CS57300
PURDUE UNIVERSITY

SEPTEMBER 15, 2021

DATA MINING

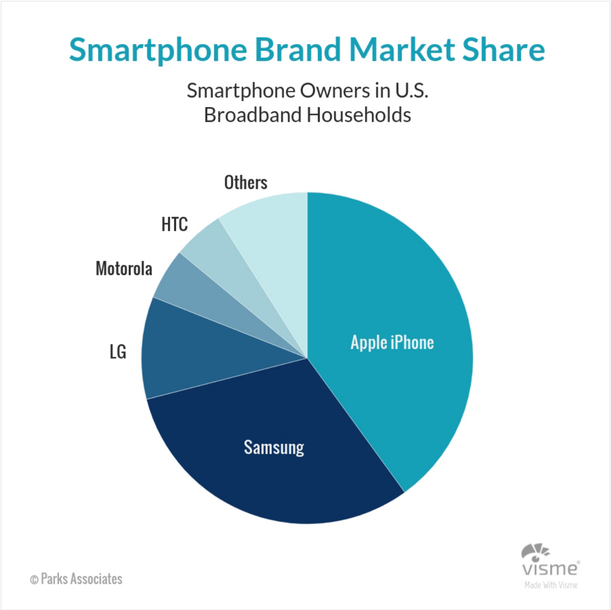
DATA VISUALIZATION

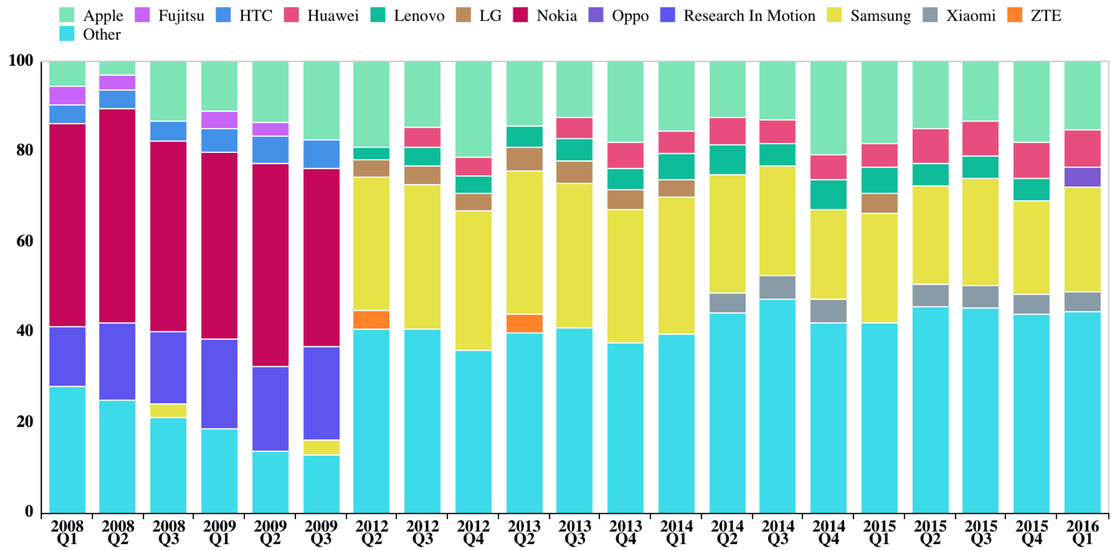
DATA VISUALIZATION

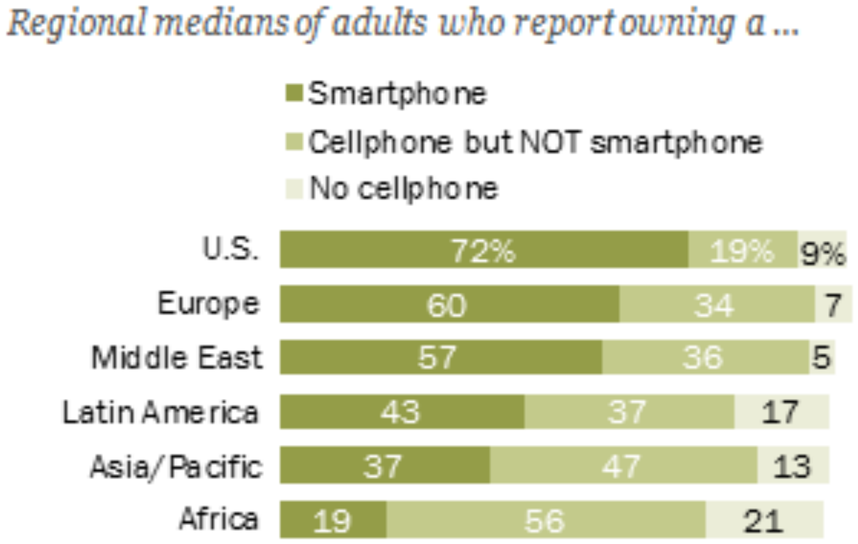
- Serve for different purposes
  - Composition: e.g., see for a discrete dimension x_i, the fraction of each values
  - Distribution: e.g., see the distribution of a continuous dimension of data x_i
  - Comparison:
    - Compare values of two continuous dimensions of the data, x_i and x_j
    - Given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

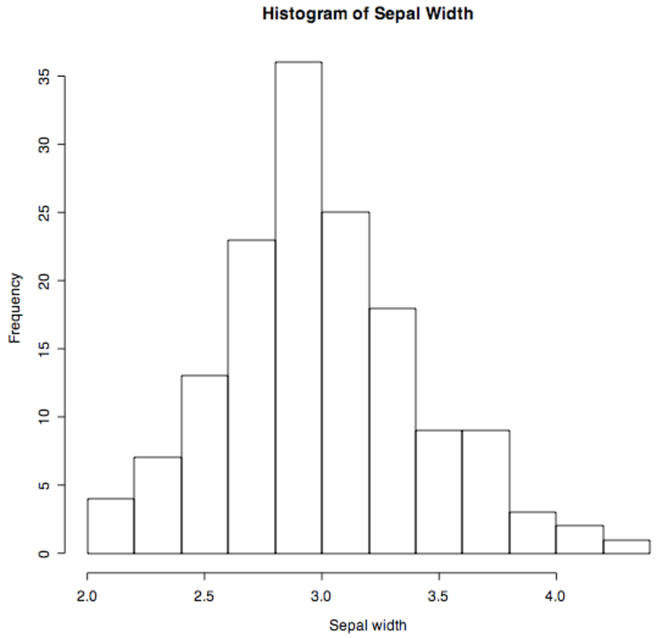






DISTRIBUTION: HISTOGRAMS (1D)

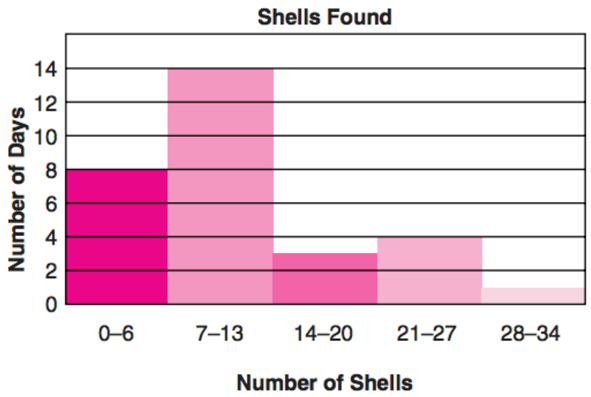
- Most common plot for univariate data
- Split data range into equal-sized bins, count number of data points that fall into each bin
- Graphically shows:
  - Center (location)
  - Spread (scale)
  - Skew
  - Outliers
  - Multiple modes



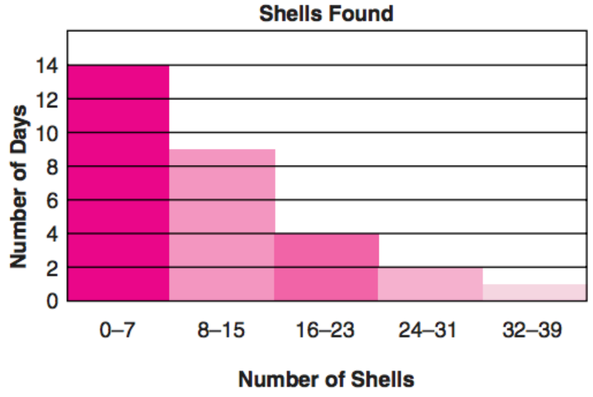
HISTOGRAM LIMITATIONS

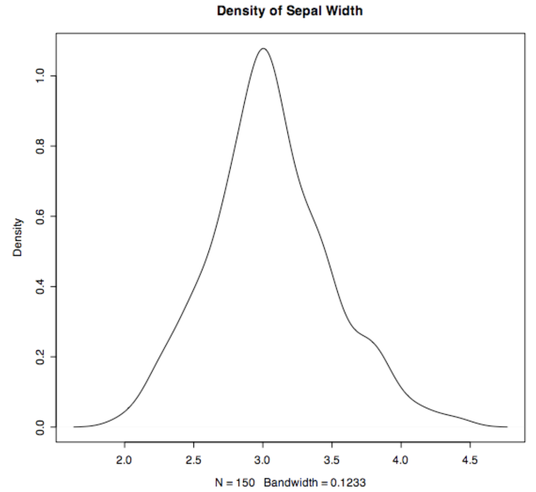
- Histograms can be misleading for small datasets
  - Slight changes in the data or binning approach can result in different histograms
- Solution: smoothed density plots
  - Use kernel function to estimate density at each point x, pools information from neighboring points

1.



2.

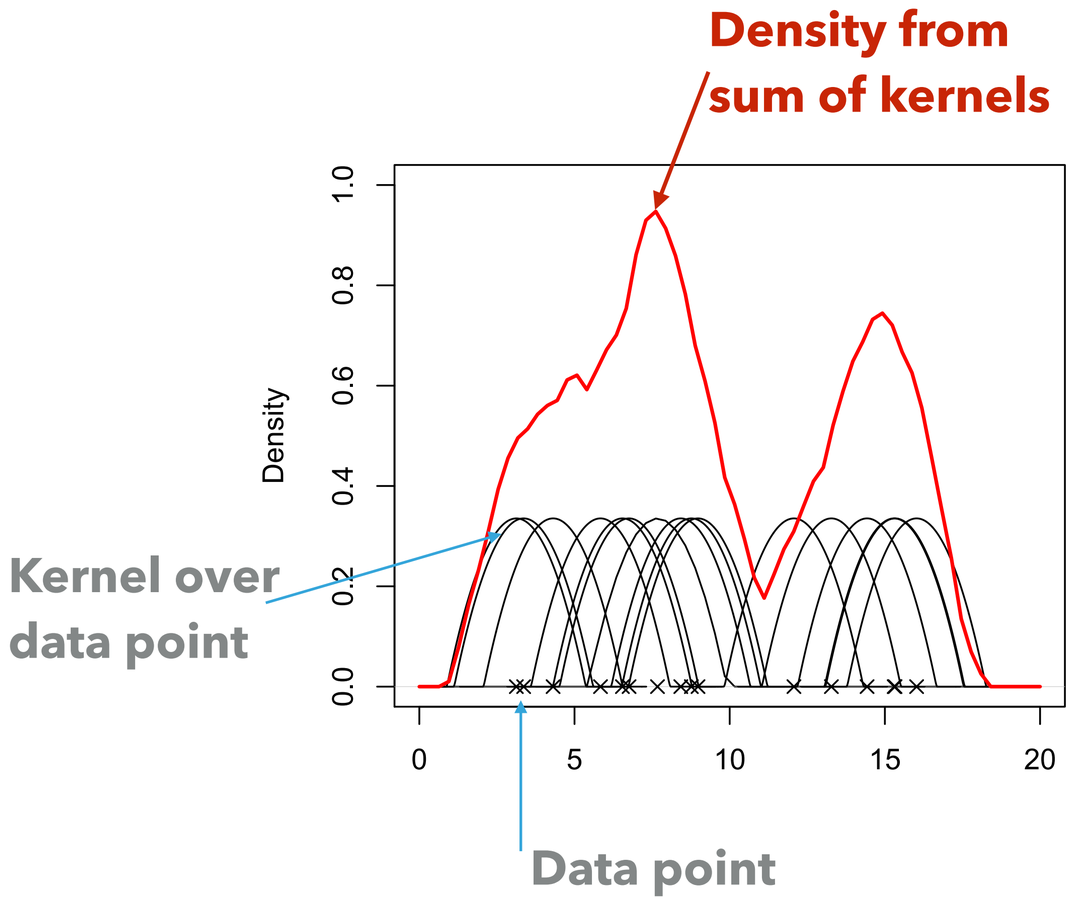




KERNEL DENSITY ESTIMATION

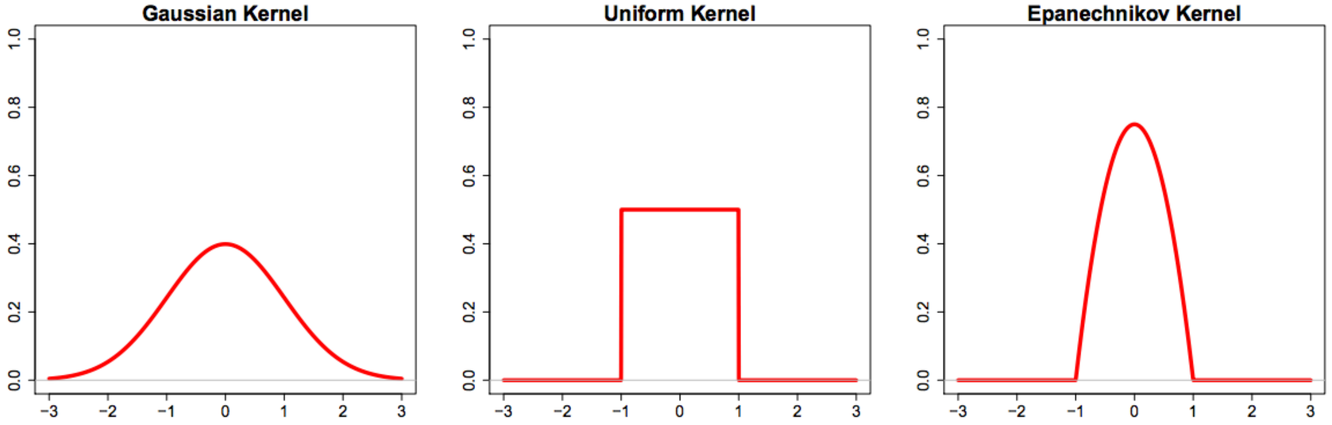
- Estimated density is:

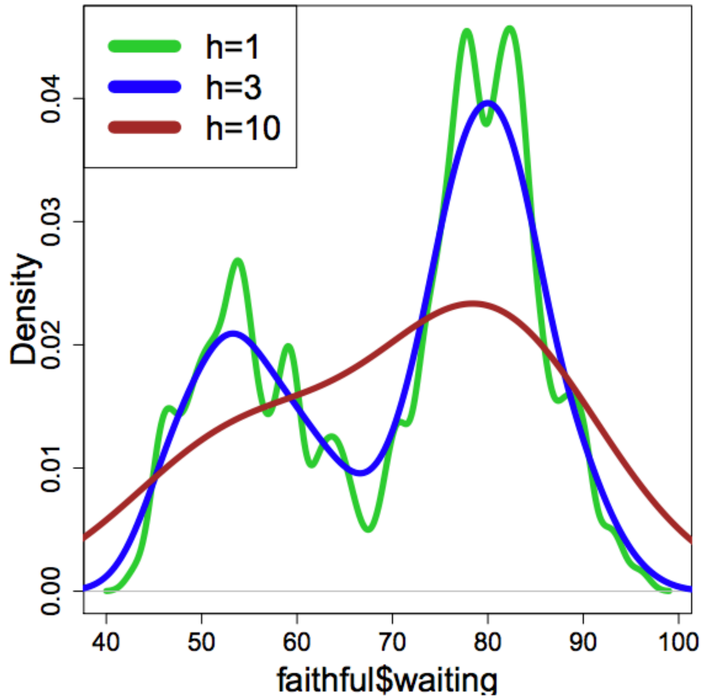
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$



KERNEL DENSITY ESTIMATION

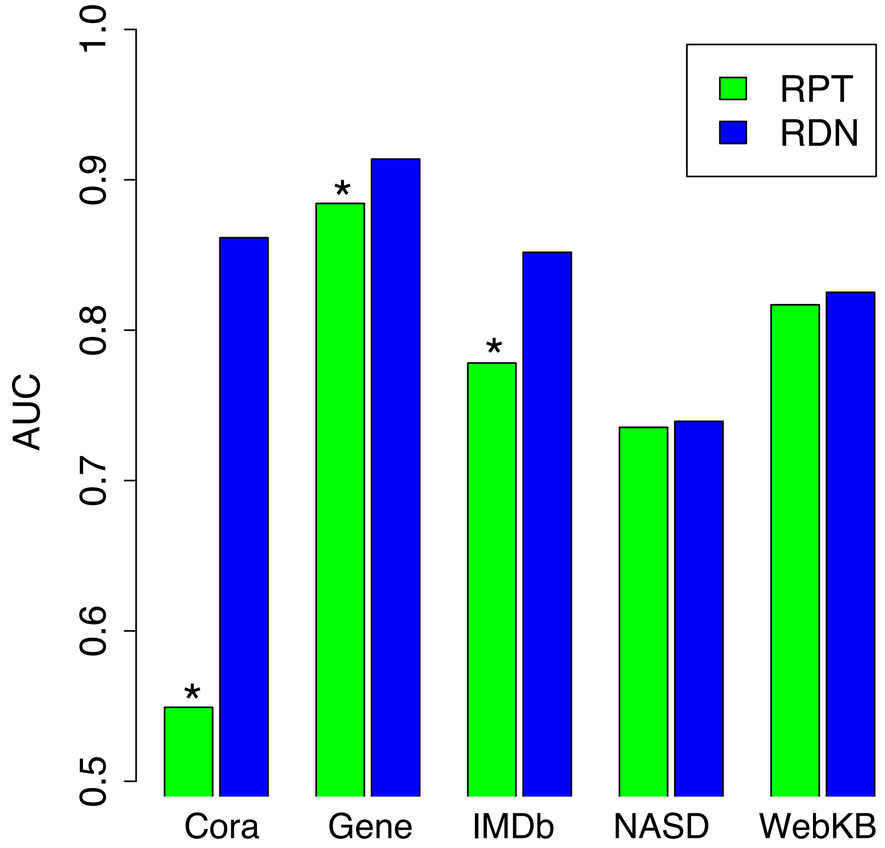
- Two parameters:
  - Kernel function K (e.g., Gaussian, Epanechnikov)
  - Bandwidth h





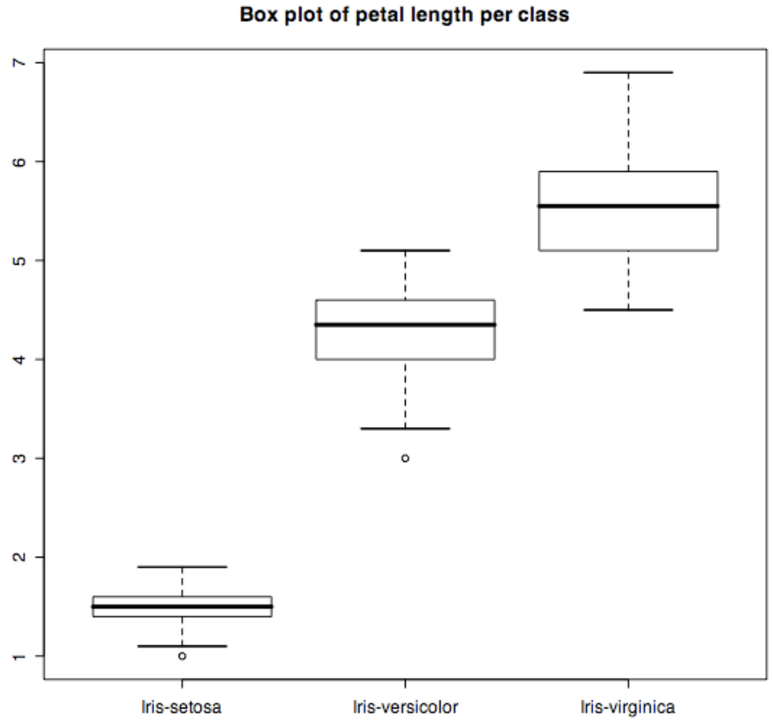
COMPARISON: BAR PLOTS

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



COMPARISON: BOX PLOT (2D)

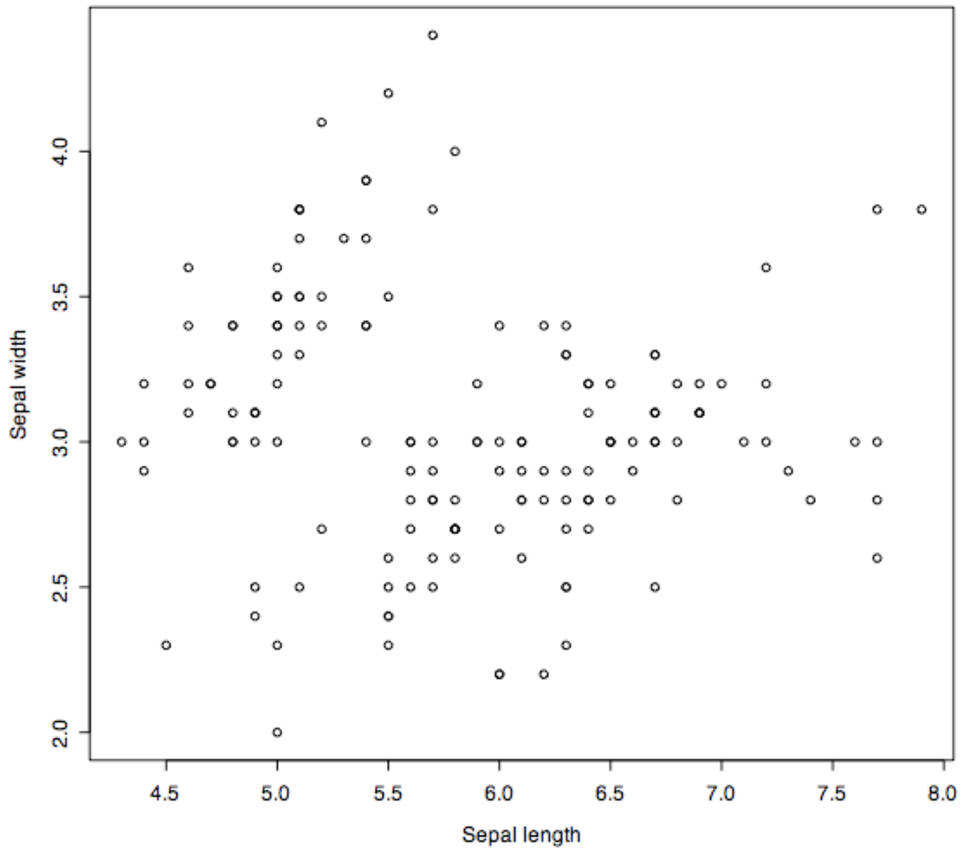
- Display relationship between discrete and continuous variables
- For each discrete value X, calculate quartiles and range of associated Y values



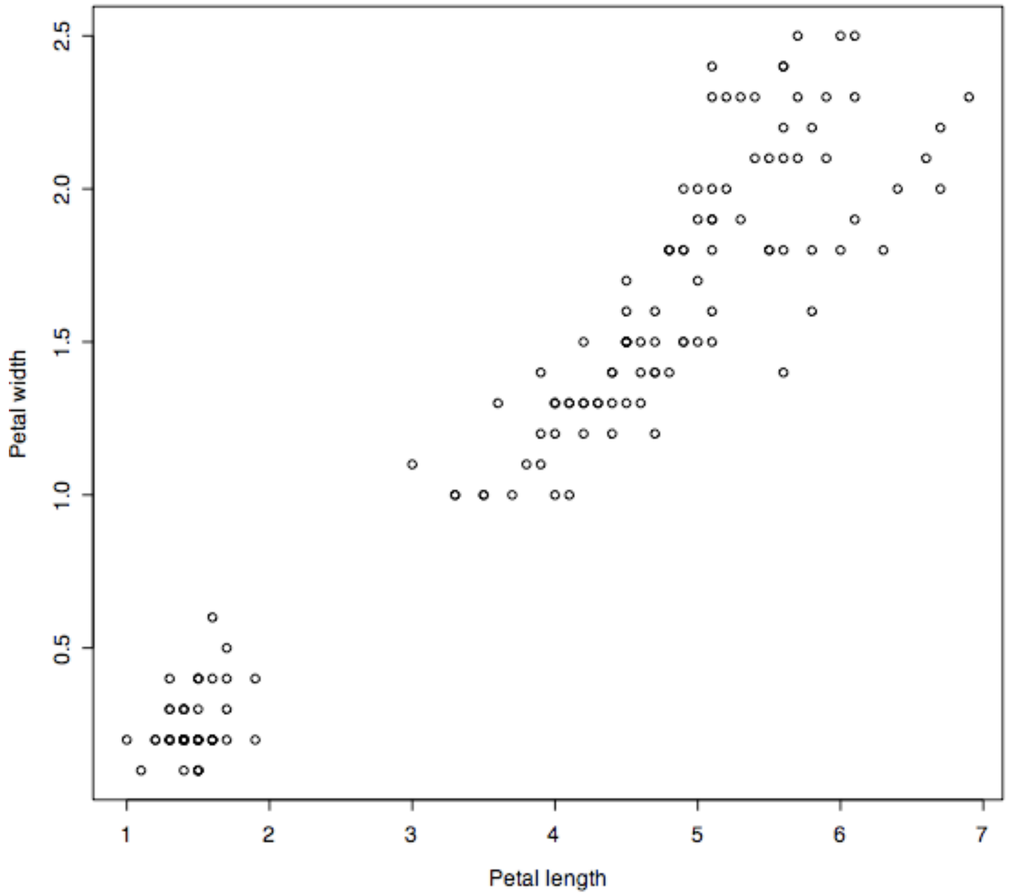
RELATIONSHIP: SCATTER PLOT (2D)

- 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

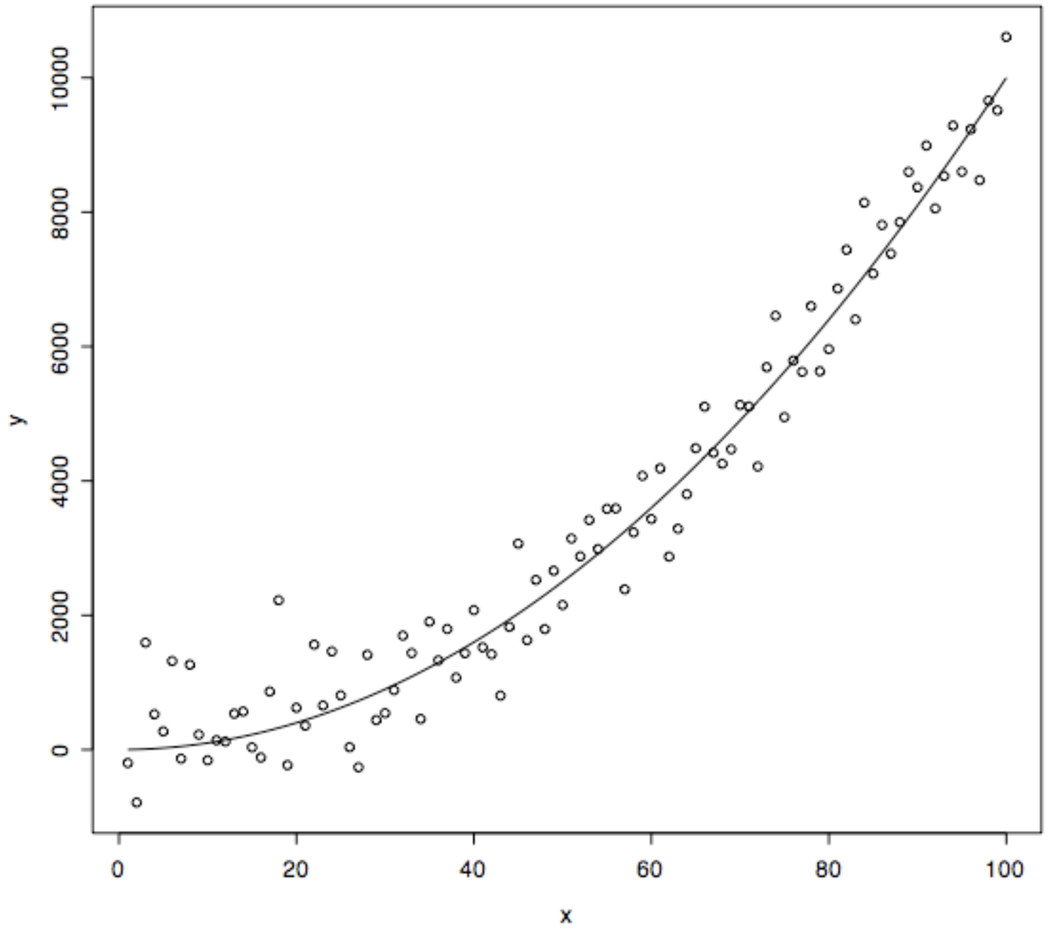
NO RELATIONSHIP



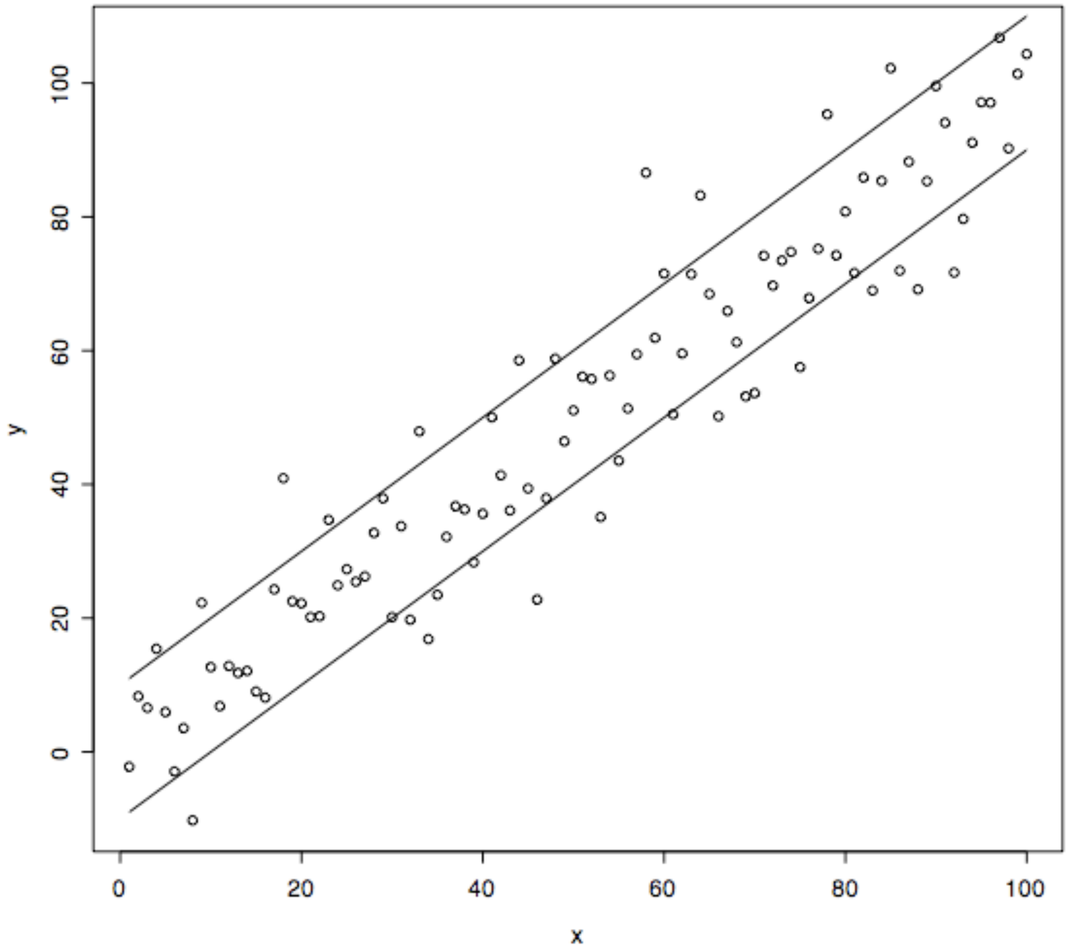
LINEAR RELATIONSHIP



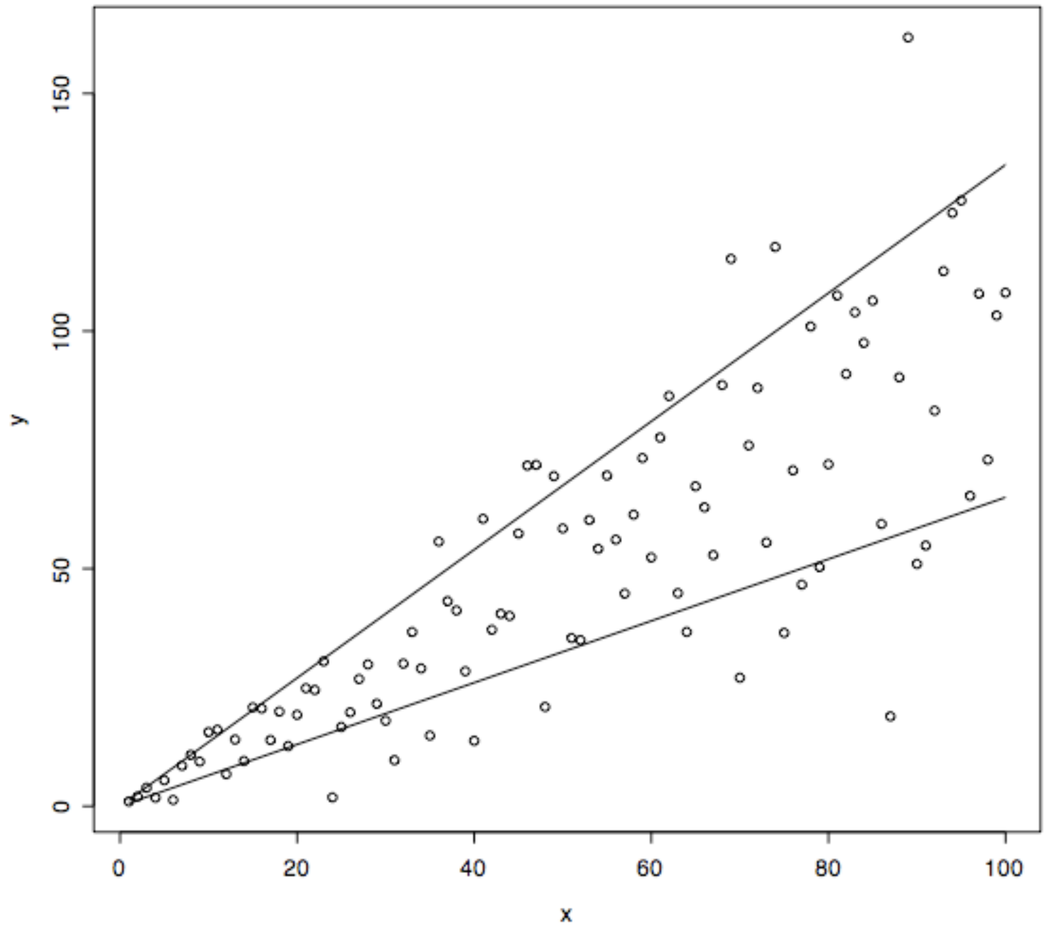
NON-LINEAR RELATIONSHIP



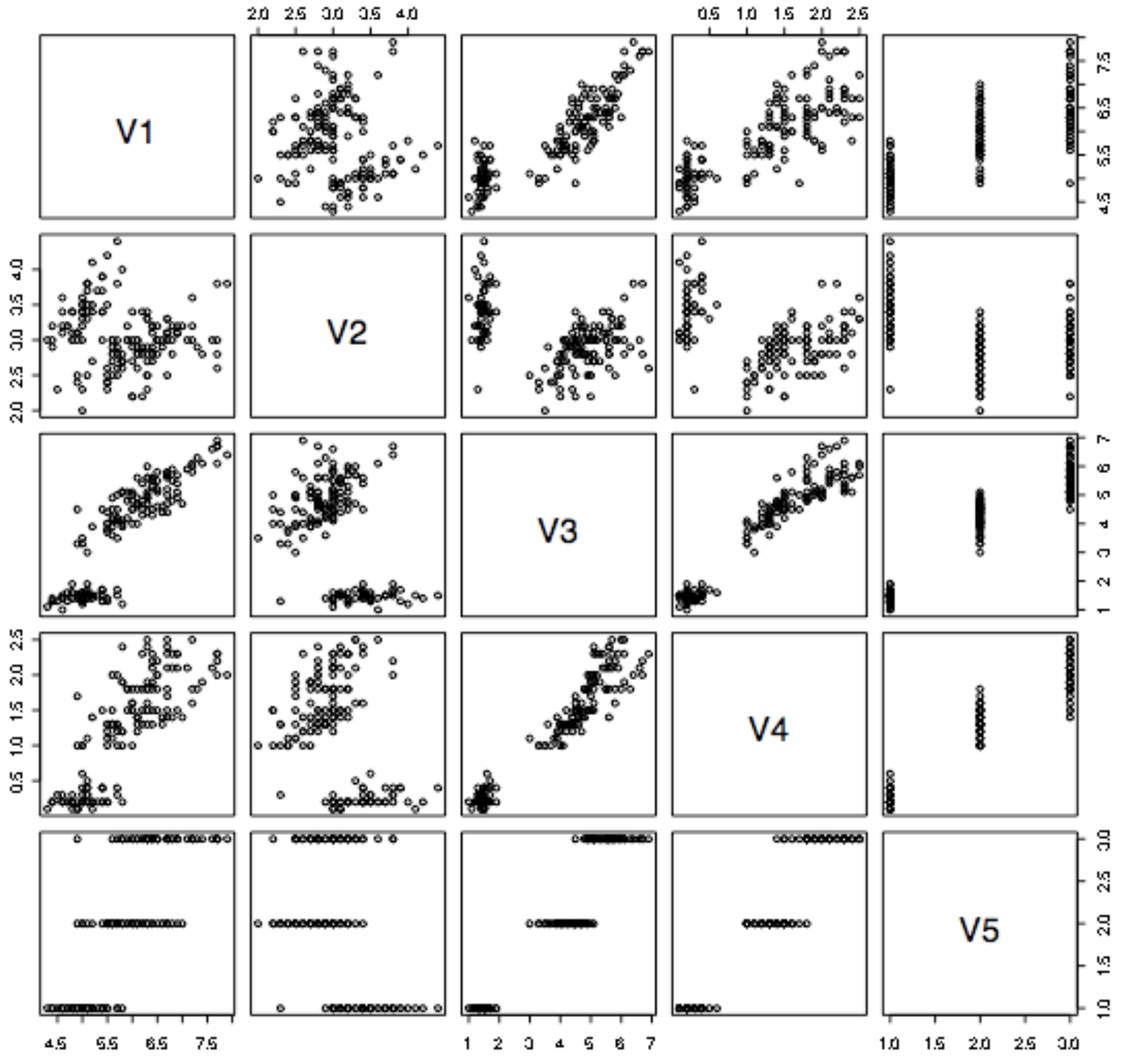
HOMOSKEDASTIC (EQUAL VARIANCE)



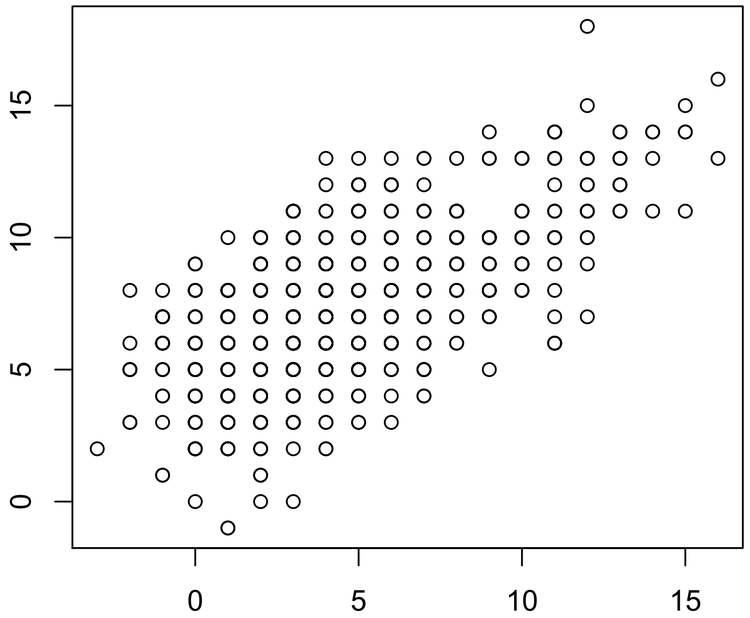
HETEROSKEDASTIC (UNEQUAL VARIANCE)



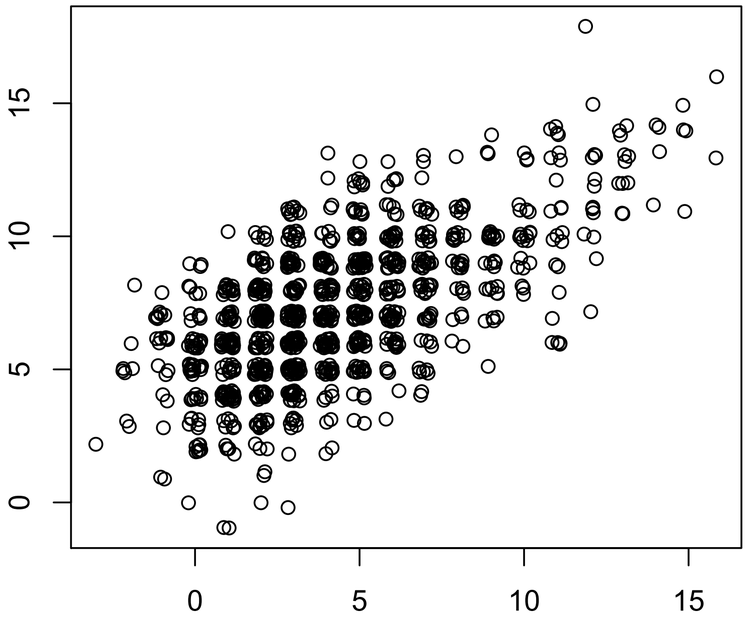
SCATTERPLOT MATRIX



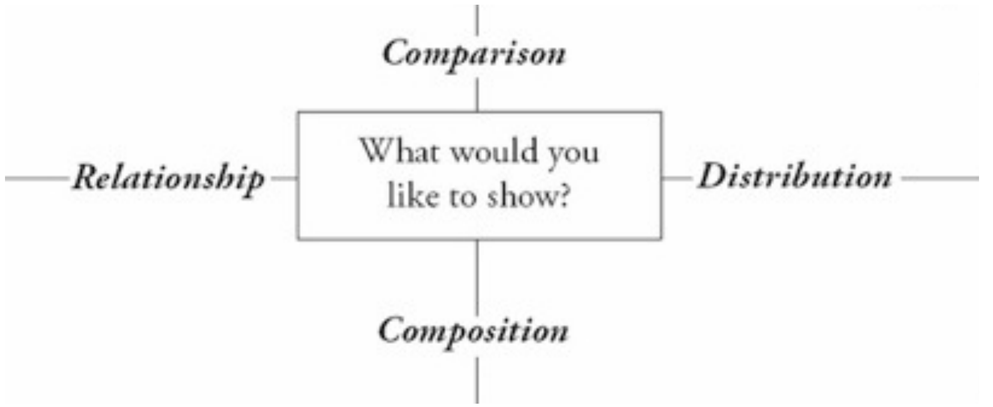
SCATTERPLOT LIMITATIONS



Overprinting

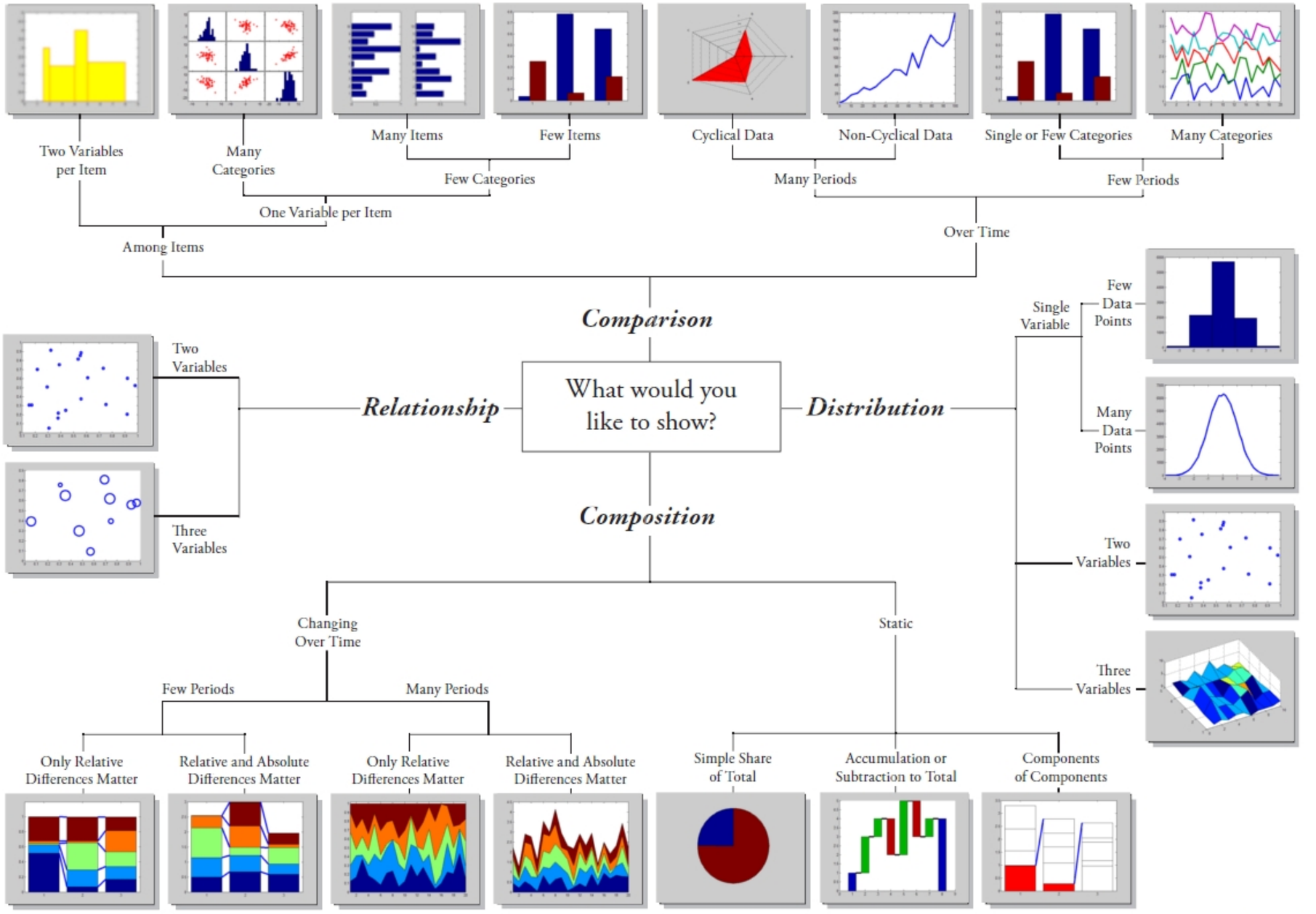


Solution: Jitter points



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?

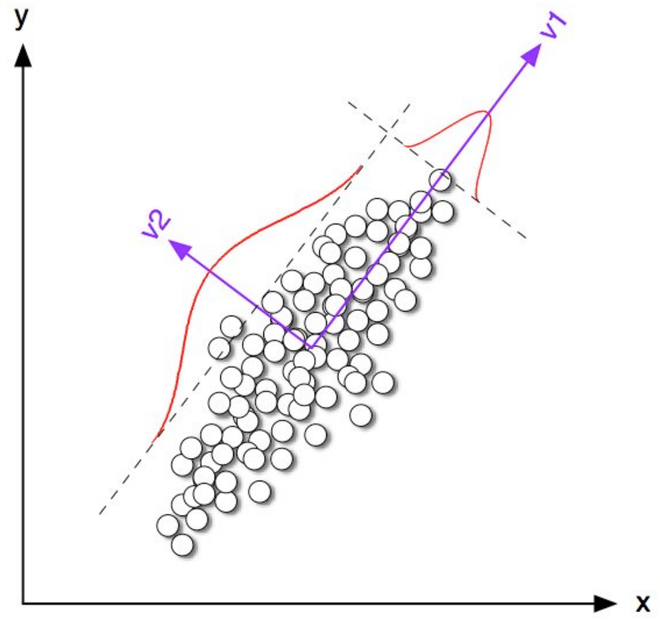
- Suppose we have a data matrix ***D*** of *n* rows and *p* columns (i.e., we have *n* data points, each data point is measured on *p* dimensions)
- If we want to decrease *p*, which dimensions can we drop?
  - Constant dimensions: 1, 1, …, 1
  - Constant dimensions with some noise: 1.001, 0.998, …,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



PRINCIPLE COMPONENT ANALYSIS (PCA)

- Input: the *n***p* data matrix ***D***
- Preprocess **D** so that the mean of each dimension is 0, call this matrix **X**

PRINCIPLE COMPONENT ANALYSIS (PCA)

- Each column vector of **Y=XA** also has a mean of 0.

PRINCIPLE COMPONENT ANALYSIS (PCA)

- Input: the *n*p* data matrix ***D***
- Preprocess **D** so that the mean of each dimension is 0, call this matrix **X**
- Goal: Find a *p*p* orthogonal transformation matrix ***A*** to conduct basis change, i.e., ***Y=XA***, such that under the new basis, the covariances between the new dimensions are low
  - The *p*p* covariance matrix $\boldsymbol{Y^TY}$ is a diagonal matrix.

PRINCIPLE COMPONENT ANALYSIS (PCA)

$$Y^TY = (XA)^T(XA)$$
$$= A^TX^TXA$$

- Notice $\Sigma = X^TX$ is the covariance matrix under the current basis, it is a symmetric square matrix!
- So, we can conduct eigendecomposition for Σ

$$Y^TY = A^T(Q\Lambda Q^T)A = (A^TQ)\Lambda(A^TQ)^T$$

PRINCIPLE COMPONENT ANALYSIS (PCA)

$$Y^TY = A^T(Q\Lambda Q^T)A = (A^TQ)\Lambda(A^TQ)^T$$

- Let $A^TQ = I$, we get $A = (Q^{-1})^T = (Q^T)^T = Q$
  - By doing so, $Y^TY = I\Lambda I = \Lambda$
- In other words, the transformation matrix A is Q, where each column is the eigenvector of the covariance matrix $\Sigma = X^TX$!
- The column vectors of A (or Q) are thus called the **principle component vectors**!

NOT DONE YET…

- So far we have only changed basis, i.e., we project the data to another *p* dimensions
  - The covariance on these *p* new dimensions is 0!
  - But we don't know which dimensions we can drop, i.e., which dimensions have smaller variance yet…
- Recall that $Y^T Y = \Lambda$, this is the covariance matrix after projecting data to the *p* new dimensions!
  - λ_i is the variance of new dimension *i* (i.e., the *i*-th column of A), $\sum_{j=1}^{p} \sigma_j^2 = \sum_{j=1}^{p} \lambda_j$

APPLYING PCA

- Order principal components according to the corresponding eigenvalues. New data vectors are formed by projecting the data onto the first few principal components (i.e., top *m* eigenvectors)

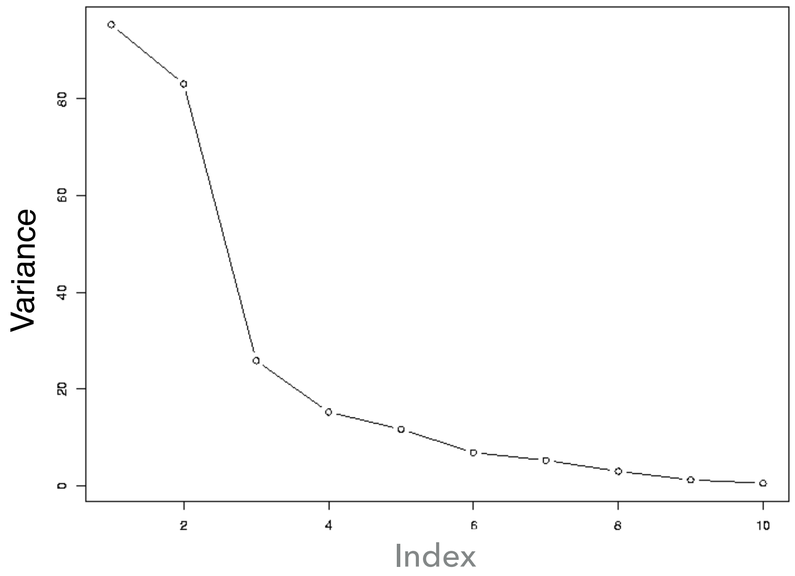
$$\mathbf{x} = [x_1, x_2, \ldots, x_p] \text{ (original instance)}$$
$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p] \text{ (principal components)}$$
$$x'_1 = \mathbf{a}_1\mathbf{x} = \sum_{j=1}^{p} a_{1j}x_j$$
$$\ldots$$
$$x'_m = \mathbf{a}_m\mathbf{x} = \sum_{j=1}^{p} a_{mj}x_j \quad for\ m < p$$
$$\mathbf{x}' = [x'_1, x'_2, \ldots, x'_m] \text{ (transformed instance)}$$

If ***m=p*** then data is transformed

If ***m<p*** then transformation is lossy and dimensionality is reduced

APPLYING PCA (CONT')

- Goal: Find a new (smaller) set of dimensions that captures most of the variability of the data
- Use **scree plot** to choose number of dimensions
  - Choose $m < p$ so projected data captures much of the variance of original data



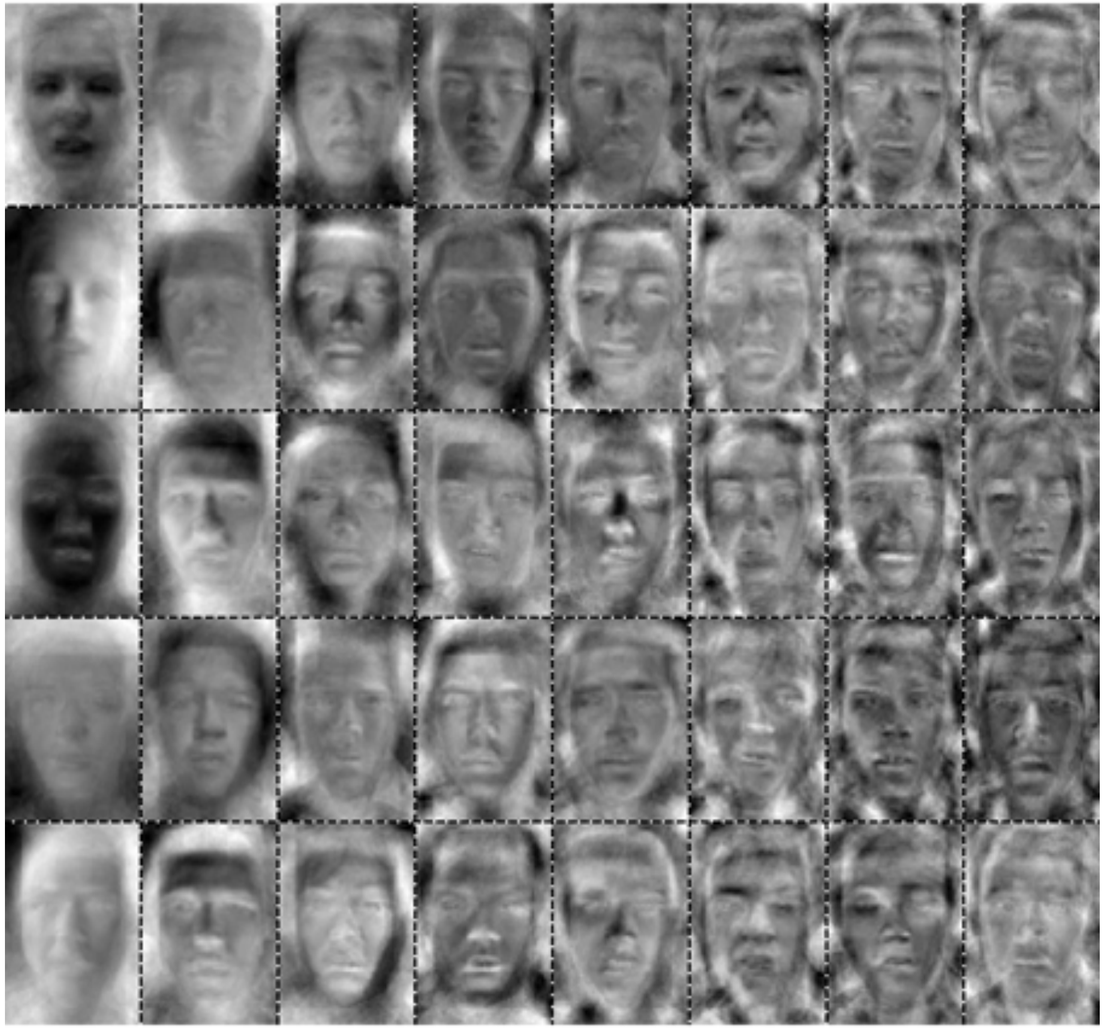
EXAMPLE: EIGENFACES

PCA applied to images of human faces.

Reduce dimensionality to set of basis images.

All other images are linear combo of these “eigenpictures”.

Used for facial recognition.



First 40 PCA dimensions