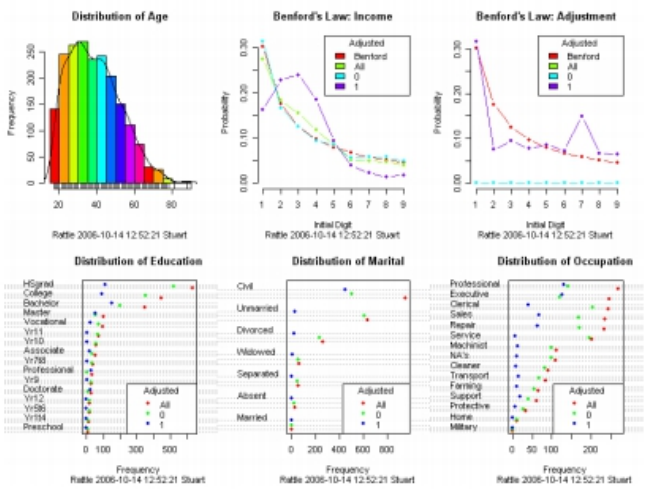
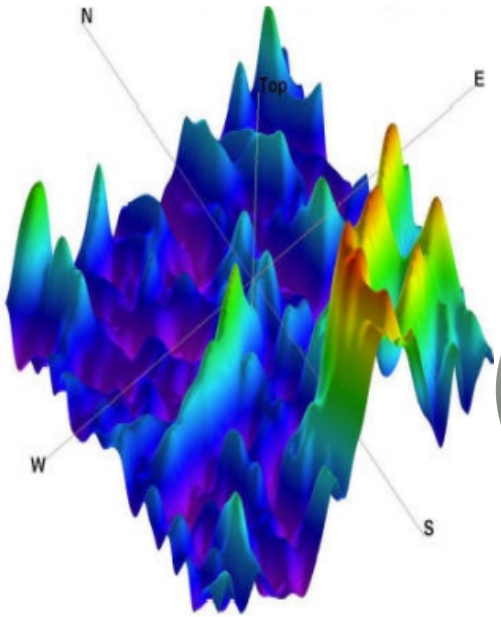
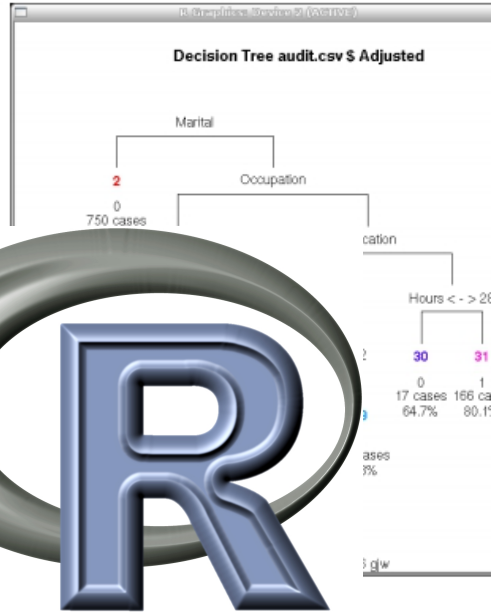
Reinforce BUSINESS INTELLIGENCE™

www.reinforcebi.com

Introduction to R Data Mining.(Rattle)







The national institute of development administration (NIDA)

M.S.Applied Management (Honors)

Naresuan University

B.Eng Mechanical

Mr.Amornthep Thongchiw

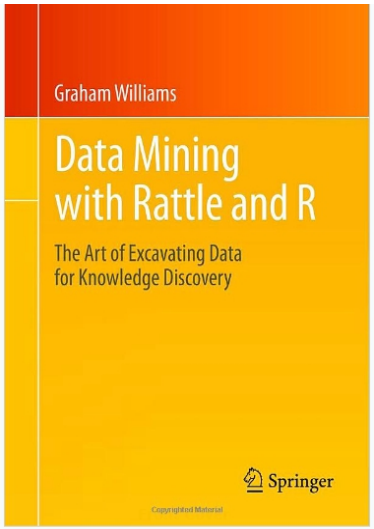
แนวคิดในการอบรม Analytic Open Source.

- 1.เราแนะนำส่งเสริม Open Source เพื่องดใช้โปรแกรมเถื่อนในการทำงาน (***แต่ส่งเสริมให้ใช้โปรแกรมลิขสิทธิ์***)
- 2.เรากำลังแนะนำความสามารถพื้นฐาน Open Source (ประมาณ 5-10%) อย่าเพิ่งไปเทียบกับ 100% Commercial ที่ใช้มานับ 10 ปี
- 3.เรากำลังแนะนำการใช้โปรแกรมสถิติและดาต้าไมนิ่ง ซึ่งล้วนแต่มีทฤษฎีอยู่เบื้องหลัง
- 4. Open Source Not Always Free ;
  Time & Learning will be need.
- 5. โปรแกรมนี้คือ R data Miner [Rattle] ซึ่งออกแบบมาให้ง่ายต่อการใช้

1.Introduction to Data Mining & CRISP-DM

Becoming an educated buyer and user of Business Analytics software

Stephen Samild
stephen@presciient.com
Presciient

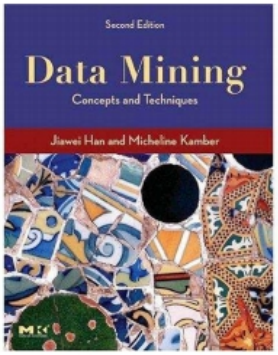


Data Mining Algorithms

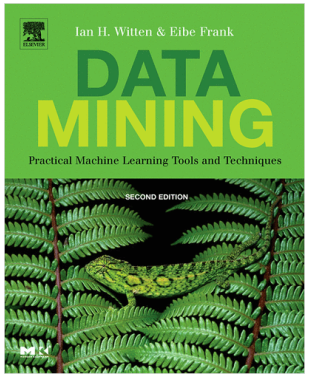
Graham Williams

Principal Data Miner, ATO
Adjunct Associate Professor, ANU

Ian H.Witten & Eibe Frank
Data Mining : Practical machine Learning Tools and Techniques



Data Mining: Concepts and Techniques
Jiawei Han, Micheline Kamber
2006, Morgan Kaufmann Publishers
ISBN: 1558609016.



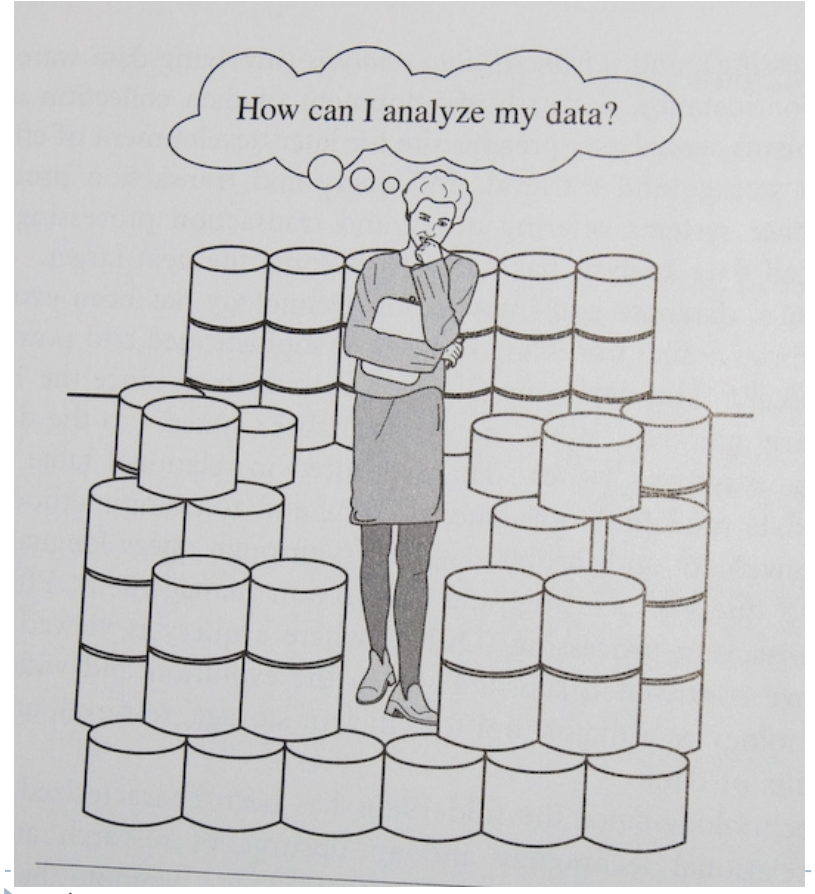
<u>เอกสารประกอบการสอน การทำเหมืองข้อมูลด้วยโปรแกรม Weka</u>

1.ผศ.ดร. กรุง สินอภิรมย์สราญ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย (2551)

2.ดร.อนันตพร ศรีสวัสดิ์ สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์ ม.พระจอมเกล้าเจ้าคุณทหารลาดกระบัง (2553)

Why Data mining?

การใช้ข้อมูลเป็นสิ่งจำเป็น ผู้ที่สามารถนำข้อมูลมาใช้ได้อย่างชาญฉลาดกว่า รวดเร็วกว่า ก็จะสร้างข้อได้เปรียบทางการแข่งขันเหนือผู้อื่น



แล้วเราจะมีวิธีการวิเคราะห์ข้อมูลอย่างไร ?

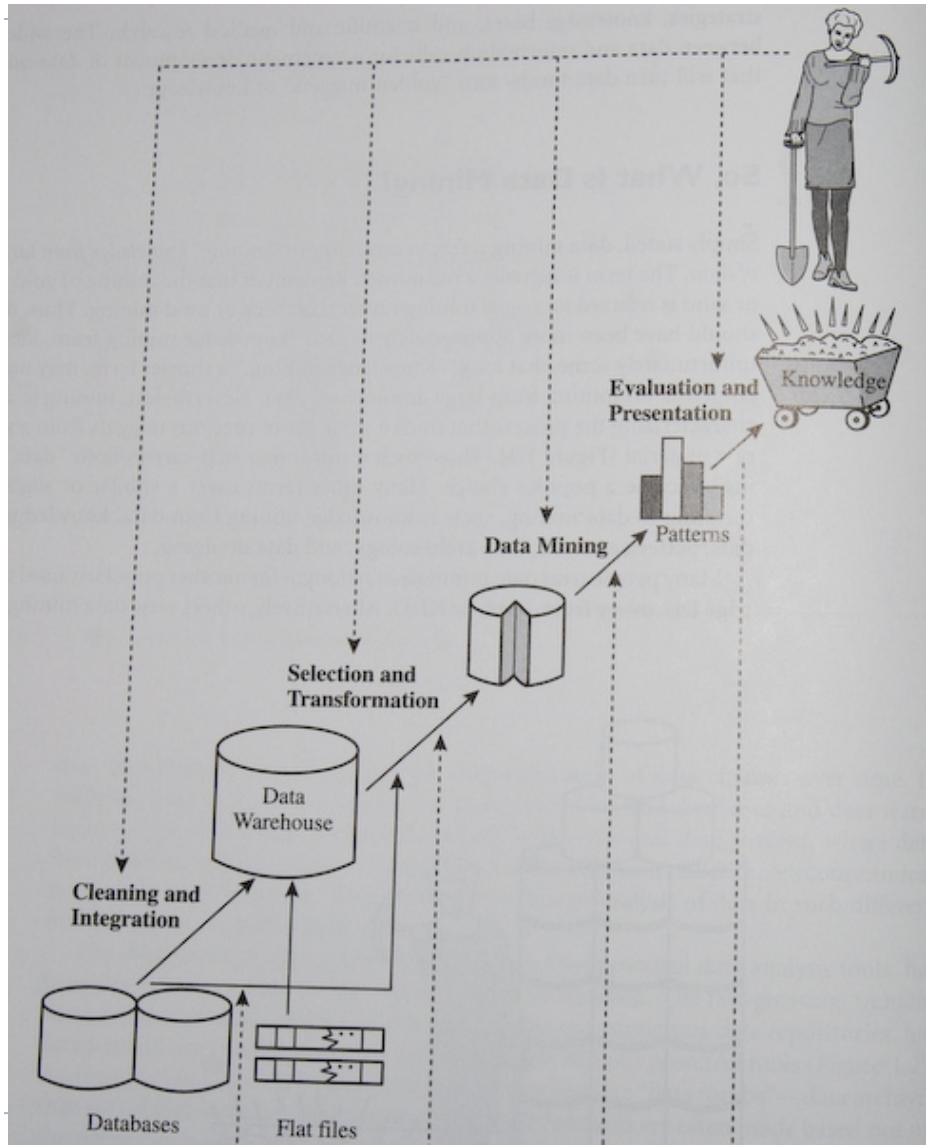
We are drowning in data,
but starving for knowledge!

Why Data mining?

การวิเคราะห์ข้อมูลสถิติและ **Data mining** ช่วยในการตัดสินใจได้เป็นอย่างดี ด้วยวิทยาศาสตร์ในการวิเคราะห์ข้อมูลสมัยใหม่ เฉกเช่นเดียวกับองค์กรระดับโลก อาทิ **Amazon ,Ebay ,google, facebook ,etc.** ที่มีระบบการวิเคราะห์ข้อมูลเพื่อนำเสนอการให้บริการแบบ **Mass-customization**

ซึ่งสร้างความประทับใจให้กับลูกค้าได้มากกว่า

ไม่ว่าคุณจะเป็นนักการตลาด ผู้จัดการฝ่ายขาย นักวิเคราะห์ข้อมูล นักวางกลยุทธ์ แพทย์ วิศวกร ฯลฯ คุณก็สามารถใช้เทคนิคการวิเคราะห์ข้อมูลทางสถิติและ **Data mining** ได้ทุกวัน



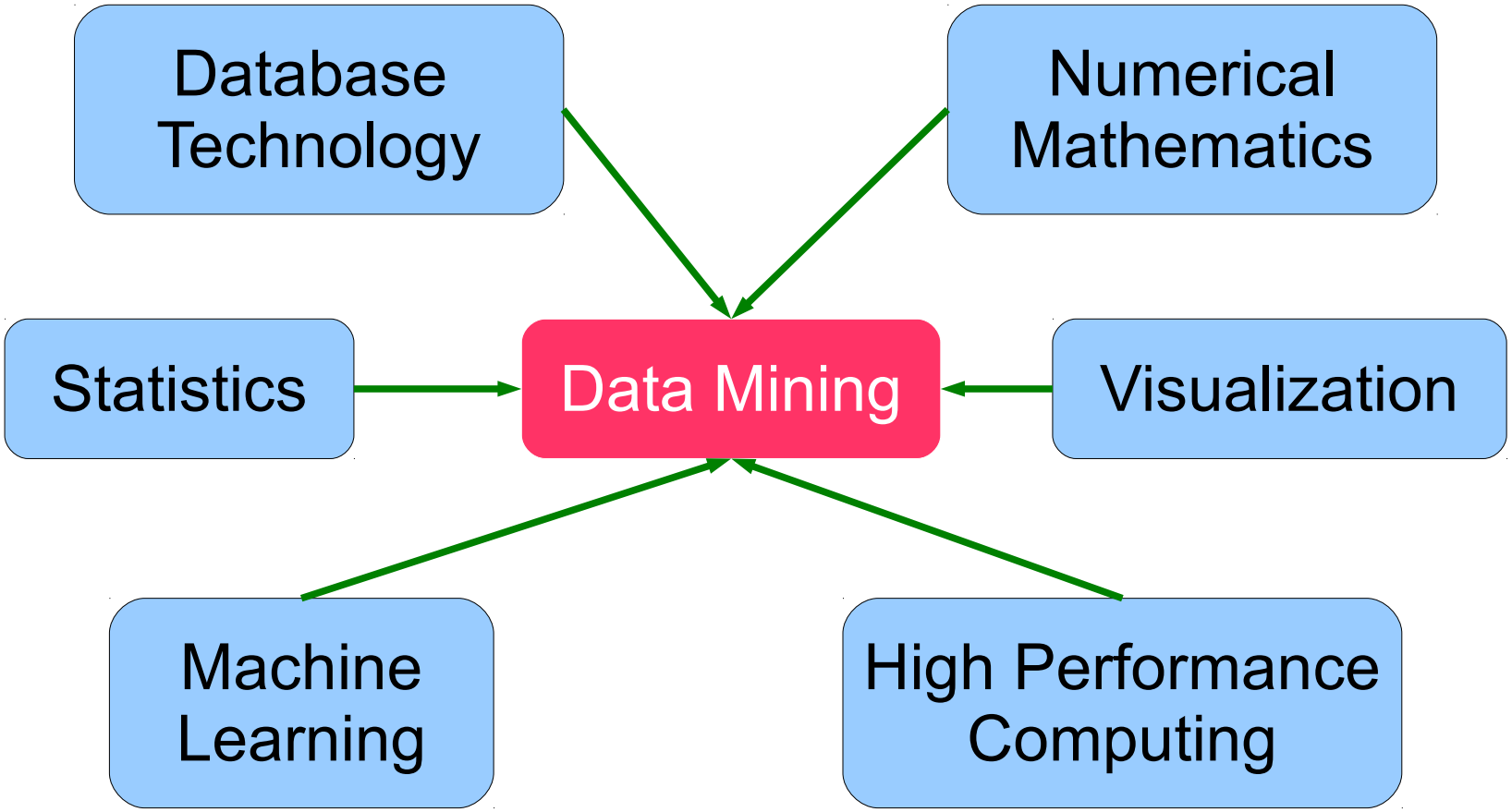
Data mining คืออะไร?

- **Knowledge discovery in database** is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
  (Fayyad, Piatetsky-Shapiro and Smyth, 1996)
- An information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.
  (http: //www.twocrows.com/glossary.htm)
  - Try also on www.google.com "data mining"

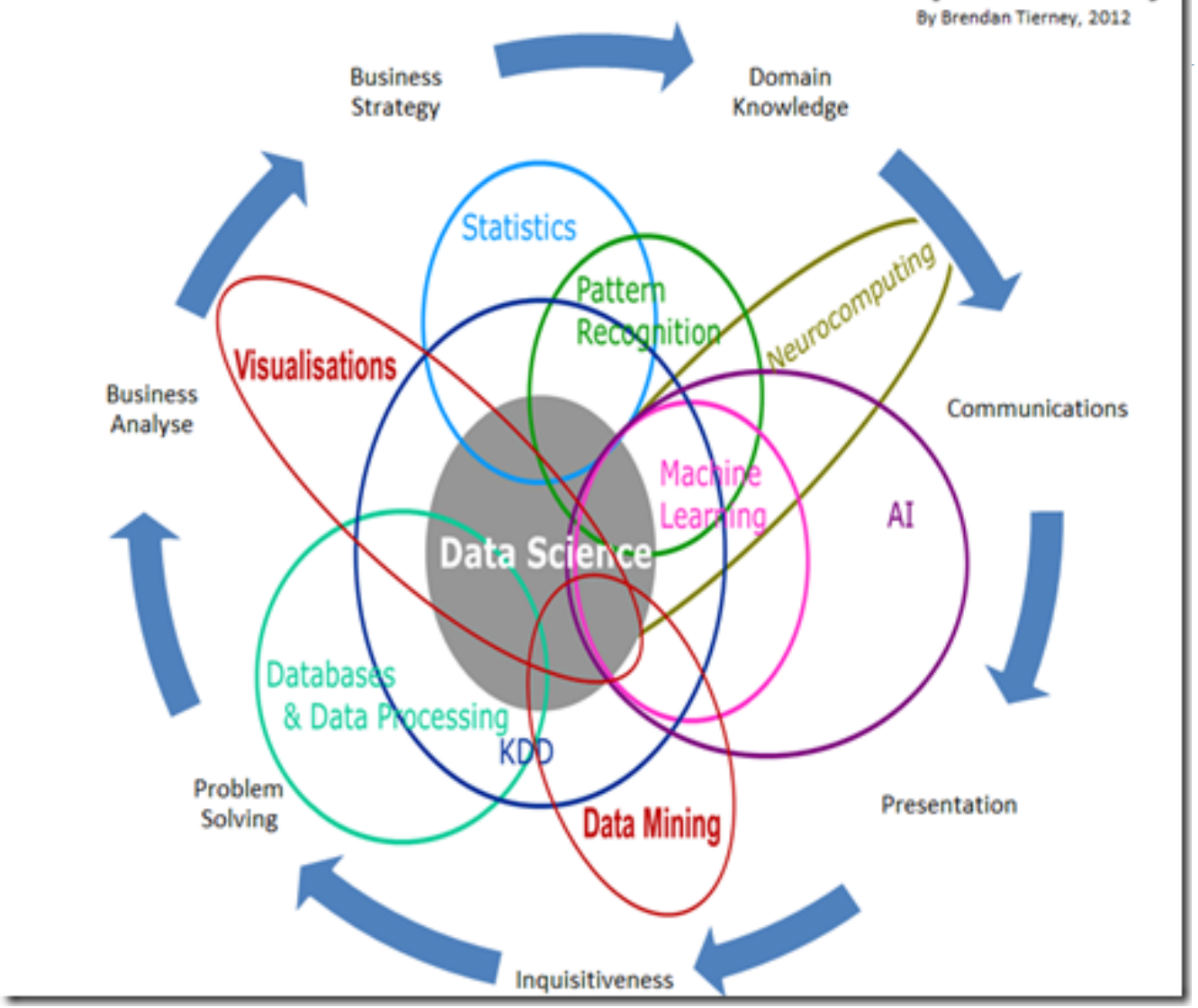
Data mining คืออะไร?

- Data mining
  - Extraction of interesting information or patterns from data in large databases
  - Data mining เป็นการค้นหาความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge discovery from very large Databases : KDD)
  - Data mining เป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ เพื่อดึงความรู้ที่สำคัญออกมา

Data mining is multi disciplinary



Data Science Is Multidisciplinary



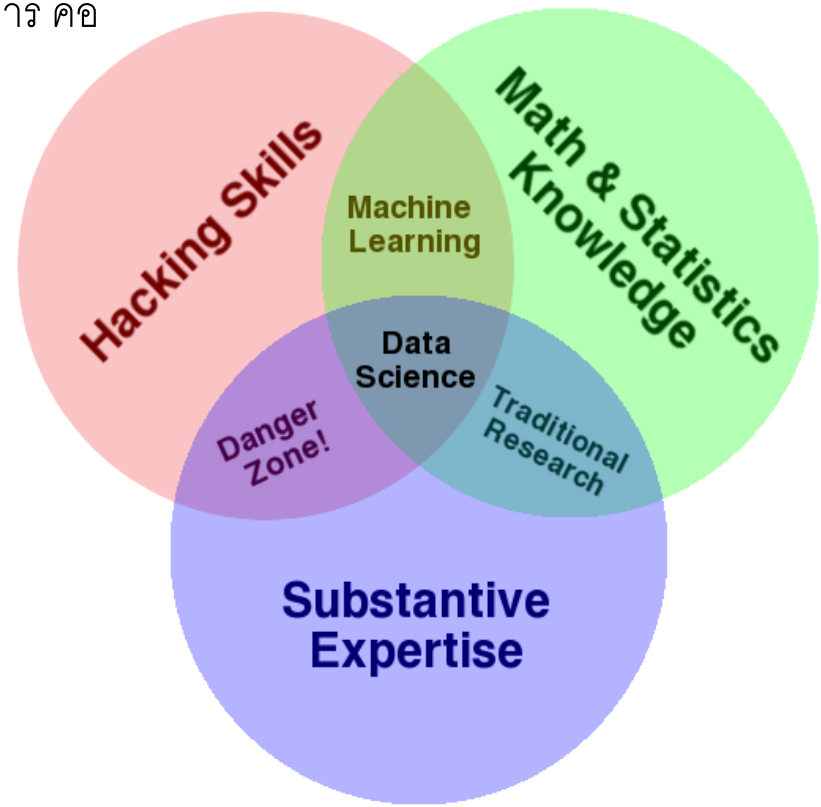
http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html

ขยายความ Data mining

เทคโนโลยี **Data mining & Analytic** นี้นิยมใช้กันมากในองค์กรชั้นนำต่างๆ บ่อยครั้งเราจะรู้จักกันในนาม **Business Intelligence** หรือ **BI Tools.** ซึ่งมีส่วนประกอบสำคัญ 2 ประการ คือ

1.ความรู้ทาง Data Science และ

2.Data analytic Software.



source : http://www.dataists.com/2010/09/the-data-science-venn-diagram/

ขยายความ Data mining

- Data Science ต้องใช้ความรู้ 3 ด้าน

  1. Hacking คือ ความรู้ด้านคอมพิวเตอร์ ฐานข้อมูล เพียงแค่รู้จักนำเข้าข้อมูล / ใช้ Filter /ใช้ Query ที่มีจาก Excel ,Accessโปรแกรมบัญชี ,SQL,Oracle ฯลฯ ก็ใช้ได้แล้ว

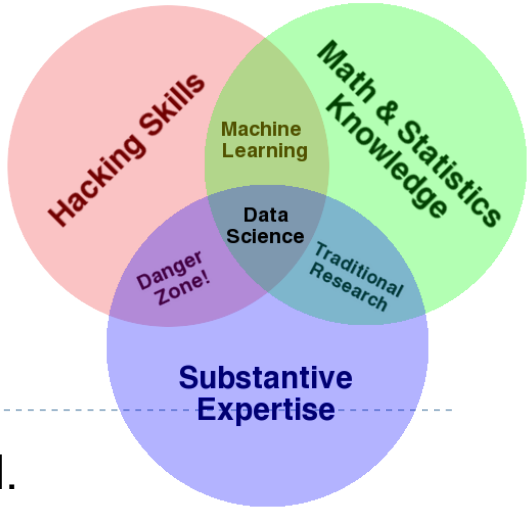
  2. Math & Statistic knowledge คือ ความรู้ด้านคณิตศาสตร์และสถิติ เพียงแค่สามารถ Run multiple regression ออกมาตีความหมายได้ก็เพียงพอแล้ว

  3. Substantive expertise คือความรู้ด้านการตีความ ความเข้าใจในปรากฏการณ์ในแบบ Real life Real world ถ้าเราอยู่กับธุรกิจหรือหน่วยงานใดๆ มาเกิน 5 ปี เราจะพบว่าตัวเรามีความชำนาญ

ขยายความ Data mining

จาก Venn diagram นี้เราจะสังเกตเห็นส่วนซ้อนทับ (Intersection) 4 ส่วน คือ

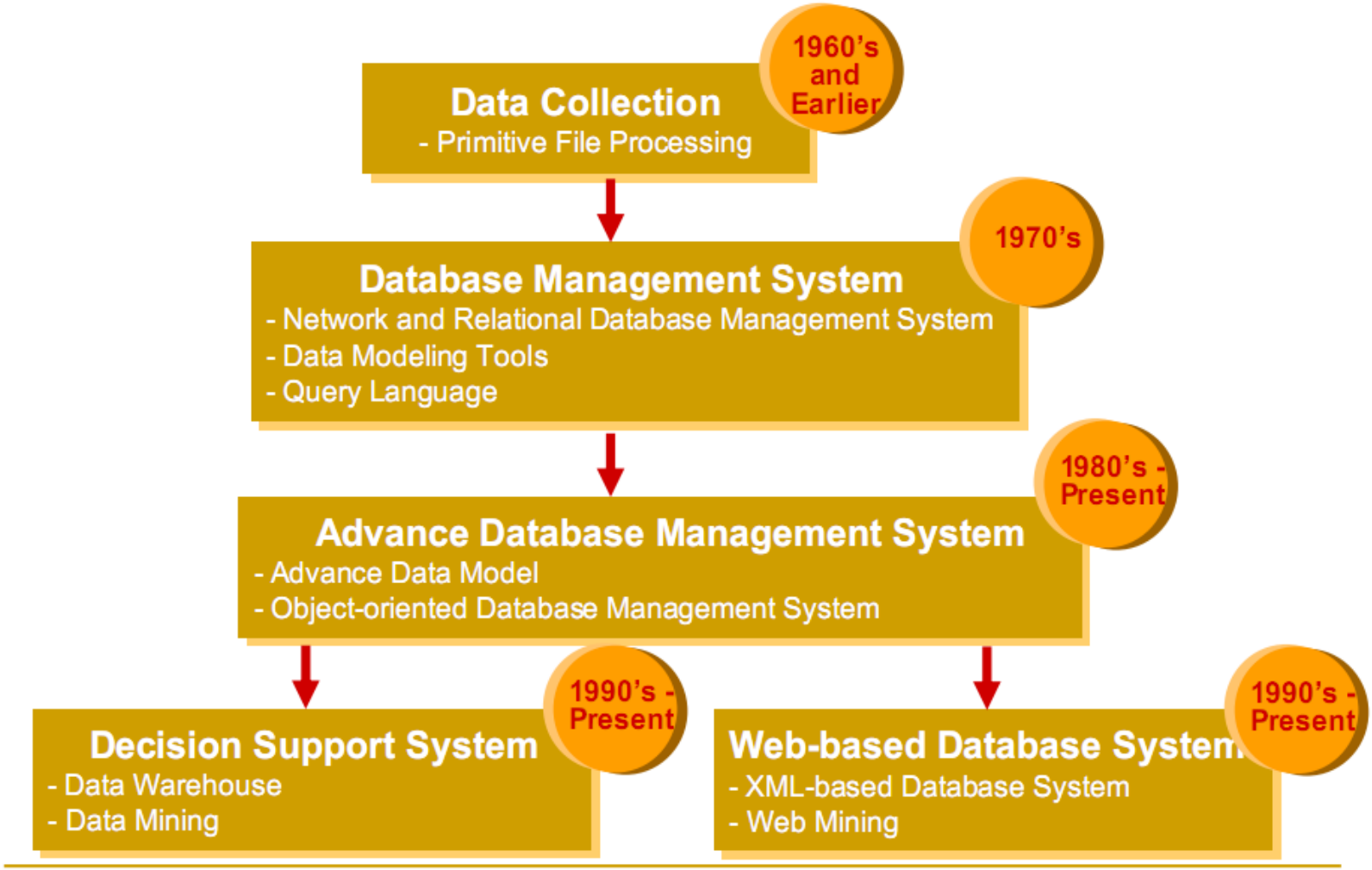
1. Machine learning คือส่วนของ Hacking กับ Math & Statistic knowledge
2. Traditional Research คือส่วนของ Math & Statistic knowledge กับ Substantive expertise
3. Danger Zone คือส่วนของ Substantive expertise กับ Hacking ผู้ที่มีความรู้แค่ 2 ด้านนี้ตกอยู่ในโซนอันตราย
4. Data science ต้องการความรู้ทั้ง 3 ด้าน คือ Hacking ,Math & Statistic knowledge & Substantive expertise



สิ่งที่ได้จาก Data mining.

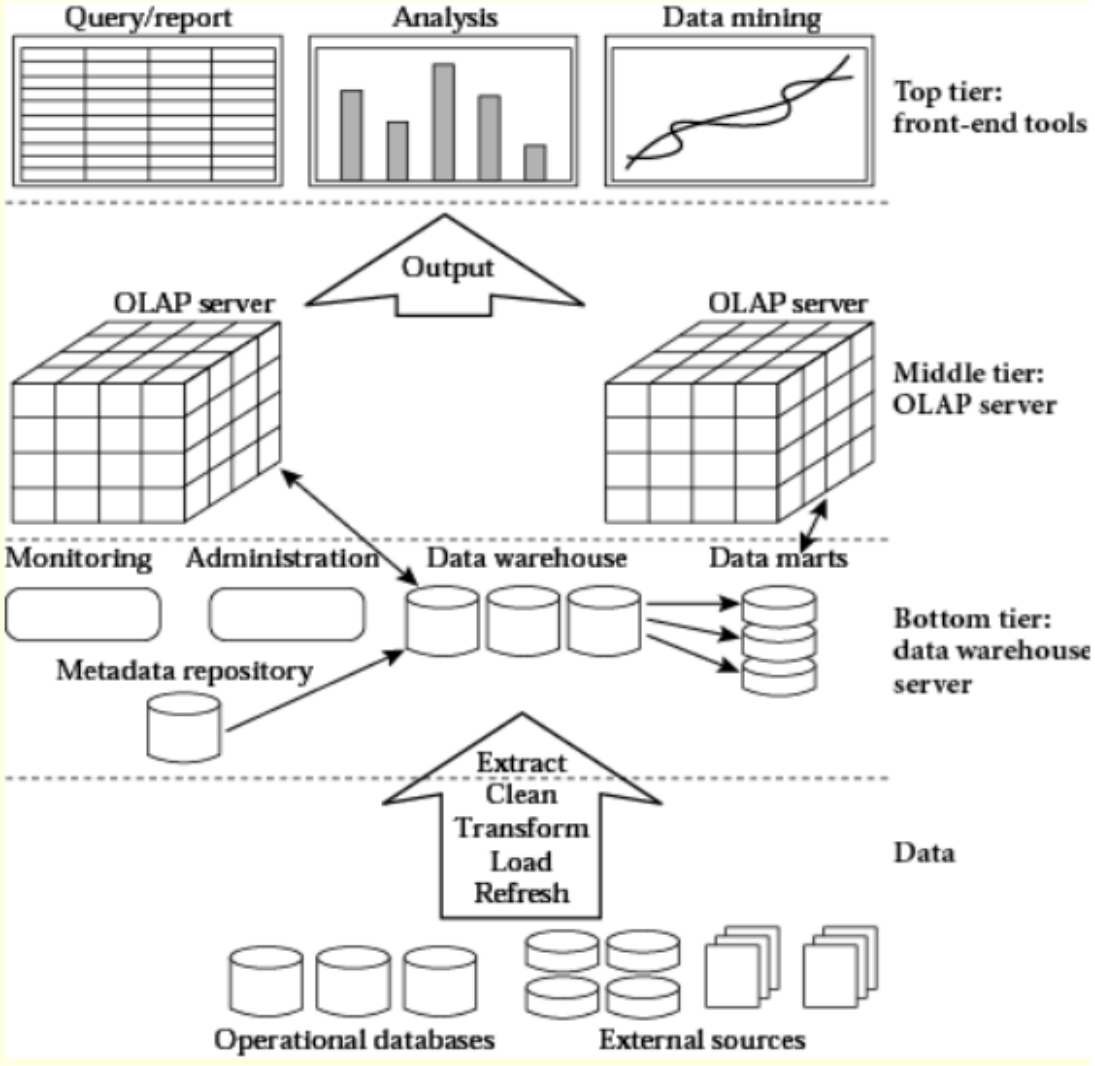
- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a timely manner.
- Anytime Data Mining – give me the answer now.
  - Data exploration
  - Data visualization
  - Very large data extractions
  - Data analysis
  - Hypothesis generation
  - Model building
- **Knowledge.**

วิวัฒนาการของการวิเคราะห์ข้อมูล บนฐานข้อมูล



Source: refer to slide of Dr.Anantaporn Srisawat
Reinforce business intelligence Co.,Ltd.

สถาปัตยกรรมของการวิเคราะห์แบบ BI.



Data mining and Business Intelligence.

"Business Intelligence"

Query and reporting

Planning and budgeting

Data Warehousing

OLAP

Forecasting

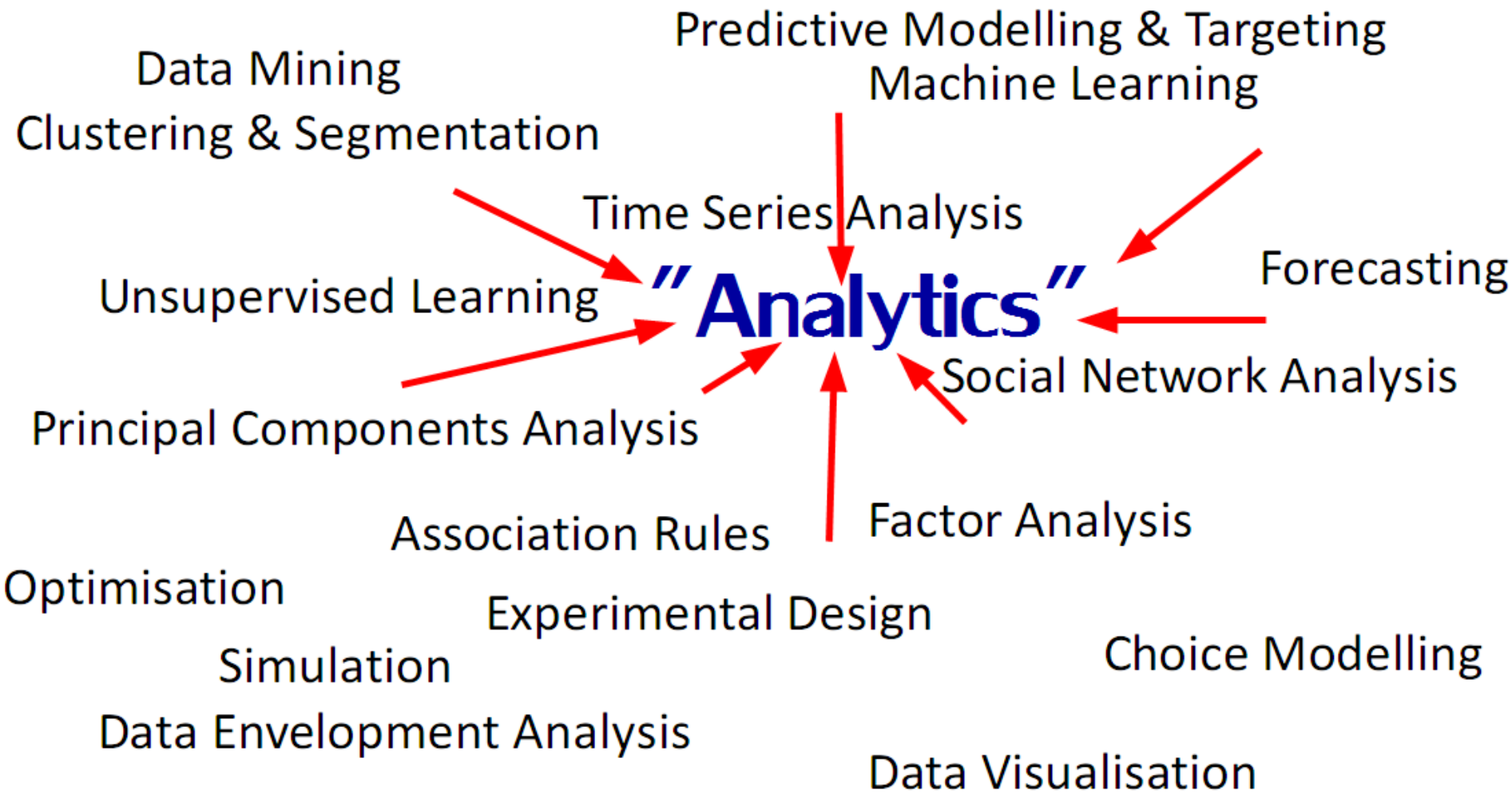
Scorecarding

Dashboarding

ETL

Master Data Management

Data mining and Analytics.



Data mining Algorithms.

The CRISP-DM Model

- Cross Industry Standard Process for Data Mining

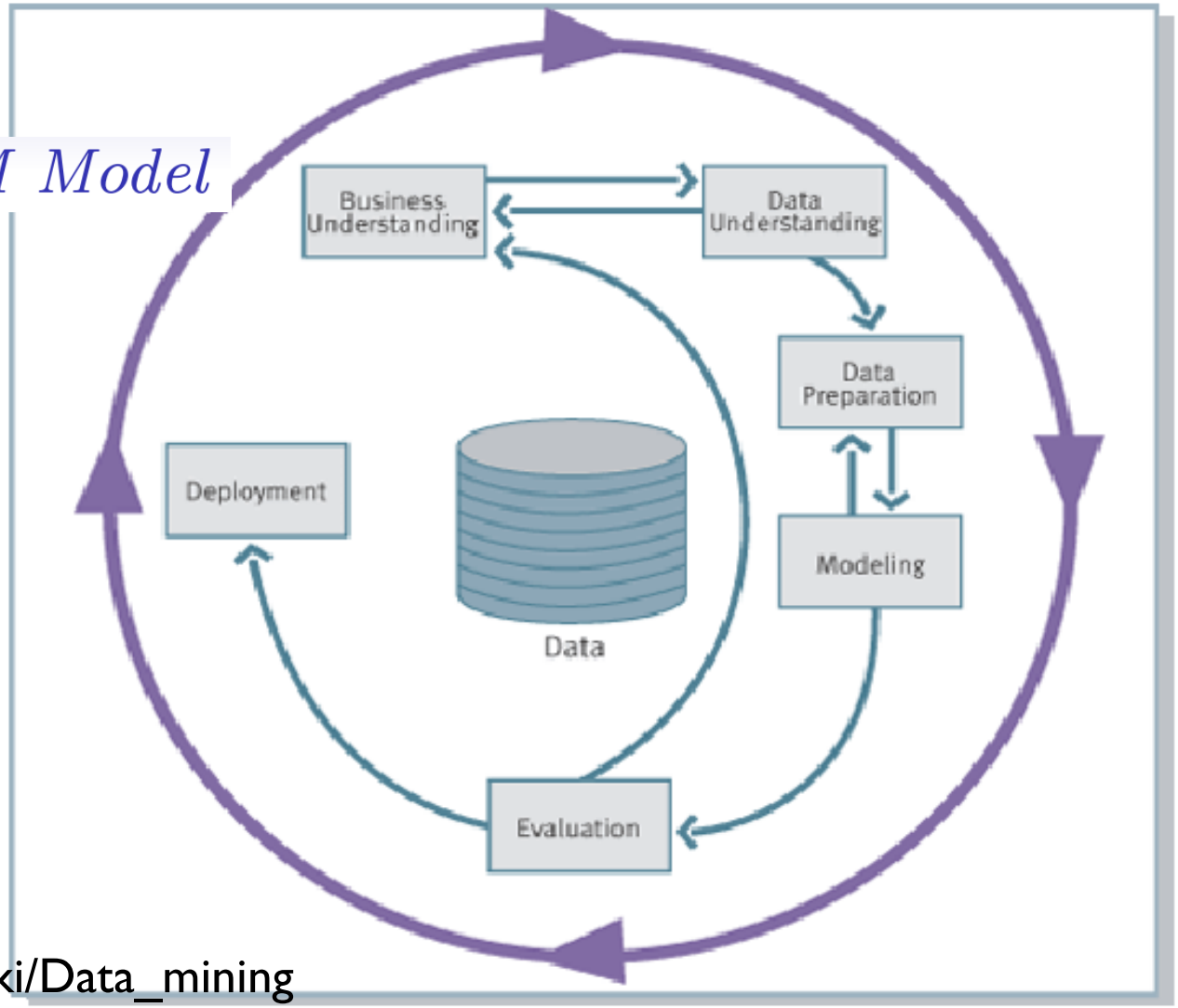
http://www.crisp-dm.org/

- Developed by NCR, Daimler-Benz, ISL, OHRA
- Define and validate a Data Mining Process Model.
  - Applicable in diverse industry sectors
  - Industry and tool neutral
  - Large data mining projects executed faster, cheaper, more reliably and more manageably
- Life cycle of six (iterative) phases.

Source: togaware

Cyclic nature of data mining.

The CRISP-DM Model

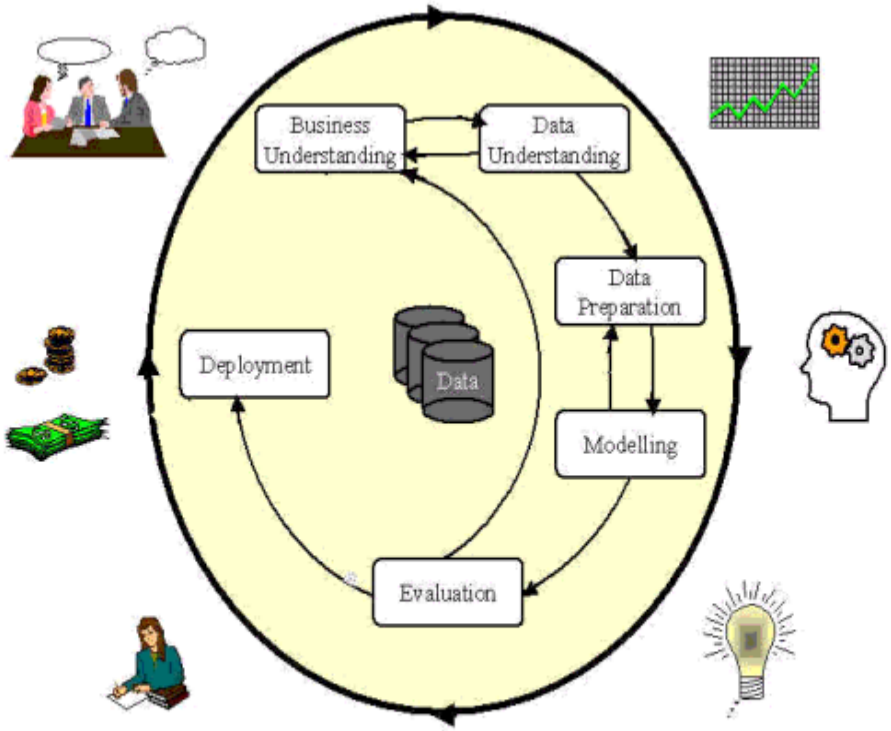


http://en.wikipedia.org/wiki/Data_mining

Cyclic nature of data mining.

Every step of a data mining process can lead to revisiting any one of the previous steps.

A DM process continues after a solution has been deployed. The lessons learnt can trigger new, often more focused business questions. Subsequent data mining processes benefit from experiences of previous ones.



Source: togaware

Six steps of CRISP-DM

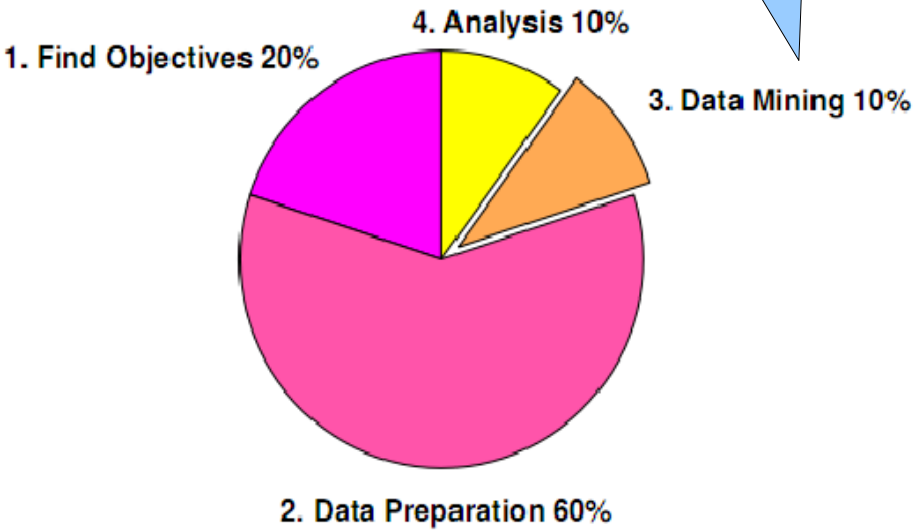
1. Business Understanding (25%)
2. Data Understanding (20%)
3. Data Preparation (25%)
4. Modelling (10%)
5. Evaluation (20%)
6. Deployment

Graham Williams

Principal Data Miner, ATO
Adjunct Associate Professor, ANU

Source: togaware

Data mining is a small Process.



Six steps of CRISP-DM

1.Business Understanding

- We had better make sure we are addressing a real business problem.
  - Initial phase focuses on understanding project objectives and requirements from a business perspective.
  - This knowledge is converted into a data mining problem definition.
  - Develop a preliminary plan designed to achieve the objectives.

Source: togaware

Six steps of CRISP-DM

2. Data Understanding

- Understand what data is available and its semantics.
  - Initial data collection
  - Familiarisation with the data
    - Identify data quality problems.
    - Discover first insights into the data
    - Detect interesting subsets to from hypotheses for hidden information

Source: togaware

Six steps of CRISP-DM

3. Data Preparation.

- Bring together the data – get it into shape for mining.
  - Construct the mining dataset
  - Derived from the initial raw datasets
  - Data preparation tasks:
    - Table, record, and attribute selection
    - Generation of derived features
    - Data transformation
    - Data cleaning

Source: togaware

Six steps of CRISP-DM

3. Data Preparation (con't)

- Issues to be dealt with include:
  - Data Quality
    - Missing data
    - Noisy data
    - Lead to inconsistent or too general / specific discoveries.
  - Data Cleaning
    - Duplicates
    - Inconsistencies
    - Identify and merge the same entities. Source: togaware

Six steps of CRISP-DM

4. Modeling

- Now the "data mining" begins!!
  - Select various modeling techniques
  - Apply and calibrate modeling techniques
  - Typically there are several techniques for the same data mining problem
  - Some techniques have specific requirements on the form of data and require stepping back to the data preparation phase

Source: togaware

Six steps of CRISP-DM

5. Evaluation

- How do we know we have a useful outcome?
  - Evaluate the model and review the steps executed to construct the model
  - Does the model properly achieve the business objectives?
  - Is there some important business issue that has not been sufficiently considered?
  - Decide on the use of the data mining results

Source: togaware

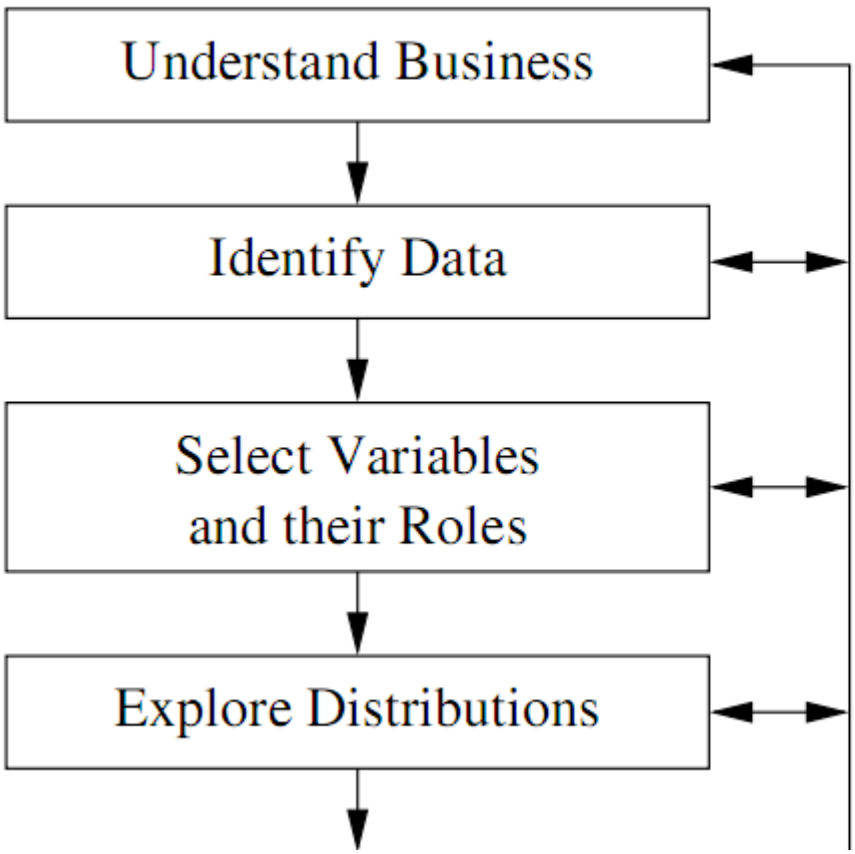
Six steps of CRISP-DM

6. Deployment

- No point to data mining unless we action the outcomes.
  - Deployment may be:
    - Generate a report of the discoveries made
    - Implement changes in the processes of the organization
    - Implement a repeatable data mining process
  - For successful deployment the customer must understand the actions to be carried out in order to actually make use of the created models.

Source: togaware

Six steps of CRISP-DM (application1-2)



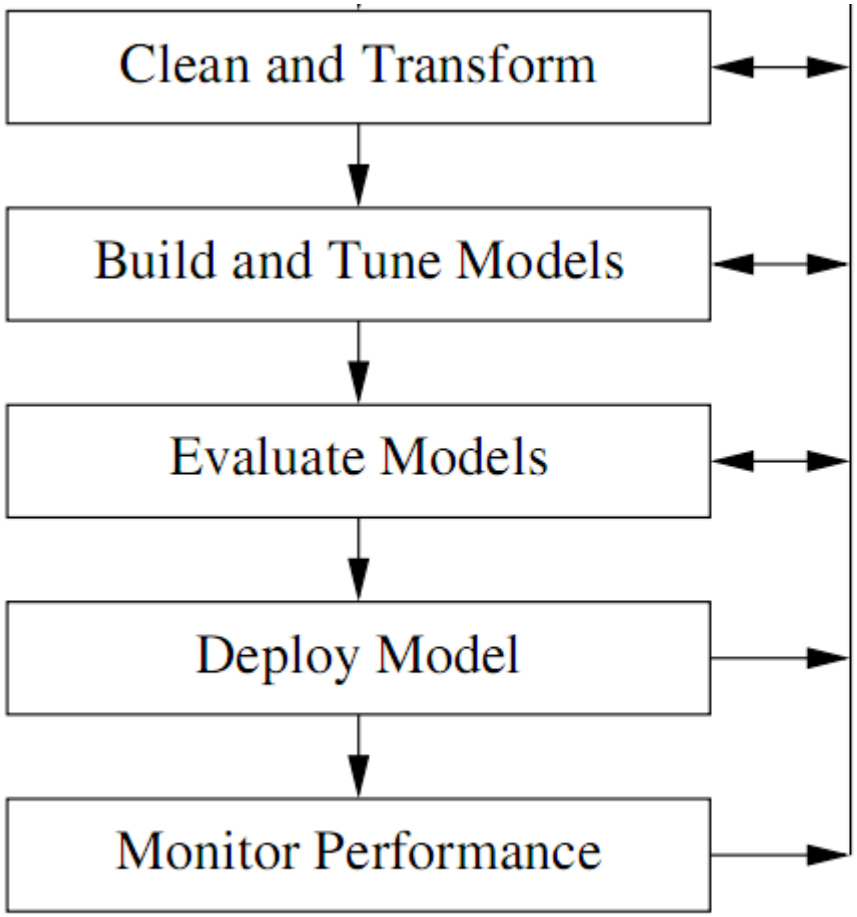
Understaning the business may itself entail a mini data mining project of a few days.

Start with getting as much data as you can and then cull.

Exploring data is important for us to understand its shape, size, and content.

Source: togaware

Six steps of CRISP-DM (application2-2)



We may loop around here many times as we clean, transform, then build and tune our model.

Evaluate performance, structure, complexity, and deployability.

Is the model run manually on demand or automatic schedule? Is it deployed standalone or integrated into current systems?

Source: togaware

Data Mining Strategies

- Supervised Learning
  - Process of building classification models using data instances of known origin to learning with predefined classes
  - Build models by using input attributes (independent variables) to predict output or dependent variables (class label)

- Unsupervised Learning
  - To discover natural grouping or concept structures in data
  - Without a dependent variable to guide learning process
  - Rather, it builds knowledge structures by using some measures of cluster quality to group instances into classes

Data mining techniques

- Supervised Learning (มีการระบุค่าการทำนาย เช่น ยอดขาย (บาท) ,ซื้อ ไม่ซื้อ)
  - Classification: response is categorical
  - Regression: response is continuous
  - Time Series: dealing with observations across time
  - Optimization: minimize or maximize some characteristic

- Unsupervised Learning (ไม่มีการระบุค่าของผลลัพธ์ เช่น ลักษณะของดอกไม้)
  - Principal Component Analysis : feature reduction
  - Clustering: grouping like objects together
  - Association and link Analysis: descriptive approach to exploring data that helps identify relationships among values in a database, e.g. Market basket analysis, those customers that buy hammers also buy nails. Examine conditional probabilities.

A machine learning perspective

- Artificial Intelligence to model the world
- Unsupervised Learning
  - Cluster Analysis
  - Association Analysis
- Supervised Learning
  - Decision Trees
  - Random forest, Ensembles
  - Support Vector Machines
  - Neural Networks

Data mining Applications (1 of 3)

- Database analysis and decision support
  - Market analysis and management
    - Target marketing
      - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
    - Market basket analysis
      - Find the relationship between product sales
    - Customer profiling
      - Data mining can tell you what types of customers buy what products (clustering or classification)
    - Identifying customer requirements
      - Identifying the best products for different customers
      - Use prediction to find what factors will attract new customers

Data mining Applications (2 of 3)

- Database analysis and decision support
  - Risk analysis and management
    - Resource planning
    - Competition
      - Monitor competitors and market directions
      - Set pricing strategy in a highly competitive market
    - Finance planning and asset evaluation
      - Time series analysis
  - Fraud detection and management
    - Auto insurance
    - Money laundering
    - Medical insurance
    - Credit card fraud
    - Detecting telephone fraud

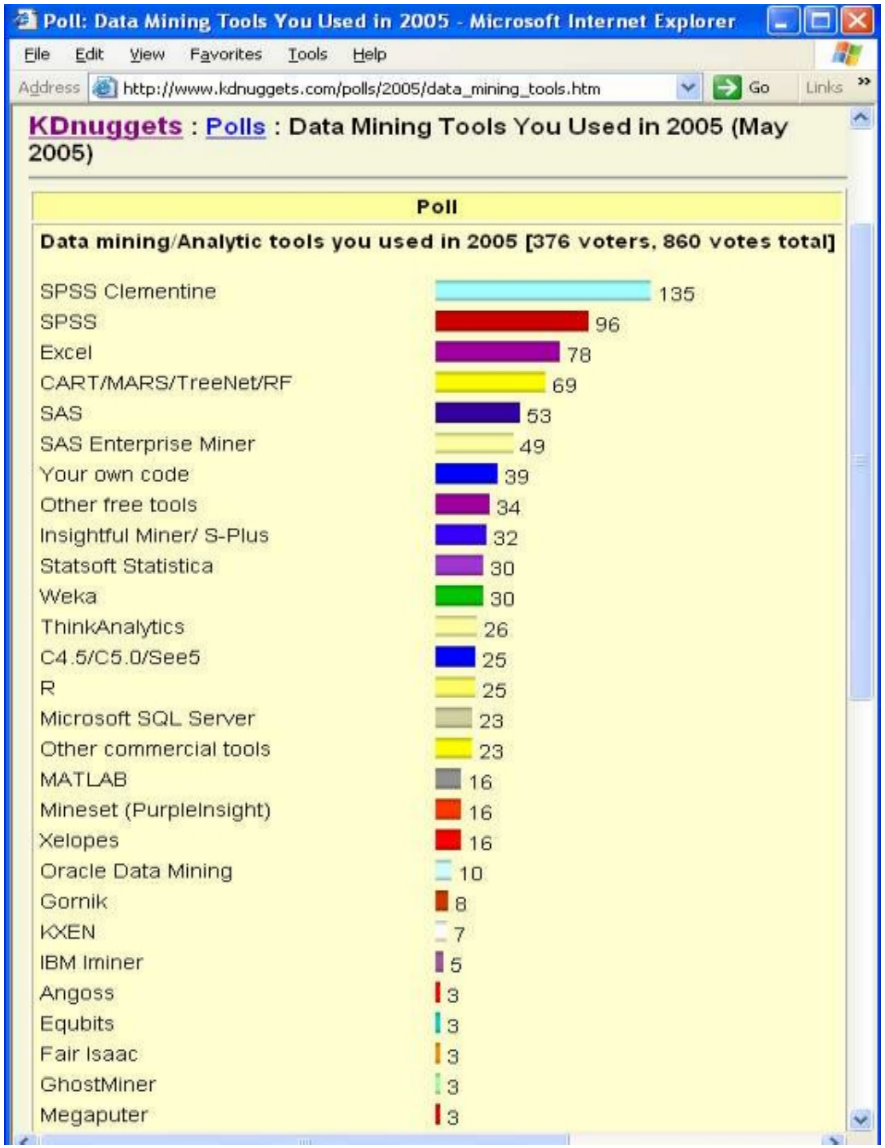
Data mining Applications (3 of 3)

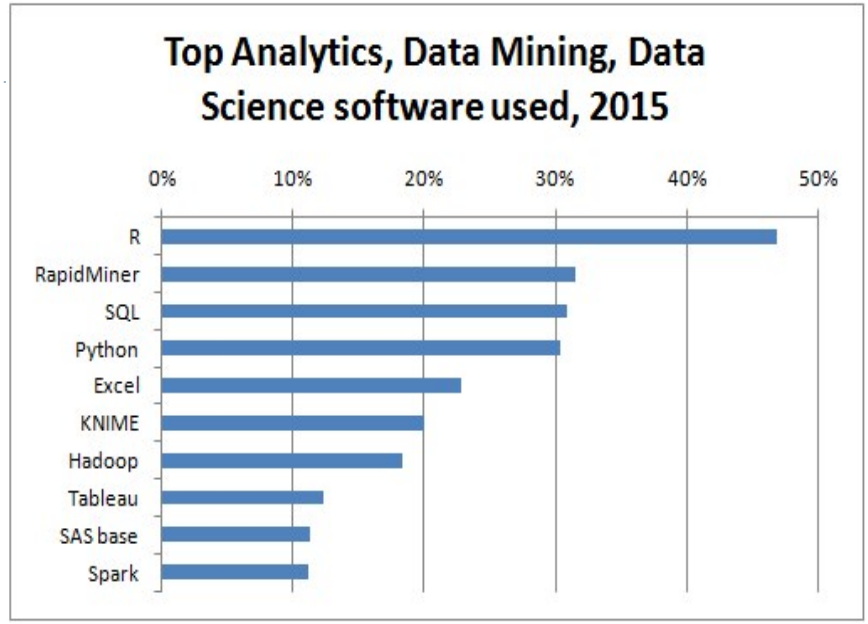
- Other Applications
  - Text mining (news group, email, documents) and Web analysis.
  - Medicine: disease outcome, effectiveness of treatments
    - Analyze patient disease history: find relationship between diseases
  - Sports
    - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
  - Astronomy
    - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
  - Web site/store design and promotion
    - Find affinity of visitor to pages and modify layout
  - Molecular/Pharmaceutical
    - Identify new drugs

Data mining/analytic tools.

The little known Open Source.

Data mining/analytic tools at 2015





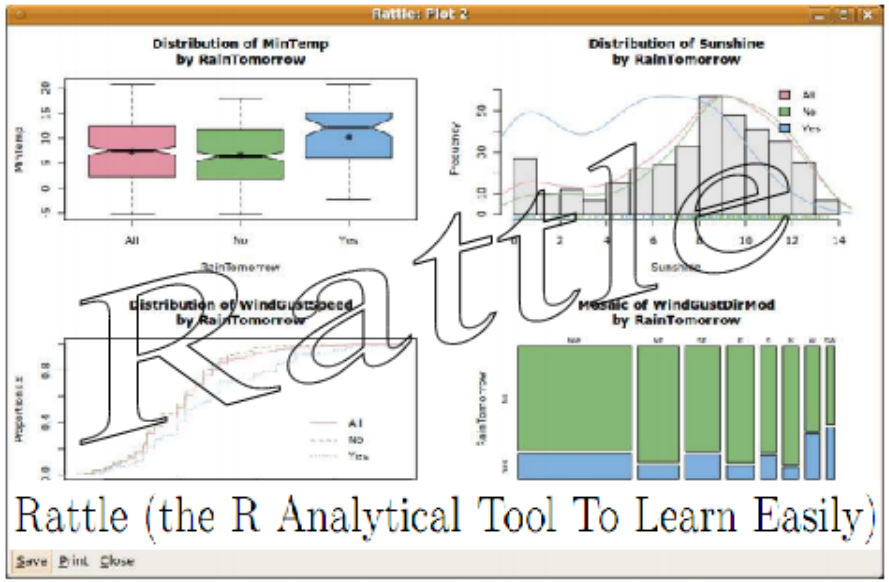
The top 10 tools by share of users were

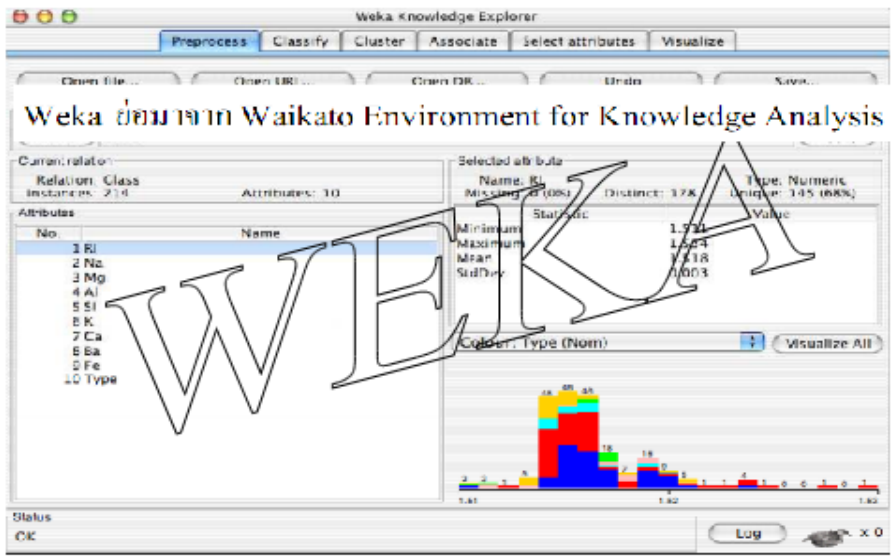
1. **R**, 46.9% share (38.5% in 2014)
2. **RapidMiner**, 31.5% (44.2% in 2014)
3. **SQL**, 30.9% (25.3% in 2014)
4. **Python**, 30.3% (19.5% in 2014)
5. **Excel**, 22.9% (25.8% in 2014)
6. **KNIME**, 20.0% (15.0% in 2014)
7. **Hadoop**, 18.4% (12.7% in 2014)
8. **Tableau**, 12.4% (9.1% in 2014)
9. **SAS**, 11.3 (10.9% in 2014)
10. **Spark**, 11.3% (2.6% in 2014)

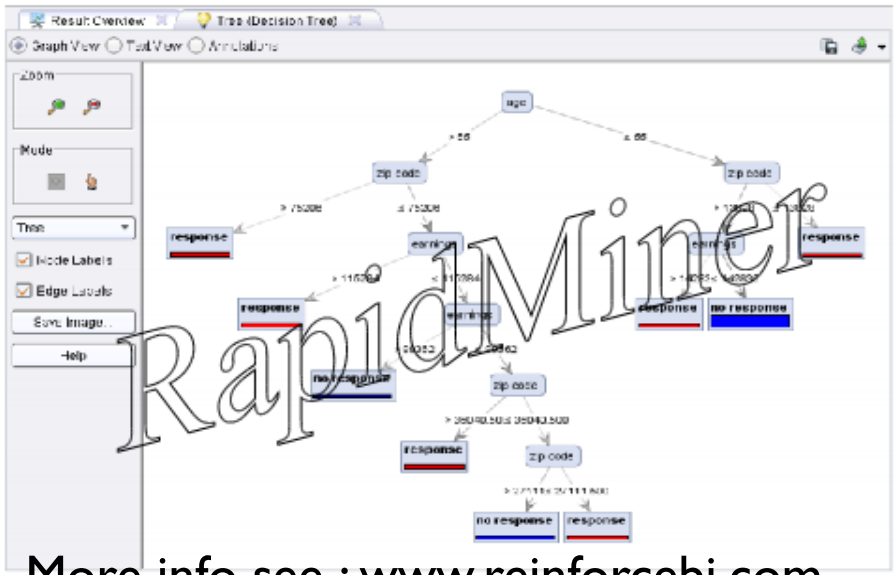
http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

Data mining/analytic tools. (Open source)

Reinforce BUSINESS INTELLIGENCE™
www.reinforcebi.com







More info see : www.reinforcebi.com

Why you learn Data Mining?.

Data Mining is now one of the most important core technologies being developed today, along with medical research, energy research, and environmental research.

(source: togaware)

2.Introduction to R ,Rattle program.

RGui

File Edit View Misc Packages Windows Help

R Console

```
R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-min

R is free software an
You are welcome to re
Type 'license()' or '

R is a collaborative
Type 'contributors()'
'citation()' on how t

Type 'demo()' for som
'help.start()' for an
Type 'q()' to quit R.

> library(rattle)
Rattle: A free graphi
Version 2.6.15 Copyri
Type 'rattle()' to sh
> rattle()
Rattle timestamp: 201
>
```

R Data Miner - [Rattle]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: (ไม่มี) 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 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 2.6.15. Copyright 2006-2011 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
```

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

2.Introduction to R ,Rattle program.

- **R** is the most sophisticated statistical software available, easily installed ,instructional, state-of-the-art, and it is free and open source.

- **Rattle** is a graphical data mining application built upon the statistical language R.

- **Rattle** is simple to use, quick to deploy, and allows us to rapidly work through the data processing , modeling ,and evaluation phases of a data mining project.

2.Introduction to R ,Rattle program.

Technology.

- Originally tooled up with commercial, expensive, data mining tools (SAS/EM, Teradata Warehouse Miner) and hardware (Big Iron MS/Windows 32 bit).
- But data mining needs skilled people, not off the shelf solutions (yet).
- Also data mining technology is rapidly developing, and commercial vendors have difficulty keeping up.

2.Introduction to R ,Rattle program.

Technology.

- Originally tooled up with commercial, expensive, data mining tools (SAS/EM, Teradata Warehouse Miner) and hardware (Big Iron MS/Windows 32 bit).
- But data mining needs skilled people, not off the shelf solutions (yet).
- Also data mining technology is rapidly developing, and commercial vendors have difficulty keeping up.

2.Introduction to R ,Rattle program.

New Approaches Ensembles.

Commercial software is lagging behind advances in Data Mining

- Current best off the shelf technology includes random forests, boosting and support vector machines - SAS/EM?
- Open source solutions allow investment in people, not software.

2.Introduction to R ,Rattle program.

Build a network of DataMining Nodes:

- 1 CPU (2 Cores), AMD64, 16GB RAM, 300GB Disk
- 4 CPU (8 Cores), AMD64, 32GB RAM, 1TB Disk (Optimal)
- 8 CPU (16 Cores), AMD64, 128GB RAM, 10TB Disk (Near Term)

- Best of class open source operating system (Debian GNU/Linux)
- Open Source data mining tools R, Rattle, Weka, AlphaMiner
- Open Source does deliver quality software

2.Introduction to R ,Rattle program.

R Data Miner - [Rattle]

Project Tools Settings Help — Rattle Version 2.6.8 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

DSN: Table: No tables found. Row Limit: 0 Believe Num Rows

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 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 2.6.8. Copyright 2006-2011 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
```

Rdata Miner [Rattle]

2.Introduction to R ,Rattle program.

A self contained tool for data mining

1. Start Rattle
2. Explore the interface
3. Load sample audit dataset
4. Explore the data: Summary, Plots, GGobi, Correlations
5. Transform the data: Rescale, Impute, Remap
6. Cluster, Associate
7. Predictive Model
8. Evaluate and Score
9. Log

Rdata Miner [Rattle]

2.Introduction to R ,Rattle program.

- Togaware
  http://datamining.togaware.com

Other Data Mining Open Source.

- Tools:
  - rattle.togaware.com
  - www.cs.waikato.ac.nz/ml/weka/
  - www.knime.org
  - rapid-i.com

Reinforce BI.
เปิดอบรมทุกโปรแกรม
08 1491 0909

Rattle Training

- 1. Getting Started with Rattle
- 2. Working with Data
- 3. Loading Data
- 4. Exploring Data
  - 4.1 Summarizing Data
  - 4.2 Visualizing Distributions
  - 4.3 Correlation Analysis
- 5. Testing Data
- 6. Transforming Data
- 7. Cluster Analysis
- 8. Association Analysis
- 9. Classifier Models
- 10. Model Performance Evaluation.

Rattle Training 1. Getting Started with Rattle

R Console

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Found
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386

R is free software and comes with
You are welcome to redistribute
Type 'license()' or 'licence()'

R is a collaborative project wit
Type 'contributors()' for more i
'citation()' on how to cite R or

Type 'demo()' for some demos, 'h
'help.start()' for an HTML brows
Type 'q()' to quit R.

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

R-2.13.1-win	24/7/2554 10:47	Application
RStudio-0.94.92	13/8/2554 12:42	Application

ดับเบิ้ลคลิ๊กเพื่อติดตั้งโปรแกรม R

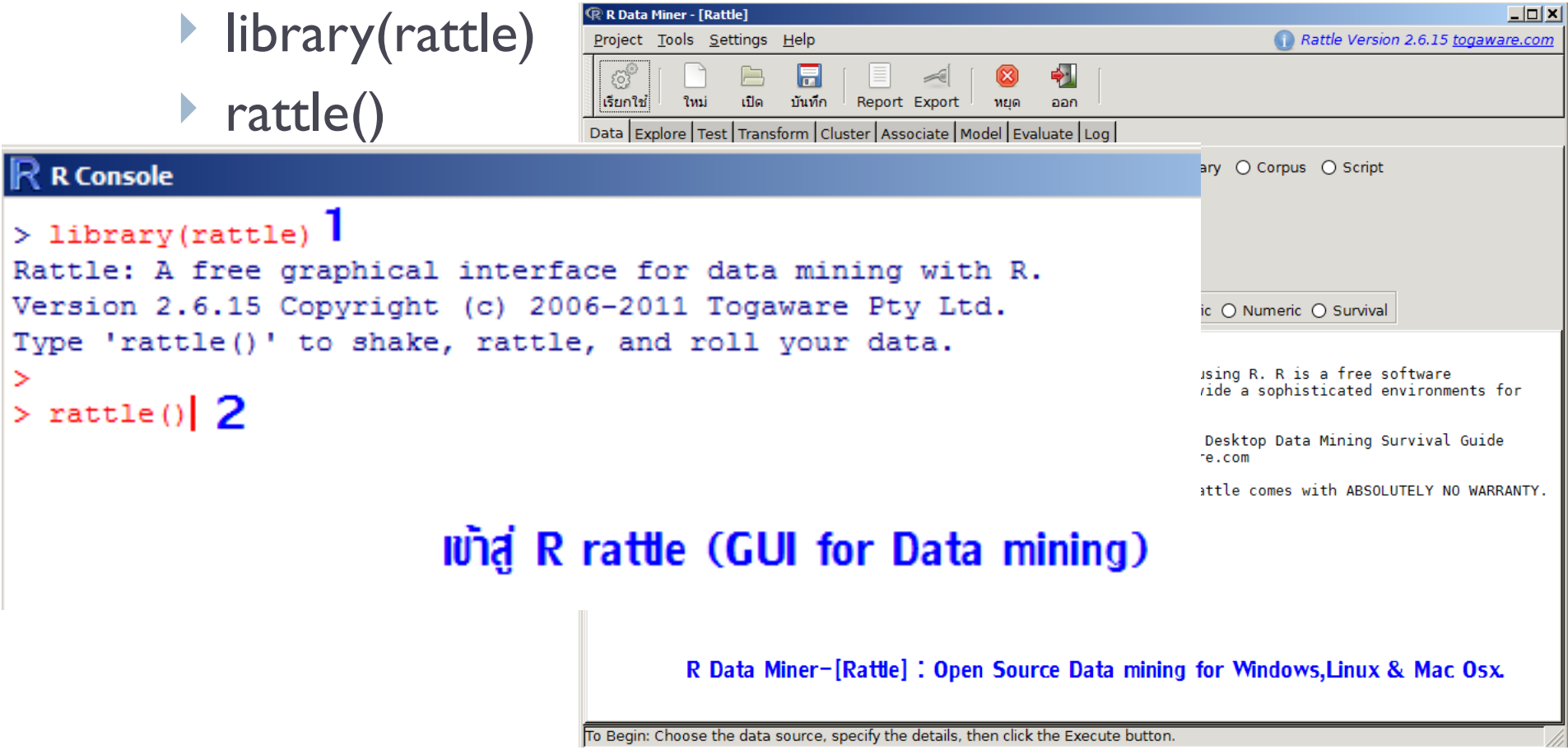
การติดตั้ง rattle นั้นต้องพึ่งพา หรือขึ้นกับ package อื่นๆ อีกหลายๆ package เพื่อให้ชุดโปรแกรม Rattle มีความสมบูรณ์จึงควรลง package ที่เกี่ยวข้องทั้งหมดเลยทีเดียว ด้วย

> install.packages ("rattle", dependencies = TRUE)

ติดตั้ง package "rattle"

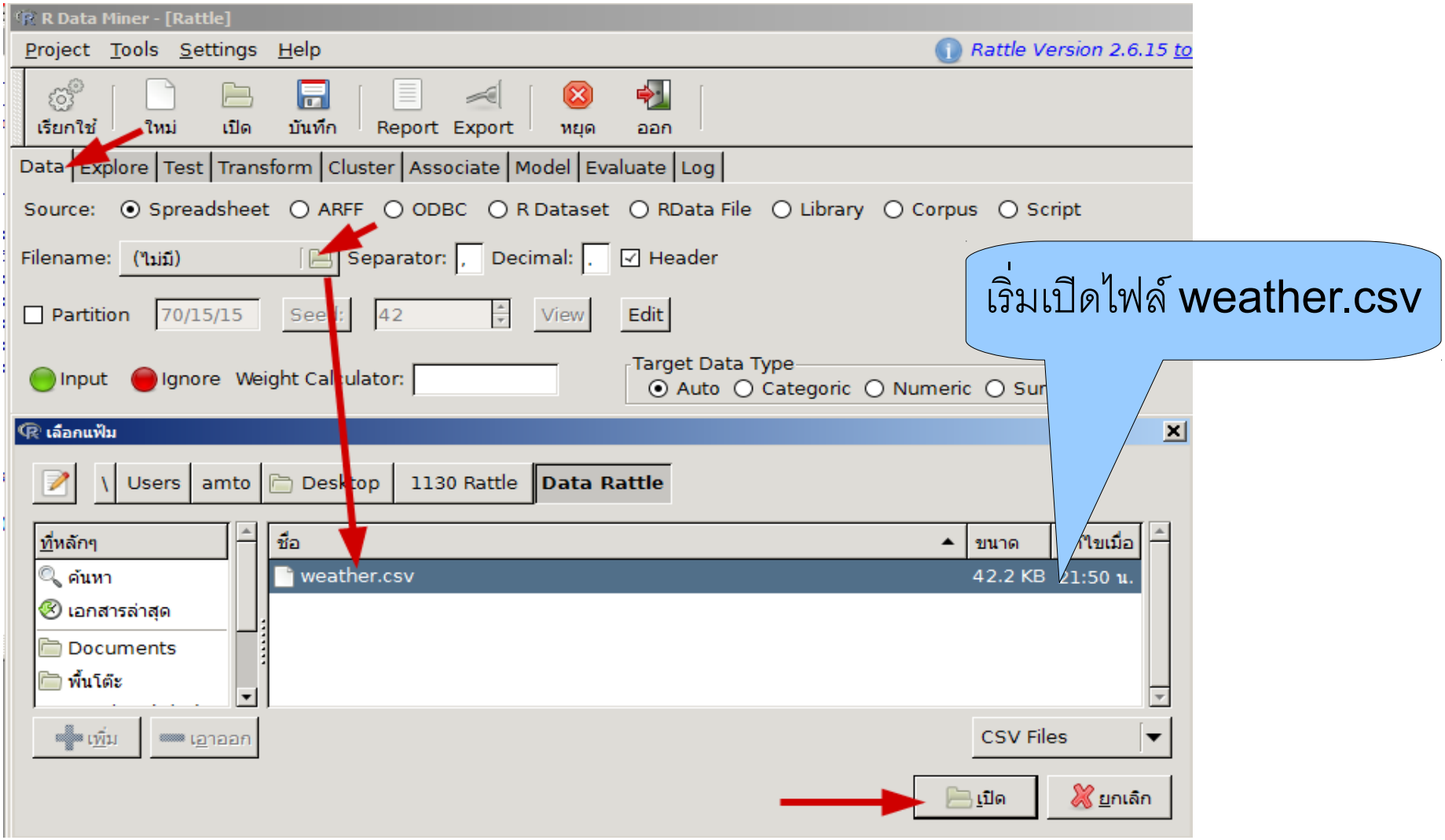
Rattle Training 1. Getting Started with Rattle

- เมื่อติดตั้ง Packages ต่างๆ เรียบร้อยแล้วก็เข้าสู่ Rattle ด้วย
- library(rattle)
- rattle()

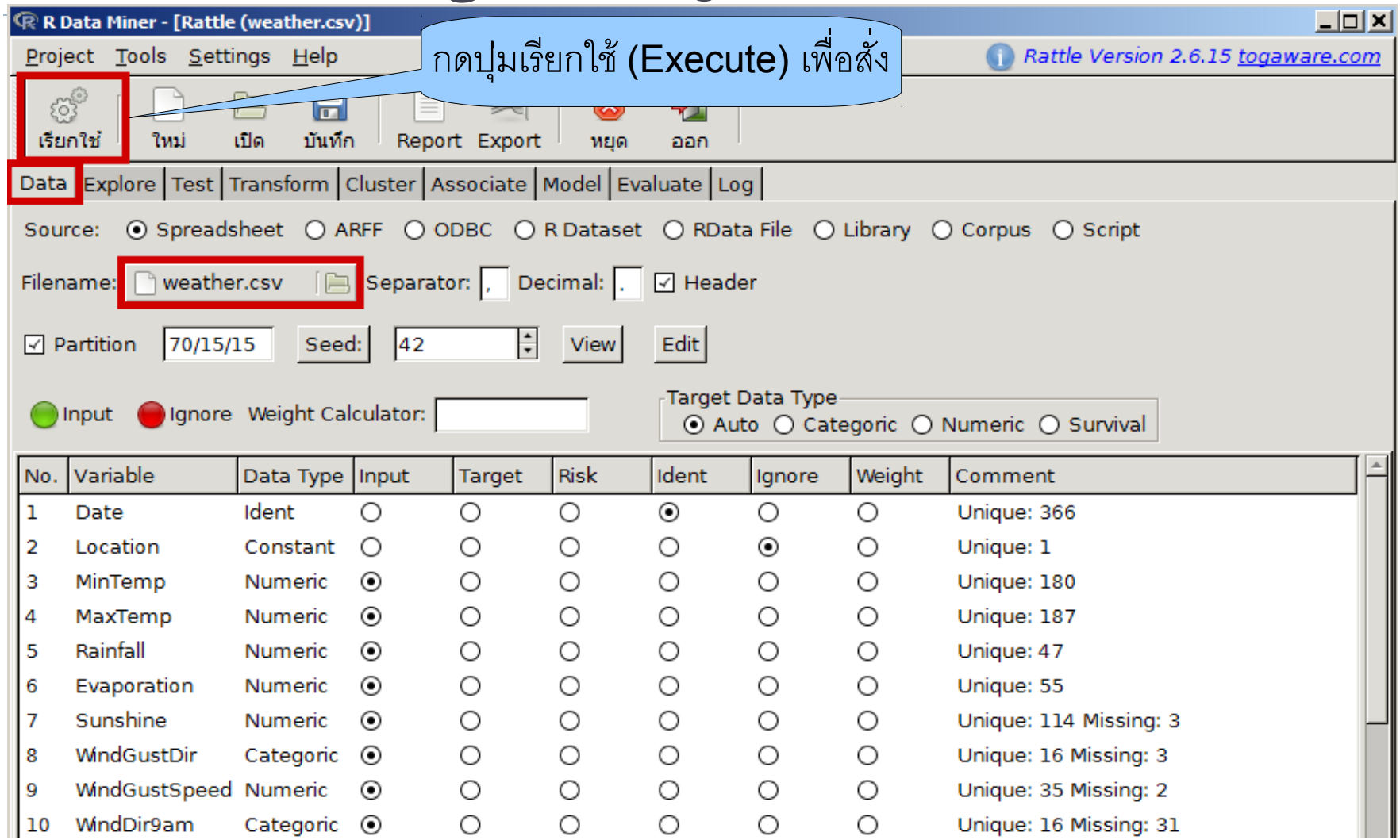


เข้าสู่ R rattle (GUI for Data mining)

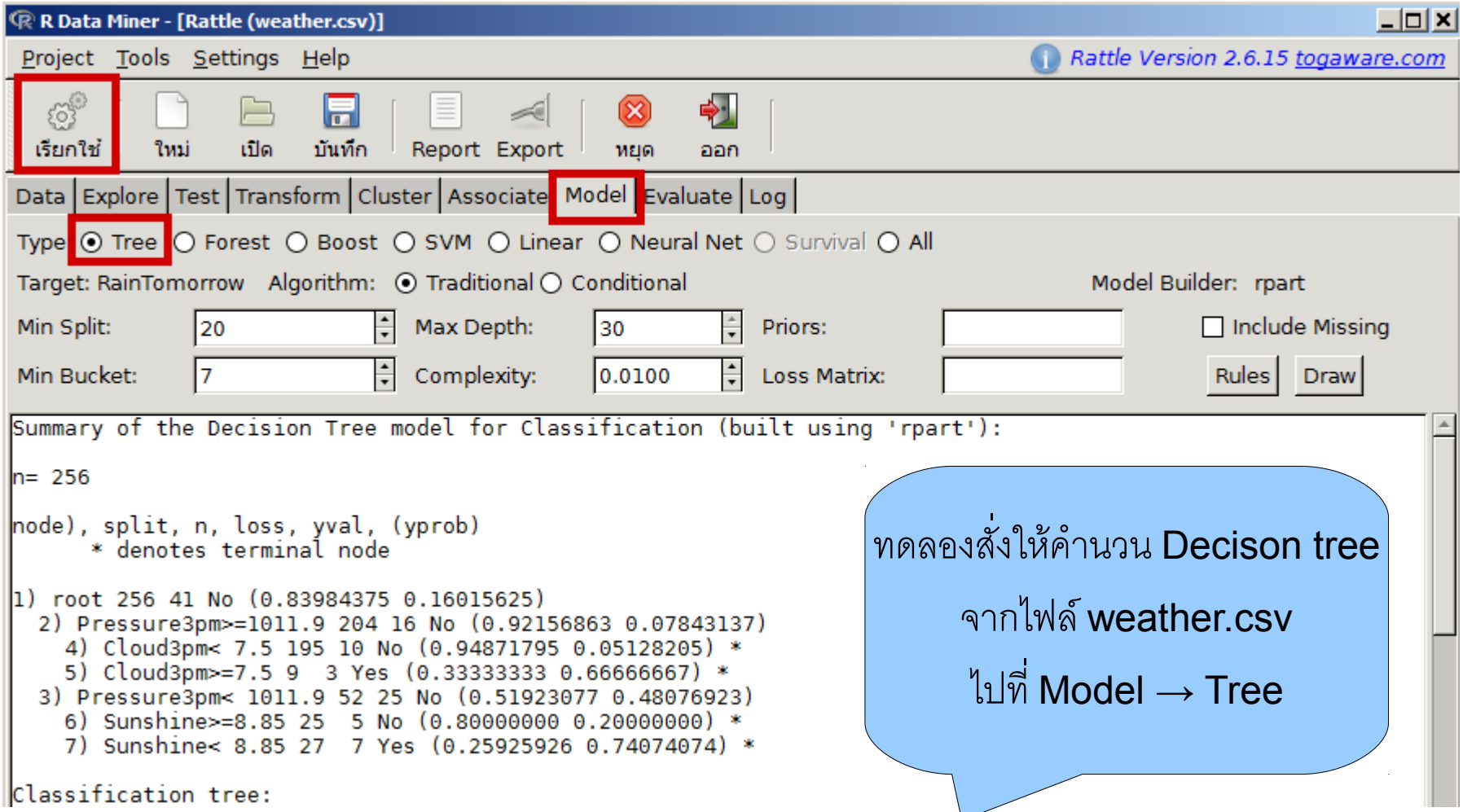
Rattle Training 1. Getting Started with Rattle



Rattle Training 1. Getting Started with Rattle



Rattle Training 1. Getting Started with Rattle



ทดลองสั่งให้คำนวน Decison tree
จากไฟล์ weather.csv
ไปที่ Model → Tree

Rattle Training 1. Getting Started with Rattle

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

Type: (•) Tree ( ) Forest ( ) Boost ( ) SVM ( ) Linear ( ) Neural Net ( ) Survival ( ) All

Target: RainTomorrow Algorithm: (•) Traditional ( ) Conditional Model Builder: rpart

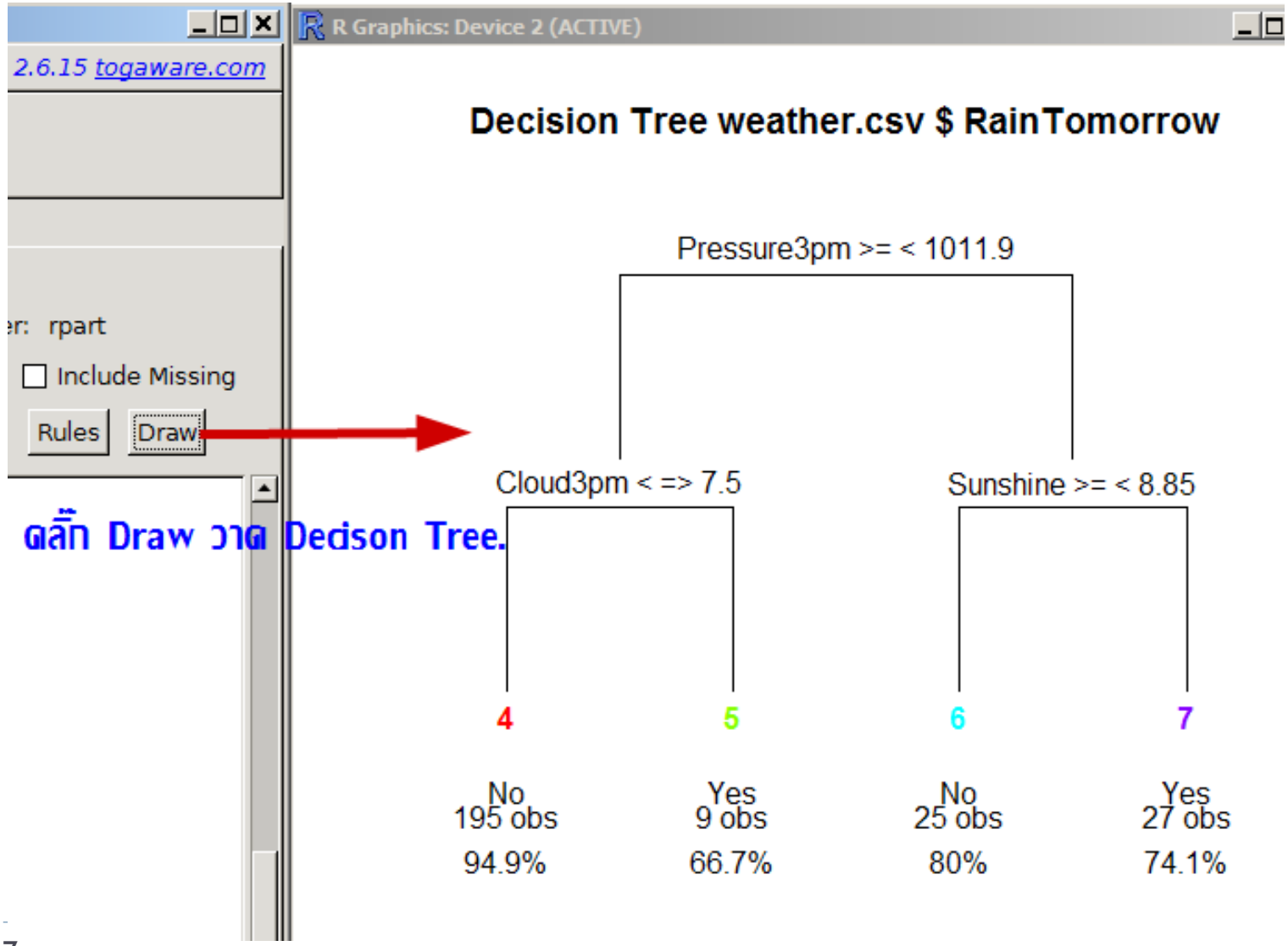
Min Split: 20 Max Depth: 30 Priors: [ ] Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix: [ ] Rules Draw

คลิ๊กแสดง Rules

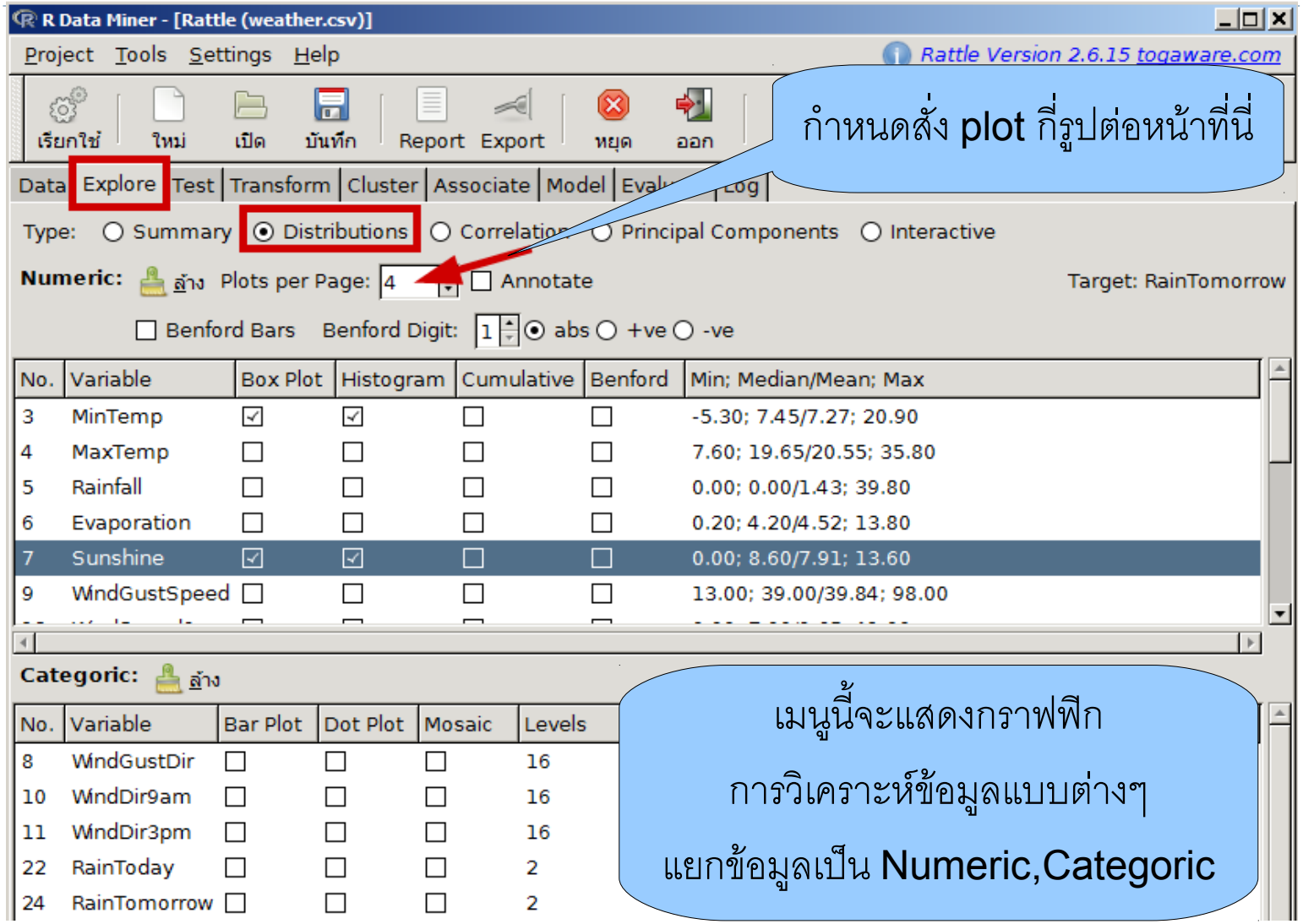
```
==========================================================================
Tree as rules:

 Rule number: 7 [RainTomorrow=Yes cover=27 (11%) prob=0.74]
   Pressure3pm< 1012
   Sunshine< 8.85

 Rule number: 5 [RainTomorrow=Yes cover=9 (4%) prob=0.67]
   Pressure3pm>=1012
   Cloud3pm>=7.5

 Rule number: 6 [RainTomorrow=No cover=25 (10%) prob=0.20]
   Pressure3pm< 1012
   Sunshine>=8.85

 Rule number: 4 [RainTomorrow=No cover=195 (76%) prob=0.05]
   Pressure3pm>=1012
   Cloud3pm< 7.5

[1] 7 4 5 6 1 2 3

Rattle timestamp: 2012-01-06 22:26:33 amto
==========================================================================
```

Rattle Training 1. Getting Started with Rattle



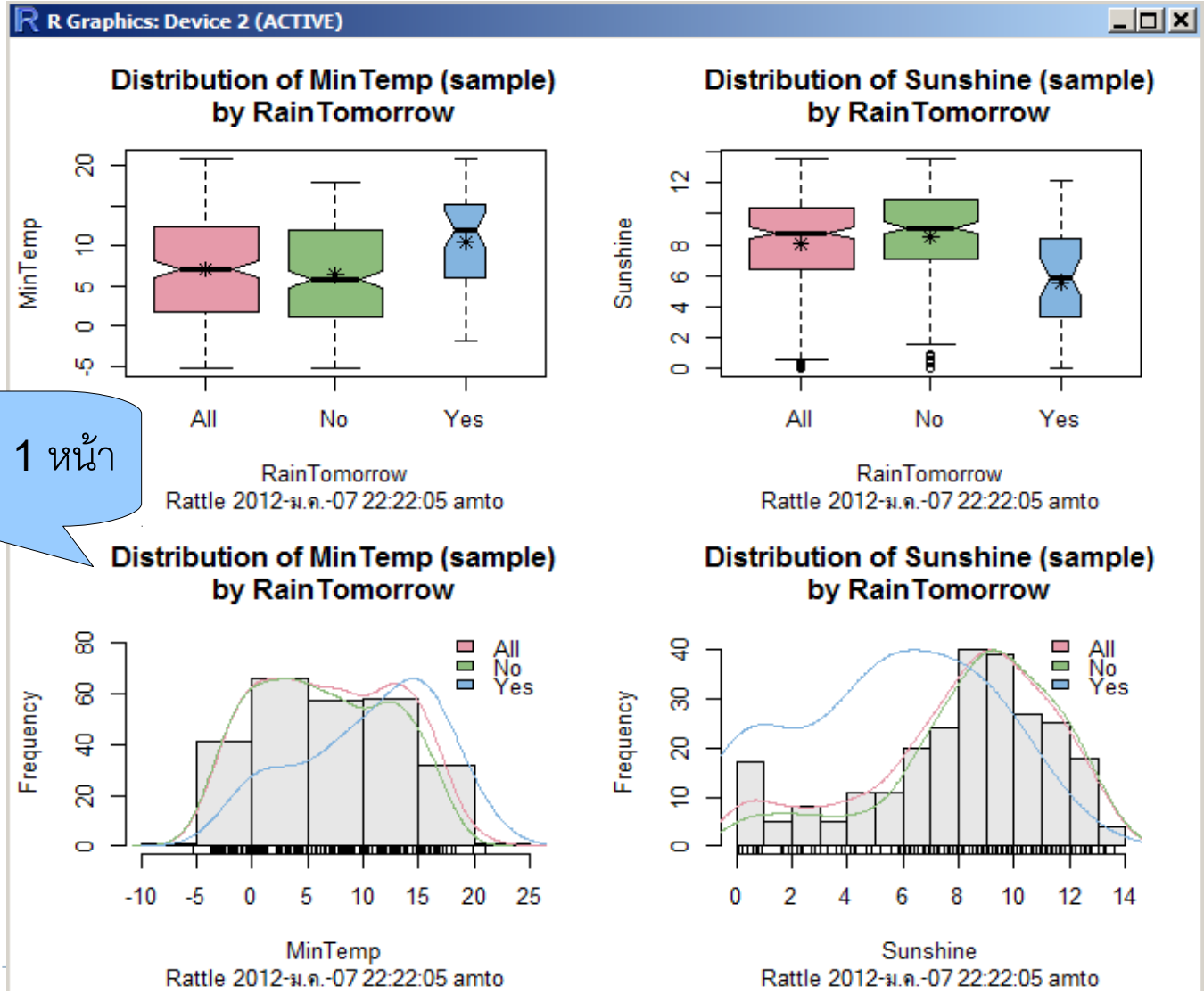
คลิ๊ก Draw วาด Decison Tree.

Rattle Training 1. Getting Started with Rattle

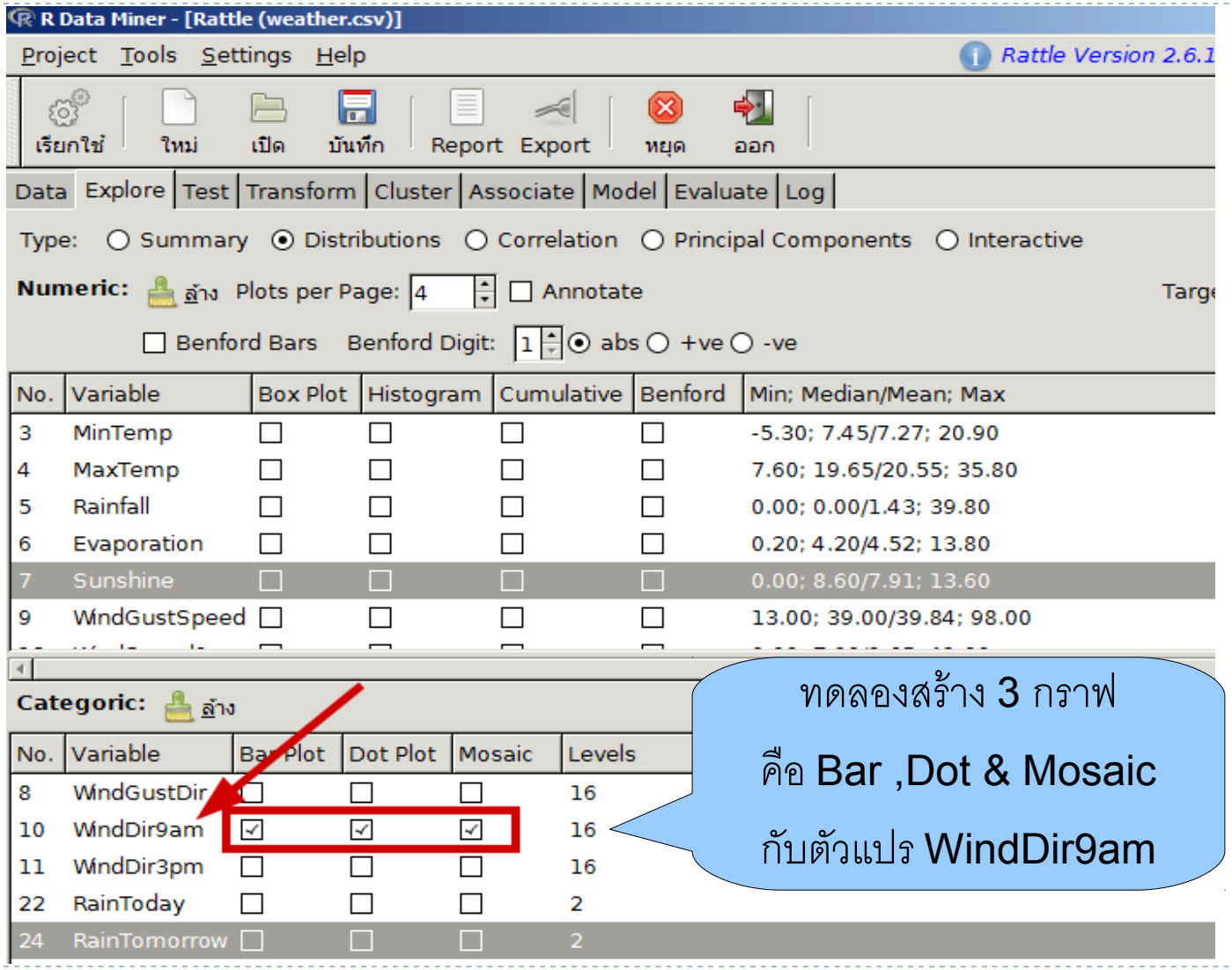


Rattle Training 1. Getting Started with Rattle

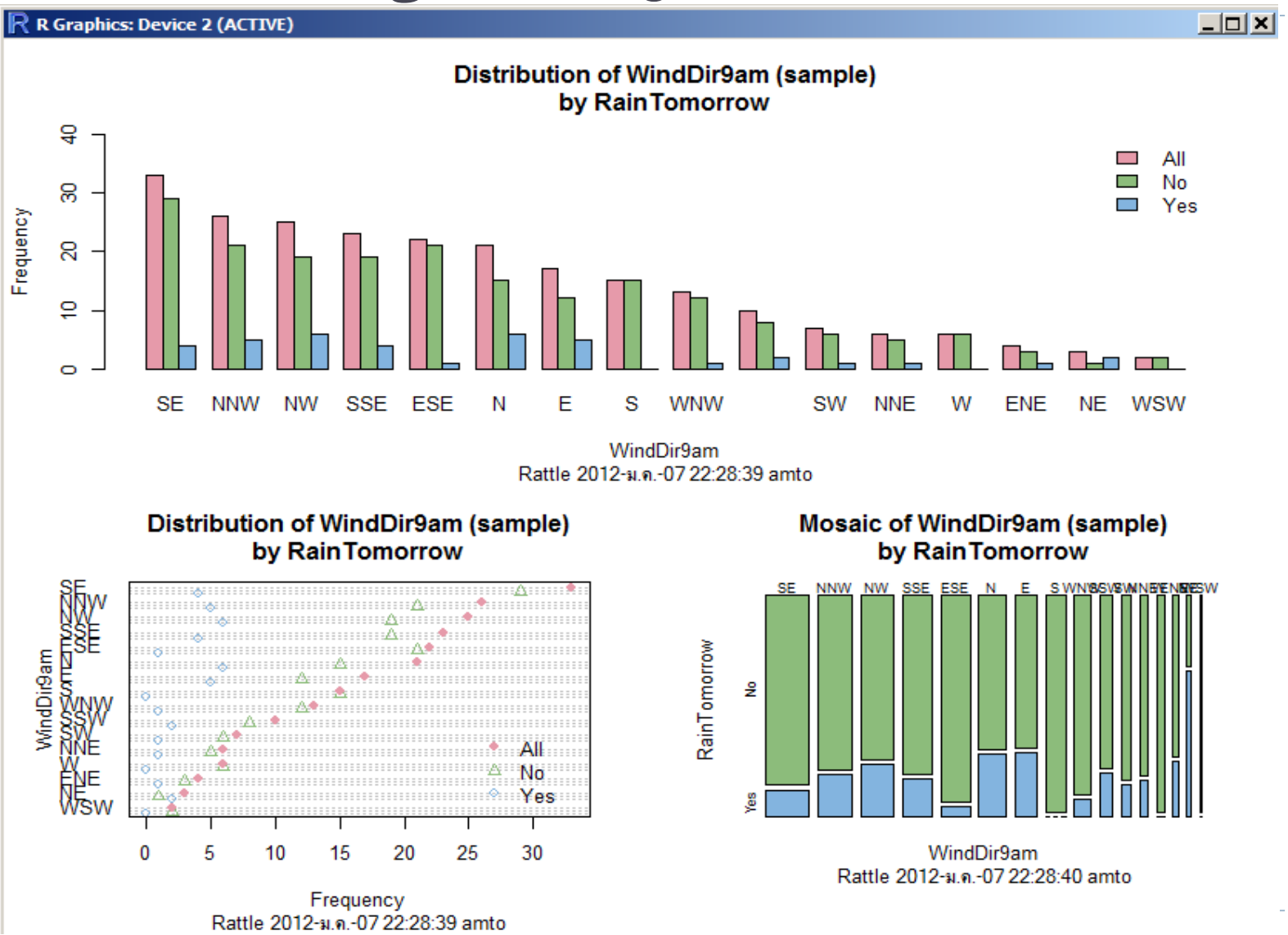
แสดงกราฟ 4 รูปต่อ 1 หน้า



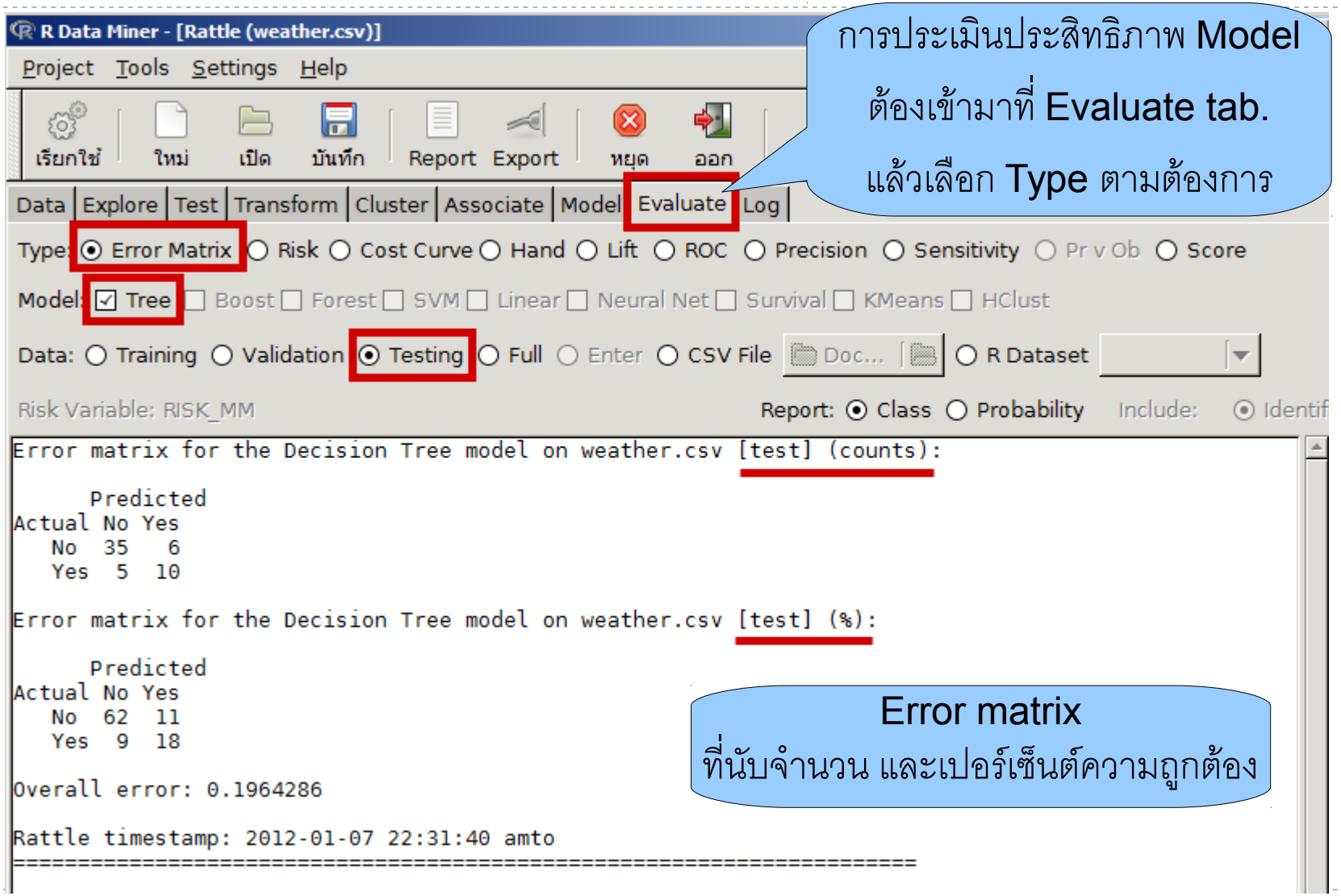
Rattle Training 1. Getting Started with Rattle



Rattle Training 1. Getting Started with Rattle



Rattle Training 1. Getting Started with Rattle



การประเมินประสิทธิภาพ Model ต้องเข้ามาที่ Evaluate tab. แล้วเลือก Type ตามต้องการ

Error matrix ที่นับจำนวน และเปอร์เซ็นต์ความถูกต้อง

Rattle Training 1. Getting Started with Rattle

Data: Training Validation Testing Full Enter CSV File Doc... R Dataset

Risk Variable: RISK_MM Report: Class Probability Include:

```
Error matrix for the Decision Tree model on weather.csv [**train**] (counts):

      Predicted
Actual  No Yes
   No  205  10
   Yes  15  26

Error matrix for the Decision Tree model on weather.csv

      Predicted
Actual No Yes
   No  80   4
   Yes  6  10

Overall error: 0.09765625

Rattle timestamp: 2012-01-07 22:34:28 amto
==========================================================================
```

เราสามารถเลือก Data ที่จะมาประเมิน ตามกำหนด เช่น 70/15/15 Training,Validation,Testing,Full หรือ Supply Test set ตามกำหนด

Rattle Training 1. Getting Started with Rattle

Data: ○ Training ◉ Validation ○ Testing ○ Full ○ Enter ○ CSV File Doc... ○ R Dataset

Risk Variable: RISK_MM Report: ◉ Class ○ Probability Include:

```
Error matrix for the Decision Tree model on weather.csv [validate] (counts):

      Predicted
Actual No Yes
   No  39   5
   Yes  5   5

Error matrix for the Decision Tree model on weather.csv [validate] (%):

      Predicted
Actual No Yes
   No  72   9
   Yes  9   9

Overall error: 0.1851852

Rattle timestamp: 2012-01-07 22:35:22 amto
======================================================================
```

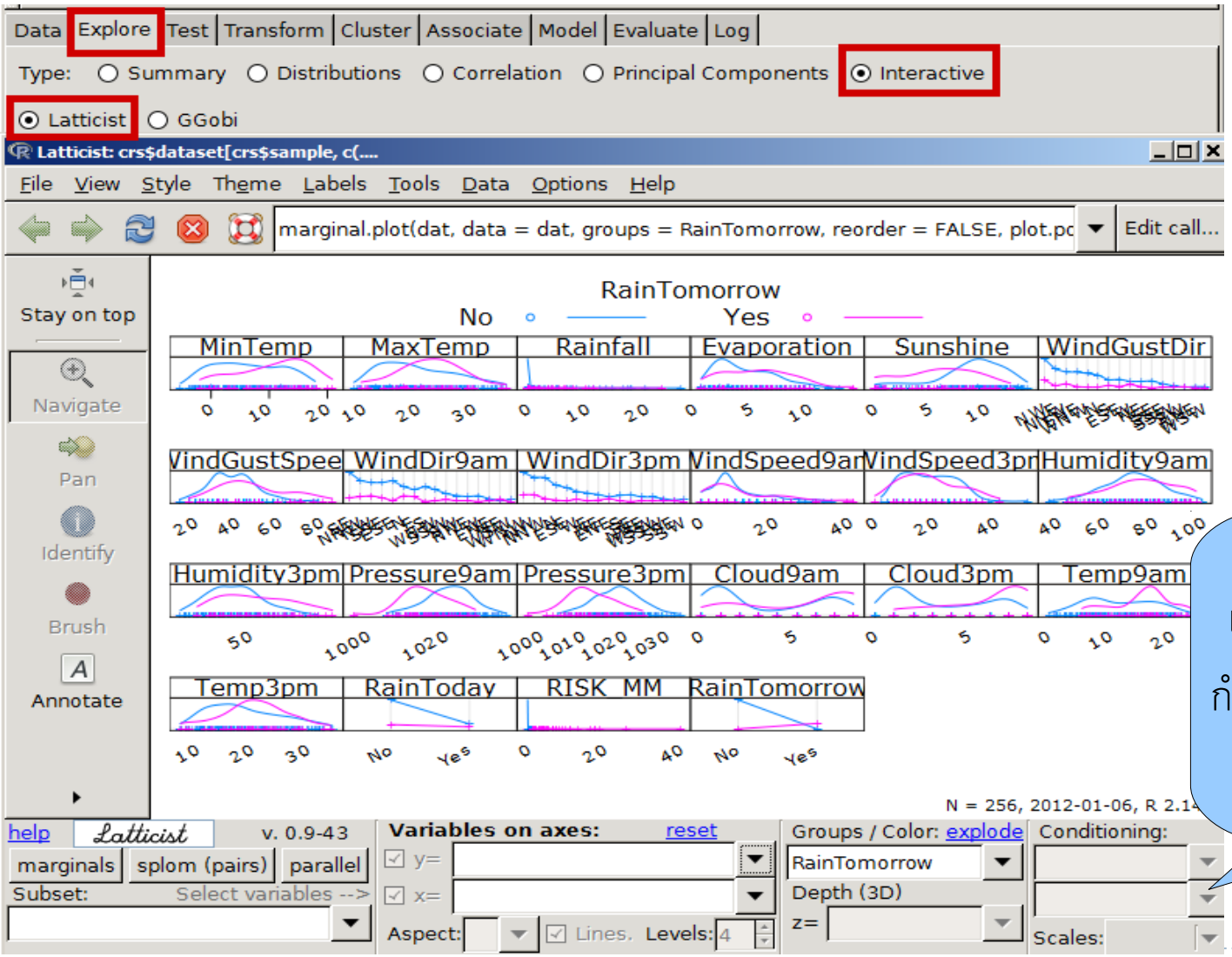
Rattle Training 1. Getting Started with Rattle

Data: ○ Training ○ Validation ○ Testing ◉ Full ○ Enter ○ CSV File Doc... ○ R Dataset

Risk Variable: RISK_MM Report: ◉ Class ○ Probability Include:

```
Error matrix for the Decision Tree model on weather.csv (counts):

      Predicted
Actual  No Yes
   No  279  21
   Yes  25  41

Error matrix for the Decision Tree model on weather.csv (%):

      Predicted
Actual No Yes
   No  76   6
   Yes  7  11

Overall error: 0.1256831

Rattle timestamp: 2012-01-07 22:36:54 amto
=====================================================================
```

Rattle Training 1. Getting Started with Rattle



Interactive graphic เป็นรูปแบบที่เราสามารถกำหนดตัวแปร และเงื่อนไขตัวแปรเชิงกลุ่มเข้าไปได้

Rattle Training 2. Working with Data

Variables

Observations

Missing

	Numeric	Categoric	Numeric	Categoric
Date	Temp	WindDir9am	Pressure3pm	Rain Tomorrow
2007-11-01	8	SW	1015	Yes
2007-11-02	14	E	1008.4	Yes
2007-11-03	13.7	N	1007.2	Yes
2007-11-04	13.3	WNW	1007	Yes
2007-11-05	7.6	SSE	1018.5	No
2007-11-06	6.2	SE	1021.7	No
2007-11-07	6.1	SE	1022.2	No
2007-11-08	8.3		1024.2	No
2007-11-09	9.8	E	1022.7	Yes
2007-11-10	8.4	S	1020.7	No
2007-11-11	9.1	SE	1021.1	No
2007-11-12	8.5	E	1019.9	No
2007-11-13	10.1	S	1017.1	No
2007-11-14	12.1	WNW	1013.1	No
2007-11-15	10.1	S	1013.7	No
2007-11-16	12.4	SE	1012.8	No

Identifier

Output

Input

Rattle Training 2. Working with Data

- R Command-line คือพื้นฐานสำคัญที่ควรจะต้องรู้ควบคู่ไปกับ GUI.

R Console

```
> head(names(weather))
[1] "Date"        "Location"    "MinTemp"     "MaxTemp"     "Rainfall"
[6] "Evaporation"
> names(weather)[1:6]
[1] "Date"        "Location"    "MinTemp"     "MaxTemp"     "Rainfall"
[6] "Evaporation"
> head(weather[2])
  Location
1 Canberra
2 Canberra
3 Canberra
4 Canberra
5 Canberra
6 Canberra
> head(weather$MinTemp)
[1]  8.0 14.0 13.7 13.3  7.6  6.2
>
```

สำรวจข้อมูลเบื้องต้นด้วย R command-line

Rattle Training 3. Loading Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

เปิดไฟล์ weather.csv

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

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

Filename: weather.csv Separator: , Decimal: . Header

การเปิดไฟล์ *.txt ต้องใส่ Separator เป็น “\t”

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: Auto Categoric Numeric Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	●	○	○	Unique: 366
2	Location	Constant	○	○	○	○	●	○	Unique: 1
3	MinTemp	Numeric	●	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	●	○	○	○	○	○	Unique: 187
5	Rainfall	Numeric	●	○	○	○	○	○	Unique: 47
6	Evaporation	Numeric	●	○	○	○	○	○	Unique: 55
7	Sunshine	Numeric	●	○	○	○	○	○	Unique: 114 Missing: 3
8	WindGustDir	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 3
9	WindGustSpeed	Numeric	●	○	○	○	○	○	Unique: 35 Missing: 2
10	WindDir9am	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 31
11	WindDir3pm	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 1
12	WindSpeed9am	Numeric	●	○	○	○	○	○	Unique: 22 Missing: 7
13	WindSpeed3pm	Numeric	●	○	○	○	○	○	Unique: 26

Rattle Training 3. Loading Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: weather.csv Separator: , Decimal: . Header

เลือกแฟ้ม

\ Users amto Desktop 1130 Rattle **Data Rattle**

ที่หลักๆ: ค้นหา, เอกสารล่าสุด, Documents, พื้นโต๊ะ

ชื่อ	ขนาด	แก้ไขเมื่อ
Data rattle.xls	7.0 KB	เมื่อวาน 01:07 น.
Spradsheet.xls	17.0 KB	20/11/2553

CSV Files
TXT Files
Excel Files
Excel 2007 Files
All Files

You have chosen to load a new dataset. This will clear the current project (dataset and models). If you choose not to continue you can then save the current project before loading the new dataset.

Do you wish to continue and so overwrite the current project?

ถามว่าจะเปิดข้อมูลทับหรือไม่

ใช่ ไม่

Rattle Training 3. Loading Data

R Data Miner - [Rattle (Spradsheet.xls)]

Project Tools Settings Help — Rattle Version 2.6.15 togaw

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: Spradsheet.... Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

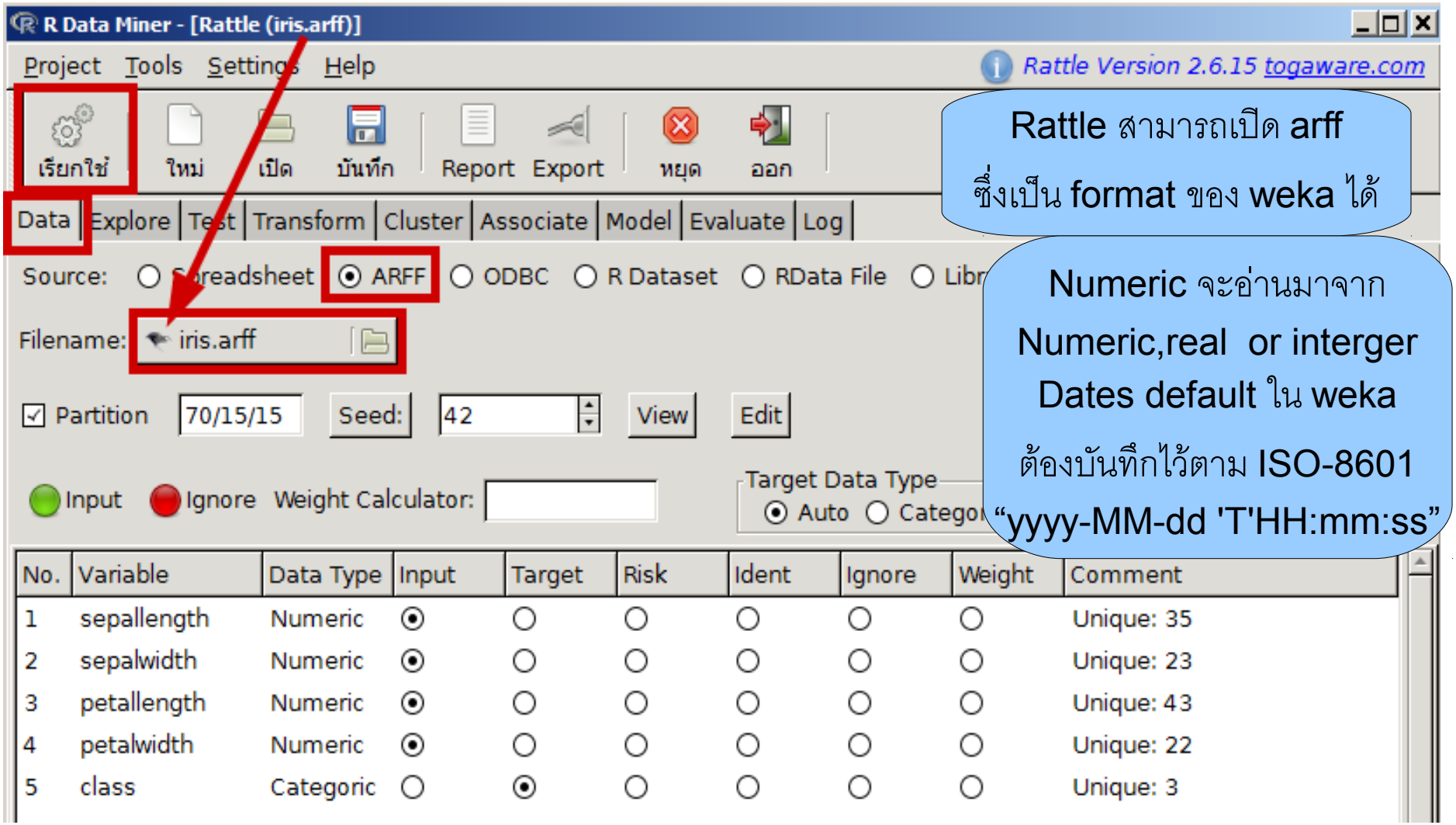
Input Ignore Weight Calculator:

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

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	DV	Numeric	◉	○	○	○	○	○	Unique: 6
2	Factor	Categoric	○	◉	○	○	○	○	Unique: 3

Rattle สามารถเปิดไฟล์ Excel ได้ทันที (ข้อมูลอยู่ใน sheet1)

Rattle Training 3. Loading Data



Rattle สามารถเปิด arff ซึ่งเป็น format ของ weka ได้

Numeric จะอ่านมาจาก Numeric,real or interger

Dates default ใน weka ต้องบันทึกไว้ตาม ISO-8601 “yyyy-MM-dd 'T'HH:mm:ss”

Rattle Training 3. Loading Data

- Databases & data warehouse. (SQL,Oracle,MySQL..)
- ต้องเปิดจาก The Open Database Connectivity (ODBC)
- พัฒนามาจาก Structured Query Language (SQL) technology
- R & Rattle use RODBC (Ripley and Lapsley,2010)
  - ตัวอย่างคำสั่งบน R
  - library(RODBC)
  - Channel ←odbcConnect("x", uid="xx",pwd="xxxx")
  - tables <- sqlColumns(channel ,"clients")
  - query <- "SELECT * FROM clients WHERE cost > 5000"
  - myds <- sqlQuery (channel, query)

Rattle Training 3. Loading Data

R Data Miner - [Rattle (iris.arff)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

DSN: weather Table: No tables found. Row Limit: 0 Believe Num Rows

Partition 70/15/15 Seed: 42 View Edit

A connection to an ODBC data source name (DSN) has not been established. Please enter the DSN and press the Enter key. This will also populate the list of tables to choose from. After establishing the connection you can choose a table or else enter a specific SQL query to retrieve a dataset.

Rattle 2.6.15

ถ้ายังไม่ได้ตั้งค่า DSN โปรแกรมจะฟ้อง

ปิด

ata Type
Categoric Numeric Survival

Ignore	Weight	Comment
		Unique: 35
		Unique: 23
		Unique: 43
		Unique: 22
		Unique: 3

Rattle Training 3. Loading Data

- การเชื่อมต่อกับ Microsoft SQL Server จะต้อง Download SQL ODBC มาติดตั้งเสียก่อน จึงจะสามารถใช้ R เชื่อมต่อได้
- เข้าไปที่ http://www.microsoft.com/download

Microsoft SQL Server 2008 Native Client

Microsoft SQL Server 2008 Native Client (SQL Native Client) is a single dynamic-link library (DLL) containing both the SQL OLE DB provider and SQL ODBC driver. It contains run-time support for applications using native-code APIs (ODBC, OLE DB and ADO) to connect to Microsoft SQL Server 2000, 2005, or 2008. SQL Native Client should be used to create new applications or enhance existing applications that need to take advantage of new SQL Server 2008 features. This redistributable installer for SQL Native Client installs the client components needed during run time to take advantage of new SQL Server 2008 features, and optionally installs the header files needed to develop an application that uses the SQL Native Client API.

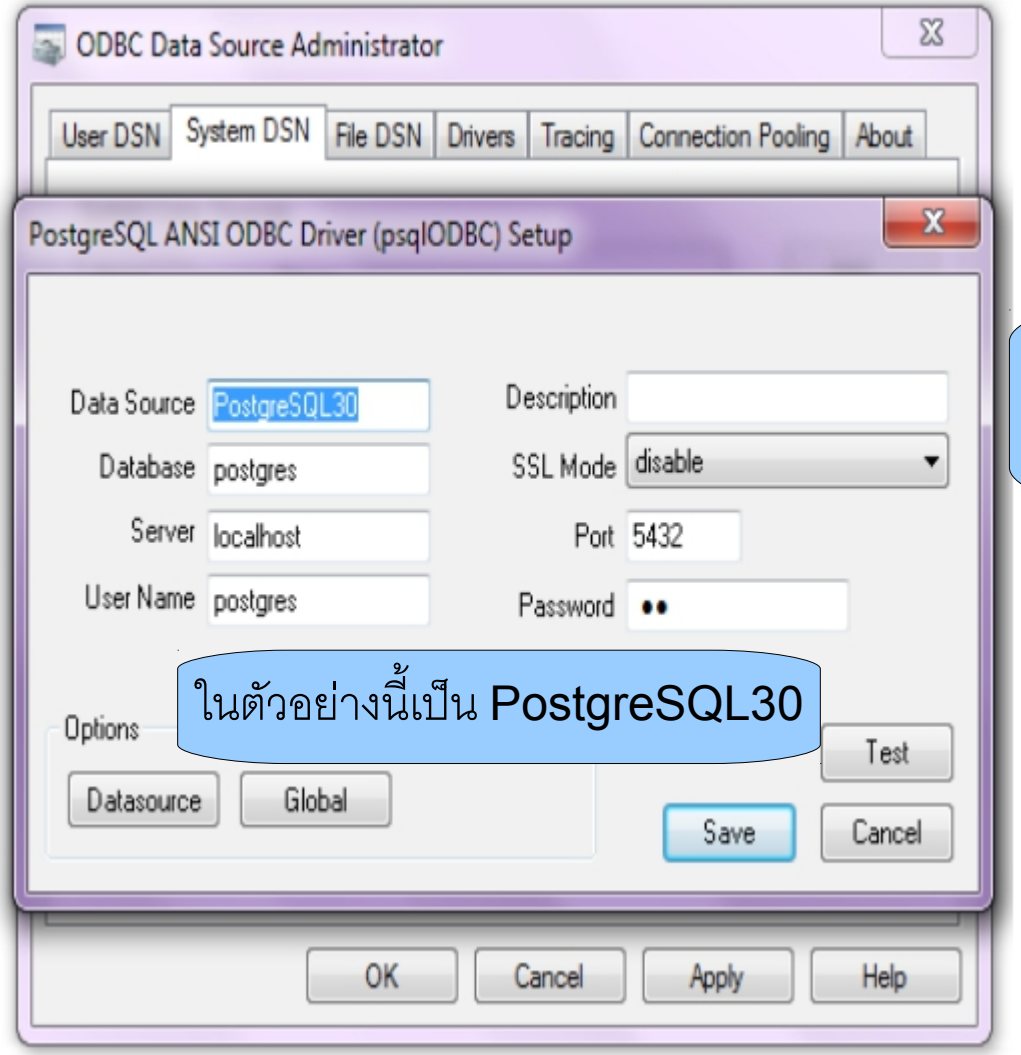
Audience(s): **Customer, Partner, Developer**

X86 Package (sqlncli.msi) - 4549 KB
X64 Package (sqlncli.msi) - 7963 KB
IA64 Package (sqlncli.msi) - 11112 KB

รายละเอียดการตั้งค่าดูหน้าถัดไป

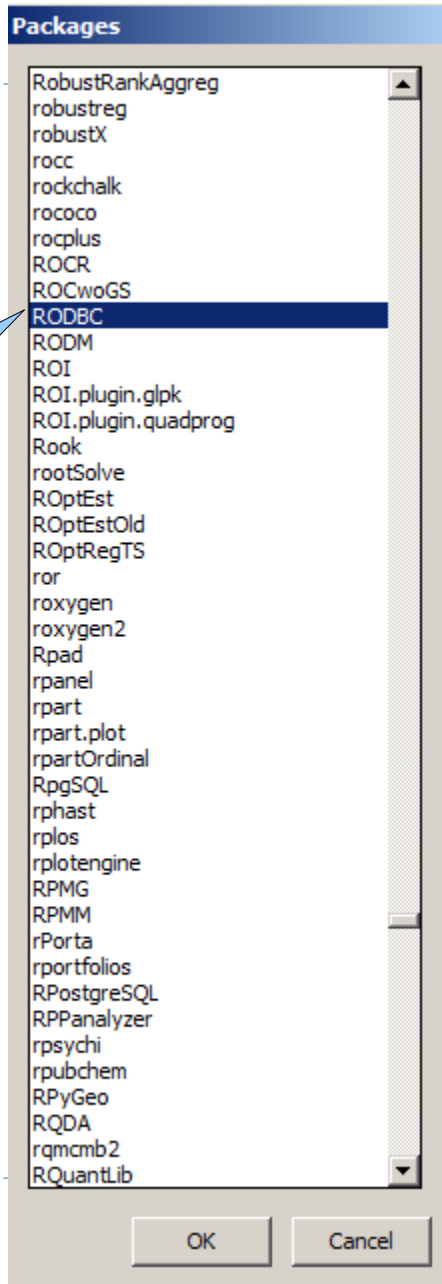
http://www.microsoft.com/download

Rattle Training 3. Loading Data



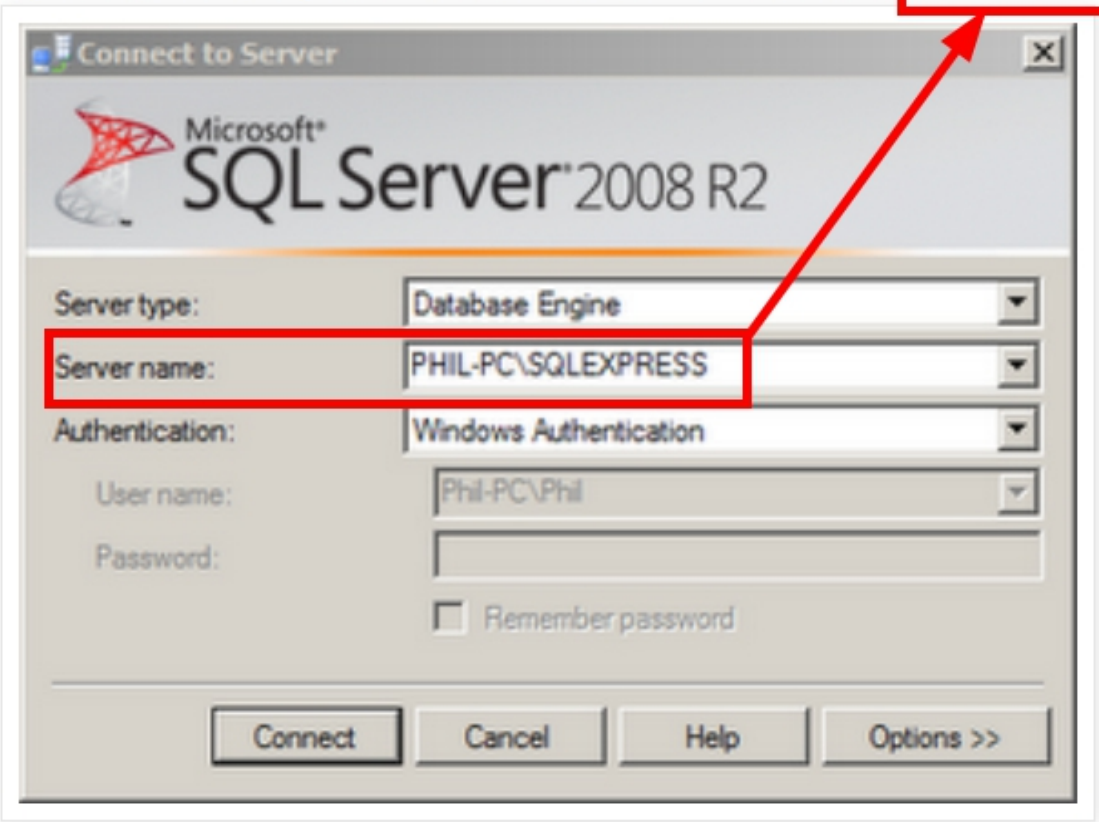
ในตัวอย่างนี้เป็น PostgreSQL30

ใน R ให้ติดตั้ง RODBC



Rattle Training 3. Loading Data

1. the driver name (**SQL Server** in our case)
2. the database server name. This can be tricky to remember. For SQL Server the easiest way to find it is when you start up the SQL Server Management Studio. In the picture below you will see the server name is PHIL-PC\SQLEXPRESS.



Rattle Training 3. Loading Data

- ระวังในเรื่อง R script เฉพาะเช่น '\' ที่เกี่ยวข้องกับ Path ที่อ้างฐานข้อมูล

There is one *gotcha* here. You will notice in the R script we have an extra '\'

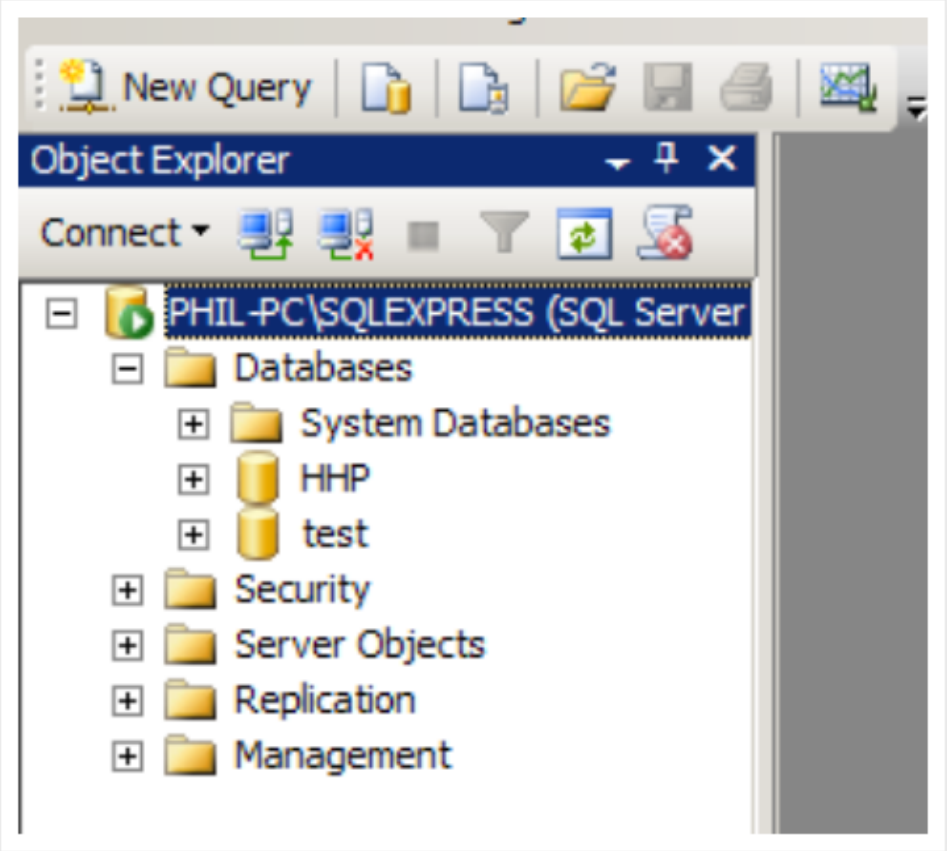
server=PHIL-PC\\SQLEXPRESS

This is because '\' is a special charater in R, so you need to put two of them together to tell it that you really do mean '\'. This is useful to know when you have paths of files etc.

Another thing to note above, we are in Windows Authentication mode, which means we won't have to worry with passwords.

Rattle Training 3. Loading Data

3. The name of the database in the database server that you want to connect to. You can discover your options from within the Management Studio. Here we have *HHP* and *test*



Rattle Training 3. Loading Data

- การเชื่อมต่ออาจจะใช้คำสั่ง odbcDriverConnect (“....”)

So when we string all this together we get,

```
conn <- odbcDriverConnect("driver=SQL Server;database=HHP;server=PHIL-PC\\SQLEXPRESS;")
```

- หรือ sqlFetch (“...”)

Once the channel is created, I can load the data from the table via the sqlFetch command in R as follows:

```
> loanstatus <- sqlFetch(ch,"tblElPasoCountyLoanStatus")
```

Rattle Training 3. Loading Data

- R connect ORACLE.
- Must use ODBC too.
- มีให้เลือก Windows,Linux

R and Oracle HR Part I - Set Up and Connect

There has been recent interest in the attention that Oracle has been giving to R. This has been related to the use of Oracle Data Mining which is a paid option to Oracle. This is the first in a series of posts that can be run using the Oracle HR Schema which is available in various versions or Oracle including its free Oracle XE version.

Rattle Training 3. Loading Data

ORACLE

(Sign In/Register for Account | Help) United States Communities I am a... I want to...

Products and Services | Solutions | Downloads | Store | Support | Training | Partners | About

Oracle Technology Network › Database › Windows › Downloads

- Database 11g
- Database Focus Areas
- Database Options
- Database Express Edition
- Database Upgrade
- Berkeley DB
- NoSQL Database
- TimesTen In-Memory Database
- Exadata
- Database Mobile Server
- RDB
- Audit Vault
- Secure Backup
- Real Application Clusters
- Clusterware
- Cloud Storage

What's New | For Beginners | Products | Technical Information | **Downloads**

OracleODBC Drivers for Windows (32bit) Download Page

NOTE: The following ODBC Drivers are ODBC Version 3.52, Level 2 compatible drivers.

- Oracle ODBCDrivers for Oracle 10.2

Version	Date	Size	
10.2.0.3.0	12-Dec-2006	340 KB	New
10.2.0.2.0	08-Sep-2006	2.3 MB	

- Oracle ODBC Drivers for Oracle 10.1

Version	Date	Size
10.1.0.5.0	25-Apr-2006	2.1 MB

- Oracle ODBC Drivers for Oracle 9.2.0

Version	Date	Size
9.2.0.8.0	20-Dec-2006	280 KB

More Downloads

- 64-bit Oracle Data Access Components (ODAC)
- ODBCInstant Client Driver for
  - Windows 32-bit
  - Windows x86-64bit
  - Linux x86-32bit
  - Solaris SPARC 64bit
  - Solaris SPARC 32bit

Oracle ODBC DriverResources

- Documentation
- Discussion Forum

http://www.oracle.com/technetwork/database/windows/downloads/

Installation and Configuration

1) Install Oracle XE (or get access to another Oracle database). There is a fair amount of documentation, but in practice the installation is pretty straightforward on Windows with a user with Admin rights.

2) Install the sample schema if needed. This would require a database user with special privileges - so you might need to get a DBA involved if you aren't using your own machine.

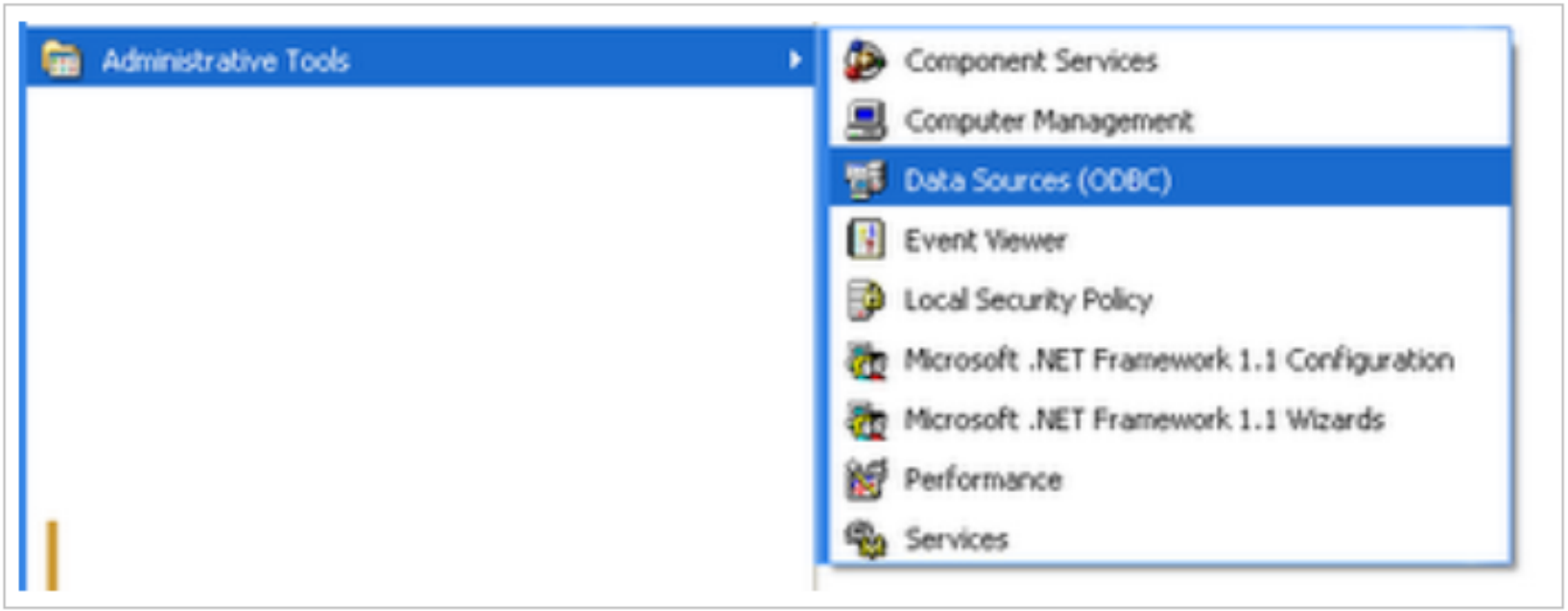
3) Unlock the HR user account. This can also be done by logging in as the system user at a SQL*Plus command prompt and executing

Rattle Training 3. Loading Data

Configure ODBC Connection

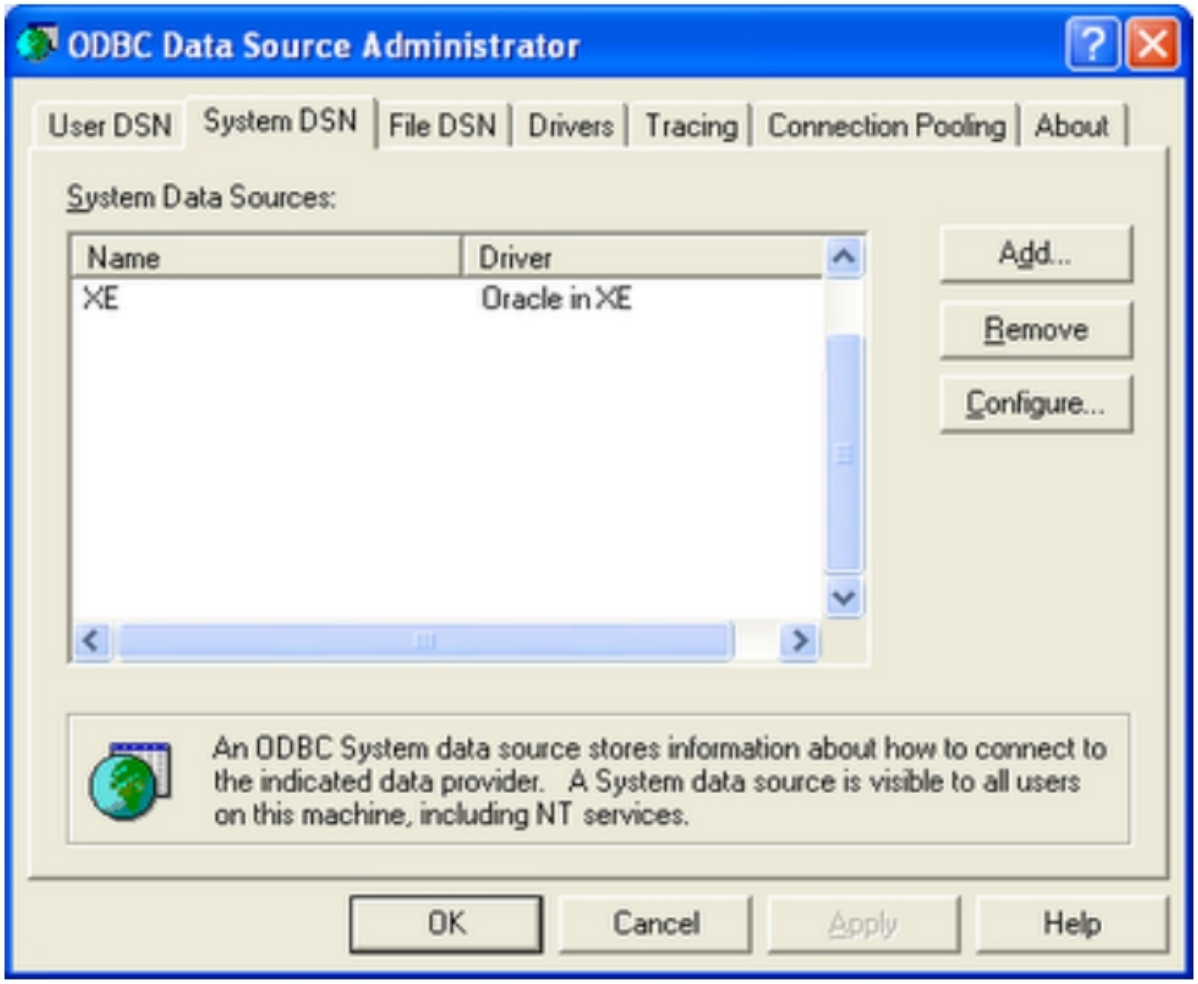
These examples will rely upon the RODBC package.

1) Find the Data Sources (ODBC) option under Administrative Tools.



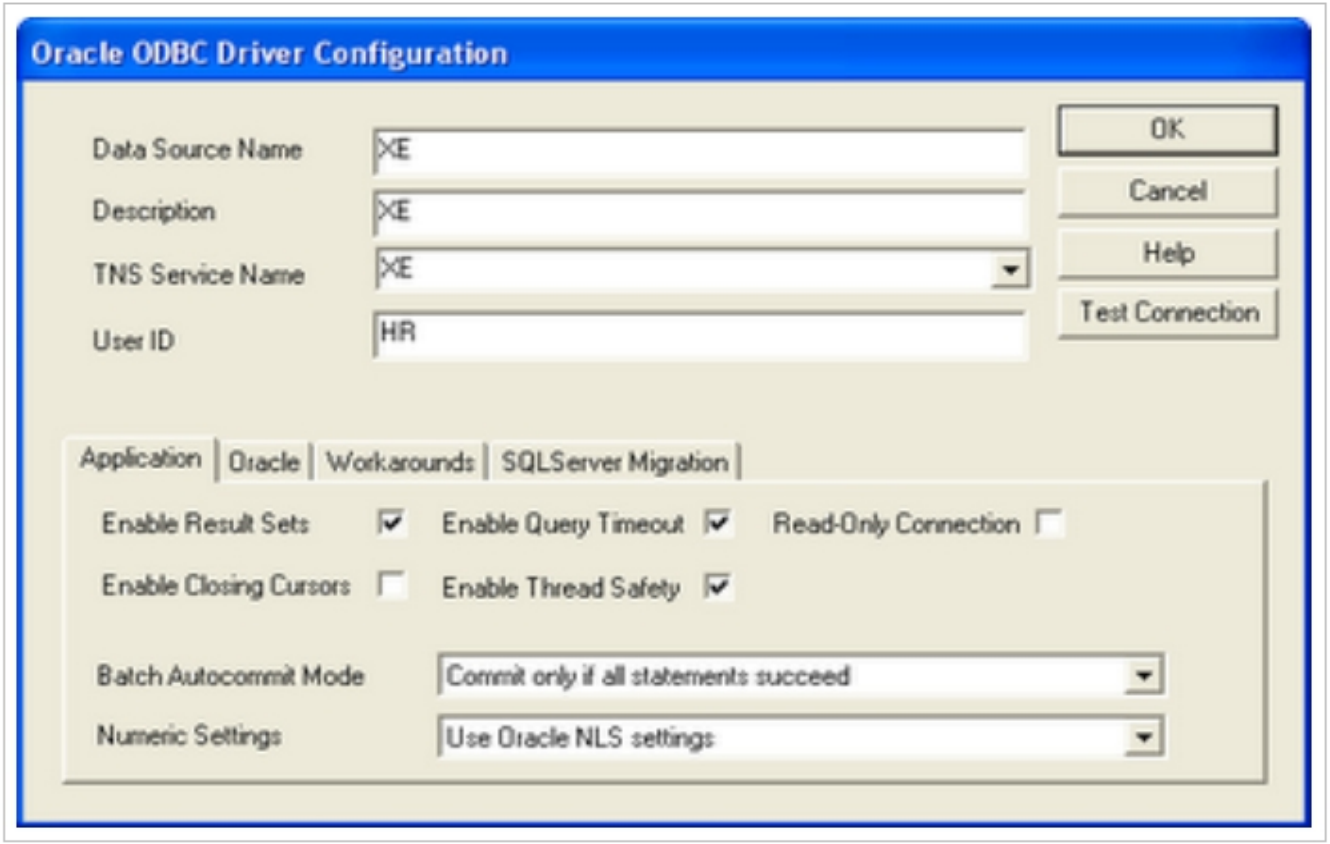
Rattle Training 3. Loading Data

2) Add a connection to your database if needed. In the case of Oracle XE, choose the add button if you do not have an XE configured.



Rattle Training 3. Loading Data

Configure the details in the popup window. You probably will only need to enter the Data Source Name, Description, TNS Service Name and User ID. Just leave other settings as defaulted.



Rattle Training 3. Loading Data

Test R connection with Oracle

Install RODBC if you have not already done so. Run the following program from R.

```
library(RODBC)
ch <- odbcConnect("XE",uid="HR",pwd="HR")

sqlTables(ch,schema='HR')
```

You can execute queries using the connection as well.

```
sqlQuery(ch, "select
sys_context('USERENV','DB_NAME') db from dual")
```

Rattle Training 3. Loading Data

- R Connect MySQL.

การเชื่อมต่อ R กับ MySQL.

1. Install ODBC Driver for MySQL at:
http://dev.mysql.com/downloads/connector/odbc/5.1.html

2. Open Microsoft Windows Control Panel and then double-click the **Administrative Tools** icon. Double-click on the **Data Sources (ODBC)** icon. Add a new User DSN called pfx with a MySQL ODBC 5.1 Diver.

3. Once a new User DSN is created, select it and press Configure

Set the following values (I am not 100% sure these are the same for all Window machines)
Server: localhost
port: 3306
user: root
Database: php2

These are the same as the ones on the SQLyog login and the Pitch F/X dataset installed

Rattle Training 3. Loading Data

- MySQL Community Server
- MySQL Cluster
- MySQL Workbench (GUI Tool)
- MySQL Proxy
- **MySQL Connectors**
  - Connector/ODBC
  - Connector/J
  - Connector/Net
  - Connector/MXJ
  - Connector/C++
  - Connector/C
  - MySQL Native Driver for PHP
- MySQL Installer (Windows)

MySQL Newsletter
Subscribe Today!
Over 1 Million Subscribers

Download Connector/ODBC

Connector/ODBC is a standardized database driver for Windows, Linux, Mac OS X, and Unix platforms

Online Documentation:

- Connector/ODBC Documentation and Change History

Please report any bugs or inconsistencies you observe to our Bugs Database.

Thank you for your support!

MySQL open source software is provided under the GPL License.

OEMs, ISVs and VARs can purchase commercial licenses.

Generally Available (GA) Releases

http://dev.mysql.com/downloads/connector/odbc

Connector/ODBC 5.1.9

Looking for previous GA versions?

Select Platform:

Microsoft Windows | Select

Windows (x86, 64-bit), MSI Installer Connector-ODBC (mysql-connector-odbc-5.1.9-winx64.msi)	5.1.9	4.7M	Download
MD5: e8a4ccc1e7d0b0427ea26a6453673238 | Signature			
Windows (x86, 32-bit), MSI Installer Connector-ODBC	5.1.9	2.7M	Download

Rattle Training 3. Loading Data

ในตัวอย่างนี้ตั้งชื่อ Data Source Name ว่า pfx

Rattle Training 3. Loading Data

- ใช้ library(RODBC) และคำสั่งข้างล่าง

4. Open R again

In the R console run:

```
library(RODBC)
channel <- odbcConnect('pfx')
odbcGetInfo(channel)
sqlTables(channel)
```

All 4 lines can be copy and paste at once. These will connect R to database and let you know that they are connect. No errors should be seen and the Pitch FX tables should be listed.

Rattle Training 3. Loading Data

- การกรองข้อมูลแบบ SQL สามารถทำได้ด้วยคำสั่ง sqlQuery (“...”)

5. Running a query. First you need to tell R that you are plotting a query.

Here is the query for Joe Mauer's home runs. I was having some problems, so I added "as x" and "as y" to describe the px and pz coordinates. Note, the entire query needs to be on a single line.

query <- sqlQuery(channel, "select p.px as x, p.pz as y from atbats a join pitches p on p.ab_id = a.ab_id join games g on g.game_id = a.game_id join players pl on pl.eliasid = a.batter where pl.last = 'Mauer' and pl.first = 'Joe' and g.date like '2009%' and a.event = 'Home Run' and p.type = 'X';")

Also the extents of data can be set at this time with the xlim and ylim commands

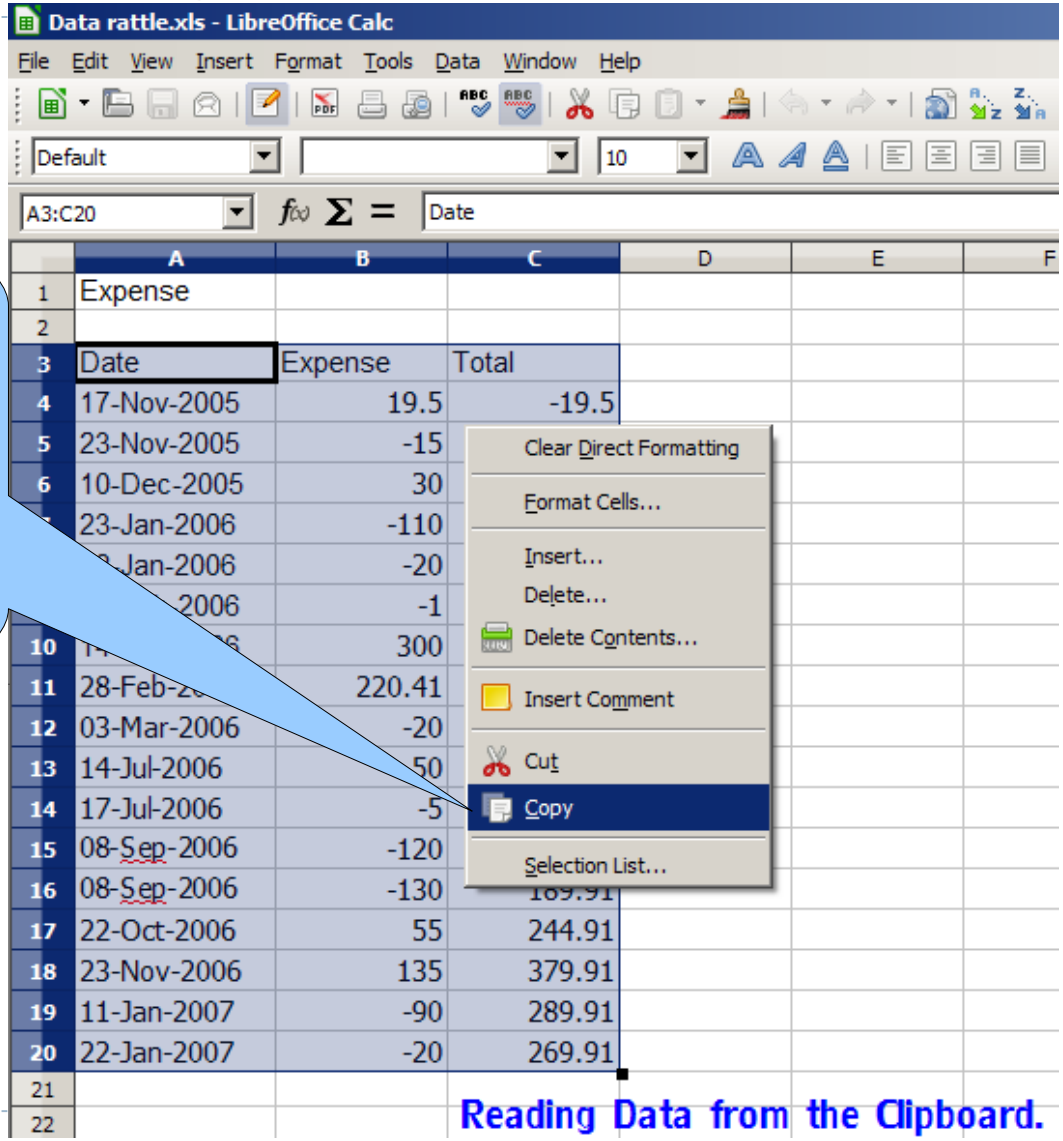
Rattle Training 3. Loading Data

- แหล่งความรู้เพิ่มเติมเกี่ยวกับการเชื่อมต่อ R กับ OBDC

- R vs Microsoft SQL Server 2012
  - http://blog.sqltrainer.com/2011/12/statistical-analysis-with-r-and.html
- Fun with Decision Trees using R and SQL Server
  - http://blog.sqltrainer.com/
- Knowledge Discovery in Databases -KDD using PostgreSQL and #Rstats
  - http://www.decisionstats.com/tag/knowledge-discovery/
- R and Oracle HR Part I - Set Up and Connect
  - http://www.r-chart.com/2010/07/r-and-oracle-hr-part-i-set-up-and.html
- Using R with MySQL #rstats
  - http://www.decisionstats.com/using-r-with-mysql/

Rattle Training 3. Loading Data

เปิดไฟล์ spreadsheet ขึ้นมา แล้ว Copy ไว้ข้อมูลก็จะถูกเก็บไว้ที่หน่วยความจำ Clipboard ซึ่ง R สามารถนำมาใช้ได้



	A	B	C
1	Expense		
2			
3	Date	Expense	Total
4	17-Nov-2005	19.5	-19.5
5	23-Nov-2005	-15	
6	10-Dec-2005	30	
	23-Jan-2006	-110	
	Jan-2006	-20	
	2006	-1	
10		300	
11	28-Feb-	220.41	
12	03-Mar-2006	-20	
13	14-Jul-2006	50	
14	17-Jul-2006	-5	
15	08-Sep-2006	-120	
16	08-Sep-2006	-130	189.91
17	22-Oct-2006	55	244.91
18	23-Nov-2006	135	379.91
19	11-Jan-2007	-90	289.91
20	22-Jan-2007	-20	269.91
21			
22			

Rattle Training 3. Loading Data

R Console

```
> expense <- read.table ('clipboard',header=T)
> expense
         Date Expense   Total
1  17-Nov-2005   19.50  -19.50
2  23-Nov-2005  -15.00    4.50
3  10-Dec-2005   30.00   34.50
4  23-Jan-2006 -110.00  -75.50
5  28-Jan-2006  -20.00  -95.50
6  14-Feb-2006   -1.00 -105.50
7  14-Feb-2006  300.00  194.50
8  28-Feb-2006  220.41  414.91
9  03-Mar-2006  -20.00  394.91
10 14-Jul-2006   50.00  444.91
11 17-Jul-2006   -5.00  439.91
12 08-Sep-2006 -120.00  319.91
13 08-Sep-2006 -130.00  189.91
14 22-Oct-2006   55.00  244.91
15 23-Nov-2006  135.00  379.91
16 11-Jan-2007  -90.00  289.91
17 22-Jan-2007  -20.00  269.91
>
```

R Data Miner - [Rattle (expense)]

Project Tools Settings Help

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Source: ○ Spreadsheet ○ ARFF ○ ODBC ◉ R Dataset ○ RData File ○ Library ○ Cor

Data Name: expense

☑ Partition 70/15/15 Seed: 42 View Edit

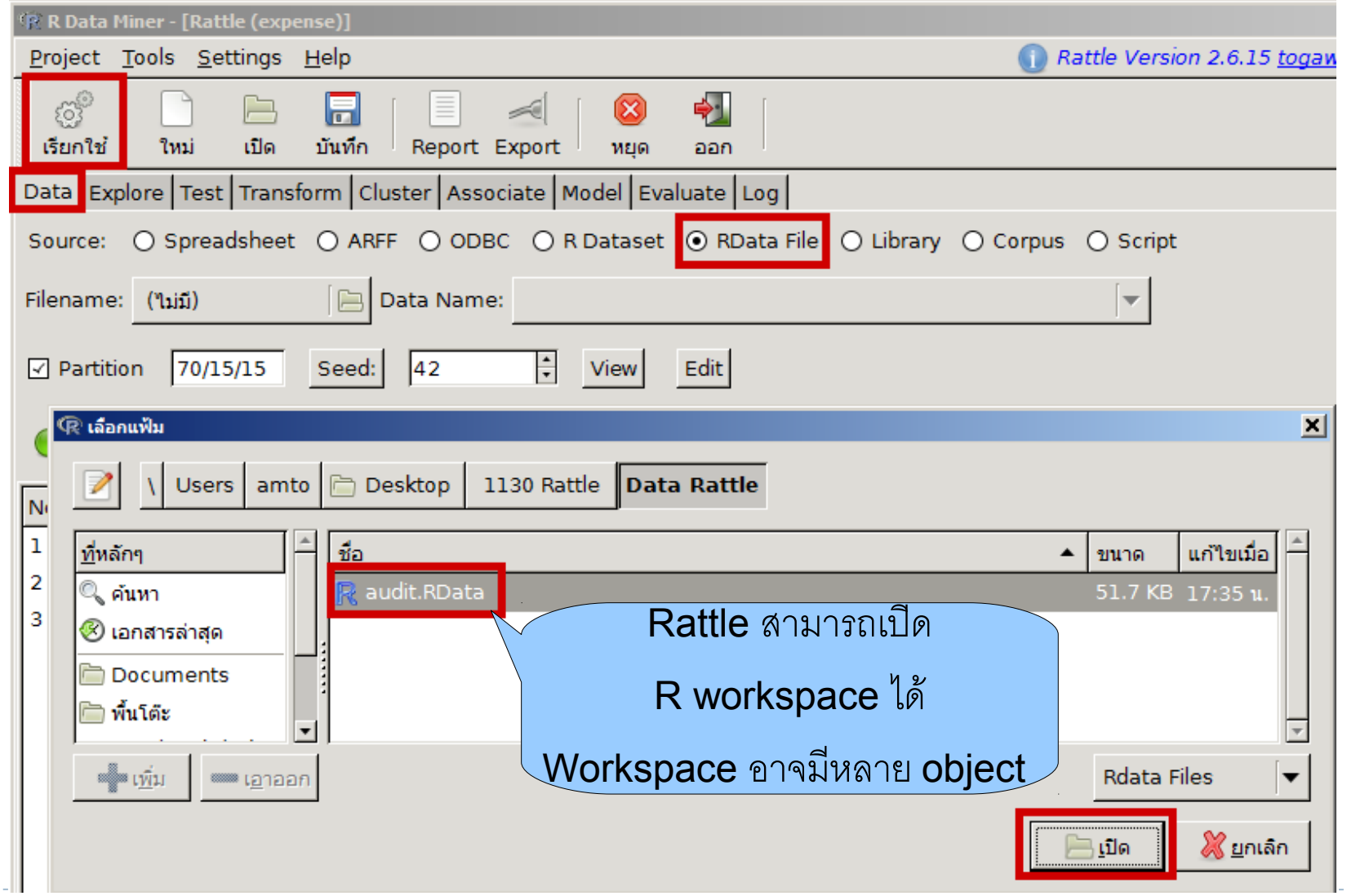
Input Ignore Weight Calculator:

Target Data Type ◉ Auto ○ Categoric ○ Nume

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Categoric	◉	○	○	○	○	○	Unique: 15
2	Expense	Numeric	◉	○	○	○	○	○	Unique: 15
3	Total	Numeric	◉	○	○	○	○	○	Unique: 17

Rattle สามารถดึงข้อมูลจาก R object environment ได้

Rattle Training 3. Loading Data



Rattle Training 3. Loading Data

R Data Miner - [Rattle (audit)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: audit.RData Data Name: audit

☑ Partition 70/15/15 Seed: 42 View Edit

ถ้า R workspace มีหลายไฟล์ จะต้องเลือกที่นี่

Input Ignore Weight Calculator:

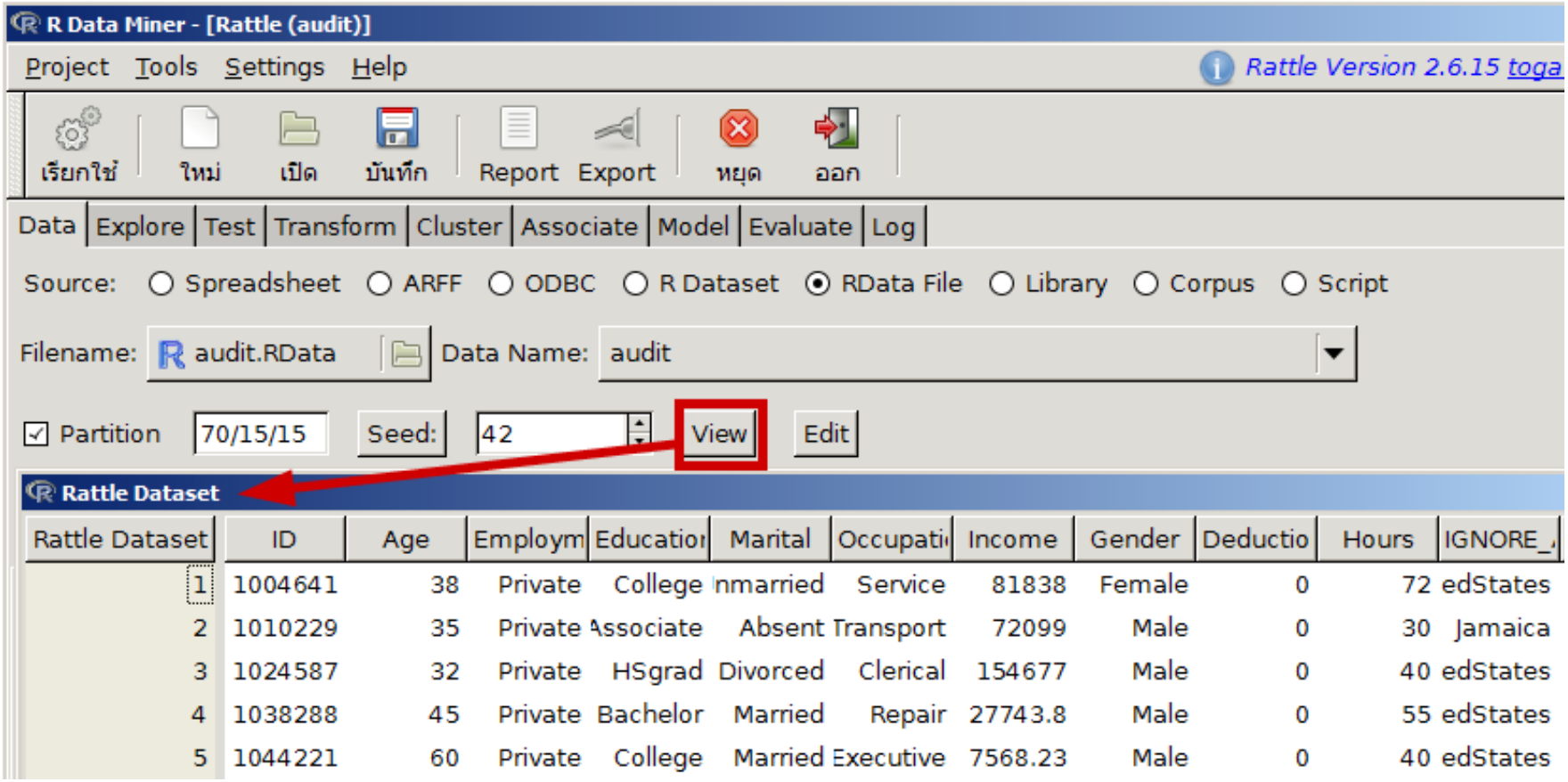
Target Data Type: ◉ Auto ○ Categoric ○ Numeric ○ Survival

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 Missing: 100
4	Education	Categoric	◉	○	○	○	○	○	Unique: 16
5	Marital	Categoric	◉	○	○	○	○	○	Unique: 6
6	Occupation	Categoric	◉	○	○	○	○	○	Unique: 14 Missing: 101
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 Missing: 43
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 enabled.

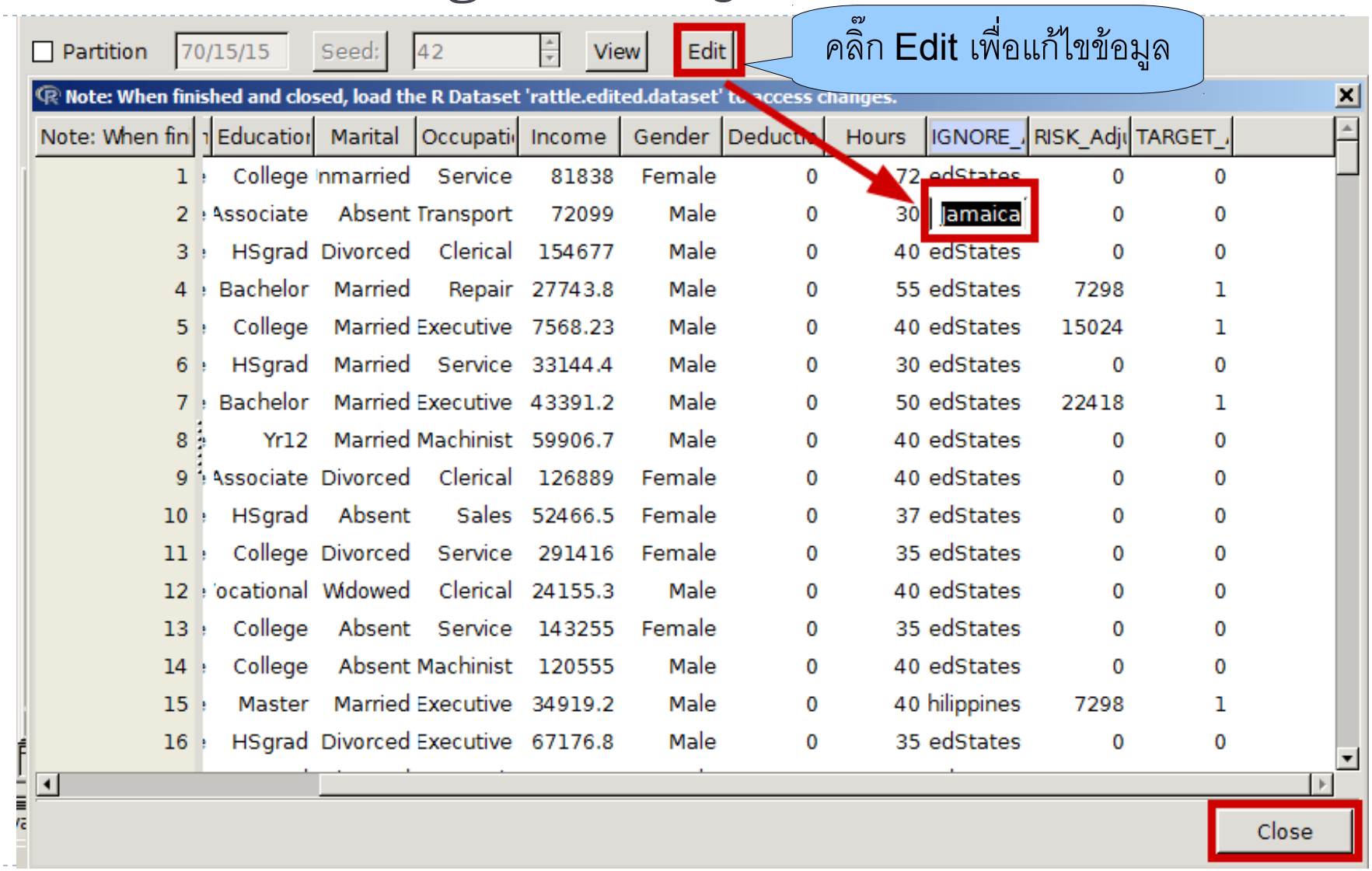
Rattle Training 3. Loading Data

- กดปุ่ม View เพื่อดูข้อมูลในรูปของตาราง



Rattle Dataset	ID	Age	Employm	Educatio	Marital	Occupati	Income	Gender	Deductio	Hours	IGNORE_
1	1004641	38	Private	College	nmarried	Service	81838	Female	0	72	edStates
2	1010229	35	Private	Associate	Absent	Transport	72099	Male	0	30	Jamaica
3	1024587	32	Private	HSgrad	Divorced	Clerical	154677	Male	0	40	edStates
4	1038288	45	Private	Bachelor	Married	Repair	27743.8	Male	0	55	edStates
5	1044221	60	Private	College	Married	Executive	7568.23	Male	0	40	edStates

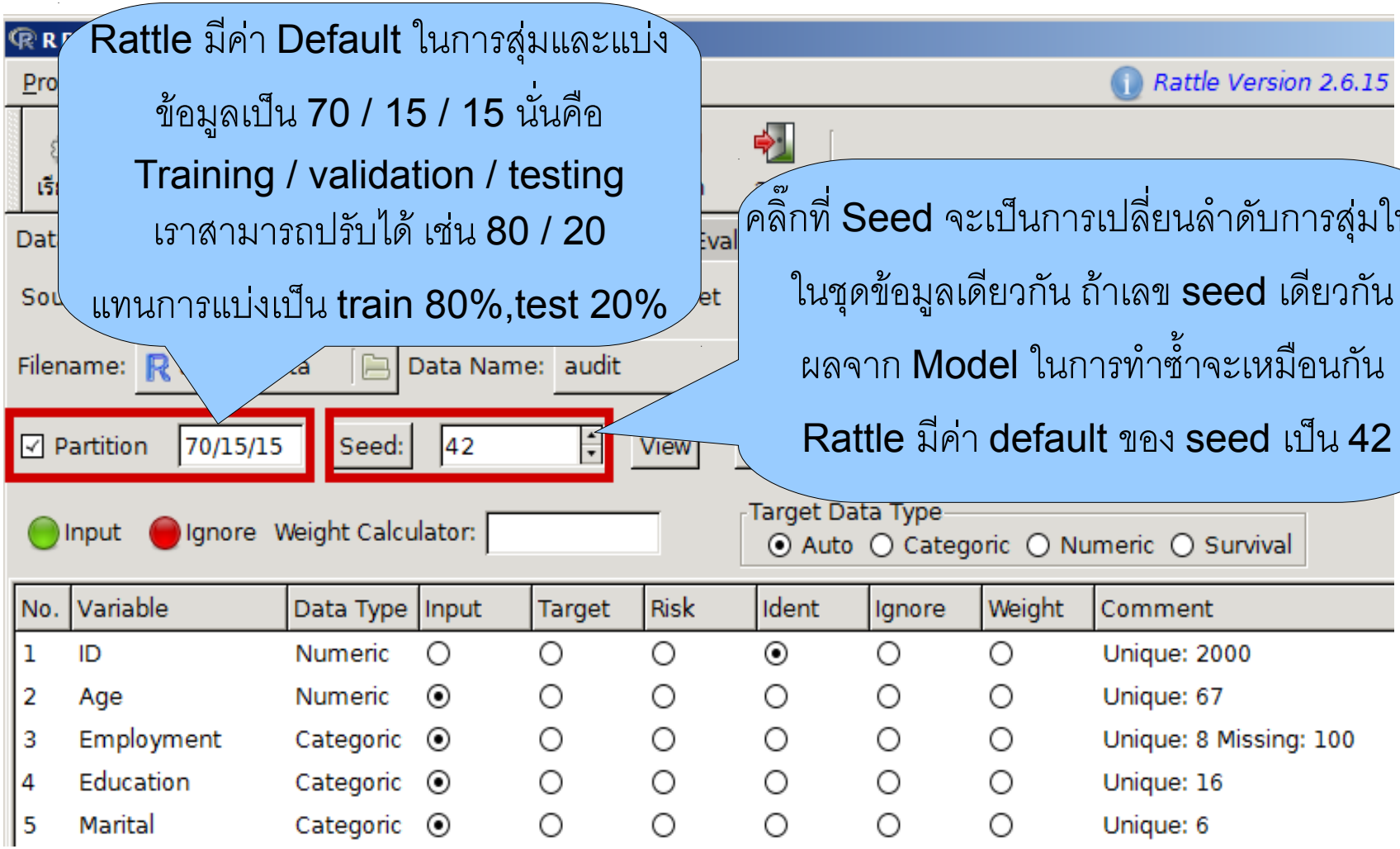
Rattle Training 3. Loading Data



Rattle Training 3. Loading Data : (Data Option)

Rattle มีค่า Default ในการสุ่มและแบ่งข้อมูลเป็น 70 / 15 / 15 นั่นคือ Training / validation / testing เราสามารถปรับได้ เช่น 80 / 20 แทนการแบ่งเป็น train 80%,test 20%

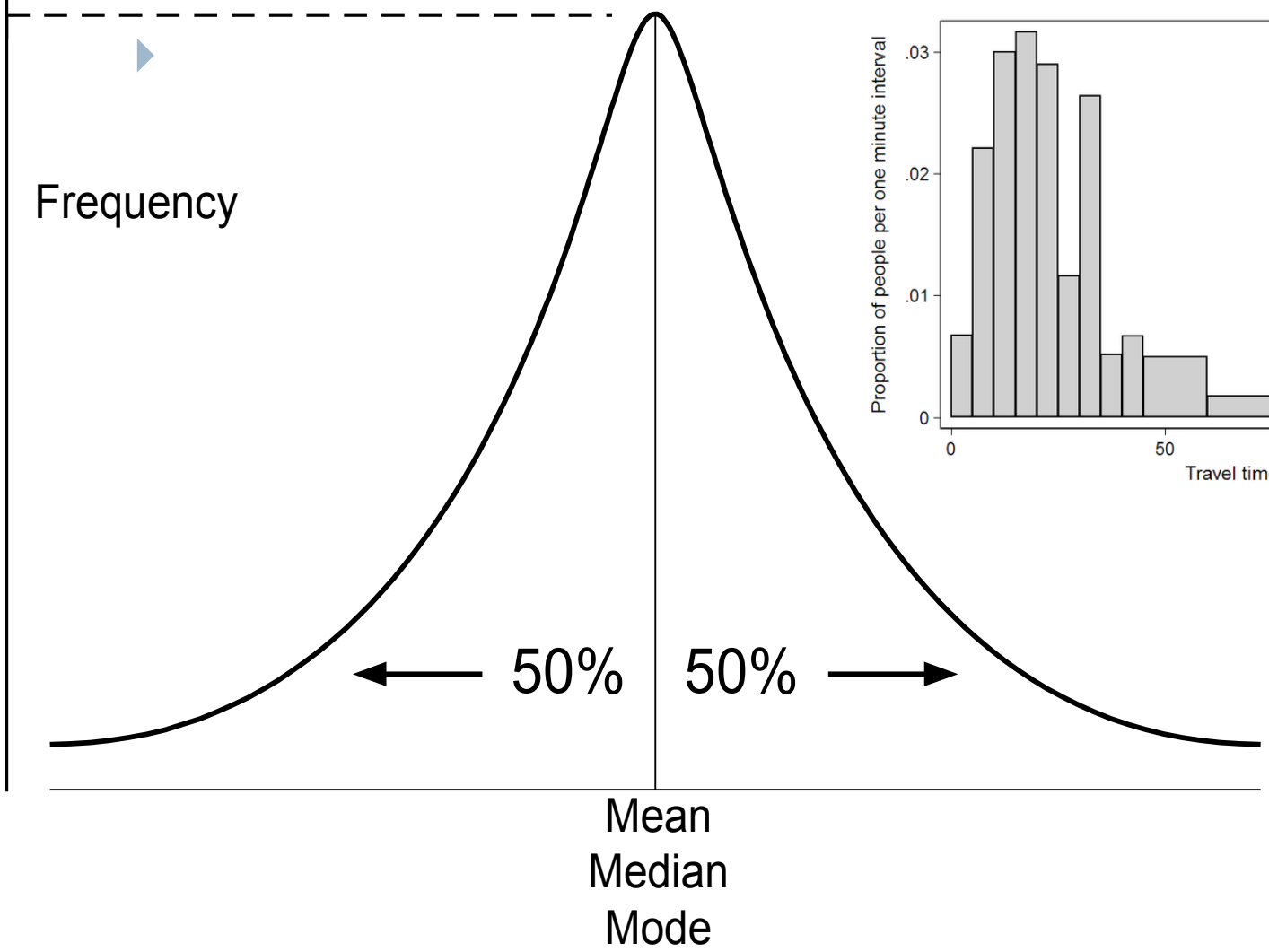
คลิ๊กที่ Seed จะเป็นการเปลี่ยนลำดับการสุ่มใหม่ในชุดข้อมูลเดียวกัน ถ้าเลข seed เดียวกัน ผลจาก Model ในการทำซ้ำจะเหมือนกัน Rattle มีค่า default ของ seed เป็น 42

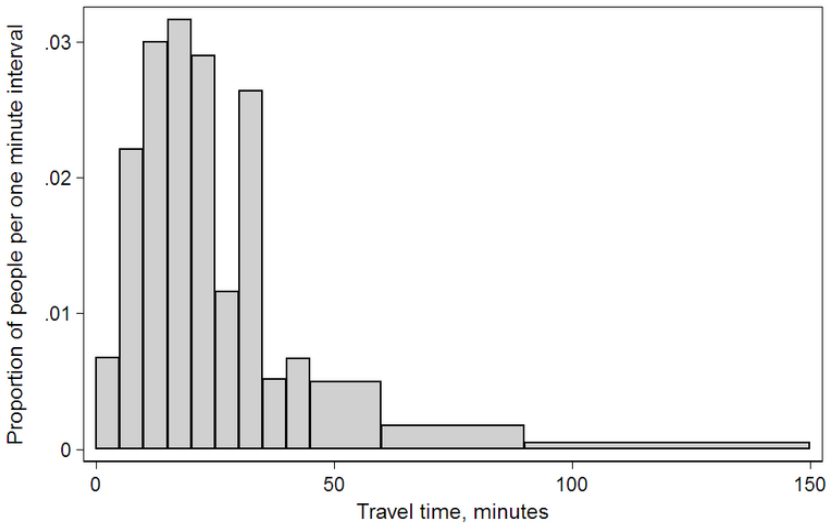


Rattle Training 3. Loading Data : (Variable Roles)

- Data Type (Rattle จะเดาประเภทของข้อมูลที่นำเข้าโดยอัตโนมัติ)
  - Ident : แทนข้อมูลที่เป็น Unique ไม่ซ้ำกันเลยทั้งชุดข้อมูล
  - Constant : แทนข้อมูลที่มี Unique : 1 (และจะเปลงเป็น Ignore โดยอัตโนมัติ)
  - Numeric: แทนข้อมุลเชิงปริมาณ
  - Categoric : แทนข้อมูลเชิงกลุ่ม
- Input :ข้อมูลที่สามารถนำเข้าไปวิเคราะห์ใน rattle ได้
- Target : ข้อมูลที่เป็น Class
- Risk : ข้อมูลที่บันทึกขนาดความเสี่ยงของแต่ละ case ไว้ (แปรตาม class)
- Ident : ระบุประเภทข้อมูลที่ไม่ซ้ำกันเลย
- Ignore : ใช้ระบุเมื่อไม่ต้องการนำตัวแปรนั้นๆ มาวิเคราะห์
- Weight : ใช้ถ่วงน้ำหนักกรณีที่แต่ละ case สำคัญไม่เท่ากัน
- Comment : เป็นข้อสรุปข้อมูลเบื้องต้น เช่น Unique & Missing

Rattle Training 4. Exploring Data 4.1 Summarizing Data





Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

ใช้ไฟล์ weather.csv

สรุปข้อมูลคร่าวๆ

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

Type: (•) Summary ( ) Distributions ( ) Correlation ( ) Principal Components ( ) Interactive

[x] Summary [ ] Describe [ ] Basics [ ] Kurtosis [ ] Skewness [ ] Show Missing [ ] Cross Tab

```
   MinTemp          MaxTemp          Rainfall
Min.   :-5.300   Min.   : 8.40   Min.   : 0.000
1st Qu.: 1.925   1st Qu.:14.70   1st Qu.: 0.000
Median : 7.100   Median :19.20   Median : 0.000
Mean   : 7.011   Mean   :20.29   Mean   : 1.386
3rd Qu.:12.400   3rd Qu.:24.95   3rd Qu.: 0.200
Max.   :20.900   Max.   :35.80   Max.   :25.800

   Sunshine       WindGustDir  WindGustSpeed      W
Min.   : 0.000   NW      : 51   Min.   :13.00   SE
1st Qu.: 6.425   E       : 29   1st Qu.:31.00   NNW
Median : 8.750   NNW     : 28   Median :39.00   NW
Mean   : 8.063   ENE     : 24   Mean   :39.72   SSE
3rd Qu.:10.400   WNW     : 21   3rd Qu.:46.00   ESE
Max.   :13.600   (Other):101    Max.   :85.00   (Ot
NA's   : 2.000   NA's    :  2   NA's   : 2.00   NA'
```

R Console

```
> library(rattle)
> dim(weather)
[1] 366  24
>
> summary(weather[7:9])
    Sunshine        WindGustDir   WindGustSpeed
 Min.   : 0.000   NW      : 73   Min.   :13.00
 1st Qu.: 5.950   NNW     : 44   1st Qu.:31.00
 Median : 8.600   E       : 37   Median :39.00
 Mean   : 7.909   WNW     : 35   Mean   :39.84
 3rd Qu.:10.500   ENE     : 30   3rd Qu.:46.00
 Max.   :13.600   (Other):144    Max.   :98.00
 NA's   : 3.000   NA's    :  3   NA's   : 2.00
>
```

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

หาค่าสถิติเชิงพรรณา
(Descriptive Statistic)

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

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

```
Below is a description of the dataset.
The data is limited to the training dataset.

crs$dataset[crs$sample, c(crs$input, crs$risk, crs$target)]

 22  Variables      256  Observations
---------------------------------------------------------------------------------------
MinTemp
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90     .95
    256       0     154   7.011  -2.525  -1.150   1.925   7.100  12.400  15.350  16.500

lowest : -5.3 -3.7 -3.5 -3.4 -3.3, highest: 17.9 18.0 18.2 19.9 20.9
---------------------------------------------------------------------------------------
MaxTemp
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90     .95
    256       0     158   20.29   11.30   12.05   14.70   19.20   24.95   29.70   32.50

lowest :  8.4  8.8  9.3  9.5  9.6, highest: 34.9 35.0 35.2 35.7 35.8
---------------------------------------------------------------------------------------
```

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Console

```
> library(Hmisc)
Loading required package: survival
Loading required package: splines

Attaching package: 'Hmisc'

The following object(s) are masked from 'package:survival':

    untangle.specials

The following object(s) are masked from 'package:base':

    format.pval, round.POSIXt, trunc.POSIXt, units

> describe(weather[7])
weather[7]

 1  Variables      366  Observations
---------------------------------------------------------------------------
Sunshine
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90
    363       3     114   7.909    0.60    2.04    5.95    8.60   10.50   11.80
    .95
  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 numeric variables.

ตัวอย่างการวิเคราะห์ข้อมูลที่เป็นตัวเลขเชิงปริมาณใน R

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Console

```
> library(fBasics)
> basicStats(weather$Sunshine)
            X..weather.Sunshine
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
```

library (fBasics) ใน R

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: (•) Summary ( ) Distributions ( ) Correlation ( ) Principal Components ( ) Interactive

[ ] Summary [ ] Describe [x] Basics [ ] Kurtosis [ ] Skewness [ ] Show Missing [ ] Cross Tab

```
Basic statistics for each numeric variable of the dataset.

$MinTemp
               X...X.1
nobs        256.000000
NAs           0.000000
Minimum      -5.300000
Maximum      20.900000
1. Quartile   1.925000
3. Quartile  12.400000
Mean          7.011328
Median        7.100000
Sum        1794.900000
SE Mean       0.386658
LCL Mean      6.249878
UCL Mean      7.772778
Variance     38.273165
Stdev         6.186531
Skewness      0.053472
Kurtosis     -1.131089
```

Option นี้จะแสดงข้อมูลสถิติเชิงพรรณา ได้ละเอียดมากขึ้น ในแต่ละตัวแปร

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Console

For categoric variables

```
> describe(weather[8])
weather[8]

 1  Variables      366  Observations
---------------------------------------------------------------------------
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
---------------------------------------------------------------------------
> fix(weather)
```

ตัวอย่างการวิเคราะห์เชิงกลุ่มใน R

Data Editor

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine (7)	WindGustDir (8)
1	2007-11-01	Canberra	8	24.3	0	3.4	6.3	NW
2	2007-11-02	Canberra	14	26.9	3.6	4.4	9.7	ENE
3	2007-11-03	Canberra	13.7	23.4	3.6	5.8	3.3	NW
4	2007-11-04	Canberra	13.3	15.5	39.8	7.2	9.1	NW

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

Data Explore Test Transform Cluster Associate Model Evaluate Log

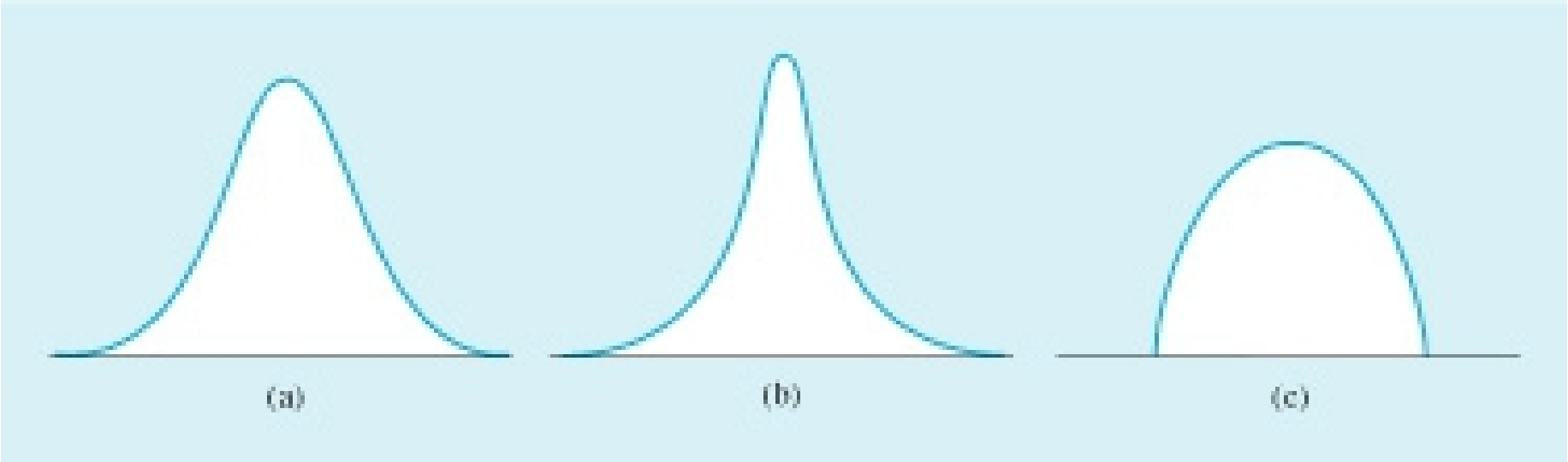
Type: Summary Distributions Correlation Principal Components Interactive

Summary Describe Basic Kurtosis Skewness Show Missing Cross Tab

```
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
  -1.13108886   -0.75164229   14.14542417   -0.09867312   -0.04982753
WindGustSpeed  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
   0.91838741    1.51304515   -0.31219908   -0.20414069   -0.01172661
  Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
  -0.04629149   -0.06395112   -1.70005241   -1.59243625   -1.02217157
      Temp3pm       RISK_MM
  -0.70472399   37.51316689

Rattle timestamp: 2012-01-11 00:13:21 amto
==========================================================================
```

Rattle Training 4. Exploring Data 4.1 Summarizing Data

- Degree to which tails of the distribution are "heavy" or "light"
  - heavy tails = higher Kurtosis(b)
  - Light tails = lower Kurtosis(c)
  - Normal distribution= Zero Kurtosis (a)



Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

Data Explore Test Transform Cluster Associate Model Evaluate Log

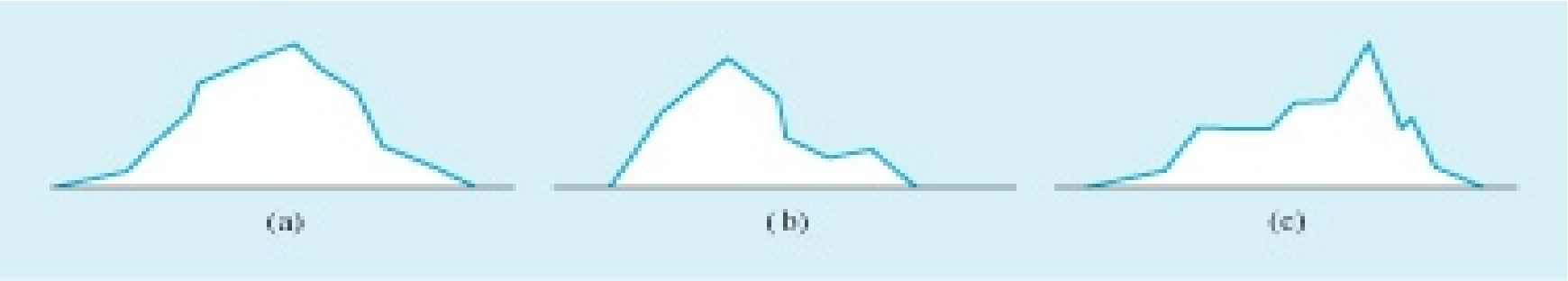
Type: Summary Distributions Correlation Principal Components Interactive

Summary Describe Basics Kurtosis Skewness Show Missing Cross Tab

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

      MinTemp       MaxTemp      Rainfall   Evaporation      Sunshine
   0.05347228    0.38465167    3.63725482    0.70497219   -0.79116605
WindGustSpeed  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
   0.78587838    1.32714547    0.40699979   -0.14866994    0.53559601
  Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
  -0.25824393   -0.13262711    0.20623961    0.17779326    0.01035714
      Temp3pm       RISK_MM
   0.36769008    5.42320876

Rattle timestamp: 2012-01-11 00:14:15 amto
==========================================================================
```

Rattle Training 4. Exploring Data 4.1 Summarizing Data

- Skewed right (b)
  - Fewer scores right of the peak
  - Positively skewed
  - Can be caused by a floor effect
- Skewed left (c)
  - Fewer scores left of the peak
  - Negatively skewed
  - Can be caused by a ceiling effect



Rattle Training 4. Exploring Data 4.1 Summarizing Data

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

```
 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
    RainTomorrow WindDir3pm WindGustSpeed Sunshine WindGustDir WindSpeed9am
328            1          1              1        1           1            1
  3            1          1              1        0           1            1
  1            1          1              1        1           0            1
 24            1          1              1        1           1            1
  1            1          0              1        1           1            1
  2            1          1              0        1           0            1
  7            1          1              1        1           1            0
               0          1              2        3           3            7
    WindDir9am
328          1  0
  3          1  1
  1          1  1
 24          0  1
  1          1  1
  2          1  2
  7          0  2
            31 47
```

การแสดง missing value

ใช้ md.pattern จาก mice package

Rattle Training 4. Exploring Data 4.1 Summarizing Data

```
>
> library(mice)
Loading required package: nnet
Loading required package: lattice
mice 2.11 2011-11-21
> md.pattern(weather[,7:10])
    WindGustSpeed Sunshine WindGustDir WindDir9am
329             1        1           1          1  0
  3             1        0           1          1  1
  1             1        1           0          1  1
 31             1        1           1          0  1
  2             0        1           0          1  2
                2        3           3         31 39
>
>
>
>
```

1 indicates a value is present.

0 indicates a value is missing.

observations with no missing.

The number of missing values over all observations.

Rattle Training 4. Exploring Data 4.1 Summarizing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: (•) Summary ( ) Distributions ( ) Correlation ( ) Principal Components ( ) Interactive

[ ] Summary [ ] Describe [ ] Basics [ ] Kurtosis [ ] Skewness [ ] Show Missing [x] Cross Tab

```
Cross tabulations:

CrossTab of WindGustDir by target variable RainTomorrow

   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


=========================================
                    crs$dataset[[crs$target]]
crs$dataset[[i]]       No      Yes    Total
-----------------------------------------
E                      33        4       37
                     30.3      6.7
                    0.246    1.106
                    0.892    0.108    0.102
                    0.111    0.061
                    0.091    0.011
-----------------------------------------
```

แสดงภาพรวมว่าตัวแปรเชิงกลุ่มใด จะมีความสัมพันธ์กัน Target บ้าง ด้วย สถิติ Chi-square.

นำข้อมูลเชิงกลุ่มมา crosstab กับ target

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Numeric: ล้าง Plots per Page: 4 ☐ Annotate — Target: RainTomorrow

☐ Benford Bars Benford Digit: 1 ◉ abs ○ +ve ○ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
3	MinTemp	☑	☐	☐	☐	-5.30; 7.45/7.27; 20.90
4	MaxTemp	☐	☐	☐	☐	7.60; 19.65/20.55; 35.80
5	Rainfall	☐	☐	☐	☐	0.00; 0.00/1.43; 39.80
6	Evaporation	☐	☐	☐	☐	0.20; 4.20/4.52; 13.80
7	Sunshine	☐	☑	☐	☐	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	☐	☐	☑	☐	13.00; 39.00/39.84; 98.00

Categoric: ล้าง

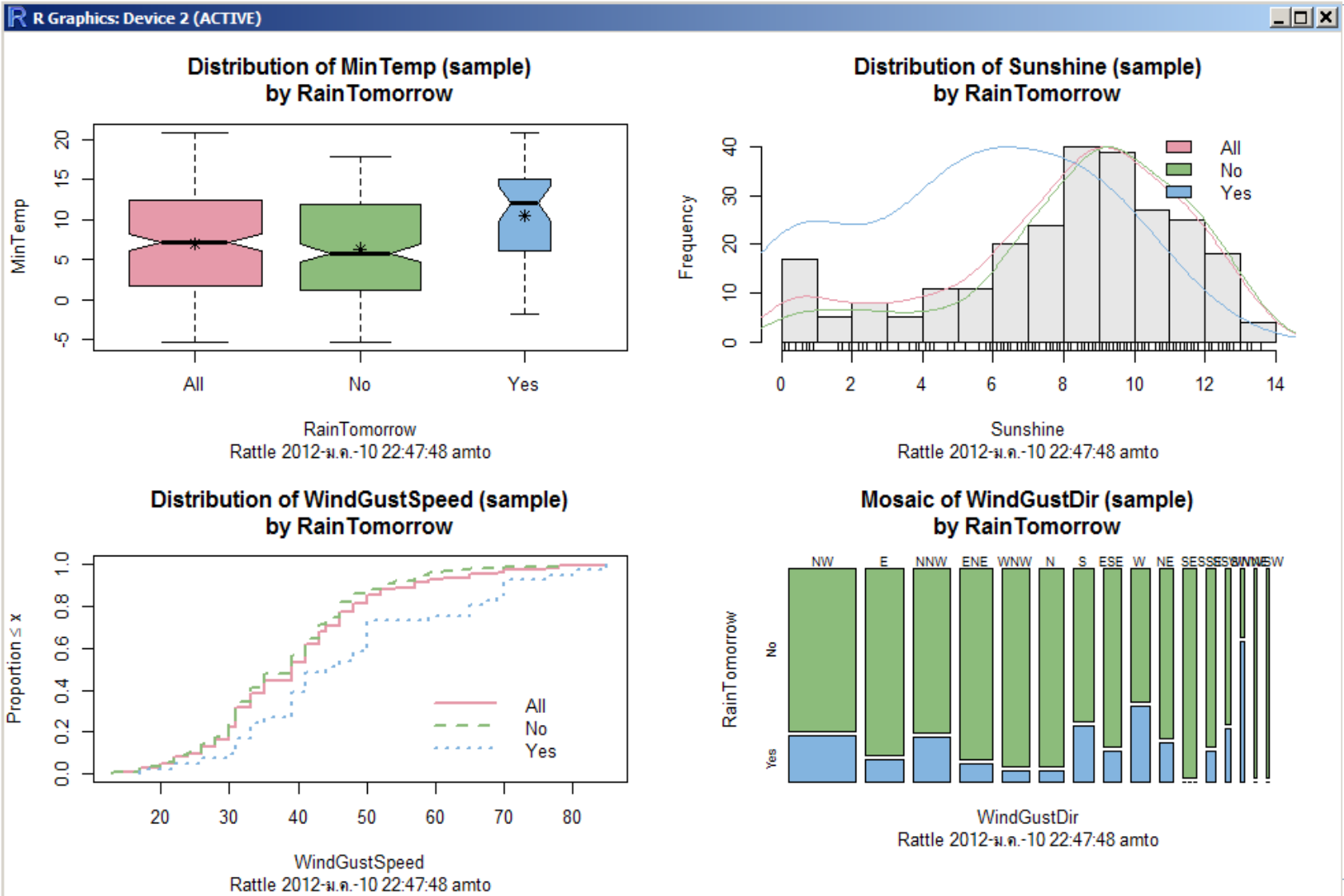
No.	Variable	Bar Plot	Dot Plot	Mosaic	Levels
8	WindGustDir	☐	☐	☑	16
10	WindDir9am	☐	☐	☐	16
11	WindDir3pm	☐	☐	☐	16
22	RainToday	☐	☐	☐	2
24	RainTomorrow	☐	☐	☐	2

Distribution จะสร้างกราฟที่น่าสนใจดังนี้

สำหรับ ***Numeric***:Box Plot,Histrogram Cumulative,Benford & Min;Median/Mean;Max

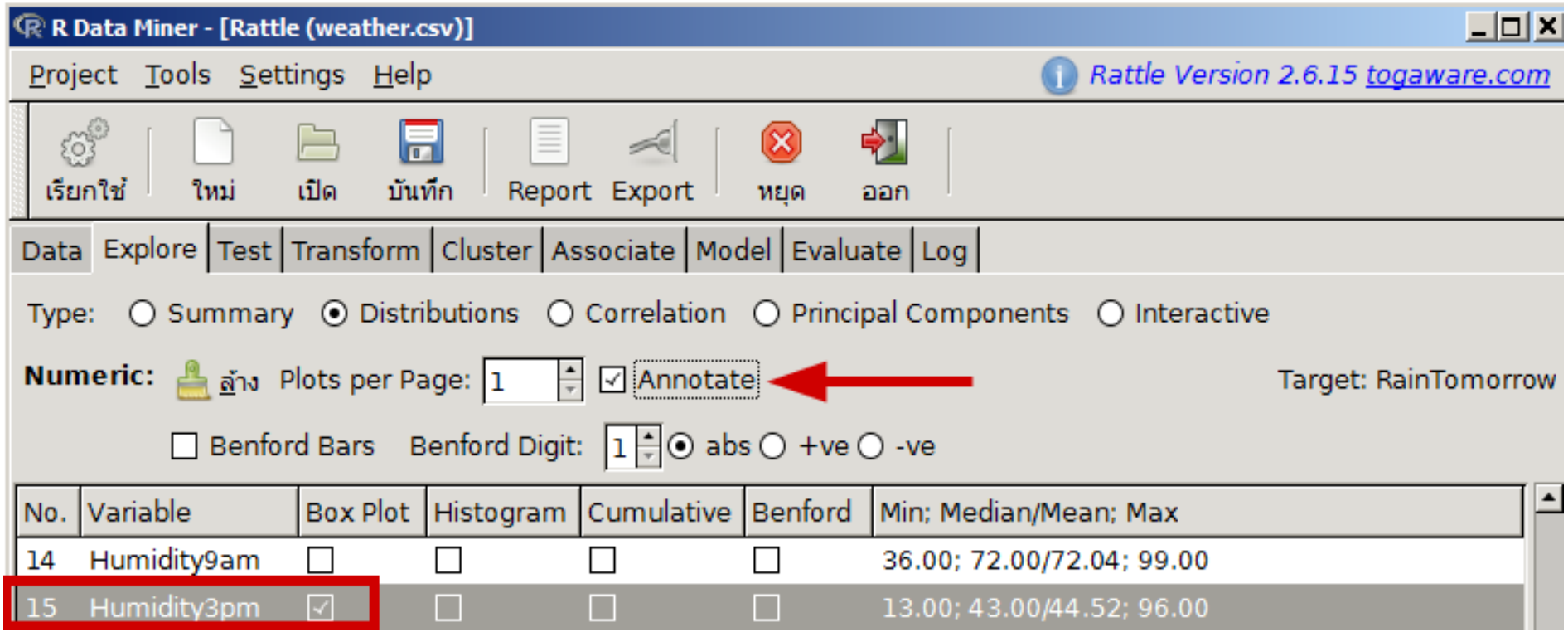
Categoric : Bar Plot,Dot Plot,Mosaic & Level

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions



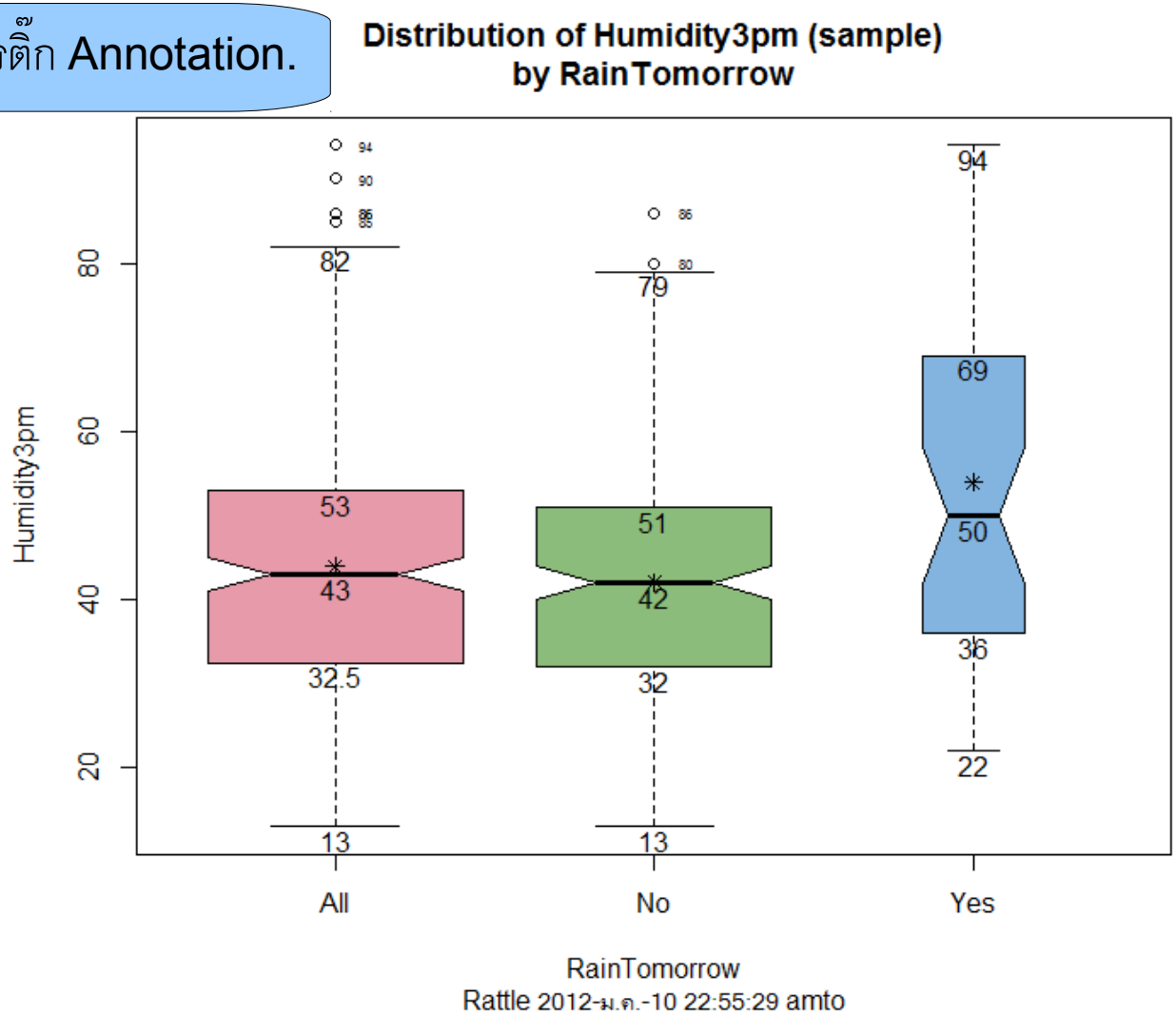
Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

- เราสามารถเลือกได้ว่าจะ Plot กี่รูปต่อ Page ที่ Plots per Page
- ถ้าต้องการให้แสดงค่าบนกราฟ ให้ติ๊ก Annotate.



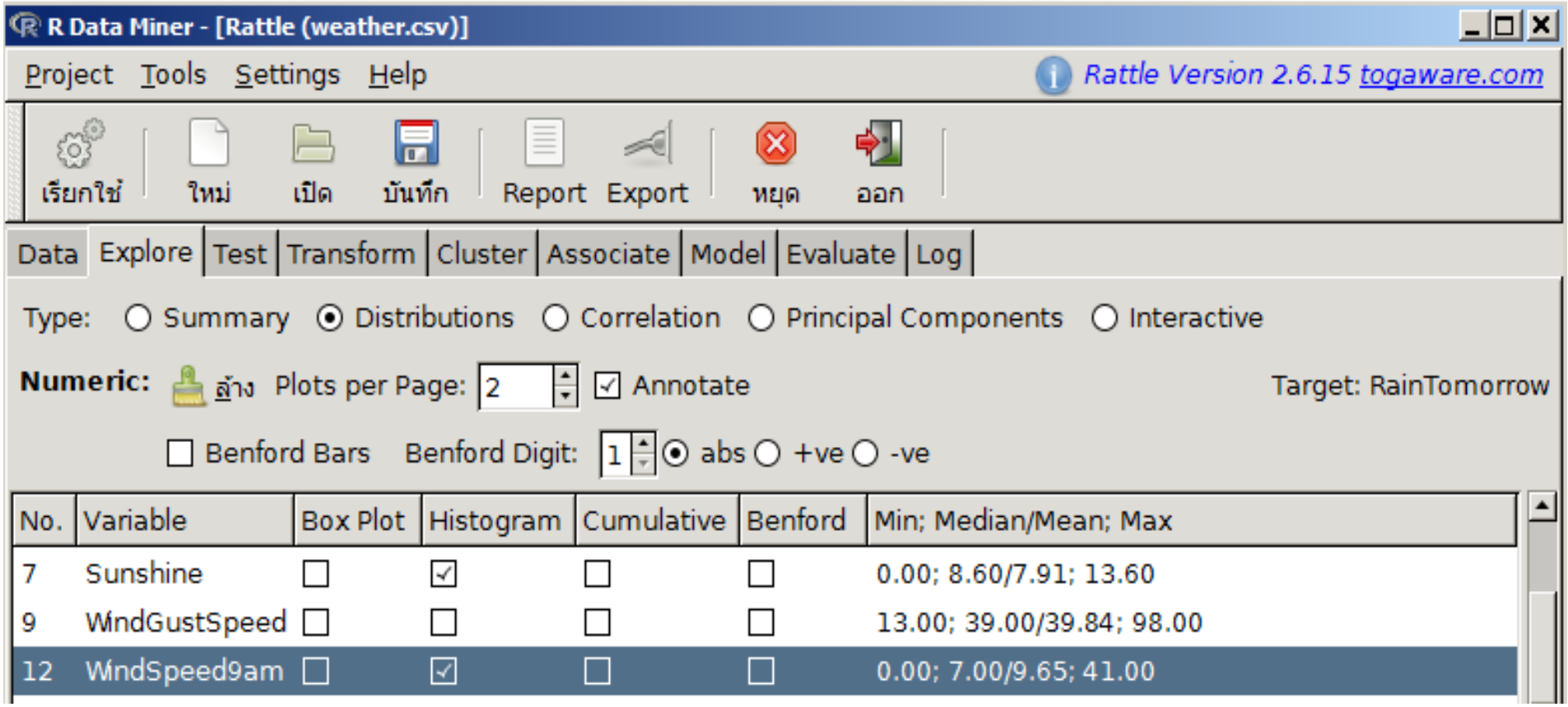
Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

ผลลัพธ์จากการติ๊ก Annotation.

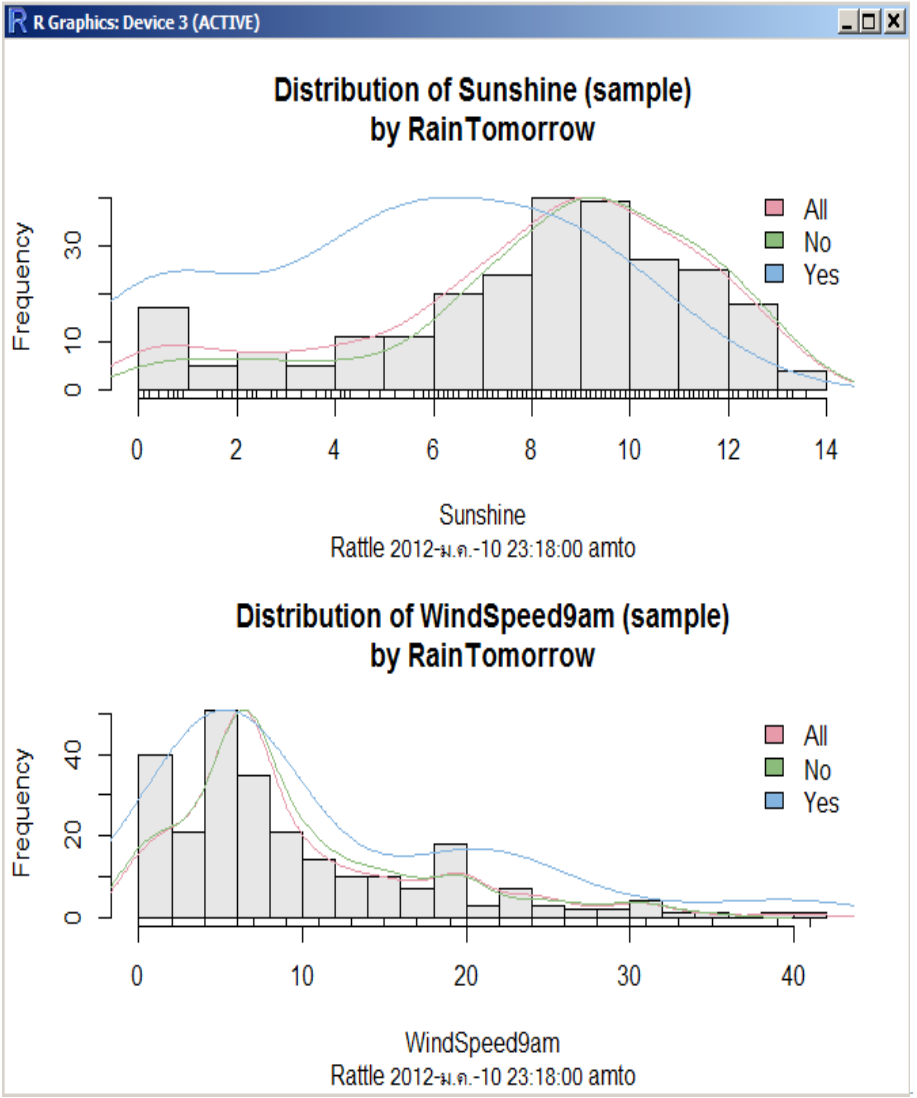


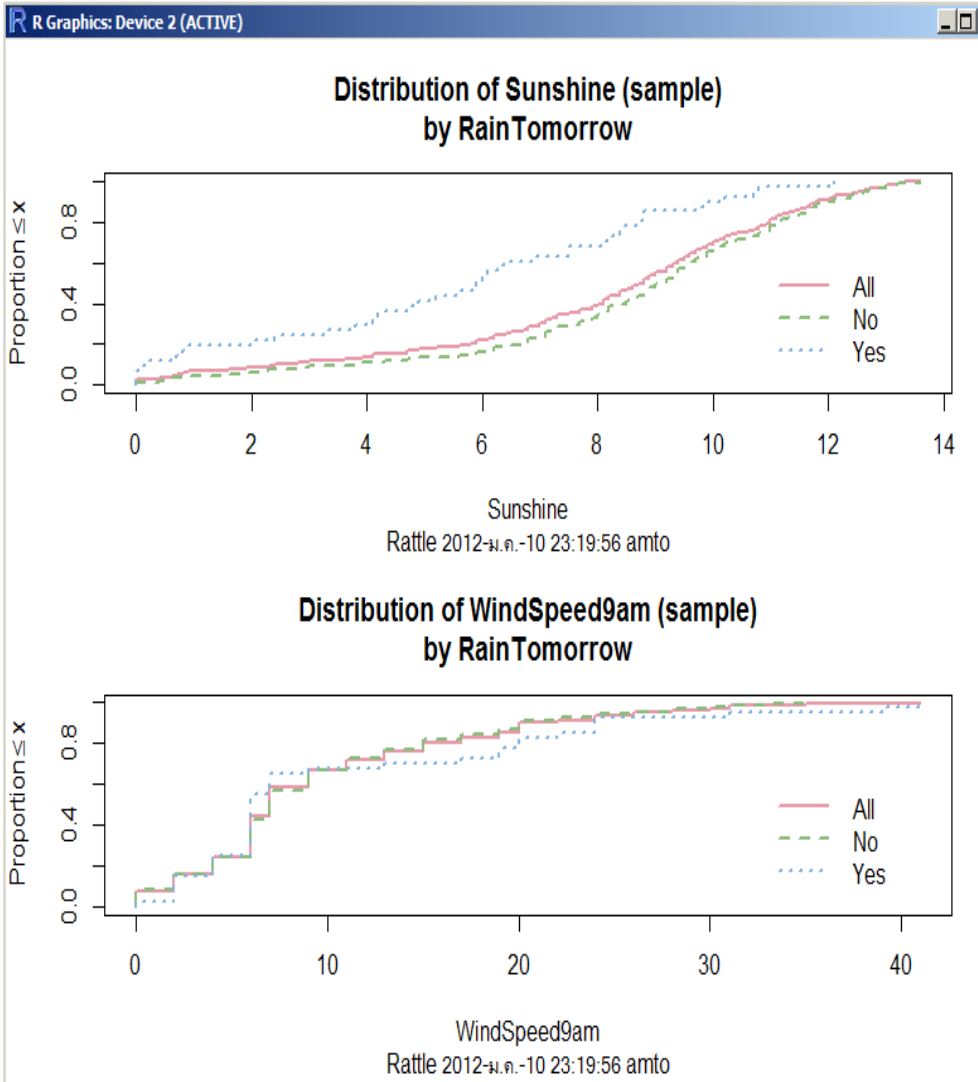
Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

- 1.ใช้ตัวแปร Sunshine & WindSpeed9am สร้าง Histogram.
- 2. ใช้ตัวแปร Sunshine & WindSpeed9am สร้าง Cumulative.



Rattle Training 4. Exploring Data 4.2 Visualizing Distributions





Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

- Histogram
  - จะให้ภาพการกระจายตัวของข้อมูลได้อย่างรวดเร็ว
  - พิจารณา shape ของข้อมูลซึ่งสะท้อนมาจาก skewness & kurtosis
  - Rattle มีเส้นการกระจายตัวของ Class ให้พิจารณาด้วย
- Cumulative Distribution Plot
  - แสดงสัดส่วนของข้อมูลที่มีค่าน้อยกว่าหรือเท่ากับค่าที่แสดงในแกน x
  - สำหรับ WindSpeed9am ข้อมูลส่วนใหญ่อยู่ในช่วง 0-10
  - ข้อมูลไม่ค่อยสะสมเพิ่มในค่าที่เกิน 10 แสดงว่าอาจมี outliers บ้าง
  - สำหรับ Sunshine ข้อมูลจำนวนมากอยู่ในช่วงกลางๆ (6-10)
  - มีความแตกต่างที่เห็นได้ชัดระหว่าง All & Yes observations
  - แสดงว่าที่ค่า Sunshine today ต่ำๆ สัมพันธ์กับ rains tomorrow

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Numeric: ล้าง Plots per Page: 4 ☐ Annotate Target: RainTomorrow

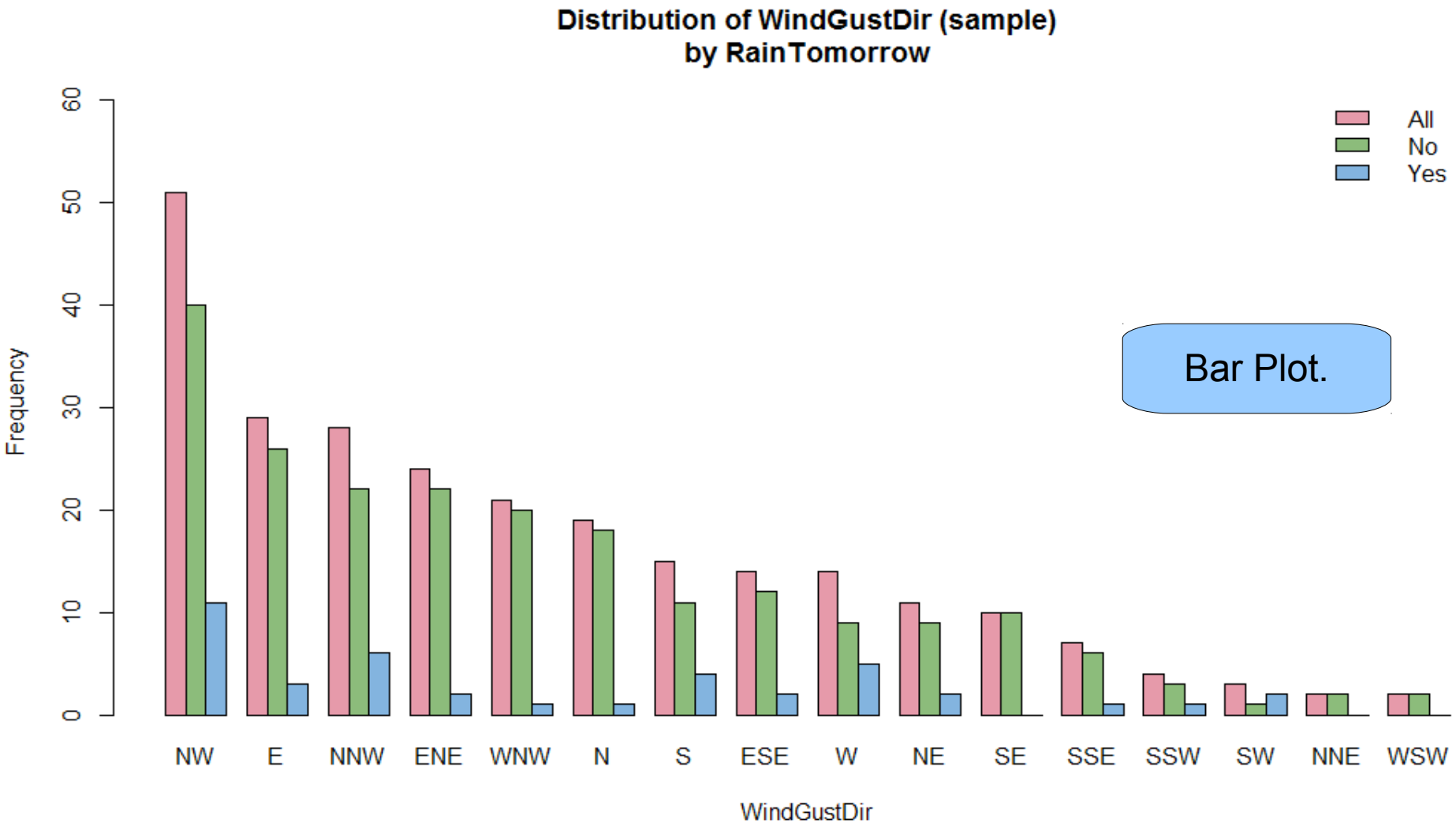
☐ Benford Bars Benford Digit: 1 ◉ abs ○ +ve ○ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
3	MinTemp	☐	☐	☐	☐	-5.30; 7.45/7.27; 20.90
4	MaxTemp	☐	☐	☐	☐	7.60; 19.65/20.55; 35.80
5	Rainfall	☐	☐	☐	☐	0.00; 0.00/1.43; 39.80
6	Evaporation	☐	☐	☐	☐	0.20; 4.20/4.52; 13.80
7	Sunshine	☐	☐	☐	☐	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	☐	☐	☐	☐	13.00; 39.00/39.84; 98.00

Categoric: ล้าง

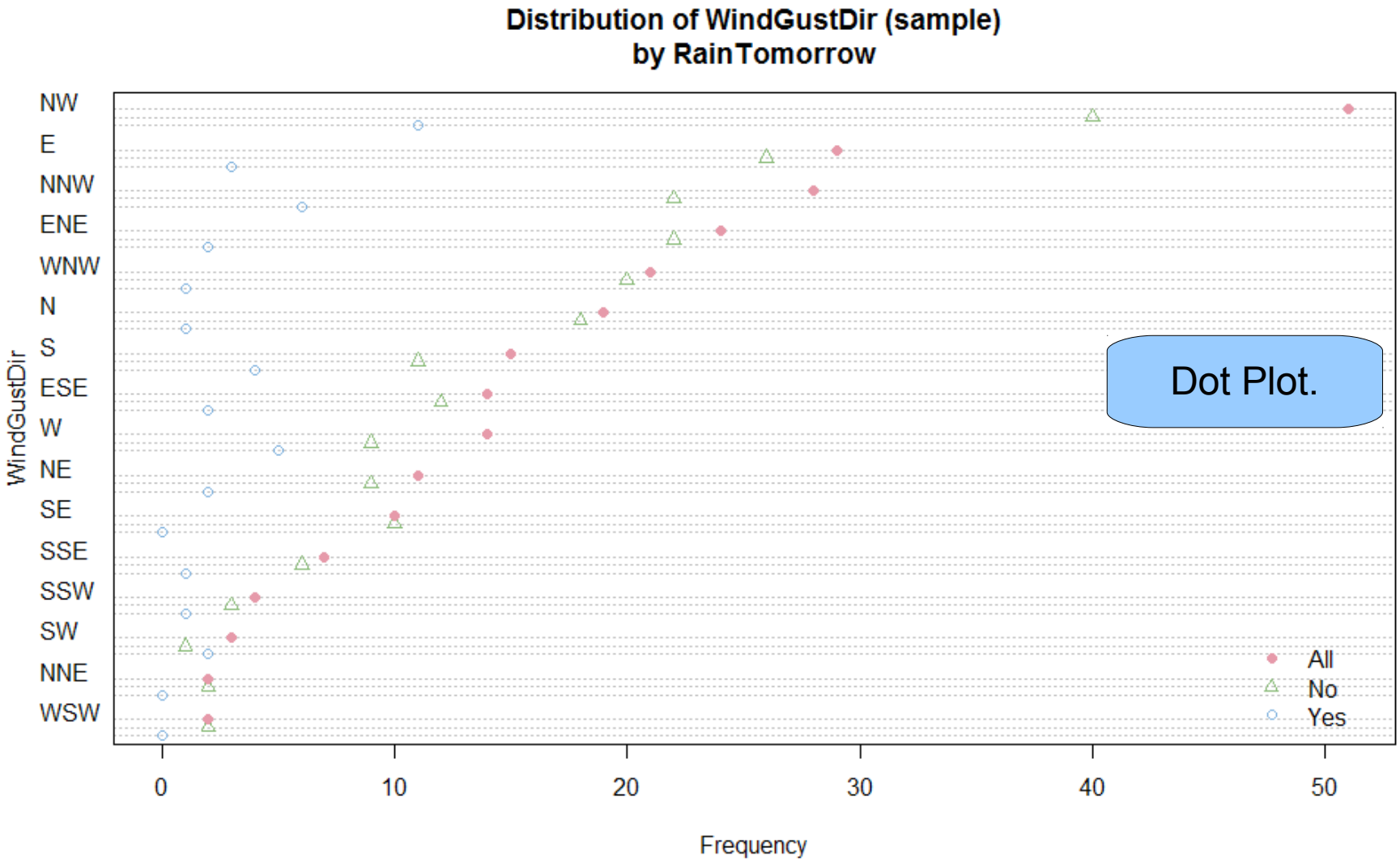
No.	Variable	Bar Plot	Dot Plot	Mosaic	Levels
8	WindGustDir	☑	☑	☑	16

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions



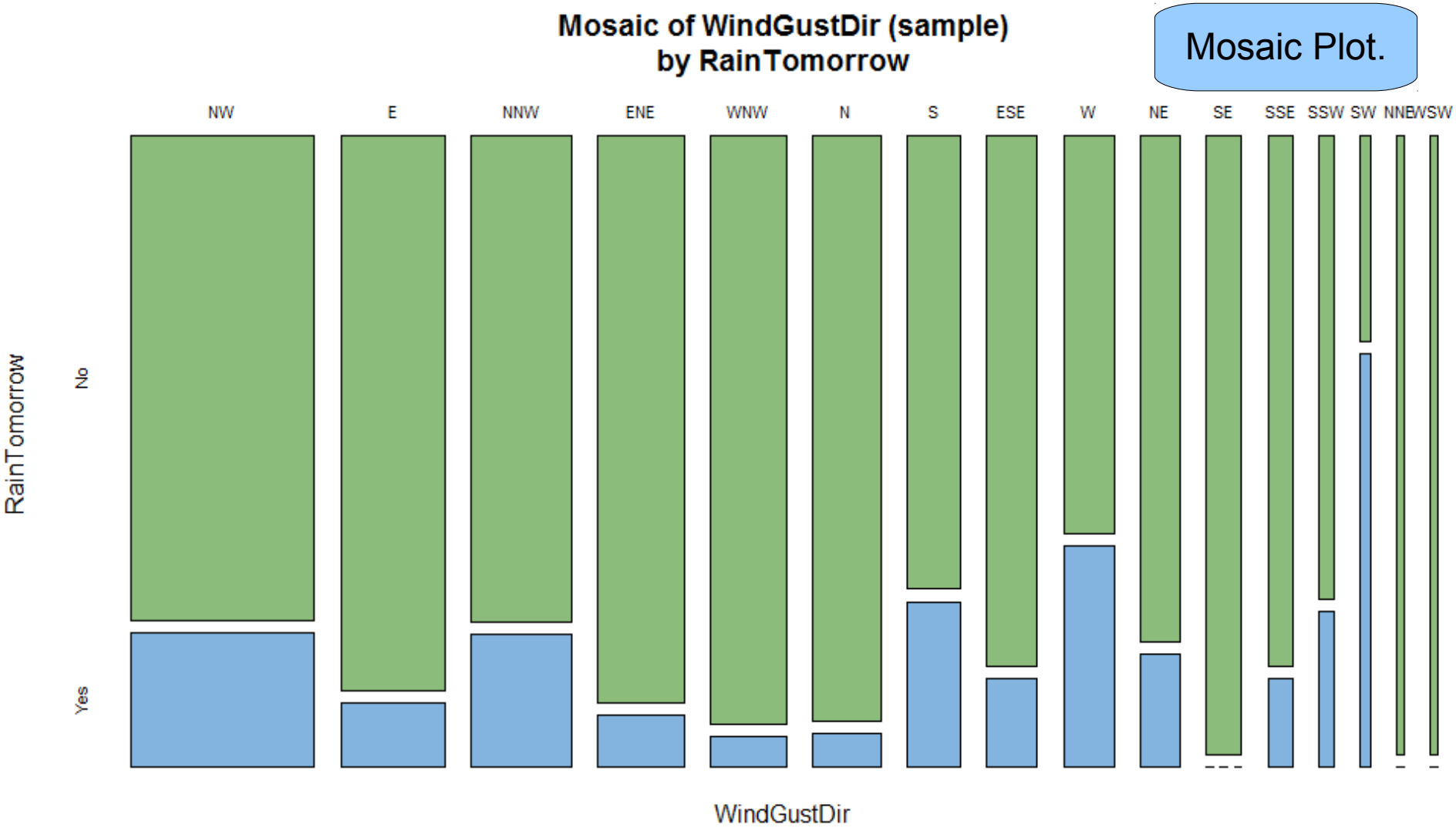
Bar Plot.

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

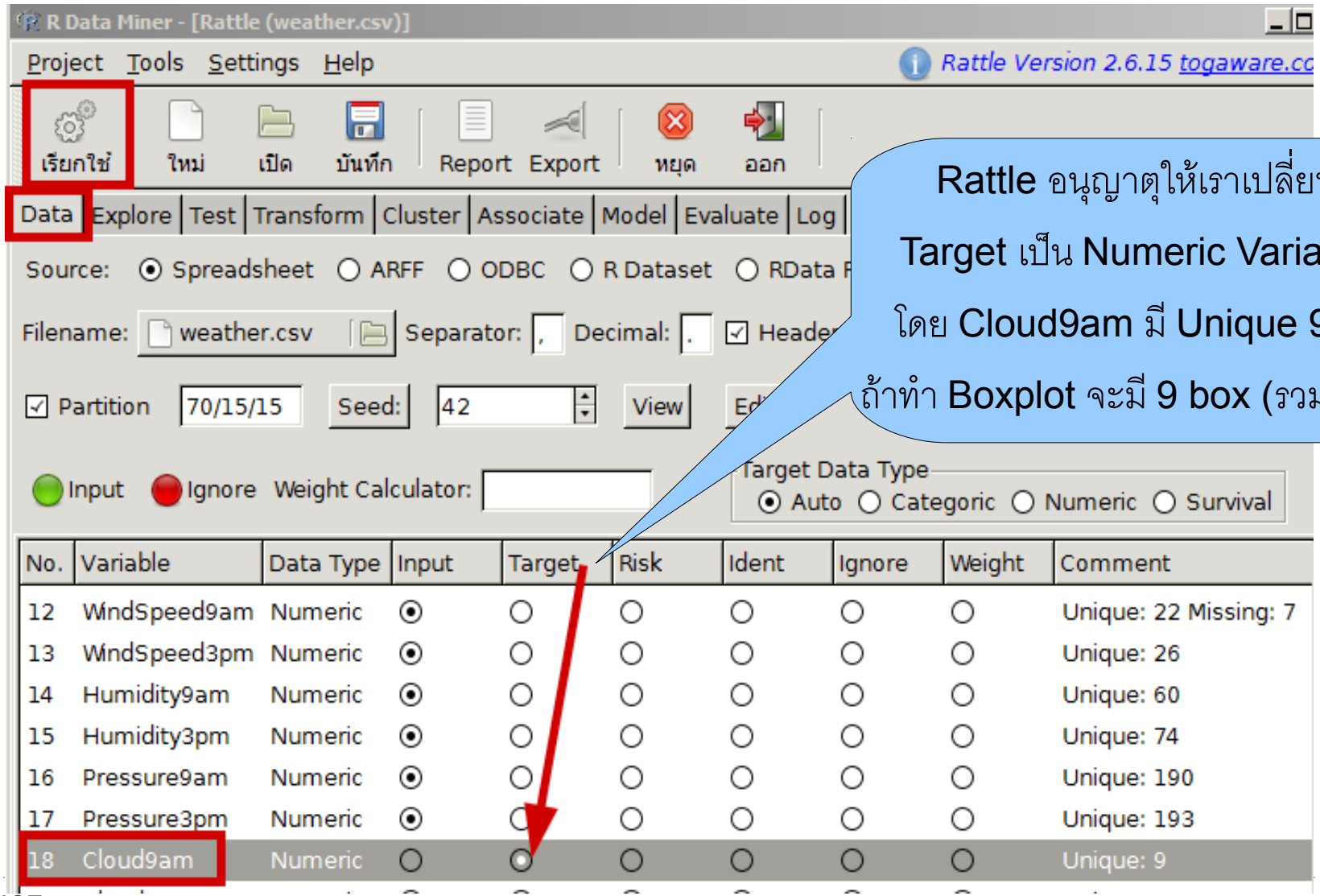


Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Mosaic Plot.

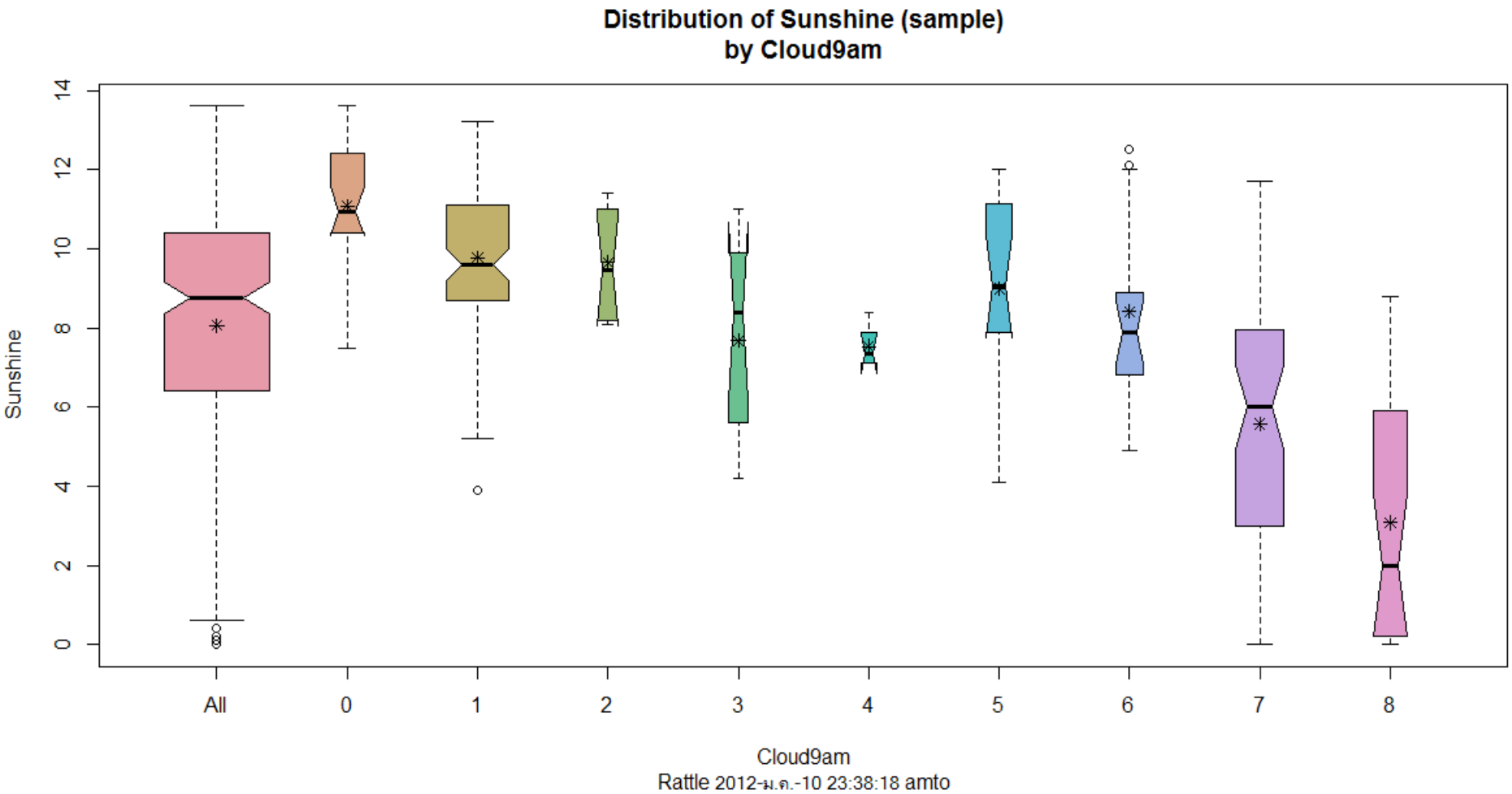


Rattle Training 4. Exploring Data 4.2 Visualizing Distributions



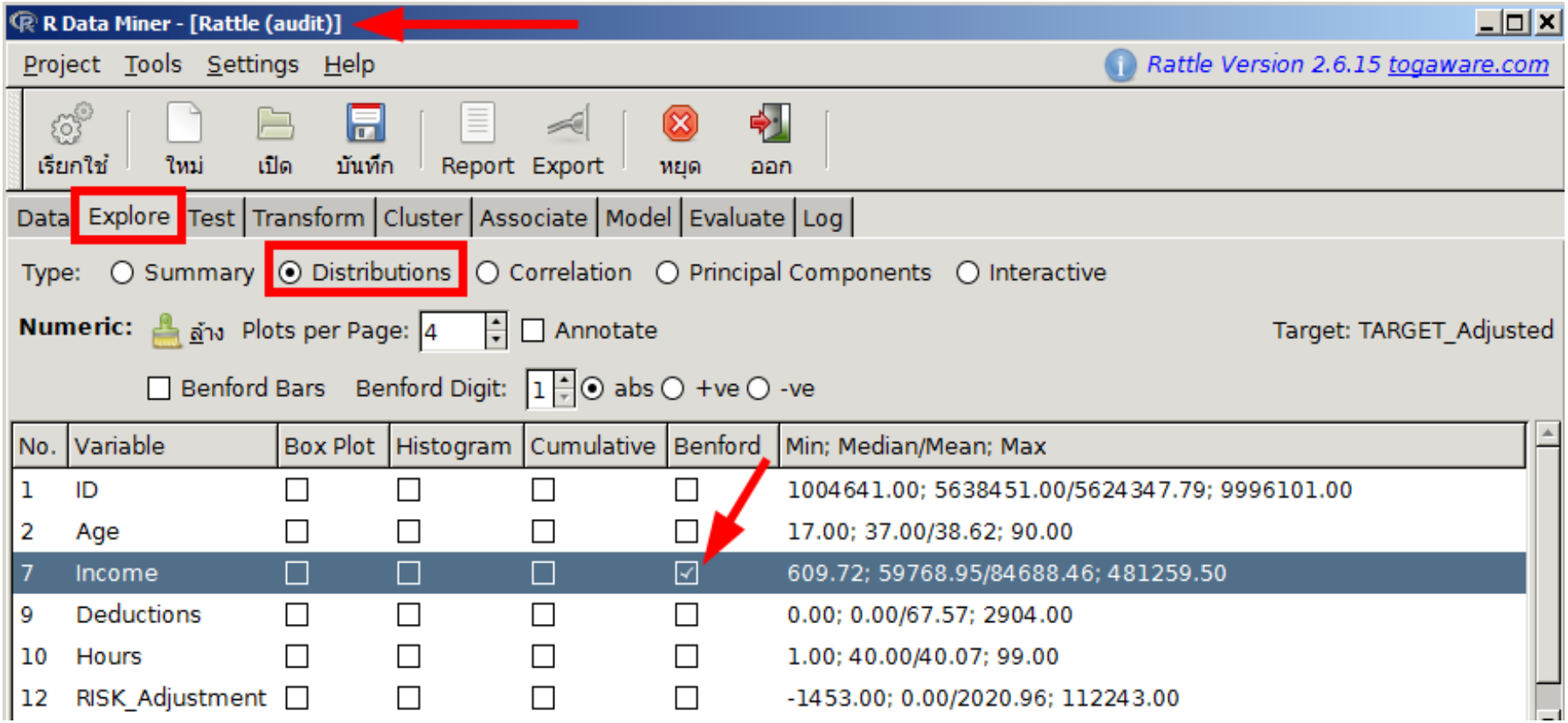
Rattle อนุญาตุให้เราเปลี่ยน Target เป็น Numeric Variable โดย Cloud9am มี Unique 9 ตัว ถ้าทำ Boxplot จะมี 9 box (รวมอีก 1)

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions



Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

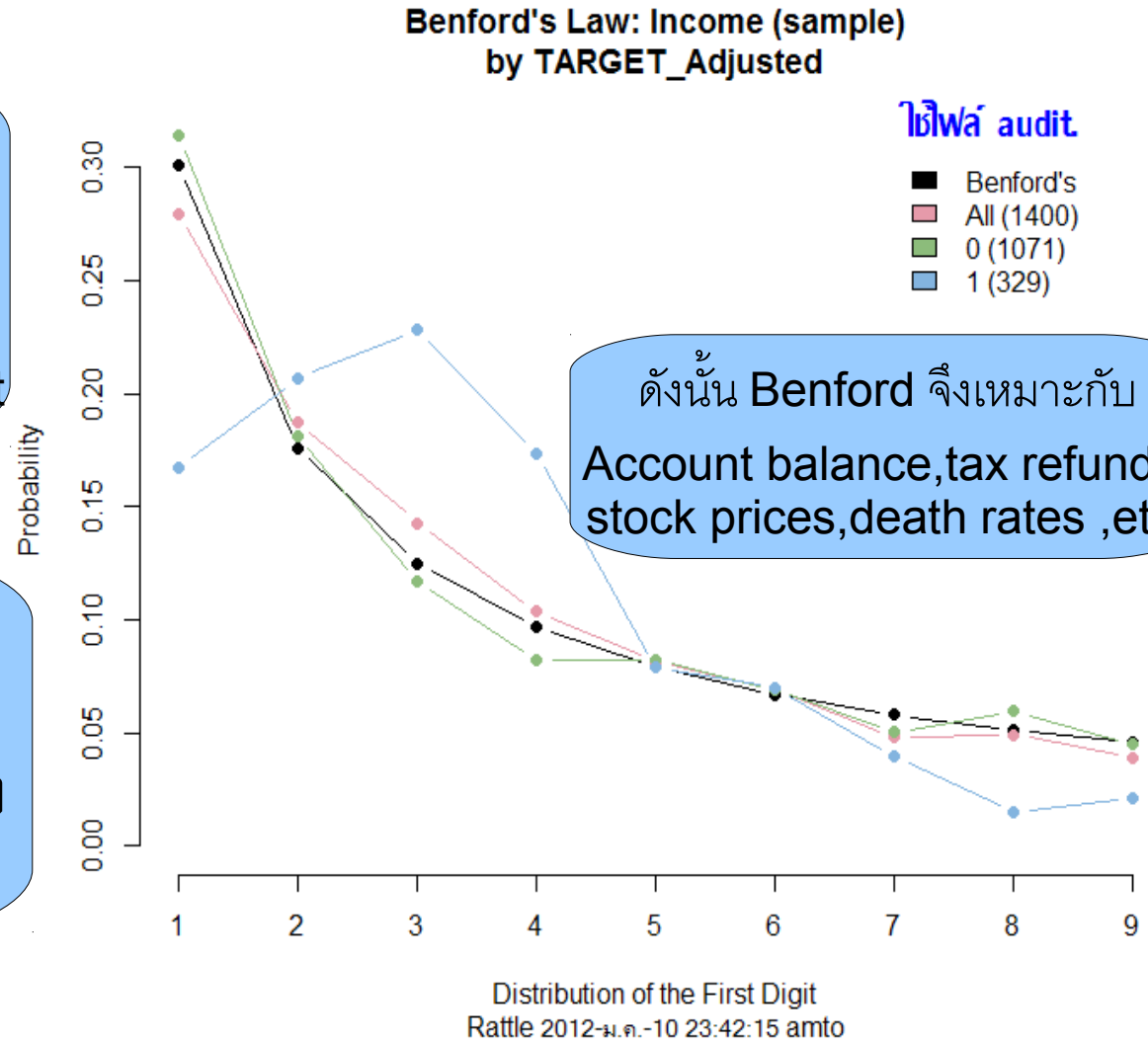
- ใช้ไฟล์ audit.RData สร้าง Benford graph กับตัวแปร Income



Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Benford's Law ใช้ในการตรวจสอบข้อมูลแปลกๆ หรือโกง หรือสิ่งน่าสนใจ โดยนำค่าตัวเลขตัวหน้า เช่น หลัก 1 10,100,1000... มา Plot

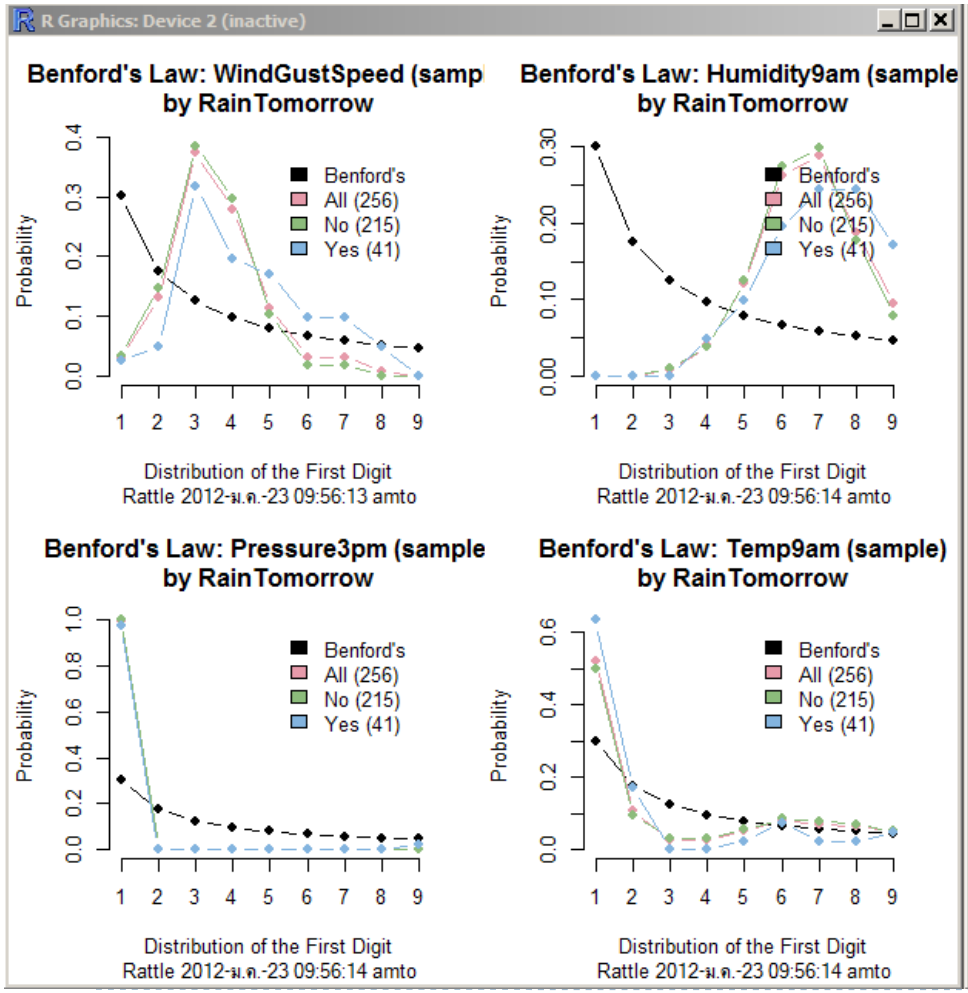
กฎนี้เริ่มจากเลข 1 มีค่า prob 30% ไล่ไปจนถึงเลข 9 มีค่า prob 5% ช่วงใดที่กราฟมีความแตกต่างไปจากกฎ แสดงว่าข้อมูลมีอยู่มากในช่วงนั้น

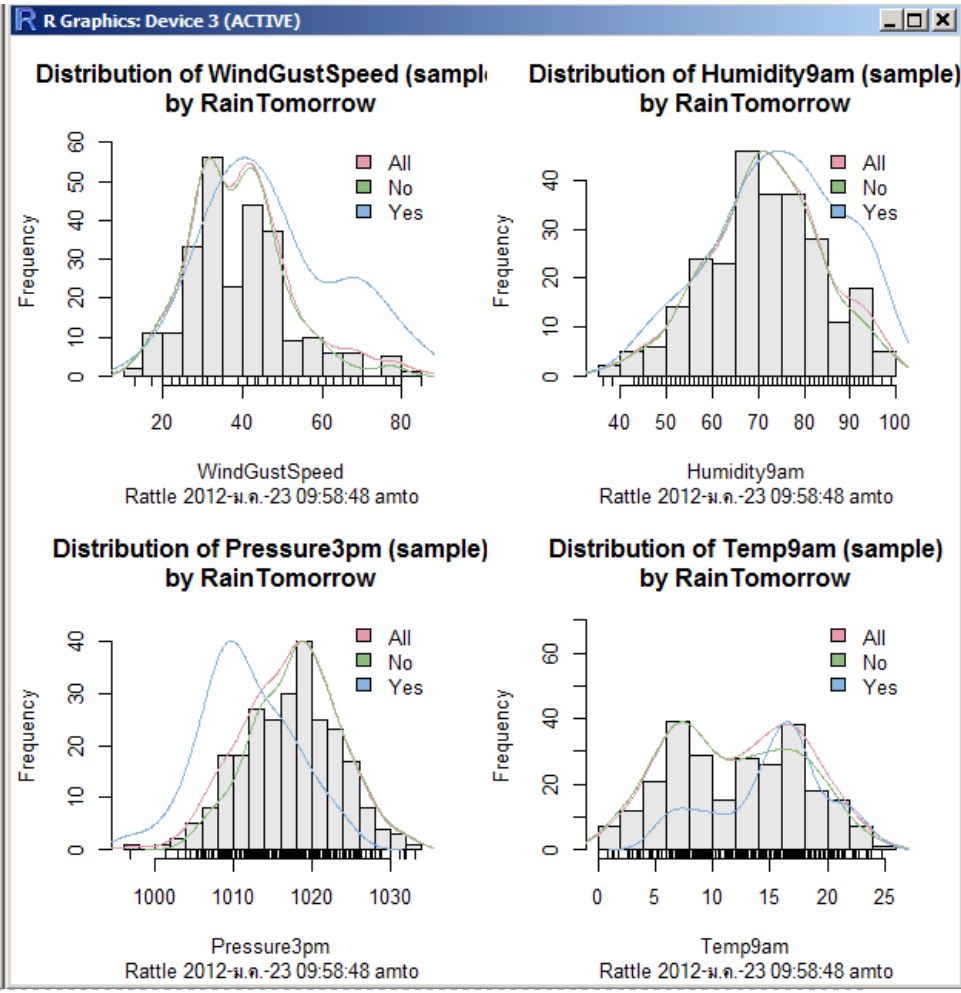


ดังนั้น Benford จึงเหมาะกับ Account balance,tax refunds, stock prices,death rates ,etc.

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

- **Benford's Law** ไม่เหมาะกับบางข้อมูล เช่น อายุ,เบอร์โทร ลองดู ตย.ข้างล่าง





Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

◉ Latticist ○ GGobi

```
Interactive Data Exploration

Latticist and GGobi are independent applications which provide highly
dynamic and interactive graphic data visualisation for exploratory
data analysis.

Latticist is based on R whilst GGobi is written in C.

Specialist tools include tours, scatterplots, barcharts and parallel
coordinates plots.

Plots allow points to be identified and linked with brushing across
multiple plots.
```

ไปที่ Interactive : Latticist

เป็นเครื่องมือพิเศษในการ

สร้างกราฟหลากหลายรูปแบบ

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Latticist: crs$dataset[crs$sample, c(...

File View Style Theme Labels Tools Data Options Help

marginal.plot(dat, data = dat, groups = RainTomorrow, reorder = FALSE, plot.points = "jitter", type = c("p", "l"), pch = "+", auto.key = list(lines = TRUE, title = Edit call...

Stay on top
Navigate
Pan
Identify
Brush
Annotate
Arrow
Panel
Plot settings

RainTomorrow
No Yes

MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir
WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am
Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am
Temp3pm, RainToday, RISK_MM, RainTomorrow

N = 256

help Latticist v. 0.9-44
marginals splom (pairs) parallel
Subset: Select variables -->
Variables on axes: reset
y=
x=
Aspect: Lines. Levels: 4
Groups / Color: explode
RainTomorrow
Depth (3D)
z=
Conditioning:
Scales:

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Latticist: crs$dataset[crs$sample, c(...

File View Style Theme Labels Tools Data Options Help

qqmath(~Sunshine, data = dat, groups = RainTomorrow, main = "Distribution of Sunshine by RainTomorrow", ylab = "Sunshine", type = c("g", "o"), distributic

Edit call...

Stay on top
Navigate
Pan
Identify
Brush
Annotate
Arrow
Plot settings

Distribution of Sunshine by RainTomorrow

RainTomorrow
No Yes

Sunshine: 0, 5, 10

Proportion ≤ y: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0

N = 254, 2012-01-11, R 2.14.1

help Latticist v. 0.9-44

marginals splom (pairs) parallel

Subset:

Variables on axes: switch reset

y= Sunshine discr.

x=

Aspect: Lines. Levels: 4

Groups / Color: explode RainTomorrow

Depth (3D)

z=

Conditioning:

Scales:

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Latticist: crs$dataset[crs$sample, c(....

File View Style Theme Labels Tools Data Options Help

xyplot(Sunshine ~ Temp9am, data = dat, groups = RainTomorrow, main = "Sunshine vs Temp9am by RainTomorrow", xlab = "Temp9am", ylab = "Sunshine" Edit call...

Stay on top
Navigate
Pan
Identify
Brush
Annotate
Arrow
Plot settings

Sunshine vs Temp9am by RainTomorrow

RainTomorrow
No Yes

Sunshine

Temp9am

N = 254, 2012-01-11, R 2.14.1

help Latticist v. 0.9-44 **Variables on axes:** switch hexbin reset Groups / Color: explode Conditioning:

marginals splom (pairs) parallel ☑ y= Sunshine discr. RainTomorrow

Subset: ☑ x= Temp9am discr. Depth (3D) or ☐ Segments (x--z)

Aspect: ☑ Lines. Levels: 4 z= Scales:

Rattle Training 4. Exploring Data 4.2 Visualizing Distributions

Latticist: crs$dataset[crs$sample, c(....

File View Style Theme Labels Tools Data Options Help

useOuterStrips(xyplot(Sunshine ~ Temp9am | equal.count(Rainfall, 4) + RainToday, data = dat, groups = RainTomorrow, main = "Sunshine vs Temp9am by R Edit call...

Stay on top
Navigate
Pan
Identify
Brush
Annotate
Arrow
Panel
Plot settings

Sunshine vs Temp9am by Rainfall and RainToday and RainTomorrow

RainTomorrow
No Yes

equal.count(Rainfall, 4) | equal.count(Rainfall, 4) | equal.count(Rainfall, 4)

Sunshine: Yes / No

Temp9am

N = 254, 2012-01-11, R 2.14.1

help Latticist v. 0.9-44
marginals splom (pairs) parallel
Subset:

Variables on axes: switch hexbin reset
y= Sunshine discr.
x= Temp9am discr.
Aspect: Lines. Levels: 4

Groups / Color: explode
RainTomorrow
Depth (3D) or Segments (x--z)
z=

Conditioning: superpose
Rainfall
RainToday
Scales: x same, y free

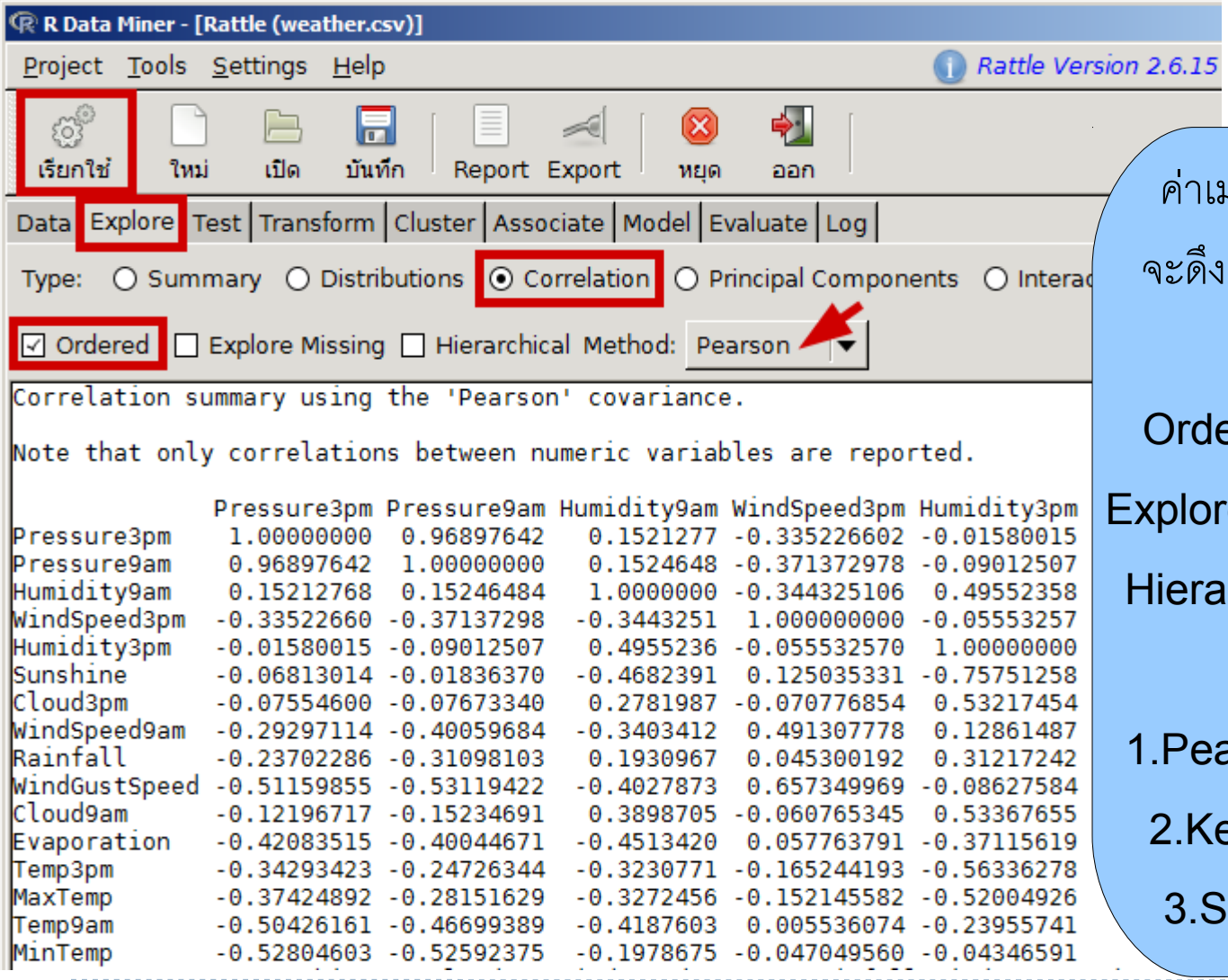
Rattle Training 4. Exploring Data

- Workshop 1: ให้สร้าง Histogram แยกค่า Skewness กับ Kurtosis ตามชื่อตัวแปรข้างล่างนี้

R Console

```
> library(Hmisc)
> skewness (weather[,c(7,9,12,13)], na.rm = TRUE)
     Sunshine WindGustSpeed  WindSpeed9am  WindSpeed3pm
   -0.7234543     0.8361055     1.3601713     0.5912721
>
> kurtosis (weather[,c(7,9,12,13)], na.rm = TRUE)
     Sunshine WindGustSpeed  WindSpeed9am  WindSpeed3pm
   -0.2706248     1.4761027     1.4758254     0.1963276
>
```

Rattle Training 4. Exploring Data 4.3 Correlation Analysis



ค่าเมตริกสหสัมพันธ์ใน Rattle จะดึงเฉพาะ numeric variable มี option ต่างๆ ดังนี้

Ordered : ให้เรียงตามค่า corr.

Explore Missing: เฉพาะ missing

Hierarchical Method: ดูคู่ลำดับ

Method มี 3 วิธี

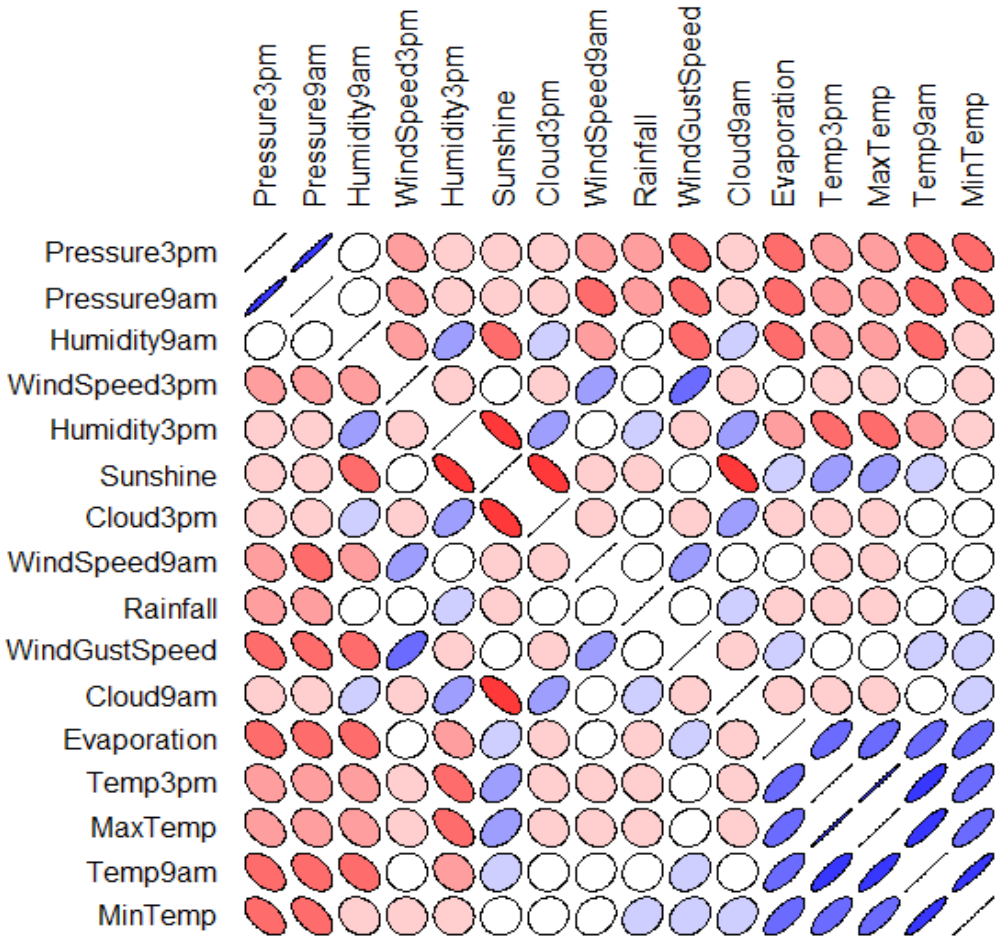
1.Pearson : ตัวเลข parametric

2.Kendall : ตัวเลข non-para

3.Spearman :ตัวเลข order

Rattle Training 4. Exploring Data 4.3 Correlation Analysis

Correlation weather.csv using Pearson



ยิ่งมีรูปร่างรี ยิ่งสัมพันธ์กัน

ยิ่งแดงเข้ม ยิ่งสัมพันธ์ทาง(−)
ยิ่งน้ำเงินเข้ม ยิ่งสัมพันธ์ทาง(+)

Analytic software today
Must shown Corr. Like this.

Rattle Training 4. Exploring Data 4.3 Correlation Analysis

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

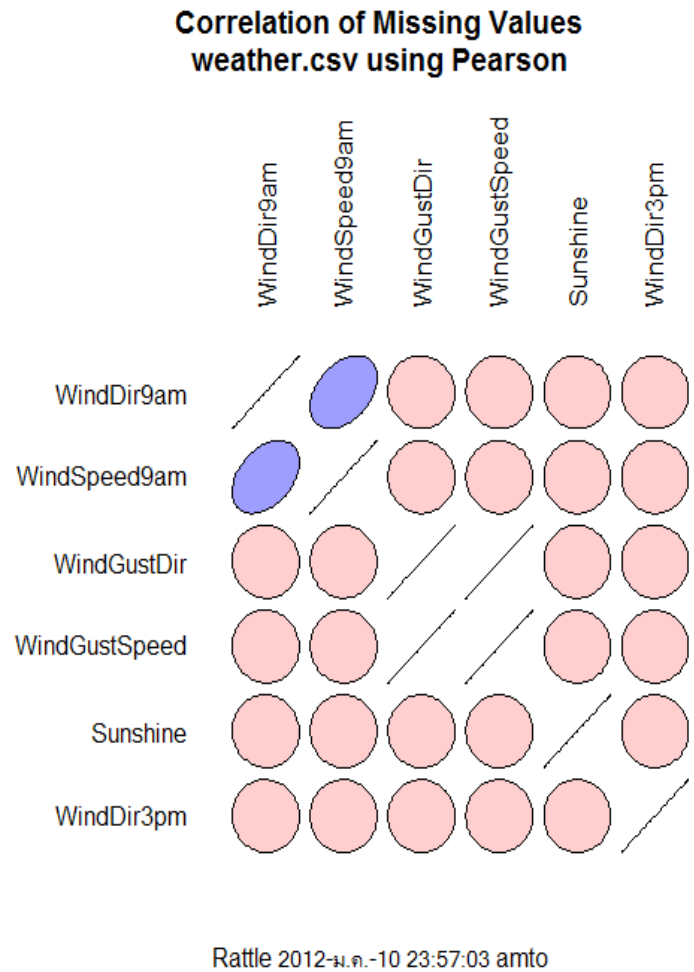
Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

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

☐ Ordered ☑ Explore Missing ☐ Hierarchical Method: Pearson

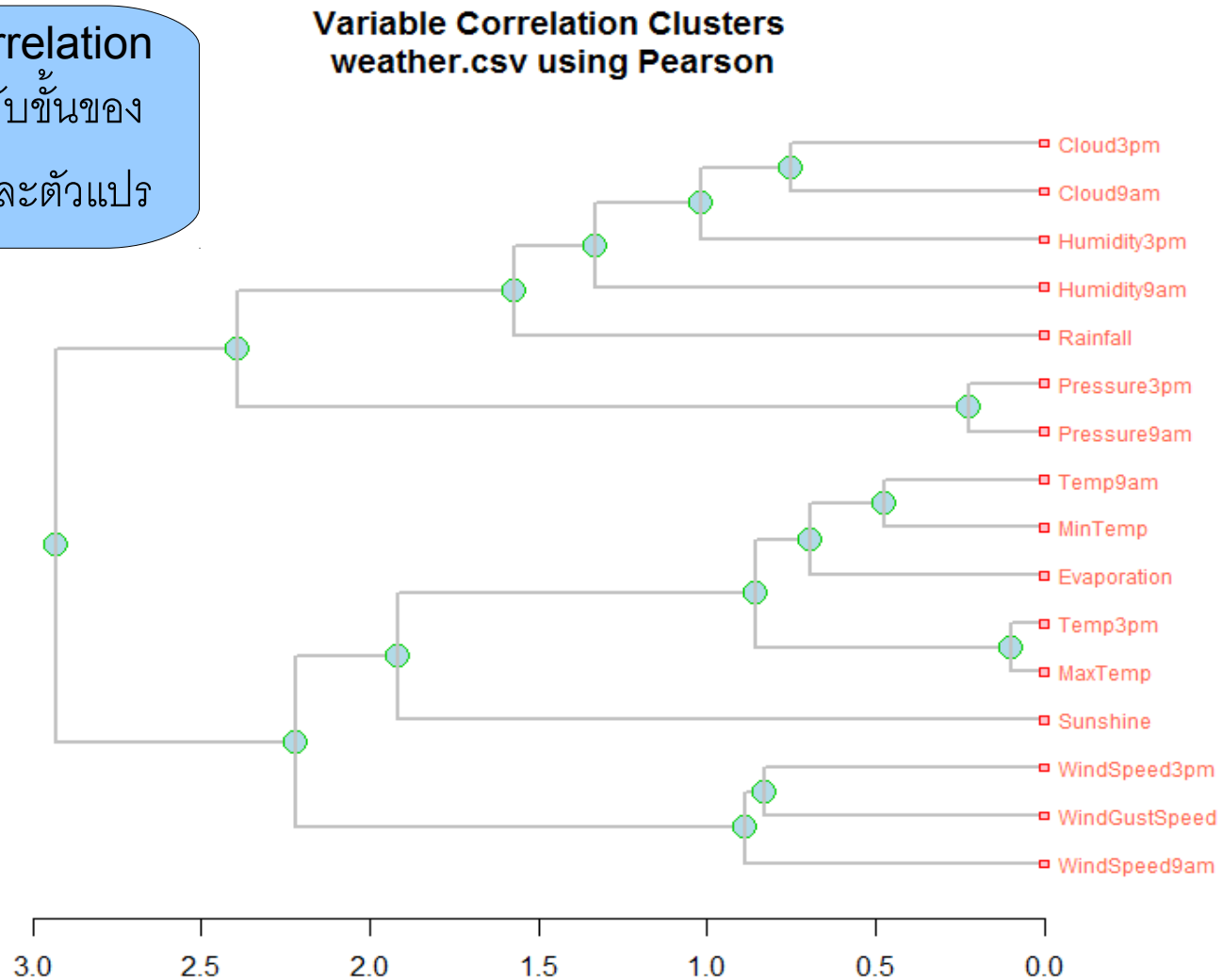
```
Missing values correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

               WindDir9am WindSpeed9am  WindGustDir WindGustSpeed     Sunshine
WindDir9am     1.00000000  0.400999442 -0.027879443  -0.027879443 -0.027879443
WindSpeed9am   0.40099944  1.000000000 -0.011179641  -0.011179641 -0.011179641
WindGustDir   -0.02787944 -0.011179641  1.000000000   1.000000000 -0.007874016
WindGustSpeed -0.02787944 -0.011179641  1.000000000   1.000000000 -0.007874016
Sunshine      -0.02787944 -0.011179641 -0.007874016  -0.007874016  1.000000000
WindDir3pm    -0.01967505 -0.007889684 -0.005556842  -0.005556842 -0.005556842
                WindDir3pm
WindDir9am    -0.019675051
WindSpeed9am  -0.007889684
WindGustDir   -0.005556842
WindGustSpeed -0.005556842
Sunshine      -0.005556842
WindDir3pm     1.000000000

Count of missing values:
   WindDir9am  WindSpeed9am   WindGustDir WindGustSpeed      Sunshine
           23             4             2             2             2
   WindDir3pm
            1
```

Explore Missing:
จะเลือกเฉพาะตัวแปรที่มีค่า
Missing มาหาค่าสหสัมพันธ์



Rattle Training 4. Exploring Data 4.3 Correlation Analysis

Hierarchical Correlation
จะแสดงให้เห็นลำดับขั้นของ
ความสัมพันธ์ในแต่ละตัวแปร



Rattle Training 5.Testing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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: MinTemp Sample 2: ☑ Group By Target: RainTomorrow

```
Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a non-parametric test of the
similarity of two distributions. The null hypothesis is that the
two samples are drawn from the same distribution. The two-sided and
the two one-sided tests are performed.

The STATISTIC calculated is the so called D statistic.
For similar distributions the statistic converges to zero.

If the p-value is less than 0.05 then we reject the null hypothesis and
accept the alternative hypothesis, that the distributions differ, at the 95% level of confidence.

The two samples being compared come from the 'MinTemp' variable, grouped by
 'RainTomorrow', with values 'No' and 'Yes'
Title:
 Kolmogorov-Smirnov Two Sample Test

Test Results:
  STATISTIC:
    D | Two Sided: 0.3012
       D^- | Less: 0
    D^+ | Greater: 0.3012
  P VALUE:
    Alternative        Two-Sided: 0.0001091
    Alternative Exact Two-Sided: 0.0001091
    Alternative             Less: 1
    Alternative          Greater: 0.00005454
```

Kolmogorov-Smirnov Test เป็น non-parametric test

ใช้ทดสอบความเหมือนของ 2 การกระจายตัว (target)

ถ้า p-value ต่ำกว่า 0.05 แสดงว่าแตกต่างกัน

Rattle Training 5.Testing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.c

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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: MinTemp Sample 2: ☑ Group By Target: RainTomorrow

```
Wilcoxon Rank Sum Test

The two-sample non-parametric Wilcoxon rank sum test (equivalent to
the Mann-Whitney test) is performed on the two specified samples. The null
hypothesis is that the distributions are the same (i.e., there is no
shift in the location of the two distributions) with an alternative
hypothesis that they differ on location (based on median).

This test does not assume that the two samples are normally distributed
but does assume they have distributions of the same shape.

If the p-value is less than 0.05 then we reject the null hypothesis and
accept the alternative hypothesis, that the two samples have different medians,
at the 95% level of confidence.

The two samples being compared come from the 'MinTemp' variable, grouped by
 'RainTomorrow', with values 'No' and 'Yes'
        Wilcoxon rank sum test with continuity correction

data:  na.omit(crs$dataset[crs$dataset[["RainTomorrow"]] == "No", "MinTemp"]) and na.omit(crs$dataset[crs$dat
W = 6502.5, p-value = 0.00001268
alternative hypothesis: true location shift is not equal to 0
```

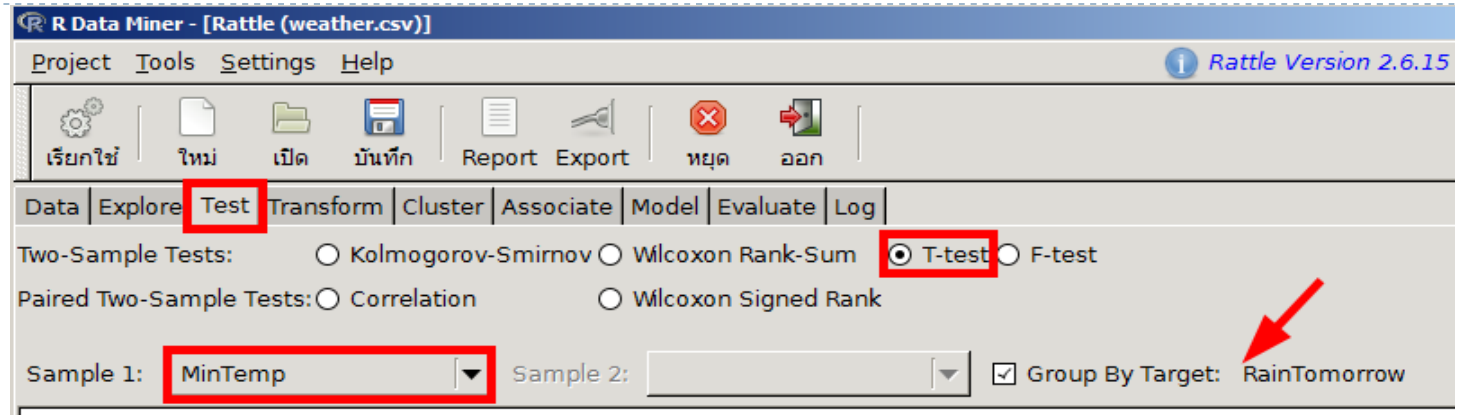
Wilcoxon Rank-Sum เป็นอีกวิธี

ที่ใช้ตรวจสอบการกระจายตัว

2 sample by target

p-value < 0.05 (diff)

Rattle Training 5.Testing Data




```
Test Results:
  PARAMETER:
    x Observations: 300
    y Observations: 66
    mu: 0
  SAMPLE ESTIMATES:
    Mean of x: 6.6073
    Mean of y: 10.2576
    Var  of x: 33.6066
    Var  of y: 38.216
  STATISTIC:
                T: -4.391
    T | Equal Var: -4.5756
  P VALUE:
    Alternative Two-Sided: 0.00003015
    Alternative      Less: 0.00001507
    Alternative   Greater: 1
    Alternative Two-Sided | Equal Var: 0.000006523
    Alternative      Less | Equal Var: 0.000003262
    Alternative   Greater | Equal Var: 1
  CONFIDENCE INTERVAL:
    Two-Sided: -5.3013, -1.9992
         Less: -Inf, -2.2689
      Greater: -5.0315, Inf
    Two-Sided | Equal Var: -5.219, -2.0814
         Less | Equal Var: -Inf, -2.3347
      Greater | Equal Var: -4.9658, Inf
```

T-test .ใช้ทดสอบความแตกต่างของค่า Mean.

2 กลุ่ม (Group by Target)

p-value < 0.05 (diff)

Rattle Training 5.Testing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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: MinTemp Sample 2: ☑ Group By Target: RainTomorrow

```
variances, at the 95% level of confidence.

The two samples being compared come from the 'MinTemp' variable, grouped by
 'RainTomorrow', with values 'No' and 'Yes'
Title:
 F Test of Variances

Test Results:
  PARAMETER:
    Hypothesized Ratio: 1
    Numerator   df: 299
    Denumerator df: 65
  SAMPLE ESTIMATES:
    Ratio of Variances: 0.8794
  STATISTIC:
    F: 0.8794
  P VALUE:
    Alternative Two-Sided: 0.4755
    Alternative      Less: 0.2378
    Alternative   Greater: 0.7622
  CONFIDENCE INTERVAL:
    Two-Sided: 0.5862, 1.2583
         Less: 0, 1.1885
      Greater: 0.6264, Inf
```

F-test .ใช้ทดสอบความแตกต่างของ Variances.

2 กลุ่ม (Group by Target)

p-value < 0.05 (diff)

Rattle Training 5.Testing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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: MinTemp Sample 2: Rainfall ☑ Group By Target: RainTomorrow

```
Pearson's product moment correlation coefficient is used.

If the p-value is less than 0.05 then we reject the null hypothesis and
accept the alternative hypothesis that the samples are correlated,
at the 95% level of confidence.

The two samples being compared are the two variables, 'MinTemp' and 'Rainfall'
Title:
 Pearson's Correlation Test

Test Results:
  PARAMETER:
    Degrees of Freedom: 364
  SAMPLE ESTIMATES:
    Correlation: 0.2019
  STATISTIC:
    t: 3.9338
  P VALUE:
    Alternative Two-Sided: 0.0001001
    Alternative      Less: 0.9999
    Alternative   Greater: 0.00005007
  CONFIDENCE INTERVAL:
    Two-Sided: 0.1015, 0.2983
         Less: -1, 0.2831
      Greater: 0.1179, 1
```

Paired Two-Sample Tests: Correlation

ให้เลือกคู่ตัวแปรที่ต้องการทดสอบความสัมพันธ์

ตัวเลือกนี้ใช้ pearson ดังนั้นจึงเป็น parametric

p-value < 0.05 (Cor)

Rattle Training 5.Testing Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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: MinTemp Sample 2: Rainfall ☑ Group By Target: RainTomorrow

```
Wilcoxon Signed Rank Test

The paired sample non-parametric Wilcoxon signed rank test is
performed on the two specified samples. The two samples are expected to be
paired (two observations for the same entity). The null
the distributions are the same.

This test does not assume that the two samples are are

If the p-value is less than 0.05 then we reject the nu
accept the alternative hypothesis, that the distributi
95% level of confidence.

The two samples being compared are the two variables, 'Min
        Wilcoxon signed rank test with continuity correction

data:  na.omit(crs$dataset[, "MinTemp"]) and na.omit(crs$dataset[, "Rainfall"])
V = 59889.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon Signed Rank: Correlation

ให้เลือกคู่ตัวแปรที่ต้องการทดสอบความสัมพันธ์

ตัวเลือกนี้เป็น non-parametric rank test.

p-value < 0.05 (Cor)

Data Preprocessing (6.Transforms data)

- Data preprocessing
- Data cleaning
- Data integration
- Data transformation
- Data reduction

Why Data Preprocessing?

- Data in the real world is dirty
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation = " "
  - Noisy: containing errors or outliers
    - e.g., salary = "-10 "
  - Inconsistent: containing discrepancies in codes or names
    - e.g., Age = "42" Birthday = "03/07/1997"
    - e.g., Was rating "1,2,3" , now rating "A, B , C"
    - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/ hardware/ software problems.
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data , no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
    - Data warehouse needs consistent integration of quality data
  - Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Major tasks in data preprocessing.

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, of files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

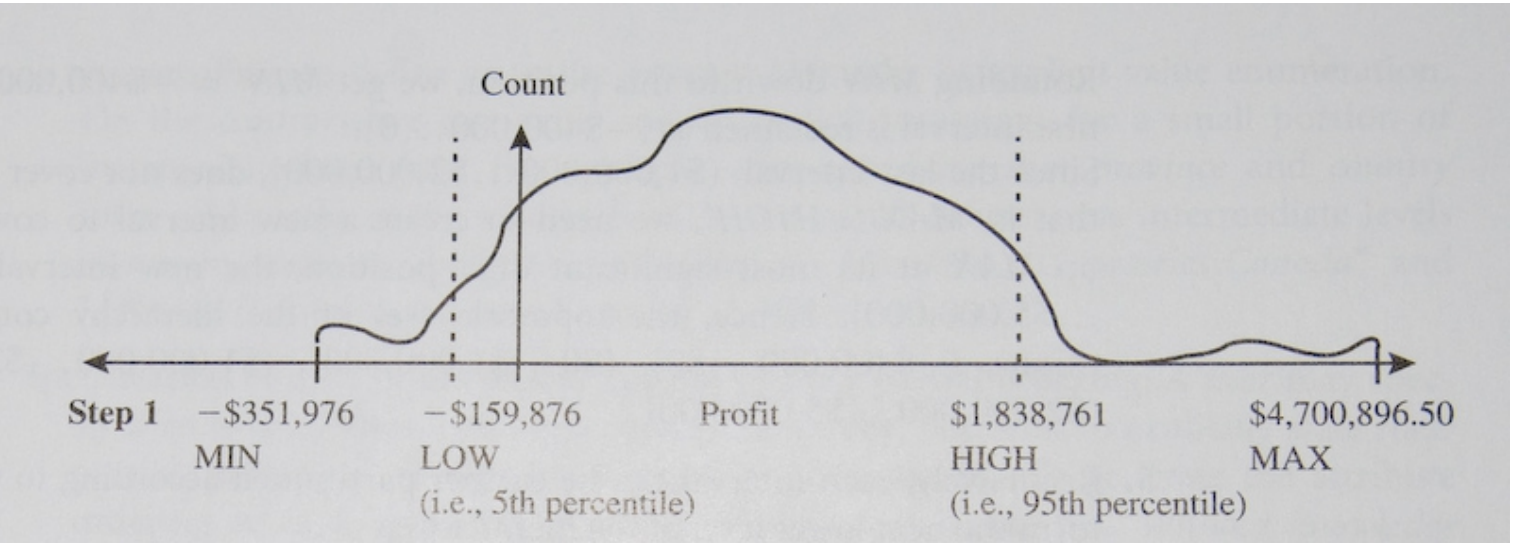
ผลที่ได้จาก Postprocessing กับ PreProcessing จะแตกต่างกันอย่างสิ้นเชิง

เทคนิค การแปลงลักษณะประจำที่เป็นตัวเลข

- การแปลงค่าที่อยู่ในลักษณะประจำ ให้อยู่ในสเกลมาตรฐาน
  - ใช้วิธี max-min normalization
  - ใช้วิธี z-score normalization
  - ใช้วิธี decimal scaling
- การปรับค่า ไม่ให้เป็นค่าลบหรือศูนย์ เพื่อสามารถนำไปเป็นตัวหารได้
  - ใช้วิธีบวก 1 กับตัวแปร binary
  - ใช้วิธีบวก 1-min {A} กับตัวแปรที่มีค่าลบ แต่ต้องการให้มีค่าเป็นบวก
- การคูณด้วยค่าถ่วงน้ำหนักที่เหมาะสม ใช้เมื่อต้องการให้ลักษณะประจำใหม่มีความสำคัญกว่าลักษณะประจำอื่น

เทคนิค การแปลงจำนวนให้เป็นค่าไม่ต่อเนื่อง

- การใช้กล่องชุดข้อมูล (Binning)
- การใช้ Histogram (เป็นกรรมวิธีทางสถิติ)
- การใช้ผลการจัดกลุ่มจากการวิเคราะห์ clustering
- การใช้ Entropy-based discretization (ใช้หลักการทาง Physic)
- การใช้ผลแบ่งกั้นที่ง่ายต่อความเข้าใจ (Natural partitioning)



ข้อมูลที่ผ่าน Data cleaning คือ

- ประกอบด้วยสดมภ์และแถว
  - สดมภ์ คือ ลักษณะประจำ = ตัวแปรในทางสถิติ ซึ่งมีทั้งตัวแปรเชิงกลุ่ม และตัวแปรที่เป็นตัวเลข
  - แถว คือ **records** ที่เก็บข้อมูลที่กำหนดค่าในแต่ละลักษณะประจำ
- ไม่มี **missing value** หรือข้อมูลที่ผิดปกติ **(Outlier)**
- ในการทำ **data mining** ชนิดของลักษณะประจำและระเบียนต้องถูกปรับเปลี่ยนเพื่อความเหมาะสมสำหรับแต่ละเทคนิคการทำ **data mining**

  (ดูรายละเอียดในหัวข้อ **associate rule, classification & clustering.**)

ชนิดของลักษณะประจำ หรือตัวแปร (variables)

- ตัวแปรเชิงกลุ่ม (nominal) คือ ตัวแปรที่ไม่มีลำดับ ไม่มีค่าต่อเนื่องในแต่ละระเบียน เช่น เพศ เชื้อชาติ อาชีพ ฯลฯ
- ตัวแปรลำดับ (ordinal) คือ ตัวแปรที่มีลำดับที่มีค่าไม่ต่อเนื่อง แต่เป็นลักษณะประจำที่บ่งบอกถึงลำดับ เช่น กลุ่มอายุ {เด็ก,ผู้ใหญ่,คนชรา} การศึกษา {ประถม,มัธยม,ป.ตรี}
- ตัวแปรตัวเลข Numerical (Continuous/Interval) เป็นลักษณะประจำที่มีค่าต่อเนื่อง เช่น อายุ น้ำหนัก ส่วนสูง ระดับน้ำตาลในเลือด ราคา ฯลฯ

Data mining vs. variables

- กรรมวิธีทาง **data mining** โจทย์ปัญหา วัตถุประสงค์และความรู้ที่ต้องการ ที่หลากหลาย ทำให้เราจำเป็นต้องใช้รูปแบบของข้อมูลที่แตกต่างกัน
- ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลแบบทำนาย **(Predicted)** แบ่งได้เป็น
  - กฎเชื่อมโยง **(Association rule)**
  - การจัดจำแนกประเภท **(Classifier)**
  - การวิเคราะห์การเกาะกลุ่ม **(Cluster)**
- ลักษณะของข้อมูลที่ต้องการจะมีความแตกต่างไปตามขั้นตอนวิธีและผลลัพธ์

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Rescale Impute Recode Cleanup

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

Order: Rank Interval Number of Groups: 100

No.	Variable	Data Type and Number Missing
9	WindGustSpeed	Numeric [13 to 98; unique=35; mean=39; median=39; miss=2].
10	WindDir9am	Categorical [16 levels; miss=31].
11	WindDir3pm	Categorical [16 levels; miss=1].
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
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].
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].

บางเทคนิค data mining เช่น Cluster ใช้การสร้าง model ตามระยะห่างข้อมูล ตัวแปร Age, Income อาจมี scale ห่าง ต้องทำ Normalize ให้มี scale เดียวกัน เช่น ***Recenter*** ไปที่ค่า 0 ผู้ใช้ทั่วไปเรียกวิธีนี้ว่า Z score.

Rattle Training 6.Transforming Data

19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
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.

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

Source: (•) Spreadsheet ( ) ARFF ( ) ODBC ( ) R Dataset ( ) RData File ( ) Library ( ) Corpus ( ) Script

Filename: weather.csv Separator: , Decimal: . [x] Header

[x] Partition 70/15/15 Seed: 42 View Edit

RRC คือ Recenter transformed

สามารถดูข้อมูลได้ที่ tab Data--> View

Rattle Dataset

Rattle Dataset	Pressure	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToda	RISK_MM	RainTomo	RRC_Temp3pm
1	1015	7	7	14.4	23.6	No	3.6	Yes	0.657967
2	1008.4	5	3	17.5	25.7	Yes	3.6	Yes	0.974215
3	1007.2	8	7	15.4	20.2	Yes	39.8	Yes	0.145945
4	1007	2	7	13.5	[illegible]	Yes	2.8	Yes	[illegible]

Rattle Training 6.Transforming Data

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

Type: (•) Rescale ( ) Impute ( ) Recode ( ) Cleanup

Normalize: ( ) Recenter (•) Scale [0-1] ( ) -Median/MAD ( ) Natural Log ( ) Matrix

Order: ( ) Rank ( ) Interval Number of Groups: 100

R01 คือ Normalize : Scale [0-1]

No.	Variable	Data Type and Number Missing
11	WindDir3pm	Categorical [16 levels; miss=1].
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17; ignored].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
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_WindSpeed3pm	Numeric [0.00 to 1.00; unique=26; mean=0.35; median=0.33]. No code export.

Normalized variables added to the dataset with 'R01' prefix.

Rattle Training 6.Transforming Data

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

Type: (•) Rescale ( ) Impute ( ) Recode ( ) Cleanup

Normalize: ( ) Recenter ( ) Scale [0-1] (•) -Median/MAD ( ) Natural Log ( )

Order: ( ) Rank ( ) Interval Number of Groups: 100

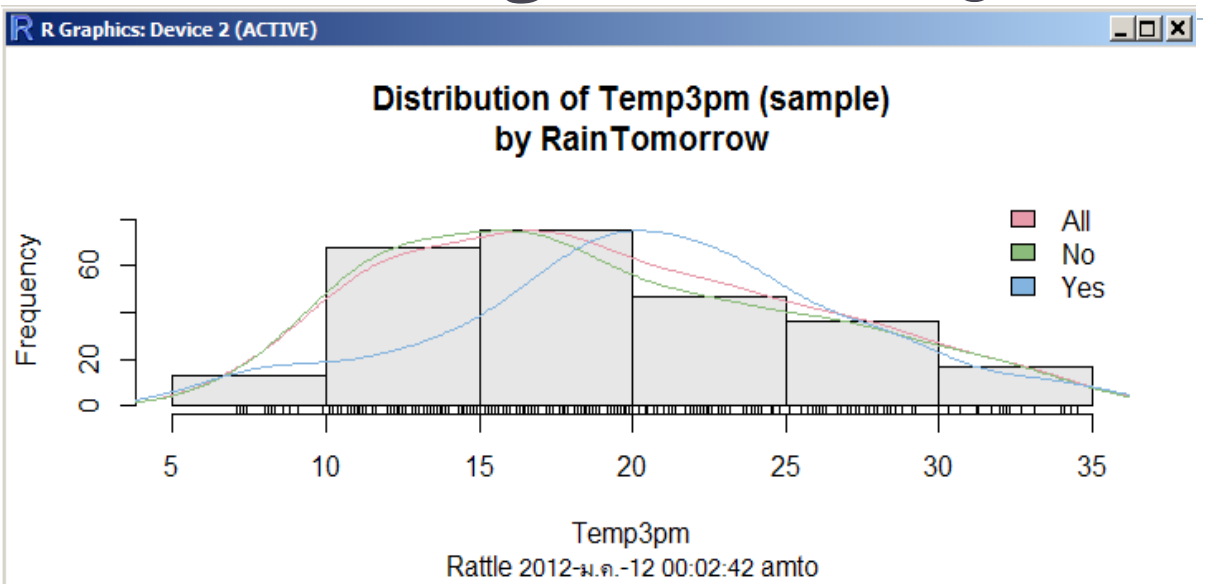
RMD คือ Normalize : Median/MAD

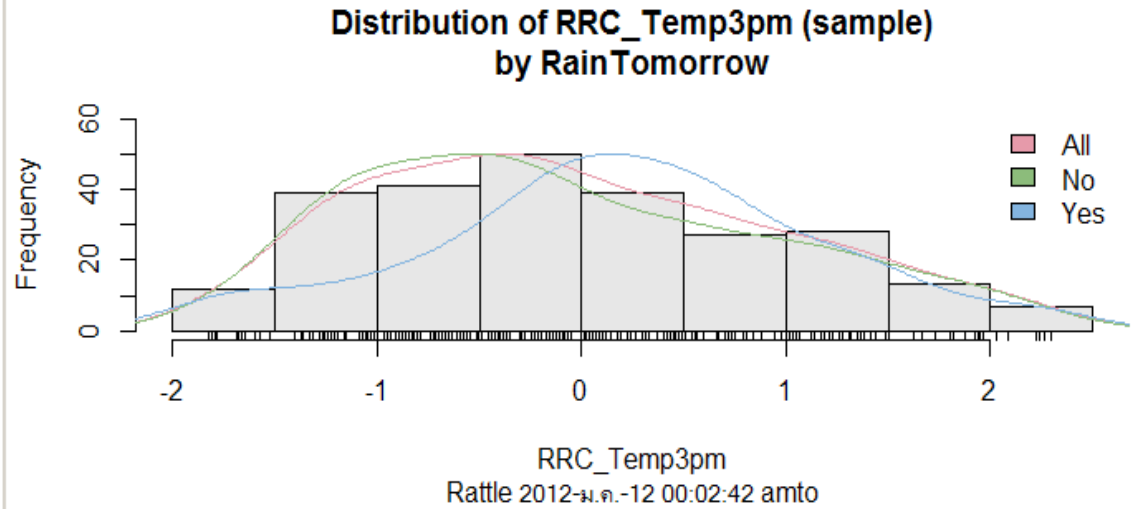
ข้อมูลในตัวแปรนั้นจะมีค่า Median = 0

No.	Variable	Data Type and Number Missing
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17; ignored].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
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_WindSpeed3pm	Numeric [0.00 to 1.00; unique=26; mean=0.35; median=0.33]. No code export.
27	RMD_Temp3pm	Numeric [-1.87 to 2.22; unique=200; mean=0.09; median=0.00]. No code export.

Normalized variables added to the dataset with 'RMD' prefix.

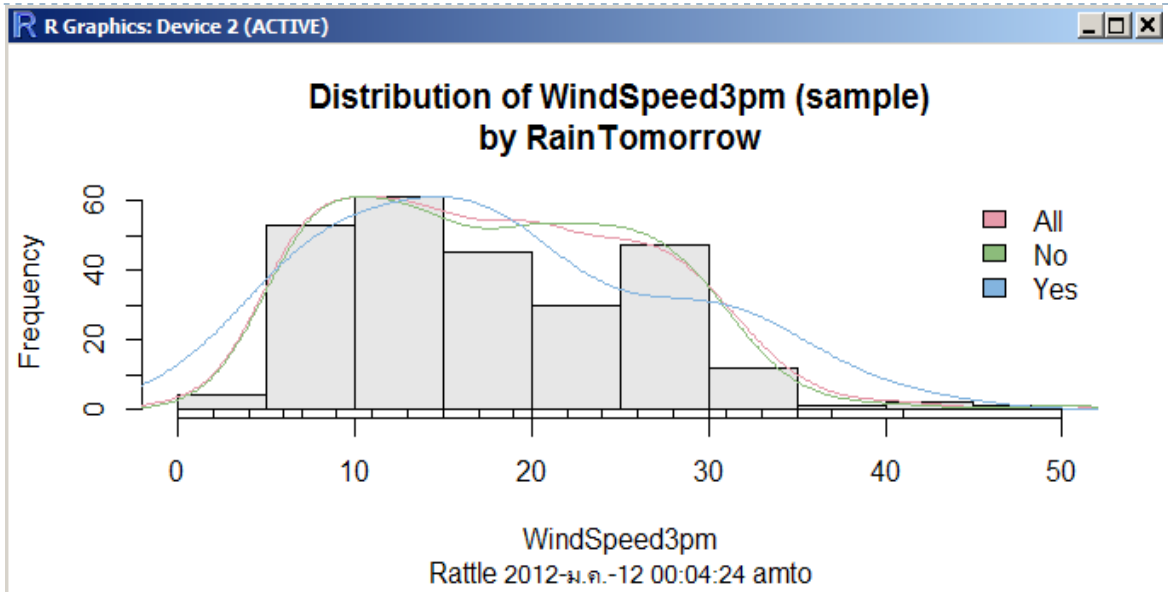
Rattle Training 6.Transforming Data

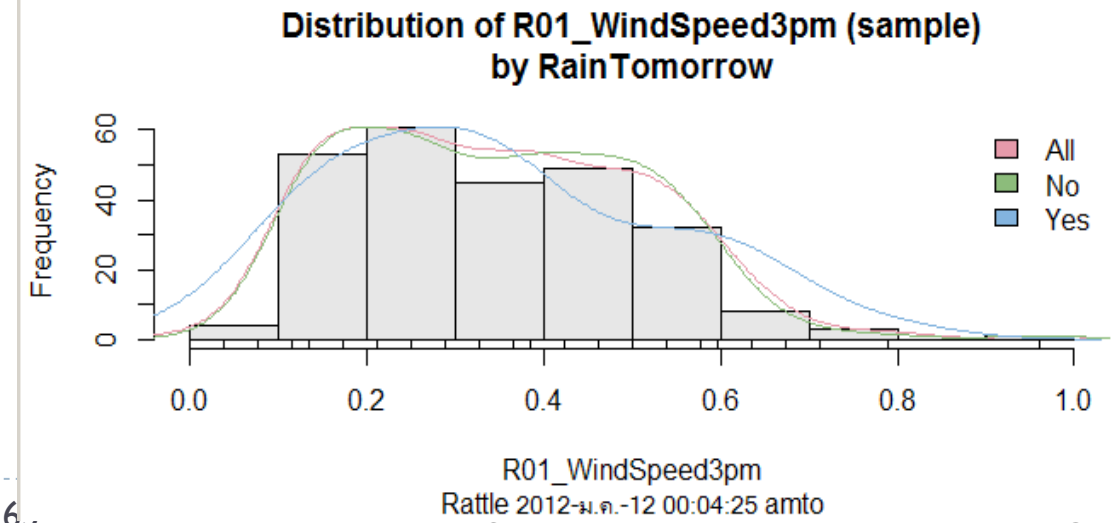




ทดลองสร้าง Histogram ระหว่างข้อมูลตัวแปร Temp3pm กับ RRC_Temp3pm

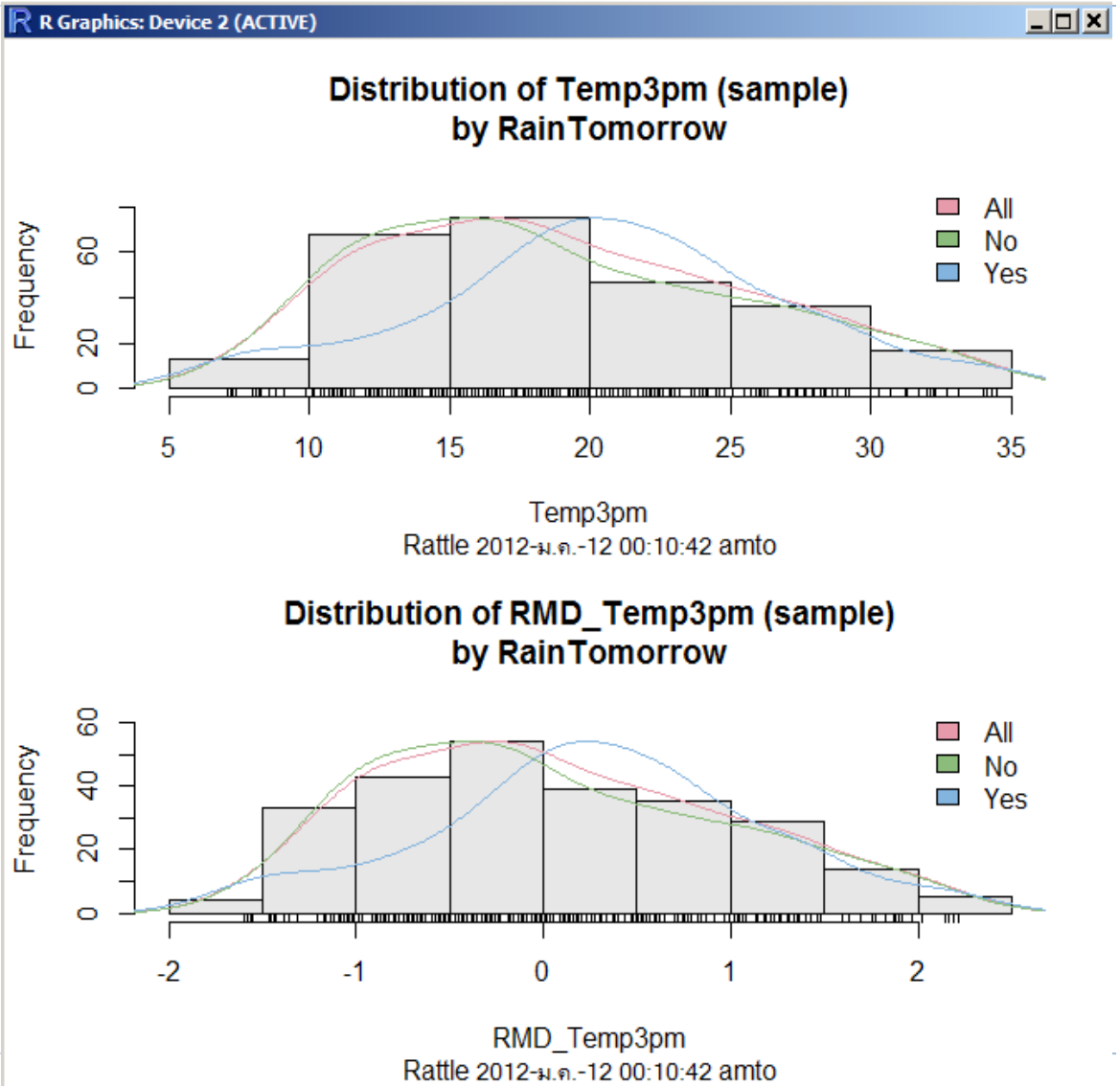
Rattle Training 6.Transforming Data





ทดลองสร้าง Histogram ระหว่างข้อมูลตัวแปร Temp3pm กับ R01_Temp3pm

Rattle Training 6.Transforming Data



ทดลองสร้าง Histogram ระหว่างข้อมูลตัวแปร Temp3pm กับ RMD_Temp3pm

Rattle Training 6.Transforming Data

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

Type: (•) Rescale ( ) Impute ( ) Recode ( ) Cleanup

Normalize: ( ) Recenter ( ) Scale [0-1] ( ) -Median/MAD (•) Natural Log ( ) Matrix

Order: ( ) Rank ( ) Interval Number of Groups: 100

No.	Variable	Data Type and Number Missing
10	WindDir9am	Categorical [16 levels; miss=31].
11	WindDir3pm	Categorical [16 levels; miss=1].
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; media
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; media
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
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	RLG_Temp3pm	Numeric [1.63 to 3.54; unique=200; mean=2.89; median=2.92]. No code export.

Normalized variables added to the dataset with 'RLG' prefix.

ในบางประเภทข้อมูล เช่น รายได้ บางครั้งจะไปกองอยู่แถวๆ 15,000 บาท ซึ่ง Logarithm จะทำให้การกระจายตัวที่ดีขึ้น (กรณีข้อมูลมีค่าบวก) เมื่อ take log

เช่น 10,000 แปลงเป็นค่า 4

100,000 แปลงเป็นค่า 5

150,000 แปลงเป็นค่า 5.176

1,000,000 แปลงเป็นค่า 6

Rattle Training 6.Transforming Data

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

Type: (•) Rescale ( ) Impute ( ) Recode ( ) Cleanup

Normalize: ( ) Recenter ( ) Scale [0-1] ( ) -Median/MAD ( ) Natural Log (•) Matrix

Order: ( ) Rank ( ) Interval Number of Groups: 100

No.	Variable	Data Type and Number Missing
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
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	RLG_Temp3pm	Numeric [1.63 to 3.54; unique=200; mean=2.89; median=2.92]. No code export.
26	RMA_Temp3pm	Numeric [0.00 to 0.00; unique=200; mean=0.00; median=0.00]. No code export.

RMA คือ Normalize : Matrix

Rattle Training 6.Transforming Data

R = Rescale

RRC = Recenter

R01= Scale [0-1]

RMD = Median/MAD

RLG = Natural Log

RMA = Matrix

Temp3pm	RainToda	RISK_MM	RainTomorrow	RRC_Temp3pm	R01_Temp3pm	RMD_Temp3pm	RLG_Temp3pm	RMA_Temp3pm
23.6				0.657967	0.629252	0.702305	3.16125	0.00335299
25.7				0.974215	0.70068	0.994352	3.24649	0.00365135
20.2				0.145945	0.513605	0.229466	3.00568	0.00286993
14.1				-0.772682	0.306122	-0.618863	2.64617	0.00200327
15.4				-0.576909	0.35034	-0.438071	2.73437	0.00218797
14.8				-0.667266	0.329932	-0.521513	2.69463	0.00210272
17.3				-0.290779	0.414966	-0.173838	2.85071	0.00245791
15.5				-0.561849	0.353741	-0.424164	2.74084	0.00220217
18.9				-0.0498279	0.469388	0.0486746	2.93916	0.00268523
21.7				0.371837	0.564626	0.438071	3.07731	0.00308304
24				0.718204	0.642857	0.757933	3.17805	0.00340982
26				1.01939	0.710884	1.03607	3.2581	0.00369397
27.1				1.18505	0.748299	1.18905	3.29953	0.00385025
30.7				1.72719	0.870748	1.6897	3.42426	0.00436172
30.4				1.68201	0.860544	1.64798	3.41444	0.0043191
30.7				1.72719	0.870748	1.6897	3.42426	0.00436172
29.8				1.59165	0.840136	1.56454	3.39451	0.00423386
28.6				1.41094	0.79932	1.39766	3.35341	0.00406337

Rattle Training 6.Transforming Data

- ผลลัพธ์ของแปลงข้อมูลให้อยู่ในรูปของ Rank หรือแบ่งเป็น Group ตามกำหนด

Rattle Dataset

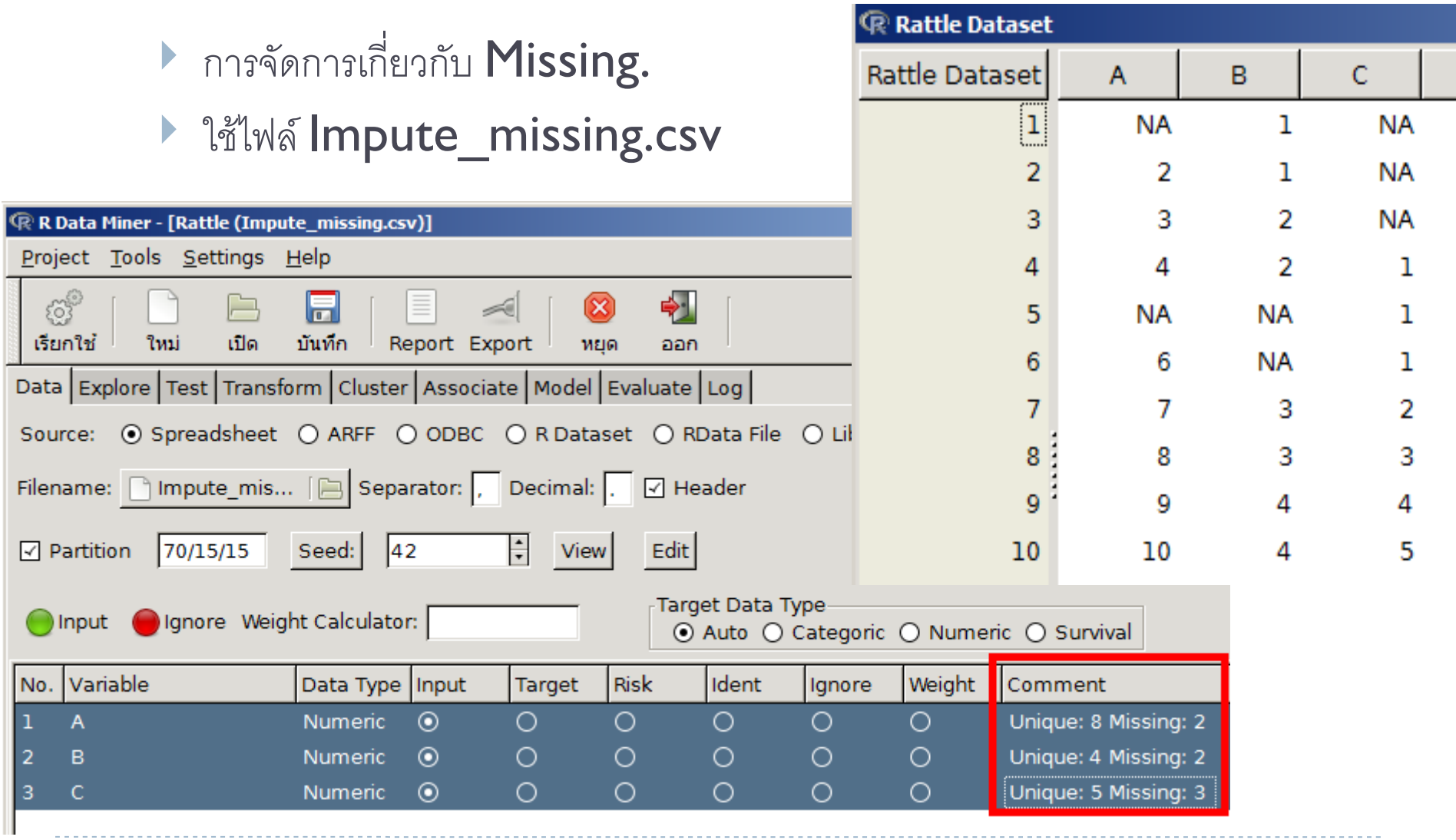
Rattle Dataset	essure	Cloud9ar	Cloud3pr	Temp9ar	Temp3pm	RainToda	RISK_MM	RainTom	RRK_Temp3pm	RIN_Temp3pm_10
352	1020.7	6	1	11.6	21.9	No	0	No	243.5	5
353	1022.2	1	7	12.7	23.7	No	0	No	269.5	6
354	1019.8	1	6	16.8					317	7
355	1017.5	1	1	16.4					302.5	6
356	1018.6	7	5	11.4	18.5	No	0.8	No	182	4
357	1023.1	1	5	8.3	14.3	No	0	No	94	3
358	1022.8	1	2	9.1	16.3	No	0	No	136	3
359	1018.4	3								4
360	1014.6	0								6
361	1014.2	5	8	16.3	25.9	No	0	No	296.5	6
362	1010.8	1	3	20.4	30	No	0	No	336.5	8
363	1016.9	0	1	17.2	28.2	No	0	No	323.5	7
364	1022.8	3	2	14.5	18.3	No	0	No	178	4
365	1016.2	6	7	15.8	25.9	No	0	No	296.5	6
366	1009.2	1	1	23.8	28.6	No	0	No	328.5	7

RRK คือเรียง Order แบบ Rank.

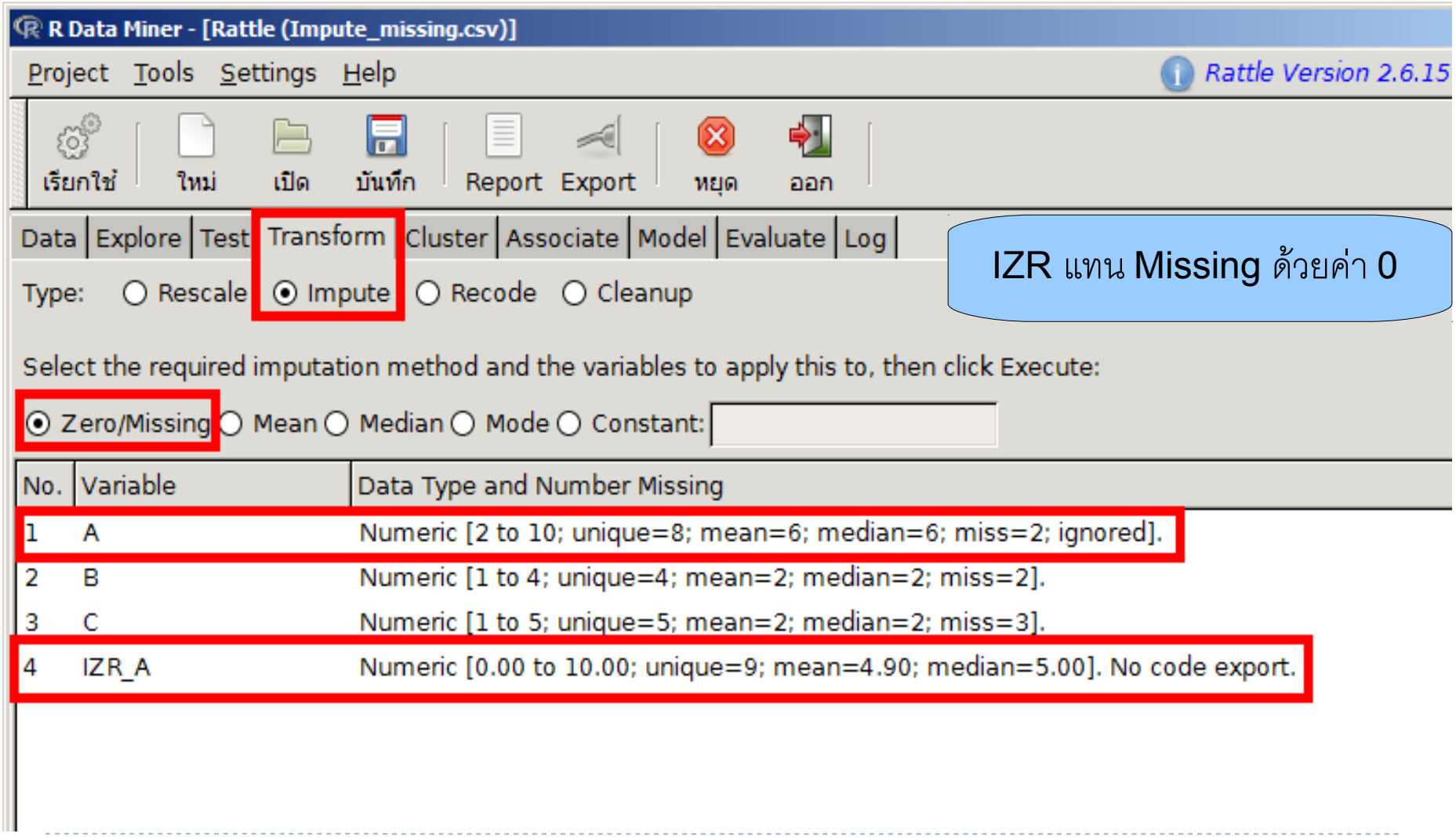
RIN คือเรียง Order แบบ Interval ในที่นี้ Number of Group = 10

Rattle Training 6.Transforming Data

- การจัดการเกี่ยวกับ Missing.
- ใช้ไฟล์ Impute_missing.csv



Rattle Training 6.Transforming Data



IZR แทน Missing ด้วยค่า 0

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Impute_missing.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: ○ Rescale ◉ Impute ○ Recode ○ Cleanup

IMN แทน Missing ด้วยค่า mean

Select the required imputation method and the variables to apply this to, then click Execute:

○ Zero/Missing ◉ Mean ○ Median ○ Mode ○ Constant:

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2; ignored].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2; ignored].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3].
4	IZR_A	Numeric [0.00 to 10.00; unique=9; mean=4.90; median=5.00]. No code export.
5	IMN_B	Numeric [1.00 to 4.00; unique=5; mean=2.50; median=2.50]. No code export.

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Impute_missing.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: ○ Rescale ◉ Impute ○ Recode ○ Cleanup

IMD แทน Missing ด้วยค่า median

Select the required imputation method and the variables to apply this to, then click Execute:

○ Zero/Missing ○ Mean ◉ Median ○ Mode ○ Constant:

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2; ignored].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2; ignored].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3; ignored].
4	IZR_A	Numeric [0.00 to 10.00; unique=9; mean=4.90; median=5.00]. No code export.
5	IMN_B	Numeric [1.00 to 4.00; unique=5; mean=2.50; median=2.50]. No code export.
6	IMD_C	Numeric [1 to 5; unique=5; mean=2; median=2]. No code export.

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Impute_missing.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Rescale Impute Recode Cleanup

IMO แทน Missing ด้วยค่า mode เหมาะกับ categoric variable

Select the required imputation method and the variables to apply this to, then click Execute:

Zero/Missing Mean Median Mode Constant:

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2; ignored].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2; ignored].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3; ignored].
4	IZR_A	Numeric [0.00 to 10.00; unique=9; mean=4.90; median=5.00]. No code export.
5	IMN_B	Numeric [1.00 to 4.00; unique=5; mean=2.50; median=2.50]. No code export.
6	IMD_C	Numeric [1 to 5; unique=5; mean=2; median=2]. No code export.
7	IMO_C	Numeric [1 to 5; unique=5; mean=2; median=1]. No code export.

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Impute_missing.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: ○ Rescale ◉ Impute ○ Recode ○ Cleanup

ICN แทน Missing ด้วยค่าคงที่

Select the required imputation method and the variables to apply this to, then click Execute:

○ Zero/Missing ○ Mean ○ Median ○ Mode ◉ Constant: 8

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2; ignored].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2; ignored].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3; ignored].
4	IZR_A	Numeric [0.00 to 10.00; unique=9; mean=4.90; median=5.00]. No code export.
5	IMN_B	Numeric [1.00 to 4.00; unique=5; mean=2.50; median=2.50]. No code export.
6	IMD_C	Numeric [1 to 5; unique=5; mean=2; median=2]. No code export.
7	IMO_C	Numeric [1 to 5; unique=5; mean=2; median=1]. No code export.
8	ICN_C	Numeric [1.00 to 8.00; unique=6; mean=4.10; median=3.50]. No code export.

Rattle Training 6.Transforming Data

Rattle Dataset

Rattle Dataset	A	B	C	IZR_A	IMN_B	IMD_C	IMO_C	ICN_C
1	NA	1	NA	0	1	2	1	8
2	2	1	NA	2	1	2	1	8
3	3	2	NA	3	2	2	1	8
4	4	2	1	4	2	1	1	1
5	NA	NA	1	0	2.5	1	1	1
6	6	NA	1	6	2.5	1	1	1
7	7	3	2	7	3	2	2	2
8	8	3	3	8	3	3	3	3
9	9	4	4	9	4	4	4	4
10	10	4	5	10	4	5	5	5
	NA	NA	NA	NA	NA	NA	NA	NA

IZR แทน missing ด้วย 0 ; IMN แทน missing ด้วย mean.

IMD แทน missing ด้วย median ; IMO แทน missing ด้วย mode

ICN แทน missing ด้วย constant (ค่าคงที่ใน ตย.นี้ คือ 8)

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (weather.csv)]

ใช้ไฟล์ weather.csv

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ◉ Quantiles ○ KMeans ○ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ○ As Categoric ○ As Numeric

BO4 แทน Recode ข้อมูล จาก Numeric เป็น Categorical ให้เป็นกล่องๆ ของค่า Quartiles กำหนดจำนวนกล่องได้ที่ Number

No.	Variable	Data Type and Number Missing
10	WindDir9am	Categorical [16 levels; miss=31].
11	WindDir3pm	Categorical [16 levels; miss=1].
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
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	BO4_Temp9am	Categorical [4 levels]. No code export.

Rattle Training 6.Transforming Data

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ○ Quantiles ◉ KMeans ○ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ○ As Categoric ○ As Numeric

BK4 แทน Recode ข้อมูล ให้เป็นกล่องๆ ตามค่า KMeans กำหนดจำนวนกล่องได้ที่ Number

No.	Variable	Data Type and Number Missing
11	WindDir3pm	Categorical [16 levels; miss=1].
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
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	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.

Rattle Training 6.Transforming Data

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ○ Quantiles ○ KMeans ◉ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ○ As Categoric ○ As Numeric

BE_ แทน Recode ข้อมูล
ให้เป็นกล่องเท่าๆ กัน Equal Width
กำหนดจำนวนกล่องได้ที่ Number

No.	Variable	Data Type and Number Missing
12	WindSpeed9am	Numeric [0 to 41; unique=22; mean=9; median=7; miss=7].
13	WindSpeed3pm	Numeric [0 to 52; unique=26; mean=17; median=17].
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
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	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.
27	BE4_Temp9am	Categorical [4 levels]. No code export.

Rattle Training 6.Transforming Data

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

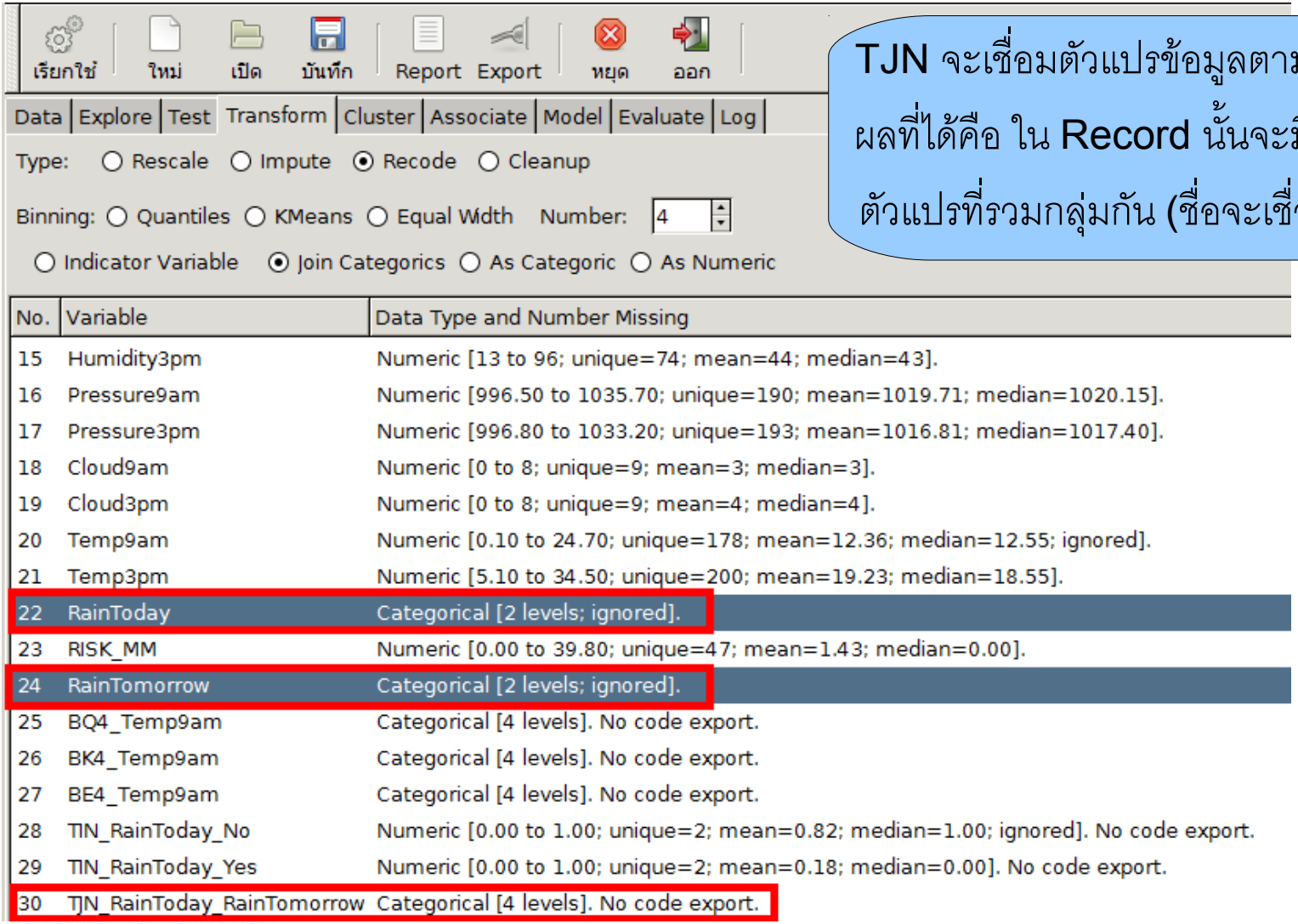
Binning: ○ Quantiles ○ KMeans ○ Equal Width Number: 4

◉ Indicator Variable ○ Join Categorics ○ As Categoric ○ As Numeric

TIN จะแปลงข้อมูล Categoric ไปเป็น Binary Numeric
ตัวแปรใหม่จะเพิ่มตามจำนวนเชิงกลุ่มที่มี

No.	Variable	Data Type and Number Missing
14	Humidity9am	Numeric [36 to 99; unique=60; mean=72; median=72].
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
22	RainToday	Categorical [2 levels; ignored].
23	RISK_MM	Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
24	RainTomorrow	Categorical [2 levels].
25	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.
27	BE4_Temp9am	Categorical [4 levels]. No code export.
28	TIN_RainToday_No	Numeric [0.00 to 1.00; unique=2; mean=0.82; median=1.00; ignored]. No code export.
29	TIN_RainToday_Yes	Numeric [0.00 to 1.00; unique=2; mean=0.18; median=0.00]. No code export.

Rattle Training 6.Transforming Data



No.	Variable	Data Type and Number Missing
15	Humidity3pm	Numeric [13 to 96; unique=74; mean=44; median=43].
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
22	RainToday	Categorical [2 levels; ignored].
23	RISK_MM	Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
24	RainTomorrow	Categorical [2 levels; ignored].
25	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.
27	BE4_Temp9am	Categorical [4 levels]. No code export.
28	TIN_RainToday_No	Numeric [0.00 to 1.00; unique=2; mean=0.82; median=1.00; ignored]. No code export.
29	TIN_RainToday_Yes	Numeric [0.00 to 1.00; unique=2; mean=0.18; median=0.00]. No code export.
30	TJN_RainToday_RainTomorrow	Categorical [4 levels]. No code export.

TJN จะเชื่อมตัวแปรข้อมูลตามที่เลือก ผลที่ได้คือ ใน Record นั้นจะมีค่าของตัวแปรที่รวมกลุ่มกัน (ชื่อจะเชื่อมกัน)

Rattle Training 6.Transforming Data

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ○ Quantiles ○ KMeans ○ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ◉ As Categoric ○ As Numeric

TFC จะแปลงข้อมูล Numeric ไปเป็น Categorical ที่มี levels ตาม unique

No.	Variable	Data Type and Number Missing
16	Pressure9am	Numeric [996.50 to 1035.70; unique=190; mean=1019.71; median=1020.15].
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4; ignored].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
22	RainToday	Categorical [2 levels; ignored].
23	RISK_MM	Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
24	RainTomorrow	Categorical [2 levels; ignored].
25	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.
27	BE4_Temp9am	Categorical [4 levels]. No code export.
28	TIN_RainToday_No	Numeric [0.00 to 1.00; unique=2; mean=0.82; median=1.00; ignored]. No code export.
29	TIN_RainToday_Yes	Numeric [0.00 to 1.00; unique=2; mean=0.18; median=0.00]. No code export.
30	TIN_RainToday_RainTomorrow	Categorical [4 levels]. No code export.
31	TFC_Cloud3pm	Categorical [9 levels]. No code export.

Rattle Training 6.Transforming Data

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

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ○ Quantiles ○ KMeans ○ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ○ As Categoric ◉ As Numeric

TNM จะแปลงข้อมูล Categorical ไปเป็น Numeric variable ค่าตัวเลขแปลงจาก levels เดิมที่มี

No.	Variable	Data Type and Number Missing
17	Pressure3pm	Numeric [996.80 to 1033.20; unique=193; mean=1016.81; median=1017.40].
18	Cloud9am	Numeric [0 to 8; unique=9; mean=3; median=3].
19	Cloud3pm	Numeric [0 to 8; unique=9; mean=4; median=4; ignored].
20	Temp9am	Numeric [0.10 to 24.70; unique=178; mean=12.36; median=12.55; ignored].
21	Temp3pm	Numeric [5.10 to 34.50; unique=200; mean=19.23; median=18.55].
22	RainToday	Categorical [2 levels; ignored].
23	RISK_MM	Numeric [0.00 to 39.80; unique=47; mean=1.43; median=0.00].
24	RainTomorrow	Categorical [2 levels; ignored].
25	BQ4_Temp9am	Categorical [4 levels]. No code export.
26	BK4_Temp9am	Categorical [4 levels]. No code export.
27	BE4_Temp9am	Categorical [4 levels]. No code export.
28	TIN_RainToday_No	Numeric [0.00 to 1.00; unique=2; mean=0.82; median=1.00; ignored]. No code export.
29	TIN_RainToday_Yes	Numeric [0.00 to 1.00; unique=2; mean=0.18; median=0.00]. No code export.
30	TJN_RainToday_RainTomorrow	Categorical [4 levels]. No code export.
31	TFC_Cloud3pm	Categorical [9 levels]. No code export.
32	TNM_RainToday	Numeric [1.00 to 2.00; unique=2; mean=1.18; median=1.00]. No code export.

Rattle Training 6.Transforming Data

- สรุปรูปแบบการ Recode ข้อมูล (ลองพิจารณาข้อมูลในแถวที่ 1)
  - BQ4 แปลงข้อมูลเป็นเชิงกลุ่ม 4 ช่วง ข้อมูลแรกอยู่ในช่วง 12.6-17
  - BK4 แปลงข้อมูลด้วย k-means 4 กลุ่ม ข้อมูลแรกอยู่ในกลุ่ม 12.4-18.6
  - TIN แปลงข้อมูลเชิงกลุ่มไปเป็น Binary numeric (1,0)
  - TJN เชื่อมค่าตัวแปร categoric ของตัวแปรที่เลือก
    - ในที่นี้มาจากตัวแปร RainToday กับ RainTommorow เชื่อมเป็น No_Yes
  - TFC แปลง numeric ไปเป็น categoric [6.7]
  - TNM แปลง categoric ไปเป็น numeric (1)

Rattle Dataset

Rattle Dataset	MM	RainTomo	BQ4_Temp9am	BK4_Temp9am	BE4_Temp9am	TIN_RainToday_No	TIN_RainToday_Yes	TJN_RainToday_RainTomorrow	TFC_Cloud3pm	TNM_RainToday
1	3.6	Yes	(12.6,17]	(12.8,18]	(12.4,18.6]	1	0	No_Yes	(6,7]	1
2	3.6	Yes	(17,24.7]	(12.8,18]	(12.4,18.6]	0	1	Yes_Yes	(2,3]	2
3	9.8	Yes	(12.6,17]	(12.8,18]	(12.4,18.6]	0	1	Yes_Yes	(6,7]	2
4	2.8	Yes	(12.6,17]	(12.8,18]	(12.4,18.6]	0	1	Yes_Yes	(6,7]	2
5	0	No	(7.6,12.6]	(7.4,12.8]	(6.24,12.4]	0	1	Yes_No	(6,7]	2
6	0.2	No	(7.6,12.6]	(7.4,12.8]	(6.24,12.4]	1	0	No_No	(4,5]	1

Rattle Training 6. Transforming Data

- การจัดการเกี่ยวกับ Cleaning missing
- ใช้ไฟล์ Cleanup.csv

Rattle Dataset

Rattle Dataset	A	B	C	D	E	F
1	NA	1	NA	1	11	10
2	2	1	NA	2	22	9
3	3	2	NA	3	33	8
4	4	2	1	4	44	7
5	NA	NA	1	5	55	6
6	6	NA	1	6	66	5
7	7	3	2	7	77	4
8	8	3	3	8	88	3
9	9	4	4	9	99	2
10	10	4	5	10	100	1

ใช้ไฟล์ Cleanup.csv

Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Cleanup.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: Cleanup.csv Separator: , Decimal: . Header

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: Auto Categoric ... ric Survival

ใน Data tab. เลือก Ignore ตัวแปร D

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	A	Numeric	●	○	○	○	○	○	Unique: 8 Missing: 2
2	B	Numeric	●	○	○	○	○	○	Unique: 4 Missing: 2
3	C	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 3
4	D	Numeric	○	○	○	○	●	○	Unique: 10
5	E	Numeric	●	○	○	○	○	○	Unique: 10
6	F	Numeric	●	○	○	○	○	○	Unique: 10

Rattle Training 6.Transforming Data

- ไปที่แท็บ **Transform** เลือก **Type** เป็น **Cleanup**
- เลือก **Delete Ignored** เพื่อลบตัวแปรดังกล่าว

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

Type: ○ Rescale ○ Impute ○ Recode ◉ Cleanup

Select the type of cleanup you want and then Execute.

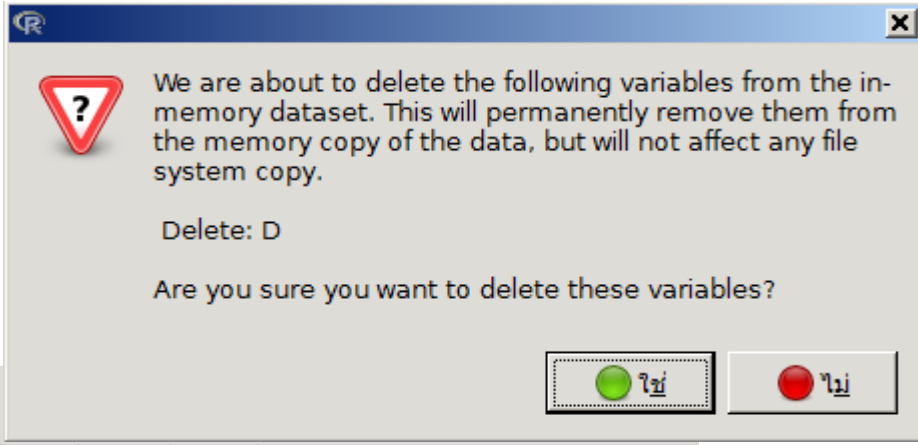
◉ Delete Ignored ○ Delete Selected ○ Delete Missing ○ Delete Obs with Missing

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3].
4	D	Numeric [1 to 10; unique=10; mean=5; median=5].
5	E	Numeric [11 to 100; unique=10; mean=59; median=60].
6	F	Numeric [1 to 10; unique=10; mean=5; median=5].

Ignored

Rattle Training 6.Transforming Data

- Rattle จะถามว่ายืนยันลบ D ?



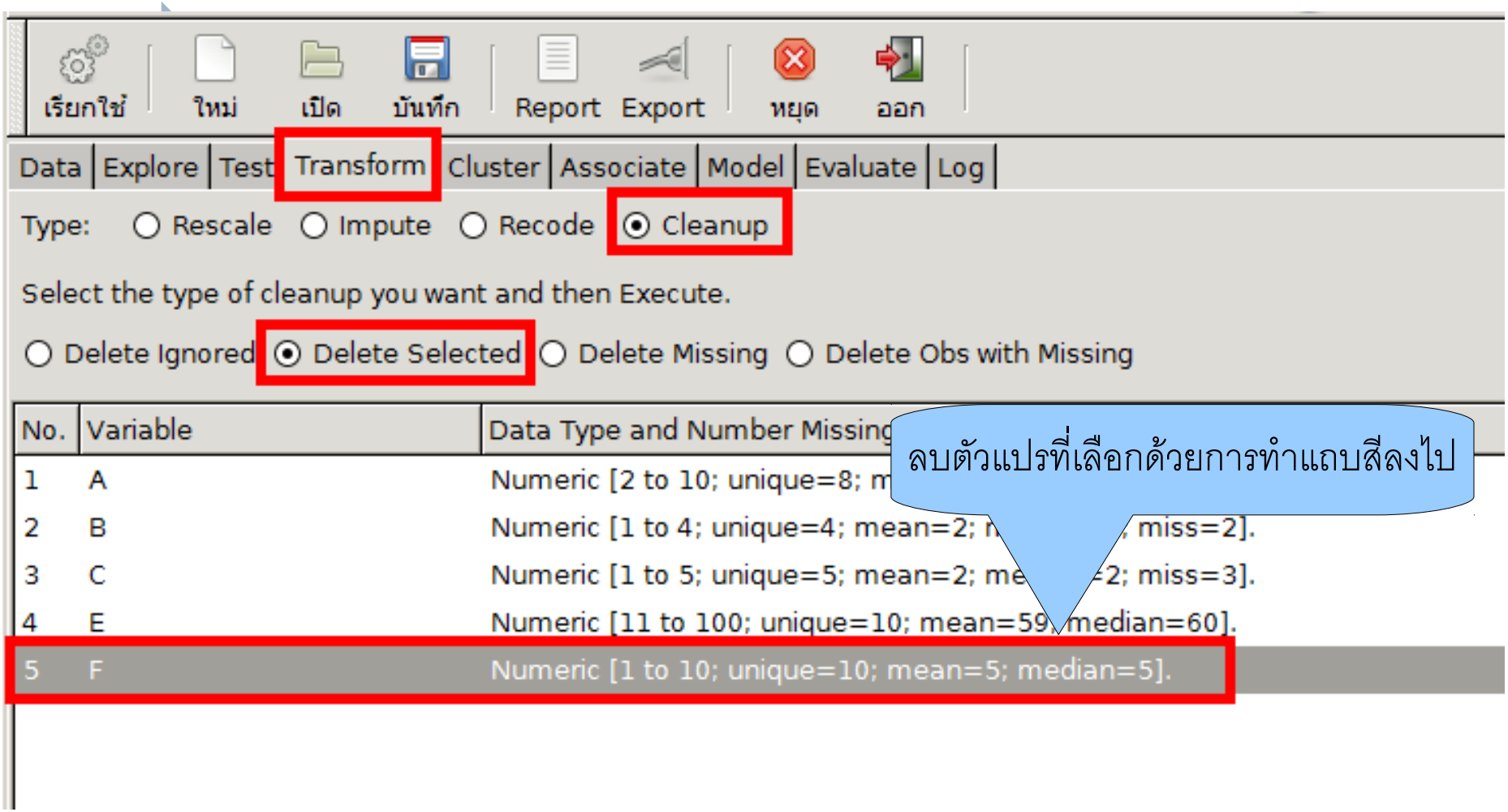
Select the type of cleanup you want and then Execute.

(•) Delete Ignored ( ) Delete Selected ( ) Delete Missing ( ) Delete Obs with Missing

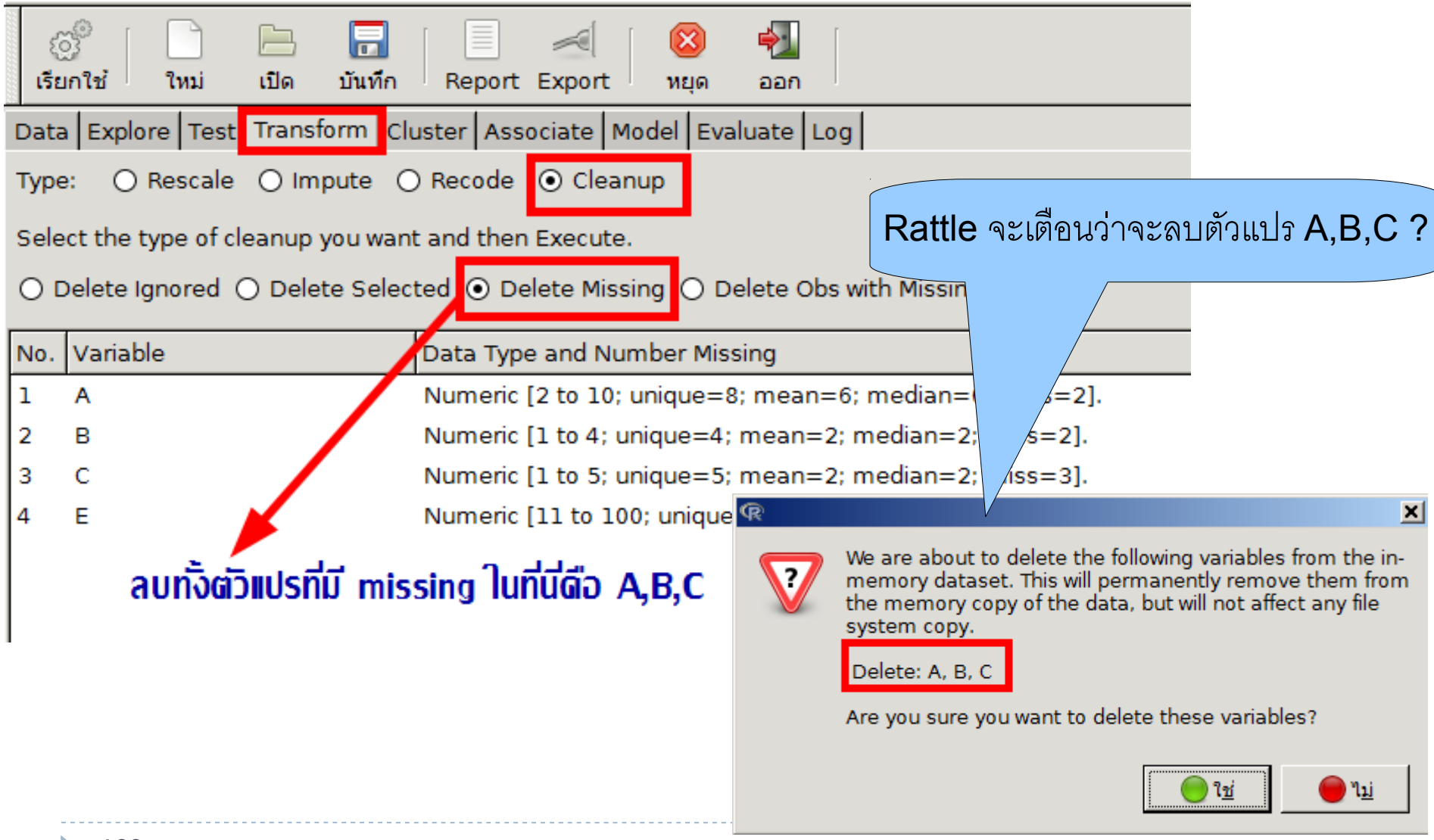
No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3].
4	E	Numeric [11 to 100; unique=10; mean=59; median=60].
5	F	Numeric [1 to 10; unique=10; mean=5; median=5].

จากการลบ Ignored variable ทำให้ D หายไปจาก Dataset

Rattle Training 6.Transforming Data



Rattle Training 6.Transforming Data



Rattle Training 6.Transforming Data

R Data Miner - [Rattle (Cleanup.csv)]

Project Tools Settings Help Rattle Ve

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: ○ Rescale ○ Impute ○ Recode ◉ Cleanup

Select the type of cleanup you want and then Execute.

○ Delete Ignored ○ Delete Selected ○ Delete Missing ◉ Delete Obs with Missing

No.	Variable	Data Type and Number Missing
1	A	Numeric [2 to 10; unique=8; mean=6; median=6; miss=2].
2	B	Numeric [1 to 4; unique=4; mean=2; median=2; miss=2].
3	C	Numeric [1 to 5; unique=5; mean=2; median=2; miss=3].
4	E	Numeric [11 to 100; unique=10; n

ลบแถวที่มี missing ในที่นี้มี 5 แถวคือ 1,2,3,5,6

Rattle จะเตือนว่า

จะลบข้อมูล 5 entities ?

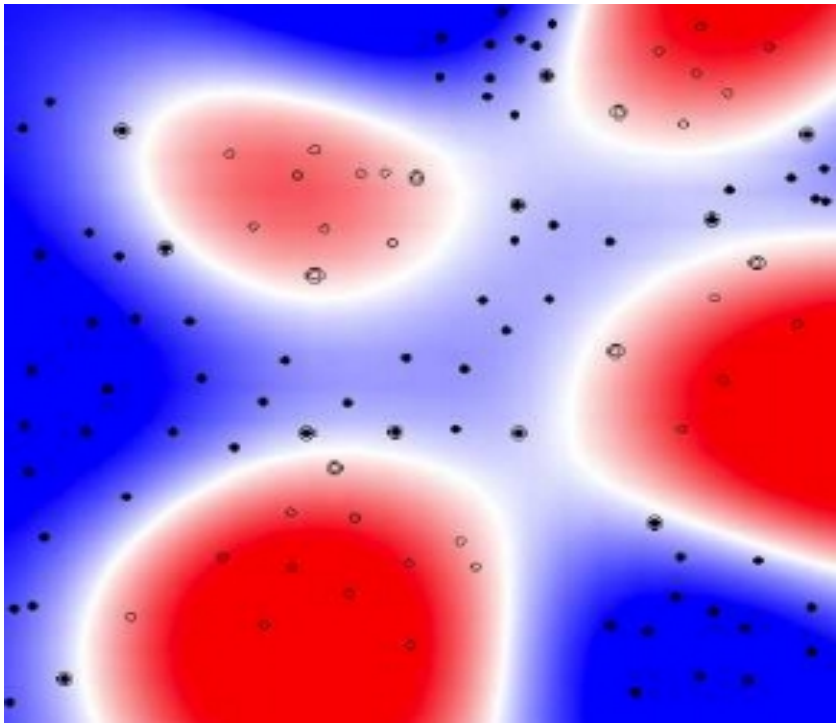
We are about to delete 5 entities from the in-memory dataset. These have missing values for some of the non-Ignore variables.

Are you sure you want to delete these entities?

ใช่ ไม่

Clustering

- Type of data in cluster analysis
  - K-means clustering
  - Hierarchical clustering

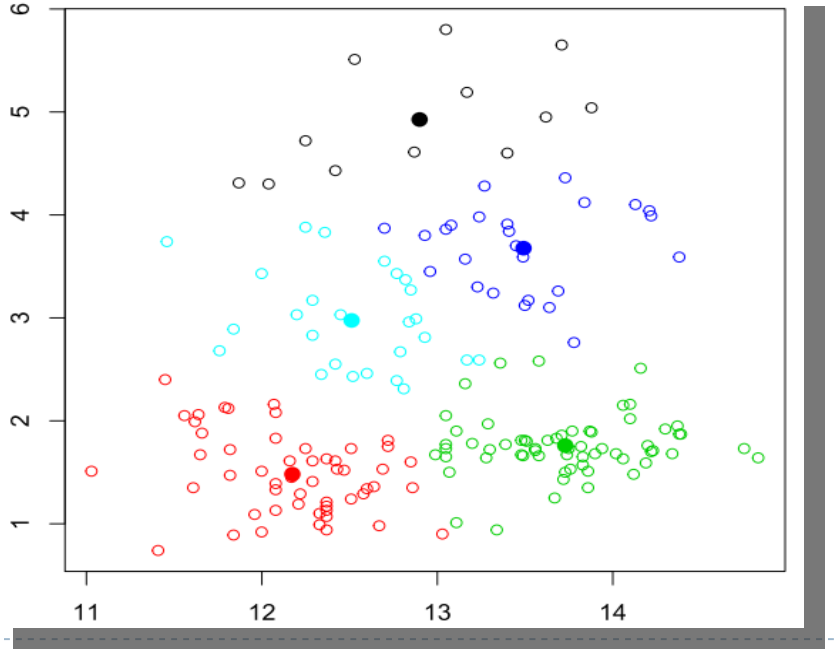


- Cluster Analysis
  - Analyze data without consulting a known class label
  - Maximizing the interclass similarity and minimizing the interclass similarity

Clustering คืออะไร?

- เรานิยามแบ่งกลุ่มคนเป็น {คนชั่ว-คนดี} {คนรวย-คนจน}
- ในทางธุรกิจเราแบ่งกลุ่มลูกค้าเป็น {คนที่ซื้อ-คนที่ไม่ซื้อ} {พึงพอใจ-ไม่พึงพอใจ}
- อื่น ๆ เช่น {เครดิตดี-เครดิตแย่} {โกง-ไม่โกง} {เลือกพรรค ก.-เลือกพรรค ข.} เป็นต้น

โลกของเรามักจะแบ่งกลุ่มไว้ในธรรมชาติ อยู่ที่เราจะมองออกหรือไม่?



Clustering คือ

- **Cluster :** คือการหาวิธีการแบ่งข้อมูลออกเป็นกลุ่มตามลักษณะประจำที่มีของแต่ละระเบียน โดยไม่มีลักษณะกลุ่ม แต่ใช้ระยะในการบ่งบอกความคล้ายและความแตกต่างของระเบียน
- ระยะห่างระหว่างระเบียนที่คล้ายกันจะมีน้อย ส่วนระยะห่างของระเบียนที่ต่างกันจะมีมากกว่า
- **Cluster** มีลักษณะเป็น **unsupervised classification**

การประยุกต์ใช้ Clustering

- Pattern Recognition
- Spatial Data Analysis
  - Create thematic map in GIS by clustering feature spaces
  - Detect spatial cluster and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- www
  - Document (text) classification
  - Question Categorization
  - Weblog Access Patterns

การประยุกต์ใช้ Clustering (ต่อ)

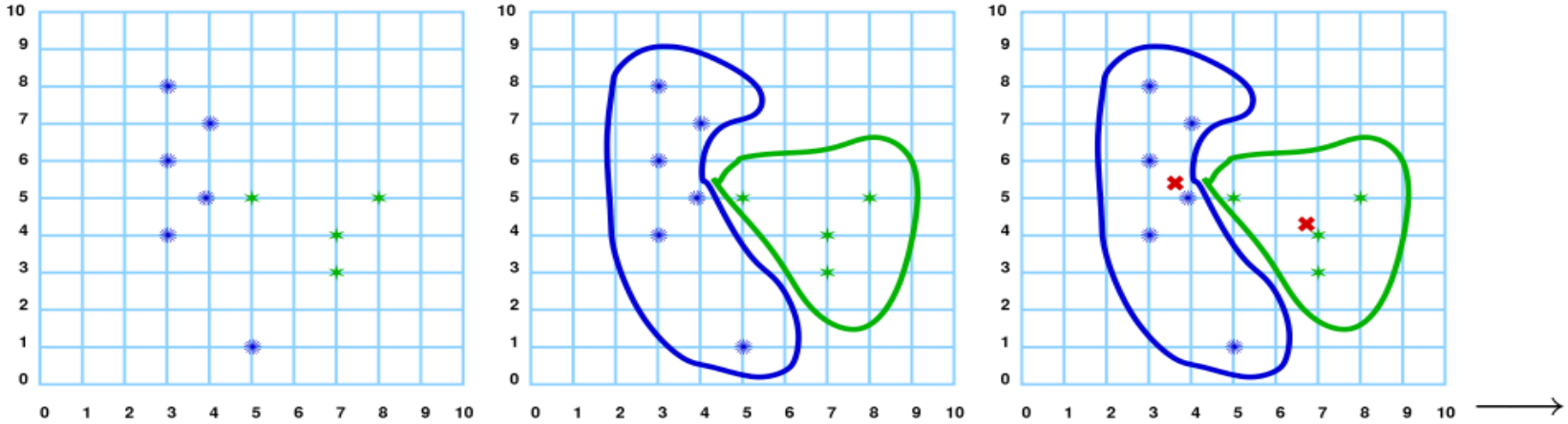
- **ด้านการตลาด** : clustering สามารถช่วยนักการตลาดค้นพบกลุ่มของลูกค้าใหม่ๆ แล้วนำความรู้นี้ไปพัฒนาสินค้ามาตอบสนองให้ตรงตามกลุ่ม อาจจะมองในแง่ของการบริหาร คือ จัดหน่วยงานขึ้นมาบริการกลุ่มลูกค้าแต่ละกลุ่ม ได้อย่างโดนใจก็ได้
- **ด้านประกันภัย** : clustering ช่วยระบุกลุ่มของผู้ถือกรมธรรม์ที่มีความเสี่ยงสูง หรือกลุ่มที่มีการเคลมประกันบ่อย (เอาไว้ออกกฎเกณฑ์หรือแนวปฏิบัติ)
- **ด้านการวางผังเมือง** : clustering ช่วยระบุกลุ่มของที่อยู่อาศัย ที่แบ่งตามลักษณะของบ้าน ,ราคา และที่ตั้งทางภูมิศาสตร์
- **ด้านวิทยาศาสตร์** : clustering ช่วยให้นักวิทยาศาสตร์จัดกลุ่มของดอกไม้ ,ต้นไม้ สัตว์ โปรตีน ดีเอ็นเอ ตามลักษณะประจำของแต่ละระเบียนได้

รูปแบบการวิเคราะห์ clustering

- วิธีแบ่งกั้น (Partitioning) : ระเบียนข้อมูลถูกแบ่งกั้นเป็นกลุ่มที่ไม่มีสมาชิกร่วมกันเลย โดยใช้การกั้นระหว่างกลุ่มด้วยระยะทาง
- ใช้ระดับชั้น (Hierarchy) : ระเบียนข้อมูลถูกรวมกลุ่มโดยการใช้ระดับชั้น (hierarchical decomposition) ที่ได้มาจากการคำนวนระยะห่างของแต่ละกลุ่มข้อมูล
- ใช้ความหนาแน่น (Density-based) : ระเบียนข้อมูลถูกรวมกันโดยใช้การเชื่อมกันและความหนาแน่น
- ใช้กริด (Grid-based) : ระเบียนข้อมูลถูกแบ่งตามกริด
- ใช้ตัวแบบ (Model-based) : เป็นการตั้งสมมติฐานของตัวแบบ แล้วเลือกตัวแบบที่แบ่งกั้นด้วยพารามิเตอร์ที่ดีที่สุด

ขั้นตอนวิธีการ clustering ด้วยวิธีแบ่งกั้น

- วิธีการแบ่งกั้น : ข้อมูล n ระเบียนแบ่งเป็น k กลุ่มที่ไม่มีสมาชิกร่วมกัน
- ขั้นตอนวิธีการแบ่งกลุ่มตามค่าเฉลี่ย k ตัว (k mean)
  - แบ่งข้อมูลเป็น k กลุ่มที่ไม่ใช่เซตว่าง
  - คำนวณจุดกึ่งกลาง (centroid) ของกลุ่มโดยใช้ค่าเฉลี่ยเลขคณิต (mean)
  - สำหรับแต่ละระเบียน นำระเบียนเทียบกับจุดกึ่งกลางเพื่อกำหนดกลุ่มให้กับระเบียน โดยเลือกระยะจากระเบียนไปจุดกึ่งกลางที่ใกล้ที่สุด
  - วนซ้ำจนกระทั่งไม่มีการเปลี่ยนกลุ่มของระเบียน หรือครบจำนวนรอบสูงสุดที่กำหนดไว้

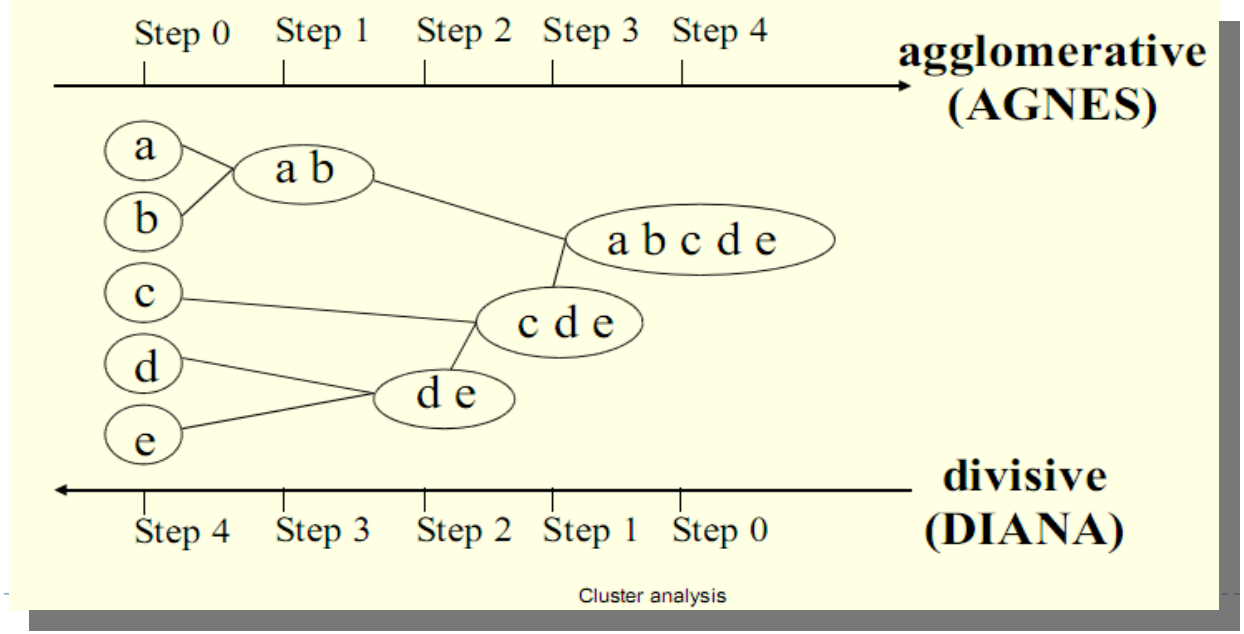


K-means Clustering

- ผลลัพธ์ที่ได้จากการวิเคราะห์ **clustering** ได้มาจากโครงสร้างในข้อมูล ดังนั้นการ **clustering** จึงไม่มีลักษณะการบอกว่าการเกาะกลุ่มนั้นถูกหรือผิด
- <u>ข้อดีของ **k-means**</u>
  - เหมาะกับข้อมูลจำนวนมาก (200 ขึ้นไป)
  - มีประสิทธิภาพสูง ขั้นตอนใช้วิธี O(tkn), เมื่อ n คือจำนวนข้อมูล k คือจำนวนกลุ่ม และ t คือจำนวนรอบที่ต้องวนซ้ำโดยปกติแล้วค่า k และ t จะน้อยกว่า n มาก
- <u>ข้อเสียของ **k-means**</u>
  - ใช้ได้เฉพาะลักษณะประจำที่เป็นจำนวนต้องใช้ค่าเฉลี่ย
  - ต้องมีการกำหนดค่า k ก่อนเริ่มขั้นตอนวิธี
  - ข้อมูลที่ผิดปกติ (outlier) จะทำให้ค่าเฉลี่ยมีค่าที่มากหรือน้อยเกินไป
  - ไม่เหมาะกับ non-convex clusters (ข้อมูลที่กระจายกันทั่ว)

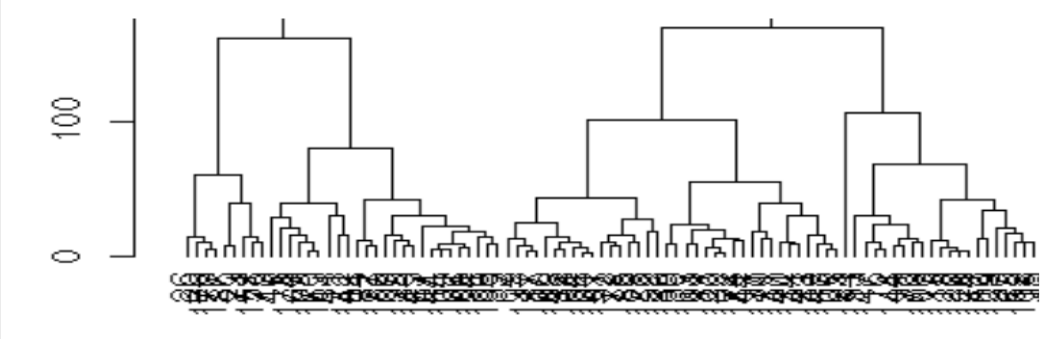
ขั้นตอนวิธีการ clustering ด้วยวิธีระดับชั้น (Hierarchy)

- ใช้ **distance matrix** เป็นเงื่อนไขในการเกาะกลุ่ม วิธีการนี้ไม่ต้องการให้กำหนดค่า **k** แต่ต้องการกำหนดเงื่อนไขในการหยุด
- แยกกลุ่มข้อมูลเป็นระดับชั้นแบบต้นไม้ **(tree of clusters)** เรียกว่า **dendrogram**
- วิธีการเกาะกลุ่มของข้อมูลคือการตัดผัง **dendrogram** ตามระดับที่ต้องการ ส่วนประกอบที่เชื่อมกันแสดงถึงการเกาะกลุ่มที่เกิดขึ้น



Hierarchical Cluster

- ข้อดี
  - ไม่จำเป็นต้องทราบจำนวนกลุ่มมาก่อน กรรมวิธีจัดกลุ่มให้
  - ใช้กับข้อมูลได้ทั้ง **Interval scale, ratio scale ,count data** หรือ **binary**
- ข้อเสีย
  - ไม่เหมาะกับข้อมูลปริมาณมาก ความซับซ้อนของเวลาคือ $O(n^2)$ เมื่อ **n** แทนจำนวนข้อมูลทั้งหมด
  - ไม่มีการแยกกลุ่มออกจากกัน เมื่อรวมแล้วข้อมูลจะอยู่ในกลุ่มนั้นตลอด



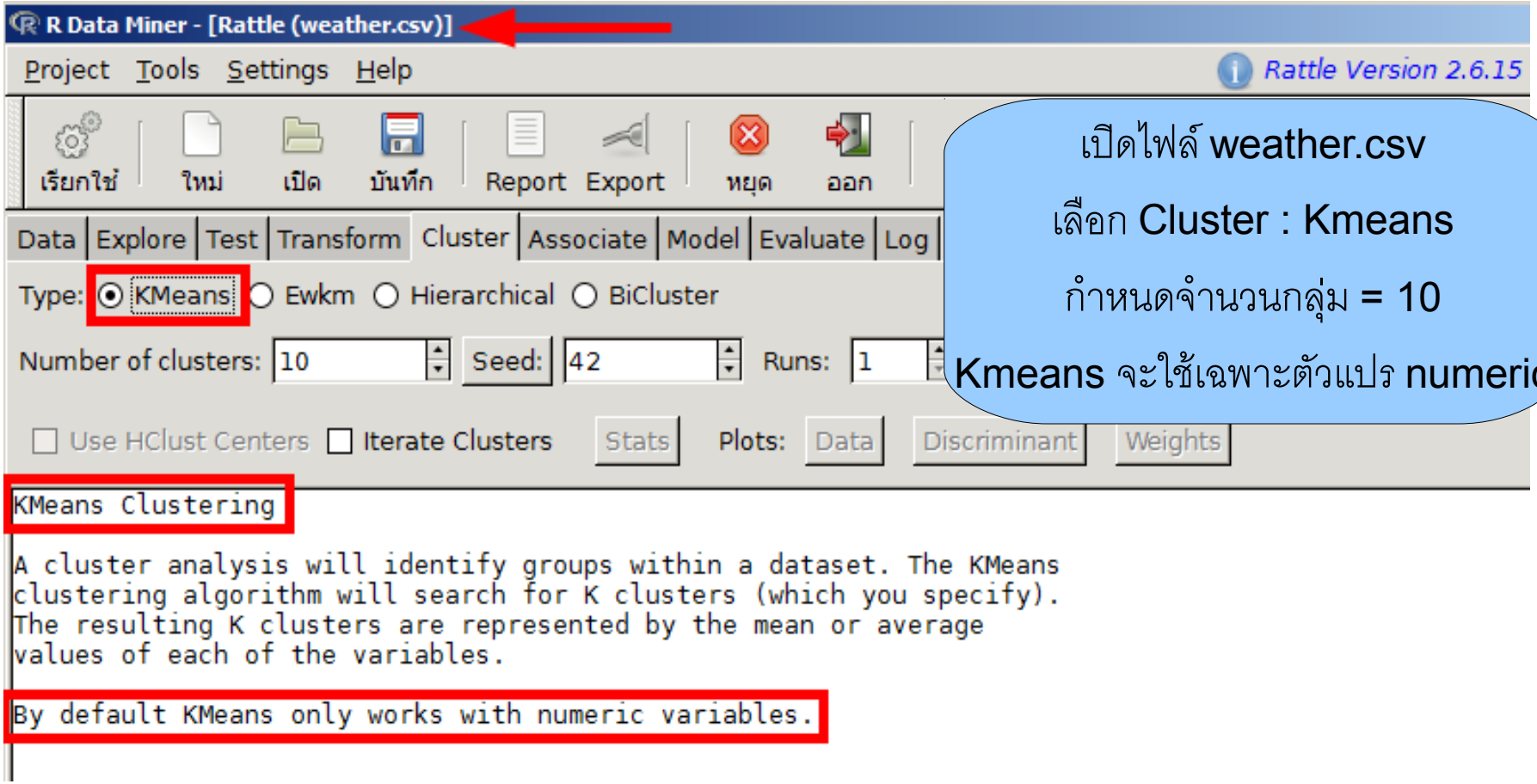
ลักษณะข้อมูลที่ต้องการของ Clustering

- การวิเคราะห์การเกาะกลุ่ม **(Cluster analysis)** ต้องใช้ระยะห่างระหว่างแต่ละ **records (centroid distance)** เพื่อบ่งบอกความเหมือน หรือต่างกัน
- การวัดระยะระหว่างแต่ละ **records** ใช้ผลรวมของการคำนวณระยะระหว่างลักษณะประจำแต่ละตัว
  - ในกรณีที่ลักษณะประจำเป็นจำนวน เราสามารถใช้ค่ากำลังสองของผลต่าง
  - ในกรณีที่ลักษณะประจำเป็นค่าที่ไม่ต่อเนื่อง เราสามารถใช้การนับความแตกต่างระหว่างลักษณะประจำที่ต่างกันได้
  - ลักษณะประจำนำเข้าอาจเป็นค่าต่อเนื่องหรือค่าไม่ต่อเนื่องก็ได้ เพราะใช้วิธีการพิจารณาการแยกกลุ่ม จากการคำนวณระยะห่างด้วยสเกลที่เหมาะสม

สรุป Clustering

- **Cluster analysis** เป็นวิธีการทำเหมืองข้อมูลแบบ **unsupervised**
- มีประโยชน์สำหรับการจัดกลุ่มประชากรขนาดใหญ่ อาจใช้ประโยชน์จากแต่ละ **sub-population** มาทำ **data mining** ด้วยวิธีอื่นอีกได้
- ชุดข้อมูลที่ต้องการเป็นเชิงกลุ่มหรือตัวเลข ,ลำดับ หรือ **Binary (0,1)**
- เนื่องจากต้องใช้ความเข้าใจด้านเทคนิคสูง ดังนั้นวิธีนี้จะมีประสิทธิภาพมากถ้ามีผู้ชำนาญการคอยให้คำแนะนำ

Rattle Training 7. Cluster Analysis



เปิดไฟล์ weather.csv

เลือก Cluster : Kmeans

กำหนดจำนวนกลุ่ม = 10

Kmeans จะใช้เฉพาะตัวแปร numeric

Rattle Training 7. Cluster Analysis

```
Cluster sizes:

[1] "23 17 22 22 17 36 23 34 22 32"

Data means:

      MinTemp       MaxTemp      Rainfall   Evaporation      Sunshine
     7.145565     20.372177      1.377419      4.544355      8.082661
WindGustSpeed  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
    39.943548      9.818548     18.056452     71.471774     43.858871
  Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
  1019.748387   1016.978629      3.689516      3.850806     12.269355
      Temp3pm
    19.081048

Cluster centers:

      MinTemp  MaxTemp   Rainfall Evaporation  Sunshine WindGustSpeed
1   8.5000000 21.04783 1.27826087    6.330435 10.495652      57.95652
2  11.6058824 30.94706 0.11764706    7.647059 11.276471      43.64706
3  13.4136364 28.76818 1.02727273    6.200000  9.463636      45.22727
4   9.1818182 16.90000 4.94545455    3.800000  2.190909      30.90909
5   7.7411765 15.19412 3.58823529    3.305882  5.658824      61.47059
6   2.7666667 17.16389 0.66111111    2.655556  7.688889      25.47222
7  -0.7913043 13.70870 0.03478261    1.921739  7.495652      39.30435
8  11.3088235 26.37353 0.50000000    6.288235 10.258824      35.88235
9   1.5045455 17.55000 0.23636364    3.500000 10.222727      33.72727
10  7.8625000 17.59687 2.21875000    4.518750  6.121875      41.50000
```

สรุปค่าต่างๆ ที่ได้จาก

Kmeans 10 กลุ่ม

Rattle Training 7. Cluster Analysis

- Model Quality อาจจะประเมินจากค่าความแตกต่างกันของค่า means
- Tuning Option
  - Number of clusters ระบุเป็นตัวเลขที่มากกว่า 1 แต่ไม่มากกว่าจำนวน cases
  - Iterate Clusters จะแสดงค่า sum of within cluster sum of squares
  - Iterate Cluster จะคำนวน sum SS ไปจนถึง cluster สูงสุดเพื่อเลือกว่ากี่กลุ่มดี
  - Seed ใช้ปรับจุดเริ่มต้นในการสุ่ม (ถ้าแต่ละครั้งค่า seed ต่างกัน model มักจะต่างกัน)

```
Within cluster sum of squares:

 [1] 14459.755  5469.311  8061.781 11733.534 11062.079  9582.717  7258.203
 [8]  7806.196  6146.200 11528.942

Time taken: 0.02 secs

Rattle timestamp: 2012-01-14 14:14:44 amto
==================================================================
```

Rattle Training 7. Cluster Analysis

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

Type: (•) KMeans ( ) Ewkm ( ) Hierarchical ( ) BiCluster

Number of clusters: 10 Seed: 42 Runs: 1

[ ] Use HClust Centers [x] Iterate Clusters Stats Plots: Data Discriminant Weights

```
General cluster statistics:

$n
[1] 248

$cluster.number
[1] 10

$cluster.size
 [1] 23 17 22 22 17 36 23 34 22 32

$diameter
 [1] 63.51638 41.82105 44.71666 54.22546 58.27735 46.76227 45.12139 36.60615
 [9] 37.39813 49.31318

$average.distance
 [1] 34.62206 25.18234 26.69300 32.27777 35.63701 22.57469 24.76224 21.19047
 [9] 23.33485 26.42843

$median.distance
 [1] 33.59747 24.66780 25.94475 32.55580 34.60015 22.34670 24.32324 20.98309
 [9] 23.57265 26.22680
```

Rattle Training 7. Cluster Analysis

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

Type: (•) KMeans ( ) Ewkm ( ) Hierarchical ( ) BiCluster

Number of clusters: 10 Seed: 42 Runs: 1

[ ] Use HClust Centers [ ] Iterate Clusters Stats Plots: Data Discriminant Weight

Rattle ถามว่าจะ Plot แค่ 5 ตัวแปรแรกหรือไม่

```
10  7.8625000 17.59687 2.21875000    4.518750  6.121875        41.50000
   WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm
1     21.652174    30.217391    51.91304    33.43478    1015.300    1013.274
2      6.294118    18.411765
3      6.136364    17.63636
4      7.500000    12.40909
5     24.470588    26.82352
6      5.055556     9.55555
7      4.565217    22.30434
8      7.205882    11.64705
9      5.590909    17.18181
10    14.343750    22.56250
   Cloud9am Cloud3pm   Temp
1  1.913043 2.739130 14.482
2  1.352941 1.647059 19.429
3  4.500000 3.954545 17.295
4  7.318182 7.090909 11.577273 15.37727
```

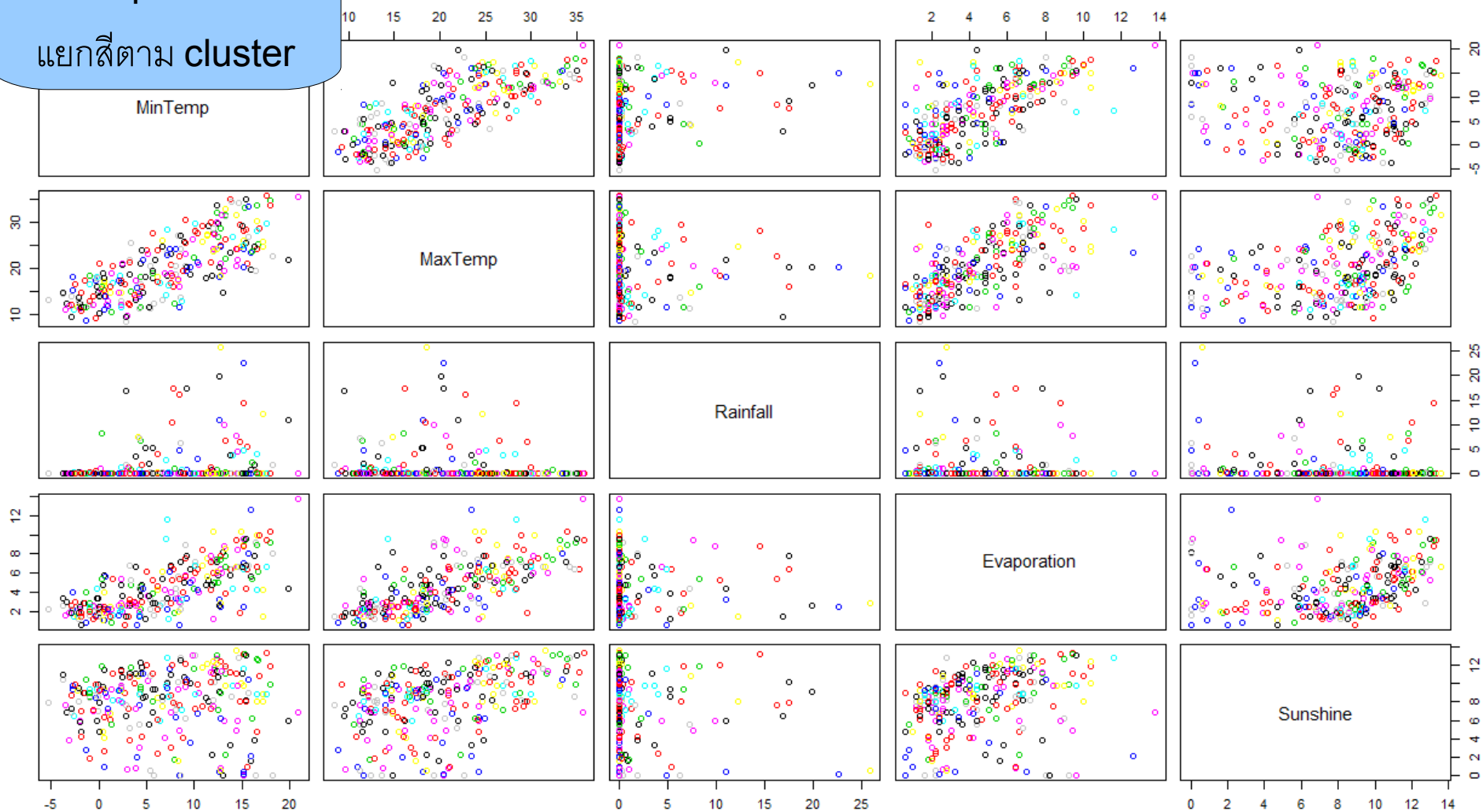
The dataset contains many variables. For more than 5 variables the plot will be cluttered. The application will also wait whilst the plot is drawn. If you continue, the first 5 variables will be used.

Would you like to continue with the reduced variables?

ใช่ ไม่

Rattle Training 7. Cluster Analysis

Scatter plot 5 ตัวแปร แยกสีตาม cluster



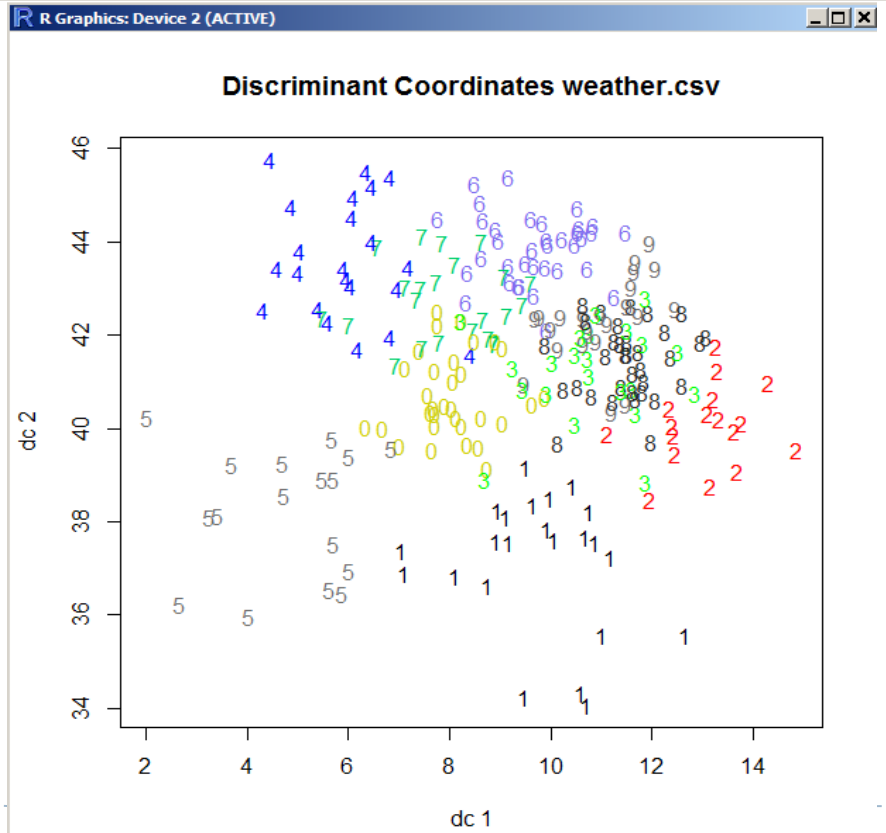
Rattle Training 7. Cluster Analysis

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

Type: (•) KMeans ( ) Ewkm ( ) Hierarchical ( ) BiCluster

Number of clusters: 10 Seed: 42 Runs: 1

[ ] Use HClust Centers [ ] Iterate Clusters Stats Plots: Data Discriminant Weights



Rattle Training 7. Cluster Analysis

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score

Model: Tree Boost Forest SVM Linear Neural Net Surviva KMeans HClust

Data: Training Validation Testing Full Enter CSV File Doc... R Dataset

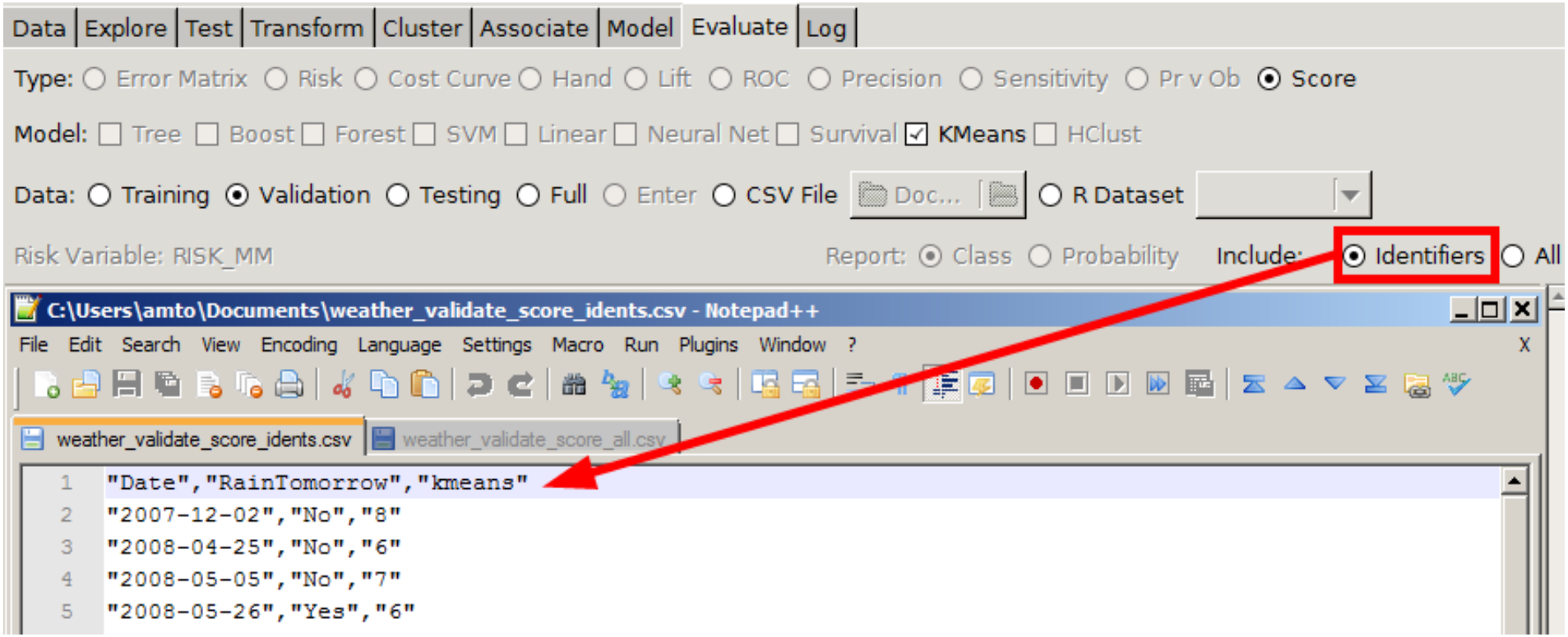
Risk Variable: RISK_MM Report: Class Probability Include: Id

```
Score

A model can be deployed on a dataset to obtain scores or
classifications for each observation in the dataset.

By default the testing dataset (if any) will be scored. O
training dataset is scored. As an alternative, a CSV file can be
loaded and scored. This choice of what is scored is controlled by the
radio button options.

For binary models a probability score can be recorded. For regression
models a value is recorded for each observation. Otherwise a class
will be recorded for each observation. This can be controlled by the
Class and Probability radio buttons.

The resulting CSV file will include just those variables having a role
as Identifier (plus the Target and the Score), or else all of the
variables.

The name of a CSV file into which the results will be written will be
prompted for.
```

เราสามารถบันทึกผลลัพธ์การจัดกลุ่มได้ที่นี่

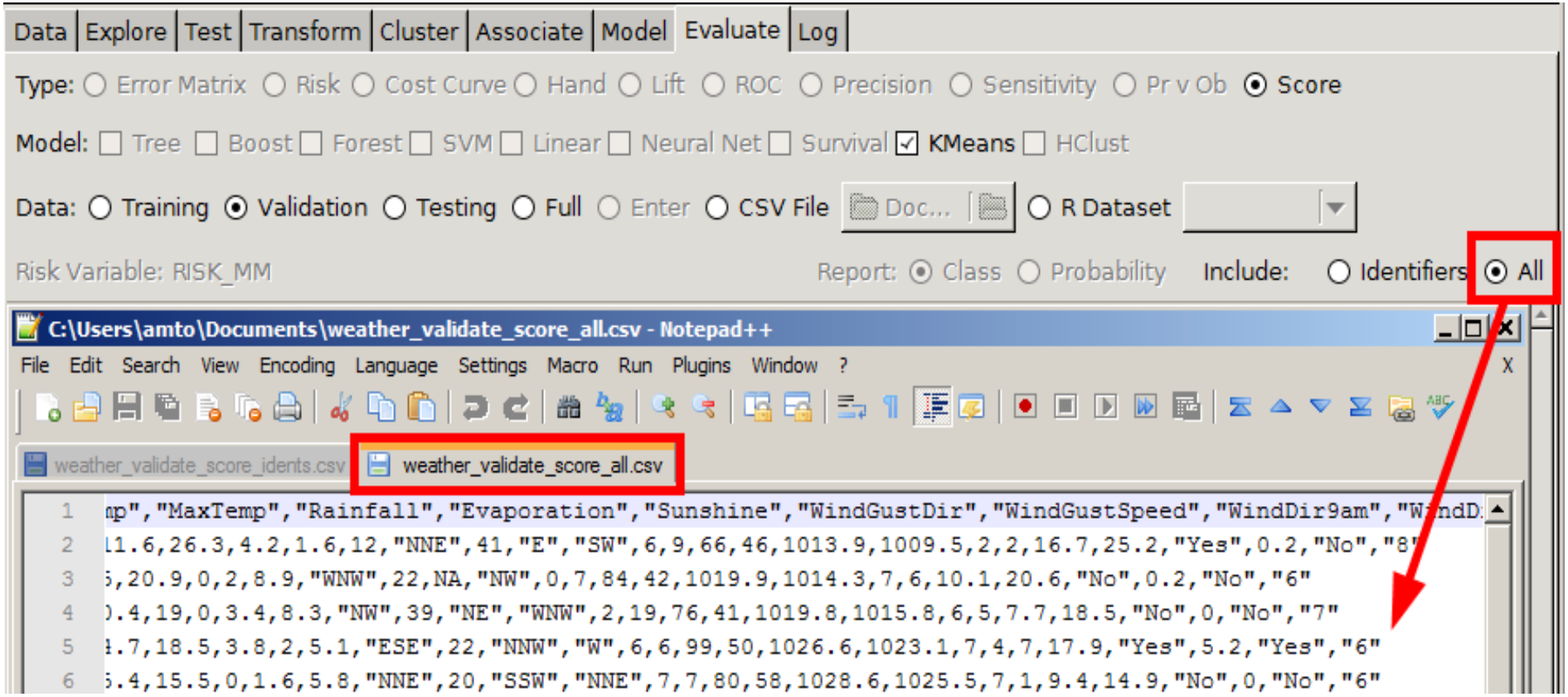
Rattle Training 7. Cluster Analysis

- Rattle สามารถแสดงกลุ่มของข้อมูลในไฟล์ซึ่งสามารถบันทึกค่า score ออกมา
- Include : Identifiers จะแสดงแบบย่อ
  - คือ Identifiers (Date), Target (RainTomorrow) & kmeans.



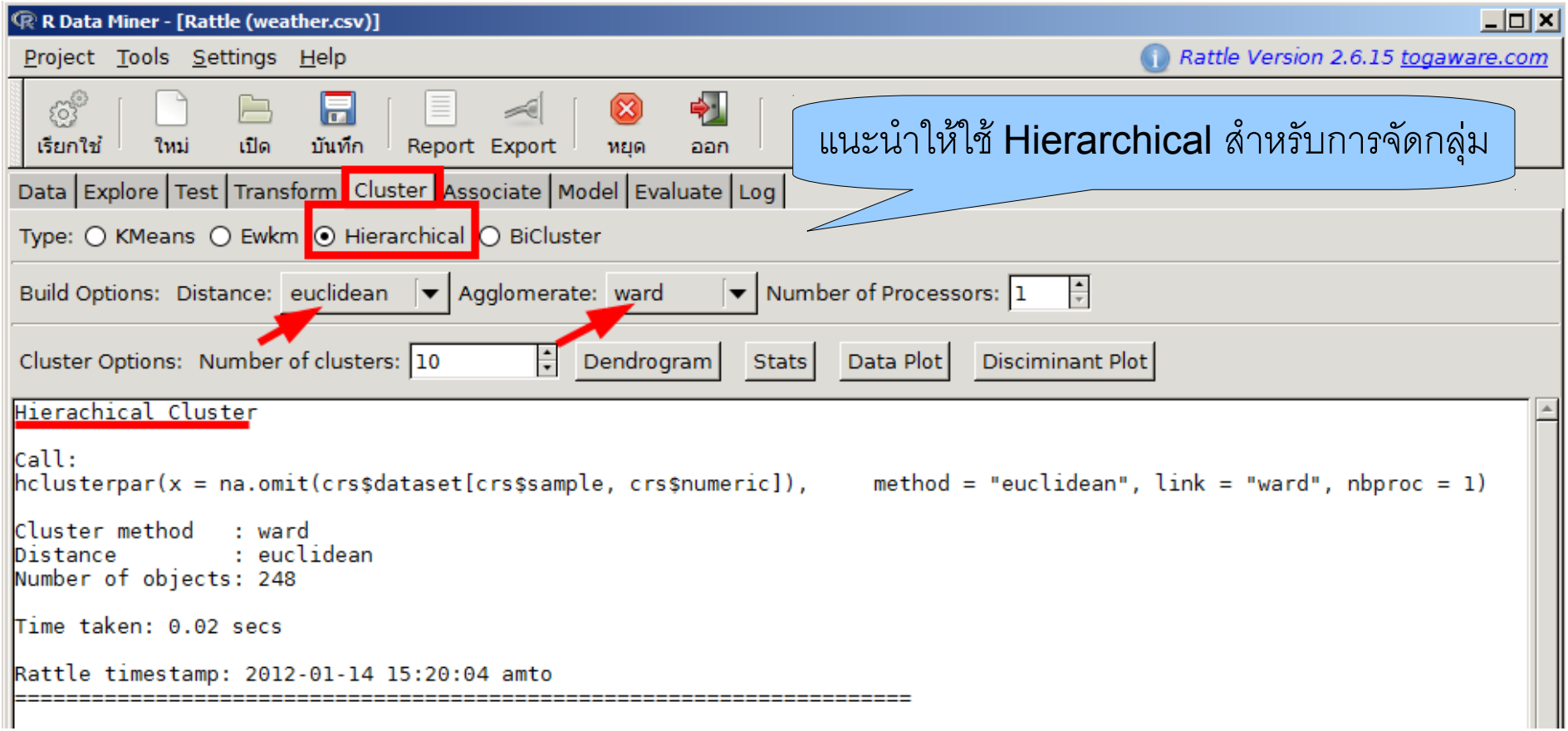
Rattle Training 7. Cluster Analysis

- กรณีที่เลือก **Include** เป็น **All** จะบันทึกข้อมูลทั้งไฟล์พร้อมกับ **kmeans num.**



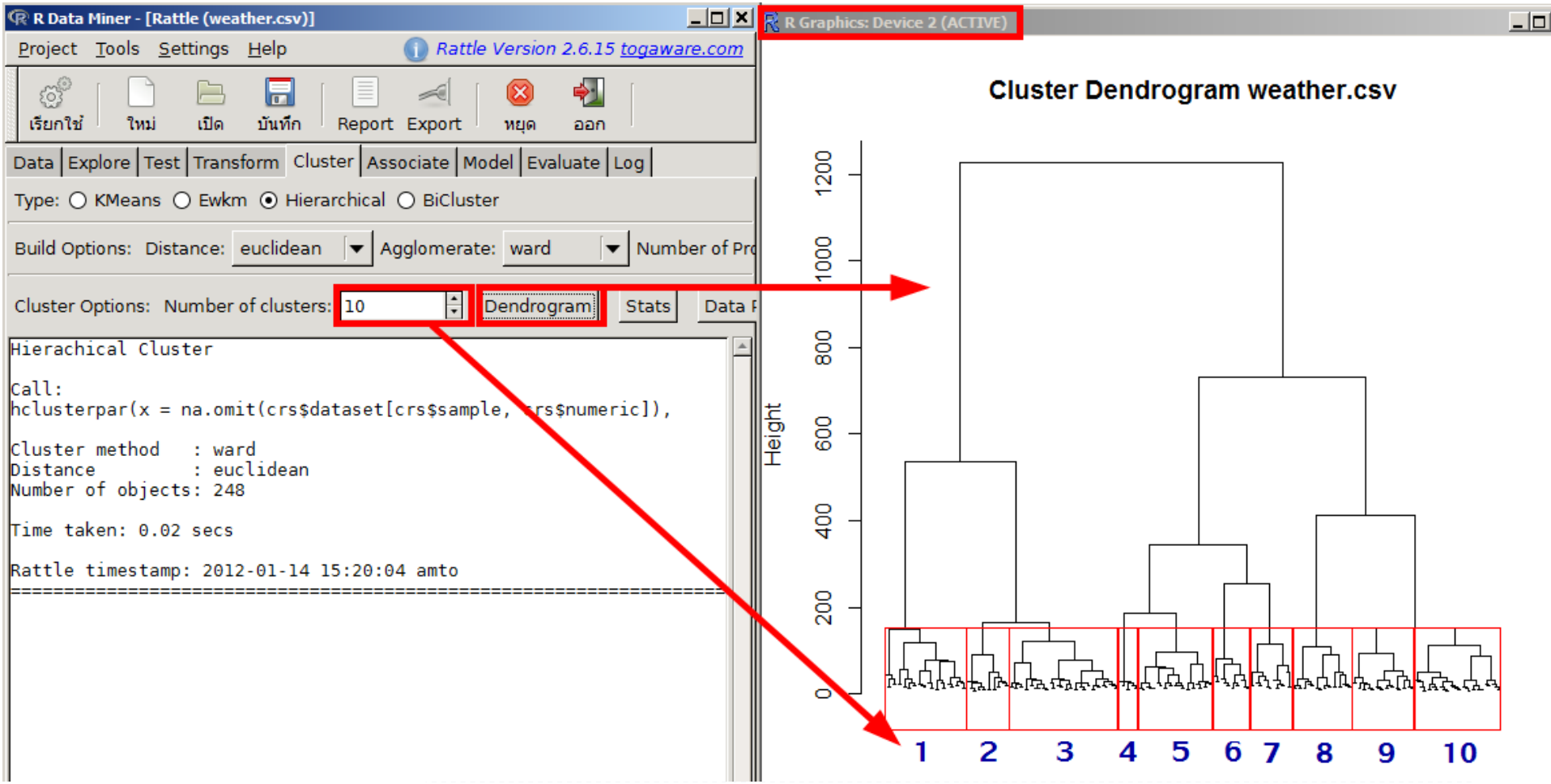
Rattle Training 7. Cluster Analysis

- วิธี Hierarchical จะมี option ในการเลือก Distance & Agglomerate
- นอกจากนี้ยังมีการสร้าง Dendrogram ให้เห็นลำดับขั้นความสัมพันธ์ของข้อมูลด้วย



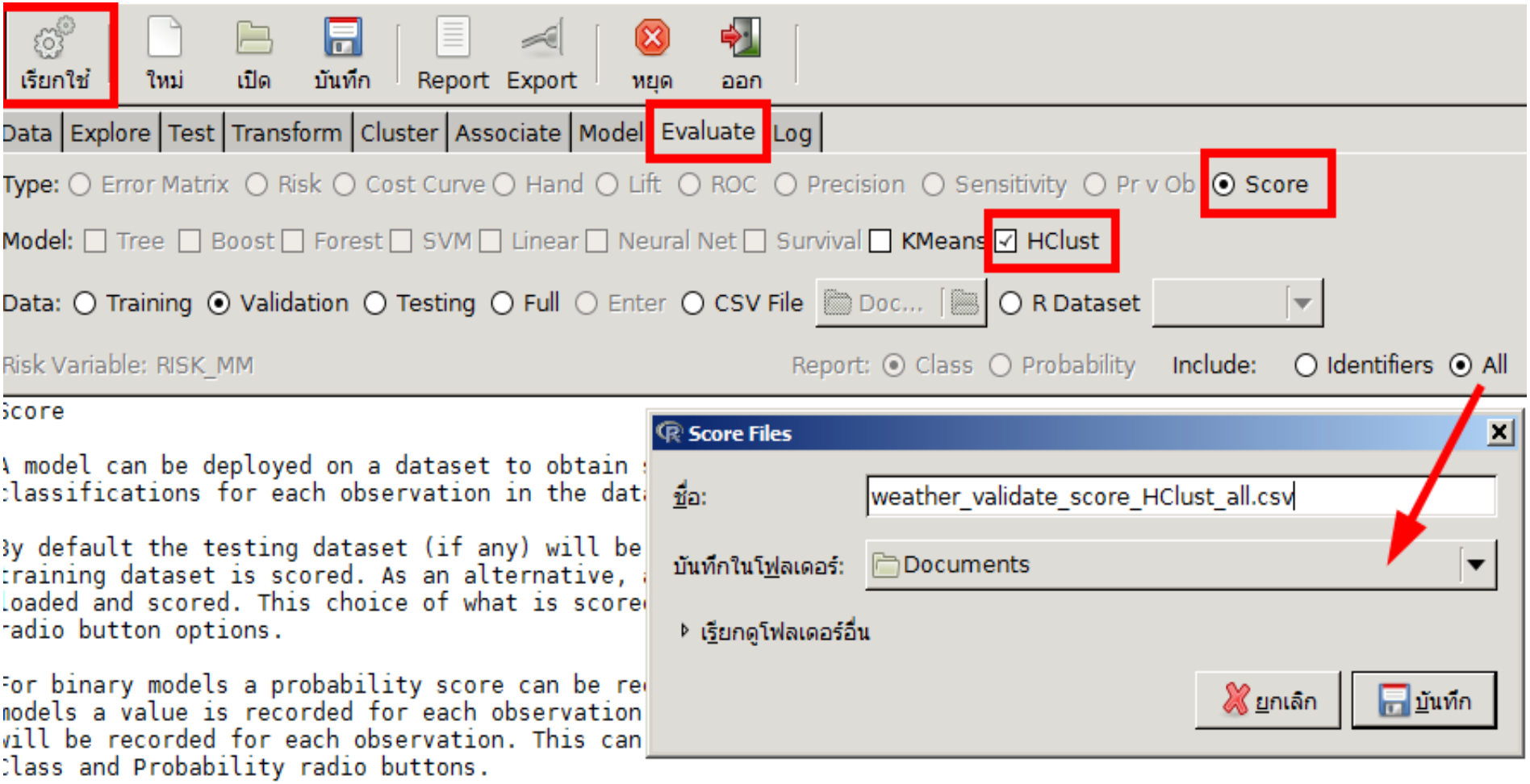
Rattle Training 7. Cluster Analysis

- กรอบสีแดงจะแบ่งกลุ่มตาม Number of clusters. (ในที่นี้คือ 10 กรอบ)



Rattle Training 7. Cluster Analysis

- การบันทึก Score ในส่วนของ Hierarchical ให้เลือกที่ HClust

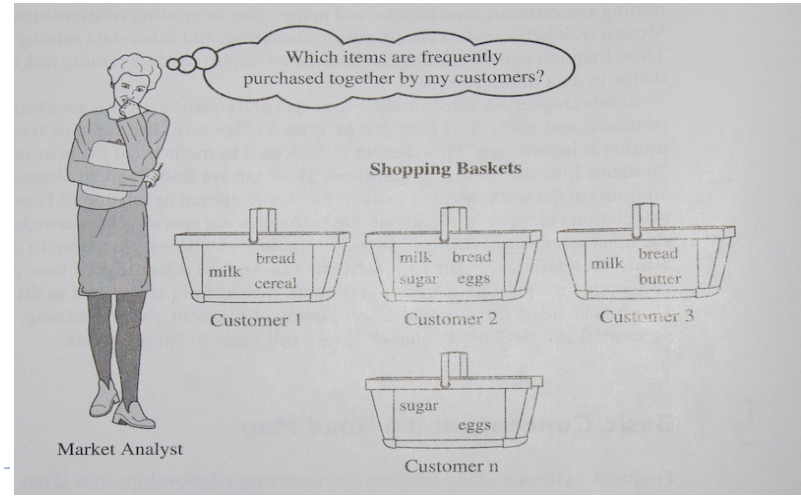


Association Rule คือ

- การทำเหมืองข้อมูลแบบกฎเชื่อมโยง คือ การหารูปแบบที่เกิดบ่อย (frequent pattern) ความเชื่อมโยงที่เกิดขึ้น (association) หรือสหสัมพันธ์ (correlation) ของกลุ่ม item จากข้อมูลที่อยู่ในรูป transaction
- นำไปใช้กับ
  - การวิเคราะห์ตะกร้าการซื้อ (Market basket analysis) การทำตลาดข้ามสายผลิตภัณฑ์ (cross-marketing) การวางรูปแบบและออกแบบ catalog การทำรายการส่งเสริมการขาย การวางแผนผังภายในร้าน เป็นต้น
- ผลลัพธ์อยู่ในรูปแบบกฎ :"Body → Head [support, confidence]"
  - Friday ∩ Nappies → Beer [0.5% , 60%]
  - Age ∈ [20,30] ∩ Income ∈ [20k, 30k] → MP3Player [2% , 60%]
  - Gladiator ∩ Patriot → Sixth Sense [0.1% , 90%]

คำถามของ Market basket

- ผู้ซื้อปฏิเสธหรือยอมรับการพยายามขายสินค้าต่างลักษณะหรือไม่?
- สินค้าใดที่ปรากฏร่วมกันบ่อยในหนึ่งใบเสร็จ?
- สินค้าใดที่ปรากฏร่วมกันบ่อยในกลุ่มสินค้าที่กำหนด?
- อะไรคือสินค้าที่พบว่าถูกขายบ่อยให้กับลูกค้าที่มารับบริการซ้ำๆ ?
- ลักษณะการขายของสินค้ามีการเปลี่ยนไปตามเวลาหรือไม่?
- ลักษณะการขายของสินค้ามีการเปลี่ยนไปตามที่อยู่ของลูกค้าหรือไม่ อย่างไร?



Association rule concept

- กฎเชื่อมโยง คือการวัดค่าสิ่งที่น่าสนใจของ
  - Support : รายการเหล่านี้เกิดขึ้นด้วยกันบ่อยแค่ไหน (ซื้อ x แล้วซื้อ y ด้วย)
  - Confidence : เงื่อนไขเหล่านี้เกิดขึ้นบ่อยแค่ไหน (บางที่ซื้อ xแต่ซื้อ m,หรือ bหรือ a)
- $X \rightarrow Y$ [s% , c%]
  - กฎนี้จะต้องมี s% ทุก transactions
  - Support $(X \rightarrow Y) = P(X \cup Y)$
  - ถ้าหาก X อยู่ในตระกร้าซื้อ แล้วกรณีที่มี Y ด้วยมี c% อยู่เท่าไหร่เมื่อเทียบกับทั้งหมด
  - Confidence $(X \rightarrow Y) = P(Y|X) = P(X \cup Y) / P(X)$

ตัวอย่างการหา association rule

- เซตของกลุ่มสินค้าที่มีค่าสนับสนุน **min support** เรียกว่า **frequent itemset**
- พิจารณากฎ A → C
- คุณ

Transaction	Items
12345	A B C
12346	A C
12347	A D
12348	B E F

⇒

Itemset	Support
A	75%
B	50%
C	50%
A, C	50%

iset

Parameters:
support = 50%
confidence = 50%

Rule $A \rightarrow C$

- support(A, C) = 50%
- confidence$(A \rightarrow C)$ = support(A, C)/support(A) = 66.6%

=50 / 75

การหา frequent itemset

- **frequent itemset** คือกลุ่มของ **item** ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่า **minimum support**
  - สับเซตของ **frequent itemset** เป็น **frequent itemset** คือ ถ้า **{A, B}** เป็น **frequent itemset** แล้ว **{A}** และ **{B}** ต้องเป็น **frequent itemset**
  - ดังนั้นการหา **frequent itemset** ที่มี **k** ตัวทำได้จากการพิจารณา **k-1** ตัว กล่าวคือเราหา **frequent itemset** จาก **1** ไปถึง **k (k-itemset)**
- หลังจากที่พิจารณาทุก **frequent itemset** ที่มีจำนวนมากที่สุดแล้ว เราสร้างกฎเชื่อมโยงจาก **frequent itemset** ทั้งหมดที่ได้
- ดูขั้นตอนของวิธี **Apriori** ในหน้าถัดไป

การทำงานของขั้นตอนวิธี Apriori (ค่อยๆ ตัดออก)

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D →

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

→ L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D ←

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

← L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D →

L_3

itemset	sup
{2 3 5}	2

ลักษณะข้อมูลที่ต้องการของ Association rule

- กฎเชื่อมโยง (Association rule) เป็นการวิเคราะห์หาความสัมพันธ์ที่มากกว่าหนึ่งสิ่งที่มีการบันทึกร่วมกันในหนึ่งแถวของข้อมูล
- ข้อมูลที่นำมาใช้วิเคราะห์ มักจะเป็น transaction นั่นคือ 1 record จะมีความสัมพันธ์ของลักษณะประจำ (ตัวแปร) หนึ่งรูปแบบ
- การวิเคราะห์รูปแบบนี้ไม่มีลักษณะประจำเป้าหมาย
- ลักษณะประจำ ต้องเป็นตัวแปร binary หรือตัวแปรที่มีค่าไม่ต่อเนื่อง
- ในกรณีที่ลักษณะประจำมีค่าที่ต่อเนื่อง จะต้องแปลงให้เป็นค่าไม่ต่อเนื่องก่อนการคำนวน อาจใช้คน หรือการจัดการข้อมูลด้วยโปรแกรมคอมพิวเตอร์อัตโนมัติก็ได้

สรุป Associate rule

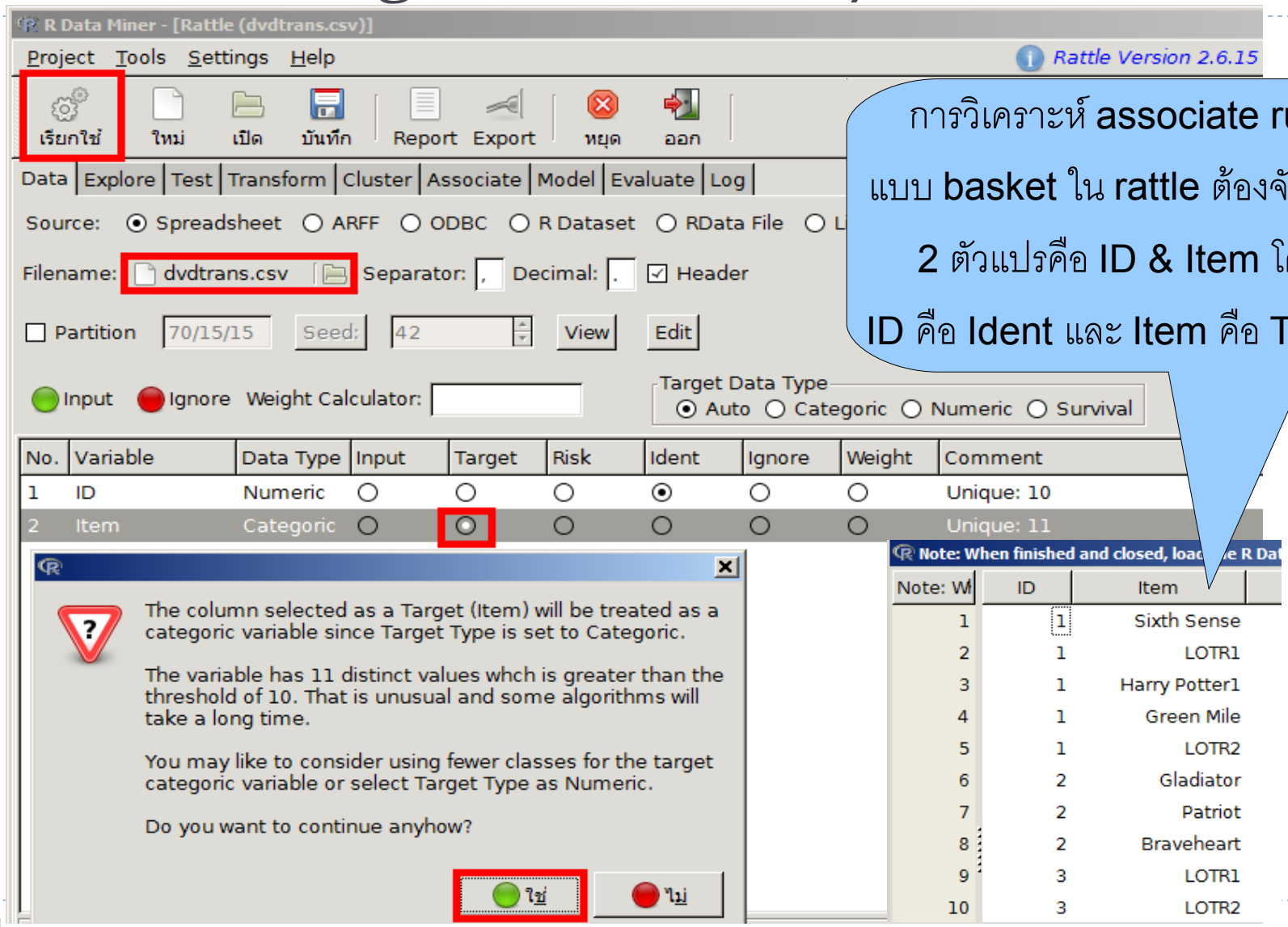
- **Associate rule** คือจุดเริ่มต้นการเกิดขึ้นของ **data mining algorithm!**
- ปรากฏในหมู่ของนักวิจัย KDD และมีการนำเสนองานส่วนนี้จำนวนมากที่สุด
- มีประโยชน์ในการหาความเชื่อมโยง บอกพฤติกรรมที่มีค่า ในฐานข้อมูลลูกค้าขนาดใหญ่
- ปกติใช้กับฐานข้อมูลเชิงสัมพันธ์ที่บันทึกเป็น **Transaction** โดยที่แต่ละระเบียนคือการซื้อสินค้าในหนึ่งครั้ง
- ผลลัพธ์ที่ได้คือกฎ กฎแสดงความสัมพันธ์ของการซื้อสินค้าต่างชนิดกันโดยไม่ขึ้นกับลูกค้า
-

Transaction	Items
12345	A C D
12346	B C E
12347	A B C E
12348	B E

$$BC \rightarrow E \ [50\%, 100\%]$$
$$BE \rightarrow C \ [50\%, 66\%]$$
$$C \rightarrow A \ [50\%, 66\%]$$

จงแสดงผลการคำนวน Support& confidence

Rattle Training 8.Association Analysis



การวิเคราะห์ associate rule แบบ basket ใน rattle ต้องจัดเป็น 2 ตัวแปรคือ ID & Item โดย ID คือ Ident และ Item คือ Target

Rattle Training 8.Association Analysis

R Data Miner - [Rattle (dvdtrans.csv)]

Project Tools Settings Help — Rattle Version 2.6.15

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

☑ Baskets | Freq Plot | Support: 0.1000 | Confidence: 0.1000 | Show Rules | Plot

```
Summary of the Transactions:

      Length        Class         Mode
          10 transactions          S4

Summary of the Apriori Association Rules:

Number of Rules: 130

Summary of the Measures of Interestingness:

    support         confidence          lift
 Min.   :0.1000   Min.   :0.1000   Min.   : 0.8333
 1st Qu.:0.1000   1st Qu.:0.5000   1st Qu.: 1.6667
 Median :0.1000   Median :1.0000   Median : 2.5000
 Mean   :0.1415   Mean   :0.7769   Mean   : 3.5120
 3rd Qu.:0.1000   3rd Qu.:1.0000   3rd Qu.: 5.0000
 Max.   :0.6000   Max.   :1.0000   Max.   :10.0000

Summary of the Execution of the Apriori Command:

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target
        0.1    0.1    1 none FALSE            TRUE     0.1      1     10  rules
   ext
```

The Association Rules model has been built. Time taken: 0.09 secs

Hahsler พัฒนา apriori for R package เพื่อให้ผู้สนใจได้ใช้ฟรี โดยในอดีต apriori นั้นเป็นส่วนหนึ่งในซอฟท์แวร์ Clementine. (Hahsler เป็นผู้พัฒนา apriori ตั้งแต่แรก)

Rattle Training 8.Association Analysis

R Data Miner - [Rattle (dvdtrans.csv)]

Project Tools Settings Help Rattle Version 2.6.15 togaware.com

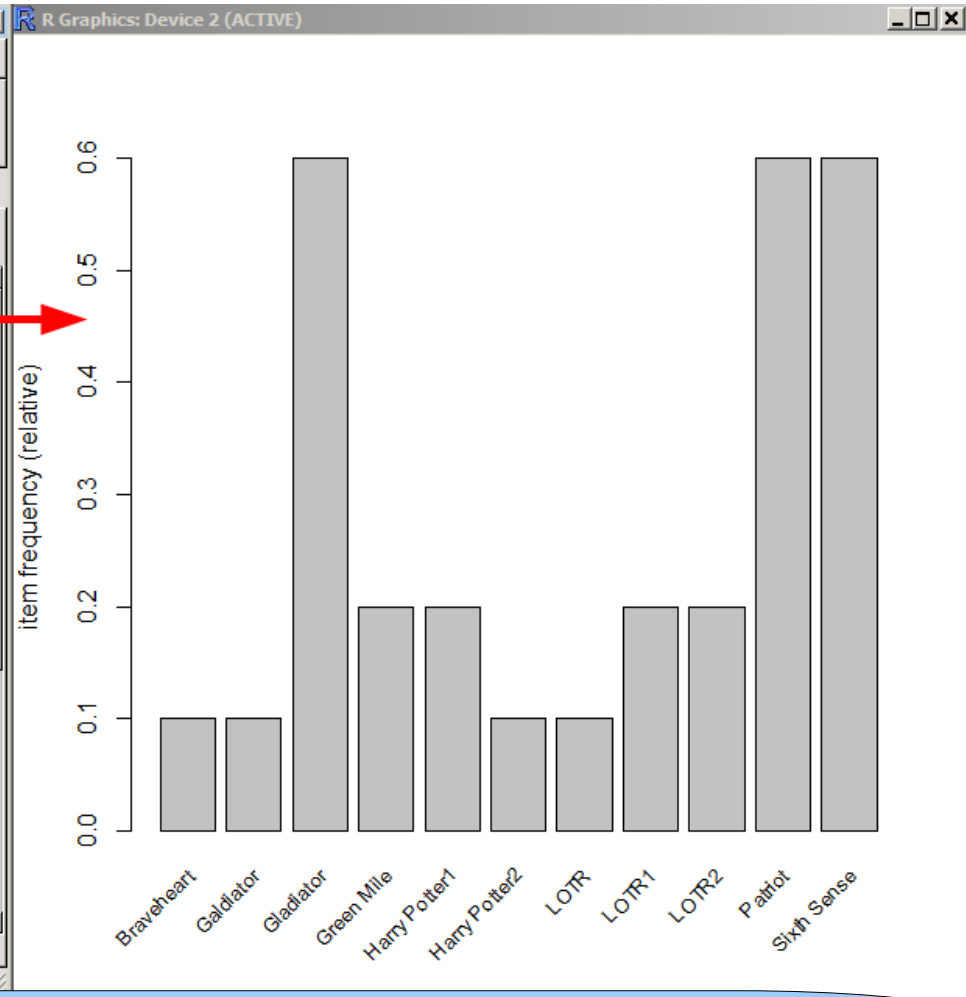
เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

☑ Baskets Freq Plot Support: 0.1000 Confidence: 0.1000 Show Rules Plot

```
Summary of the Transactions:

      Length        Class         Mode
          10 transactions           S4

Summary of the Apriori Association Rules:

Number of Rules: 130

Summary of the Measures of Interestingness:

    support         confidence          lift
 Min.   :0.1000   Min.   :0.1000   Min.   : 0.8333
 1st Qu.:0.1000   1st Qu.:0.5000   1st Qu.: 1.6667
 Median :0.1000   Median :1.0000   Median : 2.5000
 Mean   :0.1415   Mean   :0.7769   Mean   : 3.5120
 3rd Qu.:0.1000   3rd Qu.:1.0000   3rd Qu.: 5.0000
 Max.   :0.6000   Max.   :1.0000   Max.   :10.0000

Summary of the Execution of the Apriori Command:

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen
        0.1    0.1    1 none FALSE            TRUE     0.1      1     10
   ext
 FALSE

algorithmic control:
```

Generated the relative frequency plot.



เป็นการพิจารณาดูว่าข้อมูล **Item** มีจำนวนกลุ่มมากน้อยเพียงใด

Rattle Training 8.Association Analysis

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
> library(arules)     # attach package 'arules'
Loading required package: Matrix
Loading required package: lattice

Attaching package: 'Matrix'

The following object(s) are masked from 'package:base':

    det


Attaching package: 'arules'

The following object(s) are masked fr
```

Building a Model Using R Command-Line

Arules (Hahsler et al.,2011) Provides apriori()

These R Packages are widely used, and freely

Its

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
> library(arulesViz)  # attach package 'arulesVis'
Loading required package: MASS
Loading required package: scatterplot3d
Loading required package: vcd
Loading required package: grid
Loading required package: colorspace
Loading required package: seriation
Loading required package: cluster
Loading required package: TSP
Loading required package: gclus

Attaching package: 'seriation'
```

associate rule ใน Command-line

สามารถทำกราฟ ได้ดีกว่า Rattle

เราจึงแนะนำให้ใช้วิธีดังต่อไปนี้

Rattle Training 8.Association Analysis

dvd=read.transactions (file="dvdtrans.csv",rm.duplicates=FALSE, format="single",sep=",",cols=c(1,2))

```
> dvd = read.transactions (file="dvdtrans.csv", rm.duplicates= FALSE, format="single",sep=",",cols =c(1,2))
> dvd
transactions in sparse format with
 11 transactions (rows) and
 14 items (columns)
>
> # Run the apriori algorithm
> basket_market <- apriori(dvd, parameter = list (sup = 0.1, conf = 0.1, target="rules"))

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.1    0.1    1 none FALSE            TRUE     0.1      1     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Warning in apriori(dvd, parameter = list(sup = 0.1, conf = 0.1, target = "rules")) :
  You chose a very low absolute support count of 1. You might run out of memory! Increase minimum support.

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[14 item(s), 11 transaction(s)] done [0.00s].
sorting and recoding items ... [6 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [17 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

วิเคราะห์ aprori

ดึงข้อมูลเข้าหน่วยความจำของ R

basket_market <-apriori(dvd,parameter=list (sup=0.1, conf=0.1, target="rules"))

อย่าลืม Change directory เพื่อความสะดวกในการ access ข้อมูลด้วย R Command-line.

Rattle Training 8. Association Analysis

- Inspect (...) จะแสดงกฎความรู้ทั้งหมดที่ค้นพบ

```
> # Check the generated rules using inspect
> inspect(basket_market)
   lhs              rhs                support confidence     lift
1  {}            => {Harry Potter1} 0.1818182  0.1818182 1.000000
2  {}            => {LOTR1}         0.1818182  0.1818182 1.000000
3  {}            => {LOTR2}         0.1818182  0.1818182 1.000000
4  {}            => {Sixth Sense}   0.4545455  0.4545455 1.000000
5  {}            => {Gladiator}     0.5454545  0.5454545 1.000000
6  {}            => {Patriot}       0.5454545  0.5454545 1.000000
7  {LOTR1}       => {LOTR2}         0.1818182  1.0000000 5.500000
8  {LOTR2}       => {LOTR1}         0.1818182  1.0000000 5.500000
9  {Sixth Sense} => {Gladiator}     0.3636364  0.8000000 1.466667
10 {Gladiator}   => {Sixth Sense}   0.3636364  0.6666667 1.466667
11 {Sixth Sense} => {Patriot}       0.3636364  0.8000000 1.466667
12 {Patriot}     => {Sixth Sense}   0.3636364  0.6666667 1.466667
13 {Gladiator}   => {Patriot}       0.5454545  1.0000000 1.833333
14 {Patriot}     => {Gladiator}     0.5454545  1.0000000 1.833333
15 {Gladiator,
    Sixth Sense} => {Patriot}       0.3636364  1.0000000 1.833333
16 {Patriot,
    Sixth Sense} => {Gladiator}     0.3636364  1.0000000 1.833333
17 {Gladiator,
    Patriot}     => {Sixth Sense}   0.3636364  0.6666667 1.466667
>
```

```
> inspect(dvd)
   items                    transactionID
1  {Green Mile,
    Harry Potter1,
    LOTR1,
    LOTR2,
    Sixth Sense}            1
2  { Galdiator,
     Green Mile,
     LOTR,
     Sixth Sense}           10
3  {Braveheart,
    Gladiator,
    Patriot}                2
4  {LOTR1,
    LOTR2}                  3
5  {Gladiator,
    Patriot,
    Sixth Sense}            4
6  {Gladiator,
    Patriot,
    Sixth Sense}            5
7  {Gladiator,
    Patriot,
    Sixth Sense}            6
8  {Harry Potter1,
    Harry Potter2}          7
9  {Gladiator,
    Patriot}                8
10 {Gladiator,
    Patriot,
    Sixth Sense}            9
11 {Item}                   ID
>
```

Rattle Training 8.Association Analysis

File Edit View Misc Packages Windows Help

```
> #If huge number of rules are generated specific rules can read using index
> inspect(basket_market[1])
  lhs    rhs                support confidence lift
1 {}  => {Harry Potter1} 0.1818182  0.1818182    1

> inspect(basket_market[5])
  lhs    rhs            support confidence lift
1 {}  => {Gladiator} 0.5454545  0.5454545    1
> inspect(basket_market[10])
  lhs            rhs              support confidence     lift
1 {Gladiator} => {Sixth Sense} 0.3636364  0.6666667 1.466667
>
```

เราสามารถระบุว่าจะดูข้อมูล apriori ในลำดับกฎที่เท่าไหร่ โดยระบุต่อท้ายได้

Rattle Training 8. Association Analysis

- คำสั่ง **sort** กฎที่ค้นพบตามค่า **lift** และคัดเลือกเฉพาะตัวแรกๆ เช่น **5** ลำดับแรก
- ทำให้กฎที่ได้ มีหน้าตาคล้ายกับการแสดงผลใน **Data mining software** ทั่วไป

```
> ## select and inspect rules with highest lift.
> rules_high <- head(sort(basket_market, by = "lift"),5)
> inspect(rules_high)
  lhs                 rhs            support confidence     lift
1 {LOTR1}          => {LOTR2}      0.1818182          1 5.500000
2 {LOTR2}          => {LOTR1}      0.1818182          1 5.500000
3 {Gladiator}      => {Patriot}    0.5454545          1 1.833333
4 {Patriot}        => {Gladiator}  0.5454545          1 1.833333
5 {Gladiator,
   Sixth Sense}    => {Patriot}    0.3636364          1 1.833333
>
>
```

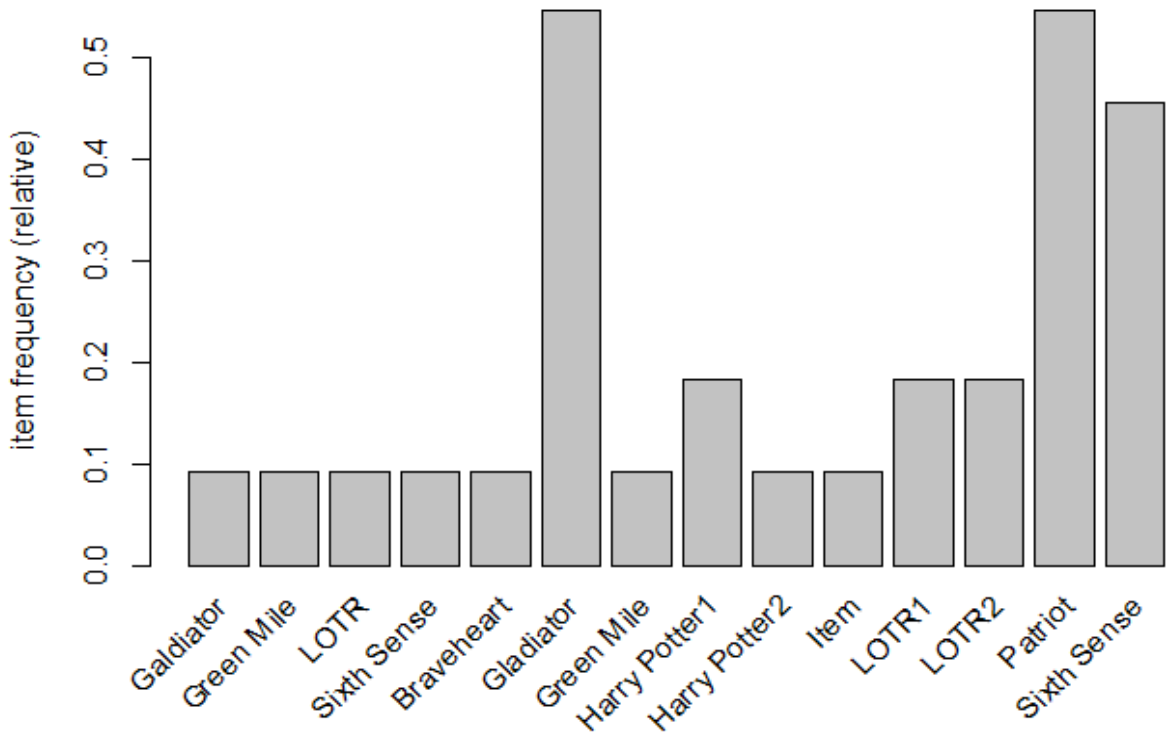
Rattle Training 8. Association Analysis

R Console

```
> #To visualize the item frequency in dvd file
> itemFrequencyPlot(dvd)
```

ItemFrequencyPlot (dvd)

R Graphics: Device 2 (ACTIVE)



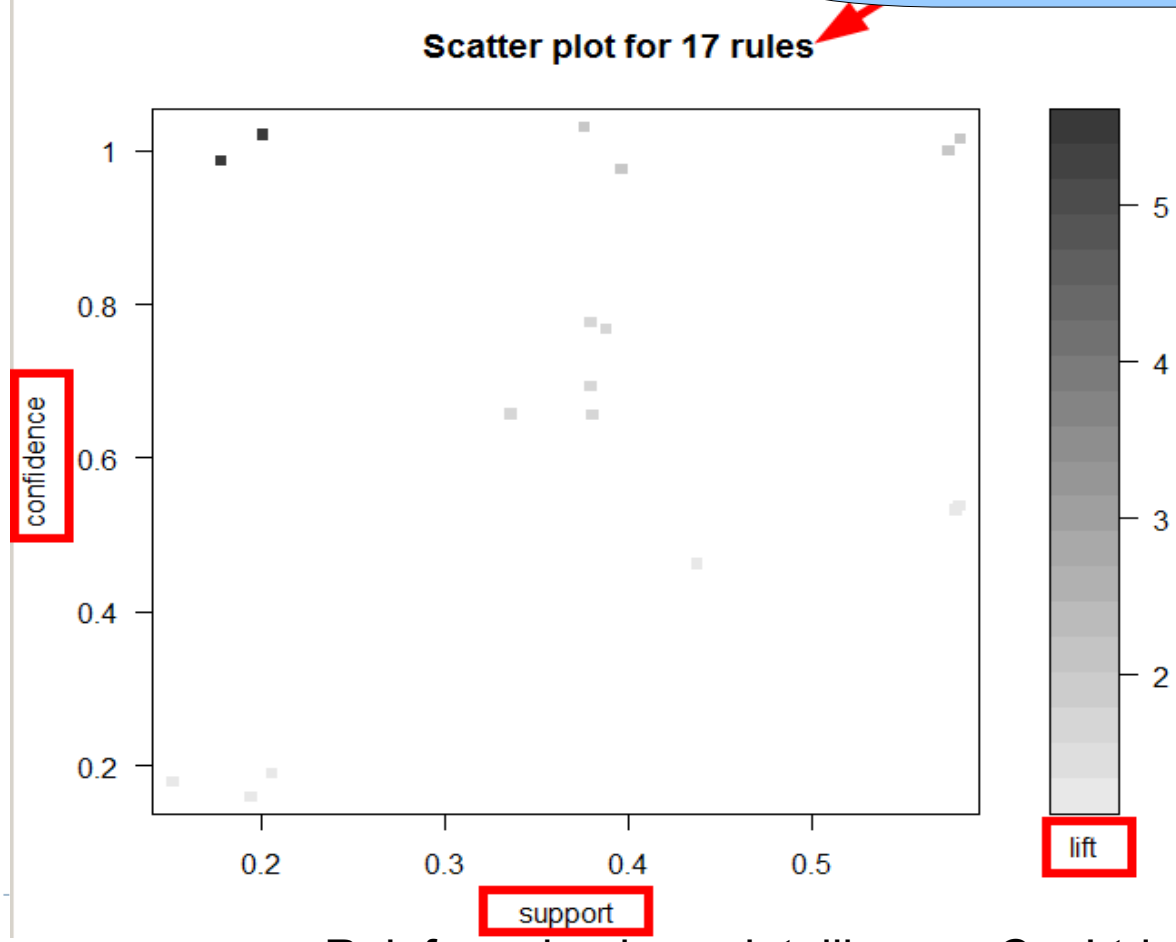
Rattle Training 8. Association Analysis

R Console

```
> #To visualize the item frequency in dvd file
> itemFrequencyPlot(dvd)
> plot(basket_market, control = list (jitter = 2))
```

plot(basket_market, control = list (jiter=2))

R Graphics: Device 2 (ACTIVE)



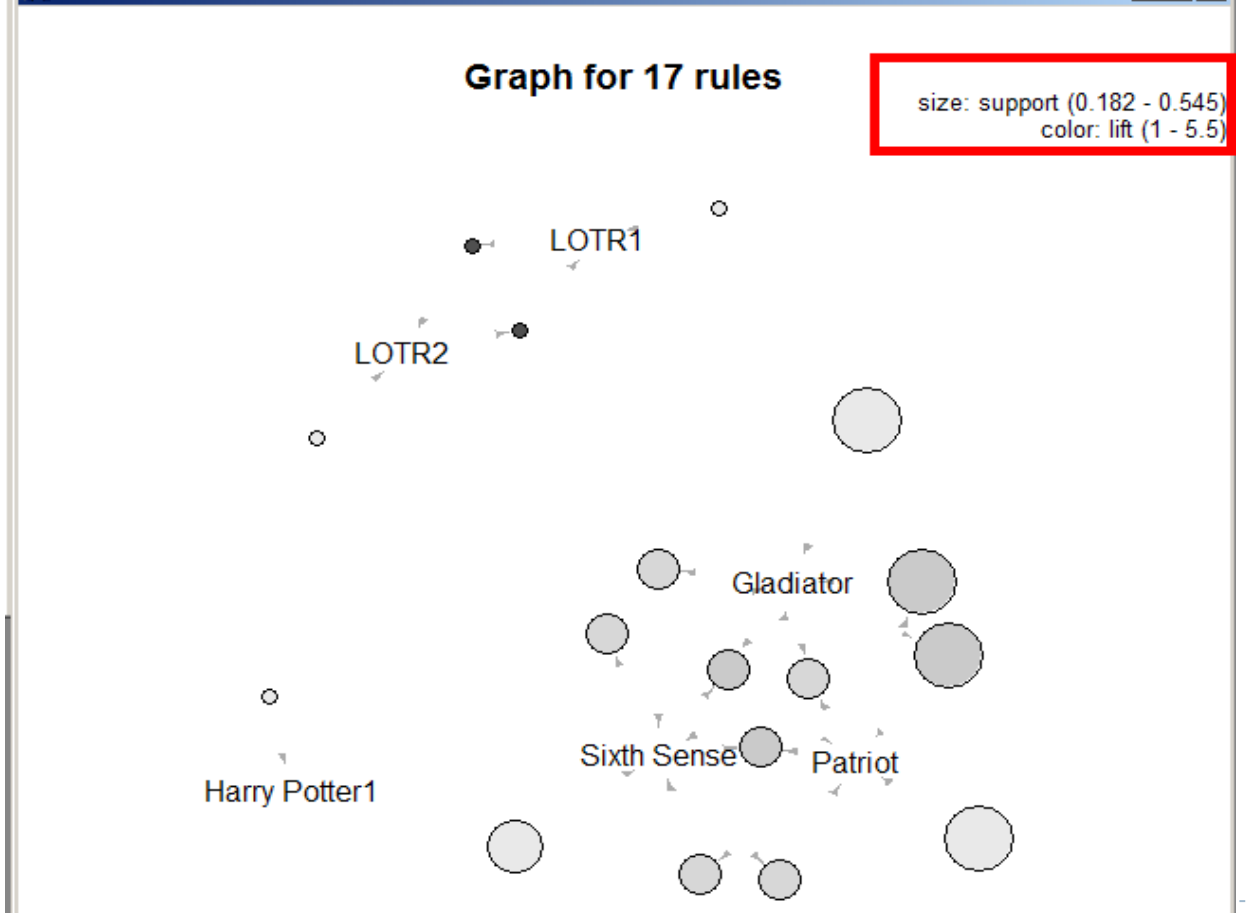
Rattle Training 8. Association Analysis

R Console

```
> #To visualize the item frequency in dvd file
> itemFrequencyPlot(dvd)
> plot(basket_market, control = list(jitter = 2))
> plot(basket_market, method="graph",control=list(type="items"))
```

plot(basket_market, method="graph", control=list(type="items"))

R Graphics: Device 2 (ACTIVE)



Rattle Training 8.Association Analysis

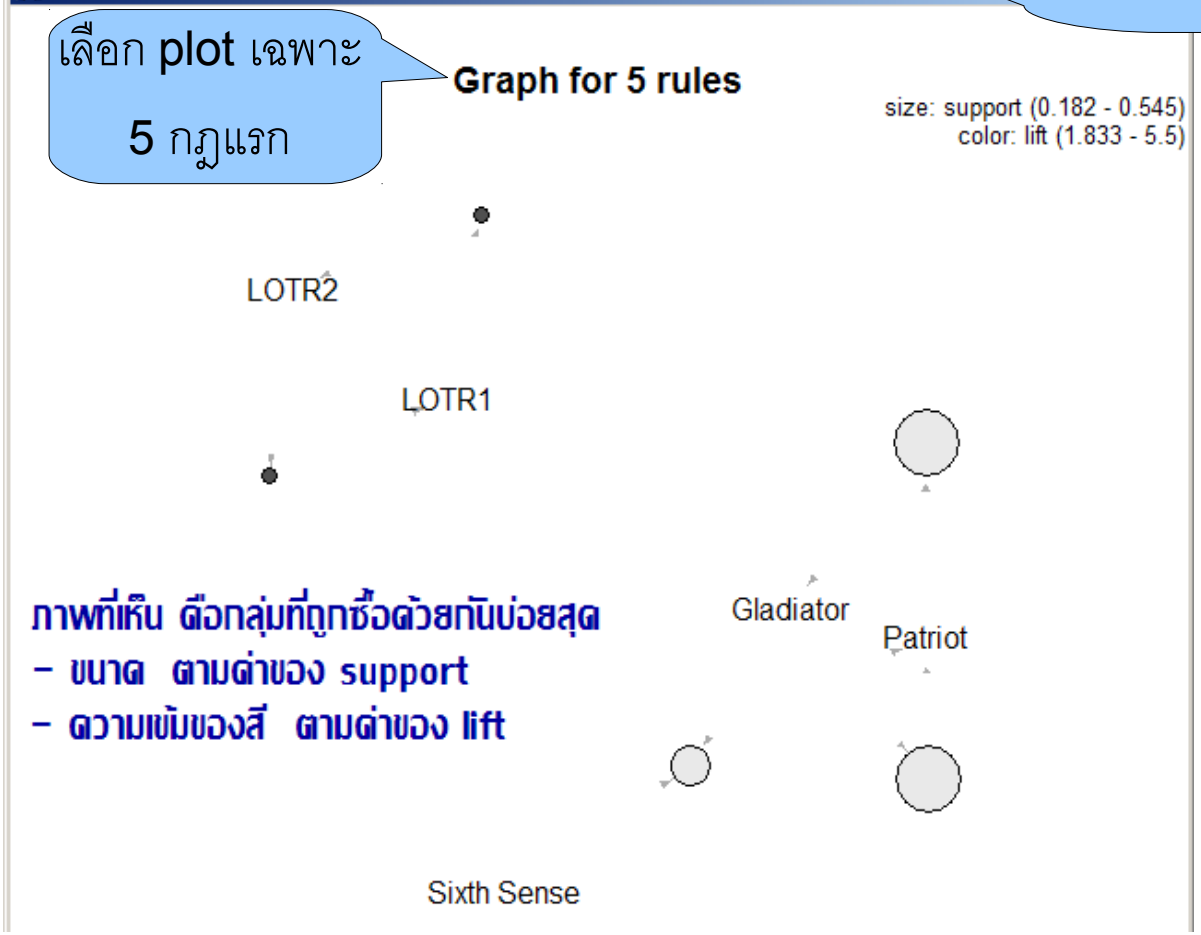
R Console

```
> plot(rules_high, method="graph",control=list(type="items"))
```

plot (rules_high, method="graph", control= list(type="items"))

R Graphics: Device 2 (ACTIVE)

เลือก plot เฉพาะ 5 กฎแรก



ภาพที่เห็น คือกลุ่มที่ถูกซื้อด้วยกันบ่อยสุด
- ขนาด ตามค่าของ support
- ความเข้มของสี ตามค่าของ lift

9.Classification Model

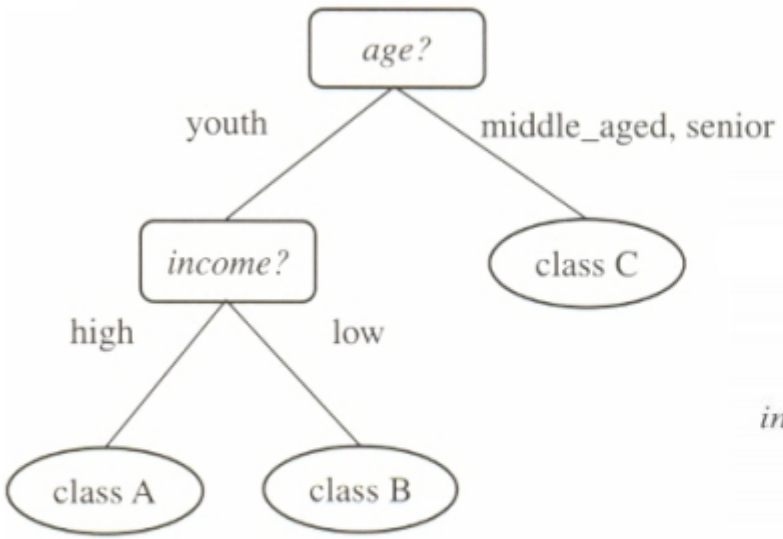
- Classification and Prediction
  - Finds a set of models to predict the class of objects
  - The models is based on a set of training data

age(X, "youth") AND income(X, "high")	→	class(X, "A")
age(X, "youth") AND income(X, "low")	→	class(X, "B")
age(X, "middle_aged")	→	class(X, "C")
age(X, "senior")	→	class(X, "C")

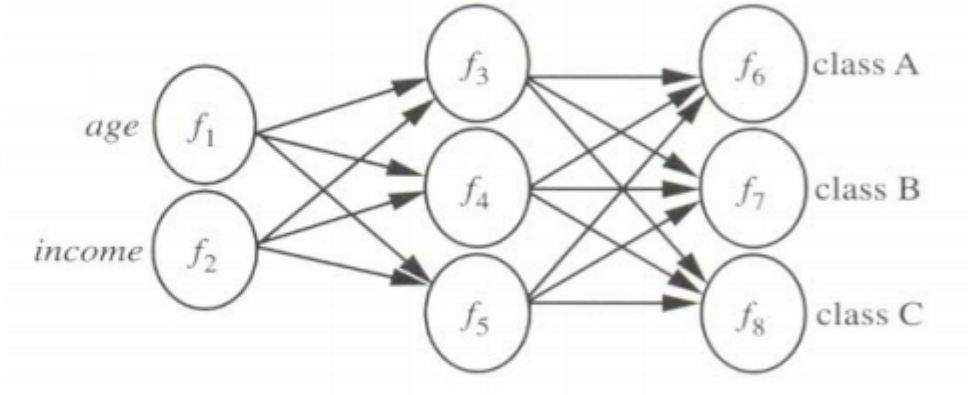
The set of classification rule model

Data Mining functionality

- Classification and Prediction



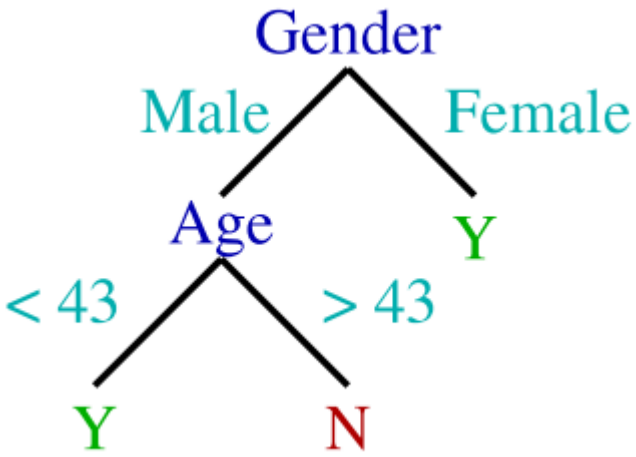
decision tree model



neural network model

Decision tree → Classification

- วัตถุประสงค์ของ **classification** คือการสร้างโมเดล เพื่อสรุปข้อมูล และข้อค้นพบที่ได้มักเป็นข้อมูลที่ไม่เคยเห็นมาก่อน
- **Classification** เป็นการเรียนรู้แบบ **supervised**
- นอกจาก **decision tree** แล้วยังมี **neural networks ; logistic regression : support vector machines** เป็นประเภทเดียวกัน
- สิ่งที่ได้ของ **decision tree** คือ **flow-chart**
  - มี **node** จุดยอดภายใน แทนลักษณะประจำที่ไม่ใช่ **class**
  - มีกิ่งก้านแทนผลลัพธ์ของการทดสอบ
  - และใบแทน **class labels** หรือ **class distribution**



Classification vs. Prediction

- **Classification**
  - เป็นวิธีการแยกข้อมูลออกเป็นกลุ่มตามลักษณะประจำเป้าหมาย หรือ **class** โดยมากใช้กับลักษณะประจำที่มีค่าไม่ต่อเนื่อง
  - ผลลัพธ์ที่ได้คือ ตัวแบบ **classifier** เพื่อนำไปใช้แยกประเภทของระเบียนข้อมูล ที่ไม่ทราบ **class**
- **Prediction**
  - เป็นวิธีการคำนวณหาค่าที่ต้องการทราบ เพื่อทำนายค่าของข้อมูล โดยมากใช้กับลักษณะประจำที่มีค่าต่อเนื่อง
  - ผลลัพธ์ที่ได้คือ ตัวแบบ **predictor** เพื่อนำไปใช้ประมาณค่าของระเบียนข้อมูล ที่สนใจว่าค่าที่ขาดหายไปจะเป็นเท่าไหร่ เช่น สมการ **multiple regression**

Decision tree model

- การจัดจำแนกด้วย **decision model** เป็น **machine learning** จากกลุ่มตัวอย่างที่กำหนดให้
- **Decision model** ที่ได้อาจอยู่ในรูปของกฎ ต้นไม้ ตารางหรือเครือข่าย
- การประเมินความถูกต้องของ **model**
  - การวัดความถูกต้องของ **decision tree model** ใช้เทคนิค **train-test set** หรือ **train-validate test set** ซึ่ง **model** จะยอมรับได้ก็ต่อเมื่อตัวอย่างที่ทำนายจาก **test set** มีความถูกต้องผ่านเกณฑ์ที่กำหนด
  - บางครั้งเราใช้ **decision tree model** กับหลักการ **train-test set** แล้วผลออกมาอย่างไม่มีข้อผิดพลาด แต่เมื่อนำไปใช้กับ **test set** กลับได้ผลไม่ดี ปรากฏการณ์นี้เรียกว่าปัญหาการเข้ากันเกินไป **(overfitting problem)**
  - วิธีการหลีกเลี่ยงปัญหาดังกล่าว คือการใช้ข้อมูลอีกชุด **validate set** เพื่อให้ได้ตัวแบบที่นำไปใช้กับตัวอย่างในอนาคตได้ดี

แนวคิดการสร้าง Decision tree

- ใช้หลักการละโมบ (A greedy algorithm) เพื่อหาแนวทางที่ดีที่สุด ที่มีอยู่
- สร้างต้นไม้ด้วยวิธี top-down , recursive, divide-and-conquer
  - เริ่มต้นทุกตัวอย่างที่ราก
  - ใช้กับลักษณะประจำที่ไม่ต่อเนื่อง categorical
  - แบ่งกั้นตัวอย่างตามค่าลักษณะประจำโดยการเรียกซ้ำ
  - ลักษณะประจำถูกเลือกจากลำดับของตัววัด เช่น information gain
- เงื่อนไขในการหยุดแบ่ง
  - ตัวอย่างทุกตัวมีค่า class เหมือนกันหมด
  - ไม่มีลักษณะประจำเหลือในการแบ่ง บอกค่า class โดยใช้ majority voting
  - ไม่มีตัวอย่างเหลือ ให้ใช้ค่า class ของจุดยอดที่อยู่เหนือขึ้นไป

ตัววัดที่ใช้เลือกลักษณะประจำ

- Information gain (ID3)
  - ใช้กับลักษณะประจำที่เป็น categorical
  - ลักษณะประจำที่เป็นค่าต่อเนื่องให้แปลงเป็นค่าไม่ต่อเนื่องก่อน
- Gain ration (C4.5)
  - ได้จากค่า information gain หารด้วยค่าสรสนเทศของลักษณะประจำ
- Information gain (ID3/C4.5)
- เลือกลักษณะประจำที่มีค่า information gain สูงที่สุด
- พิจารณากลุ่มข้อมูลที่แบ่งเป็น 2 class คือ P หรือ N
  - จำนวนตัวอย่างใน class P คือ p ตัวและจำนวนตัวอย่างใน class N คือ n ตัว
  - ค่าสารสนเทศของกลุ่มข้อมูลคือค่าคาดคะเนที่กลุ่มตัวอย่าง ต้องใช้จำนวนบิตในการแยก class P และ N นิยามโดย

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

การคำนวณค่า Information gain

กำหนด A คือลักษณะประจำที่แบ่ง S ออกเป็น v cells: $\{S_1, S_2, \ldots S_V\}$

ถ้า S_i มีตัวอย่างจาก class P จำนวน p_i และตัวอย่างจาก class N จำนวน ni, ค่า entropy หรือค่าคาดคะเนของข้อมูลที่แยกโดยการใช้ลักษณะประจำ A คือ

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$Gain(A) = I(p, n) - E(A)$$

- ค่า Information gain ที่ได้จากการแยกข้อมูลจากลักษณะประจำ A

การเลี่ยงปัญหา Overfitting problem

- **Decision tree** ที่มีปัญหา **overfitting** มักเป็นต้นไม้ที่มีกิ่งก้านแตกไปมากมาย แต่ละกิ่งมีข้อมูลอยู่ปริมาณน้อย
- วิธีการเลี่ยงปัญหา **overfitting** ใน **decision tree** ทำได้สองลักษณะ
  - การตัดแต่งกิ่งระหว่างสร้างต้นไม้ : หยุดการสร้างต้นไม้เมื่อตัววัดในการแตกกิ่งมีค่าต่ำกว่าเกณฑ์ที่กำหนด เช่นจำนวนข้อมูลในจุดยอดน้อยจนเกินไป
  - การตัดแต่งกิ่งหลังสร้างต้นไม้ : การลดจำนวนกิ่ง กำจัดกิ่งโดยยังคงความถูกต้องของการจำแนกประเภทในระดับที่ยอมรับได้
    - วิธีการใช้วิธีแบ่ง **validation data** เพื่อใช้ทดสอบการตัดแต่งกิ่งเลือกต้นไม้ที่ดีที่สุด "**best pruned tree**" สำหรับข้อมูล **validation data**

ลักษณะข้อมูลที่ต้องการของการวิเคราะห์ Classification

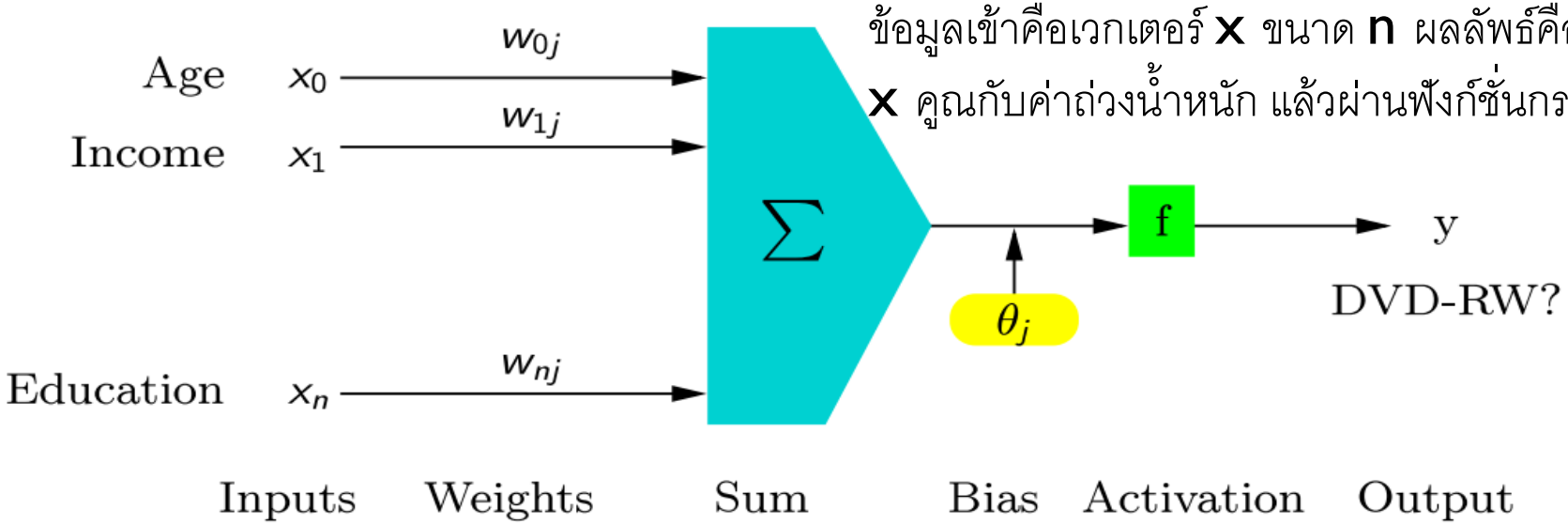
หมายรวมถึง Decision Tree, Naïve Bayesian และ Neural Network

- การจัดจำแนกประเภท (Classification) เป็นการสร้าง model เพื่อใช้จำแนก records ที่พบในอนาคตว่าจัดอยู่ในประเภทใด
- ข้อมูลที่ใช้ต้องมีลักษณะประจำเป้าหมาย (Class) ที่มีค่าไม่ต่อเนื่อง ส่วนมากเป็น binary แต่ถ้าเป็นตัวแปร nominal มากกว่า 2 ค่าก็สามารถวิเคราะห์ข้อมูลได้
- บาง algorithm เช่น ID3 ต้องการลักษณะประจำนำเข้าที่มีค่าไม่ต่อเนื่องเท่านั้น และบาง algorithm เช่น Logistic regression ต้องการลักษณะประจำที่มีค่าต่อเนื่องเท่านั้น
- แต่ algorithm ในปัจจุบันสามารถใช้ได้ทั้งลักษณะประจำที่มีค่าไม่ต่อเนื่อง หรือค่าต่อเนื่องก็ได้ จึงได้รับความนิยมมากขึ้น

Neural Networks

- เป็นแนวคิดที่เทียบเคียงกันระหว่างโครงสร้างข้อมูล กับการเรียนรู้ของระบบประสาทในสมอง
  - สมองมนุษย์มีเซลล์ประสาทจำนวนมาก ที่รับสัญญาณกระตุ้นแล้วตัดสินใจส่งสัญญาณแปรผลออกไป คล้ายๆ กับมามองดูว่าลักษณะเชิงประชากรศาสตร์ของลูกค้าเป็นแบบใด ก็จะมี **model** มาตัดสินใจให้ทราบว่า สินค้าที่ต้องการซื้อคืออะไร....
- ข้อมูลเข้าคือเวกเตอร์ **x** ขนาด **n** ผลลัพธ์คือ **y**

  x คูณกับค่าถ่วงน้ำหนัก แล้วผ่านฟังก์ชั่นกระตุ้น



$$y = f(w_{0j}x_0 + w_{1j}x_1 + \ldots + w_{nj}x_n + \theta_j)$$

Multi-Layer Perceptron

- Activation Functions

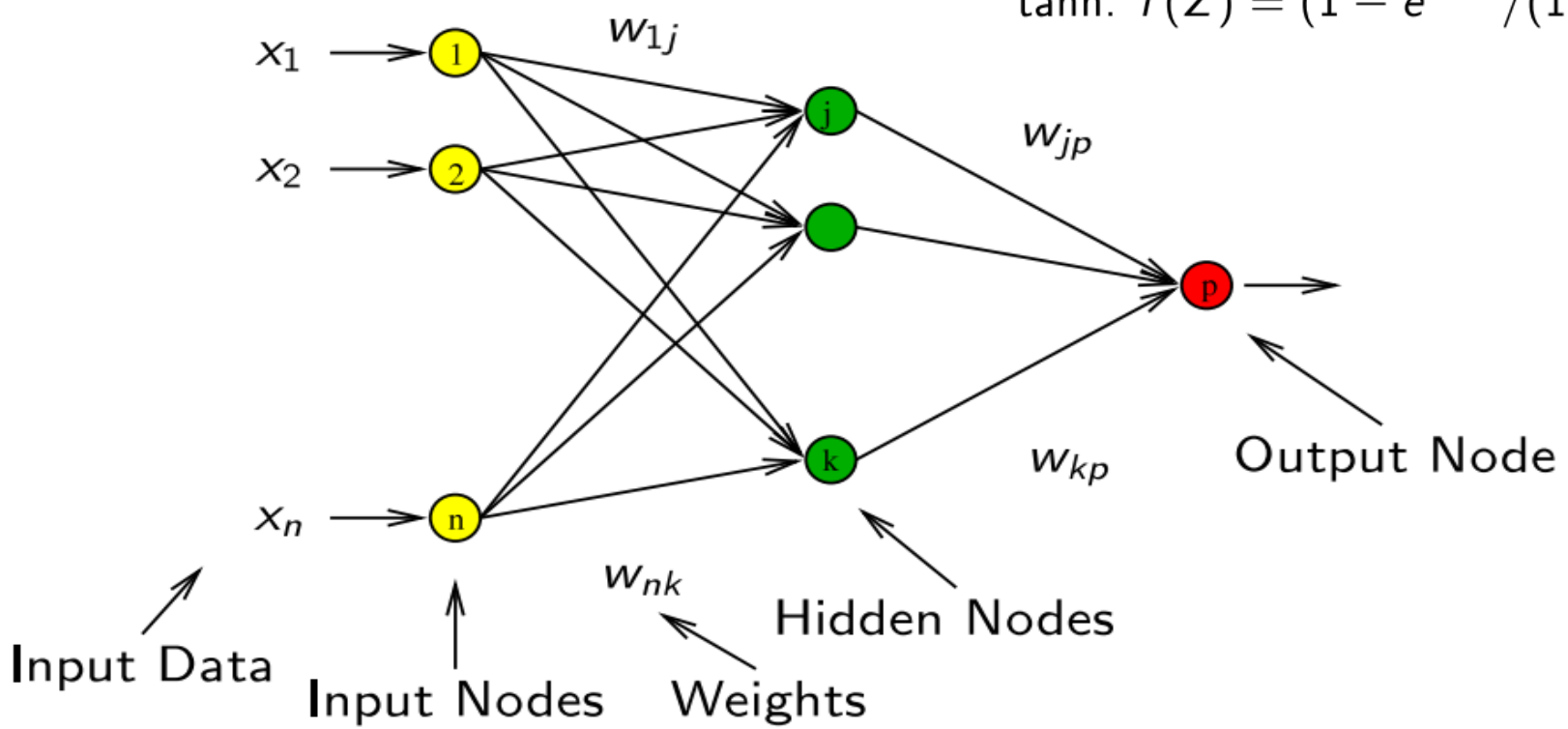
threshold

$Z > threshold : f(Z) = 1$

$Z \leq threshold : f(Z) = 0$

sigmoid: $f(Z) = 1/(1 + e^{-Z})$

tanh: $f(Z) = (1 - e^{-2Z}/(1 + e^{-2Z}))$



การสร้าง Neural Networks

- สร้าง network
  - นำเข้าข้อมูล เลือกจำนวนของ layers, จำนวน nodes ในแต่ละ layer
- Training the network ด้วย training data
  - เริ่มกำหนดค่าถ่วงน้ำหนักอย่างสุ่ม
  - วนซ้ำกับชุดของตัวอย่างทีละรอบ (Epoch)
  - คำนวณหาผลลัพธ์ที่ได้จากการใช้ network โดยคำนวณแต่ละหนึ่งหน่วยสมอง (neuron) คำนวณค่าผลรวมถ่วงน้ำหนักของข้อมูลเข้าฟังก์ชั่นกระตุ้น (activation)
  - ถ้าค่าที่คำนวณผิดจากผลลัพธ์ที่ต้องการให้ปรับค่าถ่วงน้ำหนักให้เหมาะสม
  - ทำซ้ำจนกระทั่งครบตามเงื่อนไขการหยุด เช่น การเรียนรู้เริ่มไม่ดีขึ้น หรือทำซ้ำจนครบจำนวนการทำซ้ำสูงสุดที่กำหนด (Max epochs)

การปรับปรุง Neural Networks ให้ดีขึ้น

- กำจัดสายที่เชื่อมโยงระหว่างหนึ่ง **neuron** เพราะการใช้สายเชื่อมโยงทั้งหมดในเครือข่ายทำให้การคำนวณช้าลง กล่าวคือถ้าข้อมูลมีขนาด **N** และจำนวน **node** ใน **hidden layer** คือ **h** ซึ่งโยงไปจุดยอดผลลัพธ์ **m** ตัว เราต้องใช้ค่าถ่วงน้ำหนักทั้งหมด **h(m+N)** ค่า
- สังเกตว่าการมีจุดยอดใน **hidden layer** ที่มากจะทำให้การเรียนรู้นาน และยังก่อให้เกิดปัญหา **ovefitting problem** ด้วย

Inputs	I_j	$= \sum_i w_{ij} O_i + \theta_j$
Outputs	O_j	$= \frac{1}{1+e^{-I_j}}$
Output Error	Err_j	$= O_j(1 - O_j)(T_j - O_j)$
Node Errors	Err_j	$= O_j(1 - O_j) \sum_k Err_k w_{jk}$
Propagate	w_{ij}	$= w_{ij} + l \times Err_j \times O_i$
	θ_j	$= \theta_j + l \times Err_j$

Neural Networks

- ข้อดี
  - การทำนายให้ผลแม่นยำกว่าวิธีทั่วไป
  - Model ที่ได้ยังคงใช้ได้ดี แม้ว่าข้อมูลที่ใช้มี error อยู่ด้วย
  - ผลลัพธ์ที่ต้องการเป็นได้ทั้ง ค่าต่อเนื่องหรือค่าไม่ต่อเนื่องก็ได้
  - การหาผลลัพธ์ทำได้รวดเร็ว หลังจากผ่านกรรมวิธีการเรียนรู้แล้ว
- ข้อเสีย
  - ใช้เวลา training นาน (หาค่าถ่วงน้ำหนักที่ดีที่สุดในการเรียนรู้)
  - ยากที่จะทำความเข้าใจเกี่ยวกับการเลือกพารามิเตอร์ที่เหมาะสม (best topology)
  - ยากที่จะทำความเข้าใจเกี่ยวกับ learned function

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Tree Forest Boost SVM Linear Neural Net Surviva

Target: RainTomorrow Algorithm: Traditional Conditional Model Builder: rpart

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

Min Bucket: 7 Complexity: 0.0100 Loss Matrix: Rules Draw

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 256 41 No (0.83984375 0.16015625)
  2) Pressure3pm>=1011.9 204 16 No (0.92156863 0.07843137)
    4) Cloud3pm< 7.5 195 10 No (0.94871795 0.05128205) *
    5) Cloud3pm>=7.5 9  3 Yes (0.33333333 0.66666667) *
  3) Pressure3pm< 1011.9 52 25 No (0.51923077 0.48076923)
    6) Sunshine>=8.85 25  5 No (0.80000000 0.20000000) *
    7) Sunshine< 8.85 27  7 Yes (0.25925926 0.74074074) *

Classification tree:
rpart(formula = RainTomorrow ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] Cloud3pm    Pressure3pm Sunshine

Root node error: 41/256 = 0.16016
```

The Decision Tree model has been built. Time taken: 0.16 secs

ตัวแบบการจัดจำแนกประเภท (Classifier) จะอยู่ที่ Model

Model Tree ใน rattle นั้นคือ ID3 ซึ่งแต่เดิมอยู่ในซอฟท์แวร์เชิงพาณิชย์ที่ชื่อ C4.5

Rattle Training 9. Classifier Models

- คุณภาพของ **Model** ที่ได้ดูได้จาก **Root node error**
- ส่วนขนาดของ **cp** จะเป็นตัวกำหนดขนาดของ **decision tree** ที่ **optimal**

```
Classification tree:
rpart(formula = RainTomorrow ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] Cloud3pm    Pressure3pm Sunshine

Root node error: 41/256 = 0.16016

n= 256

        CP nsplit rel error  xerror    xstd
1 0.158537      0   1.00000 1.00000 0.14312
2 0.073171      2   0.68293 0.80488 0.13077
3 0.010000      3   0.60976 0.95122 0.14024

Time taken: 0.07 secs

Rattle timestamp: 2012-01-15 13:31:07 amto
=========================================================================
```

Rattle Training 9. Classifier Models

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

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

Target: RainTomorrow Algorithm: ⦿ Traditional ○ Conditional Model Builder: rpart

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

Min Bucket: 7 Complexity: 0.0100 Loss Matrix: Rules Draw

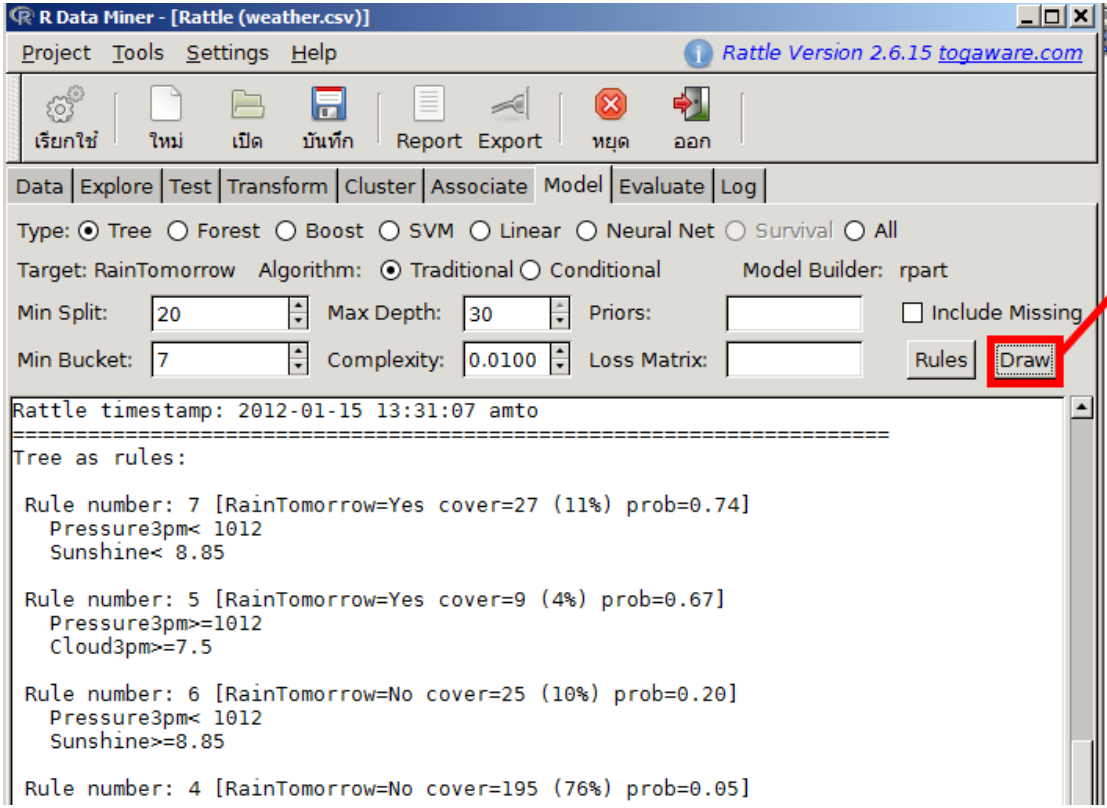
```
Rattle timestamp: 2012-01-15 13:31:07 amto
==========================================================================
Tree as rules:

 Rule number: 7 [RainTomorrow=Yes cover=27 (11%) prob=0.74]
   Pressure3pm< 1012
   Sunshine< 8.85

 Rule number: 5 [RainTomorrow=Yes cover=9 (4%) prob=0.67]
   Pressure3pm>=1012
   Cloud3pm>=7.5

 Rule number: 6 [RainTomorrow=No cover=25 (10%) prob=0.20]
   Pressure3pm< 1012
   Sunshine>=8.85

 Rule number: 4 [RainTomorrow=No cover=195 (76%) prob=0.05]
   Pressure3pm>=1012
   Cloud3pm< 7.5

[1] 7 4 5 6 1 2 3

Rattle timestamp: 2012-01-15 13:41:11 amto
==========================================================================
```

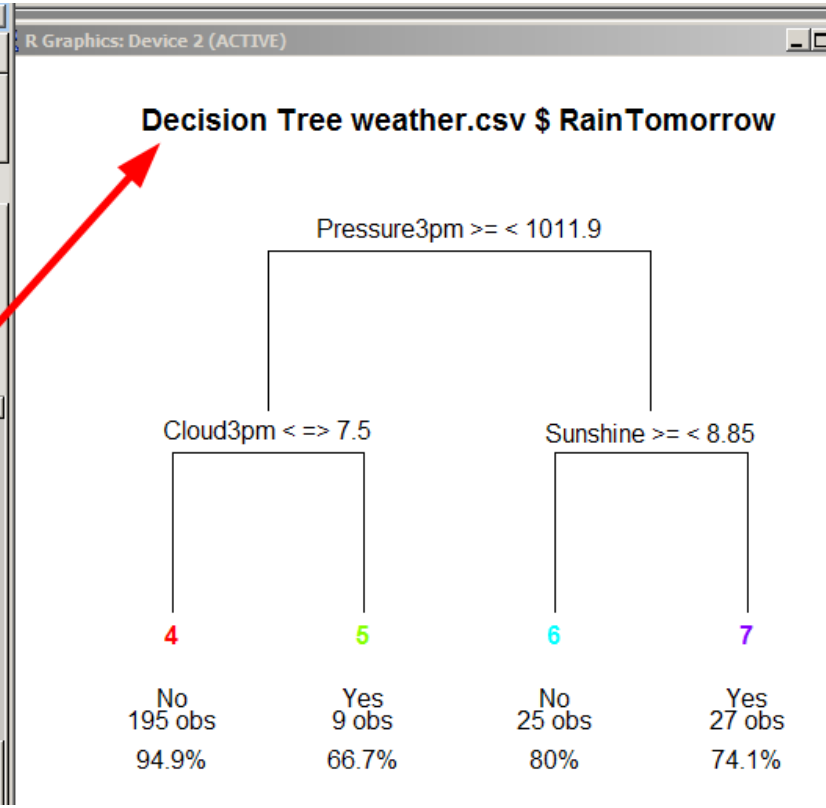
The corresponding rules have been listed. You may need to scroll the textview to view them.

เมื่อ Run model แล้วคลิ๊ก Rules โปรแกรมจะแสดงรายละเอียดของแต่ละกฎ (rules) หรือแต่ละกิ่งนั่นเอง เช่น Rule nuber 7 คือ ทำนายว่าฝนจะตก = yes มี 27 obs คิดเป็น 11% และมีค่า prob=0.74

Rattle Training 9. Classifier Models

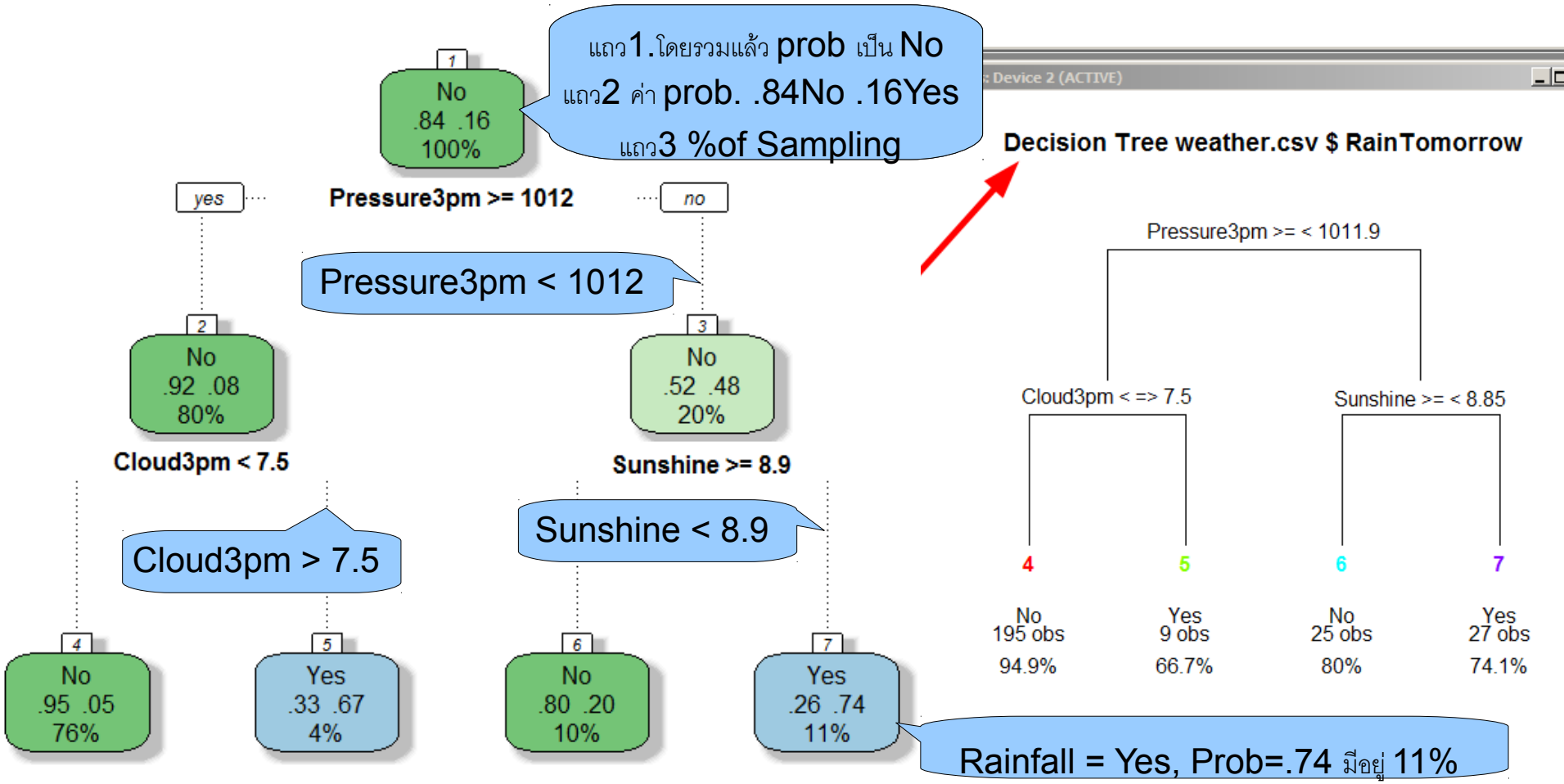
- คลิ๊กที่ **Draw** โปรแกรมจะแสดง **Decision Tree** ขึ้นมา



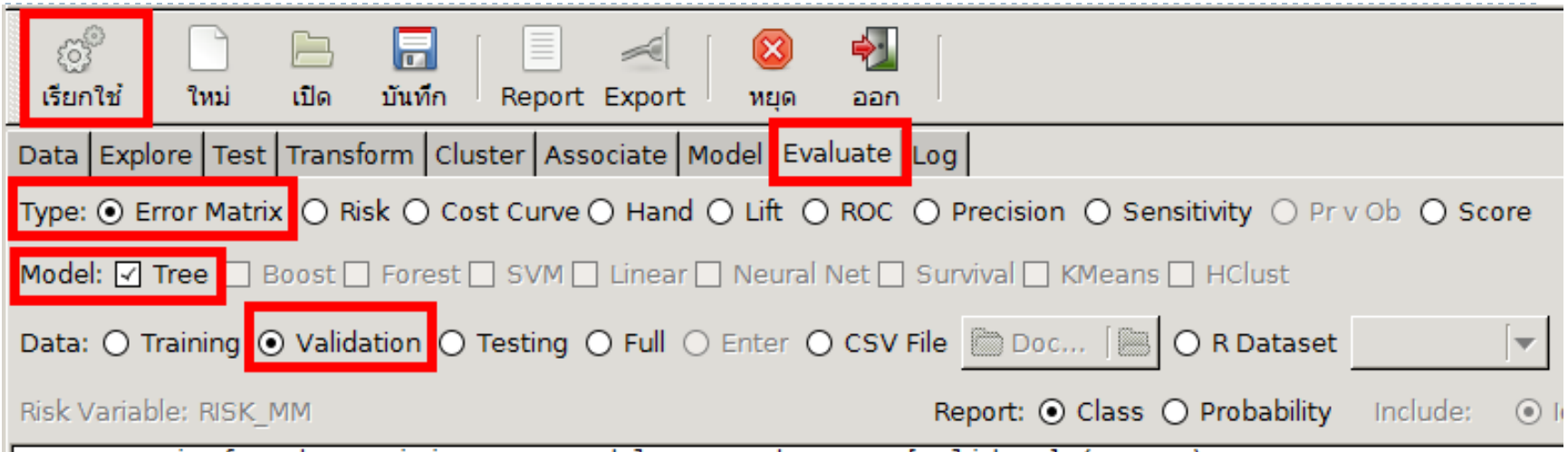


Rattle Training 9. Classifier Models

- เปรียบเทียบเวอร์ชั่นใหม่-เก่า (กราฟที่แสดงมาจาก Testing)



Rattle Training 9. Classifier Models




```
Error matrix for the Decision Tree model on weather.csv [validate] (counts):

      Predicted
Actual No Yes
   No  39   5
   Yes  5   5

Error matrix for the Decision Tree model on weather.csv [validate] (%):

      Predicted
Actual No Yes
   No  72   9
   Yes  9   9

Overall error: 0.1851852

Rattle timestamp: 2012-01-15 13:49:19 amto
=====================================================================
```

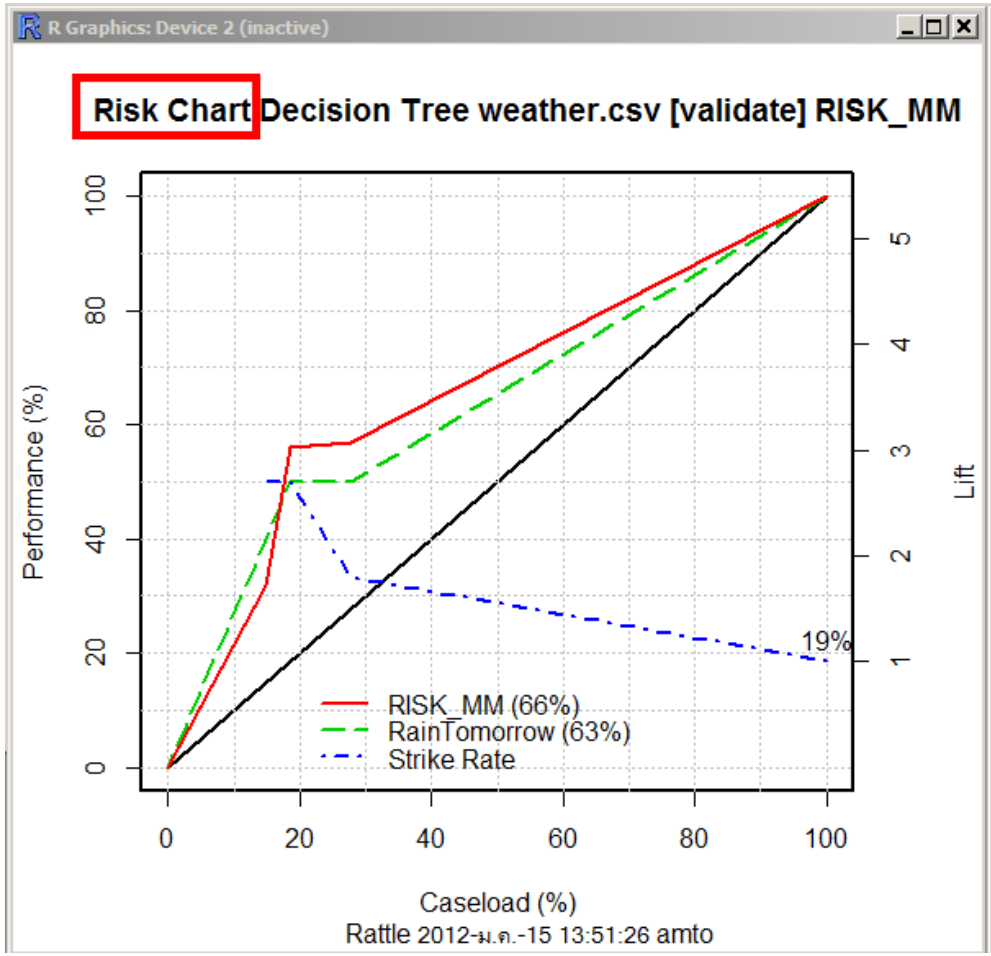
การประเมินประสิทธิภาพของ Model ที่ Evaluate tab.

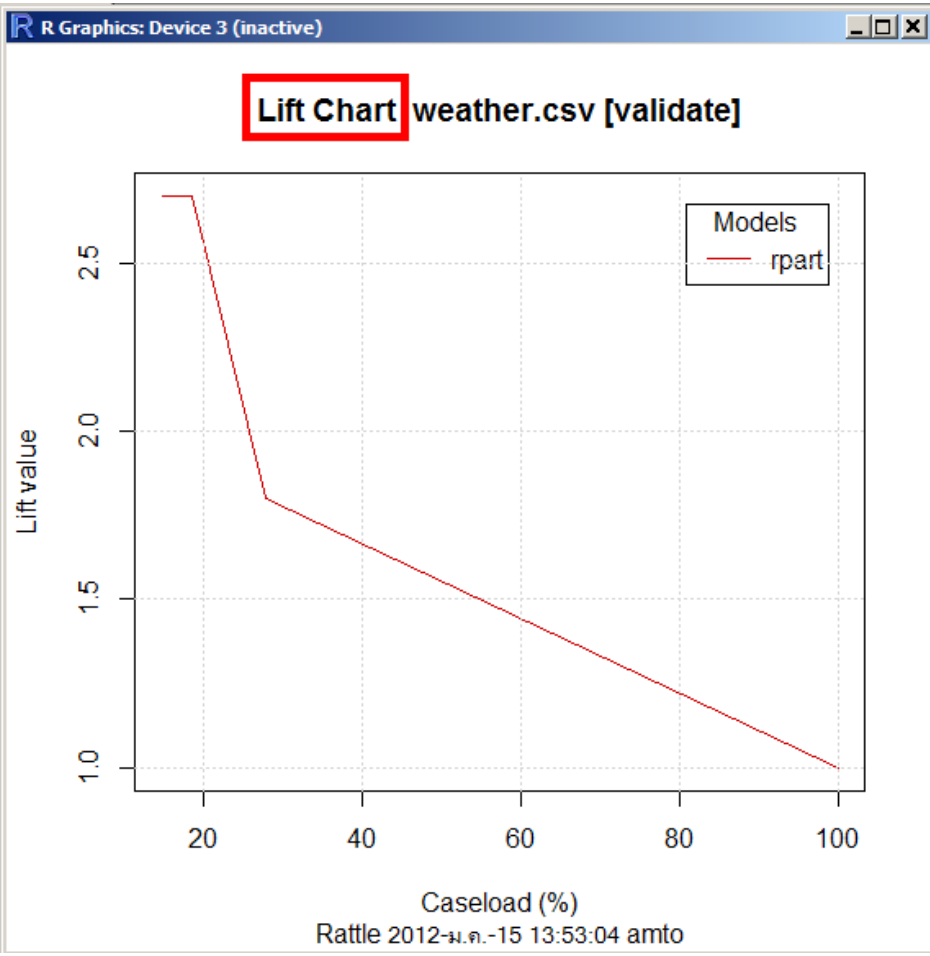
กรณี Error Matrix เลือกประเภท model เป็น Tree

เลือก Data (ในที่นี้คือ Validation) จะได้ 18.5%

Rattle Training 9. Classifier Models

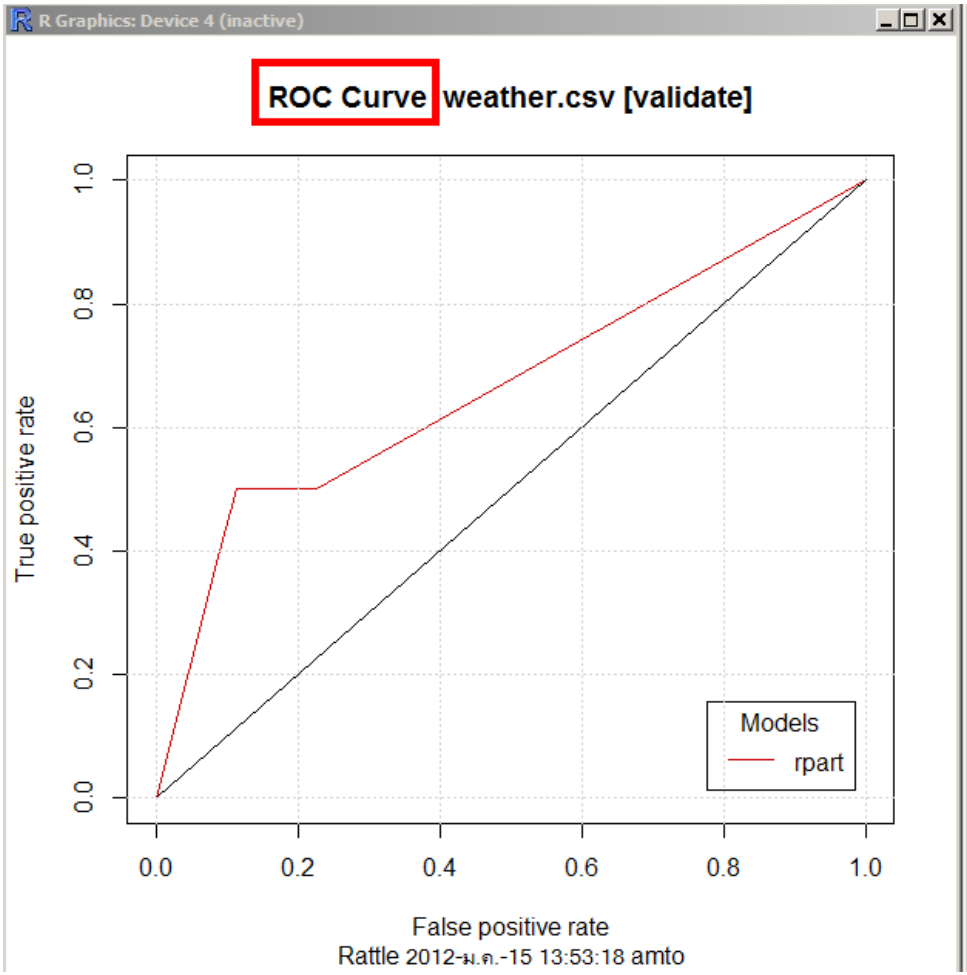
- เลือก Risk & Lift Chart (Rattle จะนำข้อมูลที่เราเลือกมาสร้างกราฟ)

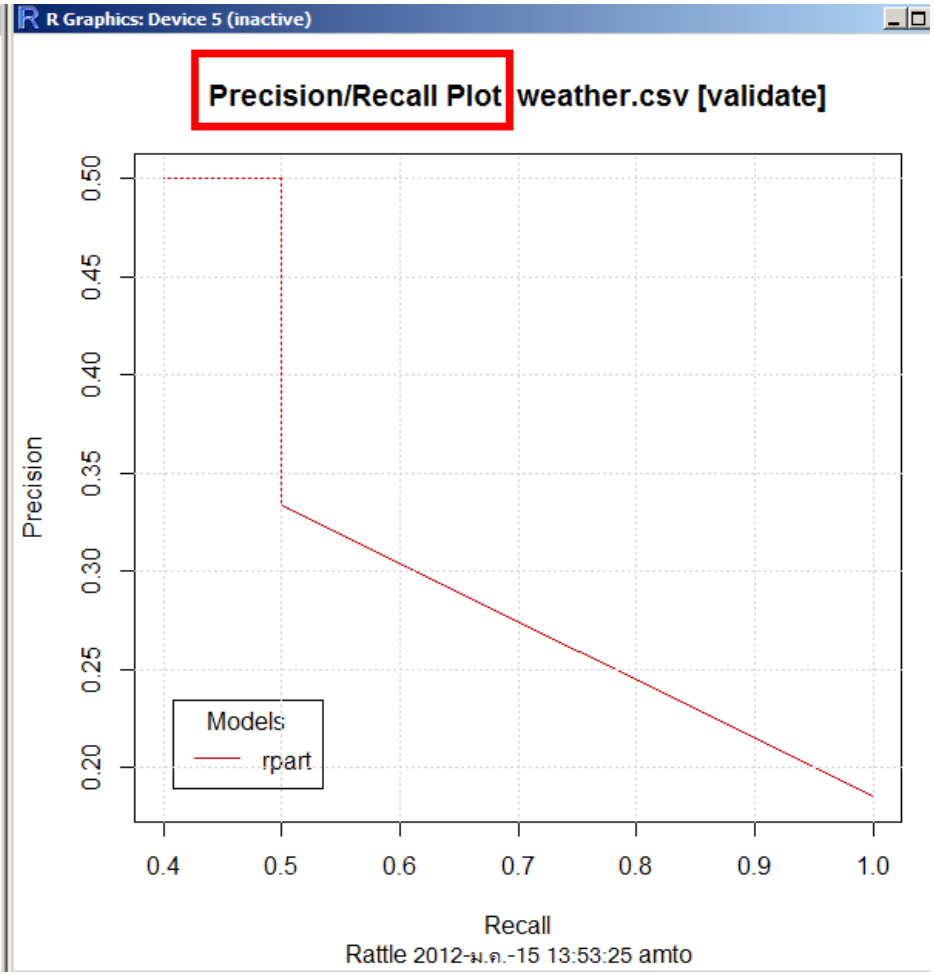




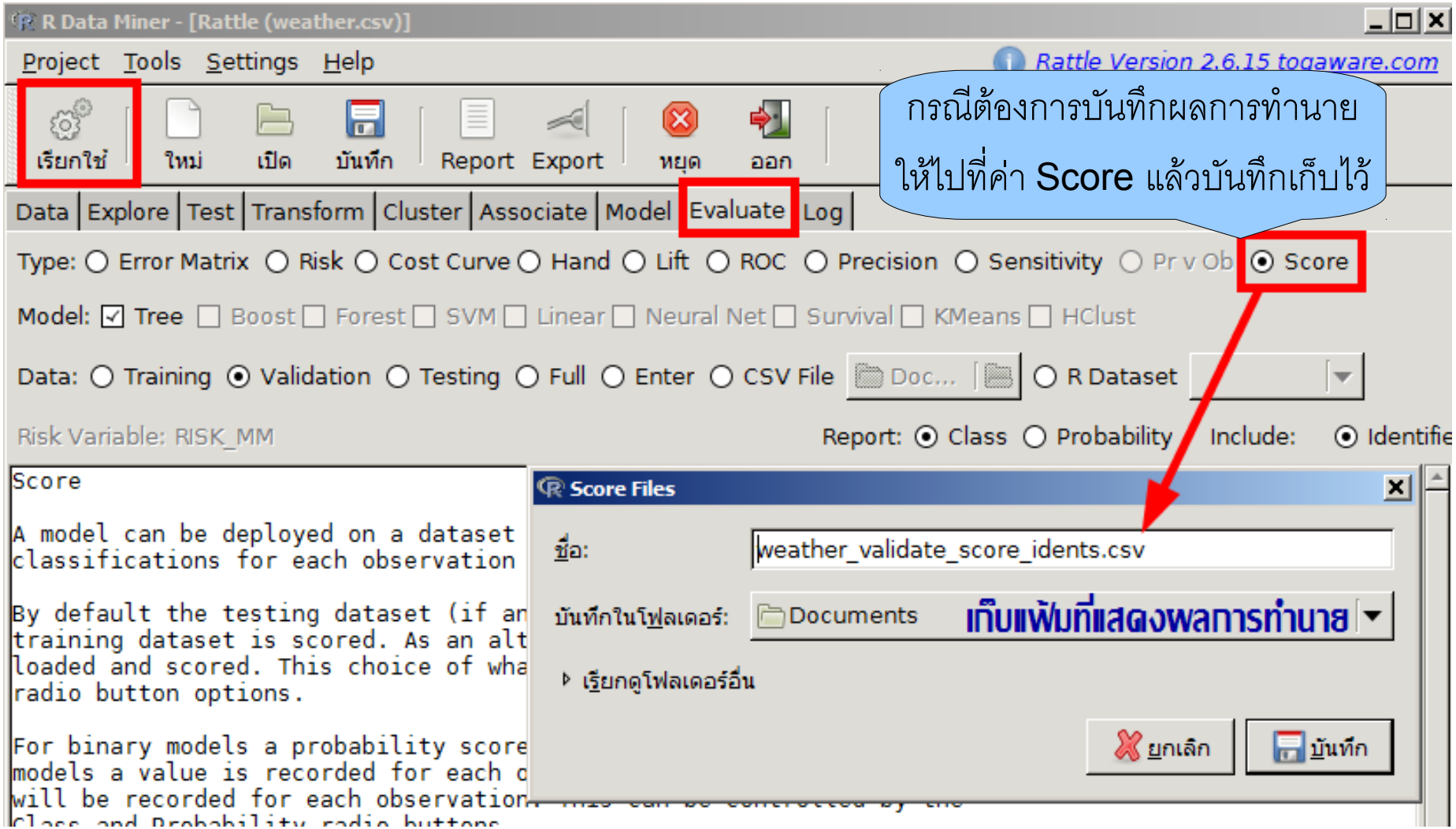
Rattle Training 9. Classifier Models

- เลือก ROC & Precision/ Recall Plot.





Rattle Training 9. Classifier Models

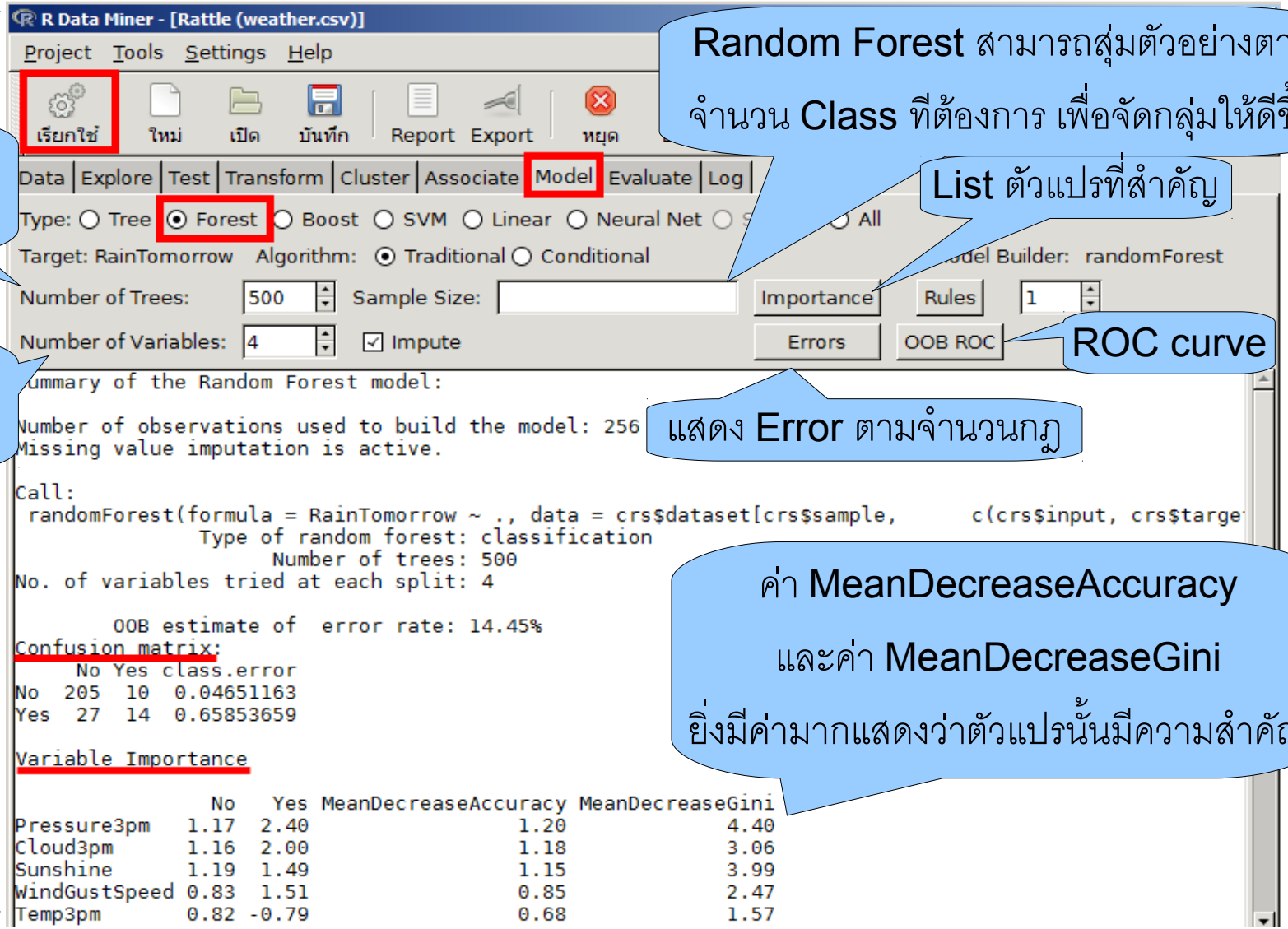


Rattle Training 9. Classifier Models

Rattle ออกแบบมาให้เหมาะกับ Binary classification.

- สรุปเกี่ยวกับการปรับแต่ง Decision Tree
  - **Minimum Split** คือจำนวนต่ำสุดของ obs ที่จะเกิดขึ้นเป็น node (ถ้าข้อมูลที่ทดสอบมีแค่ 50 cases แต่กำหนดไป 51 โปรแกรมจะไม่สร้าง model
  - **Minimum Bucket Size** คือ จำนวนต่ำสุดของ obs ในแต่ละใบ (leaf node) โดยปกติควรตั้งค่า min bucket size ไว้ที่ 1/3 ของ Minimum Split
  - **Cp** ย่อมาจาก Complexity Parameter ซึ่งค่าดังกล่าวเราสามารถนำมาควบคุมขนาดของ decision tree ว่าจะมีขนาดเป็นเท่าไร จึงจะเป็นจุดที่เหมาะสม (optimal size) ค่าปริยายคือ 0.01 โดยค่า 0 หมายถึงกรณี unpruned นั่นคือ complete decision tree ซึ่งจะมีค่า maximum depth โดยทั่วไปกรณีค่า cp ต่ำๆ จะเรียนรู้ train ได้ดีเกินจริงหรือที่เราเรียกว่า Over-fitting ดังนั้นเราอาจปรับแต่งค่า cp สำหรับ tree size ที่เหมาะสม โดยพิจารณาดู xerror ควบคู่กัน เพื่อจะสามารถนำไปใช้ predict unseen data set ได้ดีขึ้น
  - **Priors** ใช้กำหนดสัดส่วนของ class ให้สะท้อนถึงสัดส่วนจริงๆ ใน population ซึ่งผลรวมต้องเท่ากับ 1 เช่น กรณีมี 3 class อาจกำหนดสัดส่วนเป็น 0.3,0.5,0.2 เป็นต้น
  - **Loss Matrix** ใช้เป็นตัว Weight ค่าผลลัพธ์การทำนาย class เช่น กรณีมีแค่ 2 class จะต้องใส่ 4 ค่า คือ 0,10,1,0 โดย 4 ตัวนี้แทนค่า (TN,FN,FP,TP)

Rattle Training 9. Classifier Models



กำหนด จำนวน Tree

Random Forest สามารถสุ่มตัวอย่างตาม จำนวน Class ที่ต้องการ เพื่อจัดกลุ่มให้ดีขึ้น

List ตัวแปรที่สำคัญ

ROC curve

กำหนดจำนวน Variables

แสดง Error ตามจำนวนกฎ

ค่า MeanDecreaseAccuracy และค่า MeanDecreaseGini ยิ่งมีค่ามากแสดงว่าตัวแปรนั้นมีความสำคัญ

Rattle Training 9. Classifier Models

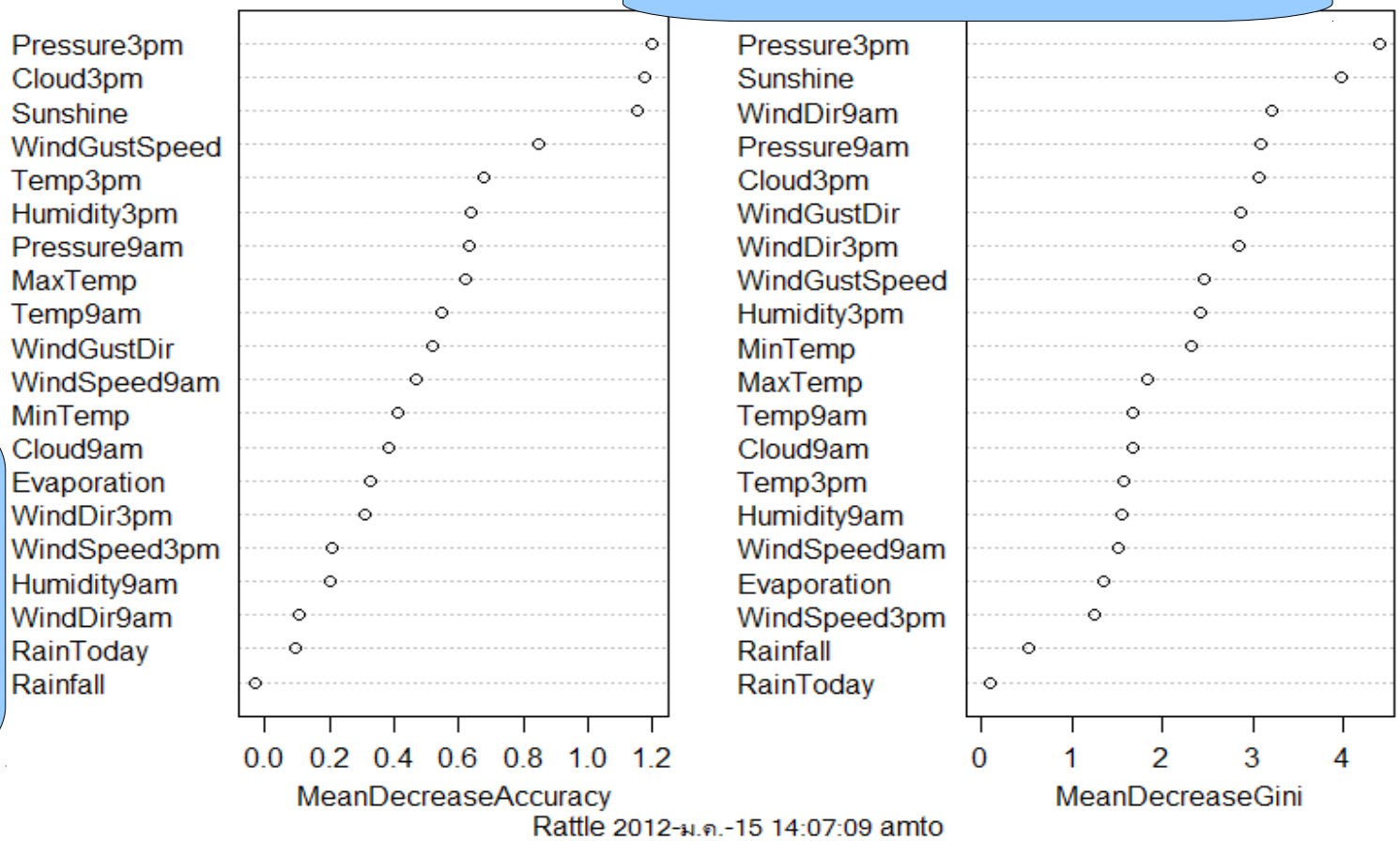
R Graphics: Device 2 (ACTIVE)

Variable Importance Random Forest weather.csv

เมื่อคลิ๊กปุ่ม Importance

Graph จากการคลิ๊กปุ่ม Importance

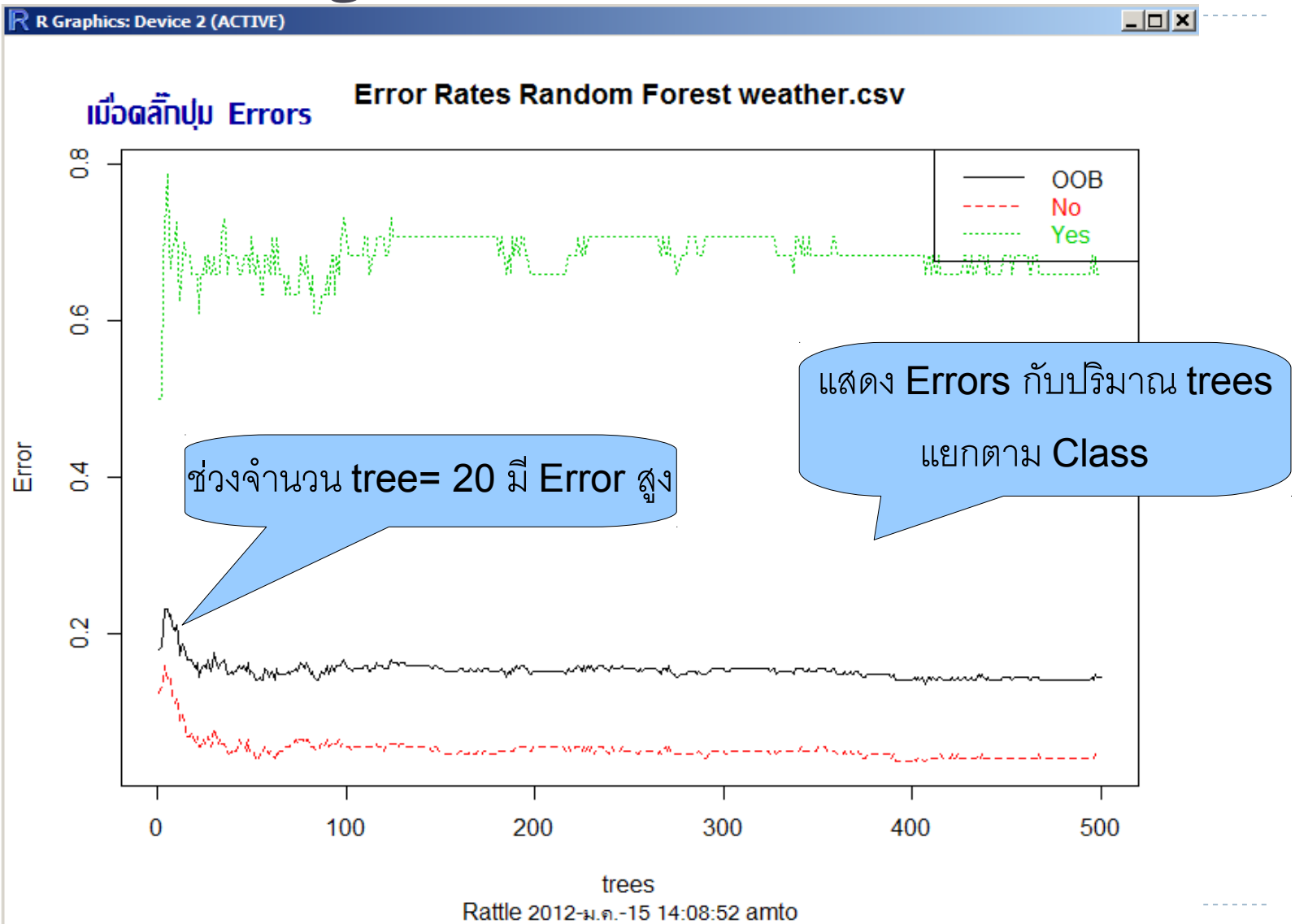
2 Graph นี้ อาจแตกต่างกันบ้าง แต่พอเห็นตัวหลักๆ



Rattle 2012-ม.ค.-15 14:07:09 amto

ความน่าสนใจอยู่ที่ การนำเอาตัวแปรเชิงปริมาณและกลุ่มมาสรุปอยู่ด้วยกันได้

Rattle Training 9. Classifier Models



Rattle Training 9. Classifier Models

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

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

Target: RainTomorrow Algorithm: ◉ Traditional ○ Conditional Model Builder: randomForest

Number of Trees: 500 Sample Size: Importance Rules 1

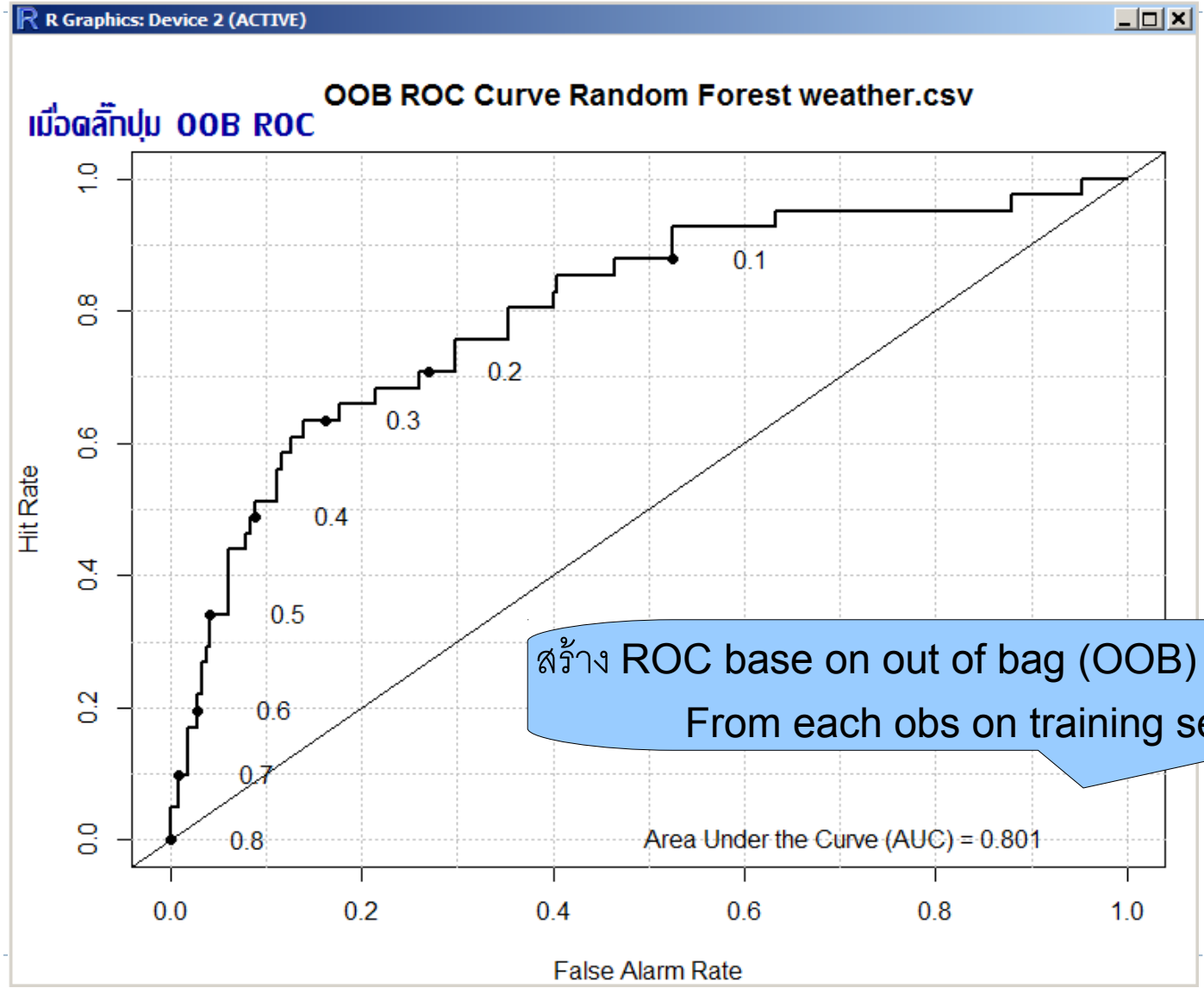
Number of Variables: 4 ☑ Impute Errors OOB ROC

```
-----------------------------------------------------------------
Tree 1 Rule 21 Node 39 Decision No

1: Sunshine > 6.45
2: Pressure3pm > 1011.9
3: MaxTemp <= 34.8
4: WindDir3pm IN ("E", "ESE", "N", "NE", "NNE", "NNW", "NW", "S", "SE", "SSE", "SSW", "W", "WNW", "W
5: MinTemp > -0.15
-----------------------------------------------------------------
Tree 1 Rule 22 Node 26 Decision No

1: Sunshine > 6.45
2: Pressure3pm > 1011.9
3: MaxTemp > 34.8
4: Cloud3pm <= 3.5
-----------------------------------------------------------------
Tree 1 Rule 23 Node 27 Decision Yes

1: Sunshine > 6.45
2: Pressure3pm > 1011.9
3: MaxTemp > 34.8
4: Cloud3pm > 3.5
-----------------------------------------------------------------
Number of rules in Tree 1: 23
```

แสดงรายละเอียดของ trees ในแต่ละกฎที่ Model สร้างขึ้น

Rules from tree 1 have been added to the textview. You may need to scroll the textview to view them.

Rattle Training 9. Classifier Models



สร้าง ROC base on out of bag (OOB) Prediction
From each obs on training set

Rattle Training 9. Classifier Models

- สรุปเกี่ยวกับการปรับแต่ง Random Forest
  - **Number of Trees** คือจำนวน trees มี default = 500 แนะนำให้กำหนดขั้นต่ำที่ 100 เพราะยิ่งสร้างจำนวน trees ให้มากขึ้น model มีแนวโน้ม error ลดลง
  - **Number of Variables** คือจำนวนตัวแปรที่คัดเลือกเข้าไปสร้าง decision tree โดยทั่วไปค่าของจำนวนตัวแปรที่เหมาะสมคือ รากที่สองของจำนวนตัวแปรที่มีทั้งหมด
  - **Sample Size** การกำหนดกลุ่มตัวอย่างใน Random forest มีวัตถุประสงค์ ดังตัวอย่างต่อไปนี้ สมมุติในไฟล์ weather ซึ่งเป็น training set มี 5,000 obs. แต่มีข้อมูล rain tomorrow แค่ 500 obs. กรณีเช่นนี้ เราสามารถกำหนดให้มี rain & no rain เป็น 400,400 ก็จะทำให้ algorithm มีประสิทธิภาพขึ้น (มี probabilities)
  - **Variable Importance** เราสามารถกำหนดจำนวนตัวแปรที่สำคัญๆ เพื่อทำนายผลลัพธ์ ซึ่งพิจารณาได้จากค่า MeanDecreaseAccuracy, MeanDecreaseGini

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Target: RainTomorrow — Model Builder: ada

Number of Trees: 50 ☐ Stumps Defaults Importance Errors List Draw 1

Max Depth: 30 Min Split: 20 Complexity: 0.0100 X Val: 10 Continue

```
Summary of the Ada Boost model:

Call:
ada(RainTomorrow ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], control = rpart.control(maxdepth = 30, cp = 0.01,
    minsplit = 20, xval = 10), iter = 50)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value  No Yes
       No  213   2
       Yes  15  26

Train Error: 0.066

Out-Of-Bag Error:  0.094  iteration= 41

Additional Estimates of number of iterations:

train.err1 train.kap1
        47         47
```

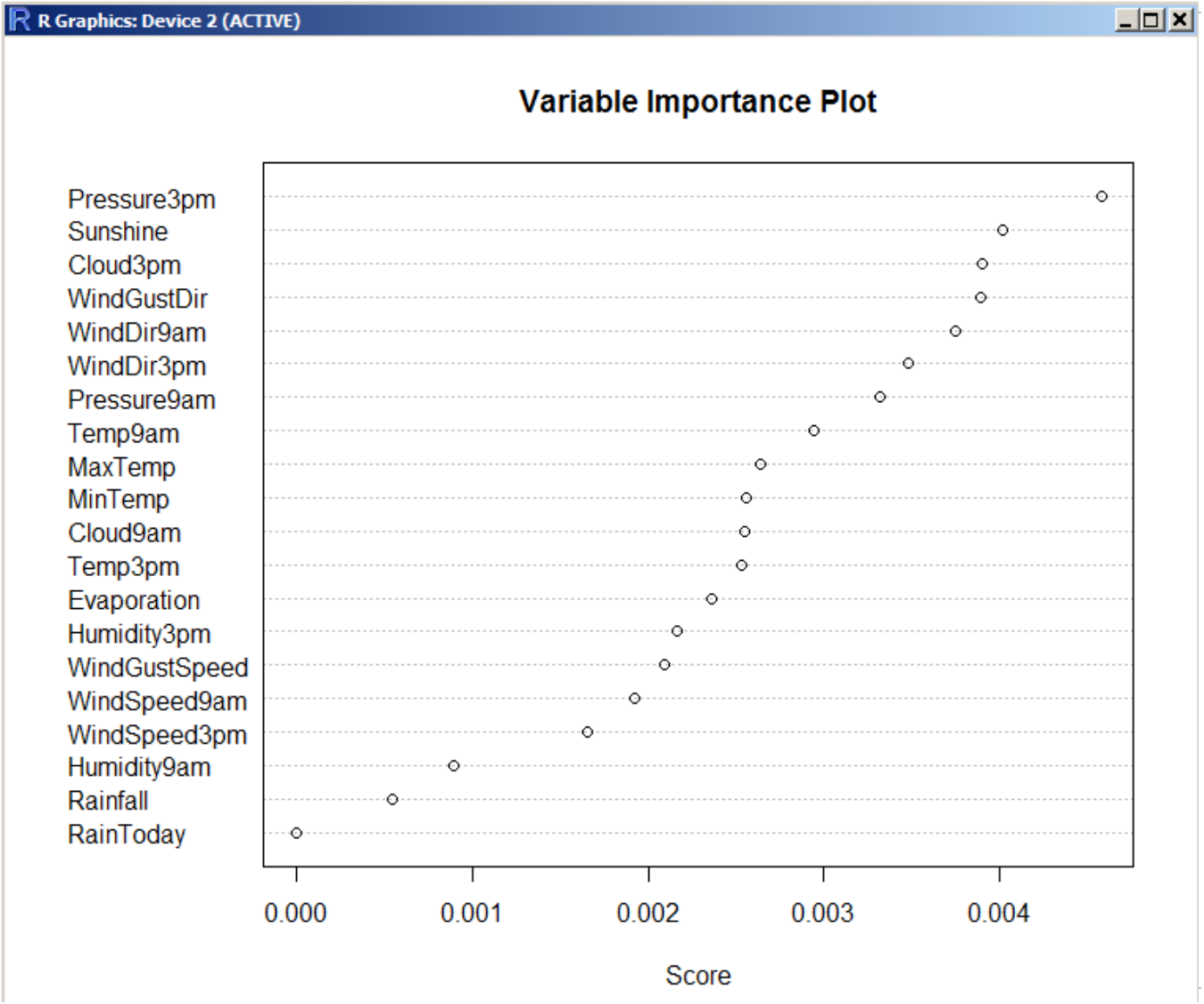
Boosting เป็น meta-algorithm คือผนวกหลายๆ model เข้ามาประยุกต์เพื่อให้เป็น good learner

Rattle Training 9. Classifier Models

- **Boost** ใน **Rattle** ออกแบบมาให้ง่ายในการใช้งาน เราปรับแค่จำนวน **trees**

```
Variables actually used in tree construction:
 [1] "Cloud3pm"      "Cloud9am"      "Evaporation"   "Humidity3pm"
 [5] "Humidity9am"   "MaxTemp"       "MinTemp"       "Pressure3pm"
 [9] "Pressure9am"   "Rainfall"      "Sunshine"      "Temp3pm"
[13] "Temp9am"       "WindDir3pm"    "WindDir9am"    "WindGustDir"
[17] "WindGustSpeed" "WindSpeed3pm"  "WindSpeed9am"

Frequency of variables actually used:

   WindDir9am    WindDir3pm      Sunshine   Pressure3pm   WindGustDir
           34            27            26            25            25
     Cloud3pm       MaxTemp       Temp9am WindGustSpeed      Cloud9am
           13             7             7             5             4
  Humidity3pm       MinTemp       Temp3pm  WindSpeed9am   Evaporation
            4             3             3             3             2
  Humidity9am  WindSpeed3pm   Pressure9am      Rainfall
            2             2             1             1

Time taken: 3.37 secs

Rattle timestamp: 2012-01-15 15:51:43 amto
==========================================================================
```

Rattle Training 9. Classifier Models



Rattle Training 9. Classifier Models

จากรูปนี้แสดงให้ว่า error นั้นค่อยๆ ลดลง มาคงที่อยู่ในช่วง 30-40 และ ลดลงมากในช่วง 50 trees

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

No Target — Model Builder: ada

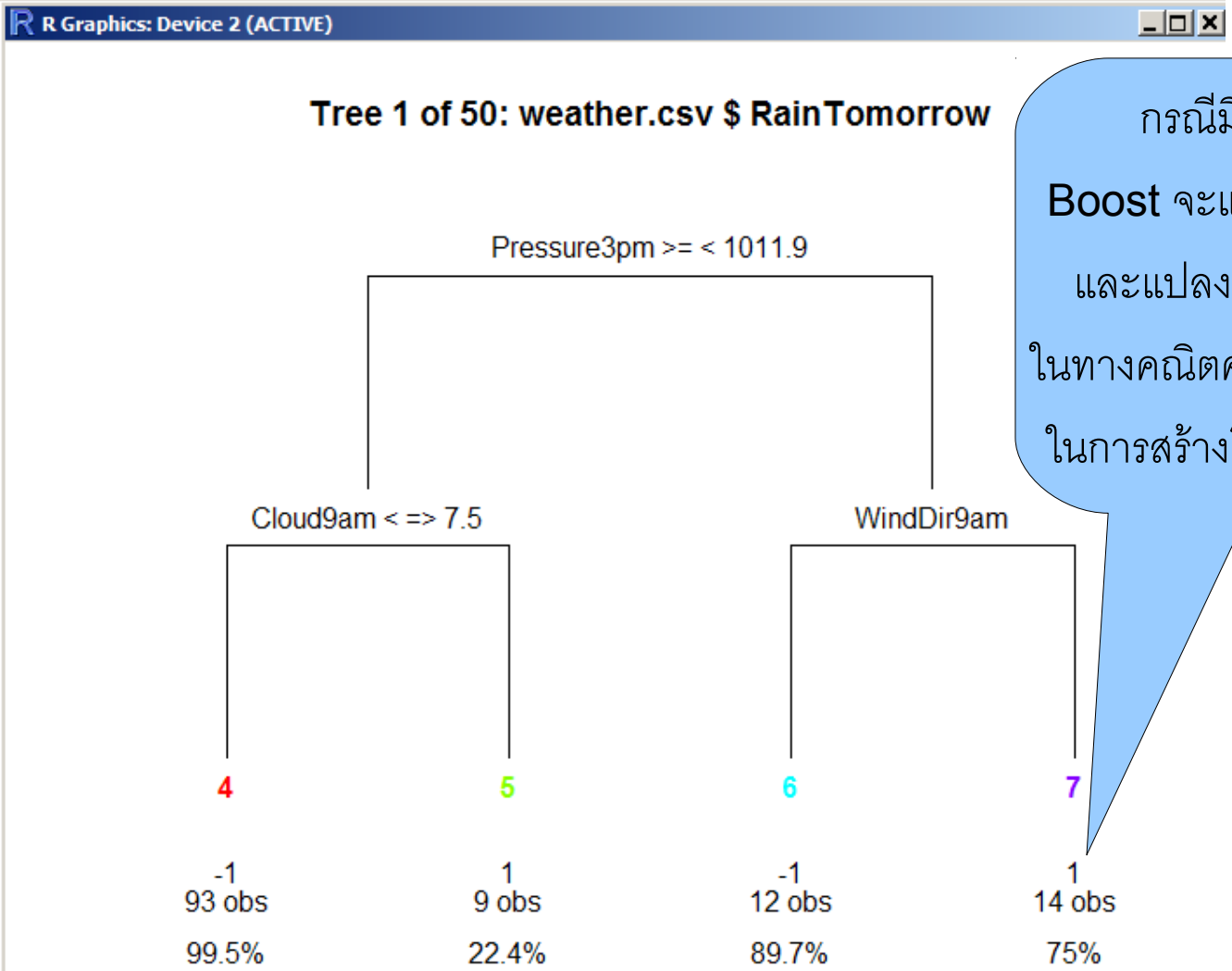
Number of Trees: 50 ☐ Stumps Defaults Importance Errors List Draw 1

Max Depth: 30 Min Split: 20 Complexity: 0.0100 X Val: 10 Continue

```
            2            2            1            1

Time taken: 3.37 secs

Rattle timestamp: 2012-01-15 15:51:43 amto
=====================================================================

Tree 1 of 50:
n= 128

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 128 0.09375000 -1 (0.949438202 0.050561798)
  2) Pressure3pm>=1011.9 102 0.02734375 -1 (0.983280255 0.016719745)
    4) Cloud9am< 7.5 93 0.00781250 -1 (0.994953743 0.005046257) *
    5) Cloud9am>=7.5 9 0.01562500 1 (0.776119403 0.223880597) *
  3) Pressure3pm< 1011.9 26 0.03515625 1 (0.696428571 0.303571429)
    6) WindDir9am=E,NNE,S,SSE,WNW 12 0.01562500 -1 (0.896551724 0.103448276) *
    7) WindDir9am=ENE,N,NNW,NW 14 0.00390625 1 (0.250000000 0.750000000) *
NULL

Rattle timestamp: 2012-01-15 15:55:40 amto
=====================================================================
```

List : สามารถแสดง tree & root ได้เช่นเดียวกับ decision tree

Rattle Training 9. Classifier Models



กรณีมีแค่ 2 class

Boost จะแปลงค่า yes เป็น 1

และแปลงค่า no เป็น -1 ซึ่ง

ในทางคณิตศาสตร์ให้ความหมาย

ในการสร้างโมเดลได้ดีกว่า 0 /1

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Target: RainTomorrow Model Builder: ksvm

Kernel: Radial Basis (rbfdot) Options:

```
Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0394657153475283

Number of Support Vectors : 106

Objective Function Value : -59.2538
Training error : 0.100877
Probability model included.

Time taken: 0.25 secs

Rattle timestamp: 2012-01-15 15:59:14 amto
==========================================================================
```

- Radial Basis (rbfdot)
- Polynomial (polydot)
- Linear (vanilladot)
- Hyperbolic Tangent (tanhdot)
- Laplacian (laplacedot)
- Bessel (besseldot)
- ANOVA RBF (anovadot)
- Spline (splinedot)

นักวิเคราะห์ควรทำกราฟ Scatter plot matrix ดูก่อน แล้วเลือก Kernel ให้เหมาะสม

SVM ใช้หลักการแบ่งข้อมูลตามรูปแบบที่เหมาะสม

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

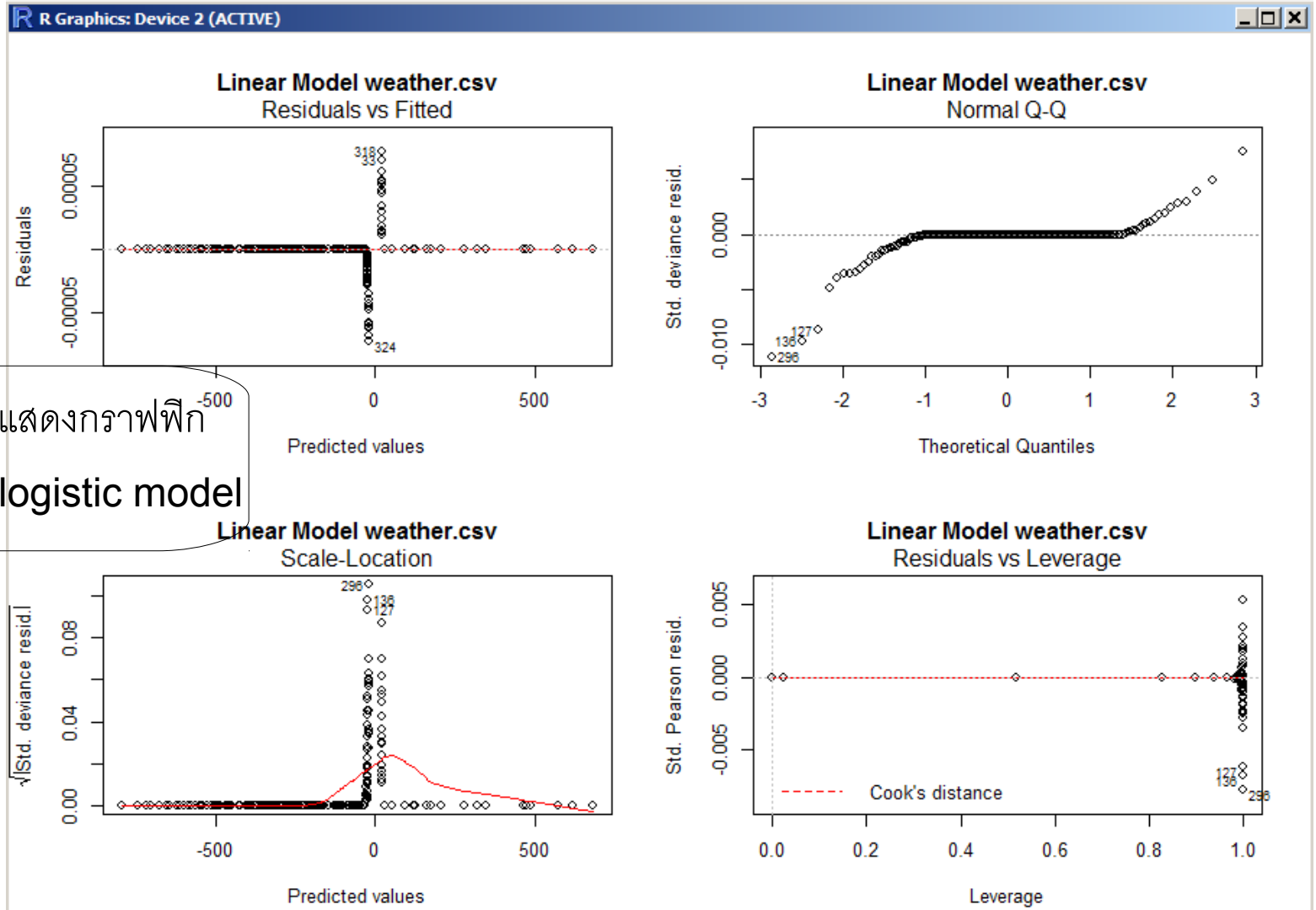
○ Numeric ○ Generalized ○ Poisson ◉ Logistic ○ Probit ○ Multinomial — Model Builder: glm (logit)

Plot

```
Summary of the Logistic Regression model (built using glm):

Call:
glm(formula = RainTomorrow ~ ., family = binomial(link = "logit"),
    data = crs$dataset[crs$train, c(crs$input, crs$target)])

Deviance Residuals:
         Min            1Q        Median            3Q           Max
-0.000072506  -0.000000021  -0.000000021  -0.000000021   0.000077052

Coefficients:
                    Estimate    Std. Error z value Pr(>|z|)
(Intercept)       22744.4759 10127171.8292   0.002    0.998
MinTemp             -24.4362    11907.6746  -0.002    0.998
MaxTemp             -53.1118    64763.4150  -0.001    0.999
Rainfall             14.7010    12647.7823   0.001    0.999
Evaporation           0.9518    12277.7891   0.000    1.000
Sunshine            -25.0170     8639.2648  -0.003    0.998
WindGustDirENE       83.8258   166966.2104   0.001    1.000
WindGustDirESE      -26.4217   254732.3762   0.000    1.000
WindGustDirN         66.0339   186147.7874   0.000    1.000
WindGustDirNE       -43.6612    87675.8241   0.000    1.000
WindGustDirNNE     -318.5608   374359.7885  -0.001    0.999
```

Logistic ใน Rattle ใช้ในกรณีที่ มี Target แค่ 2 ค่า สำหรับ Input สามารถใช้ได้ ทั้ง categoric & numeric

Rattle Training 9. Classifier Models



เมื่อคลิ๊ก plot จะแสดงกราฟฟิก

ในการ diagnosis logistic model

Rattle Training 9. Classifier Models

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.15 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Target: RainTomorrow

Hidden Layer Nodes: 10

Model Builder: nnet (0/1)

```
Summary of the Neural Net model (built using nnet):

A 62-10-1 network with 703 weights.
Inputs: MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDirENE, WindGustDirESE, WindGustDirN,
WindGustDirNE, WindGustDirNNE, WindGustDirNNW, WindGustDirNW, WindGustDirS, WindGustDirSE, WindGustDirSSE,
WindGustDirSSW, WindGustDirSW, WindGustDirW, WindGustDirWNW, WindGustDirWSW, WindGustSpeed, WindDir9amENE,
WindDir9amESE, WindDir9amN, WindDir9amNE, WindDir9amNNE, WindDir9amNNW, WindDir9amNW, WindDir9amS,
WindDir9amSE, WindDir9amSSE, WindDir9amSSW, WindDir9amSW, WindDir9amW, WindDir9amWNW, WindDir9amWSW,
WindDir3pmENE, WindDir3pmESE, WindDir3pmN, WindDir3pmNE, WindDir3pmNNE, WindDir3pmNNW, WindDir3pmNW,
WindDir3pmS, WindDir3pmSE, WindDir3pmSSE, WindDir3pmSSW, WindDir3pmSW, WindDir3pmW, WindDir3pmWNW,
WindDir3pmWSW, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am,
Cloud3pm, Temp9am, Temp3pm, RainTodayYes.
Output: as.factor(RainTomorrow).
Sum of Squares Residuals: 0.0123.

Neural Network build options: skip-layer connections; entropy fitting.

In the following table:
   b  represents the bias associated with a node
   h1 represents hidden layer node 1
   i1 represents input node 1 (i.e., input variable 1)
   o  represents the output node

Weights for node h1:
    b->h1    i1->h1    i2->h1    i3->h1    i4->h1    i5->h1    i6->h1    i7->h1
    -0.66      0.23      0.29     -0.31     -0.68     -0.36      0.27      0.23
   i8->h1    i9->h1   i10->h1   i11->h1   i12->h1   i13->h1   i14->h1   i15->h1
```

The Neural Net model has been built. Time taken: 1.58 secs

จำนวน Hidden Layer ที่เหมาะสมคือ (input+ target)/2 เศษปัดขึ้น

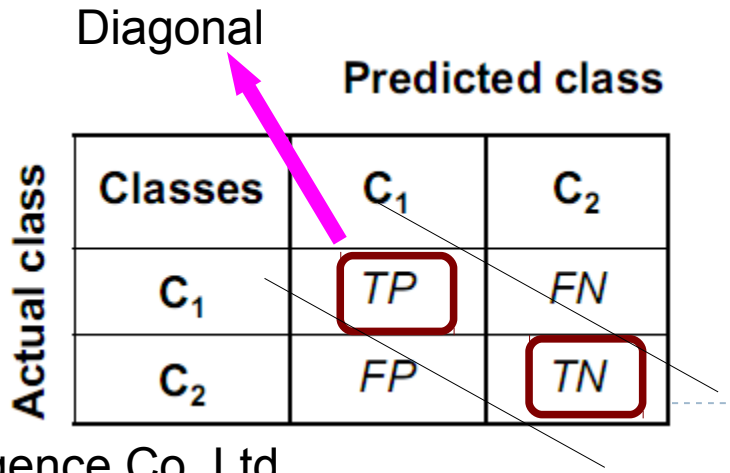
10.Evaluation

- Train & Test
  - Performance on training data → ไม่ได้ประเมินถึงข้อมูลที่ไม่เคยเจอ
  - บางครั้งผลการประเมินดีมาก เพราะ machine learning ได้เห็น training data มาก่อน
  - ถ้าจะประเมินให้มั่นใจได้ ✳ ต้องมี training data & testing data ที่มากพอ
- Cross valid (เหมาะสำหรับข้อมูลขนาดเล็ก)
  - 10-fold cross-validation → unlucky or lucky results
  - จะได้ผลที่ดีขึ้นเมื่อ ✳ repeated cross validation (เช่น ใน experimenter ของ weka)
  - Bootstrapping → sampling done with replacement to form a training dataset.

Confusion matrix

- เป็นตารางที่มีขนาด **m x m** โดยที่ **m** คือจำนวนของ **class** เป้าหมาย
- มีจำนวนที่สอดคล้องกับสี่เงื่อนไข คือ **TP, FP, TN, FN**
  - TP (True positive) คือ จำนวนตัวอย่างที่ model ทำนายว่า yes และ class เป้าหมายเป็น yes
  - TN (True negative) คือ จำนวนตัวอย่างที่ model ทำนายว่า no และ class เป้าหมายเป็น no
  - FP (False positive) คือ จำนวนตัวอย่างที่ model ทำนายว่า yes และ class เป้าหมายเป็น no
  - FN (False negative) คือ จำนวนตัวอย่างที่ model ทำนายว่า no และ class เป้าหมายเป็น yes
- ใน **WEKA** จำนวนในแถวเป็นของ **class** เป้าหมายจริง ส่วนจำนวนในสดมภ์เป็นของ **Models** ที่ทำนาย
-

```
=== Confusion Matrix ===

 a b   <-- classified as
 2 3 | a = no
 4 5 | b = yes
```



Rattle Training 10. Model Performance Evaluation.

- ตัววัดหลักๆ ล้วนแต่พัฒนามาจากค่าในส่วนของ Confusion Matrix

$$TP + FN = P$$

$$TN + FP = N$$

```
=== Confusion Matrix ===

 a b   <-- classified as
 2 3 | a = no
 4 5 | b = yes
```

Diagonal

Actual class \ Predicted class	Classes	C_1	C_2
	C_1	TP	FN
	C_2	FP	TN

Recall รู้จักกันในนามของ sensitivity ของ Model

$$FPrate = \frac{FP}{N} \qquad TPrate = \frac{TP}{P} = Recall \qquad Yrate = \frac{TP + FP}{P + N}$$

TPrate = Recall
แนวนอน:แม่นเพียงใด

$$Precision = \frac{TP}{TP + FP} \qquad Accuracy = \frac{TP + TN}{P + N}$$

Precision แนวดิ่ง
ดูความเที่ยงตรงของ tool

ส่วนการวัด TNrate = TN / N
รู้จักกันในชื่อของ specificity

Rattle Training 10. Model Performance Evaluation.

- ในกรณีที่ข้อมูลมี Class มากกว่า 2 ค่าขึ้นไป
- จะคำนวนค่าต่างๆ แยกตาม Class (ดูตัวอย่างการคำนวนในหน้าถัดไป)

```
=== Detailed Accuracy By Class ===
TP Rate    FP Rate    Precision    Recall   F-Measure    Class
  0.8        0.053       0.8         0.8        0.8       soft
  0.25       0.1         0.333       0.25       0.286     hard
  0.8        0.444       0.75        0.8        0.774     none
=== Confusion Matrix ===
  a  b  c   <-- classified as
  4  0  1 |  a = soft
  0  1  3 |  b = hard
  1  2 12 |  c = none
```

Rattle Training 10. Model Performance Evaluation.

Recall = TP / (TP + FN)

- Soft = 4 / (4 + (0 + 1)) = .8
- Hard = 1 / (1 + (0 + 3)) = .25
- None = 12 / (12 + (1 + 2)) = .8

Precision = TP / (TP + FP)

- Soft = 4 / (4 + (0 + 1)) = .8
- Hard = 1 / (1 + (0 + 2)) = .33
- None = 12 / (12 + (1 + 3)) = .75

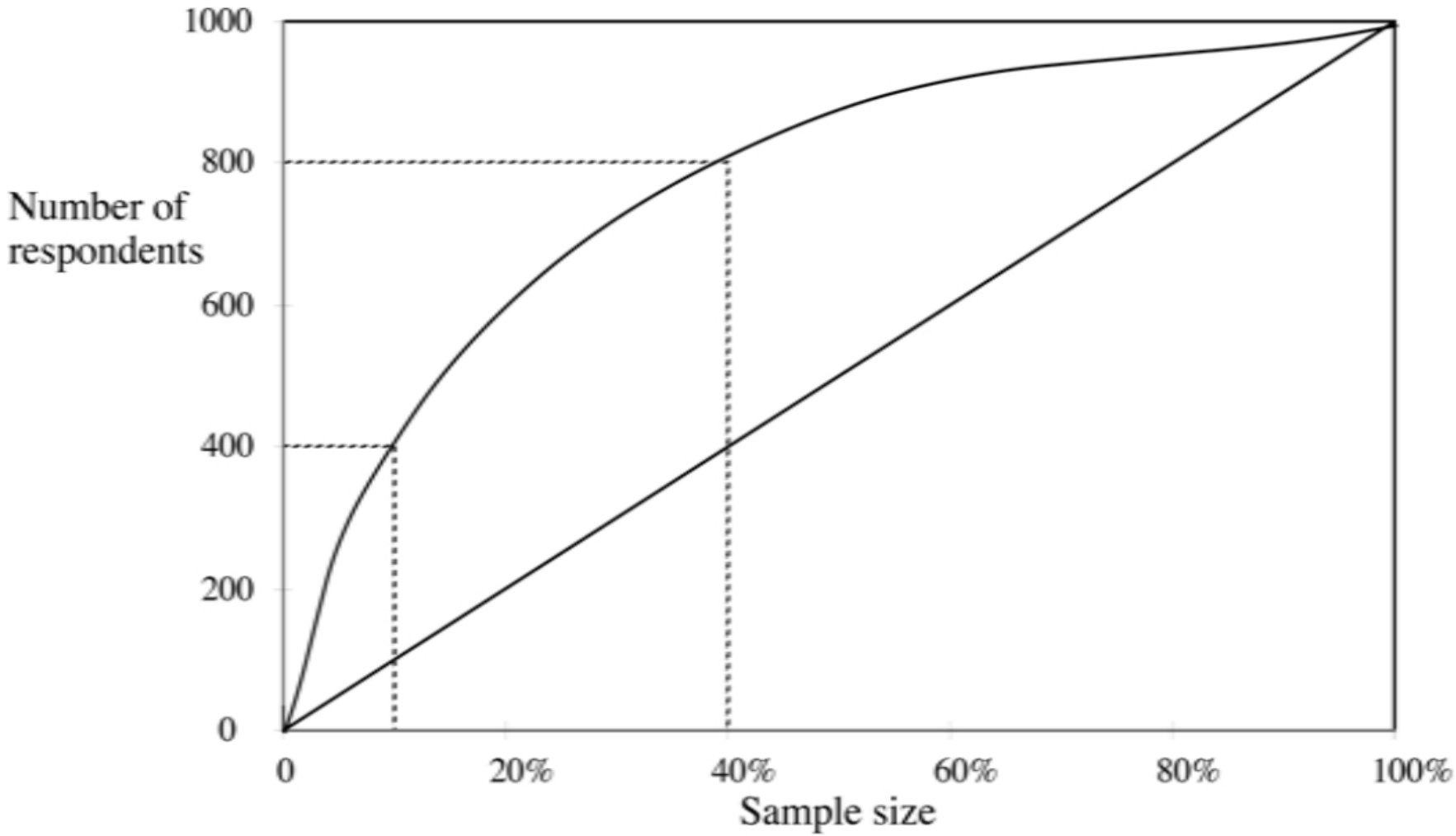
F-measure = 2TP / (2TP + FP + FN)

- Soft = 2 * 4 / (2*4 + (0 + 1) + (0 + 1)) = 8/10 = .8
- Hard = 2 * 1 / (2*1 + (0 + 2) + (0 + 3)) = 2 / 7= .286
- None = 2 * 12 / (2*12 + (1 + 3) + (1 + 2)) = 24 / 31 = .774

Lift Charts

- ในทางปฏิบัติ เรามีค่าใช้จ่ายแฝงอยู่ (ค่าปฏิบัติการ)
- ต้องตัดสินใจโดยเปรียบเทียบจาก possible scenarios
- ตย.เช่น การทำโปรโมชั่นด้วยการส่งแผ่นพับทางไปรษณีย์ไป 1,000,000 ราย
  - กรณีที่ 1 ส่งไปทั้งหมด 1,000,000 ราย จากข้อมูลที่ผ่านมาพบว่ามีผู้ตอบกลับ 0.1% (1,000 ราย)
  - กรณีที่ 2 ส่งไปทั้งหมด 100,000 ราย ตัวแบบ classifier predicts ว่ามีผู้สนใจ 0.4% (400 ราย)
  - กรณีที่ 3 ส่งไปทั้งหมด 400,000 ราย ตัวแบบ classifier predicts ว่ามีผู้สนใจ 0.2% (800 ราย)
  - การเพิ่มขึ้นของ response rate คือ Lift (0.4 / 0.1 = 4 ในกรณีที่ 2 เมื่อเทียบกับการส่งไปทั้งหมด)
- Lift Charts จะเปรียบเทียบให้เห็นสถานการณ์ต่าง ประกอบการตัดสินใจ

Lift Charts

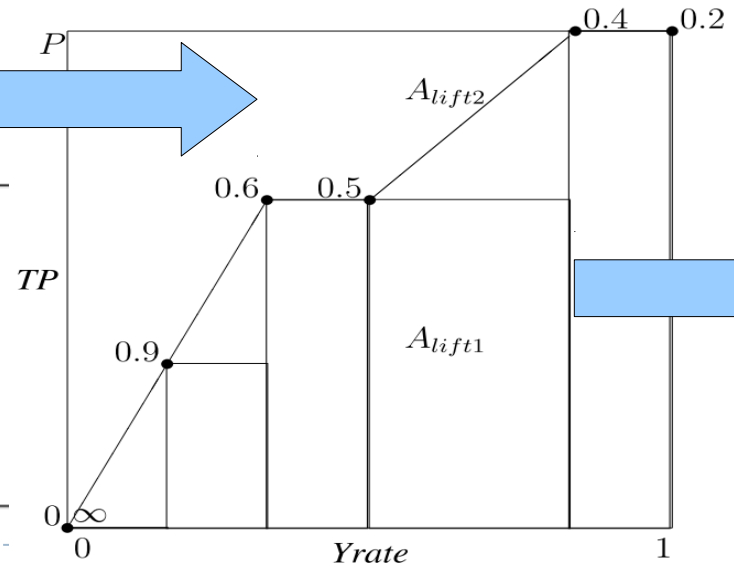

1000
800
600
400
200
0
Number of respondents
0
20%
40%
60%
80%
100%
Sample size

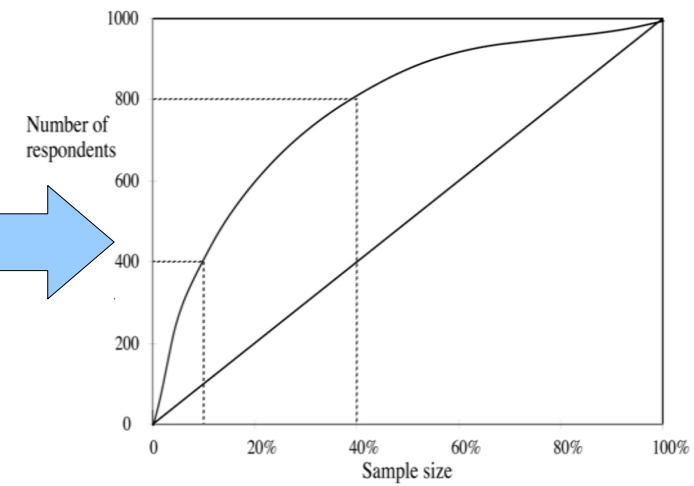
Lift Charts

- ขั้นตอนการสร้าง **Lift charts** จะได้จากค่า **prob** การทำนายของ **classifier**
  - นำค่า **probability** ของ **class** เป้าหมายที่เราสนใจมาเรียงลำดับจากมากไปน้อย (เช่น คนที่จะซื้อสินค้าจากแผ่นพับ = yes)
  - เมื่อทายถูกลากเส้นขึ้นในแนวดิ่ง (Number of respondents)
  - ทายผิดลากเส้นในแนวนอน (Sample size)

given test set.

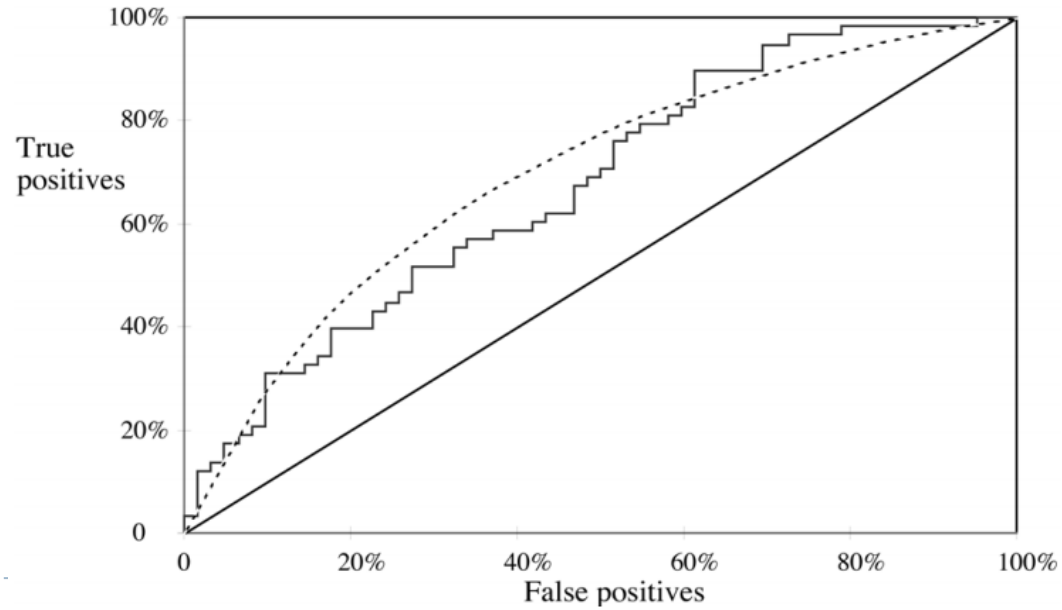
Num	Class	Score
1	p	0.9
2	p	0.6
3	n	0.5
4	n	0.4
5	p	0.4
6	n	0.2





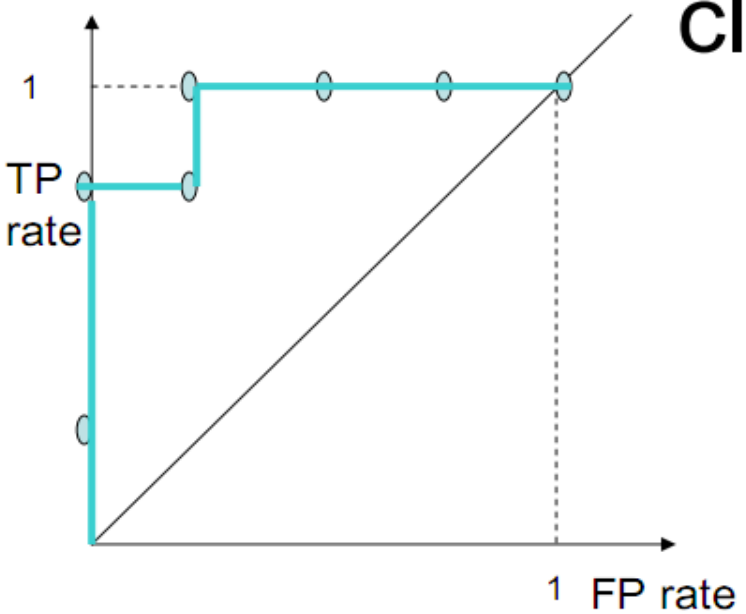
ROC Curves

- คล้ายๆ กับ **Lift Charts**
  - แกน **y** แสดง **% of true positives** จากกลุ่มตัวอย่าง (ใน **lift charts** จะเป็นจำนวนของเป้าหมาย)
  - แกน **x** แสดง **% of false positives** จากกลุ่มตัวอย่าง (ใน **lift charts** จะเป็นจำนวน กลุ่มตัวอย่าง)
- **ROC Curves** ใช้ประเมินเปรียบเทียบ **classifiers** ที่ต่างกันบนชุดข้อมูลเดียวกัน
- **ROC Curves** ไม่ได้คำนึงถึง **class distribution & costs**
  -



Rattle Training 10. Model Performance Evaluation.

ROC curve of a probabilistic classifier



For different threshold values we get different points in the ROC space

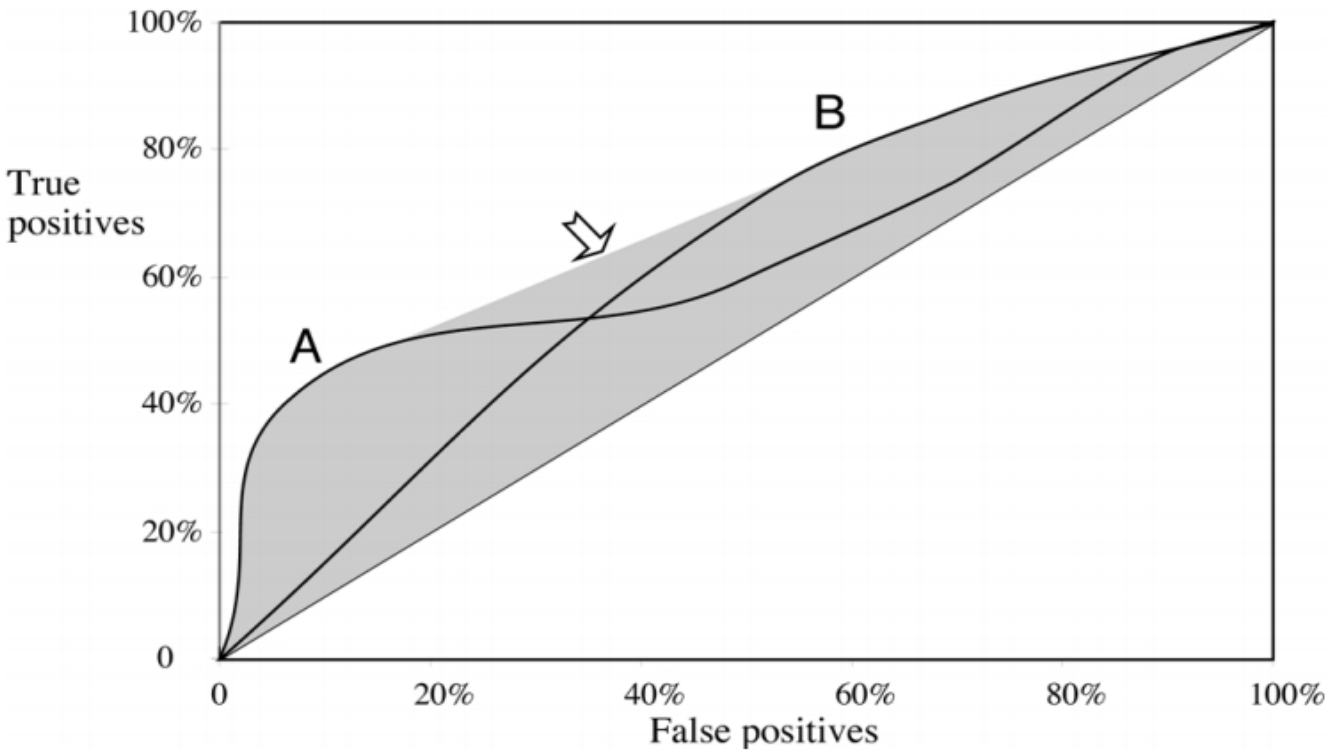
Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	YES	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: FP/N=1/4=0.25

TP rate: TP/P=3/3=1.0, etc…

ROC Curves

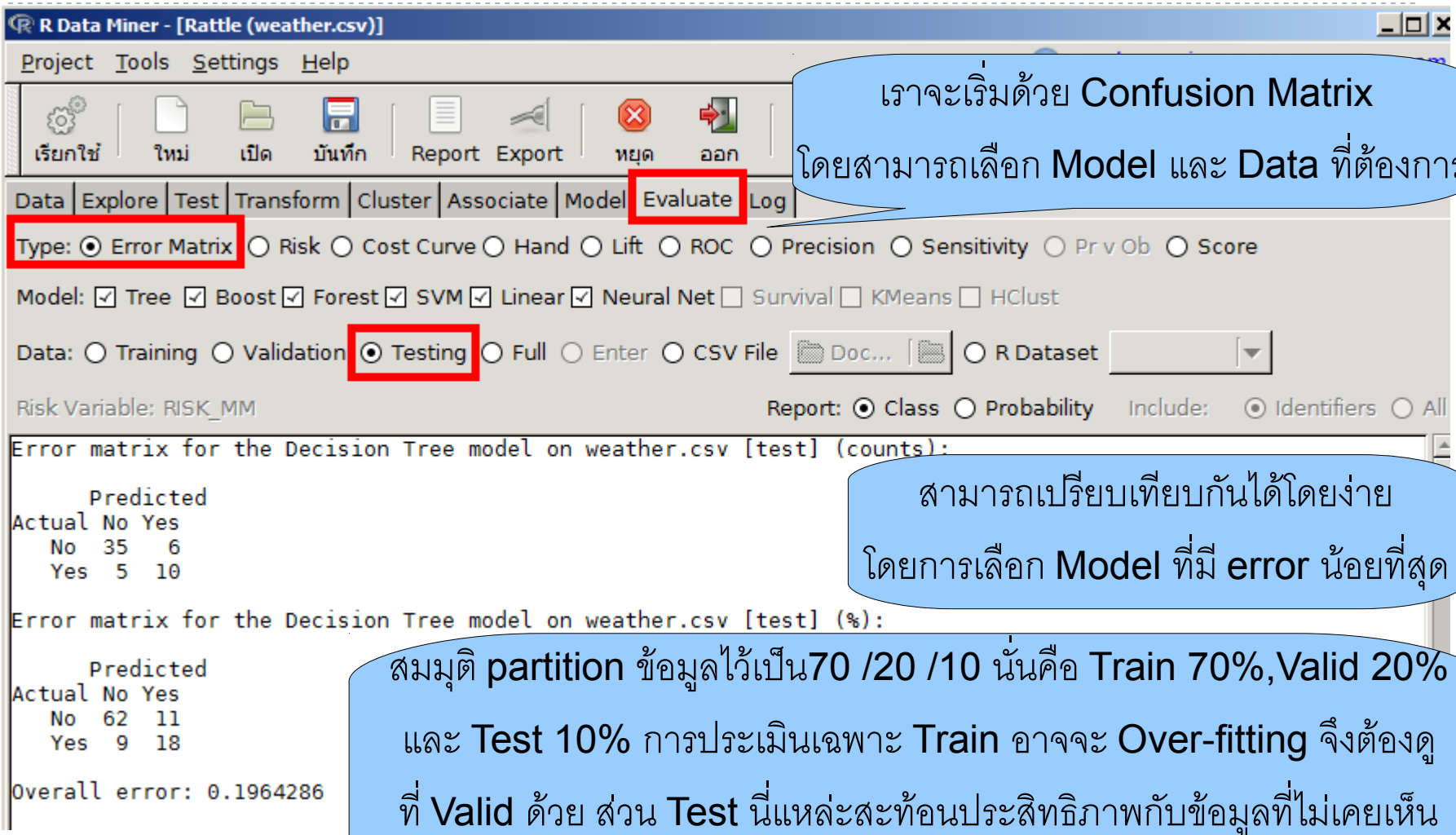
- การเปรียบเทียบ 2 method ด้วย ROC Curves ต้องมองภาพรวม เพราะ
- อย่างเช่น วิธี A อาจจะดูดีเมื่อพิจารณาจากกลุ่มตัวอย่างขนาดเล็ก
- วิธี B ได้ผลที่ดีกว่าเมื่อพิจารณาที่ช่วงกลุ่มตัวอย่างจำนวนที่มากขึ้น (ดูภาพด้านล่าง)



Summery of Evaluating

	Domain	Plot	Explanation
Lift chart	Marketing	TP Subset size	TP (TP+FP)/ (TP+FP+TN+FN)
ROC curve	Communications	TP rate FP rate	TP/(TP+FN) FP/(FP+TN)
Recall-precision curve	Information retrieval	Recall Precision	TP/(TP+FN) TP/(TP+FP)

Rattle Training 10. Model Performance Evaluation.



เราจะเริ่มด้วย Confusion Matrix โดยสามารถเลือก Model และ Data ที่ต้องการ

สามารถเปรียบเทียบกันได้โดยง่าย โดยการเลือก Model ที่มี error น้อยที่สุด

สมมุติ partition ข้อมูลไว้เป็น70 /20 /10 นั่นคือ Train 70%,Valid 20% และ Test 10% การประเมินเฉพาะ Train อาจจะ Over-fitting จึงต้องดูที่ Valid ด้วย ส่วน Test นี่แหล่ะสะท้อนประสิทธิภาพกับข้อมูลที่ไม่เคยเห็น แต่ถ้าเลือกใช้ Full ก็จะใช้ข้อมูลทั้ง 100% เพื่อเรียนรู้

Rattle Training 10. Model Performance Evaluation.

```
Error matrix for the Decision Tree

      Predicted
Actual No Yes
   No  35   6
   Yes  5  10
```

```
Error matrix for the SVM

      Predicted
Actual No Yes
   No  35   1
   Yes 10   4
```

```
Error matrix for the Ada Boost

      Predicted
Actual No Yes
   No  37   4
   Yes  6   9
```

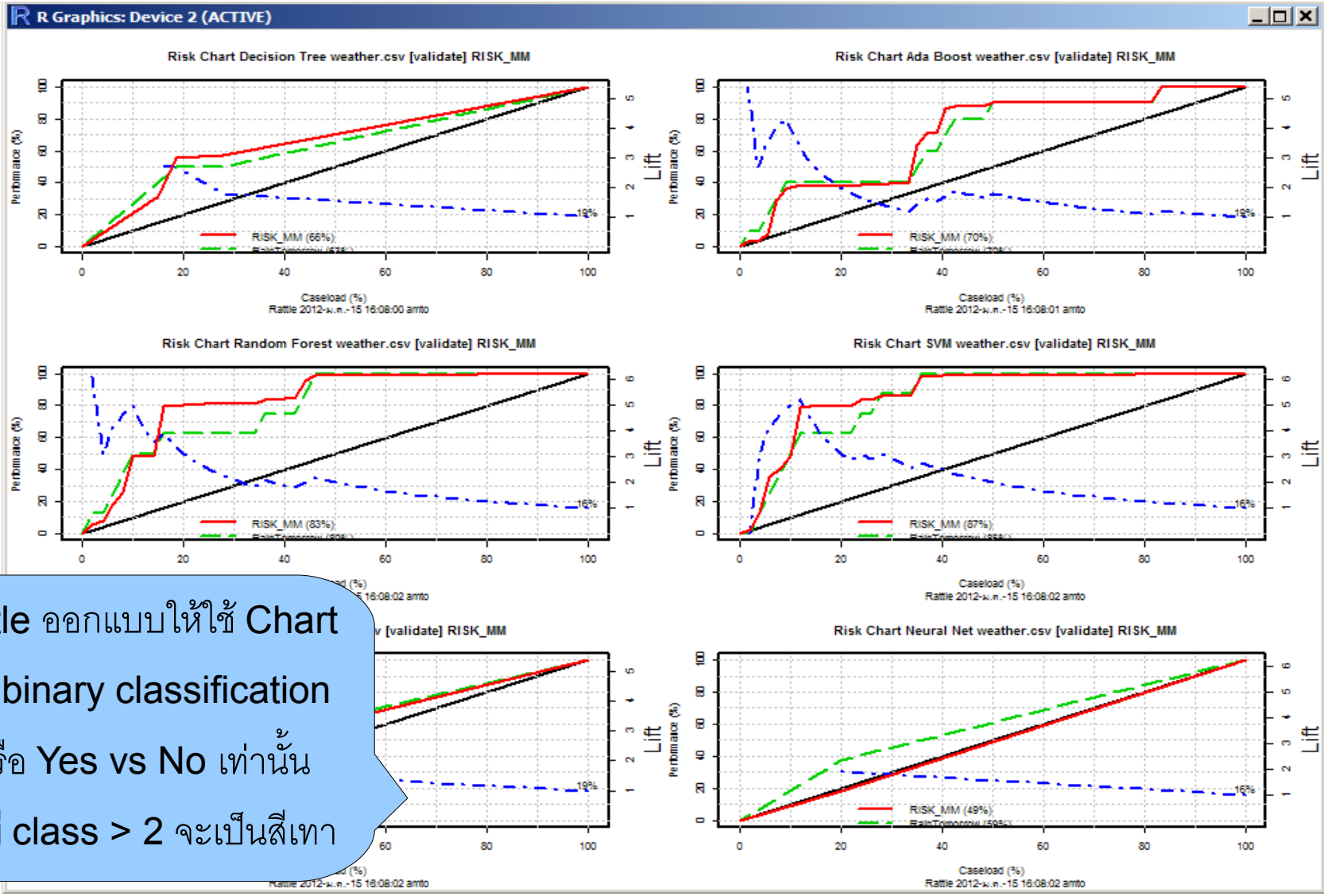
```
Error matrix for the Linear

      Predicted
Actual No Yes
   No  26  10
   Yes  4  10
```

```
Error matrix for the Random Forest

      Predicted
Actual No Yes
   No  34   2
   Yes  6   8
```

```
Error matrix for the Neural Net

      Predicted
Actual No Yes
   No  29   7
   Yes  6   8
```

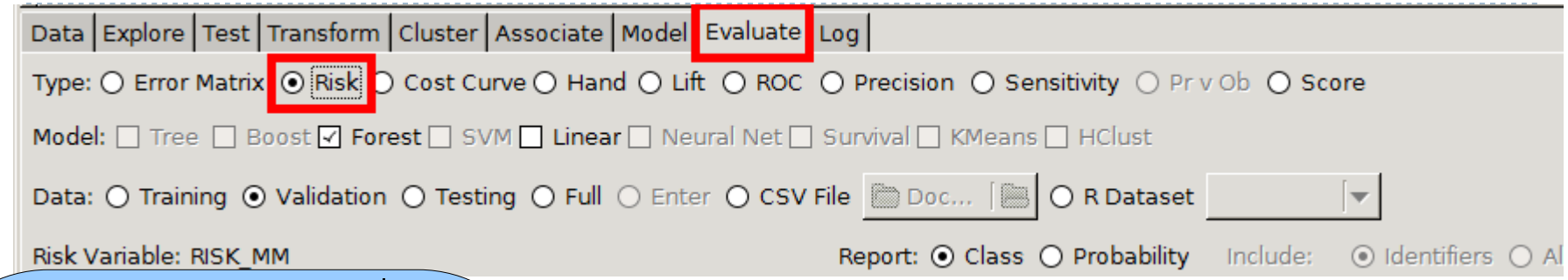
Data ที่ใช้คือ Test set.

Rattle Training 10. Model Performance Evaluation.



Rattle ออกแบบให้ใช้ Chart กับ binary classification หรือ Yes vs No เท่านั้น ถ้ามี class > 2 จะเป็นสีเทา

Rattle Training 10. Model Performance Evaluation.

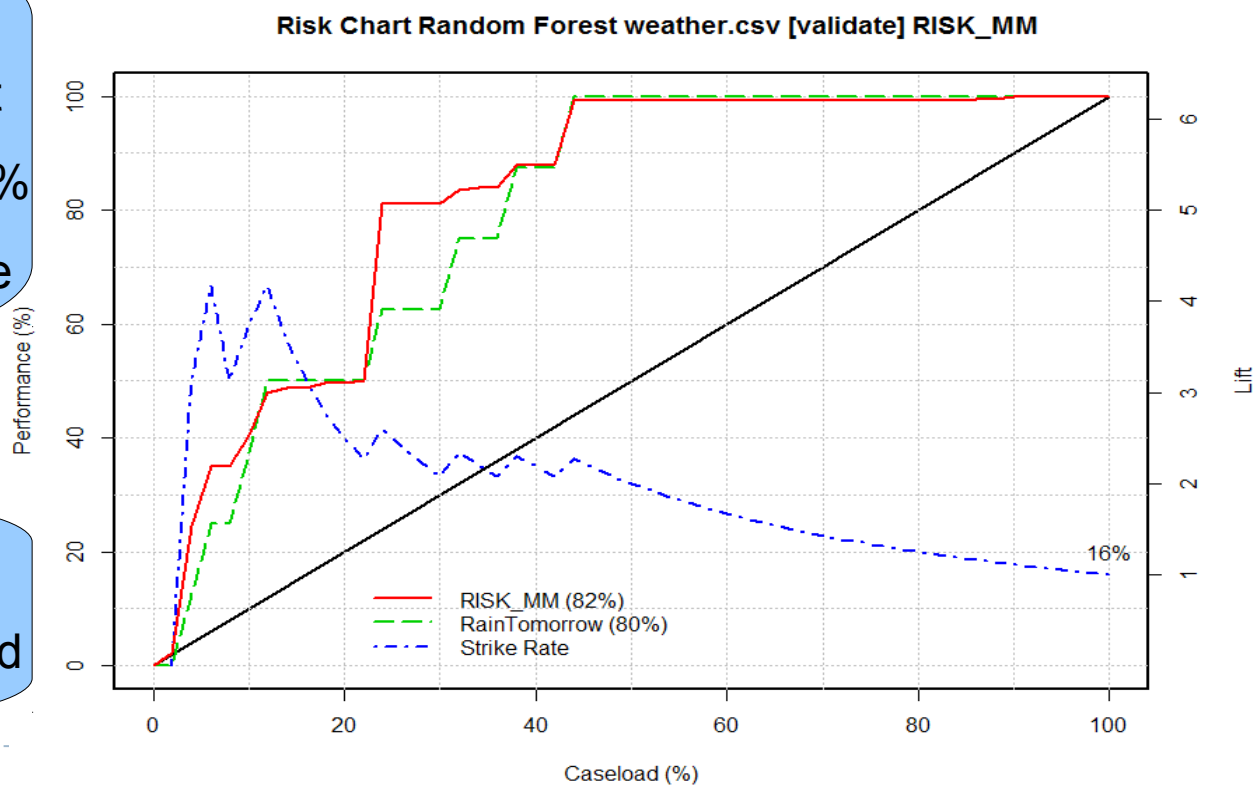


เส้นสีแดงคือค่าความเสี่ยง

เส้นสีเขียวคือ true respondent

เส้นสีดำแทนการสุ่มจึงถูกเพียง 50%

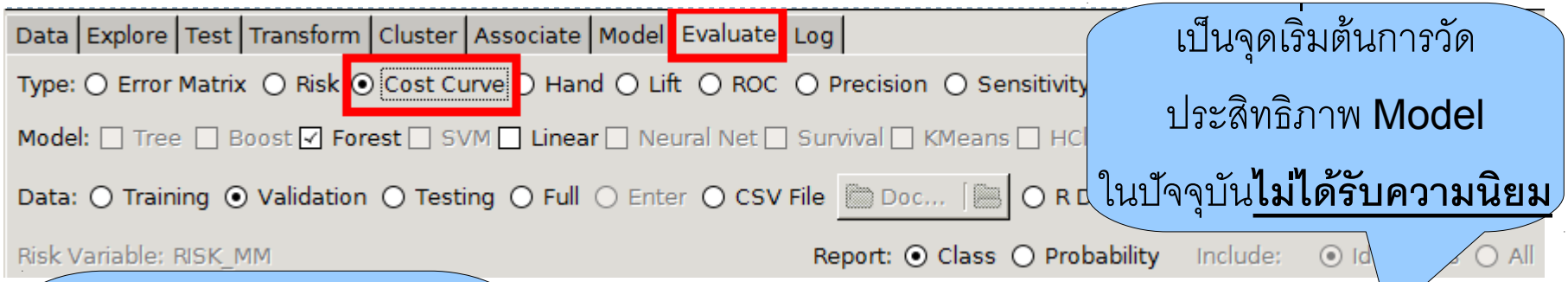
เส้นสีน้ำเงินคือ lift performance



จากกราฟนี้ความเสี่ยงเพิ่มขึ้น

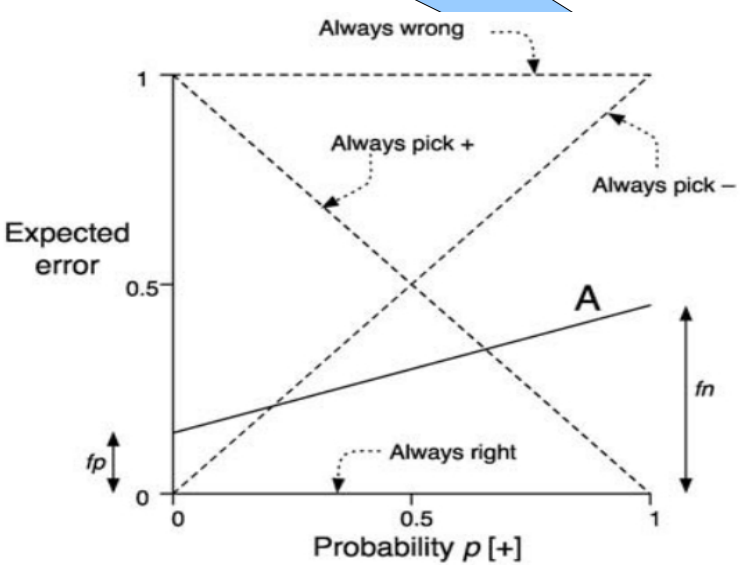
อย่างรวดเร็วที่ 20-30% caseload

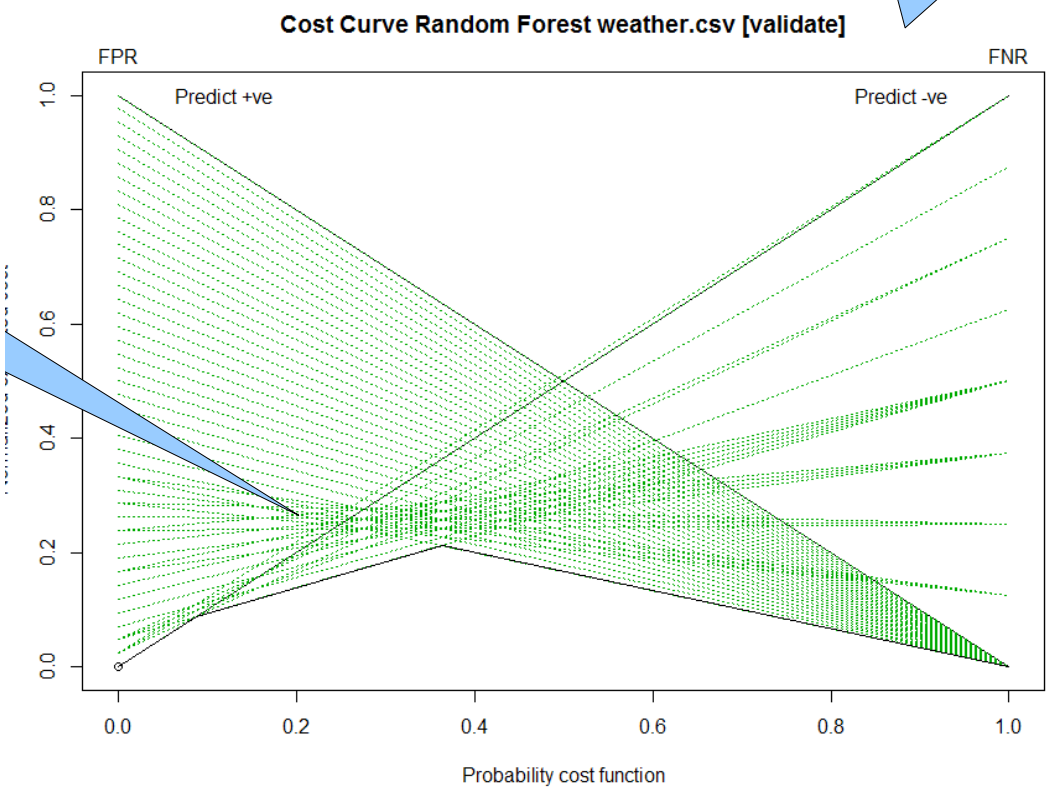
Rattle Training 10. Model Performance Evaluation.



เป็นจุดเริ่มต้นการวัดประสิทธิภาพ Model ในปัจจุบัน**ไม่ได้รับความนิยม**

สังเกตความหนาแน่นของเส้นแบนเข้าสู่ศูนย์กลาง (Normal line :A)





Rattle Training 10. Model Performance Evaluation.

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

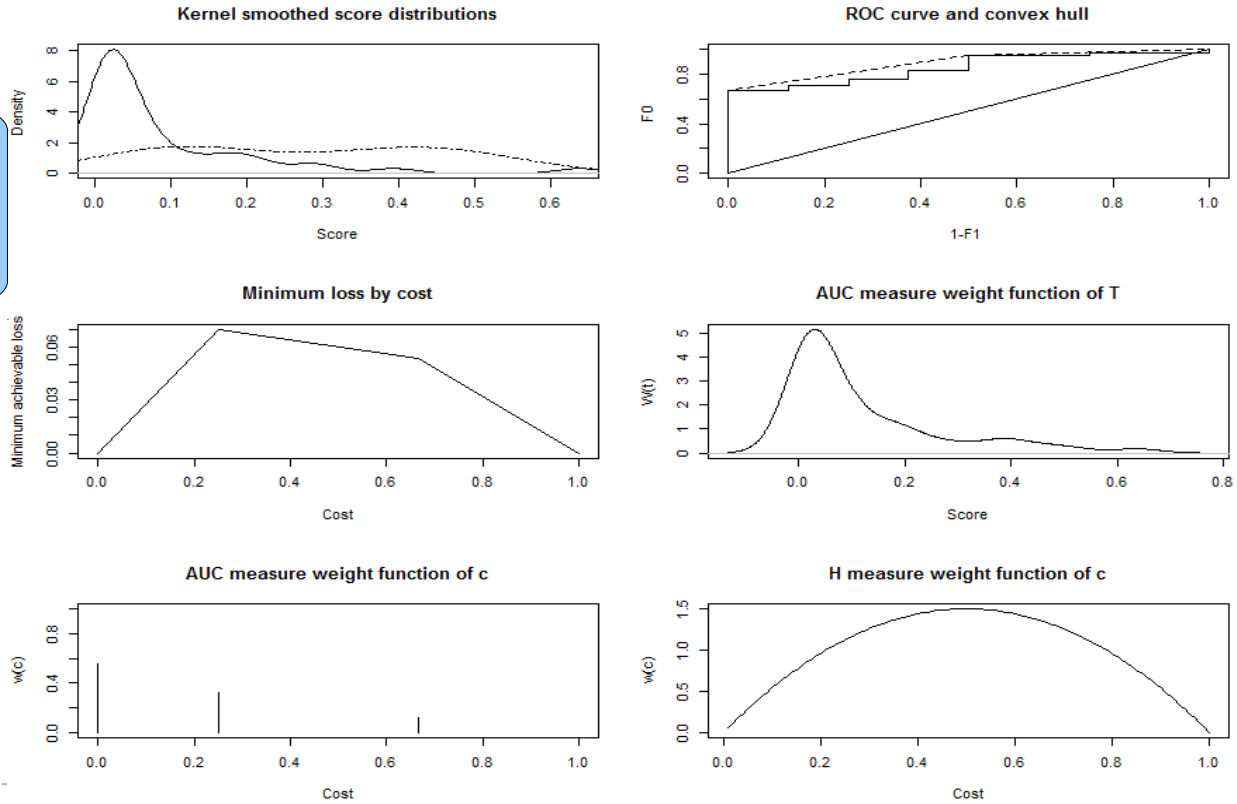
Type: ○ Error Matrix ○ Risk ○ Cost Curve ◉ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

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

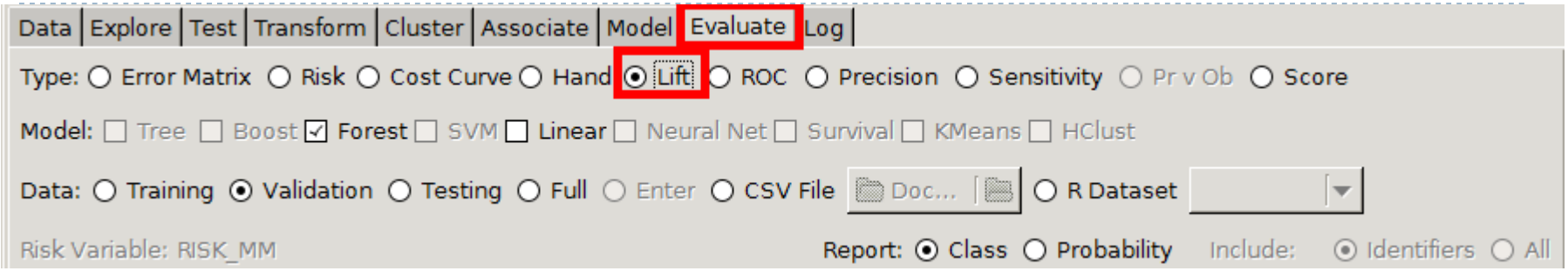
Data: ○ Training ◉ Validation ○ Testing ○ Full ○ Enter ○ CSV File Doc... ○ R Dataset

Risk Variable: RISK_MM Report: ◉ Class ○ Probability Include: ◉ Identifiers ○ All

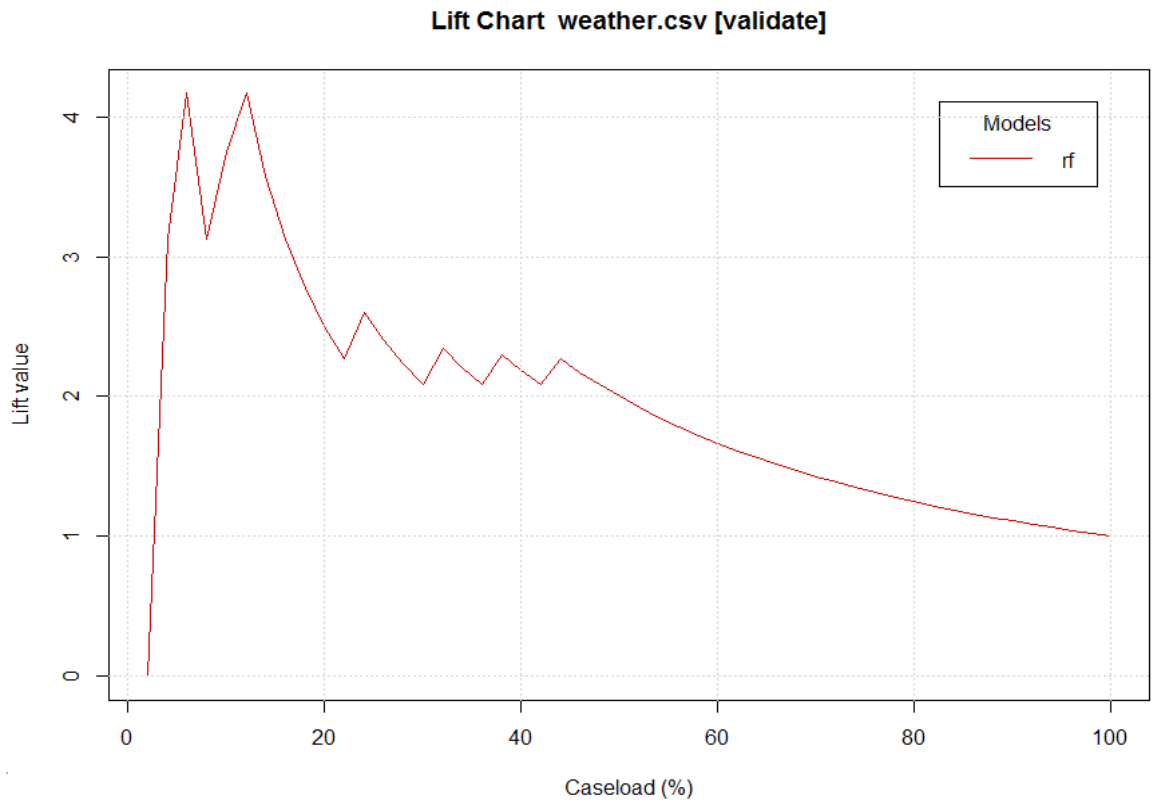
Hand chart นั้นเป็นเทคนิค
การวัดทางสถิติ โดย David Hand



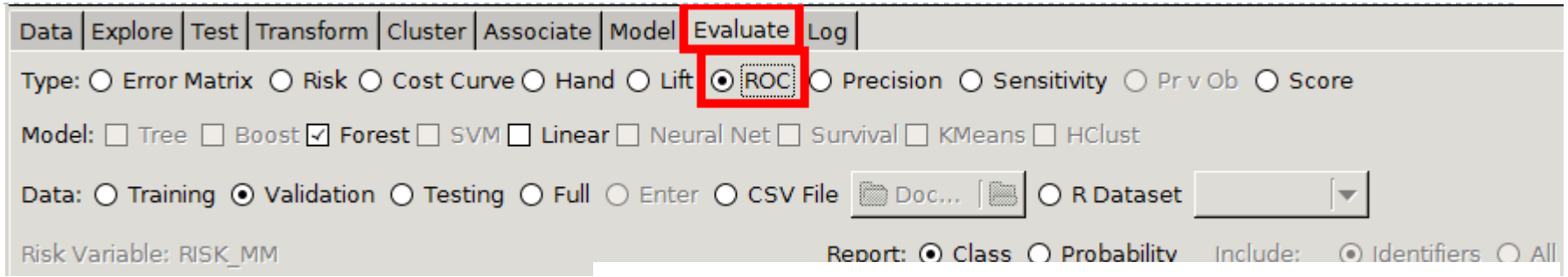
Rattle Training 10. Model Performance Evaluation.



Lift Charts แสดงให้เห็นโอกาส และความเชื่อมั่นต่อ Classifiers ในแต่ละสถานการณ์ ใน Risk chart คือเส้น strike rate



Rattle Training 10. Model Performance Evaluation.



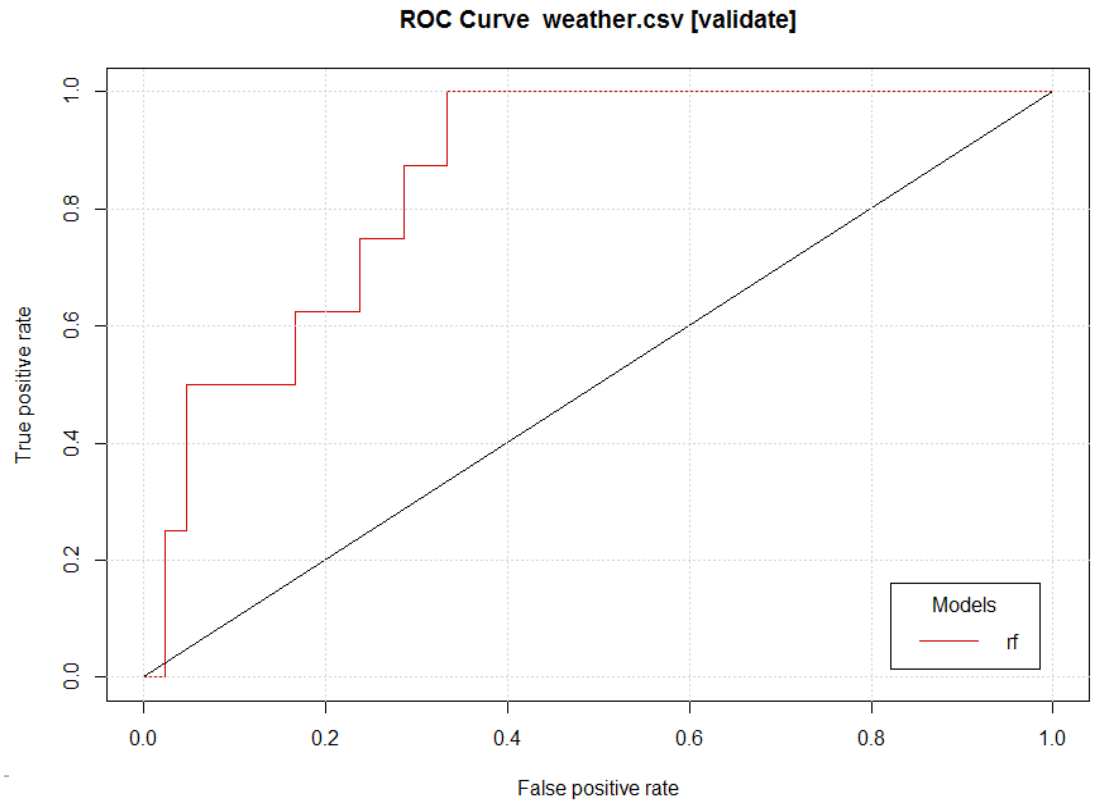
ROC Curve

-แกน y แสดง % of true positives

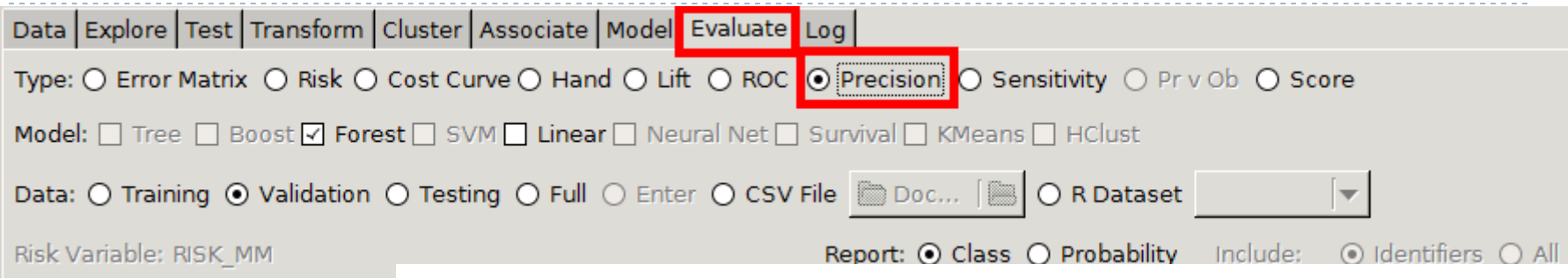
-แกน x แสดง % of false positives

-ROC ใช้ประเมินเปรียบเทียบ Classifiers ต่างกันบนข้อมูลเดียวกัน

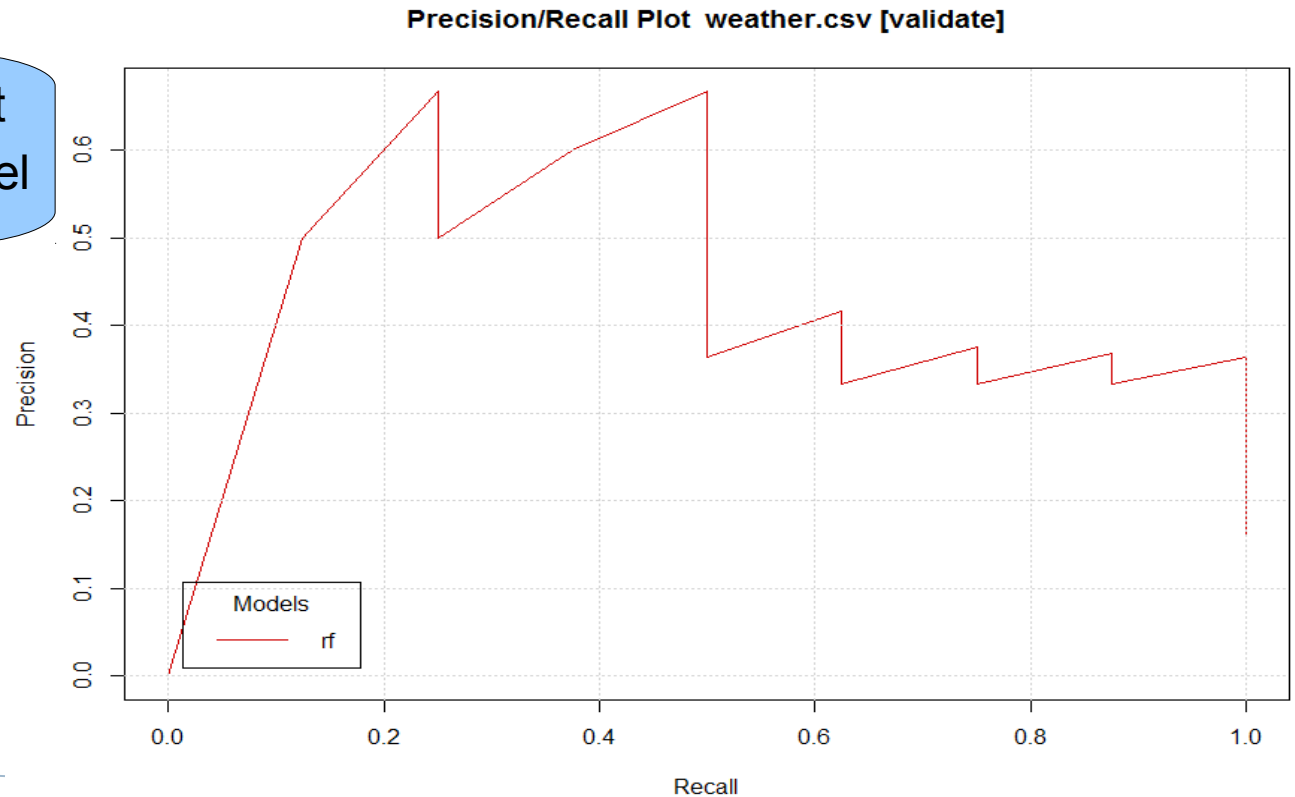
-ROC Curves ไม่ได้คำนึงถึง class distribution & costs



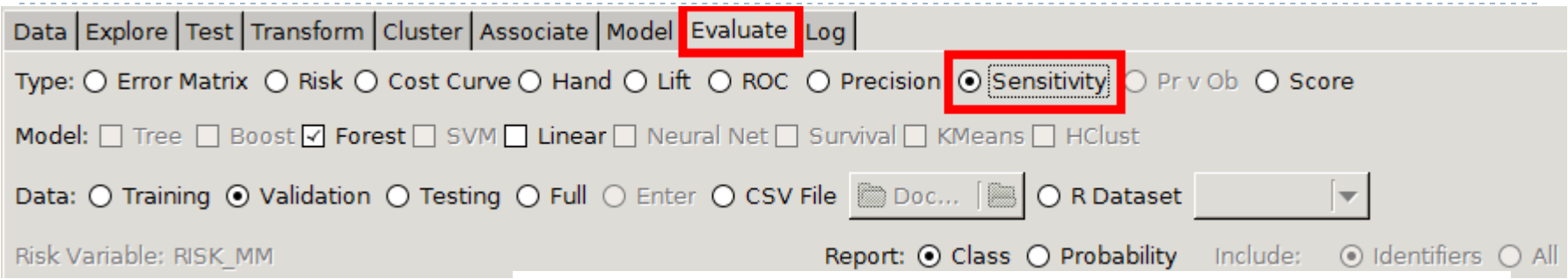
Rattle Training 10. Model Performance Evaluation.



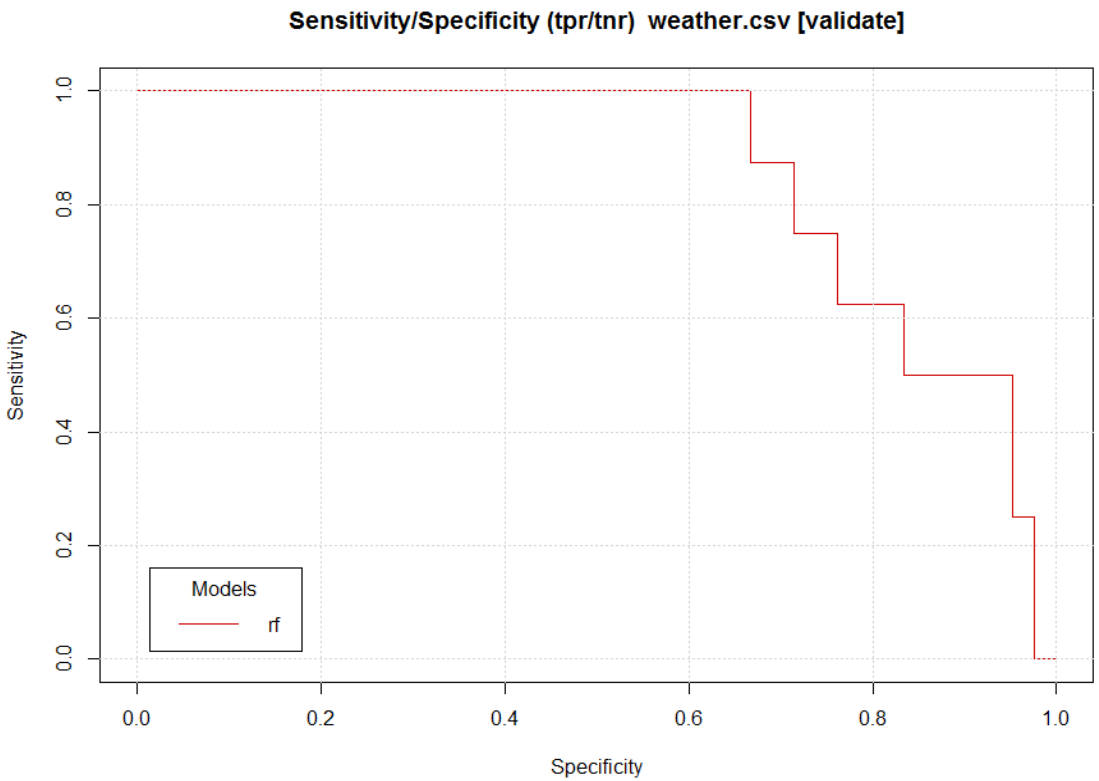
Precision /Recall chart
ใช้ประเมินคุณภาพของ Model



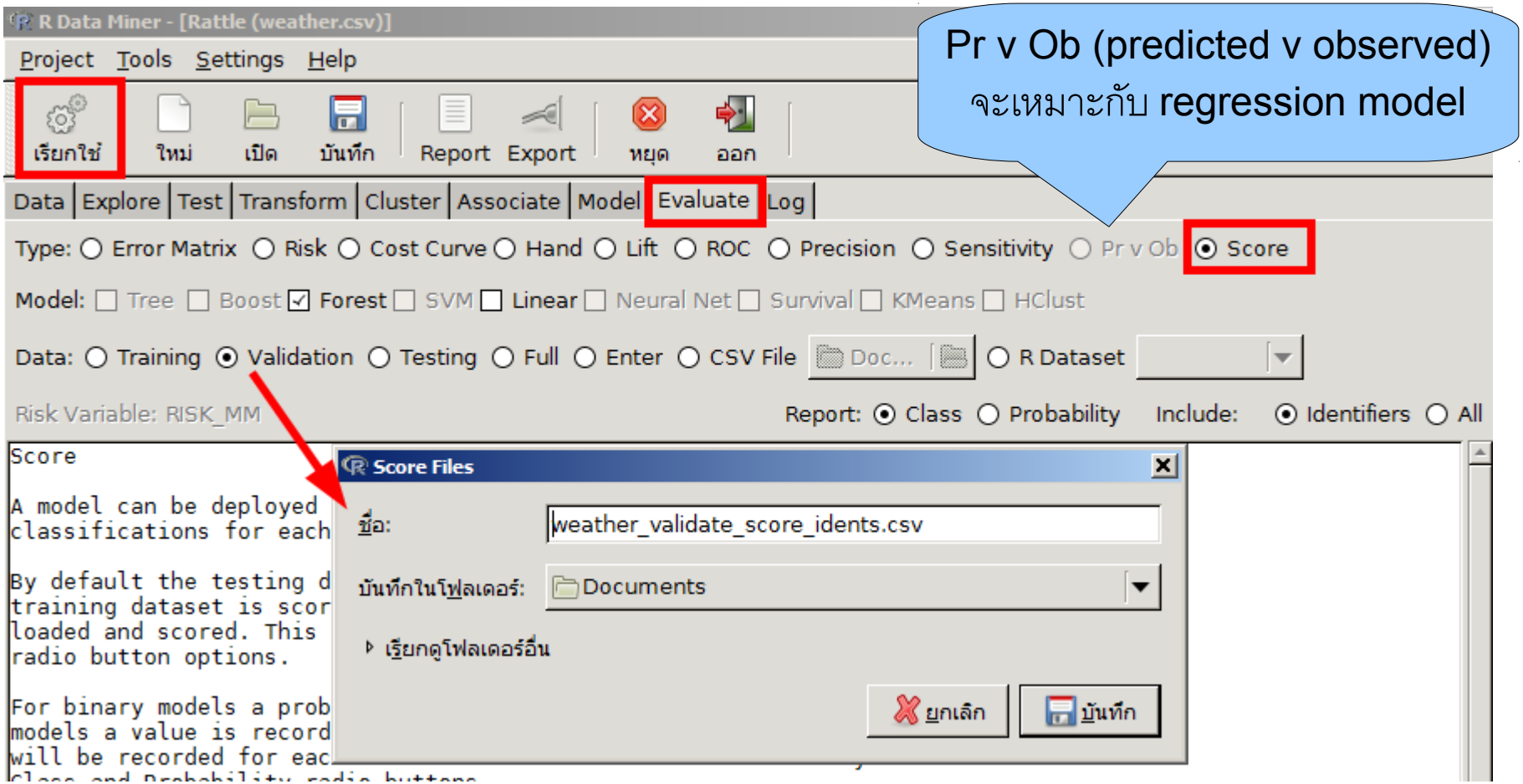
Rattle Training 10. Model Performance Evaluation.



Sensitivity /Specificity
ก็คือ True Positive Rate
หารด้วย True Negative Rate



Rattle Training 10. Model Performance Evaluation.

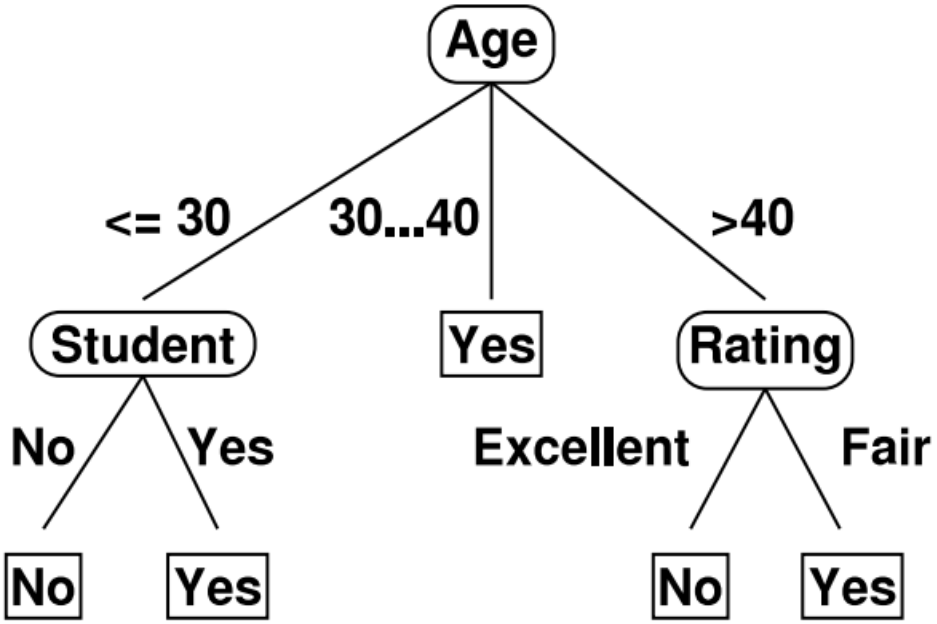


Pr v Ob (predicted v observed)
จะเหมาะกับ regression model

Workshop1 : decision tree for Buys Computer

buying computer.csv

Age	Income	Student	Credit	Buys
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	Yes
31 . . . 40	High	No	Fair	Yes
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
31 . . . 40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	No
> 40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31 . . . 40	Medium	No	Excellent	Yes
31 . . . 40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No



ใช้ Tree,Forest,SVM & Neural Net.

1.train ข้อมูล 100%

2.แบ่งข้อมูลเป็น train 60% และ test 40%

Workshop 1. Hint.

R Data Miner - [Rattle (buying computer.csv)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Source: (•) Spreadsheet ( ) ARFF ( ) ODBC ( ) R Dataset ( ) RData File ( ) Library ( ) Corpus ( ) Script

Filename: buying com... Separator: , Decimal: . [x] Header

[ ] Partition 70/15/15 Seed: 42 View Edit → **train 100%**

Input Ignore Weight Calculator:

Target Data Type: (•) Auto ( ) Categoric ( ) Numeric ( ) Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Age	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 3
2	Income	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 3
3	Student	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 2
4	Credit	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 2
5	Buys	Categoric	( )	(•)	( )	( )	( )	( )	Unique: 2

Workshop 1. Hint.

R Data Miner - [Rattle (buying computer.csv)]

Project Tools Settings Help

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด

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

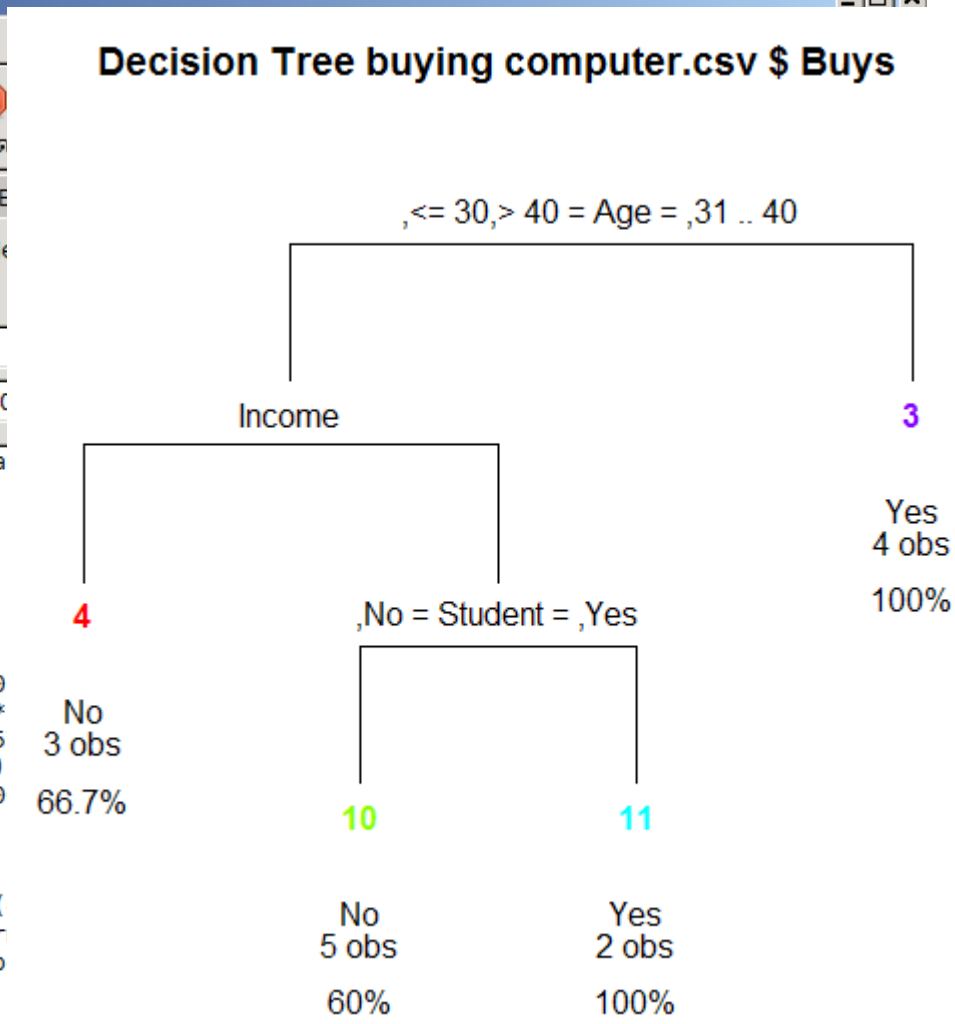
Type: Tree Forest Boost SVM Linear Ne

Target: Buys Algorithm: Traditional Conditional

Min Split: 6 Max Depth: 30

Min Bucket: 2 Complexity: 0.010

```
Summary of the Decision Tree model for Classifica

n= 14

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 14 5 Yes (0.3571429 0.6428571)
   2) Age=<= 30,> 40 10 5 No (0.5000000 0.5000000
     4) Income=Low 3 1 No (0.6666667 0.3333333) *
     5) Income=High,Medium 7 3 Yes (0.4285714 0.5
      10) Student=No 5 2 No (0.6000000 0.4000000)
      11) Student=Yes 2 0 Yes (0.0000000 1.000000
   3) Age=31 .. 40 4 0 Yes (0.0000000 1.0000000)

Classification tree:
rpart(formula = Buys ~ ., data = crs$dataset[, c(
    method = "class", parms = list(split = "infor
        minbucket = 2, usesurrogate = 0, maxsurro

Variables actually used in tree construction:
[1] Age     Income  Student

Root node error: 5/14 = 0.35714
```



Workshop 1. Hint.

R Data Miner - [Rattle (buying computer.csv)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score

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

Data: Training Validation Testing Full Enter CSV File Doc... R Dataset

Risk Variable: Report: Class Probability Include: Id

```
Error matrix for the Decision Tree model on buying computer.csv [**train**] (counts):

      Predicted
Actual No Yes
   No   5   0
   Yes  3   6

Error matrix for the Decision Tree model on buying computer.csv [**train**] (%):

      Predicted
Actual No Yes
   No  36   0
   Yes 21  43

Overall error: 0.2142857

Rattle timestamp: 2012-01-31 21:07:51 amto
=====================================================================
```

ถ้าปรับ min split =3 min bucket =1 error จะลดลง

Workshop 2.

- Workshop แบ่งเป็น 4 กลุ่ม
- กลุ่ม 1 : Boost
- กลุ่ม 2 : Random Forest
- กลุ่ม 3 : SVM
- กลุ่ม 4 : Neural Network

เงื่อนไข ทดสอบกับ Test set เพื่อหาผู้ชนะ

- 1.ใช้ไฟล์ weather.csv Test กับ weather_test.csv (use default seed)
- 2.ทดลองปรับแต่งค่าต่างๆ ตามกลุ่มแล้วหาผู้ชนะที่สามารถทำนายได้แม่นที่สุด
- 3.audit.Rdata แบ่ง Train 50% Test 50%

Workshop 2. Hint.

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.16 togaware.com

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Source: (•) Spreadsheet ( ) ARFF ( ) ODBC ( ) R Dataset ( ) RData File ( ) Library ( ) Corpus ( ) Script

Filename: weather.csv Separator: , Decimal: . [x] Header

[ ] Partition 70/15/15 Seed 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: (•) Auto ( ) Categoric ( ) Numeric ( ) Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	•	○	○	Unique: 366
2	Location	Constant	○	○	○	○	•	○	Unique: 1
3	MinTemp	Numeric	•	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	•	○	○	○	○	○	Unique: 187
5	Rainfall	Numeric	•	○	○	○	○	○	Unique: 47
6	Evaporation	Numeric	•	○	○	○	○	○	Unique: 55
7	Sunshine	Numeric	•	○	○	○	○	○	Unique: 114 Missing: 3
8	WindGustDir	Categoric	•	○	○	○	○	○	Unique: 16 Missing: 3
9	WindGustSpeed	Numeric	•	○	○	○	○	○	Unique: 35 Missing: 2
10	WindDir9am	Categoric	•	○	○	○	○	○	Unique: 16 Missing: 31
11	WindDir3pm	Categoric	•	○	○	○	○	○	Unique: 16 Missing: 1

Workshop 2. Hint.

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score

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

Data: Training Validation Testing Full Ente CSV File weat... R Dataset

Risk Variable: RISK_MM — Report: Class Probability Include: Id

```
Error matrix for the Decision Tree model on weather_test.csv (counts):

      Predicted
Actual No Yes
   No  33   1
   Yes 15  17

Error matrix for the Decision Tree model on weather_test.csv (%):

      Predicted
Actual No Yes
   No  50   2
   Yes 23  26

Overall error: 0.2424242

Rattle timestamp: 2012-01-31 21:48:24 amto
=====================================================================
```

weather_test.csv

Workshop 2. Hint.

R Data Miner - [Rattle (audit)]

Project Tools Settings Help — Rattle Version 2.6.16 togav

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

Filename: audit.RData Data Name: audit

☑ Partition 50/0/50 Seed: 42 View Edit

train/valid/test

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 Missing: 100
4	Education	Categoric	◉	○	○	○	○	○	Unique: 16
5	Marital	Categoric	◉	○	○	○	○	○	Unique: 6
6	Occupation	Categoric	◉	○	○	○	○	○	Unique: 14 Missing: 101
7	Income	Numeric	◉	○	○	○	○	○	Unique: 2000

Workshop 2. Hint.

R Data Miner - [Rattle (audit)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

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

Data: ○ Training ○ Validation ⦿ Testing ○ Full ○ Enter ○ CSV File Doc... ○ R Dataset

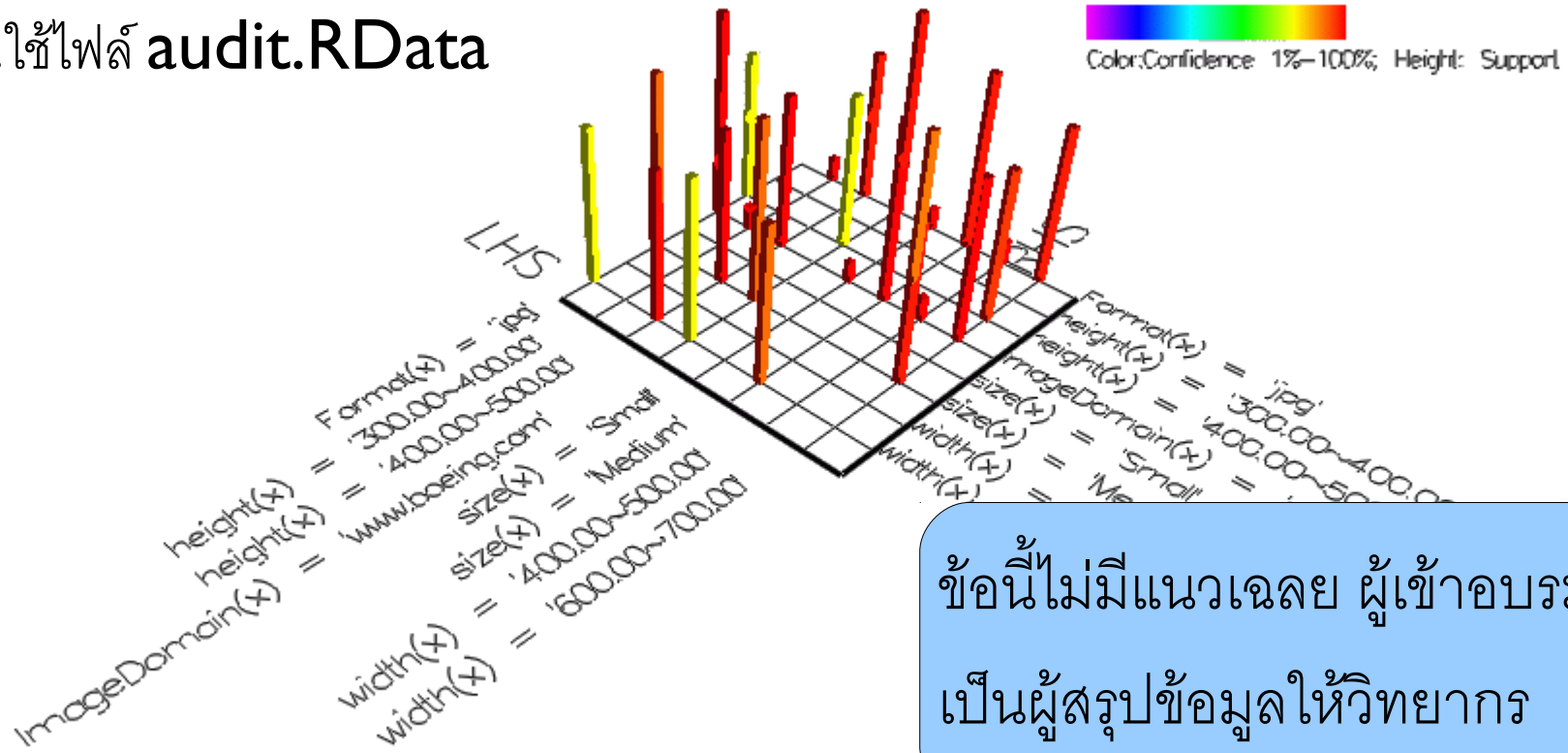
Risk Variable: RISK_Adjustment Report: ⦿ Class ○ Probability Include: ⦿ Id

```
Error matrix for the Decision Tree model on audit [test] (counts):

      Predicted
Actual   0   1
     0 728  41
     1 126 105

Error matrix for the Decision Tree model on audit [test] (%):

      Predicted
Actual  0  1
     0 73  4
     1 13 10

Overall error: 0.167

Rattle timestamp: 2012-01-31 21:58:23 amto
========================================================================
```

Workshop 3.

ใช้ **Associate rule** ค้นหาความรู้จาก

- 1. weather.csv
- 2.ใช้ไฟล์ audit.RData



ข้อนี้ไม่มีแนวเฉลย ผู้เข้าอบรม เป็นผู้สรุปข้อมูลให้วิทยากร

Workshop 4.

- 1.นำเข้าไฟล์ heart_for_rattle.txt เข้าสู่ Rattle
- 2.แบ่ง Partition ข้อมูลเป็น train 70 test 30 (disease=target)
- 3.สร้าง Box Plot & Histogram ของตัวแปร age & chol
- 4.สร้าง Mosaic of sex & restecg By disease
- 5.ใช้ Hierarchical Method: pearson ระบุว่าคู่ตัวแปรใดที่สัมพันธ์กัน
- 6.ใช้ค่า default ของ Decision tree สร้าง Model
- 7.ใช้ค่า default ของ Boost สร้าง Model
- 8.ใช้ค่า default ของ Random Forest สร้าง Model
- 9.เปรียบเทียบ Overall error จาก Confusion Matrix ของ 3 model
- 10.เปรียบเทียบ ROC ของทั้ง 3 model

Workshop 4. Hint.

Rattle Dataset

Rattle Dataset	age	sex	chest_pa	trestbps	chol	fbs	restecg	thalach	exang	disease
1	31	male	asympt	120	270	f	normal	153	yes	positive
2	33	female	asympt	100	246	f	normal	150	yes	positive
3	34	male	p_angina	140	156	f	normal	180	no	positive
4	35	male	p_angina	110	257	f	normal	140	no	positive
5	36	male	p_angina	120	267	f	normal	160	no	positive
6	37	male	asympt	140	207	f	normal	130	yes	positive
7	38	male	asympt	110	196	f	normal	166	no	positive
8	38	male	asympt	120	282	f	normal	170	no	positive
9	38	male	asympt	92	117	f	normal	134	yes	positive
10	41	male	asympt	110	289	f	normal	170	no	positive

Workshop 4. Hint.

R Data Miner - [Rattle (heart_for_rattle)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Source: (•) Spreadsheet ( ) ARFF ( ) ODBC ( ) R Dataset ( ) RData File ( ) Library ( ) Corpus ( ) Script

Filename: heart_for_ra... Separator: \t Decimal: . [x] Header

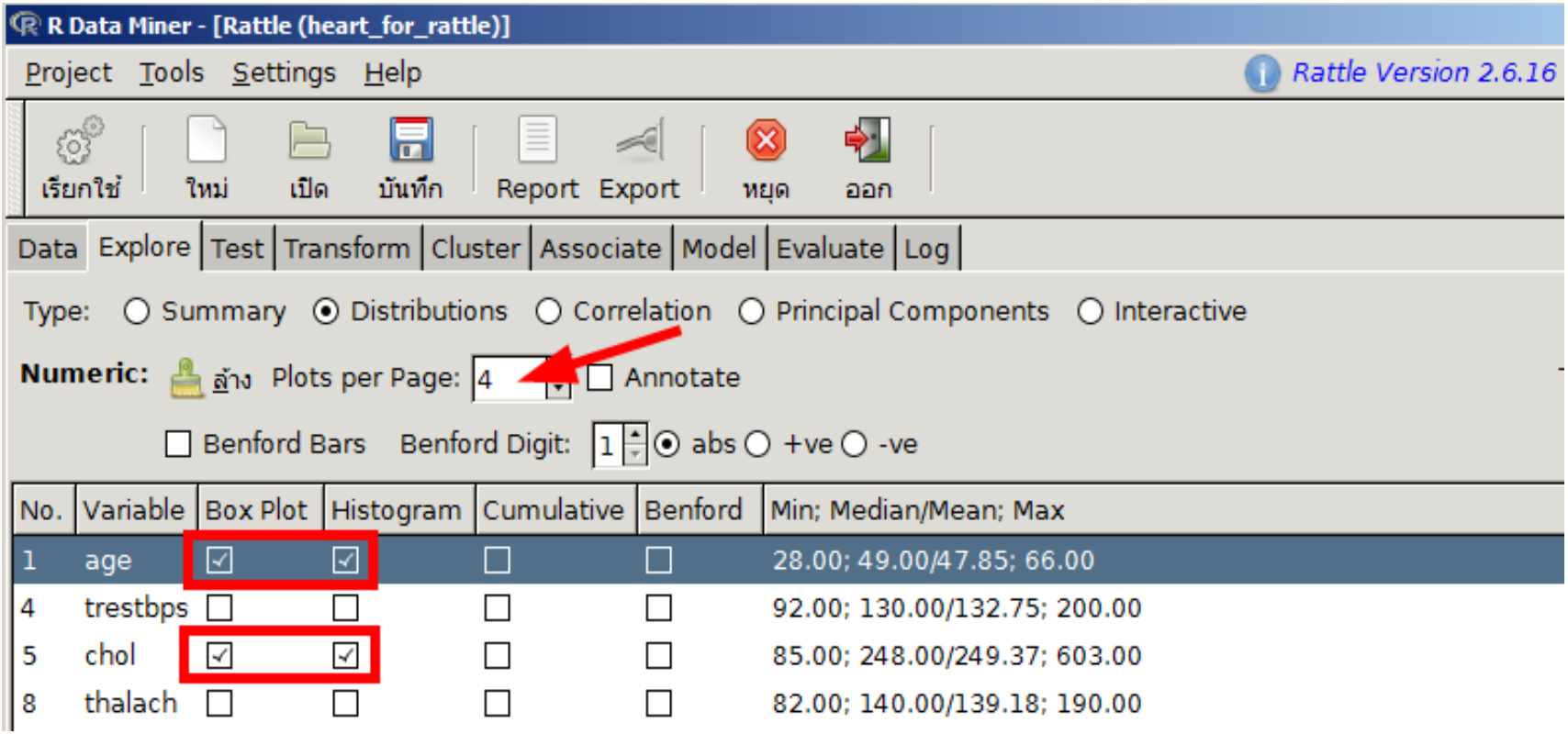
[x] Partition 70/0/30 Seed: 42 View Edit

Input Ignore Weight Calculator:

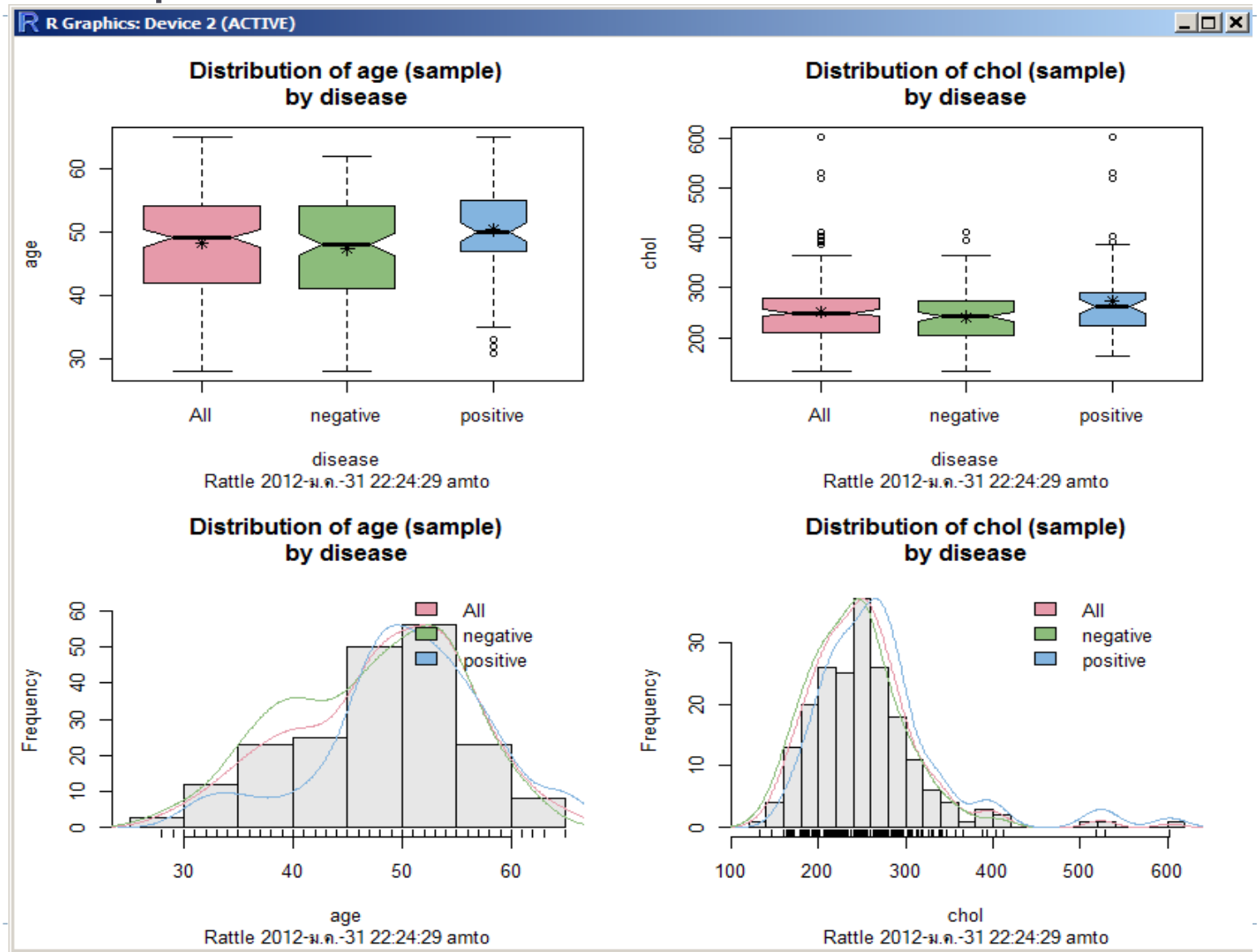
Target Data Type: (•) Auto ( ) Categoric ( ) Numeric ( ) Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	age	Numeric	(•)	( )	( )	( )	( )	( )	Unique: 38
2	sex	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 2
3	chest_pain	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 4
4	trestbps	Numeric	(•)	( )	( )	( )	( )	( )	Unique: 31
5	chol	Numeric	(•)	( )	( )	( )	( )	( )	Unique: 150
6	fbs	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 2
7	restecg	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 3
8	thalach	Numeric	(•)	( )	( )	( )	( )	( )	Unique: 69
9	exang	Categoric	(•)	( )	( )	( )	( )	( )	Unique: 2
10	disease	Categoric	( )	(•)	( )	( )	( )	( )	Unique: 2

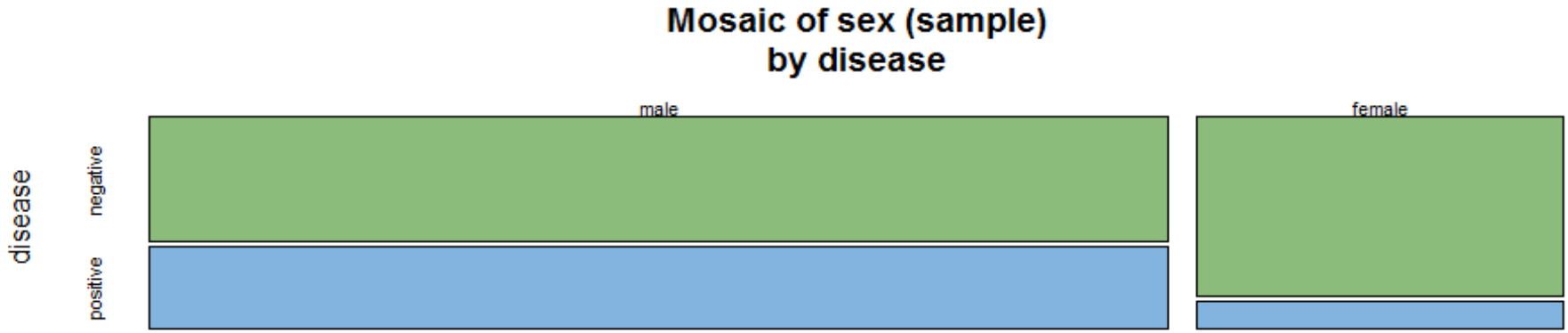
Workshop 4. Hint.



Workshop 4. Hint.

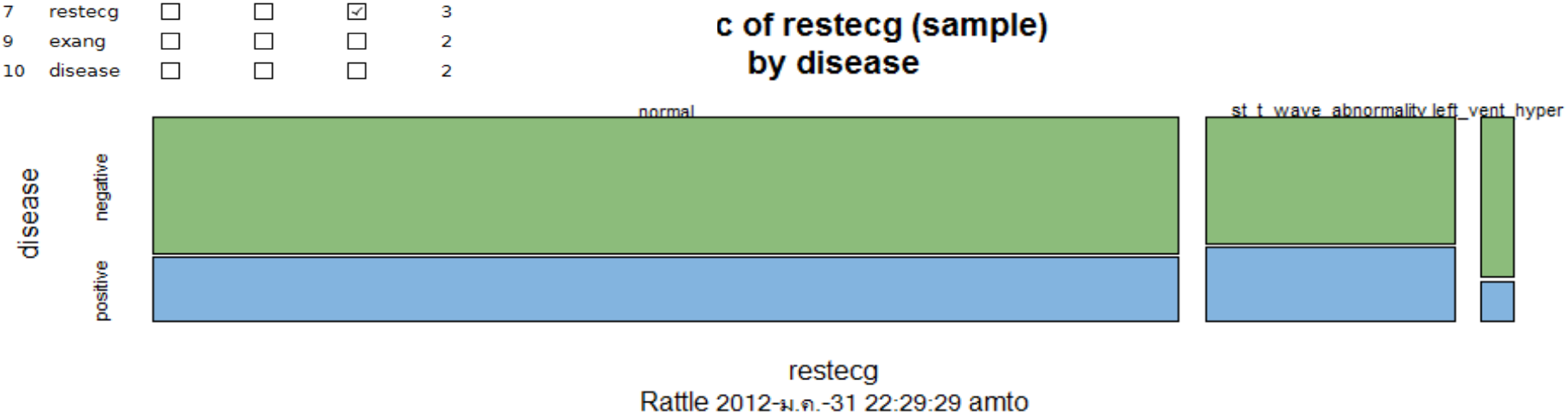


Workshop 4. Hint.

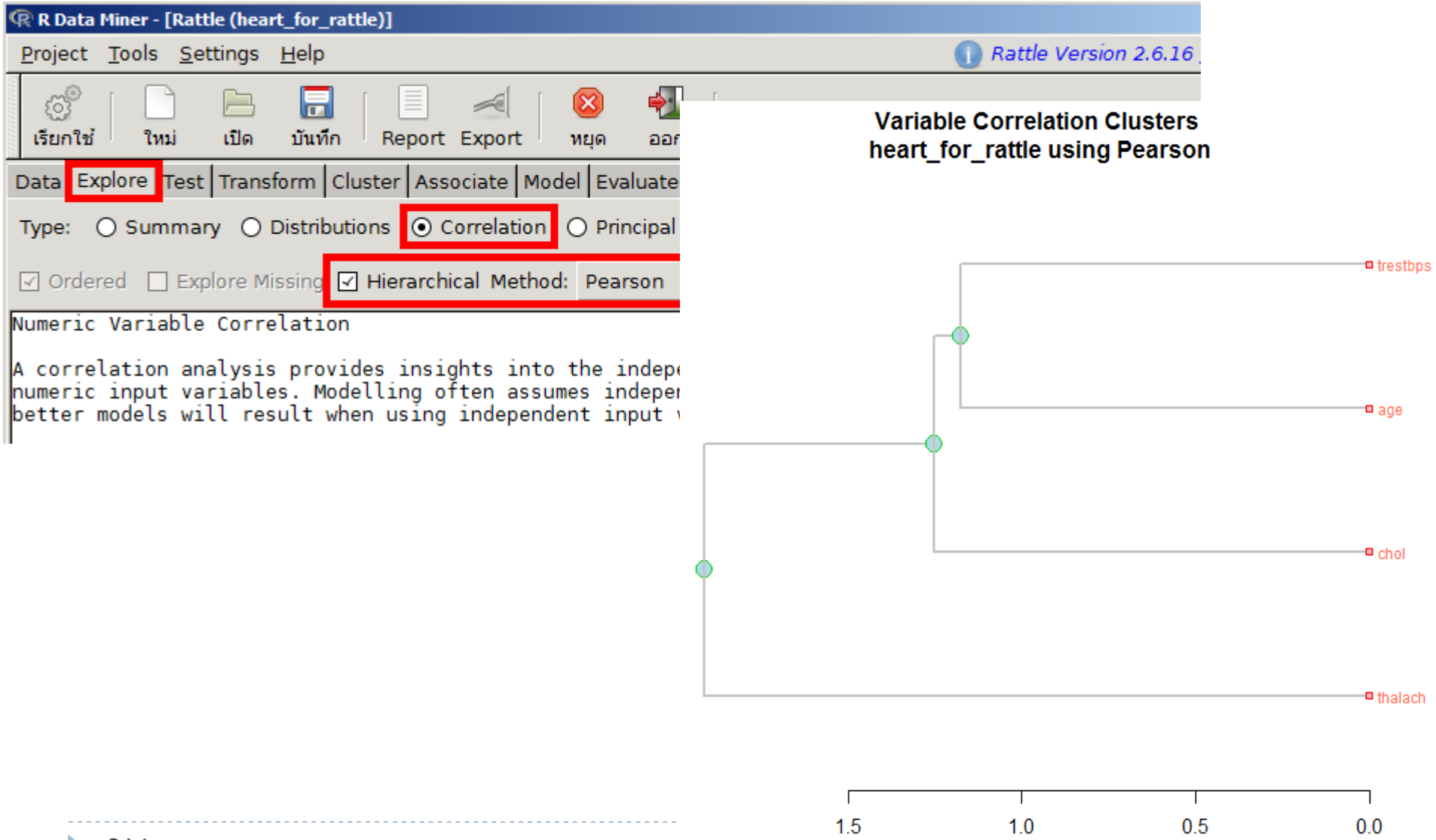


Categoric: ล้าง

No.	Variable	Bar Plot	Dot Plot	Mosaic	Levels
2	sex	☐	☐	☑	2
3	chest_pain	☐	☐	☐	4
6	fbs	☐	☐	☐	2
7	restecg	☐	☐	☑	3
9	exang	☐	☐	☐	2
10	disease	☐	☐	☐	2



Workshop 4. Hint.



Workshop 4. Hint.

R Data Miner - [Rattle (heart_for_rattle)]

Project Tools Settings Help Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score

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

Data: Training Validation Testing Full Enter CSV File Doc... R Dataset

Risk Variable: Report: Class Probability Include: Ide

```
Error matrix for the Decision Tree model on heart_for_rattle [test] (counts):

          Predicted
Actual     negative positive
  negative       45        2
  positive       19       20

Error matrix for the Decision Tree model on heart_for_rattle [test] (%):

          Predicted
Actual     negative positive
  negative       52        2
  positive       22       23

Overall error: 0.244186
```

Workshop 4. Hint.

R Data Miner - [Rattle (heart_for_rattle)]

Project Tools Settings Help — Rattle Version 2.6.16

เรียกใช้ ใหม่ เปิด บันทึก Report Export หยุด ออก

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

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

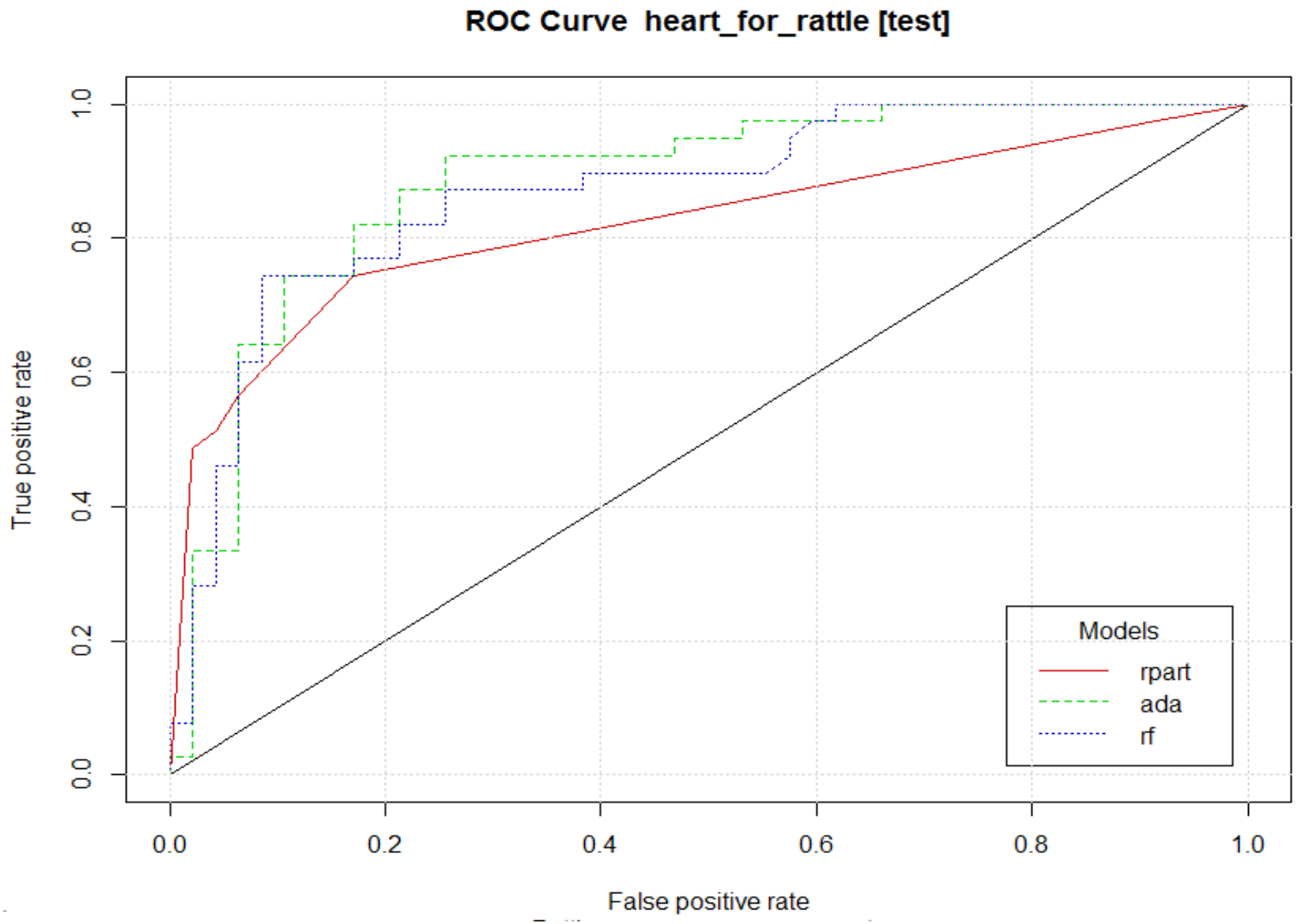
Model: ☑ Tree ☑ Boost ☑ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ○ Training ○ Validation ◉ Testing ○ Full ○ Enter ○ CSV File Doc... ○ R Dataset

Risk Variable: Report: ◉ Class ○ Probability Include: ◉ Ide

```
Area under the ROC curve for the rpart model on heart_for_rattle [test] is 0.8205

Rattle timestamp: 2012-01-31 22:38:01 amto
========================================================================
Area under the ROC curve for the ada model on heart_for_rattle [test] is 0.8833

Rattle timestamp: 2012-01-31 22:38:02 amto
========================================================================
Area under the ROC curve for the rf model on heart_for_rattle [test] is 0.8691

Rattle timestamp: 2012-01-31 22:38:02 amto
========================================================================
```

Workshop 4. Hint.



จบการนำเสนอ

Data mining and statistic tools is the way to find knowledge and get wealthy!

Welcome to use R , Rattle, RapidMiner , KNIME,and WEKA
the world leading analytic open source software.

- www.reinforcebi.com
- info@reinforcebi.com
- 08 1491 0909