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

ANS)

```
SELECT *
FROM table
```

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. Checkin = 493
vi. Photo = 6493

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

ANS)
i.

```
SELECT COUNT(distinct id)
FROM business
```

ii.

```
SELECT COUNT(distinct business_id)

FROM hours
```

iii.

```
SELECT COUNT(distinct business_id)

FROM category
```

iv.

```
SELECT COUNT(distinct business_id)
FROM attribute
```

v.

```
SELECT COUNT(distinct business_id)
FROM Checkin
```

vi.

```
SELECT COUNT(distinct business_id)
FROM Photo
```

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

Answer:

NO

```
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: 3.65

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

ANS)

i.

```
SELECT MIN(stars),MAX(stars),AVG(stars)
FROM review
```

ii.

```
SELECT MIN(stars),MAX(stars),AVG(stars)
FROM  business
```

iii.

```
SELECT MIN(Likes),MAX(Likes),AVG(Likes)
FROM  tip
```

iv.

```
SELECT MIN(Count),MAX(Count),AVG(Count)

FROM  Checkin
```

v.

```
SELECT MIN(Review_count),MAX(Review_count),AVG(Review_count)

FROM  user
```

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

```
SELECT city, Count(review_count) AS Total_review_count

FROM  business
group by city
order by Total_review_count desc
```

```
+-----------------+--------------------+
| city            | Total_review_count |
+-----------------+--------------------+
| 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 -   10

```
SELECT city, Count(stars) as star_rating
FROM  business
where City ="Avon"
```

```
+------+-------------+
| city | star_rating |
+------+-------------+
| Avon |          10 |
+------+-------------+
```

ii. Beachwood-   14

```
SELECT city, Count(stars) as star_rating
FROM  business
where City ="Beachwood"
```

```
+-----------+-------------+
| city      | star_rating |
+-----------+-------------+
| Beachwood |          14 |
+-----------+-------------+
```

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

```
      SELECT name, review_count

      FROM user
      order by review_count desc
      limit 3
```

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

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

8. Does posing more reviews correlate with more fans?

```
SELECT name, review_count, fans

FROM user
order by review_count desc
```

```
+-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Gerald    |         2000 |  253 |
| Sara      |         1629 |   50 |
| Yuri      |         1339 |   76 |
| .Hon      |         1246 |  101 |
| William   |         1215 |  126 |
| Harald    |         1153 |  311 |
| eric      |         1116 |   16 |
| Roanna    |         1039 |  104 |
| Mimi      |          968 |  497 |
| Christine |          930 |  173 |
| Ed        |          904 |   38 |
| Nicole    |          864 |   43 |
| Fran      |          862 |  124 |
| Mark      |          861 |  115 |
| Christina |          842 |   85 |
| Dominic   |          836 |   37 |
| Lissa     |          834 |  120 |
| Lisa      |          813 |  159 |
| Alison    |          775 |   61 |
| Sui       |          754 |   78 |
```

We can see from the above data that it is not necessary that possessing more reviews correlated with more fans

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

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:

```
SELECT name, fans

FROM user
order by fans desc
limit 10
```

```
+-----------+------+
| 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

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 result, we can see that there is a correlation between the location of the business and its rating. For example, we can infer that the location having postal code 85210 has a 5-star rating in most of the categories whereas the location having postal code 85206 has a 2star rating in most of the categories. Hence we can infer that the business that is located in the same neighbourhood has closer ratings. Also, they have similar working hours. Moreover, the business that has longer working hours usually have a higher rating

```
SELECT business.city, business.stars, category.category,business.review_count,
hours.hours
FROM business
JOIN category
on category.business_id = business.id
JOIN hours
ON hours.business_id= business.id
WHERE business.city = "Mesa"
GROUP BY stars
ORDER BY stars 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:

The number of reviews for the business that is open is comparatively high as compared to the business that are closed

ii. Difference 2:

The total number of business stars for the business that are closed is 2126 whereas the total number of business stars for the business that are open is 9921

```
SELECT business.city, SUM(business.stars), category.category,business.review_count,
hours.hours, business.postal_code,business.is_open
FROM business
JOIN category
on category.business_id = business.id
JOIN hours
ON hours.business_id= business.id
GROUP BY 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:

I choose to analyze the business based on the various attributes such as good for the kid, free WIFI, and various other factors to indicate the number of stars and rating

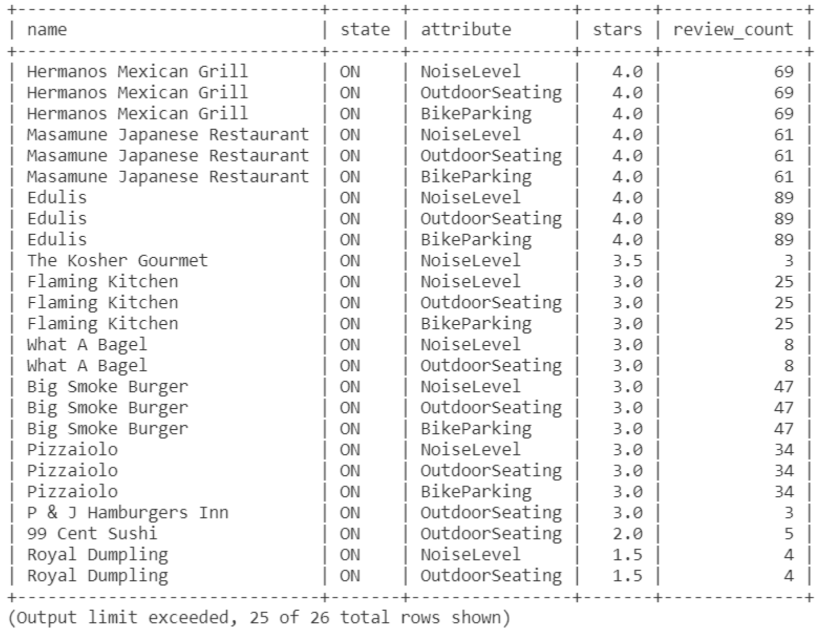
ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

I used 3 different tables namely business, attributes, and category. To connect all the tables I made use of the JOIN function and used the primary key like id from the business table, business_id from category, and business_id from attributes.

Then I selected the following column from the tables namely business.name, business.state, attribute.name, category.category, business.stars, business.review_count

I wanted to know if the following attributes are present in the restaurant namely “Noise Level”, “bike parking” and “ Outdoor seating ” and what are the rating and stars given to the Restaurant according to the attributes.

iii. Output of your finished dataset:



```
+-----------------------------+-------+----------------+-------+--------------+
| name                        | state | attribute      | stars | review_count |
+-----------------------------+-------+----------------+-------+--------------+
| Hermanos Mexican Grill      | ON    | NoiseLevel     |   4.0 |           69 |
| Hermanos Mexican Grill      | ON    | OutdoorSeating |   4.0 |           69 |
| Hermanos Mexican Grill      | ON    | BikeParking    |   4.0 |           69 |
| Masamune Japanese Restaurant | ON   | NoiseLevel     |   4.0 |           61 |
| Masamune Japanese Restaurant | ON   | OutdoorSeating |   4.0 |           61 |
| Masamune Japanese Restaurant | ON   | BikeParking    |   4.0 |           61 |
| Edulis                      | ON    | NoiseLevel     |   4.0 |           89 |
| Edulis                      | ON    | OutdoorSeating |   4.0 |           89 |
| Edulis                      | ON    | BikeParking    |   4.0 |           89 |
| The Kosher Gourmet          | ON    | NoiseLevel     |   3.5 |            3 |
| Flaming Kitchen             | ON    | NoiseLevel     |   3.0 |           25 |
| Flaming Kitchen             | ON    | OutdoorSeating |   3.0 |           25 |
| Flaming Kitchen             | ON    | BikeParking    |   3.0 |           25 |
| What A Bagel                | ON    | NoiseLevel     |   3.0 |            8 |
| What A Bagel                | ON    | OutdoorSeating |   3.0 |            8 |
| Big Smoke Burger            | ON    | NoiseLevel     |   3.0 |           47 |
| Big Smoke Burger            | ON    | OutdoorSeating |   3.0 |           47 |
| Big Smoke Burger            | ON    | BikeParking    |   3.0 |           47 |
| Pizzaiolo                   | ON    | NoiseLevel     |   3.0 |           34 |
| Pizzaiolo                   | ON    | OutdoorSeating |   3.0 |           34 |
| Pizzaiolo                   | ON    | BikeParking    |   3.0 |           34 |
| P & J Hamburgers Inn        | ON    | OutdoorSeating |   3.0 |            3 |
| 99 Cent Sushi               | ON    | OutdoorSeating |   2.0 |            5 |
| Royal Dumpling              | ON    | NoiseLevel     |   1.5 |            4 |
| Royal Dumpling              | ON    | OutdoorSeating |   1.5 |            4 |
+-----------------------------+-------+----------------+-------+--------------+
(Output limit exceeded, 25 of 26 total rows shown)
```

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

```
SELECT business.name,business.state, attribute.name as attribute,business.stars,business.review_count
FROM business
INNER JOIN attribute
ON business.id = attribute.business_id
INNER JOIN category
ON attribute.business_id= category.business_id
where ( attribute.name like  'NoiseLevel' or
        attribute.name like 'OutdoorSeating' or
        attribute.name like 'BikeParking') and
        category = 'Restaurants' and business.state="ON"
order by stars desc
```