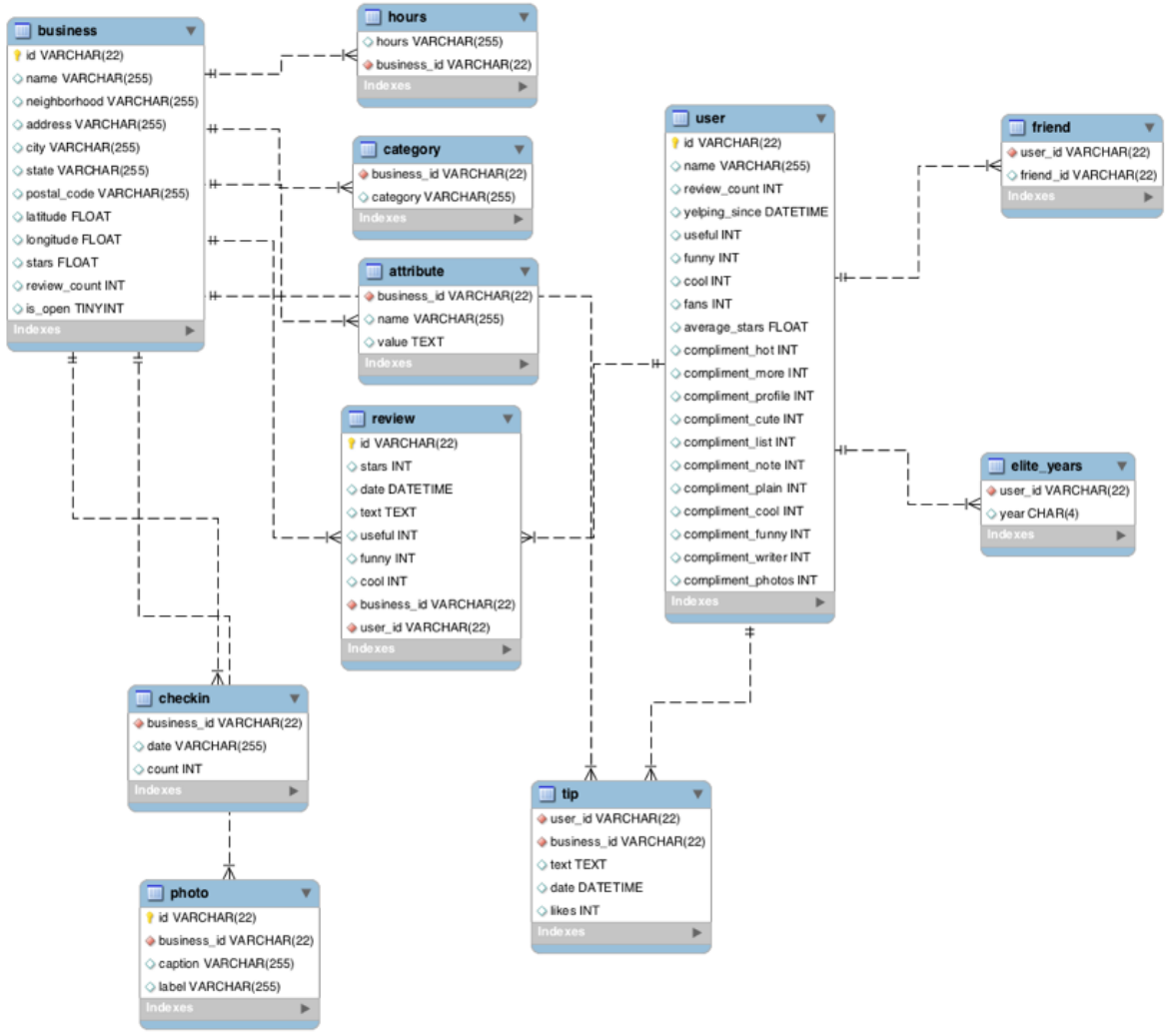
Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset
Coursera Worksheet

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 total of records
ii. Business table = 10000 total of records
iii. Category table = 10000 total of records
iv. Checkin table = 10000 total of records
v. elite_years table = 10000 total of records
vi. friend table = 10000 total of records
vii. hours table = 10000 total of records
viii. photo table = 10000 total of records
ix. review table = 10000 total of records
x. tip table = 10000 total of records
xi. user table = 10000 total of records

CODE:

```
SELECT *

FROM user


--This same query was used for each table asked, just changing the talbe
--name   in FROM  statement. The COUNT function wasn´t
--used because this would have led to a possible incorrect number of records
-- due to the exclusion of null values
```

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 = 10000 total of distinct records

```
SELECT COUNT(DISTINCT(id))
FROM business
```

ii. Hours = 1562 total of distinct records

```
SELECT COUNT(DISTINCT(business_id))
FROM hours
```

iii. Category = 2643 total of distinct records

```
SELECT COUNT(DISTINCT(business_id))
FROM category
```

iv. Attribute = 1115 total of distinct records

```
SELECT COUNT(DISTINCT(business_id))
FROM attribute
```

v. Review = 10000 total of distinct records

```
SELECT COUNT(DISTINCT(id))
FROM review
```

vi. Checkin = 493 total of distinct records

```
SELECT COUNT(DISTINCT(business_id))
FROM checkin
```

vii. Photo = 10000 total of distinct records

```
SELECT COUNT(DISTINCT(id))
FROM photo
```

viii. Tip = 3979 total of distinct records

```
SELECT COUNT(DISTINCT(business_id))
FROM Tip
```

ix. User = 10000 total of distinct records

```
SELECT COUNT(DISTINCT(id))
FROM user
```

x. Friend = 11

```
SELECT COUNT(DISTINCT(user_id))
FROM friend
```

xi. Elite_years = 2780

```
SELECT COUNT(DISTINCT(user_id))
FROM Elite_years
```

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_writer is null or compliment_photos is null
--this same query was used for each column name in the table,
```

```
--just changing the column name in
--the WHERE  statement.
```

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 MIN(Review_count), MAX(Review_count), AVG(Review_count)
FROM user
--Same code for the other tables and columns,
--just modifying names of tables and columns
```

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

SQL code used to arrive at answer:

```
SELECT CITY,

       (SELECT COUNT (id)
       FROM review) as num_reviews
FROM business
ORDER BY num_reviews DESC
```

Copy and Paste the Result Below:

city	num_reviews
Richmond Hill	10000
Huntersville	10000
Gilbert	10000
Las Vegas	10000
Tempe	10000
Tempe	10000
Pittsburgh	10000
Pittsburgh	10000
Charlotte	10000
Toronto	10000
Las Vegas	10000
Las Vegas	10000
Charlotte	10000
Henderson	10000
Gilbert	10000
Phoenix	10000
Canonsburg	10000
Bay Village	10000
Streetsboro	10000
Charlotte	10000
Charlotte	10000
Toronto	10000
Scottsdale	10000
Henderson	10000
Edinburgh	10000

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

i. Avon

SQL code used to arrive at answer:

```
SELECT
    name,
    stars,
    review_count
FROM business
WHERE city = 'Avon'
```

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

name	stars	review_count
Helen & Kal's	2.5	3
Marc's	4.0	4
Hoban Pest Control	5.0	3

Light Salon & Spa	3.5	7
Portrait Innovations	1.5	10
Winking Lizard Tavern	3.5	31
Dervish Mediterranean & Turkish Grill	4.5	31
Mulligans Pub and Grill	3.5	50
Mr. Handyman of Cleveland's Northwest Suburbs	2.5	3
Cambria hotel & suites Avon - Cleveland	4.0	17

+ii. Beachwood

SQL code used to arrive at answer:

```
SELECT
    name,
    stars,
    review_count
FROM business
WHERE city = 'Beachwood'
```

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

name	stars	review_count
Maltz Museum of Jewish Heritage	3.0	8
Charley's Grilled Subs	3.0	3
Sixth & Pine	4.5	14
Beechmont Country Club	5.0	6
Hyde Park Prime Steakhouse	4.0	69
Origins	4.5	3
Fyodor Bridal Atelier	5.0	4
College Planning Network	2.0	8
Lucky Brand Jeans	3.5	3
American Eagle Outfitters	3.5	3
Shaker Women's Wellness	5.0	6
Avis Rent A Car	2.5	3
Cleveland Acupuncture	5.0	3
Studio Mz	5.0	4

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

SQL code used to arrive at answer:

```
SELECT id, name, review_count

FROM user
ORDER BY review_count DESC
LIMIT 3
```

Copy and Paste the Result Below:

```
+------------------------+--------+--------------+
| id                     | name   | review_count |
+------------------------+--------+--------------+
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald |         2000 |
| -3s52C4zL_DHRK0ULG6qtg | Sara   |         1629 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri   |         1339 |
+------------------------+--------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

It does to some extent. The correlation is measured between -1 and 1, if the value is positive it means there is a positive correlation, i.e, when one variable increases the other one does it as well. If the value of correlation is zero, that means the two variables don´t correlate at all. For this scenario correlation was 0.4371, which means the more reviews the more fans. The code used for this was:

```
select avg( (review_count - avg_x) * (fans - avg_y) )*avg( (revi
ew_count - avg_x) * (fans - avg_y) )/(var_x*var_y) as R2
from user, (select
      avg_x,
      avg_y,
      avg((review_count - avg_x)*(review_count - avg_x)) as var_
x,
      avg((fans - avg_y)*(fans - avg_y)) as var_y
      from user, (select
          avg(review_count) as avg_x,
          avg(fans) as avg_y
          from user)
  );
```

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

Answer: 232 hate, 1780 love, then there are more reviews with the word love than with the word hate

SQL code used to arrive at answer:

```
SELECT COUNT (id)
FROM review
WHERE text like '%hate%'

SELECT COUNT (id)
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, 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?

Yes, they have different distribution hours but the difference is not very conclusive. This is because one of the businesses with high rating opens every day in the evening until late at night. The other business with high rating opens six days a week and is the business with less hours open. In contrast the business with a low rating opens for more hours than both of the two high rated businesses but still has low stars.

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

Yes, in general the higher the rating for business, the more that business has.

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

The neighborhood and postal code are insufficient to provide any analysis that can find a relationship between location and rating. Further information would be needed to accomplish this objective.

SQL code used for analysis:

```
SELECT
       CASE
       WHEN stars>=4 THEN '4-5 Stars'
       WHEN (stars>=2 AND stars<=3.9) THEN '2-3 Stars'
       END as rating,

       b.postal_code,
       b.review_count,
       h.hours,
       b.name,
       b.neighborhood
FROM business as b INNER JOIN category as c
     ON b.id=c.business_id

     INNER JOIN hours as h
     ON b.id=h.business_id
WHERE city='Toronto'
      AND category = 'Food'
      AND (stars>=4 OR (stars <=3.9 and stars>=2))
ORDER BY stars DESC, hours DESC
```

rating	postal_code	review_count	hours	name	neighborhood
4-5 Stars	M6P 1A6	26	Wednesday|18:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Tuesday|18:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Thursday|18:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Sunday|16:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Saturday|16:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Monday|16:00-2:00	Cabin Fever	High Park
4-5 Stars	M6P 1A6	26	Friday|18:00-2:00	Cabin Fever	High Park
4-5 Stars	M6H 1V5	15	Wednesday|15:00-21:00	Halo Brewery	Wallace Emerson
4-5 Stars	M6H 1V5	15	Tuesday|15:00-21:00	Halo Brewery	Wallace Emerson
4-5 Stars	M6H 1V5	15	Thursday|15:00-21:00	Halo Brewery	Wallace Emerson
4-5 Stars	M6H 1V5	15	Sunday|11:00-21:00	Halo Brewery	Wallace Emerson

4-5 Stars	M6H 1V5	15	Saturday|11:00-21:00	Halo Brewery	Wallace Emerson
4-5 Stars	M6H 1V5	15	Friday|15:00-21:00	Halo Brewery	Wallace Emerson
2-3 Stars	M6R 1X3	10	Wednesday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Tuesday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Thursday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Sunday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Saturday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Monday|8:00-22:00	Loblaws	Roncesvalles
2-3 Stars	M6R 1X3	10	Friday|8:00-22:00	Loblaws	Roncesvalles

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 about 5 times more businesses opened than closed.

Open Businesses: 8480
Closed Businesses: 1520

ii. Difference 2:

Opened businesses receive 37 % more reviews than closed businesses

Open Businesses reviews: 269300
Closed Businesses reviews: 35261

SQL code used for analysis:

```
SELECT
    COUNT(DISTINCT(id)) as tot_business,
    ROUND(AVG(review_count),1) as avg_review,
    SUM(review_count) as tot_reviews, ROUND(AVG(stars),1) as
avg_stars, is_open,
    (SELECT COUNT(DISTINCT(id)) as total_of_business
    FROM business
    WHERE is_open = 1)/
    (SELECT COUNT(DISTINCT(id)) as total_of_business
    FROM business
    WHERE is_open = 0) as times,
    ROUND((
        (SELECT ROUND(AVG(review_count),1)
        FROM business
```

```
                    WHERE is_open=1) -
                    (SELECT ROUND(AVG(review_count),1)
                    FROM business
                    WHERE is_open=0)
                    )
                    /
                    (SELECT ROUND(AVG(review_count),1)
                    FROM business
                    WHERE is_open=0) * 100,2) as percent_dif
            FROM business
            GROUP BY is_open
```

```
+--------------+------------+-------------+-----------+---------+-------+-------------+
| tot_business | avg_review | tot_reviews | avg_stars | is_open | times | percent_dif |
+--------------+------------+-------------+-----------+---------+-------+-------------+
|         1520 |       23.2 |       35261 |       3.5 |       0 |     5 |       37.07 |
|         8480 |       31.8 |      269300 |       3.7 |       1 |     5 |       37.07 |
+--------------+------------+-------------+-----------+---------+-------+-------------
+
```

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 used the category table and the business table to determine if the category is related to the number of reviews a type of business receives.

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

For my analysis I need:

-The total number of businesses per category
-The average number of reviews per category of business
-The maximum and the minimum number of reviews each type of business receives.

This data will help discovering if there is correlation between the category and the number of reviews.

I found that the more general the name of the category is the more reviews it gets. Also, The top 5 businesses that have more competence are businesses related to social activities, such as restaurants, shopping, nightlife and bars.

iii. Output of your finished dataset:

category	num_business	avg_reviews	max_reviews	min_reviews
Restaurants	71	63.44	768	3
Shopping	30	32.57	723	3
Food	23	77.43	723	3
Nightlife	20	67.55	431	3
Bars	17	77.76	431	3
Health & Medical	17	11.94	30	4
Home Services	16	5.88	14	3
Beauty & Spas	13	9.15	27	3
Local Services	12	8.33	32	3
American (Traditional)	11	102.55	431	3
Active Life	10	13.1	32	3
Automotive	9	22.0	63	3
Hotels & Travel	9	42.33	223	3
Burgers	8	37.13	94	3
Sandwiches	8	121.75	361	3
Arts & Entertainment	7	55.43	223	6
Fast Food	7	26.43	83	3
Mexican	7	46.71	103	4
American (New)	6	80.17	168	3
Event Planning & Services	6	19.67	69	4
Hair Salons	6	10.83	27	3
Bakeries	5	47.8	162	5
Doctors	5	11.0	18	5
Indian	5	12.6	32	3
Japanese	5	30.4	75	3

(Output limit exceeded, 25 of 257 total rows shown)

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

```
SELECT DISTINCT(c.category),
COUNT(business_id) as num_business,
ROUND(AVG (review_count),2) as avg_reviews,
MAX(review_count) as max_reviews,
MIN(review_count) as min_reviews
FROM category as c
JOIN business as b ON c.business_id = b.id
GROUP BY category
ORDER BY --max_reviews DESC,
num_business DESC
```