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
```
SELECT COUNT(*)FROM attribute
```

ii. Business table = 10000
```
SELECT COUNT(*) FROM business
```

iii. Category table = 10000
```
SELECT COUNT(*) FROM category
```

iv. Checkin table = 10000
```
SELECT COUNT(*) FROM checkin
```

v. elite_years table = 10000
```
SELECT COUNT(*) FROM elite_years
```

vi. friend table = 10000
```
SELECT COUNT (*) FROM friend
```

vii. hours table = 10000
```
SELECT COUNT(*) FROM hours
```

viii. photo table = 10000
```
SELECT COUNT(*) FROM photo
```

ix. review table = 10000
```
SELECT COUNT(*) FROM review
```

x. tip table = 10000
```
SELECT COUNT(*) FROM tip
```

xi. user table = 10000
```
SELECT COUNT(*) FROM user
```

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

ii. Hours = 10000
```
SELECT COUNT(DISTINCT business_id) FROM hours;
```

iii. Category = 10000
```
SELECT COUNT(DISTINCT business_id) FROM category;
```

iv. Attribute = 10000
```
SELECT COUNT(DISTINCT business_id) FROM attribute;
```

v. Review = 10000
```
SELECT COUNT(DISTINCT id) FROM review;
```

vi. Checkin = 10000
```
SELECT COUNT(DISTINCT business_id) FROM checkin;
```

vii. Photo = 10000
```
SELECT COUNT(DISTINCT id) FROM photo;
```

viii. Tip = 537(user_id)
```
SELECT COUNT(DISTINCT user_id) FROM tip;
```

ix. User = 10000
```
SELECT COUNT(DISTINCT id) FROM user;
```

x. Friend = 11
```
SELECT COUNT(DISTINCT user_id) FROM friend;
```

xi. Elite_years =
```
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 DISTINCT COUNT(*)
FROM user
WHERE (name OR review_count OR yelping_since OR useful OR funny
OR cool OR fans OR average_stars OR compliment_hot OR
compliment_more OR compliment_profile OR compliment_cute OR
compliment_list OR compliment_note OR compliment_plain OR
compliment_cool OR compliment_funny OR compliment_writer 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: 3.7082

```
SELECT MIN(stars), MAX(stars) ,AVG(stars)
FROM review;
```

ii. Table: Business, Column: Stars

min: 1 max: 5 avg: 3.6549

```
SELECT MIN(stars), MAX(stars) ,AVG(stars)
FROM business;
```

iii. Table: Tip, Column: Likes

min: 0 max: 2 avg: 0.0144

```
SELECT MIN(likes), MAX(likes) ,AVG(likes)
FROM tip;
```

iv. Table: Checkin, Column: Count

min: 1 max: 53 avg: 1.9414

```
SELECT MIN(count), MAX(count) ,AVG(count)
FROM checkin;
```

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;
```

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

SQL code used to arrive at answer:

```
SELECT COUNT(review_count) as total_reviews, city
FROM business
GROUP BY city
ORDER BY total_reviews DESC
```

Copy and Paste the Result Below:

total_reviews	city
1561	Las Vegas
1001	Phoenix
985	Toronto
497	Scottsdale
468	Charlotte
353	Pittsburgh
337	Montréal
304	Mesa
274	Henderson
261	Tempe
239	Edinburgh
232	Chandler
189	Cleveland
188	Gilbert
188	Glendale
176	Madison
150	Mississauga
141	Stuttgart
105	Peoria
80	Markham
71	Champaign
70	North Las Vegas
64	North York
60	Surprise
54	Richmond Hill

(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, review_count
FROM business
WHERE city = 'Avon'
```

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

stars	review_count
2.5	3
4.0	4
5.0	3
3.5	7
1.5	10

```
|   3.5 |           31 |
|   4.5 |           31 |
|   3.5 |           50 |
|   2.5 |            3 |
|   4.0 |           17 |
+-------+--------------+
```

ii. Beachwood

SQL code used to arrive at answer:

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

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

```
+-------+--------------+
| stars | review_count |
+-------+--------------+
|   3.0 |            8 |
|   3.0 |            3 |
|   4.5 |           14 |
|   5.0 |            6 |
|   4.0 |           69 |
|   4.5 |            3 |
|   5.0 |            4 |
|   2.0 |            8 |
|   3.5 |            3 |
|   3.5 |            3 |
|   5.0 |            6 |
|   2.5 |            3 |
|   5.0 |            3 |
|   5.0 |            4 |
+-------+--------------+
```

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

SQL code used to arrive at answer:

```
SELECT name, id, review_count
FROM user
ORDER BY review_count DESC;
```

Copy and Paste the Result Below:

```
+-----------+------------------------+--------------+
| name      | id                     | review_count |
+-----------+------------------------+--------------+
| Gerald    | -G7Zkl1wIWBBmD0KRy_sCw |         2000 |
| Sara      | -3s52C4zL_DHRK0ULG6qtg |         1629 |
| Yuri      | -81bUNlXVSoXqaRRiHiSNg |         1339 |
| .Hon      | -K2Tcgh2EKX6e6HqqIrBIQ |         1246 |
| William   | -FZBTkAZEXoP7CYvRV2ZwQ |         1215 |
| Harald    | --2vR0DIsmQ6WfcSzKWigw |         1153 |
| eric      | -gokwePdbXjfS0iF7NsUGA |         1116 |
| Roanna    | -DFCC64NXgqrxlO8aLU5rg |         1039 |
| Mimi      | -8EnCioUmDygAbsYZmTeRQ |          968 |
| Christine | -0IiMAZI2SsQ7VmyzJjokQ |          930 |
| Ed        | -fUARDNuXAfrOn4WLSZLgA |          904 |
| Nicole    | -hKniZN2OdshWLHYuj21jQ |          864 |
```

Fran	-9da1xk7zgnnfO1uTVYGkA	862
Mark	-B-QEUESGWHPE_889WJaeg	861
Christina	-kLVfaJytOJY2-QdQoCcNQ	842
Dominic	-kO6984fXByyZm3_6z2JYg	836
Lissa	-1h59ko3dxChBSZ9U7LfUw	834
Lisa	-g3XIcCb2b-BD0QBCcq2Sw	813
Alison	-l9giG8TSDBG1jnUBUXp5w	775
Sui	-dw8f7FLaUmWR7bfJ_Yf0w	754
Tim	-AaBjWJYiQxXkCMDlXfPGw	702
L	-jt1ACMiZljnBFvS6RRvnA	696
Angela	-IgKkE8JvYNWeGu8ze4P8Q	694
Crissy	-hxUwfo3cMnLTv-CAaP69A	676
Lyn	-H6cTbVxeIRYR-atxdielQ	675

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

Posing more reviews does not have a correlation with the number of fans. This is represented very clearly in table below, which shows how the person with the highest number of fans has half as many reviews but twice the number of fans of the person with the highest number of fans.

name	id	review_count	fans
Gerald	-G7Zkl1wIWBBmD0KRy_sCw	2000	253
Sara	-3s52C4zL_DHRK0ULG6qtg	1629	50
Yuri	-8lbUNlXVSoXqaRRiHiSNg	1339	76
.Hon	-K2Tcgh2EKX6e6HqqIrBIQ	1246	101
William	-FZBTkAZEXoP7CYvRV2ZwQ	1215	126
Harald	--2vR0DIsmQ6WfcSzKWigw	1153	311
eric	-gokwePdbXjfS0iF7NsUGA	1116	16
Roanna	-DFCC64NXgqrxlO8aLU5rg	1039	104
Mimi	-8EnCioUmDygAbsYZmTeRQ	968	497
Christine	-0IiMAZI2SsQ7VmyzJjokQ	930	173
Ed	-fUARDNuXAfrOn4WLSZLgA	904	38
Nicole	-hKniZN2OdshWLHYuj21jQ	864	43
Fran	-9da1xk7zgnnfO1uTVYGkA	862	124
Mark	-B-QEUESGWHPE_889WJaeg	861	115
Christina	-kLVfaJytOJY2-QdQoCcNQ	842	85
Dominic	-kO6984fXByyZm3_6z2JYg	836	37
Lissa	-1h59ko3dxChBSZ9U7LfUw	834	120
Lisa	-g3XIcCb2b-BD0QBCcq2Sw	813	159
Alison	-l9giG8TSDBG1jnUBUXp5w	775	61
Sui	-dw8f7FLaUmWR7bfJ_Yf0w	754	78
Tim	-AaBjWJYiQxXkCMDlXfPGw	702	35
L	-jt1ACMiZljnBFvS6RRvnA	696	10
Angela	-IgKkE8JvYNWeGu8ze4P8Q	694	101
Crissy	-hxUwfo3cMnLTv-CAaP69A	676	25
Lyn	-H6cTbVxeIRYR-atxdielQ	675	45

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

```
SELECT name, id, review_count, fans
FROM user
ORDER BY review_count DESC;
```

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

Answer:
There are more reviews with the word “love” (1780) than with the word %hate% (232).

SQL code used to arrive at answer:

```
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 used to arrive at answer:

```
SELECT name, id, fans
FROM user
ORDER BY fans DESC;
```

Copy and Paste the Result Below:

name	id	fans
Amy	-9I98YbNQnLdAmcYfb324Q	503
Mimi	-8EnCioUmDygAbsYZmTeRQ	497
Harald	--2vR0DIsmQ6WfcSzKWigw	311
Gerald	-G7Zkl1wIWBBmD0KRy_sCw	253
Christine	-0IiMAZI2SsQ7VmyzJjokQ	173
Lisa	-g3XIcCb2b-BD0QBCcq2Sw	159
Cat	-9bbDysuiWeo2VShFJJtcw	133
William	-FZBTkAZEXoP7CYvRV2ZwQ	126
Fran	-9da1xk7zgnnfO1uTVYGkA	124
Lissa	-lh59ko3dxChBSZ9U7LfUw	120
Mark	-B-QEUESGWHPE_889WJaeg	115
Tiffany	-DmqnhW4Omr3YhmnigaqHg	111
bernice	-cv9PPT7IHux7XUc9dOpkg	105
Roanna	-DFCC64NXgqrxlO8aLU5rg	104
Angela	-IgKkE8JvYNWeGu8ze4P8Q	101
.Hon	-K2Tcgh2EKX6e6HqqIrBIQ	101
Ben	-4viTt9UC44lWCFJwleMNQ	96
Linda	-3i9bhfvrM3F1wsC9XIB8g	89
Christina	-kLVfaJytOJY2-QdQoCcNQ	85
Jessica	-ePh4Prox7ZXnEBNGKyUEA	84
Greg	-4BEUkLvHQntN6qPfKJP2w	81
Nieves	-C-l8EHSLXtZZVfUAUhsPA	80
Sui	-dw8f7FLaUmWR7bfJ_Yf0w	78
Yuri	-8lbUNlXVSoXqaRRiHiSNg	76
Nicole	-0zEEaDFIjABtPQni0XlHA	73

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

11. Is there a strong relationship (or correlation) between having a high number of fans and being listed as "useful" or "funny?" Out of the top 10 users with the highest number of fans, what percent are also listed as “useful” or “funny”?

Key:
0% - 25% - Low relationship
26% - 75% - Medium relationship
76% - 100% - Strong relationship

SQL code used to arrive at answer:

```
SELECT name, id, fans, useful, funny
FROM user
ORDER BY fans DESC;
```

Copy and Paste the Result Below:

name	id	fans	useful	funny
Amy	-9I98YbNQnLdAmcYfb324Q	503	3226	2554
Mimi	-8EnCioUmDygAbsYZmTeRQ	497	257	138
Harald	--2vR0DIsmQ6WfcSzKWigw	311	122921	122419
Gerald	-G7Zkl1wIWBBmD0KRy_sCw	253	17524	2324
Christine	-0IiMAZI2SsQ7VmyzJjokQ	173	4834	6646
Lisa	-g3XIcCb2b-BD0QBCcq2Sw	159	48	13
Cat	-9bbDysuiWeo2VShFJJtcw	133	1062	672
William	-FZBTkAZEXoP7CYvRV2ZwQ	126	9363	9361
Fran	-9da1xk7zgnnfO1uTVYGkA	124	9851	7606
Lissa	-1h59ko3dxChBSZ9U7LfUw	120	455	150
Mark	-B-QEUESGWHPE_889WJaeg	115	4008	570
Tiffany	-DmqnhW4Omr3YhmnigaqHg	111	1366	984
bernice	-cv9PPT7IHux7XUc9dOpkg	105	120	112
Roanna	-DFCC64NXgqrxlO8aLU5rg	104	2995	1188
Angela	-IgKkE8JvYNWeGu8ze4P8Q	101	158	164
.Hon	-K2Tcgh2EKX6e6HqqIrBIQ	101	7850	5851
Ben	-4viTt9UC44lWCFJwleMNQ	96	1180	1155
Linda	-3i9bhfvrM3F1wsC9XIB8g	89	3177	2736
Christina	-kLVfaJytOJY2-QdQoCcNQ	85	158	34
Jessica	-ePh4Prox7ZXnEBNGKyUEA	84	2161	2091
Greg	-4BEUkLvHQntN6qPfKJP2w	81	820	753
Nieves	-C-l8EHSLXtZZVfUAUhsPA	80	1091	774
Sui	-dw8f7FLaUmWR7bfJ_Yf0w	78	9	18
Yuri	-8lbUNlXVSoXqaRRiHiSNg	76	1166	220
Nicole	-0zEEaDFIjABtPQni0XlHA	73	13	10

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

Please explain your findings and interpretation of the results:

Based on the table above sorting the users based on their number of fans does not show descending or ascending trend in “useful” or “funny” columns. Therefore, there should not be a strong correlation between having a high number of fans and being listed as "useful" or "funny”.

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: Mesa Category: Food

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

Yes

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

Yes

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

Based on the results, we can see that there seems to be a correlation between the location of the business and their rating. The business that are probably located in the same neighbor have close rating. Also they have similar working hours. Moreover, the business that have longer working hours usually have higher rating.

SQL code used for analysis:

```
SELECT business.name, business.city, category.category ,
business.stars,
hours.hours, business.review_count, business.postal_code
FROM (business INNER JOIN category ON business.id =
category.business_id) INNER
JOIN hours ON hours.business_id = category.business_id
WHERE business.city = 'Mesa'
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:

The business that are still open have higher rating.

ii. Difference 2:

The business that are still open have more reviews.

SQL code used for analysis:

```
SELECT business.name , business.is_open , category.category,
business.stars,
```

```
hours.hours, business.review_count
FROM (business INNER JOIN category ON business.id =
category.business_id) INNER
JOIN hours ON hours.business_id = category.business_id
WHERE business.city = 'Mesa'
GROUP BY business.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:

Finding correlation between the likes with the given rates and using “like” in the 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 need two sources of data (tables). First, I join these two tables based on users and business. Then I sort them based on rating to see if there is a correlation between the number of stars and likes.
The reason I chose this analysis and thus, the data sets is that psychologists have shown that how people think about something can completely change even after a few minutes and they think that how people think just after occurrence of an event is a better representative for the quality of that event compared to what they say after thinking about it. Because tip table is related to the occurrence of the event (shopping) and they write a review after hours or even days, comparing these two tables can help us to explore the validity what psychologists claim. As the result shows there is a slight correlation between the number of likes and stars, but this correlation is not strong. So what psychologists claim seems to be fairly valid.

iii. Output of your finished dataset:

stars	likes
3	2
5	2
5	1
5	1
5	1
5	1
5	1
5	1

```
|     5 |     1 |
|     5 |     1 |
|     3 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
|     4 |     1 |
+-------+-------+
(Output limit exceeded, 25 of 1227 total rows shown)
```

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

```
SELECT review.stars, tip.likes
FROM review INNER JOIN tip ON review.user_id = tip.user_id
ORDER BY tip.likes DESC;
```