BY: Anshal Singh

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 =10,000
ii. Business table =10,000
iii. Category table =10,000
iv. Checkin table =10,000
v. elite_years table =10,000
vi. friend table = 10,000
vii. hours table =10,000
viii. photo table = 10,000
ix. review table = 10,000
x. tip table = 10,000
xi. user table =10,000

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 =10,000
ii. Hours =1562
iii. Category =2643
iv. Attribute =1115
v. Review =8090(business_id)
vi. Checkin =493
vii. Photo =6493
viii. Tip = 3979(business_id)
ix. User = 10,000
x. Friend = 11
xi. Elite_years =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 COUNT(*)
        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

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

SQL code used to arrive at answer:

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

Copy and Paste the Result Below:

```
+-----------------+--------------+
| city            | totalReviews |
+-----------------+--------------+
| 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 DISTINCT(stars),
```

```
count(stars)
FROM business
WHERE city='Avon'
GROUP BY stars
```

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

| stars | count(stars) |
|---|---|
| 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 DISTINCT(stars),
count(stars)
FROM business
WHERE city='Beachwood'
GROUP BY stars
```

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

| stars | count(stars) |
|---|---|
| 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 posting more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

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

```
+--------------+------+
| review_count | fans |
+--------------+------+
|         2000 |  253 |
|         1629 |   50 |
|         1339 |   76 |
|         1246 |  101 |
|         1215 |  126 |
|         1153 |  311 |
|         1116 |   16 |
|         1039 |  104 |
```

```
|           968 |  497 |
|           930 |  173 |
|           904 |   38 |
|           864 |   43 |
|           862 |  124 |
|           861 |  115 |
|           842 |   85 |
|           836 |   37 |
|           834 |  120 |
|           813 |  159 |
|           775 |   61 |
|           754 |   78 |
|           702 |   35 |
|           696 |   10 |
|           694 |  101 |
|           676 |   25 |
|           675 |   45 |
+-------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

Not really as fans are not consistently distributed opposite to number of reviews.

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

Answer:

The Number of reviews with word love(1780) are much more than hate(232).

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
City=Phoenix , Category=Restaurants

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?
ANS: Not really the hours don't actually determine the rating here.

ii. Do the two groups you chose to analyze have a different number of reviews?
ANS: Avg. review count for 1-4: 125.6
Avg review count for 4-5: 219

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
ANS : No, All of them seem to have different postal code and different rating and working hours.

SQL code used for analysis:

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

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 have higher working hours.

ii. Difference 2:
The no of reviews are higher for open businesses.

SQL code used for analysis:

```
SELECT b.is_open,
b.name,
b.city,
c.category,
b.stars,
h.hours,
b.postal_code,
b.review_count
FROM (business b INNER JOIN
category c ON b.id=c.business_id)
```

```
INNER JOIN hours h ON h.business_id=c.business_id
GROUP BY b.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 want to predict whether the business is likely to be open or closed.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
To do so I need various attributes of the data like name, location, stars, state, latitude, longitude , review_count etc.
By getting the above data I can implement any classification model on the data to predict the required output(i.e open or not).

iii. Output of your finished dataset:

```
+-----------------------------------------------+-----------+-------+-------------+-
---------+-----------+-------+--------------+---------------------------------+-------
---------------------+---------+
| name                                          | city      | state | postal_code |
latitude | longitude | stars | review_count | category                        |
name  | is_open |
+-----------------------------------------------+-----------+-------+-------------+-
---------+-----------+-------+--------------+---------------------------------+-------
---------------------+---------+
| A Desert Custom Cycles                        | Mesa      | AZ    | 85201       |
33.4231 |   -111.84 |   5.0 |            4 | Motorcycle Dealers              |
None |       1 |
| Adobe Montessori                              | Chandler  | AZ    | 85226       |
33.3153 |  -111.954 |   5.0 |            5 | Montessori Schools              |
None |       1 |
| Apache Wash Trailhead                         | Phoenix   | AZ    | 85085       |
33.768 |  -112.045 |   5.0 |            9 | Active Life                     |
None |       1 |
```

| Back-Health Chiropractic | Phoenix | AZ | 85016 | 33.5028 | -112.013 | 5.0 | 19 | Health & Medical | BusinessAcceptsCreditCards | 1 |
| Big City Grill | Charlotte | NC | 28215 | 35.2495 | -80.7788 | 5.0 | 4 | Hot Dogs | None | 0 |
| Brandi Gilstrap | Henderson | NV | 89012 | 36.0161 | -115.058 | 5.0 | 5 | Hair Salons | None | 1 |
| Buddy's Muffler & Exhaust | Gastonia | NC | 28056 | 35.2772 | -81.06 | 5.0 | 4 | Auto Repair | BusinessAcceptsCreditCards | 1 |
| Camden Fairview | Charlotte | NC | 28226 | 35.1526 | -80.7952 | 5.0 | 6 | Home Services | ByAppointmentOnly | 1 |
| Christian Brothers Automotive | Chandler | AZ | 85249 | 33.248 | -111.837 | 5.0 | 63 | Transmission Repair | None | 1 |
| Clean Colonic | Tempe | AZ | 85283 | 33.3501 | -111.915 | 5.0 | 5 | Health & Medical | BusinessAcceptsCreditCards | 1 |
| Desert Medical Equipment | Las Vegas | NV | 89118 | 36.0964 | -115.187 | 5.0 | 4 | Shopping | None | 1 |
| Dollar Mania | Chandler | AZ | 85225 | 33.3497 | -111.858 | 5.0 | 4 | Event Planning & Services | BikeParking | 0 |
| Frankie Fettuccine Food Truck | Oakville | ON | L6J 6T4 | 43.5056 | -79.6611 | 5.0 | 7 | Food | RestaurantsPriceRange2 | 1 |
| Green Corner Restaurant | Mesa | AZ | 85210 | 33.3944 | -111.854 | 5.0 | 267 | Restaurants | None | 1 |
| Haggard Chiropractic | Phoenix | AZ | 85037 | 33.5085 | -112.268 | 5.0 | 18 | Doctors | None | 1 |
| Halo Plumbing | Henderson | NV | 89074 | 36.0376 | -115.076 | 5.0 | 5 | Plumbing | None | 1 |
| Innercity MMA | Toronto | ON | M5T 1G6 | 43.6536 | -79.3947 | 5.0 | 3 | Active Life | None | 1 |
| Jon Petrick, DC - Las Vegas Pain Relief Center | Las Vegas | NV | 89119 | 35.9985 | -115.109 | 5.0 | 5 | Doctors | None | 1 |
| Journey's Dry Carpet Cleaning | Charlotte | NC | 28270 | 35.1476 | -80.7499 | 5.0 | 3 | Carpet Cleaning | None | 0 |
| Kelsey's Pet Sitting & Dog Walking | Surprise | AZ | 85374 | 33.6573 | -112.465 | 5.0 | 3 | Pet Services | None | 0 |
| Lifestyles Fitness Personal Training | Tempe | AZ | 85282 | 33.3817 | -111.941 | 5.0 | 17 | Active Life | ByAppointmentOnly | 1 |

```
| Middleton Art and Framing                          | Middleton  | WI    | 53562       |
43.0967 |  -89.4983 |   5.0 |           8 | Framing                          |
None |       1 |
| Motors & More                                      | Las Vegas  | NV    | 89102       |
36.1465 |  -115.167 |   5.0 |           7 | Heating & Air Conditioning/HVAC |
ByAppointmentOnly |       1 |
| PC Savants                                         | Sun City   | AZ    | 85373       |
33.6901 |  -112.319 |   5.0 |          11 | Mobile Phone Repair              |
BusinessAcceptsBitcoin |       1 |
| PRO BIKE+RUN                                       | Pittsburgh | PA    | 15205       |
40.4521 |   -80.165 |   5.0 |           8 | Shopping                         |
WheelchairAccessible |       1 |
+----------------------------------------------------+------------+-------+-------------+-
---------+-----------+-------+--------------+----------------------------------+-------
---------------------+---------+
(Output limit exceeded, 25 of 183 total rows shown)
```

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

```sql
SELECT
b.name,
b.city,
b.state,
b.postal_code,
b.latitude,
b.longitude,
b.stars,
b.review_count,
c.category,
a.name,
b.is_open
FROM (business b
INNER JOIN category c ON b.id=c.business_id)
LEFT JOIN attribute a ON a.business_id=c.business_id
GROUP BY b.name
ORDER BY b.stars DESC;
```