CEFET/RJ

Laboratório Nacional de Computação Científica
UNIDADE DE PESQUISA DO MCTI

EXPLORATORY ANALYSIS

Eduardo Ogasawara
eogasawara@ieee.org
https://eic.cefet-rj.br/~eogasawara

Types of Data Sets

- Record
  - Relational datasets
- Matrix
  - numerical matrix, crosstabs
- Documents
  - texts, term-frequency vector
- Transactions
- Graph and network
  - World Wide Web
  - Social or information networks
- Ordered
  - Temporal data: time-series
  - Sequential data: transaction sequences
- Spatial, image, and multimedia
  - Spatial data: maps
  - Images
  - Videos

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Documents	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

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

Data	PIB - R$ (milhões)
1990.01	0.2
1990.02	0.4
1990.03	0.8
1990.04	0.7
1990.05	0.8

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Important Characteristics of Structured Data

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

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Relational data

- Data sets are made up of data objects
- A data object represents an entity
  - sales database: customers, store items, sales
  - medical database: patients, treatments, illness
  - university database: students, professors, courses
- Attributes describe data objects
- Database
  - rows: data objects (tuples)
  - columns: attributes

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Attributes

- Attribute (or dimensions, features, variables)
  - a data field, representing a characteristic or feature of a data object
  - E.g., customer_ID, name, address
- Types: Nominal, Binary, Ordinal, Numeric

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Attribute Types

- Nominal: categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- Binary
  - Attribute with only two 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 the most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking), but magnitude between successive values is not known
  - Size = {small, medium, large}, grades, army rankings

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - E.g., the 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., the temperature in Kelvin, length, counts, monetary quantities

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Sometimes, represented as integer variables
- 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

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

Iris Dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
numeric	numeric	numeric	numeric	factor

	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
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica

[1] Kaggle, 2020, *Iris Species*, https://www.kaggle.com/uciml/iris.

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data:
    - central tendency, variation and spread
- Data centrality and dispersion characteristics
  - median, max, min, quantiles, outliers, variance
- Numerical dimensions correspond to sorted intervals
  - Boxplot or quantile analysis on sorted intervals

[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Descriptive Measures

- Centrality
  - Mean (algebraic measure)
    - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Median
    - Middle value if an odd number of values, or weighted average of the middle two values otherwise
  - Mode
    - The value that occurs most frequently in the data
    - Unimodal, bimodal, trimodal
    - Empirical formula:
      - $mean - mode = 3 \cdot (mean - median)$
- Dispersion
  - Variance and standard deviation
    - Variance: (algebraic, scalable computation)
    - Standard deviation (σ): square root of the variance (σ^2)
      - $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} {x_i}^2}{n} - \mu^2$

[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five numbers summary: min, Q_1, median, Q_3, max
  - Boxplot

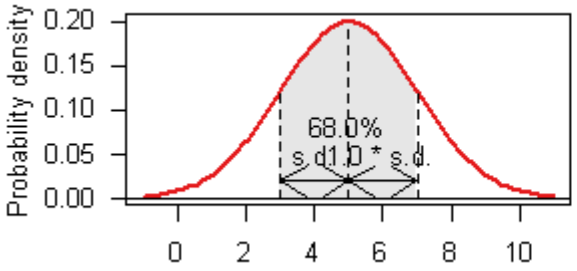
Statistics	Freq
Min.	4.300000
1st Qu.	5.100000
Median	5.800000
Mean	5.843333
3rd Qu.	6.400000
Max.	7.900000

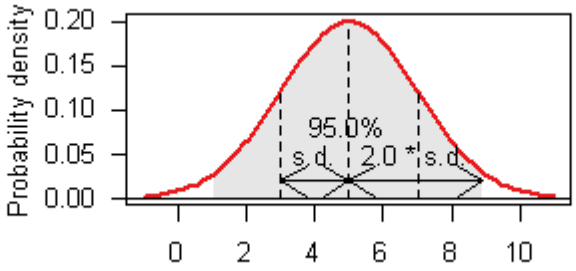
```
[1] "IQR=1.3"
```

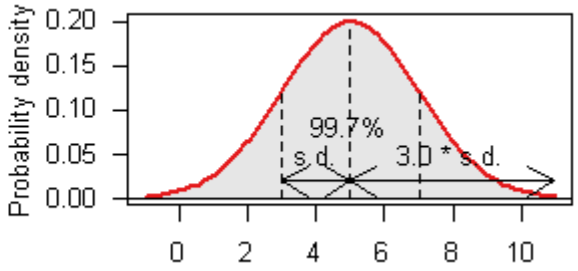
[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it
  - When distribution is normal, values below $-2.698\boldsymbol{\sigma}$ or greater than $2.698\boldsymbol{\sigma}$ are considered outliers







[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Exercises

- Read CSV file presented at:
  - https://eic.cefet-rj.br/~eogasawara/data/wine.csv
- Read the description of columns:
  - http://archive.ics.uci.edu/ml/datasets/Wine
- Present the first rows of the dataset
- What is the class label?
- Attributes can be binary, categorical, ordinal, numeric. What are the types of the attributes?
  - Hint: use command class
  - Hint: since you need to check all attribute. Combine class with sapply.
- Compute the distribution of attributes
  - Hint use command summary
- By looking on the results of previous step:
  - Do we have missing values?
  - Do we have outliers?

Take some time to practice the examples
https://nbviewer.jupyter.org/github/eogasawara/mylibrary/blob/master/ExploratoryAnalysis.ipynb

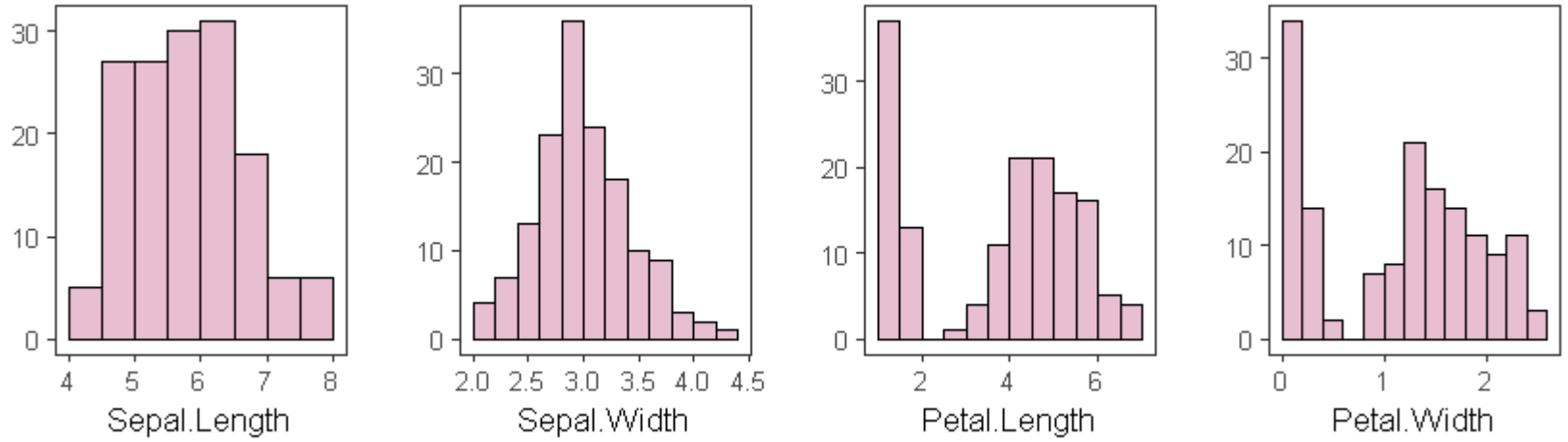
Graphic Displays of Basic Statistical Descriptions

- Histogram
- Boxplot
- Density distribution

[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.

Histogram Analysis

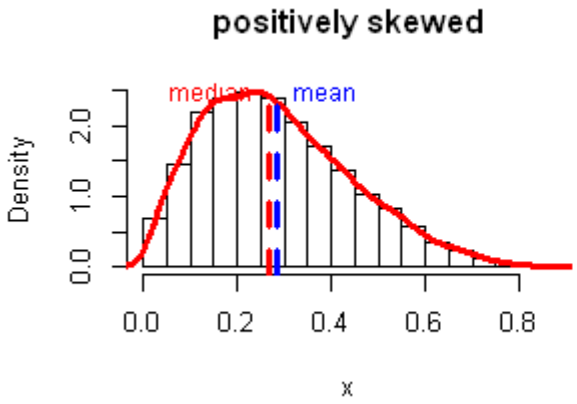
- The histogram displays values of tabulated frequencies
- It shows what proportion of cases into each category
- The area of the bar that denotes the value
  - It is a crucial property when the categories are not of uniform width
- The categories specify non-overlapping intervals of some variable
- The categories (bars) must be adjacent

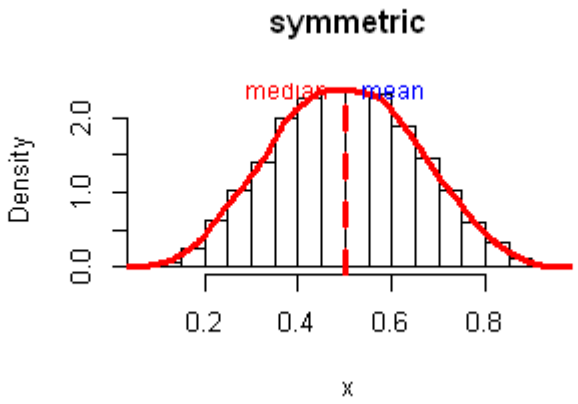


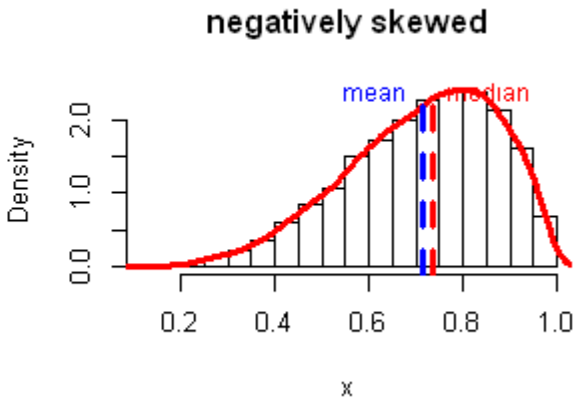
[1] R.J. Larsen and M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Symmetric vs. Skewed Data

- Median and mean for:
  - positive, symmetric, and negatively skewed data



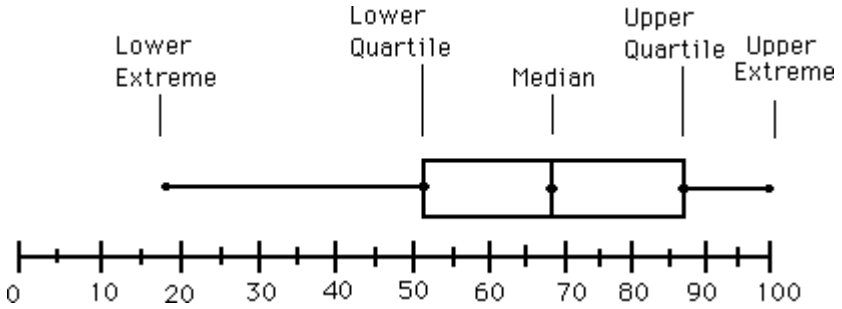




[1] R.J. Larsen and M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

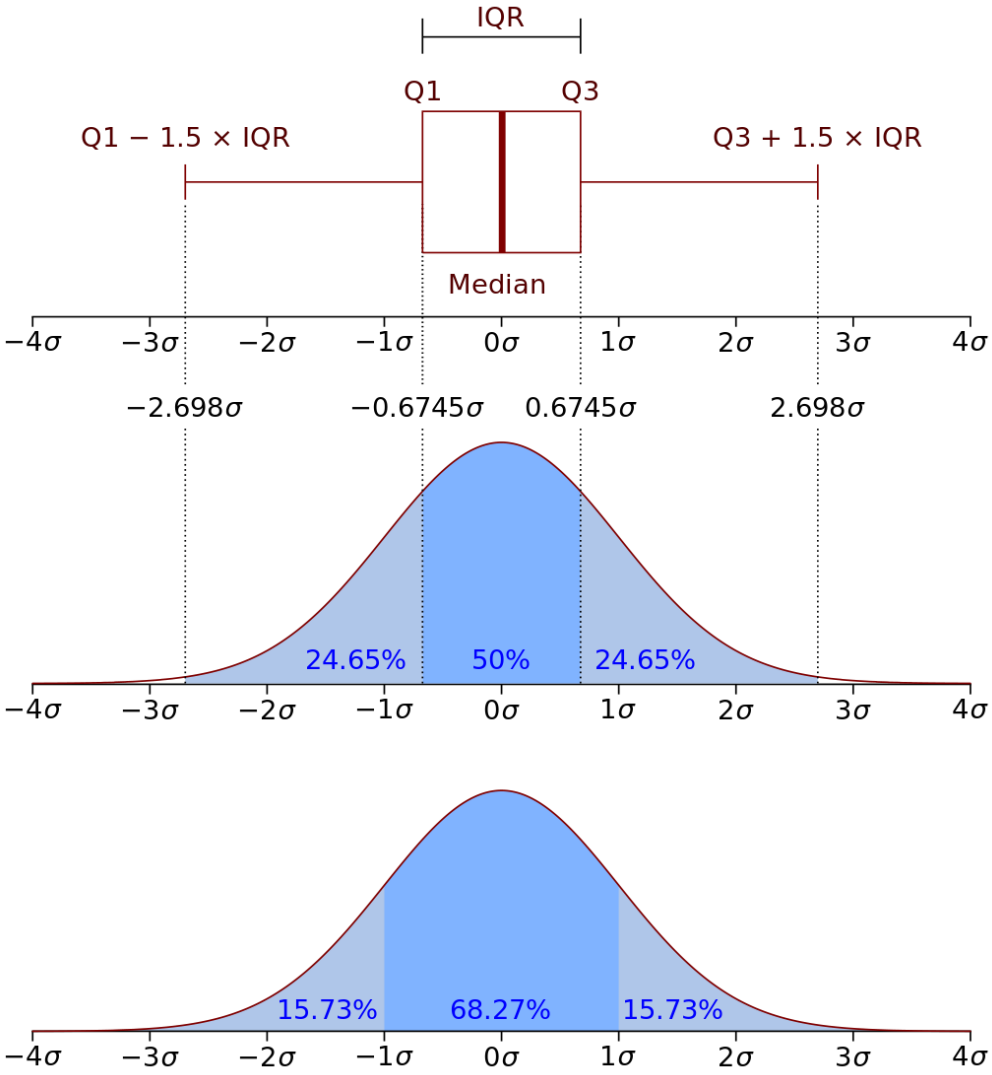
Boxplot Analysis

- Five-number summary of a distribution
  - Min., Q1, Median, Q3, Max.
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - **A line within the box marks the median**
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers are values:
    - higher than Q3 + 1.5 x IQR
    - lower than Q1 - 1.5 x IQR



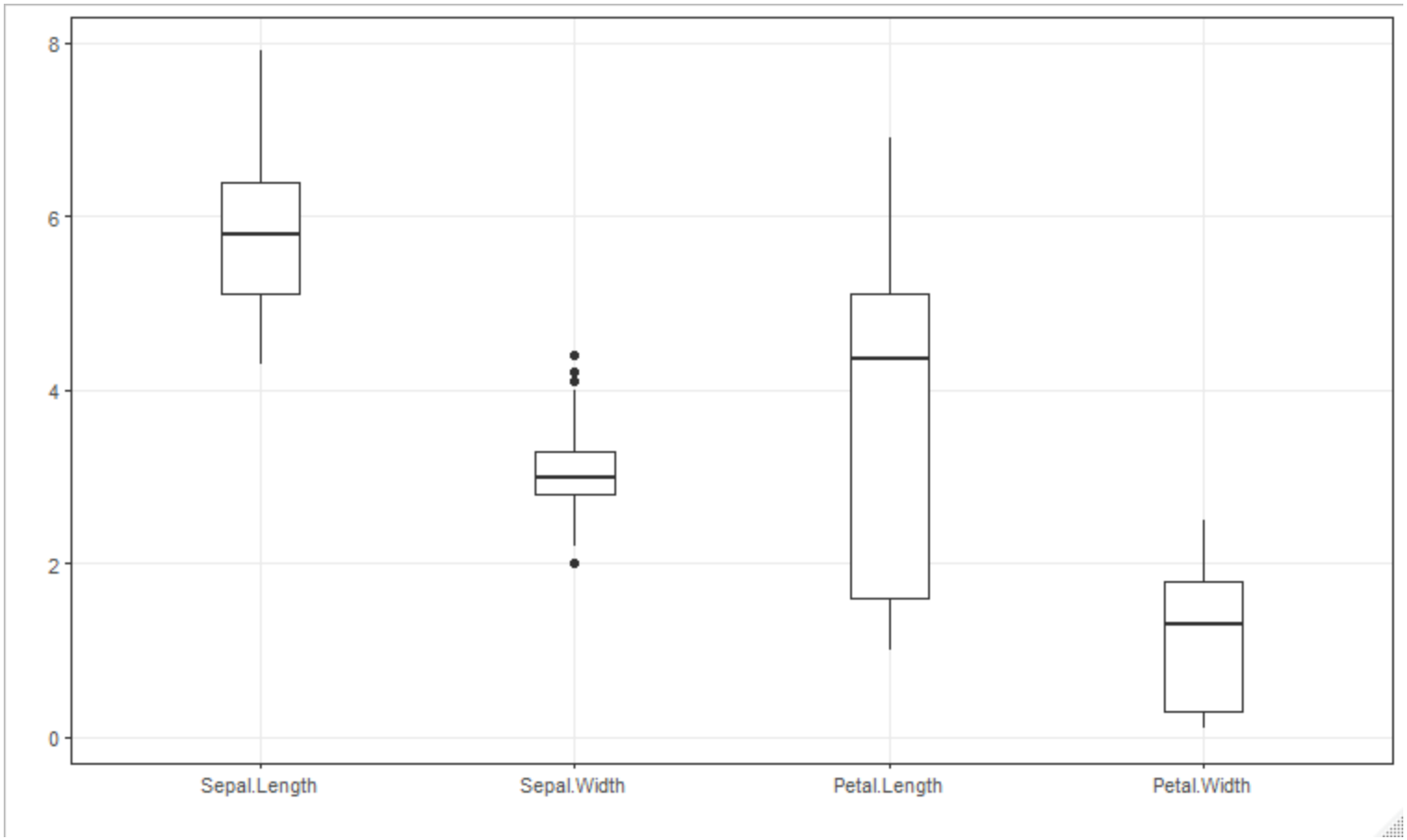
[1] R. McGill, J.W. Tukey, and W.A. Larsen, 1978, Variations of box plots, *American Statistician*, v. 32, n. 1, p. 12–16.

Outliers



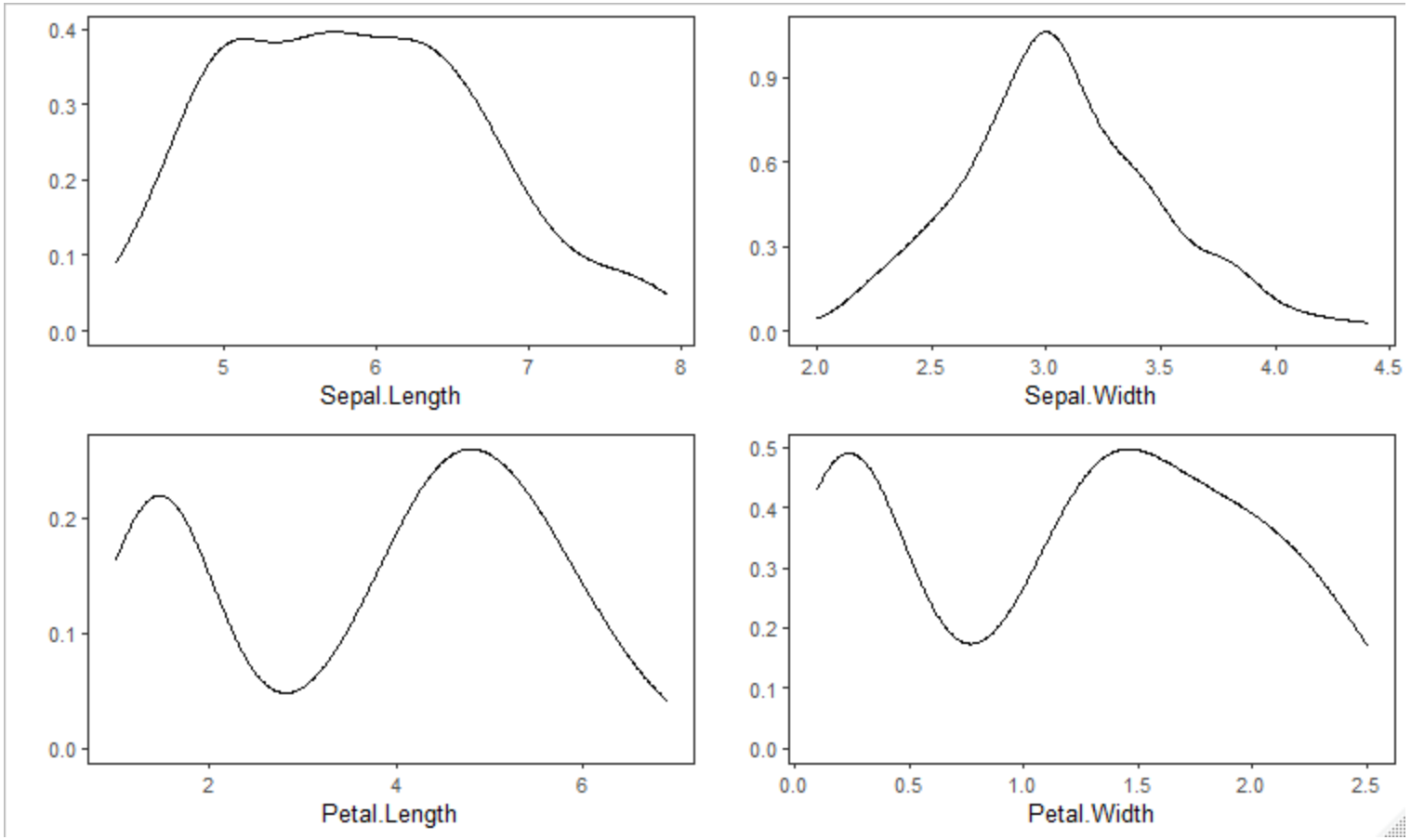
[1] R.J. Larsen and M.L. Marx, 2017, *An Introduction to Mathematical Statistics and Its Applications*. Pearson Education.
https://en.wikipedia.org/wiki/Box_plot

Boxplot for all variables



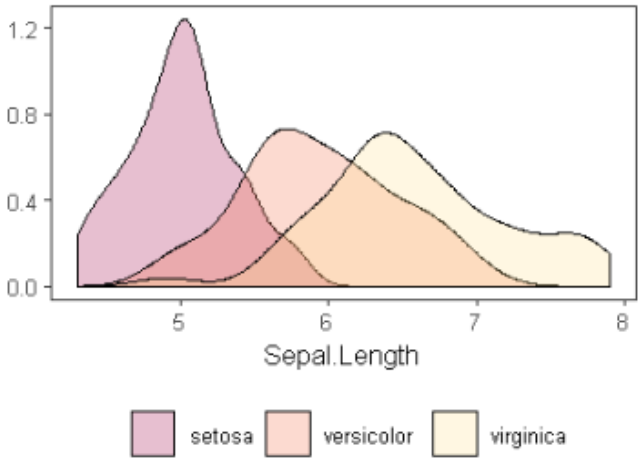
[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

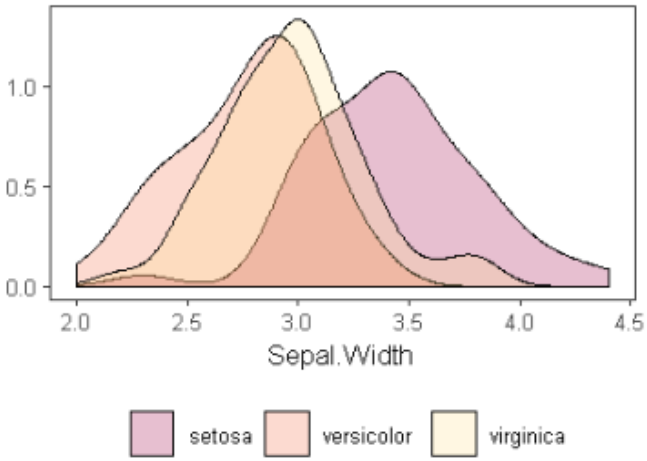
Probability density function

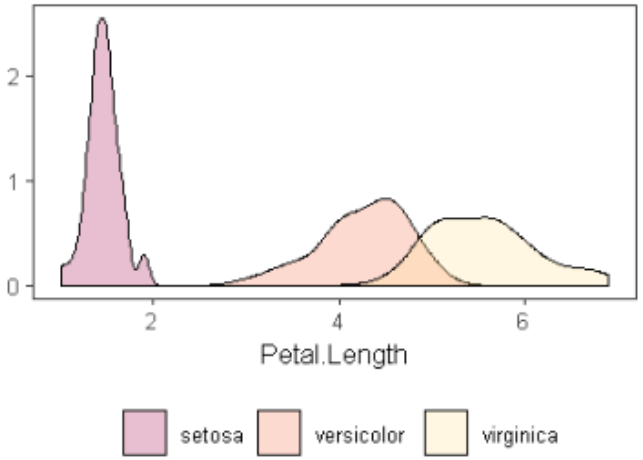


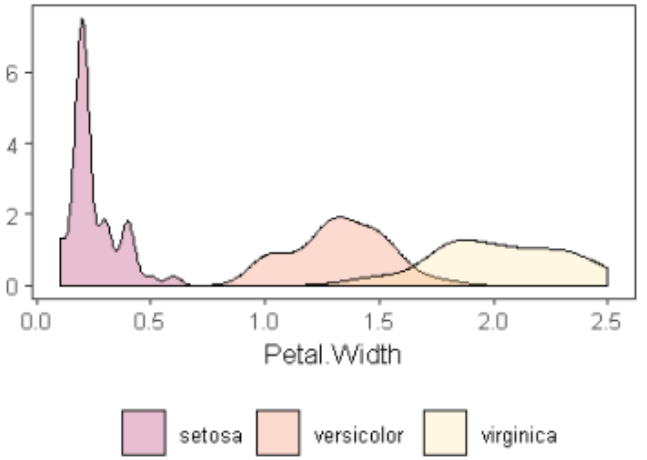
[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

Density distributions per class label



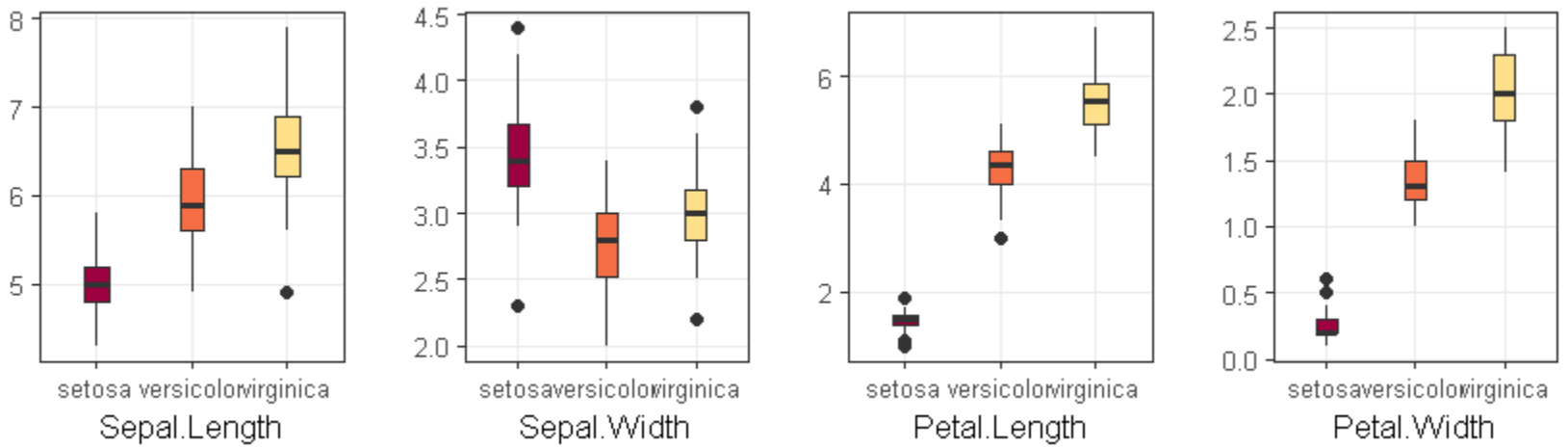






[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

Boxplot per class label



[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

Exercises

- Plot the density distribution for each attribute
  - Explain each one
- Plot the density distribution for each attribute differencing the class label (use colors)
  - Explain each one
- Plot the boxplot for each attribute
  - Explain each one
- Plot the boxplot for each attribute differencing the class label (use colors)
  - Explain each one
- Considering the problem of trying to find the class label:
  - Which attributes are interesting?
  - Why?

Take some time to practice the examples
https://nbviewer.jupyter.org/github/eogasawara/mylibrary/blob/master/ExploratoryAnalysis.ipynb

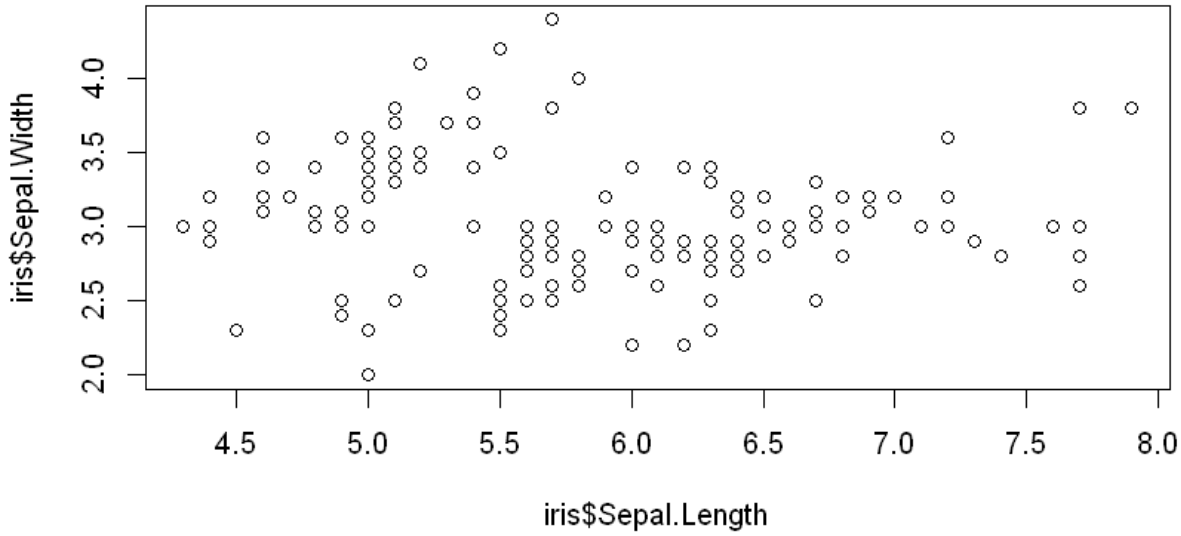
Graphic Displays of Basic Statistical Descriptions

- Scatter plot
- Correlation analysis
- Scatter matrix

[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

Scatter plot

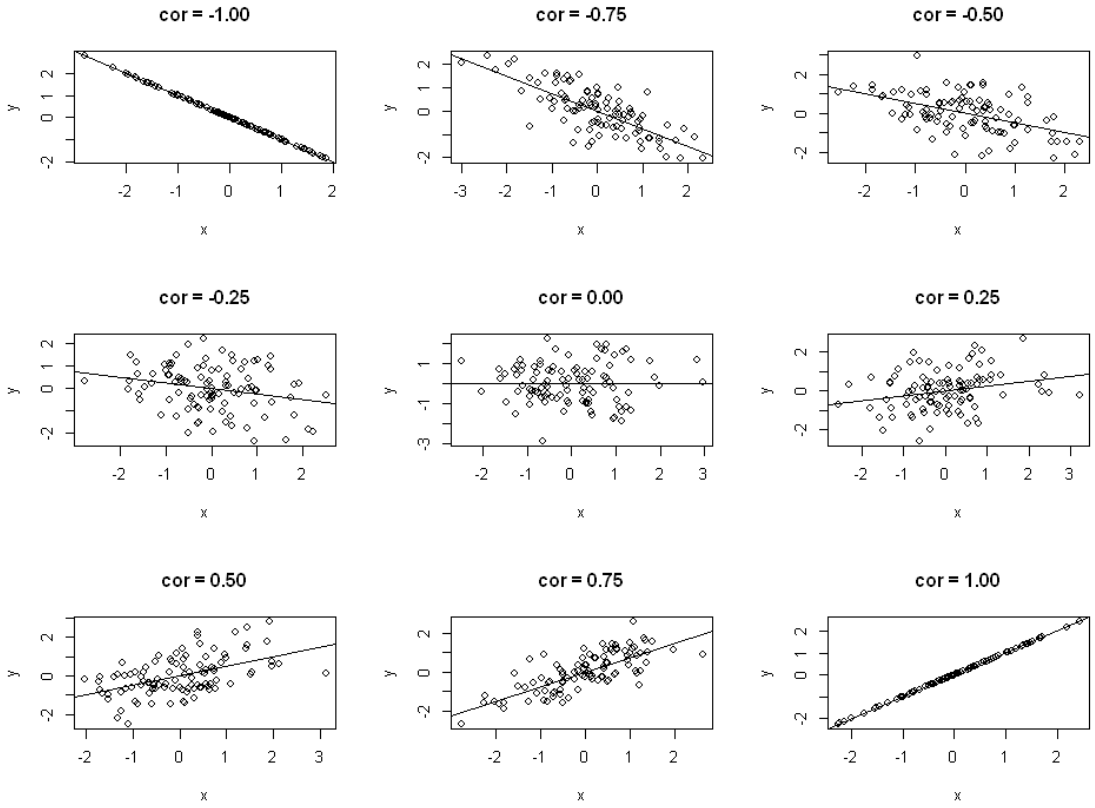
- Provides the first look at bivariate data to see clusters of points, outliers
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

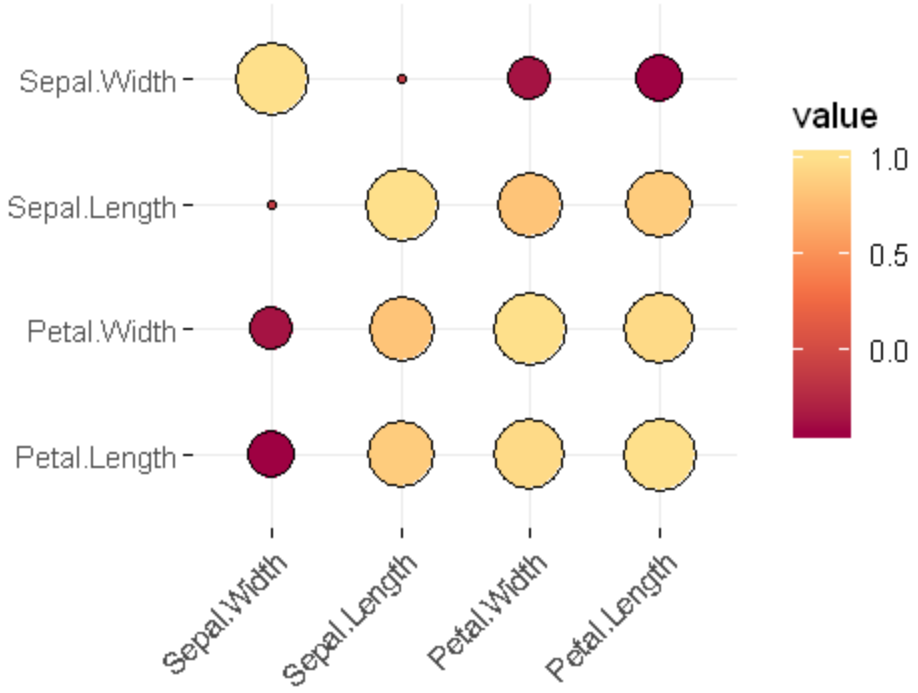
Data correlation

- The first row presents negatively correlated data
- The second row presents uncorrelated data
- The third row presents positively correlated data



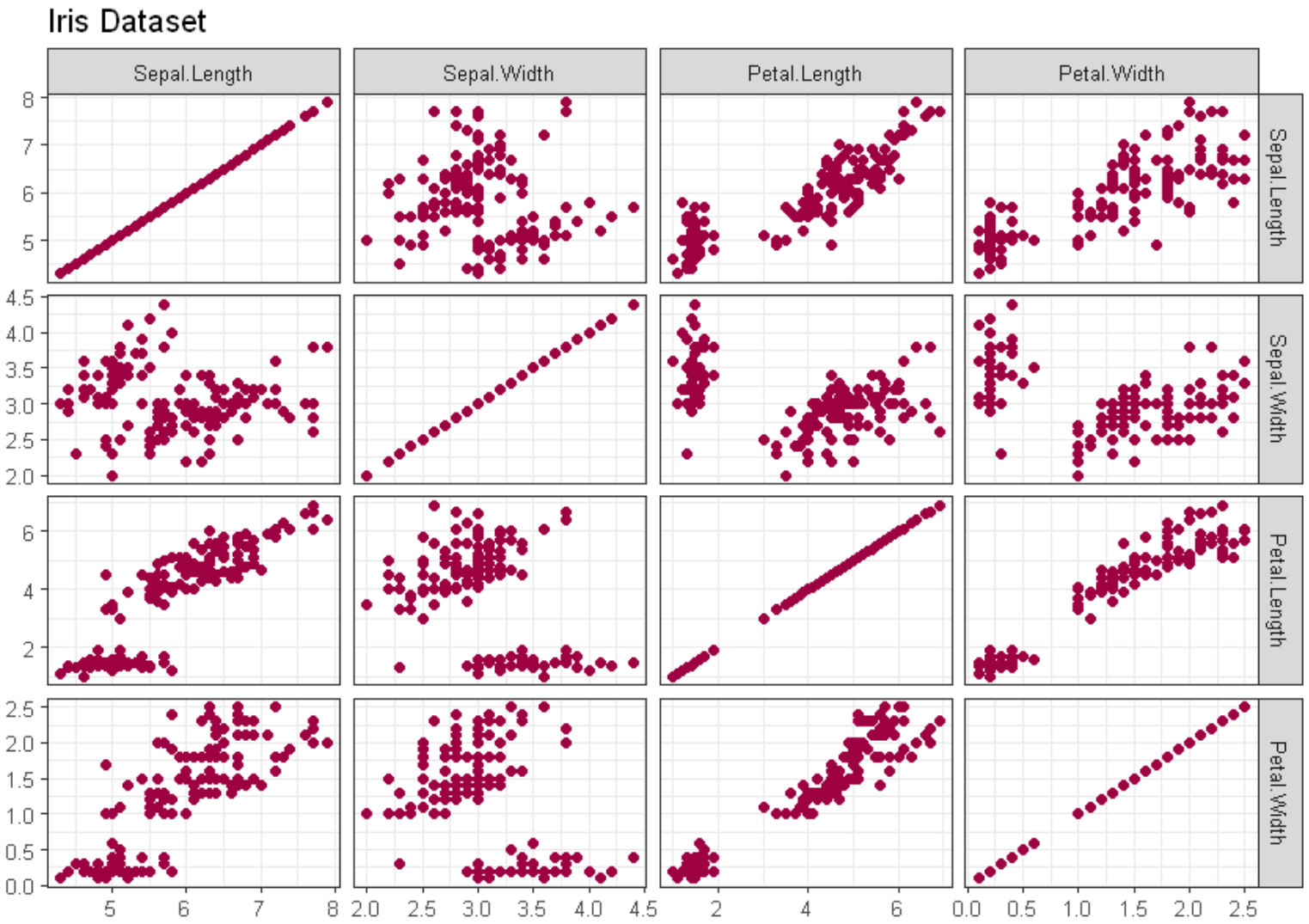
[1] K.J. Keen, 2018, *Graphics for Statistics and Data Analysis with R*. CRC Press.

Matrix plot: Correlation analysis



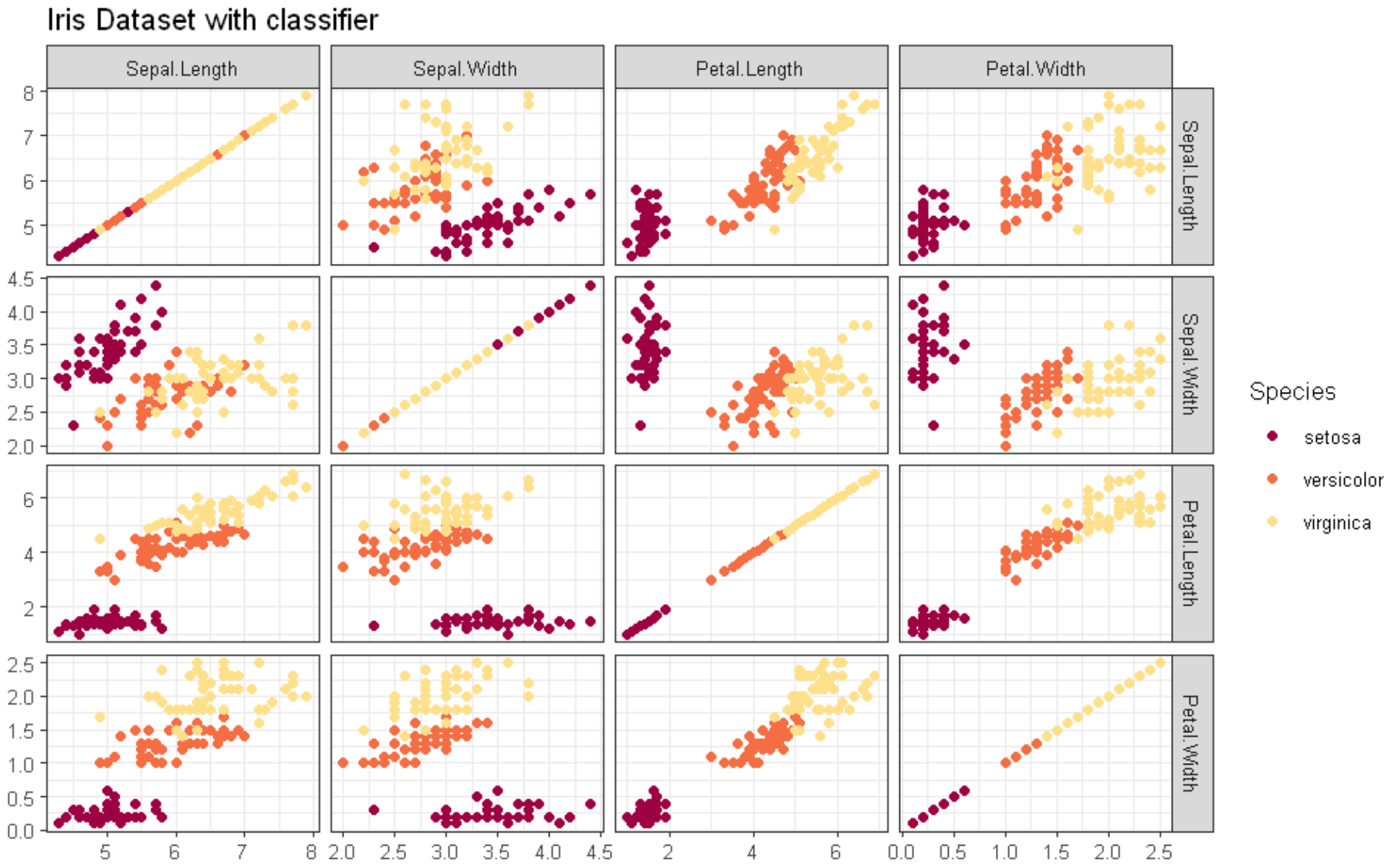
[1] M. Friendly, 2002, Corrgrams: Exploratory displays for correlatigon matrices, *American Statistician*, v. 56, n. 4, p. 316–324.

Scatter Matrix plot



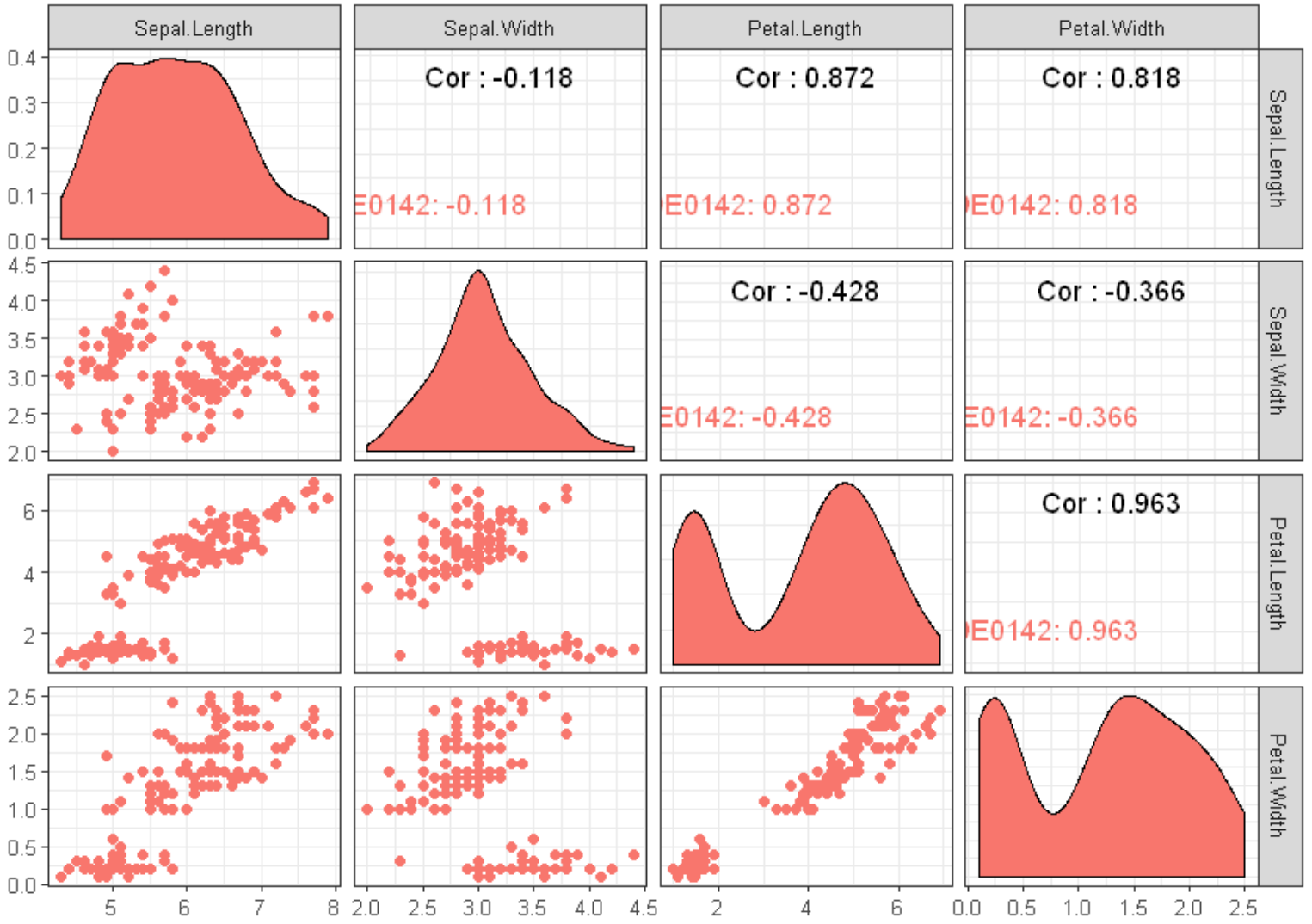
[1] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, 2008, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 6, p. 1141–1148.

Scatter Matrix plot with a class label



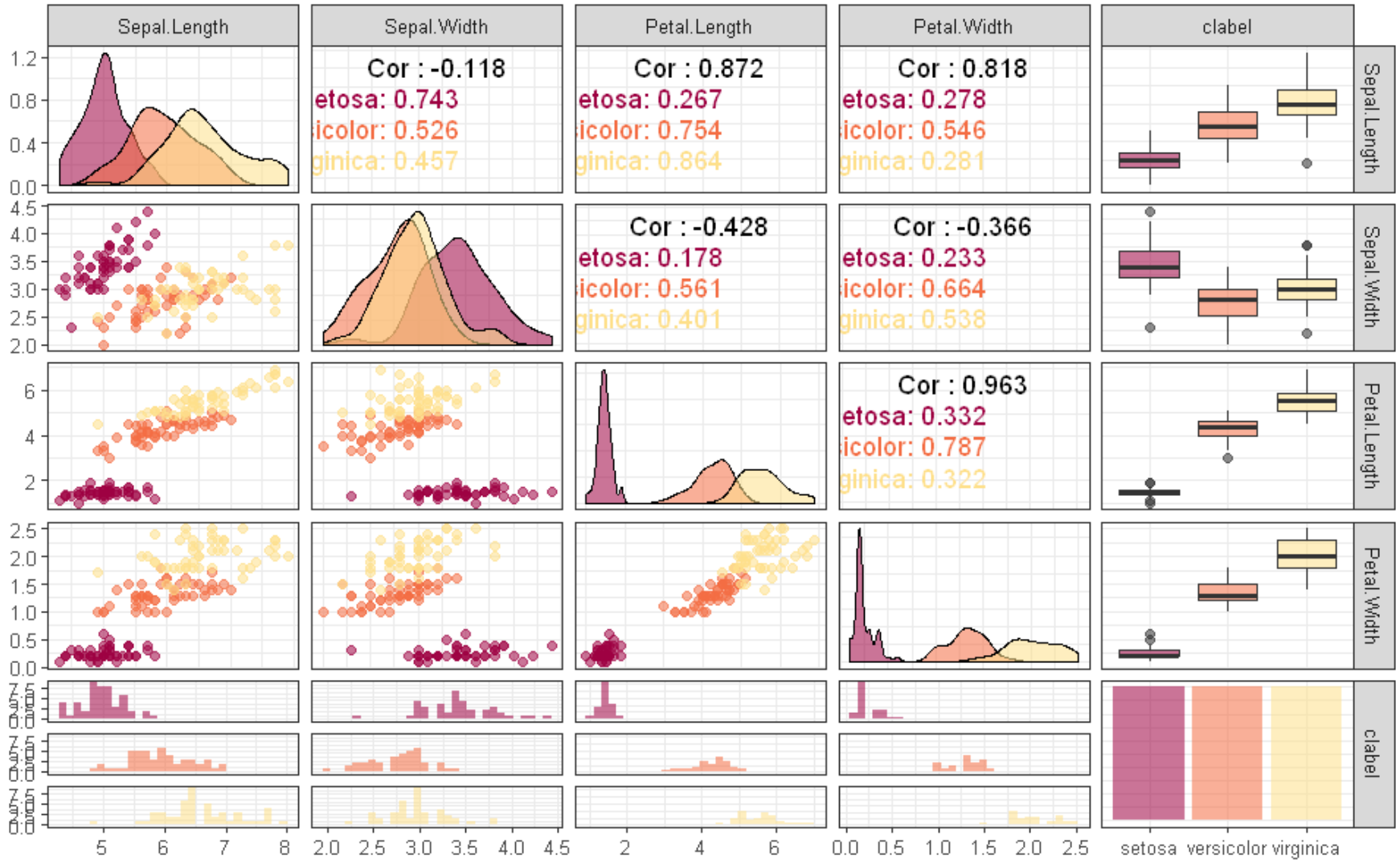
[1] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, 2008, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, IEEE Transactions on Visualization and Computer Graphics, v. 14, n. 6, p. 1141–1148.

Advanced Scatter Matrix plot



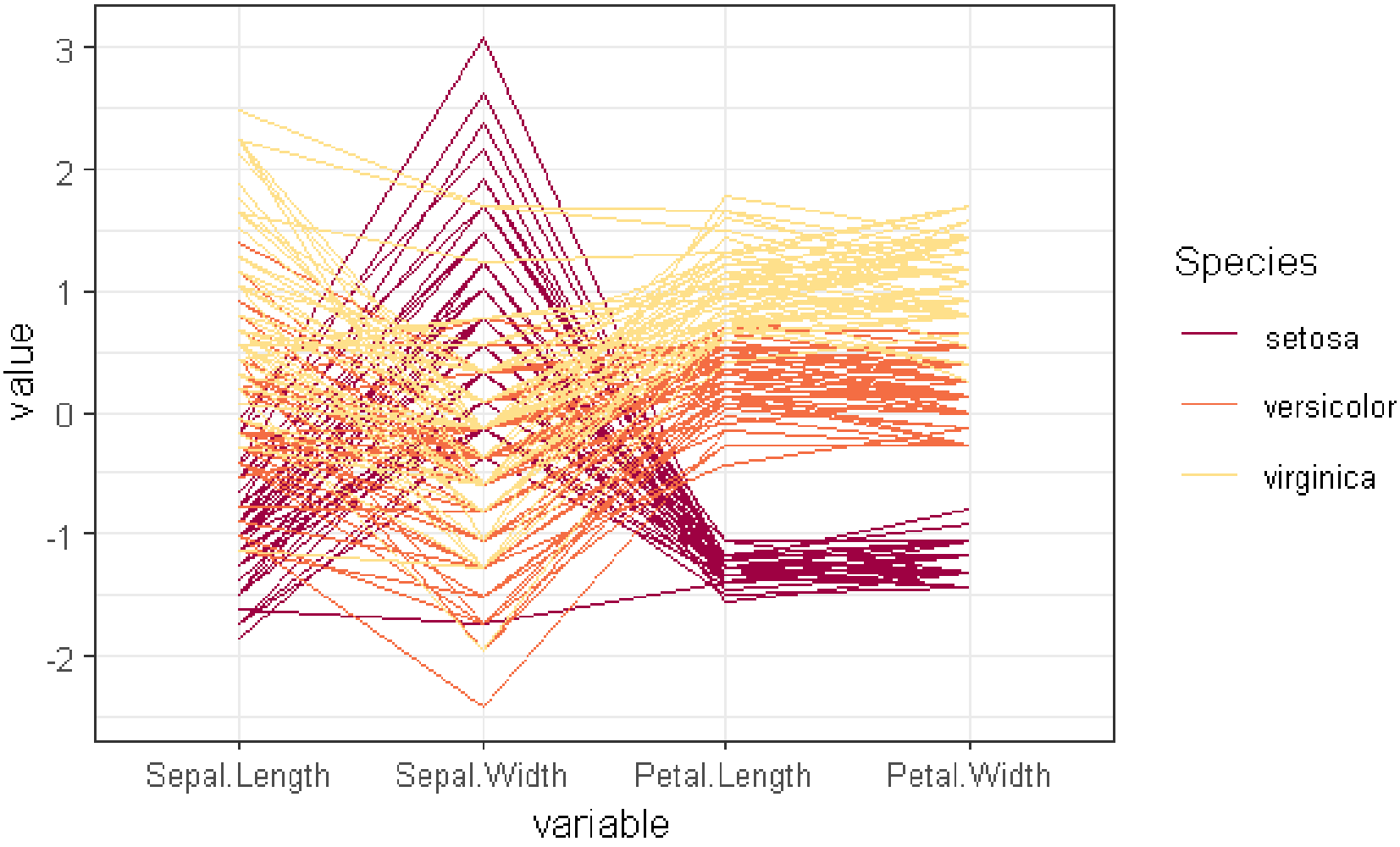
[1] D.A. Keim, M.C. Hao, U. Dayal, H. Janetzko, and P. Bak, 2010, Generalized scatter plots, Information Visualization, v. 9, n. 4, p. 301–311.

Advanced Scatter Matrix plot with a class label



[1] D.A. Keim, M.C. Hao, U. Dayal, H. Janetzko, and P. Bak, 2010, Generalized scatter plots, Information Visualization, v. 9, n. 4, p. 301–311.

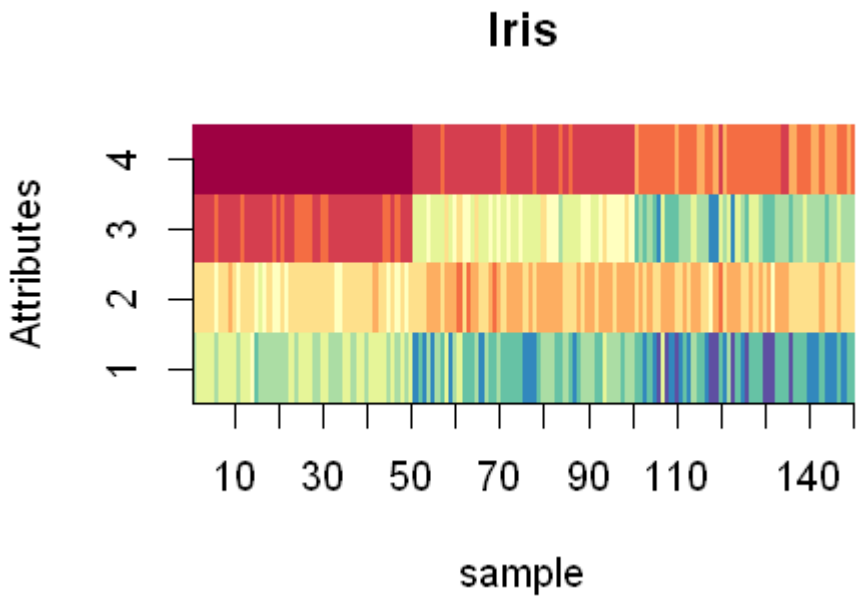
Parallel Coordinates of a Data Set



[1] A. Inselberg and B. Dimsdale, 1990, Parallel coordinates: A tool for visualizing multi-dimensional geometry, In: *IEEE Conference on Visualization - Visualization '90*, p. 361–378

Pixel-Oriented Visualization Techniques

- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



[1] D.A. Keim, 2000, Designing pixel-oriented visualization techniques: theory and applications, IEEE Transactions on Visualization and Computer Graphics, v. 6, n. 1, p. 59–78.

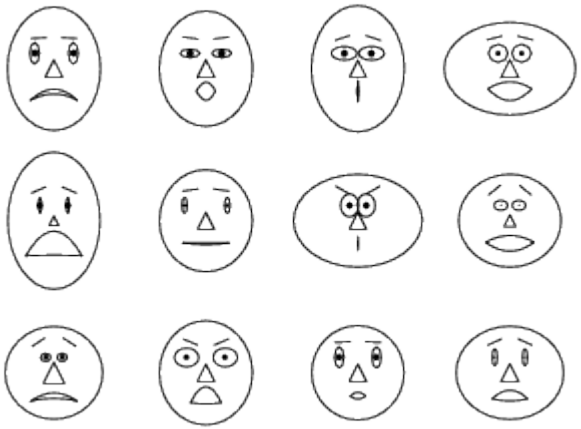
Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff faces
  - Salience
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

[1] J. Han, J. Pei, and M. Kamber, 2011, *Data Mining: Concepts and Techniques*. Elsevier.

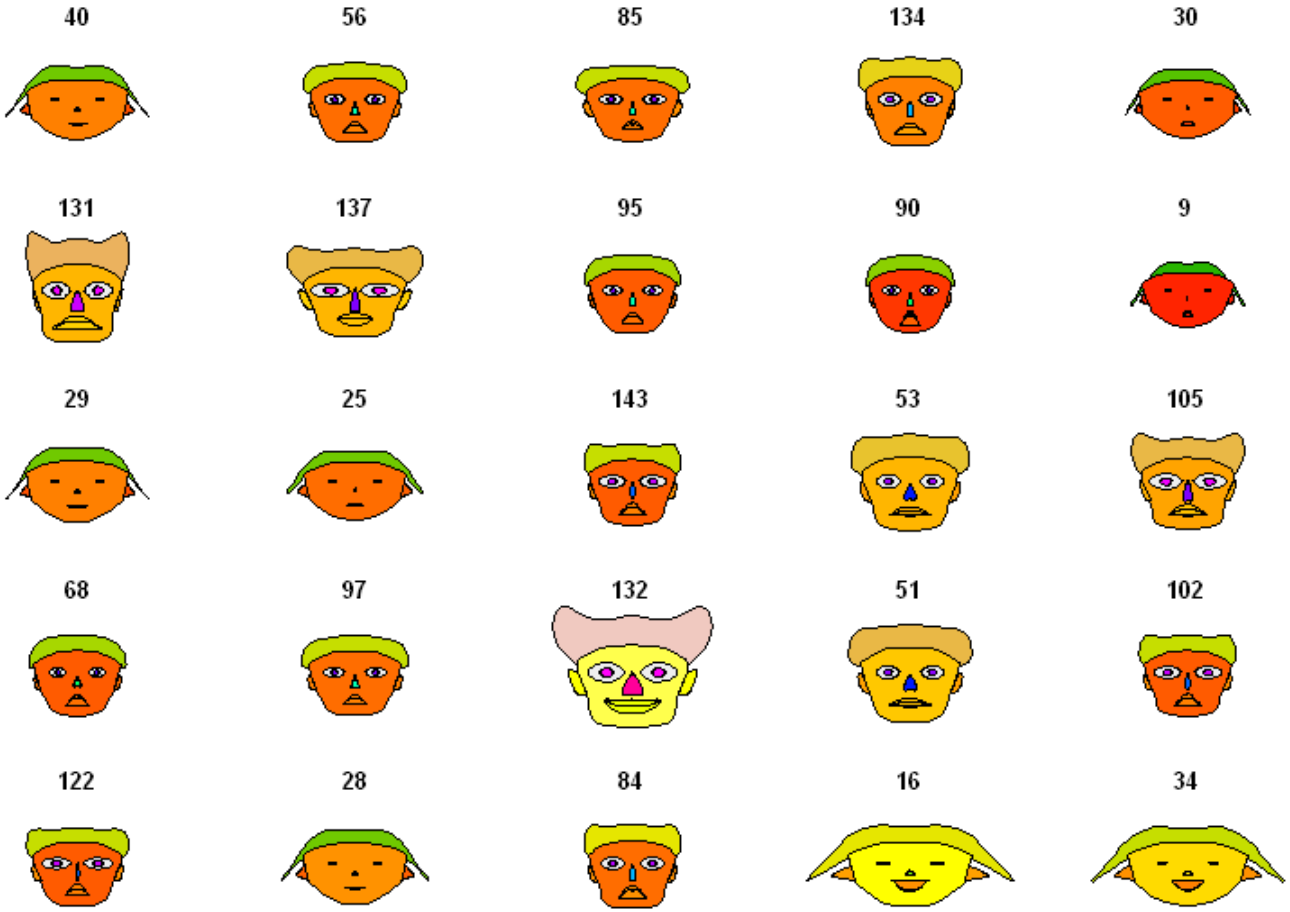
Chernoff Faces

- A way to display variables on a two-dimensional surface
  - Let x be eyebrow slant, y be eye size, z be nose length
- The figure shows faces produced using ten characteristics: head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening):
  - Each assigned one of 10 possible values

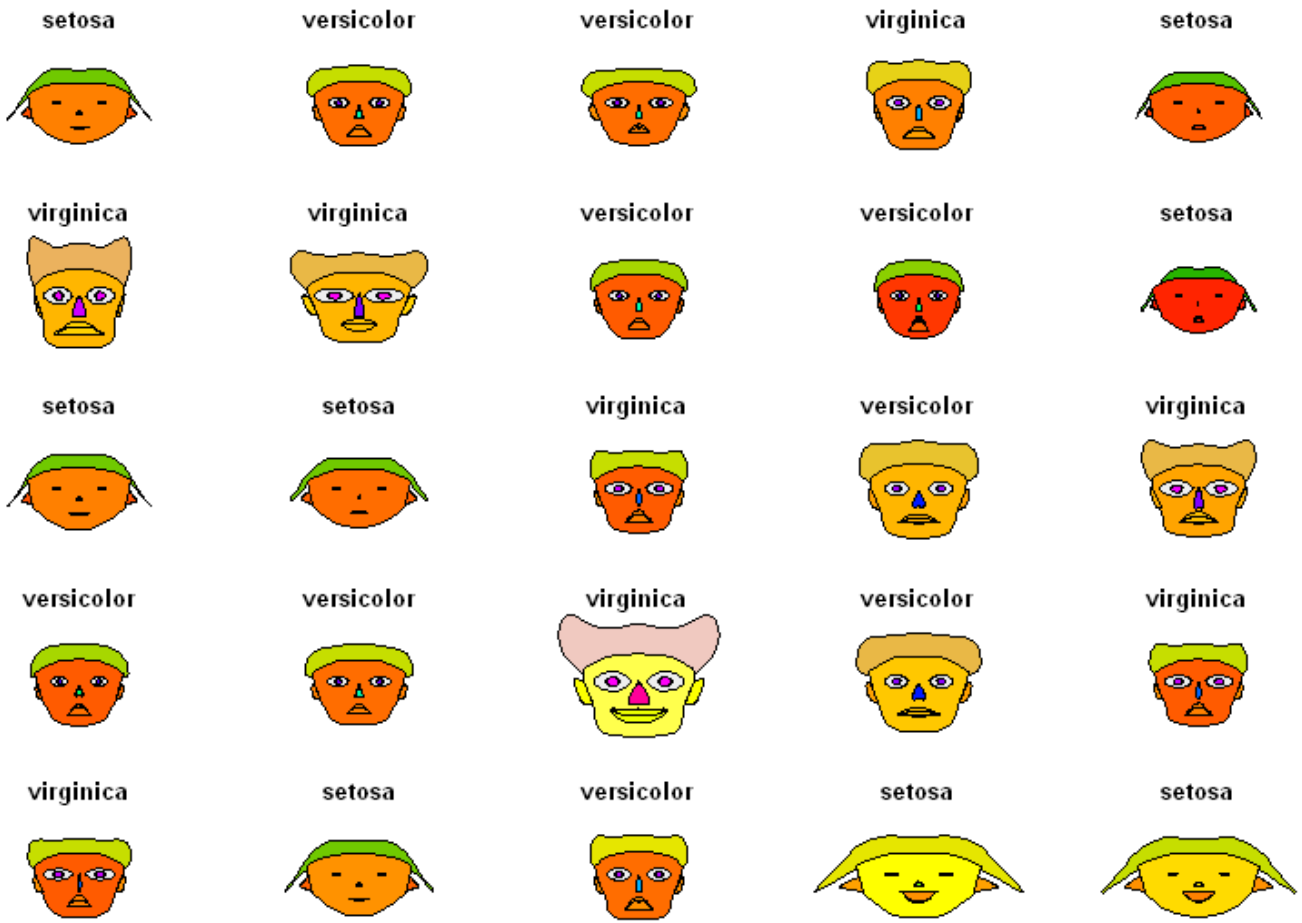


Gonick, L. and Smith, W. The Cartoon Guide to Statistics. New York: Harper Perennial, p. 212, 1993
Weisstein, Eric W. "Chernoff Face." From MathWorld -A Wolfram Web Resource. mathworld.wolfram.com/ChernoffFace.html

Chernoff Faces example with the Iris dataset
Can you see any pattern?



Chernoff Faces example with the Iris dataset (displaying class label)



Exercises

- Compute the correlation between two attributes
  - Try to explain it
- Compute and plot the correlation among all attributes
  - Hint use cor function and corrplot
  - Explain the plot
- Plot the Parallel coordinates targeting the class label
  - Explain it
- Use other methods for studying the wine dataset
  - Explain them

Take some time to practice the examples
https://nbviewer.jupyter.org/github/eogasawara/mylibrary/blob/master/ExploratoryAnalysis.ipynb

Main References