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

select count(\*) from attribute;

ii. Business table = 10000

select count(\*) from business;

iii. Category table = 10000

select count(\*) from category;

iv. Checkin table = 10000

select count(\*) from Checkin;

v. elite\_years table = 10000

select count(\*) from elite\_years;

vi. friend table = 1 0000

select count(\*) from elite\_years;

vii. hours table = 10000

select count(\*) from friend;

viii. photo table = 10000

select count(\*) from photo;

ix. review table = 10000

select count(\*) from review;

x. tip table = 10000

select count(\*) from tip;

xi. user table =10000

select count(\*) from user;

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 =10000 (primary key “id”)

ii. Hours = 1562 (foreign key “business\_id”)

iii. Category = 2643 (“Business id “is foreign key)

iv. Attribute =1115 (“Business id “is foreign key)

v. Review = 10000. (primary key “id”)

vi. Checkin = 493. (“Business id “is foreign key)

vii. Photo = 10000 (primary key “id”)

viii. Tip = 537 ( “user\_id “ is foreign )

ix. User = 10000 (primary key “id”)

x. Friend = 11 ( “user\_id “ is foreign )

xi. Elite\_years = 2780 ( “user\_id “ is foreign )

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:

SQL code used to arrive at answer:

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:

/\* by grouping with city counted the number of reviews in each city\*/

SELECT city, SUM (review\_count) reviews

FROM business

GROUP BY city

ORDER BY reviews DESC

Copy and Paste the Result Below:

A screenshot of a cell phone

Description automatically generated

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,

SUM(review\_count) count

FROM business

WHERE city == 'Avon'

GROUP BY stars

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

A screenshot of a cell phone

Description automatically generated

ii. Beachwood

SQL code used to arrive at answer:

SELECT stars,

SUM (review\_count) count

FROM business

WHERE city == ‘Beachwood’

GROUP BY stars

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

A screenshot of a cell phone

Description automatically generated

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:

A screenshot of a cell phone

Description automatically generated

8. Does posing more reviews correlate with more fans?

Yes, there is some correlation but not all reviews correlated with fans.

Please explain your findings and interpretation of the results:

SELECT id,

name,

review\_count,

fans,

yelping\_since

FROM user

ORDER BY fans DESC

A picture containing text, newspaper

Description automatically generated

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

Answer:

 A screenshot of a cell phone

Description automatically generated

SQL code used to arrive at answer:

select count (text) select count(text)

from review from review

where text like "%love%" 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

LIMIT 10

Copy and Paste the Result Below:

A close up of text on a white background

Description automatically generated

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

SELECT name,

fans,

useful,

funny,

review\_count,

yelping\_since

FROM user

ORDER BY fans DESC

Copy and Paste the Result Below:

A close up of text on a white background

Description automatically generated

Please explain your findings and interpretation of the results:

Yes, it looks there are correlation between fans and being listed as useful and funny. It looks there are some outliers in the useful and funny column especially at Herald, Sui, Nicole.

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.

select c.category,b.\*

from business b

inner join category c

on b.id = c.business\_id

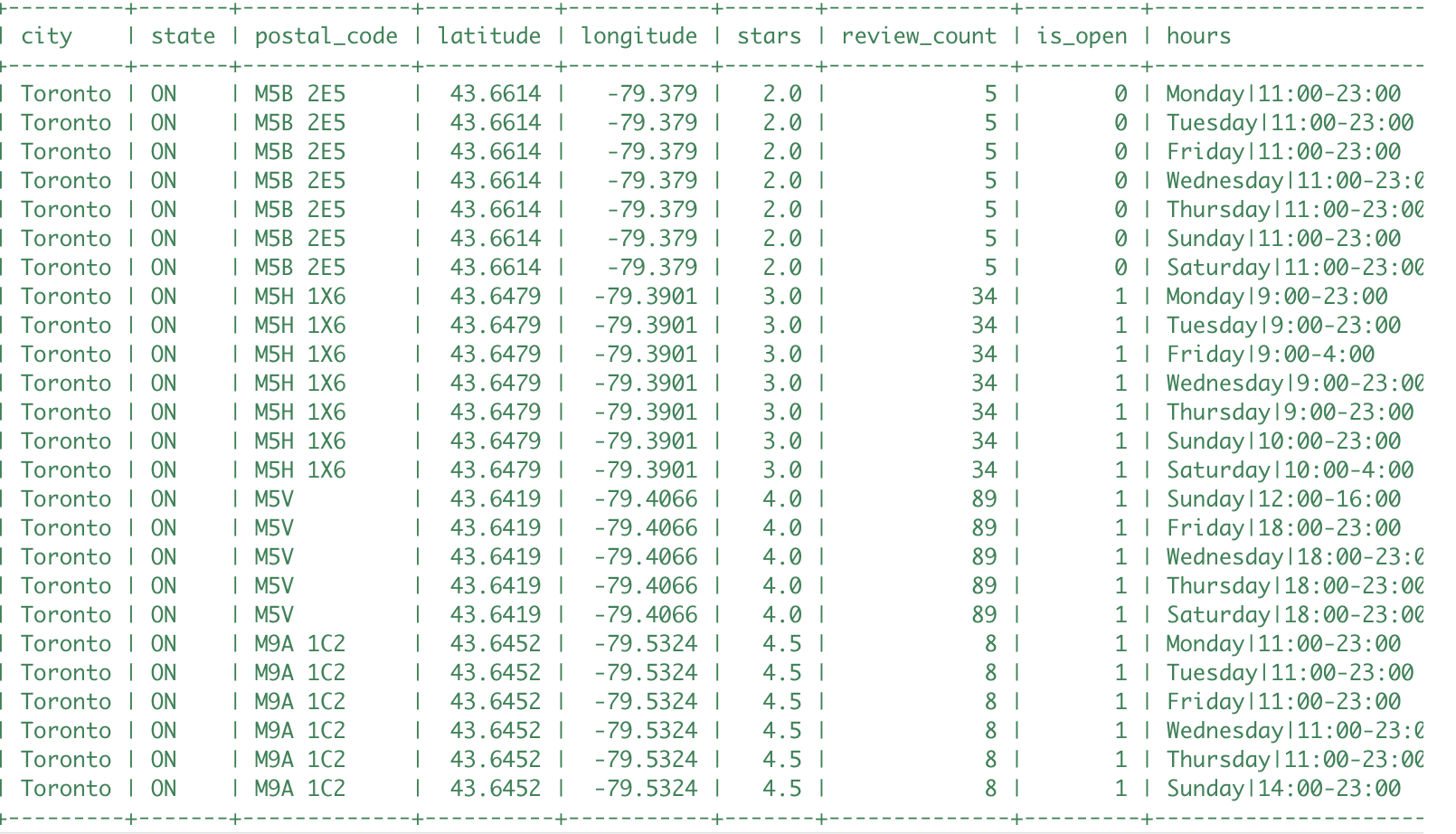
where city like 'toronto'

and category like 'Restaurants'

group by stars

order by stars

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



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

No, lowest and highest rating groups review count is less or no much difference.but the review count for 3 to 4 rating are high.

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

It looks like in core downtown the ratings are very less but the high stars given to Etobicoke area.

SQL code used for analysis:

select buss.\*,

h.hours

from hours h

inner join

(select c.category,

b.\*

from business b

inner join category c

on b.id = c.business\_id

where city like ‘Toronto'

and category like 'Restaurants'

group by stars

order by stars) buss

on h.business\_id = buss.id

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 business that are open have more number go reviews compared to the business that are closed

Average (review count) :

0: 35261

1: 269300

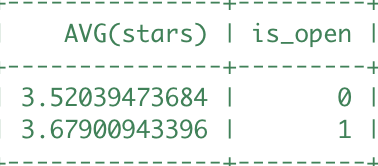
ii. Difference 2:

The average star rating is bit high for the business that are open compared to closed business.

Avg (Stars):

0 : 3.52

1 : 3.67



SQL code used for analysis:

SELECT COUNT(DISTINCT(id)),

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:

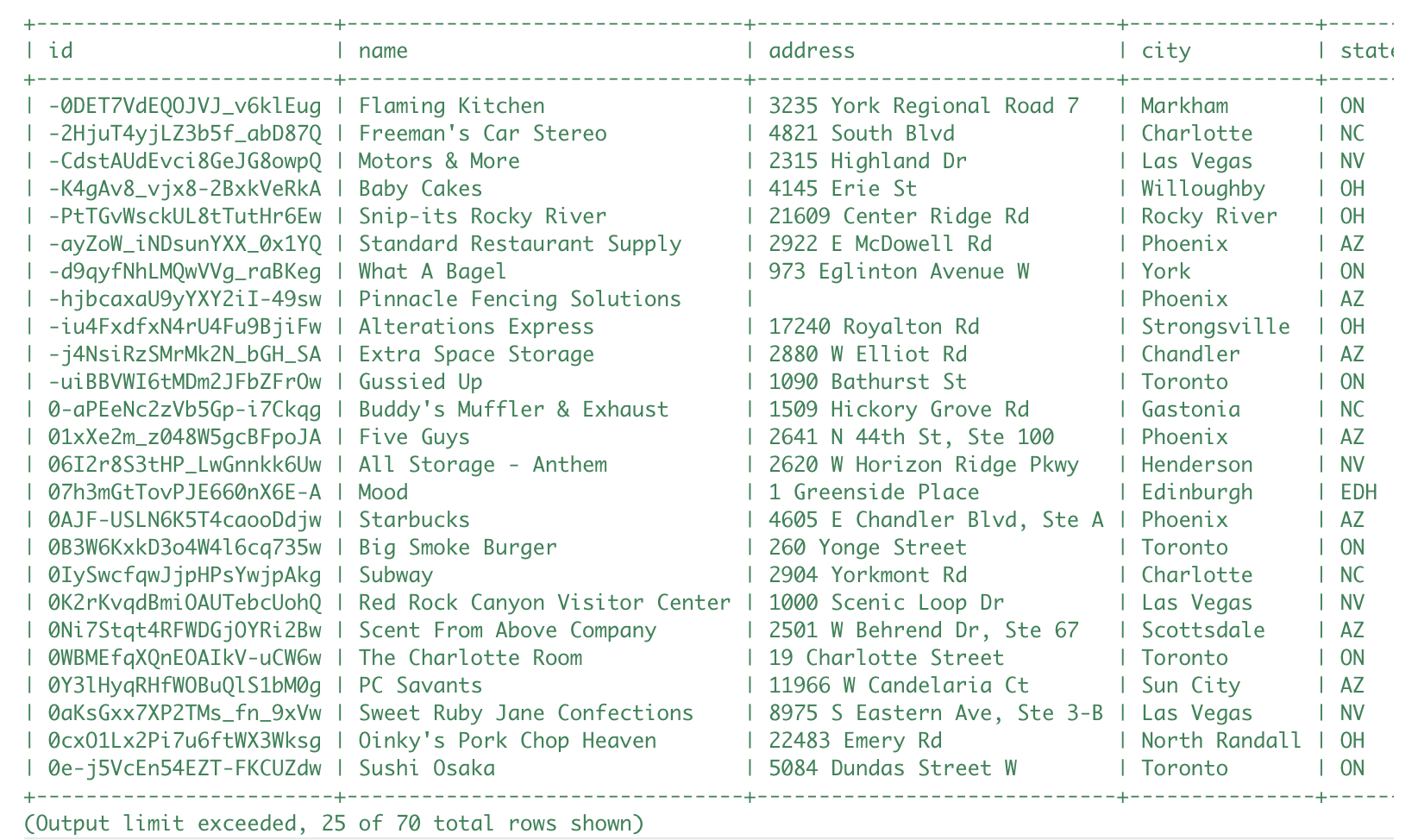
i. Indicate the type of analysis you chose to do:

Predicting whether a business will stay open or close. We wish not to explicitly examine the text of the reviews, but this would be an interesting analysis.

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

To better help businesses understand the importance of different factors which will help their business stay open. Some data that may be important; number of reviews, star rating of business, hours open, and of course location location location. We will gather the latitude and longitude as well as city, state, postal\_code, and address to make processing easier later on. Categories and attributes will be used to better distinguish between different types of businesses. `is\_open` will determine which business is open and which business have closed (not hours) but permanently.

iii. Output of your finished dataset:



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,

B.stars,

MAX(CASE

WHEN H.hours LIKE "%monday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS monday\_hours,

MAX(CASE

WHEN H.hours LIKE "%tuesday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS tuesday\_hours,

MAX(CASE

WHEN H.hours LIKE "%wednesday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS wednesday\_hours,

MAX(CASE

WHEN H.hours LIKE "%thursday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS thursday\_hours,

MAX(CASE

WHEN H.hours LIKE "%friday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS friday\_hours,

MAX(CASE

WHEN H.hours LIKE "%saturday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS saturday\_hours,

MAX(CASE

WHEN H.hours LIKE "%sunday%" THEN TRIM(H.hours,'%MondayTuesWednesThursFriSatSun|%')

END) AS sunday\_hours,

GROUP\_CONCAT(DISTINCT(C.category)) AS categories,

GROUP\_CONCAT(DISTINCT(A.name)) AS attributes,

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