I. Course Description

Statistical computing and data visualization are increasingly important and necessary aspects of a data analyst job. Whether they are dealing with small, big, structured or unstructured data, data analysts are expected to be able to access data from various sources, apply the latest statistical methodologies, and communicate their findings to others in novel visualizations.

In this course, students will learn how to make sense of data, and not the specifics of modeling. The course starts with statistical computing, and students will gain experience with a programming language called R (http://cran.r-project.org), a GNU-license statistical package. They will learn the practice of data cleaning, reshaping of data, basic tabulations, and aggregations in order to be able to produce high quality visualizations. Then, the course proceeds with graphics that are critical elements of modern data analysis and presentation. From initial exploration of a data set to the final presentation of results to the end user, statistical graphics play a very vital role in shaping our understanding of our data. Through proper use of graphics, we can make critical discoveries, and communicate them clearly. Conversely, poor use or misuse of graphics can seriously mislead (by accident or design).

II. Learning Objectives

In this course, the students will gain explicit experience with programming language concepts such as variables, assignments, flow control, functions, parameters, data structures, input and output, error handling, debugging, and so on. In addition, students will learn how to write a computer program in R to create a graph from scratch and manipulate its different attributes (axis, title, orientation, color, etc.), identify appropriate data visualization techniques given particular requirements imposed by the data, analyze and criticize examples of visualizations from newspapers, scientific papers, business reports, and journals.

III. Prerequisites

I expect basic “operational” knowledge from an introductory stats course such as GSBA-524. “Operational” means that you still recall sufficient details from regression, ANOVA, hypothesis tests, etc. Prior R knowledge is not required, but you will rather learn the basics of R in this course.
IV. Software

This is a hand-on course and it is computationally intensive. We will primarily be using R (http://cran.r-project.org), a GNU-license statistical package. We will also be using GGobie, Mondrian for high dimensional data and interactive graphics. All tools are available on the Web for free downloa

There are many reasons for us to focus on R:

- **The cost:** While commercial distributions exist, open-source R is free.
- **The rich features:** R has an estimated user community of 2 million, which includes thousands of contributors from different domains expanding the language’s capabilities through new libraries.
- **The quality:** R libraries are enhanced by domain experts and field-tested by the large user base including other experts with real datasets in real analysis scenarios.
- **The learning resources:** Thanks to the active user community, plenty of tutorials and sample code are readily available.

Rstudio is a recommended interface for the R software. It is also free, and it runs on Windows, Mac, and Linux operating systems. http://www.rstudio.org

Students are expected to bring their laptops to class during all class sessions.

V. Methods

This class is taught through a combination of lectures, computer labs, hands on computing tasks in homework, and group project and presentations. The projects are key because they integrate multiple computational topics in the content of a modern data problem. Students gain hand on experience with statistical concepts flowing from contextual problem solving with data, and they make their own discoveries by posing and answering questions rather than solely fitting models or using “this week’s lecture’s methodology” as a computing exercise.

VI. Suggested Books

Adler, J. (2010), R in a Nutshell, O’Reilly.


VII. Structure

Each topic in the course is motivated by a data problem. Some of the data sets we will use are:

• Facebook data
• Twitter data
• Behavioral risk survey data
• Los Angeles Collision data
• Fuel economy of US cars
• Characteristics and prices of 50,000 diamonds
• Mortality in Mexico
• US baby names from 1880 to 2008

We’ll also use some other interesting datasets for two projects, and you’ll have the opportunity to clean and compile your own data for the final project.

VIII. Assessment

• Homework
  To do well in this course you will need to spend 4-5 hours a week (outside of class!), and the homework are designed to encourage you do that. For each homework assignment, you will need to revise the week’s work, as well synthesize some new information, from the help pages or the web.

• Package Presentation
  For the presentation, you will need to find an interesting R package that we did not cover in class, and present it to your peers with examples and handouts.

• Midterms
  There will be two midterms in this class.

• Final project
  For your final project, you’ll be expected to find your own dataset. As well as writing a report, you’ll present at a formal poster session.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>% of Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>35%</td>
</tr>
<tr>
<td>Presentation</td>
<td>15%</td>
</tr>
<tr>
<td>Midterms</td>
<td>50%</td>
</tr>
</tbody>
</table>

IX. Students with Disabilities

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in STU 301 and is open 8:30 am - 5:00 pm, Monday through Friday. The phone number for DSP is 213 740-0776.
X. Topics:
- Data Science Tools: R, R Markdown, Rstudio, Command Line, Git, and Github
- Principles of good data display
- Exploratory data analysis: one variable, two variables, multivariate, time series, and spatial data
- Cleaning Data: dplyr, reshape2, tidyr
- Visualizing data: ggplot2, ggvis, and D3
- Interactive graphics: ggvis
- Spatial graphics: ggmap
- Time series: lubridate
- Web computing: shiny
- Web scraping: xml
- Regular Expressions and Messy data: stringr
- Other topics: Data Analysis Using Python

Statement for Students with Disabilities

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to your TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776. For more information visit www.usc.edu/disability.

Statement on Academic Integrity

USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one’s own academic work from misuse by others as well as to avoid using another’s work as one’s own. All students are expected to understand and abide by these principles. SCampus, the Student Guidebook, (www.usc.edu/scampus or http://scampus.usc.edu) contains the University Student Conduct Code (see University Governance, Section 11.00), while the recommended sanctions are located in Appendix A.

Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/. Failure to adhere to the academic conduct standards set forth by these guidelines and our programs will not be tolerated by the USC Marshall community and can lead to dismissal.

Emergency Preparedness/Course Continuity

In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.

Please activate your course in Blackboard with access to the course syllabus. Whether or not you use Blackboard regularly, these preparations will be crucial in an emergency. USC's Blackboard learning management system and support information is available at blackboard.usc.edu.