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

```
## SELECT COUNT(*)
## FROM user;
```

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 = business_id = 8090, user_id = 9581
vi. Checkin = 493
vii. Photo = business_id = 6493, id = 10000
viii. Tip = business_id = 3979, user_id = 537
ix. User = 10000
x. Friend = 11
xi. Elite_years = 2780

```
## SELECT COUNT(DISTINCT(user_id))
## FROM elite_years;
```

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 compliment_photos ISNULL;
## I went through this code with every column in user, only changing "compliment_photos" to the correct column
```

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, review_count
FROM business
ORDER BY review_count DESC;
```

Copy and Paste the Result Below:

```
+------------+--------------+
| city       | review_count |
+------------+--------------+
| Las Vegas  |         3873 |
| MontrÃ©al   |         1757 |
| Gilbert    |         1549 |
```

Las Vegas	1410
Las Vegas	1389
Las Vegas	1252
Las Vegas	1116
Las Vegas	1084
Las Vegas	961
Gilbert	902
Las Vegas	864
Scottsdale	823
Las Vegas	821
Las Vegas	786
Henderson	785
Toronto	778
Las Vegas	768
Las Vegas	758
Scottsdale	726
Cleveland	723
Las Vegas	720
Charlotte	715
Phoenix	711
Las Vegas	706
Phoenix	700
+------------+--------------+

6. Find the distribution of star ratings to the business in the following cities:

i. Avon

SQL code used to arrive at answer:

```
SELECT review_count, stars
FROM business
WHERE city = "Avon"
ORDER BY stars ASC;
```

Copy and Paste the Resulting Table Below (2 columns Ã¢â‚¬â€œ star rating and count):

review_count	stars
10	1.5
3	2.5
3	2.5
7	3.5
31	3.5
50	3.5
4	4.0
17	4.0
31	4.5
3	5.0

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT review_count, stars
FROM business
WHERE city = "Beachwood"
ORDER BY stars ASC;
```

Copy and Paste the Resulting Table Below (2 columns Ã¢â‚¬â€œ star rating and count):

review_count	stars
8	2.0
3	2.5
8	3.0
3	3.0
3	3.5
3	3.5
69	4.0
14	4.5
3	4.5
6	5.0
4	5.0
6	5.0
3	5.0
4	5.0

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

SQL code used to arrive at answer:

```
SELECT review_count
FROM user
ORDER BY review_count DESC
LIMIT 3;
```

Copy and Paste the Result Below:

review_count
2000
1629
1339

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:
Yes it appears that it does. The reviewers with the lowest reviews tended to have the least fans while the reviewers with the most reviews tended to have the most fans.

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

Answer: hate = 232
love = 1780

SQL code used to arrive at answer:

```
SELECT COUNT(DISTINCT text)
FROM review
WHERE text LIKE '%love%';


SELECT COUNT(DISTINCT 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. The higher rated restaurants (upper class) tend to be open less than the lower rated restaurants(Lower class).

ii. Do the two groups you chose to analyze have a different number of reviews?
Upper class = (89,8,26)
Lower class = (5,34,47)
They're pretty similar when it comes to reviews.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
No. Every restaurant comes from a unique neighborhood in Toronto. With no pattern between upper and lower class restaurants we can't infer anything.

SQL code used for analysis:

```
SELECT review_count, neighborhood, category, name, stars, city, hours, class FROM
(SELECT neighborhood, review_count, name, stars, category, hours,stars, city,
case
WHEN stars in (2.0,2.5,3.0,3.5) THEN 'lower'
else 'upper' end as class
FROM
(SELECT review_count, neighborhood, name, stars, hours, category, city, stars
FROM business
INNER JOIN category ON business.id = category.business_id
INNER JOIN hours ON business.id = hours.business_id
WHERE city = "Toronto" AND category = "Restaurants" AND stars > 1.5
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:
Open businesses have more total stars.

ii. Difference 2:
Open businesses have more total reviews.

SQL code used for analysis:

```
SELECT SUM(review_count), SUM(stars) ,is_open
FROM business
GROUP BY is_open
```

SUM(review_count)	SUM(stars)	is_open
35261	5351.0	0
269300	31198.0	1

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 will be predicting the number of fans a user will have.

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 use the data from the user and elite_years table. From these two tables I will use the review_count, yelping_since, and year columns to predict. I will use id, and name to distinguish each user. I believe review_count, yelping_since, and year columns correlate most with the number of fans a user will have because only the best yelpers get Elite status, higher review counts equal more people seeing the user, and yelping_since equals more time for users to see the users reviews. With this data I would plug it into python and run a correlation matrix to see if my hypothesis was correct that the predictive collumns correlate most with fans. If I was right, then I would train a machine learning algorithm like linear regression or decision tree, whichever performs better, to predict a users fans based on review_count, yelping_since, and year columns.

iii. Output of your finished dataset:

id	name	review_count	yelping_since	year	fans
-9I98YbNQnLdAmcYfb324Q	Amy	609	2007-07-19 00:00:00	None	503
-8EnCioUmDygAbsYZmTeRQ	Mimi	968	2011-03-30 00:00:00	None	497
--2vR0DIsmQ6WfcSzKWigw	Harald	1153	2012-11-27 00:00:00	None	311
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	2000	2012-12-16 00:00:00	None	253
-0IiMAZI2SsQ7VmyzJjokQ	Christine	930	2009-07-08 00:00:00	None	173
-g3XIcCb2b-BD0QBCcq2Sw	Lisa	813	2009-10-05 00:00:00	None	159
-9bbDysuiWeo2VShFJJtcw	Cat	377	2009-02-05 00:00:00	None	133
-FZBTkAZEXoP7CYvRV2ZwQ	William	1215	2015-02-19 00:00:00	None	126

-9da1xk7zgnnfO1uTVYGkA	Fran	862	2012-04-05 00:00:00	None	124
-lh59ko3dxChBSZ9U7LfUw	Lissa	834	2007-08-14 00:00:00	2013	120
-B-QEUESGWHPE_889WJaeg	Mark	861	2009-05-31 00:00:00	None	115
-DmqnhW4Omr3YhmnigaqHg	Tiffany	408	2008-10-28 00:00:00	None	111
-cv9PPT7IHux7XUc9dOpkg	bernice	255	2007-08-29 00:00:00	None	105
-DFCC64NXgqrxlO8aLU5rg	Roanna	1039	2006-03-28 00:00:00	None	104
-IgKkE8JvYNWeGu8ze4P8Q	Angela	694	2010-10-01 00:00:00	None	101
-K2Tcgh2EKX6e6HqqIrBIQ	.Hon	1246	2006-07-19 00:00:00	None	101
-4viTt9UC44lWCFJwleMNQ	Ben	307	2007-03-10 00:00:00	None	96
-3i9bhfvrM3F1wsC9XIB8g	Linda	584	2005-08-07 00:00:00	None	89
-kLVfaJytOJY2-QdQoCcNQ	Christina	842	2012-10-08 00:00:00	None	85
-ePh4Prox7ZXnEBNGKyUEA	Jessica	220	2009-01-12 00:00:00	None	84
-4BEUkLvHQntN6qPfKJP2w	Greg	408	2008-02-16 00:00:00	None	81
-C-l8EHSLXtZZVfUAUhsPA	Nieves	178	2013-07-08 00:00:00	2014	80
-dw8f7FLaUmWR7bfJ_Yf0w	Sui	754	2009-09-07 00:00:00	None	78
-8lbUNlXVSoXqaRRiHiSNg	Yuri	1339	2008-01-03 00:00:00	None	76
-0zEEaDFIjABtPQni0XlHA	Nicole	161	2009-04-30 00:00:00	None	73

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

```
SELECT id, name, review_count, yelping_since, year, fans
FROM user LEFT JOIN elite_years on user.id = elite_years.user_id
GROUP BY id
ORDER BY fans DESC;
```