Final Project Description
Working with Big Data - VSP 2024

Project Goals

This class will culminate in a group project and presentation. This project will give you experience

- Asking interesting research questions
- Finding the best data to answer those questions
- Contextualizing research using summary statistics and visualizations
- Analyzing data with the purpose of learning something specific
- Working in a group
- Presenting and writing in English

Your project should be one big narrative (or many small ones) – it should not just be a collection of code and figures. That is, your Jupyter Notebook must include exposition that ties all of the data, figures, and analyses together into a focused and clear report on your chosen topic(s).

Project Details

For this project, you will work in groups of **two people**. The specific research question for the project is open-ended. The only requisite is that is has to involve working with data. You need to find the data that is useful to answer your question, clean it, and prepare a summary of your results.

The first option for the project is to work with one of the following datasets:

1. Political Violence and Conflict ACLED
    - This is a dataset that records reports of political violence and social unrest for all countries around the world. Importantly, it contains geo-spatial information on the location of several events.
2. World Bank Data

- This repository contains a large set of variables at the country level. It cover a wide range of outcomes: literacy, access to public services, financial indicators.

Your group has the freedom to explore these datasets in any way you think it is convenient. Alternatives can be to add other variables from other sources to expand the information. Or to dig deeper in one of these datasets.

Pick your own data

As an alternative, you are free to work on other datasets. Here the range is big. In general, most datasets will be fine for the project, but if your group decides to work on a separate dataset, talk with me to get approval.

Logistical Details

Project Deliverables & Dates

Each group will submit:

- A one page project proposal that details their topics/questions, data sources, and planned approach (due the beginning of class on July 19th)
- An executed Jupyter notebook that contains their entire report including code, figures, and exposition (due before class on August 2nd)
- A README.md file detailing each group member's name and their contribution to the project
- A GitHub Repository containing the executed Jupyter Notebook, data files (that obey GitHub's size limits), a README.md file, and any other relevant files (due before class on August 2nd)

Presentation

Groups will also prepare a 12-15 minute presentation on their project, which they will give on August 2nd. Groups should use their submitted Jupyter Notebook as a visual aid. Do not, however, simply read off of the Jupyter notebook. Even when presenting the exact same information, a presentation should use different words than a report. **Each group member must present for at least 5 minutes.**

These are the only deadlines, but I highly encourage students to communicate with me about the state of their project if they need guidance or an opinion. I will be actively engaging with students during the labs to see how their projects are coming along and answering any questions. Students can also see me after class or email me to schedule appointments.

Grading

The Jupyter Notebook is worth 30% of your final grade, the presentation is worth 15%, and the proposal is worth 5%. Below, I describe the hallmarks of a good project and what I look for below.

Project Grading

A successful project will

- Have an introduction, data description, descriptive statistics, visualizations, analysis, and conclusion
- Have clear, well-defined goals and accomplish them
- Integrate code and figures with text to create a cohesive report
- Reflect effort (many analyses, difficult analyses, custom data sets, well-researched background, well-written, strong aesthetics, etc.)
- will include a data wrangling component
- will include a map
- will include a regression or classification component

Presentation Grading

A successful presentation

- Demonstrates effort and preparation
- Gives the audience a good understanding of what lies in the report
- Uses the Jupyter Notebook effectively as a visual aid but not as a crutch
- involves all group members participating equally
- engages the audience

Proposal Grading

Once I accept a proposal as your project, you get full marks, but I will only accept the proposal if I can understand what your project entails from reading it. If I cannot understand or I think you should pursue another topic, you may have to meet with me. Once we have met and I understand your project, you will still get the full 5%.

English

English grammar and spelling *will never be graded*. We are focused on coding and data analysis in this class. That being said, I do need to understand what you write and what you say, so if your English is not as strong, it might help to have someone in your group who is more proficient in English. I also do not mind if you use chatGPT to edit your writing. Please review what it writes as it can make up things that sound good.

Your Obligation to Your Group

As previously mentioned, it will be difficult to do a good job on this project if you only work in class. Whether you split up the work or try and work on everything together, you will have to communicate and meet with your group outside of class. Please make yourself available to your group members because they will be counting on you to do your share of the work. Do not ghost your group or just work on the project during class while they handle the rest. If you fail to communicate with your group or contribute to your project, you can and will receive a worse grade on the project.