**SQL For Data Science (Profiling and Analyzing)**

**Udayakumar Routhu**

**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

Including code I have executed against:

Select count(\*) as num\_of\_records from Attribute;

Note: I ran similar for other tables

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 = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10,000

viii. Tip = 537 #specified user\_id

ix. User = 10000

x. Friend = 11

xi. Elite\_years = 2780

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

Code: select count(distinct id) from Business ;

Note: I ran similar for other tables

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

Answer: Zero rows in the answer shows that there is no null values in the User table

SQL code used to arrive at answer:

select id, name, review\_count, yelping\_since, useful, funny, cool, fans, average\_stars, compliment\_hot, compliment\_more, compliment\_profile, compliment\_cute, compliment\_list, compliment\_note, compliment\_plain, compliment\_cool, compliment\_funny, compliment\_writer, compliment\_photos from user where id = NULL or name = NULL or review\_count = NULL or yelping\_since = NULL or useful = NULL or funny = NULL or cool = NULL or fans= NULL or average\_stars= NULL or compliment\_hot= NULL or compliment\_more= NULL or compliment\_profile= NULL or compliment\_cute= NULL or compliment\_list= NULL or compliment\_note= NULL or compliment\_plain = NULL or compliment\_cool= NULL or compliment\_funny= NULL or compliment\_writer= NULL or compliment\_photos= 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

**code:** select min(stars),max(stars), avg(stars) from review;

Note: I ran similar for other tables

**5. List the cities with the most reviews in descending order**:

SQL code used to arrive at answer:

select city, count(review\_count) as most\_review

from business group by city order by most\_review desc;

Copy and Paste the Result Below:

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

| city | most\_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(\*) as distribution from business

where city='Avon' group by stars ;

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

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

| stars | distribution |

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

| 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(\*) as distribution from business

where city='Beachwood' group by stars ;

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

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

| stars | distribution |

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

| 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

, id

, review\_count

from user

order by review\_count desc LIMIT 3;

Copy and Paste the Result Below:

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

| name | id | review\_count |

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

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 2000 |

| Sara | -3s52C4zL\_DHRK0ULG6qtg | 1629 |

| Yuri | -8lbUNlXVSoXqaRRiHiSNg | 1339 |

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

8. **Does posing more reviews correlate with more fans?**

Please explain your findings and interpretation of the results:

Ans: Based on the provided output, we can observe that some users who have posted more reviews also tend to have more fans, but this does not necessarily imply a direct correlation between the two factors for all users.

Output:

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

| name | id | review\_count | fans |

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

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 2000 | 253 |

| Sara | -3s52C4zL\_DHRK0ULG6qtg | 1629 | 50 |

| Yuri | -8lbUNlXVSoXqaRRiHiSNg | 1339 | 76 |

| .Hon | -K2Tcgh2EKX6e6HqqIrBIQ | 1246 | 101 |

| William | -FZBTkAZEXoP7CYvRV2ZwQ | 1215 | 126 |

| Harald | --2vR0DIsmQ6WfcSzKWigw | 1153 | 311 |

| eric | -gokwePdbXjfS0iF7NsUGA | 1116 | 16 |

| Roanna | -DFCC64NXgqrxlO8aLU5rg | 1039 | 104 |

| Mimi | -8EnCioUmDygAbsYZmTeRQ | 968 | 497 |

| Christine | -0IiMAZI2SsQ7VmyzJjokQ | 930 | 173 |

| Ed | -fUARDNuXAfrOn4WLSZLgA | 904 | 38 |

| Nicole | -hKniZN2OdshWLHYuj21jQ | 864 | 43 |

| Fran | -9da1xk7zgnnfO1uTVYGkA | 862 | 124 |

| Mark | -B-QEUESGWHPE\_889WJaeg | 861 | 115 |

| Christina | -kLVfaJytOJY2-QdQoCcNQ | 842 | 85 |

| Dominic | -kO6984fXByyZm3\_6z2JYg | 836 | 37 |

| Lissa | -lh59ko3dxChBSZ9U7LfUw | 834 | 120 |

| Lisa | -g3XIcCb2b-BD0QBCcq2Sw | 813 | 159 |

| Alison | -l9giG8TSDBG1jnUBUXp5w | 775 | 61 |

| Sui | -dw8f7FLaUmWR7bfJ\_Yf0w | 754 | 78 |

| Tim | -AaBjWJYiQxXkCMDlXfPGw | 702 | 35 |

| L | -jt1ACMiZljnBFvS6RRvnA | 696 | 10 |

| Angela | -IgKkE8JvYNWeGu8ze4P8Q | 694 | 101 |

| Crissy | -hxUwfo3cMnLTv-CAaP69A | 676 | 25 |

| Lyn | -H6cTbVxeIRYR-atxdielQ | 675 | 45 |

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

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

Answer: total number of reviews 10000, out of that only got 1958 reviews with the word “love” or “hate”

However only love reviews are 1780 and only hate reviews are 232 so compared hate hate, love reviews are more

select

count(text) as "love\_or\_hate\_reviews" from review where text LIKE '%love%' ;

select

count(text) as "love\_or\_hate\_reviews" from review where text LIKE '%hate%' ;

SQL code used to arrive at answer:

select

id,count(text) from review where text LIKE '%love%' OR text LIKE '%hate%';

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

| love\_or\_hate\_reviews |

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

| 1958 |

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

10. **Find the top 10 users with the most fans:**

SQL code used to arrive at answer:

Select name, sum(fans) total from user

group by name order by total DESC LIMIT 10;

Copy and Paste the Result Below:

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

| name | total |

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

| Amy | 519 |

| Mimi | 498 |

| Harald | 311 |

| Gerald | 256 |

| Lisa | 207 |

| Nicole | 200 |

| Christine | 187 |

| Mark | 156 |

| Jen | 148 |

| Linda | 148 |

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

**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.

Category='Shopping'

City = 'Las Vegas'

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

Ans: Yes

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

Ans: Yes

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

Businesses with higher star ratings (4.5 and 5.0) seem to be more popular and have longer hours of operation compared to those with lower star ratings (2.5 and 3.5). Customers may prefer businesses with higher ratings and more reviews, leading to increased popularity and longer

SQL code used for analysis:

select city,stars,count(\*) as total\_businesses,

SUM(review\_count) as total\_review\_business,inr.hours

from business out

JOIN

(select a.business\_id,category,hours

from category a

JOIN hours b

ON a.business\_id = b.business\_id

where a.category = 'Shopping') inr

ON out.id = inr.business\_id

WHERE out.city='Las Vegas'

GROUP BY city,stars,inr.hours;

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:

Open businesses (8,480) outnumber closed businesses (1,520) by a considerable margin. This suggests that there is a higher demand for services provided by open businesses compared to closed ones. Customers may have more options and greater accessibility to services offered by open businesses.

ii. Difference 2:

The large disparity in the number of open and closed businesses indicates the economic activity within the area. Higher numbers of open businesses suggest a vibrant and thriving economy, while a significant number of closed businesses may indicate economic challenges or shifts in market dynamics. This difference in business status could have implications for employment rates, local spending, and overall economic health.

SQL code used for analysis:

select

 CASE

    WHEN is\_open = 1 THEN 'OPEN'

    ELSE 'CLOSED'

 END as business\_status,

 count(\*) as total\_business

 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:

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

Ans: predicting the overall star rating for business

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

**Trend Analysis:** You can observe how the average review count and average is\_open status change as the overall star rating increases. This can help identify trends in customer behavior and business performance.

**Customer Behavior Analysis**: Analyzing the relationship between review counts and star ratings can provide insights into customer behavior and preferences. Businesses with higher star ratings may attract more reviews, indicating a higher level of customer engagement.

**Impact of Open Status**: Analyzing the average is\_open status across different star ratings can reveal how business openness correlates with star ratings. For example, businesses with higher star ratings may have a higher likelihood of being open.

iii. Output of your finished dataset:

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

| avg\_review\_count | avg\_is\_open | overall\_star\_rating |

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

| 6.160256410256411 | 0.9102564102564102 | 1.0 |

| 12.922330097087379 | 0.883495145631068 | 1.5 |

| 13.816254416961131 | 0.833922261484099 | 2.0 |

| 18.885393258426966 | 0.8112359550561797 | 2.5 |

| 25.999283667621775 | 0.8051575931232091 | 3.0 |

| 40.115298087739035 | 0.8340832395950506 | 3.5 |

| 48.40099750623441 | 0.8374064837905237 | 4.0 |

| 37.50904033379694 | 0.868567454798331 | 4.5 |

| 11.315015974440895 | 0.9118210862619809 | 5.0 |

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

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

SELECT

    AVG(review\_count) AS avg\_review\_count,

    AVG(is\_open) AS avg\_is\_open,

    stars AS overall\_star\_rating

FROM

    business

GROUP BY

    overall\_star\_rating;