DTSC660: Data and Database Management with SQL
Module 8
Assignment 7

Purpose

This assignment allows you as the student to demonstrate your skills and exercise some choice in the assignment. For this assignment, you will be cleaning a data set from one of these options from Kaggle, which are available in the Assignment 7 folder: *Air BnB, Data Scientist Salaries, or Netflix.*

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 in 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>_Assignment7. This document must include ALL queries requested in the instructions below.

- You will submit the file using the Assignment 7 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).

PART 1 Creating the Table and Importing the Data

*****PLEASE MAKE SURE THAT YOU INCLUDE THE TABLE CREATION AND COPY STATEMENTS IN YOUR FILE*****

1. Select a dataset of your choice from the list in the "Assignment 7 CSV Files" subfolder in Brightspace and download the file.
2. Place this file in a public folder on your computer
3. Take note of the path to this file (copy the path)
4. Utilize the COPY command you learned in Module 2 (revisit this module if necessary) to import the data.
5. Run a basic select statement that verifies the data is present and matches what is in the csv file.

In your sql file, please include the CREATE TABLE and COPY commands you used to import the data. While you will not be graded on this part, it's important that we are able to set up the same data structure you used so we can run and test your queries. **If you omit these items, you will receive a zero for the assignment, as we will not be able to assess your work.**

Part 2 (Cleaning)

This is the part you will be graded on. To complete this part, create a SQL file using the naming convention: <LastName>_Assignment7. There is no template for this assignment. Please make sure that it is clear using commented out text, each question that you are answering, so the grader can easily follow your work. Once you are done, submit the document to the Assignment 6 CodeGrade link.

Before starting, in the top comments, complete your name, the data set you chose, and why you chose that data set.

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. Failure to include comments will result in loss of points for that part of the assignment (see grading rubric).

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 different 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. Pick one other method of cleaning demonstrated in the book or videos that you feel will make the data more useful. If you are unsure whether your method meets the criteria, please request feedback **PRIOR TO SUBMITTING.**

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

This assignment will be graded on the following rubric. Incorrect syntax, extraneous results, errors, 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