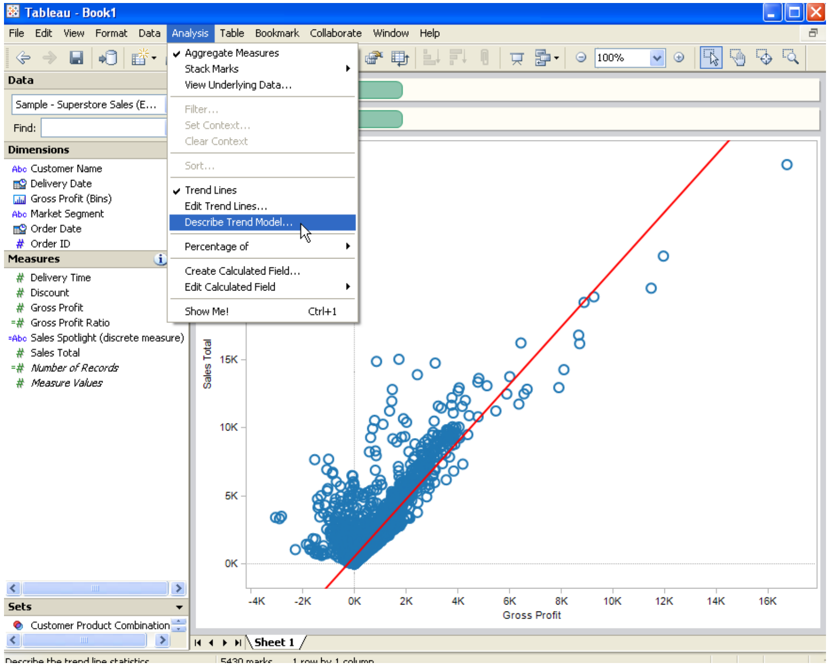


IAT 355
Visual Analytics

It's all about the numbers:
Data and Statistical Models

Lyn Bartram

SCHOOL OF INTERACTIVE
ARTS + TECHNOLOGY

Administrivia

- Teams/Project
- Assignment 1
  - Datasets
- Project areas

DATA ??

- Data Models
- Types
- Metadata

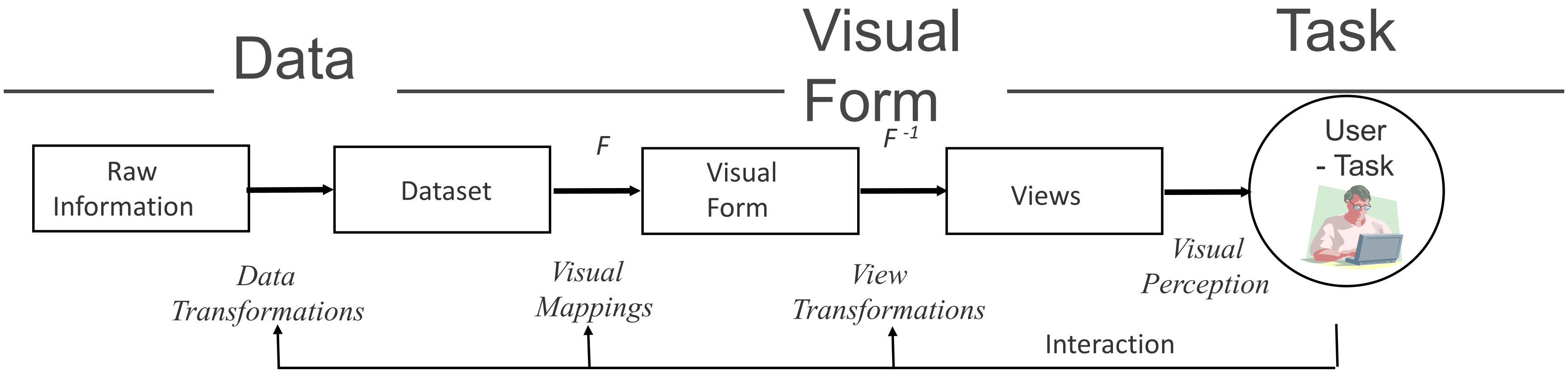
- What they tell us
- Descriptive Statistics
  - Distribution
  - Clusters
- Inferential Statistics
  - Trends
  - Patterns
  - (co)relations

Data and Data Sets

- *Data* are facts and figures collected, summarized, typed, analyzed, and interpreted.

- Collected/organized data are referred to as a *data set.*

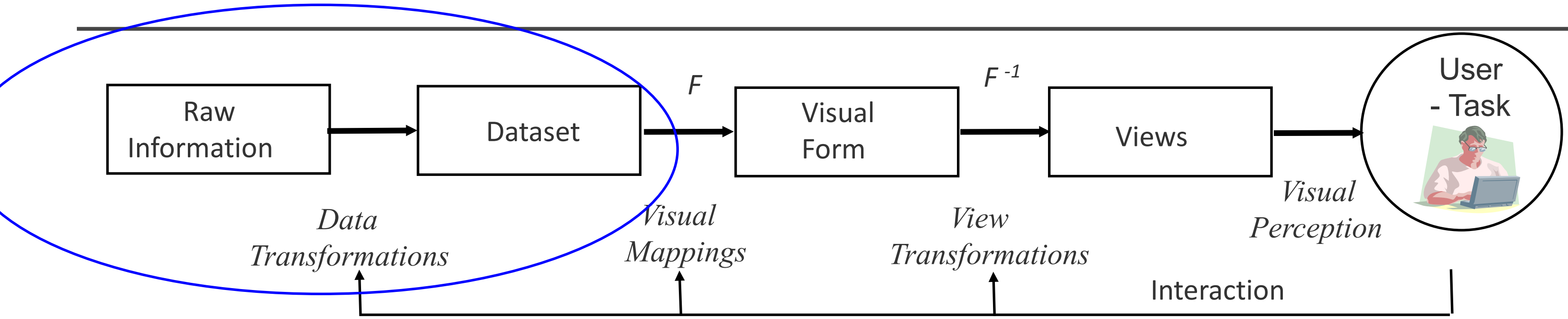
Visualization Stages



- House data
  - Price
  - Type
  - #bedrooms
  - Neighbourhood
  - …

Visualization Stages

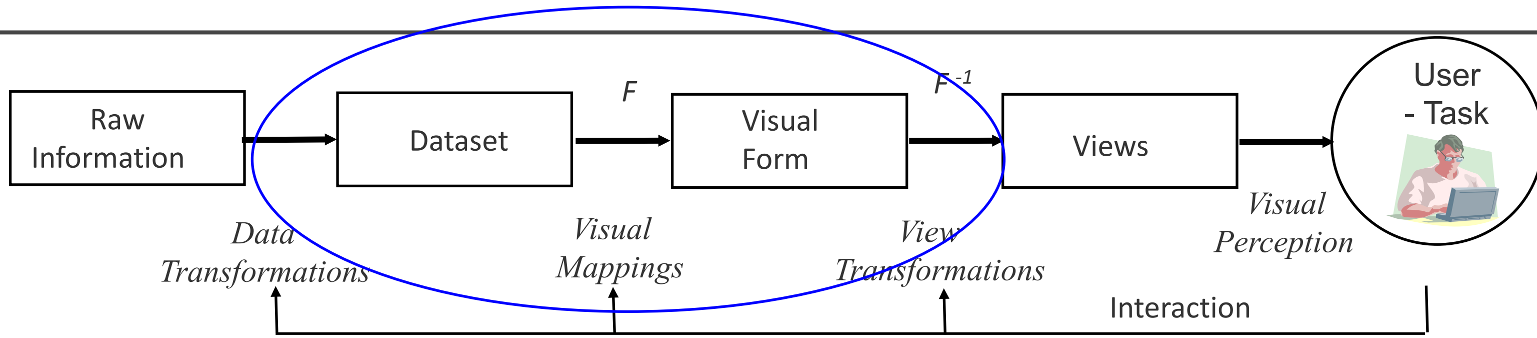
Data transformation – create a structural model (schema)



- Data transformation
  - Map raw data into data model/form (set properties)
  - Choose data
  - House type – category (text)
  - Price – currency
  - ..
  - Location (geocodes) - neighbourhood

Visualization Stages

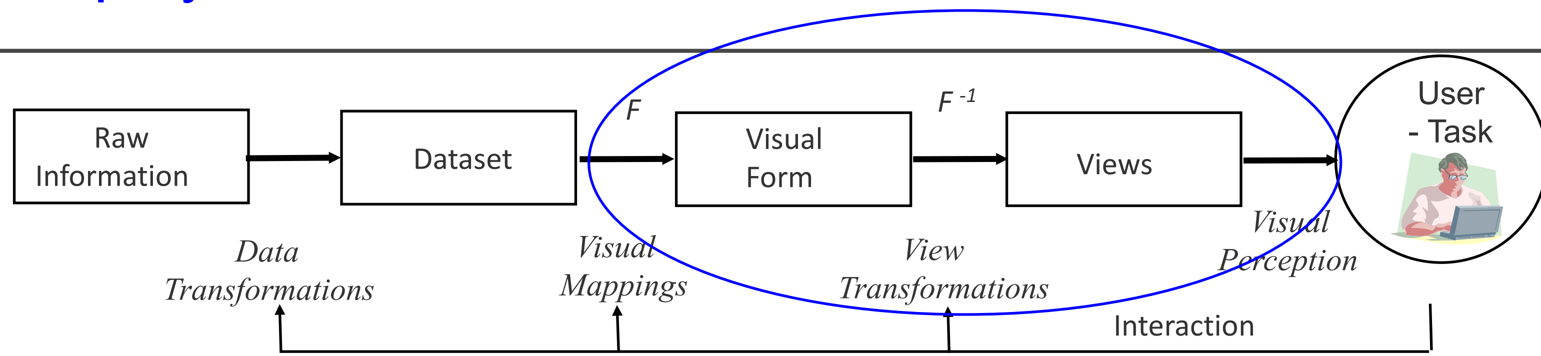
Visual mapping– create a visual spatial model



- Data transformation
  - Map raw data into data tables – e.g. text to similarity matrix
- Visual Mappings:
  - Transform data tables into visual structures
  - e.g., house price, #bedrooms to 2 dims – x, y

Visualization Stages

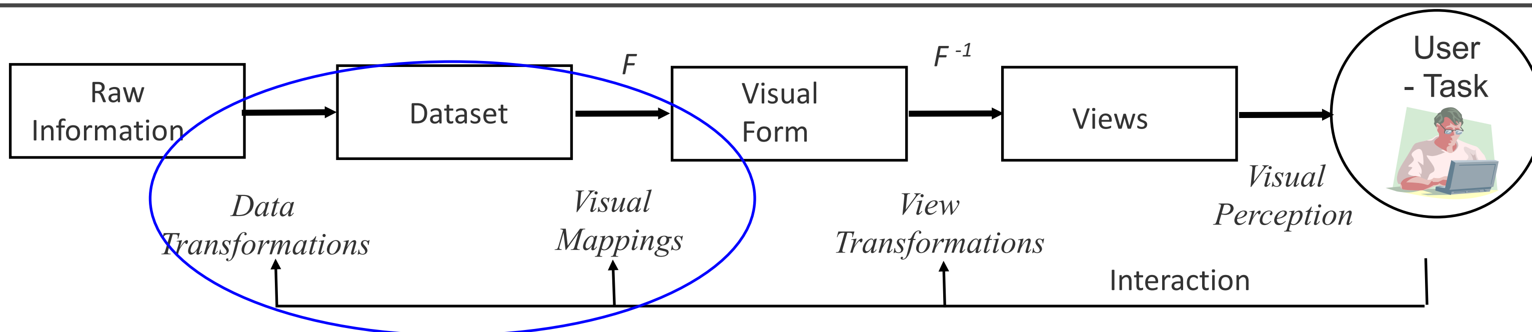
Display the data that now have visual form



- Data transformation
  - Map raw data into data tables – e.g. text to similarity matrix
- Visual Mappings:
  - Transform data tables into visual structures – e.g. 2 dims – x, y
- View Transformations:
  - Create views of the Visual Structures by specifying graphical forms, combinations

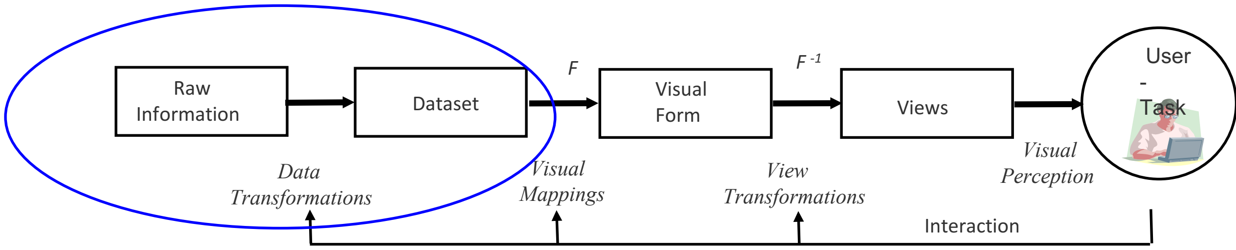
Visualization Stages

The user may change transformations and mappings



- Data transformation
  - Map raw data into data tables – e.g. text to similarity matrix
- Visual Mappings:
  - Transform data tables into visual structures – e.g. 2 dims – x, y
- View Transformations:
  - Create views of the Visual Structures by specifying graphical forms, combinations

Data models



- take raw data and transform it into a form that is more workable
  - Main idea: build a *model*
- Individual items are *cases*, *records, elements*
- Cases have *attributes* :
  - an attribute is also called a *variable, factor or observation*
  - In vis terms, a *dimension*
- The *value* of an *dimension* may differ for each *case*
- The *schema* is the "blueprint" of the data model and describes how the data are organized.

Terminology: relational (tabular) data

Data set

Record, case, item

Dimension, Property, Attribute, Variable, factor

Name	Mfr	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sug
Cheerios	General Mills	110	6	2	290	2.00000	17.0000	
Cinnamon Toast Crun...	General Mills	120	1	3	210	0.00000	13.0000	
Count Chocula	General Mills	110	1	1	180	0.00000	12.0000	
Honey Nut Cheerios	General Mills	110	3	1	250	1.50000	11.5000	
Maypo	AlphaBits	100	4	1	0	0.00000	16.0000	
Raisin Bran	Kellogg	120	3	1	210	5.00000	14.0000	
Rice Krispies	Kellogg	110	2	0	290	0.00000	22.0000	
Shredded Wheat	NutriGrain	80	2	0	0	3.00000	16.0000	
Special K	Kellogg	$\boxed{110}$	6	0	230	1.00000	16.0000	
Special K	Kellogg	110	6	0	230	1.00000	16.0000	
Special K	Kellogg	110	6	0	230	1.00000	16.0000	
Cinnamon Toast Crun...	General Mills	120	1	3	210	0.00000	13.0000	

Value

Unique Attribute/identifier

Terminology:

Levels (range) of a dimension

Name	Mfr	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sug
Cheerios	General Mills	110	6	2	290	2.00000	17.0000	
Cinnamon Toast Crun...	General Mills	120	1	3	210	0.00000	13.0000	
Count Chocula	General Mills	110	1	1	180	0.00000	12.0000	
Honey Nut Cheerios	General Mills	110	3	1	250	1.50000	11.5000	
Maypo	AlphaBits	100	4	1	0	0.00000	16.0000	
Raisin Bran	Kellogg	120	3	1	210	5.00000	14.0000	
Rice Krispies	Kellogg	110	2	0	290	0.00000	22.0000	
Shredded Wheat	NutriGrain	80	2	0	0	3.00000	16.0000	
Special K	Kellogg	110	6	0	230	1.00000	16.0000	
Special K	Kellogg	110	6	0	230	1.00000	16.0000	
Special K	Kellogg	110	6	0	230	1.00000	16.0000	
Cinnamon Toast Crun...	General Mills	120	1	3	210	0.00000	13.0000	

Data

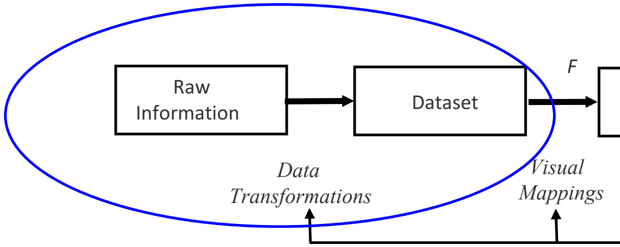
- Data Models
- Types
- Metadata
- Aggregates

- Descriptive Statistics
- Inferential Statistics
  - Distribution
  - Clusters

Transforming Data

- **Data implementation models are low level descriptions**
  - Storage, low-level functions
- **Conceptual models are mental constructions**
  - Include semantics and support reasoning
- **Data types reflect how the data can be used**
  - (1D floats) vs. Temperature vs. “Hot,Warm,Cold”
  - (3D vector of floats) vs. Space vs “Near,far, top, bottom..”

Data Models

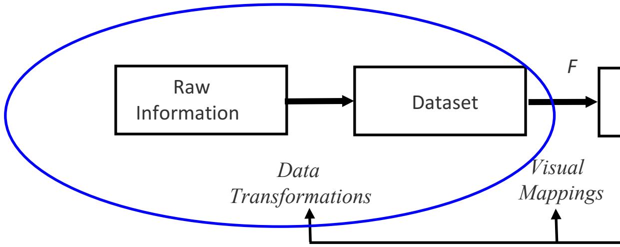


Abstract	Conceptual	Conceptual
• Low level • Numeric • Computational	• framework • Have meaning attached • Mapped into framework	• Semantic • Interpreted by reasoning
• 26. 53 (1D float) • {255,0,0}	• 79.754 ° (temp) • 	• Cool, WARM, hot • Bright red

From data model to type

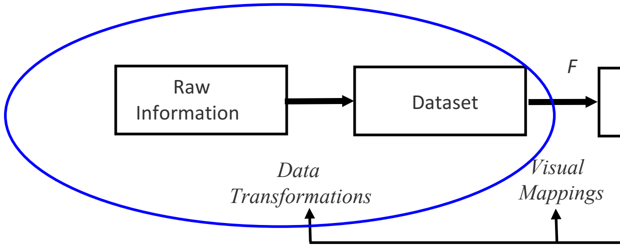
- Data storage model
  - 32.5, 54.0, -17.3, …
  - floats
- Conceptual model
  - Temperature (° C)
- Data type
  - Burned vs. Not burned (Nominal)
  - Hot, warm, cold (Ordinal)
  - Continuous range of values (Quantitative: C or D)

Data Types



- **Nominal**: categorical
  - Example: gender, Student Number
  - No concept of relative relation other than inclusion in the set
  - **$=$, $\in$, $\notin$**
- **Ordinal** : sequential (ordered set)
  - Example: Size of car, speed settings on road
  - Example: mild, medium, hot, suicide
  - Distance is **not uniform**
  - **$>$, $<$, $\leq$, $\geq$**

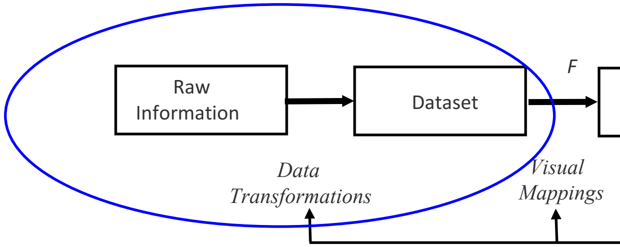
Data Types



- **Interval** : Relative measurements, no fixed zero point.
  - Data can be reasoned about numerically
  - Example: height above sea level, hours in a day
  - Distance is uniform : -2, -1, 0, +1, +2
  - Add, subtract operators (2 days away = +)

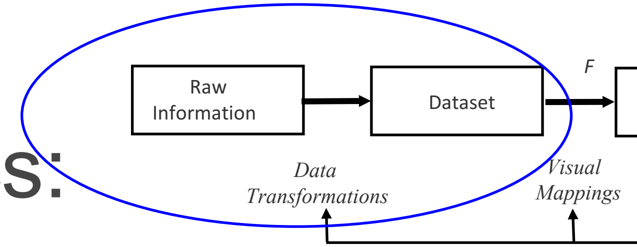
- **Ratio**: (absolute zero)
  - Ratio of two values is meaningful
  - Example: account balance
  - **Usually sampled level** (eg, nearest decimal)
  - Full arithmetic functions

Data types



- Quantitative i.e. numerical
  - Continuous (e.g. pH of a sample, patient cholesterol levels)
  - Discrete (e.g. number of bacteria colonies in a culture)
- Non-quantitative (! just qualitative)
  - Nominal (e.g. gender, blood group)
  - Ordinal (ranked e.g. mild, moderate or severe illness).
  - Assigned not sensed or measured

Quantitative Data characteristics:



- Continuous
  - Data can take any value within the range
  - Number grade (92.75%)
- Discrete: data can take only certain values
  - Example: number of students in a class (no half students)
  - Letter grade (A+)

- Time
- Spatial

Data types

- **Quantitative** i.e. numerical
  - **Continuous** (number grade)
  - **Discrete** (e.g. number of students in a class)

- **Categorical**
  - **Nominal** (e.g. gender, country of origin)
  - **Ordinal** (level of programming ease: poor, good, great). Often ordinal variables are re-coded to be quantitative.
  - This assigns *weights* or priorities.

From data model to type

Sepal Length	Sepal Width	Petal Langth	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
6.1	3	4.9	1.8	virginica
6.4	2.8	5.6	2.1	virginica
7.2	3	5.8	1.6	virginica
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa

Q

N

Sepia and petal length for three species of iris [Fisher 1936]

Microsoft Excel - fischer.iris.2.colored.xls

	A	B	C	D	E	F	G
1	ID	Case	Species_No	Species	Organ	Width	Length
2	1	1	1	I. Setosa	Petal	2	14
3	2	1	3	I. Verginica	Petal	24	56
4	3	1	2	I. Versicolor	Petal	13	45
5	4	1	1	I. Setosa	Sepal	33	50
6	5	1	3	I. Verginica	Sepal	31	67
7	6	1	2	I. Versicolor	Sepal	28	57
8	7	2	1	I. Setosa	Petal	2	10
9	8	2	3	I. Verginica	Petal	23	51
10	9	2	2	I. Versicolor	Petal	16	47
11	10	2	1	I. Setosa	Sepal	36	46
12	11	2	3	I. Verginica	Sepal	31	69
13	12	2	2	I. Versicolor	Sepal	33	63
14	13	3	1	I. Setosa	Petal	2	16
15	14	3	3	I. Verginica	Petal	20	52
16	15	3	2	I. Versicolor	Petal	14	47
17	16	3	1	I. Setosa	Sepal	31	48
18	17	3	3	I. Verginica	Sepal	30	65
19	18	3	2	I. Versicolor	Sepal	32	70
20	19	4	1	I. Setosa	Petal	1	14
21	20	4	3	I. Verginica	Petal	19	51
22	21	4	2	I. Versicolor	Petal	12	40
23	22	4	1	I. Setosa	Sepal	36	49
24	23	4	3	I. Verginica	Sepal	27	58
25	24	4	2	I. Versicolor	Sepal	26	58
26	25	5	1	I. Setosa	Petal	2	13
27	26	5	3	I. Verginica	Petal	17	45
28	27	5	2	I. Versicolor	Petal	10	33
29	28	5	1	I. Setosa	Sepal	32	44
30	29	5	3	I. Verginica	Sepal	25	49
31	30	5	2	I. Versicolor	Sepal	23	50
32	31	6	1	I. Setosa	Petal	2	16

fischer.iris

Ready

Q

O

N

Relational (tabular) data model

- Represent data as a table (*relation*)
- Each row (*tuple*) represents a single record (case)
- Each record is a fixed length tuple
- Each column (attribute, dimension) represents a single *variable*
- Each attribute has a *name* and a *data type*
- A table's *schema* is the set of names and data types
- A *database* is a collection of tables (relations)

Data Table Format

Dimensions

	Case1	Case2	Case3
Variable1	Value11	Value21	Value31
Variable2	Value12	Value22	Value32
Variable3	Value13	Value23	Value33

- Think of this as a function
  - if(case1) = <Val11, Val12,…>

Example: Student Data

Case

Unique identifier

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	2.3	3.04

Example: Student Data

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	2.3	3.04

Attribute/Dimension

Example: What kinds of data? Types?

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	[illegible]	[illegible]

Date, interval, point, term – underlying data, current format

Example: drug trial experimental data

- Variables (dimensions) are classified as:
  - **Dependent.** Variable of primary interest (e.g. blood pressure in an antihypertensive drug trial). What we want to know about.
  - **Independent/Predictor**
    - **Attribute** controlled by experimenter (also called **factor**).

- These are experimental terms: how do they apply to analysis?

Data Wrangling

- Data comes in many different forms
- Typically, not in the way you want it

- Data concerns
  - Formats and types
    - Marshalling
    - joining
  - Structure and relations
  - Purpose /analytical aggregation

Tables of observations

Month	Control	Placebo	300 mg	450 mg
March	165	163	166	168
April	162	159	161	163
May	162	161	158	160
June	166	158	160	148
July	163	158	157	150

Blood Pressure Study (4 treatments, 5 months)

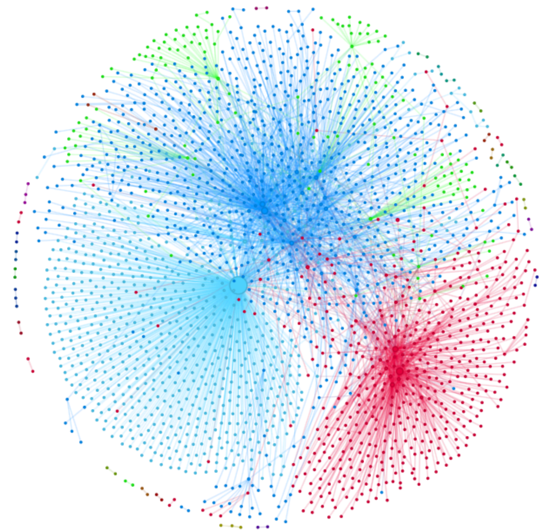
Refactoring and restructuring

Month	Treatment group	Measure
March	Control	165
March	Placebo	163
March	300 mg	166
March	450 mg	168

…..

Blood Pressure Study (4 treatments, 5 months)

Not just formats



Data wrangling

- Missing and bad values (data cleaning)

- Multiple data sources
- Problems of integration
  - Bank example, variance: naming convention, attributes for data item, account no, account type, size, currency

- inconsistency

1. Decide complete schema
2. Decide UNIQUE IDENTITIFIER
   - Index or id
3. Extract
4. Merge
5. Transform
6. load

But wait, there's more! Metadata

Mary	Tom	Louise
65432101	98765651	89846251
20	22	19
Sep 2006	Jan 2004	Sep 2005
4.0	2.3	3.04

- Descriptive information about the data
- Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example: if term >= 22, then GPA can only be above 3
- Missing values, uncertainty or importance are all examples of metadata

But wait…. There's more !

- Raw data
- Metadata
  - Data about the data
- **Frequency data**
  - "more than half the respondents smoked before 16"
  - clustering
- **Derived data**
  - Summaries, observations, inferences, predictions
  - "the odds of you getting ill from this pizza were 5 to 1"

Data analysis

- Qualitative
- **Descriptive**. Used to describe the distribution of a single variable or the relationship between two nominal variables (mean, frequencies, cross-tabulation)
- **Inferential** (Used to establish relationships among variables; assumes random sampling and a normal *distribution*)

Simplest descriptive: Min, Max, Range

- (73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100)

- Range
  - Difference between minimum and maximum values in a data set
  - Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.

Frequency analyses

basic type of descriptive statistic

- *Frequency Distribution* presents the counts of observations grouped within pre-specified groups

- *Relative Frequency Distribution* presents the corresponding proportions within the groups

- 40% of respondents are male.

- The mean level of income was $35,000

- 60% of younger voters cast their vote for Trudeau compared to 52% of voters over 50.

Understanding data: descriptive statistics

calculate the "average"

Central tendency measures

- mean
- median
- mode

calculate the "spread"

Dispersion measures

- range
- variance
- Standard deviation

Understanding data: descriptive statistics

calculate the "average"

Central tendency measures

- mean
- median
- mode

Measures of centrality allow us to summarize the dataset based on a central tendency

calculate the "spread"

Dispersion measures

- range
- variance
- Standard deviation

The “Average” ???

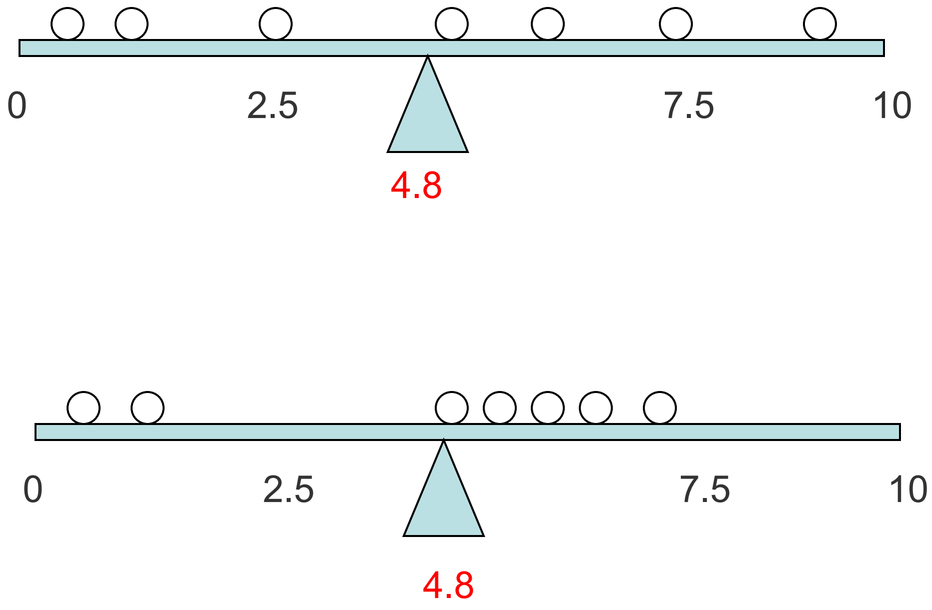
- The Average (Mean)
  - Sum of all values divided by the number of values in the data set.
  - One measure of central **location** in the data set.

Average=(73+66+69+67+49+60+81+71+78+62+53+87+74+65+74+50+85+45+63+100)/20 = 68.6

When might you not want to use the mean?

SFU

The mean is vulnerable to problems



The data may or may not be symmetrical around its average value

The Median

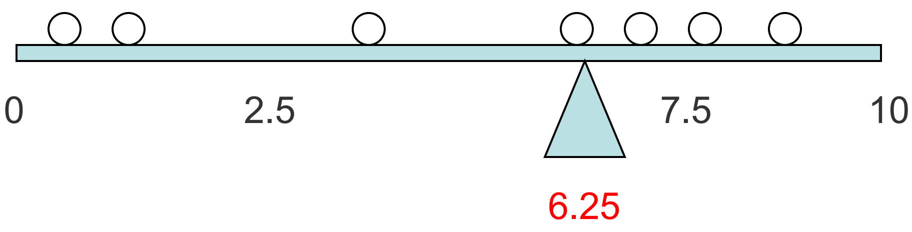
- The middle value in a sorted data set. Half the values are greater and half are less than the median.
- Another measure of central location in the data set.

(45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

Median: 68

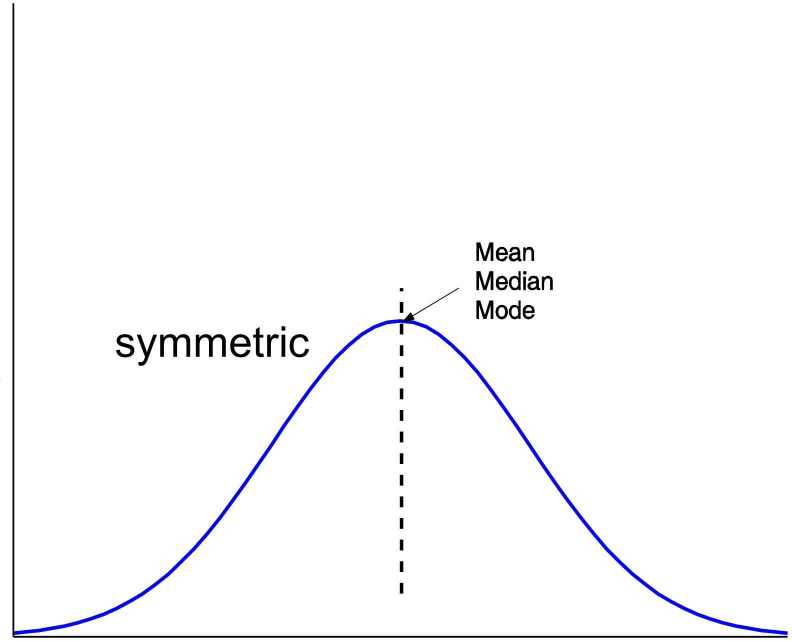
(1, 2, 4, 7, 8, 9, 9)

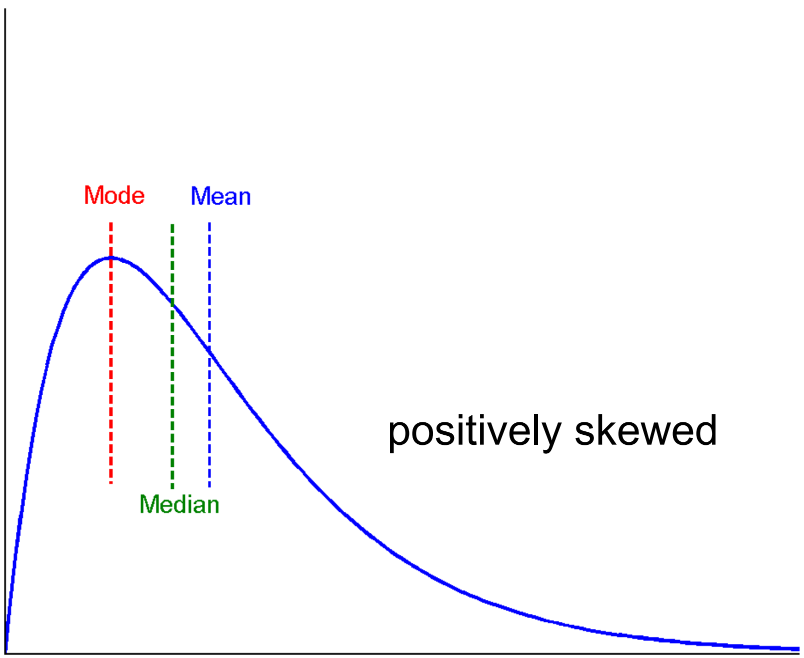
- The Median
  - May or may not be close to the mean.
  - Combination of mean and median are used to define the *skewness* of a distribution.

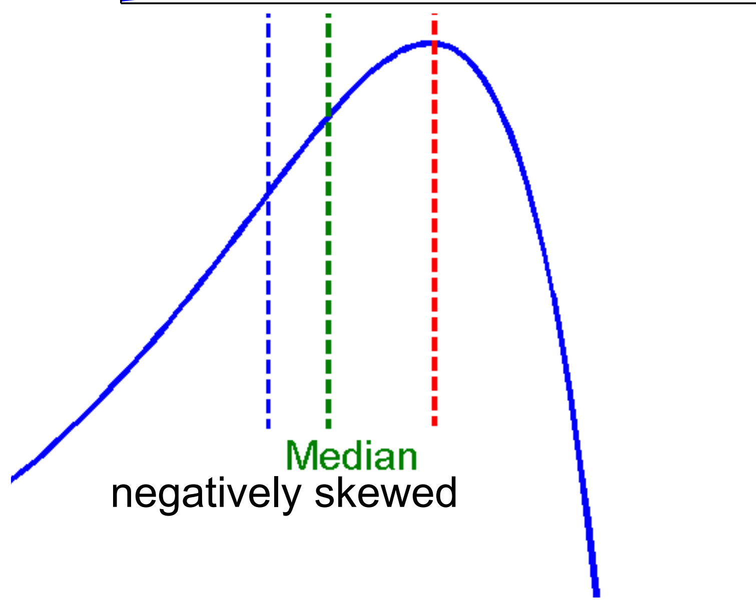


Distribution and symmetry

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







Adapted from Han, Kamber and Pei 2013

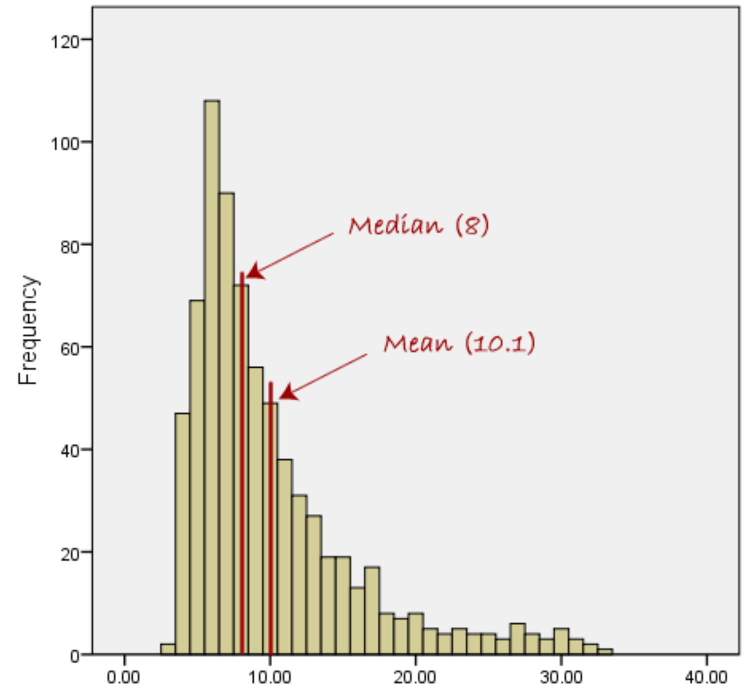
The Mode

- The most frequent occurring value.
- Another measure of central location in the data set.
- (45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)
- Mode: 74
  - Generally not all that meaningful unless a larger percentage of the values are the same number
  - BUT useful for nominal (categorical) data!
  - *Most common social media tool used by students is FB*

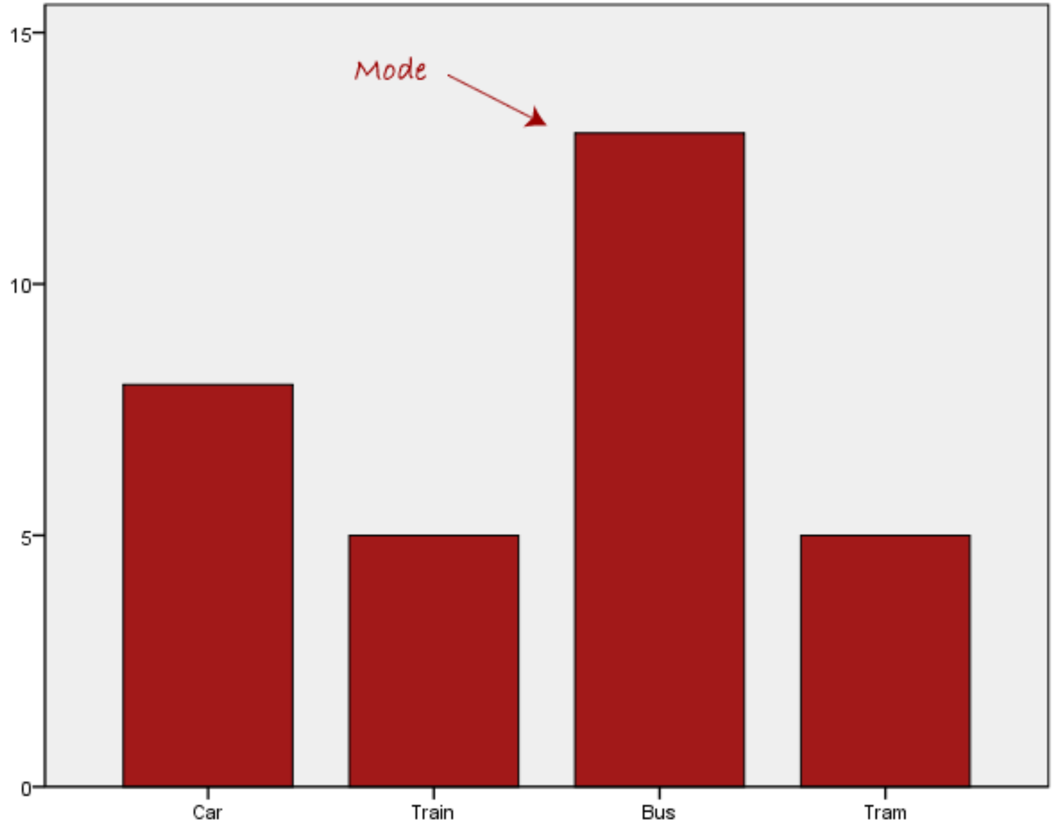
When do we use what?

- Dependent on how the data are **distributed**
  - Note if mean=median=mode then the data are said to be *symmetrical*

- Rule of thumb:
  - use mean if data are normally distributed and variance is within constraints
  - Use median to reduce effects of **outliers**



Mode for categorical frequency



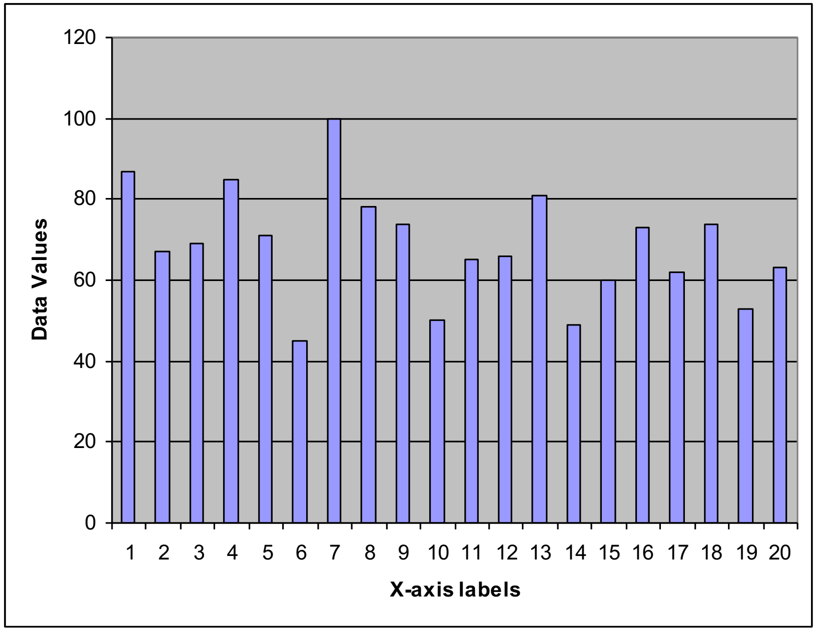
Summary

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

http://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

Centrality

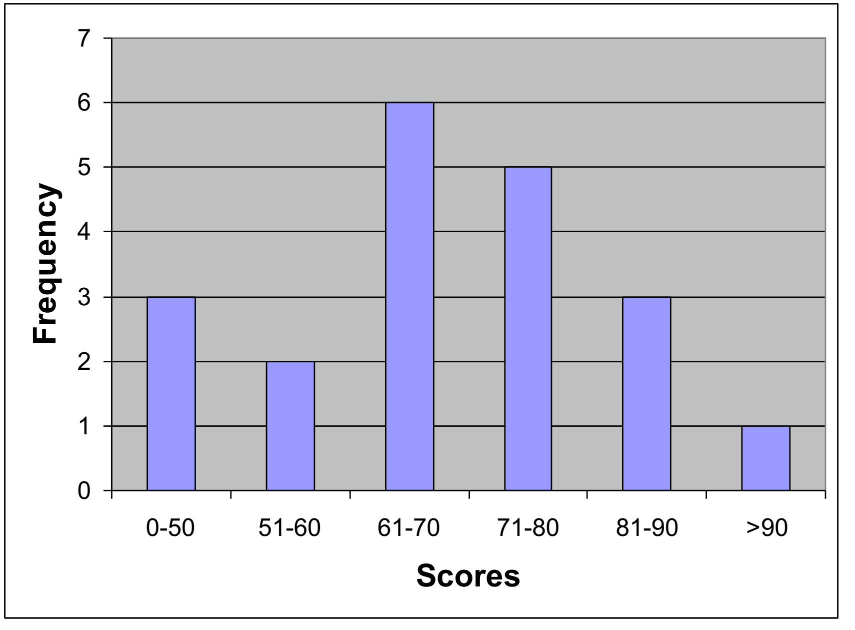
- We can’t really tell much about this data set
- Even Min and Max are hard to see



Plot the distribution

- Determine a frequency table (bins, buckets)
- A histogram is a column chart of the frequencies

Category Labels	Frequency
0-50	3
51-60	2
61-70	6
71-80	5
81-90	3
>90	1



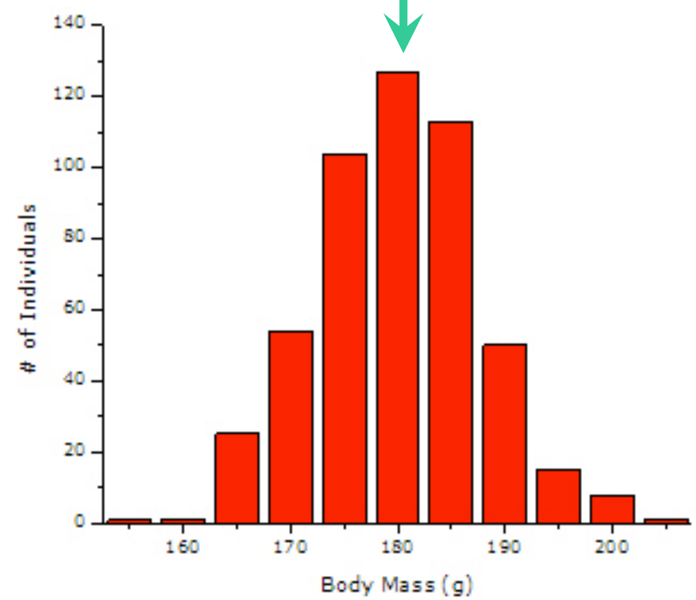
Histogram

Q → Q' → N,O

Most common form: split data range into equal-sized bins and count the number of points from the data set that fall into the bin.

- Vertical axis: Frequency (i.e., counts)
- Horizontal axis: binned variable

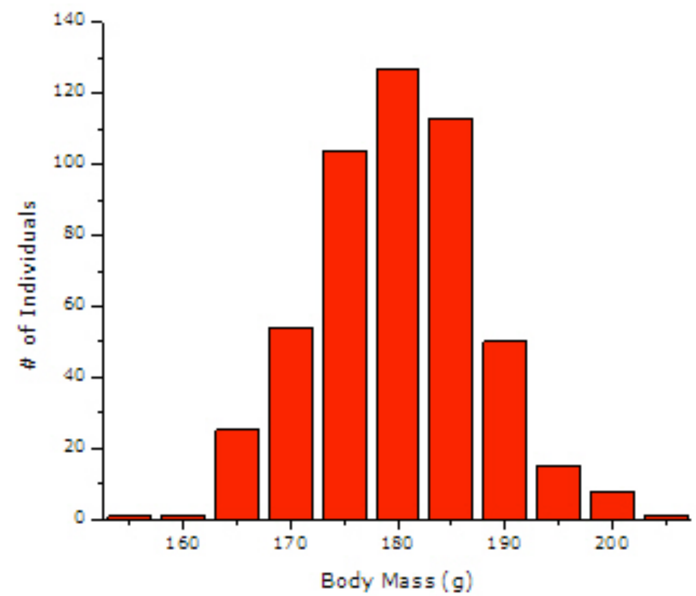
- Often use ordered bins



Histogram

The histogram graphically shows the following:

- Centrality (i.e., the location) of the data;
- spread (i.e., the scale);
- skewness;
- outliers; and
- multiple modes.

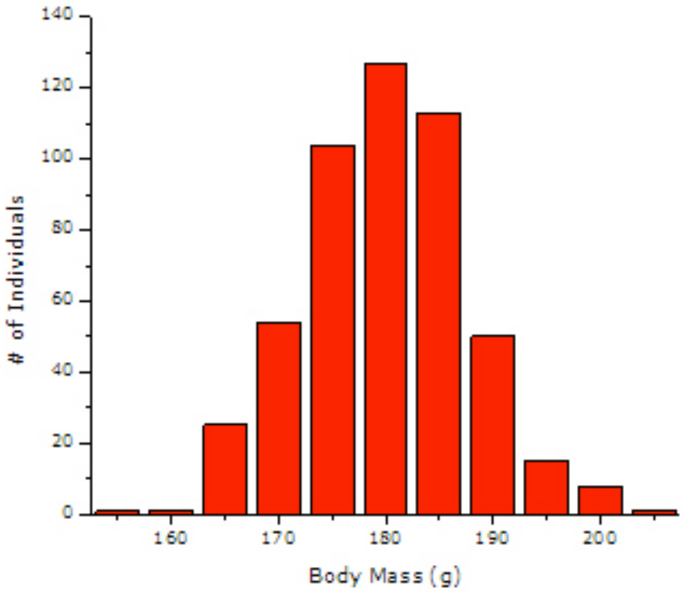


Discovering centrality

- Visualization can indicate which centrality measure to use(mean, median)

- Unimodal (one main central point
- Symmetric data

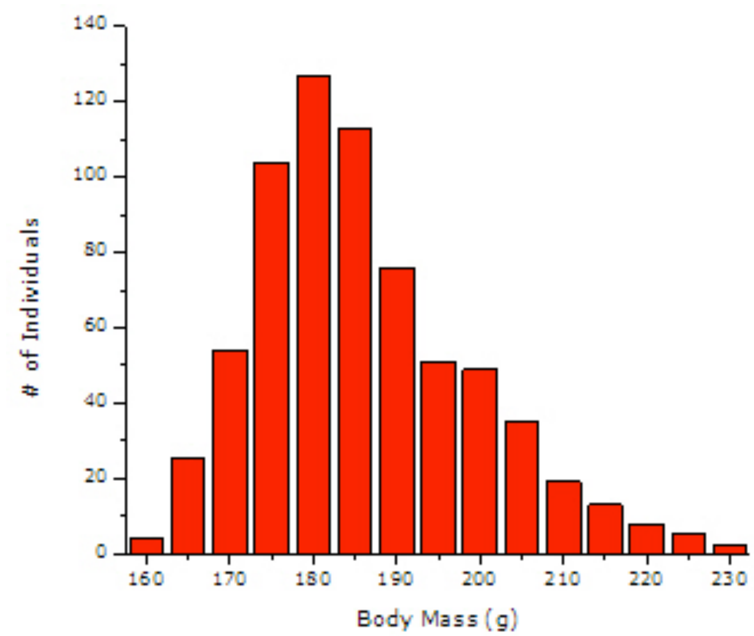
- Use the **mean**



http://www.sciencebuddies.org/science-fair-projects/project_data_analysis_summarizing_data.shtml

Discovering centrality

- Visualization can indicate which centrality measure to use(mean, median)
- Unimodal (one main central point)
- **Skewed** data
- Use the **median**

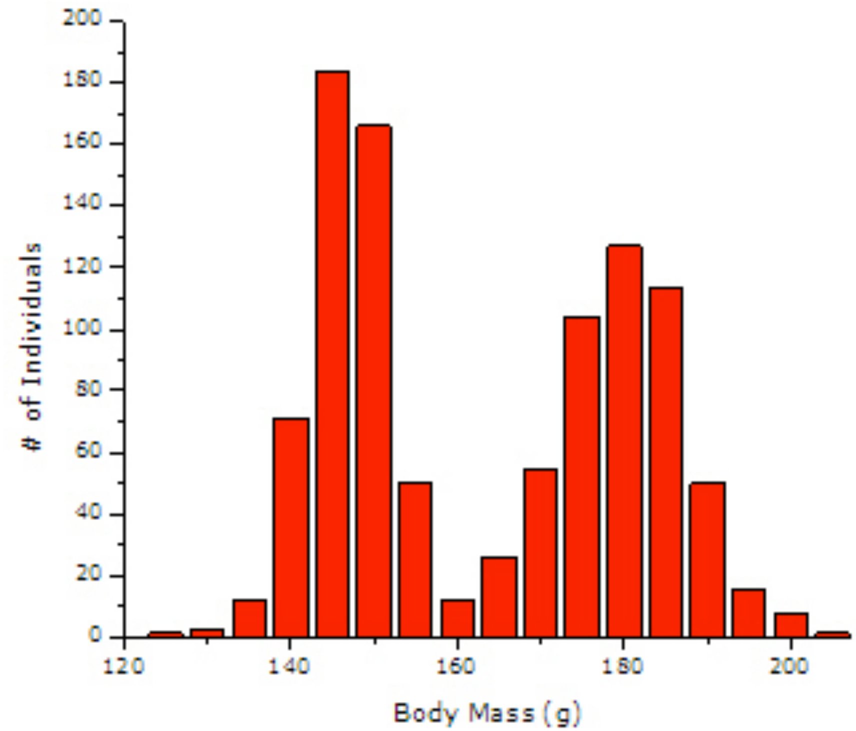


http://www.sciencebuddies.org/science-fair-projects/project_data_analysis_summarizing_data.shtml

Histograms quickly show distribution clusters

- **Bimodal** distribution
- two clusters/groups, each with its own separate central tendency.

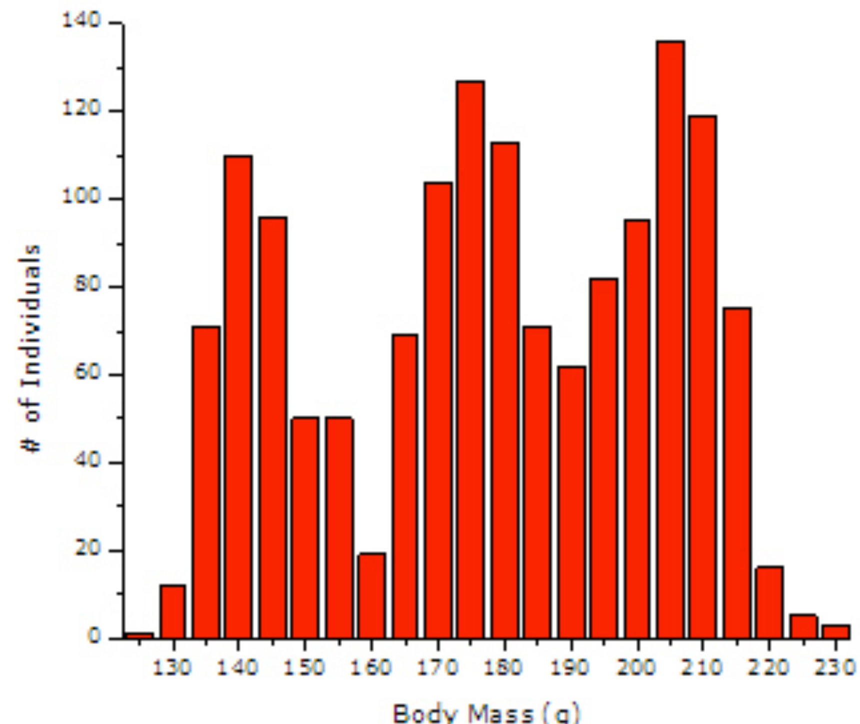
- Don’t use centrality, or at least not for overall data set



http://www.sciencebuddies.org/science-fair-projects/project_data_analysis_summarizing_data.shtml

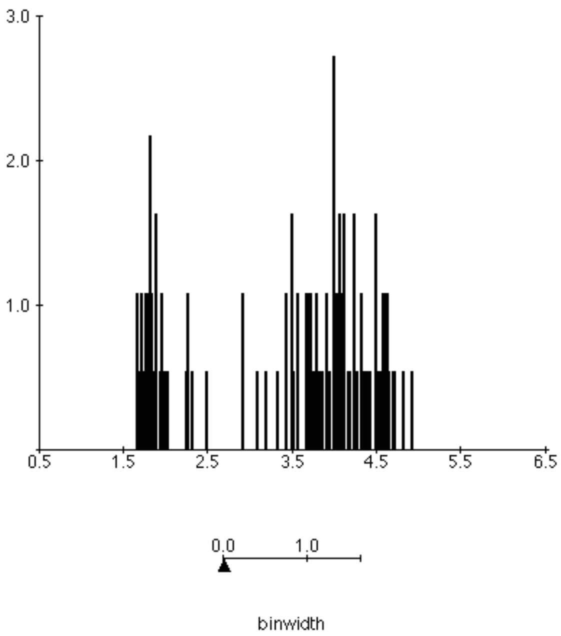
Histograms quickly show distribution clusters

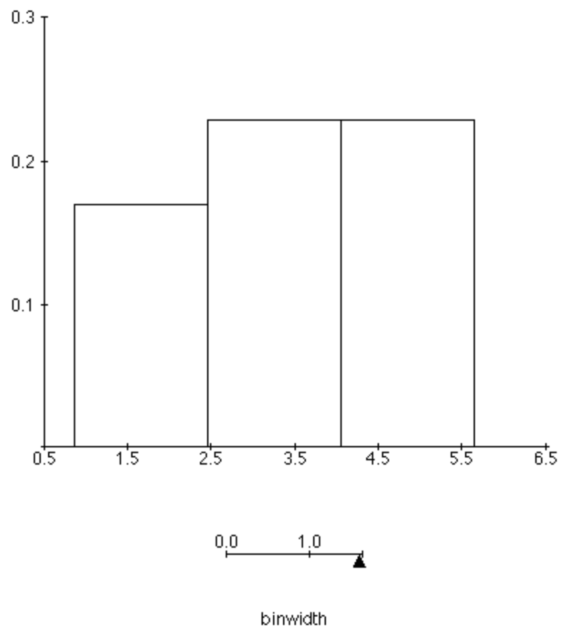
- **Multimodal** data
- No separate central tendency.
- Can't assume centrality tendencies

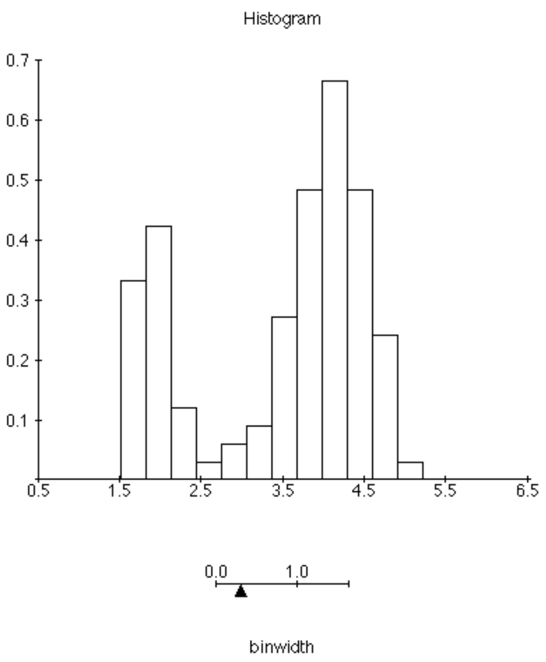


http://www.sciencebuddies.org/science-fair-projects/project_data_analysis_summarizing_data.shtml

- Bin size matters







Issues with Histograms

- For small data sets, histograms can be misleading. Small changes in the data or to the bin boundaries can result in very different histograms.
- Interactive bin-width example (online applet)
  - http://www.stat.sc.edu/~west/javahtml/Histogram.html

- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.

- Histograms effectively only work with 1 variable at a time
  - Difficult to extend to 2 dimensions, not possible for >2
  - So histograms tell us nothing about the relationships among variables

Understanding data distribution

calculate the "average"

Central tendency measures

- mean
- median
- median

calculate the "spread"

Dispersion measures

- range
- variance
- Standard deviation
- quartiles

Measures of dispersion characterise how spread out the distribution is, i.e., how variable the data are

Dispersion

- data set 1: 3, 4, 4, 5, 6, 8
- mean: 5
- data set 2: 1, 2, 4, 5, 7, 11
- Mean: 5

Dispersion: how scattered (far from the mean) the data are

- Proportional to scale of scatter
- Independent of data set size

Measures of variance

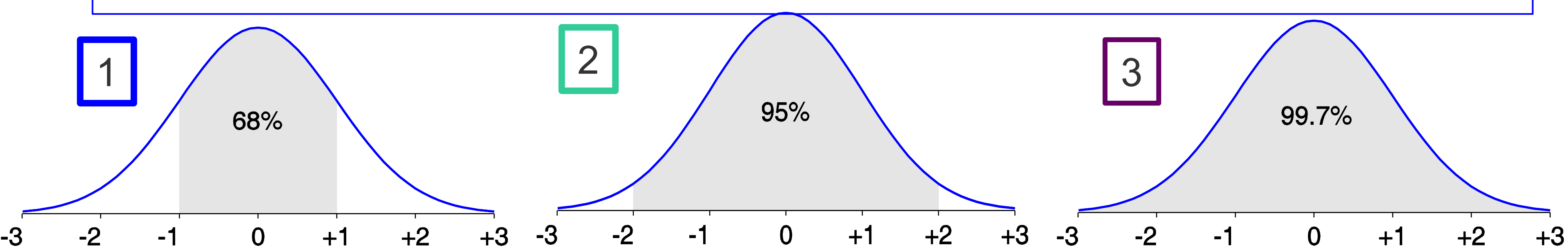
- Variance
  - How far each value in the data set is from the mean
- Standard Deviation
  - Commonest measure
  - Square root of the variance

- Variance and SD are critical in analysing your data distribution and determining how “meaningful” is the chosen average

http://www.sciencebuddies.org/science-fair-projects/project_data_analysis_variance_std_deviation.shtml

Normal Distribution

- The normal (distribution) curve
  1. 1 std dev: contains about 68% of the measurements
  2. 2 std dev: contains about 95% of it
  3. 3 std dev: contains about 99.7% of it



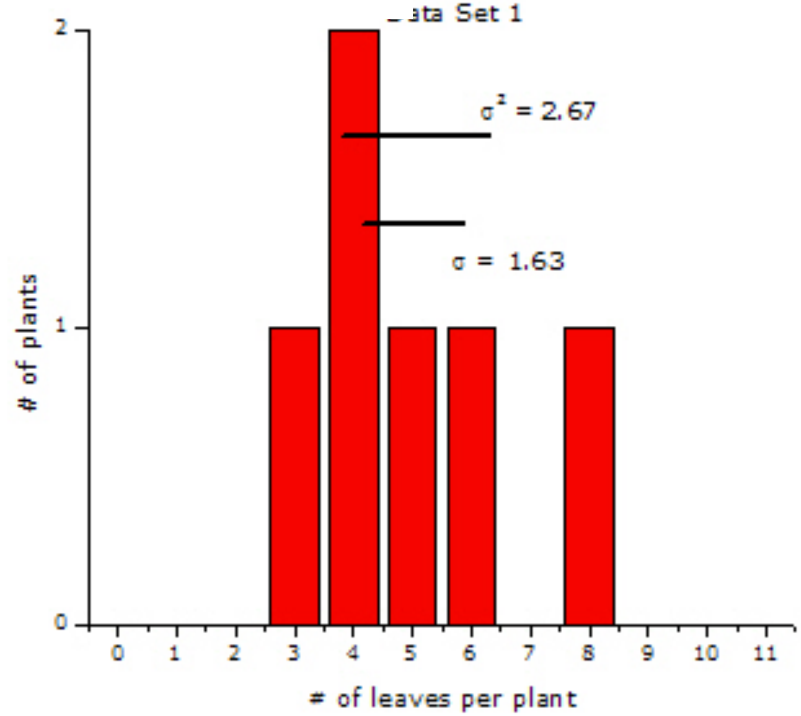
Inter-quartile range

- The Median divides a distribution into two halves.
- The **first** and **third** quartiles (denoted $\mathbf{Q_1}$ and $\mathbf{Q_3}$) are defined as follows:
  - 25% of the data lie below Q_1 (and 75% is above Q_1),
  - 25% of the data lie above Q_3 (and 75% is below Q_3)

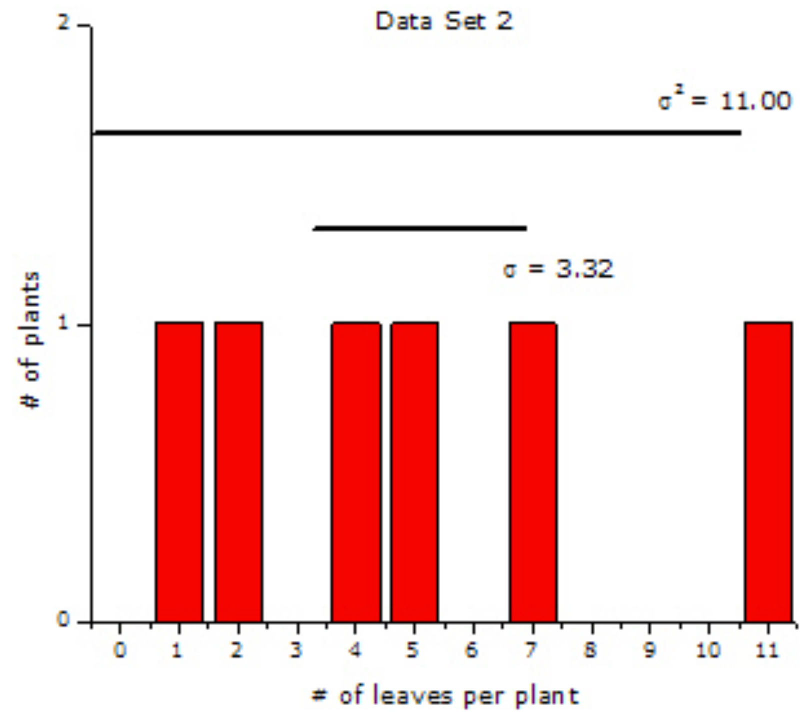
- The **inter-quartile range (IQR)** is the difference between the first and third quartiles, i.e.
  $\mathbf{IQR = Q_3 - Q_1}$

Can we see the magnitude of the spread?

Small scatter
SD = 1.63



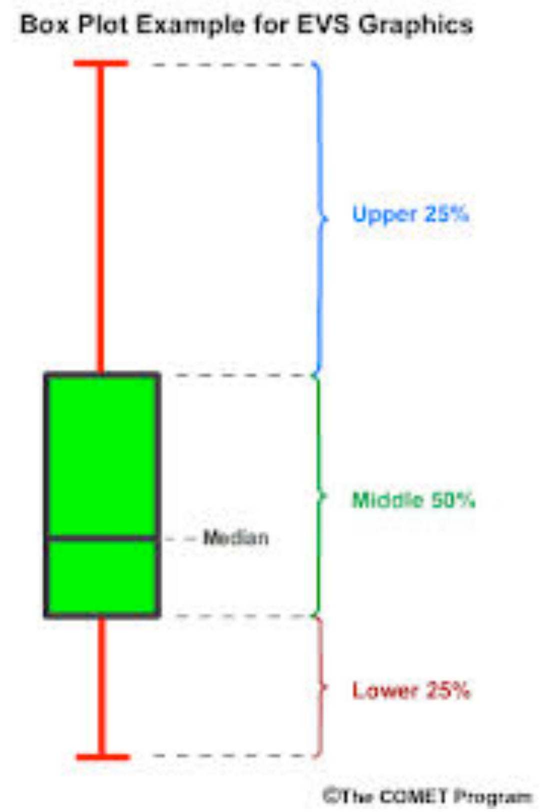
Big scatter
SD = 3.32



Box-plots

- A box-plot is a visual description of the distribution based on
  - Minimum
  - Q1
  - Median
  - Q3
  - Maximum

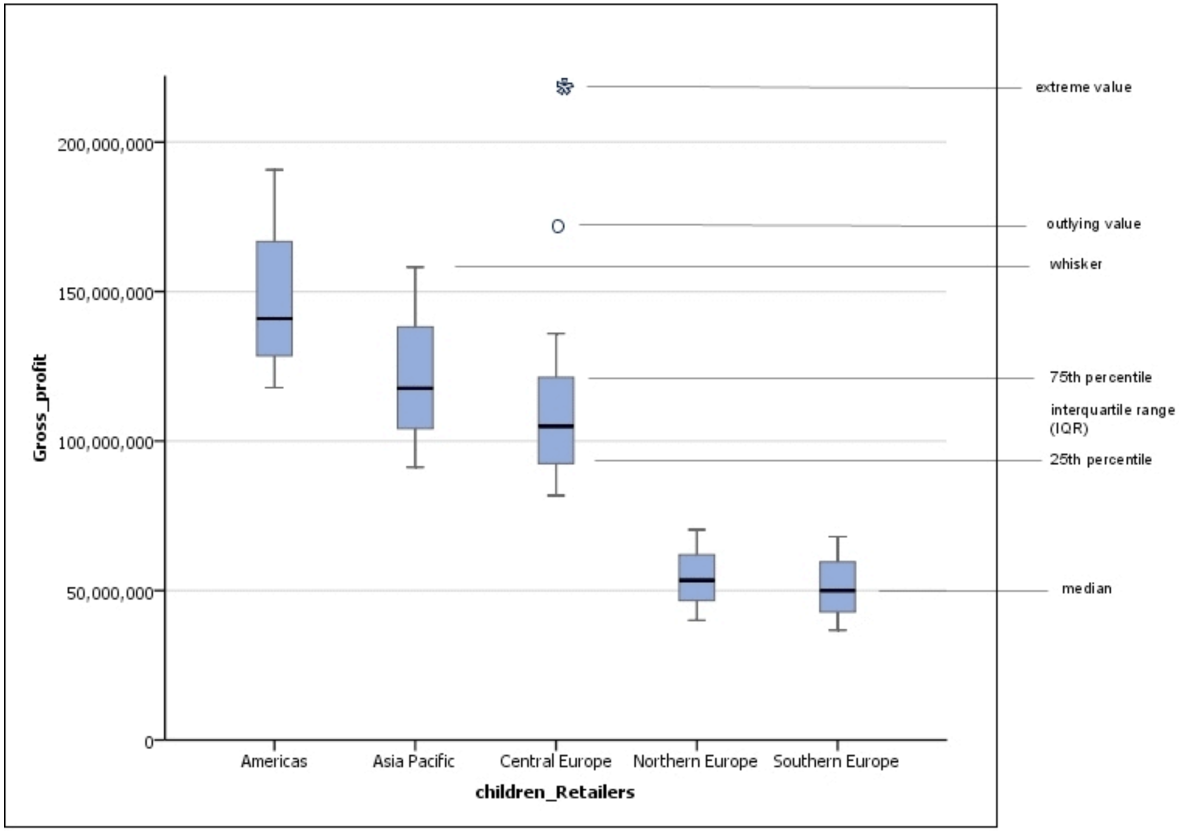
- Useful for comparing large sets of data



Outliers

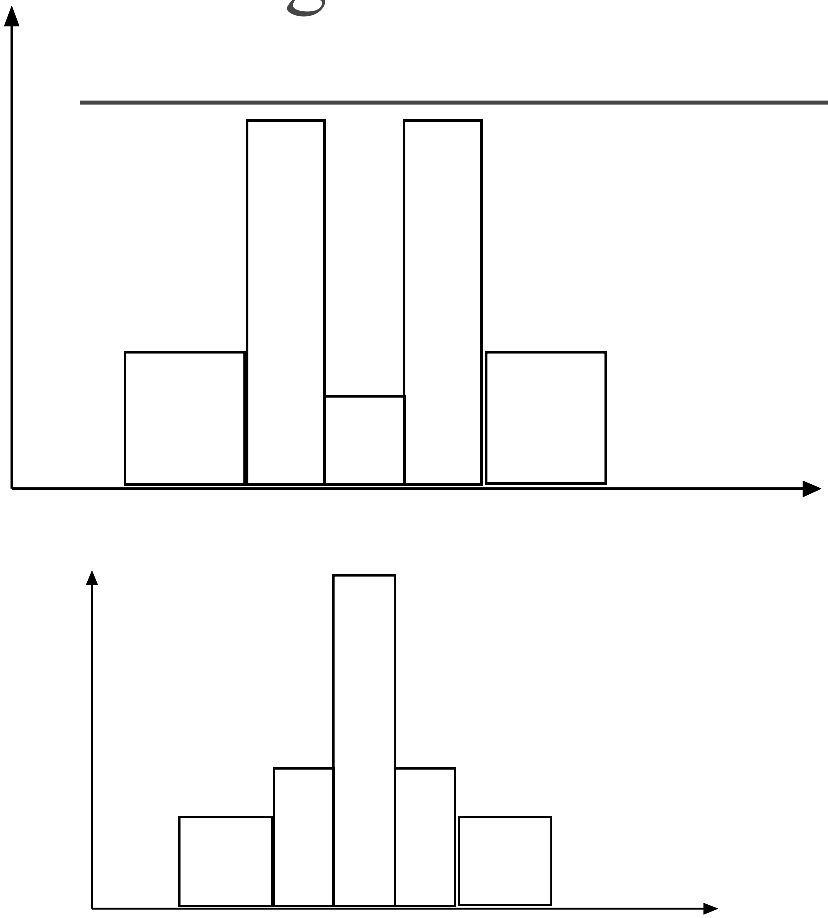
- An **outlier** is data that does not appear to belong with the other data
- Outliers can arise because of a measurement or recording error or because of equipment failure during an experiment, etc.
- An outlier might be indicative of a sub-population, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.

Box plots are useful for comparing dimensions as well as seeing outliers



[IBM Cognos Business Intelligence User Guide]()

Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions
- Use both !

Recap: Distribution is important for understanding aggregate measures

- Visualization helps us see relations – or the trends of them - as visual patterns
- a lot of what we visualize are the descriptive statistics
  - Example: mean income vs median income
  - Need to ensure that the univariate units of visualization are legit

- Rule: check your core units /variables.
- If they are summative/centrality-based, look at the distribution

Exploring data

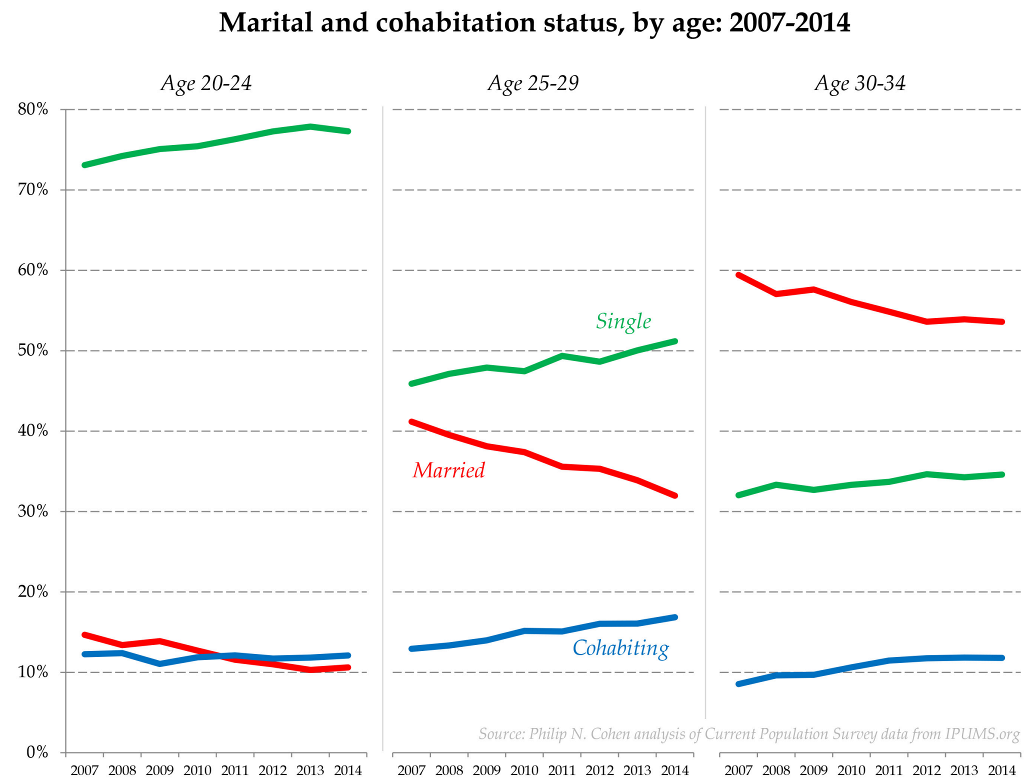
Example: US Census

- People # of people in group
- Year # 1850 – 2000 (every decade)
- Age # 0 – 90+
- Sex (Gender) # Male, female
- Marital status # Single, Married, Divorced, …

Census data: What type (N, O, Q)?

Example: US Census

- People Q- Ratio
- Year Q- interval
- Age Q - ?? O
- Sex (Gender) N
- Marital status N



Census data: what purpose?

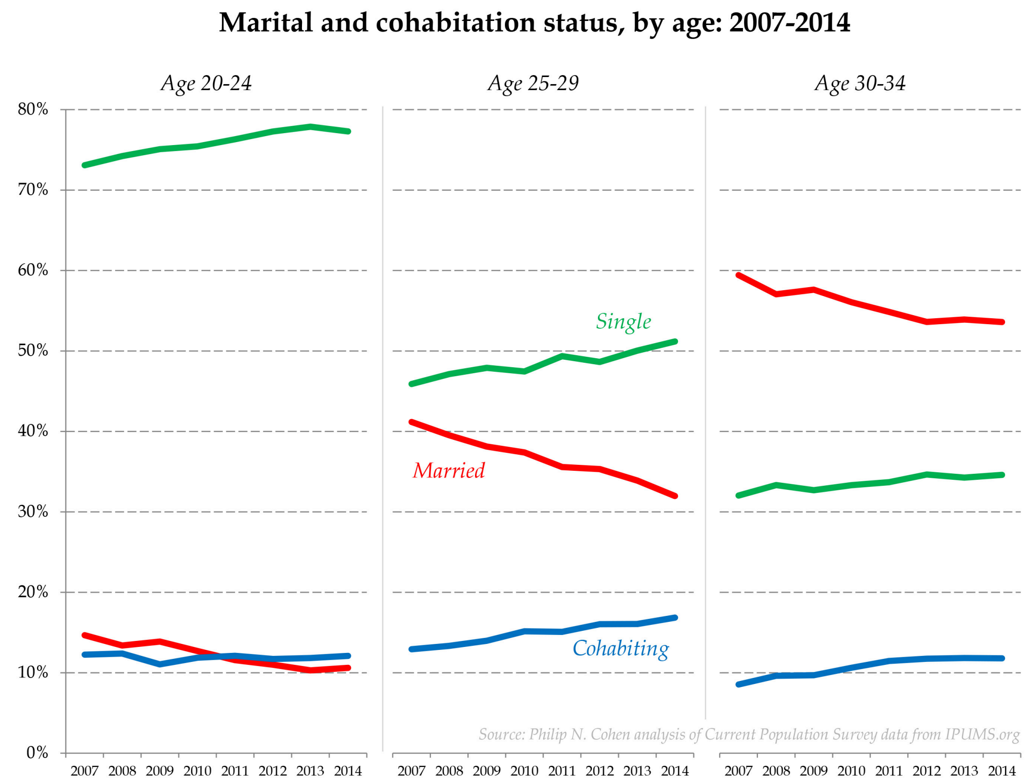
Example: US Census

Measure:

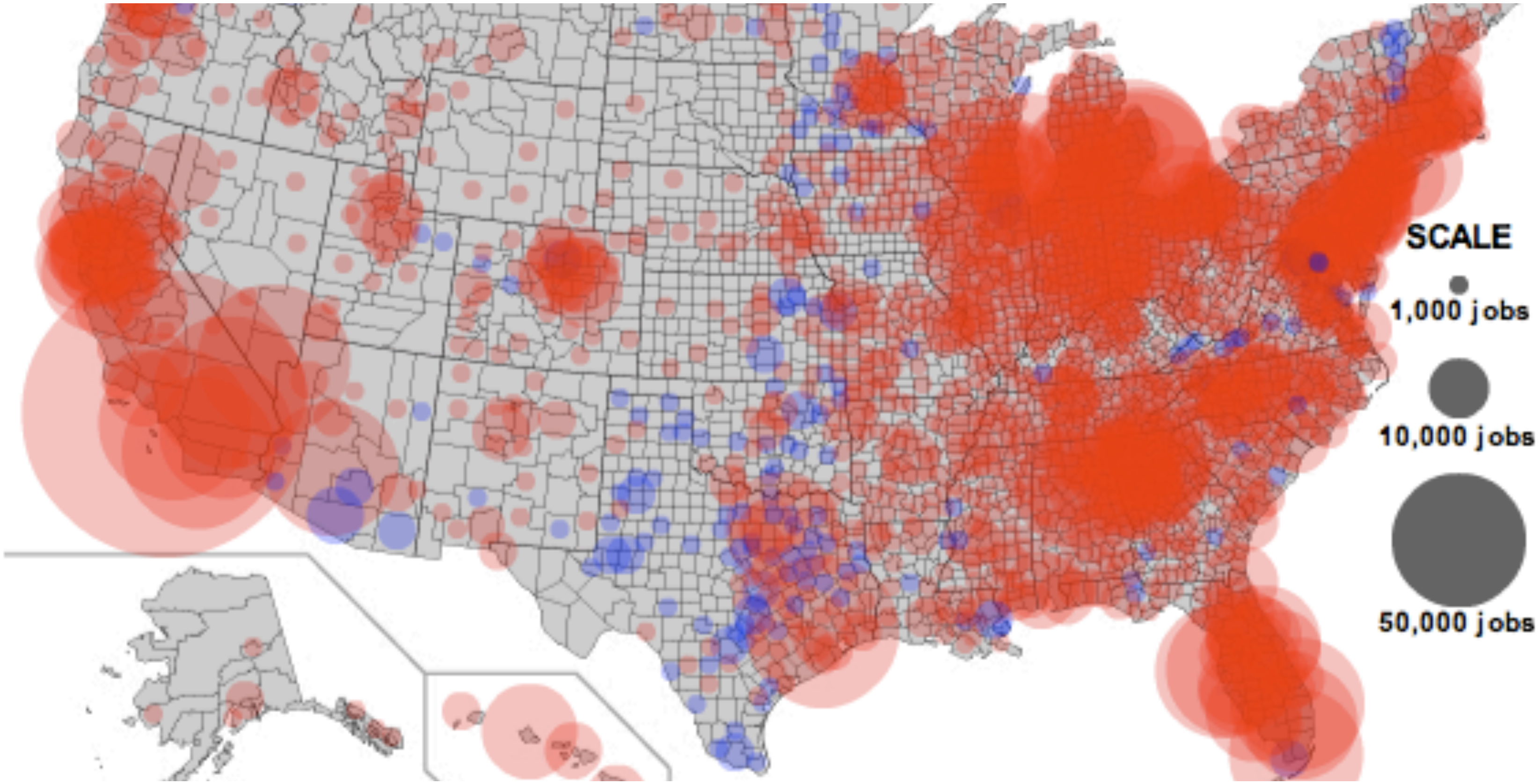
- People Q
- (dependent variable)

Dimensions

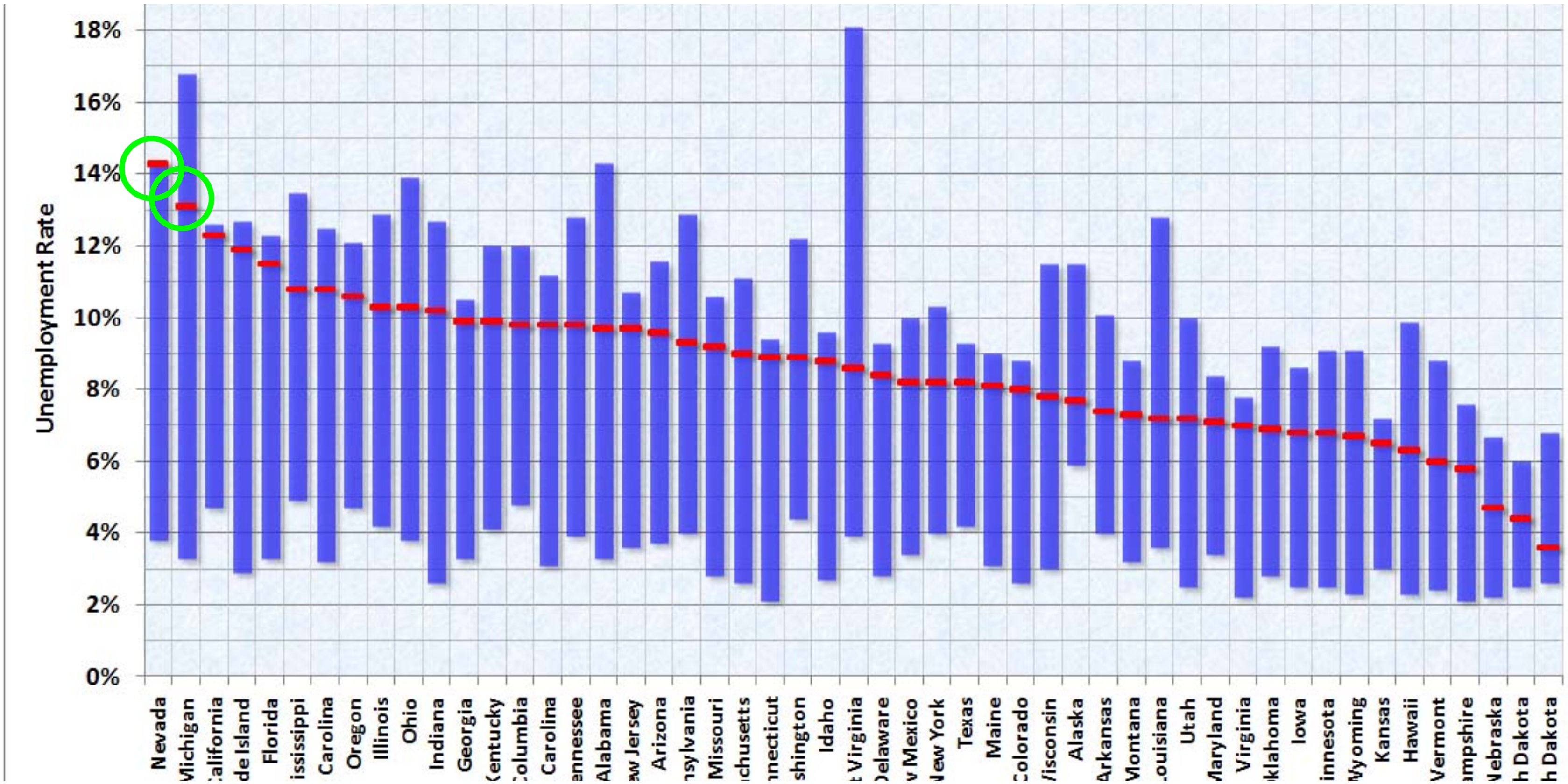
- Year sequence O
- Age sequence+category O
- Sex (Gender) N
- Marital status category O

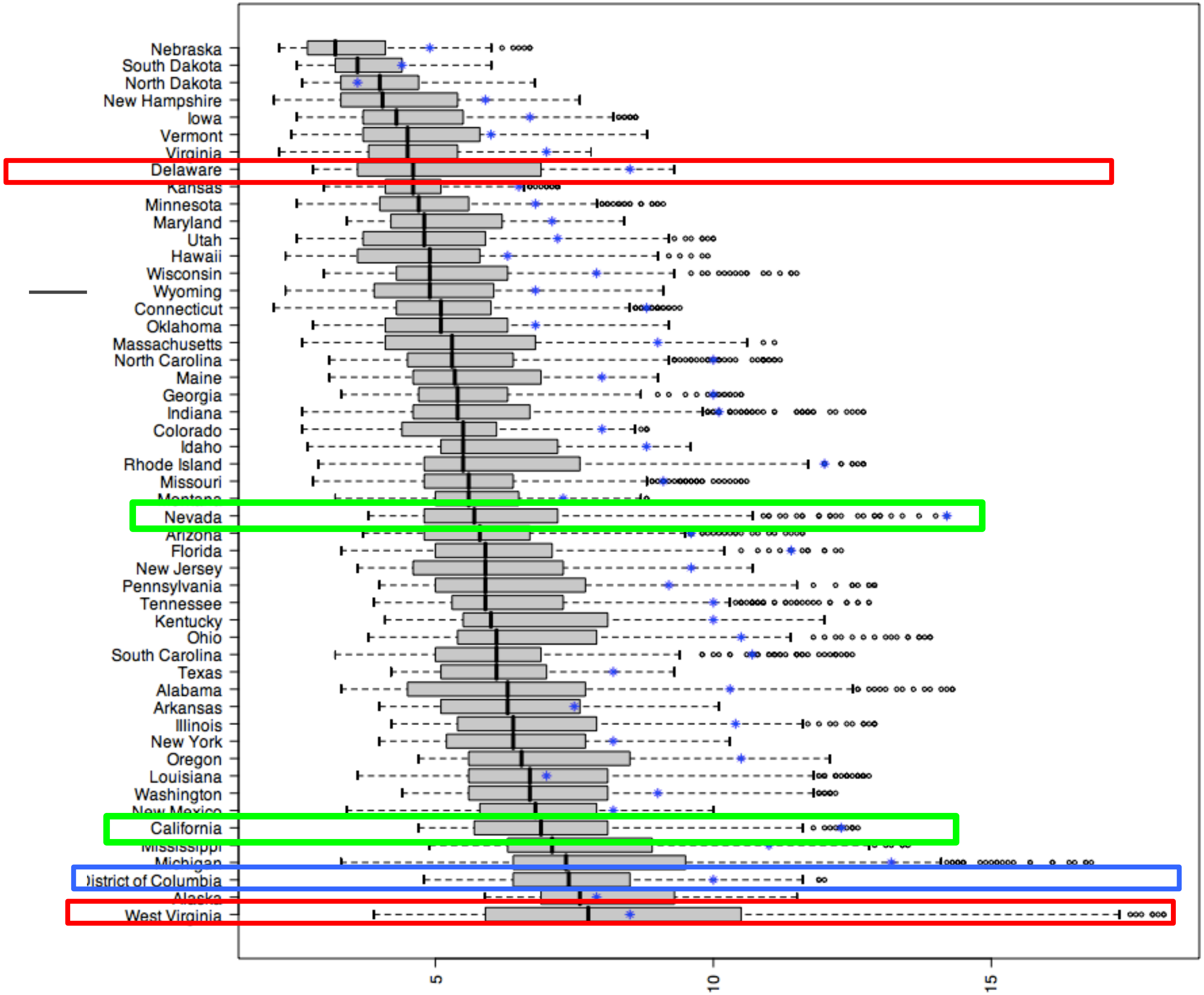


Using the right data



Example: job losses in US over time





Nebraska
South Dakota
North Dakota
New Hampshire
Iowa
Vermont
Virginia
Delaware
Kansas
Minnesota
Maryland
Utah
Hawaii
Wisconsin
Wyoming
Connecticut
Oklahoma
Massachusetts
North Carolina
Maine
Georgia
Indiana
Colorado
Idaho
Rhode Island
Missouri
Montana
Nevada
Arizona
Florida
New Jersey
Pennsylvania
Tennessee
Kentucky
Ohio
South Carolina
Texas
Alabama
Arkansas
Illinois
New York
Oregon
Louisiana
Washington
New Mexico
California
Mississippi
Michigan
District of Columbia
Alaska
West Virginia
5
10
15

Inferential statistics

- Trends, interactons, patterns
- Not a description within dimensions but relations and patterns across them
- Correlation
- Analysis of variance

Correlation

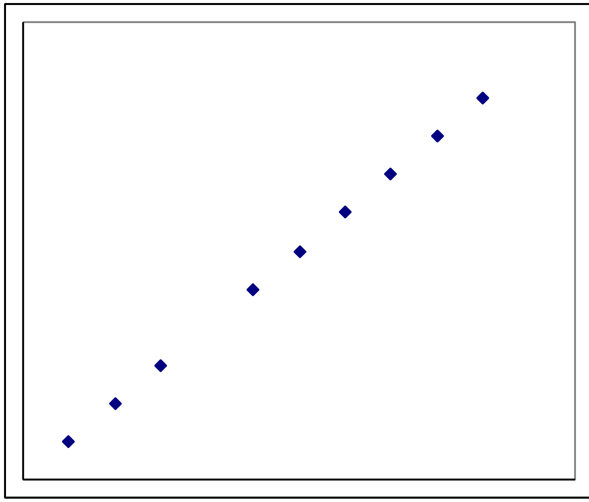
A **correlation** exists between two variables when one of them is related to the other in some way.

A **scatterplot** is a graph in which the paired (*x*,*y*) sample data are plotted on a graph.

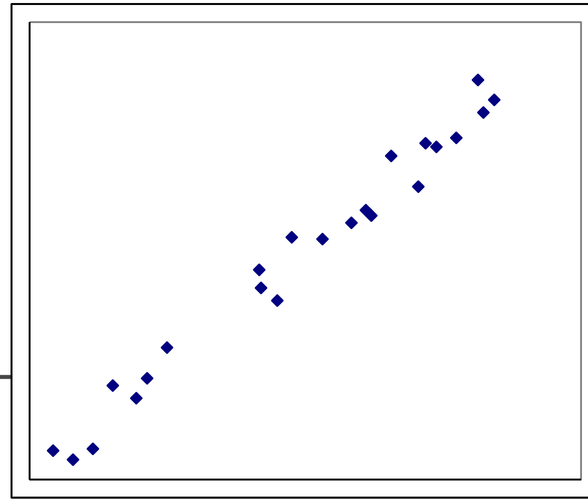
The **linear correlation coefficient *r*** measures the strength of the linear relationship.

- Also called the Pearson correlation coefficient.
- Ranges from -1 to 1.
  - r = 1 represents a perfect positive correlation.
  - r = 0 represents no correlation
  - r = -1 represents a perfect negative correlation

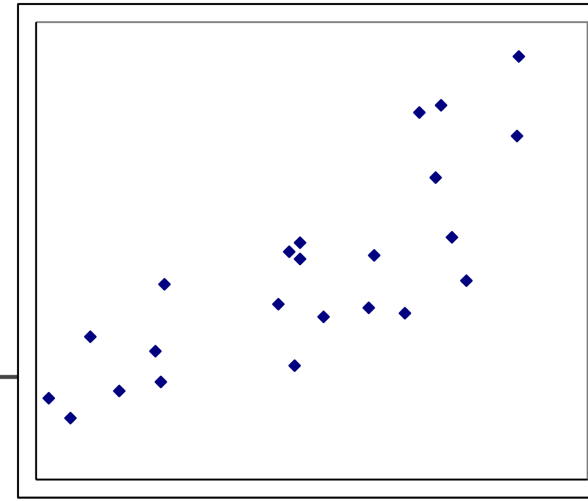
DOES NOT MEAN CAUSATION



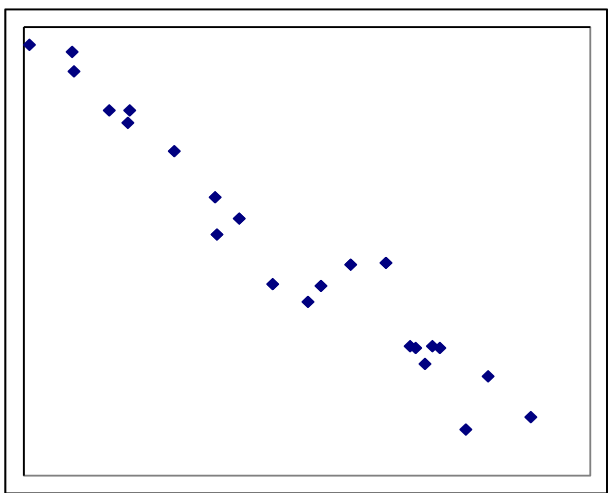
Perfect positive
correlation r = 1



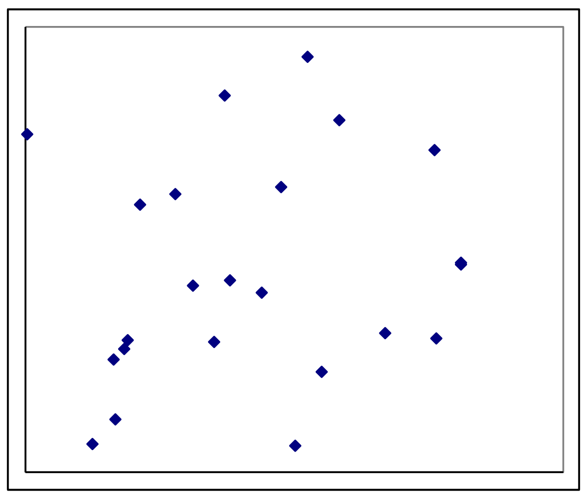
Strong positive
correlation r = 0.99



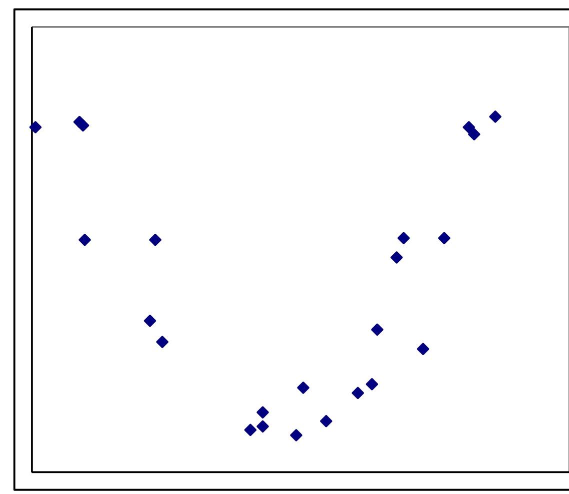
Positive
correlation r = 0.80



Strong negative
correlation r = -0.98



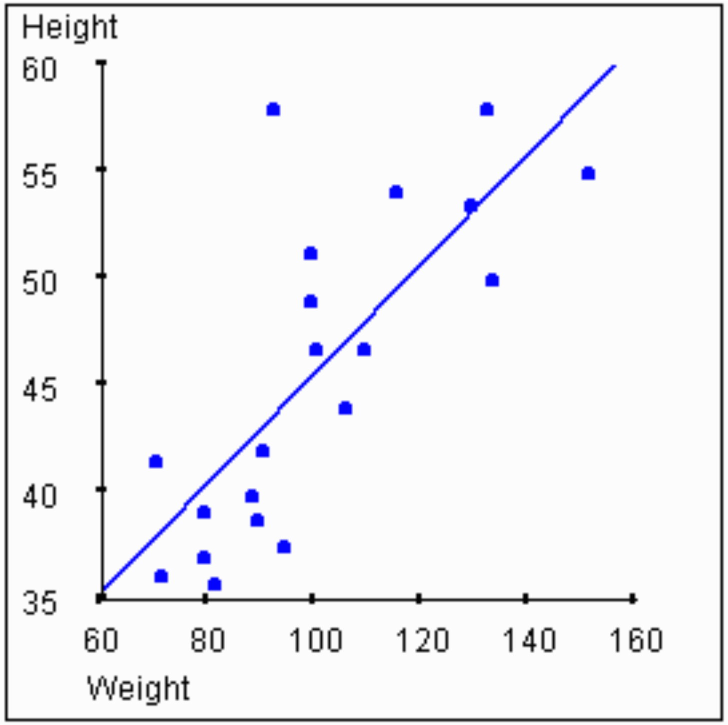
No Correlation
r = 0.16



Non-linear
relationship

Correlation

- about **63.4%** of the peoples' weight can be explained by the relationship between height and weight. This suggests that **36.6%** of the variation in weights cannot be explained by height.
- Outliers can significantly affect trend



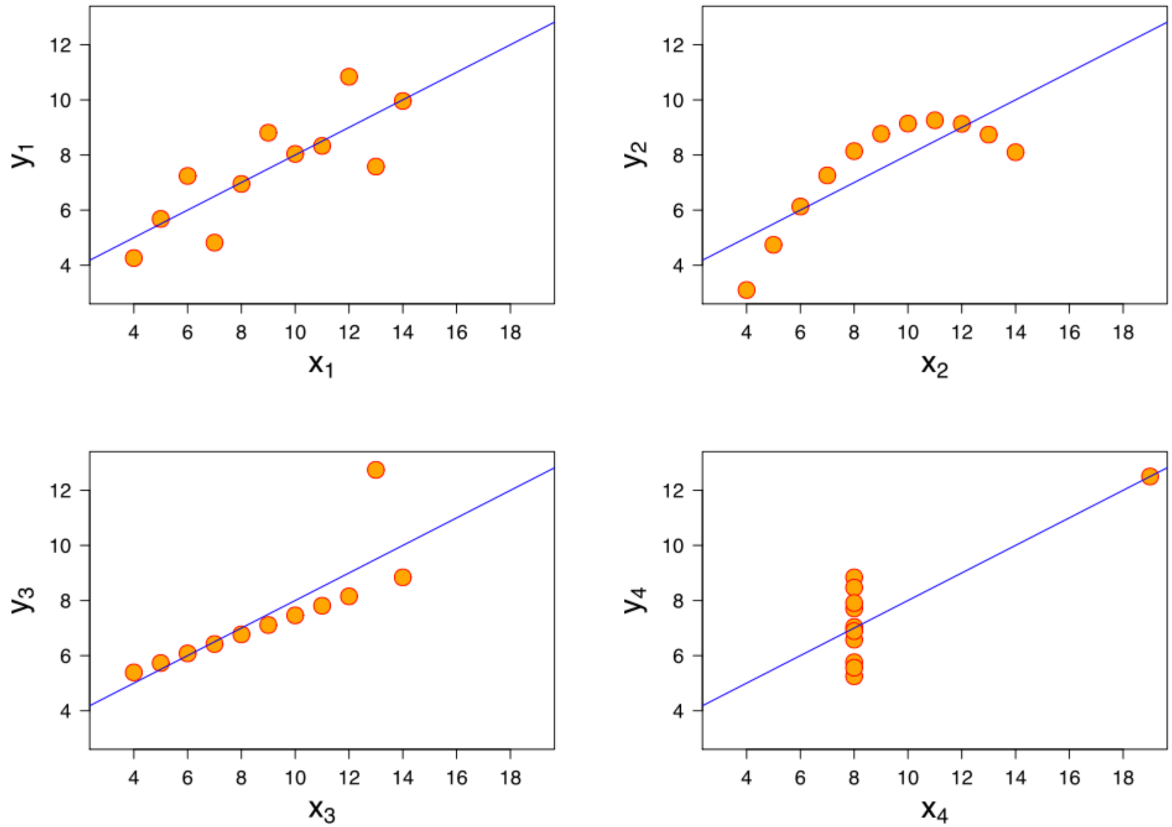
r^2 represents the proportion of the variation in y that is explained by the linear relationship between x and y.
$r^2 = 0.634$

Why we need to look at the data

4 datasets with similar properties (Anscombe's quartet)

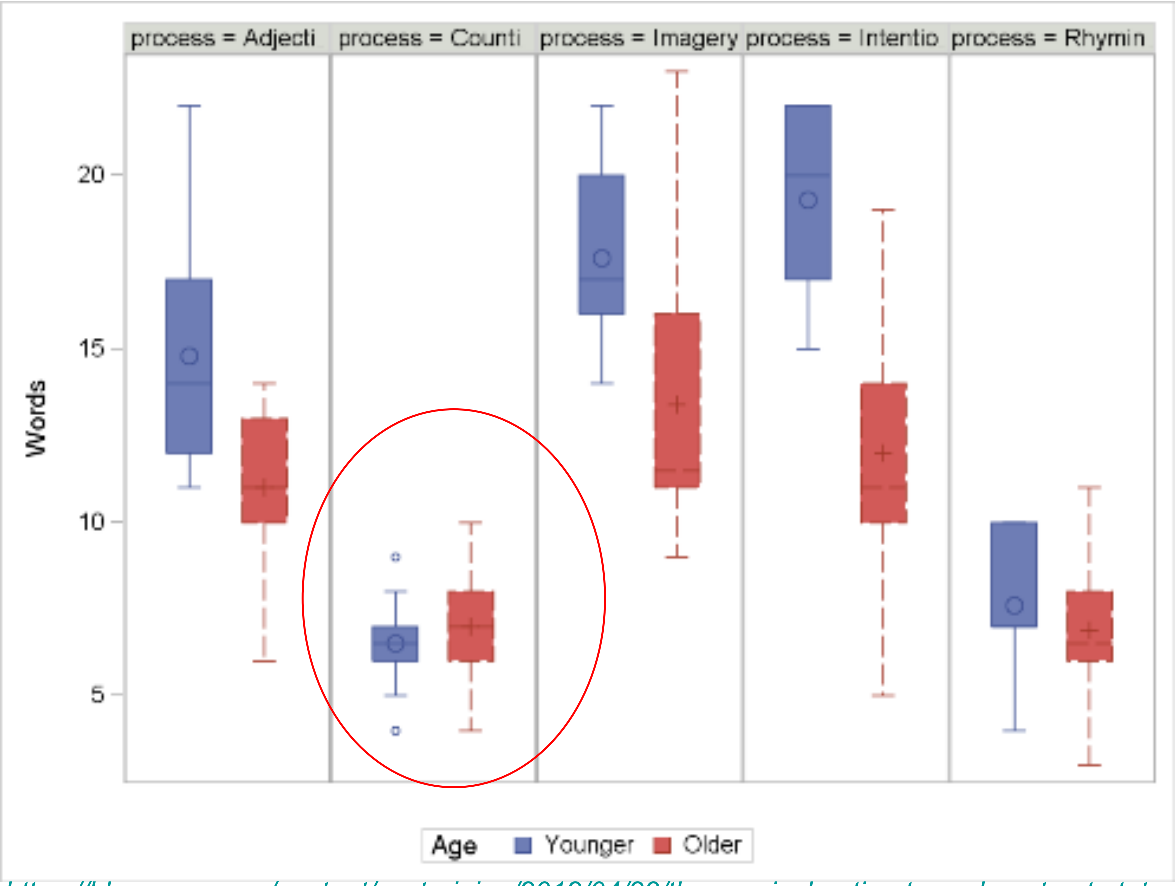
- Mean of the x values = 9.0
- Mean of the y values = 7.5
- Equation of the least-squared regression line: y = 3 + 0.5x
- Sums of squared errors (about the mean) = 110.0
- Regression sums of squared errors (variance accounted for by x) = 27.5
- Residual sums of squared errors (about the regression line) = 13.75
- Correlation coefficient = 0.82
- Coefficient of determination = 0.67

What the data look like …



Visualization for statistical modeling

- Word memory study
- Age (**Y**, **O**)
- process (5)
- Visualization helps determine the statistical questions to ask
- ANOVA



https://blogs.sas.com/content/sastraining/2012/04/23/the-magical-estimate-and-contrast-statements/

Summary

- Statistical models serve to inspect and categorise the nature of trends and relations between data
- Distribution is a critical element in deciding what statistical measures to use
  - lens by which you determine the appropriate metric
- “eyeballing” your distribution is a first step in forming your next queries