**Syllabus for**

**Methods for Data Analysis (Course 141030/141033)**

**Data Science (Program 4693/4694)**

Seattle (Puget Sound Plaza) and online

June 22 – August 24, 2014, Mondays 6:00 – 9:00 P.M.

**Nick McClure**

[nickmc@uw.edu](nickmc@uw.edu%0c)

**Course Description:**

This course is designed to build on what students have learned so far about structuring and manipulating data. By introducing core statistical techniques, students will gain tools that data scientists use to extract insights from data. Students will be asked to apply the course content to real-life scenarios and think creatively as well as critically through issues.

**Course Learning Objectives:**

By the end of the class, students will be able to apply these methods to data and interpret and communicate their results. Topics include:

* Statistical tools for data exploration
* The use of R to apply these tools to real data
* Using inferential statistics to interrogate data
* Testing and experimental design
* Bayesian and classical statistics

**Course Format:**

Each course session will be a mixture of lecture and in-class exercises. Typically the materials for each evening include presentation slides, one or more data sets, and R scripts with illustrations and exercises related to the material. There will also be weekly homework assignments which will include a combination of programming and reading.

**Course Materials:**

No required textbooks. All required reading will be available online as articles or pdf’s. There will also be additional optional reading if students wish to read more on a subject, which may include books or textbooks.

* Required Reading Sources:
  + “An Introduction to Data Science.” By Jeffrey Stanton. <https://drive.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit>
  + “Statistical Thinking for Programmers.” By Allen B. Downey. <http://greenteapress.com/thinkstats/thinkstats.pdf>
* Additional Resources (optional):
  + “Computational Statistics Using R and R Studio: An Introduction for Scientists” by Randall Pruim. <http://www.calvin.edu/~rpruim/talks/SC11/Seattle/RatSC11/Master-StatsForScience.pdf>
* Recommended Reading (optional):
  + “The Signal and the Noise.” By Nate Silver. Penguin Press HC, 2012.
  + “If you have too much data, then good enough is good enough.” By Pat Helland. ACM 2011.

**Technical Requirements:**

Students will be expected to use personal machines in class that are able to:

* Connect to the internet
* Run R (<http://cran.r-project.org/>) and R-Studio IDE (<http://www.rstudio.com/>)
* We will spend one day exploring other tools such as: (All are free and available on Windows/Linux/OSX)
  + SQLite (<http://sqlitebrowser.org/>)
  + Python V2.X ([https://www.python.org/](https://www.python.org/download/releases/2.7.8/))
  + Gephi (<http://gephi.github.io/>) (requires >=Java 1.6)

**Program Webpage:**

Moodle. <http://moodle.extn.washington.edu/course/view.php?id=6268>

**Course Topics and Assignments by Date:**

Topics and Dates are tentative and subject to change.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Date** | **Topic** | **Chapter** | **Assignment** |
| **Week 1** | 2015-06-22 | Introduction; Data Exploration; R overview | Intro DS: Ch.3,9  StatThink: Ch.2 | Data Exploration; Start thinking about project; Vote on preferred extra topics. |
| **Week 2** | 2015-06-29 | Probability Distributions; Conditional Probability; Missing Data; Getting/Storing Data | Intro DS: Ch.7,10  StatThink: Ch.4 | Conditional Probability and Outliers |
| **Week 3** | 2015-07-06 | Outliers and Missing Data;  Introduction to Hypothesis Testing | Intro DS: Ch.6 | Conditional Probability and  Topic Chosen |
| **Week 4** | 2015-07-13 | Hypothesis Testing Continued; Modeling Exercise; The Central Limit Theorem | StatThink: Ch.6,7 | Hypothesis Testing and Modeling Mini Project |
| **Week 5** | 2015-07-20 | Hypothesis Testing Continued; Confidence Intervals; Graph Algorithms | StatThink: Pg.93-97 | Hypothesis Testing |
| **Week 6** | 2015-07-27 | Regression; Feature Selection | Intro DS: Ch.16 | Regression |
| **Week 7** | 2015-08-03 | Feature Selection Continued; Simpson’s Paradox; Intro To Bayes |  | Data Analysis and Feature Selection |
| **Week 8** | 2015-08-10 | Bayesian Statistics | StatThink Pg. 97-101 | Bayesian Analysis |
| **Week 9** | 2015-08-17 | Bayesian Inference with R; Computational Statistics; Model Selection |  | Finish Project up this week! |
| **Week 10** | 2015-08-24 | Review and Possible Extra Topics (Graph Databases, Time Series, Spatial Statistics, NLP, Regex, Basket Analysis, ...) |  |  |

**Student Assessment:**

Students **must** attend at least 8 of 10 classes. Your grade will be based on eight homework assignments and one individual project. Details on these will be handed out/distributed on the first day.

For each homework, students should submit a report that includes:

* Working code which implements the procedures specified by the assignment. Code should be easy to read and commented well.
* Appropriate text and figures/graphs describing the results. All graphs and figures should be **labeled.**

**Policies and Values:**

Your gain from this course is highly dependent on your attendance and completion of the exercises. I fully expect students to actively participate (asking questions, doing the homework, helping others).

Students are expected to behave professionally and abide by all student policies outlined by The University of Washington Student Conduct Code. (http://www.washington.edu/cssc/