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:

i. Business = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 3979 (business_id was used as foreign key)

ix. User = 10000

x. Friend = 11

xi. Elite_years = 2780

3. Are there any columns with null values in the Users table?

Answer: No

SQL code used to arrive at answer:

```
SELECT *
FROM  user
WHERE  id 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_ 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 result
FROM business
GROUP BY city
ORDER BY result DESC
```

Result:

city	result
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

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, review_count
FROM business
WHERE city='Avon'
```

Result:

stars	review_count
2.5	3
4.0	4
5.0	3
3.5	7
1.5	10
3.5	31
4.5	31
3.5	50
2.5	3
4.0	17

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars, review_count
FROM business
WHERE city='Beachwood'
```

Result:

stars	review_count
3.0	8
3.0	3
4.5	14
5.0	6
4.0	69
4.5	3
5.0	4
2.0	8
3.5	3
3.5	3
5.0	6

```
|   2.5 |            3 |
|   5.0 |            3 |
|   5.0 |            4 |
+-------+--------------+
```

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
```

Result:

```
+-----------+--------------+
| name      | review_count |
+-----------+--------------+
| Gerald    |         2000 |
| Sara      |         1629 |
| Yuri      |         1339 |
+-----------+--------------+
```

8. Does posing more reviews correlate with more fans?

Yes, posing more reviews corelates with more fans. When selecting both fans and review_count from the user table with descending order of review_count, most of the 25 shown records have fans above 30. On the other hand, when we order review_count in ascending order, all the 25 shown records have 0 fan. Moreover, the average number of fans across all the records is 1.4896. Hence, these two quantities are positively correlated.

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

There are more reviews with the word “love” (1780 records) than “hate” (232 records).

SQL code used to arrive at answer:

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

CROSS JOIN

(SELECT count(text) AS Hate

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

Result:

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. Selected the city Pittsburgh and group the businesses in that city by their overall star rating. Compare the businesses with 2-3 stars to the businesses with 4-5 stars and answer the following questions.

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

Yes. For businesses with 2-3 stars, the working hours are normally from noon or afternoon to midnight. On the other hand, businesses with 4-5 stars have working hours starting from morning to evening. The total working hours seem to be comparable.

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

Yes. For businesses with 2-3 stars, they have more than 20 reviews, more than businesses with 4-5 stars who have less than 10 reviews. Hence, 4-5 stars might be due to randomness.

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

There is no obvious correlation between stars and the location data. However, we can see that those businesses having 2-3 stars are more closed to each other than from businesses having 4-5 stars, based on their postal codes.

SQL code used for analysis:

```
SELECT b.*, h.hours

FROM business b INNER JOIN hours h ON b.id=h.business_id

WHERE city = 'Pittsburgh' AND (stars BETWEEN 2 AND 3 OR stars BETWEEN 4 AND 5)

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:

Businesses that are still open tend to have more reviews than those are closed. This is intuitive because open business tends to have more consumers.

ii. Difference 2:

Businesses that are still open tend to have reviews being more useful (0.923076923077) and cool (0.384615384615), compared with business not open (0 on both criteria). This is also intuitive because reviews about closed businesses normally get few reply.

SQL code used for analysis (i):

```
SELECT AVG(review_count)

FROM business

GROUP BY is_open

ORDER BY is_open
```

(ii)

```
SELECT is_open, AVG(useful), AVG(funny), AVG(cool)
FROM business b INNER JOIN review r ON b.id=r.id
GROUP BY is_open
ORDER BY is_open
```

3. Independent analysis:

i. Indicate the type of analysis:

I chose to analyze differences in star ratings between different states in the business table and find possible variables that may cause this difference. Differences in ratings may reflect the degree of satisfaction of businesses in different states, and hence could indicate business quality.

ii. Type of data needed for analysis:

Since we need to do quantitative analysis, most of the data we need would be real numbers. Except the states which is a categorical variable, all the other variables are numbers: mean value of stars, number of distinct business stores, latitude, longitude, whether businesses are still open, etc.

Intuitively, these variables are all related to stars and are all numbers, so we could infer whether there are correlations between stars and these variables. These data are accessible from business and checkin tables.

iii. Output of finished dataset:

By calculating the average of stars group by states, we could conclude that stars differ a lot across states, from 3 to 5 stars. State NI has the lowest mean value of stars of 3 and state ST has the highest 5 stars. On the other hand, many of them are within the 3.5-4.0 range.

We finally selected two variables that seem to be correlated with stars. The new variable COUNT(distinct name) could represent the number of businesses in Yelp. States having more stars tend to have more businesses in Yelp. Furthermore, on the two extreme sides of stars, those states tend to have less than 10 businesses. This is intuitive because less businesses may cause the statistics to be more random. The variable is_open is positively correlated with stars: more open businesses in a state tend to have more stars. This result is self-explanatory because a higher proportion of open businesses will make customers to be more satisfied.

iv. SQL code:

```
SELECT state, AVG(stars) as mean, COUNT(distinct name), AVG(is_open)
FROM business
GROUP BY state
ORDER BY mean
```