Stat 628: Data Science Practicum

Module 1 Guidelines

Groups and Deliverables:

You will work in groups of three. Groups will be randomly assigned by the instructor.

Each group will be responsible for (1) the Github repo containing your analysis and write-up (2) the presentation.

Deadlines and Due Dates:

All deliverables are due on Monday, Feb. 5th, 2018. For the Github repo, each group must have pushed/committed all the files before Monday, by 2:25pm CST. For the presentations, each group must **e-mail** the presentation slides (in .ppt,, .pptx, .pdf, or html format) to the TA **before** Monday. It is **your responsibility** to check with the TA in advance that your slides can be displayed properly on the classroom projector before your presentation time.

You may not change or revise your deliverables (presentation slides and Github repos) after the stated due dates above.

Presentations:

Your group will prepare a 5.5 minute (i.e. 5 minutes, 30 seconds), in-class presentation of your data analysis, followed by questions from the audience. The goal of the presentation is to practice presenting your statistical findings in a concise and clear manner. The presentation should include key evidence (e.g. plots, tables, inferential methods, etc.) that support your findings. Your presentation must be clear and precise enough that **any graduate student of statistics** should be able to understand what statistical analysis you used and how you have reached your conclusion. The exact grading rubric for the presentation is outlined below.

Due to time constraints, the 5.5 minute time limit will be *strictly enforced*. To encourage this behavior, every additional 15 seconds after 5.5 minutes will incur a penalty of 1 point. While we will provide time cards and warnings for you during the presentation, it is ultimately **your responsibility** to rehearse your presentation so that it stays under six minutes.

Each member of your group must speak for at least 1 minute during the six-minute presentation. All members of the group must work on the presentation and be prepared to answer questions from the teaching staff or the students.

Each group must **e-mail** the presentation slides (in .ppt,, .pptx, .pdf, or html format) to the TA **before** the first day of presentations. Once submitted, the slides **cannot be changed** during the presentation days. It is **your responsibility** to check with the TA that your slides can be displayed properly on the classroom projector before your presentation time (e.g. during the discussion sections on Tuesday).

Presentation Time & Location:

Presentations will be on Feb. 5, 2018 (Monday) or Feb. 7, 2018 (Wednesday) during lecture time, 2:25pm-3:15pm, at the lecture hall, 494 Van Hise.

The exact time of your group’s presentation will be determined randomly on the first day of the presentation.

Github Repository and Contents

Your group must publish a Github repository that contains all of the data analysis. The repo should consist of three parts: (i) a data folder containing the raw and (if relevant) cleaned data, (ii) a code folder containing all the code for your analysis (e.g. cleaning the data, running the analysis, producing figures/tables, etc.), (iii) an image folder containing any figures/images/tables produced in your analysis.

Additionally, the repository must contain (a) an executive summary folder/file containing a Jupyter Notebook file which must be readable by the Chrome web browser and (b) a README Markdown file briefly summarizing the contents of the repository.

Your repository must include all figures/tables, equations, code, and references. All figures, tables, code, and text must be legible. In particular, code must be clean enough for a data scientist to read.

Executive Summary and Jupyter Notebook

The goal of the “executive” summary of your data analysis is to provide a concise, replicable, and clear description of your statistical analysis and findings. In particular, the summary must include (i) your overall findings, (ii) relevant and important evidence for your findings (e.g. plots, tables), and (iii) important details of your statistical analysis (e.g. type of model used, inferential quantities, outliers, leverage points, modeling assumptions, etc.). Your summary should be detailed enough that any data scientist can read your summary and replicate your analysis. Your summary must include all relevant figures/tables, equations, and references and must be done using the Jupyter Notebook.

All members of the group must contribute to the executive summary. On the summary, the group must clearly indicate each member’s contribution to the project, including each member’s contribution to the presentation, code, and the image files. The final summary should not exceed more than 6 pdf pages.

The last date/time to turn in/commit your repository for grading is before the beginning of the first presentation day.

You may follow any reasonable stylistic guidelines for the references (e.g. MLA, APA, Chicago Manual of Style, etc.)

Grading Rubric:

We will use the following grading rubric to grade your deliverables.

|  |  |
| --- | --- |
|  | Possible Points |
| Presentation | 30 |
| 1. Clear, takeaway message with a “rule-of-thumb” that is easy to use and accurate (4 points) 2. Relevant, concise, and clear summary of statistical analysis (4 points) 3. Relevant (no extraneous plots!) and visually accurate plots (4 points) 4. Strengths and weaknesses of the analysis (4 points) 5. A simple, illustrative example to demonstrate the rule-of-thumb? (4 points) 6. Overall, did the group present convincing evidence for their finding? (5 points) 7. Overall, was the delivery clear and easy to understand? (5 points) |  |
| Jupyter Notebook (or equivalent) | 30 |
| 1. Introduction with clear motivation and thesis statement (2 points) 2. Background information about the data (2 points) 3. Motivation for the model used and statement of the model (3 points) 4. Concise and relevant summary about estimation and inference of relevant parameters, which may include estimated coefficients, R^2, standard errors, confidence intervals, p-values, hypothesis testing statements, and etc. No “data/printout dump” (5 points) 5. Clear, laymen’s interpretation of the estimates and inferential quantities (6 points) 6. Model diagnostics and checking modeling assumptions with plots (5 points) 7. Strengths and weakness of the group’s data analysis (5 points) 8. Conclusion (2 points) |  |
| Other Files on Github Repository | 30 |
| 1. The Readme Markdown file is concise and summarizes the contents of the repository (5 points) 2. Contains clean, readable, well-documented, and error-free code (10 points) 3. Data can be easily read and cleaned using the code provided (10 points) 4. Figures/tables are legible, concise, and clear (5 points) |  |

These points are translated into following grades:

1. To receive a high pass, you must receive at least 80 points.
2. To receive a pass, you must receive at least 65 points.
3. Anything below 65 points is a fail.