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 = 1000

ii. Business table = 1000

iii. Category table = 1000

iv. Checkin table = 1000

v. elite_years table = 1000

vi. friend table = 1000

vii. hours table = 1000

viii. photo table = 1000

ix. review table = 1000

x. tip table = 1000

xi. user table = 1000

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.

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

i. Business = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 1000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 537

ix. User = 10000

x. Friend = 11

xi. Elite_years = 2780

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

Answer:

```
+----------+-----------+
| count(*) | NullOrNot |
+----------+-----------+
|    10000 | No        |
+----------+-----------+
```

SQL code used to arrive at answer:

```
SELECT count(*),
CASE WHEN 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

THEN 'Yes'
ELSE 'No'
END AS NullOrNot
FROM user
```

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: 3.708

ii. Table: Business, Column: Stars

min: 1.0 max:5.0 avg: 3.654

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, review_count
From business
Where city = 'Avon'
Group by stars
```

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

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

ii. Beachwood

SQL code used to arrive at answer:

```
Select stars, review_count
From business
Where city = 'Beachwood'
Group by stars
```

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

stars	review_count
2.0	8
2.5	3
3.0	3
3.5	3
4.0	69
4.5	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 id, name, review_count
From user
Order by review_count DESC LIMIT 3
```

Copy and Paste the Result Below:

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

I do not find significant correlation between increased review count and number of fans. I noticed that

- Amy, who joined on 2007-07-19 has 609 reviews and 503 fans;
- while Roanna who joined on 2006-03-28 has only 104 fans despite having 1039 reviews;
- yet another user, Yuri, who joined on 2008-01-03 has 76 fans despite having 1339 reviews.

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

Answer:

There are more reviews with the word ‘love’ in them over the word ‘hate’.

```
+--------------+--------------+
| love_in_text | hate_in_text |
+--------------+--------------+
|         1780 |          232 |
+--------------+--------------+
```

SQL code used to arrive at answer:

```
Select (Select count(text)
        from review
        where text like '%love%') as love_in_text,

       (Select count(text)
       from review
       where text like '%hate%') as hate_in_text
```

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.

Answer:
After analyzing different samples based on different combinations of city and category, I choose **Las Vegas** for city and **Shopping** for category. Obtained result includes one business in the group of 2-3 star rating, and two in 4-5 star rating.

i. **Do the two groups you chose to analyze have a different distribution of hours?**
I notice that the first group runs for more number of hours (15) whereas the second group runs for lesser and similar number of hours (8.5 and 9, respectively). Thus, the groups show a clear difference in distribution of hours.

ii. **Do the two groups you chose to analyze have a different number of reviews?**
Yes, the groups do have different number of reviews. The 2-3 star group has much lesser reviews compared to the 4-5 star group. However, the sample size is not as reliable to make concrete conclusions having exactly 3 businesses.

iii. **Are you able to infer anything from the location data provided between these two groups? Explain.**
Looking at the postal codes, 89121, 89161 and 89118 the three businesses seem to be in vicinity of each other, with group two being closer to each other and farther from group one.

SQL code used for analysis:

```
Select b.name as business_name,
       b.city, c.category, h.hours,
       b.stars, b.review_count, b.postal_code
From business b
Left Join category c
On c.business_id = b.id
Inner Join hours h
On h.business_id = b.id
Where (c.category = 'Shopping'
AND b.city = 'Las Vegas')
AND (b.stars between 2 and 3 or b.stars between 4 and 5)
Group By b.stars
```

Result:

business_name	hours	stars	review_count	postal_code
Walgreens	Saturday|8:00-22:00	2.5	6	89121
Red Rock Canyon Visitor Center	Saturday|8:00-16:30	4.5	32	89161
Desert Medical Equipment	Monday|8:00-17:00	5.0	4	89118

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:**
There are a total of 8480 businesses which are open (Group I) and only 1520 (Group II) which remain closed. Difference in average review count of both groups is as follows:

Open: 31.58
Close: 23.20

ii. **Difference 2:**
Difference in average star ratings is listed below:

Open: 3.68
Close: 3.52

SQL code used for analysis:

```
SELECT COUNT(DISTINCT(id)) business_count,
                        AVG(review_count),
                        SUM(review_count),
                        AVG(stars),
                        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:**
I noticed that the businesses are not classified exclusively. For example, there are businesses categorized as 'Shopping' and 'Shopping Centers', 'Food', 'Coffee & Tea', and 'Restaurants'. This creates a lot of ambiguity in the data, thus affecting the impact-fullness of the analyses. I therefore, chose clustering businesses for my analysis.

ii. **Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:**
Businesses can be clustered on more than one basis, such as the 'hours' of operation, categories such as automotive or utility provided by 'category' attribute or user-reviews like 'hot', 'funny', etc, given by the attributes from review table. Businesses can also be clustered into 1-star, 2-star and so on for which star rating would be required. Lastly, clustering on the basis of whether the business is still running or closed permanently can improve recommendations to users avoiding feeding only the ones where 'is_open' is set to True.
The data for such an analyses must contain attributes describing above mentioned features about each business, along with name of the business, city, postal code attributes.

iii. Output of your finished dataset:

id	name	address
-0DET7VdEQOJVJ_v6klEug	Flaming Kitchen	3235 York Regional Road 7
-2HjuT4yjLZ3b5f_abD87Q	Freeman's Car Stereo	4821 South Blvd
-CdstAUdEvci8GeJG8owpQ	Motors & More	2315 Highland Dr
-K4gAv8_vjx8-2BxkVeRkA	Baby Cakes	4145 Erie St
-PtTGvWsckUL8tTutHr6Ew	Snip-its Rocky River	21609 Center Ridge Rd
-ayZoW_iNDsunYXX_0x1YQ	Standard Restaurant Supply	2922 E McDowell Rd
-d9qyfNhLMQwVVg_raBKeg	What A Bagel	973 Eglinton Avenue W
-hjbcaxaU9yYXY2iI-49sw	Pinnacle Fencing Solutions	
-iu4FxdfxN4rU4Fu9BjiFw	Alterations Express	17240 Royalton Rd
-j4NsiRzSMrMk2N_bGH_SA	Extra Space Storage	2880 W Elliot Rd
-uiBBVWI6tMDm2JFbZFrOw	Gussied Up	1090 Bathurst St
0-aPEeNc2zVb5Gp-i7Ckqg	Buddy's Muffler & Exhaust	1509 Hickory Grove Rd
01xXe2m_z048W5gcBFpoJA	Five Guys	2641 N 44th St, Ste 100
06I2r8S3tHP_LwGnnkk6Uw	All Storage - Anthem	2620 W Horizon Ridge Pkwy
07h3mGtTovPJE660nX6E-A	Mood	1 Greenside Place
0AJF-USLN6K5T4caooDdjw	Starbucks	4605 E Chandler Blvd, Ste A
0B3W6KxkD3o4W4l6cq735w	Big Smoke Burger	260 Yonge Street
0IySwcfqwJjpHPsYwjpAkg	Subway	2904 Yorkmont Rd
0K2rKvqdBmiOAUTebcUohQ	Red Rock Canyon Visitor Center	1000 Scenic Loop Dr
0Ni7Stqt4RFWDGjOYRi2Bw	Scent From Above Company	2501 W Behrend Dr, Ste 67
0WBMEfqXQnEOAIkV-uCW6w	The Charlotte Room	19 Charlotte Street
0Y3lHyqRHfWOBuQlS1bM0g	PC Savants	11966 W Candelaria Ct
0aKsGxx7XP2TMs_fn_9xVw	Sweet Ruby Jane Confections	8975 S Eastern Ave, Ste 3-B
0cxO1Lx2Pi7u6ftWX3Wksg	Oinky's Pork Chop Heaven	22483 Emery Rd
0e-j5VcEn54EZT-FKCUZdw	Sushi Osaka	5084 Dundas Street W

city	state	postal_code	latitude	longitude	review_count	star_rating
Markham	ON	L3R 3P9	43.8484	-79.3487	25	2-3 stars
Charlotte	NC	28217	35.1727	-80.8755	8	3-4 stars
Las Vegas	NV	89102	36.1465	-115.167	7	4-5 stars
Willoughby	OH	44094	41.6399	-81.4064	5	3-4 stars
Rocky River	OH	44116	41.4595	-81.8587	18	2-3 stars
Phoenix	AZ	85008	33.4664	-112.018	15	3-4 stars
York	ON	M6C 2C4	43.6999	-79.4295	8	2-3 stars
Phoenix	AZ	85060	33.4805	-111.997	13	3-4 stars
Strongsville	OH	44136	41.3141	-81.8207	3	3-4 stars
Chandler	AZ	85224	33.3496	-111.892	5	3-4 stars
Toronto	ON	M5R 1W5	43.6727	-79.4142	6	4-5 stars
Gastonia	NC	28056	35.2772	-81.06	4	4-5 stars
Phoenix	AZ	85008	33.478	-111.986	63	3-4 stars
Henderson	NV	89052	36.0021	-115.102	3	3-4 stars
Edinburgh	EDH	EH1 3AA	55.957	-3.18502	11	1-2 stars
Phoenix	AZ	85048	33.3044	-111.984	52	2-3 stars
Toronto	ON	M4B 2L9	43.6546	-79.3805	47	2-3 stars
Charlotte	NC	28208	35.1903	-80.9288	7	3-4 stars
Las Vegas	NV	89161	36.1357	-115.428	32	4-5 stars
Scottsdale	AZ	85027	33.6656	-112.111	14	4-5 stars
Toronto	ON	M5V 2H5	43.6466	-79.3938	10	3-4 stars
Sun City	AZ	85373	33.6901	-112.319	11	4-5 stars
Las Vegas	NV	89123	36.015	-115.118	30	3-4 stars
North Randall	OH	44128	41.4352	-81.5214	3	2-3 stars
Toronto	ON	M9A 1C2	43.6452	-79.5324	8	4-5 stars

categories
Asian Fusion,Restaurants
Electronics,Shopping,Automotive,Car Stereo Installation
Home Services,Solar Installation,Heating & Air Conditioning/HVAC
Bakeries,Food
Beauty & Spas,Hair Salons
Shopping,Wholesalers,Restaurant Supplies,Professional Services,Wholesale Stores
Restaurants,Bagels,Breakfast & Brunch,Food
Home Services,Contractors,Fences & Gates
Shopping,Bridal,Dry Cleaning & Laundry,Local Services,Sewing & Alterations
Home Services,Self Storage,Movers,Shopping,Local Services,Home Decor,Home & Garden
Women's Clothing,Shopping,Fashion
Automotive,Auto Repair
American (New),Burgers,Fast Food,Restaurants
Truck Rental,Local Services,Self Storage,Parking,Automotive
Dance Clubs,Nightlife
Coffee & Tea,Food
Poutineries,Burgers,Restaurants
Fast Food,Restaurants,Sandwiches
Education,Visitor Centers,Professional Services,Special Education,Local Services,Community Servi
Home Cleaning,Local Services,Professional Services,Carpet Cleaning,Home Services,Office Cleaning
Event Planning & Services,Bars,Nightlife,Lounges,Pool Halls,Venues & Event Spaces
IT Services & Computer Repair,Electronics Repair,Local Services,Mobile Phone Repair
Food,Chocolatiers & Shops,Bakeries,Specialty Food,Desserts
Soul Food,Restaurants
Sushi Bars,Restaurants,Japanese,Korean

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

```
SELECT B.id,
                B.name,
                B.address,
                B.city,
                B.state,
                B.postal_code,
                B.latitude,
                B.longitude,
                B.review_count,
                CASE
                   WHEN B.stars BETWEEN 0 AND 1 THEN '0-1 stars'
                   WHEN B.stars BETWEEN 1 AND 2 THEN '1-2 stars'
                   WHEN B.stars BETWEEN 2 AND 3 THEN '2-3 stars'
                   WHEN B.stars BETWEEN 3 AND 4 THEN '3-4 stars'
                   WHEN B.stars BETWEEN 4 AND 5 THEN '4-5 stars'
                END AS star_rating,
                GROUP_CONCAT(DISTINCT(C.category)) AS categories,
                B.is_open
        FROM business B
        INNER JOIN hours H
        ON B.id = H.business_id
        INNER JOIN category C
        ON B.id = C.business_id
        INNER JOIN attribute A
        ON B.id = A.business_id
        GROUP BY B.id
```