Political Science 392 Syllabus:
Introduction to Social Science Data Analytics

Fall 2017

Instructor:	Shahryar Minhas	**Time:**	TR 3:00pm – 4:20pm
Email:	minhassh@msu.edu	**Place:**	104 Giltner Hall
Office Hours:	TR 4:30pm – 5:30pm	**Place:**	317 South Kedzie Hall

Additional information and materials in relation to the class will be posted on the following website: http://s7minhas.com/teaching/pls392/fall2017/

Course Description:

Social science data analytics is a dynamic and fast growing field that draws from disciplines such as Statistics and Computer Science. Its growth has been prompted by the emergence of massive amounts of data that we increasingly recognize as having unique and explicit structure. This structure can take the form of texts such as political speeches, spatial structures that map where conflicts and protests take place, and networks that define how pairs of actors relate to one another within closed arrangements such as the global trading system.

This course provides a hands-on introduction to how to study these and other types of data structures to answer social science questions. To do so we will be working within the R language for Statistical Computing. This course will begin by introducing the basics of R. Then we move towards working on a number of projects to learn how to use the variety of data structures that emerge in the social sciences today to further our understanding of events such as conflict and trade. In the process, we will learn how to:

- handle computational hurdles that can arise in conducting analyses using tools such as **Rcpp**;
- and develop web-based interactive applications to assist in both conducting and sharing analyses with others using tools such as **Shiny**.

Grading

Your final grade will be comprised of the following:

Mini Projects	40%	Final Project	40%
In-Class Midterm	10%	Class Engagement	10%

The grading scale will be as follows:

- 4.0 (92-100 percent)
- 3.5 (87-91.9 percent)
- 3.0 (80-86.9 percent)
- 2.5 (76-79.9 percent)
- 2.0 (70-75.9 percent)
- 1.5 (65-69.9 percent)
- 1.0 (59.5-64.9 percent)
- 0.0 (<59.5)

Teams

Each of you will be assigned to a team of 3-4 students. You and your team will work together on both in-class assignments and the mini projects. I will construct the teams based on your previous exposure to programming topics. Unless there is some serious issue, the teams will not change throughout the semester.

Lectures

Each class will involve me giving a lecture or walkthrough of how to handle some tasks in R, but they are also designed to be interactive. Throughout each class I will ask you to apply the tools and techniques that I have introduced with your team. Learning to work in a statistical computing language like R involves doing more than listening. As such this will also require me to enforce a fairly rigorous class attendance policy. If you are unable to attend a class, you must let me know ahead of time and have a valid reason. Please note that class engagement comprises ten percent of your final grade.

Mini Projects

Forty percent of your grade will be determined by your performance on a series of mini projects that you will complete in collaboration with your team. Since you are operating as a team everyone should contribute equally to each assignment, and at the end of each assignment you will be asked to evaluate your team members. Mini projects will be assigned roughly every other week and are expected to be completed by the due date. Late assignments will be penalized as follows:

- Late, but same day: -10%
- Late, next day: -20%
- 2 days or later: no credit

Final Project

Another forty percent of your grade will be determined by your final project. This will be an open-ended exercise in which you tackle some “interesting” question related to the social sciences using the techniques that we have covered in class. You will be allowed to form your own team of 3-5 students for this assignment.

A proposal for what you plan to do is due to me on Oct 26. Following the delivery of the proposal, I will meet with you to discuss whether the proposal is satisfactory. Not turning the proposal in on time will lead to a -10% reduction in the grade for your final project.

Additionally, during the scheduled final exam period for the class (Dec. 13: 10am to 12pm in 104 Giltner Hall), you and your team will give a 15 minute presentation on your final project and whatever deliverables you have made for the final project (e.g., a 10 page research report, an interactive visualization, or an R package) are due by 10am on Dec. 13.

Midterm

There will be one in-class midterm on Oct. 5 that you are expected to complete individually and will comprise ten percent of your grade. The exam will ask you to complete a number of programming tasks related to the material presented in the class by that date. We will have a review session for this midterm on Oct. 3.

Helpful Resources

No books. Here are some helpful resources:

- Wickham's R Tutorial
- R Cookbook
- ggplot2
- stackoverflow

Class Equipment

In a class about Social Science Data Analytics, you will need a computer that has some specific software on it. You should be able to do everything we do in class using modern Mac, Linux and even Windows operating systems. If you have questions, ask me. In the course, we will be using free, open-source software. It will be necessary to have a laptop in class.

- R: This is a free program, available here.
- You should use RStudio to run R, as it is very helpful.
- Github: Set up an account here github.

Policies

Missing class or midterm: If you know that you are going to have to miss a class or the midterm, let me know at least 48 hours beforehand. If you simply skip the midterm, you will get a zero and no make-up will be offered.

Academic Integrity

Don't cheat . . . `https://ombud.msu.edu/academic-integrity/student-faq.html`:

> "Academic honesty and integrity are fundamental values in a community of scholars. As stated in the MSU Student Rights and Responsibilities and Spartan Code of Honor, students and faculty share a commitment to and responsibility for "maintaining the integrity of scholarship, grades, and professional standards." To abuse these values is to assault one's own personal integrity and character.

Don't take without attribution . . . `http://www2.stat.duke.edu/~cr173/Sta323_Sp17/syllabus/`:

> "A note on sharing / reusing code – I am well aware that a huge volume of code is available on the web to solve any number of problems. Unless I explicitly tell you not to use something the course's policy is that you may make use of any online resources (e.g. StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. The one exception to this rule is that you may not directly share code with another team in this class, you are welcome to discuss the problems together and ask for advice, but you may not send or make use of code from another team."

Approximate Class Schedule

Week 1 – Sept. 5 & Sept. 7

- Introduction to the Syllabus, R, & R Markdown
- Helpful Resources:
  - Wickham's R Style guide
  - Google R Style guide
  - Advanced R – Foundations
  - R Markdown
  - R Markdown Cheat Sheet
  - R Markdown Examples

Week 2 – Sept. 12 & Sept. 14

- Types of Data structures, how to subset, and how to manipulate using base operations in R and the **dplyr** package
- Helpful Resources:
  - Advanced R – Foundations
  - Software Carpentry – Data Types & Structures
  - **dplyr** Vignette
  - RStudio Data Wrangling Cheat Sheet

Week 3 – Sept. 19 & Sept. 21

- Plotting: base vs. **ggplot2**
- Helpful Resources & some blogs that describe effective practices for creating visualizations:
  - Plotting in base
  - Base R Plotting
  - R Cookbook for **ggplot2**
  - R for Data Science – Data Visualization
  - **ggplot2** Reference Manual
  - **ggplot2** Cheat Sheet
  - **ggplot2** Extensions
  - Color Brewer
  - "Tufte's Rules" by Sealth Reinhold
  - Information is Beautiful
  - Flowing Data
  - Junk Charts
  - Storytelling with Data
  - Ann K. Emery

Week 4 – Sept. 26 & Sept. 28

- Time-series-cross-section (TSCS) data: This week we will explore a couple of commonly used datasets in political science using the techniques we have discussed so far:
  - Polity IV tracks variation in political institutions across countries and time
  - World Bank collects information on the economic and demographic characteristics of countries across time

Week 5 – Oct. 3 & Oct. 5

- Review Session for In-class Midterm on Oct. 3.
- *In-class Midterm* on Oct. 5.

Week 6 – Oct. 10 & Oct. 12

- Scraping Data from the Web: An increasing amount of data is available on the web (e.g., election results, budget allocations, legislative speeches). These data are provided in an unstructured format: you can always copy & paste, but it's time-consuming and prone to errors. Web scraping is the process of extracting this information automatically and transform it into a structured dataset.

Week 7 – Oct. 17 & Oct. 19

- Scraping Data from APIs like Twitter & Facebook: Sites like Twitter and Facebook offer a set of structured *http* requests that return JSON or XML files, this week we will learn to work with APIs such as these and will also discuss how social scientists have used data from these sites.

Week 8 – Oct. 24 & Oct. 26

- *Networks* Network analysis is becoming an increasingly common mode of analysis in the social sciences. We will review how network analysis has been used in the social sciences and how to work with networks in R.
- *Final Project Proposal*: Due on Oct. 26.

Week 9 – Oct. 31 & Nov. 2

- *Spatial Data*: Processes such as democratization and conflict have been shown to diffuse spatially. We will review how to visualize and extract descriptive information from spatial data using R.

Week 10 – Nov. 7 & Nov. 9

- *Textual data*: Words matter in political science and in recent years political scientists have increasingly utilized data analytic techniques to study treaties, congressional bills, and legislative speeches. We will go over how to structure, visualize, and summarise textual data in R.

Week 11 – Nov. 14 & Nov. 16

- *RShiny*: Interactive visualizations are a great way to better understand and explore your data. The team at RStudio have developed a package, **Shiny**, through which we can build interactive web applications straight from R.

Week 12 – Nov. 21 & No class on Nov. 23

- *Importance of reproducible research*: For your data analysis to be trusted, others must be able to reproduce the work that you have done.

Week 13 – Nov. 28 & Nov. 30

- *Special Topic*: Making full use of your computer with parallelization.

Week 14 – Dec. 5 & Dec. 7

- *Special Topic:* Handling biggish data in SQL or Speeding up computation using **Rcpp**.

Week 14 – Dec. 5 & Dec. 7

- *Work on final project in-class or a Special Topic such as how to build an R Package.*

Dec. 13: 10am to 12pm in 104 Giltner Hall

- *Final Exam*: In class presentations of final projects.