STAT 231: Blog Project

Overview

In the Shiny 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. In the Blog Project, we'll continue to practice those same skills -- asking good questions, wrangling data, and communicating results – but, this time, we'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 interest. There are only three rules that must be followed:

(1) *Your project must be centered around data.* You are welcome to continue working with the same dataset you used in the Shiny project, or find a new dataset to work with.

(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 one 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 involve little to no statistical analysis. On the other extreme are *data mining* projects like the KDD cup, which involve 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!

The final deliverables for this project will be:

- A written report in the form of a blog post created using RMarkdown (for reproducibility) and published as a webpage using GitHub Pages. Note that 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).
- 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.

Timeline and Grade Contributions

All deliverables must be submitted by 10:00 PM on the dates provided at the appropriate submission location. Adherence to these deadlines is required to help us keep on pace for the semester.

Checkpoint	Due Date	Credit	Submission Location
Update 1 (Your Plan)	4/22	10 pts	Issue in GitHub (group blog repo)
Update 2 (Status Update)	4/29	5 pts	Reply to Update 1 issue in GitHub
Update 3 (Status Update)	5/6	5 pts	Reply to Update 1 issue in GitHub
Presentation	5/13 + 5/18	30 pts	[in class] (recorded or live)
Peer Feedback	5/13 + 5/18	7 pts	[in class]
Final Blog Post	5/19	50 pts	Website (using GitHub Pages)
Reflection II	5/19	10 pts	Gradescope

Components

Update 1: Your Plan

The first update is due by 10:00 PM on Thursday, April 22nd via an issue in your group's blog repo. In this update, you should provide details on your plan for the final blog project. Your plan should contain the following content:

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

    a. *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.

    b. *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.

We will use Update 2 and Update 3 to check in on the progress you're making based on the team schedule you come up with.

Update 2: Status Update

The second update is due by 10:00 PM on Thursday, April 29th via a REPLY to the issue you created for Update 1 (Your Plan) in your group's blog repo. 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?

Update 3: Status Update

The third update is due by 10:00 PM on Thursday, May 6th via a REPLY to the issue you created for Update 1 (Your Plan) in your group's blog repo. 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 were hoping to? Or did something in the project take much longer than you anticipated?

Blog Post

A draft of your blog post is due by the time of your oral presentation (either Thursday, May 13th or Tuesday, May 18th). You may decide to make updates based on the class's questions and feedback after your presentation. The final blog post is due by 5:00 PM on Wednesday, May 19th.

In your post, you 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 that this audience is extraordinarily diverse in terms of skills and abilities, so you should assume very little about what they might know. However, your audience is reasonably tech-savvy, so you need not "dumb-down" your analysis.

A Shiny application can be included in the blog post if a Shiny app would be useful to your project. (Shiny apps are *not* required for the final project.)

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

Content You do not need to present all of the code that you wrote throughout the process of working on this project. However, the .Rmd file should contain the minimal set of code that is necessary to reproduce and understand your results and findings. If you make a claim, it must be justified by explicit calculation. A knowledgeable reviewer should be able to compile your .Rmd file without modification, and verify every statement that you have made.

That being said, *much of this code may NOT* need to be shown on the published web page (you'll likely want to make use of the `echo = FALSE` code chunk option to suppress a lot of the code from printing to your published post). If you used some nifty, new functions and/or some old functions in a creative way, you may want to show these 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()` and `summarize()` 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 that 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 can help.

Format You don't 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:

1. Why should anyone care about this?
2. What is this about? Do not assume that your readers have any domain knowledge! The burden of explanation as to what you are talking about is on you. For example, if your project involves phylogenetic trees, do not assume that your audience has anything other than a basic, lay understanding of genetics.
3. Where did your data come from? 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 English sentences, 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 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.

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

Presentation

Each group will present their blog post to the class in an 8-10 minute oral presentation either on Thursday, May 13th or Tuesday, May 18th. Each group will have the option of presenting live to the class or preparing a pre-recorded video presentation to view 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. Whether you choose to pursue a career in academia, industry or government, the ability to communicate clearly is of paramount importance. As a data scientist, the burden of proof is on you to convince your audience that what you are saying is true. If your audience cannot understand your results or their 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 it addresses the question(s) you posed. You should ***not*** tell us everything that you did, or 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?

Peer Feedback

In recognition of the value of peer feedback (both for the people receiving the feedback *and* for the people giving the feedback), we'll incorporate peer feedback more formally into this project. Groups will be assigned to each other such that you'll provide feedback to a group on the day your group is *not* presenting (e.g., if you present on Tuesday, you'll provide feedback to a group presenting on Thursday; and vice versa).

You'll be asked to: (1) identify brief responses to the questions posed under the Presentation section (1. What was their project about?; 2. What was their data like, and what techniques did they apply to it?; and 3. What were their findings?); and (2) provide at least one compliment to the group about their work, ask at least two questions about their content, code or process, and suggest at least one constructive idea for improvement.

Reflection

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.

Assessment Criteria

Presentation (30 points)

	Score				
	Excellent	Good	Satisfactory	Poor	Unacceptable
Team					
Organization (10 points)	9-10 points: Organizational pattern (specific introduction and conclusion, sequenced material within the body, and transitions) is clearly and consistently observable and is skillful, and makes the content of the presentation cohesive	8 points: Organizational pattern (specific introduction and conclusion, sequenced material within the body, and transitions) is clearly and consistently observable within the presentation	7 points: Organizational pattern (specific introduction and conclusion, sequenced material within the body, and transitions) is intermittently observable within the presentation	3-6 points: Organizational pattern (specific introduction and conclusion, sequenced material within the body, and transitions) is barely detectable within the presentation	0-2 points: Organizational pattern (specific introduction and conclusion, sequenced material within the body, and transitions) is not observable within the presentation
Content (10 points)	9-10 points: Central message is compelling (precisely stated, appropriately repeated, memorable, and strongly supported).	8 points: Central message is clear and consistent with supporting material.	7 points: Central message is basically understandable but is not memorable and/or not supported.	3-6 points: Central message can be deduced, but is not explicitly stated in the presentation.	0-2 points: Central message is unclear from the presentation.
Individual					
Language & Delivery* (10 points)	9-10 points: Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation compelling, and speaker appears polished and confident. Language in presentation is appropriate to audience.	8 points: Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation interesting, and speaker appears comfortable. Language in presentation is appropriate to audience.	7 points: Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation understandable, and speaker appears tentative. Language in presentation is appropriate to audience.	3-6 points: Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) detract from the understandability of the presentation, and speaker appears uncomfortable. Language in presentation is appropriate to audience.	0-2 points: Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) detract from the understandability of the presentation, and speaker appears uncomfortable. Language in presentation is not appropriate to audience.

* Language & Delivery will be scored individually, while the other categories will be assessed for the team as a whole.

Blog Post (50 points)

	Score				
	Excellent	Good	Satisfactory	Poor	Unacceptable
Code					
Functionality (10 points)	9-10 points: The code is completely functional and responds correctly producing the correct outputs.	7-8 points: The program is mostly functional and responds correctly producing the correct outputs in most cases. There are minor problems with the program implementation.	4-6 points: The code is marginally functional with numerous errors and/or incomplete code sections.	2-3 points: The code is minimally functional with significant portions of the code missing or incomplete.	0-1 point: The code is not functional, producing no correct outputs, or was not attempted.
Readability (5 points)	5 points: The code is extremely well organized, properly formatted, and easy to follow.	4 points: The code is reasonably easy to read. There are minor formatting problems.	3 points: The code readable only with significant effort. There is little to no proper formatting.	1-2 points: The code is poorly organized and difficult to read. There is little to no consistency in formatting.	0 points: The code is readable only by the author or someone extremely knowledgeable with its layout and purpose.
Documentation (5 points)	5 points: The code is extremely well documented. Comments are completely consistent with the associated code. There are no spelling errors.	4 points: The code is reasonably well documented. There are minor formatting omissions that would have improved users understanding of code purpose. There may be limited spelling errors.	3 points: The code is marginally documented. There are significant portions of the code that are not documented or documented incorrectly. There are significant spelling errors that detract from the documentation.	1-2 points: The code is poorly documented. There are minimal comments and/or the comments are incorrect.	0 points: The code is not documented.
Content					
Blog Post (15 points)	18-20 points: The blog post successfully communicates the goals, procedures and results of the study. Claims are adequately supported. Text and analysis is effectively interwoven. Uses graceful language that skillfully communicates	15-17 points: The blog post successfully communicates the goals, procedures and results of the study. Most claims are adequately supported. Text and analysis is effectively interwoven. Uses straightforward language that generally conveys meaning to	10-14 points: The blog post communicates the goals, procedures and results of the study. Most claims are adequately supported. Text and analysis may seem disconnected and/or the language used generally conveys	5-9 points: The goals, procedures and/or results of the study are unclear. Most claims are not adequately supported. Text and analysis are disconnected and/or language used sometimes impedes	0-4 points: The goals, procedures, and/or results of the study are unclear. None of the claims are adequately supported. Language impedes meaning

	meaning to readers with clarity and fluency, and is virtually error-free.	readers. The language has few errors.	meaning to readers with clarity, although writing may include some errors.	meaning because of error in usage.	because of error in usage.
Visualization/ Analysis (10 points)	9-10 points: All analyses and graphical elements included are appropriate for the variables/relationships under investigation. Any underlying assumptions are met, and analyses are carried out correctly. Figures have clear labels and legends, where necessary. An original view of the data is presented.	7-8 points: All analyses and graphical elements included are appropriate for the variables/relationships under investigation. Any underlying assumptions are met, and most analyses are carried out correctly. Labeling and legends not completely clear. An original view of the data is presented.	4-6 points: All analyses and graphical elements included are appropriate for the variables/relationships under investigation. Some underlying assumptions are not met and/or analyses are not carried out correctly. Labeling and legends not completely clear. An original view of the data is presented.	2-3 points: Some analyses and graphical elements included are not appropriate for the variables/relationships under investigation. Some underlying assumptions are not met and/or analyses are not carried out correctly. Unclear labeling and/or legends.	0-1 point: Analyses and graphical elements are not appropriate for the variables/relationships under investigation and/or an original view of the data is not presented.
Originality/creativity (5 points)	5 points: The topic is interesting and substantial, and the group demonstrates considerable creativity, initiative and ambition.	4 points: The topic is interesting and substantial, and the group demonstrates some creativity, initiative and ambition.	3 points: The topic is interesting, and the group demonstrates limited creativity, initiative and ambition.	1-2 points: The topic is trite, pedantic, or trivial, and the group demonstrates limited creativity, initiative and ambition.	0 points: The topic is trite, pedantic, or trivial, and the group demonstrates no creativity, initiative or ambition.