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 = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 537

ix. User = 10000

x. Friend = 11

xi. Elite\_years = 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 there are not any Null values in the database’s table as per the below queried SQL.

SQL code used to arrive at answer:

select 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

from user

where id = NULL or name = NULL or review\_count = NULL or yelping\_since = NULL

or useful = NULL or funny = NULL or cool = NULL or fans= NULL or average\_stars= NULL or compliment\_hot= NULL or compliment\_more= NULL or compliment\_profile=NULL or compliment\_cute= NULL or compliment\_list= NULL or compliment\_note= NULL or compliment\_plain = NULL or compliment\_cool= NULL or compliment\_funny= NULL or compliment\_writer= NULL or compliment\_photos= 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, review\_count

from business

order by review\_count desc;

Copy and Paste the Result Below:

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

| city | review\_count |

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

| Las Vegas | 3873 |

| Montréal | 1757 |

| Gilbert | 1549 |

| Las Vegas | 1410 |

| Las Vegas | 1389 |

| Las Vegas | 1252 |

| Las Vegas | 1116 |

| Las Vegas | 1084 |

| Las Vegas | 961 |

| Gilbert | 902 |

| Las Vegas | 864 |

| Scottsdale | 823 |

| Las Vegas | 821 |

| Las Vegas | 786 |

| Henderson | 785 |

| Toronto | 778 |

| Las Vegas | 768 |

| Las Vegas | 758 |

| Scottsdale | 726 |

| Cleveland | 723 |

| Las Vegas | 720 |

| Charlotte | 715 |

| Phoenix | 711 |

| Las Vegas | 706 |

| Phoenix | 700 |

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

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 count

FROM business

WHERE city = 'Avon'

GROUP BY stars

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

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

| stars | count |

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

| 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 count

FROM business

WHERE city = 'Beachwood'

GROUP BY stars

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

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

| stars | count |

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

| 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 answer:

select name, review\_count

from user

order by review\_count desc

Copy and Paste the Result Below:

| name | 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:

No, it does not mean that posting more reviews correlate with more fans. As now below this I am pasting the data to prove my conclusion.

select name, review\_count, fans

from user

order by review\_count desc;

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

| name | review\_count | fans |

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

| 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 |

| Ed | 904 | 38 |

| Nicole | 864 | 43 |

| Fran | 862 | 124 |

| Mark | 861 | 115 |

| Christina | 842 | 85 |

| Dominic | 836 | 37 |

| Lissa | 834 | 120 |

| Lisa | 813 | 159 |

| Alison | 775 | 61 |

| Sui | 754 | 78 |

| Tim | 702 | 35 |

| L | 696 | 10 |

| Angela | 694 | 101 |

| Crissy | 676 | 25 |

| Lyn | 675 | 45 |

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

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” as compared to “hate”.

SQL code used to arrive at answer:

select count(\*)as loves

from review

where text like "%love%"

+-------+

| loves |

+-------+

| 1780 |

+-------+

select count(\*)as hates

from review

where text like "%hate%"

+-------+

| hates |

+-------+

| 232 |

+-------+

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

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 do have different distribution of number of hours.

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

Yes, both the groups had different number of reviews. Below I have made table to conclude my answer.

SELECT stars, sum(review\_count)

FROM business

group by stars

having stars between 2 and 3

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

| stars | sum(review\_count) |

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

| 2.0 | 7820 |

| 2.5 | 16808 | (for stars 2-3)

| 3.0 | 36295 |

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

SELECT stars, sum(review\_count)

FROM business

group by stars

having stars between 4 and 5

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

| stars | sum(review\_count) |

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

| 4.0 | 97044 |

| 4.5 | 53938 | (for stars 4-5)

| 5.0 | 17708 |

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

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

No not really. They have different locations. Therefore cannot figure out anything.

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 = 'Toronto' AND category.category = "Food"

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 businesses that are still open have more stars on an average as compared to businesses which are shut

SQL code used for analysis:

SELECT avg(stars), is\_open

FROM business

group by is\_open

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

| avg(stars) | is\_open |

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

| 3.52039473684 | 0 |

| 3.67900943396 | 1 |

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

ii. Difference 2: Businesses which are open are getting more reviews than businesses which are shut.

SQL code used for analysis:

SELECT sum(review\_count), is\_open

FROM business

group by is\_open

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

| sum(review\_count) | is\_open |

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

| 35261 | 0 |

| 269300 | 1 |

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

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:

To analyze what people like the most to eat between Indian, Japanese, Chinese or Mexican dishes/food.

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

As I want to differentiate between which food ppl like the most between Indian, Japanese, Chinese or Mexican.

I would need to have the reviews column, stars column, and I would be redefining my on rating as I have stated below in my code to compare between those foods. And ofc the category table.

iii. Output of your finished dataset:

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

| sum(stars) | sum(review\_count) | ratings | category |

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

| 12.5 | 796 | 9950.0 | Chinese |

| 24.5 | 327 | 8011.5 | Mexican |

| 19.0 | 152 | 2888.0 | Japanese |

| 18.0 | 63 | 1134.0 | Indian |

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

As we can see that the Chinese food are most preferred outside their normal dishes when compared to Indian, Mexican and Japanese dishes.

As we can see the new rating system defines by me tells us that highest rating goes to Chinese dishes whereas closely following Mexican food are 2nd most preferred.

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

select sum(stars), sum(review\_count),sum(stars)\* sum(review\_count) as ratings, category

/\* defined a new rating system to know which is best\*/

from business

INNER JOIN category ON business.id = category.business\_id

group by category

having category in ("Indian", "Chinese", "Japanese", "Mexican")

order by ratings asc