KE5107: Data Mining Methodology and Methods

Workshop: Data Exploration

Preword

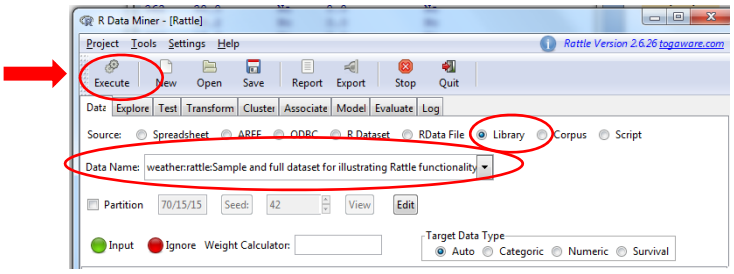
- You'll need internet connection in this workshop.
- For Rattle functions to work, many dependent R packages need to be downloaded and installed.
- We'll use GUI selection and click buttons, but please check the *Log* tab to see and learn the actual R codes.

Data Import

- Install *rattle.data* package first
- Start Rattle in R

```
> library(rattle)
Rattle: A free graphical interface for data mining with R.
Version 3.4.1 Copyright (c) 2006-2014 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
```

- Load data into Rattle
  - Let's use the "*weather*" data set from *rattle.data* package
  - **And click "Execute"!**



About the Dataset

Variable Name	Meaning	Units
Date	Date of observation	Day
Location	Location of observations	Name
Min Temps	Minimum temperature in the 24 hours to 9am.	degrees Celsius
Max Temp	Maximum temperature in the 24 hours from 9am.	degrees Celsius
Rainfall	Precipitation (rainfall) in the 24 hours to 9am.	millimeters
Evaporation	Class A pan evaporation in the 24 hours to 9am	millimeters
Sunshine	Bright sunshine in the 24 hours to midnight	hours
WindGustDir	Direction of strongest gust in the 24 hours to midnight	16 compass points
WindGustSpeed	Speed of strongest wind gust in the 24 hours to midnight	kilometers per hour
WindSpeed9am	Wind speed averaged over 10 minutes prior to 9 am	kilometers per hour
WindSpeed3pm	Wind speed averaged over 10 minutes prior to 3 pm	kilometers per hour

About the Dataset (continued)

Variable Name	Meaning	Units
Humidity9am	Relative humidity at 9 am	percent
Humidity3pm	Relative humidity at 3 pm	percent
Pressure9am	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
Pressure3pm	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals
Cloud9am	Fraction of sky obscured by cloud at 9 am	eighths
Cloud3pm	Fraction of sky obscured by cloud at 3 pm	eighths
Temp9am	Temperature at 9 am	degrees Celsius
Temp3pm	Temperature at 3 pm	degrees Celsius
RainToday	Did it rain the day of the observation	Yes/No
RISK_MM	Precipitation (rainfall) in the 24 hours to 9am.	millimeters
RainTomorrow	Did it rain the next day of the observation	Yes/No

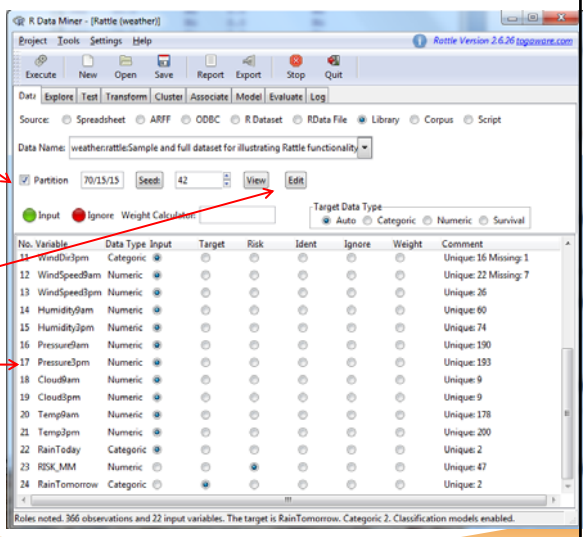
The Loaded Data

When doing predicative modeling, use this to partition data into training/validate/test sets. (Also used when it's computationally too costly to explore or visualize large dataset. Otherwise uncheck.)

To view data table or edit

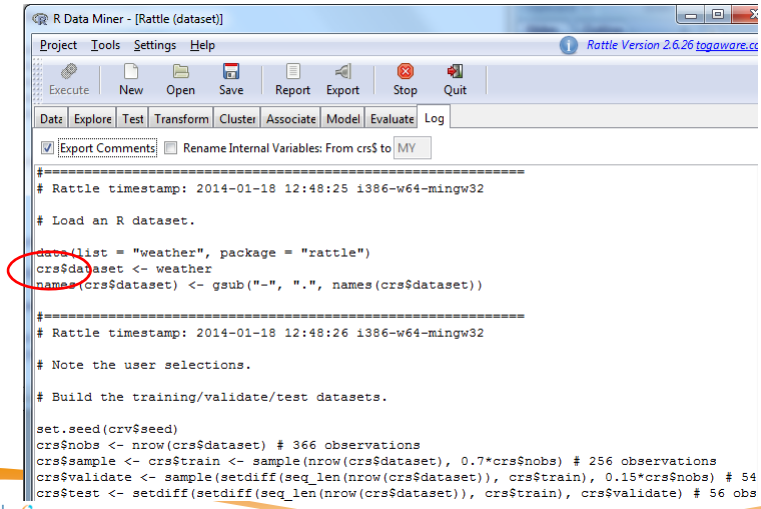
Information on variables:

- –Data type
- –Variable roles
- –Unique values



What's Going on…

- Let's see the generated codes at *Log* tab



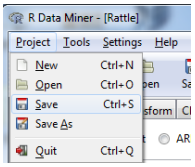
R Environment

- As a container for a collection of data within a project
- Rattle's environment is called "*crs*" by default, but can be changed.
- Information within the environment is accessed using *$* notation, e.g. *crs$dataset*
- An environment can be saved to a file for future use : Project -> Save, or use command

  save(crs, file="weather.rattle")

  which can be reloaded using Project -> Open, or

  load("weather.rattle")



- We can also use *attach* and *detach* to make the objects within an environment directly accessible without using the *crs$* notation

Data Editor

Rattle Dataset - dfedit version 0.6.1

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeec
1	13818	Canberra	8	24.3	0	3.4	6.3	NW	30
2	13819	Canberra	14	26.9	3.6	4.4	9.7	ENE	39
3	13820	Canberra	13.7	23.4	3.6	5.8	3.3	NW	85
4	13821	Canberra	13.3	15.5	39.8	7.2	9.1	NW	54
5	13822	Canberra	7.6	16.1	2.8	5.6	10.6	SSE	50
6	13823	Canberra	6.2	16.9	0	5.8	8.2	SE	44
7	13824	Canberra	6.1	18.2	0.2	4.2	8.4	SE	43
8	13825	Canberra	8.3	17	0	5.6	4.6	E	41
9	13826	Canberra	8.8	19.5	0	4	4.1	S	48
10	13827	Canberra	8.4	22.8	16.2	5.4	7.7	E	31
11	13828	Canberra	9.1	25.2	0	4.2	11.9	N	30
12	13829	Canberra	8.5	27.3	0.2	7.2	12.5	E	41
13	13830	Canberra	10.1	27.9	0	7.2	13	WNW	30
14	13831	Canberra	12.1	30.9					
15	13832	Canberra	10.1	31.2					
16	13833	Canberra	12.4	32.1					
17	13834	Canberra	13.8	31.2					

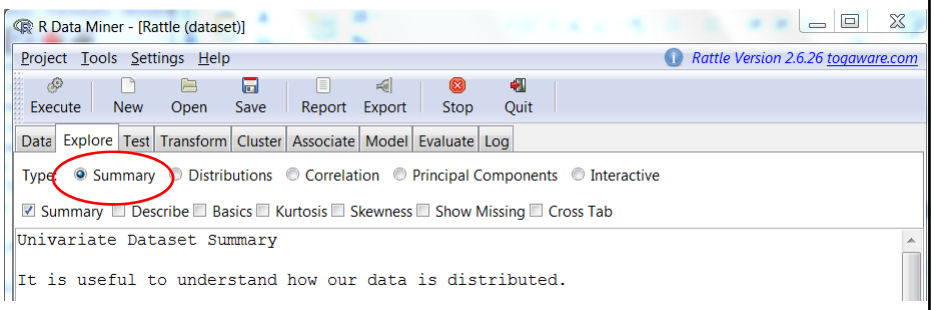
```
#============================================================
# Rattle timestamp: 2014-01-18 12:52:59 i386-w64-mingw32

# Edit the Dataset

crs$dataset <- edit(crs$dataset)
```


Exercise

- Let's do the following:
  - Uncheck "Partition" (We want to explore the whole dataset)
  - Review data types, roles, number of unique values, and missing values
    - Any variable with just one value?
    - Any variable with all unique values?
    - Any variable with numeric type but indicating categorical information?
    - ...
- Click "Execute"!

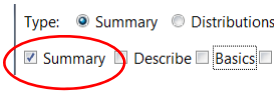
Summarizing Data

- Rattle's "Explore" tab offers a basic Summary option, which provides a **textual overview** of the data
- Select "Summary" Type, check the specific function you want, and click "Execute"



Basic Statistical Summary

- "Summary" (*summary()*) gives us the simplest text-based statistical summary of each variable in the dataset.
- Always a useful starting point in reviewing our data.
- Which variable has rather different mean and median?




```
Below we summarise the dataset.
Note that the data contains 38 observations with missing values.
Enable the 'Show Missing' check box for details.

Data frame:crs$dataset[, c(crs$input, crs$risk, crs$target)]    366 observations and 23

               Levels   Class Storage NAs
Location           46         integer   0
MinTemp                        double   0
MaxTemp                        double   0
Rainfall                       double   0
Evaporation                    double   0
Sunshine                       double   3
WindGustDir        16 ordered integer   3
WindGustSpeed                  double   2
WindDir9am         16 ordered integer  31
WindDir3pm         16 ordered integer   1
```

```
          Location      MinTemp          MaxTemp          Rainfall
 Canberra     :366   Min.   :-5.300   Min.   : 7.60   Min.   : 0.000
 Adelaide     :  0   1st Qu.: 2.300   1st Qu.:15.03   1st Qu.: 0.000
 Albany       :  0   Median : 7.450   Median :19.65   Median : 0.000
 Albury       :  0   Mean   : 7.266   Mean   :20.55   Mean   : 1.428
 AliceSprings :  0   3rd Qu.:12.500   3rd Qu.:25.50   3rd Qu.: 0.200
 BadgerysCreek:  0   Max.   :20.900   Max.   :35.80   Max.   :39.800
 (Other)      :  0
  Evaporation        Sunshine       WindGustDir  WindGustSpeed     WindDir9am
 Min.   : 0.200   Min.   : 0.000   NW     : 73   Min.   :13.00   SE     : 47
 1st Qu.: 2.200   1st Qu.: 5.950   NNW    : 44   1st Qu.:31.00   SSE    : 40
 Median : 4.200   Median : 8.600   E      : 37   Median :39.00   NNW    : 36
 Mean   : 4.522   Mean   : 7.909   WNW    : 35   Mean   :39.84   N      : 31
 3rd Qu.: 6.400   3rd Qu.:10.500   ENE    : 30   3rd Qu.:46.00   NW     : 30
 Max.   :13.800   Max.   :13.600   (Other):144   Max.   :98.00   (Other):151
                  NA's   :3        NA's   :  3   NA's   :2       NA's   : 31
```

Describe the Data

- To get more detailed summary, use "Describe" (*describe()*)
- For numerical variables

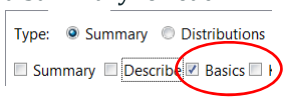

```
Evaporation
     n missing  unique    Mean     .05     .10     .25     .50     .75     .90     .95
   366       0      55   4.522     1.2     1.5     2.2     4.2     6.4     8.4     9.4

lowest :  0.2  0.6  0.8  1.0  1.2, highest: 11.4 11.6 12.4 12.6 13.8
---------------------------------------------------------------------------------------
Sunshine
     n missing  unique    Mean     .05     .10     .25     .50     .75     .90     .95
   363       3     114   7.909    0.60    2.04    5.95    8.60   10.50   11.80   12.60

lowest :  0.0  0.1  0.2  0.3  0.4, highest: 13.1 13.2 13.3 13.5 13.6
---------------------------------------------------------------------------------------
```

- For categorical variables

```
---------------------------------------------------------------------------
WindGustDir
      n missing  unique
    363       3      16

            N NNE NE ENE  E ESE SE SSE  S SSW SW WSW  W WNW NW NNW
Frequency 21   8 16  30 37  23 12  12 22   5  3   2 20  35 73  44
%          6   2  4   8 10   6  3   3  6   1  1   1  6  10 20  12
```


More Detailed Numeric Summary

- "Basics" (*basicStats()*) gives even more detailed summary for each numeric variables




```
$Sunshine
                  X...X.5
nobs           366.000000
NAs              3.000000
Minimum          0.000000
Maximum         13.600000
1. Quartile      5.950000
3. Quartile     10.500000
Mean             7.909366
Median           8.600000
Sum           2871.100000
SE Mean          0.182732
LCL Mean         7.550016
UCL Mean         8.268716
Variance        12.120962
Stdev            3.481517
Skewness        -0.723454
Kurtosis        -0.270625
```

Skewness

- How asymmetrically the data is distributed
- For a summary of skewness to compare the distribution of a number of numeric variables

  skewness() from **Hmisc** package

Type: Summary Distributions Correlation I

Summary Describe Basics Kurtosis Skewness

```
Skewness for each numeric variable of the dataset.
Positive means the right tail is longer.

      MinTemp       MaxTemp      Rainfall   Evaporation      Sunshine
 -0.003779725   0.347510625   4.552606775   0.658228261  -0.723454350
WindGustSpeed  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
  0.836105520   1.360171287   0.591272142  -0.140299824   0.589864926
  Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
 -0.345738718  -0.292831116   0.079775173   0.072301095  -0.021605052
      Temp3pm     RISK_MM
  0.301026409   4.552606775
```


Kurtosis

- How sharp or flat the peak of a distribution is

  kurtosis() from **Hmisc** package
- Larger value – sharper peak

Distributions Correlation P

ibe Basics Kurtosis Skewness

```
Kurtosis for each numeric variable of the dataset.
Larger values mean sharper peaks and flatter tails.
Positive values indicate an acute peak around the mean.
Negative values indicate a smaller peak around the mean.

      MinTemp       MaxTemp      Rainfall   Evaporation      Sunshine
  -1.12569017   -0.76360944  26.23970072   -0.20876073   -0.27062478
WindGustSpeed  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
   1.47610274    1.47582536   0.19632764   -0.20595527    0.01155850
  Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
  -0.07543813   -0.03560288  -1.71686547   -1.63023412   -0.97145358
      Temp3pm     RISK_MM
  -0.68649659  26.23970072
```

Missing Data

- To explore any structure in missing data

  md.pattern() from **mice** package

Skewness ☑ Show Missing ☐ Cros

```
    Location MinTemp MaxTemp Rainfall Evaporation WindSpeed3pm Humidity9am
328        1       1       1        1           1            1           1
  3        1       1       1        1           1            1           1
  1        1       1       1        1           1            1           1
 24        1       1       1        1           1            1           1
  1        1       1       1        1           1            1           1
  2        1       1       1        1           1            1           1
  7        1       1       1        1           1            1           1
           0       0       0        0           0            0           0
    Humidity3pm Pressure9am Pressure
328           1           1
  3           1           1
  1           1           1
 24           1           1
  1           1           1
  2           1           1
  7           1           1
              0           0
```

```
    RainToday RainTomorrow WindDir3pm WindGustSpeed Sunshine WindGustDir
328         1           1         1             1        1           1
  3         1           1         1             1        0           1
  1         1           1         1             1        1           0
 24         1           1         1             1        1           1
  1         1           1         0             1        1           1
  2         1           1         1             0        1           0
  7         1           1         1             1        1           1
            0           0         1             2        3           3
```

```
    WindSpeed9am WindDir9am
328            1          1 0
  3            1          1 1
  1            1          1 1
 24            1          0 1
  1            1          1 1
  2            1          1 2
  7            0          0 2
               7         31 47
```


Check Distribution Visually

- Select the variables and plots, then click "Execute"
  - Numerical variables: histogram, box plot

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: Summary ◉ Distributions Correlation Principal Components Interactive

Numeric: ☐ Annotate **Group By:** RainTomorrow

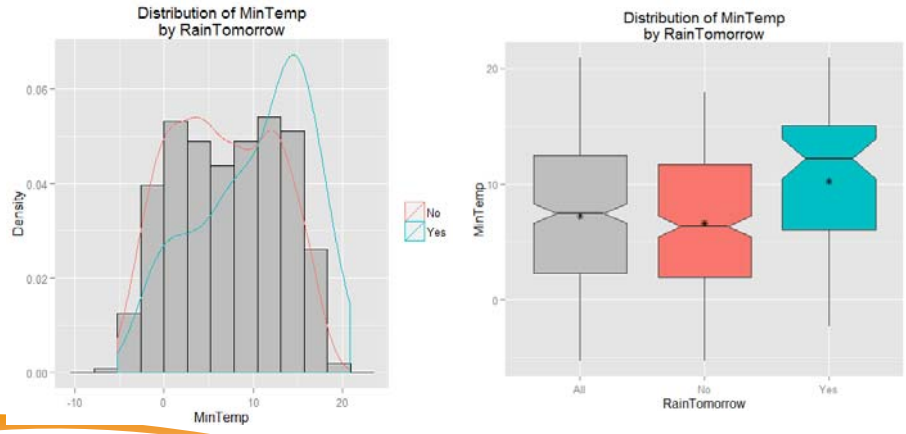
Benfords: ☐ Bars Starting Digit: 1 Number of Digits: 1 ◉ abs +ve -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Pairs	Min; Median/Mean; Max
7	Sunshine	☑	☑	☐	☐	☐	0.00; 8.60/7.91; 13.60

- Run the plots one by one.
- The generated plots include the distributions for the whole dataset, as well as the distributions for each subset of observations associated with each value (Yes and No) of the target variable.

Histogram and Boxplot

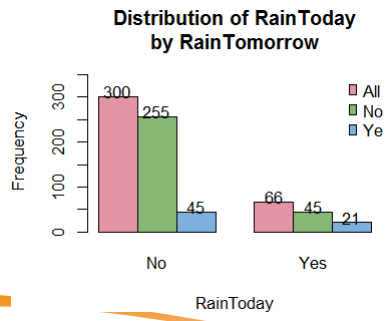
- For variable "MinTemp"

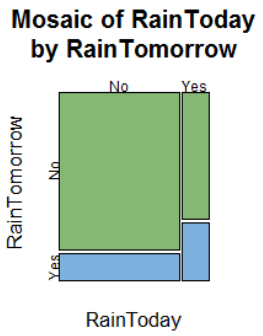


Check Distribution Visually

- Categorical variables: bar plot, mosaic

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



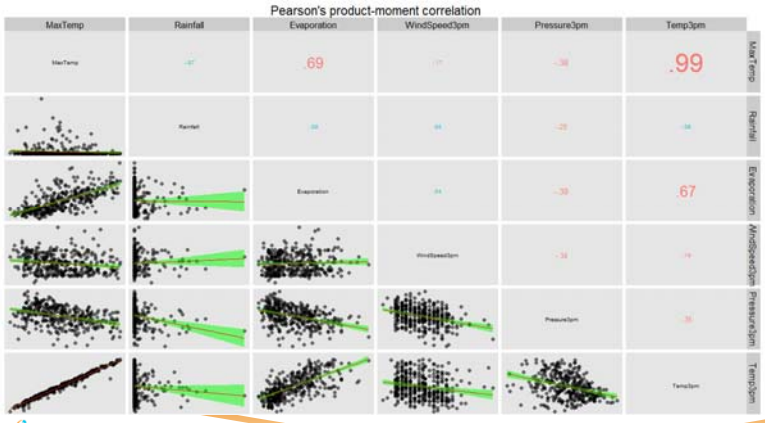


Exercise

- Visualize other variables.
- Which ones have very different distribution in the "Yes" and "No" subsets for "RainTomorrow"?

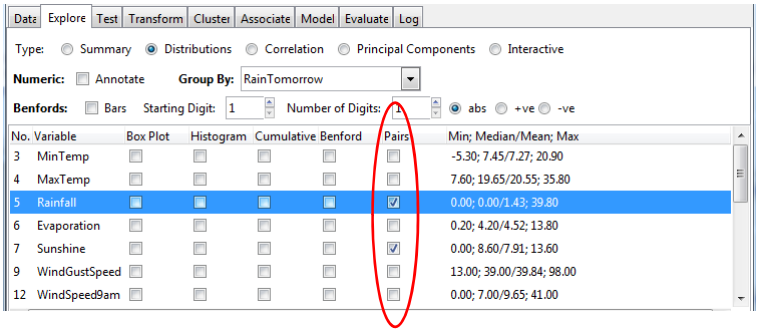
Compare Variables Pairwise

- Generate a scatter plot matrix between numeric variables
- At Explore tab, ensure no plots are selected for any variable
- Then click "Execute", 6 variables randomly selected



Compare Variables Pairwise

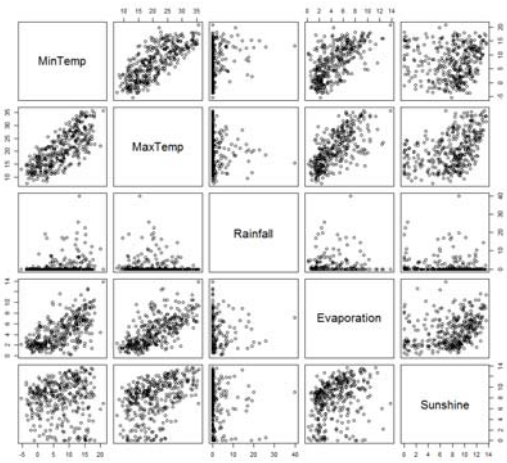
- To generate a scatter plot between specific numeric variables, select the variables in the "Pairs" column
- Then click "Execute"



Scatter Plot Matrix

- You can always go back to R console, using *plot()* or *pairs()*
- Caution! Only do it for a small number of variables, or R will crash.

```
> plot(weather[3:7])
```

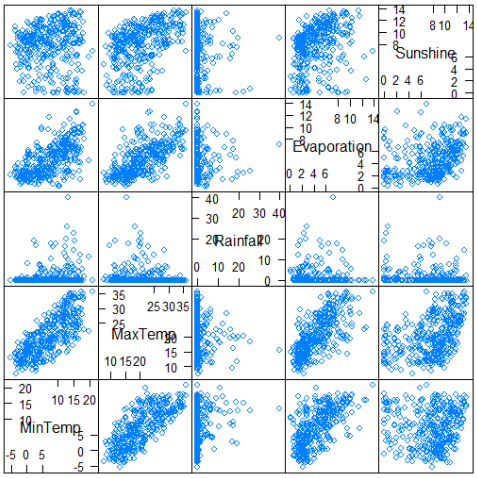


Scatter Plot Matrix

- We can also exploit the advanced plotting capabilities of R's **lattice** package
- Access it from R Console

  require(lattice)

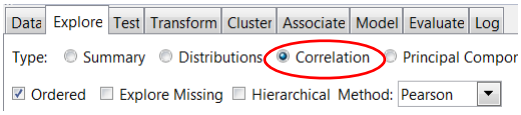
  splom(~weather[3:7])



Scatter Plot Matrix

Correlation Analysis

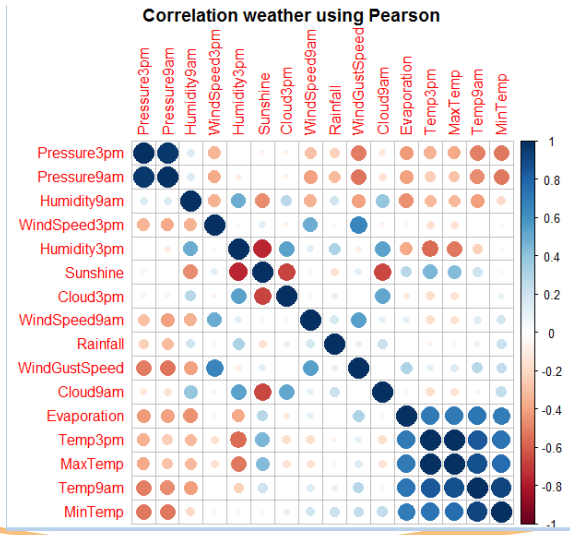
- Select "Correlation" at *Explore* tab, and click "Execute"



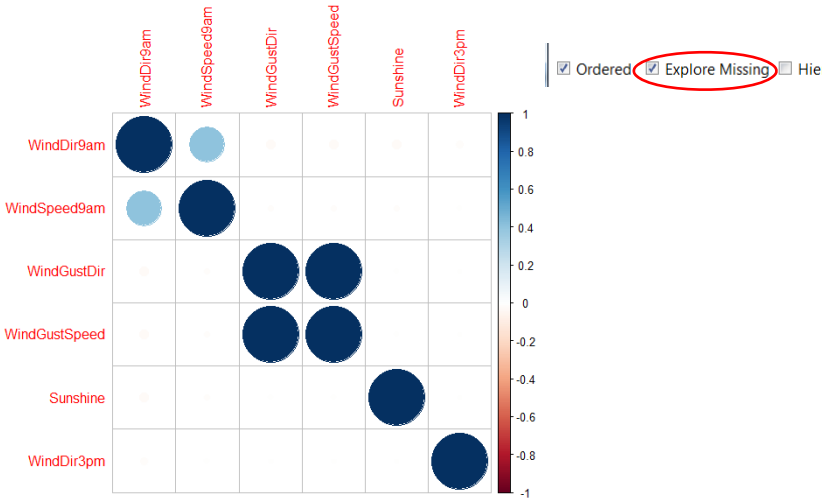

```
Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

               Pressure9am Pressure3pm Humidity9am WindSpeed3pm Humidity3pm
Pressure9am     1.000000000  0.96789496   0.1357270  -0.35980011 -0.08794614
Pressure3pm     0.967894962  1.00000000   0.1344205  -0.33732535 -0.01005189
Humidity9am     0.135726974  0.13442050   1.0000000  -0.26609247  0.54671844
WindSpeed3pm   -0.359800112 -0.33732535  -0.2660925   1.00000000 -0.02636775
Humidity3pm    -0.087946135 -0.01005189   0.5467184  -0.02636775  1.00000000
Sunshine        0.006276442 -0.03620087  -0.4990174   0.07257280 -0.75942920
Cloud3pm       -0.141000431 -0.14383718   0.2719381   0.00720724  0.51010790
WindSpeed9am   -0.356331828 -0.24795238  -0.2706229   0.47296617  0.14665712
Rainfall       -0.331581354 -0.25021761   0.1501089   0.05600849  0.28901341
Cloud9am       -0.157552787 -0.12894408   0.3928416  -0.02642642  0.55163264
WindGustSpeed  -0.540180097 -0.52688524  -0.3497931   0.69394458 -0.06943918
Evaporation    -0.381905999 -0.39109295  -0.5195867   0.04860130 -0.39177965
```

Correlation Plot

- Size and color intensity
  - degree of correlation
- Color
  - positive or negative

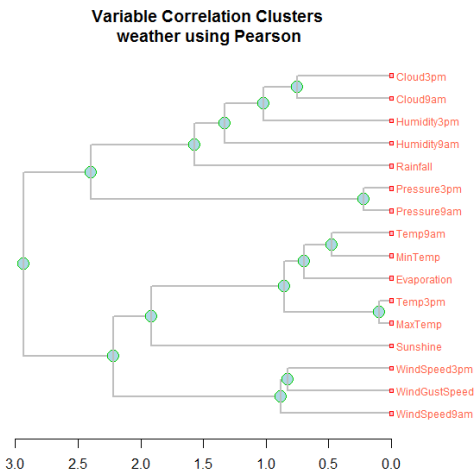


Missing Value Correlation



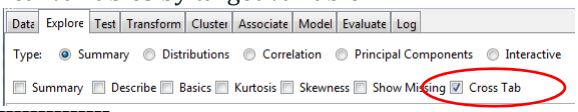
Hierarchical Correlation

- Very useful in identifying groups of correlated variables
- The height of the lines in dendrogram (along x-axis) indicates how strong the correlation is
  - Shorter height – stronger correlation



What about Categorical Variables?

- Cross Tab function from the Summary type at *Explore* Tab
- Cross tab of categorical variables by target variable




```
=========================================
                    crs$dataset[[crs$target]]
crs$dataset[[i]]        No      Yes   Total
-----------------------------------------|
No                     255       45     300
                     245.9     54.1
                     0.337    1.530
                     0.850    0.150   0.820
                     0.850    0.682
                     0.697    0.123
-----------------------------------------
Yes                     45       21      66
                      54.1     11.9
                     1.530    6.955
                     0.682    0.318   0.180
                     0.150    0.318
                     0.123    0.057
-----------------------------------------
Total                  300       66     366
                     0.820    0.180
=========================================
```

Mosaic Plot gives similar information

Tip: to explore against a categorical variable other than "RainTomorrow", simply select it as ***Target***

Exercise

- Context: a data mining project to monitor and perform an early forecast of blooms of certain harmful algae in rivers, so as to protect river lifeforms and water quality.
- Objectives:
  - To predict the frequency occurrence of several harmful algae in water samples
  - To provide a better understanding of the factors influencing the algae frequencies
- About the data:
  - Water samples were collected in different European rivers at different times during a period of approximately one year.
  - For each water sample, different chemical properties were measured as well as the frequency of occurrence of seven harmful algae. Some other characteristics of the water collection process were also stored, such as the season of the year, the river size, and the river speed.

Your tasks

- Get the data file "algae.RData"
- Load it into Rattle
- Explore your data
  - Data summary
  - Visualize the distribution of individual variables
  - Check pairwise correlation
- Any findings? (data skewness, outliers, missing data, correlation, etc.)