Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset - Coursera

SHIVANI SHREYAS

10-20-2020

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.

```
select count(distinct(key))
from table
```

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, business_id:6493
viii. Tip = user_id:537, business_id:3979
ix. User = id:10000
x. Friend = user_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: 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:1.0 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:1 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 total_reviews
FROM business
GROUP BY city
ORDER BY total_reviews DESC
```

Copy and Paste the Result Below:

```
     city            | total_reviews |
+----------------+---------------+
| Las Vegas       |         82854 |
| Phoenix         |         34503 |
| Toronto         |         24113 |
| Scottsdale      |         20614 |
| Charlotte       |         12523 |
| Henderson       |         10871 |
| Tempe           |         10504 |
| Pittsburgh      |          9798 |
| MontrÃ©al        |          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:

i. Avon

SQL code used to arrive at answer:

```
SELECT stars AS [star rating]
     ,count(stars) AS [count]
FROM business
WHERE city = 'Avon'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns Ã¢â,¬â€œ star rating and count):

star rating	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 AS [star rating]
     ,count(stars) AS [count]
FROM business
WHERE city = 'Beachwood'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns Ã¢â,¬â€œ star rating and count):

```
+-------------+-------+
| star rating | 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?

no

Please explain your findings and interpretation of the results:

```
SELECT name
     ,review_count
     ,fans
FROM user
ORDER BY fans DESC limit 10
```

```
-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Amy       |          609 |  503 |
| Mimi      |          968 |  497 |
```

Harald	1153	311
Gerald	2000	253
Christine	930	173
Lisa	813	159
Cat	377	133
William	1215	126
Fran	862	124
Lissa	834	120

As seen in the above table, the review_count of 2000 has received only 253 fans compared to review_count of 609 who has received 503 fans. Therefore, there i sno correlation between the review_count and the number of fans.

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

Answer:

love_text	hate_text
1780	232

As seen from the output table, there are more reviews with the word love than the word hate in them.

SQL code used to arrive at answer:

```
SELECT (
      SELECT count(TEXT)
      FROM review
      WHERE TEXT LIKE "%love%"
      ) AS love_text
,(
      SELECT count(TEXT)
      FROM review
      WHERE TEXT LIKE "%hate%"
      ) AS hate_text
```

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.

I picked city Toronto and category Food for my analysis.

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

yes.

name	city	category	stars	hours
Loblaws	Toronto	Food	2.5	Saturday|8:00-22:00
Halo Brewery	Toronto	Food	4.0	Saturday|11:00-21:00
Cabin Fever	Toronto	Food	4.5	Saturday|16:00-2:00

The 4-5 stars group are open late on Saturday compared to the business group with 2-3 stars..

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

Yes.

```
+--------------+---------+----------+-------+----------------------+--
------------+
| name         | city    | category | stars | hours                |
review_count |
+--------------+---------+----------+-------+----------------------+--
------------+
| Loblaws      | Toronto | Food     |   2.5 | Saturday|8:00-22:00  |
10 |
| Halo Brewery | Toronto | Food     |   4.0 | Saturday|11:00-21:00 |
15 |
| Cabin Fever  | Toronto | Food     |   4.5 | Saturday|16:00-2:00  |
26 |
+--------------+---------+----------+-------+----------------------+--
------------+
```

The 4-5 stars group have higher number of reviews compared to the businesses with 2 -3 stars.

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

No. All the business in the 2 groups are in different locations.

SQL code used for analysis:

```
SELECT b.name
     ,b.city
     ,b.address
     ,b.neighborhood
     ,b.postal_code
     ,c.category
     ,b.stars
     ,h.hours
     ,b.review_count
FROM (
     business b INNER JOIN category c ON b.id = c.business_id
     )
INNER JOIN hours h ON b.id = h.business_id
WHERE b.city = "Toronto"
     AND c.category = "Food"
GROUP BY b.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.

i. Difference 1:

The Businesses that are open tend to have higher star ratings on average compared to businesses that are closed.

Open: avg(stars) = 3.819
Closed: avg(stars) = 3.637

ii. Difference 2:

The Businesses that are open tend to have higher reviews on average compared to businesses that are closed.

Open: avg(review_count) = 31.75
Closed: avg(review_Count) = 23.19

SQL code used for analysis:

```
SELECT COUNT(DISTINCT(id)), AVG(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 of the following:

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

I chose to do an analysis on the user preferences with different kinds of cuisine on Yelp.

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

I want to analyze what kind of cuisine has higher demand on yelp based on the ratings and review count and what other factors affect that particular user preferences.
For my analysis, I chose cuisine categories in Chinese, Mexican, Barbeque, Italian, Korean, Japanese and
Indian.

iii. Output of your finished dataset:

category	Number_Of_Resturants	city	AVG(stars)	AVG(review_count)
Korean	2	Cuyahoga Falls	4.25	31.5
Japanese	5	Las Vegas	3.8	30.4
Barbeque	2	Phoenix	3.75	252.5
Indian	5	Edinburgh	3.6	12.6
Italian	2	Montréal	3.5	74.0
Mexican	7	Tolleson	3.5	46.7142857143
Chinese	4	Edinburgh	3.125	199.0

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

```
SELECT c.category
      ,COUNT(b.name) AS Number_Of_Resturants
      ,b.city
      ,AVG(stars)
      ,AVG(review_count)
FROM (
      business b INNER JOIN category c ON c.business_id = b.id
      )
WHERE c.category IN (
            "Chinese"
            ,"Mexican"
            ,"Barbeque"
            ,"Italian"
            ,"Korean"
            ,"Japanese"
            ,"Indian"
            )
GROUP BY c.category
ORDER BY AVG(stars) DESC
```

---------------------------------END---------------------------------