

Tabulation and Visualization

Department of Government
London School of Economics and Political Science

1 Getting a grip on data

2 Tabulation

3 Visualization

1 Getting a grip on data

2 Tabulation

3 Visualization

Types of Measures

- 1 Categorical
 - Binary
 - 2 Ordinal
 - 3 Interval
- Qualitative
- Quantitative

Note: *Ratio* scale measures are interval measures with a non-arbitrary zero value

Definitions

- Statistic: “a quantitative summary of a variable for a set of units”

Central Tendency

- Mean (average): $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$

Central Tendency

- Mean (average): $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
 - Trimmed mean

Central Tendency

- Mean (average): $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
 - Trimmed mean
- Median: Middle value

Central Tendency

- Mean (average): $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
 - Trimmed mean
- Median: Middle value
- Mode: Most common value

Central Tendency

- Mean (average): $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
 - Trimmed mean
- Median: Middle value
- Mode: Most common value
 - Unimodal, bimodal, multimodal

Dispersion/variation

- (Element) Variance:

$$\text{Var}(Y) = s_Y^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

Dispersion/variation

- (Element) Variance:

$$\text{Var}(Y) = s_Y^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

- Standard Deviation:

$$sd(Y) = s_Y = \sqrt{\text{Var}(Y)}$$

Dispersion/variation

- (Element) Variance:

$$\text{Var}(Y) = s_Y^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

- Standard Deviation:

$$sd(Y) = s_Y = \sqrt{\text{Var}(Y)}$$

- Median absolute deviation (MAD):

$$MAD = \text{median}(|Y_i - \text{median}(Y)|)$$

Shape

- Skewness

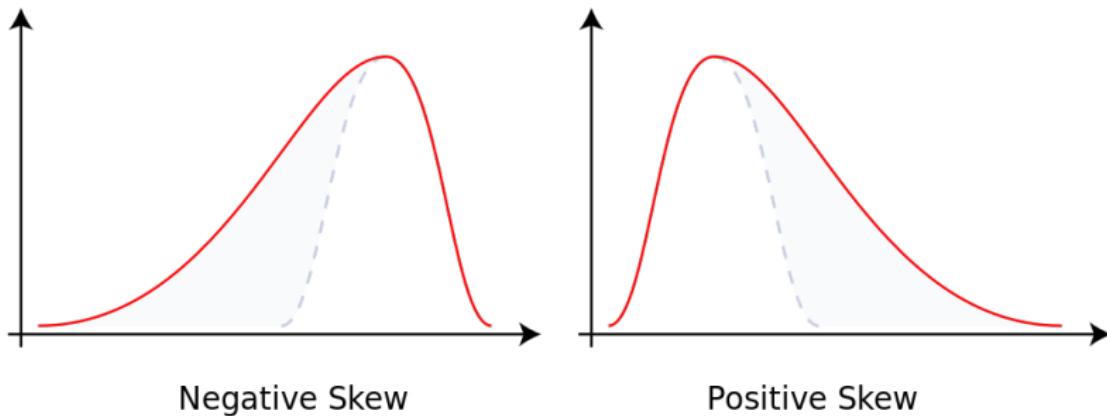
Shape

- Skewness
 - Positive/right skew
 - Symmetric
 - Negative/left skew

Shape

- Skewness
 - Positive/right skew
 - Symmetric
 - Negative/left skew
- Kurtosis: peakedness of a distribution

Skewness



Source: Rodolfo Hermans (Wikimedia)

Relationship

- Covariation:

$$\text{Cov}(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Relationship

- Covariation:

$$\text{Cov}(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- Correlation:

$$\text{Corr}(X, Y) = r_{x,y} = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)s_x s_y}$$

In R...

- `mean()`
- `median()`
- `var()`
- `sd()`
- `cov()`
- `cor()`

1 Getting a grip on data

2 Tabulation

3 Visualization

Table

- Definition: “an arrangement of information into rows and columns”
- Tables can show:
 - Values
 - Counts
 - Proportions
 - Summary statistics

In R...

- `table()`
- `prop.table()`
- `ftable()`
- `aggregate()`

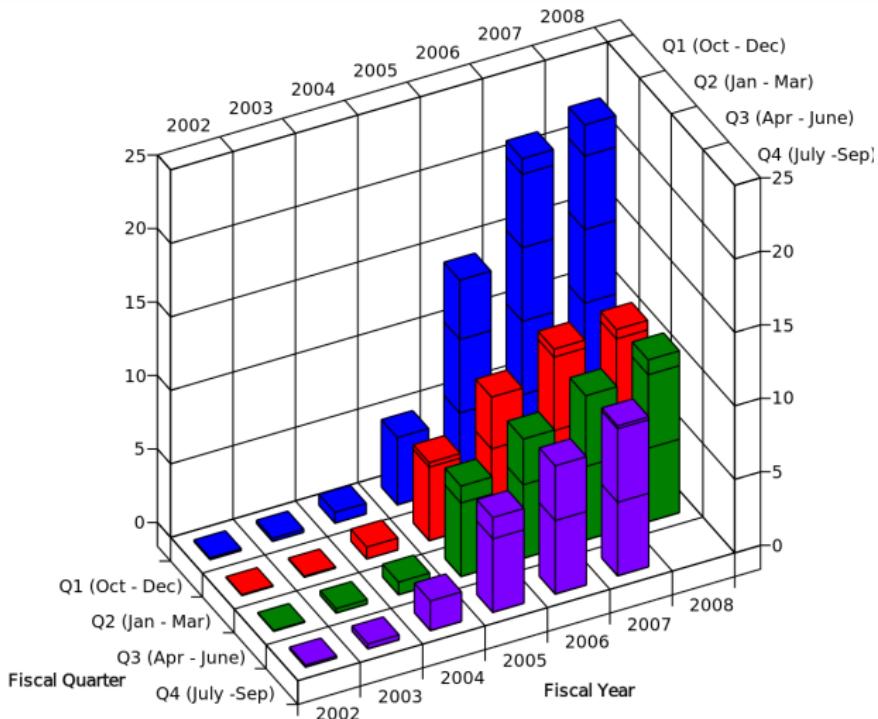
1 Getting a grip on data

2 Tabulation

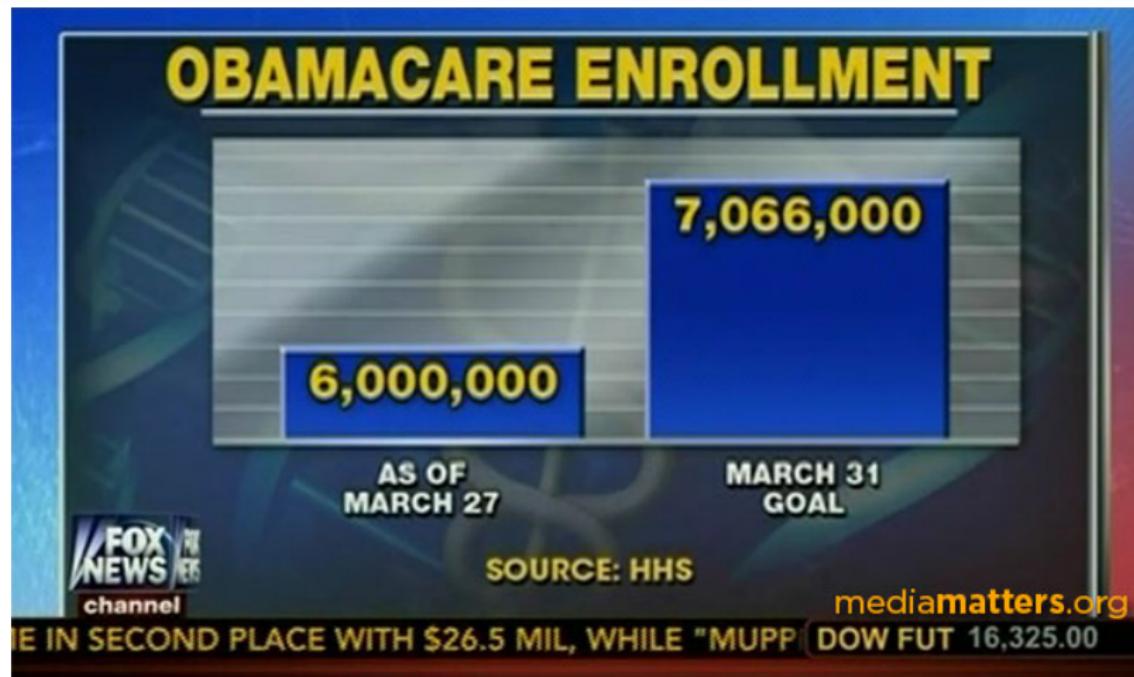
3 Visualization

Bad visualizations . . .

iPod sales per fiscal quarter till June 2008

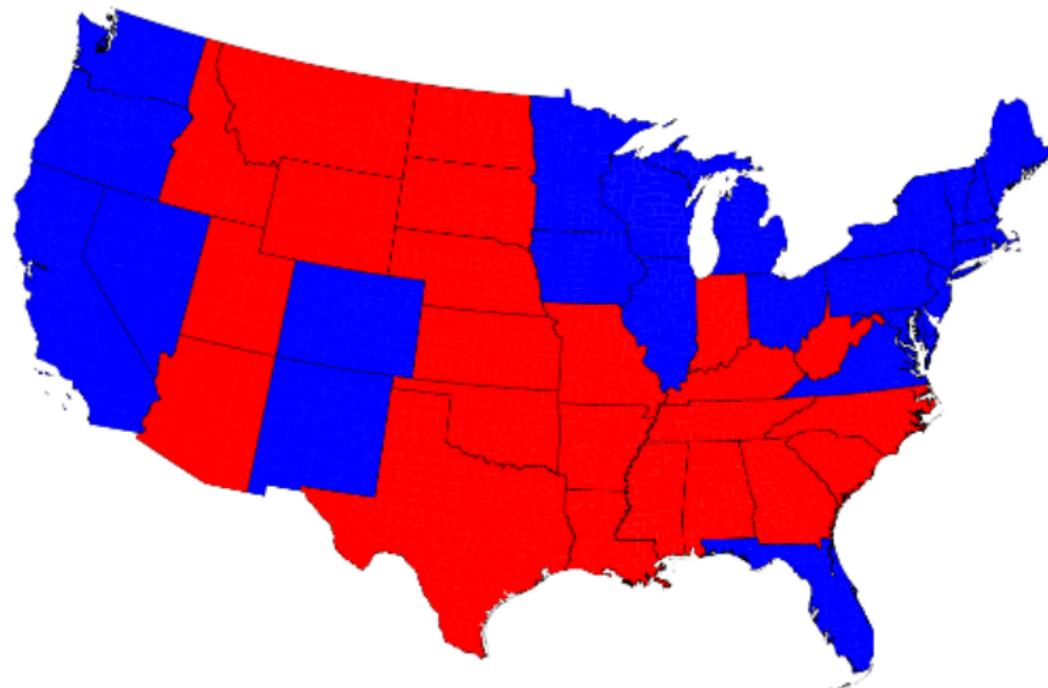


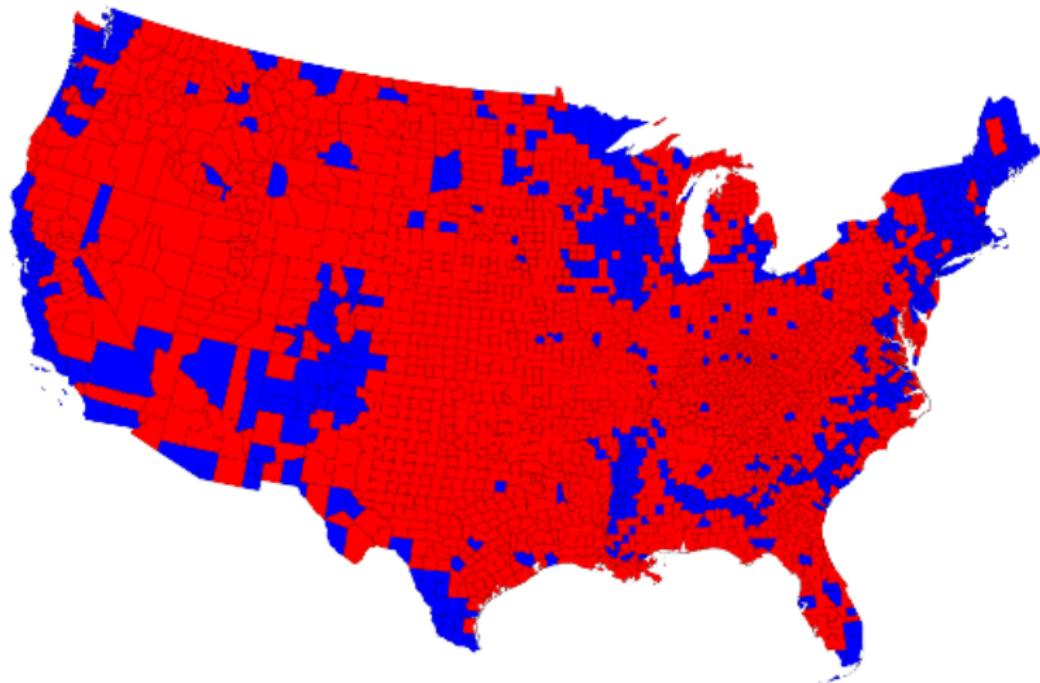


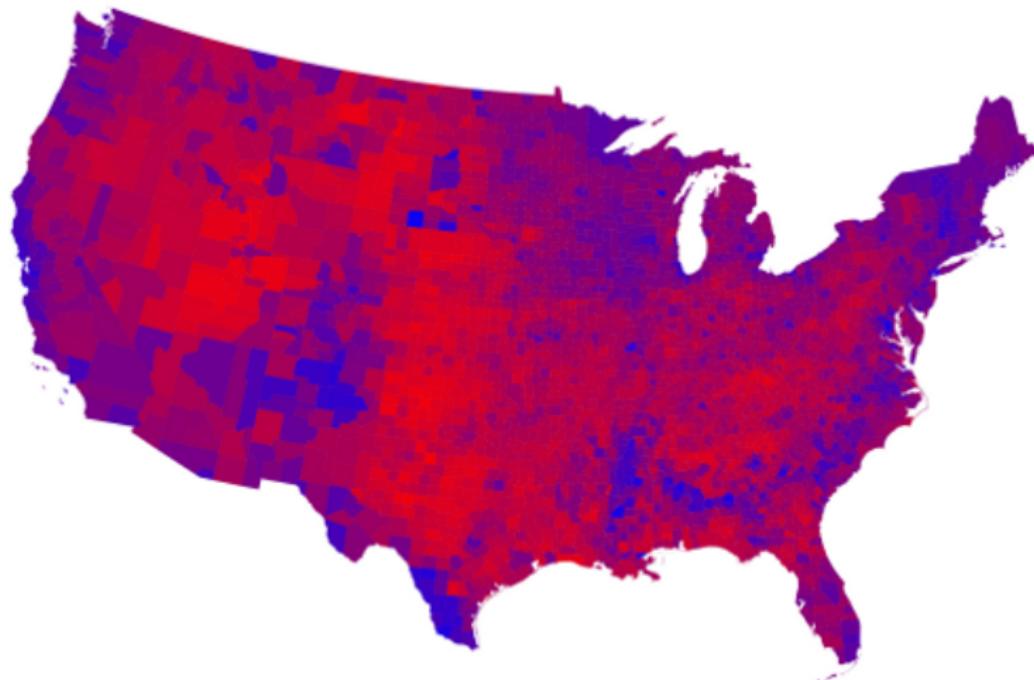


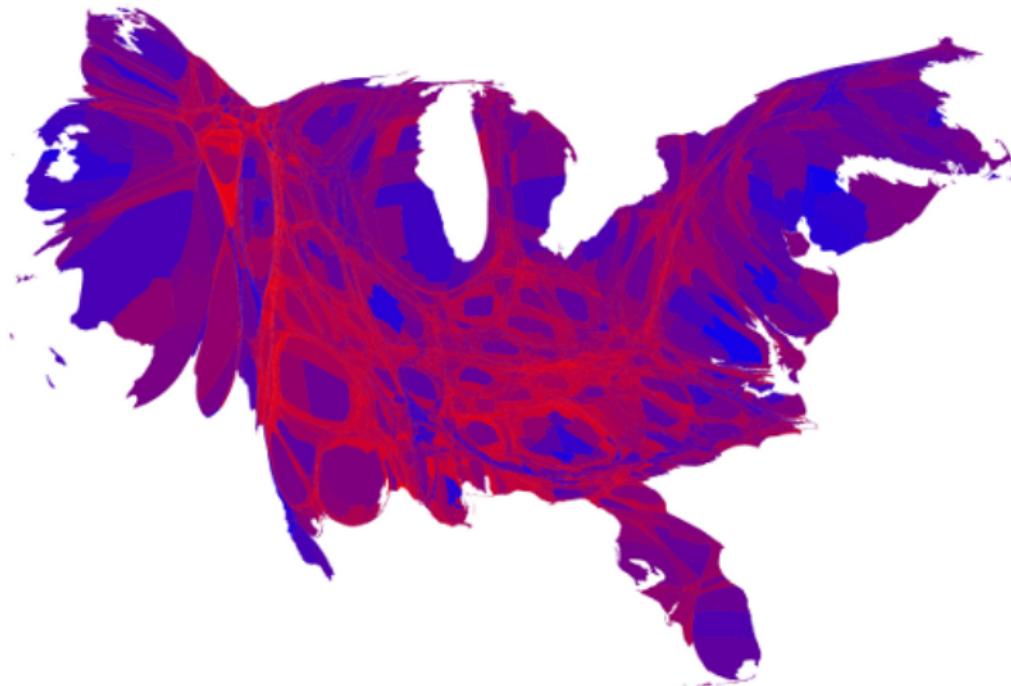
Source: MediaMatters











Visualizations

- Definition: “Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.” (Tufte, 2001)

Tufte, E. 2001. *The Visual Display of Quantitative Information*. Graphics Press.

Anscombe's Quartet

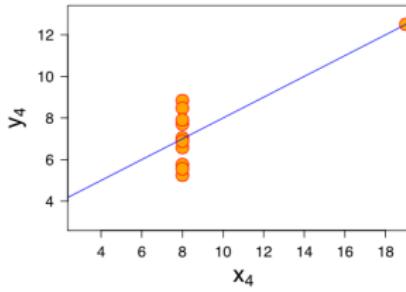
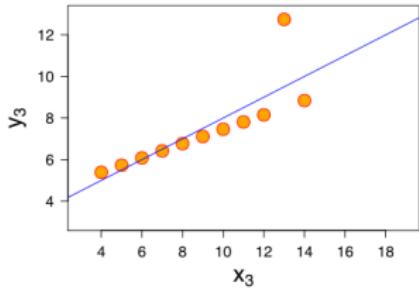
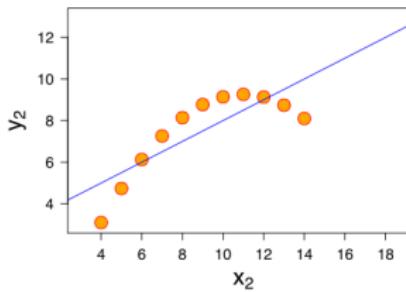
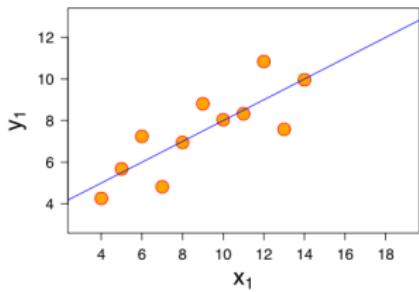
	I	II		III		IV	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

$$\bar{X} = 9, \text{Var}(X) = 11,$$

$$\bar{Y} = 7.5, \text{Var}(Y) = 4.12,$$

$$\text{Corr}(X, Y) = 0.816$$

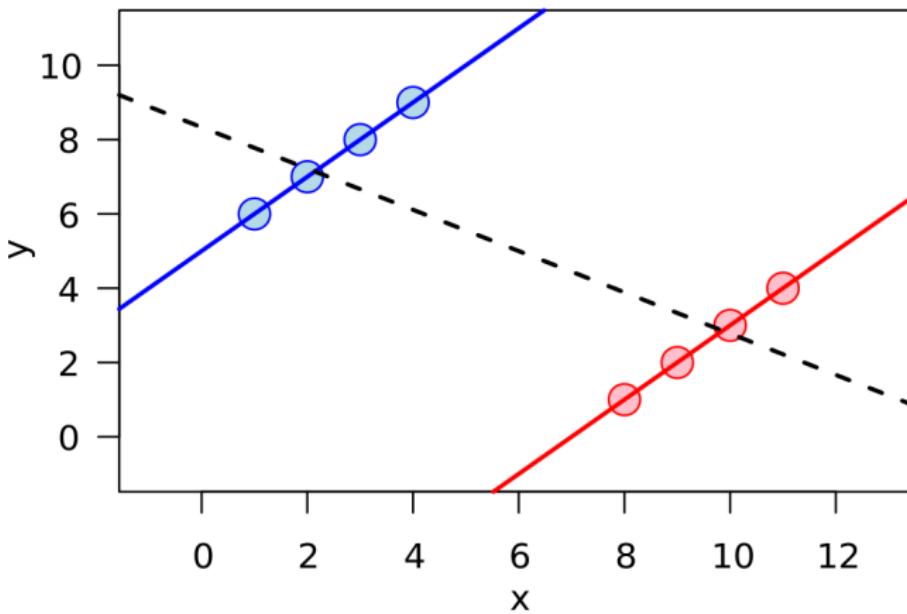
Anscombe's Quartet



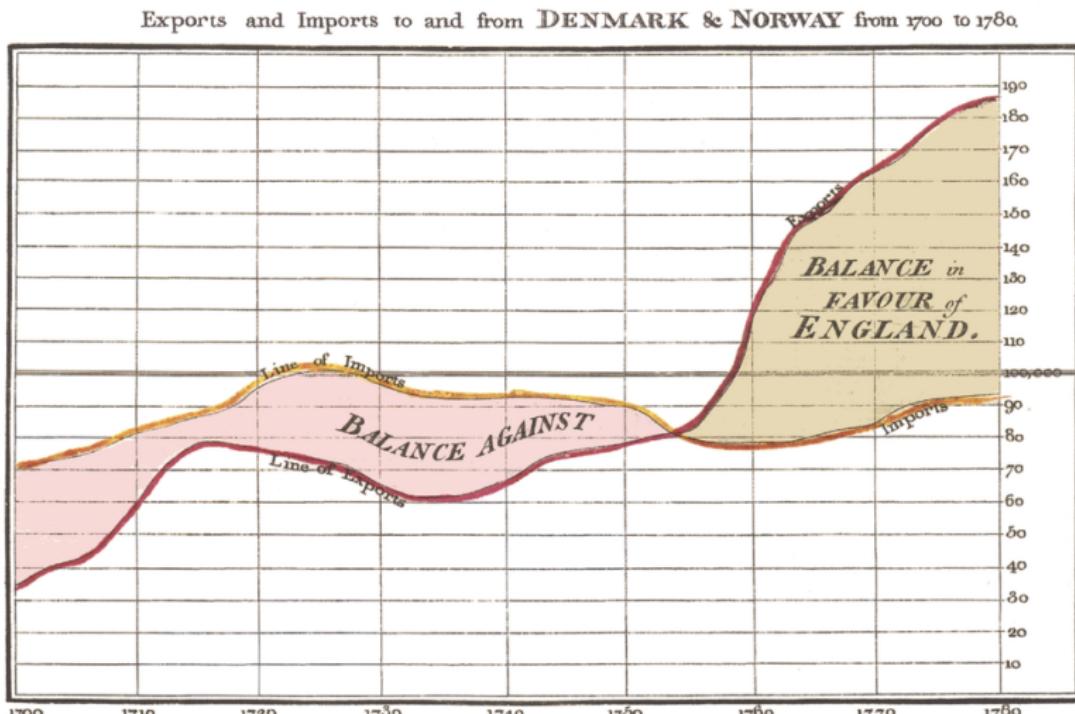
Simpson's Paradox

Source: Wikimedia

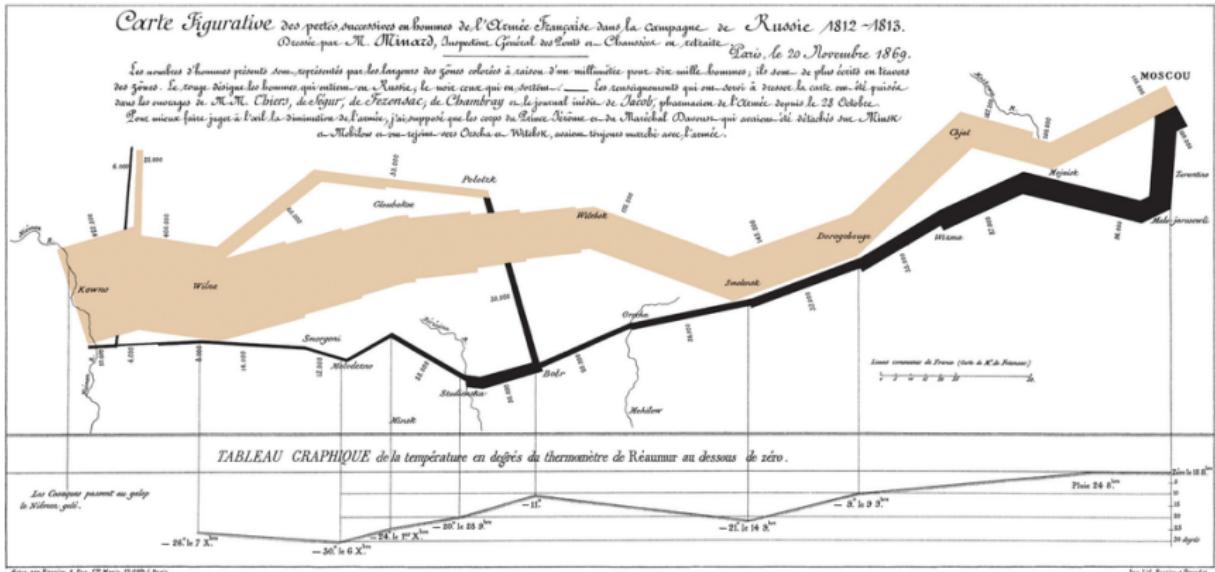
Simpson's Paradox



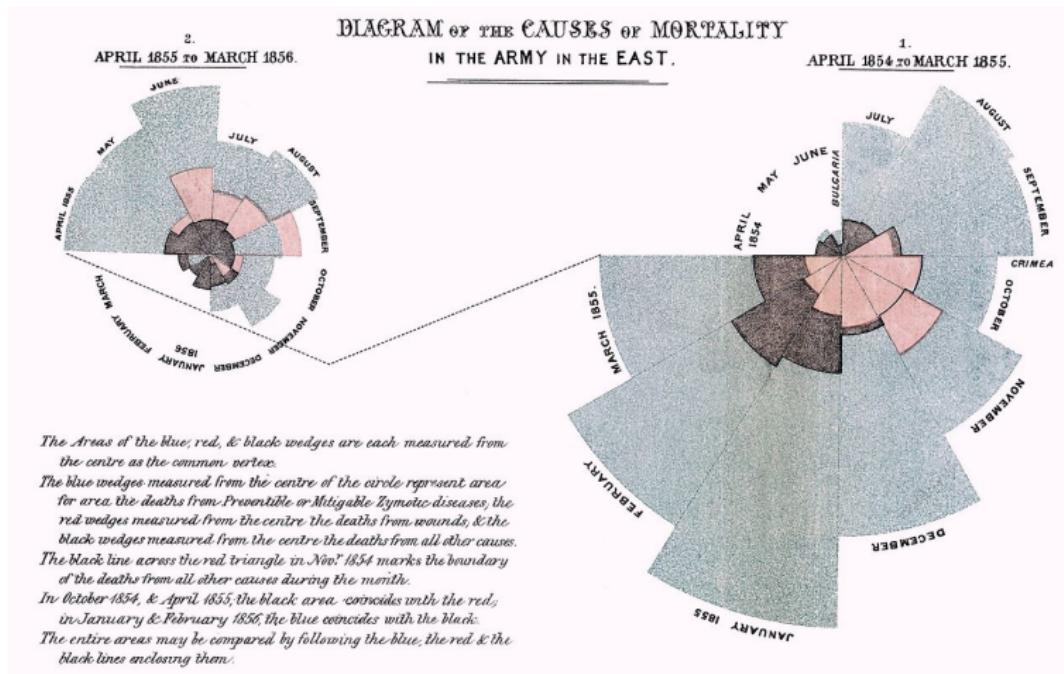
William Playfair



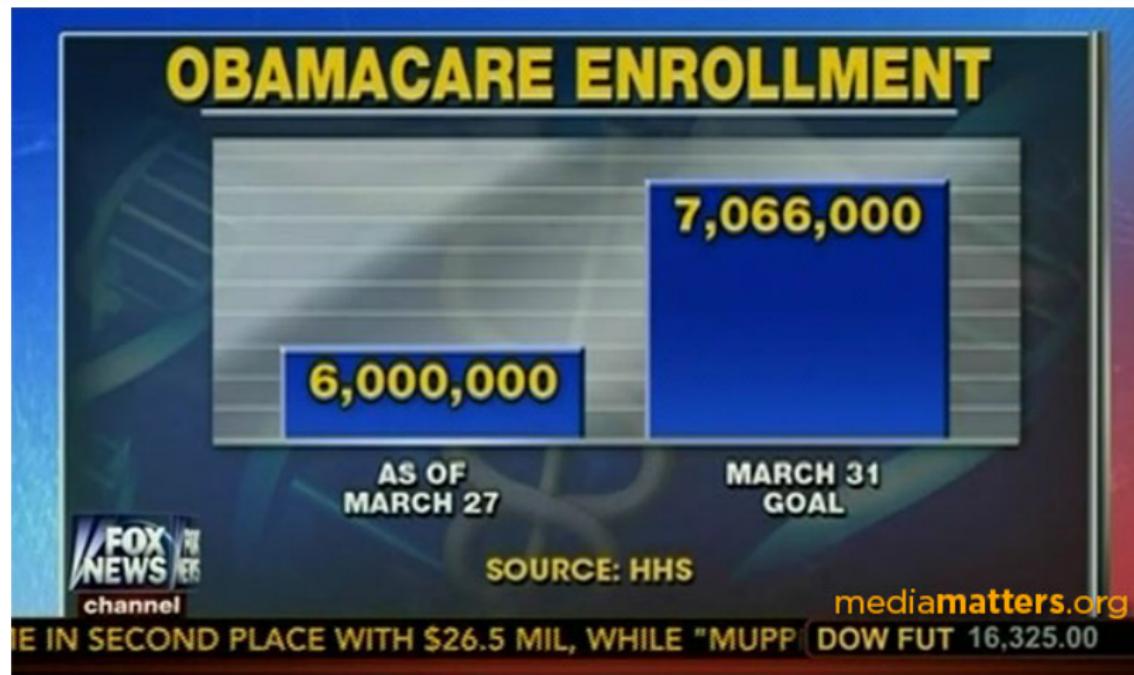
Charles Minard



Florence Nightingale

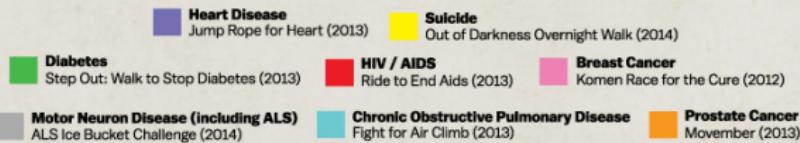


1 Be honest

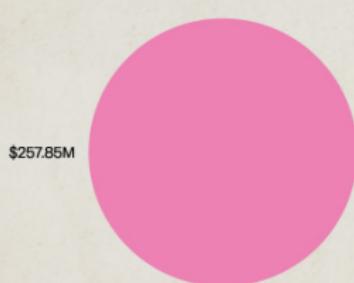




WHERE WE DONATE VS. DISEASES THAT KILL US



MONEY RAISED



DEATHS (US)



WHERE WE DONATE VS. DISEASES THAT KILL US

█ Heart Disease
Jump Rope for Heart

█ Diabetes
Step Out: Walk to Stop Diabetes

█ Motor Neuron Disease (including ALS)
ALS Ice Bucket Challenge

█ Suicide
Out of Darkness Overnight Walk

█ HIV / AIDS
Ride to End Aids

█ Chronic Obstructive Pulmonary Disease
Fight for Air Climb

█ Breast Cancer
Komen Race for the Cure

█ Prostate Cancer
Movember

MONEY RAISED

\$257.85M



\$147M



\$54.1M



\$22.9M



DEATHS (US)

596,577



142,942



73,831

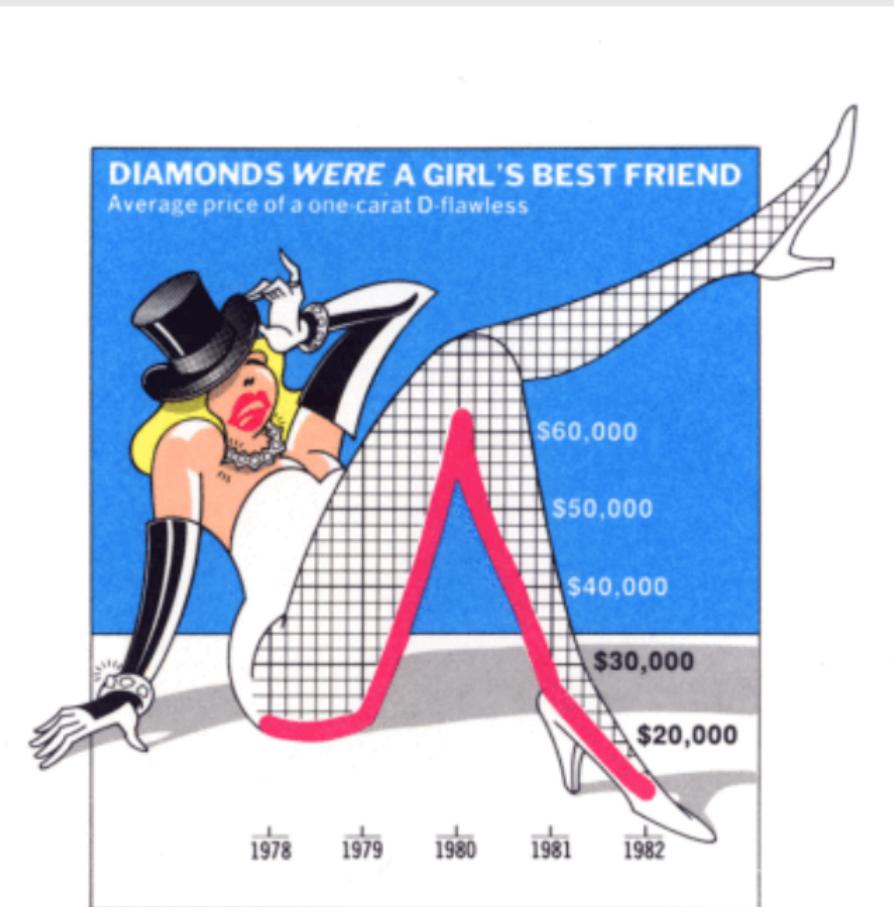


41,374



1 Be honest

- 1 Be honest
- 2 Data-Ink Ratio





max

75%

50%

25%

min



T_1

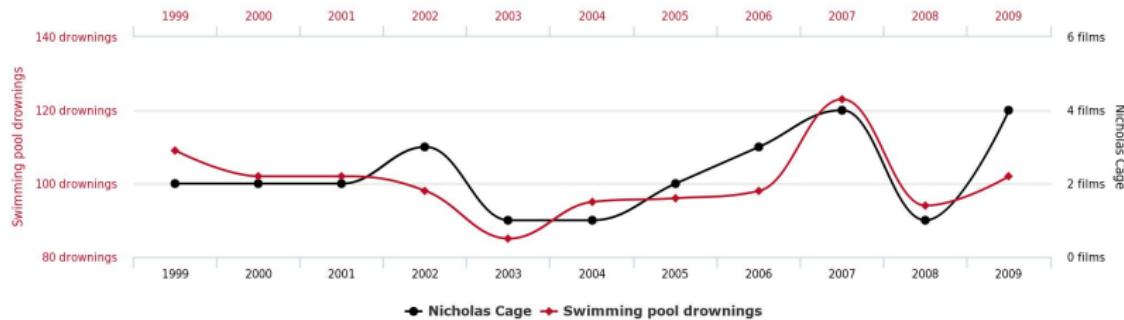
T_2 , T_4

T_3

- 1 Be honest
- 2 Data-Ink Ratio

- 1 Be honest
- 2 Data-Ink Ratio
- 3 Tell a story

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

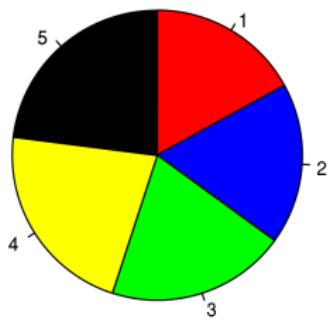
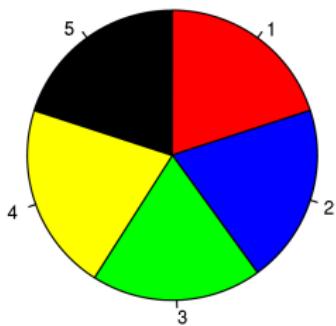
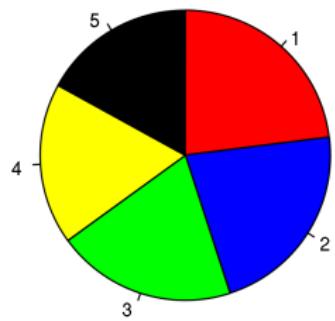


tylervigen.com

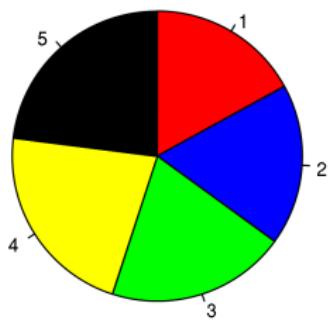
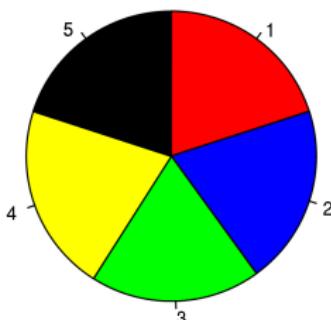
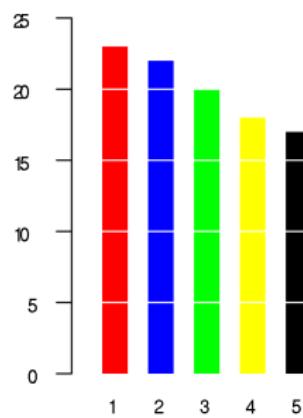
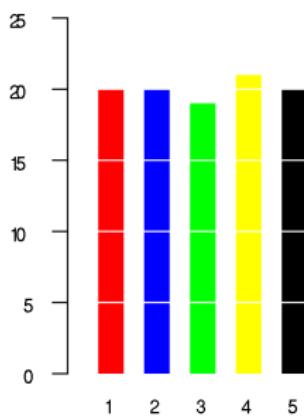
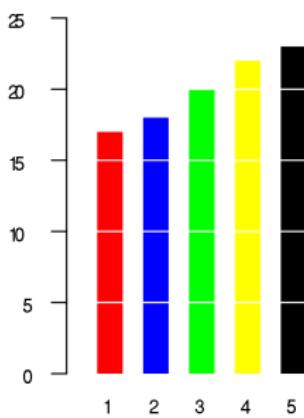
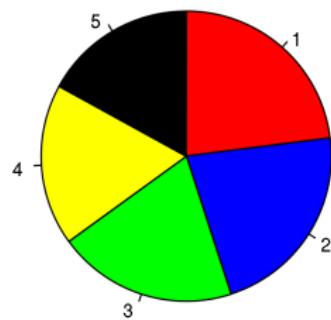
Source: CC-BY Tyler Vigen

- 1 Be honest
- 2 Data-Ink Ratio
- 3 Tell a story

- 1 Be honest
- 2 Data-Ink Ratio
- 3 Tell a story
- 4 Steer reader's attention

A**B****C**

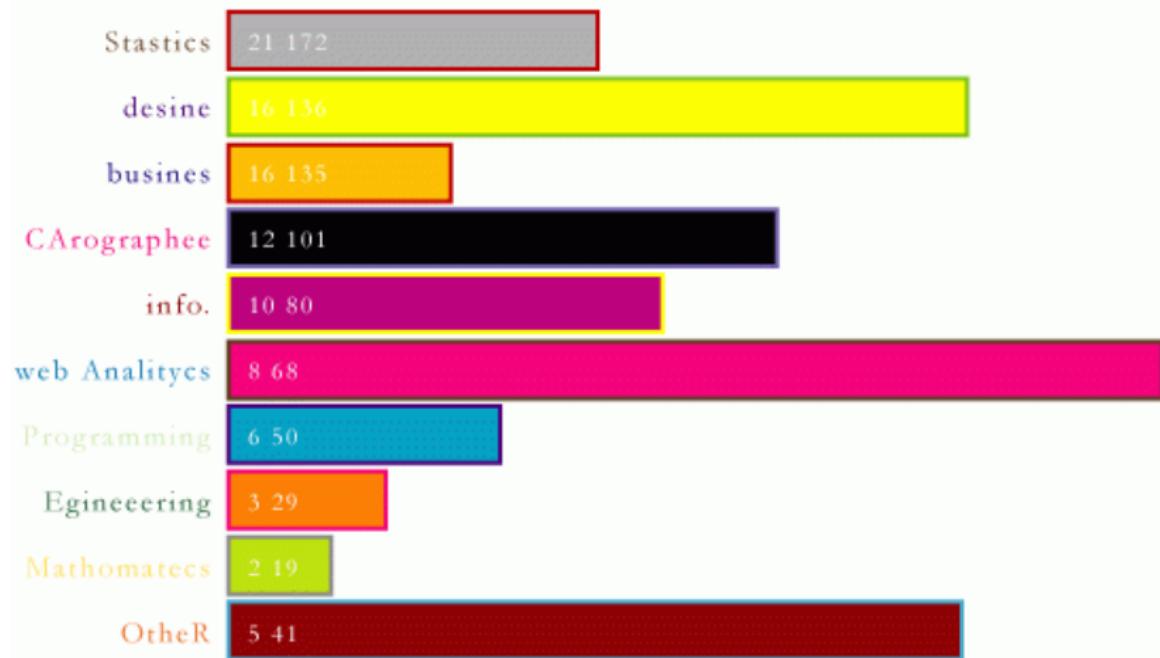
Source: Wikimedia

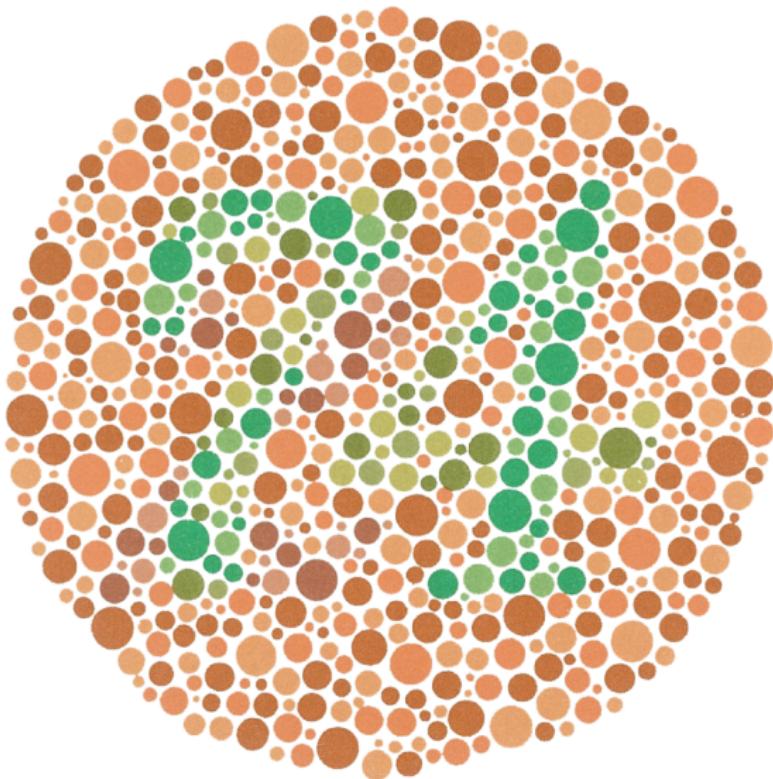
A**B****C**

- 1 Be honest
- 2 Data-Ink Ratio
- 3 Tell a story
- 4 Steer reader's attention

- 1 Be honest
- 2 Data-Ink Ratio
- 3 Tell a story
- 4 Steer reader's attention
- 5 Use balanced colour palettes

POLL R3sult: wha Data related area r u Most Interested





The bottom line

A visualization should be a display of quantitative (and/or qualitative) data that tells an information-rich story in an honest and beautiful manner.

The bottom line

A visualization should be a display of quantitative (and/or qualitative) data that tells an information-rich story in an honest and beautiful manner.

Questions?

In R...

R has 5+ graphics “systems”

- Base graphics
- The **lattice** package
- The **ggplot2** package
- The **plotrix** package
- The **htmlwidgets** package +
JavaScript's d3 library

ggplot2

- Most coherent graphics system
- Based on a “grammar” of graphics
- Easily customized using various “themes”
 - Some built-in to ggplot2
 - Some in an add-on package (**ggthemes**)

A bit about the grammar

- `ggplot()` creates a plot object
- `aes` describes a mapping of data to a visual element (e.g., color, shape, etc.)
- `geom_()` displays a particular graphical representation
- `scale_()` modifies the axes
- `coord_()` modifies the coordinate system
- `theme_()` modifies the overall look
- `facet_()` creates small multiples

```
library("rio")
d <- import("http://www.qogdata.pol.gu.se/data/qog_std_cs_jan15.dta")

summary(d$wef_lifexp) # life expectancy
summary(d$fh_polity2) # Polity scores
summary(d$gle_cgdpc) # GDP
summary(d$dpi_finter) # executive term limits
summary(d$bti_cr) # civil rights index

library("ggplot2")
p <- ggplot(d)
p + aes(x = fh_polity2) + geom_density()
p + aes(x = fh_polity2) + geom_histogram()

p + aes(x = bti_cr) + geom_bar()

p + aes(x = gle_cgdpc, y = wef_lifexp) + geom_point() +
  scale_x_log10() + scale_y_log10()

p + aes(1, fh_polity2) + geom_boxplot()
p + aes(factor(bti_cr), fh_polity2) + geom_boxplot()

p + aes(x = gle_cgdpc, y = wef_lifexp) + geom_point(aes(color = fh_polity2))
p + aes(x = fh_polity2, y = wef_lifexp) + geom_point(aes(size = gle_cgdpc))

p + aes(x = fh_polity2, y = wef_lifexp) + geom_point() + theme_bw()
```

Ways to display a variable

In a scatterplot, `geom_point()` allows us to display a variable as:

- X/Y Axis variable (via `aes(x=, y=)`)
- Colour (via `aes(color=)`)
- Alpha (via `aes(alpha=)`)
- Size (via `aes(size=)`)
- Shape (via `aes(shape=)`)
- Facets (via `facet_wrap()`)
- Animation (e.g., <http://www.gapminder.org/world>)

Tools Beyond R

- GUI-based tools:
 - Inkscape
 - GIMP
- Command line tools:
 - ImageMagick

ggplot2 Resources

- <http://docs.ggplot2.org/current/>
- <https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>
- <https://github.com/jennybc/ggplot2-tutorial>
- <http://inundata.org/2013/04/10/a-quick-introduction-to-ggplot2/>
- <http://www.cookbook-r.com/Graphs/>

General Resources

- <http://www.edwardtufte.com/tufte/>
- <http://www.informationisbeautiful.net/>
- <http://flowingdata.com/>
- <http://ourworldindata.org/>
- <http://www.thefunctionalart.com/>
- <http://www.visualisingdata.com/>
- <http://www.braumoeller.info/dataviz/>

