Quantitative methods

Week #10-11

Gergely Daróczi

Corvinus University of Budapest, Hungary

12 April 2013

Outline

There are several different averages (measures of central tendency) - with all different advantages and disadvantages:

1. **arithmetic mean**: $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+\cdots+x_n}{n}$
2. **geometric mean**: $\sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 x_2 \cdots x_n}$
3. **mode**: the most frequently occurring number/category in the sample
4. **median**: the middle number of the ranked variable
5. **midrange**: $\frac{\max x+\min x}{2}$

Which of the above would you choose to describe . . .

1. your grades in this semester,
2. the average number of students in the library,
3. the central tendency of hair color at the university,
4. the salary of people living in Budapest,
5. loss of money in a pub at Saturday night.

Judge the following statements:

1. “The average weekly earnings went up 107 percent between 1940 and 1948 in the United States Steel Corporation.”
2. “The average salary in the same corporation was $ 5.000 in 1942.”
3. “The probability of dying in a car accident is twice as much than being hit by an airplane.”
4. “Peter’s IQ is 98 and Linda’s is 101. A nice evidence of girls beeing smarter than boys.”
5. “This year I sleep twice as much than I used to last year. Should I feel happy?”

What average would you choose to describe the following variable asked in the European Values Study (Hungary, 2008):

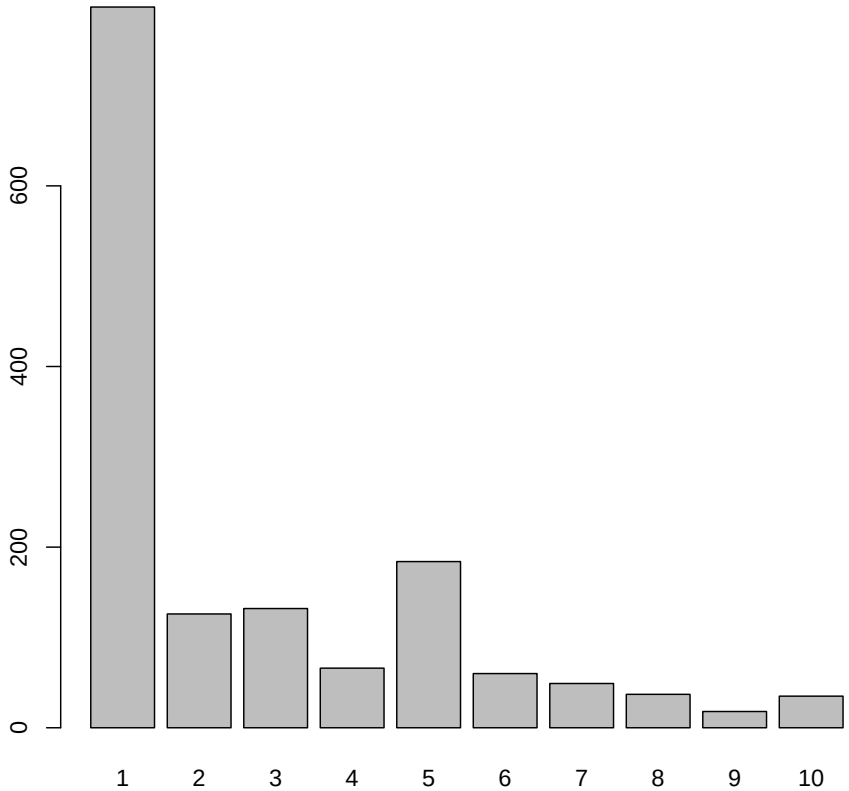
"Please tell me for each of the following statements whether you think it can always be justified (10), never (1) be justified, or something in between!"

1. Claiming state benefits which you are not entitled to
2. Abortion
3. Divorce
4. Avoiding a fare on public transport
5. Homosexuality

Descriptive statistics

Case studies

“Please tell me whether you think **Avoiding a fare on public transport** can always be justified (10), never (1) be justified, or something in between!”



Mean: 2.751

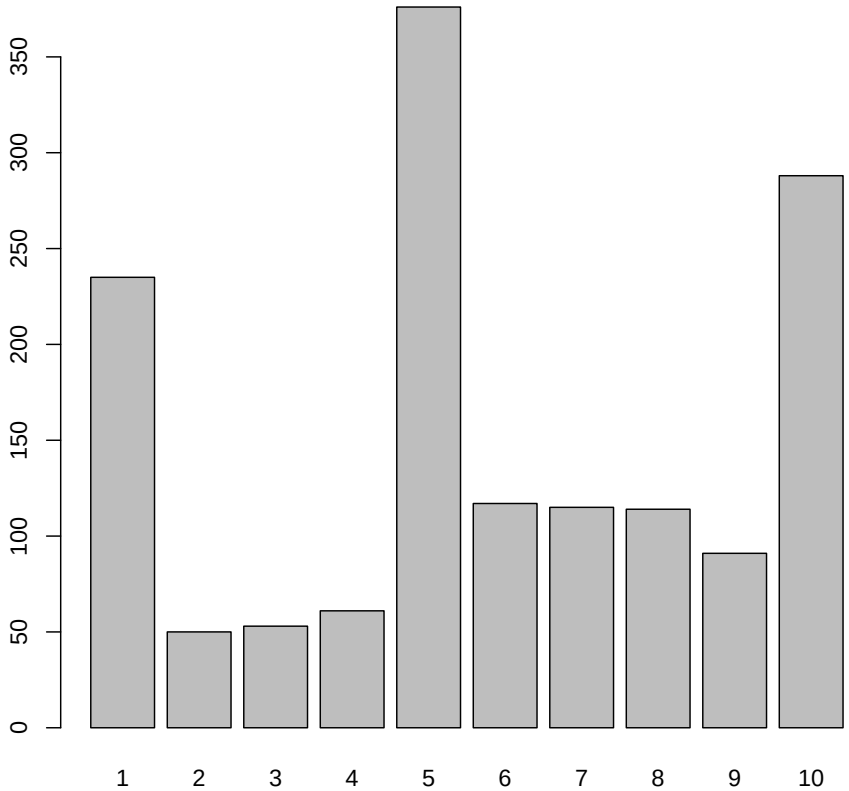
Mode: 1

Median: 1

Descriptive statistics

Case studies

"Please tell me whether you think **divorce** can always be justified (10), never (1) be justified, or something in between!"



Mean: 5.824

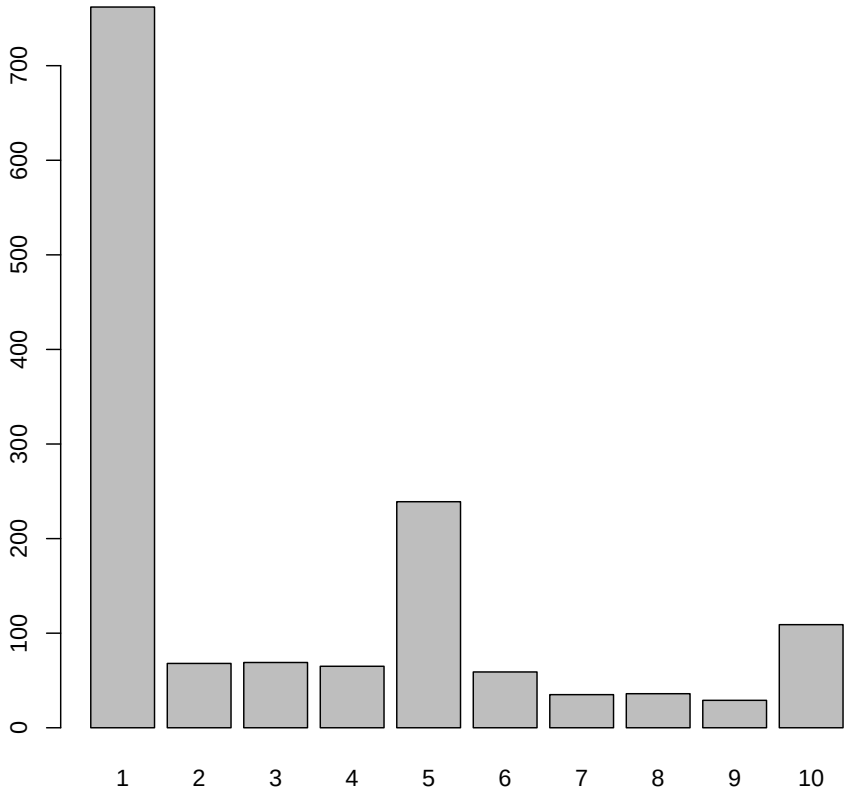
Mode: 5

Median: 5

Descriptive statistics

Case studies

"Please tell me whether you think **homosexuality** can always be justified (10), never (1) be justified, or something in between!"



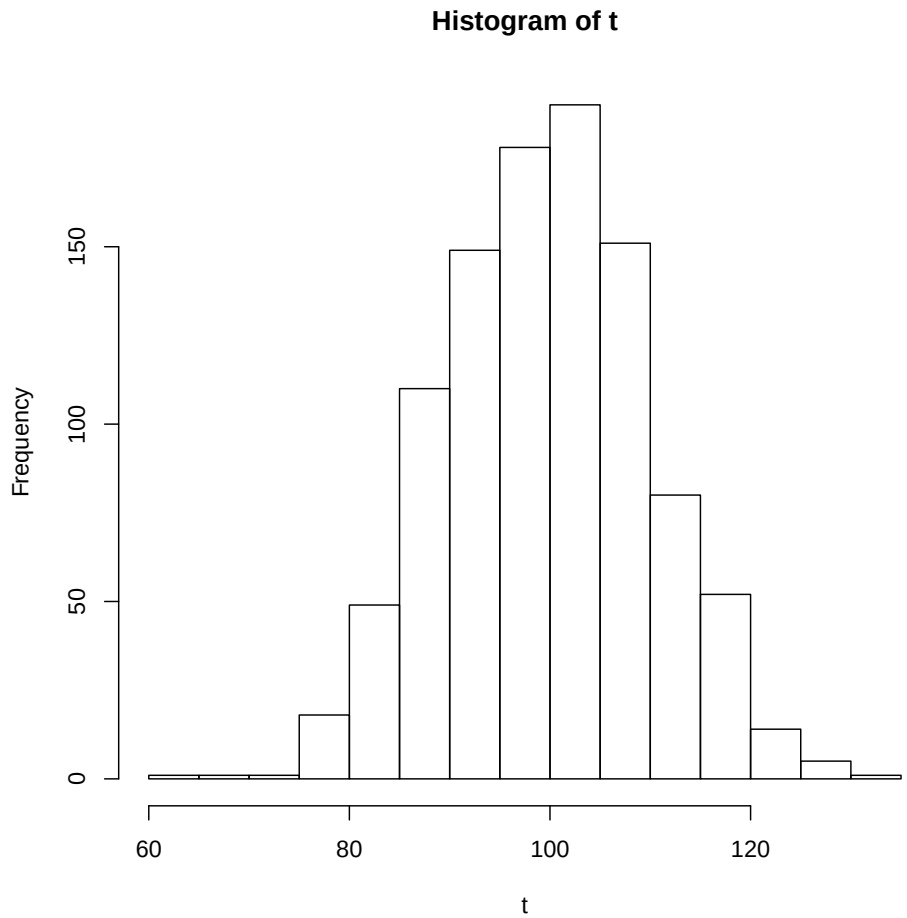
Mean: 3.261

Mode: 1

Median: 1

Research on intelligence (quotient) among students:

Research on intelligence (quotient) among students:

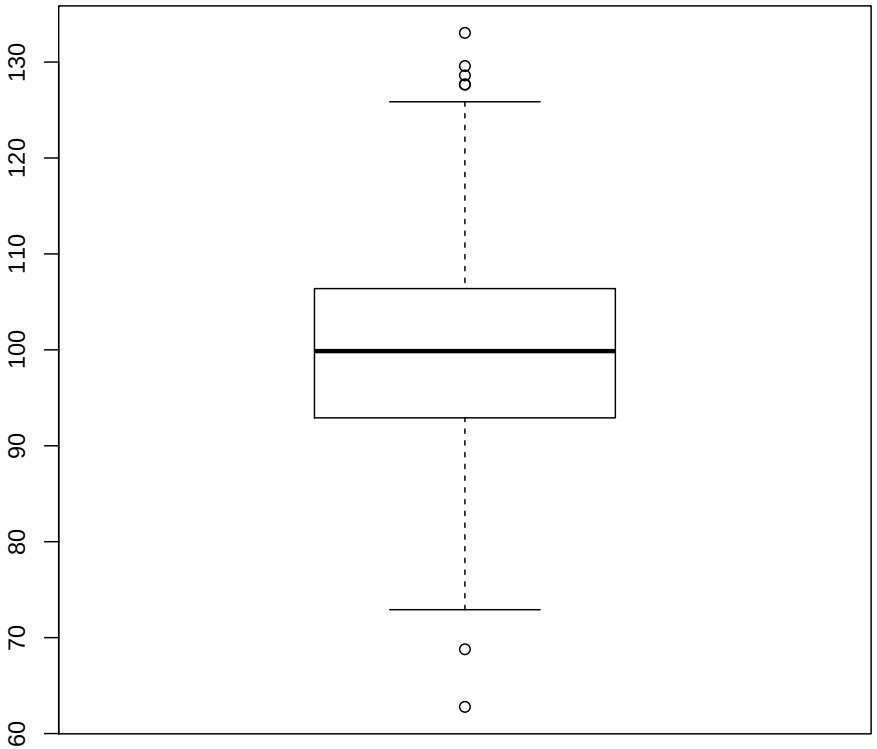


Mean: 99.6

Mode: 89.2

Median: 99.8

Research on intelligence (quotient) among students:



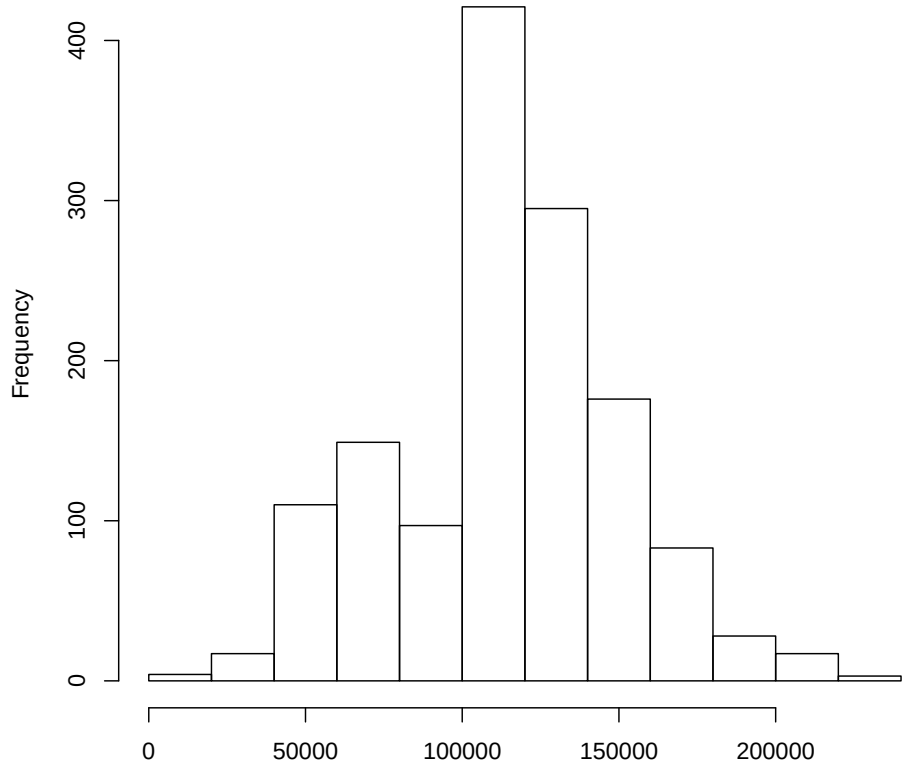
Mean: 99.6

Mode: 89.2

Median: 99.8

Research on salary of Hungarian people:

Research on salary of Hungarian people:

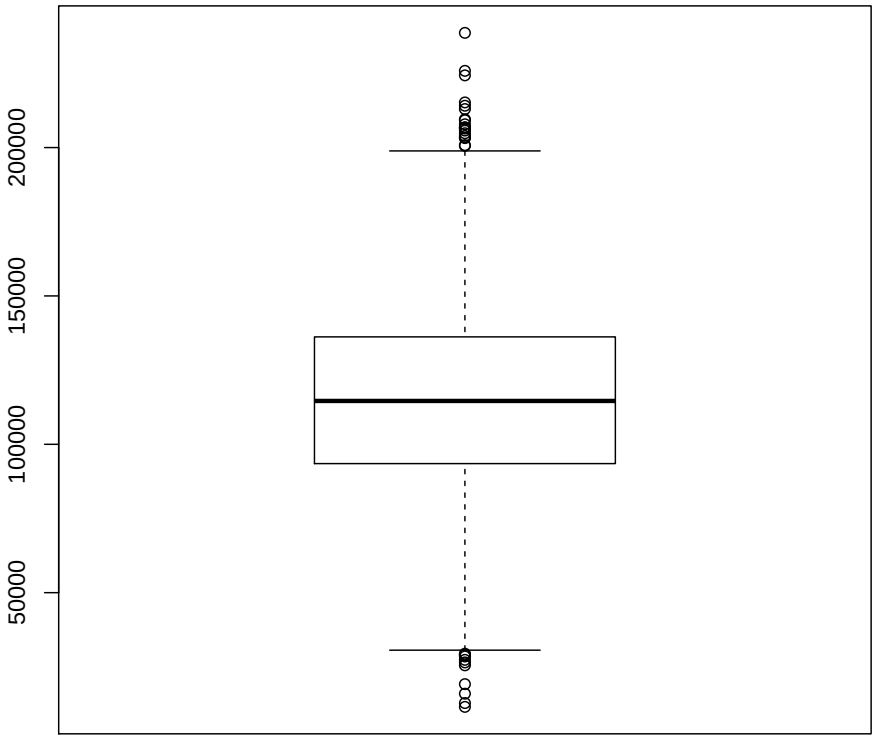


Mean: 113721

Mode: 72554

Median: 114613

Research on salary of Hungarian people:



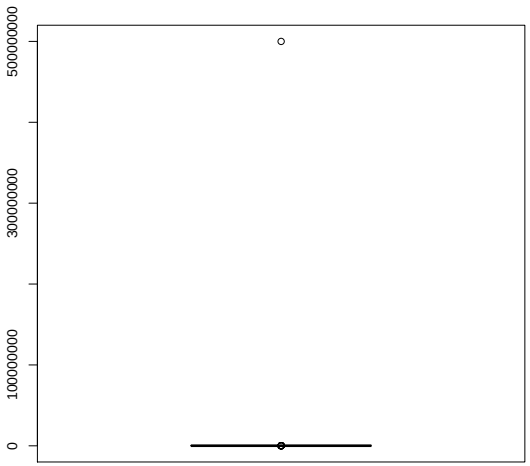
Mean: 113721

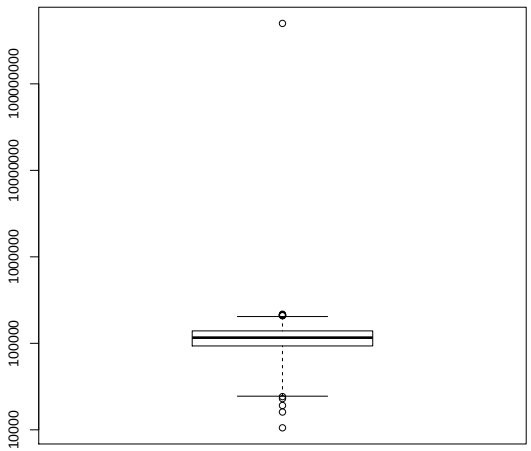
Mode: 72554

Median: 114613

What happens when we have a really rich person in the sample?

What happens when we have a really rich person in the sample?





Mean: 471150

Mode: 72554

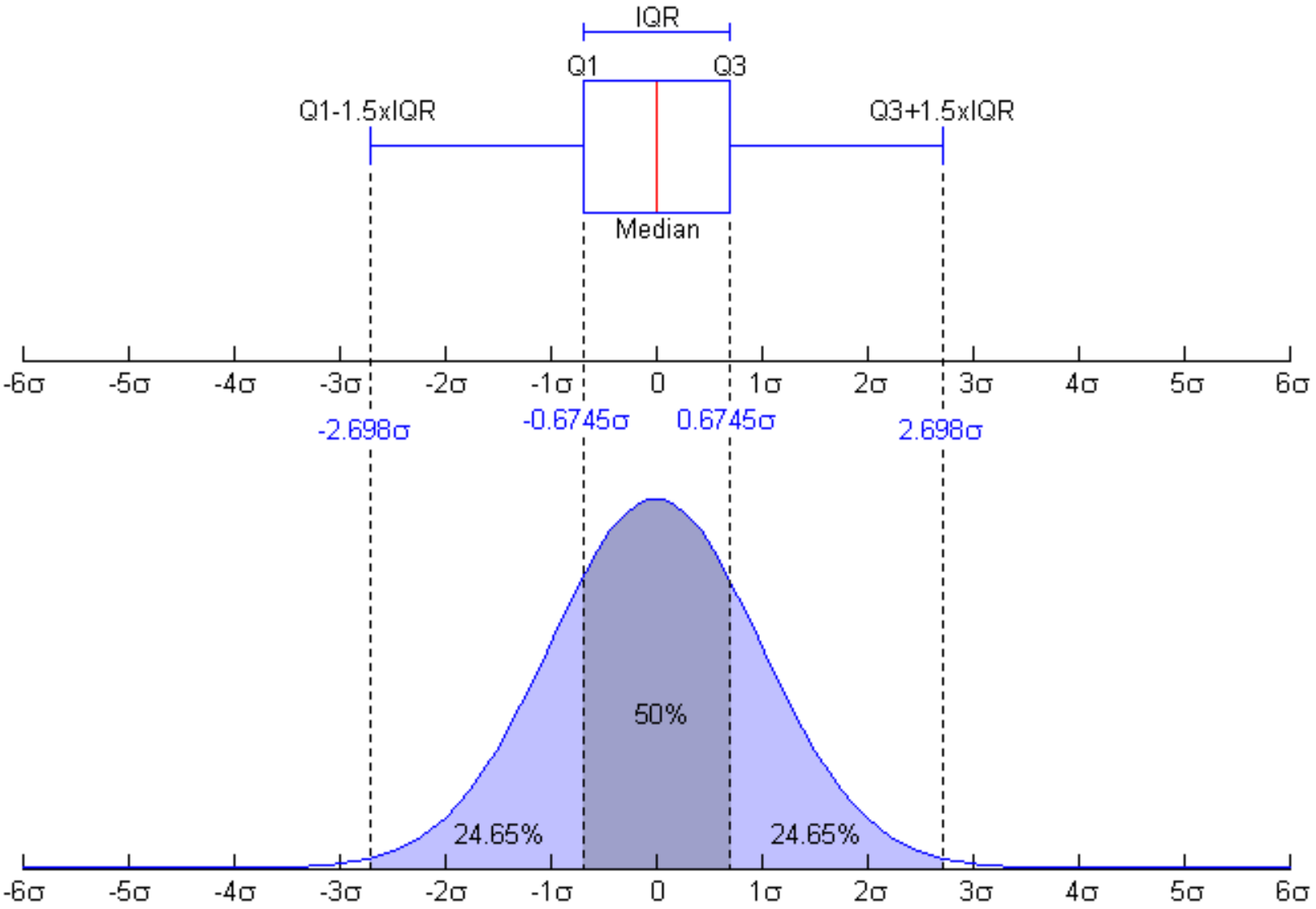
Median: 116299

There are several different statistical measures of variability or variation - with all different advantages and disadvantages:

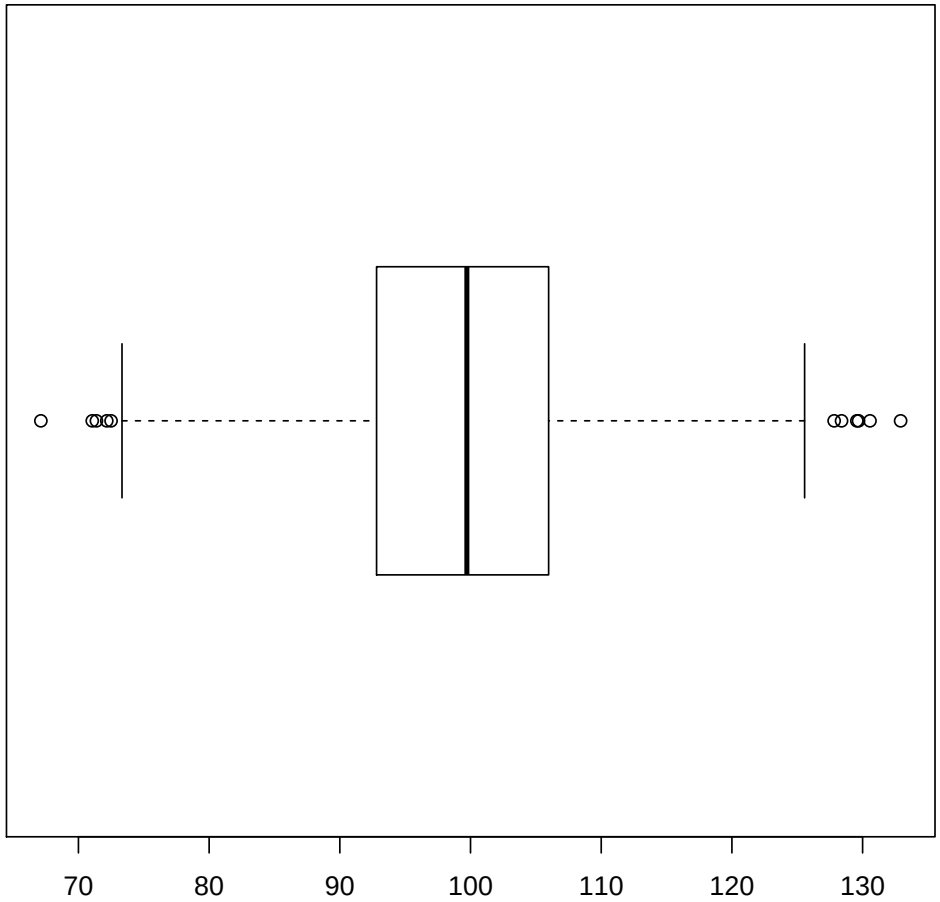
1. **range**: $\max x - \min x$
2. **standard deviation**: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{n-1}}$
3. **variance**: σ^2
4. **interquartile range (IQR)**: the difference between the third and first quartiles

Descriptive statistics

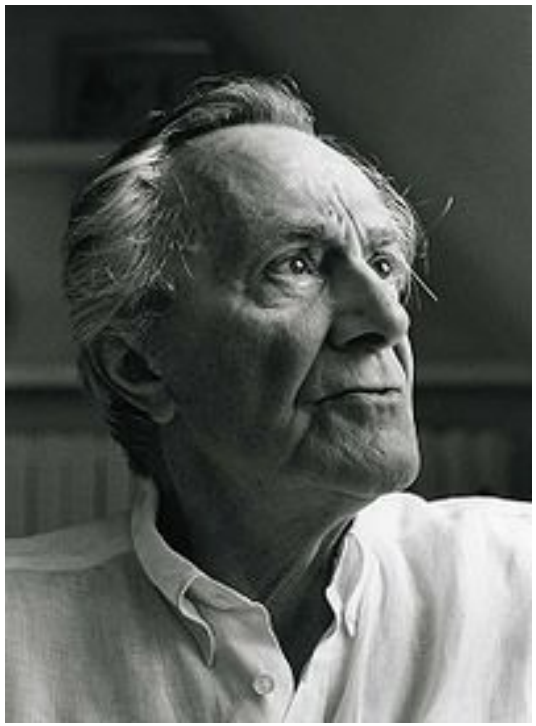
Interquartile range



Research on intelligence (quotient) among students:



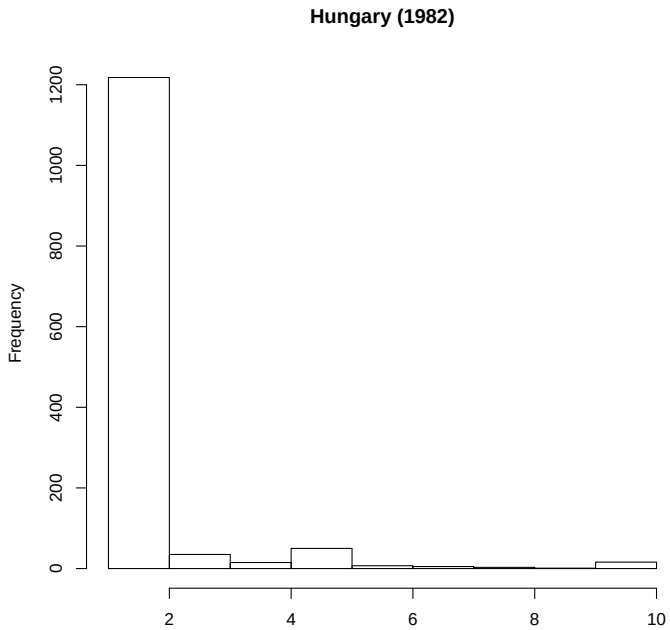
Lyotard : The Postmodern Condition. A Report on Knowledge (1979)



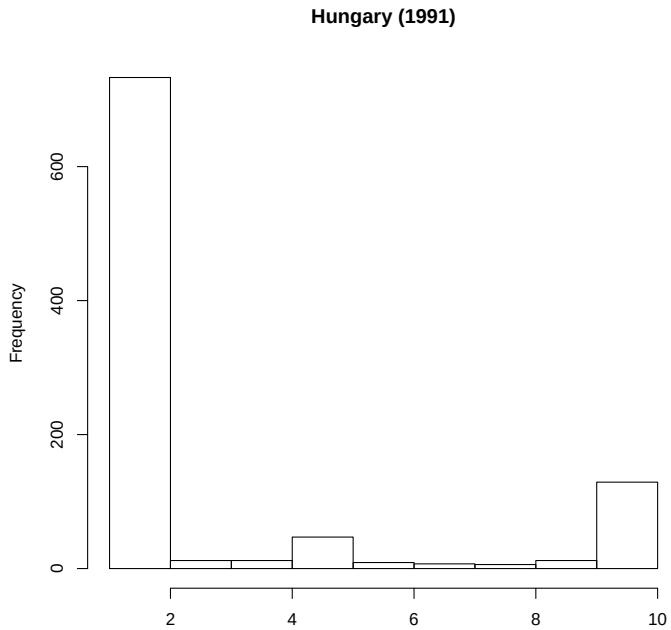
- “end of ’grand narratives’ or metanarratives”
- “anything goes”
- “postmodern and postmodern culture”

What about norms?

"Please tell me whether you think **homosexuality** can always be justified (10), never (1) be justified, or something in between!" – Hungary (1982-1991)

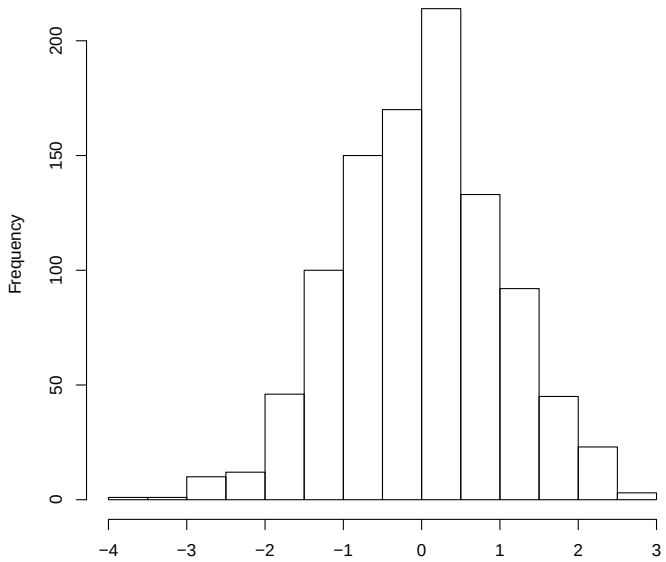


$\overline{x} = 1.447407; \sigma = 1.419384$

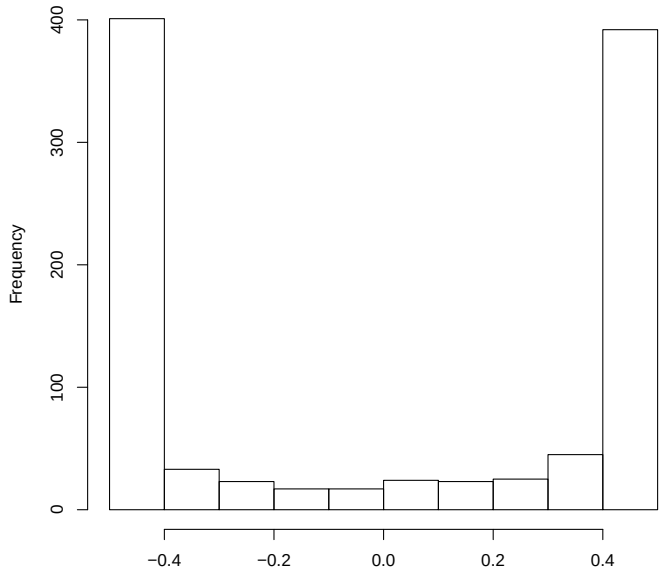


$\overline{x} = 2.713547; \sigma = 3.230236$

Check the mean and standard deviation of the following variables!

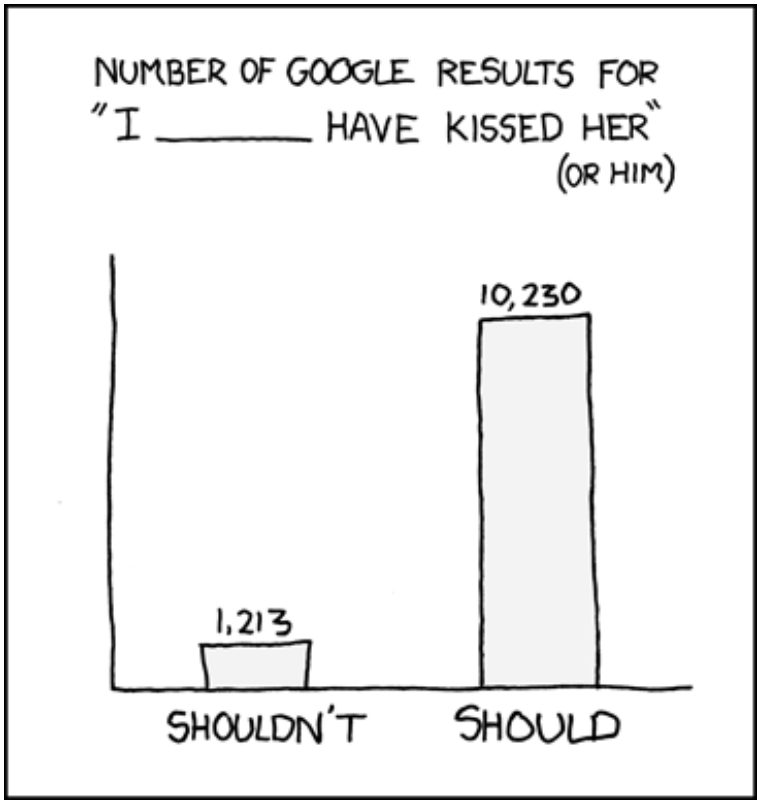


$\overline{x} = -0.1; \sigma = 1.019$



$\overline{x} = 0; \sigma = 0.453$

A new index of measurements: **sum**



What is the problem with this desciptive in this study?

Standardization and decomposition

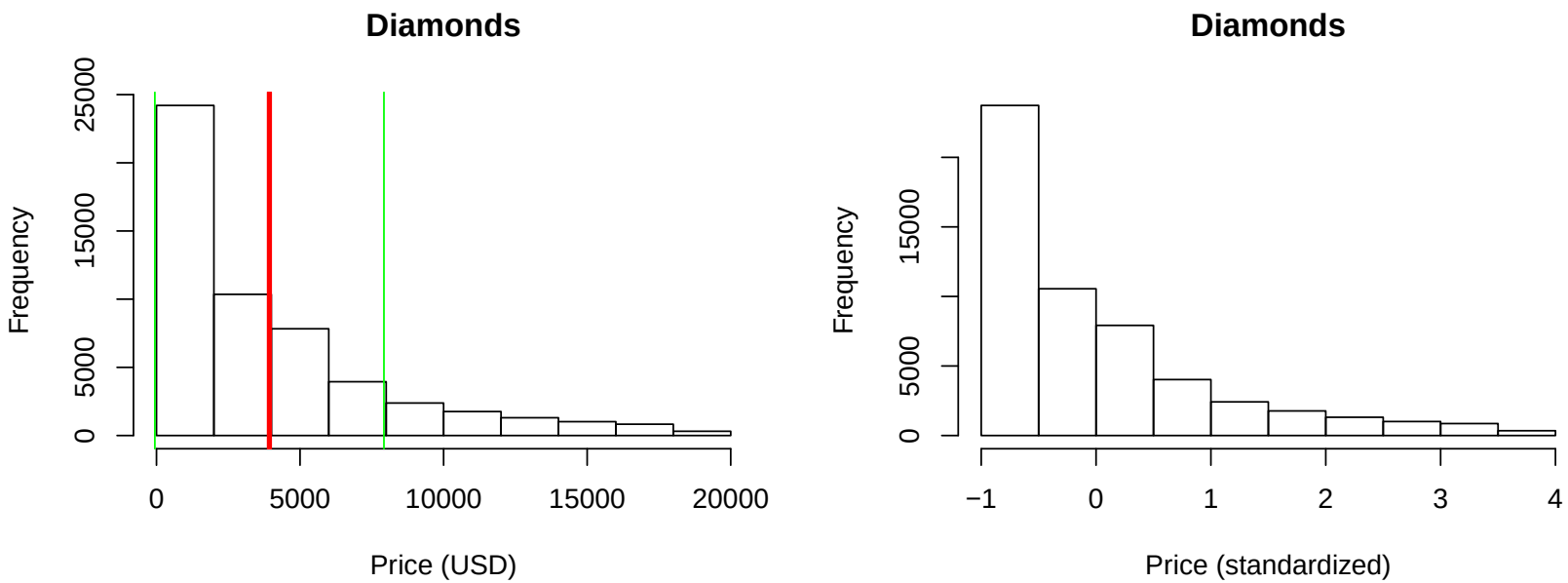
A basic example

Standardization and decomposition

A basic example

Standard score (z-values, z-scores, normal scores, standardized variables) indicates how many standard deviations an observation is above or below the mean:

$$z = \frac{x - \mu}{\sigma} = \frac{x_i - \bar{x}}{S_x^*}$$



Standardization and decomposition

Exercise

Standard score (z-values, z-scores, normal scores, standardized variables) indicates how many standard deviations an observation is above or below the mean:

$$z = \frac{x - \mu}{\sigma} = \frac{x_i - \bar{x}}{S_x^*}$$

Wages (th. forints): 100, 85, 55, 120, 65

Compute the z-score for the above variable!

Population and Deaths by Age in 1970 for White Females in Miami, Alaska, and the U.S.

	Miami			Alaska			U.S.		
Age	Pop.	Deaths	Rate*	Pop.	Deaths	Rate*	Pop.+	Deaths+	Rate*
< 15	114,350	136	1.19	37,164	59	1.59	23,961	32	1.34
15-24	80,259	57	0.71	20,036	18	0.90	15,420	9	0.58
25-44	133,440	208	1.56	32,693	37	1.13	21,353	30	1.40
45-64	142,670	1,016	7.12	14,947	90	6.02	19,609	140	7.14
65+	92,168	3,605	39.11	2,077	81	39.00	10,685	529	49.51
	562,887	5,022		106,917	285		91,028	740	
Crude death rate*			8.92			2.67			8.13

* Deaths per 1,000 population + in thousands

Definition

In direct standardization the stratum-specific rates of study populations are applied to the age distribution of a standard population.

$$Directly\ standardized\ rate = \frac{\sum stratum\ specific\ rates \times standard\ weights}{\sum standard\ weights}$$

$$Miami = \frac{(1.19x23,961) + \cdots + (39.11x10,685)}{91,208} = 6.92\ deaths/thousand$$

$$Alaska = \frac{(1.59x23,961) + \cdots + (39x10,685)}{91,208} = 6.71\ deaths/thousand$$

Definition

In indirect standardization, the standard population provides the rates and the study population provides the weights.

$$\textit{Indirectly standardized rate} = \frac{\sum \textit{observed values}}{\sum \textit{expected values}}$$

Expected values = Stratum specific rates from the standard population $\times$ stratum sizes from the study population

$$\textit{Miami} = \frac{5,022}{(1.34x114,350) + \cdots + (49.51x91,168)} 8.13 = 6.84 \textit{ deaths/th.}$$

$$\textit{Alaska} = \frac{285}{(1.34x37,164) + \cdots + (49.51x2,077)} 8.13 = 7.32 \textit{ deaths/th.}$$

Crude and Age-Standardized* 1970 Death Rates Per 1000 for White Females in Alaska, Miami, and the U.S.

	Alaska	Miami	U.S.
Crude	2.67	8.92	8.13
Direct	6.71	6.92	-
Indirect	7.23	6.84	-

*Standard population is 1970 U.S. white females

	Study population	**Standard population**
Directly-standardized rate	Rates	Weights
Indirectly-standardized rate	Weights	Rates

Standardization and decomposition

Exercise

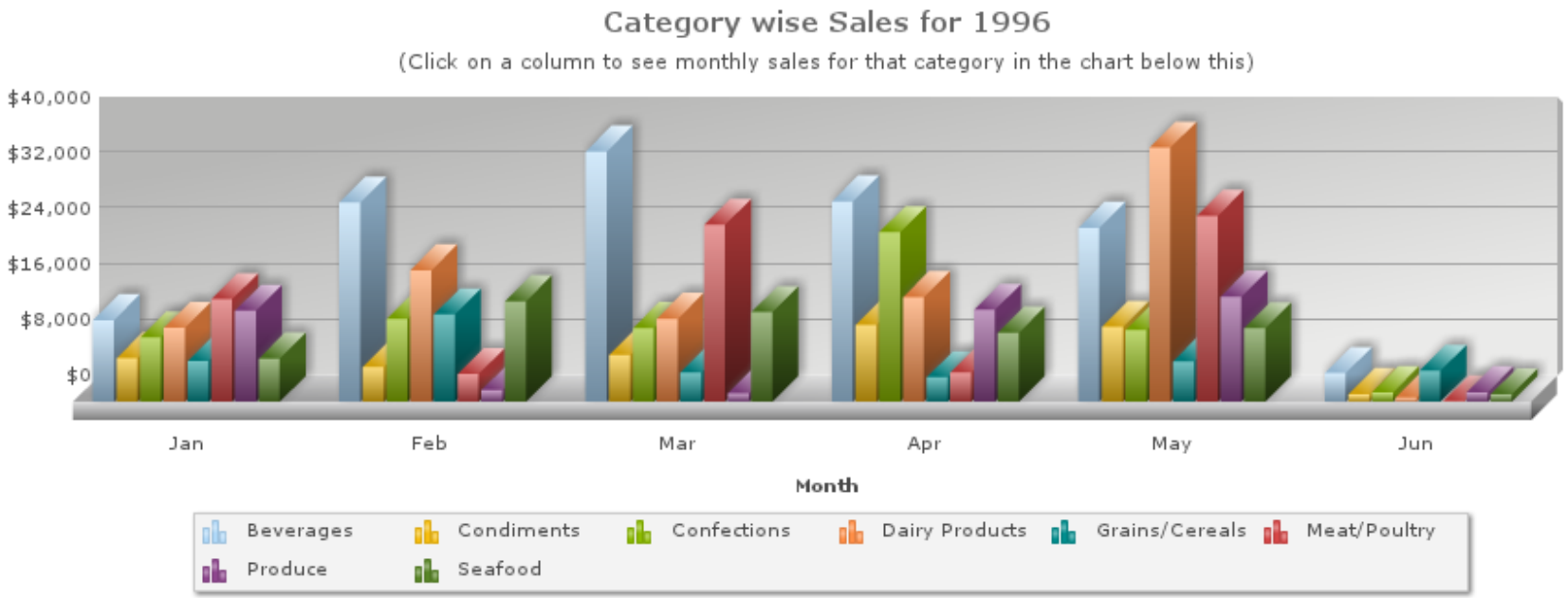
Death rates by age in two occupations and a standard population

	Occupation A			Occupation B			Standard population		
Age	Persons	Deaths	Rate	Persons	Deaths	Rate	Persons	Deaths	Rate
40-49	1,000	2	0.002	5,000	10	0.002	30,000	30	0.001
50-59	5,000	20	0.004	1,000	4	0.004	40,000	120	0.003
Total	6,000	22		6,000	14		70,000	150	

Compute the death rate for Occupation A and B!

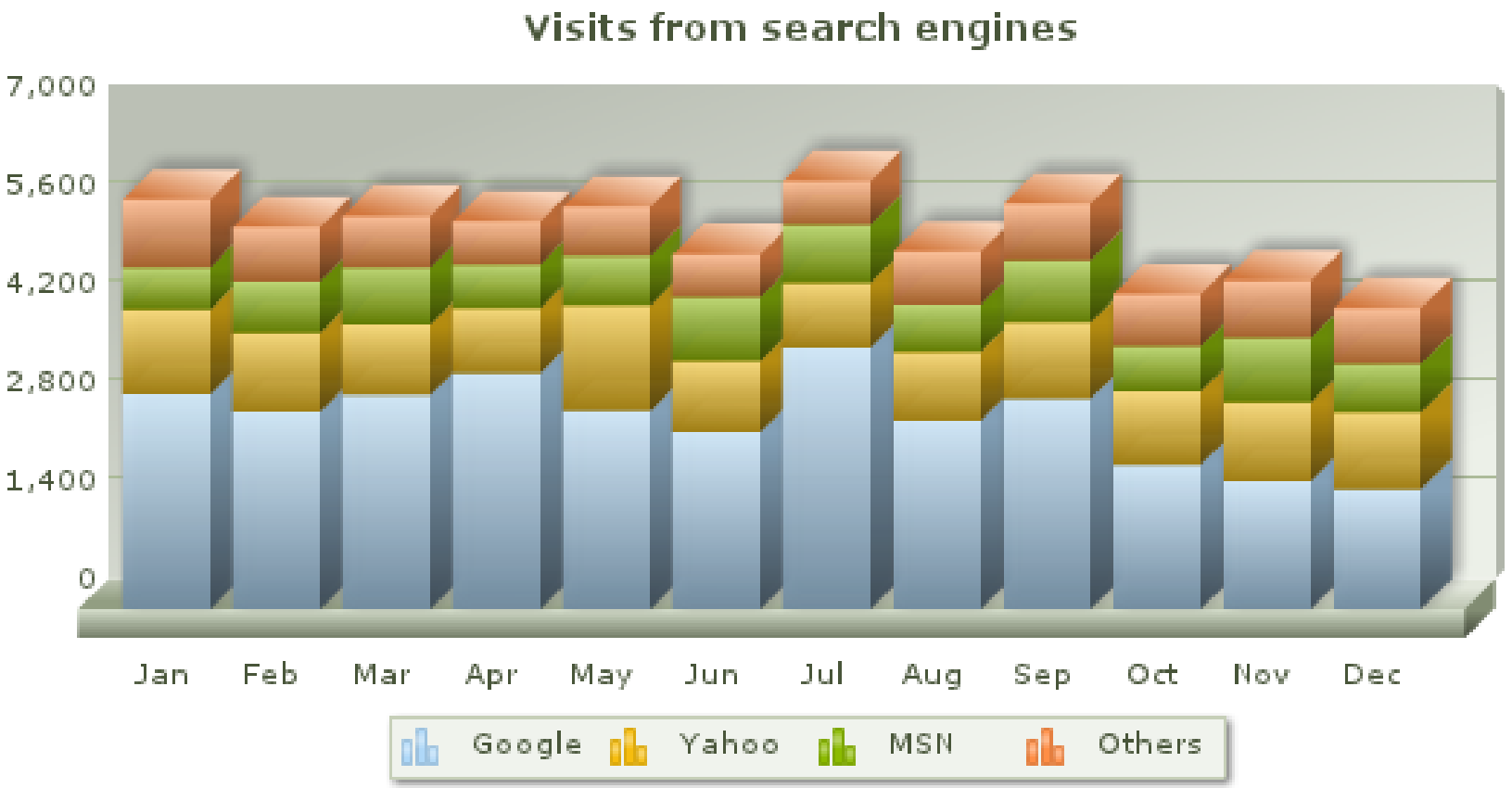
Graphs

Dodged bar



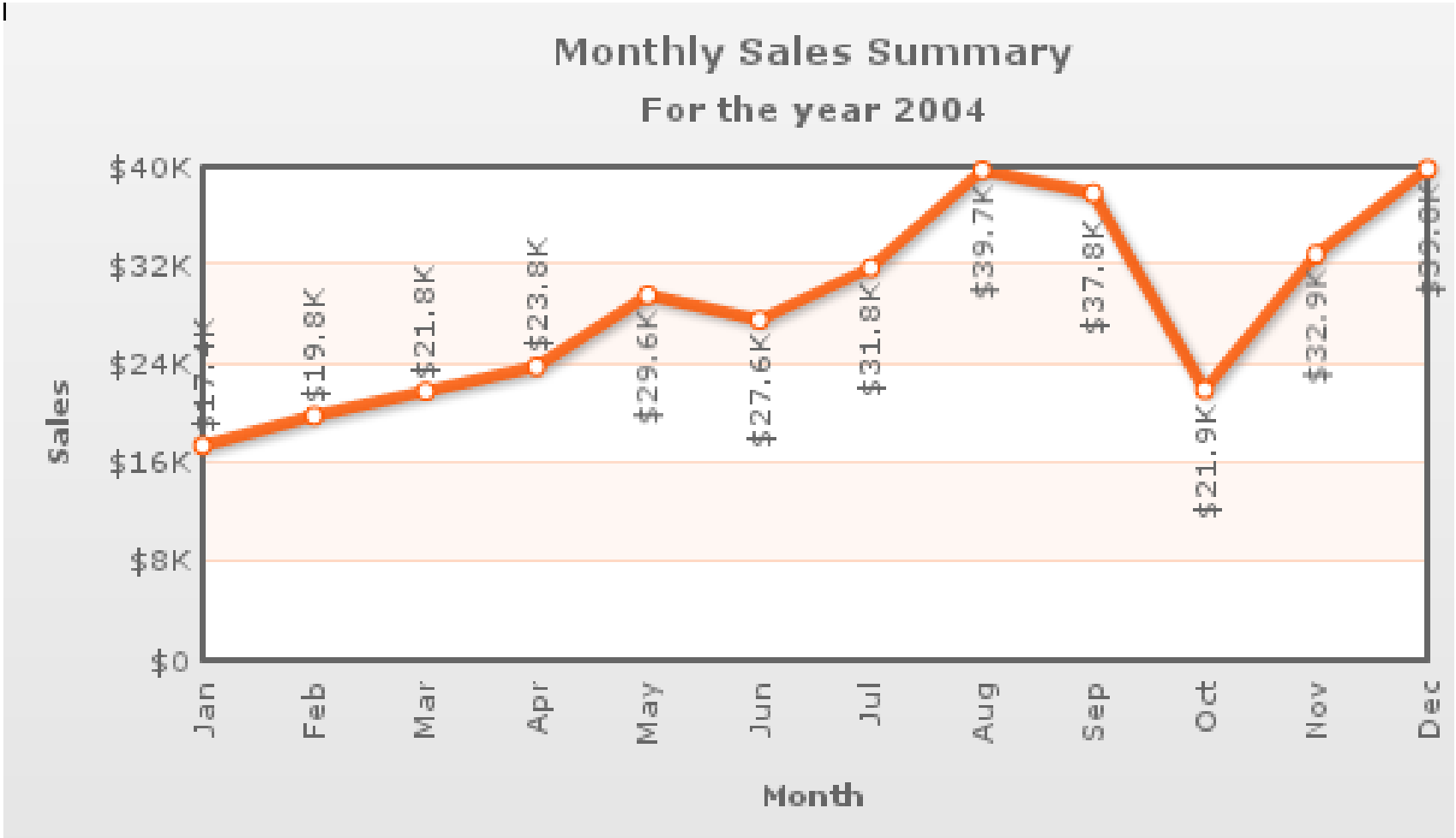
Graphs

Stacked bar



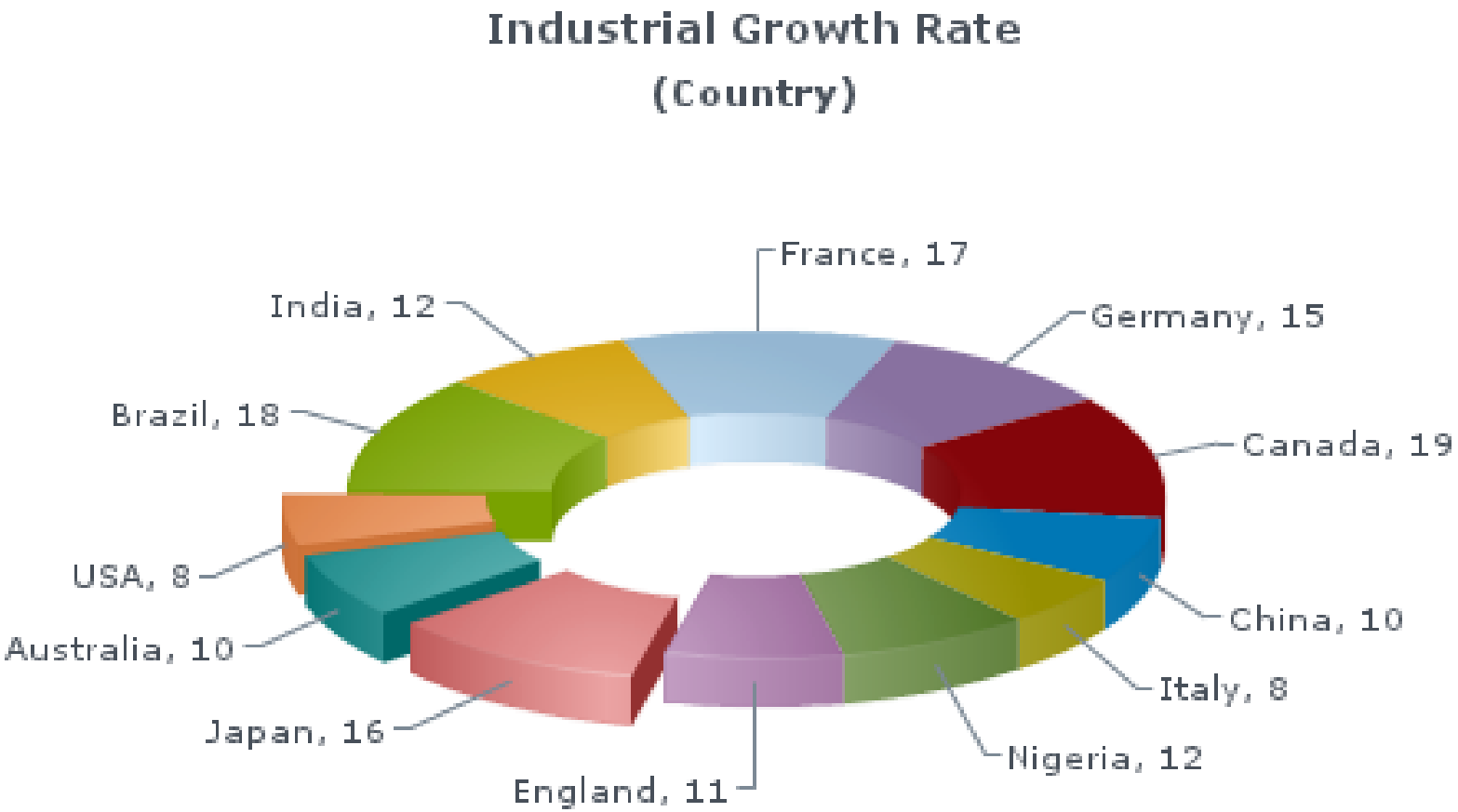
Graphs

Line


Monthly Sales Summary
For the year 2004
$40K
$32K
$24K
$16K
$8K
$0
Sales
$19.8K
$21.8K
$23.8K
$29.6K
$27.6K
$31.8K
$39.7K
$37.8K
$21.9K
$32.9K
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

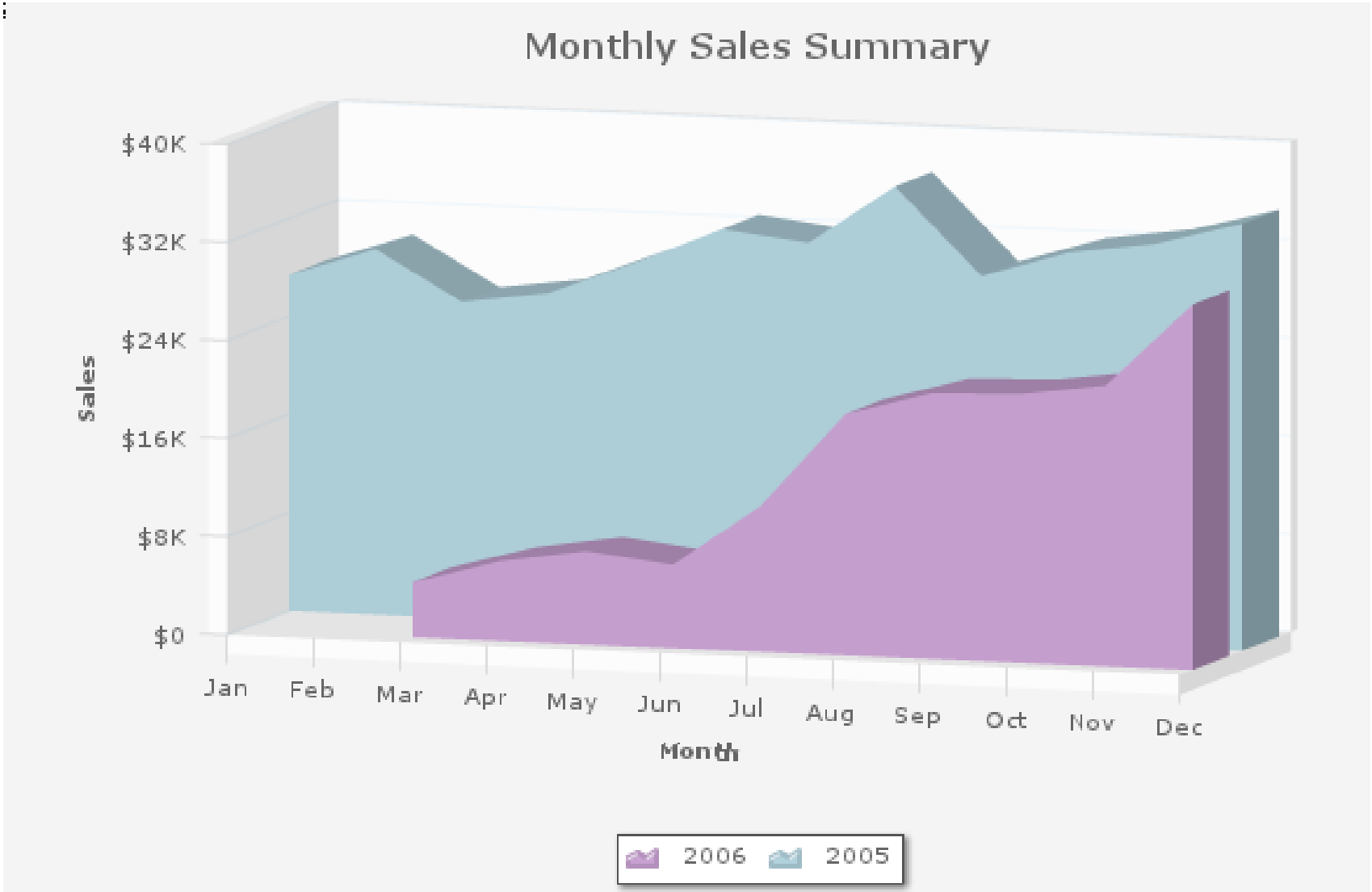
Graphs

Pie/donut



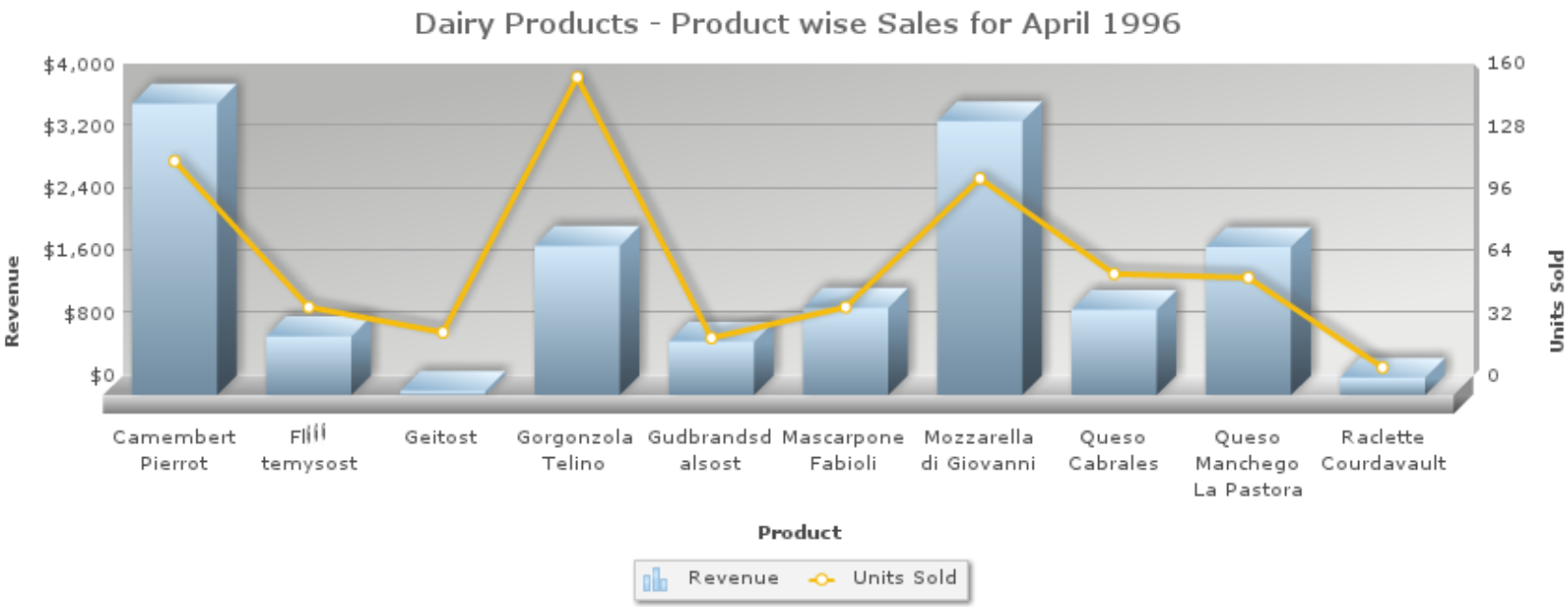
Graphs

Area


Monthly Sales Summary
$40K
$32K
$24K
$16K
$8K
$0
Sales
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
2006
2005

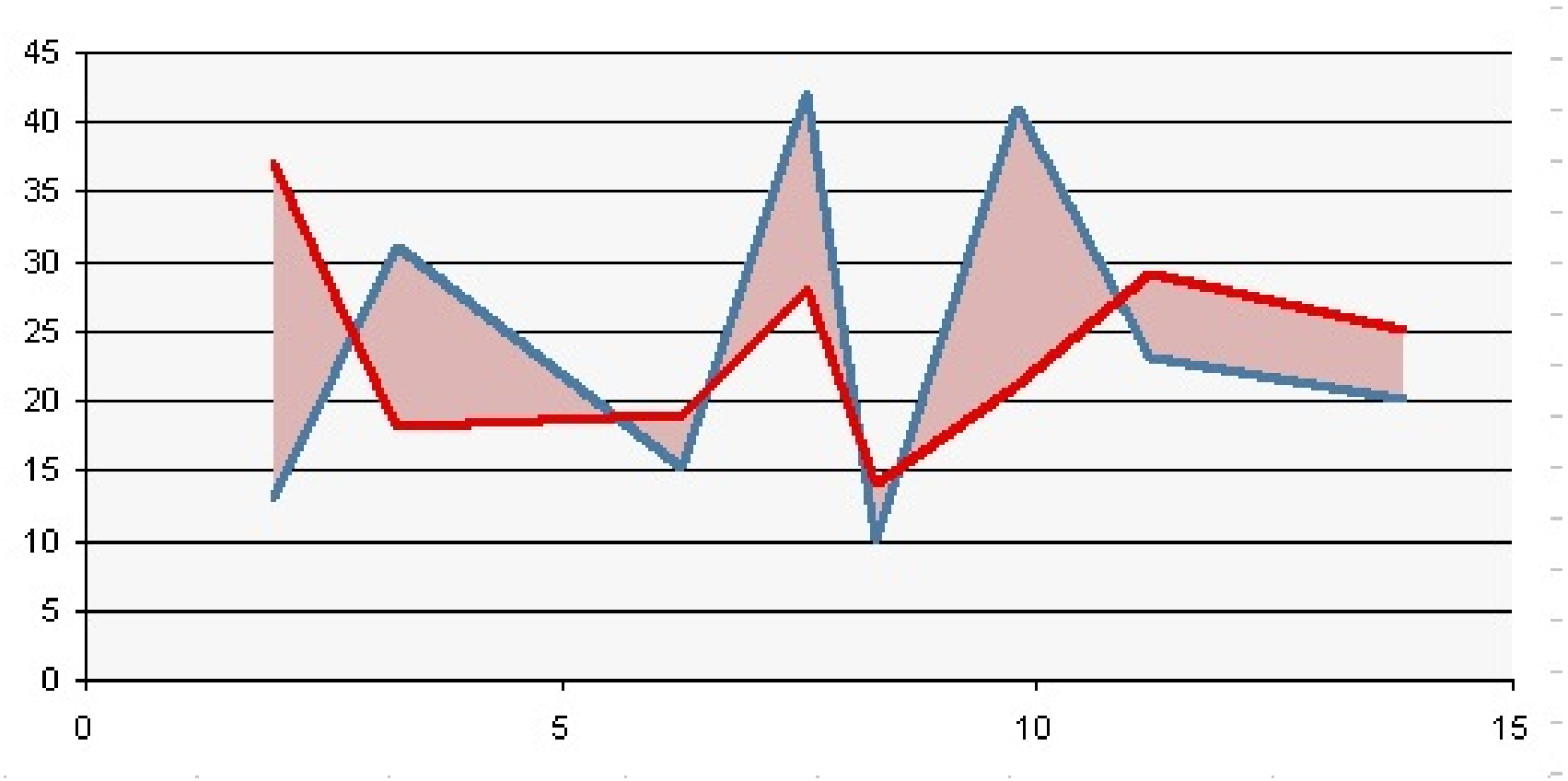
Graphs

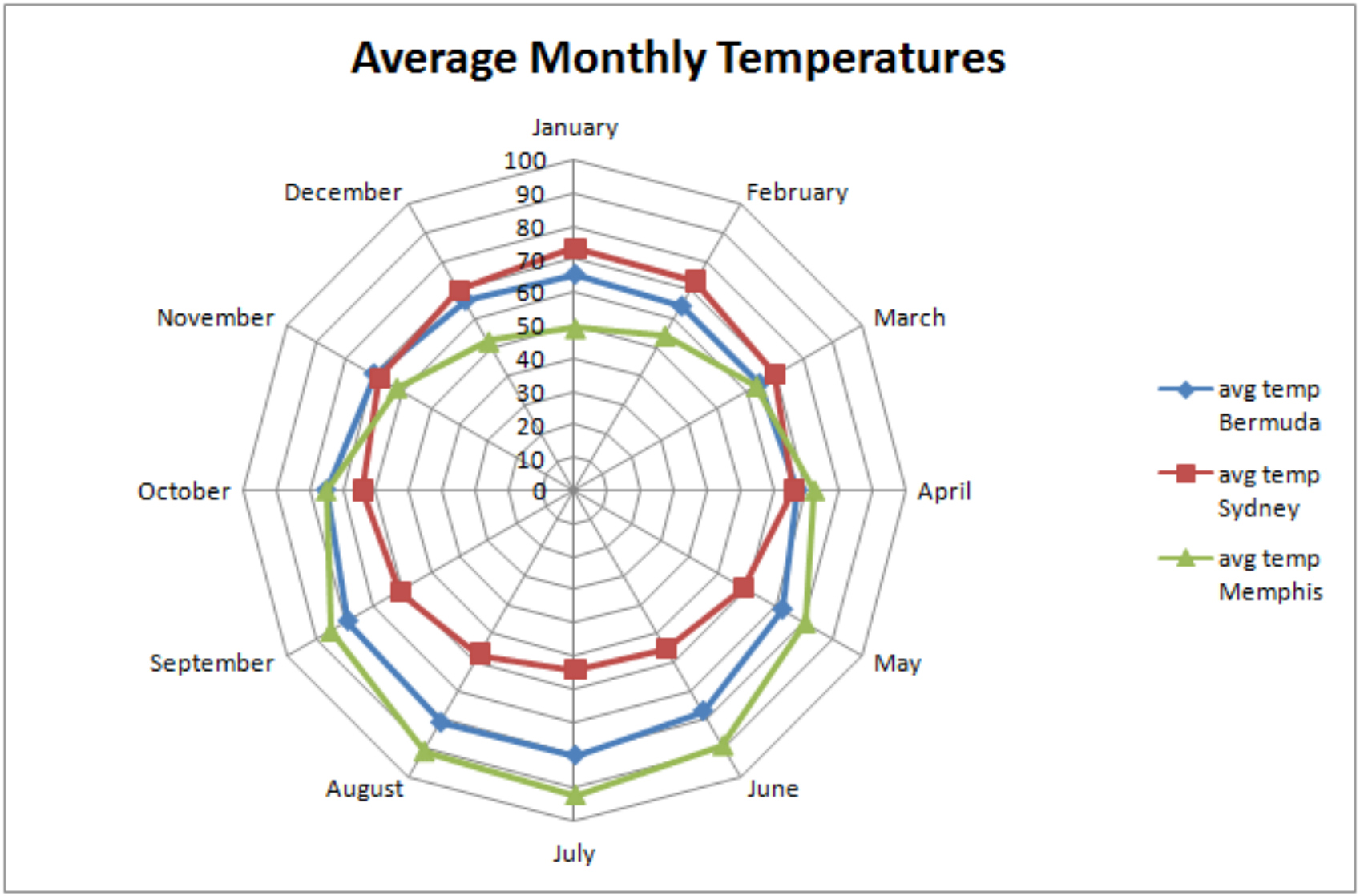
Combo



Graphs

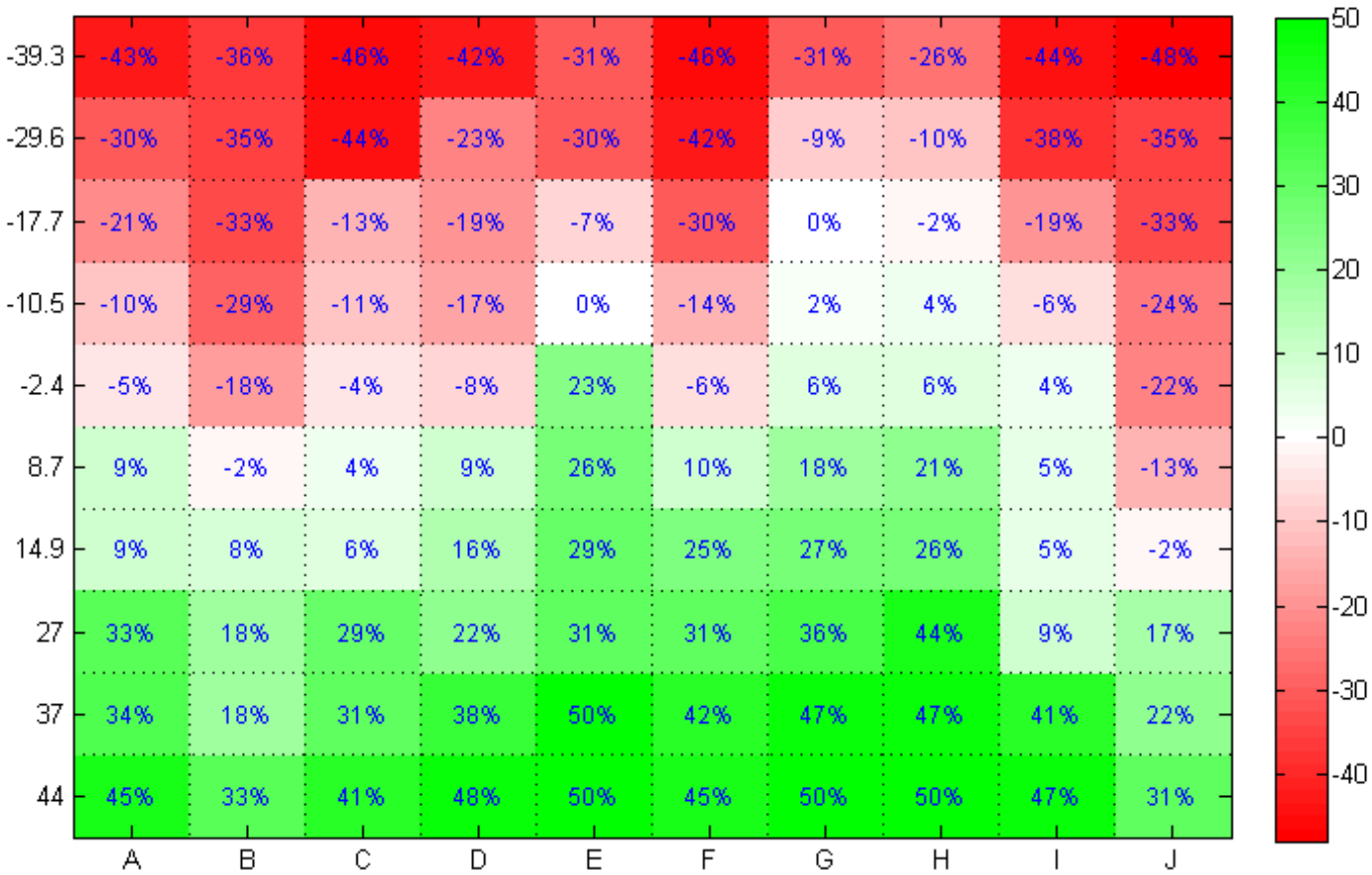
Combo




Average Monthly Temperatures
January
February
March
April
May
June
July
August
September
October
November
December
100
90
80
70
60
50
40
30
20
10
0
avg temp Bermuda
avg temp Sydney
avg temp Memphis

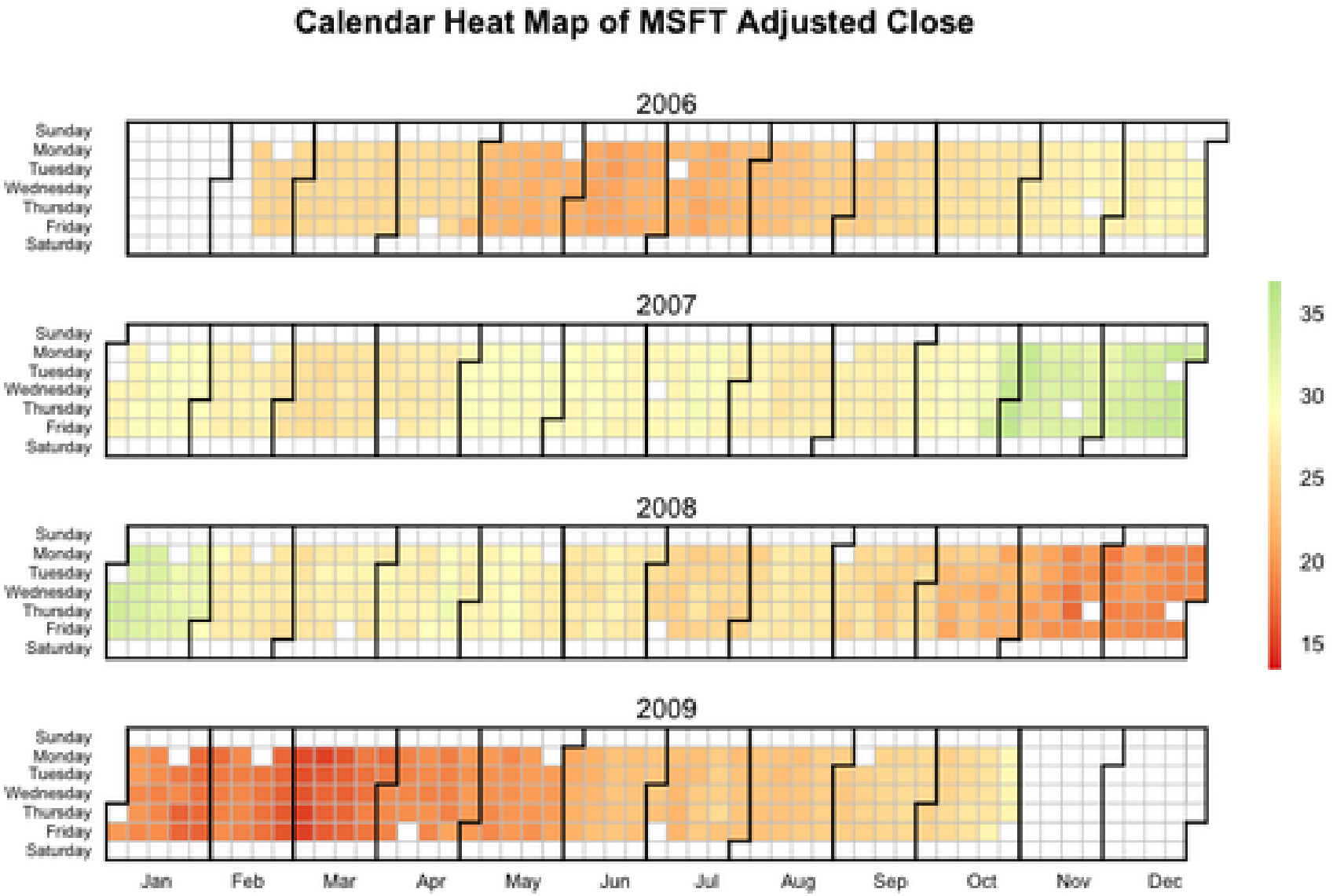
Graphs

Heatmap



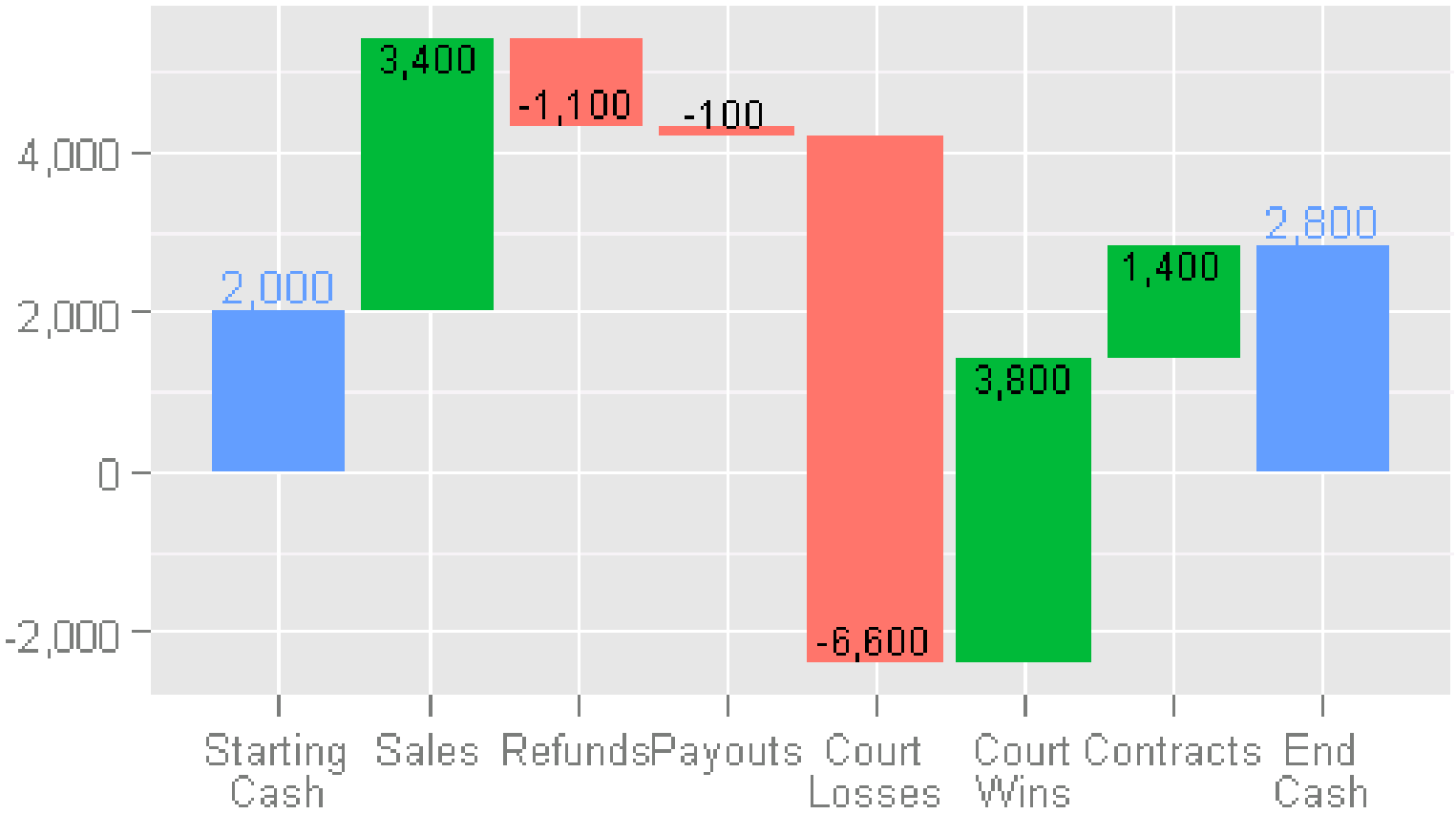
Graphs

Heatmap (calendar)



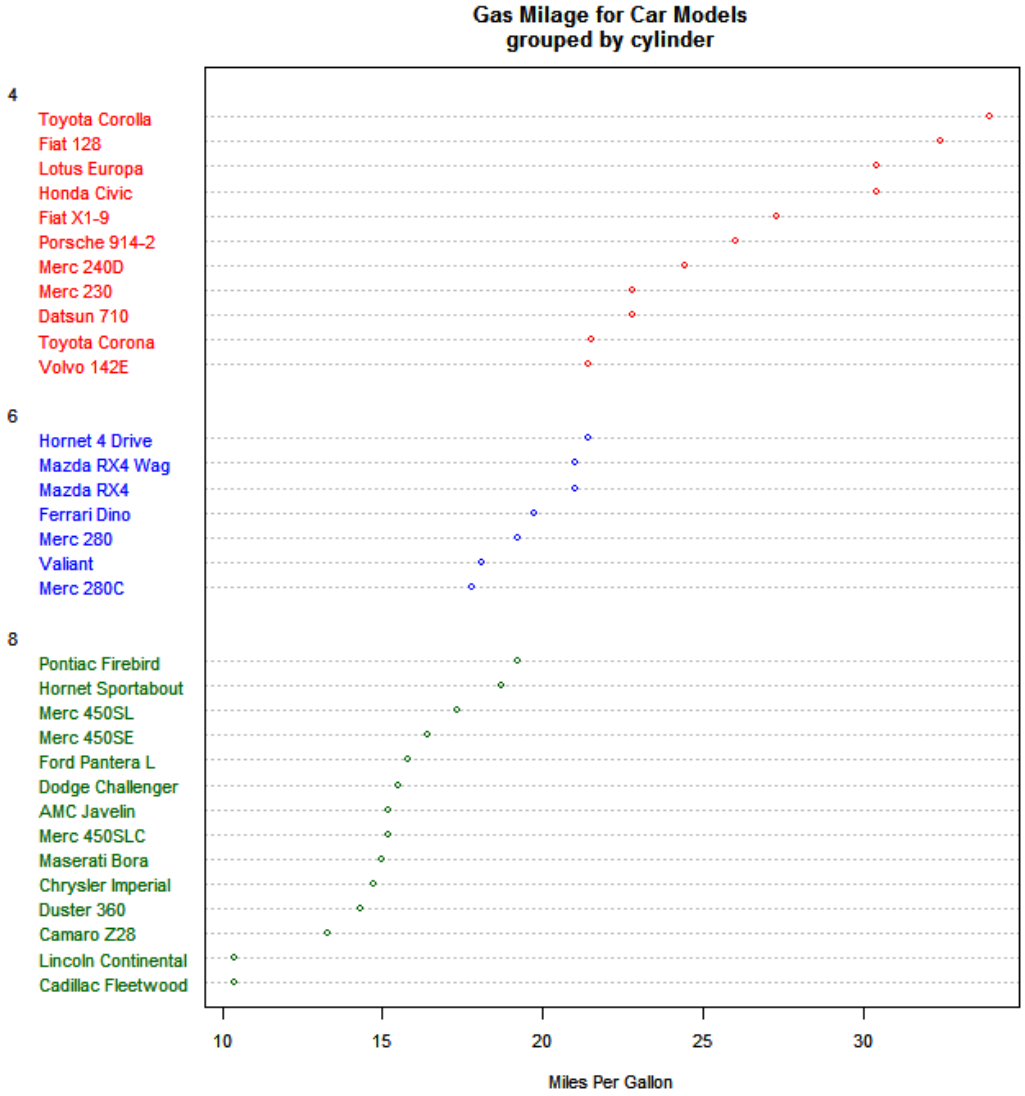
Graphs

Waterfall



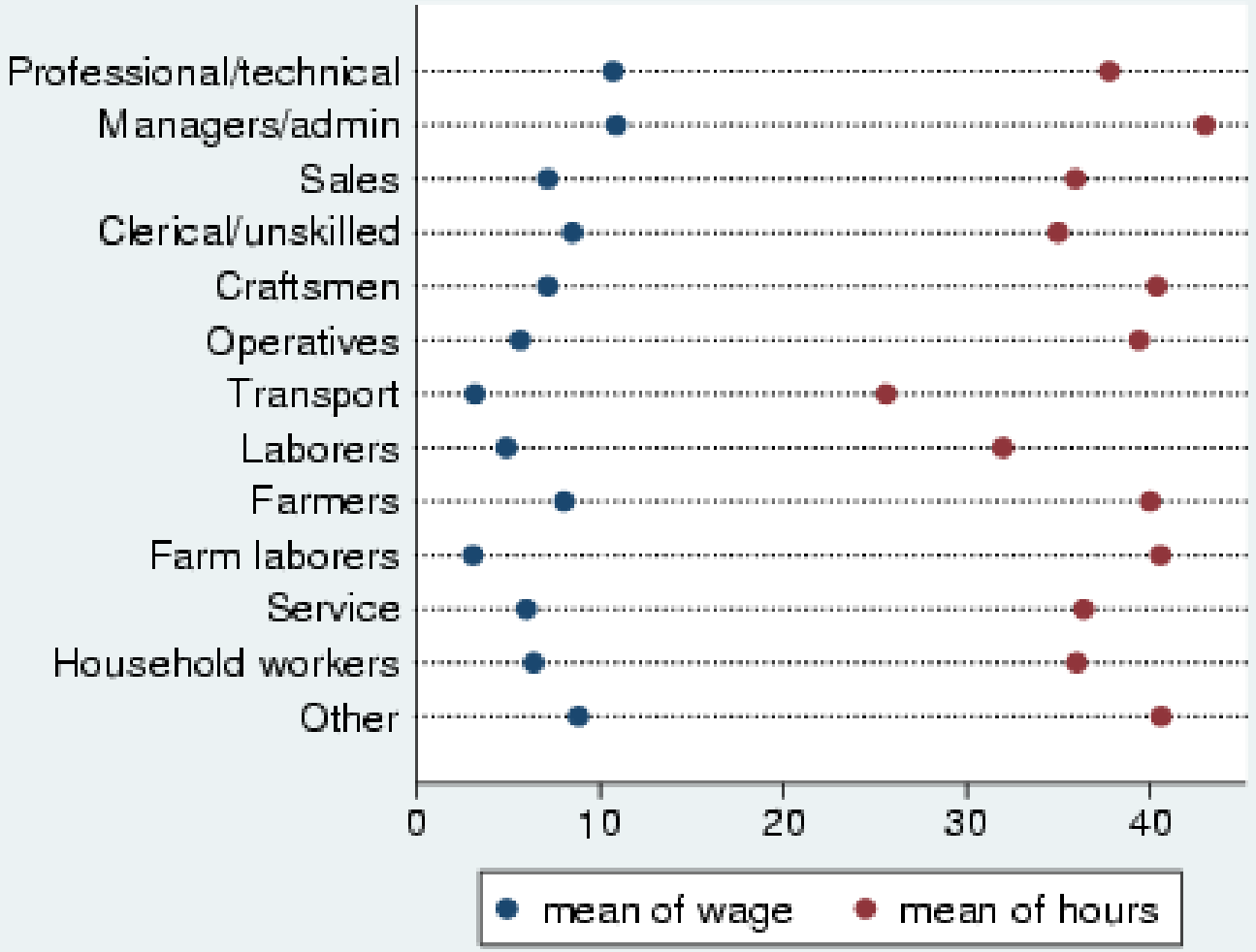
Graphs

Dot plot



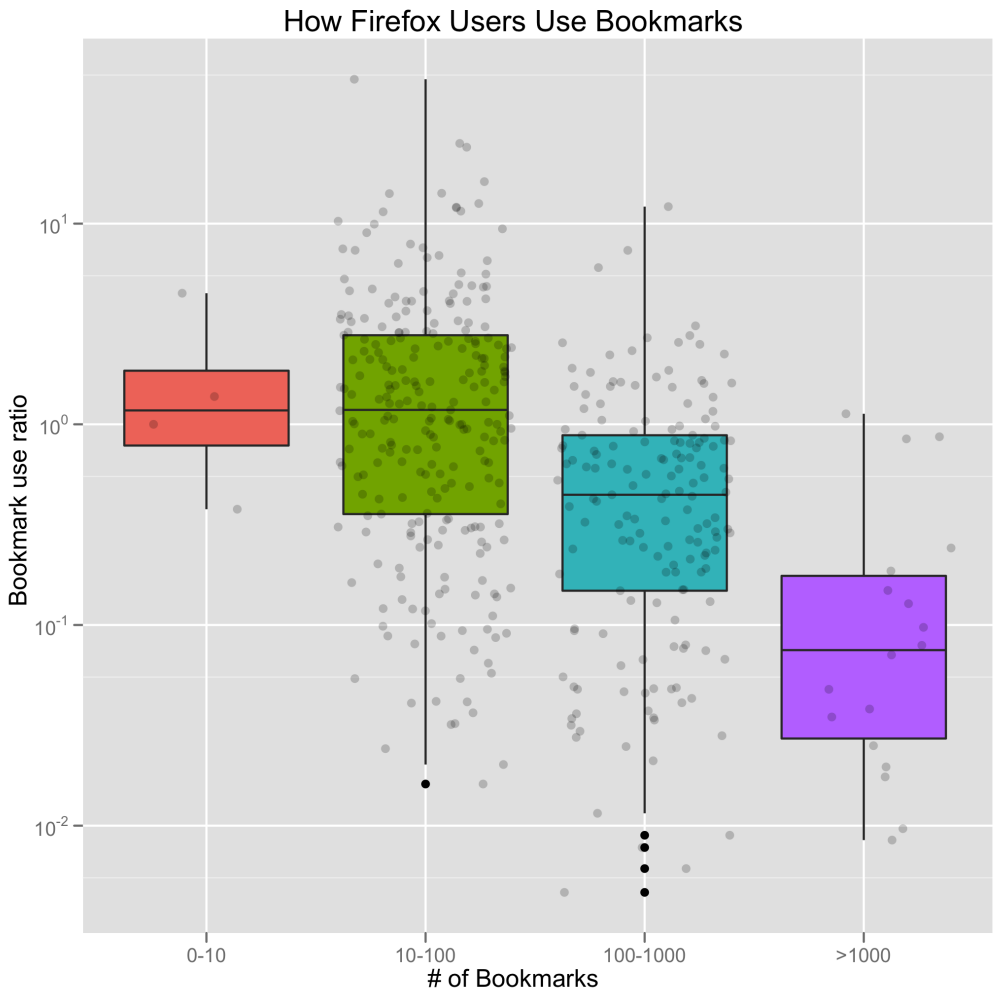
Graphs

Dot plot



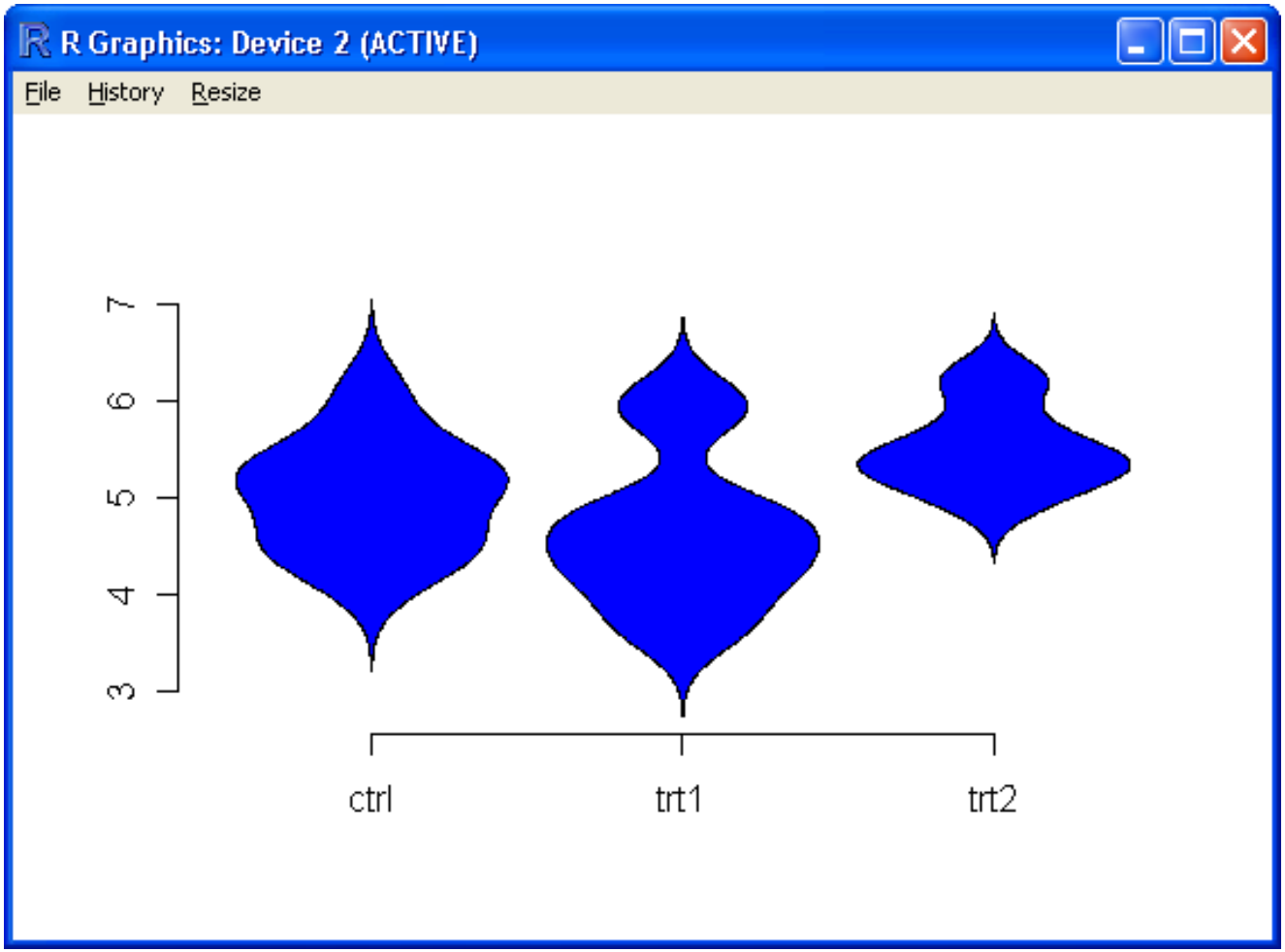
Graphs

Boxplot



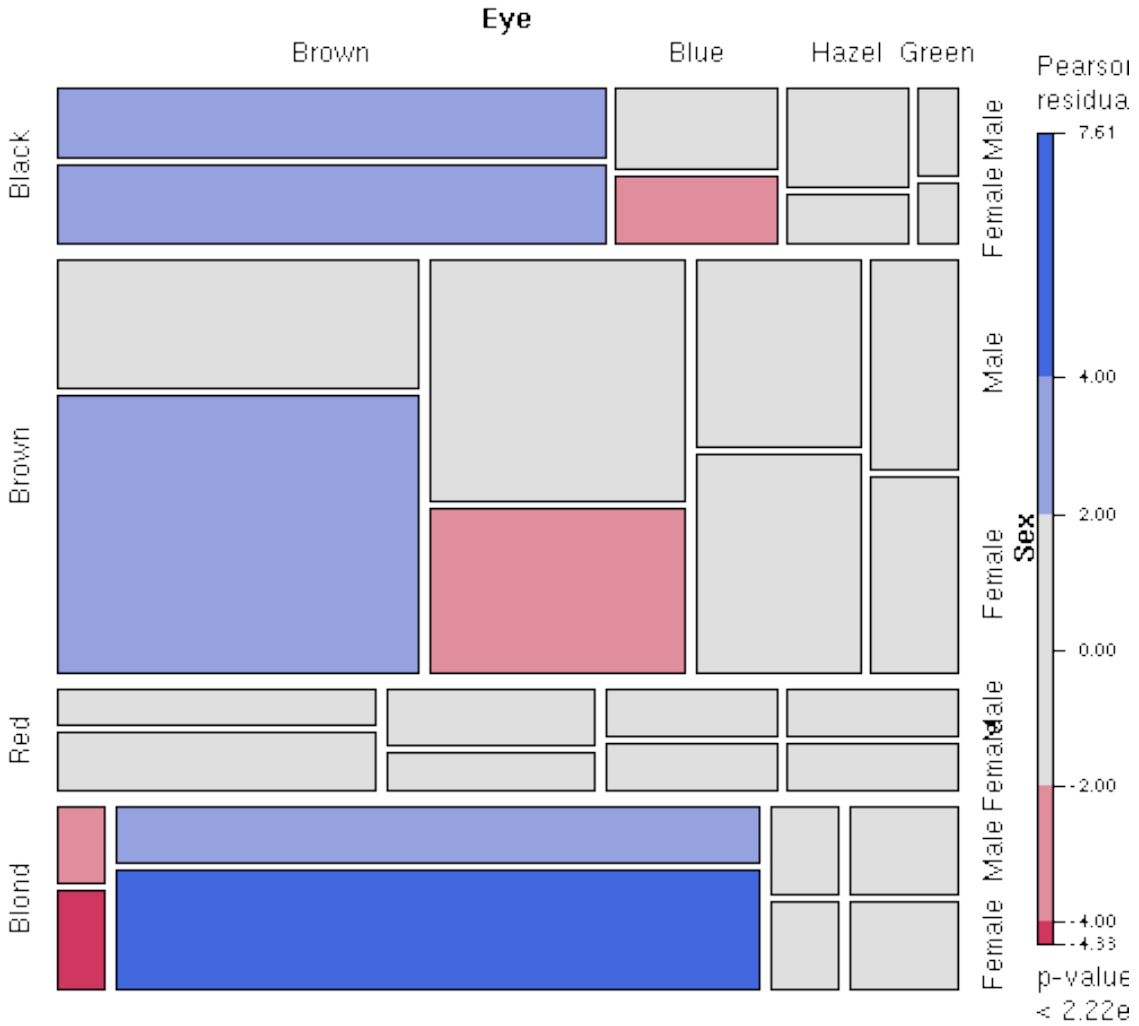
Graphs

Violin plot



Graphs

Mosaic chart

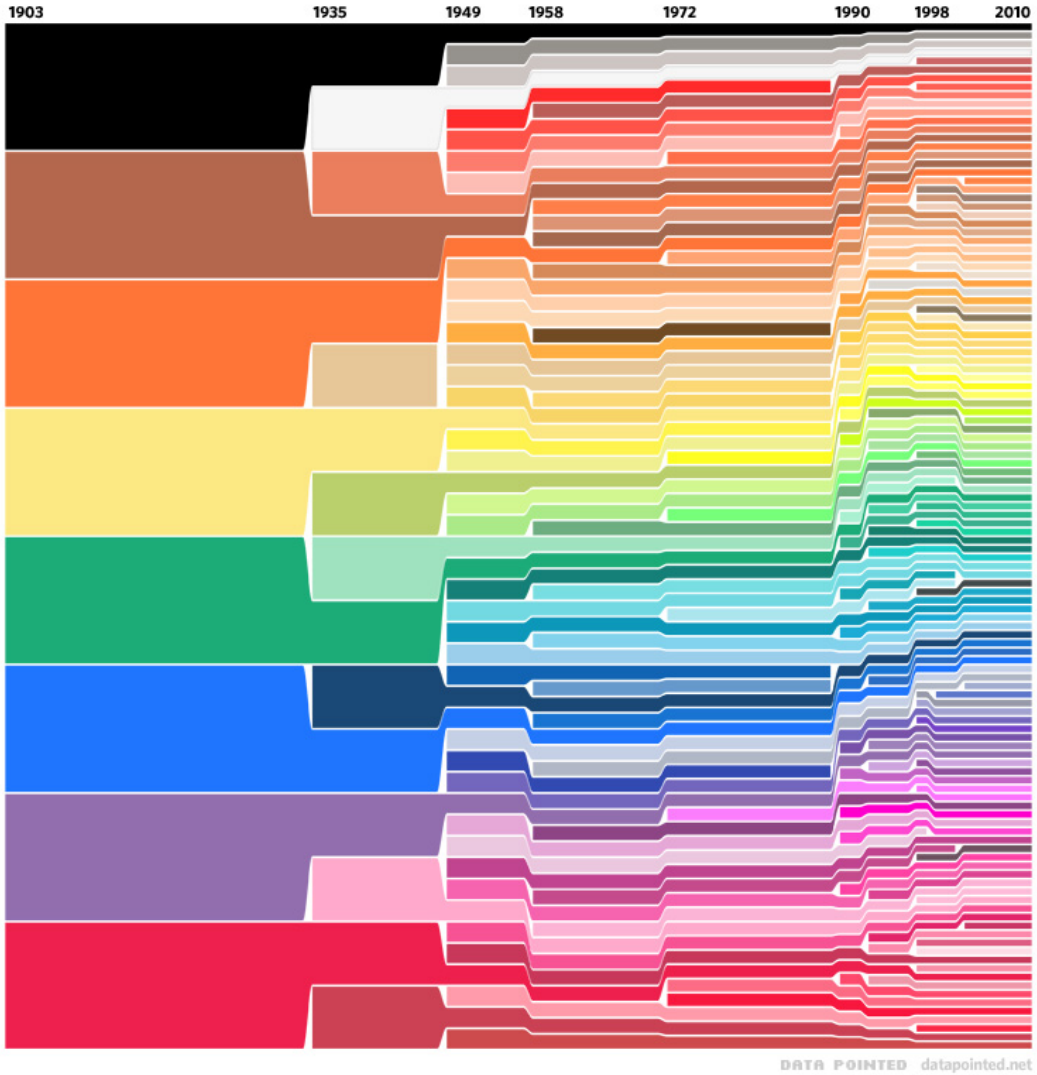


Graphs

Word cloud

Graphs

"Crayola Color Chart, 1903-2010"



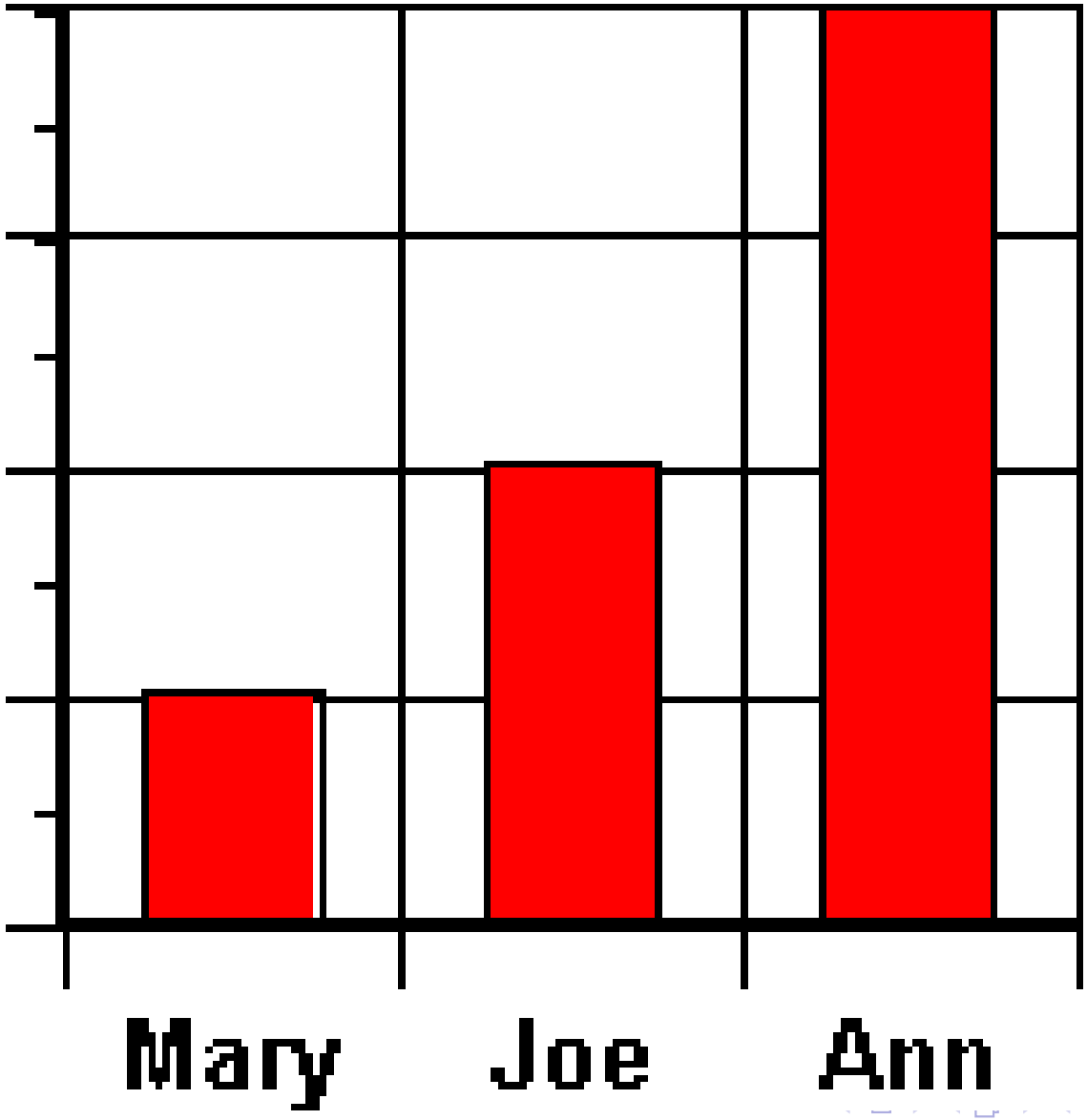
Graphs

Some interesting pages about the topic

- http://www.visual-literacy.org/periodic_table/periodic_table.html
- http://www.edwardtufte.com/tufte/
- http://www.perceptualedge.com/
- http://www.visualcomplexity.com/vc/
- http://flowingdata.com/
- http://infosthetics.com/
- http://chartsgraphs.wordpress.com/
- http://www.informationisbeautiful.net/
- http://chartporn.org/

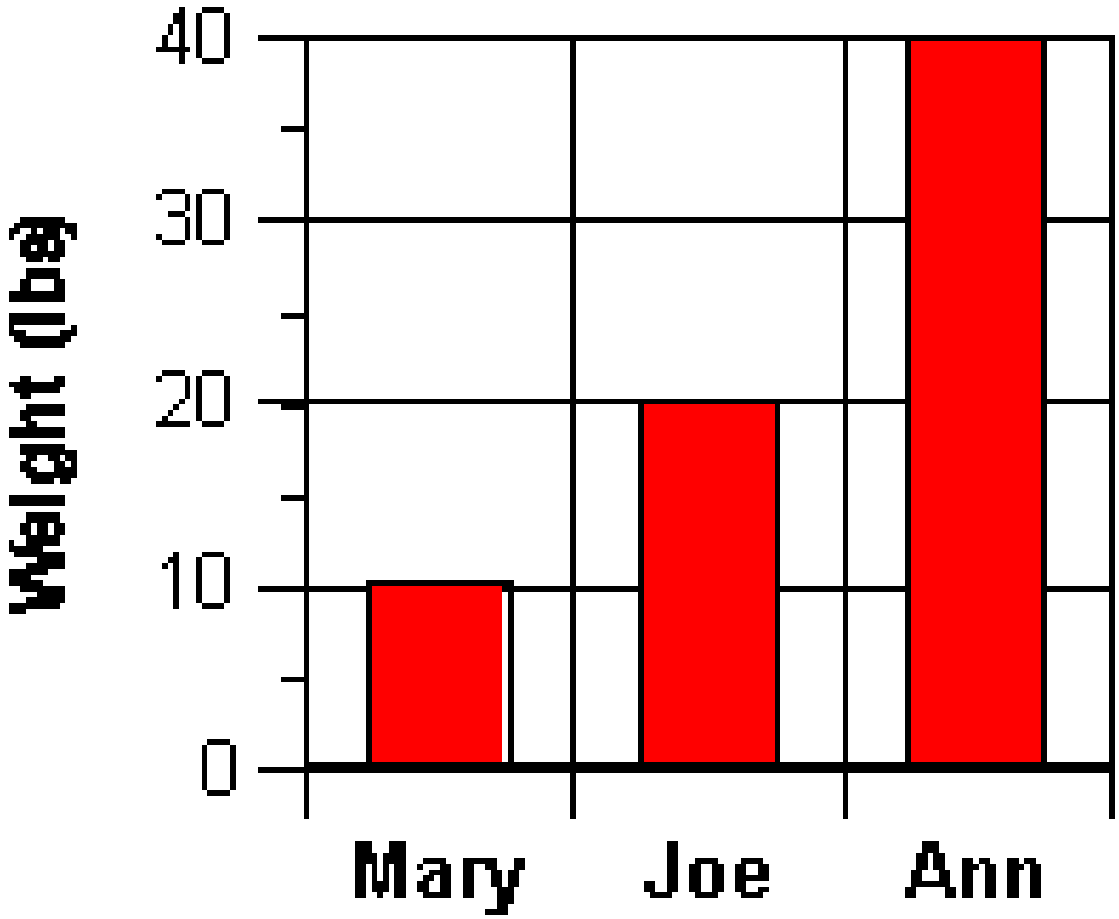
A small note on graphics

Pumpkins



A small note on graphics

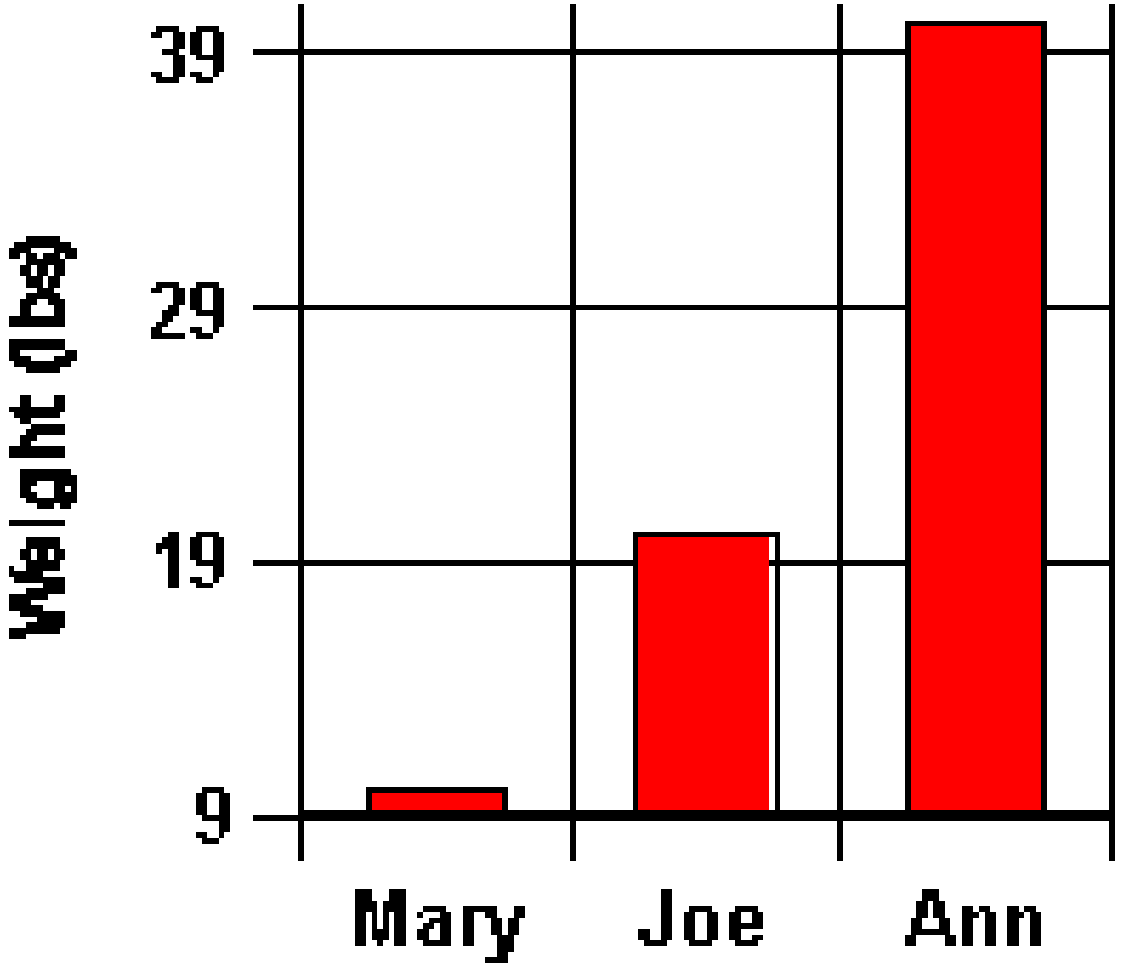
Pumpkins



Source: http://faculty.washington.edu/chudler/stat3.html

A small note on graphics

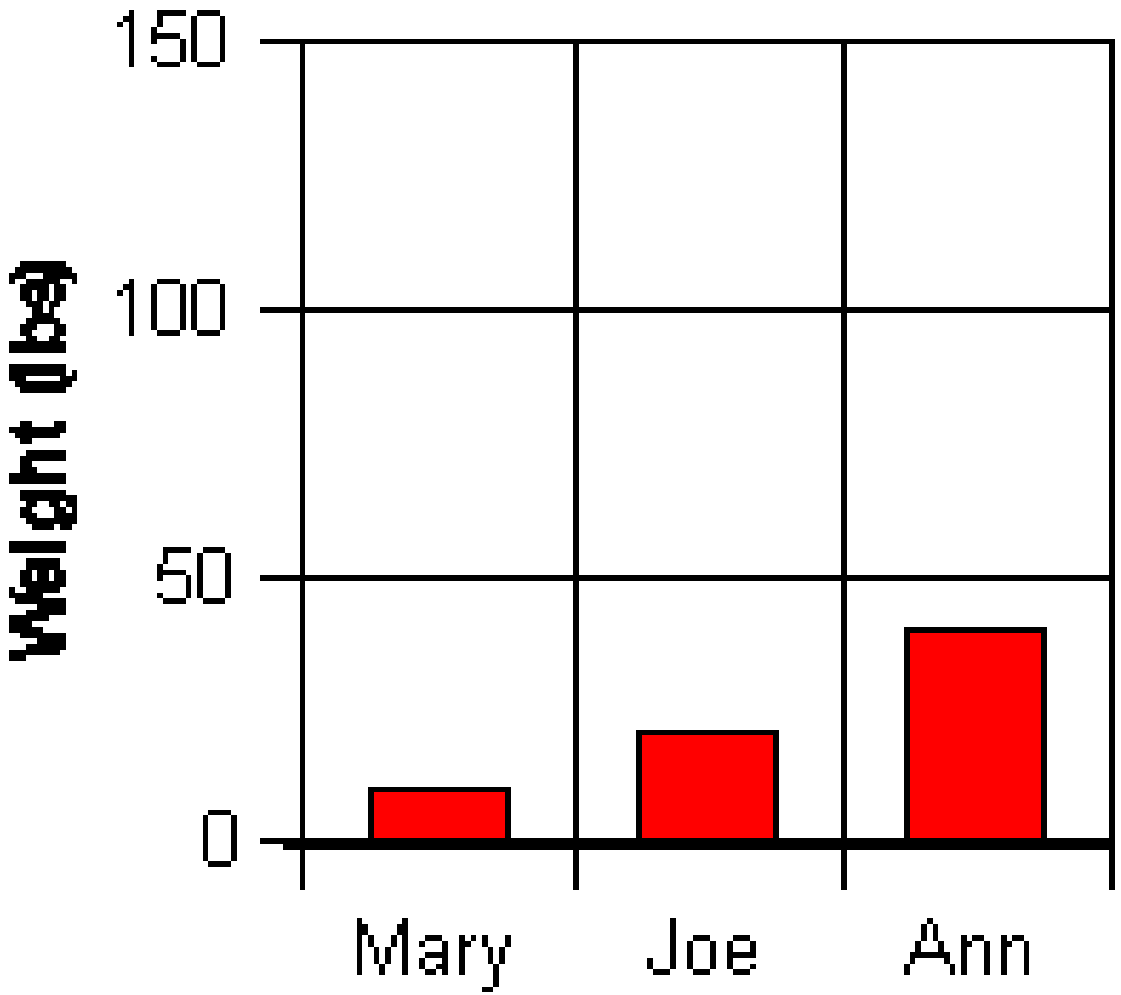
Pumpkins



Source: http://faculty.washington.edu/chudler/stat3.html

A small note on graphics

Pumpkins



Source: http://faculty.washington.edu/chudler/stat3.html

It was a pleasure!

Gergely Daróczi
daroczi.gergely@btk.ppke.hu