Stat 628: Data Science Practicum

Module 2 Grading Guidelines

Deliverables and Deadlines:

|  |  |
| --- | --- |
| Deliverables (for **Monday Group**) | Due Date (All times are on Central Time) |
| Presentation slides (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Monday, Oct. 16. | Sunday, Oct. 15, 2023, 11:59pm |
| Two-page executive summary (.pdf) | Sunday, Oct. 15, 2023, 11:59pm |
| Github repo containing code (a web link) | Sunday, Oct. 15, 2023, 11:59pm |
| Shiny (or Web-based) app (a web link) | Sunday, Oct. 15, 2023, 11:59pm |

|  |  |
| --- | --- |
| Deliverables (for **Wednesday Group**) | Due Date (All times are on Central Time) |
| Presentation slides (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Wednesday, Oct. 18. | Tuesday, Oct. 17, 2023, 11:59pm |
| Two-page executive summary (.pdf) | Tuesday, Oct. 17, 2023, 11:59pm |
| Github repo containing code (a web link) | Tuesday, Oct. 17, 2023, 11:59pm |
| Shiny (or Web-based) app (a web link) | Tuesday, Oct. 17, 2023, 11:59pm |

**All deliverables** **must be submitted to Canvas**. For the presentation and the executive summary, you must submit (i) one file for the presentation and (ii) one file for the two-page executive summary, all on Canvas. For the Github and Shiny app, you must provide (i) a link to the Github page that is hosting all the code and (ii) a link hosting your functioning Shiny App in a Shiny App server of your choosing (or an equivalent web-based app), all 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. No late submissions will be accepted. We'll only grade the latest submission that were submitted before the due date.

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.

Presentations:

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 enough that any employee 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. If you would like, you can also demo your Shiny app (see below) during the presentation. But, this is not a requirement.

Your group will prepare a **6 minute** in-class presentation of your data analysis; your presentation can be shorter than 6 minutes, but it cannot be longer than 6 minutes. All members of the group must work on the presentation and speak **for at least one minute** during the in-class presentation. The exact time of your group’s presentation will be determined randomly on the first day of the presentation.

Due to time constraints, the time limit will be**strictly enforced**. To encourage this behavior, every additional 30 seconds after the time limit will incur **a penalty of 0.5 points**. It is **your responsibility** to rehearse your presentation so that it stays under the time limit.

Each group will submit **a single** presentation slide (in .pptx, .ppt, .pdf) to Canvas.It is your responsibility to check with me or the TA that your slides can be displayed properly on the projector in the lecture hall before the presentation day.

Two-Page Executive Summary

Your group must submit a **two-page** executive summary of the data analysis, with at most **one additional page** for references & contributions. Specifically, all relevant statistical analysis, plots, tables, and figures must fit in the two- page limit. The one-page reference section **can only be used** for references & contributions (see below).

Your summary must include (i) your overall findings, (ii) relevant and important evidence to support 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 key statistical analysis. The exact grading rubric is outlined below.

On the reference page, you may follow any reasonable style for references (e.g. MLA, APA, etc.).

Also on the reference page, the group must **clearly indicate** **each member’s contribution** to the presentation, the summary, the code, and the Shiny app. For example, you can add a table such as:

|  |  |  |  |
| --- | --- | --- | --- |
| Contributions | Hyunseung Kang | Zhifeng Chen | Jane Doe |
| Presentation | Responsible for slides 1-4 (introduction and data cleaning).  Reviewed/edited slide 5-10 (results). | Responsible for slides 5-10 (results).  Reviewed/edited slides 1-4. | Reviewed/edited and provided feedback on all slides. |
| Summary | Responsible for introduction, data cleaning, conclusion, and references.  Reviewed/edited data analysis section. | Responsible for Figures 1 and 2.  Reviewed/edited and provided feedback on whole document. | Responsible for data results.  Reviewed/edited the introduction, data cleaning, and conclusion |
| Code | Responsible for data cleaning code.  Reviewed code for analysis section | Responsible for methods/results for final models and code to replicate Figures 1 and 2.  Reviewed data cleaning code. | Responsible for methods/results under different models. Reviewed data cleaning code. |
| Shiny App | Responsible for Shiny app | Reviewed/edited and provided feedback on Shiny app | Reviewed/edited and provided feedback on Shiny app |

The summary must be typed in 12-inch Times New Roman or Sans Serif font, single-spaced, with 1-inch margins. All figures and tables must be legible when printed on a standard 8x11 paper.

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 with a standard PDF viewer.

Github Repository

Your group must publish a Github repository (repo) and provide **a web link** to the repository. The Github repo must contain the following:

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, Shiny app base code, etc.)
3. an image folder containing any figures/images/tables produced in your analysis.
4. The two-page, pdf summary file above.
5. The final presentation slides above
6. **a README Markdown file** summarizing the contents of the repository and directions on how to use the code.

The code must **replicate every part of your analysis** from the beginning (i.e. reading in the data) to the end (the figures/tables/results in your presentation and executive summary). This includes, but is not limited to: data cleaning, outlier removal, model building, evaluation of different models, statistical testing, prediction, and any and all intermediary plots, tables, and analysis. Your R code must reproduce the **exact tables, plots, and other analysis** in your summary and the presentation (i.e. exact labels for axis, color shading plots, etc.). Also, your R code also must be **well-documented** so that any undergraduate statistics or data sciencestudent can read and understand your code.

It is **strongly encouraged that you use Github pull/push/commit** functions to manage your project files and to make it easy to record who contributed/worked on different parts of the code.

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

Shiny App:

In some settings, data scientists are expected to make “actionable” prototypes/products based on your data analysis. To replicate this practice in industry, you will create a Shiny (or a web-based) application that will run your body fat calculator in real-time and **submit a link to your** **live/running Shiny app** for grading.

Shiny is an easy-to-use, R-based platform to turn your R code into a web application. While you do not have to use Shiny (if you have app development experience, feel free to use other languages/platforms!), all applications must run on the latest Chrome browser. For more information about Shiny, visit: <https://shiny.posit.co/>.

We’ll leave the user-interface and other graphical specifications of the Shiny app up to you. However, your application will be graded on

1. whether it runs in real-time,
2. whether it is robust to erroneous inputs (e.g. it won’t accept text input in a numeric field)
3. whether it provides useful information to the end user, and
4. whether there is some form of a contact information if the end-user has questions about the application. Specifically, the app must also contain some contact information about who maintains the application.

Grading Rubric:

We will use the following grading rubric to grade your deliverables. We will grade the presentation, the two-page executive summary, and the Shiny app on a 0-10 scale. While the code is not graded, you must also submit the code to receive a grade for Module 2.

|  |
| --- |
| Presentation |
| 1. Clear, takeaway message/final model 2. A simple, illustrative example to demonstrate the model. 3. Relevant, concise, and clear summary of statistical analysis 4. Relevant (**no extraneous and/or “R-dump” plots**!) and visually accurate plots 5. Strengths and weaknesses of the analysis/model 6. Overall, did the group present convincing evidence for their finding? 7. Overall, was the delivery clear and easy to understand? 8. Was the presentation under the prescribed time mentioned above? |
| Executive Summary |
| 1. Introduction with clear motivation and clear statement of findings/model 2. Background information about the data 3. A clear, concise, and strong rationales behind the analysis/findings/model 4. Concise and relevant summary of estimates and inference of relevant parameters in the model. This 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 assessment of assumptions underlying your statistical analysis. 7. Strengths and weakness of the data analysis and model. 8. Conclusion 9. Does it follow instructions concerning the page limit? |
| 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? 4. Is there contact information in case if users have any questions about the app? |

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, or stealing any of the deliverables from other groups, students outside of the class, or from previous iteration of the course. In particular, you cannot ask anyone besides your own group members to clean the data, create tables, figures or plots, or provide parts of their summary. If you are unsure, you are always welcome to ask the TA or the instructor.
2. Using someone else’s analysis of the dataset, 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 API (e.g. Amazon Sagemaker), Azure ML, 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**.

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. Using ChatGPT or other generative AI to write any portion of your two-page summary or to create any part of your presentation.

However, you may use ChatGPT, Github Copilot, or related software to help you with your R code or the Shiny app, especially to debug R code, to improve existing R code that you originally wrote, or to improve your Shiny App’s performance. **If you do this, you must provide proper attributions of your code** (e.g. this portion of the code was written by ChatGPT, version BLANK, access date BLANK, prompt BLANK). Also, if you do this, please note that **you ultimately bear the responsibility of the code from the generative AI**, specifically its correctness and accuracy.

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 <https://conduct.students.wisc.edu/academic-misconduct/>