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 = 8090 , Foreign Key : business\_id

vi. Checkin = 493

vii. Photo = 6493

viii. Tip = 3979, Foreign Key : business\_id

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 NULL values in the table

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, sum(review\_count)

from business

group by city

order by sum(review\_count) desc

Copy and Paste the Result Below:

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 as [Star Rating], count(stars) as [Count]

from business b where city = 'Avon'

group by stars

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

ii. Beachwood

SQL code used to arrive at answer:

select stars as [Star Rating], count(stars) as [Count]

from business b

where city = 'Beachwood'

group by stars

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

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

| city | stars | count(stars) |

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

| Aberdour | 4.0 | 1 |

| Ahwahtukee | 5.0 | 1 |

| Ahwatukee | 3.5 | 1 |

| Ahwatukee | 4.0 | 1 |

| Ajax | 2.0 | 1 |

| Ajax | 2.5 | 1 |

| Ajax | 3.0 | 1 |

| Ajax | 3.5 | 3 |

| Ajax | 4.0 | 4 |

| Ajax | 4.5 | 1 |

| Allison Park | 1.5 | 1 |

| Allison Park | 3.0 | 1 |

| Allison Park | 3.5 | 1 |

| Allison Park | 4.0 | 1 |

| Ambridge | 4.0 | 1 |

| Ambridge | 4.5 | 1 |

| Ambridge | 5.0 | 1 |

| Amherst | 3.5 | 1 |

| Anjou | 3.5 | 2 |

| Anthem | 2.0 | 1 |

| Anthem | 3.0 | 1 |

| Anthem | 5.0 | 1 |

| Aspinwall | 3.5 | 1 |

| Aspinwall | 4.5 | 1 |

| Aurora | 2.0 | 2 |

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

(Output limit exceeded, 25 of 1183 total rows shown)

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 limit 3

Copy and Paste the Result Below:

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

| name | review\_count |

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

| Gerald | 2000 |

| Sara | 1629 |

| Yuri | 1339 |

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

8. Does posing more reviews correlate with more fans? no

Please explain your findings and interpretation of the results:

I do not see any relation between number of reviews and number of fans.

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

| name | fans | review\_count |

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

| Gerald | 253 | 2000 |

| Sara | 50 | 1629 |

| Yuri | 76 | 1339 |

| .Hon | 101 | 1246 |

| William | 126 | 1215 |

| Harald | 311 | 1153 |

| eric | 16 | 1116 |

| Roanna | 104 | 1039 |

| Mimi | 497 | 968 |

| Christine | 173 | 930 |

| Ed | 38 | 904 |

| Nicole | 43 | 864 |

| Fran | 124 | 862 |

| Mark | 115 | 861 |

| Christina | 85 | 842 |

| Dominic | 37 | 836 |

| Lissa | 120 | 834 |

| Lisa | 159 | 813 |

| Alison | 61 | 775 |

| Sui | 78 | 754 |

| Tim | 35 | 702 |

| L | 10 | 696 |

| Angela | 101 | 694 |

| Crissy | 25 | 676 |

| Lyn | 45 | 675 |

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

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

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"

SQL code used to arrive at answer:

SELECT 'love' Word, COUNT(text) [Total Count]

FROM review

WHERE text LIKE '%love%'

UNION

SELECT 'hate' Word, COUNT(text) [Total Count]

FROM review

WHERE text LIKE '%hate%'

Alt:

select (select count(text)

from review

where text like "%love%") as love\_text, (select count(text) from review

where text like "%hate%") as hate\_text

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

| count (\*) |

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

| 1780 |

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

select

count (\*)

from review

where text like '%hate%';

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

| count (\*) |

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

| 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

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? No

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.

Location is different, Village and City

SQL code used for analysis:

select \* from business where id in

(select business\_id from category where category ="Shopping")

group by stars

having stars>4 and stars<5

select \* from business where id in

(select business\_id from category where category ="Shopping")

group by stars

having stars>2 and stars<3

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 number of business that are open is higher than the number of business that are closed

ii. Difference 2:The business that are open have more number of reviews than the business that are closed

SQL code used for analysis:

select is\_open, count(is\_open)

from business

group by is\_open

select is\_open, sum(review\_count)

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:

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.

From analysis, there is a slight correlation between the number of likes and stars, but this correlation is not strong.

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;