Stat 628: Data Science Practicum

Module 2 Guidelines

Deliverables and Deadlines:

|  |  |
| --- | --- |
| Deliverables | Due Date |
| Narrated Presentation (.mp4) | Sunday, Oct. 25, 2020, 11:59pm CST |
| Two-Page Executive Summary (.pdf) | Sunday, Oct. 25, 2020, 11:59pm CST |
| Github Page Containing All Code and README | Sunday, Oct. 25, 2020, 11:59pm CST |
| Shiny (or Web-based) App | Sunday, Oct. 25, 2020, 11:59pm CST |

All deliverables will be submitted to Canvas. For the presentation and two-page summary, you must submitone file for the narrated presentation and one file the two-page executive summary. For the Github and Shiny app, you must provide (i) the link to the Github page that is hosting all the code with a README and (ii) the link hosting a functioning Shiny App loaded in a Shiny App server (or an equivalent web-based app) on Canvas. If possible, we encourage you to submit all the deliverables only once.

It is **your responsibility** to make sure that all the deliverables are submitted on time to Canvas. No late submissions will be accepted. We'll only grade the latest submission that were submitted before the due date.

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.

Each deliverable will be graded on a 1-3 scale.

Groups:

You will work in groups of three or four. Each group will be responsible for all the deliverables.

Groups will be randomly assigned and your group assignment is available on Canvas.

Narrated Presentations:

Your group will prepare a **6-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 should include key evidence (e.g. plots, tables, inferential methods, etc.) that support your findings. You may also use the presentation to briefly demonstrate your Shiny application, but this is not required.

Your presentation must be clear and precise enough that **any student** **with a quantitative background (not necessarily in 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.

The narrations must be akin to giving an in-class presentation. 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.

Two-Page Executive Summary

Your group must submit a two-page executive summary of the data analysis. The summary must include (i) your overall findings and the rule of thumb, (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 **understand** your summary and **replicate** your analysis. The exact grading rubric for the summary is outlined below.

All members of the group must contribute to the summary. On the two-page summary, the group must clearly indicate **each member’s contribution** **to the summary and the presentation**. For example, you can say that HK (initials of your group member’s name) wrote/edited (blank) part of the summary and worked on slides (blank).

The summary must be typed in 12-inch Times New Roman or Sans Serif font, single-spaced, with 1-inch margins and must include all relevant figures/tables, and equations. All figures and tables must be legible when printed on a standard 8x11 paper. Also, the summary must not exceed two pages. If you have references/bibliography section, you can use one additional page. However, this additional page must only contain references that you used in the two-page summary.

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

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 three parts:

1. a data folder containing the raw and (if relevant) cleaned data
2. a code folder containing all the code for your analysis (e.g. cleaning the data, running the analysis, producing figures/tables, etc.)
3. an image folder containing any figures/images/tables produced in your analysis.
4. The two-page summary file above.
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 undergraduate student in Statistics** can read and understand your code.

All members of the group must contribute to the R code. On the R code, the group must clearly indicate **each member’s contribution to the code**, including (i) who created the code, (ii) who revised/reviewed/maintained the code, and (iii) who is/are ultimately responsible for different portions (or all portions) of the code. For example, you can say that HK (the initial of your group member’s name) created/revised/maintained the data cleaning code, revised the lm() portion of the code, and created the code for Figures (blank) and (blank) in the summary and the presentation. It is **strongly encouraged that you use Github pull/push/commit** functions to manage your project files and to make it easy to record this information; using Github this way will be required for the final project.

Finally, we remind all students that it is **your responsibility** to make sure that the code is not copied/plagiarized/fabricated from unauthorized sources and your code produces the results reported in your other deliverables (see reminder below).

Shiny App:

Often, data science jobs expect you to make “actionable” prototypes/products based on your data analysis. To this end, you will create a Shiny (or a web-based) application that will run your body fat calculator in real-time. Shiny is an easy-to-use platform to turn your R analysis into web-based applications. For more information about Shiny, visit: <https://shiny.rstudio.com/>.

While you do not have to specifically use Shiny (if you have app development experience, feel free to use any language/platform!), all applications must run on the latest Chrome browser. This is to make sure that the application also runs on both desktop and mobile interfaces.

We’ll leave the user-interface and other graphical specifications up to you. However, your application will be graded on (i) whether it runs in real-time, (ii) whether it is robust to erroneous inputs, (iii) whether it provides useful and insightful information to the end user, and (iv) whether there is some form of a contact information if the end-user has questions about the application.

When publishing the Shiny App, the group must write **who have participated in building the App**. The App must also contain contact information about who maintains the App so that bug requests can be directed to him/her/them.

Grading Rubric:

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

|  |
| --- |
| Narrated Presentation |
| 1. Clear, takeaway message with a “rule-of-thumb” that is easy to use and accurate 2. Relevant, concise, and clear summary of statistical analysis 3. Relevant (**no extraneous and/or “R-dump” plots**!) and visually accurate plots 4. Strengths and weaknesses of the analysis 5. A simple, illustrative example to demonstrate the rule-of-thumb. 6. Overall, did the group present convincing evidence for their finding? 7. Overall, was the delivery clear and easy to understand? |
| Two-Page Summary |
| 1. Introduction with clear motivation and thesis statement/rule-of-thumb. 2. Background information about the data 3. Motivation for the model used and statement of the model 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!) 5. Clear, laymen’s interpretation of the estimates and inferential quantities 6. Model diagnostics and checking modeling assumptions with plots 7. Strengths and weakness of the group’s data analysis 8. Conclusion |
| Github Repository |
| 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 two-page summary file. |
| Shiny Application |
| 1. Does it run in real time? 2. Is the application robust to user inputs? 3. Does it provide useful and insightful information to the user? |

A Reminder: Academic Integrity

Each year in Stat 628, we have a few students who violate the academic integrity standards laid out in class. We take this opportunity to remind students of the policies regarding 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 plots/tables, from other groups, students outside of the class, or the web. 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, especially using someone else’s analysis of the dataset on the web, either in its entirety or in parts. You are also not allowed to copy, steal, plagiarize, paraphrase, or use any analysis that was already conducted on this data (or a derivative thereof) from others (e.g. data science courses online, someone’s blog post or R markdown, Google Cloud’s API platform, AWS Machine Learning library examples, or Azure ML, any pre-written software/code specific to this data or derivative thereof, etc.).

However, you are **strongly encouraged** to browse through resources on body fat, health, and other relevant information to 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. Sabotaging others’ work by deleting, copying, damaging, misrepresenting, or falsifying information about the data, any of the deliverables, or the project.
2. You may not ask someone 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/)