SI 618 Syllabus - Winter 2016

SI 618 – Exploratory Data Analysis

Wednesday 5:30pm - 8:30pm, NQ 2255

Instructor: Yuhang Wang (yuhangw@umich.edu)

Grader: Shao-Chi Wang (shaochi@umich.edu)

Instructor Office Hours: NQ 1243, Monday 6pm - 8pm

If you have questions about course material, homework, lab, or projects, please feel free to come and talk with me during my office hours. You can also contact me via email: please put "**618**" in the subject line so I can be sure to attend to it. Please note that I may not be available on email over the weekend.

Note: Some syllabus details are subject to change.

Description:

SI 618 aims to help students get started with their own data acquisition and exploratory analysis. Exploratory data analysis is crucial to evaluating and designing solutions and applications, as well as understanding information needs and use. Students in this course, who will have just completed SI 601: Data Manipulation, will learn techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data. Students will be able to make sense of and see patterns in otherwise intractable quantities of data.

More specifically, students will learn how to conduct and document an exploratory data analysis. To that end, the skills students will learn include the following:

- Converting messy data into a form that can be analyzed using R.
- Connect a database to R to simplify repeated analysis of changing data.
- Compute and visualize summary statistics of datasets.
- Master the specification of graphical displays using the 'grammar of graphics' via ggplot2.
- Combine the use of graphical aesthetics with data manipulation to visualize relationships between variables.
- Use subscripting to select subsets of data to analyze.
- Use factors to analyze categorical data.
- Produce polished information graphics for publication.

Prerequisites:

SI 601, or permission of instructor.

Texts:

Required:

Hadley Wickham, ggplot2: Elegant graphics for data analysis, Springer (2009)
http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-98140-6/contents/

Recommended:

Phil Spector, Data Manipulation with R, Springer (2008)

http://www.springer.com/statistics/computanional+statistics/book/978-0-387-74730-9

Leland Wilkinson, The Grammar of Graphics, Springer (2005)
http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-24544-7/contents/

Wes McKinney (2012). Python for Data Analysis. O'Reilly Media. ISBN: 978-1-4493-1979-3, Ebook ISBN: 978-1-4493-1978-6

Other related works:

Peter Dalgaard, Introductory Statistics with R, Springer (2008)

[This text may be downloaded free of charge using a UM connection.]

http://link.springer.com/book/10.1007%2F978-0-387-79054-1

(There will be multiple other sources used throughout the course, but I will note them in the slides)

Classroom Policy:

Students are asked to attend class on time and remain through the entire class. Students will need to bring their laptops for the in-class lab.

Original Work:

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on Academic and Professional Integrity (stated in the Master's and Doctoral Student Handbooks) will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to UMSI Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed by the Assistant Dean for Academic and Student Affairs.

Accommodations for Students with Disabilities:

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information you provide as private and confidential.

Student Mental Health and Wellbeing:

The University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (734) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a listing of other mental health resources available on and off campus, visit: http://umich.edu/~mhealth/

Course Requirements:

You are required to bring a laptop with Python 2.7.9+ and R 2.15+ installed to the class for the in-class lab assignments. For instruction on installing Python and R, please refer to the announcements on CTools.

Grading:

Lab/Homework (75 %) - There will be 5 x 100 point homework assignments during the term. Assignments will be posted on CTools.

Project (25 %) - There will be a project worth 100 points. This will involve exploratory data analysis on an interesting dataset you find. You'll put together a project proposal at the halfway point for 20 points, and then a final report (4-5 pages, 65 points) and a final presentation (to be presented during the last class) for 15 points.

Late Homework Penalty: The lab and homework assignments are due when the next class begins. No late submissions will be accepted under normal circumstances. Extensions will only be granted to students with good, documented reasons (e.g. medical grounds or other extenuating circumstances beyond the student's control) at the instructor's discretion.

Letter Grades: Assignment of the final letter grade will be done in accordance with the School of Information Masters Student Handbook guidelines.

Schedule (Tentative, some details subject to change):

Date	Subject	Assignments Due Before Beginning of Class
March 9	Introduction to SI 618: course overview Introduction to R, RStudio and R Markdown	Install software as described in the SI 618 welcome email.
March 16	How to manipulate data frames, How to use qplot and plyr. Basic statistics.	Homework 1
March 23	Smoothing and Trend-finding, Building ggplot Layer by Layer, Database Access	Homework 2 Project Proposal
March 30	ggplot2 Toolbox Finding relationships between variables Time series	Homework 3
April 6	Advanced Topics: Exploratory Cluster Analysis, Principal Component Analysis and Exploratory Factor Analysis	Homework 4
April 13	Course review Final Individual Project Presentations	Homework 5 Project slides Project report