Data Mining Fundamentals

datasciencedojo
unleash the data scientist in you

Topics

- Data and Data Types
- Data Quality
- Data Preprocessing
- Similarity and Proximity
- Data Exploration and Visualization

Data and Data Types

What is Data?

Collection of data objects and their attributes

An attribute is a property or characteristic of an object

Examples: eye color of a person, temperature, etc.

Attribute is also known as variable, field, characteristic, or feature

A collection of attributes describe an object

Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

Attribute values are numbers or symbols assigned to an attribute

Distinction between attributes and attribute values

Same attribute can be mapped to different attribute values

- Example: height can be measured in feet or meters

Different attributes can be mapped to the same set of values

- Example: Attribute values for ID and age are integers
- But properties of attribute values can be different
- ID has no limit, but age has a maximum and minimum value

Discrete and Continuous Attributes

Discrete Attribute

Has only a finite or countably infinite set of values

Examples: zip codes, counts, or the set of words in a collection of documents

Often represented as integer variables

Note: binary attributes are a special case of discrete attributes

Continuous Attribute

Has real numbers as attribute values

Examples: 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.

Types of Attributes

There are different types of attributes

Nominal

Examples: ID numbers, eye color, zip codes

Ordinal

Examples: Rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

Interval

Examples: Temperatures in Celsius or Fahrenheit

Ratio

Examples: Temperature in Kelvin, length, time, counts

Types of Data Sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

datasciencedojo
unleash the data scientist in you

Document Data

Each document becomes a "term" vector

Each term is a component (attribute) of the vector

The value of each component is the number of times the corresponding term occurs in the document

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

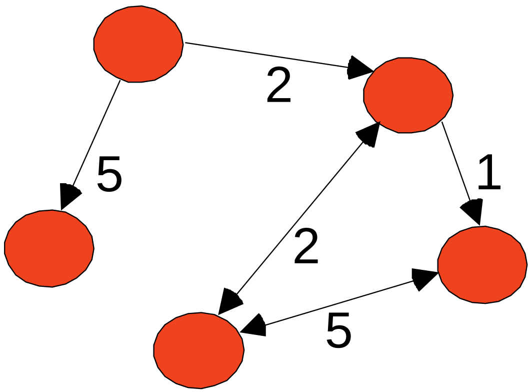
A special type of record data where each record (transaction) involves a set of items

For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitutes a transaction while the individual products that were purchased are the items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data

Examples: Generic graph and HTML Links



Data Mining

Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

datasciencedojo
unleash the data scientist in you

Ordered Data

Genomic sequence data

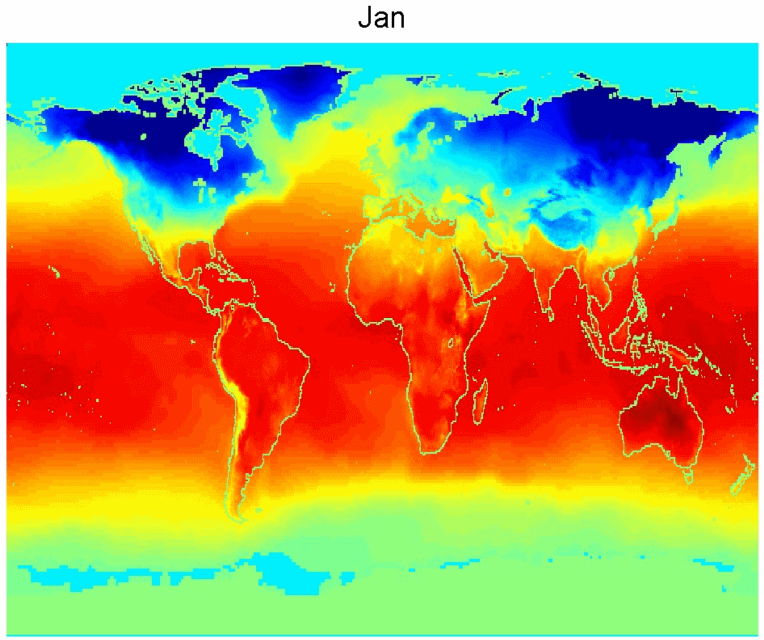
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

datasciencedojo
unleash the data scientist in you

Ordered Data

Spatial-Temporal Data

- Average Monthly Temperature of land and ocean



datasciencedojo
unleash the data scientist in you

Data Quality

Data Quality

What kinds of data quality problems are there?
How can we detect problems with the data?
What can we do about these problems?

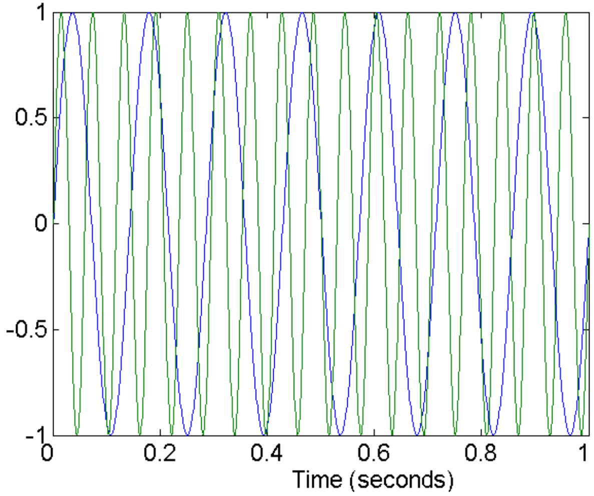
Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

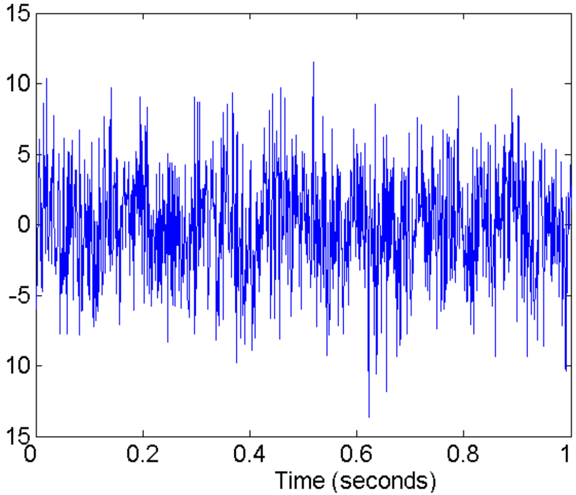
Noise

Noise refers to modification of original values

Examples: distortion of a person's voice when talking on a poor phone and "snow" on a television screen



Two Sine Waves

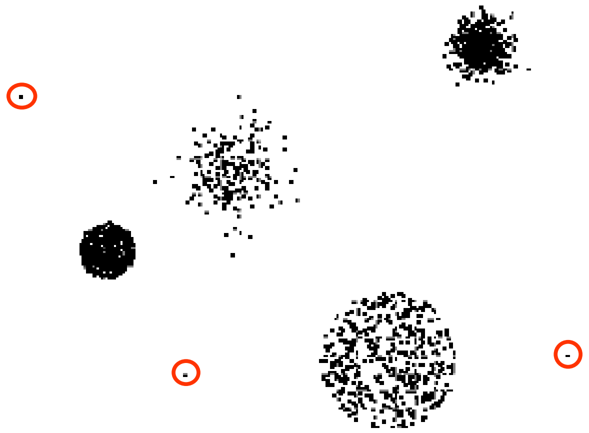


Two Sine Waves + Noise

datasciencedojo
unleash the data scientist in you

Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



datasciencedojo
unleash the data scientist in you

Missing Values

Reasons for missing values

Information is not collected
(e.g., people decline to give their age and weight)

Attributes may not be applicable to all cases
(e.g., annual income is not applicable to children)

Handling missing values

- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates, of one another

Major issue when merging data from heterogeneous sources

Example:

Same person with multiple email addresses

Data cleaning

Process of dealing with duplicate data issues

Data Preprocessing

Data Preprocessing

Aggregation

Sampling

Dimensionality Reduction

Feature Subset Selection

Feature Creation

Discretization and Binarization

Attribute Transformation

Aggregation

Combining two or more attributes (or objects) into a single attribute (or object)

Purpose

Data reduction

- Reduce the number of attributes or objects

Change of scale

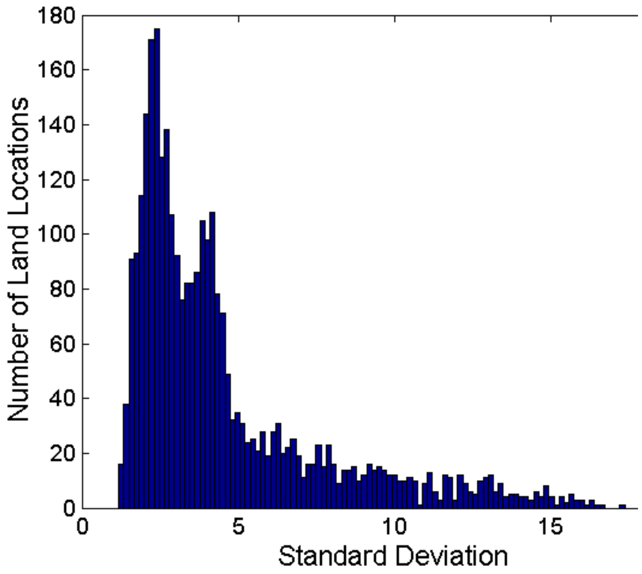
- Cities aggregated into regions, states, countries, etc.

More “stable” data

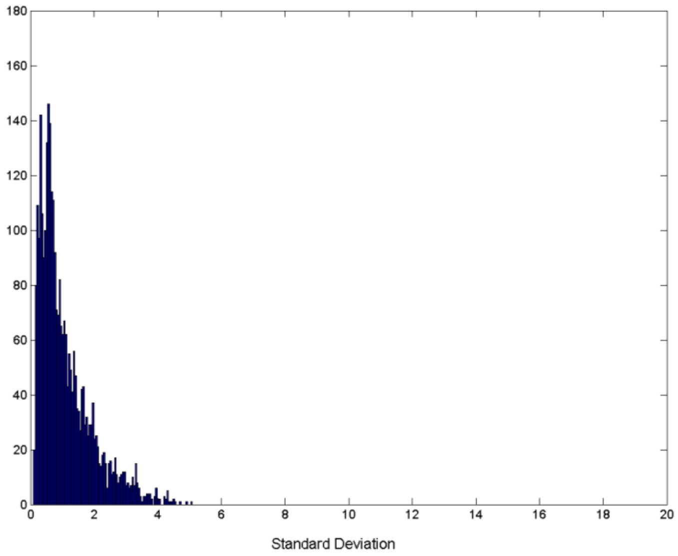
- Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

Sampling is the main technique employed for data selection

- It is often used for both the preliminary investigation of the data and the final data analysis

Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming

Sampling

The key principle for effective sampling is:

Using a sample will work almost as well as using the entire data set if the sample is representative.

Types of Sampling

Simple Random Sampling

There is an equal probability of selecting any particular item

Sampling without replacement

As each item is selected, it is removed from the population

Sampling with replacement

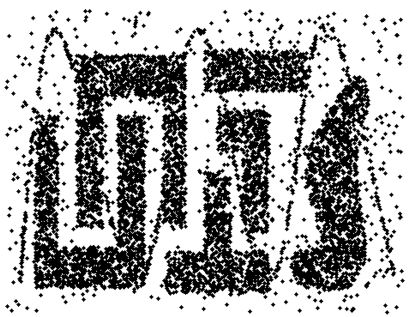
Objects are not removed from the population as they are selected for the sample

- In sampling with replacement, the same object can be picked up more than once

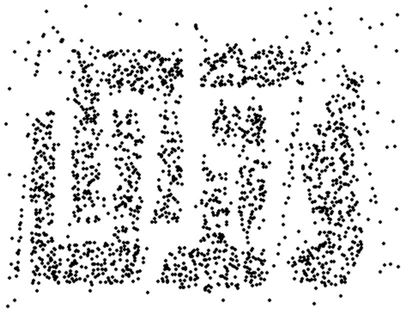
Stratified sampling

Split the data into several partitions; then draw random samples from each partition

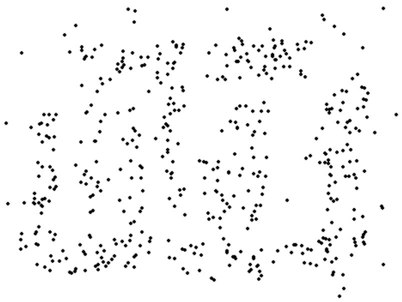
Sample Size



8000 points



2000 Points

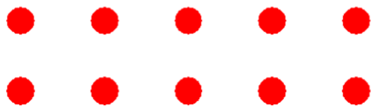


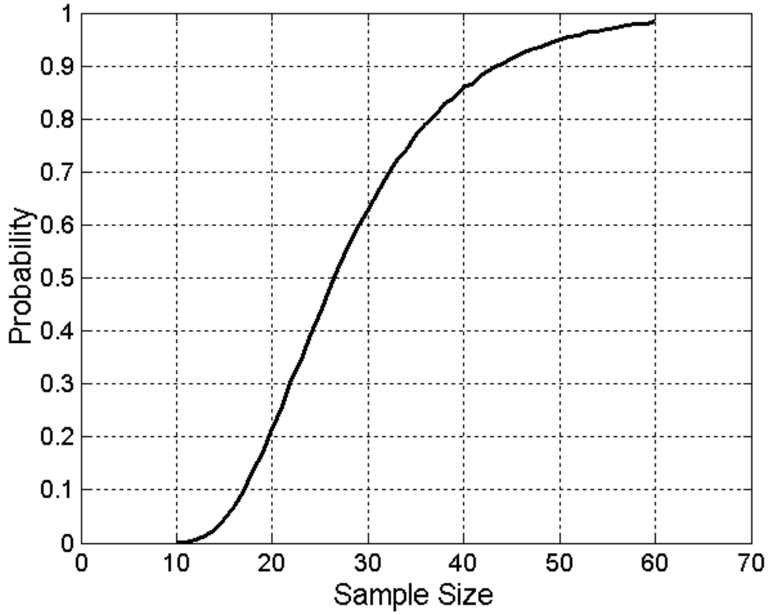
500 Points

datasciencedojo
unleash the data scientist in you

Sample Size

What sample size is necessary to get at least one object from each of 10 groups?



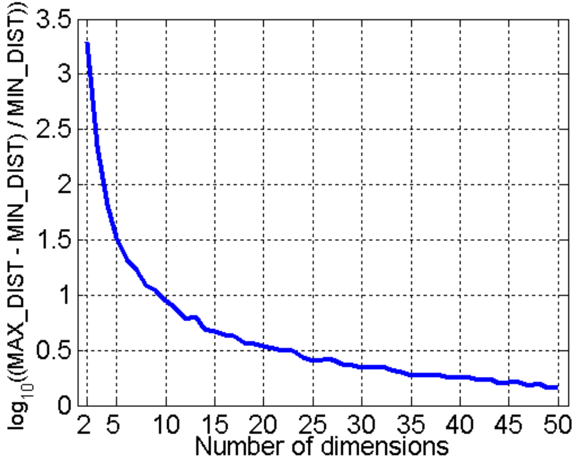


Curse of Dimensionality

When dimensionality increases, data becomes increasingly sparse in the space that it occupies

Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points



datasciencedojo
unleash the data scientist in you

Dimensionality Reduction

Purpose:

Avoid curse of dimensionality

Reduce amount of time and memory required by data mining algorithms

Allow data to be more easily visualized

May help to eliminate irrelevant features or reduce noise

Techniques:

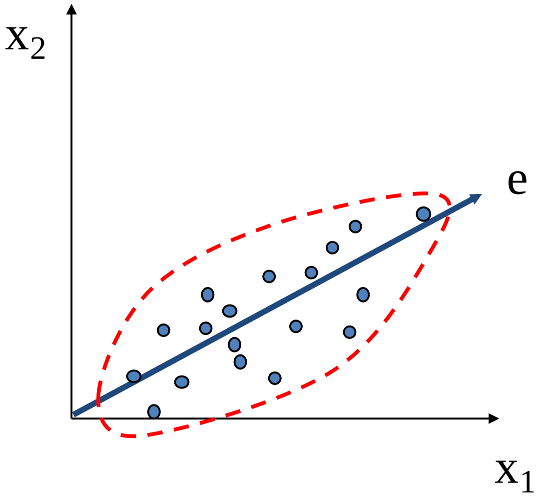
Principle Component Analysis

Singular Value Decomposition

Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

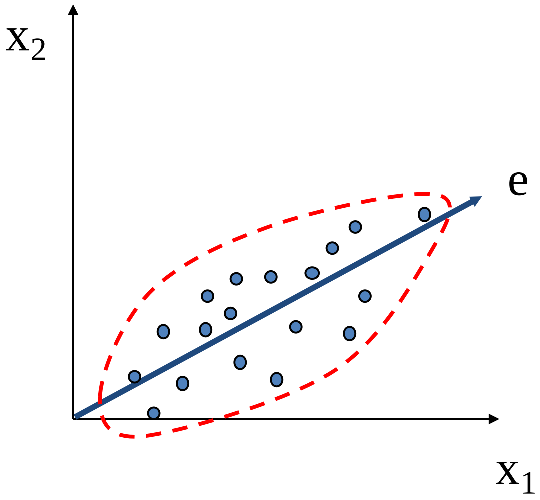
The goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

Find the eigenvectors of the covariance matrix

The eigenvectors define the new space



Feature Subset Selection

Another way to reduce dimensionality of data

Redundant features

Duplicate much or all of the information contained in one or more other attributes

Example: purchase price of a product and the amount of sales tax paid

Irrelevant features

Contain no information that is useful for the data mining task at hand

Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

Techniques:

Brute-force approach:

Try all possible feature subsets as input to data mining algorithm

Embedded approach:

Feature selection occurs naturally as part of the data mining algorithm

Filter approach:

Features are selected before data mining algorithm is run

Wrapper approach:

Use the data mining algorithm as a black box to find best subset of attributes

Feature Creation

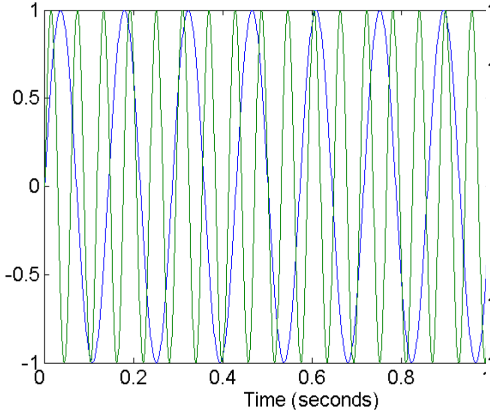
Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

Three general methodologies:

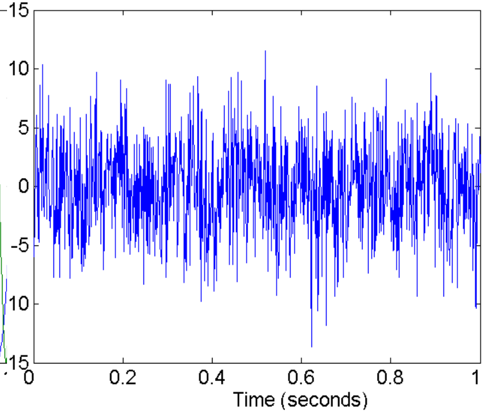
- Feature Extraction-domain specific
- Mapping Data to New Space
- Feature Construction-combining features

Mapping Data to a New Space

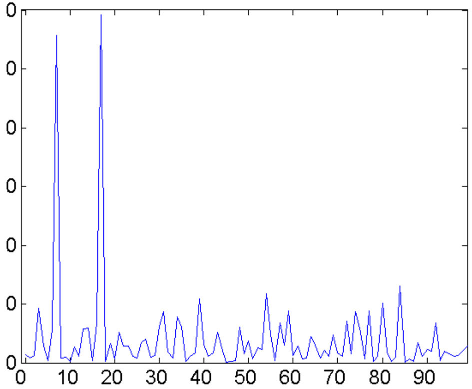
- Fourier transform
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise

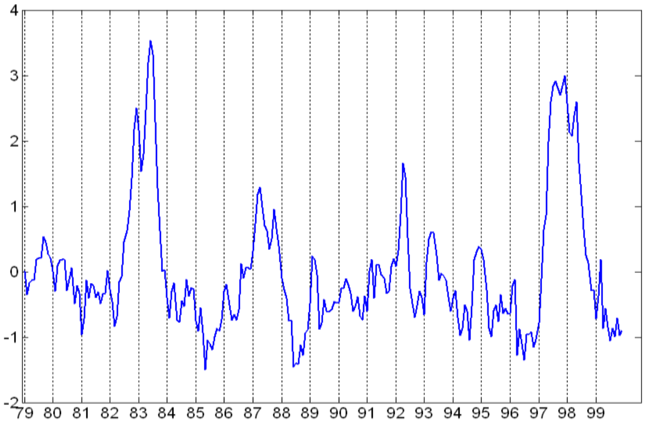


Frequency

Attribute Transformation

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

Simple functions: x^k, $\log(x)$, e^x, $|x|$



Similarity and Dissimilarity

Similarity and Dissimilarity

Similarity

Numerical measure of how alike two data objects are

Is higher when objects are more alike

Often falls in the range [0,1]

Dissimilarity

Numerical measure of how different are two data objects

Lower when objects are more alike

Minimum dissimilarity is often 0

Upper limit varies

Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

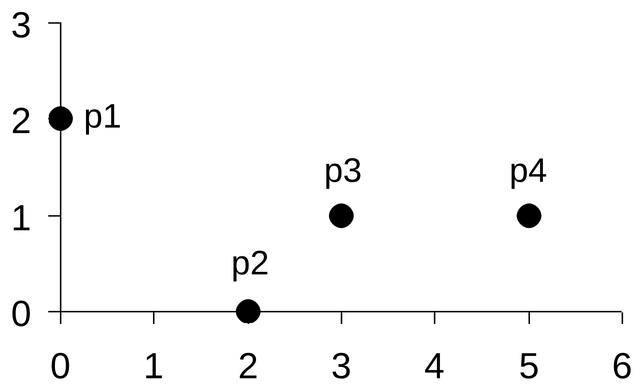
Euclidean Distance

- Euclidean Distance:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

datasciencedojo
unleash the data scientist in you

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance: Examples

- $r = 1$ City block (Manhattan, taxicab, L1 norm) distance
    - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$ Euclidean distance
- $r \to \infty$ "supremum" (L_{max} norm, L_∞ norm) distance
    - This is the maximum difference between any component of the vectors
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

Cosine Similarity

- If d_1 and d_2 are two document vectors, then:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where $\bullet$ indicates vector dot product and $|| d ||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

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

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{\mathbf{0.5}} = (6)^{\mathbf{0.5}} = 2.245$

$\cos(d_1, d_2) = .3150$

datasciencedojo
unleash the data scientist in you

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

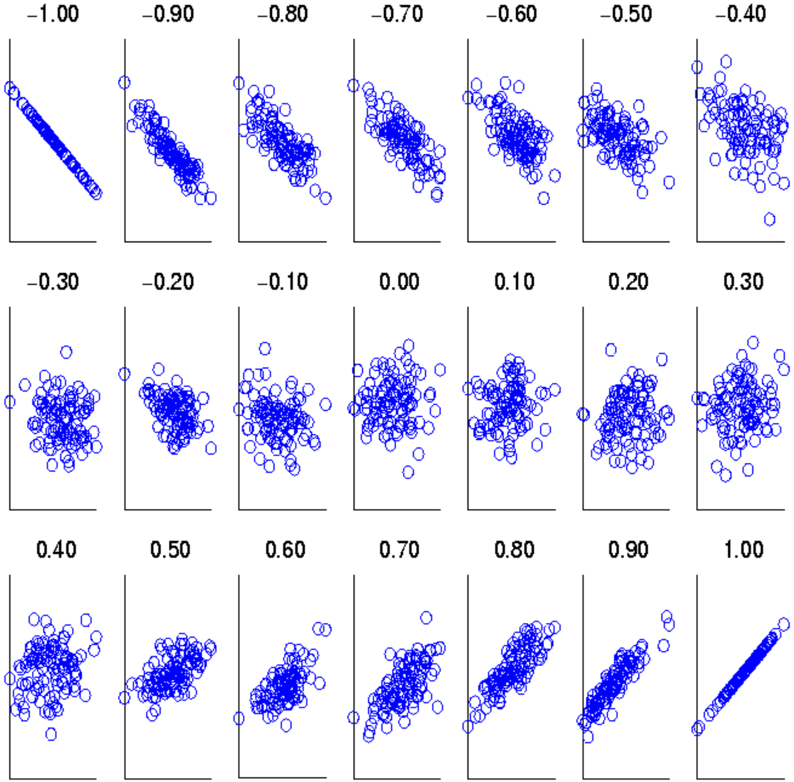
$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

datasciencedojo
unleash the data scientist in you

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1

Data Exploration

datasciencedojo
unleash the data scientist in you

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics
- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools
- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook/index.htm)

Techniques Used In Data Exploration

In EDA, as originally defined by Tukey

The focus was on visualization

Clustering and anomaly detection were viewed as exploratory techniques

In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

In our discussion of data exploration, we will focus on:

Summary statistics

Visualization

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - Sir Ronald Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location, and spread
    - Examples:
      - Location – mean
      - Spread – standard deviation
  - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

For continuous data, the notion of a percentile is more useful

Given an ordinal, or continuous, attribute *x* and a

number between 0 and 100, the *p*th percentile is a value of x such that *p*% of the observed values of x are less than X_p

For instance, the 50th percentile is the value $X_{50\%}$ such that 50% of all values of x are less than $X_{50\%}$

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points
- However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- However, this is also sensitive to outliers, so other measures are often used

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Visualization

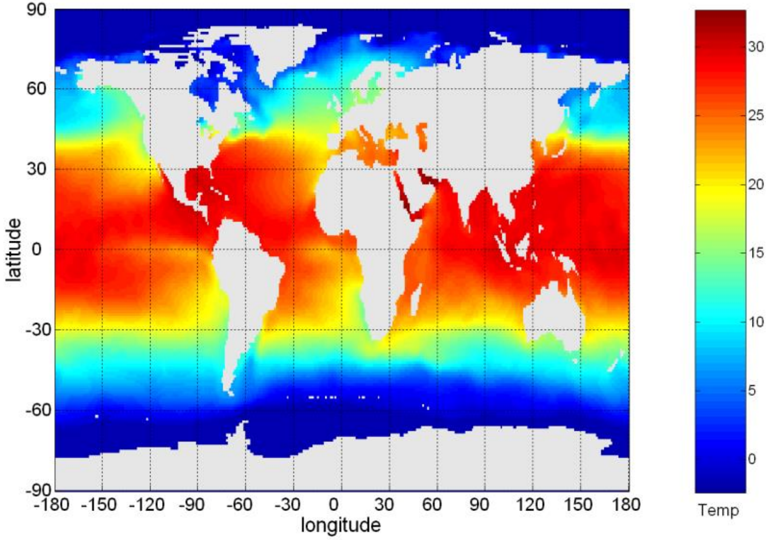
Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

Visualization of data is one of the most powerful and appealing techniques for data exploration.

- Humans have a well developed ability to analyze large amounts of information that is presented visually
- Can detect general patterns and trends
- Can detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
    - Tens of thousands of data points are summarized in a single figure



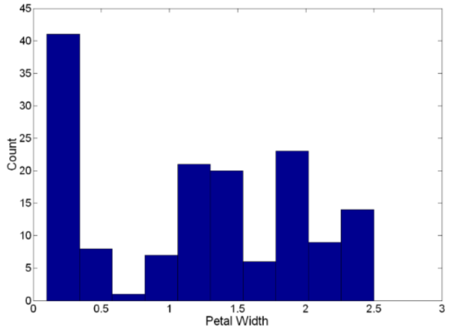
datasciencedojo
unleash the data scientist in you

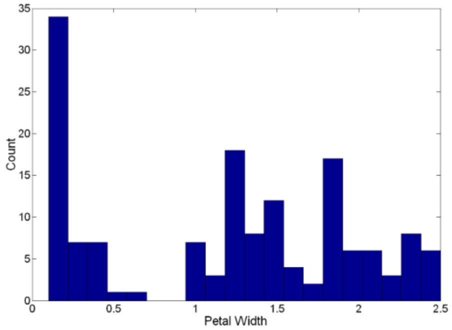
Representation

- The mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Visualization Techniques: Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)

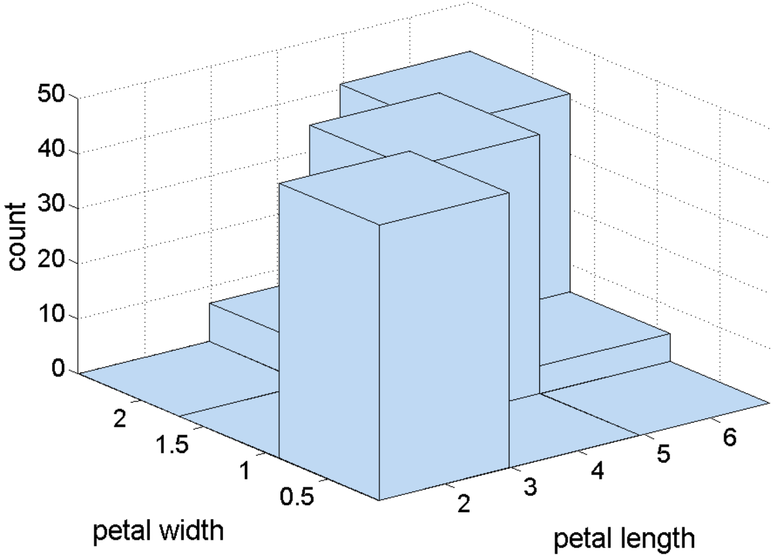




datasciencedojo
unleash the data scientist in you

Two-Dimensional Histograms

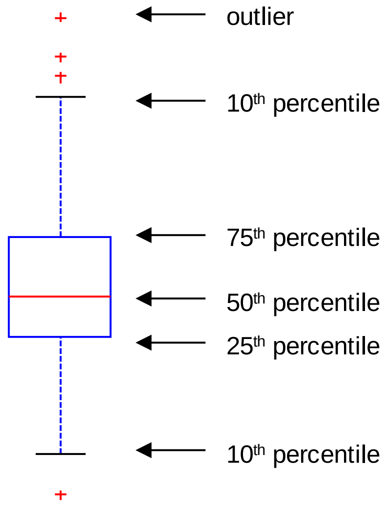
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



datasciencedojo
unleash the data scientist in you

Visualization Techniques: Box Plots

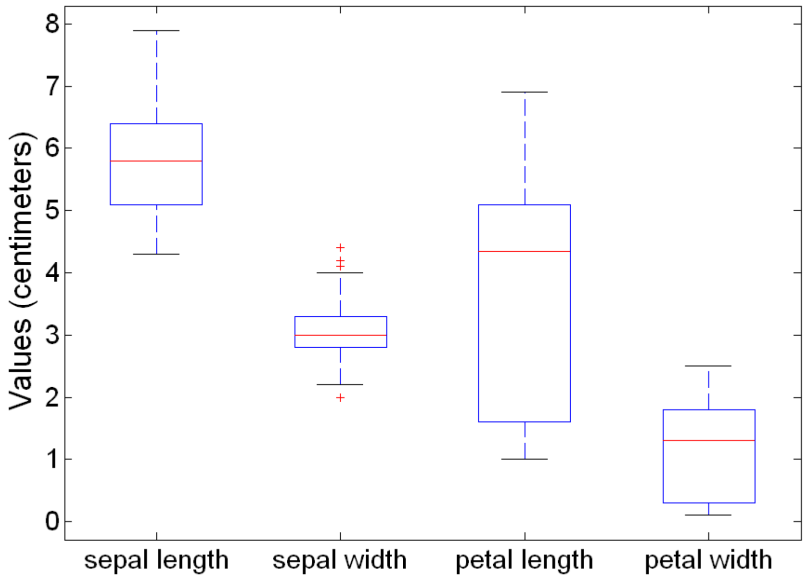
- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - The following figure shows the basic part of a box plot:



datasciencedojo
unleash the data scientist in you

Example of Box Plots

- Box plots can be used to compare attributes

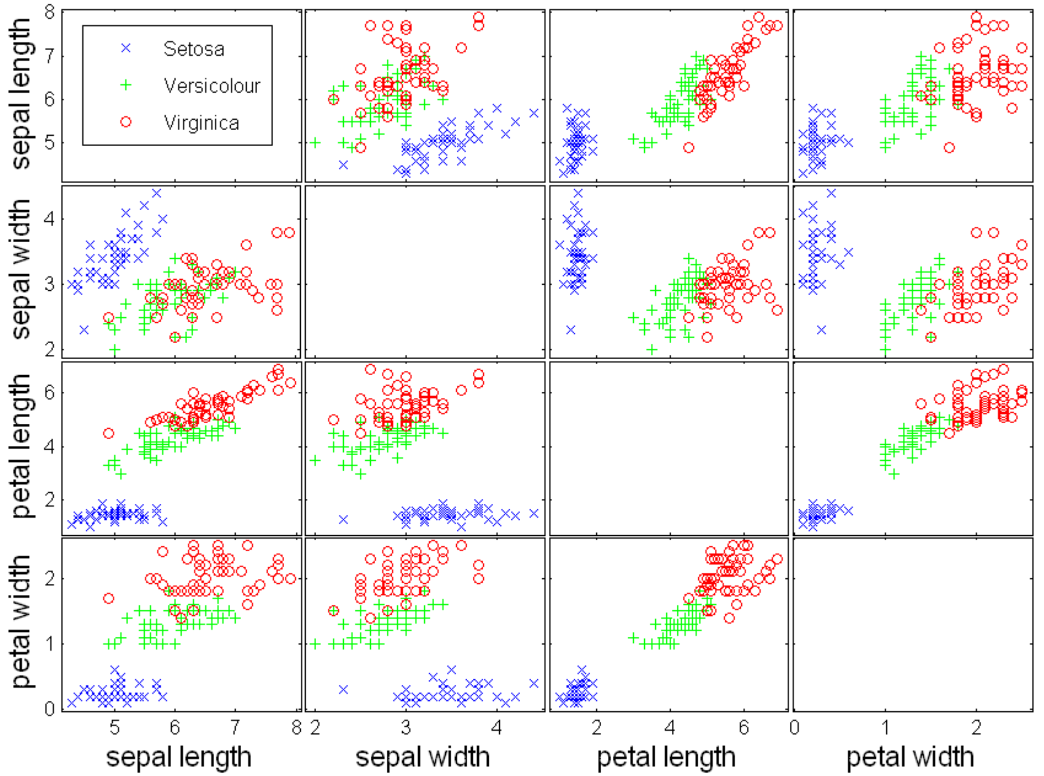


datasciencedojo
unleash the data scientist in you

Visualization Techniques: Scatter Plots

- Scatter plots
  - Attribute values determine the position
  - Two-dimensional scatter plots are the most common, but we can have three-dimensional scatter plots
  - Often additional attributes can be displayed using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots compactly summarize the relationships of several pairs of attributes

Scatter Plot Array of Iris Attributes

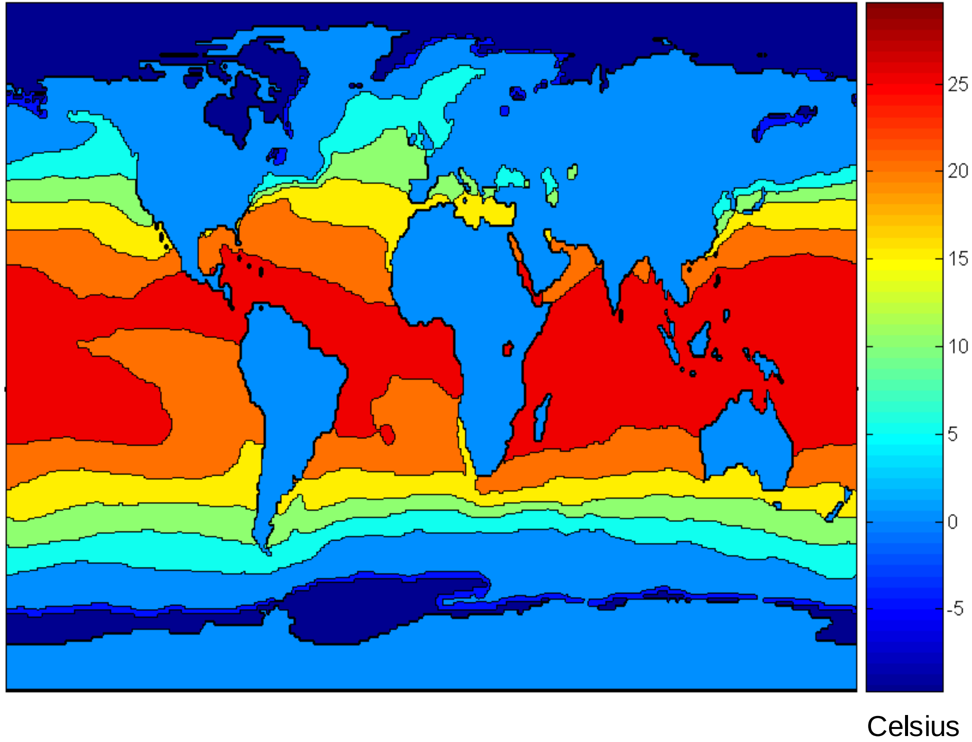


datasciencedojo
unleash the data scientist in you

Visualization Techniques: Contour Plots

- Contour plots
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values
  - The most common example is contour maps of elevation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide

Contour Plot Example: SST Dec, 1998



Questions?

datasciencedojo
unleash the data scientist in you