Data Mining: Types of Data

Disclaimer: Content taken from Han & Kamber slides, Data mining textbooks and Internet

Data objects and attribute Type

- Data sets consists of Data object that represents entity
  - Customer, sales, products
  - Patients,
  - Students, professors, courses
- Data objects = instances, samples, examples, data points, or objects
- Data objects are described by attributes
- Attributes = features, dimensions, variable
- Observed values for a given attribute are known as observations
- A set of attributes used describing a given object is called attribute vector or feature vector

Data objects and attribute Type

- Nominal / categorical attributes
- Binary Attributes
  - Symmetric
  - asymmetric
- Ordinal Attributes
- Numeric Attributes
  - Interval-scaled Attributes
  - Ratio-scaled Attributes

Nominal attributes

- Relating to names
- Values are symbols or names of things
- Each value represents a category, code or state
- Also called as categorical attribute
- Order of values is not meaningful
- E.g.
  - Hair color = brown , black, white, red
  - Marital status = married, single, divorced
- Nominal values may by numeric
- E.g. – customerID
- Nominal attributes are not quantitative
- Cannot find mean, median
- Finding mode is possible – attributes most commonly occurring value

Binary attributes

- Nominal attribute with only two states: 0 or 1
- 0 – value absent, 1 – present
- Referred as Boolean if value – True or False
- E.g. – patient cancerous = 1, non-cancerous = 0
- Medical test positive = 1, negative = 0
- Symmetric binary attributes
  - Both states are equally valuable, both have same weight
  - E.g. gender – male, female
- Asymmetric binary attributes
  - Both states are not equally valuable
  - E.g. medical test result – positive, negative
  - Student attendance – present, absent

Ordinal attributes

- Attribute values have a meaningful order or ranking
- E.g. grade – A+, A, A-, B …
- Faculty ranks – professor, associate professor, assistant professor, adhoc professors …

Numeric attributes

- Quantitative
- Measurable quantity
- Values – integer or real
- Types – interval-scaled and ratio scaled
- Interval scaled
  - Measured on scale of equal-size units
  - Values have order – provides ranking
  - Can be positive, 0, negative
  - Attributes can be compared and quantify the difference
  - Mean, median and mode measures of central tendency
  - E.g. calendar dates – 10^{th} Jan and 15^{th} Jan are 5 days apart
  - Temperature – can be ranked
    - e.g. rank as per temperature coldest to hottest

Numeric attributes

- Ratio scaled
  - Numeric attribute with an inherent zero-point
  - If measurement is ratio-scaled, one value can be a multiple of another
  - Values are ordered
  - Mean, mode and median can be calculated
  - E.g. – years_of_experience
  - Number_of_words
  - Height, weight, latitude

Numeric attributes

- True Zero-Point
  - Temperature in Celsius and Fahrenheit
  - 0°C and 0°F doesn't indicate "no temperature"
  - Difference can be computed but one temperature value is not spoken as a multiple of another
  - Without a true zero we cannot say 10°C is twice as warm as 5°C
  - Similarly for calendar dates. The year 0 does not correspond to the beginning of time.
- Ratio-scaled attributes – true zero-point exists.

Discrete and continuous attributes

- Discrete – finite number of values
- May or may not be integers
  - E.g.- hair_color, medical_test
- Countably infinite – values can grow infinite but still countable
  - customerID, zipcode
- Continuous – numeric attributes,
- Real values
  - E.g.- age, salary etc

Measure of central tendency

- Mean
- Median
- Mode
- Midrange

Measure of central tendency

- Mean
  - $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$
  - Weighted mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Median
  - Even no. of observation = middlemost two values
  - Odd no. = middle observation
  - For large observations, group the observations in intervals and frequency of each interval
  - $median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$

  Where

  L1 – lower boundary of the median interval,

  N – no. of values

  freql – sum of frequency of intervals lower than the median interval,

  freqmedian – frequency of median interval

  Width – width of median interval

Measure of central tendency

- Mode
  - Measure of central tendency
  - Mode = value that occurs more frequently
  - Can be determined by qualitative and quantitative
  - If greatest frequency corresponds to more than one value then there will be more than one mode
    - Unimodal – datasets with one mode
    - Bimodal – two modes
    - Trimodal – three modes
    - Multimodal – two or more modes
  - No mode – when each value occurs only once

- Midrange
  - Measure central tendency of numeric data set
  - Average of the largest and smallest values in the set

Examples

- Given:

60, 61, 61, 61, 62, 62, 63, 63, 63, 63, 63, 64, 64, 64, 64, 64, 64, 64, 65, 65, 66

Mean = 1326/21 = 63.1

Median = 63

Mode = 64

Examples

Age Groups	Frequency
0 - 10	40
10 - 20	53
20 - 30	58
30 - 40	64
40 - 50	72
50 - 60	49
60 - 70	36
70 - 80	25

- Grouped mean:

$$Groupedmean = \frac{\sum(f_i \times x_{im})}{n}$$

- Where :
  - n = total no. of observations
  - f_i = frequency of i^{th} observation
  - x_{im} = midpoint of i^{th} x

Grouped mean

Age Groups	Frequency	Xm	Fi * xm
0 - 10	40	(9+ 0)/2 = 4.5	40 * 4.5 = 180
10 - 20	53	(19+10)/2 = 14.5	768.5
20 - 30	58	24.5	1421
30 - 40	64	34.5	2208
40 - 50	72	44.5	3204
50 - 60	49	54.5	2670.5
60 - 70	36	64.5	2322
70 - 80	25	74.5	1862.5
Total	**397**		**14636.5**

Grouped mean = 14636.5/397 = 36.9
Mean is somewhere between 30 - 40

Grouped median

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

Where

L1 – lower boundary of the median interval,

N – no. of values

$freq_l$ – sum of frequency of intervals lower than the median interval,

$freq_{median}$ – frequency of median interval

Width – width of median interval

- Median class is the class where the middle point of the total frequency lies.
- i.e. 397/2 = 198.5 which lies in 30-40
- Hence the median class or median interval is 30-40

Age Groups	Frequency	Summation
0 - 10	40	40
10 - 20	53	40+53 = 93
20 - 30	58	40+53+58 = 151
30 - 40	64	215
40 - 50	72	287
50 - 60	49	336
60 - 70	36	372
70 - 80	25	397

Element	Value
L1 – lower boundary of the median interval,	30
N – no. of values	397
freql – sum of frequency of intervals lower than the median interval,	151
freqmedian – frequency of median interval	64
Width – width of median interval	10

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

median =

Grouped mode

$$groupedmode = L + \left(\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}\right) * width$$

Where

L – The lower limit of the group with the mode (the group with the highest frequency)

f_m – Frequency of the group with the mode

f_{m-1} – Frequency of the group before the one with the mode

f_{m+1} – Frequency of the group after the one with the mode

Width – width of the groups

Element	Value
L – The lower limit of the group with the mode	40
f_m – Frequency of the group with the mode	72
f_{m-1} – Frequency of the group before the one with the mode	64
f_{m+1} – Frequency of the group after the one with the mode	49
Width – width of groups	10

Age Groups	Frequency
0 - 10	40
10 - 20	53
20 - 30	58
30 - 40	64
40 - 50	72
50 - 60	49
60 - 70	36
70 - 80	25

Skewed distribution

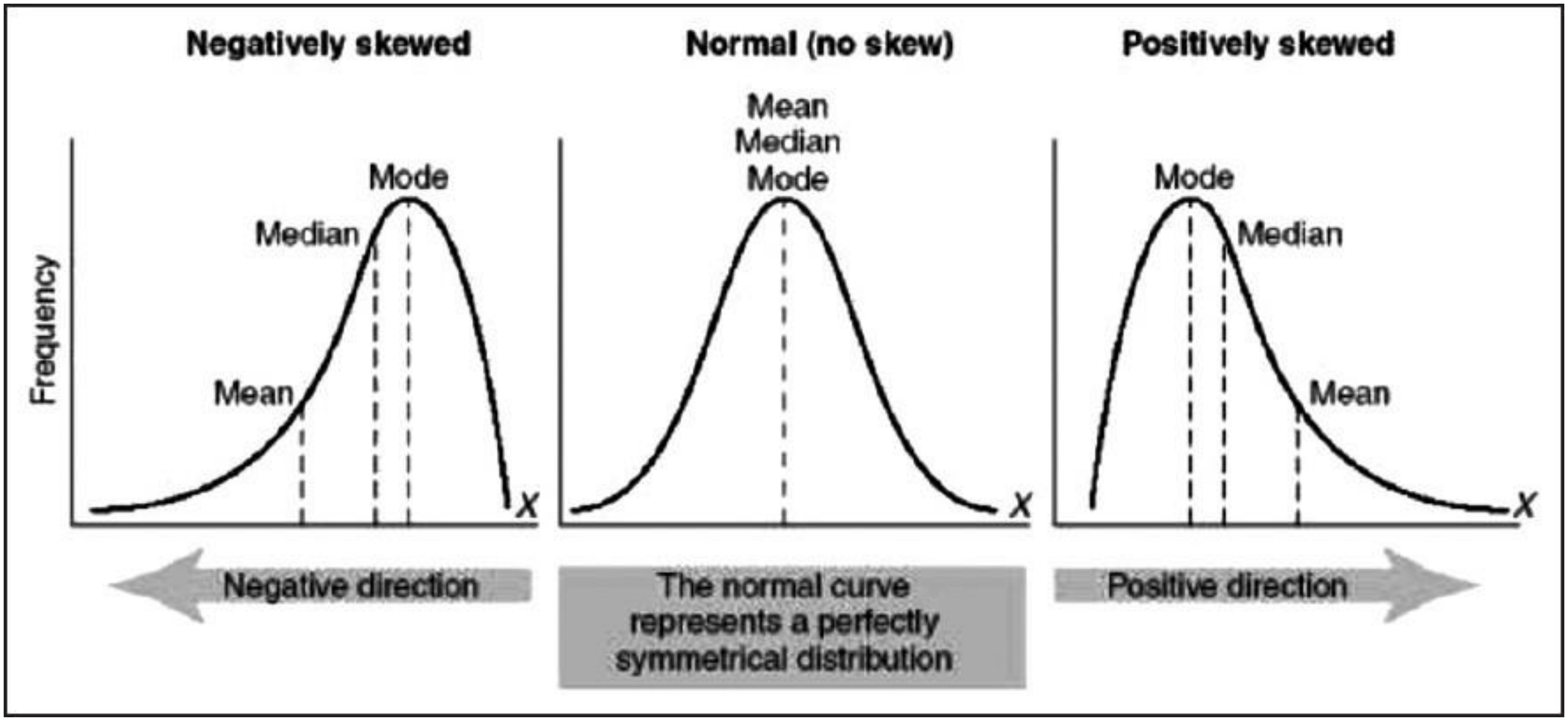
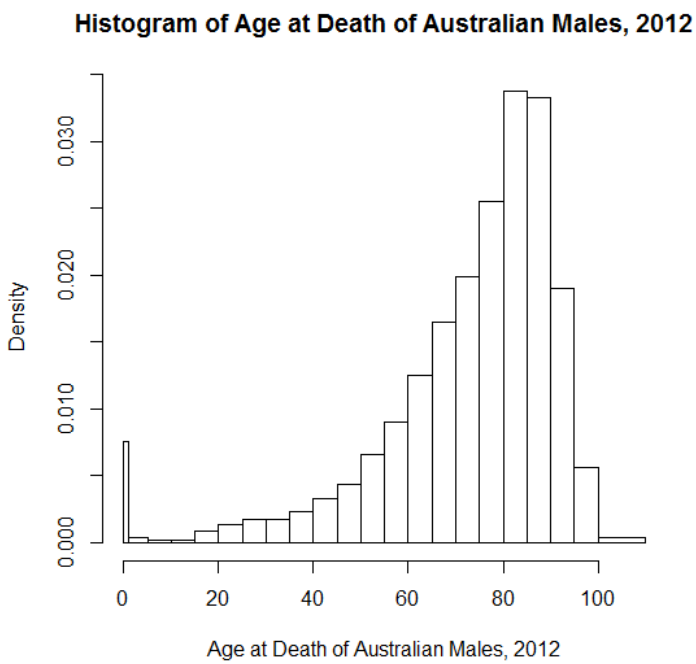


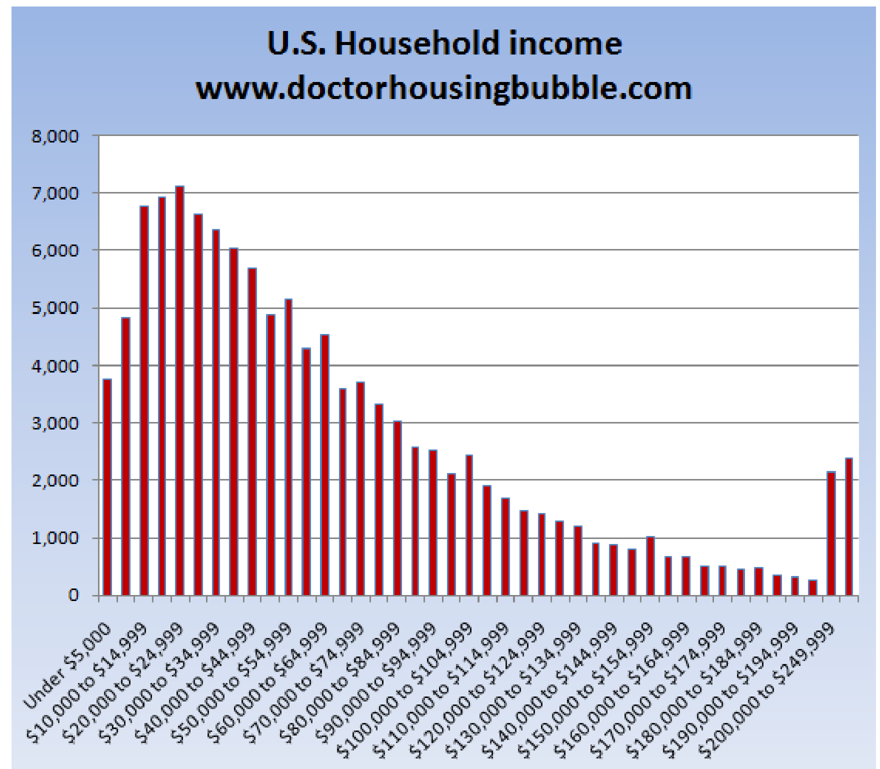
Image Source: Internet

Skewed distribution

- A left-skewed distribution has a long left tail.
  - Left-skewed distributions are also called *negatively-skewed* distributions.
  - That’s because there is a long tail in the negative direction on the number line.
  - The mean is also to the left of the peak.
  - Mean is less than median.
- A right-skewed distribution has a long right tail.
  - Right-skewed distributions are also called positive-skew distributions.
  - That’s because there is a long tail in the positive direction on the number line.
  - The mean is also to the right of the peak.
  - Mean is greater than the median.

Skewed distribution



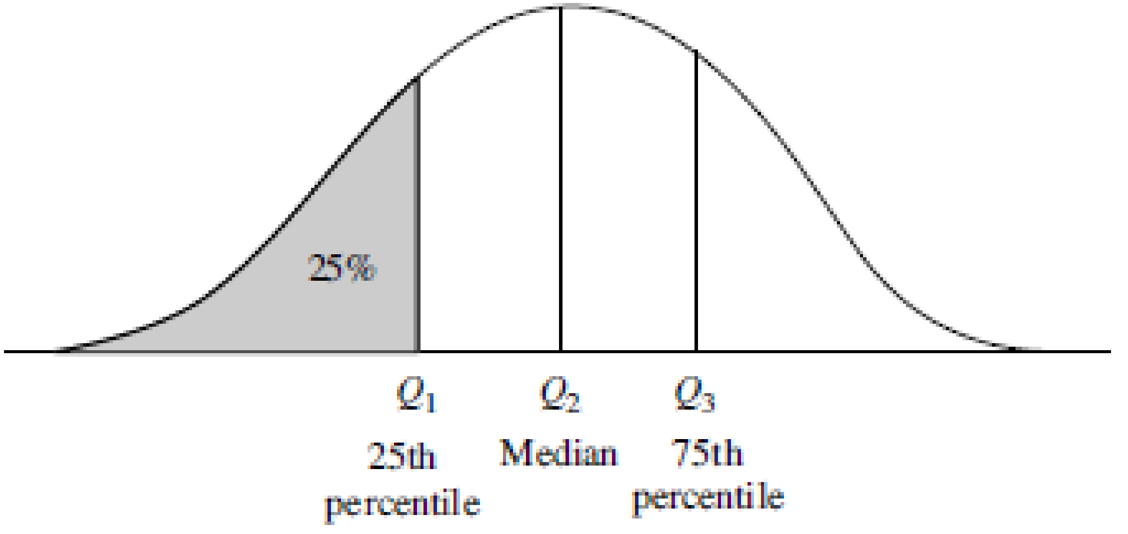


Measuring the Dispersion of Data

- Range, Quartiles and Inter quartile range
  - Range: difference between the largest max() and smallest min() values
  - Quantiles: are points taken at regular intervals of a data distribution dividing it into essentially equal size consecutive sets
  - 2-quantiles: data point dividing lower and upper halves of data distribution
  - 4-quantiles: 3 data points that split data distribution into four equal parts; each part represents one-fourth of data distribution are called **quartiles**

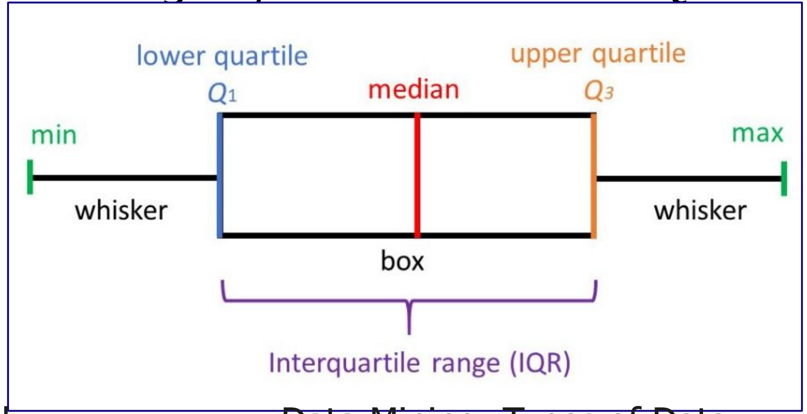
Measuring the Dispersion of Data

- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
- Percentiles: 100-quantiles divide data distribution into 100 equal sized consecutive sets
- Inter-quartile range: $IQR = Q_3 - Q_1$



Measuring the Dispersion of Data

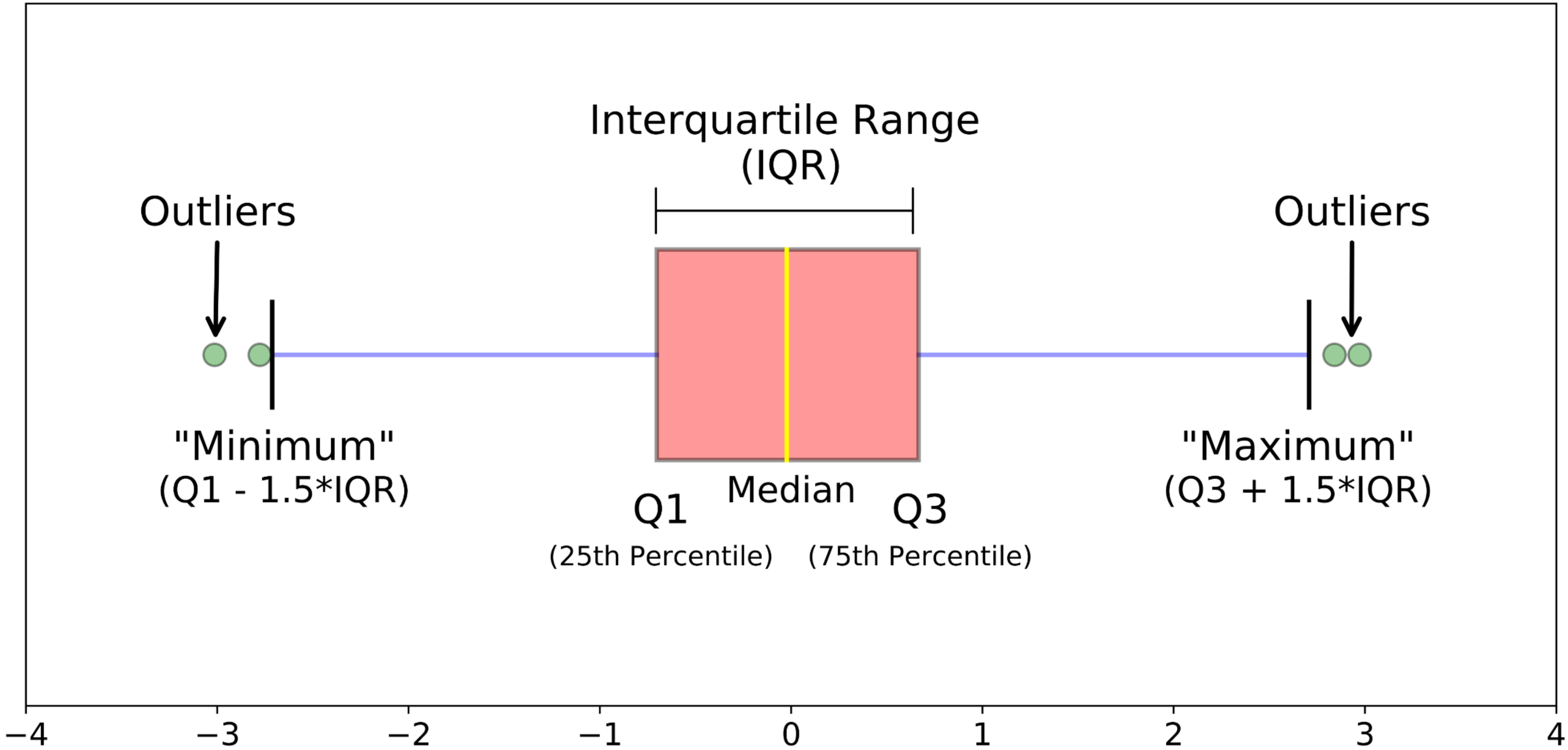
- Five number summary: min, Q_1, M, Q_3, max
- Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles,
    i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
- Outlier: usually, a value higher/lower than 1.5 x IQR



Plotting outliers

- When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually.
- To do this in a boxplot, the whiskers are extended to the extreme low and high observations *only if* these values are less than 1.5IQR beyond the quartiles.
- Otherwise, the whiskers terminate at the most extreme observations occurring within 1.5IQR of the quartiles.
- The remaining cases are plotted individually.

Box plot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Median
Q1
(25th Percentile)
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

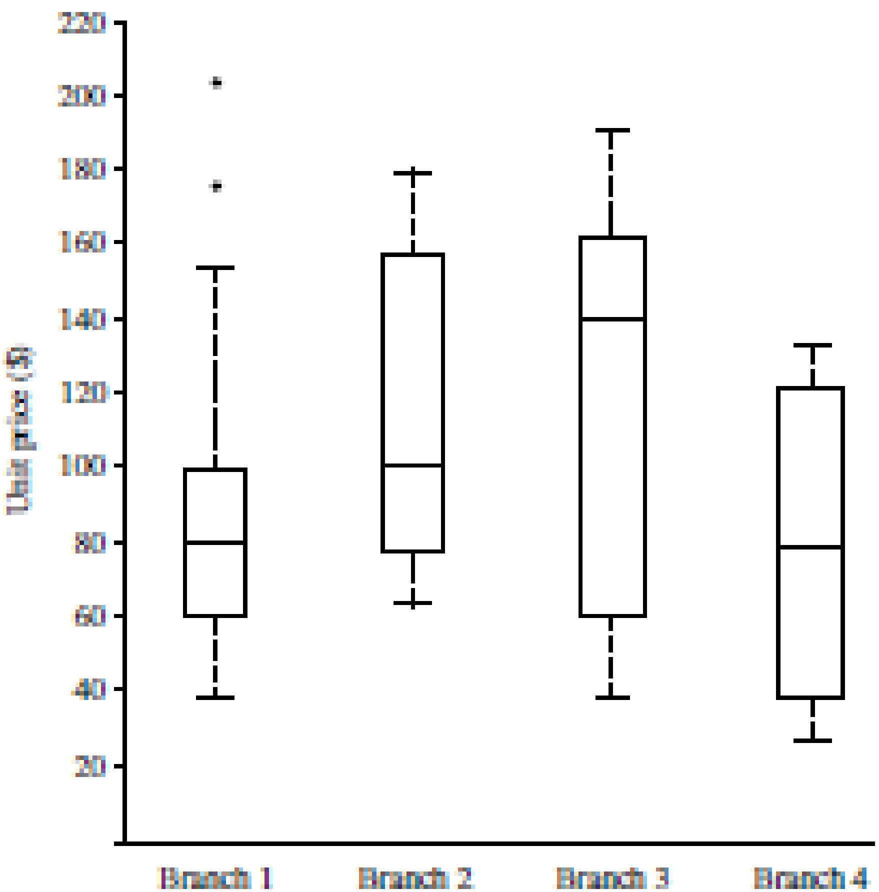


Figure shows boxplots for unit price data for items sold at four branches of *AllElectronics* during a given time period.

For branch 1, we see that the median price of items sold is $80, $Q1$ is $60, and $Q3$ is $100.

Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

Boxplot Analysis

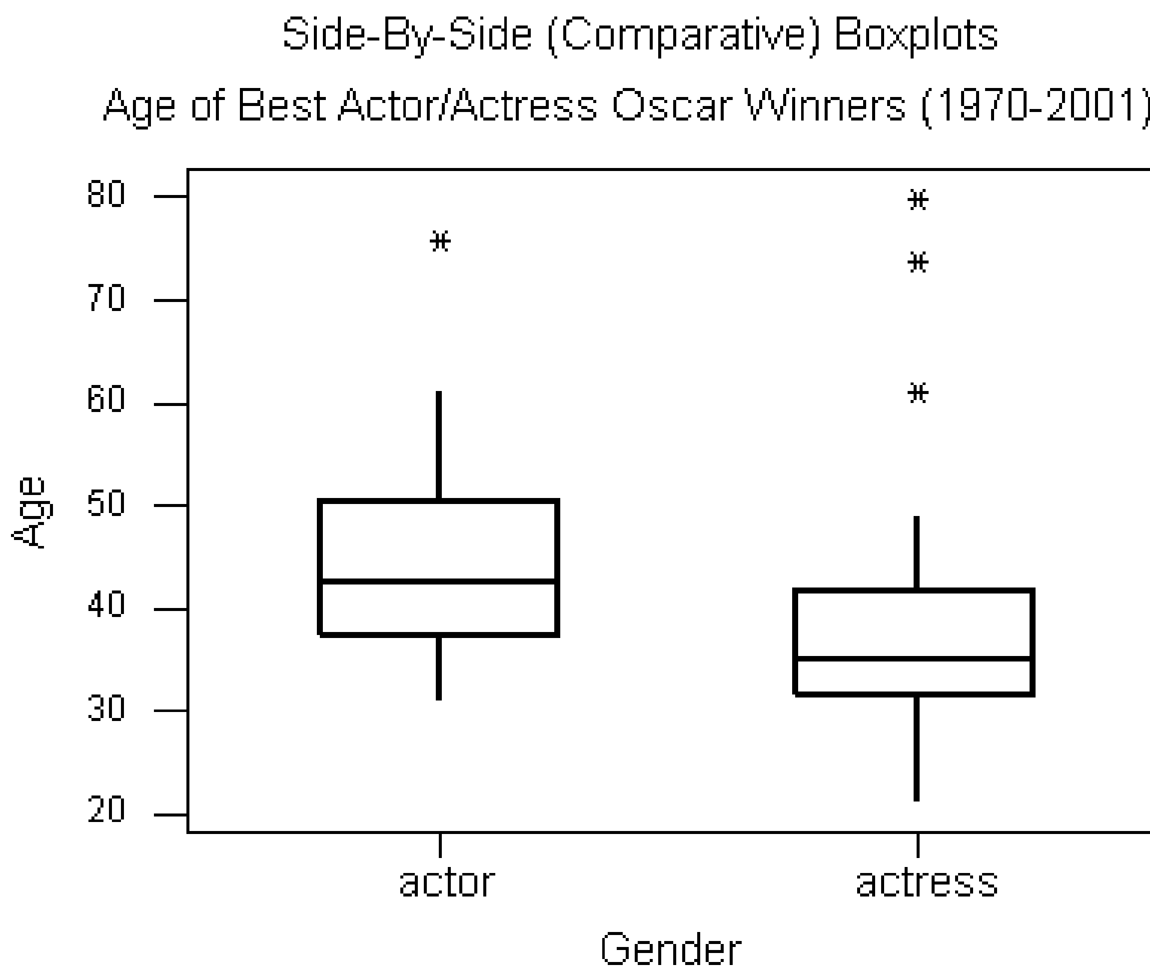


Image source: Internet

Boxplot Analysis

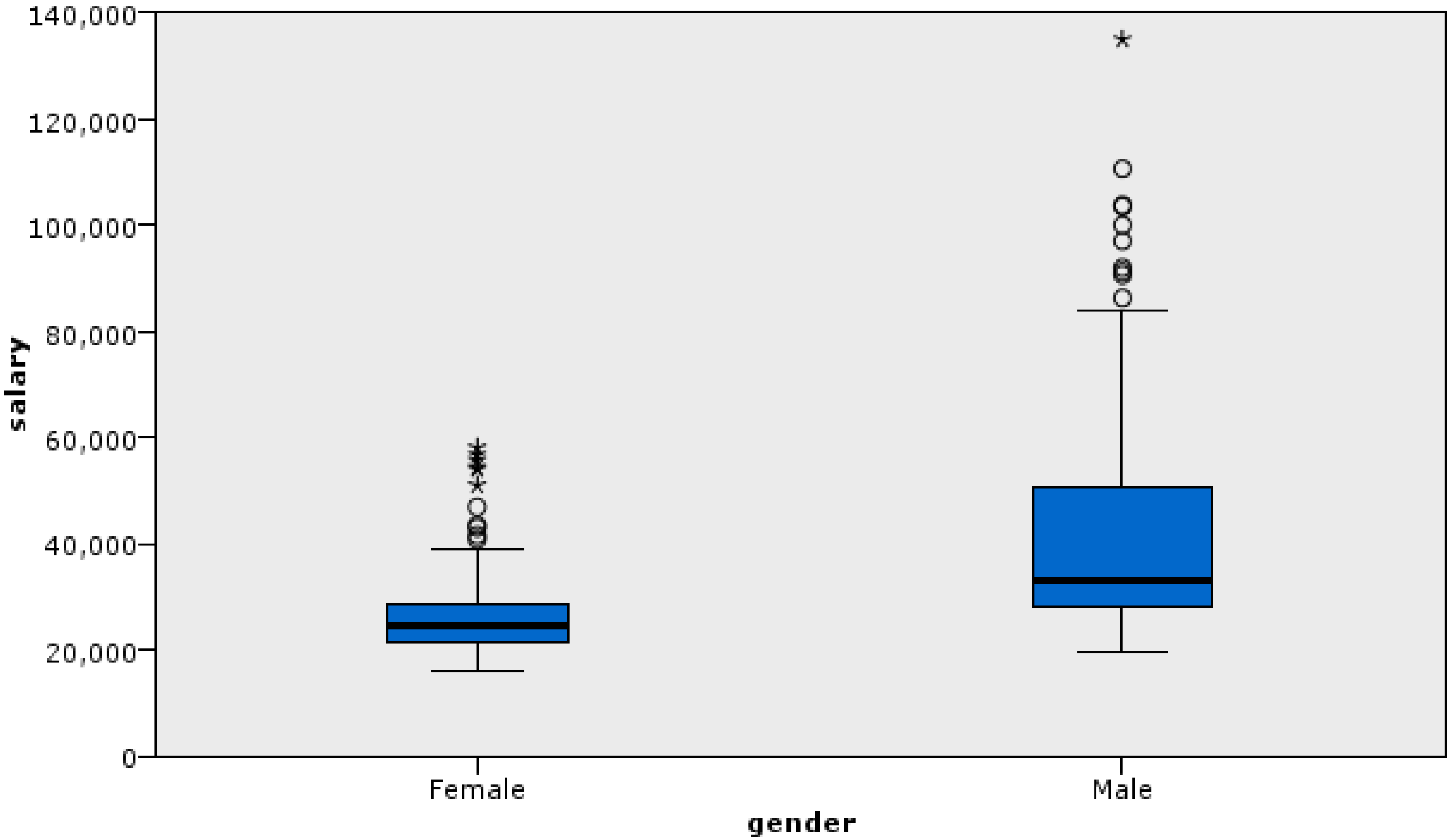
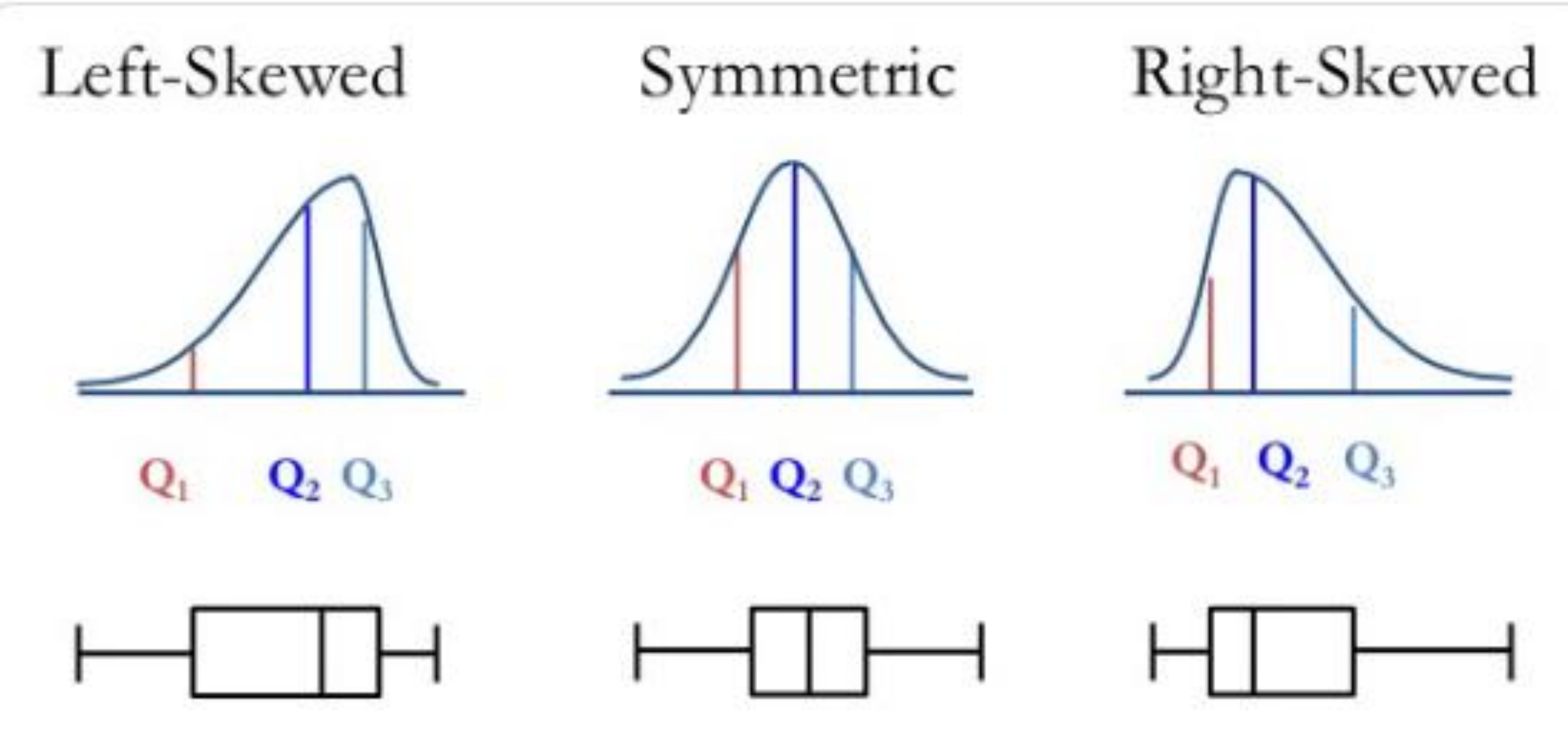


Image source:
https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/graphboard_creating_examples_boxplot.htm

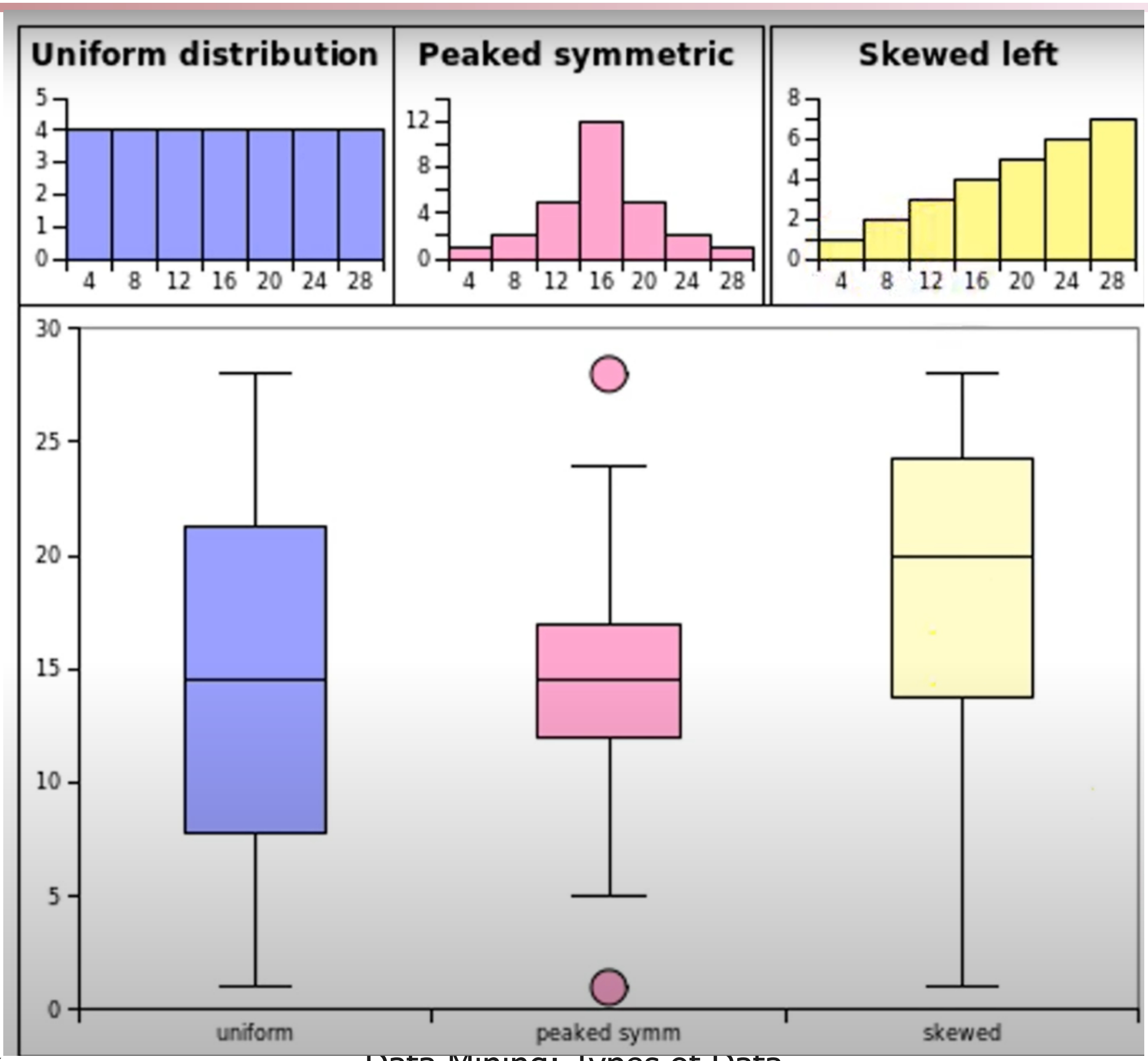
Box Plot analysis

- Step 1: Compare the medians of box plots
- Step 2: Compare the interquartile ranges and whiskers of box plots
- Step 3: Look for potential outliers
- Step 4: Look for signs of skewness

Box plot and skewness



Box plot and distribution



Let's explore the different parts of the boxplot:

- The dark line in the middle of the boxes is the median of *salary*. Half of the cases/rows have a value greater than the median, and half have a value lower. Like the mean, the median is a measure of central tendency. Unlike the mean, it is less influenced by cases/rows with extreme values. In this example, the median is lower than the mean (compare to Example: Bar Chart with a Summary Statistic). The difference between the mean and median indicates that there are a few cases/rows with extreme values that are elevating the mean. That is, there are a few employees who earn large salaries.
- The bottom of the box indicates the 25th percentile. Twenty-five percent of cases/rows have values below the 25th percentile. The top of the box represents the 75th percentile. Twenty-five percent of cases/rows have values above the 75th percentile. This means that 50% of the case/rows lie within the box. The box is much shorter for females than for males. This is one clue that *salary* varies less for females than for males. The top and bottom of the box are often called **hinges**.
- The T-bars that extend from the boxes are called **inner fences** or **whiskers**. These extend to 1.5 times the height of the box or, if no case/row has a value in that range, to the minimum or maximum values. If the data are distributed normally, approximately 95% or the data are expected to lie between the inner fences. In this example, the inner fences extend less for females compared to males, another indication that *salary* varies less for females than for males.
- The points are **outliers**. These are defined as values that do not fall in the inner fences. Outliers are extreme values. The asterisks or stars are **extreme outliers**. These represent cases/rows that have values more than three times the height of the boxes. There are several outliers for both females and males. Remember that the mean is greater than the median. The greater mean is caused by these outliers.

Measuring the Dispersion of Data

- Variance and standard deviation (sample: s, population: σ)
- Measures of data dispersion
- Indicate how spread out a data distribution is
  - Variance: (algebraic, scalable computation)
  - The **variance** of *N* observations, $x1, x2, \ldots, xN$, for a numeric attribute *X* is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\ 2} - \mu^2$$

  - Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^{\ 2} - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

    - Low standard deviation means data observations tend to be close to the mean
    - High standard deviation means data are spread out over a large range of values

Interval-valued variables

- Standardize data
  - Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- The mean absolute deviation is more robust to outliers than the standard deviation

Graphical Displays of Basic Statistical Descriptions of Data

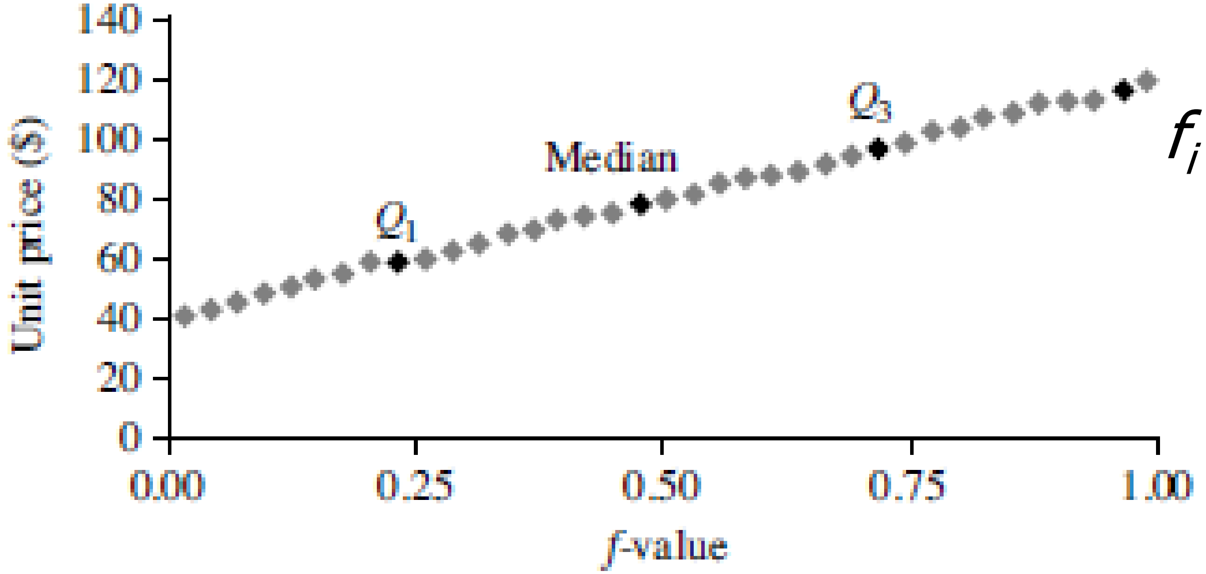
- Quantile plots – univariate distributions
- Quantile-Quantile plots – univariate distributions
- Histograms – univariate distributions
- Scatter plots – bivariate distributions

- Helpful for the visual inspection of data.
- Used in data pre-processing

Quantile Plot

- Simple and effective way to have a first look at a univariate data distribution
- Displays all of the data for a given attribute (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately f_i x 100 % of the data are below or equal to the value x_i

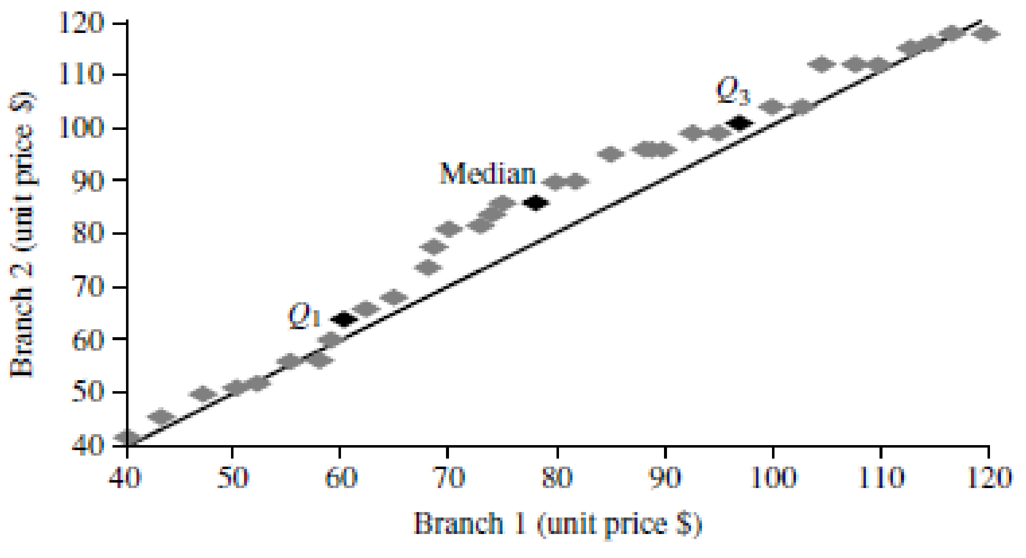
Q1: 25% of total items have unit price <60$ approx



$$f_i = (i - 0.5) / N$$

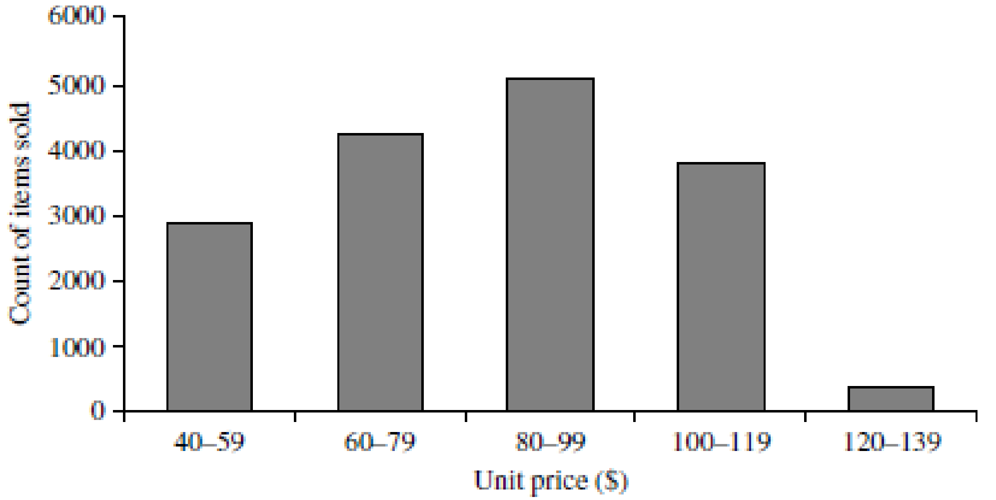
Quantile-Quantile (Q-Q) Plot

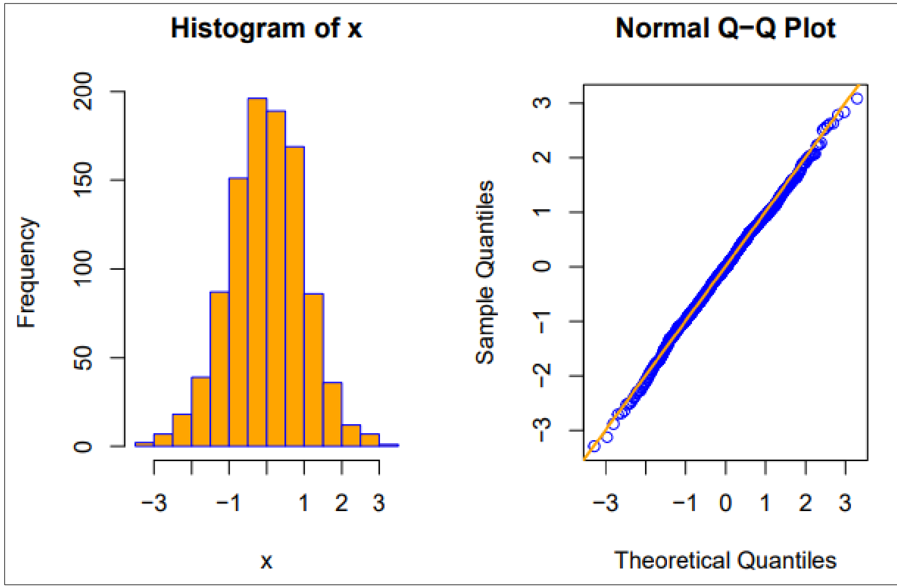
- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another

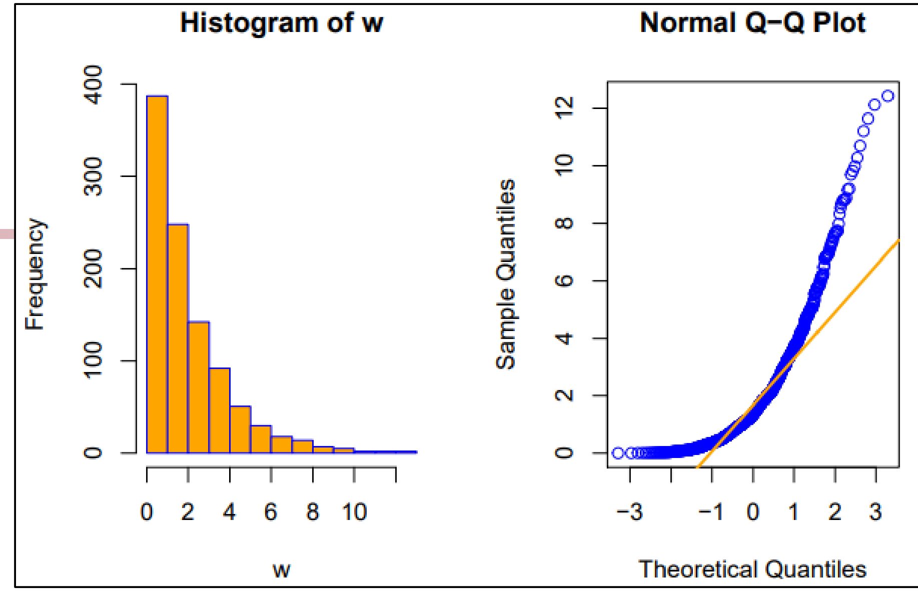


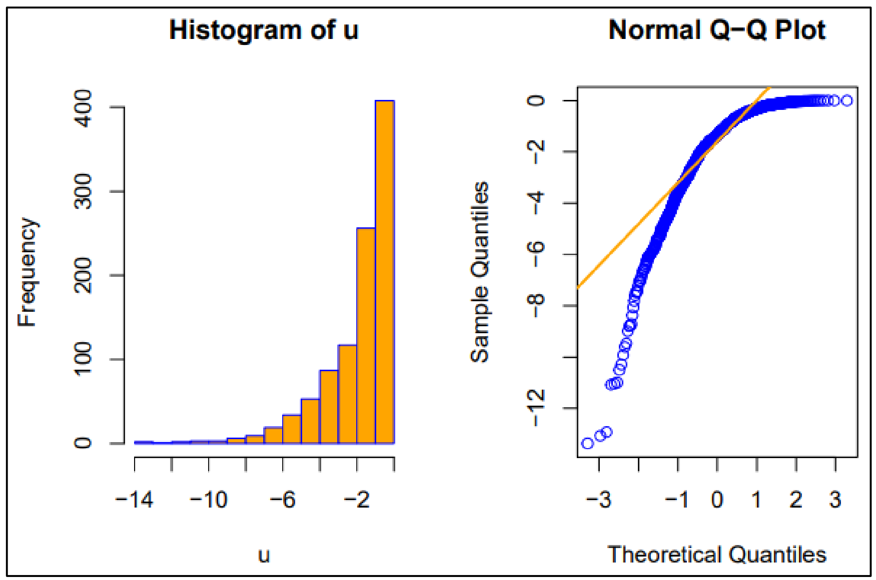
Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data





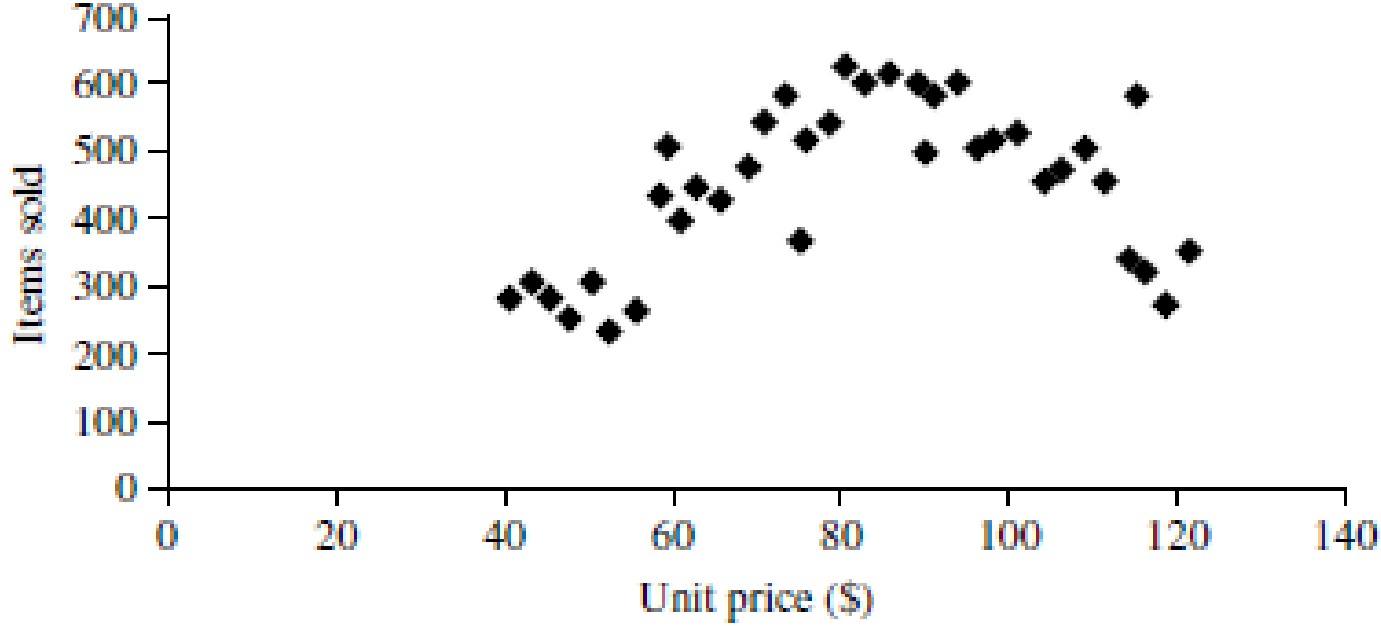




Source: https://math.illinois.edu/system/files/inline-files/Proj9AY1516-report2.pdf

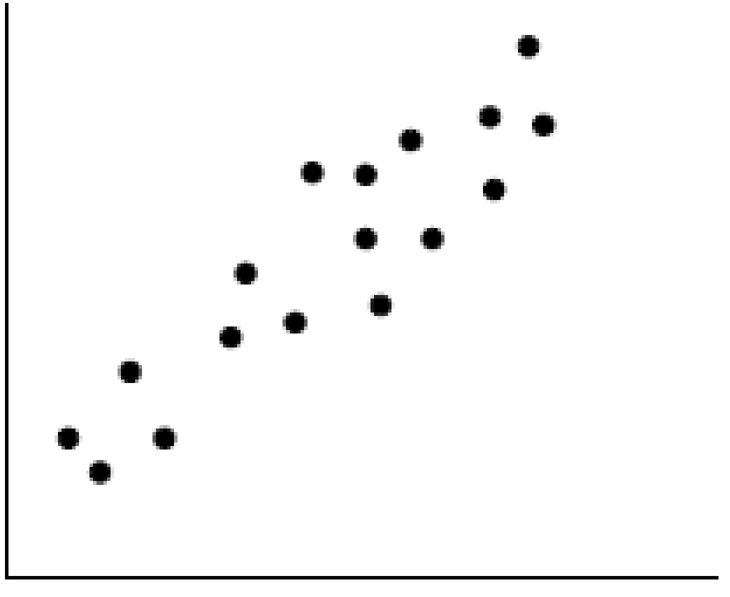
Scatter plot & Data Correlation

- 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

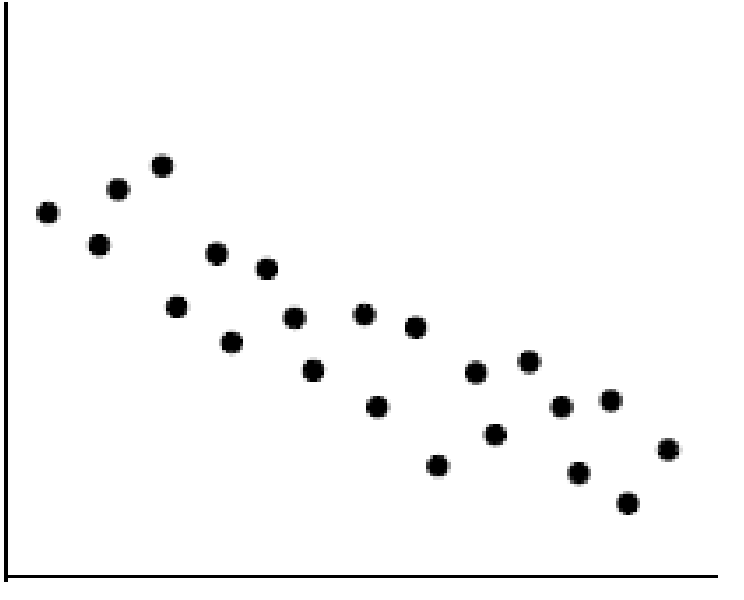


Scatter plot & Data Correlation

- Two attributes are correlated if one attribute implies the other
- Negative, positive or null (uncorrelated)

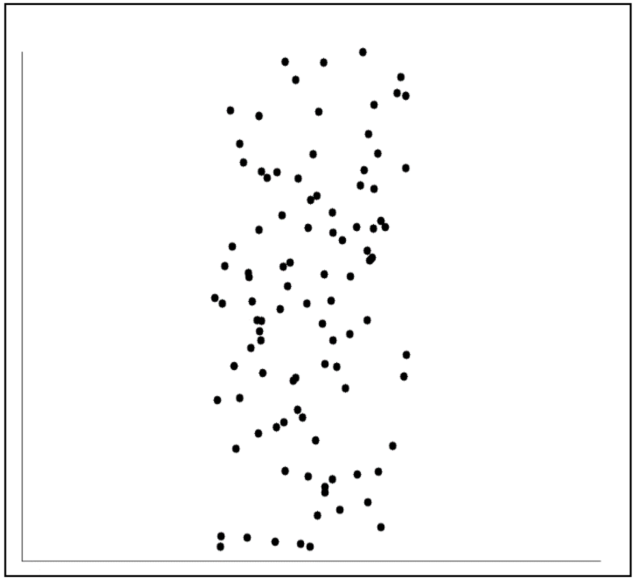


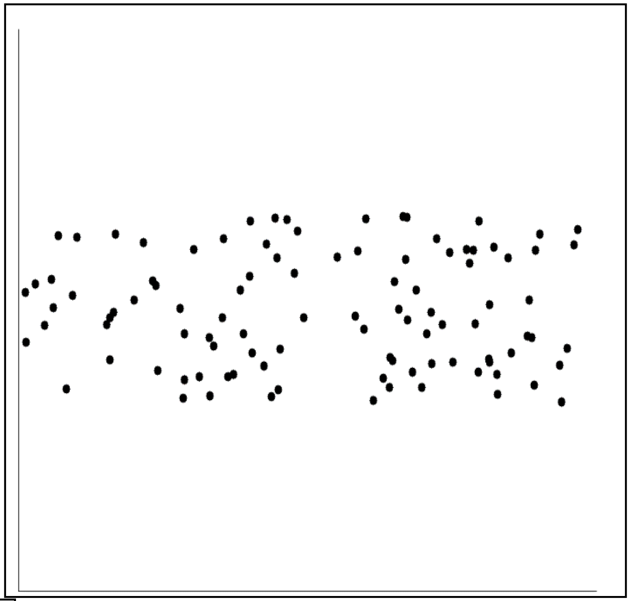
(a)

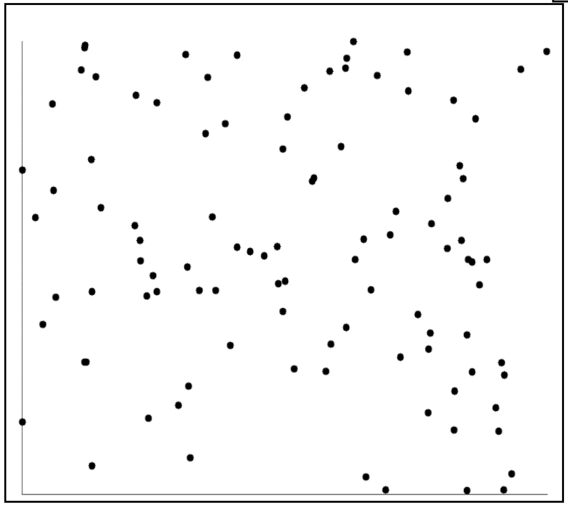


(b)

Not Correlated Data

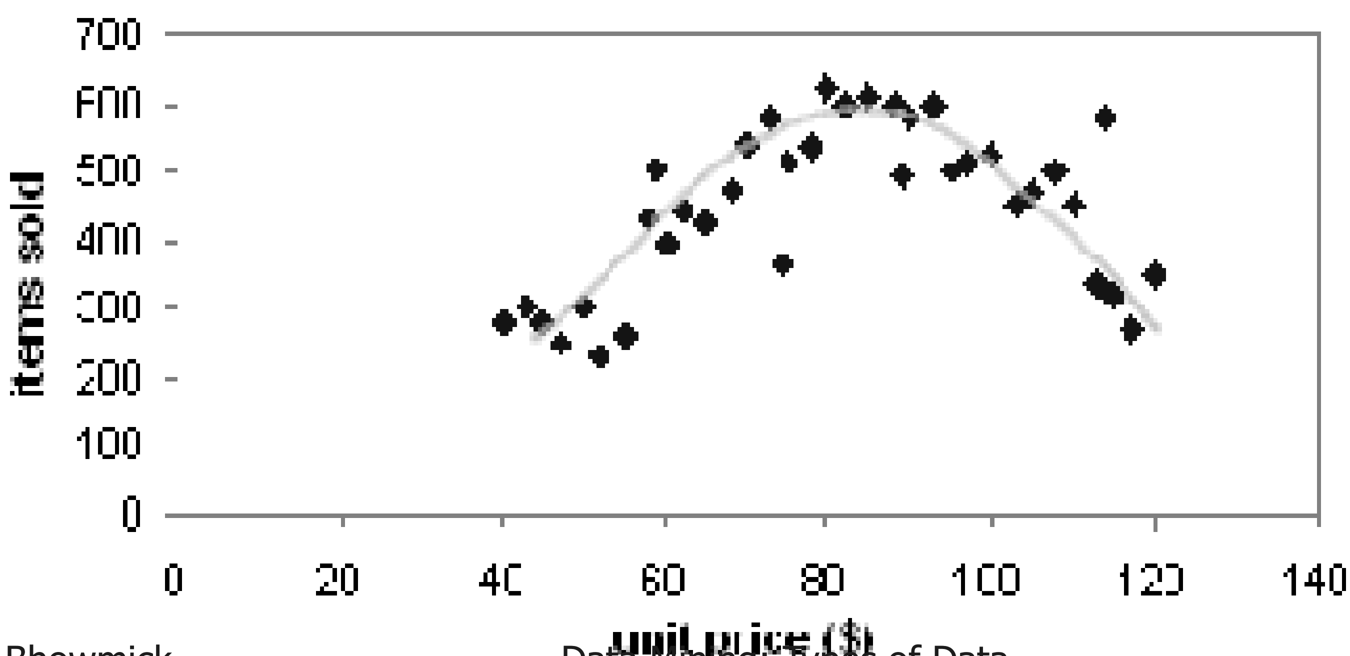






Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



Measuring data similarity and dissimilarity

- Need to assess how alike or unalike objects are in comparison to one another.
- E.g. clustering, outlier analysis and nearest neighbour classification
- Similarity and dissimilarity measures are used
- A similarity measure for two objects return the value 1 if objects are like and 0 if they are unalike.
- A dissimilarity measure works the opposite way.

Data Structures

- Data matrix (object-by-attribute structure)
  - n objects and p-variables
  - two modes

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

- Dissimilarity matrix (object-by-object structure)
  - Contains dissimilarities
  - one mode

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

Proximity measures for Nominal or Categorical Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching

$$d(i,j) = \frac{p-m}{p}$$

  - *m*: # of matches i.e. the number of attributes for which i and j are in the same state,
  - *p*: total # of variables
- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states

Example

Ob ID	Test-1	Test-2	Test-3
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1210

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

d (i, j) = 0 if objects *i* and *j* match, and
d (i, j) = 1 if the objects differ

Proximity measure for Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

  - a = no. of attributes that equal to 1 for both objects
  - b = no. of attributes that equal to 1 for object i and 0 for object j
  - c = no. of attributes that equal to 1 for object j and 0 for object i
  - d = no. of attributes that equal to 0 for both objects

Proximity measure for Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- Distance measure for asymmetric binary variables :

$$d(i,j) = \frac{b+c}{a+b+c}$$

Proximity measure for Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$$

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - gender is a symmetric attribute
  - the remaining attributes are asymmetric binary
  - let the values Y and P be set to 1, and the value N be set to 0

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$\boxed{d(i,j) = \frac{b+c}{a+b+c}}$$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$\mathbf{d(jack, mary) = \frac{0+1}{2+0+1} = 0.33}$$

$$\mathbf{d(jack, jim) = \frac{1+1}{1+1+1} = 0.67}$$

$$\mathbf{d(jim, mary) = \frac{1+2}{1+1+2} = 0.75}$$

Dissimilarity of Numeric Data

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Dissimilarity of Numeric Data

- *If q = 2, d* is Euclidean distance:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

  - Properties
    - *d(i,j)* $\geq$ 0; distance is non-negative
    - *d(i,i)* = 0; distance of object to itself is 0
    - *d(i,j) = d(j,i); distance is symmetric*
    - *d(i,j)* $\leq$ *d(i,k)* + *d(k,j); triangle inequality*

- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

$$d(i,j)=\sqrt{w_1|x_{i_1}-x_{j_1}|^2+w_2|x_{i_2}-x_{j_2}|^2+...+w_p|x_{i_p}-x_{j_p}|^2)}$$

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Example

Ob ID	Test-1	Test-2	Test-3
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1210

Ob ID	Rank	z_{if}
1	3	1
2	1	0
3	2	0.5
4	3	1

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Ratio-Scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}
- Methods:
  - treat them like interval-scaled variables—*not a good choice!* (why?—the scale can be distorted)
  - apply logarithmic transformation

$$y_{if} = log(x_{if})$$

  - treat them as continuous ordinal data treat their rank as interval-scaled

Example

Ob ID	Test-1	Test-2	Test-3
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1210

Ob ID	Log(x_{if})
1	2.65
2	1.34
3	2.21
4	3.08

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$

Numeric

Ob ID	Test-1	Test-2	Test-3
1	Code-A	Excellent	45
2	Code-B	Fair	22
3	Code-C	Good	64
4	Code-A	Excellent	28

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Max = 64, min = 22,

$$d(2,1) = \frac{(22-45)}{(64-22)} = 0.55$$

Variables of Mixed Types

- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - $\delta_{ij}^{(f)} = 0$; 1. x_{if} or x_{jf} is missing
    2. $x_{if} = x_{jf} = 0$ and f asymmetric binary
  - $\delta_{ij}^{(f)} = 1$
  - $d_{ij}^{(f)}$ dependent on its type

Example

Ob ID	Test-1	Test-2	Test-3
1	Code-A	Excellent	45
2	Code-B	Fair	22
3	Code-C	Good	64
4	Code-A	Excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

$$\frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65$$

Cosine similarity

- Measure of similarity that can be used to compare documents or give ranking of documents w.r.t a given vector of query words.

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||},$$

where $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$,

Defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$.

Cosine similarity

Document Vector or Term-Frequency Vector

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Suppose that ***x*** and ***y*** are the first two term-frequency vectors

x = (5, 0, 3, 0, 2, 0, 0, 2, 0,0)

y = (3, 0, 2, 0, 1, 1, 0, 1, 0,1)

$$\mathbf{x}^t \cdot \mathbf{y} = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

$$||\mathbf{x}|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||\mathbf{y}|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$sim(\mathbf{x}, \mathbf{y}) = 0.94$$

Vector Objects

- Vector objects: keywords in documents, gene features in micro-arrays, etc.
- Broad applications: information retrieval, biologic taxonomy, etc.
- Cosine measure $$s(\vec{X},\vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{|\vec{X}||\vec{Y}|},$$

$\vec{X}^t$ is a transposition of vector $\vec{X}$, $|\vec{X}|$ is the Euclidean normal of vector $\vec{X}$,

- A variant: Tanimoto coefficient

$$s(\vec{X},\vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$

Self Study

- Visualization