ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

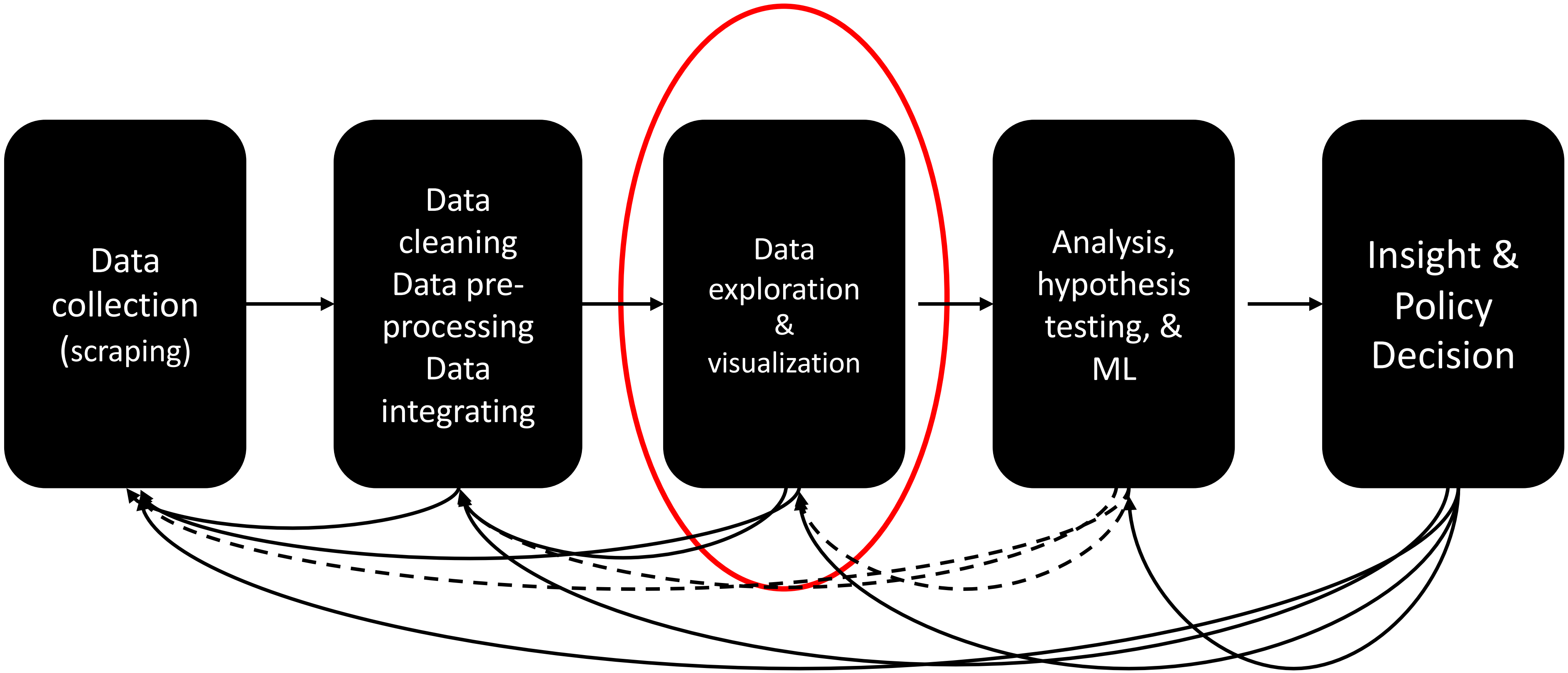
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Lecture 5
Exploratory Data Analysis
(EDA)

Learning outcomes

- Understand key elements in exploratory data analysis (EDA)
- Explain and use common summary statistics for EDA
- Plot and explain common graphs and charts for EDA

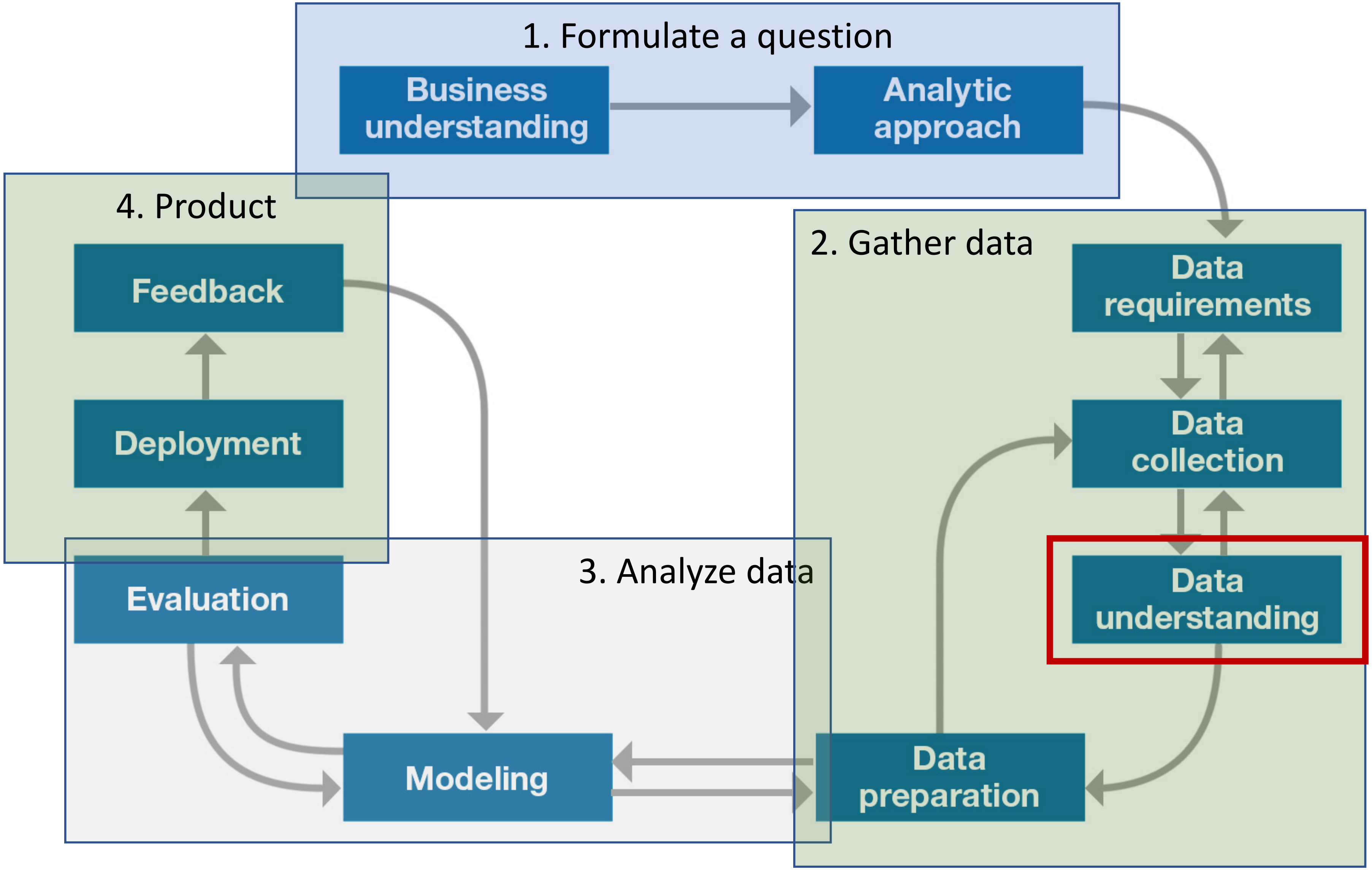
Recall: insight-driven DS methodology



Motivation

- Before making inferences from data it is essential to examine all your variables.
    - To understand your data
- Why?
    - To listen to the data:
    - to catch mistakes
    - to see patterns in the data
    - to find violations of statistical assumptions
    - to generate hypotheses
    - …and because if you don’t, you will have trouble later

Data science process



Source: *Foundational Methodology for Data Science*, IBM, 2015

Exploratory data analysis (EDA) focus

- The focus is on the data—its structure, outliers, and models suggested by the data.
- EDA approach makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.
    - Summary statistics
    - Visualization
    - Clustering and anomaly detection
    - Dimensionality reduction

EDA definition

- The EDA is precisely not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.
  - Helps to select the right tool for preprocessing or analysis
  - Makes use of humans' abilities to recognize patterns in data

EDA common questions

- What is a typical value?
- What is the uncertainty for a typical value?
- What is a good distributional fit for a set of numbers?
- Does an engineering modification have an effect?
- Does a factor have an effect?
- What are the most important factors?
- Are measurements coming from different laboratories equivalent?
- What is the best function for relating a response variable to a set of factor variables?
- What are the best settings for factors?
- Can we separate signal from noise in time dependent data?
- Can we extract any structure from multivariate data?
- Does the data have outliers?

EDA is an iterative process

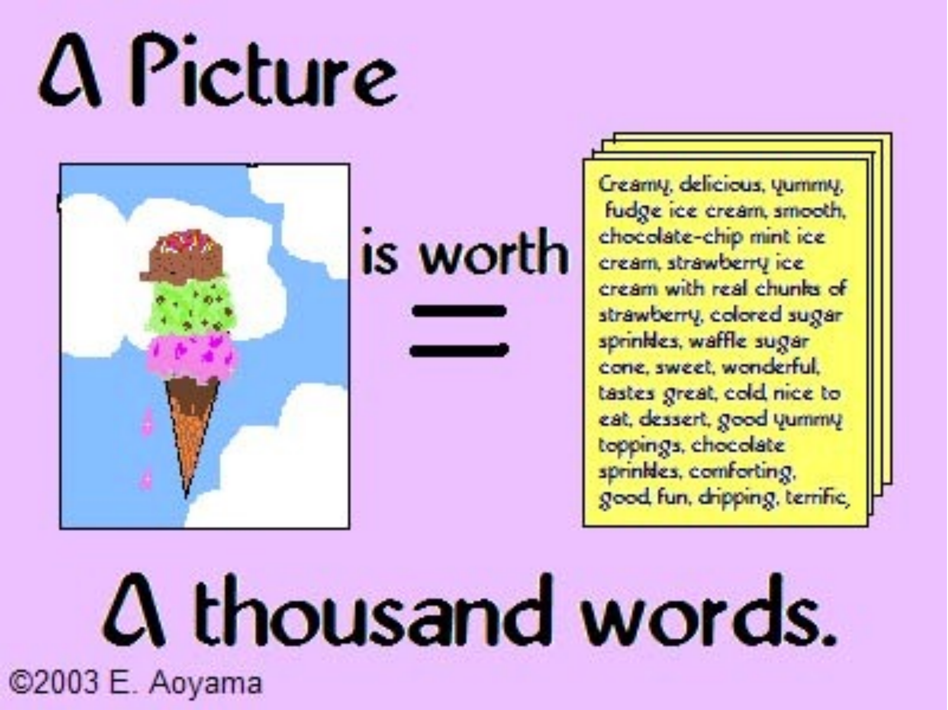
- **Repeat...**
  - Identify and prioritize relevant questions in decreasing order of importance
  - Ask questions
  - Construct graphics to address questions
  - Inspect “answer” and derive new questions

EDA strategy

- Examine variables one by one, then look at the relationships among the different variables
- Start with graphs, then add numerical summaries of specific aspects of the data
- Be aware of attribute types
  - Categorical vs. Numeric

EDA techniques

- Graphical techniques
  - scatter plots, character plots, box plots, histograms, probability plots, residual plots, and mean plots.
- Quantitative techniques



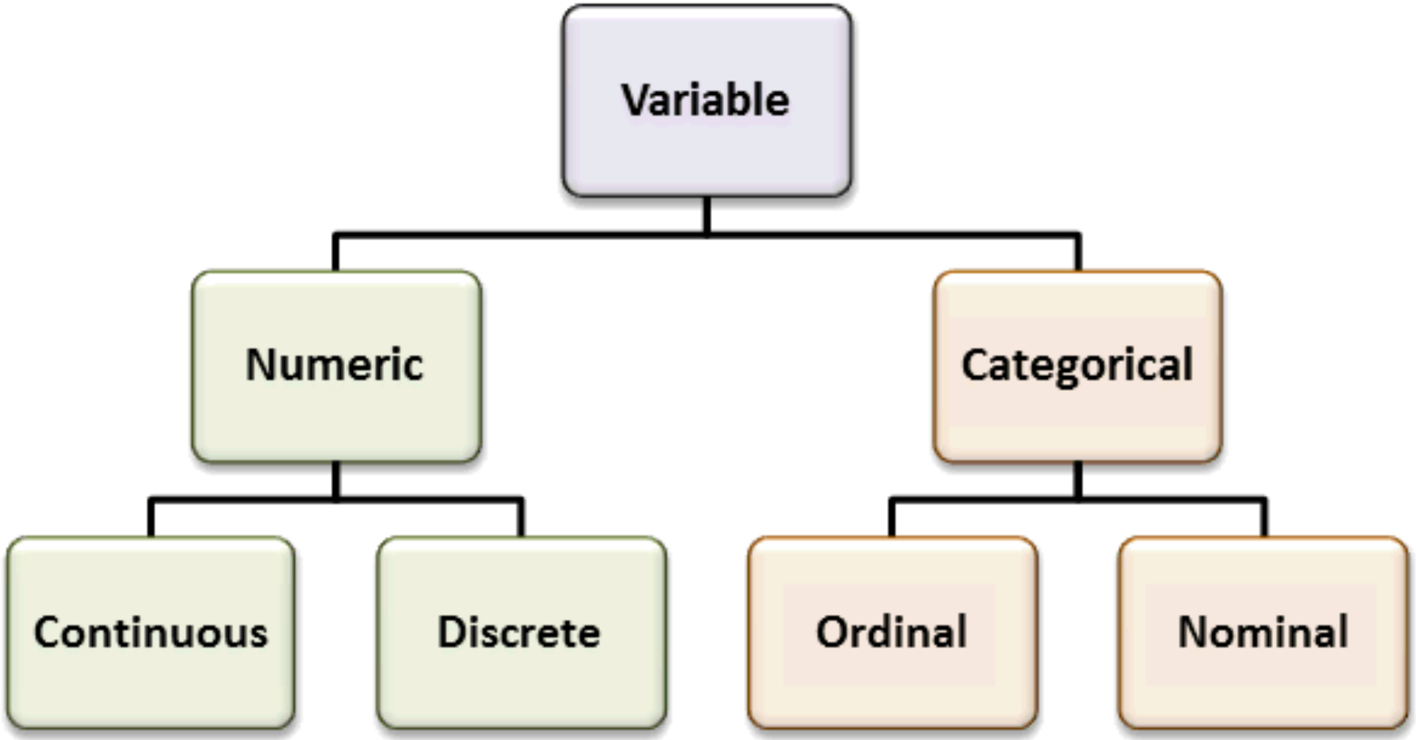
Describing univariate data

Observations and variables

- Data is an collection of observations
- an attribute is thought of as a set of values describing some aspect across all observations, it is called a variable

HR Information		Contact	
Position	**Salary**	**Office**	**Extn.**
Accountant	$162,700	Tokyo	5407
Chief Executive Officer (CEO)	$1,200,000	London	5797
Junior Technical Author	$86,000	San Francisco	1562
Software Engineer	$132,000	London	2558

Types of variables



Dimensionality of data sets

- Univariate: Measurement made on one variable per subject
- Bivariate: Measurement made on two variables per subject
- Multivariate: Measurement made on many variables per subject

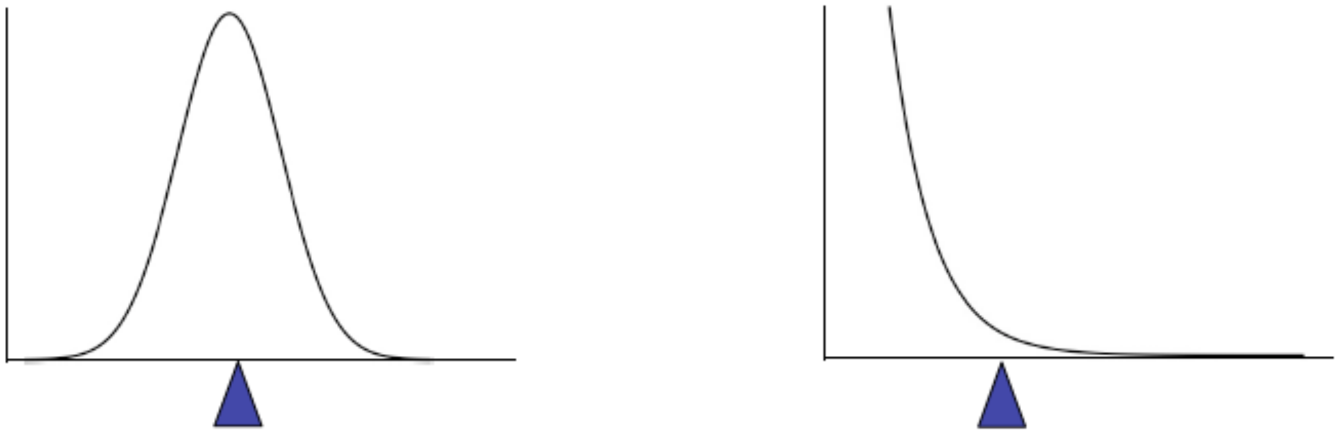
Measures of central tendency

- Measures of Location: estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.
- Measures of Scale: characterize the spread, or variability, of a data set. Measures of scale are simply attempts to estimate this variability.
- Skewness and Kurtosis

Mean

- To calculate the average value of a set of observations, sum of their values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Median

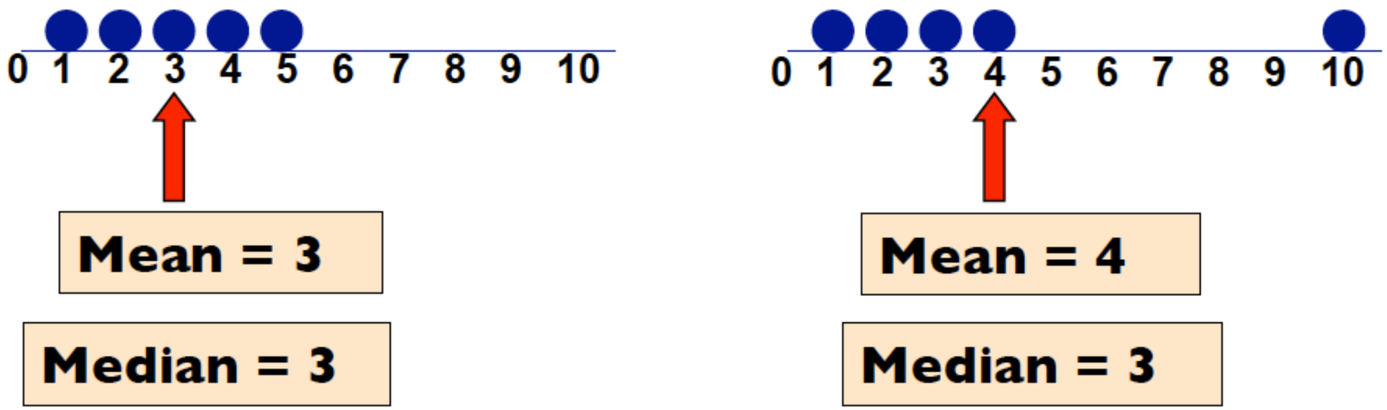
- The median is the value of the point which has half the data smaller than that point and half the data larger than that point.
- Calculation
  - If there are an odd number of observations, find the middle value
  - If there are an even number of observations, find the middle two values and average them
- Example
  - Age of participants: 17 19 21 22 23 23 23 38
  - **Median = (22+23)/2 = 22.5**

Mode

- mode is the most commonly reported value for a particular variable
  - Eg. 3, 4, 5, 6, 7, 7, 7, 8, 8, 9. Mode = 7
  - Eg. 3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9. Mode = {7, 8} = 7.5

Which location measure is best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers



Measure of scale : Variance and standard deviation

- Variance: average of squared deviations of values from the mean

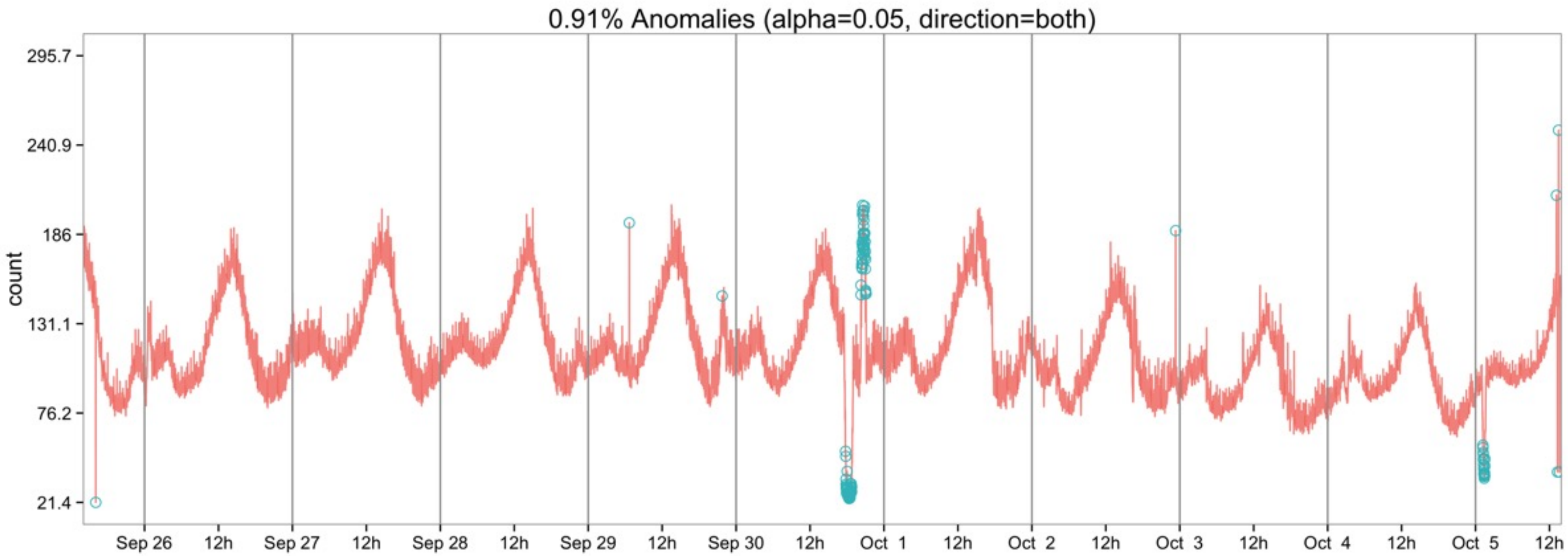
$$\hat{\sigma}^2 = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}$$

- Standard Deviation: simply the square root of the variance

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

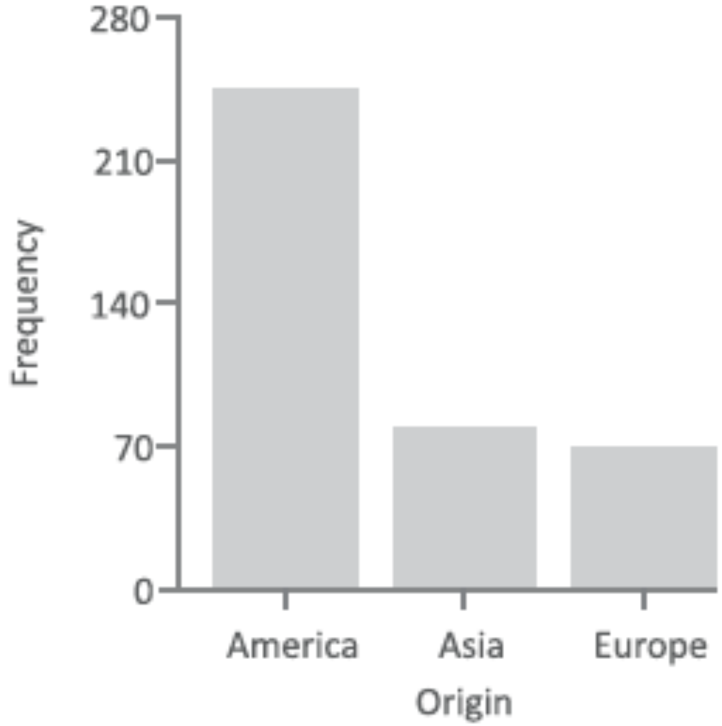
Run sequence plot

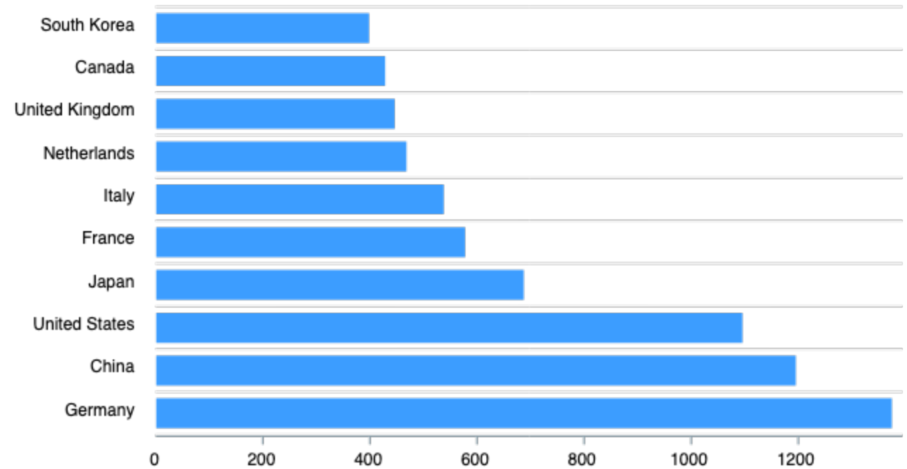
- displays observed data in a time sequence.
- The run sequence plot can be used to answer the following questions
  - Are there any shifts in location?
  - Are there any shifts in variation?
  - Are there any outliers?



Bar charts

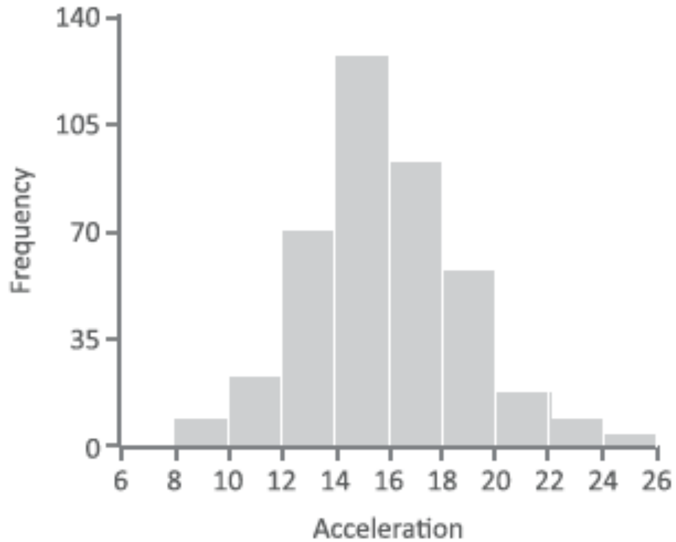
- a bar chart displays the relative frequencies for the different values.
- or a chart presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent



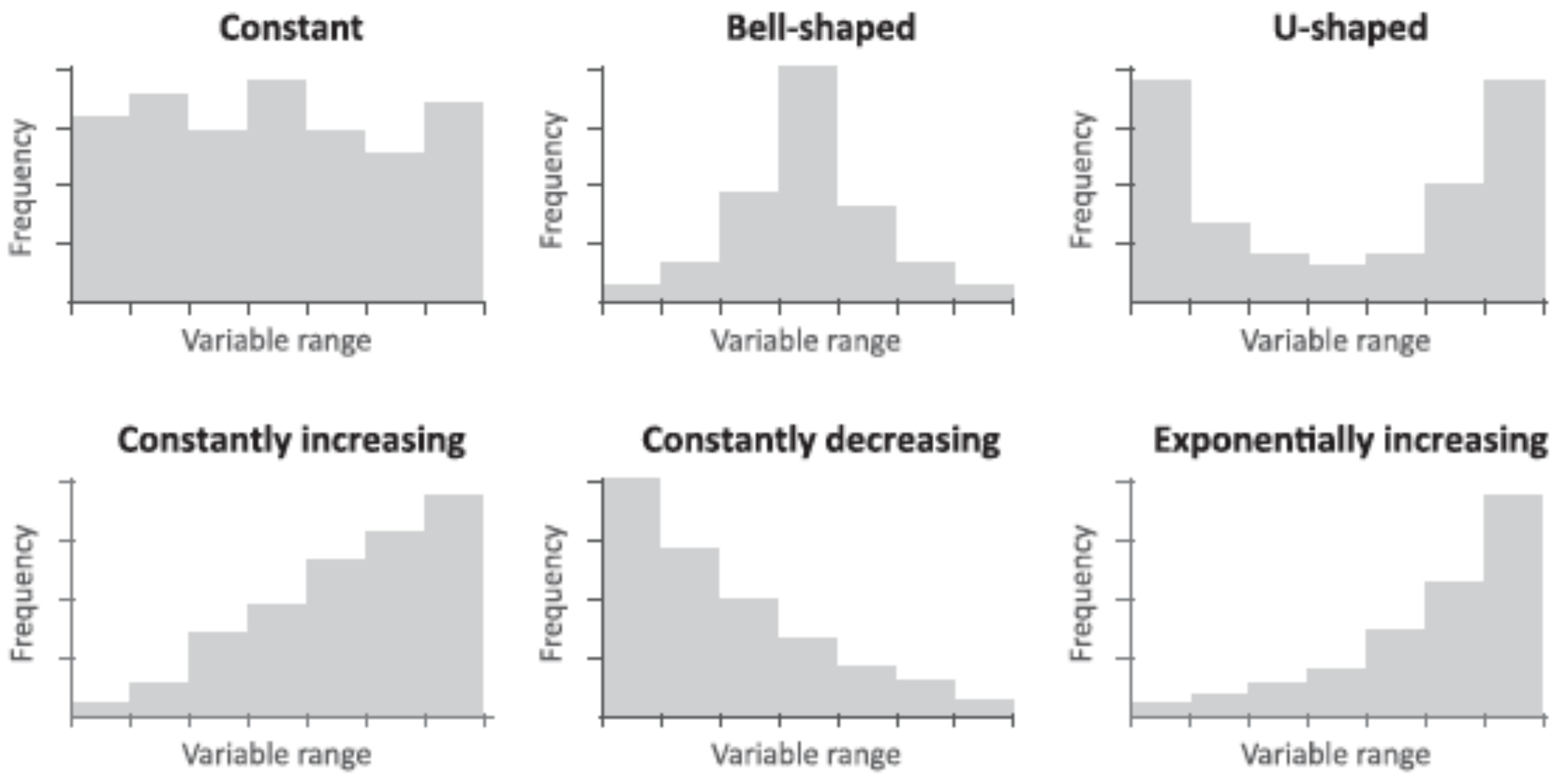


Histogram plot

- A histogram is to graphically summarize the distribution of a univariate data set.
- The histogram can be used to answer the following questions:
  - What kind of population distribution do the data come from?
  - Where are the data located?
  - How spread out are the data?
  - Are the data symmetric or skewed?
  - Are there outliers in the data?

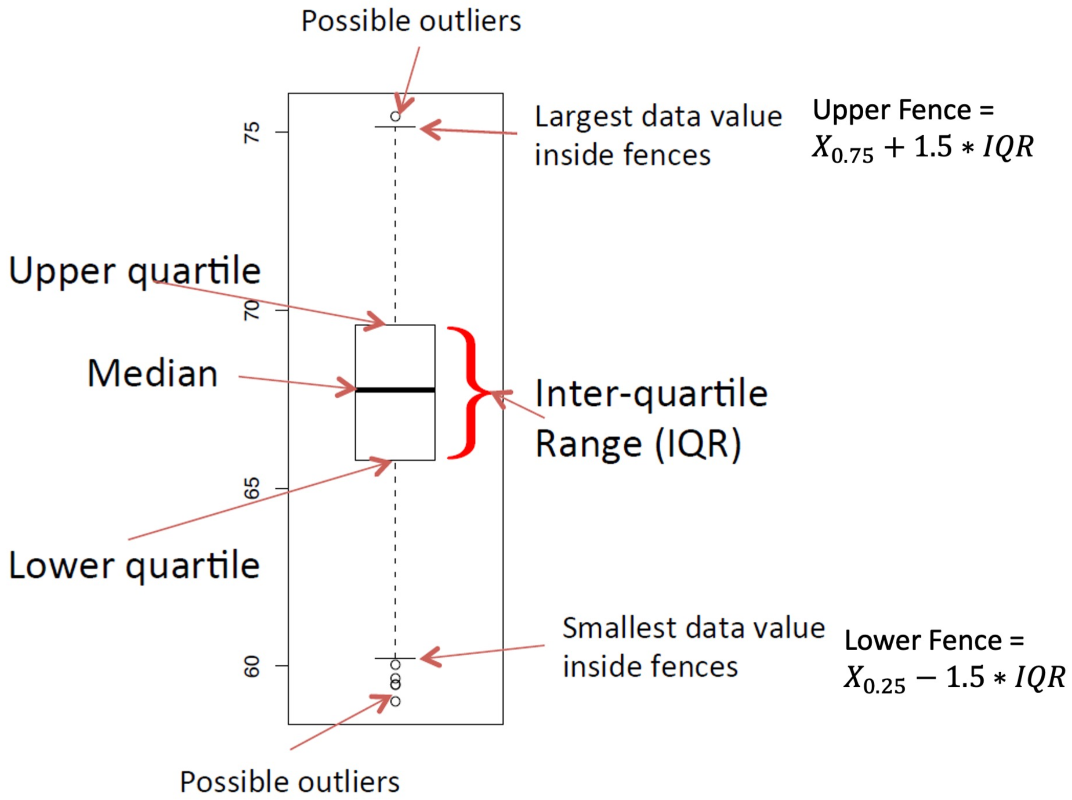


Example of frequency distributions



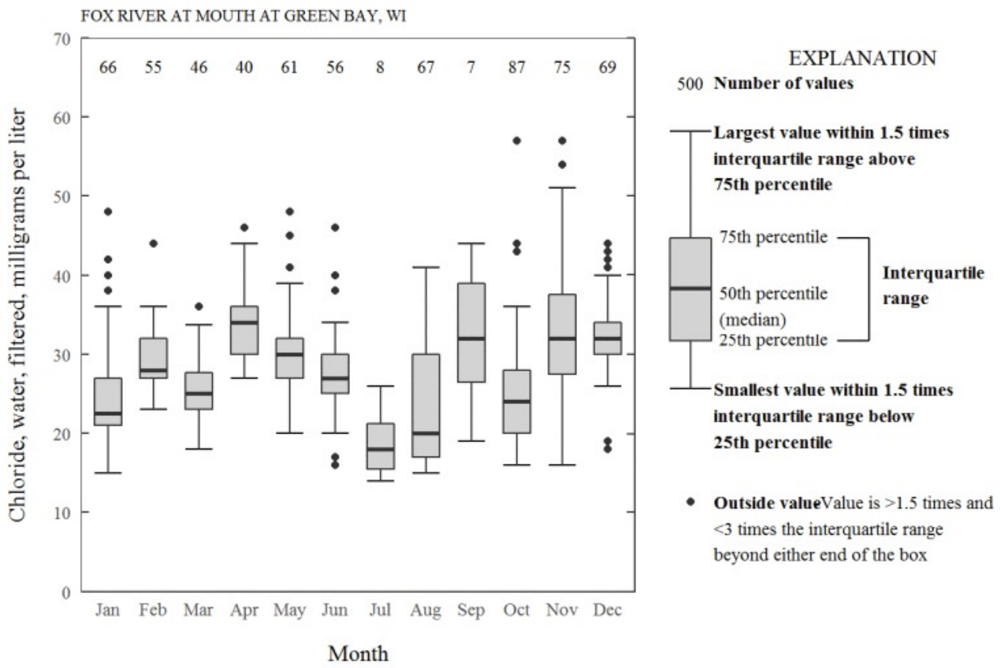
Box plot

- Box plot displayed: the lowest value, the lower quartile (Q1), the median (Q2), the upper quartile (Q3), the highest value, and the mean.



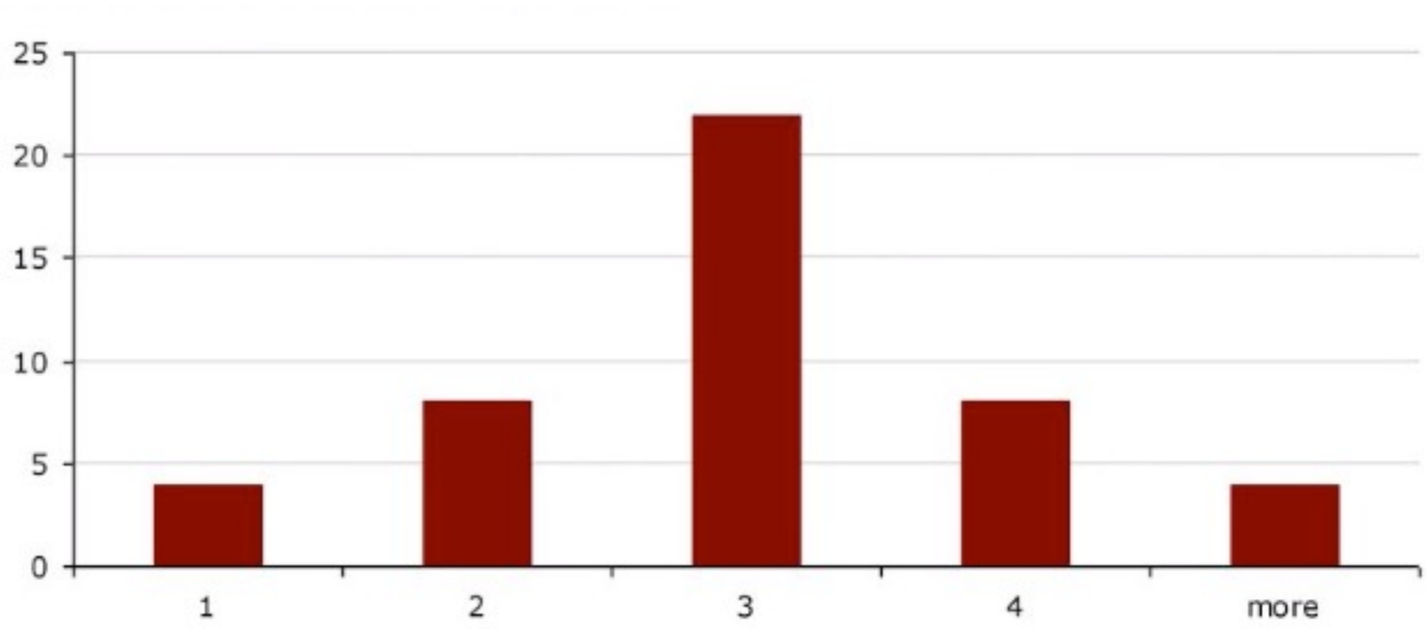
Box plot (2)

- The box plot can provide answers to the following questions:
  - Is a factor significant?
  - Does the location differ between subgroups?
  - Does the variation differ between subgroups?
  - Are there any outliers?



Skewness

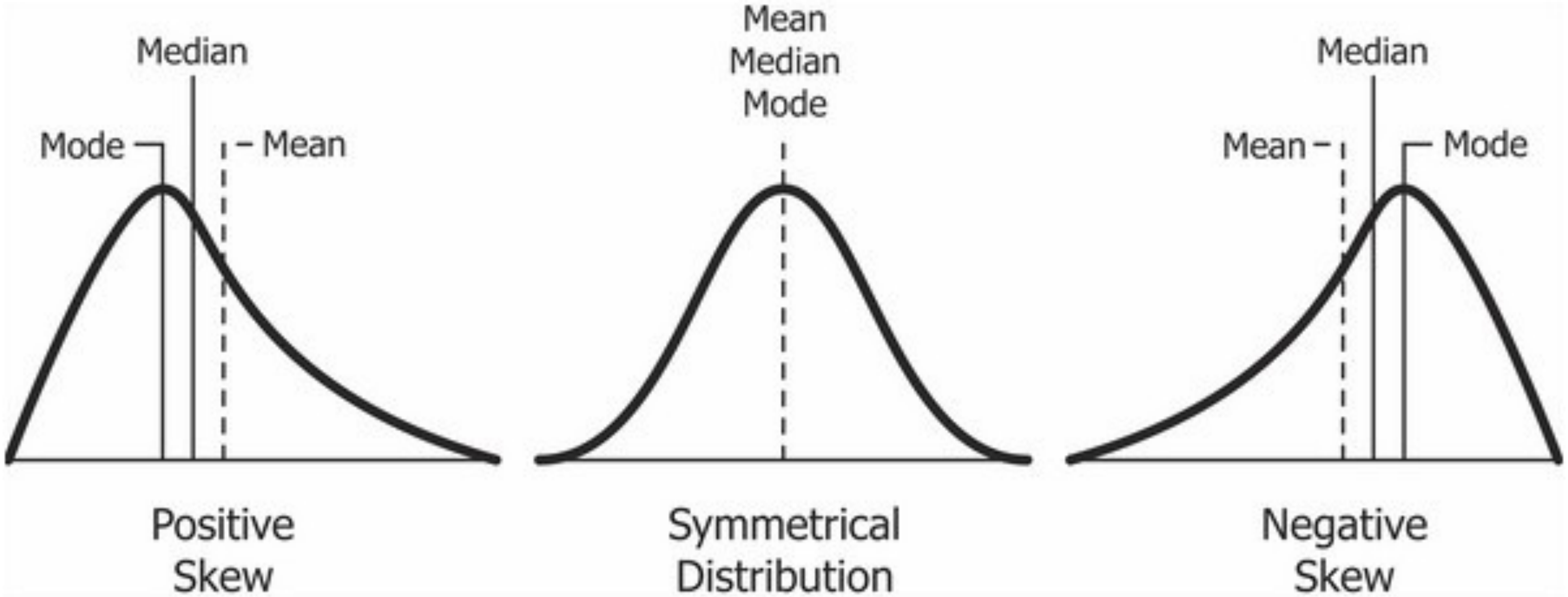
- Skewness is a measure of asymmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point
- Symetrical distribution



Mean = median = mode = 3

Negative, positive skewness

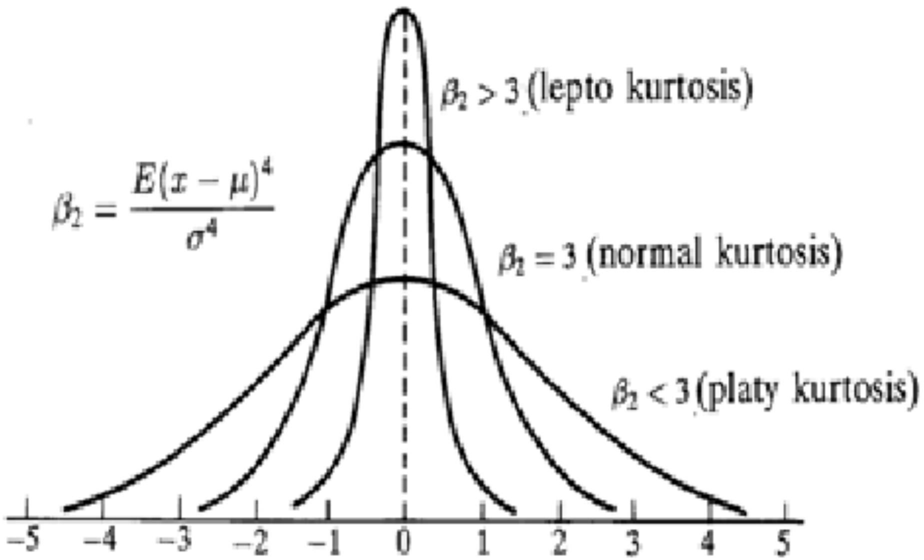
$$s_k = \frac{\sum_{i=1}^{T}\left(x_i - \bar{x}\right)^3}{\sigma^3}$$



Kurtosis

- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

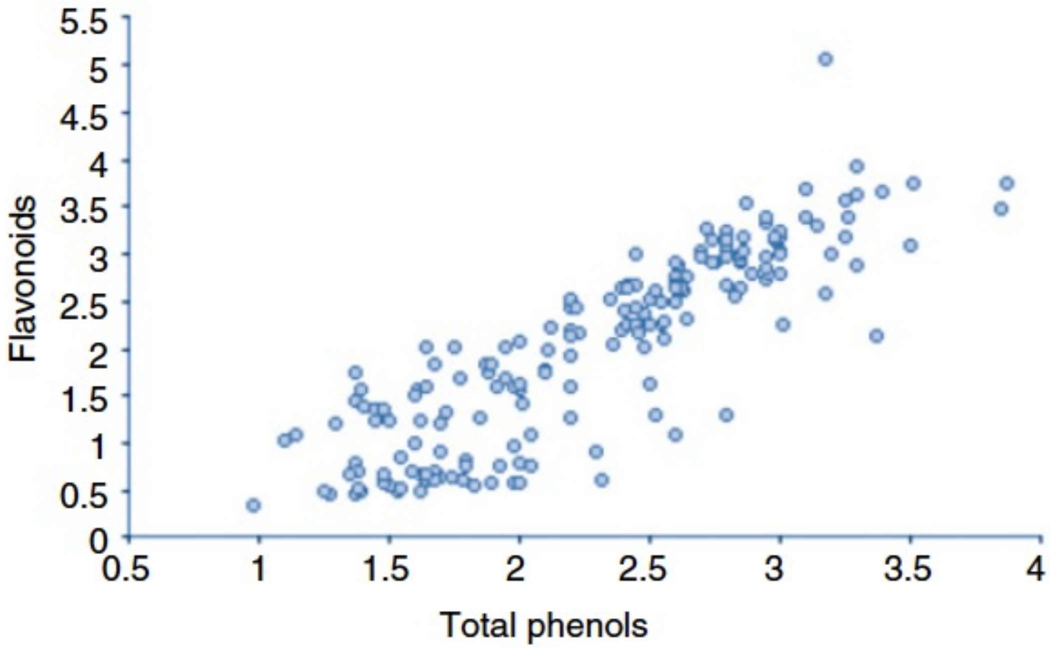
$$k = \frac{\sum_{i=1}^{T}\left(x_i - \bar{x}\right)^4}{\sigma^4}$$



Understanding relationships

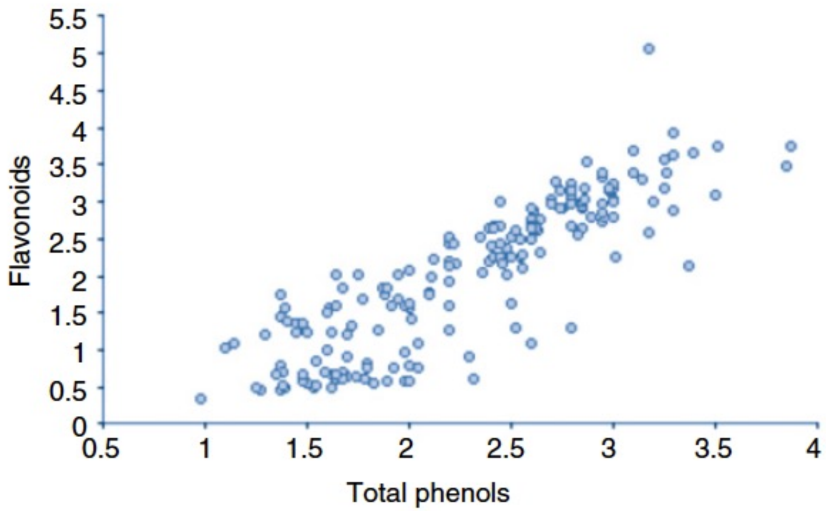
Scatter plot

- identify whether a relationship exists between two continuous variables measured on the ratio or interval scales
  - two variables are plotted on the x-and y-axis
  - each point is a single observation.

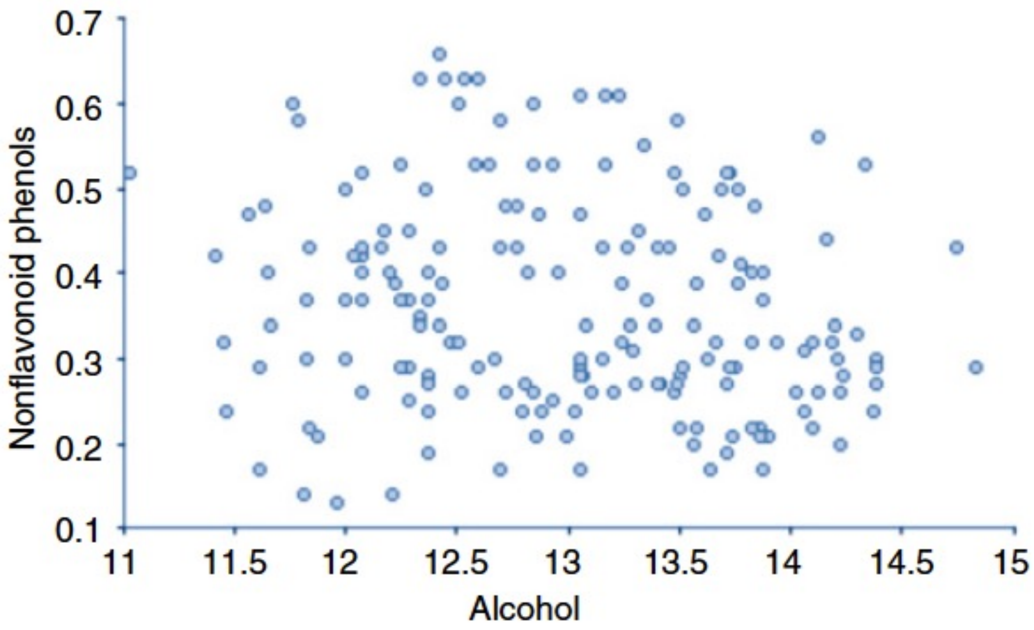


Scatter plot

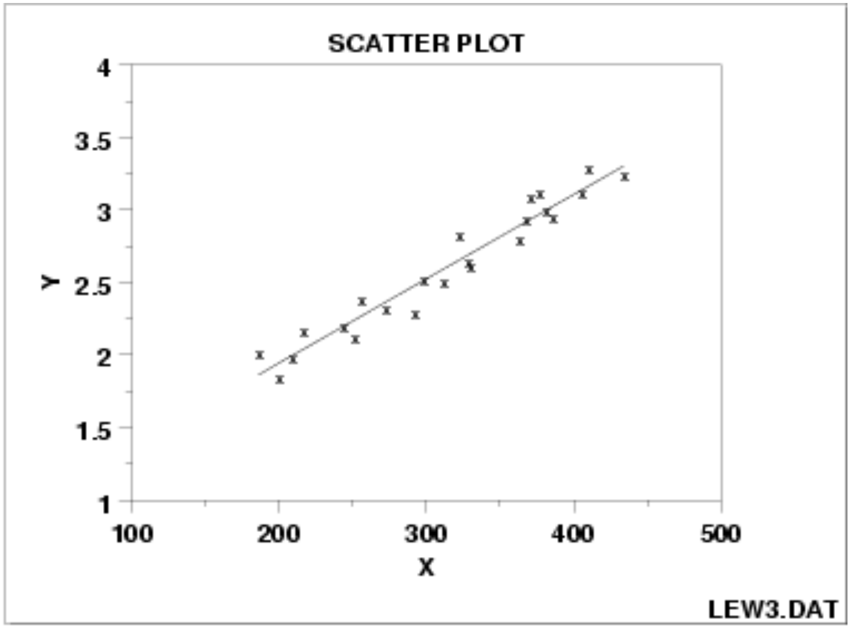
- Scatter plots can provide answers to the following questions:
  - Are variables X and Y related?
  - Are variables X and Y linearly related?
  - Are variables X and Y non-linearly related?
  - Does the variation in Y change depending on X?
  - Are there outliers?

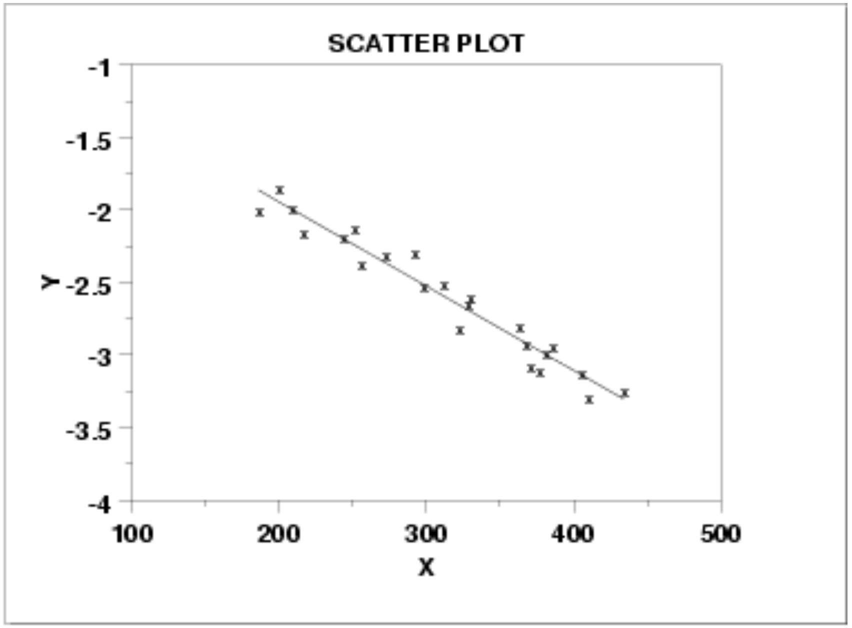


Scatter plot: No relationship

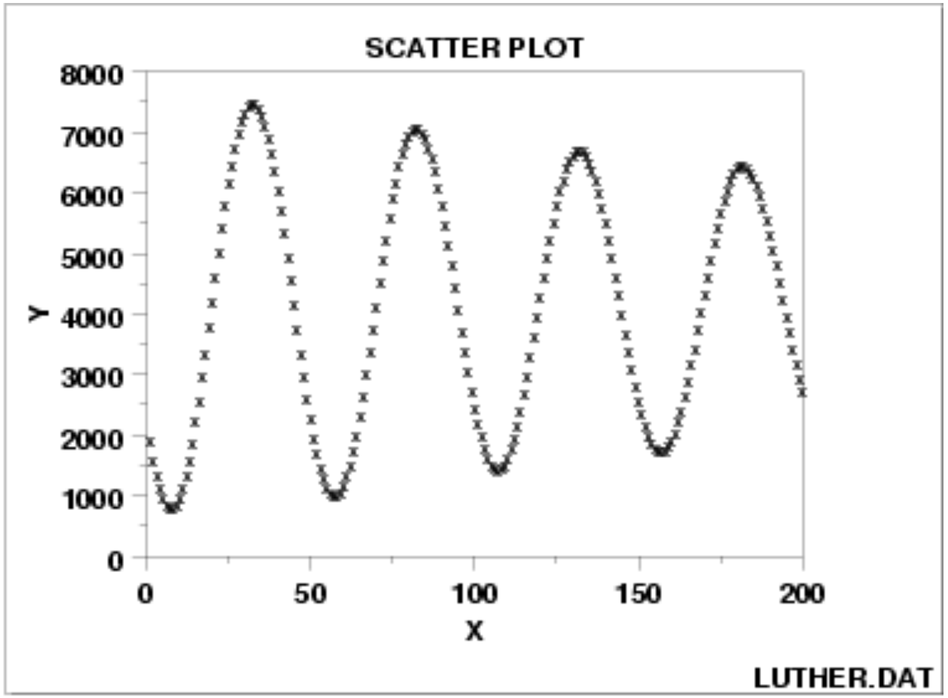


Scatter plot: Strong linear (positive - negative correlation)

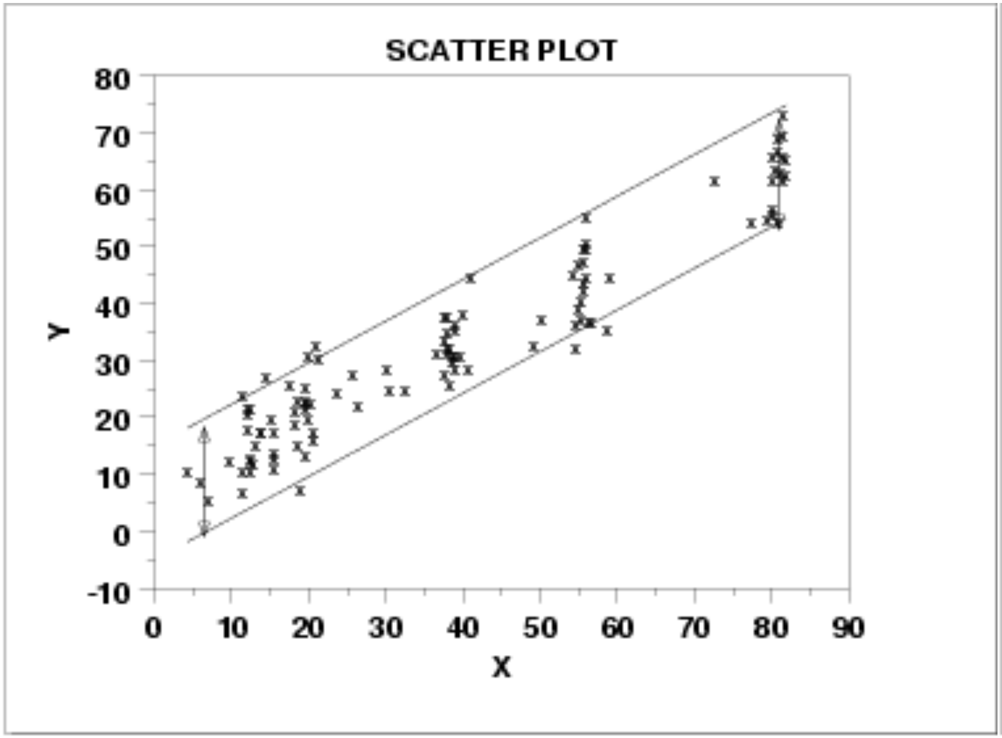


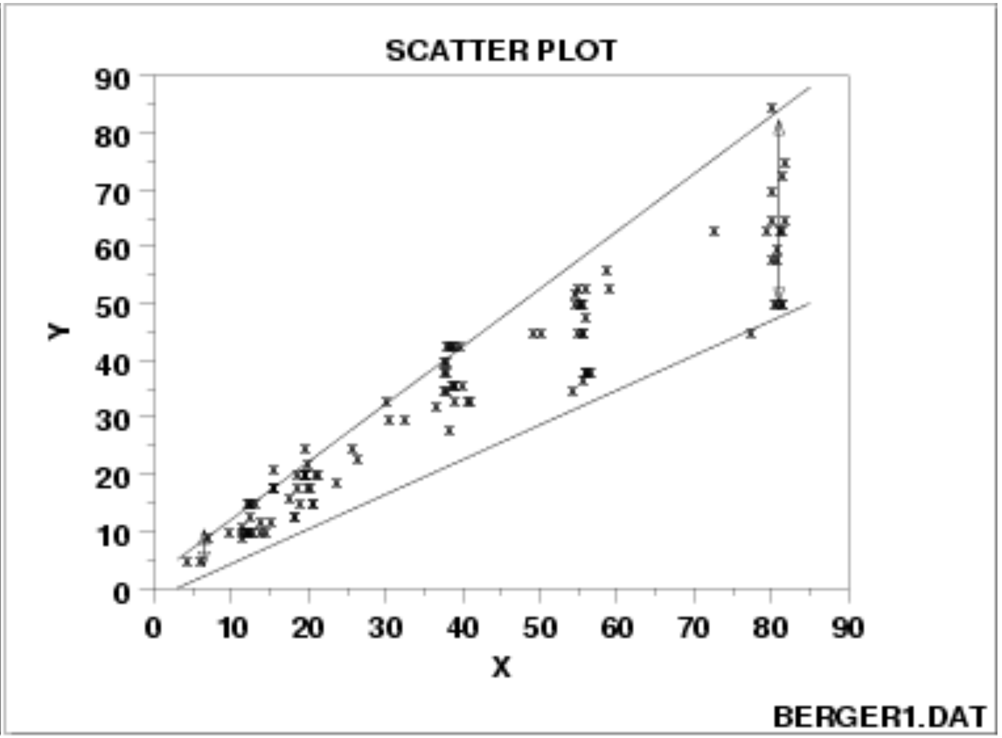


Scatter plot: Sinusoidal relationship (damped)

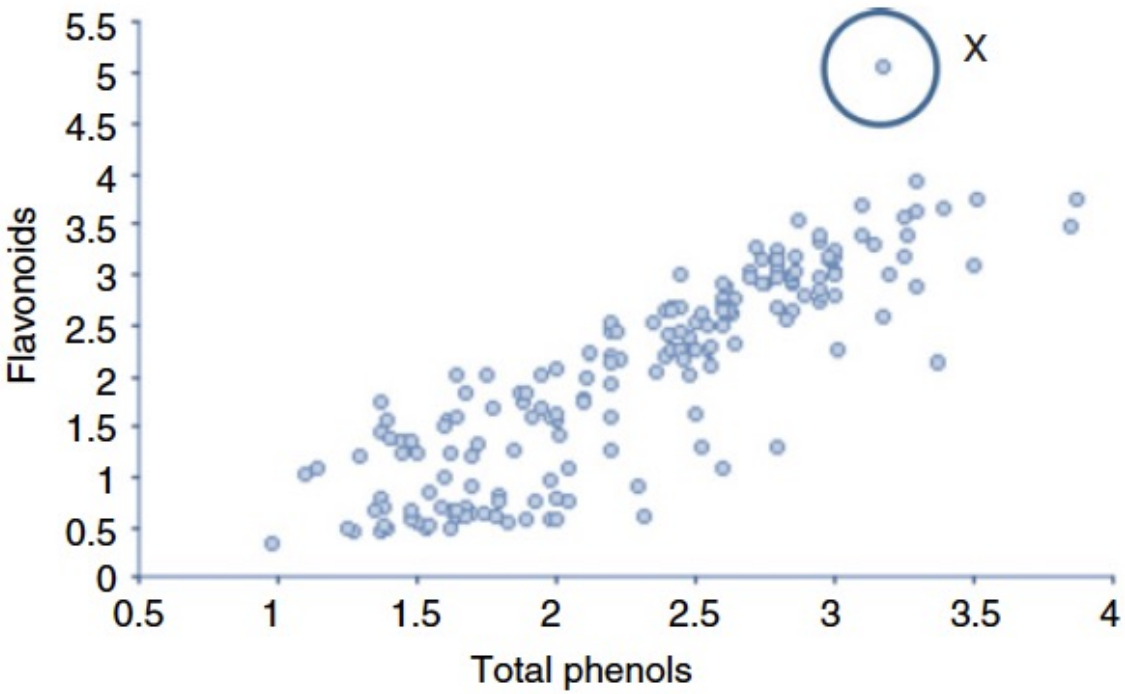


Scatter plot: variation of Y does not depend on X (homoscedastic)



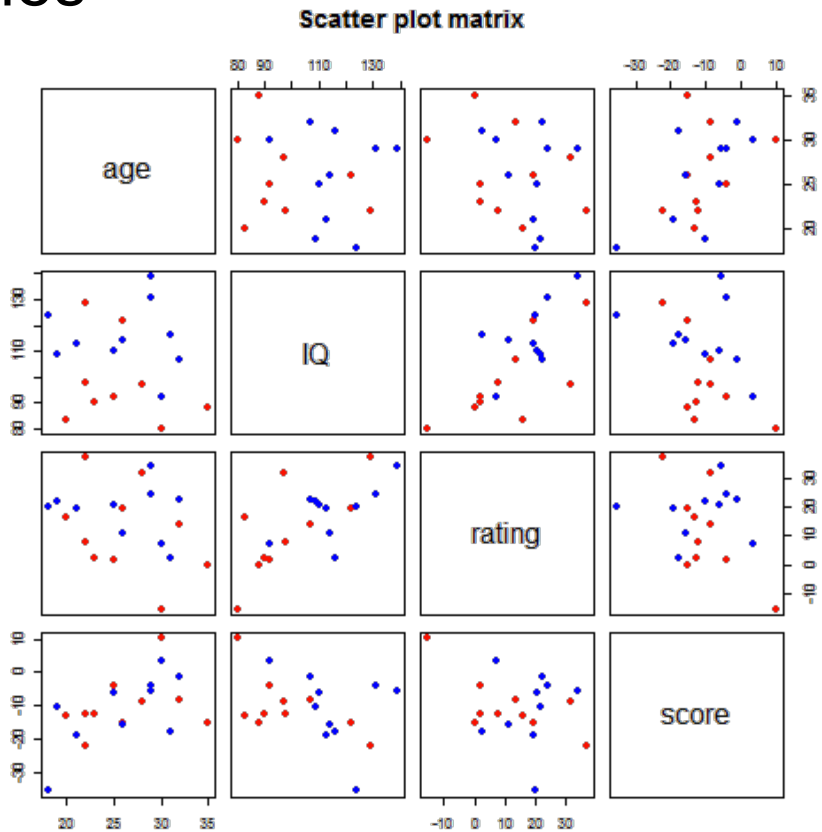


Scatter plot: Outlier



Scatterplot matrix

- a collection of **scatterplots** organized into a grid (or **matrix**).
- Each **scatterplot** shows the relationship between a pair of variables

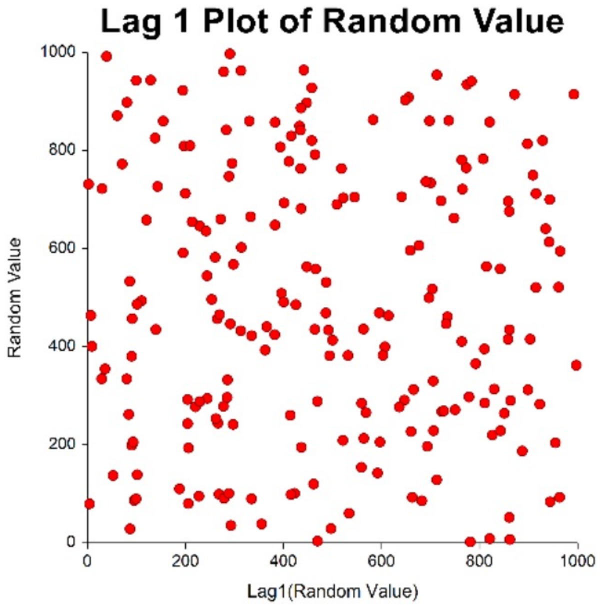


Lag plot

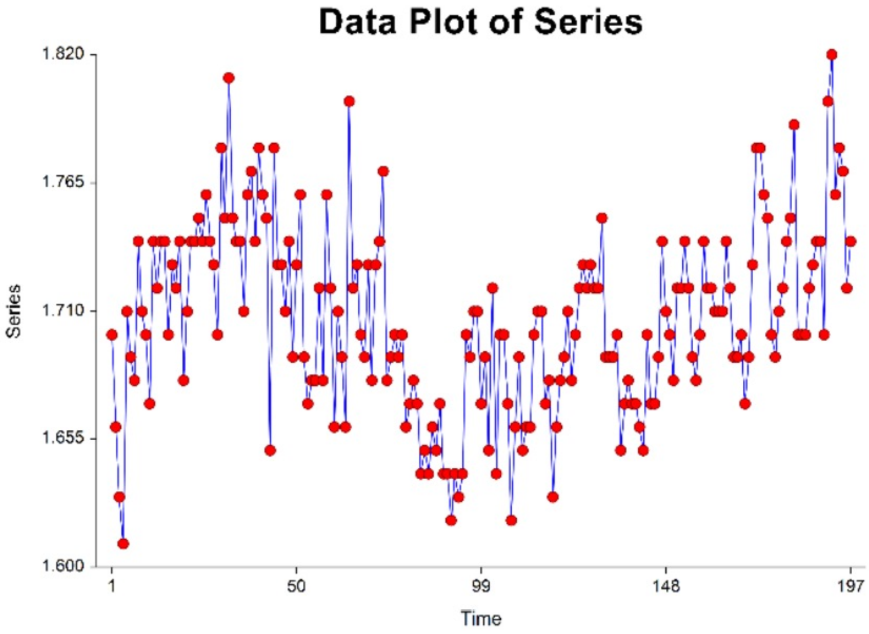
- For data values Y_1, Y_2, …, Y_N, the k-period (or k^{th}) lag of the value Y_i is defined as the data point that occurred k time points before time i. That is $\text{Lag}_k(Y_i) = Y_{i-k}$ For example, $\text{Lag}_1(Y_2) = Y_1$ and $\text{Lag}_3(Y_{10}) = Y_7$
- Lag plots can provide answers to the following questions:
  - 1. Are the data random?
  - 2. Is there serial correlation in the data?
  - 3. What is a suitable model for the data?
  - 4. Are there outliers in the data?

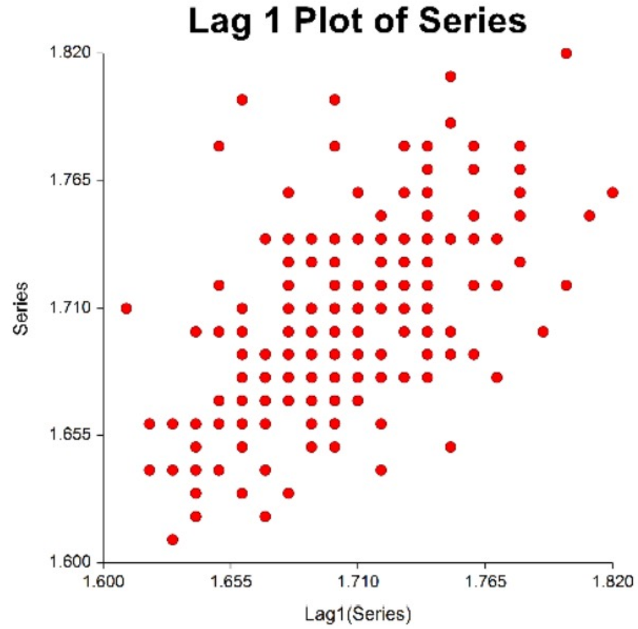
Lag plot patterns

- Random Data

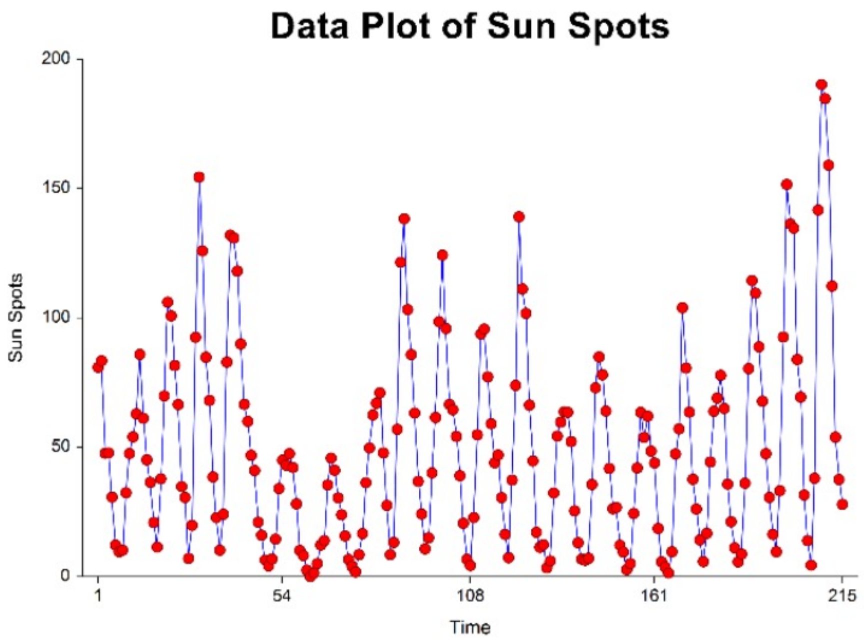


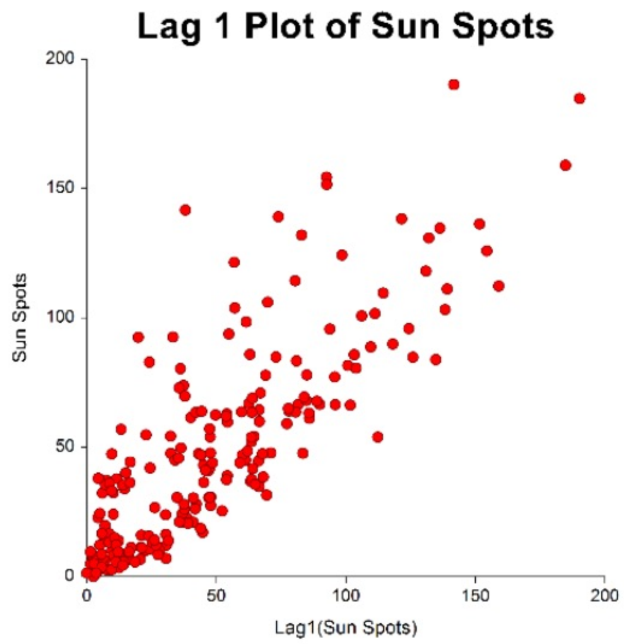
Data with weak autocorrelation



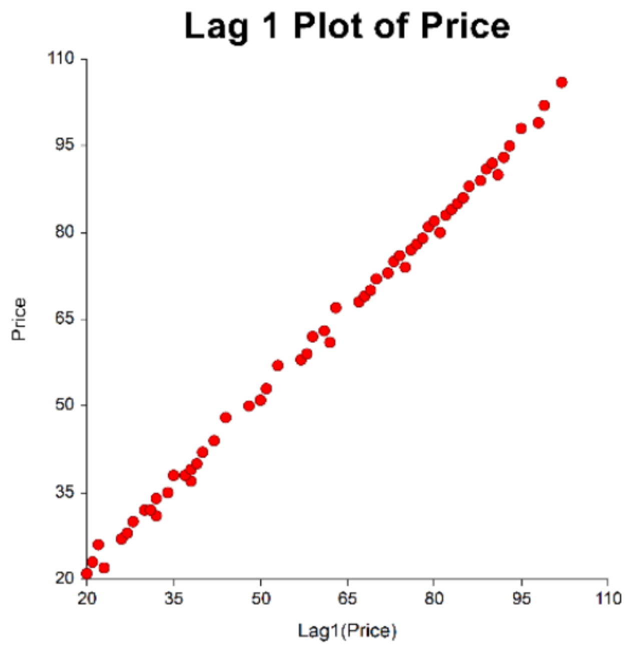


Data with moderate autocorrelation

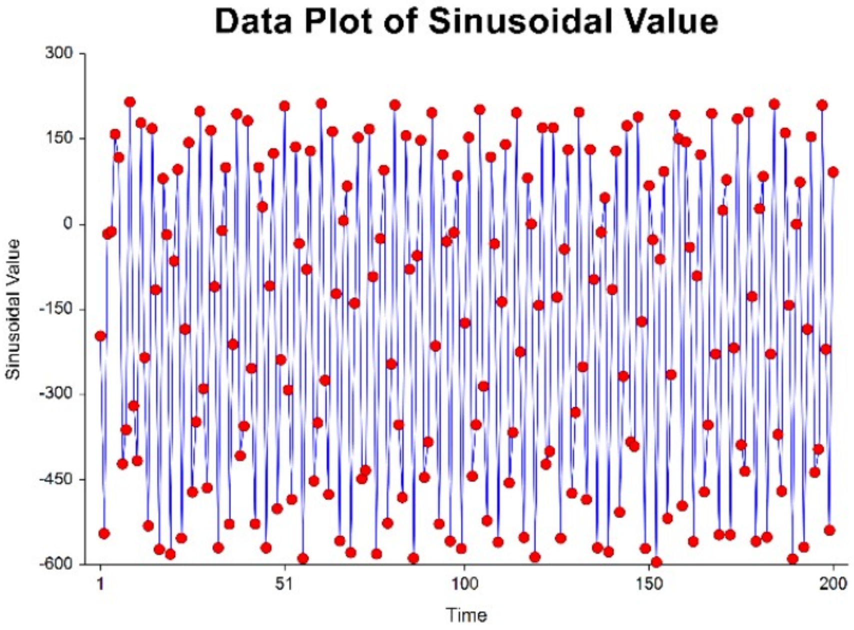


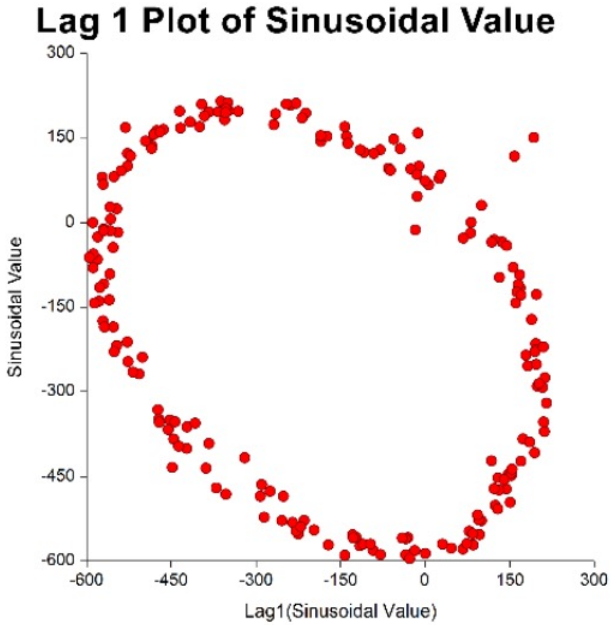


Data with high autocorrelation



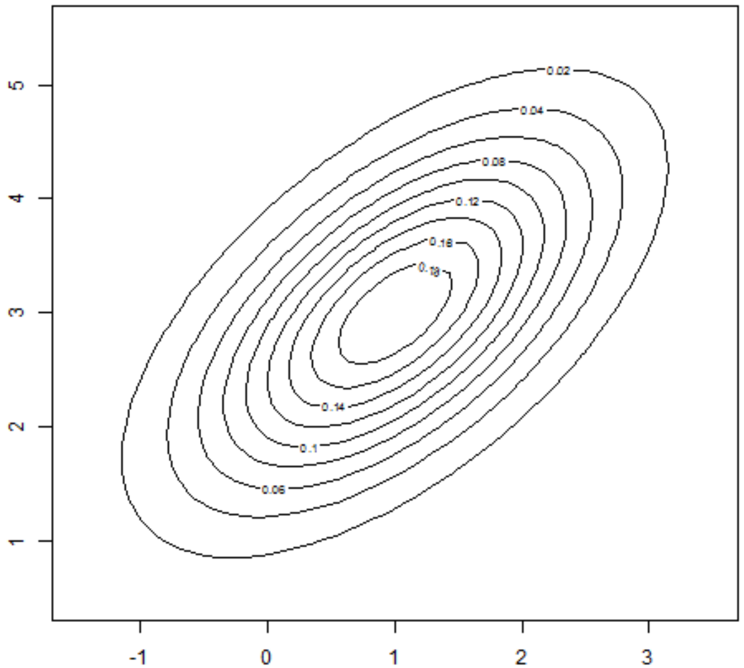
Sinusoidal data





Contour plots

- show a three-dimensional surface on a two-dimensional plane. Contour lines indicate elevations that are the same
- The contour plot is used to answer the question
  - How does Z change as a function of X and Y?



Demo

Identifing and understanding groups

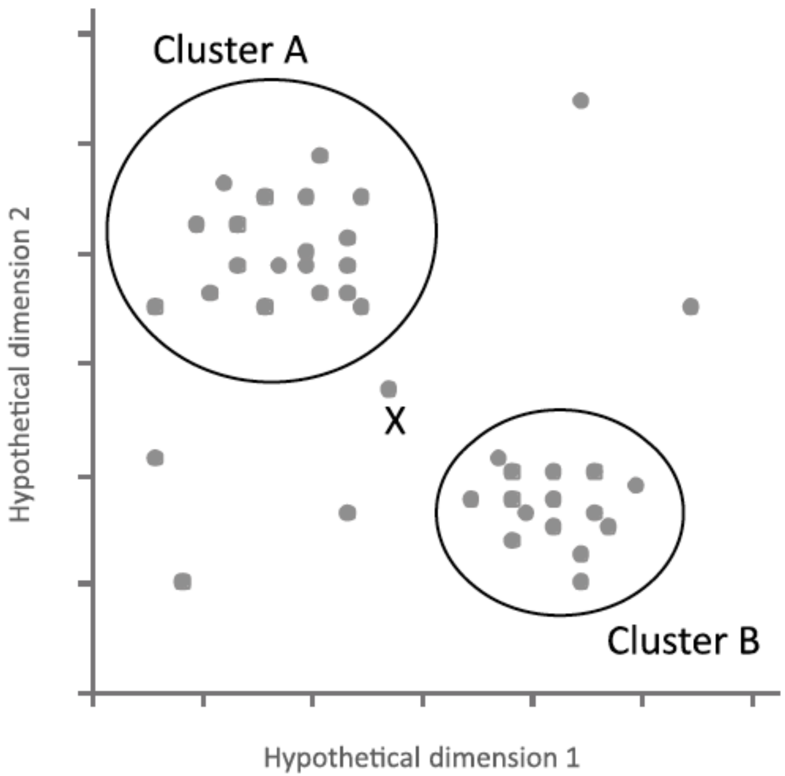
Clustering Methods in Exploratory Analysis

Motivation

- Decomposing a data set into simpler subsets helps make sense of the entire collection of observations
  - uncover relationships in the data such as groups of consumers who buy certain combinations of products
  - identify rules from the data
  - discover observations dissimilar from those in the major identified groups (possible errors or anomalies)

Clustering

- A way of grouping together data samples that are ***similar*** in some way - according to some criteria
- A form of ***unsupervised learning*** – you generally don't have examples demonstrating how the data *should* be grouped together



Can we find things that are close together?

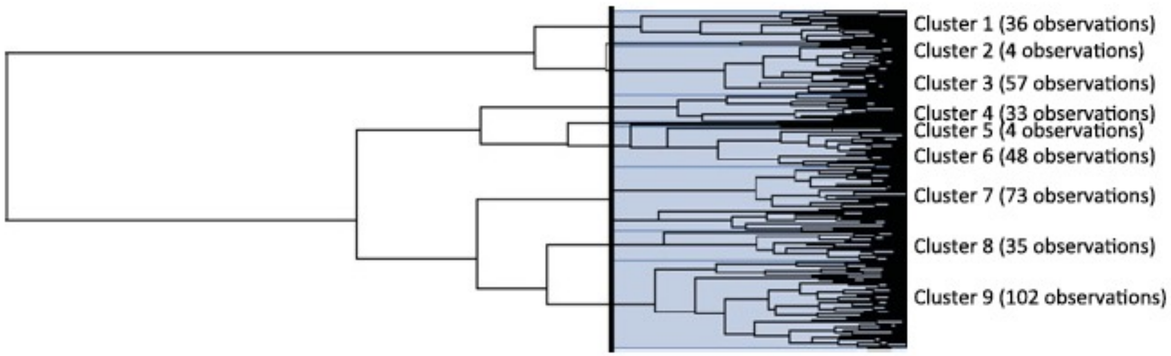
- Clustering organizes things that are close into groups
  - How do we define close?
  - How do we group things?
  - How do we visualize the grouping?
  - How do we interpret the grouping?

Types of clustering

- Hierarchical clustering
- Flat clustering

Hierarchical clustering

- An agglomerative approach
  - Find closest two things
  - Put them together
  - Find next closest
- Requires
  - A defined distance
  - A merging approach
- Produces
  - A tree showing how close things are to each other (dendrogram)



SCHO

Distances

- A method of clustering needs a way to measure how similar observations are to each other.
- Continuous - Euclidean distance
- Continuous - correlation similarity
- Binary - Manhattan distance
- Pick a distance/similarity that makes sense for the problem

Euclidean distance

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

$$d_{\mathrm{A-B}} = \sqrt{(0.7-0.6)^2 + (0.8-0.8)^2 + (0.4-0.5)^2 + (0.5-0.4)^2 + (0.2-0.2)^2}$$
$$d_{\mathrm{A-B}} = 0.17$$

$$d_{\mathrm{A-C}} = \sqrt{(0.7-0.8)^2 + (0.8-0.9)^2 + (0.4-0.7)^2 + (0.5-0.8)^2 + (0.2-0.9)^2}$$
$$d_{\mathrm{A-C}} = 0.83$$

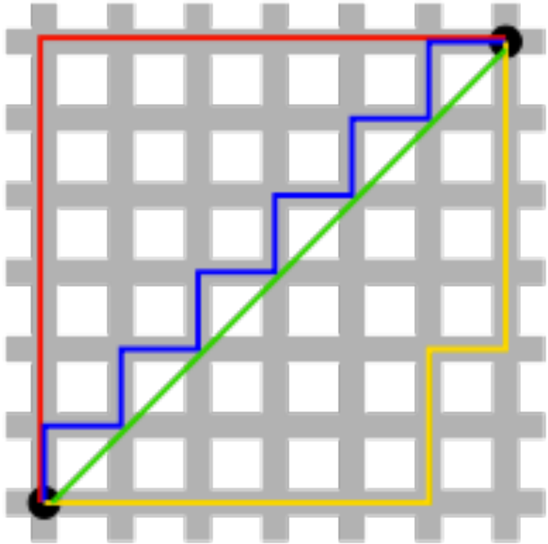
Manhattan distance

- is the sum of the lengths of the projections of the line segment between the points onto t

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

where $(\mathbf{p}, \mathbf{q})$ are vectors

$$\mathbf{p} = (p_1, p_2, \dots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \dots, q_n)$$

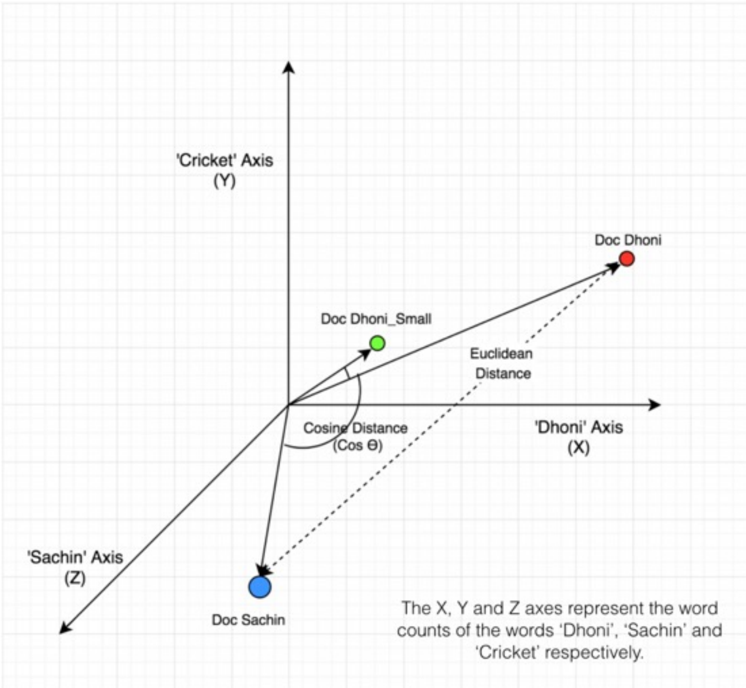


Cosine distance

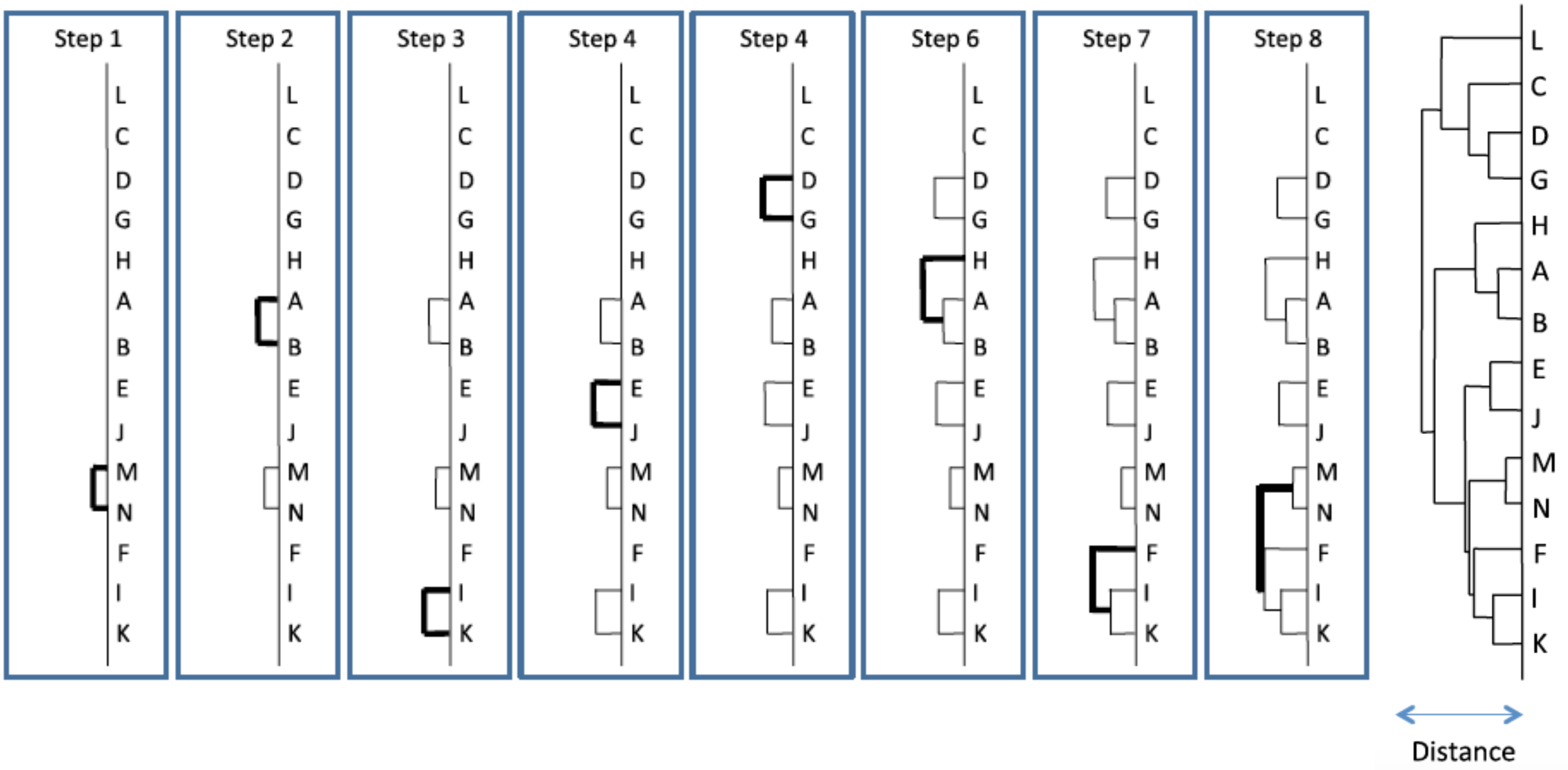
$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

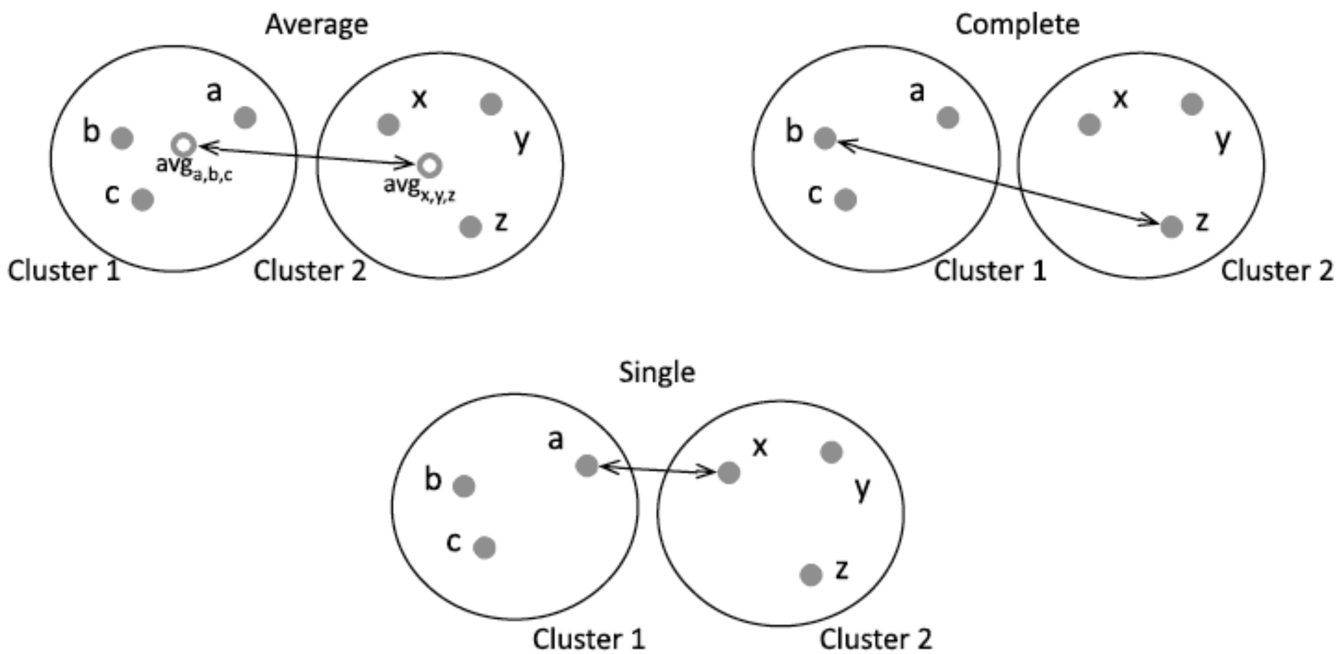
Projection of Documents in 3D Space



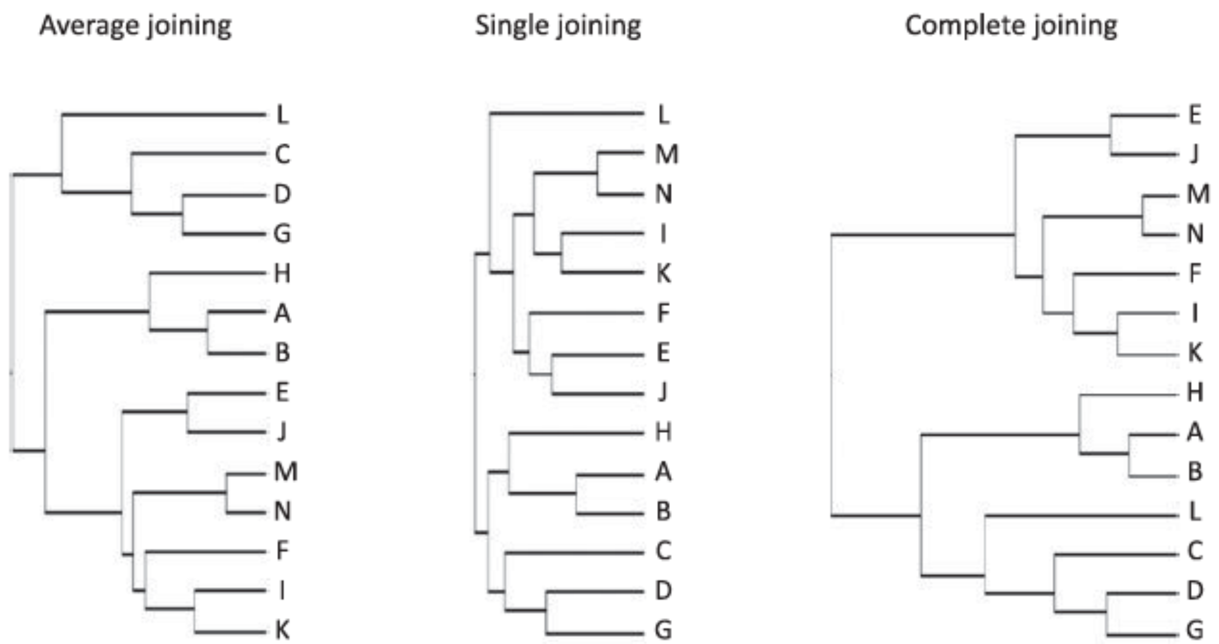
Agglomerative Hierarchical Clustering Algorithm



Linkage rules

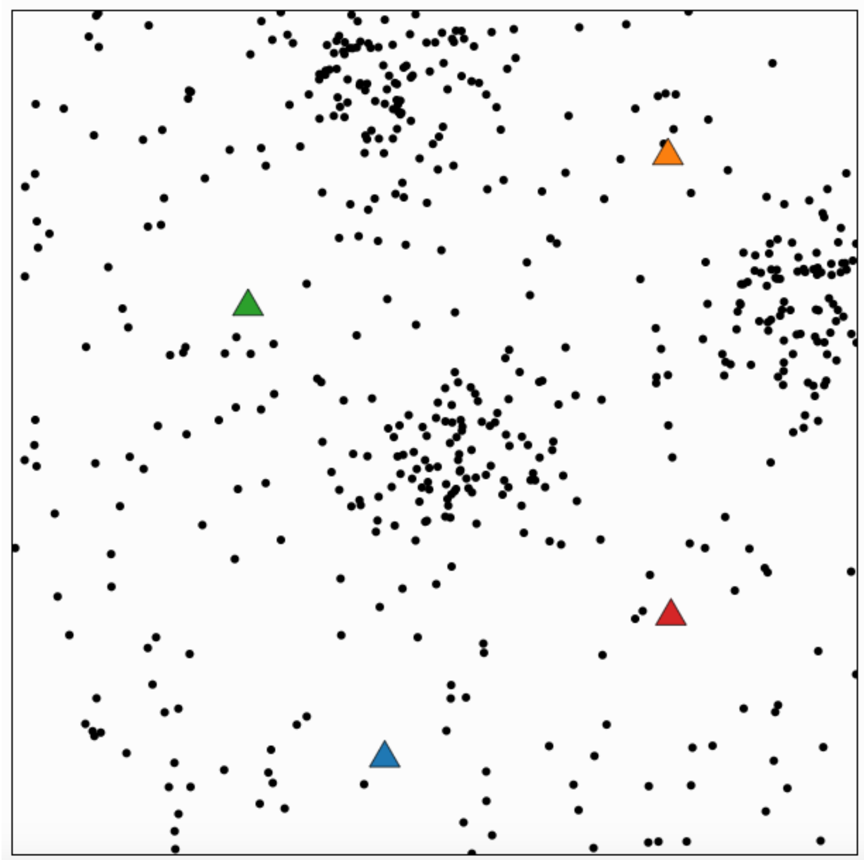


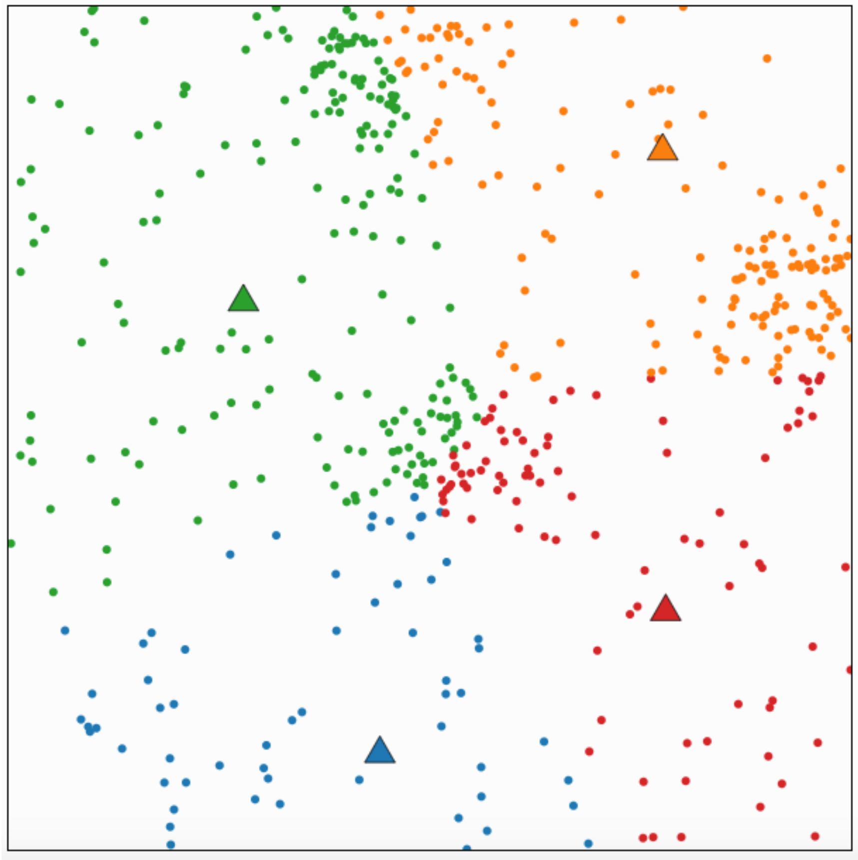
AHC result

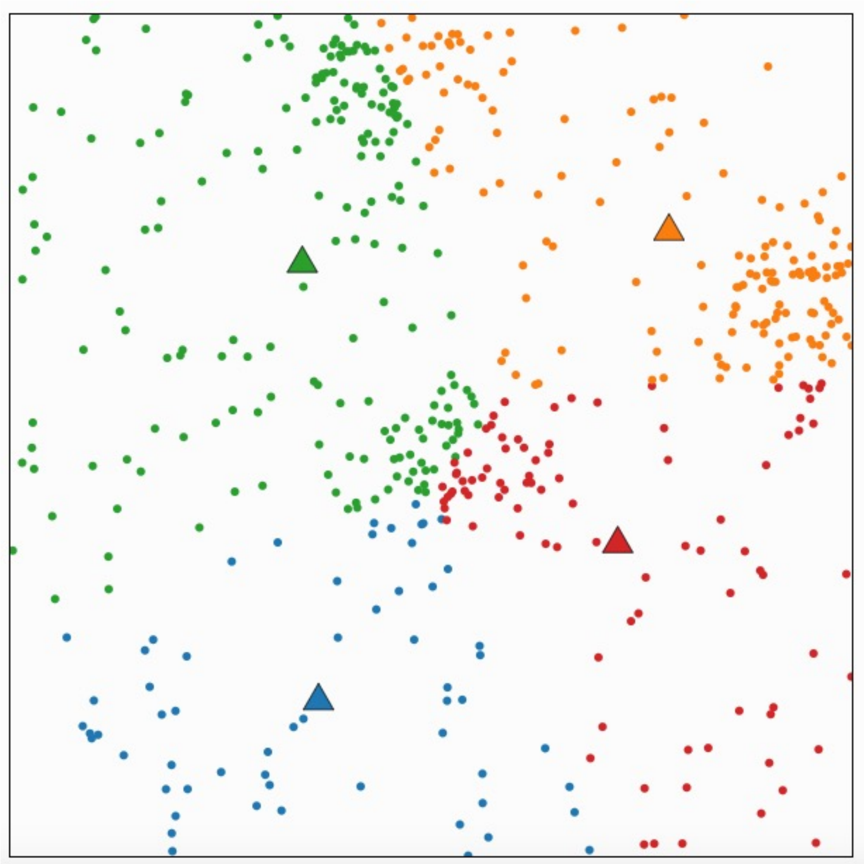

Average joining
L
C
D
G
H
A
B
E
J
M
N
F
I
K
Single joining
L
M
N
I
K
F
E
J
H
A
B
C
D
G
Complete joining
E
J
M
N
F
I
K
H
A
B
L
C
D
G

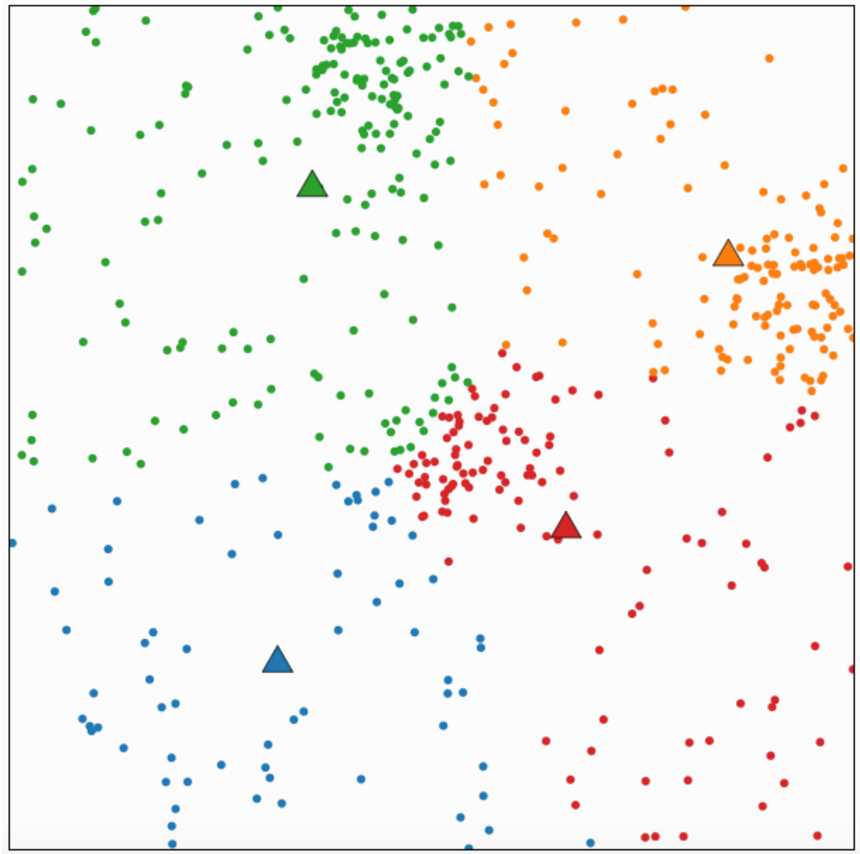
K-mean clustering

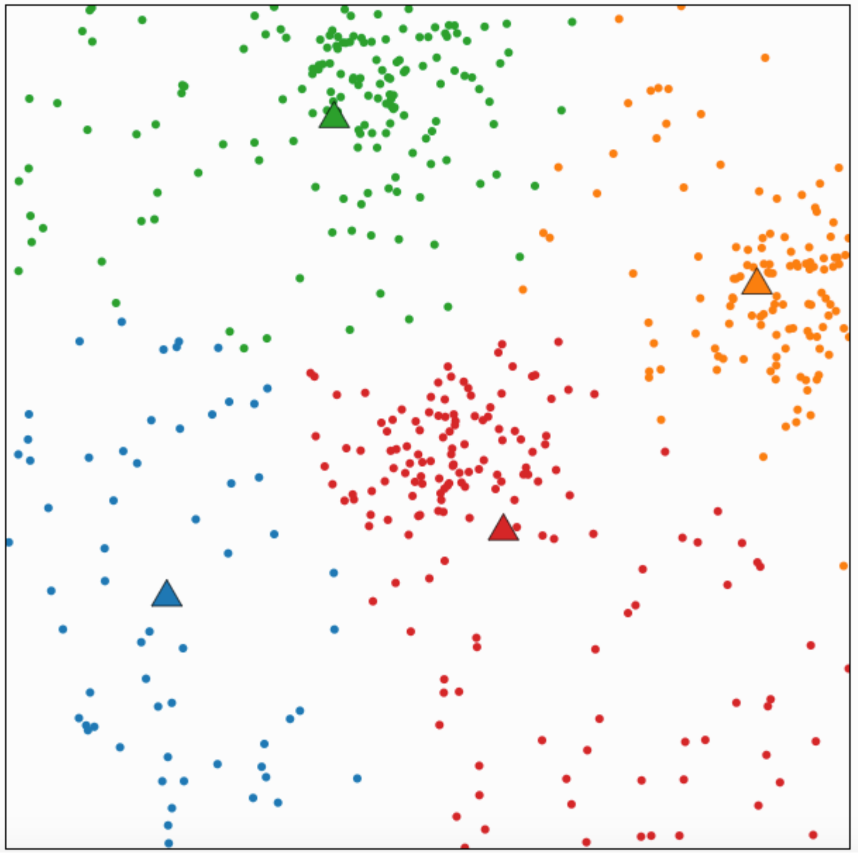
- A partitioning approach
  - Fix a number of clusters
  - Get “centroids” of each cluster
  - Assign things to closest centroid
  - Recalculate centroids
- Requires
  - A defined distance metric
  - A number of clusters
  - An initial guess as to cluster centroids
- Produces
  - Final estimate of cluster centroids
  - An assignment of each point to clusters











Dimensionality reduction

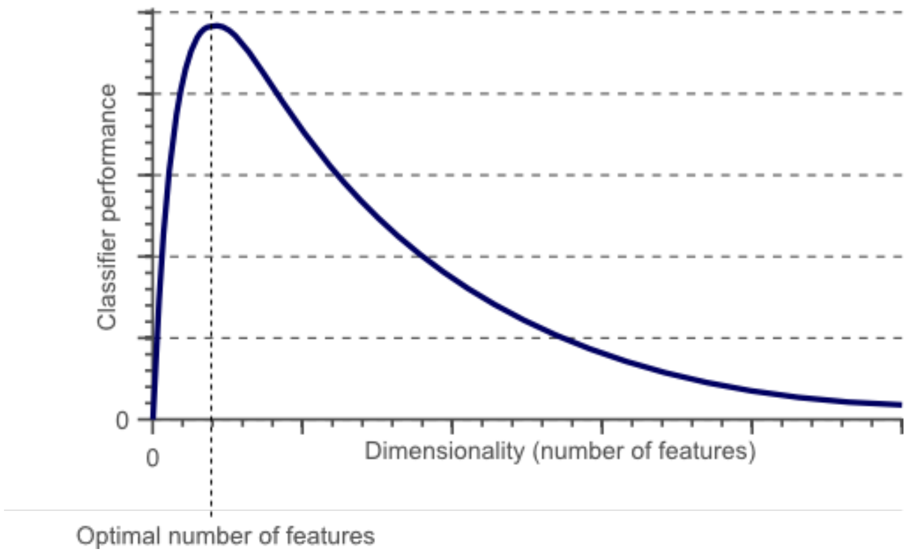
Principal Components Analysis and Singular Value Decomposition

Motivation

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality. Irrelevant and redundant features can “confuse” learners!
  - The intrinsic dimension may be small.

Curse of dimensionality

- The required number of samples (to achieve the same accuracy) grows exponentionally with the number of variables!
- In practice: number of training examples is fixed!
  - => the classifier's performance usually will degrade for a large number of features!

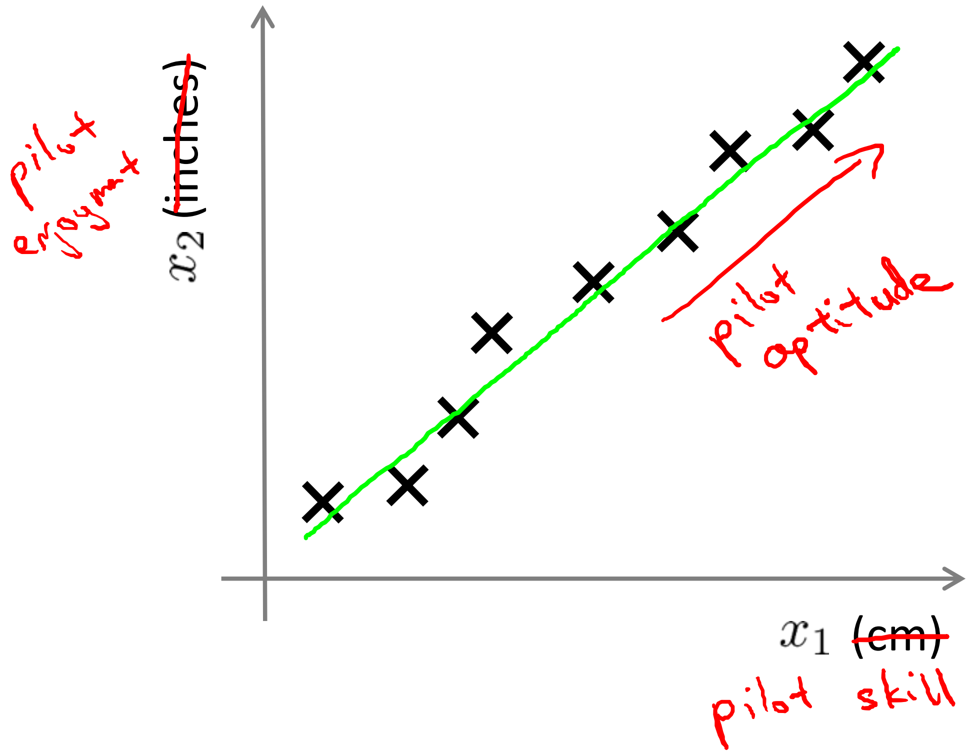


After a certain point, increasing the dimensionality of the problem by adding new features would actually degrade the performance of classifier.

Motivation

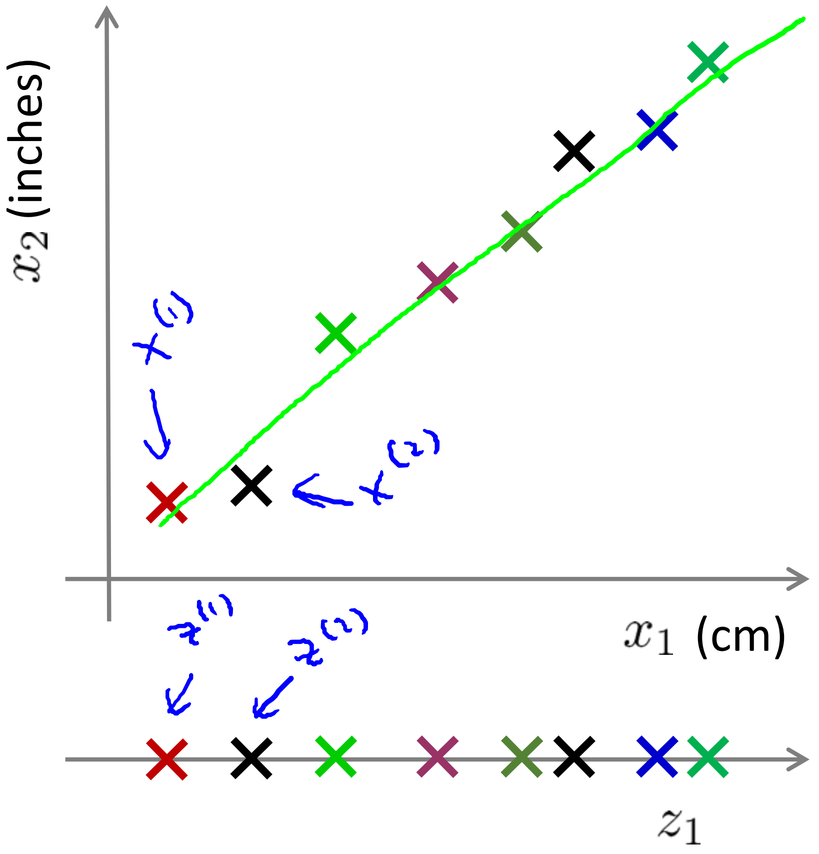
- Dimensionality reduction is an effective approach to downsizing data
  - Visualization: projection of high-dimensional data onto 2D or 3D.
  - Data compression: efficient storage and retrieval.
  - Noise removal: positive effect on query accuracy.

Data compression



Reduce data from 2D to 1D

Data compression (2)



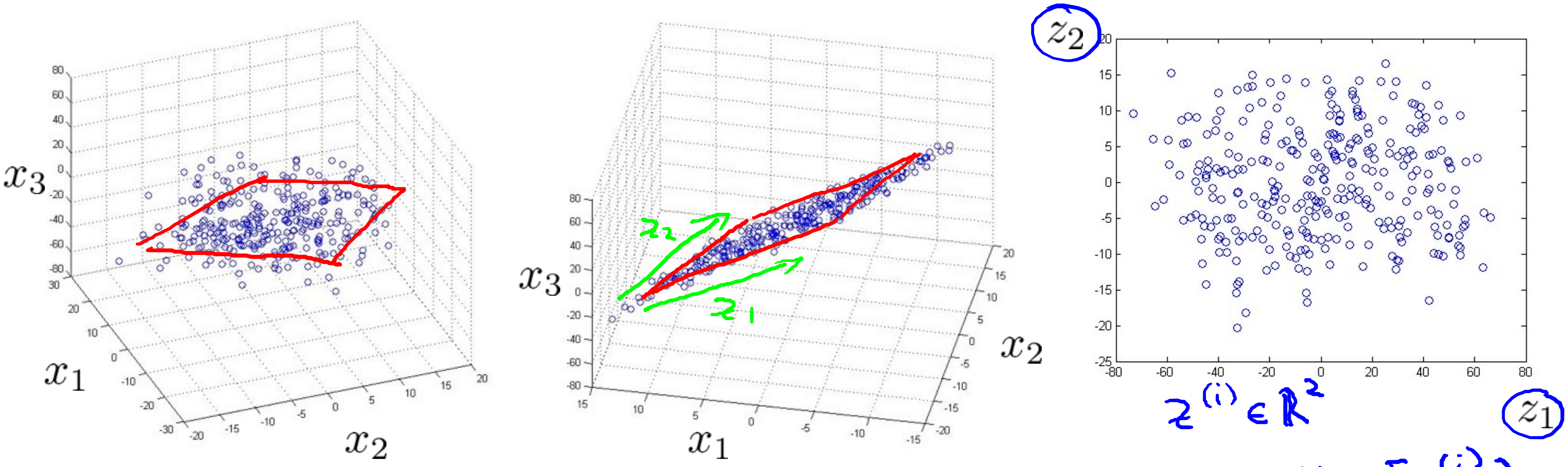
Reduce data from 2D to 1D

$$x^{(1)} \in \mathbb{R}^2 \quad \rightarrow z^{(1)} \in \mathbb{R}$$
$$x^{(2)} \in \mathbb{R}^2 \quad \rightarrow z^{(2)} \in \mathbb{R}$$
$$\vdots$$
$$x^{(m)} \in \mathbb{R}^2 \quad \rightarrow z^{(m)} \in \mathbb{R}$$

Data compression (2)

1000D → 100D

Reduce data from 3D to 2D

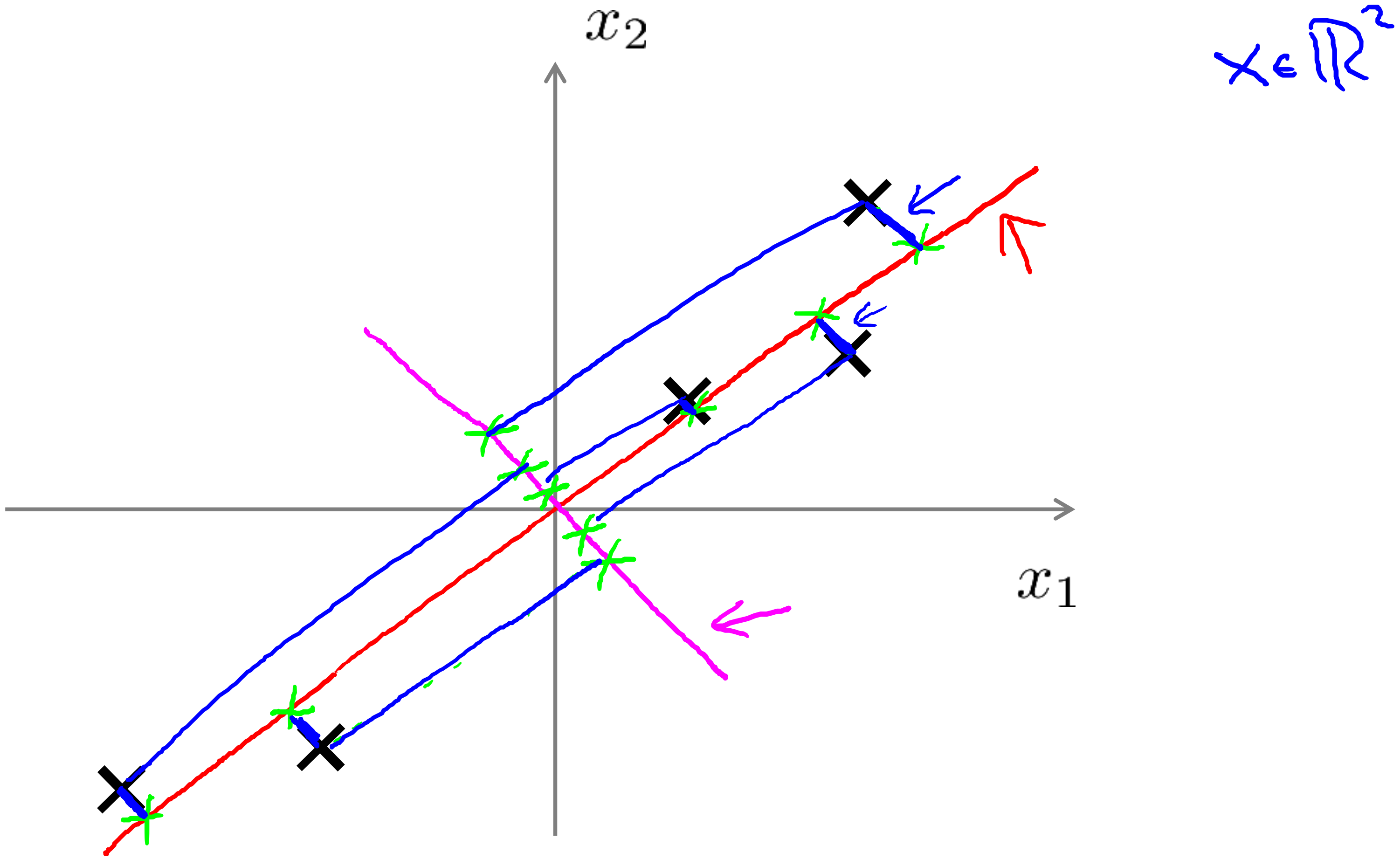


$x^{(i)} \in \mathbb{R}^3$

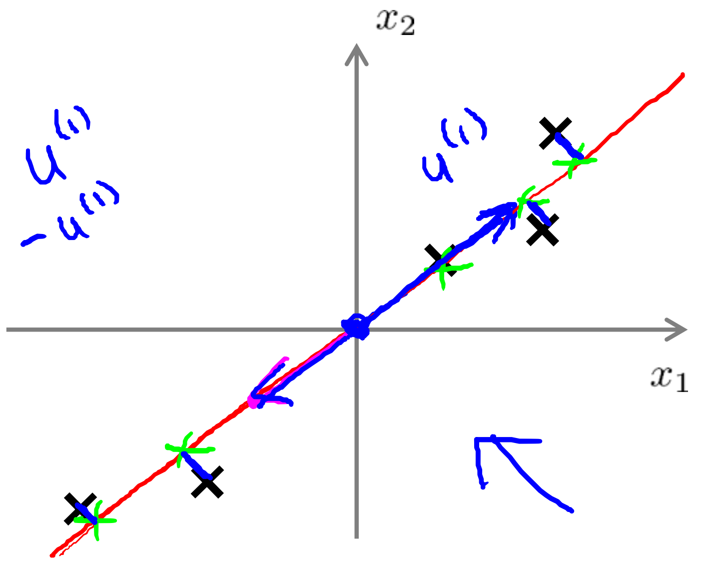
$z^{(i)} \in \mathbb{R}^2$

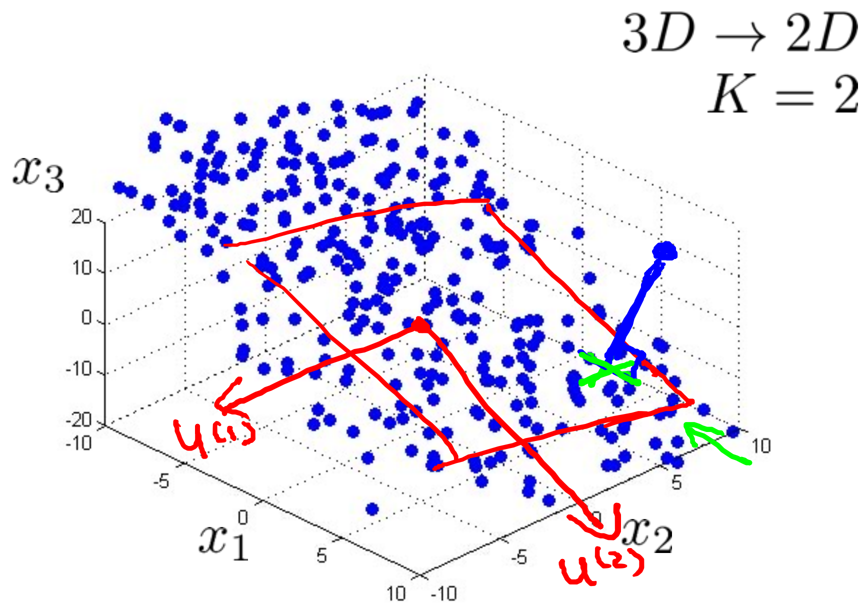
$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ $\quad z^{(i)} = \begin{bmatrix} z_1^{(i)} \\ z_2^{(i)} \end{bmatrix}$

Principal Component Analysis (PCA) problem formulation



Principal Component Analysis (PCA) problem formulation



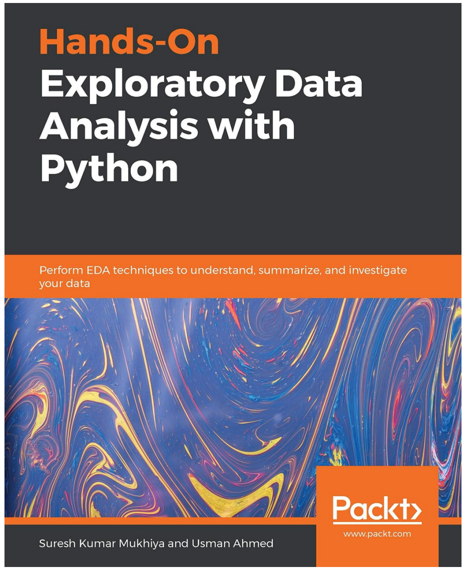


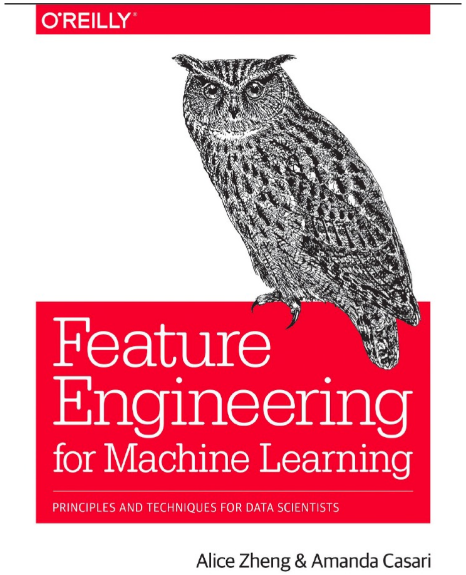
Reduce from 2-dimension to 1-dimension: Find a direction (a vector $u^{(1)} \in \mathbb{R}^n$) onto which to project the data so as to minimize the projection error.

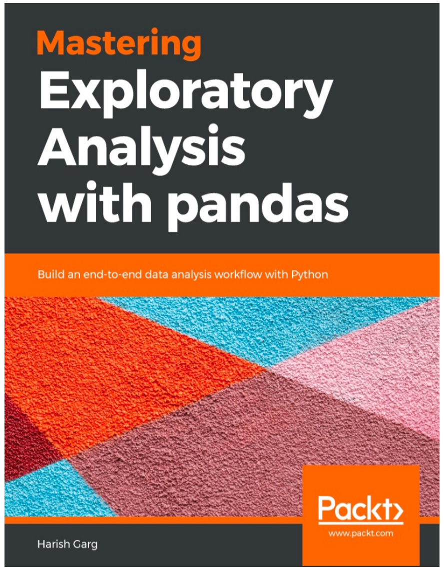
Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, so as to minimize the projection error.

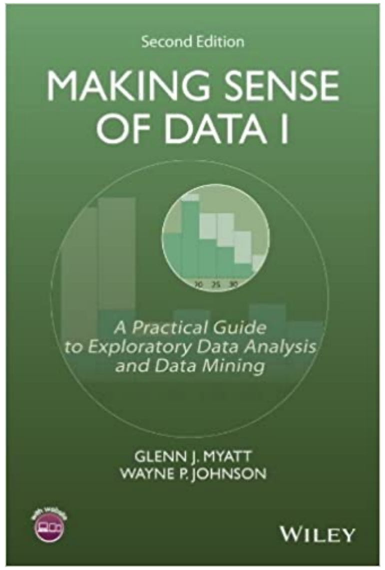
Demo

References


Hands-On
Exploratory Data Analysis with Python
Perform EDA techniques to understand, summarize, and investigate your data
Packt>
www.packt.com
Suresh Kumar Mukhiya and Usman Ahmed


O'REILLY
Feature Engineering for Machine Learning
PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS
Alice Zheng & Amanda Casari


Mastering
Exploratory Analysis with pandas
Build an end-to-end data analysis workflow with Python
Packt>
www.packt.com
Harish Garg


Second Edition
MAKING SENSE OF DATA I
A Practical Guide to Exploratory Data Analysis and Data Mining
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you
for your
attention!!!

soict.hust.edu.vn/ fb.com/groups/soict