# Data Exploration & Data Preprocessing

2 Chapter

# 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.

# Attributes

- **Attribute (**or **dimensions, features, variables**):

  a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Ordinal
  - Numeric
    - Interval-scaled
    - Ratio-scaled

# Attribute Types

1) **Nominal Attribute :**

- Related to names

- categories, states, or “names of things”
  - *Hair_color =* {*black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers

- Also called as categorical Attribute

2) **Binary Attribute**

- Nominal attribute with only 2 states (0 and 1)
- 0 means attributes absent and 1 means attribute present
- Also called Boolean if two state corresponds to true and false
- Symmetric binary: both outcomes equally important and carry same weight.
  - e.g., gender
- Asymmetric binary: outcomes not equally important
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

- **Ordinal Attribute**

  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
    - *Size =* {*small, medium, large*}*,*
    - grades=(A+, A, B, B+)

# Numeric Attribute Types

**Numeric Attribute:**

- Quantitative (integer or real-valued)
- **Interval-scaled attributes**
  - 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-scaled attributes**
  - Inherent **zero-point**
  - We can speak of values as being a multiple of another value

    (10 Kg is twice of 5 Kg).
    - e.g. *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
  - Continuous attributes are typically represented as floating-point variables

# Basic Statistical Descriptions of Data

- For successful data preprocessing it is essential to have overall picture of your data
- Statistical Descriptions can be used to identify propertied of the data
- Three area of basic SD of data-------

1) **Measuring the central tendency of data:**
   - mean, median, mode (where most of the value fall?)
2) **Measuring the dispersion of data:**

   Range, quartiles, variance, standard deviation, interquartile range, five number summery, box plot
3) **Graphic display of basic sd of data:**

   Quantile plot, quantile quantile plot, histogram, scatter plots

- Mean:
  - Effective measure of the center of a set of data
  - Let x1, x2, ….xn be set of n values for numeric attribute X like salary. The mean is given by---

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

**Mean = (x1+x2+….+xn)/n**

  - Ex. 30,36,47,50,52,52,56,60,63,70,70,110

- Weighted arithmetic mean

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

- Trimmed mean: chopping extreme values

# Median:

1. For **odd number** of values (count), Middle value is considered.

2. For **even number** of values (count), average of two middle most values is considered.

3. For **interval**---

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

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

Here

L1→ lower boundary of the median interval

N→ number of values in entire dataset

$(\sum freq)_1$→ sum of frequencies of all of the intervals that are lower than the median interval

$Freq_{median}$→ frequency of the median interval

Width→ width of the median interval

- L1→ 20
- N→ 3194
- $(\sum freq)_1$→ 950
- Freq $_{median}$→ 1500
- Width→ 30

Ans-→

Median=32.94 years

- Mode
  - A mode is defined as the value that has a higher frequency in a given set of values.
  - It is the value that appears the most number of times.
  - **Example**: In the given set of data: 2, 4, 5, 5, 6, 7, the mode of the data set is 5 since it has appeared in the set twice.

**Bimodal, Trimodal & Multimodal (More than one mode)**

- When there are two modes in a data set, then the set is called **bimodal**
  - For example, The mode of Set A = {2,2,2,3,4,4,5,5,5} is 2 and 5, because both 2 and 5 is repeated three times in the given set.
- When there are three modes in a data set, then the set is called **trimodal**
  - For example, the mode of set A = {2,2,2,3,4,4,5,5,5,7,8,8,8} is 2, 5 and 8
- When there are four or more modes in a data set, then the set is called **multimodal**

– Empirical formula for unimodal numeric data---

$$mean - mode = 3 \times (mean - median)$$

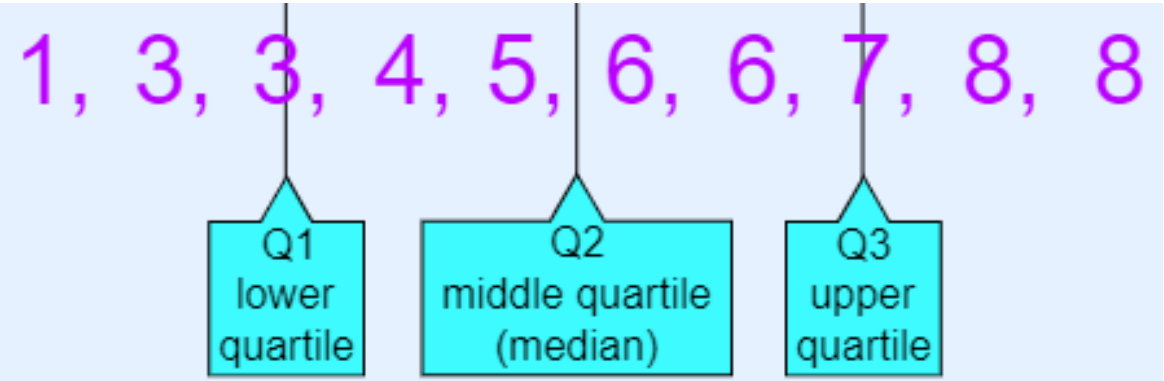
- Midrange
  - Average of largest and smallest value in dataset
  - Midrange = (Maximum Value + Minimum Value) / 2
  - **Example:** Consider the data set 110, 150, 180, 220, 270, 290, 310 and 390 as the prices of speakers. The minimum number is 110, and the maximum is 390.
  - Midrange=(390 + 110) / 2 = 250

# Measuring the Dispersion of Data

- **Quartiles, outliers and boxplots:**
  - **Quartiles**: $Q_1$ ($25^{th}$ percentile), $Q_3$ ($75^{th}$ 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 1.5 x IQR

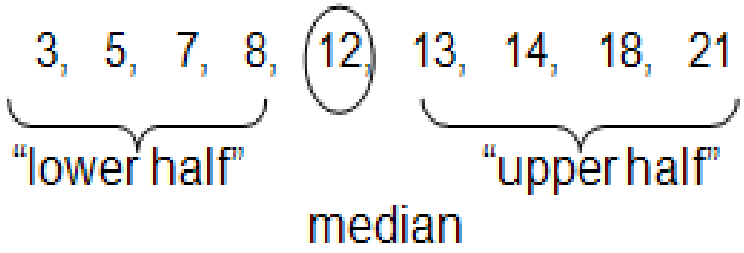
# Quartiles

- Quartiles are the values (Q1, Q2,Q3,Q4) that divide a list of numbers into quarters or four parts.



# Quartiles

- **Example 1:** Find the first and third quartiles of the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.
- **Total numbers in set=9**
- First, write data in increasing order: 3, 5, 7, 8, 12, 13, 14, 18, 21.
- The median is 12.
- The first quartile, $Q_1$, is the median of {3, 5, 7, 8}=6
- The third quartile, $Q_3$, is the median of {13,14,18,21}=16




$$Q_1 = \frac{5+7}{2} = \frac{12}{2} = 6$$

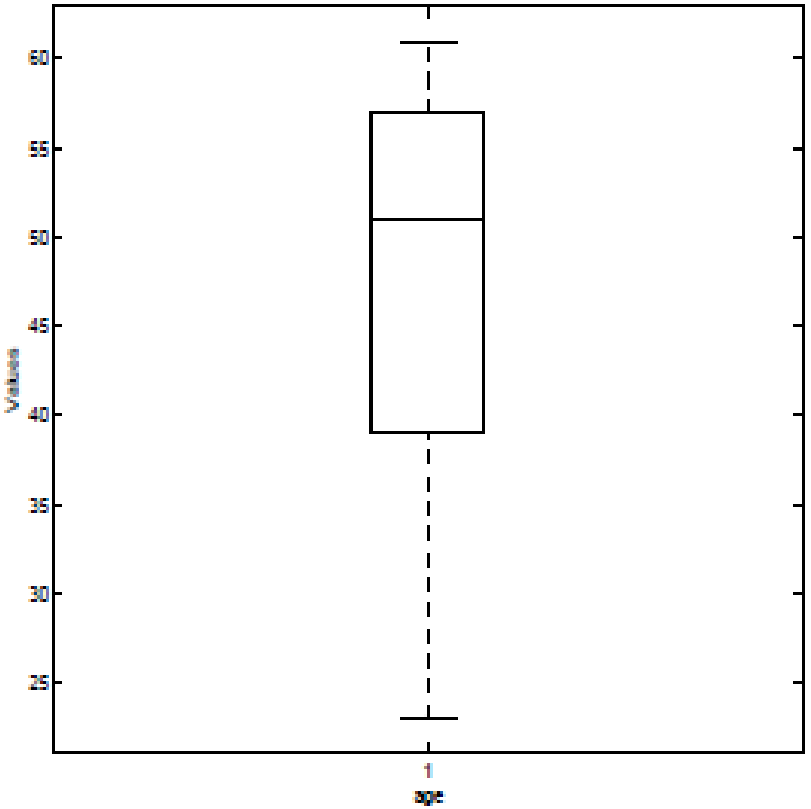
$$Q_3 = \frac{14+18}{2} = \frac{32}{2} = 16$$

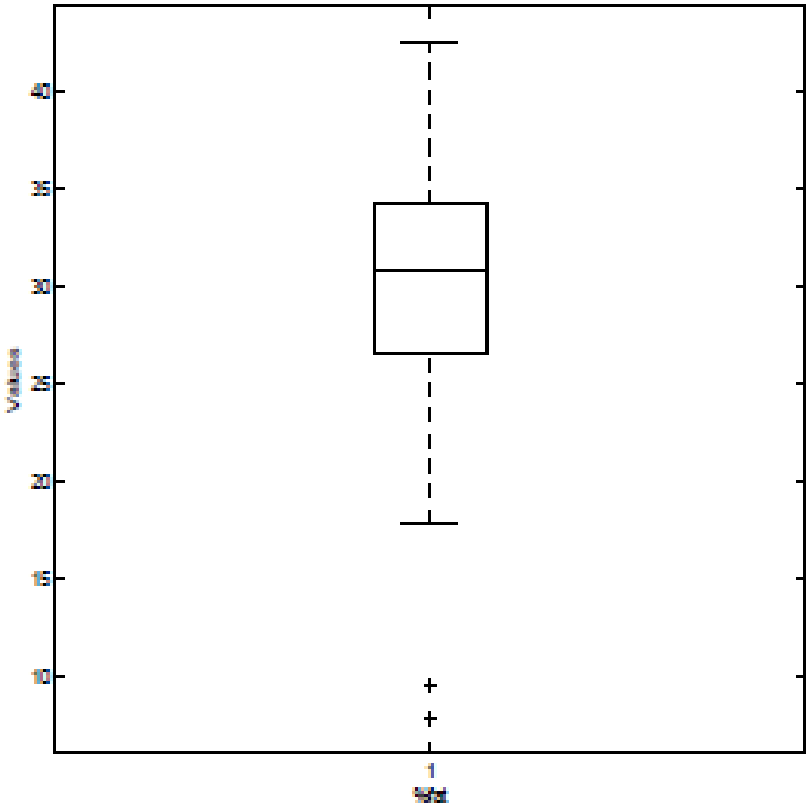
# Quartiles

- **Example 2:** Find the first and third quartiles of the set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

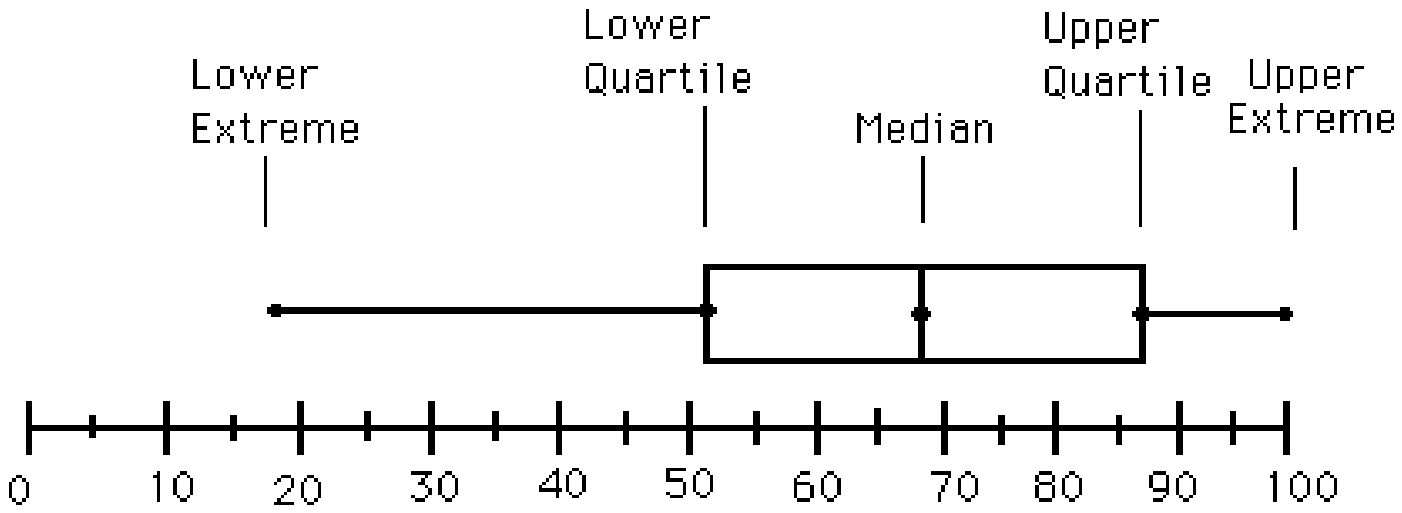
- Median (Q2) is 13 (it is the mean of 12 and 14)
- $\boldsymbol{Q_1 = 8}$
- $\boldsymbol{Q_3 = 15}$.

# Boxplot

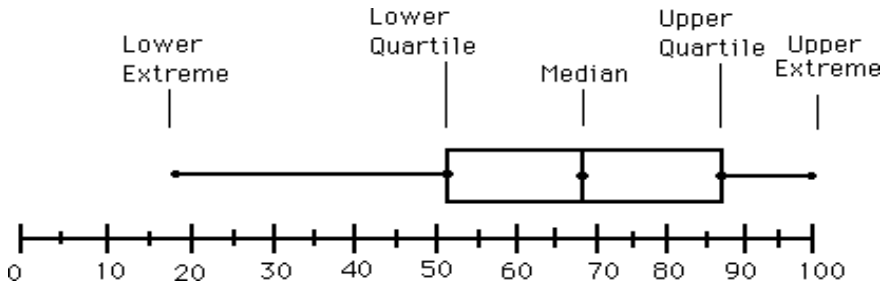

60
55
50
45
40
35
30
25
Values
1
age


40
35
30
25
20
15
10
Values
1
%fat

# Boxplot


Lower Extreme
Lower Quartile
Median
Upper Quartile
Upper Extreme
0
10
20
30
40
50
60
70
80
90
100

# 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 Example 1

Example: A sample of 10 boxes of raisins has these weights (in grams):

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

**Make a box plot of the data**

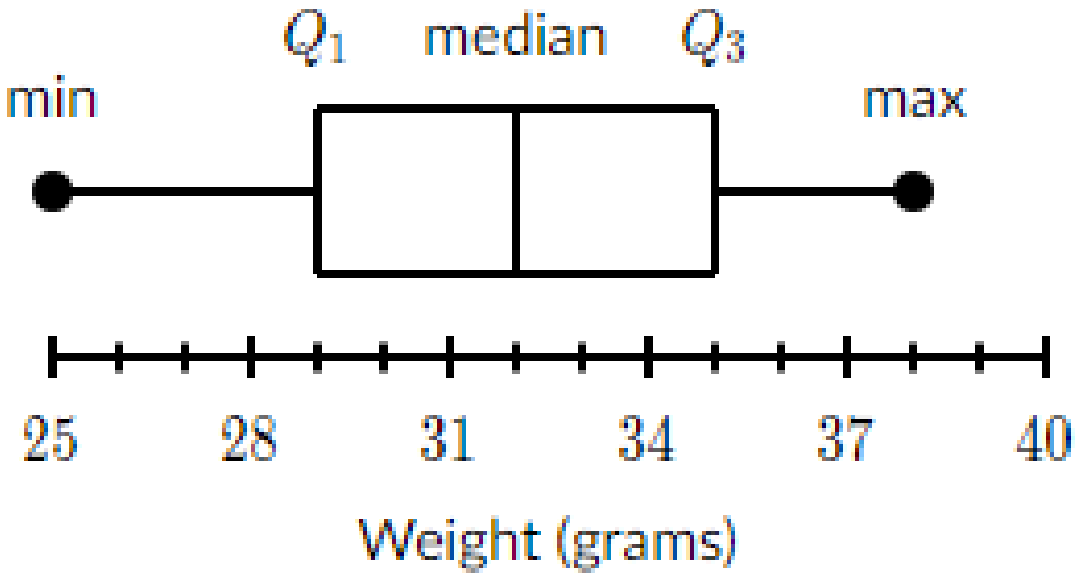
**Solution:**

**Step 1:** Order the data from smallest to largest.

Step 2: find the 5 number summary

(minimum, first quartile, median, third quartile, and maximum)

- ( min, Q1,Q2,Q3,max)
- (25,29,32,35,38)

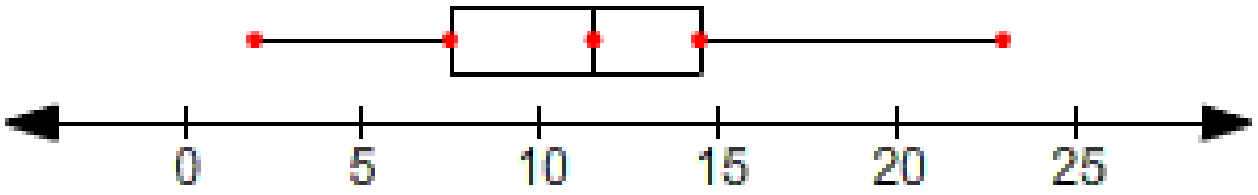



Five number summary: (25,29,32,35,38)

# Ex 2

- Find Q1 , Q2 , and Q3 for the following data set, and draw a box-and-whisker plot.
- {2,6,7,8,8,11,12,13,14,15,22,23}

- Five number summary
- (2, 7.5, 11.5,14.5,23)



# Boxplot Example 3

Ex 2. 30,36,47,50,52,52,56,60,63,70,70,110

- Draw the box plot

# Ex 4

Find $Q_1$ , $Q_2$ , and $Q_3$ for the following data set. Identify any outliers, and draw a box-and-whisker plot.

$$\{5, 40, 42, 46, 48, 49, 50, 50, 52, 53, 55, 56, 58, 75, 102\}$$

There are $15$ values, arranged in increasing order. So, $Q_2$ is the $8^{\text{th}}$ data point, $50$ .

$Q_1$ is the $4^{\text{th}}$ data point, $46$ , and $Q_3$ is the $12^{\text{th}}$ data point, $56$ .

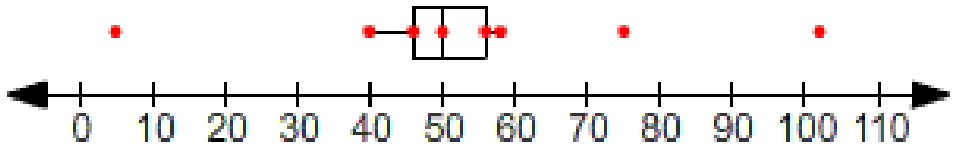
The interquartile range $\text{IQR}$ is $Q_3 - Q_1$ or $56 - 47 = 10$ .

Now we need to find whether there are values less than $Q_1 - (1.5 \times \text{IQR})$ or greater than $Q_3 + (1.5 \times \text{IQR})$

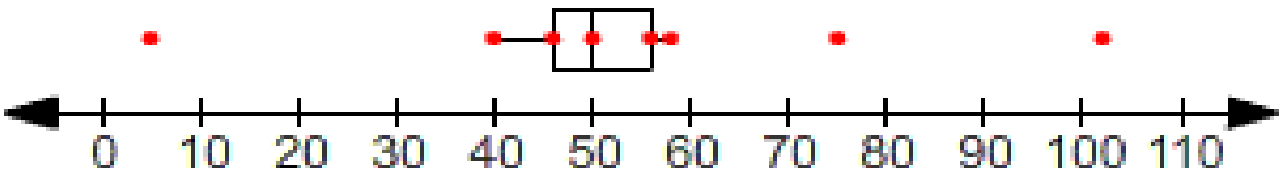
$$Q_1 - (1.5 \times \text{IQR}) = 46 - 15 = 31$$

$$Q_3 + (1.5 \times \text{IQR}) = 56 + 15 = 71$$

Since $5$ is less than $31$ and $75$ and $102$ are greater than $71$ , there are $3$ outliers.



Note that 40 and 58 are shown as the ends of the whiskers with outliers plotted separately as a dots.



# Outliers

- If a data value is very far away from the quartiles (either much less than Q1 or much greater than Q3), it is sometimes designated an outlier .
- The standard definition for an outlier is a number which is less than Q1 or greater than Q3 by more than 1.5 times the interquartile range
- IQR=Q3−Q1
- That is, an outlier is any number

less than Q1−(1.5×IQR) **or**

greater than Q3+(1.5×IQR)

# Variance & Standard Deviation

- Variance: **Variance** is the sum of squares of differences between all numbers and means.

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

- Where μ is Mean, N is the total number of elements or frequency of distribution.
- **Standard Deviation** is square root of variance. It is a measure of the extent to which data varies from the mean.
- The Standard Deviation is a measure of how spread out numbers are.

## Example 1 – Standard deviation

A hen lays eight eggs. Each egg was weighed and recorded as follows

60 g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54 g.

a. First, calculate the mean:

$$\overline{x} = \frac{\sum x}{n}$$

$$= \frac{472}{8}$$

$$= 59$$

b. Now, find the standard deviation.

**Table 1. Weight of eggs, in grams**

| Weight (x) | $(x - \bar{x})$ | $(x - \bar{x})^2$ |
|---|---|---|
| 60 | 1 | 1 |
| 56 | -3 | 9 |
| 61 | 2 | 4 |
| 68 | 9 | 81 |
| 51 | -8 | 64 |
| 53 | -6 | 36 |
| 69 | 10 | 100 |
| 54 | -5 | 25 |
| **472** | | **320** |

Using the information from the above table, we can see that

$$\sum (x - \bar{x})^2 = 320$$

In order to calculate the standard deviation, we must use the following formula:

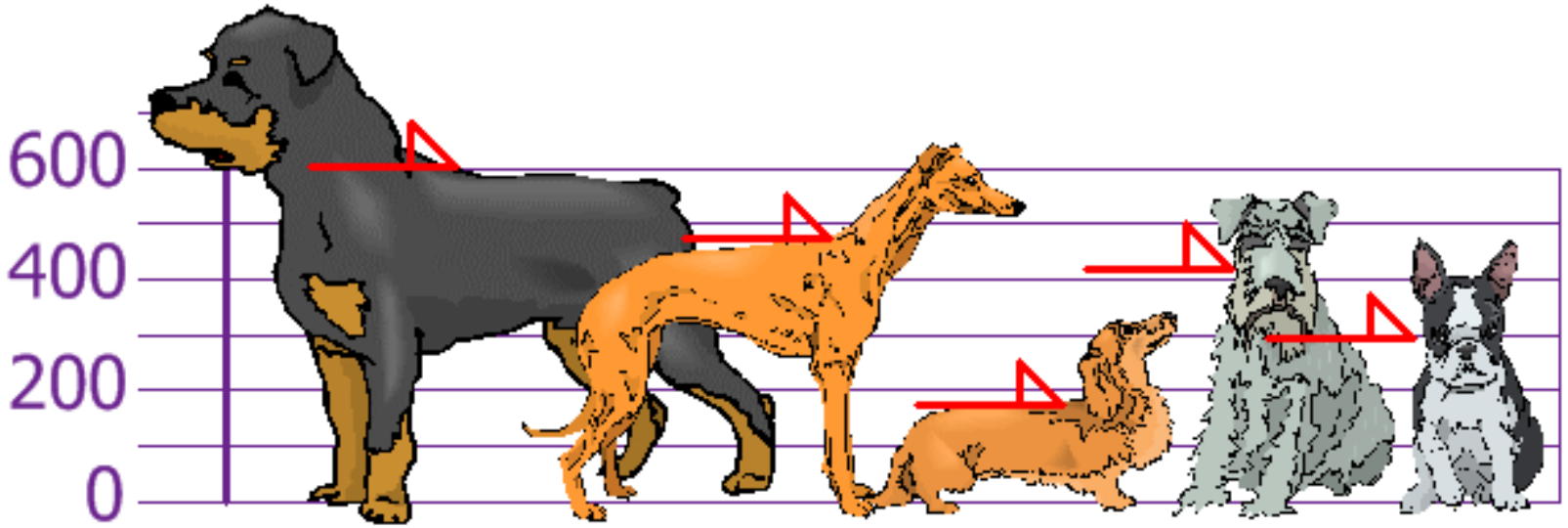
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{320}{8}}$$

$$= 6.32 \text{ grams}$$

# Example

You and your friends have just measured the heights of your dogs (in millimeters):




The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

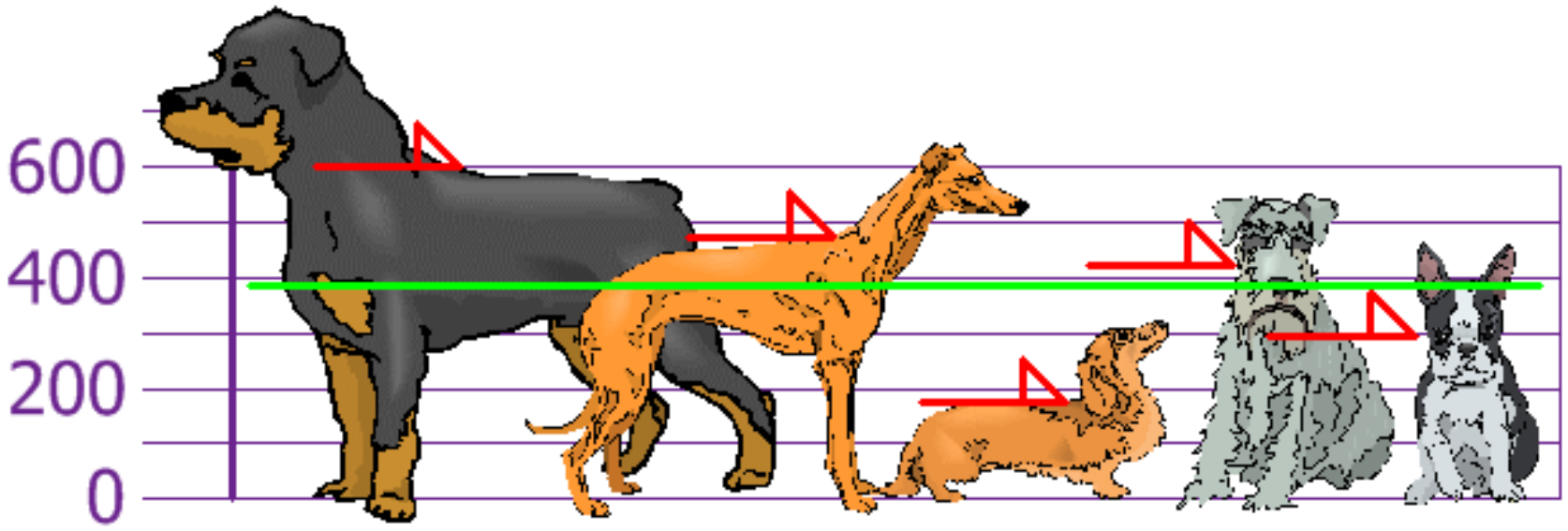
## Answer:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

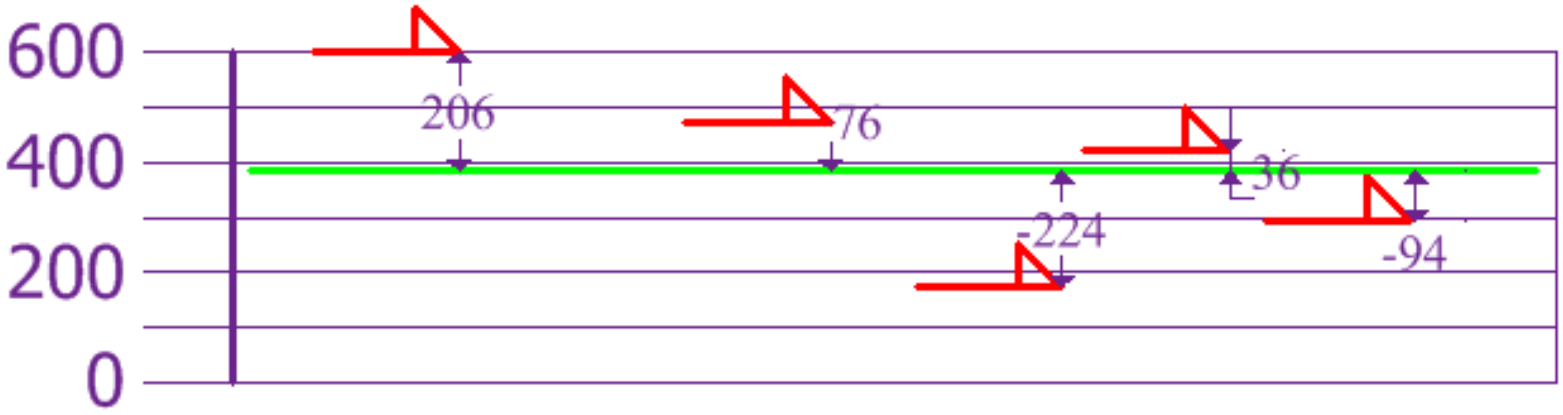
$$= \frac{1970}{5}$$

$$= 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



Now we calculate each dog's difference from the Mean:




To calculate the Variance, take each difference, square it, and then average the result:

**Variance**

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

So the Variance is **21,704**

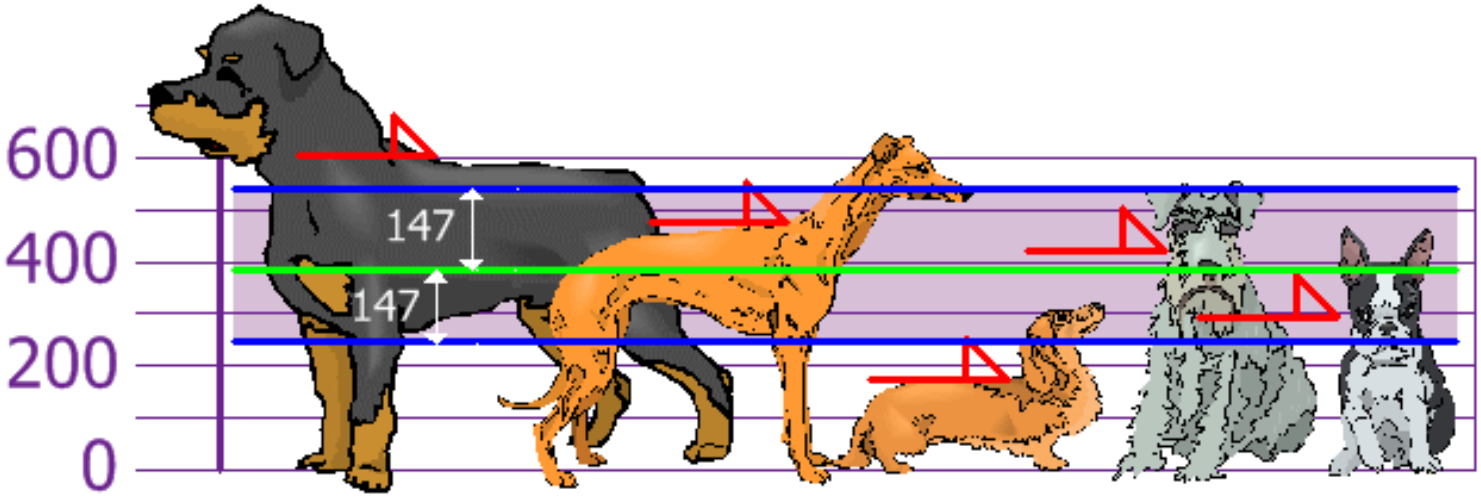
And the Standard Deviation is just the square root of Variance, so:

**Standard Deviation**

$$\sigma = \sqrt{21704}$$
$$= 147.32...$$
$$= \mathbf{147} \textit{ (to the nearest mm)}$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:




So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short, right?

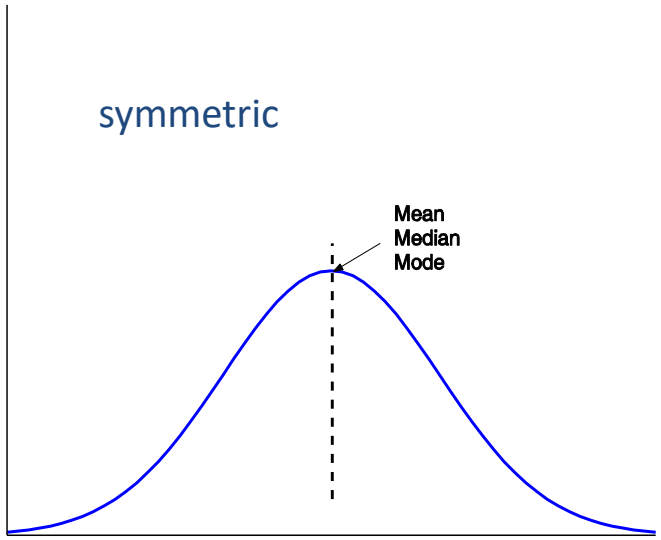
# Key points about variance/standard deviation

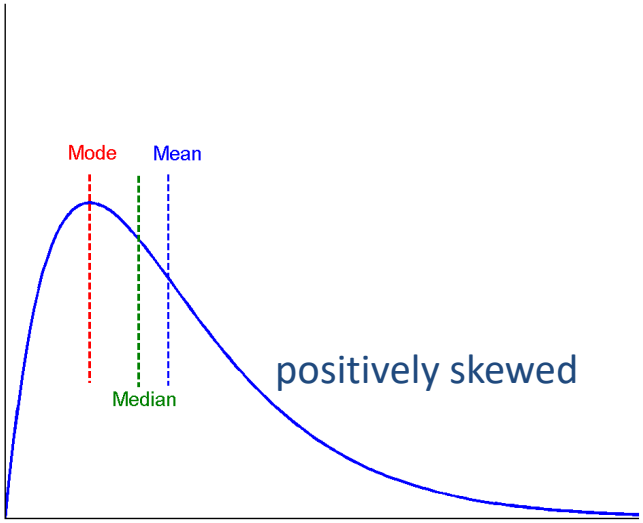
- Variance and standard deviation are measures of data dispersion
- They indicate how spread out a data distribution is.
- A **low standard** deviation means that the data observation tend to be very close to mean
- **High standard** deviation indicates data are spread out over a large range of values
- $\sigma = 0$ when there is no spread, that is , when all observation have the same value. Otherwise $\sigma > 0$

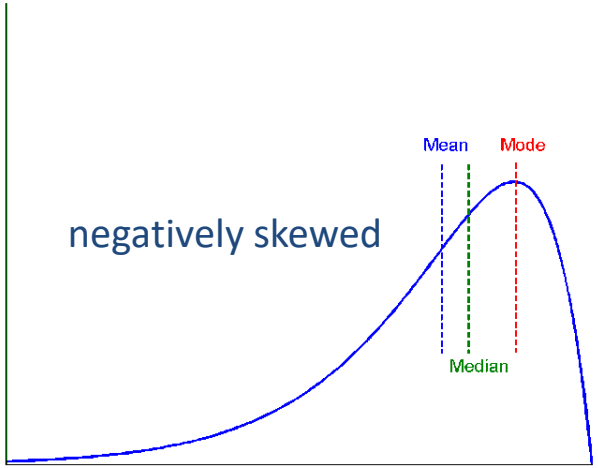
# Symmetric vs. Skewed Data

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

"skewed to the left" (the long tail is on the left hand side): negatively skewed
"skewed to the right" (the long tail is on the right hand side): positively skewed



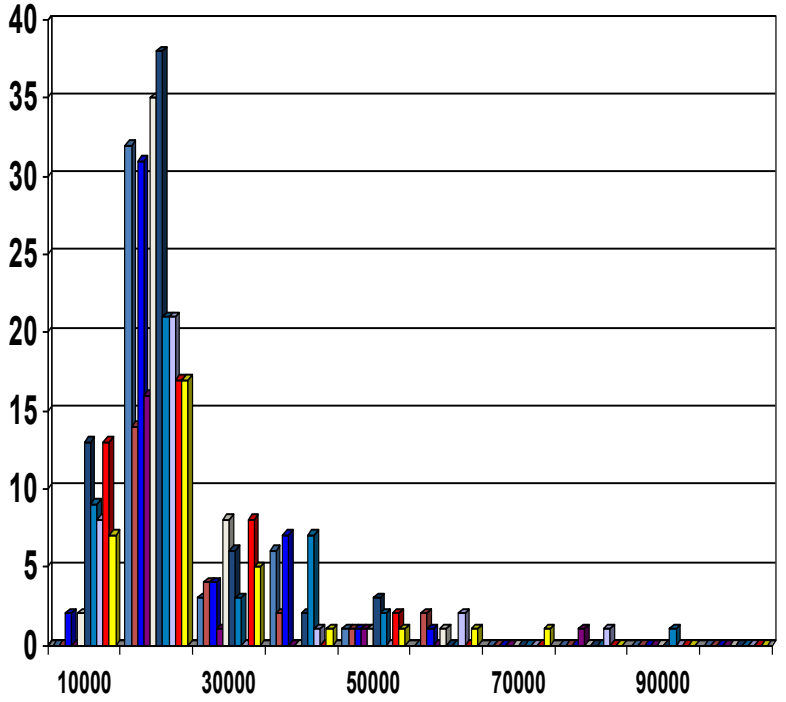



# Graphic Displays of Basic Statistical Descriptions

- **Histogram**: x-axis are values, y-axis represents frequencies
- **Quantile plot**: each value $x_i$ is paired with $f_i$ indicating that approximately 100 $f_i$ % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant 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

# 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

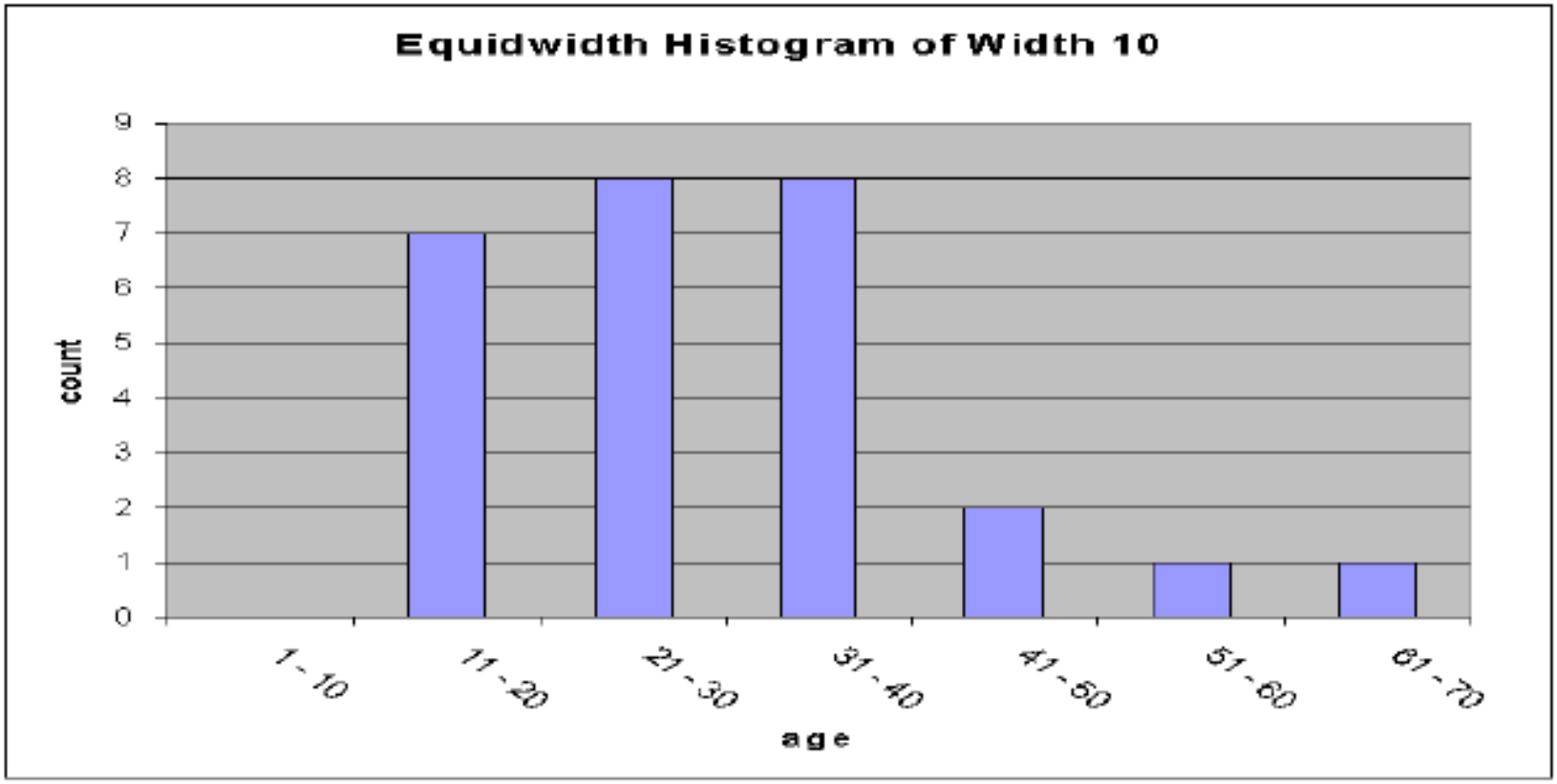


# Example Histogram

- Dataset for Age:
  13,15,16,16,20,20,21,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70

| Age | Count/Frequency |
|---|---|
| 13 | 1 |
| 15 | 1 |
| 16 | 2 |
| 20 | 2 |
| 21 | 1 |
| 22 | 1 |
| 25 | 4 |
| 30 | 1 |
| 33 | 2 |
| 35 | 4 |

# Example Equidwidth Histogram



# Example Histogram

| Unit price($) | Count of items sold |
|---|---|
| 40 | 275 |
| 43 | 300 |
| 47 | 250 |
| 74 | 360 |
| 75 | 515 |
| 78 | 540 |
| 115 | 320 |
| 117 | 270 |
| 120 | 350 |

# Quantile plot

- Used to check whether your data is Normal
- To make a QQplot:

  For a sample of size $n$: $x_1, x_2, \ldots x_n$

  1. Order the data from smallest to largest:

     $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ where $x_{(i)}$ is the $i$-th smallest

  2. Calculate the sample quantile

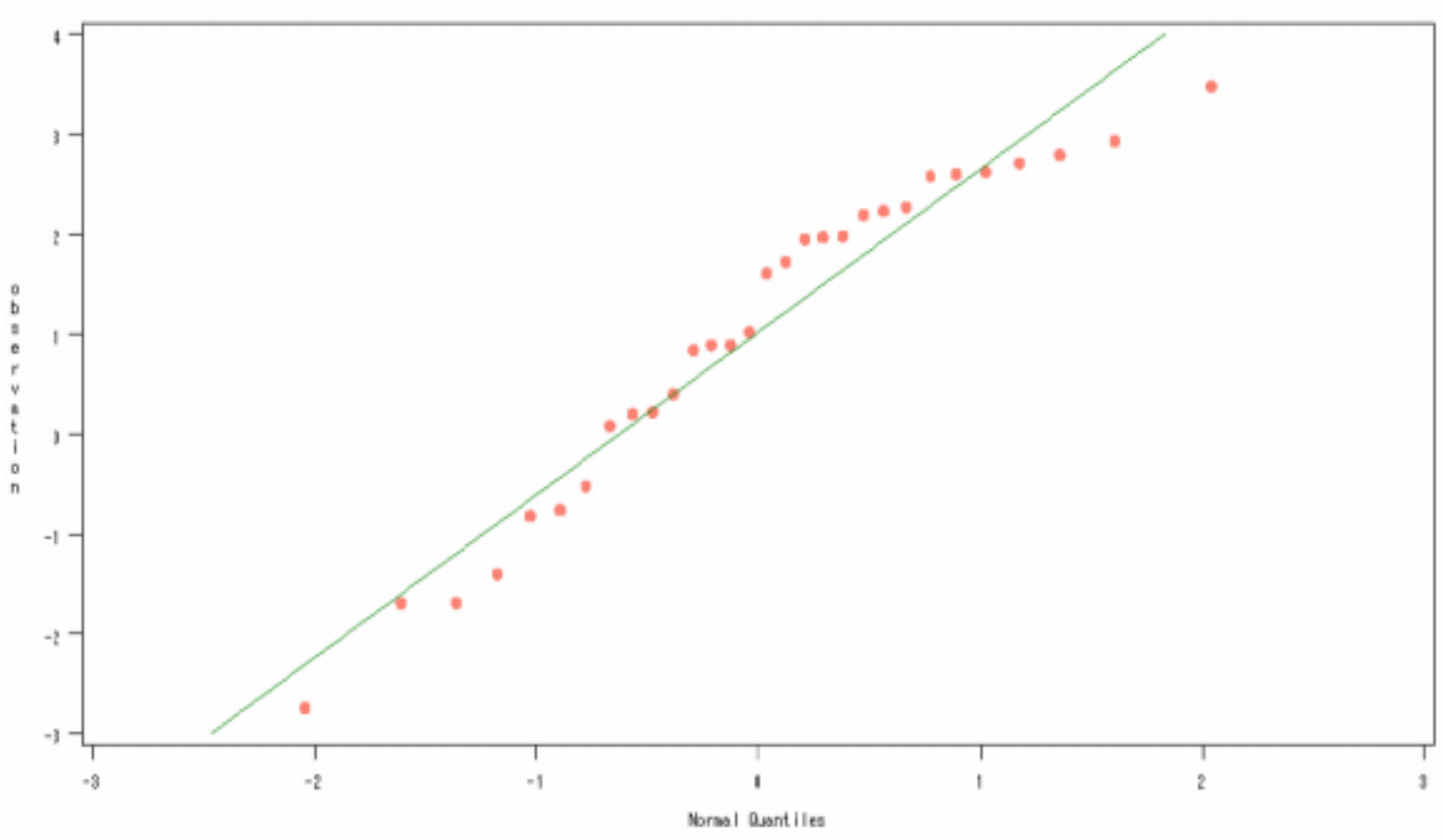
     Sample quantile is calculated as:

     $x_{(i)}$= $[(i-0.5)/n]$th sample quantile

  3. Plot the points ($[(i-0.5)/n]$th z-percentile, $x_{(i)}$)

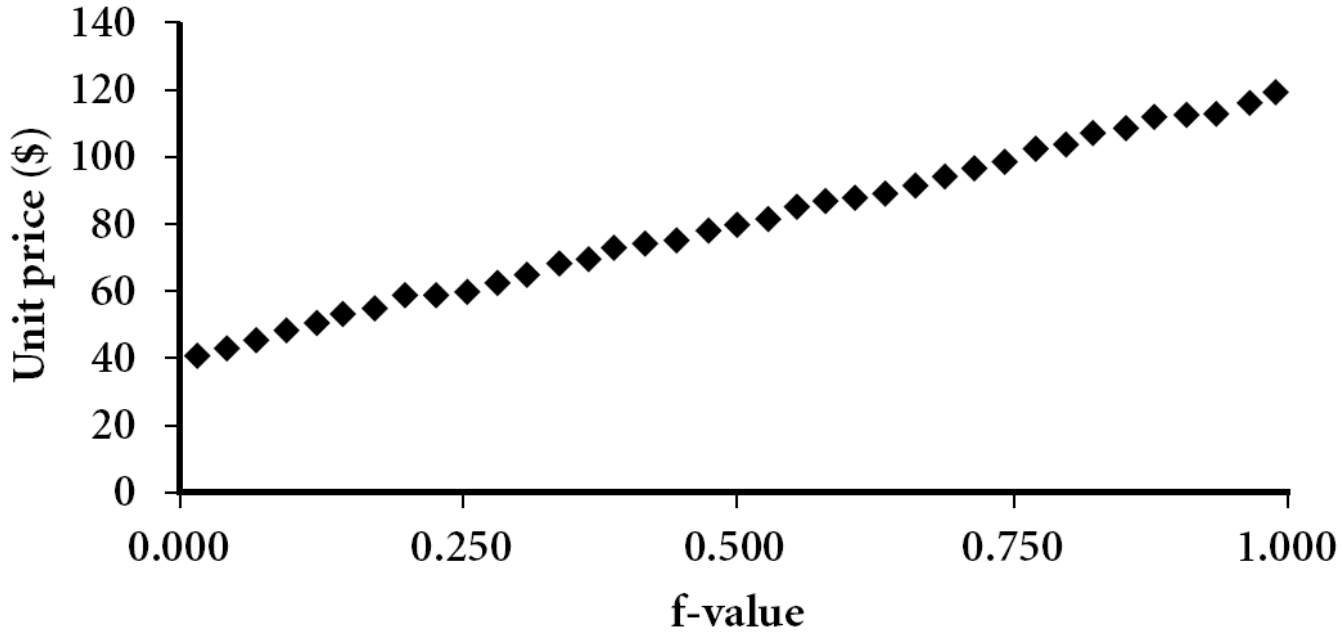
- If the data distribution is close to normal, the plotted points will lie close to a sloped straight line on the QQplot!

# Quantile plot



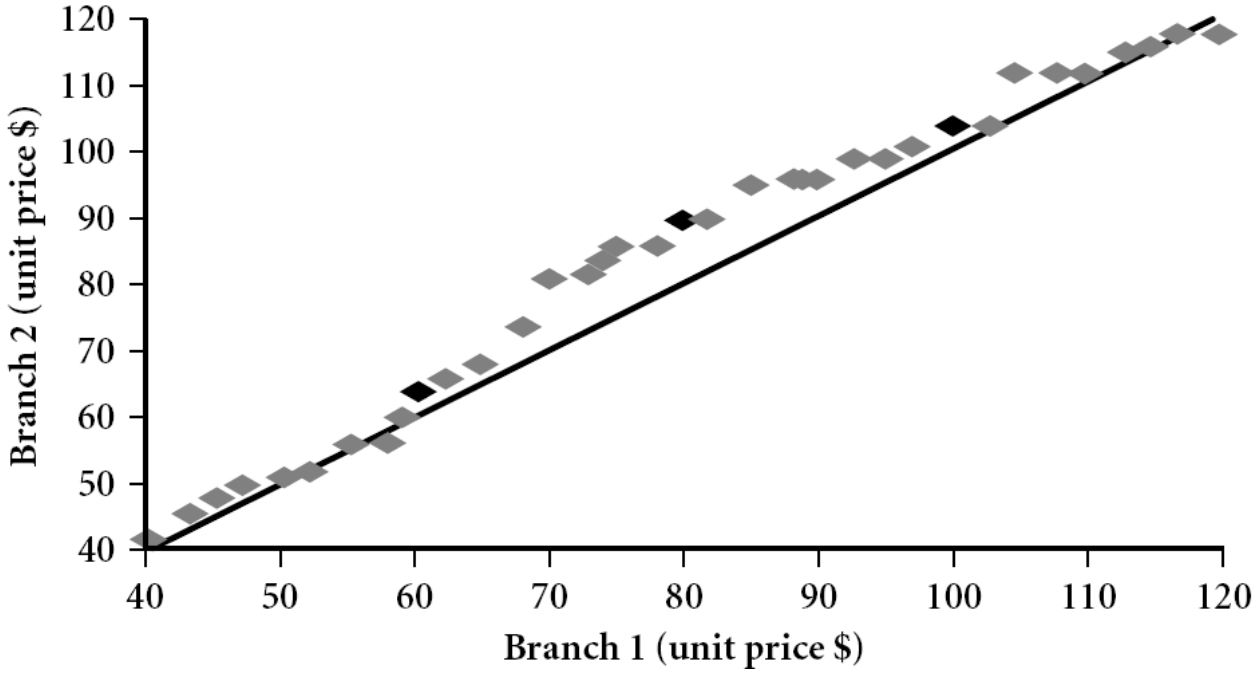
# Quantile Plot

- Displays all of the data (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 100 $f_i$% of the data are below or equal to the value $x_i$



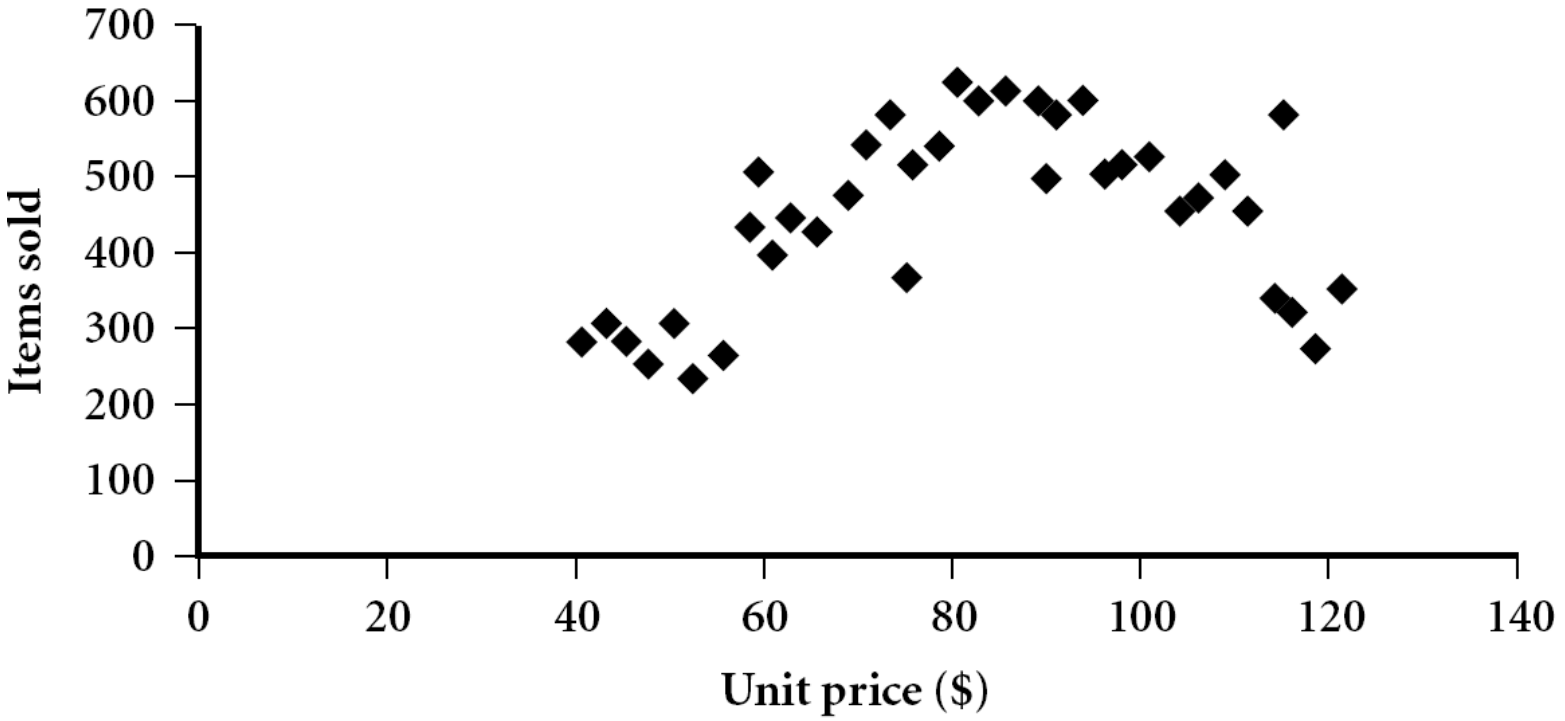
# Quantile-Quantile (Q-Q) Plot

- 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.

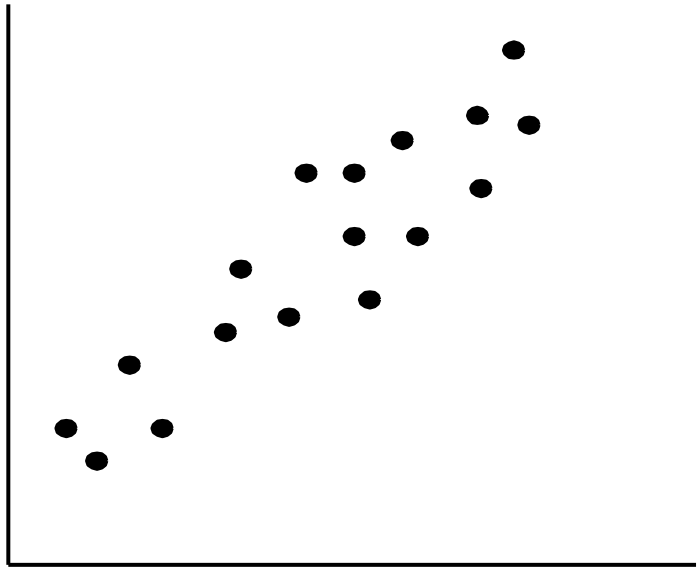


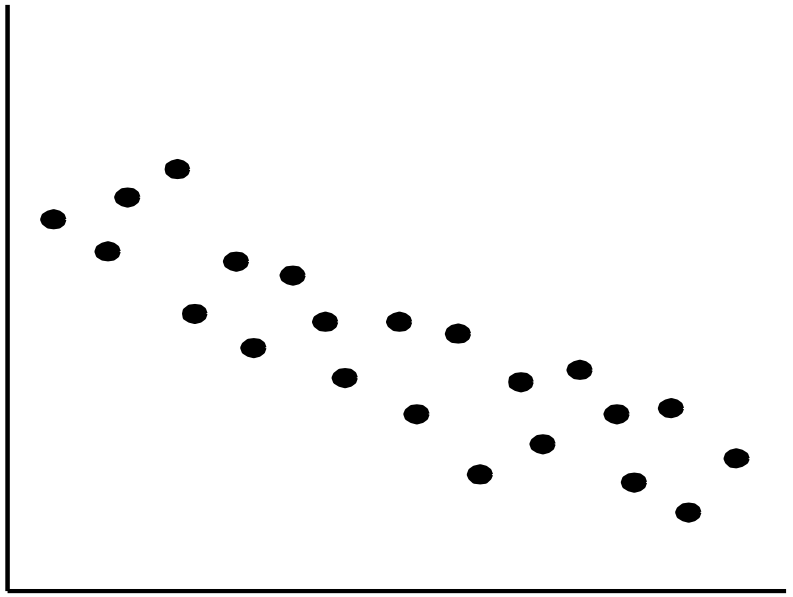
# Scatter plot

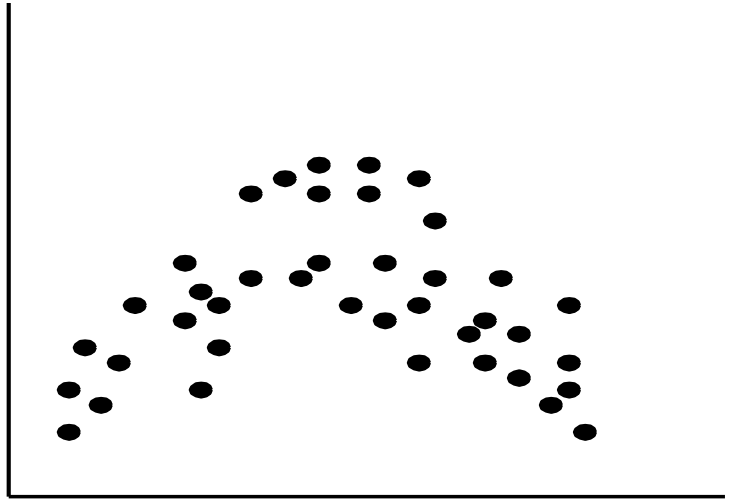
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



# Positively and Negatively Correlated Data

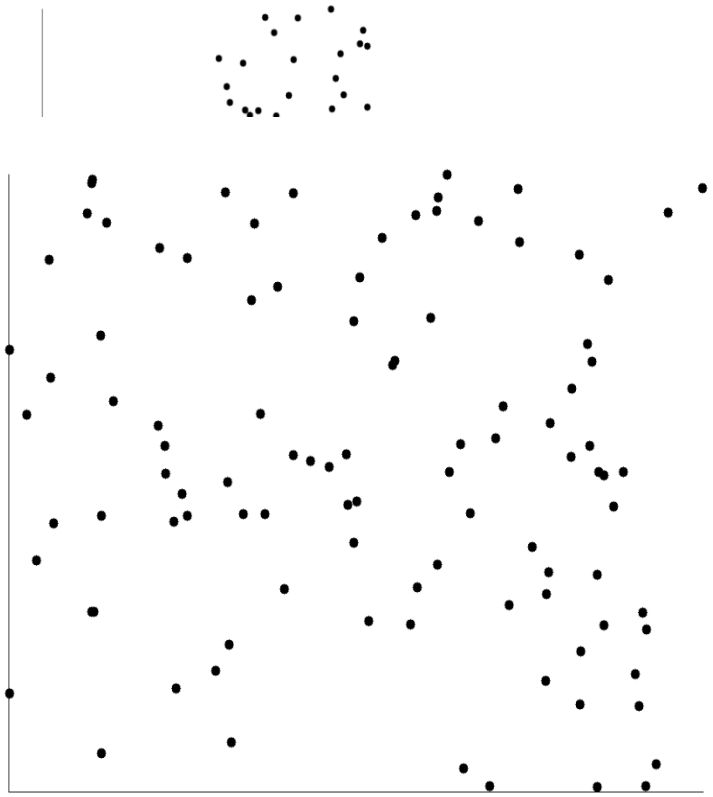


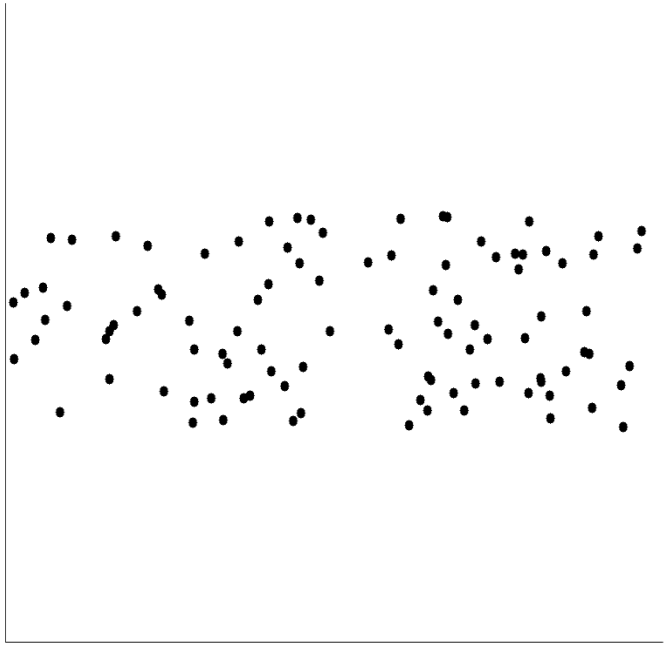




- The left half fragment is positively correlated
- The right half is negative correlated

# Uncorrelat



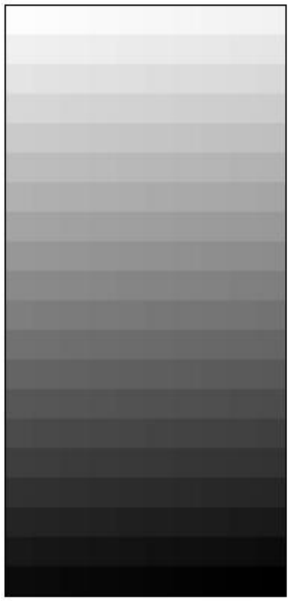


# Data Visualization

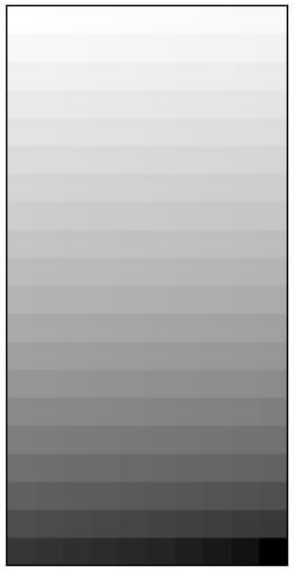
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters 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

# Pixel-Oriented Visualization Techniques

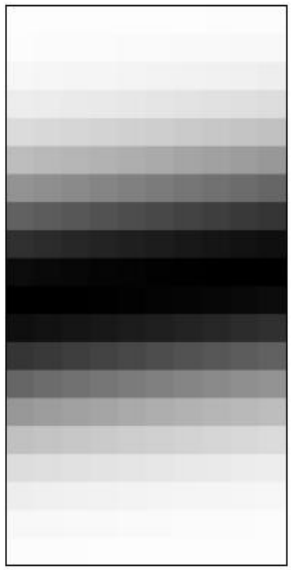
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



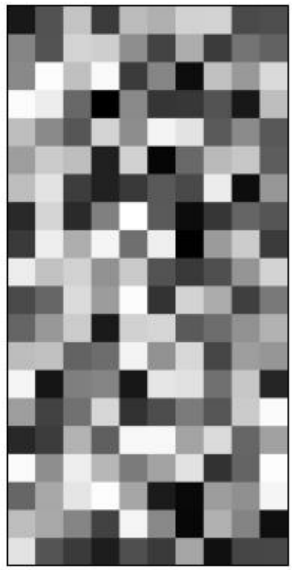
(a) Income



(b) Credit Limit



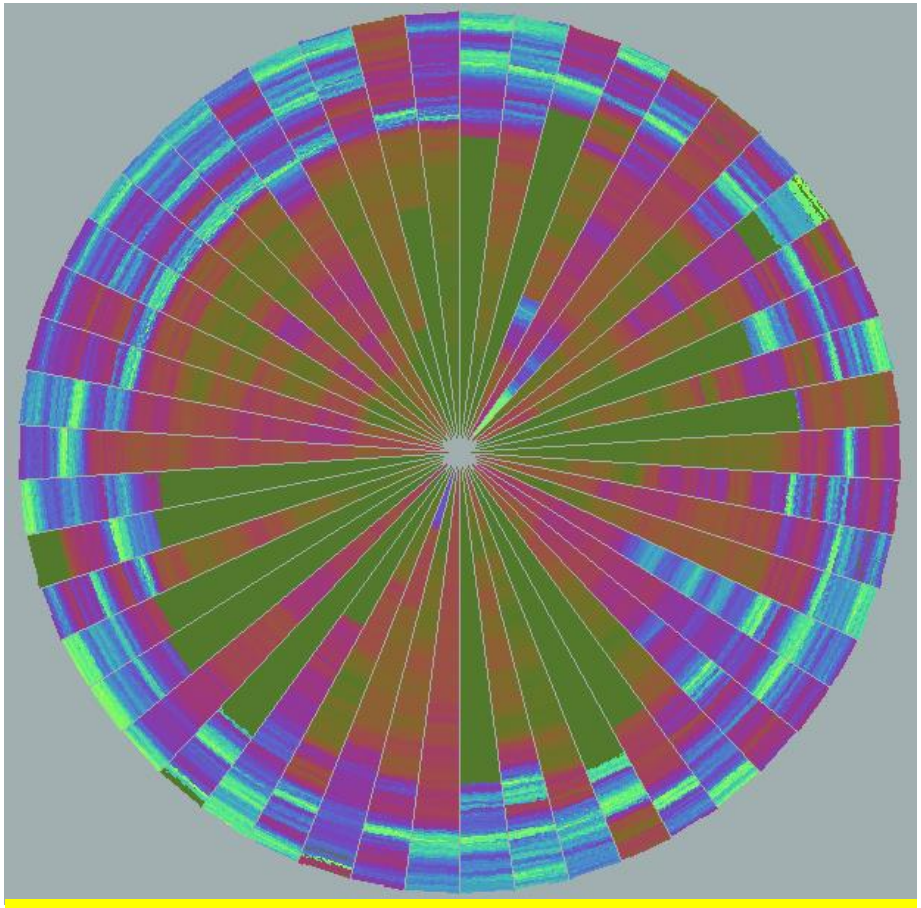
(c) transaction volume



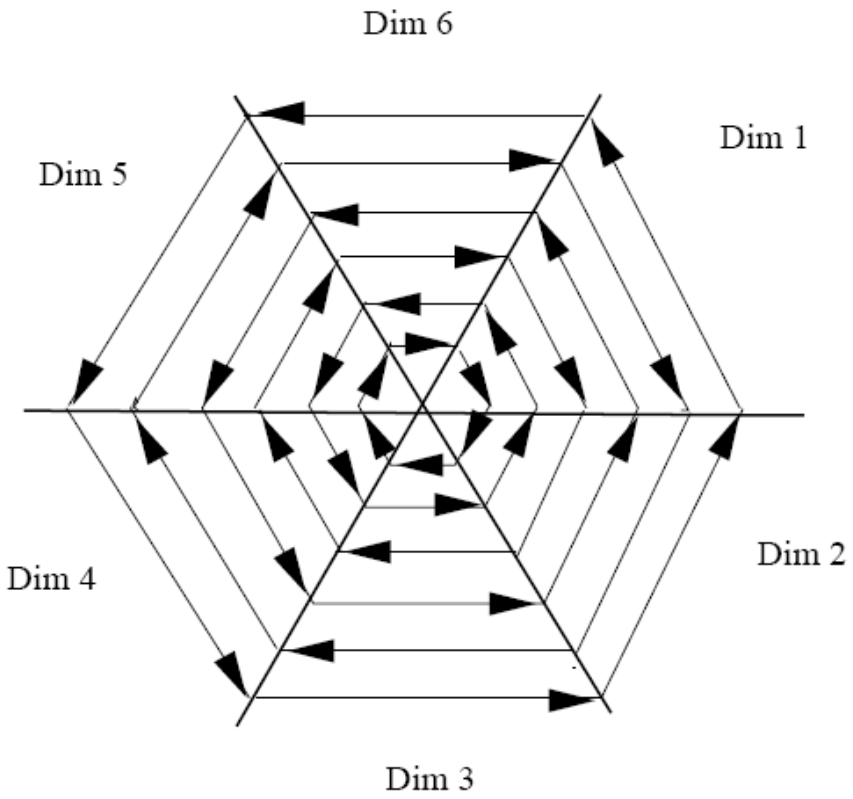
(d) age

# Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



**Representing about 265,000 50-dimensional Data Items with the 'Circle Segments' Technique**

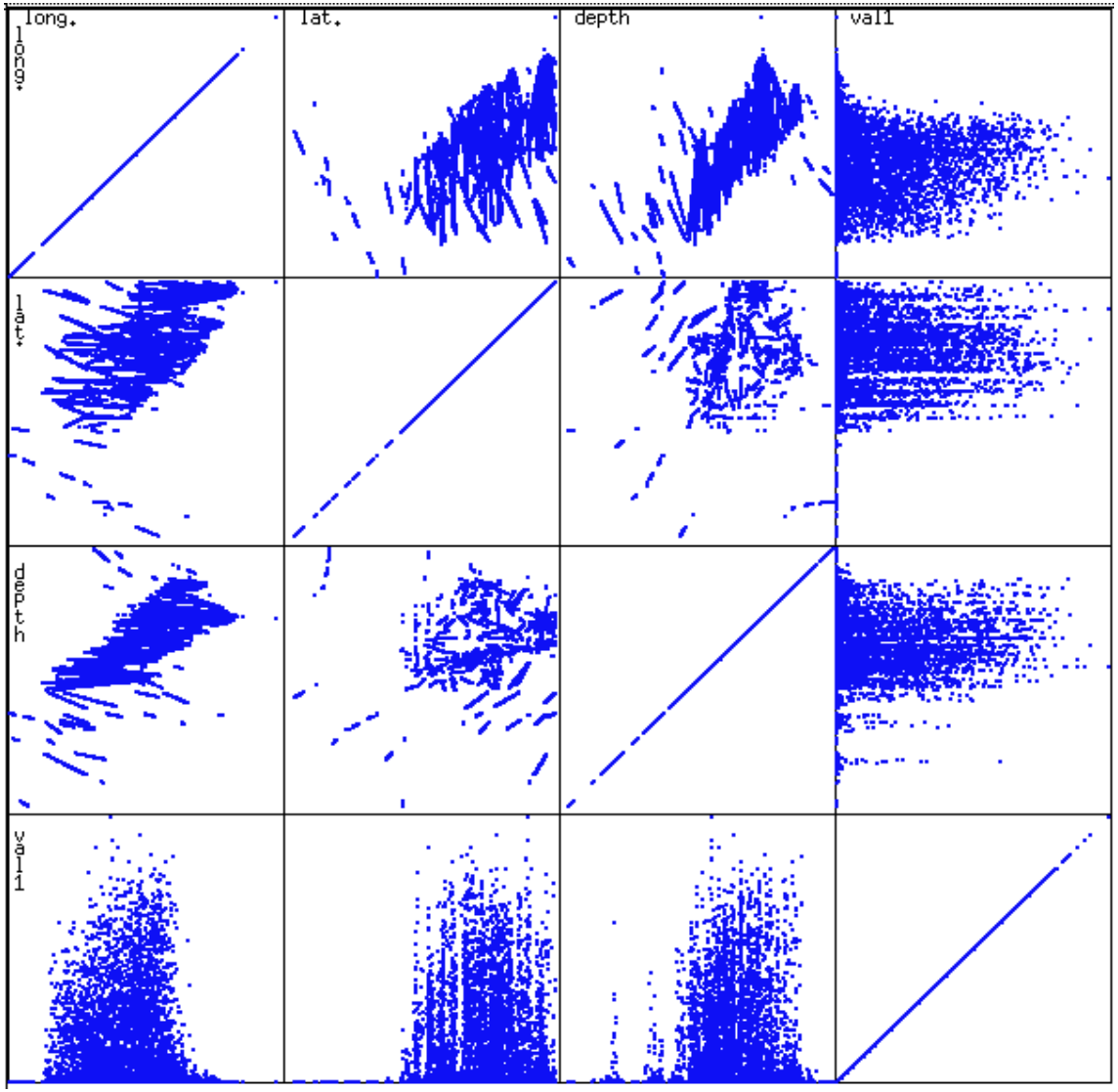



(b) Laying out pixels in circle segment

# Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  - Parallel coordinates

# Scatterplot Matrices

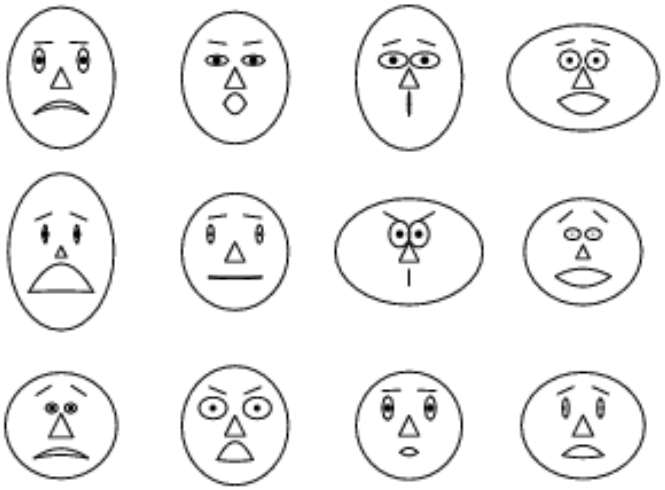





Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

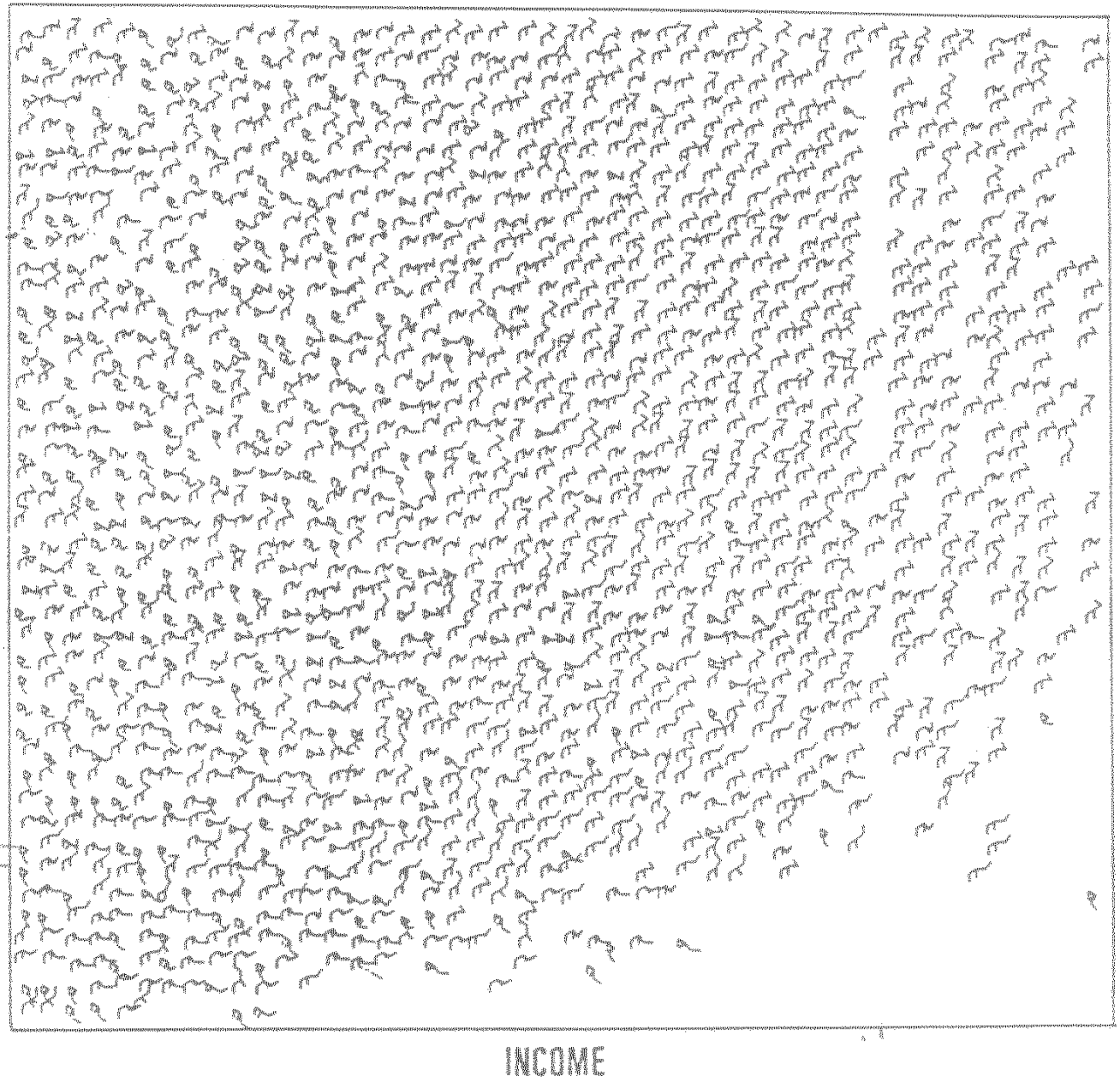
# Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)



- REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics.* New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld*--A Wolfram Web Resource. mathworld.wolfram.com/ChernoffFace.html

# Stick Figure




A census data figure showing age, income, gender, education, etc.

A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

# Hierarchical Visualization Techniques

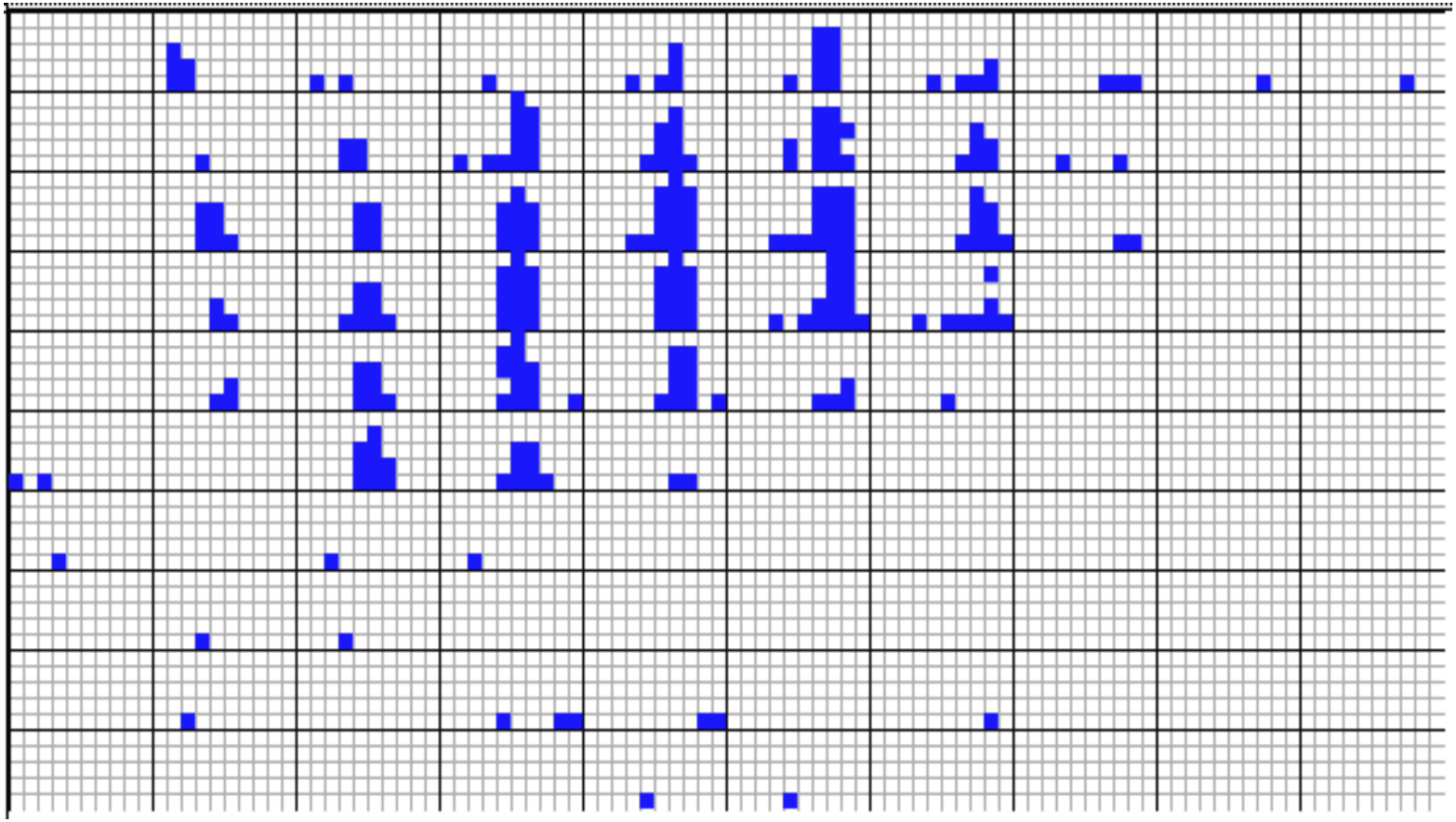
- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube

# Dimensional Stacking

- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

# Dimensional Stacking

Used by permission of M. Ward, Worcester Polytechnic Institute



Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

# Measuring data 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

# Proximity Measure for Nominal Attributes

Method 1: Simple matching

- *m*: number of matches,
- *p*: total number of variables

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

Example:

| Objects | code |
|---|---|
| 1 | Code A |
| 2 | Code B |
| 3 | Code C |
| 4 | Code A |

# Proximity Measure for Nominal Attributes

| Objects | code |
|---|---|
| 1 | Code A |
| 2 | Code B |
| 3 | Code C |
| 4 | Code A |

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

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

- **d(i,j) is 0 if objects I and j match**
- **d(i,j) is 1 if objects I and j differ**

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

# Proximity Measure for Binary Attributes

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distance measure for asymmetric binary variables:

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

Object *j*

| Object *i* | 1 | 0 | sum |
|---|---|---|---|
| 1 | $q$ | $r$ | $q+r$ |
| 0 | $s$ | $t$ | $s+t$ |
| sum | $q+s$ | $r+t$ | $p$ |

q:11
r: 10
s: 01
t: 00

# 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 be 0

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

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

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

Output: results for Jack and mary are more similar

# Dissimilarity of Numeric data

- Distance measure are commonly used for computing the dissimilarity of objects described by numeric attribute

- These measures include the Euclidean, Manhattan and Minkowski distance

# Dissimilarity of Numeric data

- Let i and j are two objects described by p numeric attributes

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

$j = (x_{j1}, x_{j2}, \ldots, x_{jp})$

- The Euclidean distance between objects I and j is defined as

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

# Dissimilarity of Numeric data

- Manhattan distance

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

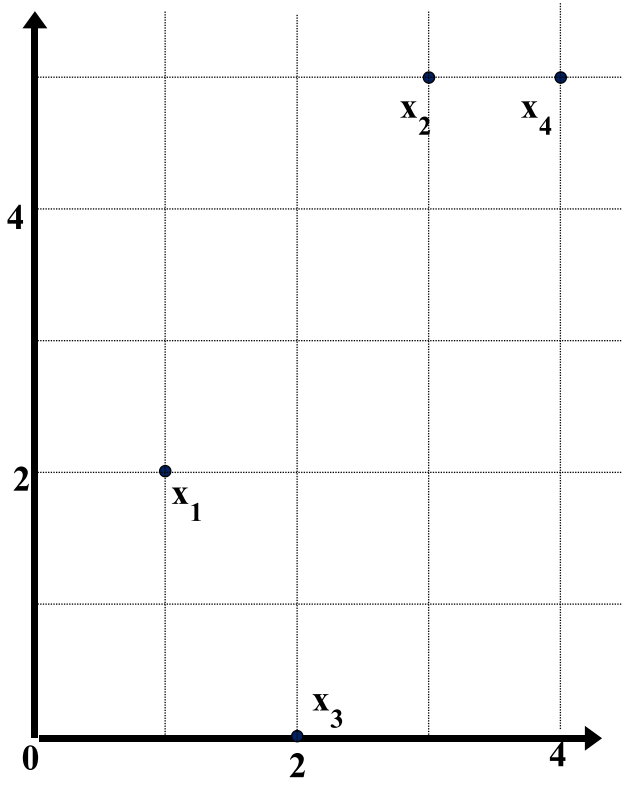
- Minkowski distance

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

# Example: Data Matrix and Dissimilarity Matrix

## Data Matrix

| 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 |

d(x1,x1)=(1-1)+(2-2)=0
d(x2,x1)=(3-1)+(5-2)=5
d(x3,x1)=(2-1)+(0-2)=3
d(x4.x1)=(4-1)+(5-2)=6

# 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.

| *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 |

- 0 value means that document do not share the word
- But this does not make them similar
- We need a measure that will focus on the words that the two documents do have in common.
- Cosine similarity is a measure of similarity that can be used to compare documents

# Example: Cosine Similarity

- Cos ($x$, $y$) =sim (x, y)= ($x \bullet y$) /||x|| ||y||

  → where • indicates vector dot product,

  →||$x$||: $(x_1*x_1+x_2*x_2+.....+x_p*x_p)^{0.5}$

- Ex: Find the **similarity** between documents 1 and 2.

  $d_1$ = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
  $d_2$ = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

  $d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25

  $||d_1||$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5}$ = 6.481
  $||d_2||$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5}$ = 4.12

  cos(d1,d2)=25/(6.481*4.12)
  cos($d_1$, $d_2$ ) = 0.94

# Cosine Similarity

- Cosine value of zero means that two vectors are at 90 degree to each other and have no match

- The closer the cosine value to 1, the smaller is the angle and greater is the match between vectors

# 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

# 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

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

  - $f$ is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - $f$ is numeric: use the normalized distance
  - $f$ is ordinal
    - Compute ranks $r_{if}$ and
    - Treat $z_{if}$ as interval-scaled

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

# Chapter 2: Getting to Know Your Data

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