Stat 628: Data Science Practicum
Spring 2018, UW-Madison
Module 2 Guidelines

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

Each group will be responsible for (1) the Github repo containing your analysis and the executive Jupyter Notebook summary, (2) **two** presentations, and (3) a Kaggle prediction submission.

Deadlines and Due Dates:
Please see the following table for due dates.

Deliverables	**Monday Lecture Group**	**Wednesday Lecture Group**
Presentation 1 slides	Sunday, March 4th, 2018 by 11:59pm CST	Tuesday, March 6th, 2018 by 11:59pm CST
Presentation 2 slides	Sunday March 11th, 2018 by 11:59pm CST	Tuesday, March 14th, 2018 by 11:59pm CST
Githut repo and final Jupyter Notebook summary	Sunday, March 11th, 2018 by 11:59pm CST	Tuesday March 14th, 2018 by 11:59pm CST
Kaggle prediction	Sunday, March 11th, 2018 by 11:59PM CST	Sunday, March 11th, 2018 by 11:59PM CST

For both presentations, each group must **e-mail** the presentation slides (in **.ppt,, .pptx, or .pdf** format) to the TA; see the times for the two presentations below.

For the final Github repo containing your Jupyter Notebook, each group must have pushed/committed all the files before this times. Each group must send the link to the Github repo to the TA by the dates specified above.

For the Kaggle predictions, each group must have made their last predictions by the dates specified above.

Once finally submitted, the slides, the Github repo, and the Kaggle predictions **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).

Presentation Time & Location:
Presentations will be on March 5nd, 2018 (Monday) and March 7th, 2018 (Monday) for the Monday lecture group and March 12th, 2018 (Wednesday) and March 14th, 2018 (Wednesday) for the Wednesday lecture group.

There are **two** presentations for this module. The first presentation (March 2nd for the Monday group and March 4th for the Wednesday group) is to describe your **data analysis plan and some preliminary analysis of the data**. The second presentation (March 9th for the Monday group and March 11th for the Wednesday group) is to describe your **final data analysis**.

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 industry professional at Yelp (with or without basic statistics background)** should be able to understand what statistical analysis you used and how you have reached your conclusion.

Due to time constraints, the 5.5 minute time limit will be *strictly enforced* for every presentation. To encourage this behavior, every additional 15 seconds after 5.5 minutes will incur a penalty of 1 point. It is ultimately **your responsibility** to rehearse your presentation so that it stays under six minutes.

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.

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 and 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 5 pdf pages.

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

Kaggle

The class Kaggle website is the following:
https://www.kaggle.com/c/uw-madison-sp18-stat628/

To participate in the Kaggle competition, use the following link:
https://www.kaggle.com/t/c8eeb4ba414a485f8431554c2d98b96c

Please do not share the participation link with anyone outside of the class.

Your group must submit your predicted Yelp ratings for the testing and the validation data by the due date. PLEASE USE THE TEAM NAME "MONDAY_GROUP(Blank)" if you are in the Monday group or "WEDNESDAY_GROUP(Blank)" if you are in the Wednesday group when you submit your Kaggle predictions. Remember to replace the Blank with your group number.

It is your responsibility to learn how to use Kaggle and make sure the submission is properly formatted.

Note that you will be graded based on both the testing AND the validation data. The public leaderboard only presents your standing based on the testing data set and is a good proxy for your performance in the validation data. The private leaderboard, which will be revealed to everyone at the end of the presentations, reveals your standings based on the validation data set.

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.

3. 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.

4. 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/.

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

	Scoring
Presentations	30 points
a. Clear, takeaway messages	

b. Relevant, concise, clear, and understandable summary of statistical analysis and plots c. Statistically correct and interpretable model(s) with an understanding of its strengths and weaknesses? d. Overall, did the group present convincing evidence for their finding? e. Overall, was the delivery clear and easy to understand?	
Github Repo and Jupyter Notebook	40 points
For the Jupyter Notebook executive summary: a. Introduction, background information, and thesis statement b. Motivation for the model(s) used and statement of the model(s) c. 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” d. Clear, laymen’s interpretation of the estimates and inferential quantities e. Correct and interpretable model with an understanding of its strengths and weaknesses by checking model assumptions and using model diagnostics f. Conclusion For the Github repo: 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	
Kaggle	30 points (1 point extra credit possible)
a. You must exceed the benchmarks on both testing and validation data to receive the first 10 points. b. After passing said benchmarks, you will receive points based on your rank in the Kaggle validation leaderboard. Specifically, you’ll receive 1.5 points * position on the leaderboard. There are 14 groups in the class. For example, if you are ranked 1st on the Kaggle leaderboard and you passed the benchmarks, you will receive 10 + 1.5 * 14 = 31 points. If you placed last in the leaderboard and you passed the benchmark, you will receive 10 + 1.5 * 1 = 11.5 points. If you did not pass the benchmark, you will receive a zero regardless of your ranking on the leaderboard.	

These points are translated into following grades:

a. To receive a high pass, you must receive at least 80 points.
b. To receive a pass, you must receive at least 65 points.

c. Anything below 65 points is a fail.