Stat 628: Data Science Practicum

Module 3 Guidelines

Deadlines and Deliverables:

|  |  |
| --- | --- |
| Deliverables (for **Tuesday Group**) | Due Date (All times are on Central Time) |
| Preliminary analysis presentation slide (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Tuesday, Nov. 15th. | Monday, Nov. 14th 2022, 11:59pm |
| Final presentation slide (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Tuesday, Dec. 6th. | Monday, Dec. 5th, 2022, 11:59pm |
| Four-page executive summary (.pdf) | Monday, Dec. 5th, 2022, 11:59pm |
| Github repo containing code (a link) | Monday, Dec. 5th, 2022, 11:59pm |
| Shiny (or Web-based) app (a link) | Monday, Dec. 5th, 2022, 11:59pm |

|  |  |
| --- | --- |
| Deliverables (for **Thursday Group**) | Due Date (All times are on Central Time) |
| Preliminary analysis presentation slide (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Thursday, Nov. 17th. | Wednesday, Nov. 16th 2022, 11:59pm |
| Final presentation slide (.pptx, .ppt, .pdf)  Note: You’ll give an in-class presentation on Thursday, Dec. 8th. | Wednesday, Dec. 7th, 2022, 11:59pm |
| Four-page executive summary (.pdf) | Wednesday, Dec. 7th, 2022, 11:59pm |
| Github repo containing code (a link) | Wednesday, Dec. 7th, 2022, 11:59pm |
| Shiny (or Web-based) app (a link) | Wednesday, Dec. 7th, 2022, 11:59pm |

**All deliverables** **will 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 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 for your lecture group (Tuesday or Thursday). No late submissions will be accepted. We'll only grade the latest submission that were submitted before the due date.

Each deliverable is 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.

Presentations:

You’ll give **two presentations** for module 3. At a high level, the goal of both presentations is to practice presenting statistical material in a concise and clear manner. The presentation should include key evidence (e.g. plots, tables, inferential methods, etc.) that support your key messages that you want to convey. Your presentation must be clear and precise enough that **any business owner listed on Yelp with limited statistical knowledge** should be able to understand your analysis and the rational behind your conclusions. In particular, these presentations are meant for business owners on Yelp who may not have taken a formal course in statistics, but are interested in making data-driven business decisions. The exact grading rubric for the presentation is outlined below.

The first presentation will primarily focus on any preliminary exploratory (or some confirmatory) analysis you may have done. More broadly, the presentation is meant to inform key stakeholders or intermediaries (e.g. business owners, managers, or data scientists at a company) about the progress of your data analysis.

The second presentation will focus on your final analysis/conclusions from the data. This should be treated as a final meeting between you and all stakeholders (e.g. business owners, managers, VPs, data scientists) where you inform them about various conclusions you have drawn from the company’s data. If you would like, you can spend majority of the time demonstrating your Shiny app, especially if the app contains key information that you want to highlight to your stakeholders. Alternatively, you can summarize key, data-driven recommendations that could improve their businesses and discuss how you arrived at the recommendations.

Your group will prepare a **5 minute** in-class presentation of your data analysis; your presentation can be shorter than 5 minutes, but it cannot be longer than 5 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 5 minutes.

For each presentation, 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 days.

Four-Page Executive Summary:

Your group must submit a **four-page** executive summary of the data analysis, with at most **one additional page** for references & contributions. That is, all the statistical analysis, plots, tables, figures must fit in the four-page limit and the one-page reference section can only be usedfor references to other works & contributions (see below).

Your 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, modeling assumptions, etc.). Your analysis should be detailed enough that **any data scientist (or their manager) can** **understand** **your summary and** **replicate** key parts of your analysis. The exact grading rubric is outlined below.

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

All members of the group must contribute to the summary. Specifically, on the reference page, the group must clearly indicate **each member’s contribution** **to the summary, the presentation, the code, and the Shiny app**. For example, you can say that

1. HK (initials of your group member’s name) wrote/edited the diagnostic part of the summary, worked on slides (blank) to (blank). HK also created code related to data cleaning , revised/ maintained the code related to Figure 3, created code for visualization histograms in the Shiny app. HK is ultimately responsible for the data cleaning portion of the code and the histogram portion of the Shiny app.
2. BK (initials of your group member’s name) wrote/edited the model building and interpretation part of the summary, worked on slides (blank) to (blank). BK also created/edited/maintained code for model building (e.g. lm(), summary(lm()) in the Shiny app, edited code related to plotting, and is ultimately responsible for the model building portion of the code.

All of the summary must be typed in 12-inch Times New Roman or Sans Serif font, single-spaced, with 1-inch margins and (again) must include all relevant figures/tables, and equations. Note that 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 in a standard PDF viewer.

Github Repository

Your group must publish a Github repository and provide **a web link** to the repository for grading. 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 executive 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 start (i.e. reading in the data) to finish (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 science** student can read and understand your code. This is important for reproducibility and to track down potential bugs/errors in your analysis.

All members of the group must contribute to the codebase. \*\*\***Also, all students are** **required to use Github’s pull/push/commit** **functions regularly**\*\*\* 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 and (again) 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 demonstrate your analysis and **submit a link to your** **live/running Shiny app** for grading.

Shiny is an easy-to-use platform to turn your R analysis into web-based applications. 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. For more information about Shiny, visit: <https://shiny.rstudio.com/>.

Your application will be graded on the following criterion

1. whether it runs in real-time,
2. whether it is robust to erroneous inputs,
3. whether **it provides useful and insightful 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.

Some examples of “useful and insightful information” include:

1. Meaningful, graphical summaries (e.g. histograms, heatmaps, bar chart, density plot) of ratings of different businesses
2. Some key information about the business (e.g. name, location, open hours, type of restaurant, etc.) or users (e.g. type of user, mean number of comments/reviews, etc.)
3. Some key statistics (e.g. mean ratings of similar restaurants with SD, quartiles of ratings in the area, some useful scores about service/quality/wait time)
4. Some useful recommendations for businesses, say
   1. Opening for additional one hour at night can drive ratings up by, on average, 0.2 stars
   2. For restaurants that are similar, offering take-out options increased ratings by 0.1 stars
   3. Users found that the service at this business was above average thanks to these attributes (BLANK, BLANK, BLANK).
   4. BLANK % of users recommended Users recommended that businesses have these attributes (BLANK, BLANK, BLANK)

While you do not have to include exactly what’s stated above, we expect that the dashboard/app to show a rich, but concise array of data analysis.

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.

|  |
| --- |
| Presentations |
| 1. Clear, takeaway analytic insights and data-driven recommendations to businesses? 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 their analytic insights + recommendations 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 recommendations based on these insights. 3. Concise, useful and relevant plots/tables summarizing the data analysis 4. Concise and relevant summary of estimates and inference of relevant parameters. 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 any inferential quantities 6. Correct and interpretable model with an understanding of its strengths and weaknesses by checking model assumptions and using model diagnostics 7. Conclusion 8. Does it follow instructions concerning the page limit? |
| 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 slides 6. Contains the executive summary 7. Contains the Shiny app link and shiny app code base   Each student must make regular contributions to the Github repo. |
| Shiny Application |
| 1. Does it run in real time? 2. Is the application robust to user inputs? 3. Does it display **clear, useful**, and **actionable recommendations** to business owners? 4. Does it provide **visually pleasing graphics** and **data-driven analytics/dashboards** to gauge the “health” of the business? |

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 <https://conduct.students.wisc.edu/academic-misconduct/student-resources-for-academic-integrity/>