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:

CODE USED:

```
SELECT COUNT(*)
FROM insert_table_name
```

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.

CODE USED:

```
SELECT COUNT(DISTINCT(key))
FROM insert_table_name
```

i. Business = primary key = id: 10000
ii. Hours = foreign key = business_id: 1562
iii. Category = foreign key = business_id: 2643
iv. Attribute = foreign key = business_id: 1115
v. Review = primary key = id: 10000; foreign key = business_id: 8090; foreign key = user_id: 9581
vi. Checkin = foreign key = business_id: 493
vii. Photo = primary key = id: 10000 ; foreign key = business_id: 6493
viii. Tip = foreign key = user_id: 537 ; foreign key = business_id: 3979
ix. User = primary key = id: 10000
x. Friend = foreign key = user_id: 11
xi. Elite_years = foreign key = user_id: 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. There are no columns with null values in the User's table.

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:

CODE USED:

```
SELECT col_name,
min(col_name) AS col_name_min,
max(col_name) AS col_name_max,
avg(col_name) AS col_name_avg
FROM insert_table_name
```

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 totalviews
FROM business
GROUP BY city
ORDER BY totalviews DESC
```

Copy and Paste the Result Below:

city	totalviews
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

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,
SUM(review_count) AS count
FROM business
WHERE city == 'Avon'
GROUP BY stars
```

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

stars	count

```
|   1.5 |    10 |
|   2.5 |     6 |
|   3.5 |    88 |
|   4.0 |    21 |
|   4.5 |    31 |
|   5.0 |     3 |
+-------+-------+
```

ii. Beachwood

SQL code used to arrive at answer:

```sql
SELECT
stars,
SUM(review_count) AS count
FROM business
WHERE city == 'Beachwood'
GROUP BY stars
```

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

```
+-------+-------+
| stars | count |
+-------+-------+
|   2.0 |     8 |
|   2.5 |     3 |
|   3.0 |    11 |
|   3.5 |     6 |
|   4.0 |    69 |
```

```
|   4.5 |    17 |
|   5.0 |    23 |
+-------+-------+
```

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

   SQL code used to arrive at answer:

```sql
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 posing more reviews correlate with more fans?

   Please explain your findings and interpretation of the

results:

According to the results, posing more reviews doesn't mean that they will have more fans. For example,
Amy has a review count of 609 and has 503 fans (most fans). Gerald, however, has a higher review count
of 2000 but only has 253 fans. Therefore, perhaps some other factors such as how long have they been yelping
or number of useful or funny reviews, etc. may also contribute to the number of fans each users have.

CODE USED:

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

RESULTS:

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

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

Answer: According to the results, there are more reviews with the word “love” than “hate”.

SQL code used to arrive at answer:

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

SELECT
```

```
COUNT(*)
FROM review
WHERE text LIKE '%hate%'
```

```
+----------+
| COUNT(*) |
+----------+
|     1780 |
+----------+
```

```
+----------+
| COUNT(*) |
+----------+
|      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 picked was Las Vegas and the category was restaurants.

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

The city that I picked was Las Vegas and the category was Food. According to the results,
I wouldn't say there's a different distribution of hours between the 2 groups. The hours for
2.5 stars is around 8:00 - 22:00 and the hours for 4.0 stars is around 10:00 - 19:00 on Saturdays.

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

Yes. The review count for 2.5 stars only has 6 review counts and 4.0 stars has 30 review counts.

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

No, their postal codes are fairly similar. The postal code for 2.5 stars is 89121 and the postal code
For 4.0 stars is 89123.

SQL code used for analysis:

```
SELECT
b.name,
b.city,
b.address,
b.postal_code,
b.stars,
b.review_count,
h.hours,
c.category
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 = "Las Vegas" AND c.category =
"Food"
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:

Businesses that are open have more total review counts (269300) than those that are closed (35261).

ii. Difference 2:

Businesses that are open have more stars (3.7) on average than those that are closed (3.5).

SQL code used for analysis:

```
SELECT
COUNT(DISTINCT(id)),
avg(stars),
avg(review_count),
sum(review_count),
is_open
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:

Here, I wish to create a data set that can help predict whether a business will stay open or close based on selected features.

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

The aim of this study is to help businesses to identify important factors such as the number of stars, review counts, open hours, category, and demographic factors such as city and state that can help determine whether their business will stay open or not. Businesses can also use this information to make adjustments and improve overall performances. In addition, this data set can also be used for businesses that are still looking for the specific city to open their stores at by conducting clustering analysis.

iii. Output of your finished dataset:

```
+----------------------+-------+-----------------
+----------------------------------------+-------+--------------
+---------------------+-----------------------+---------+
| id                    | state | city            | name
| stars | review_count | hours                | category              | is_open |
```

```
+-----------------------+-------+-----------------
+---------------------------------------+-------+-------------
+----------------------+-----------------------+---------+
| 2YmDZid3sYULrT60sRjuhA | AZ    | Chandler        | Red Apron Bakeshop
|   5.0 |            5 | Saturday|9:00-17:00  | Bakeries               |       1 |
| 0-aPEeNc2zVb5Gp-i7Ckqg | NC    | Gastonia        | Buddy's Muffler & Exhaust
|   5.0 |            4 | Saturday|9:00-15:00  | Auto Repair            |       1 |
| 0XQJbJil68MFRSZfNuo8cg | NV    | Henderson       | Brandi Gilstrap
|   5.0 |            5 | Saturday|9:00-16:00  | Hair Salons            |       1 |
| 2RhICgMZI6DK-t374VRoow | NV    | Las Vegas       | Desert Medical Equipment
|   5.0 |            4 | Monday|8:00-17:00    | Shopping               |       1 |
| 0Y3lHyqRHfWOBuQlS1bM0g | AZ    | Sun City        | PC Savants
|   5.0 |           11 | Saturday|11:00-18:00 | Mobile Phone Repair    |       1 |
| 1gD96cTZxZAZ8TsRN3EKbw | AZ    | Surprise        | Kelsey's Pet Sitting & Dog
Walking      |   5.0 |            3 | Saturday|7:00-20:00  | Pet Services           |
0 |
| 2vz3U82Sf8GgXppyjGSdbg | WI    | Cottage Grove   | Oaks Golf Course
|   4.5 |            5 | Saturday|7:00-19:00  | Active Life            |       1 |
| 2jg7v96HM3mNSUrbk3sMxg | AZ    | Mesa            | Health For Life: North Mesa
|   4.5 |           16 | Saturday|9:00-20:00  | Cannabis Clinics       |       1 |
| -KWHuAnRPvNBIh2yhBC2kg | OH    | Peninsula       | The Wine Mill
|   4.5 |           42 | Saturday|15:00-23:00 | Nightlife              |       1 |
| 0Ni7Stqt4RFWDGjOYRi2Bw | AZ    | Scottsdale      | Scent From Above Company
|   4.5 |           14 | Monday|6:00-16:00    | Window Washing         |       1 |
| 29fQtyR9EtAlA75e4jGzRw | AZ    | Tempe           | Ahn & Perez, DDS
|   4.5 |           13 | Monday|8:00-17:00    | Dentists               |       1 |
| 1D7U-KEvoQDqWJNiYTNbZg | OH    | Chesterland     | Rise and Dine Cafe
|   4.0 |           30 | Saturday|6:00-15:00  | Restaurants            |       1 |
| 2xcnolaD9e6voXJnrbu_Hg | OH    | Cleveland       | B.A. Sweetie Candy
|   4.0 |           49 | Saturday|10:00-20:00 | Candy Stores           |       1 |
| 2aiaryk7kgUBhXhVu-9vHg | EDH   | Edinburgh       | Miros Cantina Mexicana
|   4.0 |           37 | Saturday|12:00-22:30 | Restaurants            |       1 |
| 0KQJYTJuX4qDE-8dRqNl6w | OH    | Mantua          | Roundup Lake
|   4.0 |            8 | Monday|8:30-17:30    | Hotels & Travel        |       1 |
| 27nh-2hNnNkf2dBk9aeKHQ | WI    | Middleton       | C's Restaurant Bakery and Coffee
Shop   |   4.0 |           37 | Saturday|6:00-14:00  | American (Traditional) |
1 |
| 2skQeu3C36VCiB653MIfrw | AZ    | Phoenix         | Bootleggers Modern American
Smokehouse  |   4.0 |          431 | Saturday|11:00-22:00 | Barbeque               |
1 |
| -iu4FxdfxN4rU4Fu9BjiFw | OH    | Strongsville    | Alterations Express
|   4.0 |            3 | Saturday|8:00-18:00  | Sewing & Alterations   |       1 |
| 37kk0IW6jL7ZlxZF6k2QBg | ON    | Toronto         | Edulis
|   4.0 |           89 | Saturday|18:00-23:00 | Restaurants            |       1 |
| 0NDbUCHi9YsRwgG3iZO8Kg | OH    | Aurora          | Cafe Tandoor
|   3.5 |           32 | Saturday|11:30-14:00 | Restaurants            |       1 |
| 2CH7fxD6h7H1CReBrBxOgg | ON    | Brampton        | Hilight Essence Hair Studio &
Esthetics |   3.5 |            8 | Saturday|9:00-18:00  | Nail Salons            |
1 |
| 2R_z-xwaSFjuRAEWKX0oDw | NC    | Charlotte       | Gorgeous Glo
|   3.5 |           10 | Saturday|11:00-16:00 | Beauty & Spas          |       1 |
| 0kzPQQL8wVcHlBQzMdRdWQ | WI    | Fitchburg       | Thirsty Goat
|   3.5 |           74 | Saturday|11:00-2:00  | American (Traditional) |       1 |
| 1OxSzNUssdRohY5dC-kWVg | AZ    | Fountain Hills  | Ping's Cafe
|   3.5 |           21 | Saturday|11:00-21:00 | Chinese                |       1 |
| 11bhfBbcFypczdz3N_w6iw | AZ    | Litchfield Park | Senor Taco
```

```
|   3.5 |          83 | Saturday|7:00-22:00  | Restaurants           |       1 |
+----------------------+-------+----------------
+---------------------------------------+-------+-------------
+---------------------+-----------------------+---------+
```

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

```
                SELECT
                b.id,
                b.state,
                b.city,
                b.name,
                b.stars,
                b.review_count,
                h.hours,
                c.category,
                b.is_open
                FROM business b INNER JOIN hours h ON
b.id=h.business_id
                                INNER JOIN category c ON
b.id=c.business_id
                GROUP BY city
                ORDER BY stars DESC
```