Lecture #1: Exploratory Data Analysis

CS 109A, STAT 121A, AC 209A

Pavlos Protopapas Kevin Rader
Margo Levine Rahul Dave

Lecture Outline

What are Data?

Data Exploration

- Descriptive Statistics
- Visualizations

An Example

What are Data?

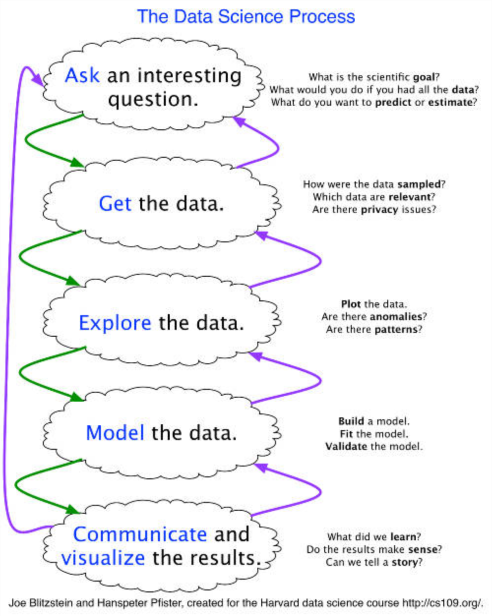
The Data Science Process

Recall the data science process.

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

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

The Data Science Process



What are data?

"A **datum** is a single measurement of something on a scale that is understandable to both the recorder and the reader. **Data** are multiple such measurements."

Claim: everything is (can be) data!

Where do data come from?

- **Internal sources**: already collected by or is part of the overall data collection of you organization.
  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
- **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?

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 Map API, Facebook API, Twitter 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.

Web Scraping

- Why do it? Older government or smaller news sites might not have APIs for accessing data, or publish RSS feeds or have databases for download. You dont want to pay to use the API or the database.
- How do you do it? See HW1
- 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?

Types of Data

What kind of values are in your data (data types)? Simple or atomic:

- **Numeric**: integers, floats
- **Boolean**: binary or true false values
- **Strings**: sequence of symbols

Types of Data

What kind of values are in your data (data types)? Compound, composed of a bunch of atomic types:

- **Date and time**: compound value with a specific structure
- **Lists**: a list is a sequence of values
- **Dictionaries**: A dictionary is a collection of key-value pairs, a pair of values x : y where x is usually a string called key representing the "name" of the value, and y is a value of any type.

  Example: Student record

  - First: Kevin
  - Last: Rader
  - Classes: [CS109A, STAT121A, AC209A, STAT139]

How is your data represented and stored (data format)?

- **Tabular Data**: a dataset that is a two-dimensional table, where each row typically represents a single data record, and each column represents one type of measurement (csv, tsp, xlsx etc.).
- **Structured Data**: each data record is presented in a form of a, possibly complex and multi-tiered, dictionary (json, xml etc.)
- **Semistructured Data**: not all records are represented by the same set of keys or some data records are not represented using the key-value pair structure.

How is your data represented and stored (data format)?

- Textual Data
- Temporal Data
- Geolocation Data

Tabular Data

In tabular data, we expect each record or observation to represent a set of measurements of a single object or event. We've seen this already in Lecture 0:

First Look At The Data

In [27]:
```
hubway_data = pd.read_csv('hubway_trips.csv', low_memory=False)
hubway_data.head()
```

Out[27]:

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_da
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0

Each type of measurement is called a **variable** or an **attribute** of the data (e.g. `seq_id`, `status` and `duration` are variables or attributes). The number of attributes is called the **dimension**.

We expect each table to contain a set of **records** or **observations** of the same kind of object or event (e.g. our table above contains observations of rides/checkouts).

Types of Data

We'll see later that its important to distinguish between classes of variables or attributes based on the type of values they can take on.

- **Quantitative variable**: is numerical and can be either:
  - **discrete** - a finite number of values are possible in any bounded interval.
    For example: "Number of siblings" is a discrete variable
  - **continuous** - an infinite number of values are possible in any bounded interval
    For example: "Height" is a continuous variable
- **Categorical variable**: no inherent order among the values For example: "What kind of pet you have" is a categorical variable

Common Issues

Common issues with data:

- Missing values: how do we fill in?
- Wrong values: how can we detect and correct?
- Messy format
- Not usable: the data cannot answer the question posed

Handling Messy Data

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

What are the variables in this dataset?

What object or event are we measuring?

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

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

Handling Messy Data

Were measuring individual deliveries; the variables are 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.

Handling Messy Data

We need to reorganize the information to make explicit the event were 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

More Messiness

What object or event are we measuring?

What are the variables in this dataset?

Delivery	Amount
On Sunday	
10:30	43
12:30	12
12:35	30
On Monday	
11:30	29
11:57	87
11.59	63
On Tuesday	
11:33	19
11:15	27
12.59	54

How do we fix?

More Messiness

Were measuring individual deliveries; the variables are Time, Day, Number of Produce:

Days	times	Amount
Sunday	10:30	43
Sunday	12:30	12
Sunday	12:35	30
Monday	11:30	29
Monday	11:57	87
Monday	11.59	63
Tuesday	11:33	19
Tuesday	11:15	27
Tuesday	12.59	54

Common causes of messiness are:

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

In general, we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation.

We want to *tabularize* the data. This makes Python happy.

Data Exploration

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:

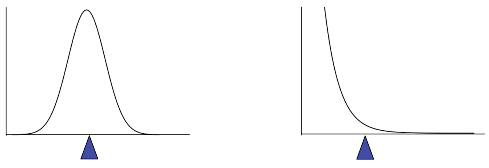
- **Selection bias**: some subjects or records are more likely to be selected
- **Volunteer/nonresponse bias**: subjects or records who are not easily available are not represented

Examples?

Sample mean

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

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



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

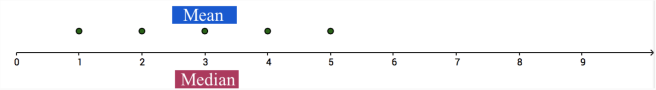
$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2}+x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

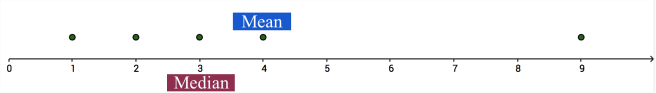
Example (already in order):
Ages: 17, 19, 21, $\underline{22,\ 23}$, 23, 23, 38
Median = (22+23)/2 = 22.5

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

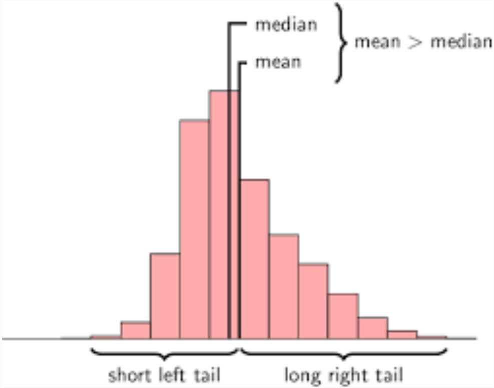
The mean is sensitive to extreme values (*outliers*)





Mean vs. Median

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



The above distribution is called *right-skewed* since the mean is greater than the median.

Measures of Centrality: computation

How hard (in terms of algorithmic complexity) is it to calculate

- the mean
- the median

Measures of Centrality: Computation Time

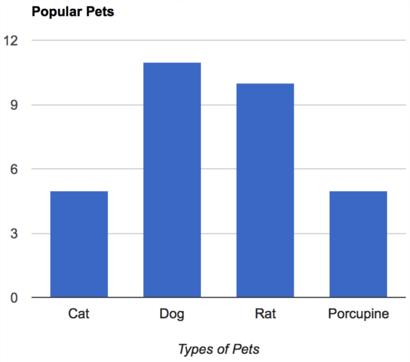
How hard (in terms of algorithmic complexity) is it to calculate

- the mean: at most $O(n)$
- the median: at least $O(n \log n)$

Note: Practicality of implementation should be considered!

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.

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**.

$$\text{Range} = \text{Maximum Value - Minimum Value}$$

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 doesnt 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!

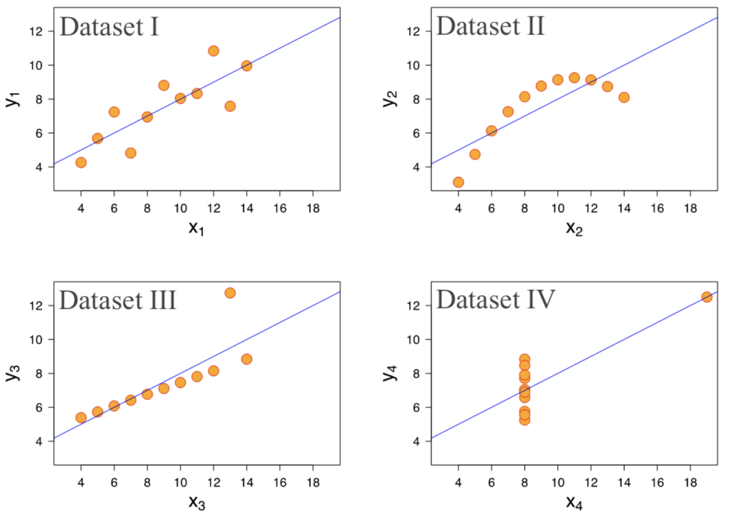
Anscombe's Data

The following four data sets comprise the Anscombes 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 2

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



More Visualization Motivation

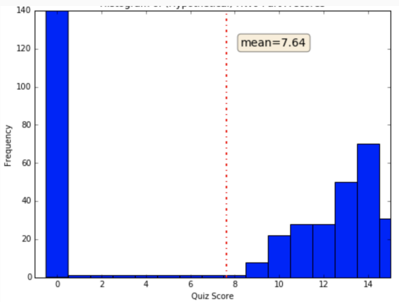
If I tell you that the average score for Homework 0 is: 7.64/15.

What does that suggest?

More Visualization Motivation

If I tell you that the average score for Homework 0 is: 7.64/15.

What does that suggest?



And what does the graph suggest?

More Visualization Motivation

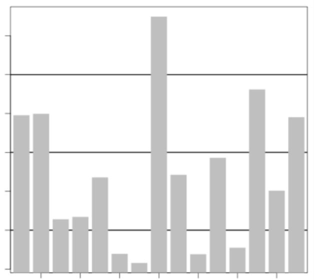
Visualizations help us to analyze and explore the data. They help to:

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

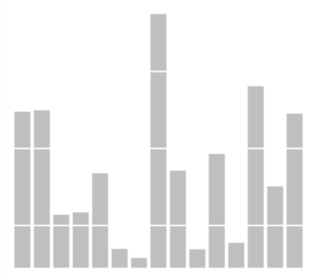
Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data

Good



Better

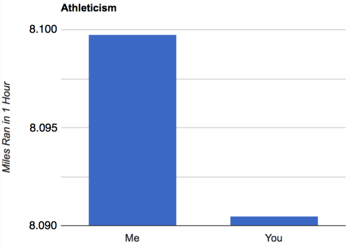


Principles of Visualizations

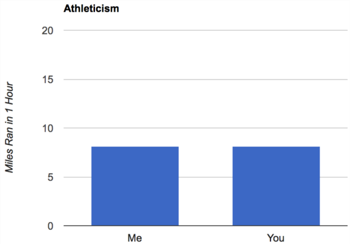
Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Dont lie with scale: minimize $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$ (Lie Factor)

Bad



Better

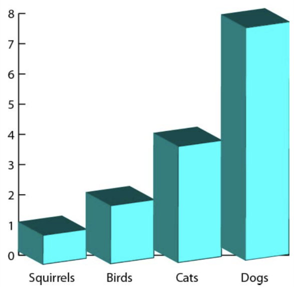


Principles of Visualizations

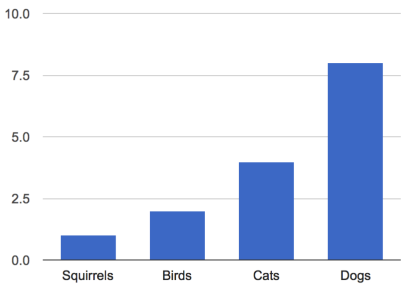
Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Dont lie with scale: minimize $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$ (Lie Factor)
3. Minimize chart-junk: show data variation, not design variation

Bad



Better



Principles of Visualizations

Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Dont lie with scale: minimize $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$(Lie Factor)
3. Minimize chart-junk: show data variation, not design variation
4. Clear, detailed and thorough labeling

More Tufte details can be found here:
http://www2.cs.uregina.ca/ rbm/cs100/notes/spreadsheets/tufte_paper.html

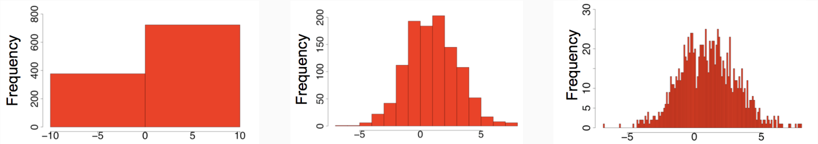
Types of Visualizations

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.

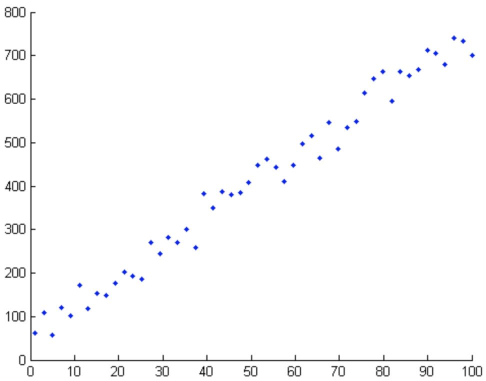


Note: Trends in histograms are sensitive to number of bins.

Scatter plots to visualize relationships

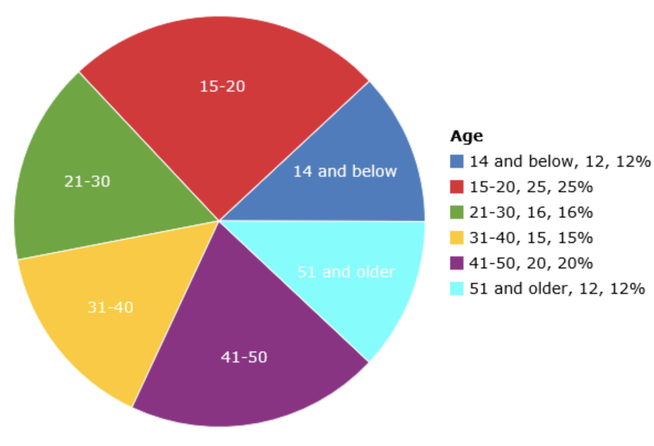
A **scatter plot** is a way to visualize how multi-dimensional data are distributed across certain values.

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



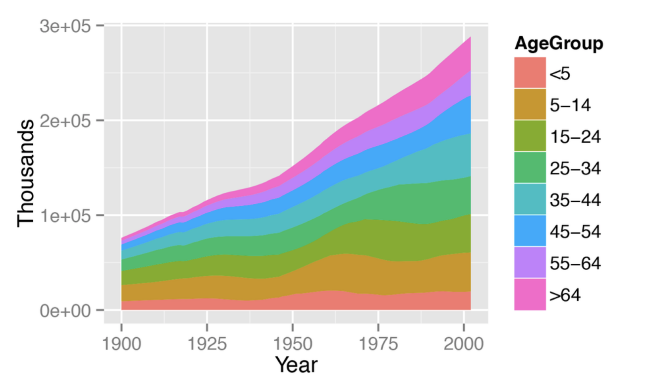
Pie chart for a categorical variable

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



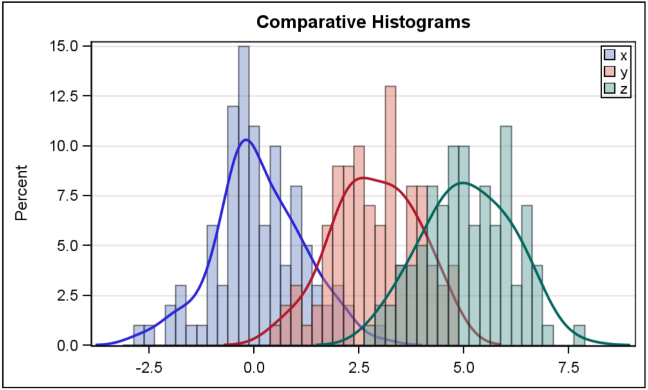
Stacked area graph to show trend over time

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).



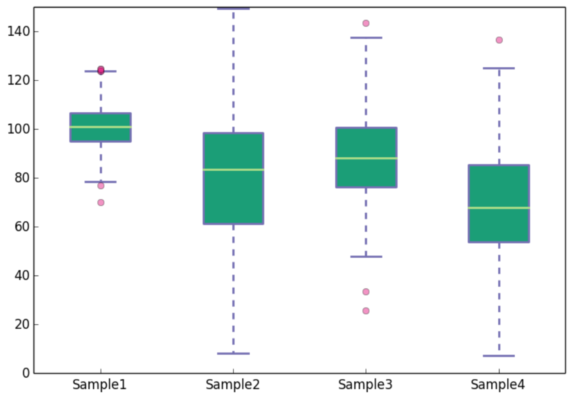
Multiple histograms

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



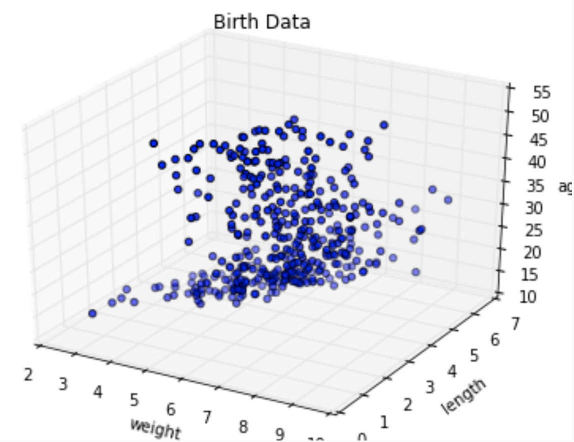
[Not] Anything is possible!

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?)

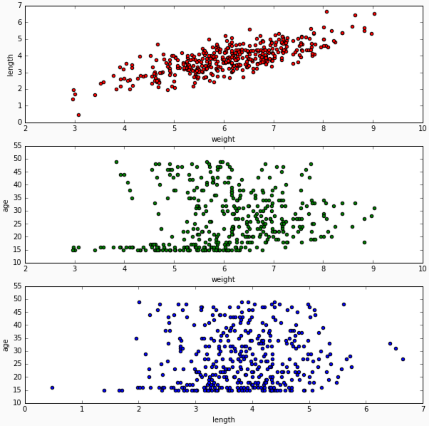
More dimensions not always better

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



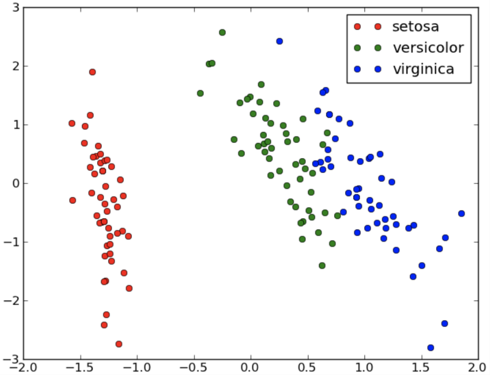
Reducing complexity

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



Adding a dimension

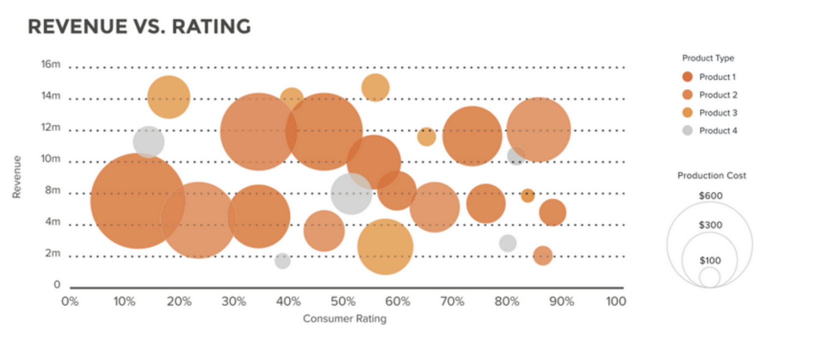
For 3D data, color coding a categorical attribute can be effective.



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

3D 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.

An Example

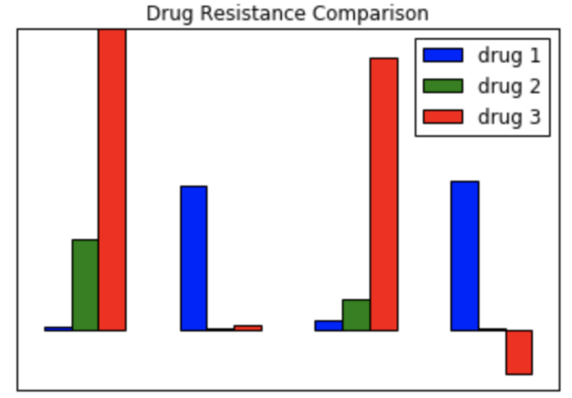
An example

Use some simple visualizations to explore 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

An example

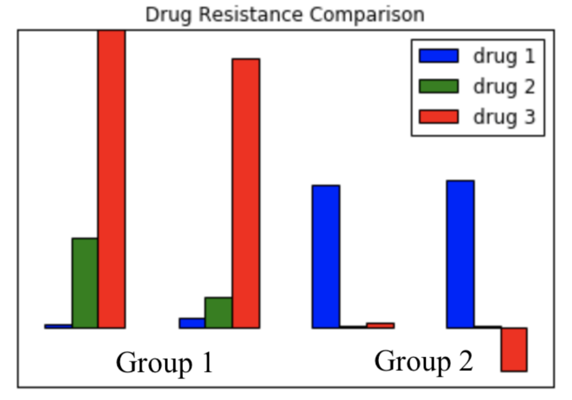
A bar graph showing resistance of each bacteria to each drug:



What do you notice?

An example

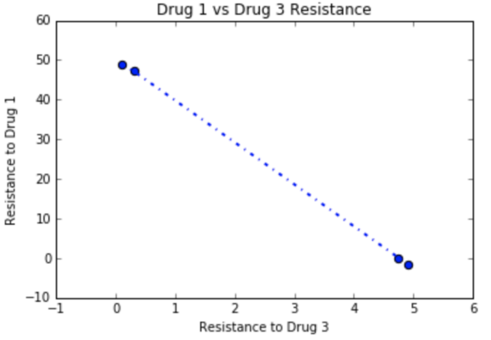
Bar graph showing resistance of each bacteria to each drug (grouped by Group Number):



Now what do you notice?

An example

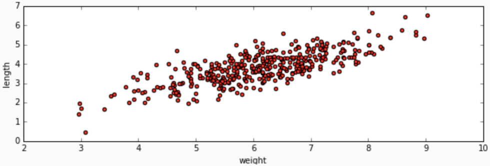
Scatter plot of Drug #1 vs Drug #3 resistance:



Key: the process of data exploration is iterative (visualize for trends, re-visualize to confirm)!

Quantifying Relationships

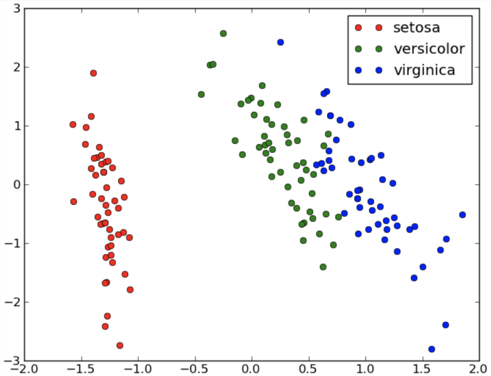
We can see that birth weight is positively correlated with femur length.



Can we quantify exactly how they are correlated? Can we predict birthweight based on femur length (or vice versa) through a statistical model?

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?