Multivariate Statistics

Chapter 1: Multivariate data

Pedro Galeano
Departamento de Estadística
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Course 2017/2018

Master in Mathematical Engineering

Introduction

- Multivariate data is a collection of data taken from several variables.
- Multivariate data analysis consists of a set of techniques for the simultaneous analysis of multivariate data.
- The main goals of the multivariate data analysis are:
  1. understand the structure in the data and summarise it in simpler ways;
  2. understand the relationship of one part of the data to another part; and
  3. make decisions and inferences based on the data.
- The early methods developed by statisticians were linear which are simple, elegant, and surprisingly powerful.
- For instance, Principal Component Analysis deals with the first topic in the preceding list, Canonical Correlation Analysis with the second, and Discriminant Analysis with the third.

Introduction

- As time moved on, more complex methods were developed.
- Nevertheless, linear methods have not lost their appeal.
- Indeed, as we have become more able to collect and handle very large and high-dimensional data, renewed requirements for linear methods have arisen.
- In high-dimensional data sets, the essential structure can often be obscured by noise, and it becomes vital to reduce the original data set in such a way that the interesting structure in the data is preserved while irrelevant features are removed.
- Principal Component Analysis and Factor Analysis have become indispensable dimension reduction tools and are often used as a first step in a more comprehensive analysis.

Introduction

- Traditionally, it is assumed that the number of variables (the dimension) is small compared with the number of elements in the sample (the sample size).
- Similarly, for the asymptotic theory, the sample size increases while the dimension remains constant.
- Many recent data sets do not fit into this framework; we encounter
  - data whose dimension is comparable to the sample size, and both are large;
  - high-dimension low sample size data whose dimension vastly exceeds the sample size; and
  - functional data whose observations are functions.

Introduction

- The Gaussian assumption will often not be useful for high-dimensional data.
- However, a deviation from normality does not affect the applicability of Principal Component Analysis or Canonical Correlation Analysis, for instance.
- Therefore, we need to take care when making inferences based on Gaussian assumptions or when we want to exploit the normal asymptotic theory.

Introduction

- In the rest of this chapter:
  - we present the general structure of a multivariate data set;
  - we review graphical techniques for visualizing a multivariate data set;
  - we introduce several multivariate descriptive measures; and
  - we briefly introduce linear transformations.

Multivariate data sets

- Suppose that we have observed a set of variables in a sample of elements from a certain population (in a wide sense).
- Traditionally, the variables are classified as:
  - quantitative, when their value is expressed numerically, such as the age of a person, their height or their income; or
  - qualitative, when their value can be attributed to a category such as gender, eye color or city of birth.

Multivariate data sets

- Quantitative variables can then be classified as:
  - continuous, when the real value can be read as an interval, such as height; or
  - discrete, when the values belonging to it are distinct and separate, such as the number of siblings.
- Qualitative variables can be classified as:
  - binaries, when there are only two possible values, such as gender (male, female); or
  - non-binaries, when many values are possible such as city of residence.

Multivariate data sets

- Usually, binary variables are coded numerically.
- For example, the gender variable converts to numerical by assigning 0 to a male and 1 to a female, or viceversa.
- It is important to note that even if a binary variable is coded, it is still a qualitative variable.

Multivariate data sets

- Non-binary variables can also be assigned a numerical value by converting them into binary variables.
- For example, consider the variable eye color (EC) and assume that the categories are blue (B), green (G), brown (Br) and black (Bl).
- Then, we can define 3 binary variables as:
  1. $x_1 = 1$, if EC=B, and $x_1 = 0$, otherwise.
  2. $x_2 = 1$, if EC=G, and $x_2 = 0$, otherwise.
  3. $x_3 = 1$, if EC=Br, and $x_3 = 0$, otherwise.

Multivariate data sets

- If the number of values of a qualitative variable is large, this procedure will lead to a great number of variables.
- Then, it is useful to group the categories in order to avoid having variables that will almost always have the same value (for instance, 0, if the category is infrequent, or 1, if it appears often).
- The variable EC could also be coded as follows, $x_1 = 1$, if EC=B, $x_1 = 2$, if EC=G, $x_1 = 3$, if EC=Br, and $x_1 = 4$, if EC=Bl.
- However, this system has the inconvenience of suggesting a gradation of values that may not exist.

Multivariate data sets

- We assume from here on that we have observed the values of p univariate random variables in a set of n elements of a population.
- We denote by $x_1, \ldots, x_p$ the p univariate random variables.
- The set of p variables forms a multivariate variable that is denoted by $x = (x_1, \ldots, x_p)'$.
- The values of the p univariate variables in each of the n elements of the population can be represented in a matrix, X, of dimensions $n \times p$, which we will call data matrix, given by:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Multivariate data sets

- Therefore, we will denote as x_{ij} the generic element of this matrix, which represents the value of the univariate random variable x_j over the individual i.
- Note that the values taken by the univariate random variable j over the n individuals are given by $x_{1j}, \ldots, x_{nj}$, for $j = 1, \ldots, p$, and can be summarized in the vector $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})'$.
- On the other hand, the values taken by the individual i for the j univariate variables are given by $x_{i1}, \ldots, x_{ip}$, for $i = 1, \ldots, n$, and can be summarized in the vector $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$.

Illustrative example (I)

- Eight univariate variables measured on the 50 states of the USA:
  - x_1: population estimate as of July 1, 1975 (in thousands).
  - x_2: per capita income (1974) (in dollars).
  - x_3: illiteracy (1970, percent of population).
  - x_4: life expectancy in years (1969 – 71).
  - x_5: murder and non-negligent manslaughter rate per 100000 population (1976).
  - x_6: percent high-school graduates (1970).
  - x_7: mean number of days with minimum temperature below freezing (1931 – 1960) in capital or large city.
  - x_8: land area in square miles.
- The data set is summarized in the following table:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Wyoming	376	4566	0.6	70.29	6.9	62.9	173	97203

Illustrative example (II)

- Five univariate variables measured on 150 flowers (50 flowers of each specie):
  - x_1: Length of the sepal (in mm.).
  - x_2: Width of the sepal (in mm.).
  - x_3: Length of the petal (in mm.).
  - x_4: Width of the petal (in mm.).
  - x_5: Specie (setosa, versicolor or virginica).
- The dataset is summarized in the following table:

	x_1	x_2	x_3	x_4	x_5
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
⋮	⋮	⋮	⋮	⋮	
150	5.9	3.0	5.1	1.8	virginica

Visualizing multivariate data sets

- Before analyzing a multivariate data set, it is important to try to visualize it.
- Often we get useful features including:
  - skewness;
  - multimodality;
  - outliers; or
  - distint groupins.

Visualizing multivariate data sets

- Graphical displays are exploratory data-analysis tools which can help to understand the data.
- Note that the insight obtained from graphical displays is more subjective than quantitative.
- However, visual cues are easier to understand and interpret than numbers alone.
- Indeed, the knowledge gained from graphical displays can complement more quantitative answers.

Visualizing multivariate data sets

- One difficulty of descriptive methods for high dimensional data is the human perceptional system.
- Point clouds in two dimensions are easy to understand and to interpret.
- We also have the possibility to see real time 3D rotations and thus to perceive also three-dimensional data.
- A qualitative jump in presentation difficulties occurs for dimensions greater than or equal to 4.
- Next, we investigate the basic descriptive and graphical techniques allowing simple exploratory data analysis.
- Note that we are going to focus in quantitative variables, while qualitative (binary) variables will be treated as additional information when making the plots.

Visualizing multivariate data sets

- We begin the exploration of a multivariate data set using the boxplot which is a simple univariate device that detects outliers variable by variable and that can compare distributions of the data among different groups.
- Two basic techniques for estimating densities are also presented: histograms and kernel densities, as they give a quick insight into the shape of the distribution of the data.
- Scatterplots are shown to be very useful for plotting bivariate or trivariate variables against each other: they help to understand the nature of the relationship among variables in a data set and allow for the detection of groups or clusters of points.
- Scatterplot matrices allow the visualization of several bivariate scatterplots on the same display.
- Parallel coordinate plots are useful to detect outliers and/or groups.

Visualizing multivariate data sets

- The boxplot is a graphical technique that display the distribution of variables.
- It help us see the location, spread, skewness, tail length and outliers.
- Let $x_{1j}, \ldots, x_{nj}$ be the n observations of the random variable x_j, for $j = 1, \ldots, p$.
- Then, the boxplot is a graphical representation of the sequence $x_{1j}, \ldots, x_{nj}$ constructed with several statistics taken from it.

Visualizing multivariate data sets

- The sample order statistics of $x_{1j}, \ldots, x_{nj}$, denoted by $x_{(1)j}, \ldots, x_{(n)j}$, are the set of sorted observations (in increasing order), where:

$$x_{(1)j} = \min \{x_{1j}, \ldots, x_{nj}\}$$

and

$$x_{(n)j} = \max \{x_{1j}, \ldots, x_{nj}\}$$

- The sample median of $x_{1j}, \ldots, x_{nj}$, denoted by M_j, typically cuts the set of observations in two equal parts, and is defined as

$$M_j = \begin{cases} x_{(\frac{n+1}{2})j} & n \text{ odd} \\ \frac{1}{2}\left\{x_{(\frac{n}{2})j} + x_{(\frac{n}{2}+1)j}\right\} & n \text{ even} \end{cases}$$

Visualizing multivariate data sets

- The sample quartiles of $x_{1j}, \ldots, x_{nj}$, denoted by Q_{Lj} and Q_{Uj}, respectively, typically cut the set into four equal parts, and are defined as

$$Q_{Lj} = x_{\left[\frac{1}{4}(n+1)\right]j}$$

and

$$Q_{Uj} = x_{\left[\frac{3}{4}(n+1)\right]j}$$

respectively, where $[\cdot]$ denote the integer part.

- Note that alternative definitions of the sample quartiles have been proposed.

Visualizing multivariate data sets

- The sample interquartile range of $x_{1j}, \ldots, x_{nj}$, denoted by IQR_j, and defined as

$$IQR_j = Q_{Uj} - Q_{Lj}$$

  is a measure of the spread of $x_{1j}, \ldots, x_{nj}$.

- The sample outside bars of $x_{1j}, \ldots, x_{nj}$, denoted by O_{Lj} and O_{Uj}, respectively, and defined as

$$\begin{aligned} O_{Lj} &= Q_{Lj} - 1.5IQR_j \\ O_{Uj} &= Q_{Uj} + 1.5IQR_j \end{aligned}$$

  are the borders beyond which a point is regarded as an outlier.

- The number 1.5 used to construct the sample outside bars is selected based on Gaussian arguments.

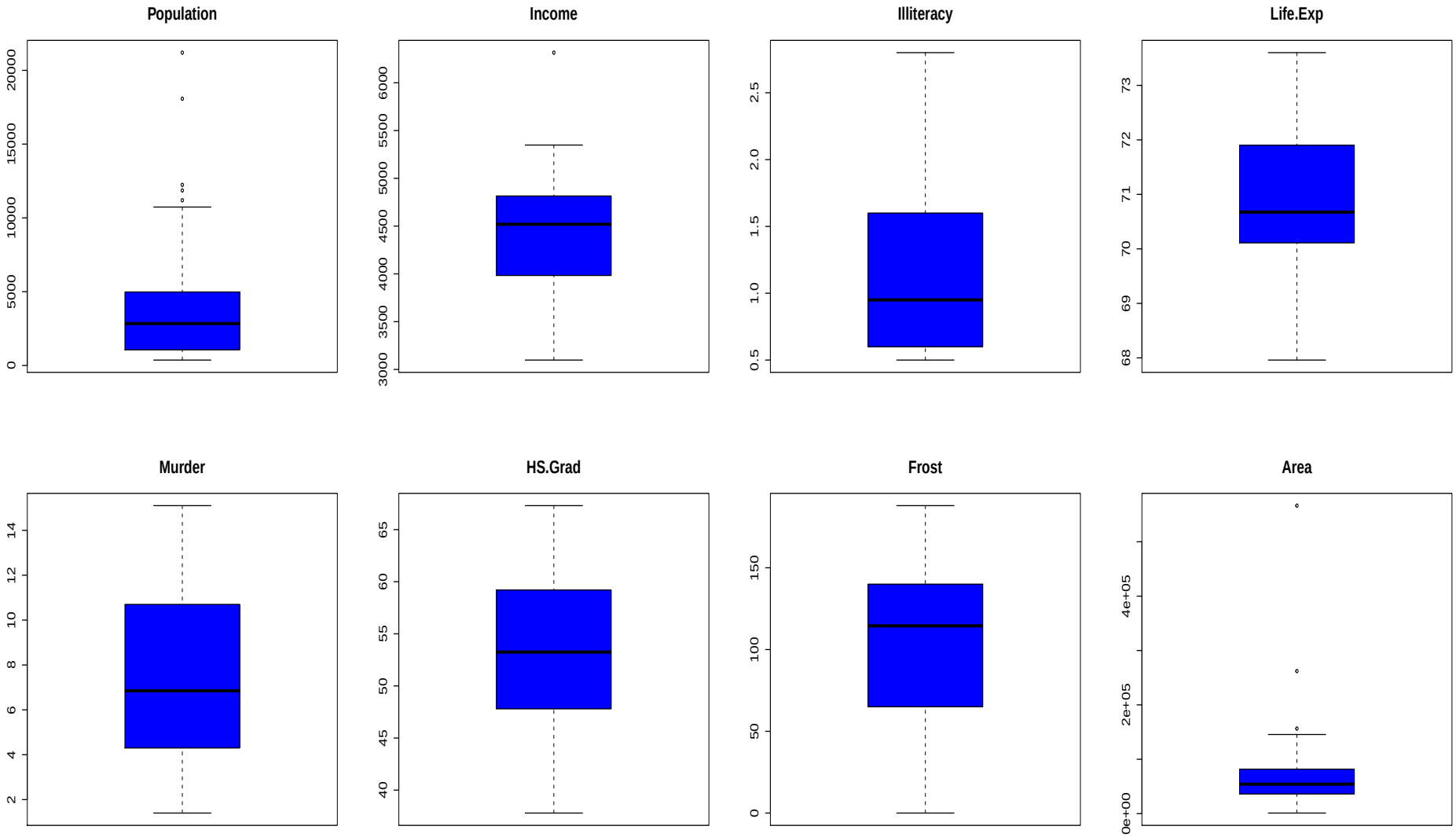
Visualizing multivariate data sets

- The boxplot is constructed in the following way:
  1. Draw a box with borders at Q_{Lj} and Q_{Uj} (i.e., 50% of the data are in this box).
  2. Draw the sample median as a solid line.
  3. Draw "whiskers" from each end of the box to the most remote point that is not an outlier.
  4. Show outliers with special characters.

Illustrative example (I)

- The next slide shows the boxplots of the eight variables of the US states data set.
- The variable "percent high-school graduates" is a quite symmetric variable while the other are clearly skewed.
- In particular, "population" and "land area" are quite skewed and containing some outliers.
- Therefore, the variables have different shapes.

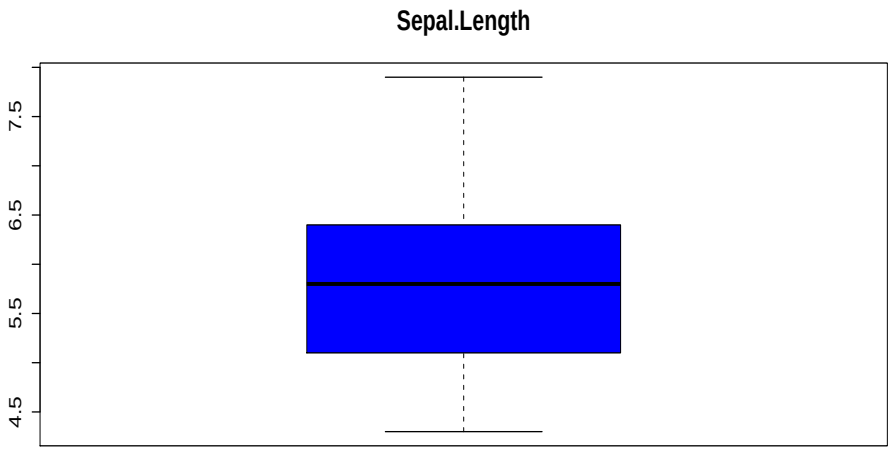
Illustrative example (I)

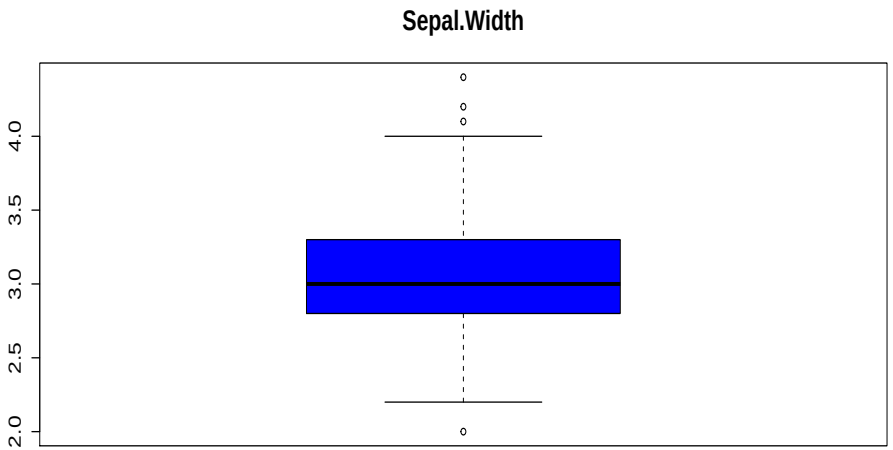


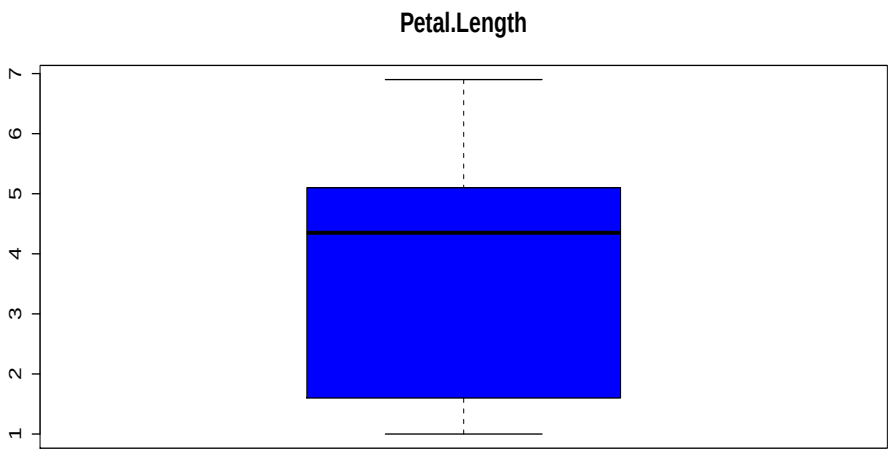
Illustrative example (II)

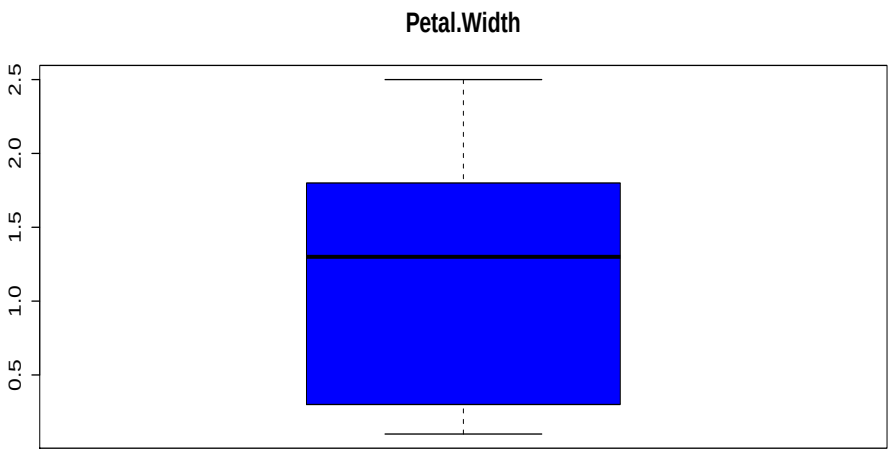
- The next two slides show boxplots of the four quantitative variables of the iris data set.
- In the first figure, it can be seen that the variables measuring the length and width of the sepal are more symmetric than the variables measuring the length and width of the petal.
- In the second figure, it can be seen that the boxplots are a useful tool to compare the values of a variable divided in different groups, as it is the case here.
- Note the different values that attains the same value when split based on the variable "specie".

Illustrative example (II)

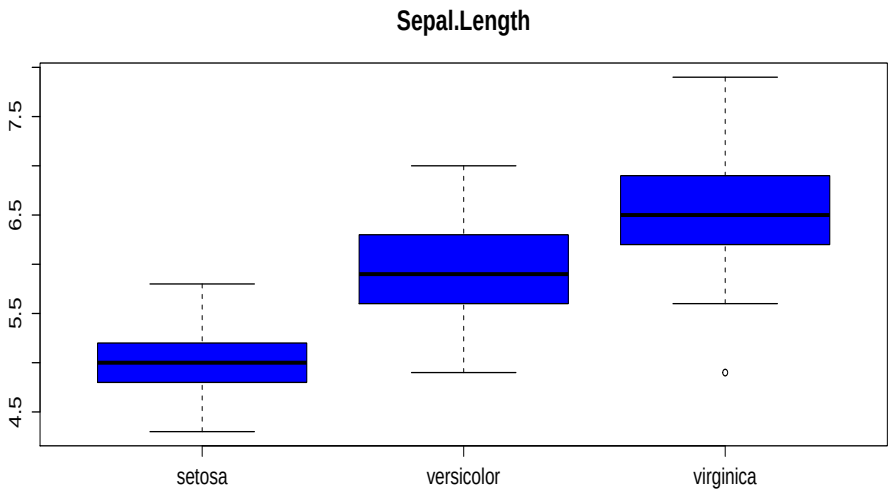


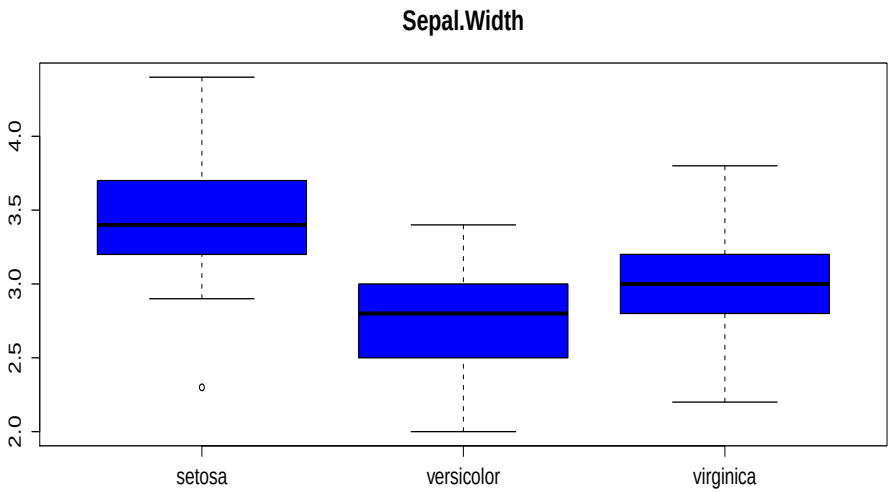


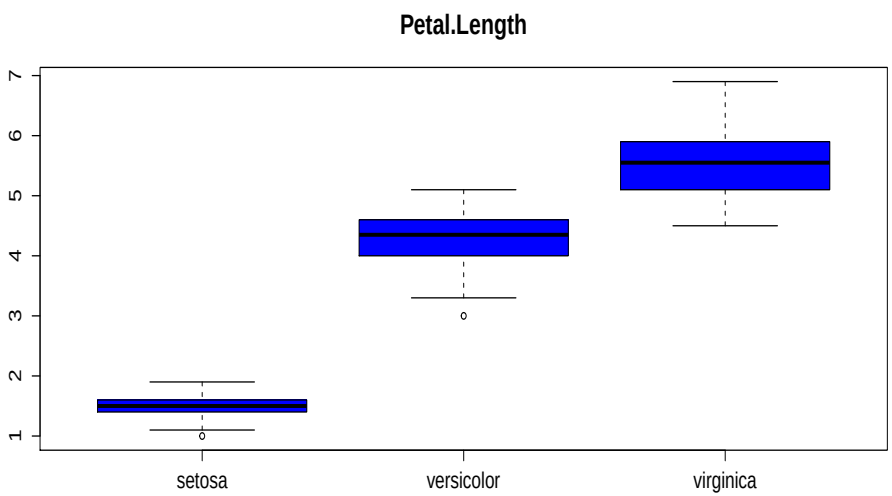


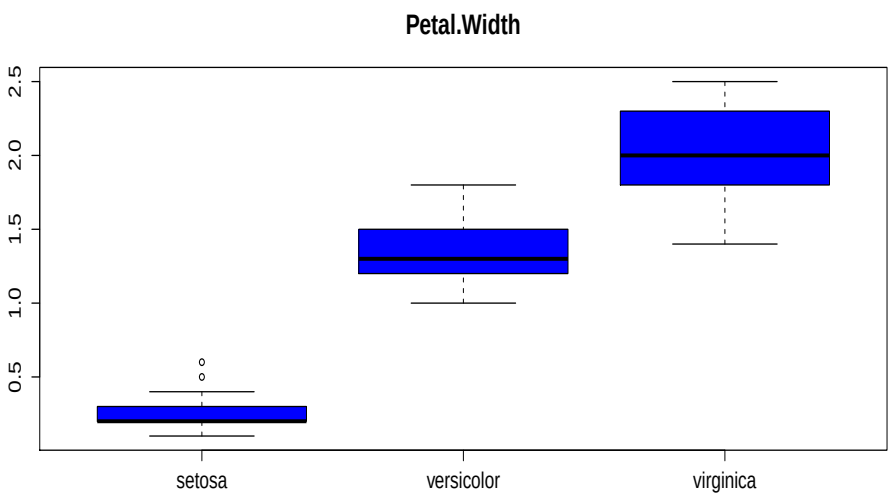


Illustrative example (II)









Visualizing multivariate data sets

- Histograms are density estimates.
- In other words, histograms gives an estimation of the distribution of the data.
- In contrast to boxplots, density estimates show possible multimodality of the data.
- The idea is to locally represent the data density by counting the number of observations in a sequence of consecutive bins.

Visualizing multivariate data sets

- Let $B_r(x_{0j}, h_j)$ denote the bin of length h_j starting at point x_{0j} and given by:

$$B_r(x_{0j}, h_j) = [x_{0j} + (r-1)h_j, x_{0j} + rh)$$

where $j = 1 \ldots, p$ and $r \in \mathbb{Z}$.

- Then, if $x_{1j}, \ldots, x_{nj}$ are observations of the variable x_j with density f_j, the histogram is defined as follows:

$$\widehat{f}_{h_j j}(x) = \frac{1}{nh_j} \sum_{r \in \mathbb{Z}} \sum_{i=1}^{n} \mathbf{I}\{x_{ij} \in B_r(x_{0j}, h_j)\} \mathbf{I}\{x \in B_r(x_{0j}, h_j)\}$$

where $\mathbf{I}\{\cdot\}$ is an indicator function such that:

$$\mathbf{I}\{x_{ij} \in B_r(x_{0j}, h_j)\} = \begin{cases} 1 & \text{if } x_{ij} \in B_r(x_{0j}, h_j) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbf{I}\{x \in B_r(x_{0j}, h_j)\} = \begin{cases} 1 & \text{if } x \in B_r(x_{0j}, h_j) \\ 0 & \text{otherwise} \end{cases}$$

Visualizing multivariate data sets

- Therefore:
  - $\mathbf{I}\{x_{ij} \in B_r(x_{0j}, h_j)\}$ counts the number of observations falling into bin $B_r(x_{0j}, h_j)$; and
  - $\mathbf{I}\{x \in B_r(x_{0j}, h_j)\}$ counts the number of observations around x.
- The parameter h_j is a smoothing parameter that controls the width of the histogram bins:
  - An h_j too large leads to very big blocks, leading to an unstructured histogram.
  - An h_j too small gives a very variable estimate with many unimportant peaks.

Visualizing multivariate data sets

- There are several methods available to select an "optimal" binwidth h_j, or equivalently, the optimal number of bins of the histogram.
- Let k_j be the number of bins, where the first bin starts at $x_{(1)j}$ and the last bin with points ends at $x_{(n)j}$.
- Then, the relationship between k_j and h_j is given by:

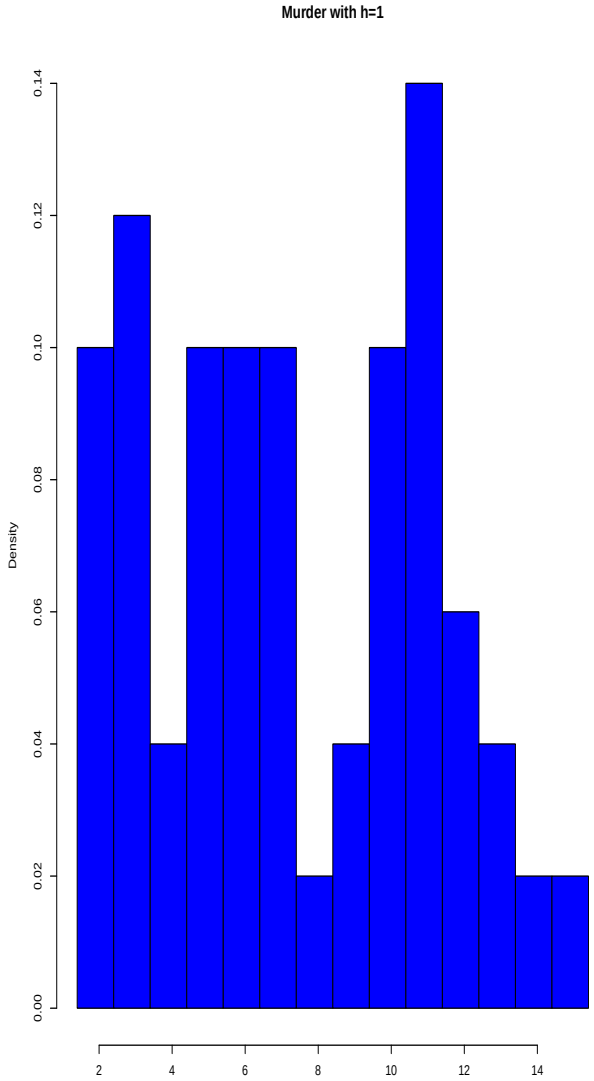
$$k_j = \frac{x_{(n)j} - x_{(1)j}}{h_j}$$

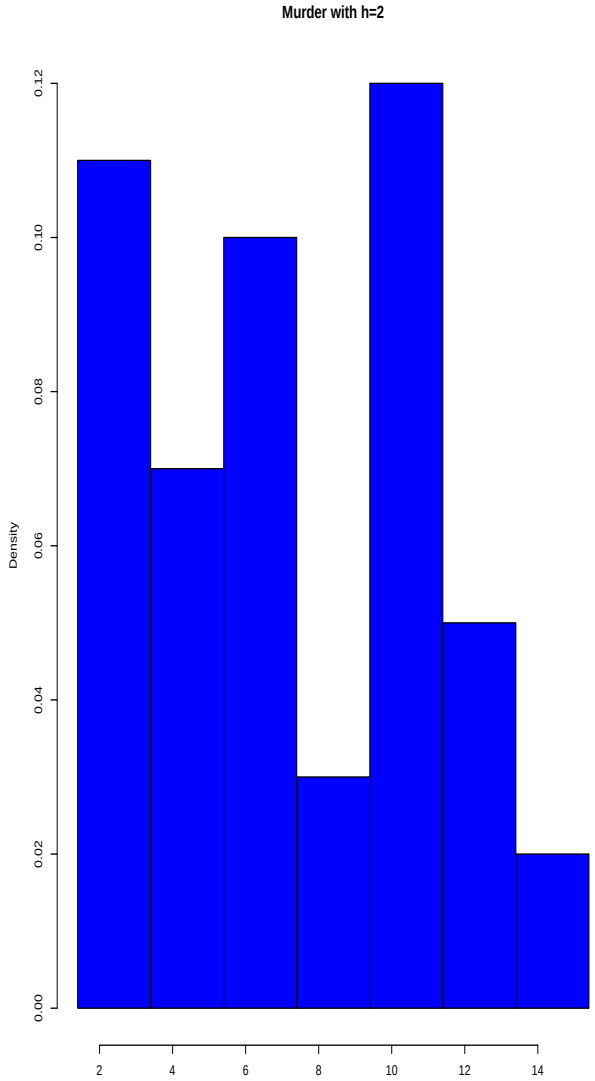
- We can select k_j (then h_j) using some of the following methods:
  - Sturges method: $k_j = [\log_2(n) + 1]$.
  - Scott method: $k_j = \frac{3.5\widehat{\sigma}_j}{n^{1/3}}$, where $\widehat{\sigma}_j$ is the sample standard deviation of the sample.
  - Freedman-Diaconis method: $k_j = 2\frac{IQR_j}{n^{1/3}}$, where IQR_j is the interquartile range of the sample.

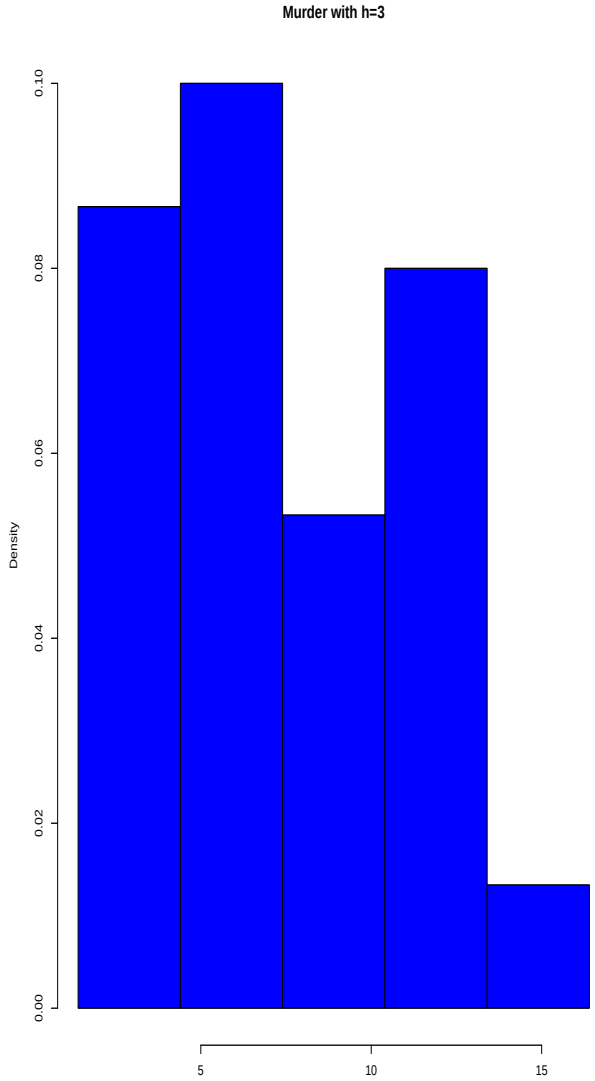
Illustrative example (I)

- The following two slides show histograms of the variables of the US states data set.
- In the first figure, it can be seen histograms of the variable "murder" taking $h = 1$, 2, and 3, respectively.
- As it can be seen, the shapes of the histograms are different.
- In the second figure, it can be seen the histograms for the eight variables with binwidth selected with the Sturges method.
- Note that the histograms show the presence of multimodality in some of the variables.
- However, the histograms are not smooth as are expected to be the distribution of the variables.

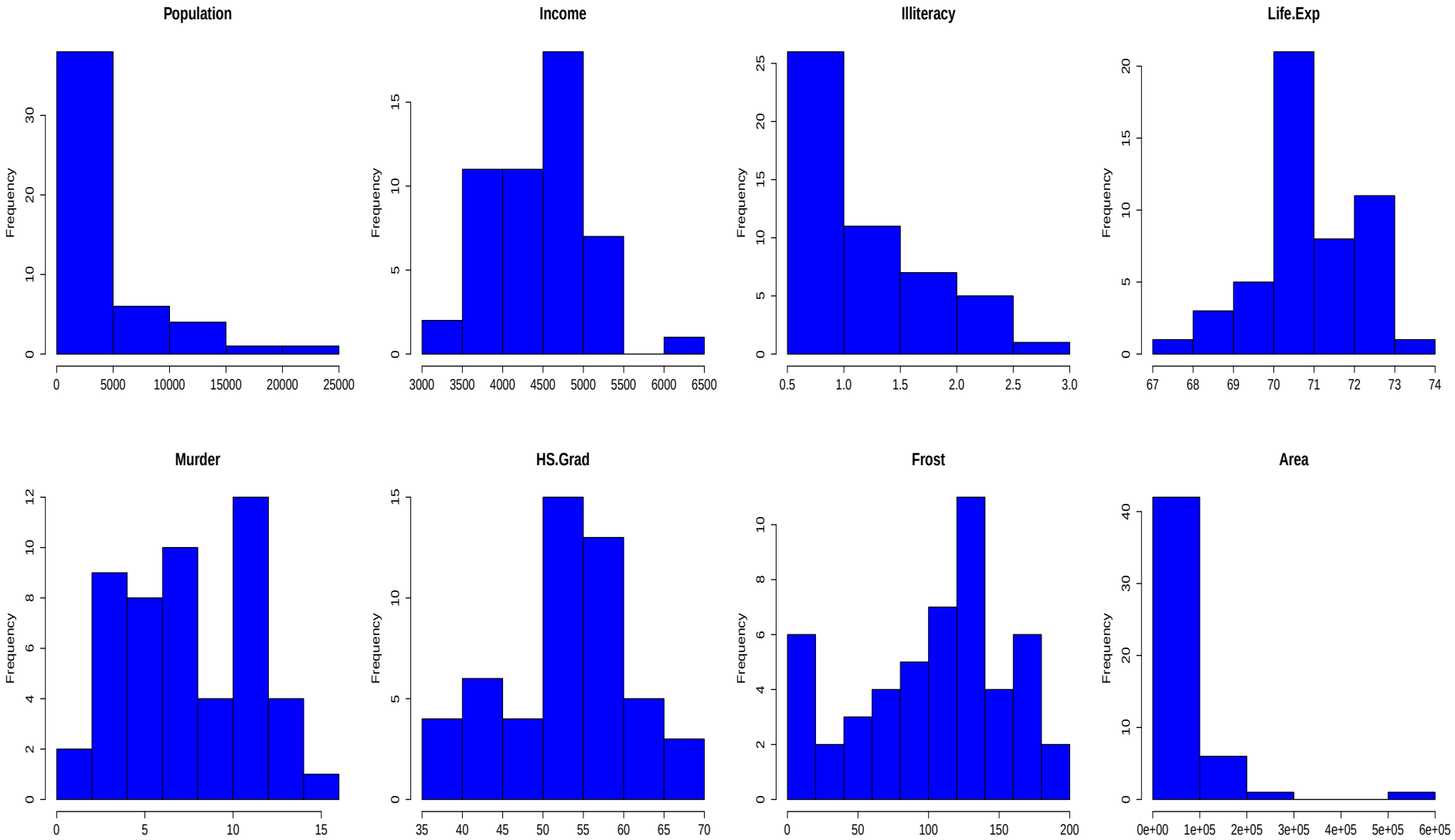
Illustrative example (I)







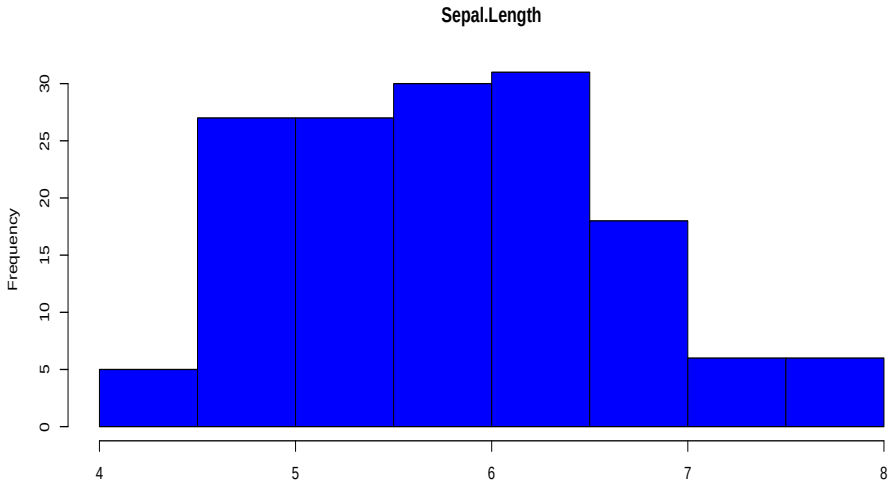
Illustrative example (I)

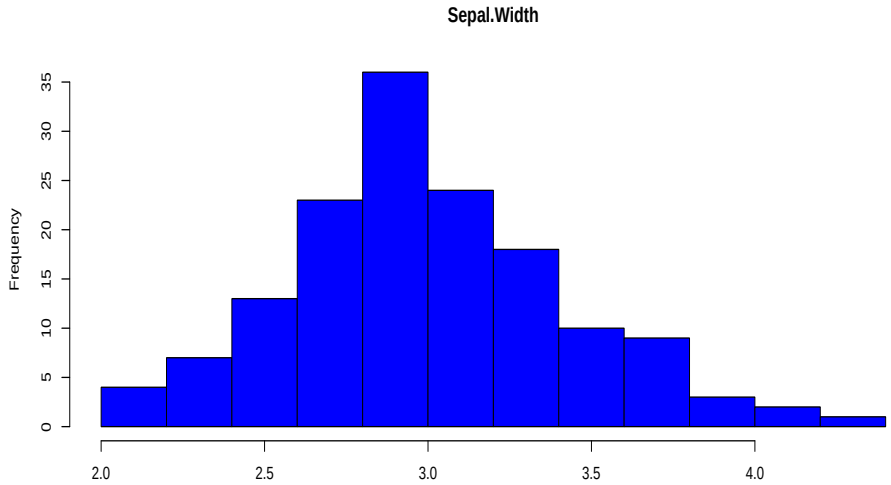


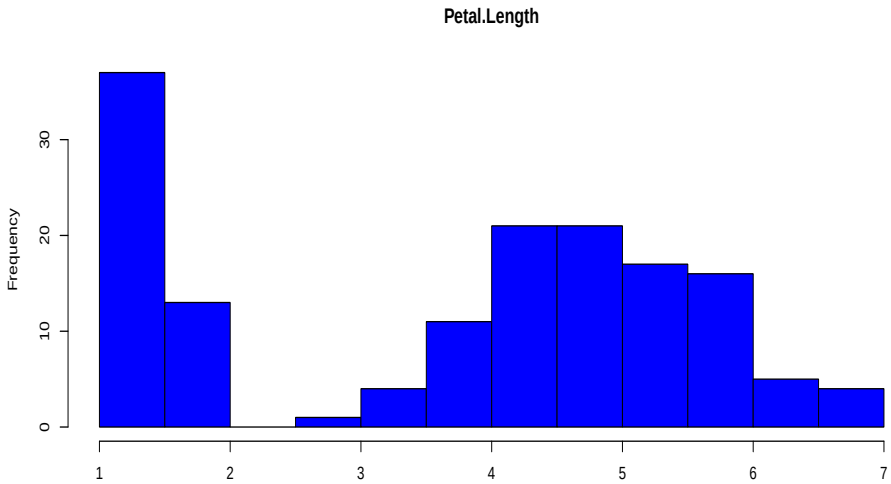
Illustrative example (II)

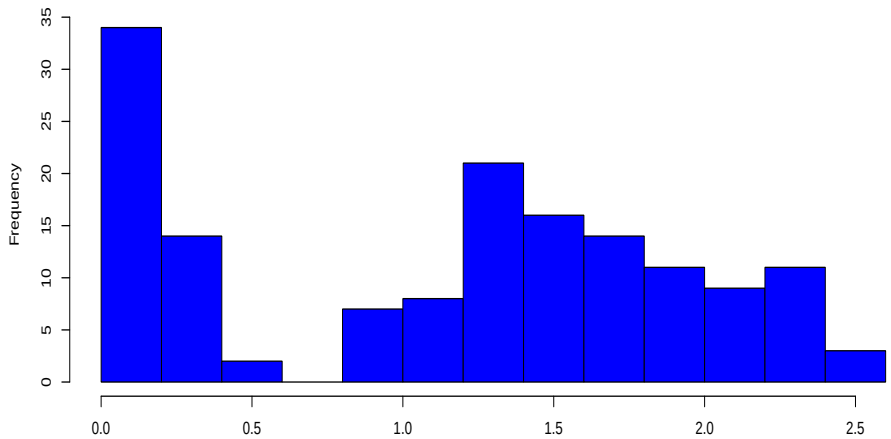
- The following two slides show histograms of the variables of the Iris data set.
- In the first figure, it can be seen histograms of the four quantitative variables with binwidth selected with the Sturges method.
- Note that the histograms show the presence of multimodality in the last two variables.
- In the second figure, it can be seen histograms of the third variable with binwidth selected with the Sturges method for the three species.
- Note that here the histograms are unimodal.
- In any case, the histograms are not smooth as are expected to be the distribution of the variables.

Illustrative example (II)

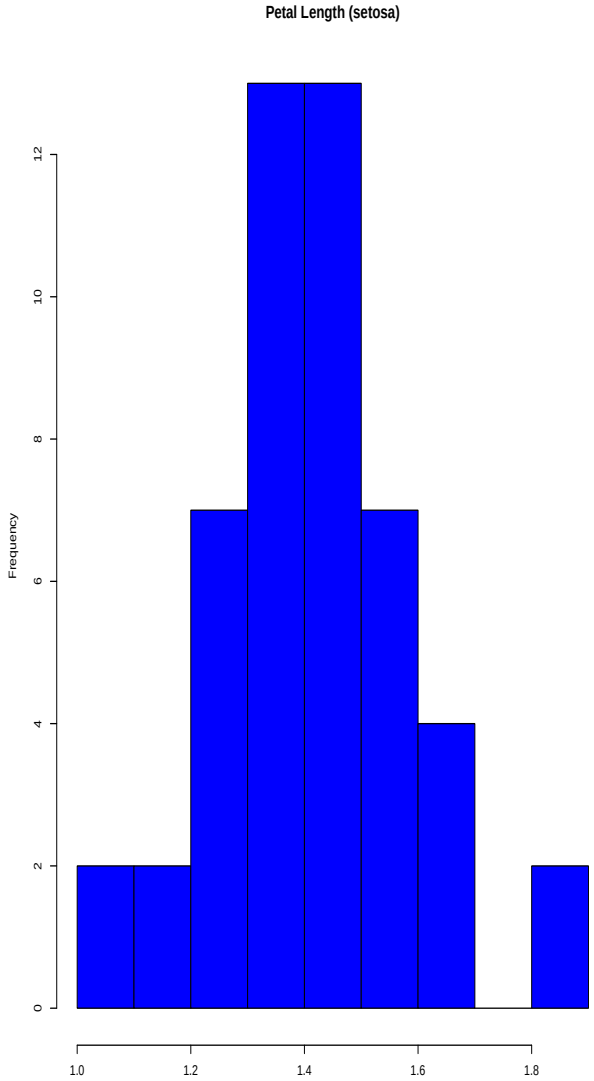


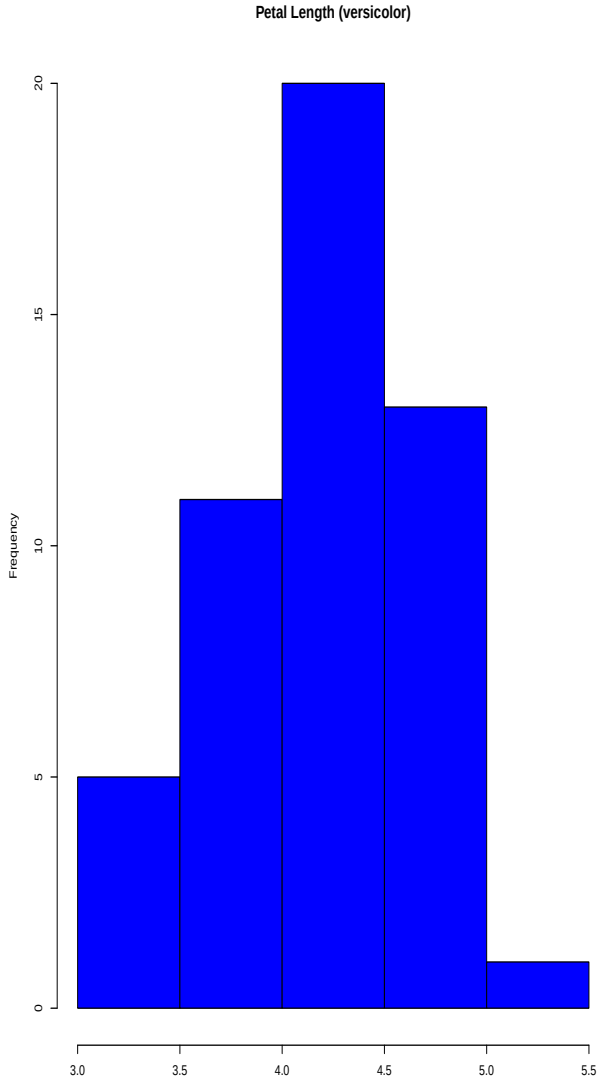


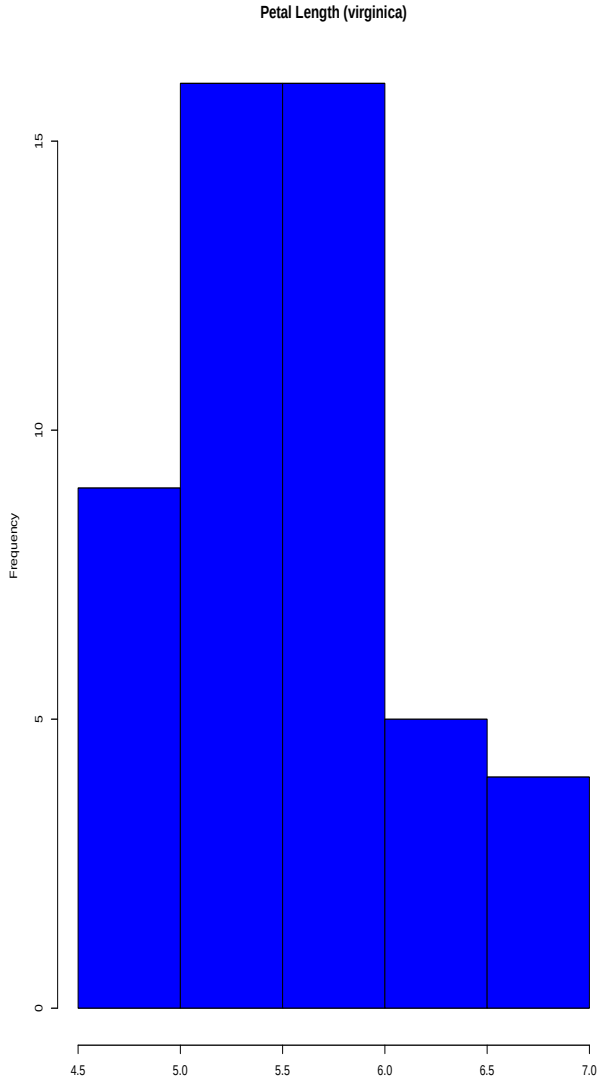




Illustrative example (II)







Visualizing multivariate data sets

- The major difficulties of histogram estimation may be summarised in four critiques:
  - determination of the binwidth.
  - choice of the bin origin.
  - loss of information since close observations are summarised with the same quantity.
  - lack of smoothness of the histogram.

Visualizing multivariate data sets

- An approach that avoids the last three difficulties is the kernel density.
- First, a smooth kernel function rather than a box is used as a basic building block.
- Second, the smooth function is centred directly over each observation.
- The general form of a kernel estimator of the density of x_j based on the sample $x_{1j}, \ldots, x_{nj}$ is given by:

$$\widehat{f}_{h_j}(x) = \frac{1}{nh_j} \sum_{i=1}^{n} K\left(\frac{x - x_{ij}}{h_j}\right)$$

where $K(\cdot)$ is a kernel function.

Visualizing multivariate data sets

- Some commonly used kernels are:
  - Uniform kernel: $K(u) = \frac{1}{2} \mathbf{I}\{|u| \leq 1\}$.
  - Triangle kernel: $K(u) = (1 - |u|) \mathbf{I}\{|u| \leq 1\}$.
  - Epanechnikov kernel: $K(u) = \frac{3}{4}\left(1 - u^2\right) \mathbf{I}\{|u| \leq 1\}$.
  - Quartic (Biweight) kernel: $K(u) = \frac{15}{16}\left(1 - u^2\right)^2 \mathbf{I}\{|u| \leq 1\}$.
  - Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$
- Different kernels generate different shapes of the estimated density.

Visualizing multivariate data sets

- The parameter h_j is called the bandwidth and can be selected by minimizing the mean integrated squared error (MISE):

$$MISE\left(h_j\right) = E\left[\int \left(\widehat{f}_{h_j}\left(x\right) - f_j\left(x\right)\right)^2 dx\right],$$

where $f_j\left(\cdot\right)$ is the density function of the univariate variable x_j.

- A lot of asymptotic theory has been done to obtain an approximation of the MISE leading to a rule useful in practice.

- If the Gaussian kernel is used, minimizing the MISE is approximately equivalent to choose the bandwith given by:

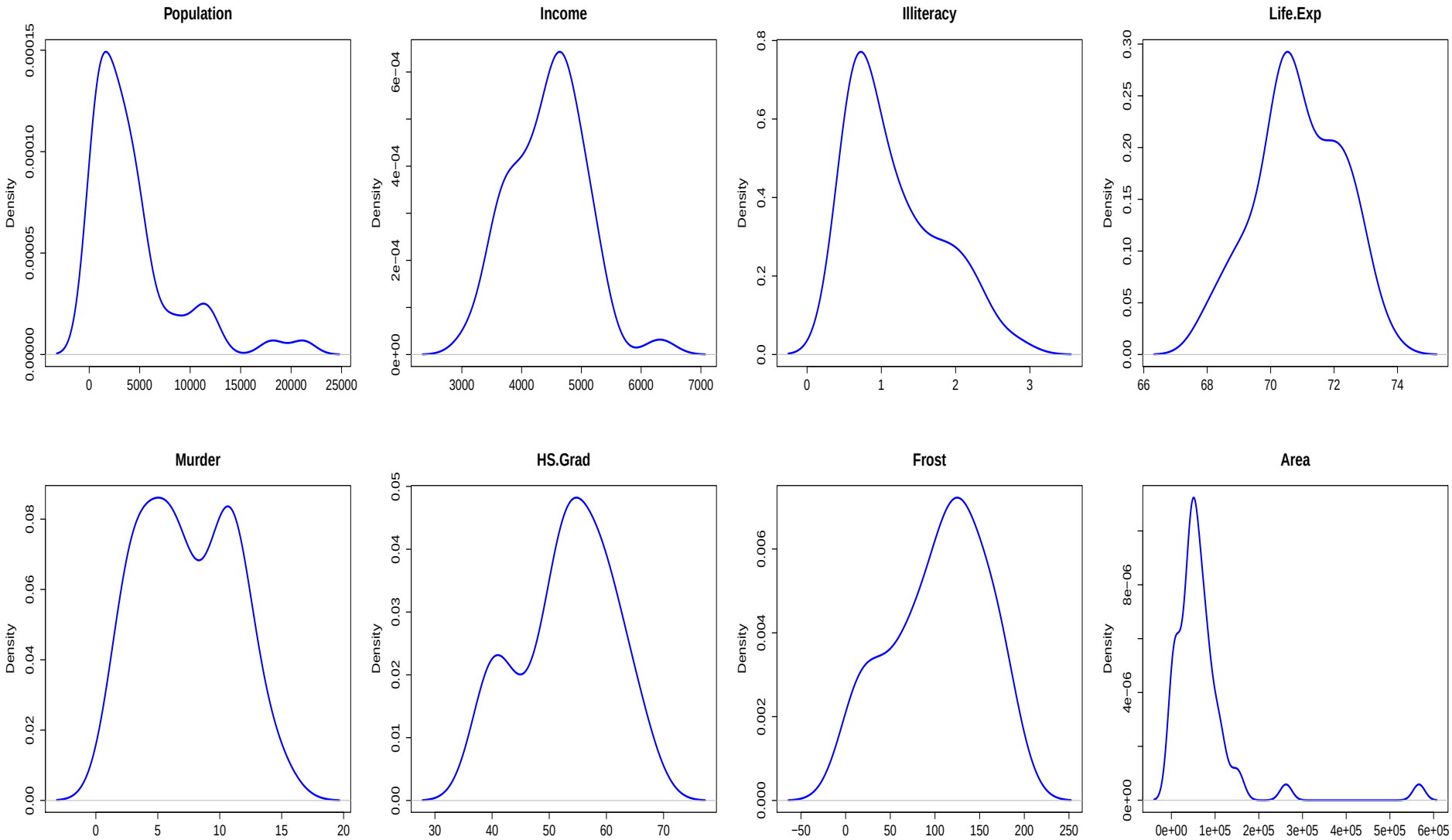
$$h_j = \left(\frac{4s_j^5}{3n}\right)^{1/5}$$

where s_j is the sample standard deviation of $x_{1j}, \ldots, x_{nj}$.

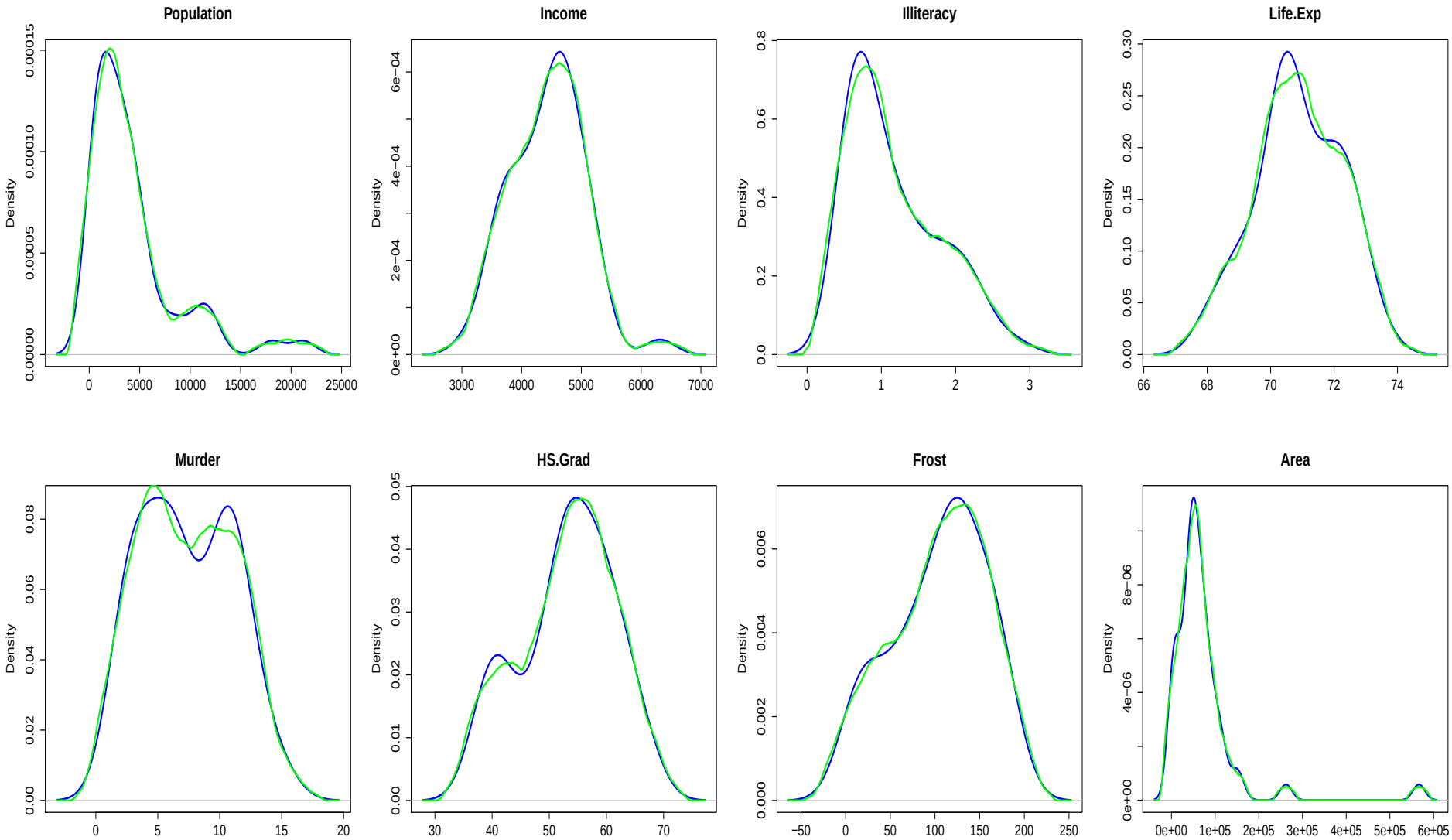
Visualizing multivariate data sets

- The following two slides show kernel densities of the variables of the US states data set.
- In the first figure, it can be seen kernel densities for the eight variables with the Gaussian kernel and bandwidth selected by minimising the integrated squared error.
- In the second figure, it can be seen a comparison of kernel densities for the eight variables with the Gaussian (in blue) and the Epanechnikov (in green) kernels and bandwidth selected by minimising the integrated squared error.
- As it can be seen both kernel densities are very close showing the presence of multimodality in some of the variables.
- The kernel estimates are very smooth, although they can be influenced by isolated observations.

Illustrative example (I)



Illustrative example (I)



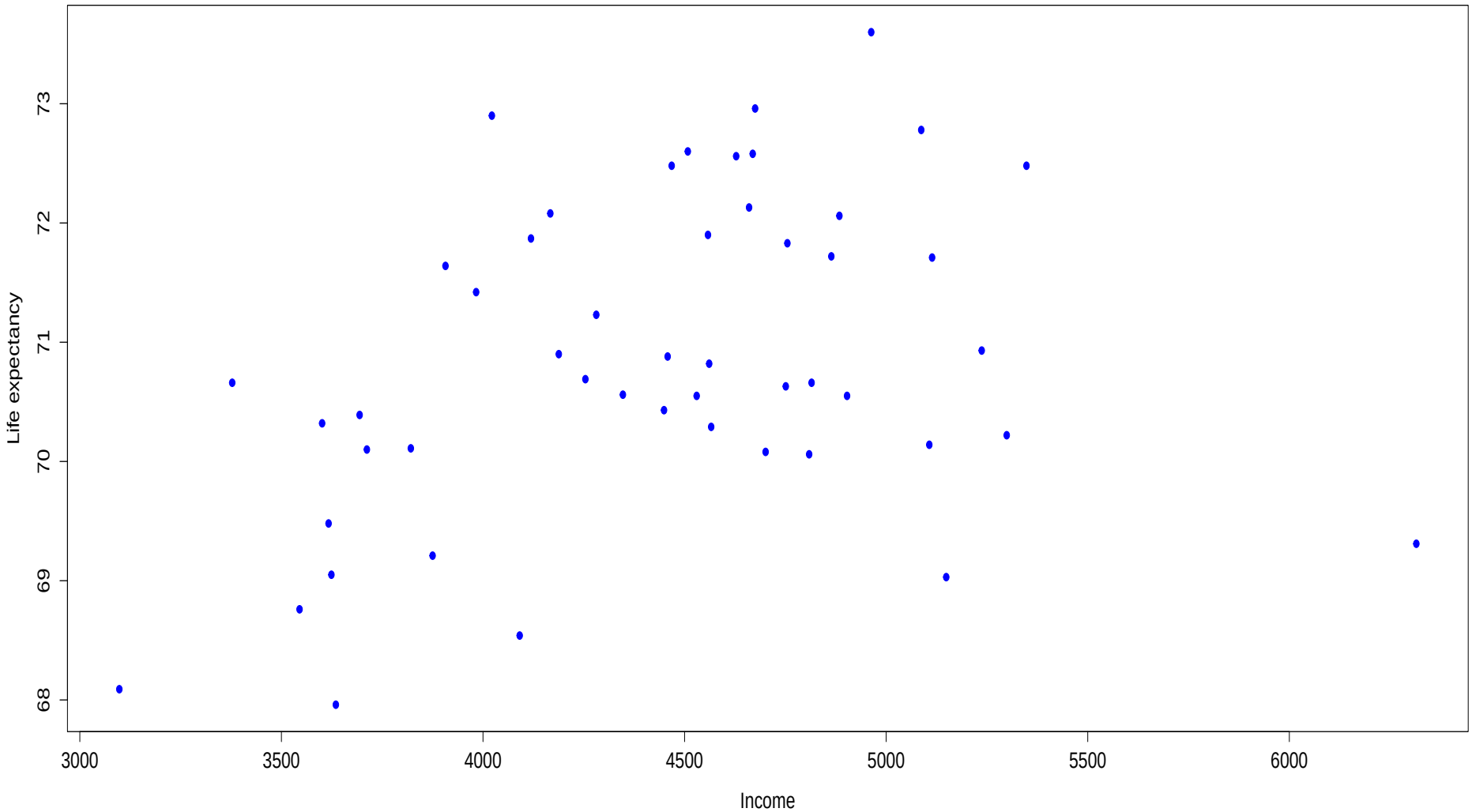
Visualizing multivariate data sets

- Scatterplots are bivariate plots of one variable against another.
- They help us to understand the relationship among the two variables.
- It is possible to extend the scatterplot by adding a third variable to obtain a 3D scatterplot.
- The scatterplot matrix draw all possible two-dimensional scatterplots of the variables.
- The scatterplot matrix helps in creating new ideas and in building knowledge about dependencies and structures among variables.

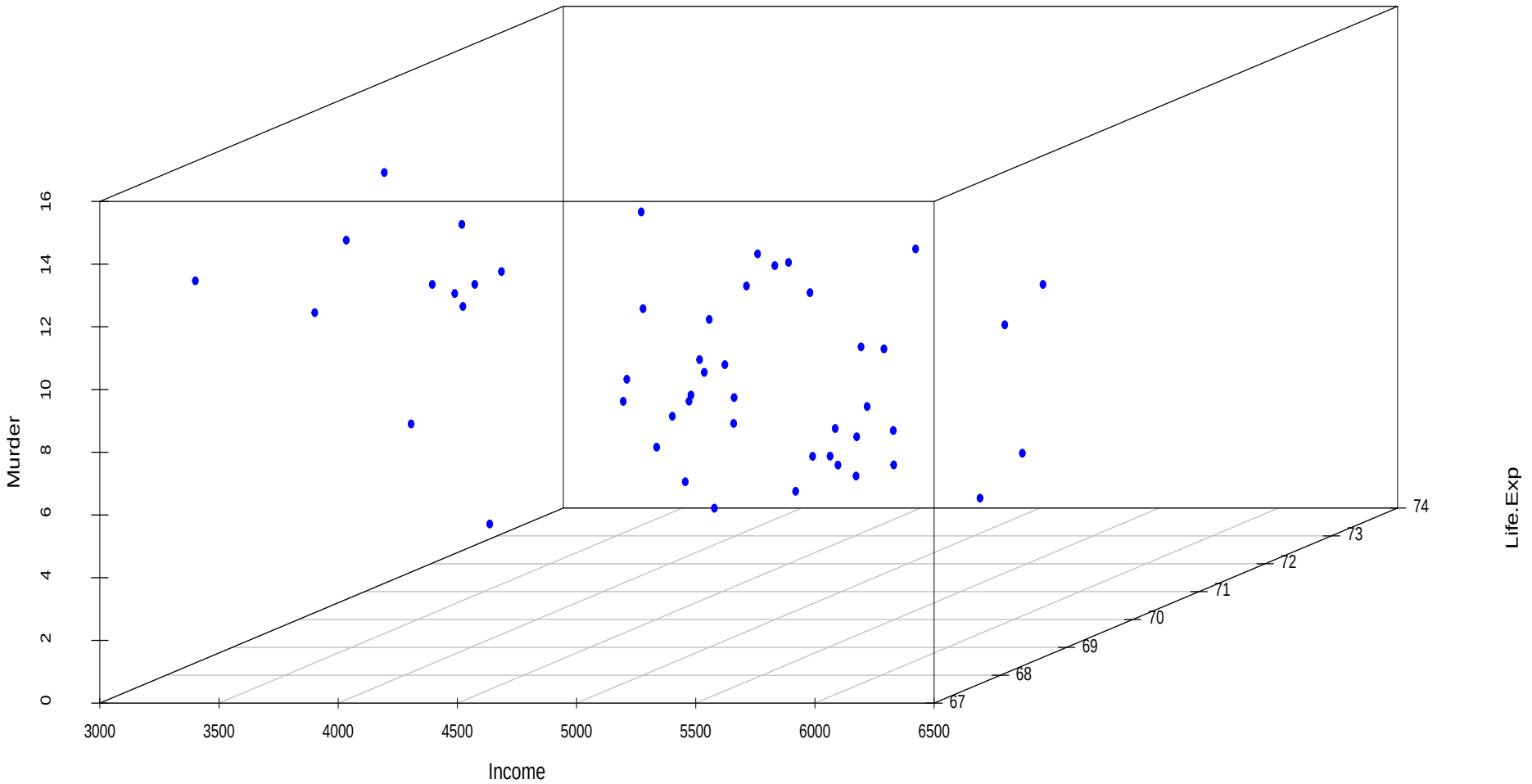
Illustrative example (I)

- The following four slides show scatterplots of the variables of the US states data set.
- In the first figure, it can be seen the scatterplot of the variables "Income" vs. "Life expectancy".
- In the second and third figures, it can be seen 3D-scatterplots of the variables "Income", "Life expectancy" and "Murder".
- Finally, the last figure shows a scatterplot matrix of the eight variables.
- As it can be seen, the relationships between variables are very different.
- There are both linear and non-linear relationships, outliers, and many other features.

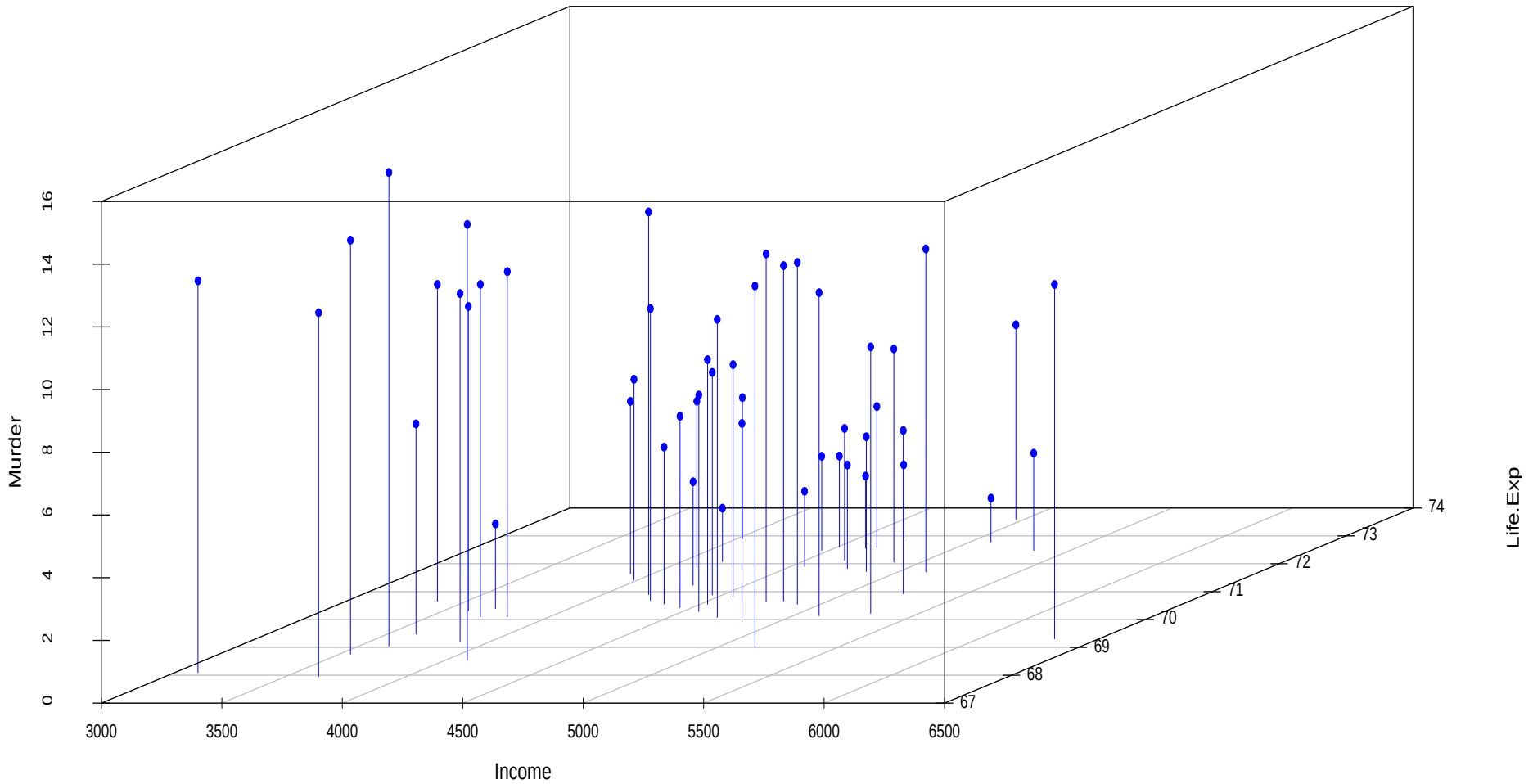
Illustrative example (I)



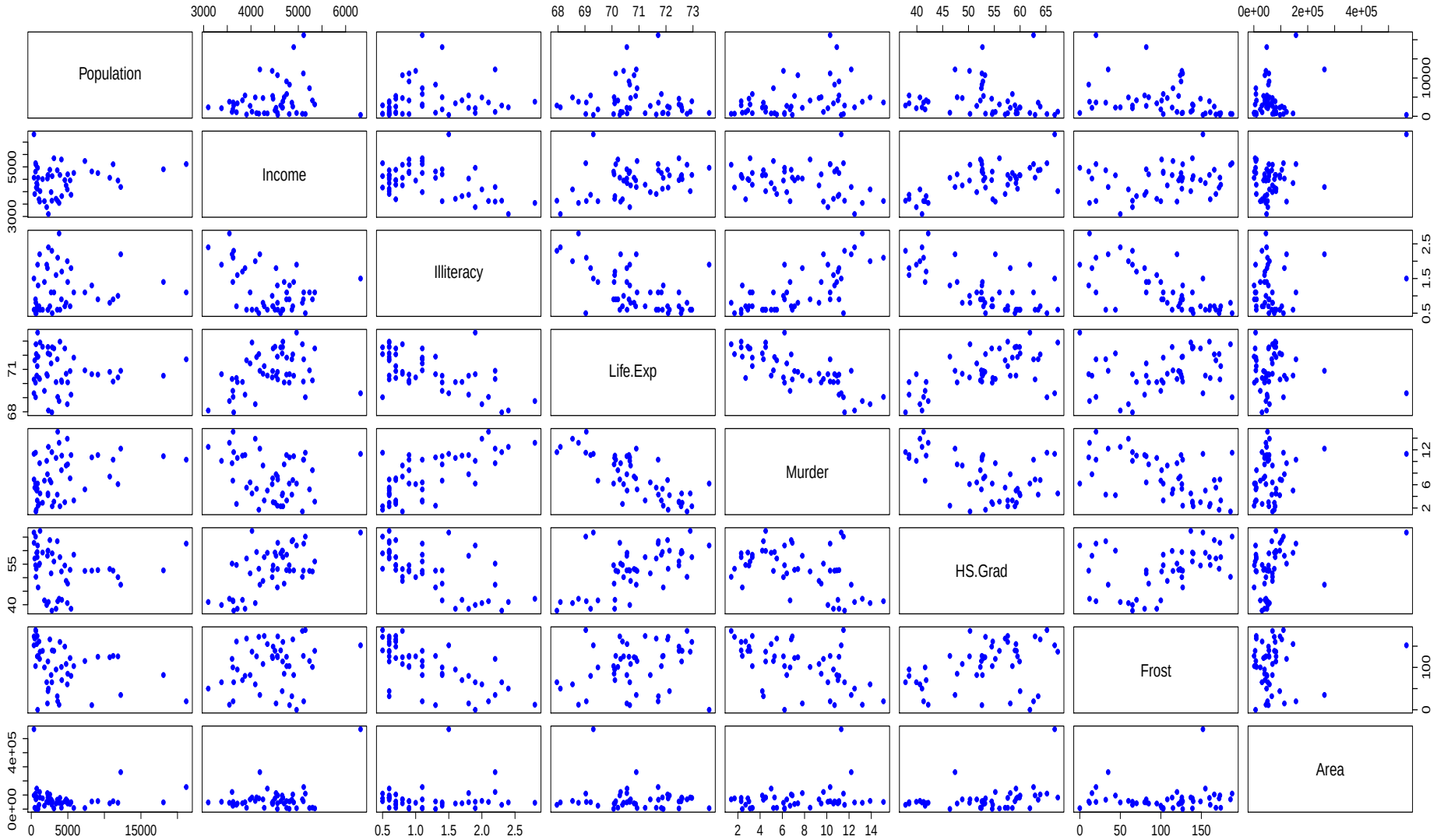
Illustrative example (I)



Illustrative example (I)



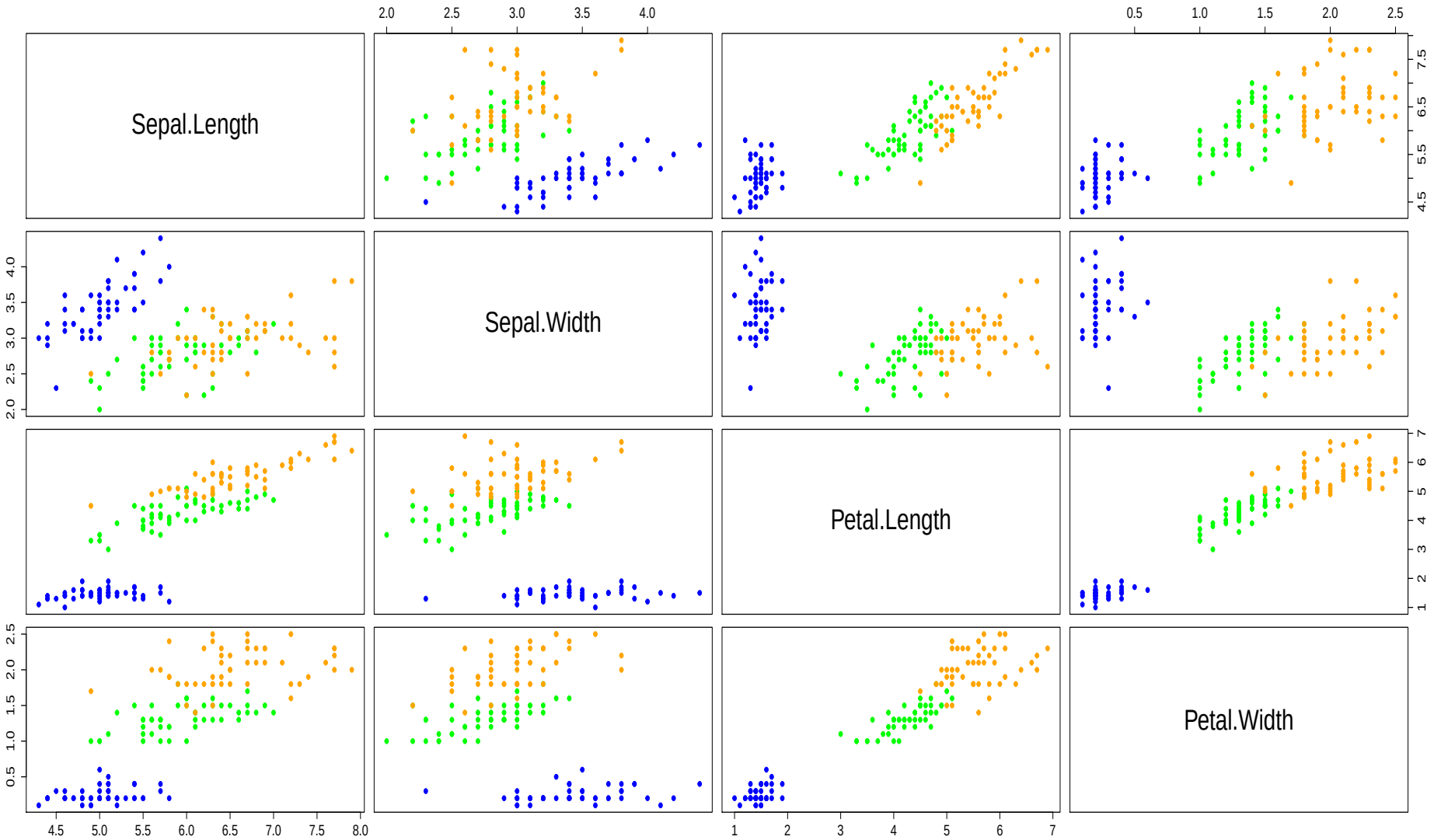
Illustrative example (I)



Illustrative example (II)

- The following slide shows a scatterplot matrix of the variables of the iris data set.
- Note that the figure represents the points with colors depending of the variable "Specie" (in blue, setosa, in green, versicolor, and in orange, virginica).
- Note how the relationship between the variables depends on the specie.

Illustrative example (II)



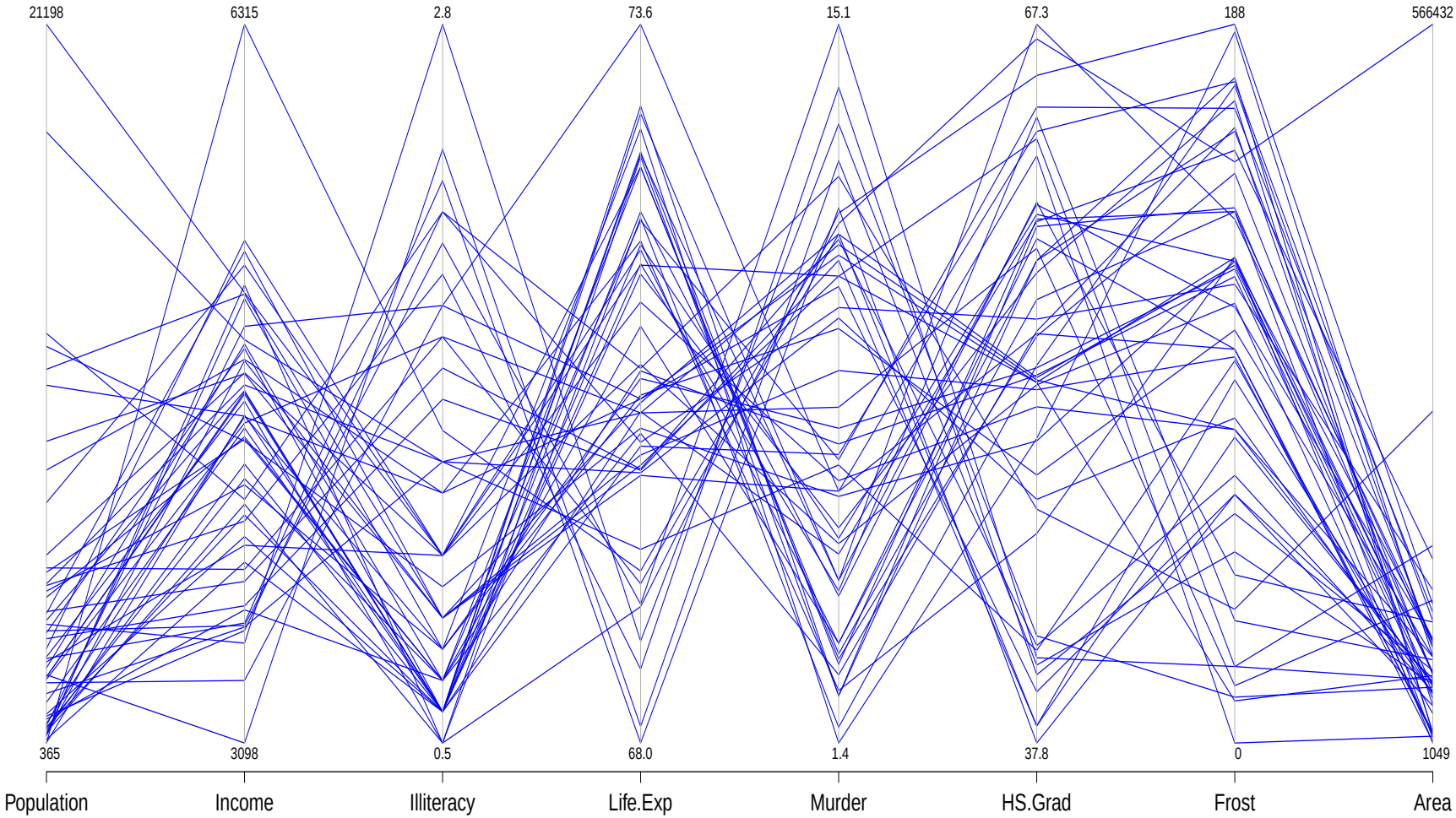
Visualizing multivariate data sets

- Parallel Coordinates Plots (PCP) is a method for representing high-dimensional data.
- Instead of plotting observations in an orthogonal coordinate system, PCP draws coordinates in parallel axes and connects them with straight lines.
- The variables are drawn into the horizontal axis, and the values of the variables are mapped onto the vertical axis.
- The PCP is very useful for high-dimensional data.
- However, it is sensitive to the order of the variables, since certain trends in the data can be shown more clearly in one ordering than in another.

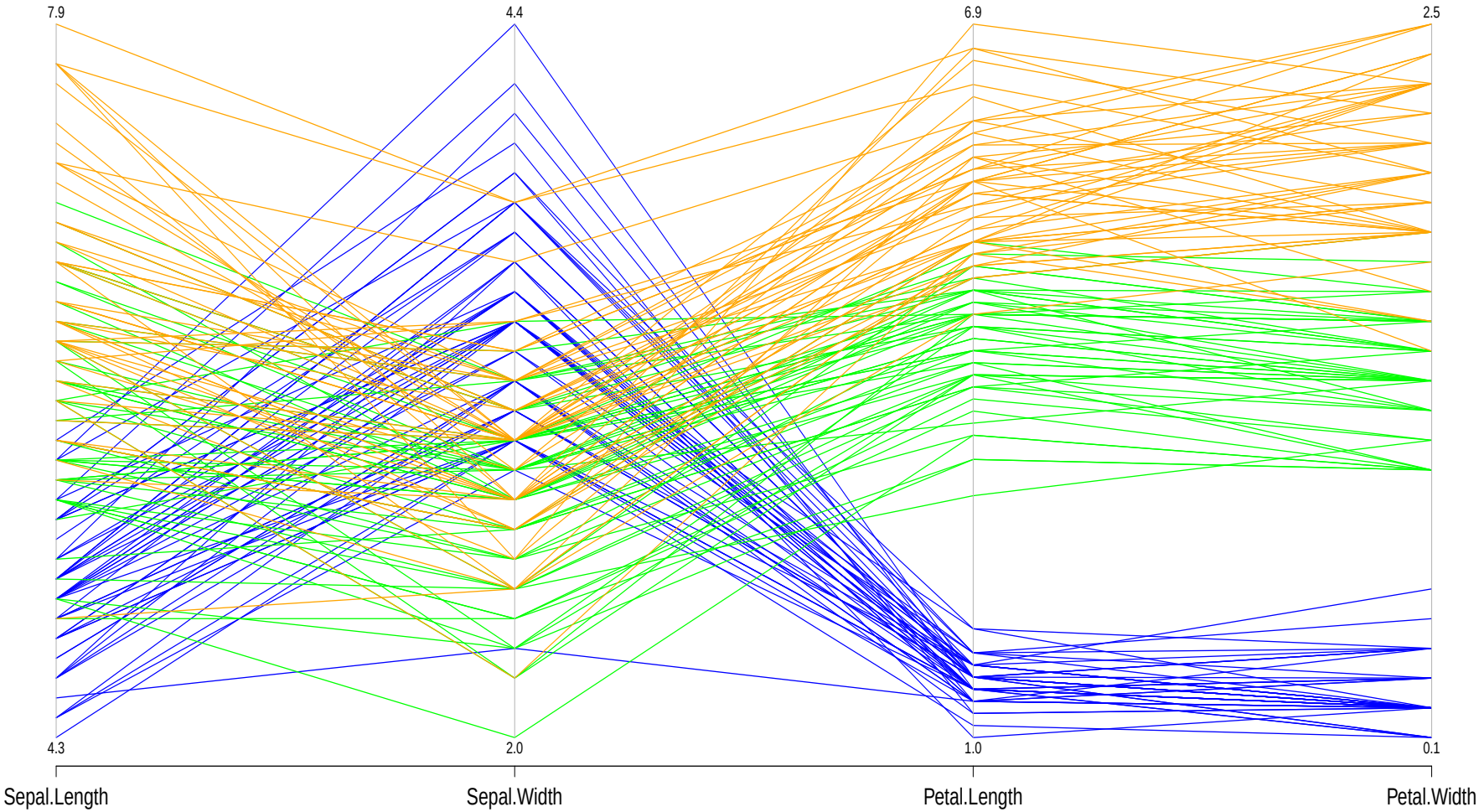
Illustrative examples (I) and (II)

- The following two slides show PCPs for the variables of the US states data set and the iris data set.
- Note that, in the first case, the PCPs help to detect outliers, at least outliers that appear in some of the variables.
- In the second case, note how the behavior of the variables strongly depends on the variable "Specie" (in blue, setosa, in green, versicolor, and in orange, virginica).

Illustrative example (I)



Illustrative example (II)



Multivariate descriptive measures

- We have seen that simple graphical devices can help in understanding the structure and dependency of data.
- The graphical tools are based on either univariate (bivariate) data representations or on transformations of multivariate information perceivable by the human eye.
- Most of the tools are extremely useful in a modelling step but do not give the full picture of the data set.
- One reason for this is that the graphical tools capture only certain dimensions of the data and do not concentrate on those dimensions or parts of the data under analysis that carry the maximum structural information.
- Chapters 3 and 4 will present powerful tools for reducing the dimension of a data set.
- Here, as a starting point, we use simple and basic tools to describe location, dispersion and dependency.

Multivariate descriptive measures

- Given our data matrix, X, we want to define in a proper way the center of the data.
- One possible criteria is to propose the point a as the center of the data if:

$$\sum_{i=1}^{n} (x_{i\cdot} - a) = 0_p$$

  where 0_p is the $p \times 1$ vector of zeros.
- Therefore, the point a is the center of balance of the data as the sum of its deviations is 0_p.

Multivariate descriptive measures

- From the previous equation,

$$a = \frac{1}{n}\sum_{i=1}^{n} x_{i\cdot} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix}$$

  where $\overline{x}_1, \ldots, \overline{x}_p$ are the sample means of the data of the variables $x_1, \ldots, x_p$.

- The point a above is called the sample mean vector, and it is denoted by $\overline{x}$.
- $\overline{x}$ is the natural extension of the sample mean of an univariate random sample.

Illustrative example (I)

- The sample mean vector for the data in the example is given by:

$$\overline{x} = (4246.42, 4435.80, 1.17, 70.87, 7.37, 53.10, 104.46, 70735.88)'$$

- Each component of the sample mean vector is the sample mean of the corresponding variable.

Multivariate descriptive measures

- On the other hand, the sample median cannot be easily generalized for multivariate variables because of the lack of a natural ranking in multivariate data (how to order multivariate observations?).
- This can be done using depth measures.
- The depth of an observation relative to the observations in the data matrix X measures how deep that observation lies in the data cloud formed by the observations in X.
- Therefore, the depths of the observations in X provides a center-outward ordering of these observations.
- Indeed, the deepest observation can be defined as a sample multivariate median.

Multivariate descriptive measures

- The halfspace depth is one of the most popular depth measures.
- The halfspace depth of an observation $x_{i\cdot}$ with respect to the observations in X is defined as the minimum fraction of observations of X contained in a closed halfspace containing $x_{i\cdot}$.
- Obviously, the problem of the halfspace depth is that its computation is very complicated when the dimension is relatively large.
- Alternatively, some approximations based on random generation of halfspaces can be done.
- Additionally, it is usual to get a certain number of ties among observations.

Illustrative example (I)

- We compute the halfspace depth of the states with 100000 random halfspaces.
- The deepest observation turns out to be the state of Iowa.
- This observation can be seen as a sample multivariate median of the states data set given by:

$$\overline{x} = (2861, 4628, 0.5, 72.56, 2.3, 59, 140, 55941)'$$

- Note the difference with respect to the sample mean vector.

Multivariate descriptive measures

- Given any univariate variable x_j of the data matrix X, the sample variance of x_j is given by:

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

- Given two univariate variables x_j and x_k of the data matrix X, the sample covariance between x_j and x_k is given by:

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

- The sample covariance measures the linear dependency between the observations of the variables x_j and x_k.

- It is very important to note that s_{jk} depends on the units of measurement of x_j and x_k.

Multivariate descriptive measures

- We can mimic the previous definition to define the sample covariance matrix of X is defined as:

$$S_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i\cdot} - \overline{x})(x_{i\cdot} - \overline{x})' = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \ddots & s_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

  where $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$.

- The sample covariance matrix contains the variances of x_j, for all $j = 1, \ldots, p$ and the covariances between any two univariate variables x_j and x_k, for all $j, k = 1, \ldots, p$ with $j \neq k$.

- Therefore, S_x contains all the information about the spread of the variables and the linear dependence between all the variables.

Multivariate descriptive measures

- Some properties of the sample covariance matrix are the following:
  - S_x is a symmetric matrix because $s_{jk} = s_{kj}$.
  - S_x can be written in terms of the centered data matrix $\widetilde{X} = X - 1_n \overline{x}'$:

$$S_x = \frac{1}{n-1} \widetilde{X}' \widetilde{X}$$

  - S_x is positive semidefinite, i.e., their eigenvalues $\lambda_1^{S_x}, \ldots, \lambda_p^{S_x}$ are non-negative, i.e., $\lambda_j^{S_x} \geq 0$, $j = 1, \ldots, p$.
  - $|S_x| = \prod_{j=1}^{p} \lambda_j^{S_x} \geq 0$.
  - When $|S_x| = 0$, there are some variables that are linear combinations of the others. Indeed, the rank of S_x is the number of linear independent variables. Then, if $|S_x| = 0$, it is necessary to delete the redundant variables.
  - $Tr(S_x) = s_1^2 + \cdots + s_p^2 = \lambda_1^{S_x} + \cdots + \lambda_p^{S_x} \geq 0$ (note that in general, $s_j^2 \neq \lambda_j^{S_x}$).

Illustrative example (I)

- The sample covariance matrix of the variables in the data set is given by:

$$S_x = \begin{pmatrix} 19.93 \times 10^6 & 57.12 \times 10^3 & 292.86 & -407.84 & 5663.52 & -3551.50 & -77.08 \times 10^3 & 8.58 \times 10^6 \\ 57.12 \times 10^3 & 37 \times 10^4 & -163.70 & 280.66 & -521.89 & 3076.76 & 7227.60 & 1.90 \times 10^7 \\ 292.86 & -163.70 & 0.37 & -0.48 & 1.58 & -3.23 & -21.29 & 4.01 \times 10^3 \\ -407.84 & 280.66 & -0.48 & 1.80 & -3.86 & 6.31 & 18.28 & -1.22 \times 10^4 \\ 5663.52 & -521.89 & 1.58 & -3.86 & 13.62 & -14.54 & -103.40 & 7.19 \times 10^4 \\ -3551.50 & 3076.76 & -3.23 & 6.31 & -14.54 & 65.23 & 153.99 & 2.29 \times 10^5 \\ -77.08 \times 10^3 & 7227.60 & -21.29 & 1828 & -103.40 & 153.99 & 2702.00 & 2.62 \times 10^5 \\ 85.87 \times 10^5 & 1.90 \times 10^7 & 4.01 \times 10^3 & -1.22 \times 10^4 & 7.19 \times 10^4 & 2.29 \times 10^5 & 2.62 \times 10^5 & 7.28 \times 10^9 \end{pmatrix}$$

- The eigenvalues of S_x are $\lambda_1^{S_x} = 7.28 \times 10^9$, $\lambda_2^{S_x} = 1.99 \times 10^7$, $\lambda_3^{S_x} = 3.12 \times 10^5$, $\lambda_4^{S_x} = 2.15 \times 10^3$, $\lambda_5^{S_x} = 36.51$, $\lambda_6^{S_x} = 6.05$, $\lambda_7^{S_x} = 0.43$ and $\lambda_8^{S_x} = 0.08$, respectively.
- Also, $Tr(S_x) = 7301060101$ and $|S_x| = 7.87 \times 10^{26}$.
- The eigenvalues are quite different. Why?
- Indeed, why is the largest eigenvalue so close to the variance of the last variable?

Multivariate descriptive measures

- The usual approach to solve the problem of having different units of measurement is to standardize the variables.
- Therefore, we can standardize the variables as follows:

$$y_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$$

- Now, the sample covariance between the data of the variables y_j and y_k is given by:

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} y_{ij}y_{ik} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_{ij} - \overline{x}_j)}{s_j}\frac{(x_{ik} - \overline{x}_k)}{s_k} = \frac{s_{jk}}{s_j s_k}$$

- This is called the sample correlation between the observations of two variables x_j and x_k of X.

Multivariate descriptive measures

- The sample correlation, as the sample covariance, measures the linear dependence between the observations of the variables x_j and x_k.
- However, the r_{jk} does not depend on the units of measurement of x_j and x_k.
- Note that r_{jk} is in absolute value always less than 1.
- In particular, the closer $|r_{jk}|$ to 1, the more linearly dependent the observations of x_j and x_k.
- In particular, $r_{jk} = 0$ if, and only if, $s_{jk} = 0$. In this case, we say that the observations of x_j and x_k are uncorrelated.

Multivariate descriptive measures

- Now, it is easy to see that the sample covariance matrix of the standardized variables in X is given by:

$$R_x = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \ddots & r_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

- R_x is called the sample correlation matrix and it is the multivariate (for more than 2 variables) counterpart of the sample correlation.
- The sample correlation matrix contains the correlations between any two univariate variables x_j and x_k, for all $j, k = 1, \ldots, p$ with $j \neq k$.
- Therefore, R_x contains all the information about the linear dependence between all the variables.

Multivariate descriptive measures

- Some properties of the sample correlation matrix are the following:
  - R_x is a symmetric matrix because $r_{jk} = r_{kj}$.
  - R_x can be written in terms of S_x as follows:

    $$R_x = D_x^{-1/2} S_x D_x^{-1/2}$$

    where D_x is the $p \times p$ diagonal matrix containing the elements of the main diagonal of S_x, i.e., the variances $s_1^2, \ldots, s_p^2$.
  - R_x is positive semidefinite, i.e., their eigenvalues $\lambda_1^{R_x}, \ldots, \lambda_p^{R_x}$ are non-negative, i.e., $\lambda_j^{R_x} \geq 0$, $j = 1, \ldots, p$.
  - $|R_x| = \prod_{j=1}^{p} \lambda_j^{R_x} \geq 0$.
  - When $|R_x| = 0$, there are some variables that are linear combinations of the others. Indeed, the rank of R_x is the number of linear independent variables. Then, if $|R_x| = 0$, it is necessary to delete the redundant variables.
  - $Tr(R_x) = 1 + \cdots + 1 = \lambda_1^{R_x} + \cdots + \lambda_p^{R_x} = p$ (note that in general, $\lambda_j^{R_x} \neq 1$).

Illustrative example (I)

- The sample correlation matrix of the variables in the data set is given by:

$$R_x = \begin{pmatrix} 1 & 0.20 & 0.10 & -0.06 & 0.34 & -0.09 & -0.33 & 0.02 \\ 0.20 & 1 & -0.43 & 0.34 & -0.23 & 0.61 & 0.22 & 0.36 \\ 0.10 & -0.43 & 1 & -0.58 & 0.70 & -0.65 & -0.67 & 0.07 \\ -0.06 & 0.34 & -0.58 & 1 & -0.78 & 0.58 & 0.26 & -0.10 \\ 0.34 & -0.23 & 0.70 & -0.78 & 1 & -0.48 & -0.53 & 0.22 \\ -0.09 & 0.61 & -0.65 & 0.58 & -0.48 & 1 & 0.36 & 0.33 \\ -0.33 & 0.22 & -0.67 & 0.26 & -0.53 & 0.36 & 1 & 0.05 \\ 0.02 & 0.36 & 0.07 & -0.10 & 0.22 & 0.33 & 0.05 & 1 \end{pmatrix}$$

- The eigenvalues of R_x are $\lambda_1^{R_x} = 3.59$, $\lambda_2^{R_x} = 1.63$, $\lambda_3^{R_x} = 1.11$, $\lambda_4^{R_x} = 0.70$, $\lambda_5^{R_x} = 0.38$, $\lambda_6^{R_x} = 0.30$, $\lambda_7^{R_x} = 0.14$ and $\lambda_8^{R_x} = 0.11$, which are not very close to 0. Why?

- Also, $Tr(R_x) = 6$ and $|R_x| = 0.0089$.

Multivariate descriptive measures

- There are other coefficients to measure the dependency between the data of two random variables.
- For instance, the Kendall's tau and the Spearman's rho are two correlation coefficients based on the ranks of the data.
- However, there are no available generalizations of these coefficients to more than two dimensions so we do not enter into details here.

Linear transformations

- In many practical applications we need to study linear transformations of the original data.
- For instance, to define the sample correlation matrix, we have standardize the data, which is a linear transformation of the variables.
- This motivates the question of how to calculate descriptive statistics after such linear transformations.

Linear transformations

- Let X be a $n \times p$ data matrix and let $c = (c_1, \ldots, c_p)'$ be a $p \times 1$ column vector.
- Then, the $n \times 1$ column vector $y = Xc$ is a linear combination of X:

$$Y = Xc = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

  where $y_i = c_1 x_{i1} + \cdots + c_p x_{ip}$, for $i = 1, \ldots, n$.

- The sample mean and the sample variance of the new variable are given by:

$$\overline{y} = c'\overline{x} \qquad s_y^2 = c' S_x c$$

  respectively, where $\overline{x}$ and S_x are the sample mean and sample covariance matrix of X.

Linear transformations

- If C is a $p \times r$ matrix, then, the $n \times r$ data matrix $Y = XC$ is a linear transformation of X:

$$Y = XC = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1r} \\ c_{21} & c_{22} & \cdots & c_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pr} \end{pmatrix} =$$

$$= \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nr} \end{pmatrix}$$

  where $y_{ij} = c_{1j}x_{i1} + \cdots + c_{pj}x_{ip}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

- Then, the sample mean vector and sample covariance matrix of Y are given by:

$$\overline{y} = C'\overline{x} \qquad S_y = C'S_xC$$

  respectively.

Illustrative example (II)

- The mean vector of the iris data set is given by:

$$\overline{x} = (5.84, 3.05, 3.75, 1.19)'$$

- The sample covariance matrix of the iris data set is given by:

$$S_x = \begin{pmatrix} 0.68 & -0.04 & 1.27 & 0.51 \\ -0.04 & 0.18 & -0.32 & -0.12 \\ 1.27 & -0.32 & 3.11 & 1.29 \\ 0.51 & -0.12 & 1.29 & 0.58 \end{pmatrix}$$

- We want to create two new variables from the variables in the data matrix X:
  - the sum of the lengths of the sepal and the petal of each flower; and
  - the sum of the widths of the sepal and the petal of each flower.

Illustrative example (II)

- The problem is to compute the mean and the covariance matrix of the new data set given by:

$$Y = XC$$

where:

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- Then,

$$\overline{y} = C'\overline{x} = \begin{pmatrix} 9.60 \\ 4.25 \end{pmatrix}$$

and,

$$S_y = C'S_xC = \begin{pmatrix} 6.35 & 1.43 \\ 1.43 & 0.52 \end{pmatrix}$$

Linear transformations

- The individual standardization of X can be written as:

$$Y = \widetilde{X} D_x^{-1/2}$$

where D_x is the $p \times p$ diagonal matrix formed by the elements of the principal diagonal of S_x, i.e., the variances $s_1^2, \ldots, s_p^2$.

- Note that:

$$\overline{y} = 0_p \qquad S_y = D_x^{-1/2} S_x D_x^{-1/2} = R_x$$

- Therefore, the univariate standardization eliminates the mean and standardises the variance of each variable.

Linear transformations

- The multivariate standardization of X is given by:

$$Y = \widetilde{X} S_x^{-1/2}$$

- Note that:

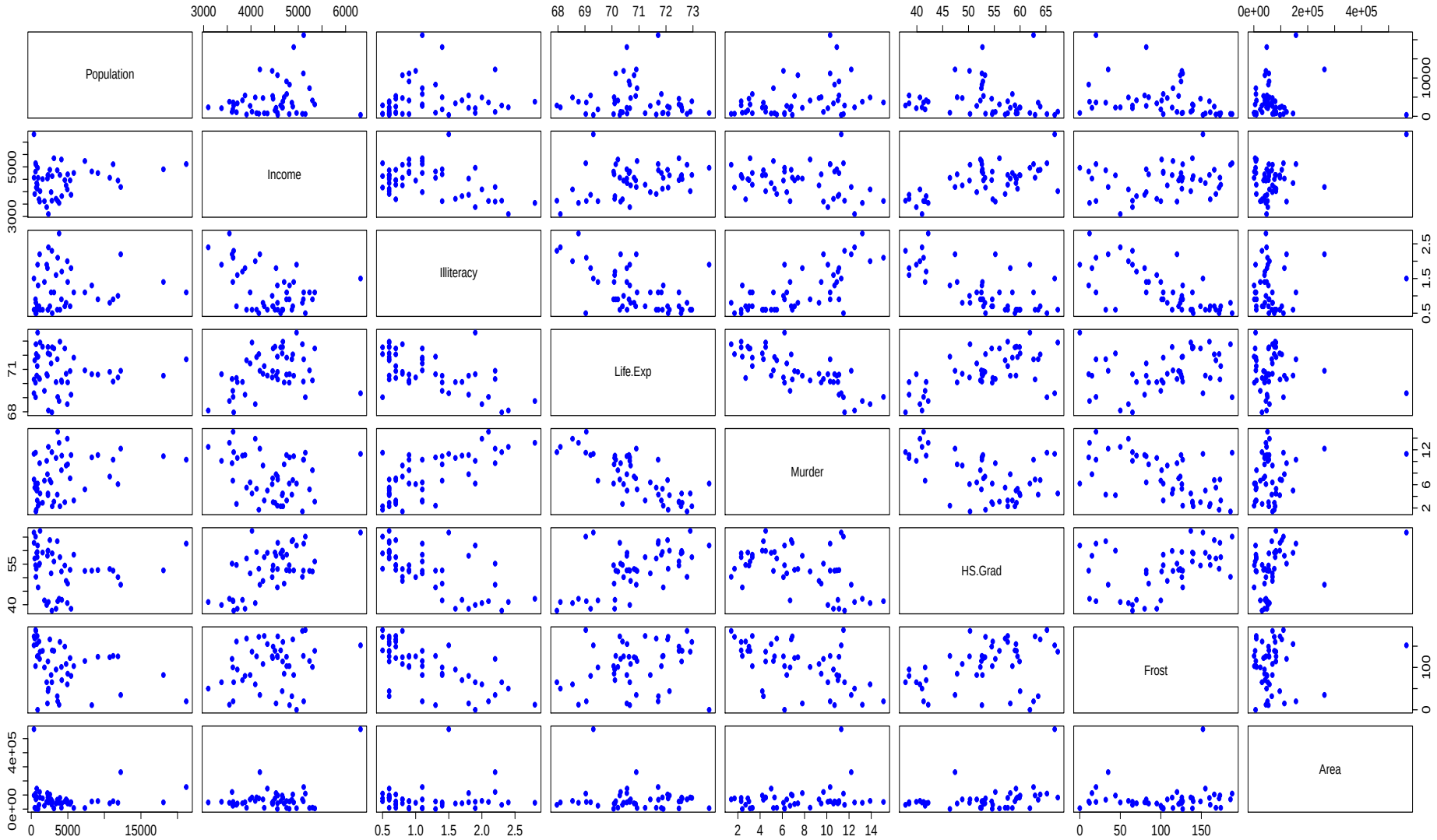
$$\overline{y} = 0_p \qquad S_y = \left(S_x^{-1/2}\right)' S_x S_x^{-1/2} = I_p$$

- Therefore, the multivariate standardization eliminates the mean and the correlation between the variables and standardises the variance of each variable.

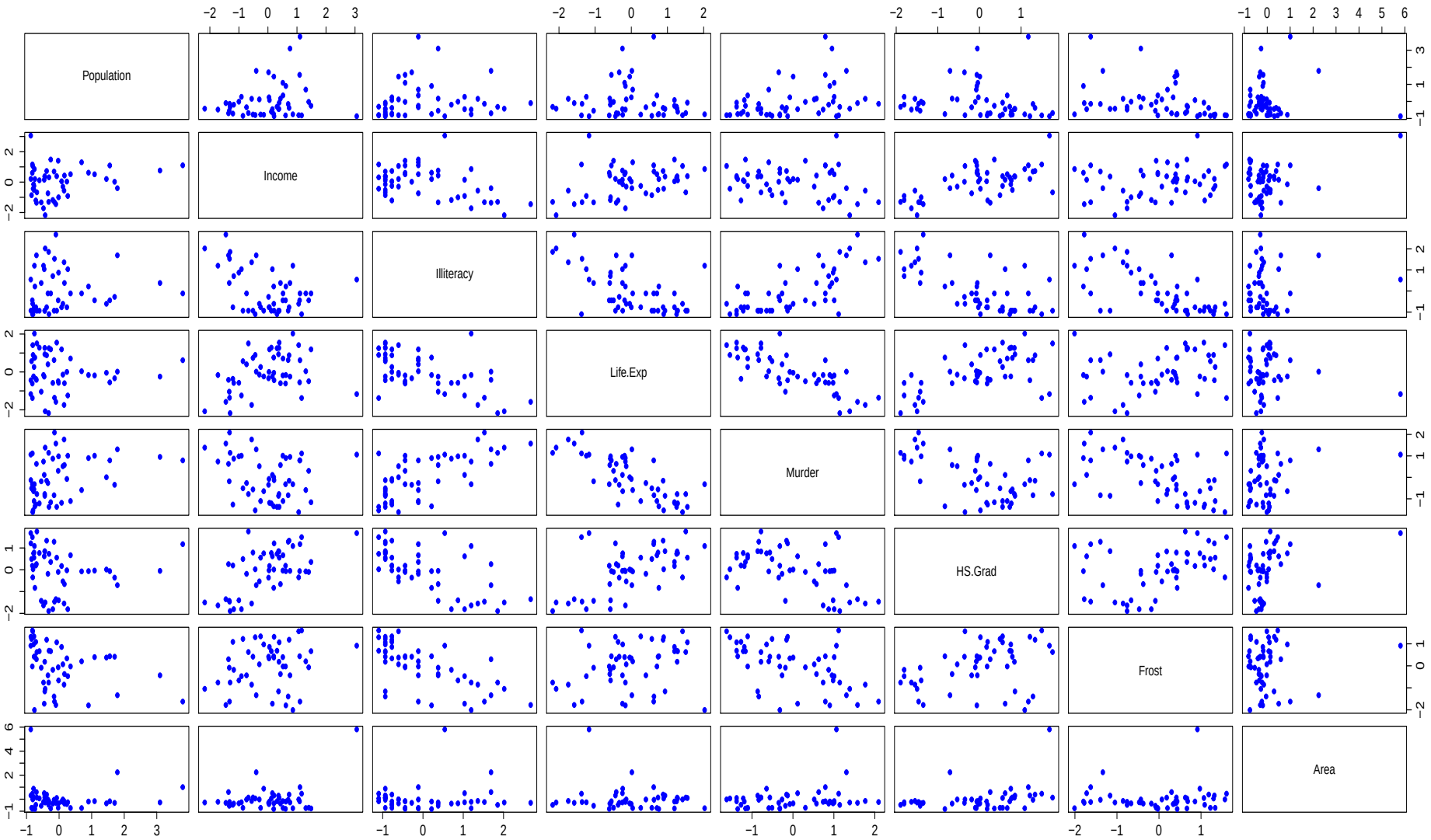
Illustrative example (I)

- Next, see the scatterplot matrices corresponding to:
  1. The original state data set.
  2. The univariate standardized data set.
  3. The multivariate standardized data set.

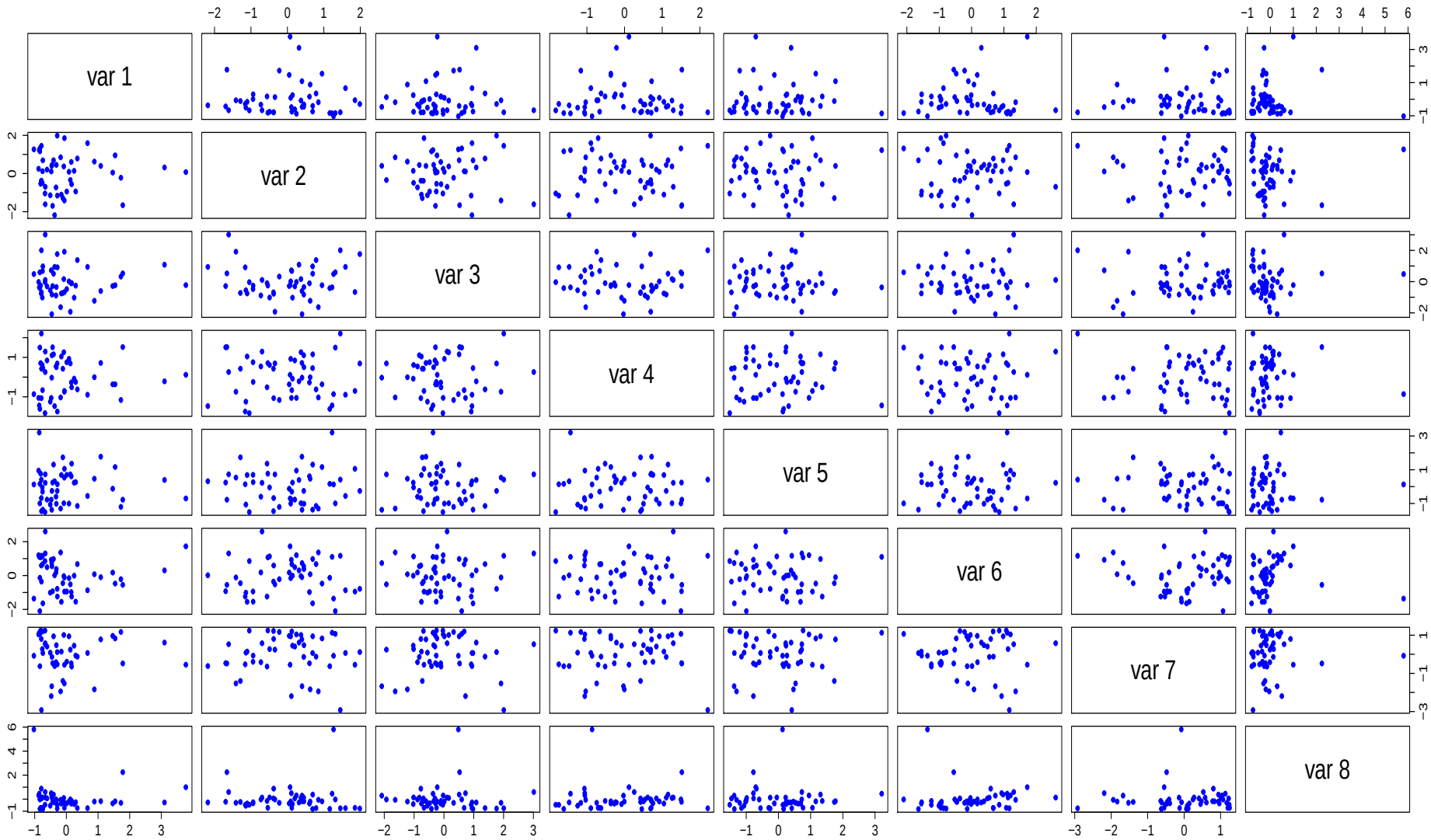
Illustrative example (I)



Illustrative example (I)



Illustrative example (I)



Conclusion

- We are ready now for:

Chapter 2: Multivariate distributions and inference