# Chapter 1: Introduction

## Statistical Approach to a Scientific Problem

1. Ask a question
2. Collect data
3. Initial, exploratory data analysis
4. Answer the question (Inferential statistics)

## Ask a question

- Describe something (What is happening?)
- Make predictions (What will happen?)
- Causal inference (What will happen if I ...?)

## Regression Analysis

Build a model to "explain" the relationship between a single variable $Y$ and other variables $X_1, \cdots, X_p$

- $Y$: response variable, output, dependent variable
- $X$: predictor variable, input, independent variable
- Example: Classification

## Goals of Regression Analysis

- Make predictions
- Effect of predictor variables
- Description of data structure
- Warning: regression analysis does not establish causation (i.e., you cannot tell whether changing X causes Y to change or the other way around)

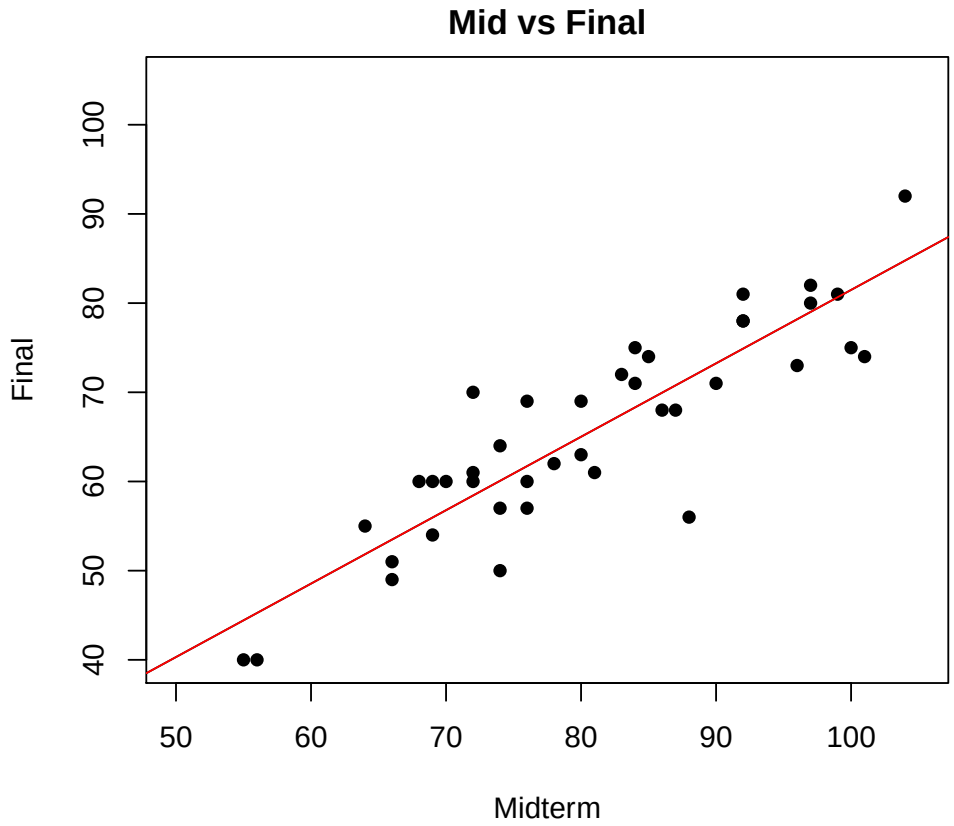
## Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Examples
  - Predicting **sales amounts** of new product based on advertising expenditure
  - Predicting **wind velocities** as a function of temperature, humidity, air pressure, etc
  - Time series prediction of **stock market indices**

# Regression Example: STATS 415 students

**Goal:** Use regression to predict final exam scores Data

- **midterm scores**



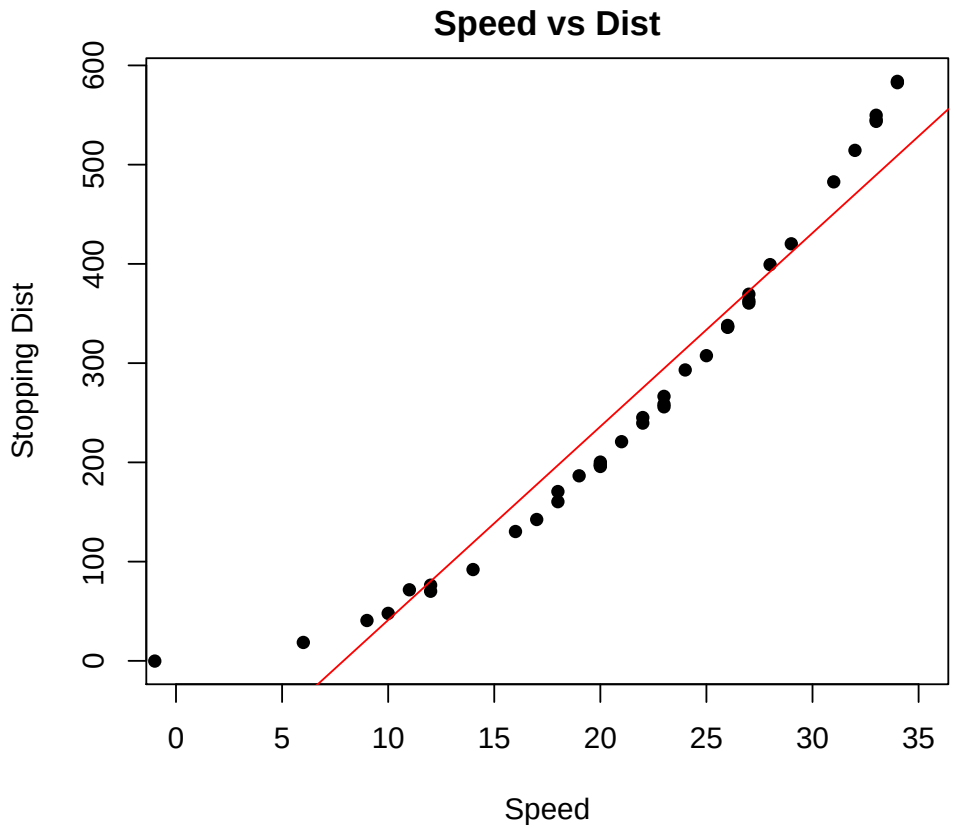
## Regression: Examples

Population: STATS 415 students Additional Data

- Categorical: ethnicity
- Binary (two values only): gender
- Discrete: # of credits, # of housemates
- Continuous: age, height

# Regression Example:

**Goal:** Use regression to find a relationship between speed and stopping distance

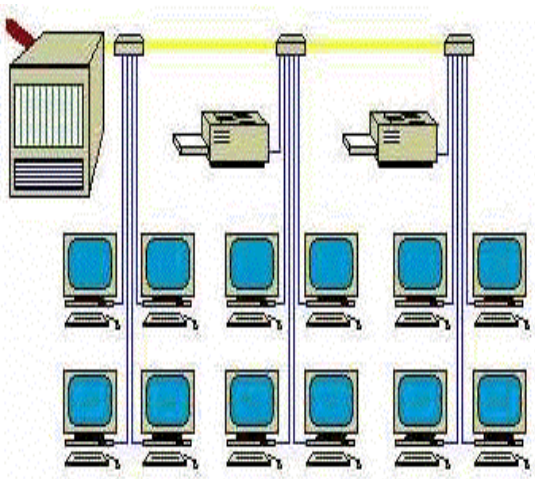


## Classification Example: Customer Scoring

- A bank has a database of 1 million past customers, 10% of whom took out mortgages
- **Goal:** Use data mining to predict whether a new customer will take out a "mortgage or not" based on the customer data
- Customer data
  - Other credit data
  - Demographic data on the customer

# Classification Example: Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit card fraud detection
  - Network intrusion detection

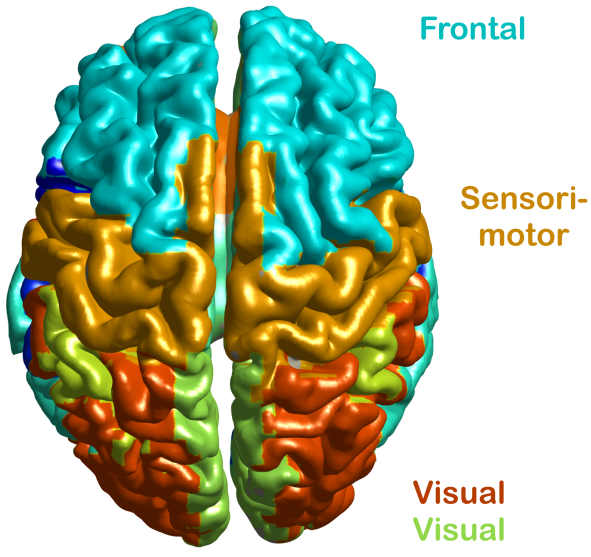
## Collect data

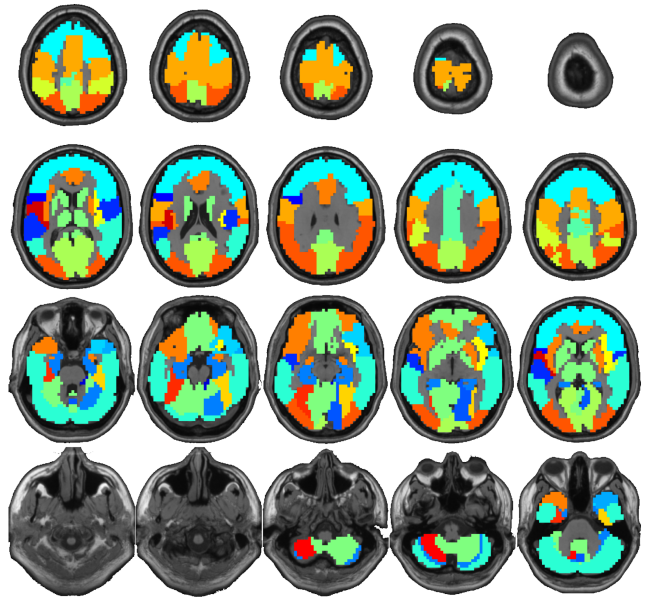
- Is the data relevant?
- Is there measurement error?
- Is there missing data?
- Is the data a sample?
  - What is the population?
  - Random sample?
- Is the data from an experiment?
  - What was the treatment?
  - How was the treatment allocated? (random?)

## Exploratory Data Analysis

- Organize data
- Display data graphically
- Summarize data
- Be alert for the unexpected

# Clustering Example: Brain



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

# Mean and Median

- The mean is the most common measure of the location of a set of points

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

- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used

$$\text{median}(x) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

# Percentiles

- For ordinal or continuous data, the notion of a percentile is also useful.
- Given an ordinal or continuous variable $x$ and a number $q$ between 0 and 100, the $q$th percentile is a value $x_q$ such that $q\%$ of the observed values of $x$ are less than $x_q$.
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of $x$ are less than $x_{50\%}$.

# Range and Variance

- Range is the difference between the max and min

$$\text{range}(x) = \max(x) - \min(x)$$

- The variance or standard deviation is the most common measure of the spread of a set of points

$$\text{Variance}(x) = s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

# Other Measures

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

$$\begin{aligned} \text{AAD}(x) &= \frac{1}{n}\sum_{i=1}^{n}|x_i - m(x)| \\ \text{MAD}(x) &= \text{median}(|x_1 - m(x)|, \ldots, |x_n - m(x)|) \\ \text{interquartile}(x) &= x_{75\%} - x_{25\%} \end{aligned}$$

# Multivariate Summary Statistics

- Location: compute the mean or median separately for each variable

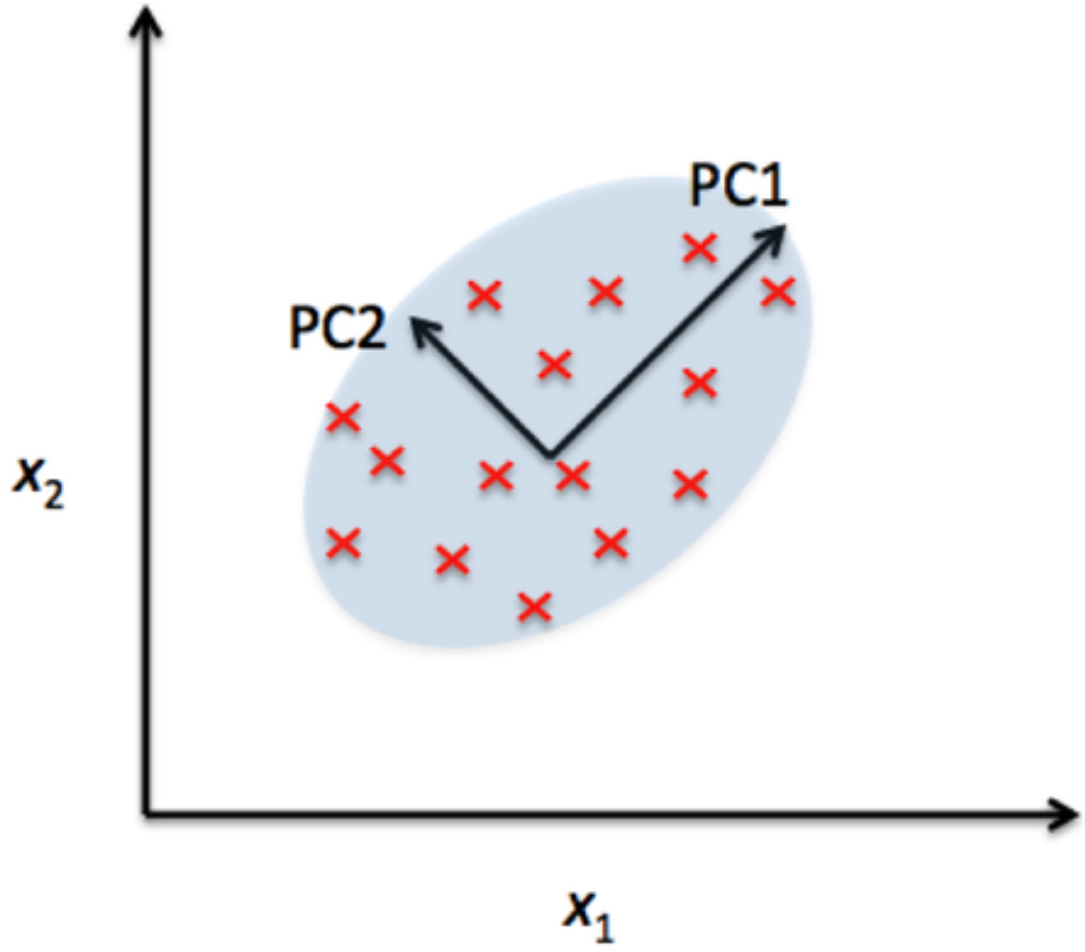
$$\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)$$

- Spread: covariance matrix

$$\text{Covariance}(x_j, x_{j'}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

- Correlation matrix

# Multivariate Summary Statistics: 2-Dims



# Other Measures

- Skewness: measures the degree to which the values are symmetrically distributed around the mean

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

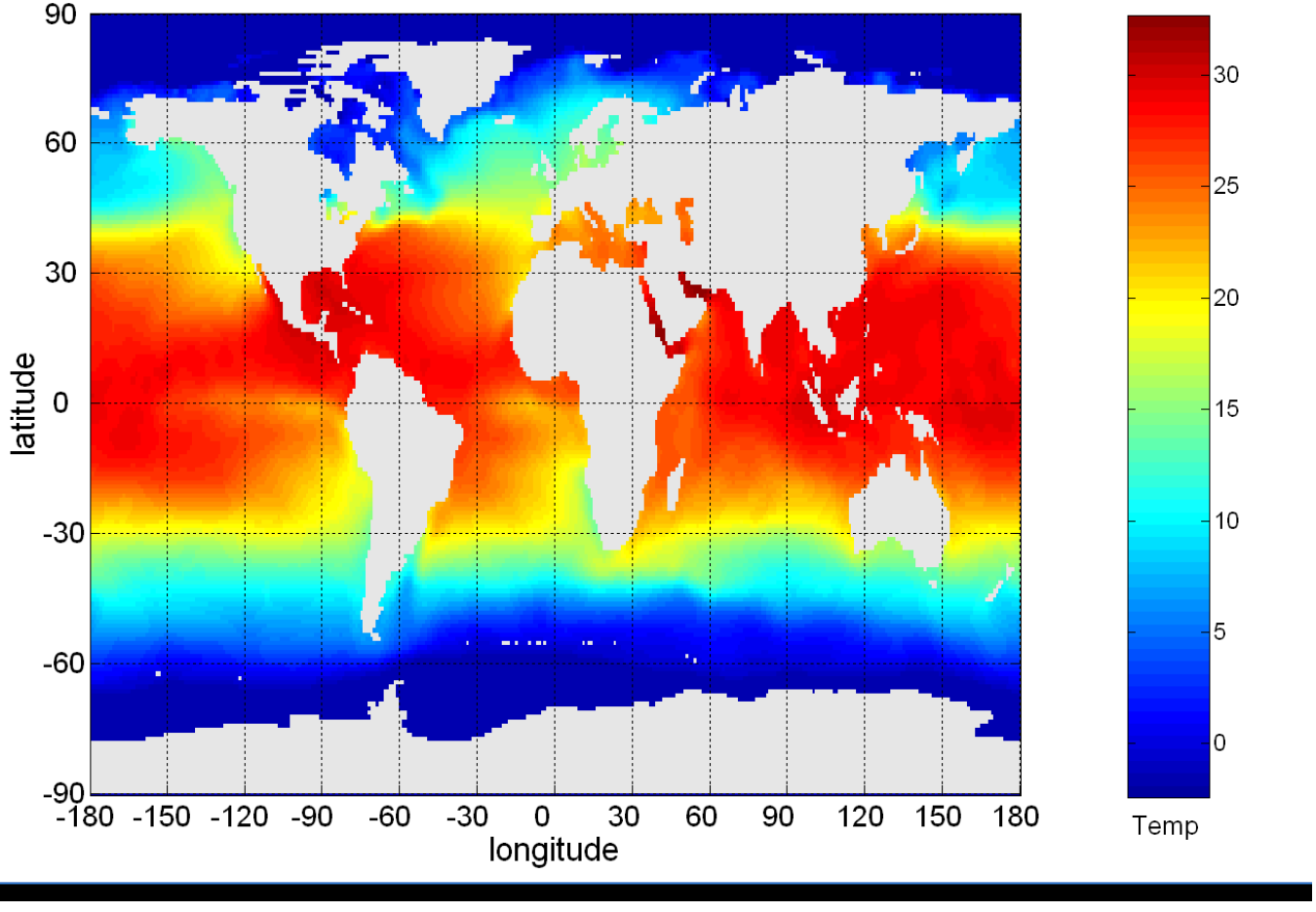
- Characteristics not easy to measure quantitatively: multi-modal

# Multivariate Summary Statistics

- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data objects or variables 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 for July 1982.
- Tens of thousands of data points are summarized in a single figure.



# Iris Data Set

Many of the exploratory data techniques are illustrated with the iris plant data set

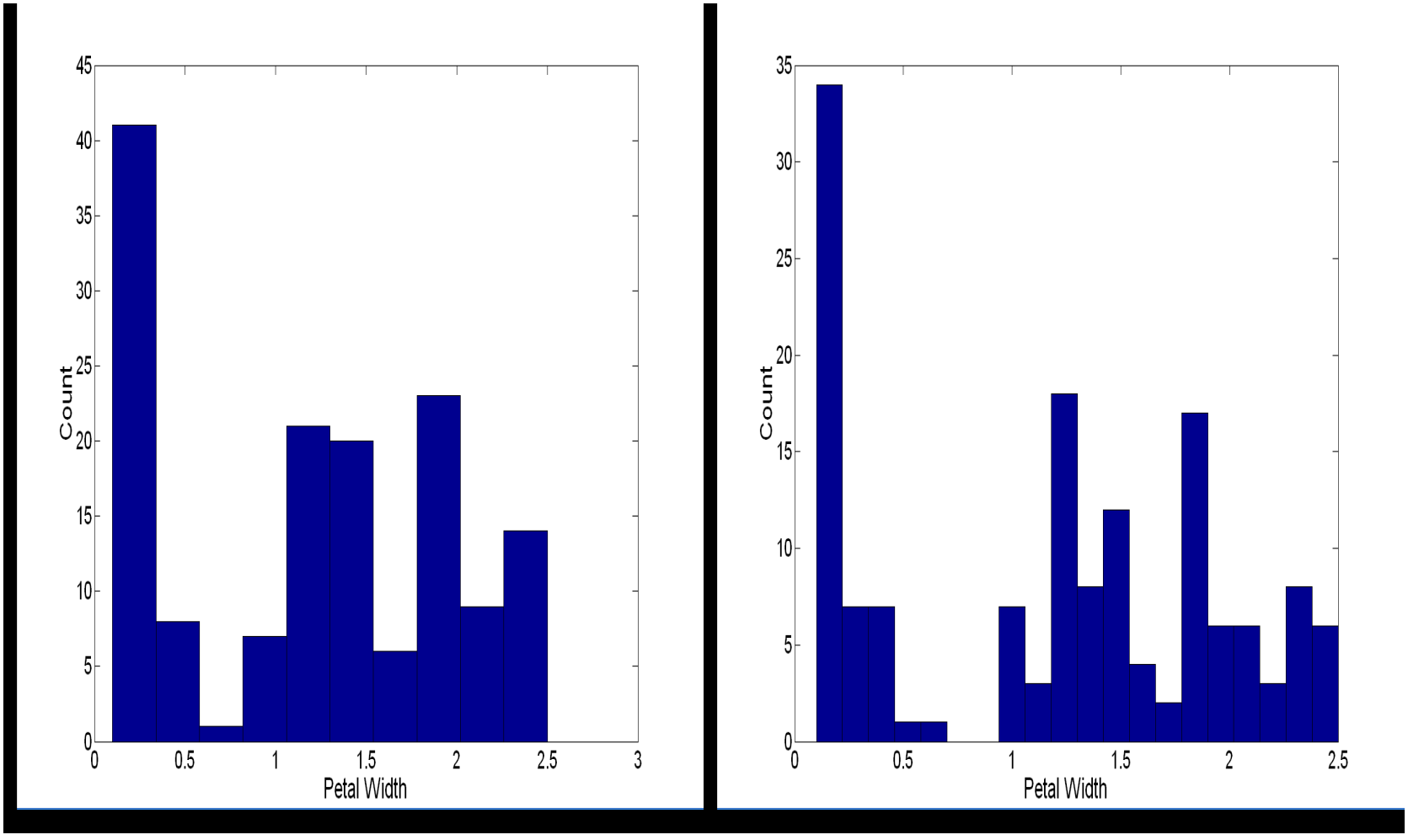
- Can be obtained from R
- Three flower types (classes)
  - Setosa
  - Virginica
  - Versicolour
- Four (non-class) variables
  - Sepal width and length
  - Petal width and length

# Iris Data Set

# Histogram

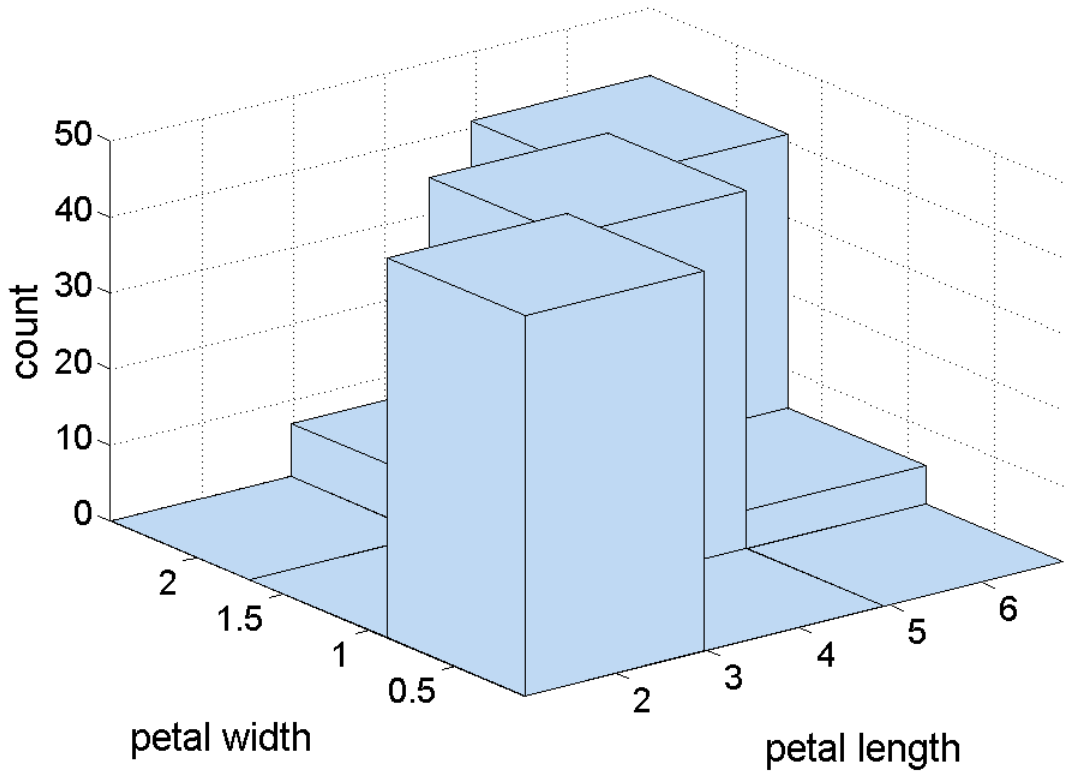
- Usually shows the distribution of values of a single variable
- The height of each bar indicates the number of objects.
- Shape of histogram depends on the number of bins.
- Example: petal width (10 and 20 bins, respectively)

# Histogram: Iris Data Set



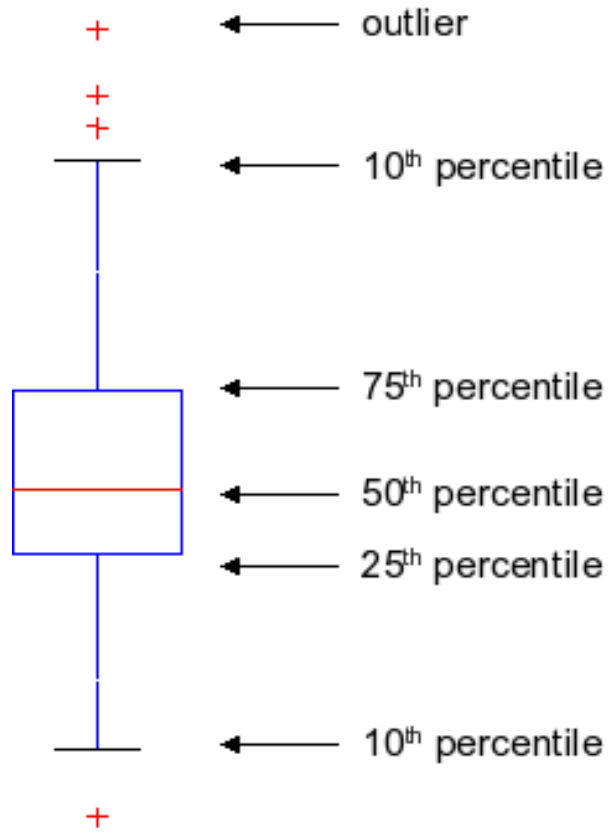
# Two-Dim Histogram: Iris Data Example

- Show the joint distribution of two attributes
- Example: petal width and petal length



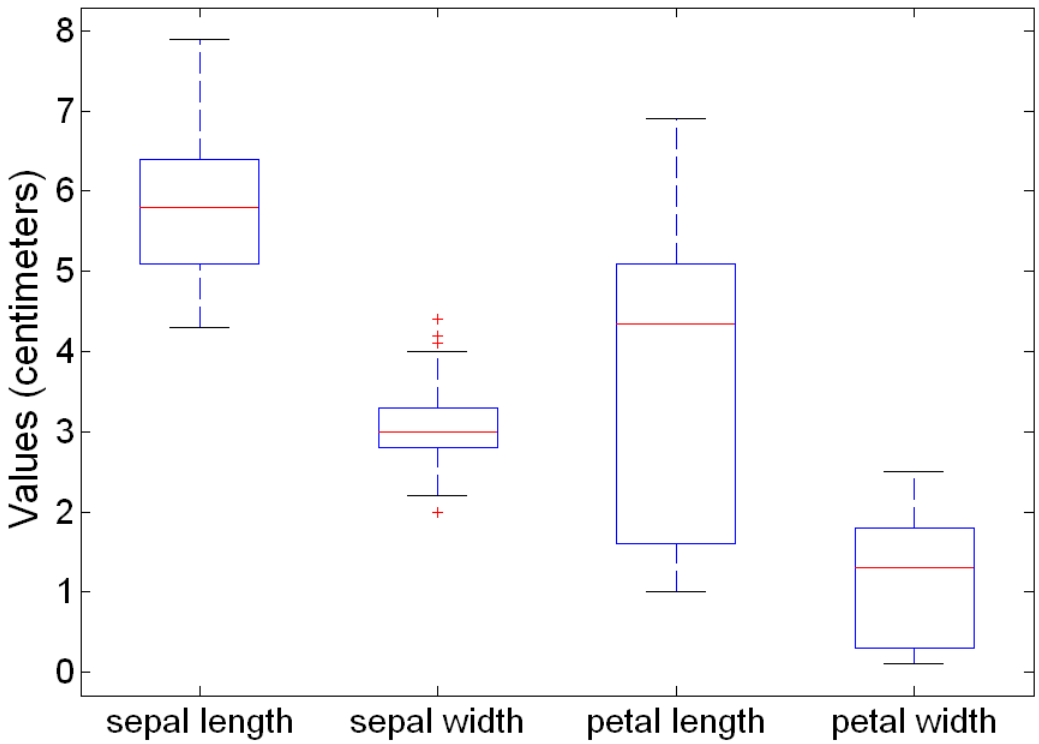
# Box Plot (1)

- Invented by Tukey
- Another way of displaying the distribution of data
- Basic part of a box plot



# Box Plot: Iris Data Example

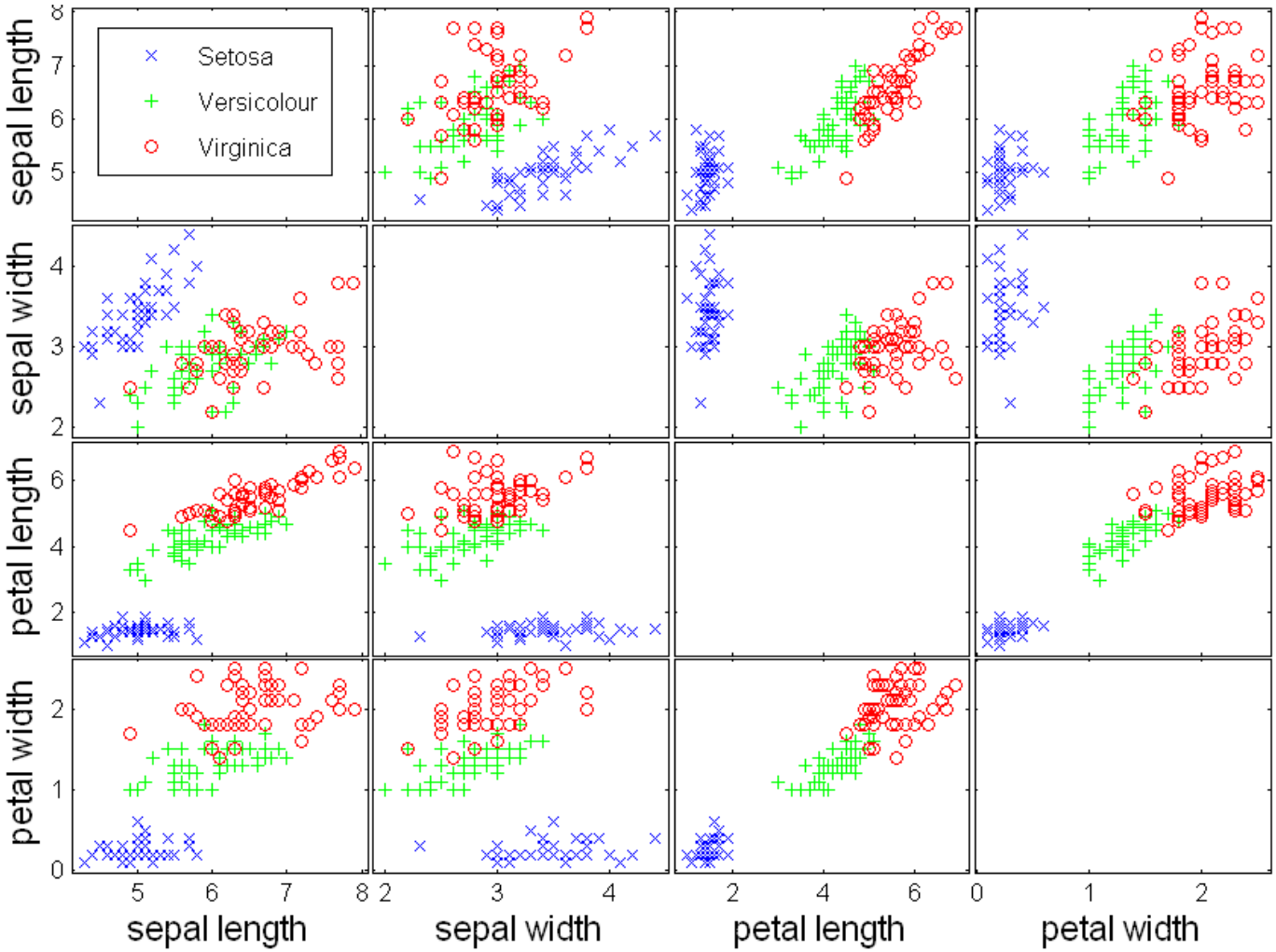
- Box plots can be used to compare variables.



# Scatter Plot

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

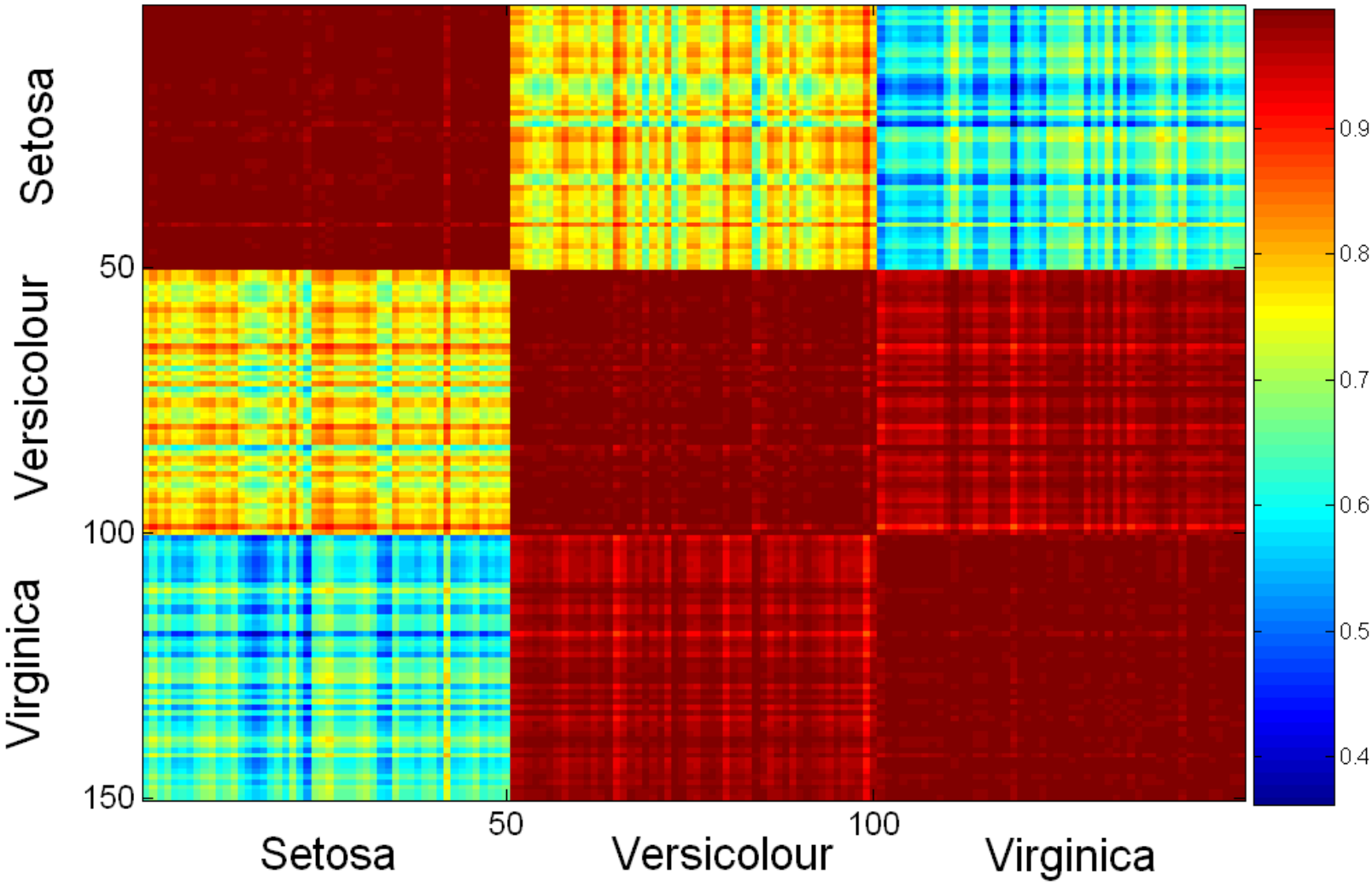
# Scatter Plot: Iris Data Example



# Matrix Plot

- Can plot the data matrix
- This can be useful when objects are sorted according to class.
- Typically, the variables are normalized to prevent one variable from dominating the plot.
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects.

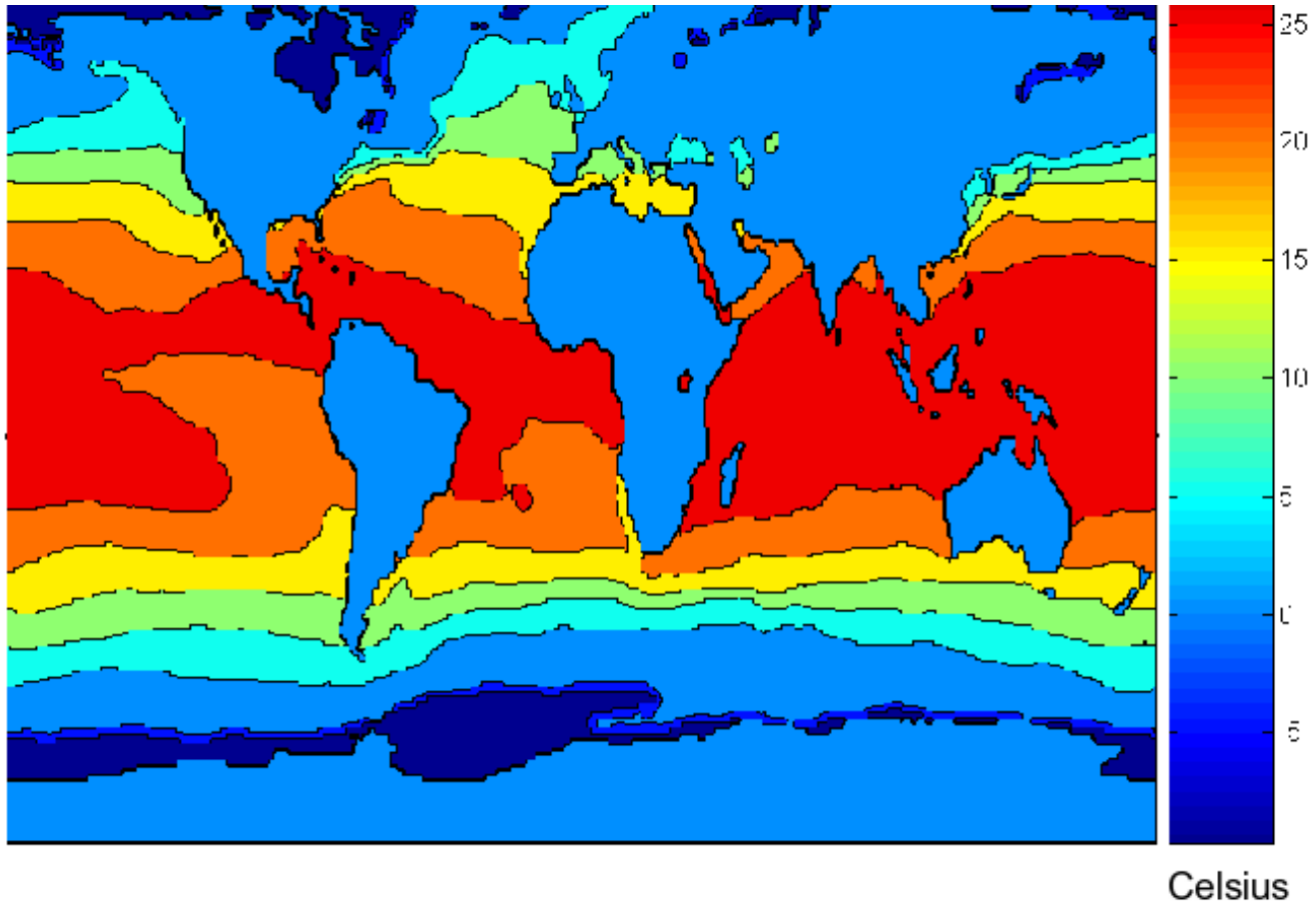
# Matrix Plot: Iris Data Example



# Contour Plot

- Useful when a continuous variable is measured on a spatial grid.
- 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: sea surface temperature.

# Contour Plot: Sea Temperature Example



## Inferential Statistics

Based on data

- Test hypotheses
- Make predictions
- Reaching decisions

via linear regression analysis or generalized linear model

## Inference

- Alternatively, we may also be interested in the **type of relationship** between $y$ and the $x$'s.
- For example
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

## Example: Housing Inference

- Wish to predict median house price based on 14 variables (river view, crime, and etc.).
- Probably want to understand which factors have the **biggest effect** on the response and how big the effect is.
- For example how much impact does a river view have on the house value.

## Types of Variables

- Qualitative, categorical: can't say one is bigger than another
- Quantitative, numerical
  - Discrete counts
  - Continuous measures

## Examples

Population: STATS 500 students

- Categorical: ethnicity
- Binary (two values only): gender
- Discrete: # of credits, # of housemates
- Continuous: age, height

## What We Will Cover

Response variable

- $Y$ is a continuous variable: linear regression

Predictor variable

- $X$: continuous, discrete or categorical

## Emphases of the Course

- Practice using linear regression models
- Learn what methods are available, and their limitations
- Many examples, less mathematical theory
- More intuition, less derivation of formulas
- Will still learn mathematical foundations behind practical tools

# Pima Data Example: Exploratory Data Analysis

# Pima Data Example

- Data collected on 768 adult female Pima Indians
- Variables: number of times pregnant, plasma glucose concentration, diastolic blood pressure, skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, age, and a test whether the patient showed signs of diabetes
- Many possibilities
  - Y : diabetes; X1,X2: diastolic, BMI
  - Y : BMI; X1,X2: diastolic, test
  - Y : test; X1,X2: diastolic, BMI
  - Y : number of times pregnant; X1,X2: age, BMI

# Pima Data Example: Exploratory Data Analysis

```
## Load the library
> library(faraway)
## Read in the data
> data(pima)
> pima
    pregnant glucose diastolic triceps insulin  bmi ...
1          6     148        72      35       0 33.6 ...
2          1      85        66      29       0 26.6 ...
3          8     183        64       0       0 23.3 ...
...
767        1     126        60       0       0 30.1 ...
768        1      93        70      31       0 30.4 ...
```

```
> help(pima)
 The dataset contains the following variables
     'pregnant' Number of times pregnant
     'glucose' Plasma glucose concentration at 2 hours
               in an oral glucose tolerance test
     'diastolic' Diastolic blood pressure (mm Hg)
     'triceps' Triceps skin fold thickness (mm)
     'insulin' 2-Hour serum insulin (mu U/ml)
     'bmi' Body mass index (weight in kg/(height in m)^2)
     'diabetes' Diabetes pedigree function
     'age' Age (years)
     'test' test whether the patient shows signs of
            diabetes (coded 0 if negative, 1 if positive)
```

```
## Dimension of the data
> dim(pima)
[1] 768   9

## Numerical Summaries
> summary(pima)
    pregnant        glucose      diastolic        triceps
 Min.   : 0.0   Min.   :  0   Min.   :  0   Min.   : 0
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 62   1st Qu.: 0
 Median : 3.0   Median :117   Median : 72   Median :23
 Mean   : 3.9   Mean   :121   Mean   : 69   Mean   :21
 3rd Qu.: 6.0   3rd Qu.:140   3rd Qu.: 80   3rd Qu.:32
 Max.   :17.0   Max.   :199   Max.   :122   Max.   :99
```

```
    insulin          bmi             diabetes          age       
 Min.   :  0   Min.   : 0.0   Min.   :0.08   Min.   :21  
 1st Qu.:  0   1st Qu.:27.3   1st Qu.:0.24   1st Qu.:24  
 Median : 31   Median :32.0   Median :0.37   Median :29  
 Mean   : 80   Mean   :32.0   Mean   :0.47   Mean   :33  
 3rd Qu.:127   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41  
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81  
      test       
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.349  
 3rd Qu.:1.000  
 Max.   :1.000
```

```
## Missing Values
> sort(pima$diastolic)
  [1]   0   0   0   0   0   0   0   0   0   0   0   0
 [13]   0   0   0   0   0   0   0   0   0   0   0   0
 [25]   0   0   0   0   0   0   0   0   0   0   0  24
 [37]  30  30  38  40  44  44  44  44 ...
> pima$diastolic[pima$diastolic == 0] = NA
> pima$glucose[pima$glucose == 0] = NA
> pima$triceps[pima$triceps == 0] = NA
> pima$insulin[pima$insulin == 0] = NA
> pima$bmi[pima$bmi == 0] =NA
```

```
## Categorical Variable
> pima$test = factor(pima$test)
> summary(pima$test)
  0   1
500 268
> levels(pima$test) = c("negative", "positive")
> summary(pima$test)
negative positive
     500      268
```

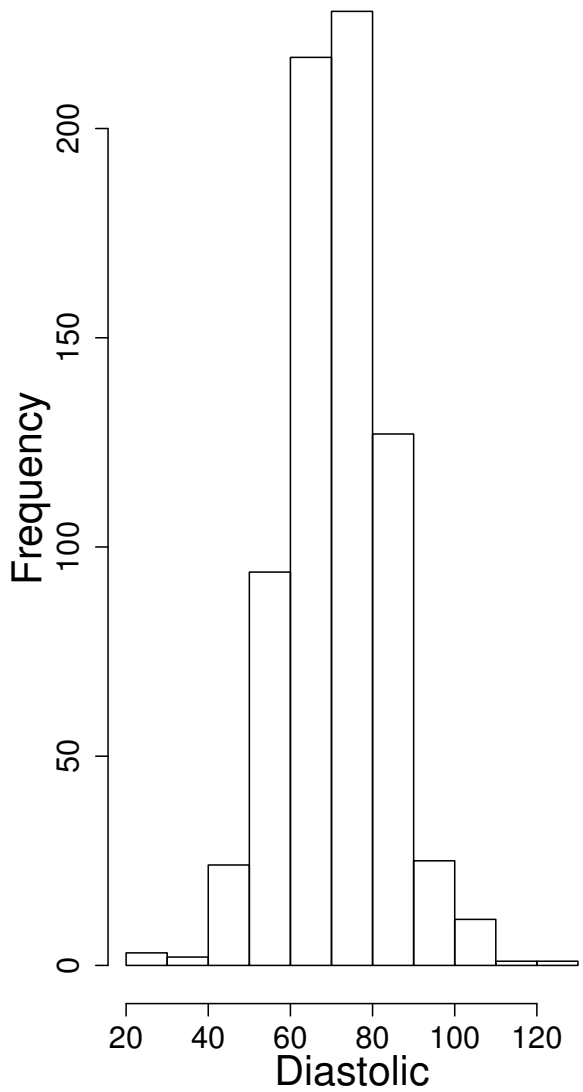
```
## New Summary
> summary(pima)
    pregnant         glucose        diastolic        triceps
 Min.   : 0.0   Min.   : 44   Min.   : 24   Min.   :  7
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 64   1st Qu.: 22
 Median : 3.0   Median :117   Median : 72   Median : 29
 Mean   : 3.8   Mean   :122   Mean   : 72   Mean   : 29
 3rd Qu.: 6.0   3rd Qu.:141   3rd Qu.: 80   3rd Qu.: 36
 Max.   :17.0   Max.   :199   Max.   :122   Max.   : 99
                NA's   :  5   NA's   : 35   NA's   :227
```

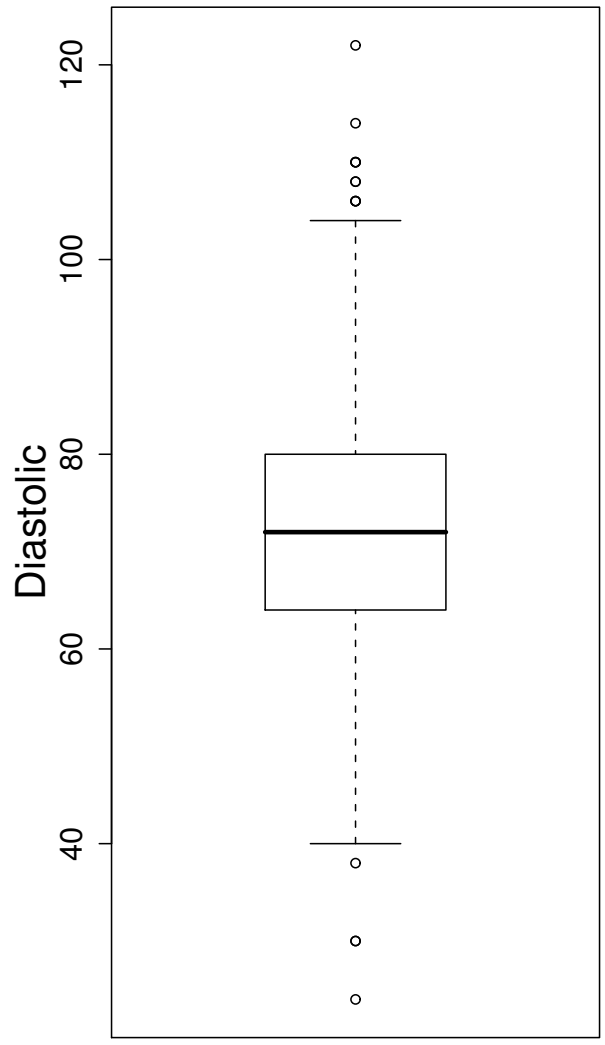
```
    insulin          bmi             diabetes            age       
 Min.   : 14   Min.   :18.2   Min.   :0.08   Min.   :21  
 1st Qu.: 76   1st Qu.:27.5   1st Qu.:0.24   1st Qu.:24  
 Median :125   Median :32.3   Median :0.37   Median :29  
 Mean   :156   Mean   :32.5   Mean   :0.47   Mean   :33  
 3rd Qu.:190   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41  
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81  
 NA's   :374   NA's   :11.0                              
      test       
 negative:500  
 positive:268  
```

```
## Individual summary functions
> mean(pima$diastolic, na.rm=T)
[1] 72.40518
> median(pima$diastolic, na.rm=T)
[1] 72
> range(pima$diastolic, na.rm=T)
[1]  24 122
> quantile(pima$diastolic, na.rm=T)
  0%  25%  50%  75% 100%
  24   64   72   80  122
## Other functions:  var(), sd()
```

## Graphical Summaries: single variable

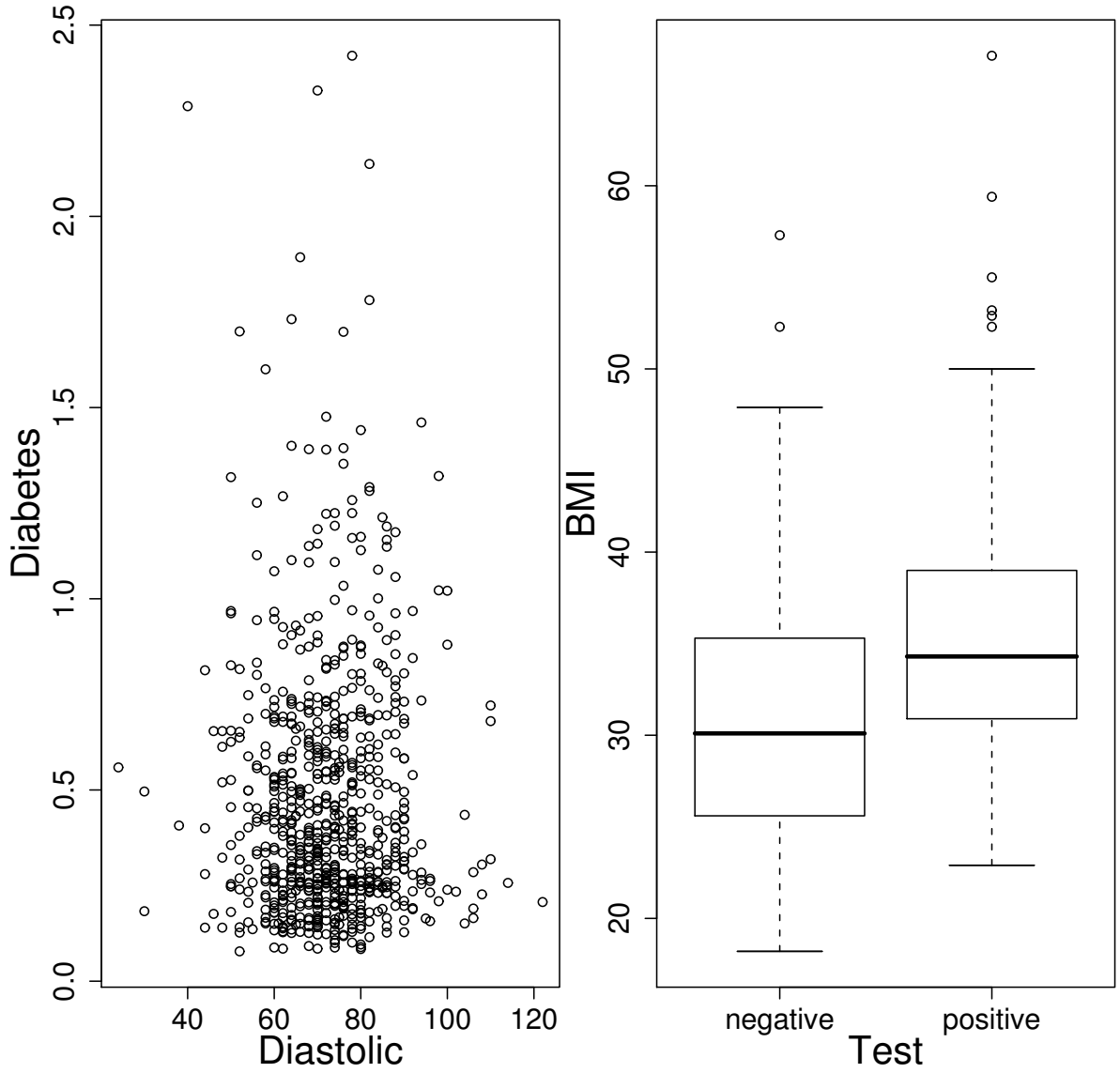
```
> hist(pima$diastolic)
> boxplot(pima$diastolic)
```



## Graphical Summaries: two variables

```
> plot(pima$diastolic, pima$diabetes)
> plot(pima$test, pima$bmi)
```



```
## Selecting Subsets of the Data
## The second row
> pima[2,]
  pregnant glucose diastolic triceps insulin
2        1      85         66      29      NA
       bmi diabetes age     test
2     26.6    0.351  31 negative
## The third column
> pima[,3]
  [1]  72  66  64  66  40  74  50  NA  70 ...
## The (2,3) element
> pima[2,3]
[1] 66
```

```
## The first, second and fourth row
> pima[c(1,2,4), ]
  pregnant glucose diastolic triceps insulin ...
1        6     148        72      35      NA ...
2        1      85        66      29      NA ...
4        1      89        66      23      94 ...
## The third through sixth rows
> pima[3:6, ]
  pregnant glucose diastolic triceps insulin ...
3        8     183        64      NA      NA ...
4        1      89        66      23      94 ...
5        0     137        40      35     168 ...
6        5     116        74      NA      NA ...
```

```
## "Everything but"
> pima[, -c(1,2)]
    diastolic triceps insulin  bmi diabetes age     test
1          72      35      NA 33.6    0.627  50 positive
2          66      29      NA 26.6    0.351  31 negative
3          64      NA      NA 23.3    0.672  32 positive
...

## Cases which have pregnant greater than 14
> pima[pima$pregnant > 14, ]
    pregnant glucose diastolic triceps insulin ...
89        15     136        70      32     110 ...
160       17     163        72      41     114 ...
```

```
## Help
> help(boxplot)
> ?boxplot
> help('*')
> help.start()
```