ERG 2050 Introduction to data analytics
Project Assignment

- < = 4 people a group
- You are to pick one of the eight problems provided at the end of this document. Among the eight problems, two of them (4 and 8) are more complex. You will have bonus if you choose those two.
- You can try out all the different statistical learning approaches that we have learned in this course to find the best way to solve these problems.
- Assessment: 10-minutes Presentation + Q&A and final paper report.
- The presentation time is on the week of Dec. 11th. Details to be announced later.
- The report of your project should be submitted on Moodle on or before Dec. 22rd 11:59pm. The codes should be also attached.

For the presentation:

The presentation should include

1. Description of the data and the question/s that you are interested in answering.
2. Review of some of the approaches that you tried or thought about trying.
3. Summary of the final approach you used and why you chose that approach.
4. Summary of the results.
5. Conclusions.

For using the data:

Files in folder

-The 'train' file contains the training data. Please use training data to build your model and do the cross validation.

-The 'test' file contains testing data. This data is used to test the performance of your final model. Don't use it until your model is well trained and fixed.

Problems

1. Letter Image Recognition Data Set (Classification)

2. Gene expression cancer RNA-Seq Data Set (Classification/Clustering)

3. Crowdsourced Mapping Data Set (Classification)

4. MoCap Hand Postures Data Set (Classification)

5. Forest Fire Data set (Regression)

6. Property's Sale Price Data Set (Regression)

7. Bikes Rented Data Set (Regression)

8. Stock portfolio performance Data Set (Regression)

Datasets can be downloaded through this link:
http://10.20.6.103/final_project