Lecture 3, August 29

Expressions

Slides created by Ani Adhikari and John DeNero
Announcements

● Office hours are in Weekly Schedule in the top menu bar of data8.org

● Homework 1 is due on Thursday at 5 pm; you get a bonus point for turning it in by Wednesday at 5 pm.

● Submission instructions are in the post “HW 1 Posted” on Piazza and have also been sent by email.
From last time ...
Key to establishing causality

If the treatment and control groups are similar apart from *the treatment*, then differences between the outcomes in the two groups can be ascribed to the treatment.
If the treatment and control groups have *systematic differences other than the treatment*, then it might be difficult to identify causality.

Such differences are often present in **observational studies**.

When they lead researchers astray, they are called **confounding factors**.
Randomize!

- If you assign individuals to treatment and control at random, then the two groups are likely to be similar apart from the treatment.

- You can account – mathematically – for variability in the assignment.

- Randomized Controlled Experiment
Regardless of what the dictionary says, in data science

Random ≠ Haphazard
Working with data
Bad. Really bad.
Chartjunk

- A term coined by Edward Tufte – look him up!
- It means **needless graphics**, especially if they distract from the main point.
Much better

Plot by Ross Ihaka, Cal Stat PhD and one of the creators of R.
Arithmetic
Python is popular both for data science & general software development

Data science requires mastering several extensions to the core language, such as tables, plots, & notebooks

Mastering the language fundamentals is also critical

Learn through immersion, not by reading the dictionary

(Demo)
## Arithmetic Operators

<table>
<thead>
<tr>
<th>Operation</th>
<th>Operator</th>
<th>Example</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addition</td>
<td>+</td>
<td>2 + 3</td>
<td>5</td>
</tr>
<tr>
<td>Subtraction</td>
<td>-</td>
<td>2 - 3</td>
<td>-1</td>
</tr>
<tr>
<td>Multiplication</td>
<td>*</td>
<td>2 * 3</td>
<td>6</td>
</tr>
<tr>
<td>Division</td>
<td>/</td>
<td>7 / 3</td>
<td>2.66667</td>
</tr>
<tr>
<td>Exponentiation</td>
<td>**</td>
<td>2 ** 0.5</td>
<td>1.41421</td>
</tr>
</tbody>
</table>
Names
Assignment Statements

- Statements don't have a value; they perform an action
- An assignment statement changes the meaning of the name to the left of the = symbol
- The name is bound to a value (not an equation)

(Demo)
Carte Figurative des pertes successives en hommes de l’Armée Française dans la campagne de Russie 1812-1813.

Les nombres d’hommes perdus sont représentés par les longues des goupilles de caisse d’un millième pour des mille hommes : ils sont de plusieurs centaines fourchus de l’armée. 

La ligne indique les hommes qui restent en Russie, qui sont ceux qui en Action, ainsi que les contingents qui on dû à travers la carte en 1813, pour dessiner les contours du territoire de l’Armée et le 25 Octobre.

Le tableau graphique de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les courbes passent au golfe de Finlande, est.

Tableau graphique de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Charles Joseph Minard, 1781-1870

Tufte called Minard’s visualization of Napoleon’s 1812 march “probably the best statistical graphic ever drawn.”