Welcome to DSCI 101

Introduction to Data Science

Week 4-5 Topics Recap

- Basic dataframe methods:
  - import with ***pd.read_csv***
  - indexing with ***df[], df.loc*** and ***df.iloc***
  - sort, rename, mutate, drop a column content
  - ***df.sample:*** get a random sample of rows

- Data summaries and manipulation:
  - ***series.value_counts, df.groupby, df.pivot_table, df.describe***
  - ***df.merge***: join two df by common column(s)
  - missing values

Week 6 Topics Preview

- Exploring data through visualization:
  - Variable types and basic plots: bar plot, histogram, boxplot, line plot and scatter plot
- Storytelling with data
  - Modern data and advanced plots
- Data visualization learning outcome
  - choose the right plots and implement in Python
  - read and interpret plots

A review

Variable

Look out for:

- ***Categorical coded as numbers***
- ***Numerical stored as strings***

Numerical

Continuous

Examples:

- temperature
- price

Discrete

Examples:

- number of kids
- years of school

Categorical

Nominal

Examples:

- ethnicity
- political party

Ordinal

Examples:

- survey ratings
- pain level

Categorical – bar plot

- X-axis:
  - categories
- Y-axis can be:
  - numerical
  - count or proportion
- Length of the bar encode values
- Width of the bar encode nothing!

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Figure 2.17: A contingency table for `app_type` and `homeownership`.

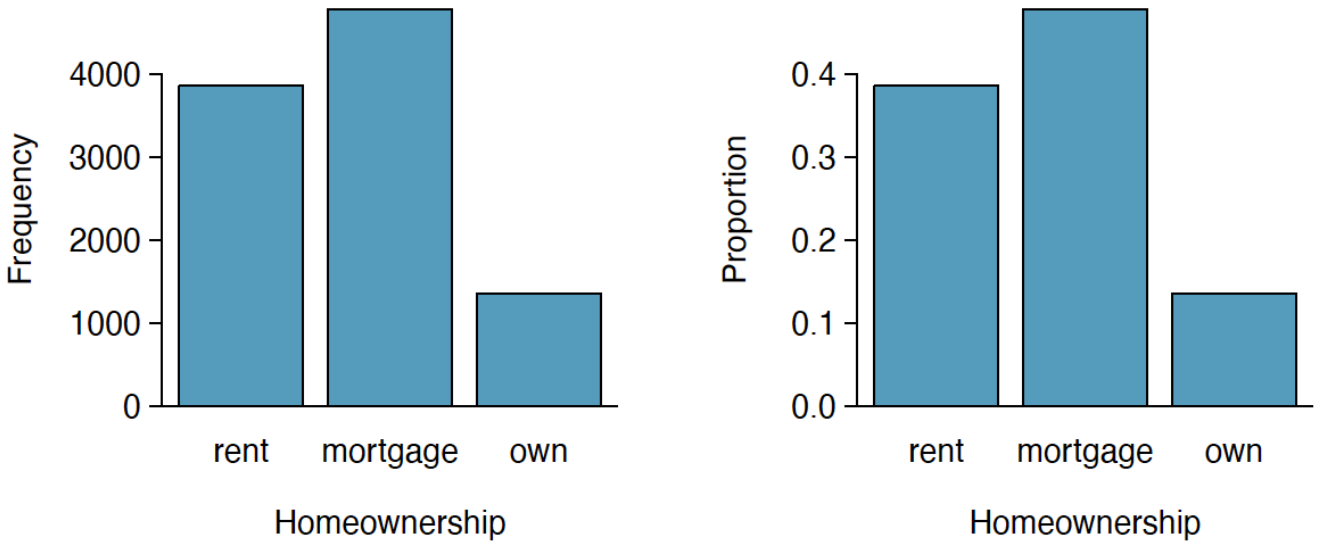
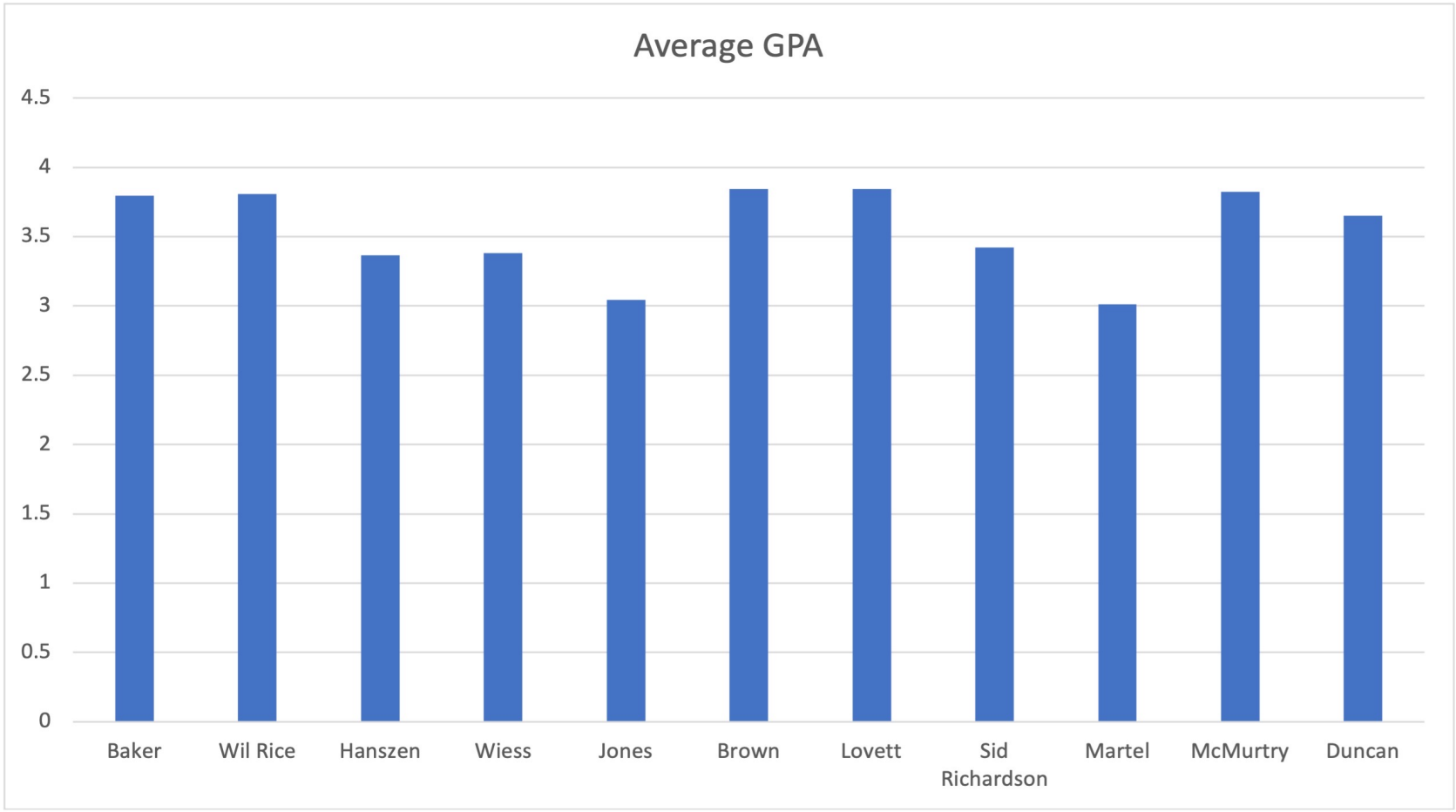


Figure 2.19: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

More bar plots

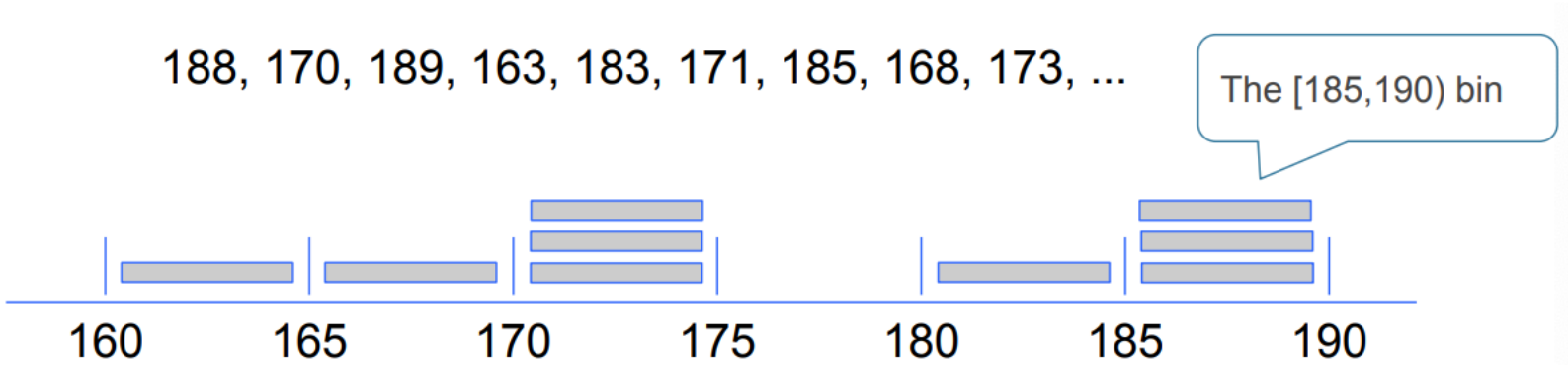
Be mindful of color choices!

More bar plots – categorical and numerical



Numerical - histogram

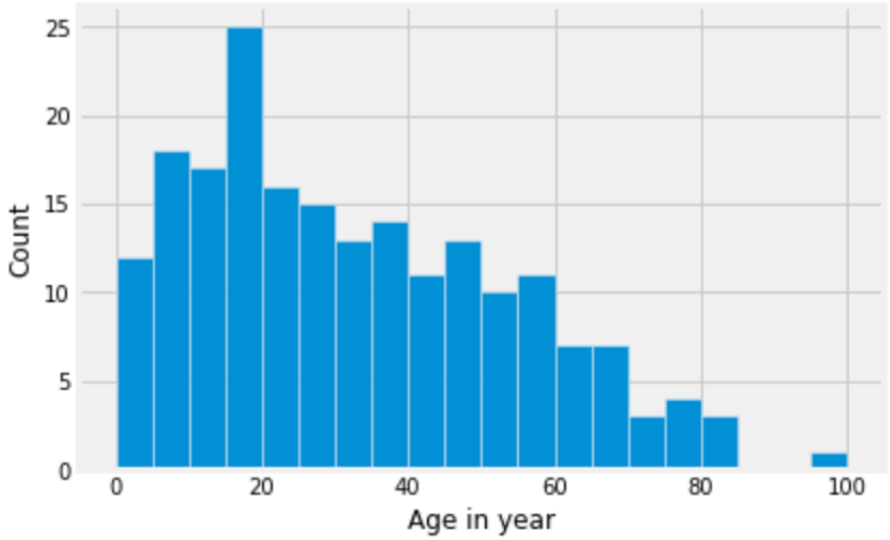
- Binning a numerical variable:
  - count the number of numerical values that lie within ranges, called bins
  - Bins are defined by their lower bounds (inclusive), and upper bounds (exclusive)
  - The upper bound is the lower bound of the next bin



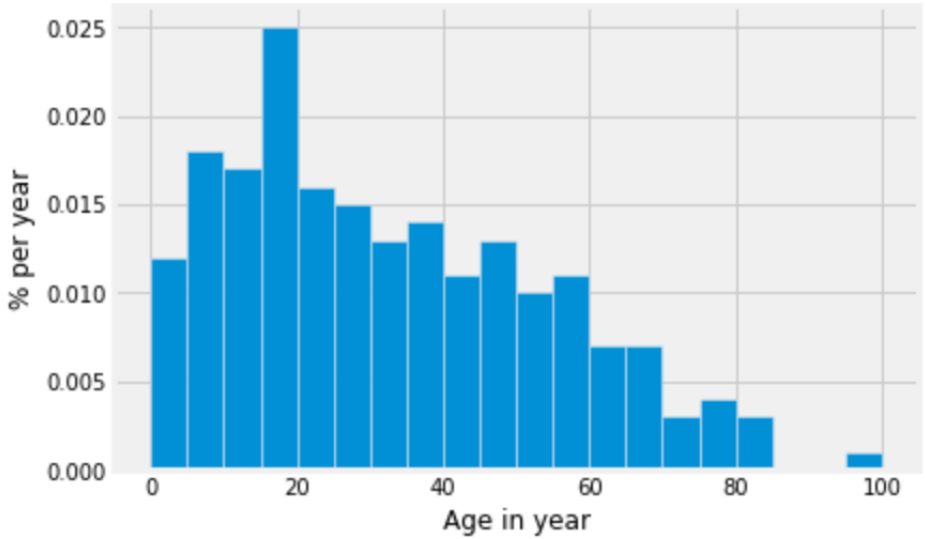
Visualizing distributions

Histogram – two ways

- Y-axis show counts
  - intuitive
  - most common (default)
  - even sized bins only!!!

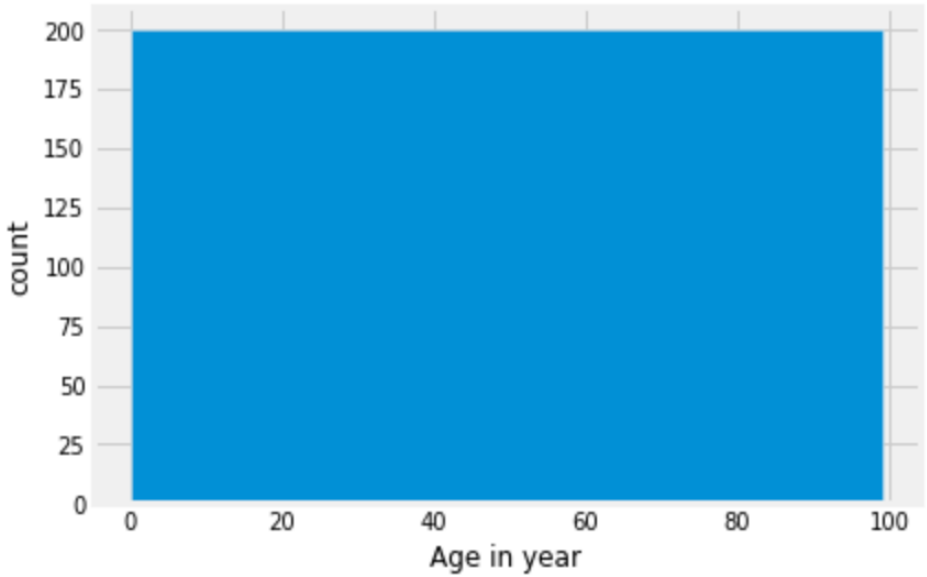


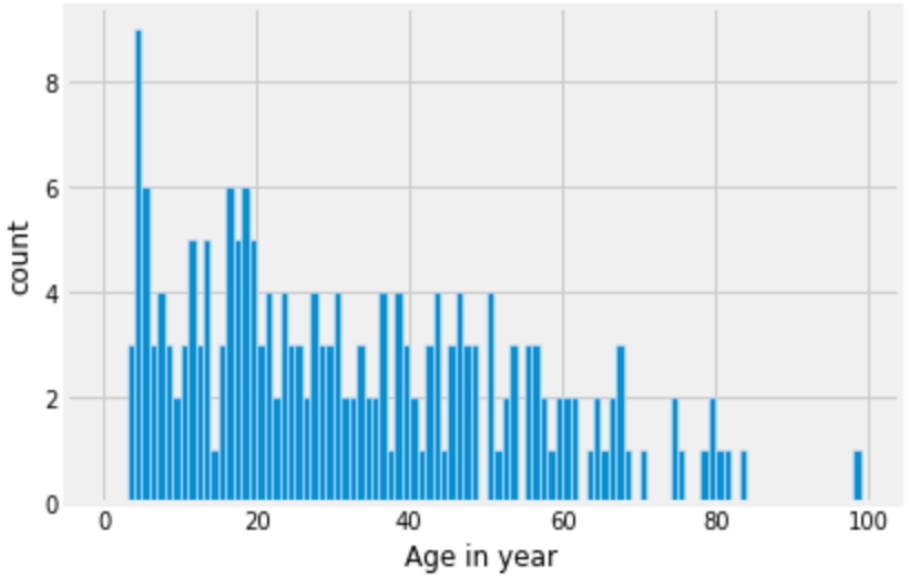
- Y-axis show density
  - area principle
  - height = % / bin width
  - can use uneven bins



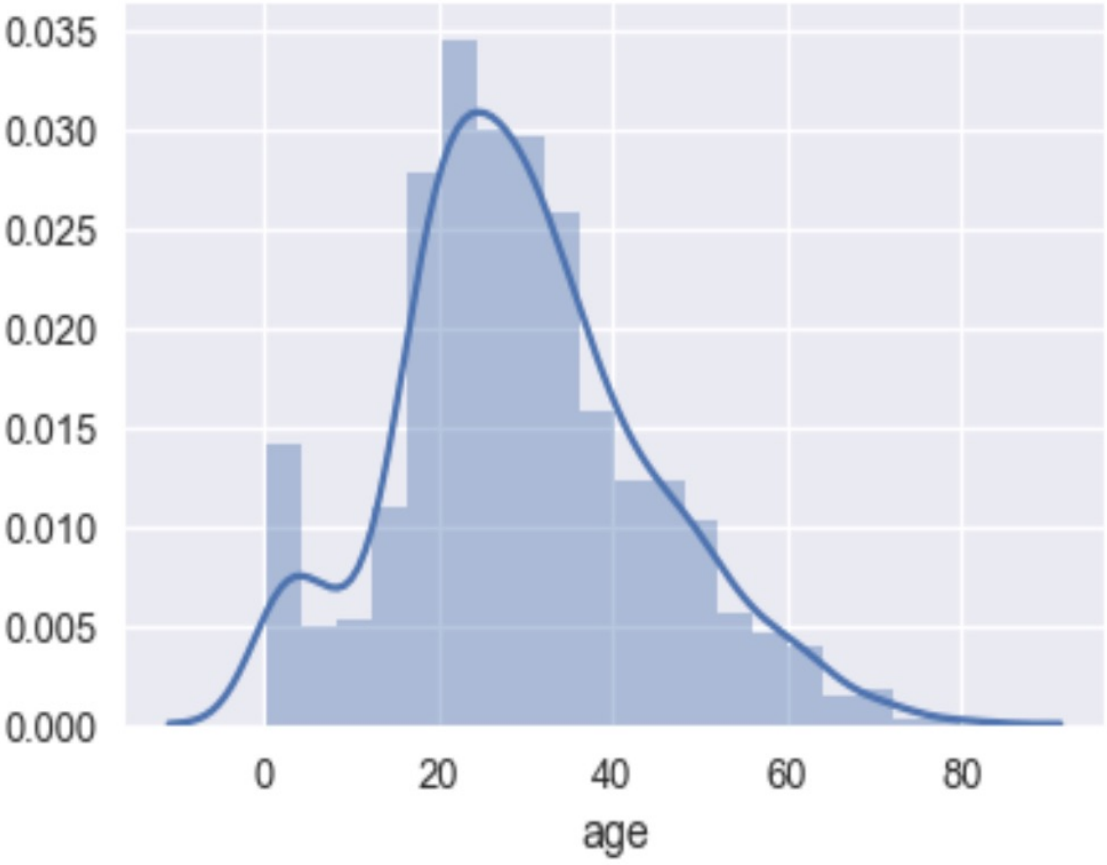
How to choose bin size?

- Think about what bin size encodes?
  - really large bin size v.s. really small bin size
  - trade off: granularity v.s. interpretability





Histogram and kernel density estimate



- Remember height of the bars in histogram measure density
- "Smooth out" the height
- KDE
  - estimate a **probability density function** (pdf) from data
  - DSCI 301 / 303

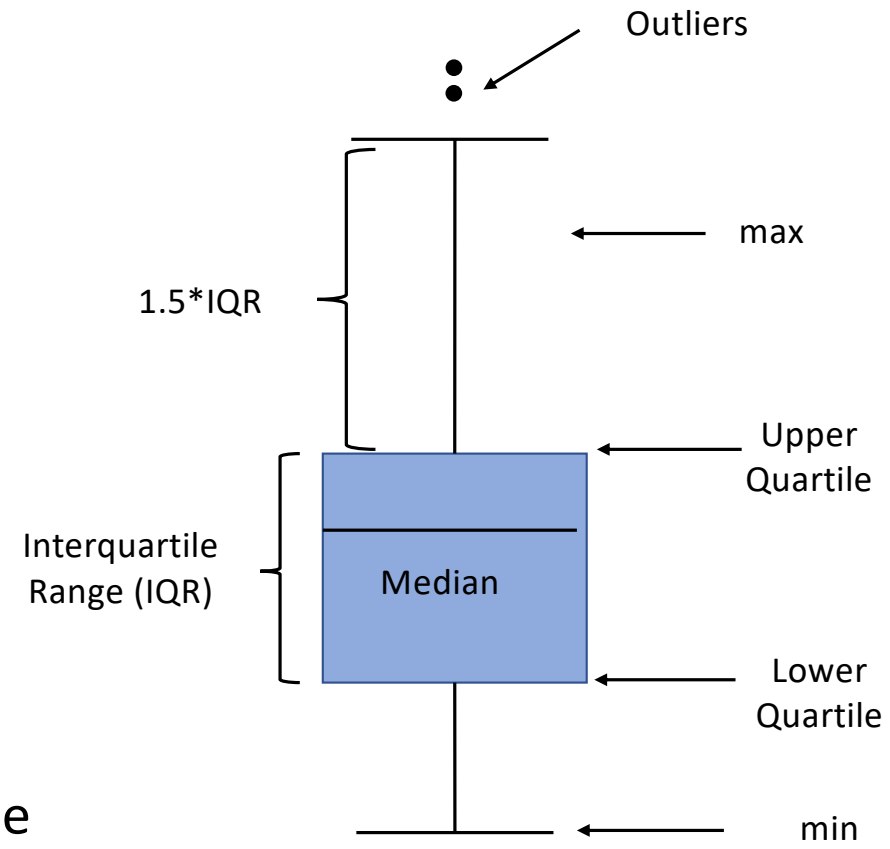
Bar plot vs. Histogram

- Distribution of one categorical variable
- Bars have arbitrary width
- Bars usually do not touch

- Distribution of one numerical variable
- Bar width defined by bins
- No gap between bars

Box plot

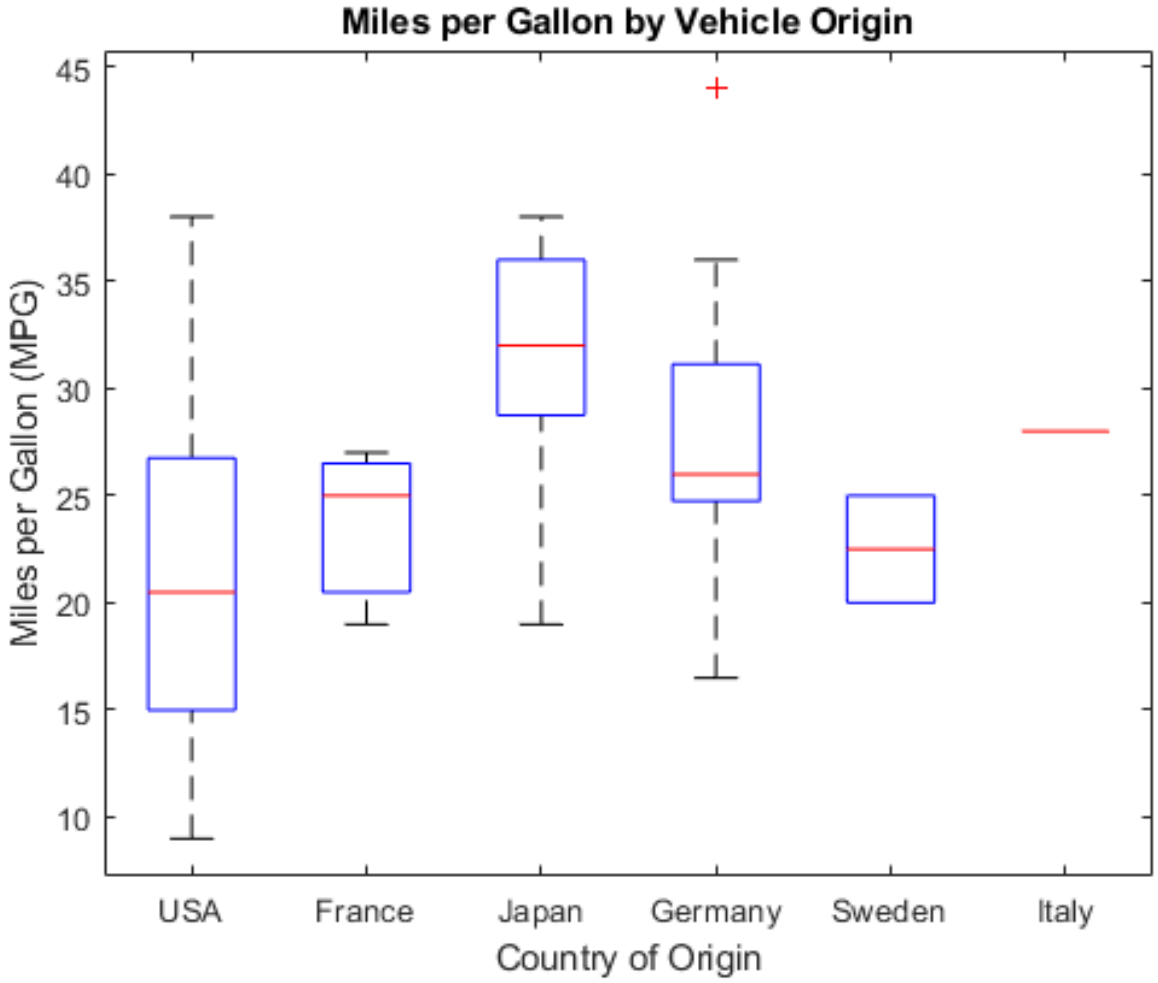
- The pth percentile:
  - Smallest value that has at least p% of the data at or below it



- Quartile: 25% & 75% percentile:
- Median: 50% percentile
- IQR: 75% percentile – 25% percentile
  (the middle 50% of the data)
- Outlier: 1.5*IQR away from lower and upper quartiles

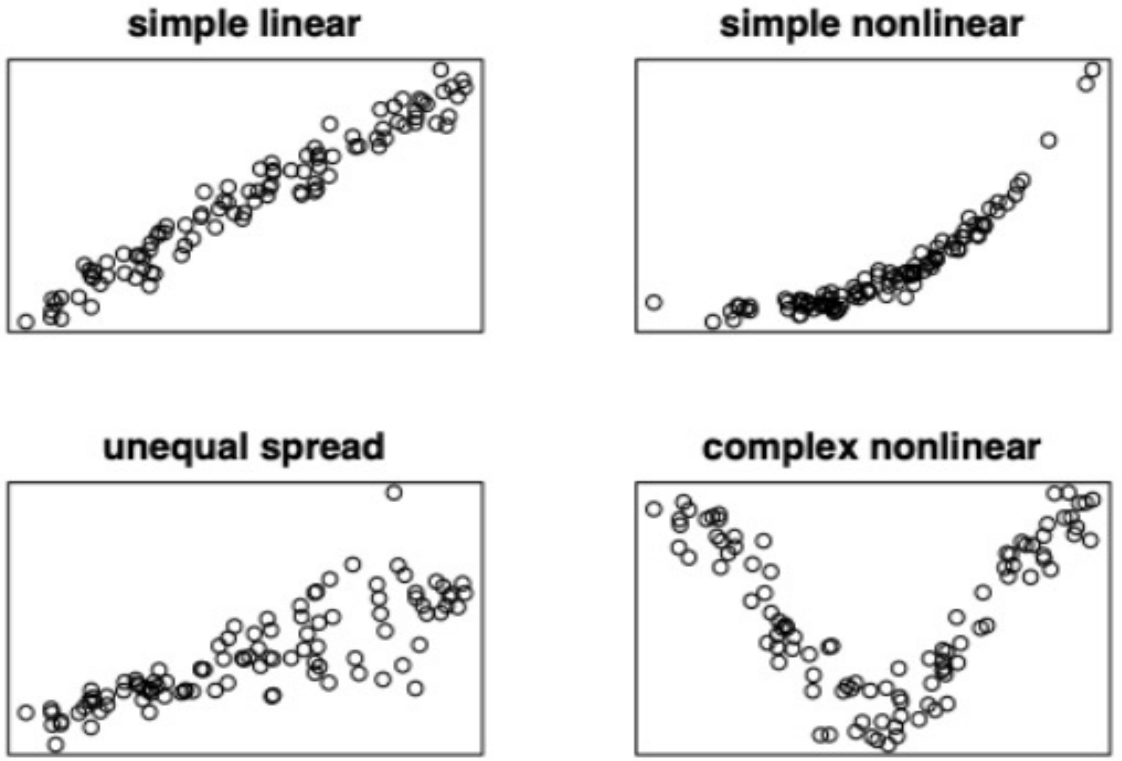
More box plot

- side-by-side box plots
- interaction between a numerical and categorical variables
- comparison of the same measure among different groups

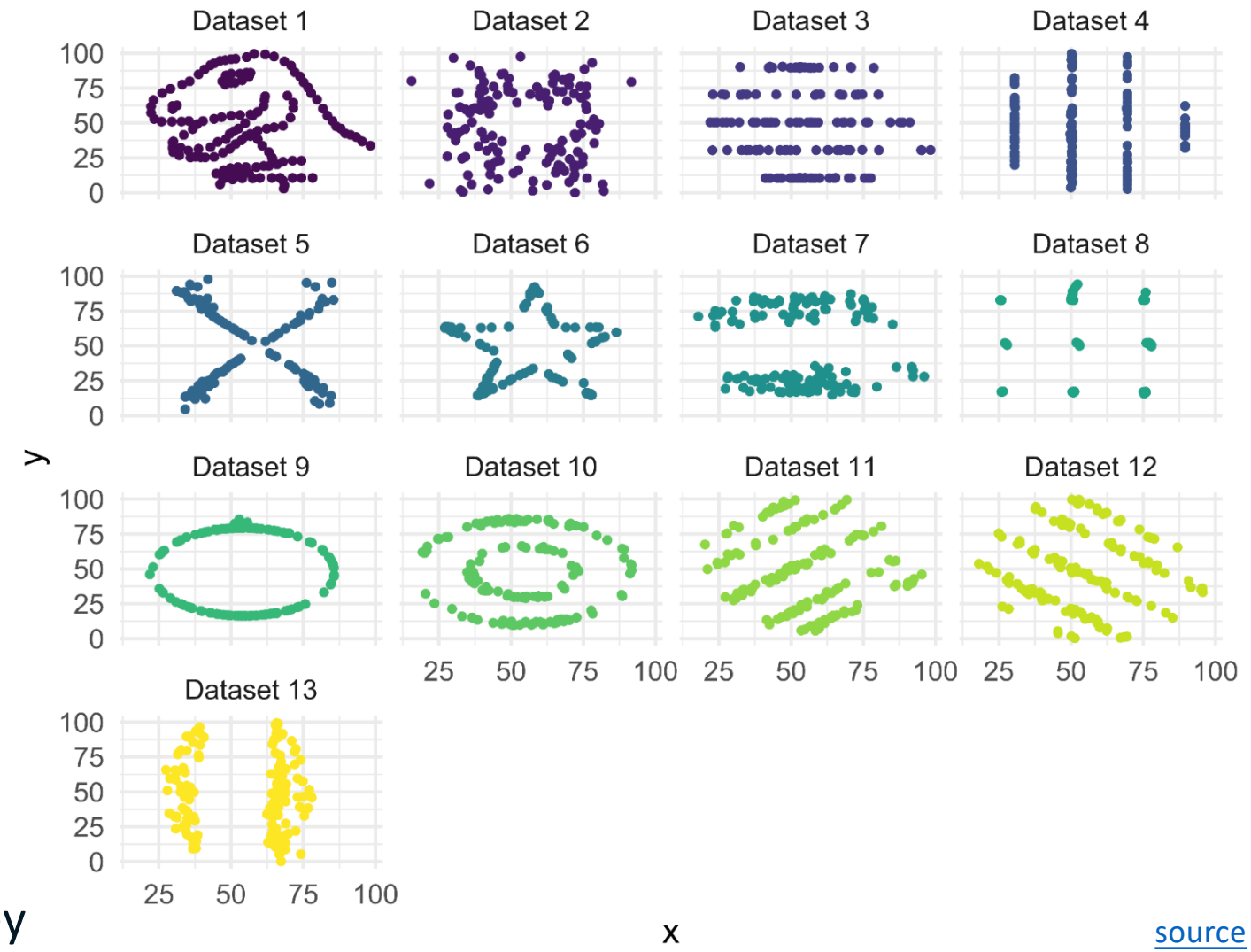


Scatter plot

- Two numerical variables
- Relationship between the pair
- Identify data pattern to help inform model choices



dataset	n	Average x	Average y	St Dev x	St Dev y	Correlation
Dataset 1	142	54.3	47.8	16.8	26.9	-0.1
Dataset 2	142	54.3	47.8	16.8	26.9	-0.1
Dataset 3	142	54.3	47.8	16.8	26.9	-0.1
Dataset 4	142	54.3	47.8	16.8	26.9	-0.1
Dataset 5	142	54.3	47.8	16.8	26.9	-0.1
Dataset 6	142	54.3	47.8	16.8	26.9	-0.1
Dataset 7	142	54.3	47.8	16.8	26.9	-0.1
Dataset 8	142	54.3	47.8	16.8	26.9	-0.1
Dataset 9	142	54.3	47.8	16.8	26.9	-0.1
Dataset 10	142	54.3	47.8	16.8	26.9	-0.1
Dataset 11	142	54.3	47.8	16.8	26.9	-0.1
Dataset 12	142	54.3	47.8	16.8	26.9	-0.1
Dataset 13	142	54.3	47.8	16.8	26.9	-0.1



"The simple graph has brought more information to the data analyst's mind than any other device."

John Tukey

[source]

Line plot

- Two numerical variables
- Often the variable on x-axis is time
- Best to show changes over time and predict future trend

- Let’s look at some plots of life expectancy and fertility rate per women over time for some countries.
- Can you guess which country?

More plots of real data to look at

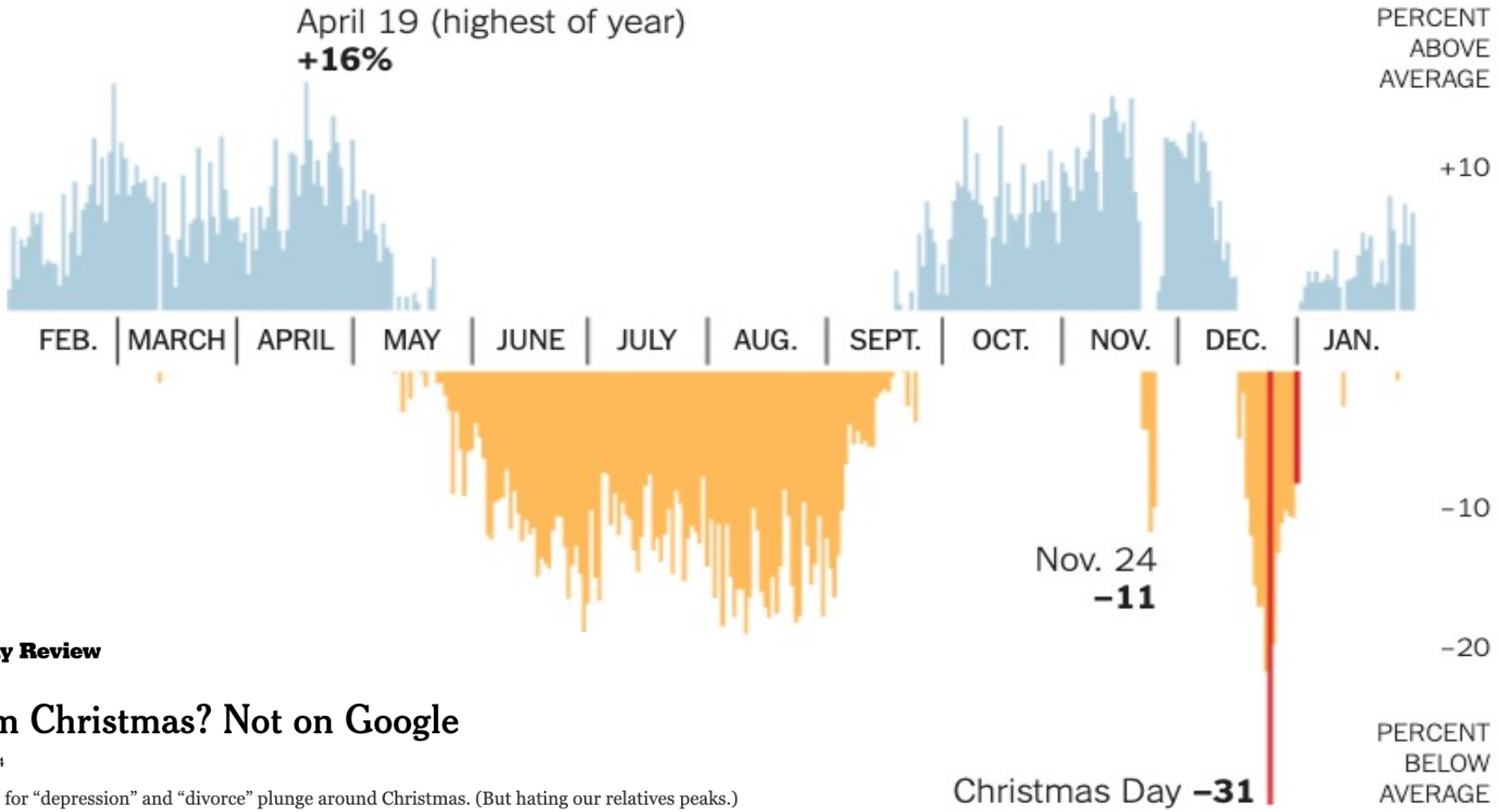
- Can you tell what kind of plot it is?

- Can you read the plot and explain what's going on?

- What do you think makes a good plot?

depression

https://www.nytimes.com/interactive/2014/12/21/sunday-review/glum-christmas-not-on-google.html?mtrref=www.google.com&assetType=REGIWALL



Sunday Review

Glum Christmas? Not on Google

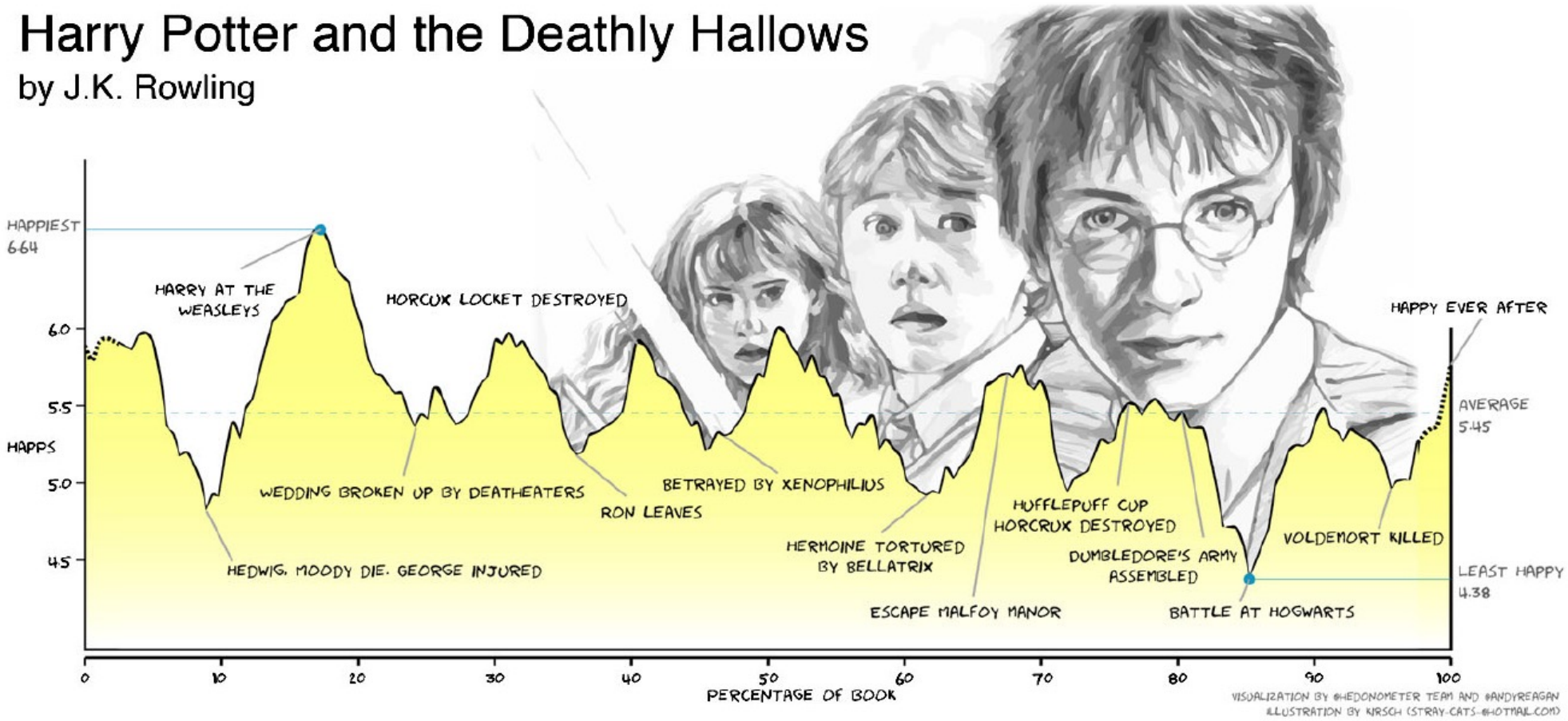
DEC. 20, 2014

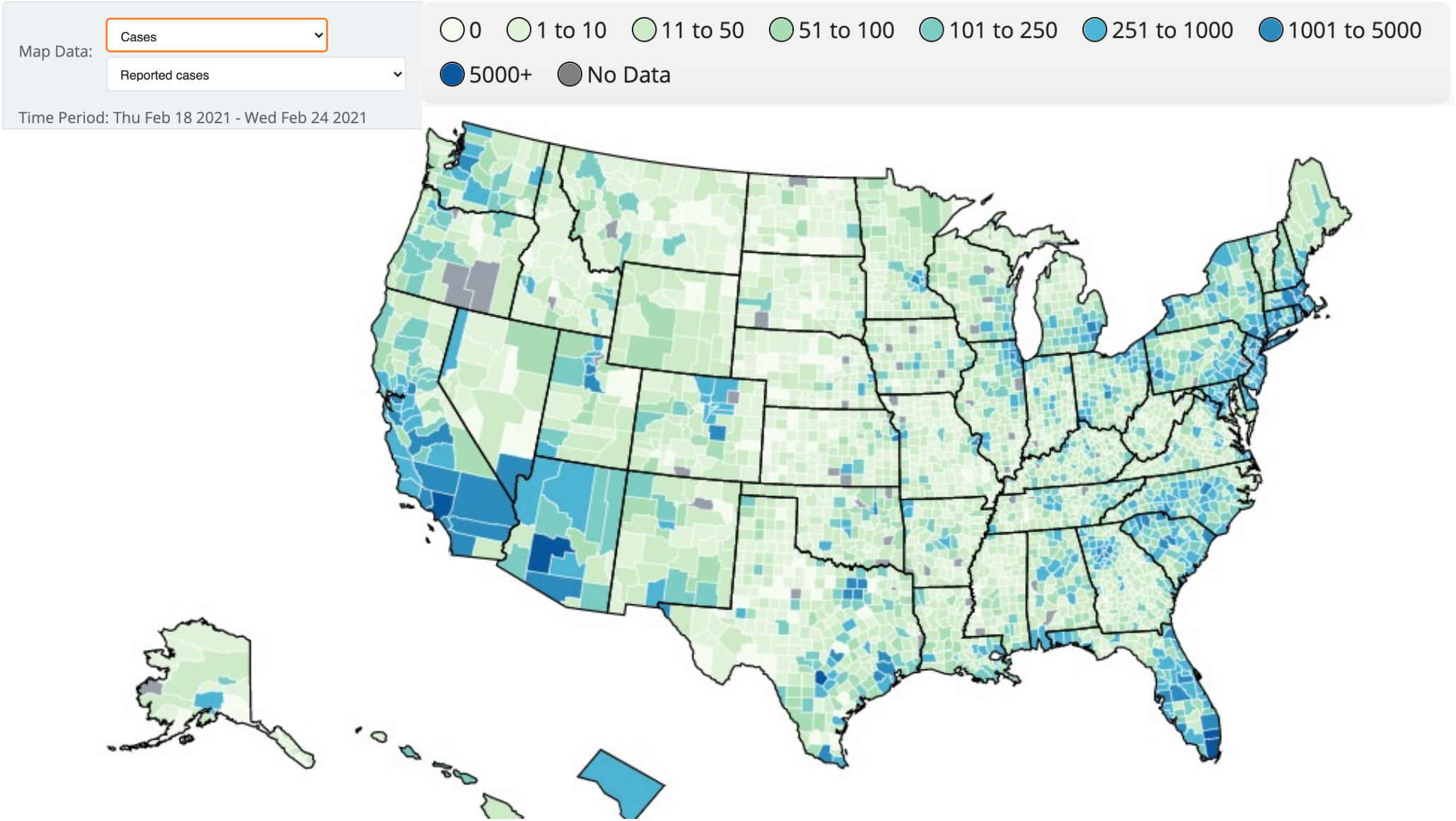
Searches for “depression” and “divorce” plunge around Christmas. (But hating our relatives peaks.)

The emotional arcs of stories are dominated by six basic shapes

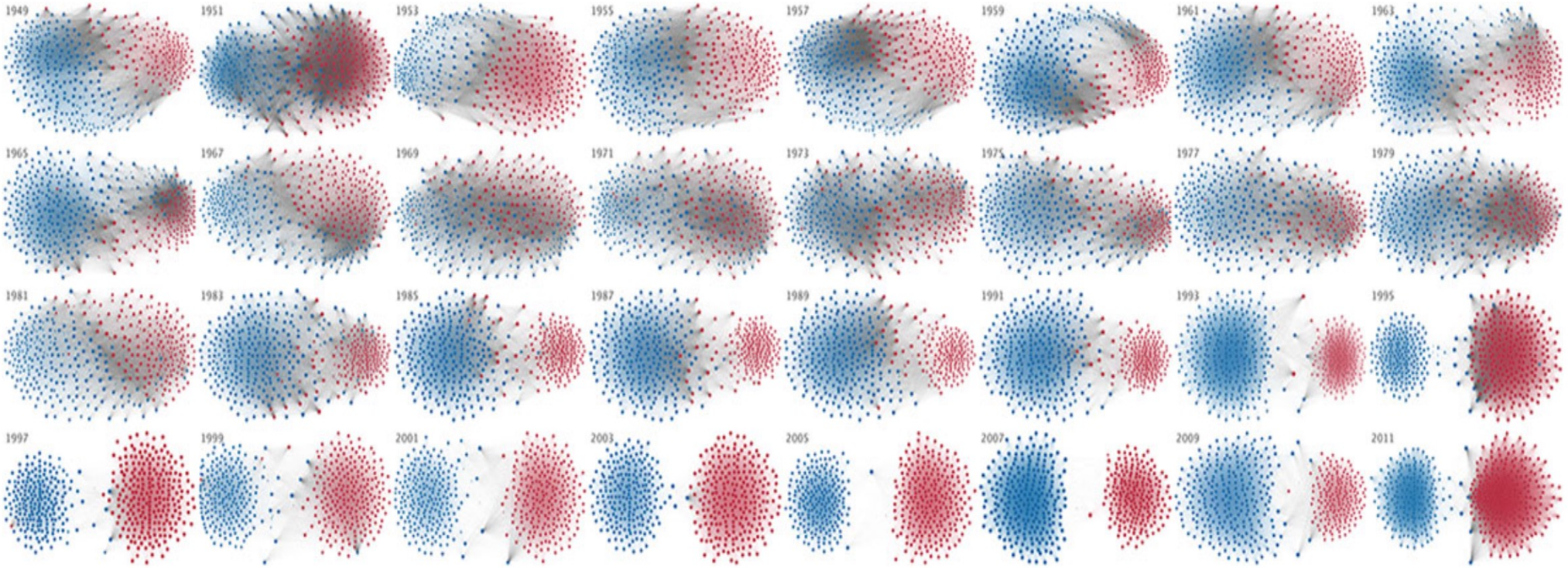
Andrew J Reagan[1*], Lewis Mitchell[2], Dilan Kiley[1], Christopher M Danforth[1] and Peter Sheridan Dodds[1]

Reagan et al. *EPJ Data Science* (2016) 5:31
DOI 10.1140/epjds/s13688-016-0093-1
EPJ.org
REGULAR ARTICLE

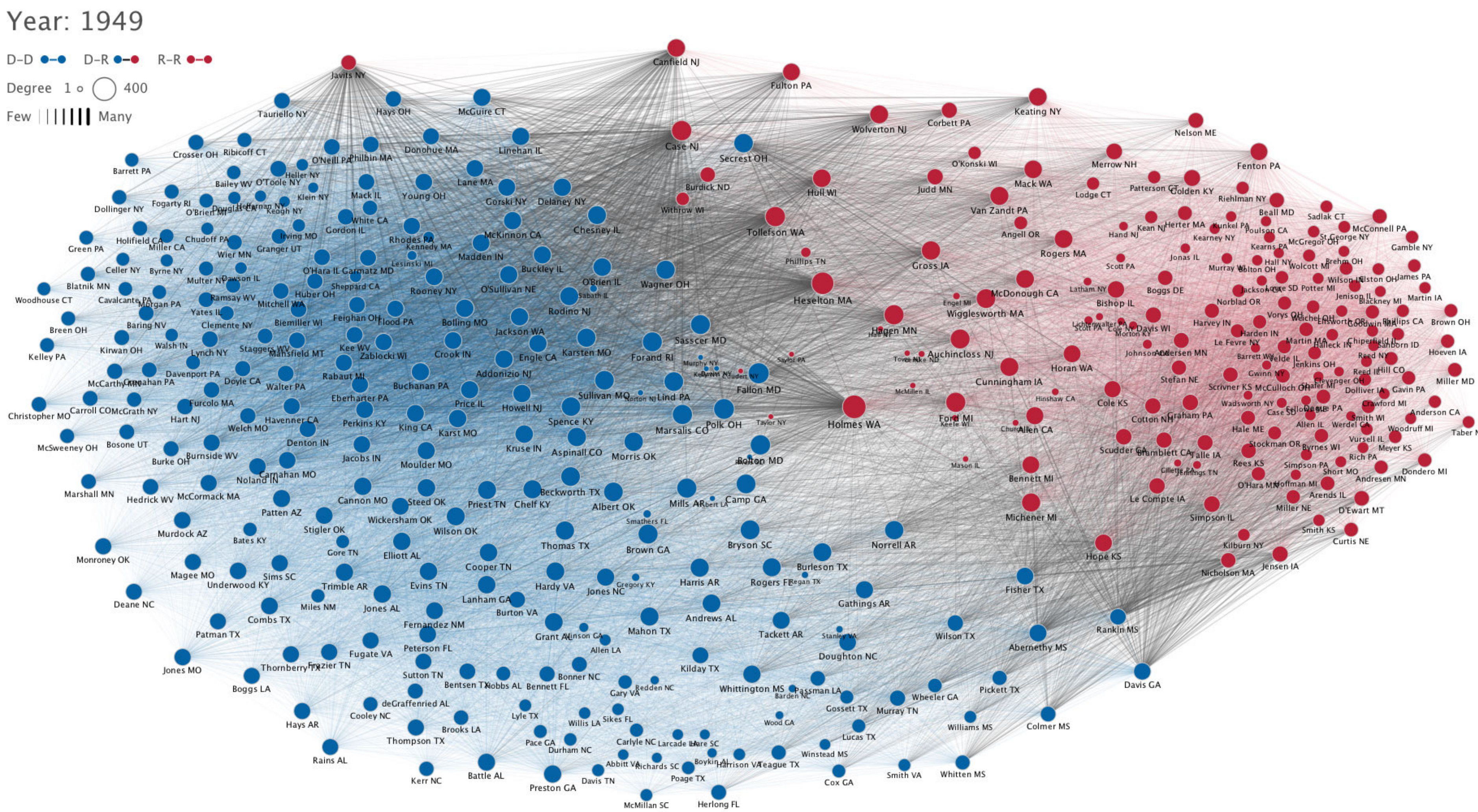



Map Data:
Cases
Reported cases
Time Period: Thu Feb 18 2021 - Wed Feb 24 2021
0
1 to 10
11 to 50
51 to 100
101 to 250
251 to 1000
1001 to 5000
5000+
No Data

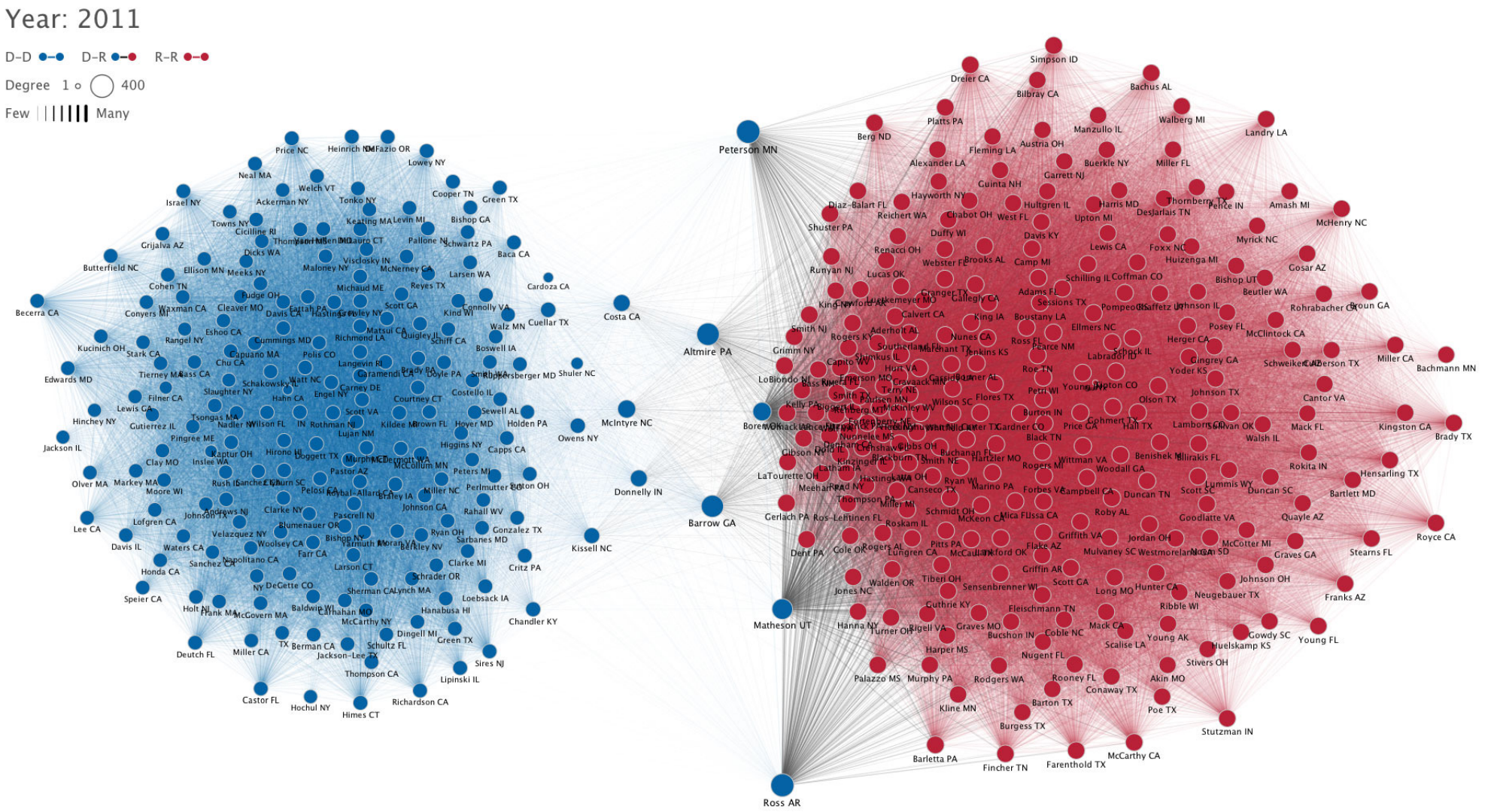
The rise of partisanship in the US House of Representatives



https://www.mamartino.com/projects/rise_of_partisanship/index.html



Year: 1949
D–D
D–R
R–R
Degree 1
400
Few
Many
Javits NY
Canfield NJ
Fulton PA
Keating NY
Wolverton NJ
Corbett PA
Nelson ME
Tauriello NY
Hays OH
McGuire CT
Case NJ
Secrest OH
Fenton PA
Merrow NH
O'Konski WI
Crosser OH
Ribicoff CT
Philbin MA
Donohue MA
Linehan IL
Burdick ND
Hull WI
Judd MN
Mack WA
Lodge CT
Patterson CT
Golden KY
Barrett PA
Lane MA
Withrow WI
Van Zandt PA
Riehlman NY
Dollinger NY
Fogarty RI
Mack IL
Young OH
Gorski NY
Delaney NY
Beall MD
Chesney IL
Tollefson WA
Angell OR
Herter MA
Sadlak CT
McConnell PA
Holifield CA
Chudoff PA
Miller CA
White CA
Gordon IL
Rhodes PA
McKinnon CA
Rogers MA
Gamble NY
Green PA
Wier MN
Granger UT
Madden IN
Buckley IL
Phillips TN
Gross IA
Celler NY
Byrne NY
O'Hara IL
Garmatz MD
O'Brien IL
Wagner OH
Heselton MA
McDonough CA
Bishop IL
Boggs DE
Blatnik MN
Multer NY
Dawson IL
Huber OH
Sheppard CA
Rooney NY
O'Sullivan NE
Woodhouse CT
Cavalcante PA
Morgan PA
Ramsay WV
Mitchell WA
Rodino NJ
Hagen MN
Wigglesworth MA
Norblad OR
Martin IA
Breen OH
Yates IL
Clemente NY
Biemiller WI
Feighan OH
Flood PA
Bolling MO
Jackson WA
Sasscer MD
Davis WI
Harvey IN
Brown OH
Baring NV
Walsh IN
Kirwan OH
Lynch NY
Staggers WV
Mansfield MT
Kee WV
Zablocki WI
Crook IN
Engle CA
Karsten MO
Forand RI
Auchincloss NJ
Horan WA
Andersen MN
Le Fevre NY
Hoeven IA
Kelley PA
Davenport PA
McCarthy MN
Crosahan PA
Doyle CA
Rabaut MI
Buchanan PA
Addonizio NJ
Fallon MD
Cunningham IA
Stefan NE
Miller MD
Walter PA
Eberharter PA
Sullivan MO
Lind PA
Cole KS
Holmes WA
Ford MI
Christopher MO
Carroll CO
McGrath NY
Furcolo MA
Price IL
Howell NJ
Polk OH
Allen CA
Cotton NH
Graham PA
Anderson CA
Taber NY
Hart NJ
Welch MO
Havenner CA
Perkins KY
King CA
Karst MO
Spence KY
Marsalis CO
McSweeney OH
Bosone UT
Denton IN
Kruse IN
Aspinall CO
Morris OK
Bolton MD
Scudder CA
Hale ME
Burke OH
Burnside WV
Jacobs IN
Moulder MO
Bennett MI
Le Compte IA
Marshall MN
Noland IN
Carnahan MO
Hedrick WV
McCormack MA
Cannon MO
Steed OK
Priest TN
Chelf KY
Beckworth TX
Albert OK
Mills AR
Camp GA
Michener MI
Simpson IL
D'Ewart MT
Patten AZ
Stigler OK
Wickersham OK
Wilson OK
Curtis NE
Murdock AZ
Bates KY
Thomas TX
Brown GA
Bryson SC
Norrell AR
Hope KS
Kilburn NY
Jensen IA
Monroney OK
Gore TN
Elliott AL
Cooper TN
Burleson TX
Nicholson MA
Magee MO
Sims SC
Trimble AR
Evins TN
Hardy VA
Jones NC
Harris AR
Rogers FL
Fisher TX
Underwood KY
Deane NC
Miles NM
Jones AL
Lanham GA
Burton VA
Gathings AR
Combs TX
Fernandez NM
Mahon TX
Andrews AL
Tackett AR
Rankin MS
Patman TX
Grant AL
Wilson TX
Abernethy MS
Peterson FL
Fugate VA
Doughton NC
Jones MO
Thornberry TX
Frazier TN
Kilday TX
Davis GA
Boggs LA
Sutton TN
Bentsen TX
Hobbs AL
Bennett FL
Bonner NC
Whittington MS
Pickett TX
Gary VA
Passman LA
Wheeler GA
deGraffenried AL
Gossett TX
Murray TN
Hays AR
Cooley NC
Lyle TX
Brooks LA
Willis LA
Sikes FL
Williams MS
Colmer MS
Thompson TX
Carlyle NC
Lucas TX
Pace GA
Durham NC
Winstead MS
Rains AL
Abbitt VA
Richards SC
Teague TX
Battle AL
Kerr NC
Preston GA
Davis TN
Poage TX
Cox GA
Smith VA
Whitten MS
McMillan SC
Herlong FL


Year: 2011
D-D
D-R
R-R
Degree 1 400
Few Many
Peterson MN
Altmire PA
Barrow GA
Matheson UT
Ross AR
Costa CA
McIntyre NC
Donnelly IN
Kissell NC
Price NC
Heinrich NM
Fazio OR
Lowey NY
Neal MA
Welch VT
Israel NY
Ackerman NY
Tonko NY
Cooper TN
Green TX
Towns NY
Keating MA
Levin MI
Bishop GA
Cicilline RI
Grijalva AZ
Dicks WA
Pallone NJ
Schwartz PA
Baca CA
Butterfield NC
Ellison MN
Meeks NY
Maloney NY
Visclosky IN
McNerney CA
Larsen WA
Cohen TN
Michaud ME
Reyes TX
Cardoza CA
Becerra CA
Conyers MI
Waxman CA
Fudge OH
Cleaver MO
Davis CA
Scott GA
Connolly VA
Kind WI
Cuellar TX
Walz MN
Kucinich OH
Rangel NY
Eshoo CA
Cummings MD
Richmond LA
Quigley IL
Schiff CA
Boswell IA
Stark CA
Capuano MA
Polis CO
Langevin RI
Chu CA
Tierney MA
Lynch MA
Caramendi CA
Doyle PA
Rothenberger MD
Shuler NC
Edwards MD
Watt NC
Schakowsky IL
Carney DE
Engel NY
Filner CA
Hahn CA
Courtney CT
Costello IL
Lewis GA
Slaughter NY
Sewell AL
Hinchey NY
Tsongas MA
Nadler NY
Wilson FL
Rothman NJ
Kildee MI
Brown FL
Hover MD
Holden PA
Gutierrez IL
Jackson IL
Pingree ME
Higgins NY
Capps CA
Clay MO
Inslee WA
Doggett TX
Olver MA
Markey MA
Pastor AZ
Peters MI
Moore WI
Pelosi CA
Miller NC
Perlmutter CO
Boyton OH
Lee CA
Lofgren CA
Andrews NJ
Clarke NY
Pascrell NJ
Johnson CA
Rahall WV
Velazquez NY
Blumenauer OR
Ryan OH
Gonzalez TX
Davis IL
Waters CA
Woolsey CA
Bishop NY
Sarbanes MD
Farr CA
Berkley NV
Napolitano CA
Sanchez CA
Clarke MI
Larson CT
Critz PA
Honda CA
Schrader OR
Speier CA
DeGette CO
Sherman CA
Lynch MA
Loebsack IA
Holt NJ
Frank MA
McGovern MA
Baldwin WI
Lahahan MO
Hanabusa HI
Chandler KY
McCarthy NY
Dingell MI
Green TX
Deutch FL
Miller CA
Berman CA
Schultz FL
Jackson-Lee TX
Sires NJ
Thompson CA
Lipinski IL
Castor FL
Hochul NY
Himes CT
Richardson CA
Simpson ID
Dreier CA
Bilbray CA
Bachus AL
Platts PA
Berg ND
Manzullo IL
Walberg MI
Landry LA
Fleming LA
Austria OH
Alexander LA
Buerkle NY
Miller FL
Garrett NJ
Guinta NH
Hayworth NY
Diaz-Balart FL
Hultgren IL
Harris MD
Thornberry TX
Amash MI
Reichert WA
Chabot OH
West FL
Upton MI
Desjarlais TN
Pence IN
McHenry NC
Shuster PA
Duffy WI
Davis KY
Myrick NC
Renacci OH
Lewis CA
Foxx NC
Runyan NJ
Webster FL
Brooks AL
Camp MI
Huizenga MI
Gosar AZ
Lucas OK
Schilling IL
Coffman CO
Bishop UT
Beutler WA
Adams FL
Sessions TX
Pompeo KS
Saffetz UT
Johnson IL
Rohrabacher CA
Broun GA
Calvert CA
King IA
Boustany LA
Ellmers NC
Posey FL
McClintock CA
Smith NJ
Rogers KY
Aderholt AL
Nunes CA
Ross FL
Herger CA
Grimm NY
Southerland FL
Merchant TX
Jenkins KS
Pearce NM
Schock IL
Labrador ID
Gingrey GA
Schweikert AZ
Culberson TX
Miller CA
Bachmann MN
Capito WV
Holt VA
Roe TN
Yoder KS
LoBiondo NJ
Emerson MO
Crawford AR
Petri WI
Upton CO
Johnson TX
Cantor VA
Smith TX
Paulsen MN
Flores TX
Olson TX
Kelly PA
McKinley WV
Wilson SC
Burton IN
Gohmert TX
Mack FL
Kingston GA
Brady TX
Nunnelee MS
Price GA
Hall TX
Lamborn CO
Sullivan OK
Walsh IL
Black TN
Gibson NY
Crenshaw FL
Gibbs OH
Buchanan FL
Benishek MI
Hillrakis FL
Rokita IN
Blackburn TN
Smith NE
Hartzler MO
Wittman VA
Rogers MI
Woodall GA
LaTourette OH
Latham IA
Hastings WA
Ryan WI
Meehan PA
Canseco TX
Marino PA
Forbes VA
Campbell CA
Duncan TN
Scott SC
Duncan SC
Lummis WY
Hensarling TX
Bartlett MD
Thompson PA
Miller MI
Gerlach PA
Ros-Lehtinen FL
Schmidt OH
Mica FL
Issa CA
Roby AL
Goodlatte VA
Quayle AZ
Royce CA
Roskam IL
McKeon CA
Griffith VA
Jordan OH
Dent PA
Cole OK
Rogers AL
Lungren CA
Pitts PA
McCaul TX
Crawford OK
Flake AZ
Mulvaney SC
Westmoreland GA
McCotter MI
Graves GA
Stearns FL
Griffin AR
Tiberi OH
Scott GA
Johnson OH
Walden OR
Sensenbrenner WI
Long MO
Hunter CA
Neugebauer TX
Franks AZ
Jones NC
Guthrie KY
Ribble WI
Fleischmann TN
Hanna NY
Turner OH
Rigell VA
Graves MO
Mack CA
Coble NC
Young AK
Gowdy SC
Young FL
Bucshon IN
Scalise LA
Huelskamp KS
Harper MS
Nugent FL
Stivers OH
Palazzo MS
Murphy PA
Rodgers WA
Rooney FL
Akin MO
Conaway TX
Kline MN
Barton TX
Poe TX
Burgess TX
Stutzman IN
Barletta PA
Fincher TN
Farenthold TX
McCarthy CA

Recipe for good data visualization

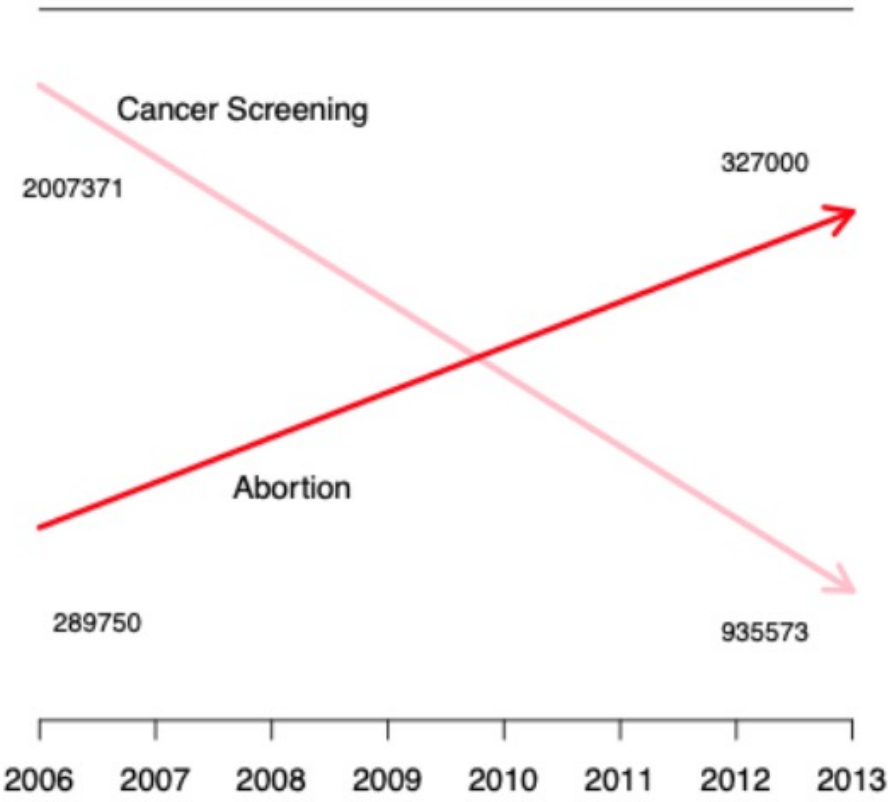
- Good understanding of your data and your purpose!
  - What does your data tell?
  - What do you want to tell?

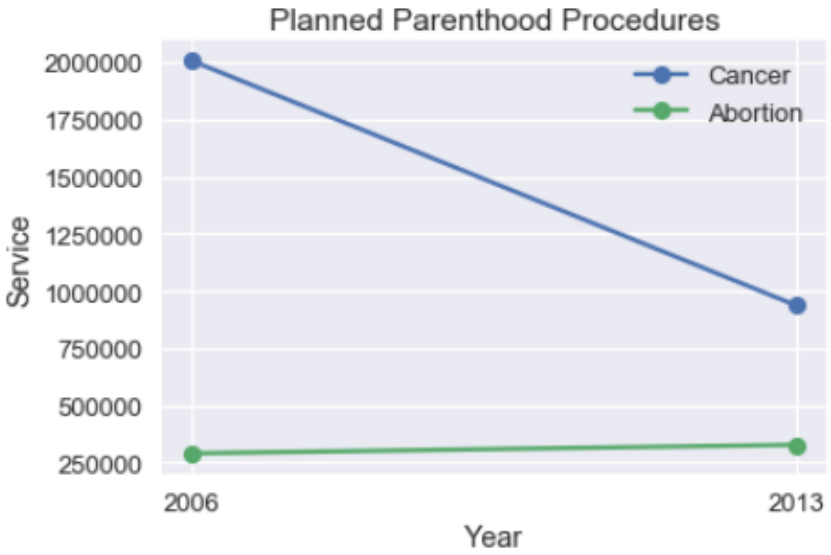
- Some principles

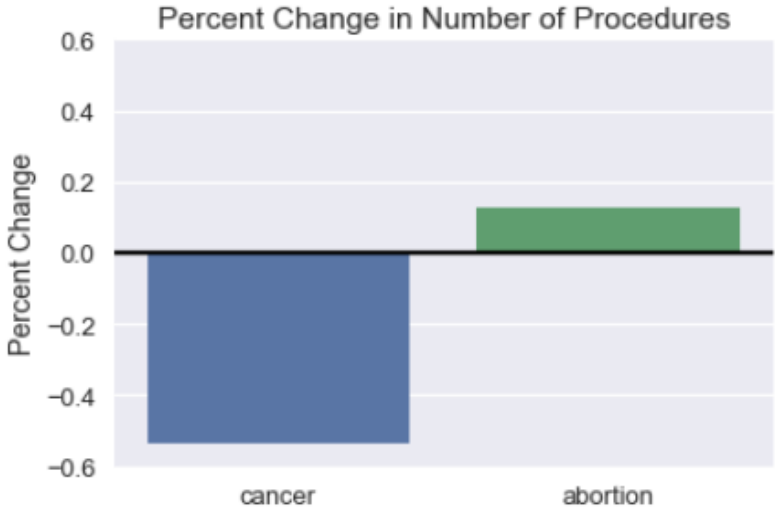
  Scale Comparison Perception Transformation Context

- A dash of creativity

Scale

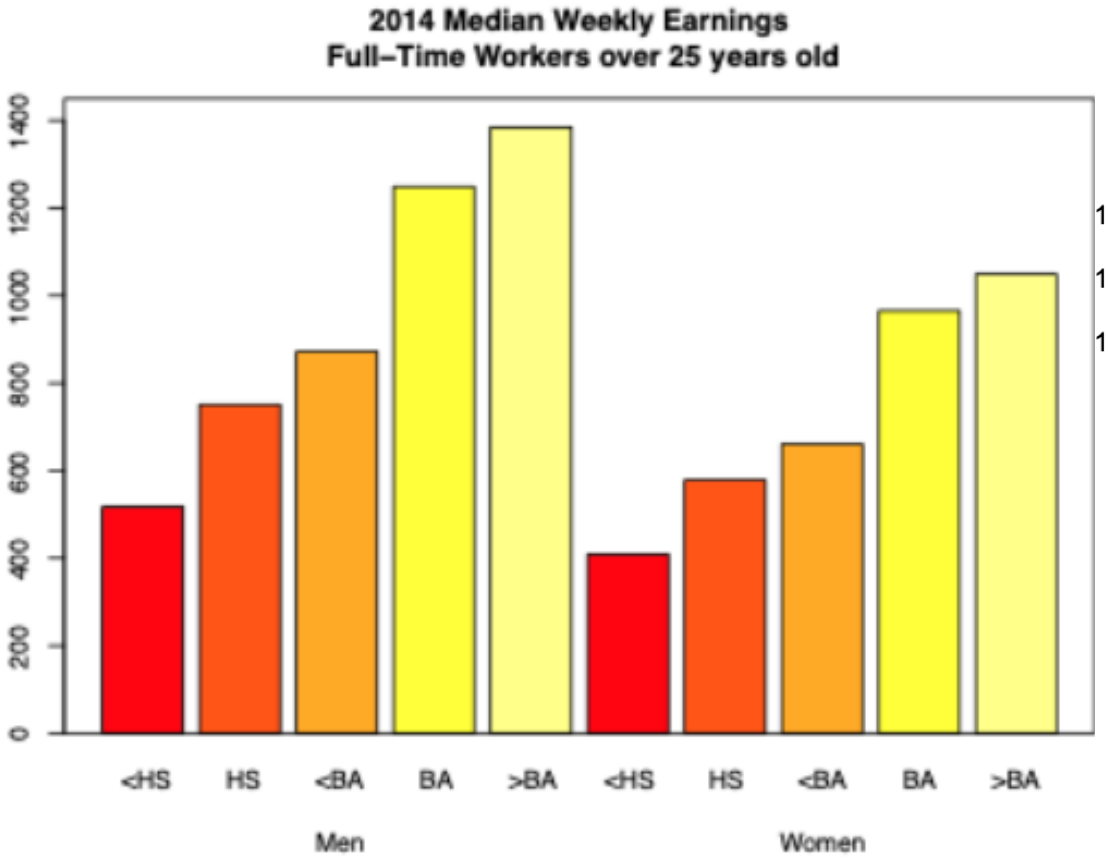


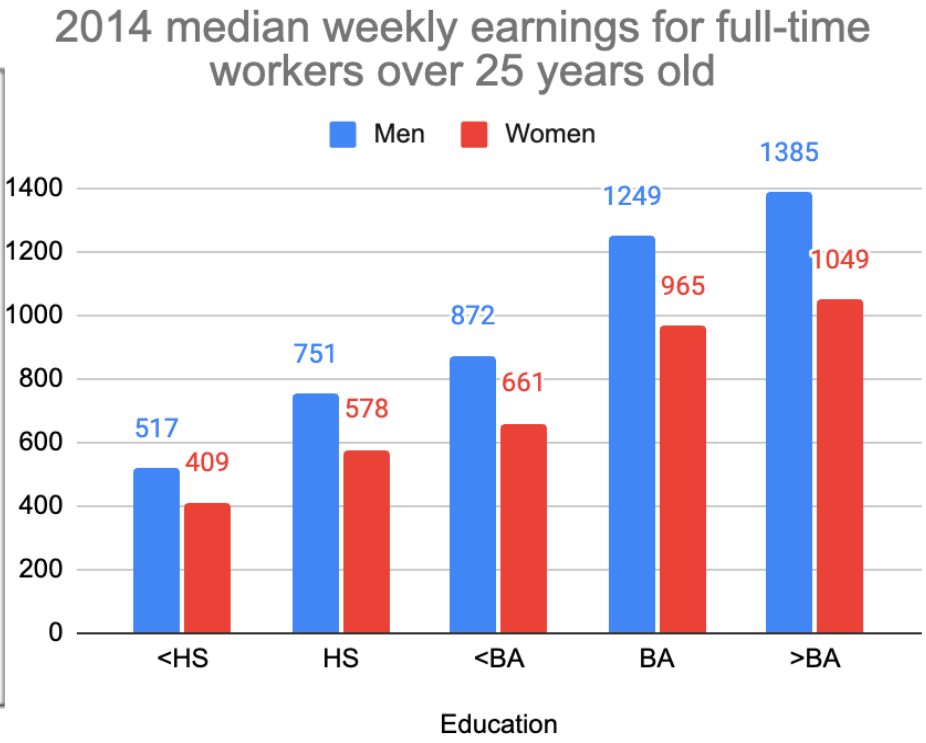




https://oversight.house.gov/interactivepage/plannedparenthood/

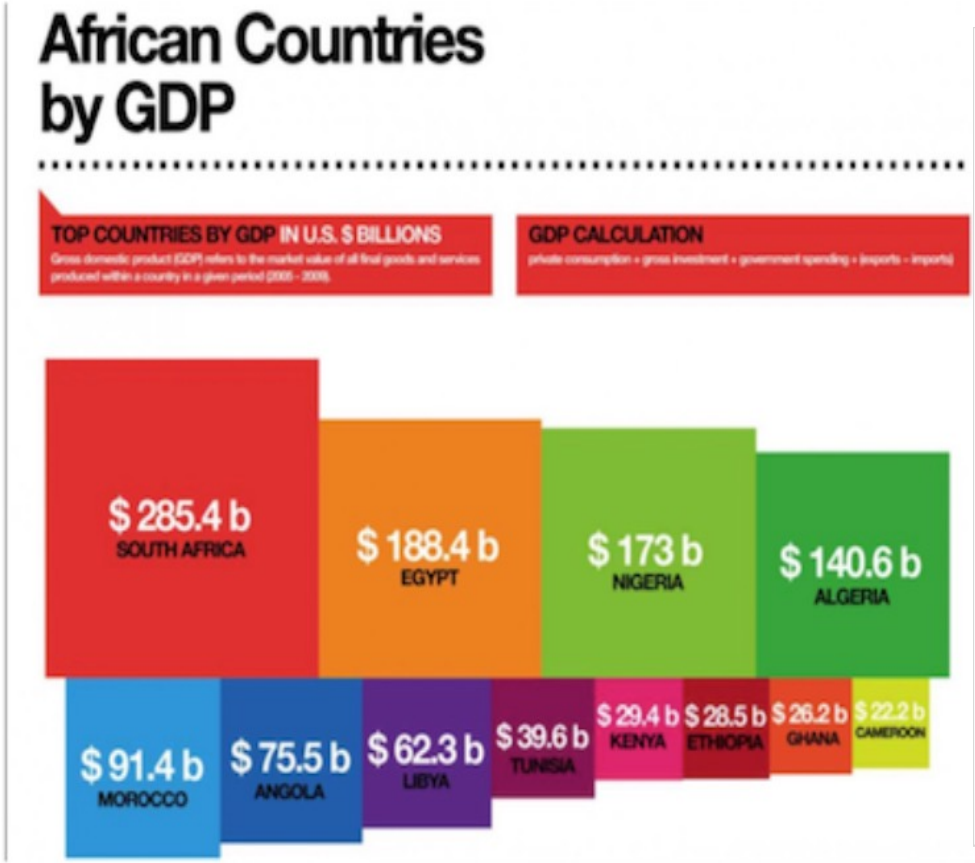
Comparison

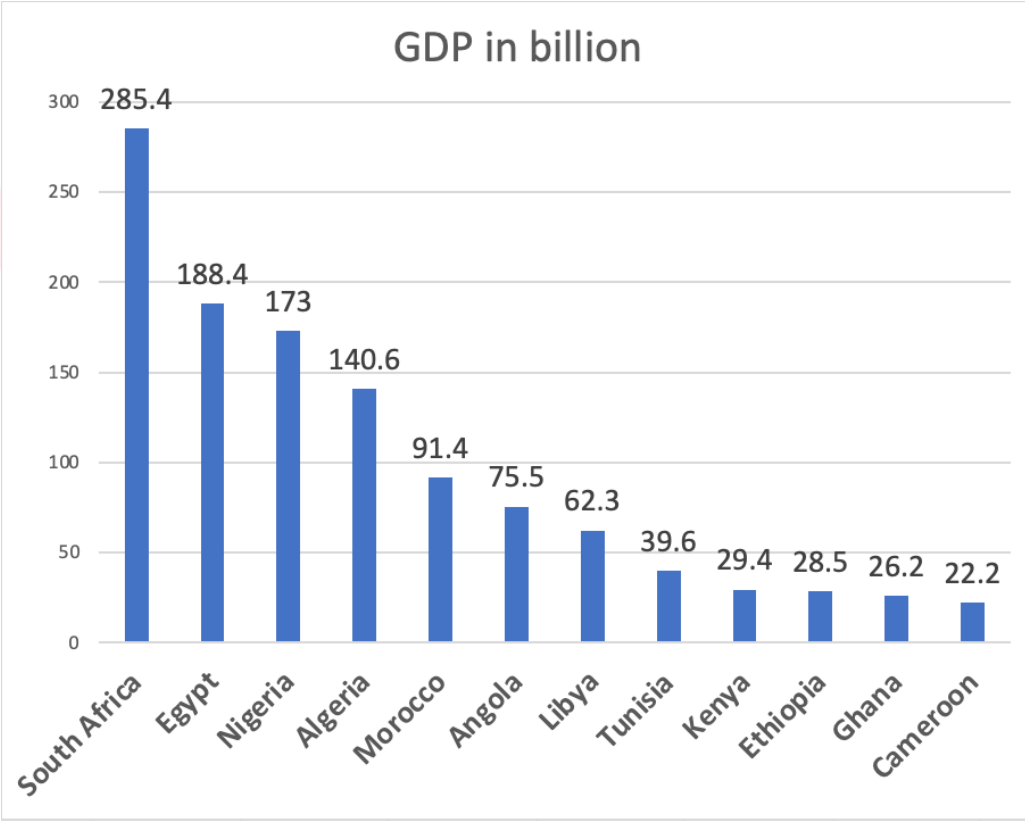




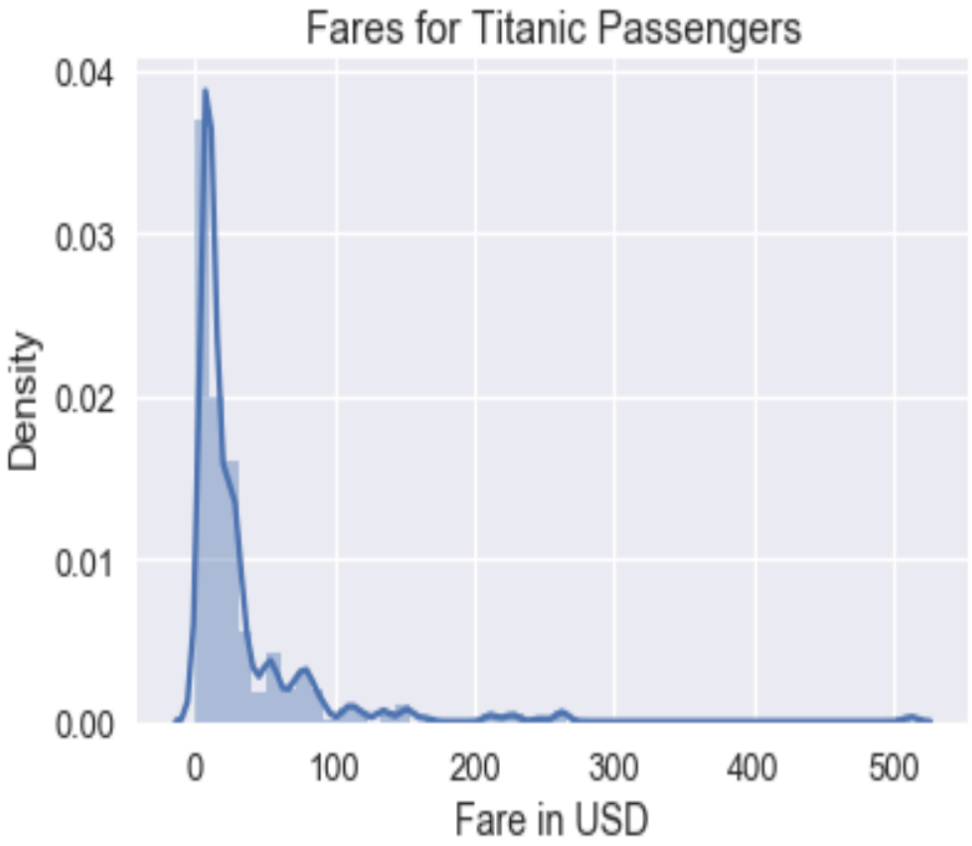
Bureau of Labor Statistics

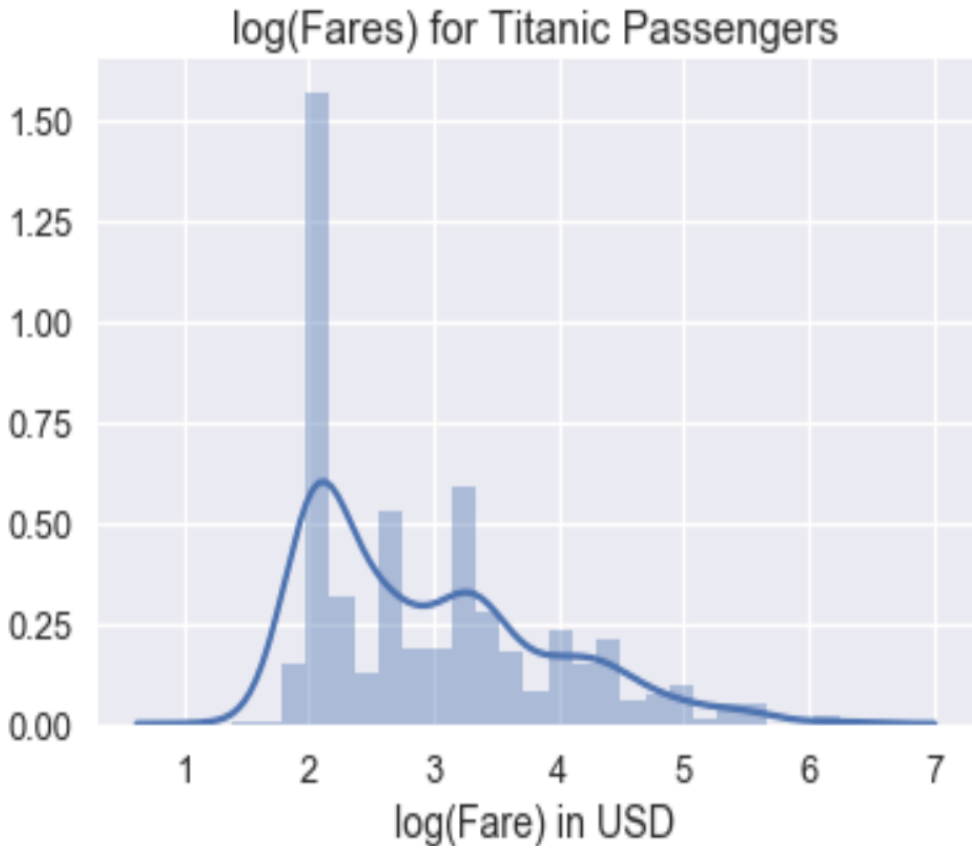
Perception


African Countries by GDP
TOP COUNTRIES BY GDP IN U.S. $ BILLIONS
GDP CALCULATION
$ 285.4 b
SOUTH AFRICA
$ 188.4 b
EGYPT
$ 173 b
NIGERIA
$ 140.6 b
ALGERIA
$ 91.4 b
MOROCCO
$ 75.5 b
ANGOLA
$ 62.3 b
LIBYA
$ 39.6 b
TUNISIA
$ 29.4 b
KENYA
$ 28.5 b
ETHIOPIA
$ 26.2 b
GHANA
$ 22.2 b
CAMEROON


GDP in billion
300
250
200
150
100
50
0
285.4
188.4
173
140.6
91.4
75.5
62.3
39.6
29.4
28.5
26.2
22.2
South Africa
Egypt
Nigeria
Algeria
Morocco
Angola
Libya
Tunisia
Kenya
Ethiopia
Ghana
Cameroon

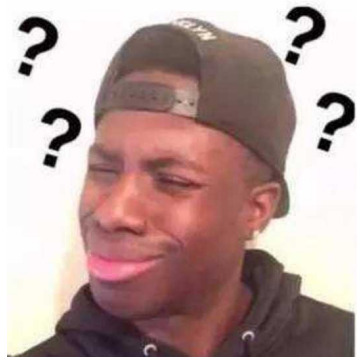
Transformation





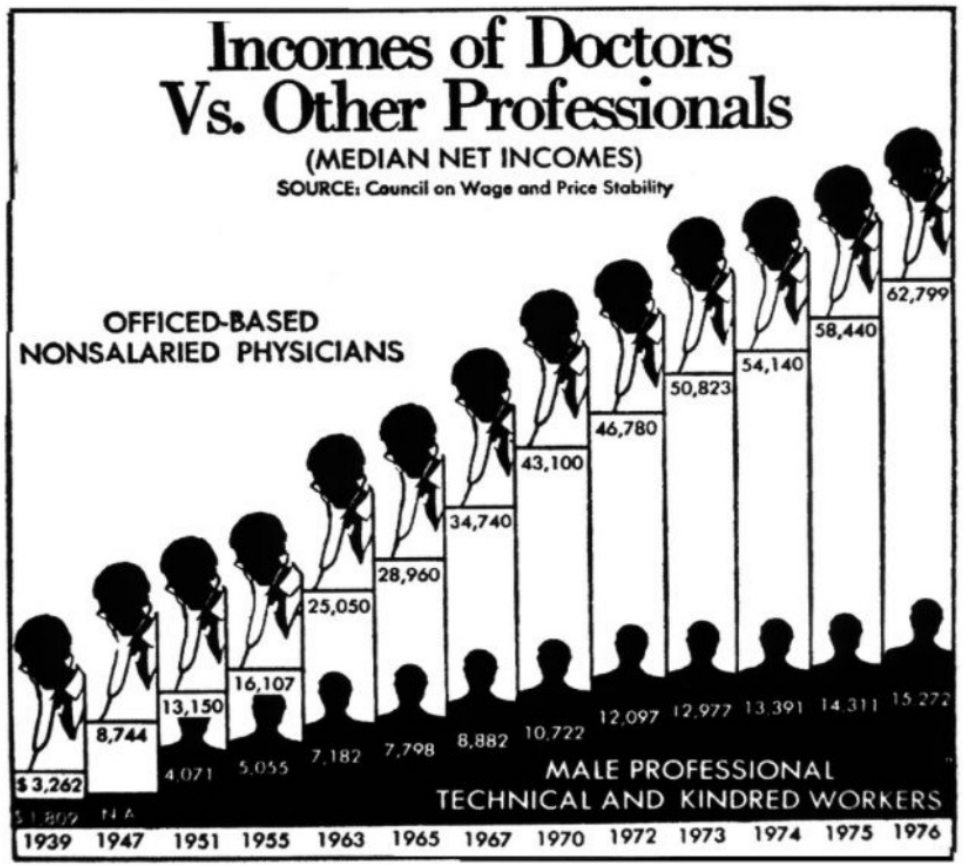
Context

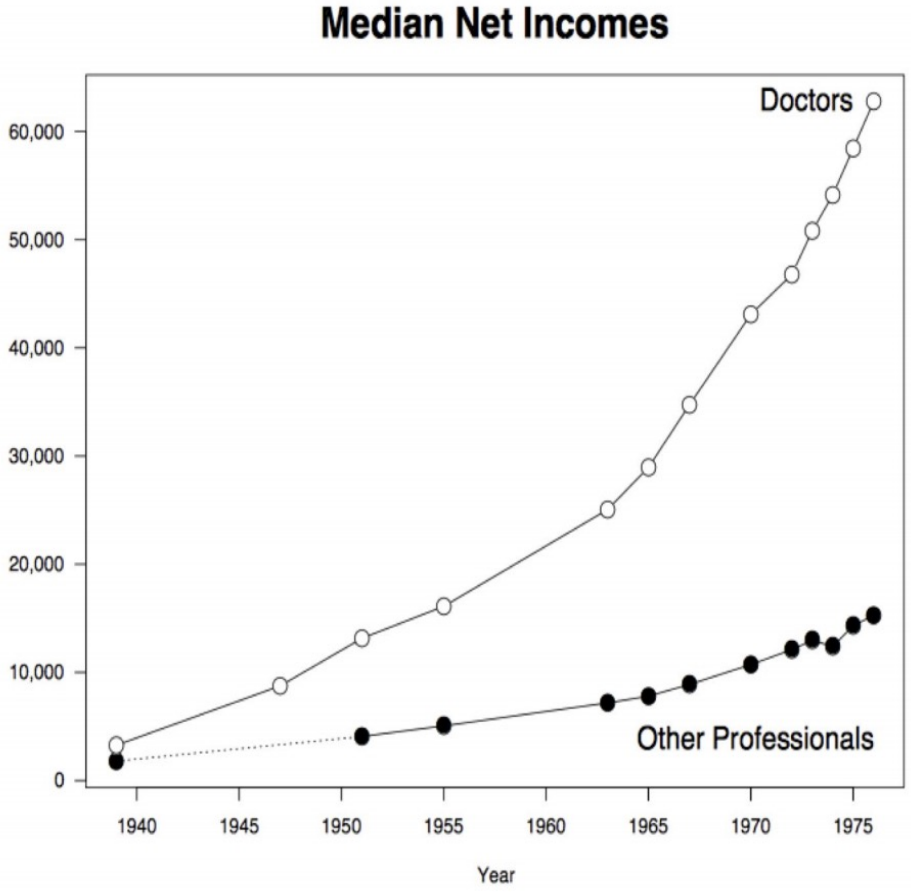
- The most common mistake: No Label



- In general, we provide context for a plot through:
  - Plot title
  - Axes labels
  - Reference lines and markers for important values
  - Labels for interesting points and unusual observations
  - Captions that describe the data and its important features

Don’t be over creative…





Storytelling with data

- “the greatest graph of all time”

- Visualized Napoleon's 1812 invasion of Russia, including
  - the number of soldiers
  - the direction of the march
  - the latitude and longitude of each city
  - the temperature on the return journey
  - dates in November and December

Charles Joseph Minard
1781 - 1870
French civil engineer

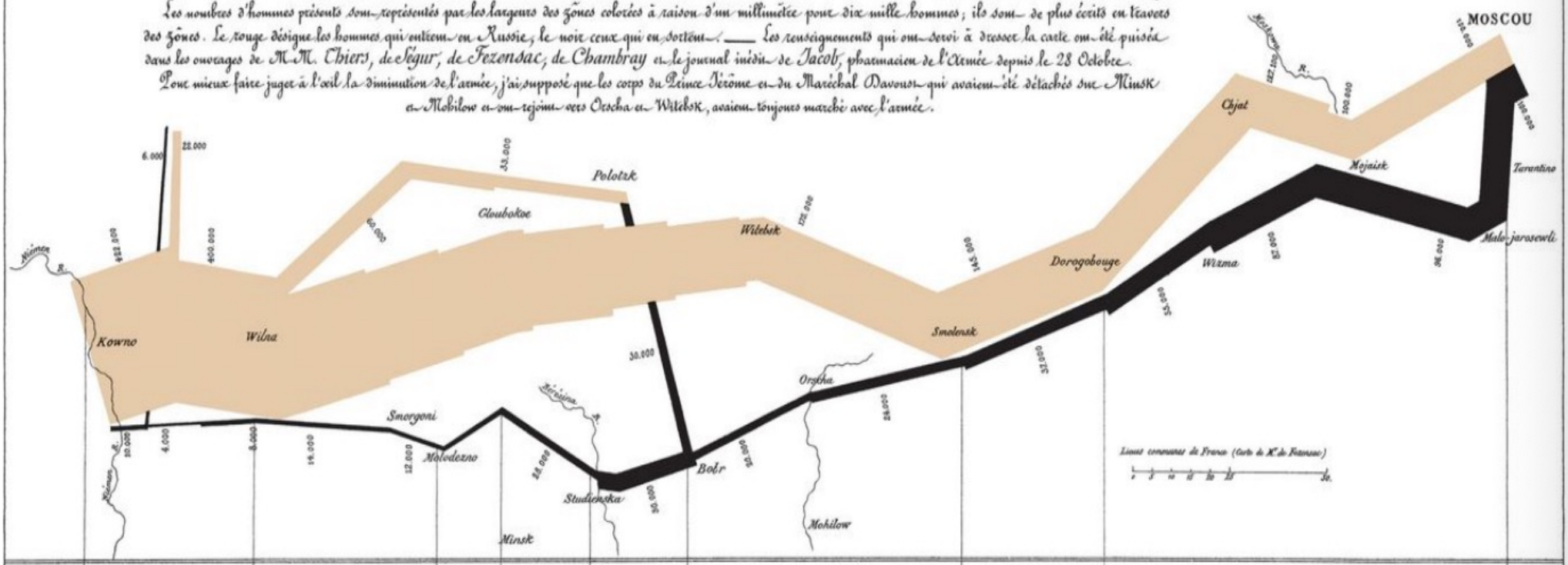
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.



Autog. par Regnier, 8 Pas. Ste Marie St Gven à Paris.

Imp. Lith. Regnier et Dourdet.

3 ways to plot in Python

- Matplotlib
  - the underlying plotting library powering all 3 of these
- Pandas.plot()
  - knows how to make some default plots for you
- Seaborn
  - allows to create sophisticated visualizations quickly
  - good default setting

- Documentation is your friend!
  - Matplotlib Gallery
  - Seaborn Example Gallery
- DSCI 304: data visualization in R

Demo