Booth Big Data: Final Project

Two *related* big data tasks:

Understanding and interpretation in high-dimensions

- Tell us a story that builds stylized facts about the data.
- Tools: data visualization, causal inference, factor models, clustering, network graphs...

Build and evaluate a prediction model.

- Raw prediction: build and evaluate a forecasting machine.
- Tools: Linear/logistic Lasso regression, causal inference, PCR, trees...

Describe analysis goals and use them as motivation.

You should bring your own data

Supply your own dataset and develop analysis goals.

- Data must be rich enough for both explore/predict tasks.
- You should be able to use many tools from class (not all).
- Make sure the data and your goals are compatible.

Your project score will have a data multiplier corresponding to level of data-difficulty (from cleaning to conceptualization).

Think of the midterm data as the baseline of one.
(i.e., you don't want anything much more simple).

See piazza for a starter list of data sources.

Group and individual projects

- As always, you can work in a group of up to 4,
- Everybody in the group receives the same project score.
- If you are 1 or 2, I don't expect the work of 4 people.

Presentation and format

- Make your analysis and conclusions clear and concise.
- Include enough R output/code to show what you've done.
- See homework and midterm guidance on presentation.

You should build a professional-grade report for this project. Make it as you would for a very statistically savvy client who was not in our classroom this quarter.

The project is due by Wednesday March 16 at 8:30am.

Early submissions are absolutely welcome.

Submit through the form link that will be posted on Piazza.

Keep your code handy in case we want to see it.