CS145: INTRODUCTION TO DATA MINING

2: Know Your Data

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

September 23, 2021

Content

- Vector/Tabular Data 
- Data Exploration: Descriptive Statistics
- Data Exploration: Data Visualization

Example

Living Area (sqft)	# of Beds	Has pool	Price (1000$)
2104	3	Yes	400
1600	3	No	330
2400	3	No	369
1416	2	No	232
3000	4	Yes	540

A matrix of n $\times$ p:

- n data objects / points
- p attributes / dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Attribute Type

- Numerical
  - E.g., height, income
- Categorical / discrete
  - E.g., Sex, Race

Categorical Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Content

- Vector/Tabular Data
- Data Exploration: Descriptive Statistics 
- Data Exploration: Data Visualization

Population vs. Sample

- A set of data points is a sample from a population:
  - A **population** is the entire set of objects or events under study.
    - E.g., population can be hypothetical "all students" or all students in this class.
    - E.g., population can be all the houses in a region
  - A **sample** is a "representative" subset of the objects or events under study. Needed because it's impossible or intractable to obtain or compute with population data.

Basic Statistical Descriptions of Data

- Central Tendency
- Dispersion of the Data

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:

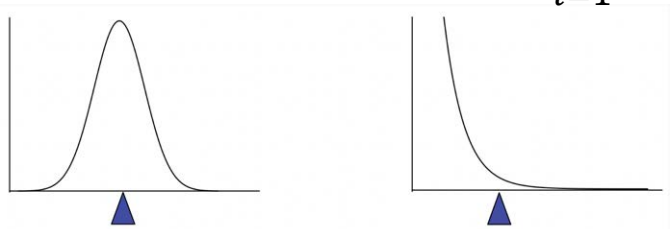
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

Sample mean

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

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



- The mean describes what a "typical" sample value looks like, or where is the "center" of the distribution of the data.
- Key theme: 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

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

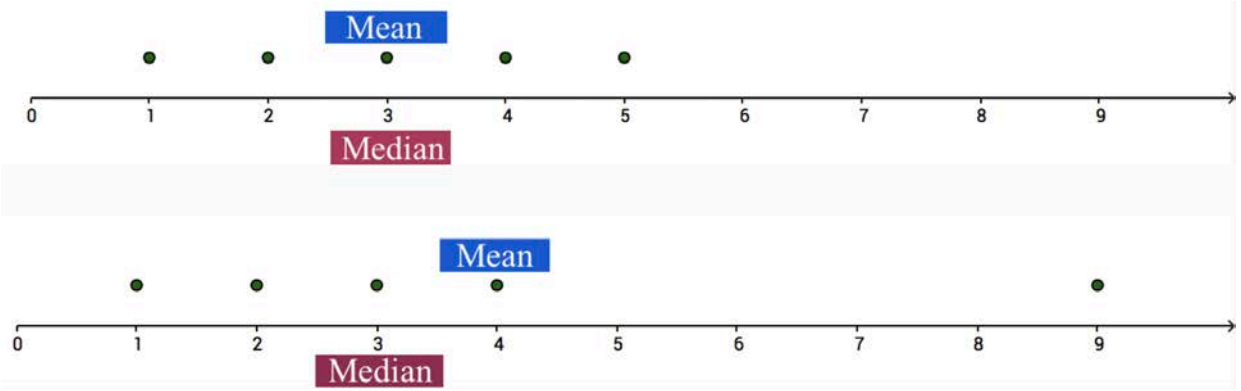
- Example (already in order):

  Ages: 17, 19, 21, 22, 23, 23, 23, 38

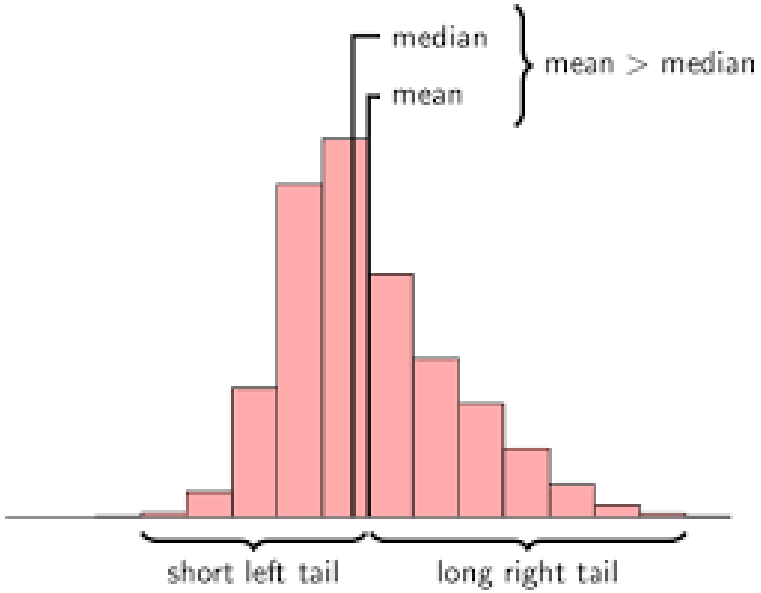
- Median = (22+23)/2 = 22.5
- The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

Mean vs. Median

- The mean is sensitive to extreme values (**outliers**)



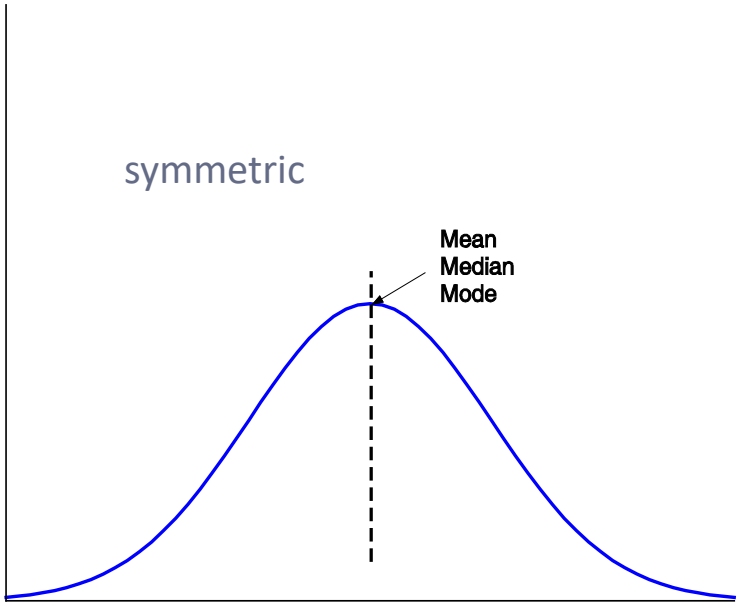
Mean, median, and skewness

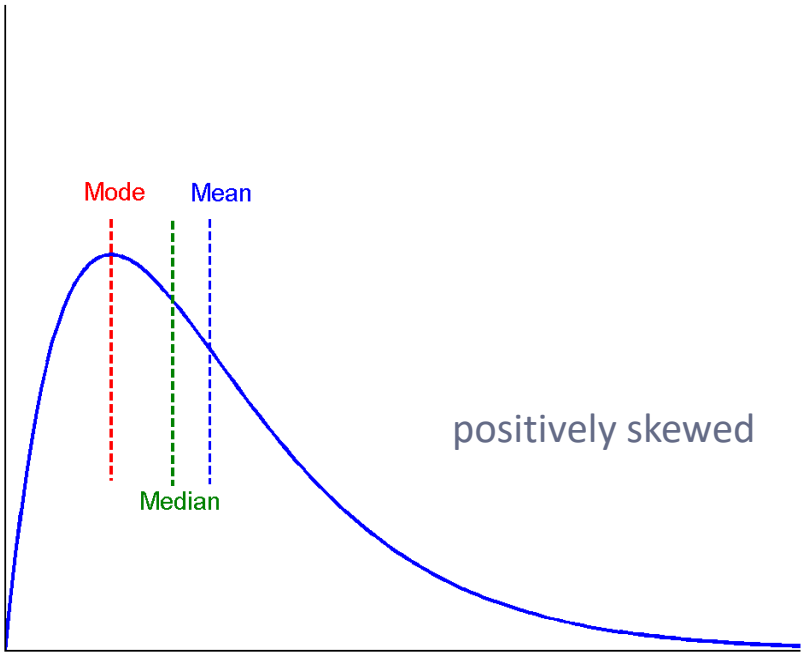


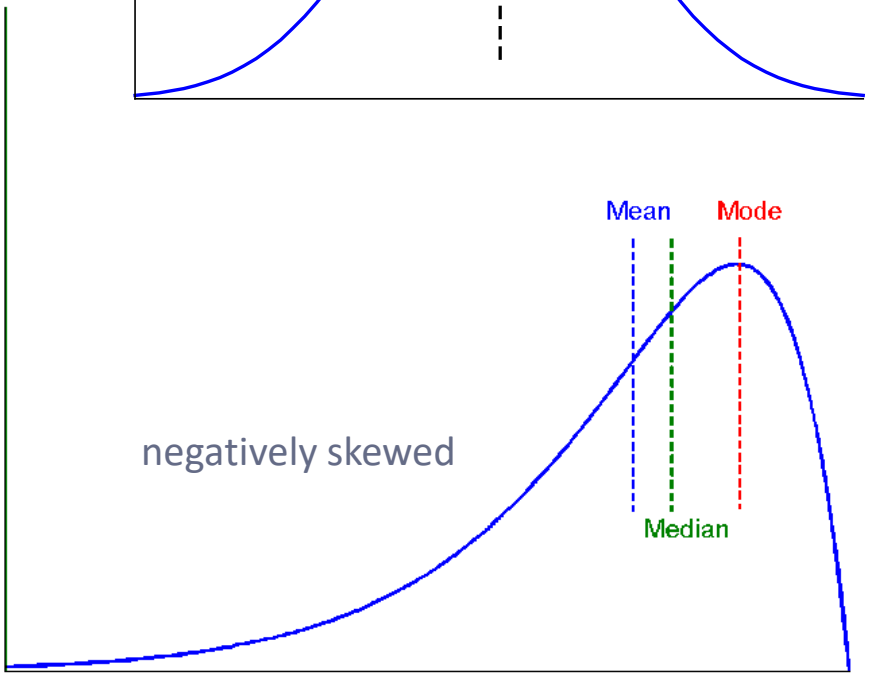
- The above distribution is called **right-skewed** since the mean is greater than the median. Note: **skewness** often "follows the longer tail".

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data





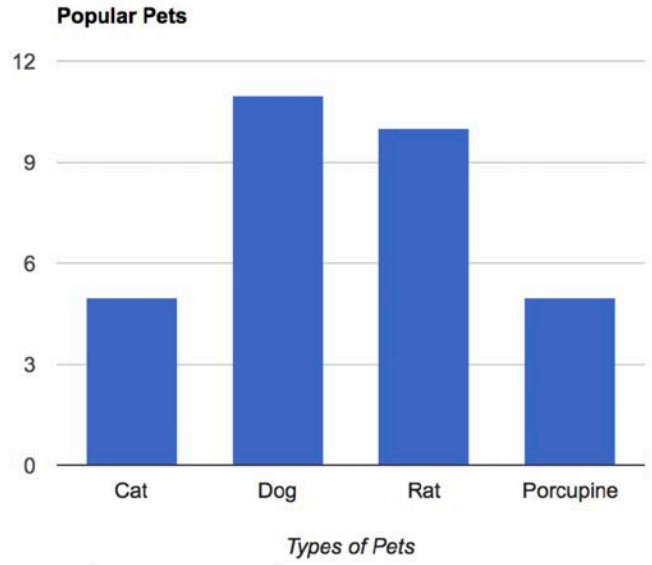


Question

- Is income positively or negatively skewed?

Regarding Categorical Variables…

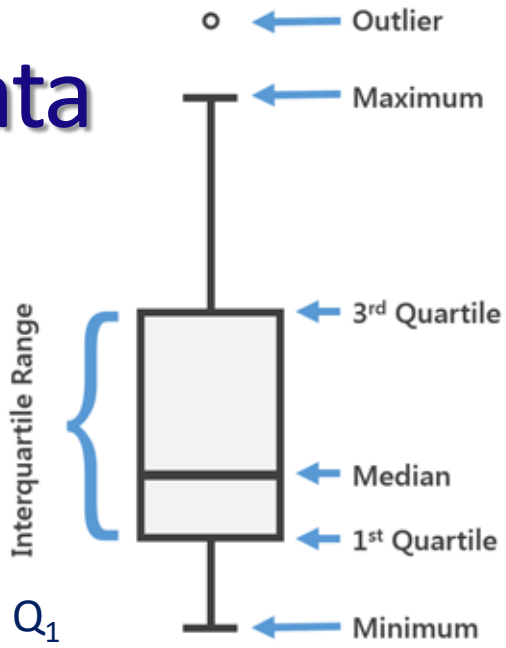
- For categorical variables, neither mean or median make sense. Why?



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

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR of Q_3 or Q_1



Box Plot

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
    - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$
    - $\sigma^2 = E[(X - E(X))^2] = E(X^2) - (E(X))^2$
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* $\sigma^{2)}$

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 mean? 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. Phew!

Content

- Vector/Tabular Data
- Data Exploration: Descriptive Statistics
- Data Exploration: Data Visualization 

Anscombe's Data

- The following four data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

G. E. M. Anscombe
FBA

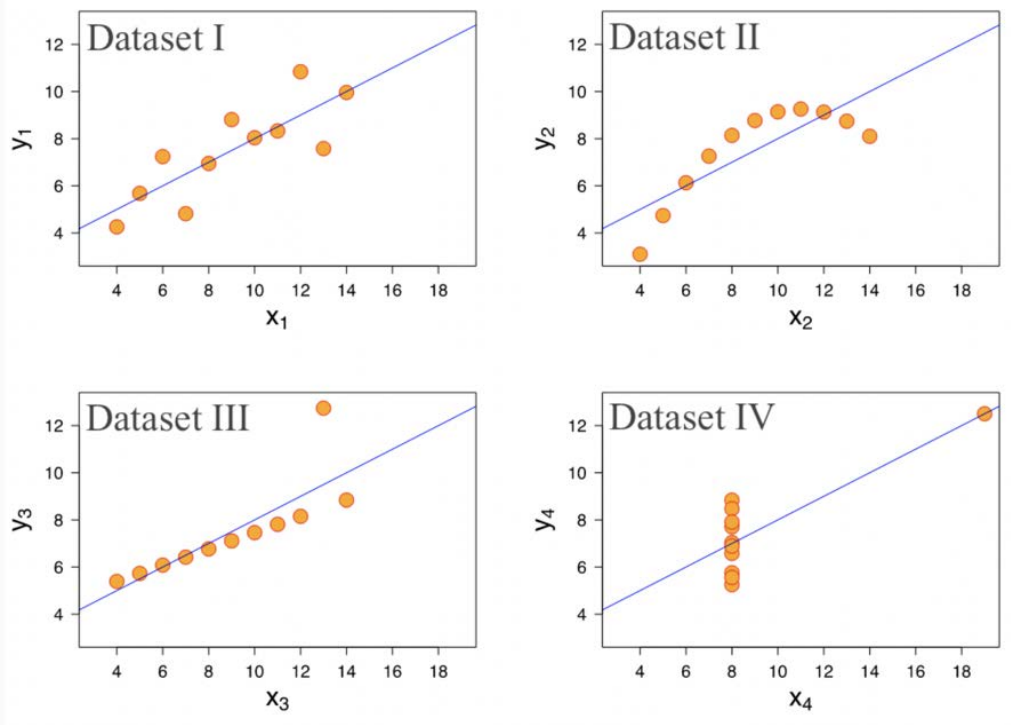
Anscombe as a young woman

Born	Gertrude Elizabeth Margaret Anscombe 18 March 1919 Limerick, Ireland
Died	5 January 2001 (aged 81) Cambridge, England

	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

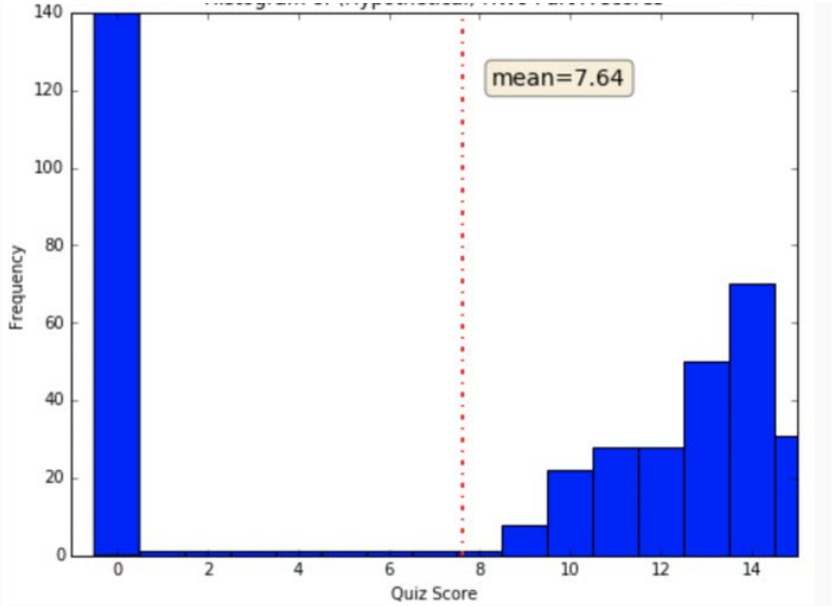
Anscombe's Data (cont.)

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



More Visualization Motivation

- If I tell you that the average score for a Homework is: 7.64/15 = 50.9%, what does that suggest?



- And what does the graph suggest?

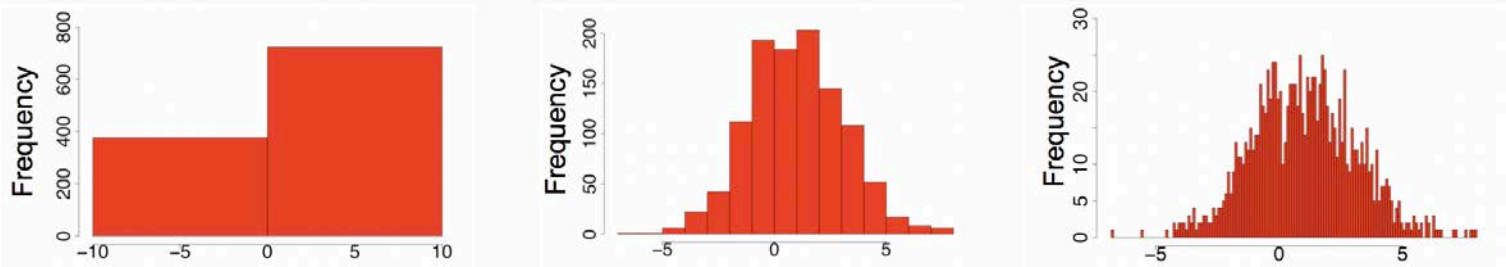
Types of Visualizations

- 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 distribution

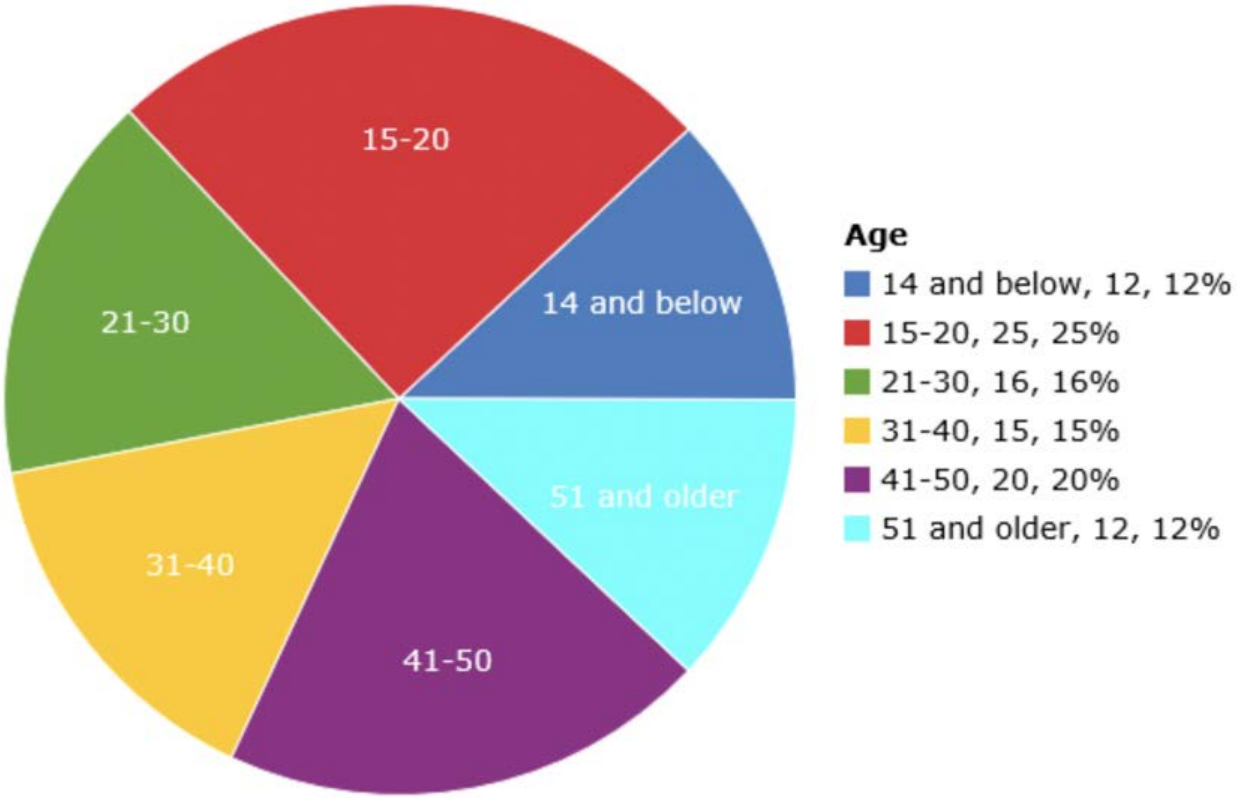
- A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.



- Note: Trends in histograms are sensitive to number of bins.

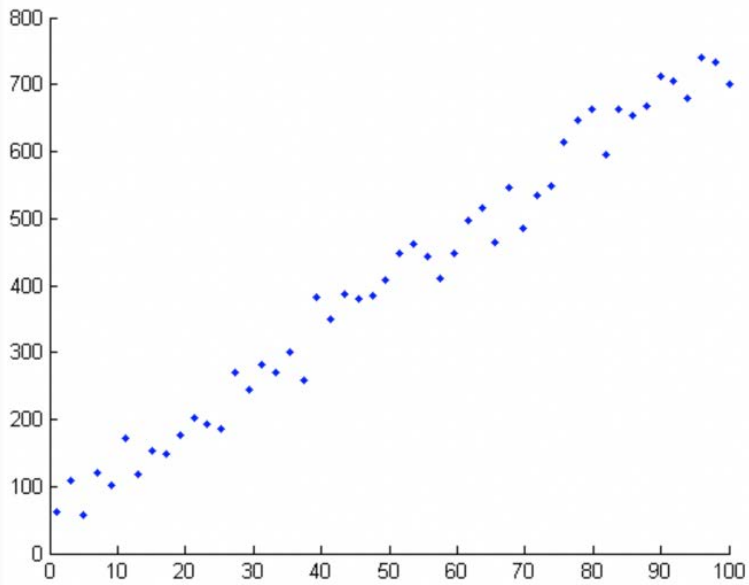
Pie chart for a categorical variable

- A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).



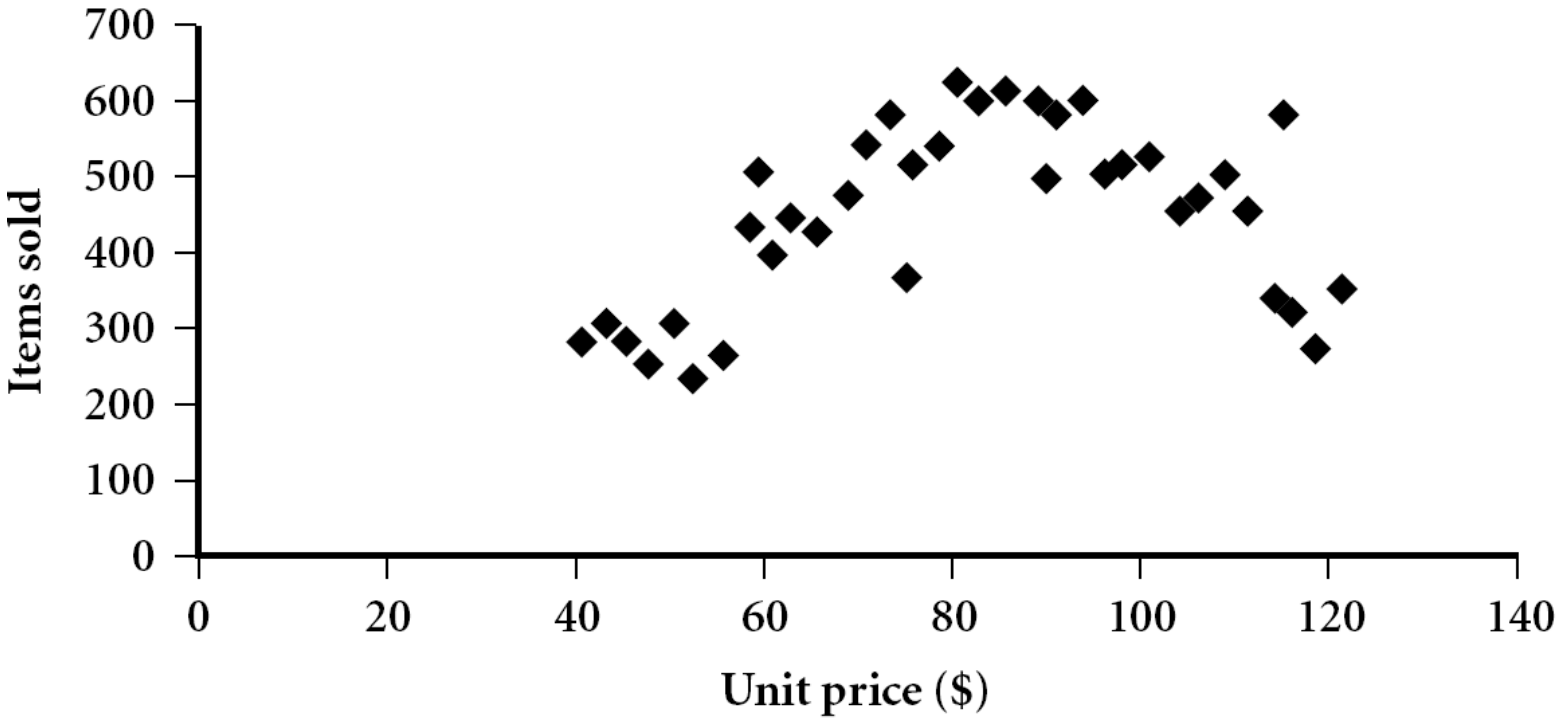
Scatter plots to visualize relationships

- A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.

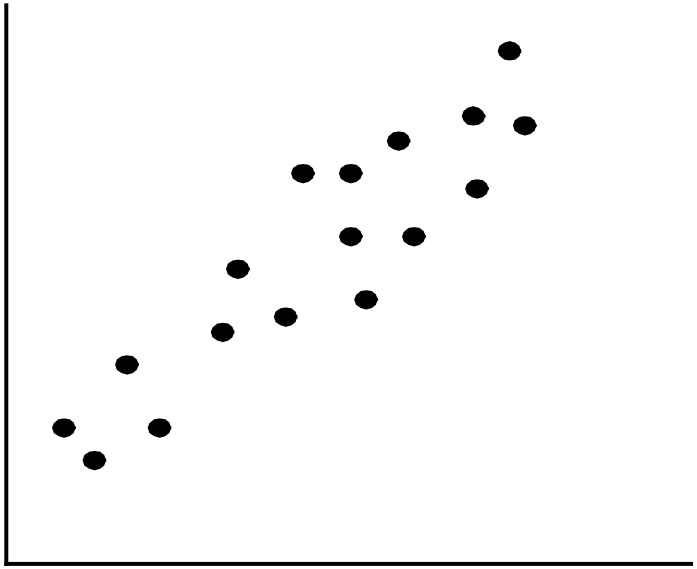


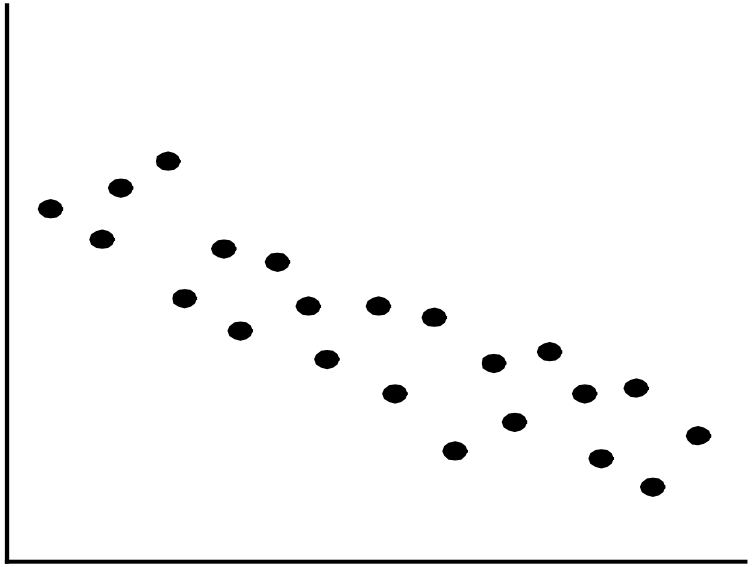
Scatter plot

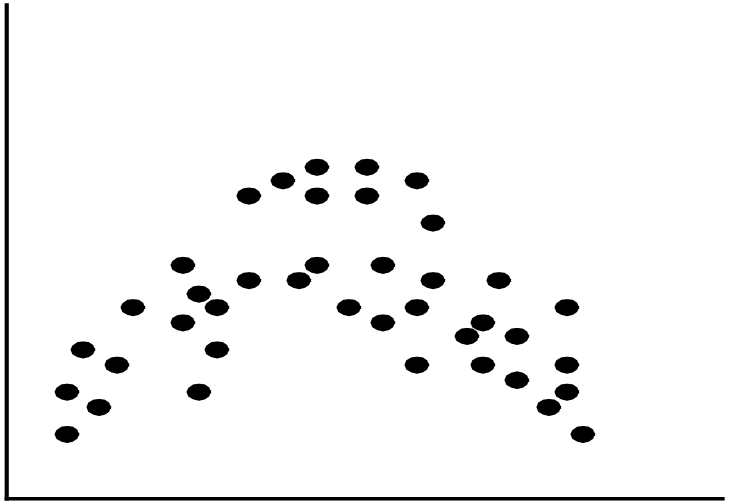
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

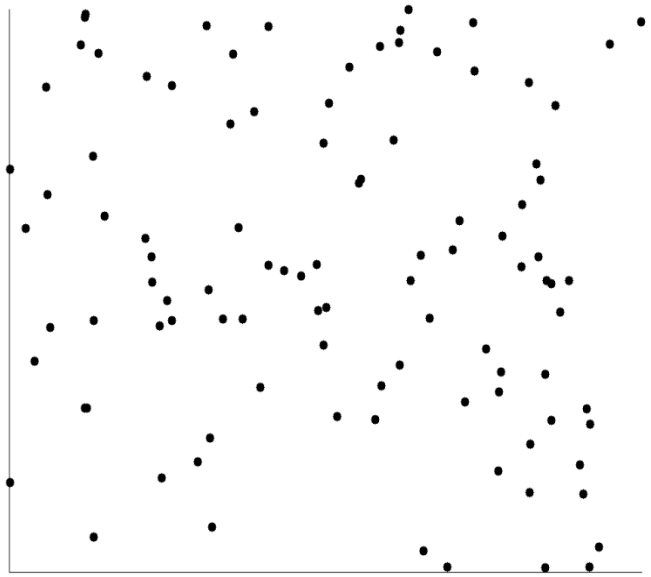


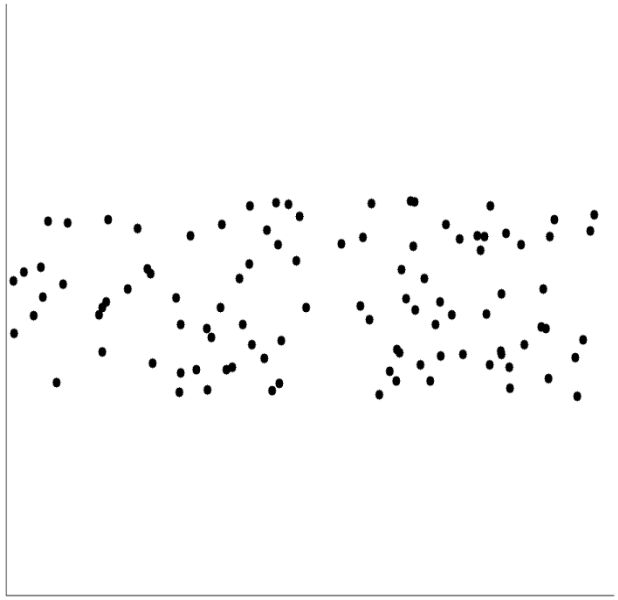


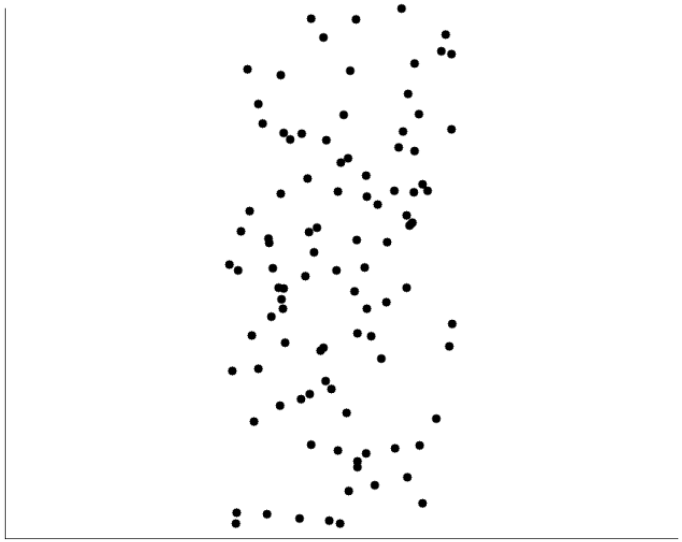


- The left half fragment is positively correlated
- The right half is negative correlated

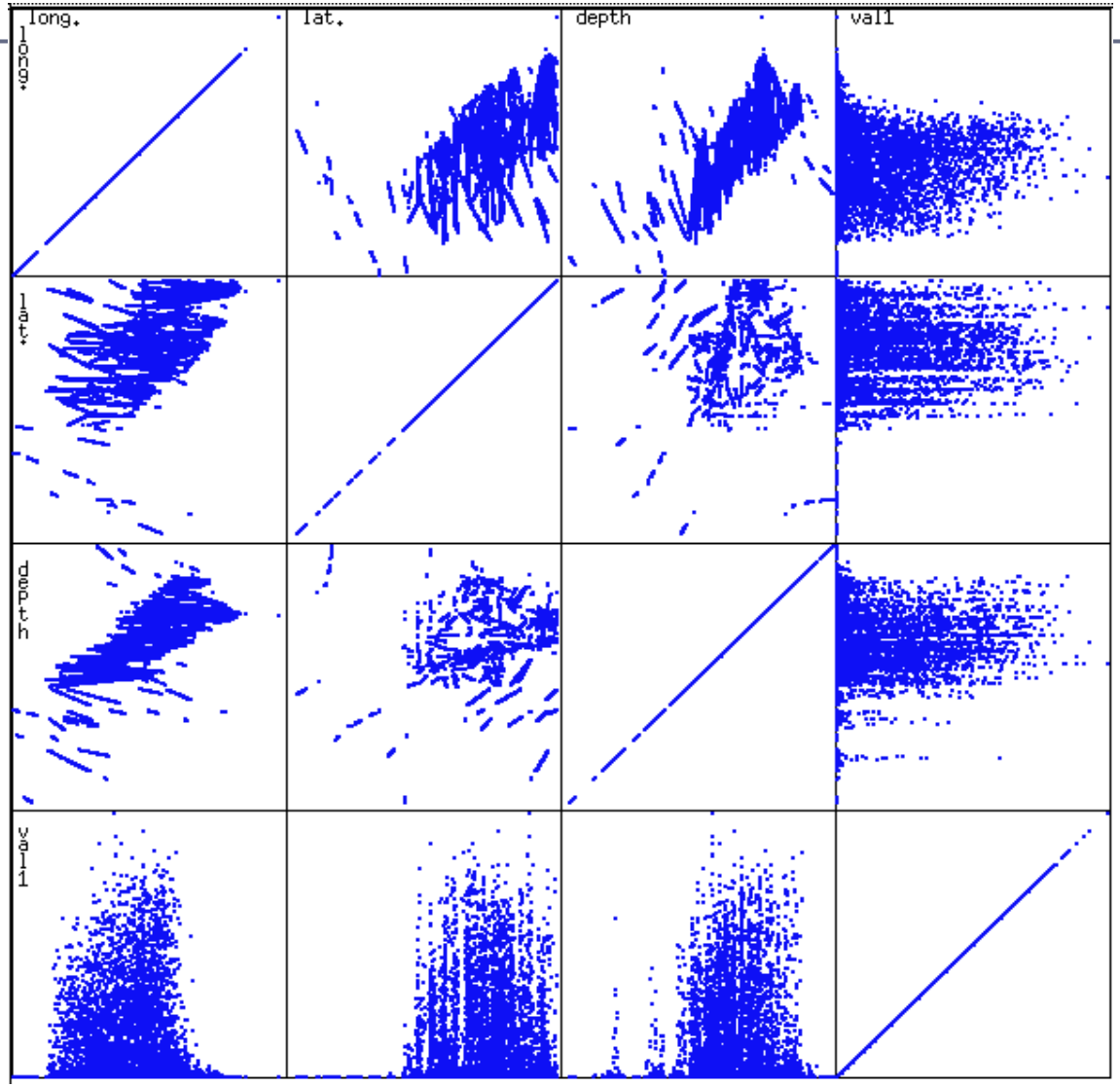
Uncorrelated Data







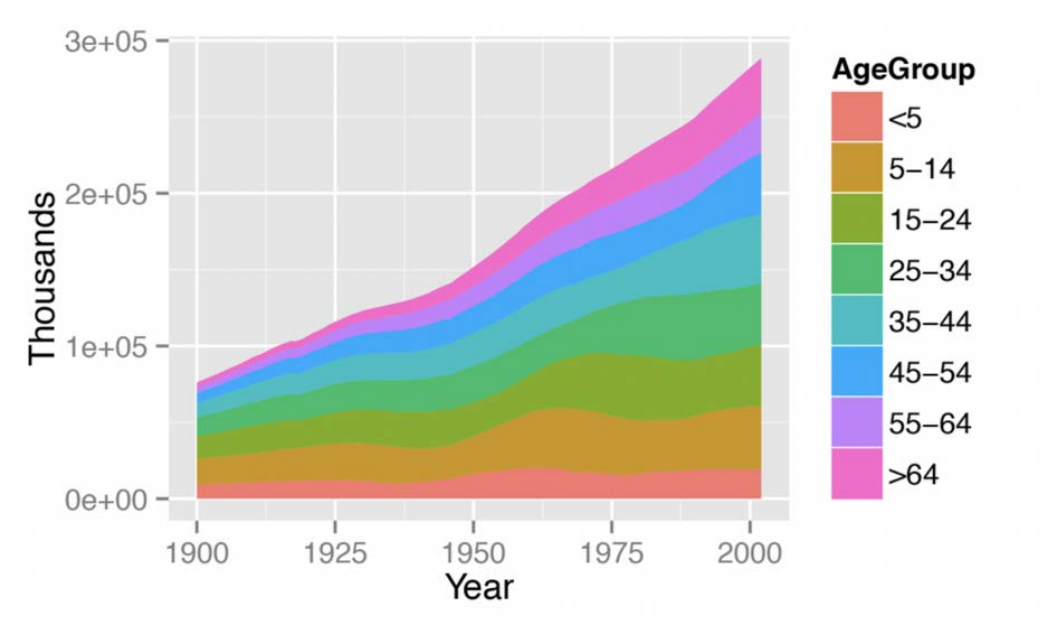
Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $\binom{k}{2} + k$ unique scatterplots]

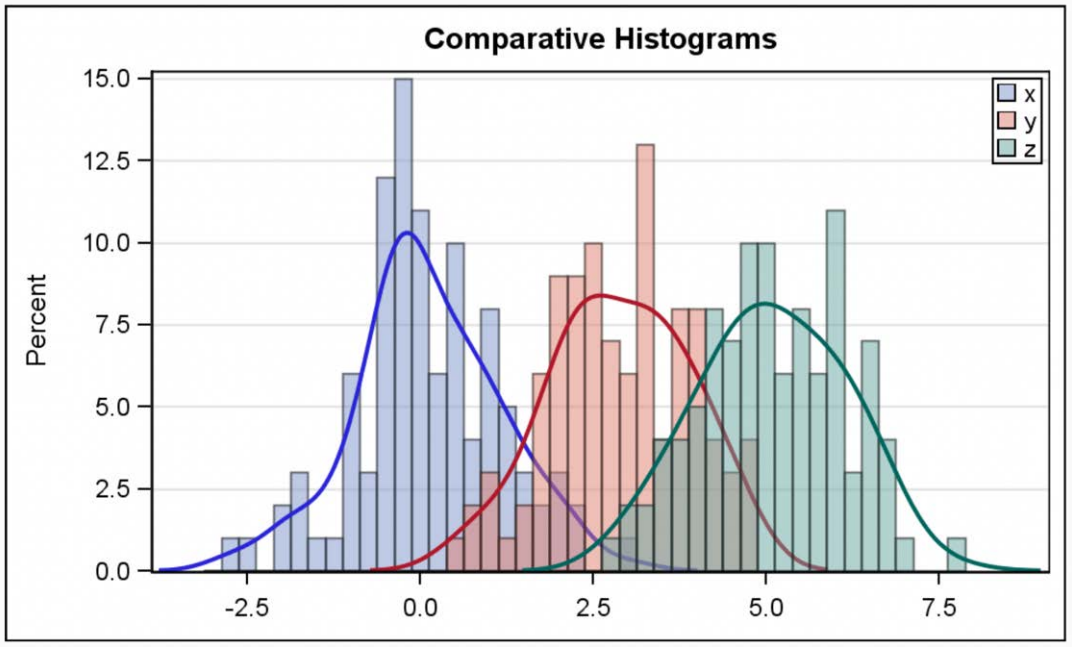
Stacked area graph to show trend over time

- 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).



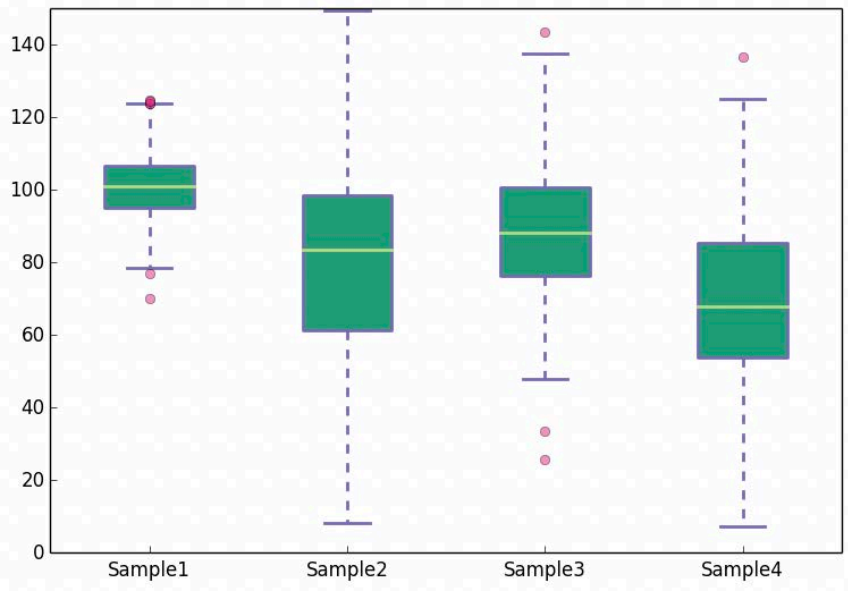
Multiple histograms

- Plotting **multiple histograms** 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 simplified visualization to compare a quantitative variable across groups. It highlights the range, quartiles, median and any outliers present in a data set.

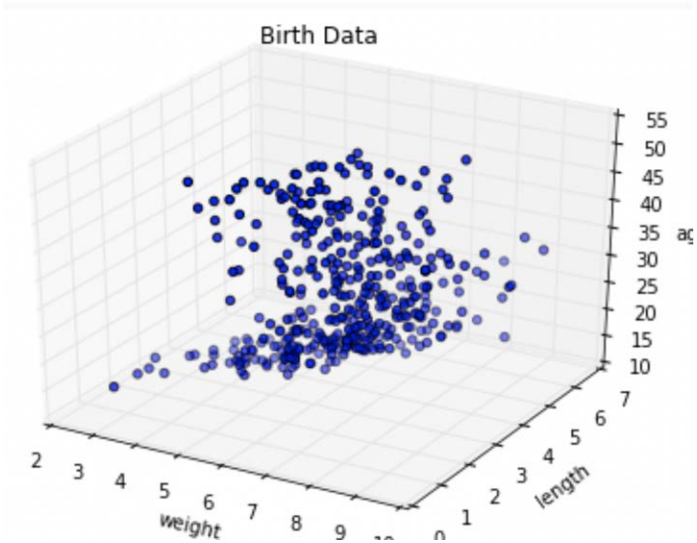


[Not] Anything is possible!

- Often your dataset seem 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?)

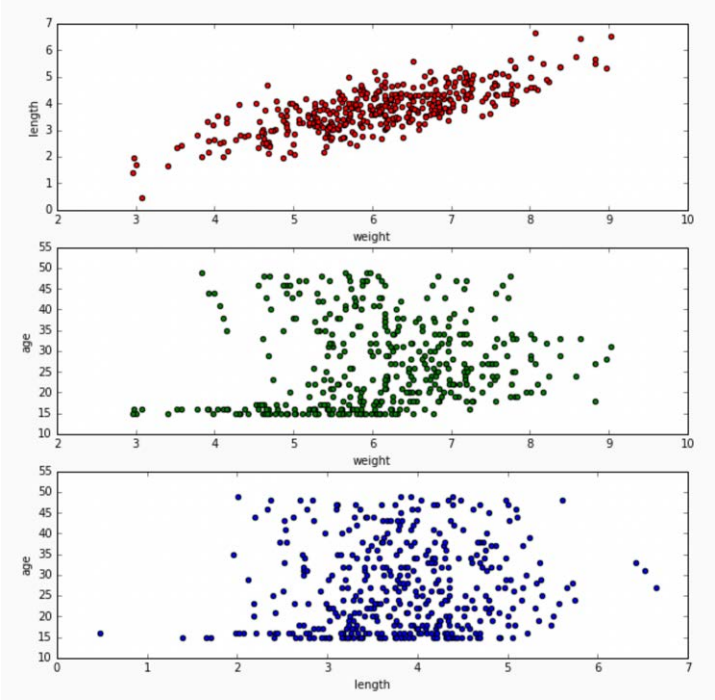
More dimensions not always better

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



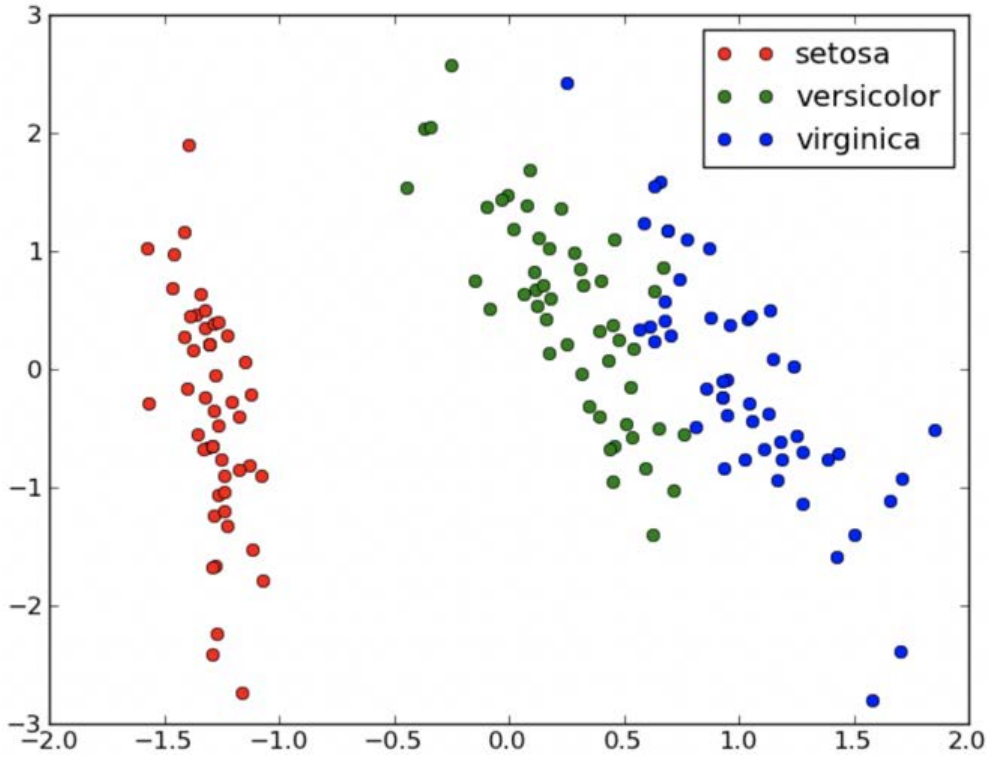
Reducing complexity

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



Reducing complexity

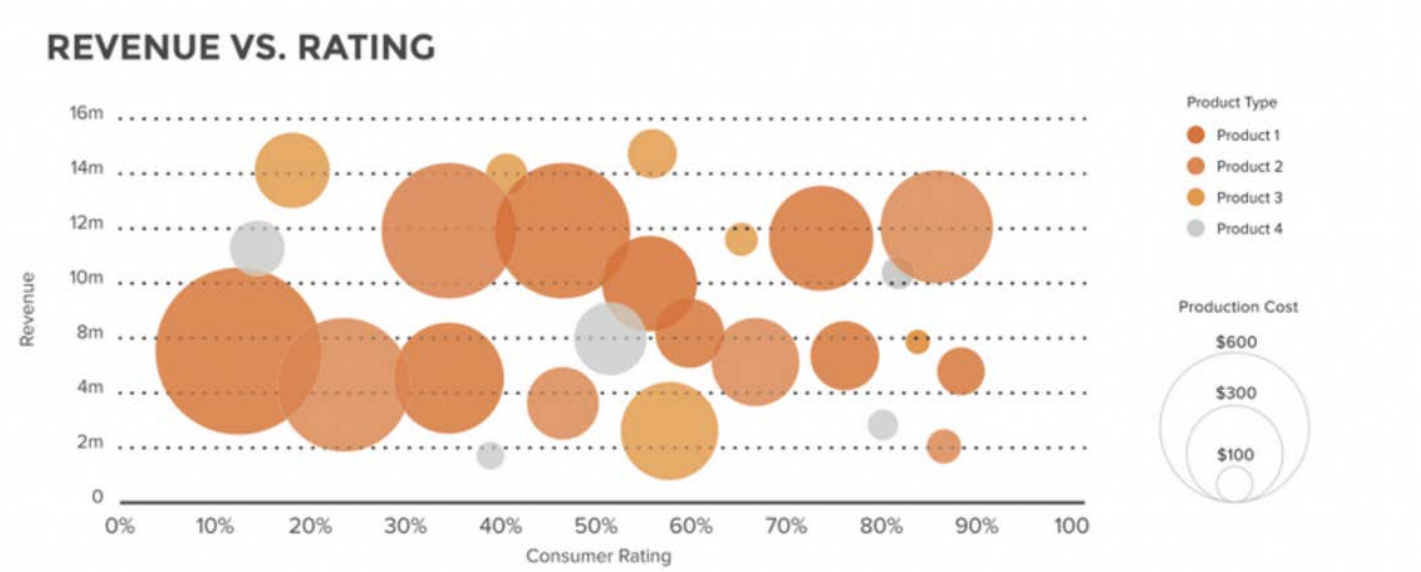
- For 3D data, color coding a categorical attribute can be “effective”



This visualizes a set of Iris measurements. The variables are: petal length, sepal length, Iris type (setosa, versicolor, virginica).

3D can work

- For 3D data, a quantitative attribute can be encoded by size in a bubble chart.



- The above visualizes a set of consumer products. The variables are: revenue, consumer rating, product type and product cost.

Summary

- Explore vector/tabular data
  - Attribute types
  - Basic statistics
  - Visualization