Course Syllabus

STAT 230/530 – Introductory Data Analysis
Spring 2017 Syllabus

Instructors:
Susan Wang (xiaofei.wang@yale.edu) 24 Hillhouse, Rm. 206
Joe Chang (joseph.chang@yale.edu) 24 Hillhouse, Rm. 211

Teaching Fellows:
Tal Sarig (tal.sarig@yale.edu)
Jonathan Harris (jonathan.harris@yale.edu)

Tutor: Addison Hu (addison.hu@yale.edu)

Office hours/TA session

- Mondays: **4-6p**, 17 Hillhouse, Rm. 07 Library Classroom (Jonathan)
- Tuesdays: **1:30-3:30p**, 24 Hillhouse, Rm. 211 (Joe)
- Wednesdays: **2:30-4:30p** Loc TBD (Tal)
- Thursdays: **10:15-12p**, 24 Hillhouse Rm. 206 (Susan); **3-5p** 24 Hillhouse Basement B06 (Addison)
- *TA Session*: Wednesdays **6-7:30p** Loc TBD

Prerequisites:

- Statistics: You should have had a course in introductory statistics, or AP Statistics. If you have not had a previous course in Statistics, you would be much better off taking Stat 100/500 this semester!
- Computer programming: Prior experience is not required, although of course if you have some it could be helpful. While this is not a "programming class," you may feel that you are spending a lot of time learning and doing programming in this class. So, although it is fine if you have no prior exposure to programming, in order to be happy with this class, you should at least like the idea of learning about programming and be willing to put in some effort to develop some programming skills.

Computing:
In this course we will examine various real-world data problems using R, a popular, open-source (and free!) statistical software. Each class will consist of a mix of lecture and R examples. The best way to follow along with these examples is to bring your laptop to class.

Topics:

- Intro to R and R markdown: installation, compiling scripts, working with vectors, functions, getting help, reading datasets
- Data cleaning
- Using APIs from R (e.g. twitteR, blsAPI, quandl, ZillowR)
- Exploratory data analysis – basic plotting, numeric summaries
- 2-sample inference type problems, parametric + simulation-based
- Simulation – with writing loops, functions, bootstrap, permutation tests
- Linear regression – inference, diagnostics, transformations, indicators, interactions/higher ordered terms
- model selection and cross-validation
- ANOVA
- Data scraping
- Advanced plotting (e.g. lattice, ggplot2, choroplethr)
- Selected advanced methods: logistic regression, k-means clustering, PCA, multilevel or mixed-effects models

Textbook and references:
There is no required textbook to buy. References, links about R and data, etc. will be posted over the course of the semester. As a start, two useful textbooks for reference on R that you can download for free from Orbis are

- *Introductory Statistics with R* (2nd ed, 2008) by Dalgaard
- *Statistical Analysis and Data Display: An Intermediate Course with Examples in R* (2nd ed, 2015) by Heiberger and Holland

Grading:
Our current plan (which may be revised over time) is for grades to be based on a weighted average of four components: homework (45%), quizzes (20%), project (30%), and participation (5%). More details about these components appear below.

Homework:
We will be assigning problem sets regularly, about one a week. Most of these assignments will not contain "pencil-and-paper" type problems (like mathematical exercises), but instead will involve data analysis in R. An important part of data analysis is communicating the results clearly and concisely. As such, some homework problems will require you, in full sentences, to explain your approach or interpret a model in context.

Your lowest homework grade will be dropped, so you can encounter unexpected complications and miss one assignment without affecting your course grade. Late homeworks will be accepted only with a Dean's Excuse.

Quizzes:
At this point we anticipate having 2 in-classes quizzes over the course of the semester.

Project:
The project will consist of an extended data analysis that integrates many of the ideas we cover in the semester. More details will be provided approximately around spring break.

Participation:
We will use the online discussion forum Piazza (Links to an external site.)Links to an external site. for questions, answers, and discussion about all aspects of the class. This is intended to be an easy 5% just to encourage you to participate early and often!

Expectations:
Forming study groups to discuss materials and homework is strongly encouraged. However, any code or words you write should be your own.

Course Summary:

Date	Details
Fri Jan 27, 2017	Homework 1 due by 11:59pm
Fri Feb 3, 2017	Homework 2 due by 11:59am
Fri Feb 10, 2017	Homework 3 due by 11:59am
Fri Feb 17, 2017	Homework 4 due by 11:59am
Fri Feb 24, 2017	Homework 5 due by 12pm
Fri Mar 3, 2017	Homework 6 due by 12pm
Tue Mar 7, 2017	Team Roster due by 11:59pm
Fri Mar 10, 2017	Homework 7 due by 12pm
Fri Mar 31, 2017	Project Proposal due by 11:59pm
Fri Apr 7, 2017	Homework 8 due by 12pm
Mon Apr 10, 2017	Dataset due by 11:59pm
Fri Apr 14, 2017	Homework 9 due by 12pm
Fri Apr 21, 2017	Homework 10 due by 12pm Outline due by 11:59pm
Fri May 5, 2017	Final Report due by 12pm Team Peer Evaluations due by 12pm Quiz 1 Quiz 2