Course Project Description

Environmental Data Science, Fall 2020

Course Project:

For an environmental topic of your choosing, you will generate a question or objective to be addressed through the analysis of publicly available data. As a part of your analysis, you will find and read scientific literature related to better understand and contextualize your data analysis. You will present this analysis in class and make a digital poster that includes visualizations, the method of analysis, and interpretations the data. The question or objective you would like to address should be able to be answered with existing data. Think about how you would need to answer this question. For example, are you comparing means? Are you looking at patterns throughout time?

This project is an opportunity for you to advance your data science skillsets in a more specific area of your interest. I will provide XX datasets that can be a starting point for a project. You are welcome to choose one of these datasets or different data that you find online. You may also conduct an analysis on data that you may have collected for another course or a research project. However, your project must be sufficiently different and new that you are not applying it for credit or wages outside of this course.

Choose a topic that you find exciting and think you would benefit from understanding more deeply. Also think about choosing a data analysis that enhances tools and skillsets you are interested in learning more closely. For example, if you'd like to learn more about spatial data, machine learning, or linear regression use this as an opportunity to do so!

Project Assessment

Project Components:

1. R script for data analysis
2. Poster that contains
    a. Background description
    b. Study objective, hypothesis, or questions
    c. Description of methods and data
    d. One statistical test/technique
    e. Summary of results with at least 3 graphical figures or tables
    f. Summary of main conclusions

Your work will be graded based on the four fundamentals of data analysis outlined in class. Descriptions and assessment of each fundamental are provided below.

The assignment will be worth a total of 70 points.

Rubrics are provided for evaluation below:

1. Reproducible: Using the same data, your script can be run to produce the same results that you present. This requires careful management of data and organized code to analyze and visualize data. I expect to be able to map to your folder in the server and run your code.

Component	A	B	C	D	F
Code reproduces results (5 points)	All results can be replicated	Most results can be replicated	Limited results can be replicated	Few results run.	Not attempted.
Ease of reproducibility (5 points)	Version control utilized. Workflow is set up to readily replicate results.	Version control utilized. Handling of data directories complicates replication.	Version control utilized. Instructor is unable to replicate results in a reasonable manner.	Version control not utilized. Instructor is unable to replicate results in a reasonable manner.	Not attempted.

2. Tractable: Your code should be easy to understand, use, or change. You should use comments throughout the majority of your code, and it should be organized. This means lines of code should be run in order from start to finish, you should delete any unnecessary code. You should also consider your workflow. For the purpose of activities, you should not need multiple scripts or to export a lot of data files.

Component	A	B	C	D	F
Documentation (5 points)	Code is well commented with insightful direct comments.	Most of code is well commented with direct comments.	Comments are limited and unclear with direct comments.	Few comments.	Not attempted
Organization (5 points)	Code is accurately ordered for analysis. No unnecessary code is included. Workflow prevents confusion.	Code is mostly accurately ordered for analysis. Limited unnecessary code is included. Workflow minimizes confusion.	Code mostly lacks order for analysis. Unnecessary code is included. Workflow causes confusion.	Code is often erroneous, unnecessary, and lacks adequate workflow.	Not attempted

3. Products: The outcomes of your data analysis should be high quality and thoroughly documented. Visualizations should be well labelled, readable, and appropriately display the data. Statistics should be fully reported. Data summaries or descriptions should be clear and accurate.

Component	A	B	C	D	F
Outcomes: (statistical analysis, figures, 20 points total)	Products of data analysis fully address assignment prompts. Products demonstrate a thorough understanding of the prompts and comprehension in course concepts. Figures represent data in a compelling and accurate manner.	Products of data analysis mostly address assignment prompts. Products mostly demonstrate an understanding and comprehension in course concepts. Figures represent data, but may need some further clarification or improved labelling.	Products of data analysis do not fully address assignment prompts. Products demonstrate more comprehension in course concepts is necessary. Figures do not adequately convey data, and need clarification and better labelling.	Products of data analysis do not address assignment prompts.	Not attempted
Reporting (results summary, conclusions: 10 points total)	All outcomes are organized, well labelled, and appropriately described.	Most outcomes are organized, well labelled, and appropriately described although clarification or improvement may be needed.	Outcomes need organization, more labelling, and do not adequately convey understanding of data or analysis in descriptions.	Outcomes are difficult to discern and need organization, more labelling, and lacking description.	Not attempted

4. Interpretation: Clear and informative communication of data analysis that addresses the initial prompt of the analysis. Interpretation should be direct but careful to not overstep limitations of data and statistics. Delivery for target audience should be considered (*e.g.* general public or experts in the field). Use of visualizations can often enhance interpretation.

Component	A	B	C	D	F
Study question, objective or hypothesis (5 points)	The goal of the study is accurately and clearly conveyed in a manner that can be fully addressed by the analysis. The focus is novel and compelling.	The goal of the study is mostly accurate, but could use refining to fully convey the goal of the analysis. The focus is novel.	The goal of the study is not very clear and needs improvement to convey the direction of the analysis.	The goal is largely lacking and does not establish the subsequent analysis well.	Not attempted.
Comprehension of outcomes (conclusions: 10 points)	Description of analyses and data conveys comprehension of course concepts. Interpretation is carefully worded to provide specific outcomes, address nuances, and synthesizes information.	Description of analyses and data mostly conveys comprehension of course concepts. Interpretation is carefully worded to provide specific outcomes, and synthesizes information.	Description of analyses and data demonstrates limited comprehension of course concepts. Interpretation does not provide specific outcomes or synthesize information.	Description of analyses and data does not adequately address comprehension of course concepts.	Not attempted
Delivery (5 points)	Poster is well written, easy to interpret, compelling, and insightful.	In general, the poster is well written, easy to interpret, compelling, and insightful. However, some of these components may not be fully met and need further improvement.	The poster mostly reads well, but some errors or issues with interpreting content.	The poster is difficult to read, disorganized, and some features may not be viewed properly.	Not attempted

Project data:
You may choose to find your own data for this project or you may choose to work with one of the datasets provided. You may also choose to use a mix of provided data and pull in an additional dataset from the internet.

Provided datasets:

New York State Weather: Data from weather stations across NY state. I have organized the data such that each station and observation type (*e.g.* precipitation, maximum temperature) is stored within its own csv. There are 2,653 weather station data files. I recommend choosing a subset of weather stations to focus on. This is also a good dataset to choose if you are looking to improve your coding skills with larger, complex data.

Arctic Sea Ice Extent: Extent of the Arctic sea ice downloaded from the National Snow and Ice Center from 1979- 2019. This is a spatial dataset.

California Wildfires: This dataset comes from CALFIRE. It is a csv that contains all wildfire information for recent years in California. The table has statistics related to wildfire size, location, and timing. More information can be found here: https://www.fire.ca.gov/incidents. Depending on your approach, you may want to examine multiple datasets that may be of interest for the wildfires. I've also included a daily air quality (pm2.5) dataset for 2020 as an example of potential complementary datasets (downloaded from: https://www.epa.gov/outdoor-air-quality-data/download-daily-data)

Keep in mind there is an abundance of datasets at the following resources. Keep in mind that bringing in additional data will occupy more of your analysis with data wrangling and management. The following websites have excellent datasets that you may find useful:

Remote sensing data
https://earthexplorer.usgs.gov/

Drought
https://droughtmonitor.unl.edu/

Weather
https://www.ncdc.noaa.gov/cdo-web/datatools/findstation

Climate Data
https://prism.oregonstate.edu/

Air Quality Data
https://www.epa.gov/outdoor-air-quality-data/download-daily-data