Part 1: Yelp Dataset Profiling and Understanding

1. Profile the data by finding the total number of records for each of the tables below:

```
SELECT COUNT(*)
FROM table;
```

i. Attribute table = 10,000
ii. Business table = 10,000
iii. Category table = 10,000
iv. Checkin table = 10,000
v. elite_years table = 10,00
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.

```
SELECT COUNT(DISTINCT (key))
FROM Table;
```

Table	# records	Key
i. Business =	10,000	id
ii. Hours =	1562	business_id
iii. Category =	2643	business_id
iv. Attribute =	1115	business_id
v. Review =	10,000, 8090, 9581	id, business_id, user_id
vi. Checkin =	493	business_id
vii. Photo =	10000, 6493	id,business_id
viii. Tip =	3979,537	business_id,user_id
ix. User =	10,000	id
x. Friend =	11	user_id
xi. Elite_years =	2780	user_id

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(*) AS Number_of_Nulls
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:

```
SELECT
MIN(Column) AS Minimum,
MAX(Column) AS Maximum,
AVG(Column) AS Average
FROM Table;
```

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:

```
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, COUNT(stars) AS frequency
FROM Business
WHERE City = 'Avon'
GROUP BY stars;
```

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

```
+-------+-----------+
| stars | frequency |
+-------+-----------+
|   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, COUNT(stars) AS frequency
FROM Business
WHERE City = 'Beachwood'
GROUP BY stars;
```

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

stars	frequency
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
name, review_count
FROM user
ORDER BY review_count DESC
LIMIT 3;
```

Copy and Paste the Result Below:

name	review_count
Gerald	2000
Sara	1629
Yuri	1339

8. Does posing more reviews correlate with more fans?

No, from the given database and the table "user", posting more reviews does not correlate with more fans.

Please explain your findings and interpretation of the results:

By running the below query,

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

The obtained result is,

id	name	review_count	fans	yelping_since
-9I98YbNQnLdAmcYfb324Q	Amy	609	503	2007-07-19 00:00:00

-8EnCioUmDygAbsYZmTeRQ	Mimi	968	497	2011-03-30 00:00:00
--2vR0DIsmQ6WfcSzKWigw	Harald	1153	311	2012-11-27 00:00:00
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	2000	253	2012-12-16 00:00:00
-0IiMAZI2SsQ7VmyzJjokQ	Christine	930	173	2009-07-08 00:00:00
-g3XIcCb2b-BD0QBCcq2Sw	Lisa	813	159	2009-10-05 00:00:00
-9bbDysuiWeo2VShFJJtcw	Cat	377	133	2009-02-05 00:00:00
-FZBTkAZEXoP7CYvRV2ZwQ	William	1215	126	2015-02-19 00:00:00
-9da1xk7zgnnfO1uTVYGkA	Fran	862	124	2012-04-05 00:00:00
-lh59ko3dxChBSZ9U7LfUw	Lissa	834	120	2007-08-14 00:00:00
-B-QEUESGWHPE_889WJaeg	Mark	861	115	2009-05-31 00:00:00
-DmqnhW4Omr3YhmnigaqHg	Tiffany	408	111	2008-10-28 00:00:00
-cv9PPT7IHux7XUc9dOpkg	bernice	255	105	2007-08-29 00:00:00
-DFCC64NXgqrxlO8aLU5rg	Roanna	1039	104	2006-03-28 00:00:00
-IgKkE8JvYNWeGu8ze4P8Q	Angela	694	101	2010-10-01 00:00:00
-K2Tcgh2EKX6e6HqqIrBIQ	.Hon	1246	101	2006-07-19 00:00:00
-4viTt9UC44lWCFJwleMNQ	Ben	307	96	2007-03-10 00:00:00
-3i9bhfvrM3F1wsC9XIB8g	Linda	584	89	2005-08-07 00:00:00
-kLVfaJytOJY2-QdQoCcNQ	Christina	842	85	2012-10-08 00:00:00
-ePh4Prox7ZXnEBNGKyUEA	Jessica	220	84	2009-01-12 00:00:00
-4BEUkLvHQntN6qPfKJP2w	Greg	408	81	2008-02-16 00:00:00
-C-l8EHSLXtZZVfUAUhsPA	Nieves	178	80	2013-07-08 00:00:00
-dw8f7FLaUmWR7bfJ_Yf0w	Sui	754	78	2009-09-07 00:00:00
-8lbUNlXVSoXqaRRiHiSNg	Yuri	1339	76	2008-01-03 00:00:00
-0zEEaDFIjABtPQni0XlHA	Nicole	161	73	2009-04-30 00:00:00

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

We can clearly witness that there is no relationship between review count and number of fans.

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

Answer:

hatecount
232

lovecount
1780

SQL code used to arrive at answer:

```
SELECT
COUNT(*) AS hatecount
FROM review
WHERE text LIKE '%hate%';

SELECT
COUNT(*) AS lovecount
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?
Ans: Yes, they follow different distribution hours, as by the results I can interpret that the restaurants with the rating range '2-3 stars' operates for less hours as compared to the restaurants with the rating range '4-5 stars'.

ii. Do the two groups you chose to analyze have a different number of reviews?
Ans: Yes they have different number of reviews overall.

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
Ans: No, as right now I am not aware about the locality that is given in location so I cannot interpret any useful information.

SQL code used for analysis:

```
SELECT
b.name,
b.city,
c.category,
h.hours,
b.review_count,
b.stars,
b.postal_code AS location,
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 rating_range
FROM business b
INNER JOIN hours h ON b.id = h.business_id
INNER JOIN category c ON c.business_id = b.id
```

```
WHERE City = 'Mississauga' AND category = 'Restaurants' AND
rating_range in ('2-3 stars','4-5 stars')
GROUP BY name
ORDER BY stars DESC;
```

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 businesses that are open are more than the businesses that are closed. Precisely the count for open businesses is 8480 and the count for closed businesses is 1520, i.e., therefore the number of open businesses are approximately 5.5 times the closed businesses.

ii. Difference 2:

The users have reviews open businesses more than the closed businesses. The average review count was 9 points more for business that are open than the business that are closed

SQL code used for analysis:

```
SELECT
count(distinct id),
avg(stars),
avg(review_count),
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 the following:

i. Indicate the type of analysis you chose to do:

The type of analysis I chose to do is descriptive analysis. I performed calculations to determine the average count of reviews, average rating, and the number of open businesses for each category of businesses in the dataset. By grouping the results by category and ordering them based on the average count of reviews, I gained insights into the distribution of these metrics across different categories. This analysis helps in understanding the overall trends and characteristics of the businesses in terms of reviews, ratings, and open status within each category.

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

For my analysis, I would need a dataset that includes information about operating businesses, such as their attributes, reviews, hours of operation, categories, and tips. This comprehensive dataset allows me to

gain insights into various aspects of the businesses, such as their characteristics, customer feedback, operational hours, and categorization. By considering multiple tables such as business, reviews, hours, category, and tips, I can leverage the relationships between them to perform more meaningful analysis and explore correlations between different factors. This diverse set of data enables a holistic understanding of the businesses and facilitates a thorough analysis of their performance, customer satisfaction, and operational patterns.

iii. Output of your finished dataset:

category	Avg_count_of_review	Rating	num_of_open_businesses	city
Malaysian	768.0	4.0	1	Las Vegas
Taiwanese	768.0	4.0	1	Las Vegas
Farmers Market	723.0	4.5	1	Cleveland
Fruits & Veggies	723.0	4.5	1	Cleveland
Market Stalls	723.0	4.5	1	Cleveland
Meat Shops	723.0	4.5	1	Cleveland
Public Markets	723.0	4.5	1	Cleveland
Seafood Markets	723.0	4.5	1	Cleveland
Smokehouse	431.0	4.0	1	Phoenix
Asian Fusion	396.5	3.5	2	Las Vegas
Soup	394.5	3.75	2	Las Vegas
Noodles	386.5	3.25	2	Las Vegas
Ethnic Food	363.0	4.0	2	Cleveland
Arabian	267.0	5.0	1	Mesa
Halal	267.0	5.0	1	Mesa
Barbeque	252.5	3.75	2	Phoenix
Architects	223.0	4.5	1	Scottsdale
Architectural Tours	223.0	4.5	1	Scottsdale
Museums	223.0	4.5	1	Scottsdale
Tours	223.0	4.5	1	Scottsdale
Chinese	199.0	3.125	3	Edinburgh
Salad	198.0	4.5	1	Mesa
Specialty Food	179.2	4.0	3	Cleveland
Vegetarian	168.0	4.0	0	Las Vegas
Mediterranean	161.0	4.5	2	Oakville

(Output limit exceeded, 25 of 257 total rows shown)

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

```
SELECT
c.category,
AVG(b.review_count) AS Avg_count_of_review,
AVG(b.stars) AS Rating,
SUM(is_open) AS num_of_open_businesses,
b.city
FROM category c
INNER JOIN business b ON b.id = c.business_id
GROUP BY c.category
ORDER BY Avg_count_of_review DESC;
```