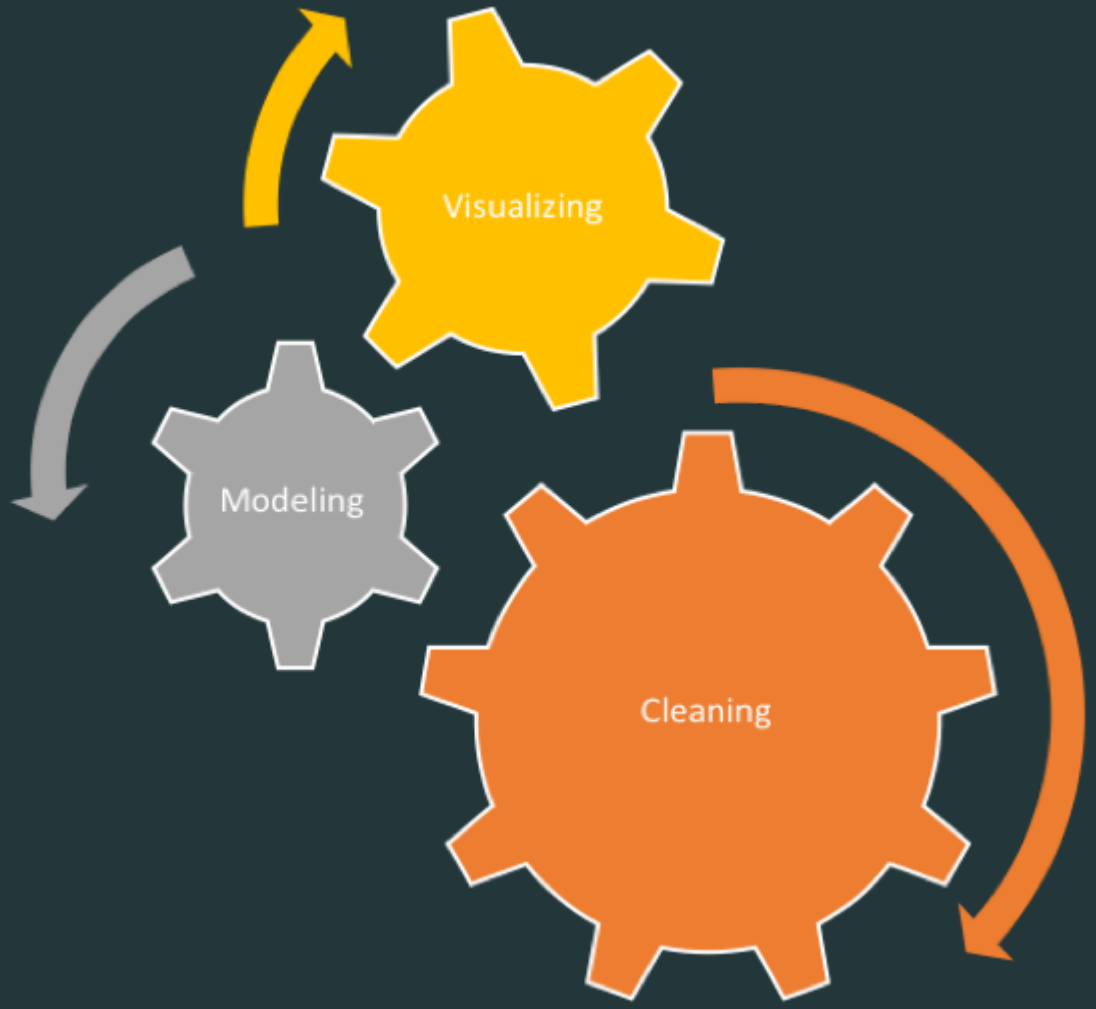
Lecture 1 - Data Science Workflow

R for Data Science

Akbar Akbari Esfahani

01/14/2019

Data Science Workflow



Today's Agenda

- Course Overview
- Programming, coding and version control - the data science workflow
- Introduction to GitLab, R, RStudio, and R-Markdown

My Background

- University of Colorado
  - Master's Degree in Stats and GIS
  - Collaboration with School of Medicine on developing Doc-In-Box
- US Geological Survey
  - Modeling of Soil based on satellite images
  - Modeling of unexploded ammunition based on radar
  - Climate change forecasting
- UCLA Center for Health Policy Research
  - Data Science with Survey research
  - NLP for Survey data
  - Developed methods in R for automated quality control of data
- Highmark Inc
  - In charge of introduction of new developments in data science
  - Automation of Reports with R and Shiny Apps
  - Democratization of data
  - Data Science evangelist
  - Overseeing product testing and procurement for department

A Data Science Project

1. Background
    - A business question that needs to be addressed for a value
2. Data Collection
    - The business question is translated to data requirements
    - Once the requirements are clear, we need to collect the data
3. The Resuls
    - We create a model based on the data product from 2
    - Once model is validated, we create a report
4. In Production
    - If the customer accepts finding from 3, we put our model into production

How this class works

- Basic understanding of programming
  - We will cover advanced methods in R and data science
  - You need to understand data structures, coding, and creating scripts
- Some stats knowledge presumed
  - We Will cover some advanced methods in data science which require understanding of Statistics
- Class attendance is mandatory
  1. If you miss a class, you will miss the explanation of a lab
  2. you'll miss some information for homework
  3. you'll miss some information for the final project
- Collaboration is expected
  - Data Science does not happen in a vaccum, we do a lot of team collaboration
- Class will be very cumulative
  - See class attendance

Class Structure - Less talk, more doing

Grading Structure

- Class participation (15%)
- Lab Work (15%)
- Assignments (20%)
- Final project (50%)

Disclaimer: Lab work and class participation can be adjusted at instructer's discretion. While late homeworks will not be accepted for grades, I highly encourage you to complete them and turn them in.

More about grading

Class participation/Labs (30%)

- Labs: Each lecture has an accompanying lab assignment that is due next day by 6pm.
- The completion of labs is needed to finish your homework

Homework assignments (20%)

- Homework: There will be four homework assignment that builds on top of your in-class labs.
  - Each homework also builds on top of the other homeworks.
- Single lowest HW score will be dropped
- HW assigned are due following week at start of class
- Late homework will not be accepted for credit

Final project (50%)

- Final Project: Instead of a final test, we will have a final project that will be in the form of a complete report that will be the accumulation of all the labs and homeworks. I encourage you to find a topic early on as to avoid last minute problems.

Course resources

Required textbook: Garrett Grolemund and Hadley Wickham, R for Data Science

Other resources

Hadley Wickham, ggplot2
RStudio Webinars
RMarkdown official website

We will be very collaborative and I will insist on version control, thus, each student is required to create a free GitHub account

Goal of this class

This class will teach you to use R, the integerated development environment (IDE) RStudio and version control and collaboration through GitHub to:

1. Create scripts and use version control
2. Generate graphical and tabular data summaries
3. Perform data science analytics: Data Wrangling, Data Visualization, and Data Modeling
4. Produce reproducible statistical reports using R Markdown
5. Integrate R with other tools (e.g., databases, web, etc.)

Why R?

- I like it
- Free (open-source)
- Programming language (not point-and-click)
- Excellent graphics
- Offers broadest range of statistical tools
- Easy to generate reproducible reports
- Easy to integrate with databases without the need to learn another language

Understanding data science problems

there are multiple approaches to solve a problem and all of them lead to the same solution, but chosing the right one for your data at hand is key

Some pointers on coding in general

- Don't reinvent the wheel
- If you have a problem, ask for help
- Most problems you have, someone else had it before you
- Write your code in a way that you can look at it in a year and know what you were doing
- Create a StackOverflow account, look for help and in return, try to help
- Use functional programming

Today's Lab: *GitLab and RStudio*