# CS/CE 457/464 - Homework Assignment 4: SQL

Due Date: Monday, September 23 at 11:59 pm

**Purpose**:
Demonstrate exploration of data via creation of statistical tables using RDBMS/SQL; connecting Python with database and perform exploratory data analysis.

**Tools:**

- PostgreSQL, Oracle, MySQL, etc. (your choice)
    - PostgreSQL: https://www.postgresql.org/download/

**Part 1** (70 points):
**Deliverables**: You can either include screenshots of your pgadmin screen showing query and output table or copy paste your queries and output table for your answers and submit PDF version.

- Create a SQL database tables for both datasets `countries.csv` and `cities.csv` separately, using a RDBMS (PostgreSQL preferred). You need to submit a create table query in the final document.
- Load/Import the dataset into the table.
- Query the database tables and interpret the results, displaying:
    1. the count of total number of records in each table.
    2. the count of number of cities for each country in descending order of count (use group by)
    3. the count of regions and sub-regions in each country. Sort them by ascending order of country name. (use group by)
    4. Top 10 most populous capital cities. Display country, city and population in descending order.
    5. Average city population of capital and non-capital cities. (use group by)
    6. Average country birth rate for each region and sub-region (use group by)

**Part 2** (30 points):

- Connect to your `countries` and `cities` database tables in Python and load into pandas dataframe
    - Perform **three** interesting analyses on this data with visualization and tell the story about interesting insights in your analysis
        - Analysis could be anything such as univariate analysis, bivariate analysis, correlation etc.

**Deliverables**: Submit your pdf file for Part 1 and Jupyter Notebook ipynb file for part 2 with all the code and analysis.