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 'Attribute', COUNT(\*) FROM Attribute;

SELECT 'Business', COUNT(\*) FROM Business;

SELECT 'Category', COUNT(\*) FROM Category;

SELECT 'Checkin', COUNT(\*) FROM Checkin;

SELECT 'elite\_years', COUNT(\*) FROM elite\_years;

SELECT 'friend ', COUNT(\*) FROM friend;

SELECT 'hours', COUNT(\*) FROM hours;

SELECT 'photo', COUNT(\*) FROM photo ;

SELECT 'review', COUNT(\*) FROM review;

SELECT 'tip', COUNT(\*) FROM tip;

SELECT 'user', 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 = id - 10000

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 , business\_id - 6493

viii. Tip = business\_id -, user\_id - 3979

ix. User = id - 10000

x. Friend = user\_id - 11, friend\_id - 9415

xi. Elite\_years = user\_id - 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 coalesce(id, name, review\_count, yelping\_since, useful,funny, cool, fans, average\_stars,compliment\_hot, compliment\_more, compliment\_profile, compliment\_cute, compliment\_list, compliment\_note, compliment\_plain, compliment\_cool, compliment\_funny, compliment\_writer, 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

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

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

SQL code used to arrive at answer:

select city, sum(review\_count) as Reviews

from business

group by city

order by Reviews desc

Copy and Paste the Result Below:

city | Reviews |

+-----------------+---------+

| 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 stars, sum(review\_count) as Reviews

from business

where city = 'Avon'

group by stars

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

stars | Reviews |

+-------+---------+

| 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 answer:

select stars, sum(review\_count) as Reviews

from business

where city = 'Beachwood'

group by stars

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

group by id

order by review\_count desc

limit 3

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

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:

From the result set, its clear that there is no correlation between the reviews and fans.

SQL code used to arrive at answer:

select name, review\_count, fans, yelping\_since

from user

order by fans desc

limit 3

Copy and Paste the Result Below:

| name | review\_count | fans | yelping\_since |

+--------+--------------+------+---------------------+

| Amy | 609 | 503 | 2007-07-19 00:00:00 |

| Mimi | 968 | 497 | 2011-03-30 00:00:00 |

| Harald | 1153 | 311 | 2012-11-27 00:00:00 |

+--------+--------------+------+---------------------

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

Answer: Love came up with more results of 1780 compared to hate which only had 232 occurrences.

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, fans

from user

group by id

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?

Businesses with 4-5 stars seems to have shorter hours than the businesses with 2-3 stars

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

Yes, the 4-5 star rating has both low and high review counts but the 2-3 star rating businesses constantly fall in the lower rating category

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

Every Business is in a different zip code.

SQL code used for analysis:

SELECT B.name,

               B.review\_count,

               H.hours,

               postal\_code,

               CASE

                  WHEN hours LIKE "%monday%" THEN 1

                  WHEN hours LIKE "%tuesday%" THEN 2

                  WHEN hours LIKE "%wednesday%" THEN 3

                  WHEN hours LIKE "%thursday%" THEN 4

                  WHEN hours LIKE "%friday%" THEN 5

                  WHEN hours LIKE "%saturday%" THEN 6

                  WHEN hours LIKE "%sunday%" THEN 7

               END AS ord,

               CASE

                  WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'

                  WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'

               END AS star\_rating

        FROM business B INNER JOIN hours H

        ON B.id = H.business\_id

        INNER JOIN category C

        ON C.business\_id = B.id

        WHERE (B.city == 'Toronto' AND C.category LIKE 'Food')

        AND

        (B.stars BETWEEN 2 AND 3 OR B.stars BETWEEN 4 AND 5)

        GROUP BY stars,ord

        ORDER BY ord,star\_rating ASC

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.

1. Difference 1:

Open: Avg(review\_count) = 31.75

Closed: Avg(review\_count) = 23.19

ii. Difference 2:

Open: Avg(stars) = 3.679

Closed: Avg(stars) = 3.52

SQL code used for analysis:

SELECT COUNT(DISTINCT(id)),

               AVG(review\_count),

               SUM(review\_count),

               AVG(stars),

               is\_open

        FROM business

        GROUP BY 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:

I would like to analyze the preferences for the type of cuisine on Yelp.

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

I will pick several types of food and I will analyze their star ratings and number of reviews to get some insights on which type of food is popular on yelp.

1. Output of your finished dataset:

+----------+-----------------+-----------+-------------+-----------+

| category | Num\_Restaurants | Avg\_Stars | Avg\_Reviews | city |

+----------+-----------------+-----------+-------------+-----------+

| Korean | 7 | 4.5 | 8.0 | Toronto |

| French | 12 | 4.0 | 135.08 | Las Vegas |

| Chinese | 13 | 3.77 | 423.23 | Las Vegas |

| Mexican | 28 | 3.63 | 73.0 | Edinburgh |

| Italian | 13 | 3.54 | 78.23 | Montréal |

| Indian | 6 | 3.5 | 32.0 | Aurora |

| Japanese | 20 | 3.48 | 22.85 | Toronto |

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

SELECT c.category, COUNT(b.name) AS Num\_Restaurants, round(AVG(stars),2) as Avg\_Stars,

round(AVG(review\_count),2) as Avg\_Reviews, b.city

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

("Chinese","Mexican","French","Italian","Korean","Japanese","Indian")

GROUP BY c.category

ORDER BY AVG(stars) DESC