#Final Assignment

SQL for Data Science

(University of California, Davis)

Himanshu Saini

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 =10,000
ii. Business table =10,000
iii. Category table =10,000
iv. Checkin table =10,000
v. elite_years table =10,000
vi. friend table =10,000
vii. hours table =10,000
viii. photo table =10,000
ix. review table =10,000
x. tip table =10,000
xi. user table =10,000

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 =10,000(primary key,'id')
ii. Hours =1,562(foreign key,'business_id')
iii. Category =2,643(foreign key,'business_id')
iv. Attribute =1,115(foreign key,'business_id')
v. Review =10,000(primary key,'id');9,581(foreign key,'user_id'),8,090(foreign key,'business_id')
vi. Checkin =493(foreign key,'business_id')
vii. Photo =10,000(primary key,'id');6,493(foreign key,'photo')
viii. Tip =3,979(foreign key,'business_id');537(foreign key,'user_id')
ix. User =10,000(primary key,'id')
x. Friend =11(foreign key,'user_id')
xi. Elite_years =2,780(foreign key,'user_id')
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 COUNT(*)
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:1.0 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 review_nums
FROM business
GROUP BY city
ORDER BY review_nums DESC;
```

Copy and Paste the Result Below:

```
+-----------------+-------------+
| city | review_nums |
+-----------------+-------------+
| 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, count(stars) as distribute_stars
from business
where city="Avon"
group by stars;
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

stars	distribute_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) as distribute_stars
from business
where city="Beachwood"
group by stars;
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

stars	distribute_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
order by review_count desc
limit 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?

Please explain your findings and interpretation of the results:

No. The more reviews does not necessarily correlate with more fans.

SQL code is listed as below:

```
select name, review_count, fans
from user
order by review_count desc;
```

And the result as below:

```
+----------+--------------+------+
| name | review_count | fans |
+----------+--------------+------+
| Gerald | 2000 | 253 |
| Sara | 1629 | 50 |
| Yuri | 1339 | 76 |
| .Hon | 1246 | 101 |
| William | 1215 | 126 |
| Harald | 1153 | 311 |
| eric | 1116 | 16 |
| Roanna | 1039 | 104 |
| Mimi | 968 | 497 |
| Christine | 930 | 173 |
| Ed | 904 | 38 |
| Nicole | 864 | 43 |
| Fran | 862 | 124 |
```

```
| Mark | 861 | 115 |
| Christina | 842 | 85 |
| Dominic | 836 | 37 |
| Lissa | 834 | 120 |
| Lisa | 813 | 159 |
| Alison | 775 | 61 |
| Sui | 754 | 78 |
| Tim | 702 | 35 |
| L | 696 | 10 |
| Angela | 694 | 101 |
| Crissy | 676 | 25 |
| Lyn | 675 | 45 |
+----------+--------------+------+
(Output limit exceeded, 25 of 10000 total rows shown)
```

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”.
SQL code used to arrive at answer:

```
select count(text) as love_review
from review
where text like "%love%";
select count(text) as hate_review
from review
where text like "%hate%";
+-------------+
| love_review |
+-------------+
| 1780 |
+-------------+
+-------------+
| hate_review |
+-------------+
| 232 |
+-------------+
```

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.

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

Yes. The businesses which have comparatively higher rating
stars open late and close at around midnight. While the businesses having lower rating stars open in the morning and close earlier.

```
select b.stars,b.city,b.name,c.category, h.hours
from (business as b inner join category as c
on b.id=c.business_id)
inner join hours as h on b.id=h.business_id
where b.city='Toronto' and c.category='Food';
```

```
+-------+--------+-------------+---------+--------------------+
| stars | city | name | category | hours
|
+-------+--------+-------------+---------+--------------------+
| 4.5 | Toronto | Cabin Fever | Food | Monday|16:00-2:00|
| 4.5 | Toronto | Cabin Fever | Food | Tuesday|18:00-2:00|
| 4.5 | Toronto | Cabin Fever | Food | Friday|18:00-2:00|
| 4.5 | Toronto | Cabin Fever | Food | Wednesday|18:00-2:00 |
| 4.5 | Toronto | Cabin Fever | Food | Thursday|18:00-2:00 |
| 4.5 | Toronto | Cabin Fever | Food | Sunday|16:00-2:00|
| 4.5 | Toronto | Cabin Fever | Food | Saturday|16:00-2:00 |
| 2.5 | Toronto | Loblaws | Food | Monday|8:00-22:00|
| 2.5 | Toronto | Loblaws | Food | Tuesday|8:00-22:00|
| 2.5 | Toronto | Loblaws | Food | Friday|8:00-22:00|
| 2.5 | Toronto | Loblaws | Food | Wednesday|8:00-22:00 |
| 2.5 | Toronto | Loblaws | Food | Thursday|8:00-22:00 |
| 2.5 | Toronto | Loblaws | Food | Sunday|8:00-22:00|
| 2.5 | Toronto | Loblaws | Food | Saturday|8:00-22:00 |
| 4.0 | Toronto | Halo Brewery | Food | Tuesday|15:00-21:00 |
| 4.0 | Toronto | Halo Brewery | Food | Friday|15:00-21:00|
| 4.0 | Toronto | Halo Brewery | Food | Wednesday|15:00-21:00 |
| 4.0 | Toronto | Halo Brewery | Food | Thursday|15:00-21:00 |
| 4.0 | Toronto | Halo Brewery | Food | Sunday|11:00-21:00 |
| 4.0 | Toronto | Halo Brewery | Food | Saturday| 11:00-21:00 |
+-------+--------+-------------+---------+--------------------+
```

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

The restaurants with higher rating stars tend to have more reviews while lower-rated restaurants tend to have fewer reviews.

----------SQL code----------

```
select b.stars,b.city,b.name,c.category, b.review_count
from (business as b inner join category as c
on b.id=c.business_id)
inner join hours as h on b.id=h.business_id
where b.city='Toronto' and c.category='Food'
group by b.stars;
```

```
+-------+--------+-------------+---------+--------------+
| stars | city | name | category | review_count |
+-------+--------+-------------+---------+--------------+
| 2.5 | Toronto | Loblaws | Food | 10 |
| 4.0 | Toronto | Halo Brewery | Food | 15 |
| 4.5 | Toronto | Cabin Fever | Food | 26 |
+-------+--------+-------------+---------+--------------+
```

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

The address of these three restaurants are different, thus cannot infer anything from that.

----------SQL code used for analysis-------------

```
select b.stars,b.city,b.name,c.category, b.address
from (business as b inner join category as c
on b.id=c.business_id)
inner join hours as h on b.id=h.business_id
where b.city='Toronto' and c.category='Food'
group by b.stars;
```

stars	city	name	category	address
2.5	Toronto	Loblaws	Food	2280 Dundas Street W
4.0	Toronto	Halo Brewery	Food	247 Wallace Avenue
4.5	Toronto	Cabin Fever	Food	1669 Bloor Street W

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 ones that are still open have higher rating stars in average than those that are closed.

ii. Difference 2:
The ones that are closed have fewer reviews in average than those those that are still open.

----------SQL code used for analysis-----------

```
select avg(stars) as avg_stars, avg(review_count) as
avg_review, is_open
from business
group by is_open;
```

```
+--------------+--------------+---------+
| avg_stars | avg_review | is_open |
+--------------+--------------+---------+
| 3.52039473684 | 23.1980263158 | 0 |
| 3.67900943396 | 31.7570754717 | 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 chose to find the correlation between hours and stars for coffeeshops.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:
The reason of analyzing the correlation between hours and stars for coffeeshops is that people have different preferences when it comes to the time of having coffee&tea, I would like to find whether the opening hours will effect people's attitude towards coffeeshops.

I will analyze the star ratings and opening hours for the category of "Coffee & Tea".

iii. Output of your finished dataset:

```
+-------------+-------+---------------------+
| name | stars | hours |
+-------------+-------+---------------------+
| Starbucks | 3.0 | Saturday|5:00-20:00 |
| Koko Bakery | 4.0 | Saturday|9:00-20:00 |
| Cabin Fever | 4.5 | Saturday|16:00-2:00 |
+-------------+-------+---------------------+
```

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

```
select business.name,
business.stars,
hours.hours
FROM (business inner join CATEGORY
on business.id=CATEGORY.business_id) inner join hours
on business.id=hours.business_id
WHERE CATEGORY.CATEGORY='Coffee & Tea'
group by business.stars
order by business.stars;
```