**CS 59000/ STAT 46700 Topics in Data Science Spring 2025**

**Project Outline**

The purpose of this project is to apply what you've learned about both supervised and unsupervised learning and related topics to a dataset of your interest.

1. Begin searching the net and other sources for a dataset that captivates your interest

You may find the links provided in Brightspace useful (If you have trouble finding data sources, ask me for assistance!).

1. Dataset from kaggle.com is not accepted without instructor’s permission.
2. Data should be related to a real study. Published article is acceptable but you should carry all statistical analysis of the data.
3. Large date with multiple variables (both qualitative and quantitative) is preferred
4. You should use R to perform all necessary tasks.
5. Include output analysis as well as a written description of statistics and interpretation of your results.
6. Provide graphical presentation whenever applicable

You'll want to begin your project with a general research question and identify question and provide answers. You should have an introduction to your project which includes your research question (including any reasons for being interested in your topic) and a list of the variables and what they represent etc.

Finally, conclusion and discussion are required. This should consist of a summary of what you've learned about your research question along with statistical arguments supporting your conclusions. Furthermore, critique your own methods and provide suggestions for improving your analysis and future work (wish list!!) as we move forward.

Be creative and utilize your critical thinking skills. The key is to choose a dataset that interests you

You can work in a group of two students.

This should be a fun project!!

Looking forward to reading your project report!!