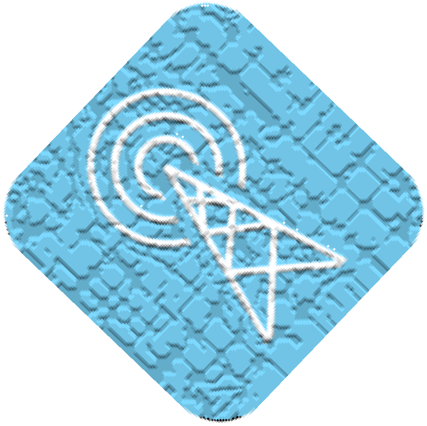
Data Science Using R

Lesson06–Introduction to R Markdown and Rattle

Objective

After completing this lesson you will be able to:



- Describe R Markdown and Rattle
- Build a basic R Markdown document
- Explain the various features of Rattle
- Run a dataset in Rattle through a set of commonly used techniques of data analysis.

R Markdown–An Introduction

- R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R.
- R markdown can be used to create reports in the following format:

Report Format	Output Format
Document	HTML, PDF, WORD
Presentation	HTML(ioslides), HTML(Slidy), PDF(Beamer)
Interactive Shiny Report	Shiny Document, Shiny Presentation

- R Markdown documents can be automatically regenerated whenever underlying R code or data changes.

R Markdown–Install Package

- The first step to use R markdown is to install the package.

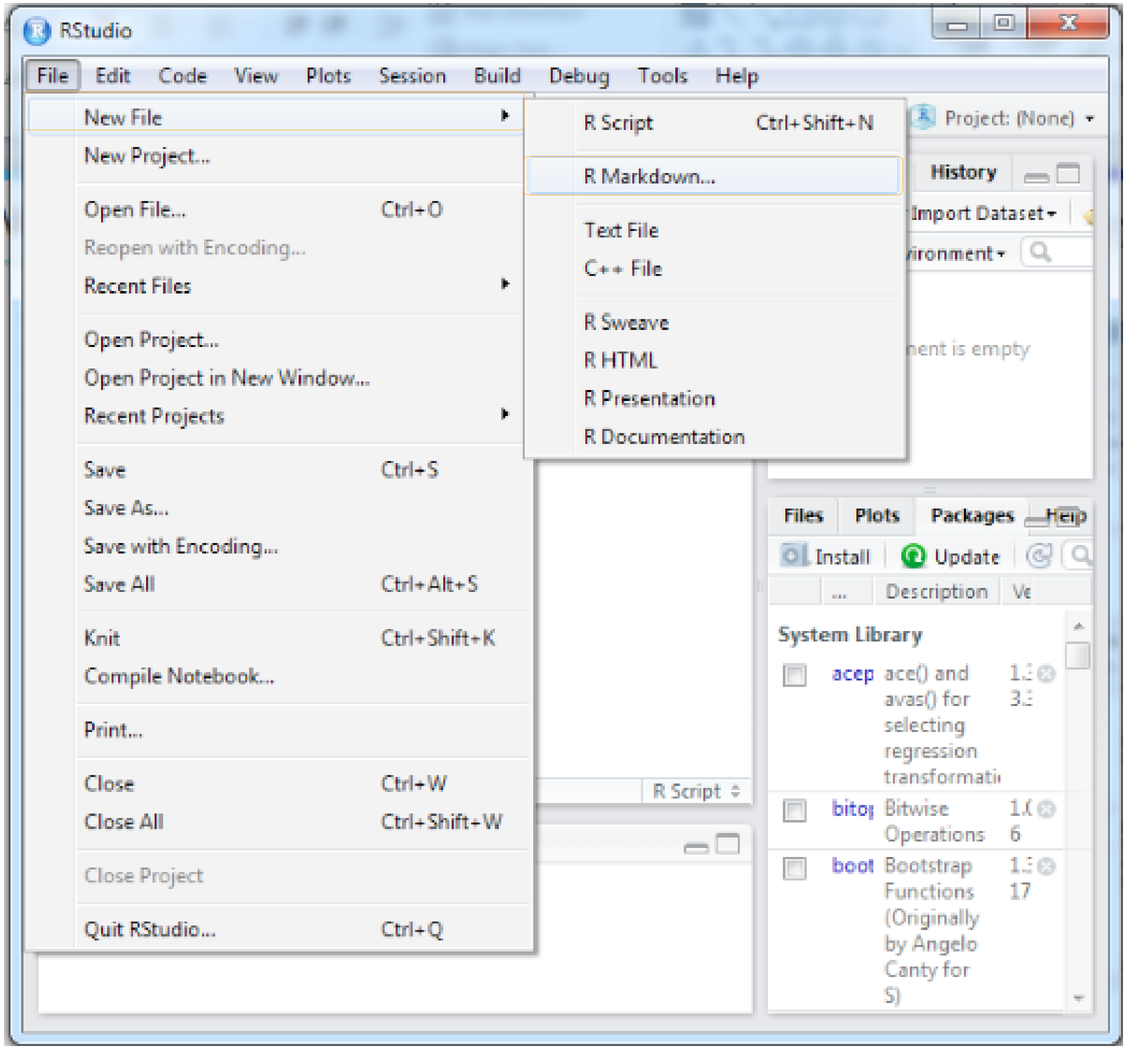
On R Studio Console:

```
>Install.packages(“rmarkdown”)
```

Or install using the Rstudio Install packages options

Working with R Markdown

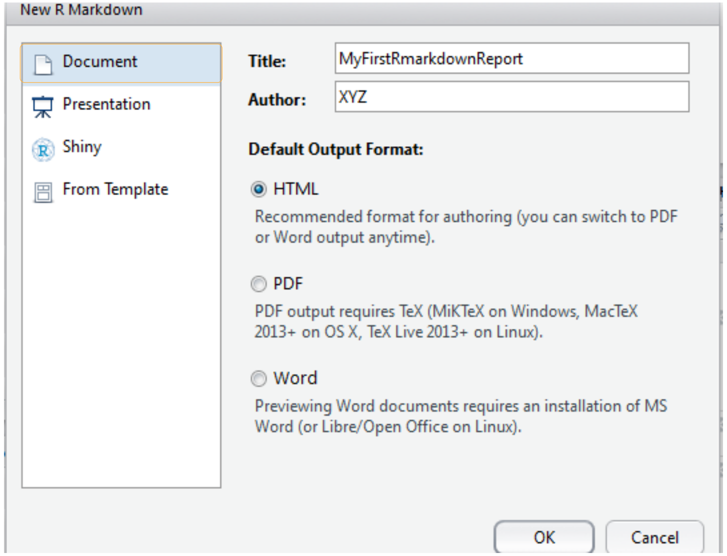
- Open a new R markdown file from the R Studio file option.



Working with R Markdown

- Select the type of report from the window that follows.

- *Select 'Document' as the report type if creating an HTML, PDF or Word document.*
- *Select 'Presentation' as the report type if creating HTML or PDF presentation.*
- *Select 'Shiny' as the report type if creating an interactive shiny report.*
- *There are specific templates which can be picked up to create report.*


New R Markdown

Title: MyFirstRmarkdownReport

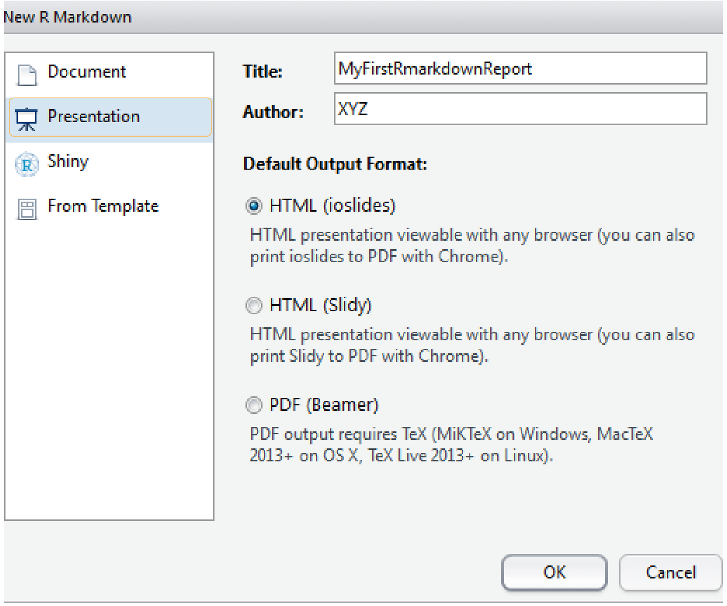
Author: XYZ

Default Output Format:

● HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).


New R Markdown

Title: MyFirstRmarkdownReport

Author: XYZ

Default Output Format:

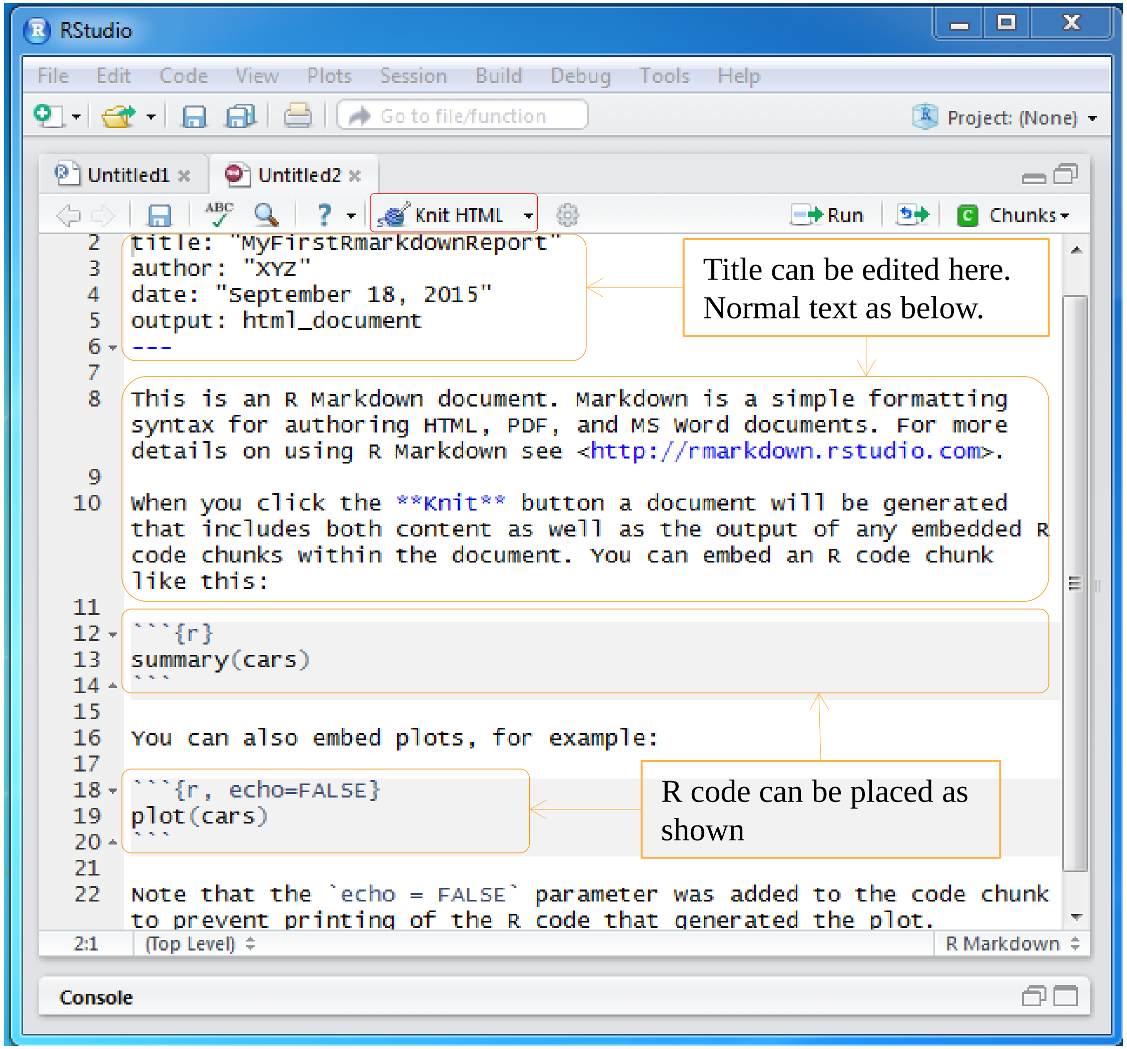
● HTML (ioslides)
HTML presentation viewable with any browser (you can also print ioslides to PDF with Chrome).

○ HTML (Slidy)
HTML presentation viewable with any browser (you can also print Slidy to PDF with Chrome).

○ PDF (Beamer)
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

My First R Markdown Code

- The R markdown code structure is simple to follow.
- Click on the Knit HTML icon to save the file.



~~~
title: "MyFirstRmarkdownReport"
author: "XYZ"
date: "September 18, 2015"
output: html_document
---

This is an R Markdown document. Markdown is a simple formatting
syntax for authoring HTML, PDF, and MS Word documents. For more
details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated
that includes both content as well as the output of any embedded R
code chunks within the document. You can embed an R code chunk
like this:

```{r}
summary(cars)
~~~

You can also embed plots, for example:

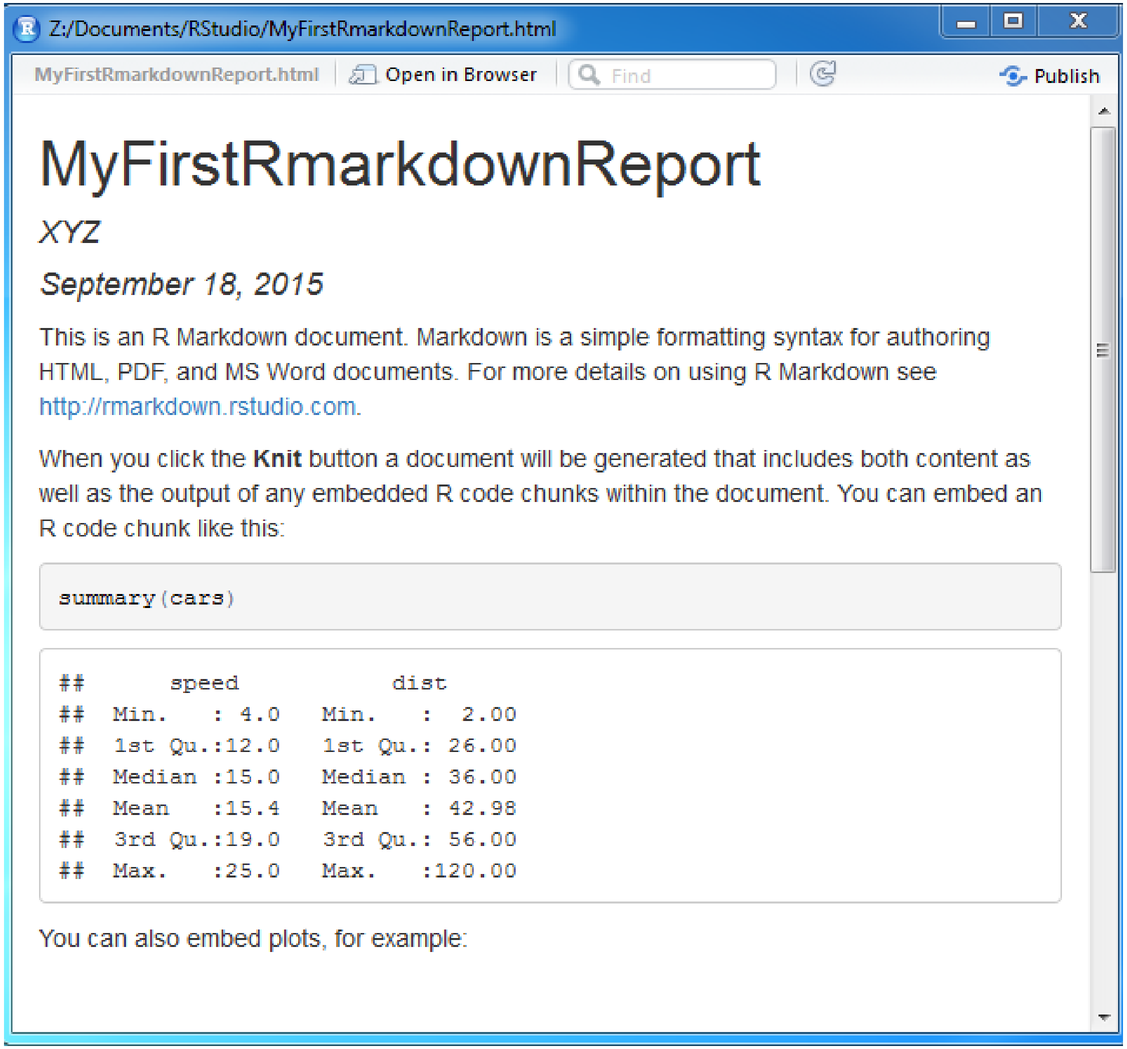
```{r, echo=FALSE}
plot(cars)
```

Note that the `echo = FALSE` parameter was added to the code chunk
to prevent printing of the R code that generated the plot.
```

Title can be edited here. Normal text as below.

R code can be placed as shown

- File gets saved with '.Rmd' extension in the current working directory.
- Report can be opened up in a separate window or inside the R Studio viewer.
```

# My First R Markdown Report

- The report will look like a formatted report.

- Very sophisticated formatting can be applied on the text including writing equations, hyperlinks, appending images etc.


Z:/Documents/RStudio/MyFirstRmarkdownReport.html
MyFirstRmarkdownReport.html
Open in Browser
Find
Publish

# MyFirstRmarkdownReport

*XYZ*

*September 18, 2015*

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

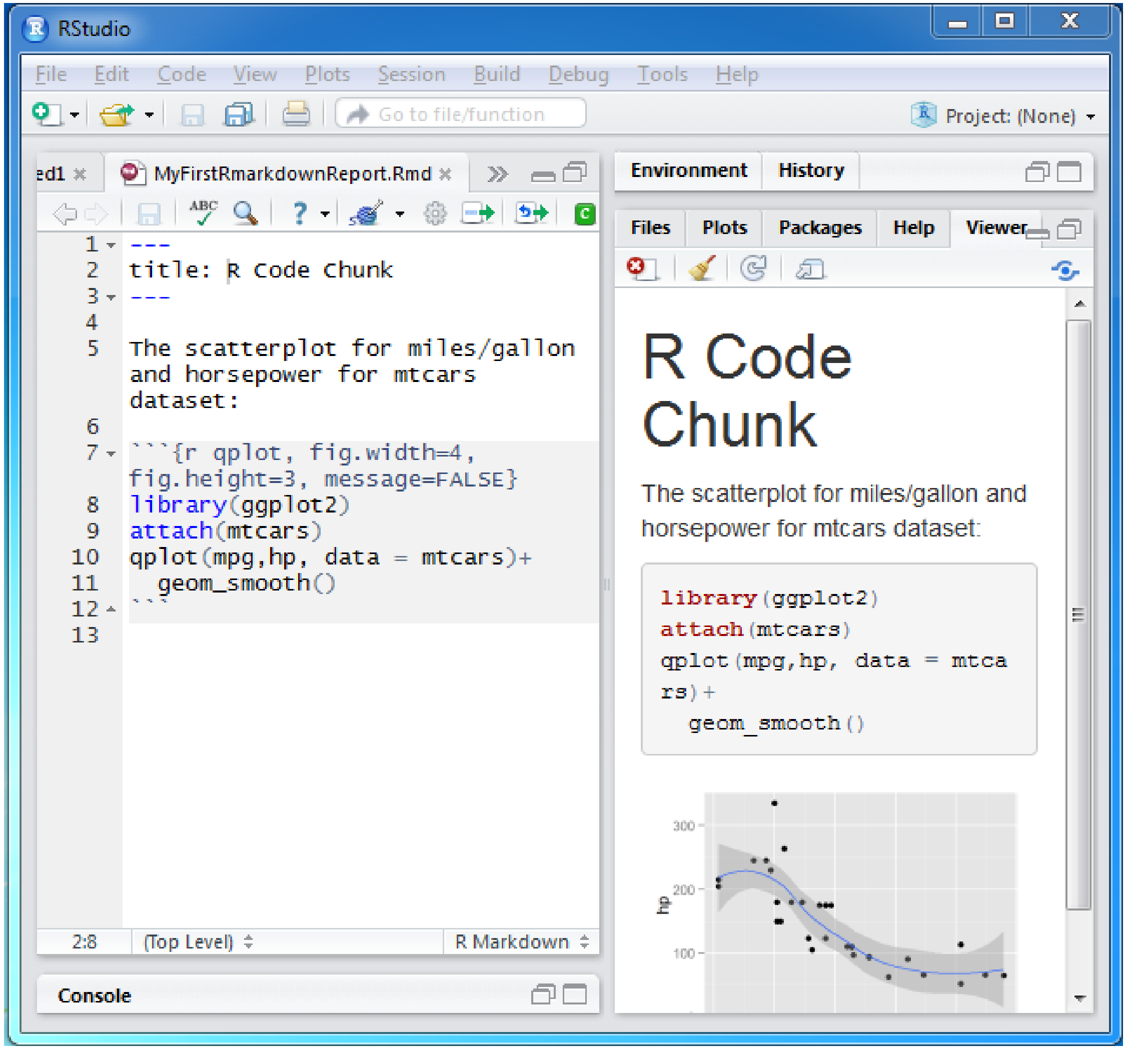
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

You can also embed plots, for example:

R Markdown Code and Viewer

- The R code and viewer can be used side by side as a regular R scripting tool.
- The code for scatter plot and resulting output in the viewer is depicted here.

More on Rmarkdown at:
http://rmarkdown.rstudio.com/



~~~
---
title: R Code Chunk
---

The scatterplot for miles/gallon
and horsepower for mtcars
dataset:

```{r qplot, fig.width=4,
fig.height=3, message=FALSE}
library(ggplot2)
attach(mtcars)
qplot(mpg,hp, data = mtcars)+
  geom_smooth()
```
~~~

R Code Chunk

The scatterplot for miles/gallon and horsepower for mtcars dataset:

```
library(ggplot2)
attach(mtcars)
qplot(mpg,hp, data = mtca
rs)+
  geom_smooth()
```

Demo of the RMarkdown using an example dataset.

Rattle–An Introduction

- R Analytical Tool to Learn Easily (Rattle) is a user interface based data mining tool built on top of R.

On R Studio Console:

```
>Install.packages(“rattle”)
```

To force the installation of all dependency:

```
>install.packages("rattle", dep=c("Suggests"))
```

Or install using the Rstudio Install packages options

- Rattle relies on extensive collection of R packages which powers the Rattle UI.

Dependent packages for Rattle are RGtk2, cairoDevice and XML. Troubleshooting at http://rattle.togaware.com/rattle-install-troubleshooting.html

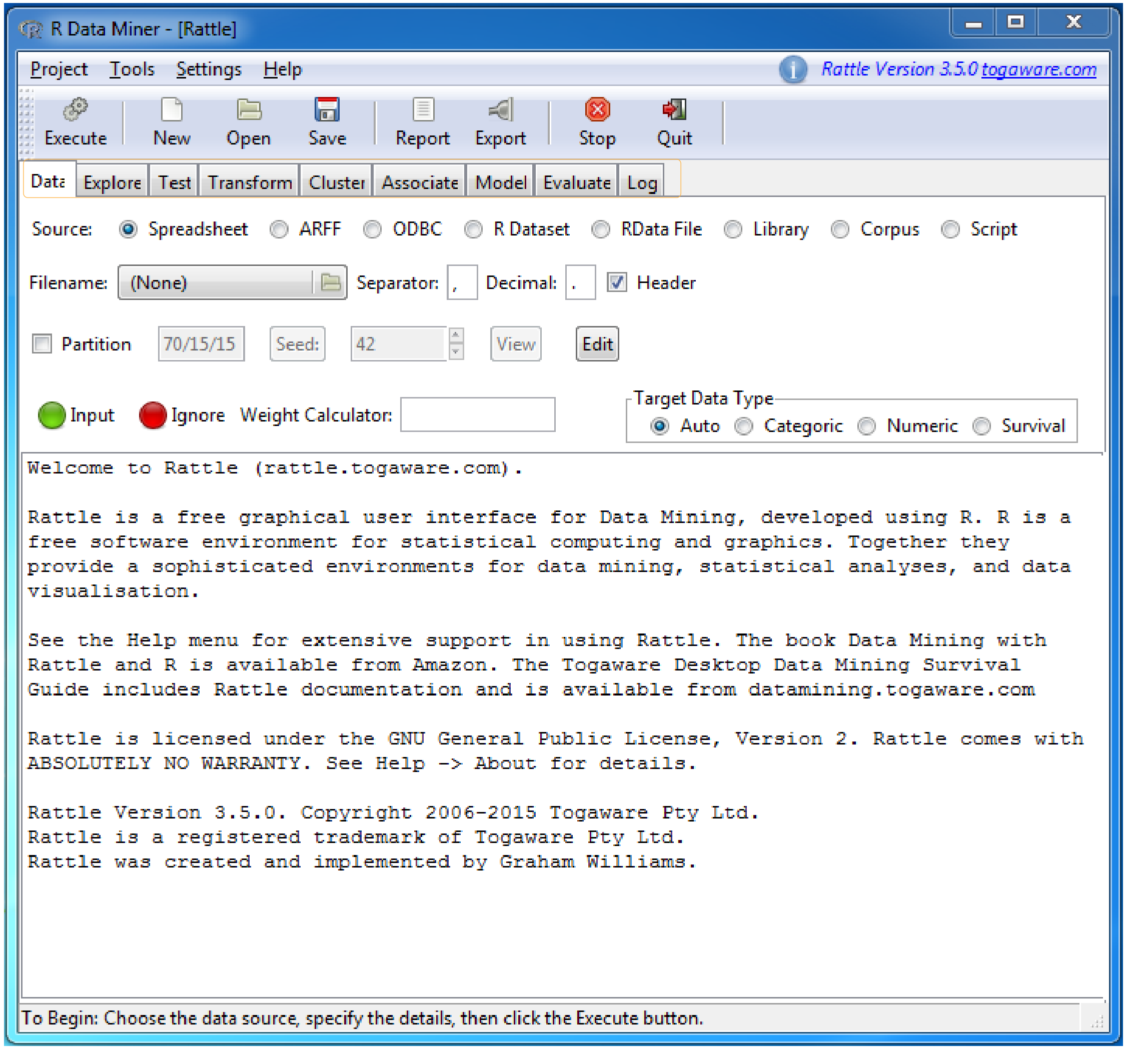
Rattle User Interface

- The user interface can be invoked as follows:

On R Studio Console:

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

- Tab based view with options to:
  - Load dataset
  - Explore dataset
  - Test distributions
  - Transform data
  - Clustering and association
  - Build models
  - Evaluate models
  - Code log



R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 3.5.0 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 book Data Mining with
Rattle and R is available from Amazon. 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 3.5.0. Copyright 2006-2015 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
```

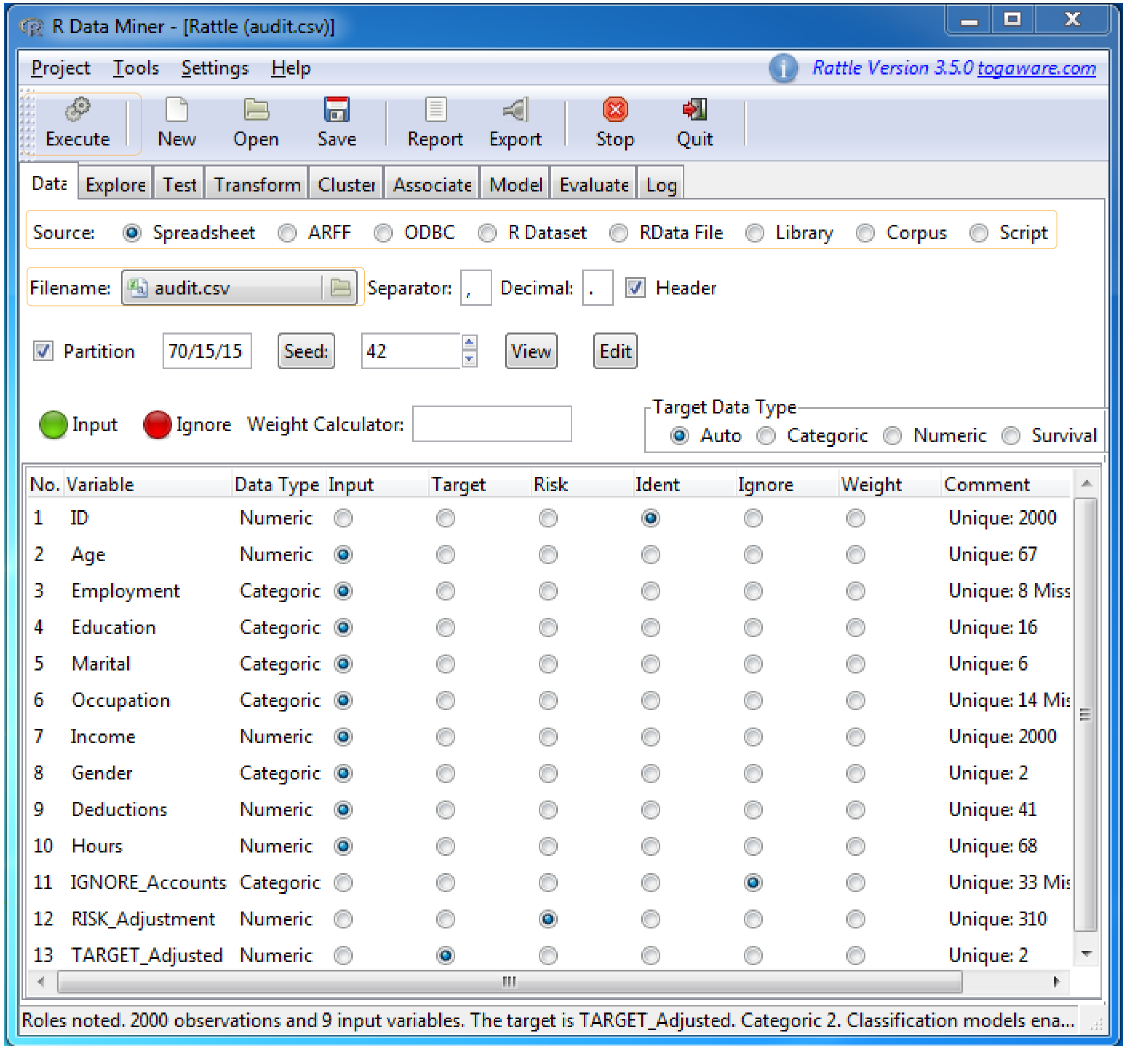
To Begin: Choose the data source, specify the details, then click the Execute button.

Rattle–Load Dataset

- A dataset is executed by the execute command.

- *If execute is clicked without any dataset, Rattle gives an option to load example dataset.*

- *Rattle recognizes special pre-fixes for default variable role*
  - *‘ID_’*
  - *‘IGNORE_’*
  - *‘RISK_’ (measure of size of the target)*
  - *‘IMP_’*
  - *‘TARGET_’*



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

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

☑ Partition 70/15/15 Seed: 42

● Input ● Ignore Weight Calculator:

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

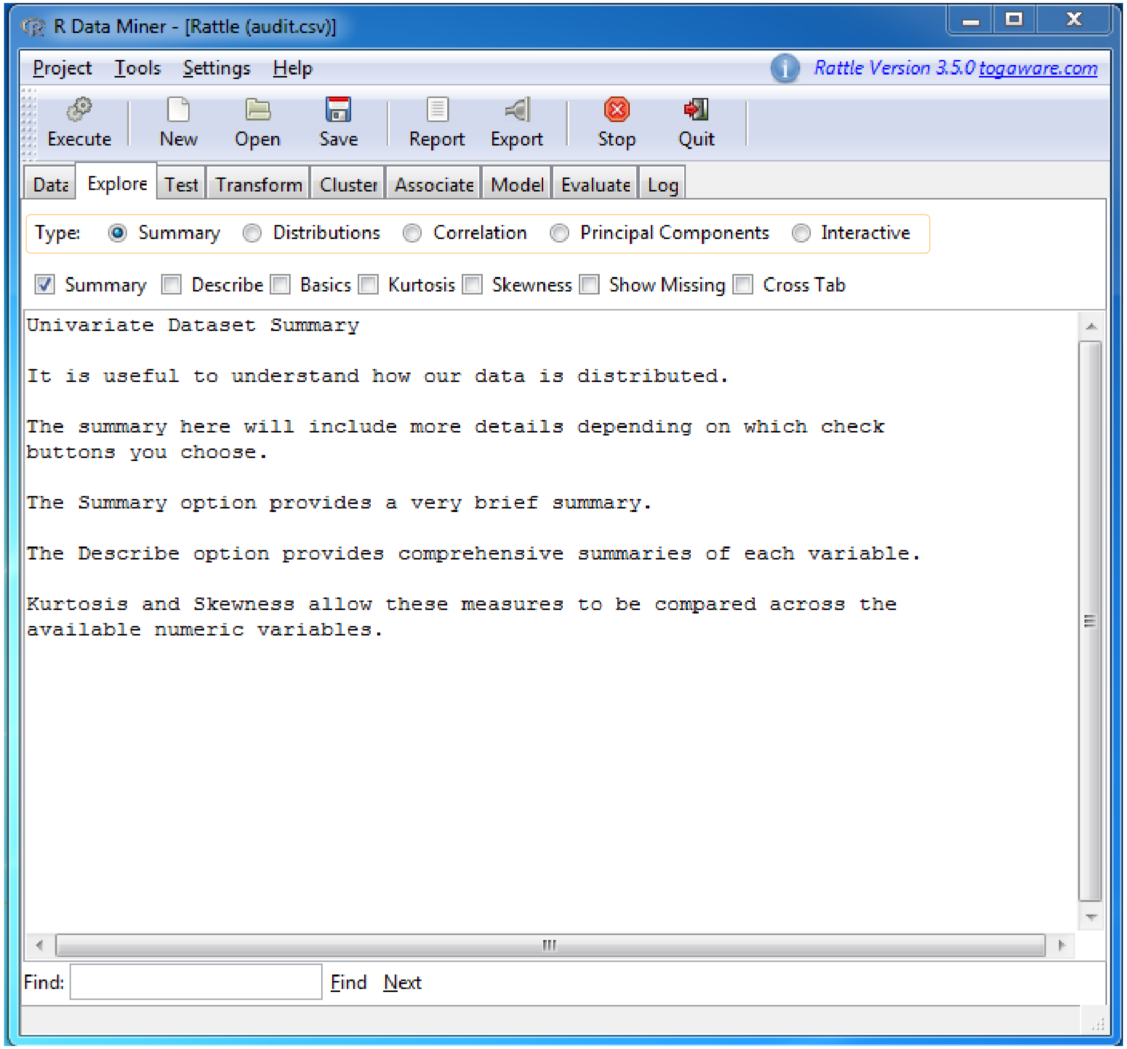
No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	ID	Numeric	○	○	○	●	○	○	Unique: 2000
2	Age	Numeric	●	○	○	○	○	○	Unique: 67
3	Employment	Categoric	●	○	○	○	○	○	Unique: 8 Miss
4	Education	Categoric	●	○	○	○	○	○	Unique: 16
5	Marital	Categoric	●	○	○	○	○	○	Unique: 6
6	Occupation	Categoric	●	○	○	○	○	○	Unique: 14 Mis
7	Income	Numeric	●	○	○	○	○	○	Unique: 2000
8	Gender	Categoric	●	○	○	○	○	○	Unique: 2
9	Deductions	Numeric	●	○	○	○	○	○	Unique: 41
10	Hours	Numeric	●	○	○	○	○	○	Unique: 68
11	IGNORE_Accounts	Categoric	○	○	○	○	●	○	Unique: 33 Mis
12	RISK_Adjustment	Numeric	○	○	●	○	○	○	Unique: 310
13	TARGET_Adjusted	Numeric	○	●	○	○	○	○	Unique: 2

Roles noted. 2000 observations and 9 input variables. The target is TARGET_Adjusted. Categoric 2. Classification models ena...

Rattle–Explore Dataset

- Explore tab provides various options for exploratory data analysis

- ***Summary****: Provides basic univariate summary and extended summary.*
- ***Distributions****: Provide various plots for numeric as well as categorical data*
- ***Correlation****: provides insights into the independence of the numeric input variables.*
- ***Principal component****: Provides insight into the importance of variables in explaining the variation.*
- ***Interactive****: Provides option for Interactive data exploration.*


R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Find: Find Next

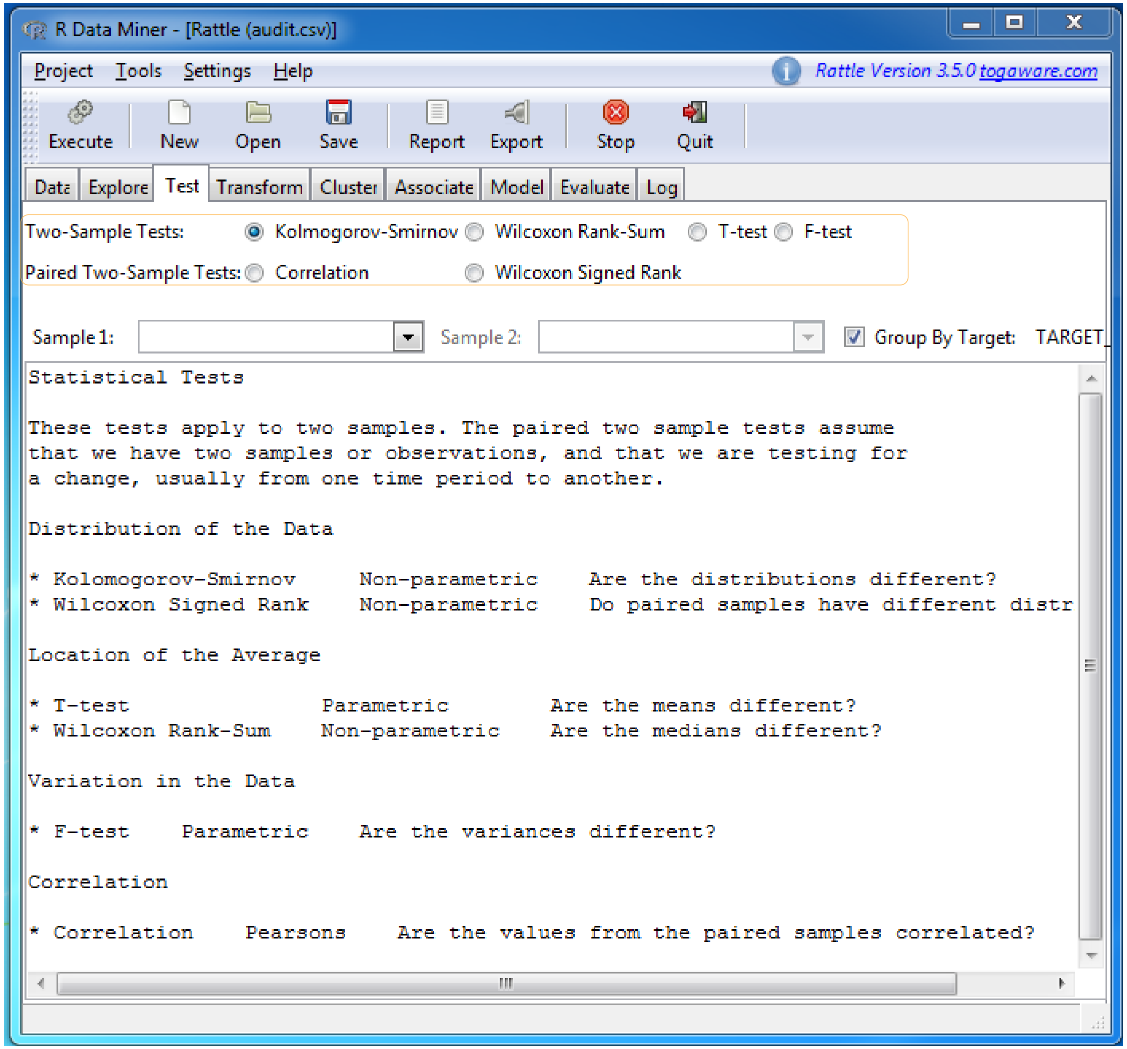
Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

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

```
Univariate Dataset Summary

It is useful to understand how our data is distributed.

The summary here will include more details depending on which check
buttons you choose.

The Summary option provides a very brief summary.

The Describe option provides comprehensive summaries of each variable.

Kurtosis and Skewness allow these measures to be compared across the
available numeric variables.
```

Rattle–Test Dataset

- Provides access to number of statistical tests of distributions.


R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute
New
Open
Save
Report
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

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

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

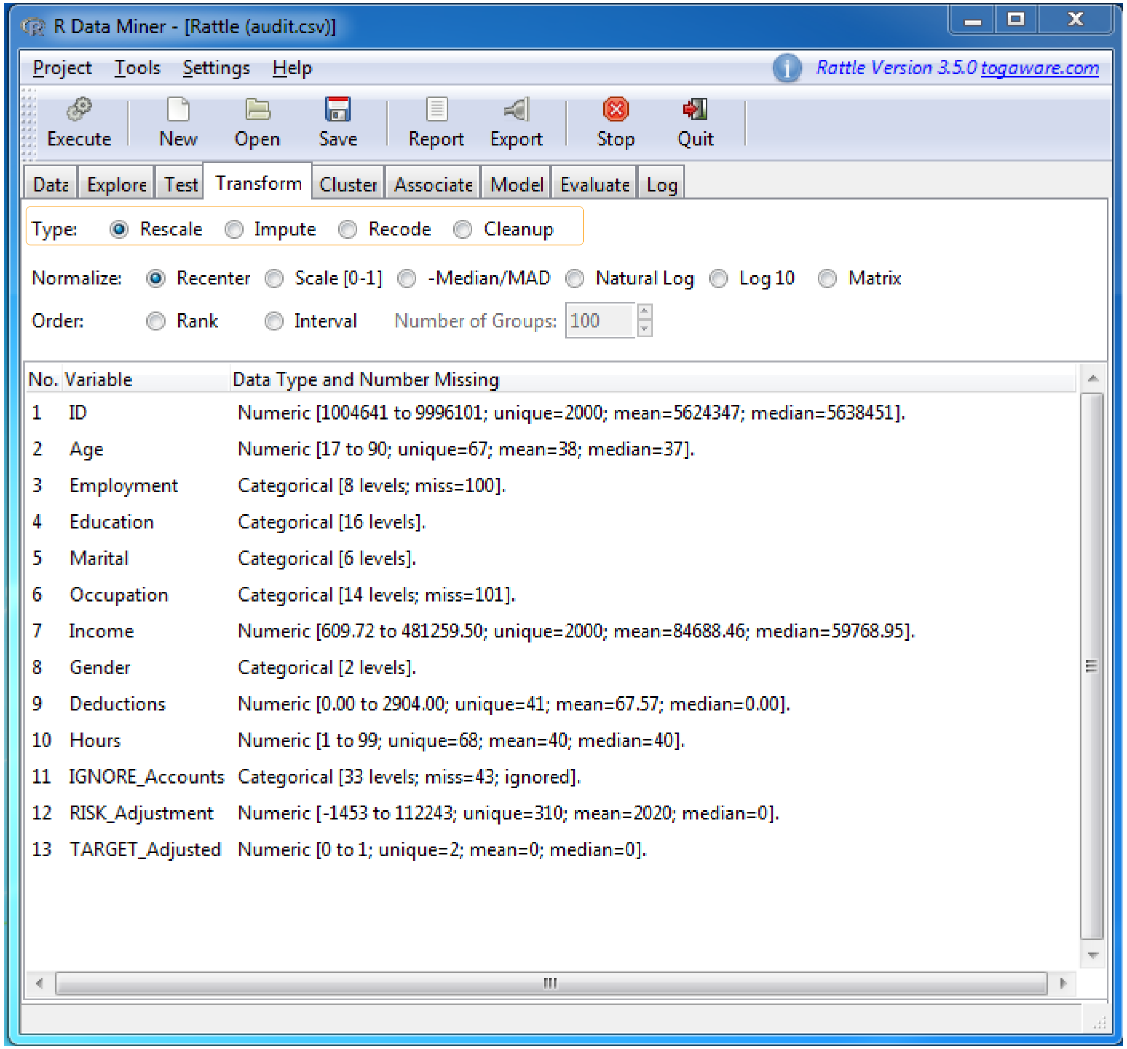
Sample 1: Sample 2: ☑ Group By Target: TARGET_

```
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 different?
* Wilcoxon Signed Rank    Non-parametric    Do paired samples have different distr

Location of the Average

* T-test                 Parametric         Are the means different?
* Wilcoxon Rank-Sum      Non-parametric     Are the medians different?

Variation in the Data

* F-test     Parametric    Are the variances different?

Correlation

* Correlation    Pearsons    Are the values from the paired samples correlated?
```

Rattle–Transform Dataset

- Cleaning data and creating new features (derived variables) takes significant time in data analysis.

- ***Rescale***: *Provides options for re-centering and scaling around zero.*
- ***Impute***: *Provides basic imputation of missing values using mean, median and mode.*
- ***Recode***: *Provides options for recoding/binning the variables with a default of 4 bins.*
- ***Cleanup***: *Provides option to treat the missing values after having tried imputation etc.*


R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ● Rescale ○ Impute ○ Recode ○ Cleanup

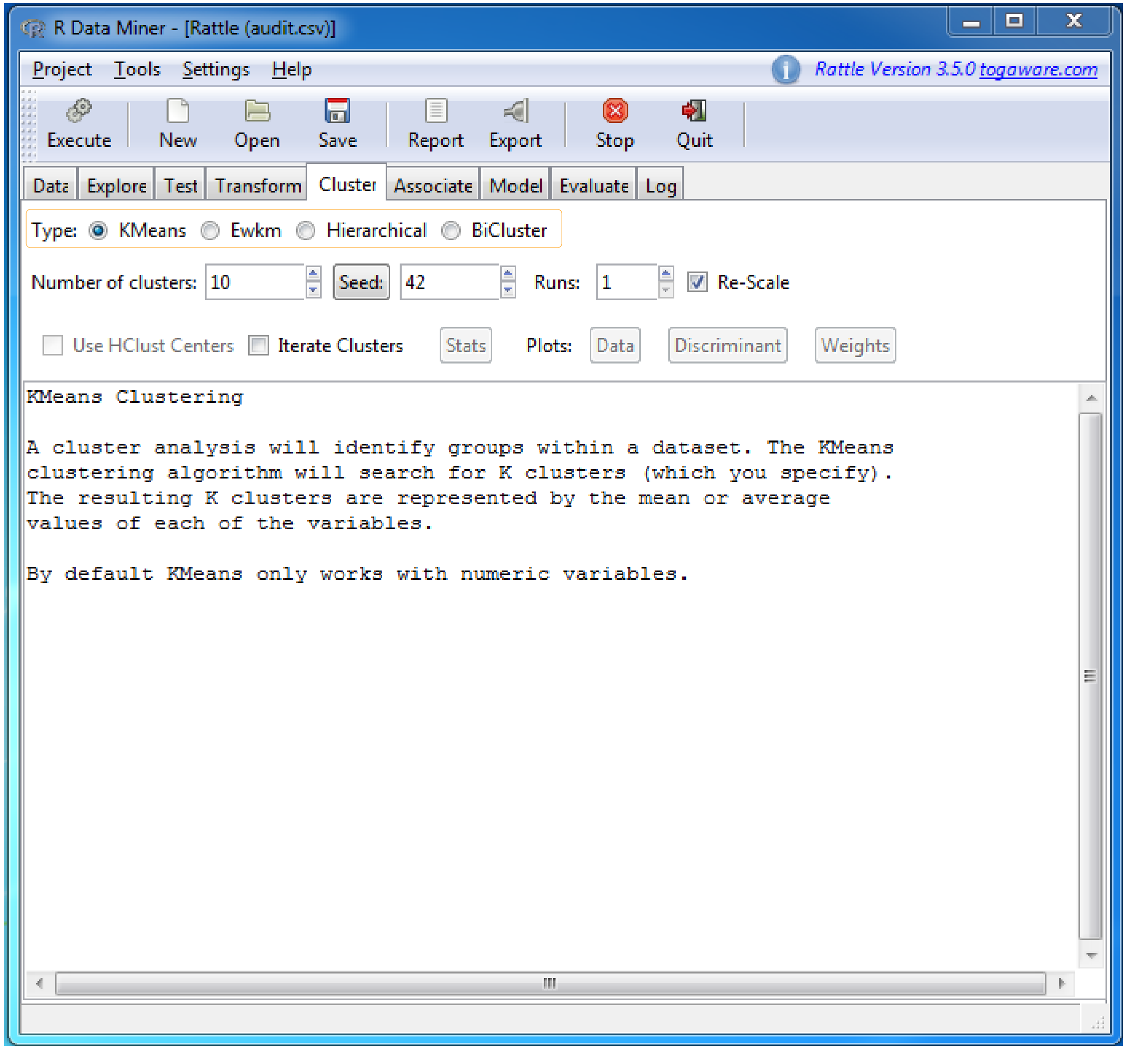
Normalize: ● Recenter ○ Scale [0-1] ○ -Median/MAD ○ Natural Log ○ Log 10 ○ Matrix

Order: ○ Rank ○ Interval Number of Groups: 100

No.	Variable	Data Type and Number Missing
1	ID	Numeric [1004641 to 9996101; unique=2000; mean=5624347; median=5638451].
2	Age	Numeric [17 to 90; unique=67; mean=38; median=37].
3	Employment	Categorical [8 levels; miss=100].
4	Education	Categorical [16 levels].
5	Marital	Categorical [6 levels].
6	Occupation	Categorical [14 levels; miss=101].
7	Income	Numeric [609.72 to 481259.50; unique=2000; mean=84688.46; median=59768.95].
8	Gender	Categorical [2 levels].
9	Deductions	Numeric [0.00 to 2904.00; unique=41; mean=67.57; median=0.00].
10	Hours	Numeric [1 to 99; unique=68; mean=40; median=40].
11	IGNORE_Accounts	Categorical [33 levels; miss=43; ignored].
12	RISK_Adjustment	Numeric [-1453 to 112243; unique=310; mean=2020; median=0].
13	TARGET_Adjusted	Numeric [0 to 1; unique=2; mean=0; median=0].

Rattle–Cluster Analysis

- Cluster tab provides option to build descriptive or unsupervised model.
- Several clustering algorithm available as options to identify groups within the dataset.



R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Stats
Plots: Data Discriminant Weights

Type: ● KMeans ○ Ewkm ○ Hierarchical ○ BiCluster

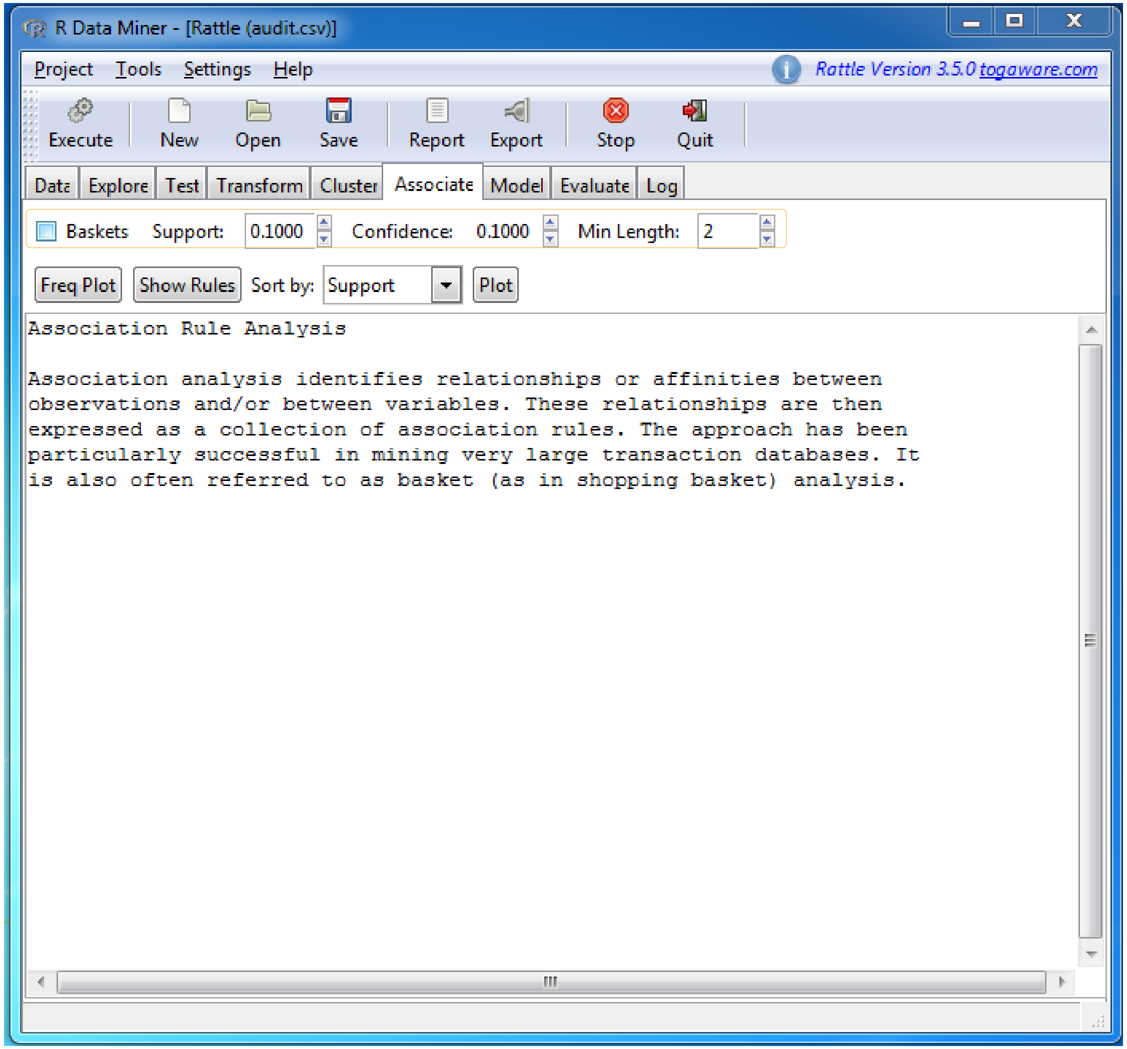
Number of clusters: 10 Seed: 42 Runs: 1 ☑ Re-Scale

□ Use HClust Centers □ Iterate Clusters

```
KMeans Clustering

A cluster analysis will identify groups within a dataset. The KMeans
clustering algorithm will search for K clusters (which you specify).
The resulting K clusters are represented by the mean or average
values of each of the variables.

By default KMeans only works with numeric variables.
```

Rattle–Basket Analysis

- Associate tab gives another option to build descriptive or unsupervised model.
- Option available for market basket analysis to identify affinities between observations and/or between variables.



R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Freq Plot Show Rules Plot

□ Baskets Support: 0.1000 Confidence: 0.1000 Min Length: 2

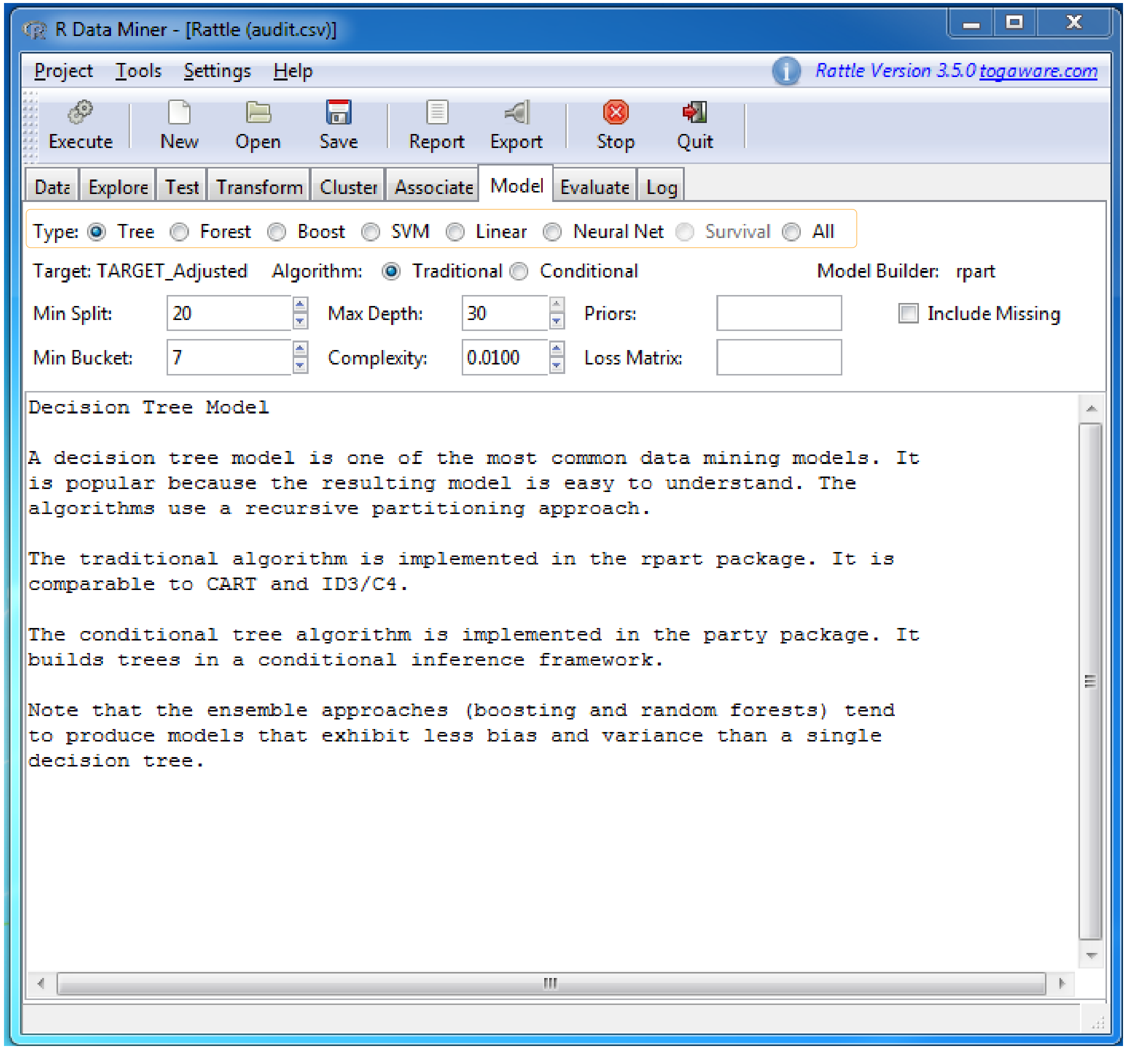
Sort by: Support

```
Association Rule Analysis

Association analysis identifies relationships or affinities between
observations and/or between variables. These relationships are then
expressed as a collection of association rules. The approach has been
particularly successful in mining very large transaction databases. It
is also often referred to as basket (as in shopping basket) analysis.
```

Rattle–Model Dataset

- Model tab provides a comprehensive list of techniques to build predictive models.

- *Provides an option to use all the model building techniques over the same dataset.*
- *The models can be evaluated for performance and the best model can be selected.*



R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

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

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

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

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

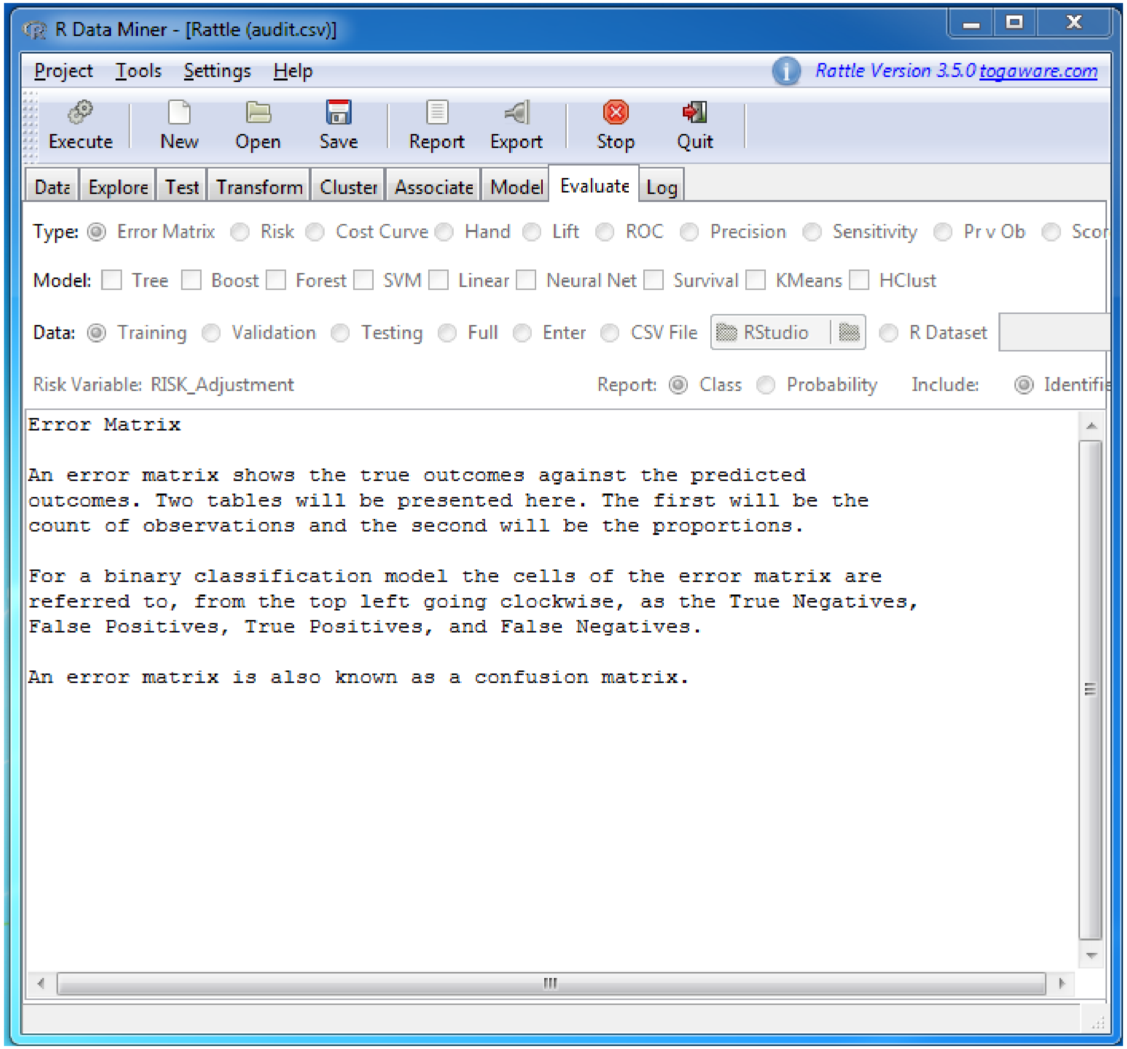
```
Decision Tree Model

A decision tree model is one of the most common data mining models. It
is popular because the resulting model is easy to understand. The
algorithms use a recursive partitioning approach.

The traditional algorithm is implemented in the rpart package. It is
comparable to CART and ID3/C4.

The conditional tree algorithm is implemented in the party package. It
builds trees in a conditional inference framework.

Note that the ensemble approaches (boosting and random forests) tend
to produce models that exhibit less bias and variance than a single
decision tree.
```

Rattle–Evaluate Model

- Evaluate tab provides a collection of techniques for evaluating the performance of models

- *Some of the commonly used techniques for model comparison can be seen as options:*
  - ○ *Error matrix*
  - ○ *ROC curve*
  - ○ *Lift Chart*

- *Rattle supports deployment of the model through the ‘Score’ option.*
  - ○ *The complete model can be saved as a Rattle project and can later be used on the new dataset to score the*



Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Scor

Model: □ Tree □ Boost □ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ● Training ○ Validation ○ Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

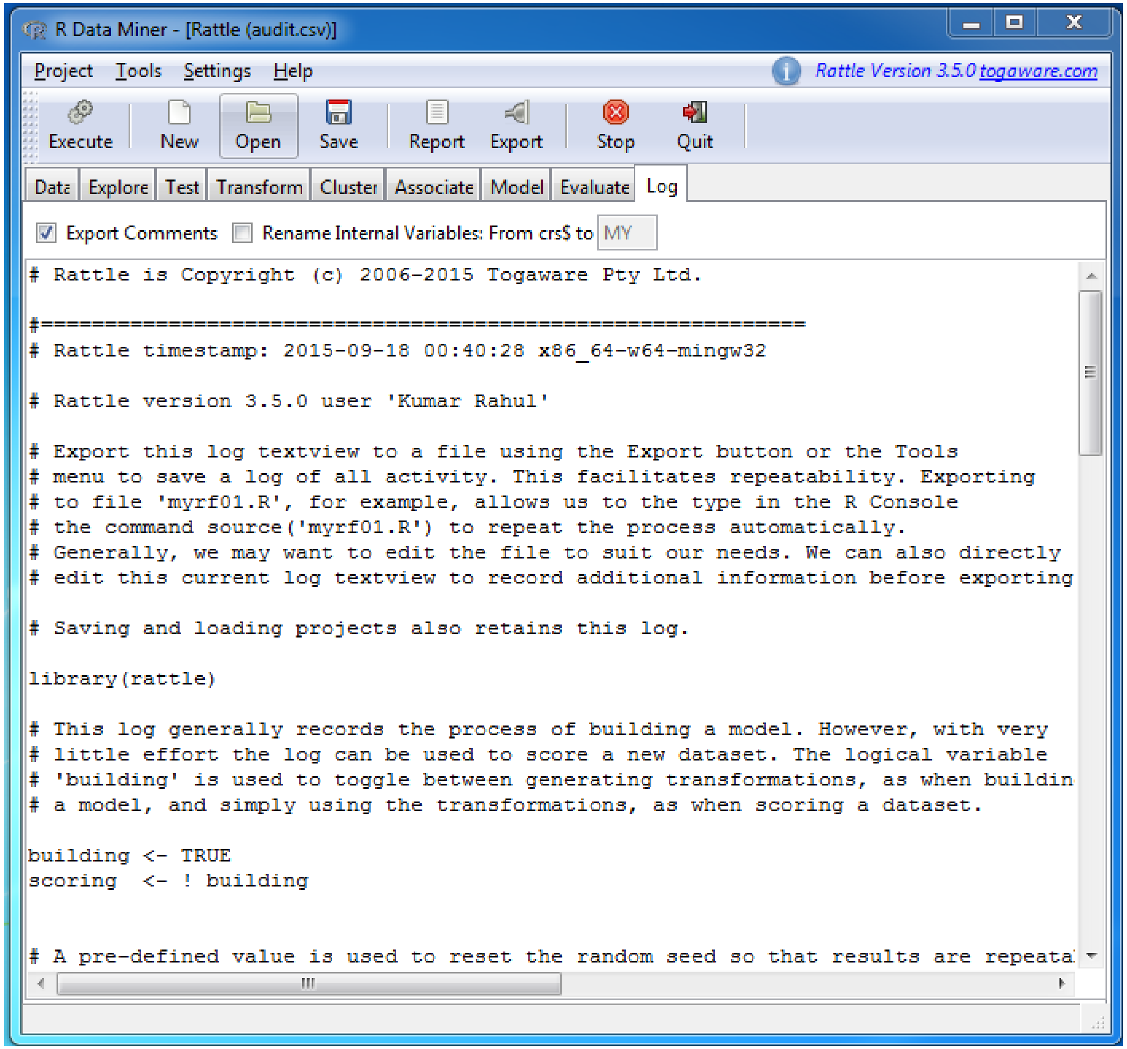
Risk Variable: RISK_Adjustment Report: ● Class ○ Probability Include: ● Identifi

```
Error Matrix

An error matrix shows the true outcomes against the predicted
outcomes. Two tables will be presented here. The first will be the
count of observations and the second will be the proportions.

For a binary classification model the cells of the error matrix are
referred to, from the top left going clockwise, as the True Negatives,
False Positives, True Positives, and False Negatives.

An error matrix is also known as a confusion matrix.
```

Rattle–Log Generation

- Log tab records the process of building the model.
- The recorded script gives the flexibility to fine tune the analysis using R directly.

- *The log can be used for deployment to score a new dataset.*


R Data Miner - [Rattle (audit.csv)]
Project Tools Settings Help
Rattle Version 3.5.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

☑ Export Comments □ Rename Internal Variables: From crs$ to MY

```
# Rattle is Copyright (c) 2006-2015 Togaware Pty Ltd.

#============================================================
# Rattle timestamp: 2015-09-18 00:40:28 x86_64-w64-mingw32

# Rattle version 3.5.0 user 'Kumar Rahul'

# Export this log textview to a file using the Export button or the Tools
# menu to save a log of all activity. This facilitates repeatability. Exporting
# to file 'myrf01.R', for example, allows us to the type in the R Console
# the command source('myrf01.R') to repeat the process automatically.
# Generally, we may want to edit the file to suit our needs. We can also directly
# edit this current log textview to record additional information before exporting

# Saving and loading projects also retains this log.

library(rattle)

# This log generally records the process of building a model. However, with very
# little effort the log can be used to score a new dataset. The logical variable
# 'building' is used to toggle between generating transformations, as when buildin
# a model, and simply using the transformations, as when scoring a dataset.

building <- TRUE
scoring  <- ! building


# A pre-defined value is used to reset the random seed so that results are repeata
```

Demo of the Rattle tool using an example dataset.

Summary

Summary of the topics covered in this lesson:

- R Analytical Tool to Learn Easily (Rattle) is a user interface based data mining tool built on top of R.
- Rattle provides a tab based options to load, explore, test, transform a dataset; followed by building and evaluating models.

QUIZ TIME

Quiz Question 1

Quiz 1

What is the command line syntax to install rattle?

Select all that apply.

a. *install.packages("rattle", dep=c("Suggests"))*

b. *install.packages("rattle“)*

c. *install.package(“rattle”)*

d. *install.package("rattle", dep=c("Suggests"))*

Quiz Question 1

Quiz 1

What is the command line syntax to install rattle?

Select all that apply.

a. *install.packages("rattle", dep=c("Suggests"))*

b. *install.packages("rattle“)*

c. *install.package(“rattle”)*

d. *install.package("rattle", dep=c("Suggests"))*

Correct answer is:

a & b

Both a and b has the correct syntax. Option a has an optional argument of forcing the dependent packages to be installed.

R is an official part of the Free Software Foundation’s GNU project.
RStudio and Shiny are affiliated projects of the Foundation for Open Access Statistics.

End of Lesson06–Introduction to R Markdown and Rattle