540 Fall 2021
Statistics Case Studies
Overview

Fan Li

Department of Statistical Science

Duke University

Class information

- Lecture times: Thursday **9:00-11:15** am **IN PERSON** (lead by Fan Li)
- Lab times: Tuesday 3:30-4:45 (lead by Bo Liu)
- Location: Perkins LINK 087 (Classroom 3)
- Lecturer/coordinator: Fan Li fli@duke.edu
- Teaching assistant: Bo Liu bo.liu1997@duke.edu
- Office hours (**All Zoom**)
  - Fan Li: Monday, Wednesday 8-9pm
  - Bo Liu: Thursday 8-9pm
- Course website:
  - Github: https://github.com/STA540-21Fall/Teaching
  - Duke Sakai (for some lecture notes and grading)

Course structure: Project-based learning

- Objectives: gain real world experience on end-to-end analysis and and hone statistical thinking
    - problem setup, acquiring data, cleaning data, analysis, visualization, communicating the results
- 4-5 group projects and 1 individual project
- Group projects
    - Each project lasts 2-3 weeks
    - Students are randomly divided into four 3-people groups, collaborate on Github
    - Oral presentations in class
    - Group written reports
- Individual project: self-selected topic
- Lectures: Descriptions of case studies data and highlights of critical issues to consider in analysis of data. Occasionally lectures on specific topics, e.g. causal inference, survival analysis, will be given. Majority of the class time will be devoted to oral presentation and critique.
- Labs: usually function as an office hour. Occasionally, TA will present topics commonly emerging from homework, particularly on computing and programming issues.

Topics to be touched upon

- Unsupervised learning: descriptive statistics of high-dimensional big data, classification
- Supervised learning, predictive modeling
- Hierarchical models
- Causal inference
- Visualization
- Communication: oral and written
- …

Programming

- Github: collaborative working platform

https://github.com/STA540-21Fall/Teaching

- R or Python, whichever one choose to use
- Python is strongly encouraged because of the popularity in industry
- Visualization: R shiny app or Tableau or anything suitable
- Project report: uploaded to Github, need to be completely reproducible; preferable format is html generated from R markdown

Grading

- Group case studies 65%; individual case study 25%; class participation 10%
- Late work policy for case study reports:
  - late, but within 24 hours of due date/time: -30%
  - any later: no credit
- You must complete the individual case study and be in class to present it in order to pass this course
- Regrade requests must be made within three days of when a report is returned. These will be honored if points were tallied incorrectly, or if you feel part of your report is correct, but it was marked wrong.

Grading criteria/learning objectives

- Originality and relevance of the topic
- Visualization
- Quality of analysis:
  - appropriateness of methods
  - creativity
  - how well the analysis supports the conclusions
- Presentation and communication
  - Clarity
  - Professionalism
  - Presence
  - Writing

Acknowledgements

- I borrowed lots of ideas and material from the Columbia Stat’s course on Applied Data Science, instructed by Professor Tian Zheng
- Amy Herring, for sharing the information on Stat 440
- Mine Çetinkaya-Rundel
- Funda Gunes