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:

1. Attribute table = SELECT Count (\*) FROM attribute Answer:10000

ii. Business table = SELECT Count (\*) FROM business Answer: 10000

iii. Category table = SELECT Count (\*) FROM category Answer:10000

iv. Checkin table = SELECT Count (\*) FROM checkin Answer:10000

v. elite\_years table = SELECT Count (\*) FROM elite\_years Answer:10000

vi. friend table = SELECT Count (\*) FROM friend Answer:10000

vii. hours table = SELECT Count (\*) FROM hours Answer:10000

viii. photo table = SELECT Count (\*) FROM photo Answer:10000

ix. review table = SELECT Count (\*) FROM review Answer:10000

x. tip table = SELECT Count (\*) FROM tip Answer:10000

xi. user table = SELECT Count (\*) FROM user Answer: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.

SELECT COUNT (DISTINCT key) FROM Table

i. Business = id (10000) SELECT COUNT (DISTINCT id) FROM business

ii. Hours = business\_id (1562) SELECT COUNT (DISTINCT business\_id) FROM hours

iii. Category = business\_id (2643) SELECT COUNT (DISTINCT business\_id) FROM category

iv. Attribute = business\_id (1115) SELECT COUNT (DISTINCT business\_id)

FROM attribute

v. Review = id (10000), business\_id (8090), user\_id (9581) SELECT COUNT (DISTINCT id)

FROM review

vi. Checkin = business\_id (493) SELECT COUNT (DISTINCT business\_id)

FROM checkin

vii. Photo = id (10000),business\_id (6493) SELECT COUNT (DISTINCT business\_id)

FROM photo

viii. Tip = user\_id (537) business\_id (3979) SELECT COUNT (DISTINCT business\_id) FROM tip

ix. User = id (10000) SELECT COUNT (DISTINCT id) FROM user

x. Friend = user\_id (11) SELECT count(DISTINCT user\_id) FROM friend

xi. Elite\_years = user\_id (2780) SELECT COUNT (DISTINCT user\_id)

FROM elite\_years

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 COUNT(\*) |
|  | 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 |
|  | funny IS NULL OR |
|  | cool IS NULL OR |
|  | fans IS NULL OR |
|  | average\_stars 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 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) AS reviews

        FROM business

        GROUP BY city

        ORDER BY reviews DESC

Copy and Paste the Result Below:

city | reviews |

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

| Las Vegas | 82854 |

| Phoenix | 34503 |

| Toronto | 24113 |

| Scottsdale | 20614 |

| Charlotte | 12523 |

| Henderson | 10871 |

| Tempe | 10504 |

| Pittsburgh | 9798 |

| Montréal | 9448 |

| Chandler | 8112 |

| Mesa | 6875 |

| Gilbert | 6380 |

| Cleveland | 5593 |

| Madison | 5265 |

| Glendale | 4406 |

| Mississauga | 3814 |

| Edinburgh | 2792 |

| Peoria | 2624 |

| North Las Vegas | 2438 |

| Markham | 2352 |

| Champaign | 2029 |

| Stuttgart | 1849 |

| Surprise | 1520 |

| Lakewood | 1465 |

| Goodyear | 1155 |

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

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,sum(stars)

        FROM business

        WHERE city = 'Avon'

        GROUP BY  city,stars

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

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

| city | stars | sum(stars) |

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

| Avon | 1.5 | 1.5 |

| Avon | 2.5 | 5.0 |

| Avon | 3.5 | 10.5 |

| Avon | 4.0 | 8.0 |

| Avon | 4.5 | 4.5 |

| Avon | 5.0 | 5.0 |

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

ii. Beachwood

SQL code used to arrive at answer:

SELECT city,stars,sum(stars)

        FROM business

        WHERE city = 'Beachwood'

        GROUP BY  city,sta

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

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

| city | stars | sum(stars) |

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

| Beachwood | 2.0 | 2.0 |

| Beachwood | 2.5 | 2.5 |

| Beachwood | 3.0 | 6.0 |

| Beachwood | 3.5 | 7.0 |

| Beachwood | 4.0 | 4.0 |

| Beachwood | 4.5 | 9.0 |

| Beachwood | 5.0 | 25.0 |

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

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

LIMIT 3

Copy and Paste the Result Below:

id | name | review\_count |

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

| -G7Zkl1wIWBBmD0KRy\_sCw | Gerald | 2000 |

| -3s52C4zL\_DHRK0ULG6qtg | Sara | 1629 |

| -8lbUNlXVSoXqaRRiHiSNg | Yuri | 1339 |

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

SELECT review\_count,fans

FROM user

ORDER BY fans DESC

Yes, there is positive correlation between reviews and fans. As he reviews are higher number of fans are also increasing according to the query.

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:

SELECT COUNT(\*) FROM review

WHERE text LIKE '%love%'

SELECT COUNT(\*) 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.

City= Charlotte Category=Restaurants

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

SELECT b.name,b.stars,b.city,c.category,h.hours

FROM business b JOIN category c

ON b.id=c.business\_id

JOIN hours h on

c.business\_id=h.business\_id

WHERE b.city = 'Charlotte' and

c.category='Restaurants'

name | stars | city | category | hours |

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

| Subway | 3.5 | Charlotte | Restaurants | Monday|6:00-22:00 |

| Subway | 3.5 | Charlotte | Restaurants | Tuesday|6:00-22:00 |

| Subway | 3.5 | Charlotte | Restaurants | Friday|6:00-22:00 |

| Subway | 3.5 | Charlotte | Restaurants | Wednesday|6:00-22:00 |

| Subway | 3.5 | Charlotte | Restaurants | Thursday|6:00-22:00 |

| Subway | 3.5 | Charlotte | Restaurants | Saturday|10:00-21:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Monday|11:00-20:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Tuesday|11:00-20:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Friday|11:00-20:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Wednesday|11:00-20:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Thursday|11:00-20:00 |

| Big City Grill | 5.0 | Charlotte | Restaurants | Saturday|11:00-20:00 |

According to the query result above, restaurants that lower ratings, opens very early in comparison with restaurants that has 5.0 rating.

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

Yes,they have different number of reviews.

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

SQL code used for analysis:

SELECT b.name,b.stars,b.city,b.postal\_code,b.stars,c.category,h.hours

FROM business b JOIN category c

ON b.id=c.business\_id

JOIN hours h on

c.business\_id=h.business\_id

WHERE b.city = 'Charlotte' and

c.category='Restaurants'

GROUP BY b.stars

They have different locations according to the query.

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

| name | city | category | stars | hours | review\_count | address | postal\_code |

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

| Subway | Charlotte | Restaurants | 3.5 | Saturday|10:00-21:00 | 7 | 2904 Yorkmont Rd | 28208 |

| Big City Grill | Charlotte | Restaurants | 5.0 | Saturday|11:00-20:00 | 4 | 4450 The Plaza, Ste E | 28215 |

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: Businesses that are open have highest reviews.

ii. Difference 2: Businesses that are closed have lower avg stars.

SQL code used for analysis:

SELECT COUNT(DISTINCT(id)),

       AVG(review\_count),

       SUM(review\_count),

       AVG(stars),

       is\_open

       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 the following:

1. Indicate the type of analysis you chose to do:

I want to find which city is better for food choice among customers and it could be good for businesses owners to open restaurants per this choice.

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

I used name, star rating, neighborhood, city category columns to get idea.

1. Output of your finished dataset:

stars | name | neighborhood | city | category | star\_rating |

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

| 5.0 | Red Apron Bakeshop | | Chandler | Food | 4-5 stars |

| 4.0 | Koko Bakery | Goodrich Kirtland | Cleveland | Food | 3-4 stars |

| 4.0 | B.A. Sweetie Candy | | Cleveland | Food | 3-4 stars |

| 4.5 | West Side Market | Ohio City | Cleveland | Food | 4-5 stars |

| 2.5 | Walgreens | Eastside | Las Vegas | Food | 2-3 stars |

| 4.0 | Sweet Ruby Jane Confections | Southeast | Las Vegas | Food | 3-4 stars |

| 4.0 | C's Restaurant Bakery and Coffee Shop | | Middleton | Food | 3-4 stars |

| 5.0 | Frankie Fettuccine Food Truck | | Oakville | Food | 4-5 stars |

| 3.5 | Food Lion | | Paw Creek | Food | 3-4 stars |

| 3.0 | Starbucks | | Phoenix | Food | 2-3 stars |

| 2.5 | Safeway Food & Drug | | Phoenix | Food | 2-3 stars |

| 4.0 | Bootleggers Modern American Smokehouse | | Phoenix | Food | 3-4 stars |

| 5.0 | Water N Ice | | Phoenix | Food | 4-5 stars |

| 4.5 | Vanilla Pastry Studio | Shadyside | Pittsburgh | Food | 4-5 stars |

| 4.0 | The Cider Mill | | Scottsdale | Food | 3-4 stars |

| 2.5 | Mango Flats | | Tempe | Food | 2-3 stars |

| 4.0 | Taqueria Y Cenaduria Culiacan | | Tolleson | Food | 3-4 stars |

| 2.5 | Loblaws | Roncesvalles | Toronto | Food | 2-3 stars |

| 3.5 | The Kosher Gourmet | | Toronto | Food | 3-4 stars |

| 4.0 | Halo Brewery | Wallace Emerson | Toronto | Food | 3-4 stars |

| 4.5 | Cabin Fever | High Park | Toronto | Food | 4-5 stars |

| 3.5 | Baby Cakes | | Willoughby | Food | 3-4 stars |

| 3.0 | What A Bagel | | York | Food | 2-3 stars |

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

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

Select b.stars, b.name, b.neighborhood,b.city, c.category,

CASE

WHEN B.stars BETWEEN 0 AND 1 THEN '2-3 stars'

WHEN B.stars BETWEEN 1 AND 2 THEN '1-2 stars'

WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'

WHEN B.stars BETWEEN 3 AND 4 THEN '3-4 stars'

WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'

END AS star\_rating

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

where c.category is 'Food'

order by city, star\_rating