SQL for DATA SCIENCE (Peer Review Project)

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 number of distinct records for the primary keys in each of the tables listed below:

i. Business = 10000
ii. Hours = 1562
iii. Category = 2643
iv. Attribute = 1115
v. Review = 10000
vi. Checkin = 493
vii. Photo = 10000
viii. Tip = 537,3979(two secondary keys)
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 *
FROM user
WHERE ('*') IS NULL; /*Selecting all the columns in the where statement*/
```

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:	max:	avg:
1	5	3.082

ii. Table: Business, Column: Stars

min:	max:	avg:
1	5	3.6549

iii. Table: Tip, Column: Likes

min:	max:	avg:
0	2	0.0144

iv. Table: Checkin, Column: Count

min:	max:	avg:
1	53	1.9414

v. Table: User, Column: Review_count

min:	max:	avg:
0	2000	24.2995

5. List the cities with the most reviews in descending order:

SQL code used to arrive at answer:

```
SELECT city, review_count /*I have included review_count to get feel of
no's*/
FROM business
ORDER BY review_count DESC;
```

Copy and Paste the Result Below:

city	review_count
Las Vegas	3873
Montréal	1757
Gilbert	1549
Las Vegas	1410
Las Vegas	1389
Las Vegas	1252
Las Vegas	1116
Las Vegas	1084
Las Vegas	961
Gilbert	902
Las Vegas	864
Scottsdale	823
Las Vegas	821
Las Vegas	786
Henderson	785
Toronto	778
Las Vegas	768
Las Vegas	758
Scottsdale	726
Cleveland	723
Las Vegas	720
Charlotte	715
Phoenix	711
Las Vegas	706
Phoenix	700

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,name /*I didn't understand the count criteria*/
FROM business
WHERE city='Avon';
```

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

stars	name
2.5	Helen & Kal's
4.0	Marc's
5.0	Hoban Pest Control
3.5	Light Salon & Spa
1.5	Portrait Innovations
3.5	Winking Lizard Tavern
4.5	Dervish Mediterranean & Turkish Grill
3.5	Mulligans Pub and Grill
2.5	Mr. Handyman of Cleveland's Northwest Suburbs
4.0	Cambria hotel & suites Avon - Cleveland

ii. Beachwood

SQL code used to arrive at answer:

```
SELECT stars,name /*I didn't understand the count criteria*/
FROM business
WHERE city= 'Beachwood';
```

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

stars	name
3.0	Maltz Museum of Jewish Heritage
3.0	Charley's Grilled Subs
4.5	Sixth & Pine
5.0	Beechmont Country Club
4.0	Hyde Park Prime Steakhouse
4.5	Origins
5.0	Fyodor Bridal Atelier
2.0	College Planning Network
3.5	Lucky Brand Jeans
3.5	American Eagle Outfitters
5.0	Shaker Women's Wellness
2.5	Avis Rent A Car
5.0	Cleveland Acupuncture
5.0	Studio Mz

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; /*Limiting the outcome to 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

Please explain your findings and interpretation of the results:

review_count	fans	
2000	253	
1629	50	
1339	76	
1246	**101**	
1215	**126**	As can be even with more number of reviews people have less
1153	311	fan following, which states that there is no correlation.
1116	16	I believe People have more fan following
1039	104	because of the genuineness of the review, so its more about
968	497	quality than quantity
930	173	
904	38	
864	43	
862	124	
861	115	
842	85	
836	37	
834	120	
813	159	
775	61	
754	78	
702	35	
696	10	
694	101	
676	**25**	
675	**45**	

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

Answer:
For love the count was 1780
For hate the count was 232,
Hence more reviews had ‘love’ in it.

SQL code used to arrive at answer:

```
SELECT COUNT(text)
FROM review
WHERE text LIKE '%hate%'; /*I used separate codes for hate and love, also the
word should also be counted when it is in between different words*/

SELECT COUNT(text)
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

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 name, fans
FROM user
WHERE name IN (SELECT name
FROM user
ORDER BY useful DESC
LIMIT 10) AND fans>=120 AND funny>=2913/*This is the lower limit for fans i.e
10th person in sorted fan list have 120 fans and similarly for funny the
count is greater than 2913 for the top ten people*/
ORDER BY fans DESC;
```

Copy and Paste the Result Below:

name	fans
Harald	311
Christine	173
William	126
Fran	124

Please explain your findings and interpretation of the results:

The relationship is a medium one, as there are four person which are common in all the three list. (40%)
This leads to the interpretation that there is less correlation between user review being funny, useful and lovable(fan). This makes sense as it happens in real life, the thing you find helpful useful might not be funny.

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

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

iii. Are you able to infer anything from the location data provided between these two groups? Explain.
Not much, since neighborhood contain NULL values for one group.

SQL code used for analysis:

```
SELECT city, category, stars, hours, review_count, neighborhood
FROM category
INNER JOIN business ON business.id=category.business_id
INNER JOIN hours    ON business.id=hours.business_id
WHERE city="Las Vegas" AND category="Shopping"
GROUP BY stars
HAVING (stars<=3 AND stars>=2) OR (stars<=5 AND stars>=4);
```

OUTPUT

city	category	stars	hours	review_count	neighborhood
Las Vegas	Shopping	2.5	Saturday|8:00-22:00	6	Eastside
Las Vegas	Shopping	4.5	Saturday|8:00-16:30	32	
Las Vegas	Shopping	5.0	Monday|8:00-17:00	4	

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:

As can be seen from the output that most of the businesses which are still open have higher number of review counts, which makes sense as well.

ii. Difference 2:

As for the second difference I have used stars, it’s visible that many businesses which are up and running have stars in the range 3-5.

SQL code used for analysis:

```
SELECT  name,stars,review_count, is_open
FROM business
```

```
WHERE stars>=3 AND stars<=5
ORDER BY review_count DESC;
/*I am not using a group by funtion as it limits the output to only 2 rows,
which is very less for any inferential statistics.
I am using stars and review_count to assess the business capability of
running
*/
```

OUTPUT

name	stars	review_count	is_open
The Buffet	3.5	3873	1
Schwartz's	4.0	1757	1
Joe's Farm Grill	4.0	1549	1
Carson Kitchen	4.5	1410	1
Delmonico Steakhouse	4.0	1389	1
Le Thai	4.0	1252	1
Scarpetta	4.0	1116	1
Diablo's Cantina	3.0	1084	1
Joyride Taco House	4.0	902	1
Yonaka Modern Japanese	4.5	864	1
Breakfast Club- Scottsdale	3.5	823	1
VegeNation	4.5	821	1
Lazy Dog Restaurant & Bar	4.0	786	1
Lucille's Smokehouse Bar-B-Que	4.0	785	1
Salad King Restaurant	3.5	778	1
Big Wong Restaurant	4.0	768	1
Picasso	4.5	758	1
Cowboy Ciao	4.0	726	1
West Side Market	4.5	723	1
Bruxie	4.5	720	1
Pinky's Westside Grill	4.0	715	1
Switch Restaurant & Wine Bar	4.0	711	1
Kinh Do	4.0	706	1
Matt's Big Breakfast	4.0	700	0
Toronto Pearson International Airport	3.0	683	1

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 would like to build a predictive model where with the help of previous star ratings, review_counts, neighborhood and whether the review has contents like “love” in it or not, I could predict whether the business will remain open or not.

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

Much of the existing will be useful for the analysis, apart from the existing data, I would really like to incorporate where a business is segregated on accounts of being cheap, expensive, mildly expensive. So a user could have identified aforementioned categories and based on that there would have been a column in the review data having a name like “Spending limit”

iii. Output of your finished dataset:
is_open would be my output, predicting whether a business will run or not.

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

```
CREATE TABLE Yelpdataset (
    star_ratings float
    is_open binary,
    review_counts int,
    name_business varchar(255),
    review varchar(255),
    neighborhood varchar(255)
);
/*After the table is created I can make use of data to built predictive
models*/
```