SYLLABUS AND CLASS DESCRPTION – DATA 101

Grading: 75% weekly challenges + final project
10% midterm
15% final

Prerequisites: Placement into Intermediate Algebra or higher, or completion of Math 025. No programing experience required.

Website: sakai.rutgers.edu

Overview:

Data is everywhere, and data literacy – understanding how to collect, visualize, analyze and interpret data – is an essential skill for success now and in the future

This class aims to provide you with a basic set of tools for data literacy as well as general view of the impact of data on society and elements of common sense data analysis and reasoning. Our high level goals specifically include learning about:

a. Drawing meaningful conclusions from data –discovering patterns in data, simple prediction using simple R. This is hands on experience, basic skills to analyze data and visualize your findings
b. Common sense probability – Bayesian common sense reasoning
c. Backing up your arguments with data: for example: is global warming caused by human activity? Do concealed guns provide more or less security?
d. Data and its impact on society – what are tradeoffs between privacy and convenience& security which we as society are willing to accept?

A significant piece of the class will be learning foundations of R. R is a statistical software environment and programming language that we'll use to analyze and visualize datasets. R is extremely flexible and powerful, and is used by many professional data scientists and statisticians. Learning simple R will take some work; however, if you're able to master the basics covered in this class, you'll gain a concrete, marketable skill that may very well be extremely useful in your academic and professional life. On the statistical side, we'll cover basic topics from statistics and probability that are required to argue persuasively using data (a list of some of the topics to be covered can be found below).

Weekly challenges: Weekly challenges will be announced in the first class of each week and will involve analyzing (possibly real or simulated) data. *Each challenge must be submitted by end of the week on* through Sakai, and each submission should contain

(i) a .ppt file for presentation, describing your results and (ii) a .R file (R code), showing how your results were obtained using R in case it is a coding assignment.
It is intensely paced class – driven by data puzzles and hands on problem solving, research on the web as well as reading assignments (for data and society part)

***The court of data**:* Each 3-hour class will have second period devoted to the court of data. During court of data a number of students will be selected to present in front of the class. These presentations will be discussed and JUDGED by Instructors and fellow students (using clicker). Students selected for the court of data won't be announced until the class, so everyone should be ready to present. You will earn extra points on your assignment by performing well in the court of data. Throughout the semester, everyone will have any opportunity to present in the court of data; however, interesting challenges submissions will be preferred.

Quizzes: Midterm and final quizzes will contain mostly basic questions on R and statistical techniques used throughout the class; they should be easy, if you're an active participant in your group and the weekly challenges.

Final project:: Group project (groups of 2-3 students) to be defined. Best projects will be presented at the end of the semester and will compete for the best project award.

Recitations: Recitations are important! This is where the most of the coding instructions and coding practice will take place. The main part of each recitation will involve going over a model solution for the current week's challenge (usually for a different dataset than the one assigned for the challenge). This will provide a very useful hint at what is expected, or what one possible approach is to your challenge solution.

<u>Textbooks and references that you might find useful:</u> None are required, all are available free online

- Statistics references
    - OpenIntro statistics: https://www.openintro.org/stat/textbook.php
        - Introductory statistics textbook (somewhat standard, but well done)
    - Various online courses, available through Coursera (https://www.coursera.org/). For example, see:
        - Statistics: Making sense of data (from University of Toronto)
        - Statistics one (from Princeton University)
- R references
    - Download R: http://www.r-project.org/, http://www.rstudio.com/
    - Learn R online (interactive): http://tryr.codeschool.com/
    - R at OpenIntro: https://www.openintro.org/stat/labs.php
    - Other resources: http://cran.r-project.org/doc/manuals/R-intro.pdf, http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Data and Society references -- *"Dragnet Nation: A quest for privacy, security and freedom in a world of relentless surveillance*", Julia Angwin 2015, "Data and Goliath: The hidden battles to collect your data and control your world", Bruce Schneier, 2014
"*Thinking, Fast and Slow*", Daniel Kahneman 2013 etc

Approximate Syllabus

Each weekly class: First 80 minutes: concepts, Second 80 minutes – court of data

Class 1-2: Introduction to Data Science and objective of the Data 101 class.. Review of basic plots. First challenge – plotting real data sets from the web.

Class 3-4: Data set transformations in R. Court of Data : Find and Plot interesting data set of your choice. Discussion in class. Next data challenge introduction.

Class 5-6: Comparing means – hypothesis testing. Court of data. Next data challenge introduction.

Class 7-8: p-values, Permutation testing, Court of Data: Next data challenge introduction

Class 9-10: Multiple comparison problem. Court of Data: Next data challenge introduction

Class 11-12: The Art of Advanced Plots. + Court of Data: Back it up or Reject

Class 13-14: Midterm + Caveats and Pitfalls of probability, Bayesian reasoning in every day life. (Kahnneman et al) + Court of Data: Next data challenge introduction

.

Class 15-16: Decision Trees, Linear Models, Prediction, Classification, Crossvalidation. Prediction Cup Launch.

Class 17-18: Pitfalls of prediction (Kahneman, Taleb, Nate Silver), Court of Data – Prediction Challenge

Class 19-20: Privacy and Personalization, How data about you is collected on the web – cookies, plugins, toolbars, GPS tracking (Angwin, Schneier) : Court of Data – Prediction Challenge

Class 21-22: Data cleaning and Big data performance challenges: Court of Data – Prediction Challenge

Class 23-24: Best final projects presentation and class discussion