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 = 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:

There are 20 columns in the user table. Namely: 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

select \*

from user

where id is null

select \*

from user

where name is null

select \*

from user

where review\_count is null

select \*

from user

where yelping\_since is null

select \*

from user

where useful is null

select \*

from user

where funny is null

select \*

from user

where cool is null

select \*

from user

where fans is null

select \*

from user

where average\_stars is null

select \*

from user

where compliment\_hot is null

select \*

from user

where compliment\_more is null

select \*

from user

where compliment\_profile is null

select \*

from user

where compliment\_cute is null

select \*

from user

where compliment\_list is null

select \*

from user

where compliment\_note is null

select \*

from user

where compliment\_plain is null

select \*

from user

where compliment\_cool is null

select \*

from user

where compliment\_funny is null

select \*

from user

where compliment\_writer is null

select \*

from user

where 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 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:

--tot\_rev is the total number of reviews from that city

select city, count(\*) as tot\_rev

from business inner join review on business.id = review.business\_id

group by city

order by tot\_rev desc

Copy and Paste the Result Below:

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

| city | tot\_rev |

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

| Las Vegas | 193 |

| Phoenix | 65 |

| Toronto | 51 |

| Scottsdale | 37 |

| Henderson | 30 |

| Tempe | 28 |

| Pittsburgh | 23 |

| Chandler | 22 |

| Charlotte | 21 |

| Montréal | 18 |

| Madison | 16 |

| Gilbert | 13 |

| Mesa | 13 |

| Cleveland | 12 |

| North Las Vegas | 6 |

| Edinburgh | 5 |

| Glendale | 5 |

| Lakewood | 5 |

| Cave Creek | 4 |

| Champaign | 4 |

| Markham | 4 |

| North York | 4 |

| Mississauga | 3 |

| Surprise | 3 |

| Avondale | 2 |

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

(Output limit exceeded, 25 of 67 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 stars, 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 | 1 |

| 2.5 | 2 |

| 3.5 | 3 |

| 4.0 | 2 |

| 4.5 | 1 |

| 5.0 | 1 |

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

ii. Beachwood

SQL code used to arrive at answer:

select stars, 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 | 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?

Please explain your findings and interpretation of the results:

No. I don’t think so.

As per the table on the right, arranging the number of reviews by the user, I noticed that while the number of reviews are in descending order, the number of fans are wildly out of order. As per my results, the number of fans are not correlated with the number of reviews posted by the user.

SQL Code:

select name, review\_count, fans

from user

order by review\_count desc

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

Answer: Review with the word ‘love’ are more than the ones with ‘hate’ (love- 1780, hate- 232)

SQL code used to arrive at answer:

select count(\*) as review\_l

from review

where text like '%love%'

select count(\*) as review\_l

from review

where text like '%hate%'

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.

Phoenix, Restaurants

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

Yes. they do. The lower rated restaurants open earlier than the higher rated ones

select city, name, category, stars, hours--review\_count as reviews

from (business inner join category on category.business\_id = business.id)

inner join hours on hours.business\_id = business.id

where category= "Restaurants" AND city= "Phoenix"

group by stars

order by stars desc

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

| city | name | category | stars | hours | +---------+----------------------------------------+-------------+-------+----------------------+

| Phoenix | Charlie D's Catfish & Chicken | Restaurants | 4.5 | Saturday|11:00-18:00 |

| Phoenix | Bootleggers Modern American Smokehouse | Restaurants | 4.0 | Saturday|11:00-22:00 |

| Phoenix | Five Guys | Restaurants | 3.5 | Saturday|10:00-22:00 |

| Phoenix | Gallagher's | Restaurants | 3.0 | Saturday|9:00-2:00 |

| Phoenix | McDonald's | Restaurants | 2.0 | Saturday|5:00-0:00 | +---------+----------------------------------------+-------------+-------+----------------------+

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

Yes.

select city, name, category, stars, review\_count as reviews

from (business inner join category on category.business\_id = business.id)

inner join hours on hours.business\_id = business.id

where category= "Restaurants" AND city= "Phoenix"

group by stars

order by stars desc

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

| city | name | category | stars | reviews | +---------+----------------------------------------+-------------+-------+---------+

| Phoenix | Charlie D's Catfish & Chicken | Restaurants | 4.5 | 7 |

| Phoenix | Bootleggers Modern American Smokehouse | Restaurants | 4.0 | 431 |

| Phoenix | Five Guys | Restaurants | 3.5 | 63 |

| Phoenix | Gallagher's | Restaurants | 3.0 | 60 |

| Phoenix | McDonald's | Restaurants | 2.0 | 8 | +---------+----------------------------------------+-------------+-------+---------+

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

No. All the restaurants are not in one zip code.

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

| city | name | category | stars | zip | +---------+----------------------------------------+-------------+-------+-------+

| Phoenix | Charlie D's Catfish & Chicken | Restaurants | 4.5 | 85034 |

| Phoenix | Bootleggers Modern American Smokehouse | Restaurants | 4.0 | 85028 |

| Phoenix | Five Guys | Restaurants | 3.5 | 85008 | | Phoenix | Gallagher's | Restaurants | 3.0 | 85024 |

| Phoenix | McDonald's | Restaurants | 2.0 | 85004 | +---------+----------------------------------------+-------------+-------+-------+

SQL code used for analysis:

select city, name, category, stars, postal\_code as zip

from (business inner join category on category.business\_id = business.id)

inner join hours on hours.business\_id = business.id

where category= "Restaurants" AND city= "Phoenix"

group by stars

order by stars desc

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.

The query results do not make sense as it returned only 2 values.

i. Difference 1:

The one with the lower rating was still open.

ii. Difference 2:

The one with the lower rating has more reviews.

SQL code used for analysis:

select city, name, category, stars, review\_count, is\_open

from business INNER JOIN category on business.id = category.business\_id

where city = "Phoenix" AND category = "Restaurants"

group by is\_open

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

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

| Phoenix | Charlie D's Catfish & Chicken | Restaurants | 4.5 | 7 | 0 | | Phoenix | Matt's Big Breakfast | Restaurants | 4.0 | 188 | 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:

I wanted to see which category of establishments on Yelp have the most reviews and photos. This will give me an estimate of what establishments are the most liked and visited.

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 need the category of establishment, the number of reviews the category has received, the average star rating and the number of photos that have been uploaded. I chose this data as in my view, it is the easiest and most effective way to see the most visited and liked establishments.

iii. Output of your finished dataset:

+---------------------------+---------+--------+--------------------+ | category | Reviews | Photos | Average\_Rating | +---------------------------+---------+--------+--------------------+ | Poutineries | 47 | 1 | 3.0 |

| Caterers | 69 | 1 | 4.0 |

| Event Planning & Services | 69 | 1 | 4.0 |

| Latin American | 69 | 1 | 4.0 |

| Fast Food | 83 | 1 | 3.5 |

| Mexican | 83 | 2 | 3.75 |

| Diners | 87 | 1 | 3.5 |

| Restaurants | 87 | 9 | 3.7777777777777777 |

| American (Traditional) | 94 | 1 | 4.0 |

| Bars | 94 | 2 | 3.5 |

| Burgers | 94 | 2 | 3.5 |

| Nightlife | 94 | 2 | 3.5 |

| Sushi Bars | 94 | 1 | 4.0 |

| Irish | 141 | 1 | 3.0 |

| Pubs | 141 | 1 | 3.0 |

| American (New) | 168 | 2 | 4.0 |

| Breakfast & Brunch | 168 | 2 | 4.0 |

| French | 168 | 2 | 4.0 |

| Gluten-Free | 168 | 2 | 4.0 |

| Sandwiches | 168 | 3 | 4.333333333333333 |

| Vegetarian | 168 | 2 | 4.0 |

| Arabian | 267 | 1 | 5.0 |

| Halal | 267 | 1 | 5.0 |

| Mediterranean | 267 | 1 | 5.0 |

| Salad | 267 | 1 | 5.0 | +---------------------------+---------+--------+--------------------+

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

select category, review\_count as Reviews, count(distinct(photo.id)) as Photos, avg(stars) as Average\_Rating

from (business inner join category on business.id = category.business\_id)

inner join photo on business.id = photo.business\_id

group by category

order by review\_count