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\_column)) from table\_name;

i. Business =id:10000

ii. Hours =business\_id:1562

iii. Category =business\_id:2643

iv. Attribute =business\_id:1115

v. Review =id: 10000 ; business\_id:8090 ; user\_id:9581

vi. Checkin = business\_id:493

vii. Photo = id:10000

viii. Tip = user\_id:537 ; business\_id:3979

ix. User = id:10000

x. Friend = id:11

xi. Elite\_years =user\_id: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:

SQL code used to arrive at answer:

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

1. Table: Review, Column: Stars

select min(stars),max(stars),avg(stars) from review

min:1 max:5 avg:3.7082

1. Table: Business, Column: Stars

select min(stars),max(stars),avg(stars) from business

min:1.0 max:5.0 avg:3.6549

1. Table: Tip, Column: Likes

select min(likes),max(likes),avg(likes) from tip

min:0 max:2 avg:0.0144

1. Table: Checkin, Column: Count

select min(count),max(count),avg(count) from checkin

min:1 max:53 avg:1.9414

1. Table: User, Column: Review\_count

select min(review\_count),max(review\_count),avg(review\_count) from user

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,count(review\_count) as "total\_review" from business

group by city

order by total\_review desc;

Copy and Paste the Result Below:

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

| city | total\_review |

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

| Las Vegas | 1561 |

| Phoenix | 1001 |

| Toronto | 985 |

| Scottsdale | 497 |

| Charlotte | 468 |

| Pittsburgh | 353 |

| Montréal | 337 |

| Mesa | 304 |

| Henderson | 274 |

| Tempe | 261 |

| Edinburgh | 239 |

| Chandler | 232 |

| Cleveland | 189 |

| Gilbert | 188 |

| Glendale | 188 |

| Madison | 176 |

| Mississauga | 150 |

| Stuttgart | 141 |

| Peoria | 105 |

| Markham | 80 |

| Champaign | 71 |

| North Las Vegas | 70 |

| North York | 64 |

| Surprise | 60 |

| Richmond Hill | 54 |

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

(Output limit exceeded, 25 of 362 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(review\_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(review\_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(review\_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(review\_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 id,name,review\_count from business

order by review\_count desc

limit  3

Copy and Paste the Result Below:

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

| id | name | review\_count |

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

| 2weQS-RnoOBhb1KsHKyoSQ | The Buffet | 3873 |

| 0W4lkclzZThpx3V65bVgig | Schwartz's | 1757 |

| 0FUtlsQrJI7LhqDPxLumEw | Joe's Farm Grill | 1549 |

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

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

Answer:

select count(\*)

from review

where text like '%love%'

=1780

select count(\*)

from review

where text like '%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 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 |

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

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 : mesa category : food

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

* yes

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

* yes

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

* Based on the result, it can be seen that there seems to be a correlation between the location of the business and their rating, the business that are probably located in the same neighbour have close rating. Also they have similar working hours. Moreover, the business that have longer working hours usually have higher rating.

SQL code used for analysis:

Select

business.name, business.city,

category.category, business.stars.

hours.hours, business.review\_Count,

business.postal\_code

from ( business inner join categoty on

business.id=category.business\_id)

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

where business.city=”mesa”

group by business.stars;

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

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.

1. Difference 1:

* The business that are still open have higher rating.

1. Difference 2:

* The business that are still open have more reviews.

SQL code used for analysis:

Select

business.is\_open,

category.category, business.stars.

hours.hours, business.review\_Count,

business.postal\_code

from ( business inner join categoty on

business.id=category.business\_id)

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

where business.city=”mesa”

group by business.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:

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

* Finding the correlation between the likes with the given rates and using like in the reviews.

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

* I need two sources of data(tables). first, I join these two tables based on users and business. The I sort them based on rating to see if there is a correlation between the number of stars and likes.
* The reason I have selected for this analysis and thus, the data sets is that psychologists have shown that how people think about something can completely change even after a few minutes and they think that how people think just after occurrence of an event is better representative for the quality of that event compared to what they say after thinking about it.
* Because tip is related to the occurrence of the event (shopping) and they write a review after hours or even days, comparing these two tables can help us to explore the validity what psychologist claim. As the result shows there is slight correlation between the number of likes and stars, but this correlation is not strong. So, what psychologists claim seems to be valid.

iii. Output of your finished dataset:

|  |
| --- |
|  |
|  | +-------+-------+ |
|  | | stars | likes | |
|  | +-------+-------+ |
|  | | 3 | 2 | |
|  | | 5 | 2 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 5 | 1 | |
|  | | 3 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | | 4 | 1 | |
|  | +-------+-------+ |
|  | (Output limit exceeded, 25 of 1227 total rows shown) |

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

* Select review.stars, tip.likes from review inner join tip on

review.user\_id = tip.user\_id

order by tip.likes desc;