Business Intelligence

© Han, J., Kamber, M. and Pei, J., 2011. Data mining concepts and techniques 3rd edition. *The Morgan Kaufmann Series in Data Management Systems*.

Lecture 3: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	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

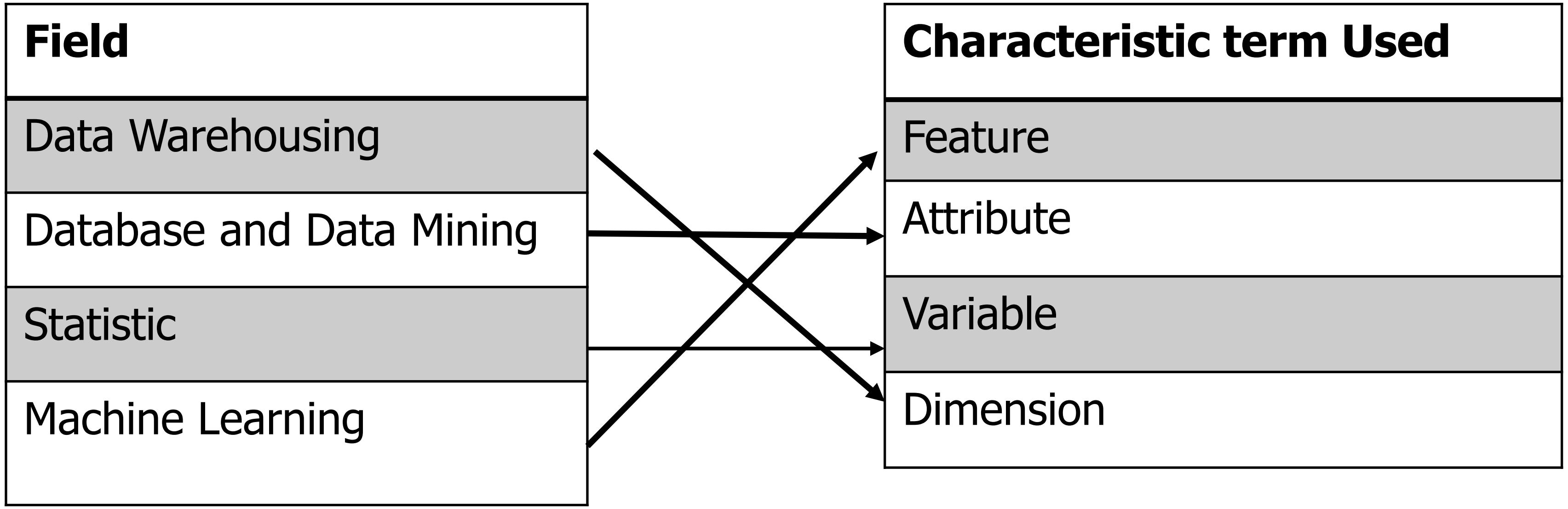
TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Data Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Attribute refers to the characteristic of the data object.
  - The nouns defining the characteristics are used interchangeably: Attribute, dimension, feature, and variable.



Field
Data Warehousing
Database and Data Mining
Statistic
Machine Learning

Characteristic term Used
Feature
Attribute
Variable
Dimension

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., cat or dog
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
    - the positive (1) and negative (0) outcomes of a **disease test.**
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Recap

- Data sets are made up of data objects. A **data object** represents an entity.
- Data objects are described by attributes. Attributes can be nominal, binary, ordinal, or numeric.
- The values of a **nominal** (or **categorical**) **attribute** are symbols or names of things, where each value represents some kind of category, code, or state.
- **Binary attributes** are nominal attributes with only two possible states (such as 1 and 0 or true and false). If the two states are equally important, the attribute is *symmetric*; otherwise it is *asymmetric*.
- An **ordinal attribute** is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
- A **numeric attribute** is *quantitative* (i.e., it is a measurable quantity) represented in integer or real values. Numeric attribute types can be *interval-scaled* or *ratio-scaled*.
- The values of an **interval-scaled attribute** are measured in fixed and equal units. **Ratio-scaled attributes** are numeric attributes with an inherent zero-point.
- Measurements are ratio-scaled in that we can speak of values as being an order of magnitude larger than the unit of measurement.

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread

- Measures of central tendency
  - measure the location of the middle or center of a data distribution.
  - Mean, Median, and Mode

- Measures the dispersion
  - how are the data spread out?
  - Range, Quartiles, Variance, Standard Deviation, and Interquartile Range
  - useful for identifying outliers

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  - Weighted arithmetic mean:

  $$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

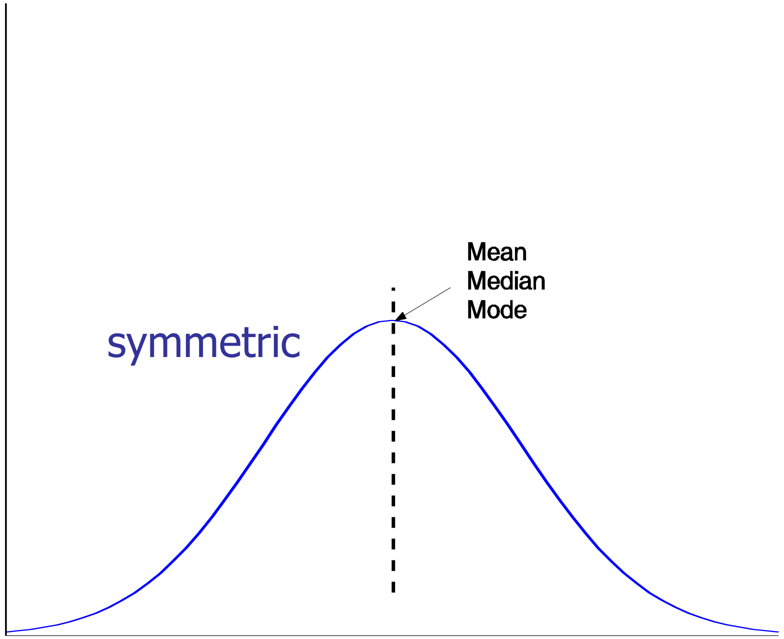
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

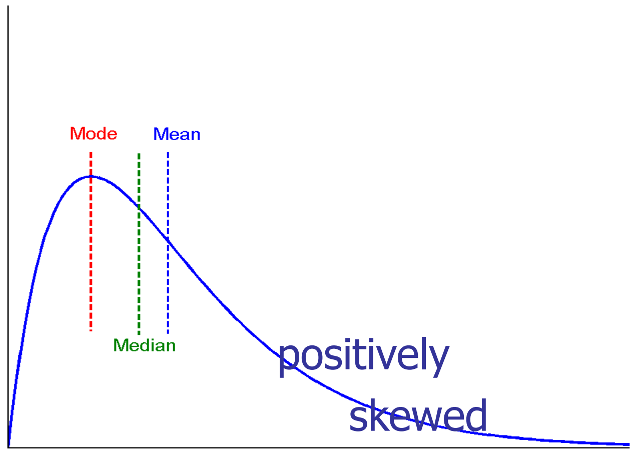
age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

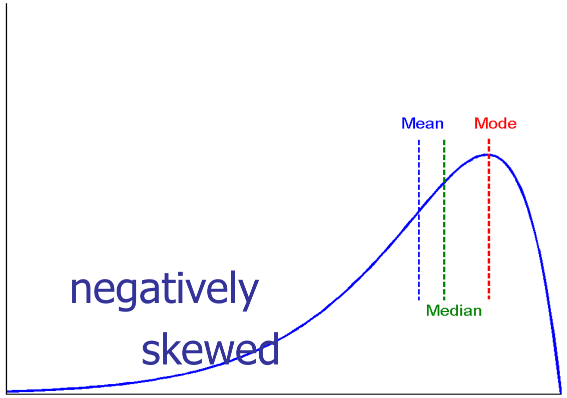
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

Symmetric vs. Skewed Data

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





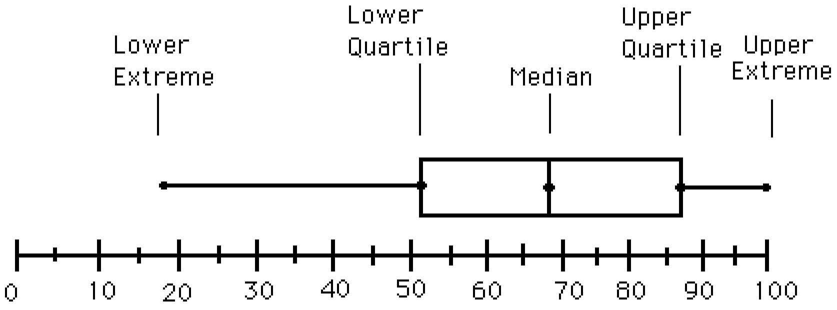


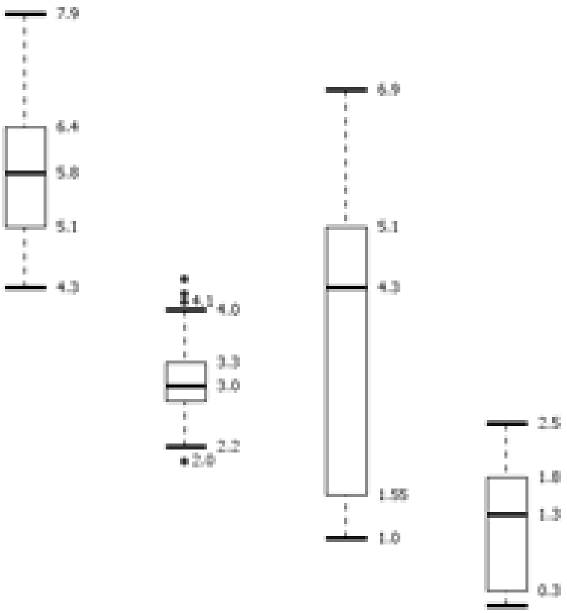
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 – Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than Q_3 + 1.5 x IQR or Q_1 – 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 (or σ^2)

Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

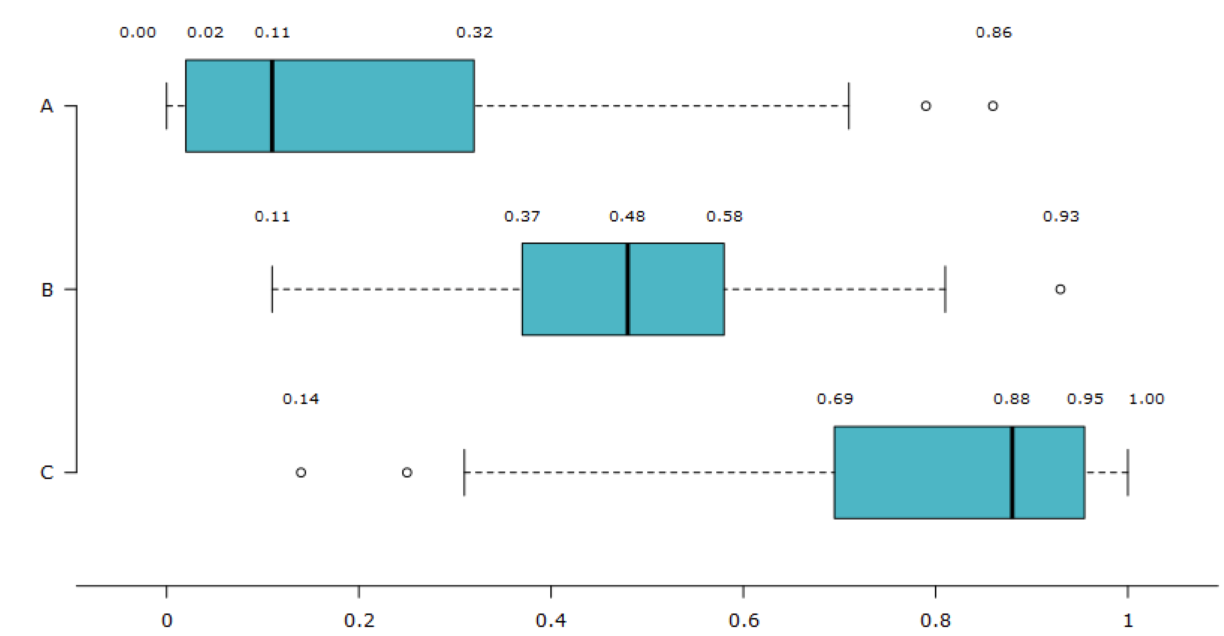




Boxplot Analysis Example

- Distribution A is positively skewed, because the whisker and half-box are longer on the right side of the median than on the left side.
- Distribution B is approximately symmetric, because both half-boxes are almost the same length (0.11 on the left side and 0.10 on the right side).
- Distribution C is negatively skewed because the whisker and half-box are longer on the left side of the median than on the right side.

Box and whisker plots and five-number summaries of distributions A, B and C



Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represent frequencies
- **Quantile plot**: a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute, then it plots quantile information
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Exercise

- Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   a. What attribute type is the variable *age*? Justify your answer.

   b. Use R code to display the *five-number summary* of the data.

   c. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.)

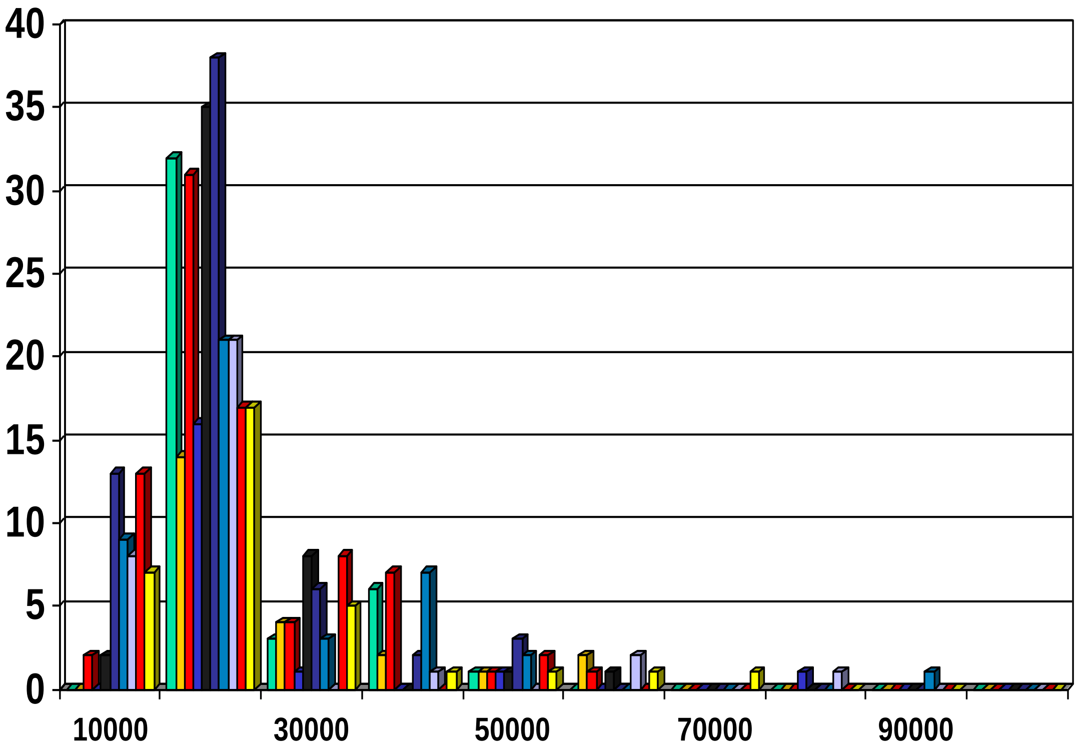
   d. Show a boxplot of the data. Comment on the skewedness of the data.

   e. Show the data in a histogram

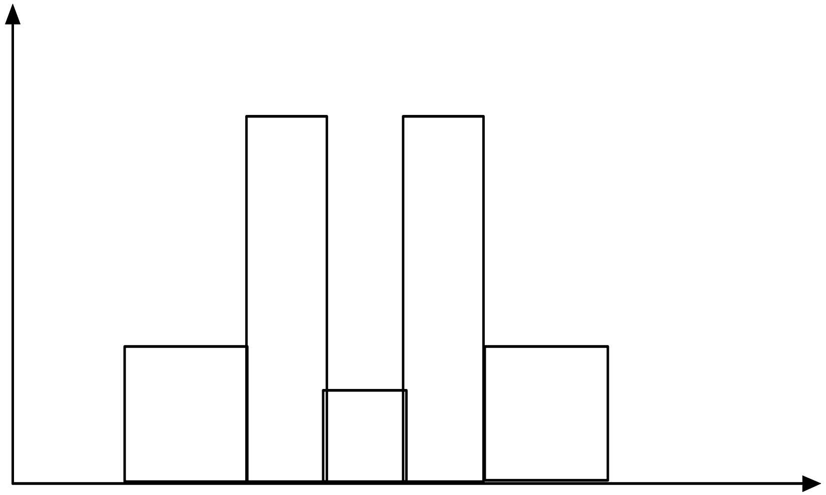
   f. How is a quantile–quantile plot different from a quantile plot?

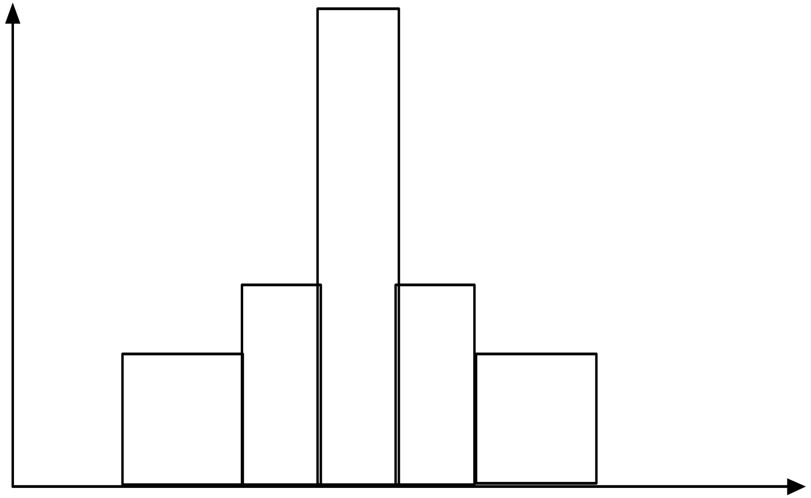
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

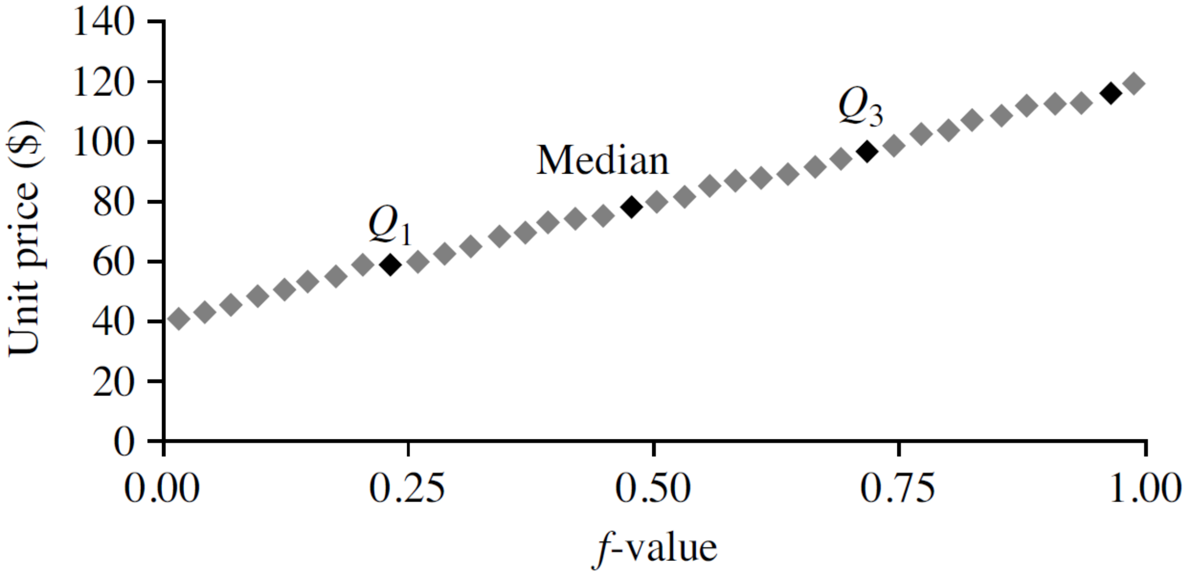




- The two histograms may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

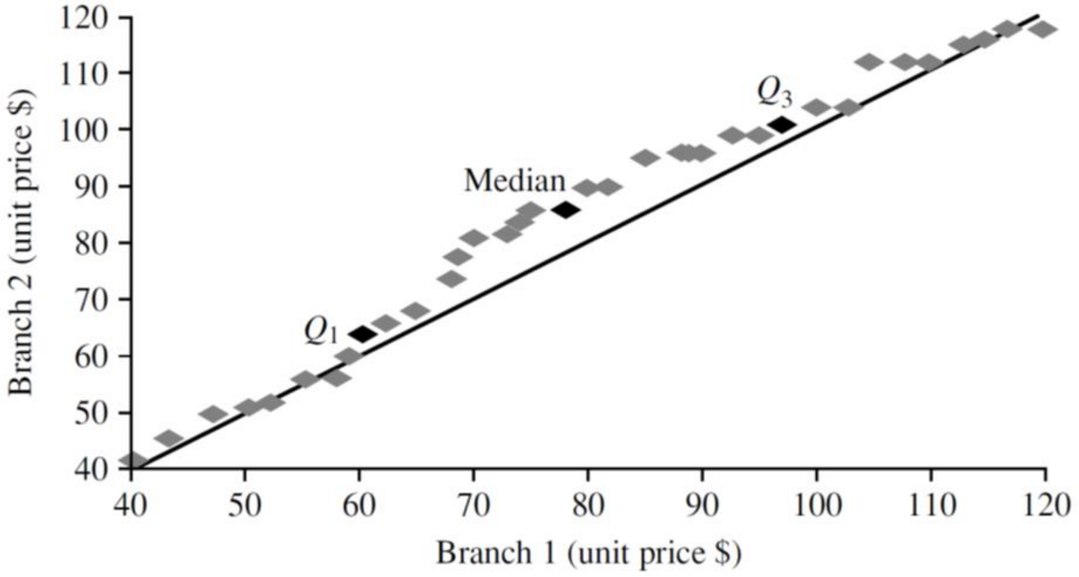
Quantile Plot

- First, it displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Second, it plots **quantile** information
  - For a data x_i data sorted in increasing order, each observation, x_i is paired with a percentage f_i which indicates that approximately f_i x 100% of the data are below or equal to the value x_i



$$f_i = \frac{i - 0.5}{N}$$

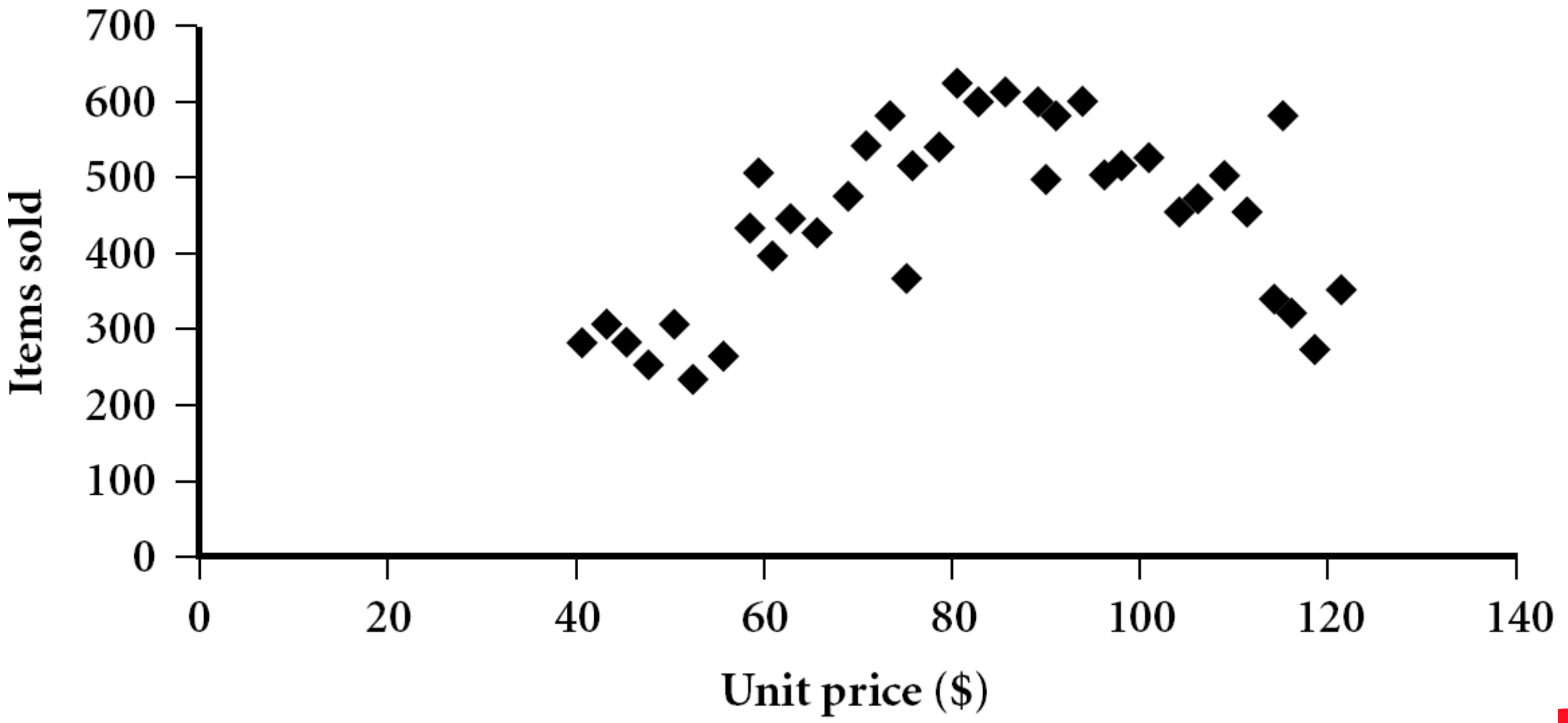
Quantile-Quantile (Q-Q) Plot



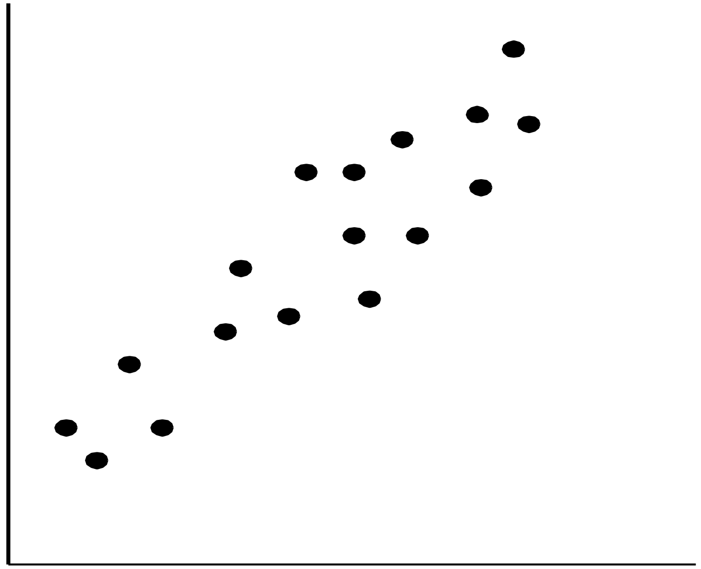
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

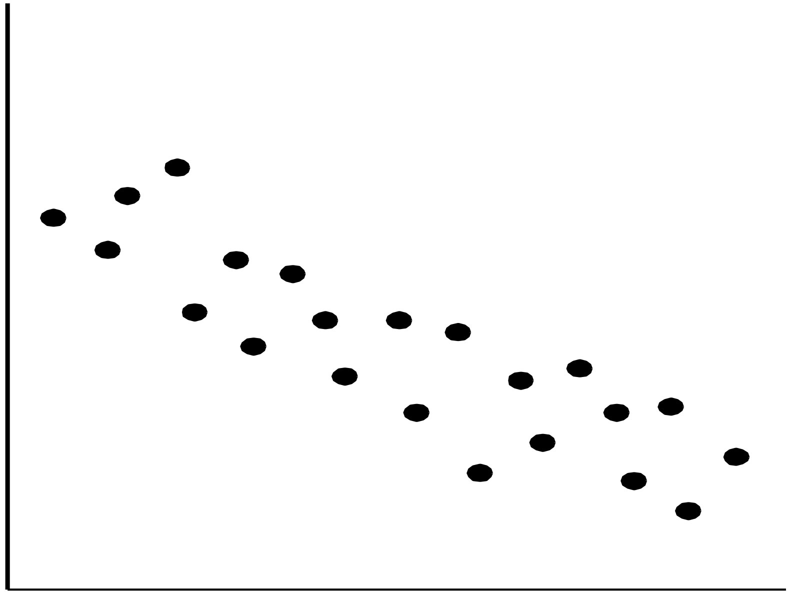
Scatter plot

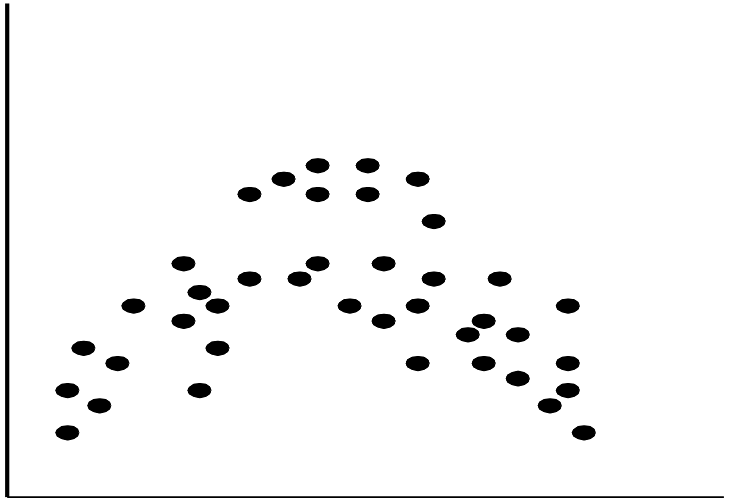
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane
- Two attributes, X, and Y, are correlated if one attribute implies the other.



Positively and Negatively Correlated Data

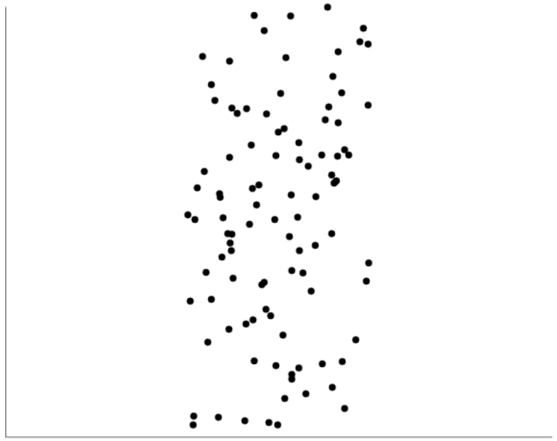


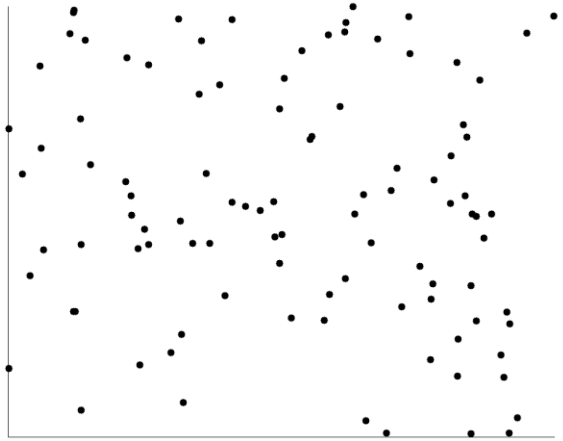


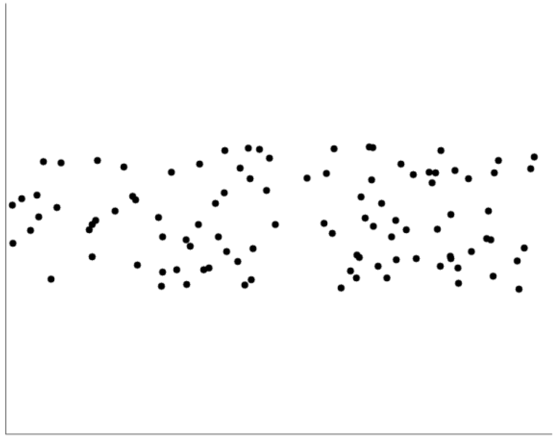


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data



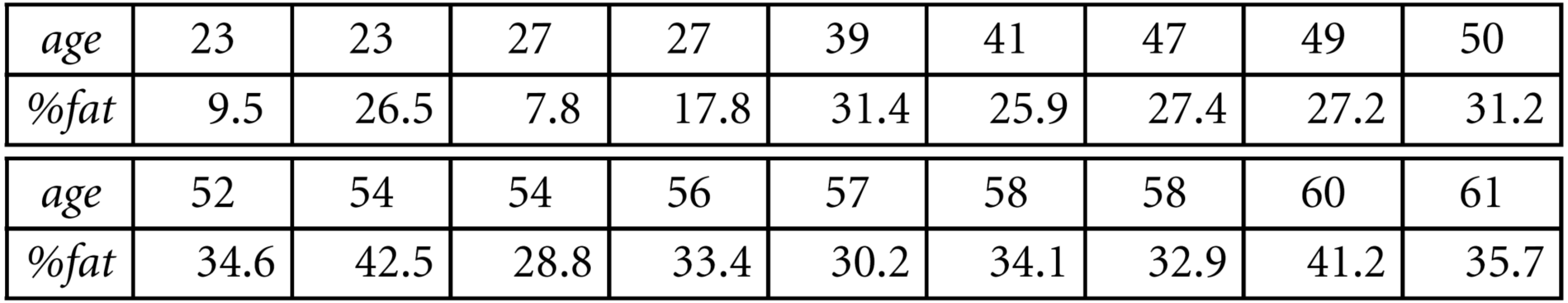




- Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

Exercise

- Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:



age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

- Use R code to calculate the mean, median, and standard deviation of *age* and *%fat*.
- Draw the boxplots for *age* and *%fat*.
- Draw a scatter plot and a q-q plot based on these two variables.

Measuring Data Similarity and Dissimilarity

- In data mining applications, such as clustering, outlier analysis, and nearest-neighbor classification, we need ways to assess how alike or unalike objects are in comparison to one another.
  - For example, a store may want to search for clusters of *customer* objects.
- A **cluster** is a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters.
- Potential **outliers** are objects that are highly dissimilar to others.
- Similarity and dissimilarity measures, are also referred to as *measures of proximity*.
- Similarity and dissimilarity are related.
- A similarity measure for two objects, *i* and *j*, will typically return the value 0 if the objects are unalike.
  - The higher the similarity value, the greater the similarity between objects. (Typically, a value of 1 indicates complete similarity, that is, the objects are identical.)
- A dissimilarity measure works the opposite way. It returns a value of 0 if the objects are the same (and therefore, far from being dissimilar).
  - The higher the dissimilarity value, the more dissimilar the two objects are.

Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - *object-by-attribute structure*
  - *n* data points with *p* dimensions
  - Two modes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - *object-by-object structure*
  - *n* data points, but registers only the distance
  - A triangular matrix
  - Single mode
  - *d(i, j)* is the measured dissimilarity or "difference" between objects *i* and *j*.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Distance Measures

- Distance measures are commonly used for computing the dissimilarity of objects described by numeric attributes.
  - Euclidean, Manhattan, and Minkowski distances
- In some cases, the data are *normalized* before applying distance calculations.
- This involves transforming the data to fall within a smaller or common range, such as [-1, 1] or [0.0, 1.0].
- Euclidean distance (i.e., straight line or "as the crow flies") - most popular distance measure

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

- Manhattan distance ("city block distance") - the distance in blocks between any two points in a city

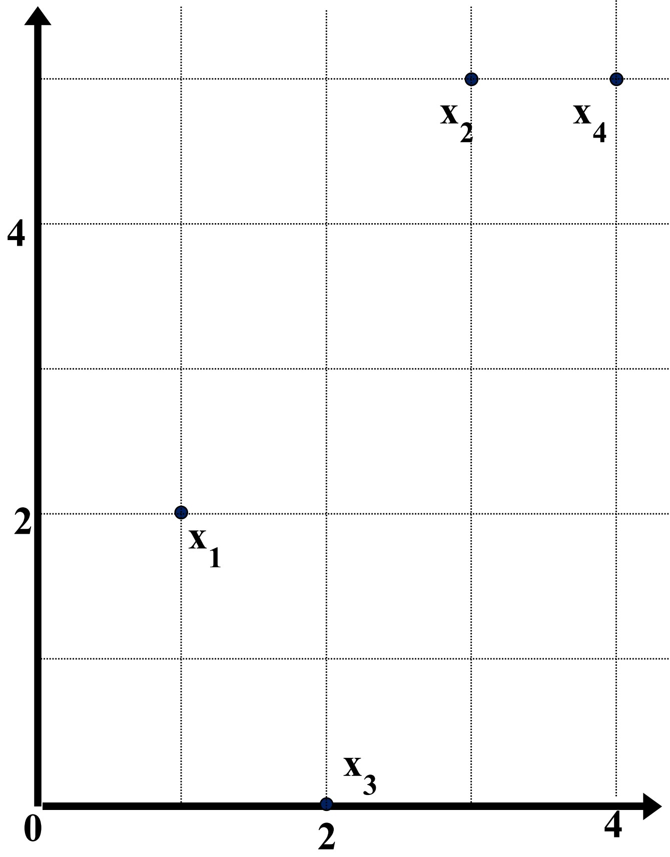
$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

- Minkowski distance - a generalization of the Euclidean and Manhattan distances.

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

- The supremum distance (also referred to as L_{max} ,L_{∞} **norm** or Chebyshev distance) is a generalization of the Minkowski distance

Example:
Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

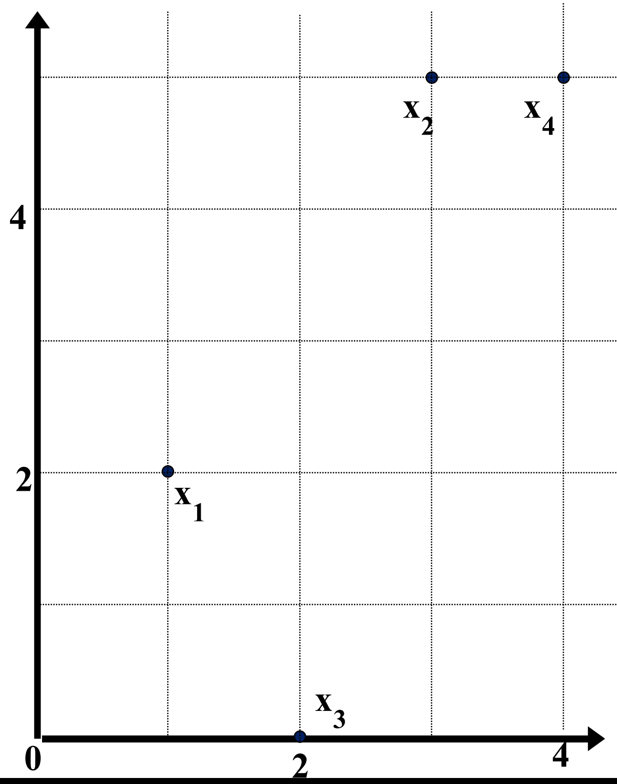
Dissimilarity Matrix

(with Euclidean Distance)

	x1	***x2***	***x3***	***x4***
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

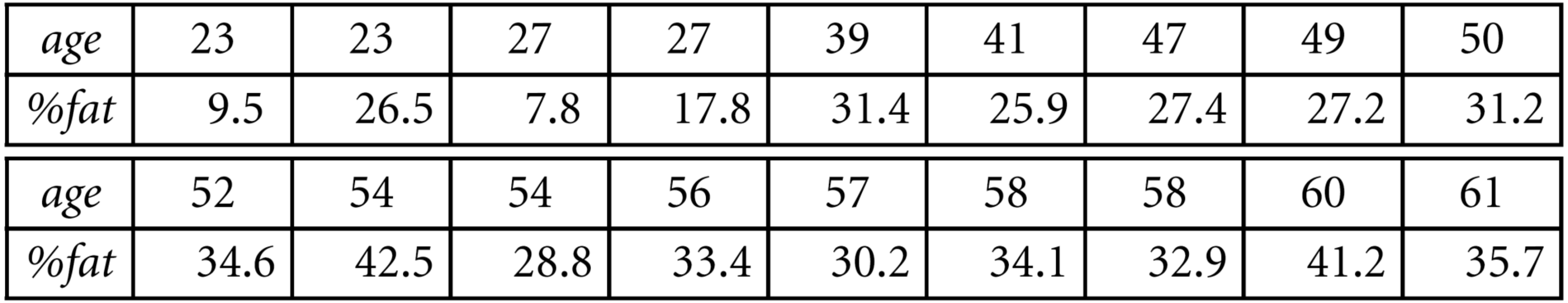
L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

Exercise

- Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:



age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

- Store this data in a data matrix.
- Compute all the pairwise dissimilarities between the observations in the data set using Euclidean metric and without standardization.
- Compare your initial solution with other distance measures (e.g. *Manhattan, and Minkowski*)

Case Study

- The `iris` dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 50 flowers from 3 different species. The `iris` dataset has been used for classification in many research publications.
- You are required to perform various analysis of the dataset to practice using sample datasets and get familiar with your data.

References

- Han, J., Kamber, M. and Pei, J., 2011. Data mining concepts and techniques 3rd edition. *The Morgan Kaufmann Series in Data Management Systems*.
- Andrea Cirillo (2017) *R Data Mining : Mine Valuable Insights From Your Data Using Popular Tools and Techniques in R*. Birmingham, UK: Packt Publishing. **(available on ebscohost)**
- Zhao, Y., 2012. *R and Data Mining: Examples and Case Studies*. Academic Press.
- Torgo, L., 2011. *Data Mining with R: Learning with Case Studies*. Chapman and Hall/CRC.
- Layton, R., 2017. *Learning Data Mining with Python*. Packt Publishing Ltd.
- Madhavan, S., 2015. *Mastering Python for Data Science*. Packt Publishing Ltd.
- Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Tan, P.N., Steinbach, M. and Kumar, V., 2016. *Introduction to data mining*. Pearson Education India.
- Weiss, S.M. and Indurkhya, N., 1998. *Predictive data mining: a practical guide*. Morgan Kaufmann.