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 = 2052
iii. Category = 2643
iv. Attribute = 1115
v. Review = 10000
vi. Checkin = 493
vii. Photo = 10000
viii. Tip = 537 user_id, 3979 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 *
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 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 city, SUM(review_count) AS Total_review_count
FROM business
GROUP BY city
ORDER BY SUM(review_count) DESC
```

Copy and Paste the Result Below:

```
| city            | Total_review_count |
+-----------------+--------------------+
```

Las Vegas	82854
Phoenix	34503
Toronto	24113
Scottsdale	20614
Charlotte	12523
Henderson	10871
Tempe	10504
Pittsburgh	9798
MontrÃ©al	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 as [star rating], count(stars) AS count
FROM business
WHERE city = "Avon"
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns â€“ 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 â€“ 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 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 doesn't.

Please explain your findings and interpretation of the results:

The result for the search for the top 3 fans of the users is:

```
+------------------------+--------+--------------+------+
| id                     | name   | review_count | fans |
+------------------------+--------+--------------+------+
| -9I98YbNQnLdAmcYfb324Q | Amy    |          609 |  503 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi   |          968 |  497 |
| --2vR0DIsmQ6WfcSzKWigw | Harald |         1153 |  311 |
+------------------------+--------+--------------+------+
```

This indicates that, top three users based on review count (answer of the last question) don't have the three most number of fans.

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

Answer: Yes, there are 1780 reviews with "love" word whereas 232 reviews with the word "hate" in them.

SQL code used to arrive at answer:

```
+------+-----------+
| Word | count(id) |
+------+-----------+
| hate |       232 |
| love |      1780 |
+------+-----------+
```

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.

City : Toronto and category : Bars

stars	Number of Bars	Open Days overall	Avg. open days	Total_reviews
2.5	1	7	7	35
3.5	1	6	6	10
4.0	1	6	6	15
4.5	1	7	7	26

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

No. These two groups have similar distribution of hours. Between 2 and 3 stars, we have 1 bar with 7 avg. open days. Whereas, 4-5 stars two bars have average 6.5 (7 and 6) open days.

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

Yes, They have different number of reviews. 2.5 star one has the most number of reviews which is 35. However, the two 4-5 star bars have an average of ~ 20 reviews per bar. The difference its not too large though.

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

Its is difficult to provide any inference from these set of data as the difference in the number oof reviews are almost in the similar range.
SQL code used for analysis:

```
SELECT stars, COUNT (DISTINCT id) AS [Number of Bars],
COUNT(hours) AS [Open Days overall], (COUNT(hours) / COUNT(DISTINCT id)) AS [Avg. open
days],
(SUM(review_count) / COUNT(hours)) AS Total_reviews
FROM
((business b
JOIN hours h
ON b.id = h.business_id)
JOIN category c
ON h.business_id = c.business_id)
WHERE category = "Bars" AND city = "Toronto"
GROUP BY 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.

is_open	Number of business	Avg_stars	Total reviews
0	1520	3.52039473684	35261
1	8480	3.67900943396	269300

i. Difference 1:

Total number of reviews for open businesses are larger than the closed ones.

ii. Difference 2:

More number of businesses are open than closed.

SQL code used for analysis:

```
SELECT is_open, COUNT(is_open) AS [Number of business], SUM(stars)/COUNT(is_open) AS
Avg_stars,
SUM(review_count) AS [Total reviews]
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:

Information about the yelp users

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

The main idea behind this analysis is to gather some information about the yelp users, including their name, how long they are using yelp, how many and what types of reviews they provided and how many fans they have etc.
The main goal will be to show how the results are related or not.

iii. Output of your finished dataset:

This is the dataset after sorting in the decreasing order with respect to review counts.

name	review_count	fans	average_stars	yelping_since	Loyalty
Gerald	2000	253	3.6	2012-12-16 00:00:00	10
Sara	1629	50	3.42	2010-05-16 00:00:00	13
Yuri	1339	76	4.11	2008-01-03 00:00:00	15
.Hon	1246	101	3.14	2006-07-19 00:00:00	17
William	1215	126	4.41	2015-02-19 00:00:00	8
Harald	1153	311	4.4	2012-11-27 00:00:00	10
eric	1116	16	3.31	2007-05-27 00:00:00	16
Roanna	1039	104	3.71	2006-03-28 00:00:00	17
Mimi	968	497	4.05	2011-03-30 00:00:00	12
Christine	930	173	3.69	2009-07-08 00:00:00	14

Inferences: 1. Number of reviews does not correspond to the number of years the user his using yelp.
2. Review count does not effect the number of fans for the users.

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

```
SELECT name, review_count, fans, average_stars, yelping_since,
(strftime('%Y', 'now') - strftime('%Y', yelping_since))
     - (strftime('%m-%d', 'now') < strftime('%m-%d', yelping_since)) AS Loyalty
FROM
user
ORDER BY review_count DESC
LIMIT 10
```