Final Project – Creative Data Analysis

Overview

The final assignment for this course is to collaboratively analyze publicly available data sets and to present some interesting plots of data that answer a specific research question. This project with be done in groups of two students (three if necessary). Individually, each student will develop software similar to the assignments done over the last few weeks for a specific data set with command-line parsing using **argparse**, logging using the **logging** module, unit testing of functions with the **unittest** module, graphical output with **matplotlib**, and data output using the **csv** module. **NumPy** data arrays must be used for numerical data processing.

As a team the students must develop analyses that draw from all the data sets identified by the team members, and that merge or combine data from multiple sources to produce the answer to an interesting research question.

Data Sources

Any publicly available data source can be used for this assignment. Each student should work with a separate data set (obviously, the data sets need to be related somehow so combined analysis can be performed).

Good places to look for data sets are:

- Data.gov
- UCI Machine Learning Repository
- Most U.S. states have their own public data repositories
- There are good sports data sets available

Whatever data sets you identify, please make sure they are truly public. Please do not scrape websites for data that are not explicitly made available.

Research Question

Each team should start with an interesting research question that it wants to answer. There are very few limitations on what this can be, but it should combine the interests of all the team members. Some interesting fields to explore: sports, science, state and local government, federal government, recreation, social media (basically anything). Most questions will be something like **"What is the relationship between X and Y?"** (e.g., "What is the relationship between voting in federal elections and the weather on voting day?"). The question should take aim at several different concepts that are explored or described by different data sets.

Analyses

Each student must perform the necessary programming to be able to produce three different plots (and CSV data) of the data set they are working with. These plots should be related to the overall research question and should show interesting trends or relationships within a data set.

Together, the team must also produce three different plots that combine the data sets and that shed light on the research question of interest.

Upload

Each student must turn in the individual processing and analysis they perform on their data set. This should include a corresponding unit-test module.

The team will also turn in the higher level module that performs the combined analysis by using the functionality available for each module.

Please combine all your files into a single ZIP file.