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

_____SQL_CODE_____
```
SELECT COUNT(*)
FROM   table_name
```
****

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 = id : 1000
ii. Hours = business_id : 1562
iii. Category = business_id : 2643
iv. Attribute = business_id : 1115
v. Review = id : 10000, business_id : 8090, user_id : 9581
vi. Checkin = business_id : 493
vii. Photo = id : 10000, photo : 6493
viii. Tip = user_id : 537, business_id : 3979
ix. User = id : 10000
x. Friend = user_id : 11
xi. Elite_years = user_id : 2780

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

_____SQL CODE_____

```
SELECT COUNT(DISTINCT Keys)
FROM table_name
```

_____

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

Answer: NO

SQL code used to arrive at the answer:

```
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
+----------+
| COUNT(*) |
+----------+
|        0 |
+----------+
```

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: 20.2995

_____SQL_CODE_____
```
SELECT MIN(COLUMN_NAME), MAX(COLUMN_NAME), AVG(COLUMN_NAME)
 FROM  TABLE_NAME
```
_____

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

SQL code used to arrive at the answer:

```
SELECT SUM(REVIEW_COUNT) AS MOST_REVIEW,CITY
FROM BUSINESS
GROUP BY CITY
ORDER BY MOST_REVIEW DESC
```

Copy and Paste the Result Below:

```
+-------------+-----------------+
| MOST_REVIEW | city            |
+-------------+-----------------+
|       82854 | Las Vegas       |
|       34503 | Phoenix         |
|       24113 | Toronto         |
|       20614 | Scottsdale      |
|       12523 | Charlotte       |
|       10871 | Henderson       |
|       10504 | Tempe           |
|        9798 | Pittsburgh      |
|        9448 | Montréal        |
|        8112 | Chandler        |
|        6875 | Mesa            |
|        6380 | Gilbert         |
|        5593 | Cleveland       |
|        5265 | Madison         |
|        4406 | Glendale        |
|        3814 | Mississauga     |
|        2792 | Edinburgh       |
|        2624 | Peoria          |
|        2438 | North Las Vegas |
|        2352 | Markham         |
|        2029 | Champaign       |
|        1849 | Stuttgart       |
|        1520 | Surprise        |
|        1465 | Lakewood        |
|        1155 | Goodyear        |
```

```
+-------------+-----------------+
(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 ,sum(review_count) as counts
from business
where city == "Avon"
group by stars
```

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

```
+-------+--------+
| stars | counts |
+-------+--------+
|   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 the answer:

```
select  stars ,sum(review_count) as counts
from business
where city == "Beachwood"
group by stars
```

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

stars	counts
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 the answer:

```
select  name,sum(review_count) as counts
from user
group by name
order by review_count desc
limit 3
```

Copy and Paste the Result Below:

name	counts
Gerald	2034
.Hon	1246
eric	1231

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

Answer: Not necessarily because e.g, Gerald has max review_count but has ‘253’ fans while Amy has only ‘609’ reviews but has ‘503’ fans which are approximately double of Gerald. Also, some of them have reviews in 3 digit of number but have 0 fans. Therefore we can’t say that more reviews correlate with more fans.

```
+--------------+-------------+------+
| name      | review_count | fans |
| 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 |
+-----------+-------------+------+
+-----------+-------------+------+
| Gerald    |         2000 |  253 |
| Sara      |         1629 |   50 |
| Yuri      |         1339 |   76 |
|.Hon      |         1246 |  101 |
| William   |         1215 |  126 |
| Harald    |         1153 |  311 |
| eric      |         1116 |   16 |
| Roanna    |         1039 |  104 |
| Mimi      |          968 |  497 |
| Christine |          930 |  173 |
+-----------+-------------+------+
| name          | review_count | fans |
+---------------+-------------+------+
| G             |          359 |    0 |
| gric          |          177 |    0 |
| Uwe           |          122 |    0 |
| Sally         |          108 |    0 |
| Marlene       |          106 |    0 |
| Jason         |          105 |    0 |
| Anand         |          104 |    0 |
```

```
| Inconspicuous |          103 |    0 |
| Ckoka         |          103 |    0 |
| Tara          |           96 |    0 |
+---------------+--------------+------+
```

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

Answer: Yes, there are ‘1780’ review with “love” and ‘232’ review with the word “hate” So, undoubtedly we can say that reviews with the word “love” are more than the word “hate”.

SQL code used to arrive at answer:

// words with “love” review

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

// words with “hate” review

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

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

SQL code used to arrive at the 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?

Answer. According to question, the two groups I choose are “Las Vegas ” and “Shopping”. And, I found that place with low rating like 2.5 closes at 22:00 while shopping places with a High rating of more than or equal to 4.5 closes before 17:00. And, Interesting thing is that they open at the same time(8:00). So, yes groups have different distribution time.

```
+-------------------------------+-----------+----------+-------+---------------------
-+
| name                          | city      | category | stars | hours
|
+-------------------------------+-----------+----------+-------+---------------------
-+
| Walgreens                     | Las Vegas | Shopping |   2.5 | Saturday|8:00-22:00
|
| Wooly Wonders                 | Las Vegas | Shopping |   3.5 | Saturday|10:00-16:00
|
| Red Rock Canyon Visitor Center | Las Vegas | Shopping |   4.5 | Saturday|8:00-16:30
|
| Desert Medical Equipment      | Las Vegas | Shopping |   5.0 | Monday|8:00-17:00
|
+-------------------------------+-----------+----------+-------+---------------------
-+
```

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

Answer: Yes, The group which has less and high rating has less review while groups with average rating have more reviews

```
+-------------------------------+-----------+----------+-------+----------------------+------
--
| name                          | city      | category | stars | hours
|review_count
+-------------------------------+-----------+----------+-------+----------------------+------
---
| Walgreens                     | Las Vegas | Shopping |   2.5 | Saturday|8:00-22:00  |   6
|
| Wooly Wonders                 | Las Vegas | Shopping |   3.5 | Saturday|10:00-16:00 |   11
|
| Red Rock Canyon Visitor Center | Las Vegas | Shopping |   4.5 | Saturday|8:00-16:30  |   32
|
| Desert Medical Equipment      | Las Vegas | Shopping |   5.0 | Monday|8:00-17:00    |   4
|
+-------------------------------+-----------+----------+-------+----------------------+------
--
```

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

Answer: No nothing. Although some have the same postal_code but have different addresses.

SQL code used for analysis:

```
SELECT
business.name,
business.city,
category.category,
business.stars,
hours.hours,
business.review_count,
business.address,
business.postal_code
FROM
        ( business INNER JOIN category ON business.id = category.business_id )
INNER JOIN
                ( hours ON hours.business_id =business.id )
WHERE
                business.city = 'Las Vegas' AND category.category = "Shopping"
GROUP BY
```

business.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: There are more reviews on businesses that are open that the businesses which are closed.

ii. Difference 2: And, the business which is open has more stars than the close businesses.

iii. Difference 3: Also, las vegas has more open and close businesses than any other city.

SQL code used for analysis:

```
SELECT    AVG(b.stars),SUM(b.review_count),AVG(b.review_count),
          COUNT(r.useful)+COUNT(r.funny),is_open
FROM business b INNER JOIN review r ON b.id = r.id
GROUP BY b.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:

Answer: I choose to study preferences among different type of night out places on yelp dataset.

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

Answer: I pick different night out places like Hotels, Restaurants and, Bars. Then I analyze their overall average rating and reviews by category. And, I find out that all 7 Hotels are closed also their rating and reviews were also less. Also, I find out that Bars and Restaurants have an approximately the same average rating but the number of restaurants is more as compared to Bars and most are open. So, I can get the insight which type of night out I can choose.

iii. Output of your finished dataset:

category	no_of_shop	avg(stars)	avg(review_count)	city	sum(is_open)
Hotels	7	1.5	9.0	Oakwood Village	0
Bars	94	3.52659574468	82.4893617021	Peninsula	61
Restaurants	289	3.53979238754	87.0034602076	Chesterland	222

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

```
select
	c.category, count(name) as no_of_shop, avg(stars),    avg(review_count), b.city,
sum(is_open)
from
	( business b inner join hours h on  b.id = h.business_id )
inner join
	( category c on c.business_id = b.id )
where
	c.category in("Hotels","Restaurants","Bars","Clubs")
group by
	c.category
order by
	avg(stars)
```