Data Visualization II

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- This is a natural continuation of the material from the last "lecture".
- In the last lecture, most of the figures were based on examples in Unwin (2015)[a].
- The same is true of this lecture, but the figures from the first example are based on figures in McNicholas (2016)[b].

[a] Unwin, A. (2015). *Graphical Data Analysis with R*. Boca Raton: Chapman & Hall/CRC Press.

[b] McNicholas, P.D. (2016). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.

Bivariate Gaussian Distributions

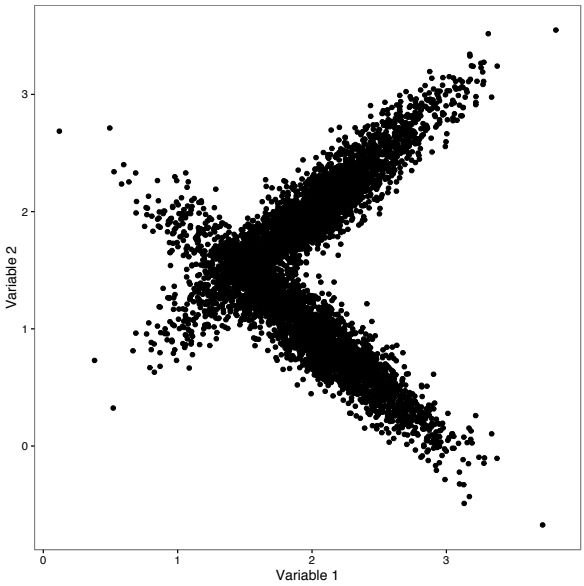
- McNicholas (2016) uses the example of two overlapping bivariate Gaussian densities to illustrate why the number of clusters should not be taken equal to the number of modes.

- These data contain 6,000 observations, each simulated from one of two bivariate Gaussian distributions.

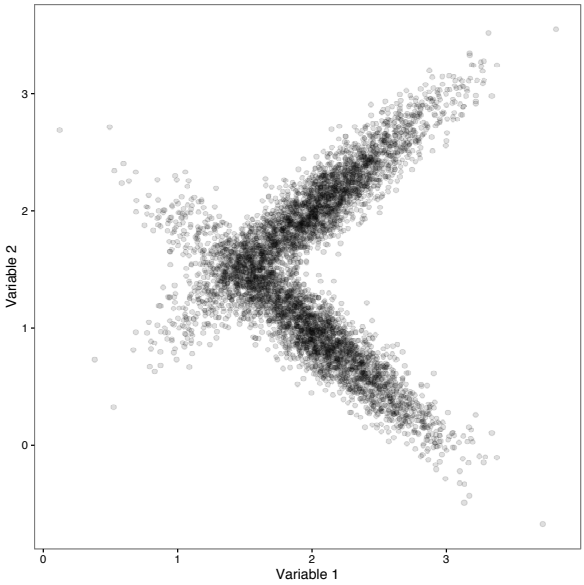
- This presents a data visualization challenge — a standard scatter plot looks like a blob.

- McNicholas (2016) uses semi-transparent points to overcome this challenge.

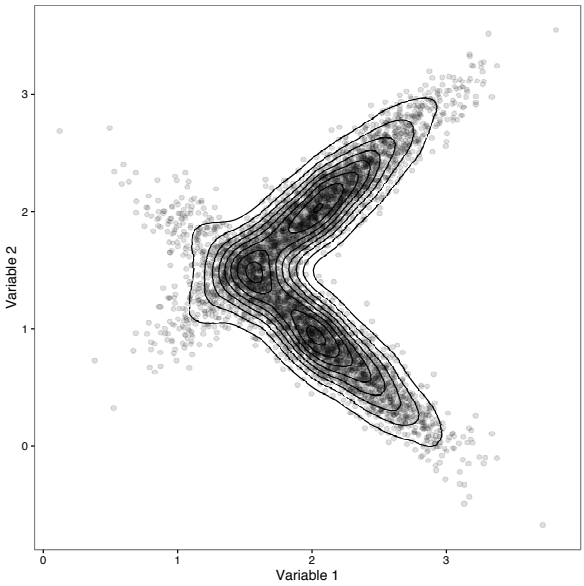
Scatter Plot



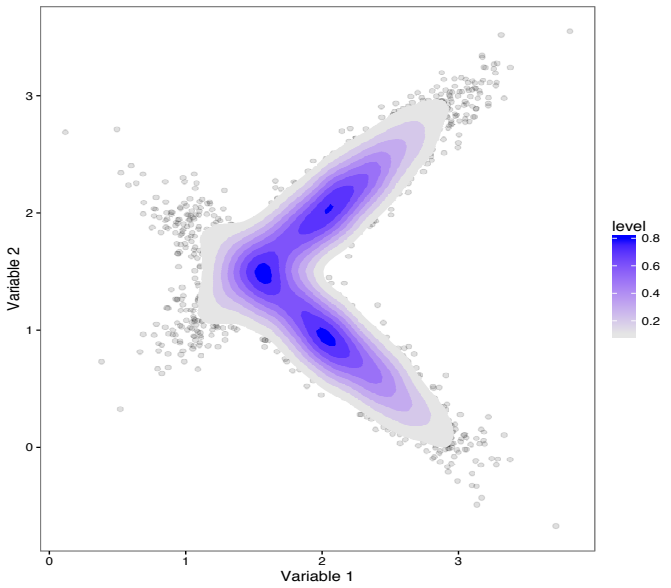
With Semi-Transparent Points



With Countours



With Shaded Contours



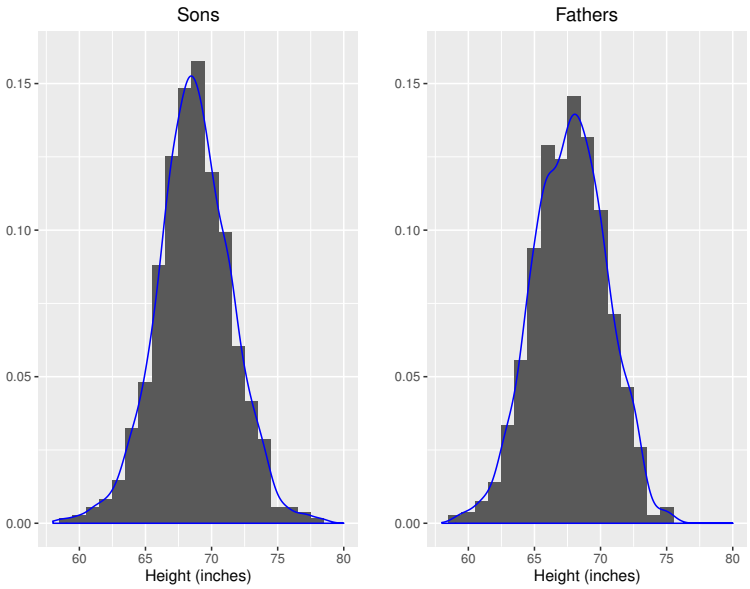
Comments

- Using semi-transparent points can be very convenient.
- In this example, we were able to see areas of high-density on the scatter plot (rather than a blob).
- Let's look at a very famous data set...

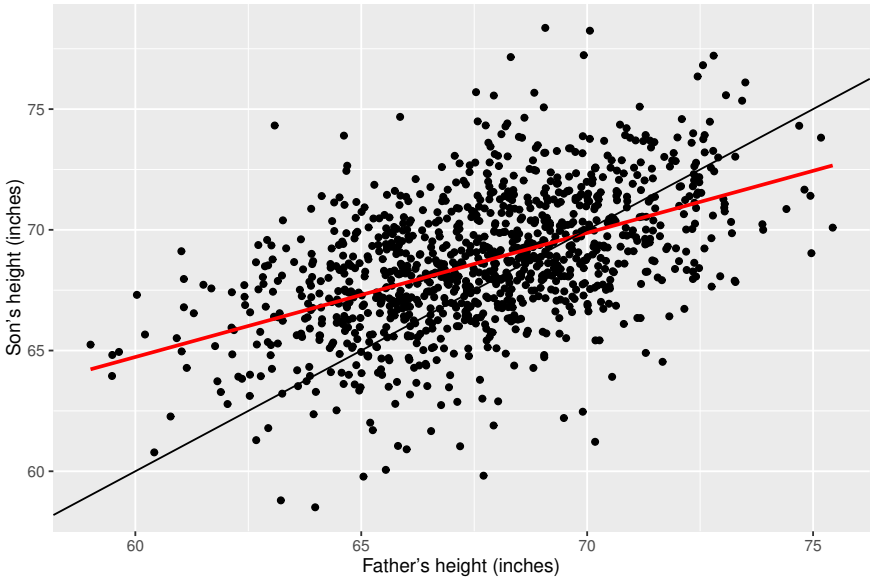
Fathers & Sons

- Height for 1,078 fathers and sons (in inches).
- Very famous example used by Pearson.
- One of the fundamental examples of regression — perhaps the fundamental example.

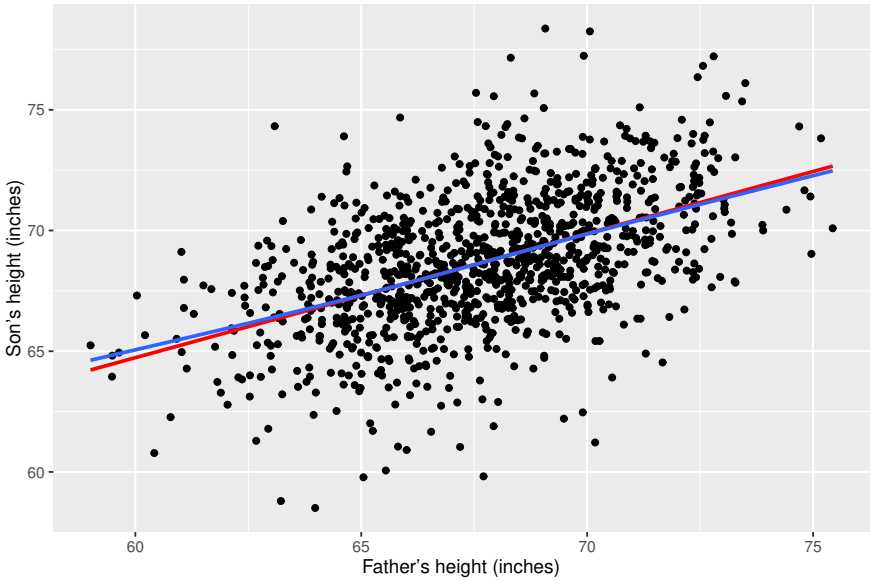
Fathers & Sons: Histograms



Fathers & Sons: Scatter + Line



Fathers & Sons: Line + Smooth Line

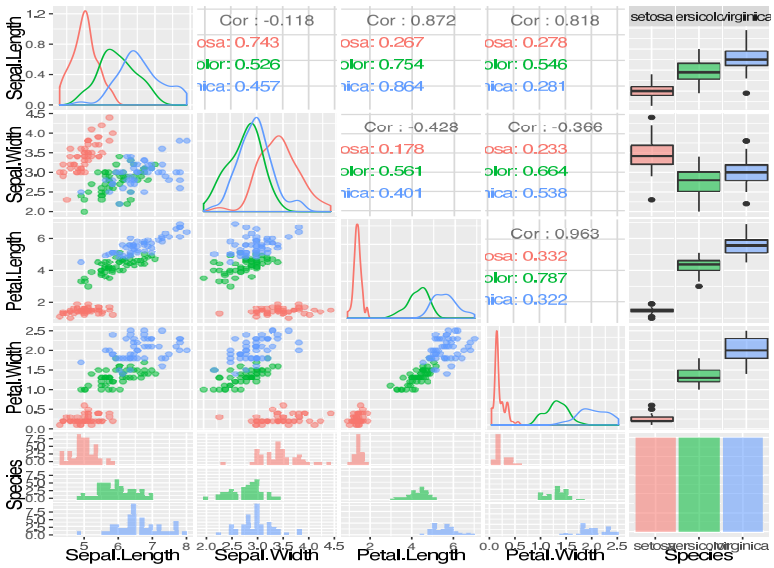


Iris Data

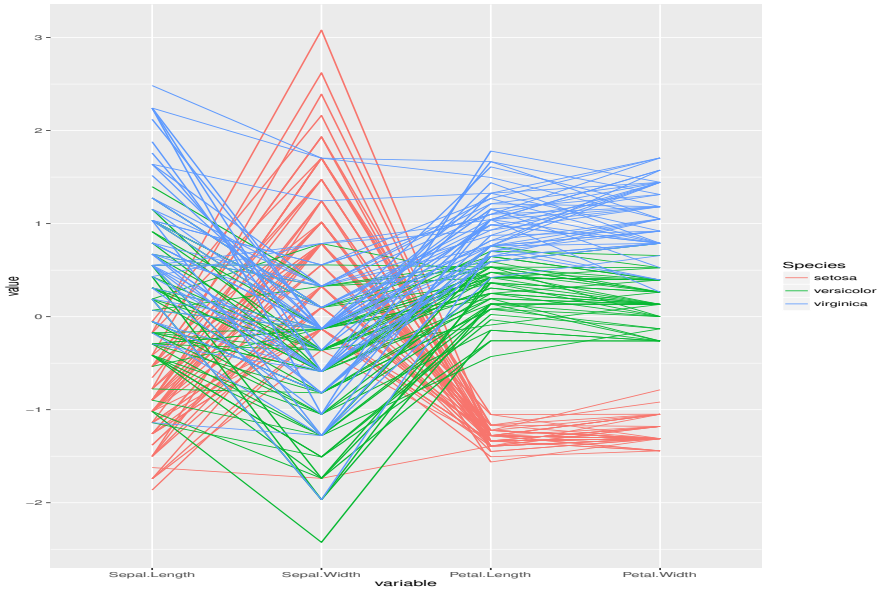
- Very famous data collected by Anderson (1935)[a]
- Measurements (in cm) of four variables — sepal length and width as well as petal length and width — for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.
- There are 50 flowers of each species.

[a] Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society **59**, 2–5.

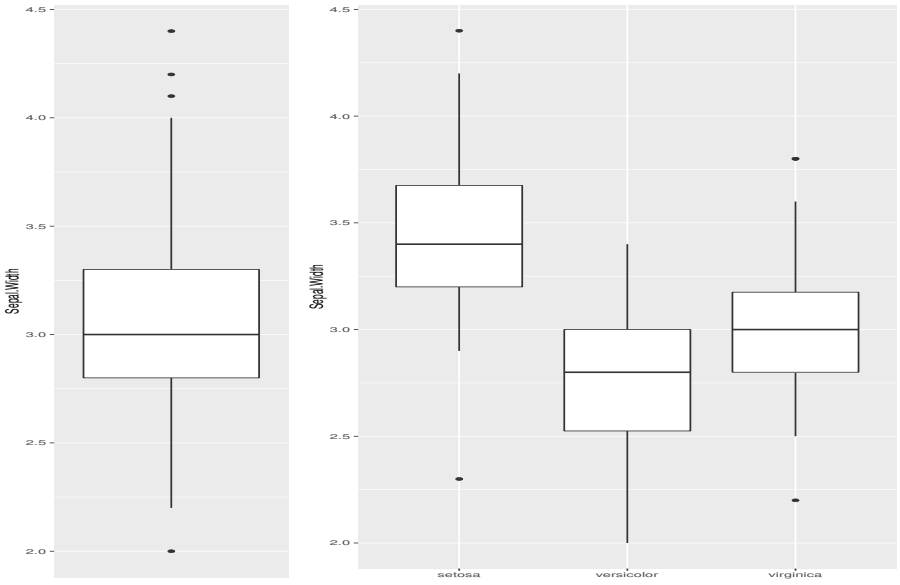
Iris: Pairs+



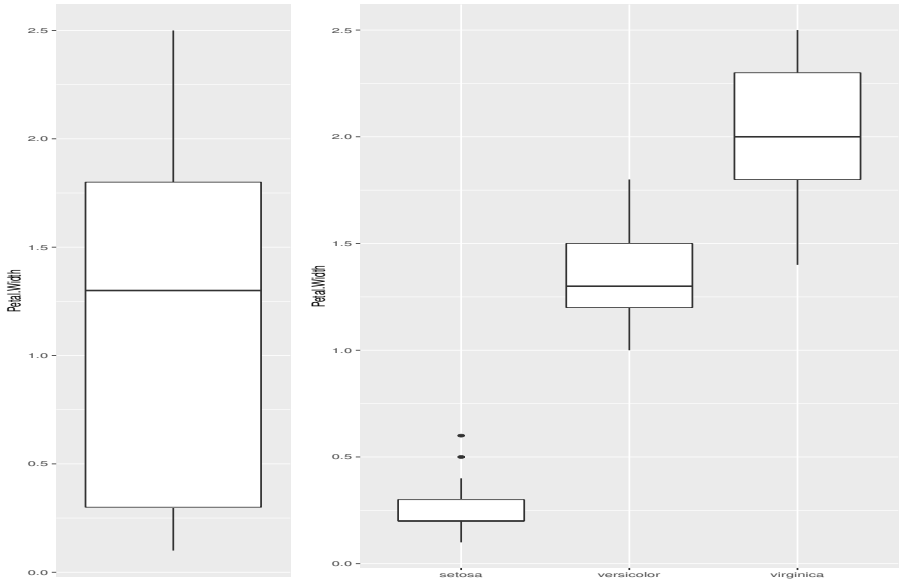
Iris: Parallel Coordinates



Iris: Some Box Plots



Iris: Some More Box Plots



Body Data

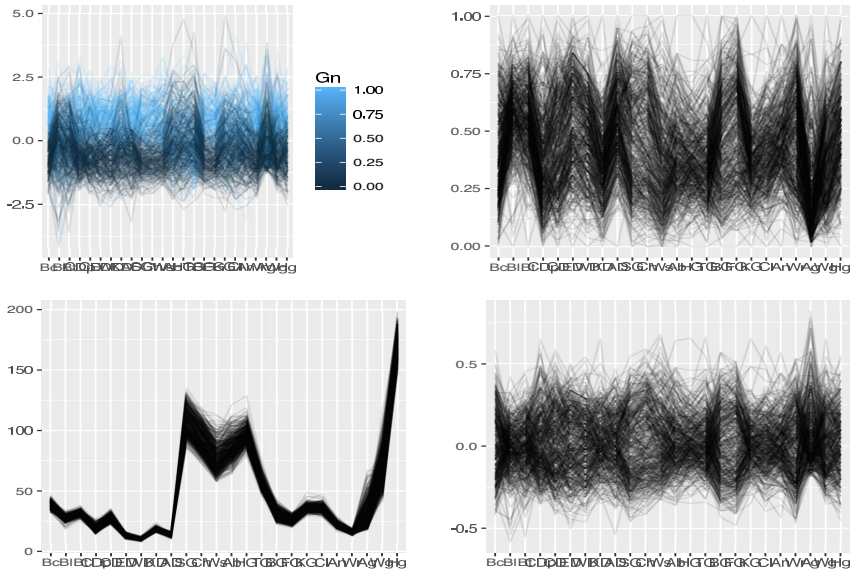
- The `body` dataset from `gclus` contains 21 body dimension measurements as well as age (in years), weight (in kg), height (in cm), and gender (binary) on 507 individuals.

- These break down as 247 men and 260 women.

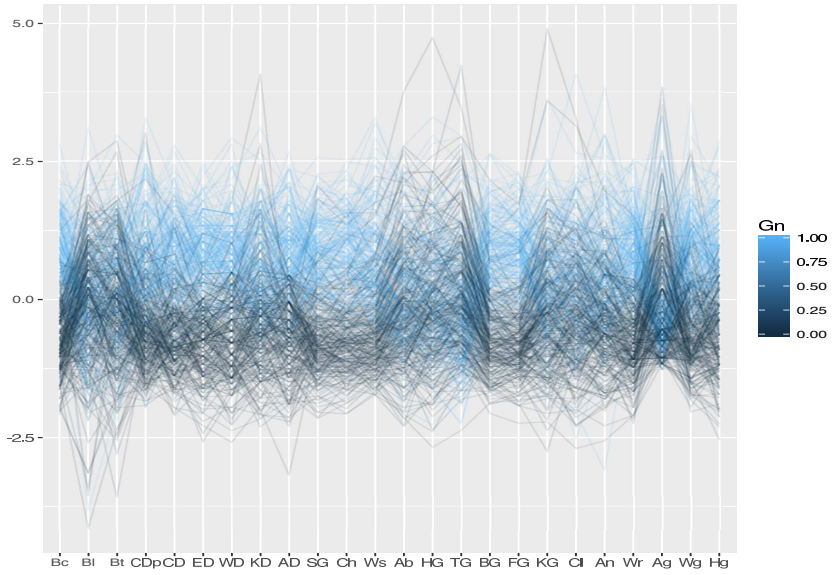
- Mostly, these are people in their twenties and thirties, with some older men and women.

- All exercise several hours a week.

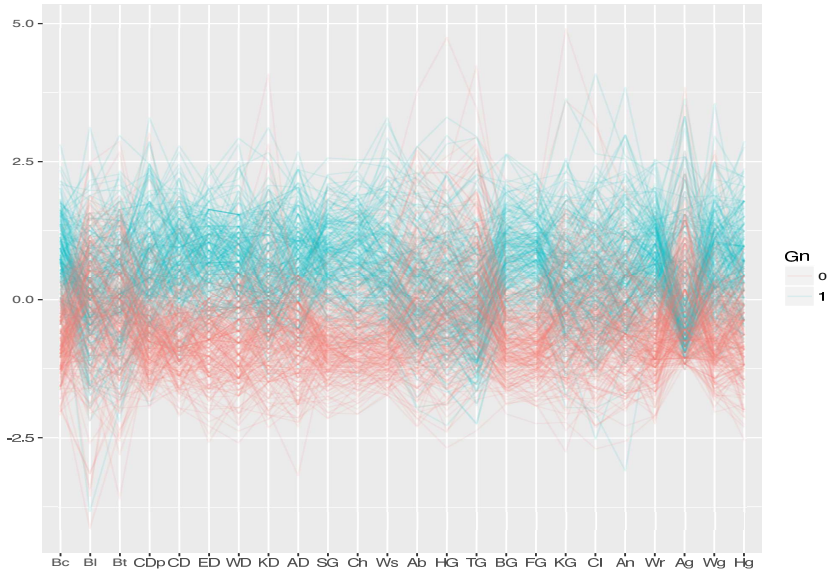
Body Data: Four Parallel Coord.



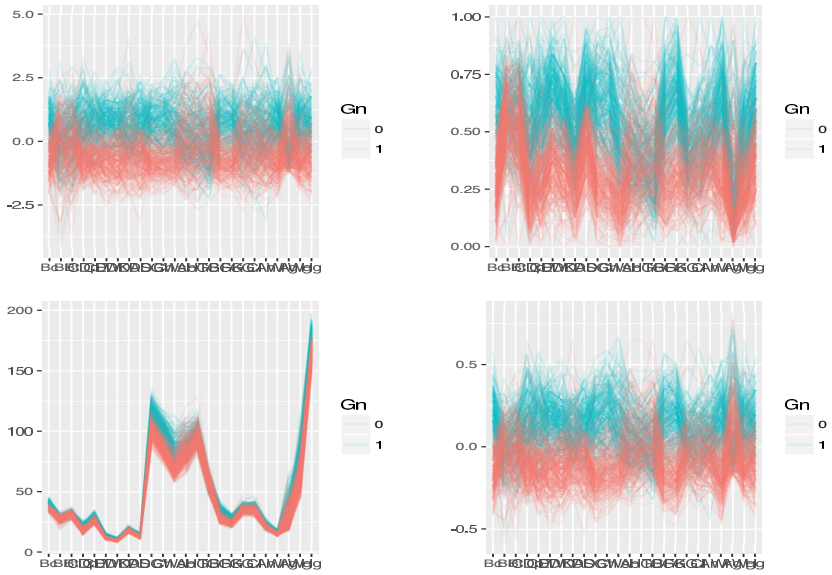
Body Data: Top-Left by Gender (??)



Body Data: Top-Left by Gender



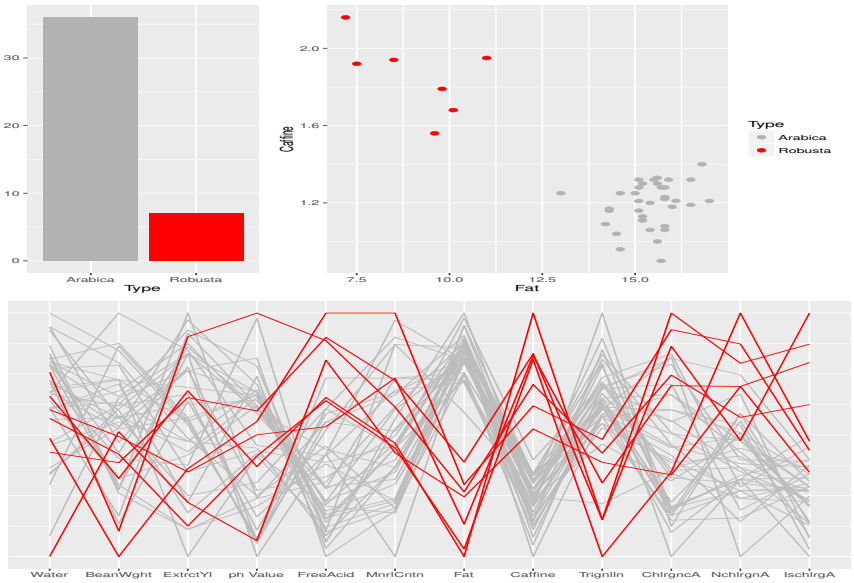
Body Data: All by Gender



Coffee Data

- Data on the chemical composition of coffee samples collected from around the world.
- A total of 43 samples from 29 countries.
- Each sample is either of the Arabica or Robusta variety.
- Twelve chemical constituents available in the `pgmm` package.

Coffee Data: Summary Panel



Comments

- We have seen a lot in this "lecture".
- We have learned, *inter alia*, that scale is very important for parallel coordinate plots (and in general).
- For further reading, see Unwin (2015).
- Now, let's play around a bit in R...