2024

Data, Statistics, and Visualization

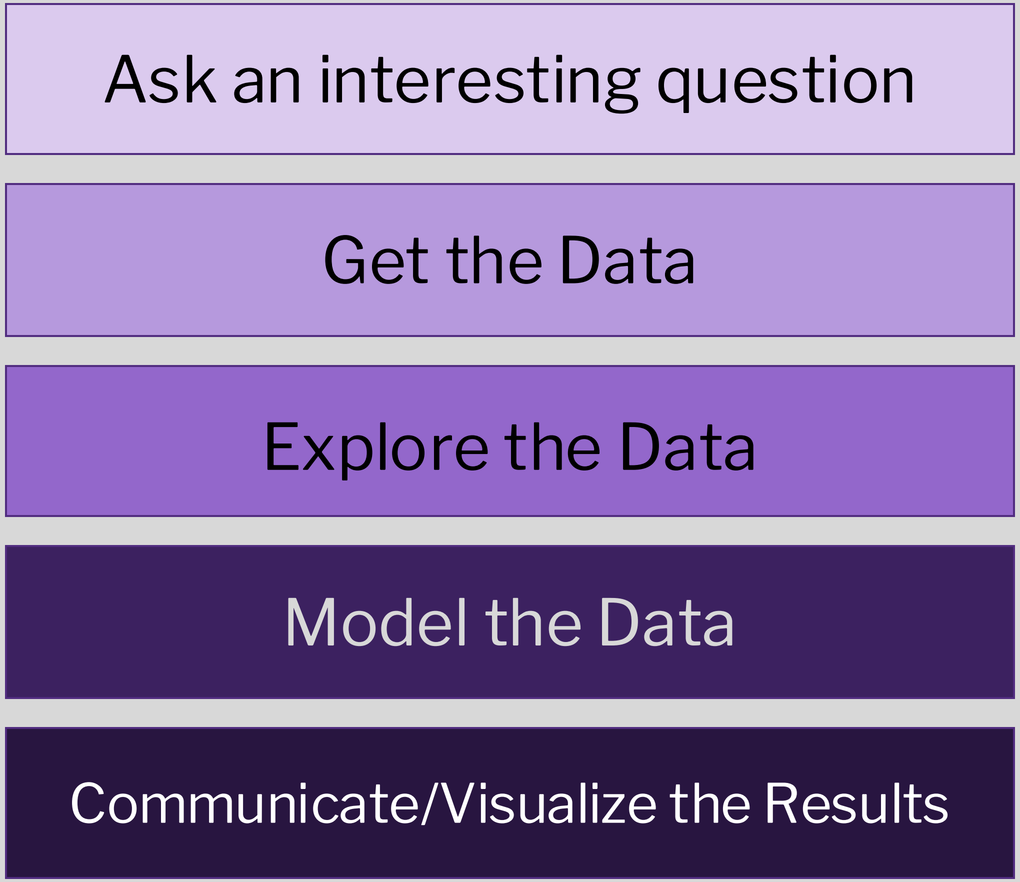
CPSC 4300/6300: Applied Data Science

Dr. Aaron J. Masino

Associate Professor, School of Computing

College of
ENGINEERING, COMPUTING
AND APPLIED SCIENCES

The Data Science Process



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Today we will begin introducing the data collection and data exploration steps.

Data Defined

What is data?

- Collection of “measurements” or “observations” that can be recorded and convey information about some entity of interest
- Just about any information that can be recorded in some physical medium (e.g. written down, stored electronically, etc) can be considered data
- *datum* : individual value in collection of data. Notion of individual is *context specific*, e.g. consider whether a single image or a single pixel is the datum.
- *data* : collection of related datum

Data Types

Eye color
(nominal)

Educational attainment
(ordinal)

Number of hospital admissions
(discrete)

EKG output
(continuous)

- **Qualitative**: measures are subjective
  - Nominal / categorical –possible values (labels) have no objective ordering
  - Ordinal
    - ➢ Possible values can be ordered
    - ➢ Unit changes in attribute values are not uniform
- **Quantitative**: measures are objective
  - Discrete
    - ➢ Possible values can be ordered
    - ➢ All possible values can be enumerated
    - ➢ Unit changes in attribute values are uniform
  - Continuous
    - ➢ Values can be ordered
    - ➢ Possible values cannot be enumerated

Structured Data

Eye Color	Pain Level	Age (years)	Temp (F)	Medicatio n
Brown	Low	28	97.5	Tylenol
Blue	High	47	99.2	Advil
Green	Low	36	98.2	Tylenol
Hazel	Medium	19	100.2	Tylenol
Gray	High	77	101.3	Aleve
Hazel	High	64	96.5	Aleve

- Has a standardized format (typically tabular) with well defined attributes (aka features)
- Every instance (e.g., a patient entry) has the same attribute types (e.g., height)
  - ➢ Dimension = number of attributes
- Each attribute is of a known data type with the same precision and possible values
- Every instance has the same organization
  - ➢ For example, in tabular data a given column in a data instance always represents the same attribute
  - ➢ Implies every instance has the same number of attributes

Unstructured Data

Lacks one or more of the properties of structured data

- Organization of data does not guarantee that values in each position represent the same concept
- Instances may have different number of attributes

How can we use unstructured data in data science?

- Feature engineering
- Neural networks

Computing with Data - Atomic data types

Python (and all other programming languages) have atomic types from which all other higher order data types are constructed

- **Numeric:**
  - Integers: 1, 2, 3, ..
  - Floating point (decimal): 3.14159, 2.718, 1.618
- **Boolean:**
  - Binary: 0 or 1
  - Boolean: True or False
- **Characters:** a, b, c, …, !, #, …
  - **Strings:** sequence of characters (*not an atomic type*)

Computing with Data - Compound data types

Python uses atomic types to build compound data types. Examples:

- **Date and time:** includes a year, month, date, hour, minute, second
- **Lists:** a list is a sequence of values (i.e., data)

  [“ford”, “chevy”, “Toyota”, …]
- **Dictionaries:** A dictionary is a collection of key-value pairs

  x : *y*

  where *x* is allowed to be any *hashable* type* (e.g., strings, integers) representing the “name” of the entry, and *y* is a value of any type.

  {”cars” : [“ford”, “chevy”, “Toyota”, …], “years”:[1903, 1911, 1937, ..]

  *hashable : there is some function (the hash function) that always maps a given input to fixed-size byte string

Computing with Data – Python Class Objects

- Python compound data types are instances of a *Class*
- *Class* – a formalism for composing data types and adding functions
- Example

```
Class Complex:
      def __init__(self, real, img):
           self.r = real
           self.i = img

      def magnitude(self):
           return self.r*self.r + self.i*slef.i
x = Complex(2,4)
print(x.magnitude())
4.472
```

Storing Atomic Data Types

- Usually stored in a text file using Python *i/o* operations
- Tabular data usually stored in a *.csv* (comma separated value) file
  - Each row is an entry
  - Columns are features

Storing Compound Data Types

How to save compound data types to files?

- Serialization for compound objects : the process of converting an object to a byte stream, and the inverse of which is converting a byte stream back to a Python object hierarchy
  - JSON (javascript object notation)
  - XML (extensible markup language)
  - Python pickle / dill libraries

Acquiring Data

Where do data come from?

- **Internal sources:** already collected by or is part of the overall data collection of your organization (business data).
  For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources:** available in ready to read format from an outside source for free or for a fee.
  For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference.
- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing.
  For example: data appearing only in print form, or data on websites.

Where do data come from? More examples

- **Internal sources:** The data you collect from the users who access your website. Important: *International regulations apply to the data collection management.* i.e., https://www.termsfeed.com/blog/legal-requirements-collect-personal-data/
- **Existing External Sources:**
  - ➢ CommonCrawl (data from the web https://commoncrawl.org/)
  - ➢ X API (https://developer.x.com/en)
- **External Sources Requiring Collection Efforts**: Online reviews
  - ➢ (https://www.sitejabber.com/) Listings on websites
  - ➢ (https://www.simplyrecipes.com/) Collecting recipes
  - ➢ (https://cars.com)

Ways to gather online data

How to get data generated, published or hosted online:

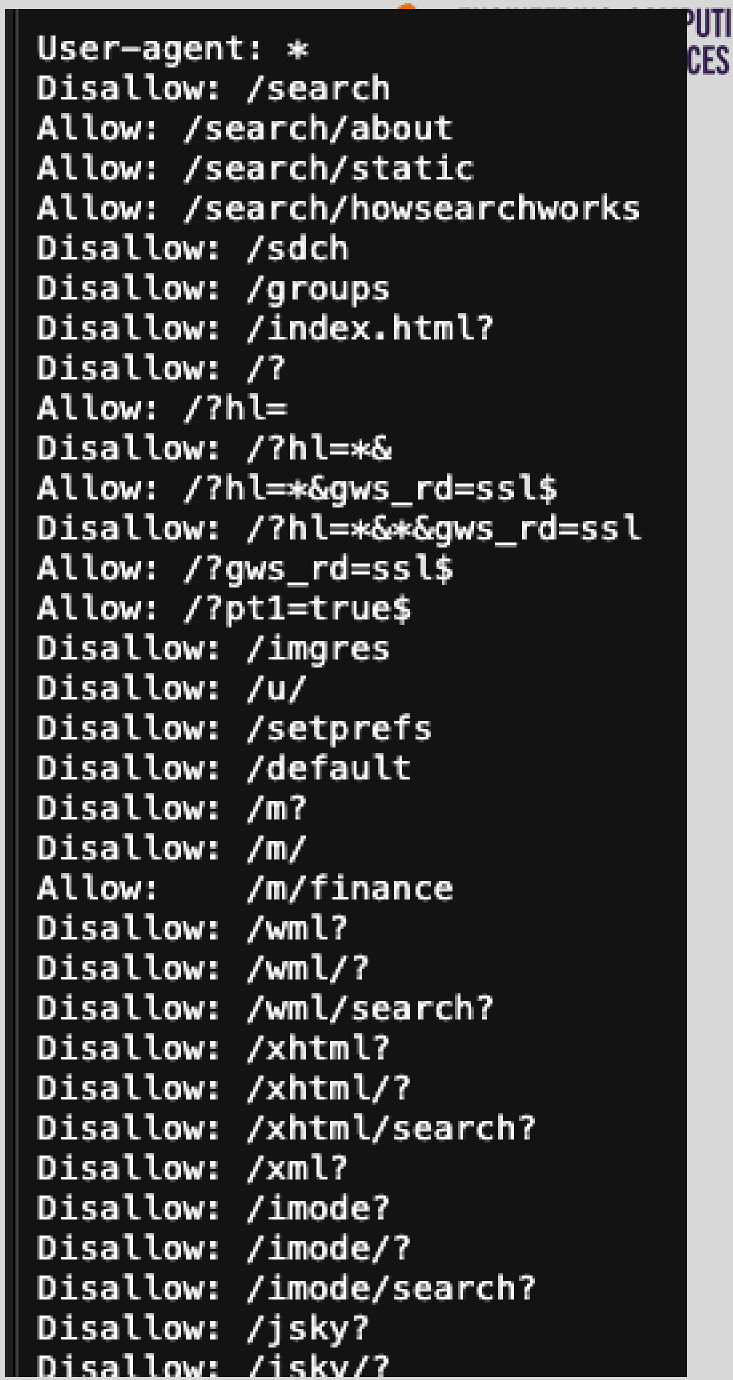
- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use. For example: Google Maps API, Facebook API, X API
- **RSS (Rich Site Summary):** summarizes frequently updated online content in standard format. Free to read if the site has one. For example: news-related sites, blogs
- **Web scraping:** using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file (often in tables).

Web Scraping

- **Why do it?** Older government or smaller news sites might not have APIs for accessing data. Or, you don’t want to pay to use the API or the database.
- **How do you do it?** There are many tools available. (e.g., beautifulsoup, selenium, puppeteer, apify, cheerio, scrapy, etc.)
- **Should you do it?**
  - You just want to explore: Are you violating their terms of service? Privacy concerns for website and their clients?
  - You want to publish your analysis or product: Do they have an API or fee that you are bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

Web Scraping

- **Rules for scraping (untold):**
- If you had to log in to a website to scrape the data, then you cannot make the extracted data public.
- Behave like a human, crawl at human pace (do not perform thousands URL calls per minute).
- Don’t overload the site: Only scrape the directories allowed in the robots.txt file from the web server root:
    - https://www.google.com/robots.txt
    - https://stock.adobe.com/robots.txt



```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
Disallow: /setprefs
Disallow: /default
Disallow: /m?
Disallow: /m/
Allow:    /m/finance
Disallow: /wml?
Disallow: /wml/?
Disallow: /wml/search?
Disallow: /xhtml?
Disallow: /xhtml/?
Disallow: /xhtml/search?
Disallow: /xml?
Disallow: /imode?
Disallow: /imode/?
Disallow: /imode/search?
Disallow: /jsky?
Disallow: /jsky/?
```

Web Scraping

- **Tools** (mostly for python, javascript, or java)
  - Only HTML: beautifulsoup, scrapy
  - Interaction is required (data is loaded via javascript or you have to perform a mouse or keyboard action): Puppeteer, Selenium
- **Platforms** (most are fee for access)
  - Cloud: apify,import.io, webscraper.io
  - Installed (also used for bot development, RPA: Robot Process Automation): PathUI, BluePrism

Data Wrangling

Tabular Data Is Often “Messy”

For tabular data, we typically want:

- A single file to correspond to a dataset
- Each column to represent a single variable
- Each row to represent a single observation.

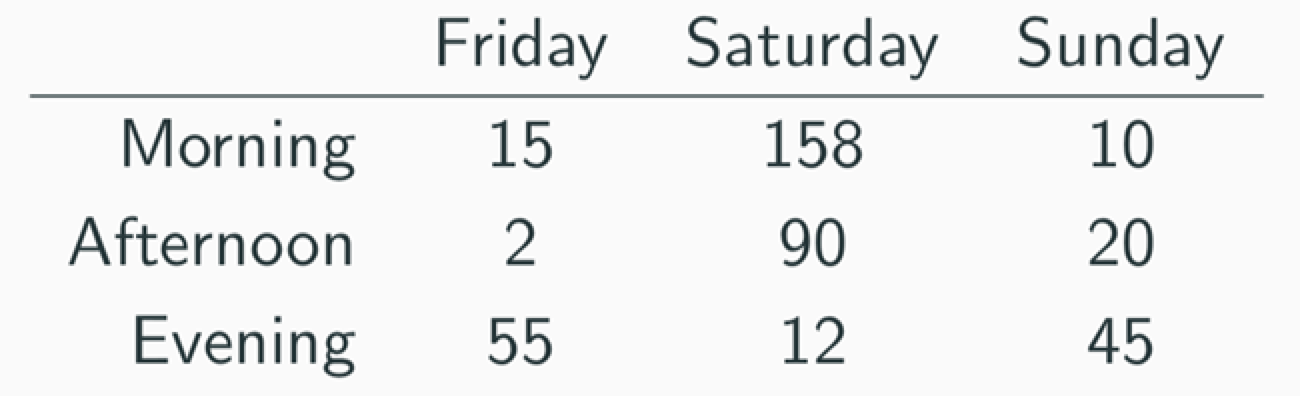
The data we ”get” is often “messy”

- Column headers are values, not variable names
- Variables are stored in both rows and columns
- Multiple variables are stored in one column/entry
- Multiple types of experimental units stored in same table

Messy Data

The following is a table accounting for **the number of produce deliveries over a weekend**.

This table is not well structured for analysis. What are the variables?

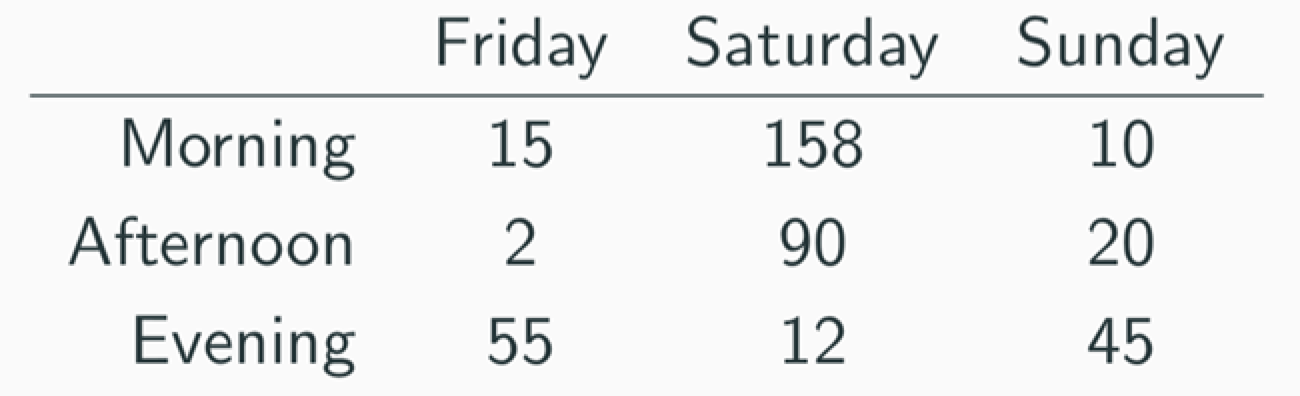


	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

What's the issue? How do we fix it?

Messy Data

Useful variables: Time, Day, Number of Produce.

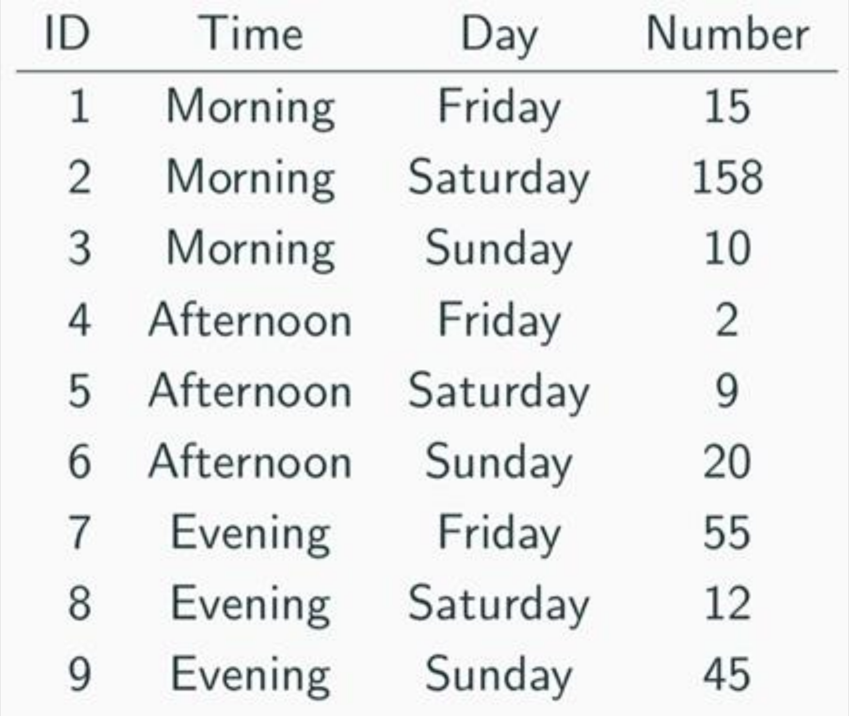


	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Problem: each column header represents a single value rather than a variable. Row headers are “hiding” the Day variable. The values of the variable, “Number of Produce”, is not recorded in a single column.

Fixing Messy Data

We need to reorganize the information to make explicit the event we’re observing, and the variables associated to this event.



ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

Other Issues We'll Discuss This Semester

- Missing values: how do we fill in?
- Wrong values: how can we detect and correct?
- Is the data suitable to answer the question(s) of interest?

Data Exploration: Descriptive Statistics

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 it’s impossible or intractable to obtain or compute with population data.

Basics of Sampling

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?

Participants included in an influenza vaccine trial may be healthy young adults, whereas those who are most likely to receive the intervention in practice may be elderly and have many comorbidities.

Sample Mean

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

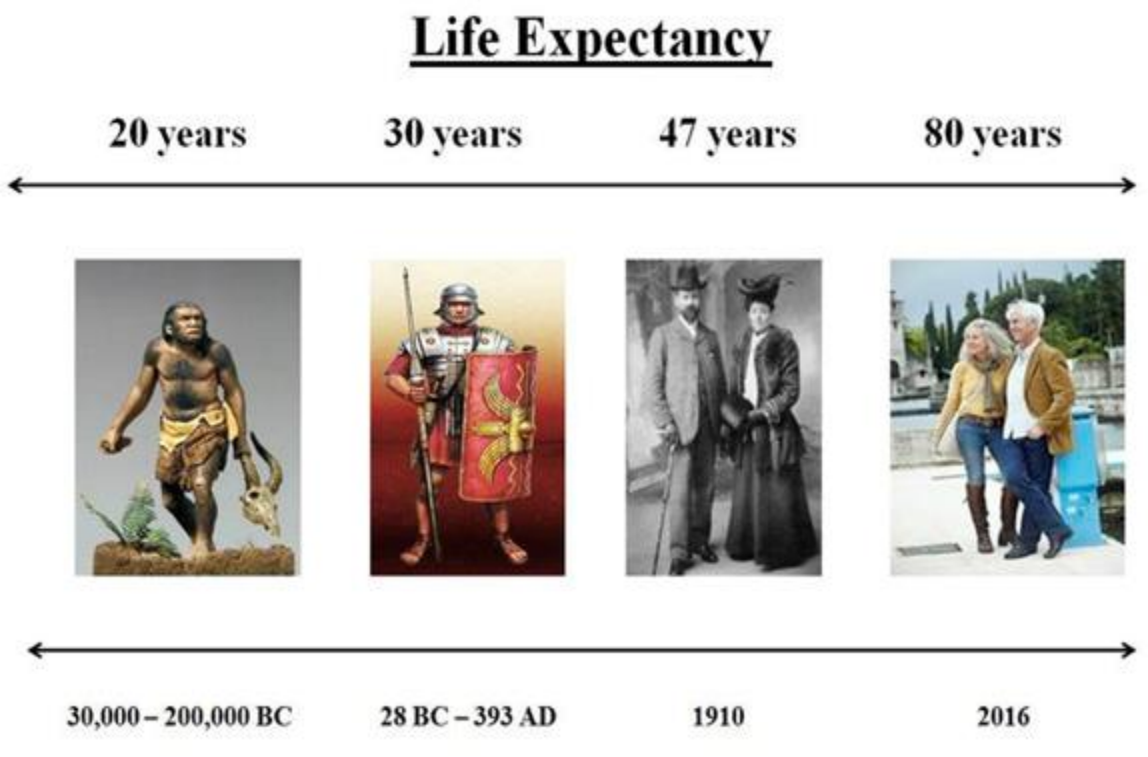
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

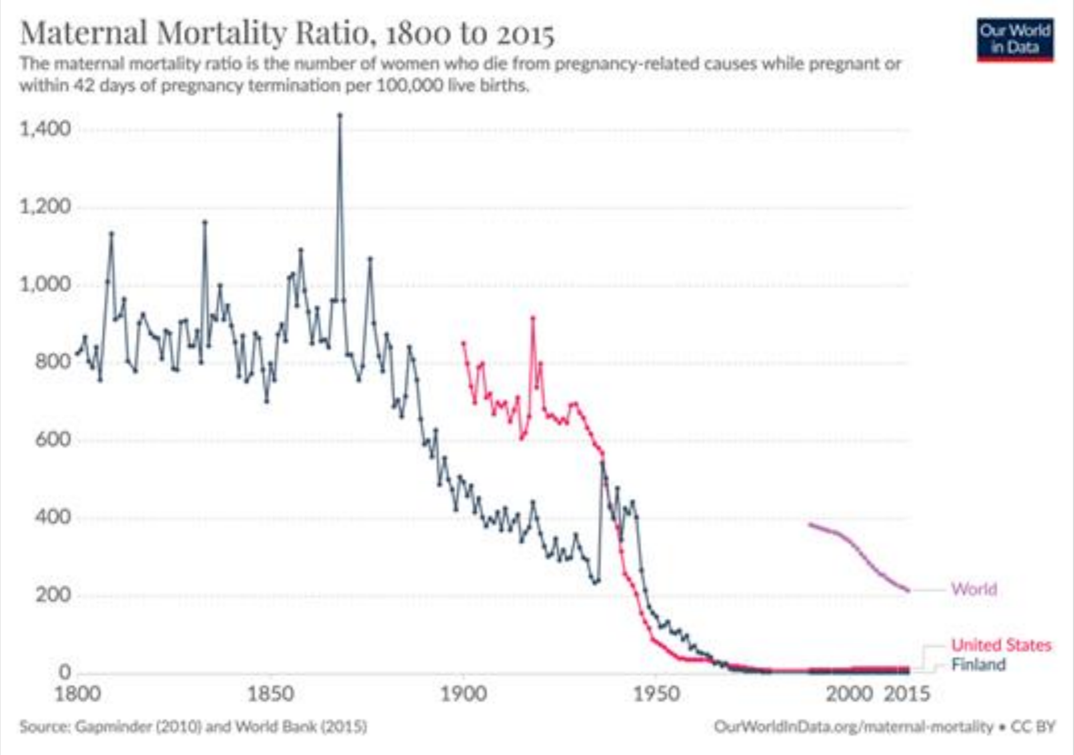
The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.

Key theme: there is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample Mean

The mean is likely to be influenced by data with high variability (sensitive to extreme values)
"The problem of life expectancy" https://www.youtube.com/watch?v=i2qckcs_tml




Maternal Mortality Ratio, 1800 to 2015

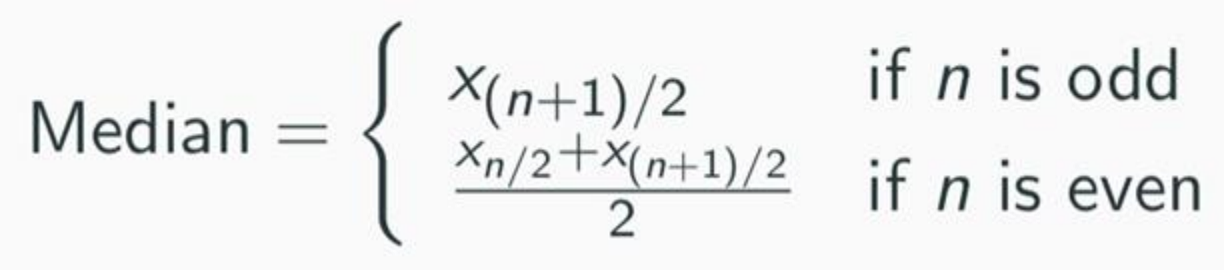
The maternal mortality ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births.

Source: Gapminder (2010) and World Bank (2015)

OurWorldInData.org/maternal-mortality • CC BY

Sample Median

The ***median*** of a set of *n* numerical observations in a sample, ordered by value is defined by



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

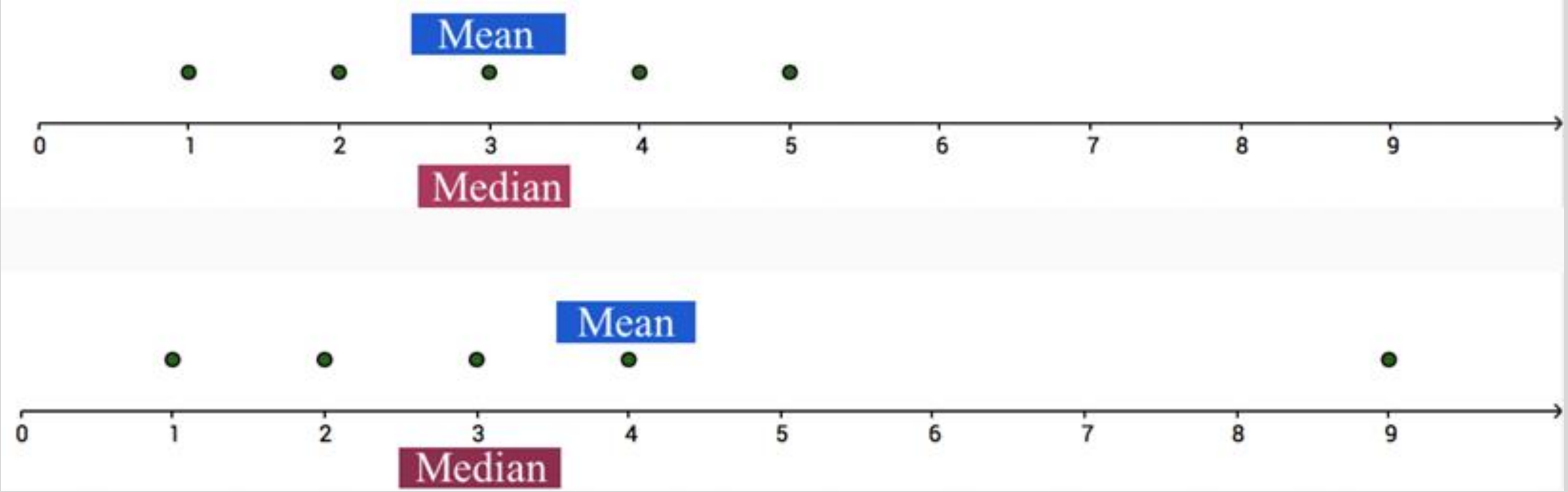
Example (already in order):

- 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.

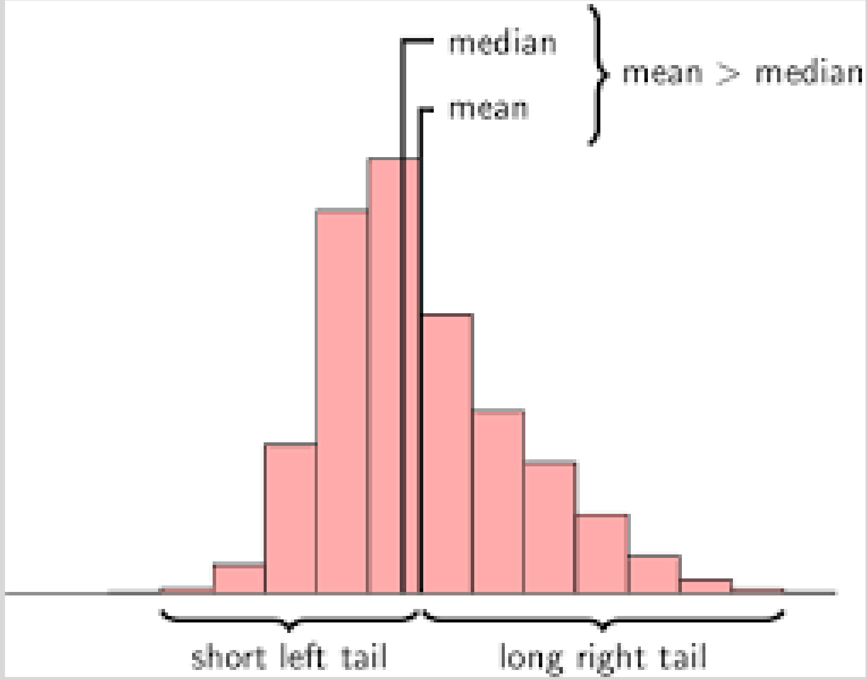
Mean vs. Median

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



Mean, Median, and Skewness

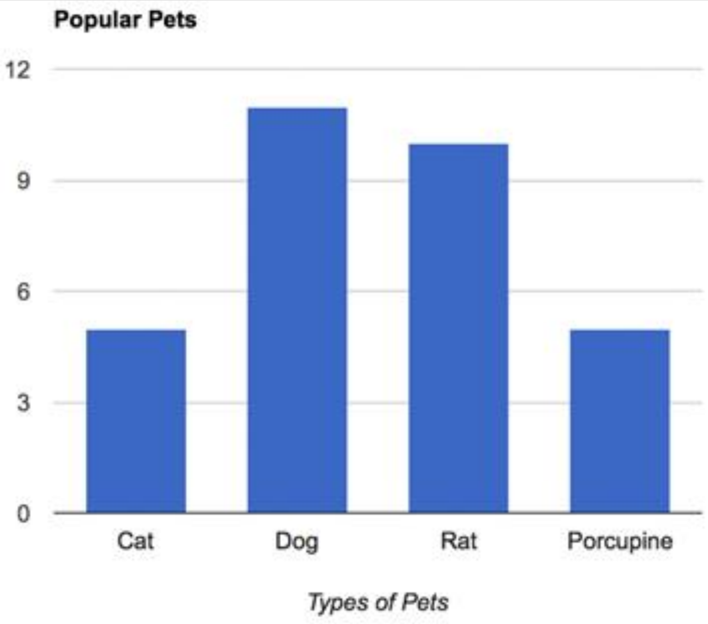
The mean is sensitive to outliers:



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

Regarding Categorical Variables

- For categorical variables, neither mean or median make sense. Why?
- Mode : the value that occurs most in the sample



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

Example:

- Sample: 2.1, 5.8, 9.6
- Max = 9.6
- Min = 2.1
- Range = 9.6 – 2.1 = 7.5

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

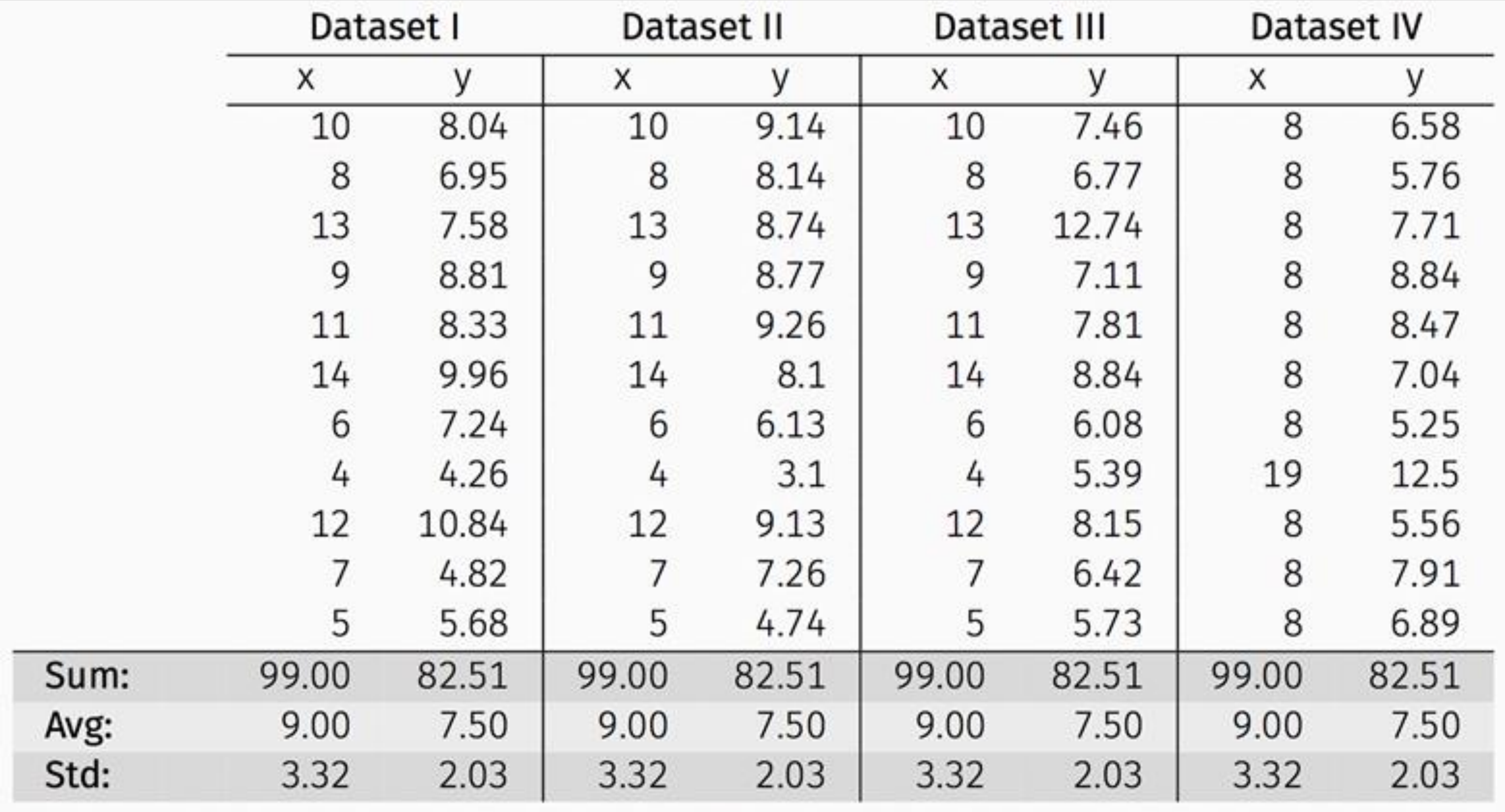
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Note: *s* does have the same units as the x_i!

Data Exploration: Visualizations

Anscombe’s Data

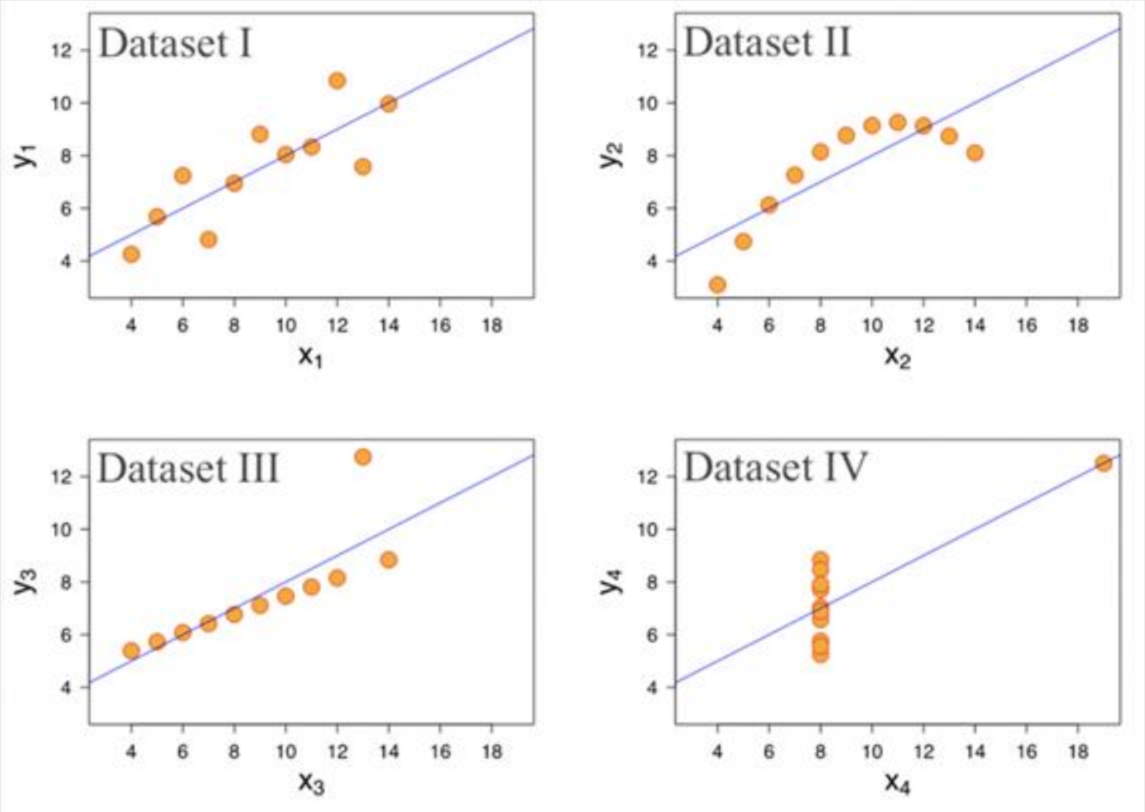
The following four data sets comprise the Anscombe’s Quartet; all four sets of data have identical simple summary statistics.



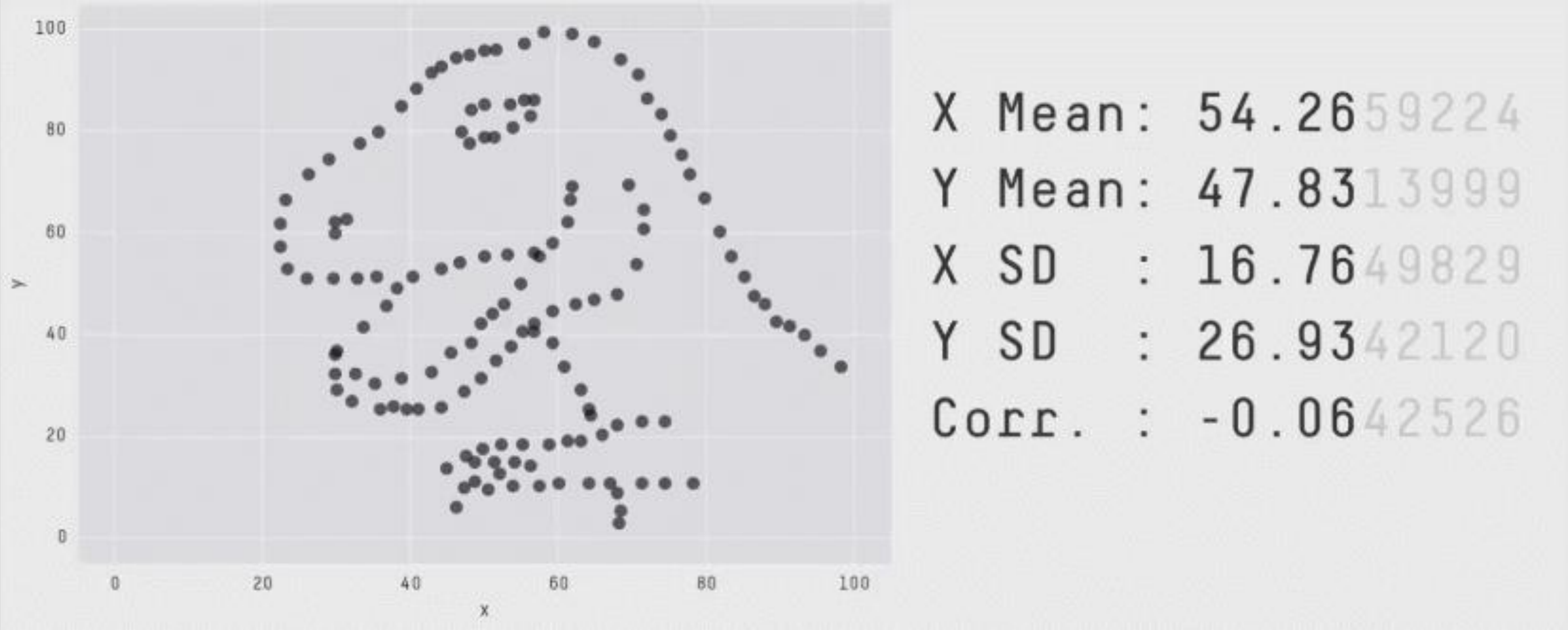
	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	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.



Another, more extreme example

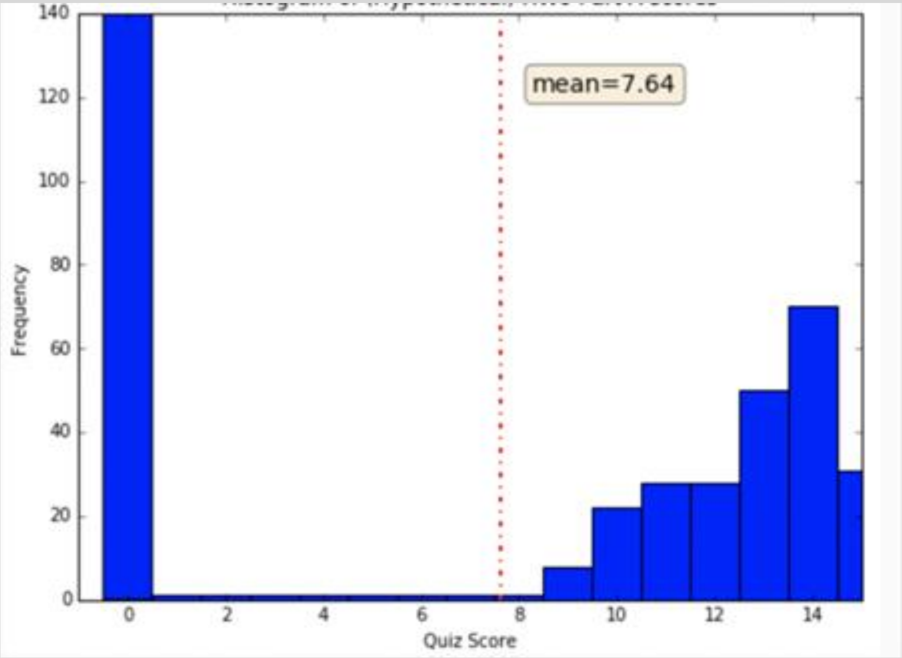


```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

More Visualization Motivation

If I tell you that the average score for Homework 0 was: 7.64/15 = 50.9% last year, what does that suggest?

And what does the graph suggest?



More Visualization Motivation

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

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

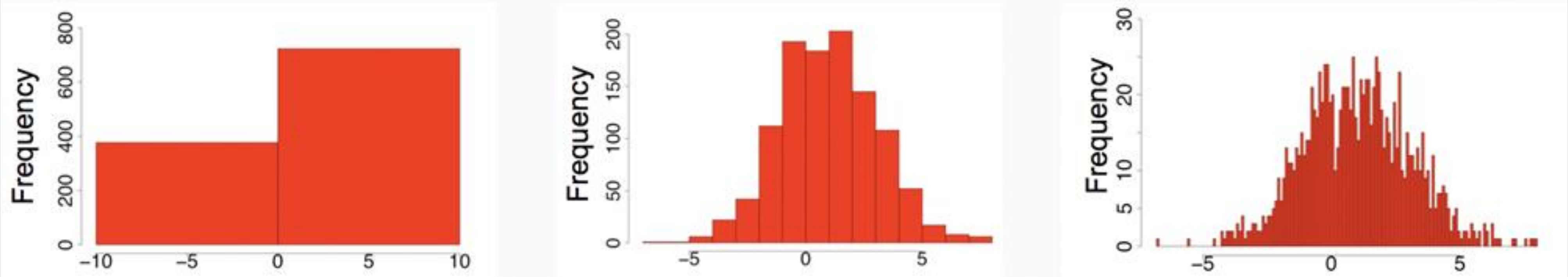
Types of Visualizations

Visualizations can illustrate:

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

Histograms to Visualize Distribution

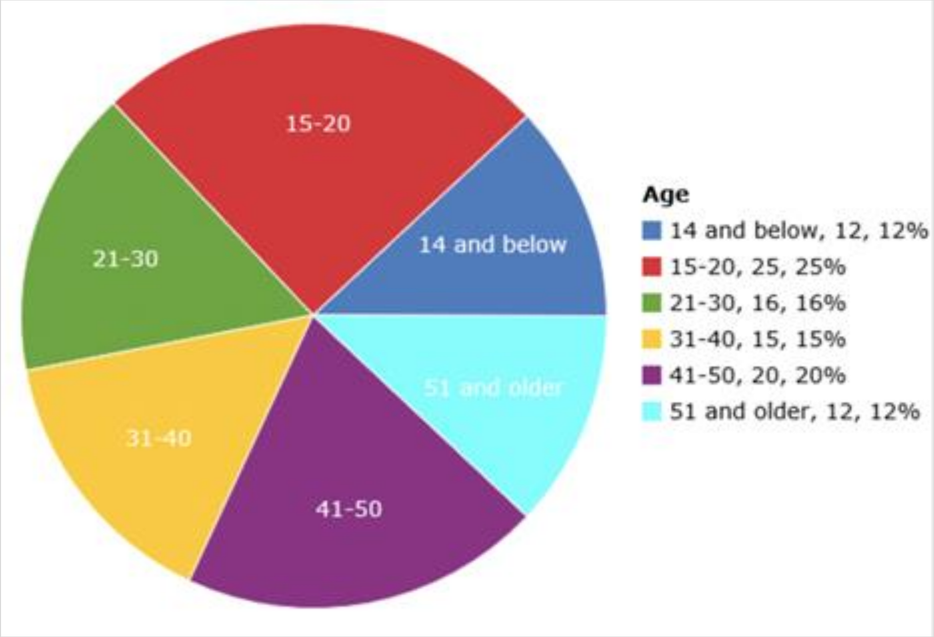
A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.



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

Pie Chart for a Categorical Variable?

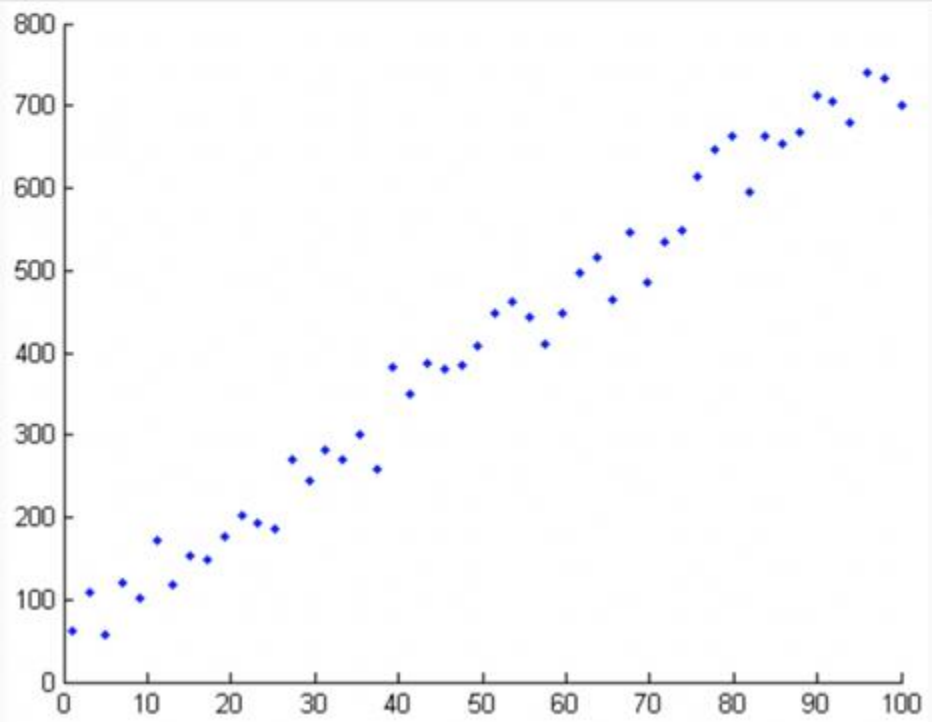
A **pie chart** is a (BAD) way to visualize the static composition (aka, distribution) of a variable (or single group).



Bar charts should be used instead. Why?

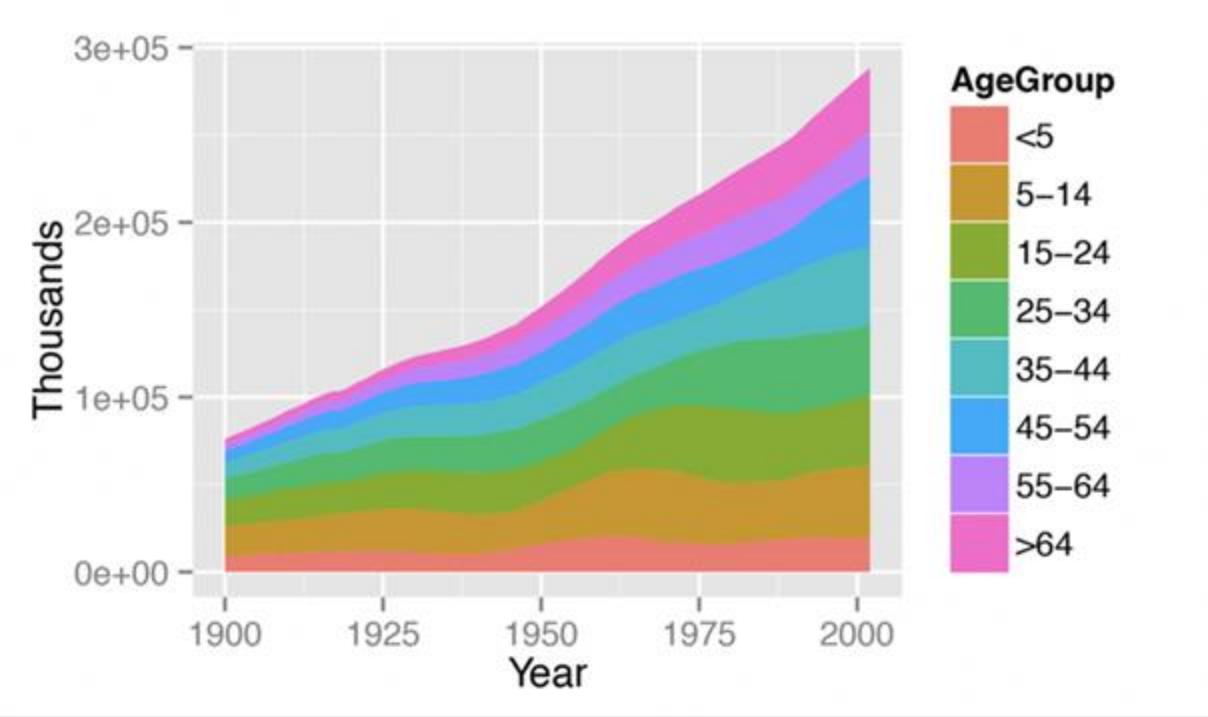
Scatter Plots to Visualize Relationships

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



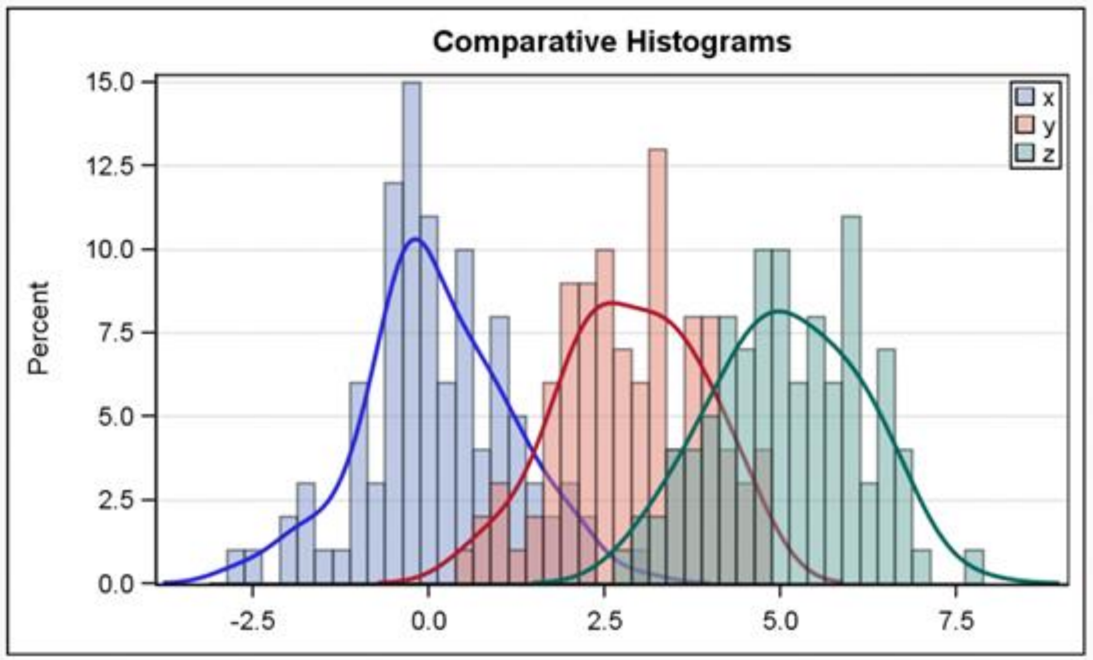
Stacked Area Graph for Trends Over Time

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).



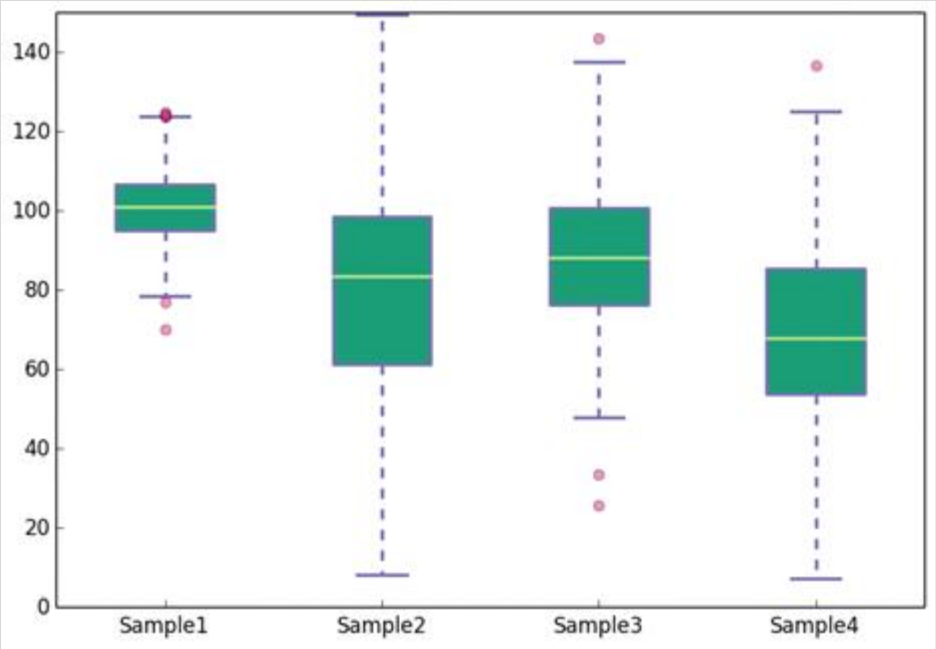
Multiple Histograms

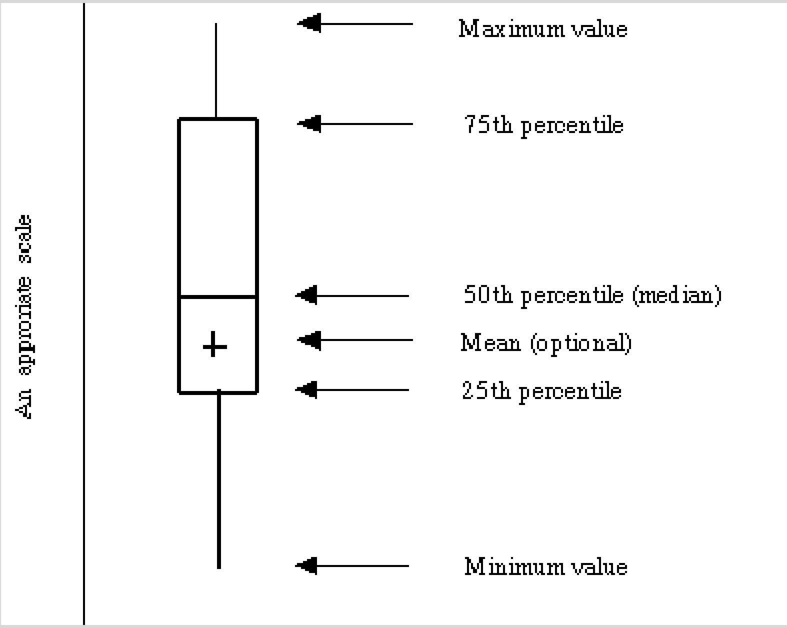
Plotting **multiple histograms** (and **kernel density estimates** of the distribution) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).

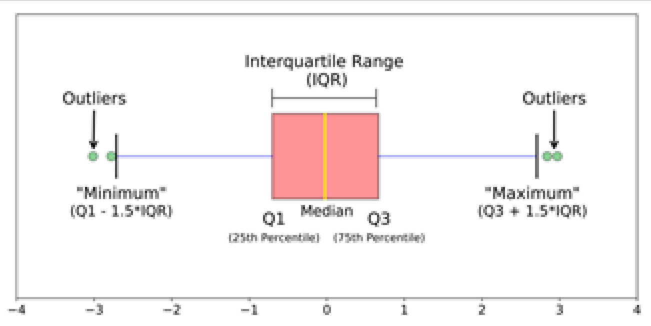


Boxplots

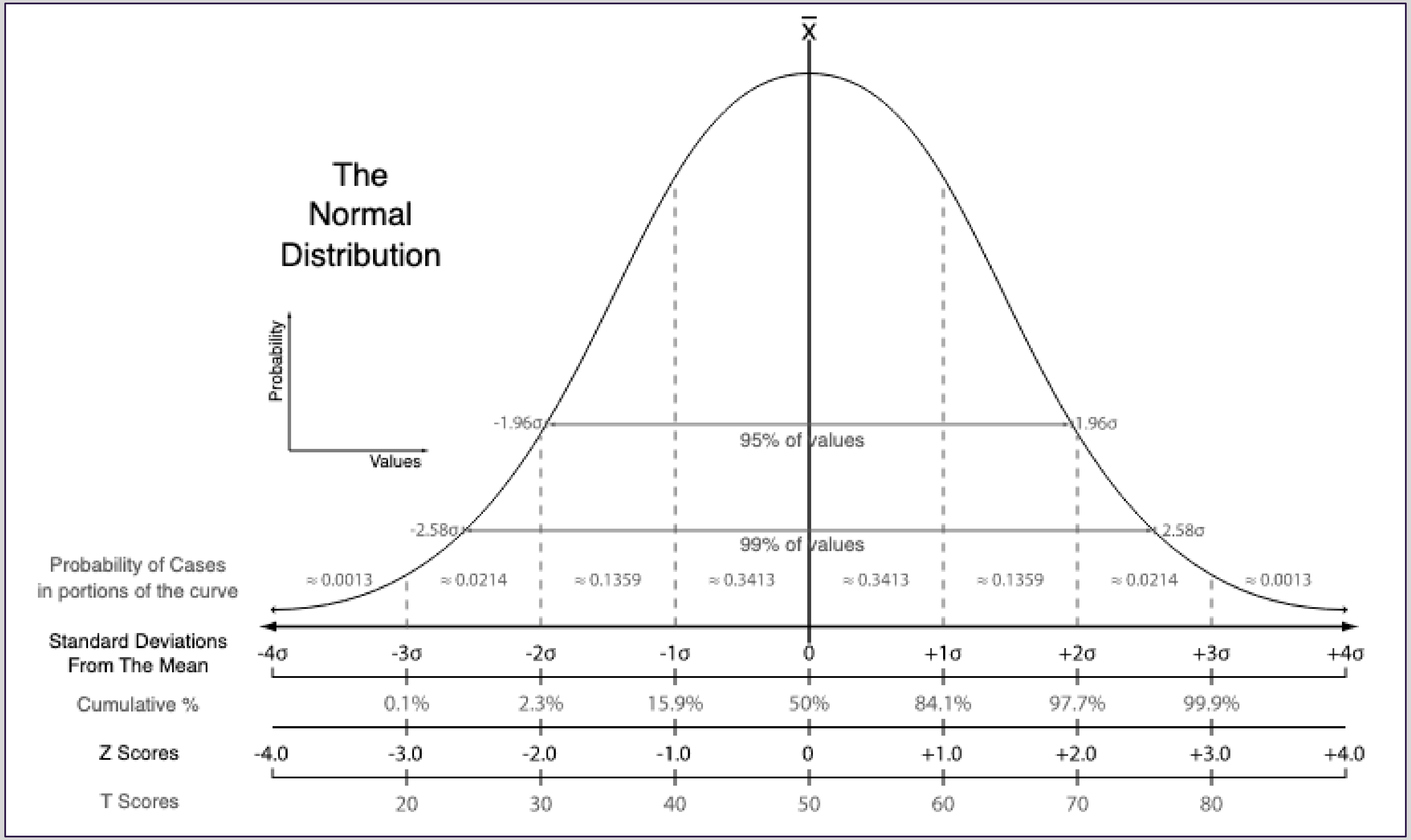
A **boxplot** is a simplified visualization to compare variable distributions across groups. It highlights the range, quantiles, median and any outliers present in a data set.







Outliers



$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

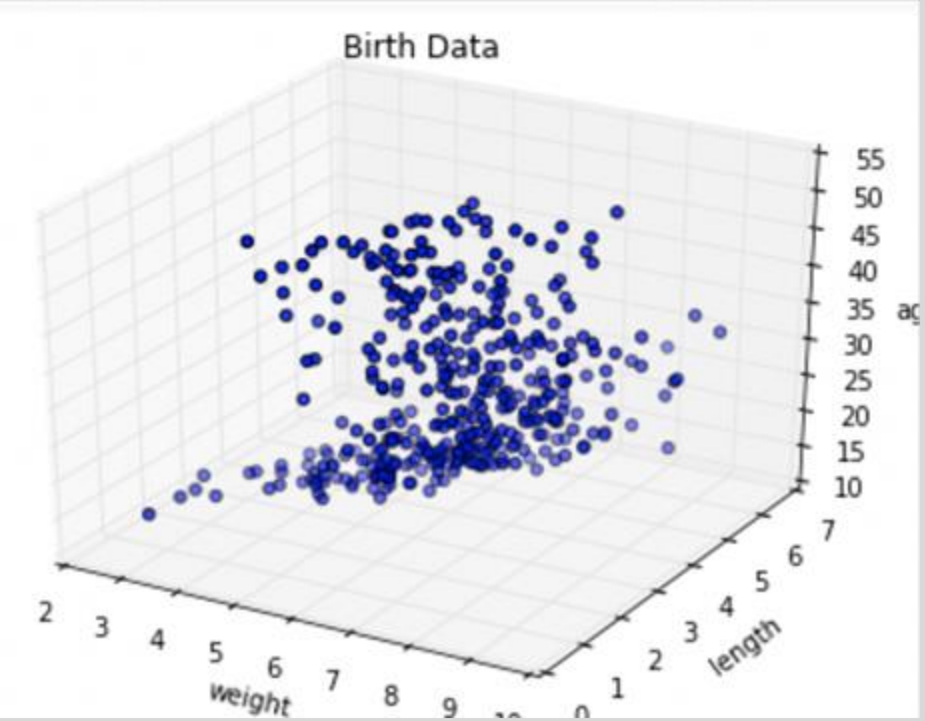
[Not] Anything is possible!

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?)

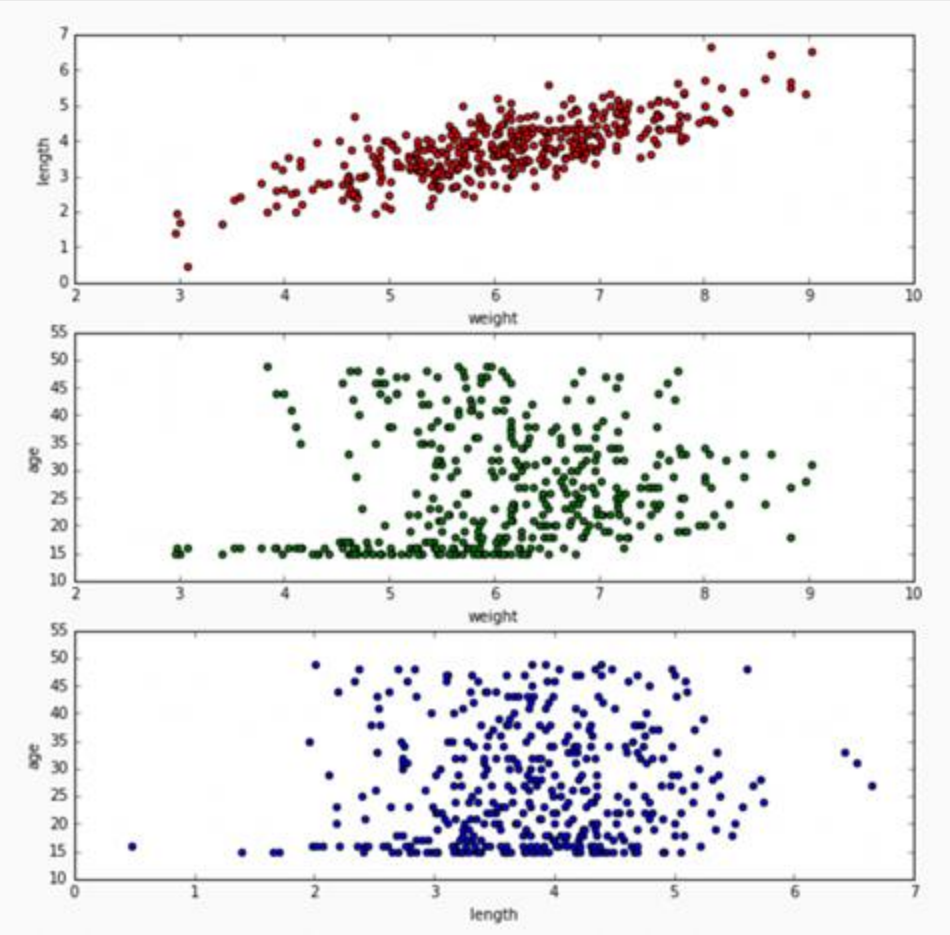
More dimensions is not always better

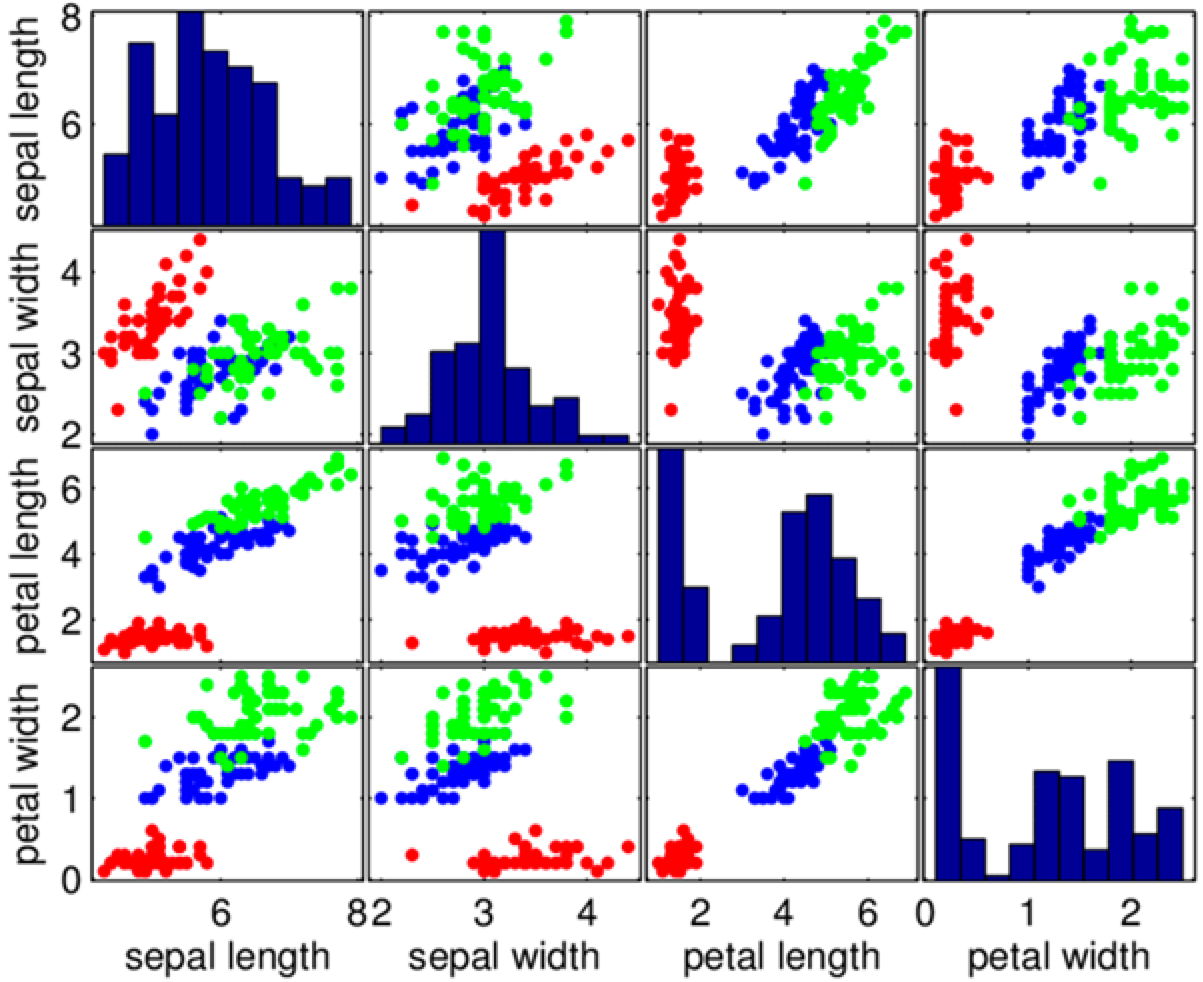
When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful



Reducing Complexity

- Relationships may be easier to spot by producing multiple plots of lower dimensionality.
- More advanced methods attempt to “project” the data into lower dimensions




sepal length
sepal width
petal length
petal width
8
6
4
3
2
6
4
2
2
1
0
6
8
2
3
4
2
4
6
0
1
2
sepal length
sepal width
petal length
petal width

Questions?