DS502/MA543: Statistical Methods for Data Science
Final Group Project
Fall Semester 2021

The idea of this project is that you are to pick a real data set for which you believe there are interesting questions to answer. You will then try out at least **two** statistical learning approaches that we have covered in this course to try to find the best way to answer these questions. An important aspect of your results will be **appropriate error assessments** for the results you provide. The project can include the collaboration of 3-4 students per project.

Deliverables

This project includes four deliverables:

1. A proposal idea 1-2 paragraphs **(Due Sept 28)**
2. A proposal for the project- 1-2 pages long (**Due Tuesday, Nov. 2**)
    a. Members' names
    b. Description of the problem
    c. Description of the dataset (dimensions, names of variables with their description)
    d. Regression or classification?
    e. The methods you plan to try.
    f. The error metrics you plan to use and the algorithms for assessing them.
    g. Comments and/ or concerns?

3. Slides for a 15 minute presentation (**Due one night before project presentation**).
    a. Description of the data and the questions that you attempted to answer.
    b. Review of the approaches that you tried or thought about trying. It is interesting and useful to discuss both successes and failures!
    c. Summary of the final approach you thought worked best and why you chose that approach.
    d. Summary of the results.
    e. Conclusions.

    Points will be allocated for the explanation of the question of interest, the descriptions of approach you used, the reasons you chose your final approach, and the conclusions you were able to draw, both positive and negative.

4. A report to be sent via email. (**Due Dec 7**)

    The report will contain a summary of the material covered in the presentation (between 3 and 5 pages). Submit the R code you developed by email to Prof. Emdad (femdad@wpi.edu).

Data Repositories that you might consider

1. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
2. Statlib datasets: http://lib.stat.cmu.edu/
3. Kaggel: www.kaggle.com
4. Open Gov. Data: www.data.gov, www.data.gov.uk, www.data.gov.fr, http://opengovernmentdata.org/data/catalogues/