LECTURE #1: EXPLORATORY DATA ANALYSIS

CS 109A, STAT 121A, AC 209A: Data Science

Weiwei Pan, Pavlos Protopapas, Kevin Rader

Fall 2016

Harvard University

ANNOUNCEMENTS

- How to optimally communicate with us (Piazza then help line)
- Issues with Gutenberg in HW0 (for now **stop using Gutenberg**)
- Extension students: labs stream on Friday (posted same day), interactive lab via zoom, TF office hours, extra day on HW
- Cross-registered students already have or will get full access to Canvas today (**please check!**)
- HW0 must be submitted (**through Vocareum**)!
- HW0 submission deadline extended to Wednesday (today) at 11:59pm.
- Time and effort required to complete HW0 will vary depending on familiarity with programming (this is the learning curve), and is a good indicator of fit of the class (the programming will not get easier)
- HW1 is released

THE DATA SCIENCE PROCESS

On the first day of class you were introduced to the "data science" process.

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

Note: This process is not linear!

LECTURE OUTLINE

What Is Data?

Exploring Data

- Descriptive Statistics
- Data Visualization

An Example

What Next?

WHAT IS DATA?

WHAT IS DATA?

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

Provocative claim: everything is (can be) data!

WHERE DOES IT 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 DOES IT 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 don't want to pay to use the API or the database.
- **How do you it?** See HW0
- **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're bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

WHAT DOES IT LOOK LIKE?

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

WHAT DOES IT LOOK LIKE?

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: Weiwei
  - Last: Pan
  - Classes: [CS109A, STAT121A, AC209A]

WHAT DOES IT LOOK LIKE?

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.

WHAT DOES IT LOOK LIKE?

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

- Textual Data
- Temporal Data
- Geolocation Data

MORE ON TABULAR DATA

In tabular data, we expect each record or **observation** to represent a set of measurements of a single object or event.

	Hight	Radius	Do I Like It?
Cylinder # 1	10	5	Yes
Cylinder # 2	3	7.5	No

Each type of measurement is called a **variable** or an **attribute** of the data (e.g. Height, Radius and "Do I Like It?" are variables or attributes). The number of attributes is called the **dimension** of the data.

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 cylinders).

MORE ON TABULAR DATA

You'll see later that it's 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
  - ⇒ **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

IS THE DATA ANY GOOD?

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 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

We're 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.

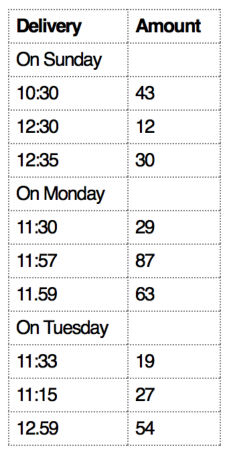
We need to reorganize the information to make explicit the event we're observing and the variables associated to this event.

Delivery	Time	Day	No. of Produce
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	90
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

HANDLING MESSY DATA

What object or event are we measuring?

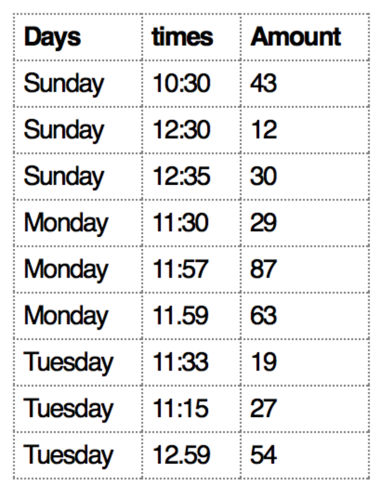
What are the variables in this dataset?



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

HANDLING MESSY DATA

We’re 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

HANDLING MESSY DATA

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**.

EXPLORING DATA

TALK OUTLINE

What Is Data?

Exploring Data

- Descriptive Statistics
- Data Visualization

An Example

What Next?

BASIC TERMS

Population versus sample:

- **Population** is the entire set of objects or events under study. Population can be hypothetical “all students” or all students in this class.
- **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.

Biases in samples:

- **Selection:** some subjects or records are more likely to be selected
- **Volunteer/nonresponse:** subjects or records who are not easily available are not represented
  For example: I usually only hear from students for whom something has gone terribly wrong in the course.

DESCRIBING DATA

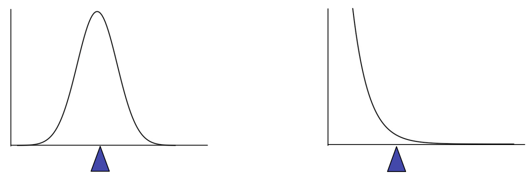
Given some large dataset, we'd like to compute a few quantities that intuitively summarizes the data. To begin with we'd like to know

- what are typical values for our variables or attributes?
- how representative are these typical values?

CENTRALITY

The **mean** of a set of n number of samples of a variable is denoted $\overline{x}$ and is defined by

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



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

CENTRALITY

The **median** of a set of *n* number of samples, ordered by value, of a variable is is defined by

$$\text{Median} = \begin{cases} x_{\lfloor n/2 \rfloor + 1}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

Example:

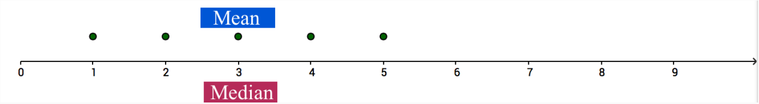
Ages: $17, 19, 21, \underline{22, 23,}\ 23, 23, 38$

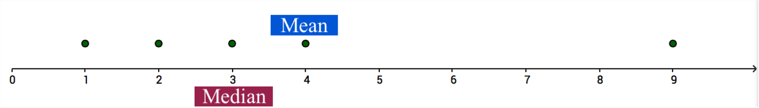
Median $= \frac{22+23}{2} = 22.5$

The median describes what a “typical” sample looks like, or where is the “center” of the distribution of the samples.

CENTRALITY

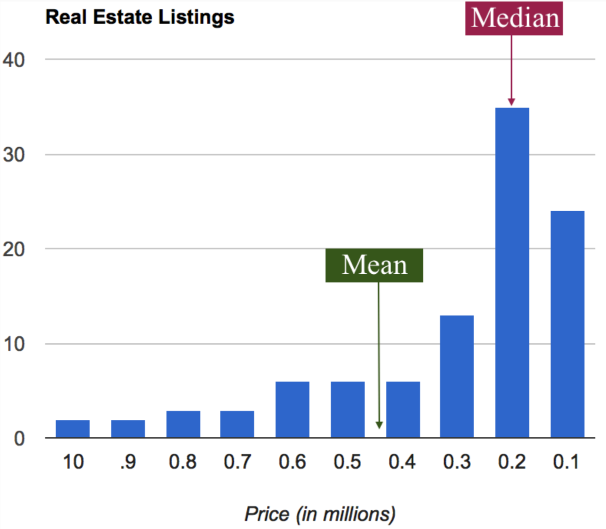
The mean is *sensitive to outliers*.





CENTRALITY

The mean is *sensitive to skewness (asymmetry) of distributions.*



CENTRALITY

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

- the mean
- the median

CENTRALITY

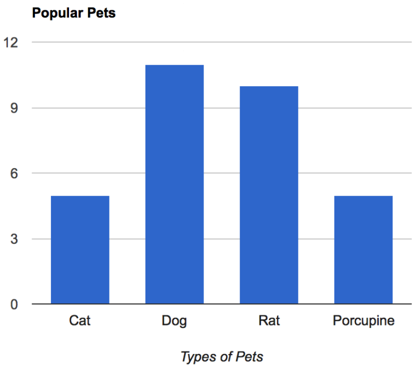
How hard (in terms of algorithmic complexity) is it to calculate

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

Note: Practicality of implementation has to be considered!

CENTRALITY

For samples of categorical variables, neither mean or median make sense.



The **mode** might be a better way to find the most “representative” value.

SPREAD

The spread of samples measures how well the mean or median describes the sample set.

One way to measuring spread of a set of samples is via the **range**.

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

SPREAD

The (sample) **variance**, denoted s^2, measures how much on average the sample values "deviates" from the mean

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

Note: the term $|x_i - \overline{x}|$ measure the amount by which x_i deviates from the mean $\overline{x}$. Squaring these deviation means that s^2 is sensitive to extreme values (outliers).

Note: s^2 doesn't have the same units as x_i! What does a variance of $1,008$ mean? Or 0.0001?

SPREAD

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

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

Note: s has the same units as x_i!

TALK OUTLINE

What Is Data?

Exploring Data

- Descriptive Statistics
- Data Visualization

An Example

What Next?

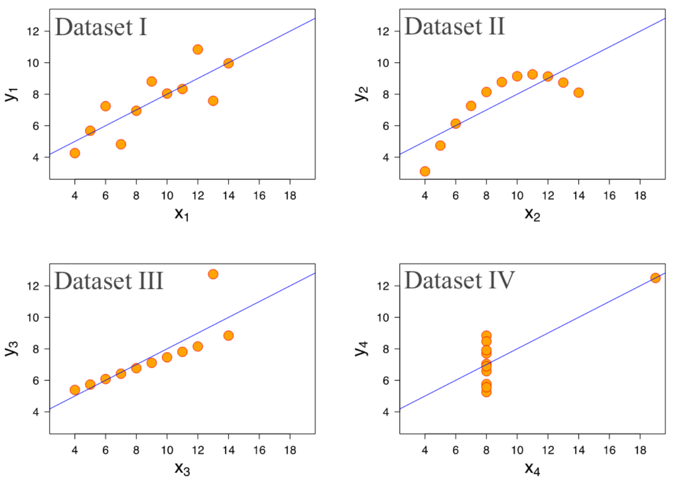
WHY DATA VISUALIZATION?

The following 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

WHY DATA VISUALIZATION?

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



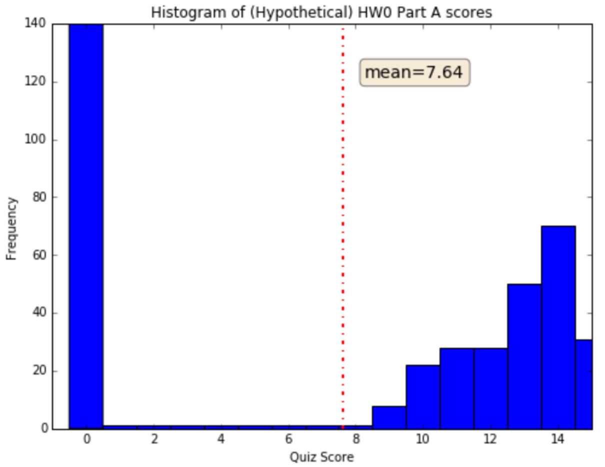
WHY DATA VISUALIZATION?

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

What does that suggest?

WHY DATA VISUALIZATION?

If I then show you the following graph, what does it suggest?



WHAT IS DATA VISUALIZATION GOOD FOR?

Analyze:

- Identify hidden patterns and trends
- Help formulate/test hypothese
- Help determine the next step in analysis/modeling

WHAT IS DATA VISUALIZATION GOOD FOR?

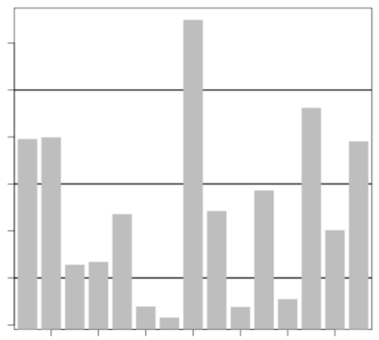
Communicate:

- Present information and ideas succinctly
- Provide evidence and support
- Influence and persuade

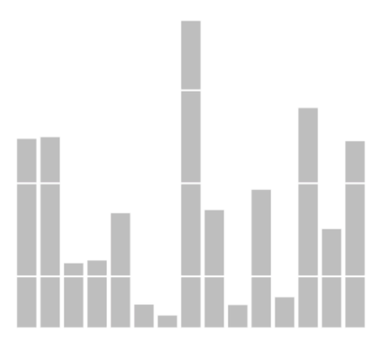
Basic data visualization guidelines from Edward Tufte:

- Maximize data to ink ratio: show the data

Bad



Better

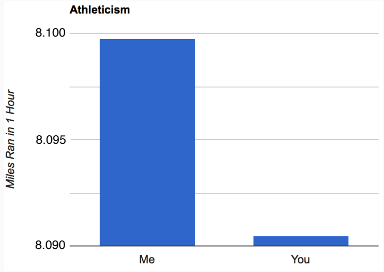


VISUALIZATION DESIGN PRINCIPLES

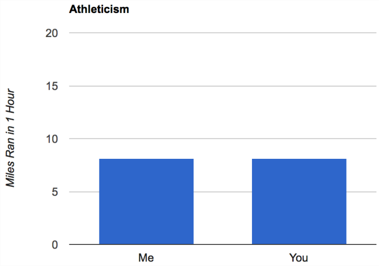
Basic data visualization guidelines from Edward Tufte:

- Maximize data to ink ratio: show the data
- Don't lie with scale: minimize $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$ (Lie Factor)

Bad



Better

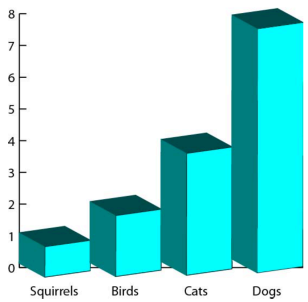


VISUALIZATION DESIGN PRINCIPLES

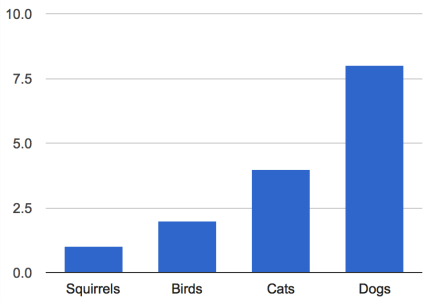
Basic data visualization guidelines from Edward Tufte:

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

Bad



Better



The number of visual parameters should not exceed the dimension of the data!

VISUALIZATION DESIGN PRINCIPLES

Basic data visualization guidelines from Edward Tufte:

- Maximize data to ink ratio: show the data
- Don't lie with scale: minimize $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$ (Lie Factor)
- Minimize chart-junk: show data variation, not design variation
- Clear, detailed and thorough labeling (including important events)

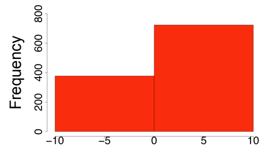
TYPES OF DATA VISUALIZATIONS

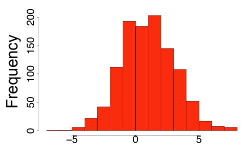
What do you want your visualization to show about your data?

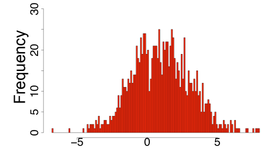
- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how the dataset breaks down into subgroups
- **Comparison:** how trends in multiple variable or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

DISTRIBUTION

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



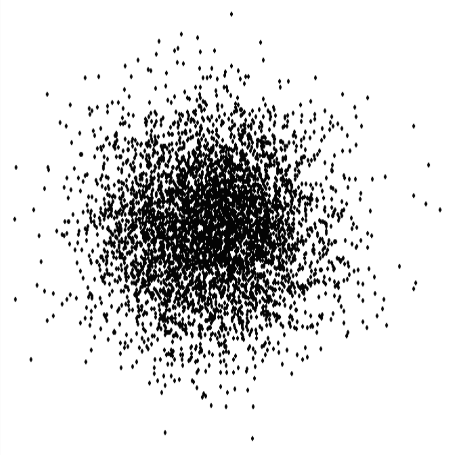




Note: Trends in histograms are sensitive to number of bins.

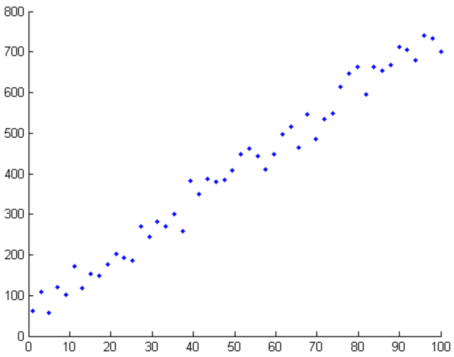
DISTRIBUTION

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



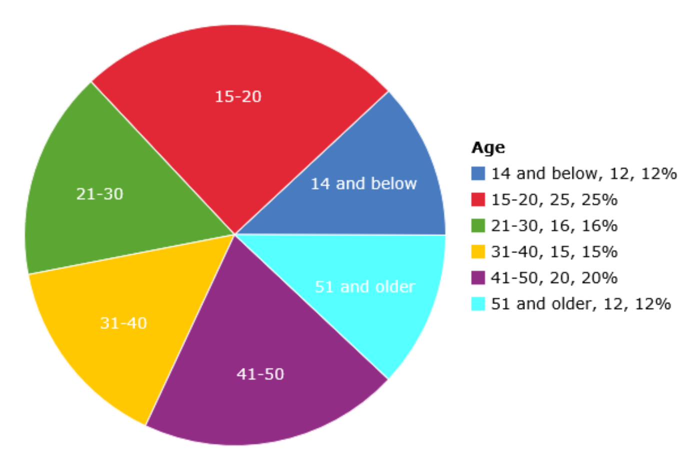
RELATIONSHIPS

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



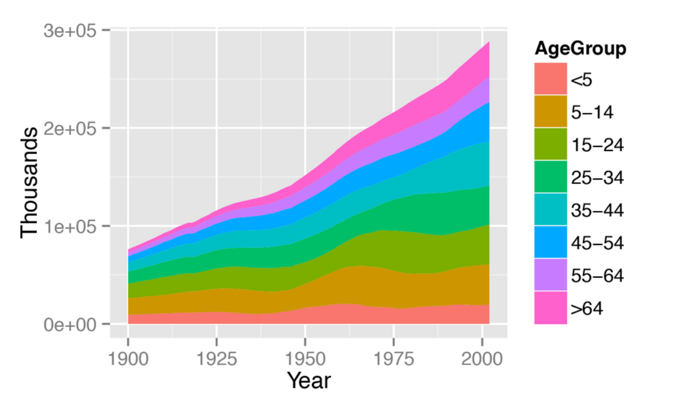
COMPOSITION

A **pie chart** is a way to visualize the static composition of a group.



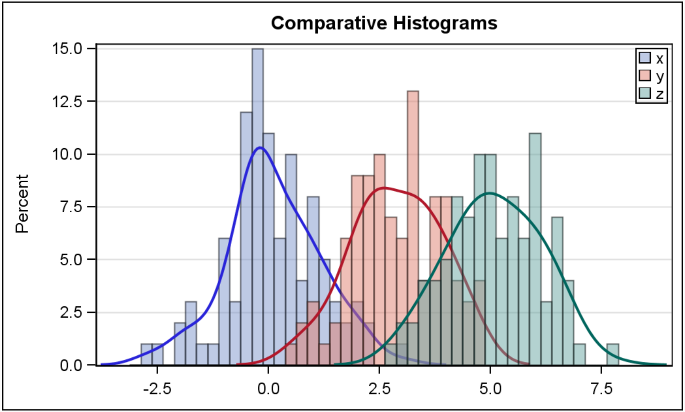
COMPOSITION

A **stacked area graph** is a way to visualize the composition of a group as it changes over time.



COMPARISONS

Plotting multiple histograms or curves on the same axes is a way to visualize how different variables compare.



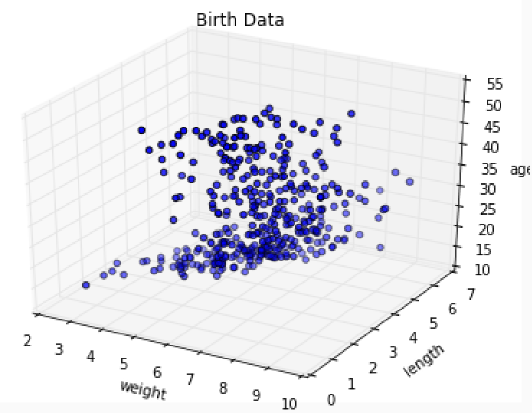
VISUALIZING THE IMPOSSIBLE

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”?)

REDUCING THE DIMENSION

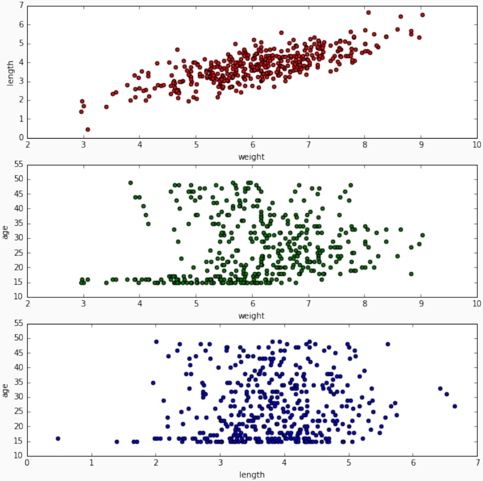
When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful.



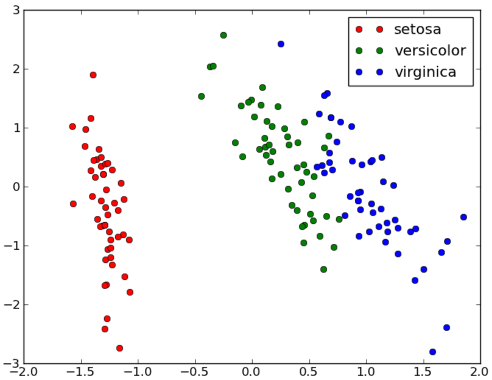
The above is the data from Homework set #0!

REDUCING THE DIMENSION

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



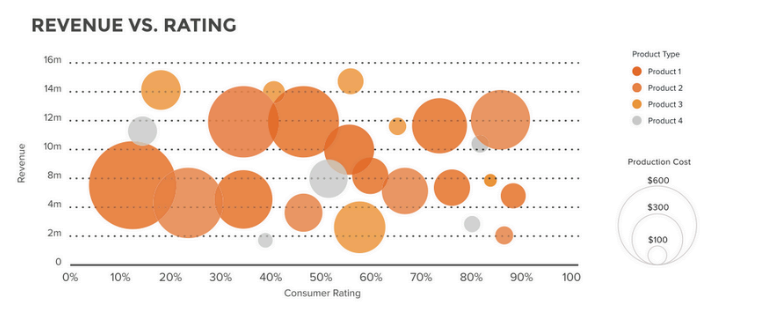
For 3D data, *color coding* a categorical attribute can be effective.



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

ADDING EXTRA DIMENSIONS

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

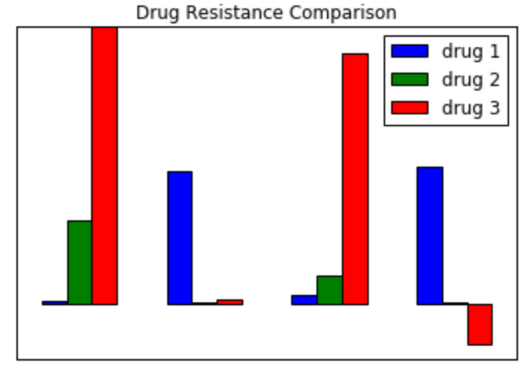
EFFECTIVENESS OF DRUGS

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

EFFECTIVENESS OF DRUGS

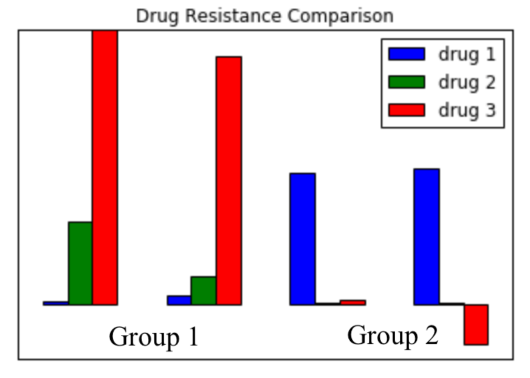
Bar graph showing resistance of each bacteria to each drug:



Any patterns?

EFFECTIVENESS OF DRUGS

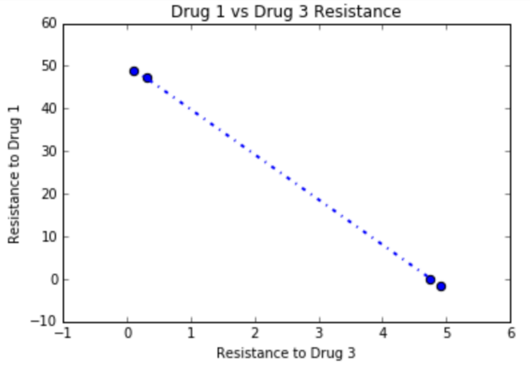
Bar graph showing resistance of each bacteria to each drug (grouped by Group Number):



Any patterns?

EFFECTIVENESS OF DRUGS

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

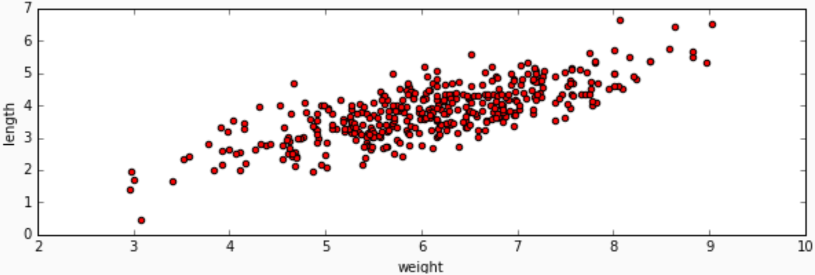


Note: The process of data exploration is iterative (visualize for trends, re-visualize to confirm)!

WHAT NEXT?

EXPLAIN

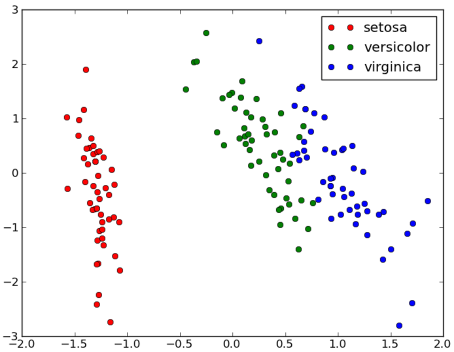
We can see that birth weight is positively correlated with femur length.



Can we describe exactly how they are correlated?

PREDICT

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?