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 = 10,000

ii. Business table = 10,000

iii. Category table = 10,000

iv. Checkin table =10,000

v. elite\_years table = 10,000

vi. friend table = 10,000

vii. hours table = 10,000

viii. photo table = 10,000

ix. review table = 10,000

x. tip table = 10,000

xi. user table = 10,000

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.

1. Business = 10,000 (id)

ii. Hours = 1562 (business\_id)

iii. Category = 2643 (business\_id)

iv. Attribute = 1115 (business\_id)

v. Review = 10,000 id, 8090 business\_id, 9581 user\_id

vi. Checkin = 493 checkin

vii. Photo = 10,000 id, 6493 business\_id

viii. Tip = 537 user\_id,3979 business\_id

ix. User = 10000 id

x. Friend = 11 user\_id

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

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 is null

or name is null

or review\_count is null

or yelping\_since is null

or useful is null

or cool is null

or funny is null

or average\_stars is null

or fans is null

or compliment\_hot is null

or compliment\_more is null

or compliment\_profile is null

or compliment\_cute is null

or compliment\_list is null

or compliment\_note is null

or compliment\_plain is null

or compliment\_cool is null

or compliment\_funny is null

or compliment\_writer is null

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

ii. Table: Business, Column: Stars

min:1.0 max:5.0 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 As rating, count(stars) as count

from business

where city ='Avon'

group by stars;

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

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

| rating | count |

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

| 1.5 | 1 |

| 2.5 | 2 |

| 3.5 | 3 |

| 4.0 | 2 |

| 4.5 | 1 |

| 5.0 | 1 |

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

1. Beachwood

SQL code used to arrive at answer:

select

stars As rating, count(stars) as count

from business

where city ='Beachwood'

group by stars;

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

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

| rating | count |

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

| 2.0 | 1 |

| 2.5 | 1 |

| 3.0 | 2 |

| 3.5 | 2 |

| 4.0 | 1 |

| 4.5 | 2 |

| 5.0 | 5 |

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

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:

As we see higher number of reviews per user we do not see a correlating effect with number of fans.

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

Answer: More reviews with love.

SQL code used to arrive at answer:

select (select count(text)

from review

where text like '%Love%') As Love,

(Select count(text)

from review

where text like '%Hate%') As Hate

Results:  
  
+------+------+

| Love | Hate |

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

| 1780 | 232 |

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

10. Find the top 10 users with the most fans:

SQL code used to arrive at answer:

select

name AS User, fans As fans

from

User

Order by fans DESC limit 10;

Copy and Paste the Result Below:

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

| User | 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.

1. 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.

Yes. For Las Vegas it seems certain locations such as China town attract more patrons willing to submit reviews.Seems as though Chinese food here is very popular.

SQL code used for analysis:

select

b.name, b.city, b.neighborhood, c.category, h.hours, b.review\_count, b.stars,

CASE

When b.stars between 2 and 3 then "Low Rating"

When b.stars between 4 and 5 then "High Rating"

End AS High\_Low\_Rating,

CASE

WHEN h.hours like "%Monday%" then 1

WHEN h.hours like "%Tuesday%" then 2

WHEN h.hours like "%Wednesday%" then 3

WHEN h.hours like "%Thursday%" then 4

WHEN h.hours like "%Friday%" then 5

WHEN h.hours like "%Saturday%" then 6

WHEN h.hours like "%Sunday%" then 7

END AS hourdistr

from business b inner join category c on b.id = c.business\_id

inner join hours h on c.business\_id = h.business\_id

WHere city = "Las Vegas" AND category = "Restaurants"

Group by High\_Low\_Rating

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:

Desserts and general food business are closed while East Asian restaurants are very popular with high reviews and open.

1. Difference 2:

More likely to find the word “Love” in the higher rated reviews of open businesses.

SQL code used for analysis:

Select

b.review\_count, b.name, b.stars,

CASE

When r.text like "%Love%" then "Love Review"

When r.text like "%Hate%" then "Hate Review"

End AS Love\_Hate,

CASE

When is\_open is 0 then "Closed"

When is\_open is 1 then "Open"

End AS Open\_Closed

from business b inner join review r ON b.id = r.business\_id inner join category c on b.id = c.business\_id

group by category

Order by b.stars DESC

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:

Analysis for determining best auto repair shops to recommend in each geographic location for travelers or locals in need.

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

For this analysis I will need location data, auto repair category, Review rating, hours whether the business is open and review count. This data combined gives a user the insight on many aspects a patron would consider when in need of auto repair. More and more modern-day app, web users will search for businesses nearby with the best reviews and the most volume. Also, the I chose the left join because it provided the most records. I found that suing inner join this time limited the amount of data that could be used which would not be useful.

iii. Output of your finished dataset:

1. Provide the SQL code you used to create your final dataset: Copy and paste error occurred. Had to paste image instead.A close-up of a document

   Description automatically generated with medium confidence

Table

Description automatically generated

select

Distinct b.id, b.name, b.city, b.stars, b. review\_count, h.hours,

CASE

When b.is\_open is 0 then "Closed"

When b.is\_open is 1 then "Open"

End as Open

from business b left join review r on b.id = r.business\_id left join hours h on b.id = h.business\_id

Where b.name like "%Auto%"

order by b.review\_count desc;