Introduction to Data Mining

Web Chapter Exploring Data

by Michael Hahsler

Based in Slides by Tan, Steinbach, Karpatne, Kumar

R Code Examples
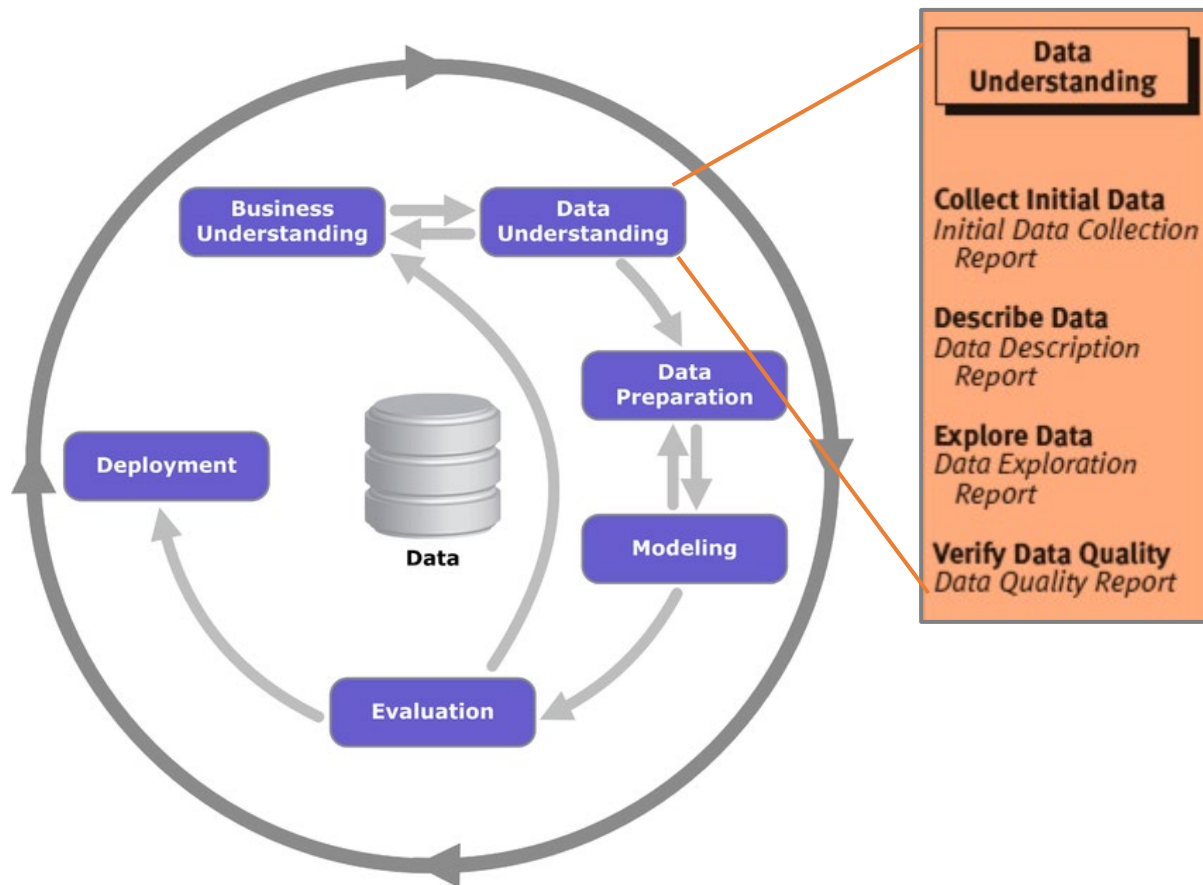
- Available R Code examples are indicated on slides by the R logo

- The Examples are available at https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/

Exploring Data in the Data Mining Process

Topics

- **Exploratory Data Analysis**
- Summary Statistics
- Visualization

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

Topics

- Exploratory Data Analysis
- **Summary Statistics**
- Visualization

Summary Statistics

Summary statistics are numbers that summarize properties of the data

Summarized properties include location and spread for continuous data

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.**

Measures of Location: Mean and Median

- For quantitative features.

- 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 that other measures are often used.

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

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

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

Percentiles

- Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value $x_{p\%}$ of x such that p% of the observed values of x are less than $x_{p\%}$.

- Example: the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

Percentiles

Median – 50% of the cases has a smaller value & 50% are larger
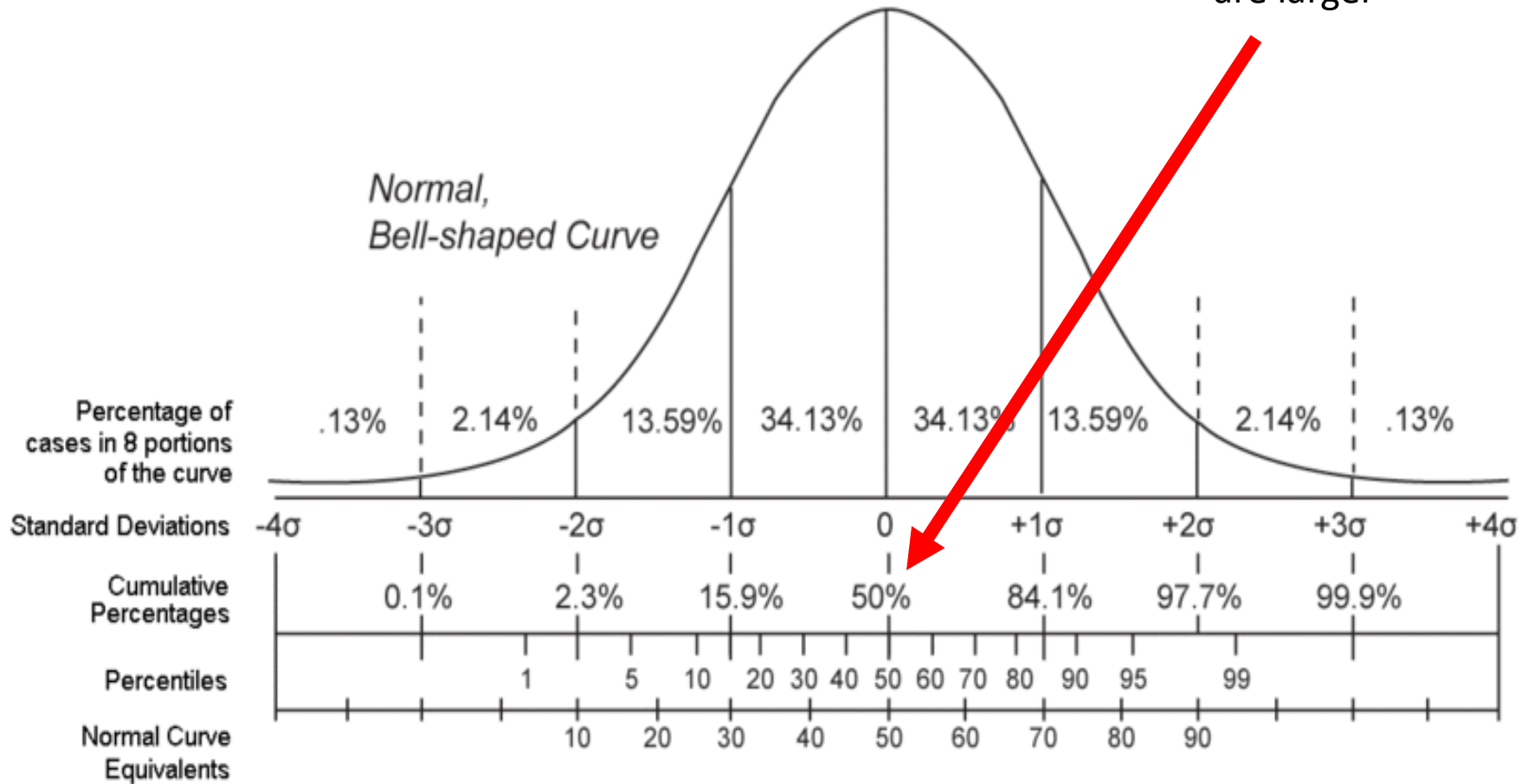
Multivariate Summary Statistics

Object	x_1	x_2
1	12	15
2	2	4
...	...	...
m	18	4

- Covariance between features i and j

$$s_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

- Correlation

$$r_{ij} = \frac{s_{ij}}{s_i \, s_j}$$

- s_i is the variance of feature i

Correlation

- Correlation measures the (linear) relationship between two variables.

- To compute Pearson correlation (Pearson's Product Moment Correlation), we standardize data objects, p and q, and then take their dot product

$$\rho = \frac{cov(X,Y)}{sd(X)\,sd(Y)}$$

- Estimation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Correlation is often used as a measure of similarity.

Visually Evaluating Correlation

Scatter plots showing the similarity from −1 to 1.
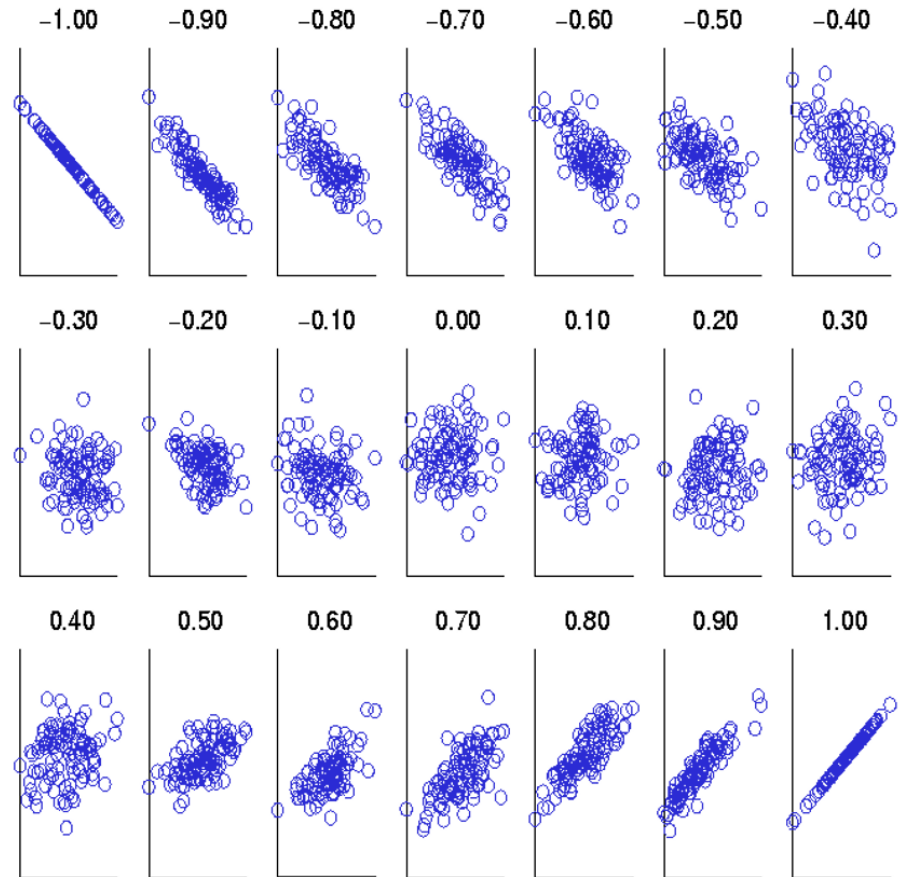
Rank Correlation

- Measure the degree of similarity between two ratings (e.g., ordinal data).

- Is more robust against outliers and does not assume normality of data or linear relationship like Pearson Correlation.

- Measures (all are between -1 and 1)
  — Spearman's Rho: Pearson correlation between ranked variables.
  — Kendall's Tau

$N_s...concordant\ pair$
$N_d...discordant\ pair$

$$\tau = \frac{N_s - N_d}{\frac{1}{2}n(n-1)}$$

  — Goodman and Kruskal's Gamma

$$\gamma = \frac{N_s - N_d}{N_s + N_d}$$

Topics

- Exploratory Data Analysis
- Summary Statistics
- **Visualization**

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
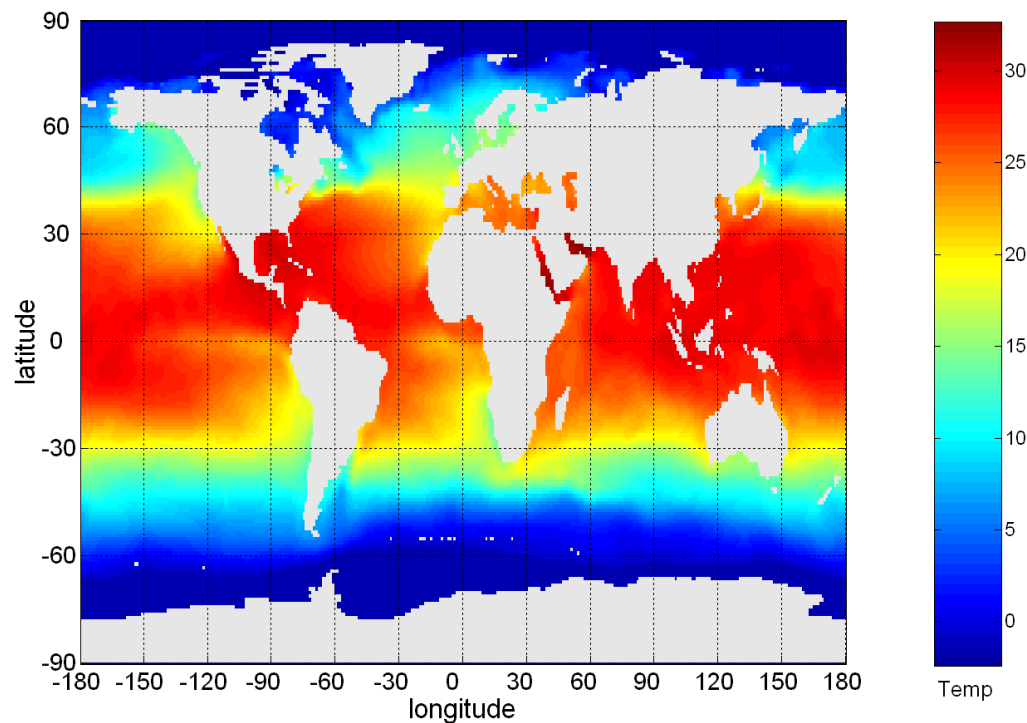
Representation

- Is 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.

Ararrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data

Example:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

The Iris Dataset

Many of the exploratory data techniques are illustrated with the Iris Plant data set.

- Included as a demo datasert in many tools (R, scikit-learn, Rapidminer, …).

- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

- From the statistician R.A. Fisher

- 150 flowers, three types (classes).

- Four (non-class) attributes

Iris Versicolor

Iris Virginica

Iris Setosa

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa

. . .

Distribution: Histograms

- 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)
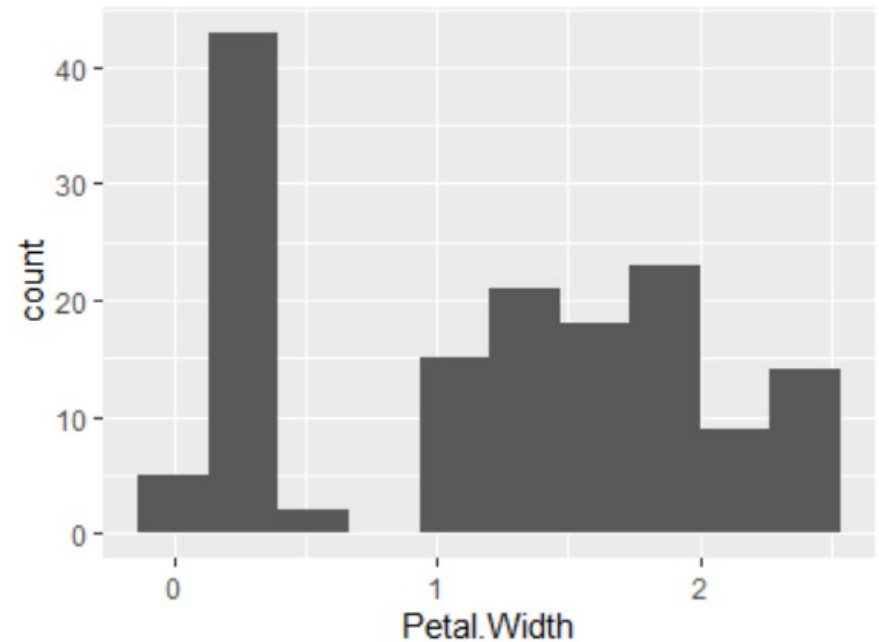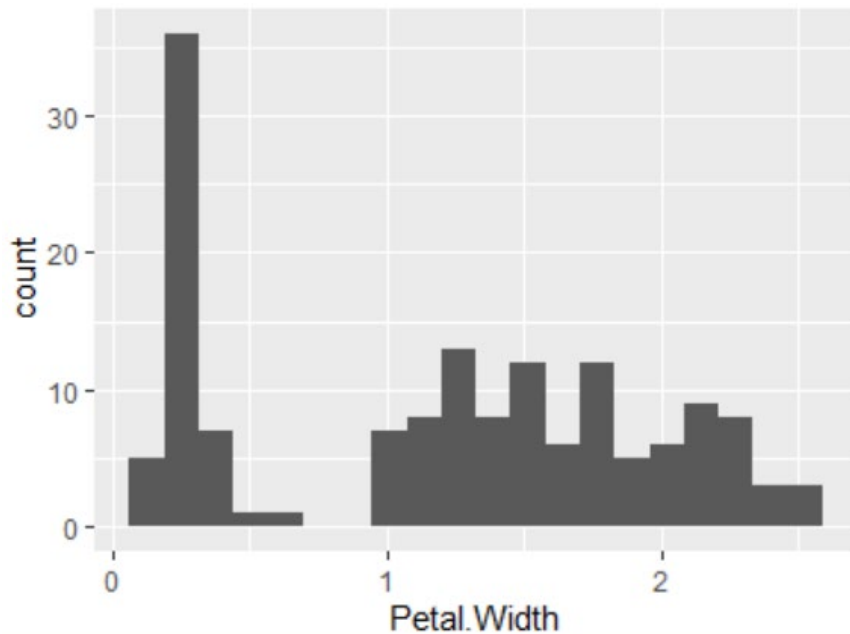
Distribution Box Plots

- Invented by J. Tukey
- Simplified version of a PDF/histogram.
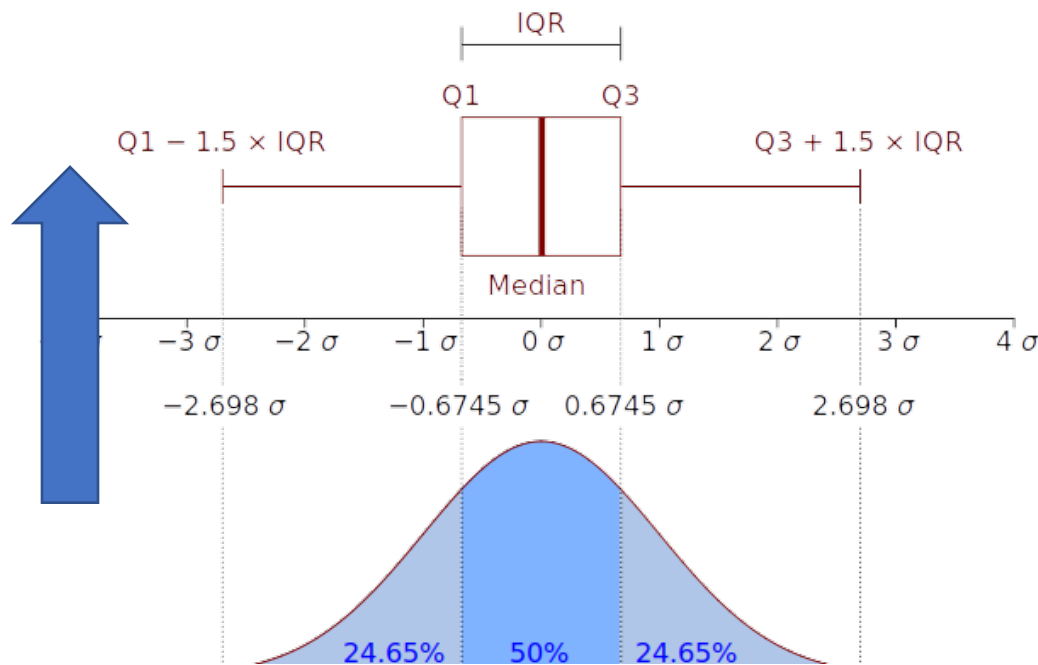
PDF or histogram
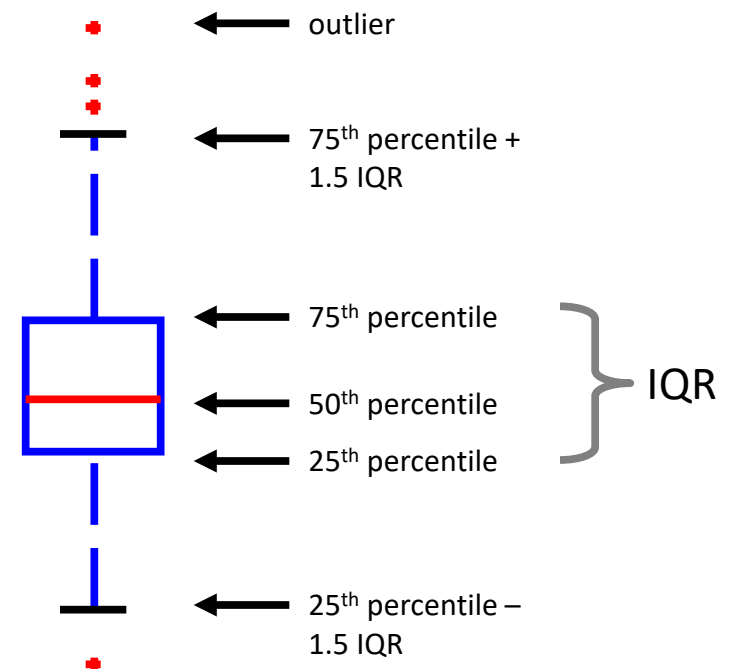
Examples of Box Plots

- Box plots can be used to compare attributes or subgroups.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
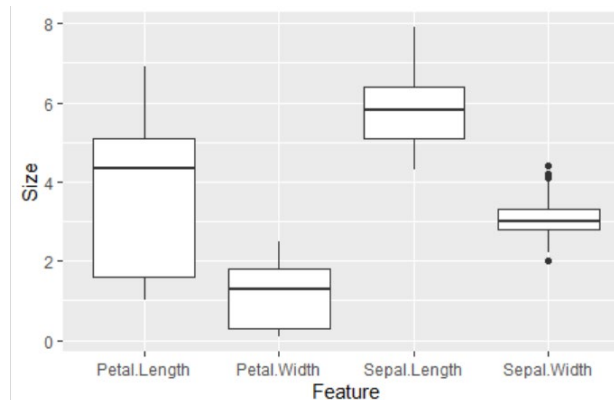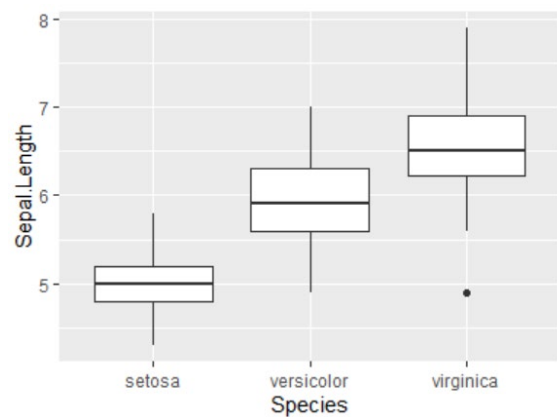
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes

Example: petal width and petal length. What does this tell us?
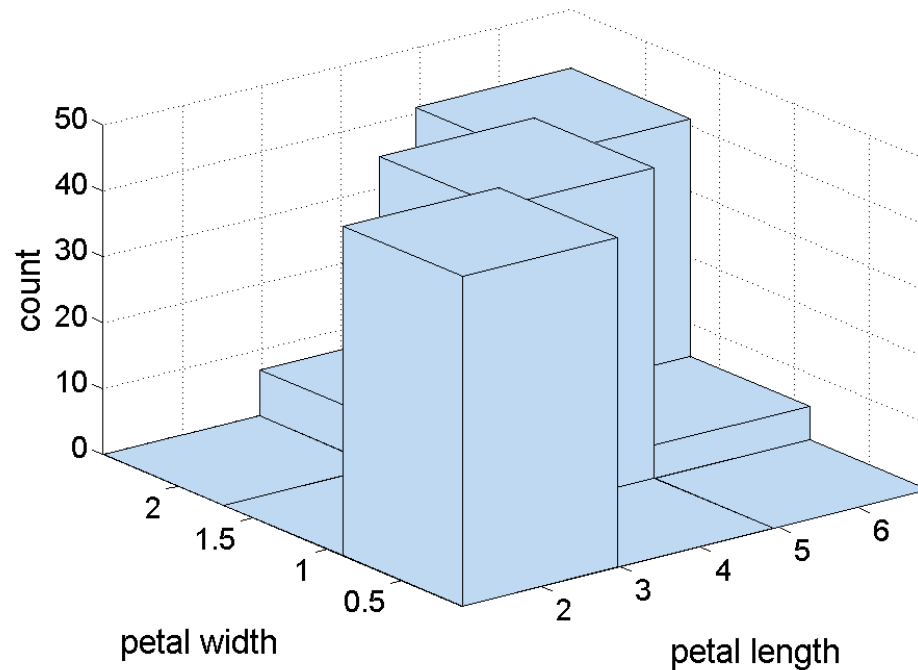
Scatter Plots

- Attributes values determine the position

- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots

- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
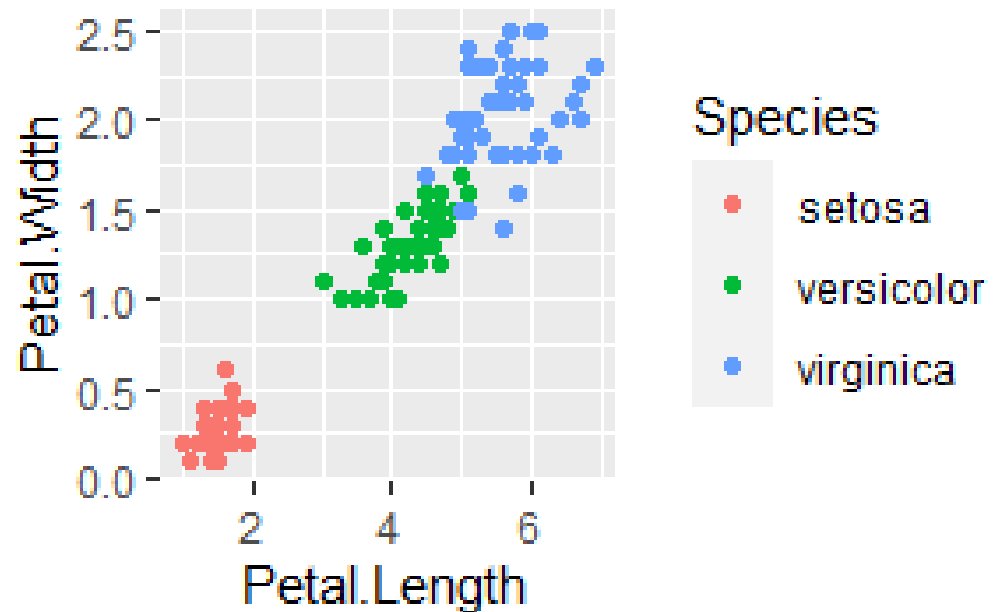
Scatter Plot Array of Iris Attributes
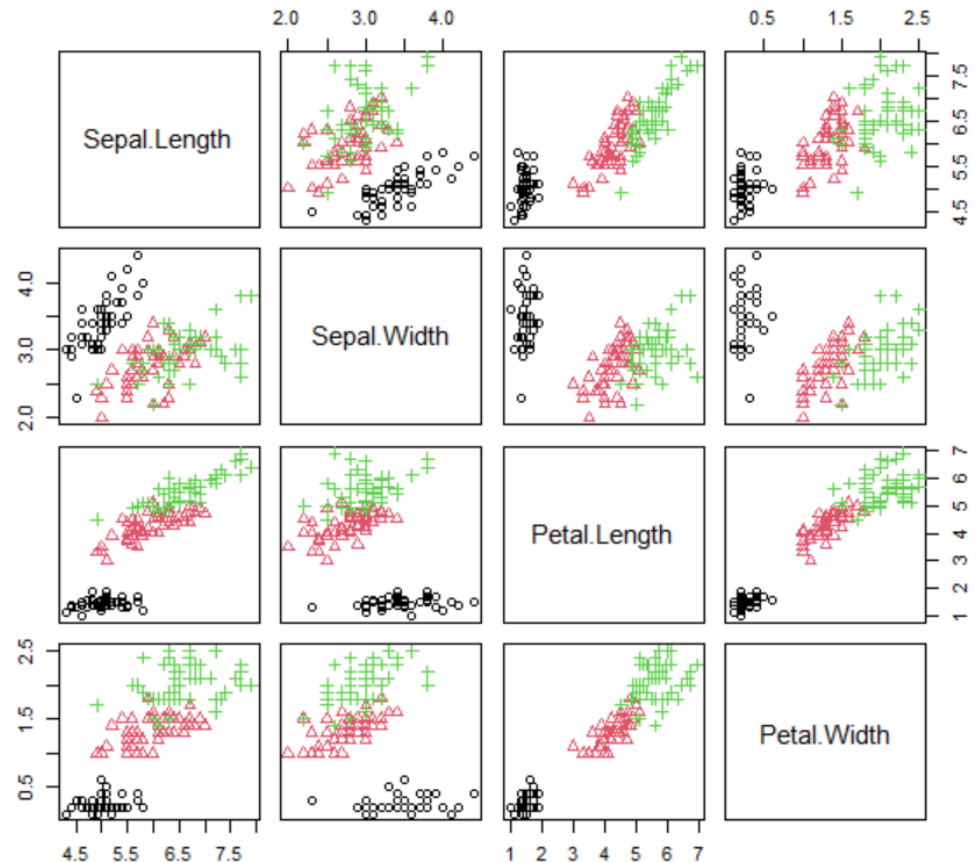
Contour Plots

- 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.

Celsius
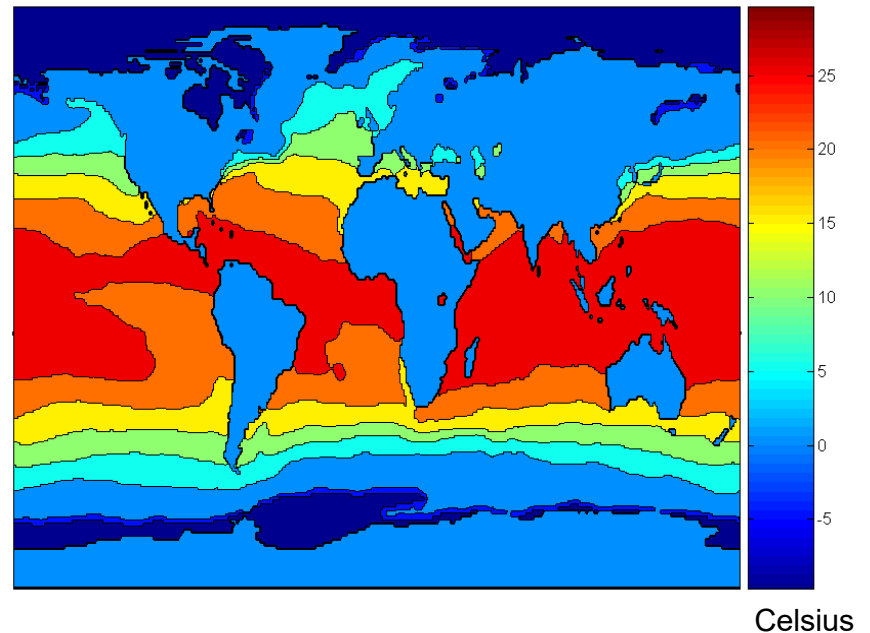
Matrix Plots

- Can plot a data matrix

- Can be useful when objects are sorted according to class

- Typically, the attributes are normalized to prevent one attribute from dominating the plot

- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects

Example: The Iris Data Matrix

Deviation form feature mean
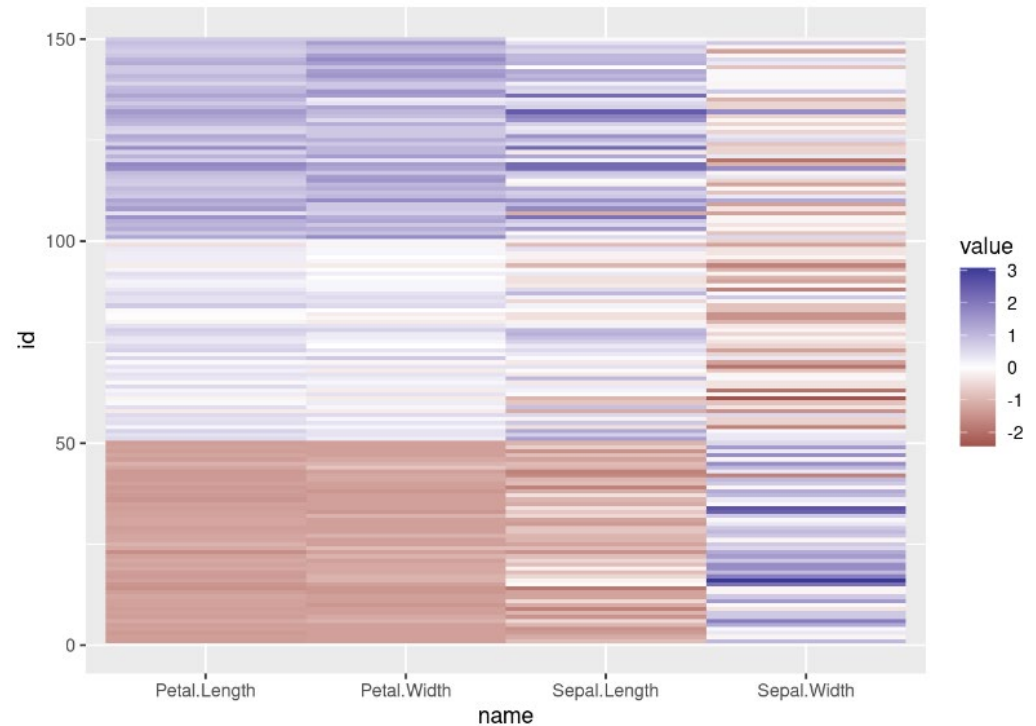
Example: The Iris Correlation Matrix
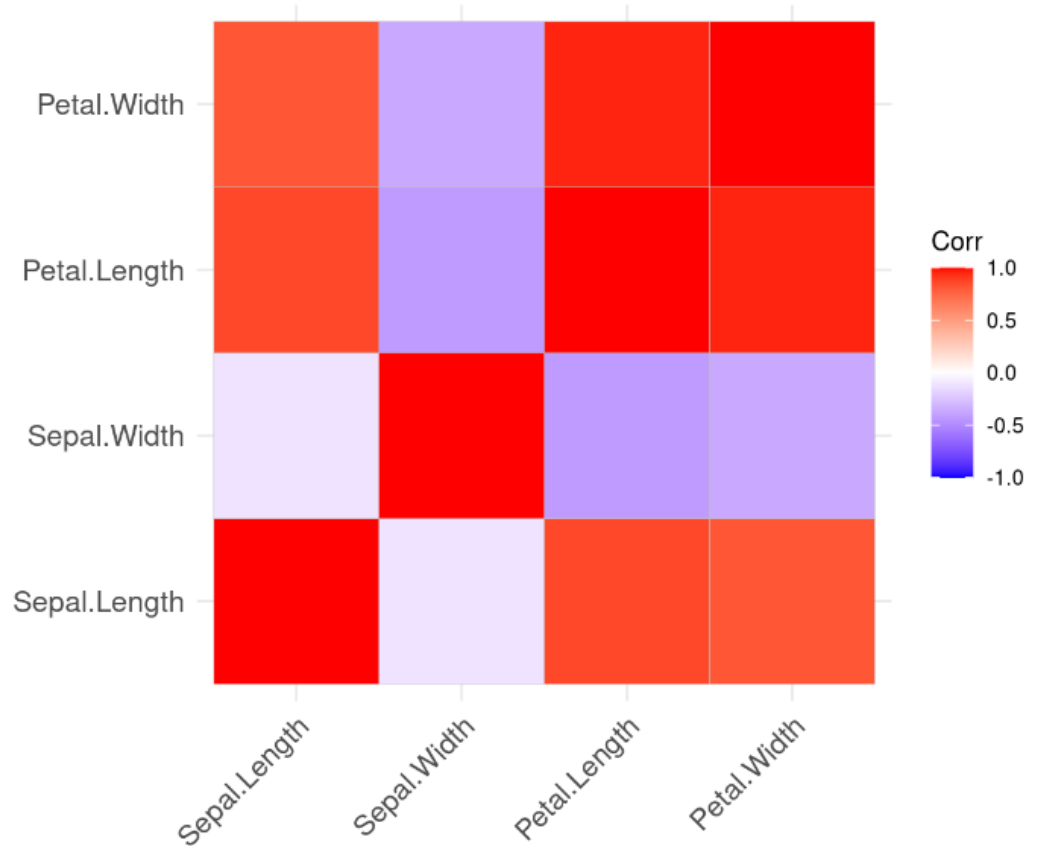
Parallel Coordinates

- Used to plot the attribute values of high-dimensional data

- Instead of using perpendicular axes, use a set of parallel axes

- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line

- Thus, each object is represented as a line

- Often, the lines representing a distinct class of objects group together, at least for some attributes

- Ordering of attributes is important in seeing such groupings

Example: Parallel Coordinates Plots for Iris Data

Reordered features
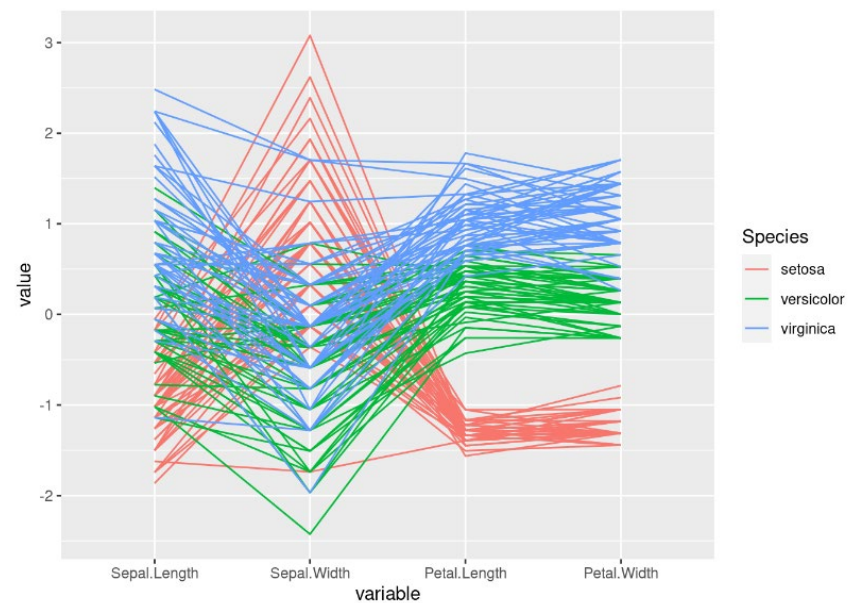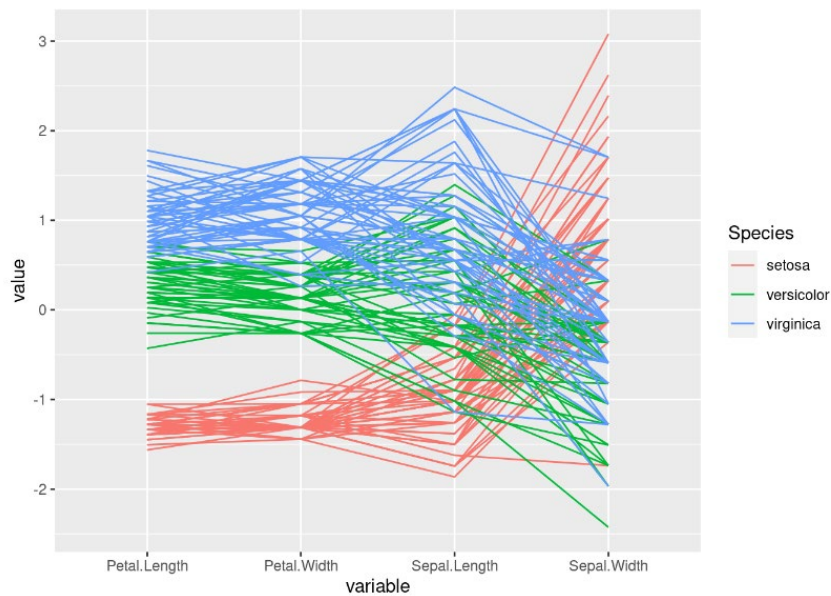
Other Visualization Techniques

- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon

- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Example: Star Plots for Iris Data
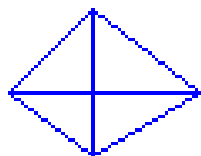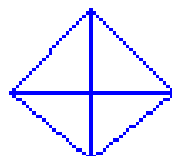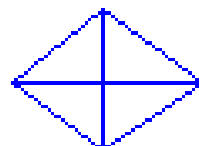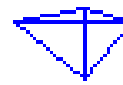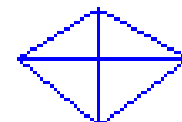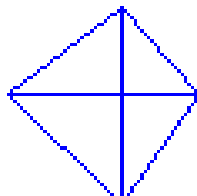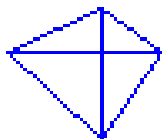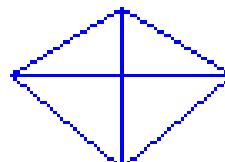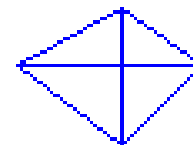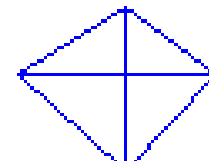
Example: Chernoff Faces for Iris Data

- Setosa

- Versicolor

- Virginica

Conclusion

- Exploring data is the first step when working with data.

- The goal is to:
    1. Understand what data is available.
    2. Assess data distributions and how variables relate to each other.
    3. Assess data quality.

- Understanding the data is necessary to decide on data preparation and modeling.