

可视化数据挖掘工具Rattle

周世祥

2020/5/28

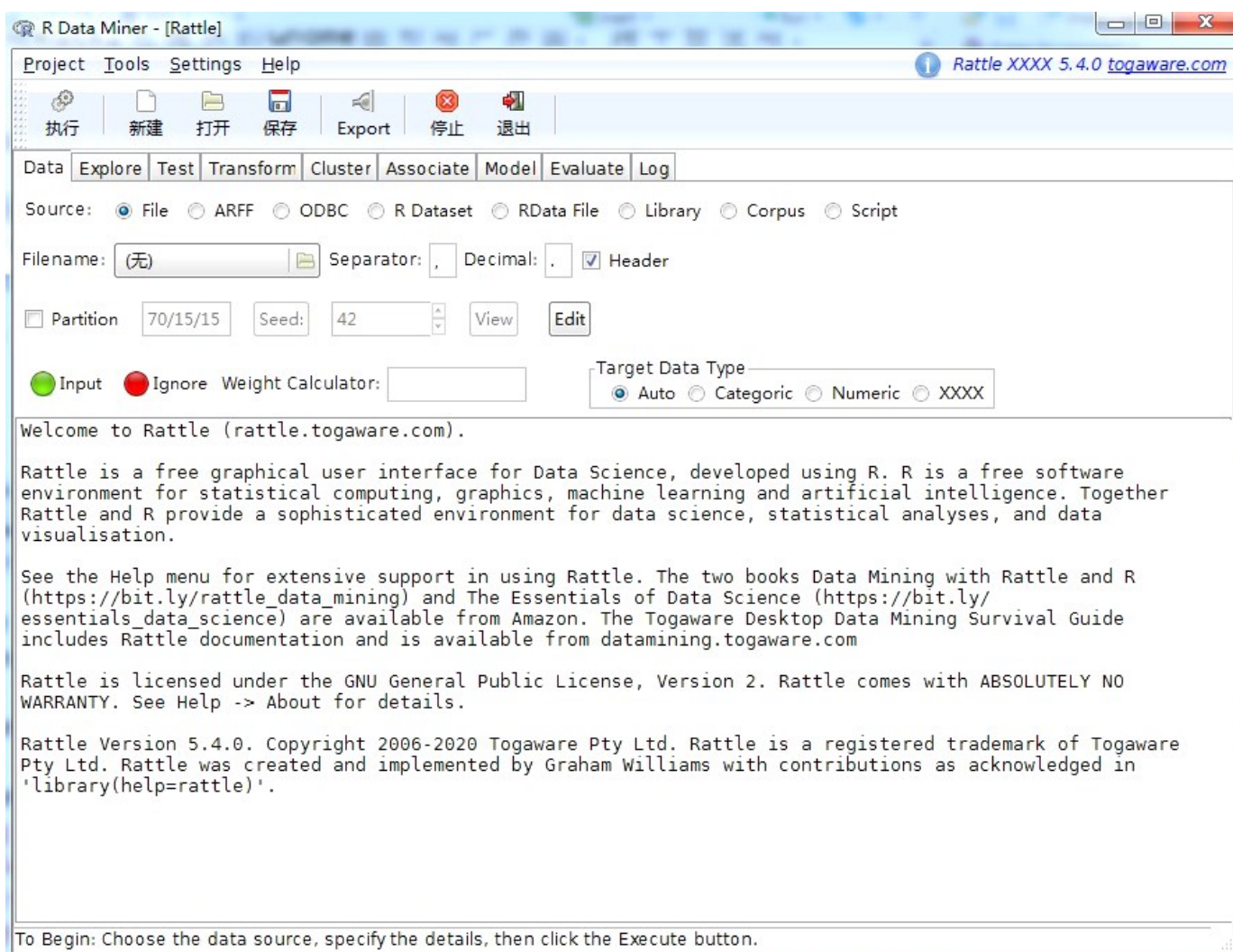
了解并安装Rattle

是一个用于数据挖掘的R语言的图形交互界面，从数据整理到模型评价，完整的解决方案。

使用RGtk2包提供的Gnome图形用户界面，跨平台使用。

```
install.packages("RGtk2")

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



rattle功能

界面从上到下，一次排列的是菜单栏，工具栏，标签栏。

Data选择数据源，输入数据

Explore执行数据探索，理解数据分布情况

Test提供各种统计检验

Transform可以变换数据的形式

Cluster为数据聚类，包括k-means,系统聚类和双聚类

Associate为关联规则方法

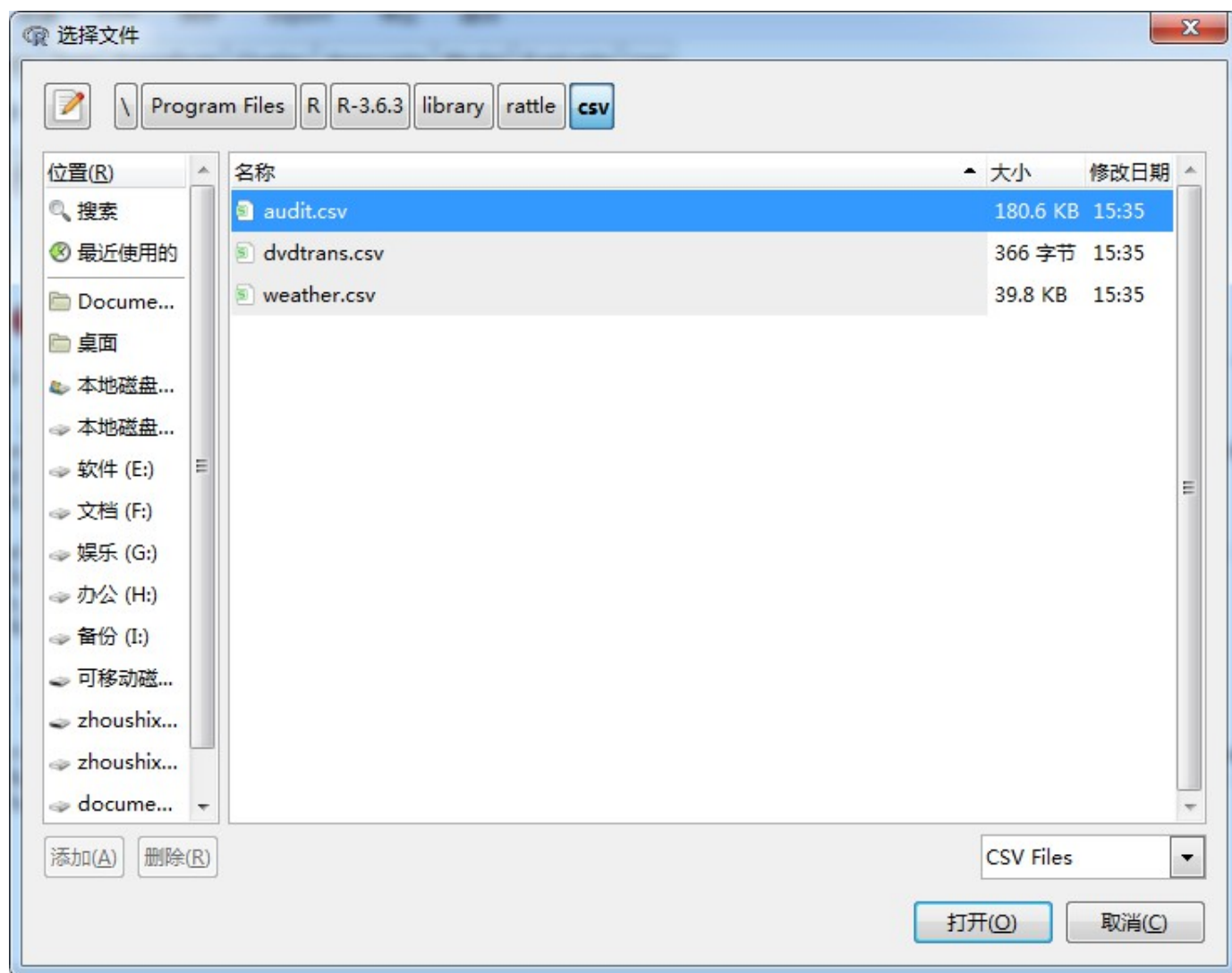
Model是内容最丰富的一个标签，包括多种算法：决策树，随机森林，组合算法，支持向量机，线性模型，人工神经网络，生存分析，。

Evaluate为模型评估。有一系列模型评估标准，混淆矩阵Error Matrix，模型风险表Risk,模型ROC曲线，得分表Score等

导入数据

有内置数据集，也能从各种各样的来源中读取数据，支持大量的文件格式。

单击“Filename”按钮可以打开“选择文件”对话框，选择需要的csv文件，此处选择自带的天气数据集。



按“执行”按钮，或F2，将数据从文件中导入。

数据导入后，Rattle会利用sample函数进行随机抽样，将样本按70：15：15的比例分成训练集，验证集，测试集。可以通过Partition选项调整各部分比例。通过seed选项改变随机种子。查看log记录的实例。

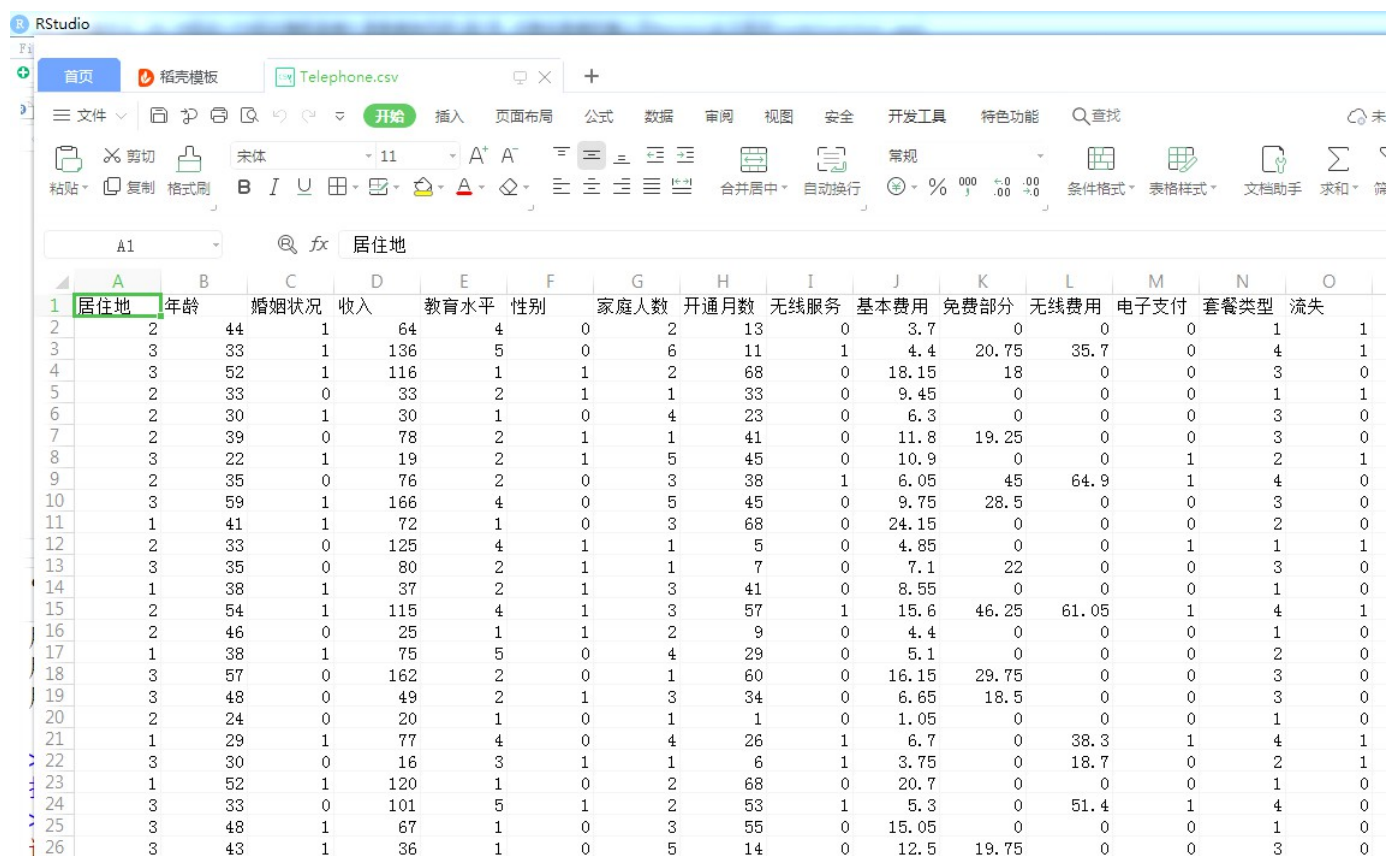
可以看到此数据集参数，nobs=366 train=256 validate=55 test=55，一共有366个样本，可单击“View”或“edit”，对weather数据集进行查看或修改。

修改工作，单击确定后，提示安装RGtk2Extra扩展包。

右下角有一个选择文本文件格式的选项，默认是csv，还有excel文件。

xlsx包依赖于rjava包，需要首先在本地安装好java环境，才能安装rjava和xlsx包。

从剪贴板中导入数据。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	居住地	年龄	婚姻状况	收入	教育水平	性别	家庭人数	开通月数	无线服务	基本费用	免费部分	无线费用	电子支付	套餐类型	流失
2	2	44	1	64	4	0	2	13	0	3.7	0	0	0	1	1
3	3	33	1	136	5	0	6	11	1	4.4	20.75	35.7	0	4	1
4	3	52	1	116	1	1	2	68	0	18.15	18	0	0	3	0
5	2	33	0	33	2	1	1	33	0	9.45	0	0	0	1	1
6	2	30	1	30	1	0	4	23	0	6.3	0	0	0	3	0
7	2	39	0	78	2	1	1	41	0	11.8	19.25	0	0	3	0
8	3	22	1	19	2	1	5	45	0	10.9	0	0	1	2	1
9	2	35	0	76	2	0	3	38	1	6.05	45	64.9	1	4	0
10	3	59	1	166	4	0	5	45	0	9.75	28.5	0	0	3	0
11	1	41	1	72	1	0	3	68	0	24.15	0	0	0	2	0
12	2	33	0	125	4	1	1	5	0	4.85	0	0	1	1	1
13	3	35	0	80	2	1	1	7	0	7.1	22	0	0	3	0
14	1	38	1	37	2	1	3	41	0	8.55	0	0	0	1	0
15	2	54	1	115	4	1	3	57	1	15.6	46.25	61.05	1	4	1
16	2	46	0	25	1	1	2	9	0	4.4	0	0	0	1	0
17	1	38	1	75	5	0	4	29	0	5.1	0	0	0	2	0
18	3	57	0	162	2	0	1	60	0	16.15	29.75	0	0	3	0
19	3	48	0	49	2	1	3	34	0	6.65	18.5	0	0	3	0
20	2	24	0	20	1	0	1	1	0	1.05	0	0	0	1	0
21	1	29	1	77	4	0	4	26	1	6.7	0	38.3	1	4	1
22	3	30	0	16	3	1	1	6	1	3.75	0	18.7	0	2	1
23	1	52	1	120	1	0	2	68	0	20.7	0	0	0	1	0
24	3	33	0	101	5	1	2	53	1	5.3	0	51.4	1	4	0
25	3	48	1	67	1	0	3	55	0	15.05	0	0	0	1	0
26	3	43	1	36	1	0	5	14	0	12.5	19.75	0	0	3	0

```
actionuser <- read.table("clipboard",header = T)
```

```
## Warning in read.table("clipboard", header = T): incomplete final line found by  
## readTableHeader on 'clipboard'
```

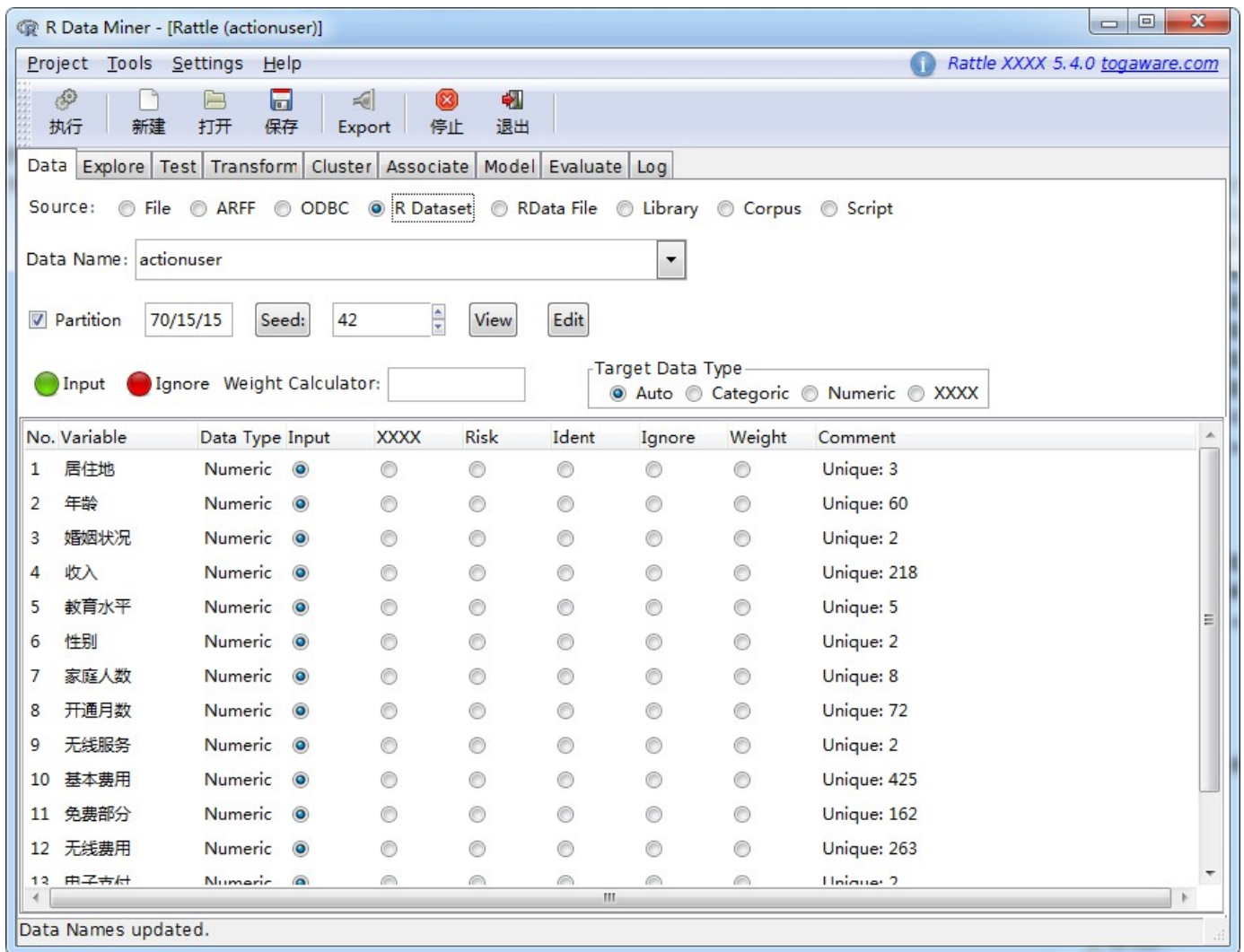
```
dim(actionuser)
```

```
## [1] 0 1
```

```
head(actionuser)
```

```
## [1] hierarchical.png  
## <0 rows> (or 0-length row.names)
```

接下来就可以在R Dataset的Data Name中选择数据对象“actionuser”，然后单击“执行”按钮就可以将actionuser数据导入到Rattle中。



library中自带了很多数据集，也可以导入分析。

探索数据

Explore选项，输出数据总体概况，数据分布情况，数据的相互关系矩阵，数据集的主成分分析，各变量之间的交互作用。

R Data Miner - [Rattle (actionuser)]

Project Tools Settings Help Rattle XXXX 5.4.0 togaware.com

执行 新建 打开 保存 Export 停止 退出

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Summary ☐ Distributions ☐ Correlation ☐ Principal Components ☐ Interactive

☒ Summary ☐ Describe ☐ Basics ☐ Kurtosis ☐ Skewness ☐ Show Missing ☐ Cross Tab

Storage
 居住地 integer
 年龄 integer
 婚姻状况 integer
 收入 integer
 教育水平 integer
 性别 integer
 家庭人数 integer
 开通月数 integer
 无线服务 integer
 基本费用 double
 免费部分 double
 无线费用 double
 电子支付 integer
 套餐类型 integer
 流失 integer

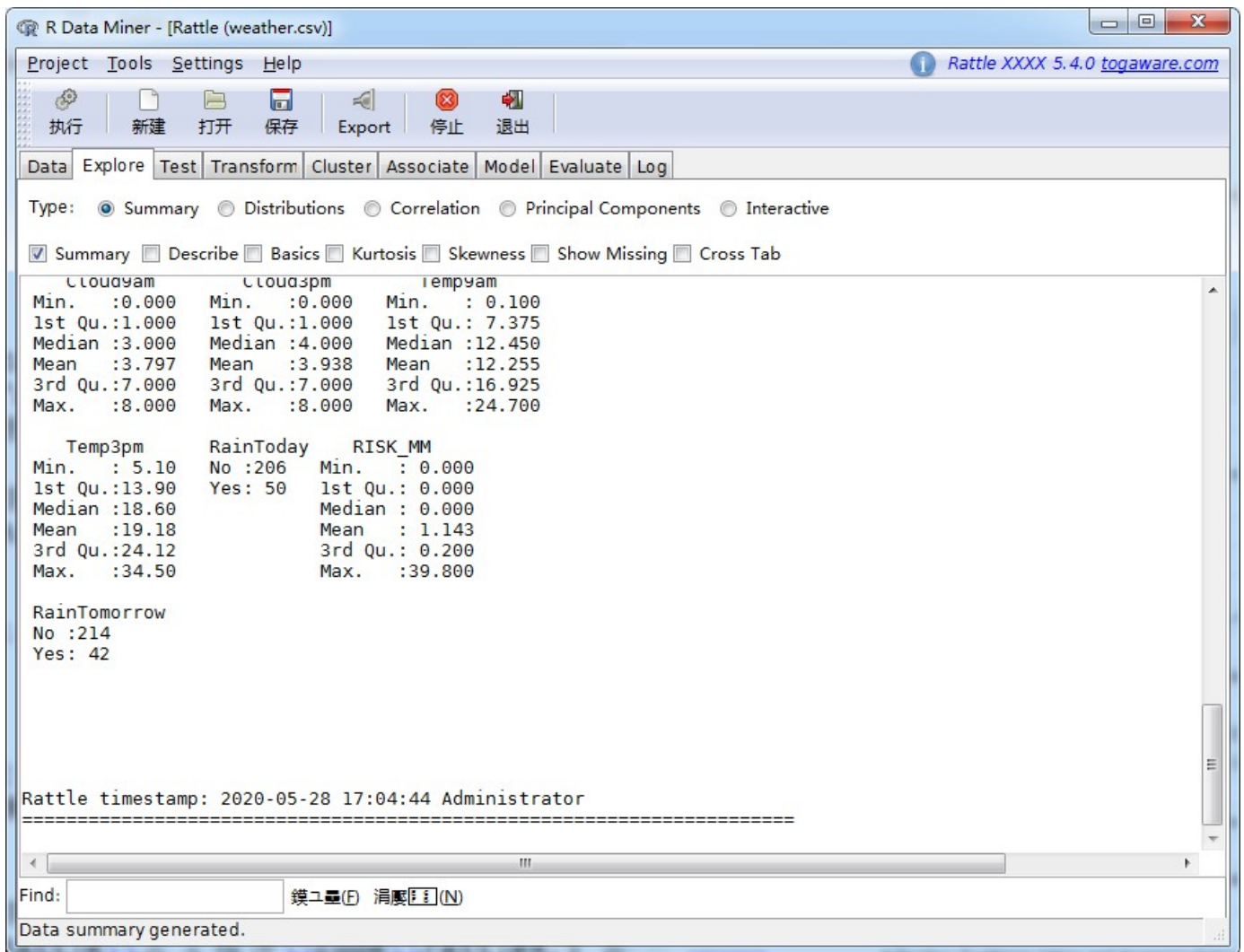
For the simple distribution tables below the 1st and 3rd Qu. refer to the first and third quartiles, indicating that 25% of the observations have values of that variable which are less than or greater than (respectively) the value listed.

居住地	年龄	婚姻状况
Min. :1.000	Min. :18.00	Min. :0.00
1st Qu.:1.000	1st Qu.:32.00	1st Qu.:0.00
Median :2.000	Median :40.00	Median :0.00
Mean :2.037	Mean :41.58	Mean :0.49
3rd Qu.:3.000	3rd Qu.:51.00	3rd Qu.:1.00

Find: 模式(F) 消费(F)(N)

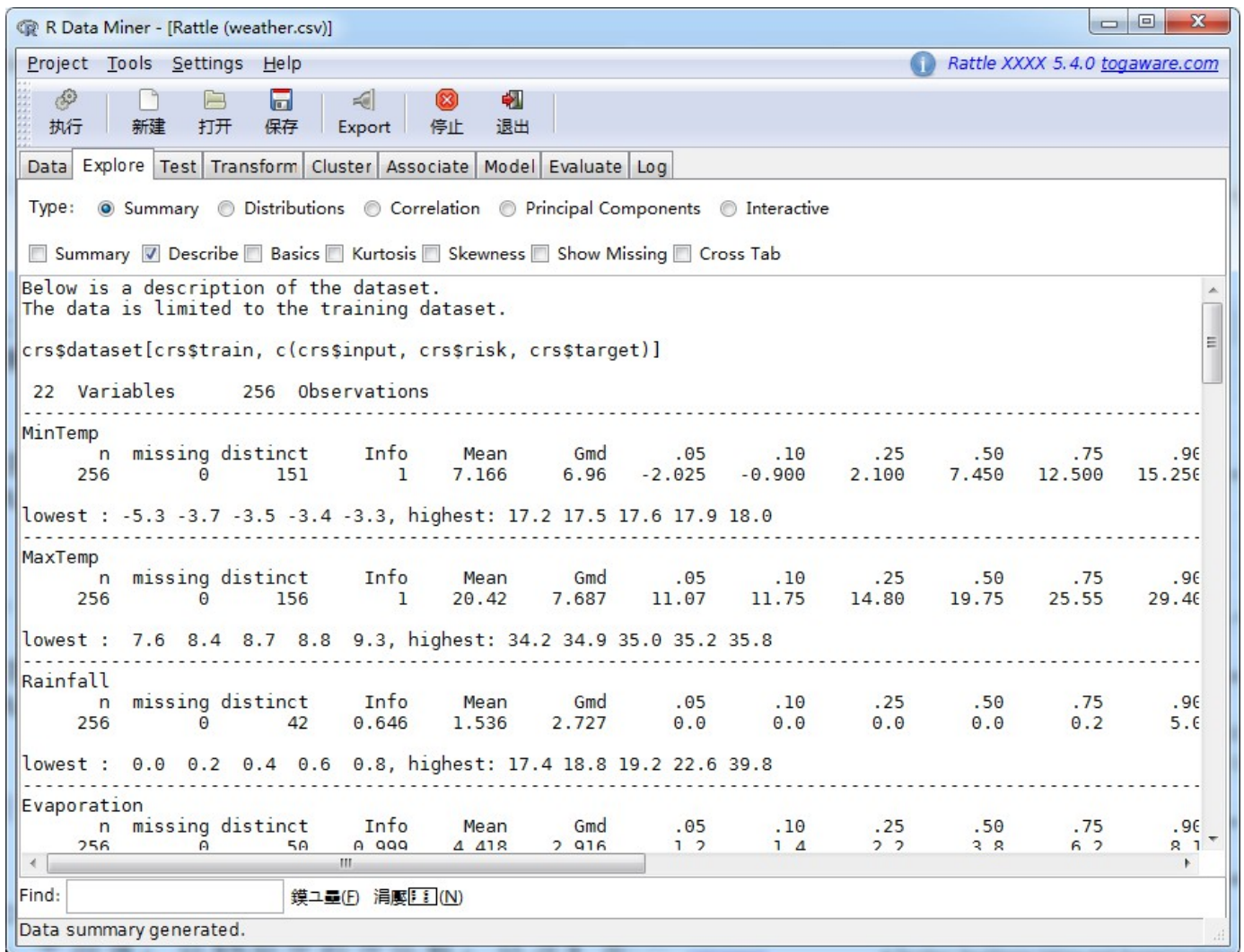
Data summary generated.

对天气数据进行分析：



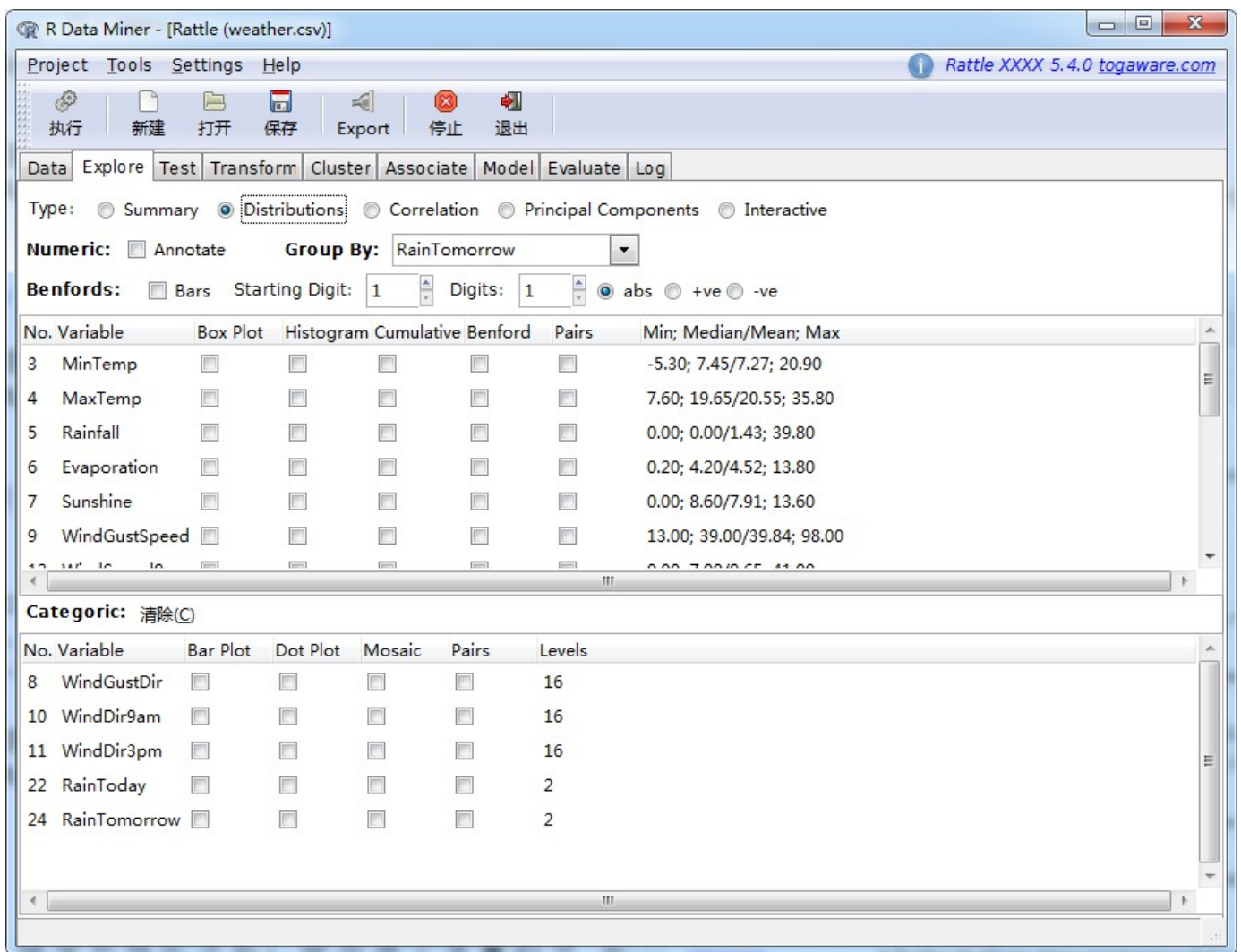
发现RainTomorrow-明天是否会下雨，有215天是晴天，41天是雨天(yes).

利用Hmisc包中的describe函数返回变量和观测的数量，缺失值和唯一值的数目，平均值，以5%划分的分位数，以及5个较大的值和5个较小的值。

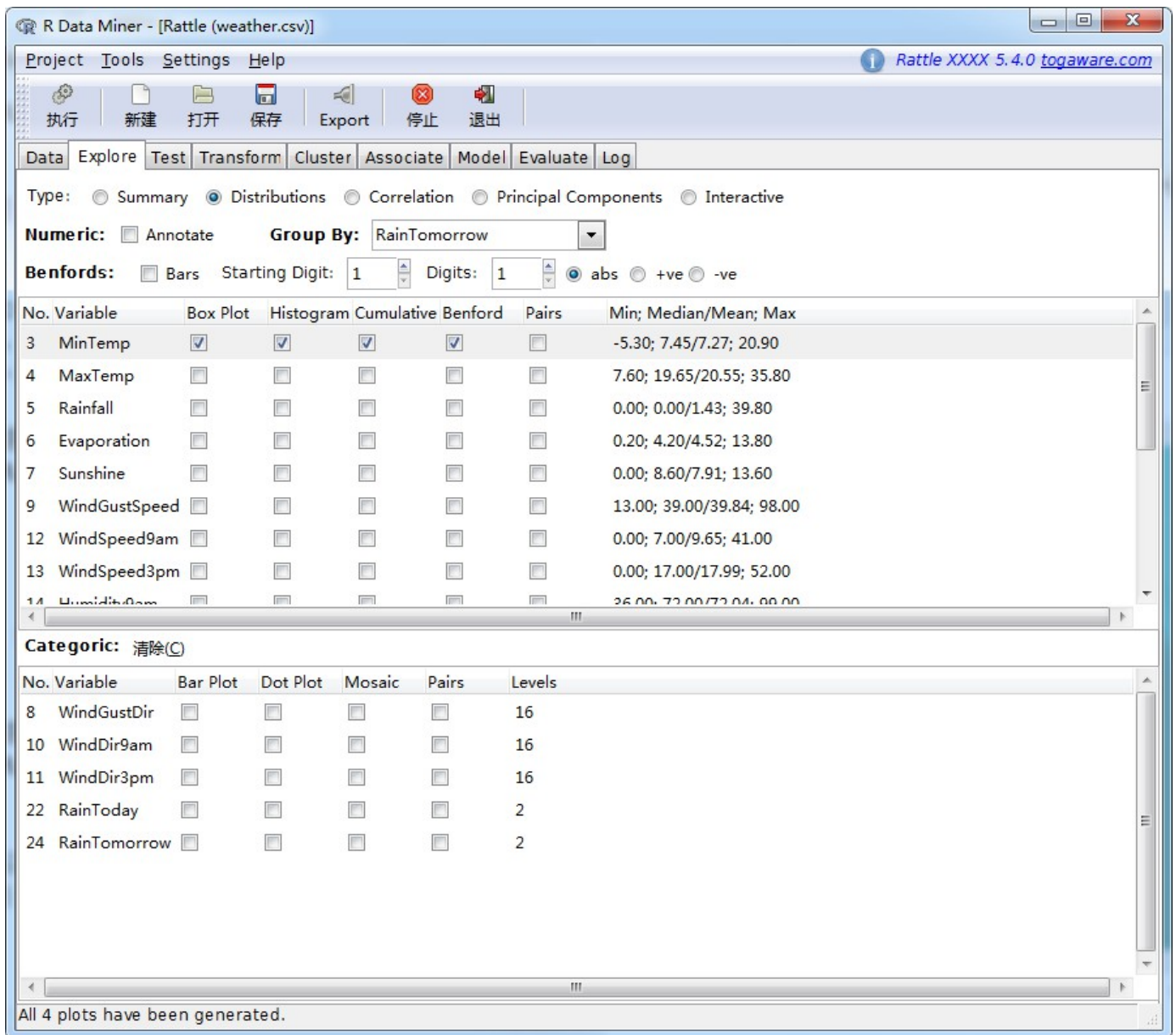


数据分布探索

Distributions选项，结合可视化方式，给出各个变量的分布特征。

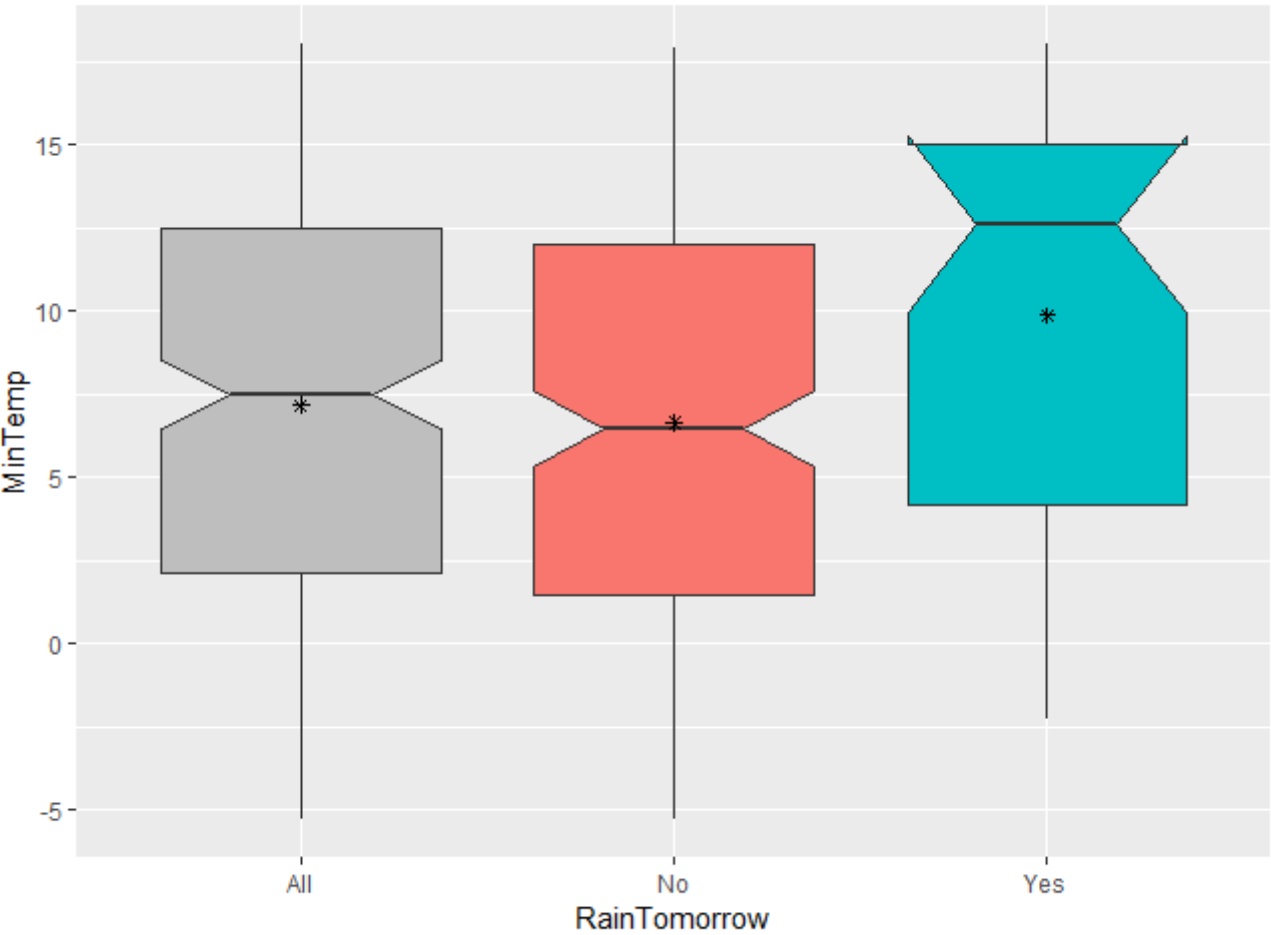


对于数值型变量，可以绘制箱线图，直方图，累积分布图和Benford图，对于类别变量，可以绘制条形图，点图，马赛克图。

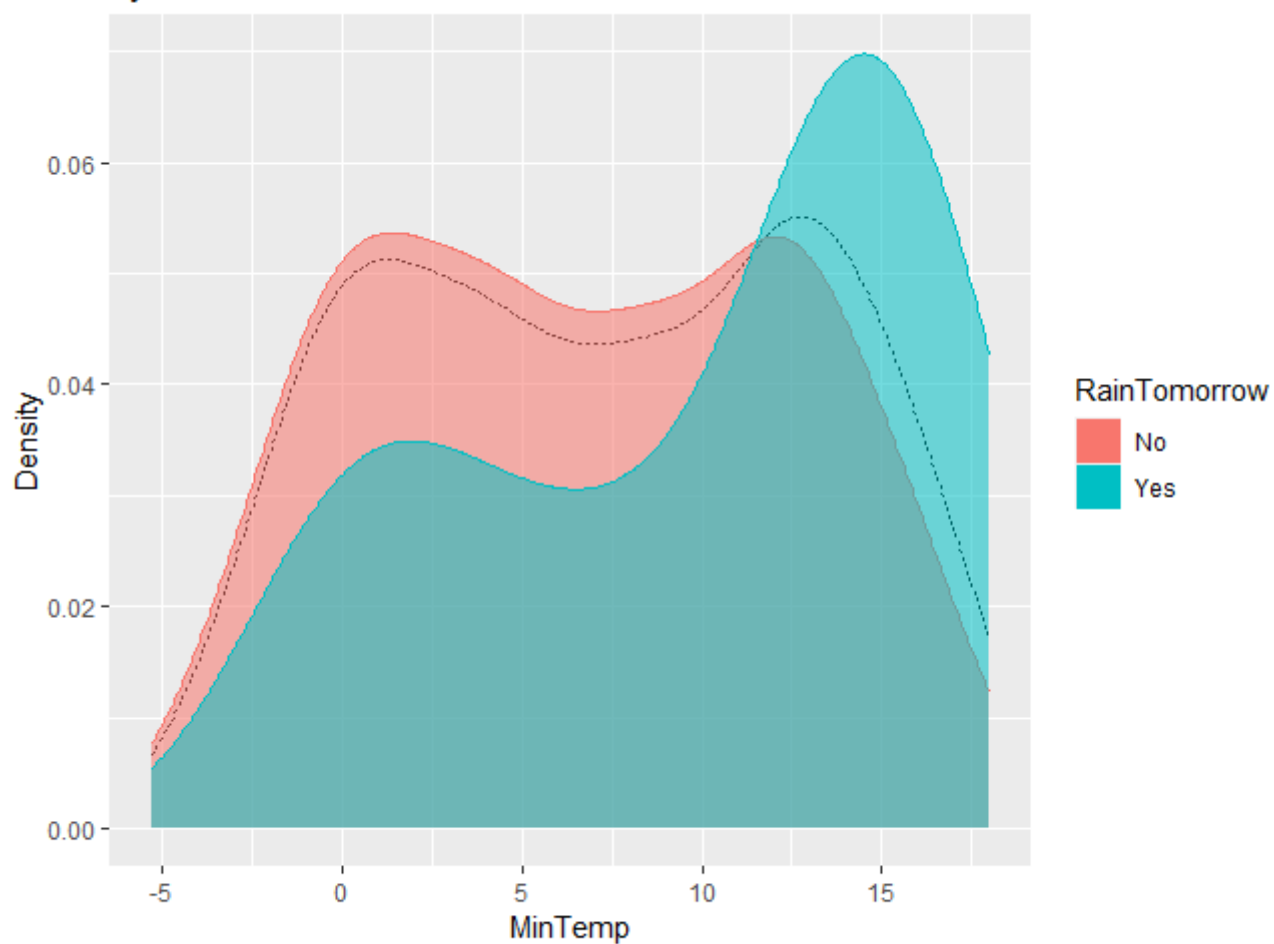


以weather数据集，RainTomorrow为分组变量，绘制出MinTemp变量的箱线图，直方图，累积分布图和Benford图。

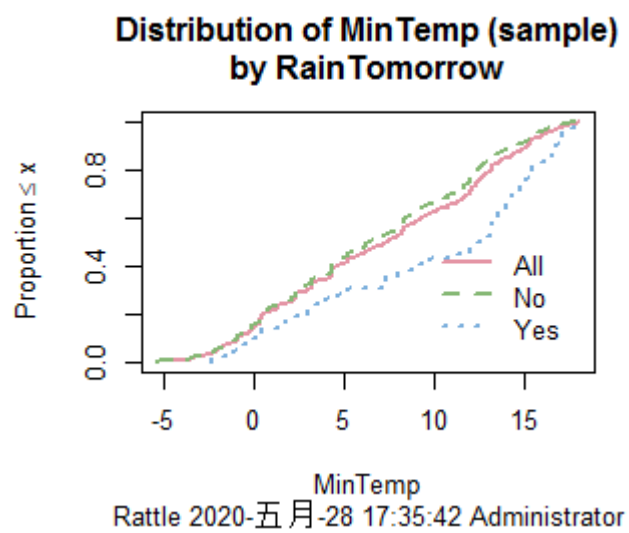
Distribution of MinTemp (sample)
by RainTomorrow



Distribution of MinTemp (sample)
by RainTomorrow

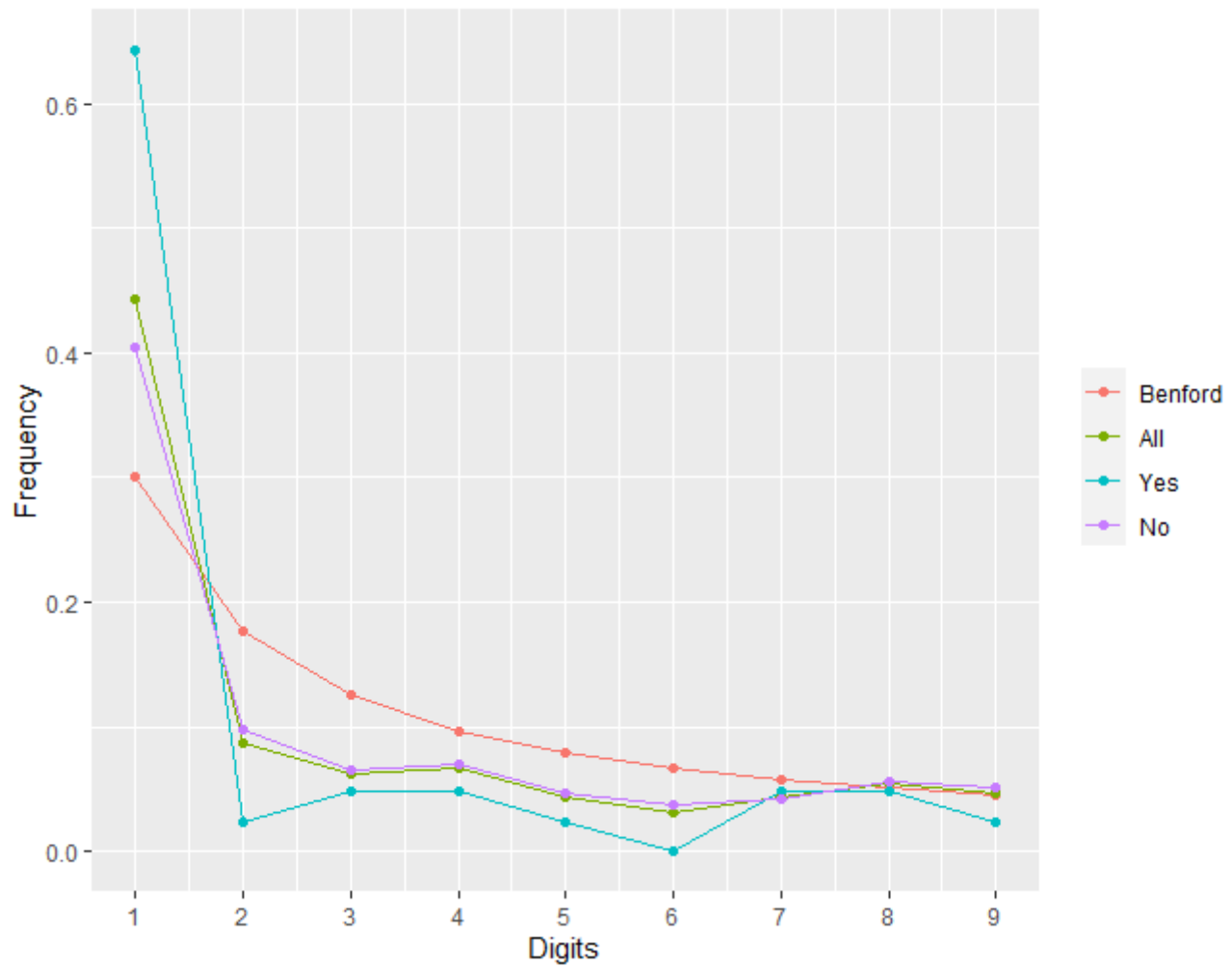


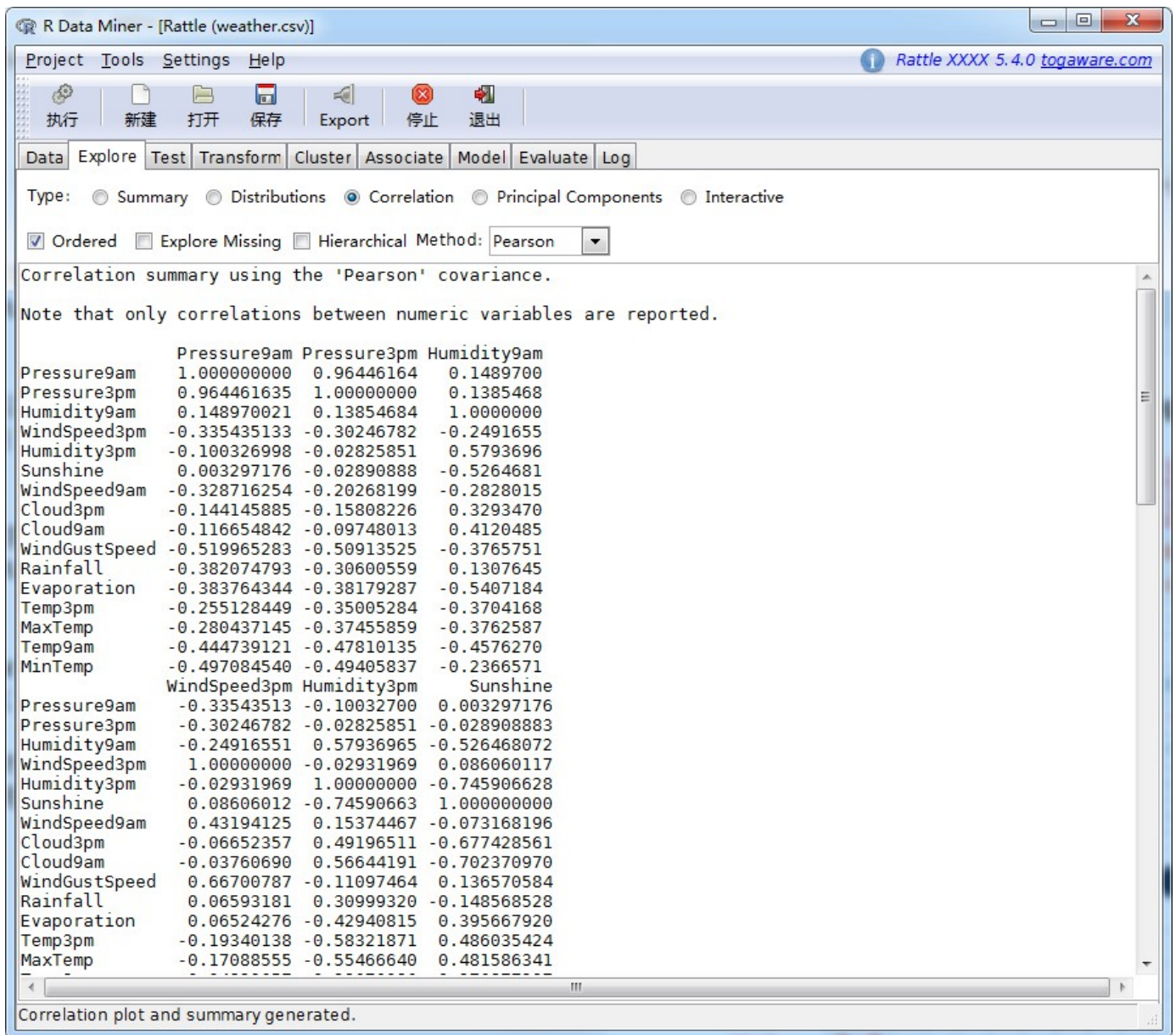
Rattle 2020-五月-28 17:35:41 Administrator



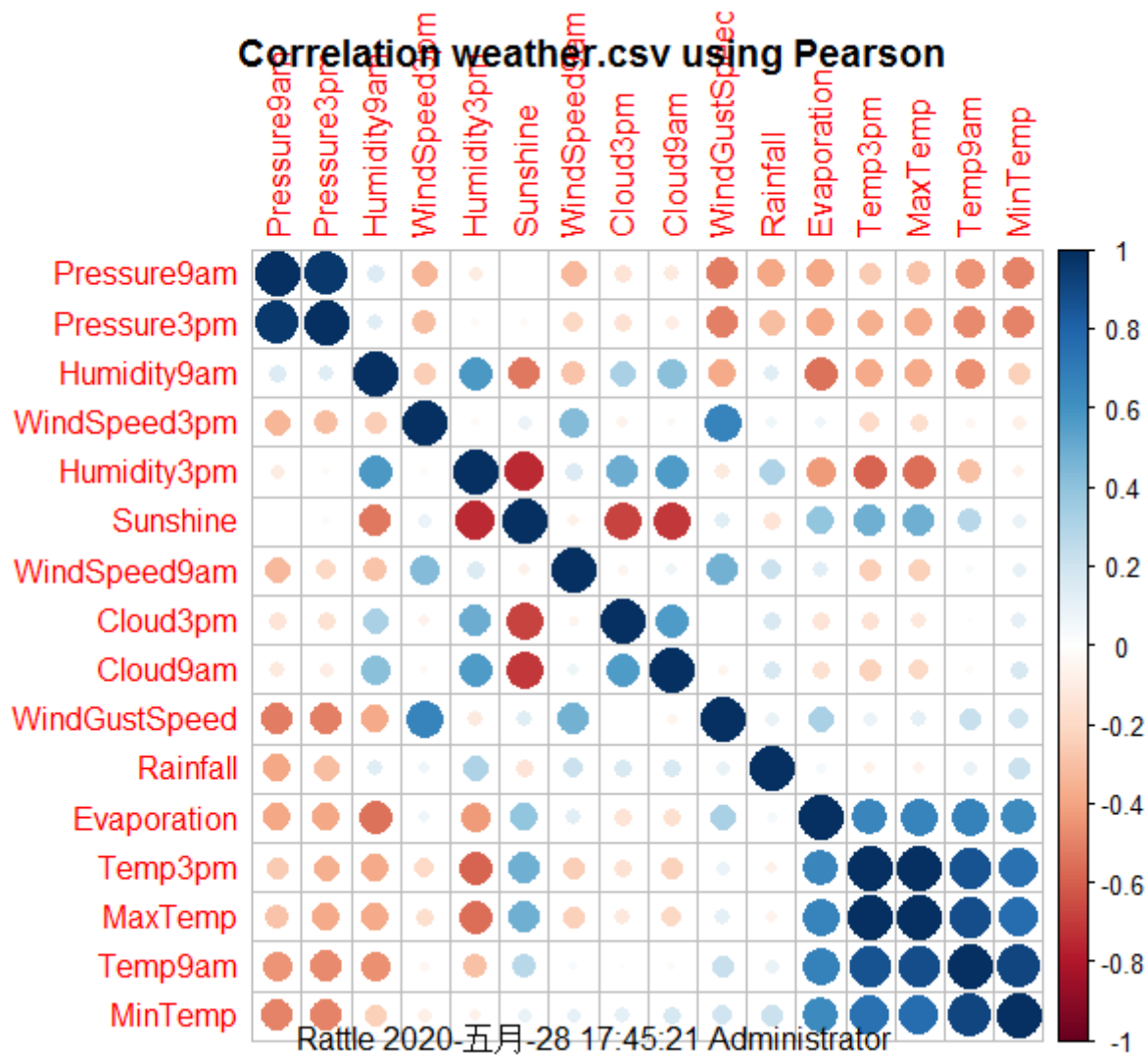
当RainTomorrow为No时，数据处于右偏的正偏态分布状态。直方图中也能看到趋势。

Digital Analysis of First Digit
of MinTemp by RainTomorrow





计算出数值型变量的相关系数。



可视化输出在Rstudio的plots窗口中。

交互图

GGobi和GGRaptR两种方法以交互式方式探索数据。需要安装GGobi软件以及相应的rggobi包。

到<http://www.ggobi.org/downloads/ggobi-2.1.8.exe> ,

下载最新的ggobi并安装，我安装在D:/Program Files/ggobi目录。安装后点击桌面的图标，就能够使用GGobi了。

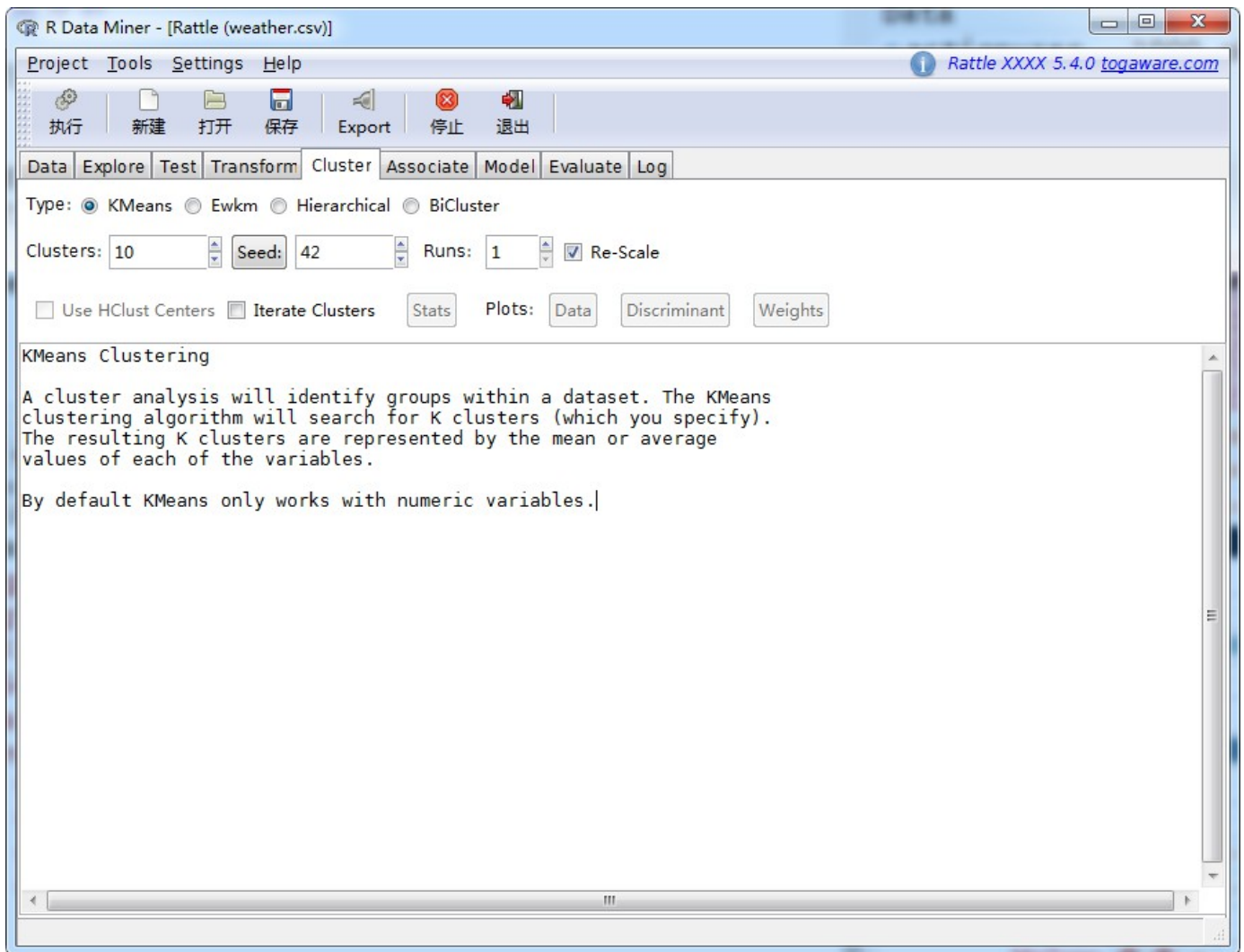
Rggobi只是GGobi在R中的一个接口。这需要在R中安装一些packages。

聚类分析

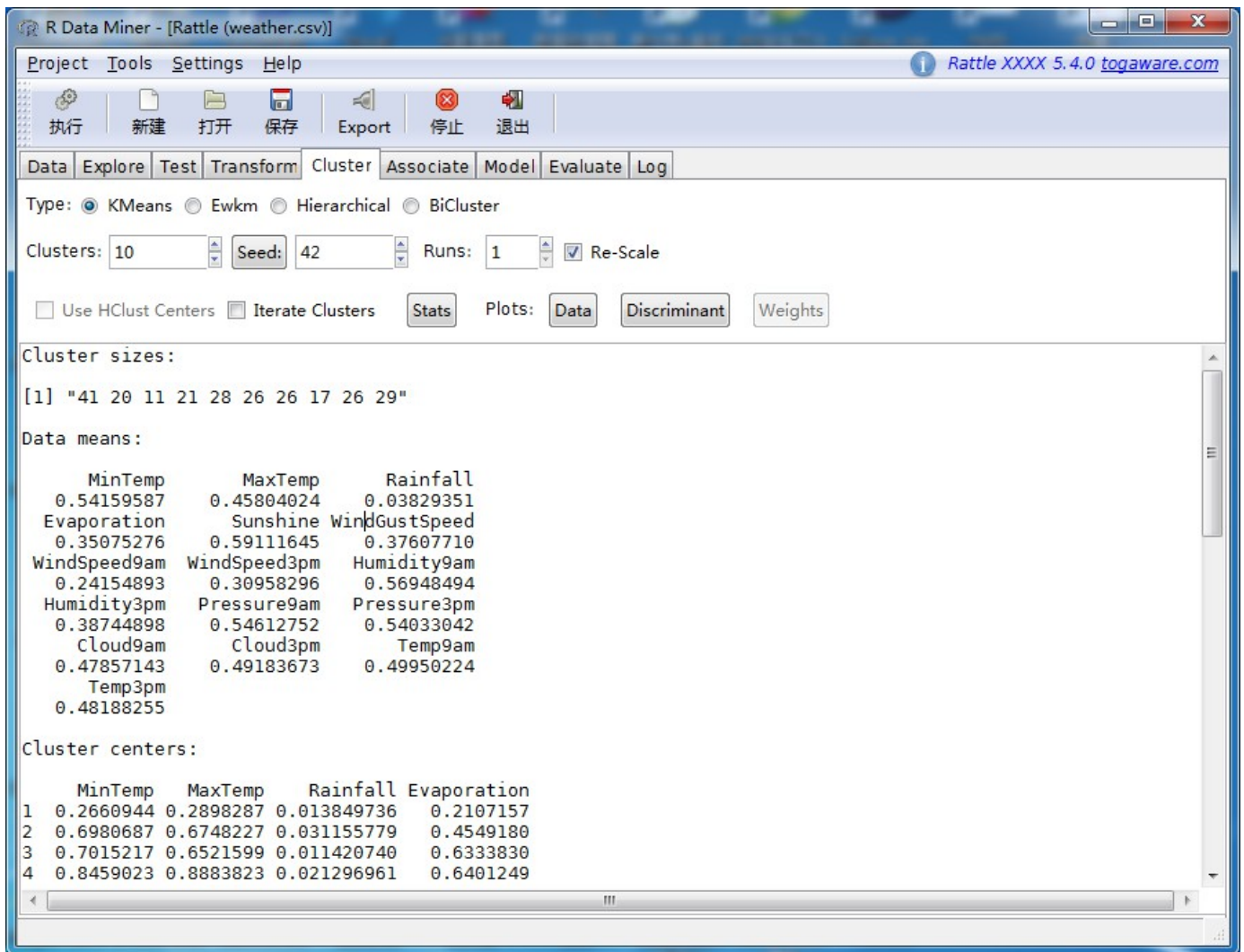
将观测对象的群体按照相似性和相异性进行不同的群组划分。

聚类算法种类很多，Rattle可以实现最常用的k-means聚类和层次聚类。

Rattle通过Cluster选项可以建立k-means聚类和层次聚类。



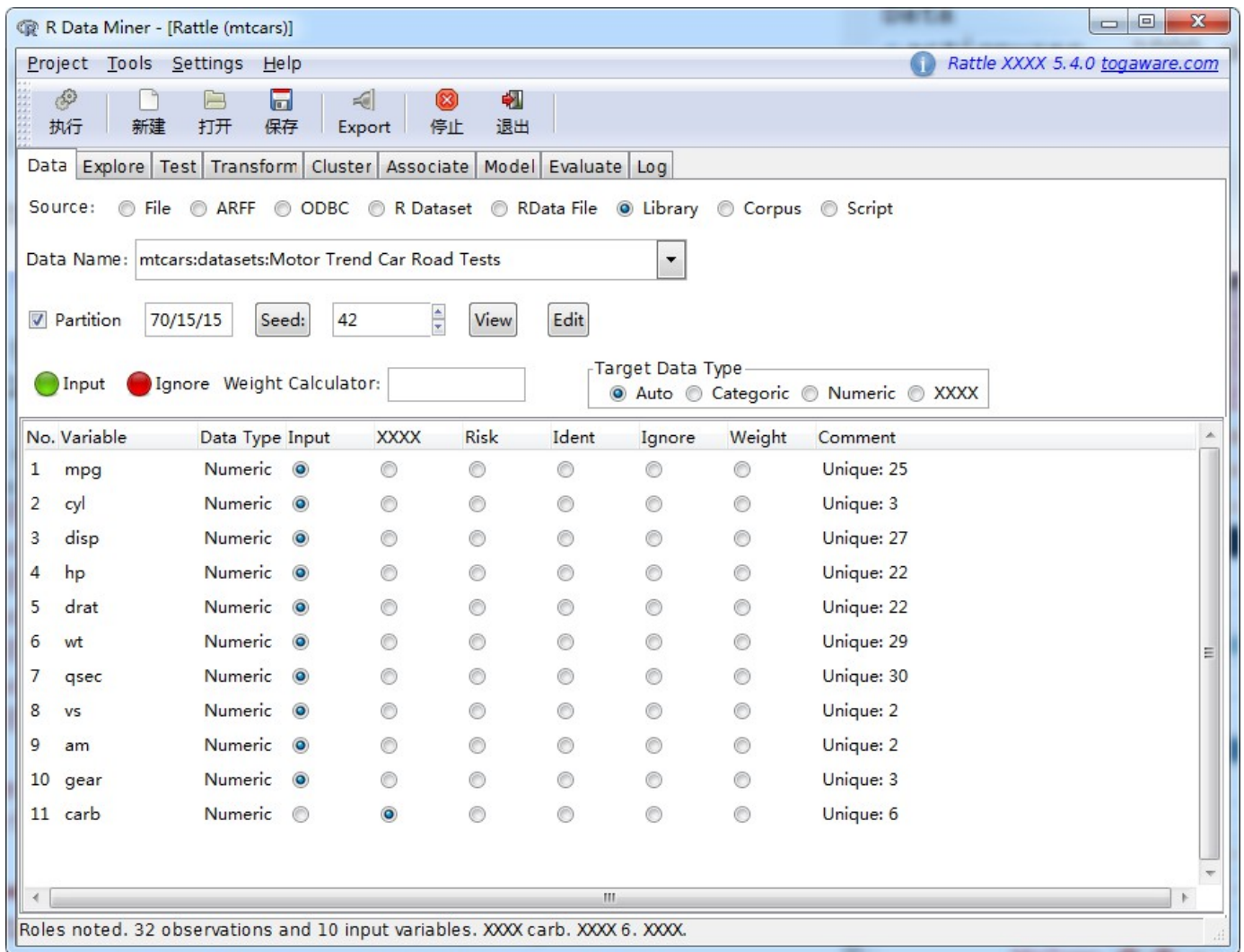
单击“执行”，模型结果会先后输出各类别所包含的样本数(cluster sizes)，训练数据集各变量的均值(data means)，各类别的均值(cluster centers)，各类别的组内平方和。



单击“data”按钮会打印出前5个数值变量 的散点图矩阵，用不同颜色区分不同类别的样本。

```
#出现错误
Error in eval(parse(text = top.vars.cmd)) :
  无法改变被锁定的联编'vars' 的值
```

单击“Discriminant”按钮，则会生成样本投影图。

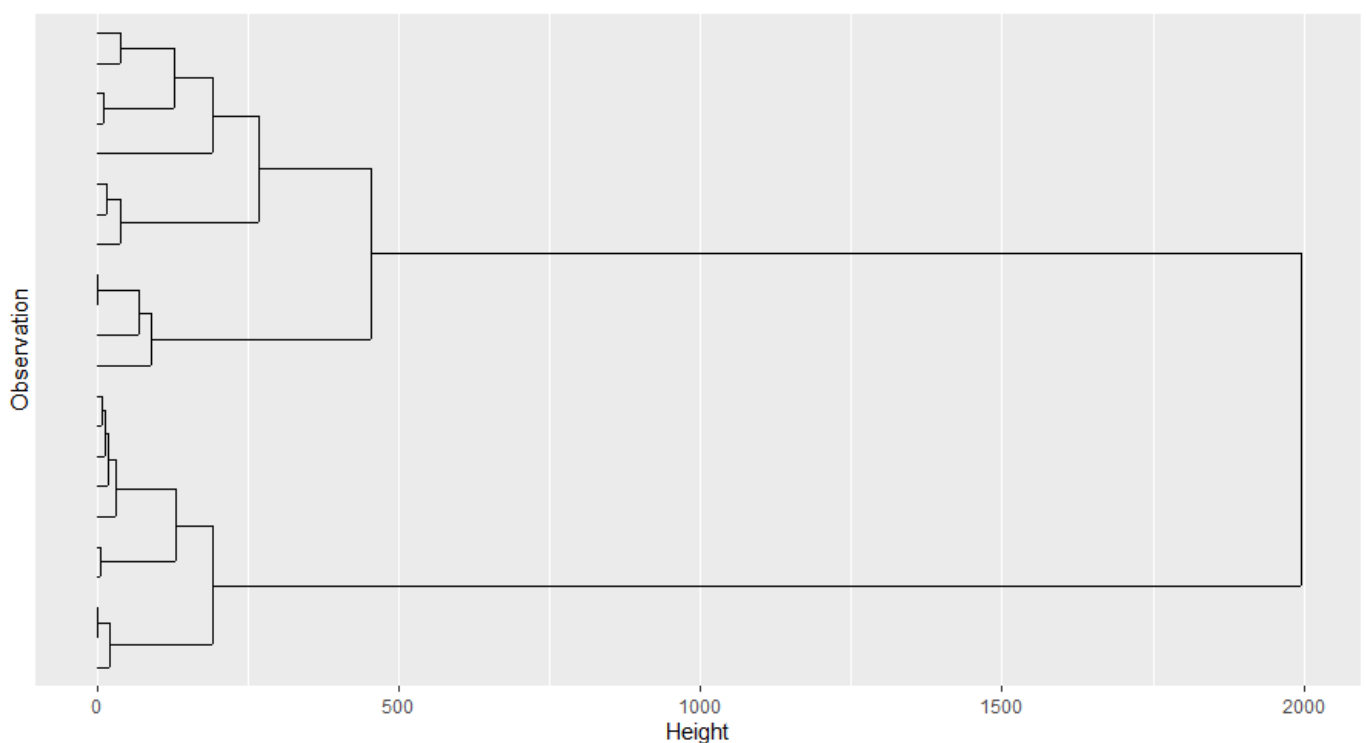


单击“Hierarchical”，执行，层次聚类模型。

单击“Data Plot”按钮生成数值变量的散点图矩阵，单击“Discimininant”按钮生成投影图，单击“Dendrogram”按钮生成系统聚类树图。

Cluster Dendrogram mtcars

Rattle 2020-五月-28 21:03:51 Administrator



关联规则

Apriori算法。

略。

决策树

略

随机森林

略

模型评估

略

混淆矩阵

略

风险图

略

ROC

参考文献

1、林智章，张良均，R语言编程基础，人民邮电出版社，2019年1月。