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

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 = 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

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

ii. Hours = 1,562

iii. Category = 2,643

iv. Attribute = 1,115

v. Review = 10,000

vi. Checkin = 493

vii. Photo = 10,000

viii. Tip = 537

ix. User = 10,000

x. Friend = 11

xi. Elite\_years = 2,780

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: No

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 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:2,000 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 NUM

FROM business

GROUP BY city

ORDER BY NUM DESC

Copy and Paste the Result Below:

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

| city | NUM |

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

| 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 |

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

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

i. Avon

SQL code used to arrive at answer:

SELECT SUM(review\_count)

AS Numbers, stars FROM business WHERE city == "Avon"

GROUP BY stars

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

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

| Numbers | stars |

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

| 10 | 1.5 |

| 6 | 2.5 |

| 88 | 3.5 |

| 21 | 4.0 |

| 31 | 4.5 |

| 3 | 5.0 |

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

ii. Beachwood

SQL code used to arrive at answer:

SELECT SUM(review\_count) AS Numbers, stars FROM business

WHERE city == "Beachwood"

GROUP BY stars

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

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

| Numbers | stars |

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

| 8 | 2.0 |

| 3 | 2.5 |

| 11 | 3.0 |

| 6 | 3.5 |

| 69 | 4.0 |

| 17 | 4.5 |

| 23 | 5.0 |

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

7. Find the top 3 users based on their total number of reviews:

SQL code used to arrive at answer:

SELECT review\_count, name

FROM user

ORDER BY review\_count DESC

LIMIT 3

Copy and Paste the Result Below:

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

| review\_count | name |

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

| 2000 | Gerald |

| 1629 | Sara |

| 1339 | Yuri |

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

No, posing more reviews does not correlate with more fans. There are some individuals with a higher number of reviews but when comparing the number of fans to other individuals with less reviews, they do not have the higher number of fans.

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” in them.

SQL code used to arrive at answer:

SELECT COUNT(\*)

FROM review

WHERE text LIKE "%love%"

SELECT COUNT(\*)

FROM review

WHERE text LIKE "%love%"

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

Yes the two groups I chose to analyze have different distribution of hours.

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

Yes the two groups I chose to analyze have a different number of reviews.

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

I was not able to infer anything from the location data provided between the two groups.

SQL code used for analysis:

SELECT business.name

, business.city

, category.category

, business.stars

,hours.hours

, business.review\_count

, business.address

, business.postal\_code

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

WHERE business.city = 'Las Vegas' AND category.category = "Food"

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: There are more businesses listed with useful or funny opened in comparison to those without.

ii. Difference 2: The open businesses have higher ratings.

SQL code used for analysis:

SELECT

AVG(b.stars),SUM(b.review\_count),AVG(b.review\_count),COUNT(r.use ful)+COUNT(r.funny),is\_open FROM business b INNER JOIN review r ON b.id = r.id

GROUP BY b.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:

The analysis I chose to do was to determine where businesses were located and how many were in the same cities.

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

The data needed to complete this analysis will include data related to all the different businesses and their names. I will also need to access data related to the cities these business are located in.

iii. Output of your finished dataset:

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

| name | city |

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

| Las Vegas Trail Riding | nboulder city |

| Jake's Auto Repair Service | Youngtown |

| Mattress Shack | Youngtown |

| ZM Cycle & Fitness | York |

| What A Bagel | York |

| Paf & Dap Best Jerk Kitchen In Town | York |

| Raps Authentic Jamaican | York |

| Ryding Auto Body | York |

| Mainsha Restaurant | York |

| LCBO - Dundas and Jane | York |

| Corky and Lenny's | Woodmere Village |

| The Black Orchid | Woodmere |

| Eton: Chagrin Boulevard | Woodmere |

| Firehouse Subs | Woodmere |

| Bravo! Cucina Italiana | Woodmere |

| August Moon Chinese Restaurant | Woodbridge |

| Wicker Emporium | Woodbridge |

| Traction Tire | Woodbridge |

| Gran Gusto Ristorante | Woodbridge |

| Restaurant Little Mexico | Woodbridge |

| Michaels | Woodbridge |

| Roy Foss Woodbridge | Woodbridge |

| Piper Studios | Woodbridge |

| Cameo Lounge | Woodbridge |

| Winners | Woodbridge |

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

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

SELECT name

,city

FROM business

ORDER BY city DESC