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 = id

ii. Hours = Business_id
iii. Category = Business_id
iv. Attribute = Business_id
v. Review = id
vi. Checkin = Business_id
vii. Photo = id/Business_id
viii. Tip = user_id/Business_id
ix. User = id
x. Friend = user_id
xi. Elite_years = user_id

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 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
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: 3.7

ii. Table: Business, Column: Stars

min: 1.0 max: 5.0 avg: 3.7

iii. Table: Tip, Column: Likes

min: 0 max: 2 avg: 0.014

iv. Table: Checkin, Column: Count

min: 1 max: 53 avg: 1.94

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_reviews_by_City
from Business
    group by city
    order by Total_reviews_by_City desc
```

Copy and Paste the Result Below:

| city | Total_reviews_by_City |
|---|---|
| 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)

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

i. Avon

SQL code used to arrive at answer:

```
select sum(review_count) as Total,stars from business
where city like '%Avon'
group by stars
```

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

| Total | stars |
|---|---|
| 10 | 1.5 |
| 6 | 2.5 |
| 88 | 3.5 |
| 21 | 4.0 |
| 31 | 4.5 |
| 3 | 5.0 |

ii. Beachwood

SQL code used to arrive at answer:

```
select sum(review_count) as Total,stars from business
where city like '%Beachwood'
group by stars
```

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

| Total | 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 used to arrive at answer:

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

Copy and Paste the Result Below:

| name | sum(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:

| name | sum(review_count) | fans |
|---|---|---|
| Amy | 609 | 503 |
| Mimi | 968 | 497 |
| Harald | 1153 | 311 |
| Gerald | 2000 | 253 |

Not really, by adding then grouping the column ‘Fans’ to the table, you can clearly see that Amy has more fans but lesser reviews than Gerald who has less fans but more reviews so clearly these 2 variables or columns don’t correlate.

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

Answer: yes,1780 of the word ‘Love’ and 232 of the word ‘Hate’ .

SQL code used to arrive at answer:

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

Or

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

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

SQL code used to arrive at answer:

```sql
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.

| city | category | stars | hours | is_open |
|---|---|---|---|---|
| Toronto | Restaurants | 1.5 | None | 1 |
| Toronto | Restaurants | 2.0 | Saturday\|11:00-23:00 | 0 |
| Toronto | Restaurants | 3.0 | Saturday\|10:00-4:00 | 1 |
| Toronto | Restaurants | 3.5 | None | 1 |
| Toronto | Restaurants | 4.0 | Saturday\|18:00-23:00 | 1 |
| Toronto | Restaurants | 4.5 | Saturday\|11:00-23:00 | 1 |

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

I have chosen Toronto as city and restaurant as category and the hours almost do correlate.

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

Yes, the 2-3 stars group have an total reviews of 46 and 4-5 stars group have an total reviews of 97.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
Yes ,It looks like the restaurants are close to the Coast.

SQL code used for analysis:

```
select City,Latitude,Longitude,review_count,Category,stars,hours
from Business b
left join Category c on b.id = c.business_id
left join hours h on  h.business_id = b.id
where city like '%Toronto%' and Category like '%Restaurant%'
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.

i. Difference 1:
looking at open and closed is normalise as an INT ‘1’ OR ‘2’

ii. Difference 2:
The average stars and reviews are correlating somehow.

SQL code used for analysis:

```
select count(is_open) as T_Business,
is_open as is_open_or_closed,
avg(stars),review_count
from business
group by is_open
```

| T_Business | is_open_or_closed | avg(stars) | review_count |
|---|---|---|---|
| 1520 | 0 | 3.52039473684 | 19 |
| 8480 | 1 | 3.67900943396 | 25 |

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:
I would have look at the closed businesses specially the restaurant with high stars and reviews.

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

I will create a temporary view table, name it and gather all possible columns to find insight on why these restaurant with 4-5 stars and enough reviews were shut down.What was the key cause of the closure?

iii. Output of your finished dataset:

| name | city | T_Businesses | Open_Closed | stars | review_count |
|---|---|---|---|---|---|
| Eklectic Pie - Mesa | Mesa | 1 | 0 | 4.0 | 129 |
| The Cider Mill | Scottsdale | 1 | 0 | 4.0 | 91 |
| Nabers Music, Bar & Eats | Chandler | 1 | 0 | 4.0 | 75 |

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

```sql
select b.name,city,count(*) as T_Businesses, is_open as Open_Closed,
b.stars,b.review_count
    from business b
    left join user u on b.id = u.id
    left join category c on b.id = c.business_id
    left join review r on b.id = r.id
    where category like '%Restaurant%' and is_open =='0'
    and b.stars between '4.0' and '5.0'
    group by city
    order by b.review_count desc
    limit 3
```