95-868: Exploring and Visualizing Data, Spring 2016 mini 3

Meeting Times

**Lectures:** MW 9-10:20 am, Hamburg Hall 1000
**Review Session:** F 10:30-11:00 am, Hamburg Hall 1000

Instructors and Office Hours

Prof: David Choi
HBH 2101C
davidch@andrew.cmu.edu

TA: Rahul Ladhania
rladhania@cmu.edu

TA: Philipp Burkhardt
pburkhardt@cmu.edu

All office hours TBD (we will take a poll to determine best times)

Course Description

This course covers the fundamentals of statistical exploration and visualization of data. We will fit models and produce specialized graphs to explore data in a detailed and statistics-oriented manner. This course also serves as an introduction to R, a widely used statistical programming language.

In this class, students will learn:

1. How to use R to perform basic data manipulation such as filtering, aggregating, and organizing data sets
2. How to produce graphics in R
3. How transformations, model fits, and residuals can be used to explore and check assumptions about data

Prerequisites

A first course in statistics is required, such as either 95-796 or 90-711.

Laptop policy

Bring your laptop to class. We may sometimes have hands on exercises.

How does this 95-868 compare to 94-842?

94-842 (R for Policy Analytics) is an excellent course which teaches R at a gentler pace than this one. Some students prefer to take 94-842 before this one, while others prefer to jump straight into this course. Both types of students have done equally well in the past. It really depends on whether you tend to get comfortable with new software tools quickly, or slowly.

There is about 1-2 weeks of overlap between the two courses at the beginning, and then they diverge. 94-842 teaches how to write careful and thoughtful reports, which answer questions using linear/logistic regression. 95-868, on the other hand, teaches how to explore the data to discover new insights, which may suggest new questions that you would not have even thought of beforehand. This is sometimes called hypothesis generation. Often, you will discover that the data is complex in ways which go beyond the basic linear model.

Textbooks
Helpful references for using R. These aren't necessary but might be helpful references for you in the future.

- R Graphics Cookbook, by Winston Chang
- R for Everyone, by Jared Lander
- R Cookbook, by Paul Teetor

Helpful references for effective usage of infographics. Note that we'll hardly be covering this topic in class, as your objectives will generally be different (i.e., more analytical) than those of a blog or newspaper. In my opinion, this subject is also better suited for self-study than lecture, i.e., it will be more enjoyable for you to read the books (and look at the pictures!), rather than listening to me recite them to you. They are light reads.

- Creating More Effective Graphs, by Naomi Robbins
- Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations, by Isabel Meirelles

Finally, if you are interested in a deeper discussion of the statistical methods presented in class, you may want to check out

- Visualizing Data, by William Cleveland

**Coursework**

Your grade in this course will be based on 5 homework assignments, and a mini-project.

- **Homework (60%)**
  - There are 5 homeworks, equally weighted towards your grade.
  - Homework should be submitted online via blackboard, by end of day on the scheduled due date.
  - Each student has 5 late days. You may use them at your discretion, to cover travel for interviews, illness, or general business. Otherwise, late homework will not be accepted.

- **Mini-Project (40%)**
  - We will give you a data set and an exploratory objective
  - The project will not necessarily cover the entire course
  - Your objective will be to discover interesting and non-obvious aspects of the data. The goal is exploration, discovery, and hypothesis generation. Not inference.
  - You will be responsible for structuring the analysis yourself and deciding what tools should be used.
  - You will also need to present your analysis and results in a clear and concise manner.

Grades will be curved according to Heinz college standards.

<table>
<thead>
<tr>
<th></th>
<th>Release Date</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hw 1</td>
<td>Wed Jan 13</td>
<td>Wed Jan 20</td>
</tr>
<tr>
<td>Hw 2</td>
<td>Wed Jan 20</td>
<td>Wed Jan 27</td>
</tr>
<tr>
<td>Hw 3</td>
<td>Wed Jan 27</td>
<td>Wed Feb 3</td>
</tr>
<tr>
<td>Hw 4</td>
<td>Wed Feb 3</td>
<td>Wed Feb 10</td>
</tr>
<tr>
<td>Hw 5</td>
<td>Wed Feb 10</td>
<td>Wed Feb 17</td>
</tr>
<tr>
<td>Mini-Project</td>
<td>Wed Feb 17</td>
<td>Fri Feb 24</td>
</tr>
</tbody>
</table>
Homework rubric (12 points total)

- 1 pts: the code compiles on grader’s first attempt
  - 1/1 if the grader runs “knit HTML” and an HTML document is returned instead of errors
  - 0/1 if we ask you to resubmit because the homework is not gradeable for any reason
  - if you do not resubmit in 36 hours: homework is LATE.

Assuming the HW eventually is compile-able:

- 8 pts: correctness
  - 8/8 all problems are 100% correct or have inconsequential errors
  - negative 2 for each significant error in the analysis (up to -8 in total penalties)
  - negative 2.5 for each problem that is not attempted (up to -8 in total penalties)

If the same error is repeated multiple times in the HW, we will try to deduct only once if the rest of the solution can be separated from the error.

- 3 pt: clarity and readability
  - 3/3 if the homework is clear and well written:
    - approach is logical, avoids needless complication
    - variables are well named
    - avoids needless replication
    - good commenting within R code
    - reasonable coding style
    - non-code writing is clear, to the point, and not comprised of “filler material”
  - 0/3 if there are significant coding or writing issues, and we don't understand what you are doing

Email

Homework questions should be sent at least 24 hours before the deadline to assure a timely response. Please include the course code 95868 in the subject line of your emails.

Collaboration

You are encouraged to discuss general approaches and clarification questions with your fellow students. However, you should do your homework yourself.

- Do not look at (or copy) another student's homework.
- Do not copy from another student's homework.

If you receive any help from another student or outside the class (such as stackexchange or other forums or websites), you must clearly identify where you received help. The expectation is that your grade must reflect the work that you alone did.

Tentative Schedule

1. Introduction to R, Rstudio, and RMarkdown/ Data cleaning and aggregation
2. Graphics
3. Averages and sample sizes
4. Univariate distributions
5. Modeling noisy relationships
6. Interactions in multivariate data
7. Data cleaning for larger data sets