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 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

*******SQL CODE ***********

```
Select COUNT (*)
From Table
**************************
```

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. Attribute table = business_id: 1115
ii. Business table = id: 10000
iii. Category table = business_id: 2643
iv. Checkin table = business_id: 493
v. elite_years table = user_id: 2780
vi. friend table = user_id: 11, friend_id: 9415
vii. hours table = business_id: 1562
viii. photo table = business_id: 6493, id: 10000
ix. review table = id: 10000, business_id: 8090, user_id: 9581
x. tip table = business_id: 3979, user_id: 537
xi. user table = id: 10000

*******SQL CODE ***********

```
Select COUNT (DISTINCT keys)
From Table
```

****

3. Are there any columns with null values in the Users table? Indicate "yes," or "no."

No

*******SQL CODE ******************

```
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:

Table Name	Column Name	Min	Max	Average
Review	Stars	1	5	3.7082
Business	Stars	1	5	3.6549
Tip	Likes	0	2	0.0144
Checkin	Count	1	53	1.9414
User	Review_count	0	2000	24.2995

*******SQL CODE ***********

```
Select  min(col_name),
        max(col_name),
        avg (col_name)
from table
**************************
```

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

```
+-----------------+--------------+
| city            | total_review |
+-----------------+--------------+
| 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)
```

*******SQL CODE ***********

```
Select  city,
        SUM (review_count) as total_review
from business
group by city
order by total_review desc
**************************
```

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

i. Avon

*******SQL CODE ***********

```
Select SUM (review_count) as totat_count
      , stars
from business
where city= "Avon"
group by stars
**************************
```

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

ii. Beachwood

*******SQL CODE ***********

```
Select SUM (review_count) as totat_count
      , stars
from business
where city= "Beachwood"
group by stars
**************************
```

```
+-------------+-------+
| totat_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 ***********

```
Select name,
     review_count
from user
order by review_count desc
limit 3
**************************
```

```
+--------+--------------+
| name   | review_count |
+--------+--------------+
| Gerald |         2000 |
| Sara   |         1629 |
| Yuri   |         1339 |
+--------+--------------+
```

8. Does posing more reviews correlate with more fans?

*******SQL CODE **********

```
Select name,
     , review_count
     , fans
from user
order by fans desc
**************************
```

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

(Output limit exceeded, 25 of 10000 total rows shown)

After looking at the results, I cannot find any correlation between review_count and fans.
Like Gerald has the highest review_count with very less fans. Amy, who has the most fans, has only 609 reviews.

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

There are 1780 reviews with the word love. However, the word hate came just 232 times.

*******SQL CODE **********

```
Select count (*)
from review
where text like "%love%"

Select count (*)
from review
where text like "%hate%"

**************************
```

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

*******SQL CODE ***********

```
Select name
    , fans
from user
order by fans desc
limit 10
```

****

```
+-----------+------+
| 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 analyse have a different distribution of hours?

I choose Charlotte as my city and category is Shopping. The shopping complex with 3.5 stars opens from 10:00-15:00. The place with higher rating of 4.0 open till late night that is, 12:00 – 22:00.

*******SQL CODE **********

```
select b.name, b.city , b.stars, c.category,h.hours
from
        (business b inner join category c
          on b.id = c.business_id
        )
        inner join hours h
        on b.id = h.business_id
where city = "Charlotte" and category="Shopping"
group by b.stars
```

****

```
+-------------------------+-----------+-------+----------+----------------------+
| name                    | city      | stars | category | hours                |
+-------------------------+-----------+-------+----------+----------------------+
| Dilworth Custom Framing | Charlotte |   3.5 | Shopping | Saturday|10:00-15:00 |
| HighLife North Tryon    | Charlotte |   4.0 | Shopping | Saturday|12:00-22:00 |
```

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

Yes, the two groups have different number of reviews. The Business with the higher stars has a lower review count as compared to the other business.

*******SQL CODE **********

```
select b.name, b.city , b.stars, c.category,h.hours, b.review_count
from
        (business b inner join category c
          on b.id = c.business_id
        )
        inner join hours h
        on b.id = h.business_id
where city = "Charlotte" and category="Shopping"
group by b.stars
```

****

```
+------------------------+-----------+-------+----------+----------------------+--------------+
| name                   | city      | stars | category | hours                | review_count |
+------------------------+-----------+-------+----------+----------------------+--------------+
| Dilworth Custom Framing | Charlotte |   3.5 | Shopping | Saturday|10:00-15:00 |            6 |
| HighLife North Tryon    | Charlotte |   4.0 | Shopping | Saturday|12:00-22:00 |            5 |
+------------------------+-----------+-------+----------+----------------------+--------------+
```

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

No, I cannot see any change with respect to location.

*******SQL CODE **********

```
select b.name, b.city , b.stars, c.category,h.hours, b.review_count, b.location, b.postal_code
from
        (business b inner join category c
          on b.id = c.business_id
        )
        inner join hours h
        on b.id = h.business_id
where city = "Charlotte" and category="Shopping"
group by b.stars
```

****

```
+------------------------+-----------+-------+----------+----------------------+--------------
+-------------------------+-------------+
| name                   | city      | stars | category | hours                | review_count |
address                 | postal_code |
+------------------------+-----------+-------+----------+----------------------+--------------
+-------------------------+-------------+
| Dilworth Custom Framing | Charlotte |   3.5 | Shopping | Saturday|10:00-15:00 |            6 |
125 Remount Rd, Ste C-2 | 28203       |
| HighLife North Tryon    | Charlotte |   4.0 | Shopping | Saturday|12:00-22:00 |            5 |
9605 N Tryon St, Ste C  | 28262       |
+------------------------+-----------+-------+----------+----------------------+--------------
+-------------------------+-------------+
```

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.

Difference 1: The business that are open have more average review count and stars as compared to the ones that are closed.

*******SQL CODE **********

```
select COUNT (Distinct b.id),
    b.is_open ,
    avg(b.stars),
    avg(b.review_count)
from business b
group by is_open
```

****

```
+----------------------+---------+--------------+---------------------+---------------------+
| COUNT (Distinct b.id) | is_open |  avg(b.stars) | sum(b.review_count) | avg(b.review_count) |
+----------------------+---------+--------------+---------------------+---------------------+
|                 1520 |       0 | 3.52039473684 |               35261 |       23.1980263158 |
|                 8480 |       1 | 3.67900943396 |              269300 |       31.7570754717 |
+----------------------+---------+--------------+---------------------+---------------------+
```

Difference 2: The businesses that are open are more useful and funnier as compared to the ones that are closed now. This shows that the business that are closed were not working properly.

*******SQL CODE **********

```
select b.is_open ,
    sum(review_count) as review,
    avg(b.stars) avg_stars,
    avg(b.review_count) as avg_count ,
    count(r.useful) as useful,
    count(r.funny) as funny
from business b inner join review r
on b.id = r.id
group by b.is_open
```

****

```
+---------+--------+---------------+---------------+--------+-------+
| is_open | review |     avg_stars |     avg_count | useful | funny |
+---------+--------+---------------+---------------+--------+-------+
|       0 |      4 |           2.0 |           4.0 |      1 |     1 |
|       1 |    504 | 2.96153846154 | 38.7692307692 |     13 |    13 |
+---------+--------+---------------+---------------+--------+-------+
```

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.

i. Indicate the type of analysis you chose to do:

I preferred to study the different types of restaurants that provides different foods. Then I selected specific type of food categories on yelp

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

Firstly, I preferred to see the names that contains restaurants as a keyword. Then I picked up few types of food in which I want to perform the analysis. I wanted to see the star rating, number of reviews so that I can get some insights on which type of food is most popular on yelp.

iii. Output of your finished dataset:

category	number_of_restaurants	avg(b.stars)	avg(review_count)	city	hours
Arabian	7	5.0	267.0	Mesa	Saturday|10:30-22:00
Mediterranean	7	5.0	267.0	Mesa	Saturday|10:30-22:00
Korean	7	4.5	8.0	Toronto	Saturday|11:00-23:00
French	12	4.0	135.083333333	Las Vegas	Saturday|11:00-20:00
Chinese	13	3.76923076923	423.230769231	Las Vegas	Saturday|10:00-23:00
Mexican	28	3.625	73.0	Edinburgh	Saturday|12:00-22:30
Italian	13	3.53846153846	78.2307692308	Montréal	Saturday|11:30-0:00
Indian	6	3.5	32.0	Aurora	Saturday|11:30-14:00

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

```
Select c.category,
        count(b.name) as number_of_restaurants,
        avg( b.stars),
        avg(review_count),
        city,
        h.hours
from (business b inner join category c
            on b.id = c.business_id)
inner join hours h
            on b.id = h.business_id
where c.category in
("Chinese","Arabian","Italian","Japenese","Mediterranean","Mexican","French","Korean","Indian")
group by category
ORDER BY avg(STARS) DESC
```