Exploratory Data Analysis, Stats, and Visualization

Jay Urbain, PhD

Data Exploration Topics

- Sampling
- Distributions
- Aggregate statistics
- Principles of visualization

Data Science Process

- Ask questions – hypothesize
- Data collection
- **Data exploration**
- Data modeling
- Data analysis
- Visualization and communication of results

Basics of sampling

Population versus sample:

- A population is the entire set of objects or events under study. Population can be hypothetical “all students” or “all students in this class.”
- A sample is a “representative” subset of the objects or events under study. Needed because its impossible or intractable to obtain or compute with population data.

Biases in samples:

- Selection bias: some subjects or records are more likely to be selected.
- Volunteer/nonresponse bias: subjects or records who are not easily available are not represented.

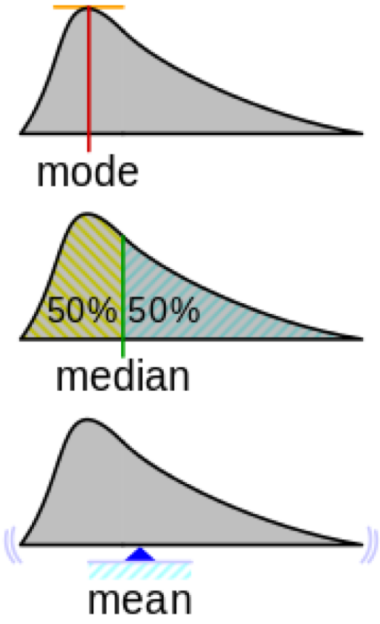
Examples?

Distributions

- Distribution: Set of all possible random variables
- Example distributions:
  - Flipping coins for heads and tails
    - binomial (two possible outcomes, e.g., heads or tails)
    - discrete (categories of students, no real numbers)
    - evenly weighted (each category of students is equally likely; heads have same probability of tails)
  - Tornado events in Milwaukee?
    - binomial
    - discrete
    - evenly weighted (note: tornadoes are rare events)

Central measures

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**



Sample mean

- The mean of a set of *n* observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.
- There is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample median

- The median of a set of *n* number of observations in a sample, ordered by value, of a variable is is defined by:

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example (already in order):

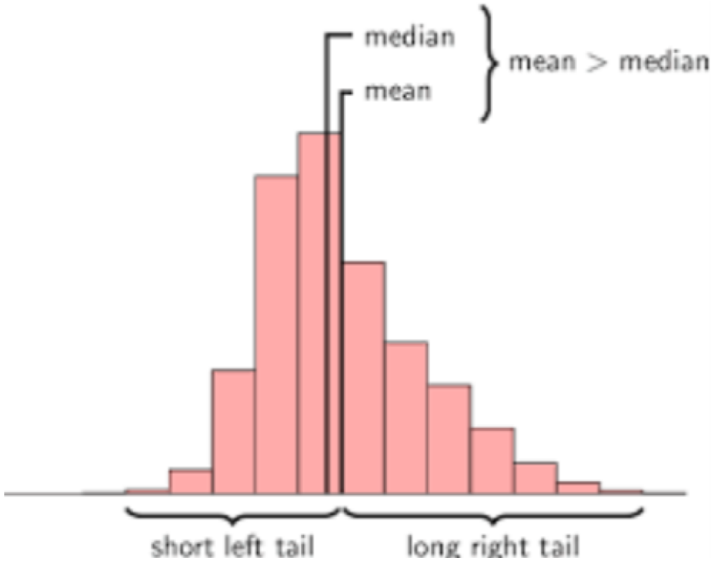
- Ages: 17, 19, 21, 22, 23, 23, 23, 38
- Median = (22+23)/2 = 22.5
- The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

Mean vs. Median

- The mean is sensitive to extreme values (outliers)

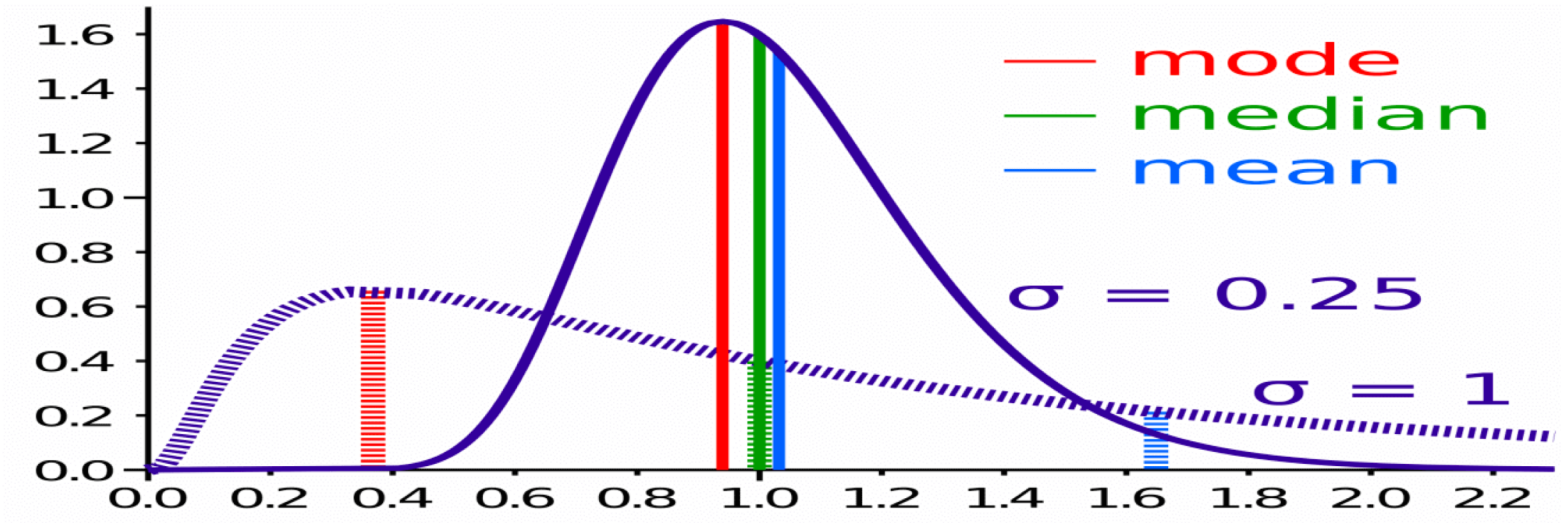
Mean vs. Median

- The mean is sensitive to extreme values (outliers)



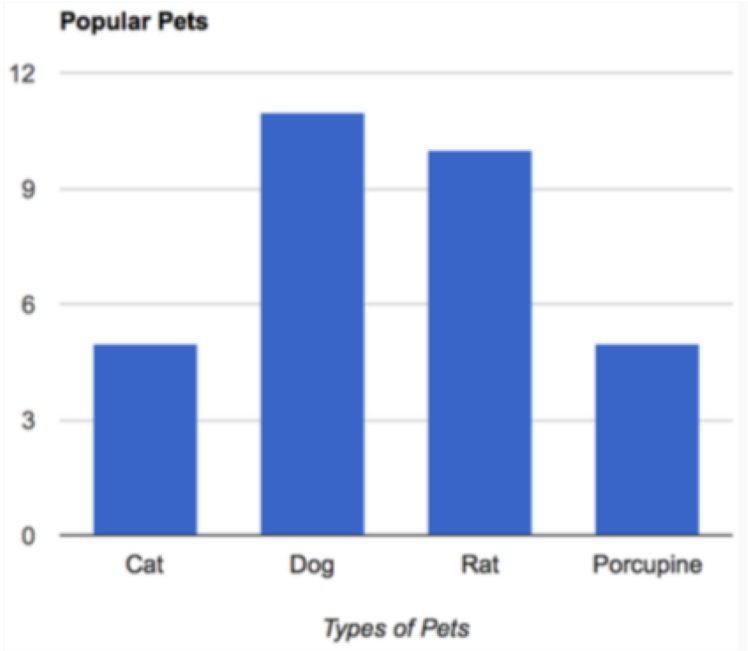
- The above distribution is called right-skewed since the mean > median.

Mode, median, mean



Central measures: categorical variables

- For categorical variables, neither mean or median make sense.
- *Why?*



- The mode might be the way to find the most representative value.

Central measures: computation

Computational time complexity:

- Mean: *O(n)*
- Median: *O(n*lg(n))* # with comparison sort
- Mode: *O(n)* to *O(n*lg(n))* # dependent on data structure

Measures of Spread: Range

- The spread of a sample of observations measures how well the mean or median describes the sample.
- One way to measure spread of a sample of observations is via the range.

Range = Maximum Value - Minimum Value

Measures of Spread: Variance

- The (sample) variance, denoted s^2, measures how much on average the sample values deviate from the mean:

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

- Note: the term $|x_i - \bar{x}|$ measures the amount by which each x_i deviates from the mean $\bar{x}$. Squaring these deviations means that s^2 is sensitive to extreme values (outliers).
- Note: s^2 doesn't have the same units as the x_i. What does a variance of 1,008 or 0.0001?

Measures of Spread: Standard Deviation

- The (sample) standard deviation, denoted *s*, is the square root of
- the variance:

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

- Note: *s* does have the same units as the x_i .

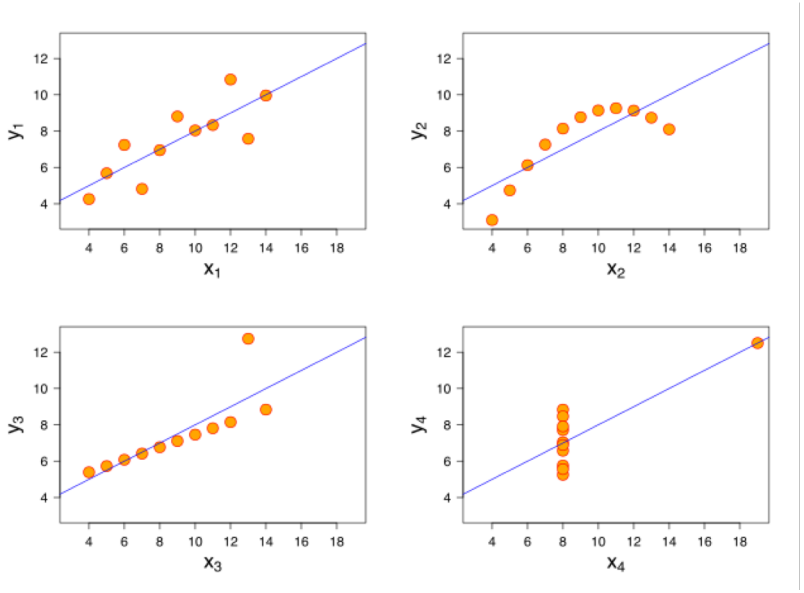
Anscombe's Data

- The following four data sets comprise the *Anscombes Quartet.*
- All four sets of data have almost identical simple summary statistics.

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

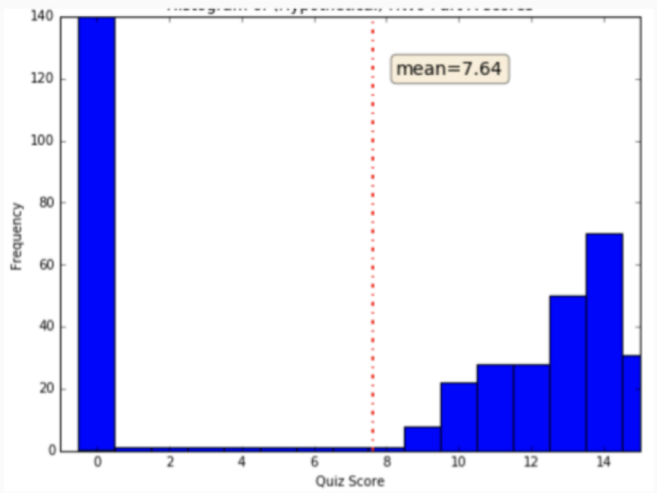
Anscombe's Data

- Summary statistics clearly don't tell the story of how they differ.
- But a picture often can:



Visualization Motivation

- The average score for the quiz is 7.37/15. What does this suggest?



- What does the graph suggest?

Principals of Visualization

Visualizations help us analyze and explore the data. They help:

- Identify hidden patterns and trends
- Formulate/test hypotheses
- Communicate:
  - Present information and ideas succinctly
  - Provide evidence and support
  - Influence and persuade
- Determine the next step in analysis/modeling

More visualization motivation

Visualizations help us to analyze and explore the data:

- Identify hidden patterns and trends
- Formulate/test hypotheses
- Communicate modeling results
- Present information and ideas succinctly
- Provide evidence and support
- Influence and persuade
- Determine the next step in analysis/modeling

Principals of Visualization

Some basic data visualization guidelines from *Edward Tufte*:

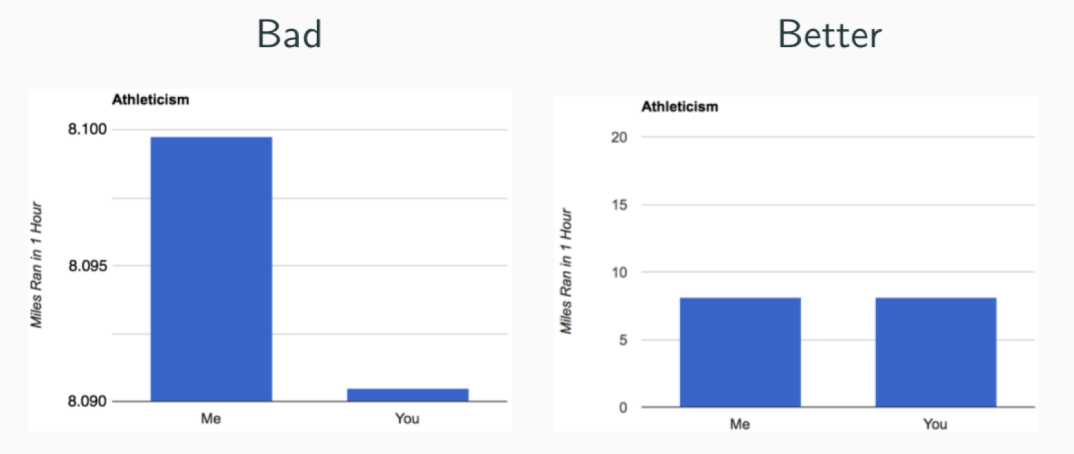
1. Maximize data to ink ratio: show the data

Principals of Visualization

Some basic data visualization guidelines from *Edward Tufte*:

1. Maximize data to ink ratio: show the data.
2. Don't lie with scale: minimize size of (Lie Factor) :

   (effect in graph)/(size of effect in data)



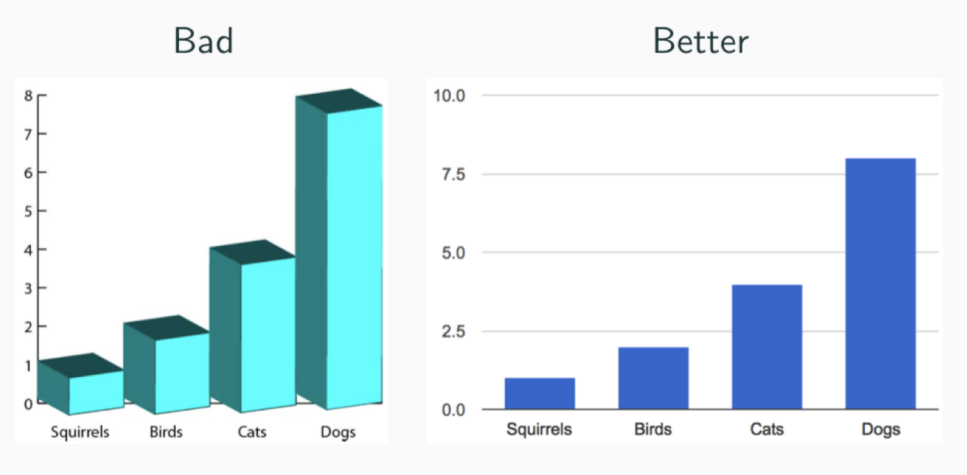
Principals of Visualization

Some basic data visualization guidelines from *Edward Tufte*:

1. Maximize data to ink ratio: show the data.
2. Don't lie with scale: minimize size of (Lie Factor) :

   (effect in graph)/(size of effect in data)

3. Minimize chart-junk: show data visualizations, not design variation



Principals of Visualization

Some basic data visualization guidelines from *Edward Tufte*:

1. Maximize data to ink ratio: show the data.
2. Don't lie with scale: minimize size of (Lie Factor) :

   (effect in graph)/(size of effect in data)

3. Minimize chart-junk: show data visualizations, not design variation
4. Clear, detailed and thorough labeling

More on Tufte principals of visualization:

http://www2.cs.uregina.ca/ rbm/cs100/notes/spreadsheets/tufte paper.html

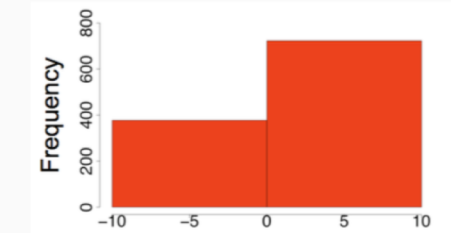
Types of Visualization

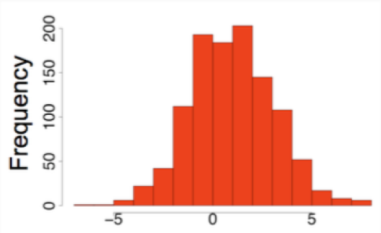
What do you want your visualization to show about your data?

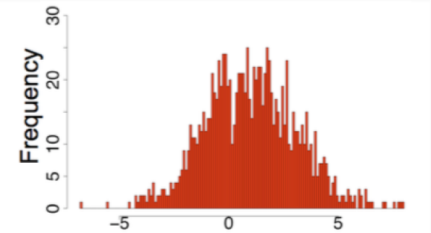
- **Distribution**: how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship**: how the values of multiple variables in the dataset relate.
- **Composition**: how the dataset breaks down into subgroups.
- **Comparison**: how trends in multiple variable or datasets compare.

Histograms to visualize distributions

- A histogram is a way to visualize how 1-dimensional data is distributed across certain values.
- Trends in histograms are sensitive to number of bins.

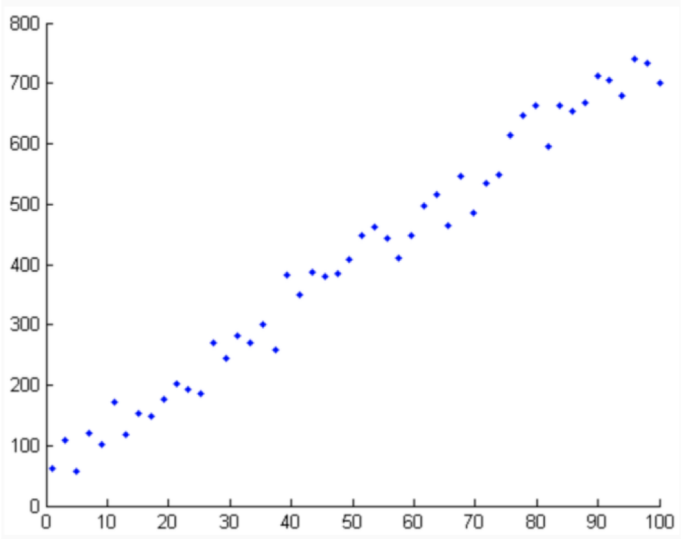






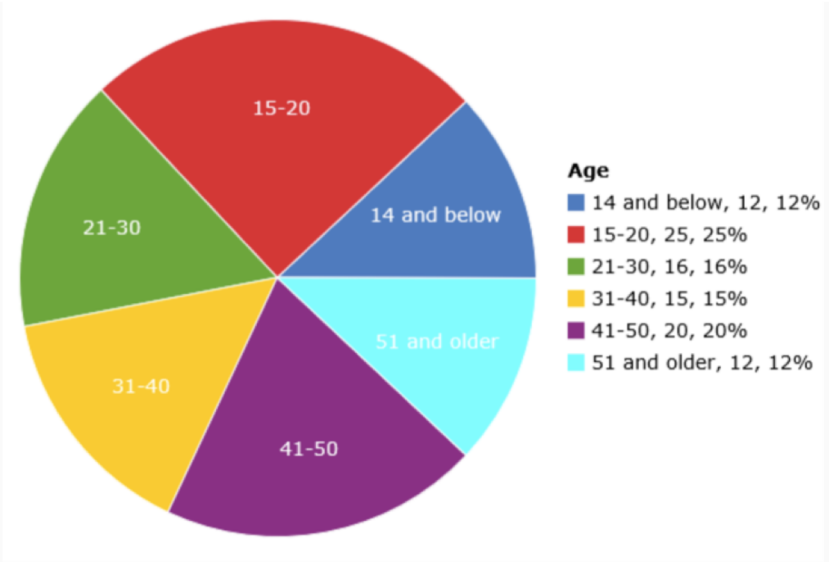
Scatter plots to visualize relationships

- A scatter plot is a way to visualize how multi-dimensional data are distributed across certain values.
- A scatter plot is also a way to visualize the relationship between two different attributes of multi-dimensional data.



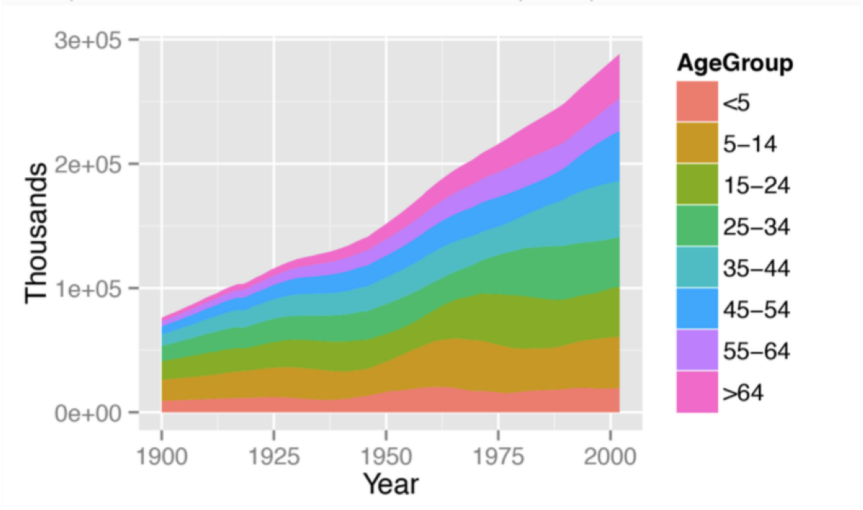
Pie chart for categorical variable

- A pie chart is a way to visualize the static composition (i.e., distribution) of a variable (or single group).
- Tufte says use bar chart.



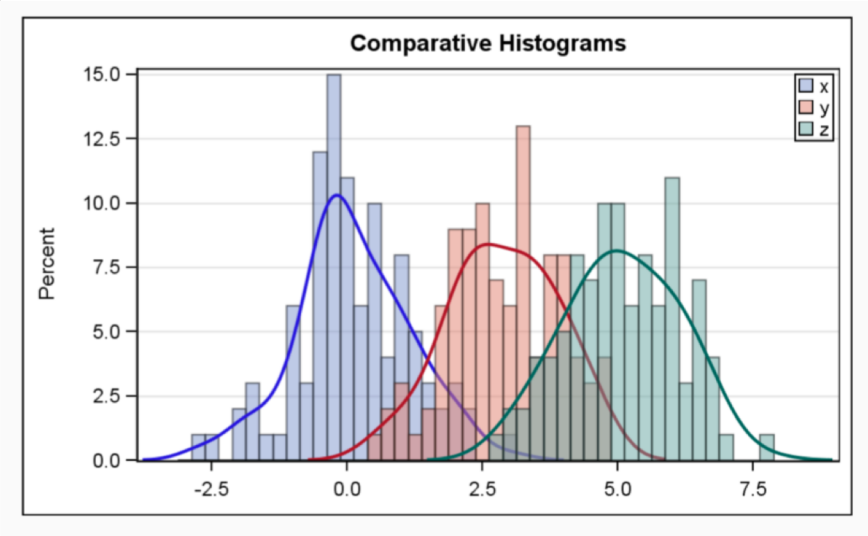
Stacked area graph to show trend over time

- A stacked area graph is a way to visualize the composition of a group as it changes over time (or some other quantitative variable).
- This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year).



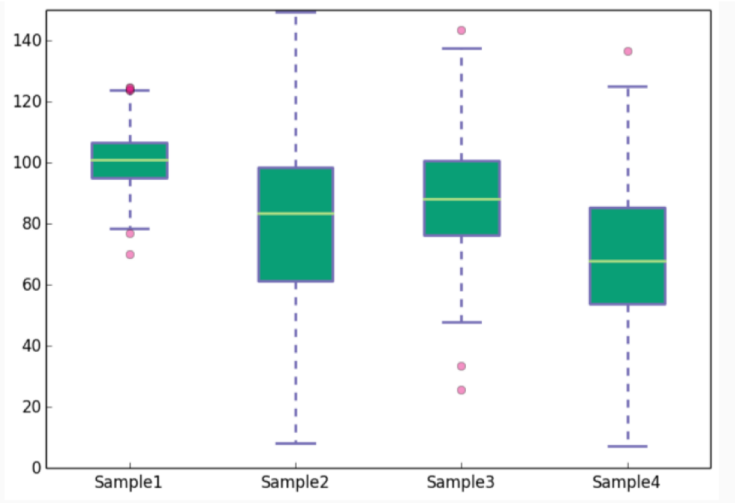
Multiple histograms

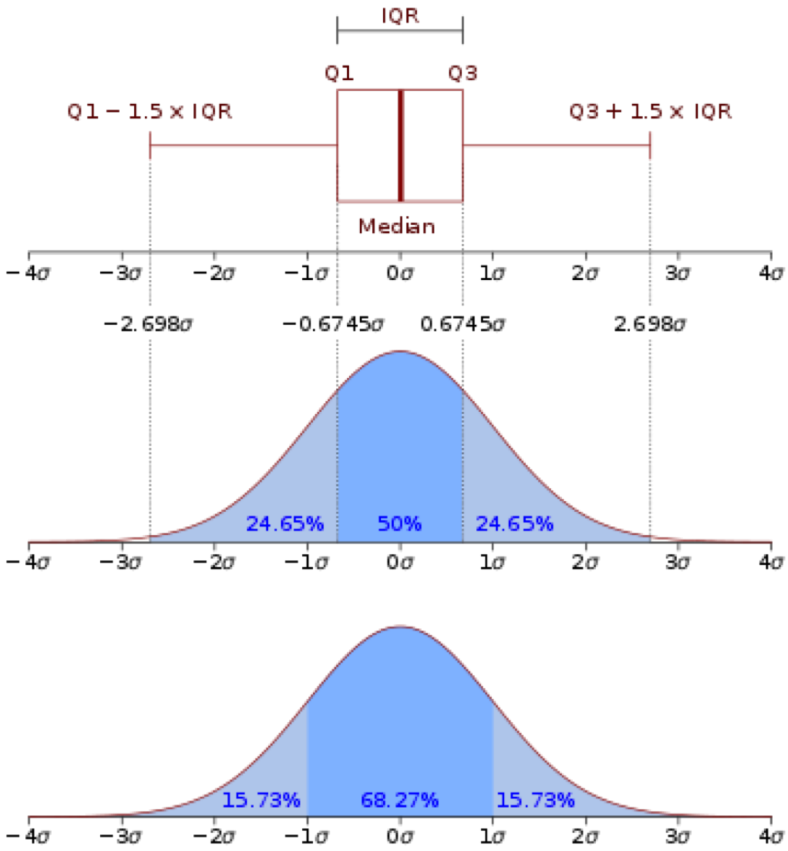
- Plotting multiple histograms and/or distribution curves on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups)



Boxplots

- A boxplot is a visualization to compare a quantitative variable across groups.
- It highlights the range, quartiles, median and any outliers present in a data set.





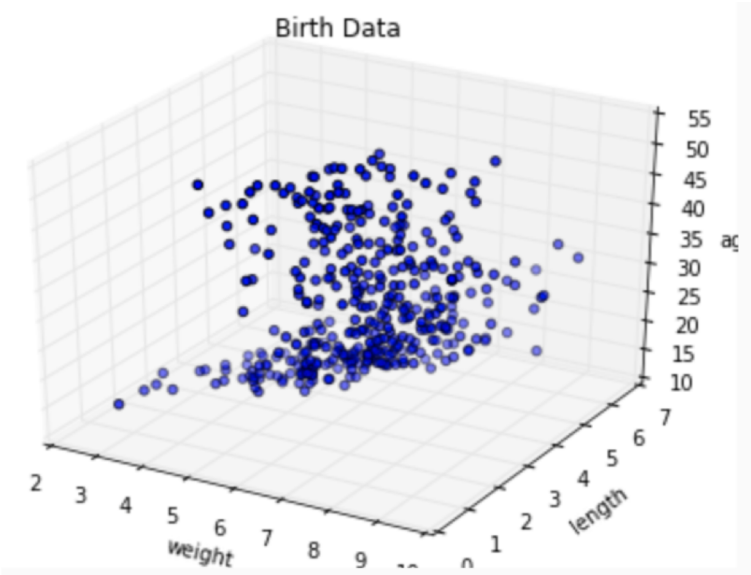
Data complexity

Often your dataset seems too complex to visualize:

- Data is too high dimensional
  - How do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like Cat or No?)

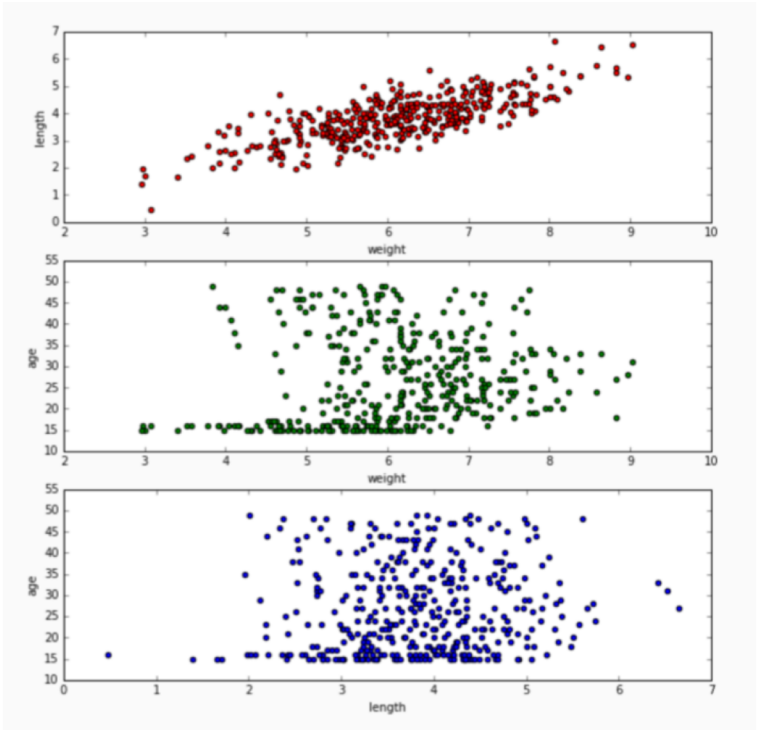
More dimensions not always better

- When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful.



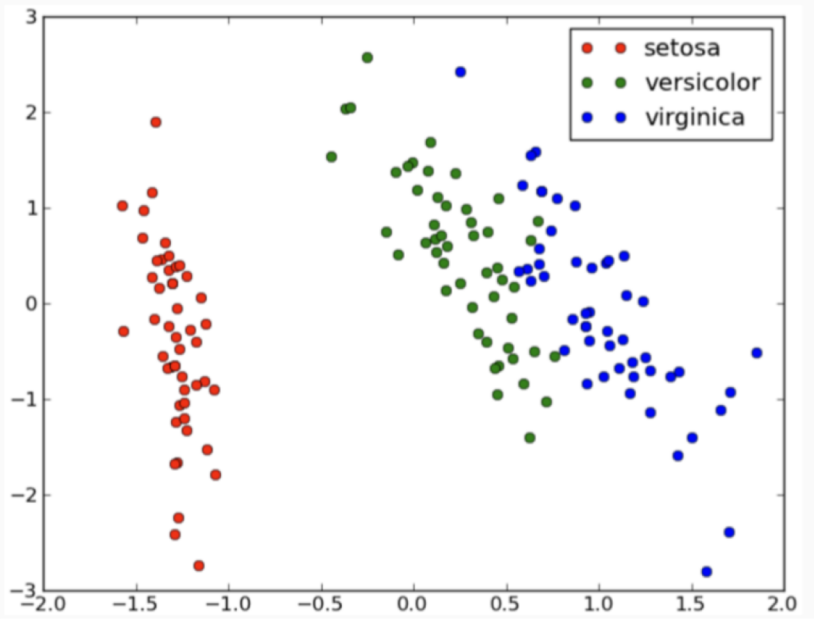
Reducing complexity

- Relationships may be easier to spot by producing multiple plots of lower dimensionality.



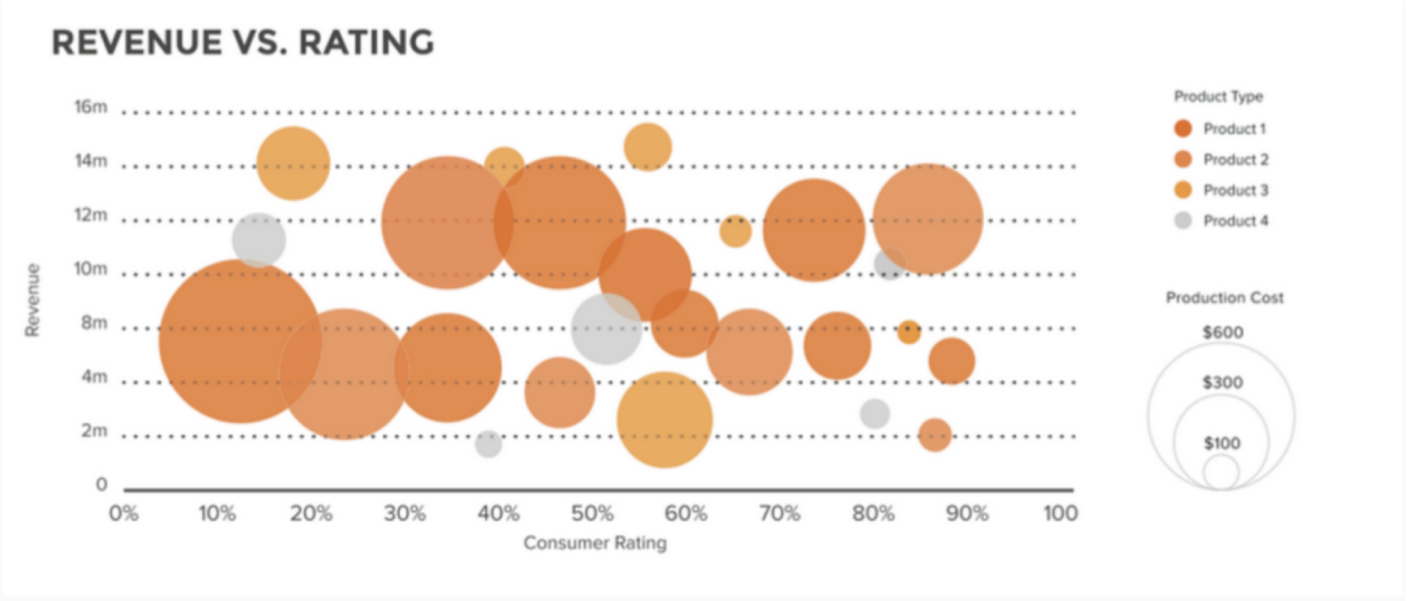
Adding a dimension

- For 3D data, color coding a categorical attribute can be effective.
- The above visualizes a set of Iris measurements. The variables are: petal length, sepal length, Iris type (setosa, versicolor, virginica).



3D

- For 3D data, a quantitative attribute can be encoded by size in a bubble chart.
- Visualization of a set of consumer products. The variables are: revenue, consumer rating, product type and product cost.



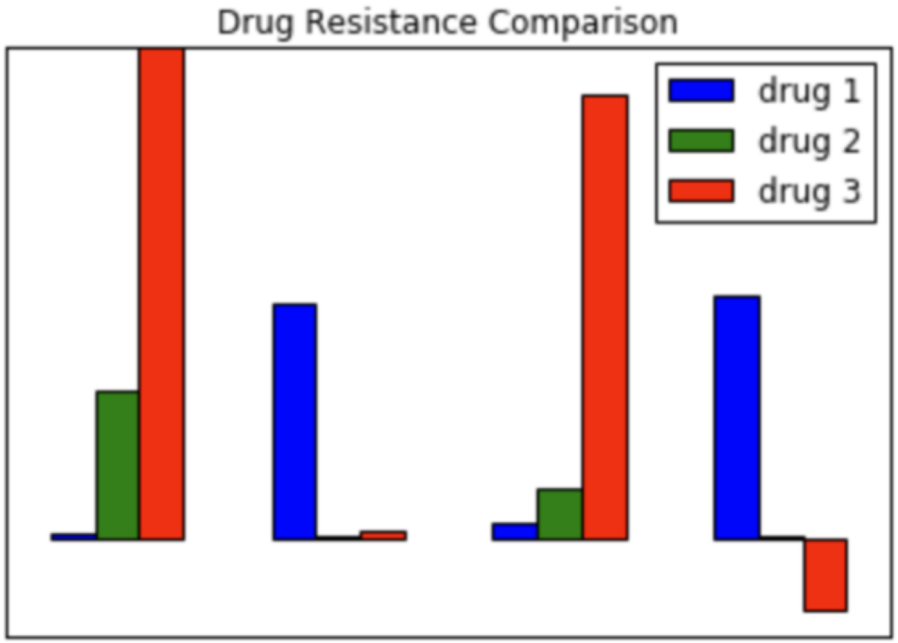
Example dataset

- Some visualizations for the following dataset:

Bacteria Name	Group No.	Res. to Drug 1	Res. to Drug 2	Res. to Drug 3
Brucella abortus	1	0.1	3	49
Diplococcus pneumoniae	2	4.75	0.007	0.125
Aerobacter aerogenes	1	0.3	1	47.2
Streptococcus viridans	2	4.9	0.03	-1.45

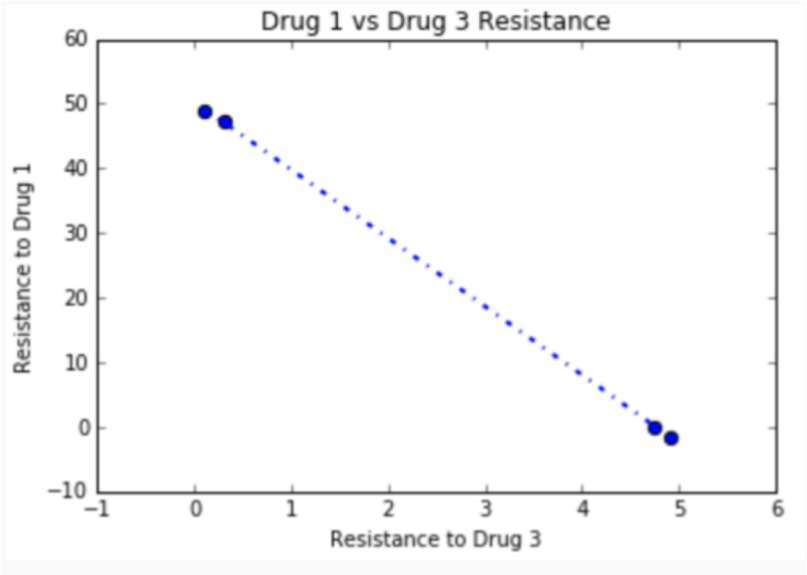
Example

- A bar graph showing resistance of each bacteria to each drug:
- Issues?



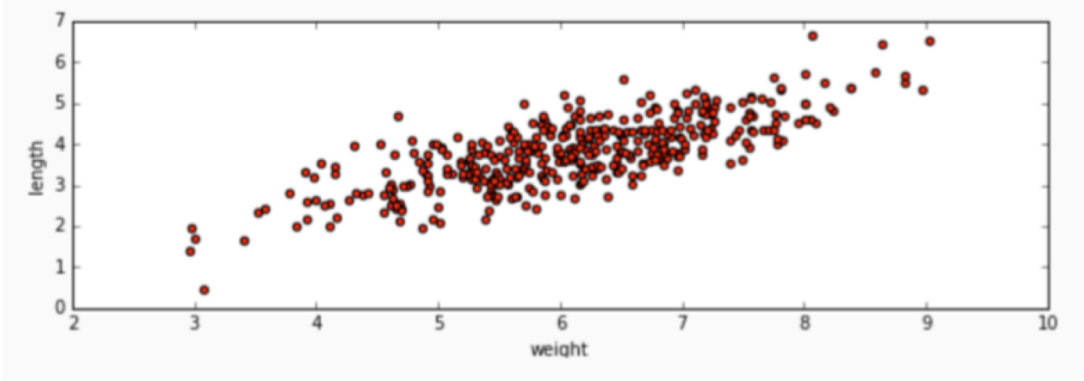
Example

- A scatter plot of drug #1 vs. drug #3 resistance.
- Key idea: the process of data exploration is iterative.



Quantifying Relationships

- We can see that birth weight is positively correlated with femur length.
- Can we quantify exactly how they are correlated? Can we predict birthweight based on femur length (or vice versa) through a statistical model? See notebook.



Prediction

- We can see that types of iris seem to be distinguished by petal and sepal lengths.
- Can we predict the type of iris given petal and sepal lengths through some sort of statistical model?