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 = business_id varchar(pk)- 10000
ii. Hours = business_id varchar (fk) -1562
iii. Category = business_id varchar (fk). –2643
iv. Attribute = business_id varchar (fk) –1115
v. Review = id varchar(pk)–10000,business_id varchar (fk)–8090 ,user_id varchar (fk)..9581
vi. Checkin = business_id varchar (fk)–493
vii. Photo = id varchar (pk)10,000,business_id varchar (fk)…6493
viii. Tip = business_id varchar (fk)…3979,user_id varchar (fk)…537
ix. User = id varchar (fk)…10000
x. Friend = user_id varchar (fk)….11
xi. Elite_years = user_id varchar (fk)..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 compliment_photos is null
```

-- I did not find any better approach, sorry about that --

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

```
select max(s), min(s), avg(s)
from

(SELECT stars as s
FROM review)
```

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

SQL code used to arrive at answer:

```
select city,sum(review_count)
     from business

group by 1
order by 2 desc
```

Copy and Paste the Result Below:

```
+------------+---          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 city, stars, count(stars)
from business
where city like 'avon'

group by 2
```

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

city	stars	count(stars)
Avon	1.5	1
Avon	2.5	2
Avon	3.5	3
Avon	4.0	2
Avon	4.5	1
Avon	5.0	1

ii. Beachwood

SQL code used to arrive at answer:

```
select city, stars, count(stars)
from business
where city like 'Beachwood'

group by 2
```

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

city	stars	count(stars)
Beachwood	2.0	1
Beachwood	2.5	1
Beachwood	3.0	2
Beachwood	3.5	2
Beachwood	4.0	1
Beachwood	4.5	2
Beachwood	5.0	5

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

SQL code used to arrive at answer:

```
select name, max(review_count )
from user
group by 1
order by 2 desc
limit 3
```

Copy and Paste the Result Below:

name	max(review_count)
Gerald	2000
Sara	1629
Yuri	1339

8. Does posing more reviews correlate with more fans?
yes

Please explain your findings and interpretation of the results:

They r positively correlated. One good reviews earn you approx. 10 new fans.

```
+------+-------------------+
| fans | max(review_count) |
+------+-------------------+
|  253 |              2000 |
|   50 |              1629 |
+------+-------------------+
+------+-------------------+
| fans | min(review_count) |
+------+-------------------+
|    0 |                 0 |
|    1 |                 1 |
+------+-------------------+

+------+-------------------+
| fans | avg(review_count) |
+------+-------------------+
|   63 |               6.0 |
|   70 |               7.0 |
+------+-------------------+
```

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

Love haas been used more than hate approx.8 times more

Answer:

```
+-----++----------+
| count(*) |
+----------+
|      232 |
+----------+
| count(*) |
```

```
+----------+
|     1780 |
+----------+
```

SQL code used to arrive at answer:

```
select count(*)
from
(select text
from review
where text like '%hate%' )

select count(*)
from
(select text
from review
where text like '%love%' )
```

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

SQL code used to arrive at answer:

```
select name, max(fans)
from user
group by 1
order by 2 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.

City - Las Vegas
Group one - Las Vegas with 2-3 stars
Group two - Las Vegas with 4-5 stars

i. Do the two groups you chose to analyze have a different distribution of hours?
47 restaurants are open on weekdays for 4 or 5 star restaurants in Vegas
87restaurants are open on weekdays for 4 or 5 star restaurants in Vegas

ii. Do the two groups you chose to analyze have a different number of reviews?
Las Vegas with 4 and 5 stars have double the reviews as compared to the ones with 2 and 3 stars.

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

MOST OF THE RESTAUENTS ON THE STRIP ARE HAVE A STAR RATING OF 2-3 RATHER THAN 4 AND 5.

SQL code used for analysis:

PART 1
```
select h.hours, count(h.hours) as count_weekday_open
from  hours h
join business b
on h.business_id = b.id
where  ( b.city = "Las Vegas" ) and (b.stars = 4 or  5) and
(h.hours like 'm%' or h.hours like 'T%' or h.hours like 'W%' or
h.hours like  'F%')
group by 1
```
Stars 4 and 5
```
+-+--------------------+
```

```
| count_weekday_open |
+--------------------+
|                 47 |
+--------------------+
```

Stars 2 and 3

```
+--------------------+
| count_weekday_open |
+--------------------+
|                 87 |
+--------------------+
```

PART 2

```
select review_count
from business b
where  ( b.city = "Las Vegas" ) and (b.stars = 4 or b.stars = 5)
group by 1
```

LAS_VEGAS

Stars 4 and 5

```
+------------------+
| review_count_4_5 |
+------------------+
|                3 |
|                4 |
|                5 |
|                6 |
|                7 |
|                8 |
|                9 |
|               10 |
|               11 |
|               12 |
|               13 |
|               14 |
|               15 |
|               16 |
|               17 |
|               18 |
|               19 |
```

```
|                20 |
|                21 |
|                22 |
|                23 |
|                24 |
|                25 |
|                26 |
|                27 |
+-------------------+
(Output limit exceeded, 25 of 125 total rows shown)

Stars 2 and 3
+-------------------+
| review_count_2_3 |
+-------------------+
|                 3 |
|                 4 |
|                 5 |
|                 6 |
|                 7 |
|                 8 |
|                 9 |
|                10 |
|                11 |
|                12 |
|                13 |
|                14 |
|                15 |
|                16 |
|                17 |
|                18 |
|                19 |
|                20 |
|                21 |
|                22 |
|                23 |
|                24 |
|                25 |
```

```
|               26 |
|               27 |
+------------------+
(Output limit exceeded, 25 of 83 total rows shown)
```

PART 3

```
select neighborhood, count(*) in_las_vegas
from
(select neighborhood
from business b
where  ( b.city = "Las Vegas" ) and (b.stars = 2 or b.stars =
3) )
group by 1
order by 2 desc
limit 5
```

```
+--------------+--------------+
| neighborhood | in_las_vegas_2_3 |
+--------------+--------------+
|              |           45 |
| The Strip    |           41 |
| Southeast    |           34 |
| Westside     |           32 |
| Eastside     |           26 |
+--------------+--------------+
```

```
+---------------+------------------+
| neighborhood  | in_las_vegas_4_5 |
+---------------+------------------+
|               |              119 |
| Southeast     |               69 |
| Spring Valley |               69 |
| Westside      |               66 |
| The Strip     |               44 |
+---------------+------------------+
```

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:

There are way more eateries open in Las Vegas as compared to Toronto. About 8 times more.

ii. Difference 2:

Postal code 89109, 89103 and 89119 has maximum numbers of star rating as 4 and 5.
Postal code 89102 has maximum numbers of star rating as 2 and 3.

SQL code used for analysis:

```
Part 1
select is_open, count(*) as open_las_vegas_2_3
from business b
where city = 'Las Vegas'and (b.stars = 2 or b.stars = 3)
group by 1
```

is_open	open_las_vegas_2_3
0	63
1	215

is_open	open_las_vegas_4_5
0	66
1	506

```
Part 2
select postal_code, count(postal_code) as postal_code_2_3
from
(select *
from business b
where city = 'Las Vegas' and is_open =1 and (b.stars = 2 or
b.stars = 3)
```

```
group by 1)
group by 1
order by 2 desc
limit 5
```

postal_code	postal_code_2_3
89109	25
89119	19
89102	15
89103	14
89117	9

postal_code	postal_code_4_5
89109	47
89103	28
89118	28
89119	28
89146	27

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:
 Parsing out keywords and business attributes for sentiment

analysis

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
Predicting how long people have using yelp for a maximum years
Youngest user

iii. Output of your finished dataset:
Youngest user is 3
Maximum a customer has been yelping since 15 years

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

```
select yelping_since ,
strftime('%Y',yelping_since ) as year,
strftime('%m',yelping_since ) as month,
strftime('%d',yelping_since ) as day,
date('now')-(strftime('%Y',yelping_since )) as age


from  user
order by age
limit 1
```

```
+---------------------+------+-------+-----+-----+
| yelping_since       | year | month | day | age |
+---------------------+------+-------+-----+-----+
| 2005-05-25 00:00:00 | 2005 | 05    | 25  |  15 |
+---------------------+------+-------+-----+-----+
```

```
select yelping_since ,
strftime('%Y',yelping_since ) as year,
strftime('%m',yelping_since ) as month,
strftime('%d',yelping_since ) as day,
date('now')-(strftime('%Y',yelping_since )) as age


from  user
order by age desc
limit 1
```

```
+---------------------+------+-------+-----+-----+
```

```
| yelping_since       | year | month | day | age |
+---------------------+------+-------+-----+-----+
| 2005-05-25 00:00:00 | 2005 | 05    | 25  |  15 |
+---------------------+------+-------+-----+-----+
```