Stat 628: Data Science Practicum

Fall 2020, UW-Madison

Module 3 Guidelines

Groups:

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

\*\*\*An important note about groupwork\*\*\*: All members of the group must make some **meaningful** and **equitable** contributionsto **every deliverable** below in a professional, respectable fashion; this must be documented in the four-page executive summary. It is **everyone** in the group’s responsibility to reach out to your team members early and have **regular, two-way communication** between team members. I realize that everyone may work at a different pace and fluid communication may be difficult in light of the current situation. But, out of courtesy for your fellow team members, everyone should reach out to their team members *during the first week* when the project is released, set up a plan to divide up the work, and have regular meetings between each other sharing progress on the project.

Deadlines and Deliverables:

Please see the following table for due dates.

|  |  |
| --- | --- |
| **Deliverables** | **Thursday Lecture Group** |
| Narrated presentation (a single .mp4) | Wednesday, Dec. 2nd , 2020 by 11:59pm CST |
| Four-page executive summary (a single .pdf) |
| A link to a real-time running Shiny (or web-based) App |
| A link to Github page containing all (documented) code, README, executive summary, Shiny app link, and Shiny app server code |

Each group must all four deliverables to **Canvas**. Canvas will automatically shut down the submission website after 11:59pm on Wednesday, Dec. 2nd and no late submission will be accepted. It is **your responsibility** to make sure that all the deliverables are submitted on time to Canvas.

Please be mindful that Canvas has **a combined 500MB upload limit** and, depending on your internet connection, some files may take some time to upload to Canvas; UW Madison recommends leaving at least 24 hours to upload media files.

If possible, we encourage you to submit all the deliverables only once. If you submit multiple times, we'll only grade the **latest submission** that were submitted before the due date.

Narrated Presentations:

Your group will prepare a **10-minute** narrated presentation\*\* of your data analysis. The goal of the presentation is to practice presenting your statistical findings in a concise and clear manner. The presentation must describe how you achieved the overall goals of this project, specifically:

1. Data-driven actionable plan to help business owners improve their ratings on Yelp.
2. Key useful, analytical insights from the data (e.g. plots, tables, models, inferential methods, etc.) that support (a).
3. A demonstration of your Shiny application.

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 the statistical analysis you used and how you have reached your conclusion. In particular, these presentations are meant for business owners on Yelp who may have never taken a formal course in statistics, but are interested in making data-driven business decisions.

The narrations must be akin to giving a presentation to clients in a business meeting typical in consulting settings. In particular, as you advance through slides, your narration should match what’s on the slide at that particular moment in time; see my lecture videos as examples.

All members of the group must work on the presentation and speak **for at least one minute** during the narration. Microsoft Powerpoint, by default, records narrations slide-by-slide (see the third and fourth links below) and you can share your Powerpoint file amongst your group members to achieve this goal. You are not required to have in-person video or closed captioning as part of your narration. But, please make sure to speak loudly and clearly.

Each group will submit **a single** narrated presentation (**in .mp4**) to Canvas. It is **your responsibility** to make sure that the slides and the narrations can be uploaded on Canvas AND that it can be played (with no audio issues) in a standard MP4 video player. Also, because Canvas has an upload limit (see above), I encourage you to use 480p quality for your mp4 files so you don’t exceed the limit.

For details on how to achieve narrate slides in Powerpoint, see the following articles from UW-Madison and Microsoft:

<https://blendedtoolkit.wisc.edu/develop/creating-narrated-presentations-with-powerpoint-for-windows/>

<https://blendedtoolkit.wisc.edu/develop/creating-narrated-presentations-with-powerpoint-for-mac/>

[https://support.microsoft.com/en-us/office/video-record-a-presentation-2570dff5-f81c-40bc-b404-e04e95ffab33](https://support.microsoft.com/en-us/office/video-record-a-presentation-2570dff5-f81c-40bc-b404-e04e95ffab33?ui=en-us&rs=en-us&ad=us)

<https://support.microsoft.com/en-us/office/record-a-slide-show-with-narration-and-slide-timings-0b9502c6-5f6c-40ae-b1e7-e47d8741161c>

You’re also welcome to use any other video editing software to generate a narrated presentation.

\*\*\*An important note about time of narrated presentation\*\*\*:If your presentation exceeds more than **11 minutes**, your group will automatically get a **zero** on the presentation portion.

Four-Page Executive Summary

The goal of the “executive” summary of your data analysis is to provide a concise and clear description of your findings and how you are supporting your finding. In particular, the summary must include (i) your overall analytical insights and data-driven action plan, (ii) relevant and important evidence to support your insights and plan (e.g. plots, tables), and (iii) important details of your statistical analysis (e.g. type of model used, inferential quantities, outliers, modeling assumptions, etc.). Your summary should be detailed enough that **any quantitative-oriented person** can read your summary and understand the type of analysis you conducted.

The summary must be typed in 12-inch Times New Roman or Sans Serif font, single-spaced, with 1-inch margins. The four-page summary must include all relevant figures/tables, but excludes references and contribution section (see below).

On the 5th page of the report, you can include any references that you have used. You may follow any reasonable stylistic guidelines for the references (e.g. MLA, APA, Chicago Manual of Style, etc.).

Additionally on this 5th page, the group must clearly indicate **each member’s contribution** **to (i) the summary, (ii) the presentation, (iii) Github, and (iv) the Shiny app**. For example, you can say that HK (initials of your group member’s name) wrote/edited (blank) part of the summary, worked on slides (blank), created and revised code concerning (blank), edited (blank) on the shiny App and provided general feedback on the Shiny app.

Each group will submit **a single electronic copy (in .pdf)** to Canvas. It is your responsibility to submit the file on time and that it can be opened in a standard PDF viewer.

Github Repository and Contents

Your group must publish a Github repository that contains all of the data analysis. The repo should consist of:

1. a code folder containing all the code for your analysis (e.g. reading/cleaning the data, running the analysis, producing figures/tables, etc.)
2. an image folder containing any figures/images/tables produced in your analysis.
3. The executive summary
4. The Shiny app link and codebase
5. **a README Markdown file** summarizing the contents of the repository and directions on how to use the code.

Your code must be **well-documented** so that **any data scientist in industry** can read and understand your code**.**

Finally, **every student** **in the group will be** **required to commit/pull/push** 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. For grading, 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.

\*\*\*Important note about Github push/pull/commits near the deadline\*\*\*: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.

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.

\*\*\*Important note about individual grading for groupwork\*\*\*: If the teaching staff finds that there was significant unequitable contribution among team members and there were several attempts by team members and teaching staff to have regular two-way communication, we may **assign grades individually** based on the documented contribution on the four-page summary.

|  |
| --- |
| Presentations |
| 1. Clear, takeaway analytic insights and data-driven business plan? 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 insights + business plan? 5. Overall, was the delivery clear and easy to understand? |
| Executive Summary |
| 1. Introduction and background information 2. Clear summary of analytical insights and data-driven business plan based on these insights. 3. Concise, useful and relevant plots/tables summarizing the data analysis 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/R printout dump”** (properly format your tables/plots so they look presentable!)    1. Clear, laymen’s interpretation of the estimates and any inferential quantities    2. Correct and interpretable model with an understanding of its strengths and weaknesses by checking model assumptions and using model diagnostics 5. Conclusion |
| Github Repo |
| 1. The Readme file is concise and summarizes the contents of the repository 2. Contains clean, readable, well-documented, and error-free code 3. Data can be easily read and cleaned using the code provided 4. Figures/tables are legible, concise, and clear 5. Contains the executive summary file 6. Contains the Shiny app link and shiny app code base.   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 zero** 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? |

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, 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. Sabotaging others’ work by removing, deleting, manipulating their code base or statistical analysis or providing misleading and/or false information about the data, relevant analysis, or any information related to the module.
2. You may not ask someone or some entity to do any part of the analysis on your behalf.

Committing said acts can result in disciplinary action, which includes, but is not limited to failure on the assignment/course, disciplinary probation, or suspension. Substantial or repeated cases of misconduct will be forwarded to the Office of Student Conduct & Community Standards for additional review. For more information, refer to [students.wisc.edu/student-conduct/academic-integrity/](https://students.wisc.edu/student-conduct/academic-integrity/).