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 = 2

iii. Category = 15

iv. Attribute = 145

v. Review = 6879 (foreign key = business\_id)

vi. Checkin = 42

vii. Photo = 4727 (foreign key = business\_id)

viii. Tip = 2226 (foreign key = business\_id)

ix. User = 10000

x. Friend = 0

xi. Elite\_years = 390

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 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 name is null or review\_count IS NULL or yelping\_since is null or useful is null or funny is null or cool is null or fans is null or average\_stars is null or compliment\_hot is null or compliment\_profile 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:

Cities repeating

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

| 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 |

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

(Output limit exceeded, 25 of 10000 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 city,stars,count(stars) from business where city="Avon"

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

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

| city | stars | count(stars) |

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

| Avon | 4.0 | 10 |

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

ii. Beachwood

SQL code used to arrive at answer:

select city,stars,count(stars) from business where city="Beachwood"

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

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

| city | stars | count(stars) |

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

| Beachwood | 5.0 | 14 |

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

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 order by review\_count desc

Copy and Paste the Result Below:

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

| id | name | review\_count |

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

| -G7Zkl1wIWBBmD0KRy\_sCw | Gerald | 2000 |

| -3s52C4zL\_DHRK0ULG6qtg | Sara | 1629 |

| -8lbUNlXVSoXqaRRiHiSNg | Yuri | 1339 |

| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon | 1246 |

| -FZBTkAZEXoP7CYvRV2ZwQ | William | 1215 |

| --2vR0DIsmQ6WfcSzKWigw | Harald | 1153 |

| -gokwePdbXjfS0iF7NsUGA | eric | 1116 |

| -DFCC64NXgqrxlO8aLU5rg | Roanna | 1039 |

| -8EnCioUmDygAbsYZmTeRQ | Mimi | 968 |

| -0IiMAZI2SsQ7VmyzJjokQ | Christine | 930 |

| -fUARDNuXAfrOn4WLSZLgA | Ed | 904 |

| -hKniZN2OdshWLHYuj21jQ | Nicole | 864 |

| -9da1xk7zgnnfO1uTVYGkA | Fran | 862 |

| -B-QEUESGWHPE\_889WJaeg | Mark | 861 |

| -kLVfaJytOJY2-QdQoCcNQ | Christina | 842 |

| -kO6984fXByyZm3\_6z2JYg | Dominic | 836 |

| -lh59ko3dxChBSZ9U7LfUw | Lissa | 834 |

| -g3XIcCb2b-BD0QBCcq2Sw | Lisa | 813 |

| -l9giG8TSDBG1jnUBUXp5w | Alison | 775 |

| -dw8f7FLaUmWR7bfJ\_Yf0w | Sui | 754 |

| -AaBjWJYiQxXkCMDlXfPGw | Tim | 702 |

| -jt1ACMiZljnBFvS6RRvnA | L | 696 |

| -IgKkE8JvYNWeGu8ze4P8Q | Angela | 694 |

| -hxUwfo3cMnLTv-CAaP69A | Crissy | 676 |

| -H6cTbVxeIRYR-atxdielQ | Lyn | 675 |

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

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

No It can be seen that Harald has greater fans but lesser reviews

SQL CODE

select name,review\_count,fans from user

Answer

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

| name | review\_count | fans |

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

| Monera | 245 | 15 |

| Joe | 2 | 0 |

| Jeb | 57 | 0 |

| Jed | 8 | 0 |

| Rae | 2 | 0 |

| Carolyn | 43 | 1 |

| Talia | 26 | 2 |

| Ryan | 2 | 0 |

| Joe | 1 | 0 |

| Scott | 7 | 0 |

| John | 3 | 0 |

| Ron | 9 | 0 |

| Bryan | 5 | 0 |

| Patti | 2 | 0 |

| Gary | 23 | 0 |

| Kristin | 28 | 0 |

| Harald | 1153 | 311 |

| Cynthia | 4 | 0 |

| Benjamin | 111 | 2 |

| Mrme | 2 | 0 |

| Kristie | 213 | 10 |

| Tamaki | 239 | 23 |

| Austin | 2 | 0 |

| Kiristen | 400 | 23 |

| Mesut | 25 | 0 |

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

(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:

Love : 1780

Hate : 232

SQL code used to arrive at answer:

Love

select count(\*) from review where text like '%love%'

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

| count(\*) |

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

| 1780 |

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

Hate

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

Copy and Paste the Result Below:

Amy | 503 |

| Mimi | 497 |

| Harald | 311 |

| Gerald | 253 |

| Christine | 173 |

| Lisa | 159 |

| Cat | 133 |

| William | 126 |

| Fran | 124 |

| Lissa | 120 |

11. Is there a strong relationship (or correlation) between having a high number of fans and being listed as "useful" or "funny?" Out of the top 10 users with the highest number of fans, what percent are also listed as “useful” or “funny”?

Key:

0% - 25% - Low relationship

26% - 75% - Medium relationship

76% - 100% - Strong relationship

SQL code used to arrive at answer:

select name,fans,useful,funny from user order by fans desc

Copy and Paste the Result Below:

| name | fans | useful | funny |

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

| Amy | 503 | 3226 | 2554 |

| Mimi | 497 | 257 | 138 |

| Harald | 311 | 122921 | 122419 |

| Gerald | 253 | 17524 | 2324 |

| Christine | 173 | 4834 | 6646 |

| Lisa | 159 | 48 | 13 |

| Cat | 133 | 1062 | 672 |

| William | 126 | 9363 | 9361 |

| Fran | 124 | 9851 | 7606 |

| Lissa | 120 | 455 | 150 |

Please explain your findings and interpretation of the results:

As such there is no correlation between funny and useful with top 10 users

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, hours is not a factor

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

Yes

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

| review\_count | city | stars |

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

| 9 | Tempe | 3.0 |

| 10 | Tempe | 4.0 |

| 42 | Tempe | 2.0 |

| 3 | Tempe | 1.0 |

| 9 | Tempe | 2.5 |

| 30 | Tempe | 4.5 |

| 37 | Tempe | 2.5 |

| 10 | Tempe | 3.5 |

| 3 | Tempe | 2.5 |

| 29 | Tempe | 3.0 |

| 406 | Tempe | 4.5 |

| 10 | Tempe | 3.5 |

| 5 | Tempe | 3.5 |

| 170 | Tempe | 3.5 |

| 13 | Tempe | 4.0 |

| 7 | Tempe | 4.0 |

| 8 | Tempe | 3.0 |

| 40 | Tempe | 4.5 |

| 3 | Tempe | 2.5 |

| 53 | Tempe | 3.0 |

| 7 | Tempe | 1.0 |

| 77 | Tempe | 4.5 |

| 20 | Tempe | 5.0 |

| 163 | Tempe | 3.5 |

| 12 | Tempe | 4.0 |

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

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

SQL code used for analysis:

select review\_count,city,stars from business where city="Tempe" order by stars desc

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

| review\_count | city | stars |

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

| 20 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 4 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 4 | Tempe | 5.0 |

| 28 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 8 | Tempe | 5.0 |

| 5 | Tempe | 5.0 |

| 91 | Tempe | 5.0 |

| 4 | Tempe | 5.0 |

| 20 | Tempe | 5.0 |

| 17 | Tempe | 5.0 |

| 5 | Tempe | 5.0 |

| 13 | Tempe | 5.0 |

| 45 | Tempe | 5.0 |

| 5 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 9 | Tempe | 5.0 |

| 7 | Tempe | 5.0 |

| 11 | Tempe | 5.0 |

| 3 | Tempe | 5.0 |

| 13 | Tempe | 5.0 |

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

(Output limit exceeded, 25 of 261 total rows shown)

No conclusive data can be inferred from the location data

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:

City is different

ii. Difference 2:

name is different

SQL code used for analysis:

select name,city,stars,review\_count,is\_open from business

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

| name | city | stars | review\_count | is\_open |

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

| John's Chinese BBQ Restaurant | Richmond Hill | 3.0 | 30 | 1 |

| Primal Brewery | Huntersville | 4.0 | 42 | 1 |

| Valley Bone and Joint Specialists | Gilbert | 4.5 | 3 | 1 |

| Delmonico Steakhouse | Las Vegas | 4.0 | 1389 | 1 |

| Great Clips | Tempe | 3.0 | 9 | 1 |

| Famous Footwear | Tempe | 4.0 | 10 | 1 |

| Eazor's Auto Salon | Pittsburgh | 5.0 | 8 | 1 |

| Howl at the Moon | Pittsburgh | 3.0 | 44 | 1 |

| Pio Pio | Charlotte | 4.0 | 299 | 1 |

| Sunnyside Grill | Toronto | 3.5 | 37 | 1 |

| World Food Championships | Las Vegas | 3.0 | 5 | 1 |

| Lucky's Pet Grooming & Boutique | Las Vegas | 3.5 | 16 | 1 |

| Bath & Body Works | Charlotte | 4.0 | 3 | 1 |

| The Bar At Bermuda & St. Rose | Henderson | 4.0 | 100 | 1 |

| Welch Physical Therapy | Gilbert | 3.5 | 6 | 1 |

| Kabab House | Phoenix | 4.5 | 24 | 0 |

| Mm Mm Pizza | Canonsburg | 4.0 | 7 | 1 |

| Lake Erie Nature & Science Center | Bay Village | 4.5 | 19 | 1 |

| Happy Moose Bar and Grill | Streetsboro | 3.5 | 61 | 1 |

| Q's Nails | Charlotte | 3.5 | 3 | 1 |

| Hungry Howie's Pizza | Charlotte | 3.0 | 12 | 1 |

| The Manor - Boutique Salon | Toronto | 4.5 | 5 | 1 |

| Sam's Club | Scottsdale | 3.0 | 26 | 1 |

| SkinRN | Henderson | 5.0 | 3 | 0 |

| International Newsagents | Edinburgh | 4.0 | 10 | 1 |

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

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

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 want to do analysis about which business are most influential and in which category and location

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

I would require name, city, review\_count, is\_open, stars, category data from business and category table

I have chosen this data as based on their ratings and reviews I am able to determine which business has the most dominant influence

iii. Output of your finished dataset:

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

| name | city | review\_count | is\_open | stars | category |

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

| Motors & More | Las Vegas | 7 | 1 | 5.0 | Home Services |

| Motors & More | Las Vegas | 7 | 1 | 5.0 | Solar Installation |

| Motors & More | Las Vegas | 7 | 1 | 5.0 | Heating & Air Conditioning/HVAC |

| Frankie Fettuccine Food Truck | Oakville | 7 | 1 | 5.0 | Event Planning & Services |

| Frankie Fettuccine Food Truck | Oakville | 7 | 1 | 5.0 | Caterers |

| Frankie Fettuccine Food Truck | Oakville | 7 | 1 | 5.0 | Food Trucks |

| Frankie Fettuccine Food Truck | Oakville | 7 | 1 | 5.0 | Food |

| Halo Plumbing | Henderson | 5 | 1 | 5.0 | Home Services |

| Halo Plumbing | Henderson | 5 | 1 | 5.0 | Plumbing |

| Middleton Art and Framing | Middleton | 8 | 1 | 5.0 | Art Galleries |

| Middleton Art and Framing | Middleton | 8 | 1 | 5.0 | Shopping |

| Middleton Art and Framing | Middleton | 8 | 1 | 5.0 | Arts & Entertainment |

| Middleton Art and Framing | Middleton | 8 | 1 | 5.0 | Arts & Crafts |

| Middleton Art and Framing | Middleton | 8 | 1 | 5.0 | Framing |

| Christian Brothers Automotive | Chandler | 63 | 1 | 5.0 | Automotive |

| Christian Brothers Automotive | Chandler | 63 | 1 | 5.0 | Auto Repair |

| Christian Brothers Automotive | Chandler | 63 | 1 | 5.0 | Oil Change Stations |

| Christian Brothers Automotive | Chandler | 63 | 1 | 5.0 | Transmission Repair |

| Buddy's Muffler & Exhaust | Gastonia | 4 | 1 | 5.0 | Automotive |

| Buddy's Muffler & Exhaust | Gastonia | 4 | 1 | 5.0 | Auto Repair |

| Innercity MMA | Toronto | 3 | 1 | 5.0 | Fitness & Instruction |

| Innercity MMA | Toronto | 3 | 1 | 5.0 | Martial Arts |

| Innercity MMA | Toronto | 3 | 1 | 5.0 | Active Life |

| Journey's Dry Carpet Cleaning | Charlotte | 3 | 0 | 5.0 | Home Services |

| Journey's Dry Carpet Cleaning | Charlotte | 3 | 0 | 5.0 | Carpet Installation |

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

(Output limit exceeded, 25 of 696 total rows shown)

I will be doing top 4 analysis

It can be seen that most influential business is in Las Vegas, Oakville, Middleton, Chandler and the most influential shops are Motors & More, Frankie Fettuccine Food Truck, Middleton Art and Framing, Christian Brothers Automotive and these shops have succeeded in different department

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

select b.name,b.city,b.review\_count,b.is\_open,b.stars,c.category from business as b inner join category as c on b.id=c.business\_id order by stars desc