The Data Co.™

“Data Mining with Rattle and R”

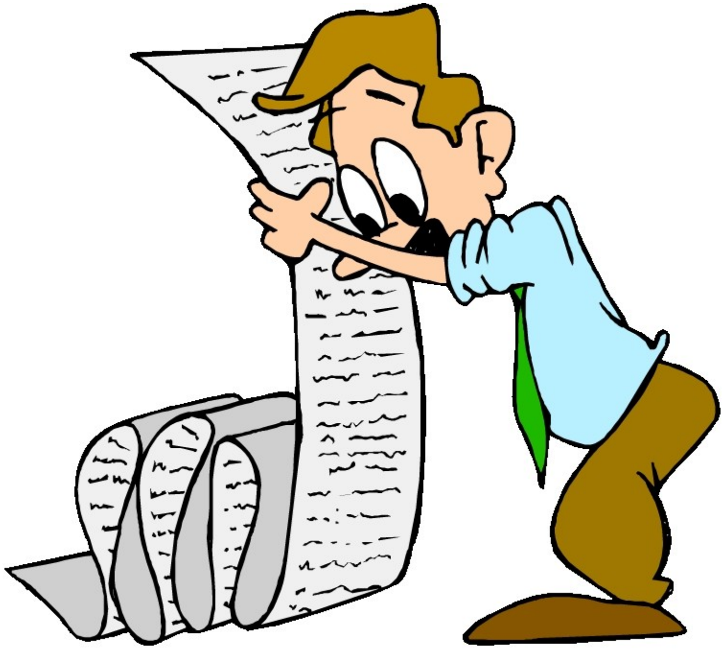
DIT Analytics Club
07 March, 2013
“Peadar Kearney's”, Dame Street

Colman McMahon
colman@thedata.co

LinkedIn
www.linkedin.com/in/colmanmcmahon

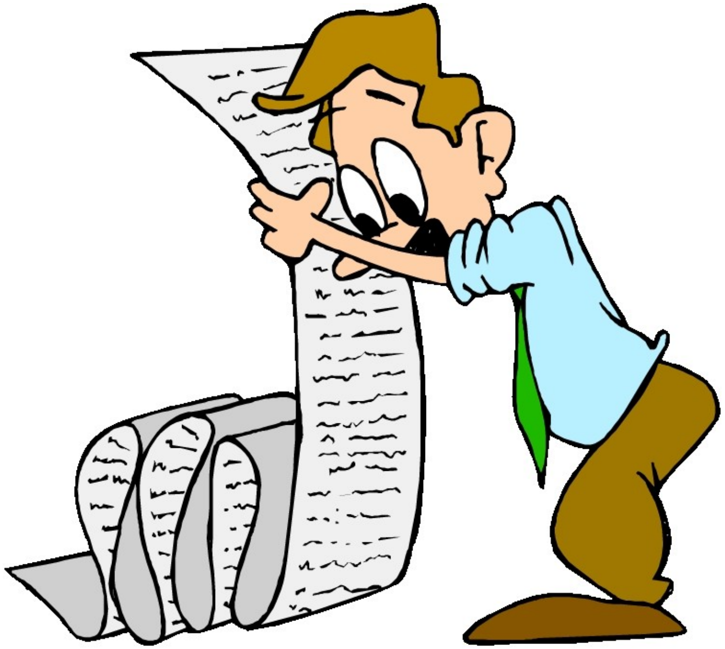
Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix



Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix

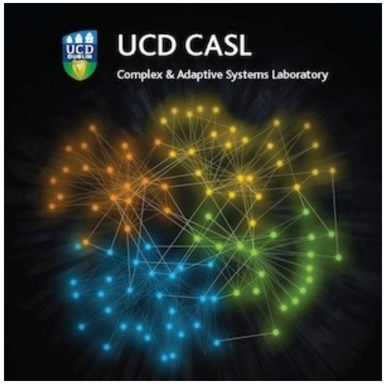


Bio

- Professional
  - Background: Film, VFX, Digital Media
  - Currently: PhD Fellow, UCD Dynamics Lab
- PhD:
  - Policy network analysis
  - Creative Industries
- Technologies:
  - Social Network Analysis
  - Agent Based Modelling
  - Data/statistical analysis

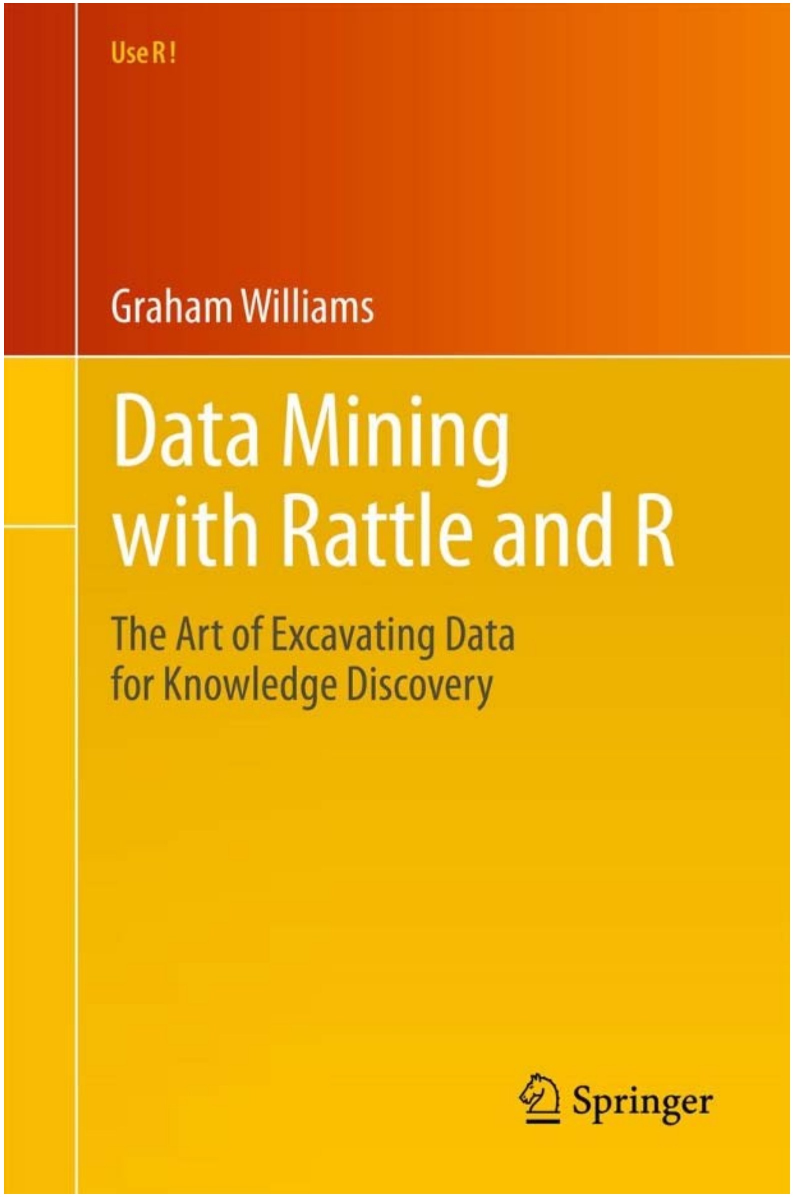
Dynamics Lab
Computational Social Science
University College Dublin - Geary Institute - Ireland

www.dl.ucd.ie



www.ucd.ie/casl/

- Data Mining with Rattle and R
  - The Art of Excavating Data for Knowledge Discovery
- Graham Williams
- Springer, 2011
- ISBN 978-1-4419-9889-7


Use R!
Graham Williams
Data Mining with Rattle and R
The Art of Excavating Data for Knowledge Discovery
Springer

Rattle

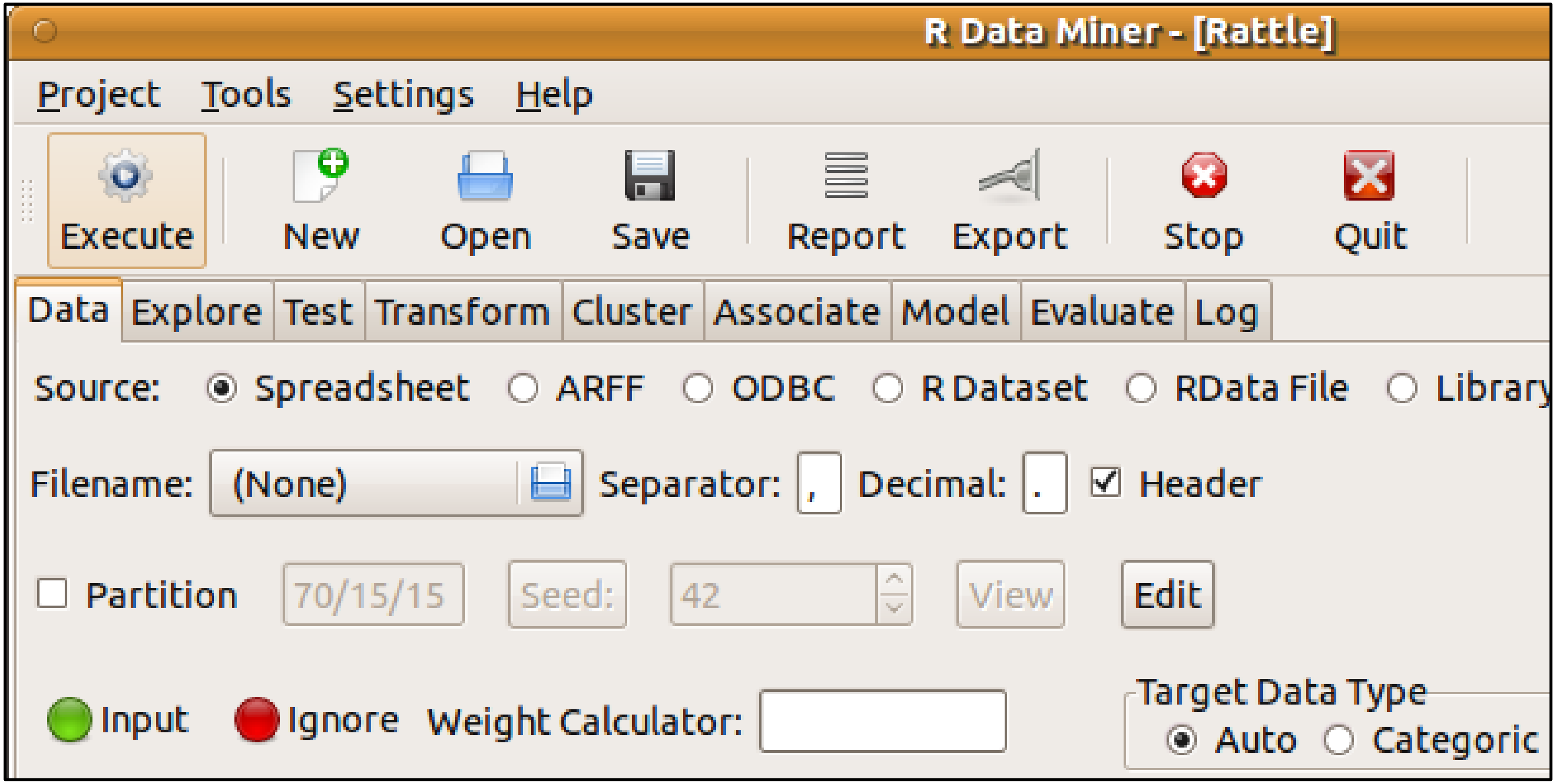
Rattle - “the R Analytical Tool To Learn Easily”

- Presents statistical & visual summaries of data
- Transforms data into forms that can be readily modelled
- Builds models (unsupervised and supervised) from the data
- Graphically presents the performance of models
- Scores new datasets

Rattle

- Built on the statistical language R
  - an understanding of R is not required in order to use it
- Simple to use, quick to deploy, and allows us to rapidly work through the data processing, modelling, and evaluation phases of a data mining project
- Can migrate from Rattle to R when we need to fine-tune and further develop our data mining projects

Rattle GUI



R Data Miner - [Rattle]
Project Tools Settings Help
Execute
New
Open
Save
Report
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library

Filename: (None) Separator: , Decimal: . ☑ Header

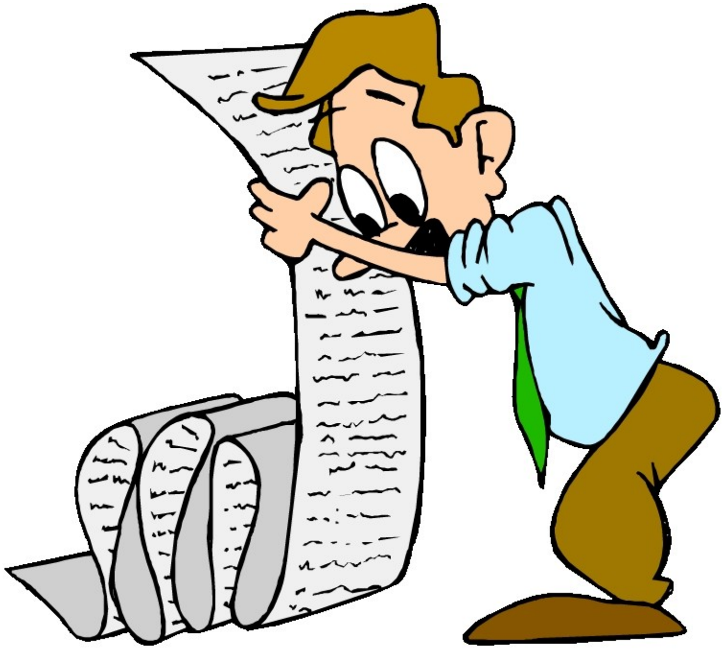
□ Partition 70/15/15 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric

Agenda

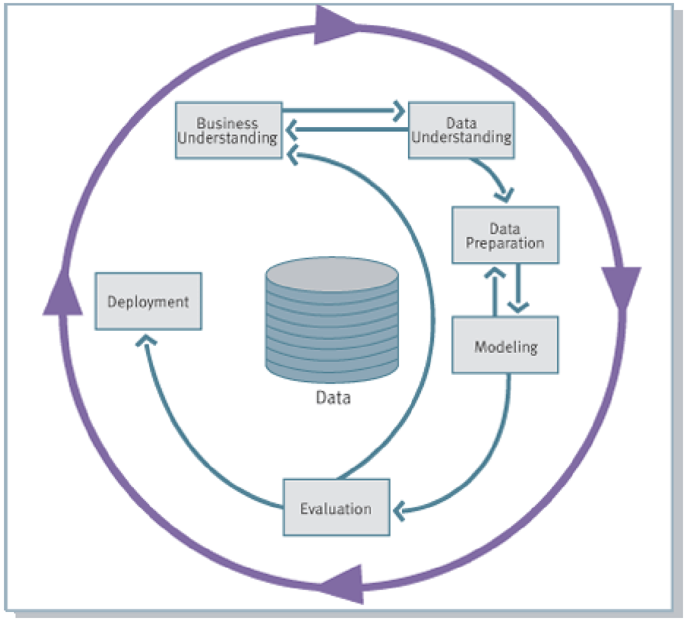
- Intro
- Data mining
- Rattle stack
- Rattle workflow
- Appendix



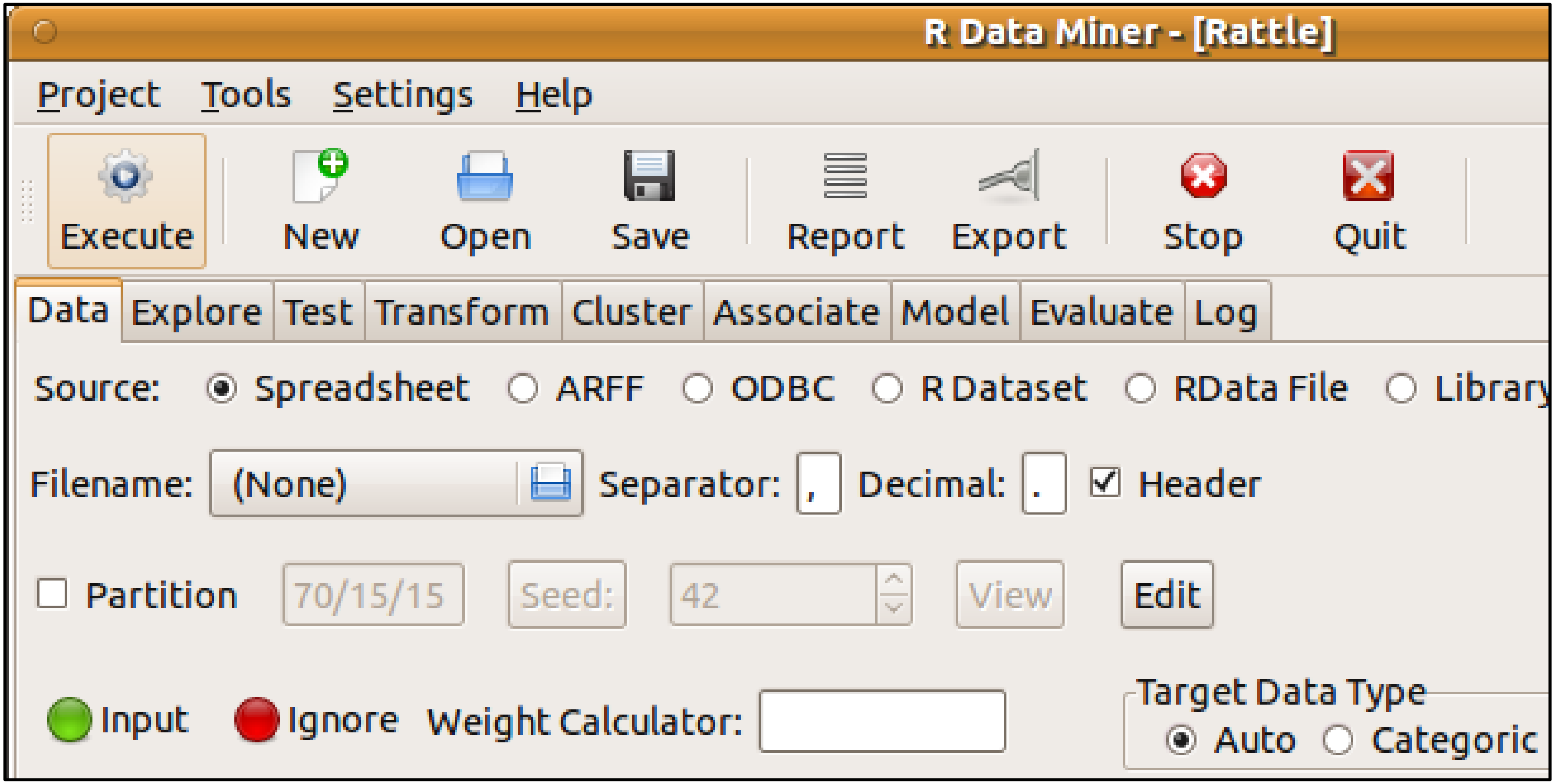
CRISP-DM

- Cross Industry Process for Data Mining (CRISP-DM, 1996)
  - » framework for delivering data mining projects.

1) Problem Understanding
2) Data Understanding
3) Data Preparation
4) Modelling
5) Evaluation
6) Deployment



Rattle GUI



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library

Filename: (None) Separator: , Decimal: . ☑ Header

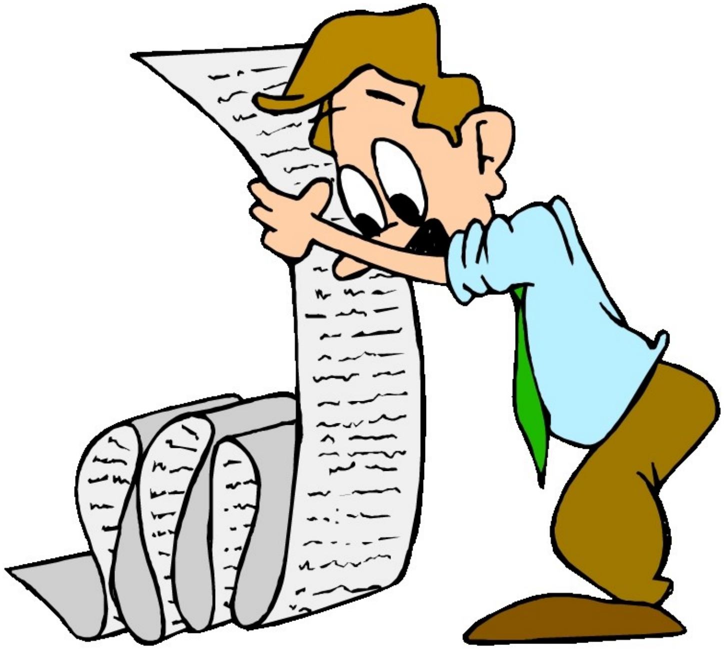
☐ Partition 70/15/15 Seed: 42

Weight Calculator:

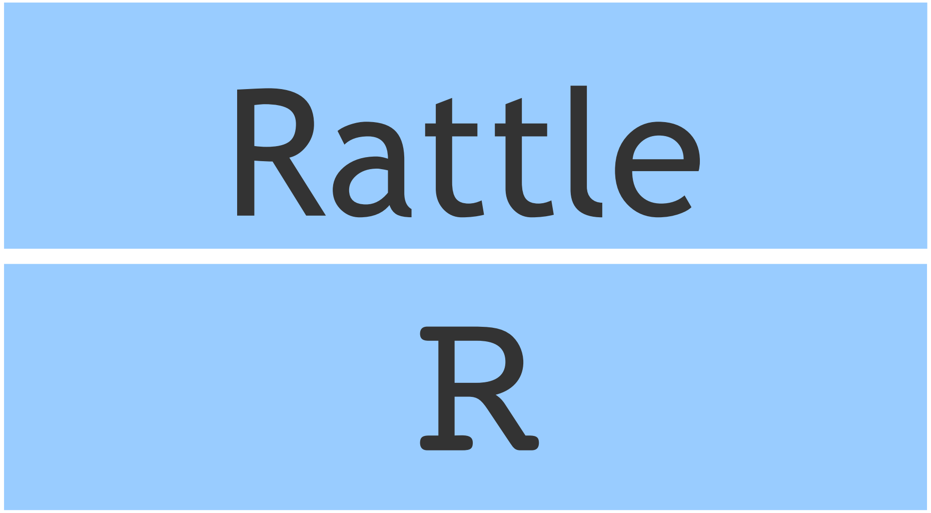
Target Data Type
● Auto ○ Categoric

Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix



Rattle stack



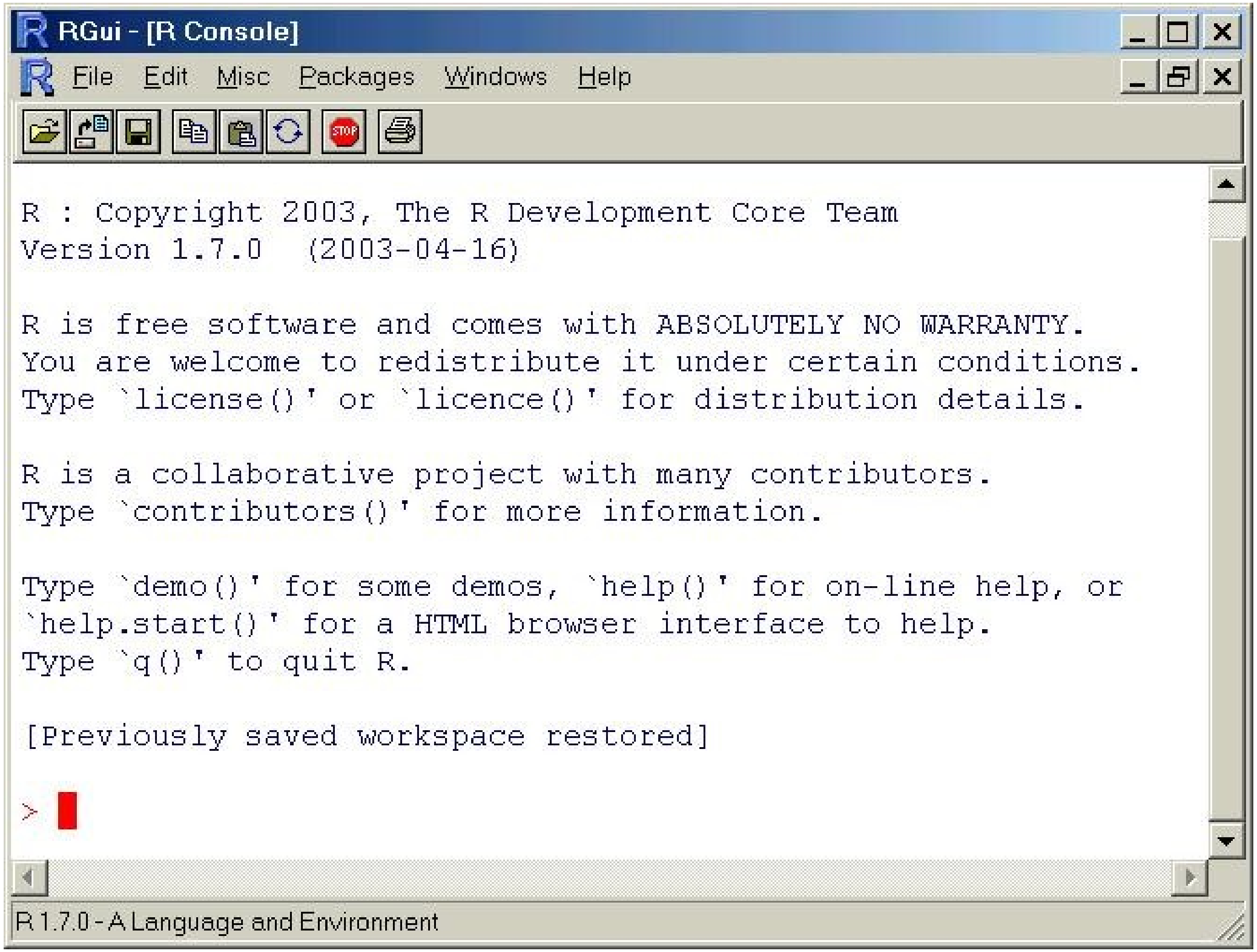
Rattle

R

R (www.r-project.org)

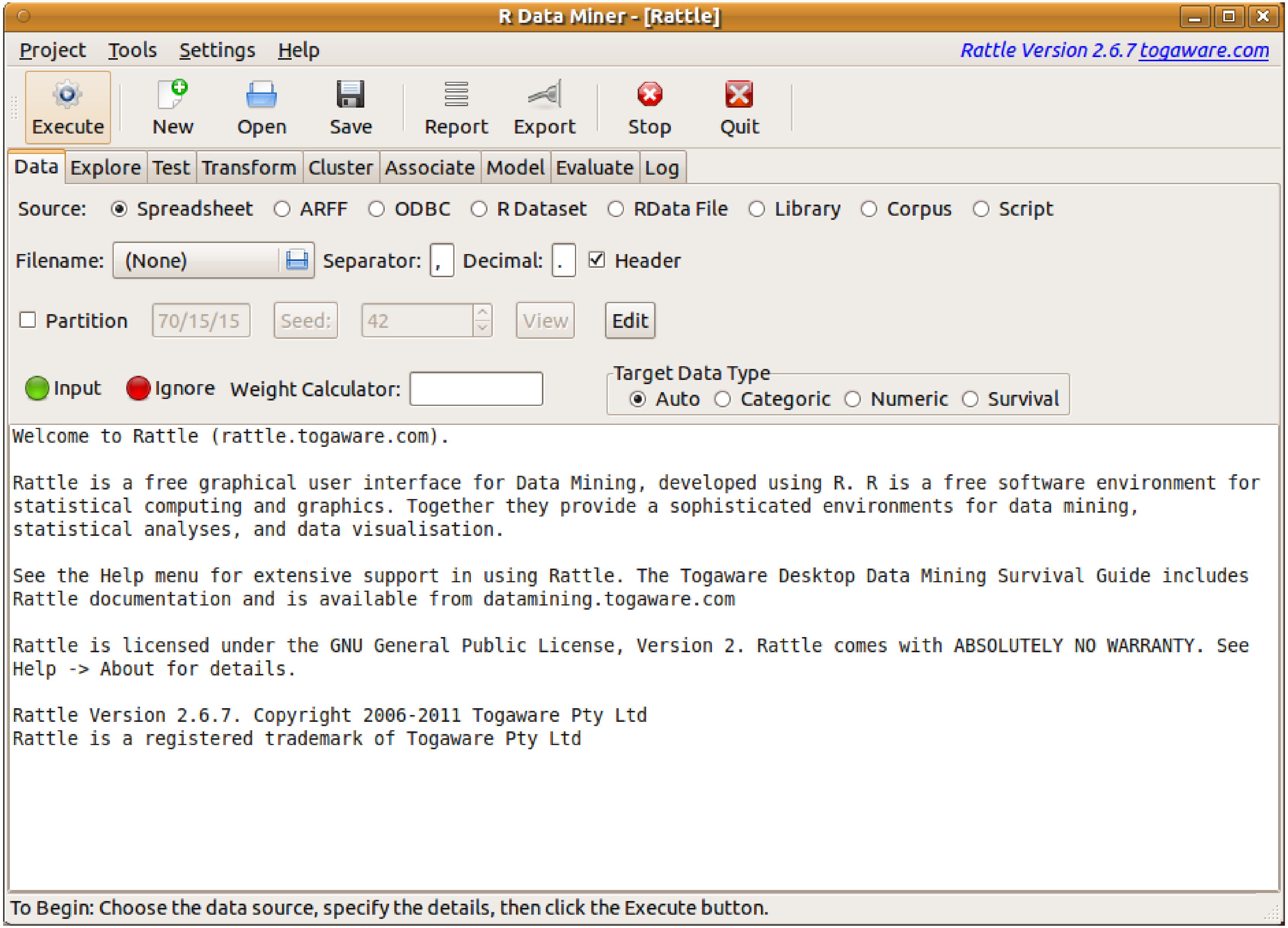
- R - a sophisticated statistical software package
  - easily installed, instructional, state-of-the-art, and it is free and open source
- The basic *modus operandi* - write scripts using the R language
- Steeper learning curve than using GUI based systems, but once over the hurdle, becomes relatively easy

R Project (www.r-project.org)



```
R : Copyright 2003, The R Development Core Team
Version 1.7.0  (2003-04-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.

R is a collaborative project with many contributors.
Type `contributors()' for more information.

Type `demo()' for some demos, `help()' for on-line help, or
`help.start()' for a HTML browser interface to help.
Type `q()' to quit R.

[Previously saved workspace restored]

>
```

Rattle (rattle.togaware.com)


R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 2.6.7 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for
statistical computing and graphics. Together they provide a sophisticated environments for data mining,
statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes
Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See
Help -> About for details.

Rattle Version 2.6.7. Copyright 2006-2011 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
```

To Begin: Choose the data source, specify the details, then click the Execute button.

Installation

Install R

```
www.r-project.org
```

```
Start R
```

Install Rattle

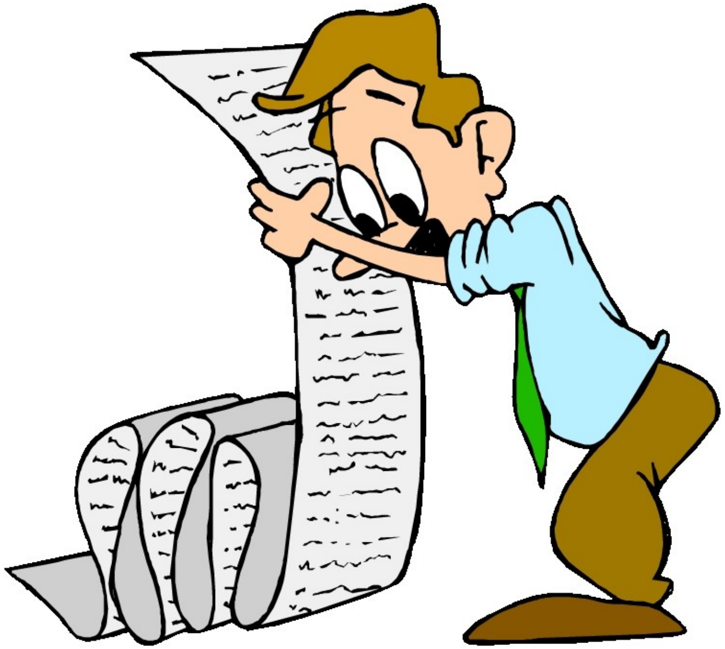
```
> install.packages("rattle")
```

Load rattle into the R library

```
> library(rattle)
> rattle()
```

Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix



Rattle workflow

1) Load a dataset

2) Select variables and entities for exploring and mining

3) Explore the data to understand how it is distributed or spread

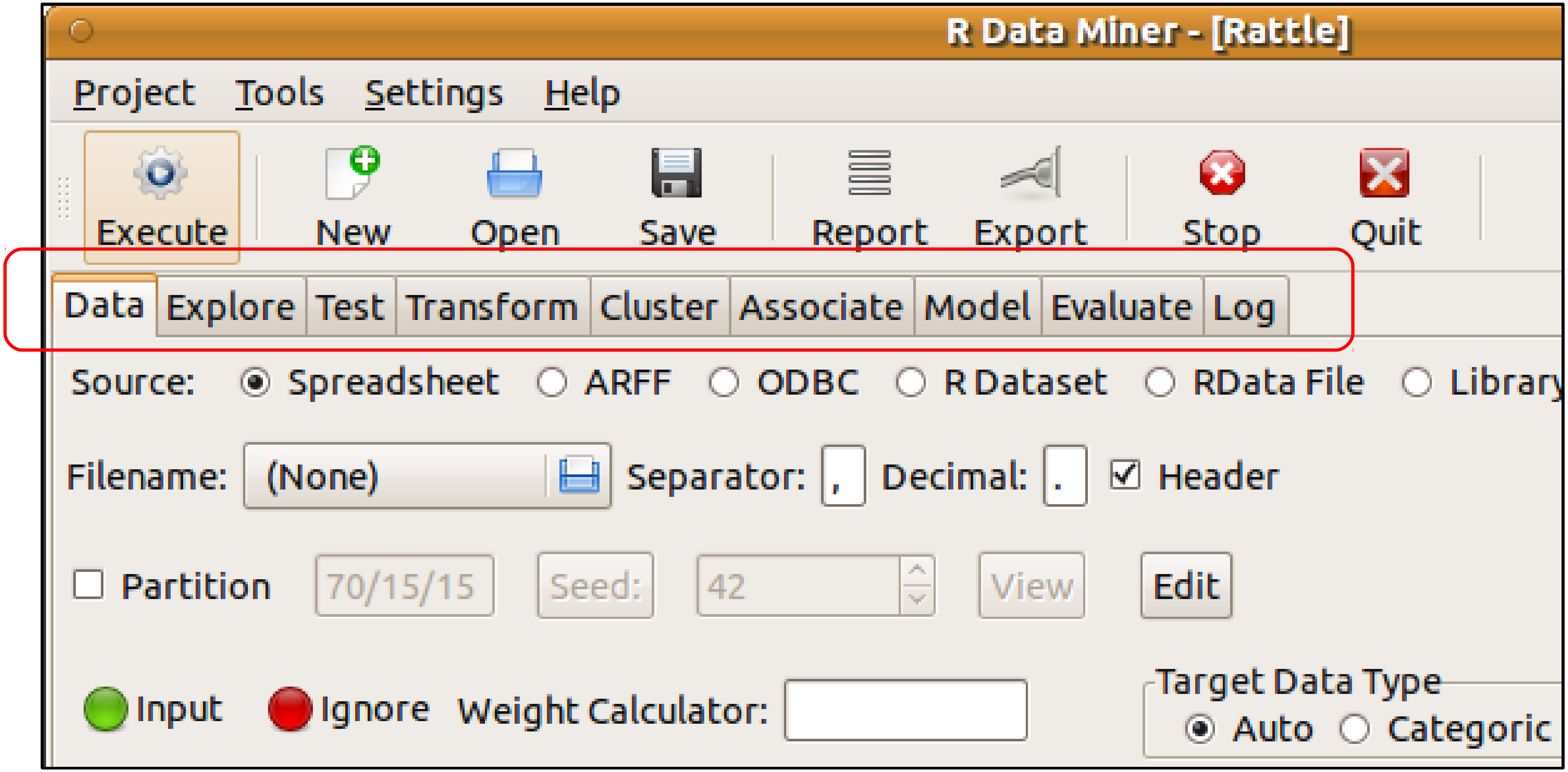
4) Transform the data to suit our data mining purposes

5) Build our models

6) Evaluate the models on other datasets

7) Export the models for deployment

Workflow - tabs



R Data Miner - [Rattle]
Project Tools Settings Help
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library

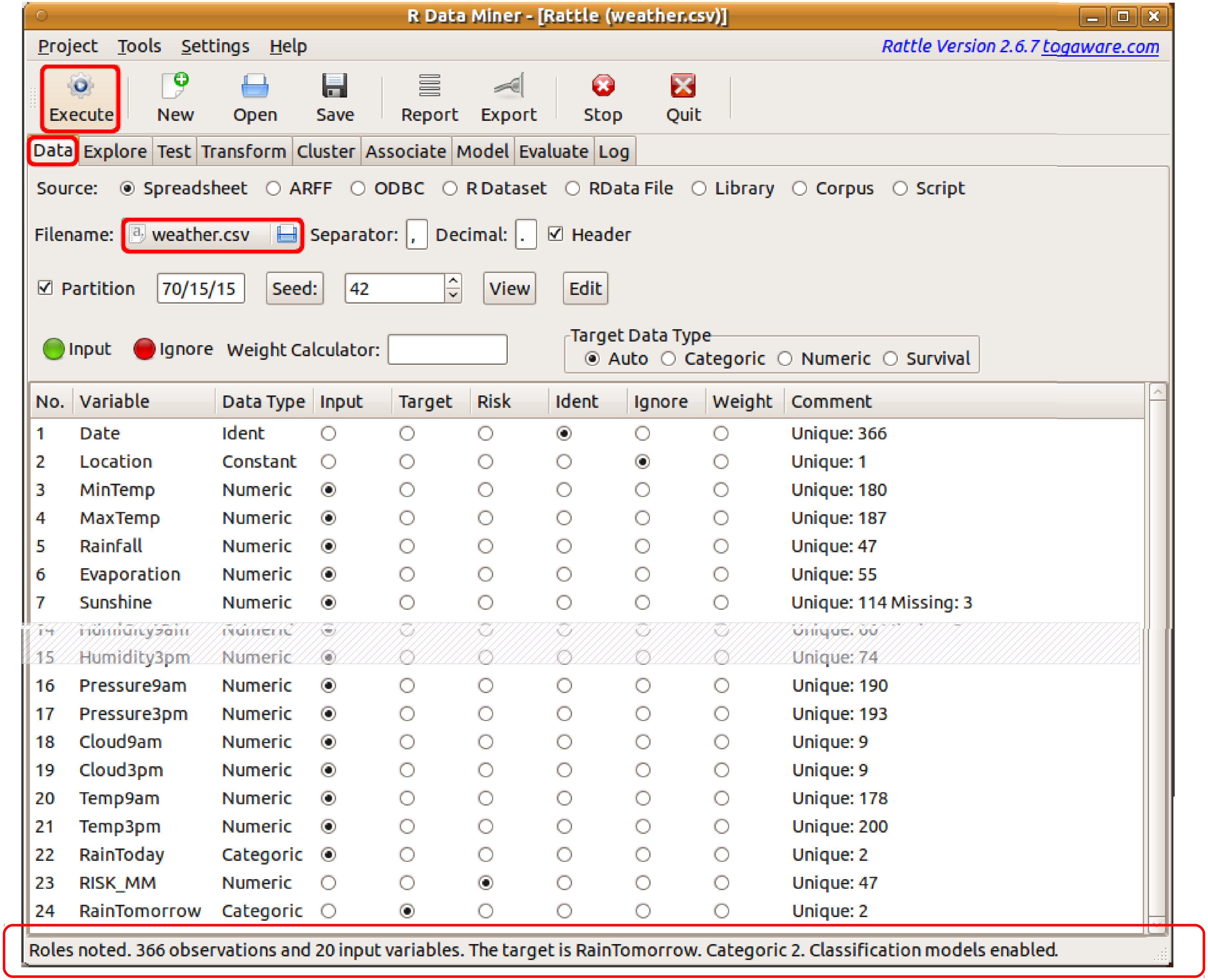
Filename: (None) Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type
● Auto ○ Categoric

The Data Co.



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: weather.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42

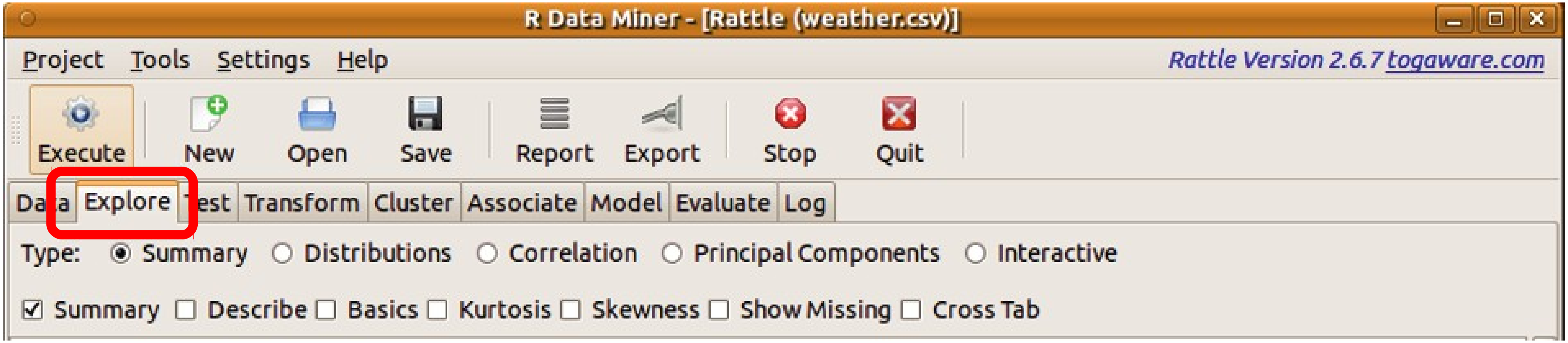
● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	●	○	○	Unique: 366
2	Location	Constant	○	○	○	○	●	○	Unique: 1
3	MinTemp	Numeric	●	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	●	○	○	○	○	○	Unique: 187
5	Rainfall	Numeric	●	○	○	○	○	○	Unique: 47
6	Evaporation	Numeric	●	○	○	○	○	○	Unique: 55
7	Sunshine	Numeric	●	○	○	○	○	○	Unique: 114 Missing: 3
14	Humidity9am	Numeric	●	○	○	○	○	○	Unique: 60
15	Humidity3pm	Numeric	●	○	○	○	○	○	Unique: 74
16	Pressure9am	Numeric	●	○	○	○	○	○	Unique: 190
17	Pressure3pm	Numeric	●	○	○	○	○	○	Unique: 193
18	Cloud9am	Numeric	●	○	○	○	○	○	Unique: 9
19	Cloud3pm	Numeric	●	○	○	○	○	○	Unique: 9
20	Temp9am	Numeric	●	○	○	○	○	○	Unique: 178
21	Temp3pm	Numeric	●	○	○	○	○	○	Unique: 200
22	RainToday	Categoric	●	○	○	○	○	○	Unique: 2
23	RISK_MM	Numeric	○	○	●	○	○	○	Unique: 47
24	RainTomorrow	Categoric	○	●	○	○	○	○	Unique: 2

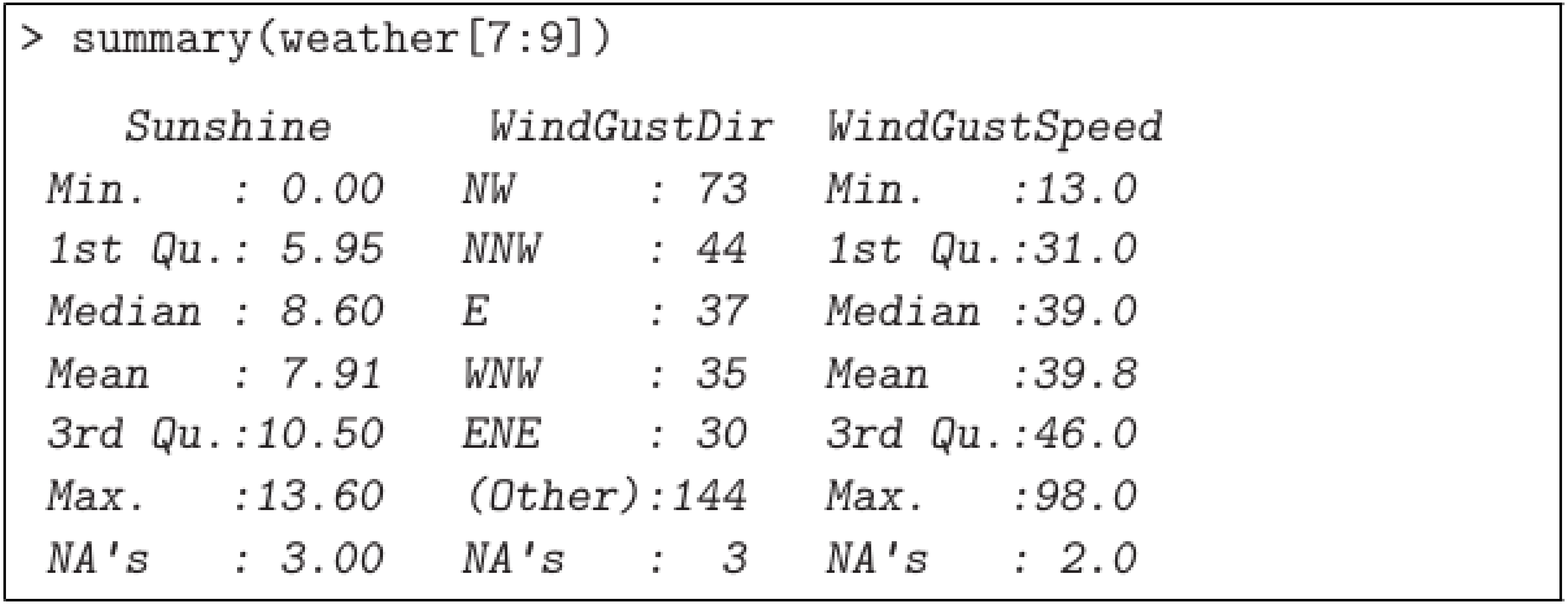
Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.

Explore



Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

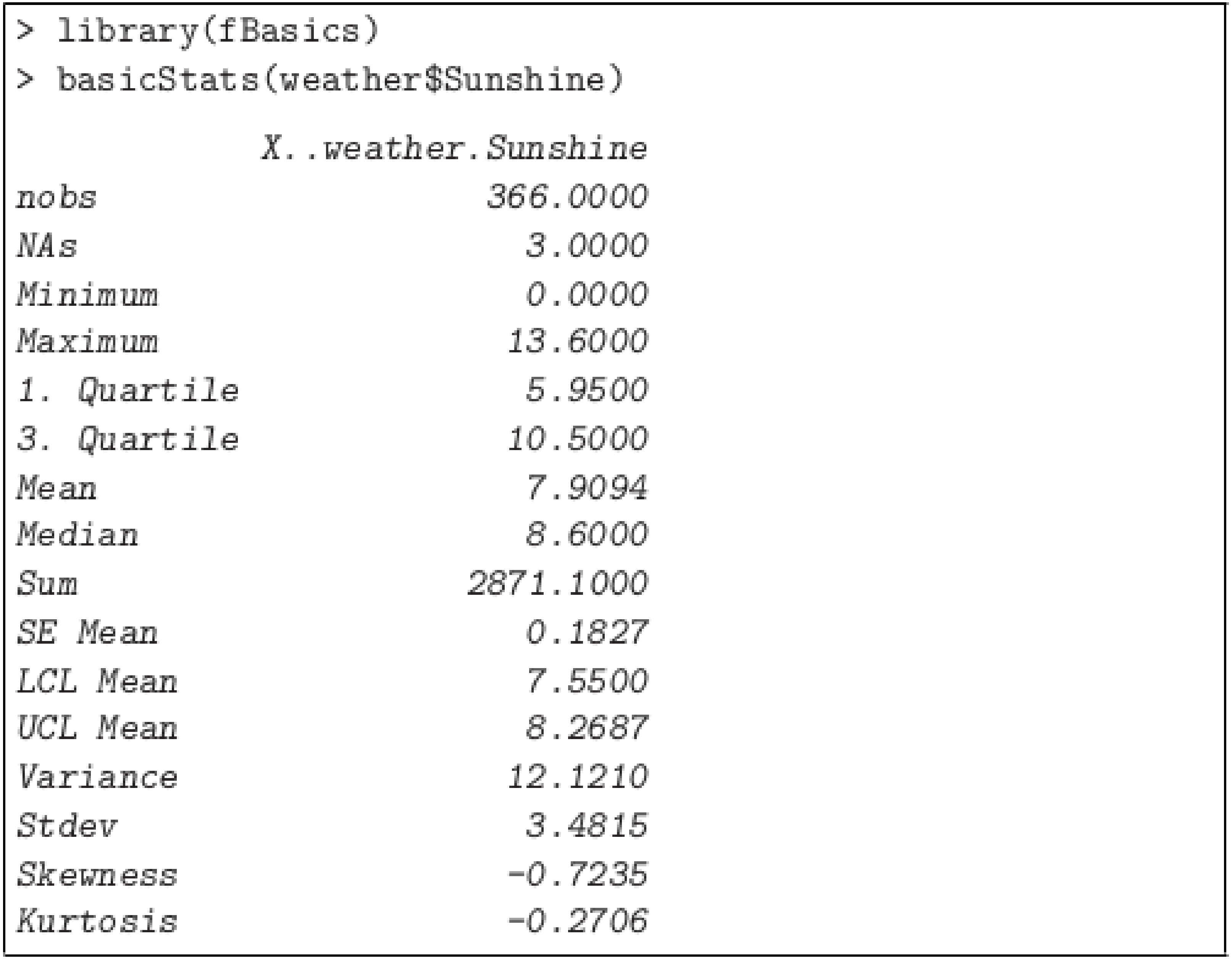
☑ Summary □ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab



```
> summary(weather[7:9])

    Sunshine         WindGustDir  WindGustSpeed
 Min.    : 0.00   NW     : 73   Min.   :13.0
 1st Qu.: 5.95   NNW    : 44   1st Qu.:31.0
 Median : 8.60   E      : 37   Median :39.0
 Mean   : 7.91   WNW    : 35   Mean   :39.8
 3rd Qu.:10.50   ENE    : 30   3rd Qu.:46.0
 Max.   :13.60   (Other):144   Max.   :98.0
 NA's   : 3.00   NA's   :  3   NA's   : 2.0
```

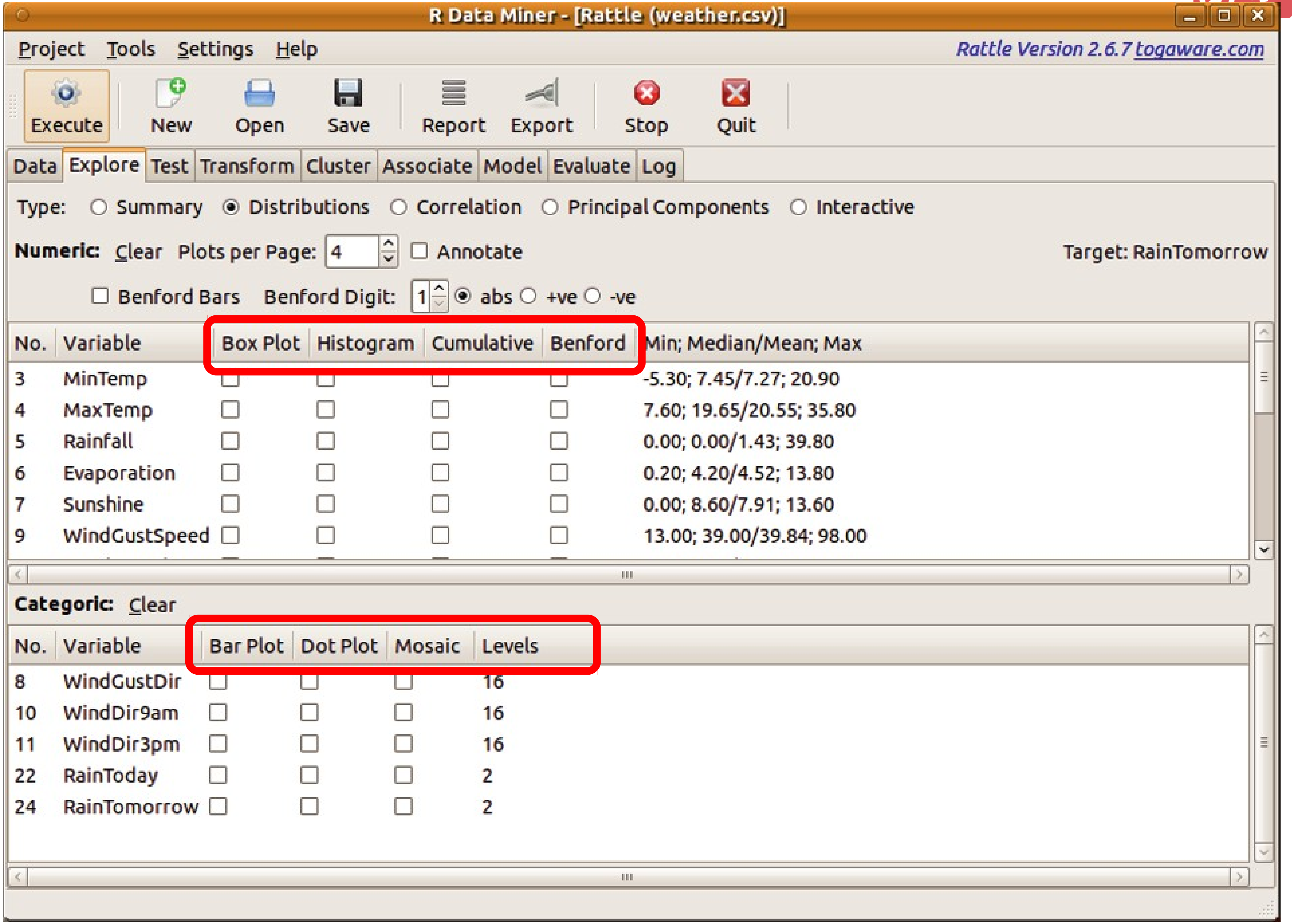
- Provides a textual overview of the data

Explore - detailed



```
> library(fBasics)
> basicStats(weather$Sunshine)

            X..weather.Sunshine
nobs                   366.0000
NAs                      3.0000
Minimum                  0.0000
Maximum                 13.6000
1. Quartile              5.9500
3. Quartile             10.5000
Mean                     7.9094
Median                   8.6000
Sum                   2871.1000
SE Mean                  0.1827
LCL Mean                 7.5500
UCL Mean                 8.2687
Variance                12.1210
Stdev                    3.4815
Skewness                -0.7235
Kurtosis                -0.2706
```

Explore (graphically)



Type: ○ Summary ● Distributions ○ Correlation ○ Principal Components ○ Interactive

Numeric: Clear Plots per Page: 4 ☐ Annotate Target: RainTomorrow

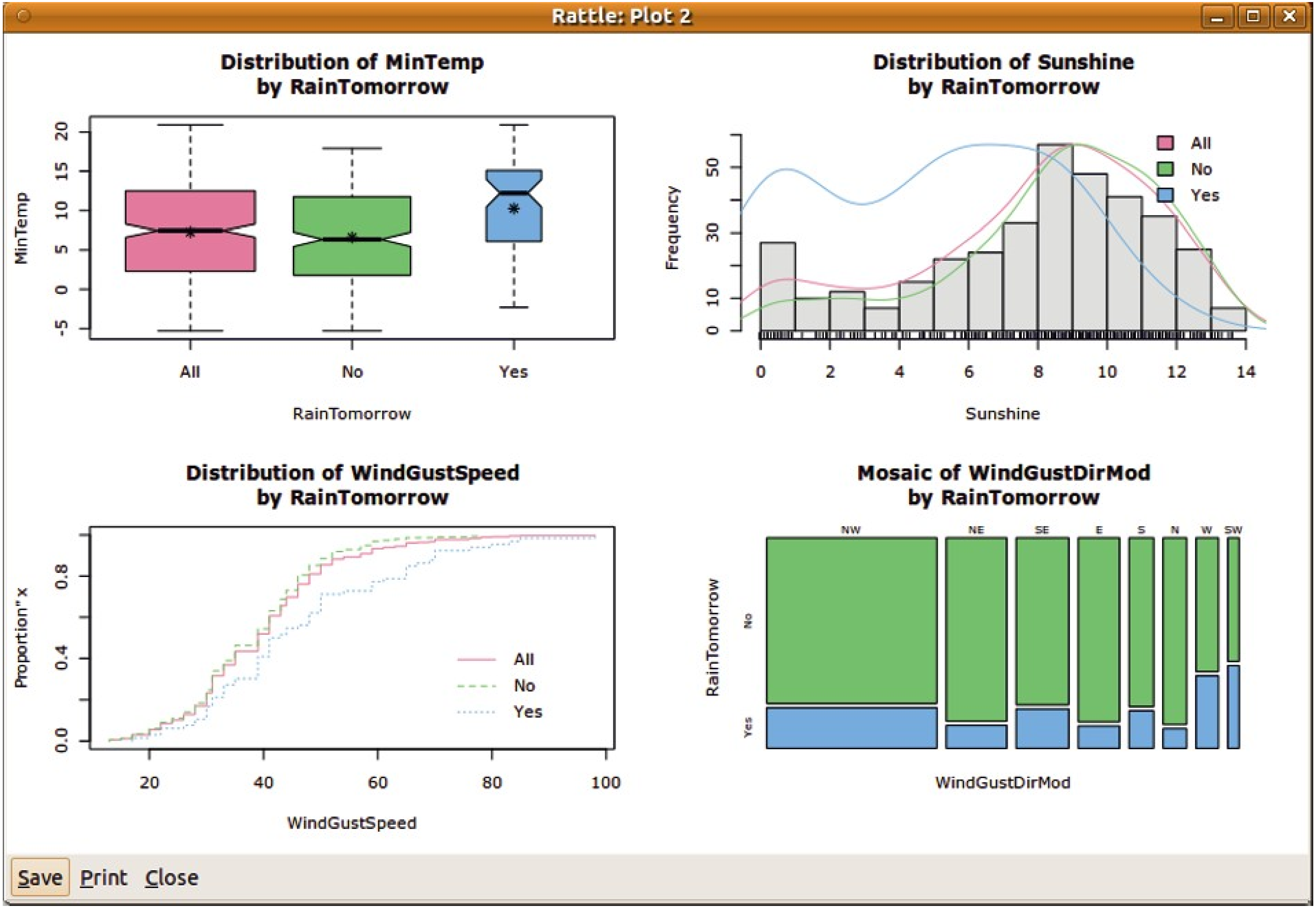
☐ Benford Bars Benford Digit: 1 ● abs ○ +ve ○ -ve

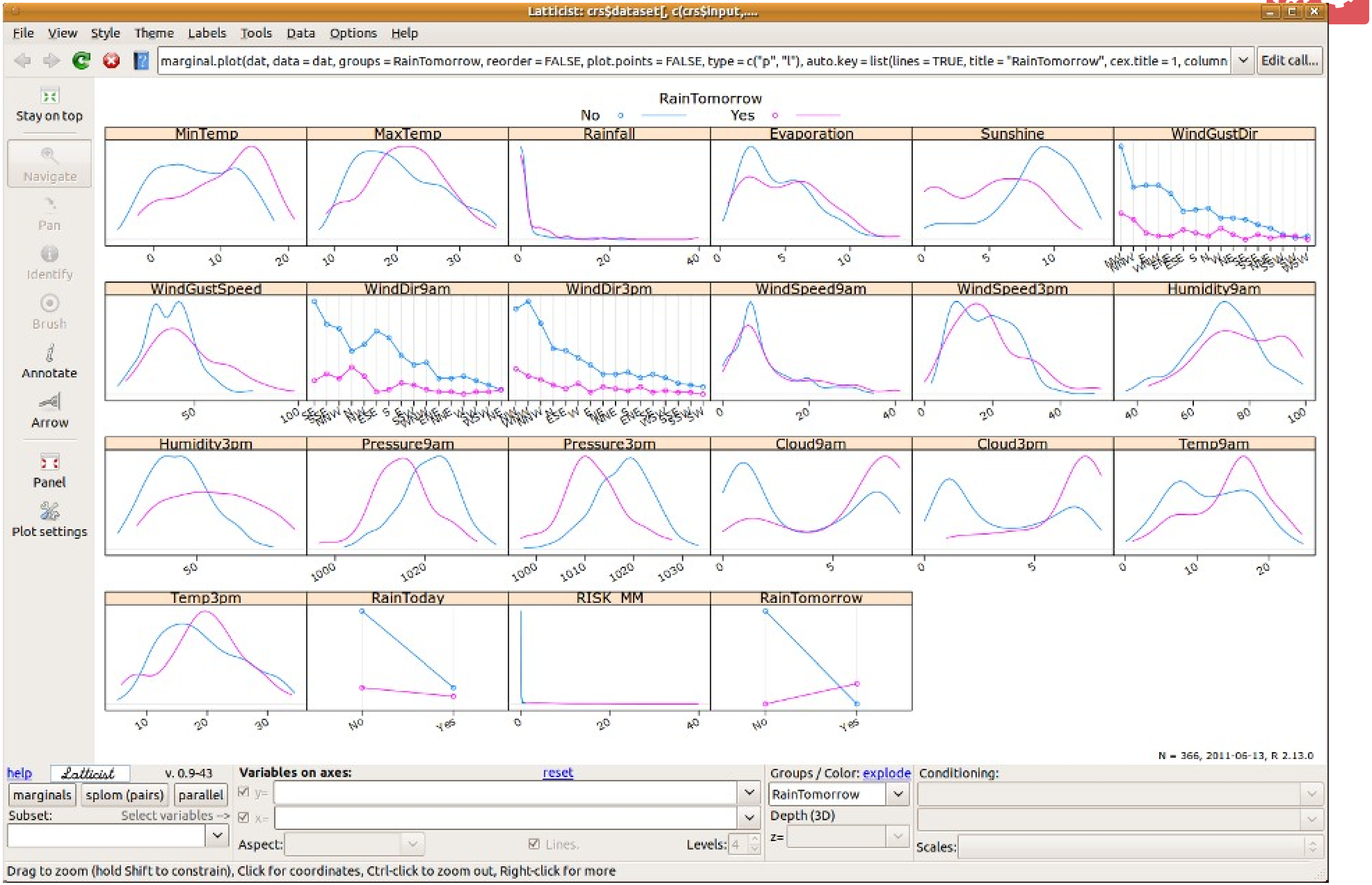
No.	Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
3	MinTemp	☐	☐	☐	☐	-5.30; 7.45/7.27; 20.90
4	MaxTemp	☐	☐	☐	☐	7.60; 19.65/20.55; 35.80
5	Rainfall	☐	☐	☐	☐	0.00; 0.00/1.43; 39.80
6	Evaporation	☐	☐	☐	☐	0.20; 4.20/4.52; 13.80
7	Sunshine	☐	☐	☐	☐	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	☐	☐	☐	☐	13.00; 39.00/39.84; 98.00

Categoric: Clear

No.	Variable	Bar Plot	Dot Plot	Mosaic	Levels
8	WindGustDir	☐	☐	☐	16
10	WindDir9am	☐	☐	☐	16
11	WindDir3pm	☐	☐	☐	16
22	RainToday	☐	☐	☐	2
24	RainTomorrow	☐	☐	☐	2

Explore (graphically)





```
marginal.plot(dat, data = dat, groups = RainTomorrow, reorder = FALSE, plot.points = FALSE, type = c("p", "l"), auto.key = list(lines = TRUE, title = "RainTomorrow", cex.title = 1, column
```

N = 366, 2011-06-13, R 2.13.0

v. 0.9-43

Variables on axes:

☑ y=

☑ x=

Subset: Select variables -->

Aspect:

☑ Lines.

Levels: 4

Groups / Color: RainTomorrow

Depth (3D)

z=

Conditioning:

Scales:

Drag to zoom (hold Shift to constrain), Click for coordinates, Ctrl-click to zoom out, Right-click for more

Figure 6.3: The parallel coordinates plot from latticist

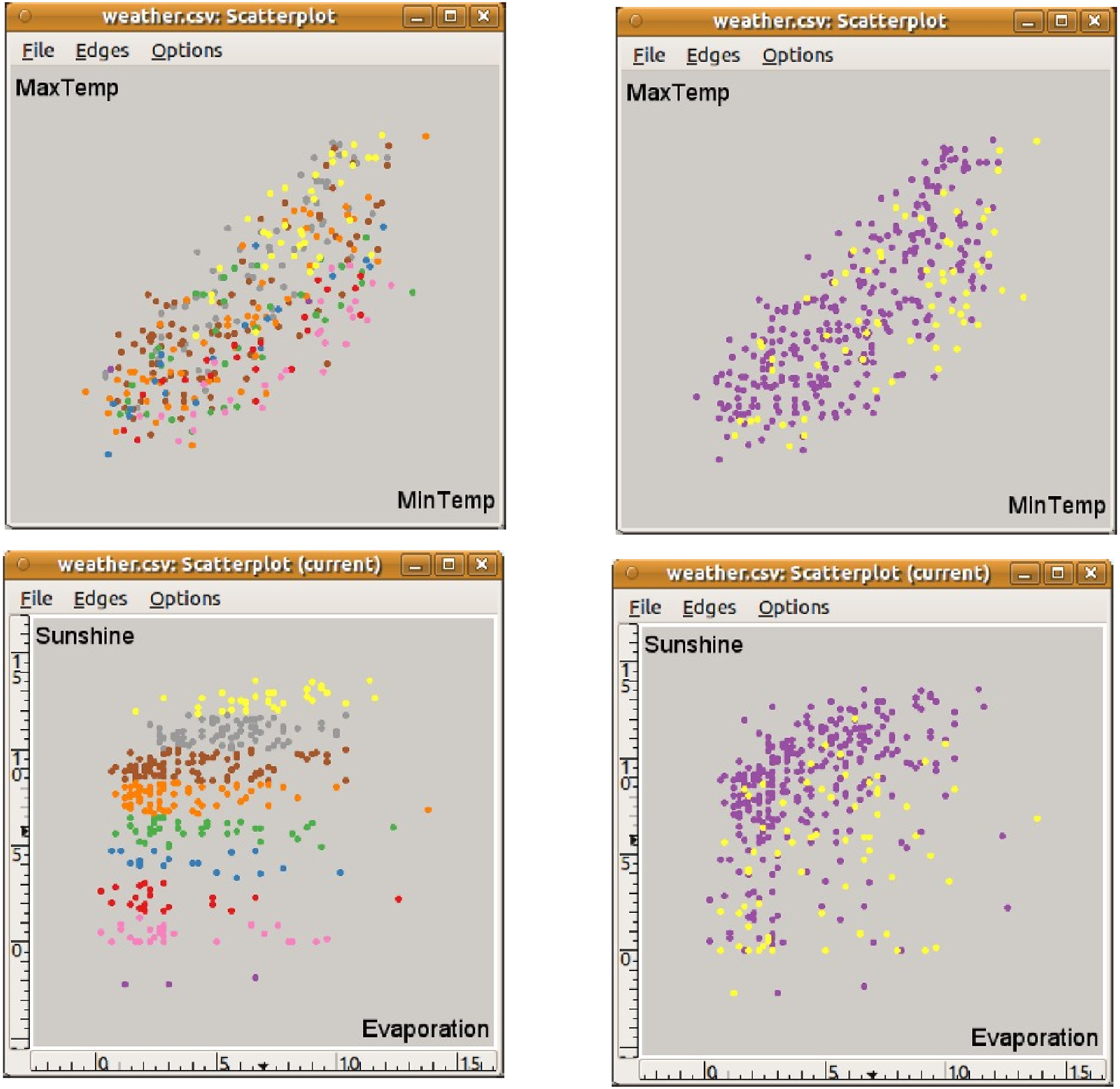


Figure 6.9: Colourful brushing of multiple scatterplots

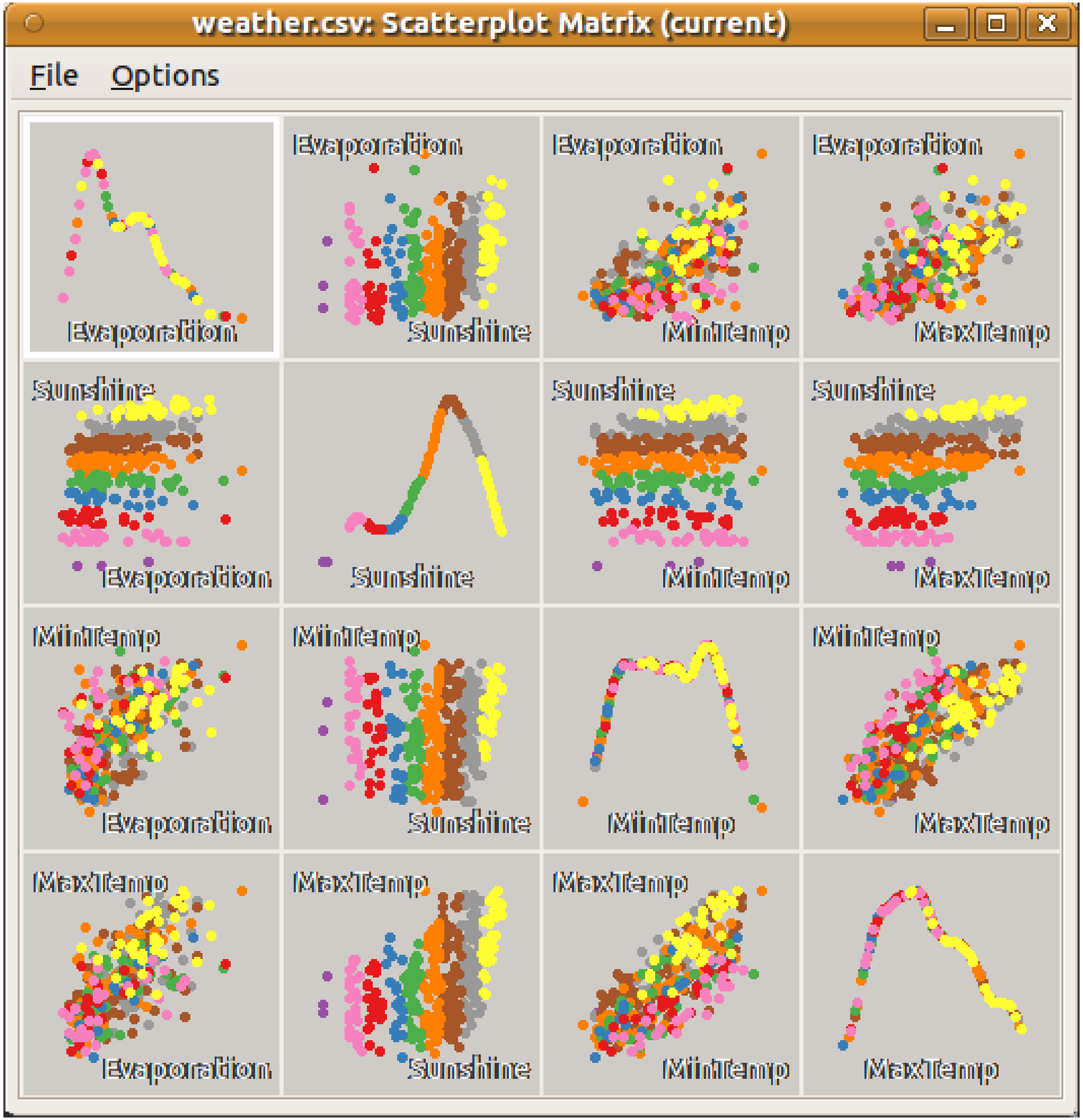


Figure 6.10: GGobi's scatter plot matrix.

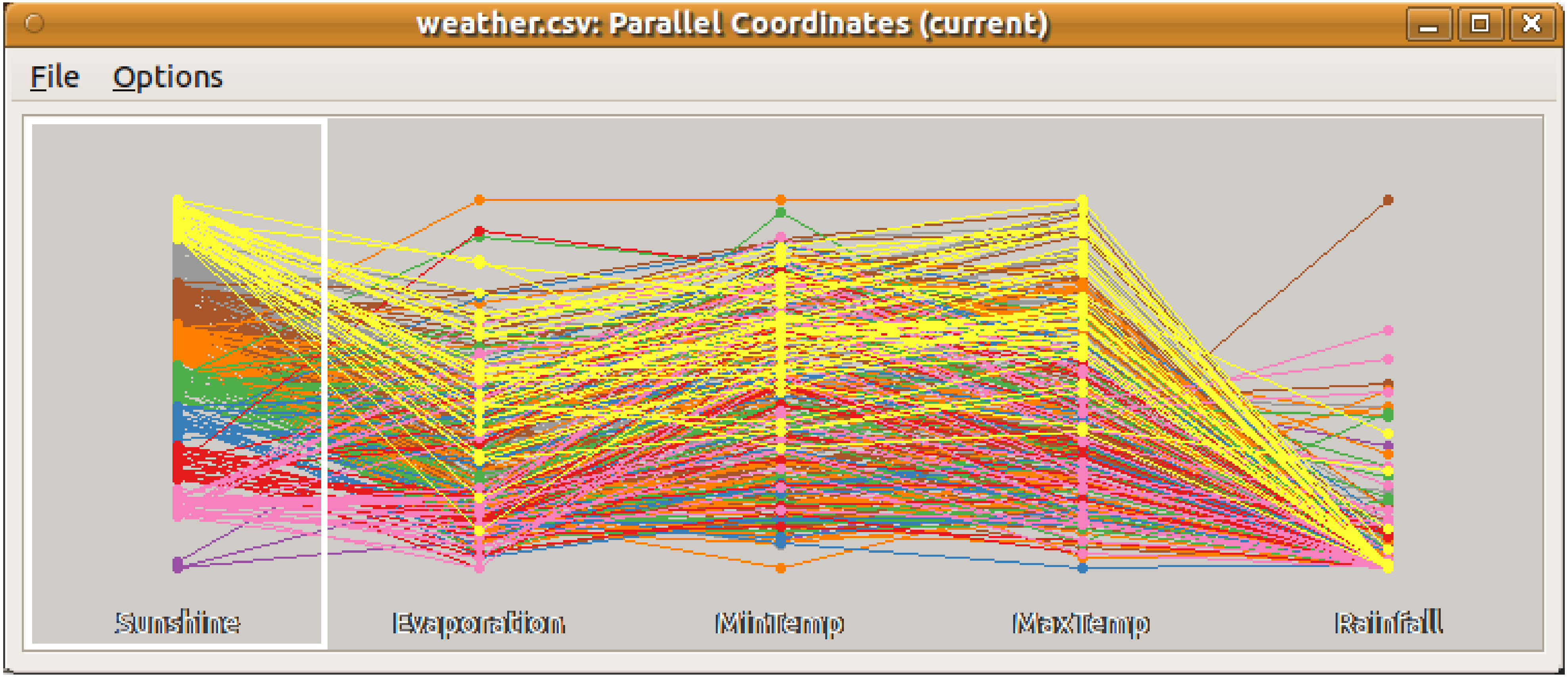
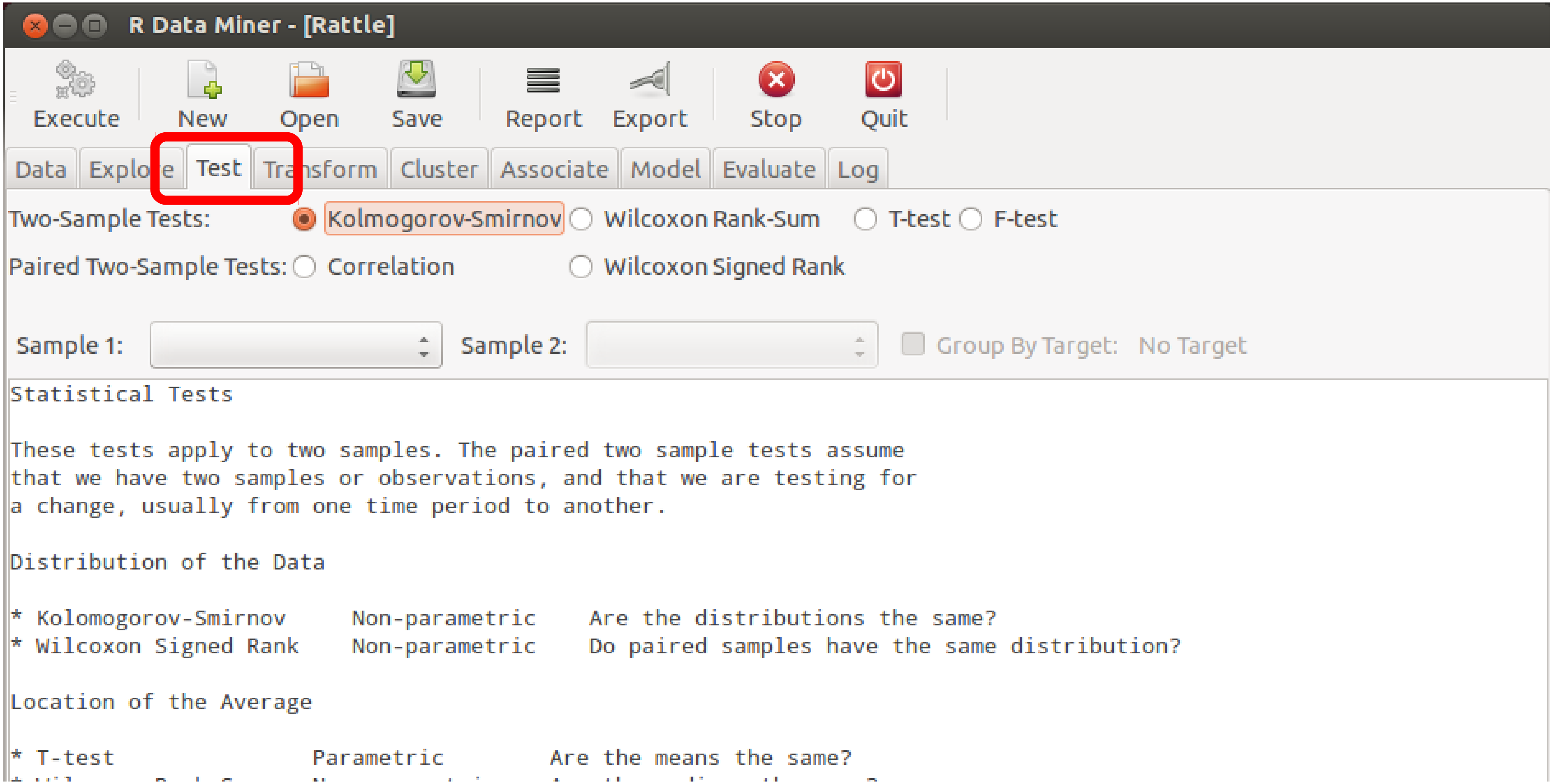


Figure 6.11: GGobi's parallel coordinates plot.

Test



Two-Sample Tests: ● Kolmogorov-Smirnov ○ Wilcoxon Rank-Sum ○ T-test ○ F-test

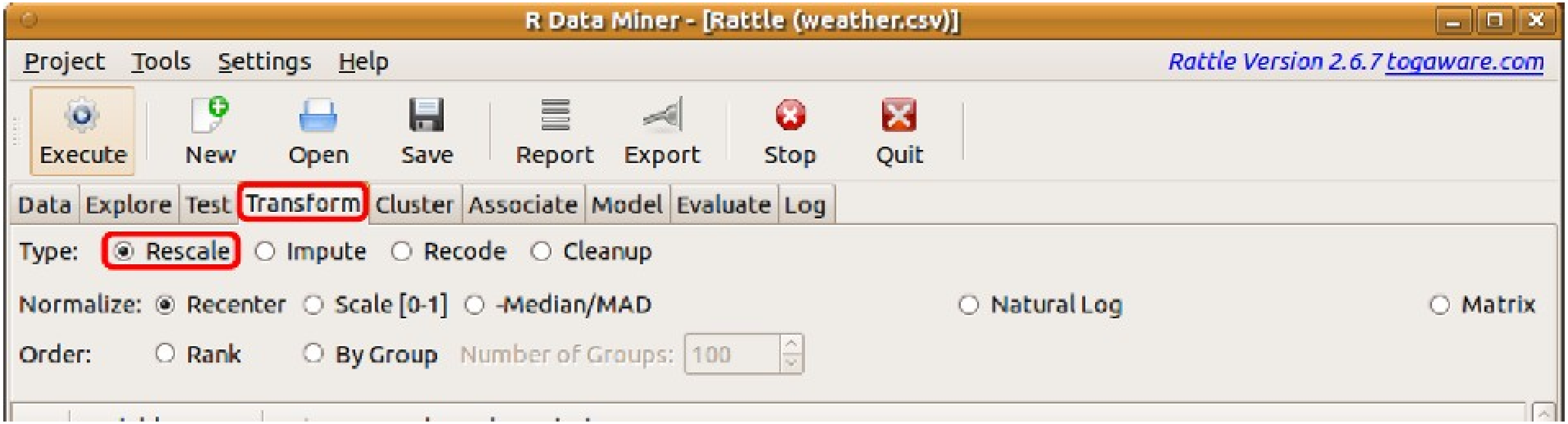
Paired Two-Sample Tests: ○ Correlation ○ Wilcoxon Signed Rank

Sample 1: Sample 2: □ Group By Target: No Target

```
Statistical Tests

These tests apply to two samples. The paired two sample tests assume
that we have two samples or observations, and that we are testing for
a change, usually from one time period to another.

Distribution of the Data

* Kolomogorov-Smirnov     Non-parametric    Are the distributions the same?
* Wilcoxon Signed Rank    Non-parametric    Do paired samples have the same distribution?

Location of the Average

* T-test               Parametric        Are the means the same?
```

- various statistical tests, e.g. the T-test and F-test

Transform



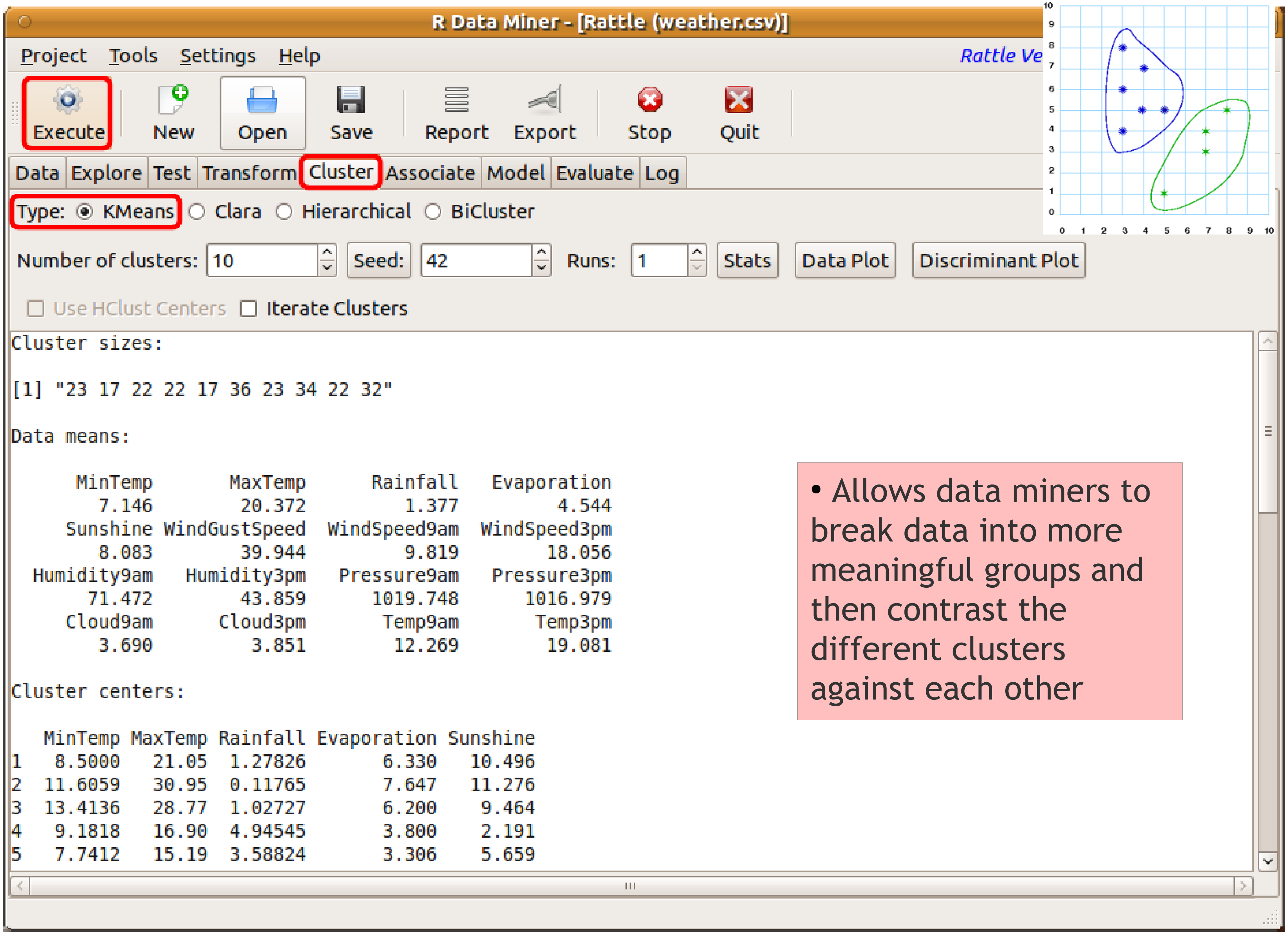
Type: ● Rescale ○ Impute ○ Recode ○ Cleanup

Normalize: ● Recenter ○ Scale [0-1] ○ -Median/MAD ○ Natural Log ○ Matrix

Order: ○ Rank ○ By Group Number of Groups: 100

- Normalising
- Filling in missing values
- Turning numeric variables into categoric variables (and vice versa)
- Dealing with outliers
- Removing variables or observations with missing values

Cluster



Type: ● KMeans ○ Clara ○ Hierarchical ○ BiCluster

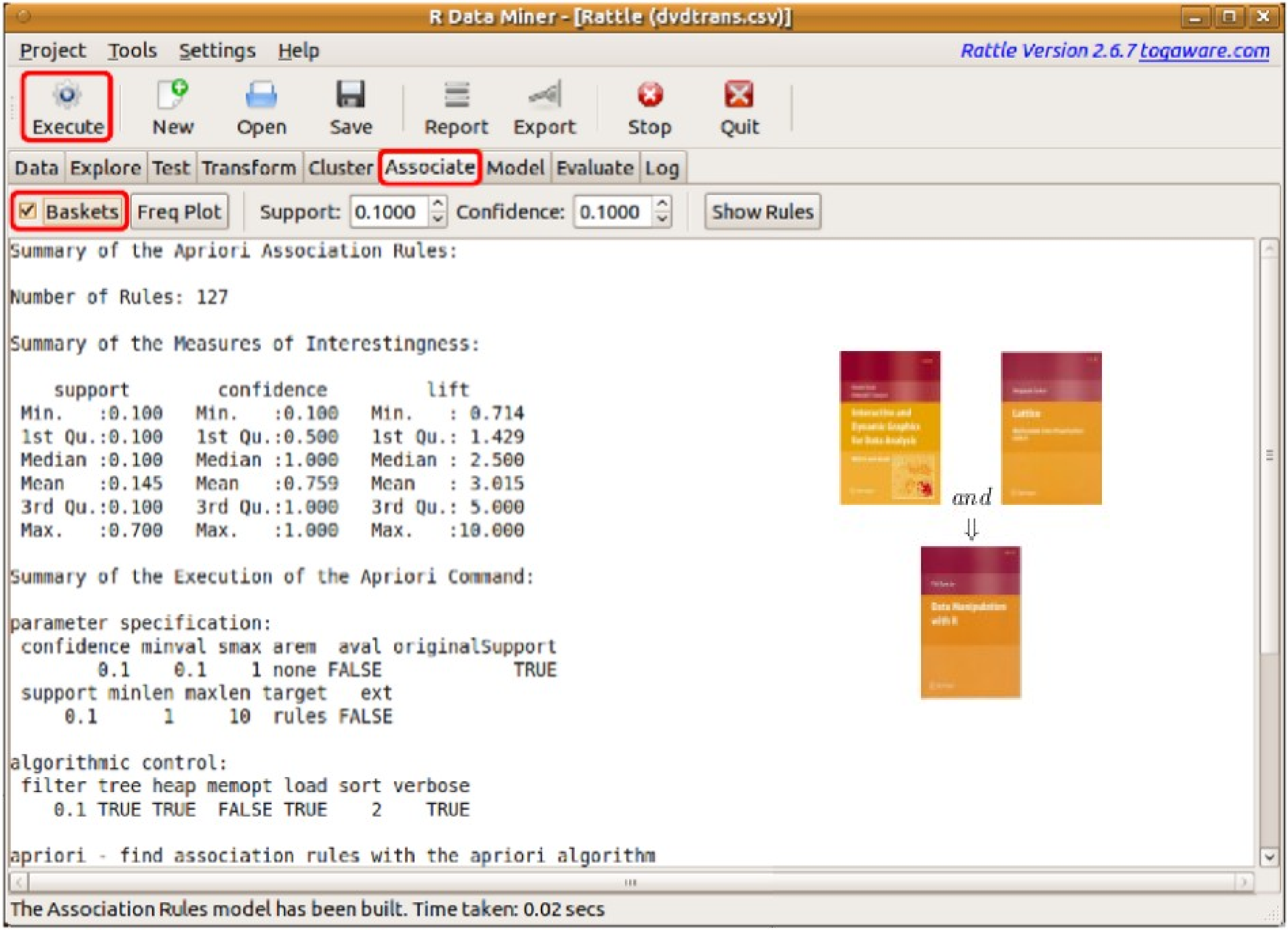
Number of clusters: 10 Seed: 42 Runs: 1

□ Use HClust Centers □ Iterate Clusters

```
Cluster sizes:

[1] "23 17 22 22 17 36 23 34 22 32"

Data means:

      MinTemp       MaxTemp       Rainfall    Evaporation
        7.146        20.372          1.377          4.544
     Sunshine WindGustSpeed   WindSpeed9am   WindSpeed3pm
        8.083        39.944          9.819         18.056
  Humidity9am   Humidity3pm    Pressure9am    Pressure3pm
       71.472        43.859       1019.748       1016.979
     Cloud9am      Cloud3pm        Temp9am        Temp3pm
        3.690         3.851         12.269         19.081

Cluster centers:

   MinTemp MaxTemp Rainfall Evaporation Sunshine
1   8.5000   21.05  1.27826       6.330   10.496
2  11.6059   30.95  0.11765       7.647   11.276
3  13.4136   28.77  1.02727       6.200    9.464
4   9.1818   16.90  4.94545       3.800    2.191
5   7.7412   15.19  3.58824       3.306    5.659
```

- Allows data miners to break data into more meaningful groups and then contrast the different clusters against each other

Associate



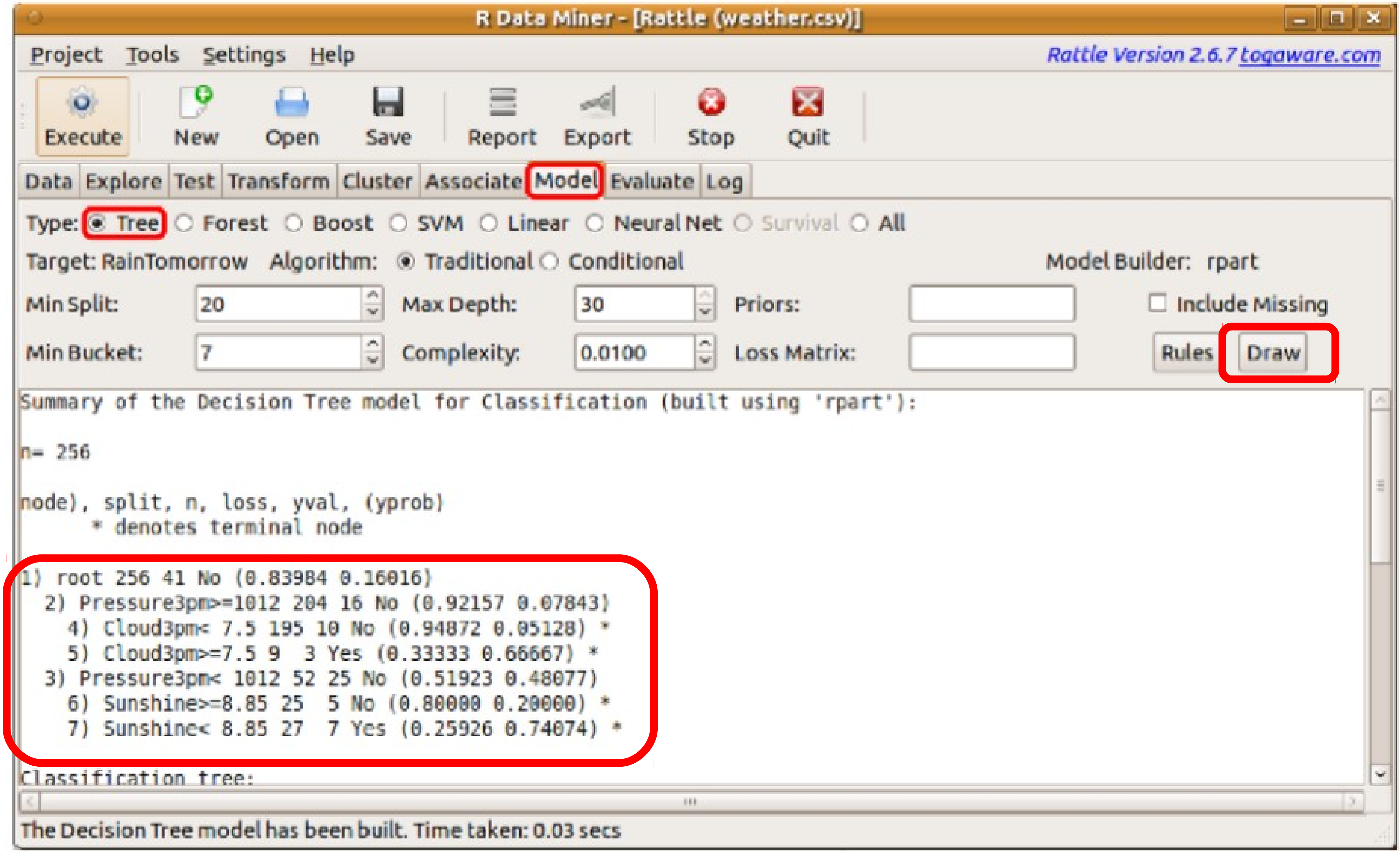
R Data Miner - [Rattle (dvdtrans.csv)]
Project Tools Settings Help
Rattle Version 2.6.7 togaware.com
Execute
New
Open
Save
Report
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Freq Plot
Show Rules
and
⇓

☑ Baskets Support: 0.1000 Confidence: 0.1000

```
Summary of the Apriori Association Rules:

Number of Rules: 127

Summary of the Measures of Interestingness:

    support          confidence          lift       
 Min.   :0.100   Min.   :0.100   Min.   : 0.714  
 1st Qu.:0.100   1st Qu.:0.500   1st Qu.: 1.429  
 Median :0.100   Median :1.000   Median : 2.500  
 Mean   :0.145   Mean   :0.759   Mean   : 3.015  
 3rd Qu.:0.100   3rd Qu.:1.000   3rd Qu.: 5.000  
 Max.   :0.700   Max.   :1.000   Max.   :10.000  

Summary of the Execution of the Apriori Command:

parameter specification:
 confidence minval smax arem  aval originalSupport
        0.1    0.1    1 none FALSE            TRUE
 support minlen maxlen target   ext
     0.1      1     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
```

The Association Rules model has been built. Time taken: 0.02 secs

Model



R Data Miner - [Rattle (weather.csv)]
Project Tools Settings Help
Rattle Version 2.6.7 togaware.com
Execute
New
Open
Save
Report
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rules
Draw

Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow Algorithm: ● Traditional ○ Conditional Model Builder: rpart

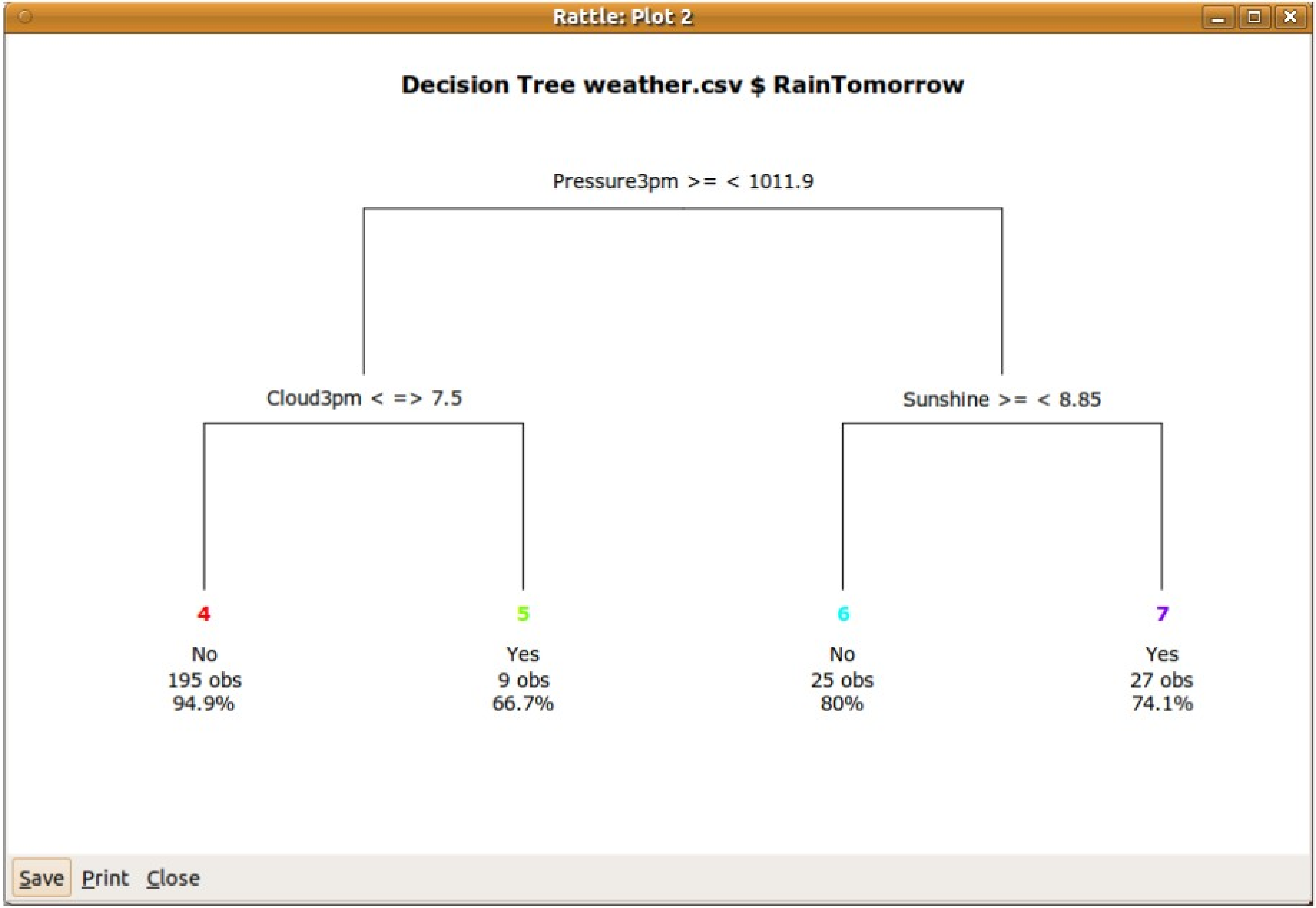
Min Split: 20 Max Depth: 30 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

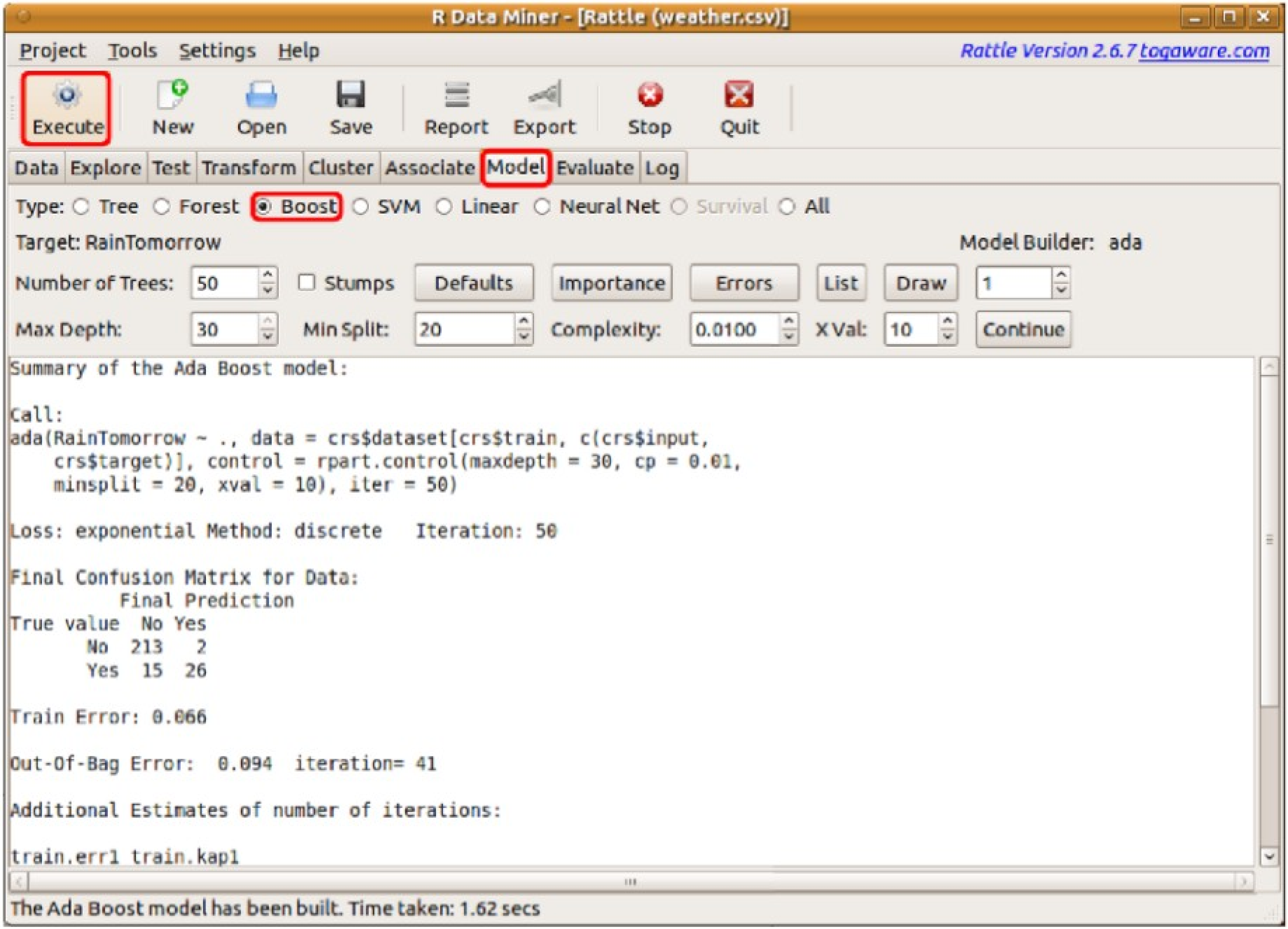
```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 256 41 No (0.83984 0.16016)
  2) Pressure3pm>=1012 204 16 No (0.92157 0.07843)
    4) Cloud3pm< 7.5 195 10 No (0.94872 0.05128) *
    5) Cloud3pm>=7.5 9  3 Yes (0.33333 0.66667) *
  3) Pressure3pm< 1012 52 25 No (0.51923 0.48077)
    6) Sunshine>=8.85 25  5 No (0.80000 0.20000) *
    7) Sunshine< 8.85 27  7 Yes (0.25926 0.74074) *

Classification tree:
```

The Decision Tree model has been built. Time taken: 0.03 secs

Decision Trees



Boosting



Type: ○ Tree ○ Forest ● Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow

Model Builder: ada

Number of Trees: 50 □ Stumps 1

Max Depth: 30 Min Split: 20 Complexity: 0.0100 X Val: 10

```
Summary of the Ada Boost model:

Call:
ada(RainTomorrow ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], control = rpart.control(maxdepth = 30, cp = 0.01,
    minsplit = 20, xval = 10), iter = 50)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value  No Yes
       No  213   2
       Yes  15  26

Train Error: 0.066

Out-Of-Bag Error:  0.094  iteration= 41

Additional Estimates of number of iterations:

train.err1 train.kap1
```

The Ada Boost model has been built. Time taken: 1.62 secs

Boosting

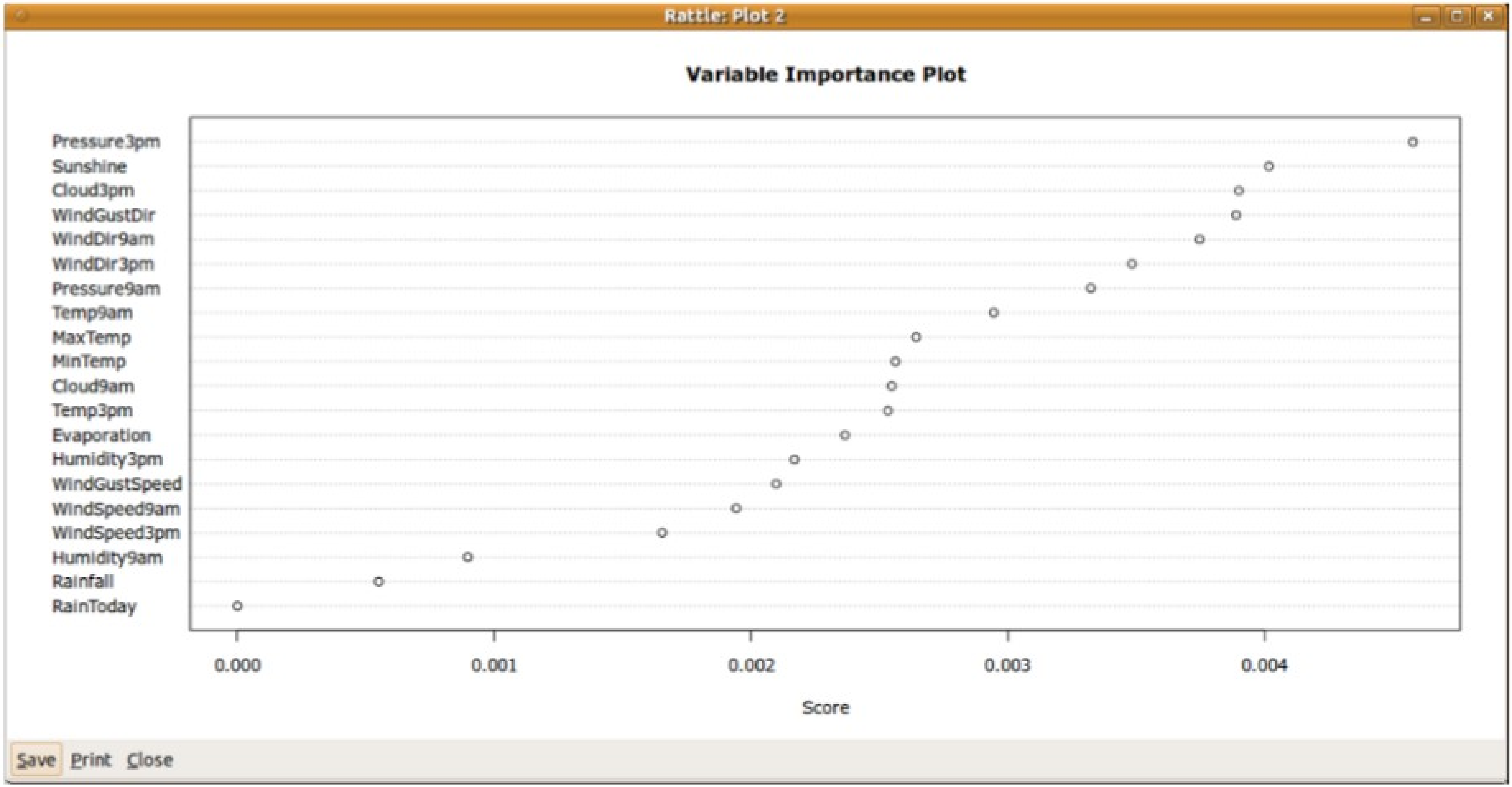
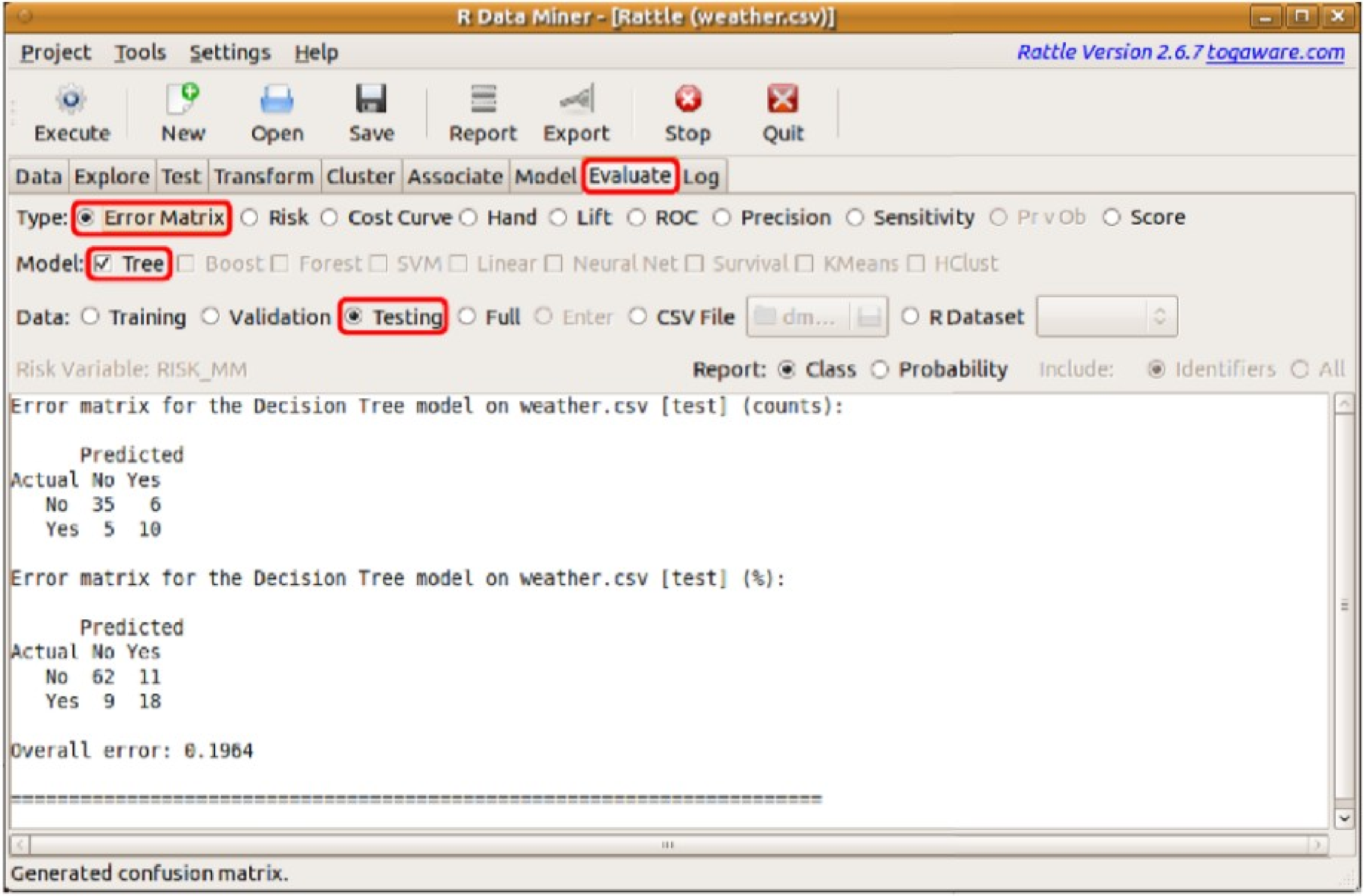


Figure 13.3: The variable importance plot for a boosted model.

Evaluate


R Data Miner - [Rattle (weather.csv)]
Project Tools Settings Help
Rattle Version 2.6.7 togaware.com
Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score
□ Boost □ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust
○ Training ○ Validation ○ Full ○ Enter ○ CSV File dm... ○ R Dataset
○ Probability Include: ● Identifiers ○ All

Type: ● Error Matrix

Model: ☑ Tree

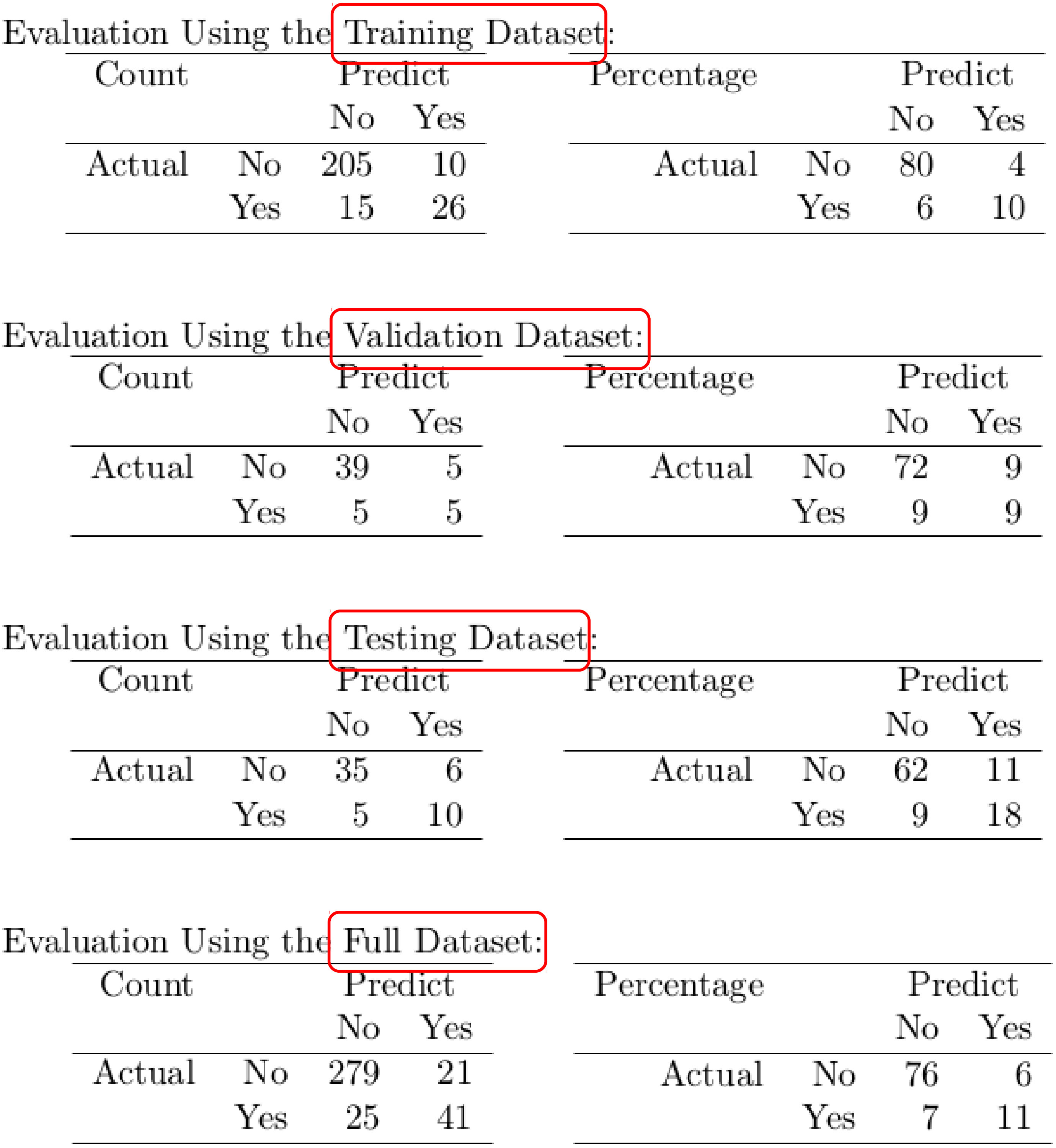
Data: ● Testing

Risk Variable: RISK_MM

Report: ● Class

```
Error matrix for the Decision Tree model on weather.csv [test] (counts):

      Predicted
Actual No Yes
   No  35   6
   Yes  5  10

Error matrix for the Decision Tree model on weather.csv [test] (%):

      Predicted
Actual No Yes
   No  62  11
   Yes  9  18

Overall error: 0.1964

======================================================================
```

Generated confusion matrix.



Evaluation Using the Training Dataset:

Count		Predict	
		No	Yes
Actual	No	205	10
	Yes	15	26

Percentage		Predict	
		No	Yes
Actual	No	80	4
	Yes	6	10

Evaluation Using the Validation Dataset:

Count		Predict	
		No	Yes
Actual	No	39	5
	Yes	5	5

Percentage		Predict	
		No	Yes
Actual	No	72	9
	Yes	9	9

Evaluation Using the Testing Dataset:

Count		Predict	
		No	Yes
Actual	No	35	6
	Yes	5	10

Percentage		Predict	
		No	Yes
Actual	No	62	11
	Yes	9	18

Evaluation Using the Full Dataset:

Count		Predict	
		No	Yes
Actual	No	279	21
	Yes	25	41

Percentage		Predict	
		No	Yes
Actual	No	76	6
	Yes	7	11

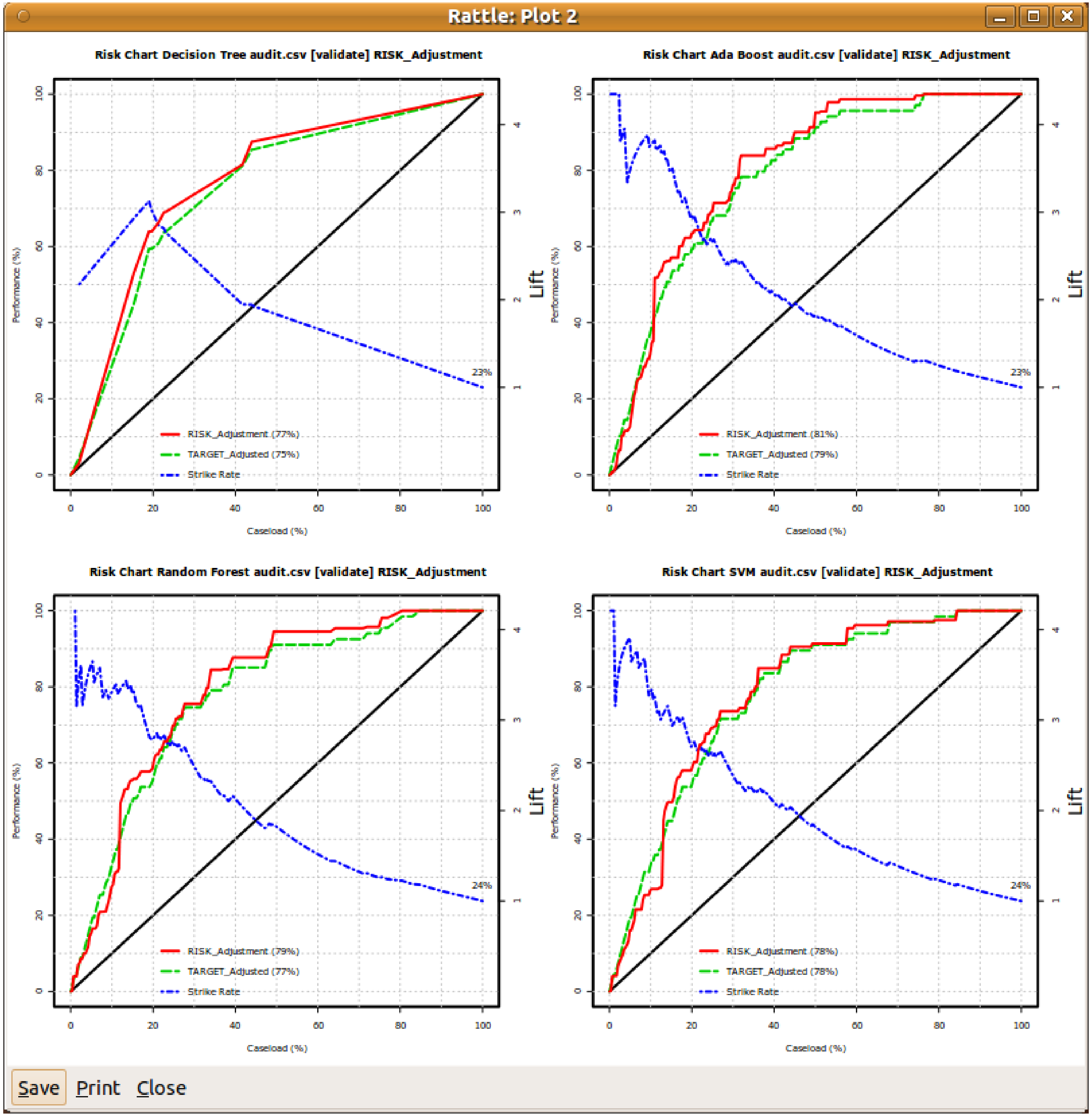


Figure 15.5: Four risk charts displayed to compare performances of multiple model builders on the audit dataset.

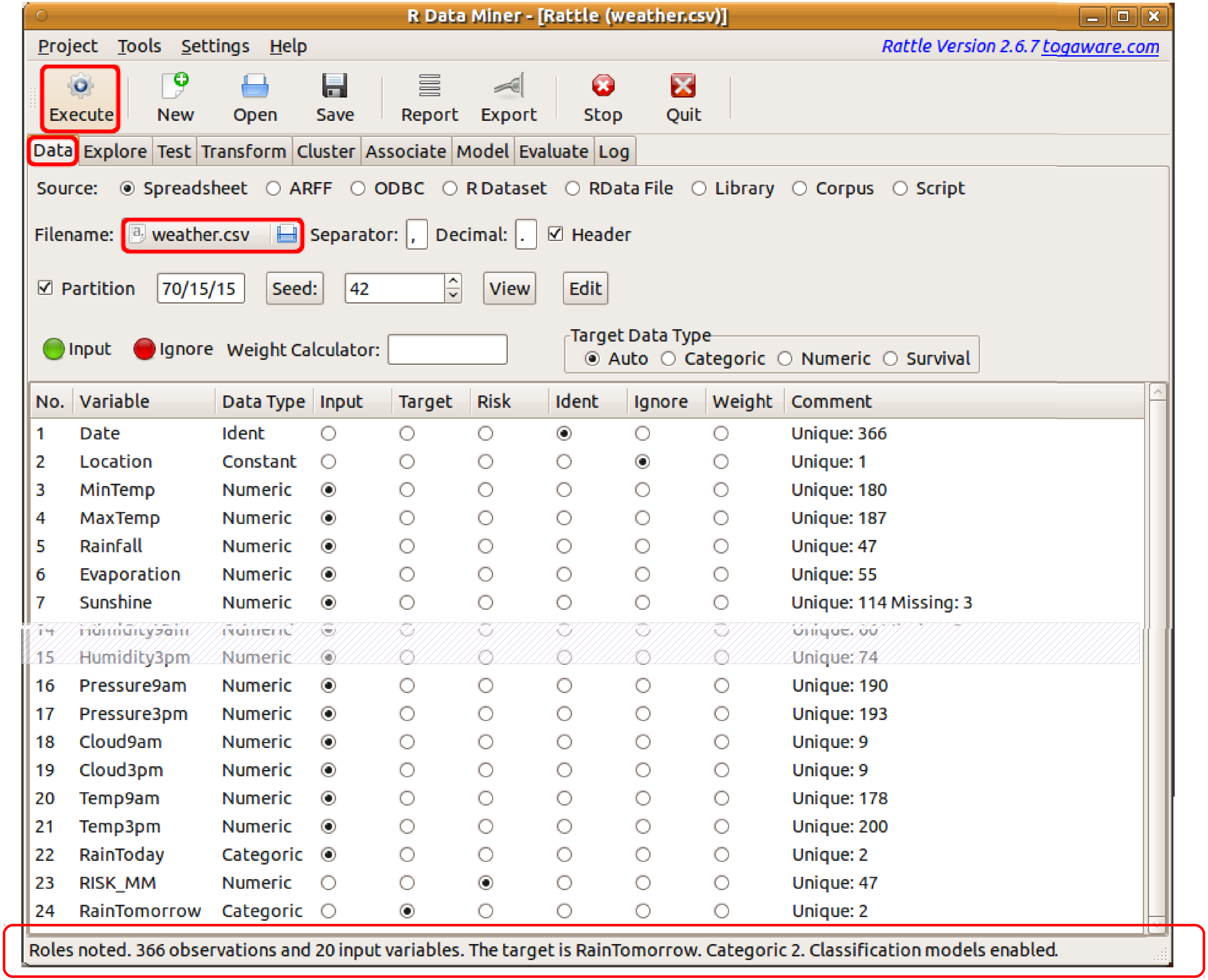
First Model

(example)

First model

- Once we have processed our data, we can build a model

1) Click on the Execute button
   - Rattle will notice that no dataset has been identified
2) The sample “weather” dataset will be offered
   - Click “Yes”
3) Click on the Model tab
   - This is where we tell Rattle what kind of model we want to build
4) Click on the Execute button.



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: weather.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42

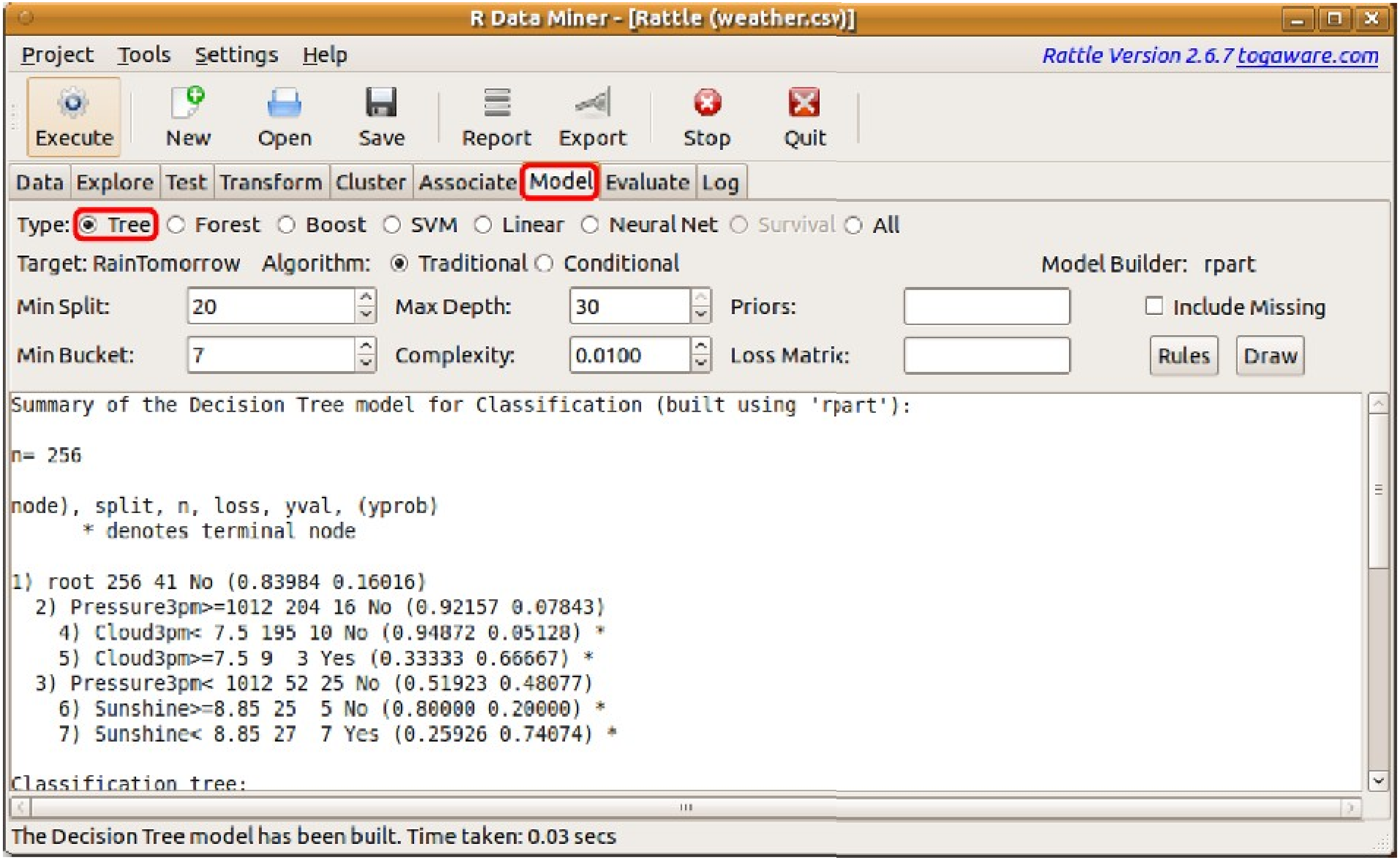
● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	●	○	○	Unique: 366
2	Location	Constant	○	○	○	○	●	○	Unique: 1
3	MinTemp	Numeric	●	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	●	○	○	○	○	○	Unique: 187
5	Rainfall	Numeric	●	○	○	○	○	○	Unique: 47
6	Evaporation	Numeric	●	○	○	○	○	○	Unique: 55
7	Sunshine	Numeric	●	○	○	○	○	○	Unique: 114 Missing: 3
14	Humidity9am	Numeric	●	○	○	○	○	○	Unique: [illegible]
15	Humidity3pm	Numeric	●	○	○	○	○	○	Unique: 74
16	Pressure9am	Numeric	●	○	○	○	○	○	Unique: 190
17	Pressure3pm	Numeric	●	○	○	○	○	○	Unique: 193
18	Cloud9am	Numeric	●	○	○	○	○	○	Unique: 9
19	Cloud3pm	Numeric	●	○	○	○	○	○	Unique: 9
20	Temp9am	Numeric	●	○	○	○	○	○	Unique: 178
21	Temp3pm	Numeric	●	○	○	○	○	○	Unique: 200
22	RainToday	Categoric	●	○	○	○	○	○	Unique: 2
23	RISK_MM	Numeric	○	○	●	○	○	○	Unique: 47
24	RainTomorrow	Categoric	○	●	○	○	○	○	Unique: 2

Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.

Building a Model



Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow Algorithm: ● Traditional ○ Conditional Model Builder: rpart

Min Split: 20 Max Depth: 30 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 256 41 No (0.83984 0.16016)
  2) Pressure3pm>=1012 204 16 No (0.92157 0.07843)
    4) Cloud3pm< 7.5 195 10 No (0.94872 0.05128) *
    5) Cloud3pm>=7.5 9  3 Yes (0.33333 0.66667) *
  3) Pressure3pm< 1012 52 25 No (0.51923 0.48077)
    6) Sunshine>=8.85 25  5 No (0.80000 0.20000) *
    7) Sunshine< 8.85 27  7 Yes (0.25926 0.74074) *

Classification tree:
```

The Decision Tree model has been built. Time taken: 0.03 secs

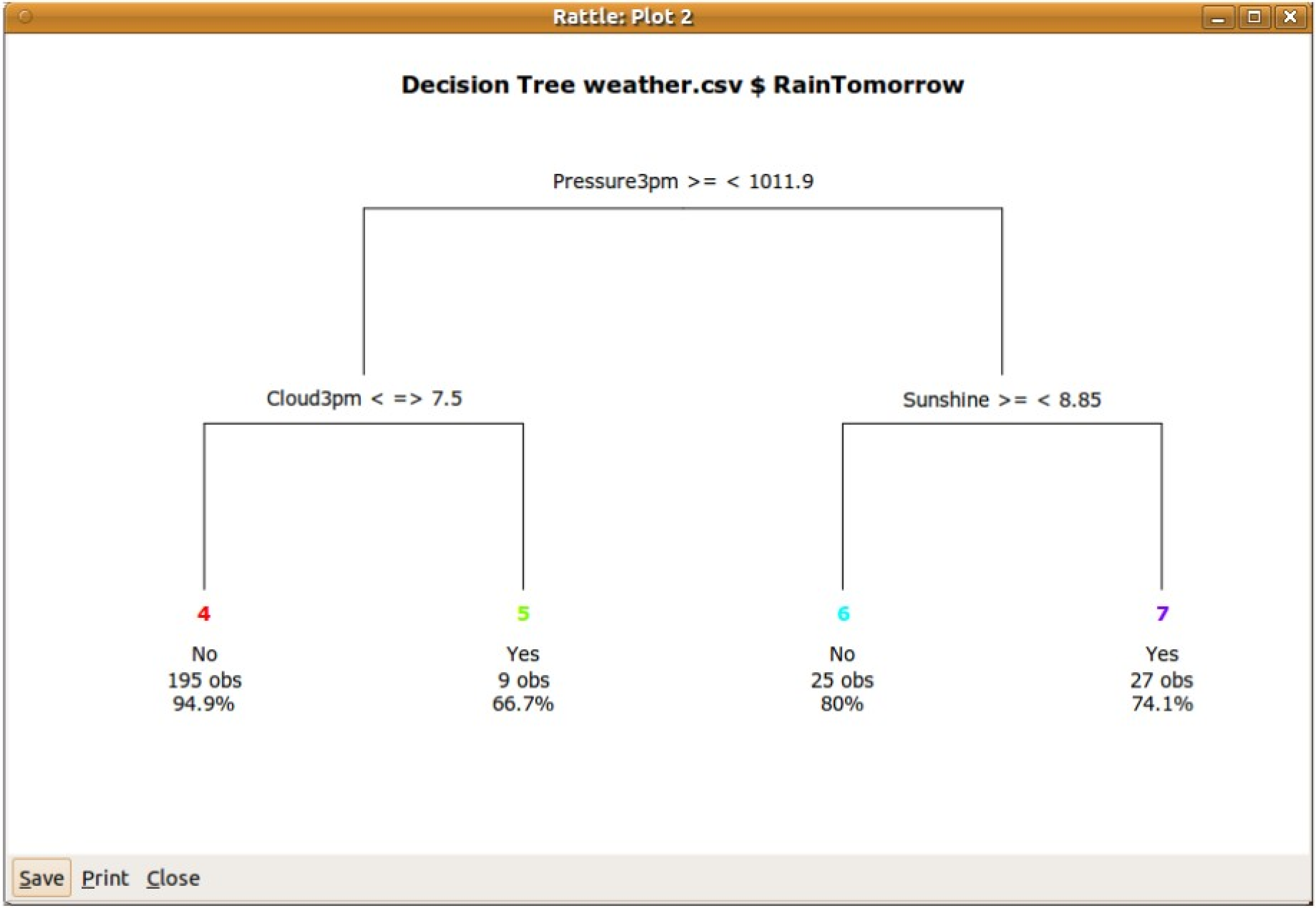

Rattle: Plot 2
Decision Tree weather.csv $ RainTomorrow
Pressure3pm >= < 1011.9
Cloud3pm < => 7.5
Sunshine >= < 8.85
4
No
195 obs
94.9%
5
Yes
9 obs
66.7%
6
No
25 obs
80%
7
Yes
27 obs
74.1%
Save
Print
Close

Figure 2.5

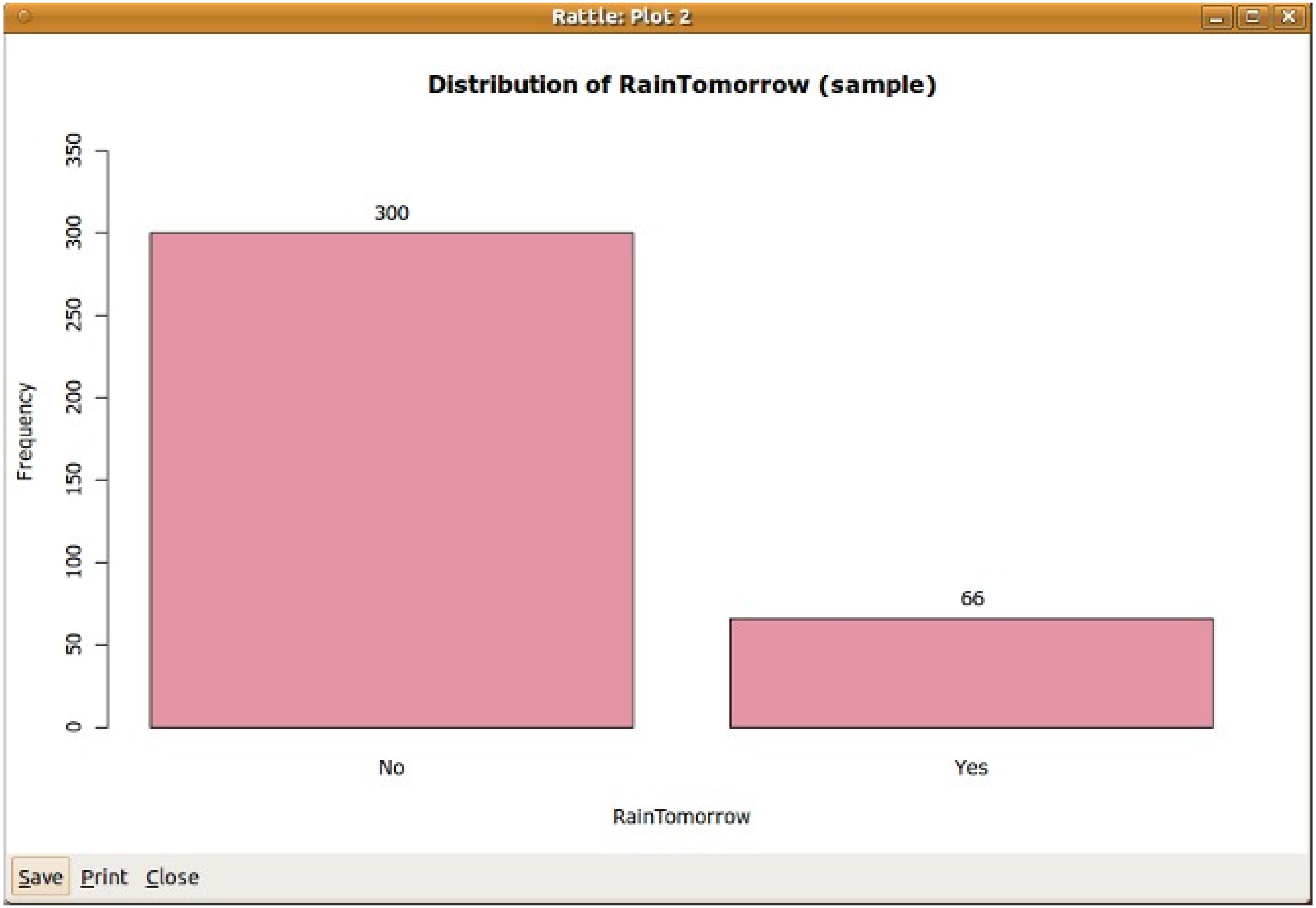


Figure 2.6: The target variable, RainTomorrow, is skewed, with Yes being quite under-represented

Appendix

Installation

Install R

www.r-project.org

Start R

Install Rattle

```
> install.packages("rattle")
```

Load rattle into the R library

```
> library(rattle)

> rattle()
```

Articles on Rattle

- **Rattle: A Data Mining GUI for R**
- http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf

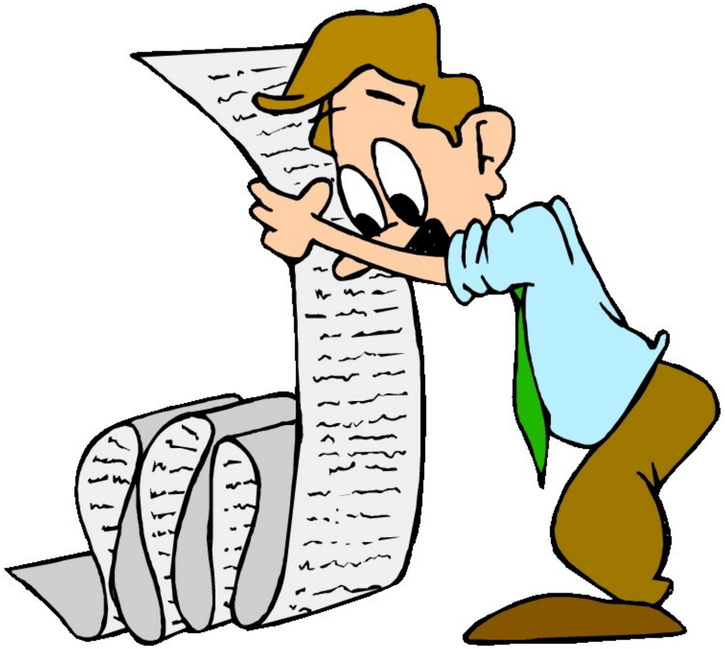
- **Getting started with data mining in R using Rattle**
- http://techpad.co.uk/content.php?sid=240

R Resources (sample)

- **Data Manipulation with R** (Spector, 2008) – covers basic data structures, reading and writing data, subscripting, manipulating, aggregating, and reshaping data
- **Introductory Statistics with R** (Dalgaard, 2008) - good introduction to statistics using R.
- **Modern Applied Statistics with S** (Venables and Ripley, 2002 - an extensive introduction to statistics using R.
- **Data Analysis and Graphics Using R** (Maindonald and Braun, 2007) - excellent practical coverage of many aspects of exploring and modelling data using R
- **The Elements of Statistical Learning** (Hastie et al., 2009) is a more mathematical treatise, covering all of the machine learning techniques discussed in this book in quite some mathematical depth.
- **R for SAS and SPSS Users** (Muenchen, 2008) is an excellent choice
- **Lattice: Multivariate Data Visualization with R** (Sarkar, 2008) - covers
- the extensive capabilities of one of the graphics/plotting packages available for R.
- **ggplot2: Elegant Graphics for Data Analysis** (Wickham, 2009) - newer graphics framework is detailed
- **Bivand et al.** (2008) cover applied spatial data analysis,
- **Kleiber and Zeileis** (2008) cover applied econometrics
- **Cowpertwait and Metcalfe** (2009) cover time series

Agenda

- Intro
- Data mining
- Rattle installation
- Rattle workflow
- Appendix



The Data Co.™

“Data Mining with Rattle and R”

DIT Analytics Club
07 March, 2013
“Peadar Kearney's”, Dame Street

Colman McMahon
colman@thedata.co

LinkedIn®
www.linkedin.com/in/colmanmcmahon