Data Visualization

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Outline

"One picture worth ten thousand words."

Frederick R. Barnard, *Printer's Ink*, March 10th, 1927.

An Oldie But Goodie

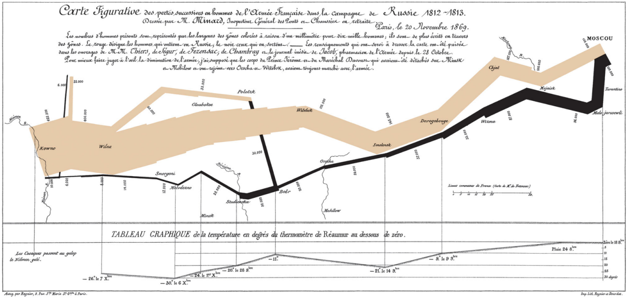

Figure 1: *Minard's representation of Napoleon's 1812 Russian Campaign.* This graph, made in 1861 by Charles Joseph Minard (1781–1870), is commonly regarded as one of the finest ever. It represents, in only two dimensions, the size of the troops, their location, their direction of movement, dates, and temperatures. https://en.wikipedia.org/wiki/Charles_Joseph_Minard.

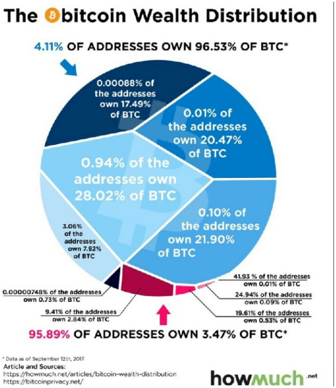

Figure 2: *Bitcoin wealth distribution.* `http://viz.wtf/image/166329900475`.

One picture worth ten thousand words.

- Only if it is a good picture.
- We tend to be more demanding with text than with graphics.
- How long does it take to write/read one thousand words? At least the same effort should be put into making/viewing a graph.

Learning Objectives

- Become a wise and effective "creator" / "maker" as well as "reader" / "viewer" of data visualization.
- Master general principles for data visualization and apply these when making your own graphs as well as when viewing others'.
- Produce the right graph for the matter at hand.
- Become aware of the variety of graphical techniques available for different types of data and purposes and understand their pros and cons.
  Go beyond histograms and pie charts!
- Think more carefully about each plot you create, consider the pros and cons of different choices, and try several different plots for a given dataset.

Learning Objectives

- Familiarize yourself with software for data visualization. Most of the examples in theses slides are based on Python's matplotlib and seaborn libraries. However, as discussed in the first lecture, other languages such a R may be better suited for certain tasks.
- Focus on what type of plot to make rather than how to make it, i.e., compose the plot conceptually before thinking of its software implementation details. Concepts are general and long-lasting, will syntax is highly specific and ephemeral.
- Avoid bad graphs!

- Data visualization is a fundamental aspect of Data Science.
- It is essential to "look at data" throughout the workflow, from exploratory data analysis (EDA) to model diagnostics and reporting the results of the inquiry.
- Visualization is valuable for detecting the main features (good or bad) of a dataset, revealing patterns, and suggesting theories or further questions.
- Visualization is also useful for quality/assessment control (QA/QC) and detecting problems with the data.
- An effective plot can be good enough to answer the question on its own. In some cases, it may even be the only appropriate type of answer.

- An effective plot can also be sufficient to convince stakeholders of the findings from a full-blown statistical inference procedure.

Data Visualization

- Although data visualization is ubiquitous and heavily relied upon, in research as well as in the media, typically not much thought is put into creating or reading plots.
  - Creators often rely on very limited subsets of plots and without proper consideration of their limitations.
  - Readers often passively absorb a message imposed on them by the graph, rather than reason and think critically about it.
- Very few Statistics, Computer Science (CS), or domain curricula offer courses in data visualization.
- Proper data visualization is non trivial. Entire courses could and should be devoted to data visualization, including discussions of vision and perception to guide the design of effective graphs.

Do We Really Need a Graph?

- When the data only comprise a handful of values, a table or a simple mention in text may be a more effective, i.e., accurate and simple, display.
- E.g. Percentage of popular vote for Trump and Clinton in 2016 presidential election:

Trump	46.1 %
Clinton	48.2 %

Do We Really Need a Graph?

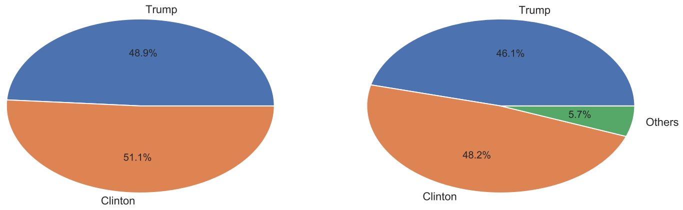

Figure 3: *US Election Results 2016.* Left: Pie chart of percentage of popular vote for Trump and Clinton. Right: Pie chart of percentage of popular vote for Trump, Clinton, and other candidates. Why the different percentages on left and right?

From Tables to Graphs

- When a table represents two or more variables, with more than a handful of values each, a graph may be more effective.
- Tables leave the interpretation to the viewer.
- Graphs provide a summary of the data and are more amenable to comparisons.
- Gelman et al. (2002). Lets Practice What We Preach: Turning Tables into Graphs. http://www.stat.columbia.edu/~gelman/research/published/dodhia.pdf.

From Tables to Graphs

Profession	Frequency of recent citations	1996 total employed (1,000)	Relative frequency
Lawyers	8101	880	9.2
Economists	1201	148	8.1
Architects	1097	160	6.9
Physicians	3989	667	6.0
Statisticians	34	14	2.4
Psychologists	479	245	2.0
Dentists	165	137	1.2
Teachers (not university)	3938	4724	0.8
Engineers	934	1960	0.5
Accountants	628	1538	0.4
Computer programmers	91	561	0.2
Total	20,657	11,034	1.9

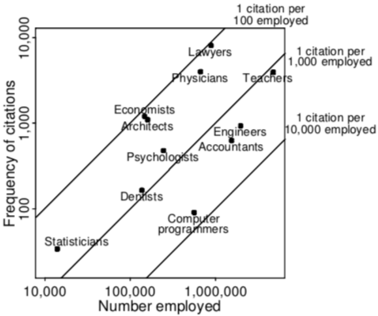

Figure 4: *Turning tables into graphs (Gelman et al., 2002, Figure 2).* Counts and rates of citations of various professions from the New York Times database. Graph: Log-log scale allows comparison across several orders of magnitude. Any 45° line indicates constant relative frequency. The relative positions of the different professions is clearer.

More Oldies But Goodies

Figure 5: *Album de Statistique Graphique (1881).* `https://www.davidrumsey.com/`.

More Oldies But Goodies: Maps

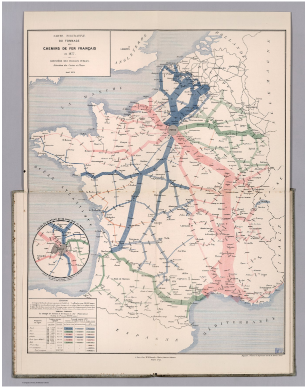

Figure 6: *Album de Statistique Graphique (1881).* Train load (scaled by length of line) is represented by thickness of bands. How would you represent this data without a graph?

More Oldies But Goodies: Graphical Timetables

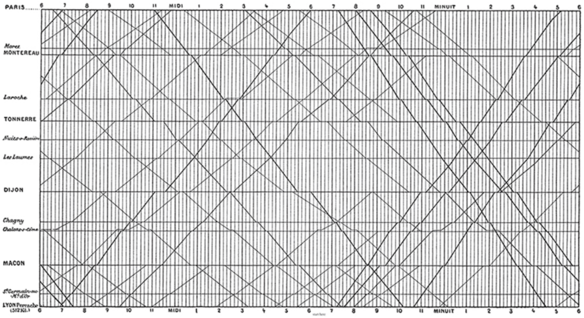

Figure 7: *Marey (1885). Train schedule Paris–Lyon, 1880s.* `https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003zP`. How would you represent this data without a graph?

More Oldies But Goodies: Graphical Timetables

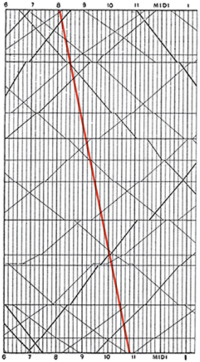

Figure 8: *Marey (1885). Train schedule Paris–Lyon with TGV, 1980s vs. 1880s.* The red line indicates the 1981 itinerary of the TGV, a new express train that cut the trip from Paris to Lyon to under three hours (vs. nine hours in the 1880s).

More Oldies But Goodies: Graphical Timetables

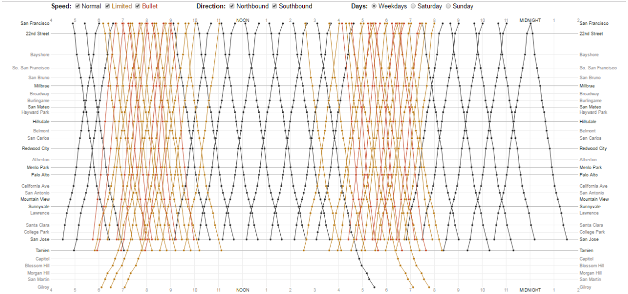

Figure 9: *Train schedule SF–Gilroy, now.* https://i.stack.imgur.com/qJ1hH.

More Oldies But Goodies: Graphical Timetables

- In Marey (1885)'s Paris–Lyon graphical train schedule in the 1880s, time is represented on the x axis and the stations and distances between stations are represented on the y axis (Tufte, 2001).
- A train's itinerary is represented by a line.
- The slope of the line reflects the speed of the train: The more nearly vertical the line, the faster the train.
- The length of a stop at a station is indicated by the length of the horizontal line.
- The intersection of two lines locates the time and place that trains going in opposite directions pass each other.
- This type of graph, known as a parallel coordinates plot, is still used today and has many other applications.

Caveats

- Graphs should attempt to summarize data in a simple, intuitive, and efficient manner, without distorting or loosing important information.
- However, not all good graphs are simple. As with text, plots conveying a lot of information (e.g., displaying multiple variables) require both a skillful creator and an educated reader.
  E.g. Minard's graph for Napoleon's Russia campaign, old graphical train schedules.
- There is no "one-size-fits-all" graph, i.e., different types of graphs should be used for different
  - types of data, e.g., quantitative, qualitative variables;
  - purposes, e.g., debugging code, EDA, reporting results;
  - media, e.g., print journal, projector.

Caveats

- Graphs typically reduce the information contained in the data.
  E.g. Histograms map n data points into $B < n$ bins; boxplots map n data points into 5 summary statistics ($+$ possibly outliers).
- By focusing on certain aspects of the data or even imposing structure on data, graphs can also be subjective.
  E.g. Choosing which variables to plot, decisions regarding axes and scales, dendrogram representation of clusters[1].
- As with text, the creator of the plot makes editorial decisions as to which data to display and which aspects of these data to show or emphasize.

Caveats

- The reader should assess the relevance and reliability of the data being displayed, as well as the appropriateness of the graph.
- Software implicitly makes many decisions for the creator of a plot, e.g., axes, scales, plotting symbols, color, ordering of data. Experiment with different settings.
- Graphs are rarely presented on their own. They should be interpreted in context of the text which they support. The reader should examine the graph-text interface and, in particular, whether the conclusions in the text are supported by the graph.

---

[1] A dendrogram is a graphical representation of hierarchical clustering results; for a given clustering of n objects, there are 2^{n-1} possible dendrograms. The various choices made in hierarchical clustering as well as the dendrogram representation impose (vs. reveal) structure on the data.

Statistical Inference

- Graphs are by definition functions of the data, i.e., statistics.
- Although not typically viewed this way, visualization can therefore be used as part of statistical inference.
- One can produce the same types of plots for a sample and for a population, in that sense, the plot for the sample can be viewed as an estimator of the plot for the population, i.e., the parameter.
- A pattern that we detect from plotting data for a sample can be used to infer properties of the population from which the sample was drawn. A formalized special case of such an approach is given by linear regression.

General Considerations

In the process of creating a plot, you should consider the following issues.

- Determine the purpose of the plot.
  E.g. EDA, debugging code, comparing distributions, model diagnostics, summarizing results, reporting results.
- Formulate the message.
- Identify the audience.
- Identify the display mode/medium (e.g., journal, projector).
- Think about the best type of graph for the purpose, message, audience, and display mode.
- Aim for efficient perception: Speed, accuracy, and minimum cognitive load for understanding the message.

General Considerations

- Apply visual perception principles.
  E.g. Angles and areas are harder to perceive/compare than lengths.
- Do not use more dimensions to represent the data than are in the data. This rules out pie charts and barplots.
- An important consideration when selecting a graphical technique is how easily it can be extended (e.g., to multiple variables) and how amenable it is to comparing distributions.
- Choose graphical parameters carefully: Aspect ratio, plotting symbols, line types, texture, axes, etc.

General Considerations

- Choose color palette carefully. E.g. Be mindful of color blindness, use different color schemes for different types of data and messages (e.g., sequential, qualitative, and diverging).
- Provide sufficient information so that the plot can be interpreted properly.
  E.g. Title, axis parameters (i.e., label, tick marks), annotation, legend, caption, etc.
  In a document, number the figures and tables.
- Do not include irrelevant information, i.e., avoid "chart junk".
- Principle of "least surprise": If you defy expectations, people may get confused. Only defy expectations if it is very important.

General Considerations

- Experiment, i.e., consider different types of plots and update the plots iteratively.
- Of course, always think about the quality of the data you plot.

General Considerations

- Sample size.
  - For small sample sizes, plot all of the data – Why loose information?
  - For larger samples sizes, plot relevant summaries of the data, that do not distort or loose important information in the data.
- Variables to display/emphasize. Depends on the purpose and message of the plot.
- Type of variables. Quantitative and qualitative variables call for different types of graphical summaries.
- Pre-processing. E.g. Transformation (e.g., log), dimensionality reduction, imputation.

Graphical Perception

- Cleveland and McGill (1985): *"Graphical perception is the visual decoding of the quantitative and qualitative information encoded on graphs. Recent investigations have uncovered basic principles of human graphical perception that have important implications for the display of data."*
- When we create a graph, we encode the data as graphical attributes.
- Possible graphical attributes are: Angles, areas, lengths, position on common aligned/unaligned scale, slopes, color properties.
- Effective graphs are those for which attributes are most easily decoded.

Graphical Perception

- There are empirical laws for perception that can be used to rank different types of graphical encodings.
- In general, such laws relate the perceived (change in) intensity in a physical stimulus to the actual (change in) intensity. This concerns stimuli to all senses, i.e., vision, hearing, taste, touch, and smell.

Graphical Perception: Weber's Law

- Weber's Law is an empirical relationship in psychophysics between the initial intensity in a stimulus (I) and the smallest perceivable difference (a.k.a., just noticeable difference) in the stimulus intensity (ΔI):

$$\frac{\Delta I}{I} = k, \tag{1}$$

where k is a proportionality constant for a given type of stimulus [2].

- In terms of length, this means we detect a 1 cm change in a 1 m length as easily as we detect a 10 m change in a 1 km length.
- Weber's Law appears to hold for many different graphical encodings.

[2] Law formulated and published by Gustav Theodor Fechner (1801–1887), a student of Ernst Heinrich Weber (1795–1878).

Graphical Perception: Stevens' Law

- Stevens (1957) Law is an empirical relationship in psychophysics between the intensity in a stimulus and the perceived magnitude of the sensation created by the stimulus:

$$\psi(I) = Ci^{\beta}, \tag{2}$$

where I is the intensity or strength of the stimulus in physical units (energy, weight, pressure, mixture proportions, etc.), $\psi(I)$ is the magnitude of the sensation, β is an exponent that depends on the type of stimulation or sensory modality, and C is a proportionality constant that depends on the units used.

- Examples of values for exponent, β

Length:	0.9 – 1.1
Area:	0.6 – 0.9
Volume:	0.5 – 0.8

Graphical Perception: Stevens' Law

- For lengths, the relationship is almost linear, thus our perception is about right.
- However, according to this power law, our perception of areas and volumes is conservative, i.e., when values are represented as areas or volumes, we underestimate the large values relative to the small ones and overestimate the small ones relative to the large ones.
- E.g. Areas, with $\beta = 0.7$.
  Consider two areas of size 1 and 2, respectively.

$$\frac{\psi(2)}{\psi(1)} = \frac{2^{0.7}}{1^{0.7}} \approx 1.62.$$

Thus, we don't see the bigger area as twice as large.

Graphical Perception: Stevens' Law

Now consider two areas of size $1/2$ and 1, respectively.

$$\frac{\psi(1/2)}{\psi(1)} = \frac{0.5^{0.7}}{1^{0.7}} \approx 0.62.$$

Thus, we don't see the smaller area as half as large.

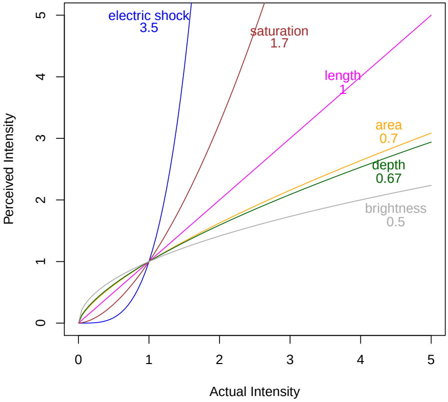

Figure 10: *Graphical perception: Steven's Law.* Stevens (1957) perceived sensory magnitude power law.

Graphical Perception: Combining Weber's and Stevens' Laws

- Consider comparing the values x and $x + w$, using length ($\beta = 1$) and area ($\beta = 0.7$) encodings.
- For length, we perceive the relative value

$$\frac{x + w}{x} = 1 + \frac{w}{x}.$$

- For area, we perceive the relative value

$$\frac{(x + w)^{0.7}}{x^{0.7}} = \left(1 + \frac{w}{x}\right)^{0.7} \approxeq 1 + \frac{0.7w}{x}.$$

- Thus, we are more likely to detect small differences using length encoding.

Graphical Perception

- Cleveland and McGill (1985) carried out an extensive study of graphical encodings to obtain a best to worst ranking.
- The encodings they examined include: position on a common aligned scale, position on a common unaligned scale, length, slope, angle, area, volume, color hue, brightness, and purity.
- One of their experiments consisted of
  - 7 graphical encodings,
  - 3 judgments per encoding,
  - 10 replications per subject,
  - 127 experimental subjects.

  Assessment criterion: error $= \|\text{perceived } p - \text{true } p\|$, where p denotes the ratio (in percentages) of the smaller to the larger magnitude.

Graphical Perception

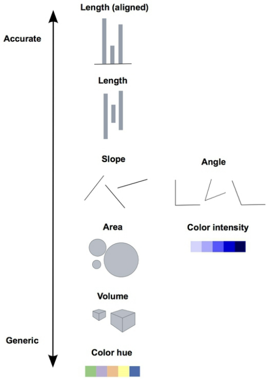

Figure 11: *Graphical perception.* Based on Table 1 in Cleveland and McGill (1985). `http://paldhous.github.io/ucb/2016/dataviz/week2.html`.

Bad Graphs

- The literature is full of "bad graphs", that, for instance, distort the data and are misleading, are too complicated, or are missing essential information.
- Karl Broman's Top Ten Worst Graphs (including one of his own!): https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/.
- Ross Ihaka's Good and Bad Graphs: https://www.stat.auckland.ac.nz/~ihaka/120/Lectures/lecture03.pdf.
- Edward Tufte: https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00040Z.
- Junk Charts: https://junkcharts.typepad.com/junk_charts/.
- WTF Visualization: http://viz.wtf.

Bad Graphs: Pie Charts

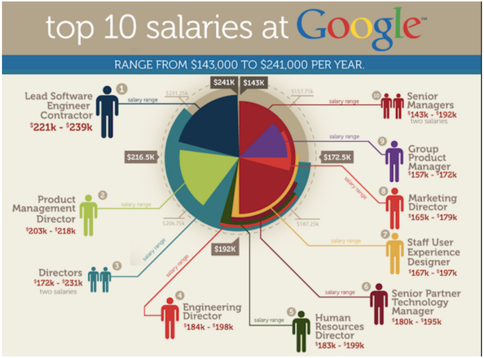

Figure 12: *Top 10 Google salaries by job category: Pie chart.* `https://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html`. What's the message? What do the angles represent? What's a better graph?

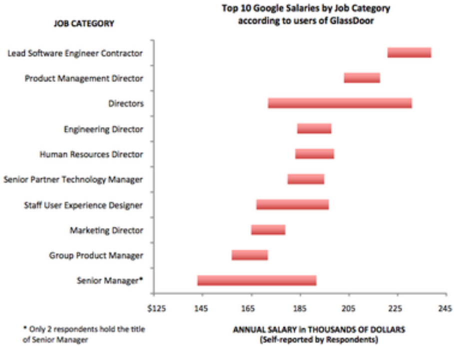

Figure 13: *Top 10 Google salaries by job category: Interval chart.* https://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html.

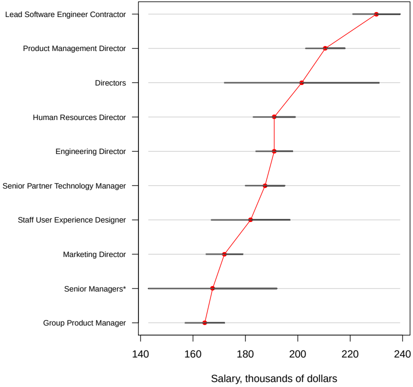

Figure 14: *Top 10 Google salaries by job category: Interval chart.* Sorted by midpoint of salary range.

Bad Graphs: Pie Charts

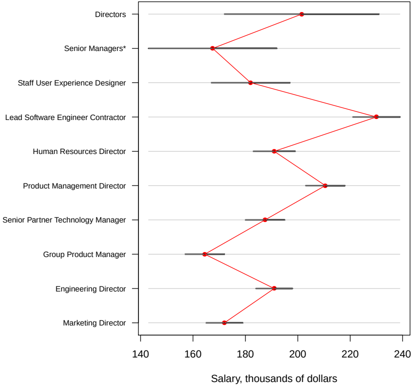

Figure 15: *Top 10 Google salaries by job category: Interval chart.* Sorted by salary range.

Bad Graphs: Pie Charts

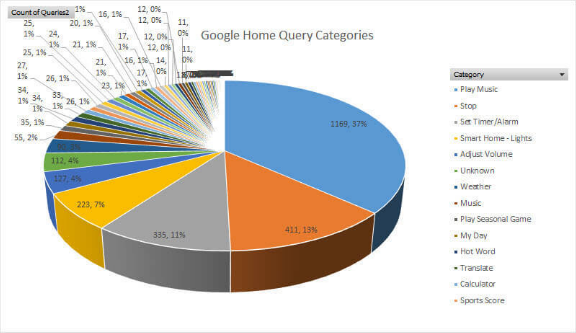

Figure 16: *Google Home query categories: Pie chart.* http://viz.wtf/image/171134950336. Unreadable. Can't match numbers to categories. What's a better graph?

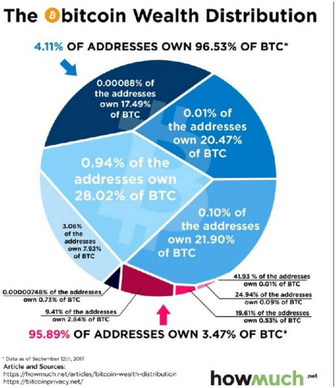

Figure 17: *Bitcoin wealth distribution: Pie chart.* `http://viz.wtf/image/166329900475`. What's the message? How to compare shapes and areas? Without text, pie uninformative. What's a better graph?

Bad Graphs: Pie Charts

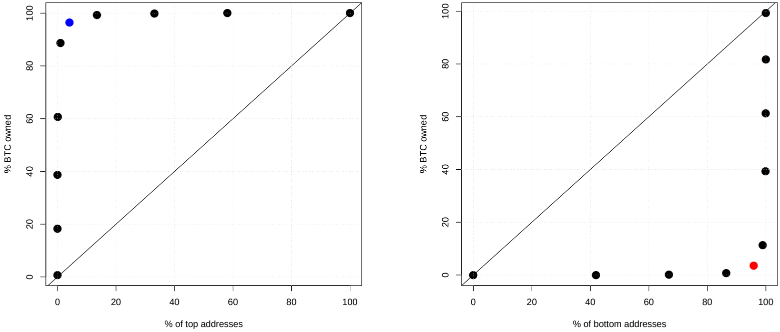

Figure 18: *Bitcoin wealth distribution: Scatterplot.*

Bad Graphs: Multilevel Donut Charts

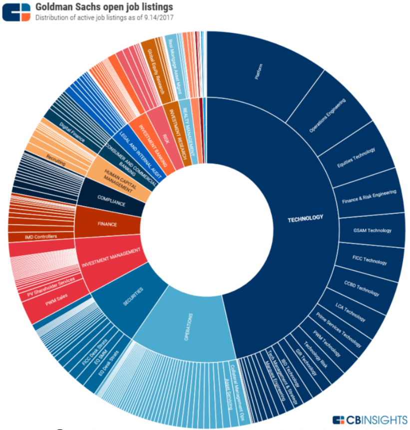

Figure 19: *Goldman Sachs job listings: Multilevel donut chart.* https://s3.amazonaws.com/cbi-research-portal-uploads/2017/09/18173935/GSteardownjobs. What's the message? Unreadable. What's a better graph?

Bad Graphs: Wordclouds

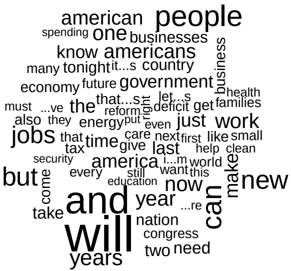

Figure 20: *State of the Union speeches 2010 and 2011: Wordcloud.* Frequency of words with at least 15 occurrences. What's the message? How to compare frequencies of words? What's a better graph?

Bad Graphs: Wordclouds

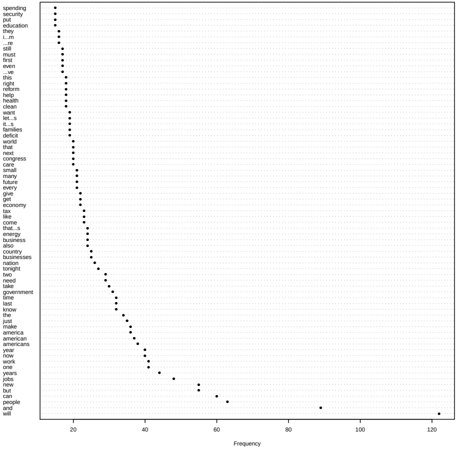

Figure 21: *State of the Union speeches 2010 and 2011: Dotplot.* Frequency of words with at least 15 occurrences.

Bad Graphs: Wordclouds

Expressions et hashtags les plus utilisés sur Twitter (Tweets et Retweets) lors des 4 premières journées de mobilisation – Données récoltées avec l'application Talkwalker

Figure 22: *Gilets jaunes: Wordcloud.* Frequency of expressions and hashtags on Twitter for first four days of gilets jaunes movement. How to compare frequencies between days? https://www.lexpress.fr/actualite/societe/gilets-jaunes-ce-qu-en-disent-les-francais_2055542.html.

Bad Graphs: Wordclouds

Figure 23: *Names: Wordcloud.* `https://www.wordclouds.com/?cloud=names`.

Bad Graphs: Wordclouds

Figure 24: *Business words: Wordcloud.*
https://www.wordclouds.com/?cloud=business

Bad Graphs

- Chart junk. The previous graphs exemplify "chart junk" , i.e., they contain superfluous elements that are not necessary to convey the information contained in the data, but instead distract the viewer from this information or even mask or distort important information.
- Pie charts.
  - Frequency represented by angle/area.
  - Angles and areas are hard to perceive and compare.
  - Pie charts quickly become unreadable for more than a handful of values.
  - Listing the values is often better – they are actually often added to a pie chart anyway!
  - How to select order of categories?
  - Not amenable to comparing distributions; side-by-side comparisons not effective.

  - Hard to extend to multiple variables.
  - A lot of junk often added to pie charts, e.g., thickness, slice explosion.
- Wordclouds/tag clouds.
  - Frequency represented by font size.
  - Neither area nor height corresponds to frequency of words.
  - How do longer words compare with shorter words?
  - How are capital letters handled?
  - How to calculate relative difference in frequency between two words?
  - How are the words ordered within the cloud (alphabetical, frequency)?
  - Not amenable to comparing distributions; side-by-side comparisons not effective.
  - How to extend to multiple variables?
  - A lot of junk often added to word clouds.

Bad Graphs

- Barcharts/barplots. Better.
  - Based on length and position on common aligned scale.
  - Add an irrelevant dimension (thickness of bar).
  - How to select order of categories?
  - Not readily amenable to comparisons.
  - Extension to multiple variables problematic.
- Dotcharts/dotplots. (And interval charts.) Even better.
  - Based on length and position on common aligned scale.
  - Display only the relevant information.
  - How to select order of categories?
  - More amenable to comparisons and extensions to multiple variables.

Gapminder

Gapminder. (`https://www.gapminder.org`)

- We will use data from Gapminder to reason through the process of data visualization, e.g., population, population density, life expectancy, income for each country.
- Note that in this case we have a census, i.e., there is no sampling involved [3].
- Gapminder is a Swedish foundation co-created in 2005 by Hans Rosling (Professor of International Health at Karolinska Institute) and family members.
- *"Gapminder is a fact tank, not a think tank."*
  "Gapminder measures ignorance about the world."
  "Gapminder makes global data easy to use and understand."
  "Gapminder promotes Factfulness, a new way of thinking."

- Gapminder developed Trendalyzer, a data visualization software providing dynamic and interactive graphics of data compiled by organizations such as the United Nations and the World Bank (acquired by Google in 2007).

[3] Some of the data could be estimates, but we won't concern ourselves with this at this point.

Gapminder

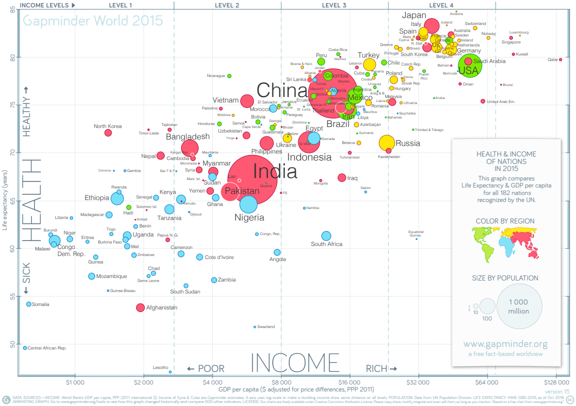

Figure 25: *Gapminder: World Poster 2015. “How Does Income Relate to Life Expectancy? Short answer - Rich people live longer.”* Bubble chart with four variables displayed in 2D.

Software

- Most of the plots below are produced with Python's seaborn library, using default arguments.
- Default settings typically do not correspond to the most basic version of the plot, but rather impose many decisions on the plot, e.g., color, legend, ordering. Experiment with different settings to make sure you get the plot you want.
- seaborn tutorial: `https://seaborn.pydata.org/tutorial.html`. Each function has many arguments to customize the plots. As usual, consult documentation.
- Datasets available at: `https://github.com/mwaskom/seaborn-data`. E.g. Titanic survival dataset, Fisher's iris dataset.

One Quantitative Variable

How would you visualize life expectancy in 2018 over all countries?

count	182.000000
mean	72.726374
std	7.237996
min	51.100000
25%	67.150000
50%	74.100000
75%	78.075000
max	84.200000

Stem-And-Leaf Plots

Key: aggr|stem|leaf

182 | **84** | *0* = **84** x0.1.0 = 84.0

182	84	02
180	83	25
178	82	1244446669
168	81	11223333458889
154	80	0125778
147	79	13446
142	78	00122367
134	77	0223466677899
121	76	0125678899
111	75	1223355578999
98	74	011238899
89	73	123448
83	72	0002334456
73	71	115569
67	70	3555679
60	69	138
57	68	0002378
50	67	113389
44	66	114689
38	65	0245788
31	64	356
28	63	14569
23	62	24599
18	61	01112269
10	60	025
7	59	57
5	58	067
2	57	
2	56	
2	55	
2	54	
2	53	
2	52	
2	51	16

Figure 26: *Life expectancy, 2018.*

Stripplots

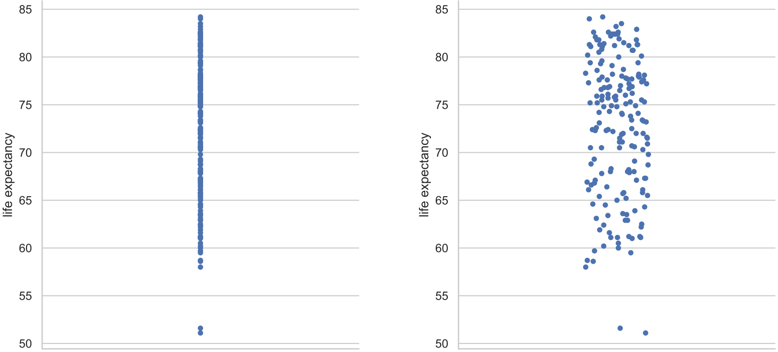

Figure 27: *Life expectancy, 2018.* Right: Jittering, i.e., adding random noise, to avoid overplotting.

Histograms

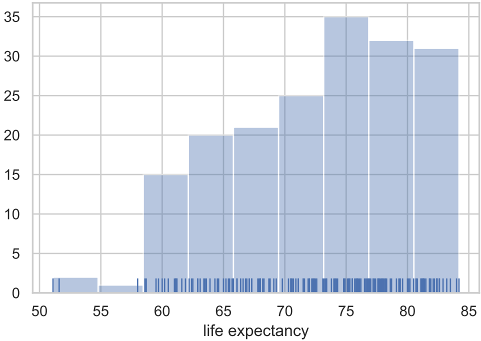

Figure 28: *Life expectancy, 2018.*

Histograms

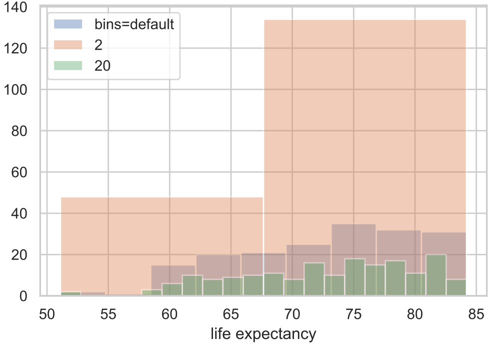

Figure 29: *Life expectancy, 2018.* Different numbers of bins.

Density Plots

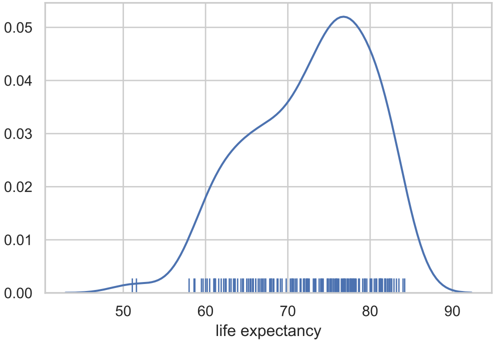

Figure 30: *Life expectancy, 2018.*

Density Plots

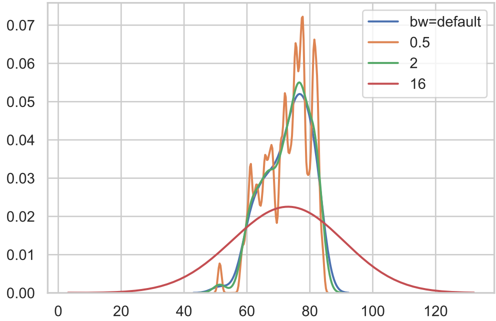

Figure 31: *Life expectancy, 2018.* Different bandwidths.

Boxplots

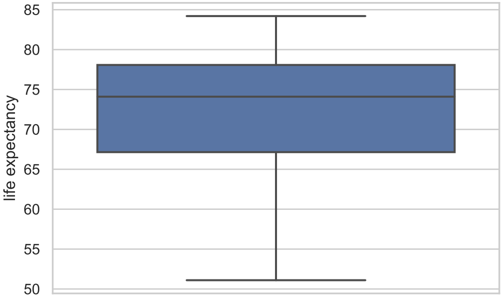

Figure 32: *Life expectancy, 2018.*

One Quantitative Variable and One Qualitative Variable

How would you visually compare life expectancy between regions?

In [52]:
```
(gm2018.groupby('six_regions'))['life expectancy'].describe()
```

Out[52]:

six_regions	count	mean	std	min	25%	50%	75%	max
america	33.0	75.827273	3.774360	64.5	73.400	76.10	78.600	82.2
east_asia_pacific	26.0	72.557692	6.833399	61.1	68.100	71.55	77.275	84.2
europe_central_asia	49.0	77.969388	4.047746	70.5	75.200	78.00	81.500	83.5
middle_east_north_africa	19.0	75.689474	4.642544	67.1	74.050	76.90	78.050	82.4
south_asia	8.0	71.675000	6.652121	58.7	68.825	72.45	75.550	80.1
sub_saharan_africa	47.0	64.157447	4.852599	51.1	61.150	63.90	66.850	74.9

Stripplots

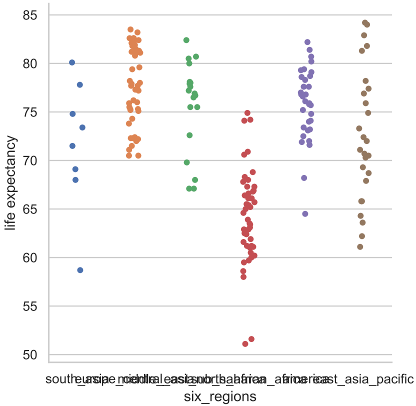

Figure 33: *Life expectancy by region, 2018.*

Histograms

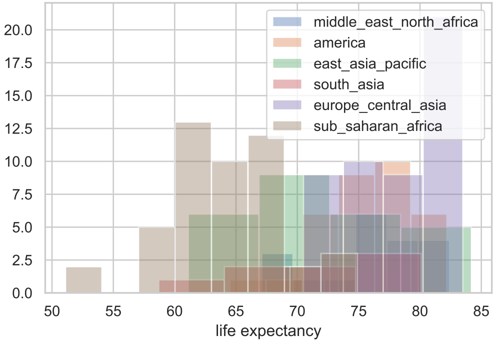

Figure 34: *Life expectancy by region, 2018.*

Density Plots

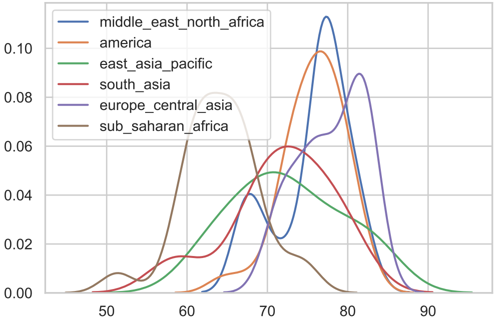

Figure 35: *Life expectancy by region, 2018.*

Boxplots

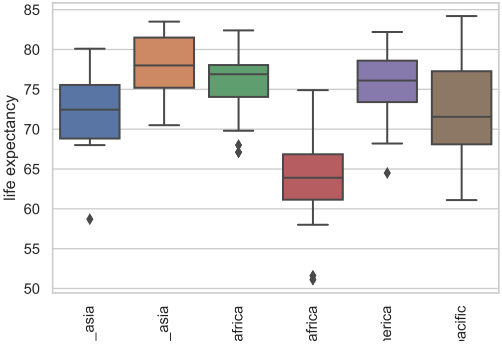

Figure 36: *Life expectancy by region, 2018.*

Violin Plots

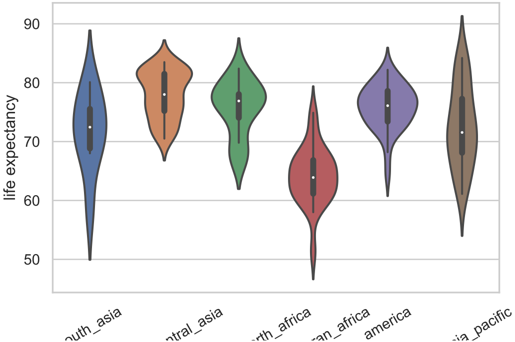

Figure 37: *Life expectancy by region, 2018.*

Log-Transformation

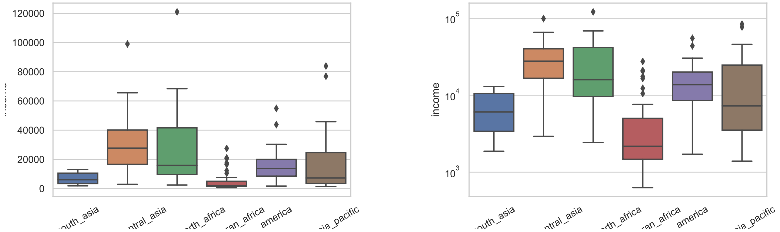

Figure 38: *Income, 2018.* Left: Income (GDP/capita, inflation-adjusted \$). Right: Log-transformed income.

How did life expectancy vary between 1800 and 2018?

Time Series

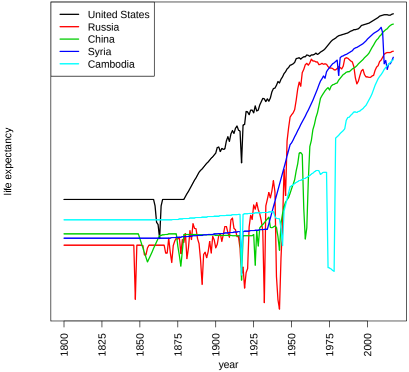

Figure 39: *Life expectancy over time for five countries.*

Time Series

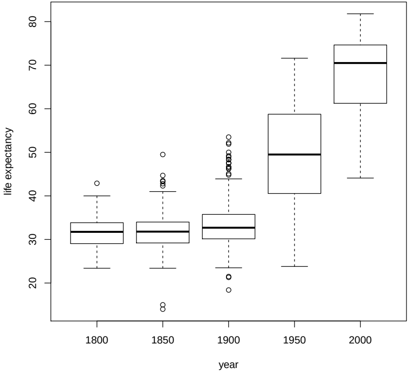

Figure 40: *Life expectancy over time.*

One Quantitative Variable: Summary

Displaying and comparing marginal distributions for quantitative data.

- Stem-and-leaf plots.
  - Simple pen-and-paper method for visualizing the distribution of all of a handful of values.
  - Not amenable to comparisons between distributions.
  - No reason to use these days.
- Stripcharts/Stripplots. (Sometimes referred to as dotcharts/dotplots, related to rug plots.)
  - Effective for visualizing the distribution of all of a moderate number of values.
  - Can use side-by-side stripplots to compare multiple distributions.
- Histograms.
  - Classical method for displaying a single distribution.

- Sensitive to bin width and bin boundaries.
- Cannot easily display and compare multiple distributions.

- Density plots.
  - Based on kernel density estimation (cf. smoothing).
  - Sensitive to bandwidth, but methods available to select bandwidth.
  - Effective for displaying and comparing multiple distributions.
- Boxplots. (A.k.a., box-and-whiskers plots.)
  - Summarize distribution by only 5 numbers (+ outliers): Median, upper and lower-quartiles, whiskers at 1.5 times inter-quartile range (IQR) above and below upper and lower-quartiles, respectively.
  - Possible loss of information, e.g., multimodality.
  - Effective for displaying and comparing multiple distributions, especially with notches.

One Quantitative Variable: Summary

- Violin plots.
  - Trendy hybrids of boxplots and density plots.
  - Redundant (twice the density plot!), unless plot different densities on each side.
  - Same limitations and issues as with boxplots and density plots.
  - Cannot compare densities as readily as with standard density plots.

How would you visually examine the relationship between life expectancy and income in 2018 over all countries?

Scatterplots

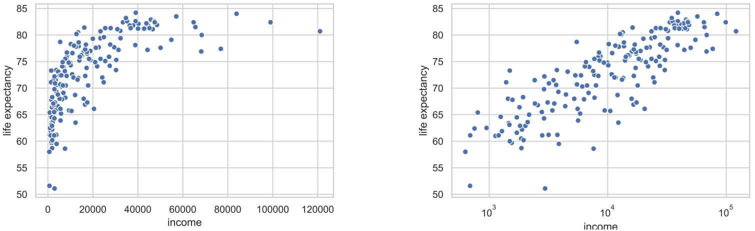

Figure 41: *Life expectancy vs. income, 2018.* Left: Income (GDP/capita, inflation-adjusted $). Right: Log-transformed income.

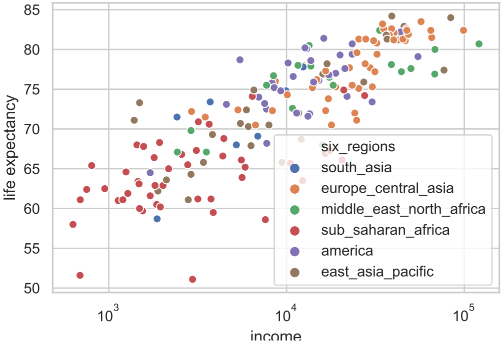

Figure 42: *Life expectancy vs. income, colored by region, 2018.*

Bubble Charts

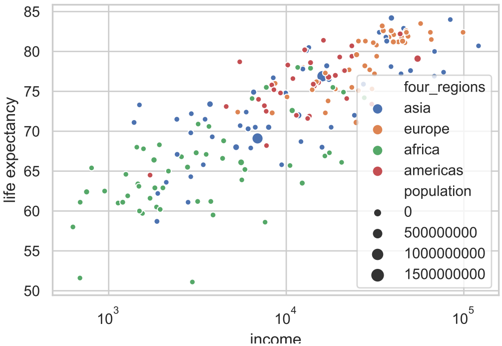

Figure 43: *Life expectancy vs. income, colored by region and with area of bubbles representing population, 2018.*

Mean-Difference Plots

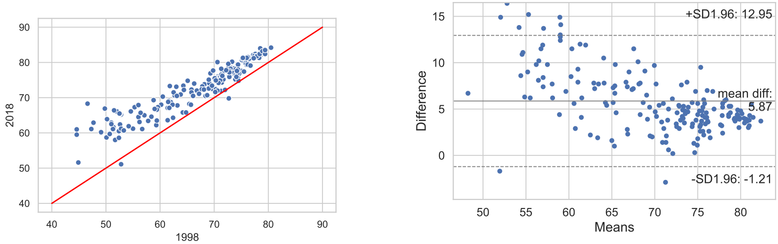

Figure 44: *Life expectancy, 2018 vs. 1998.*

Scatterplot Matrices

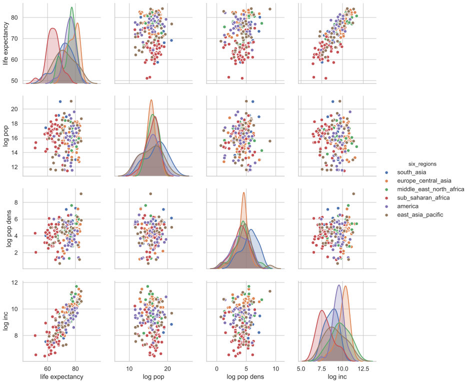

Figure 45: *Life expectancy, population, population density, and income, by region, 2018.*

Scatterplot Matrices

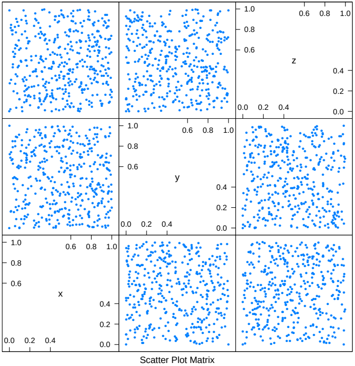

Figure 46: *RANDU RNG.* Triples of successive numbers.

3D Scatterplots

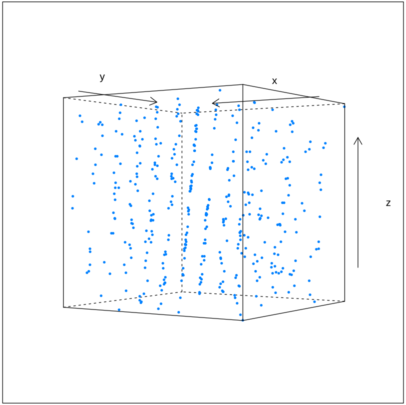

Figure 47: *RANDU RNG.* Triples of successive numbers.

RANDU RNG

RANDU random number generator. (R. Ihaka, https://www.stat.auckland.ac.nz/~ihaka/120/Lectures/lecture27.pdf.)

- The dataset consists of 400 triples of successive numbers produced by the RANDU random number generator (RNG).
- The consecutive triples produced by RANDU are constrained to lie on a series of parallel planes which cut through the unit cube.
- The planes are not aligned with the sides of the unit cube and so do not show up in any of the panels of a scatterplot matrix.

Multiple Quantitative Variables: Summary

Displaying joint distributions for quantitative data.

- While density plots and boxplots are useful for comparing two or more marginal distributions (e.g., in terms of location and scale), they do not provide any information about joint distributions and, in particular, associations between two variables.
- Scatterplots and scatterplot matrices.
  - Useful for examining linear association between two variables.
  - Can extend beyond two variables by using color and plotting symbol area, as in bubble charts.
  - However, can miss important higher-dimensional patterns (cf. RANDU example).
- Mean-difference plots.

Multiple Quantitative Variables: Summary

- Rotated and scaled version of scatterplot.
- Better for looking at differences vs. associations.

- Bubble charts. A bubble chart is a type of scatterplot that displays one or two extra dimensions using area and color.
- Parallel coordinates plots.
  - Natural for visualizing time series data, i.e., same variable measured across time.
    Cf. Train schedules.
  - Can also be used for visualizing the relationship between multiple variables, but trickier: Each line corresponds to an observation and each axis to a variable.
  - Three important considerations, that can affect interpretation of the plot: The order, the rotation, and the scaling of the axes.

Overplotting

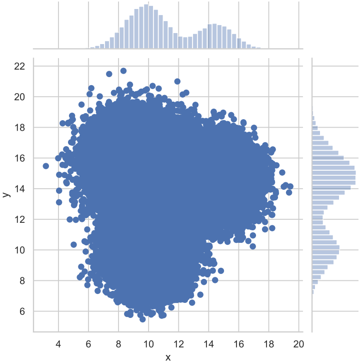

Figure 48: *Simulated data, $n = 60,000$: Scatterplot.*

Overplotting: Hexagonal Binning

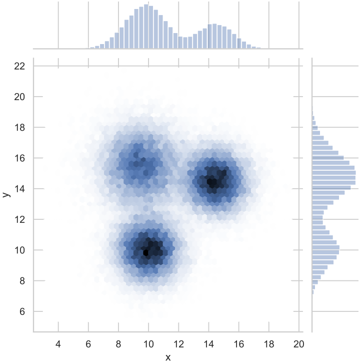

Figure 49: *Simulated data, $n = 60,000$: Hexagonal binning.*

Overplotting: Scatterplot Smoothing

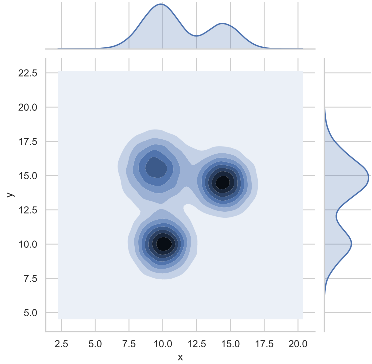

Figure 50: *Simulated data, $n = 60,000$: Scatterplot smoothing.*

Overplotting

Overplotting issues can be reduced by the following approaches.

- Changing plotting symbol.
- Jittering, i.e., adding random noise.
- Smoothing.
- Hexagonal binning.

Qualitative Variables

How would you visualize the 2017 UK election results?

```
Number of seats for each of 13 parties.
Party MPs
0 CON 318
1 LAB 261
2 SNP 35
3 LIB DEM 12
4 DUP 10
5 SF 7
6 PC 4
7 GREEN 1
8 IND 1
9 OTHER 1
10 UKIP 0
11 SDLP 0
12 UUP 0
```

Pie Charts

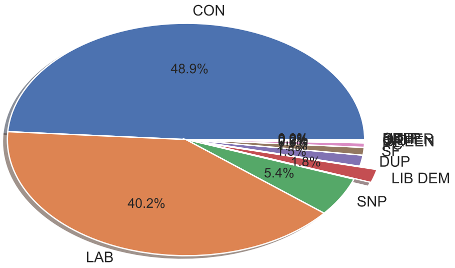

Figure 51: *UK Election Results 2017.* Number of seats for each of 13 parties.

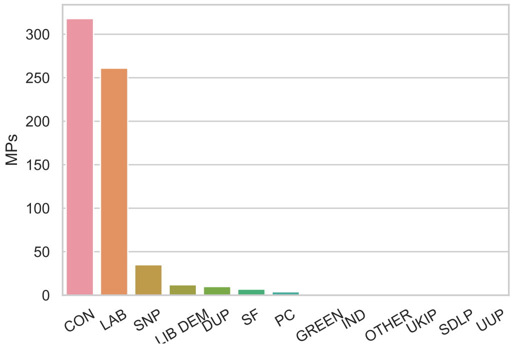

Figure 52: *UK Election Results 2017.* Number of seats for each of 13 parties.

Dotplots

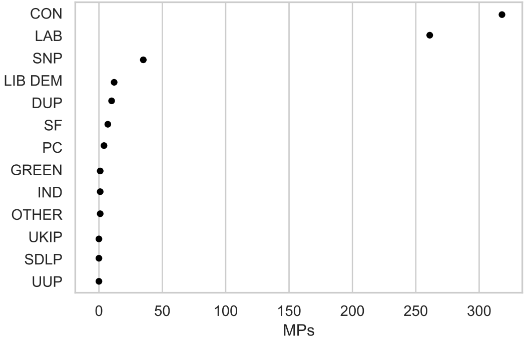

Figure 53: *UK Election Results 2017.* Number of seats for each of 13 parties.

Lollipop Plots

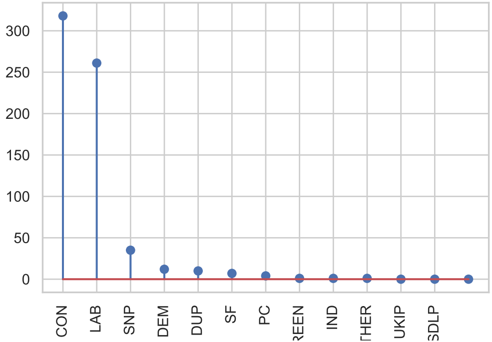

Figure 54: *UK Election Results 2017.* Number of seats for each of 13 parties.

One Qualitative Variable: Summary

- Pie charts.
  - Frequency represented by angle/area.
  - Angles and areas are hard to perceive and compare.
  - Pie charts quickly become unreadable for more than a handful of values.
  - Listing the values is often better – they are actually often added to a pie chart anyway!
  - How to select order of categories?
  - Not amenable to comparing distributions; side-by-side comparisons not effective.
  - Hard to extend to multiple variables.
  - A lot of junk often added to pie charts, e.g., thickness, slice explosion.
- Wordclouds/tag clouds.
  - Frequency represented by font size.

One Qualitative Variable: Summary

- Neither area nor height corresponds to frequency of words.
- How do longer words compare with shorter words?
- How are capital letters handled?
- How to calculate relative difference in frequency between two words?
- How are the words ordered within the cloud (alphabetical, frequency)?
- Not amenable to comparing distributions; side-by-side comparisons not effective.
- How to extend to multiple variables?
- A lot of junk often added to word clouds.

• Barcharts/barplots.
  - Based on length and position on common aligned scale.
  - Add an irrelevant dimension (thickness of bar).
  - How to select order of categories?
  - Not readily amenable to comparisons.

One Qualitative Variable: Summary

- Extension to multiple variables problematic.

- Dotcharts/dotplots. (And interval charts.)
  - Based on length and position on common aligned scale.
  - Display only the relevant information.
  - How to select order of categories?
  - More amenable to comparisons and extensions to multiple variables.
- Lollipop plots.
  - Similar to dotcharts/dotplots (with added stem) and barcharts/barplots.
  - Stem is redundant.
  - How to select order of categories?
  - Not readily amenable to comparisons.
  - Extension to multiple variables problematic.

How would you display survival data on the Titanic according to class, gender, and age?

```
In [107]: pd.crosstab(index=titanic['survived'], columns=[titanic['class'], titanic['who']], margins=True)
```

Out[107]:

class	First			Second			Third			All
who	child	man	woman	child	man	woman	child	man	woman	
survived										
0	1	77	2	0	91	6	33	281	58	549
1	5	42	89	19	8	60	25	38	56	342
All	6	119	91	19	99	66	58	319	114	891

Barplots

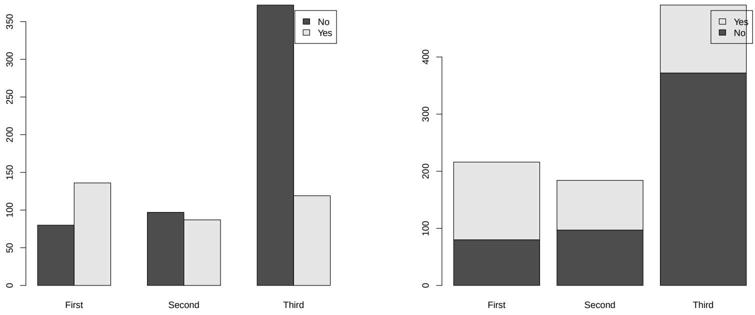

Figure 55: *Titanic: Survival by class.*

Dotplots

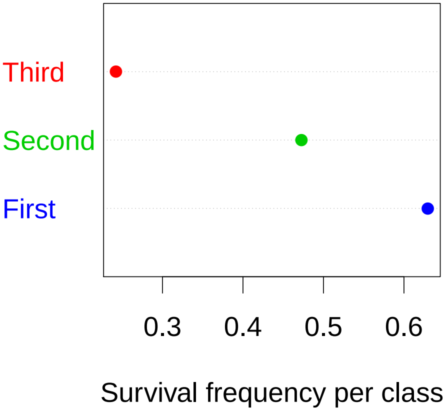

Figure 56: *Titanic: Survival by class.*

Dotplots

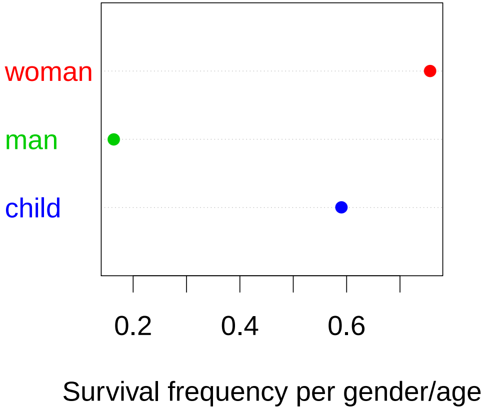

Figure 57: *Titanic: Survival by gender/age.*

Dotplots

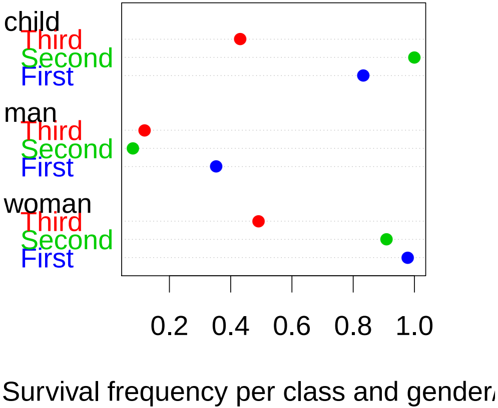

Figure 58: *Titanic: Survival by class and gender/age.*

Mosaic Plots

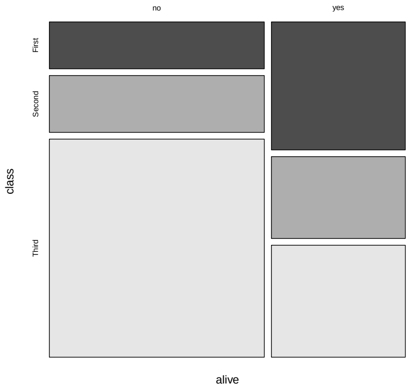

Figure 59: *Titanic: Survival and class.*

Mosaic Plots

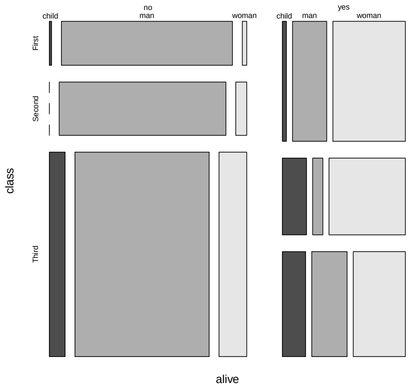

Figure 60: *Titanic: Survival, class, and gender/age.*

Multiple Qualitative Variables: Summary

The following types of plots are used to represent conditional distributions for multiple qualitative variables or counts for hierarchical categories.

- Multilevel donut/pie/sunburst plots.
  - Same or worse perception issues as with univariate pie charts.
  - Which variable to choose for "outer" layer?
- Barcharts/barplots.
  - For two categorical variables, a barchart/barplot displays the counts (or percentages) for each category of the second variable within each category of the first variable., i.e., conditional distribution of second variable given first.
  - Which variable to choose as "first"?
  - In a side-by-side barplot, the frequencies for the second variable are displayed as juxtaposed bars.

Multiple Qualitative Variables: Summary

- In a stacked/segmented barplot, the bars for the second variable are staked, so that their total height is the total count for the category of the first variable or 100 percent.
- Hard to compare frequencies between categories of first variable with both types of barplots.
- Hard to compare frequencies of second variable within categories of first variable with stacked barplot.
- Circular barcharts/barplots: Eye-catching, but even harder to compare frequencies.

- Treemap. The hierarchical or conditional frequencies are represented using nested figures, usually rectangles.
- Mosaic plots.
  - A mosaic plot is a graphical display of the counts in a contingency table (a.k.a., cross-tabulation or crosstab), where each cell is represented by a tile (i.e., rectangle) whose area is proportional to the cell frequency.

Multiple Qualitative Variables: Summary

- Color and shading of the tiles can be used to represent unusually large or small counts, the sign and magnitude of residuals (deviations) for particular models (e.g., independence).
- For two categorical variables, the width of each tile is proportional to the marginal frequency of the category for the first variable and the height of the tile to the conditional frequency of the category for the second variable given the first.
- Can be hard to read mosaic plots for more than two variables.

Conditional Plots

Conditional plots/coplots/faceting/panels/small multiples.

- Collection of plots, where each plot represents the conditional distribution of one or more variables given a conditioning variable.
- Each plot corresponds to a value or set of values for the conditioning variable. For a quantitative conditioning variable, the ranges are typically chosen so that there are equal numbers of observations in each panel.
- The scales on the axes, plotting symbols, colors, and legends have to be the same for all panels.
- How to arrange/order the panels?
  - Based on value of conditioning variable. Often natural sequence, e.g., chronological, alphabetical.

Conditional Plots

- Based on distributions displayed in each panel, e.g., decreasing slope of Y vs. X relationship, increasing overall mean/median of a time series.
- Geographically, e.g., on a map.
- Nodes in a network.

• E.g. Scatterplots of life expectancy vs. income for each of the six world regions.

Conditional Plots

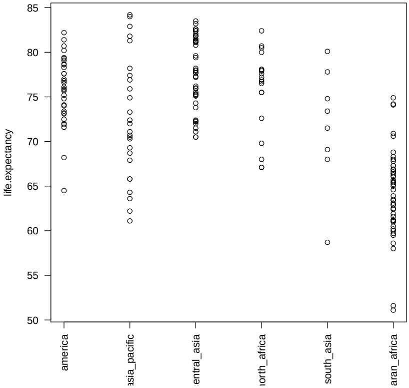

Figure 61: *Life expectancy by region, 2018.*

Conditional Plots

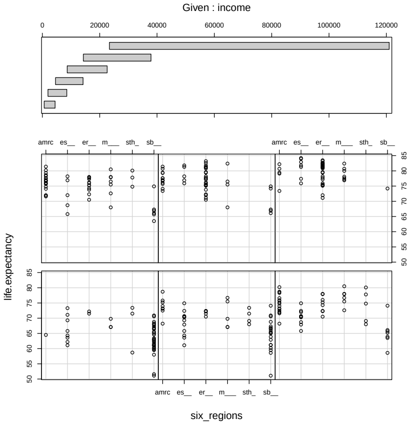

Figure 62: *Life expectancy by region conditioning on income, 2018.*

Quantile-Quantile Plots

- A quantile-quantile plot or QQ-plot is a scatterplot of the quantiles of one distribution against the corresponding quantiles of the other distribution.
- Quantile-quantile plots provide a detailed comparison of two distributions and can reveal in what way (simple or complicated) the two distributions differ, e.g., in terms of location, scale, skewness, kurtosis, outliers.
- Boxplots and density plots allow direct comparisons of commensurate values.
  For boxplots: Median (notch overlap), upper and lower-quartiles, IQR, outliers.
  For density plots: Mode, tails.

Quantile-Quantile Plots

- Compared to boxplots and density plots, QQ-plots make it easier to pinpoint in which ways the two distributions differ. They also provide better resolution than boxplots.
- Linear QQ-plot: Same distribution, possibly different means and variances, depending on intercept and slope. U-shaped QQ-plot: One distribution is skewed relative to the other.
- S-shaped QQ-plot: One distribution has heavier tails than the other, i.e., kurtosis.
- One can use a mean-difference version of a QQ-plot to compare quantiles, e.g., to see which distribution has larger values (i.e., is stochastically larger).
- Useful for model diagnostics, e.g., examining the distribution of residuals (observed minus fitted values).

Specialized Plots

Just scratching the surface, much more to be said about each of the topics below.

- Heatmaps/pseudocolor images. Graphical representation of a matrix, where the value of each element is represented using color. Choice of color palette is important.
- Correlation plots. Graphical representation of a correlation matrix using ellipse-shaped glyphs for each entry (i.e., level curve of a bivariate Gaussian density with the matching correlation), pseudocolor images.
- Spatial data/surfaces.
  - Spatial data/surfaces can be represented by triples (x, y, z), where x and y are coordinates on the plane and z is the variable of interest.

Specialized Plots

- Spatial data/surfaces can be displayed using perspective plots, pseudo color images, or contour plots.

- Clustering. Dendrograms, heatmaps.
- Graphs/networks.
  - Graph with vertices and edges.
  - E.g. Gene pathway, gene ontologies. Social networks.
- Maps and cartograms.
  - Many perception issues, e.g., perception of the size of areas is affected by the shape and color of those areas.
  - A cartogram is a map in which some thematic mapping variable (e.g., population, GNP) is substituted for land area or distance. This can lead to extreme distortion of the geometry of the map.
  - E.g. Gapminder (`https://www.gapminder.org/tools#$chart-type=map`).

Networks

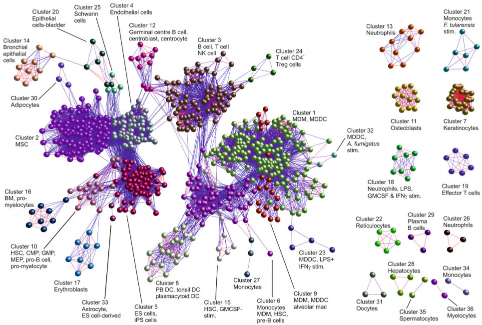

Figure 63: *Biological network.* `http://vizbi.org/blog/tag/data-visualization/`.

Maps

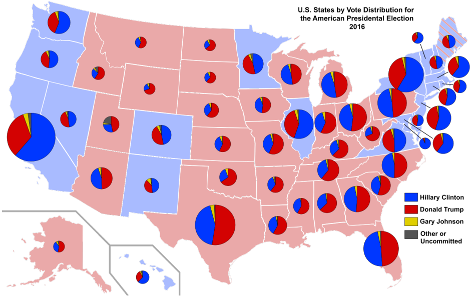

Figure 64: *2016 US presidential election.* `https://en.wikipedia.org/wiki/2016_United_States_presidential_election#Maps`.

Maps

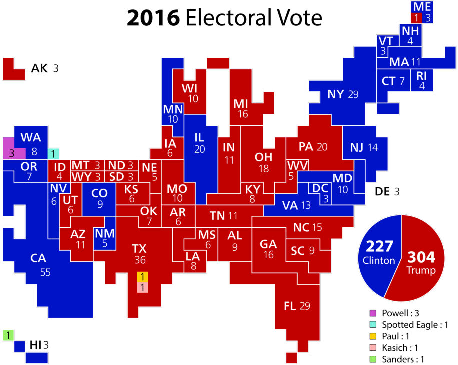

Figure 65: *2016 US presidential election.* https://en.wikipedia.org/wiki/2016_United_States_presidential_election#Maps.

Maps

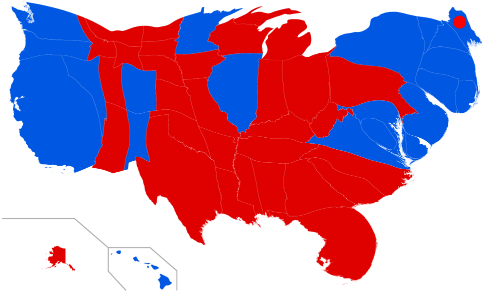

Figure 66: *2016 US presidential election.* `https://en.wikipedia.org/wiki/2016_United_States_presidential_election#Maps`.

Maps

Figure 67: *2016 US presidential election.* https://vanderbei.princeton.edu/JAVA/election2016/.

Maps

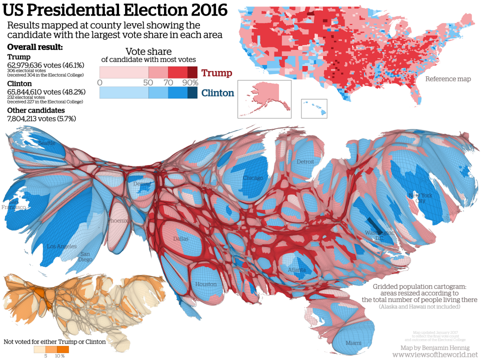

Figure 68: *2016 US presidential election.* http://www.viewsoftheworld.net/wp-content/uploads/2016/11/USelection2016Cartogram.png.

Customizing Plots

- Choose graphical parameters carefully as these can have a large impact on the plot.
  E.g. Aspect ratio, plotting symbols, line types, texture, color, axes, fonts, magnification, margins, grid, background, etc.
- The software makes a lot of decisions (not always the best) concerning graphical parameters.
- Experiment with different choices.

Customizing Plots

Provide sufficient information so that the plot can be interpreted properly.

- Annotation. Highlight points with text, arrows, boxes. Add auxiliary data, e.g., lines or rectangles indicating events or periods in time.
- Legend.
- Title.
- Caption.

Texture

Instead of color, one can use gray-scale shading lines to add texture to a plot.

- Orientation.
- Density.
- Contrast.

File Formats

- Raster graphics. Array of pixels, color each pixel (e.g., RGB).
  E.g. GIF, JPEG, PNG, TIFF.
- Vector graphics. Based 2D points, which are connected by lines and curves to form polygons and other shapes (i.e., objects).
  E.g. PDF, Postscript, SVG.
- As we zoom in/out, vector graphics scale, whereas raster graphics become highly pixelated.
- However, vector graphics file size increases with number of points and the graphics can be slow to render.

Dynamic and Interactive Graphics

- So far, we've only discussed static graphics.
- Graphs can be extended in several ways to add information and allow the viewer to interact with elements of the graph.
- Dynamic graphics. Allow addition of other variable (e.g., time, spatial location) to examine trends.
- Interactive graphics. Allow the viewer to focus on certain aspects of the graph and examine the data in an interactive and iterative manner.

Dynamic and Interactive Graphics

- D3.js. (https://d3js.org) JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards.
- Bokeh. (https://bokeh.pydata.org/) Python interactive visualization library that targets web browsers for presentation. Allows the creation of interactive plots, dashboards, and data applications. Bokeh has interfaces in Scala, Julia, and R.
- Trendalyzer. https://www.gapminder.org https://en.wikipedia.org/wiki/Trendalyzer.

Data Visualization

Dudoit

Motivation

Principles of Data Visualization

Do We Really Need a Graph?

General Considerations

Graphical Perception

Bad Graphs

Survey of Data Visualization Techniques

One Quantitative Variable

Multiple Quantitative Variables

One Qualitative Variable

Multiple Qualitative Variables

Conditional Plots

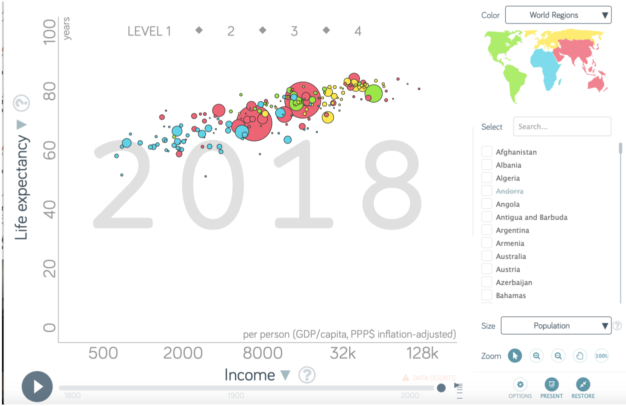

Figure 69: *Gapminder: Dynamic bubble chart.* Animation adds time as a fifth variable. `https://www.gapminder.org/tools/?from=world#$state$time$value=2018;;&chart-type=bubbles`.

Gapminder

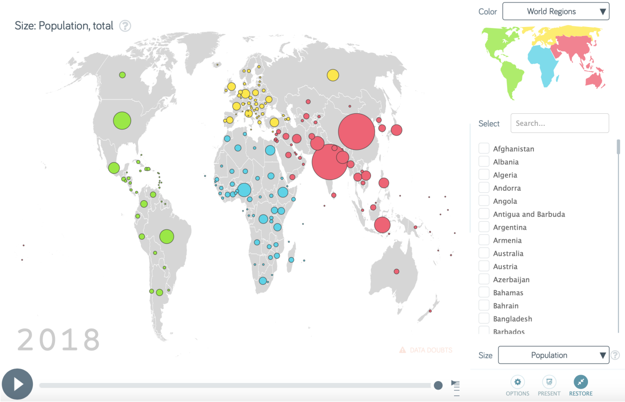

Figure 70: *Gapminder: Dynamic map.* Total population by country. `https://www.gapminder.org/tools#$state$time$value=2018;;&chart-type=map`.

General Considerations

- First, "get it right in black and white".
  E.g. Use luminance for grayscale version.
- Do not add color gratuitously.
- Think about the purpose of adding color.
  E.g. Representing different categories or groups in the data (e.g., by coloring plotting symbols, lines, boxplots, barplots), representing continuous values in a heatmap or on a map.
- Apply visual perception principles to make proper color choices for the type of message and data.
- Ideally, aim for linearity between actual and perceived signal, but not easy because of Stevens' Law.

General Considerations

- Color perception and context. Can distinguish fewer colors with small areas. Consider surrounding colors and contrasts.
- Medium.
  - Does the medium on which the graph is to be shown support color?
  - If so, does it render color properly?
  - Color calibration software is available for output devices.
  - E.g. Black-and-white print journal or book, computer monitor, LCD projector.
- Audience. Can the audience perceive color? If not, what are the constraints on color choices?
  E.g. Color blindness.

Color Vision

- Color vision is the ability to distinguish objects based on the wavelengths of the light they emit, reflect, or transmit.
- Perception of color is a subjective process whereby the brain responds to the stimuli that are produced when incoming light reacts with cone and rod cells in the eye.
- Accordingly, color can be quantified in various ways.
- There are two main complementary theories of color vision, the trichromatic theory and the opponent process theory.
- The trichromatic or RGB theory (YoungHelmholtz, 1802, 1850) states that the retina's three types of cones are preferentially sensitive to red, green, and blue.
- The opponent process theory (Hering, 1892) states that colors are perceived in an antagonistic way: Red vs. green, blue vs. yellow, black vs. white.

Color Vision

- There are two types of photoreceptor cells in the human retina: Rod cells and cone cells. These cells absorb light and convert it into an electrical signal that is passed to the brain through the optic nerve.
- Rod cells are responsible for black-and-white, peripheral, and night vision.
- Cone cells or cones are responsible for visual details, color, central, and day vision.
- There are three types of cone cells, each of which detects colored light of a different wavelength:
    - the L cone type responds the most to long-wavelength light, peaking at about 560 nm, i.e., red;
    - the M cone type responds the most to medium-wavelength light, peaking at 530 nm, i.e., green;

Color Vision

- the S cone type responds the most to short-wavelength light, peaking at 420 nm, i.e., blue and violet.

Color Vision

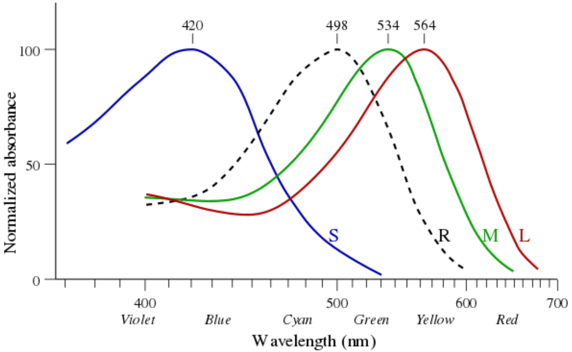

Figure 71: *Cone and rod cells.* Wavelength responsiveness of short (S), medium (M), and long (L) cones and rods (R).

Color Systems

There are different types of systems/models for representing and quantifying color, each using three dimensions that focus on different aspects of color perception.

- RGB, for red, green, and blue.
- HSL, for hue, saturation, and lightness.
- HSB/HSV, for hue, saturation, and brightness/value.
- CIELAB or LAB.
- CMYK, for cyan, magenta, yellow, and key/black.

Demo for color system conversion: `http://colorizer.org`.

Color Systems

Figure 72: *Color system conversion.* `http://colorizer.org`.

RGB Color System

- In the RGB model, a color is described by the amount of red, green, and blue it contains, each typically normalized to the range $[0, 1]$ or $[0, 255]$.
- The set of possible colors thus corresponds to points (R, G, B) lying in a cube.

RGB Color System

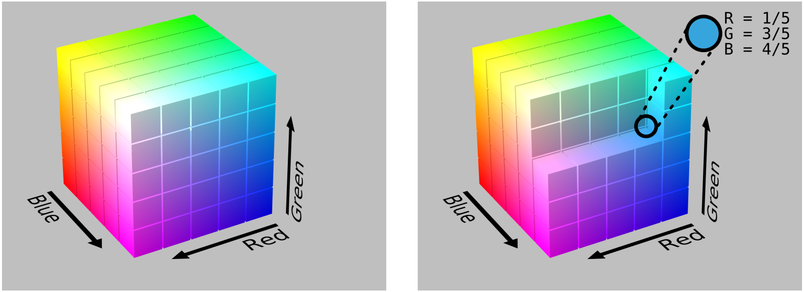

Figure 73: *RGB cube.*
https://en.wikipedia.org/wiki/HSL_and_HSV.

Opponent Process Color System

- The opponent process color theory states that the human visual system interprets information about color by processing signals from cones and rods in an antagonistic manner.
- There are three opponent channels: Red vs. green, blue vs. yellow, and black vs. white (the latter is achromatic and detects light-dark variation or luminance).
- Responses to one color of an opponent channel are antagonistic to those to the other color. That is, one cannot simultaneously perceive both redness and greenness or blueness and yellowness.

Opponent Process Color System

- The opponent process system describes color in terms of three parameters: Where it lies on a light-dark scale, where it lies on a red-green scale, and where it lies on a yellow-blue scale.
- The opponent process model is followed by the CIELAB (CIE 1976 L*a*b* or LAB) system, developed by the International Commission on Illumination. This system expresses color as three numerical values, L* for the lightness and a* and b* for the green-red and blue-yellow color components, respectively.

Opponent Process Color System

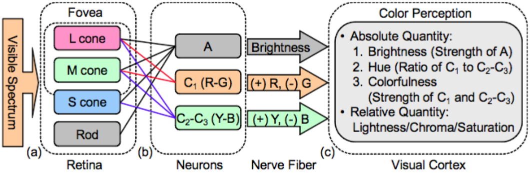

Figure 74: *Opponent process color system.* https://en.wikipedia.org/wiki/Opponent_process.

RGB and Opponent Process Color Systems

- The RGB system is external, i.e., corresponds to the stimulation of the receptors in our eyes and can thus be examined by applying external stimuli.
- The opponent process system is internal and is less subject to experimentation.
- Although the RGB and opponent models provide a full description of color, they are not natural for thinking about color.

Color Perception

We more naturally think of color in terms of the following three types of perception parameters which can be defined as functions of the non-perceptually-based RGB coordinates.

- Hue, H.
  - Hue is a property of color corresponding to the wavelength of light.
  - It is the degree to which a stimulus can be described as similar to or different from stimuli that are described as red, green, blue, and yellow, i.e., "pure" colors found in the spectrum and not mixed with black or white.
  - Hue is 0 for red, 120 for green, and 240 for blue.
- Saturation/Purity, S.

Color Perception

- In the HSB/HSV system, saturation is the amount of white mixed with a pure color (0 is white, 1 is pure color). For example, pink is red mixed with white, i.e., partly desaturated red.
- In the HSL system, saturation is the amount of gray mixed with a pure color (0 is gray, 1 is pure color).

- Brightness/Value, B/V. Amount of black mixed with a pure color (0 is black, 1 is pure color).
  Lightness, L. Amount of black and white (grayscale) mixed with pure color (0 is black, 1 is white, pure color is 0.5).

Color Perception

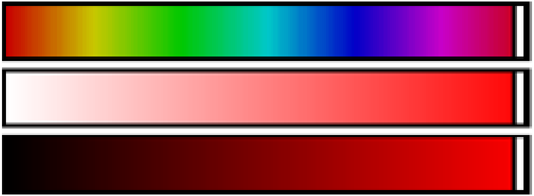

Figure 75: *Hue, saturation, and value.*

HSB/HSV and HSL Color Systems

- As the RGB system, the HSB/HSV and HSL systems represent color using three coordinates.
- HSB/HSV: Color represented by hue, saturation, and brightness/value.
- HSL: Color represented by hue, saturation, and lightness.
- The HSB/HSV and HSL geometries are transformations of the RGB unit cube.

HSB/HSV and HSL Color Systems

Let R, G, and $B \in [0,1]$ denote, respectively, the amounts of red, green, and blue light in a color.

- Let $M = \max(R, G, B)$ and $m = \min(R, G, B)$.
- Chroma.

$$C = M - m \in [0,1].$$

- Hue. The color type (such as red, blue, or yellow).

$$H = 60^\circ \begin{cases} \frac{G-B}{C} \mod 6, \text{if } M = R \\ \frac{B-R}{C} + 2, \text{if } M = G \\ \frac{R-G}{C} + 4, \text{if } M = B \end{cases} \in [0^\circ, 360^\circ).$$

If $C = 0$, H is either undefined or set to 0° for neutral colors.
Each value corresponds to one color: 0 for red, 120 for green, and 240 for blue.

HSB/HSV and HSL Color Systems

- Value/Brightness.

$$V = M = \max(R, G, B) \in [0, 1].$$

0 is always black. Depending on saturation, 1 may be white or a more or less saturated color.

- Lightness.

$$L = \frac{1}{2}(M + m) \in [0, 1].$$

0 for black, 1 for white.

- Intensity.

$$I = \frac{1}{3}(R + G + B) \in [0, 1].$$

- Saturation/Purity. The exact definition depends on the color system.

$$
\begin{aligned}
S_{HSV} &= \begin{cases} 0, \text{if } V = 0 \\ \frac{C}{V} = 1 - \frac{m}{M}, \text{ow} \end{cases} \\
S_{HSL} &= \begin{cases} 0, \text{if } L = 1 \\ \frac{C}{1-|2L-1|}, \text{ow} \end{cases} \\
S_{HSI} &= \begin{cases} 0, \text{if } I = 0 \\ = 1 - \frac{m}{I}, \text{ow} \end{cases} .
\end{aligned}
$$

Ranges from 0 to 1.
0 means no color, that is a shade of grey between black and white; 1 means intense color.

HSB/HSV Color System

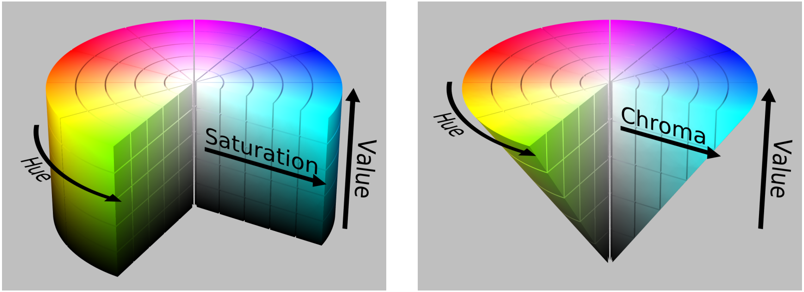

Figure 76: *HSB/HSV geometries.*
`https://en.wikipedia.org/wiki/HSL_and_HSV`.

HSL Color System

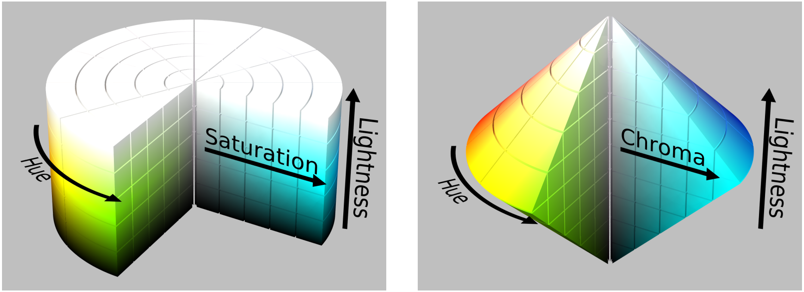

Figure 77: *HSL geometries.*
https://en.wikipedia.org/wiki/HSL_and_HSV.

Color Palettes

- A color palette is a set of colors (cf. a painter's palette).
- Different palettes are appropriate for different types of data.
- Categorical data.
    - Use hue to provide unique identity.
    - How many categories can we readily discriminate on a single plot?
      6–12, including background, default colors, text color.
    - Use fully saturated, opponent colors which have "names"; makes it easier to discuss.
    - When there are too many categories, explicitly collapse/group categories in a contextually meaningful manner.
- Ordinal data.

Color Palettes

- Map monotonic range in data to monotonic lightness range: Typically, low values in data map to lighter values, and high values in data map to darker values.
- Use medium saturation for large areas and high saturation for more important, small features (e.g., points, lines).
- Using 2 hues emphasizes large-scale structure.
- Using more hues emphasizes mid-scale structure (e.g., rainbow of hues), but serious perceptual drawbacks.
- Ideally match the end points (and others) to semantically meaningful colors corresponding to features of the data, e.g., white for snow-capped mountains, blue for water, green for land.

Color Palettes

ColorBrewer provides the following three types of palettes (`colorbrewer2.org/`).

- Sequential palettes are suited to ordinal data that progress from low to high. Lightness steps dominate the look of these palettes, with light colors for low data values and dark colors for high data values.
- Qualitative palettes are suited to categorical data. They do not imply magnitude differences between classes. Hues are used to create the primary visual differences between classes.

Color Palettes

- Diverging palettes put equal emphasis on mid-range and extremes values at both ends of the data range. The critical class or break for middle values is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues.

Color Palettes

- seaborn color palettes, including ColorBrewer: `https://seaborn.pydata.org/tutorial/color_palettes.html`.
- matplotlib color palettes: `https://matplotlib.org/users/colormaps.html`

Color Palettes

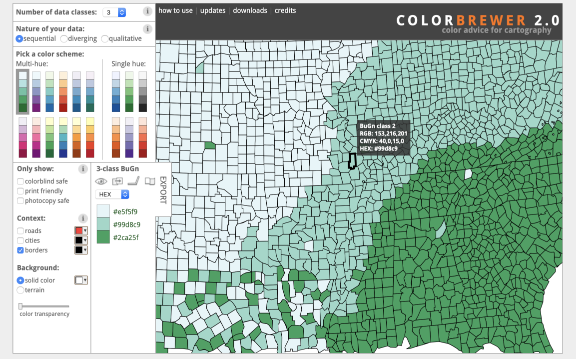

Figure 78: *ColorBrewer*. http://colorbrewer2.org/.

Color Palettes

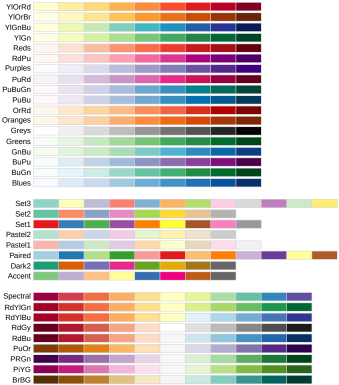

Figure 79: *ColorBrewer.* Sequential, qualitative, and diverging color palettes.

Color Blindness

- Color blindness affects about 8% of the population, primarily males, but also .5% of females (X-linked gene). `http://www.color-blindness.com`.
- Different types of color blindness: Most common can't distinguish red and green, others blue and yellow. But can distinguish lightness.
- Don't rely on hue, but use luminance and saturation instead.
- Avoid red-green divergence colormaps.
- Tools available to convert images to how they look for different types of color vision deficiencies. E.g. `http://www.vischeck.com/vischeck/`.

References

- Peter Aldhous. Data visualization: basic principles. http://paldhous.github.io/ucb/2016/dataviz/week2.html.
- Ross Ihaka. Statistics 120 – Information Visualisation. https://www.stat.auckland.ac.nz/~ihaka/120/.
- Duncan Temple Lang. Data Visualization Workshops. http://dsi.ucdavis.edu/tag/data-visualization.html.

W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985.

A. Gelman, C. Pasarica, and R. Dodhia. Lets practice what we preach: Turning tables into graphs. *The American Statistician*, 56(2):121–130, 2002.

E. J. Marey. *La Mthode Graphique*. Librairie de l'Académie de Médecine, 1885.

References

S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3): 153–181, 1957.

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd edition, 2001.