DTSC660: Data and Database Management with SQL
Module 8
Assignment 6

Purpose

For this assignment, you will be learning to interpret and clean data. It is important to note that the most time consuming and difficult part of this assignment is digging through and interpreting the data. The majority of your time should be spent searching through the data for inconsistencies, errors, and misrepresentations. This can be done in pgAdmin or in Excel/Sheets if you are more comfortable with that environment. The queries and data modifications will be much simpler in this assignment than in the other assignments. This is expected, as the focus is more on your ability to consider the data thoroughly and then perform tasks to make the data more useful. To get started choose one of the data sets in the Assignment 6 folder (*Air BnB, Data Scientist Salaries, Netflix, or food choices)* and begin exploring the data.

Once you have determined which data set you are interested in working with, you will need to wrangle (clean) the data to make it useful for analysis.

Submission

You will submit a total of **1** sql files to CodeGrade. Files should be named appropriately and be in .sql format. Each file must use the postgres standards taught in the course. Use of other SQL languages such as T-SQL will result in an automatic 0 for the assignment. **Ensure your file runs in its entirety in pgAdmin.** You will have **one submission attempt** for this assignment.

- ***File 1***: You must submit a SQL document called <LastName>_Assignment6. This document must include ALL queries requested in the instructions below.

- You will submit the file using the Assignment 6 Submission link to CodeGrade.

Instructions

As described in the videos and textbook, you will be cleaning a dataset to make it useful for analysis. Complete the steps below carefully and ensure that you save the assignment as a SQL file as indicated in the submission instructions. There is no template for this assignment. Also, make sure you review the rubric to ensure that you have met all the requirements for the assignment. It is expected that if you have questions or difficulties with any portion of this assignment that you utilize the assignment discussion board or email the GAs to gain clarity (dtsc_ga_660@eastern.edu).

Make sure to complete the initial steps below to prepare the data file for importing:

- Select a dataset of your choice from the list in the “Assignment 6 CSV Files” subfolder in Brightspace and download the file.
- Place this file in a public folder on your computer
- Take note of the path to this file (copy the path)

PART 1 Creating the Table and Importing the Data

1. Create the table reusing the column titles from the csv, *do not change these*
    a. When selecting data types for your tables, ensure they accommodate the full range of values and avoid truncating any values. Ensure that decimals, lengths, and other specifications are sufficient for all data to be imported completely.
2. Write the copy statement to bring the data into the database
    a. Remember that if you choose an incompatible data type, you can enter the *DROP TABLE* command to remove the table and restart.
3. Run a basic select statement that verifies the data is present and matches what is in the csv file.

Make sure to include the table creation and copy statement in your code as this is a necessary component for the grader to grade your assignment. Assignments without this portion may result in a 0.

To complete this next part, create a SQL file using the naming convention:
<LastName>_Assignment6.

There is no template for this assignment. Please indicate which question you are answering by using commented out text so the grader can easily follow your work. If you need guidance on how to comment in SQL, see this link. Once you are done, submit the document to the Assignment 6 Submission CodeGrade link.

Before you start, ensure to include your name, the chosen data set, and the reason for choosing that data set in a multi-line comment at the top of your submission.

For each part below, you will be required to include the code used to clean your data as well as a rationale included in the comments section for each part. Comment rationales should be 2-3 sentences explaining the purpose of the modification and your logic behind your choice. Failure to include comments will result in loss of points for that part of the assignment (see grading rubric).

Part 2 (Cleaning)

1. Create a backup of your imported table (no comments required)

2. Create a duplicate column in the table (no comments required)

3. Locate and update values representing missing data in one column and perform **ONE** of the following modifications:

    a. Change values so that they are correctly labeled and recognized by SQL as NULL values
    b. Change their values to another value that accurately represents or reflects the data (such as substituting the mean of the column for the value)
    c. Remove the data containing null values

4. Perform step 3 using a second method (e.g. a, b or c from above) on a different column

5. Group similar values (i.e. - Sr., Senior, Sr from the video), misspelled, or inconsistent data for one column such that the data is correct and consistent. Only one group of similar values need to be cleaned, not the entire column.

6. Repeat step 5 on another column.

7. Choose an additional method of data cleaning, not previously utilized in this assignment, that was demonstrated either in the textbook or in the videos. Your selected method should be done with the purpose to make the data more useful. If you're unsure whether your chosen method satisfies these criteria, please request feedback **PRIOR TO SUBMITTING** the assignment.

*******************************GRADING RUBRIC ON NEXT PAGE*********************************

This assignment will be graded on the following rubric. Please see *DTSC 660 Assignment Grading* in **Module 0** for our course's guidelines on point allocation and deduction. Incorrect syntax, extraneous results, or incorrectly addressing all question requirements will result in loss of points. Graders will NOT attempt to correct malformed sql code. :

STEP	**Step Points**	**Comment Points**
1	5	N/A
2	5	N/A
3	10	5
4	10	5
5	10	5
6	10	5
7	20	10
Total (100)	70	30