Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset Coursera 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.

For both parts of this assignment, use this "worksheet." It provides all the questions you are being asked, and your job will be to transfer your answers and SQL coding where indicated into this worksheet so that your peers can review your work. You should be able to use any Text Editor (Windows Notepad, Apple TextEdit, Notepad ++, Sublime Text, etc.) to copy and paste your answers. If you are going to use Word or some other page layout application, just be careful to make sure your answers and code are lined appropriately.
In this case, you may want to save as a PDF to ensure your formatting remains intact for you reviewer.

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 =

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.

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: max: avg:

ii. Table: Business, Column: Stars

min: max: avg:

iii. Table: Tip, Column: Likes

min: max: avg:

iv. Table: Checkin, Column: Count

min: max: avg:

v. Table: User, Column: Review_count

min: max: avg:

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

SQL code used to arrive at answer:

Copy and Paste the Result Below:

6. Find the distribution of star ratings to the business in the following cities:

i. Avon

SQL code used to arrive at answer:

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

ii. Beachwood

SQL code used to arrive at answer:

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

7. Find the top 3 users based on their total number of reviews:

SQL code used to arrive at answer:

Copy and Paste the Result Below:

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

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

Answer:

SQL code used to arrive at answer:

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

SQL code used to arrive at answer:

Copy and Paste the Result Below:

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?

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

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

SQL code used for analysis:

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:

ii. Difference 2:

SQL code used for analysis:

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:

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

iii. Output of your finished dataset:

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