Introduction to Data Science: Part 1
Universidad del Rosario

Instructor: Kevin Munger
km2713@nyu.edu
Department of Politics
New York University
June, 2018

Course Description: In this course (the first half of a course co-taught with Profesor Jorge Gallego), we will cover the basics of data science. This course is aimed at students with an interest in applying recent developments in data science to any number of possible endeavors, including academic research, government work, the non-profit world or the corporate sector. That said, I am coming from an academic background, and the course will be presented with a rigor commensurate with academic research. This portion of the course will begin with a theoretical overview of the scope of data science and the basic understanding necessary to make valid inferences using the increasingly powerful methods that constitute it. We will then delve into areas of data science with which I have the most experience working: textual analysis, unsupervised models, and social media.

Methods: The course is both theoretical and practical. We will discuss high-level concepts as well as specific data manipulation and statistical techniques; the latter will entail both formal mathematical explanations and implementation in the statistical language R. That said, this course is an overview, and we will not focus too much on deriving each formula or writing programs from scratch. Rather, the goal is that students will be able to employ basic data science methods correctly and productively by the end of the course. All of the course materials can be found on my github: https://github.com/kmunger/Intro_Data_Science_Rosario

Before the course: In order to take the best advantage of our time together, I encourage all students to have the latest versions of R and R studio installed.

Evaluation: Final grades for the course will be a simple average of the score on the final exam (one for each section of the course) and a project proposal (one for the overall course).

- Final Exam: 50%
- Project Proposal: 50%

Weekly Schedule

Note that the specific daily schedule may vary based on our progress through the course material. A more detailed breakdown of the code/slides we'll be using can be found on the course github.

1: Introduction to Data Science

- What is Data Science?
- Thinking like a data scientist
- Ethics

2: Introduction to R

- Why are we using R?
- Getting around in R

3: Introduction to Using Text as Data

- Making text into data
- Getting to know the text
- R text package: quanteda

4: Unsupervised Models

- Making sense of noisy data
- Dimension reduction
- Clustering

5: Text as Data Applications

- Applying PCA to detect a mystery author

6: Social Network Analysis

- Graphing networks
- Calculating network statistics

7. Web Data – Scraping and APIs

- Scraping unstructured web data
- Using for structured data

8. Web Data – Twitter

- Using the Twitter APIs
- Plotting tweets over time and space

9: Exam