Data Mining

Know your Data

Getting to Know Your Data

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

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
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color = {black, blond, brown, grey, red, white*}{0,1,2…}
  - marital status, occupation, ID numbers (numbers but math operations can't be done on them), zip codes
  - No mean and median but mode can be used
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - 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**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large*}*,* grades, army rankings

Numeric Attribute Types

- Quantitative (integer or real-valued). Can do math on them i.e. mean, median and mode, etc.
- **Interval Scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C°or F°, calendar dates*
  - No true zero-point. 0 ° *C is not showing "no temp"*
- **Ratio Scaled**
  - Inherent **zero-point**
    - e.g., *length, counts, monetary quantities, years of experience, word counts, weight, height etc.*

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
  - Attributes Hair_Color, Smoker, Med Test, Drink_Size each have a finite number of values, thus are discrete.
  - Discrete attributes may have numeric values 0 and 1 for binary attributes
  - Age have values from 0 to 110
  - Customer_ID is countably infinite
  - Zip codes

Discrete vs. Continuous Attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

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

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals

Measuring the Central Tendency

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

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

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

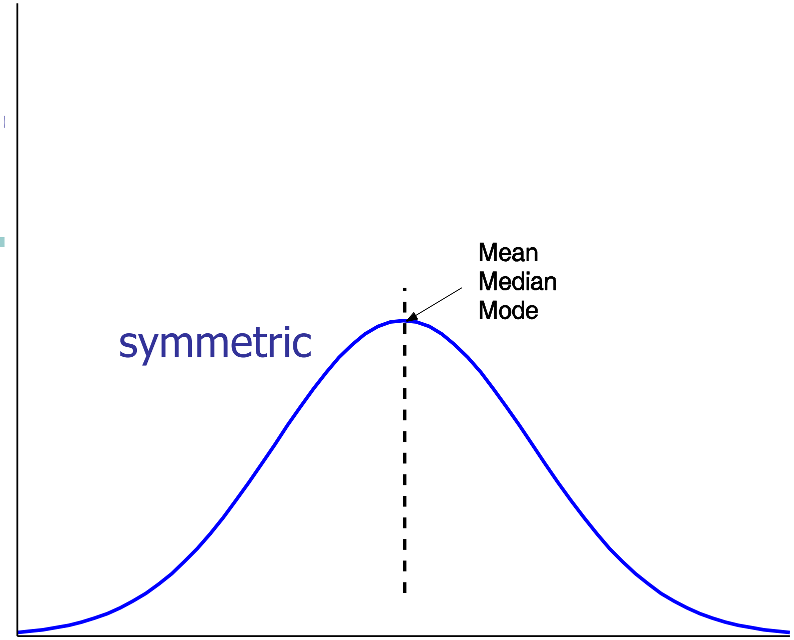
  - Weighted arithmetic mean:

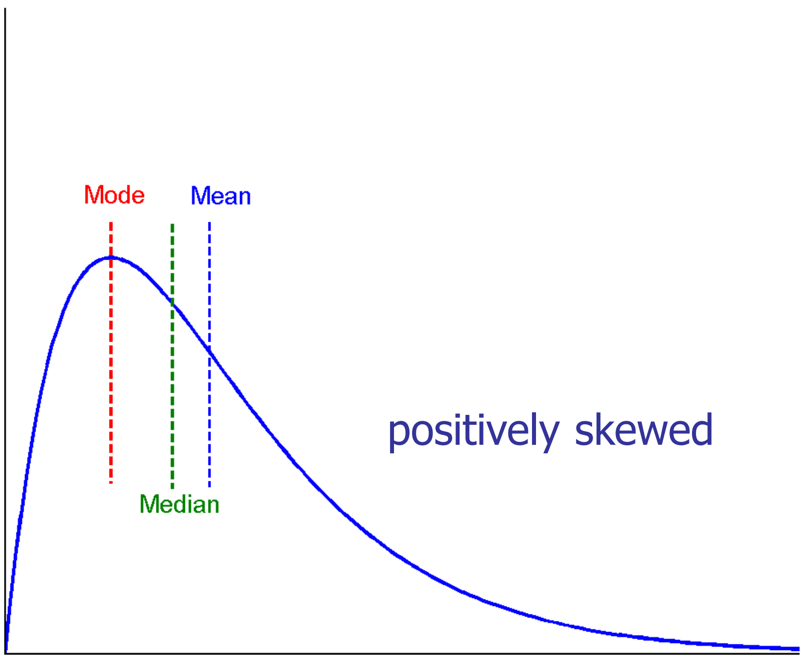
    $$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

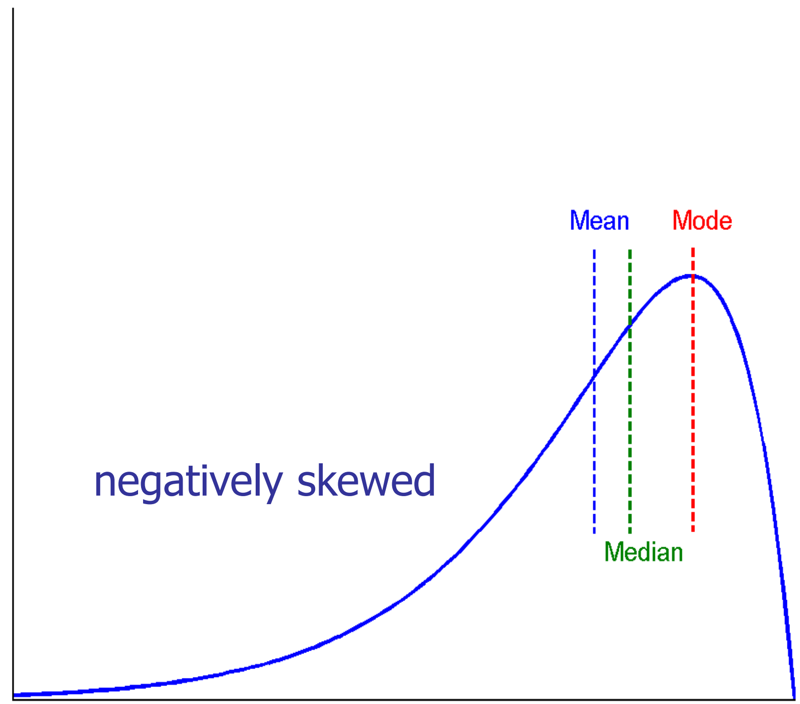
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $$mean - mode \approx 3 \times (mean - median)$$
  - Mode for unimodal frequencies can be approximated if mean and median values are known

Symmetric vs. Skewed Da

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



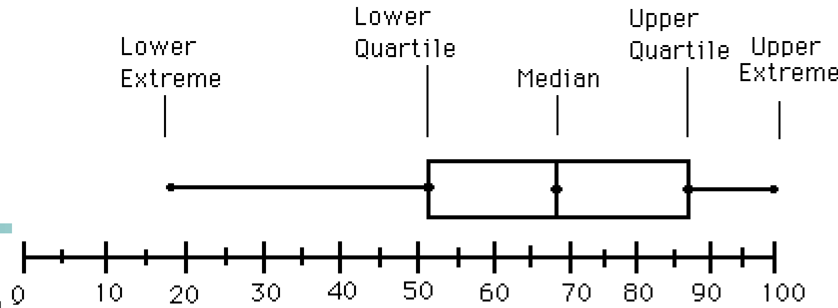




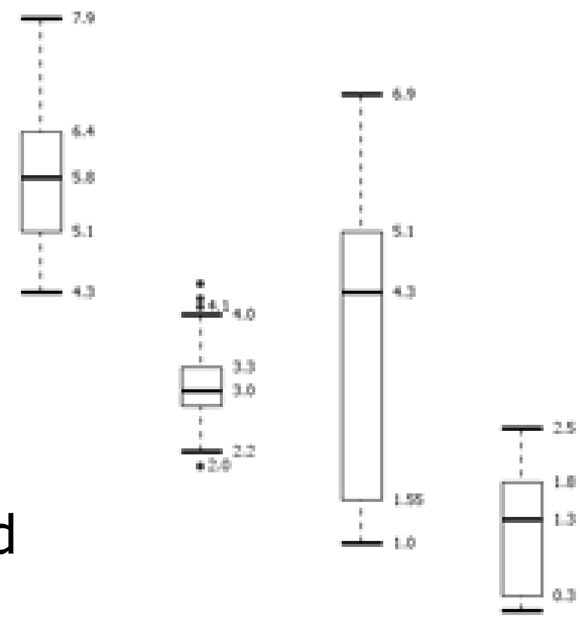
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value lower/higher
    - than o = (1.5 x IQR) of (Q1-o)/(Q3+o)

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



Exercise 1

- Given the following data
- 7, 8, 7, 10, 7, 2, 8, 2, 7, 30
- Median =
- Q1 (index=2.5~3) =
- Q3 (index=7.5~8) =
- IQR = Q3 – Q1
- 5 number summary =
- Outliers based on IQR =

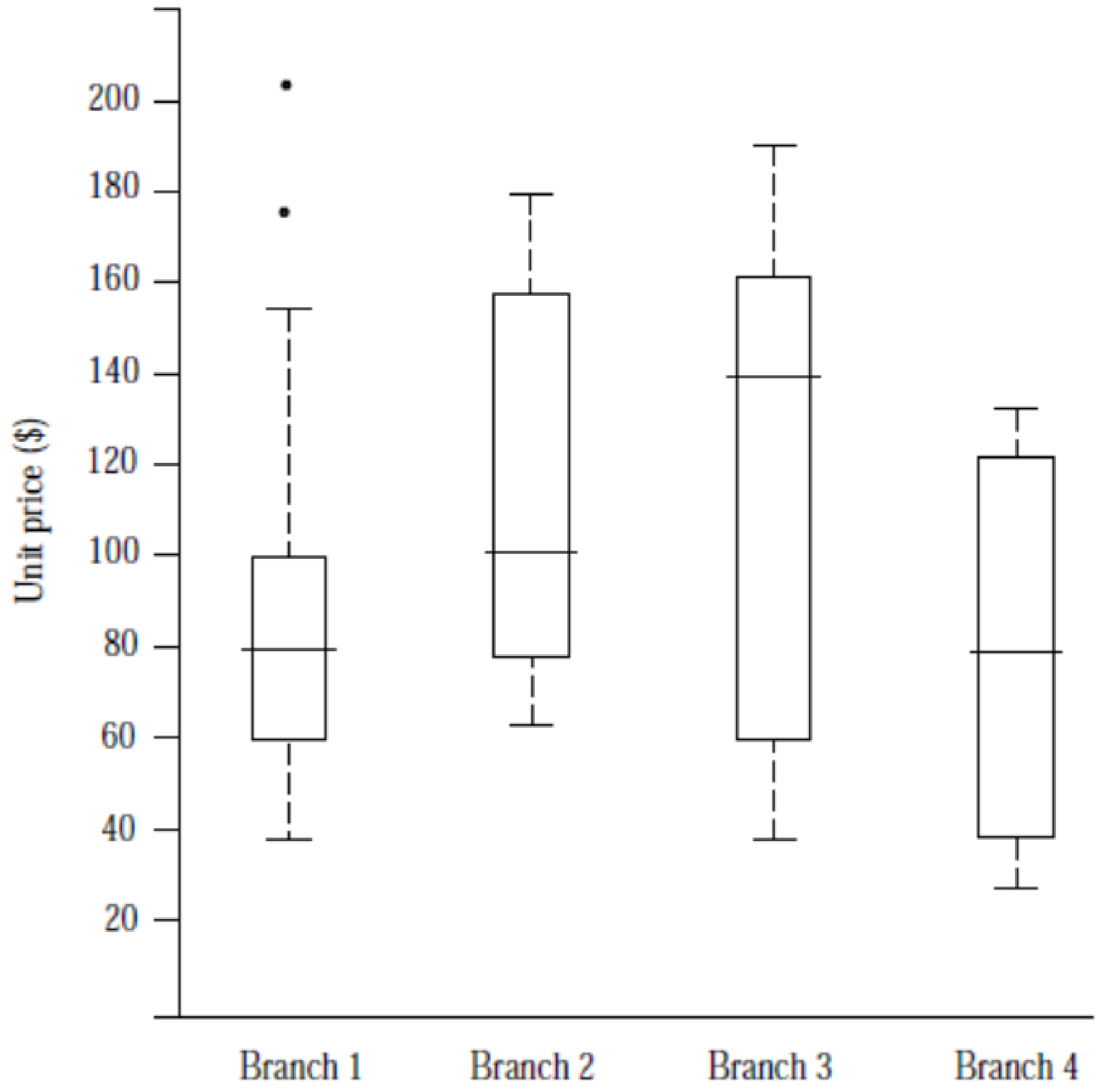
Exercise 1 – Solution

- Given the following data
- 7, 8, 7, 10, 7, 2, 8, 2, 7, 30
- Sort the data 2,2,7,7,**7,7**,8,8,10,30
- Median = 7
- Q1 (index=10*.25=2.5~3) = 7
- Q3 (index=10*.75=7.5~8) = 8
- IQR = Q3 – Q1 = 8 – 7 = 1
- **Five number summary**: min, Q_1, median, Q_3, max
- 5 number summary = 2, 7, 7, 8, 30
- Outliers based on IQR =
  - IQR * 1.5 =1* 1.5=1.5 = x
  - Values which are less than (Q1 – x) Or greater than (Q3 + x)
  - Q1 – 1.5 = 7 – 1.5 = 5.5
  - Q3 + 1.5 = 8 + 1.5 = 9.5
  - Outliers = 2, 10, 30

Exercise 2

- Find the outliers in the following using IQR
  - 2,2,3,4,7,7,8,9,10,30
- Find the outliers based on IQR
- Q1 = 3
- Q3 = 9
- IQR = Q3 – Q1 = 9 – 3 = 6
- x = 1.5 * IQR = 9
- Outliers < Q1 – x = 3 – 9 = -6
- Outliers > Q3 + x = 9 + 9 = 18
- Outliers = 30

Boxplot Analysis



Lab Task 1

- Import data sales_data.csv
- See the metadata view to check the attribute type, statistics (mean, mode, etc.), range and no of missing values.
- Create Boxplot for suitable fields

Python Hint

- #Import Basic Libraries
- import numpy as np # linear algebra
- import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
- import matplotlib.pyplot as plt #data visualization
- import seaborn as sns #data visualization

Variance and Standard Deviation

- Variance and standard deviation
- **Variance**: (algebraic, scalable computation)

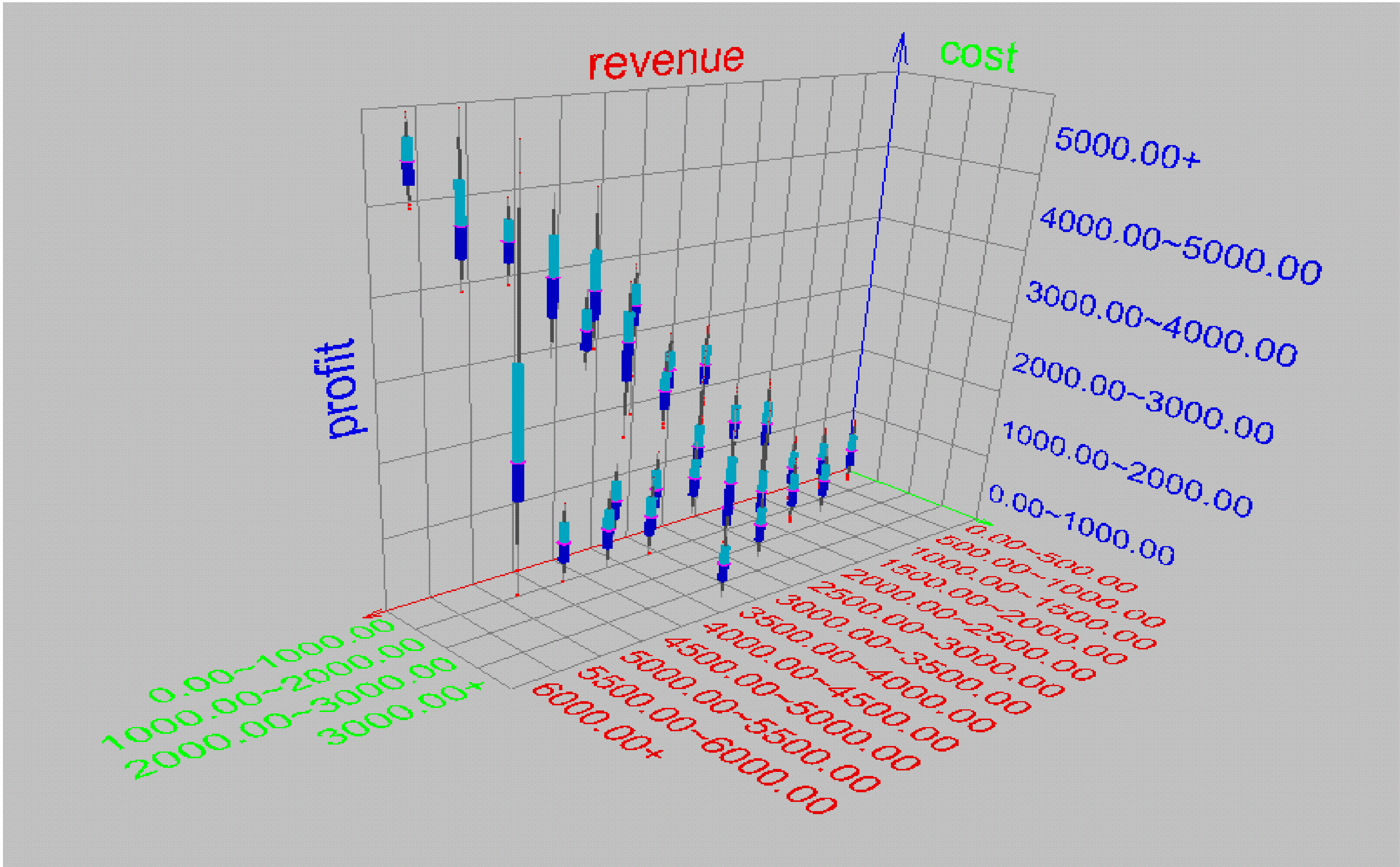
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2) - \bar{x}^2,$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 $(or\ \sigma^2)$
- Variance and standard deviation are measures of data dispersion.
- Low SD means observations are close to the mean
- High SD means the data are spread out over a large range of values

Exercise 3: Find Variance and SD

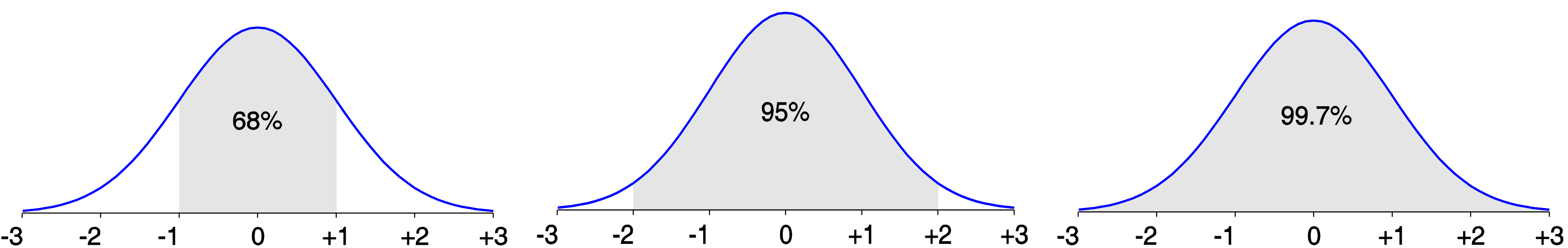
- 5, 10, 15
- N = 3
- Mean = (5+10+15)/3 = 10
- Var = ((5-10)^2 + (10-10)^2 + (15-10)^2) / N
- = (-5^2 + 0^2 + 5^2) / 3
- = (25 + 0 + 25) / 3 = 50/3 = 16.7
- StDev = sqrt(Var) = sqrt(16.7) ~ 4

Visualization of Data Dispersion: 3-D Boxplots



Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



Lab Task2

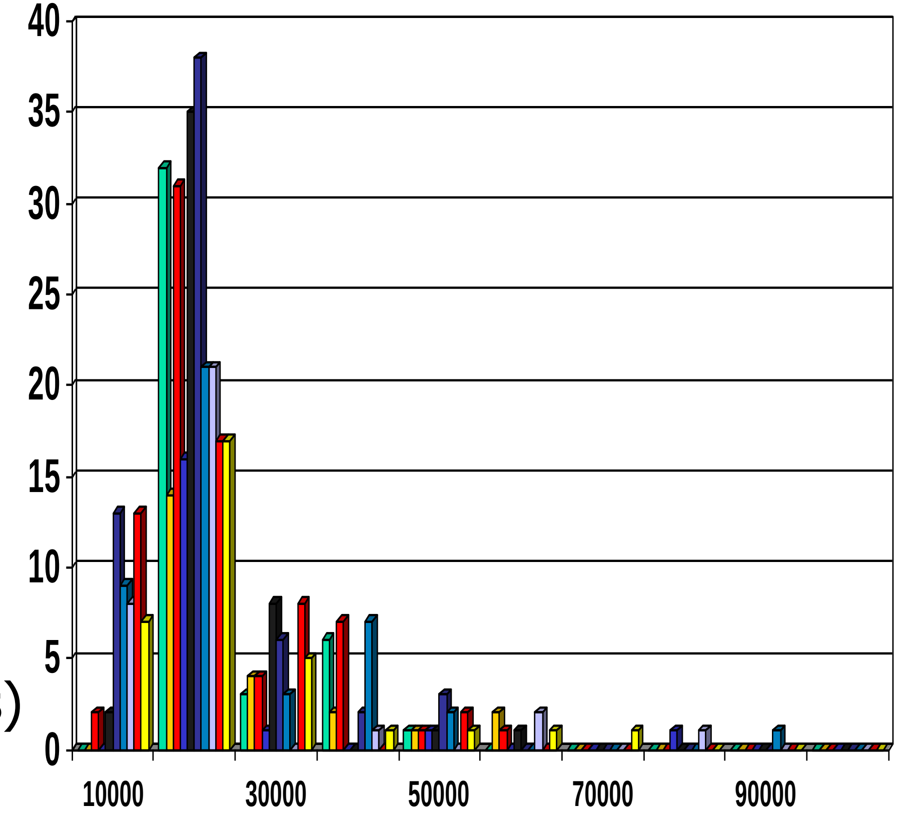
- Find normal distribution curve of usable attributes from your sales dataset

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **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
- Bar chart represents categorical data while histogram represents quantitative data
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

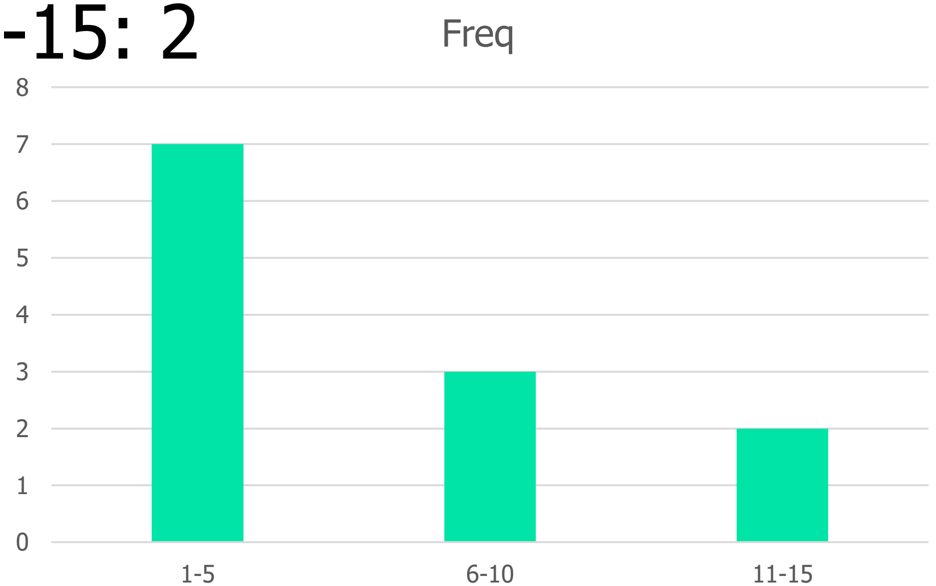


Exercise 4

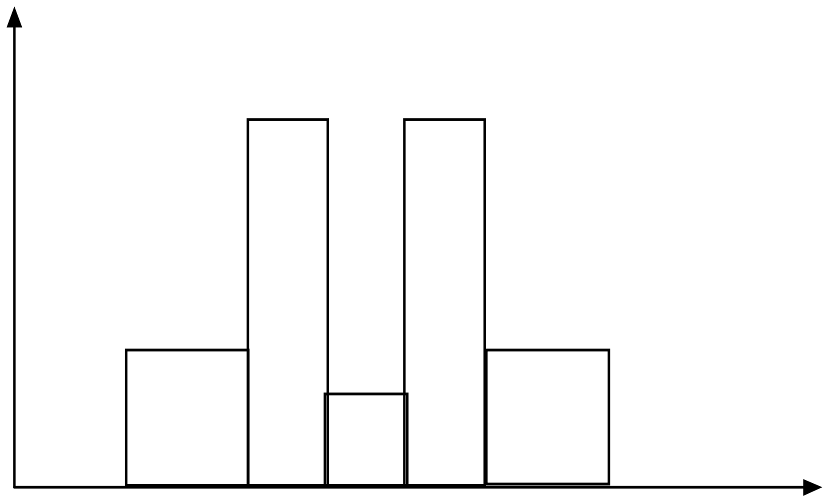
- Given the ages of the children:
  - 9,7,12,10,5,4,8,2,4,3,1,2,8,14
- Convert the data into ranges
- Count the frequency for each range
- Create a histogram

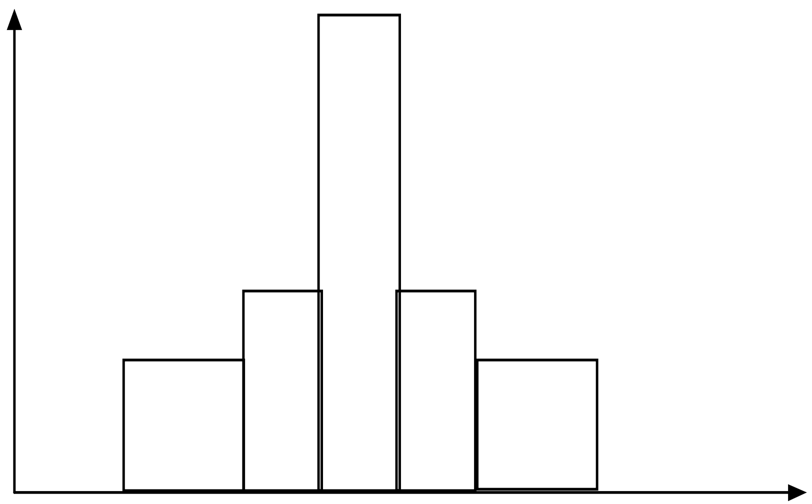
Exercise 4

- Given the ages of the children:
  - 9,7,12,10,5,4,8,2,4,3,1,2,8,14
- Convert the data into ranges
  - 1-5,6-10,11-15
- Count the frequency for each range
  - 1-5: 7, 6-10: 3, 11-15: 2
- Create a barchart



Histograms Often Tell More than Boxplots





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

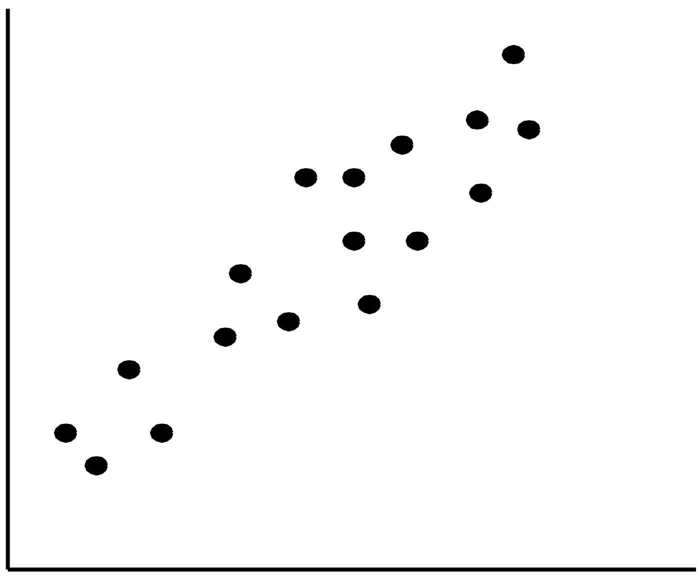
Lab work

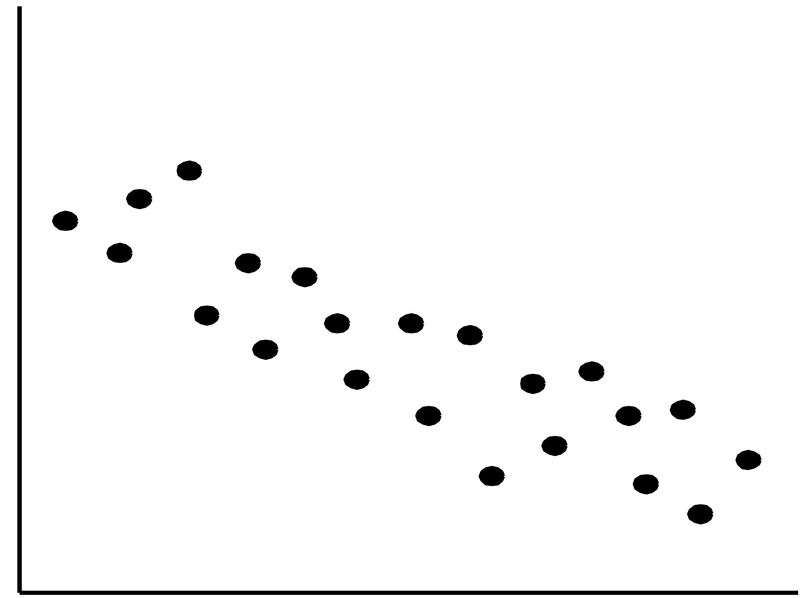
- Create a histogram of amount
- Create a histogram of single_price
- Create a bar chart of product_category and single_price

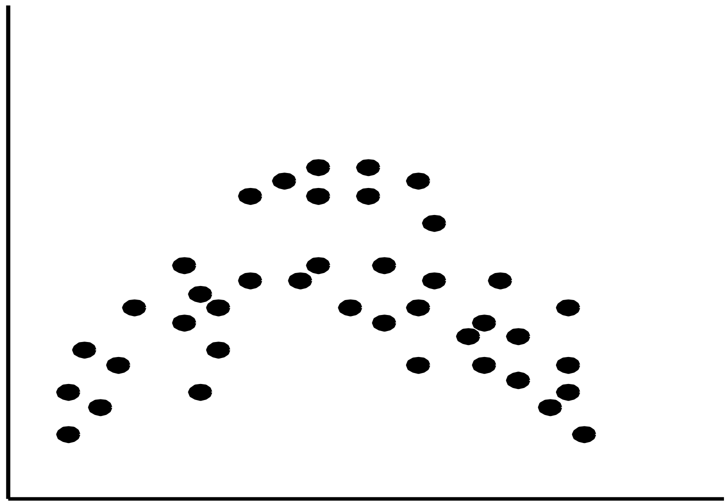
Scatter plot

- Provides a first look at bivariate data (involving two attributes) 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
- Helps in finding if there is a correlation between two attributes

Positively and Negatively Correlated Data

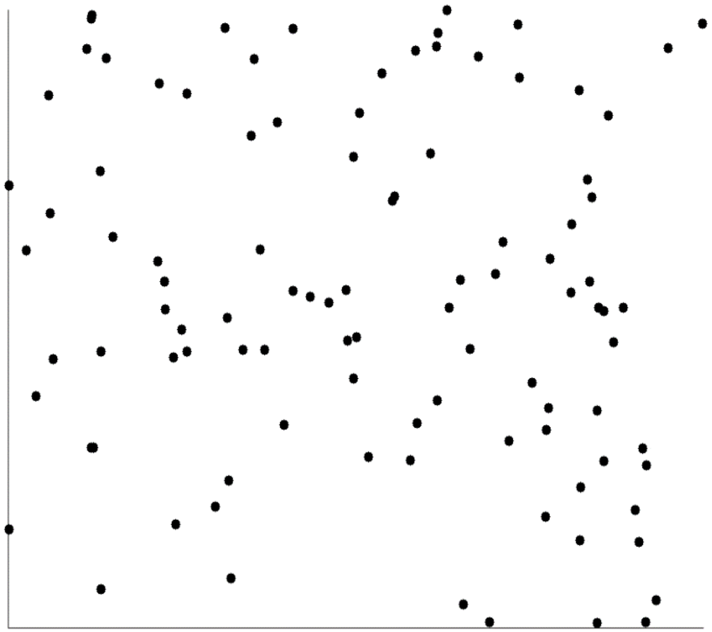


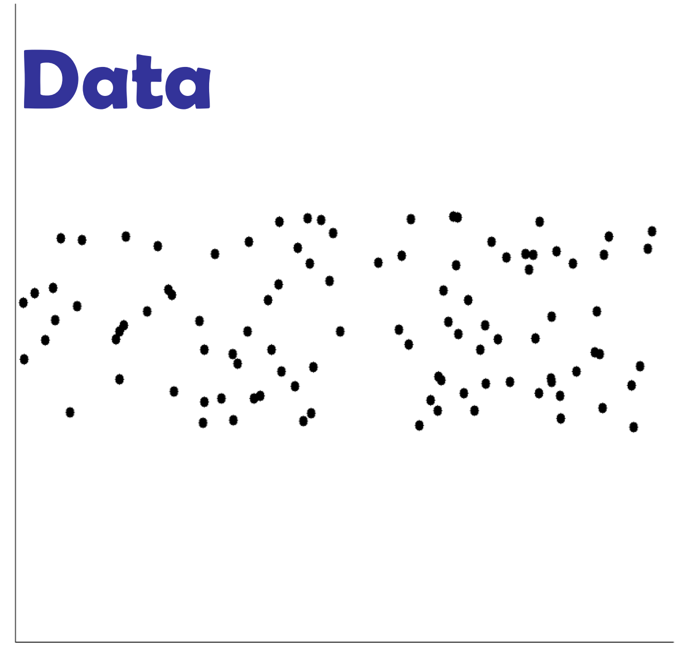


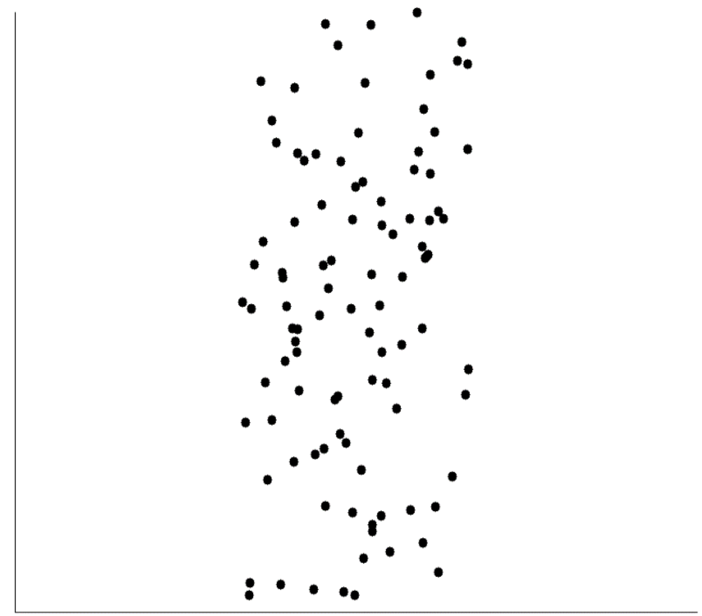


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







Exercise 6

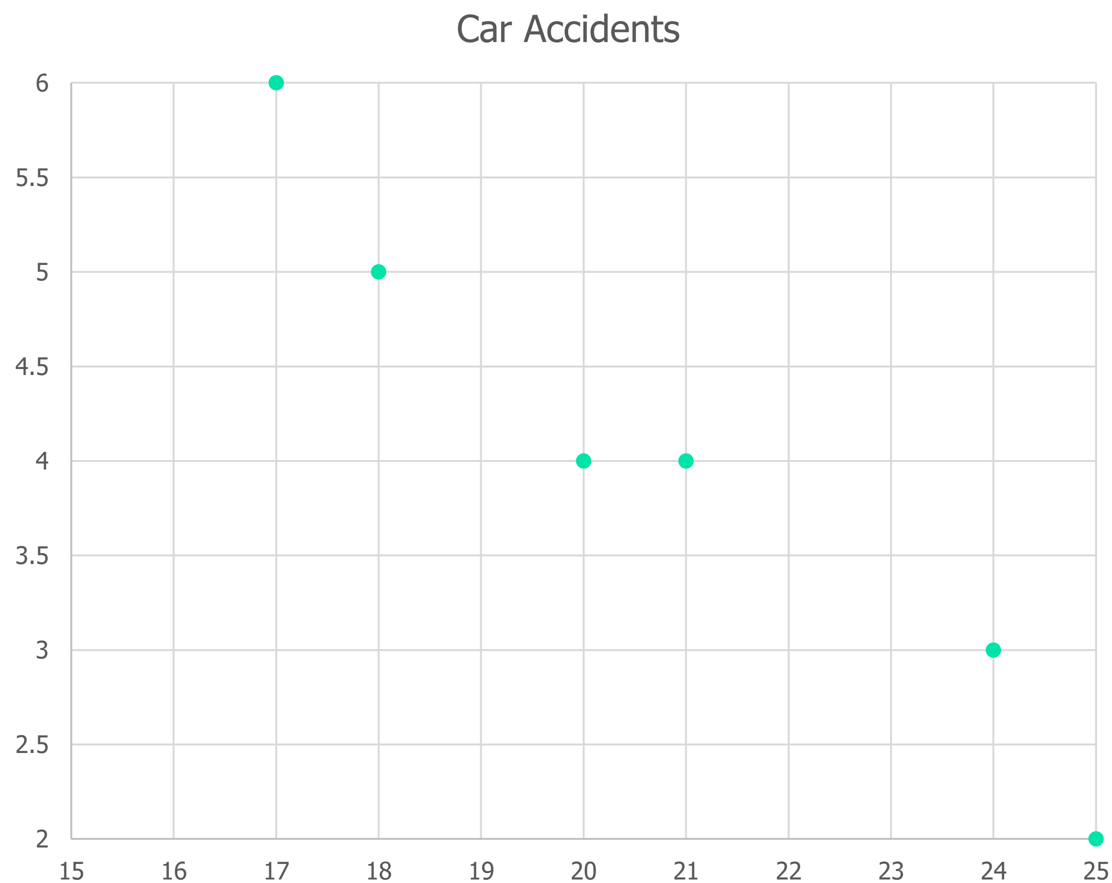
- Create a scatter plot for the following data

Age	Car Accidents
17	6
21	4
18	5
25	2
20	4
24	3

Exercise 6

- Create a scatter plot for the following data

Age	Car Accidents
17	6
21	4
18	5
25	2
20	4
24	3



Chapter 2: Getting to Know Your Data

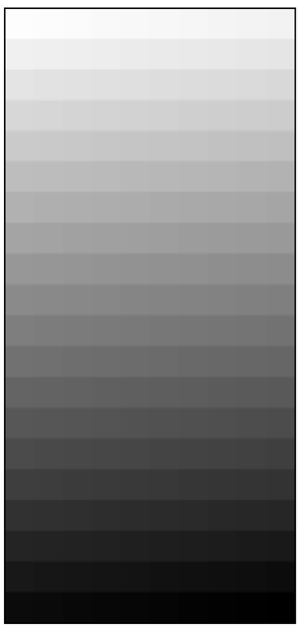
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

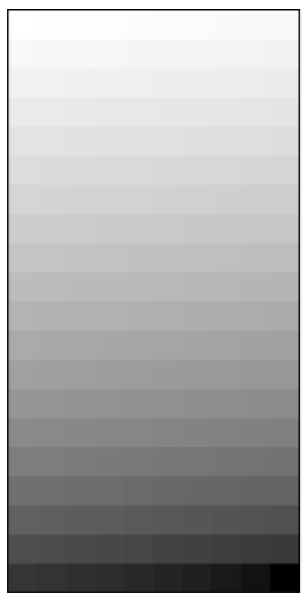
Data Visualization

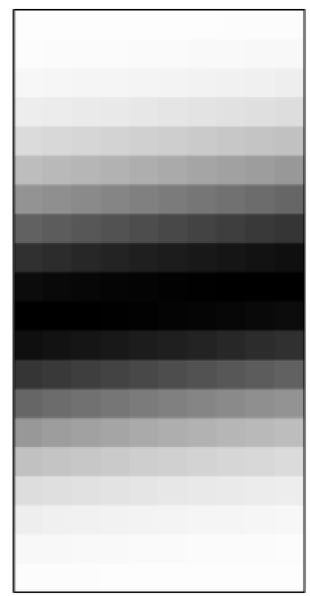
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters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

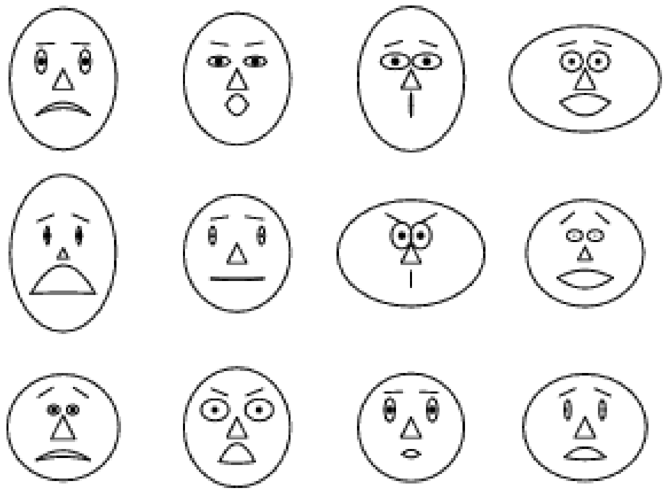

(d) age

Icon-Based Visualization Techniques

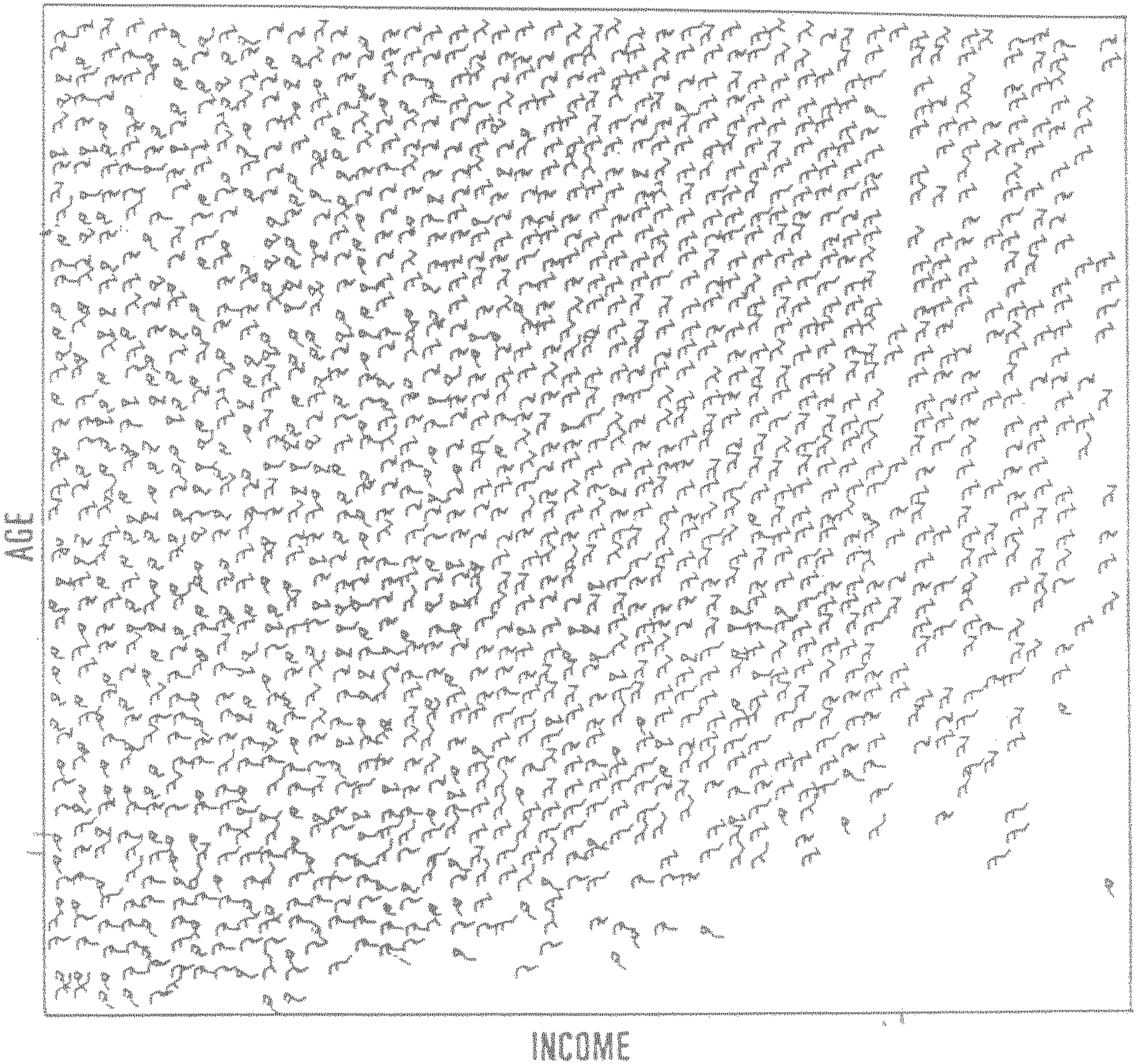
- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding

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.

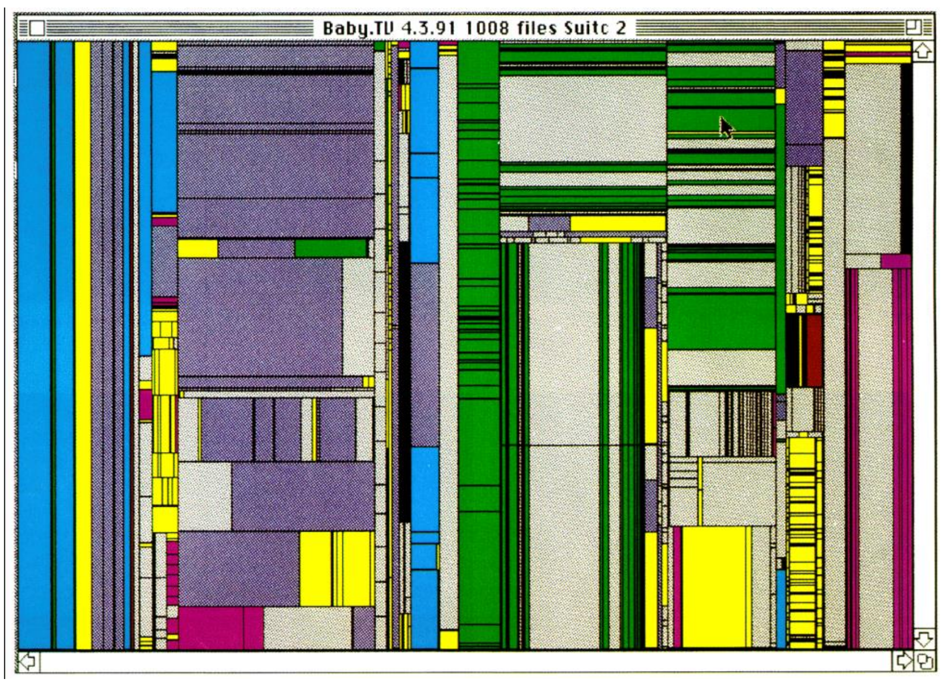
A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

Hierarchical Visualization Techniques

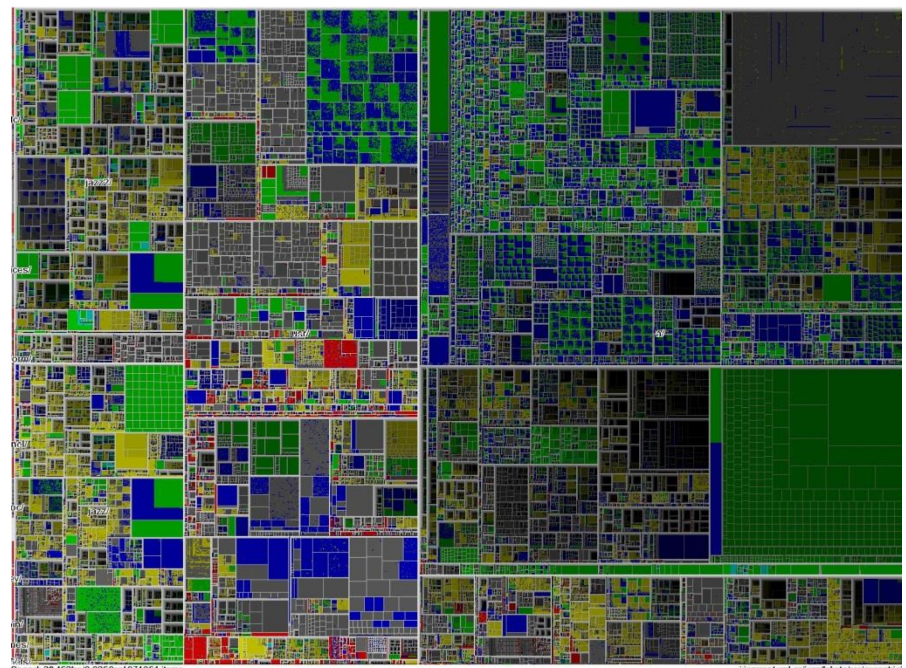
- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Tree-Map

Tree-Map

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values



Schneiderman@UMD: Tree-Map of a File System

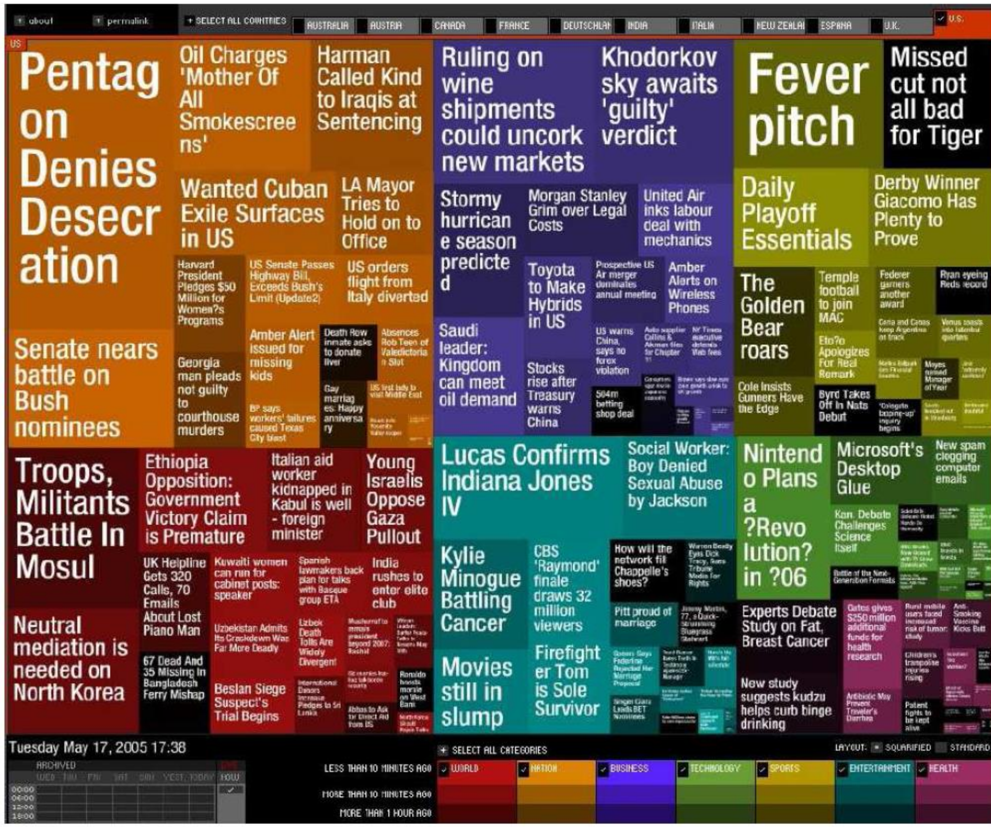


Schneiderman@UMD: Tree-Map to support large data sets of a million items

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
  - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks

World Population Tag Cloud



Newsmap: Google News Stories in 2005

Assignment 1

- Find appropriate datasets from kaggle for the following visualization techniques and apply them
  - Pixel-oriented visualization
  - Chernoff faces
  - Stick figures
  - Tree map

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

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
- **Proximity** refers to a similarity or dissimilarity

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

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

Dissimilarity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the *M* nominal states. E.g. to encode nominal attr 'color', a binary attr can be created for each of the colors listed. Yellow will have 1, others 0

Exercise 7

- Find the distance based on the attributes "Favorite Color" and "Favorite Food" between
  - Ali and Bilal
  - Ali and Faris

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grad e
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

Exercise 7

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

- Find the distance based on the attributes "Favorite Color" and "Favorite Food" between
  - Ali and Bilal
  - (2-1)/2= 0.5
  - Ali and Faris
  - (2-0)/2 = 1

	Favorite Color	**Favorite Food**	**Plays Chess**	**Plays Football**	**Age**	**Salary (1000s)**	**Grad e**
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Dissimilarity measure for symmetric binary variables:

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

- Dissimilarity measure for asymmetric binary variables:

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

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

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

Dissimilarity between Binary Variables

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Example

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

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

  - d(jack,mary), d(jack,jim)
  - d(jim,mary)
  - Let the values Y and P be 1, and the value N 0

Dissimilarity between Binary Variables

- Example

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

  - Let the values Y and P be 1, and the value N 0

		Mary 1	Mary 0	Σ_{row}
Jack	1	2	0	2
Jack	0	1	3	4
	Σ_{col}	3	3	6

		Jim 1	Jim 0	Σ_{row}
Jack	1	1	1	2
Jack	0	1	3	4
	Σ_{col}	2	4	6

		Mary 1	Mary 0	Σ_{row}
Jim	1	1	1	2
Jim	0	2	2	4
	Σ_{col}	3	3	6

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

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

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

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

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Exercise 8

- Find the distance based on the "symmetric" attributes "Plays Chess" and "Plays Football" between
  - Ali and Bilal
  - Ali and Faris

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grade
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Exercise 8

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

- Find the distance based on the "symmetric" attributes "Plays Chess" and "Plays Football" between
  - Ali and Bilal
    - dist(Ali,Bilal) = 0/2 = 0
  - Ali and Faris
    - dist(Ali,Faris) = 2/2 = 1

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grade
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

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 is also called L-h norm)

Special Cases of Minkowski Distance

- h = 1: Manhattan (L_1 norm) distance

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

- h = 2: (L_2 norm) Euclidean distance

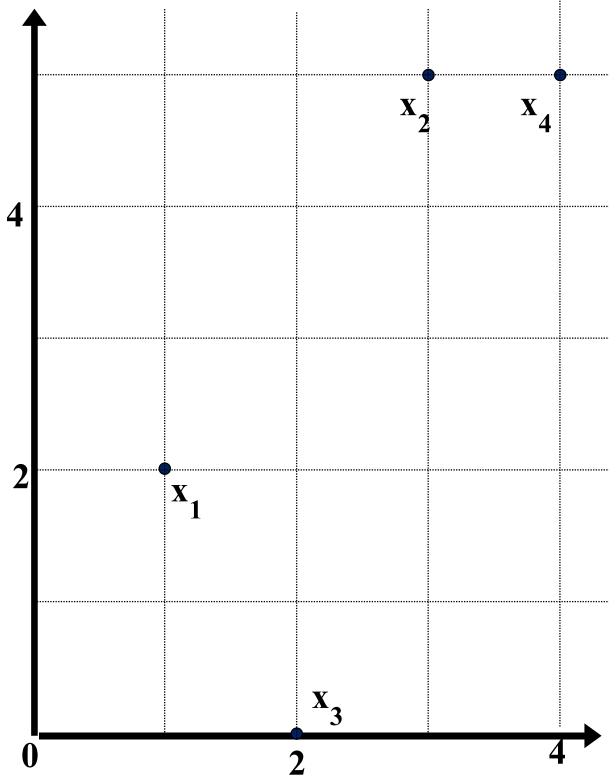
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$ "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

Exercise 9

- Find the distance based on the attributes “Age” and “Salary” between
  - Ali and Bilal (Euclidean)
  - Ali and Bilal (Supremum)

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grad e
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

Exercise 9

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

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

- Find the distance based on the attributes "Age" and "Salary" between
  - Ali and Bilal (Euclidean)
    - dist(Ali,Bilal) = $\sqrt{5^2 + 9^2} = \sqrt{106}$
  - Ali and Bilal (Supremum)
    - dist(Ali,Bilal) = max(5,9) = 9

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grade
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

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 attribute by

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

  - compute the dissimilarity using methods for interval-scaled variables.

Exercise 10

- Find the distance based on the attributes “Age”, “Salary”, and “Grade” between
  - Ali and Bilal (Euclidean)
  - Ali and Bilal (Supremum)
  - Set A=1, B=2, C=3

	Favorite Color	**Favorite Food**	**Plays Chess**	**Plays Football**	**Age**	**Salary (1000s)**	**Grade**
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes	Yes	25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

Exercise 10

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

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

- Find the distance based on the attributes "Age", "Salary", and "Grade" between
  - Ali and Bilal (Manhattan)
    - dist(Ali,Bilal) = |20-25|+|34-25|+|1-0.5|=14.5
  - Ali and Bilal (Supremum)
    - dist(Ali,Bilal) = max(|20-25|,|34-25|,|1-0.5|)=9

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

	Favorite Color	**Favorite Food**	**Plays Chess**	**Plays Football**	**Age**	**Salary (1000s)**	Grade	Grade (N)
Ali	Blue	Cake	Yes	Yes	20	34	C	1
Bilal	Yellow	Cake	Yes	Yes	25	25	B	0.5
Ehsan	Yellow	Pasta	Yes	No	20	25	C	1
Faris	Yellow	Burger	No	No	20	25	A	0

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 $\quad z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - Treat z_{if} as interval-scaled
  - $\delta_{ij}^{(f)} = 0$
    - if x_{if} or x_{jf} is misssing
    - or x_{if} = $x_{jf} = 0$ for binary asymmetric attributes

Exercise 11

- Find the distance based on all the attributes (mixed type)
  - Ali and Bilal
  - Ali and Faris

	Favorite Color	**Favorite Food**	**Plays Chess**	**Plays Football**	**Age**	**Salary (1000s)**	**Grade**
Ali	Blue	Cake	Yes	Yes	20	34	C
Bilal	Yellow	Cake	Yes		25	25	B
Ehsan	Yellow	Pasta	Yes	No	20	25	C
Faris	Yellow	Burger	No	No	20	25	A

- 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
-
  - if x_{if} or x_{jf} is misssing
  - or $x_{if} = x_{jf} = 0$ for binary asymmetric attributes

Exercise 11

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

- Find the distance based on all the attributes (mixed type). For numeric attributes, use Manhattan distance.
  - Ali and Bilal
    - dist(Ali,Bilal) = (1*1+1*0+1*0+**0***1+1*5+1*9+1*0.5)/(1+1+1+**0**+1+1+1)=15.5/6=2.58
  - Ali and Faris
    - dist(Ali,Faris) = (1*1+1*1+1*1+**1***1+1*0+1*9+1*1)/(1+1+1+**1**+1+1+1)=14/7=2

	Favorite Color	Favorite Food	Plays Chess	Plays Football	Age	Salary (1000s)	Grade
Ali	Blue	Cake	Yes	Yes	20	34	1
Bilal	Yellow	Cake	Yes		25	25	0.5
Ehsan	Yellow	Pasta	Yes	No	20	25	1
Faris	Yellow	Burger	No	No	20	25	0

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

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: Euclidean norm of vector d.

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity:

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

  where • indicates vector dot product, $||d||$: Euclidean norm of vector d.

- 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)

  - First, calculate vector dot product

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

  - Then, calculate $||d_1||$ and $||d_2||$

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

  - Calculate cosine similarity: cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94

Example

- D1: A red apple
- D2: I like red apple
- D3: Apple computers are good
- D4: Red apple red apple red apple red apple

Term by Document Matrix

	This	Red	Apple	I	Like	Comp	Good
D1	0	1	1	0	0	0	0
D2	0	1	1	1	1	0	0
D3	0	0	1	0	0	1	1
D4	0	4	4	0	0	0	0

Cosine(d1,d4) =

(0*0+1*4+1*4+0*0+0*0+0*0+0*0)/(sqrt(2)*sqrt(32))= 8/8 = 1

Exercise 12

- Create a term by document matrix for the following documents (only consider green terms)
  - D1: I like to eat red apples
  - D2: Red apples are sweet
  - D3: Apple computers are easy to use computers
- Find the distance cosine similarity between
  - D1 and D2
  - D1 and D3

Exercise 12

- Create a term by document matrix for the following documents
  - D1: I like to eat red apples
  - D2: Red apples are sweet
  - D3: Apple computers are easy to use computers

	Like								
D1									
D2									
D3									

- Find the distance cosine similarity between
  - D1 and D2
  - D1 and D3

Exercise 12

- Create a term by document matrix for the following documents
  - D1: I like to eat red apples
  - D2: Red apples are sweet
  - D3: Apple computers are easy to use computers

	like	eat	red	apples	sweet	computers	easy	use
D1	1	1	1	1	0	0	0	0
D2	0	0	1	1	1	0	0	0
D3	0	0	0	1	0	2	1	1

- Find the distance cosine similarity between

  ||d1|| = sqrt(1^2 + 1^2+ 1^2 + 1^2) = sqrt(4)

  ||d2|| = sqrt(1^2 + 1^2 + 1^2) = sqrt(3)

  ||d3|| = sqrt(1^2 + 2^2+ 1^2 + 1^2) = sqrt(7)

  - D1 and D2
    - sim(d1,d2) = (1+1)/(sqrt(4)*sqrt(3)) = 2/sqrt(12)
  - D1 and D3
    - sim(d1,d3) = 1/(sqrt(4)*sqrt(7)) = 1/sqrt(28)

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

Summary

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

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009