KE5107: Data Mining Methodology and Methods

Workshop: Data Preparation & Transformation

What we know about the dataset

- We are working on "weather" dataset from Rattle Library
- From our earlier workshop, we've discovered
  - There are missing values in the data
  - Some variables are skewed (e.g. Rainfall, WindSpeed9am)
  - There are variables with duplicated information (e.g. Temp3pm and MaxTemp, Pressure9am and Pressure3pm)
  - Some variables have more different distribution in records with "RainTomorrow =Yes" and records with "RainTomorrow =No"
  - "RISK_MM" is actually describing the amount of rain TOMORROW!

Missing Values in R

- In R, missing values are represented by the symbol **NA**
- Impossible values – NaN (not a number, e.g. dividing by 0)
- Testing of missing values – *is.na()* returns TRUE or FALSE
- Some functions have an option of ignoring missing values, like

  mean(mpg, na.rm=TRUE)
- To check which cases are complete, use function complete.cases(), returning a logical vector
  - Very useful for finding rows with missing values

    weather[!complete.cases(weather),]

Missing Values

- The situation of missing data in our dataset: most of the records are fine, and those with missing values are only missing 1 to 2 values.

	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	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
328	1	1	1	1	1	1	1
3	1	1	1				
1	1	1	1				
24	1	1	1				
1	1	1	1				
2	1	1	1				
7	1	1	1				
	0	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

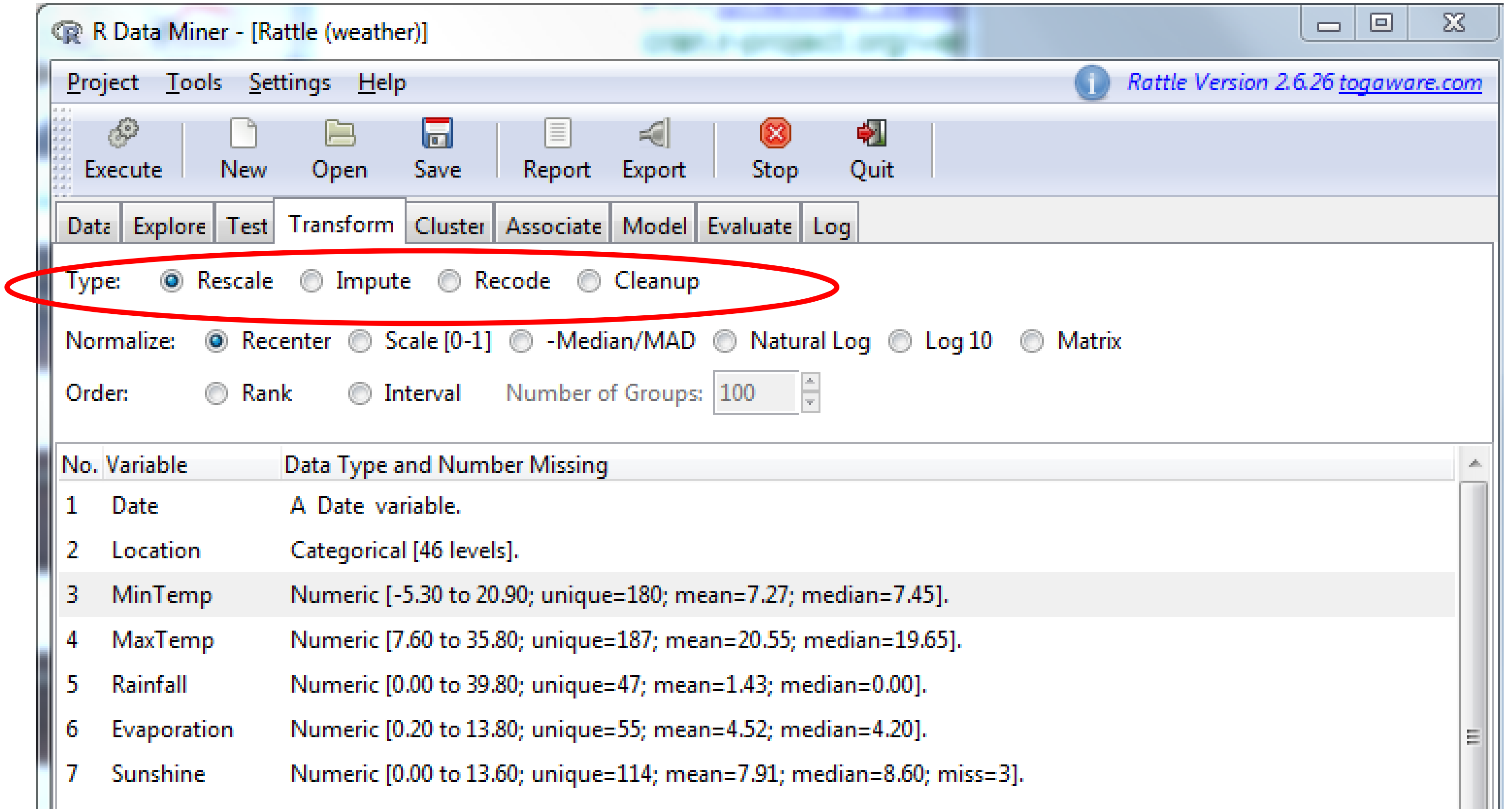
	WindSpeed9am	WindDir9am	
328	1	1	0
3	1	1	1
1	1	1	1
24	1	0	1
1	1	1	1
2	1	1	2
7	0	0	2
	7	31	47

Missing Values

- One variable have more missing values than others (WindDir9am)
- If this variable is very useful for prediction, we can try imputation. Otherwise, it can be ignored.

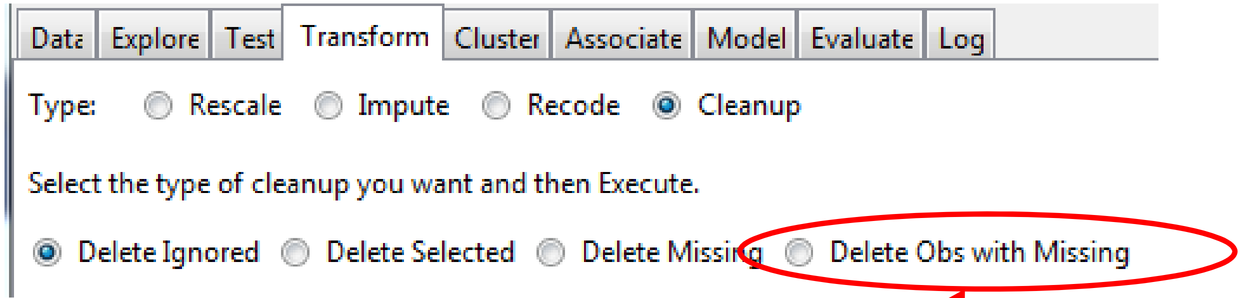
```
    RainToday RainTomorrow WindDir3pm WindGustSpeed Sunshine WindGustDir
328         1            1          1             1        1           1
  3         1            1          1             1        0           1
  1         1            1          1             1        1           0
 24 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

Rattle Transform Tab



To Cleanup Missing Values

- To delete columns and observations
  - Delete Ignored: remove any variables set as Ignore
  - Delete Selected: remove any selected variables
  - Delete Missing: remove any variables with missing values
  - Delete Obs with Missing: remove rows with missing values



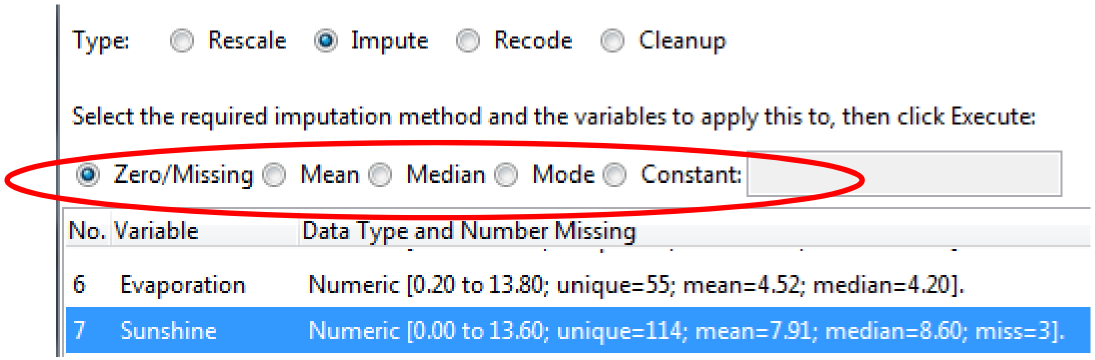
- If we want to remove observations with any missing values
- Code to remove rows with missing values

new_weather <- na.omit(weather)

Should we do that?

Imputation

- Dealing with missing values
  - Zero/Missing: replacing missing value with 0 (for numerical), or Missing (for categorical)
  - Mean/Median/Mode: use "central" value of the variable to reduce impact on the distribution (*mean* for generally normally distributed data, *median* for skewed data, *mode* for categorical)
  - Constant: to use your own default value



- Which method is better here?

Impute Sunshine

- Try Zero/Missing, Mean, Median imputation on Sunshine, which has 3 missing values
- Original Sunshine

  7 Sunshine Numeric [0.00 to 13.60; unique=114; mean=7.91; median=8.60; miss=3;

- After imputation. Which one is better?

  31 IZR_Sunshine Numeric [0.00 to 13.60; unique=114; mean=7.84; median=8.60].
  32 IMN_Sunshine Numeric [0.00 to 13.60; unique=115; mean=7.91; median=8.60].
  33 IMD_Sunshine Numeric [0.00 to 13.60; unique=114; mean=7.92; median=8.60].

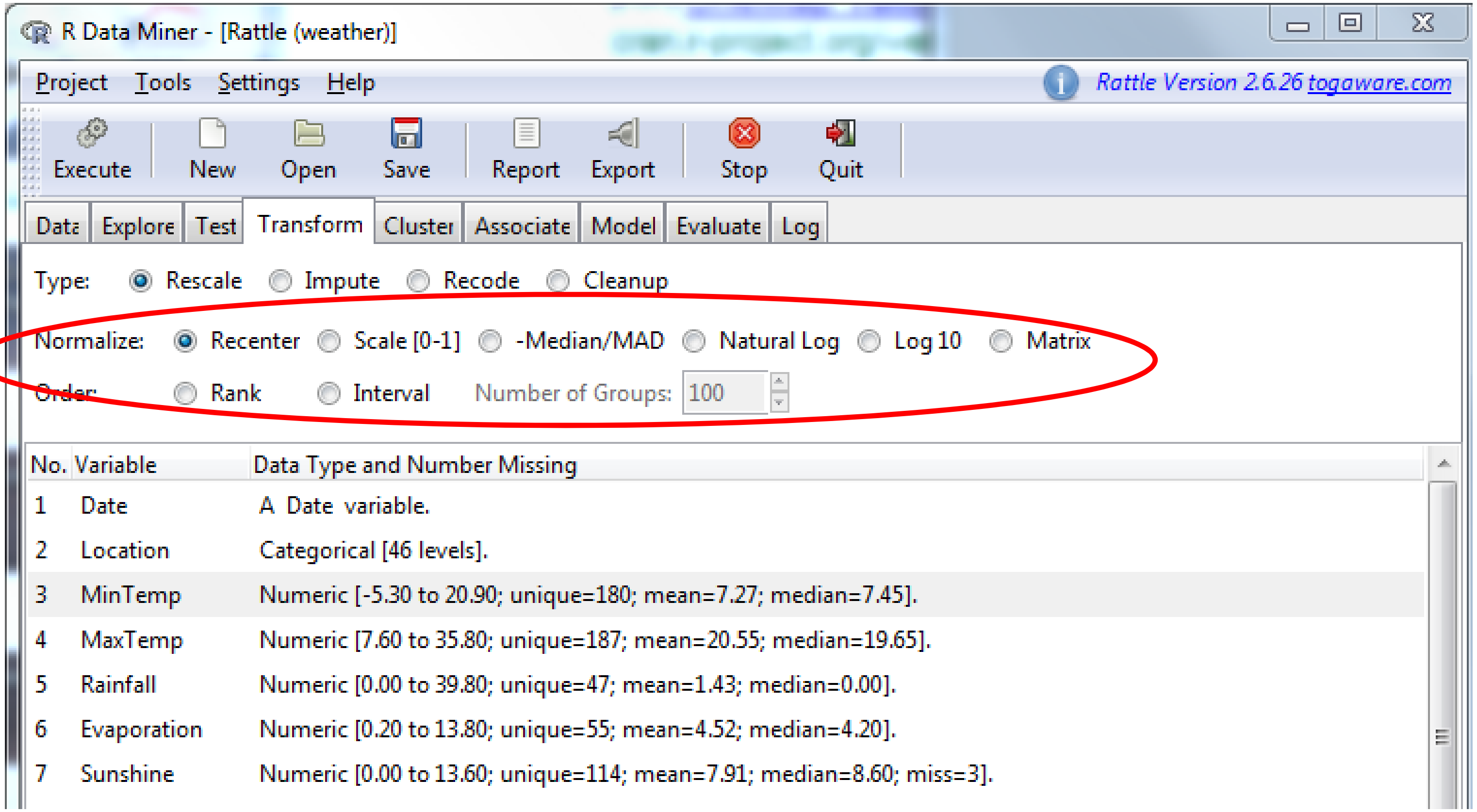
- Notice that after transformation, the original variable is automatically set to "Ignore"
- Exercise: Handle some other variables with missing values

View the Code

```
# Impute Sunshine.

crs$dataset[["IMD_Sunshine"]] <- crs$dataset[["Sunshine"]]

# Change all NAs to the median (not advisable).

if (building)
{
  crs$dataset[["IMD_Sunshine"]][is.na(crs$dataset[["Sunshine"]])] <- median(crs$dataset[["Sunshine"]], na.rm=TRUE)
}

# When scoring, transform using the training data parameters:

if (scoring)
{
  crs$dataset[["IMD_Sunshine"]][is.na(crs$dataset[["Sunshine"]])] <- 8.6
}
```

```
# Impute Sunshine.

crs$dataset[["IZR_Sunshine"]] <- crs$dataset[["Sunshine"]]

# Change all NAs to 0.

if (building)
{
  crs$dataset[["IZR_Sunshine"]][is.na(crs$dataset[["Sunshine"]])] <- 0
}

# When scoring, transform using the training data parameters:

if (scoring)
{
  crs$dataset[["IZR_Sunshine"]][is.na(crs$dataset[["Sunshine"]])] <- 0
}
```

Rescaling


R Data Miner - [Rattle (weather)]
Project Tools Settings Help
Rattle Version 2.6.26 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 Date A Date variable.
2 Location Categorical [46 levels].
3 MinTemp Numeric [-5.30 to 20.90; unique=180; mean=7.27; median=7.45].
4 MaxTemp Numeric [7.60 to 35.80; unique=187; mean=20.55; median=19.65].
5 Rainfall Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
6 Evaporation Numeric [0.20 to 13.80; unique=55; mean=4.52; median=4.20].
7 Sunshine Numeric [0.00 to 13.60; unique=114; mean=7.91; median=8.60; miss=3].

Rescaling

- Normalization
  - Recenter: Z score, the mean of scaled data is 0
  - Scale [0-1]: normalized to be in the range from 0 to 1
  - Median/MAD: robust rescaling around 0 using the median
  - Natural Log
  - Log 10
  - Matrix: transform multiple variables with one divisor
- Order
  - Rank: convert numbers into a rank ordering
  - Interval: rescale a variable according to some group that the observation belongs to
- Let's try a few methods on one variable *Temp3pm* for comparison (select one, click "Execute". Then repeat with another method.)

Rescaled Variables

- Rescaled variables are inserted into the table as new columns, with prefix indicating the kind of transformation

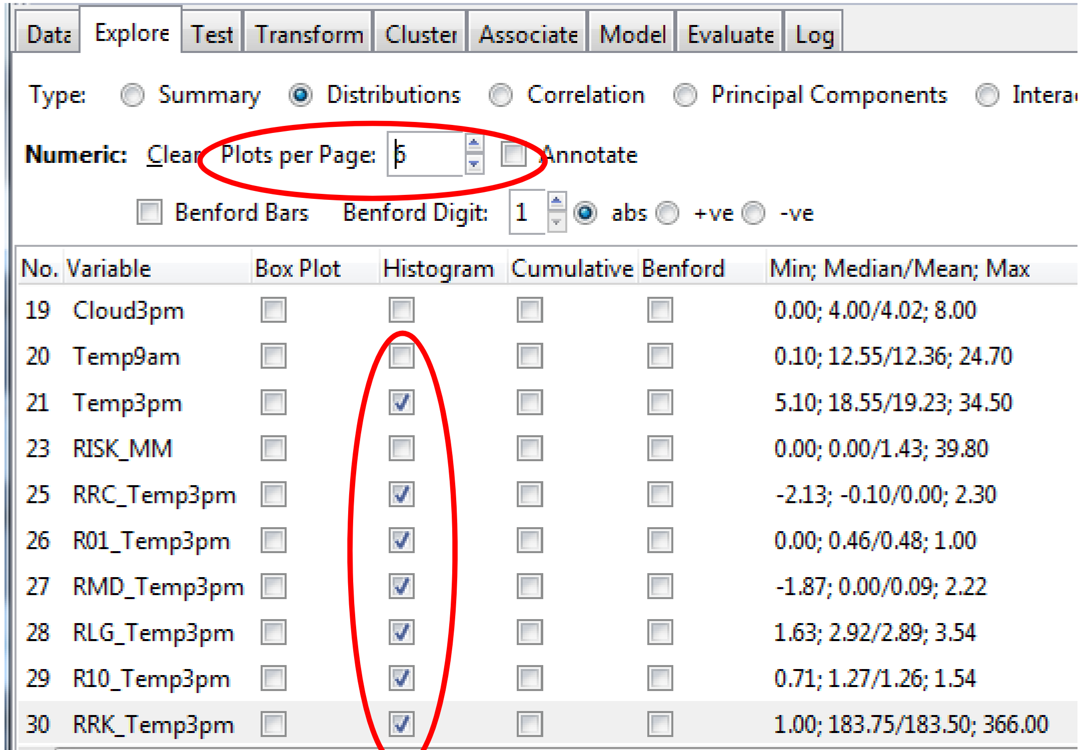
Original

Transformed

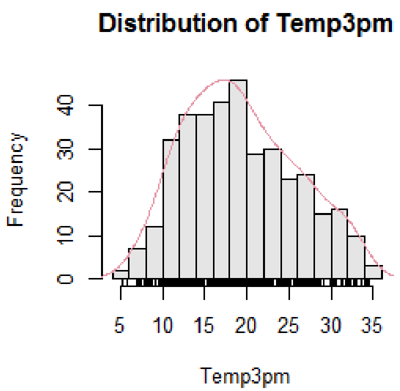
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55; ignored].
22	RainToday	Categorical [2 levels].
23	RISK_MM	Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
24	RainTomorrow	Categorical [2 levels].
25	RRC_Temp3pm	Numeric [-2.13 to 2.30; unique=200; mean=0.00; median=-0.10]. No code export.
26	R01_Temp3pm	Numeric [0.00 to 1.00; unique=200; mean=0.48; median=0.46]. No code export.
27	RMD_Temp3pm	Numeric [-1.87 to 2.22; unique=200; mean=0.09; median=0.00]. No code export.
28	RLG_Temp3pm	Numeric [1.63 to 3.54; unique=200; mean=2.89; median=2.92]. No code export.
29	R10_Temp3pm	Numeric [0.71 to 1.54; unique=200; mean=1.26; median=1.27]. No code export.
30	RRK_Temp3pm	Numeric [1.00 to 366.00; unique=200; mean=183.50; median=183.75]. No code export.

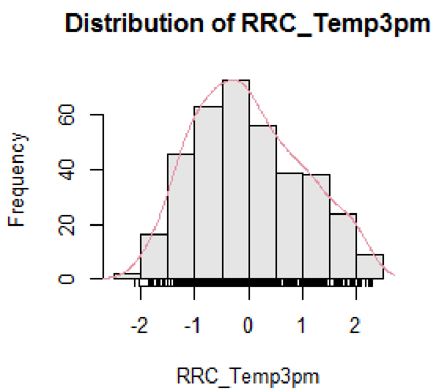
Check Distribution

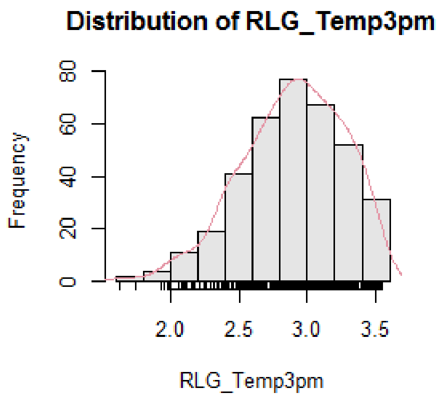
- At *Explore* Tab, select "Distributions" Type, and check the original Temp3pm variable, and its scaled versions. Change "Plots per Page" to 6. Click "Execute"

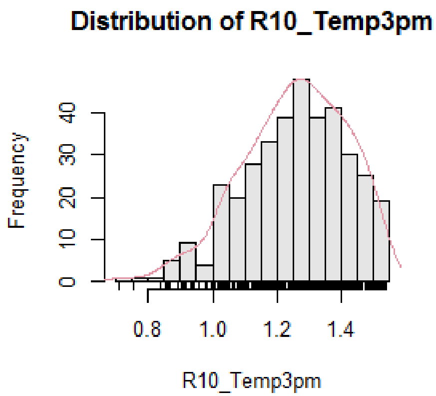


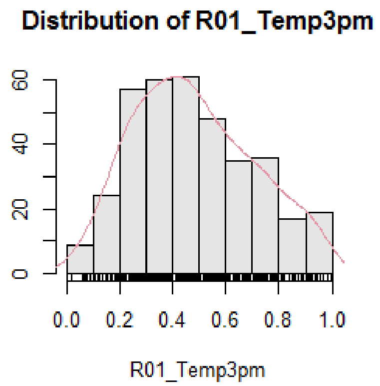
Distribution of Rescaled Variables

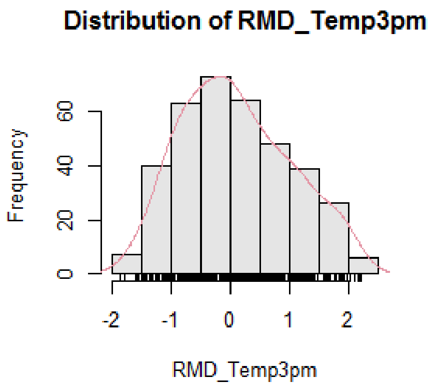


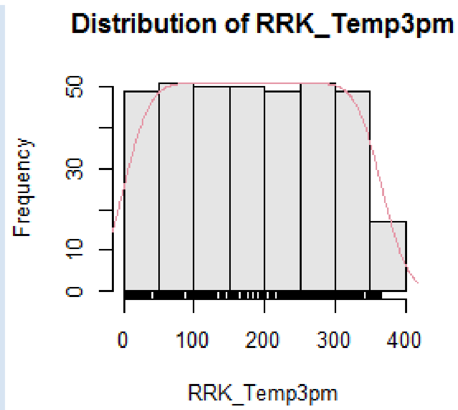












Rescaling Exercise

- Rescale the two variables that were found skewed (Rainfall, WindSpeed9am)

Sidetrack (you can do this after class)

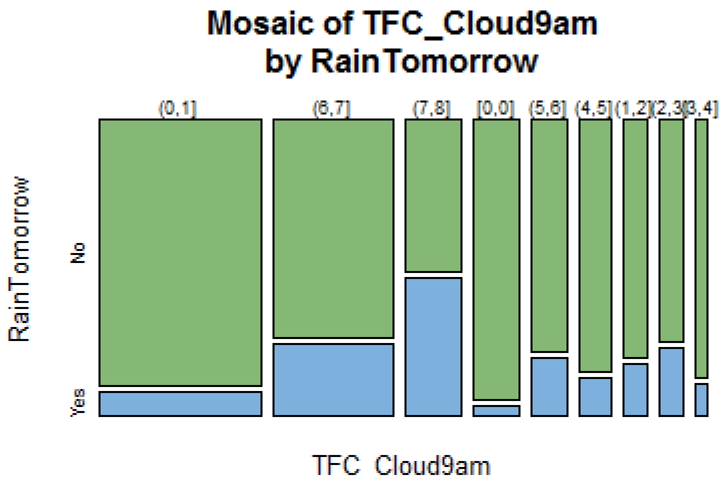
- Remember when we did Principle Component Analysis on mtcars dataset with method Eigen, the biplot doesn't look so right with original variables?
- Let's try rescaling all the variables to the same range [0-1]
- Do the PCA with Eigen again.
- Any change?

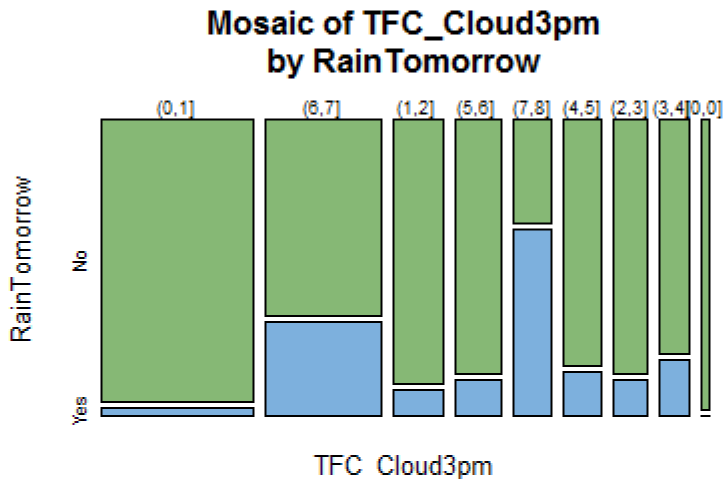
Recoding

- Binning
  - Transforming a continuous numeric variable into categorical values based on the numeric values, e.g. from Age to Age Groups
  - Can be useful in simplifying models, or for visualization
- Indicator Variables
  - Transform a categorical variable into a set of indicator(1/0) variables
  - Some model builders in Rattle (like Linear) do this automatically
- Join Categories
  - Stratify the dataset based on multiple categorical variables, e.g. from RainToday(yes/no) and RainTomorrow (yes/no), generate a new variable (yes_no/yes_yes/no_yes/no_no)
- Type conversion: As Categoric, As Numeric

Recoding

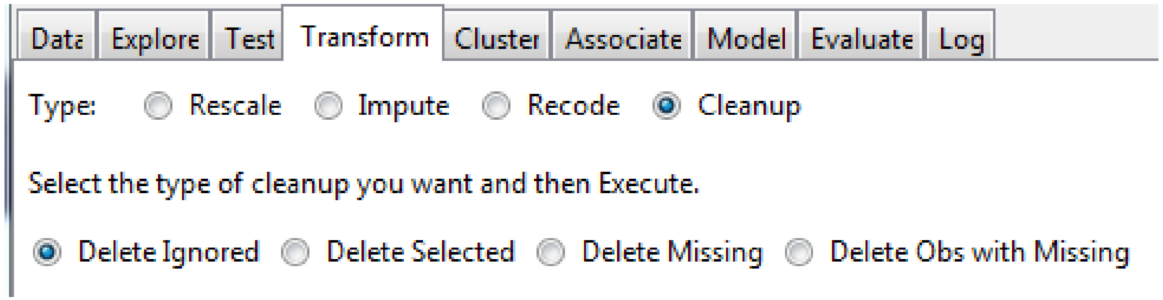
- Remember “Cloud9am” and “Cloud3pm” with <10 unique values? Would it be better to treat them as categorical variables?
- Let’s convert them using “As Categoric”
- Plot the converted variables. Any discovery? Should we keep the change?





Cleanup

- To delete columns and observations
  - Delete Ignored: remove any variables set as Ignore
  - Delete Selected: remove any selected variables
  - Delete Missing: remove any variables with missing values
  - Delete Obs with Missing: remove rows with missing values



Cleaning up

- Let's clean up the dataset by removing unwanted variables
- The code of removing a variable is straight forward

```
# CLEANUP the Dataset

# Remove specific variables from the dataset.

crs$dataset$R01_Temp3pm <- NULL
crs$dataset$RMD_Temp3pm <- NULL
crs$dataset$RLG_Temp3pm <- NULL
crs$dataset$R10_Temp3pm <- NULL
crs$dataset$RRK_Temp3pm <- NULL
crs$dataset$IMN_Sunshine <- NULL
crs$dataset$IMD_Sunshine <- NULL
```

- Removing rows with missing values

  new_weather <- na.omit(weather)

Exporting Transformed Data

- Then click “Export” button, and save the transformed dataset as “weather_transformed.csv”