Stat231: Blog Project

Project Description

Summary

For the final project in this course, you'll continue working in your Power Up Group (PUG) to practice asking good questions, wrangling data, and communicating results. This time, however, you'll take the data analysis a step further and incorporate some of the exploratory and descriptive data analysis techniques introduced in this class.

This project, again, is 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.** 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 methods introduced in the second half of the semester: spatial data, text analysis, network science, and/or unsupervised learning.** 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 thought-provoking and 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, and expectations may be different depending on where you are on the scale. Your project should go beyond downloading a single data source and summarizing it with with some perfunctory visualization. Make sure there is depth to the statistical exploration (Rule 2) and meaning.

There are two final deliverables for this project:

1. **Blog:** A written report in the form of a website created using Quarto markdown (for reproducibility) and published as a webpage using GitHub Pages.
2. **Project Showcase:** An oral presentation delivered to the class summarizing your project.

Timeline

All deliverables must be submitted by 11:59 PM (unless otherwise indicated) on the dates provided at the appropriate submission location. Adherence to these deadlines is required to help us keep on pace for the semester.

Activity	Date	Submission
Blog plan	Thursday, April 10	New GitHub Issue (Blog plan)
Status Update 1	Thursday, April 17	New comment on GH Issue (Blog plan)
Status Update 2	Thursday, April 24	New comment on GH Issue (Blog plan)
Project Showcase	Thursday, May 1	In class
Final blog post	Tuesday, May 6 by 5pm	Website (using GitHub Pages)
Reflection II	Tuesday, May 6 by 5pm	Gradescope

Getting started

Review the project description on your own and review the American Association of Colleges and Universities' (AACU) Teamwork Value Rubric.

Project Repo Setup

Within the our course's GitHub organization, one team member should follow these steps to create a **New repository** from the provided template:

1. In the **Repository template** section, use the drop-down menu to select **acstat231-f25/blog-template**.
2. **Name** the repo using the convention *blog-team-name*, e.g. *blog-mighty-mammoths*
3. List your full names in the repo **Description** to help me keep track of things, e.g. "Brittney Bailey, Kara Yacoubou Djima, and Kat Correia"
4. Make the repo **Public**
5. Add a **.gitignore** template for **R**
6. Hit **Create repository**
7. Go to the repo **Settings** and add team members (only) as collaborators with **Admin** responsibilities. (Professor Correia has access by default so you do not need to add her.)
8. Still in the repo **Settings**, click on **Pages** in the menu on the left-hand side of the page.
9. Under the **Branch** section where it currently says "None", click on the dropdown menu there to select **main**.
10. In the second dropdown that currently says "/(root)", click on the dropdown menu there to choose **/docs**. Then click **Save**. In a couple minutes, you will have a published website!
11. When the link pops up, copy the link and return to the main landing page of your repo to edit the **About** section.

12. Make sure the names of your team members are in the **Description** section and add the link to the website in the **Website** section.

All members should then take the usual steps to open the shared repo in GitHub Desktop and clone the repo to their STAT 231 course folder (where all your other course repos are).

Blog website starter

1. Open `index.qmd` and add your names to the author line. Save your changes and render the file to html.
2. In GitHub Desktop, **commit** both the `index.qmd` and the `index.html` files and **push** your changes back to github.com.
3. Visit your blog website link (e.g., https://acstat231-s25.github.io/blog-mighty-mammoths) to see that the author information (bottom left of page) updated correctly

Note: Do **NOT** change the name of `index.qmd` or `index.html`! These files must keep these names for your blog website to work.

Components

Blog plan

The details on your plan for the final blog project should be submitted as a new issue to your GitHub blog repo titled **Blog plan**.

Your plan should contain the following content:

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

    a. *If Yes*: Identify specific ideas for how you will extend your mid-semester project. The more details the better here. Are you answering additional, related questions? Do you plan to incorporate additional data? Be sure to include which topic(s) you will incorporate: spatial data, text analysis, network science, and/or unsupervised learning.

    b. *If No*: Include details regarding the new general topic/phenomena you want to explore and the questions you hope to address.

2. What are your data sources? (How does it fulfill Rule 2?)

    Identify your data sources. If you are obtaining new data, how you will acquire the data (web scrape? download? specific packages? API?).

3. What do you hope to deliver as a final product? (How does it fulfill Rule 2 and 3?)

    Identify which of the methods from the second half of the semester you plan to incorporate (spatial data, text analysis, network science, and/or unsupervised learning) and how. Discuss any other graphs, tables, or statistical content you hope to incorporate. Will any of your visualization or tables be dynamic? Provide as much detail as possible.

4. Outline a schedule for your group's progress that will take you from now (ideas phase) to final blog post and showcase at the end of the semester. During the last project, we had specific checkpoints for different phases on 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 hour 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 Status Update 1 and Status Update 2 to check in on the progress you're making based on the team schedule you come up with.

Status Update 1

To submit the first update, REPLY to the issue you created for your Blog 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?

Status Update 2

To submit the second update, REPLY to the issue you created for your Blog 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

References

All data sources, any key R packages, and any other sources used in developing your blog should be cited in full with relevant links in a list of references at the end of your blog. Your blog post should also link to these sources as they are discussed as either direct links or in-text citations.

For examples of data citations, see this resource on Citing data sources from Columbia University. Many data repository websites will provide details on how to cite specific datasets, but otherwise you should follow the guidance at the link.

Write-up

You do not need to follow a specific format in the blog post, but you should start with an introduction and finish with a conclusion and list of references. With the exception of the list of references, these components do not need to 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.

Your write-up should address the following questions in some way:

- ☐ **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.
- ☐ **Why should anyone care about this?**
- ☐ **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? All data sources should be cited (include a list of references at the end of your blog).
- ☐ **What are your methods and findings?** What kind of statistical analysis did you do? What are your results/conclusions? Again, even if you display code showing how some of the analyses were performed, it is up to you to interpret, in simple terms, the results of these calculations (avoid statistical jargon to the degree possible when writing conclusions). Do not forget about units, axis labels, etc.

  > **BAD methods/interpretation example (too much jargon!!)**: We fit the regression model $\widehat{\text{Price}} = 120000 + 8(\text{square footage}) + 32152(\text{bedrooms})) - 5(\text{days on market})$ tested the null hypothesis $H_0 : \beta_2 = 0$ against the alternative hypothesis $H_a : \beta_2 \neq 0$ and our test statistics was $t = 2.47$ and our p-value, which is the probability of getting data as extreme or more extreme than the data we observed under the assumption the null hypothesis is true, was only 0.03 which is less than our $\alpha = 0.05$ signifiance level. So we reject the null hypotheis of $H_0 : \beta_2 = 0$ and conclude we have evidence for the alternative hypothesis $H_a : \beta_2 \neq 0$. And we conclude that `bedrooms` is a signficant predictor of `price`. $\hat{\beta}_2 = 31152$ mean for each additional bedroom the home price goes up $32,152.
  >
  > **GOOD methods/interpretation example**: We used linear regression to model the sale price of homes in Amherst, MA based on the square footage, number of bedrooms of the home, and days on the market (see results in Table 2 [*Table 2 in this case would be a nicely formatted table of coefficients, standard errors and/or CIs, and p-values*]). We found that, after adjusting for square footage of the home and days on the market, for each additional bedroom a house has, the expected sales price increases by $32,152 ($p = 0.03$). Additionally, after adjusting for number of bedrooms and days on the market, the expected sale price increases by $8 per square foot ($p < 0.001$). The number of days the house has been on the market does not seem to have a significant effect on the sale price ($\hat{\beta} = -5$, $p = 0.53$).
- ☐ **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.

Code

Similar to the Shiny app, the website's qmd file(s) should contain the minimal set of code necessary for your published website, and any extensive wrangling should be in separate R scripts.

The code does **not** need to be shown on the published web page. The default code chunk option will hide your code (`echo: false`). Only set `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 showcase

some of 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 (unless the wrangling is part of the exposition!).

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.

As usual, I will be running all of your code and reproducing your analyses so please be sure that the process is reproducible from a clean environment, with relative filepaths, and all files needed to reproduce your work are available.

Visualizations and tables

Visualizations, particularly interactive ones, will be well-received. That said, do not over-use 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.

If you want to include dynamic visualizations or tables, you should explore your options from packages that are built from **htmlwidgets**. These **htmlwidgets**-based packages offer ways to build lighter-weight, dynamic visualizations or tables that don't require an R server to run! A more complete list of packages is available on the linked website, but a short list includes:

- **DT**: Tabular data via DataTables
- **plotly**: Interactive graphics with D3
- **leaflet**: Interactive maps with OpenStreetMap
- **dygraphs**: Interactive time series visualization
- **visNetwork**: Network graph visualization vis.js
- **sparkline**: Small inline charts
- **threejs**: Interactive 3D graphics

I generally discourage the use of Shiny apps within your blog. If your blog project is closely related to the Shiny app you already created, consider linking to it somewhere in the webpage. If an embedded Shiny app is truly imperative to your blog, be aware that there is a limited window size for embedded objects which tends to makes the user experience of the app worse relative to a dedicated Shiny app page. Additionally, Shiny apps will go idle after a few minutes and have to be reloaded by the user, which may also affect the user experience. Any Shiny apps embedded in your blog should be accompanied by the link to the published Shiny app.

More on Quarto website features

There are many many ways to customize your Quarto website. Many of the options are listed under the **Documents –> HTML** section of the Quarto Guide, but here are the main ones you might use frequently:

- HTML Basics for Quarto
- Changing a default theme
- How to link between pages within the website
- Customizing website navigation in the _quarto.yml file
- Creating an "About" page
- Using R HTML widgets in Quarto
- Bootstrap icons for Navigation bar

Project Showcase

Details still in progress but may be a similar process to what we did with the Shiny app demos.

Each person will additionally be asked to read and provide feedback on at least one other blog. Rubrics will be provided.

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.