Peer Group Project Part 2: Blog

For your Shiny app, 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 blog, you'll continue to practice those same skills—asking good questions, wrangling data, and communicating results—but this time, you'll take the data analysis a step further and incorporate some of the exploratory data analysis techniques introduced in this 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.** Preferably, you will work with large, complex and/or messy data. Alternatively, you may work with a dataset that isn't very messy, but is challenging to obtain (e.g., a challenging web scraping task).
2. **You must incorporate (at least) one of the topics introduced in the second half of the semester: text analysis, network science, unsupervised learning, and/or spatial data.** Although we only spend a week or less 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.
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 to no statistical analysis. On the other extreme are data mining projects like the KDD Cup annual data mining competition, which may involve little to no visualization. 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!

Learning objectives

The learning objectives of the Shiny App project are to demonstrate your ability to:

1. identify a set of questions that can be addressed with data available to you;
2. wrangle a large, messy dataset (including gathering, reshaping, and cleaning the data) into a format necessary to answer the question(s) at hand;
3. deploy interactive data visualizations using Shiny;
4. effectively communicate results via visualizations and oral presentation;
5. effectively collaborate with your peers and identify the value in teamwork; and
6. demonstrate awareness of ethical considerations related to data acquisition, management, and communication.

Components

The final deliverables for this project will be:

1. **Report:** A written report in the form of a blog post created using RMarkdown (for reproducibility) and published as a webpage using GitHub Pages. If interactivity would be useful to your project, you're welcome to incorporate a Shiny application into your blog post (either by linking to it, or embedding it directly).
2. **Presentation:** An 8-10-minute oral presentation delivered to the class (either live or recorded).

I'll create a main Data Science webpage for our course that is publicly available and will include a summary and link to each group's blog post.

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

Project plan (Tuesday, November 2)

I will set up the initial blog project repo for you in our course organization on GitHub. The details on your plan for the final blog project should be submitted as a new issue to your GitHub blog repo titled **Project plan**.

Your plan 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. Do you plan to add additional data? Be sure to include which topic(s) you will incorporate: text analysis, network science, unsupervised learning, and/or spatial 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. Describe what you hope to deliver as a final product. Will your blog 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?

3. Outline a schedule for your group's progress that will take you from now (ideas phase) to final blog post and presentation at the end of the semester. During the last project, we had specific checkpoints for different phases of the project. Based on what you envision for your final blog post, identify checkpoints for your group and dates by which you plan to reach those checkpoints. Hold each other accountable, so you're not waiting until the last minute to do things! In particular, you should have at least one checkpoint each week (ideally two) identifying what work you expect to complete by then.

Status update 1 (Tuesday, November 9)

Reply to the **Project plan** issue with an update. In this 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 group 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?

Status update 2 (Tuesday, November 16)

Reply to the **Project plan** issue with a second update, following the same instructions as your first status update.

Blog post (Final version due Wednesday, December 9)

A nearly final draft of your blog post is due by the time of your oral presentation. You may decide to make updates to your blog based on the questions and feedback you receive from your peers after your presentation. The final blog post is due by 5pm ET on the last day of the semester.

Your blog post 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 may be included in the blog post if it would be useful to your project.

Your blog post 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

You do not need to present all of the code you wrote throughout the process of working on this project. However, the 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 Rmd file without modification and verify every statement you have made.

Although the necessary code must be included within your Rmd file, much of this code should not need to be shown on the published web page. You may want to set `echo = FALSE` as the default code chunk option for the 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 on the post 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.

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 blog post but has very little time to read it. Can you give them a reason to continue reading? A cool visualization or result may help.

Format

You do not need to follow a specific format in the blog post, but you should start with an introductory paragraph and finish with a conclusion. These paragraphs need not follow the formal writing style that you would use in most other classes. Here, a colloquial style that is accessible to a lay reader is appropriate. Nevertheless, your write-up should address the following questions in some way:

1. **Why should anyone care about this?**
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)? 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?
4. **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.
5. **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.

Style

The Markdown format is designed to be an interactive document (not dissimilar to a blog entry). Take advantage of this by including hyperlinks, figures, videos, etc. to provide context for the reader. Use Markdown elements like links, lists, LaTeX, and images as needed. Include a bibliography that includes citations for your data and key packages that you are using.

Visualizations, particularly interactive ones, 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 wellthought 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.

I will be reproducing your analyses so please be sure that the process is reproducible from a clean environment.

Presentation (Tuesday, November 30)

Each group will present their blog post to the class in an 8- to 10-minute oral presentation, with the option of presenting live to the class or preparing a pre-recorded video presentation to play during class. You will be assessed on both a group and individual basis for this portion of the project.

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.

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 a clear understanding of your topic, along with a basic understanding of your project, and how well the Shiny app addresses the question(s) you posed. 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 your data like, and what techniques did you apply to it?
3. What were your findings?

Reflection (Wednesday, December 8)

The reflection will be completed individually, and consists of a series of questions (different from the mid-semester project reflection) designed to help you reflect on the trajectory of your group's work together.

Timeline and grade

All project components except the final reflection will be due on **Tuesdays** by 10pm ET so that I may review your material on Wednesdays. See the rubric for additional details.

Activity	Points	Timeline
Initial plan	10 points	Tuesday, November 2 by 10pm ET
Status update 1	5 points	Tuesday, November 9 by 10pm ET
Status update 2	5 points	Tuesday, November 16 by 10pm ET
Presentation	30 points	Tuesday, November 30 by start of class
Blog	60 points	Wednesday, December 9 by 5pm ET
Peer group reflection	10 points	Wednesday, December 9 by 5pm ET