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
ii. Hours = 1562
iii. Category = 2643
iv. Attribute = 1115
v. Review = 10000

vi. Checkin = 493
vii. Photo = 10000
viii. Tip = 3979 (Foreign Key: Business_ID)
ix. User = 10000
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(*) AS NullRecordNum
FROM USER
WHERE id IS NULL OR                    /* Restricting rows to those with at least one NULL field*/
      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 SUM(review_count) AS ReviewNum, City
FROM business
GROUP BY City
ORDER BY ReviewNum DESC;
```

Copy and Paste the Result Below:

ReviewNum	city
82854	Las Vegas
34503	Phoenix
24113	Toronto
20614	Scottsdale
12523	Charlotte
10871	Henderson
10504	Tempe
9798	Pittsburgh
9448	Montréal
8112	Chandler
6875	Mesa
6380	Gilbert
5593	Cleveland
5265	Madison
4406	Glendale
3814	Mississauga
2792	Edinburgh
2624	Peoria
2438	North Las Vegas
2352	Markham
2029	Champaign
1849	Stuttgart
1520	Surprise
1465	Lakewood
1155	Goodyear

(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 SUM(review_count), STARS
FROM business
WHERE City = 'Avon'
GROUP BY STARS;
```

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

SUM(review_count)	stars
10	1.5
6	2.5
88	3.5
21	4.0
31	4.5
3	5.0

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT SUM(review_count), STARS
FROM business
WHERE City = 'Beachwood'
GROUP BY STARS;
```

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

SUM(review_count)	stars
8	2.0
3	2.5
11	3.0
6	3.5
69	4.0
17	4.5
23	5.0

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?

No, it does not seem to be that more reviews are in proportion with higher fans.

Please explain your findings and interpretation of the results:

```
SELECT id, name, fans, review_count, yelping_since
FROM user
ORDER BY fans DESC
LIMIT 10;
```

id	name	fans	review_count	yelping_since
-9I98YbNQnLdAmcYfb324Q	Amy	503	609	2007-07-19 00:00:00
-8EnCioUmDygAbsYZmTeRQ	Mimi	497	968	2011-03-30 00:00:00
--2vR0DIsmQ6WfcSzKWigw	Harald	311	1153	2012-11-27 00:00:00
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	253	2000	2012-12-16 00:00:00
-0IiMAZI2SsQ7VmyzJjokQ	Christine	173	930	2009-07-08 00:00:00
-g3XIcCb2b-BD0QBCcq2Sw	Lisa	159	813	2009-10-05 00:00:00
-9bbDysuiWeo2VShFJJtcw	Cat	133	377	2009-02-05 00:00:00
-FZBTkAZEXoP7CYvRV2ZwQ	William	126	1215	2015-02-19 00:00:00
-9da1xk7zgnnfO1uTVYGkA	Fran	124	862	2012-04-05 00:00:00
-lh59ko3dxChBSZ9U7LfUw	Lissa	120	834	2007-08-14 00:00:00

Many of the highes reviewers (for eg. Sara, Yuri and Eric) have less than 100 fans.

Also, Amy, with the highest fans, has less than 1/3 reviews as Gerald, who has the fourth most reviews.
This means that user age (yelping_since) plays a significantly important role in predicting fans.

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

Answer: "love" (1780 > 232)

SQL code used to arrive at answer:

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

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

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

SQL code used to arrive at answer:

```
SELECT 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 picked Toronto and restaurants!

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

Not exactly, the lowest and best rated restaurants have the same open hours on Saturday while middle restaurant have more specialised times.

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

Yes, the higher rated restaurants had more reviews in total, but the highest rated one had comparitively few.

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

Similarly rated restaurants seem to be in nearby locations.

SQL code used for analysis:

```
SELECT business.name, business.city, category.category,
    business.stars, business.review_count, hours.hours,
business.postal_code
FROM (business INNER JOIN category ON
    business.id = category.business_id) INNER JOIN hours ON
    hours.business_id = category.business_id
WHERE City = 'Toronto' AND category.category = 'Restaurants'
    AND (Stars between 2.0 and 3.0 OR Stars between 4.0 and 5.0)
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:
All but 9 of all businesses that closed had no photos.

ii. Difference 2:
All businesses with more than 2 reviews stayed open.

SQL code used for analysis:

```
SELECT business.id, business.is_open, business.name,
business.postal_code,
    COUNT(photo.id) AS Pics, Count(Review.id) AS comments
From (business INNER JOIN photo ON
    business.id = photo.business_id) INNER JOIN Review ON
    review.business_id = business.id
GROUP BY business.id
/*HAVING Pics > 1 AND is_open = 1*/      /* Previously removed from
commentsset to 0 to find almost no closed shop with any photos */
HAVING comments > 2 AND is_open = 0;
```

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 higher rated stores have more photos.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
I will need to compare the number of photos of highly rated businesses and to the number of photos and gain insight as to whether there is a directly proportional trend. I can do this for many cities.

iii. Output of your finished dataset:

Pics	rating	city
52634.1052632	3.60526315789	Pittsburgh
5337.58333333	3.875	Madison
613.666666667	3.5	Cleveland Heights
83.3333333333	3.5	Thornhill
74.25	3.875	Gastonia
66.0	3.5	Newmarket
29.0	3.5	Fort Mill
19.0	3.0	Ludwigsburg
12.0	3.5	North Olmsted
10.6666666667	3.5	Davidson

All cities with high numbers of photos of images sported decent ratings of nearly all >= 3.5, but none of them are exceptionally high.

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

```
SELECT AVG(photo.id) AS Pics,
    AVG(business.stars) AS rating, business.City
From (business INNER JOIN photo ON
    business.id = photo.business_id)
GROUP BY business.city
ORDER BY Pics DESC
LIMIT 10;
```