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

Answer

(i)

Code

```
SELECT count(*) as

total_records

FROM attribute;
```

Output

```
+---------------+
| total_records |
+---------------+
|       10000 |
+---------------+
```

(ii)

Code

```
SELECT count(*) as
total_records
FROM business;
```

Output

```
+---------------+
| total_records |
+---------------+
|       10000 |
+---------------+
```

(iii)

Code

```
SELECT count(*) as
total_records
FROM Category;
```

Output

```
+---------------+
| total_records |
+---------------+
|        10000 |
+---------------+
```

(iv)

Code

```
SELECT count(*) as
total_records
FROM Checkin;
```

Output

```
+---------------+
| total_records |
+---------------+
|        10000 |
+---------------+
```

(V)

Code

```
SELECT count(*) as
total_records
FROM elite_years;
```

Output

```
+---------------+
| total_records |
+---------------+
|         10000 |
+---------------+
```

(Vi)

Code

```
SELECT count(*) as
total_records
FROM friend;
```

Output

```
+---------------+
| total_records |
+---------------+
|         10000 |
+---------------+
```

(Vii)

Code

```
SELECT count(*) as
total_records
FROM hours;
```

Output

```
+---------------+
| total_records |
+---------------+
|        10000 |
+---------------+
```

(Viii)

Code

```
SELECT count(*) as
total_records
FROM photo;
```

Output

```
+---------------+
| total_records |
+---------------+
|        10000 |
+---------------+
```

(ix)

Code

```
SELECT count(*) as
total_records
FROM review;
```

Output

```
+---------------+
| total_records |
+---------------+
|        10000 |
+---------------+
```

(x)

Code

```
SELECT count(*) as
total_records
FROM tip;
```

Output

```
+---------------+
| total_records |
+---------------+
|         10000 |
+---------------+
```

(xi)

Code

```
SELECT count(*) as
total_records
FROM user;
```

Output

```
+---------------+
| total_records |
+---------------+
|         10000 |
+---------------+
```

2. Find the total number of distinct records for each of the keys listed below:

1. Business = 10,000

2. Hours = 1562

3. Category = 2643

4. Attribute = 1115

5. Review = 10,000

6. Checkin = 493

7. Photo = 10,000

8. Tip = 537

9. User = 10,000

10. Friend = 11

11. Elite_years = 2780

Answer

Sample Code

```
SELECT count(distinct name)  + count(distinct business_id)

+ count(distinct value)

AS

total_records

FROM attribute;
```

```
+---------------+

| total_records |

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

|         1115 |

+---------------+
```

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

Answer: Zero rows in the output means that there is no null values in the User table

Code

```
select  *
from user
where '*' is NULL;
```

Output

```
+----+------+-------------+--------------+--------+-------+------+------+--------------
+----------------+----------------+-------------------+-----------------+----------------
+-----------------+-----------------+----------------+-----------------+------------------
-+-------------------+

| 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 |

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

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

(Zero rows)
```

4. Find the minimum, maximum, and average 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

Answer

(i)code

```
SELECT min(stars)

,max(stars)
```

```
,avg(stars)

FROM review;
```

```
+------------+------------+------------+

| min(stars) | max(stars) | avg(stars) |

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

|          1 |          5 |     3.7082 |
```

(ii)code

```
select min(stars)
,max(stars)
,avg(stars)
from Business;
```

OUTPUT

```
+------------+------------+------------+
| min(stars) | max(stars) | avg(stars) |
+------------+------------+------------+
|        1.0 |        5.0 |     3.6549 |
+------------+------------+------------+
```

(iii)code

```
select min(Likes)
,max(Likes)
,avg(Likes)
from tip;
```

OUTPUT

```
+------------+------------+------------+
| min(Likes) | max(Likes) | avg(Likes) |
+------------+------------+------------+
|          0 |          2 |     0.0144 |
+------------+------------+------------+
```

(iv)code

```
select min(Count)
,max(Count)
,avg(Count)
from Checkin;
```

OUTPUT

```
+------------+------------+------------+
| min(Count) | max(Count) | avg(Count) |
+------------+------------+------------+
|          1 |         53 |     1.9414 |
```

(v)code

```
select min(Review_count)
,max(Review_count)
,avg(Review_count)
from user;
```

OUTPUT

```
+-------------------+-------------------+-------------------+
| min(Review_count) | max(Review_count) | avg(Review_count) |
+-------------------+-------------------+-------------------+
|                 0 |              2000 |           24.2995 |
+-------------------+-------------------+-------------------+
```

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

Answer

code

```
SELECT

city

, count(review_count) as total_review

FROM business

group by city

order by total_review desc;
```

OUTPUT

```
+-----------------+--------------+

| city            | total_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:

Answer

i. Avon

CODE

```
select
name
, stars
, review_count
from business
where city = 'Avon';
```

OUTPUT

StarRating	Count
0	0
1	0
1.5	1
2	0
2.5	2
3	1
3.5	2
4	2
4.5	1
5	1

ii. Beachwood

code

```
select
name
, stars
, review_count
from business
where city = 'Beachwood';
```

OUTPUT

```
+---------------------------------+-------+--------------+
| name                            | stars | review_count |
+---------------------------------+-------+--------------+
| Maltz Museum of Jewish Heritage |   3.0 |            8 |
| Charley's Grilled Subs          |   3.0 |            3 |
| Sixth & Pine                    |   4.5 |           14 |
| Beechmont Country Club          |   5.0 |            6 |
| Hyde Park Prime Steakhouse      |   4.0 |           69 |
| Origins                         |   4.5 |            3 |
| Fyodor Bridal Atelier           |   5.0 |            4 |
| College Planning Network        |   2.0 |            8 |
| Lucky Brand Jeans               |   3.5 |            3 |
| American Eagle Outfitters       |   3.5 |            3 |
| Shaker Women's Wellness         |   5.0 |            6 |
| Avis Rent A Car                 |   2.5 |            3 |
| Cleveland Acupuncture           |   5.0 |            3 |
| Studio Mz                      |   5.0 |          4 |

+--------------------------------+-------+--------------+
```

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

Answer

CODE

```
select

name

, id

, review_count

from user
```

```
order by review_count desc;
```

OUTPUT

```
+-----------+------------------------+--------------+
| name      | id                     | review_count |
+-----------+------------------------+--------------+
| Gerald    | -G7Zkl1wIWBBmD0KRy_sCw |         2000 |
| Sara      | -3s52C4zL_DHRK0ULG6qtg |         1629 |
| Yuri      | -8lbUNlXVSoXqaRRiHiSNg |         1339 |
| .Hon      | -K2Tcgh2EKX6e6HqqIrBIQ |         1246 |
| William   | -FZBTkAZEXoP7CYvRV2ZwQ |         1215 |
| Harald    | --2vR0DIsmQ6WfcSzKWigw |         1153 |
| eric      | -gokwePdbXjfS0iF7NsUGA |         1116 |
| Roanna    | -DFCC64NXgqrxlO8aLU5rg |         1039 |
| Mimi      | -8EnCioUmDygAbsYZmTeRQ |          968 |
| Christine | -0IiMAZI2SsQ7VmyzJjokQ |          930 |
| Ed        | -fUARDNuXAfrOn4WLSZLgA |          904 |
| Nicole    | -hKniZN2OdshWLHYuj21jQ |          864 |
| Fran      | -9da1xk7zgnnfO1uTVYGkA |          862 |
| Mark      | -B-QEUESGWHPE_889WJaeg |          861 |
| Christina | -kLVfaJytOJY2-QdQoCcNQ |          842 |
| Dominic   | -kO6984fXByyZm3_6z2JYg |          836 |
| Lissa     | -lh59ko3dxChBSZ9U7LfUw |          834 |
| Lisa      | -g3XIcCb2b-BD0QBCcq2Sw |          813 |
| Alison    | -l9giG8TSDBG1jnUBUXp5w |          775 |
| Sui       | -dw8f7FLaUmWR7bfJ_Yf0w |          754 |
| Tim       | -AaBjWJYiQxXkCMDlXfPGw |          702 |
| L         | -jt1ACMiZljnBFvS6RRvnA |          696 |
| Angela    | -IgKkE8JvYNWeGu8ze4P8Q |          694 |
| Crissy    | -hxUwfo3cMnLTv-CAaP69A |          676 |
| Lyn       | -H6cTbVxeIRYR-atxdielQ |          675 |
+-----------+------------------------+--------------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

Answer

As table below illustrates, posing more reviews does not necessarily correlate with more fans. For example, although, Gerald and sara have posed the most reviews, they have fewer fans in comparison with Harald. Therefore, sorting the users in descending order based on their total number of reviews does not sort the fans in the same order, meaning that there is not a correlation between the total number of reviews and number of fans.

```
select

name

, id

, review_count

, fans

from user
```

```
order by review_count desc;

+------------+------------------------+--------------+------+
| 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 |
+-----------+-----------------------+-------------+------+
```

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

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

Answer:

As the tables below show there are more reviews with the word □love□ in them compared to the word □hate□.

code

```
select

count (*)

from review

where text like '%love%';
```

Output

```
+-----------+

| count (*) |

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

|      1780 |

+-----------+
```

code

```
select

count (*)

from review

where text like '%hate%';
```

Output

```
+-----------+

| count (*) |

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

|      232 |

+-----------+
```

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

Answer

CODE

```
select

name
```

```
, id

, fans

from user

order by fans desc;
```

OUTPUT

```
+-----------+------------------------+------+
| name      | id                     | fans |
+-----------+------------------------+------+
| Amy       | -9I98YbNQnLdAmcYfb324Q |  503 |
| Mimi      | -8EnCioUmDygAbsYZmTeRQ |  497 |
| Harald    | --2vR0DIsmQ6WfcSzKWigw |  311 |
| Gerald    | -G7Zkl1wIWBBmD0KRy_sCw |  253 |
| Christine | -0IiMAZI2SsQ7VmyzJjokQ |  173 |
| Lisa      | -g3XIcCb2b-BD0QBCcq2Sw |  159 |
| Cat       | -9bbDysuiWeo2VShFJJtcw |  133 |
| William   | -FZBTkAZEXoP7CYvRV2ZwQ |  126 |
| Fran      | -9da1xk7zgnnfO1uTVYGkA |  124 |
| Lissa     | -lh59ko3dxChBSZ9U7LfUw |  120 |
| Mark      | -B-QEUESGWHPE_889WJaeg |  115 |
| Tiffany   | -DmqnhW4Omr3YhmnigaqHg |  111 |
| bernice   | -cv9PPT7IHux7XUc9dOpkg |  105 |
| Roanna    | -DFCC64NXgqrxlO8aLU5rg |  104 |
| Angela    | -IgKkE8JvYNWeGu8ze4P8Q |  101 |
| .Hon      | -K2Tcgh2EKX6e6HqqIrBIQ |  101 |
| Ben       | -4viTt9UC44lWCFJwleMNQ |   96 |
| Linda     | -3i9bhfvrM3F1wsC9XIB8g |   89 |
| Christina | -kLVfaJytOJY2-QdQoCcNQ |   85 |
| Jessica   | -ePh4Prox7ZXnEBNGKyUEA |   84 |
| Greg      | -4BEUkLvHQntN6qPfKJP2w |   81 |
| Nieves    | -C-l8EHSLXtZZVfUAUhsPA |   80 |
| Sui       | -dw8f7FLaUmWR7bfJ_Yf0w |   78 |
| Yuri      | -8lbUNlXVSoXqaRRiHiSNg |   76 |
| Nicole    | -0zEEaDFIjABtPQni0XlHA |   73 |
+-----------+------------------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

11. Is there a strong correlation between having a high number of fans and being listed as "useful" or "funny?"

Answer

CODE

```
select

name

, id

, fans

, useful

, funny

from user

order by fans desc;
```

Copy and Paste the Result Below:

```
    +-----------+----------------------+------+--------+--------+

| name      | id                   | fans | useful |  funny |

+-----------+----------------------+------+--------+--------+
| Amy       | -9I98YbNQnLdAmcYfb324Q |  503 |   3226 |   2554 |

| Mimi      | -8EnCioUmDygAbsYZmTeRQ |  497 |    257 |    138 |

| Harald    | --2vR0DIsmQ6WfcSzKWigw |  311 | 122921 | 122419 |

| Gerald    | -G7Zkl1wIWBBmD0KRy_sCw |  253 |  17524 |   2324 |

| Christine | -0IiMAZI2SsQ7VmyzJjokQ |  173 |   4834 |   6646 |

| Lisa      | -g3XIcCb2b-BD0QBCcq2Sw |  159 |     48 |     13 |

| Cat       | -9bbDysuiWeo2VShFJJtcw |  133 |   1062 |    672 |

| William   | -FZBTkAZEXoP7CYvRV2ZwQ |  126 |   9363 |   9361 |

| Fran      | -9da1xk7zgnnfO1uTVYGkA |  124 |   9851 |   7606 |

| Lissa     | -lh59ko3dxChBSZ9U7LfUw |  120 |    455 |    150 |

| Mark      | -B-QEUESGWHPE_889WJaeg |  115 |   4008 |    570 |

| Tiffany   | -DmqnhW4Omr3YhmnigaqHg |  111 |   1366 |    984 |

| bernice   | -cv9PPT7IHux7XUc9dOpkg |  105 |    120 |    112 |

| Roanna    | -DFCC64NXgqrxlO8aLU5rg |  104 |   2995 |   1188 |

| Angela    | -IgKkE8JvYNWeGu8ze4P8Q |  101 |    158 |    164 |
| .Hon      | -K2Tcgh2EKX6e6HqqIrBIQ |  101 |   7850 |   5851 |
| Ben       | -4viTt9UC44lWCFJwleMNQ |   96 |   1180 |   1155 |
| Linda     | -3i9bhfvrM3F1wsC9XIB8g |   89 |   3177 |   2736 |
| Christina | -kLVfaJytOJY2-QdQoCcNQ |   85 |    158 |     34 |
| Jessica   | -ePh4Prox7ZXnEBNGKyUEA |   84 |   2161 |   2091 |
| Greg      | -4BEUkLvHQntN6qPfKJP2w |   81 |    820 |    753 |
| Nieves    | -C-l8EHSLXtZZVfUAUhsPA |   80 |   1091 |    774 |
| Sui       | -dw8f7FLaUmWR7bfJ_Yf0w |   78 |      9 |     18 |
| Yuri      | -8lbUNlXVSoXqaRRiHiSNg |   76 |   1166 |    220 |
| Nicole    | -0zEEaDFIjABtPQni0XlHA |   73 |     13 |     10 |
+-----------+------------------------+------+--------+--------+

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

Please explain your findings and interpretation of the results:

Based on the table above sorting the users based on their number of fans doesn�t show descending or ascending trend in �useful� or �funny� columns. Therefore, there shouldn�t be 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.

City:Pittsburgh Category: Pizza

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

Yes

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

Yes

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

Based on the results, we can see that there seems to be a correlation between the location of the business and their rating. The business that are probably located in the same neighbor have close rating. Also they have similar working hours. Moreover, the business that have longer working hours usually have higher rating.

code

```
select

business.name

, business.city

, category.category

, business.stars

, hours.hours

, business.review_count

, business.postal_code

from (business inner join category on business.id = category.business_id) inner join hours on hours.business_id = category.business_id

where business.city = 'Pittsburgh’

 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 open has low rating.

ii. Difference 2:

The business open not more reviews.

iii. Difference 3:

The business open has less working hours.

code

```
select

business.name

, business.is_open

, category.category

, business.stars

, hours.hours

, business.review_count

, business.postal_code

from (business inner join category on business.id = category.business_id) inner join
hours on hours.business_id = category.business_id

where business.city = 'Pittsburgh'

 group by business.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: