Worksheet-Data Analysis using SQL

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 ; foreign key= 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 *                 --selecting all columns
from user                    ---table name
where id is null or          ----condition for null, checking on all columns
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:  1.0   max: 5.0         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 total_reviews
from business
group by city
order by total_reviews desc
```

Copy and Paste the Result Below:

```
+-----------------+---------------+
| city            | total_reviews |
+-----------------+---------------+
| 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)
```

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 as star_rating,count(stars) as Count
```

```
from business
where city='Avon'
group by star_rating;
```

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

```
+-------------+-------+
| star_rating | Count |
+-------------+-------+
|         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 as star_rating,count(stars) as Count
from business
where city='Beachwood'
group by star_rating;
```

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

```
+-------------+-------+
| star_rating | Count |
+-------------+-------+
|         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 id as user_id,name, review_count
from user
order by  review_count desc                      -- highest review_count users will be listed first as it is a
descending order
limit 3;                                         -- Limit to 3 rows in output
```

Copy and Paste the Result Below:

```
+------------------------+--------+--------------+
| user_id                | name   | review_count |
+------------------------+--------+--------------+
| -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:

No. Because higher review_count doesn't lead to more fans.

SQL Code:

```
select id as user_id, review_count,fans
from user
order by  fans desc;
```

Output:

```
+------------------------+--------------+------+
| user_id                | review_count | fans |
+------------------------+--------------+------+
| -9I98YbNQnLdAmcYfb324Q |          609 |  503 |
| -8EnCioUmDygAbsYZmTeRQ |          968 |  497 |
| --2vR0DIsmQ6WfcSzKWigw |         1153 |  311 |
| -G7Zkl1wIWBBmD0KRy_sCw |         2000 |  253 |
| -0IiMAZI2SsQ7VmyzJjokQ |          930 |  173 |
| -g3XIcCb2b-BD0QBCcq2Sw |          813 |  159 |
| -9bbDysuiWeo2VShFJJtcw |          377 |  133 |
| -FZBTkAZEXoP7CYvRV2ZwQ |         1215 |  126 |
| -9da1xk7zgnnfO1uTVYGkA |          862 |  124 |
| -lh59ko3dxChBSZ9U7LfUw |          834 |  120 |
| -B-QEUESGWHPE_889WJaeg |          861 |  115 |
| -DmqnhW4Omr3YhmnigaqHg |          408 |  111 |
| -cv9PPT7IHux7XUc9dOpkg |          255 |  105 |
| -DFCC64NXgqrxlO8aLU5rg |         1039 |  104 |
| -IgKkE8JvYNWeGu8ze4P8Q |          694 |  101 |
| -K2Tcgh2EKX6e6HqqIrBIQ |         1246 |  101 |
| -4viTt9UC44lWCFJwleMNQ |          307 |   96 |
| -3i9bhfvrM3F1wsC9XIB8g |          584 |   89 |
| -kLVfaJytOJY2-QdQoCcNQ |          842 |   85 |
| -ePh4Prox7ZXnEBNGKyUEA |          220 |   84 |
| -4BEUkLvHQntN6qPfKJP2w |          408 |   81 |
| -C-l8EHSLXtZZVfUAUhsPA |          178 |   80 |
| -dw8f7FLaUmWR7bfJ_Yf0w |          754 |   78 |
| -8lbUNlXVSoXqaRRiHiSNg |         1339 |   76 |
| -0zEEaDFIjABtPQni0XlHA |          161 |   73 |
+------------------------+--------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

AS we can see from the result (here fans is arranged in descending order) fans and review_count are not positively correlated. There are a quite a few cases wherein less reviews by users have more fans.

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

Answer: "love"

```
      SQL code used to arrive at answer:
select count(id) as Love_count                                --count the number of reviews in which the word
love is used
from review
where text like '%love%' or text like 'Love%' or text like '%Love%';    --assuming no upper case form of the
word "love" is not used
Output:
+------------+
| Love_count |
+------------+
|       1780 |
+------------+

Sql Code for Hate:
select count(id) as Hate_count                                       --count the number of reviews in which
the word hate is used
from review
where text like '%hate%' or text like 'Hate%' or text like '%Hate%';    --assuming no upper case form of the
word "hate" is not used

+------------+
| Hate_count |
+------------+
|        232 |
+------------+
```

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

```
      SQL code used to arrive at answer:

select id as user_id, name, fans
from user
order by  fans desc               -- arrange user_id in descending order of number of fans
limit 10;                         -- limit to first 10 rows

      Copy and Paste the Result Below:
+------------------------+-----------+------+
| user_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 |
+------------------------+------------+------+
```

11. 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 id,name,fans,useful,funny,(useful+funny) as total
from user
order by fans desc
limit 10;
```

Copy and Paste the Result Below:

```
+------------------------+------------+------+--------+--------+--------+
| id                     | name       | fans | useful |  funny |  total |
+------------------------+------------+------+--------+--------+--------+
| -9I98YbNQnLdAmcYfb324Q | Amy        |  503 |   3226 |   2554 |   5780 |
| -8EnCioUmDygAbsYZmTeRQ | Mimi       |  497 |    257 |    138 |    395 |
| --2vR0DIsmQ6WfcSzKWigw | Harald     |  311 | 122921 | 122419 | 245340 |
| -G7Zkl1wIWBBmD0KRy_sCw | Gerald     |  253 |  17524 |   2324 |  19848 |
| -0IiMAZI2SsQ7VmyzJjokQ | Christine  |  173 |   4834 |   6646 |  11480 |
| -g3XIcCb2b-BD0QBCcq2Sw | Lisa       |  159 |     48 |     13 |     61 |
| -9bbDysuiWeo2VShFJJtcw | Cat        |  133 |   1062 |    672 |   1734 |
| -FZBTkAZEXoP7CYvRV2ZwQ | William    |  126 |   9363 |   9361 |  18724 |
| -9da1xk7zgnnfO1uTVYGkA | Fran       |  124 |   9851 |   7606 |  17457 |
| -lh59ko3dxChBSZ9U7LfUw | Lissa      |  120 |    455 |    150 |    605 |
+------------------------+------------+------+--------+--------+--------+
```

Please explain your findings and interpretation of the results:

Based on the results, Harald has the highest number of useful and/or funny and he has the 3rd highest number of fans.
After going through the result table, I think there is a medium relationship between having high number of fans and being listed as useful or funny.
Medium because as we go down in the table with the monotonic decrease in number of fans, the useful and/or funny count is not decreasing monotonically.

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.

City = Toronto
Group 1= star rating between 2 and 3
Group 2= star rating between 4 and 5

Group 1 have majority of the business which are open all day.

```
+-------+---------+-------------+--------------------+
| stars | city    | hours       | number_of_business |
+-------+---------+-------------+--------------------+
|   2.0 | Toronto | 11:00-23:00 |                  7 |
|   2.5 | Toronto | |10:00-2:00 |                  2 |
|   2.5 | Toronto | |11:00-2:00 |                  5 |
|   2.5 | Toronto | |8:00-22:00 |                  7 |
|   3.0 | Toronto | 10:00-23:00 |                  1 |
|   3.0 | Toronto | 10:30-21:00 |                  6 |
|   3.0 | Toronto | 11:00-19:00 |                  1 |
|   3.0 | Toronto | y|9:00-4:00 |                  1 |
|   3.0 | Toronto | |10:00-4:00 |                  1 |
|   3.0 | Toronto | |6:00-21:00 |                  1 |
|   3.0 | Toronto | |6:00-22:00 |                  4 |
|   3.0 | Toronto | |8:00-18:00 |                  1 |
|   3.0 | Toronto | |8:00-20:00 |                  1 |
|   3.0 | Toronto | |9:00-23:00 |                  4 |
+-------+---------+-------------+--------------------+
```

Group 2 have majority of the business which are open either only during first half of the day or only during second half of the day.

```
+-------+---------+-------------+--------------------+
| stars | city    | hours       | number_of_business |
+-------+---------+-------------+--------------------+
|   4.0 | Toronto | 11:00-21:00 |                  2 |
|   4.0 | Toronto | 12:00-16:00 |                  1 |
|   4.0 | Toronto | 15:00-21:00 |                  4 |
|   4.0 | Toronto | 18:00-23:00 |                  4 |
|   4.5 | Toronto | 10:00-14:00 |                  2 |
|   4.5 | Toronto | 10:00-17:00 |                  1 |
|   4.5 | Toronto | 11:00-17:00 |                  1 |
|   4.5 | Toronto | 11:00-19:00 |                  4 |
|   4.5 | Toronto | 11:00-23:00 |                  6 |
|   4.5 | Toronto | 11:30-18:00 |                  2 |
|   4.5 | Toronto | 12:00-16:00 |                  1 |
|   4.5 | Toronto | 14:00-19:00 |                  2 |
|   4.5 | Toronto | 14:00-23:00 |                  1 |
|   4.5 | Toronto | |16:00-2:00 |                  3 |
|   4.5 | Toronto | |18:00-2:00 |                  4 |
|   4.5 | Toronto | |9:00-19:00 |                  3 |
|   5.0 | Toronto | 17:00-22:00 |                  3 |
|   5.0 | Toronto | 18:00-22:00 |                  2 |
+-------+---------+-------------+--------------------+
```

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

City = Toronto
Group 1= star rating between 2 and 3
Group 2= star rating between 4 and 5

Group 1 output:

```
+-------+---------+-------------------+
| stars | city    | number_of_reviews |
+-------+---------+-------------------+
|   2.0 | Toronto |               470 |
|   2.5 | Toronto |              1690 |
|   3.0 | Toronto |              3833 |
+-------+---------+-------------------+
```

Group 2 output:

```
+-------+---------+-------------------+
| stars | city    | number_of_reviews |
+-------+---------+-------------------+
|   4.0 | Toronto |              6775 |
|   4.5 | Toronto |              2425 |
|   5.0 | Toronto |               751 |
+-------+---------+-------------------+
```

As it can be seen from the above tables, the number of reviews in group 1= 5993 and number of reviews in group 2= 9951 are different.

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

Yes.
City=Toronto
Neighborhood=Downtown Core
Group 1= star rating between 2 and 3
Group 2= star rating between 4 and 5

```
+-------+---------------+---------+--------------------+
| stars | neighborhood  | city    | number_of_business |
+-------+---------------+---------+--------------------+
|   2.0 | Downtown Core | Toronto |                  8 |
|   2.5 | Downtown Core | Toronto |                  6 |
|   3.0 | Downtown Core | Toronto |                 12 |
|   3.5 | Downtown Core | Toronto |                 19 |
|   4.0 | Downtown Core | Toronto |                 23 |
|   4.5 | Downtown Core | Toronto |                  5 |
|   5.0 | Downtown Core | Toronto |                  9 |
+-------+---------------+---------+--------------------+
```

From the above results, the number of business in group 2= 37 are more than that in group 1= 26 in Downtown Core area of Toronto City.

So, there are more higher rating business in Donwtown Core area.

SQL code used for analysis:

i.

Group 1:

```
		select b.stars,b.city,substr(h.hours,-11) as hours,count(b.id) as number_of_business     --
substring to remove the day from the hours column
		from business as b
		inner join hours as h                                                                  --inner
join to fetch data from business and hours table
		on b.id=h.business_id
		group by b.stars,b.city,substr(h.hours,-11)
		having city='Toronto'
		and stars between 2.0 and 3.0
		order by stars ;
```

Group 2:

```
		select b.stars,b.city,substr(h.hours,-11) as hours,count(b.id) as number_of_business     --
substring to remove the day from the hours column
		from business as b
		inner join hours as h                                                                  --inner
join to fetch data from business and hours table
		on b.id=h.business_id
		group by b.stars,b.city,substr(h.hours,-11)
		having city='Toronto'
		and stars between 4.0 and 5.0
```

```
                order by stars ;
```

ii.

Group 1:

```
                select stars,city,sum(review_count) as number_of_reviews
                from business
                group by stars,city
                having city='Toronto' and stars between 2.0 and 3.0;
```

Group 2:

```
                select stars,city,sum(review_count) as number_of_reviews
                from business
                group by stars,city
                having city='Toronto' and stars between 4.0 and 5.0;
```

iii.

```
                select stars,neighborhood,city,count(id) as number_of_business
                from business
                group by stars,neighborhood,city
                having city='Toronto'and neighborhood='Downtown Core'
                and stars between 2.0 and 5.0
                order 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 dataset contains information 10000 business in total out of whihc 1520 business are closed and 8480 are open.

ii. Difference 2:

For city=Concord percentage of closed business's = (3/49)*100= 6.12 whereas
for city = Charlotte it is = (70/468)*100=   14.96%

SQL code used for analysis:

```
--For difference 1
```

```
select is_open, count(name) as Number_of_business
from business
group by is_open                                         -- grouping by whether open or closed

--For difference 2

select is_open,city,count(name) as Number_of_business
from business
group by is_open,city                           -- grouping by whether open or closed and also city where the
business is located
having city='Concord'

select is_open,city,count(name) as Number_of_business
from business
group by is_open,city                         -- grouping by whether open or closed and also city where the
business is located
having city='Charlotte'
```

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:

Determine whether an elite user would have more number of fans than a non elite user or vice versa.

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

Classify the users in user table into elite user and non elite user. Then determine the number of elite users, sum of fans of elite users, sum of fans of non elite users, number of non elite users.

Using these values calculate the average number of fans per elite user and non elite user.

Results are-

	Sum of Fans	No: of users	Average no: of fans	
Total	14896	10000		
Elite	447	21	21.28571429	average no: of fans per elite user
Non-Elite	14449	9979	1.447940675	average no: of fans per non elite user

As we can see from the results table, the average number of fans for an eliter user is > than that for a non elite user

This was done to test human psychology on a historical dataset. Generally, people prefer to follow users which have elite status so chances of elite users greater number of fans is higher.

iii. Output of your finished dataset:

```
+------------------------+---------+--------------+------+------------+
| id                     | name    | review_count | fans | Elite_user |
+------------------------+---------+--------------+------+------------+
| --BumyUHiO_7YsHurb9Hkw | Sapna   |           38 |    1 | Yes        |
| --Qh8yKWAvIP4V4K8ZPfHA | Dixie   |          503 |   41 | Yes        |
| -0HhZbPBlB1YZx3BhAfaEA | Tasha   |          250 |    8 | Yes        |
| -50XWnmQGqBgEI-9ANvLlg | Lalena  |          224 |   25 | Yes        |
| -5e4VTnu_pR4Gpv3VSncaw | Justin  |          177 |   13 | Yes        |
| -9RU4LuI_TfYgv9rBijJoQ | Keith   |           61 |    3 | Yes        |
| -9SoHrhiiUVmx6-MkyR4RA | Brad    |          182 |    1 | Yes        |
| -a0LRFr94D9ohyBJCKVvXQ | Elaine  |          332 |   18 | Yes        |
| -aAgfEUH4UoFDRXZCfJSUA | Matt    |          476 |   14 | Yes        |
| -C-l8EHSLXtZZVfUAUhsPA | Nieves  |          178 |   80 | Yes        |
| -cvrhCPCKHUkEsDak_fY4g | Jamie    |           95 |    4 | Yes        |
| -d2daWmftYumOaYpbD5D8Q | Jia      |          228 |    8 | Yes        |
| -dbWm5L_Ol2hZeLRoQOK7w | Mel      |          156 |    9 | Yes        |
| -EWQZjRHAKMddHW_dZTvdw | Chris    |           70 |    2 | Yes        |
| -fUARDNuXAfrOn4WLSZLgA | Ed       |          904 |   38 | Yes        |
| -ga7pQvnJcMB1_pIapHQRQ | Tracy    |           71 |    5 | Yes        |
| -GD0XVUKRj96vf6TP68Evw | Maung    |           54 |    0 | Yes        |
| -HLE-x7Lpkfprd6er-JFGg | Danial   |          136 |    5 | Yes        |
| -hYYjAXSAa657rY0ANtTGQ | Kristen  |          428 |   15 | Yes        |
| -kO6984fXByyZm3_6z2JYg | Dominic  |          836 |   37 | Yes        |
| -lh59ko3dxChBSZ9U7LfUw | Lissa    |          834 |  120 | Yes        |
| ---1lKK3aKOuomHnwAkAow | Monera   |          245 |   15 | No         |
| ---94vtJ_5o_nikEs6hUjg | Joe      |            2 |    0 | No         |
| ---cu1hq55BP9DWVXXKHZg | Jeb      |           57 |    0 | No         |
| ---fhiwiwBYrvqhpXgcWDQ | Jed      |            8 |    0 | No         |
+------------------------+----------+--------------+------+------------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

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

```
--To classify users in user table into elite user and non elite user

select id,name,review_count,fans,
(case when id in (select user_id from elite_years)
then 'Yes'
else 'No'
end) as Elite_user
from user
order by Elite_user desc


--sum of fans 14896 amongst  10000 users
select count(distinct id)   -- number of users
from user

select sum(fans)    ---sum of fans
from user
```

```
--sum of fans of elite_users = 447 amongst 21 elite_users

select count(id)           --number of elite users in user table
from user
where id in (select user_id from elite_years)


select sum(fans)            ---sum of fans of elite users
from user
where id in (select user_id from elite_years)
```