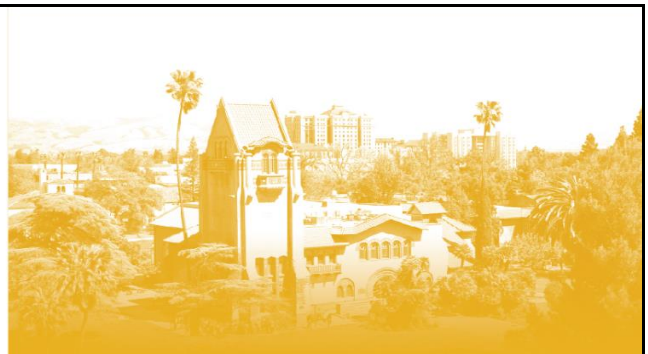
SJSU SAN JOSÉ STATE UNIVERSITY

Know Your Data
(Data Exploration)



1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

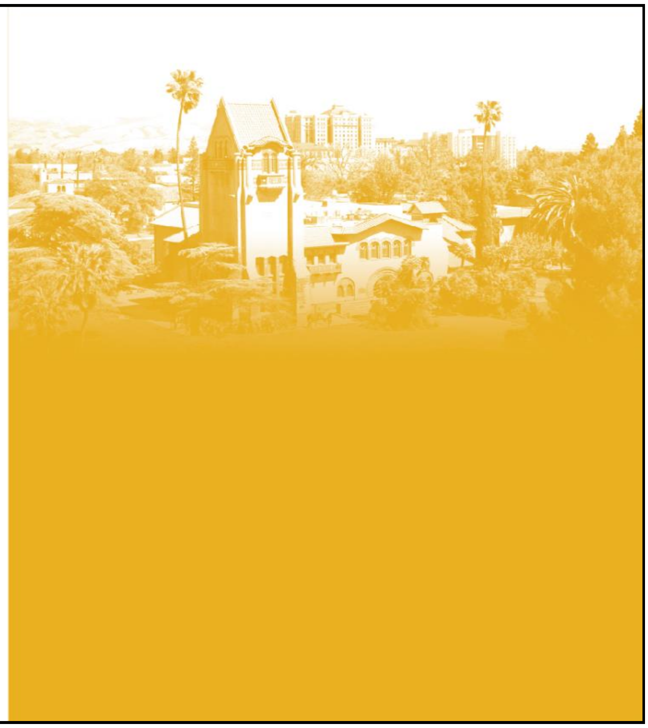
- Review of the Data Types (from DATA270)
- Characteristics of Structured Data
- Basic Statistical Descriptions of Data
- Similarity and Dissimilarity Measures

2

2

SJSU SAN JOSÉ STATE UNIVERSITY

Recall from DATA270…



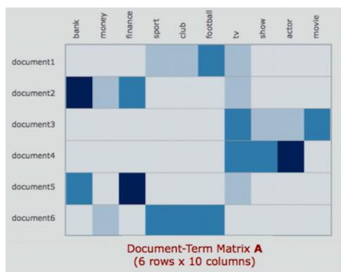
3

SJSU SAN JOSÉ STATE UNIVERSITY

Record or Tabular Dataset

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Record or Transaction Data



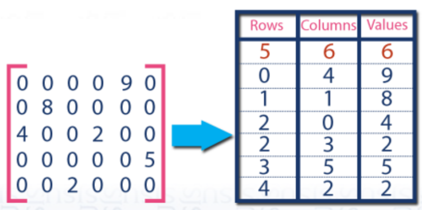
Document-Term Matrix

Variables

Cases

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

Data Matrix



Sparse Data Matrix

Images adopted from various internet pages

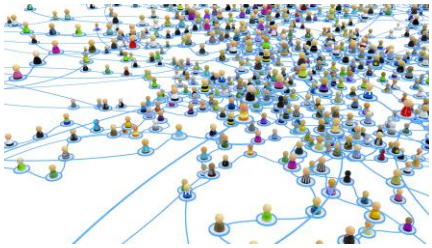
4

4

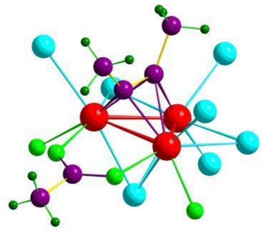
SJSU SAN JOSÉ STATE UNIVERSITY

Graph-Based Dataset

- Transportation network
- World Wide Web



Data with Relationships among Objects



Data with Objects that are Graphs

- Molecular Structures
- Social or information networks

5

SJSU SAN JOSÉ STATE UNIVERSITY

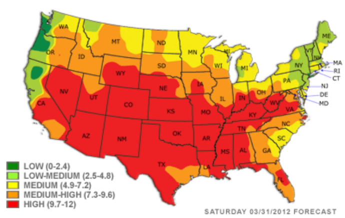
Ordered Dataset

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

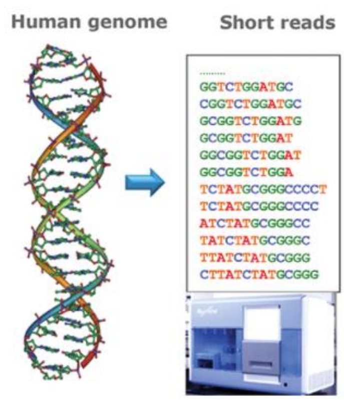
Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Sequential Data

Time Series Data



Spatial Data

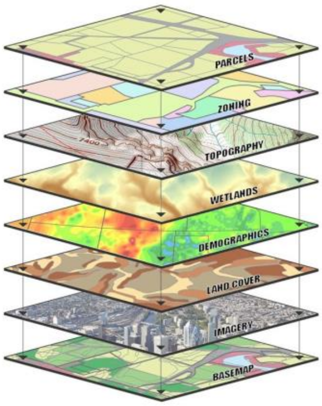


Sequence Data

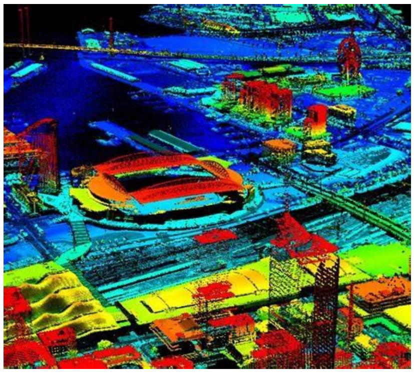
Images adopted from various sources.

6

Other Ordered Dataset



GIS Data



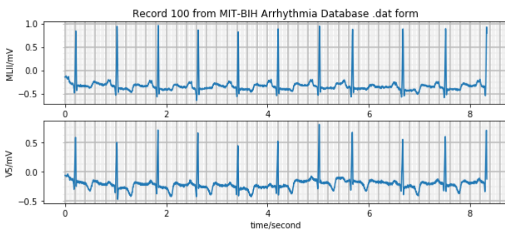
LiDAR Data

Satellite Data

Images adopted from various sources.

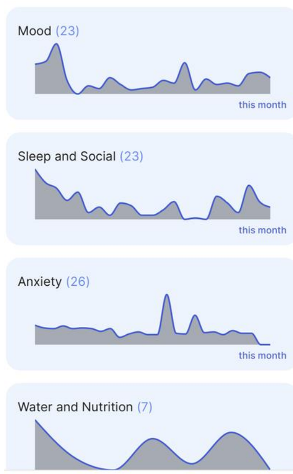
7

More Ordered Dataset



ECG Data

Video Data



Wellness Data

Images adopted from various sources.

8

SJSU SAN JOSÉ STATE UNIVERSITY

Characteristics of Structured Data

9

SJSU SAN JOSÉ STATE UNIVERSITY

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

10

SJSU SAN JOSÉ STATE UNIVERSITY

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**
- In database lingo: database rows ➔ data objects; database columns ➔ attributes

11

11

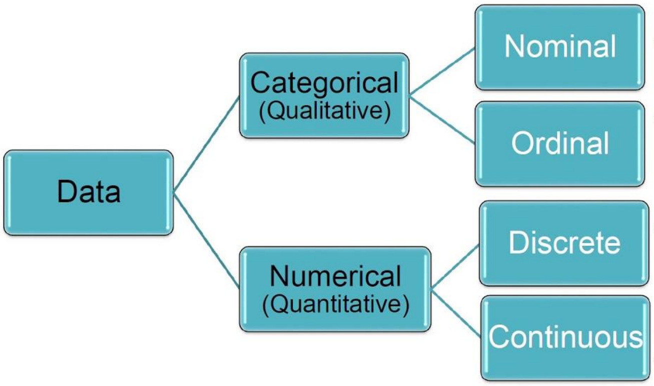
SJSU SAN JOSÉ STATE UNIVERSITY

Data Attributes

- **Attributes** (or **dimensions, features, variables**)
  - A data field, representing a characteristic or feature of a data object.
  - *e.g., customer _ID, name, address*
- **Attribute Types:**
  - Categorical
  - Numerical
  - Continuous vs Discrete

12

Type of Data Attributes



- Named Categories
- Categories with an implied order
- Only takes certain values along an Interval
- Takes on a value at any

Categorical Attribute Types

- **Nominal**: categories, states, or "names of things"
  - Values without any meaningful order or ranking
  - hair_color = {auburn, black, blond, brown, grey, red, white}, marital status, occupation, zip codes
- **Binary (special case of nominal)**
  - 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 or ranking <u>but magnitude between successive values is not known</u>
  - Size = {small, medium, large}, grades, army rankings

SJSU SAN JOSÉ STATE UNIVERSITY

Numerical Attribute Types

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

15

15

SJSU SAN JOSÉ STATE UNIVERSITY

Discrete vs. Continuous Attributes

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

16

16

SJSU SAN JOSÉ STATE UNIVERSITY

Questions

What are the data attribute types (nominal or ordinal) of the following data types?

- Course letter grades
- Gender
- Customer satisfaction level
- Marital status

O-N-O-N

SJSU SAN JOSÉ STATE UNIVERSITY

Question

Classify the following as Categorical (nominal or ordinal) or Numerical (interval or ratio).

I/O - I - O - I - O - I/R - I/R - O

- Time in terms of AM or PM.
- Brightness as measured by a light meter.
- Brightness as measured by people's judgments.
- Angles as measured in degrees between 0 and 360.
- Bronze, Silver, and Gold medals as awarded at the Olympics.
- Height above sea level.
- Number of patients in a hospital.
- Military rank.

SJSU SAN JOSÉ STATE UNIVERSITY

Descriptive Statistics

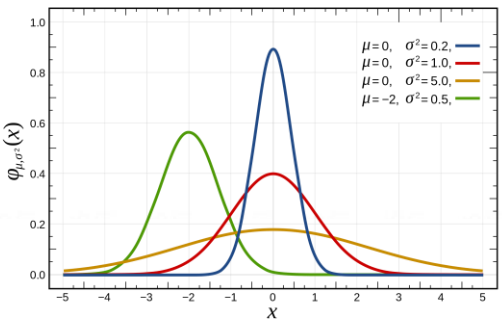
19

SJSU SAN JOSÉ STATE UNIVERSITY

Important Measurements of Data

To better understand the data, here are some important measures:

- Central Tendency
- Dispersion
- Graphic Displays of Basic Statics of Data
- Covariance and Correlation Analysis



20

SJSU SAN JOSÉ STATE UNIVERSITY

Population vs Sample

A set of data points is a sample from a population:

- A population is the entire set of objects or events under study.
  - e.g., population can be hypothetical "all students" or all students in this class.
  - e.g., population can be all the houses in a region

- A sample is a "representative" subset of the objects or events under study. This is needed because it's impossible or intractable to obtain or compute with population data.

SJSU SAN JOSÉ STATE UNIVERSITY

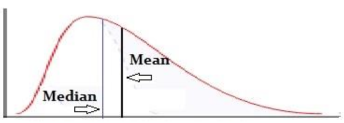
Measuring the Central Tendency

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

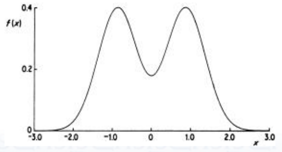
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

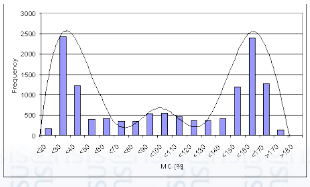
sample vs population   weighted mean

- **Median**: middle value (odd # of values) or average of the middle 2 values (otherwise)



- **Mode**: Value that occurs most frequently in the data
  - unimodal
  - bimodal
  - trimodal

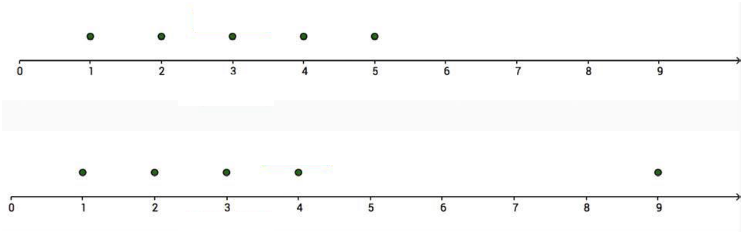




SJSU SAN JOSÉ STATE UNIVERSITY

Mean vs Median

- Which is more sensitive to extreme values or outliers? Mean or Median?



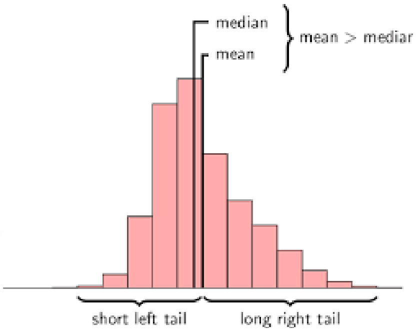
23

23

SJSU SAN JOSÉ STATE UNIVERSITY

Mean, Median, and Skewness

The following distribution is called right-skewed since the mean is greater than the median.



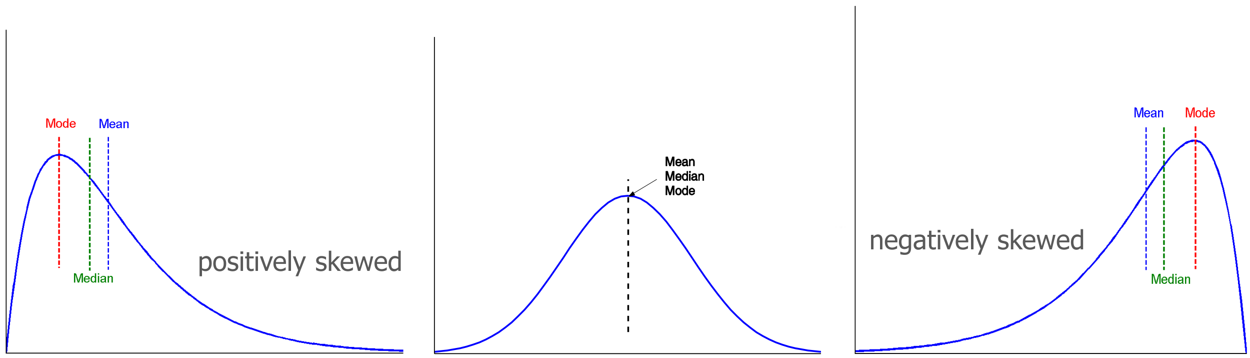
Note: skewness often "follows the longer tail".

24

24

SJSU SAN JOSÉ STATE UNIVERSITY

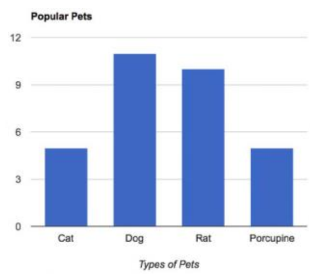
Symmetric vs. Skewed Data



SJSU SAN JOSÉ STATE UNIVERSITY

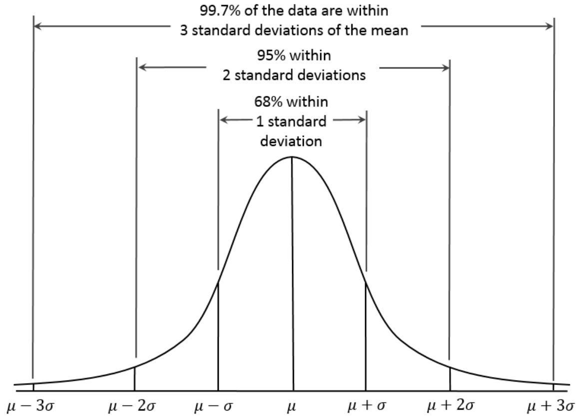
Questions

- Is income positively or negatively skewed?
- For categorical variables, does mean, median or mode make sense. Why?



Properties of Normal Distribution Curve

←———Represent data dispersion, spread ———→



Represent central tendency

27

Measuring Dispersion of Data

- **Variance**

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

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

Note: The subtle difference of formulae for sample vs. population
- n : the size of the sample
- N : the size of the population

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

28

SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Displays of Basic Statistical Descriptions

- Boxplot: graphic display of five-number summary
- Histogram: x-axis are values, y-axis repres. frequencies
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are less than or equal to 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

29

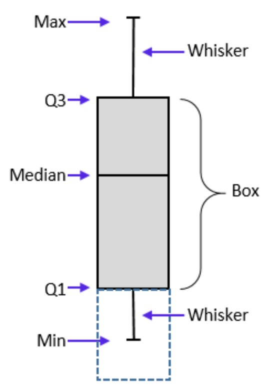
29

SJSU SAN JOSÉ STATE UNIVERSITY

Measuring the Dispersion of Data

Quartiles, **Outliners** 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
- Outliner: a value higher/lower than 1.5x IQR of Q_1 or Q_3

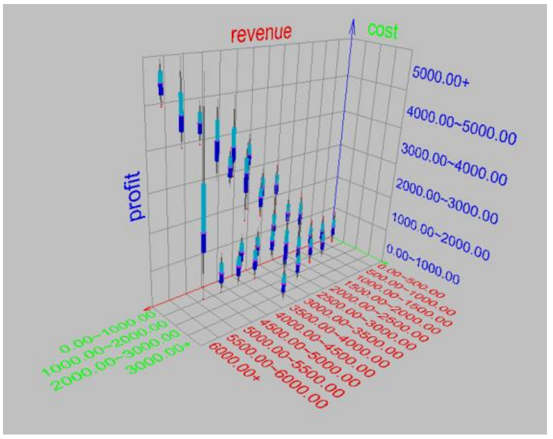


30

30

SJSU SAN JOSÉ STATE UNIVERSITY

Visualization of Data Dispersion



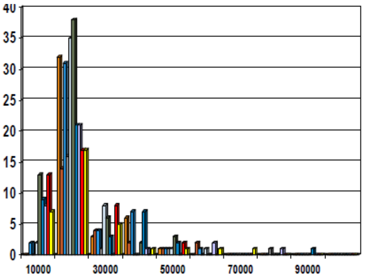
31

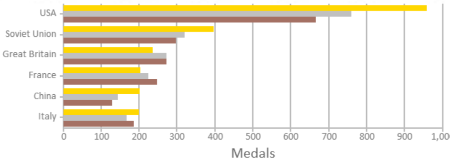
31

SJSU SAN JOSÉ STATE UNIVERSITY

Histogram Analysis

- Display of tabulated frequencies, shown as bars.
- Differences between histograms and bar charts:
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot binned quantitative data while bar charts plot categorical data
  - Bars can be reordered in bar charts but not in histograms





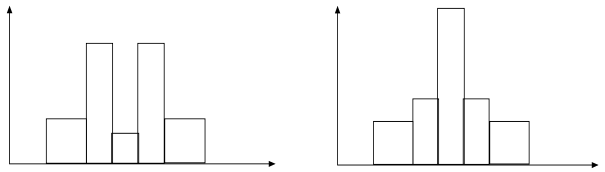
32

32

SJSU SAN JOSÉ STATE UNIVERSITY

Histograms Often Tell More than Boxplots

- Consider the following histograms:



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

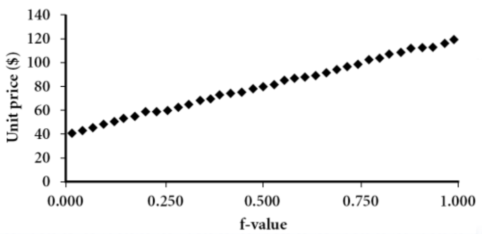
33

33

SJSU SAN JOSÉ STATE UNIVERSITY

Quantile Plot

- Displays all of the data (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



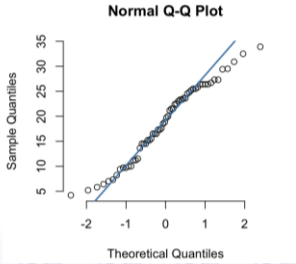
34

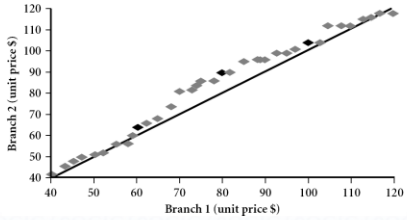
34

SJSU SAN JOSÉ STATE UNIVERSITY

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

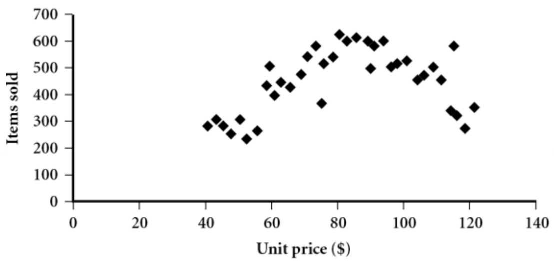


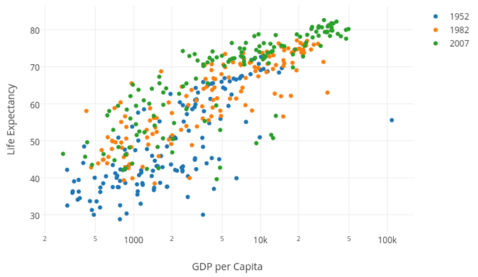


SJSU SAN JOSÉ STATE UNIVERSITY

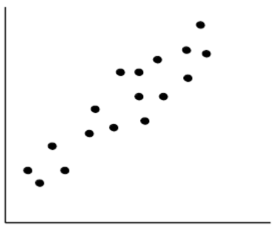
Scatter Plot

- 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

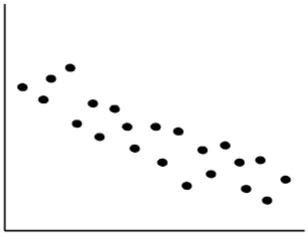




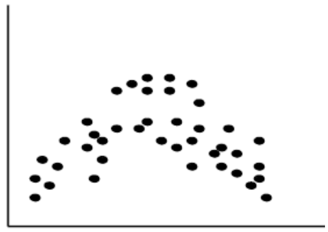
Positively and Negatively Correlated Data



Positively correlated



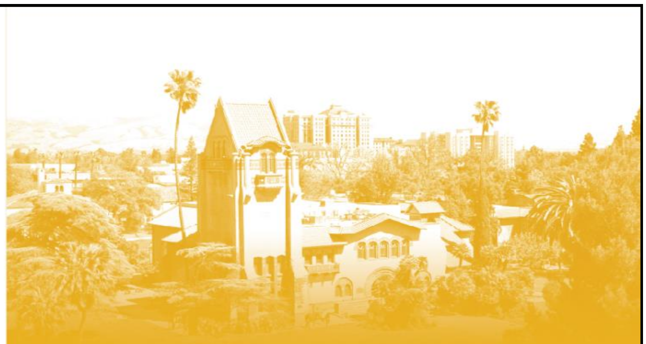
Negatively correlated



The left half is positively correlated
The right half is negatively correlated

More Scattered Plots

- What about these scattered plots?



SJSU SAN JOSÉ STATE UNIVERSITY

Data Similarity and Dissimilarity

39

SJSU SAN JOSÉ STATE UNIVERSITY

Similarity and Dissmilarity Measures

- Data Matrix versus Dissimilarity Matrix
- Similarity Measures for:
  - Binary Attributes
  - Nominal Attributes
  - Ordinal Attributes
- Dissimilarity Measures for:
  - Numeric Data: Minkowski Distance
- Cosine Similarity of 2 Vectors
- Capturing Hidden Semantics in Similarity Measures

40

40

SJSU SAN JOSÉ STATE UNIVERSITY

Similarity, Dissimilarity, and Proximity

- Similarity measure or function
  - a real-valued function that quantifies the similarity between two objects
  - measure how two data objects are alike: higher value ⇒ more alike
  - usually falls in the range [0, 1]: 0: no similarity; 1: completely similar
- Dissimilarity (or Distance) measure
  - numerical measure of how different two data objects are
  - similar to the inverse of similarity: The lower, the more alike
  - minimum dissimilarity is often 0 (i.e., completely similar)
  - range [0, 1] or [0, ∞), depending on the definition
- Proximity usually refers to either similarity or dissimilarity

41

41

SJSU SAN JOSÉ STATE UNIVERSITY

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with l dimensions

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

- Dissimilarity (distance) matrix
  - n data points, but registers only the distance d(i, j)
  - Usually symmetric → only need a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

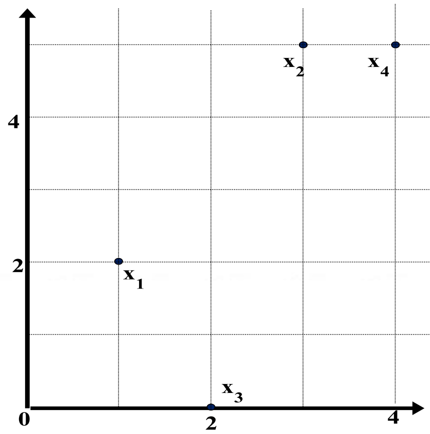
$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

42

42

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1		
x2		
x3		
x4		

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1				
x2				
x3				
x4				

SJSU SAN JOSÉ STATE UNIVERSITY

Distance on Numeric Data: Minkowski Distance

- Minkowski Distance:

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

  where $i = (x_{i1}, x_{i2}, ..., x_{il})$ and $j = (x_{j1}, x_{j2}, ..., x_{jl})$ are two l-dimensional data objects, and p is the order (the distance so defined is also called L-p norm)

- Distance Properties:
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) <= d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

- p = 1: (L_1 norm) Manhattan (or city block) distance
  - e.g. the Hamming distance: # of bits that are different between two binary vectors

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

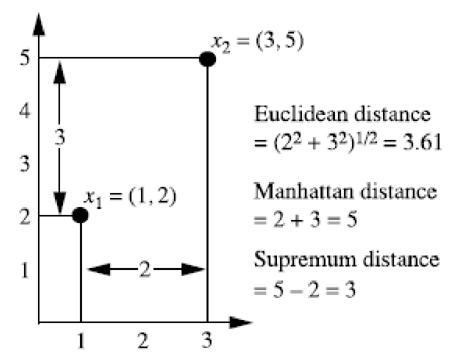
- p = 2: (L_2 norm) Euclidean distance

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

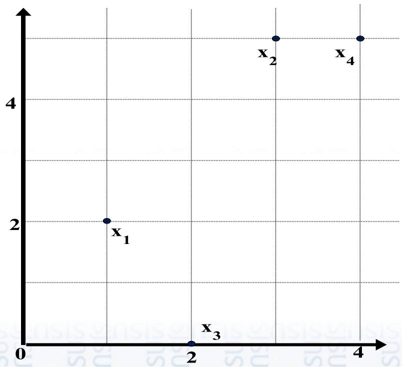
Illustration of Manhantan, Euclidean and Chebyshev Distances



SJSU SAN JOSÉ STATE UNIVERSITY

Example: Different Minkowski Distance

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



Manhattan (L_1)

L	x1	x2	x3	x4
x1				
x2				
x3				
x4				

Euclidean (L_2)

L2	x1	x2	x3	x4
x1				
x2				
x3				
x4				

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1				
x2				
x3				
x4				

47

SJSU SAN JOSÉ STATE UNIVERSITY

Proximity Measures for Binary Attributes

- To compute proximity for binary attributes, we utilize contingency tables
- Here's a contingency table for 2 binary data objects:

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

q : # attributes =1 for both *i* and *j*

r: # attributes = 1 for *i* but = 0 for j

s: # attributes = 0 for i but = 1 for j

t: # attributes = 0 for both i and j

48

48

SJSU SAN JOSÉ STATE UNIVERSITY

Proximity Measures for Binary Attributes

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

- Distance measure for symmetric binary variables (2 states are equally important):

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

- Distance measure for asymmetric binary variables (2 states aren't equally important):

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

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

$$sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j)$$ aka symmetric binary similarity

49

49

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Dissimilarity between Asymmetric Binary Variables

- Consider the following patient record table:

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, all others are asymmetric binary
- Let Y & P be 1 and N be 0 ➔ create contingency tables!!!
- Distance:

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

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$$

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

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

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

50

50

SJSU SAN JOSÉ STATE UNIVERSITY

Proximity Measures for Nominal Attributes

- For nominal attributes such as:
  e.g.: color (red, yellow, blue, green), profession, etc.

- Simple matching:

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

m: # of matches, p: total # of attributes

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}$$

- Use a large number of binary attributes
  - create a new binary attribute for each of the M nominal states

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Dissmilarity between Nominal Attributes

- Consider the sample data with nominal attributes as shown:
- Compute the dissimilarity matrix using the measure:

Object Identifier	Test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

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

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

SJSU SAN JOSÉ STATE UNIVERSITY

Proximity Measures for Ordinal Attributes

- Ordinal attributes can be discrete or continuous where order is important
  e.g. size (small, medium, large), class (freshman, sophomore, junior, senior)
- Can be treated as interval-scaled
  - Replace an ordinal variable value by its 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

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

e.g. freshman: 0; sophomore: 1/3; junior: 2/3; senior 1

➔ distance: d(freshman, senior) = 1, d(junior, senior) = 1/3

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Dissmilarity between Ordinal Attributes

- Consider the sample data as shown (only consider ordinal):

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)
1	code A	excellent
2	code B	fair
3	code C	good
4	code A	excellent

- Compute the dissimilarity matri:
  - 3 states (fair, good, excellent) ➔ $M_f = 3$
  - replace ordinal attribute values with rank ➔ $\{3, 1, 2, 3\}^T$
  - Normalize the ranking to [0, 1] using $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ ➔ $\{1, 0, 0.5, 1\}^T$
  - compute Euclidean distance ➔ dissimilarity matrix

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Document as a Data Matrix

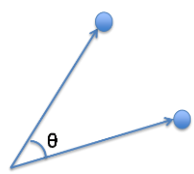
- A document can be represented by document vector, with each attribute recording the frequency of a particular term (such as keyword, or phrase) in the document:

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, etc.

Cosine Similarity of Two Vectors

- Cosine Similarity : x and y are two vectors (e.g., term-frequency vectors)



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

$||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Cosine Similarity of Documents

Find the (cosine) similarity between documents 1 and 2 from following document table.

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

- The document vectors are:

$$\boldsymbol{x} = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \qquad \boldsymbol{y} = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

- First, calculate vector dot product

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

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Cosine Similarity of Documents

- Then, calculate the Euclidean norms of $\boldsymbol{x}$ and $\boldsymbol{y}$:

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

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

- Cosine similarity:

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||} = 0.94$$

SJSU SAN JOSÉ STATE UNIVERSITY

Capturing Hidden Semantics in Similarity Measures

- The similarity measures discussed so far cannot capture hidden semantics
  - Which pairs are more similar: Geometry, algebra, music, politics?
- The same bags of words may express rather different meanings
  - "The cat bites a mouse" vs. "The mouse bites a cat"
  - This is beyond what a vector space model can handle
- Moreover, objects can be composed of rather complex structures and connections (e.g., graphs and networks)
- New similarity measures needed to handle complex semantics
  - Distributive representation and representation learning

59

SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- Data attribute types: nominal, binary, ordinal, numerical (interval-scaled, ratio-scaled)
- Many types of data sets

  e.g., numerical, text, graph, web, image.
- Gain insight into the data by:
  - Descriptive Statistics: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Similarity Measurements: distance between data objects
- All these are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

60

60