Getting to Know Your Data – Descriptive Study

Knowledge about Data

Knowledge about your data (Descriptive analysis) is very useful for data pre-processing

- What are the types of attributes or fields that make up your data?
- What kind of values does each attribute have?
- Which attributes are discrete, and which are continuous-valued?
- What do the data look like?
- How are the values distributed?
- Are there ways we can visualize the data to get a better sense of it all?
- Can we spot any outliers? Can we measure the similarity of some data objects with respect to others?

All these types of information about data helps a lot in subsequent analysis

Data Objects and Attributes

Data sets are made up of Data Objects

- **A data object represents an entity / Class / Concept**
  - In a tabular view of data, a *row* stores information for a data object, for example
    - a customer
    - a sales Record
  - Data Objects are often referred to as:
    - Samples (Stats)
    - Examples
    - Instances (Weka)
    - Data points
    - Objects (programming, object-oriented design)
- **An attribute is a data field, representing a characteristic of feature of a data object**
  - In a tabular view of data, a column value describes a particular feature of the data object.
  - Attributes are often referred to as:
    - Dimensions (data warehouse)
    - Features (data mining)
    - Variables (stats)
    - Attributes (programming, object-oriented design, data mining)
  - The distribution of data involving
    - One attribute – univariate
    - Two attributes – bivariate
    - Multiple attributes – multi-variate
  - The set of attributes describing a type of objects is called
    - Attribute vectors or feature vectors.

Type of an Attribute

- There are four measurement scales (or types of data):

Nominal,

Ordinal,

Interval and

Ratio

Qualitative Attributes

Quantitative Attributes

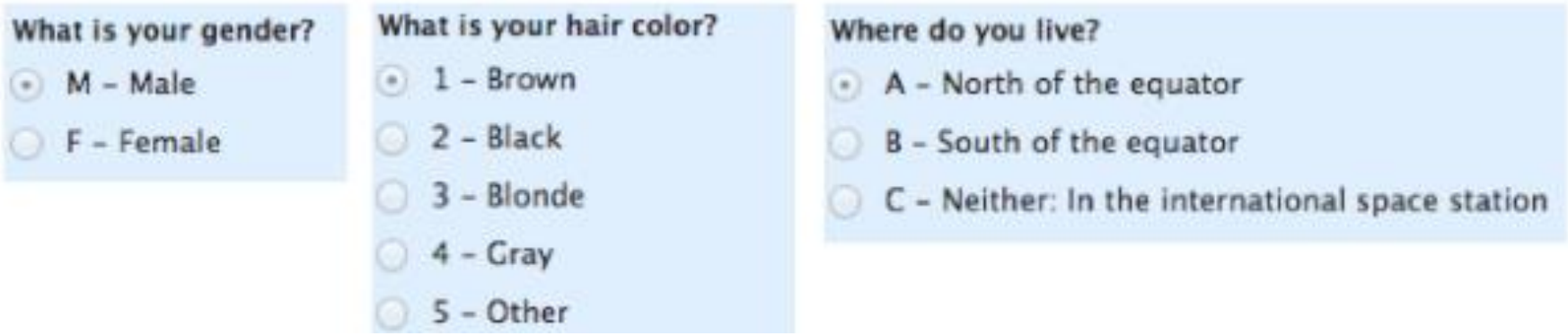
Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Nominal

- Nominal scales are used for labelling variables, without any quantitative value.
- “Nominal” scales could simply be called “labels.”
- Note that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance.
- –--Ex. ID numbers, Eye Colour, Zip Codes

Nominal Attributes

- Here are some detail examples.



Nominal Attributes

- Nominal scales are the lowest levels of measurement.
- We can use numbers to represent labels within a category, but the number does not have quantities of a true number--just a category label.

Democrat = 1	Freshman = 1	Team 1
Republican = 2	Sophomore = 2	Team 2
Independent = 3	Junior = 3	Team 3
	Senior = 4	Team 4

Binary

Binary attributes are Nominal attribute with only 2 states or labels (0 and 1)

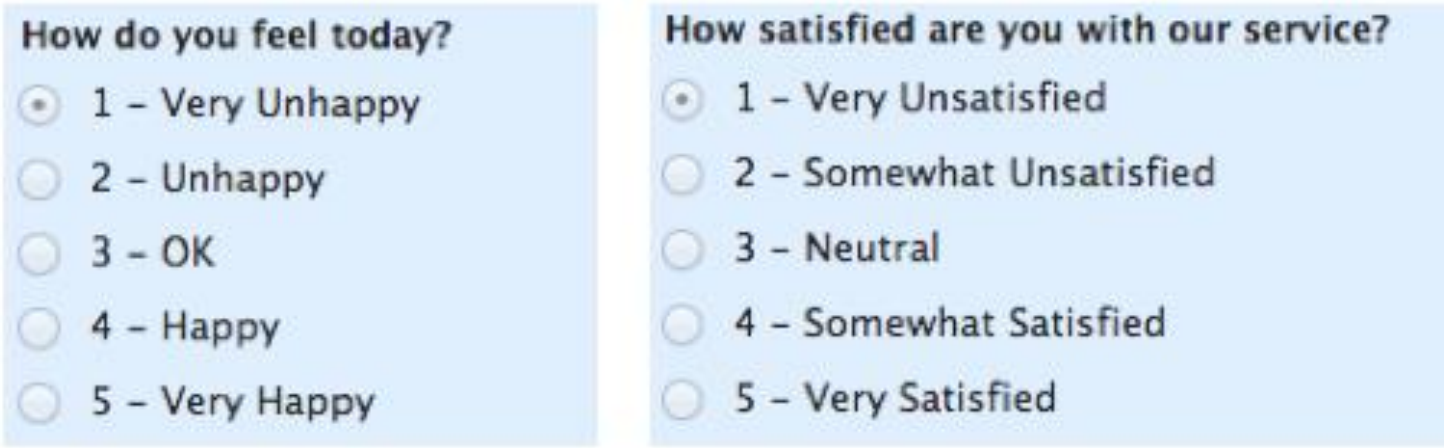
- **Symmetric binary**: both states or labels are equally valuable and carry the same weight
  - e.g., gender
- **Asymmetric binary**: both states or labels are not equally important.
  - e.g., medical test (positive or negative)
  - e.g., HIV positive (yes or no)
  - Convention: assign 1 or Y to most important state or label

Ordinal

- With ordinal scales, the *order* of the values is important and significant, but the differences between each one is not really known.
- Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

Ordinal

- Example of Ordinal Scales



Ordinal scales tell us relative order, but give us no information regarding differences between the categories.

is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.

Interval

- Interval scales are numeric scales in which we know both the order and the exact differences between the values.
- The classic example of an interval scale is **Celsius temperature** because the difference between each value is the same.

For example, the difference between **60** and **50** degrees is a measurable **10** degrees, as is the same difference between **80** and **70** degrees.

Interval

- “Interval” itself means “space in between,” which is the important thing to remember– interval scales not only tell us about order, but also about the value between each item.
- In Interval scale, the increments are known, consistent, and measurable.

Interval

- Here's the problem with interval scales: they don't have a "**true zero**."
- For example, there is no such thing as "no temperature," with **Celsius**. In the case of interval scales, zero doesn't mean the absence of value, but is actually another number used on the scale, like 0 degrees Celsius.
- Without a true zero, it is impossible to compute ratios. With interval data, we can **add and subtract, but cannot multiply or divide**.

Interval

- consider this: 10 degrees C + 10 degrees C = 20 degrees C. No problem there.

  But, 20 degrees C is not twice as hot as 10 degrees C, however, because there is no such thing as “no temperature” when it comes to the Celsius scale.

  When converted to Fahrenheit, it’s clear: 10C=50F and 20C=68F, which is clearly not twice as hot.

  With Interval scales, we cannot calculate ratios

A person with an IQ score of 160 is not twice smarter than a person with an IQ score 80.

Ratio

- Ratio scales are the ultimate measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied.
- Everything about interval data applies to ratio scales, plus ratio scales have a clear definition of zero. Good examples of ratio variables include **height and weight**.

Ratio

- Ratio scales provide a wealth of possibilities when it comes to statistical analysis.
- These variables can be meaningfully **added, subtracted, multiplied, divided (ratios)**.
- Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.
- *** 10 miles is twice as long as 5 miles. 0 miles is no distance.**

Summary

- In summary, **nominal** variables are used to "*name*," or label a series of values.
- **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey.
- **Interval** scales give us the order of values plus the ability to quantify *the difference between each one*.
- Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Types of Attributes

- Other way of organizing attributes:

- **Discrete Attribute :** a finite or countably infinite set of values, which is mostly represented as integers.

->The attributes hair_color, medical_test, and grades each have a finite number of values, and so are discrete

->An attribute is countably infinite if the set of possible values is infinite but the values can be put in a one-to - one correspondence with natural numbers. customer_ID, Zip codes are countable infinite attributes.

Types of Attributes

- **Continuous Attribute**

  Continuous attributes are typically represented as floating-point variables.

Examples are Salary of an employee, Weight, Height, etc.

Object Identifier	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

- **Basic Statistical Descriptions of Data**
  - Measuring Central Tendency of Data
  - Measuring Dispersion of Data
  - Graphical Displays for Basic Statistical Description of Data

Measures of Central Tendency

- It is used to **measure the location of the middle or centre of a data distribution**
- Knowing such basic statistics regarding each attribute makes it easier **to fill in missing values, smooth noisy values, and spot outliers during data pre-processing**.
- Knowledge of the attributes and attribute values can also help in **fixing inconsistencies incurred during data integration**.

Measures of Central Tendency

- Measures of central tendency are the measure that are used for describing the data using a **single value .**
- **Mean, Median** and **Mode** are the three measures of central tendency.

Measuring the central tendency - Mean

- **Mean**
  - Average or arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

  - Weighted arithmetic mean or weighted average

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

  - Trimmed mean: because the tradition meanl is sentitive to extreme values (outliers)
    - e.g. order the values and remove the top and bottom 2%

Measuring the central tendency - Mean

- $\bar{X}$ and **μ** respectively denotes mean of sample and mean of population respectively.
- Mean can be interpreted as the centre of gravity of the distribution of the data.
- It is the most frequently used measure since it uses all the observations in the data set, but it is significantly **affected** by presence of **outliers**.
- An important property of mean is that the summation of deviation of observations from the mean is zero,

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

- Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$
$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000.

- What will be mean for following values:

1,2,70,80,90,80,89,87,90,87 = 67.6

Trimmed Mean = 82.66

Measuring the central tendency - Median

- **The middle value in a set of ordered data values**
  - The value that seperates the higher half from the lower half.
  - Useful for skewed (asymmetric) data
- **For a N value dataset**
  - If N is odd, then there is a unique median
  - If N is even
    - If the data are nominal, then median is not unique, any of the two middlemost values can be the median
    - If the data are numeric, the median is the average of the two middlemost values.
- **With large number of observations, median is hard to calculate. Approximation with the support of**
  - Binning and then find the median frequency
  - Let the interval containing the median frequency be the median interval (L_1, $L_1 + width$)
  - N is the total number of values in the dataset and $(\sum freq)_l$ is the total frequency values that are lower than the median frequency interval.

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

Measuring the central tendency - Median

- **Approximate Median**,

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

L_1 – Lower boundary of median interval

N – Number of values in the entire set

$(\sum freq)_l$ - Sum of the frequencies of all the intervals lower than the median interval

$freq_{median}$ - Frequency of median interval

width - Width of median interval.

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

- Suppose that the values for a given set of data are grouped into intervals. The intervals and using Equation, we have L1 = 21, N= 3194, $(\sum freq)_l$ = 950, $freq_{median}$ = 1500, width = 30, median = 32.94 years. Corresponding frequencies are as follows

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

Compute an approximate median value for the data.

we have $L_1 = 20$, $N = 3194$, $(\sum freq)_l = 950$, $freq_{median} = 1500$, $width = 30$, $median$ = 32.94 years.

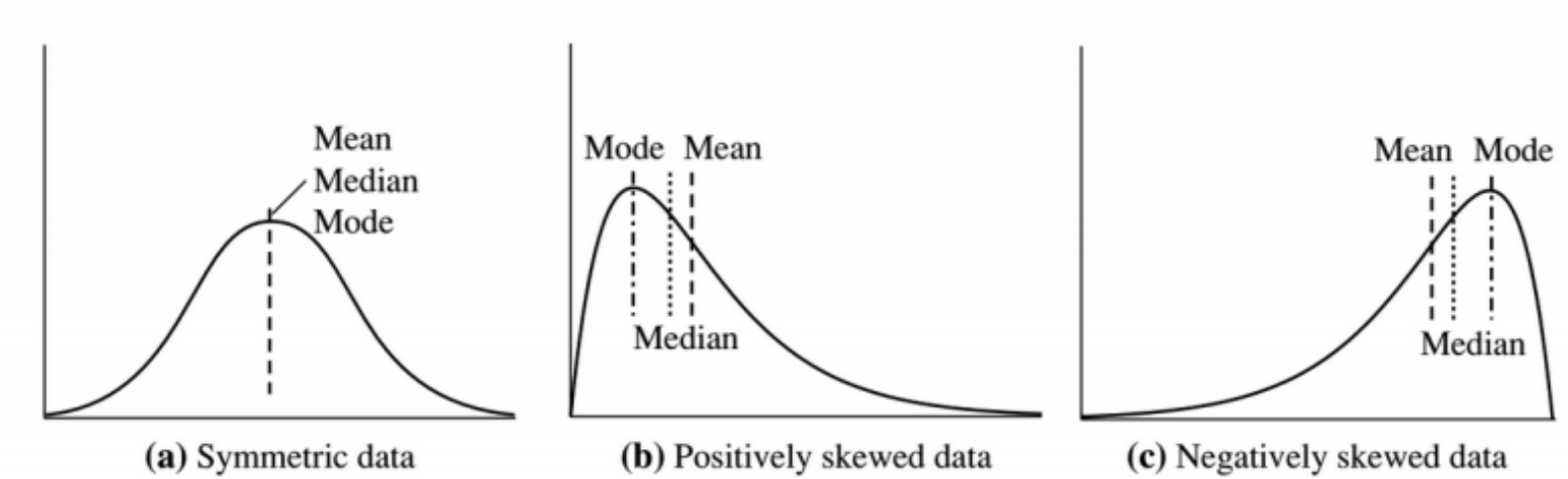
Measuring the central tendency - Mode

- **The mode for a set of data is the value that occurs most frequently in the set.**
  - It can be determined for qualitative and quantitative attributes.
  - The greatest frequency may correspond to several different values, which results in more than one mode.
    - Unimodal – one mode
    - Bimodal – two modes
    - Trimodal – three modes
    - Multimodal – multiple modes
    - No mode when each value only occurs once
- **For unimodal numeric data that moderately skewed, we have the following empirical relation:**

$$mean - mode \approx 3 \times (mean - median).$$

Measuring the central tendency - Midrange

- **Midrange is the average of min and max of the dataset**
  - $midrange = \frac{\boldsymbol{min()+max()}}{\mathbf{2}}$
- **skewed**
  - Positively: when $mode < median$
  - Negatively: when $mode > median$



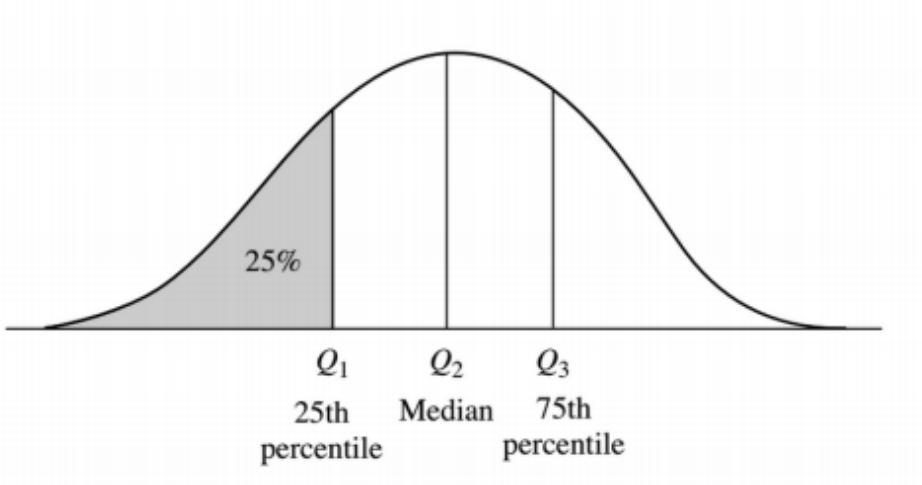
(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Measuring the Dispersion of Data

- Plotting the measures of central tendency shows us if the data are **symmetric or skewed**. This is also called as **dispersion of the data**.
- The most common data dispersion measures are the **range, quantiles, quartiles, percentiles and inter-quartile range.**
- **The five-number summary** and **boxplots** can be useful to identify the **Outliers** along with **the variance and standard deviation of the data.**

Measuring the Dispersion of Data

- **Range: max()-min()**
- **Quantiles: dividing the data into exactly equal-sized subsets**
  - Median is 2-quantile
  - Quartiles: 4-quantiles
  - Percentiles: 100-quantiles
- **Interquartile Range (IQR)**
  - $IQR = Q_3 - Q_1$



Measuring the Dispersion of Data

Let $x_1, x_2, \ldots, x_N$ be a set of observations for some numeric attribute, X. The **range** of the set is the difference between the largest (`max()`) and smallest (`min()`) values.

- Suppose that the data for attribute X are sorted in increasing numeric order. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets.
- These data points are called **quantiles**. **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

Quantiles-Percentile, Decile and Quartile

- The **2-quantile** is the data point dividing the lower and upper halves of the data distribution. It corresponds to the **median**.
- The **4-quantiles** are the **three data points** that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**.
- The **100- quantiles** are more commonly referred to as percentiles; they divide the data distribution into 100 equal-sized consecutive sets and commonly called as **percentiles**.
- **The decile, median, quartiles, and percentiles are the most widely used forms of quantiles.**

Quantiles-Percentile, Decile and Quartile

- The quartiles give an indication of a distribution's **centre, spread, and shape**.
- The first quartile, denoted by Q1, is the 25th percentile. It cuts off the lowest 25% of the data.
- The third quartile, denoted by Q3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data.
- The second quartile is the 50th percentile. As the median, it gives the centre of the data distribution

Quantiles-Percentile, Decile and Quartile

- Percentile, decile and quartile are frequently used to identify the **position of the observation in the data set**.
- Value at P_x is the position in the data set and calculated as -

$$P_x = \frac{x(n+1)}{100}$$

Value at P_x in the data set

- with Rounding or Approximations

- Ex. 30, 36,47, 50, 52,52,56,60,63,70,70,100

➢ with Rounding

P_{25} = 3.3 which is 47

P_{50} = 6.5 which is 52 and

P_{75} = 9.8 which is 70

➢ With Approximations

P_{25} = 3.3 which is 47 + 0.25 (value at 4^{th} – value at 3^{rd}) = 47+0.25(3)= 47.75

P_{50} = 6.5 which is 54 and

P_{75} = 9.8 which is 68.25

The five-number summary

- The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of **Minimum, Q1, Median, Q3, Maximum**.

Boxplots

- Boxplots are a popular way of visualizing a distribution.
- A boxplot incorporates the five-number summary as follows:

■ Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.

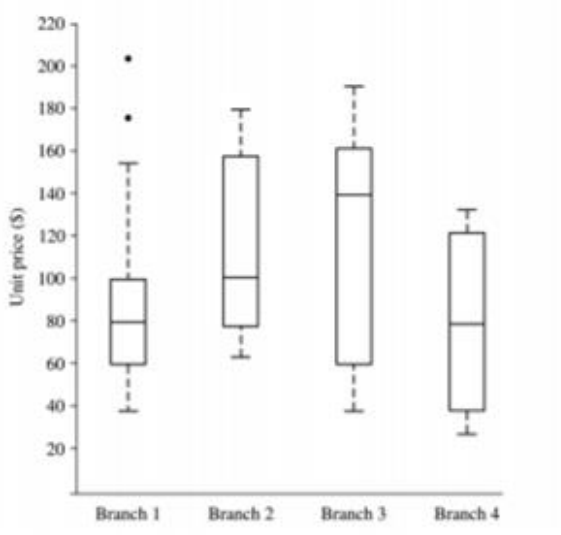
■ The median is marked by a line within the box.

■ Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Five Number Summary and Box Plots

- **The five-number summary of a distribution consists of**
  - the median (Q_2), the quartiles Q_1 and Q_3, and the smallest and largest individual observations,
  - written in the order of Minimum, Q_1, Median, Q_3, Maximum.
- **A common rule of thumb for identifying suspected outliers is to single out values**

Lower Limit = Q1 – 1.5 IQR
Upper Limit = Q3 + 1.5 IQR



- 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.
- Give the five-number summary of the data.
- Show a boxplot of the data.

Standard Deviation and Variance

- **Variance measures the dispersion of data around mean**
  - Used when mean is considered as the centre of the datasets
  - Most data are only several standard deviation away from the centre - another way of finding outliers

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

The standard deviation, σ, of the observations is the square root of the variance, σ^2.

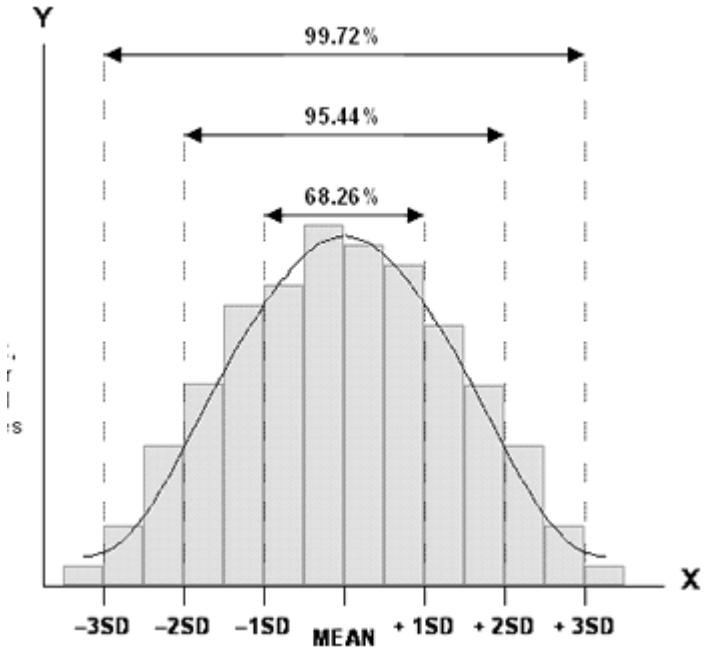
Standard Deviation and Variance

- When the values in a dataset are pretty tightly bunched together the **standard deviation is small**.
- When the values are spread apart the **standard deviation** will be **relatively large**.
- The standard deviation is usually presented in conjunction with the mean and is measured in the same units.

Standard Deviation and Variance

- For the normal distributions it is always the case that 68% of values are less than one standard deviation (1SD) away from the mean value
- That 95% of values are less than two standard deviations (2SD) away from the mean and
- That 99% of values are less than three standard deviations (3SD) away from the mean.

Standard Deviation and Variance



If the mean of a dataset is 25 and its standard deviation is 1.6, then

- 68% of the values in the dataset will lie between **MEAN-1SD** (25-1.6=23.4) and **MEAN+1SD** (25+1.6=26.6)
- 99% of the values will lie between **MEAN-3SD** (25-4.8=20.2) and **MEAN+3SD** (25+4.8=29.8).

Graphic Displays

- **Quantile(Q)** plots, **Quantile- Quantile(Q-Q)** plots, **Histograms**, and **Scatter** plots are other graphic displays of basic statistical descriptions. Many times we are using all such graphical displays **to visually inspect our data**
- These can all be useful during data pre-processing and can provide insight into areas for mining.
- The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Quantile Plot

- A **quantile** plot is 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 (allowing the user to assess both the overall behavior and unusual occurrences).
- Second, it plots quantile information

Quantile Plot

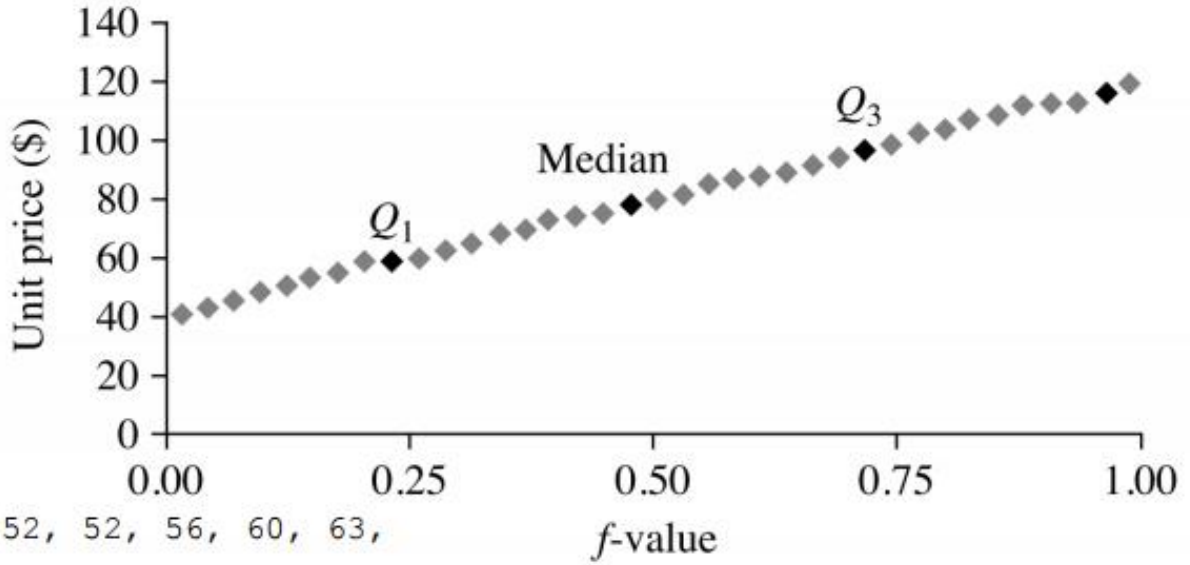
- Sort the dataset into ascending order $\{x_1, x_2, \ldots, x_N\}$, and construct a matching list for f_i

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

Unit price data :
40,43,47,74,75,78,115,117,120

For actual plot we are using modified f_i,

$$f_i = \frac{i - 1}{n - 1}, \quad (i = 1, \ldots, n)$$

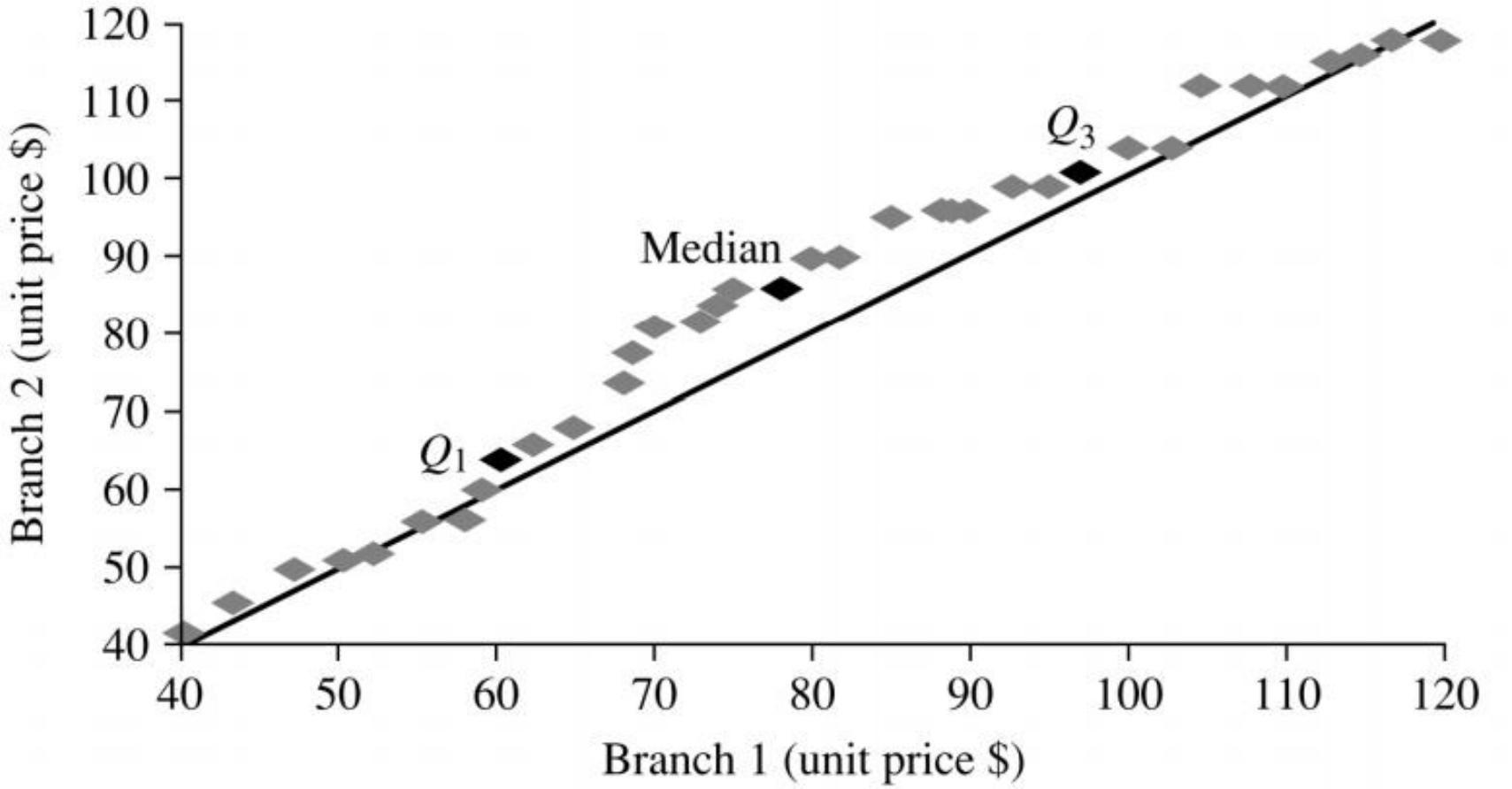



```
data <- c(30, 36, 47, 50, 52, 52, 56, 60, 63,
70, 70, 110)
N <- length(data)
z <- 1:N
f <- (z-0.5)/N
plot(f,data)
```

Quantile–Quantile Plot

- A quantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Quantile–Quantile Plot



Quantile–Quantile Plot

- Let be the data from the first branch $X_i = X_1, X_2 \ldots X_N$, and be the data from the second, $Y_i = Y_1, Y_2 \ldots Y_M$.
- If M = N (i.e., the number of points in each set is the same), then we simply plot y_i against x_i, where y_i and x_i are both $(i - 1)/N\text{-}1$ quantiles of their respective data sets.
- If M < N (i.e., the first branch has fewer observations than the second), there can be only M points on the q-q plot. Here, y_i is the $(i - 1)/M$ quantile of the y data, which is plotted against the $(i - 1)/M$ quantile of the x_i data.

Histograms (equal frequency)

- Histograms (or frequency histograms) is a graphical method for **summarizing the distribution of a given attribute**, X.
- If X is **nominal**, such as automobile_model or item_type, then a pole or vertical bar is drawn for each known value of X.
- The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a **bar chart**.

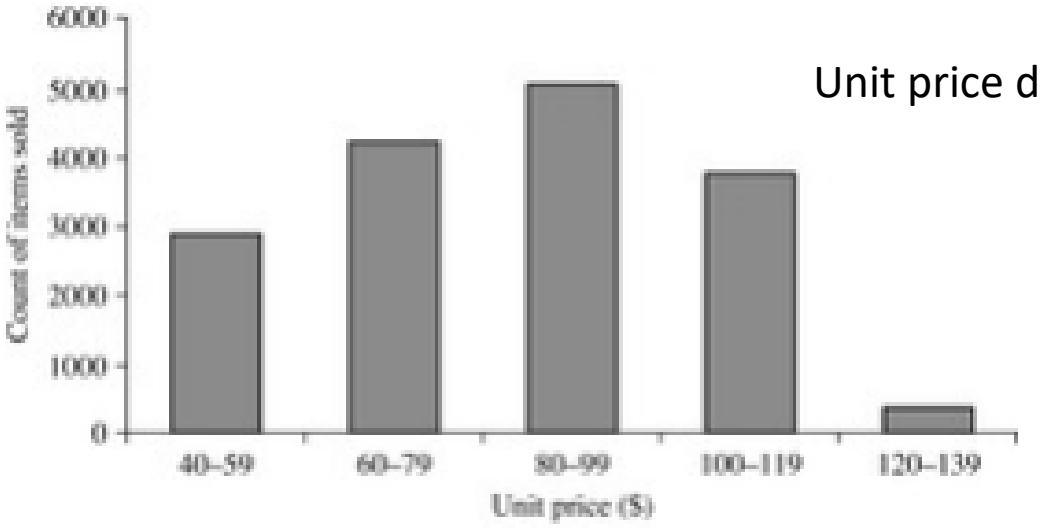
Histograms (equal width)

- If X is **numeric**, the term **histogram** is preferred. The range of values for X is partitioned into disjoint consecutive subranges.
- The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X.
- The range of a bucket is known as the **width**. Typically, the buckets are of equal width.

$$N = \frac{X_{max} - X_{min}}{W} \qquad w = 1 + 3.322 \log_{10}(n)$$

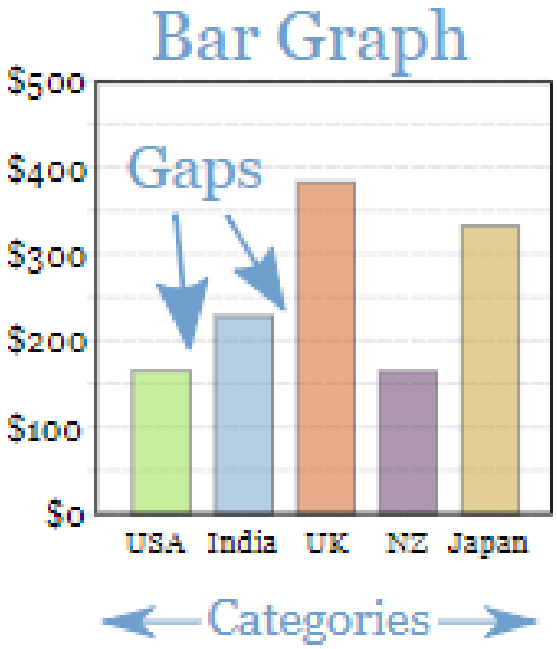
Histograms- equal -width

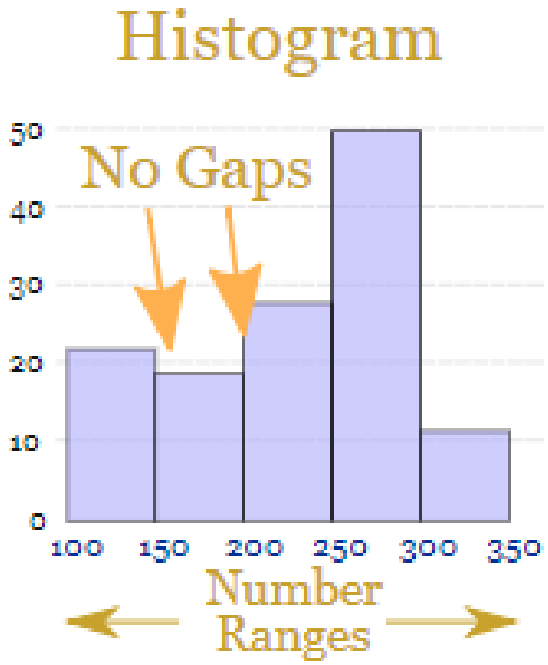
- Histogram for the data set, where buckets (or bins) are defined by equal -width ranges representing $20 increments and the frequency is the count of items sold.



Unit price data : 40.....130

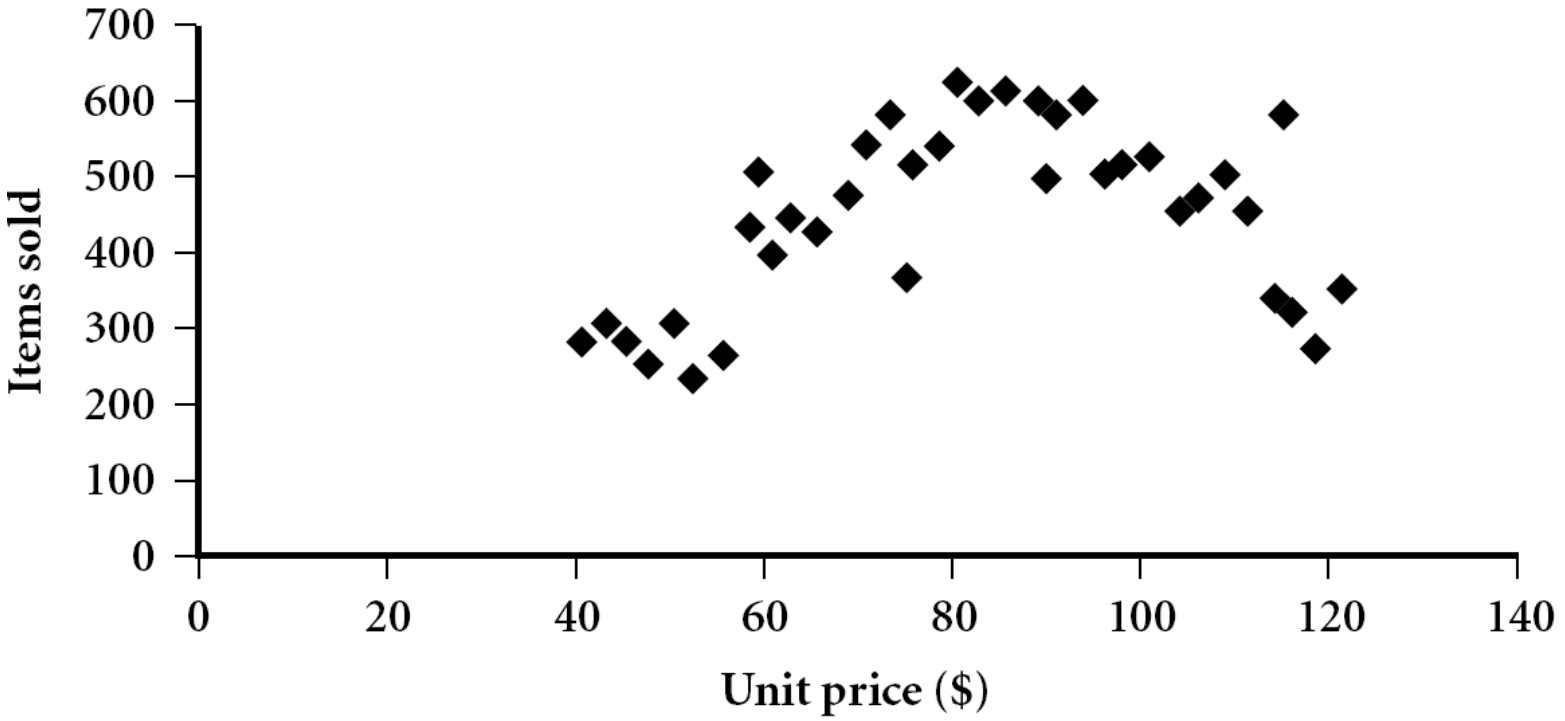
BAR and Histograms





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

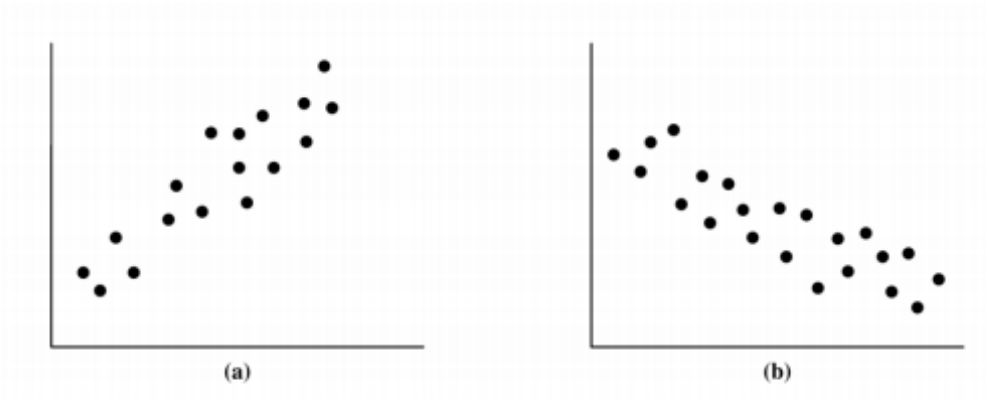


Scatter Plot

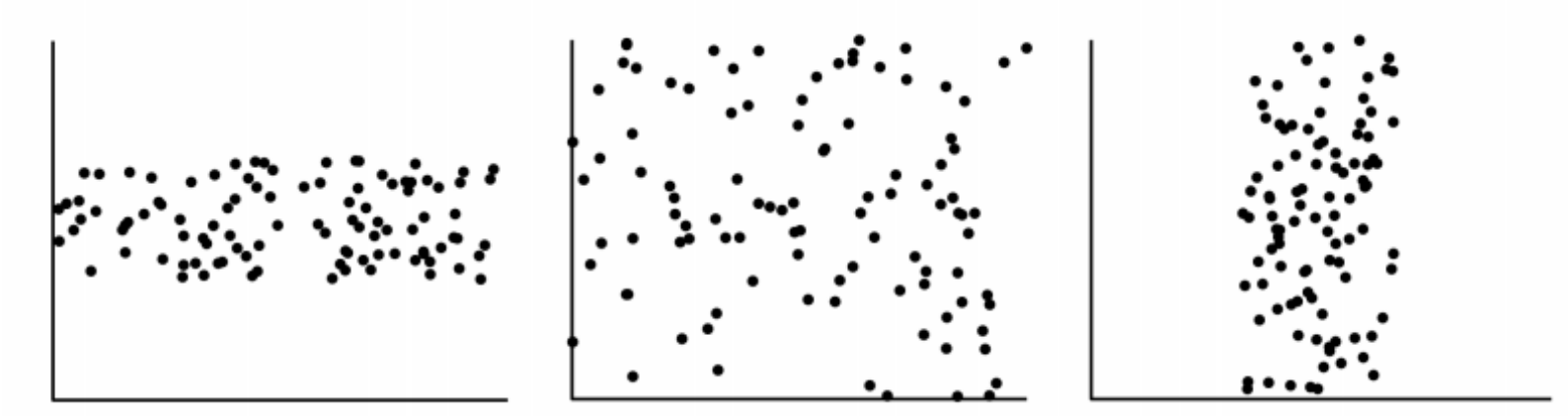
- The scatter plot is a useful method for providing a first look at bivariate data to see **clusters of points** and **outliers**, or **to explore the possibility of correlation relationships**.
- Two attributes, X, and Y, are correlated if one attribute implies the other.
- Correlations can be positive, negative, or null (uncorrelated).

Scatter Plot

- **Positive (a) and Negative (b) Correlation**



- **No correlation**



Data Visualization

- Why data visualization?
  - Aims to communicate data clearly and effectively through graphical representations.
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Measuring Data Similarity and Dissimilarity

- Clustering, Outlier analysis, and Nearest-Neighbour classification, we need ways to assess how alike or unalike objects are in comparison to one another.
- There are many measures for assessing similarity and dissimilarity.
- In general, such measures are referred to as **proximity measures**.
- The proximity of two objects is a distance between their attribute values.

Measuring Data Similarity and Dissimilarity

- 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

Attributes of Mixed Type

- A database may contain all attribute types
  - **Nominal, symmetric binary, asymmetric binary, numeric, ordinal**
- One may use a weighted formula to combine their effects

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

  - where the indicator $\delta_{ij}^{(f)}$=0 if either –

(1) x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or object j), or

(2) $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary;

(3) otherwise, $\delta_{ij}^{(f)}$=1 .

Example

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

Data Matrix and Dissimilarity Matrix

- Data matrix
  - 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
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode

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

Data Matrix versus Dissimilarity Matrix

- Data matrix (or object-by-attribute structure): This structure stores the n data objects in the form of a relational table, or **n-by-p** matrix

  (n objects × p attributes):

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

- Each row corresponds to an object. As part of our notation, we may use ***f*** to index through the ***p*** attributes.

Data Matrix versus Dissimilarity Matrix

- Dissimilarity matrix (or object-by-object structure): This structure stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table

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

where d(i, j) is the measured dissimilarity or “difference” between objects i and j. In general, d(i, j) is a nonnegative number that is close to 0 when objects i and j are highly similar or “near” each other, and becomes larger the more they differ. Note that d(i, i) = 0; that is, the difference between an object and itself is 0.

Measures of Similarity

- Measures of similarity can often be expressed as a function of measures of dissimilarity.

$$sim(i,j) = 1 - d(i,j)$$

where sim(i, j) is the similarity between objects i and j.

Example

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

Proximity Measures for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Simple matching
  - *m*: # of matches, *p*: total # of attributes

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

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

Proximity Measures for Binary Attributes

A contingency table for binary data

		Object *j*	
		1	0
Object *i*	1	q	r
	0	s	t

- Distance measure for symmetric binary variables: $$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distance measure for asymmetric binary variables: $$d(i, j) = \frac{r + s}{q + r + s}$$

Proximity Measures for Binary Attributes

- Jaccard coefficient

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

i.e. 1-d (i,j)

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 1, and the value N 0

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

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

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

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

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

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}.$$

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

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

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- h = 1: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

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

- h = 2: (L_2 norm) Euclidean distance

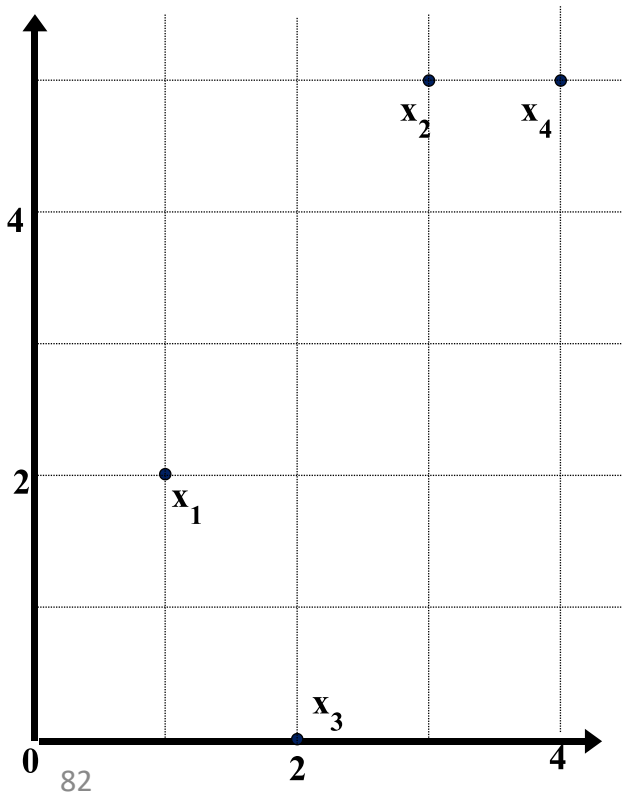
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

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

Note: f **is numeric,use the normalized distance**

Mixed Type Attributes

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, otherwise, $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

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

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

Dissimilarity for Attributes of Mixed Types

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

The resulting dissimilarity matrix obtained for the data described by the three attributes of mixed types is:

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

Normalization

- **Min-max normalization**:Min-max normalization performs a linear transformation on the original data. Suppose that minA and maxA are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, vi, of A to v'i in the range [new_minA, new_maxA] by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range 12,000 to 98,000 normalized to [0.0, 1.0]. Then 73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.
- each document is an object represented by what is called a 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

- **Cosine similarity** is a measure of **similarity** that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let **A** and **B** be two vectors for comparison. Using the cosine measure as a similarity function, we have

Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

where ||A|| and ||B|| are the Euclidean norm of vector

A cosine value of **0** means that the two vectors are at **90 degrees** to each other (orthogonal) and have **no match**. The closer the cosine value to **1**, the **smaller the angle** and the **greater the match between vectors**.

Cosine Similarity

$x = (5,0,3,0,2,0,0,2,0,0)$ and $y = (3,0,2,0,1,1,0,1,0,1)$. How similar are x and y? compute the cosine similarity between the two vectors, we get:

$$x \cdot y = 5\times3+0\times0+3\times2+0\times0+2\times1+0\times1+0\times0+2\times1 +0\times0+0\times1 = 25$$

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

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

$$sim(x,y) = 0.94$$

Therefore, if we were using the cosine similarity measure to compare these documents, they would be considered quite similar.