Stat 470/670: Exploratory Data Analysis

Meeting time: Tuesdays and Thursdays, 4-5:15pm
Meeting location: Fine Arts 102
Website: jfukuyama.github.io/teaching/stat670

Instructor: Prof. Julia Fukuyama
jfukuyam at iu dot edu
Office hours: Wednesdays 2:30-4:30pm
Office: Informatics East, Room 201

Associate Instructor: Ms. Fatma Palak
fparlak at iu dot edu
Office hours: Thursdays 9-11am
Office: Informatics East, Room 103

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

The primary textbook for the course will be Cleveland's *Visualizing Data*.

Also useful will be Hadley Wickham's *ggplot2: Elegant Graphics for Data Analysis*, available for free through the IU library as an ebook, and *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 bring a laptop with R installed so you can follow along. Slides or notes, with R code, will be posted to the class website the day before each lecture.

Assessment

Grades will be assigned based on:

- Homeworks, 25% of the grade. These will be short, assigned most weeks, and due Fridays at 5pm. They should be formatted as a pdf and uploaded to canvas. Done individually.

- Mini projects, two, together worth 20% of the grade. These will involve more substantial data analysis and a writeup. Done in groups of up to three.
- In-class midterm, November 15, 15% of the grade. Note that this is the Thursday before Thanksgiving break.
- Final project and presentation, 40% of the grade. The components are a proposal, due November 9, final presentations in the last week of classes, and a final report due December 7. 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 December 7.

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

Late homework and mini projects will be penalized at 10% per 24 hours, and no assignments will be accepted more than three days late (e.g. a homework due on Friday at 5pm will not be accepted later than the following Monday at 5pm). No more than three late days can be used over the course of the semester. 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.