**Introduction:**

We will be using a dataset from a US-based organization called Yelp, which provides a platform for users to provide reviews and rate their interactions with a variety of organizations – businesses, restaurants, health clubs, hospitals, local governmental offices, charitable organizations, etc. Yelp has made a portion of this data available for personal, educational, and academic purposes.

**Data Scientist Role Play:**

Profiling and Analyzing the Yelp Dataset Worksheet

This is a 2-part assignment. In the first part, you are asked a series of questions that will help you profile and understand the data just like a data scientist would. For this first part of the assignment, you will be assessed both on the correctness of your findings, as well as the code you used to arrive at your answer. You will be graded on how easy your code is to read, so remember to use proper formatting and comments where necessary.

In the second part of the assignment, you are asked to come up with your own inferences and analysis of the data for a particular research question you want to answer. You will be required to prepare the dataset for the analysis you choose to do. As with the first part, you will be graded, in part, on how easy your code is to read, so use proper formatting and comments to illustrate and communicate your intent as required.

**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 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

ii. Hours = 1562 (foreign key -> business\_id)

iii. Category = 2643 (foreign key -> business\_id)

iv. Attribute = 1115 (foreign key -> business\_id)

v. Review = 10000

vi. Checkin = 493 (foreign key -> business\_id)

vii. Photo = 10000

viii. Tip = 3979 (foreign key -> business\_id)

ix. User = 10000

x. Friend = 11 (foreign key -> user\_id)

xi. Elite\_years = 2780 (foreign key -> user\_id)

*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 coalesce(id, name, review\_count, yelping\_since, useful, funny, cool, fans, average\_stars,

compliment\_hot, compliment\_more, compliment\_profile, compliment\_cute, compliment\_list,

compliment\_note, compliment\_plain, compliment\_cool, compliment\_funny, compliment\_writer,

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 city, count(review\_count) as review\_cnt

from business

group by city

order by review\_cnt desc

Copy and Paste the Result Below:

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

| city | review\_cnt |

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

| Las Vegas | 1561 |

| Phoenix | 1001 |

| Toronto | 985 |

| Scottsdale | 497 |

| Charlotte | 468 |

| Pittsburgh | 353 |

| MontrÃ©al | 337 |

| Mesa | 304 |

| Henderson | 274 |

| Tempe | 261 |

| Edinburgh | 239 |

| Chandler | 232 |

| Cleveland | 189 |

| Gilbert | 188 |

| Glendale | 188 |

| Madison | 176 |

| Mississauga | 150 |

| Stuttgart | 141 |

| Peoria | 105 |

| Markham | 80 |

| Champaign | 71 |

| North Las Vegas | 70 |

| North York | 64 |

| Surprise | 60 |

| Richmond Hill | 54 |

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

(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 city, min(stars) as min\_star, max(stars) as max\_star, avg(stars) as avg\_star

from business

where city = 'Avon'

Copy and Paste the Resulting Table Below (2 columns Ã¢â‚¬â€œ star rating and count):

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

| city | min\_star | max\_star | avg\_star |

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

| Avon | 1.5 | 5.0 | 3.45 |

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

ii. Beachwood

SQL code used to arrive at answer:

select city, min(stars) as min\_star, max(stars) as max\_star, avg(stars) as avg\_star

from business

where city = 'Beachwood'

Copy and Paste the Resulting Table Below (2 columns Ã¢â‚¬â€œ star rating and count):

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

| city | min\_star | max\_star | avg\_star |

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

| Beachwood | 2.0 | 5.0 | 3.96428571429 |

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

*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. As we can see in table below, sara review count are more than yuri but yuri has more fans.

select name, fans, review\_count

from user

order by review\_count desc

limit 20

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

| name | fans | review\_count |

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

| Gerald | 253 | 2000 |

| Sara | 50 | 1629 |

| Yuri | 76 | 1339 |

| .Hon | 101 | 1246 |

| William | 126 | 1215 |

| Harald | 311 | 1153 |

| eric | 16 | 1116 |

| Roanna | 104 | 1039 |

| Mimi | 497 | 968 |

| Christine | 173 | 930 |

| Ed | 38 | 904 |

| Nicole | 43 | 864 |

| Fran | 124 | 862 |

| Mark | 115 | 861 |

| Christina | 85 | 842 |

| Dominic | 37 | 836 |

| Lissa | 120 | 834 |

| Lisa | 159 | 813 |

| Alison | 61 | 775 |

| Sui | 78 | 754 |

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

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

Answer: love(454)

SQL code used to arrive at answer:

select count(text) as count

from tip

where text like "%love%"

union all

select count(text) as count

from tip

where text like "%hate%"

*10. Find the top 10 users with the most fans:*

SQL code used to arrive at answer:

select name, count(fans) as fan\_count

from user

group by name

order by fan\_count desc

limit 10

Copy and Paste the Result Below:

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

| name | fan\_count |

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

| John | 102 |

| David | 90 |

| Chris | 74 |

| Mike | 74 |

| Michael | 72 |

| Jennifer | 63 |

| Mark | 59 |

| Lisa | 58 |

| Melissa | 58 |

| Sarah | 55 |

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

**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. In "Cleveland", we can clearly observe different operation timings for different days of food stalls.

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

Yes. Number of reviews in "Cleveland" are more as compared to "Phoenix".

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

Food lovers are more active in "Phoenix", but food socialist are more active in "Cleveland".

SQL code used for analysis:

select biz.city, cat.category, rev.stars, hr.hours, biz.review\_count

from business as biz

inner join category as cat on cat.business\_id = biz.id

inner join review as rev on rev.business\_id = biz.id

inner join hours as hr on hr.business\_id = biz.id

where cat.category = 'Food'

Table:

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

| city | category | stars | hours | review\_count |

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

| Cleveland | Food | 5 | Sunday|10:00-16:00 | 723 |

| Cleveland | Food | 5 | Friday|7:00-18:00 | 723 |

| Cleveland | Food | 5 | Wednesday|7:00-16:00 | 723 |

| Cleveland | Food | 5 | Monday|7:00-16:00 | 723 |

| Cleveland | Food | 5 | Saturday|7:00-18:00 | 723 |

| Cleveland | Food | 4 | Sunday|10:00-16:00 | 723 |

| Cleveland | Food | 4 | Friday|7:00-18:00 | 723 |

| Cleveland | Food | 4 | Wednesday|7:00-16:00 | 723 |

| Cleveland | Food | 4 | Monday|7:00-16:00 | 723 |

| Cleveland | Food | 4 | Saturday|7:00-18:00 | 723 |

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

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

| city | category | stars | hours | review\_count |

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

| Phoenix | Food | 4 | Monday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Tuesday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Friday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Wednesday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Thursday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Sunday|11:00-22:00 | 431 |

| Phoenix | Food | 4 | Saturday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Monday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Tuesday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Friday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Wednesday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Thursday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Sunday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Saturday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Monday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Tuesday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Friday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Wednesday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Thursday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Sunday|11:00-22:00 | 431 |

| Phoenix | Food | 5 | Saturday|11:00-22:00 | 431 |

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

*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:

"Restaurants" business is most activ/profitable in most of the cities with count of 71, followed by "Shopping".

ii. Difference 2:

"Accessories" with other business like "Acupuncture" etc is the leat activ/profitable business with count of just 1 in cities.

SQL code used for analysis:

select distinct cat.category, count(biz.is\_open) as City\_Count

from category as cat

inner join business as biz on cat.business\_id = biz.id

group by cat.category

order by City\_Count desc

Table:

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

| category | City\_Count |

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

| Restaurants | 71 |

| Shopping | 30 |

| Food | 23 |

| Nightlife | 20 |

| Bars | 17 |

| Health & Medical | 17 |

| Home Services | 16 |

| Beauty & Spas | 13 |

| Local Services | 12 |

| American (Traditional) | 11 |

| Active Life | 10 |

| Automotive | 9 |

| Hotels & Travel | 9 |

| Burgers | 8 |

| Sandwiches | 8 |

| Arts & Entertainment | 7 |

| Fast Food | 7 |

| Mexican | 7 |

| American (New) | 6 |

| Event Planning & Services | 6 |

| Hair Salons | 6 |

| Bakeries | 5 |

| Doctors | 5 |

| Indian | 5 |

| Japanese | 5 |

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

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

| category | City\_Count |

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

| Accessories | 1 |

| Acupuncture | 1 |

| Arabian | 1 |

| Architects | 1 |

| Architectural Tours | 1 |

| Bagels | 1 |

| Banks & Credit Unions | 1 |

| Beaches | 1 |

| Beer Garden | 1 |

| Bike Repair/Maintenance | 1 |

| Bikes | 1 |

| Blow Dry/Out Services | 1 |

| Books | 1 |

| Breweries | 1 |

| Bridal | 1 |

| Cafes | 1 |

| Canadian (New) | 1 |

| Candy Stores | 1 |

| Cannabis Clinics | 1 |

| Car Rental | 1 |

| Car Stereo Installation | 1 |

| Cardiologists | 1 |

| Carpet Installation | 1 |

| Check Cashing/Pay-day Loans | 1 |

| Chicken Shop | 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:*

Which city and in what business, people are most happy with?

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

Data required for this will be business category, city, stars, review count, useful, cool, funny.

*iii. Output of your finished dataset:*

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

| city | category | avg\_stars | avg\_review\_count | avg\_useful | avg\_funny | avg\_cool |

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

| Cleveland | Ethnic Food | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Farmers Market | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Fruits & Veggies | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Market Stalls | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Meat Shops | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Public Markets | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Seafood Markets | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Shopping | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Specialty Food | 4.5 | 723.0 | 0.5 | 0.5 | 0.5 |

| Cleveland | Sandwiches | 4.5 | 361.0 | 0.0 | 0.0 | 0.0 |

| Pittsburgh | Desserts | 4.5 | 72.0 | 2.0 | 0.0 | 1.0 |

| Phoenix | Food | 4.25 | 468.5 | 0.666666666667 | 0.166666666667 | 0.333333333333 |

| Phoenix | American (Traditional) | 4.125 | 413.5 | 0.25 | 0.0 | 0.0 |

| Las Vegas | Asian Fusion | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Las Vegas | Chinese | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Las Vegas | Malaysian | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Las Vegas | Noodles | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Las Vegas | Soup | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Las Vegas | Taiwanese | 4.0 | 768.0 | 3.0 | 0.5 | 1.0 |

| Phoenix | Barbeque | 4.0 | 431.0 | 0.333333333333 | 0.0 | 0.0 |

| Phoenix | Bars | 4.0 | 431.0 | 0.333333333333 | 0.0 | 0.0 |

| Phoenix | Smokehouse | 4.0 | 431.0 | 0.333333333333 | 0.0 | 0.0 |

| Phoenix | Restaurants | 4.0 | 399.777777778 | 0.888888888889 | 0.111111111111 | 0.222222222222 |

| Phoenix | Breakfast & Brunch | 4.0 | 188.0 | 0.5 | 0.0 | 0.0 |

| Phoenix | Nightlife | 3.875 | 325.0 | 0.25 | 0.0 | 0.0 |

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

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

select biz.city, cat.category, avg(biz.stars) as avg\_stars,

avg(biz.review\_count) as avg\_review\_count, avg(rev.useful) as avg\_useful, avg(rev.funny) as avg\_funny, avg(rev.cool) as avg\_cool

from business as biz

inner join review as rev on rev.business\_id = biz.id

inner join category as cat on cat.business\_id = biz.id

group by cat.category

order by avg\_stars desc, avg\_review\_count desc