DSCI 100 - Introduction to Data Science

Lecture 2 - Getting data into R

2019-01-10

Housekeeping

- Late assignments for worksheet_01 should now be able to be uploaded to Canvas
- The course late policy (https://github.com/UBC-DSCI/dsci-100#lateabsence-policies) applies (except in exceptional circumstances)

Last week

- Introduction to R + Jupyter and a sprinkle of data analysis

Get that data into R!

- 4 most common ways to do this in Data Science
    1. **read in a plain text file where data is in the shape of a rectangle (think spreadsheet)**
    2. from a database (e.g., SQL)
    3. **scrape data from the web**
    4. use a web API to read data from a website

A note on scraping data from the web

- More and more websites don't want you scraping
- They instead are providing "easier" ways for you to access the data as opposed to scraping it (which they can regulate and know who you are)
- So, TL;DR read the Terms of Service for ANY webpage you are planning on scraping
    - they're long to read, so search for "scraping", "auto", "bot", etc to find the relevant section

A note on the scraping exercises in worksheet_02 on the server

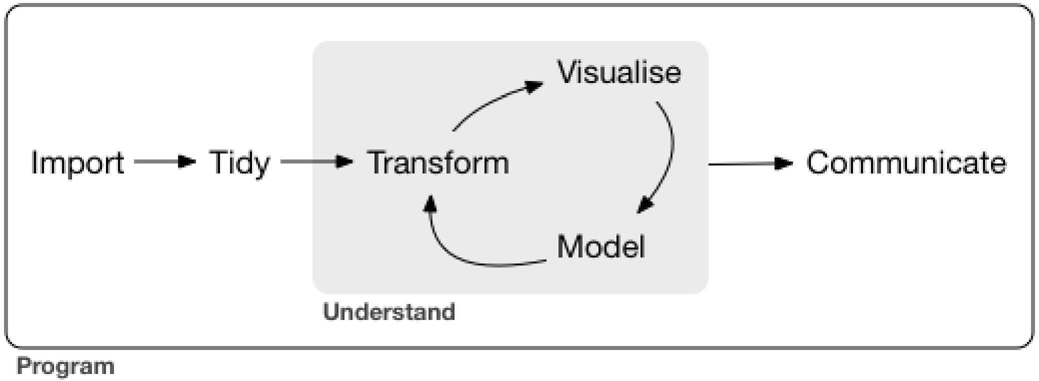
There is a known error that we have been getting, it looks like this:

```
Error in xpath_combinedselector(<S4 object of class
structure("CombinedSelector", package = "selectr")>): could not
find function "xpath_combinedselector"
```

Honest answer/solution:

- We have not been able to pinpoint exactly what is going wrong.
- BUT we have been able to "work around it" by a couple different solutions:
    - install `selectr` package and restart kernel (might take 1-2 restarts...)
    - use the XPath from the selectorgadget instead of the CSS Selector (see Aaron's code on next slide)

Put this step in context - the Data Science workflow



source: R for Data Science (https://r4ds.had.co.nz/) by Grolemund & Wickham

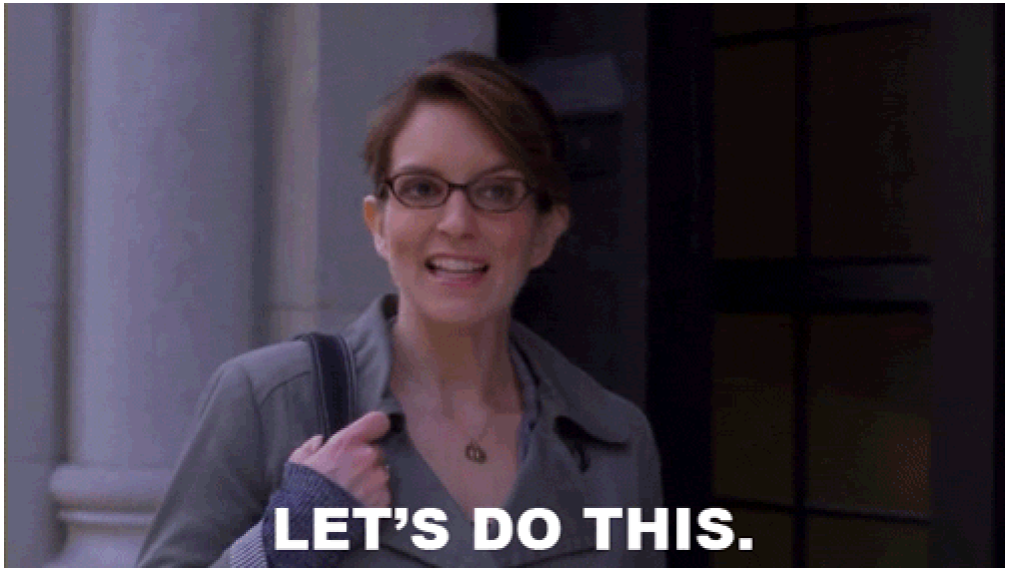
Note about data import

- It is important!
- Think of it as tying your shoes before you run, it might be "boring" but if done wrong it will trip you up later!

Questions?

- General?
- Specific?

Let's do this!



Class activity:

- In the group at your table, try to read in this dataset from the web: https://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt (https://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt)

What did we learn?

- `read_table2` allows us to read multiple whitespace delimited files
- read the terms of service before scraping
- when to use the different forms `read_delim` (or `read_*`)