Data Exploration and Visualization with R

Yanchang Zhao

http://www.RDataMining.com

30 September 2014

Outline

Data Exploration and Visualization with R [1]

Data Exploration and Visualization

- Summary and stats
- Various charts like pie charts and histograms
- Exploration of multiple variables
- Level plot, contour plot and 3D plot
- Saving charts into files of various formats

[1] Chapter 3: Data Exploration, in book *R and Data Mining: Examples and Case Studies*. `http://www.rdatamining.com/docs/RDataMining.pdf`

Outline

Size and Structure of Data

```
dim(iris)

## [1] 150   5

names(iris)

## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Wid...
## [5] "Species"

str(iris)

## 'data.frame': 150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",....
```

Attributes of Data

```r
attributes(iris)

## $names
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Wid...
## [5] "Species"
##
## $row.names
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  ...
##  [16]  16  17  18  19  20  21  22  23  24  25  26  27  28  ...
##  [31]  31  32  33  34  35  36  37  38  39  40  41  42  43  ...
##  [46]  46  47  48  49  50  51  52  53  54  55  56  57  58  ...
##  [61]  61  62  63  64  65  66  67  68  69  70  71  72  73  ...
##  [76]  76  77  78  79  80  81  82  83  84  85  86  87  88  ...
##  [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 1...
## [106] 106 107 108 109 110 111 112 113 114 115 116 117 118 1...
## [121] 121 122 123 124 125 126 127 128 129 130 131 132 133 1...
## [136] 136 137 138 139 140 141 142 143 144 145 146 147 148 1...
##
## $class
## [1] "data.frame"
```

First Rows of Data

```
iris[1:3, ]

##   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

head(iris, 3)

##   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

tail(iris, 3)

##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Spe...
## 148          6.5         3.0          5.2         2.0 virgi...
## 149          6.2         3.4          5.4         2.3 virgi...
## 150          5.9         3.0          5.1         1.8 virgi...
```

A Single Column

The first 10 values of `Sepal.Length`

```
iris[1:10, "Sepal.Length"]

##  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9

iris$Sepal.Length[1:10]

##  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

Outline

Summary of Data

Function `summary()`

- numeric variables: minimum, maximum, mean, median, and the first (25%) and third (75%) quartiles
- categorical variables (factors): frequency of every level

```
summary(iris)

##   Sepal.Length   Sepal.Width    Petal.Length   Petal.Width
##  Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1
##  1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3
##  Median :5.80   Median :3.00   Median :4.35   Median :1.3
##  Mean   :5.84   Mean   :3.06   Mean   :3.76   Mean   :1.2
##  3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8
##  Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```r
library(Hmisc)
describe(iris[, c(1, 5)])  # check columns 1 & 5

## iris[, c(1, 5)]
##
##  2  Variables      150  Observations
## ---------------------------------------------------------...
## Sepal.Length
##       n missing  unique    Info    Mean     .05     .10    ...
##     150       0      35       1   5.843   4.600   4.800   5...
##     .50     .75     .90     .95
##   5.800   6.400   6.900   7.255
##
## lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
## ---------------------------------------------------------...
## Species
##       n missing  unique
##     150       0       3
##
## setosa (50, 33%), versicolor (50, 33%)
## virginica (50, 33%)
## ---------------------------------------------------------...
```

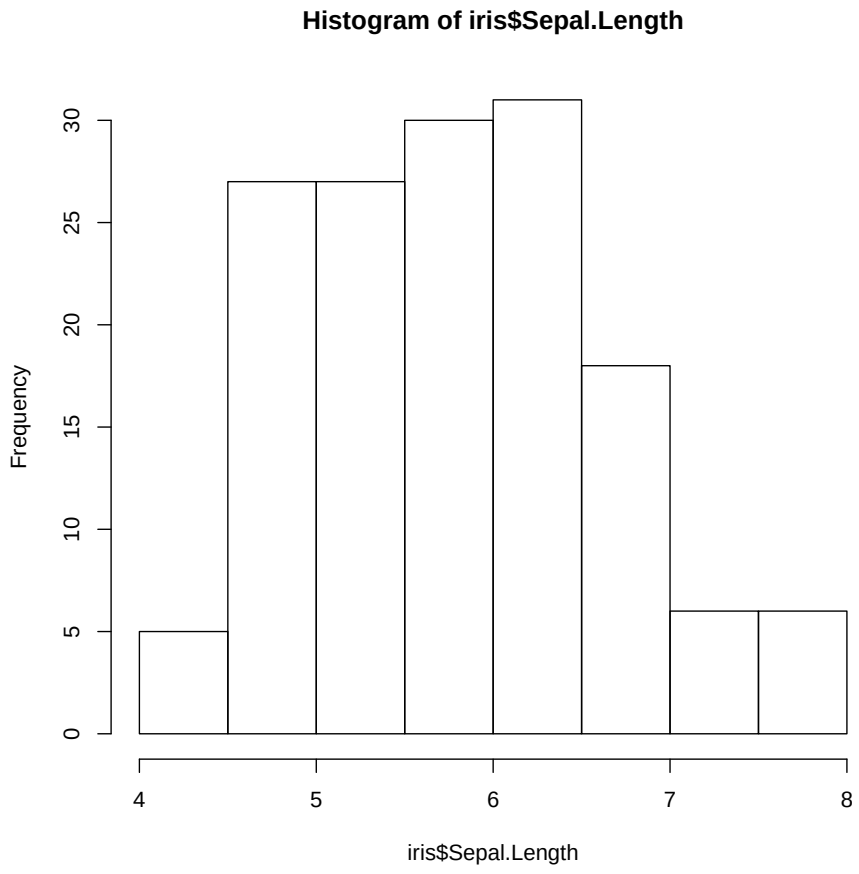
Mean, Median, Range and Quartiles

- Mean, median and range: `mean()`, `median()`, `range()`
- Quartiles and percentiles: `quantile()`

```
range(iris$Sepal.Length)

## [1] 4.3 7.9

quantile(iris$Sepal.Length)

##   0%  25%  50%  75% 100%
##  4.3  5.1  5.8  6.4  7.9

quantile(iris$Sepal.Length, c(0.1, 0.3, 0.65))

##  10%  30%  65%
## 4.80 5.27 6.20
```

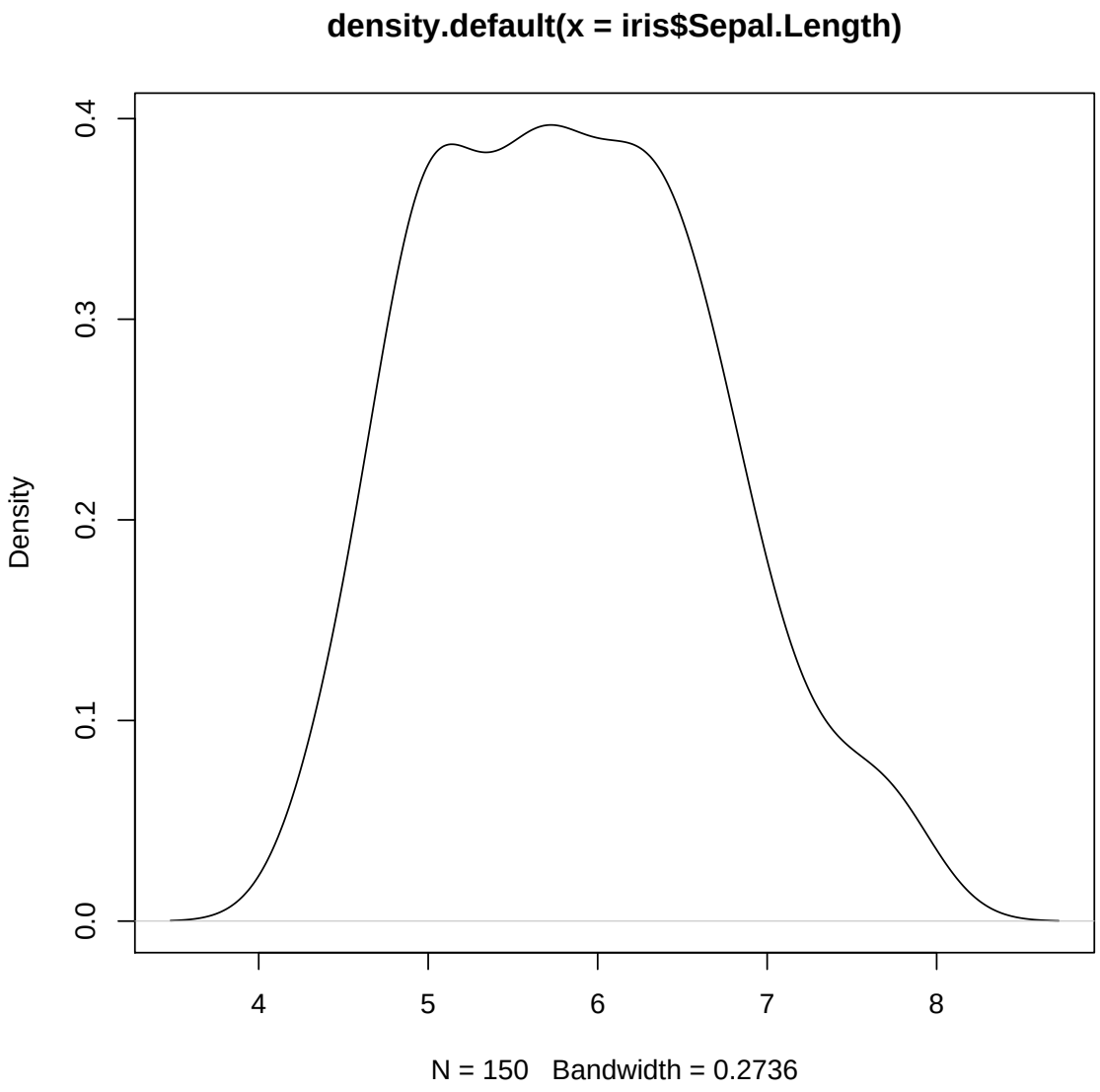
Variance and Histogram

```
var(iris$Sepal.Length)

## [1] 0.6857

hist(iris$Sepal.Length)
```



Density

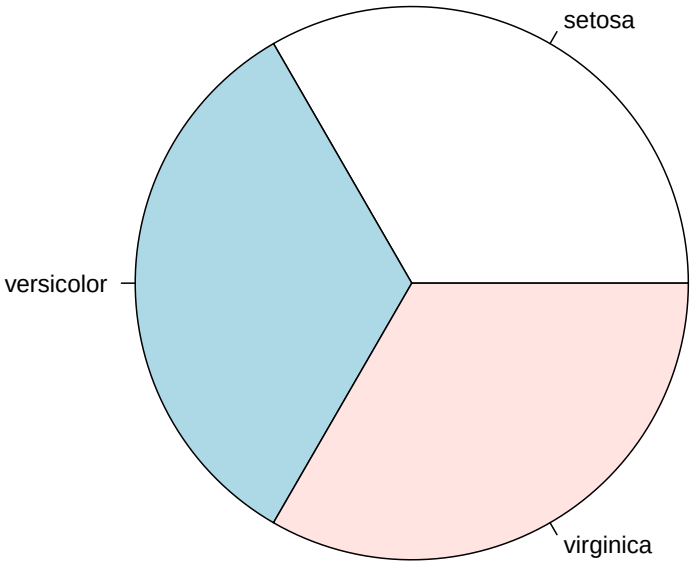
```
plot(density(iris$Sepal.Length))
```



Pie Chart

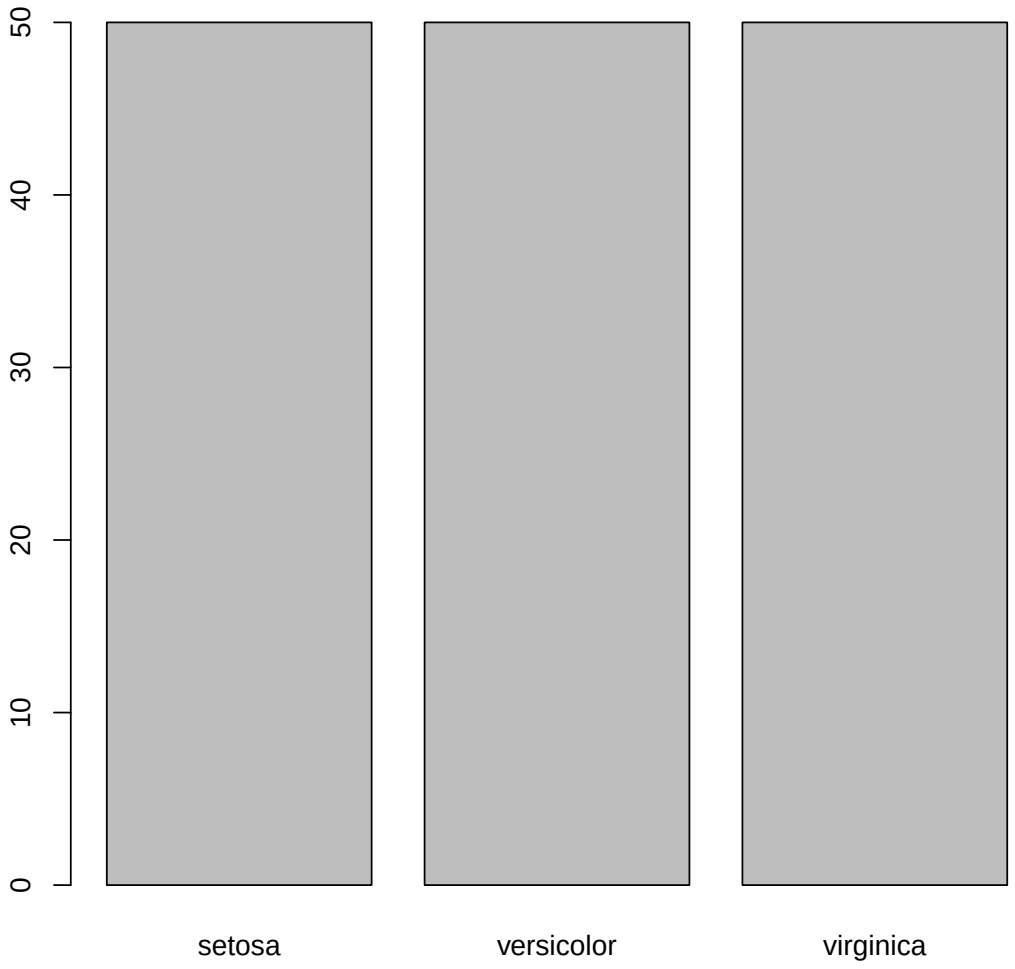
Frequency of factors: `table()`

```
table(iris$Species)

##
##     setosa versicolor  virginica
##         50         50         50

pie(table(iris$Species))
```



Bar Chart

```
barplot(table(iris$Species))
```



Outline

Correlation

Covariance and correlation: `cov()` and `cor()`

```
cov(iris$Sepal.Length, iris$Petal.Length)

## [1] 1.274

cor(iris$Sepal.Length, iris$Petal.Length)

## [1] 0.8718

cov(iris[, 1:4])

##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length      0.68569    -0.04243       1.2743      0.5163
## Sepal.Width      -0.04243     0.18998      -0.3297     -0.1216
## Petal.Length      1.27432    -0.32966       3.1163      1.2956
## Petal.Width       0.51627    -0.12164       1.2956      0.5810

# cor(iris[,1:4])
```

Aggreation

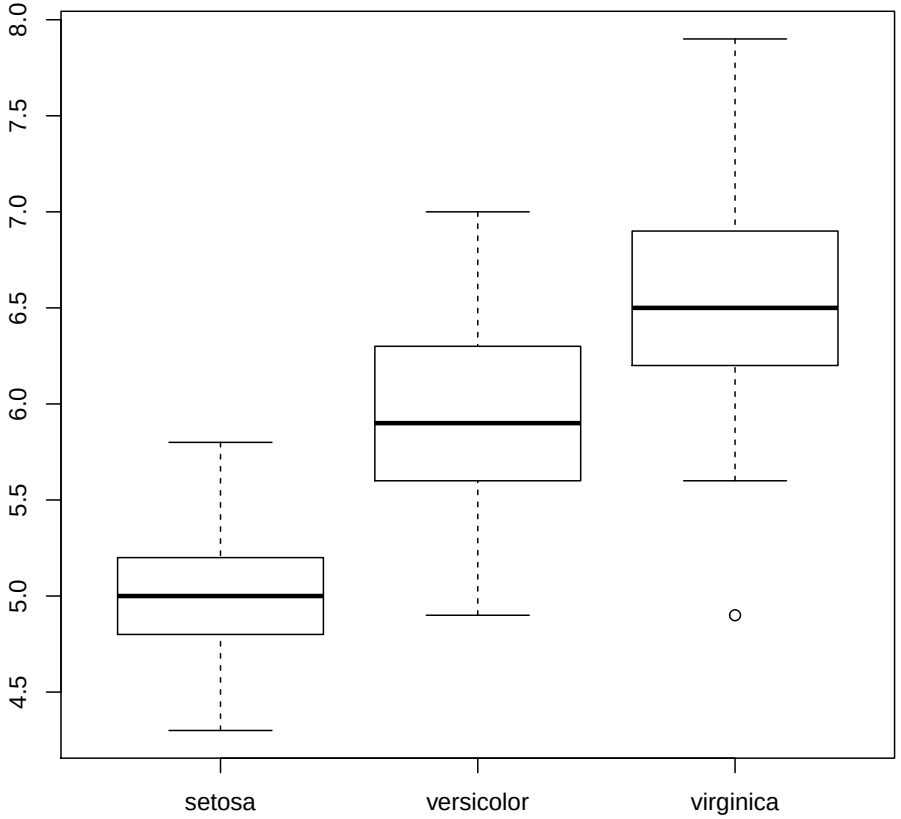
Stats of `Sepal.Length` for every `Species` with `aggregate()`

```
aggregate(Sepal.Length ~ Species, summary, data = iris)

##      Species Sepal.Length.Min. Sepal.Length.1st Qu.
## 1     setosa              4.30                 4.80
## 2 versicolor              4.90                 5.60
## 3  virginica              4.90                 6.22
##   Sepal.Length.Median Sepal.Length.Mean Sepal.Length.3rd Qu.
## 1                5.00              5.01                 5.20
## 2                5.90              5.94                 6.30
## 3                6.50              6.59                 6.90
##   Sepal.Length.Max.
## 1              5.80
## 2              7.00
## 3              7.90
```

Boxplot

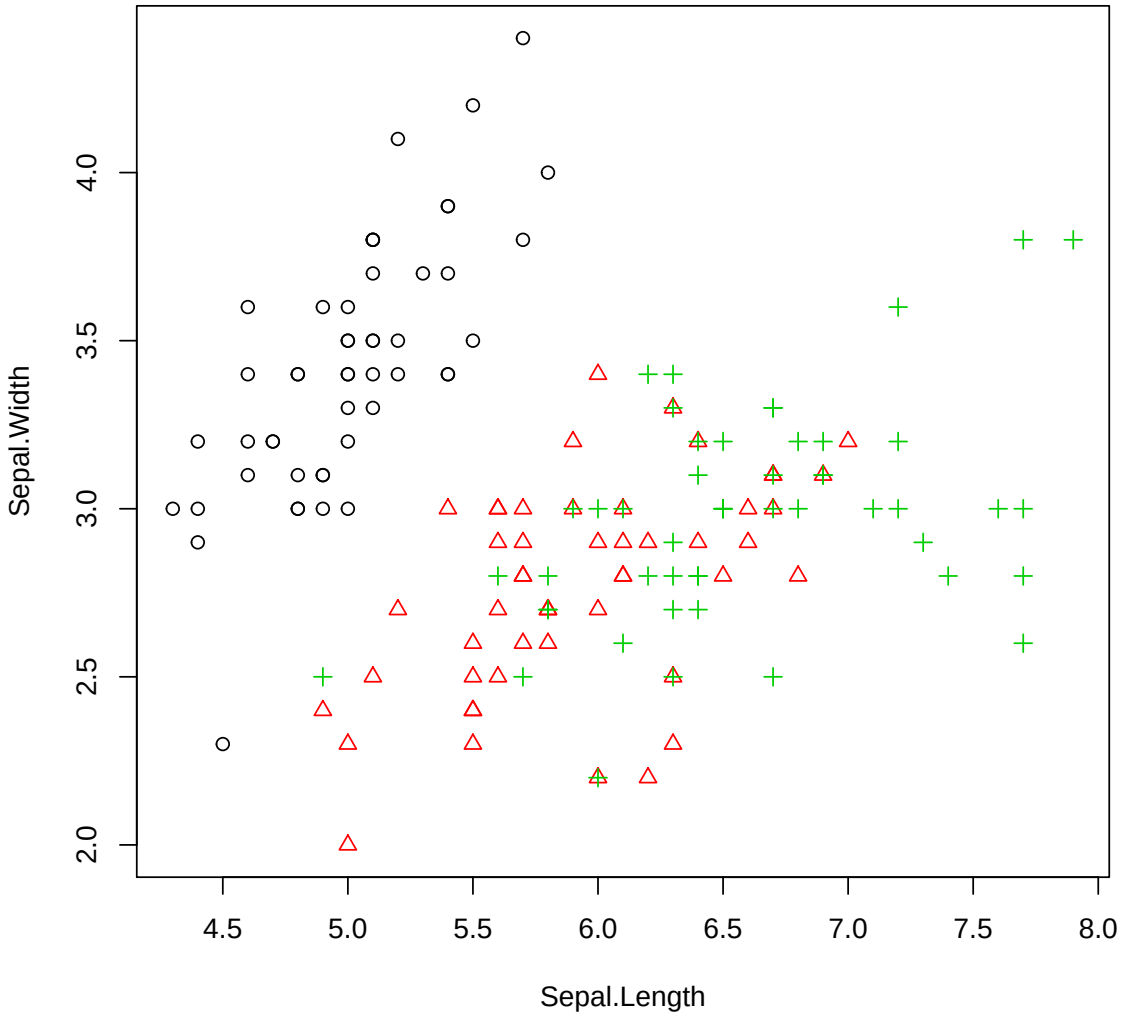
- The bar in the middle is median.
- The box shows the interquartile range (IQR), i.e., range between the 75% and 25% observation.

```
boxplot(Sepal.Length ~ Species, data = iris)
```



Scatter Plot

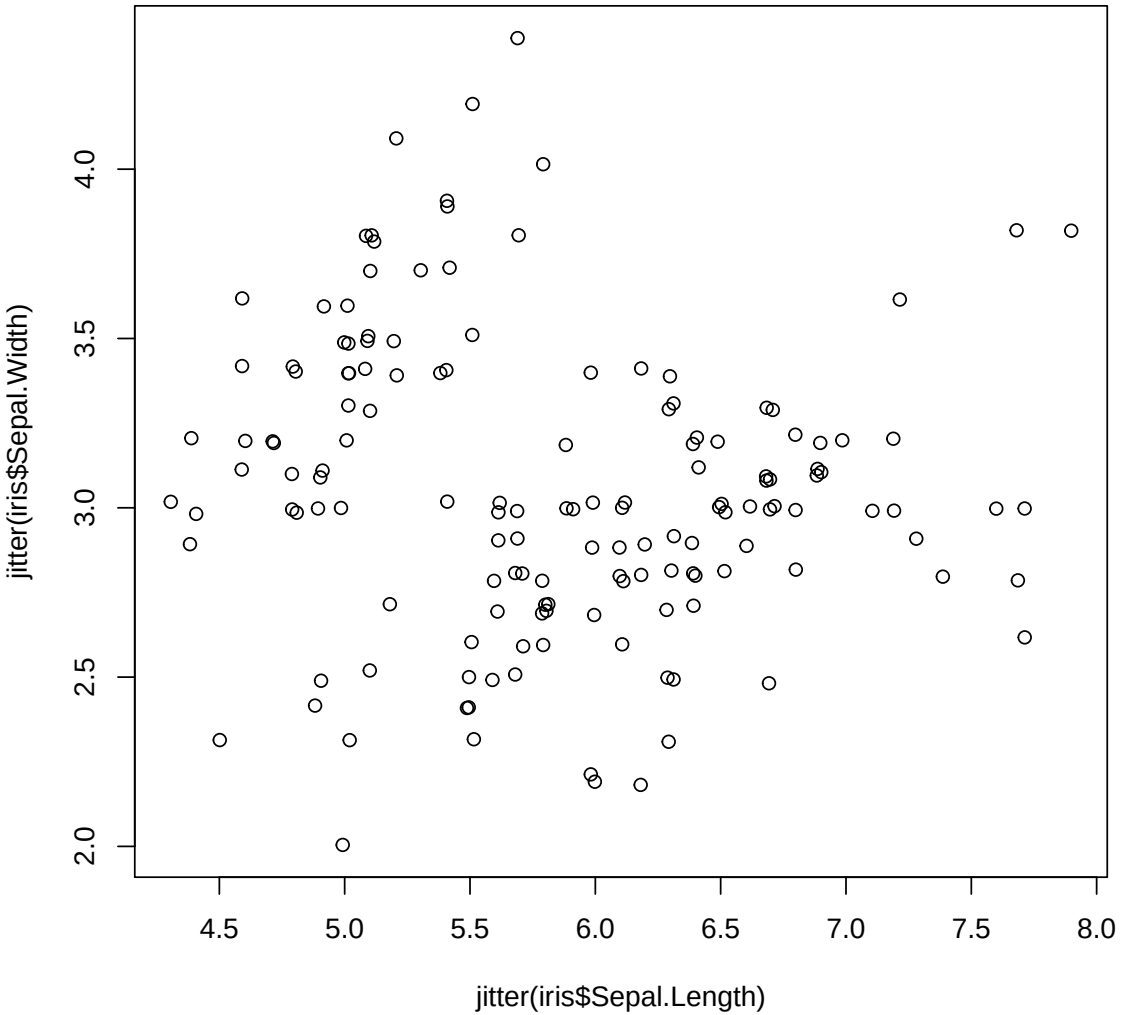
```
with(iris, plot(Sepal.Length, Sepal.Width, col = Species,
                pch = as.numeric(Species)))
```



Scatter Plot with Jitter

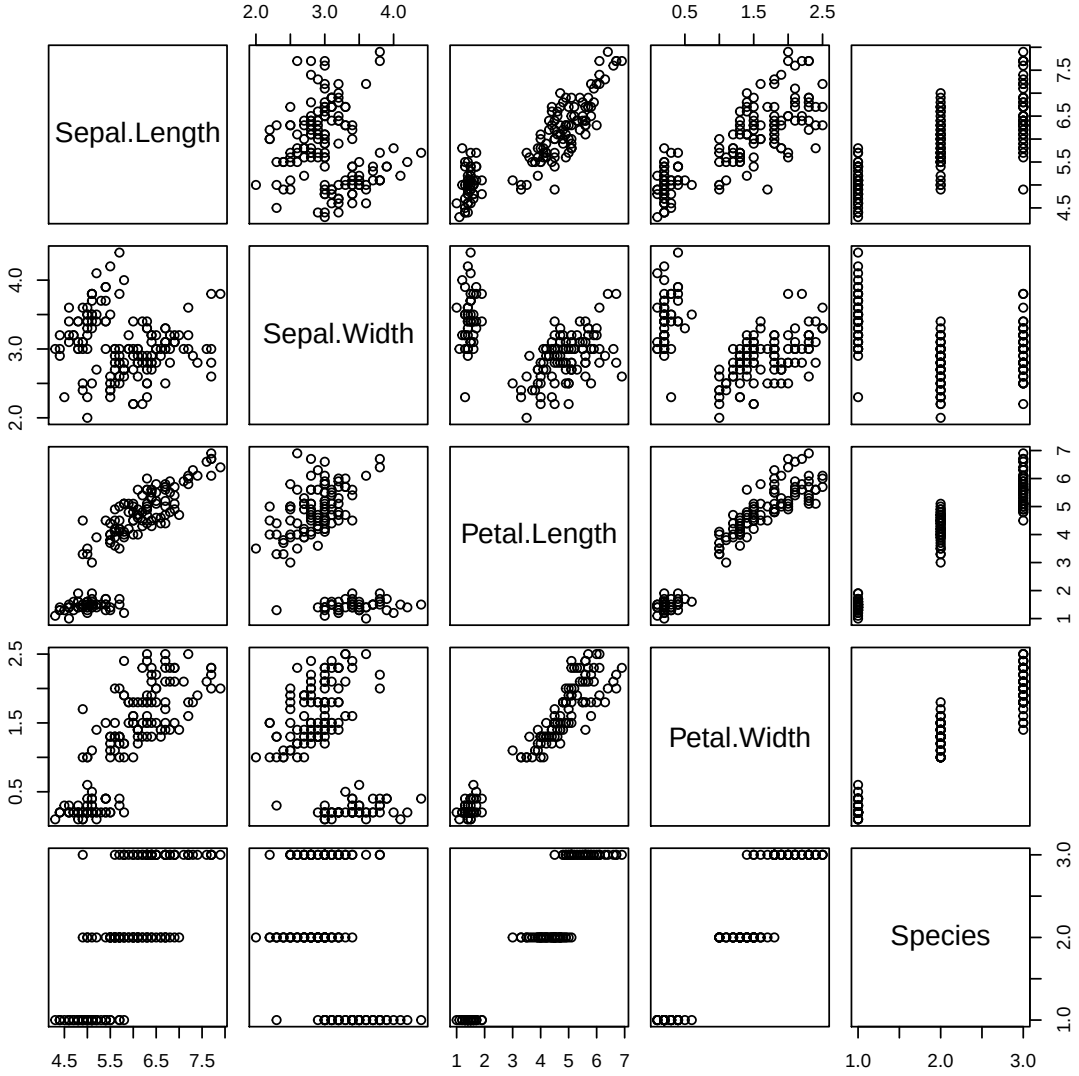
Function `jitter()`: add a small amount of noise to the data

```
plot(jitter(iris$Sepal.Length), jitter(iris$Sepal.Width))
```



A Matrix of Scatter Plots

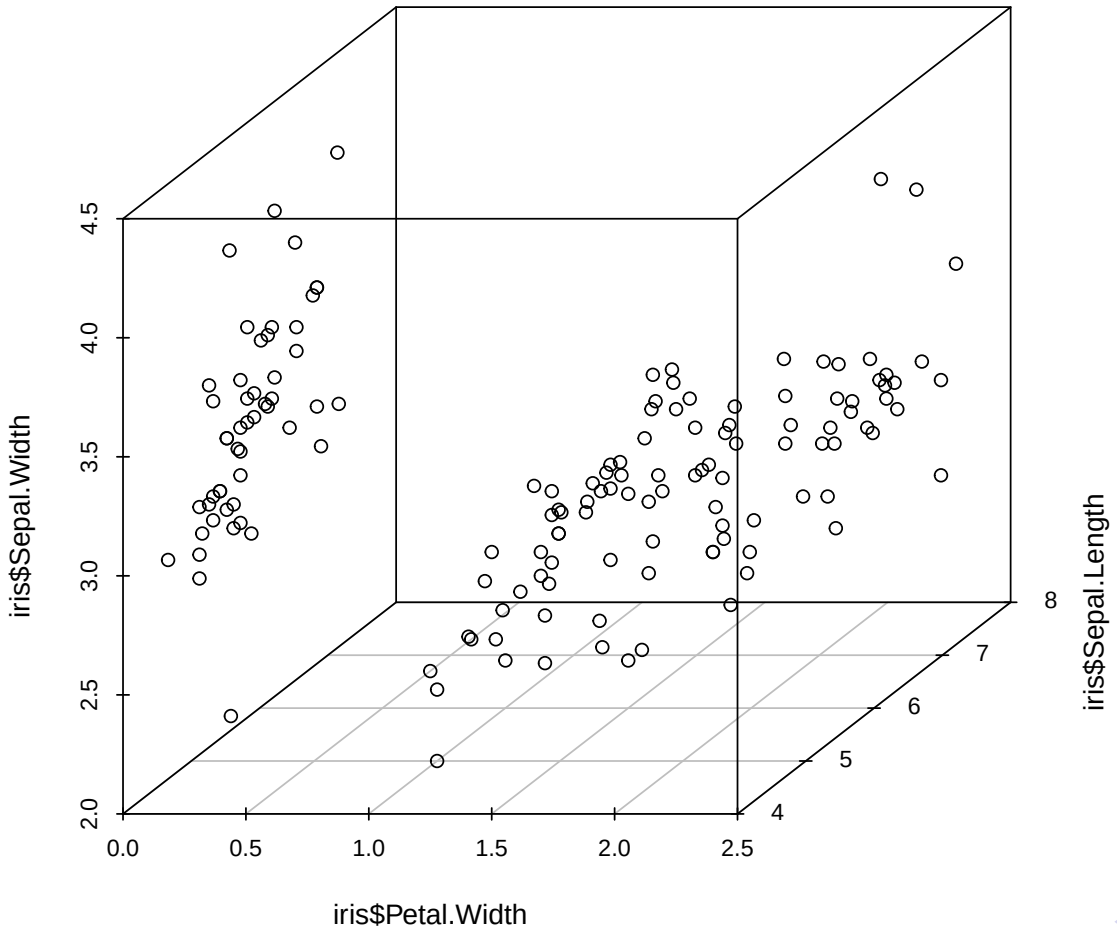
```r
pairs(iris)
```



Outline

3D Scatter plot

```
library(scatterplot3d)
scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)
```



Interactive 3D Scatter Plot

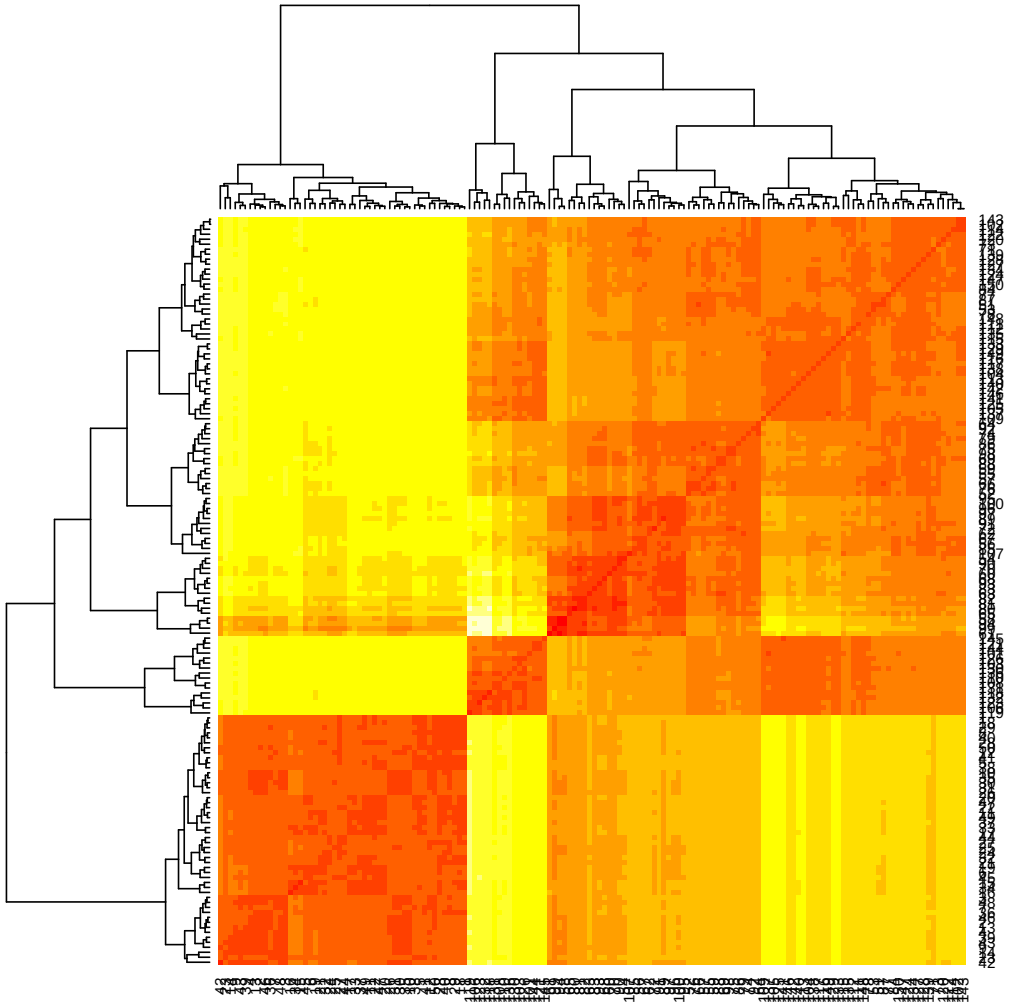
Package *rgl* supports interactive 3D scatter plot with `plot3d()`.

```
library(rgl)
plot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)
```

Heat Map

Calculate the similarity between different flowers in the `iris` data with `dist()` and then plot it with a heat map

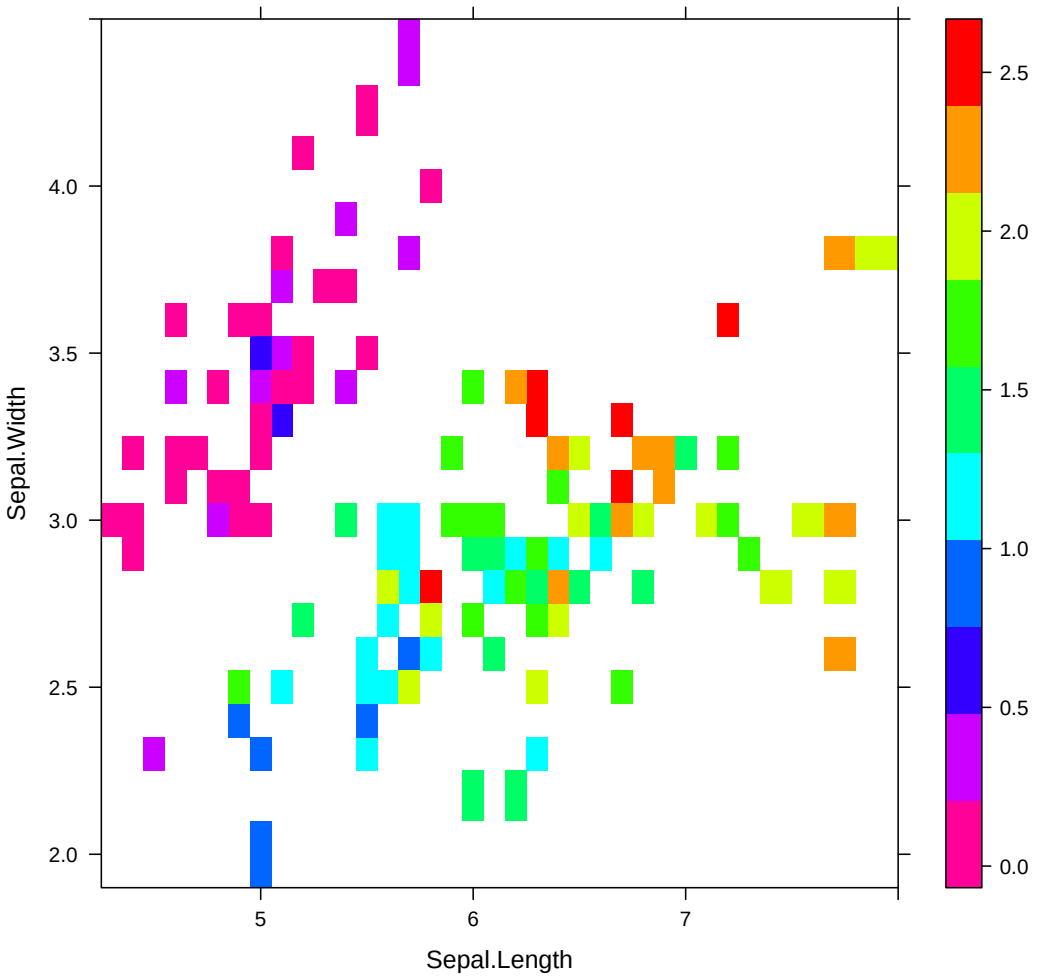
```
dist.matrix <- as.matrix(dist(iris[, 1:4]))
heatmap(dist.matrix)
```



Level Plot

Function `rainbow()` creates a vector of contiguous colors.

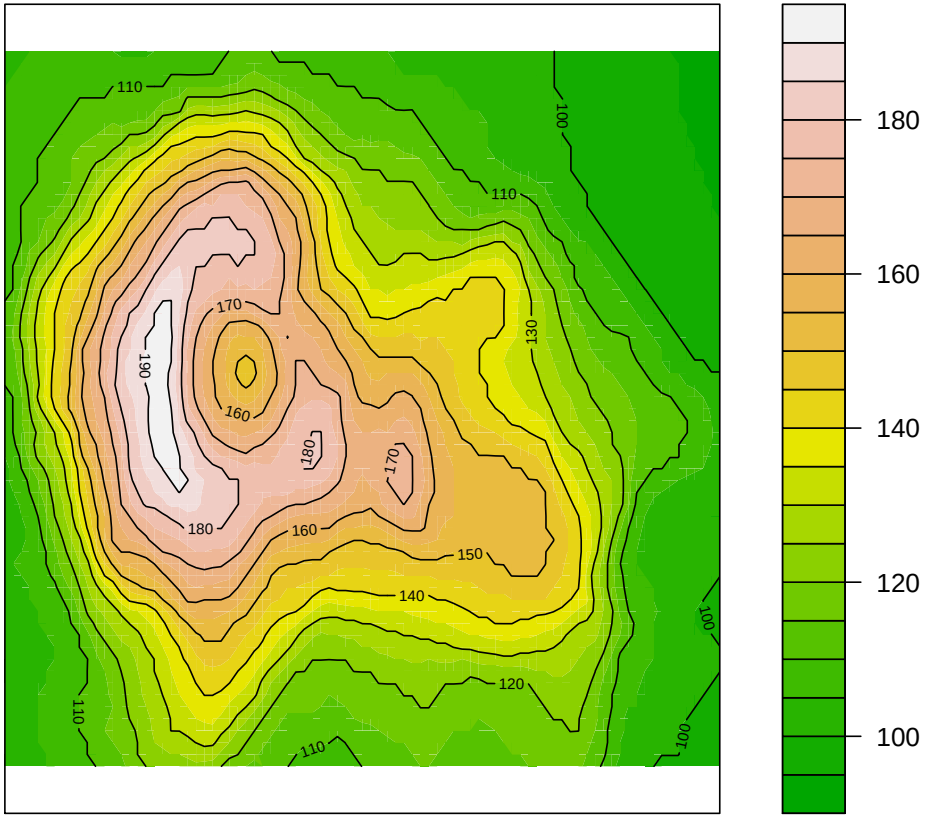
```
library(lattice)
levelplot(Petal.Width ~ Sepal.Length * Sepal.Width, iris, cuts = 9,
    col.regions = rainbow(10)[10:1])
```



Contour

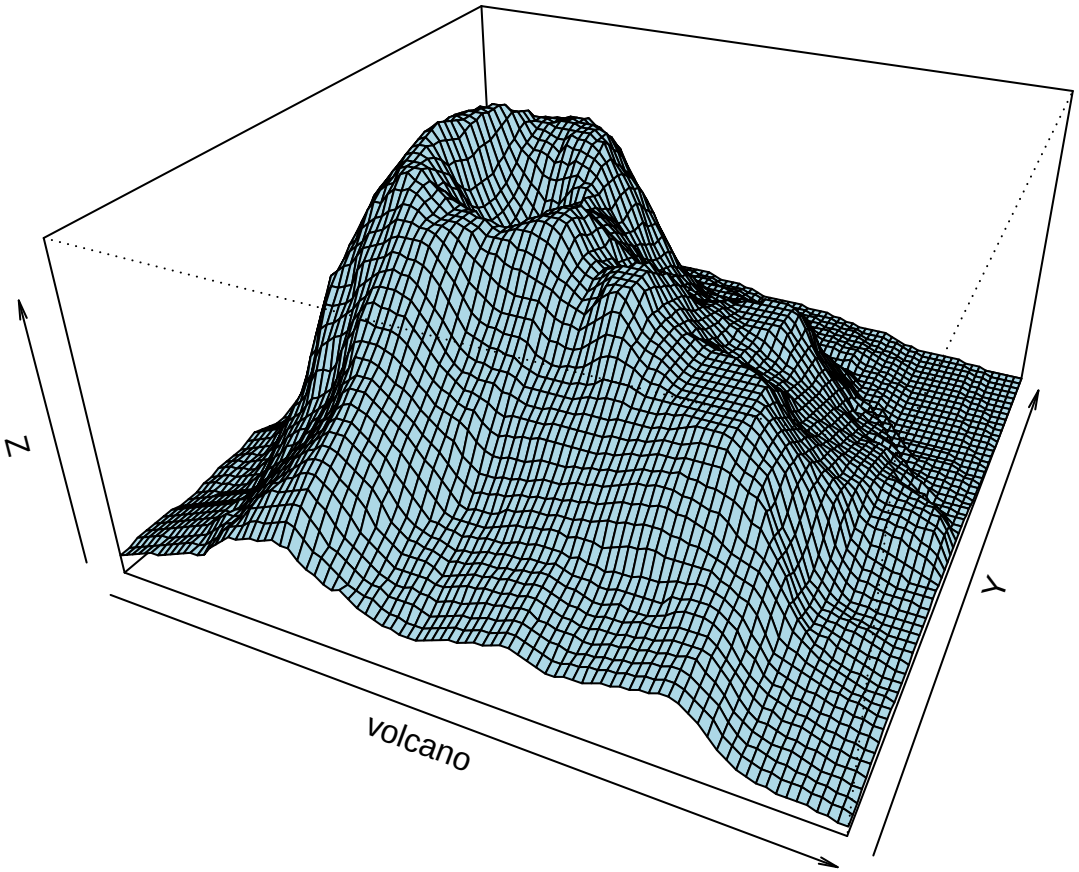
`contour()` and `filled.contour()` in package *graphics*
`contourplot()` in package *lattice*

```
filled.contour(volcano, color = terrain.colors, asp = 1, plot.axes = co
    add = T))
```



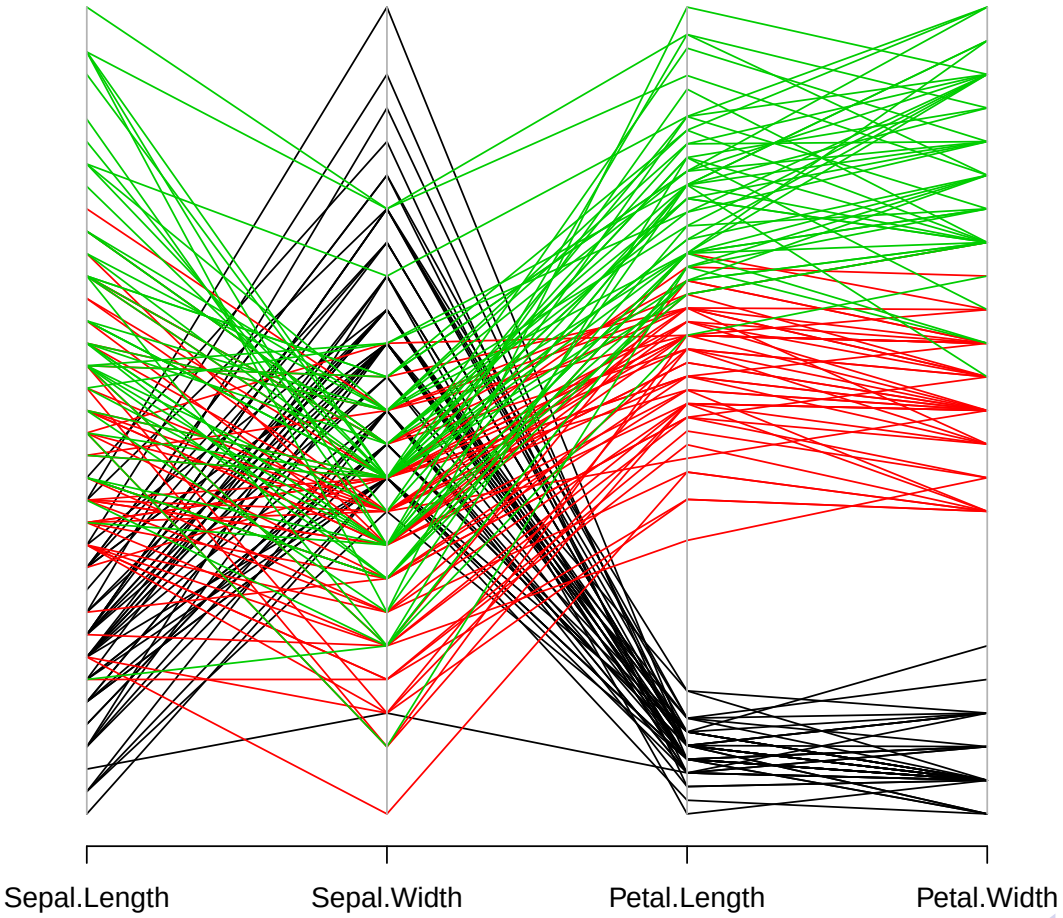
3D Surface

```
persp(volcano, theta = 25, phi = 30, expand = 0.5, col = "lightblue")
```



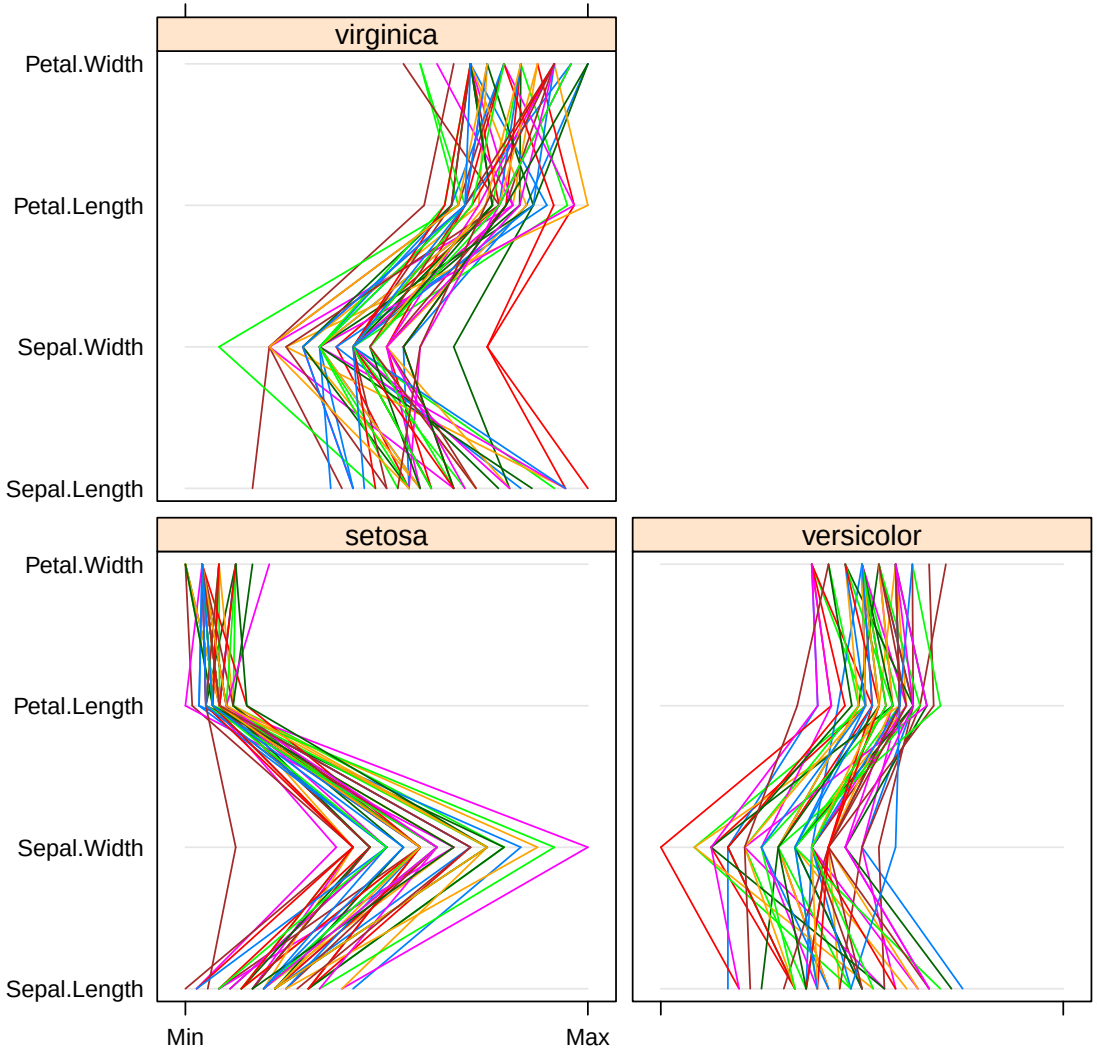
Parallel Coordinates

```
library(MASS)
parcoord(iris[1:4], col = iris$Species)
```



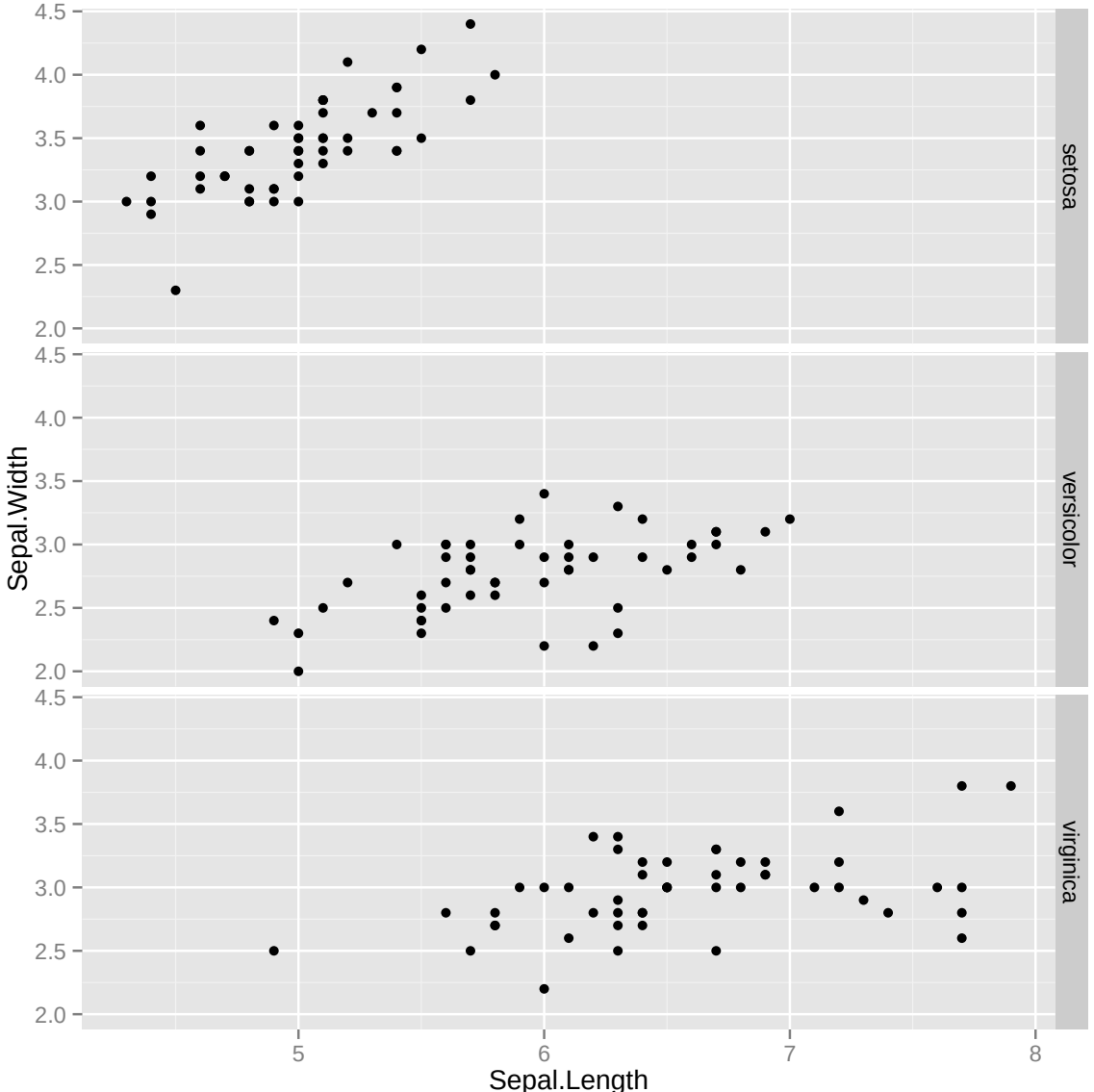
Parallel Coordinates with Package *lattice*

```
library(lattice)
parallelplot(~iris[1:4] | Species, data = iris)
```



Visualization with Package *ggplot2*

```
library(ggplot2)
qplot(Sepal.Length, Sepal.Width, data = iris, facets = Species ~ .)
```



Outline

Save Charts to Files

- Save charts to PDF and PS files: `pdf()` and `postscript()`
- BMP, JPEG, PNG and TIFF files: `bmp()`, `jpeg()`, `png()` and `tiff()`
- Close files (or graphics devices) with `graphics.off()` or `dev.off()` after plotting

```
# save as a PDF file
pdf("myPlot.pdf")
x <- 1:50
plot(x, log(x))
graphics.off()
# Save as a postscript file
postscript("myPlot2.ps")
x <- -20:20
plot(x, x^2)
graphics.off()
```

Outline

Further Readings

- Examples of ggplot2 plotting:
  `http://had.co.nz/ggplot2/`
- Package *iplots*: interactive scatter plot, histogram, bar plot, and parallel coordinates plot (iplots)
  `http://stats.math.uni-augsburg.de/iplots/`
- Package *googleVis*: interactive charts with the Google Visualisation API
  `http://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html`
- Package *ggvis*: interactive grammar of graphics
  `http://ggvis.rstudio.com/`
- Package *rCharts*: interactive javascript visualizations from R
  `http://rcharts.io/`

Online Resources

- Chapter 3: Data Exploration, in book *R and Data Mining: Examples and Case Studies*

  `http://www.rdatamining.com/docs/RDataMining.pdf`
- R Reference Card for Data Mining

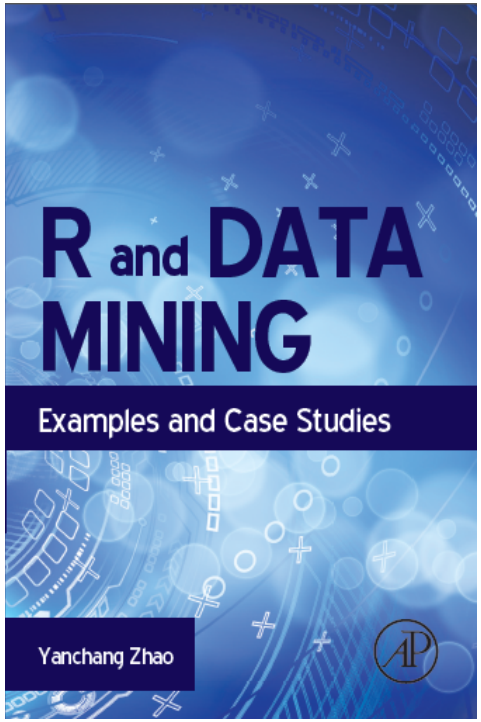
  `http://www.rdatamining.com/docs/R-refcard-data-mining.pdf`
- Free online courses and documents

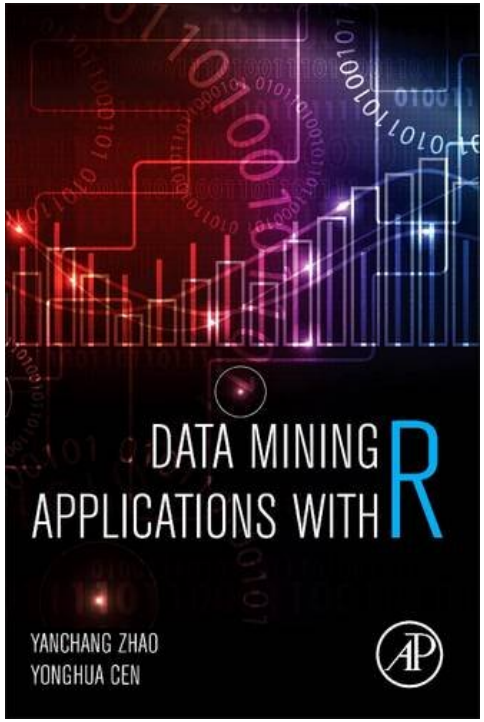
  `http://www.rdatamining.com/resources/`
- RDataMining Group on LinkedIn (7,000+ members)

  `http://group.rdatamining.com`
- RDataMining on Twitter (1,700+ followers)

  `@RDataMining`

The End





Thanks!

Email: yanchang(at)rdatamining.com