Machine Learning & Data Mining

- Exploratory Data Analysis -

Kyung-Ah Sohn

Ajou University

Content

- Data
  - Types of variables
  - Types of data
  - Data quality

- Exploratory data analysis
  - Numerical summary
  - Graphical summary

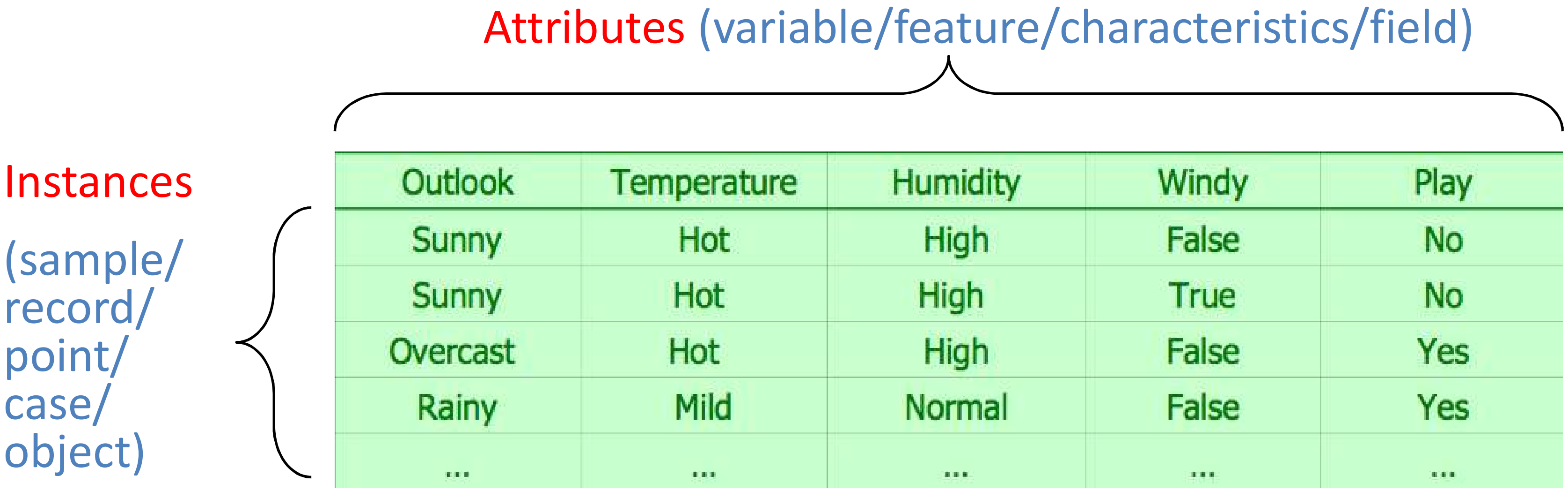
DATA

Terminology

- Components of the input (data)
  - Instances: the individual, independent examples of a concept
    - Sample/object/record/point/case
  - Attributes: measuring aspects of an instance
    - Variable/feature/characteristic/field

Data

- Collection of data instances and their attributes
- An attribute or a feature is a property or characteristic of an instance
- A collection of attributes describe an instance



Attributes (variable/feature/characteristics/field)

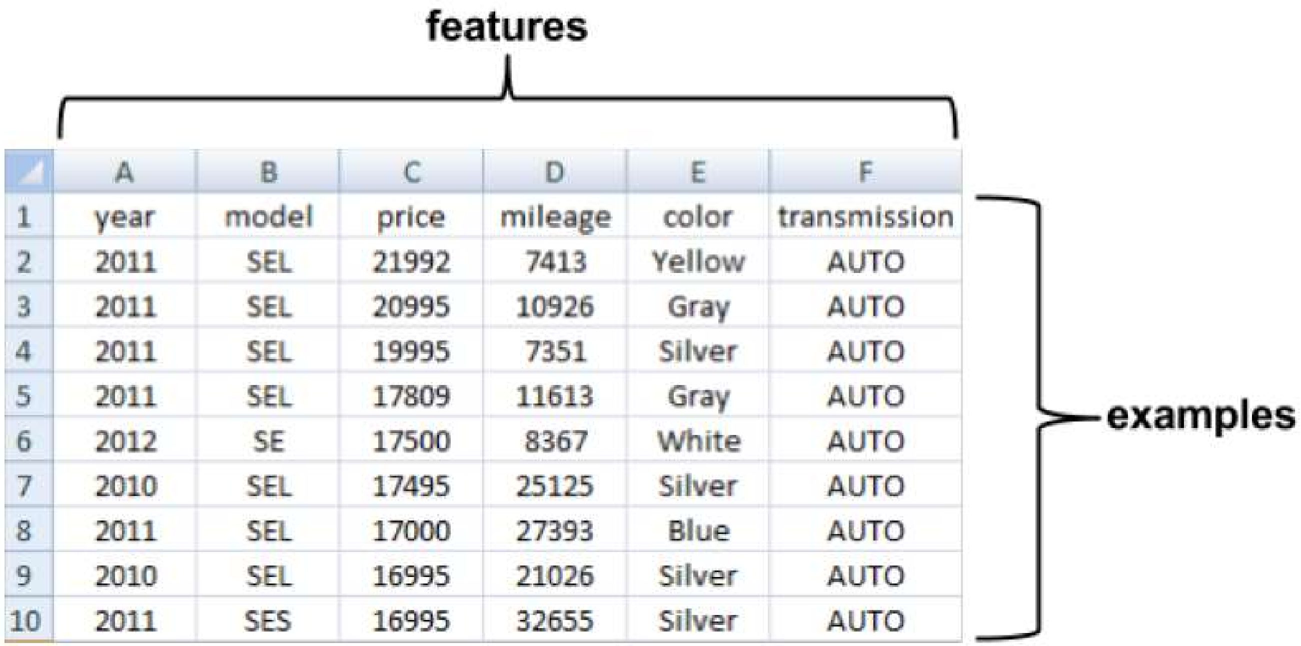
Instances

(sample/ record/ point/ case/ object)

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

Data in matrix format

- The most common form
- Mostly with numeric or categorical features



	A	B	C	D	E	F
1	year	model	price	mileage	color	transmission
2	2011	SEL	21992	7413	Yellow	AUTO
3	2011	SEL	20995	10926	Gray	AUTO
4	2011	SEL	19995	7351	Silver	AUTO
5	2011	SEL	17809	11613	Gray	AUTO
6	2012	SE	17500	8367	White	AUTO
7	2010	SEL	17495	25125	Silver	AUTO
8	2011	SEL	17000	27393	Blue	AUTO
9	2010	SEL	16995	21026	Silver	AUTO
10	2011	SES	16995	32655	Silver	AUTO

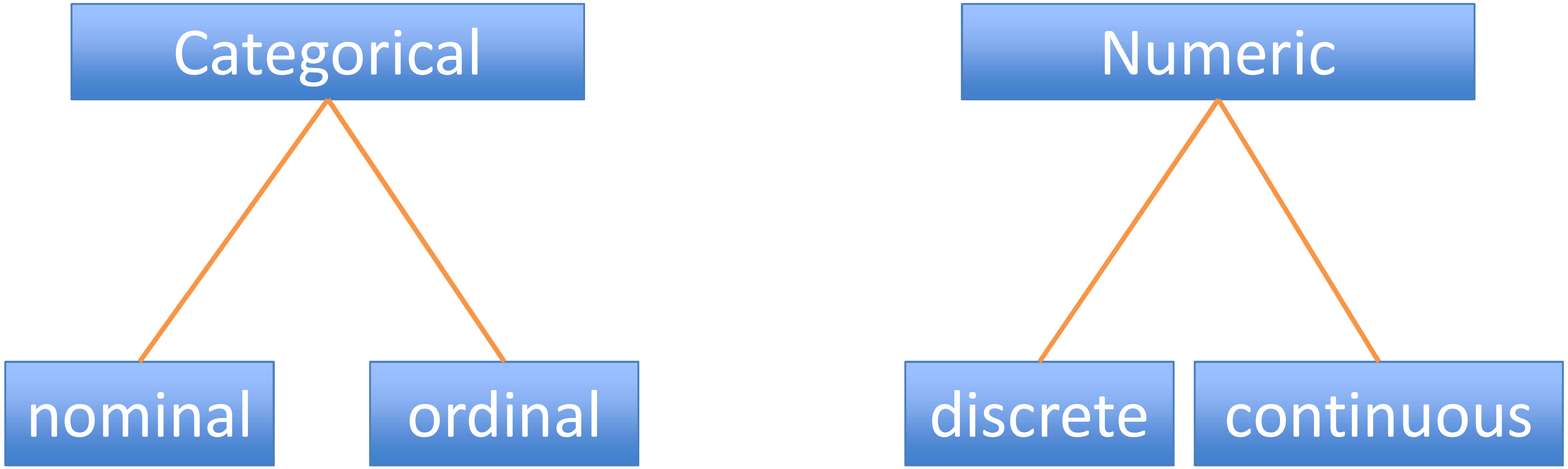
Attribute Values

- Attribute/feature values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be measured in feet or meters

- Possible attribute types ("levels of measurement"):
  - *Nominal, ordinal, interval* and *ratio, …*

Types of Variables



Categorical

Numeric

nominal

ordinal

discrete

continuous

categories

order matters

numerical

uninterrupted

Nominal quantities

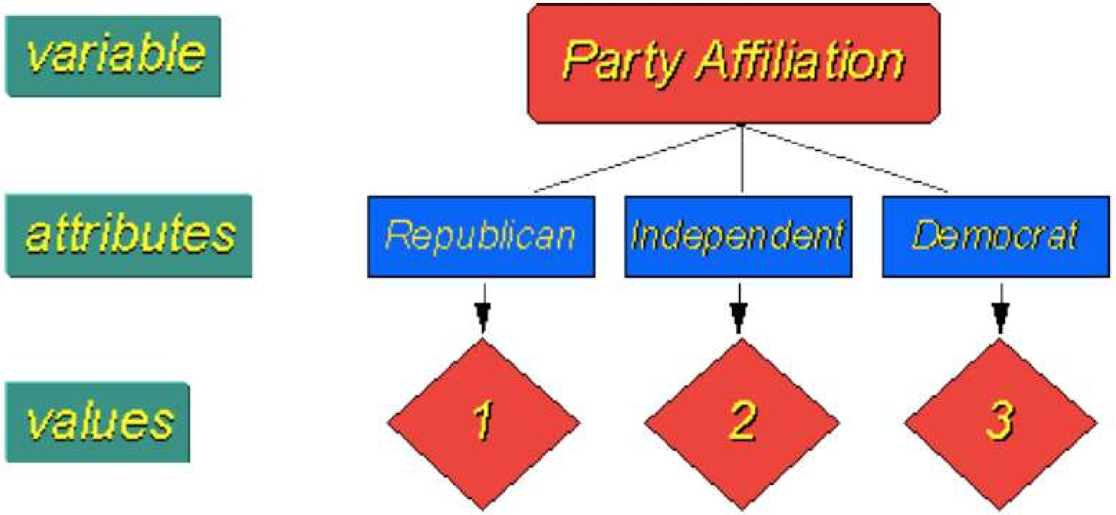
- Values are distinct symbols
  - Values themselves serve only as labels or names
  - *Nominal* comes from the Latin word for name

- Example: attribute “outlook” from weather data
  - Values: “sunny”,”overcast”, and “rainy”

- No relation is implied among nominal values (no ordering or distance measure)

- Only equality tests can be performed

Nominal Variables



- Numerical values have no semantic meaning, just indices
- No ordering implied

- Example
  - ID numbers, eye color, zip codes
  - Jersey numbers in basketball

Ordinal quantities

- Impose order on values
- But: no distance between values defined

- Example: attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”

- Note: addition and subtraction don’t make sense

- Example: Rankings, grades, height in {tall, medium, short}

- Example rule: temperature < hot ⇒ play = yes

- Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

Discrete and Continuous

- Discrete attributes
  - Has only a finite or countably infinite set of values
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

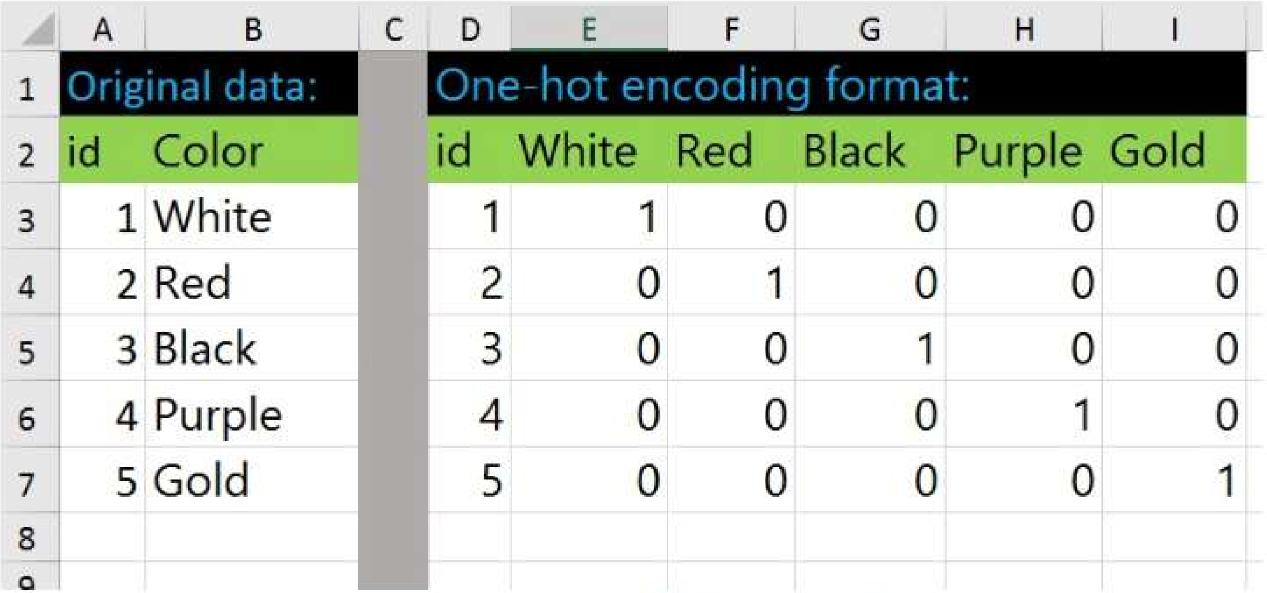
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight

Why is this important?

- Many models require data to be represented in a specific form
- e.g. real-valued vectors

- What do we do with non-real valued inputs?
  - – Nominal with M values
    - Not appropriate to “map” 1 to M (why?)
    - Could use M binary “indicator” variables

One-hot encoding



	A	B	C	D	E	F	G	H	I
1	Original data:			One-hot encoding format:					
2	id	Color		id	White	Red	Black	Purple	Gold
3	1	White		1	1	0	0	0	0
4	2	Red		2	0	1	0	0	0
5	3	Black		3	0	0	1	0	0
6	4	Purple		4	0	0	0	1	0
7	5	Gold		5	0	0	0	0	1
8									
9									

One-hot 인코딩 예시, 출처: stackoverflow

Mixed data

- Many real-world data sets have multiple types of variables
- Unfortunately, many data analysis algorithms are suited to only one type of data…

Types of Data sets

- Record
- Transaction data
- Graph
- Ordered
- …

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

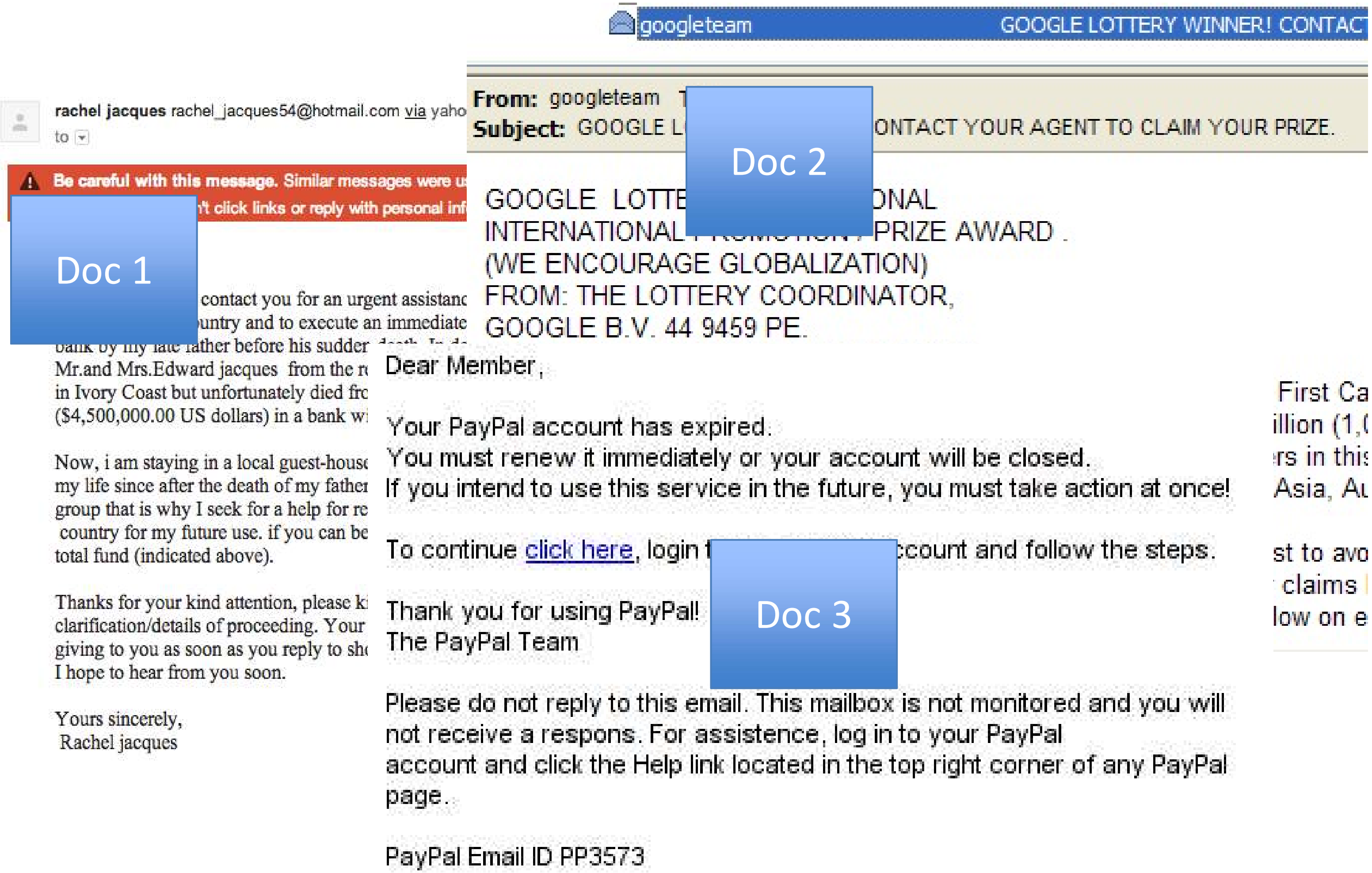
Record

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

- Tables

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Document data



Doc 1

rachel jacques rachel_jacques54@hotmail.com via yaho
to

Be careful with this message. Similar messages were u
't click links or reply with personal inf

contact you for an urgent assistanc
untry and to execute an immediate
bank by my late father before his sudder
Mr.and Mrs.Edward jacques from the r
in Ivory Coast but unfortunately died frc
($4,500,000.00 US dollars) in a bank wi

Now, i am staying in a local guest-house
my life since after the death of my father
group that is why I seek for a help for re
country for my future use. if you can be
total fund (indicated above).

Thanks for your kind attention, please ki
clarification/details of proceeding. Your
giving to you as soon as you reply to sho
I hope to hear from you soon.

Yours sincerely,
Rachel jacques

Doc 2

googleteam GOOGLE LOTTERY WINNER! CONTACT

From: googleteam T
Subject: GOOGLE L ONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTE ONAL
INTERNATIONAL PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.

Doc 3

Dear Member,

Your PayPal account has expired.
You must renew it immediately or your account will be closed.
If you intend to use this service in the future, you must take action at once!

To continue click here, login t ccount and follow the steps.

Thank you for using PayPal!
The PayPal Team

Please do not reply to this email. This mailbox is not monitored and you will not receive a respons. For assistence, log in to your PayPal account and click the Help link located in the top right corner of any PayPal page.

PayPal Email ID PP3573

First Ca
illion (1,(
rs in this
Asia, Au

st to avo
claims
low on e

Document Data

- Each document becomes a `term' vector

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

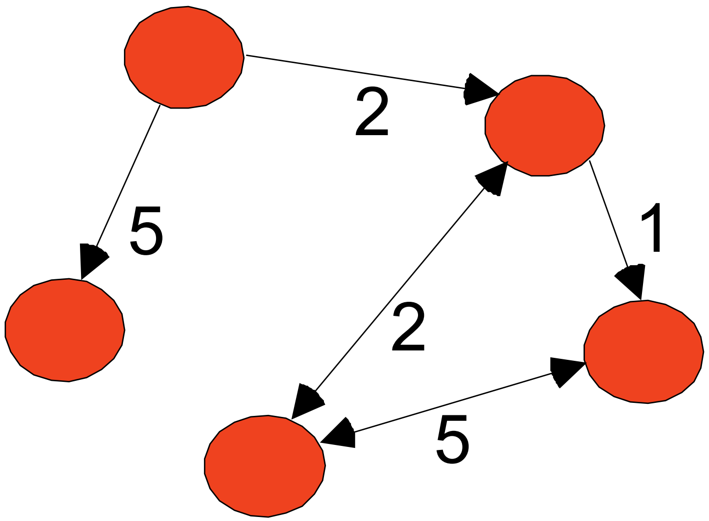
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data

- Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Either as a matrix or a list

Ordered Data

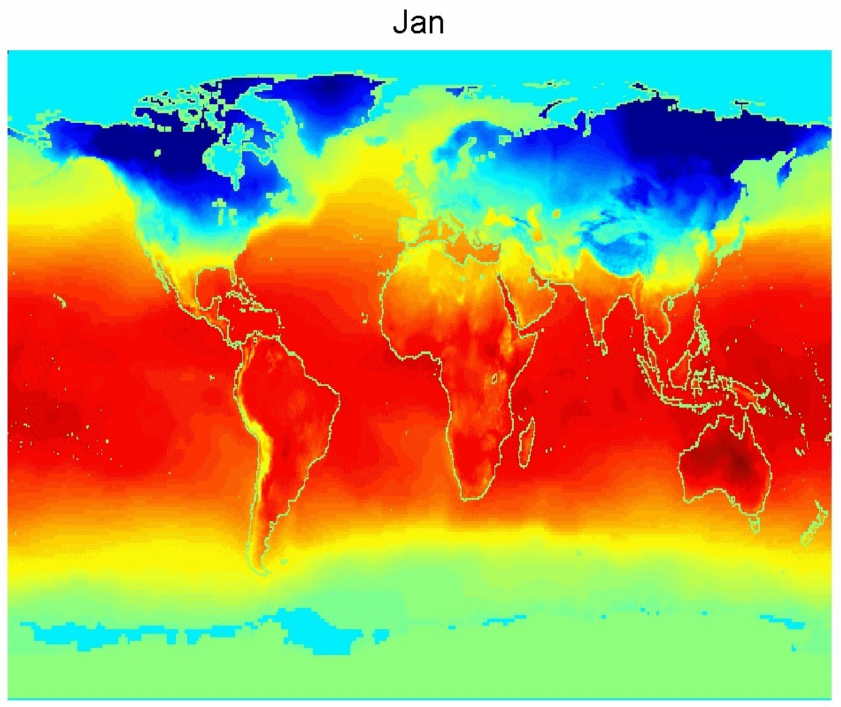
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

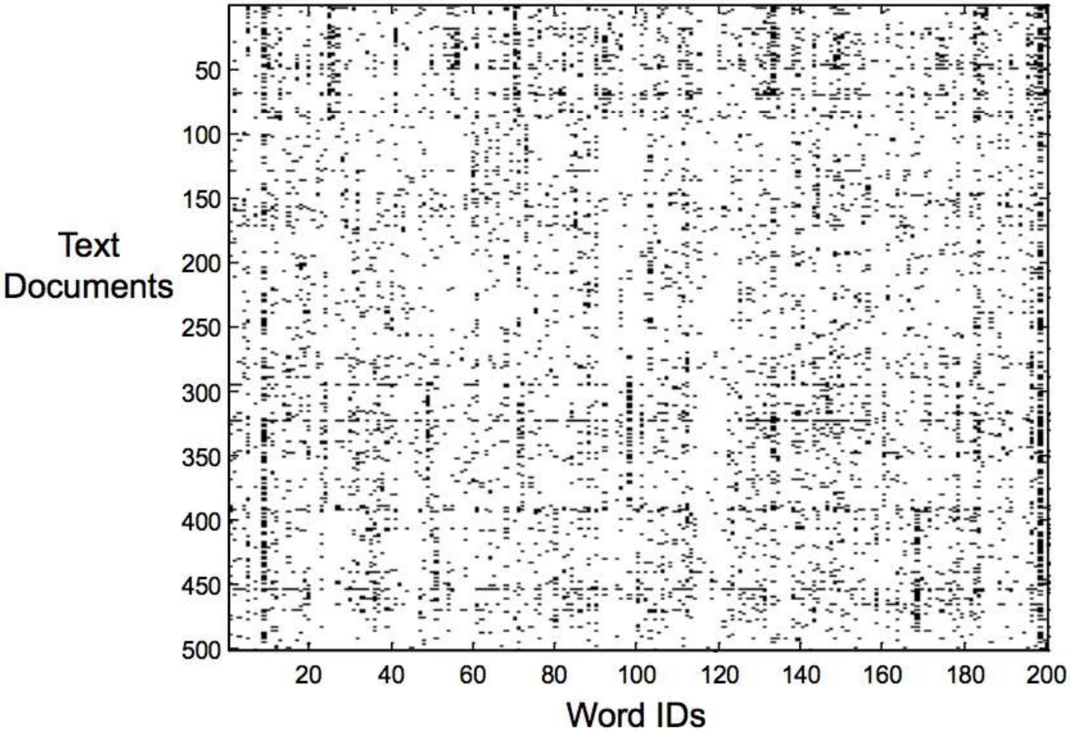
- Spatio-Temporal data

Average Monthly Temperature of land and ocean



Sparse data

- In some applications most attribute values in a dataset are zero
  - E.g. word counts in a text categorization problem



Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - missing values
  - Noise and outliers
  - duplicate data

Missing Values

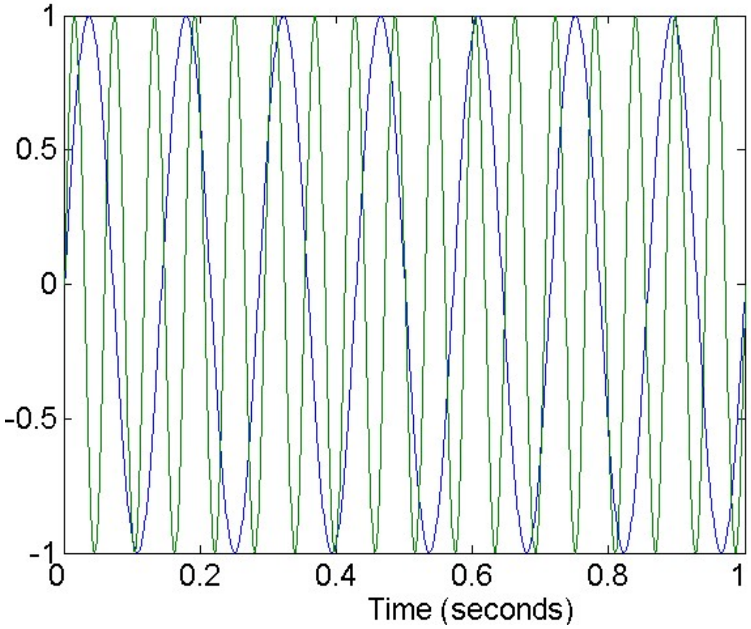
- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Missing value may have significance in itself (e.g. missing test in a medical examination)
  - Most schemes assume that is not the case: “missing” may need to be coded as additional value

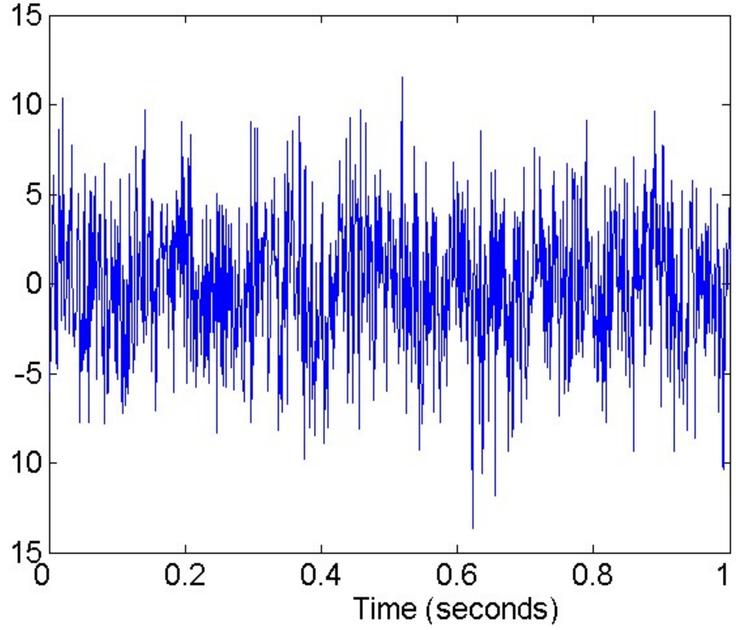
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Code the missing values with additional value
  - Ignore the Missing Value During Analysis

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone



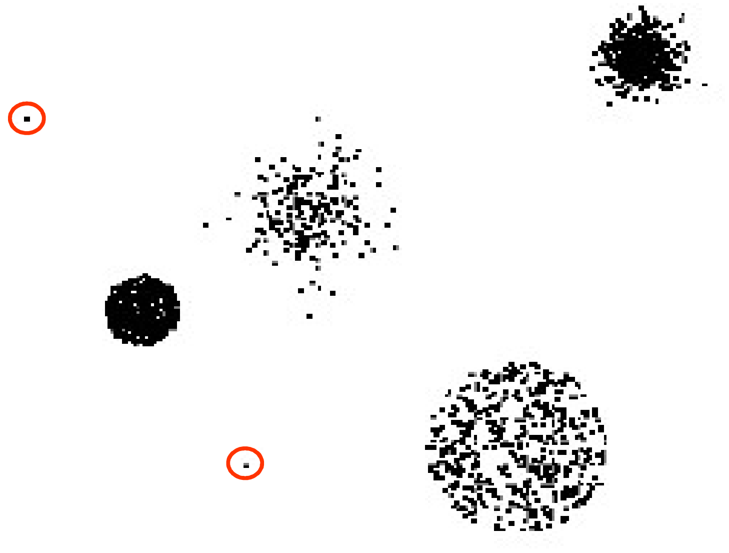
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Duplicate data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

EXPLORATORY DATA ANALYSIS

EDA

- Graphical summaries of data
  - Visualization

- Numerical summaries of data
  - Descriptive statistics

Getting to know the data

- Simple visualization tools are very useful
  - Nominal attributes: histograms (Distribution consistent with background knowledge?)
  - Numeric attributes: graphs

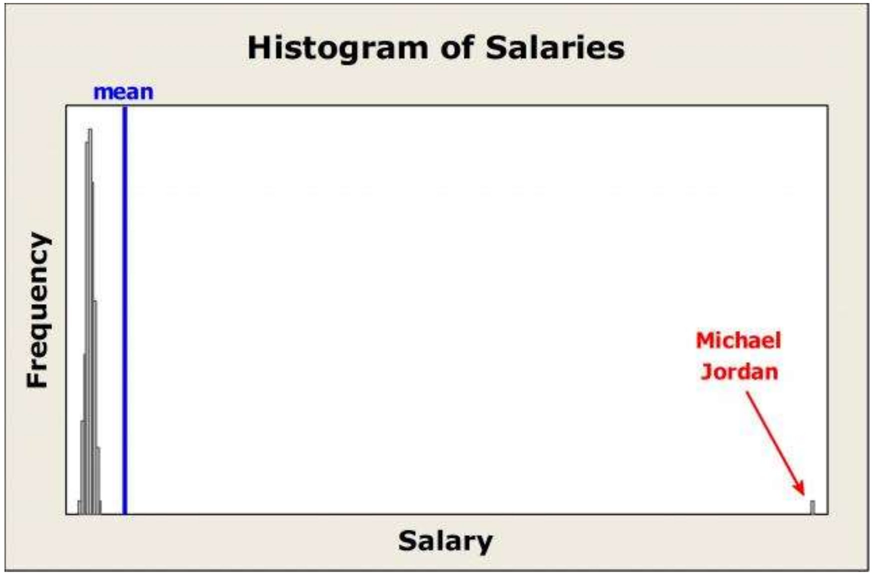
    (Any obvious outliers?)

- 2-D and 3-D plots show dependencies
- Need to consult domain experts
- Too much data to inspect? Take a sample!

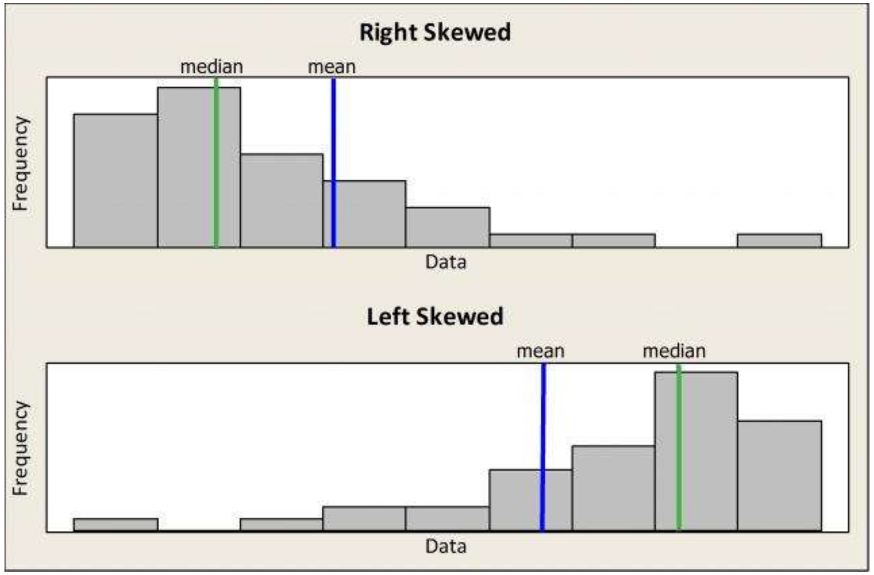
Exploratory Data Analysis (EDA)

- To get a general sense of the data
- You should always look at every variable - you will learn something!
- Data-driven (model-free)
- Think interactive and visual
  - You can use more than 2 dimensions (space, color, time, …)
- Especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))
- Bottom line: it is always well worth looking at your data!

- Always graph your data



- Compute some basic statistics, including both the mean and median



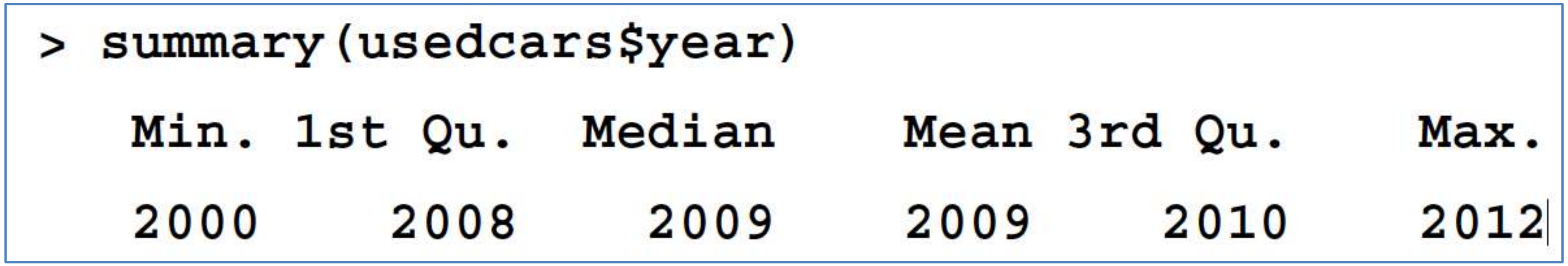
Numerical Summaries of Data

- Not visual
- Summary statistics
  - mean, median
  - mode: the most common value
  - variance, standard deviation
  - quartiles
  - Number of distinct values for a categorical variable

- Don’t need to report all of theses: Bottom line…do these numbers make sense???

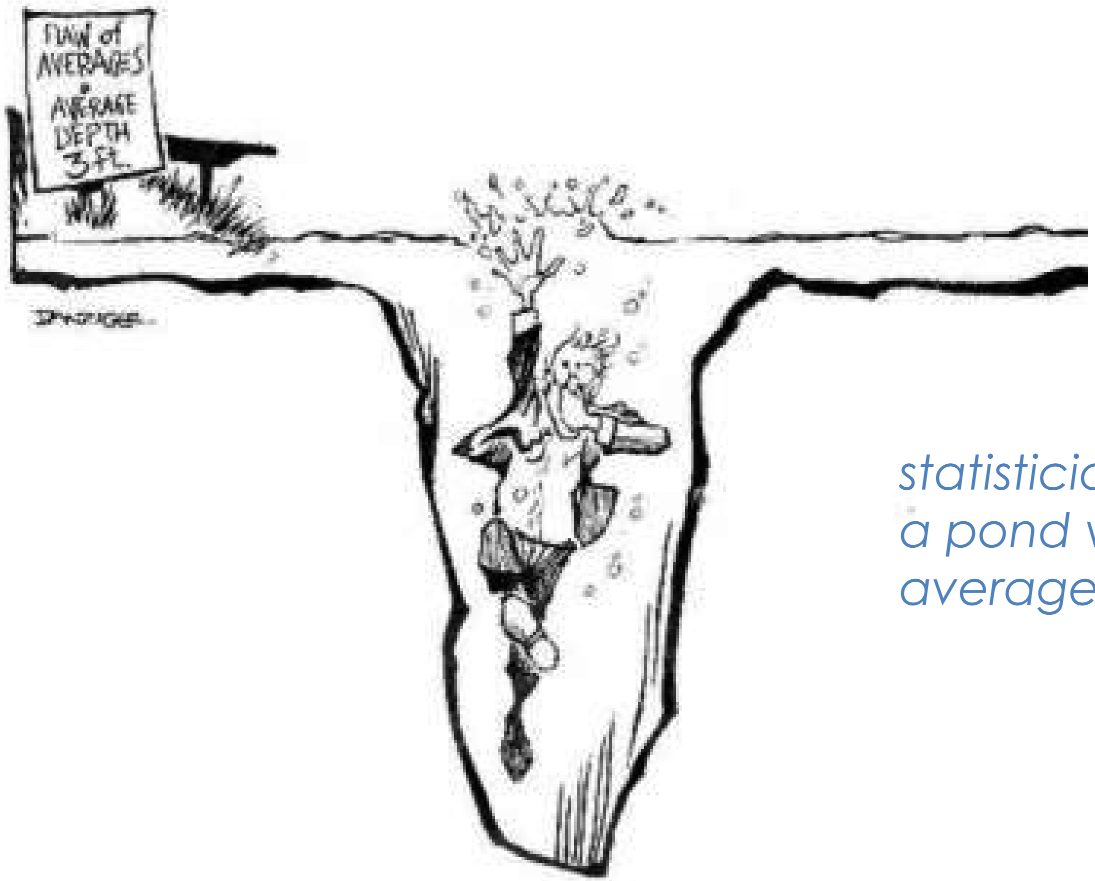
Exploring numeric variables

- Measuring the central tendency
- Measuring spread – quartiles and the five-number summary



```
> summary(usedcars$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2000    2008    2009    2009    2010    2012
```

Averages can be misleading!



statistician drowning in a pond with an average depth of 3ft

http://filipspagnoli.wordpress.com/2009/11/13/lies-damned-lies-and-statistics-21-misleading-averages/

Using the mean in data analysis

- Back in the mid-1980’s at the University of North Carolina, the average starting salary of geography students was over $100,000. Knowing that, would you have considered making a career change?

- What if I told you that basketball great Michael Jordan – formerly the world’s highest paid athlete – graduated from UNC with a degree in geography?

http://blog.minitab.com/blog/michelle-paret/using-the-mean-its-not-always-a-slam-dunk

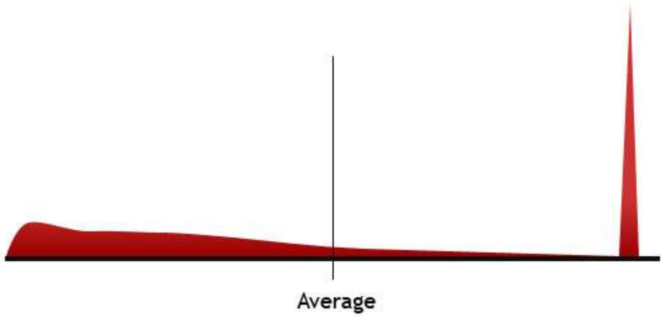
The Mean can Mislead

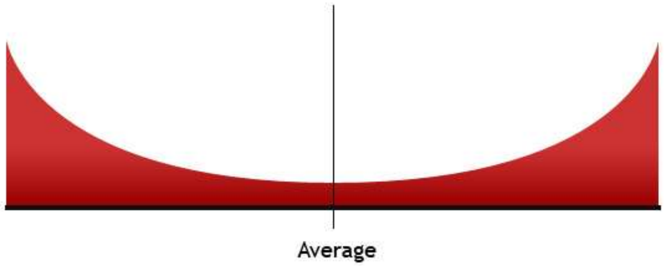
- Jordan’s earnings from his athletic career raises the “average” salary for geography graduates in a way that doesn’t accurately convey what graduates are likely to earn

- By almost any measure, Jordan’s earnings would be an outlier

- How to identify this anomaly?

The average is not a good representation of the true center of the data





http://www.baekdal.com/notes/twitter-average

Median

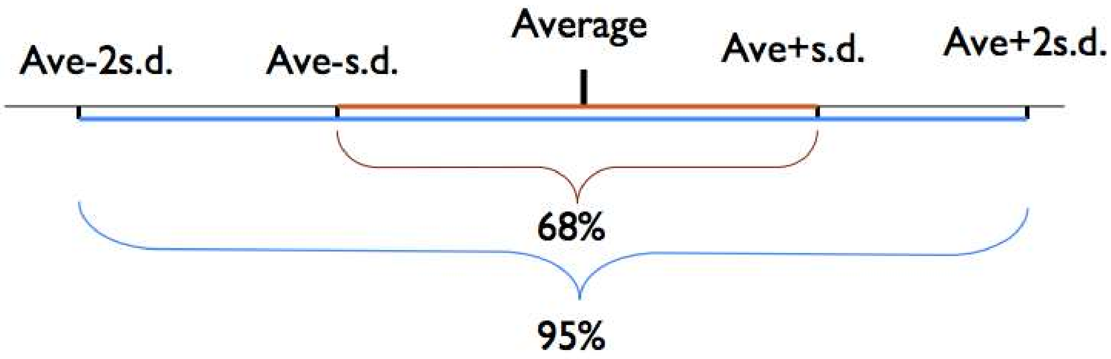
- Median: the exact middle value
  - Useful for skewed distributions or data with outliers
  - More robust than mean
  - Difficult to handle theoretically (no easy mathematical formula)
- Example

Data	1 2 3 4 5	1 2 3 4 100
Mean		
Median		

Standard deviation

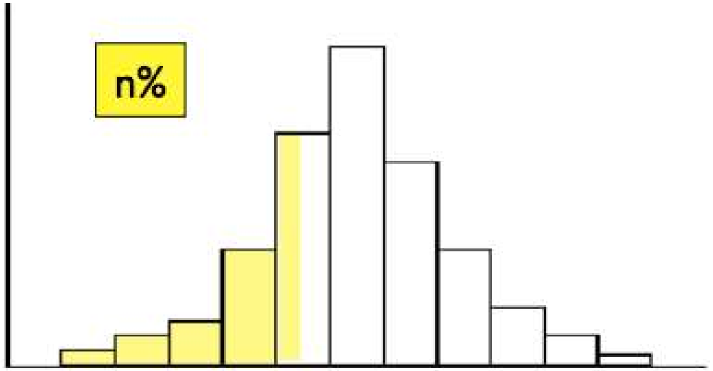
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Interesting theoretical results
  - For many lists of observations, especially if their histogram is bell-shaped



Percentiles (aka Quantiles)

- The **nth percentile** is a value such that n% of the observations fall at or below of it



- Q1 : 25th percentile
- Median: 50th percentile
- Q3 : 75th percentile
- IQR: Interquartile range (25 to 75%: Q3-Q1)

Visualizing numeric variables

- Boxplot: a common visualization of the five-number summary
- Histogram: another way to graphically depict the spread of a numeric variable

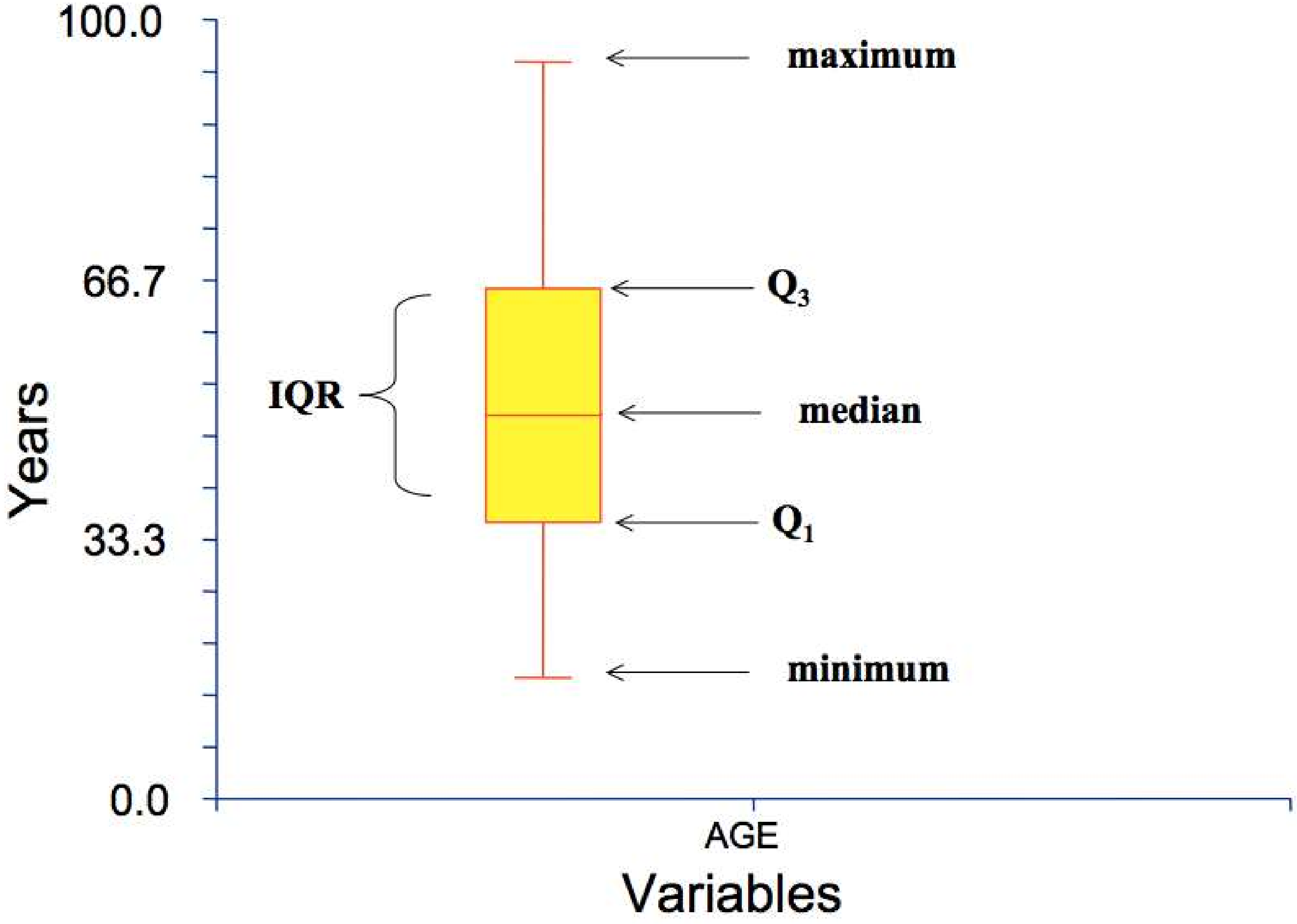
Boxplot

- x-axis: categorical variable
- y-axis: real-valued or integer variable

- For each group, the boxplot shows
  - Median
  - Interquartile range (25 to 75%) (IQR)
  - Whiskers (most extreme points not considered to be outliers)
  - Outliers

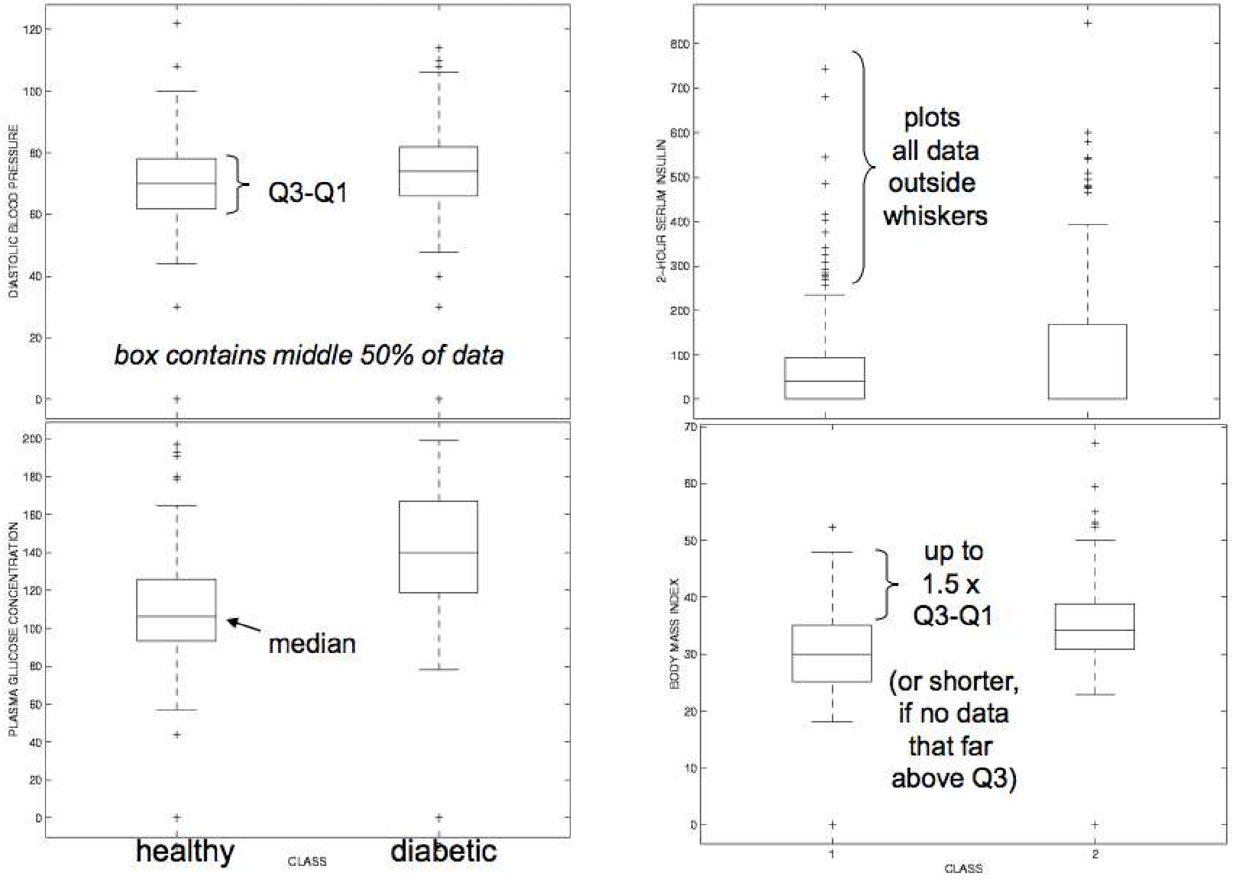
- Negatives
  - Over-plotting
  - Hard to tell distributional shape

Boxplot


100.0
66.7
33.3
0.0
Years
maximum
Q_3
IQR
median
Q_1
minimum
AGE
Variables

Box (and Whisker) Plots
- Pima Indian data-

```
> library(MASS)
> data(Pima.te)
```



Q3-Q1

box contains middle 50% of data

plots all data outside whiskers

median

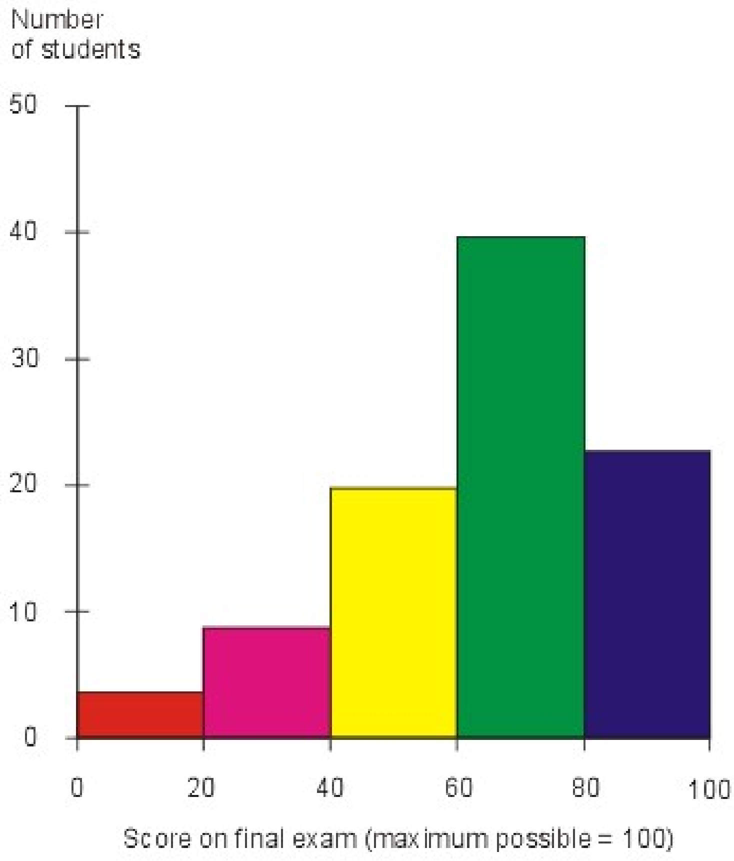
up to 1.5 x Q3-Q1

(or shorter, if no data that far above Q3)

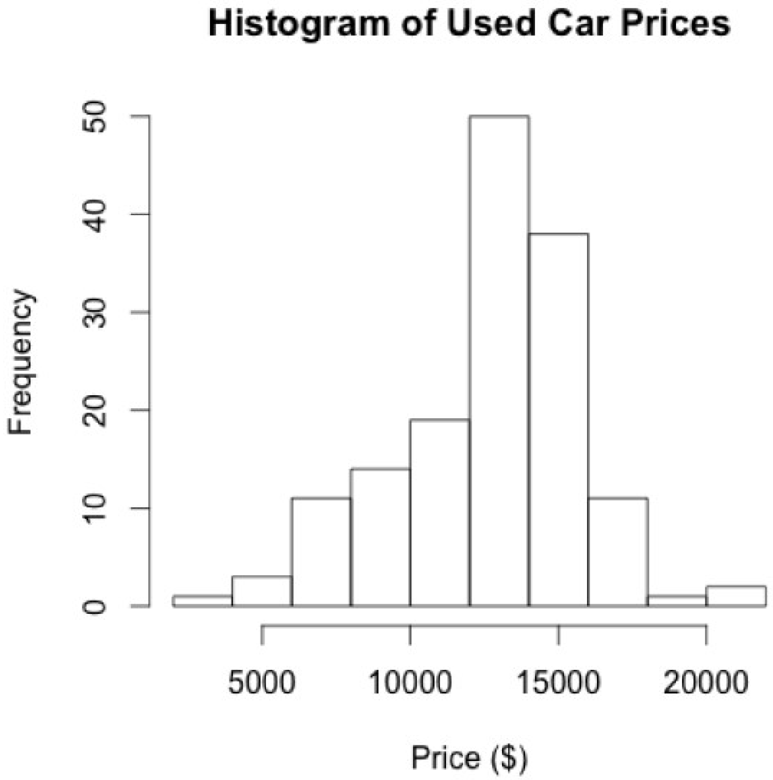
```
## e.g.
> boxplot(Pima.te$bmi ~ Pima.te$type)
```

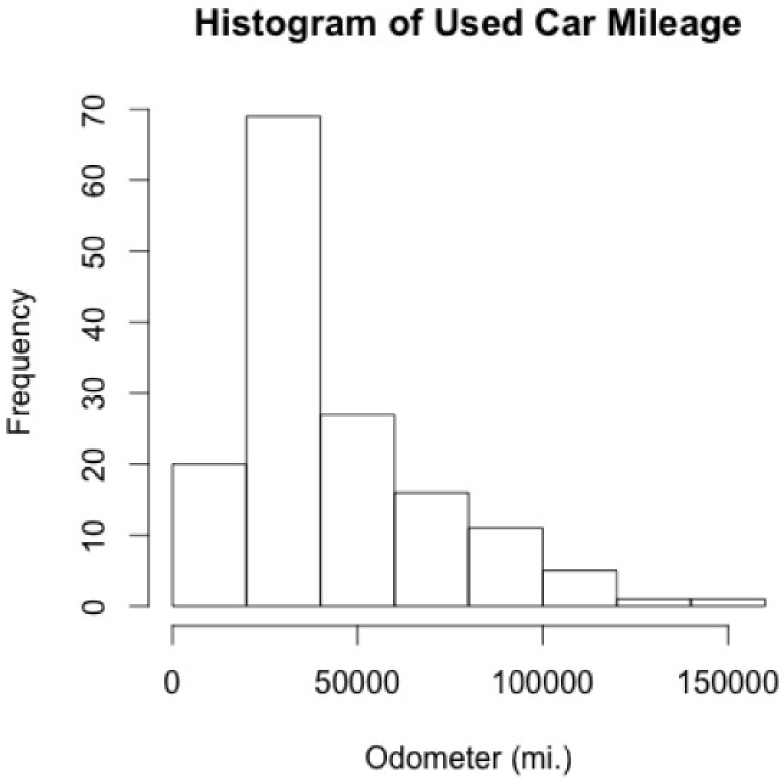
Histograms

- Histogram
  - Split data range into equal-sized bins
  - Count the number of data points falling into each bin
  - x axis: values of the variable
  - y axis: frequency (counts for each bin)

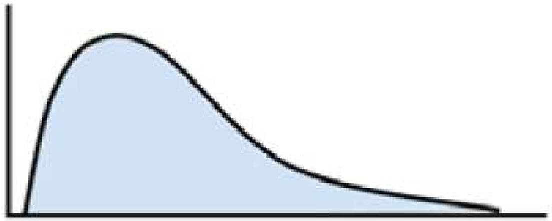


http://www.webquest.hawaii.edu/kahihi/mathdictionary/H/histogram.php

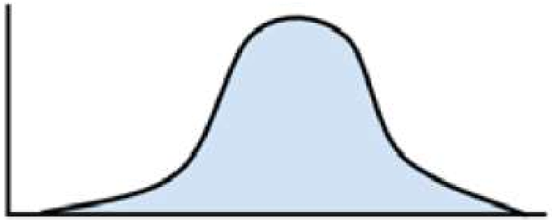

Histogram of Used Car Prices
Frequency
0
10
20
30
40
50
5000
10000
15000
20000
Price ($)


Histogram of Used Car Mileage
Frequency
0
10
20
30
40
50
60
70
0
50000
100000
150000
Odometer (mi.)

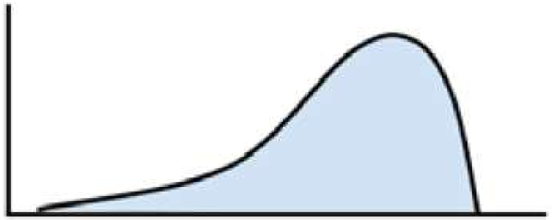
Shape of histograms



Right Skew

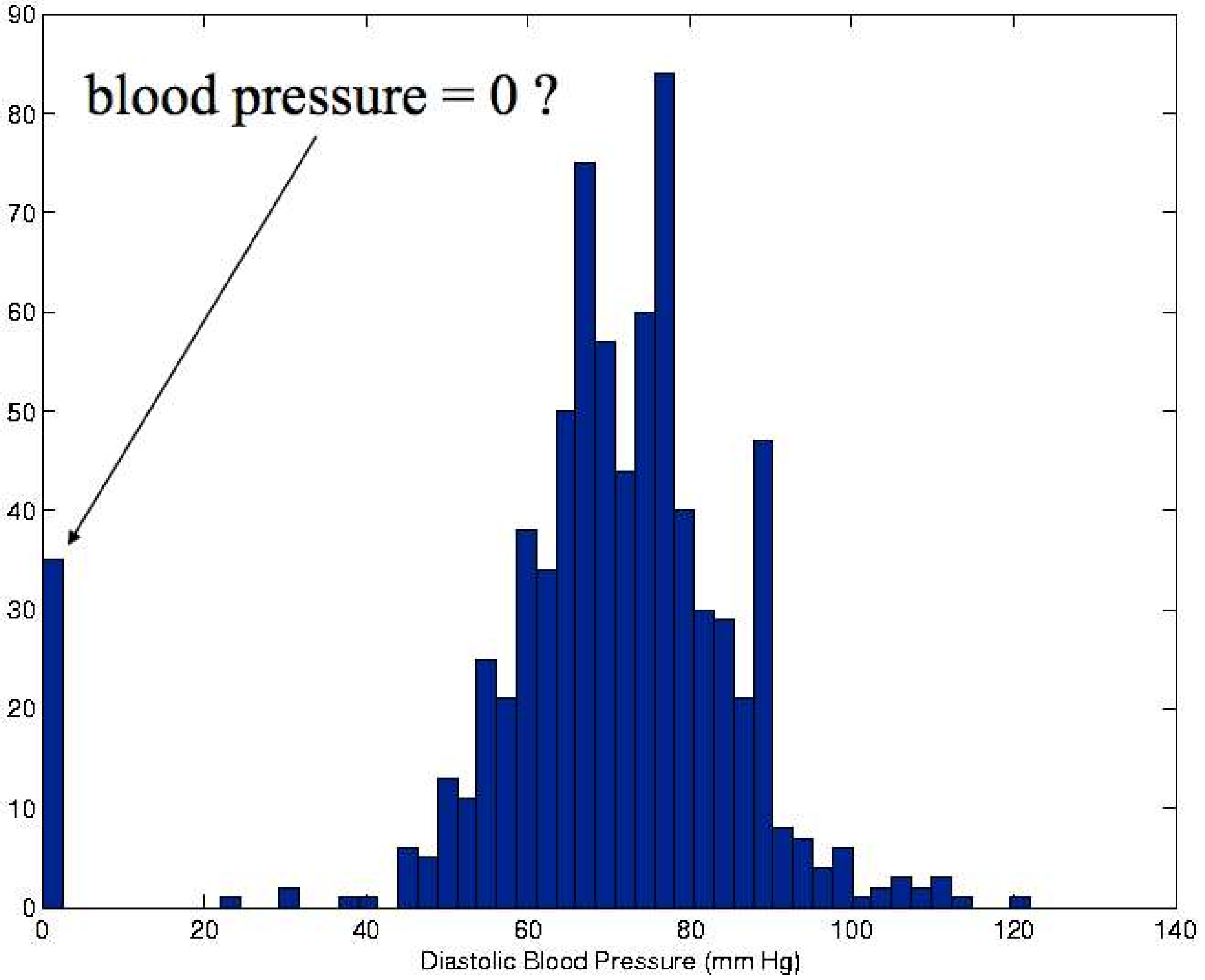


No Skew



Left Skew

Histogram detecting outliers



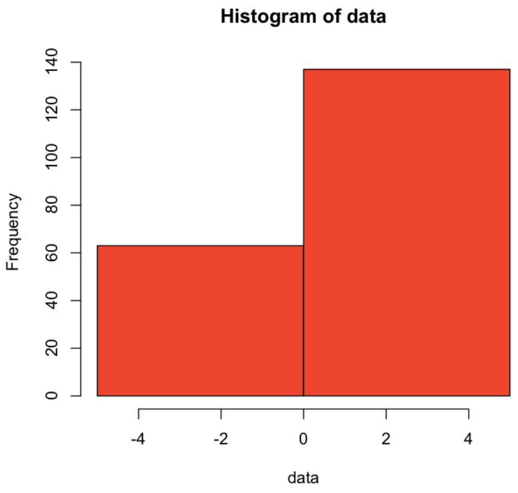
Issues with Histograms

- Histograms can be misleading for small data sets
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution
- Effective only with one variable
- Can smooth histogram using a variety of techniques

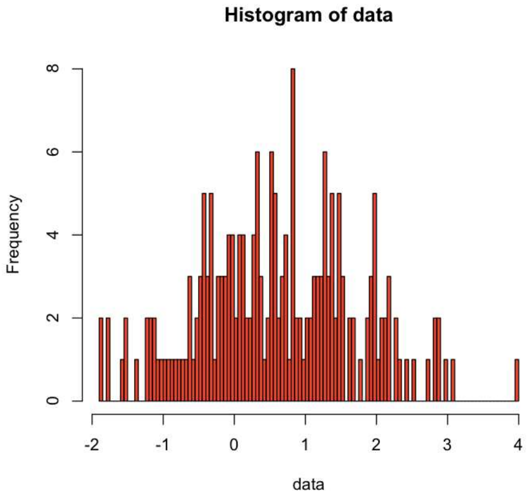
Effect of Bin Size on Histogram

```
> data <- c(rnorm(100), rnorm(100)+1)
```

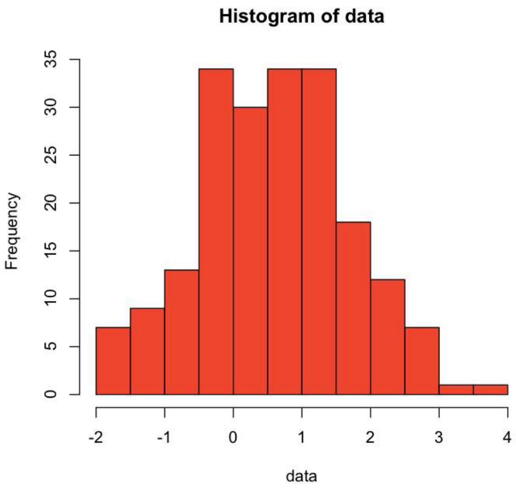
: Simulated 100 points from N(0,1) and 100 points from N(1,1)



```
> hist(data,breaks=2, col="red")
```



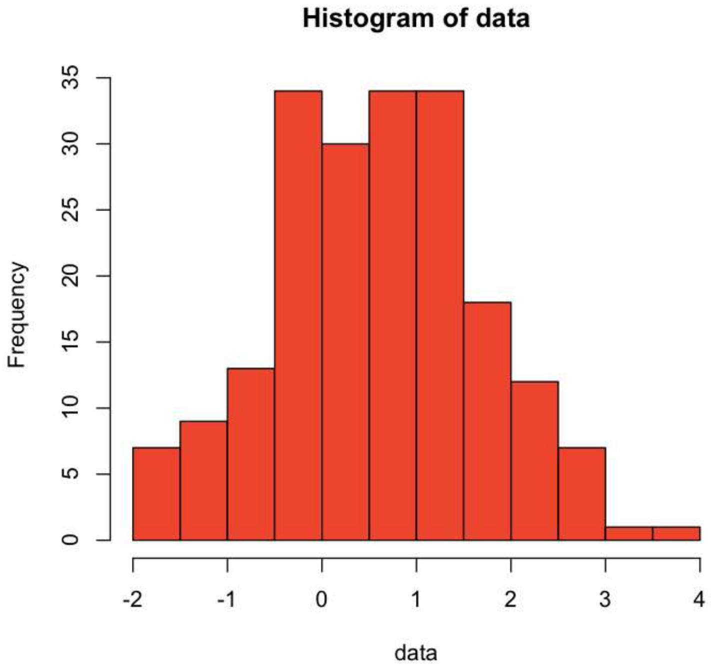
```
> hist(data,breaks=100, col="red")
```

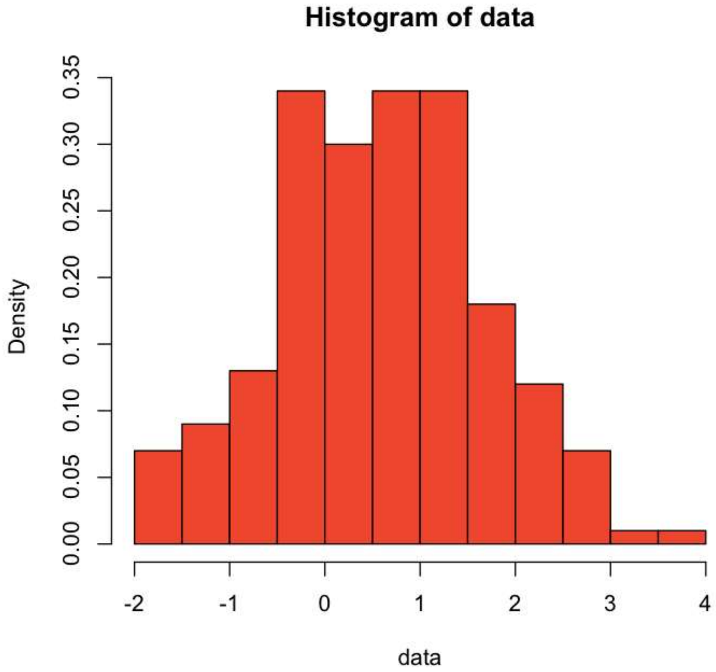


```
> hist(data,breaks=10, col="red")
```

More on Histograms

- Frequency histogram vs. density histogram





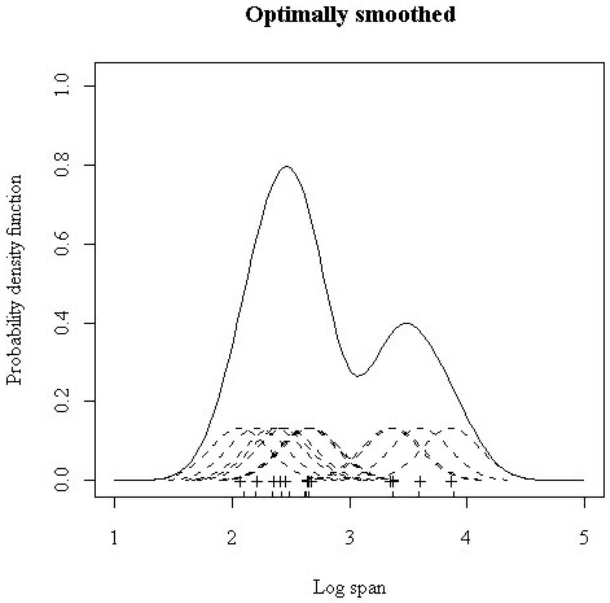
```
> hist(data,breaks=10, freq=F, col="red")
```

Smoothed Histograms - Density Estimates

- Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.

$$\hat{f}(x) = \tfrac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h is the kernel width

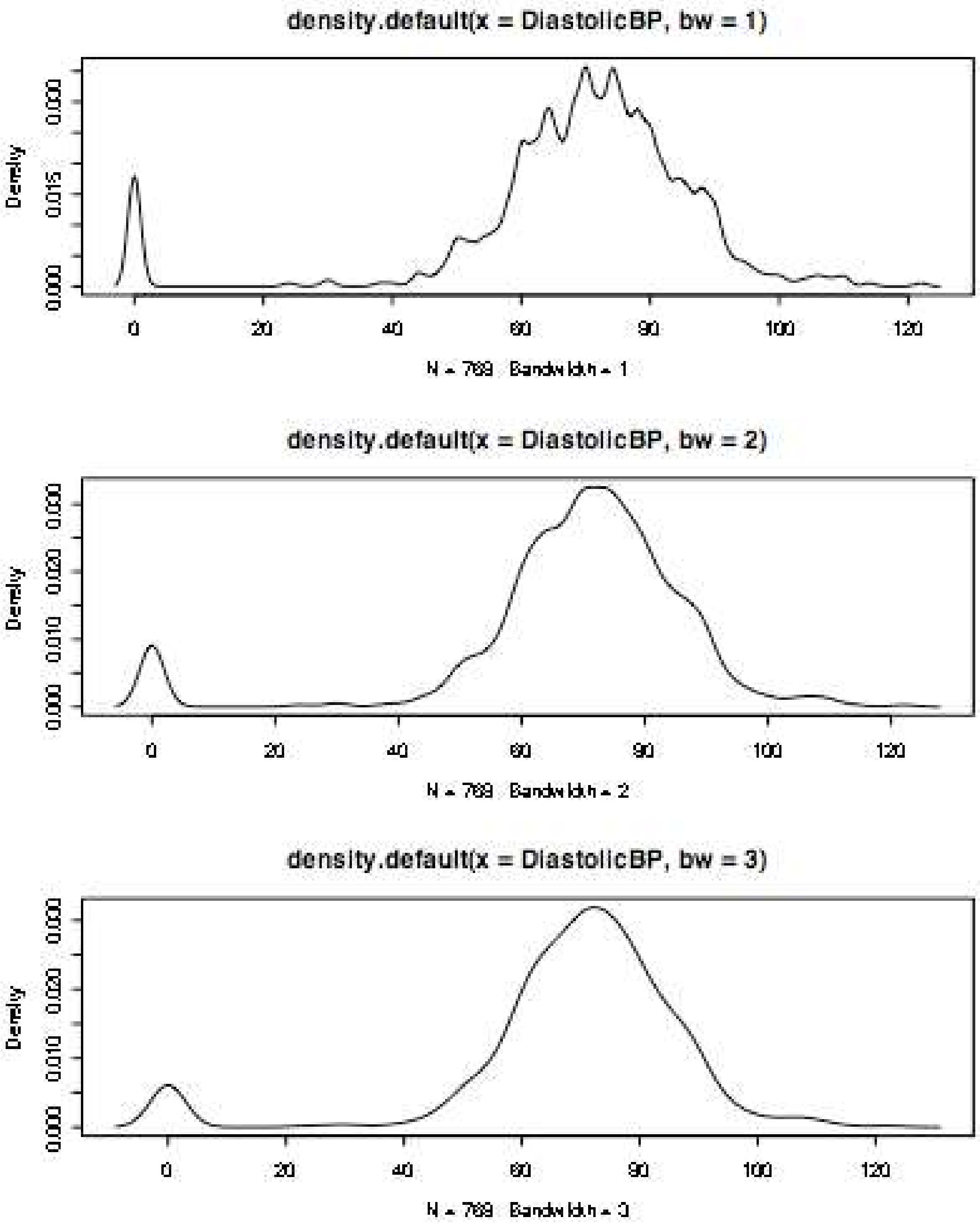


- Gaussian kernel is common:

$$Ce^{-\frac{1}{2}\left(\frac{x - x(i)}{h}\right)^2}$$

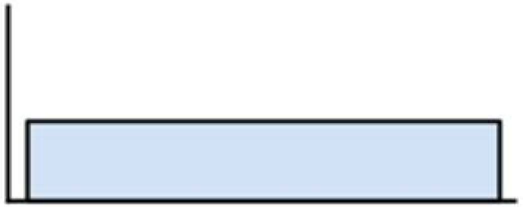
Bandwidth choice is an art

Usually want to try several



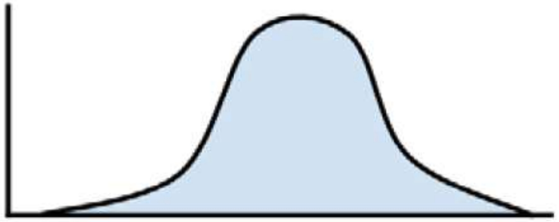
Understanding numeric data

- Uniform distribution



Uniform Distribution

- Normal distribution



Normal Distribution

Exploring categorical variables

- Categorical data is examined using tables rather than summary statistics
  - e.g. one-way table

- Measuring the central tendency – the mode
  - The value occurring most often
  - Often used for categorical data
  - e.g. in the used car data, the mode of the Year variable is 2010, the models for Color is Black, etc.

Exploring relationships between variables

- Do relationships between the model and color data provide insight into the types of cars we are examining?

- Bivariate, or multivariate relationships

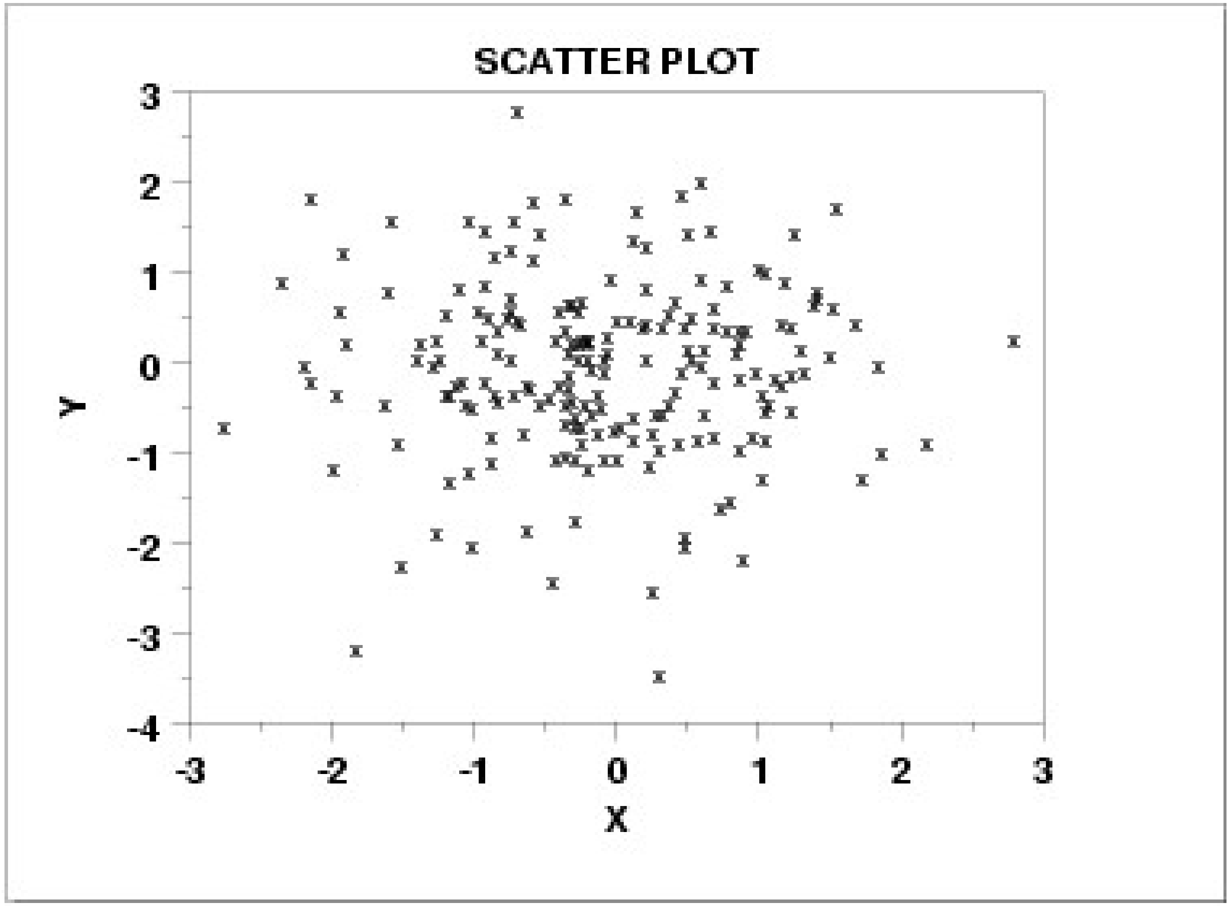
- Scatter plots, two-way cross-tabulation (contingency table)

2D Scatter plots

- standard tool to display relation between 2 variables

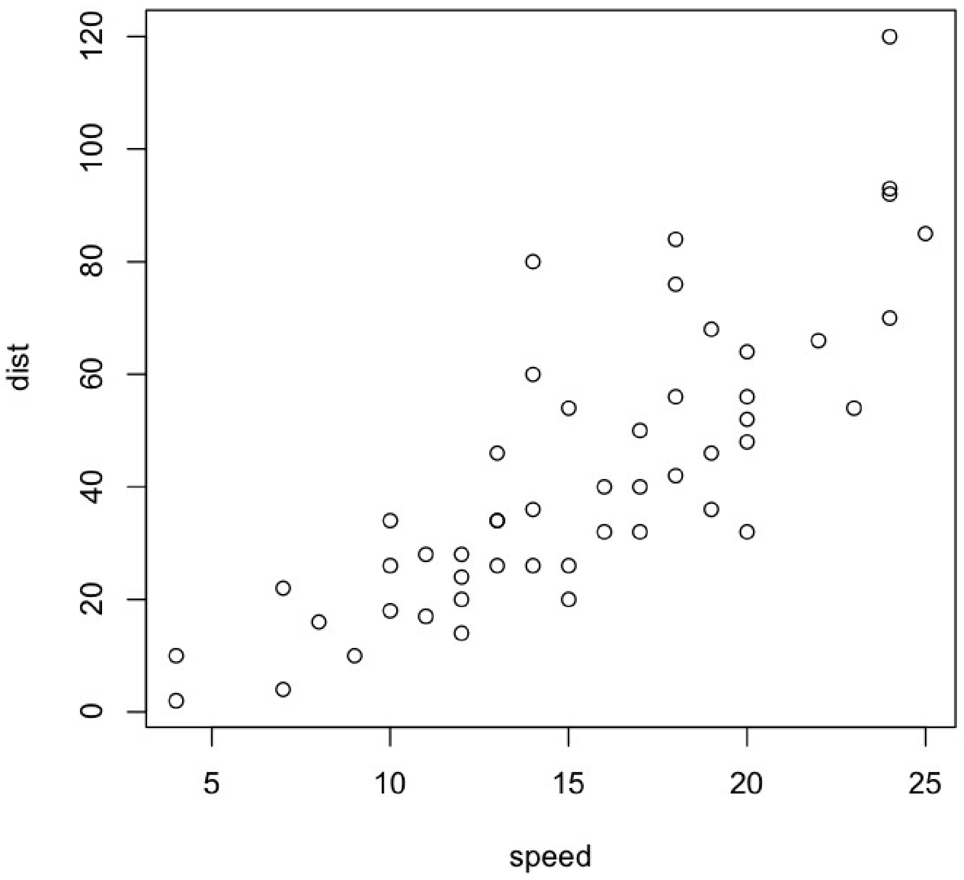
- Useful to answer
  - x and y related?
  - Variance(y) depend on x?
  - Outliers?

Scatter Plot: No apparent relationship



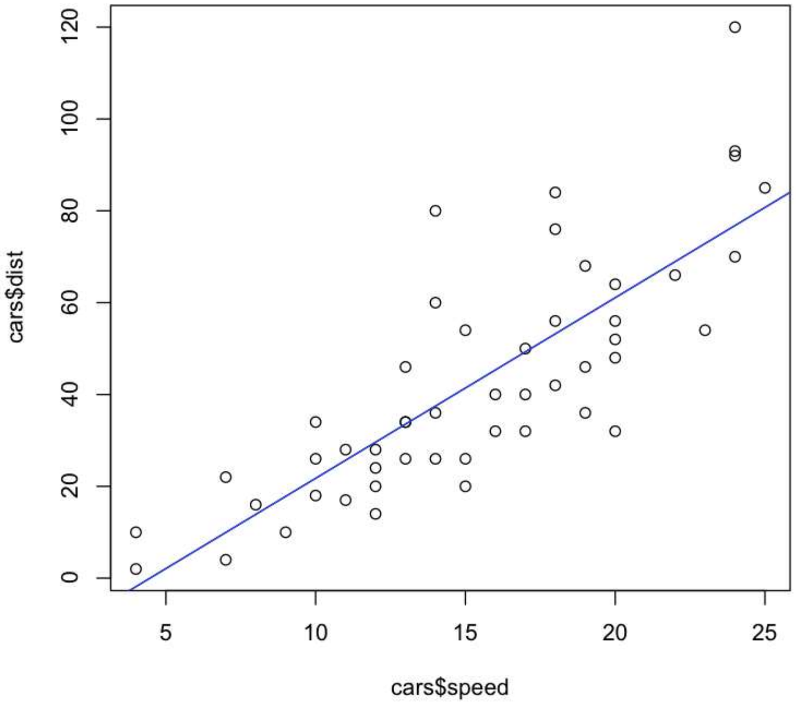
Scatter plot

- Speed and Stopping Distances of Cars -

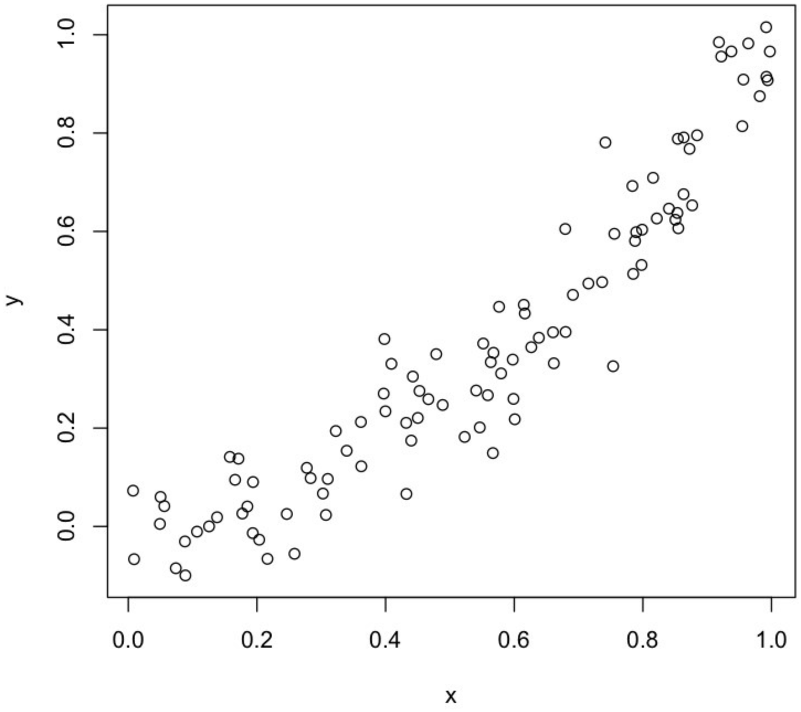


Scatter plot

linear relation

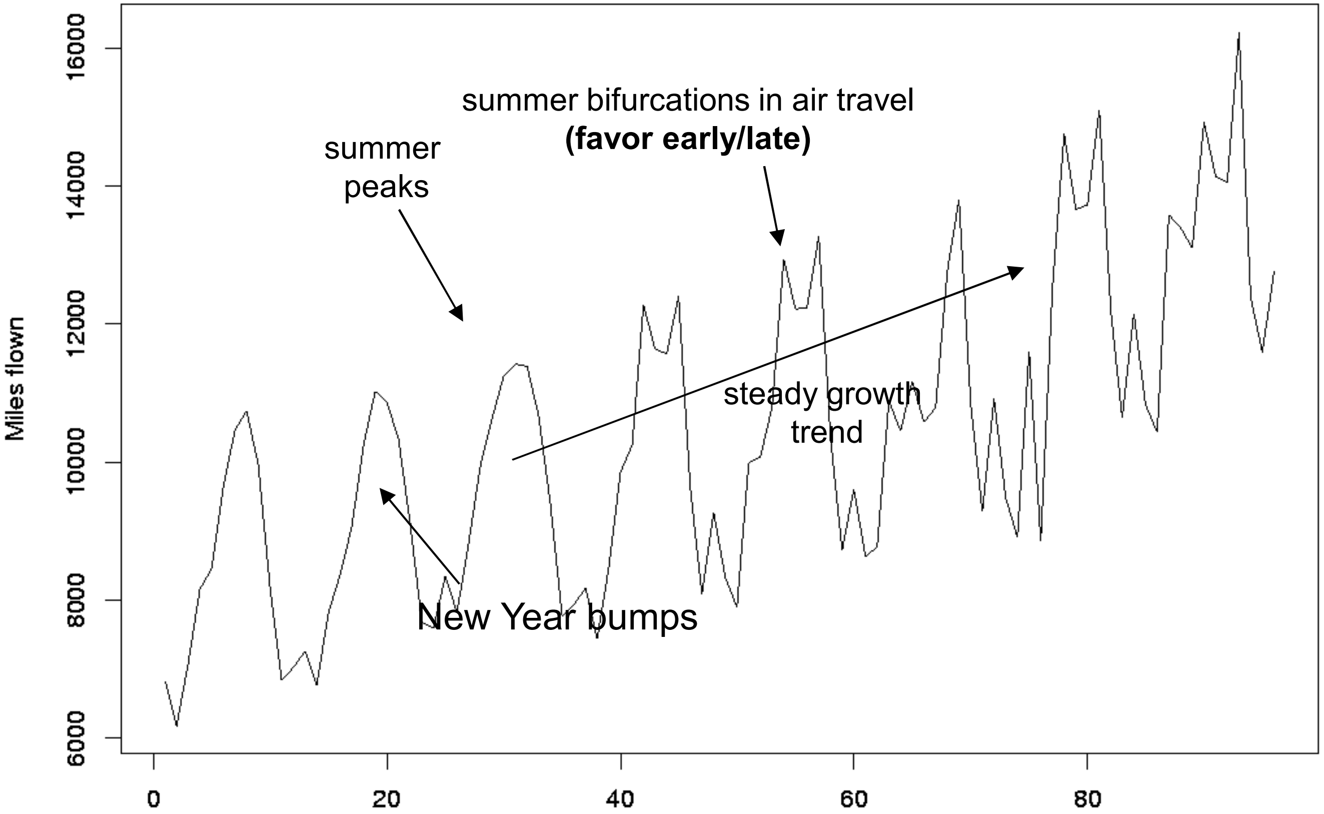


quadratic relation



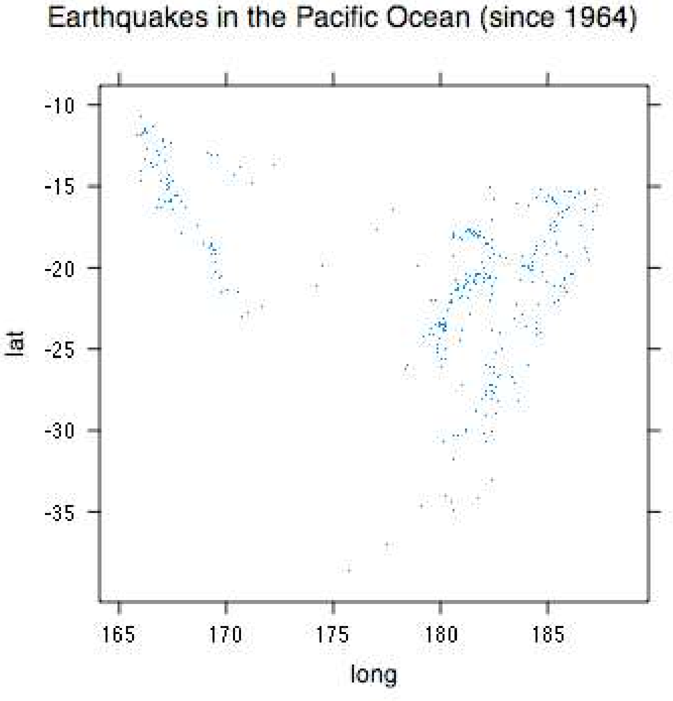
Time Series

If your data has a temporal component, be sure to exploit it



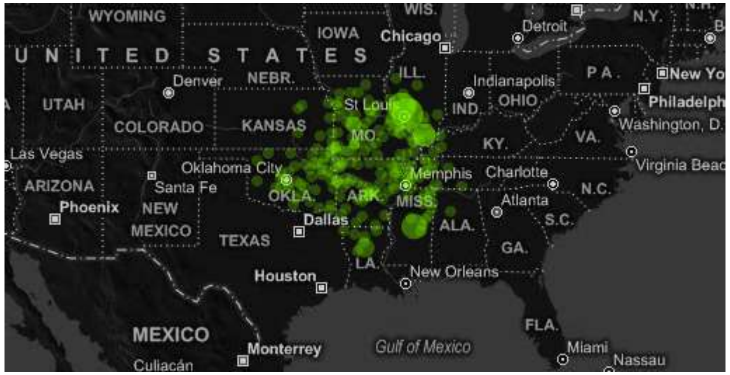
Spatial Data

- If your data has a geographic component, be sure to exploit it
- Data from cities/states/zip cods – easy to get lat/long
- Can plot as scatterplot



Spatio-temporal data

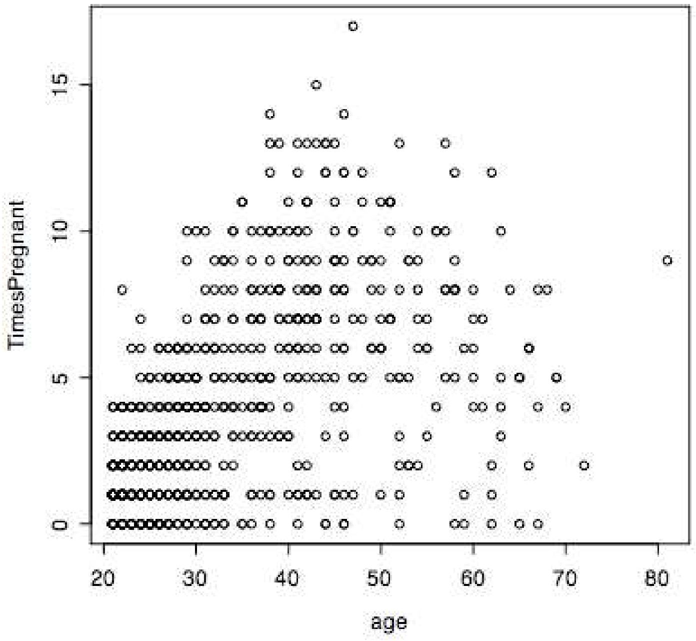
- spatio-temporal data
  - http://projects.flowingdata.com/walmart/ (Nathan Yau)

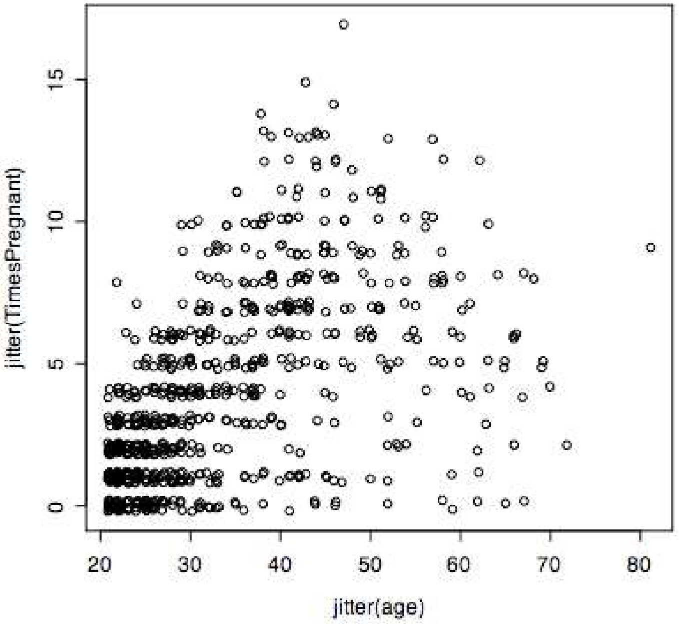


  - But, fancy tools not needed! Just do successive scatterplots to (almost) the same effect

Jittering

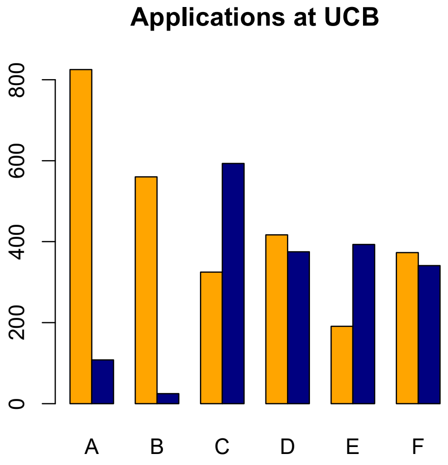
- Jittering points helps too

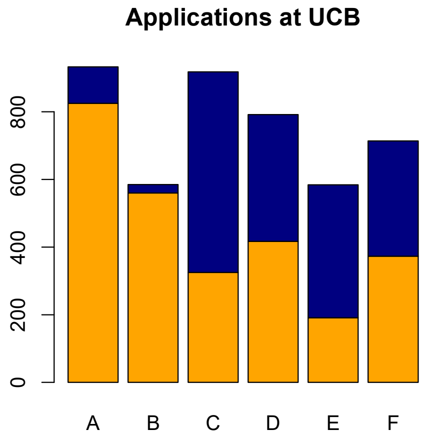




Barcharts and Spineplots

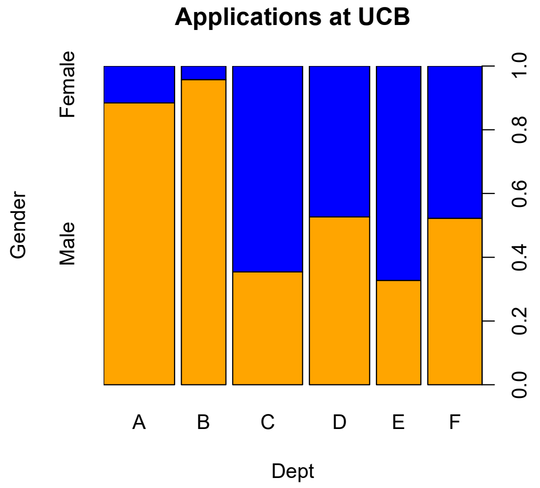
stacked barcharts can be used to compare continuous values across two or more categorical ones.





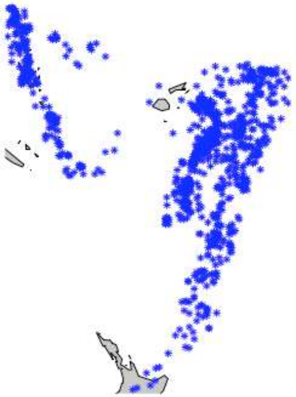
orange=M blue=F

spineplots show proportions well, but can be hard to interpret

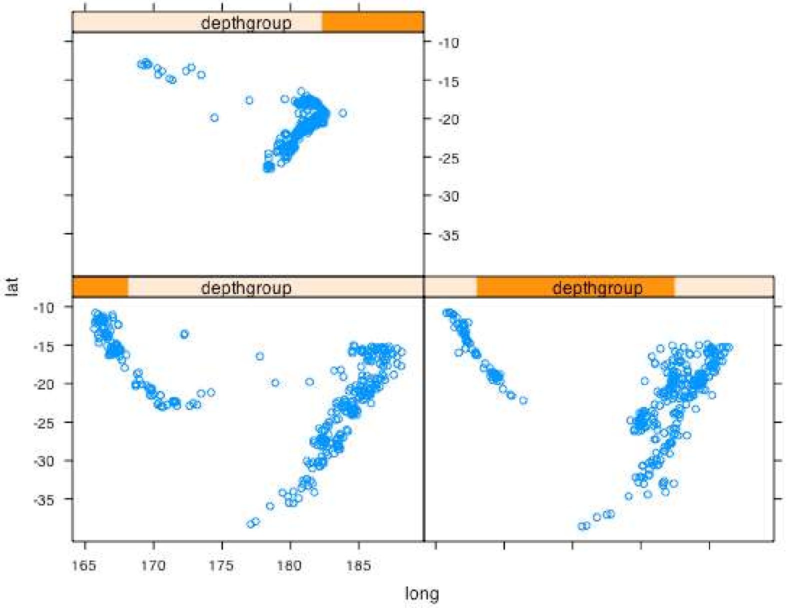


Multivariate: More than two variables

- Get creative!
- Conditioning on variables
  - trellis or lattice plots
  - Infinite possibilities

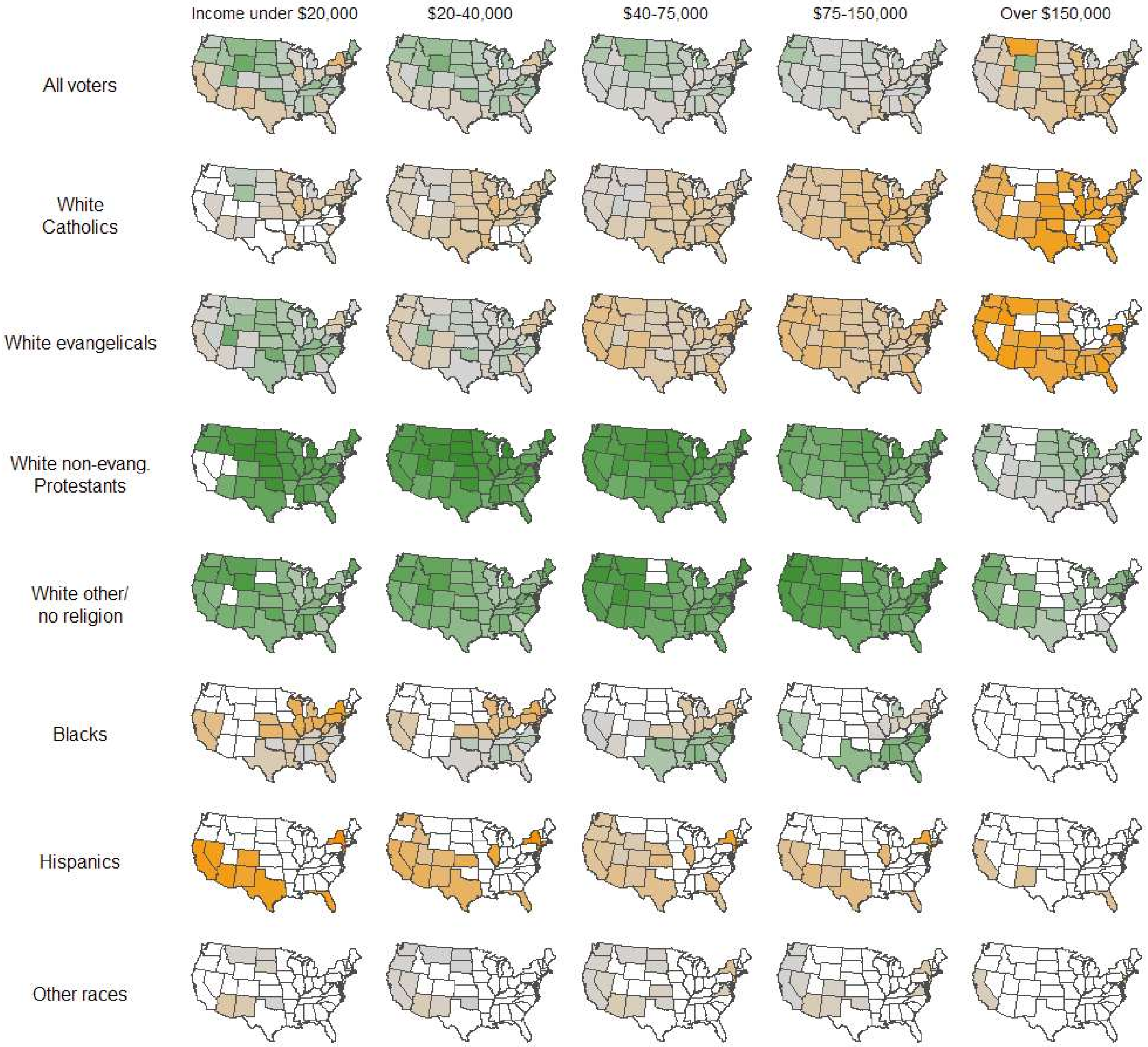


- Earthquake data:
  - locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964
  - Data collected on the severity of the earthquake



How many dimensions are represented here?

2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

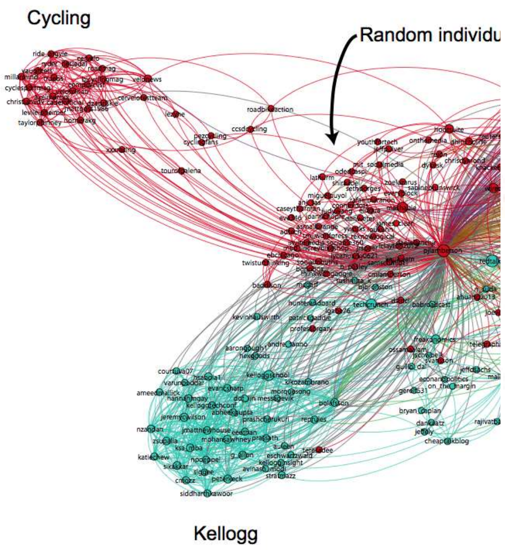


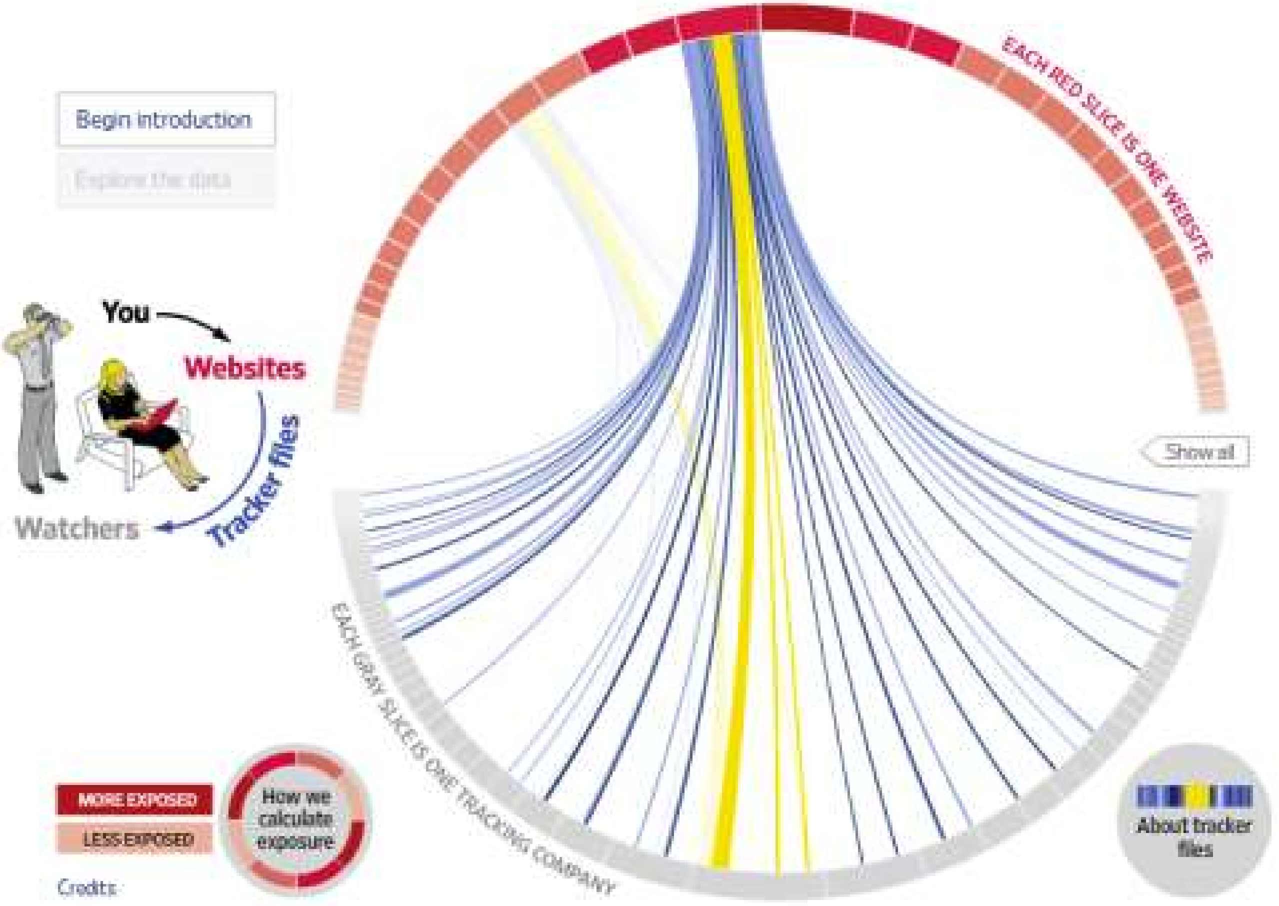
Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

Andrew Gelman blog 7/1

Networks and Graphs

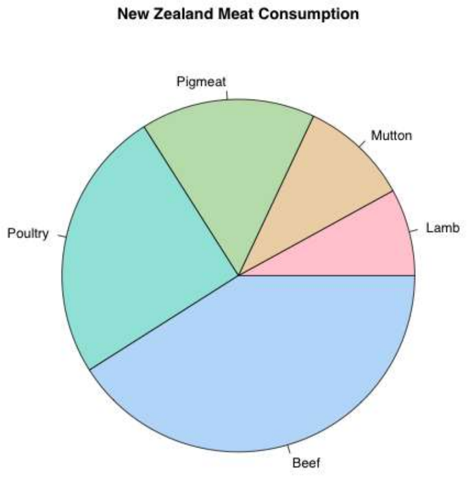
- Visualizing networks is helpful, even if is not obvious that a network exists





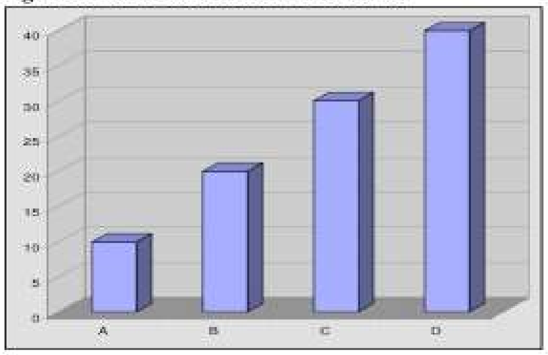
What’s missing?

- pie charts
  - very popular
  - good for showing simple relations of proportions
  - Human perception not good at comparing arcs
  - barplots, histograms usually better (but less pretty)

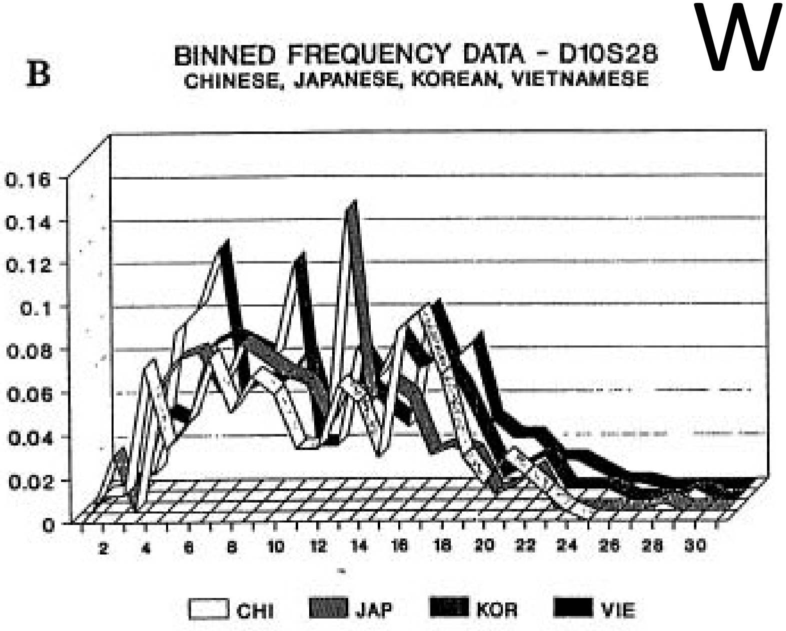


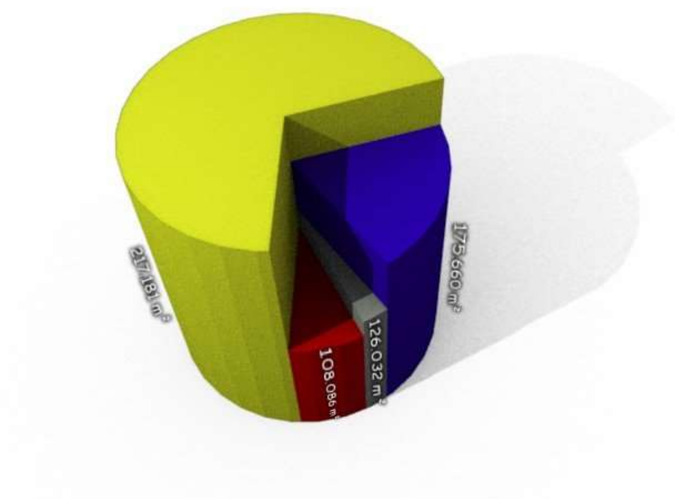
- 3D
  - nice to be able to show three dimensions
  - hard to do well
  - often done poorly
  - 3d best shown through “spinning” in 2D
    - uses various types of projecting into 2D
    - http://www.stat.tamu.edu/~west/bradley/

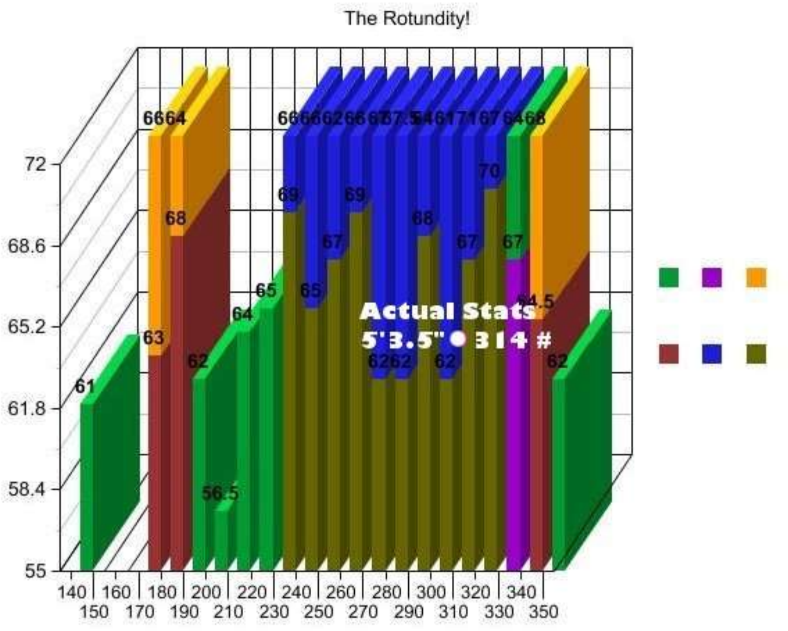
Figure 1. Three-dimensional bar chart.

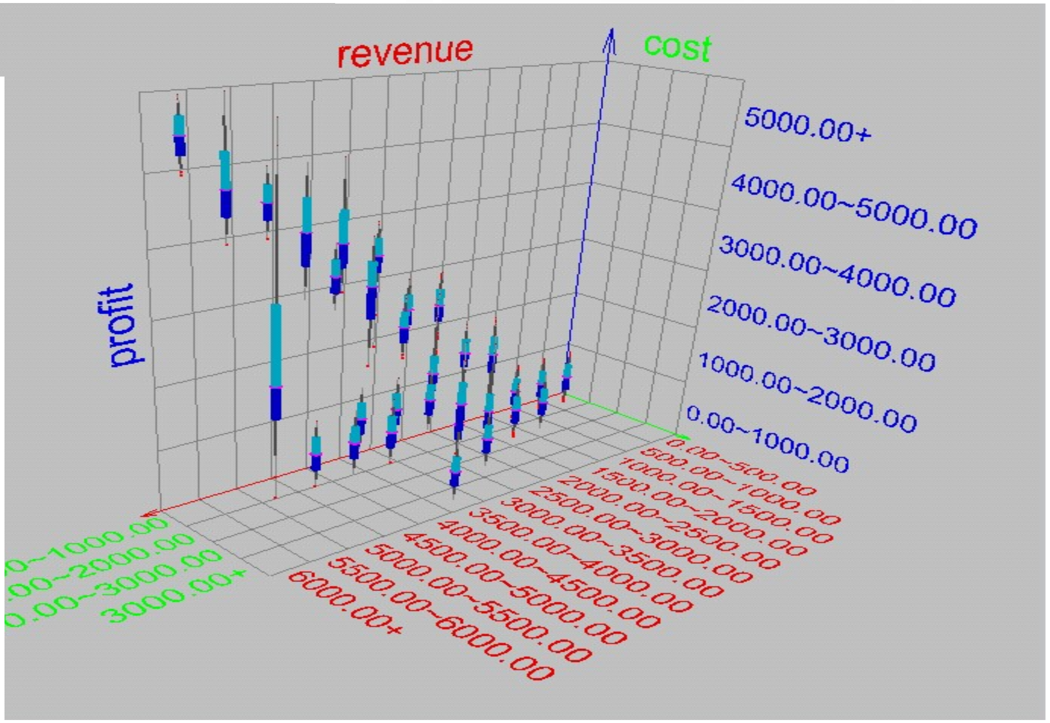


Worst graphic in the world?









Dimension Reduction

- One way to visualize high dimensional data is to reduce it to 2 or 3 dimensions
  - Variable selection
    - e.g. stepwise
  - Principle Components
    - find linear projection onto p-space with maximal variance
  - Multi-dimensional scaling, t-SNE
    - takes a matrix of (dis)similarities and embeds the points in p-dimensional space to retain those similarities

(More on this later)

Visualization done right

- Hans Rosling @ TED
- http://www.youtube.com/watch?v=jbkSRLYSojo