DATA VISUALIZATION

Polonsky/Brine Summer Internships - 2018

Prof. Deena Engel

Introduction to Data Visualization

When Mark Twain famously said …

“ [there are] lies, damned lies, and statistics”

… he did not know about data visualization!

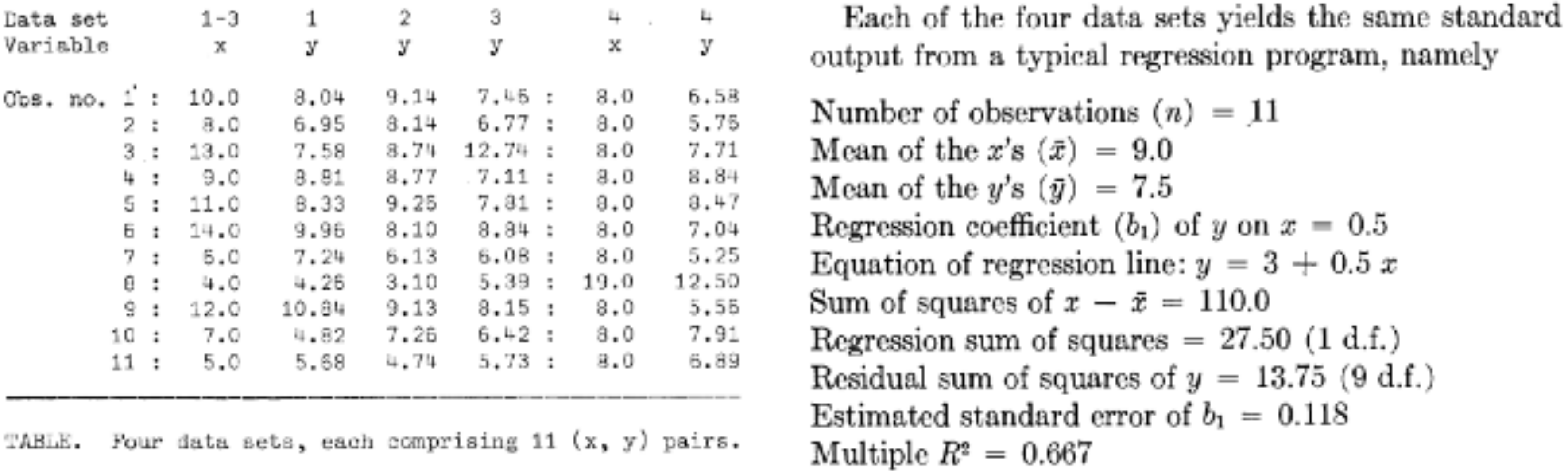
Criteria for a Good Visual Display

- Show the data
- The user should be able to focus on the data … and not the methodology or the software.
- Avoid data distortion!
- Present many numbers in a small space; ideally this is a way to compress a large amount of data into a small visual space.
- Encourage the viewer's eye to compare/contrast
- Show the data at various levels of detail
- Integrate the statistical, visual, and textual descriptions of the data

from Tufte's Visual Display of Quantitative Information

A famous example of deceptive statistics:

- This is a famous example from Anscombe's "Graphis in Statistical Analysis", *American Statistician*, 27 (February 1973), 17-21.



Data set		1-3	1	2	3	4	4
Variable		x	y	y	y	x	y
Obs. no.	1 :	10.0	8.04	9.14	7.46 :	8.0	6.58
	2 :	8.0	6.95	8.14	6.77 :	8.0	5.76
	3 :	13.0	7.58	8.74	12.74 :	8.0	7.71
	4 :	9.0	8.81	8.77	7.11 :	8.0	8.84
	5 :	11.0	8.33	9.26	7.81 :	8.0	8.47
	6 :	14.0	9.96	8.10	8.84 :	8.0	7.04
	7 :	6.0	7.24	6.13	6.08 :	8.0	5.25
	8 :	4.0	4.26	3.10	5.39 :	19.0	12.50
	9 :	12.0	10.84	9.13	8.15 :	8.0	5.56
	10 :	7.0	4.82	7.26	6.42 :	8.0	7.91
	11 :	5.0	5.68	4.74	5.73 :	8.0	6.89

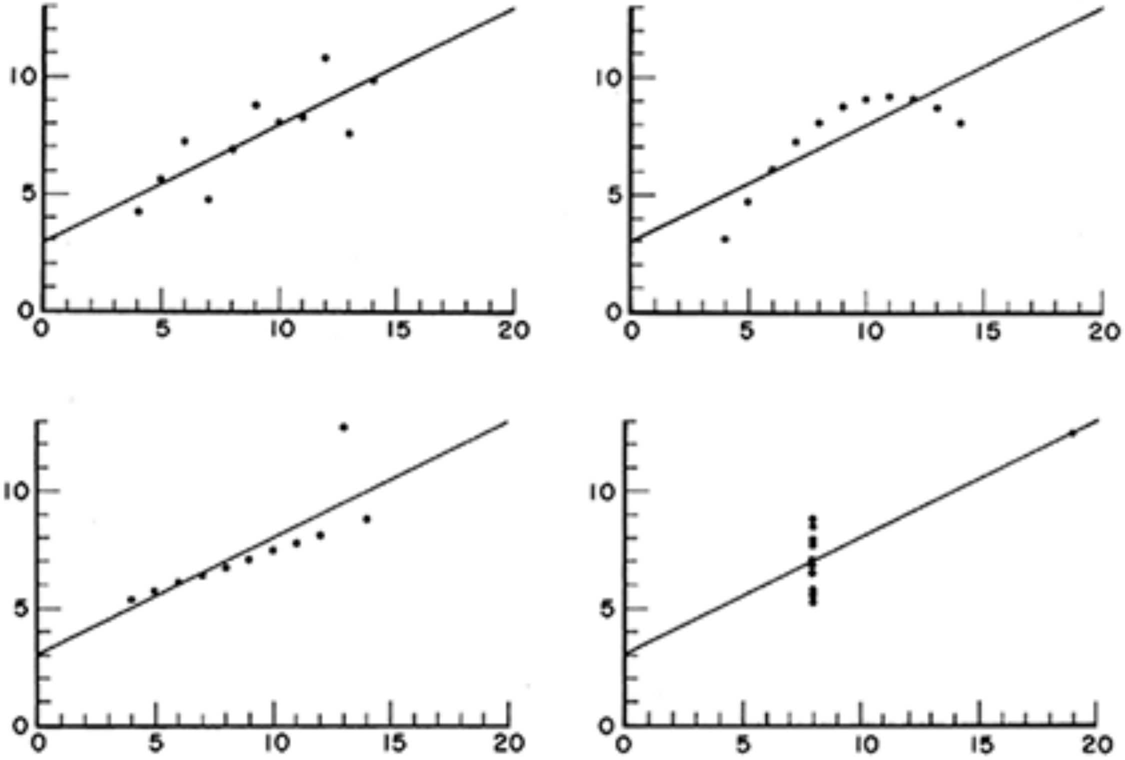
TABLE. Four data sets, each comprising 11 (x, y) pairs.

Each of the four data sets yields the same standard output from a typical regression program, namely

Number of observations (n) = 11
Mean of the x's ($\bar{x}$) = 9.0
Mean of the y's ($\bar{y}$) = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667

http://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf

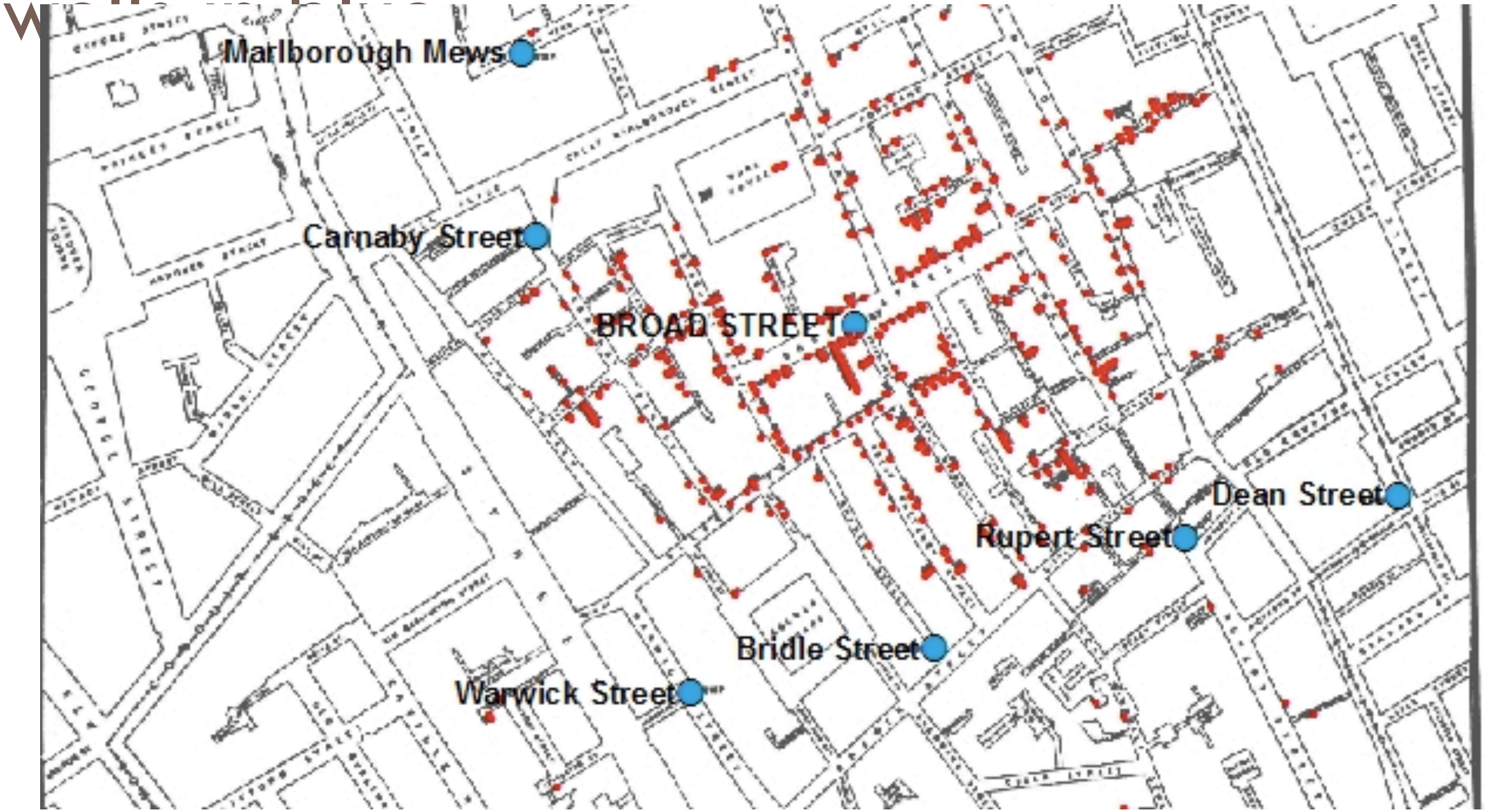
… and the graphical results which display the differences not seen in the statistical results:



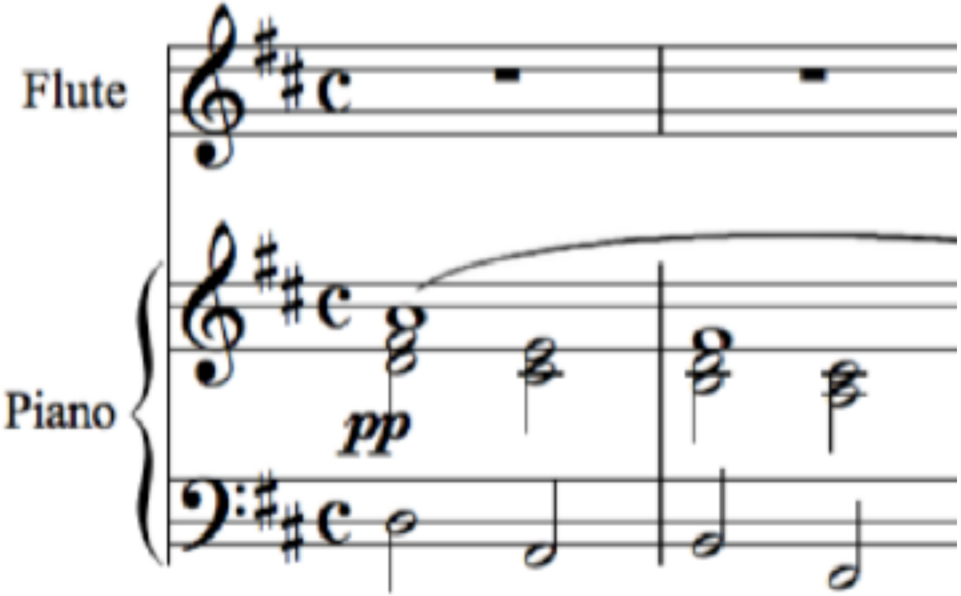
An early example of data visualization

- Cholera in London in 1854:
- Dr. Snow drew a chart
  - He juxtaposted the locations of 578 cholera deaths
  - … with 13 public water wells

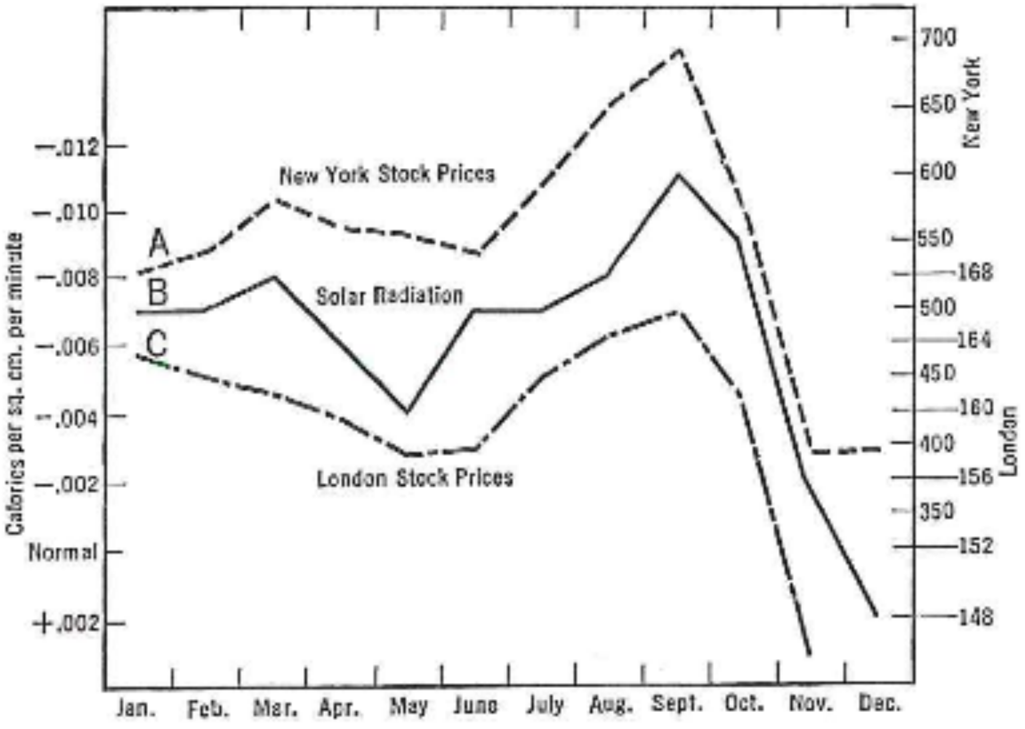
Excerpt from Dr. Snow's chart with the cholera deaths in red and the water wells in blue



Musical Scores as Data Visualization



Illogical and incorrect hypotheses … lead to errors in the results:



SOLAR RADIATION AND STOCK PRICES

A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Tufte *Visual Display*, p.15)

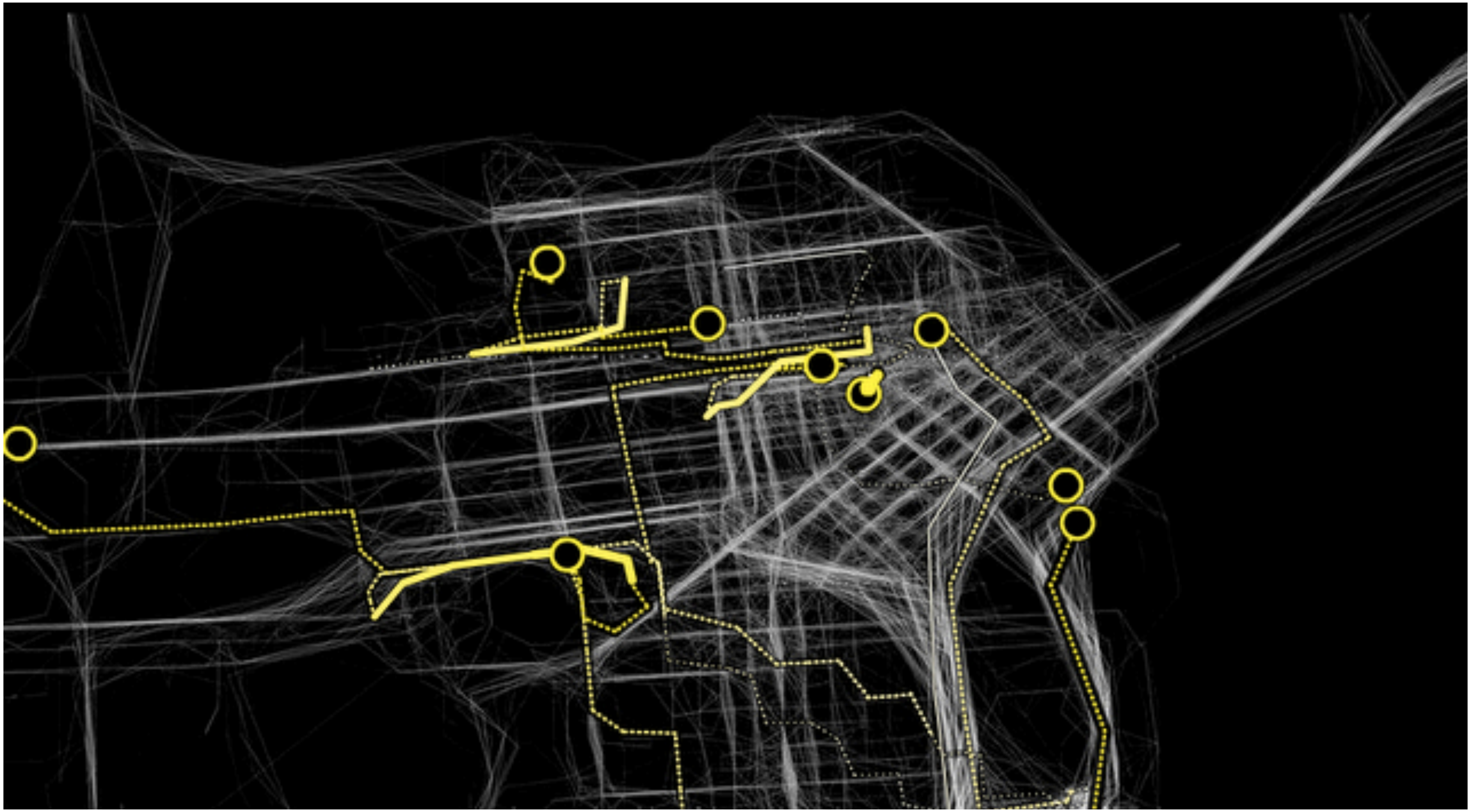
Data Visualization as Contemporary Art

- MoMA – the Museum of Modern Art in New York City – owns and has displayed a number of works of art that use data visualization, e.g. *Cab Spotting*
  - Tracks the movement of taxi cabs in the Bay Area
  - The software uses time lapses and builds a map of only the taxi cabs and their locations

"By looking at the aggregate cab data from San Francisco, Cabspotting reveals the Bay Area's economic, social, and cultural patterns. The work is meant to inspire viewers with an awareness of the vast simultaneous activities of their fellow human beings… the piece reveals the possibilities of benevolent and inspiring uses of surveillance technologies."

Cab Spotting (excerpt)

http://www.snibbe.com/projects/interactive/cabspotting



I Want You to Want Me

- MoMA also owns and has exhibited *I Want You to Want Me* by Jonathan Harris and Sep Kamvar
- This is a work of art that consists of many components and is considered “database art”:
  - MySQL
  - Java webcrawler
  - C++ front-end along with CGI components
  - Multi-media (videos, still images, music)
  - Customized hardware

“I *Want You To Want Me* explores the search for love and self in the world of online dating.”

I Want You To Want Me

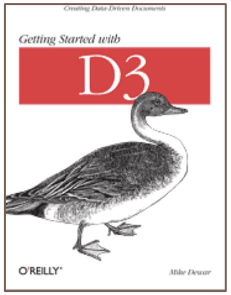
http://iwantyoutowantme.org/

Software for Charts and Data Visualization: Spreadsheets

- MS-Excel
  - Bar charts
  - Scatter plots
  - Pie charts
  - Line charts
  - Bubble charts
  - … and many others
- These charts and graphs are widely used with small datasets.

Software for Charts and Data Visualization: JavaScript / Open Source Libraries

- There are a number of JavaScript libraries that are used with data for charts and/or data visualizations:



- Rgraph - http://www.rgraph.net/
  - Uses HTML5 and JavaScript
  - Supports 22 types of charts
  - Used in social sciences, business, and other fields
- D3 – *Data Driven Documents* - http://d3js.org/
  - D3 is widely used in the arts and humanities as well as the sciences
  - O'Reilly book has tutorials using MTA subway data

Software for Charts and Data Visualization: Java, Processing, R and others

- Java
  - Mondrian - http://www.theusrus.de/Mondrian/
  - Processing - http://processing.org/
- Other
  - R - http://www.r-project.org/ (open source)
  - Origin http://www.originlab.com/index.aspx?go=Products/Origin/Statistics (proprietary)
  - SAS-JMP - http://www.jmp.com/ (proprietary)

Software for Charts and Data Visualization: Textual Analysis

- Google books – nGram Viewer https://books.google.com/ngrams
- Word Clouds http://www.wordle.net/

Software for Charts and Data Visualization: Networks, complex systems

- gephi.org - https://gephi.org/
  - Written in Java
  - Field: All types of networks
  - Types: Directed, unidrected, mixed
  - Size: <1M nodes & edges
  - Layouts: various

Questions?

- Please feel free to email me to meet if you would like to ask about data visualization specific to your projects!