SI 618 Syllabus - Winter 2014

SI 618 – Exploratory Data Analysis

Tuesdays 8:30am-11:30am, NQ1255

Instructor: Kevyn Collins-Thompson (kevynct@umich.edu)  http://www.umich.edu/~kevynct/
GSI: Danny Wu (tzuyu@umich.edu)

Instructor Office Hours: NQ4360 Tuesdays 2:00pm – 4:00pm, or by appointment. (For any questions or issues related to the course.)

GSI Office Hours: NQ4352  Fridays 10am – noon, for technical support on assignments, projects.

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'm not generally 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 subsetting 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:
http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-24544-7/contents/


(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.

I don’t take attendance in this class, but we do have in-class labs every time that must be completed for points. So it will be in your best interest to attend.

**Original Work:**

Unless otherwise specified in an assignment, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but may not copy code or other work and must indicate on your turned-in assignment who you worked with. Any excerpts 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 severe penalties, which might range from failing an assignment, to failing a course, to being expelled from the program, at the discretion of the instructor and the Associate Dean for Academic 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.

**Course Requirements:**

You are required to bring a laptop with Python 2.7.3 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 6 x 100 point lab/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 assignments** (lab, homework, projects) submitted within 24h of the stated deadline (i.e. if the normal deadline is 8:30am Tuesday, emailed to the GSI no later than 8:30am Wed) will be worth 50% of normal credit. After that, assignments can still be graded for learning purposes, but will be worth zero credit.

Assignment of final course letter grades will be done in accordance with the UMSI Masters Student Handbook.

**Schedule (Tentative, some details subject to change):**

<table>
<thead>
<tr>
<th>Date</th>
<th>Subject</th>
<th>Assignments Due Before Beginning of Class</th>
</tr>
</thead>
</table>

https://ctools.umich.edu/portal/tool/a60ffbdd-71da-4795-af0b-6ef47b50d0d0/printFriendly
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 11</td>
<td>Introduction to SI 618: course overview</td>
<td>Install software as described in the 618 welcome email.</td>
</tr>
<tr>
<td></td>
<td>Introduction to R, RStudio and R Markdown</td>
<td></td>
</tr>
<tr>
<td>Mar 18</td>
<td>How to manipulate data frames, How to use qplot and plyr. Basic statistics.</td>
<td>Homework 1</td>
</tr>
<tr>
<td>Mar 25</td>
<td>Smoothing and Trend-finding. Building ggplot Layer by Layer</td>
<td>Homework 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Project Proposal</td>
</tr>
<tr>
<td>Apr 1</td>
<td>ggplot2 Toolbox</td>
<td>Homework 3</td>
</tr>
<tr>
<td></td>
<td>Finding relationships between variables</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time series</td>
<td></td>
</tr>
<tr>
<td>Apr 8</td>
<td>Exploratory Cluster Analysis</td>
<td>Homework 4</td>
</tr>
<tr>
<td>Apr 15</td>
<td>Principle Component Analysis and Exploratory Factor Analysis</td>
<td>Homework 5</td>
</tr>
<tr>
<td>Apr 22</td>
<td>Course review</td>
<td>Homework 6</td>
</tr>
<tr>
<td></td>
<td>Final Individual Project Presentations</td>
<td>Final Presentations</td>
</tr>
</tbody>
</table>