CSCE100 Introduction to Informatics
Fall 2021

Final Project Assignment (Group): Investigating Social Issues

Points: 250 points. Assignment Date: October 21, 2021
Proposal Due Date: November 2, 2021 11:00 a.m.
Project Status Update (Presentation): November 9, 2021 (in class)
Project Handin Due Date: December 6 (Monday), 2021 11:00 a.m.
Project Demos Date: December 7, 9, 2021 (in class)

Objectives

1. To define and describe a problem in data science and informatics focused on the several datasets relevant to investigating social issues
2. To design and develop your solution to address the problem
3. To carry out the various steps of Informatics to solve the problem: data preparation, data cleaning, data pre-processing, data analysis, and data visualization
4. To implement programs in Python to perform the above steps of informatics on the data downloaded
5. To write a report to document your problem and solution, your informatics steps, outcomes and results of your informatics work, insights and lessons learned, and potential future work
6. To present your final project well using slides and a software demonstration
7. To learn about teamwork and work as a team

Datasets

There are multiple repositories with datasets relevant to social issues (including health, economy, literacy, employment, and so forth), available for download:

- https://www.racialhealthequity.org/data
- http://diversitydata.org
- https://www.icpsr.umich.edu/web/pages/
- https://nij.ojp.gov/library/datasets-nij-funded-research (searchable)
- https://www.data.gov (searchable)
- https://data.worldbank.org (searchable)
- http://uis.unesco.org (searchable) (by country)
- many more

Note also that you are encouraged to combine the above data with the Covid-19 data from the Johns Hopkins University's GitHub repository (https://github.com/CSSEGISandData/COVID-19). For example, there are country-by-country data in the above links and also in the data from Johns Hopkins University. Correlation analysis can be run to find whether socio-economic factors are correlated with the death rates, for example.

Final Project Problem

For this final project, identify a specific problem in data science or informatics that your team will investigate given the datasets. For example, are there interesting questions to answer? Are there patterns that you want to find? Are there trends to be observed? Explore the datasets to get yourself familiarized with the data first, and then form your questions that will drive your Informatics efforts. ***Note that if your team would like to use dataset(s) with a focus or theme other than social issues, please let me know as soon as possible so I will have a chance to review your datasets.***

Requirements and Handin

1. The submission deadline for all handins is December 6 (Monday), 2021, 11:00 AM, which is one day before our first Demo Day. **Late handins will *not* be accepted or graded.**
2. **Final Project Proposal Q&A** (0 points but REQUIRED):
   a. **November 2, 2021:** Each team is required to turn in a 3-page Final Project Proposal
   b. **The Proposal must include at least the following:** (1) a Team Name, (2) The names of all team members, (3) Proposed informatics problem, (4) Proposed tasks to solve the informatics problem, and (5) Specific tasks and responsibilities for each team member.
3. **Final Project Updates (0 points but REQUIRED)**
   a. **November 9, 2021:** Each team is required present their proposal and current project status (using Powerpoint slides) in class and participate in a Q&A discussion.
4. **Final Project Report** (100 points): Each team is required to handin a final project report. The following sections are required:
   a. **Introduction** (15 points): This section describes your informatics problem
   b. **Data Preparation** (15 points): This section describes your data exploration strategy and how you extract with the data that you used for your final project.
   c. **Data Cleaning & Pre-Processing** (10 points): This section describes your data cleaning and pre-processing strategies used. Must refer to any Python programs that you used to accomplish this step.
   d. **Data Analysis** (20 points): This section describes your data analysis solution. Must justify your analysis and findings. Must refer to any Python programs that you used to accomplish this step.
   e. **Data Visualization** (20 points): This section describes your data visualization strategies. Must justify your strategies. Must refer to any Python program that you used to accomplish this step. Must also provide the infographic.
   f. **Conclusions** (10 points): This section documents any insights or lessons learned from this Informatics assignment. This section also summarizes any findings from the Data Analysis section.
   g. **Appendix** (10 points):
      i. This section describes your overall approach to implement the Python programs and the list of all Python programs that you implemented and their purpose.

ii. This section must also include the 3-page proposal.

5. **Data and Programs** (50 points): For each team:
   a. You are required to handin a screen capture of your "testing session" using your programs. (10 points)
   b. You are required to handin all program files. (10 points)
   c. You are required to handin all input and output files. (5 points)
6. **Final Project Demo** (50 points): For each team:
   a. You are required to present your Final Project using Powerpoint slides and execution of your programs at real-time on your Demo Day (either **December 7** or **December 9, 2021**)
   b. Your presentation should cover all sections of your Final Project report adequately (35 points)
   c. Your demonstration of your programs should proceed smoothly showing how you obtain the output from running your programs (15 points)
7. You are required to handin online the above files to Canvas under Final Project.