DS 301: Spring 2021 Final Project

Key Due Dates:

- **Finalize Team Members** due Monday April 19 by 11:59 pm.
- **Presentation of Final Project** due May 4 by 11:59 pm (pre-recorded and submitted to Canvas). Bonus points for presenting during lecture. See details below.
- **Final Report** due Tuesday, May 4 by 11:59 pm (submitted to Canvas).
- **Participation** due Friday, May 7 by 11:59 pm.

Your final project will consist of analyzing a data set using the techniques we have learned this semester. You are expected to work in teams of 5. Together, you will need to submit a presentation and a report. There are a number of data sets available to you that you may pick from. Once your team has picked a data set, identify **two** questions of interest you hope to address from the data. The bulk of your final project will consist of your analysis of the these two questions using methodology we have covered. **Read the following instructions carefully** to have a clear idea of what's expected.

1 Finalize Team Members

Please form teams of 5. As a team, you'll need to decide:

1. Which data set you want to analyze.
2. A team name.
3. Anticipated responsibilities of each team member (in other words, how do you plan to divide the work?).

Your team should report items (1) and (2) to the Google spread sheet (see link the assignment description). Please also include your team member's names. A data set can only be claimed by 1 team. You are welcome to select your own data set as well. Item (3) will be reported as part of HW 9.

2 Presentation of Final Project (50 points)

Each team will present their project results. The presentation should be roughly 15 minutes. There are two possible options for this component:

1. Your team may record a presentation (using Zoom or any other software) and upload this to Canvas by the due date May 4.

2. Your team may present your project results during regularly scheduled lecture time on Friday April 30. If your team chooses this option, this will result in an automatic 5 points of extra credit (out of 50).

Presentations will be graded on: organization, delivery, content, and timing. Each is worth 25% of your presentation score. To get full credit, **every team member needs to introduce themself and speak in the presentation**.

The presentation should inform the audience on

1. The background of the data set.
2. What are the questions of interest?
3. A concise, big picture description of the strategy used to answer the questions of interest.
4. A more detailed discussion and justification of the methods used for analysis.
5. A critical examination of your findings and conclusions.

3 Final Report (100 points)

Your team's final report should be divided into the following components. Your team will be graded on each component. Page limit for the report (not including references and figures) is 10 pages. **Upload your report and a copy of the `R` code used in your analysis to Canvas by the deadline.**

1. **Introduction:** The introduction should include an overview of your data set. Here you should specify:
   - The two questions of interest you hope to analyze from the data.
   - Your team is expected to carry out exploratory data analysis on the data set. Do not just provide summary statistics/plots of every variable in the data set. That is not meaningful. Instead, the exploratory data analysis should help shed insight on the two questions of interest you've specified. Highlight those key results. Your team should demonstrate how your EDA helps you build a better model (instead of just blindly trying different methods).
2. **Methodology:** This section should describe the methods used to address your questions of interest. Your team should attempt a variety of methods to answer the two questions of interest. This needs to include a method that has not been covered in this course. In order to receive full credit, the methodology should be appropriately selected for the task and implemented carefully. This *may* mean considering:
   - Tuning parameters (for example, λ or k)
   - Interaction terms
   - Polynomial or transformed predictors
   - Model selection

- Training and testing your model
- Diagnostics
- ROC/AUC
- ....

Additionally, your team should implement a method not covered in this course to address at least one of the questions of interest. There are many models that are extensions of the models we've discussed (for example, regression splines, elastic net, principal components regression, support vector machines, etc etc) which build on the same fundamentals we have covered. Almost all of these have `R` packages for easy implementation and tons of tutorials. You may also choose a method that is covered in our textbook but that we did not discuss in class.

3. **Results:** For all your methods, report your results thoughtfully (no raw `R` output). As we've discussed, modeling is an iterative process. You may not want to show results for every single step, but you should summarize the thought-process and justification for how you obtained your final results.
4. **Discussion:** In this section, your team should directly answer the two questions of interest. Your team should demonstrate that the insights are well-supported by strong evidence in the data and models. Any unexpected results should be thoroughly investigated and multiple explanations should be proposed and discussed. Problematic areas of the data and/or approaches need to be identified and their impact on the results should be addressed. A critical examination of the strengths and weaknesses of the methods used should be thoroughly discussed here.
5. **Conclusion:** Wrap up your report by providing a clear summary of the project and the insights discovered by the analysis.
6. **Statement of team member contributions.** Each team member needs to write 1 -2 sentences describing their contribution to the project. Your team should also collectively decide how much each person contributed to the project and report this as a percentage. For example, if each team member feels they have contributed equally to the project, each team member would report 20%.

4 Participation

You can earn 1 participation point by watching another team's presentation video and posting a comment describing:

1. One thing you liked about their project/presentation.
2. One thing that could be improved about their project/presentation.

This activity will be in place of Discussion 3. You can earn up to 6 points (assuming there are 6 other teams) here (if you have not already maxed out your participation points).