Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset

Xiuyu Chen

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
```

**********SQL CODE**********

```
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(id)
ii. Hours       = 1562(business_id)
iii. Category   = 2643(business id)
iv. Attribute   = 1115(business_id)
v. Review       = 10000(id)
vi. Checkin     = 493(business_id)
vii. Photo      = 10000(id)
viii. Tip       = 537(user_id) or 3979(business_id)
ix. User        = 10000(id)
x. Friend       = 11(user_id)
xi. Elite_years = 2780(user_id)
```

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

**********SQL CODE**********

```
SELECT DISTINCT key
FROM table;
```

****

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: 1.0 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

**********SQL CODE**********

```
SELECT MIN(column), MAX(column), AVG(column)
FROM table;
```

****

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

SQL code used to arrive at answer:

```
SELECT city, SUM(review_count) AS NReviews
FROM business
GROUP BY city
ORDER BY NReviews DESC;
```

Copy and Paste the Result Below:

city	NReviews
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

```
SELECT stars AS [star rating]
, COUNT(stars) AS count
FROM business
WHERE city = 'Avon'
GROUP BY stars;
```

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

```
+-------------+-------+
| star rating | 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 AS [star rating]
, COUNT(stars) AS count
FROM business
WHERE city = 'Beachwood'
GROUP BY stars;
```

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

```
+-------------+-------+
| star rating | 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 name, review_count
FROM user
ORDER BY review_count
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:

Yes. Posing more reviews seem to correlate with more fans. Most of people appear in the table below have written lots of reviews on Yelp.

SQL CODE:

```
SELECT name, review_count, fans
FROM user
ORDER BY fans DESC;
```

Result:

```
+-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Amy       |          609 |  503 |
| Mimi      |          968 |  497 |
| Harald    |         1153 |  311 |
| Gerald    |         2000 |  253 |
| Christine |          930 |  173 |
| Lisa      |          813 |  159 |
| Cat       |          377 |  133 |
| William   |         1215 |  126 |
| Fran      |          862 |  124 |
| Lissa     |          834 |  120 |
| Mark      |          861 |  115 |
| Tiffany   |          408 |  111 |
| bernice   |          255 |  105 |
| Roanna    |         1039 |  104 |
| Angela    |          694 |  101 |
```

```
| .Hon      |          1246 |  101 |
| Ben       |           307 |   96 |
| Linda     |           584 |   89 |
| Christina |           842 |   85 |
| Jessica   |           220 |   84 |
| Greg      |           408 |   81 |
| Nieves    |           178 |   80 |
| Sui       |           754 |   78 |
| Yuri      |          1339 |   76 |
| Nicole    |           161 |   73 |
+-----------+---------------+------+
(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: Yes

SQL code used to arrive at answer:

```
SELECT(
    SELECT COUNT(*)
    FROM review
    WHERE text LIKE '%love%') AS love
    ,(SELECT COUNT(*)
    FROM review
    WHERE text LIKE '%hate%') AS hate;
```

Result:

```
+------+------+
| love | hate |
+------+------+
| 1780 |  232 |
+------+------+
```

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.

The city I pick is Toronto, and the category I pick is restaurants.

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

This is no obvious difference between restaurants with higher ratings and those of lower ratings in terms of hours. 99 Cent Sushi and Sushi Osaka both opens st 11:00 and closes at 23:00 on Saturday. However, the former has the lowest star rating of 2.0, while the latter enjoys the highest rating of 4.5.

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

No. Number of reviews is not necessarily correlated to star rating.

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

The location data cannot provide us with information about restaurants . Restaurants with 2-3 stars and Edulis (with 4 stars) are located in the same area.Reviews they received and opening hours also do not seem to be related to locations.

SQL code used for analysis:

```
SELECT b.name, b.city, c.category, b.stars, h.hours,
b.review_count, b.address, b.postal_code
FROM (business b INNER JOIN category c ON b.id =
c.business_id) INNER JOIN hours h ON b.id = h.business_id
WHERE b.city = 'Toronto' AND c.category = 'Restaurants'
GROUP BY b.stars;
```

Result:

```
+---------------+---------+-------------+-------
+----------------------+-------------
+-----------------------+-------------+
| name          | city    | category    | stars |
hours                 | review_count | address
| postal_code |
+---------------+---------+-------------+-------
+----------------------+-------------
+-----------------------+-------------+
| 99 Cent Sushi | Toronto | Restaurants |   2.0 |
Saturday|11:00-23:00 |            5 | 389 Church
Street     | M5B 2E5     |
| Pizzaiolo     | Toronto | Restaurants |   3.0 |
Saturday|10:00-4:00  |           34 | 270 Adelaide
Street W | M5H 1X6     |
| Edulis        | Toronto | Restaurants |   4.0 |
Saturday|18:00-23:00 |           89 | 169 Niagara
Street    | M5V         |
| Sushi Osaka   | Toronto | Restaurants |   4.5 |
Saturday|11:00-23:00 |            8 | 5084 Dundas
Street W  | M9A 1C2     |
+---------------+---------+-------------+-------
+----------------------+-------------
```

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

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 running have more reviews on average than those that are currently closed.

ii. Difference 2:
    Those who are open have higher ratings in overall compared to those that are closed.

SQL code used for analysis:

```
SELECT b.is_open, COUNT(DISTINCT b.id), AVG(b.stars),
AVG(b.review_count), COUNT(r.useful), COUNT(r.funny),
COUNT(r.cool)
FROM business b INNER JOIN review r ON b.id = r.id
GROUP BY b.is_open;
```

Result:

```
+---------+----------------------+---------------
+---------------------+-----------------
+----------------+---------------+
| is_open | COUNT(DISTINCT b.id) |  AVG(b.stars) |
AVG(b.review_count) | COUNT(r.useful) |
COUNT(r.funny) | COUNT(r.cool) |
+---------+----------------------+--------------
+---------------------+-----------------
+----------------+---------------+
|       0 |                    1 |           2.0 |
4.0 |                1 |               1 |
1 |
|       1 |                   13 | 2.96153846154 |
38.7692307692 |               13 |              13 |
13 |
```

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

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 want to find business attributes that matter to restaurants.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
Restaurants are places that serve customers with food, but food is not the only factor that attracts consumers. Since customers rate restaurants on Yelp to review services they receive, we can analyse the relationship between stars and business attributes at the restaurant level to get some insights into this question.

iii. Output of your finished dataset:

service	AVG(b.stars)
BestNights	3.66666666667
CoatCheck	3.66666666667
DogsAllowed	3.66666666667
GoodForDancing	3.66666666667
HappyHour	3.66666666667
Smoking	3.66666666667
WiFi	3.36363636364
Caters	3.34210526316
Ambience	3.3
BusinessParking	3.27777777778

```
| Alcohol                    | 3.26923076923 |
| WheelchairAccessible       | 3.26923076923 |
| RestaurantsTableService    |          3.26 |
| HasTV                      | 3.25925925926 |
| Music                      |          3.25 |
| NoiseLevel                 | 3.24074074074 |
| BikeParking                |         3.225 |
| DriveThru                  | 3.21428571429 |
| GoodForKids                | 3.21428571429 |
| RestaurantsReservations    | 3.20689655172 |
| GoodForMeal                |  3.2037037037 |
| RestaurantsAttire          | 3.18965517241 |
| RestaurantsPriceRange2     | 3.18333333333 |
| BusinessAcceptsCreditCards | 3.16666666667 |
| OutdoorSeating             | 3.15517241379 |
+----------------------------+---------------+
(Output limit exceeded, 25 of 28 total rows shown)
```

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

```
SELECT DISTINCT a.name as quality, COUNT(a.name) AS frequency,
AVG(value)
FROM attribute a INNER JOIN business b ON a.business_id = b.id
GROUP BY quality
ORDER BY frequency DESC, AVG(value) DESC;
```