José María Martínez Marín

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

This is a 2-part assignment. In the first part, you are asked a series of questions that will help you profile and understand the data just like a data scientist would. For this first part of the assignment, you will be assessed both on the correctness of your findings, as well as the code you used to arrive at your answer. You will be graded on how easy your code is to read, so remember to use proper formatting and comments where necessary.

In the second part of the assignment, you are asked to come up with your own inferences and analysis of the data for a particular research question you want to answer. You will be required to prepare the dataset for the analysis you choose to do. As with the first part, you will be graded, in part, on how easy your code is to read, so use proper formatting and comments to illustrate and communicate your intent as required.

For both parts of this assignment, use this "worksheet." It provides all the questions you are being asked, and your job will be to transfer your answers and SQL coding where indicated into this worksheet so that your peers can review your work. You should be able to use any Text Editor (Windows Notepad, Apple TextEdit, Notepad ++, Sublime Text, etc.) to copy and paste your answers. If you are going to use Word or some other page layout application, just be careful to make sure your answers and code are lined appropriately.

In this case, you may want to save as a PDF to ensure your formatting remains intact for you reviewer.

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.

i. Business = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 537 (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.

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 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

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

SQL code used to arrive at answer:

```
SELECT City, SUM(review_count) AS NoReviews
FROM business
```

```
GROUP BY City
ORDER BY NoReviews DESC
```

Copy and Paste the Result Below:

city	NoReviews
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 stars, SUM(review_count) as Sum
FROM business
WHERE City = 'Avon'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns – star rating and count):

stars	Sum
1.5	10
2.5	6
3.5	88
4.0	21
4.5	31
5.0	3

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars, SUM(review_count) as Sum
FROM business
WHERE City = 'Beachwood'
GROUP BY stars
```

Copy and Paste the Resulting Table Below (2 columns – star rating and count):

stars	Sum
2.0	8
2.5	3
3.0	11
3.5	6
4.0	69
4.5	17
5.0	23

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

SQL code used to arrive at answer:

```
SELECT id, Name, SUM(review_count) AS NoReviews
FROM user
GROUP BY id
ORDER BY NoReviews DESC
LIMIT 3
```

Copy and Paste the Result Below:

```
+------------------------+--------+-----------+
| id                     | name   | NoReviews |
+------------------------+--------+-----------+
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald |      2000 |
| -3s52C4zL_DHRK0ULG6qtg | Sara   |      1629 |
| -8lbUNlXVSoXqaRRiHiSNg | Yuri   |      1339 |
+------------------------+--------+-----------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

Yes, and also with how long have they been yelping since because, as it can be seen, the longer one user has been yelping, the more fans he/she has, and less directly correlated, the more reviews one receives.

CODE:

```
SELECT id, name, fans, review_count, yelping_since
FROM user
ORDER BY fans DESC
```

OUTPUT:

```
+------------------------+-----------+------+--------------+---------------------+
| id                     | name      | fans | review_count | yelping_since       |
+------------------------+-----------+------+--------------+---------------------+
```

| -9I98YbNQnLdAmcYfb324Q | Amy | 503 | 609 | 2007-07-19 00:00:00 |

| -8EnCioUmDygAbsYZmTeRQ | Mimi | 497 | 968 | 2011-03-30 00:00:00 |

| --2vR0DIsmQ6WfcSzKWigw | Harald | 311 | 1153 | 2012-11-27 00:00:00 |

| -G7Zkl1wIWBBmD0KRy_sCw | Gerald | 253 | 2000 | 2012-12-16 00:00:00 |

| -0IiMAZI2SsQ7VmyzJjokQ | Christine | 173 | 930 | 2009-07-08 00:00:00 |

| -g3XIcCb2b-BD0QBCcq2Sw | Lisa | 159 | 813 | 2009-10-05 00:00:00 |

| -9bbDysuiWeo2VShFJJtcw | Cat | 133 | 377 | 2009-02-05 00:00:00 |

| -FZBTkAZEXoP7CYvRV2ZwQ | William | 126 | 1215 | 2015-02-19 00:00:00 |

| -9da1xk7zgnnfO1uTVYGkA | Fran | 124 | 862 | 2012-04-05 00:00:00 |

| -lh59ko3dxChBSZ9U7LfUw | Lissa | 120 | 834 | 2007-08-14 00:00:00 |

| -B-QEUESGWHPE_889WJaeg | Mark | 115 | 861 | 2009-05-31 00:00:00 |

| -DmqnhW4Omr3YhmnigaqHg | Tiffany | 111 | 408 | 2008-10-28 00:00:00 |

| -cv9PPT7IHux7XUc9dOpkg | bernice | 105 | 255 | 2007-08-29 00:00:00 |

| -DFCC64NXgqrxlO8aLU5rg | Roanna | 104 | 1039 | 2006-03-28 00:00:00 |

| -IgKkE8JvYNWeGu8ze4P8Q | Angela | 101 | 694 | 2010-10-01 00:00:00 |

| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon | 101 | 1246 | 2006-07-19 00:00:00 |

| -4viTt9UC44lWCFJwleMNQ | Ben | 96 | 307 | 2007-03-10 00:00:00 |

| -3i9bhfvrM3F1wsC9XIB8g | Linda | 89 | 584 | 2005-08-07 00:00:00 |

| -kLVfaJytOJY2-QdQoCcNQ | Christina | 85 | 842 | 2012-10-08 00:00:00 |

| -ePh4Prox7ZXnEBNGKyUEA | Jessica | 84 | 220 | 2009-01-12 00:00:00 |

| -4BEUkLvHQntN6qPfKJP2w | Greg | 81 | 408 | 2008-02-16 00:00:00 |

| -C-l8EHSLXtZZVfUAUhsPA | Nieves | 80 | 178 | 2013-07-08 00:00:00 |

| -dw8f7FLaUmWR7bfJ_Yf0w | Sui | 78 | 754 | 2009-09-07 00:00:00 |

| -8lbUNlXVSoXqaRRiHiSNg | Yuri | 76 | 1339 | 2008-01-03 00:00:00 |

| -0zEEaDFIjABtPQni0XlHA | Nicole | 73 | 161 | 2009-04-30 00:00:00 |

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

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

Answer:

There are 1780 reviews with the word "love", and 232 with the word "hate"

SQL code used to arrive at answer:

For "love":

```
SELECT COUNT(text) AS NoReviews
FROM review
WHERE text LIKE '%love%'
```

For "hate":

```
SELECT COUNT(text) AS NoReviews
FROM review
WHERE text LIKE '%hate%'
```

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

SQL code used to arrive at answer:

```
SELECT id, name, fans
FROM user
ORDER BY fans DESC
LIMIT 10
```

Copy and Paste the Result Below:

id	name	fans
-9I98YbNQnLdAmcYfb324Q	Amy	503
-8EnCioUmDygAbsYZmTeRQ	Mimi	497
--2vR0DIsmQ6WfcSzKWigw	Harald	311
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	253
-0IiMAZI2SsQ7VmyzJjokQ	Christine	173
-g3XIcCb2b-BD0QBCcq2Sw	Lisa	159
-9bbDysuiWeo2VShFJJtcw	Cat	133
-FZBTkAZEXoP7CYvRV2ZwQ	William	126
-9da1xk7zgnnfO1uTVYGkA	Fran	124
-lh59ko3dxChBSZ9U7LfUw	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 have chosen as the city, Las Vegas, and as the category, restaurants.

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

Regarding the city, it can be seen that business with 2-3 stars open for actually longer than

the ones of 4-5 stars. Nevertheless, there are only 5 valid instances, therefore, no clear conclusions

can be infered from that. More data is needed.

name	stars	Sum	hours	star_rating
Walgreens	2.5	42	Saturday|8:00-22:00	2-3 stars
Wingstop	3.0	861	Saturday|11:00-0:00	2-3 stars
Hi Scores - Blue Diamond	3.5	801	Saturday|0:00-0:00	None
Anthem Pediatrics	4.0	6914	Saturday|8:00-12:00	4-5 stars
Red Rock Canyon Visitor Center	4.5	224	Saturday|8:00-16:30	4-5 stars
Desert Medical Equipment	5.0	62	Monday|8:00-17:00	4-5 stars

If we constrain our analysis to the category "restaurants", then there are only 2 cases and,

in fact, in both cases the opening hours are the same length.

stars	neighborhood	Sum	hours	star_rating
3.0		861	Saturday|11:00-0:00	2-3 stars
4.0	Chinatown	6552	Saturday|10:00-23:00	4-5 stars

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

Yes, in fact, the best rated has almost 8 times more reviews than the other.

iii. Are you able to infer anything from the location data provided between these

two groups? Explain.

No, since the location in terms of neighborhood is empty in one of the instances, which represents

50% of the total instances, so no conclusions can be made.

SQL code used for analysis:

i)

```
SELECT stars, SUM(review_count) as Sum, hours,
                                    CASE
```

```
                                        WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'
                                        WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'
                                    END AS star_rating
        FROM business B INNER JOIN hours H ON B.id = H.business_id
        WHERE City = 'Las Vegas'
        GROUP BY stars

        SELECT stars, neighborhood, SUM(review_count) as Sum, hours,
                                    CASE
                                        WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'
                                        WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'
                                    END AS star_rating
        FROM business B INNER JOIN hours H ON B.id = H.business_id
        INNER JOIN category C ON H.business_id = C.business_id
        WHERE City = 'Las Vegas'
        AND Category LIKE '%restaurant%'
        GROUP BY 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 first difference is the fact that the average number of reviews for open business is 31.75, while for the closed ones is 23.19 .

AverageReviews	AvgStars	is_open
23.1980263158	3.52039473684	0
31.7570754717	3.67900943396	1

ii. Difference 2:

The second difference is that the average number of stars for the open ones is 3.68, while for the closed ones is 3.52 .

SQL code used for analysis:

```
SELECT AVG(review_count) as AverageReviews,
AVG(stars) AS AvgStars,
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:

A predictive sentiment analysis will be carried out by creating a column of whether the review is good or bad, based on the kind of vocabulary used, as well as the number of stars and whether the business was labeled as "useful", "funny" or "cool".

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

The ID of the business will be needed, as well as the number of stars and the categories of the review. Also two extra columns for whether the review can be considered as good or bad based on the language used and the number of reviews are included. With all this information, it can be easy to determine the popularity of a business.

iii. Output of your finished dataset:

name	stars	useful	funny	cool	review_count	Reviews
Galaxy Cannery Theatre	1	1	1	1	251	None

```
        | Spinato's Pizza                          |     5 |      0 |     0 |    0 |          507 |
None |
        | Linda Woodson Dermatology                |     2 |      5 |     0 |    0 |           48 |
None |
        | 808 Sushi                                |     5 |      0 |     0 |    0 |          435 |
None |
        | Kimberfire                               |     5 |      0 |     0 |    0 |           27 |
None |
        | Herbal Nails & Spa -  - Happy Valley     |     4 |      0 |     0 |    0 |           49 |
None |
        | Woo Che                                  |     3 |      2 |     0 |    2 |          102 |
None |
        | Vanity Nails & Spa                       |     5 |      1 |     0 |    0 |          148 |
BAD |
        | Ocean Blue Caribbean Restaurant and Bar  |     5 |      0 |     0 |    0 |          140 |
None |
        | The Yard                                 |     4 |      0 |     0 |    0 |          168 |
None |
        | D & D Discount Motorcycles               |     5 |      0 |     0 |    0 |           11 |
None |
        | Toronto Don Valley Hotel and Suites      |     2 |      0 |     0 |    0 |           30 |
None |
        | El Fish Taco                             |     5 |      0 |     0 |    0 |          112 |
None |
        | Switch Restaurant & Wine Bar             |     4 |      2 |     2 |    2 |          711 |
None |
        | Chutney's Indian Cuisine                 |     2 |      1 |     0 |    0 |          240 |
None |
        | Mellow Mushroom                          |     3 |      0 |     0 |    0 |          244 |
None |
        | Michael Mina                             |     5 |      0 |     0 |    0 |          574 |
None |
        | Food Palace Gelato                       |     5 |      0 |     0 |    0 |           16 |
GOOD |
        | Pio Pio                                  |     3 |      1 |     0 |    0 |          299 |
None |
        | Pizza Taglio                             |     5 |      0 |     0 |    0 |           93 |
None |
        | Heart Bar                                |     4 |      1 |     0 |    1 |          108 |
None |
        | Hong Kong Garden Seafood & BBQ Cafe      |     3 |      0 |     0 |    0 |          147 |
None |
        | Nandini Indian Cuisine                   |     5 |      1 |     0 |    1 |          406 |
None |
        | Tortilla Fish                            |     5 |      1 |     0 |    0 |          102 |
None |
        | Greens and Proteins                      |     5 |      0 |     0 |    0 |          333 |
None |
        +------------------------------------------+-------+--------+-------+------+--------------+--
-------+
        (Output limit exceeded, 25 of 636 total rows shown)
```

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

```
        SELECT business.name, review.stars, useful, funny, cool, review_count,
                CASE
                        WHEN text LIKE '%super%' or '%interesting%' or '%great%' or '%good%' OR
'%amazing%'
                        THEN 'GOOD'
                        WHEN text LIKE '%bad%' OR '%awful%' OR '%terrible%'
                        THEN 'BAD'
                        END AS Reviews
        FROM review INNER JOIN business ON review.business_id = business.id
```