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 =

ii. Business table =

iii. Category table =

iv. Checkin table =

v. elite_years table =

vi. friend table =

vii. hours table =

viii. photo table =

ix. review table =

x. tip table =

xi. user table =

Solution=

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.

i. Business =

ii. Hours =

iii. Category =

iv. Attribute =

v. Review =

vi. Checkin =

vii. Photo =

viii. Tip =

ix. User =

x. Friend =

xi. Elite_years =

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

Solution=

i. Business =1000 distinct records for primary key 'id' of business table

ii. Hours =1562 distinct records for foreign key 'business_id' of hours table

iii. Category =2643 distinct records for foreign key 'business_id' of category table

iv. Attribute =1115 distinct records for foreign key 'business_id' of attribute table

v. Review =1000 distinct records for primary key 'id',8090 distinct records for primary key 'business_id' and 9581 distinct records for primary key 'user_id' from the review table

vi. Checkin = 493 distinct records for primary key 'user_id' from the checkin table

vii. Photo =6493 distinct records for foreign key 'business_id' and 1000 distinct records for primary key 'id' from the photo table

viii. Tip =3979 distinct records for foreign key 'business_id' and 537 distinct records for foreign key 'user_id' from the tip table

ix. User = 1000 distinct records for primary key 'id' of user table

x. Friend = 11 distinct records for foreign key 'id' from the friend table

xi. Elite_years =2780 distinct records for foreign key 'user_id' from the elite_years table

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 *
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:

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: SELECT SUM(review_count) AS reviews,city

FROM business

GROUP BY city

ORDER BY SUM(review_count) DESC;

Copy and Paste the Result Below:

```
+---------+-----------------+
| reviews | city            |
+---------+-----------------+
|   82854 | Las Vegas       |
|   34503 | Phoenix         |
|   24113 | Toronto         |
|   20614 | Scottsdale      |
|   12523 | Charlotte       |
|   10871 | Henderson       |
|   10504 | Tempe           |
|    9798 | Pittsburgh      |
```

| 9448 | MontrÃ©al |

| 8112 | Chandler |

| 6875 | Mesa |

| 6380 | Gilbert |

| 5593 | Cleveland |

| 5265 | Madison |

| 4406 | Glendale |

| 3814 | Mississauga |

| 2792 | Edinburgh |

| 2624 | Peoria |

| 2438 | North Las Vegas |

| 2352 | Markham |

| 2029 | Champaign |

| 1849 | Stuttgart |

| 1520 | Surprise |

| 1465 | Lakewood |

| 1155 | Goodyear |

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

(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)

FROM business

WHERE city='Avon'

GROUP BY stars;

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

| stars | COUNT(stars) |

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

| 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)

FROM business

WHERE city='Beachwood'

GROUP BY stars;

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

| stars | COUNT(stars) |

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

| 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

GROUP BY name

ORDER BY review_count DESC

LIMIT 3;

Copy and Paste the Result Below:

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

| name | review_count | id |

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

| Gerald | 2000 | -G7ZkI1wIWBBmD0KRy_sCw |

| .Hon | 1246 | -K2Tcgh2EKX6e6HqqIrBIQ |

| eric | 1116 | -gokwePdbXjfS0iF7NsUGA |

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

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

No,Posing more reviews is not directly related with more fans.It also depends on the yelping since factor.

The longer they have been yelping,the more number of genuine reviews they give,hence resulting in more number of fans.

In most of the cases,we see that with increase in the number of reviews,a corresponding increase in the nuumber of fans is not seen.

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
-4viTt9UC44lWCFJwIeMNQ	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 |

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

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

Answer:Number of reviews with the word 'love' in them =1780

Number of reviews with the word 'hate' in them =232

Hence,there are more reviews with the word love in it.

SQL code used to arrive at answer:

```
SELECT COUNT(text)
FROM review
WHERE text LIKE '%love%';
SELECT COUNT(text)
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
GROUP BY name
ORDER BY fans DESC
LIMIT 10;
```

Copy and Paste the Result Below:

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

| id | name | fans |

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

| -G7ZkI1wIWBBmD0KRy_sCw | Gerald | 253 |

| -lh59ko3dxChBSZ9U7LfUw | Lissa | 120 |

| -cv9PPT7IHux7XUc9dOpkg | bernice | 105 |

| -DFCC64NXgqrxIO8aLU5rg | Roanna | 104 |

| -K2Tcgh2EKX6e6HqqIrBIQ | .Hon | 101 |

| -C-l8EHSLXtZZVfUAUhsPA | Nieves | 80 |

| -dw8f7FLaUmWR7bfJ_Yf0w | Sui | 78 |

| -KFjONqNDuBfKDeKAoA-bg | Koizumi | 73 |

| -J06qsoD12jmjAU75sCbjA | rebecca | 69 |

| -5IOjA-DymtPai1i8toROA | Princeton | 64 |

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

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?

I chose 'Toronto' as city and 'Food' as category.Yes,they have different distribution of working hours.

Working hours for 2-3 star rating businesses are 7 and working hours for 4-5 star rating businesses are 13.

```
SELECT stars
,CASE WHEN stars>=4 THEN "4-5_stars"
WHEN stars>=2 THEN "2-3_stars"
ELSE "Below 2"
END Star_rank
,business.city,COUNT(DISTINCT hours.business_id) AS Company_count
,COUNT(hours.hours) AS working_hours
,category.category
FROM((business JOIN hours ON business.id=hours.business_id)
JOIN category ON business.id=category.business_id)
WHERE city='Toronto' AND category= 'Food'
GROUP BY Star_rank;
```

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

I chose 'Toronto' as city and 'Food' as category.The two categories have different number of reviews.

Number of reviews for 2-3 star rating businesses are 70 and number of reviews for 4-5 star rating businesses are 272.

```
SELECT stars
,CASE WHEN stars>=4 THEN "4-5_stars"
WHEN stars>=2 THEN "2-3_stars"
ELSE "Below 2"
END Star_rank
,business.city,COUNT(DISTINCT hours.business_id) AS Company_count
```

```
,COUNT(hours.hours) AS working_hours

,SUM(review_count) AS Number_of_reviews

,category.category

FROM((business JOIN hours ON business.id=hours.business_id)

JOIN category ON business.id=category.business_id)

WHERE city='Toronto' AND category= 'Food'

GROUP BY Star_rank;
```

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

2-3 star rating businesses are located near one another.4-5 star rating businesses are apart from one anothe.

SQL code used for analysis:

```
SELECT stars

,CASE WHEN stars>=4 THEN '4-5 Stars'

WHEN stars>=2 THEN '2-3 Stars'

ELSE 'below 2'

END Star_Rank

,category,city,postal_code,address,neighborhood

FROM business INNER JOIN category ON business.id=category.business_id

WHERE city='Las Vegas' AND category='Shopping'

ORDER BY Star_Rank;
```

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 have a greater number of reviews when compared to the businesses that are closed.

ii. Difference 2:The average stars given are very closed to each other 3.68 vs. 3.52.

We can assume that businesses which are closed was not solely due to poor services or poor quality.

SQL code used for analysis:

```
SELECT
CASE
WHEN is_open='0' THEN 'closed'
WHEN is_open='1' THEN 'open'
ELSE 'NO_Status'
END Status
,COUNT( DISTINCT id) AS Company_num
,SUM(review_count) AS num_of_reviews
,ROUND(AVG(review_count),2) AS AVG_REVIEW
,ROUND(AVG(stars),2) AS AVG_STARS
FROM business
GROUP BY Status;
```

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:

The analysis done is to find out which categories of business have higher stars rating and more number of companies.

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

For this analysis, we will need data such as id, stars, and review count from the business table and category table.

We are counting the numbers of companies within each category,

the average stars given by the consumers to see how they perform,

and the total reviews given to see if the data is relevant and ensure itâ€™s not biased.

We are only analysing the categories with at least 10 companies and an average of 3+ stars to reduce any irrelevant data.

iii. Output of your finished dataset:

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

| category | num_companies | total_reviews | Average_stars |

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

| Local Services | 12 | 100 | 4.21 |

| Active Life | 10 | 131 | 4.15 |

| Health & Medical | 17 | 203 | 4.09 |

| Home Services | 16 | 94 | 4.0 |

| Shopping | 30 | 977 | 3.98 |

| Beauty & Spas | 13 | 119 | 3.88 |

| American (Traditional) | 11 | 1128 | 3.82 |

| Food | 23 | 1781 | 3.78 |

| Bars | 17 | 1322 | 3.5 |

| Nightlife | 20 | 1351 | 3.48 |

| Restaurants | 71 | 4504 | 3.46 |

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

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

```
SELECT category,
COUNT(DISTINCT id) AS num_companies,
SUM(review_count) AS total_reviews,
ROUND(AVG(stars),2) AS Average_stars
FROM business INNER JOIN category
ON business.id=category.business_id
GROUP BY category
HAVING Average_stars>=3 AND num_companies >=10
ORDER BY Average_stars DESC;
```