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:

SELECT COUNT (\*)

FROM table\_name;

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

1. 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 =10000

ii. Hours =1562

iii. Category =2643

iv. Attribute =1115

v. Review =10000

vi. Checkin =493

vii. Photo =10000

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

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

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\_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

|  |
| --- |
|  |
|  |  |

1. For each table and column listed below, display the smallest (minimum), largest (maximum), and average (mean) value for the following fields:

SELECT MIN(column), MAX(column), AVG(column)

FROM table;

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, 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 |

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

(Output limit exceeded, 25 of 362 total rows shown)

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

1. Avon

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

| stars | reviews |

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

| 1.5 | 10 |

| 2.5 | 6 |

| 3.5 | 88 |

| 4.0 | 21 |

| 4.5 | 31 |

| 5.0 | 3 |

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

SQL code used to arrive at answer:

SELECT stars, SUM(review\_count) AS reviews

FROM Business

WHERE city == 'Avon'

GROUP BY stars;

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

1. Beachwood

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

| stars | reviews |

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

| 2.0 | 8 |

| 2.5 | 3 |

| 3.0 | 11 |

| 3.5 | 6 |

| 4.0 | 69 |

| 4.5 | 17 |

| 5.0 | 23 |

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

SQL code used to arrive at answer:

SELECT stars, SUM(review\_count) AS reviews

FROM Business

WHERE city == 'Beachwood'

GROUP BY stars;

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

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:

Posting more reviews does correlate with more fans. However, more fans also correlates with time posting or how long the user has been yelping. Posting more reviews does correlate with more fans but does not certainly cause it.

SELECT id, name, review\_count, fans, yelping\_since

FROM user

ORDER BY fans DESC;

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

| id | name | review\_count | fans | yelping\_since |

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

| -9I98YbNQnLdAmcYfb324Q | Amy | 609 | 503 | 2007-07-19 00:00:00 |

| -8EnCioUmDygAbsYZmTeRQ | Mimi | 968 | 497 | 2011-03-30 00:00:00 |

| --2vR0DIsmQ6WfcSzKWigw | Harald | 1153 | 311 | 2012-11-27 00:00:00 |

| -G7Zkl1wIWBBmD0KRy\_sCw | Gerald | 2000 | 253 | 2012-12-16 00:00:00 |

| -0IiMAZI2SsQ7VmyzJjokQ | Christine | 930 | 173 | 2009-07-08 00:00:00 |

| -g3XIcCb2b-BD0QBCcq2Sw | Lisa | 813 | 159 | 2009-10-05 00:00:00 |

| -9bbDysuiWeo2VShFJJtcw | Cat | 377 | 133 | 2009-02-05 00:00:00 |

| -FZBTkAZEXoP7CYvRV2ZwQ | William | 1215 | 126 | 2015-02-19 00:00:00 |

| -9da1xk7zgnnfO1uTVYGkA | Fran | 862 | 124 | 2012-04-05 00:00:00 |

| -lh59ko3dxChBSZ9U7LfUw | Lissa | 834 | 120 | 2007-08-14 00:00:00 |

| -B-QEUESGWHPE\_889WJaeg | Mark | 861 | 115 | 2009-05-31 00:00:00 |

| -DmqnhW4Omr3YhmnigaqHg | Tiffany | 408 | 111 | 2008-10-28 00:00:00 |

| -cv9PPT7IHux7XUc9dOpkg | bernice | 255 | 105 | 2007-08-29 00:00:00 |

| -DFCC64NXgqrxlO8aLU5rg | Roanna | 1039 | 104 | 2006-03-28 00:00:00 |

| -IgKkE8JvYNWeGu8ze4P8Q | Angela | 694 | 101 | 2010-10-01 00:00:00 |

| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon | 1246 | 101 | 2006-07-19 00:00:00 |

| -4viTt9UC44lWCFJwleMNQ | Ben | 307 | 96 | 2007-03-10 00:00:00 |

| -3i9bhfvrM3F1wsC9XIB8g | Linda | 584 | 89 | 2005-08-07 00:00:00 |

| -kLVfaJytOJY2-QdQoCcNQ | Christina | 842 | 85 | 2012-10-08 00:00:00 |

| -ePh4Prox7ZXnEBNGKyUEA | Jessica | 220 | 84 | 2009-01-12 00:00:00 |

| -4BEUkLvHQntN6qPfKJP2w | Greg | 408 | 81 | 2008-02-16 00:00:00 |

| -C-l8EHSLXtZZVfUAUhsPA | Nieves | 178 | 80 | 2013-07-08 00:00:00 |

| -dw8f7FLaUmWR7bfJ\_Yf0w | Sui | 754 | 78 | 2009-09-07 00:00:00 |

| -8lbUNlXVSoXqaRRiHiSNg | Yuri | 1339 | 76 | 2008-01-03 00:00:00 |

| -0zEEaDFIjABtPQni0XlHA | Nicole | 161 | 73 | 2009-04-30 00:00:00 |

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

(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: There are more reviews with the word “love” than the word “hate”.

SQL code used to arrive at answer:

SELECT COUNT(\*)

FROM review

WHERE text LIKE '%love%';

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

| COUNT(\*) |

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

| 1780 |

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

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 id, name, fans

FROM user

ORDER BY fans DESC

LIMIT 10;

Copy and Paste the Result Below:

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

| id | name | fans |

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

| -9I98YbNQnLdAmcYfb324Q | Amy | 503 |

| -8EnCioUmDygAbsYZmTeRQ | Mimi | 497 |

| --2vR0DIsmQ6WfcSzKWigw | Harald | 311 |

| -G7Zkl1wIWBBmD0KRy\_sCw | Gerald | 253 |

| -0IiMAZI2SsQ7VmyzJjokQ | Christine | 173 |

| -g3XIcCb2b-BD0QBCcq2Sw | Lisa | 159 |

| -9bbDysuiWeo2VShFJJtcw | Cat | 133 |

| -FZBTkAZEXoP7CYvRV2ZwQ | William | 126 |

| -9da1xk7zgnnfO1uTVYGkA | Fran | 124 |

| -lh59ko3dxChBSZ9U7LfUw | 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, review\_count, yelping\_since

FROM user

ORDER BY fans DESC;

Copy and Paste the Result Below:

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

| name | fans | useful | funny | review\_count | yelping\_since |

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

| Amy | 503 | 3226 | 2554 | 609 | 2007-07-19 00:00:00 |

| Mimi | 497 | 257 | 138 | 968 | 2011-03-30 00:00:00 |

| Harald | 311 | 122921 | 122419 | 1153 | 2012-11-27 00:00:00 |

| Gerald | 253 | 17524 | 2324 | 2000 | 2012-12-16 00:00:00 |

| Christine | 173 | 4834 | 6646 | 930 | 2009-07-08 00:00:00 |

| Lisa | 159 | 48 | 13 | 813 | 2009-10-05 00:00:00 |

| Cat | 133 | 1062 | 672 | 377 | 2009-02-05 00:00:00 |

| William | 126 | 9363 | 9361 | 1215 | 2015-02-19 00:00:00 |

| Fran | 124 | 9851 | 7606 | 862 | 2012-04-05 00:00:00 |

| Lissa | 120 | 455 | 150 | 834 | 2007-08-14 00:00:00 |

| Mark | 115 | 4008 | 570 | 861 | 2009-05-31 00:00:00 |

| Tiffany | 111 | 1366 | 984 | 408 | 2008-10-28 00:00:00 |

| bernice | 105 | 120 | 112 | 255 | 2007-08-29 00:00:00 |

| Roanna | 104 | 2995 | 1188 | 1039 | 2006-03-28 00:00:00 |

| Angela | 101 | 158 | 164 | 694 | 2010-10-01 00:00:00 |

| .Hon | 101 | 7850 | 5851 | 1246 | 2006-07-19 00:00:00 |

| Ben | 96 | 1180 | 1155 | 307 | 2007-03-10 00:00:00 |

| Linda | 89 | 3177 | 2736 | 584 | 2005-08-07 00:00:00 |

| Christina | 85 | 158 | 34 | 842 | 2012-10-08 00:00:00 |

| Jessica | 84 | 2161 | 2091 | 220 | 2009-01-12 00:00:00 |

| Greg | 81 | 820 | 753 | 408 | 2008-02-16 00:00:00 |

| Nieves | 80 | 1091 | 774 | 178 | 2013-07-08 00:00:00 |

| Sui | 78 | 9 | 18 | 754 | 2009-09-07 00:00:00 |

| Yuri | 76 | 1166 | 220 | 1339 | 2008-01-03 00:00:00 |

| Nicole | 73 | 13 | 10 | 161 | 2009-04-30 00:00:00 |

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

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

Please explain your findings and interpretation of the results:

There is a strong correlation for being listed as “useful” or “funny”. The higher the counts of useful and funny the more likely that user has more fans. An exception to this would be the third reviewer “Harald”. He has the greatest number of “useful” and “funny” reviews but does not have the most fans.

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?

The 4-5 star group seems to have shorter hours than the 2-3 star group. This is not a great sample size (3 entries).

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

Yes they do. One of the 4-5 star groups has a much greater number of reviews and the other 4-5 star group has a slightly greater number of reviews.

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

We cannot infer anything from the location data provided between the two groups because each business is in a different zip code.

SQL code used for analysis:

SELECT B.name,

B.review\_count,

H.hours,

postal\_code,

CASE

WHEN hours LIKE "%monday%" THEN 1

WHEN hours LIKE "%tuesday%" THEN 2

WHEN hours LIKE "%wednesday%" THEN 3

WHEN hours LIKE "%thursday%" THEN 4

WHEN hours LIKE "%friday%" THEN 5

WHEN hours LIKE "%saturday%" THEN 6

WHEN hours LIKE "%sunday%" THEN 7

END AS ord,

CASE

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

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

END AS star\_rating

FROM business B INNER JOIN hours H

ON B.id = H.business\_id

INNER JOIN category C

ON C.business\_id = B.id

WHERE (B.city == 'Las Vegas'

AND

C.category LIKE 'shopping')

AND

(B.stars BETWEEN 2 AND 3

OR

B.stars BETWEEN 4 AND 5)

GROUP BY stars,ord

ORDER BY ord,star\_rating ASC

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:

AVG(review\_count) = 31.757

ii. Difference 2:

AVG(review\_count) = 23.198

SQL code used for analysis:

SELECT AVG(review\_count) FROM business

WHERE is\_open = 1;

SELECT AVG(review\_count) FROM business

WHERE is\_open = 0;

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 see what makes a business most successful by rating.

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

I chose to join business data with categories to see if company classification contributed to ratings. With this, location and the business being open were also factors.

1. Output of your finished dataset:

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

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

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

| Horton Carpet Cleaning | Ahwahtukee | None | 5.0 | 1 |

| ODBQ | Ambridge | None | 5.0 | 1 |

| Goodyear Chiropractic | Avondale | None | 5.0 | 1 |

| Studio Mz | Beachwood | None | 5.0 | 1 |

| Deni's Door Service | Broadlands | None | 5.0 | 1 |

| SYNERGY HomeCare | Broadview Heights | None | 5.0 | 1 |

| Brooklin Bakery | Brooklin | None | 5.0 | 0 |

| Samuels Income Tax | Brunswick | None | 5.0 | 1 |

| L.C's World Tae Kwon Do | Cave Creek | None | 5.0 | 1 |

| 12th House Interiors | Chagrin Falls | None | 5.0 | 1 |

| Tootsie's Tap | Dane | None | 5.0 | 1 |

| Parks Automotive | De Forest | None | 5.0 | 1 |

| First Impressions Dental | East Gwillimbury | None | 5.0 | 1 |

| Fort Mill Vision Center | Fort Mill | None | 5.0 | 1 |

| Trattoria Gallo Nero Beccofino | Freyburg | None | 5.0 | 1 |

| The Family Pet Clinic | Garfield Heights | None | 5.0 | 1 |

| Shutter Envy | Gilbert | None | 5.0 | 1 |

| Nicole Podrat, D.M.D | Glenshaw | None | 5.0 | 0 |

| Inkz Tattoo Company | Guadalupe | None | 5.0 | 1 |

| Colette's Body & Sole Day Spa | Houston | None | 5.0 | 1 |

| Dani's Hair Loft | Kennedy Township | None | 5.0 | 1 |

| Deli La Trattoria | Kirkland | None | 5.0 | 1 |

| Boulangerie De L'ile Bizard | L'ile-Bizard | None | 5.0 | 1 |

| Le Shack du Pêcheur | La Prairie | None | 5.0 | 1 |

| Nikki's Collar Club | Lyndhurst | None | 5.0 | 1 |

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

(Output limit exceeded, 25 of 362 total rows shown)

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

SELECT b.name, b.city, c.category, b.stars, b.is\_open

FROM business b

LEFT JOIN category c

ON b.id = c.business\_id

GROUP BY b.city

ORDER BY b.stars DESC