Stat 628: Data Science Practicum

Fall 2019, UW-Madison

Module 3 Guidelines

Deliverables and Deadlines:

You will work in groups of three or four. Groups will be randomly assigned by the instructor. You’ll turn in each deliverable as a group.

Please see the following table for due dates.

|  |  |  |
| --- | --- | --- |
| **Deliverables** | **Tuesday Lecture Group** | **Thursday Lecture Group** |
| Presentation 1 slides | Monday, Oct. 28th, 2019 by 5:00pm CST | Wednesday, Oct. 30th, 2019 by 5:00pm CST |
| Presentation 2 slides | Monday, Nov. 18th, 2019 by 5:00pm CST | Monday, Nov. 20th, 2019 by 5:00pm CST |
| Jupyter Notebook summary | Monday, Nov. 18th, 2019 by 5:00pm CST | Monday, Nov. 20th, 2019 by 5:00pm CST |
| Githup repo final commit | Monday, Nov. 18th, 2019 by 5:00pm CST | Monday, Nov. 20th, 2019 by 5:00pm CST |
| Shiny (or web-based) Application | Monday, Nov. 18th, 2019 by 5:00pm CST | Monday, Nov. 20th, 2019 by 5:00pm CST |

Each group must **submit to Canvas**

1. the presentation slides (in .ppt, .pptx, or pdf format)
2. the Jupyter notebook (in .pdf format)
3. a link to your Githup repository
4. a link to a real-time running Shiny app (or Web-Based) app.

Canvas **will automatically** shut down the submission website after 5:00pm on Monday (for Tuesday group) and 5:00pm on Wednesday (for Thursday). Once submitted, the slides, the Jupyter notebook (in pdf format), the Github repo (including your Jupyter notebook in .ipynb format), and the Shiny app **cannot be changed**.

The Jupyter notebook (in .ipynb format) must also be **in** your Github repo.

Presentation Time & Location:

There are **two** presentations for this module. The first presentation is to describe your **data analysis plan and some preliminary analysis of the data**. The second presentation is to describe your **final data analysis**. Presentations will be on Oct. 29th (Tuesday) and Nov. 19th (Tuesday) for the Tuesday lecture group and Oct 31st (Thursday) and Nov. 21st (Thursday) for the Thursday lecture group.

The goal of both presentations 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 business owner listed on Yelp with limited statistical knowledge** (i.e. at Stat 101-level) should be able to understand what statistical analysis you used and how you have reached your conclusion.

**Note**: Remember, these presentations are meant for non-statistical audience, including business owners on Yelp who may have never taken a formal course in statistics, but are interested in making data-driven business decisions.

Due to time constraints, the 6 minute time limit will be *strictly enforced* for every presentation. To encourage this behavior, every additional 15 seconds after 6 minutes will incur a penalty of 2 points. It is ultimately **your responsibility** to rehearse your presentation so that it ends on time.

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

All presentations will be videotaped.

**IT IS YOUR RESPONSIBILTY**, not the **TA** or the **professor**, to make sure that your presentation works properly on the presentation laptop **before each presentation day** (not during the presentation day).

Executive Summary with the 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.

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

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 data in JSON and cleaned data in csv, (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.

Finally, **every student** in the group will be required to commit and contribute to the code basis in the central repo containing your analysis from their individual Github accounts. We also *strongly encourage* you to use the Github platform to create issue requests. We will use (i) the number of commits over time and (ii) the type of commits in the central repo to evaluate each students’ contribution to the code base.

**Note**:If a student’s entire pull and commits for the project are near the deadlines above (e.g. 50 commits a few days from the deadlines), the student will receive **a zero** for this portion.

Generally speaking, industry professionals use your individual Github account as an extension of your resume, specifically to evaluate your ability as a collaborative data scientists and/or software engineer.

Rules and Academic Integrity

Each student assumes the responsibilities of an active participant in UW-Madison’s community of scholars in which everyone’s academic work and behavior are held to the highest academic integrity standards. Academic misconduct compromises the integrity of the university. Cheating, fabrication, plagiarism, sabotaging other groups’ work, unauthorized collaboration, and helping others commit these acts are examples of academic misconduct. Specific examples include, but are not limited to,

1. Copying, plagiarizing, stealing, fabricating any of the deliverables, especially the code or the predictions on Kaggle, from other groups, students outside of the class, or the Internet. In particular, while you may ask other groups for general ideas and questions, you cannot ask for help cleaning the data set, analyzing the dataset, and doing other activities that would be inconsistent with the academic integrity at UW-Madison. If you are unsure, you are always welcome to ask the TA or the professor.
2. Using unauthorized sources, including the original Yelp dataset on Yelp’s website or the original ratings (or summaries of ratings of businesses) which can be derived from Yelp’s website. You are also not allowed to directly copy, steal, plagiarize, paraphrase, or use any analysis that was already conducted on the Yelp data by others (e.g. data science courses online, someone’s blog post or R markdown, Google Cloud’s API platform for sentiment analysis, any pre-written software/code that does sentiment analysis automatically, etc.).

However, you are **strongly encouraged** to browse through Yelp, resources on natural language processing (NLP), sentiment analysis, and other researchers’ analysis of the Yelp data and gather **background information**. You are strongly encouraged to use the information from your background research **to complement** your own analysis and **provide proper attributions**. In short, your analysis of the data must be **original** and **must be your own work**. Or, in industry-lingo, you should not be stealing others’ intellectual property.

If you have any questions about this, please come talk to the TA or the professor.

1. Attempting to gain an unfair advantage by recreating the original Yelp data and using predictors that are not part of the data set. You must only work with the data set you were provided with.

You are strongly encouraged to create your own predictors based on the data set you were given. Again, please come talk to the TA or the professor if you have any questions about this.

1. You may not ask someone to do any part of the analysis on your behalf.

Failure to follow these guidelines will result in an **automatic F** **for the course** **and other school-level disciplinary actions**. For more information, refer to [students.wisc.edu/student-conduct/academic-integrity/](https://students.wisc.edu/student-conduct/academic-integrity/).

Grading Rubric:

We will use the following grading rubric to grade your deliverables. Each deliverable will be based on a score from 1 to 3. We’ll take the average of these scores as your final score for module 3.

|  |
| --- |
| Presentations |
| 1. Clear, takeaway messages 2. Relevant, concise, clear, and understandable summary of statistical analysis and plots 3. Statistically correct and interpretable model(s) with an understanding of its strengths and weaknesses? 4. Overall, did the group present convincing evidence for their finding? 5. Overall, was the delivery clear and easy to understand? |
| Jupyter Notebook Summary |
| For the Jupyter Notebook executive summary:   1. Introduction, background information, and thesis statement 2. Motivation for the model(s) used and statement of the model(s) 3. 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” 4. Clear, laymen’s interpretation of the estimates and any inferential quantities 5. Correct and interpretable model with an understanding of its strengths and weaknesses by checking model assumptions and using model diagnostics 6. Conclusion |
| Github Repo |
| 1. The Readme Markdown file is concise and summarizes the contents of the repository 2. Contains clean, readable, well-documented, and error-free code (in Markdown format or well-commented R/Python Code) 3. Data can be easily read. 4. Code that can produce a cleaned CSV data for final analysis. 5. Figures/tables are legible, concise, and clear   Each student contributed to the central Github repo; if a student provided minimal contributions to the repo or only provided “last-minute” contributions described above, **the student** will automatically **get a one** for this entire portion. |
| Shiny Application |
| 1. Does it run in real time? Is it responsive to user input? 2. Is the application robust to user inputs? 3. Does it display **clear, useful**, and **actionable** information to business owners (and maybe to customers of the restaurant?) 4. Does it provide **visually pleasing graphics** and **data-driven analytics/dashboards** to gauge the “health” of the restaurant? |