Instructions

As a hands on Data Science Assignment, you will be working on a real world dataset provided by the Chicago Data Portal. You will be asked questions that will help you understand the data just like a data scientist would. You will be assessed both on the correctness of your SQL queries and results.

A Jupyter notebook has been provided to help with completing this assignment. Follow the instructions to complete all the problems. Then share the Queries and Results with your peers for reviewing.

Review criteria

The total points possible for this assignment are **25**. Here is the breakdown:

- Problem 1: 1 Points
- Problem 2: 1 Points
- Problem 3: 2 Points
- Problem 4: 2 Points
- Problem 5: 2 Points
- Problem 6: 2 Points
- Problem 7: 2 Points
- Problem 8: 4 Points
- Problem 9: 4 Points
- Problem 10: 5 Points

For each problem points will be awarded as follows:

- Full points: Used a correct SQL query that yielded a correct result
- Half or partial points: The query and results are not fully correct
- No points: Did not attempt the problem or did not upload any solution

Step-By-Step Assignment Instructions

Assignment Topic:

In this assignment, you will download the datasets provided, load them into a database, write and execute SQL queries to answer the problems provided, and upload a screenshot showing the correct SQL query and result for review by your peers. A Jupyter notebook is provided in the preceding lesson to help you with the process.

This assignment involves 3 datasets for the city of Chicago obtained from the Chicago Data Portal:

1. Chicago Socioeconomic Indicators

This dataset contains a selection of six socioeconomic indicators of public health significance and a *hardship index*, by Chicago community area, for the years 2008 – 2012.

2. Chicago Public Schools

This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year.

3. Chicago Crime Data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

Instructions:

1. Review the datasets

Before you begin, you will need to become familiar with the datasets. Snapshots for the three datasets in .CSV format can be downloaded from the following links:

- Chicago Socioeconomic Indicators: Click here
- Chicago Public Schools: Click here
- Chicago Crime Data: Click here

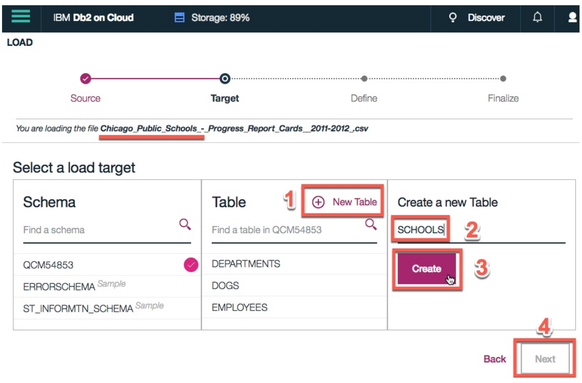
NOTE: Ensure you have downloaded the datasets using the links above instead of directly from the Chicago Data Portal. The versions linked here are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment. The CSV file provided above for the Chicago Crime Data is a very small subset of the full dataset available from the Chicago Data Portal. The original dataset is over 1.55GB in size and contains over 6.5 million rows. For the purposes of this assignment you will use a much smaller sample with only about 500 rows.

2. Load the datasets into a database

Perform this step using the LOAD tool in the Db2 console. You will need to create 3 tables in the database, one for each dataset, named as follows, and then load the respective .CSV file into the table:

1. CENSUS_DATA
2. CHICAGO_PUBLIC_SCHOOLS
3. CHICAGO_CRIME_DATA

To load the data into the tables the steps are similar to Week 2 Lab 1 Part II. The only difference with that lab is that in Step 5 of the instructions you will need to click on create *(+) New Table* and specify the name of the table you want to create and then click *Next*.



3. Write and execute queries

Perform this step in the Jupyter notebook provided in the previous section. Carefully read and understand each problem. Compose and execute the appropriate SQL queries to answer each of the problems. Take a screenshot of each query and its results and save it as a jpg file..

Problem 1: Find the total number of crimes recorded in the crime table.

Problem 2: Retrieve first 10 rows from the CRIME table.

Problem 3: How many crimes involve an arrest.

Problem 4: Which unique types of crimes (e.g. THEFT) have been recorded at a GAS STATION locations?

Problem 5: In the CENUS_DATA table list all community areas whose names start with the letter 'B'.

Problem 6: List the schools in community areas 10 to 15 that are healthy school certified.

Problem 7: What is the average school Safety Score?

Problem 8: Find the top 5 Community Areas by average College Enrollment [number of students].

Problem 9: Use a sub-query todeterminewhich Community Area has the least value for school Safety Score?

Problem 10: [Without using an explicit JOIN operator] Find the Per Capita Income of the Community Area which has a school Safety Score of 1.

How to submit:

A screenshot in JPEG format is required to be submitted for solution to each of the problems. The screenshot for each problem should clearly show the SQL query and results for the query. The screenshots will be uploaded in the following sections.

Example Submissions

Here is an example of a submission clearly showing both the SQL Query and its Results/output, when executed from a Jupyter notebook.

Important aspects of the submission are highlighted below for illustrative purposes:

Author(s)

Rav Ahuja

Change log

Date	Version	Changed by	Change Description
2020-09-05	2.0	Malika Singla	Migrated to GitLab