Stat 470/670: Exploratory Data Analysis

Meeting time: Tuesdays and Thursdays, 1:15-2:30 Meeting location: M C141
Website: jfukuyama.github.io/teaching/stat670

Instructor: Prof. Julia Fukuyama jfukuyam at iu dot edu
Office hours: Mondays 4-5 (on Zoom, link on canvas), Fridays 2-3 (in person)

Associate Instructor: Mr. Paul Hunt pchunt at iu dot edu
Office hours: TBA

Course Overview

Graphical and modeling techniques for exploring data, with an emphasis on visualization, interpretation, and clear communication of findings. Use of modern software tools for data manipulation and visualization. Connections to traditional statistical methods.

Textbooks

We will be drawing heavily on Cleveland's *Visualizing Data* and Hadley Wickham's *ggplot2: Elegant Graphics for Data Analysis*. Both of these are available online through the IU library.

Also useful will be *R for Data Science* by Wickham and Grolemund, available online at http://r4ds.had.co.nz.

Readings and notes for topics not covered in the textbooks will be posted to the course website and to canvas.

Class Structure

Classes will be a combination of lecture and tool demonstration. It will generally be helpful for you to have an R session open to follow along wth the code. Slides or notes, with R code, will be posted to the class website before each lecture.

Assessment

Grades will be assigned based on:

- Homeworks, 30% of the grade.
- Mini projects, two, together worth 30% of the grade. These will involve more substantial data analysis and a writeup. Done in groups of up to three.
- Final project, 40% of the grade. Done in groups of up to three.

There will be no final exam; the last responsibility for the course will be the report for the final project due on the last day of class.

All the assignments will be graded on how well the material is presented in addition to accuracy. This means there should be no extraneous material, plots should be readable, and text and figures should be formatted nicely.

Topics

There are two categories of topics: *what* to do and *how* to do it. In the *what* to do category, we will cover:

- Univariate data: measures of center and spread, transformations, visualization.
- Bivariate data: Simple regression, curve fitting,
- Trivariate/Hypervariate data: Multiple regression, model selection, principal components.
- Binary responses: Logistic regression, residuals.
- Categorical data: Contingency tables, correspondence analysis.
- Distance data: Multi-dimensional scaling, non-linear dimensionality reduction.
- Graph data: Descriptive statistics, spectral methods, visualization.
- Dangers of EDA and remedies: Multiple comparisons, data splitting, cross validation.
- Other topics according to time and interest.

In the *how* to do it category, we will cover

- ggplot2 for plotting.
- tidy-verse methods for data wrangling.

By the end of the course, you should feel comfortable using R to visualize and model many kinds of data. Given a dateset, you should be able to visualize the data, generate hypotheses about the relationships among the variables, investigate those hypotheses, and communicate your results.

Course Policies

Late Policy

You have four "free" late days to use for assignments and mini projects. After that, late work will be penalized at 10% per 24 hours. Special accommodations may be granted if you ask very early.

Academic Integrity

You are expected to abide by the guidelines of the IU Code of Student Rights, Responsibilities, and Conduct (http://studentcode.iu.edu/responsibilities/academic-misconduct.html) regarding cheating and plagiarism. Any ideas or materials taken from another source must be fully acknowledged and cited.

Disability Accommodation

Please contact me if you require assistance or academic accommodations for a disability. You should establish your eligibility for disability support services through the Office of Disability Services for Students in Wells Library W302, 812-855-7578.