Final Project: Open-Ended Project

Presentations + Main Deliverables due at end of semester (details below)

For your Shiny app project, you were tasked with wrangling a messy dataset and creating an effective interactive Shiny application. The intended audience for your presentation was this Data Science class.

For your final project, you'll continue to practice those same skills—asking good questions, wrangling data, and communicating results (the data analysis cycle)—but with some differences. First, you'll be on your own (individual project) and second, this time, you'll take the data analysis a step further and incorporate some of the exploratory data analysis (EDA) techniques introduced in this class, and then expanding on what we've learned in class.

This project is, again, deliberately open-ended to allow you to explore your creativity and interests. There are only three main rules that must be followed:

1. **Your project must be centered around data.** While you can use data related to the data set from your Shiny app project, you must use at least one new data set in some way. In other words, you cannot use only the wrangled data from the Shiny project as your data set. You could start with it and join something to it, or get information from different years that might have a similar wrangling process, or just use an entirely new data set for a new topic, etc. We want to practice wrangling more!
2. **You must incorporate (at least) one of these four topics introduced in the second half of the semester: text analysis, network science, unsupervised learning (clustering), and/or spatial data, and go beyond what we did in class.** Although we only spend a short time introducing each of these topics, there are entire courses dedicated to each one! This project provides you an opportunity to take a deeper dive into one or more of these exploratory analyses. The idea is for you to go beyond what we did in class - try new methods, make more detailed plots, etc.

Examples of *going beyond*:

- Clustering - We focused on k-means in class, but the text illustrates hierarchical clustering. Try hierarchical clustering or a different clustering method (ask me for ideas), or investigate the problem of choosing the number of clusters, or cluster validation (how do you know you have a good clustering solution?).
- Text Analysis - We focused on analysis of single words in class. What about bigrams? We didn't run any models looking at relationships between characteristics of documents, nor did we explore methods to determine themes/topics (you can look at latent dirichlet allocation).
- Network Science - We explored a few ways to work with edge weights and explore centrality. But there are other descriptive statistics, other centrality measures, etc. There is more to dive into with directed or weighted graphs, concepts of small world graphs, modeling, or even random graph generation. We saw two clustering (community detection) methods on networks - there are more, plus ways of assessing the solution, to explore.
- Spatial data - We focused on static plots and only toyed with leaflet a little bit. What can you do if you combine leaflet and Shiny? Is there anything you want to do exploring different projections or their impact on analysis? We didn't explore Chapter 18's material either. Some basic spatial modelling may be possible as well.

3. **Your project must tell us something meaningful.** On one extreme are *data art* projects like the Dear Data project or Memo Akten's *Forms*, which may involve little statistical analysis and are predominantly visualizations. On the other extreme are data mining projects like the KDD Cup annual

data mining competition, which may involve little visualization and much more analysis. Your project can be anywhere on this spectrum, but expectations may be different depending on where you are on the scale. An example of a project that doesn't tell us anything, would be something that downloads a single data source and summarizes it, with some perfunctory visualization. Make sure that your project is thought-provoking and has some underlying meaning!

Aside: Organization

You will be using your individual course repos for this project. This may mean that you need to "clean up" your repo considerably - it will need a final project folder at a minimum. Let me know if I can assist with this (IMO, using the Github Desktop app makes this much easier than within RStudio). If you end up wanting to use data files from the Shiny project, you will need to copy them into your repo (be sure you have everything needed for your work to be reproducible. In some cases, that may mean additional cleaning of your wrangling files to be sure the sources are listed there for all the files!).

Importantly, if you decide to move old files into a project folder, be sure you keep file structures the same, so that you don't break data read in commands. In other words, don't change where files/folders are relative to each other.

Ask me if you have any concerns about your file organization.

In the end, it should be absolutely clear which files are for which individual project (this final one or calendar query), along with all our other classwork, whether that's through folder organization or file names.

While it may seem odd to focus on this (we could just create another repo after all), keeping organized is a valuable skill to practice. After all, if you write a major report for your boss in your job after Amherst one day, save it to your computer, but can't find it the next day to email it, or can't find it two years in the future when someone asks for it as evidence of your work product for a big promotion, that won't serve you well.

Some points for the project are allocated for organization. These should be easy points to earn! See the rubric for details.

Components

The final deliverables for this project will be:

1. **Report:** A written report created using RMarkdown (for reproducibility). If interactivity would be useful to your project, you are welcome to incorporate a Shiny application into your report (you can link to it in the report, and provide a link in the readme in your repo). You could also incorporate an executive summary or anything else you'd like to practice. But you must have a report to submit.
2. **Presentation:** A 6-8-minute oral presentation delivered to the class.

As with the previous project, there will be several checkpoints along the way.

Project Proposal (Thursday, April 11th by midnight)

As with the Shiny app project, we will use a project proposal so that I can get a sense of your projects and help direct you. I will work with individuals if major revisions are needed, but hope that your proposals will let you get into the necessary work quickly.

The project proposal will be submitted to your repo in a way so that it's obvious it's for the final project. (See organization above.) Please be sure it is a .pdf file so that I can pull it into my note-taking software for comments easily. You are responsible for creating the file (no template).

Your proposal should contain the following content:

1. Do you plan for your final project to be an extension of the mid-semester project?

- *If Yes*: Identify specific ideas for how you will extend your mid-semester project. The more details the better here. In particular, what is your plan to add additional data?
- *If No*: Include details regarding the new general topic / phenomena you want to explore and the questions you hope to address. Identify reasonable data sources and how you will acquire the data (web scrape? download? specific packages? API?).

2. Exploratory topic chosen. Be sure to include which topic(s) you will incorporate: text analysis, network science, unsupervised learning (clustering), and/or spatial data. What, if any, questions do you have about the topic heading into the project? What are potential ways you might *go beyond*? (This will help me provide you with resources and references, or allow me to tell you an option is beyond the scope of what you should attempt for the project.)

3. Describe what you hope to deliver as a final product. Obviously, you will generate a report. But what else? Will your project include a published Shiny application? Will it incorporate an interactive map? Will it involve a predictive model that forecasts future values of some quantity using data that you've integrated? Do you want to write an executive summary? Or a pamphlet to distribute for informational purposes? Etc. (You can have whatever deliverables you want on top of the report.)

Don't forget you can use statistical methods from outside of class that you are comfortable with. Maybe you want to combine text analysis with regression analysis (intro stat / stat 230), or network analysis with a nonparametric method to compare groups (stat 225), for example. Use your statistics knowledge!

4. Outline a schedule for your progress that will take you from now (ideas phase) to final report and presentation at the end of the semester. (See timeline below for assistance.) During the shiny project, we had specific checkpoints for different phases of the project. Based on what you envision for your final project, identify checkpoints and dates by which you plan to reach those checkpoints. Feel free to share these with your classmates so you can hold each other accountable, so you're not waiting until the last minute to do things! In particular, I suggest you aim to have at least one checkpoint each week (ideally two) identifying what work you expect to complete by then.

You may want to start using the option of creating issues in the repo to help keep to the schedule. For example, suppose your schedule has the following major components apart from the report and presentation:

- identify new data source,

- wrangle data and add to shiny app data,
- research *go beyond* topic,
- run network statistics,
- run regression models.

You could create an issue with each of these and a deadline for yourself. Then, when you finish the task you can close the issue.

Status update 1 (Monday, April 22nd by midnight)

This checkpoint is here especially to allow for revisions/updates to your project proposal as you get going. In particular, if I tell you in the feedback on the proposal that you must adjust something, this is the date you need to have it adjusted by. Submit the revision to your repo just like the original proposal, but with revise or revision in the filename to distinguish it.

After addressing that (as needed), create an issue in your course Github repo with a name like "Status Update" or "Status Update 1".

In the update, you should provide details on the progress you've made and whether or not you've achieved the work you expected to by this point in your schedule. If you're behind schedule, adjust your checkpoints and come up with a plan to get back on track. Consider why you got behind schedule: were you unable to dedicate as much time to this project as you had hoped to? Or did something in the project take much longer than you anticipated?

This may help motivate you to create other issues and determine how to work on various tasks.

In addition, you may ask questions you have for me. For example, if there is a question about your *go beyond* topic or similar, ask away!

If there are no questions asked of me, I will close this issue after reviewing it. You can view closed issues in Git, if you need to look back.

If you ask questions, I will reply with answers, and you should close the issue after reviewing the reply.

Status update 2 (Monday, April 29th by end of our class period)

After working on the project in class, before class ends, create a new issue in your Github repo with a name like "Status Update, continued" or "Status Update 2".

(Everyone should be set with proposals by this point, so this is focused on the update part.)

As before, you should provide details on the progress you've made and whether or not you've achieved the work you expected to by this point in your schedule. If you're behind schedule, adjust your checkpoints and come up with a plan to get back on track. Consider why you got behind schedule: were you unable to dedicate as much time to this project as you had hoped to? Or did something in the project take much longer than you anticipated?

The presentations are later that week and week after (and you may be randomly assigned to either day of the week), with reports due by 5 pm on the last day of classes, Tuesday May 7th. This will allow you to incorporate feedback from the class from your presentations into your report (though it is not much time for those with presentations on Monday). Bear this in mind as you adjust your checkpoints.

As before, you may ask questions you have for me. If there are no questions asked of me, I will close this issue after reviewing it. If you ask questions, I will reply with answers, and you should close the issue after reviewing the reply.

Report (Final version due Tuesday, May 7th by 5 pm)

You should have a nearly final version of your report by the time of your oral presentation. Start early and don't leave the writing till the end! For example, you may be able to write a "data description" section early on while working on the project, but "results" come later.

After your presentation, you may decide to make updates to your report based on the questions and feedback you receive from your peers (described below). The final report is due by 5pm ET on the last day of the semester (not including make-up days).

Your report should tell a data science audience about your project, why they should care about it, and what you have discovered. Assume the readers/audience will be people like you—current or aspiring data scientists. Keep in mind this audience is extraordinarily diverse in terms of skills and abilities, but you may assume some level of familiarity with introductory-level computer science and statistics.

Although not required, a Shiny application or other data product may be included as a deliverable if it would be useful to your project.

Your report should make it clear to me and any other student in the class which methods and techniques you have used to produce your finished product.

Content

In the report, you do not need to present all of the code you wrote throughout the process of working on this project. Indeed, you can follow the same process for data wrangling as used for the Shiny app, making separate files that you then read in. You can describe the wrangling in a few sentences, rather than putting it all in the report. If you go this route, the wrangling file should be clearly labeled, etc. Also, be absolutely certain the entire process is reproducible - if it's not clear where you pulled the .csv files from, or there are many files and the order to access them is not clear, fix that!

The report .Qmd/.Rmd file should contain the minimal set of code necessary to reproduce and understand your results and findings. If you make a claim, it must be justified explicitly in the analysis. A knowledgeable reviewer should be able to compile your file without modification (so, the paths should read in data correctly from your repo without adjustment) and verify every statement you have made.

For the first time, you are also not required to show all the code to the screen. You should experiment to give your report a professional look. Although the necessary code *must* be included within your .Qmd/.Rmd file, much of this code may not need shown. You may want to set `echo: false` as the default code chunk option for the .Qmd/.Rmd file, and only use `echo: true` for code you wish to show the audience. For example, if you used some nifty, new functions and/or some old functions in a creative way, you may want to show the code as a way to teach the audience about these functions and techniques. However, the audience does not need to see every `filter()`, `mutate()`, `summarize()` etc. you use in the course of your analysis. Ask yourself: "Would someone in class know how to do this without seeing the code? Or is it new / creative enough that they should see it?" (Note, your .Qmd/.Rmd must end up in the repo with the compiled "nice" pdf, because I still need to see and check over all this code, including your formatting and documentation.)

You are expected to show the code associated with your *go beyond* part of the analysis, as your classmates will not necessarily be familiar with it. Again, imagine you are trying to teach the class about this new "thing", so that includes the code. You may similarly find it useful to describe all the inputs, etc., particularly if it is from a package that requires a very different sort of data format to run the function.

Bearing all this in mind, you will still need to summarize what is going on with the code/data as you proceed in the analysis. For example, if you tell us the data is for years 1990-2017, but one part of the analysis is only for 2010, you need to state that, but we don't need to see your `filter()`.

Don't forget to describe the wrangling and any challenges you encountered in the report! Since code may be hidden here, we will not know what you did without you stating it. Motivating example: years ago, before the many recent changes to stat comps, a major submitted a report that was very short for their comprehensive evaluation. Their project was focused on a very complicated wrangling task but simple analysis afterwards.

They didn't describe the wrangling at all (even though this was the bulk of their work!), and so, any reader would not know the scope of difficulty in the wrangling process, and thus, the report didn't demonstrate all the skills of the analyst. (Yes, they revised it and it was fine, but this is a reminder to tell us about your wrangling!)

Motivation

Be sure to motivate your topic at the beginning of your write-up. You should try to hook the reader early on. Assume your audience is a skeptical data scientist who has stumbled across your report but has very little time to read it. Can you give them a reason to continue reading? A cool visualization or foreshadowing of results may help.

Format

You do not need to follow a specific format in the report, but you should start with an introduction and finish with a conclusion. Abstracts are not required, but you can include them if you like. Your write-up should address the following questions in some way:

1. Why should anyone care about this? Share your motivations. You may reference external sources (properly cited) to help motivate the topic.
2. What is this about? - Do not assume your readers have any specific knowledge of your subject matter! For example, if your project involves phylogenetic trees, you should assume your audience has only a very simple understanding of genetics.
3. What are the data? - What was the source of the data (who collected it, when, in what way, and why)? Provide appropriate citation. What kind of data was it? Is there a link to the data or some other way for the reader to follow up on your work? You need to reference your wrangling file here if you used one, in addition to providing these details.
4. In what way did you *go beyond*? - What new method/analysis, etc. did you explore? You will need to describe the way in which you *went beyond.* This could include references to new R packages for analysis, citations of textbooks, blogs, or external sources, etc. You don't need to give us a major expository review of your *beyond* topic (no derivations, proofs, or anything of the sort), but give the reader a sense of it, and references for where they could go explore it more.
5. What are your findings? - What kind of statistical computations (if any) have you done to support those conclusions? (Again, even if you display code showing how some of the calculations were performed, it is up to you to interpret, in simple terms, the results of these calculations.) Do not forget about units, axis labels, etc.
6. What are the limitations of your work? - Be clear so that others do not misinterpret your findings. To what population do your results apply? Do they generalize? Could your work be extended with more data or computational power or time to analyze? How could your study be improved? Suggesting plausible extensions doesn't weaken your work; it strengthens it by connecting it to future work.
7. What are your references? - It's been mentioned above that you should be citing things. Thus, you should have a references (or bibliography) section, usually after the conclusion. The data sources must be listed here (and in your wrangling files too), as well as any external sources used. You could have sources for your chosen topic, such as a text analysis textbook. You should also cite any R packages that were not introduced in class materials. (Several groups mentioned these during Shiny app presentations.) This doesn't replace in-line citation that should also be occurring when you use these references. For example, in the report, you could use (Author, Year) when pulling from a particular reference, and then be sure the complete reference is in the bibliography at the end. I am not insisting on a particular citation format (such as MLA or APA), just be consistent. However, do not use footnotes for citations. These are not considered accessible due to their font size.

An example report framework that would incorporate all these is:

- Introduction (Items 1 and 2)
- Data (Item 3)
- Methods (Item 4)
- Results (Item 5)
- Conclusion (Item 6)
- References/Bibliography (Item 7)

Other variants (merging, etc.) are possible and some re-arranging is possible as well.

Style

The Markdown format is designed to be an interactive document, and there is no limit to the length of the report. It should be however long it needs to be to convey your analysis and take home messages. You can include hyperlinks, figures, suggested videos, etc. to provide context for the reader. Just be sure you cite appropriately!

Use Markdown elements like links, lists, LaTeX, and images as needed.

Visualizations, particularly interactive ones if you incorporate a Shiny app, will be well-received. That said, do not overuse visualizations. You may be better off with one complicated but well-crafted visualization as opposed to many quick-and-dirty plots. Any plots should be well thought out, properly labeled, informative, and visually appealing!

The code is there to support the technical reader who wishes to dig into your work – not to substitute for written explanation. Do not present long unbroken chunks of code without offering written explanations. Even though the code may be hidden, someone looking back at your source .Qmd/.Rmd should be able to follow along with everything you do.

Remember that your analysis must be reproducible. Please be sure that the process is reproducible from a clean environment (meaning it will run after you hit the broom to clean the R environment).

Presentation (Wednesday and Monday, May 1st or 6th in class)

Each student will present their report to the class in an 6- to 8-minute oral presentation during class. You will be assessed on the presentation in terms of presentation skills, and will receive feedback about statistical issues to address (those are not graded at this point). You will be randomly assigned to the various days to spread out the presentations to allow for peer feedback/commentary (described below).

An effective oral presentation is an integral part of this project. Communication is key as a data scientist. In their book *Build a Career in Data Science*, Emily Robinson and Jacqueline Nolis emphasize the importance of communication. Here are just three quotes:

- "employers are first and foremost looking for evidence that you can code and communicate about data" (page 59)
- "Much of a data scientist's job is conveying information to nontechnical peers" (page 141)
- "A data scientist needs to be able to communicate. Over and over, people we interviewed for the book mentioned that their success came from communicating their work effectively." (page 280)

You want to show you can communicate your results clearly (with the audience in mind) and concisely. If your audience cannot understand your results or interpretations, then the technical merit of your project is irrelevant.

As with the Shiny project, the intended audience for this presentation is our actual audience: a class of data science students. Your goal should be to convey to the audience what your research questions were and why, what your data was, what you learned after applying your chosen technique(s)/analysis, and your conclusions. You should not tell us everything that you did, nor should you show a bunch of things that you tried that didn't work well.

After hearing your talk, each student in the class should be able to answer:

1. What was your project about?

2. What was the chosen exploratory topic and associated *go beyond* element?

3. What were your findings?

Peer Feedback

This project provides an opportunity for us to explore the value of peer feedback on the report content, presentation, etc., before you submit the final report. Often, as your peers are more removed from the project, others may offer a different take on a graph or ask a question about an analysis that will help you craft the report.

To benefit from the peer feedback, we need to allow ourselves time to include it on the presentation days. As such, our format for presentations will go like this for each presenter:

- Your presentation (6-8 minutes)
- Peer feedback time (max 2 minutes)

You will be provided with a sheet labeled with the presenters for each day. The three questions above will be presented to you for each:

1. What was the project about?

2. What was the chosen exploratory topic and associated *go beyond* element?

3. What were their findings?

There will also be a space for questions/comments.

Filling in the form will provide valuable feedback for your peers. Constructive comments are most useful. For example, "I didn't understand variable X, please explain." or "Check the size of the graphic in your report, it was hard to read." or "Did you have any concerns about assumptions for method X?" are all potentially useful comments for a peer. They don't require a student to completely redo an analysis, but give useful ideas for edits to the report: add a sentence explanation, adjust figure size, add a sentence about assumption concerns, if any.

I will take the feedback and collate it and distribute it back to everyone anonymously, along with my own comments. By reviewing your answers to the provided questions, I can point out other issues to you. For example, if you had a really challenging wrangling experience, but several comments say it looked really easy to wrangle, I can remind you to detail the wrangling more in the report.

Your grade on this portion (10 points, see below) is based on completeness and sharing relevant comments split over both days. For example, if you only fill out the form for one presenter on one day, that isn't going to earn many points. If you fill out the form for everyone, but the only comment for everyone is "Great job!" that's not helpful for your peers. Yes, you should still encourage your peers and if you really enjoyed a presentation, please do say so! Just remember that you are trying to provide feedback to help the student prepare their report, so don't be afraid to ask for clarification on a topic.

Due to doing individual projects, we won't have time to ask questions of the speakers (unless they give really short presentations). Thus, if you have questions for the speaker, be sure to get them on the form.

Also, this means your presentation really does need to be 6-8 minutes. If you go over 8 minutes, I will need to end the presentation so we can fit in everyone (we'll need to do 8 presentations one day and 7 presentations on the other).

Reflection (Wednesday, May 15th by 5 pm)

The reflection consists of a series of questions (different from the Shiny project reflection) designed to help you reflect on the trajectory of your work and development through Data Science this semester. This will again be posted as a Google form, accessible through Moodle. It will open after reports are submitted and run through this day in exam week. It is the only assignment due during finals week. All other coursework is completed before reading period begins.

Timeline and grading details

Remember to see the rubric for details, especially for the presentation and report. Other component points are usually based on items being submitted on-time and complete (i.e. proposal should have all items listed above, etc.), and may not be listed in the rubric if solely based on these items.

Days for project work in class are included to help you in your planning.

Activity	**Points, if any**	**Date**
Project Class Time	-	Monday, April 8th
Project Proposal	10 points	Thursday, April 11th by midnight
Project Class Time	-	Monday, April 22nd
Status Update 1	5 points	Monday, April 22nd by midnight
Project Class Time	-	Wednesday and Monday, April 24th and 29th
Status Update 2	5 points	Monday, April 29th by end of our class period
Peer Feedback	10 points	In class during presentations
Presentation	30 points	Either Wednesday, May 1st or Monday, May 6th in class
Report - Code	30 points	Tuesday, May 7th by 5 pm
Report - Rest	80 points	Tuesday, May 7th by 5 pm
Repo Organization	10 points	Tuesday, May 7th by 5 pm
Reflection	10 points	Wednesday, May 15th by 5pm ET