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 = 2052(Hours) ; 1562 (Business\_Id)

iii. Category = 2643(Business\_Id) ; 712 (Category)

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 537 (User\_Id); 3979(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

SQL code used to arrive at answer:

Select \*

from User As u

Where u.name is Null

(Applied it to different columns)

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 b.city

from business as b

order by b.review\_count Desc

Copy and Paste the Result Below:

city |

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

| Las Vegas |

| Montréal |

| Gilbert |

| Las Vegas |

| Las Vegas |

| Las Vegas |

| Las Vegas |

| Las Vegas |

| Las Vegas |

| Gilbert |

| Las Vegas |

| Scottsdale |

| Las Vegas |

| Las Vegas |

| Henderson |

| Toronto |

| Las Vegas |

| Las Vegas |

| Scottsdale |

| Cleveland |

| Las Vegas |

| Charlotte |

| Phoenix |

| Las Vegas |

| Phoenix |

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

6. Find the distribution of star ratings to the business in the following cities:

1. Avon

SQL code used to arrive at answer:

Select b.stars

,count(b.stars)

from business as b

Where b.city ="Avon"

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

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

| stars | count(b.stars) |

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

| 4.0 | 10 |

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

ii. Beachwood

SQL code used to arrive at answer:

Select b.stars

,count(b.stars)

from business as b

Where b.city ="Beachwood"

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

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

| stars | count(b.stars) |

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

| 5.0 | 14 |

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

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

SQL code used to arrive at answer:

Select u.name

from user as u

Inner join business as b on

b.id = u.id

Order by b.review\_count Desc

Copy and Paste the Result Below:

+--------+

| name |

+--------+

| Irene |

| Jimmy |

| Mandy |

\_\_\_\_\_\_\_\_\_\_

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

No, it does not show a correlation relationship. As you can tell from the table below, some high value count like 449 and 170 have 0 fans while small value count like 4 has 10 fans.

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

| review\_count | fans |

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

| 449 | 0 |

| 170 | 0 |

| 61 | 1 |

| 49 | 0 |

| 29 | 0 |

| 13 | 0 |

| 7 | 0 |

| 7 | 0 |

| 6 | 0 |

| 6 | 0 |

| 5 | 0 |

| 5 | 0 |

| 4 | 10 |

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

Answer:

Yes, there are 1780 reviews with the word “love” while 232 with the word “hate”.

SQL code used to arrive at answer:

Select count(r.id)

from review as r

where r.text like "%love%"

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

SQL code used to arrive at answer:

Select u.name

from user as u

Order by u.fans Desc

limit 10

Copy and Paste the Result Below:

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

| name |

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

| Amy |

| Mimi |

| Harald |

| Gerald |

| Christine |

| Lisa |

| Cat |

| William |

| Fran |

| Lissa |

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

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 = Markham Category = Asian Fusion

Select b.\*

from business as b

where City = "Markham"

Group by b.stars

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

Select b.city

,b.stars

,h.hours

from business as b

inner join hours as h

on b.id = h.business\_id

where City = "Markham"

Group by b.stars

Yes, the businesses with 2-3 stars hours is on Saturday|12:00-23:00

while the businesses with 4-5 stars has no record for hours.

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

Select b.city

,b.stars

,b.review\_count

from business as b

where City = "Markham"

Group by b.stars

Yes, the two groups are different.

The businesses with 2-3 stars’ review count has min 6, max 12, mean 9.67

while the businesses with 4-5 stars’ review count has min 14, max 216, mean 142.

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

| city | stars | review\_count |

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

| Markham | 1.0 | 12 |

| Markham | 1.5 | 3 |

| Markham | 2.0 | 11 |

| Markham | 2.5 | 6 |

| Markham | 3.0 | 12 |

| Markham | 3.5 | 54 |

| Markham | 4.0 | 195 |

| Markham | 4.5 | 216 |

| Markham | 5.0 | 14 |

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

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

SQL code used for analysis:

Select b.city

,b.stars

,b.latitude

,b.longitude

from business as b

where City = "Markham"

Group by b.stars

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

| city | stars | latitude | longitude |

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

| Markham | 1.0 | 43.8464 | -79.3577 |

| Markham | 1.5 | 43.8708 | -79.2139 |

| Markham | 2.0 | 43.8599 | -79.341 |

| Markham | 2.5 | 43.8291 | -79.3533 |

| Markham | 3.0 | 43.8266 | -79.4264 |

| Markham | 3.5 | 43.879 | -79.2618 |

| Markham | 4.0 | 43.8068 | -79.4219 |

| Markham | 4.5 | 43.8437 | -79.3778 |

| Markham | 5.0 | 43.8964 | -79.3155 |

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

The difference of latitude and longitude between two groups are minimum. Locations are similar(close to each other).

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:

Open’s name: ki-Bento Japanese Express

Closed’s name: Scott Roofing Company

ii. Difference 2:

Open’s address: 403 Yonge Street

Closed’s address: 3601 N 36th Ave

SQL code used for analysis:

Select b.\*

from business as b

Group by b.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:

predicting the overall star rating for a business “Babe's Cabaret”

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

I will select stars from review dataset, likes from tip dataset.

I assume that the more likes a business has, the higher rating it is going to have .

iii. Output of your finished dataset:

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

| name | stars | stars | likes |

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

| Babe's Cabaret | 3.0 | 3 | 0 |

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

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

Select b.name

,b.stars

, r.stars

,t.likes

from business as b

inner join review as r

on b.id = r.business\_id

inner join tip as t

on b.id = t.business\_id

group by b.name

having b.name ="Babe's Cabaret"