

STAT400, Fall 2021: **Group R-Project**

October 7, 2021

This document will outline the guidelines for the Group R-Project. Please read carefully and reach out to me if you have any questions.

1 Project Outline and Datasets

You can work with different datasets that can be found in Kaggle, for example you can analyse datasets for the 2019 Presidential Elections.

The suggested dataset is the COVID-19 dataset curated by John Hopkins University: JHU COVID-19 dataset.

The goal of this project would be to get experience in handling a current real-world dataset.

The scope of this project is broad, and here are some general guidelines:

1. Understand the dataset/datasets. Note that there are different datasets in the repository. You would want to understand the different variables that are involved, identify which ones are categorical/quantitative, and how the variables interact with each other.
2. You want to find appropriate visualizations for the data.
3. You want to come with different questions/hypothesis that you want to ask/test. Please keep all your methods within the scope of this course.
Note that this is often the most difficult step, you want to think about your questions very carefully. It can be a good idea to have multiple questions that are considered.
4. Use the data to derive answers and/or insights into the questions that you asked. Make sure to work within the scope of the tools learnt in this class.

2 Project Groups

I have already assigned you into groups of 6, based on your discussion section and teaching assistant. Your TA will mentor you on the project and you are required to meet the TA three times (more if you like) before the final submission, and at least one of those meetings have to be with the entire group present.

The TA's are aware about the project guidelines, and appropriate logistics and meeting times must be planned in advance with them.

Each group will have an assigned group leader. If you are interested in being a group leader please reach out to the concerned TA and copy me on the email.

3 Resources

Please note that you have multiple resources available.

1. Your teaching assistant will be your primary mentor. It would be very helpful if you reach out to your mentors with specific questions, cross check your ideas, and get feedback on your methods. Please make the best use of their experience and knowledge.
2. Feel free to talk to your GSS leaders about your project ideas. The GSS leaders are encouraged to pursue project directions in their sessions.
3. I will be available to answer questions, if you need any help, consider coming to office hours. I am happy to talk about project ideas before or after class.
4. This dataset is well studied, feel free to checkout any examples online. Please keep in mind that copy-pasting ideas is plagiarism. It is important (to the extent possible) that you provide relevant citations/credits for things you use from the internet.

4 Grading

This project will comprise of 10% of the total course grade and will be graded out of a total of 200 points.

There will be two required submissions of the project, grading criteria for it will be as follows:

1. **Preliminary Submission:**

You will need to submit a basic outline of your project idea. This submission will be graded out of 50 points.

You must meet your project mentor (discussion TA) prior to submitting your preliminary submission, making sure that you get appropriate feedback on your choice of project idea/ideas. Points will be deducted for submissions that have not been vetted by your mentor.

2. **Final Submission:** This will be graded out of 150 points. Final submission must involve the following four components, each uploaded separately as part of your single group submission. It is recommended that you divvy up completion tasks across group members.

- (a) **HTML/PDF-Rendered R-Notebook** This is your main submission. It must have four sections:

- i. **Introduction:** Introduce the data set, the source, with appropriate links. Provide brief description about what a reader can expect to find in the project. Take time to get the reader interested in your project, state why the questions you ask might be interesting/relevant.
- ii. **Analysis/Visualizations:** Provide relevant analysis and visualizations to in your pursuit to address questions raised in the introduction.

- iii. **Conclusions:** Draw relevant conclusions from your analysis and visuals. Be sure to identify the strengths and limitations of your methods.
- iv. **Future work:** Identify what more you could/would have done if you had more time and tools at your disposal.

Your document should not be more than 5 pages in length. You might need to pick and choose between various things your group has tried.

- (b) **Rmd of the Notebook** You need to submit the .Rmd file that rendered the notebook above. Make sure that it works on every group member's computer.
- (c) **Document of Ideas and Participation** This document should document (at some broad level) the participation of individual group members and the tasks they were responsible for). For example: meeting notes, suggested ideas, etc. Note that all members might not be participating equally, but they do need to have participated somehow.
- (d) **Project Video Presentation** A 120 second video presentation by at most two members of the group. Please make sure that you have a well written script and that there are not more than 5 slides in the presentation.
You can either (1) upload your video to Youtube and submit link as part of submission, or (2) submit a .mp4 video file as part of submission.

We will have presentations on the last day of classes, perhaps play some of the pre-recorded submissions, and have questions afterwards.

- 3. **Group Participation:** Each group has a group page on Canvas. It is recommended that you use that to participate. If you prefer, it is possible to create groups in Piazza too, let me know if that is something that you would prefer. Depending on your participation and feedback, I might request to meet with your group to get a sense of how you are progressing with the project.

5 Important Notes

The goal of this project is to **study a real-world dataset in a collaborative setting within the scope and abilities explored in this class**. This is **NOT** a competition between the different groups.

To this end please understand the following:

- 1. It is important every group member contributes to the project. The group leader must inform me if there are problems organizing the group.
- 2. The ideas and questions you seek to answer do not have to be complicated. Please make sure you are sticking to the scope of this class.
- 3. Even though there is a component of using R, this is **NOT** just a coding project. Please leverage the diverse individual skills of all your group members in this collaborative setup.
- 4. Everyone in the group should feel comfortable voicing their ideas. Please make sure that you create an inclusive and non-threatening environment when having group discussions.

5. This is not a competition, so don't spend too much time judging your peers ideas, there is room for many questions and different answers. This is not to say that you cannot have disagreements, just make sure you keep them civil.
6. Any form of misbehaviour will not be tolerated. Please inform me right away if there are any issues with members of your group.