**Assignment #3: Working with data in SQLite**

**The goals in this assignment are as follows:**

* Select a new dataset in either a .TXT or a .CSV file format (text file that is comma-delimited,tab-delimited or other standard plain text format) or you may use the "clean" version of the dataset you worked on for assignment #2.
* Your dataset should have a minimum of 5,000 records. If you are using a smaller dataset which you used for assignment #2 however, that is fine. Many students in the class used large datasets for assignment #2 and extracted a selection of the records for assignment #2; in those cases, feel free to use the entire file for this assignment as we discussed in class. If you wish to use a dataset smaller than 5,000 records which you did not use for assignment #2, please see your instructor.
* Write a .SQL script to create a table for your data. Use .import to load the data into the SQLite table using your SQLite database that you created or use FILE / IMPORT in DB-Browser. (Note: If you write a Python script to run INSERT queries, be sure to submit the Python script.)
* Write 12 queries to analyse your data as noted below.

**Once you have successfully imported the data into SQLite, write the following 12 queries:**

1. Write a query to count how many records there are in your table.
2. Write a query to list the first 10 records in your table based on a sort in a reverse order.
3. Display three important fields; at least one should be in alphabetical order; show only the first 10 records of an appropriate selection using WHERE.
4. Use GROUP BY to find averages on a numerical field in a reasonable breakdown. Note: If you only have textual data columns, use the COUNT function instead of averages.
5. Write a query that includes a HAVING clause.
6. Create a user-friendly listing of the first 15 records in alphabetical order by creating a text field using concatenation, e.g.:

Tom Sawyer (Twain, Mark; written in 1862)

1. and 8. Write two queries using at least two different SQLite functions listed here: <https://www.sqlite.org/lang_corefunc.html> that we have not used in class.

9 – 12. Write four more queries that are relevant to your data and and yield results that you find interesting. For each of these queries, write the question that you are posing in English as a comment and then write and execute the query in SQL. For example:

-- What is the value of the books in our

-- collection for each publisher

-- for all books that were written after 1850?

sqlite> SELECT edition,SUM(price)

...> FROM books

...> WHERE year\_written > 1850

...> GROUP BY edition

...> ORDER BY edition;

**Write up your results as follows:**

Write a brief report containing the following information using the text-editor of your choice and submit the report as a .PDF file or plain text file:

* the URL where you obtained the data
* one or several sentences describing the nature of the data that you downloaded (e.g. *These are records listing the name, address and contact information for wi-fi hotspots across New York City.)*
* Write several paragraphs (or approximately one page double-spaced) describing what you have learned from running your queries. Please answer the following questions:
  + What are your conclusions? (For example, *"There are 3 times as many wi-fi hotspots in the wealthiest neightborhoods in Manhattan than in the poorest neighborhoods based on my study. This could have an impact on ...")*;
  + Now that you have queried that table, do you believe that the dataset that you worked with is reliable? Why or why not?
  + If you were to do further research with these data, what would be your approach? Would you need additional data?
  + Do the results of your analysis reflect what you anticipated you would find?

**What to submit:**

Go to NYU/Classes and submit the following:

**Note** - Be sure to compress all of your files into one .zip file for submission and include both the assignment number and your netid in every filename. Here are two examples; please follow this model: *asg3\_de123.zip* for the .zip file and *asg3\_de123\_script.sql* as a sample SQL script file.

1. Submit your report about the data (including the URL and comments as noted above)
2. Submit your SQLite script (.sql file) that you used to create the table.
   * Note: If you worked in DB-Browser, you may use FILE / EXPORT / DATABASE TO SQL FILE to generate an SQL script to create the table for submission.
   * Under "Export Everything", select EXPORT SCHEMA ONLY to obtain a script that creates the table but does not recreate the dataset.
3. Submit your SQLite script (.sql file) that you used to run the queries and that contains all of the twelve queries along with your "natural language" (English) description for #9-12.
4. You may *also* submit a copy of your Terminal session as a text file with your queries if you wish but you are not required to do so. The grader and I will re-create your database and run all of the queries.
5. Submit your data file that you are working with and imported into SQLite. If your data file is very large, you may submit the URL along with a sample of the data (e.g. 5,000 records) and the python program you wrote to select data for this assignment (if you did so).
6. If you wish to write and submit your work in ONE .sql script that includes your CREATE TABLE, .import and SELECT queries, you are welcome to do so.

**Resources:**

* [Data sources to consider](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_resources_python_dataInTheWild_fa19.html)(but you are not limited to these)
* [Resources on SQLite](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_resources_sqlite_fa19.html)
* [SQLite Class Notes](https://cs.nyu.edu/courses/fall19/CSCI-UA.0060-001/notes/dbw_readings_sqlite_fa19.html)