Data Scientist Role Play: Profiling and Analysing the Yelp Dataset Coursera Worksheet

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. Check-in 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(Primary Key-id)

ii. Hours = 1562(Foreign Key-Business_id)

iii. Category = 2643(Foreign Key-Business_id)

iv. Attribute = 1115(Foreign Key-Business_id)

v. Review = 8090(Foreign Key-Business_id), 9581(Foreign Key-Business_id),

10000(Primary Key-id)

vi. Checkin = 493(Foreign Key-Business_id)

vii. Photo =6493(Foreign Key-Business_id),

10000(Primary Key-id)

viii. Tip = 3979(Foreign Key-Business_id),

537(Foreign Key-Business_id),

ix. User = 10000(Primary Key-id)

x. Friend = 11(Foreign Key-User_id)

xi. Elite_years =2780 (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 nu
ll or compliment_note is null or compliment_plain is null or compliment_cool i
s 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	MAX	AVG
1	5	3.7082

ii. Table: Business, Column: Stars

MIN	MAX	AVG
1.0	5.0	3.6549

iii. Table: Tip, Column: Likes

MIN	MAX	AVG
0	2	0.0144

iv. Table: Checkin, Column: Count

MIN	MAX	AVG
1	53	1.9414

v. Table: User, Column: Review_count

MIN	MAX	AVG
0	2000	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 RC
      from
            business
      group by
            city
      order by
            RC desc;
```

Copy and Paste the Result Below:

```
+------------------+-------+
| city             |    RC |
+------------------+-------+
| 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(*) as count_of_businesses
FROM business
WHERE city = "Avon"
GROUP BY stars
ORDER BY stars;
```

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

```
+-------+---------------------+
| stars | count_of_businesses |
+-------+---------------------+
|   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(*) as count_of_businesses
FROM business
WHERE city = "Beachwood"
GROUP BY stars
ORDER BY stars;
```

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

```
+-------+---------------------+
| stars | count_of_businesses |
+-------+---------------------+
|   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?

Please explain your findings and interpretation of the results:

- The quality of reviews matters; fans are drawn to valuable and insightful content.
- Consistent reviews within a specific niche engage audiences and cultivate loyalty.
- Interaction, effective promotion, and adapting to feedback amplify fanbase expansion.

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

Answer: More reviews with love

SQL code used to arrive at answer:

```
 Select COUNT(text) from review
  where text like "love%";


Select COUNT(text) from review
  where text like "hate%";
```

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. 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.

The differences between the two groups of restaurants (2-3 star rated vs. 4-5 star rated) based solely on their location. It got more good ratings in famous and tourist areas compared to un popular areas.

SQL code used for analysis:

```
SELECT
    b.name,
    b.city,
    b.stars,
    h.hours,
    c.category
FROM
    business b
JOIN
    hours h ON b.id = h.business_id
JOIN
    category c ON c.business_id = h.business_id
WHERE
    b.city = "Toronto"
    AND c.category = "Restaurants"
    AND (
        (b.stars BETWEEN 2 AND 3)
        OR (b.stars BETWEEN 4 AND 5)
    );
```

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 businesses tend to have a higher average review count than closed businesses. This could be due to their popularity and larger customer base, potentially contributing to their sustained operation.

ii. Difference 2:

Open businesses often exhibit slightly higher average ratings than closed businesses, possibly attributed to better service quality, customer satisfaction, and overall value. Elevated ratings could potentially enhance customer attraction and contribute to the business's long-term viability.

SQL code used for analysis:

```
SELECT COUNT(DISTINCT(id)),

         AVG(review_count),
         SUM(review_count),
         AVG(stars),
         is_open
  FROM business
  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:

Sentiment Analysis:

My chosen analysis involves sentiment analysis of Yelp reviews. This approach aims to extract insights from customer reviews by discerning sentiment—whether positive, negative, or neutral. By parsing review text and attributes, this analysis sheds light on the overall sentiment of various businesses' customer feedback.

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

In our sentiment analysis, we're leveraging the comprehensive Yelp dataset, which includes crucial elements like reviews, business attributes, and star ratings. By focusing on the review text, we're deciphering the sentiment behind customer feedback, helping businesses decode the underlying opinions. Additionally, we're tapping into business categories to provide a contextual backdrop, allowing us to understand sentiment in relation to the type of business. This category-based analysis empowers businesses to grasp sentiment nuances unique to their industry.

Star ratings, acting as a benchmark, enhance our analysis by offering a comparative reference. This holistic approach aids businesses in extracting actionable insights from customer feedback. It not only assists in pinpointing areas for improvement but also aids in evaluating overall customer satisfaction. By aligning sentiment analysis with business categories and star ratings, this methodology provides a comprehensive tool for businesses to make informed decisions and elevate their customer experience.

iii. Output of your finished dataset:

The completed dataset will comprise essential columns: business_id, business_name, business_category, review_id, review_text, star_rating, sentiment_score, and sentiment_label. In this configuration, sentiment_score, a numerical representation, gauges the sentiment polarity of each review text. Meanwhile, sentiment_label allocates sentiments into categories: positive, negative, or neutral. This enriched dataset effectively captures review sentiments across different businesses.

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

```sql
select b.id AS ID,b.name AS Name,c.category AS Category,b.stars AS Star_Ratings,r.text AS Reviews
CASE
        WHEN sentiment_score > 0 THEN 'Positive'
        WHEN sentiment_score < 0 THEN 'Negative'
        ELSE 'Neutral'
    END AS sentiment_label,
    sentiment_score

from business b
    JOIN review r ON r.business_id=b.id
        JOIN category c ON c.business_id=b.id
```