CS57300
PURDUE UNIVERSITY

JANUARY 24, 2019

DATA MINING

ADMINISTRATIVE NOTE

- Start to think about your project now!
  - Form a team of 2-4 people, e.g., via Piazza
  - Decide upon a topic
    - Topics that you are interested in
    - Will be great if it connects to your research
    - Kaggle competition is ok
  - Collect your data now if needed

OVERVIEW

- Task specification
- Knowledge representation
- **Learning technique**
  - Search + scoring
- Prediction and/or interpretation

LEARNING TECHNIQUE

- Method to construct model or patterns from data
- **Model space**
  - Choice of knowledge representation defines a set of possible models or patterns
- **Scoring function**
  - Associates a numerical value (score) with each member of the set of models/patterns
- **Search technique**
  - Defines a method for generating members of the set of models/patterns, determining their score, and identifying the ones with the “best” score

EXAMPLE LEARNING PROBLEM

Task: Devise a rule to classify items based on the attribute ***x***

Knowledge representation:

If-then rules

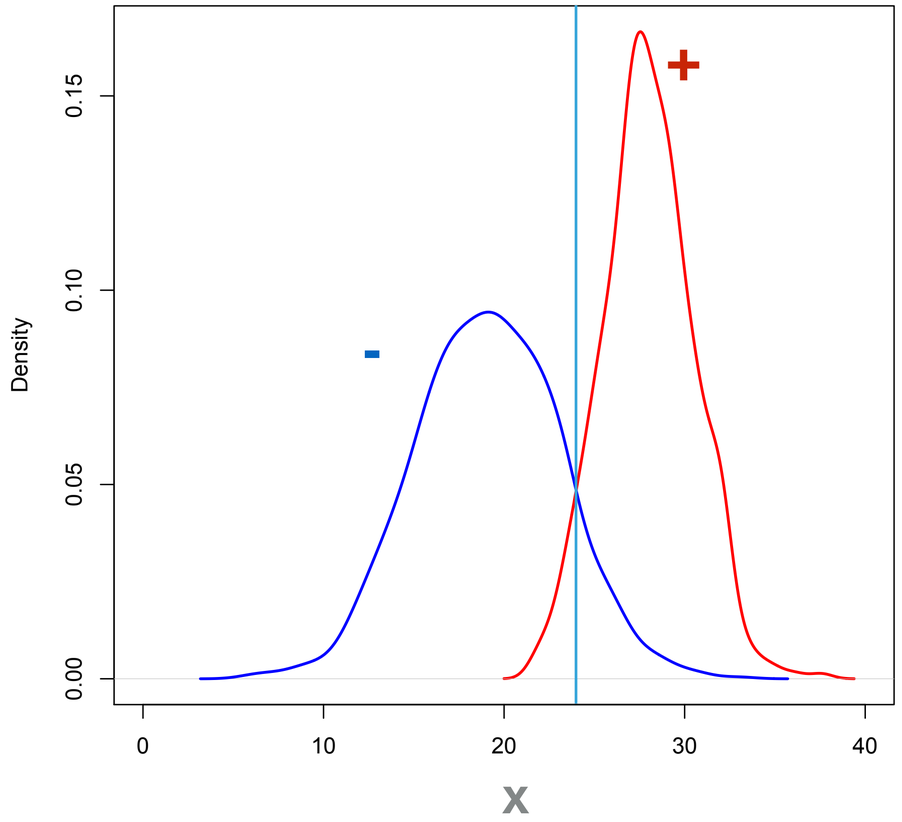
Example rule:

If x > 24 then **+**

Else **-**

What is the model space?

All possible thresholds



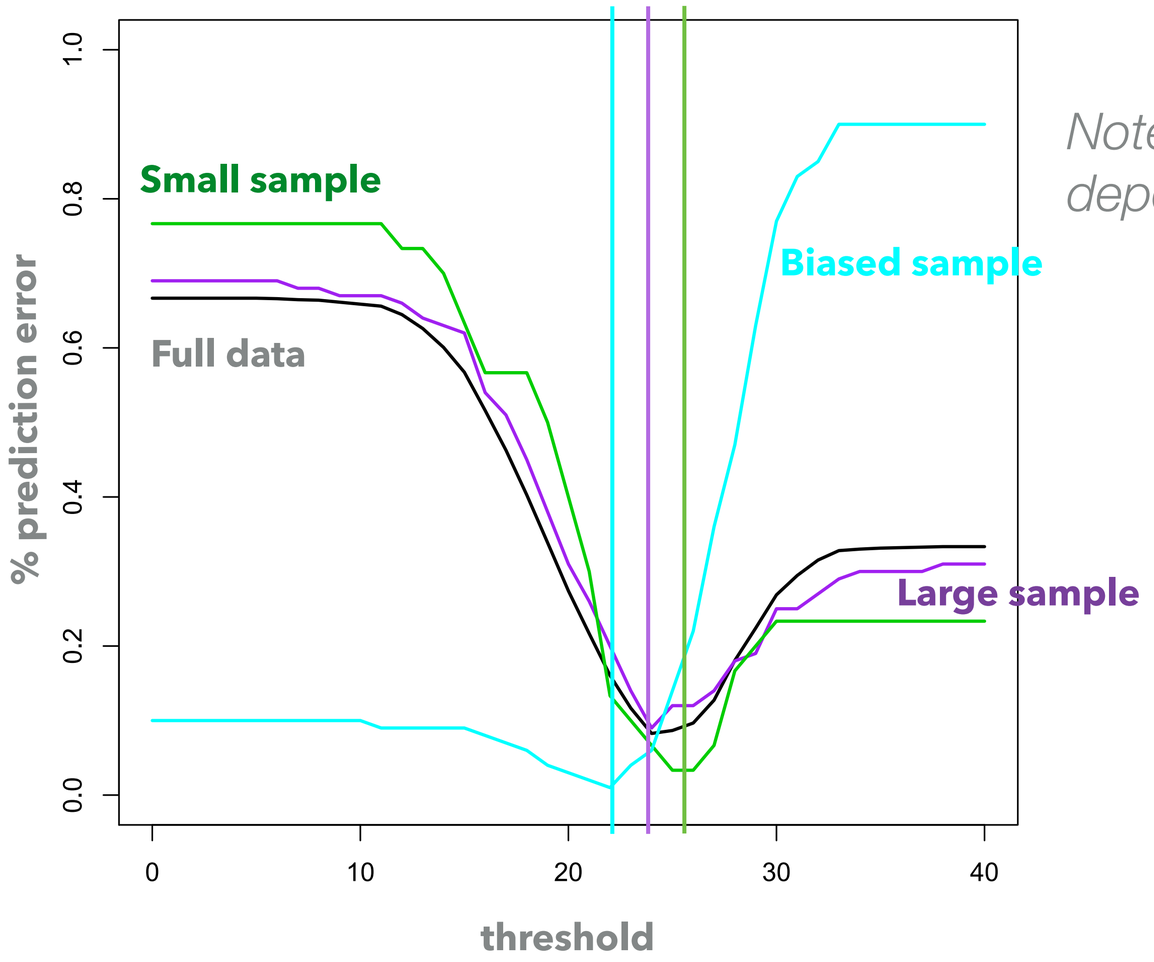
What score function?

Prediction error rate

SCORE FUNCTION OVER MODEL SPACE

Search procedure?

Try all thresholds, select one with lowest score



*Note: learning result depends on **data***

OVERVIEW

- Task specification
- Knowledge representation
- Learning technique
  - Search + Evaluation
- **Prediction and/or interpretation**

INFERENCE AND INTERPRETATION

- Prediction technique
  - Method to apply learned model to new data for prediction/analysis
  - Only applicable for predictive and some descriptive models
  - Prediction is often used during **learning** (i.e., search) to determine value of scoring function
- Interpretation of results
  - Objective: significance measures
  - Subjective: importance, interestingness, novelty

EXAMPLE: IDENTIFYING EMAIL SPAM

- Task
  - Design automatic spam detector that can differentiate between labeled emails
- Data
  - Table of relative word/ punctuation frequencies
- Knowledge representation
  - If/then rules with conjunctions of features
- Learning technique
  - **Search** over set of rules, **select** rule with maximum accuracy on training data

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

if (%george < 0.6) & (%you > 1.5) then spam
else email.

EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

- Data analysis approach that employs a number of (mostly graphical) techniques to:
  - Maximize insight into data
  - Uncover underlying structure
  - Identify important variables
  - Detect outliers and anomalies
  - Test underlying modeling assumptions
  - Develop parsimonious models
  - Generate hypotheses from data

DATA VISUALIZATION

VISUALIZATION

- Human eye/brain have evolved powerful methods to detect structure in nature
- Display data in ways that exploit human pattern recognition abilities
- Limitation: Can be difficult to apply if data size (number of dimensions or instances) is large

VISUALIZING/SUMMARIZING DATA

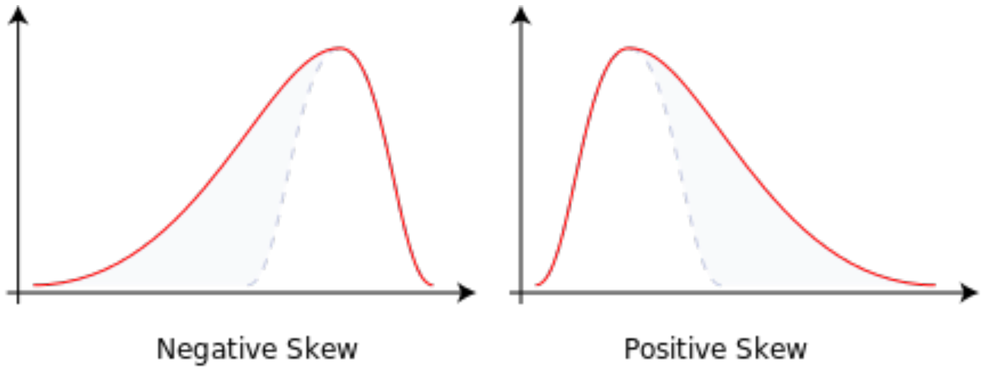
- Low-dimensional data
  - Summarizing data with simple statistics
  - Plotting raw data (1D, 2D, 3D)
- Higher-dimensional data
  - Dimensionality reduction: e.g., principal component analysis

DATA SUMMARIZATION

- Measures of location
  - Mean: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x(i)$
  - Median: value with 50% of points above and below
  - Quartile: value with 25% (75%) points below
  - Mode: most common value

DATA SUMMARIZATION

- Measures of dispersion or variability
  - Variance: $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2$
  - Standard deviation: $\hat{\sigma}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2}$
  - Range: difference between max and min point
  - Interquartile range: difference between 1st and 3rd Q
  - Skew: $\frac{\sum_{i=1}^{n}(x(i)-\hat{\mu})^3}{(\sum_{i=1}^{n}(x(i)-\hat{\mu})^2)^{\frac{3}{2}}}$

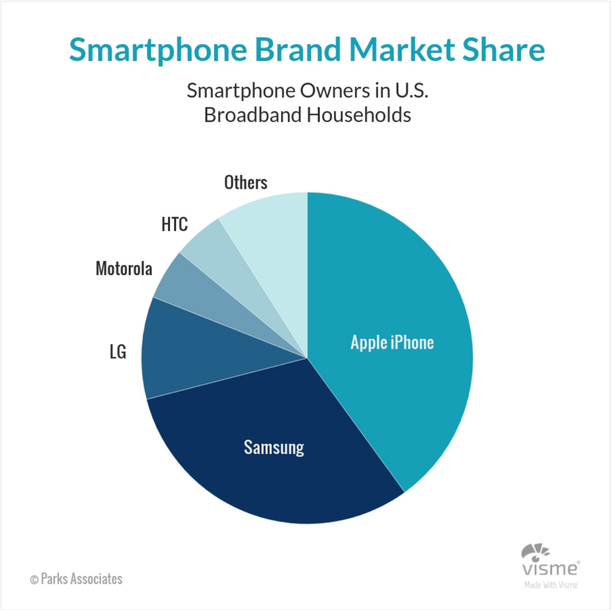


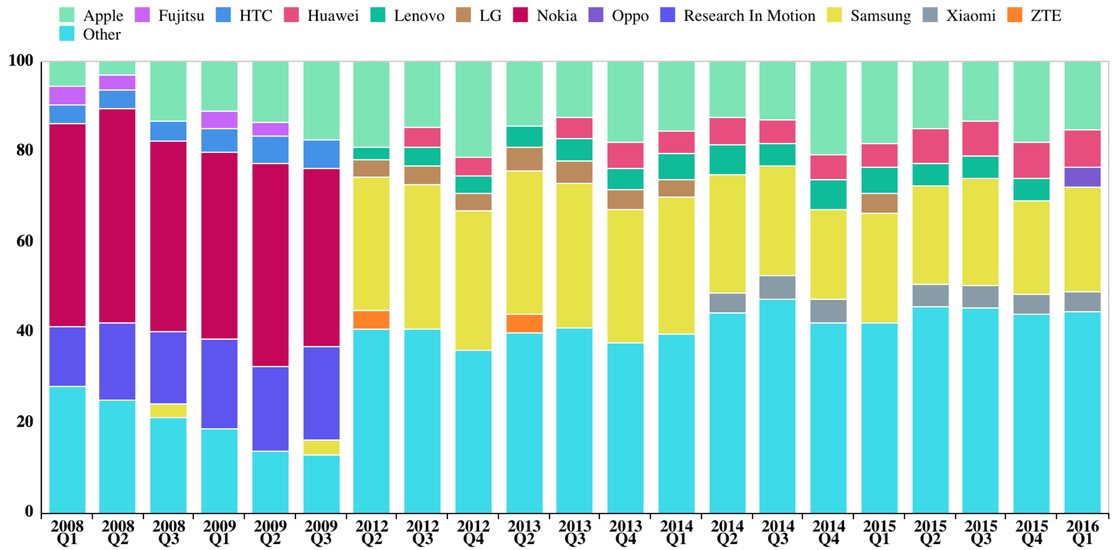
DATA VISUALIZATION

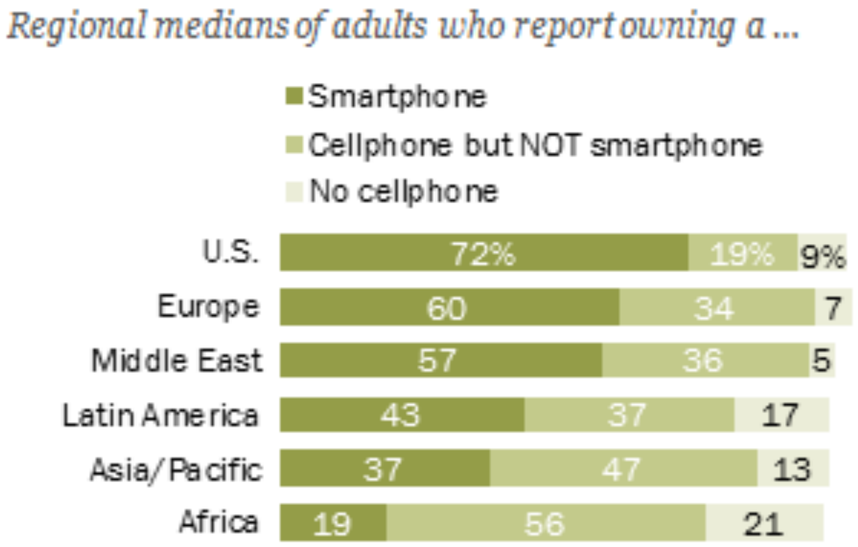
- Serve for different purposes
  - Composition: e.g., see for a discrete dimension x_i, the fraction of each values
  - Distribution: e.g., see the distribution of a continuous dimension of data x_i
  - Comparison:
    - Compare values of two continuous dimensions of the data, x_i and x_j
    - Given discrete x_i, compare the values of x_j when x_i takes different values.
  - Relationship: e.g., examine the relationship between x_i and x_j

COMPOSITION

- Pie charts
- Stacked bars
  - Temporal trends
  - Compare across groups

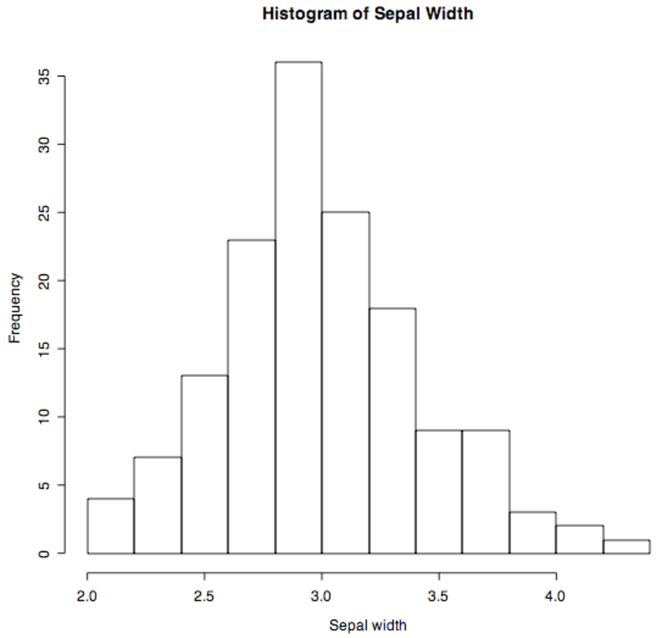






DISTRIBUTION: HISTOGRAMS (1D)

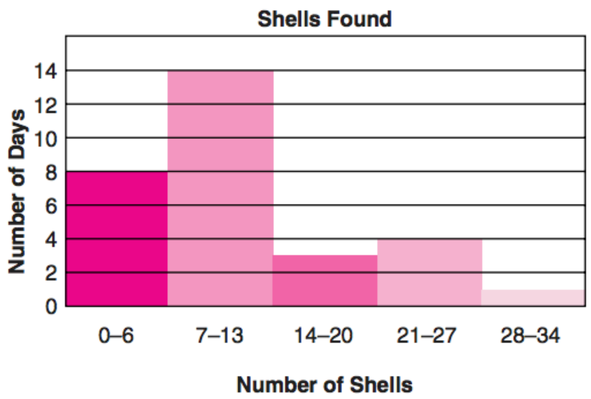
- Most common plot for univariate data
- Split data range into equal-sized bins, count number of data points that fall into each bin
- Graphically shows:
  - Center (location)
  - Spread (scale)
  - Skew
  - Outliers
  - Multiple modes



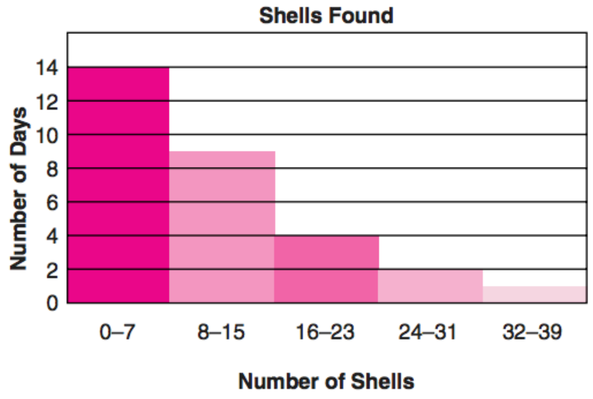
HISTOGRAM LIMITATIONS

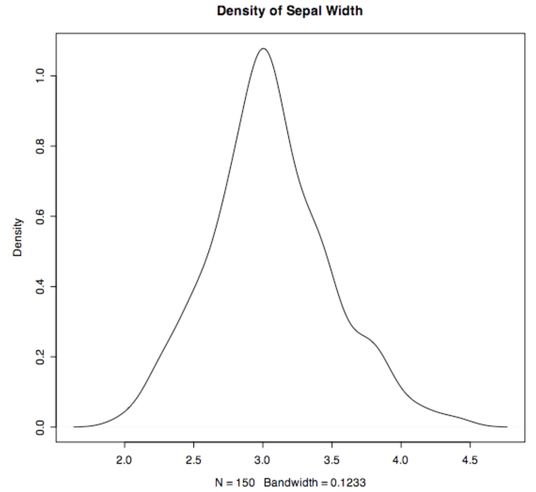
- Histograms can be misleading for small datasets
  - Slight changes in the data or binning approach can result in different histograms
- Solution: smoothed density plots
  - Use kernel function to estimate density at each point x, pools information from neighboring points

1.



2.

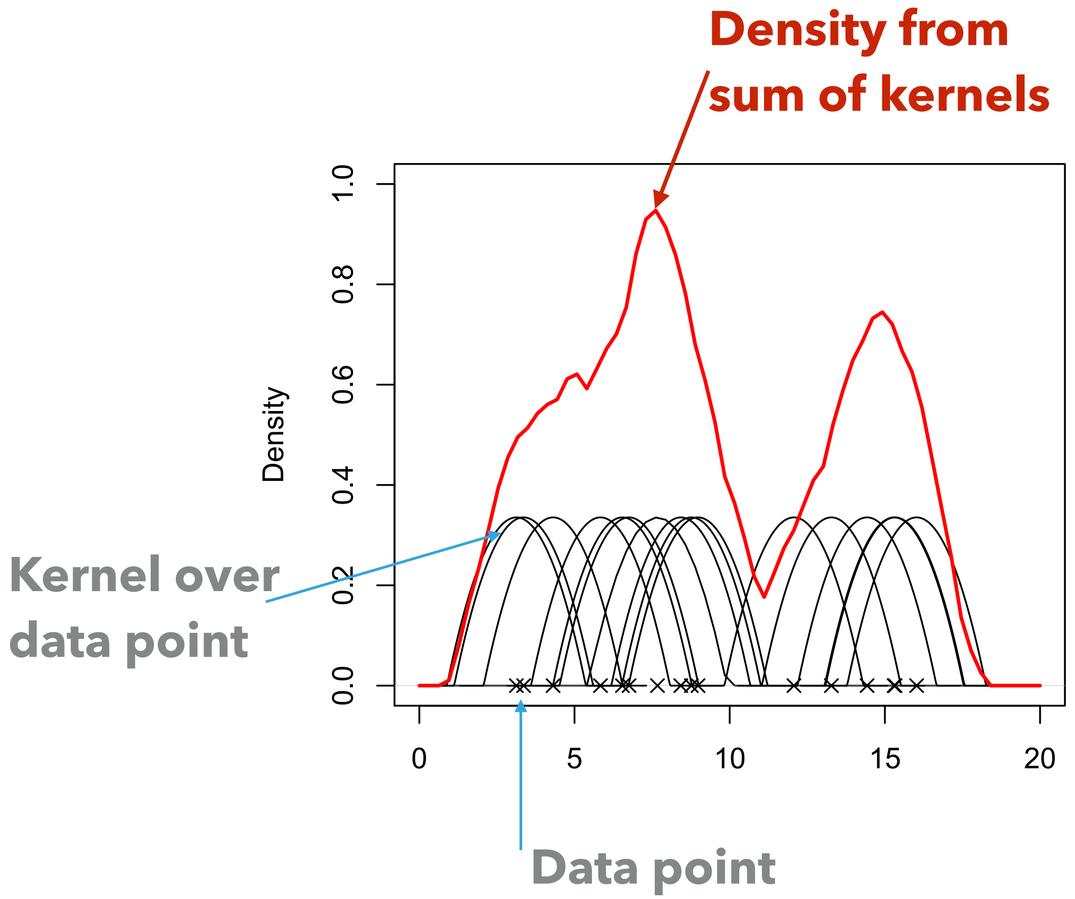




KERNEL DENSITY ESTIMATION

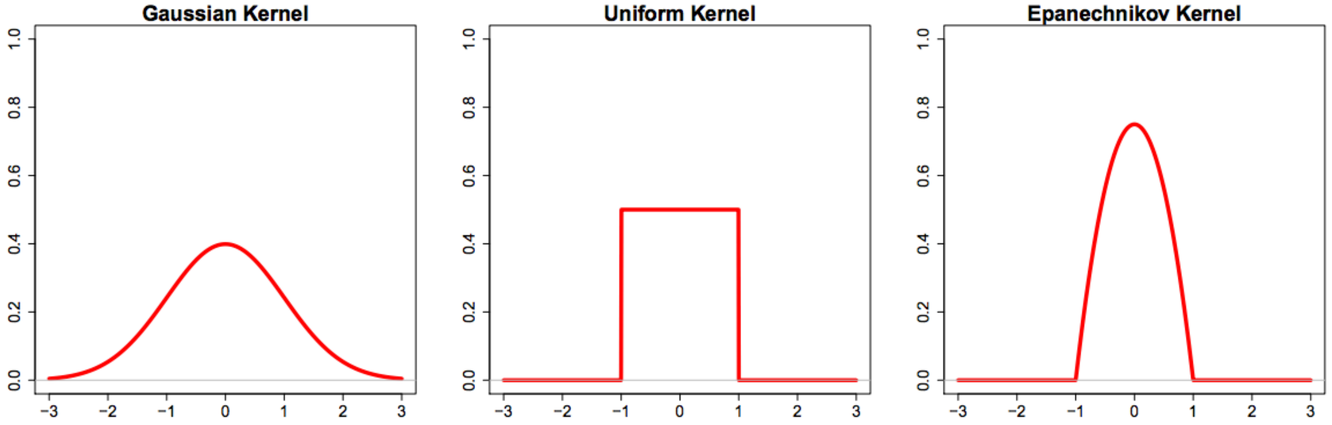
- Estimated density is:

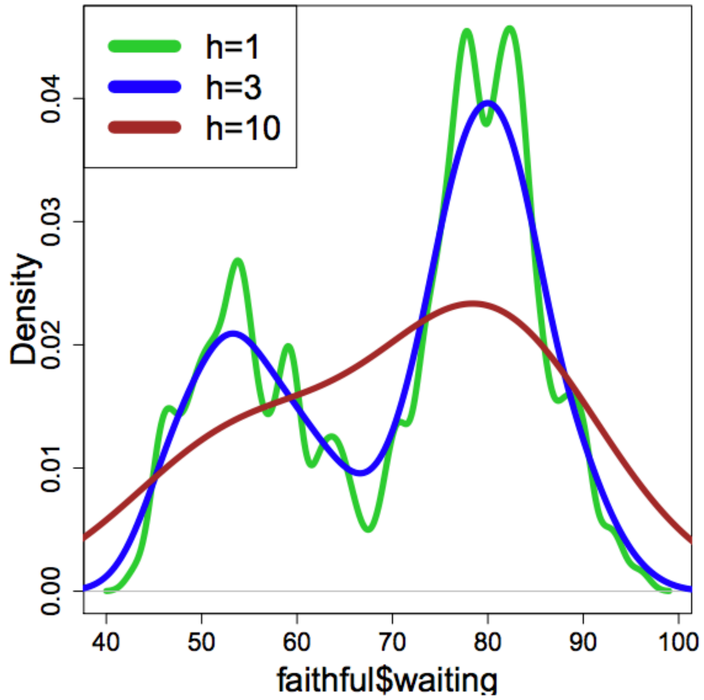
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$



KERNEL DENSITY ESTIMATION

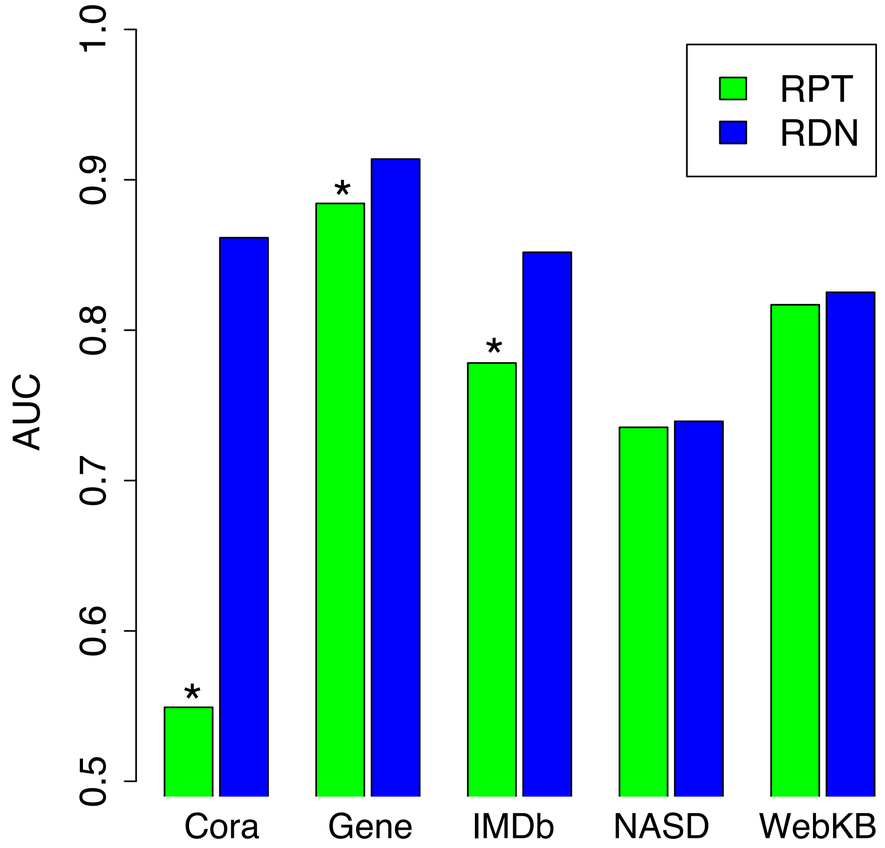
- Two parameters:
  - Kernel function K (e.g., Gaussian, Epanechnikov)
  - Bandwidth h





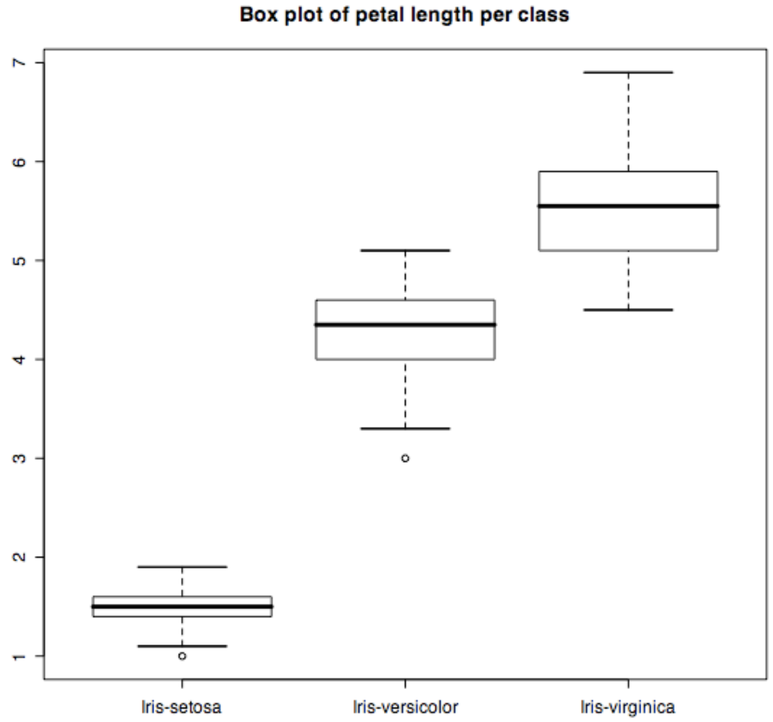
COMPARISON: BAR PLOTS

- Compare values of two continuous dimensions of the data, x_i and x_j



COMPARISON: BOX PLOT (2D)

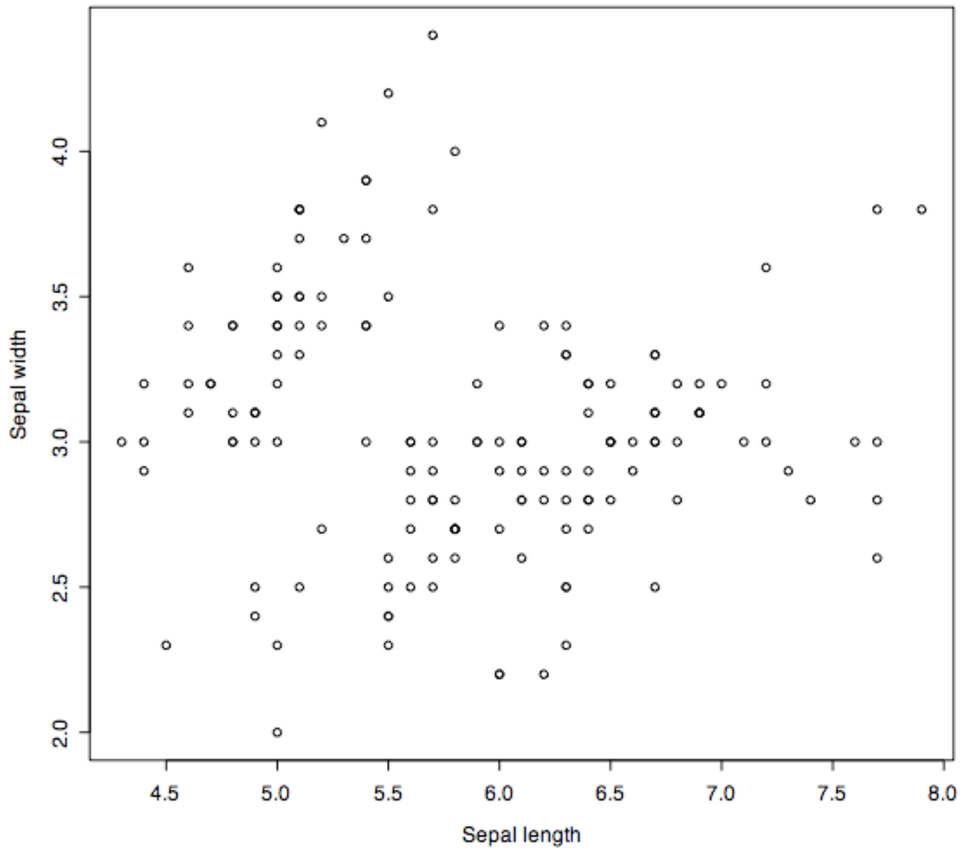
- Display relationship between discrete and continuous variables
- For each discrete value X, calculate quartiles and range of associated Y values



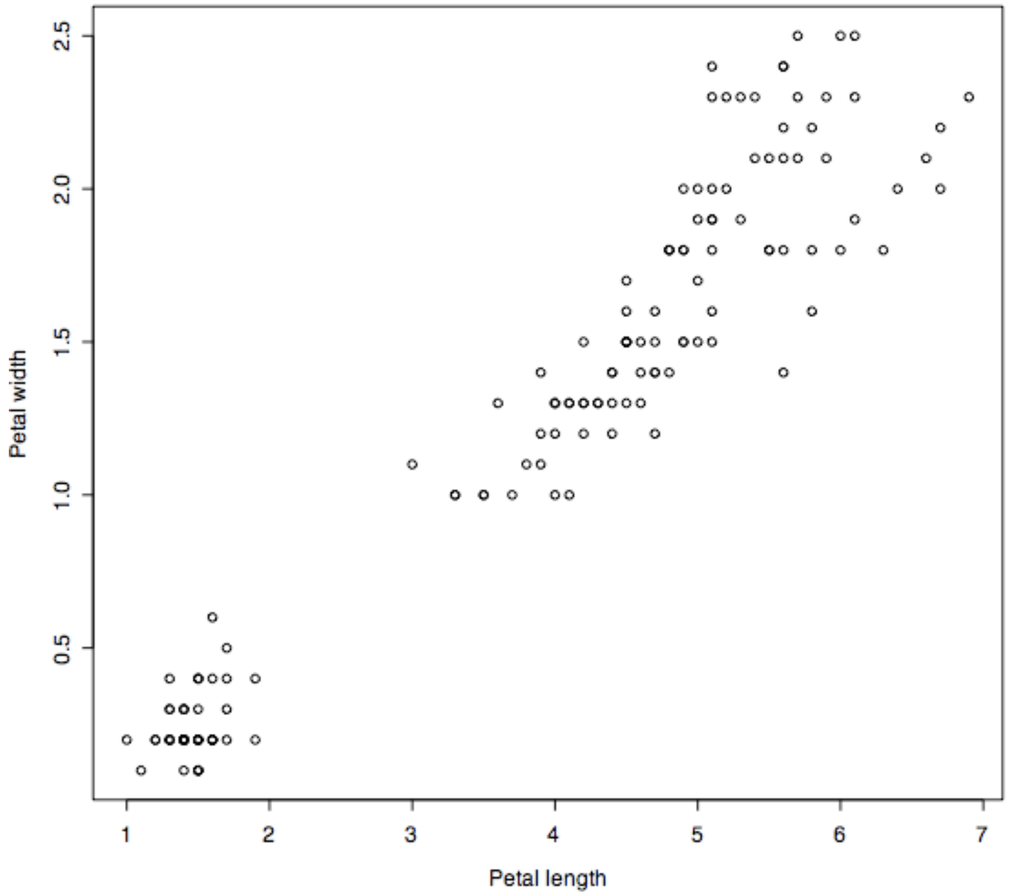
RELATIONSHIP: SCATTER PLOT (2D)

- Most common plot for bivariate data
  - Horizontal X axis: the suspected **independent** variable
  - Vertical Y axis: the suspected **dependent** variable
- Graphically shows:
  - If X and Y are related
  - Linear or non-linear relationship
  - If the variation in Y depends on X
  - Outliers

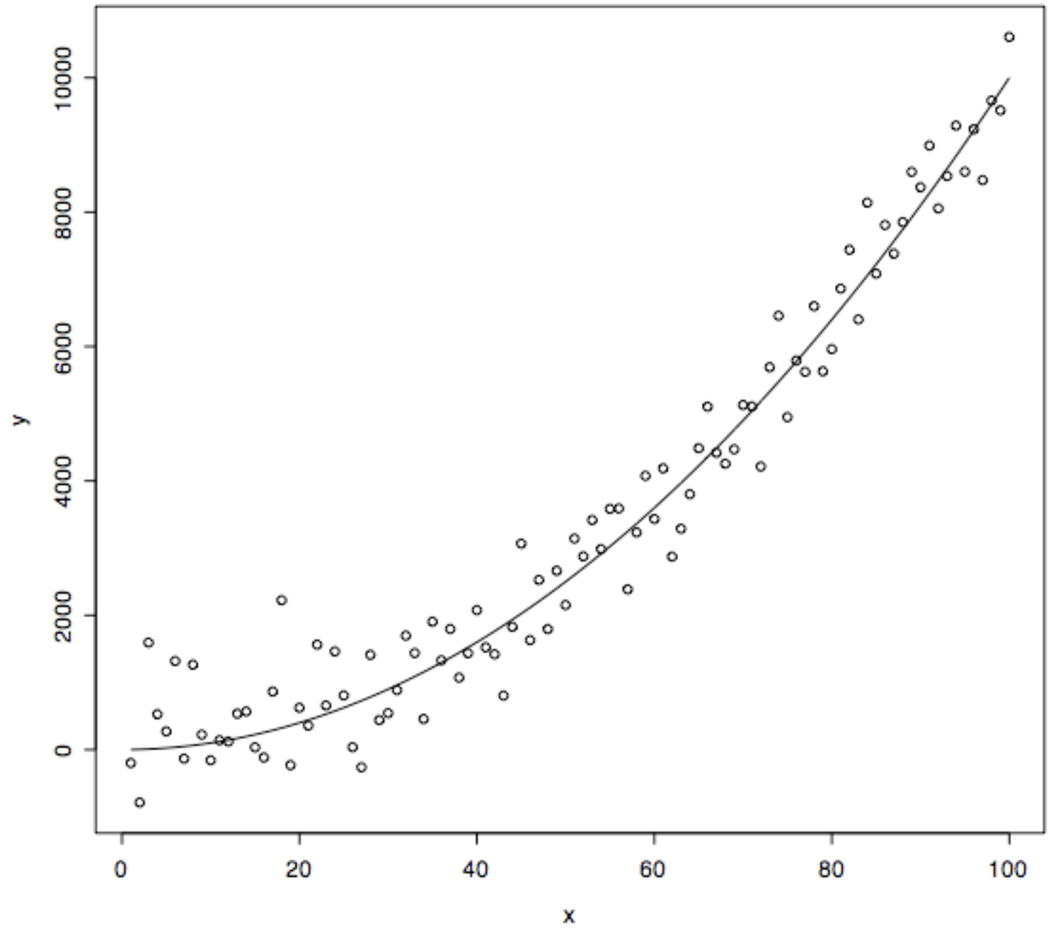
NO RELATIONSHIP



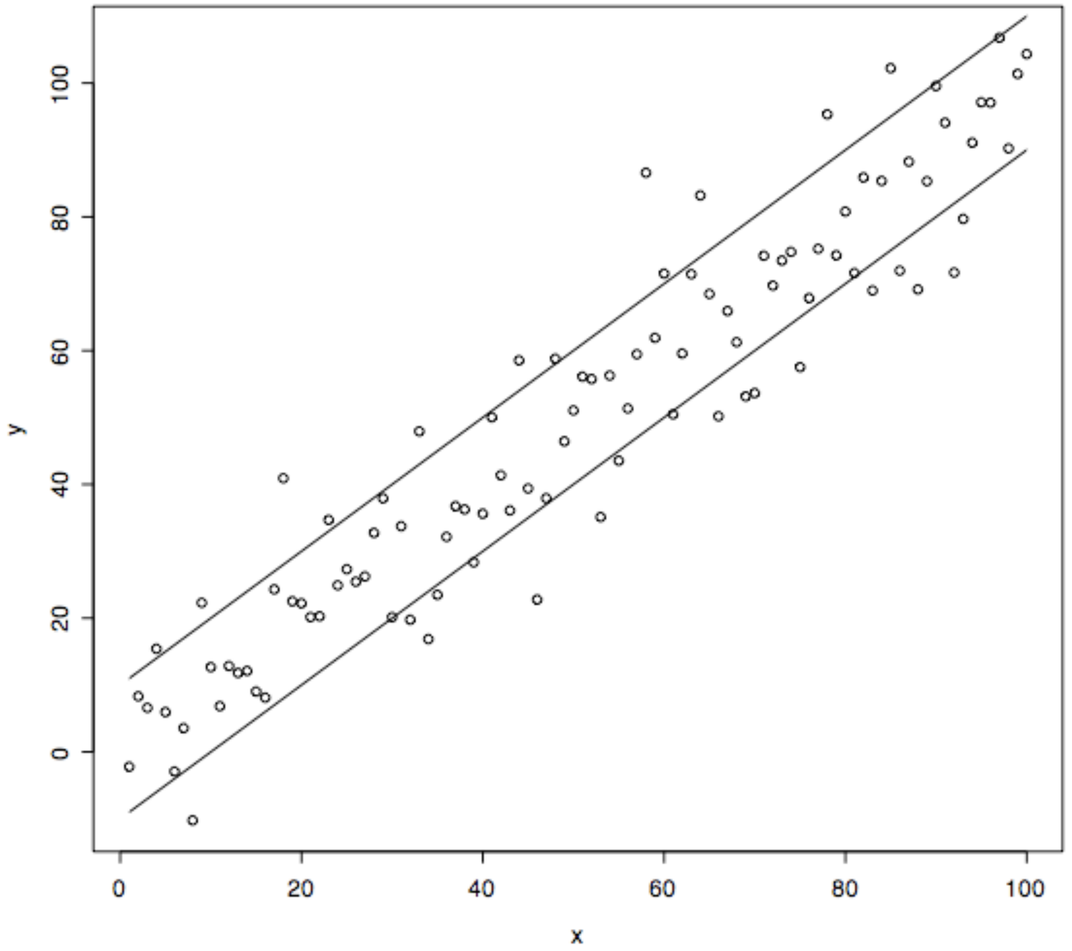
LINEAR RELATIONSHIP



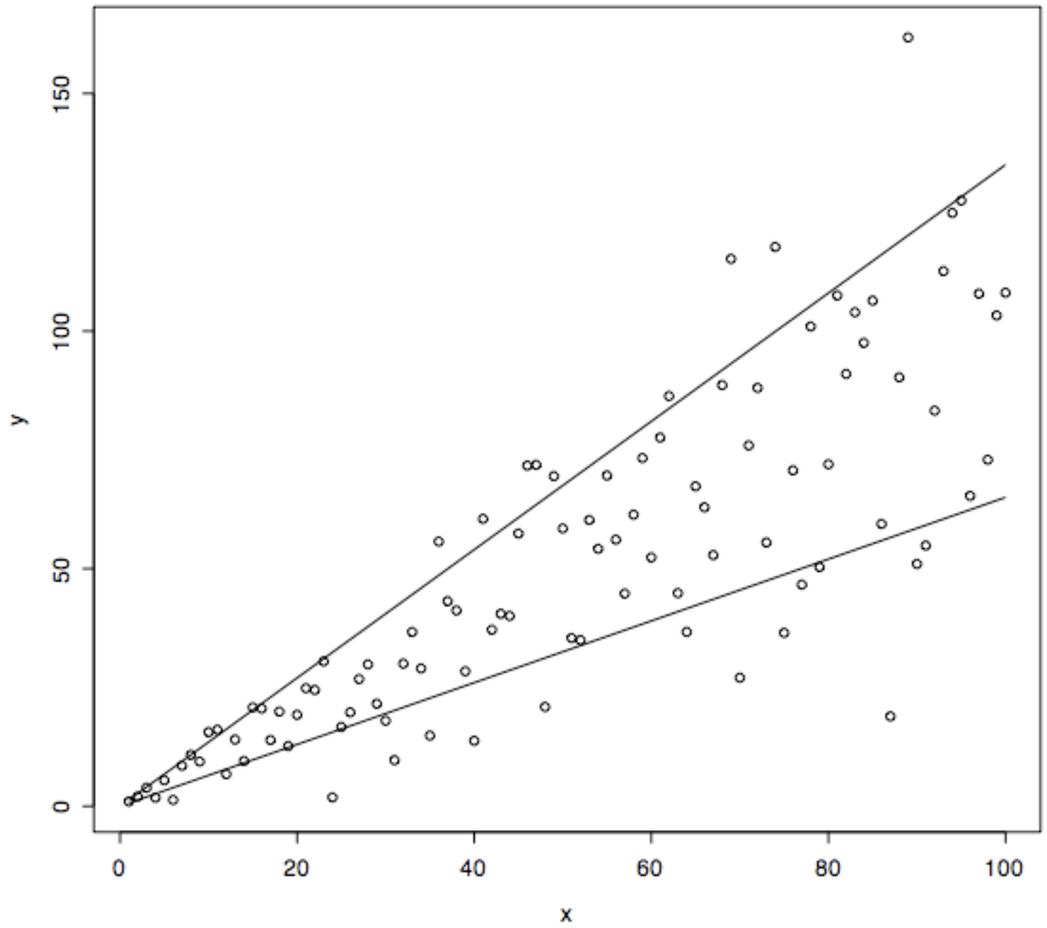
NON-LINEAR RELATIONSHIP



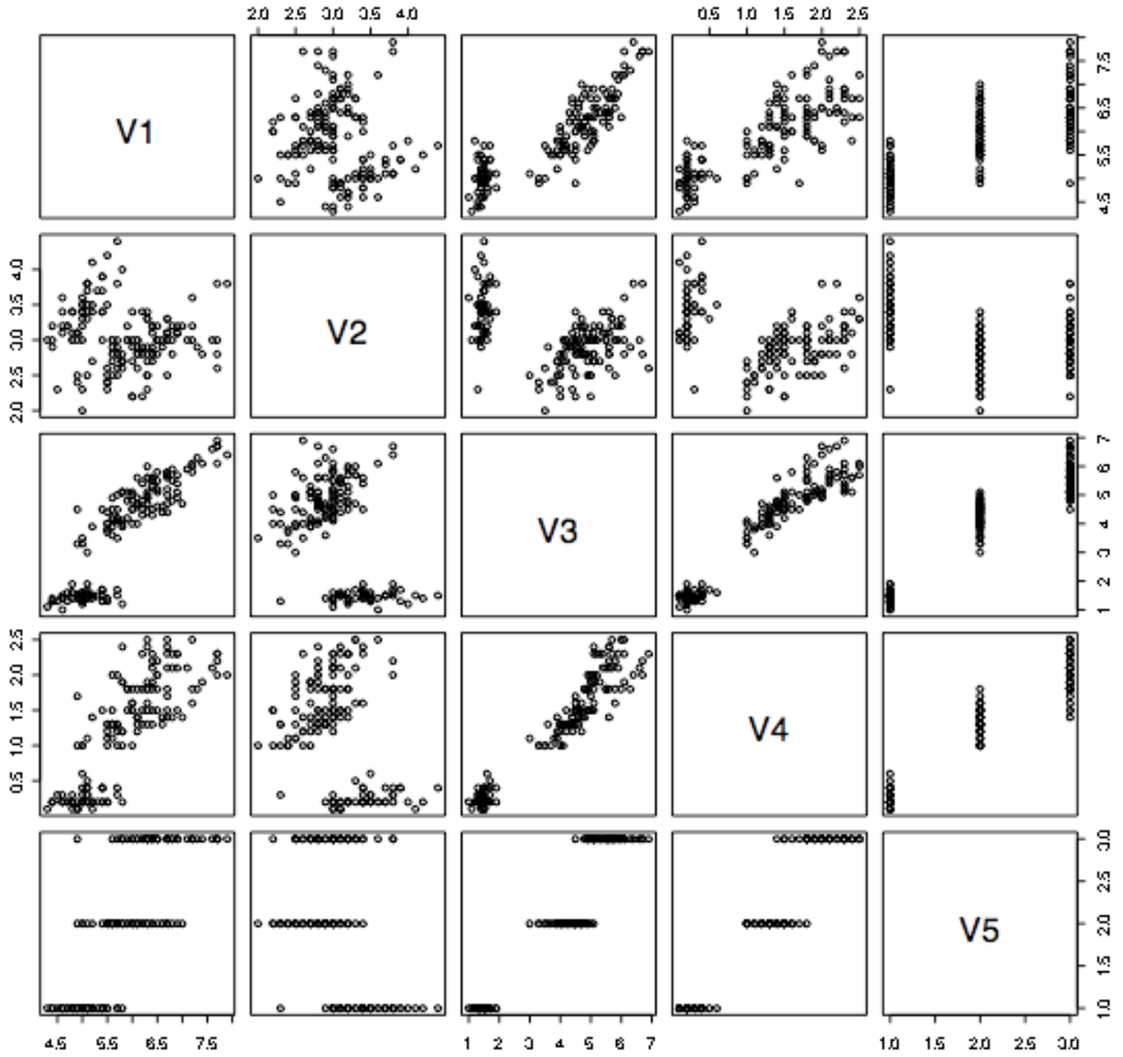
HOMOSKEDASTIC (EQUAL VARIANCE)



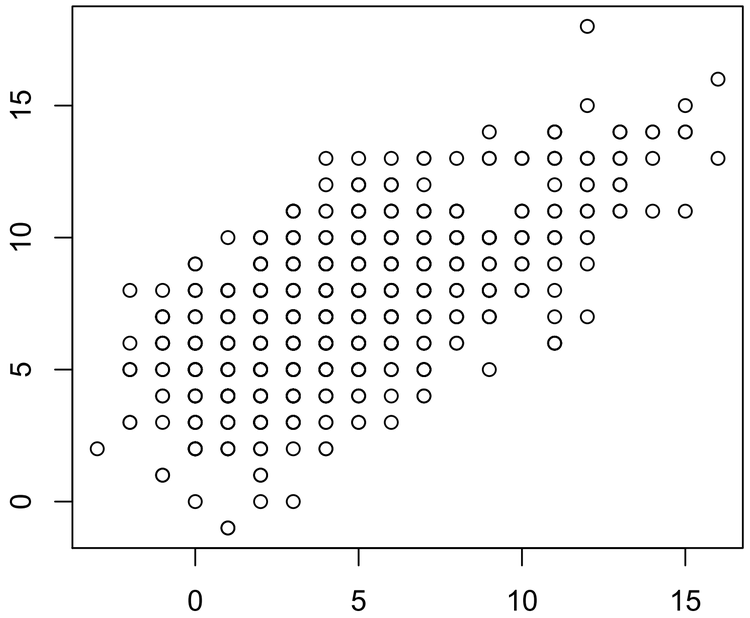
HETEROSKEDASTIC (UNEQUAL VARIANCE)



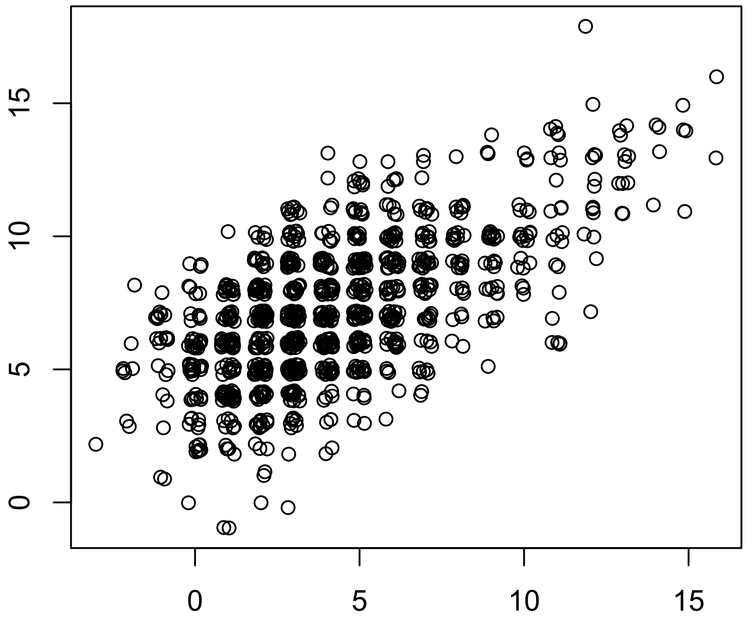
SCATTERPLOT MATRIX



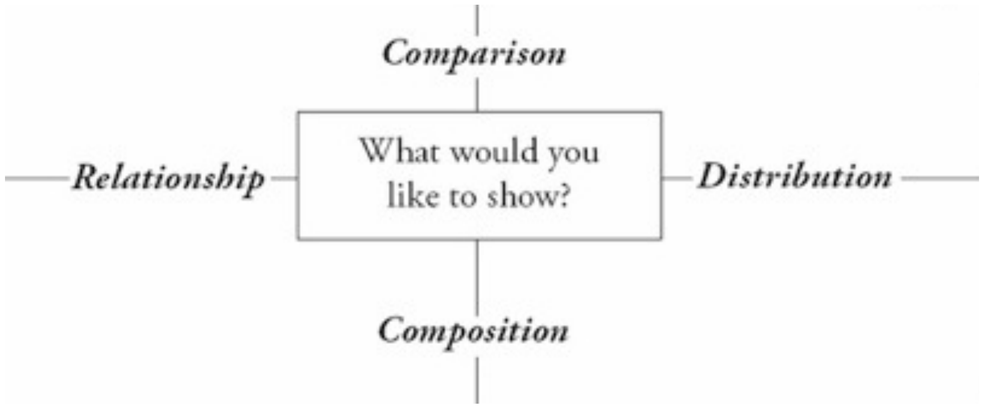
SCATTERPLOT LIMITATIONS



Overprinting

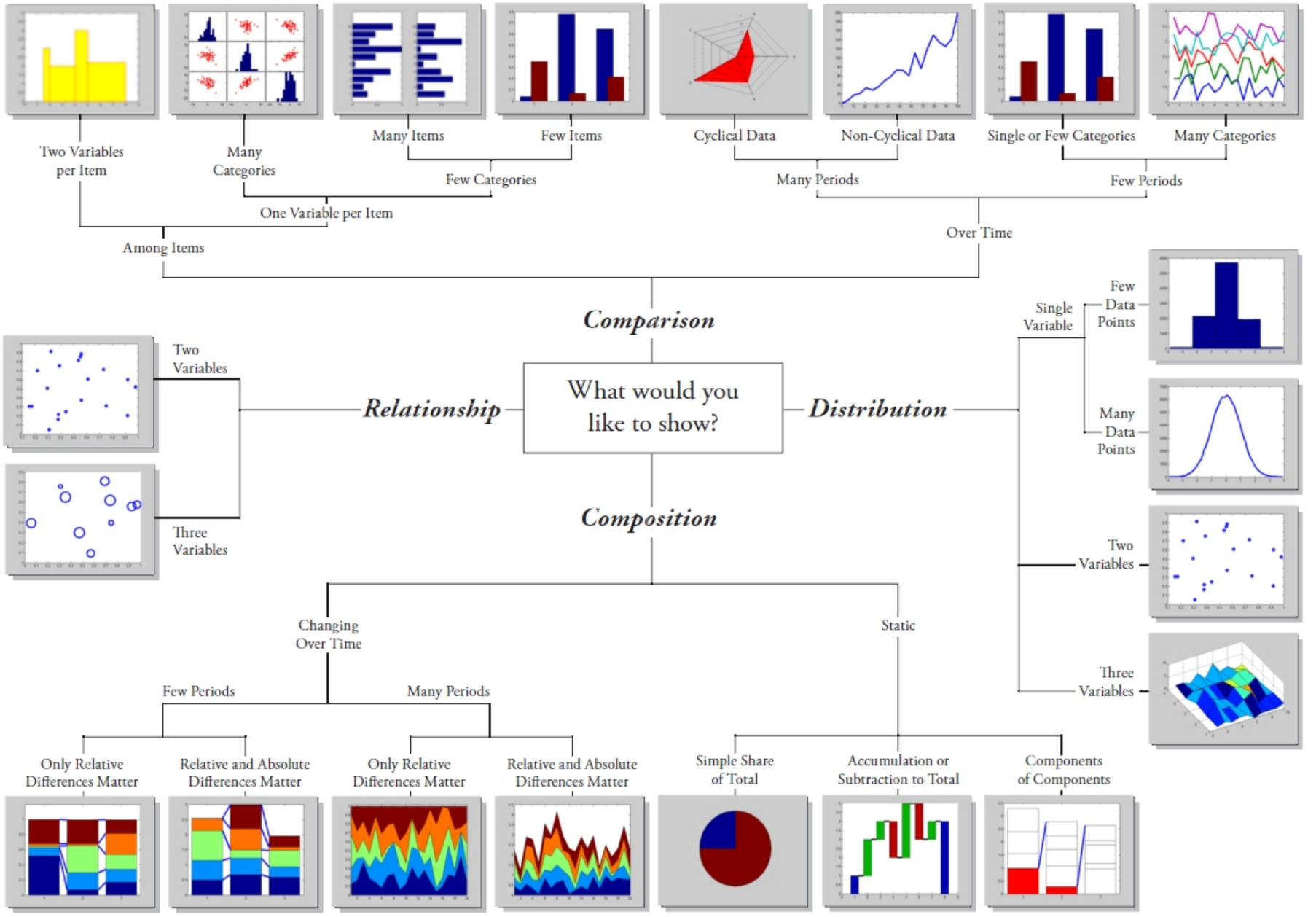


Solution: Jitter points



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

DIMENSIONALITY REDUCTION

DIMENSIONALITY REDUCTION

- Identify and describe the “dimensions” that underlie the data
  - May be more fundamental than those directly measured but hidden to the user
- Reduce dimensionality of modeling problem
  - Benefit is simplification, it reduces the number of variables you have to deal with in modeling
- Can identify set of variables with similar behavior
- Principal component analysis (PCA)

WHAT DIMENSION CAN BE DROPPED?

- Suppose we have a data matrix ***D*** of *n* rows and *p* columns (i.e., we have *n* data points, each data point is measured on *p* dimensions)
- If we want to decrease *p*, which dimensions can we drop?
  - Constant dimensions: 1, 1, …, 1
  - Constant dimensions with some noise: 1.001, 0.998, …, 1.003
  - Dimensions that is linearly dependent on other dimensions: Z=aX+bY

HIGH VARIANCE!

LOW COVARIANCE!

CHANGE OF BASIS

- But the dimension with highest variance may not necessarily be the dimension that we have measured
- Need change of basis such that:
  - The largest amount of variability of the data can be reflected by projecting the data to some basis vector in the new basis
  - After projecting the data to the new basis (or "new dimensions"), the covariances between new dimensions are low