# Tabulation and Visualization

Department of Government
London School of Economics and Political Science

1. Getting a grip on data

2. Tabulation

3. Visualization

# Preview: Analysis

Analysis is the “systematic and detailed examination of data.”

Two broad categories of analytic strategies:

1. Quantitative analysis
2. Qualitative analysis

# Preview: Quantitative Analysis

- *Quantitative analysis* involves calculation of statistic(s)
    - Statistic: “a quantitative summary of a variable for a set of units”

# Preview: Quantitative Analysis

- *Quantitative analysis* involves calculation of statistic(s)
  - Statistic: "a quantitative summary of a variable for a set of units"
- Examples
  - Total: Count, sum, proportion
  - Centrality: Mean, median, mode
  - Dispersion: Variance, standard deviation
  - Relationship: Correlation, etc.

# Preview: Qualitative Analysis

- *Qualitative analysis* involves typically narrative characterisations of phenomena

# Preview: Qualitative Analysis

- *Qualitative analysis* involves typically narrative characterisations of phenomena
- Examples
  - Typologies
  - Hierarchies
  - Accounts or interpretations

# Preview: Qualitative Analysis

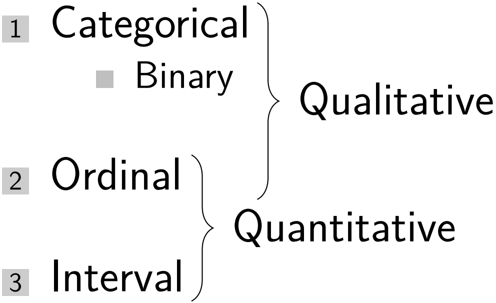
- *Qualitative analysis* involves typically narrative characterisations of phenomena
- Examples
  - Typologies
  - Hierarchies
  - Accounts or interpretations
- *Qualitative analysis* is more general and fluidic than quantitative

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”
- Three parts:
  - A set of units
  - A variable measured for those units
  - An estimator (i.e., aggregation procedure)

```
            country continent lifeExp      pop
            Austria    Europe    79    8199783
  Equatorial Guinea    Africa    51     551201
            Iceland    Europe    81     301931
               Iran      Asia    70   69453570
             Kuwait      Asia    77    2505559
            Lesotho    Africa    42    2012649
             Serbia    Europe    74   10150265
              Sudan    Africa    58   42292929
             Sweden    Europe    80    9031088
Trinidad and Tobago  Americas    69    1056608
```

# Central Tendency

# Central Tendency

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

# Mean/average

```
             country continent lifeExp      pop
             Austria    Europe      79  8199783
   Equatorial Guinea    Africa      51   551201
             Iceland    Europe      81   301931
                Iran      Asia      70 69453570
              Kuwait      Asia      77  2505559
             Lesotho    Africa      42  2012649
              Serbia    Europe      74 10150265
               Sudan    Africa      58 42292929
              Sweden    Europe      80  9031088
 Trinidad and Tobago  Americas      69  1056608
```

$Sum = 79 + 51 + 81 + 70 + 77 + 42 + 74 + 58 + 80 + 69 = 681$

$Mean = 681/10 = 68.1$

# Central Tendency

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

# Central Tendency

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

- Sort-based statistics:
  - Range
  - Minimum
  - Median (middle value)
  - Maximum
  - Percentiles

# Median, Min, Max, etc.

```
            country continent lifeExp      pop
            Austria    Europe     79   8199783
  Equatorial Guinea    Africa     51    551201
            Iceland    Europe     81    301931
               Iran      Asia     70  69453570
             Kuwait      Asia     77   2505559
            Lesotho    Africa     42   2012649
             Serbia    Europe     74  10150265
              Sudan    Africa     58  42292929
             Sweden    Europe     80   9031088
Trinidad and Tobago  Americas     69   1056608
```

# Median, Min, Max, etc.

```
             country continent lifeExp      pop
             Lesotho    Africa      42  2012649
   Equatorial Guinea    Africa      51   551201
               Sudan    Africa      58 42292929
 Trinidad and Tobago  Americas      69  1056608
                Iran      Asia      70 69453570
              Serbia    Europe      74 10150265
              Kuwait      Asia      77  2505559
             Austria    Europe      79  8199783
              Sweden    Europe      80  9031088
             Iceland    Europe      81   301931
```

# Central Tendency

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

- Sort-based statistics:
    - Range
    - Minimum
    - Median (middle value)
    - Maximum
    - Percentiles

# Central Tendency

- Mean (average): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sort-based statistics:
  - Range
  - Minimum
  - Median (middle value)
  - Maximum
  - Percentiles
- Mode: Most common value

# Mode

```
            country continent lifeExp      pop
            Austria    Europe     79   8199783
  Equatorial Guinea    Africa     51    551201
            Iceland    Europe     81    301931
               Iran      Asia     70  69453570
             Kuwait      Asia     77   2505559
            Lesotho    Africa     42   2012649
             Serbia    Europe     74  10150265
              Sudan    Africa     58  42292929
             Sweden    Europe     80   9031088
Trinidad and Tobago  Americas     69   1056608
```

# Mode

```
            country continent lifeExp      pop
  Equatorial Guinea    Africa     51    551201
            Lesotho    Africa     42   2012649
              Sudan    Africa     58  42292929
Trinidad and Tobago  Americas     69   1056608
               Iran      Asia     70  69453570
             Kuwait      Asia     77   2505559
            Austria    Europe     79   8199783
            Iceland    Europe     81    301931
             Serbia    Europe     74  10150265
             Sweden    Europe     80   9031088
```

# Central Tendency

- Mean (average): $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Sort-based statistics:
  - Range
  - Minimum
  - Median (middle value)
  - Maximum
  - Percentiles
- Mode: Most common value

# Dispersion/variation

- Variance:
  $Var(x) = s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

# Dispersion/variation

- Variance:
  $Var(x) = s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

- Standard Deviation:
  $sd(x) = s_x = \sqrt{Var(x)}$

```
              country continent lifeExp      pop
              Austria    Europe      79  8199783
   Equatorial Guinea    Africa      51   551201
              Iceland    Europe      81   301931
                 Iran      Asia      70 69453570
               Kuwait      Asia      77  2505559
              Lesotho    Africa      42  2012649
               Serbia    Europe      74 10150265
                Sudan    Africa      58 42292929
               Sweden    Europe      80  9031088
 Trinidad and Tobago  Americas      69  1056608
```

$$Mean = 68.1$$

$$Variance = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad = \frac{1620.9}{10-1} = 180.1$$

$$SD = \sqrt{Var(x)} \qquad = 13.42$$

# Shape

- Skewness

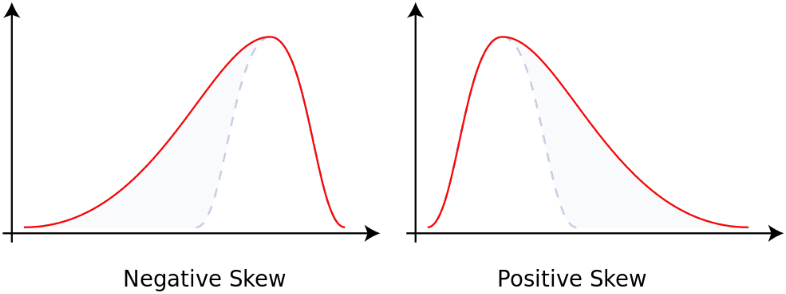
# Shape

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

# Shape

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

# Skewness



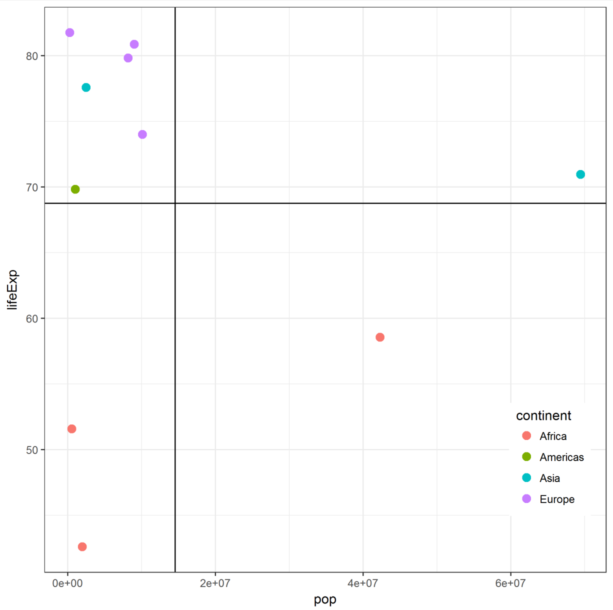
Source: Rodolfo Hermans (Wikimedia)

# Relationship

- Covariation:
  $Cov(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

# Relationship

- Covariation:
  $Cov(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- Correlation:
  $Corr(x, y) = r_{x,y} = \sum\limits_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$


80
70
60
50
lifeExp
0e+00
2e+07
4e+07
6e+07
pop
continent
Africa
Americas
Asia
Europe

# In R. . .

- `mean()`
- `median()`, `min()`, `max()`, `quantile()`
- `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

```
            country continent lifeExp      pop
            Austria    Europe      79  8199783
  Equatorial Guinea    Africa      51   551201
            Iceland    Europe      81   301931
               Iran      Asia      70 69453570
             Kuwait      Asia      77  2505559
            Lesotho    Africa      42  2012649
             Serbia    Europe      74 10150265
              Sudan    Africa      58 42292929
             Sweden    Europe      80  9031088
Trinidad and Tobago  Americas      69  1056608
```

# Tabulation (Counts/Totals)

| Continent | Count |
|---|---:|
| Africa | 3 |
| Americas | 1 |
| Asia | 2 |
| Europe | 4 |
| Total | 10 |

# Tabulation (Proportions)

| Continent | Count |
|---|---:|
| Africa | 0.3 (30%) |
| Americas | 0.1 (10%) |
| Asia | 0.2 (20%) |
| Europe | 0.4 (40%) |
| Total | 1.0 (100%) |

# Tabulation (Aggregations)

| Continent | Mean Population |
|---|---:|
| Africa | 14952260 |
| Americas | 1056608 |
| Asia | 35979565 |
| Europe | 6920767 |
| Grand Mean | 14555558 |

# In R. . .

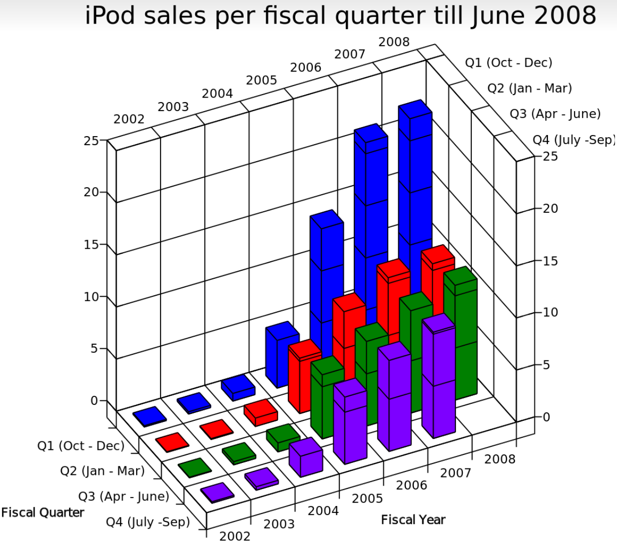
- `table()`
- `prop.table()`
- `aggregate()`
- `dplyr::summarize()`

1 Getting a grip on data

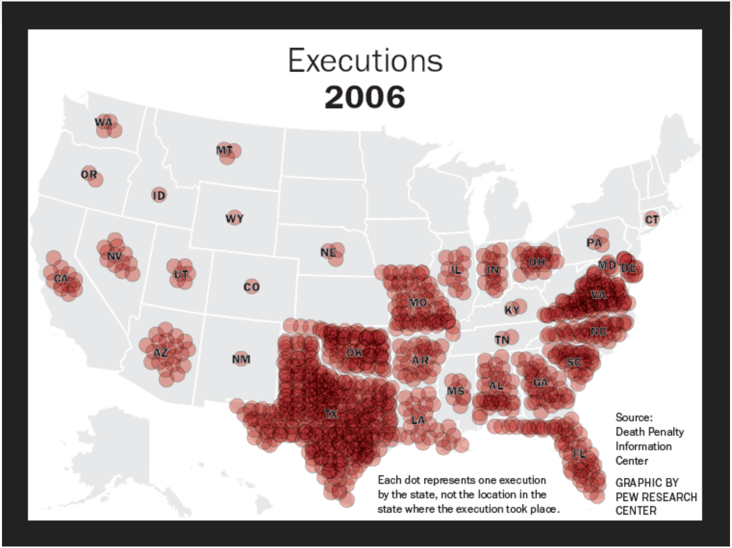
2 Tabulation

3 Visualization

# Bad visualizations. . .



Source: Wikimedia

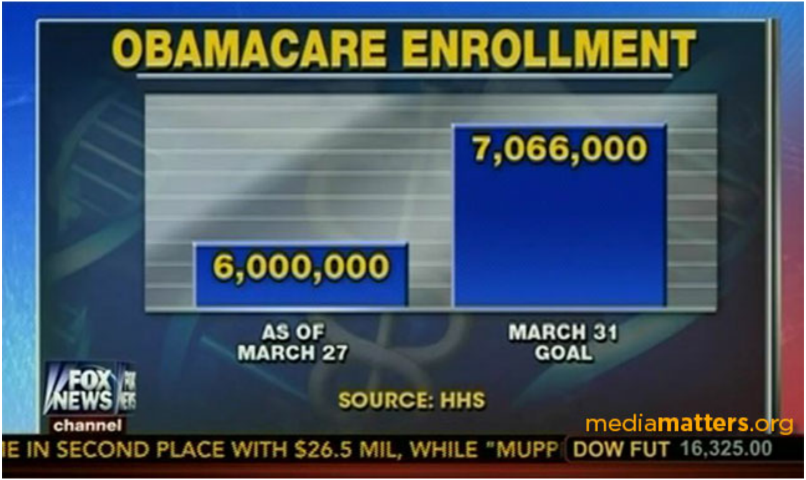


Source: Death Penalty Information Center

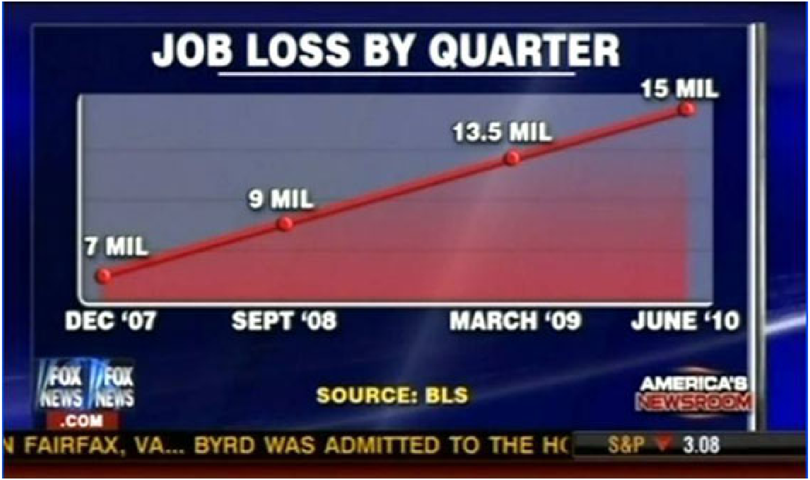
Each dot represents one execution by the state, not the location in the state where the execution took place.

GRAPHIC BY PEW RESEARCH CENTER

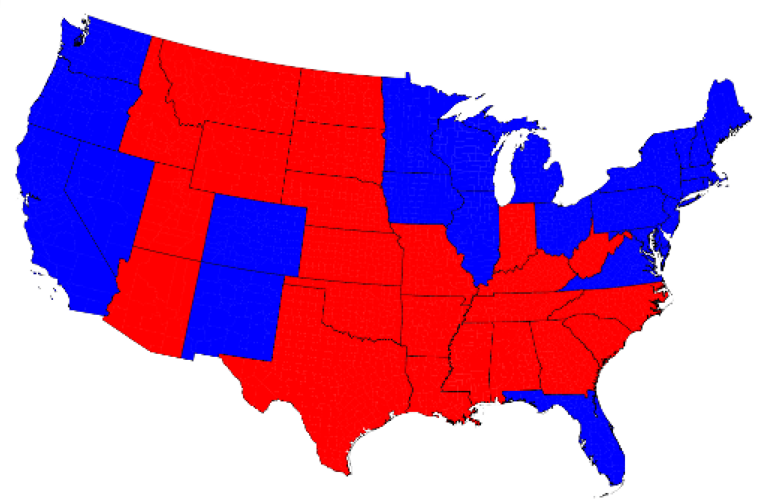
Source: JunkCharts



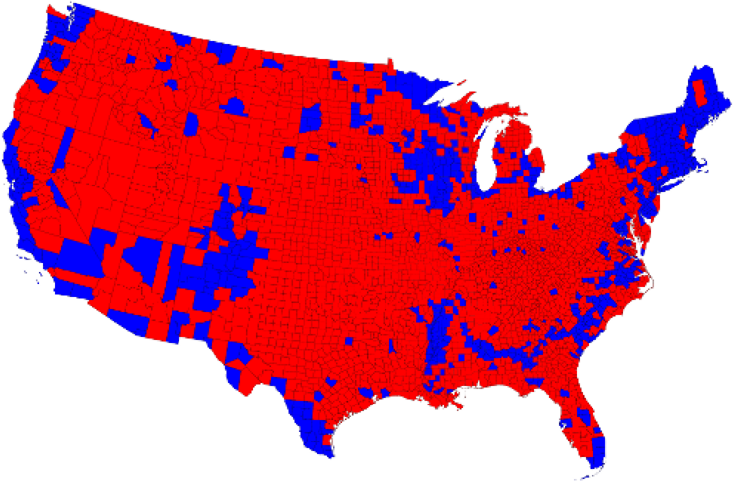
Source: MediaMatters



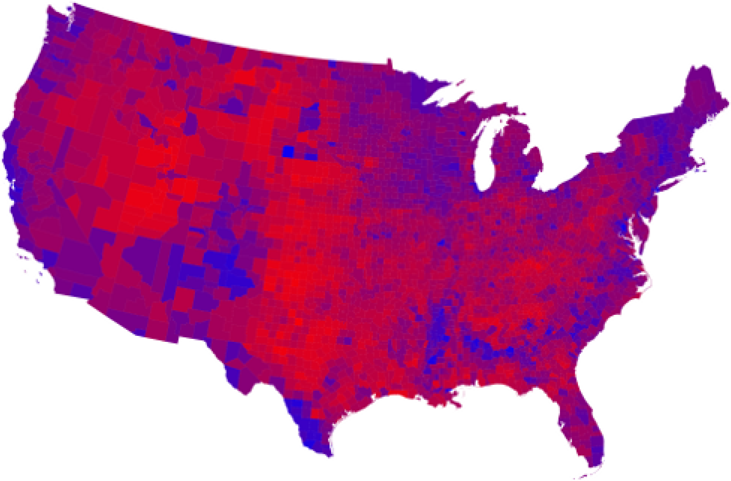
Source: MediaMatters



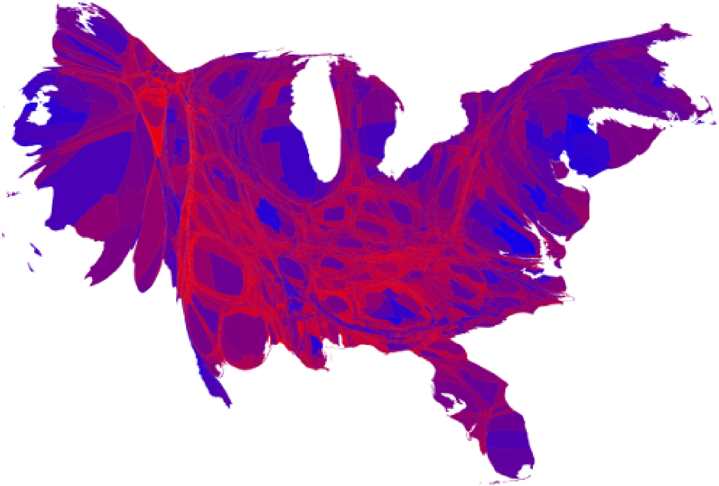
Source: (c) Mark Newman



Source: (c) Mark Newman



Source: (c) Mark Newman



Source: (c) Mark Newman

# 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

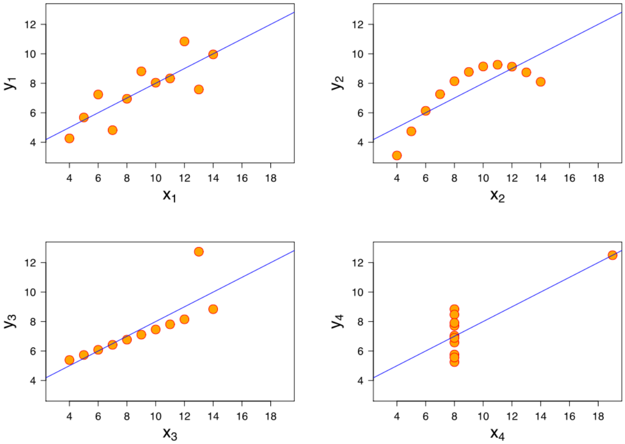
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 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 |

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

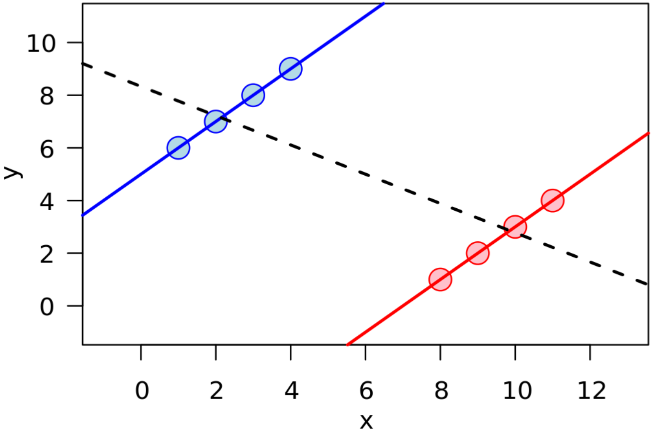
$$\bar{y} = 7.5,\ Var(y) = 4.12,$$

$$Corr(x, y) = 08.16$$

# Anscombe’s Quartet



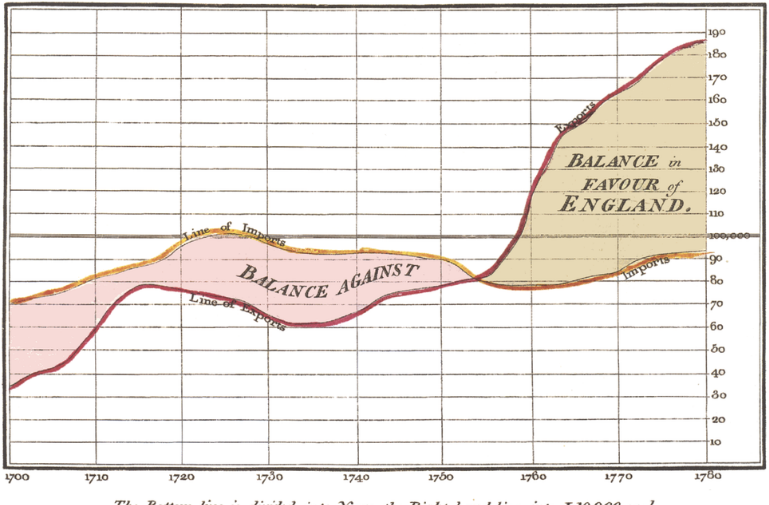
# Simpson’s Paradox



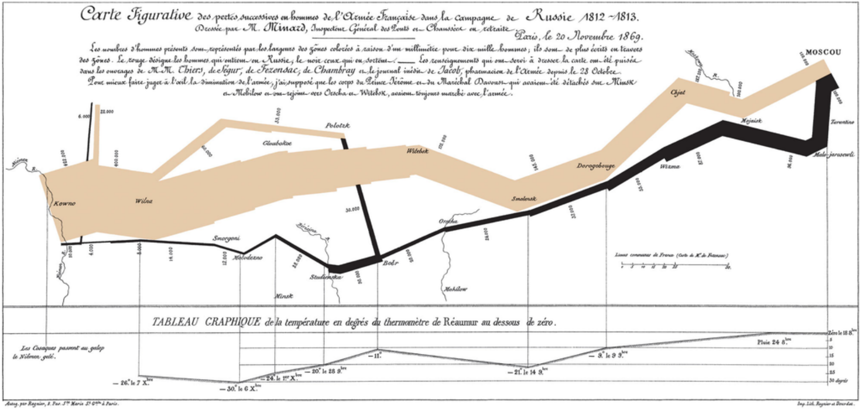
Source: Wikimedia

# William Playfair

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

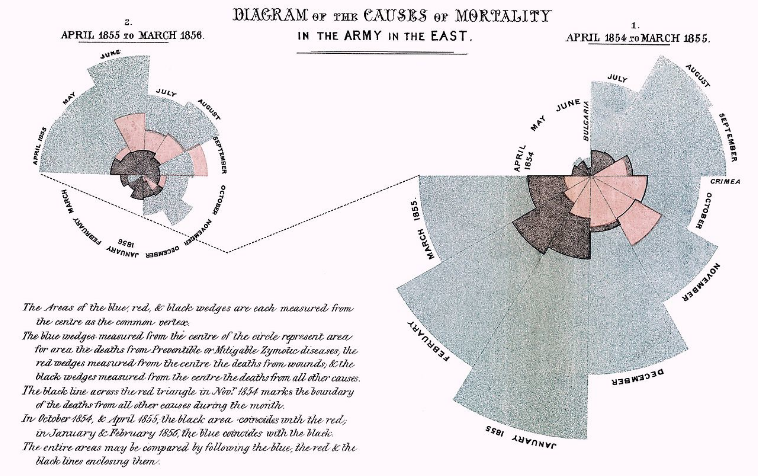


# Charles Minard


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 MM. 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.
MOSCOU
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wizma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Smorgoni
Molodezno
Studienska
Minsk
Bobr
Orscha
Mohilow
Niemen
Berezina
6.000
22.000
422.000
400.000
60.000
33.000
30.000
175.000
145.000
127.000
100.000
55.000
37.000
24.000
20.000
50.000
28.000
12.000
14.000
8.000
4.000
10.000
96.000
87.000
Lieues communes de France
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8.bre
Pluie 24 8.bre
— 9° le 9 9.bre
— 11°
— 21° le 14 9.bre
— 20° le 28 9.bre
— 24° le 1er X.bre
— 30° le 6 X.bre
— 26° le 7 X.bre
5
10
15
20
25
30 degrés

Source: Wikimedia

# Florence Nightingale



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

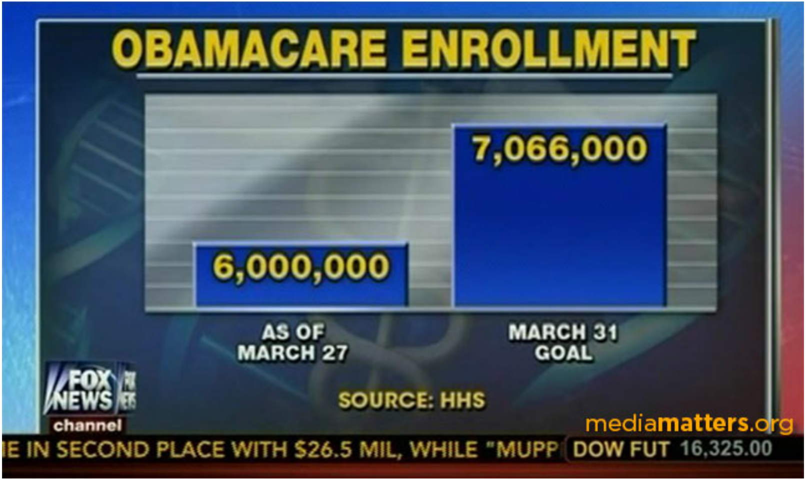
In October 1854, & April 1855; the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

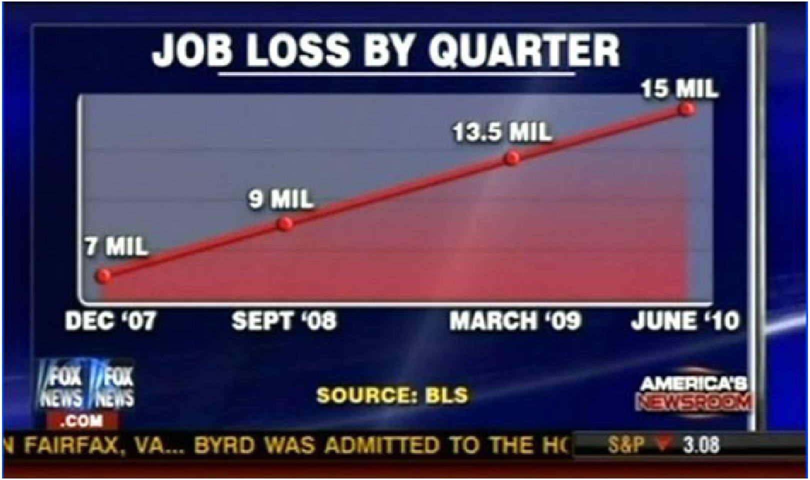
Source: Wikimedia

# Some Basic Principles

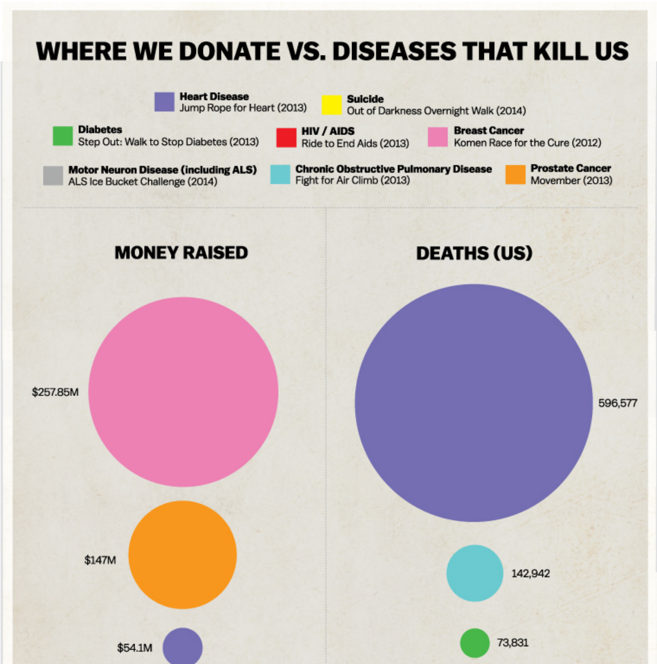
1. Be honest

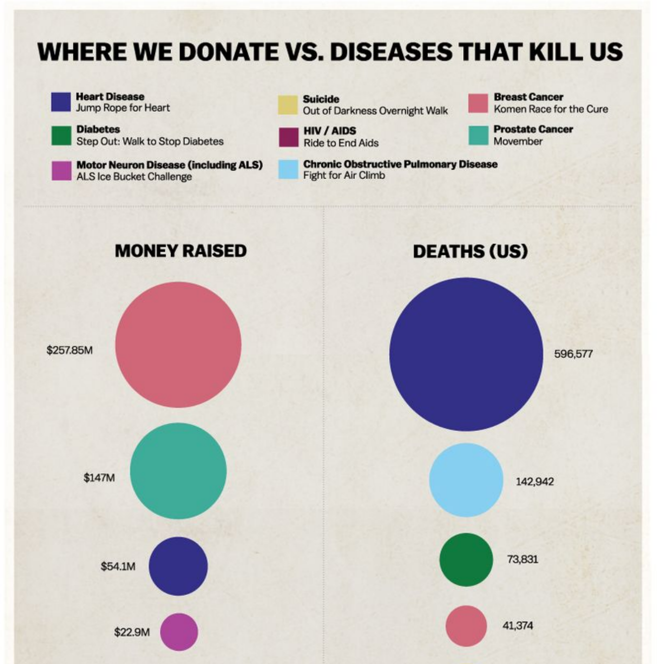


Source: MediaMatters



Source: MediaMatters


WHERE WE DONATE VS. DISEASES THAT KILL US
Heart Disease
Jump Rope for Heart (2013)
Suicide
Out of Darkness Overnight Walk (2014)
Diabetes
Step Out: Walk to Stop Diabetes (2013)
HIV / AIDS
Ride to End Aids (2013)
Breast Cancer
Komen Race for the Cure (2012)
Motor Neuron Disease (including ALS)
ALS Ice Bucket Challenge (2014)
Chronic Obstructive Pulmonary Disease
Fight for Air Climb (2013)
Prostate Cancer
Movember (2013)
MONEY RAISED
$257.85M
$147M
$54.1M
DEATHS (US)
596,577
142,942
73,831

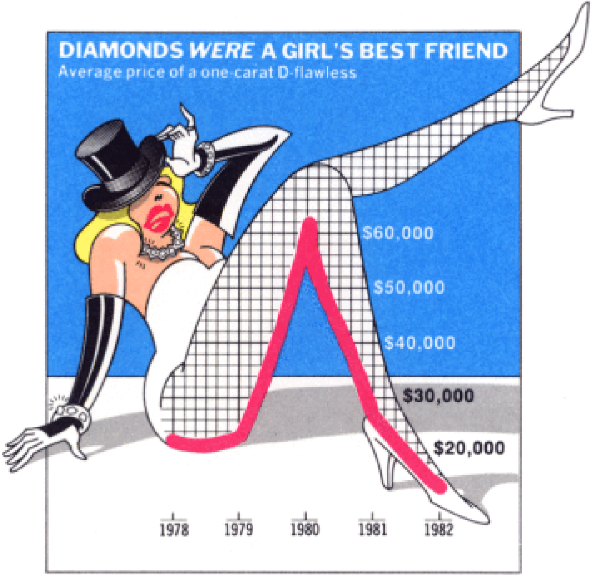

WHERE WE DONATE VS. DISEASES THAT KILL US
Heart Disease
Jump Rope for Heart
Suicide
Out of Darkness Overnight Walk
Breast Cancer
Komen Race for the Cure
Diabetes
Step Out: Walk to Stop Diabetes
HIV / AIDS
Ride to End Aids
Prostate Cancer
Movember
Motor Neuron Disease (including ALS)
ALS Ice Bucket Challenge
Chronic Obstructive Pulmonary Disease
Fight for Air Climb
MONEY RAISED
$257.85M
$147M
$54.1M
$22.9M
DEATHS (US)
596,577
142,942
73,831
41,374

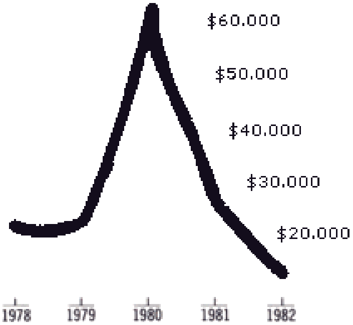
# Some Basic Principles

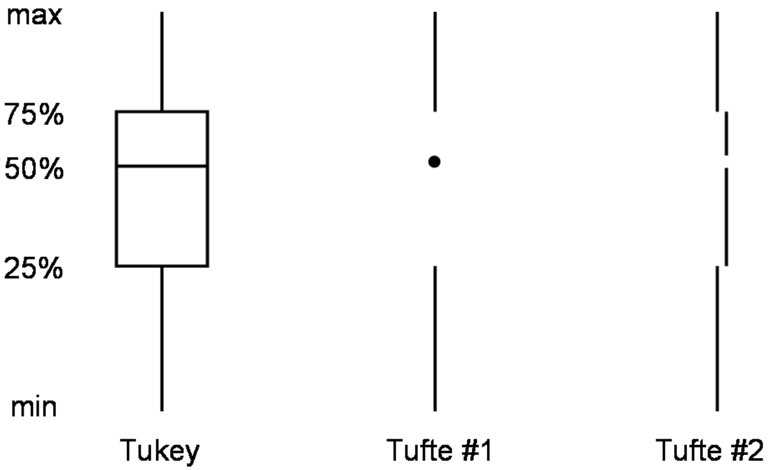
1. Be honest

# Some Basic Principles

1. Be honest
2. Data-Ink Ratio


DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982


$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982



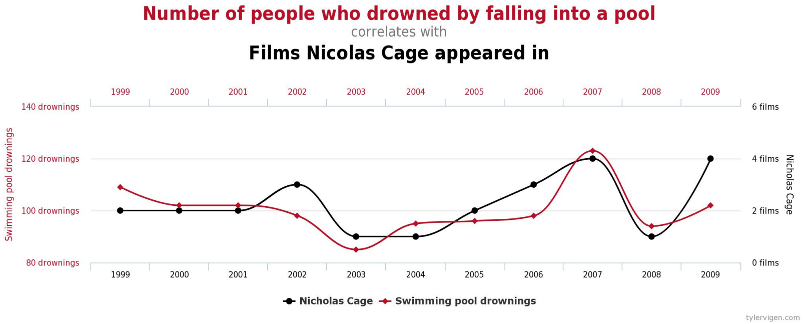
Source: StackOverflow

# Some Basic Principles

1. Be honest
2. Data-Ink Ratio

# Some Basic Principles

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



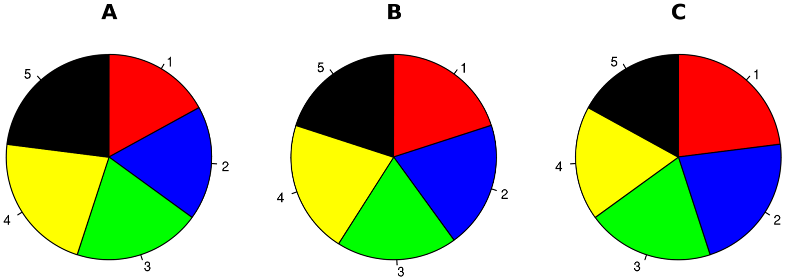
Source: CC-BY Tyler Vigen

# Some Basic Principles

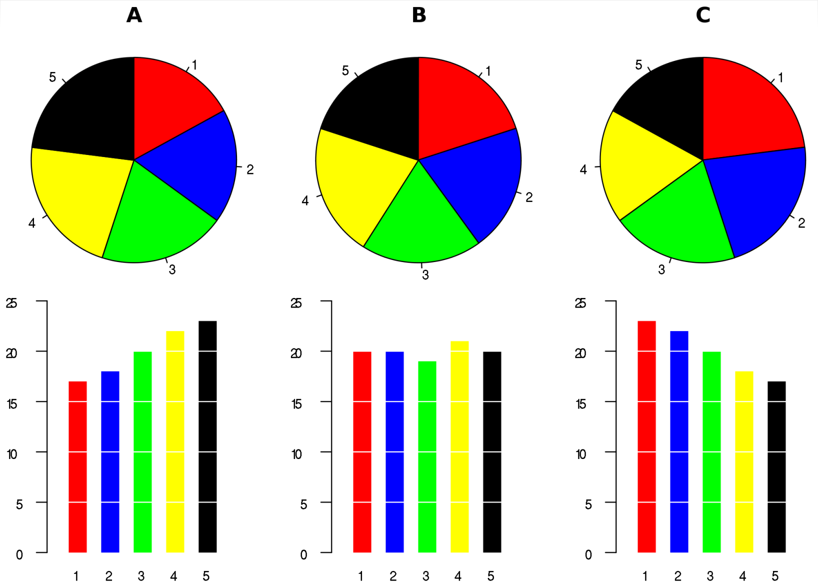
1. Be honest
2. Data-Ink Ratio
3. Tell a story

# Some Basic Principles

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



Source: Wikimedia



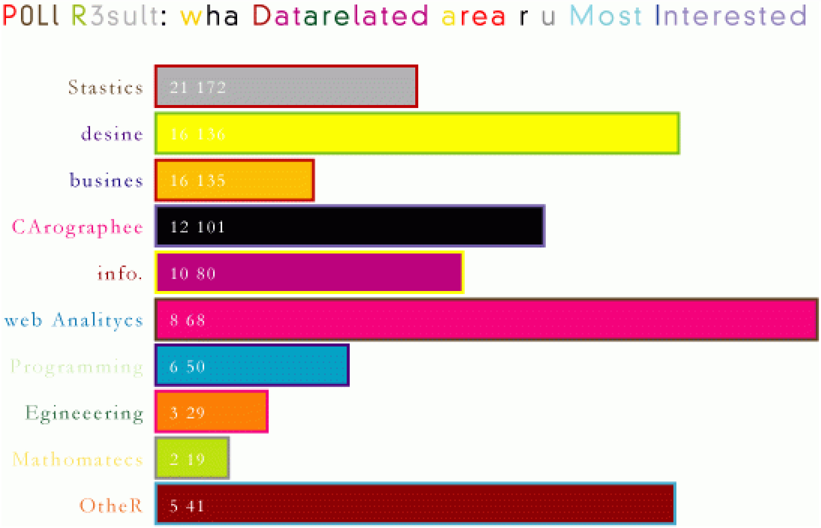
Source: Wikimedia

# Some Basic Principles

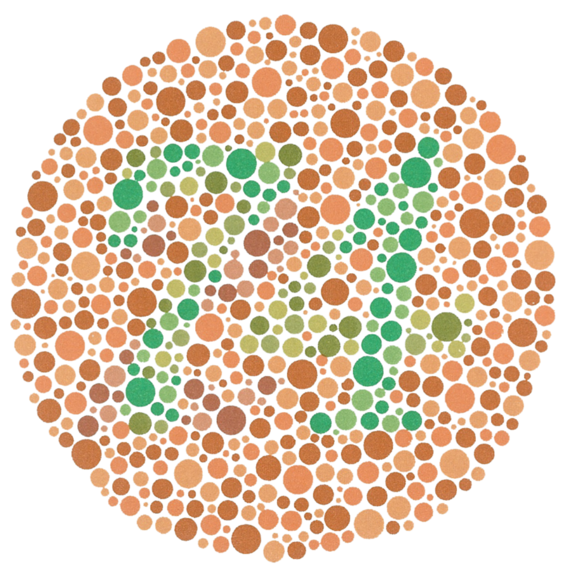
1. Be honest
2. Data-Ink Ratio
3. Tell a story
4. Steer reader's attention

# Some Basic Principles

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



Source: Flowing Data



Source: Wikimedia

# 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?

# Homework

1. Find a visualization anywhere on the internet.
2. Post a link to the visualization to the Moodle forum.
3. Include the visualization as an image.
4. Describe:
   - What is being visualized
   - Strengths of the visualization
   - Weaknesses of the visualization

# In R. . .

R has 5+ graphics "systems"

- Base graphics
- The **ggplot2** package
- The **lattice** 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

# 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`)

```
library("rio")
d <- import("http://www.qogdata.pol.gu.se/data/qog_std_cs_jan17.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()
```

# 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/