Course 1: Yelp Dataset SQL Lookup

Part 1

Answer 1:

Profile the data by finding the total number of records for each of the tables below:

i. Attribute table = 10,000

ii. Business table = 10,000

iii. Category table = 10,000

iv. Checkin table = 10,000

v. elite\_years table = 10,000

vi. friend table = 10,000

vii. hours table = 10,000

viii. photo table = 10,000

ix. review table = 10,000

x. tip table = 10,000

xi. user table = 10,000

SQL Code:

SELECT \*

FROM business

## I just kept changing the table name in the above code and the total number of records were written at the end of the table.

Answer 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 = 10,000

## primary key is id and the following code used to generate the answer

SELECT DISTINCT id

FROM business

ii. Hours = 1562

## foreign key business\_id used to generate the above answer and the following is the code

SELECT DISTINCT business\_id

FROM hours

iii. Category = 2643

## foreign key business\_id used to generate the above answer and the following is the code

SELECT DISTINCT business\_id

FROM category

iv. Attribute =1115

## foreign key business\_id used to generate the above answer and the following is the code

SELECT DISTINCT business\_id

FROM Attribute

v. Review =10000

## primary key is id and the following code used to generate the answer

SELECT DISTINCT id

FROM review

vi. Checkin = 493

## foreign key business\_id used to generate the above answer and the following is the code

SELECT DISTINCT business\_id

FROM checkin

vii. Photo = 10000

## primary key is id and the following code used to generate the answer

SELECT DISTINCT id

FROM photo

viii. Tip = 537

## foreign key user\_id used to generate the above answer and the following is the code

SELECT DISTINCT user\_id

FROM tip

ix. User = 10000

## primary key is id and the following code used to generate the answer

SELECT DISTINCT id

FROM user

x. Friend = 11

## foreign key user\_id used to generate the above answer and the following is the code

SELECT DISTINCT user\_id

FROM friend

xi. Elite\_years = 2780

## foreign key user\_id used to generate the above answer and the following is the code

SELECT DISTINCT user\_id

FROM elite\_years

Answer 3

Are there any columns with null values in the Users table? Indicate "yes," or "no."

No, none of the columns in the user table have null values.

The result in the table showed zero rows.

SQL code used to arrive at answer:

SELECT name

FROM user

WHERE name IS NULL

## I just changed the column names in the ‘where’ condition to check for all columns. Ww can also use OR condition the the ‘where’ clause to arrive at this answer.

Answer 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

SELECT MIN (stars) as Minvalue, MAX(stars) as Maxvalue, AVG(stars) AS Averagevalue

FROM Review

ii. Table: Business, Column: Stars

min: 1.0 max: 5.0 avg: 3.6549

SELECT MIN (stars) as Minvalue, MAX(stars) as Maxvalue, AVG(stars) AS Averagevalue

FROM Business

iii. Table: Tip, Column: Likes

min: 0 max: 2 avg: 0.0144

SELECT MIN (likes) as Minvalue, MAX(likes) as Maxvalue, AVG(likes) AS Averagevalue

FROM Tip

iv. Table: Checkin, Column: Count

min: 1 max: 53 avg: 1.9414

SELECT MIN (count) as Minvalue, MAX(count) as Maxvalue, AVG(count) AS Averagevalue

FROM Checkin

v. Table: User, Column: Review\_count

min: 0 max: 2000 avg: 24.2995

SELECT MIN (review\_count) as Minvalue, MAX(review\_count) as Maxvalue, AVG(review\_count) AS Averagevalue

FROM User

Answer 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

FROM Business

GROUP BY City

ORDER BY total DESC

Copy and Paste the Result Below:

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

| city | total |

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

| 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 |

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

Answer 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, review\_count

FROM Business

WHERE City = 'Avon'

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

stars | review\_count |

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

| 2.5 | 3 |

| 4.0 | 4 |

| 5.0 | 3 |

| 3.5 | 7 |

| 1.5 | 10 |

| 3.5 | 31 |

| 4.5 | 31 |

| 3.5 | 50 |

| 2.5 | 3 |

| 4.0 | 17 |

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

ii. Beachwood

SQL code used to arrive at answer:

SELECT Stars, review\_count

FROM Business

WHERE City = 'Beachwood'

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

stars | review\_count |

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

| 3.0 | 8 |

| 3.0 | 3 |

| 4.5 | 14 |

| 5.0 | 6 |

| 4.0 | 69 |

| 4.5 | 3 |

| 5.0 | 4 |

| 2.0 | 8 |

| 3.5 | 3 |

| 3.5 | 3 |

| 5.0 | 6 |

| 2.5 | 3 |

| 5.0 | 3 |

| 5.0 | 4 |

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

Answer 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

Copy and Paste the Result Below:

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

| name | review\_count |

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

| Gerald | 2000 |

| Sara | 1629 |

| Yuri | 1339 |

| .Hon | 1246 |

| William | 1215 |

| Harald | 1153 |

| eric | 1116 |

| Roanna | 1039 |

| Mimi | 968 |

| Christine | 930 |

| Ed | 904 |

| Nicole | 864 |

| Fran | 862 |

| Mark | 861 |

| Christina | 842 |

| Dominic | 836 |

| Lissa | 834 |

| Lisa | 813 |

| Alison | 775 |

| Sui | 754 |

| Tim | 702 |

| L | 696 |

| Angela | 694 |

| Crissy | 676 |

| Lyn | 675 |

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

Answer 8:

Does posing more reviews correlate with more fans?

SELECT Name, review\_count, fans

FROM user

ORDER BY review\_count DESC

Please explain your findings and interpretation of the results:

It is clearly seen that posing more reviews does not relate to more fans. For example Harald has 1153 reviews and 311 fans where as Gerald has 2000 reviews and 253 fans.

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

There are more reviews for the word ‘love’ in them.

SQL code used to arrive at answer:

SELECT COUNT (text)

FROM review

WHERE text LIKE '%love%'

## the result of the above code is 1780

SELECT COUNT (text)

FROM review

WHERE text LIKE '%hate%'

## the result of the above code is 232

Answer 10: 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

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 |

| Mark | 115 |

| Tiffany | 111 |

| bernice | 105 |

| Roanna | 104 |

| Angela | 101 |

| .Hon | 101 |

| Ben | 96 |

| Linda | 89 |

| Christina | 85 |

| Jessica | 84 |

| Greg | 81 |

| Nieves | 80 |

| Sui | 78 |

| Yuri | 76 |

| Nicole | 73 |

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

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

Part 2:

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: Las Vegas

Category: Shopping

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

Yes, the distribution of hours is differnt.

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

Yes, the number of reviews are different.

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

There is a relation between the location and the rating

SQL code used for analysis:

SELECT business.name, business.city, business.stars, business.review\_count, business.address, category.category, hours.hours

FROM business

INNER JOIN category ON (category.business\_id= business.id)

INNER JOIN hours ON (hours.business\_id=business.id)

WHERE City = 'Las Vegas'

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.

i. Difference 1:

The business that has closed has more number of review than the one that is open. The closed business has 168 reviews and the business that is open has 4 reviews.

ii. Difference 2:

The rating of the business that is closed is lower than the rating of the business the is open. The closed business has a rating of 4 whereas the business that is open has a rating of 5.

SQL code used for analysis:

SELECT business.name, business.stars, business.review\_count, business.address, business.is\_open, category.category, hours.hours

FROM business

INNER JOIN category ON (category.business\_id= business.id)

INNER JOIN hours ON (hours.business\_id=business.id)

WHERE City = 'Las Vegas'

GROUP BY business.is\_open

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

| name | stars | review\_count | address | is\_open | category | hours |

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

| Jacques Cafe | 4.0 | 168 | 1910 Village Center Cir, Unit 1 | 0 | Gluten-Free | Saturday|11:00-20:00 |

| Desert Medical Equipment | 5.0 | 4 | 3555 W Reno Ave, Ste F | 1 | Shopping | Monday|8:00-17:00 |

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

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:

How does the number of working hours depend on the number of reviews and the star rating of the business that are currently open. Also, what is the observation on the category of the business.

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

To analyse the number of reviews and the star rating of a business, it’s important to have the data on business, number of reviews and the stars. Also, data on number of working hours for each business is required.

To know which category does the business belong to the category data is also needed. All these analysis are done for the businesses that are open.

iii. Output of your finished dataset:

Observation 1: The less number of working hours leads to higher reviews.

Observation 2: All business that are open during the evening hours have a higher star rating. They all have a rating of 3.5 and above out of 5.

Observation 3: Higher number of working hours leads to lesser reviews and lower star rating. For example, McDonald’s has working hours from 5:00am and to 12:00am and the reviews are 8 and the star rating is 2.0 out of 5.

Observation 4: It is also observed that Bakeries, Restaurants, fast food joints, beauty & Spas are mostly open on weekends hence its possible that the sale is more on weekends.

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

SELECT business.name, business.stars, business.review\_count, category.category, hours.hours

FROM business

INNER JOIN category ON (category.business\_id= business.id)

INNER JOIN hours ON (hours.business\_id=business.id)

WHERE is\_open = '1'

GROUP BY stars

ORDER BY review\_count DESC

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

| name | stars | review\_count | category | hours |

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

| Edulis | 4.0 | 89 | Restaurants | Saturday|18:00-23:00 |

| Papa Da Vinci | 2.5 | 28 | Pizza | Saturday|11:00-3:00 |

| Poutine Lafleur | 3.5 | 11 | Restaurants | Saturday|11:00-19:00 |

| Cardiac Solutions | 3.0 | 9 | Doctors | Monday|8:00-16:30 |

| McDonald's | 2.0 | 8 | Fast Food | Saturday|5:00-0:00 |

| Fresh Bonsai Nails & Spa | 1.5 | 7 | Beauty & Spas | Saturday|9:30-18:00 |

| Oaks Golf Course | 4.5 | 5 | Active Life | Saturday|7:00-19:00 |

| Red Apron Bakeshop | 5.0 | 5 | Bakeries | Saturday|9:00-17:00 |

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