Stat 628: Data Science Practicum
Module 2 Guidelines

Groups and Deliverables:
You will work in groups of three or four. Groups will be randomly assigned by the instructor.

Each group will be responsible for (1) the Github repo containing your analysis, (2) the executive Jupyter Notebook summary, (3) the presentation, and (4) a Shiny (or web-based) App that runs your body fat calculator in real-time.

Due Dates:
Please see the following table for due dates. No late submissions are accepted.

Deliverables	**Tuesday Lecture Group**	**Thursday Lecture Group**
Presentation slides	Monday Oct. 7th, 2019 by 5:00pm CST	Wednesday, Oct. 9th, 2019 by 5:00pm CST
Jupyter Notebook summary	Monday, Oct. 7th, 2019 by 5:00pm CST	Wednesday, Oct. 9th, 2019 by 5:00pm CST
Githup repo final commit	Monday, Oct. 7th, 2019 by 5:00pm CST	Wednesday, Oct. 9th, 2019 by 5:00pm CST
Shiny (or Web-Based) App	Monday, Oct. 7th, 2019 by 5:00pm CST	Wednesday, Oct. 9th, 2019 by 5:00pm CST

Each group must **submit to Canvas** (i) the presentation slides (in .ppt, .pptx, or pdf format), (ii) the Jupyter notebook (in .pdf form), (iii) a link to your Githup repository, and (iv) 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).

Important: The Jupyter notebook (in .ipynb format) should also be **in** your Github repo.

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**.

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).

Presentations:
Your group will prepare a 6.5 minute, 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 6.5 minute time limit will be *strictly enforced*. To encourage this behavior, every additional 15 seconds after 6.5 minutes will incur a penalty of 1 point. It is **your responsibility** to rehearse your presentation so that it stays under seven minutes.

Each member of your group must speak for at least 1 minute during the presentation. 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 for review.

Presentation Time & Location:
Presentations of your work will be on Oct. 8th, 2019 (Tuesday) for the Tuesday lecture group and Oct. 10th, 2019 (Thursday) for the Thursday lecture group at the lecture hall.

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 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 3 pdf pages**.

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

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.

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

Presentation
a. Clear, takeaway message with a "rule-of-thumb" that is easy to use and accurate and a simple illustrative demonstration of the rule-of-thumb. b. Relevant, concise, and clear summary of statistical analysis c. Relevant (no extraneous plots!) and visually accurate plots d. Strengths and weaknesses of the analysis e. Overall, did the group present convincing evidence for their finding? f. Overall, was the delivery clear and easy to understand?
Jupyter Notebook
a. Introduction with clear motivation and thesis statement b. Background information about the data c. Motivation for the model used and statement of the model d. 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" e. Clear, laymen's interpretation of the estimates and inferential quantities f. Model diagnostics and checking modeling assumptions with plots

g. Strengths and weakness of the group's data analysis h. Conclusion
Other Files on Github Repository
a. The Readme Markdown file is concise and summarizes the contents of the repository b. Contains clean, readable, well-documented, and error-free code c. Data can be easily read and cleaned using the code provided d. Figures/tables are legible, concise, and clear
Shiny Application
a. Does it run in real time? b. Is the application robust to user inputs? c. Does it provide useful and insightful information to the user?