COSC 481 – Intro to Data Science
Major Project

Purpose: To complete a full data science project that benefits society

Specifications:

This major project in data science is about two things: executing a full data science study, similar to ones we have seen throughout the course; and determining where a data science approach can be beneficial to a societal need or movement.

Needless to say, this second part is somewhat vague. What I am looking for is some question that can be answered through the power of data that can help a community of people better themselves. This can be as broad or as narrow as you'd like. I'd recommend taking a look at the various studies posted by the Data Science for Social Good group at the University of Chicago (https://dssg.uchicago.edu/) for general ideas of what I'm looking for. From there your team should look at the various data sets available to you at the link posted on the course website and identify a broad category (i.e. weather, politics, etc.) that you'd like to explore. You may also attempt a current competition available through Kaggle's Data Science for Good (https://www.kaggle.com/competitions). NOTE: there are a lot of Kaggle competitions, only the Data Science for Good ones are automatically eligible.

Question Proposal: Your team will need to narrow down a central question you plan on exploring very early on in the process. Once you have established your question, you should prepare a one-page paper detailing your question and its relevance to a societal need. You should also include what data set you'll be using and any citations you need to make your team's case. Note: citations are not necessary, but your team's argument for societal need may be rejected due to lack of evidence and your team will have to redo the paper (with a grade penalty).

Data: Once your team's question has been approved, you should start to clean up your data and do your data analysis. There is no specific turnin for this (other than notification of the data set selected via email) but you should include information about how you formatted your data and your data analysis in the final paper (see below). Once you have clean data, you should create whatever models you need to answer the question you have chosen in Python, using whatever third-party modules you need. There is no restriction beyond the language for your programs.

Final Paper: In addition to your team's code, there will be a final paper documenting your team's process. The first page of this is your justification paper, but the remaining parts of the paper should include information about: how your approach or the question is unique; what data and conclusions are already available; how the data has been altered (cleaned); what inferences you were able to make during the data analysis portion of the project; what successes and failures you had during the model creation process; what conclusions you are able to make in reference to your question; and what future work is available in further exploration of your question. Including appropriate diagrams of appropriate size (but not including any significant amount of code, or the required bibliography, title, and abstract), this should be at least 10 pages, single-spaced with 1” margins, 12pt Times New Roman font, no excessive spacing, and you should have at least 10 in-line cited sources. You may use MLA, Chicago, or AMS citation styles.

Be very careful to pay attention to all deadlines noted below. There will be no allowances for late turnins, including slip days. Also be sure to read and understand the rubrics associated with the various portions of this project (also below).

Team organization:

Your team must select one member of your group to be the team lead. This person will be in charge of communicating with me at various points during the remainder of the semester as well as creating and maintaining the github repository.

Github Use:

Github is required for storing and updating your code, paper, and presentations. Each member must do their own pull requests (you are not allowed to have a central person handling all changes). This is done for two reasons: to make sure everyone gets experience using a source control mechanism and to make sure that everyone is contributing equally. We will be dedicating a class to the ins and outs of github and git.

NOTE: you will likely not be able to store your data files on github due to their requirement that every file be <100MB in size. A Dropbox or Google Drive link will be sufficient for distribution of your data.

Presentation:

During the final class you and your team will be giving a professional, 20 minute presentation on your question, methodology, and results. Your presentation must include the following components:

* Your team’s question and the motivation for answering said question
* Establishing what results are already available
* Your team’s methodology, including justification for the methodology
* Your results delivered in a clear, concise manner
* What the next steps in this project would be

Presentations will be evaluated by the instructor and the class. You will be evaluated on clarity, content, ability to answer questions, presentation ability (this includes being able to deliver a clean, well-rehearsed lecture, with appropriate body language, and no technical hiccups), and professional appearance (both attire and slides). After the presentation, there will be a 10-minute Q&A session.

Individual Rubric for Presentation:

35% - Class average evaluation
35% - Instructor average evaluation
10% - Duration
10% - Discussion of all required components
10% - Even distribution of speaking across all members of the team

Deductions:
-25 presentation slides not in correct format/turned in on time
-25 substantive changes to presentation slides after deadline
-100 not participating in the presentation

Deliverables:

Justification paper – by the deadline listed below, your team leader should email me your justification paper in docx or pdf format.

Project code + data – by the deadline listed below, you will need to finalize your project code and README.md file in github. You will also need to send me a link to your clean data so that I can reliably run your project. Your README.md file needs to include the following information:

* Project title
* Team member list
* Project abstract (250-500 words)
* Target language and version
* List of any additional packages required with installation instructions

Project paper – by the deadline listed below, you will need to finalize your paper in docx or pdf format in github.

Presentation slides – your presentation slides will be posted to github and cloned after the deadline noted below and should be in either pptx or pdf. You are allowed to make minor changes, but substantive changes in the slides will result in a deduction.

Team member reviews – by the deadline listed below, each team member should send me an email containing a numeric (0-100) grade for each of the other team members complete with justification.

Due dates:

All team submission emails should come from the team leads.
3/21, 5pm - Link to the github page for your project via email

3/28, 5pm – Selection notification of data set(s) via email
4/7, 5pm – Selection of question to be answered, justification paper via email
4/27, 5pm – Final clone of the github repository for all code and the final paper
4/27, 5pm – Link for the clean dataset(s) via email or dataset available via github
4/27, 11:59pm – Team member reviews via email
4/27, 11:59pm – Presentation slides final clone
4/28, 2pm – Final presentations

Individual Rubric:

25% - team evaluations
35% - code and model correctness
25% - final paper (content, length, grammar, and citations)
15% - successful completion of early deadlines

Deductions:
-10 team evaluations not in correct format/turned in on time
-25 lack of github activity
-15 lack of comments in the code
-100 not attending the final presentations