ASDS Code Camp

Day 1: Introduction

Martyn Egan (Teaching Fellow)

September 6 - 10, 2021

Schedule

- Monday: Introduction
- Tuesday: R basics
- Wednesday: Python basics
- Thursday: Good practices
- Friday: How to write up and report

Today's class

What is data science?

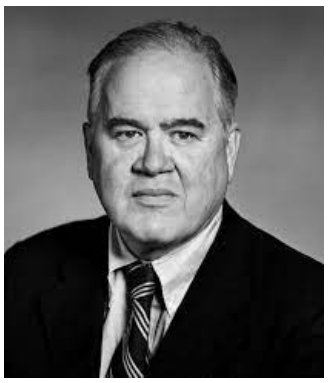
R and Python

Expectations

What is data science?

- 'The science of learning from data' *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
  avoid the hype

A brief history...



John Tukey, 1915-2000

'All in all, I have come to feel that my central interest is in **data analysis**, *which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.'*

The Future of Data Analysis, 1962.

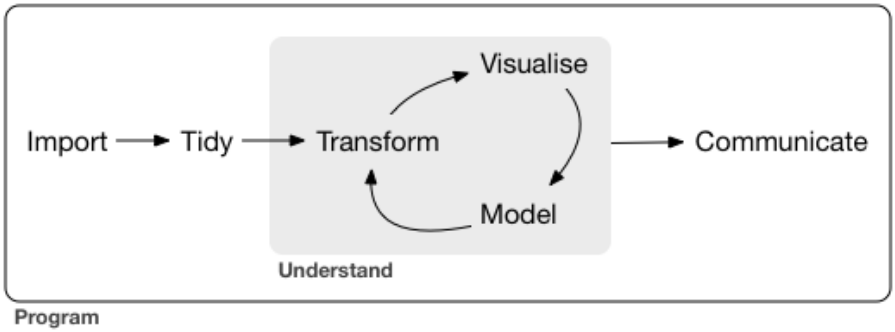
Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey
- 1990s: 'data mining' and 'knowledge discovery' emerge as terms for finding patterns in increasingly large datasets
- 1996: 'data science' included for first time in International Federation of Classification Societies (IFCS) conference title
- 2000s: analytics becomes increasingly important to businesses, 'big data' becomes a thing

'I keep saying the sexy job in the next ten years will be statisticians [...] The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.

Hal Varian, Google Chief Economist, Jan. 2009

Data science is a process



Data science tools and workflow, c/o Hadley Wickham (R for Data Science)

Things to keep in mind

- Data science and statistics are essentially the same, but in practice they are coming to mean different things
  - **Statistics**: the mathematics associated with inference
  - **Data science**: the practices associated with working with data
  - Not everyone agrees with this distinction...
- There is a difference between **scientific** and **engineering** mindsets
  - the scientific mindset seeks to understand the underlying process
  - the engineering mindset looks to find the best prediction
- In the social sciences, we often want to understand what's inside the 'black box', but not all data science methods are designed for this.

R and Python

What is R?

R is a programming language for statistical data manipulation and analysis. It is based on the S language developed by AT&T in the 1970s. R supports both object-oriented and functional programming. It is free to download and use, and is supported by a large and active user community. There are a great many packages available in R which extend the language's capabilities to all manner of statistical methods.

What is Python?

Python is a general-purpose programming language developed by Guido van Rossum in the late 1980s. Its underlying philosophy emphasises readability. Like R, it supports both object-oriented and functional programming. Unlike R, its applications extend beyond statistics and data analysis. In recent years Python has developed a large data science user base, primarily due to the data science capabilities of the numpy, pandas and matplotlib libraries.

Expectations

- There is a lot to learn
- There is a steep learning curve
- Slow and steady wins the race
- We are here to help

Some useful resources

- David Donoho (2017) "50 Years of Data Science", Journal of Computational and Graphical Statistics, 26:4, 745-766
- Leo Breiman (2001) "Statistical Modeling: The Two Cultures", Statistical Science, 16:3, 199-231
- Hadley Wickham and Garrett Grolemund (2017) "R for Data Science", O'Reilly
- Wes McKinney (2018) "Python for Data Analysis", O'Reilly (2nd. ed.)
- John Tukey (1962) "The Future of Data Analysis", The Annals of Mathematical Statistics, 33:1, 1-67
- John Tukey (1977) "Exploratory Data Analysis", Pearson