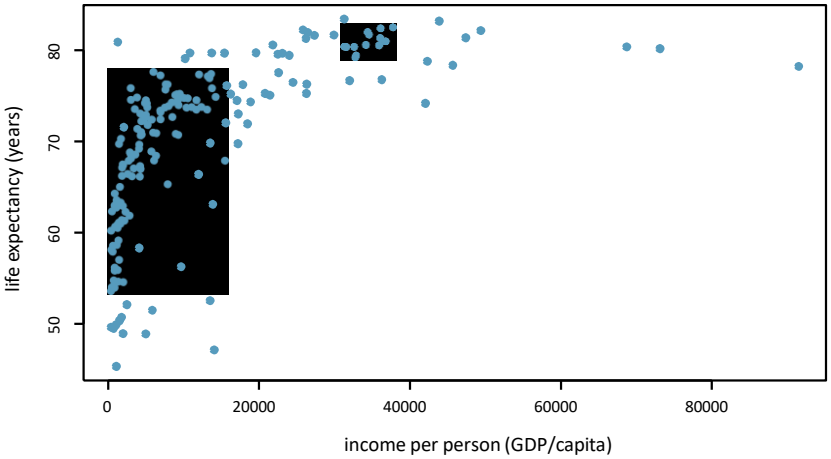
Introduction to Data Science

Statistical Charts

Visualizing Numerical Data

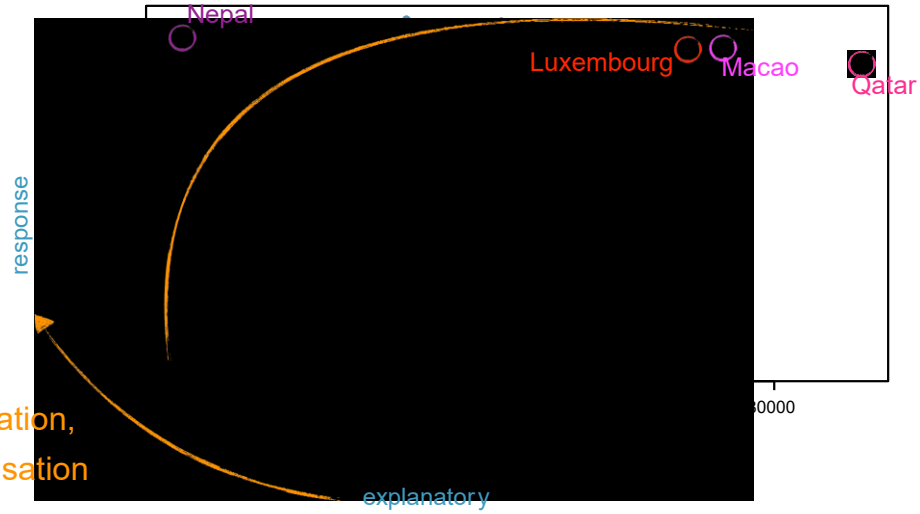
Scatterplot

data	income /person	life expectancy
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
⋮	⋮	⋮
Zimbabwe	545.3	58.142



- *Scatterplots* are useful for visualizing the relationship between two numerical variables.

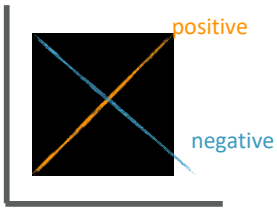
Scatterplot



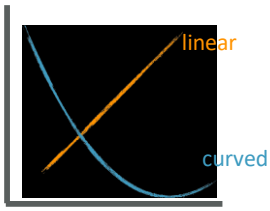
correlation,
not causation

Evaluating the relationship

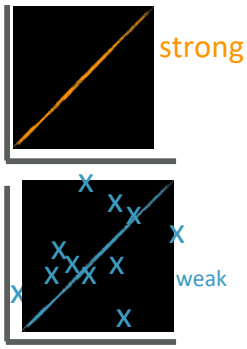
direction



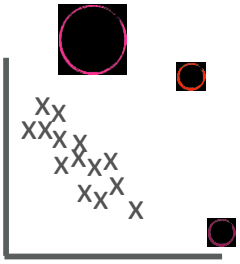
shape



strength



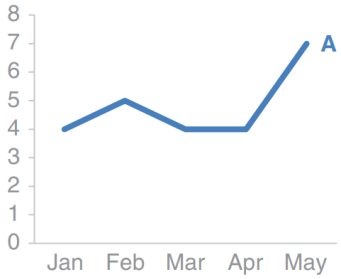
outliers



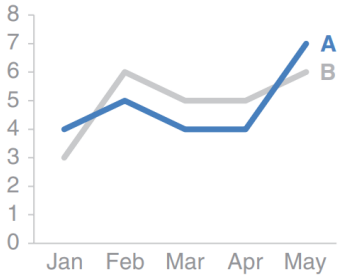
Line Graph

- Line graphs are used to plot continuous data often in some unit of time.

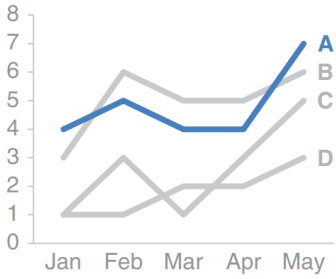
Single series



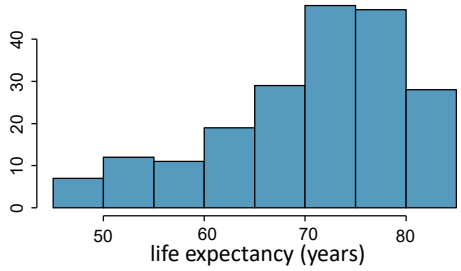
Two series

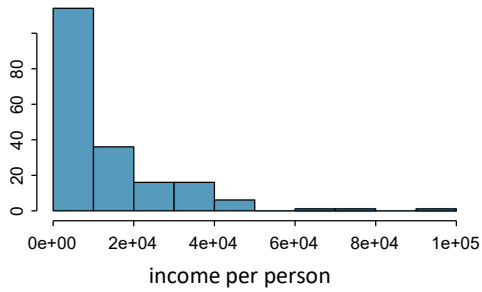


Multiple series



Histogram

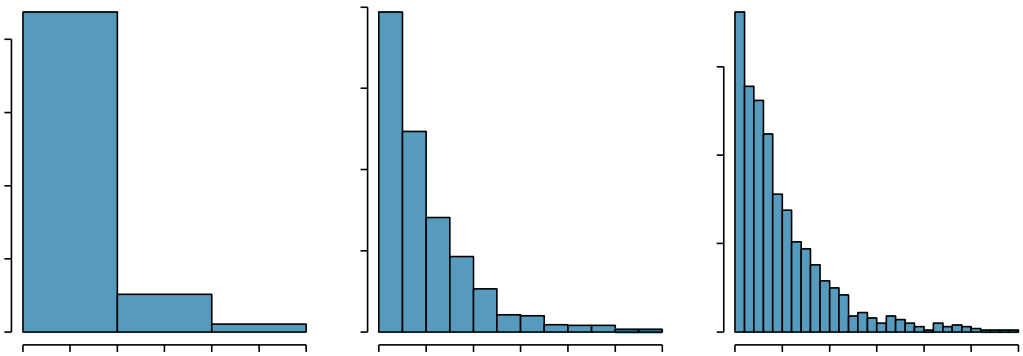




- Histograms provide a view of the data density.
- Histograms are especially convenient for describing the shape of the data distribution.
- The chosen bin width can alter the story the histogram is telling

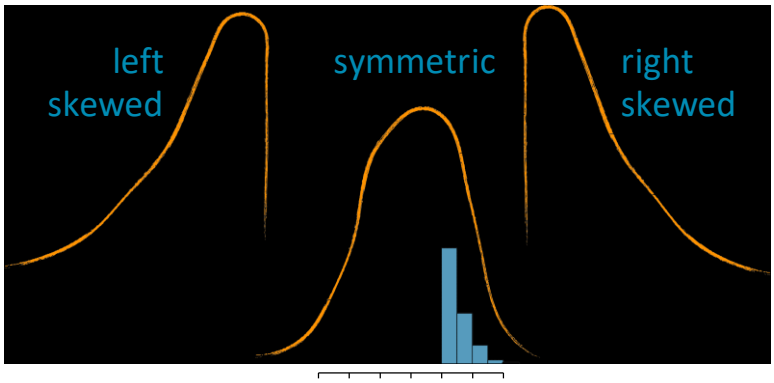
Bin Width

- When the bin width is too wide, we might lose interesting details.
- When the bin width is too narrow, it might be difficult to get an overall picture of the distribution.

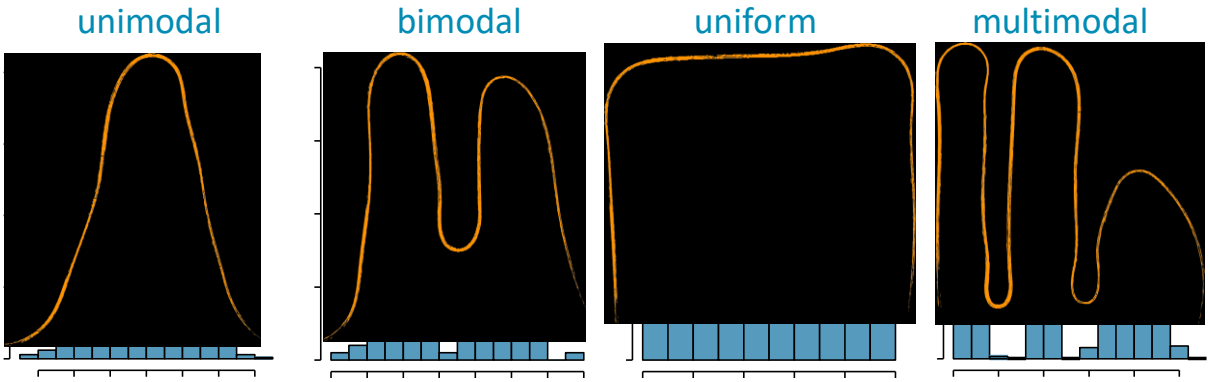


Skewness

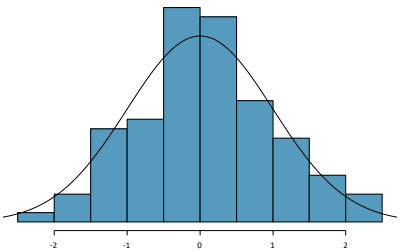
- Distributions are skewed to the side of the long tail



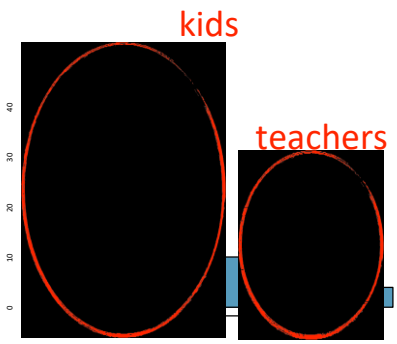
Modality



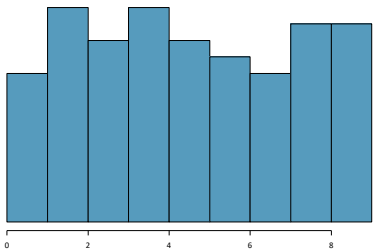
Modality



normal

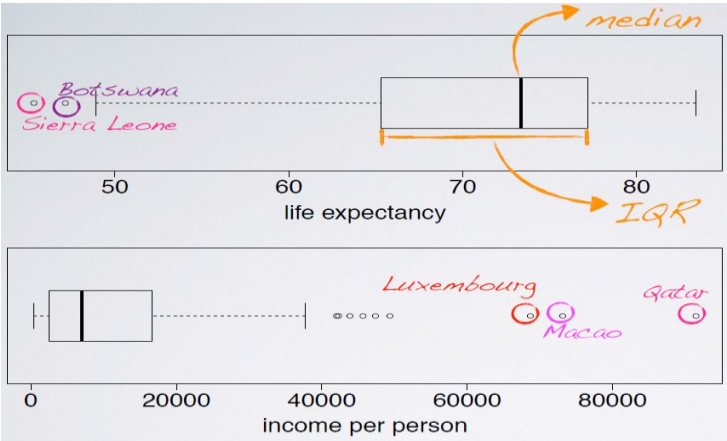


heights at elementary school



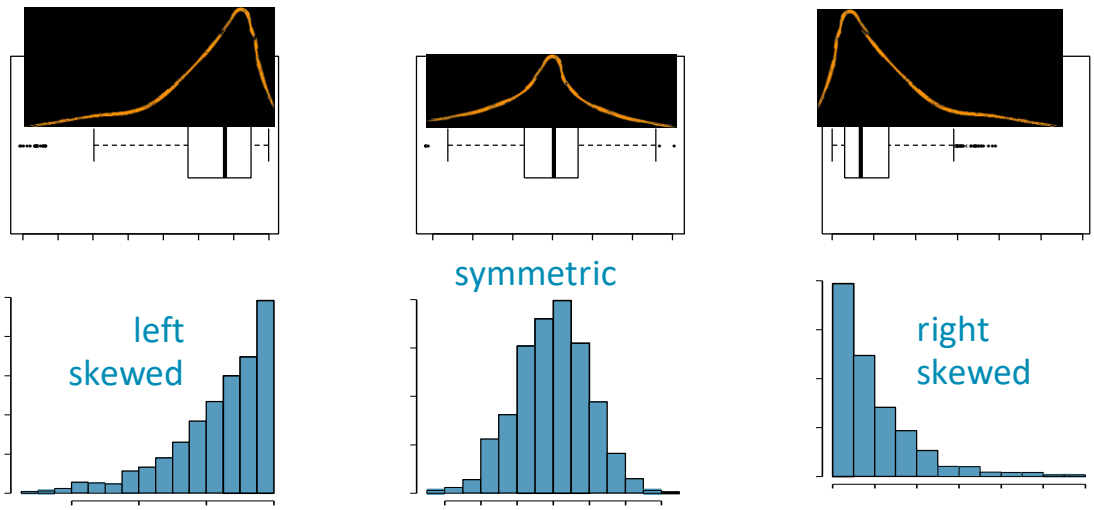
last digit of national ID

Box plot

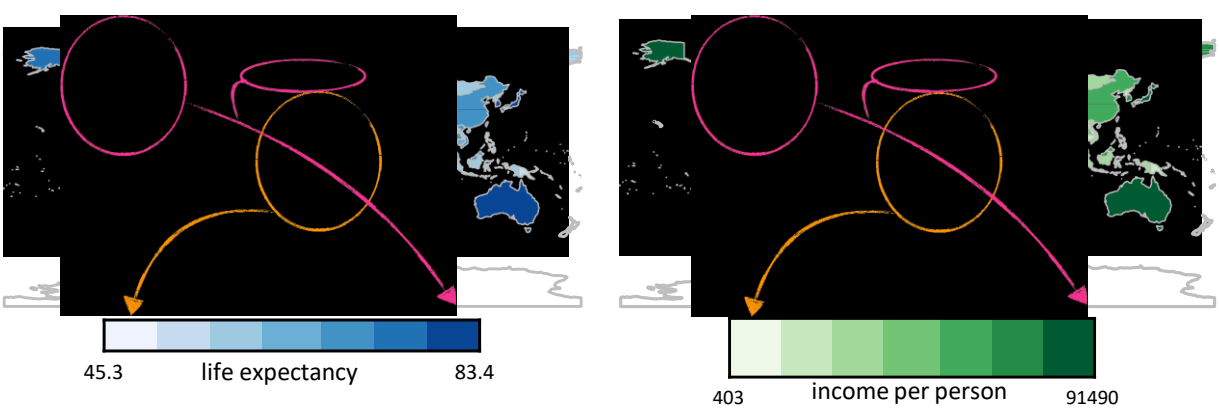


- Useful for highlighting outliers, median, IQR.

Determining the skewness from a box plot



Intensity Map



- Useful for highlighting the spatial distribution.

Measures of Center

- Mean: arithmetic average
  - Sample mean: $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$
  - Population mean: μ
- Median: midpoint of the distribution
  - 50th percentile
- Mode: most frequent observation

sample statistic

point estimate

population parameter

Example

- Nine students exam score:

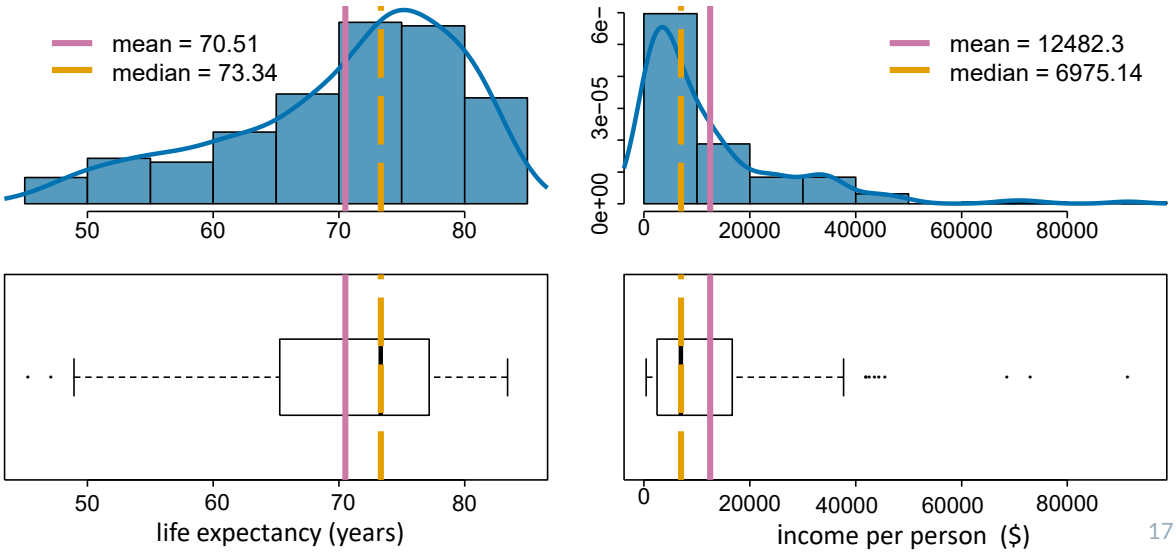
75, 69, 88, 93, 95, 54, 87, 88, 27

mean: $\frac{75+69+88+93+95+54+87+88+27}{9} = 75.11$

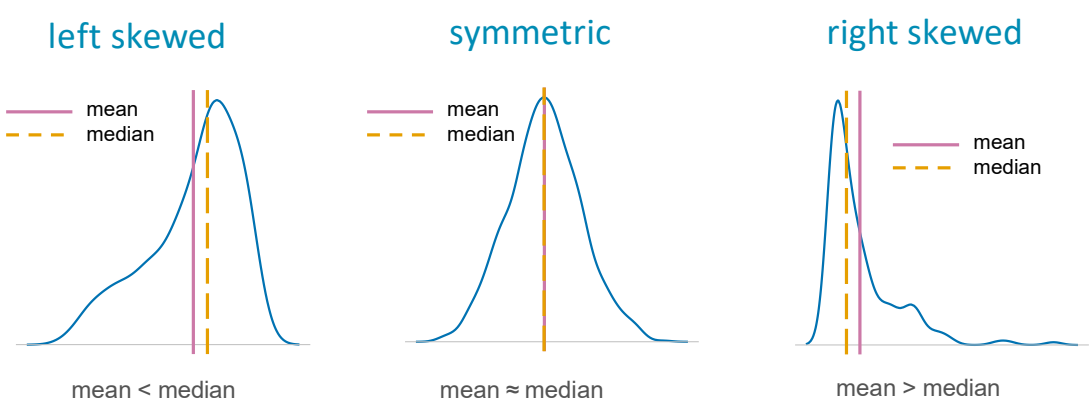
mode: 88

median: 27, 54, 69, 75, 88, 88, 93, 95

Relation between Mean and Median

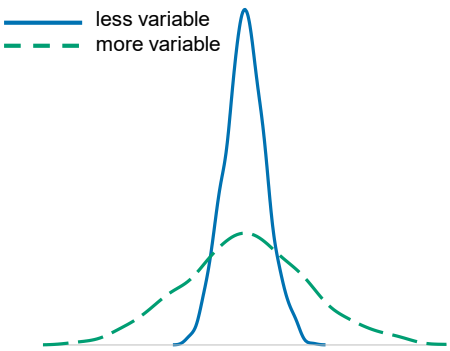


Skewness vs. Measures of Center



Measures of Spread

- In other words, statistics that tell us about the variability in the data:
  - Range = (*max* – *min*)
  - Variance
  - Standard deviation
  - Inter-quartile range



Variance

- Variance: roughly the average squared deviation from the mean
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - Population variance: σ^2

- Example: Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

	data	life expectancy
1	Afghanistan	60.254
2	Albania	77.185
3	Algeria	70.874
⋮	⋮	⋮
201	Zimbabwe	58.142

$$s^2 = \frac{(60.3 - 70.5)^2 + (77.2 - 70.5)^2 + \ldots + (58.1 - 70.5)^2}{201 - 1}$$

$$= 83.06 \text{ years}^2$$

Standard Deviation

- Standard deviation: roughly the average deviation from the mean that has the same units as the data
  - Sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

square root of the variance

  - Population standard deviation: σ

- Example: Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

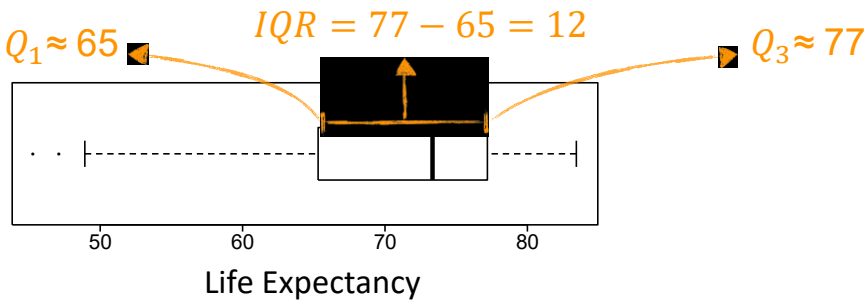
$$s = \sqrt{83.06} = 9.11 \text{ years}$$

Interquartile Range

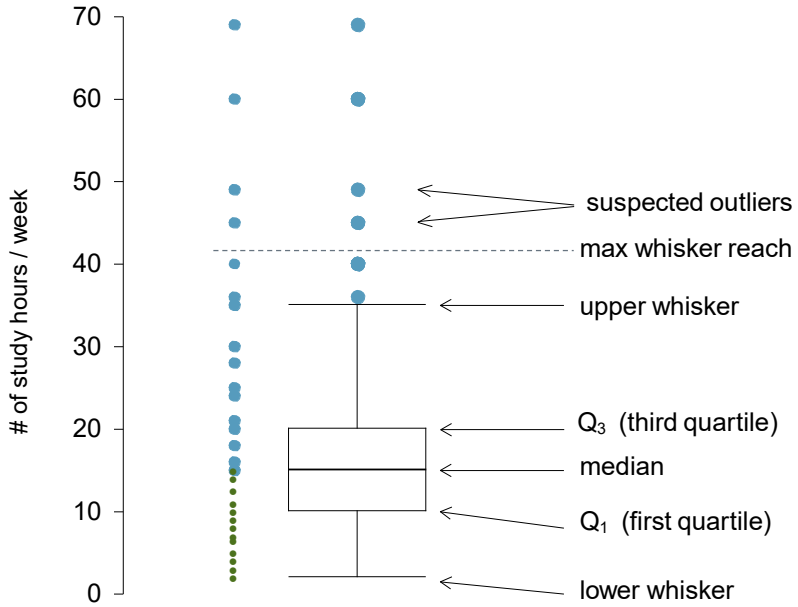
- Range of the middle 50% of the data, distance between the first quartile (25th percentile) and third quartile (75th percentile):

$$IQR = Q_3 - Q_1$$

$$IQR = 77 - 65 = 12$$



Boxplot



Whiskers

- The whiskers attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times \mathrm{IQR}$:

  max upper whisker reach = $Q_3 + 1.5 \times \mathrm{IQR}$

  max lower whisker reach = $Q_1 - 1.5 \times \mathrm{IQR}$

- Example:

  IQR : 20 − 10 = 10

  max upper whisker reach = 20 + 1.5 × 10 = 35

  max lower whisker reach = 10 − 1.5 × 10 = −5

- A potential outlier is defined as an observation beyond the maximum reach of the whiskers.
  - An observation that appears extreme relative to the rest of the data.

Outliers

- Why it is important to look for outliers?

- Examination of data for possible outliers serves many useful purposes, including:
  1. Identifying strong skew in the distribution.
  2. Identifying data collection or entry errors.
  3. Providing insight into interesting properties of the data.

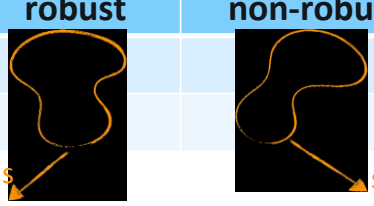
Robust Statistics

- We define robust statistics as measures on which extreme observations have little effect.

- Example:

Data	Mean	Median
1, 2, 3, 4, 5, 6	3.5	3.5
1, 2, 3, 4, 5, 1000	169	3.5

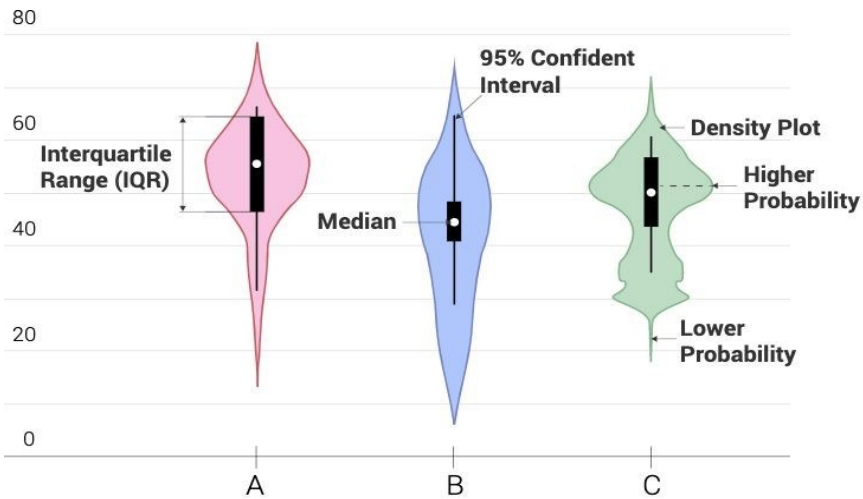
	robust	non-robust
center		
spread		



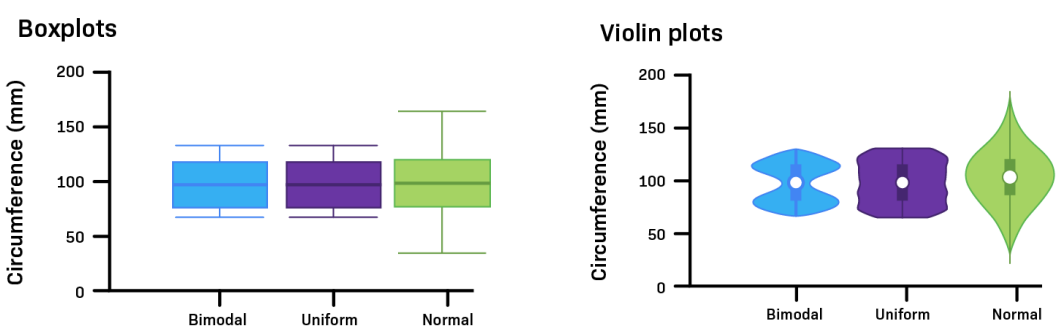
skewed, with extreme observations

symmetric

Violin Plot



Violin Plot vs. Box Plot



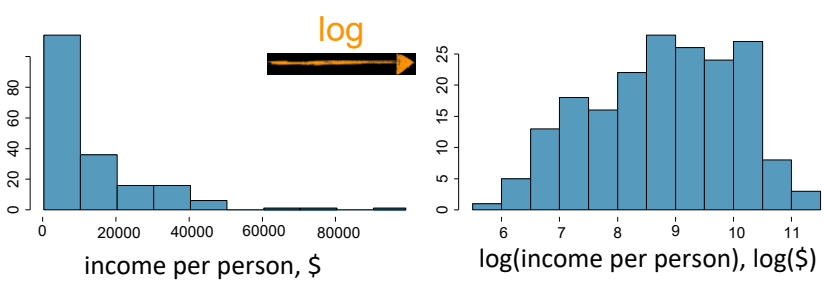
Data Transformation

- A transformation is a rescaling of the data using a function.
  - Log transformation
  - Square root transformation
  - Inverse transformation

- When data are very strongly skewed, we sometimes transform them so they are easier to model.

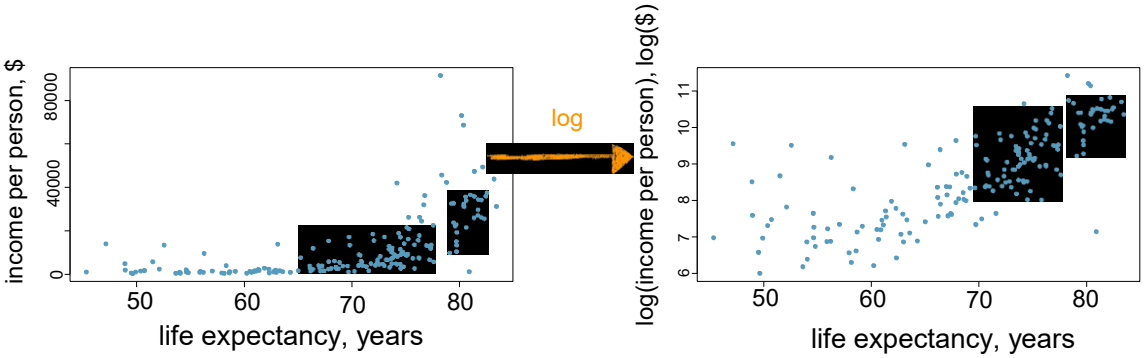
Log Transformation

- Often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive.

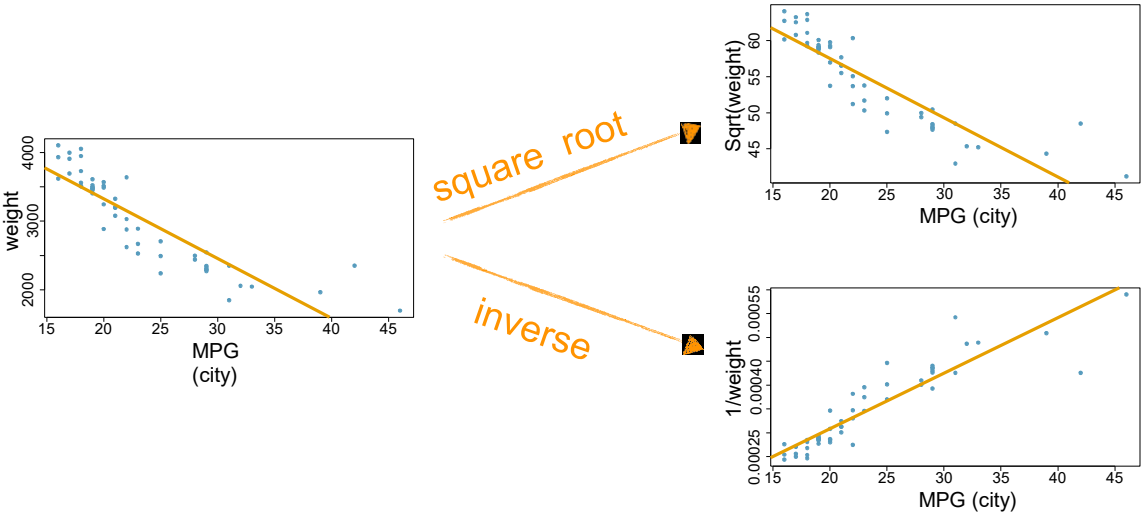


Log Transformation

- To make the relationship between the variables more linear, and hence easier to model with simple methods



Other Transformations



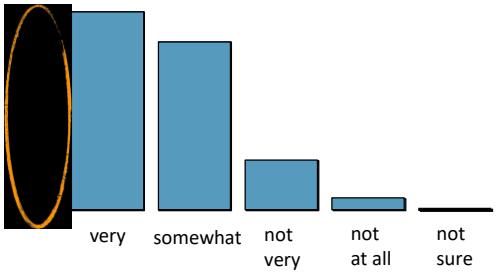
Goals of Transformation

- To see the data structure differently.
- To reduce skew and assist in modeling.
- To straighten a nonlinear relationship in a scatterplot.
- To model the relationship with simpler methods.

Describing Categorical Variables

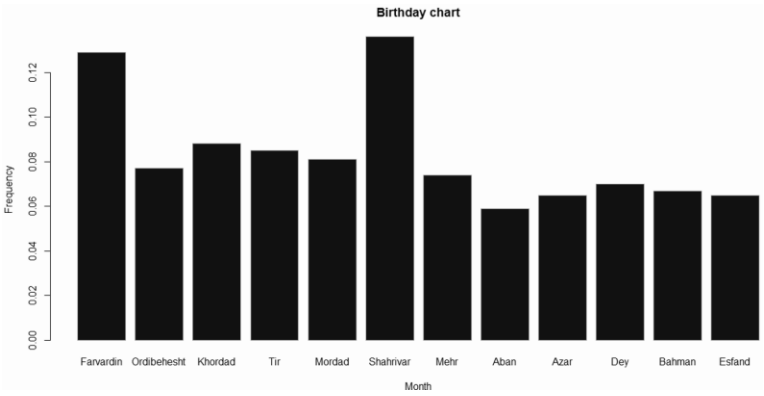
Frequency Table & Bar Plot

Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%

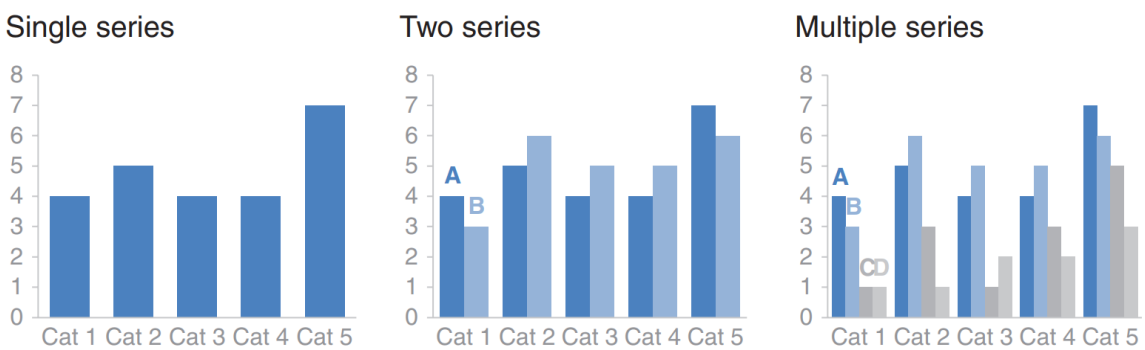


Birthdays in Iran

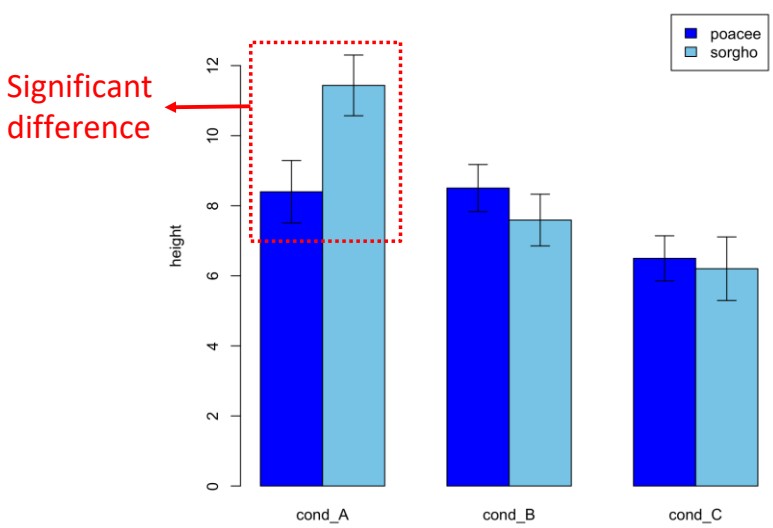
- Based on 1395 Census (A sample of 1,048,575 individuals)
  - Total number of valid data with Persian calendar: 1,000,222



Grouped Bar Chart

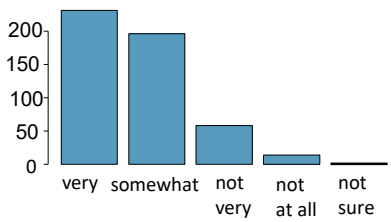


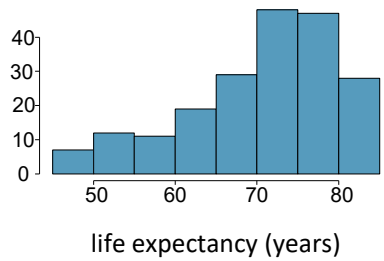
Bar Plot + Error Bar



Bar Plots vs. Histograms

- Barplots for categorical variables, but histograms for numerical variables.
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable.





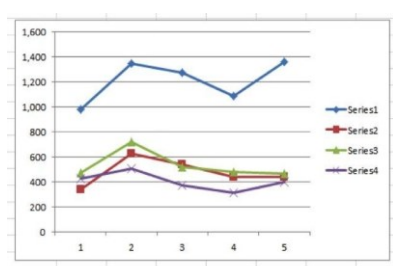
Bar Plots vs. Line Charts

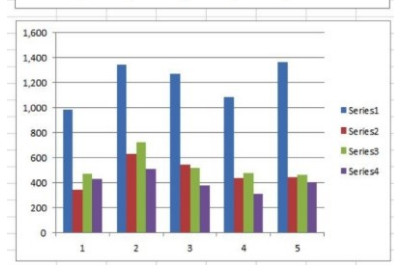
Continuous values

[redacted]

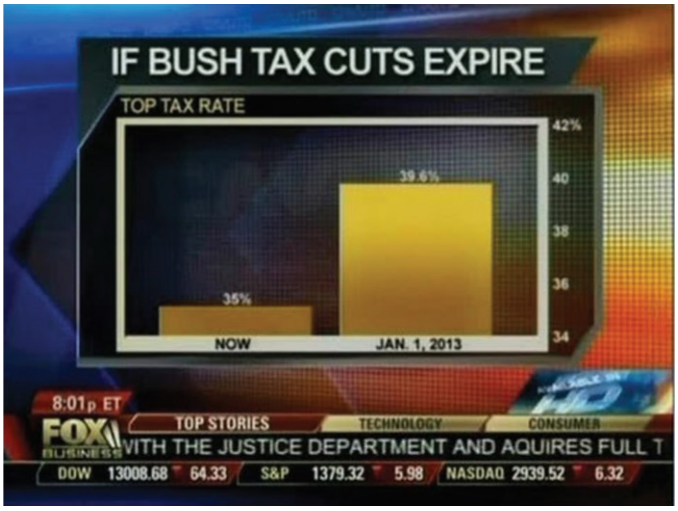
Discrete values

[redacted]





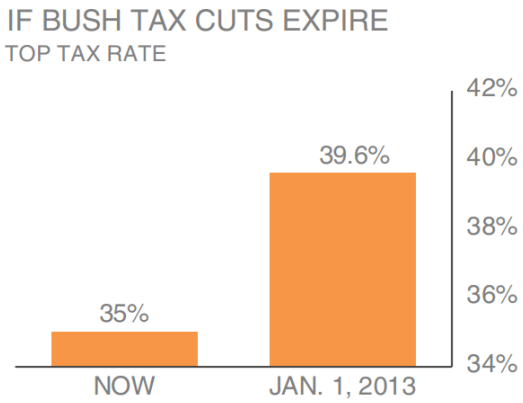
Bar Plot Abuse



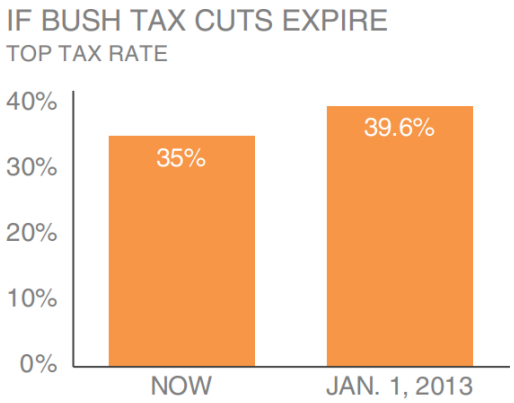
41

Bar Plot Abuse

Non-zero baseline: as originally graphed



Zero baseline: as it should be graphed



42

Contingency Table

		Income				
		< $40K	$40-80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

- A table that summarizes data for two categorical variables is called a contingency table.

43

Relative Frequency

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

< $40K: 128/202 = 63% find it very difficult to save

$40K-$80K: 63/148 = 43%

$80K: 31/124 = 25%

Refused: 9/26 = 35%

→ feelings about difficulty of saving money and income are associated (dependent)

44

Heatmap

- A heatmap is a way to visualize data in tabular format, where in place of (or in addition to) the numbers, you leverage colored cells that convey the relative magnitude of the numbers.

Table

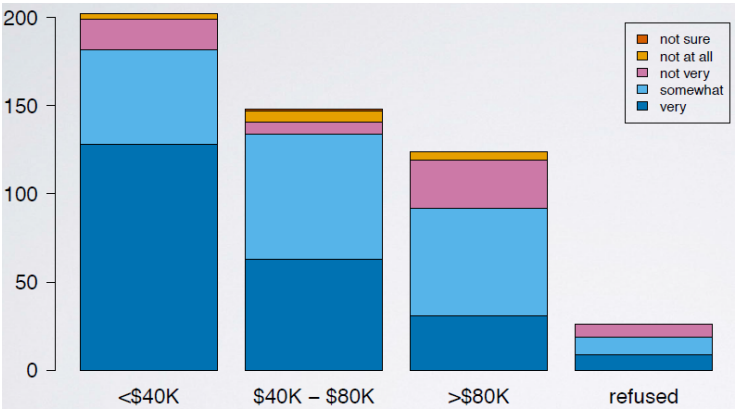
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

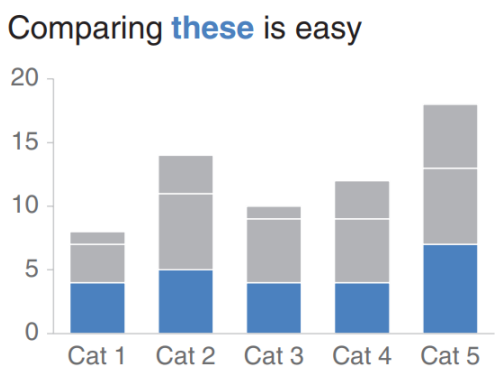
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

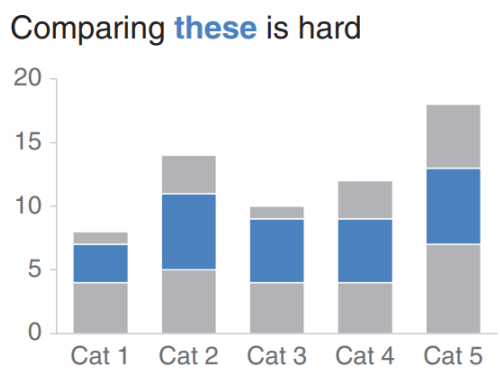
Segmented (Stacked) Bar Plot



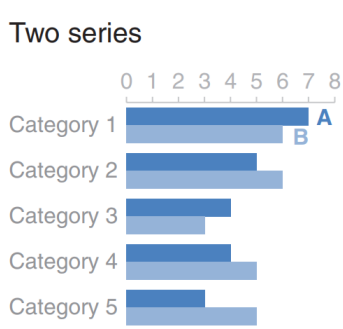
- Useful for visualizing conditional frequency distributions
- Compare relative frequencies to explore the relationship between the variables

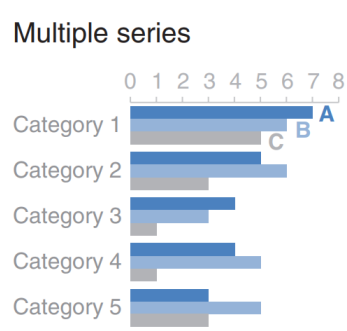
Stacked Bar Plot



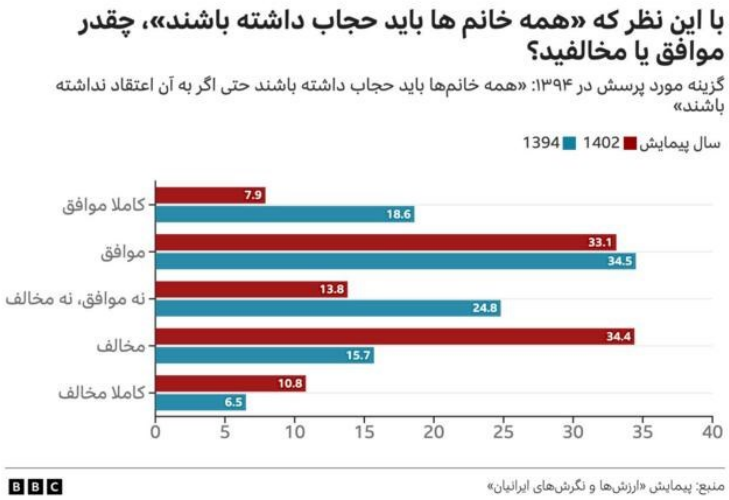


Horizontal Bar Plot

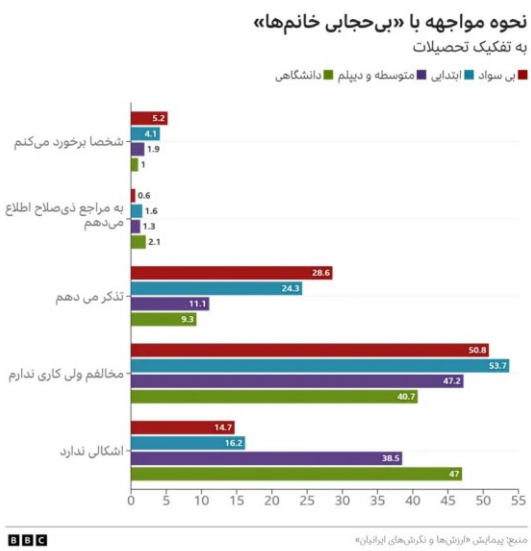




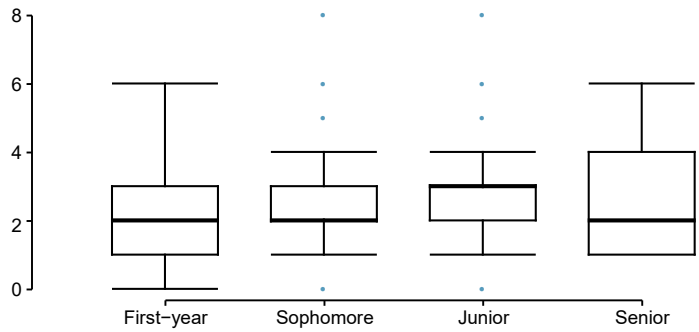
Relative Frequency Bar Plot



A Common Mistake in Comparing Relative Frequencies

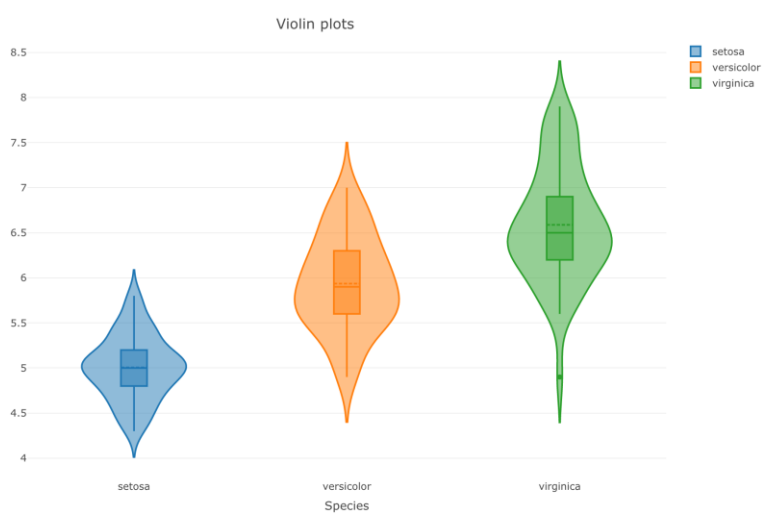


Side-by-side box plots



- Does there appear to be a relationship between class year and number of societies students are in?

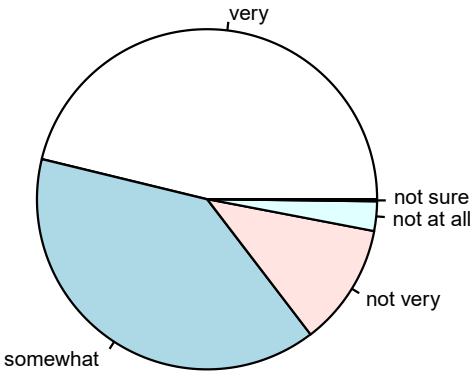
Side-by-side Violin Plot

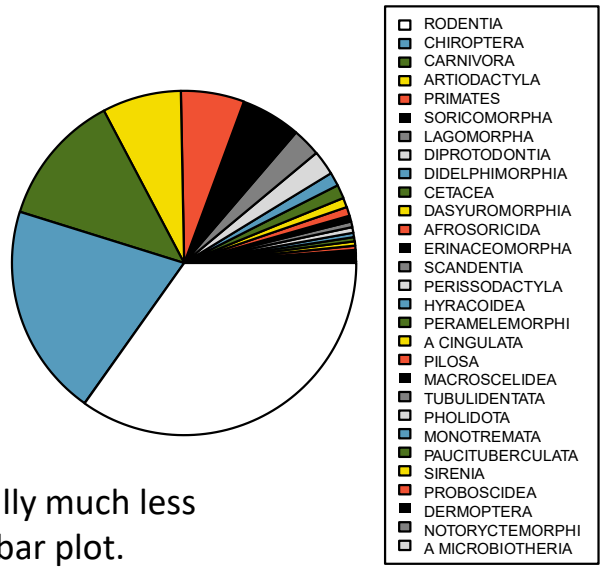


Violin Plots for Comparison

To Be Avoided

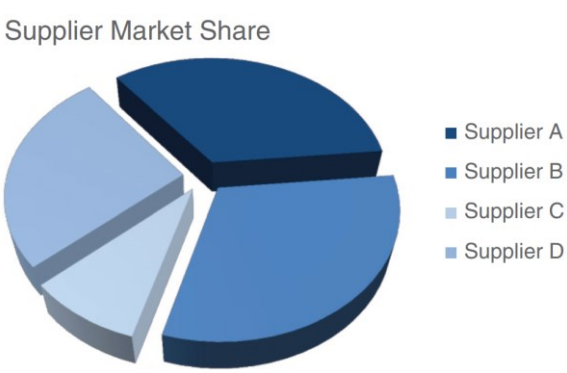
~~Pie Chart?~~ NO!

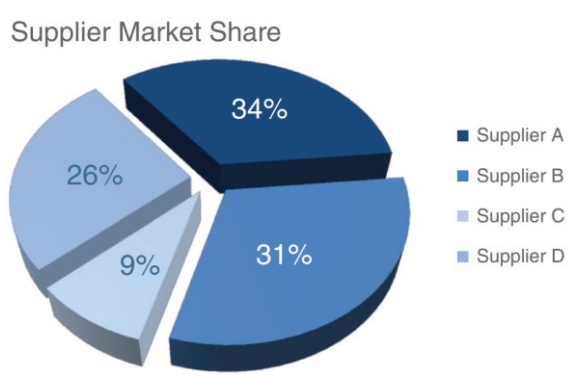




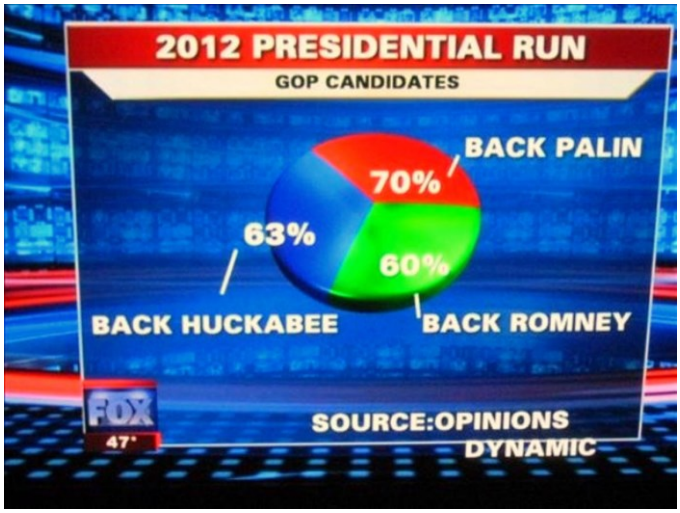
- A pie chart is actually much less informative than a bar plot.

3D Pie Charts

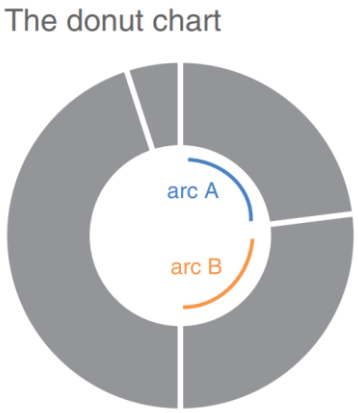




Terrible Pie Chart

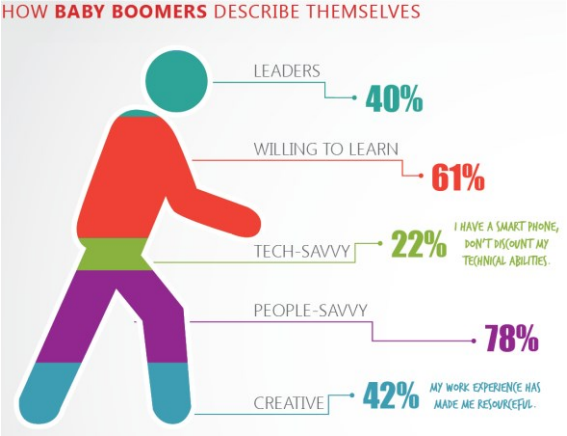


Donut Chart

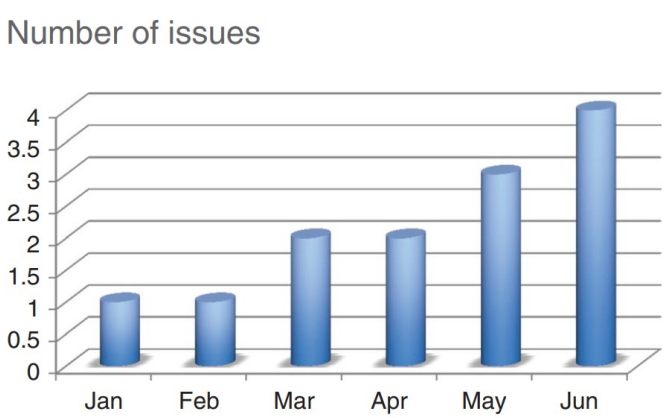


Area Graphs

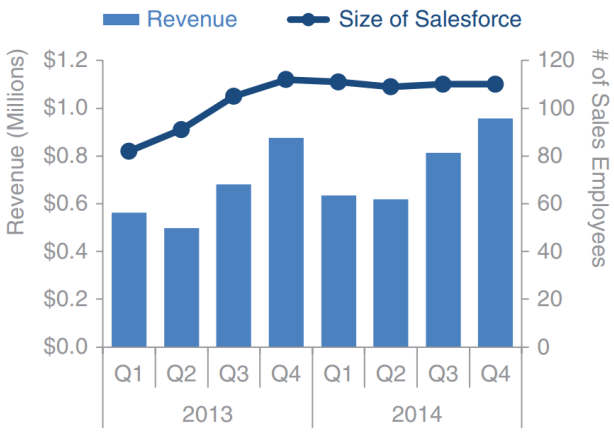
- Humans' eyes don't do a great job of attributing quantitative value to two-dimensional space.



Never use 3D



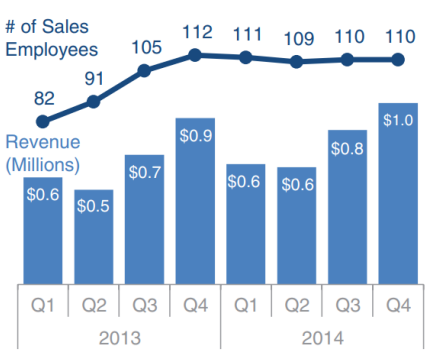
Secondary *y*-axis



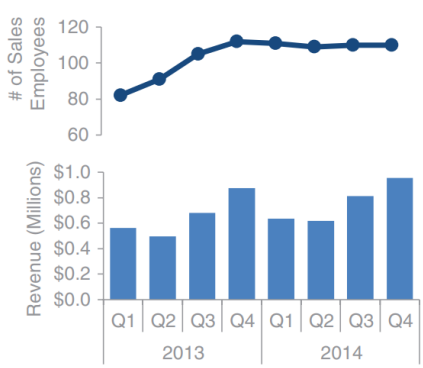
61

Alternatives for Secondary *y*-axis

Alternative 1: **label directly**

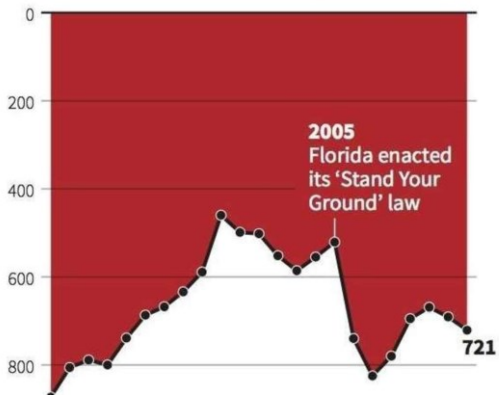


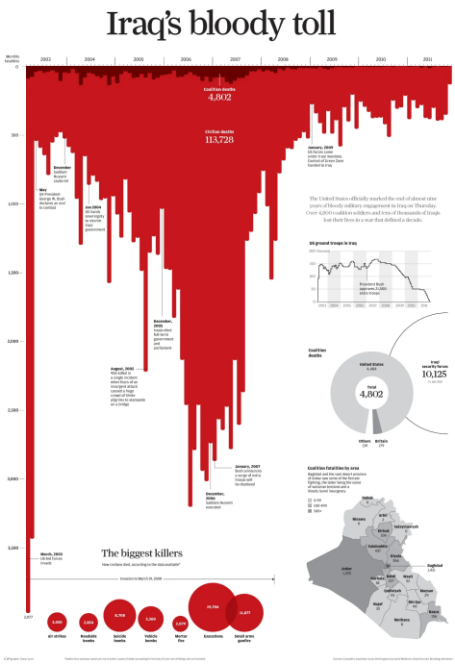
Alternative 2: **pull apart vertically**



62

Inverse Charts





Cumulative Charts

Cumulative Charts