants.

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

For shopping category, only Walgreens qualified for the 2-3 group and only Desert Medical Equipment and Red Rock Canyon Visitor Centre qualified in the 4-5 group. In comparison, Walgreens is open everyday from 8am to 10pm whereas the ones in 4-5 rating are open from 8 am to 4:30pm and 5 pm respectively. Hence, there is a drastic difference in the distribution of hours.

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

Walgreens has 6 reviews whereas the 4-5-star businesses have 4 and 32 reviews respectively.

SQL code used for analysis:

```
     Select
     Case
     When stars>=4 Then ‘4-5 stars’
     When (stars>=2 And stars<=3) Then ‘2-3 stars’
     End as rating,

     postal_code,
     review_count,
     hours.hours,
     name,
     neighborhood

     From business INNER JOIN category
     On business.id=category.business_id INNER JOIN hours
     On business.id=hours.business_id

     Where city=‘Mesa’
    And category = 'Shopping'
    And (stars>=4 OR (stars <3 and stars>2))

    Order by stars DESC, hours DESC
```

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:
Open: total reviews = 269300
Closed: AVG(review_count) = 35261

ii. Difference 2:
Open: AVG(stars) = 3.68
Closed: AVG(stars) = 3.52

SQL code used for analysis:

```
     Select Count(DISTINCT(id)) As Number_of_business,
     Round(AVG(review_count),2) As avg_review,
     Sum(review_count) As total_review,
     is_open
     From business
     Group by is_open
```

```
+--------------------+------------+--------------
+---------+
| Number_of_business | avg_review | total_review |
is_open |
+--------------------+------------+--------------
+---------+
|               1520 |       23.2 |        35261 |
0 |
|               8480 |      31.76 |       269300 |
1 |
+--------------------+------------+--------------
+---------+
```

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 think observing the average star rating of Restaurants according to WiFi availability would be interesting.

After studying the ER diagram, there are several factors that can be taken into consideration. Namely, WiFi availability, restaurant's average rating, number of businesses on each category and average reviews.

iii. Output of your finished dataset:

```
+------+-------+----------------+---------------
+---------------+
| name | value | total_business |     avg_stars |
avg_reviews |
+------+-------+----------------+---------------
+---------------+
| WiFi | no    |             13 | 3.65384615385 |
86.9230769231 |
| WiFi | free  |              9 | 2.94444444444 |
34.7777777778 |
+------+-------+----------------+---------------
+---------------+
```

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

```
Select DISTINCT att.name
         ,att.value
         ,COUNT(att.business_id) total_business
         ,AVG(bu.stars) avg_stars
         ,AVG(bu.review_count) avg_reviews
```

```
From attribute att
LEFT JOIN business bu ON att.business_id = bu.id
LEFT JOIN category c ON c.business_id = bu.id
Where (att.name = 'WiFi')
     And (c.category IS NOT NULL)
     And (c.category = 'Restaurants')
Group By att.value
Order By total_business DESC
```