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:
   1. Attribute table = 10000
   2. Business table = 10000
   3. Category table = 10000
   4. Checkin table = 10000
   5. elite\_years table = 10000
   6. friend table = 10000
   7. hours table = 10000
   8. photo table = 10000
   9. review table = 10000
   10. tip table = 10000
   11. user table = 10000

\*\*\*\*\*\*\*\*\*\*\*SQL CODE\*\*\*\*\*\*\*\*\*\* SELECT COUNT(\*)

FROM table

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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.
   1. Business = id: 10000
   2. Hours = business\_id: 1562
   3. Category = business\_id: 2643
   4. Attribute = business\_id: 1115
   5. Review = id: 10000, business\_id: 8090, user\_id: 9581
   6. Checkin = business\_id: 493
   7. Photo = id: 10000, photo: 6493
   8. Tip = user\_id: 537, business\_id: 3979
   9. User = id: 10000
   10. Friend = user\_id: 11
   11. Elite\_years = user\_id: 2780

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

\*\*\*\*\*\*\*\*\*\*\*SQL CODE\*\*\*\*\*\*\*\*\*\* SELECT COUNT(DISTINCT Keys)

FROM table

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. 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

1. 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

min: 1 max: 5 avg: 3.7082

* 1. Table: Business, Column: Stars min: 1.0 max: 5.0 avg: 3.6549
  2. Table: Tip, Column: Likes

min: 0 max: 2 avg: 0.0144

* 1. Table: Checkin, Column: Count min: 1 max: 53 avg: 1.9414
  2. Table: User, Column: Review\_count min: 0 max: 2000 avg: 24.2995

\*\*\*\*\*\*\*\*\*\*\*SQL CODE\*\*\*\*\*\*\*\*\*\*

SELECT MIN(Column),MAX(Column),AVG(Column) FROM table

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. 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 | | |

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

(Output limit exceeded, 25 of 362 total rows shown)

1. Find the distribution of star ratings to the business in the following cities:
   1. 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 – star rating and count):

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

| Numbers | stars |

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

| | | 10 | | | 1.5 | | |
| --- | --- | --- | --- | --- |
| | | 6 | | | 2.5 | | |
| | | 88 | | | 3.5 | | |
| | | 21 | | | 4.0 | | |
| | | 31 | | | 4.5 | | |
| | | 3 | | | 5.0 | | |

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

* 1. 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 – 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 | | |

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

1. 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 |

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

1. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

Not necessarily correlated. Amy, who has the most fans, only has 609 reviews. Yuri has only 76 fans, but has the third most reviews. Some other factors should also be considered.

\*\*\*\*\*\*\*\*\*\*\*SQL CODE\*\*\*\*\*\*\*\*\*\* SELECT name,review\_count,fans FROM user

ORDER BY fans DESC

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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

| 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 | | |
| | | Mark | | | 861 | | | 115 | | |
| | | Tiffany | | | 408 | | | 111 | | |
| | | bernice | | | 255 | | | 105 | | |
| | | Roanna | | | 1039 | | | 104 | | |
| | | Angela | | | 694 | | | 101 | | |
| | | .Hon | | | 1246 | | | 101 | | |
| | | Ben | | | 307 | | | 96 | | |
| | | Linda | | | 584 | | | 89 | | |
| | | Christina | | | 842 | | | 85 | | |
| | | Jessica | | | 220 | | | 84 | | |
| | | Greg | | | 408 | | | 81 | | |
| | | Nieves | | | 178 | | | 80 | | |
| | | Sui | | | 754 | | | 78 | | |
| | | Yuri | | | 1339 | | | 76 | | |
| | | Nicole | | | 161 | | | 73 | | |

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

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

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

Answer: Yes. There are 1780 reviews with “love” and 232 reviews with “hate”

SQL code used to arrive at answer: SELECT COUNT(\*)

FROM review

WHERE text LIKE "%love%"

SELECT COUNT(\*)

FROM review

WHERE text LIKE “%hate%”

1. 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 | | |

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

1. Is there a strong relationship (or correlation) between having a high number of fans and being listed as "useful" or "funny?" Out of the top 10 users with the highest number of fans, what percent are also listed as “useful” or “funny”?

Key:

0% - 25% - Low relationship

26% - 75% - Medium relationship

76% - 100% - Strong relationship

SQL code used to arrive at answer:

SELECT name,fans,useful,funny FROM user

ORDER BY fans DESC LIMIT 10

Copy and Paste the Result Below:

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

| name | fans | useful | funny |

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

| | | Amy | | | 503 | | | 3226 | | | 2554 | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | Mimi | | | 497 | | | 257 | | | 138 | | |
| | | Harald | | | 311 | | | 122921 | | | 122419 | | |
| | | Gerald | | | 253 | | | 17524 | | | 2324 | | |
| | | Christine | | | 173 | | | 4834 | | | 6646 | | |
| | | Lisa | | | 159 | | | 48 | | | 13 | | |
| | | Cat | | | 133 | | | 1062 | | | 672 | | |
| | | William | | | 126 | | | 9363 | | | 9361 | | |
| | | Fran | | | 124 | | | 9851 | | | 7606 | | |
| | | Lissa | | | 120 | | | 455 | | | 150 | | |

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

Please explain your findings and interpretation of the results:

Out of the top 10 users with the highest number of fans, 100% are also listed as either “useful” or “funny”. So there is a strong correlation between having a high number of fans and being listed as “useful” or “funny”.

# 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.

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

I picked Toronto and Food for this question. Yes. The restaurants with only 2.5 stars open from 8:00-22:00 on Saturday. The places with higher rating stars open late on Saturday.

The results from SQL is shown below:

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

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

| 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 |

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

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

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

Yes. The group with 2-3 stars has less review (10) compared with the group with higher rating stars.

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

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

| 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 |

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

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

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

No really. They have different locations.

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

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

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

| name | city | category | stars | hours

| review\_count | address | postal\_code |

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

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

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

| Loblaws | Toronto | Food | 2.5 | Saturday| 8:00-22:00 | 10 | 2280 Dundas Street W | M6R 1X3

|

| Halo Brewery | Toronto | Food | 4.0 | Saturday| 11:00-21:00 | 15 | 247 Wallace Avenue | M6H 1V5

|

| Cabin Fever | Toronto | Food | 4.5 | Saturday| 16:00-2:00 | 26 | 1669 Bloor Street W | M6P 1A6

|

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

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

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

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 = 'Toronto' AND category.category = "Food" GROUP BY business.stars;

1. 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 ones that are still open have more reviews on average than ones that are closed

* 1. Difference 2:

There are more business that are still open listed as “useful” or “funny”.

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

Results:

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

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

| AVG(b.stars) | SUM(b.review\_count) | AVG(b.review\_count) | COUNT(r.useful)+COUNT(r.funny) | is\_open |

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

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

| 2.0 | 4 | 4.0 |

2 | 0 |

| 2.96153846154 | 504 | 38.7692307692 |

26 | 1 |

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

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

1. 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:

Here I chose to study the preference among different types of food on yelp.

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

I will pick several types of food including “Chinese”,”Mexican”,”Korean”,”French”,”Italian”,”Japanese” and “Indian”. Then I will analyze their star ratings and number of reviews so that I can get some insights on which type of food is popular on yelp.

* 1. Output of your finished dataset:

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

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

| category | Number\_Of\_Resturants | AVG(stars) | AVG(review\_count) | city |

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

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

| Korean | 7 | 4.5 |

| 8.0 | Toronto |  | | |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | French | |  |  |  |  | 12 | | | 4.0 | | |
| 135.083333333 | | | Las | Vegas | | |  |  |  |  |
| | Chinese | |  |  |  |  | 13 | | | 3.76923076923 | | |
| 423.230769231 | | | Las | Vegas | | |  |  |  |  |
| | Mexican | |  |  |  |  | 28 | | | 3.625 | | |
| 73.0 | Edinburgh | | | | |  |  |  |  |  |
| | Italian | | |  | |  | 13 | | | 3.53846153846 | | |
| 78.2307692308 | | | Montréal | | | |  |  |  |  |
| | Indian | | |  | |  | 6 | | | 3.5 | | |
| 32.0 | Aurora | | | | |  |  |  |  |  |
| | Japanese | | |  | |  | 20 | | | 3.475 | | |

22.85 | Toronto |

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

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

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

SELECT c.category,COUNT(b.name) AS Number\_Of\_Resturants,AVG(stars),AVG(review\_count),b.city FROM (business b INNER JOIN hours h ON b.id = h.business\_id) INNER JOIN category c ON c.business\_id = b.id

WHERE c.category IN ("Chinese","Mexican","French","Italian","Korean","Japanese","Ind ian")

GROUP BY c.category ORDER BY AVG(stars) DESC