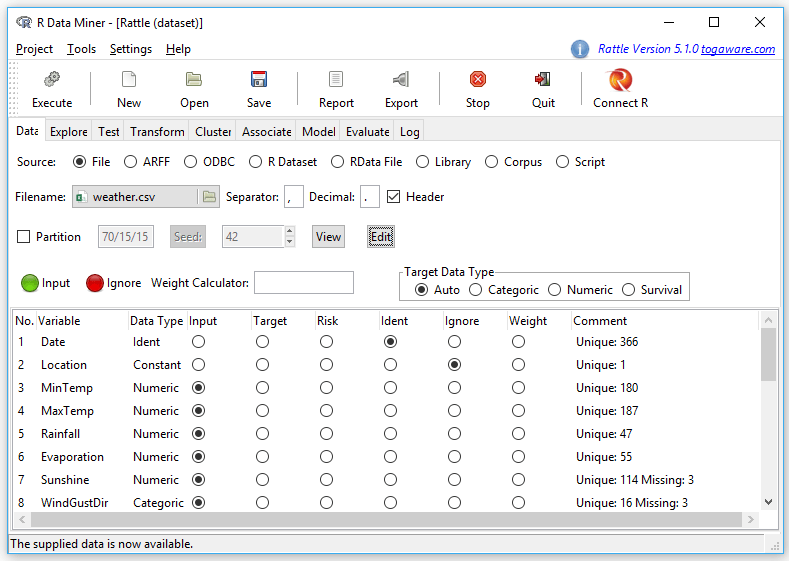
**Experiment 8: Explore, visualise, transform and summarise input datasets for building classification models.**

**Requirement:** Laptop or Desktop with Python installed

**Theory**

Rattle is a popular GUI for data mining using R. It presents statistical and visual summaries of data, transforms data so that it can be readily modelled, builds both unsupervised and supervised machine learning models from the data, presents the performance of models graphically, and scores new datasets for deployment into production. A key features is that all of your interactions through the graphical user interface are captured as an R script that can be readily executed in R independently of the Rattle interface. Use it as a tool to learn and develop your skills in R and then to build your initial models in Rattle to then be tuned in R which provides considerably more powerful options.



Since R GUIs are using R to do the work behind the scenes, they often include the ability to read a wide range of files, including SAS, SPSS, and Stata. Some, like BlueSky Statistics, also include the ability to read directly from SQL databases. Of course you can always use R code to import data from any source and then continue to analyze it using any GUI, but the point of GUIs is to avoid programming.

Rattle skips many common statistical data formats, but it includes a couple exclusive ones, such as the Attribute-Relation File Format used by other data mining tools. It also includes “corpus” which reads in text documents, and it then it performs the popular tf-idfcalculation to prepare them for analysis using the other numerically-based analysis methods.

On its “Data” tab, Rattle offers several formats:

* File: CSV
* File: TXT
* File: Excel
* Attribute-Relation File Format (ARFF)
* Open Database Connectivity (ODBC)
* R Dataset
* RData File
* Library
* Corpus (for text analysis)
* Script

**Data Management**

It’s often said that 80% of data analysis time is spent preparing the data. Variables need to be transformed, recoded, or created; strings and dates need to be manipulated; missing values need to be handled; datasets need to be stacked or merged, aggregated, transposed, or reshaped (e.g. from wide to long and back). A critically important aspect of data management is the ability to transform many variables at once. For example, social scientists need to recode many survey items, biologists need to take the logarithms of many variables. Doing these types of tasks one variable at a time can be tedious. Some GUIs, such as jamovi and RKWard handle only a few of these functions. Others, such as BlueSky Statistics or the R Commander can handle all, or nearly all, of these tasks.

Rattle provides minimal data management tools. Its designer chose to focus on reading a single data set, and making transformations that are common in data mining projects quick and easy. More complex data management tasks are left to other tools such as SQL in a database before the data set is read in, or using R programming.

**Graphics**

The various GUIs available for R handle graphics in several ways. Some, such as RKWard, focus on R’s built-in graphics. Others, such as BlueSky Statistics, focus more on graphics done by the popular ggplot2 package.

GUIs also differ quite a lot in how they control the style of the graphs they generate. Ideally, you could set the style once, and then all graphs would follow it. That’s how jamovi works, but then jamovi is limited to its custom graph functions, as nice as they may be.

Rattle uses a wide variety of packages to create graphics (e.g. Figure 6), so their appearance differs and there is no way to standardize their style within Rattle. You could use R code to change their look or style by editing the R code in the “Log” tab and submitting it in an R console.

The types of graphs that Rattle provides follow its focus on data mining, and you create them in the data mining process flow:

* Boxplots
* Histogram / Density
* Cumulative
* Benford
* Pairs (correlation plot)
* Hierarchical (dendogram of correlations)
* Missing Values
* Principal Components Importance Barplot
* Principal Components Biplot
* Cluster Data (scatterplots of clusters)
* Cluster Discriminant Coordinates
* Cluster Weights Heatmap
* Associate Frequency Plot
* Associate Rule Plot
* Decision Tree Plot
* R’s plot Function on Generalized Linear Models
* Risk Chart
* Cost Curve
* Hand Curve
* Lift Plot
* ROC Plot
* Precision Plot
* Sensitivity Plot
* Predicted vs. Observed

Rattle doesn’t include a graphic display window, so the graphs appear in window provided by the R software that you use to start Rattle.