1

การค้นหาความรู้โดยการวิเคราะห์ข้อมูลด้วย R และ Weka

Knowledge Discovery by Data Analysis with R and Weka

สุณี รักษาเกียรติศักดิ์

การค้นหาความรู้

การค้นหาความรู้ในสิ่งที่เราอยากทราบหรือเราอาจจะเรียกว่ากระบวนการวิจัย สามารถดำเนินการได้ 2 แนวทางคือ forward และ backward

Forward เริ่มจากวัตถุประสงค์/คำถามการวิจัย หรือโจทย์หรือปัญหาการวิจัยนั่นเอง (Objective หรือ Problem definititions) จากนั้นก็จะเป็นวิธีการดำเนินการ (ซึ่งอาจจะต้องมีกระบวนการทบทวนวรรณกรรม – review literature หรือไม่ต้องมีก็ได้ เพราะโจทย์อาจจะเป็นปัญหาที่เกี่ยวกับการปฏิบัติงาน ซึ่งก็ดำเนินการได้เลย) ในวิธีการดำเนินงาน จะเกี่ยวข้องกับการออกแบบและ/หรือการพัฒนาเพื่อให้ได้คำตอบของโจทย์ที่ต้องการ ในกระบวนการนี้อาจจะต้องมีการออกแบบเครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลเพื่อตอบโจทย์ เครื่องมืออย่างง่ายที่ใช้กันคือการออกแบบสอบถาม จากนั้นก็เป็นการกำหนดกลุ่มตัวอย่างที่จะตอบแบบสอบถาม (sample) ซึ่งตามหลักการแล้วก็ต้องมีการนิยามประชากรที่เราสนใจจะศึกษา (population) ด้วย และระบุว่ากลุ่มตัวอย่างที่ได้มาและเป็นตัวแทน (representative) ของประชากรนั้น ได้มาด้วยวิธีการสุ่มตัวอย่าง (sampling) แบบใด เช่น การสุ่มอย่างง่าย (random sampling) ซึ่งคือการสุ่มที่ทุกหน่วยของประชากรมีโอกาสได้รับการสุ่มเข้ามาด้วยความน่าจะเป็นที่เท่ากัน หรือการสุ่มแบบแบ่งชั้น (stratify sampling) ซึ่งคือการสุ่มที่ให้ได้สัดส่วนของตัวแปรที่ต้องการแบ่งชั้นตามต้องการ เป็นต้น แต่การวิเคราะห์ข้อมูลเพื่อตอบโจทย์ในการทำงานมักใช้กลุ่มตัวอย่างตามสะดวก (convenient sample) ตัวอย่างเช่น ในการปฏิบัติงานทุกอย่าง ต้องมีการประเมินผลการปฏิบัติงานของกิจกรรมใด กิจกรรมหนึ่ง ก็ใช้กลุ่มตัวอย่างของผู้ที่เข้ามาร่วมกิจกรรม เป็นตัวแทนของประชากรที่ต้องการ

ในการวิเคราะห์ข้อมูลนั้นเราศึกษาจากกลุ่มตัวอย่าง (sample) ที่เรามีข้อมูล โดยคำนวณค่าสถิติ (statistics) ต่าง ๆ จากกลุ่มตัวอย่าง ที่เราเรียกว่าสถิติพรรณนา (descriptive statistics) ทำให้เราได้ความรู้เกี่ยวกับลักษณะของกลุ่มตัวอย่างที่เราศึกษา แต่เราจะขยายผลของการศึกษาจากกลุ่มตัวอย่างนี้ไปอนุมาน (inference) เพื่อให้ได้ความรู้เกี่ยวกับประชากร (parameters) ได้โดยกระบวนการทดสอบสมมุติฐานทางสถิติ (statistical hypothesis testing) ซึ่งเป็นกระบวนการใช้สถิติอนุมาน/สถิติอ้างอิง (inferential statisitcs) ซึ่งคำนวณจากสถิติพรรณนา (descriptive statistics) สำหรับการทดสอบคุณลักษณะของประชากร (parameters) ที่ต้องการ โดยใช้หลักการความน่าจะเป็น (probability distribution) ของการแจกแจง (distribution) ของสถิติอ้างอิงนั้น เพื่อสรุปความจริงเกี่ยวกับประชากร

สถิติอนุมานมี 2 แบบ คือ สถิติพาราเมตริก (parametric) และสถิตินอนพาราเมตริก (non parametric) สถิติพาราเมตริกจะมีข้อตกลงเบื้องต้น (assumption) ข้อมูล/ตัวแปร (variable) ต้องมีการแจกแจงความน่าจะเป็นแบบปกติ (normal distribution) ซึ่งโดยธรรมชาติแล้ว ข้อมูลที่เกิดโดยธรรมชาติจะมีการแจกแจงแบบปกติ ถ้าข้อมูลไม่มีการแจกแจงแบบปกติควรจะต้องใช้สถิตินอนพาราเมตริกแทน ในการคำนวณของสถิตินอนพาราเมตริกจะไม่ใช้ข้อมูลดิบของตัวแปรมาคำนวณ แต่จะใช้อันดับ (rank) ของข้อมูลมาคำนวณ ซึ่งเป็นวิธีการที่หยาบกว่า ดังนั้นหากสามารถจะใช้วิธีการทางสถิติ (statistical methods) ที่เรียกว่าพาราเมตริกได้ ก็ควรทำในเบื้องแรก หากไม่สามารถทำได้ เช่น ข้อมูลไม่มีการแจกแจงแบบปกติ หรือข้อตกลงเบื้องต้นของวิธีการทางสถิตินั้นไม่เป็นจริง ก็ค่อยไปใช้วิธีการแบบนอนพาราเมตริกแทน ข้อได้เปรียบของวิธีการพาราเมตริกคือให้กำลังการทดสอบ (power) สูงกว่าวิธีการนอนพาราเมตริก

วิธีการทาง backward มี 2 กรณี คือ

กรณีที่ 1: เราอาจจะไม่ได้ระบุคำถามการวิจัยที่ชัดเจนนักหรือเพียงระบุวัตถุประสงค์กว้าง ๆ ไว้ แล้วเก็บข้อมูลต่าง ๆ ที่คาดว่าจะเกี่ยวข้องกับวัตถุประสงค์ของโจทย์ เช่น **วัตถุประสงค์ คือ ต้องการศึกษาการฝึกงานของนิสิตเอกวิทยาการคอมพิวเตอร์ (วิชา CP493 Internship)** ซึ่งเป็นวิชาบังคับของหลักสูตร ข้อมูลและเครื่องมือที่จะใช้ในการเก็บรวบรวมข้อมูลที่จะตอบโจทย์ยังไม่ชัดเจน (ซึ่งจะเป็นโจทย์ให้ผู้อบรมได้ฝึกปฏิบัติต่อไป)

กรณีที่ 2: การค้นหาความรู้จากข้อมูลที่อยู่ในฐานข้อมูลการปฏิบัติงาน (Knowledge Discovery in Database หรือ KDD) การค้นหาความรู้ในกรณีนี้จะใช้วิธีการทางสถิติ (statistical methods) หรือวิธีการทางการทำเหมืองข้อมูล (data mining) ด้วยวิธีการทางการเรียนรู้ของเครื่อง (machine learning) ก็ได้

การค้นหาความรู้โดยวิธีการเรียนรู้ของเครื่องจะตั้งอยู่บนสมมติฐานว่าจะแบ่งข้อมูลออกเป็น 2 ชุด ชุดหนึ่งคือชุดในการสร้างโมเดล (training set) ซึ่งจะเป็นข้อมูลส่วนใหญ่ และข้อมูลอีกชุดหนึ่งเรียกว่าชุดทดสอบโมเดล (test set) ข้อมูลสองชุดนี้ต้องเป็นอิสระต่อกัน (independent) คือต้องไม่ซ้ำกัน

เมื่อทำการ train แล้ว จะได้ผลลัพธ์เป็นความรู้หรือโมเดล ซึ่งอาจจะอยู่ในรูปสมการ หรือ กฎ เป็นต้น โมเดลที่ได้จะมีความน่าเชื่อถือเพียงใดต้องมีการประเมินโมเดล โดยจะใช้ test set (จะกล่าวในรายละเอียดต่อไป)

วัตถุประสงค์ของการอบรม

1. แสดงการใช้เครื่องมือออนไลน์ Google WORK (FORMS และ SHEETS) ในการเก็บรวบรวมข้อมูล
2. แสดงการวิเคราะห์ข้อมูลอย่างง่ายโดย Google สำหรับข้อมูลที่จัดเก็บโดย FORMS และวิเคราะห์ข้อจำกัดของการประมวลผลโดย Google จากโจทย์ **“การฝึกงานของนิสิต”**
3. แสดงการนำข้อมูลจาก Google WORK มาวิเคราะห์ข้อมูลทางสถิติโดยใช้โปรแกรม R โดยมีการสร้างตัวแปรใหม่ วิเคราะห์สถิติพรรณา สถิติอนุมาน เช่น การเปรียบเทียบค่าเฉลี่ย 2 กลุ่มด้วย t-test การ

เปรียบเทียบค่าเฉลี่ยมากกว่า 2 กลุ่มด้วย One way ANOVA และการทดสอบความแตกต่างของความถี่ของกลุ่มด้วยไคร์สแคว์

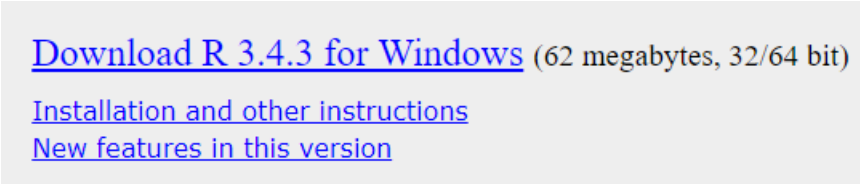
4. ใช้ข้อมูลของ Weka เป็นตัวอย่างในสร้างโมเดลโดยใช้การวิเคราะห์ Regression และ Logistic Regression โดยวิเคราะห์ด้วยโปรแกรม R
5. ใช้ Weka ในการทำ data mining โดยใช้ข้อมูลตัวอย่างของ Weka และวิเคราะห์โจทย์ประเภท Prediction/Regression, Classification, Association, Clustering โดยใช้ Explorer Interface
6. เปรียบเทียบการวิเคราะห์โจทย์ประเภท Classification ด้วย data set หลายชุด และใช้หลาย algorithm โดยใช้ Experimenter Interface

การใช้โปรแกรม R ในการวิเคราะห์ข้อมูลทางสถิติ

สุณี รักษาเกียรติศักดิ์

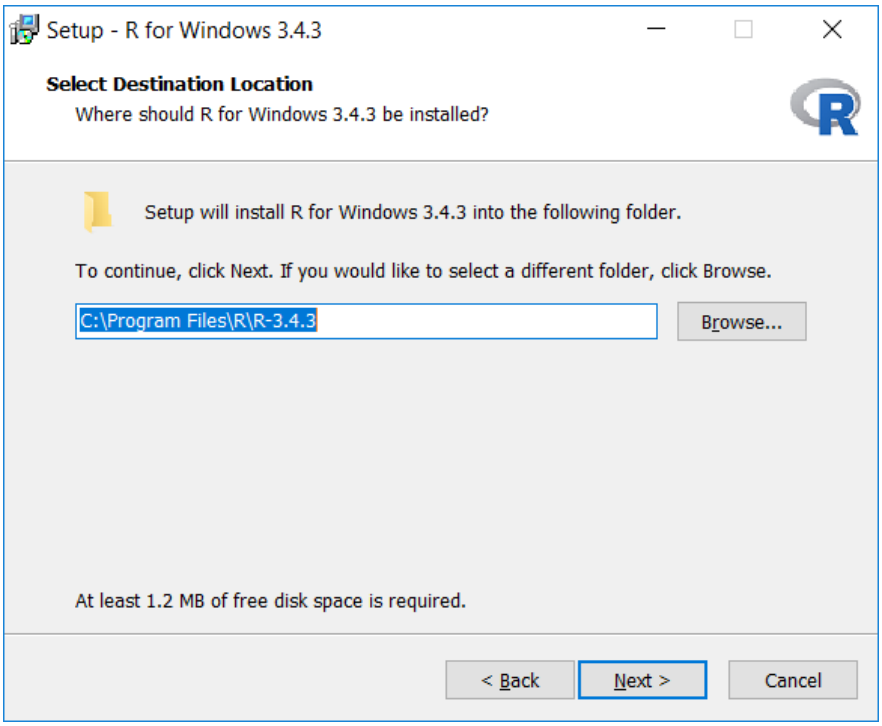
การ Install โปรแกรม R และ Package RcmdR

1. ไป download โปรแกรมที่ R Project http://www.r-project.org/
   เลือก **Download R** แล้วเลือก CRAN Mirrors เป็น **Thailand**
2. เลือก **Download R for Windows** แล้วเลือก **base** จะได้ดังรูป



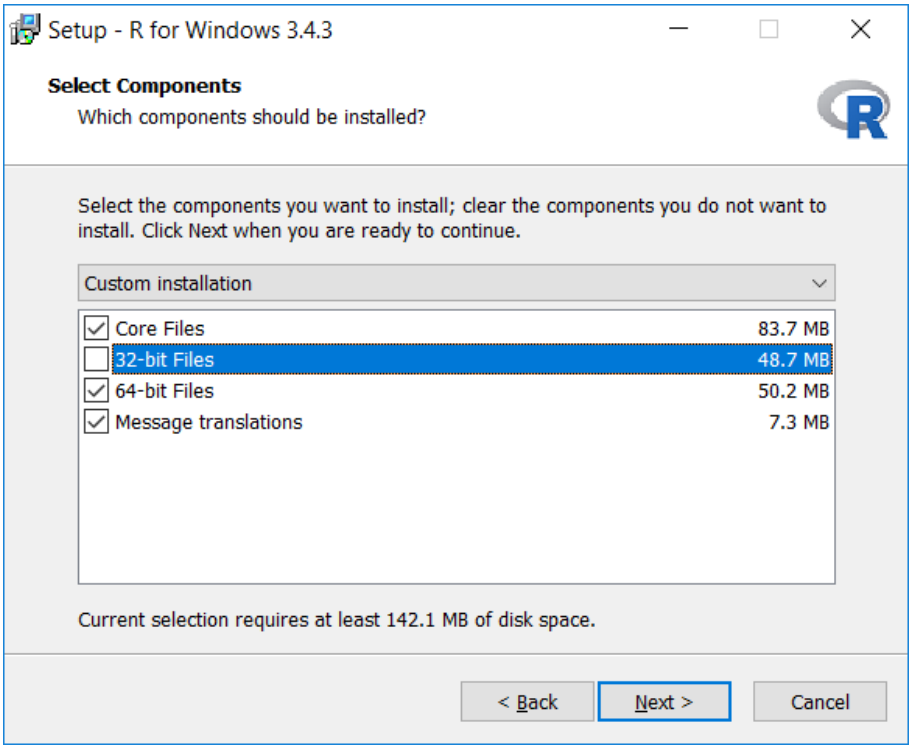
3. เลือก **Download R 3.4.3 for Windows** ให้ save Application ที่ download (R-3.4.3-win.exe)

4. ทำการ run โปรแกรม (install) เลือกโฟลเดอร์ตาม default

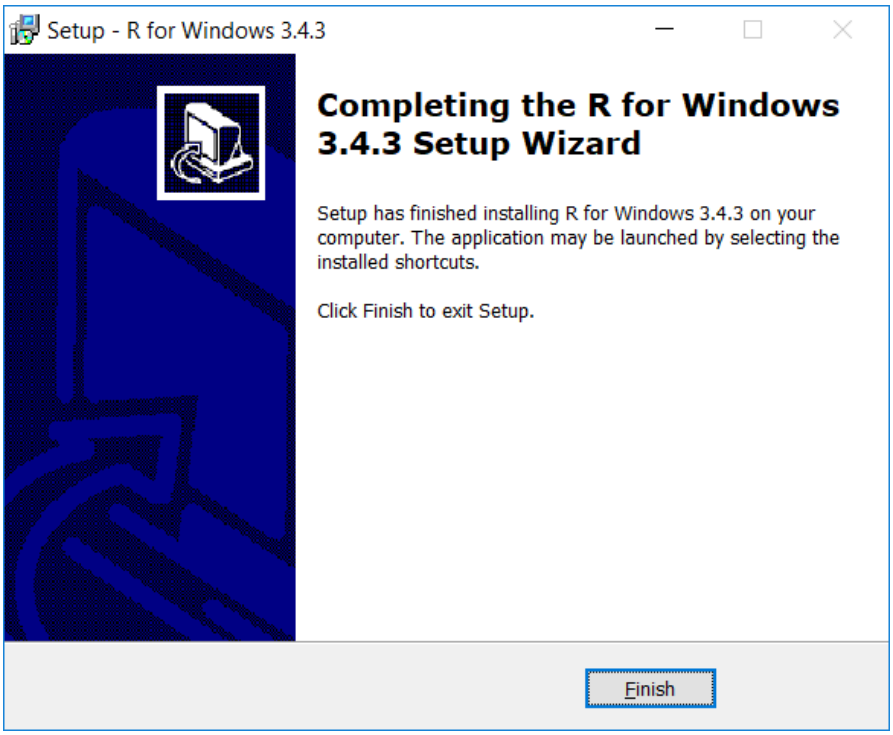


5. Select Components

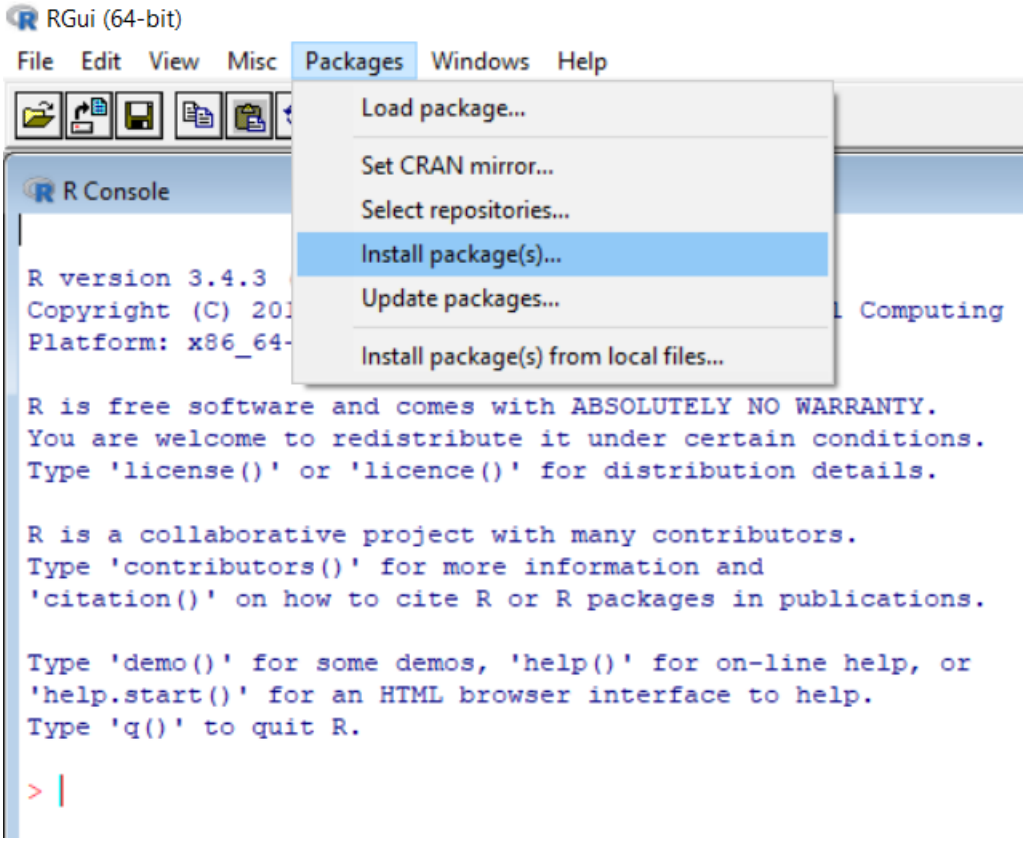
ให้เอา 32-bit Files ออกสำหรับเครื่อง 64 bits



6. เลือกตาม default และ Next จนระบบติดตั้งเสร็จ

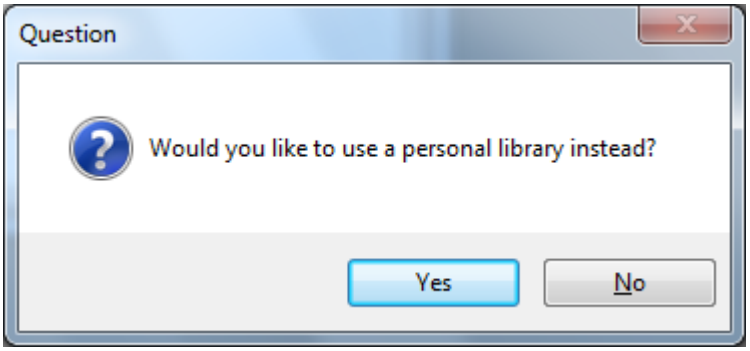


7. เมื่อ Install เรียบร้อยแล้วจะปรากฏไอคอนของโปรแกรม R บน Desktop
8. ให้ double click โปรแกรมเพื่อใช้งาน
   โดยต้อง Install package เพิ่มสำหรับเมนูการวิเคราะห์ข้อมูลทางสถิติ
   โดยไปที่ Packages **> Install package(s)**

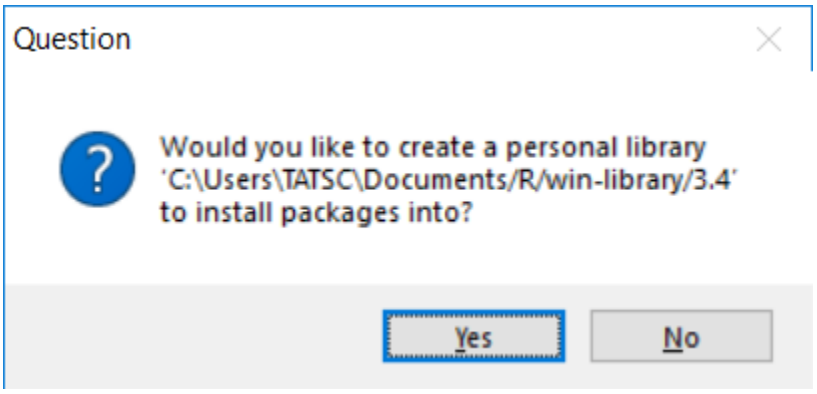


9. เลือก CRAN mirror
10. เลือก Packages เป็น **Rcmdr**

ระบบจะถาม

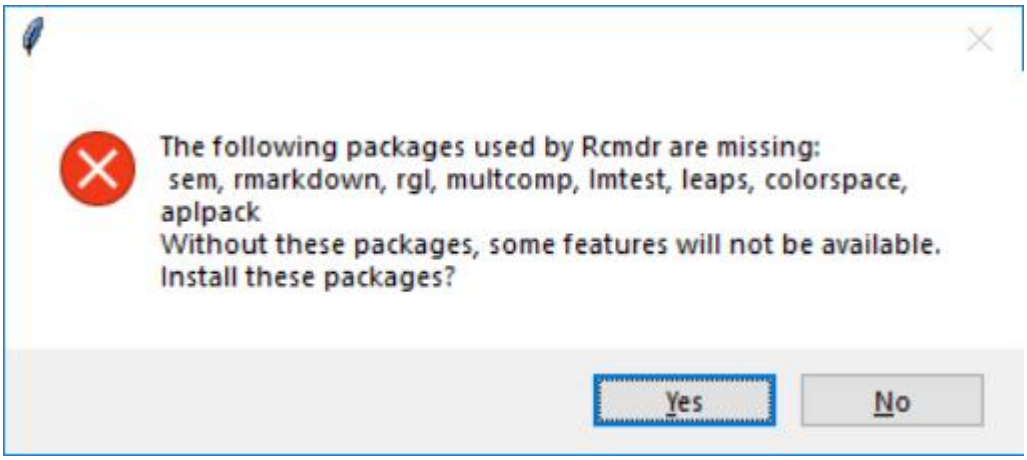


ให้ตอบ Yes

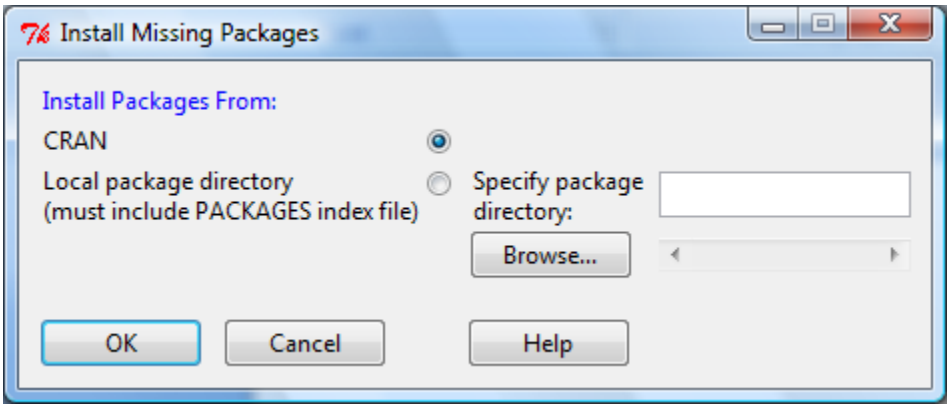


ระบบจะทำการ download packages ต่าง ๆ มาเก็บไว้

11. เลือก Packages > Load Packages จะขึ้นดังรูป

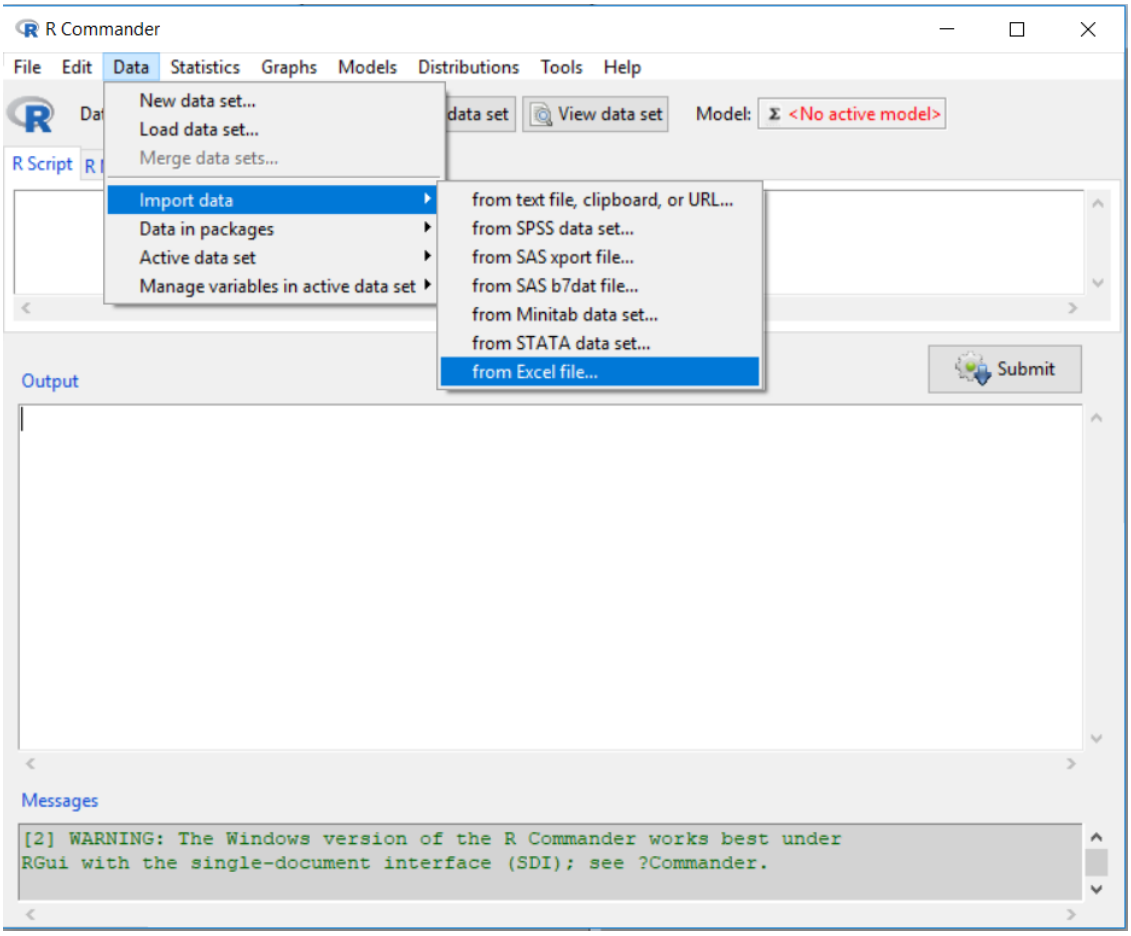


12. ให้กด Yes แล้วกด OK คือให้ Install จาก CRAN



13. ช่วงนี้จะใช้เวลาสักพัก รอจน download เสร็จ (ระบบจะ download ทุก packages ที่จำเป็น)

14. เมื่อเรียบร้อยจะขึ้นหน้าต่าง R Commander สามารถทดสอบโดยอ่านข้อมูลจาก Excel ได้ โดยไปที่เมนู Data > Import data > from Excel



15. เมื่อเลิกใช้งานให้ปิด R Commander และ RGui

16. เมื่อต้องการใช้งานใหม่ เมื่อเปิด RGui ขึ้นมาต้อง Load Package Rcmdr มาใช้งานเสมอ

<u>หมายเหตุ</u> เนื่องจากการติดตั้งต้องใช้ Internet ดังนั้นจึงได้เตรียมโปรแกรม R ที่ install เรียบร้อยแล้วใน Virtual Machine สำหรับภาคปฏิบัติ

การใช้เครื่องมือออนไลน์ Google WORK (FORMS & SHEETS) ในการเก็บรวบรวมข้อมูล

Google SHEETS

เครื่องมือในการเก็บรวบรวมข้อมูล ในแบบเดิมมักจะใช้แบบสอบถามที่เป็นกระดาษ เช่น ดังตัวอย่าง **แบบรายงานการฝึกงานของนิสิต ชั้นปีที่ 3 ....** ดังแสดงในไฟล์ ***Report_Internship*** จากนั้นนำข้อมูลมาคีย์ข้อมูลลงใน Excel หรือ Goggle SHEETS ได้ดังแสดงในไฟล์ชื่อ ***Report_Internship***

Google FORMS

การเก็บรวบรวมข้อมูลออนไลน์ด้วย Google FORMS กำลังเป็นที่นิยมอย่างมากเนื่องด้วยการสร้างฟอร์ม (FORM) ทำได้ง่าย และสามารถจะเก็บข้อมูลแบบออนไลน์ได้ทันที ทำให้ประหยัดค่าใช้จ่าย และ Google FORM ยังประมวลผลข้อมูลอย่างง่ายให้ในระดับตัวแปรเดียว เช่น จำนวนและร้อยละ เป็นต้น หากผู้วิจัยต้องการวิเคราะห์ข้อมูลเพิ่มเติมด้วยโปรแกรมสำเร็จรูปทางสถิติอื่น ๆ เช่น โปรแกรม SPSS หรือโปรแกรม R ก็สามารถ download ข้อมูลจาก Form_responses มาใน Excel แล้วดำเนินการต่อตามต้องการได้

จากโจทย์ “การฝึกงานของนิสิต” ผู้วิจัยได้สร้างเครื่องมือในการเก็บข้อมูลเพิ่มเติมจากการฝึกงานของนิสิต โดยใช้ Google FORMS ดังแสดงในไฟล์ ***Form*** และข้อมูลจากการตอบแบบสอบถาม ดังแสดงในไฟล์ ***Form_Responses*** โดย Google จะประมวลผลข้อมูลอย่างง่ายให้ดังแสดงในไฟล์ ***Summary_Responses***

การจัดเตรียมข้อมูลสำหรับการวิเคราะห์

ผู้วิจัยได้จัดเตรียมข้อมูลสำหรับการวิเคราะห์โดยนำตัวแปรที่สนใจจากไฟล์ ***Report_Internship*** และจากไฟล์ ***Form_Responses*** โดยทำการ join ด้วย ID ซึ่งก็คือหลักการทำ ETL นั้นเอง แต่เนื่องจากข้อมูลมีจำนวนน้อยและผู้วิจัยรู้จักข้อมูลดี จึงทำด้วยมือ โครงสร้างข้อมูลที่ได้ดังแสดงในตารางที่ 1 และข้อมูลดังแสดงในไฟล์ ***DataAll***

ตารางที่ 1 โครงสร้างข้อมูลของ “การฝึกงานของนิสิต”

ชื่อตัวแปร	คำอธิบายตัวแปร	ค่าของตัวแปร (Domain)	ระดับการวัด (**Measure**)
ID	รหัสนิสิต	รหัส 4 หลักสุดท้าย	-
GENDER	เพศ	ชาย, หญิง	
GPA	เกรดเฉลี่ย	1.00-4.00	
ORG_SIZE	ขนาดขององค์กรที่ฝึกงาน	ขนาดเล็ก (น้อยกว่า 15 คน) ขนาดกลาง (15 - 50 คน) ขนาดใหญ่ (มากกว่า 50 คน)	
PAY	ค่าเบี้ยเลี้ยง	ไม่ได้รับเบี้ยเลี้ยง น้อยกว่า 100 100 - 200 201 - 300 301 - 400 มากกว่า 400	
DAYS	จำนวนวันที่ทำงาน	20 – 60	
HOURS	จำนวนชั่วโมงที่ทำงาน	130 - 500	
CP121, …, CP444	ความรู้ที่ใช้ในวิชาที่เรียนมา	1 – 5 (น้อย – มาก)	
WSK1, …, WSK6	ความรู้และทักษะในการทำงาน	4=ดีมาก, 3=ดี, 2=ปานกลาง, 1=น้อย, 0=น้อยมาก	
USK1, …, USK5	อัตลักษณ์นิสิตของมหาวิทยาลัย “มีทักษะการสื่อสาร”	4=ดีมาก, 3=ดี, 2=ปานกลาง, 1=น้อย, 0=น้อยมาก	
Letter	ใบรับรองการฝึกงาน	0=ไม่มี, 1=มี	
Stamp	ใบประทับตราบริษัท	0=ไม่มี, 1=มี	

ระดับการวัดของข้อมูล (Level of Measurement)

ข้อมูลหรือตัวแปรหรือแอททริบิวต์ (Attribute) ที่จะวิเคราะห์เพื่อหาค่าสถิติของข้อมูลในเบื้องต้นนั้น เราต้องทราบว่าข้อมูลของนั้นมีระดับการวัด (measure) เป็นอย่างไร

ระดับการวัดแบ่งเป็น 2 ระดับใหญ่ ๆ คือ

1. ข้อมูลเชิงคุณลักษณะหรือข้อมูลเชิงคุณภาพ (Qualitative data)
   แบ่งออกเป็น 2 ระดับคือ
   - Nominal เป็นข้อมูลเชิงกลุ่มหรือนามบัญญัติ เช่น เพศ หรือ Gender มี 2 กลุ่ม/ค่า คือ Male และ Female หรือ ชาย และ หญิง เป็นต้น
   - Ordinal เป็นข้อมูลเชิงกลุ่มแบบมีอันดับ เช่น วุฒิการศึกษา หรือ EDU มี 3 กลุ่ม/ค่า คือ ต่ำกว่าปริญญาตรี ปริญญาตรี สูงกว่าปริญญาตรี เป็นต้น ซึ่งการศึกษามีอันดับจากต่ำไปสูง (มีปริมาณ หรือ magnitude แต่ไม่มี scale หรือหน่วยวัด)
2. ข้อมูลเชิงปริมาณ (Quantitative data) คือข้อมูลที่มีค่าต่อเนื่อง (จำนวนจริง แต่ในทางปฏิบัติอาจจะบันทึกเป็นจำนวนเต็ม) มีหน่วยวัด (scale) แบ่งเป็น 2 ระดับคือ
   - Interval เป็นข้อมูลแบบช่วง มีหน่วยวัด เช่น คะแนน เป็นข้อมูลที่ไม่มีศูนย์สัมบูรณ์ มักเป็นเครื่องมือวัดทางจิตวิทยา (psychological measurement) เครื่องมือสอบถามความคิดเห็น ข้อสอบวัดต่าง ๆ หรือเครื่องมือวัด KPI ต่าง ๆ
   - Ratio เป็นข้อมูลแบบอัตราส่วน มีหน่วยวัด และมักเป็นเครื่องมือวัดทางกายภาพ (physical measurement) ที่เป็นสากล มีศูนย์สัมบูรณ์หรือศูนย์แท้ ทำให้สามารถเปรียบเทียบค่าของข้อมูลแบบอัตราส่วนได้ เช่น อายุ (ปี) น้ำหนัก (กก.) ส่วนสูง (ซม.) เป็นต้น

หมายเหตุ ในการวิเคราะห์ข้อมูลทางสถิติ สถิติที่เลือกใช้จะขึ้นอยู่กับลักษณะของข้อมูล 3 ระดับเท่านั้น คือ Nominal, Ordinal, Scale (Interval/Ratio)

Measure ในโปรแกรม SPSS, Excel, R, Weka เป็นดังนี้

	nominal	ordinal	Scale (interval/ratio)
SPSS	nominal	ordinal	scale
Excel	text	text	number
R	character, factor	character, factor	numeric
Weka	nominal	nominal	numeric

Workshop 1: จงใส่ระดับการวัดของข้อมูล (Measure: nominal, ordinal, scale) ลงในตารางที่ 1

สถิติและพารามิเตอร์

สถิติ (statistics) เป็นตัวเลขเป็นค่าที่คำนวณได้จากกลุ่มตัวอย่าง เป็นคุณลักษณะของกลุ่มตัวอย่าง (sample)

พารามิเตอร์ (parameter) เป็นคุณลักษณะของประชากร (population) ซึ่งเราไม่ทราบค่า

สถิติ เป็น**ตัวประมาณค่า (estimator)** ของพารามิเตอร์

Workshop 2: จงเขียนสัญลักษณ์แทนสถิติและพารามิเตอร์ต่อไปนี้ และแสดงสูตรการคำนวณของค่าสถิติ (ตัวอย่างสัญลักษณ์: $\overline{X}$, SD, r, μ, เป็นต้น)

	สถิติ (สัญลักษณ์และสูตรการคำนวณ)	พารามิเตอร์ (สัญลักษณ์)
ค่าเฉลี่ยหรือ Mean		
ค่าเบี่ยงเบนมาตรฐาน หรือ Standard deviation		
ค่าสัมประสิทธิ์สหสัมพันธ์ หรือ Correlation		

การหาค่าสถิติสำหรับตัวแปรเดียว

ตัวแปรเชิงคุณลักษณะ: ค่าความถี่ (จำนวนและร้อยละ) หรือ Frequency & percentage

กราฟ: กราฟแท่ง (bar chart) หรือกราฟพาย (pie chart)

ตัวแปรเชิงปริมาณ: ค่ากลาง และค่าการกระจาย หรือค่าเปอร์เซ็นไทล์ต่าง ๆ ได้แก่

ค่าเฉลี่ย (Mean) และค่าเบี่ยงเบนมาตรฐาน (Standard deviation)

ค่ามัธยฐาน (Median) และค่า Interquartile Range (Q3 – Q1)

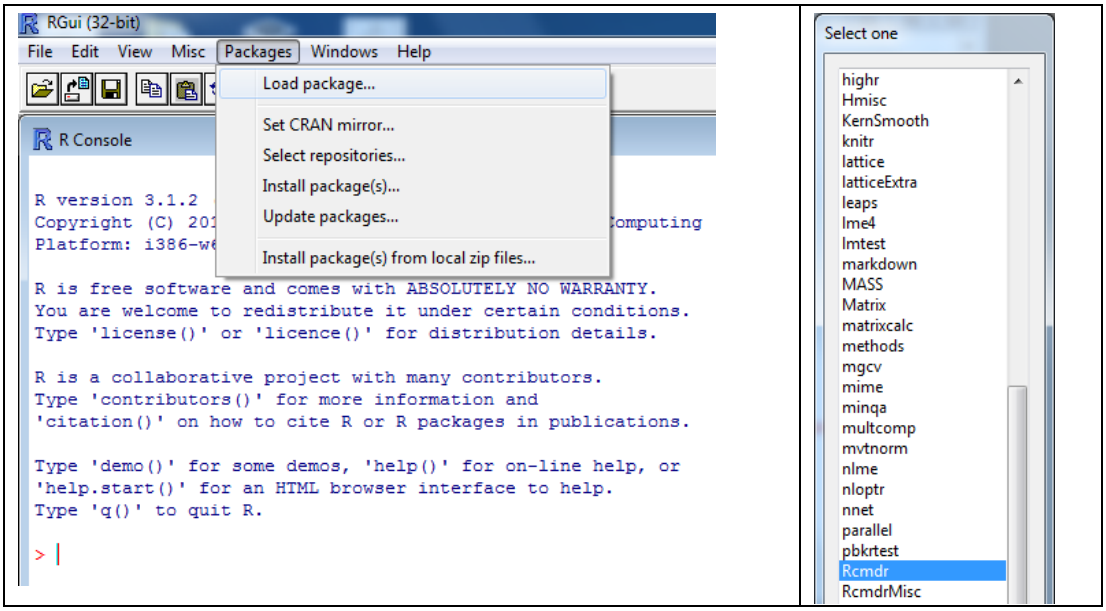
ค่าเปอร์เซ็นไทล์ (Percentile) ต่าง ๆ เช่น P10, P25=Q1, P50=Q2=Median, P75, P90 เป็นต้น

กราฟ: ฮีสโตรแกรม (histogram) แสดงการแจกแจง (distribution) ของข้อมูล หรือ box plot (แสดงค่า Q1, Q2=Median, Q3)

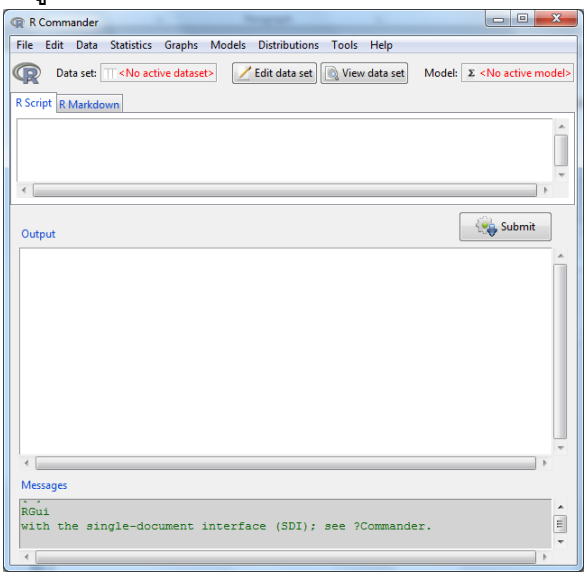
Workshop 3: ใช้โปรแกรม R หาค่าสถิติเบื้องต้นของโจทย์ “การฝึกงานของนิสิต” ข้อมูลอยู่ในไฟล์ DataAll และสร้างตัวแปรใหม่

1. เปิดโปรแกรม R โดยดับเบิลคลิก  บน desktop มาทำงาน

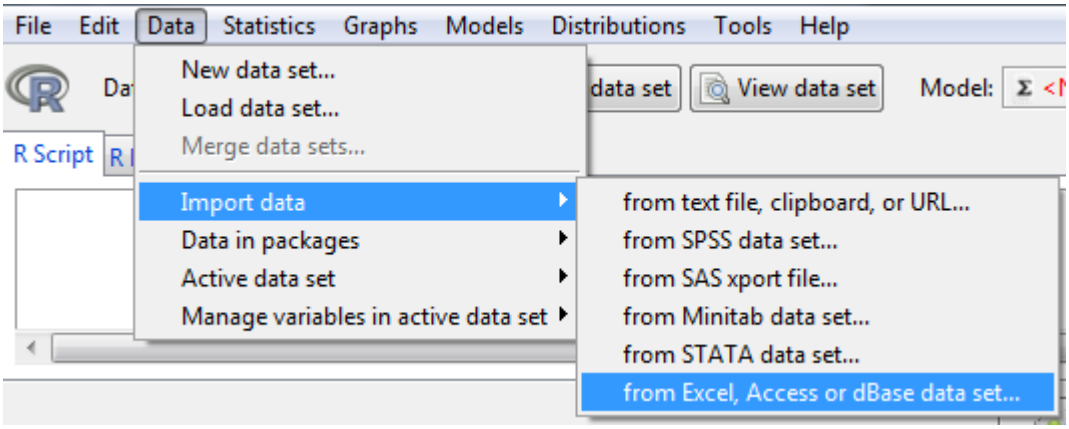
2. เลือก **Packages > Load package** (รูปซ้าย) แล้วเลือก **Rcmdr** (รูปขวา)



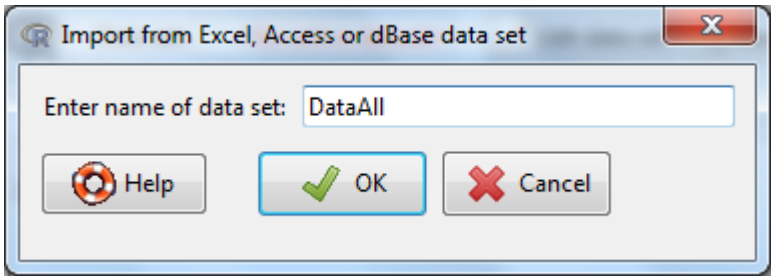
จะได้ R Commander ซึ่งเป็นเมนูที่ผูกฟังก์ชันการทำงานของโปรแกรมคำนวณค่าสถิติต่าง ๆ (เหมือนเมนูของ SPSS) ทำให้ง่ายต่อการใช้งาน



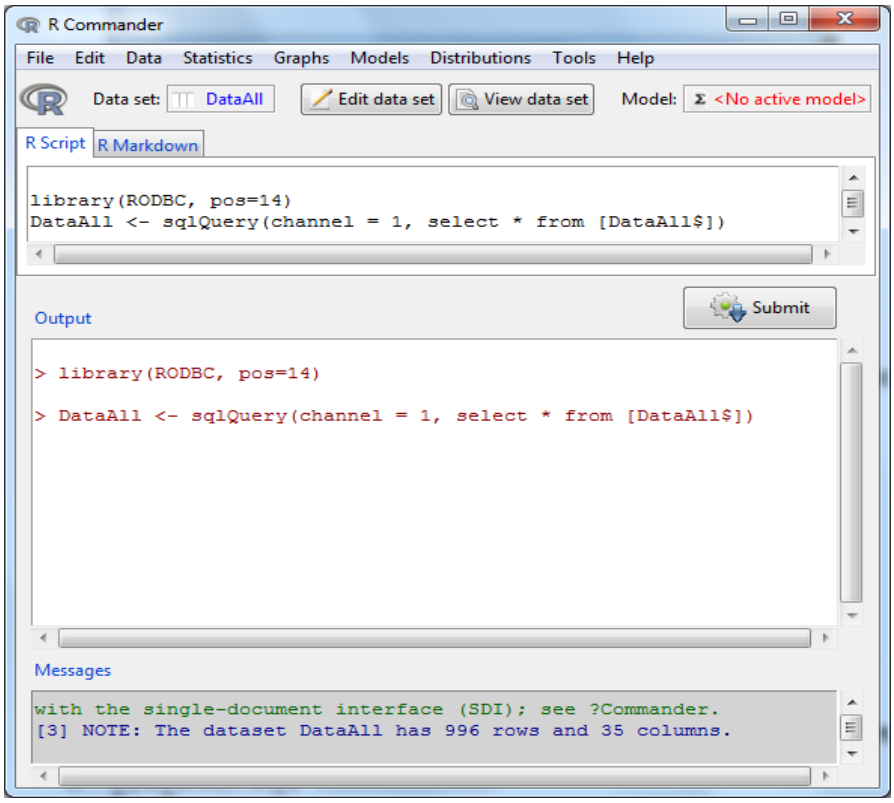
3. อ่านข้อมูล Excel ด้วยคำสั่ง Data > **Import data > from Excel**, ..



ตั้งชื่อ dataset เป็น **DataAll** แล้วกด OK



เลือกไฟล์ DataAll.xlsx จาก CD (เลือกไฟล์ type เป็น MS Excel 2007 file (.xlsx) แล้วเลือก Sheet DataAll จะได้



หน้าจอการทำงานของโปรแกรม R จะแบ่งเป็น 3 ส่วนคือ

- R Script เป็นคำสั่ง R จากเมนูที่เลือก
- Output แสดงผลลัพธ์การประมวลผล
- Messages แสดงข้อความสำหรับการทำงานนั้น (อ่านข้อมูลเข้า 996 rows 35 columns) ถ้าเกิด error ข้อความก็จะเป็นสีแดง

R Markdown **R Markdown** is an authoring format that enables easy creation of dynamic documents, presentations, and reports from **R**. It combines the core syntax of **markdown** (an easy-to-write plain text format) with embedded **R** code chunks that are run so their output can be included in the final document. R Markdown — Dynamic Documents for R **rmarkdown**.rstudio.com/	หมายเหตุ: R Markdown เป็น feature ใหม่ใน version นี้ (ยังไม่ได้ทดสอบการใช้งาน) ไม่เกี่ยวกับการประมวลผลทางสถิติ

4. ดูข้อมูลโดยกดปุ่ม View data set
   ถ้าไม่มีข้อมูล (missing value) R จะแทนด้วยอะไร? .....
5. **Clear window** (ทำเป็นครั้งคราวตามต้องการ) โดยคลิกขวาที่ window ที่ต้องการ clear เลือก Clear window

6. หาค่าสถิติเบื้องต้น (descriptive statistics) ของแต่ละตัวแปรเพื่อสรุปลักษณะของข้อมูล

   Qualitative variables: **GENDER, ORG_SIZE, PAY, Letter, Stamp**

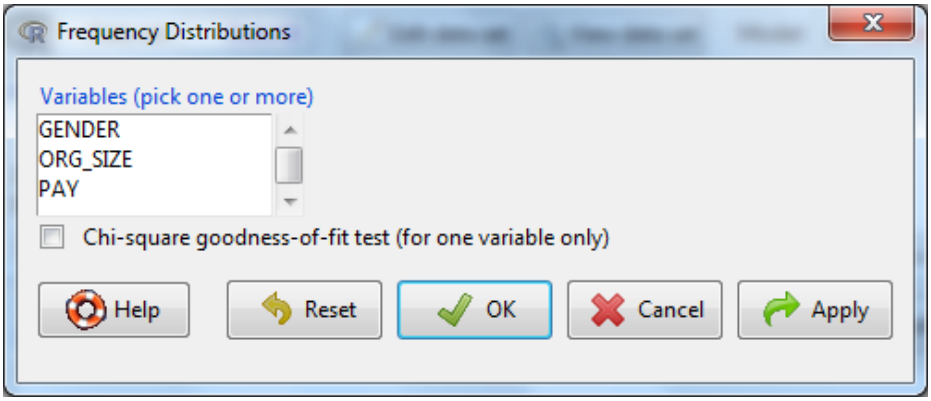
   ด้วยการหาการแจงแจงความถี่ (จำนวนและร้อยละ) ด้วยคำสั่ง **Frequency distributions** และแสดงกราฟวงกลม ด้วยคำสั่ง **Pie Chart**

   Quantitative variables: **GPA, DAYS, HOURS, CP121, , CP444, WSK1, , WSK6, USK1, USK6**

   ด้วยการหาค่ากลางและค่าการกระจาย ด้วยคำสั่ง **Numerical summaries**

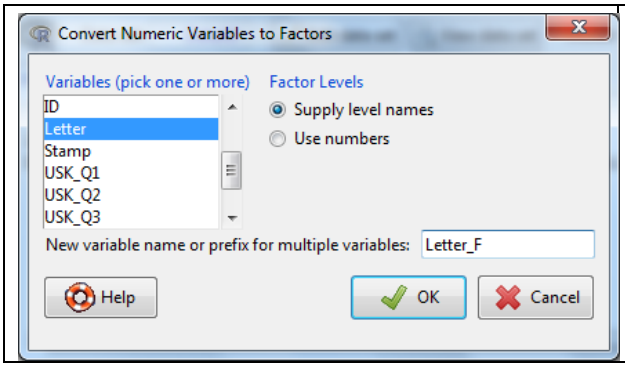
สำหรับ Qualitative variables:

6.1 เลือกคำสั่ง Statistics > Summaries > **Frequency distributions** จะได้



จะเห็นได้ว่ามีแค่ 3 ตัวแปร ที่เป็น factor (ใน R ตัวแปร character จะเป็น factor โดยอัตโนมัติ) ไม่มีตัวแปร Letter และ Stamp เนื่องด้วยค่าของตัวแปรเป็นตัวเลข (จาก Excel) R เลยให้ชนิดเป็น numeric

6.2 แปลงตัวแปร Letter, Stamp ให้เป็น factor ด้วยคำสั่ง Data > Manages variables in active data set > **Convert numeric variables to factors**

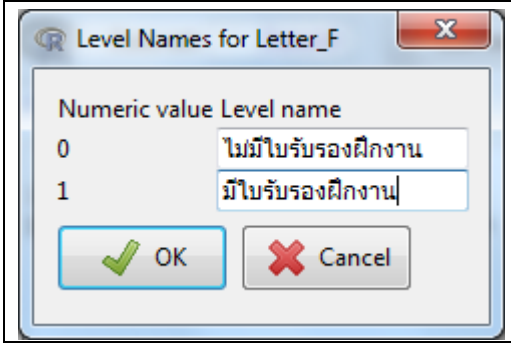


Variables: เลือก Letter

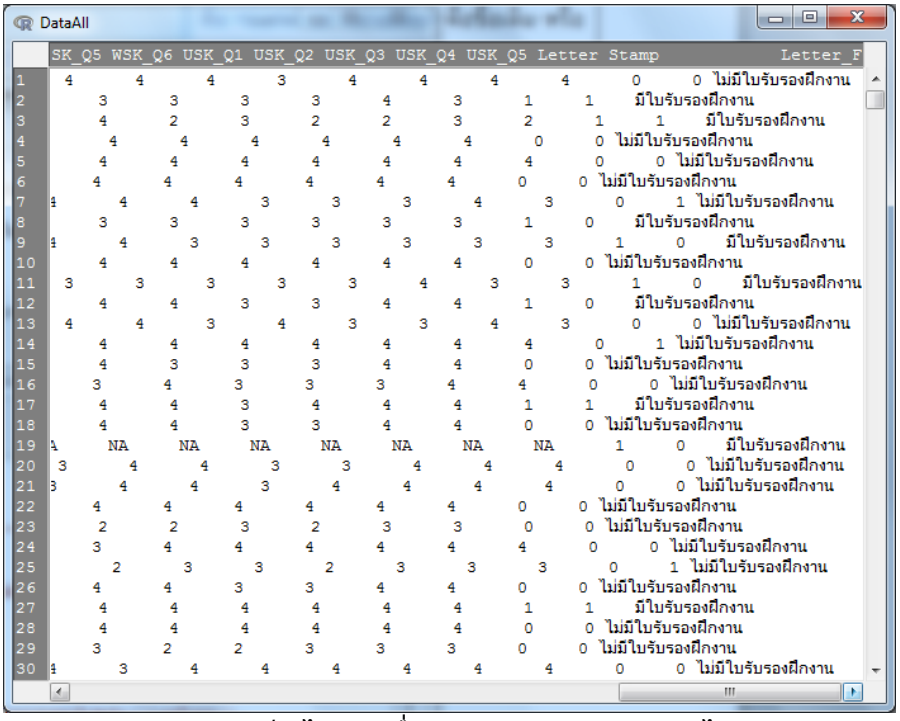
Factor Levels: เลือก Supply level names

New variable name .....: เลือก default คือ <same as variables> คือชื่อเดิม หรือจะตั้งชื่อใหม่ ตั้งใหม่เป็น Letter_F

กด OK

	ใส่ชื่อค่าของข้อมูล (Numeric value Level name) ดังภาพ แล้วกด OK

ลอง View data set ดู จะได้



DataAll

	SK_Q5	WSK_Q6	USK_Q1	USK_Q2	USK_Q3	USK_Q4	USK_Q5	Letter	Stamp	Letter_F
1	4	4	4	3	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน
2	3	3	3	3	4	3	1	1		มีใบรับรองฝึกงาน
3	4	2	3	2	2	3	2	1	1	มีใบรับรองฝึกงาน
4	4	4	4	4	4	4	0	0		ไม่มีใบรับรองฝึกงาน
5	4	4	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน
6	4	4	4	4	4	4	0	0		ไม่มีใบรับรองฝึกงาน
7	4	4	4	3	3	3	4	3	0	1 ไม่มีใบรับรองฝึกงาน
8	3	3	3	3	3	3	3	1	0	มีใบรับรองฝึกงาน
9	4	4	3	3	3	3	3	3	1	0 มีใบรับรองฝึกงาน
10	4	4	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน
11	3	3	3	3	3	4	3	3	1	0 มีใบรับรองฝึกงาน
12	4	4	3	3	4	4	4	1	0	มีใบรับรองฝึกงาน
13	4	4	3	4	3	3	4	3	0	0 ไม่มีใบรับรองฝึกงาน
14	4	4	4	4	4	4	4	0	1	ไม่มีใบรับรองฝึกงาน
15	4	3	3	3	4	4	0	0		ไม่มีใบรับรองฝึกงาน
16	3	4	3	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน
17	4	4	3	4	4	4	1	1		มีใบรับรองฝึกงาน
18	4	4	3	3	4	4	0	0		ไม่มีใบรับรองฝึกงาน
19	A	NA	NA	NA	NA	NA	NA	NA	1	0 มีใบรับรองฝึกงาน
20	3	4	4	3	3	4	4	4	0	0 ไม่มีใบรับรองฝึกงาน
21	3	4	4	3	4	4	4	4	0	0 ไม่มีใบรับรองฝึกงาน
22	4	4	4	4	4	4	0	0		ไม่มีใบรับรองฝึกงาน
23	2	2	3	2	3	3	0	0		ไม่มีใบรับรองฝึกงาน
24	3	4	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน
25	2	3	3	2	3	3	3	0	1	ไม่มีใบรับรองฝึกงาน
26	4	4	3	3	4	4	0	0		ไม่มีใบรับรองฝึกงาน
27	4	4	4	4	4	4	1	1		มีใบรับรองฝึกงาน
28	4	4	4	4	4	4	0	0		ไม่มีใบรับรองฝึกงาน
29	3	2	2	3	3	3	0	0		ไม่มีใบรับรองฝึกงาน
30	4	3	4	4	4	4	4	4	0	0 ไม่มีใบรับรองฝึกงาน

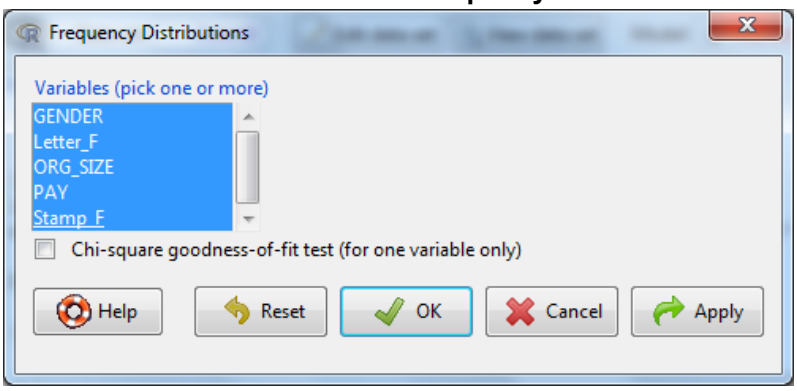
ข้อมูลของแต่ละคอลัมน์จะไม่ตรงเนื่องจากการแสดงผลภาษาไทย

แปลงตัวแปร Stamp ให้เป็น factor โดยใช้ **Use Number** สำหรับ Factor Levels และแทนที่ตัวแปรเดิม (<same as variables>) เพื่อความรวดเร็ว หรือจะตั้งชื่อตัวแปรใหม่เป็น Stamp_F

หมายเหตุ ถ้าเราต้องการใช้ตัวแปรเดิมที่เป็น numeric เราต้องสร้างตัวแปรใหม่ให้เป็น factor จะได้มีตัวแปร 2 ตัว เลือกใช้ได้ตามความเหมาะสม

หมายเหตุ ใน R มี data type ได้หลายแบบ ได้แก่ scalars, vectors (numerical, character, logical), matrices, data frames, and lists. ซึ่ง data frame ก็คือชนิดข้อมูลที่เป็นตาราง (table หรือ relation) นั่นเอง คือแต่ละคอลัมน์จะเป็นตัวแปร ซึ่งจะมีชนิดเป็น numeric หรือ character (ซึ่งก็คือ string) ถ้าข้อมูลเป็น character R จะถือเป็นnominal อัตโนมัติแล้วให้ type เป็น **factor** ส่วนตัวแปรที่เป็น numeric ถ้าจะให้เป็น factor ก็ต้องดำเนินการ ดังตัวอย่างข้างต้น

6.3 เลือกคำสั่ง Statistics > Summaries > **Frequency distributions** ใหม่ เลือกตัวแปรทั้งหมด (5 ตัว)

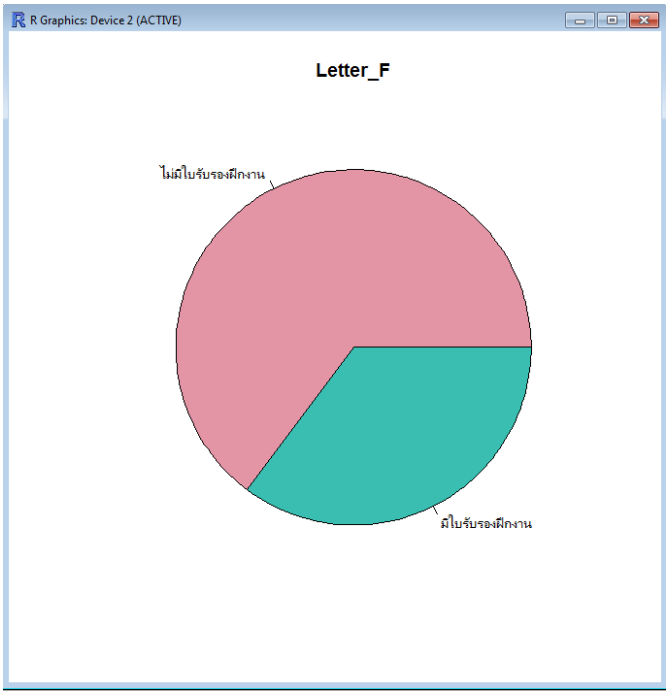


คลิก OK

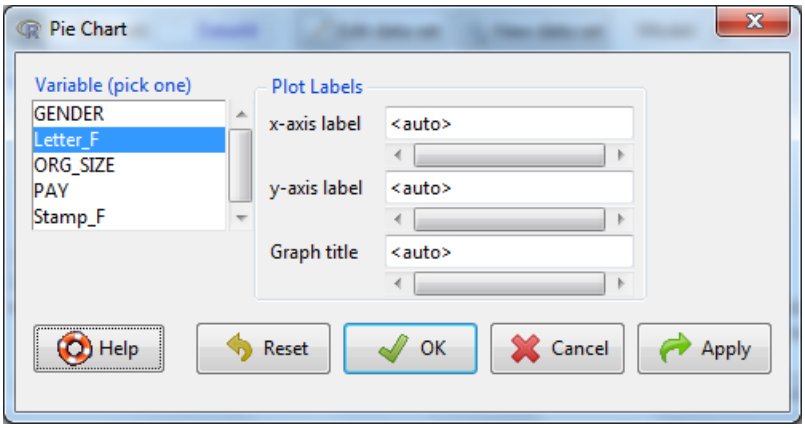
จะได้

R Scrip	**Output**
local({ .Table <- with(DataAll, table(Stamp_F)) cat("\ncounts:\n") print(.Table) cat("\npercentages:\n") print(round(100*.Table/sum(.Table), 2))	counts: Stamp_F 0 1 58 12 percentages: Stamp_F 0 1 82.86 17.14
local({ .Table <- with(DataAll, table(GENDER)) cat("\ncounts:\n") print(.Table) cat("\npercentages:\n") print(round(100*.Table/sum(.Table), 2)) })	counts: GENDER ชาย หญิง 40 32 percentages: GENDER ชาย หญิง 55.56 44.44
หมายเหตุ หากข้อมูลมีมากกว่า 3 ค่า การแสดง value label ที่เป็นภาษาไทยจะเรียงตามตัวอักษร ถ้าข้อมูลเป็น Ordinal จะสับสน เพราะมันจะไม่เรียงอันดับให้ หากต้องการให้เรียงอันดับก็ควรใส่ 1_, 2_, 3_, ... ไว้ข้างหน้า	

6.4 แสดงกราฟพายด้วยคำสั่ง Graphs > **Pie Chart** ของตัวแปร Letter_F ตาม default ได้



ให้ระวังและสังเกตว่าผลลัพธ์ของกราฟจะแสดงอยู่ใน RGUI ดังนั้นต้องยุบหน้าต่าง R Commnader ก่อน ลองเรียก Pie Chart อีกครั้ง



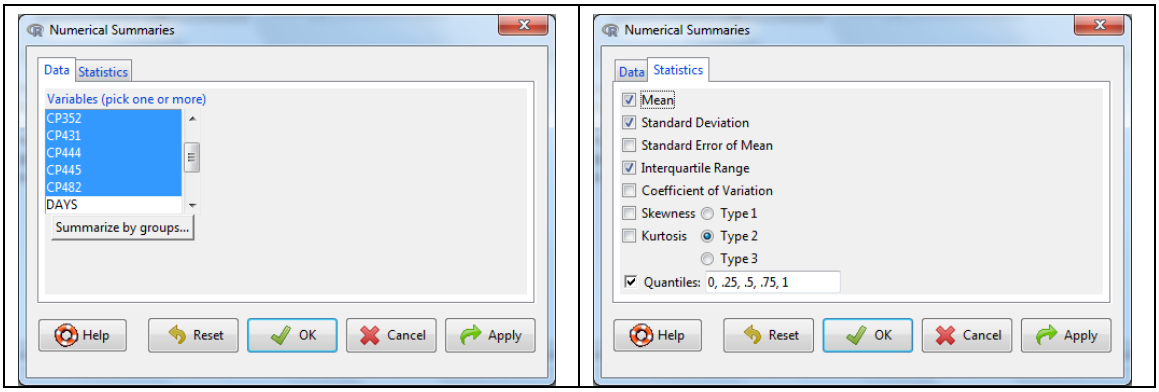
คลิกที่ 

เพื่ออ่านเอกสารว่าจะมีการปรับแต่งอะไรได้บ้าง

สำหรับ Quantitative variables:

6.5 เลือกคำสั่ง Statistics > Summaries > **Numerical summaries** แล้วเลือกกลุ่มวิชา CP…

ตาม default จะให้ค่า Statistics เป็น Mean, Standard Deviation, Quantiles: 0, .25, .5, .75, 1



R Scrip:

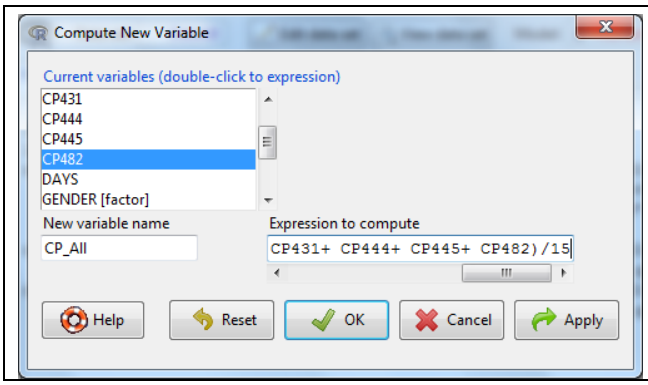
```
numSummary(DataAll[,c("CP111", "CP121", "CP201", "CP212", "CP241", "CP251",
"CP316", "CP322", "CP323", "CP342", "CP352", "CP431", "CP444", "CP445",   "CP482")],
statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

Output:

```
          mean        sd  IQR 0%  25% 50%  75% 100%  n
CP111 2.666667 1.3737991 3.00  1 1.00   3 4.00    5 72
CP121 2.041667 1.2830147 2.00  1 1.00   1 3.00    5 72
CP201 1.944444 0.9913554 1.25  1 1.00   2 2.25    5 72
CP212 2.888889 1.4394400 2.00  1 2.00   3 4.00    5 72
CP241 2.805556 1.2292089 2.00  1 2.00   3 4.00    5 72
CP251 3.750000 1.3188770 2.00  1 3.00   4 5.00    5 72
CP316 2.638889 1.3871194 3.00  1 1.00   3 4.00    5 72
CP322 2.055556 1.0991532 2.00  1 1.00   2 3.00    5 72
CP323 2.291667 1.1800871 2.00  1 1.00   2 3.00    5 72
CP342 3.388889 1.3692350 2.00  1 2.00   4 4.00    5 72
CP352 2.625000 1.2608682 2.00  1 2.00   3 4.00    5 72
CP431 2.777778 1.3862730 2.25  1 1.75   3 4.00    5 72
CP444 1.972222 1.1864520 2.00  1 1.00   1 3.00    5 72
CP445 2.083333 1.2189813 2.00  1 1.00   2 3.00    5 72
CP482 2.305556 1.1705168 2.00  1 1.00   2 3.00    5 72
```

7. สร้างตัวแปรตัวใหม่เพื่อวัด “ระดับความรู้ที่ใช้” ตั้งชื่อว่า **CP_All** โดยเป็นค่าเฉลี่ยของ CP ทั้งหมด

   เลือกคำสั่ง Data > Manage variables in active data set > **Compute new variable**



Expression to compute =

(CP111+ CP121+ CP201+ CP212+ CP241+ CP251+ CP316+ CP322+ CP323+ CP342+ CP352+ CP431+ CP444+ CP445+ CP482)/15

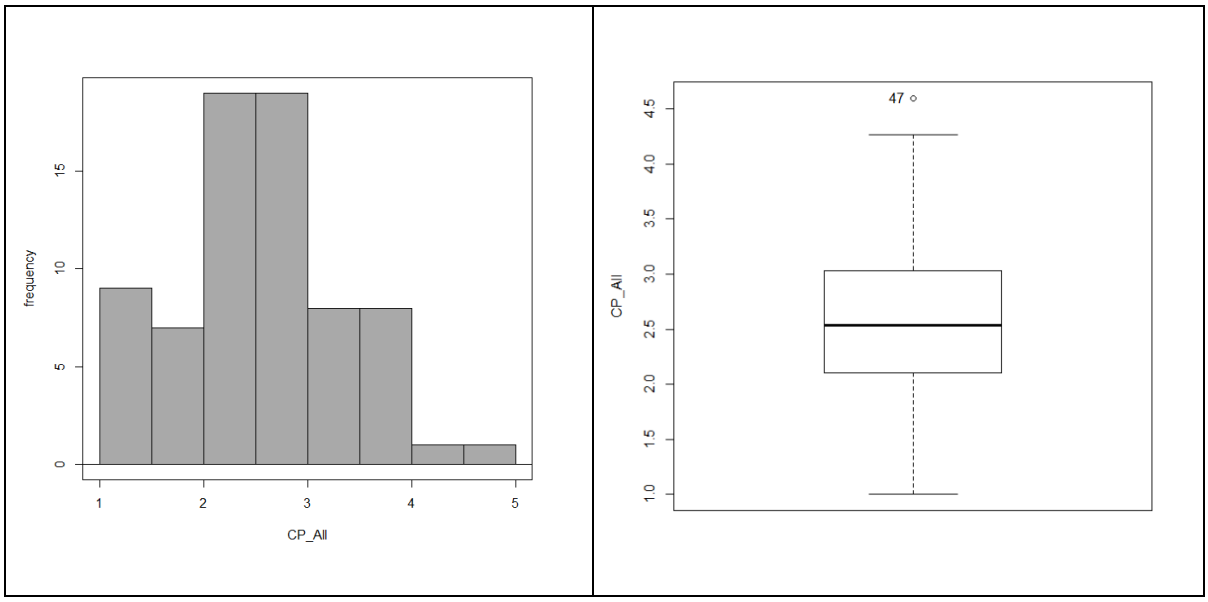
(ให้ double click ตัวแปรที่ต้องการ)

   ลอง **View Data Set** เพื่อดูค่าตัวแปรตัวใหม่ ต้องมีค่าอยู่ระหว่าง 1 – 5

8. ลอง plot กราฟ ของตัวแปรเชิงปริมาณ ด้วย histogram และ box plot

   คำสั่ง: Graphs > **Histogram** (รูปซ้าย

   หรือ Graphs > **Boxplot** (รูปขวา)



   และหากจะสำเนากราฟก็คลิกขวาที่รูปแล้วเลือก **Copy as bitmap**

9. หาค่าสถิติเบื้องต้นของ CP_All คำสั่งที่ใช้ .....
และให้ผลสรุป

Script & Output

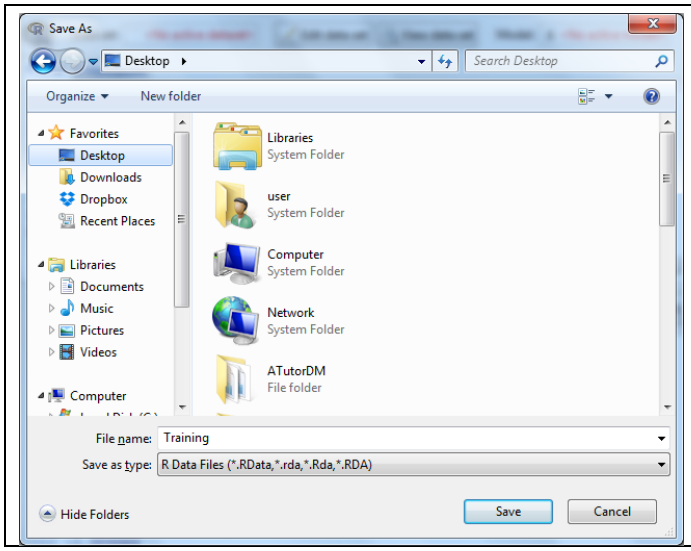
```
> numSummary(DataAll[,"CP_All"], statistics=c("mean", "sd", "IQR",
+   "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean        sd IQR 0%      25%      50%      75% 100%  n
 2.549074 0.8238611 0.9  1 2.116667 2.533333 3.016667  4.6 72
```

ถ้าให้ความหมายของ Rating scale 5 ระดับ เป็น
1=น้อยที่สุด, 2=น้อย, 3=ปานกลาง, 4=มาก, 5=มากที่สุด ก็จะได้เกณฑ์การประเมิน 2 แบบ คือ
แบบที่ 1: ใช้ค่ากลางในการให้ความหมาย จะได้เกณฑ์ดัง**รูปซ้าย** ซึ่งช่วงของหัวท้ายจะไม่เท่าช่วงตรงกลาง
แบบที่ 2: ใช้แบบช่วงเท่าในการให้ความหมาย คือ แบ่ง max – min ออกเป็น 5 ช่วงเท่า ๆ กัน (ช่วงละ .8) จะได้เกณฑ์ดัง**รูปขวา**

1.00 – 1.50 = น้อยที่สุด	1.00 – 1.80 = น้อยที่สุด
1.51 – 2.50 = น้อย	1.81 – 2.60 = น้อย
2.51 – 3.50 = ปานกลาง	2.61 – 3.40 = ปานกลาง
3.51 – 4.50 = มาก	3.41 – 4.20 = มาก
4.51 – 5.00 = มากที่สุด	4.21 – 5.00 = มากที่สุด

<u>แบบฝึกหัด</u> ลองสร้างตัวใหม่ตัวใหม่ เพื่อวัด “ด้านความรู้และทักษะการทำงาน” WSK และ “ด้านอัตลักษณ์นิสิตของมหาวิทยาลัย – มีทักษะการสื่อสาร” USK ว่าอยู่ในระดับใด

10. จัดเก็บผลลัพธ์การทำงานไว้ใช้ในโอกาสต่อไป
ใน R Commander: File > Save R Workspace as ตั้งชื่อไฟล์ว่า **Training**



จะได้ไฟล์ type เป็น R Data Files

หากต้องการกลับมาใช้งานใหม่

- ใน R Commander
  Data > **Load data set**

การทดสอบสมมุติฐานทางสถิติ

การทดสอบสมมุติฐานทางสถิติเป็นกระบวนสรุปความจริงเกี่ยวกับประชากร (population) โดยศึกษาจากกลุ่มตัวอย่าง (sample) ที่เป็นตัวแทนของประชากร สถิติที่ใช้จะเรียกว่าสถิติอนุมาน (inferential statistics)

ขั้นตอนการทดสอบสมมุติฐานทางสถิติ (Hypothesis Testing Template)

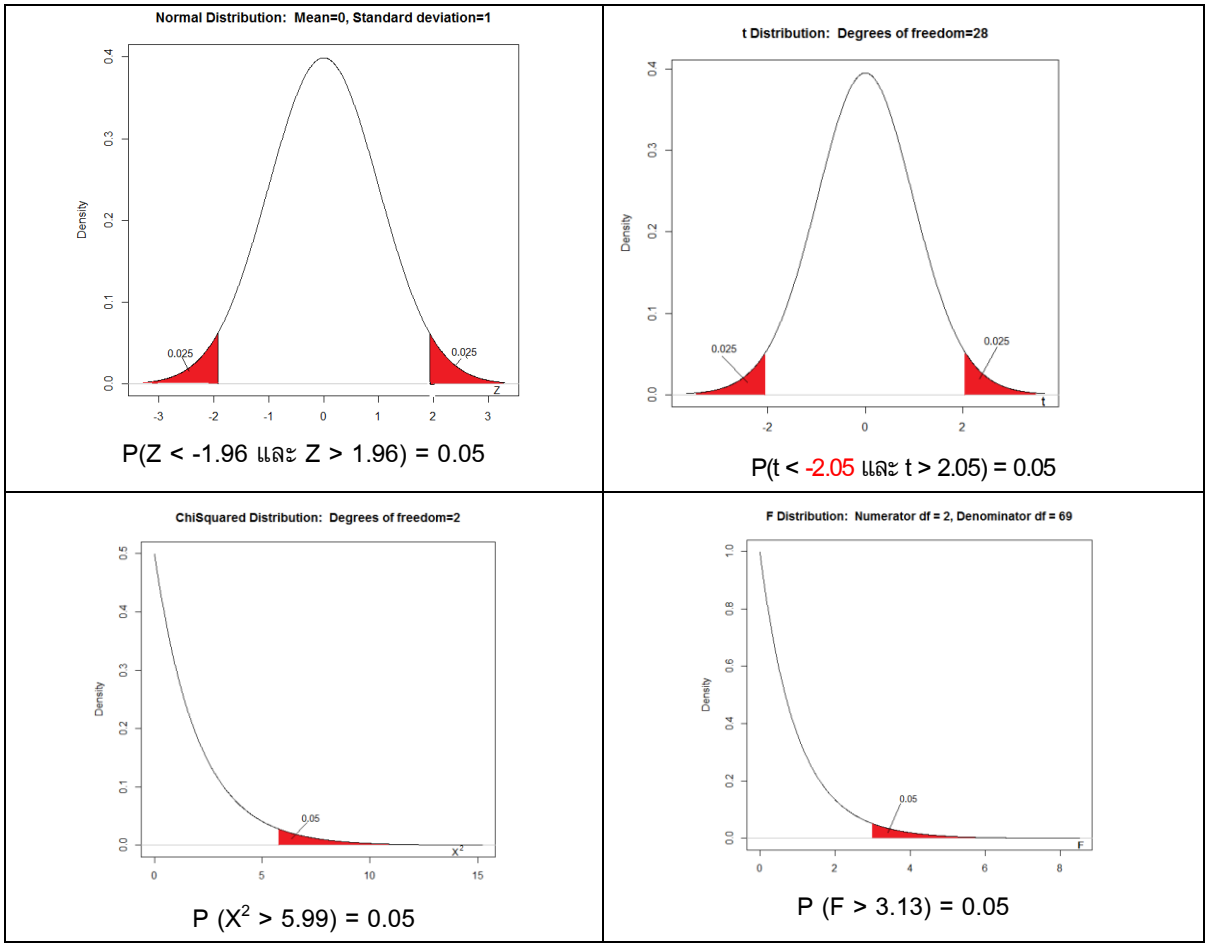
1. สมมุติฐานที่ทดสอบ H_0 และ H_1 (คุณลักษณะของประชากร ต้องเป็นพารามิเตอร์เสมอ)
   H_0: .....
   H_1: .....
2. กำหนดระดับนัยสำคัญของการทดสอบ α = ... (ปกติจะเป็น .05 หรือ .01 หรือ .10)
3. เลือกสถิติที่เหมาะสมกับสมมุติฐานที่ทดสอบในข้อ 1
   สถิตินี้คือสถิติอนุมาน/สถิติอ้างอิง และตัวสถิตินี้มีการแจกแจงความน่าจะเป็นแบบใด?
   Z = ...(สูตร)..... มีการแจกแจงแบบ Z (ปกติมาตรฐาน)
   t = ...(สูตร)..... มีการแจกแจงแบบ t ที่ df = ...
   X^2 = ...(สูตร)..... มีการแจกแจงแบบ X^2 ที่ df = ...
   F = ...(สูตร).... มีการแจกแจงแบบ F ที่ df_1 = ... และ df_2 = ...
4. กฎการตัดสินใจ ปฏิเสธ H_0 ถ้า p-value < α
5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง (โปรแกรมสำเร็จรูปจะให้ p-value ด้วย)
   เช่น t = ... p-value = ...
6. ตัดสินใจ ตีความและให้ความหมาย
   ตัดสินใจ: ปฏิเสธ H_0 / ไม่ปฏิเสธ H_0
   ตีความและให้ความหมาย .....

<u>อธิบายเพิ่มเติม</u>

2. ระดับนัยสำคัญของการทดสอบ หรือ Type I error คือ ความน่าจะเป็นที่ปฏิเสธ H_0 แต่ H_0 เป็นจริง

	ความจริงเกี่ยวกับสมมุติฐานที่ทดสอบ (Actual/Model)	
การตัดสินใจ ***(Predicted/Test)***	H_0 เป็นจริง (Positive)	H_0 เป็นเท็จ (Negative)
ไม่ปฏิเสธ H_0 / H_0 เป็นจริง *(Positive)*	ตัดสินใจถูก (True Positive: TP) Confidence Level (1 – α)	ตัดสินใจผิด (Type II error = β) (False Negative: FN)
ปฏิเสธ H_0 / H_0 เป็นเท็จ *(Negative)*	ตัดสินใจผิด (Type I error = α) (False Positive: FP)	ตัดสินใจถูก True Negative: TN) Power of the test (1 – β)

3. การแจกแจงความน่าจะเป็นของสถิติทดสอบ และค่าวิกฤตที่ระดับ α = .05



P(Z < -1.96 และ Z > 1.96) = 0.05

P(t < -2.05 และ t > 2.05) = 0.05

P (X^2 > 5.99) = 0.05

P (F > 3.13) = 0.05

4. กฎการตัดสินใจ

<u>แบบเดิม</u>: หาค่าวิกฤติ โดยเปิดตารางการแจกแจงของข้อ 3 เช่น

สำหรับการแจกแจงแบบ Z ค่าวิกฤตคือ -1.96 และ 1.96 บริเวณวิกฤตคือ Z < -1.96 และ Z > 1.96

ปฏิเสธ H_0 ถ้า ค่าสถิติในข้อ 3 ตกอยู่ในบริเวณวิกฤติ (บริเวณปฏิเสธ) ซึ่งมีความน่าจะเป็น = α

<u>แบบใหม่</u>: ใช้ค่าพี (p-value) (คือความน่าจะเป็นจากค่าสถิติจนถึงหางของการแจกแจง) เช่น

ถ้า Z = 2.00 p-value คือ P(Z >= 2.00) + P(Z <= 2.00) = 0.02275013 + 0.02275013 = 0.0455

6. ถ้าการตัดสินใจเป็น ปฏิเสธ H_0 เรียกว่า การทดสอบมีนัยสำคัญ (the test is significant)

แสดงว่า H_0 (ความรู้เดิม) ไม่จริง H_1 จริง (H_1 คือความรู้ใหม่ที่นักวิจัยค้นพบ)

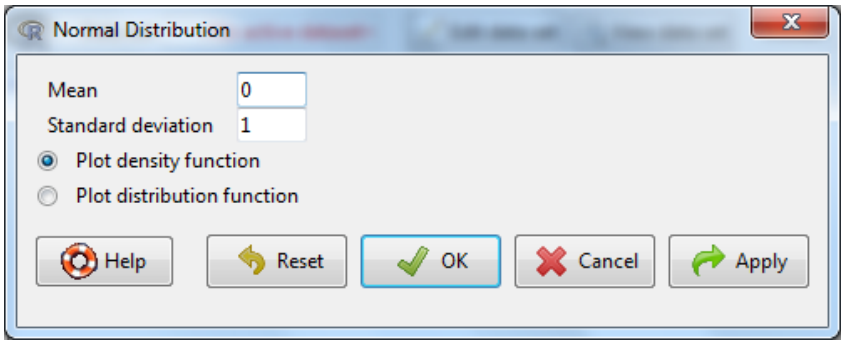
ดังนั้นเป้าหมายของการวิจัยคือ การปฏิเสธ H_0 ซึ่งก็คือต้องการให้การทดสอบมีนัยสำคัญ

ซึ่งต่างจากการทดสอบข้อตกลงเบื้องต้น (assumption) เช่น การทดสอบ Normal distribution หรือการทดสอบการเท่ากันของ Variance คือ<u>ไม่</u>ต้องการปฏิเสธ H_0

Workshop 4: การแจกแจงความน่าจะเป็นแบบต่าง ๆ การหาค่าวิกฤตที่ระดับนัยสำคัญต่าง ๆ และการหาค่าพี (p-value)

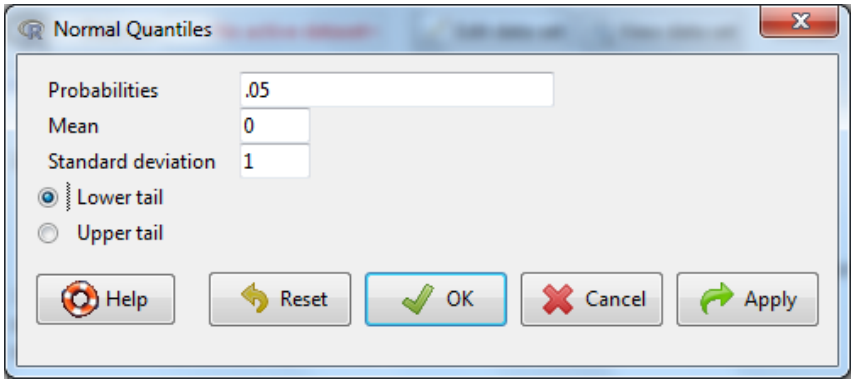
1. Plot การแจกแจงความน่าจะเป็นของการแจกแจงปกติมาตรฐาน N(0,1)

   Distributtion > Continuous distributions > Normal distribution > **Plot normal distribution**



2. หาค่าวิกฤต (ค่าสถิติ) ที่ระดับนัยสำคัญต่าง ๆ (α)

   Distributtion > Continuous distributions > Normal distribution > **Normal quantiles**



   Probabilities คือค่า α (one tail)

   ผลลัพธ์ที่ได้คือค่าวิกฤต (ค่าสถิติ)

   Script & Output

```
> qnorm(c(0.05), mean=0, sd=1, lower.tail=TRUE)
[1] -1.644854
```

   หากต้องการค่าวิกฤตของ α 2 tails ต้องหาร 2 คือต้องใส่ Probabilites = .025 จะได้ -1.959964 หรือ -1.96

   ลองดำเนินการด้วยค่า α อื่น ๆ

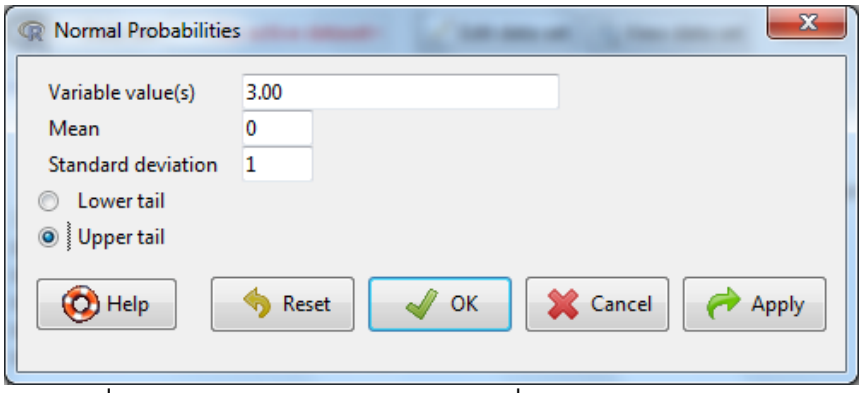
<u>สรุป</u> P(Z < -1.96 and Z > 1.96) = 0.05

P(Z < -1.64 and Z > 1.64) = 0.10

P(Z < -2.58 and Z > 2.58) = 0.01

3. หาค่าพี (p-value) จากค่าสถิติที่คำนวณได้

Distributtion > Continuous distributions > Normal distribution > **Normal probabilities**



ผลลัพธ์ที่ได้คือค่าพี (p-value) ของ**ค่าสถิติ**ที่ใส่ใน Variable value(s)

Script & Output

```
> pnorm(c(3.0), mean=0, sd=1, lower.tail=FALSE)
[1] 0.001349898
```

ดังนั้น p-value ของ 3.00 one tail คือ 0.001349898 ถ้าต้องการ 2 tails ต้องคูณ 2 จะได้ 0.002699796

สามารถลองเล่นกับ distribution อื่น ๆ เช่น t distribution, Chi-squared distribution, F distribution

การแจกแจงแบบปกติ

โดยธรรมชาติแล้วข้อมูลเชิงปริมาณส่วนใหญ่จะมีการแจกแจงแบบปกติ (Normal distribution) เช่น X1, X2, X3, …. เป็นค่าของข้อมูลของกลุ่มตัวอย่างที่มาสุ่มมาจากประชากรที่มีการแจกแจงปกติ เขียนสัญลักษณ์ $X_i \sim N(\mu, \sigma^2)$ และ ค่าเฉลี่ยของ X คือ $\overline{X}$ ซึ่งจะมีการแจกแจงแบบ $N(\mu, \sigma^2/N)$

การแจกแจงแบบปกติเป็นข้อตกลงเบื้องต้นของสถิติ**พาราเมตริก** (parametric statistics) หากข้อมูลไม่มีการแจกแจงแบบปกติก็ต้องไปใช้สถิติแบบ**นอนพาราเมตริก** (nonparametric statistics)

การทำ transformation ของข้อมูลอาจจะมีความจำเป็นในบางกรณีในกระบวนการวิเคราะห์ข้อมูลด้วย data mining ที่สำคัญ 2 อย่างคือ การทำให้เป็นปกติมาตรฐาน (**standardization)** และการ **normalization** เพื่อให้ค่าของตัวแปรอยู่ในมาตรวัดเดียวกัน

การทำให้เป็นปกติมาตรฐานคือการลบด้วยค่าเฉลี่ยและหารด้วยค่าเบี่ยงเบนมาตรฐาน ค่าปกติมาตรฐานจะมีค่าเฉลี่ยเป็น 0 และค่าเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งหลักการนี้เป็นหลักการที่สำคัญในสถิติอนุมาน

การ **normalization** คือการที่ทำให้ข้อมูลมีค่าอยู่ระหว่าง 0 และ 1 ทำได้โดยลบข้อมูลด้วย min และหารด้วย max – min

Workshop 5 จงทดสอบว่า CP_All มีการแจกแจงปกติหรือไม่

Statistics > Summaries > **Shapio-Wilk test for normality** เลือก Variable > CP_All

Script & Output

```
> with(DataAll, shapiro.test(CP_All))

        Shapiro-Wilk normality test

data:  CP_All
W = 0.9796, p-value = 0.292
```

การทดสอบสมมุติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H_0 และ H_1

   H_0: CP_All มีการแจกแจงแบบปกติ

   H_1: CP_All ไม่มีการแจกแจงแบบปกติ

2. กำหนดระดับนัยสำคัญของการทดสอบ α = 0.05
3. สถิติทดสอบ

   Shapiro-Wilk test for normality

4. กฎการตัดสินใจ ปฏิเสธ H_0 ถ้า p-value < α
5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง

   ค่าสถิติ W = 0.9796 p-value = 0.292

6. ตัดสินใจ ตีความและให้ความหมาย

   ไม่ปฏิเสธ H_0

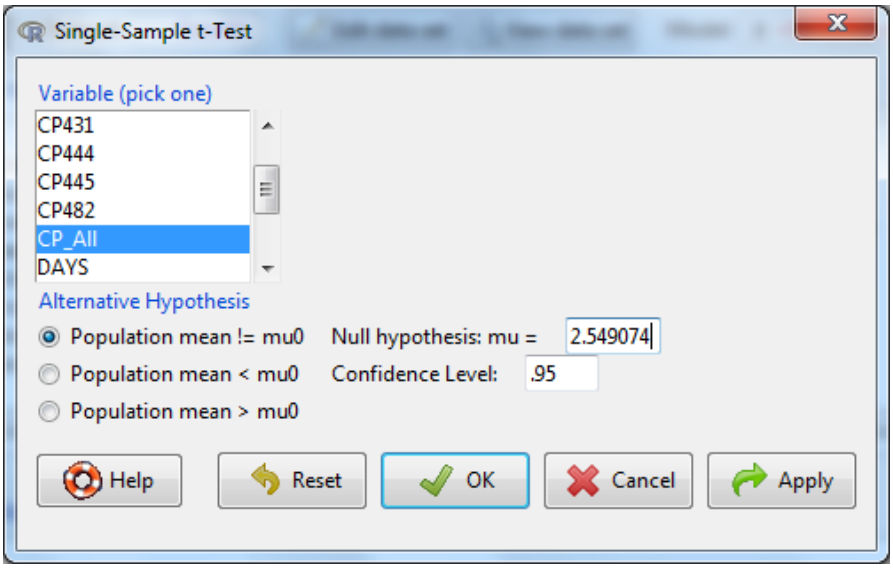
   CP_All มีการแจกแจงปกติ

ทดสอบค่าเฉลี่ยของประชากร 1 กลุ่ม

การหา 95% ช่วงความเชื่อมั่นของ $\mu = \bar{X} \pm t_{0.025}$*SD/sqrt(N)

Workshop 6: จงหา 95% ช่วงความเชื่อมั่นของค่าเฉลี่ยของ CP_All (1 sample t-test)

คำสั่ง: Statistics > Mean > **Single-Sample t-test** ใส่ค่า Null hypothesis mu = 2.549074



Script & Output

```
> with(DataAll, (t.test(CP_All, alternative='two.sided', mu=2.549074,
+   conf.level=.95)))

        One Sample t-test

data:  CP_All
t = 0, df = 71, p-value = 1
alternative hypothesis: true mean is not equal to 2.549074
95 percent confidence interval:
 2.355476 2.742672
sample estimates:
mean of x
 2.549074
```

95% ช่วงความเชื่อมั่นของค่าเฉลี่ยของ CP_All

จงทดสอบให้เห็นจริง โดยการแทนค่าในสูตร 2.549074 ± 1.993943*(0.8238611/sqrt(72))

= 2.549074 ± 0.193598

ทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 2 กลุ่ม

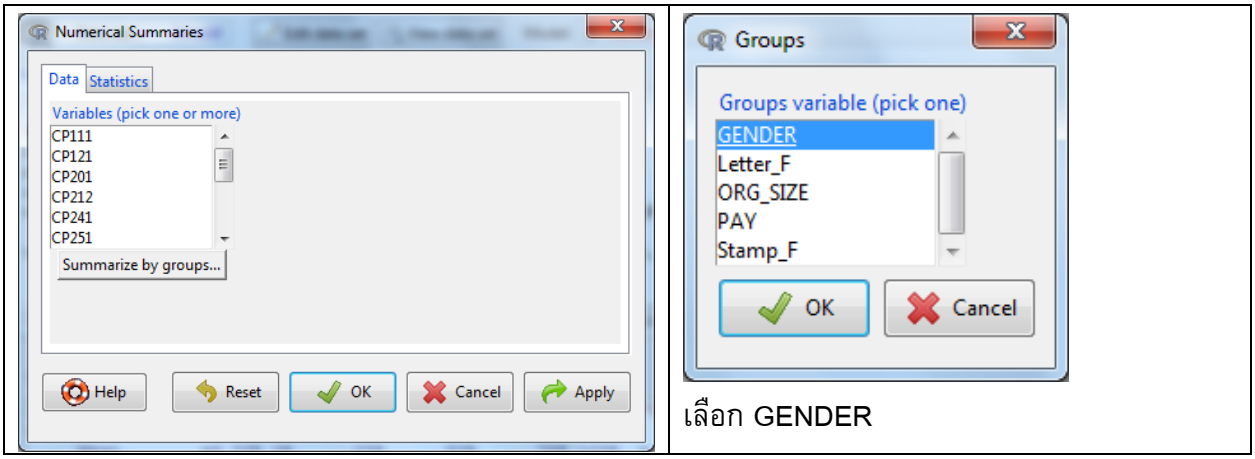
Workshop 7: นิสิตหญิงและนิสิตชายประเมินการใช้ความรู้ที่เรียนในชั้นเรียนกับการทำงาน (CP_All) แตกต่างกันหรือไม่

หมายเหตุ 1. ตัวแปรที่เราต้องการจะศึกษา/ต้องการที่จะทดสอบ (CP_All) เรียกว่าตัวแปรตาม (Dependent variable) ตัวแปรกลุ่ม (GENDER) เรียกว่าตัวแปรต้น (Independent variable) หรือตัวแปรกลุ่ม (Group variable)

หมายเหตุ 2. ใน Data warehouse Measure attribute ก็คือ Dependent variable และ Dimension attribute ก็คือ Independent variable หรือ Group variable นั่นเอง

1. หาค่าสถิติพรรณนา Mean, SD, N ของ CP_ALL ของกลุ่ม ชาย และ หญิง

คำสั่ง: Statistics > Summaries > **Numerical Summaries** (ซ้าย) คลิก Summarize by groups (ขวา)



เลือก GENDER

Script & Output

```
> numSummary(DataAll[,"CP_All"], groups=DataAll$GENDER, statistics=c("mean",
+   "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
         mean        sd      IQR 0%  25%      50%      75% 100% data:n
ชาย  2.660000 0.9011294 1.133333  1 2.20 2.533333 3.333333  4.6     40
หญิง 2.410417 0.7051494 0.750000  1 2.05 2.366667 2.800000  4.0     32
```

ให้สังเกตว่า ผลลัพธ์ข้างต้น ก็คือ เมเซอร์ 8 ตัว (mean, sd, IQR, 0%, 25%, 50%, 75%, 100%, data:n) ซึ่ง data:n aggregation function คือ count นั่นเอง ส่วนเมเซอร์ตัวอื่น ๆ ก็คือ CM (Calculated Member)

2. ทดสอบค่าเฉลี่ย 2 ค่า

คำถาม แบบที่ 1: Mean ของ CP_All ของชาย (2.660000) และของหญิง (2.410417) แตกต่างกันอย่างมีนัยสำคัญหรือไม่?

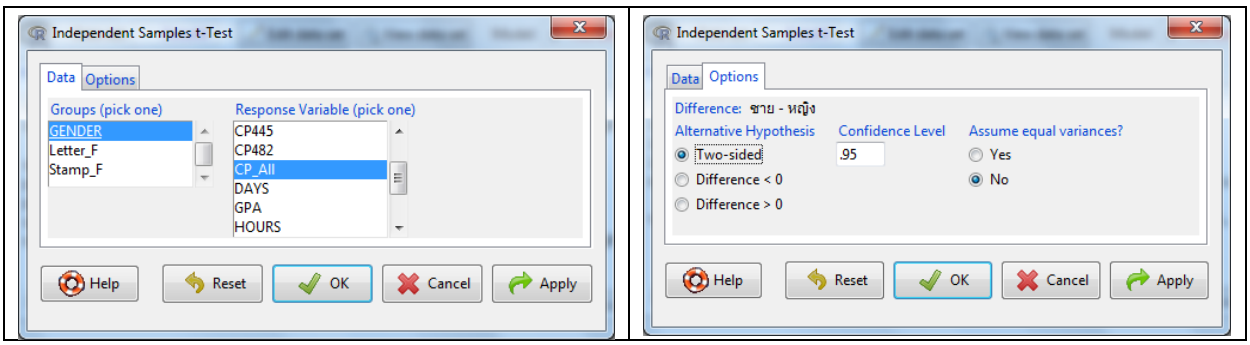
คำถาม แบบที่ 2: Mean ของ CP_All ของชาย และของหญิง แตกต่างกันหรือไม่?

ในแบบที่ 1 Mean คือ Sample Mean ($\bar{X}_{ชาย}$ และ $\bar{X}_{หญิง}$)

ในแบบที่ 2 Mean คือ Population Mean ($\mu_{ชาย}$ และ $\mu_{หญิง}$)

ค้นหาคำตอบด้วยการทดสอบ Independent samples t-test

คำสั่ง: Statistics > Means > **Independent samples t-test**

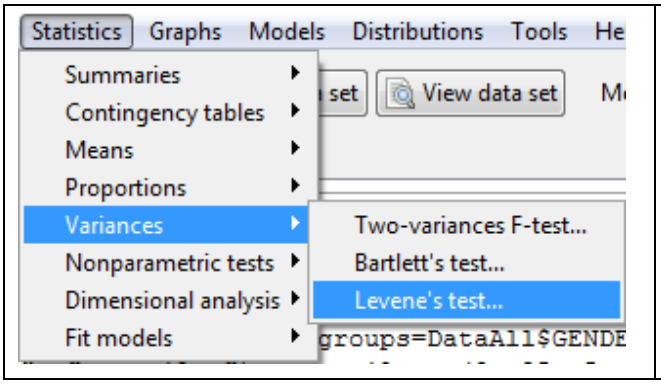


ซึ่ง **Independent samples t-test** มี 2 สูตร คือ สูตร Variances เท่า และ ไม่เท่า

ดังนั้นจึงต้องทดสอบ Variances ก่อน

3. ทดสอบ Variance 2 ค่า

คำสั่ง: Statistics > Variances > **Lavene's test**



มีสถิติให้เลือก 3 ตัว คือ

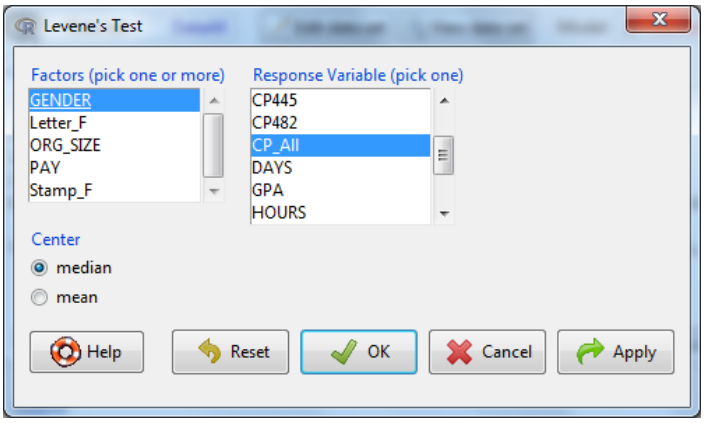
1.Two-variances F-test

2.Bartlett's test

3.Levene's test

1 และ 2 มีสูตรคำนวณปรากฏในตำราสถิติพื้นฐานทั่วไป

3. SPSS ใช้ตัวนี้



Script & Output

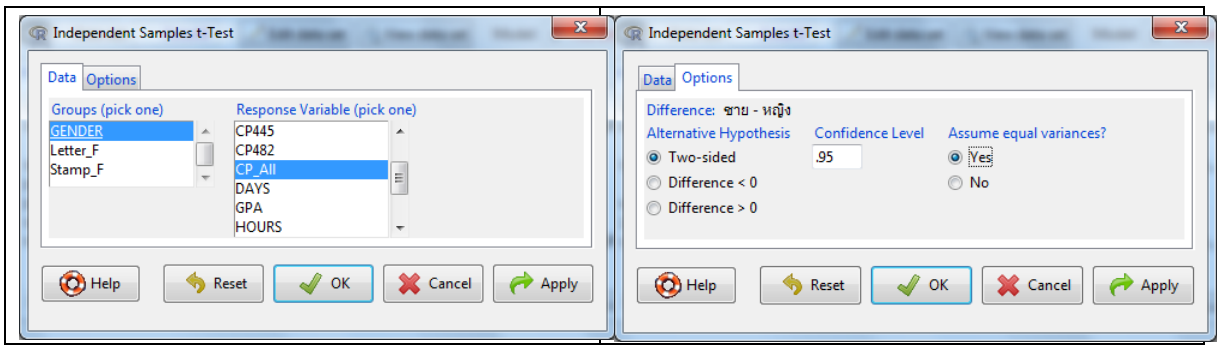
```
> leveneTest(CP_All ~ GENDER, data=DataAll, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  1  2.3079 0.1332
      70
```

ทดสอบสมมุติฐานในใจ: p=value = 0.1332 แสดงว่าไม่ปฏิเสธ H_0, H_0 คือ Variances เท่า

ดังนั้น สรุปได้ว่า Variance เท่า

4. ทดสอบค่าเฉลี่ย 2 ค่า โดยใช้ **Independent samples t-test** สูตร Variance เท่า

   คำสั่ง: Statistics > Means > **Independent samples t-test** (ใน Options: Assume equal variances? = **Yes**)



Script & Output

```
> t.test(CP_All~GENDER, alternative='two.sided', conf.level=.95,
+   var.equal=TRUE, data=DataAll)

        Two Sample t-test

data:  CP_All by GENDER
t = 1.2831, df = 70, p-value = 0.2037
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1383594  0.6375260
sample estimates:
 mean in group ชาย mean in group หญิง
          2.660000           2.410417
```

การทดสอบสมมุติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H_0 และ H_1

   H_0: $\mu_{ชาย} = \mu_{หญิง}$ หรือ $\mu_{ชาย} - \mu_{หญิง} = 0$ เมื่อ μ คือค่าเฉลี่ยของ CP_All

   H_1: $\mu_{ชาย} \neq \mu_{หญิง}$ หรือ $\mu_{ชาย} - \mu_{หญิง} \neq 0$

2. กำหนดระดับนัยสำคัญของการทดสอบ $\alpha = 0.05$

3. สถิติทดสอบ

   Independent samples t-test

   $$t = \frac{(\bar{x}1 - \bar{x}2)}{\sqrt{\frac{(n1-1)s1^2+(n2-1)s2^2}{n1+n2-2}(\frac{1}{n1}+\frac{1}{n2})}}$$ มีการแจกแจงแบบ t ที่ df = 40+32 -2 = 70 เมื่อ H_0 เป็นจริง

4. กฎการตัดสินใจ ปฏิเสธ H_0 ถ้า p-value < α

5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง

   ค่าสถิติ t = 1.2831 p-value = 0.2037

6. ตัดสินใจ ตีความและให้ความหมาย

   ไม่ปฏิเสธ H_0

   คำตอบแบบที่ 1: Mean ของ CP_All ของชาย และของหญิง แตกต่างกันอย่างไม่มีนัยสำคัญ

   คำตอบแบบที่ 2: Mean ของ CP_All ของชาย และของหญิง ไม่แตกต่างกัน

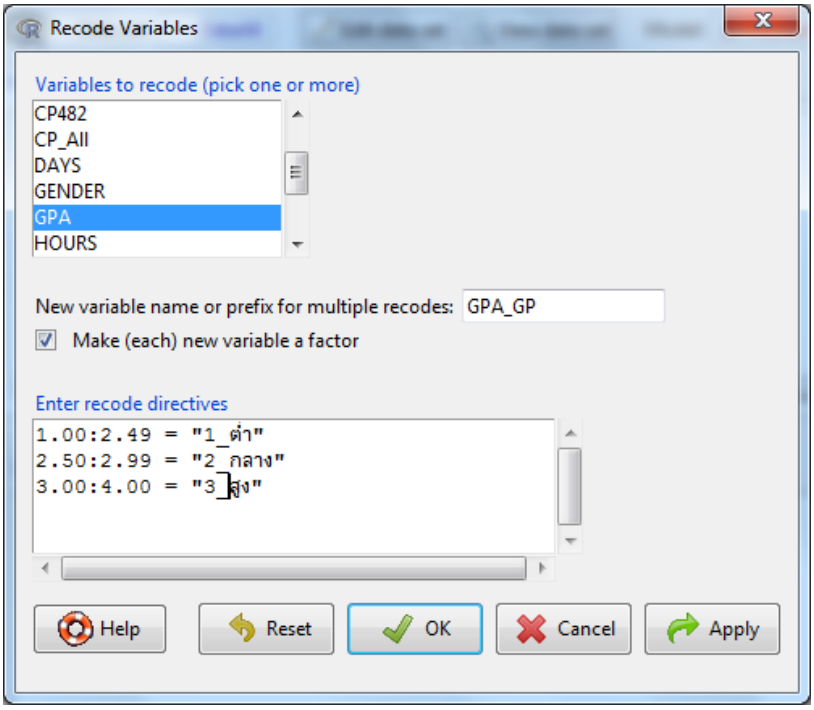
หมายเหตุ **Independent samples t-test** เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องต้นว่าข้อมูลของประชากร 2 กลุ่มต้องมีการแจกแจงปกติ หากข้อตกลงเบื้องต้นนี้ไม่เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริกที่คู่กัน ด้วยคำสั่ง Statistics > Nonparametric tests > **Two-sample Wilcoxon test** การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

ทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 3 กลุ่ม

Workshop 8: นิสิต 3 กลุ่ม ที่มี GPA 3 ระดับคือ ต่ำ (< 2.50) ปานกลาง (2.50 – 2.99) สูง (>= 3.00) ประเมินการใช้ความรู้ที่เรียนในชั้นเรียนกับการทำงาน (CP_All) แตกต่างกันหรือไม่ (One Way ANOVA)

1. สร้างตัวแปรตัวใหม่ (transformation) ชื่อ GPA_GP โดยแปลงค่า GPA ที่มีค่าต่อเนื่องให้มีค่าเป็น 3 กลุ่ม (discretization) ตามโจทย์

คำสั่ง: Data > Manages variables in active data set > **Recode variables**

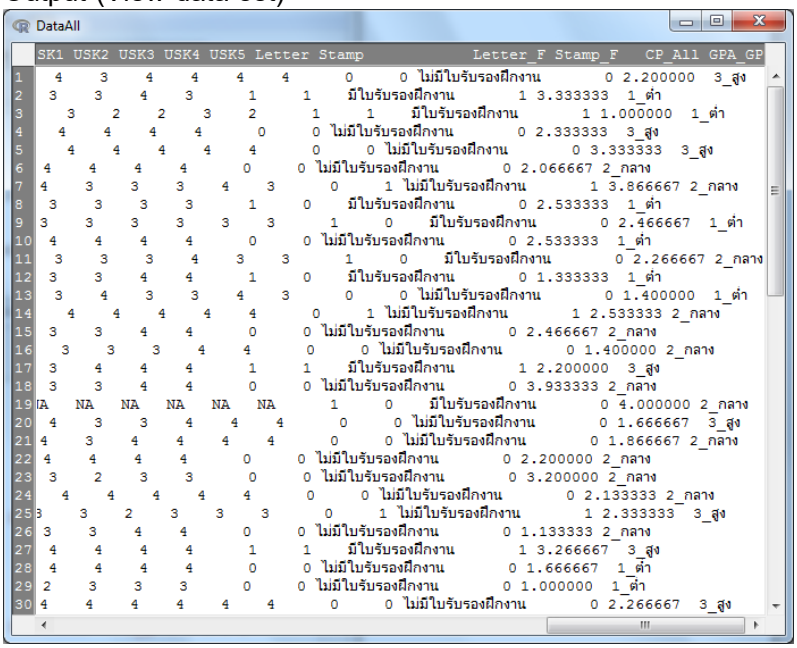


หมายเหตุ หากไม่ทราบว่าจะ Enter recode directives อย่างไร ให้กดปุ่ม Help

Script

```
> DataAll <- within(DataAll, {
+   GPA_GP <- Recode(GPA,
+   '1.00:2.49 = "1_ต่ำ"; 2.50:3.00 = "2_กลาง"; 3.00:4.00 = "3_สูง"; ; ;',
+   as.factor.result=TRUE)
+ })
```

Output (View data set)



DataAll

	SK1	USK2	USK3	USK4	USK5	Letter	Stamp	Letter_F	Stamp_F	CP_All	GPA_GP
1	4	3	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.200000	3_สูง
2	3	3	4	3	1	1	มีใบรับรองฝึกงาน	1	3.333333	1_ต่ำ	
3	3	2	2	3	2	1	1	มีใบรับรองฝึกงาน	1	1.000000	1_ต่ำ
4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.333333	3_สูง	
5	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	3.333333	3_สูง
6	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.066667	2_กลาง	
7	4	3	3	3	4	3	0	1	ไม่มีใบรับรองฝึกงาน	1	3.866667	2_กลาง
8	3	3	3	3	1	0	มีใบรับรองฝึกงาน	0	2.533333	1_ต่ำ	
9	3	3	3	3	3	1	0	มีใบรับรองฝึกงาน	0	2.466667	1_ต่ำ
10	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.533333	1_ต่ำ	
11	3	3	3	4	3	1	0	มีใบรับรองฝึกงาน	0	2.266667	2_กลาง
12	3	3	4	4	1	0	มีใบรับรองฝึกงาน	0	1.333333	1_ต่ำ	
13	3	4	3	3	4	3	0	0	ไม่มีใบรับรองฝึกงาน	0	1.400000	1_ต่ำ
14	4	4	4	4	4	0	1	ไม่มีใบรับรองฝึกงาน	1	2.533333	2_กลาง
15	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.466667	2_กลาง	
16	3	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	1.400000	2_กลาง
17	3	4	4	4	1	1	มีใบรับรองฝึกงาน	1	2.200000	3_สูง	
18	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	3.933333	2_กลาง	
19	NA	NA	NA	NA	NA	NA	1	0	มีใบรับรองฝึกงาน	0	4.000000	2_กลาง
20	4	3	3	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	1.666667	3_สูง
21	4	3	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	1.866667	2_กลาง
22	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.200000	2_กลาง	
23	3	2	3	3	0	0	ไม่มีใบรับรองฝึกงาน	0	3.200000	2_กลาง	
24	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.133333	2_กลาง
25	3	3	2	3	3	3	0	1	ไม่มีใบรับรองฝึกงาน	1	2.333333	3_สูง
26	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	1.133333	2_กลาง	
27	4	4	4	4	1	1	มีใบรับรองฝึกงาน	1	3.266667	3_สูง	
28	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	1.666667	1_ต่ำ	
29	2	3	3	3	0	0	ไม่มีใบรับรองฝึกงาน	0	1.000000	1_ต่ำ	
30	4	4	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน	0	2.266667	3_สูง

2. ตรวจสอบผลการ Discritization

คำสั่ง: Statistics > Summaries > **Frequency distributions**

```
> local({
+   .Table <- with(DataAll, table(GPA_GP))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })

counts:
GPA_GP
 1_ต่ำ 2_กลาง  3_สูง
    17      33      22

percentages:
GPA_GP
 1_ต่ำ 2_กลาง  3_สูง
 23.61   45.83   30.56
```

3. หาค่าสถิติพรรณนา Mean, SD, N ของ CP_ALL ของ 3 กลุ่ม (GPA_GP)

คำสั่ง: Statistics > Summaries > **Numerical Summaries** Variables: เลือก CP_All, Summarize by groups: เลือก GPA_GP, Statistics เลือก Mean, Standard Deviation

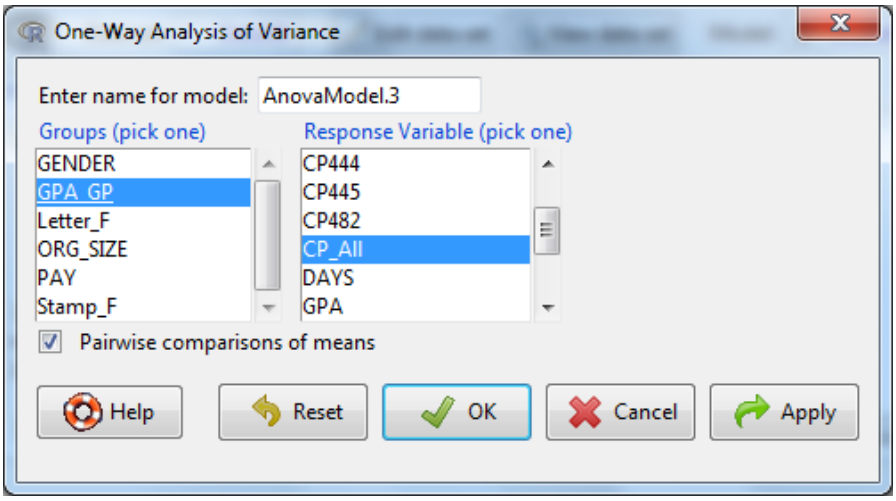
Script & Output

```
> numSummary(DataAll[,"CP_All"], groups=DataAll$GPA_GP, statistics=c("mean",
+   "sd"), quantiles=c(0,.25,.5,.75,1))
           mean        sd data:n
1_ต่ำ  2.309804 0.9055205     17
2_กลาง 2.618182 0.8980338     33
3_สูง  2.630303 0.6167768     22
```

4. One Way ANOVA มีข้อตกลงเบื้องต้นว่า Variance ของ 3 กลุ่มต้องเท่ากัน ถ้าไม่เท่าต้องไปใช้ Nonparametric ที่คู่กัน

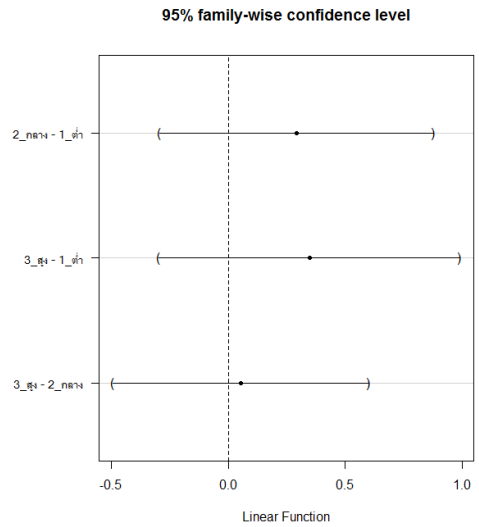
ผลการทดสอบ Variance เท่า ดังนั้นจึงใช้ One Way ANOVA ได้ (ลองทดสอบ)

คำสั่ง: Statistics > Means > **One-way ANOVA** (เลือก **Pairwise comparisons of means**)



Script & Output

```
> AnovaModel.1 <- aov(CP_All ~ GPA_GP, data=DataAll)

> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value Pr(>F)
GPA_GP        2   1.31  0.6568   0.967  0.385
Residuals    69  46.88  0.6794

> with(DataAll, numSummary(CP_All, groups=GPA_GP, statistics=c("mean", "sd")))
           mean        sd data:n
1_ต่ำ  2.309804 0.9055205     17
2_กลาง 2.601961 0.8893663     34
3_สูง  2.657143 0.6187032     21
```




```
> local({
+    .Pairs <- glht(AnovaModel.1, linfct = mcp(GPA_GP = "Tukey"))
+    print(summary(.Pairs)) # pairwise tests
+    print(confint(.Pairs)) # confidence intervals
+    print(cld(.Pairs)) # compact letter display
+    old.oma <- par(oma=c(0,5,0,0))
+    plot(confint(.Pairs))
+    par(old.oma)
+ })

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = CP_All ~ GPA_GP, data = DataAll)

Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
2_กลาง - 1_ต่ำ == 0  0.30838    0.24617   1.253    0.425
3_สูง - 1_ต่ำ == 0   0.32050    0.26627   1.204    0.454
3_สูง - 2_กลาง == 0  0.01212    0.22696   0.053    0.998
(Adjusted p values reported -- single-step method)
```

```
         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = CP_All ~ GPA_GP, data = DataAll)

Quantile = 2.3927
95% family-wise confidence level


Linear Hypotheses:
                     Estimate lwr      upr
2_กลาง - 1_ต่ำ == 0  0.30838 -0.28064  0.89740
3_สูง - 1_ต่ำ == 0   0.32050 -0.31663  0.95762
3_สูง - 2_กลาง == 0  0.01212 -0.53093  0.55517

 1_ต่ำ 2_กลาง  3_สูง
   "a"    "a"    "a"
```

การทดสอบสมมติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H_0 และ H_1

   H_0: $\mu_1 = \mu_2 = \mu_3$ เมื่อ μ_i คือค่าเฉลี่ยของ CP_All ของ GPA_GP แต่ละกลุ่ม (ต่ำ, กลาง, สูง)

   H_1: มี $\mu_i \neq \mu_j$ อย่างน้อย 1 คู่ (มีกี่คู่?)

2. กำหนดระดับนัยสำคัญของการทดสอบ α = 0.05

3. สถิติทดสอบ

   One way ANOVA (ANalysis Of Variances)

   $F = \frac{MS_b}{MS_w} = \frac{SS_b/(k-1)}{SS_w/(N-k)}$; มีการแจกแจงแบบ F ที่ df = 3-1=2 และ df = 72-3 =69 ($\sim F_{2,69}$)

   เมื่อ H_0 เป็นจริง

4. กฎการตัดสินใจ ปฏิเสธ H_0 ถ้า p-value < α

5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง

   ค่าสถิติ F = 0.967 p-value = 0.385

6. ตัดสินใจ ตีความและให้ความหมาย

   ไม่ปฏิเสธ H_0

   ค่าเฉลี่ยของ CP_All ของนิสิต 3 กลุ่ม (GPA_GP) ไม่แตกต่างกัน

   ซึ่งสอดคล้องกับการเปรียบเทียบรายคู่ (Miltiple Cpmparisons of Means) ด้วยวิธี Tukey

   ไม่มีคู่ใดที่แตกต่างกัน ดังนั้นค่าเฉลี่ยของทุกกลุ่มจะอยู่ใน set เดียวกัน (สัญลักษณ์ “a”)

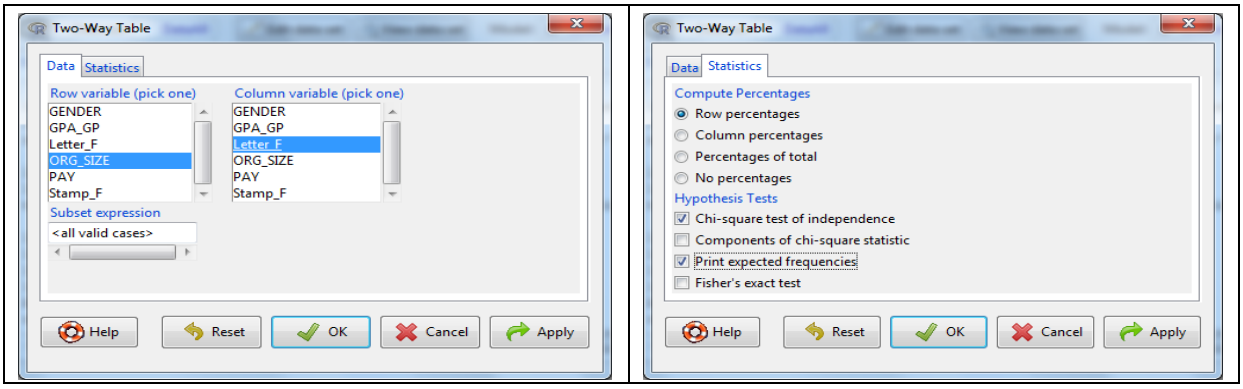
หมายเหตุ **One-way ANOVA** เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องต้นว่าข้อมูลของประชากรทุกกลุ่มต้องมีการแจกแจงปกติ และความแปรปรวนของประชากรทุกกลุ่มต้องเท่ากัน หากข้อตกลงเบื้องต้นนี้ไม่เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริกที่คู่กัน ด้วยคำสั่ง Statistics > Nonparametric tests > **Kruskal-Wallis test** การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

ทดสอบความเป็นอิสระของตัวแปรเชิงคุณลักษณะ

Workshop 9: การออกใบรับรองการฝึกงาน (Letter) เป็นอิสระหรือสัมพันธ์กับขนาดขององค์กรที่ฝึกงาน (ORG_SIZE) (Chi-square test for independent)

1. หาค่าสถิติพรรณนา จะได้เป็นตารางแจกแจงความถี่ 2 ทาง (Cross tab หรือ Table)

คำสั่ง: Statistics > **Contingency tables > Two-way tables**



Script & Output

```
> local({
+   .Table <- xtabs(~ORG_SIZE+Letter_F, data=DataAll)
+   cat("\nFrequency table:\n")
+   print(.Table)
+   cat("\nRow percentages:\n")
+   print(rowPercents(.Table))
+   .Test <- chisq.test(.Table, correct=FALSE)
+   print(.Test)
+   cat("\nExpected counts:\n")
+   print(.Test$expected)
+ })

Frequency table:
                        Letter_F
ORG_SIZE                 ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน
  ขนาดกลาง (15 - 50 คน)                  10                1
  ขนาดเล็ก (น้อยกว่า 15 คน)               7                1
  ขนาดใหญ่ (มากกว่า 50 คน)                29               23

Row percentages:
                        Letter_F
ORG_SIZE                 ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน Total Count
  ขนาดกลาง (15 - 50 คน)                90.9              9.1   100    11
  ขนาดเล็ก (น้อยกว่า 15 คน)             87.5             12.5   100     8
  ขนาดใหญ่ (มากกว่า 50 คน)              55.8             44.2   100    52
```

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 6.9529, df = 2, p-value = 0.03092


Expected counts:
                          Letter_F
ORG_SIZE                   ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน
  ขนาดกลาง (15 - 50 คน)                7.126761         3.873239
  ขนาดเล็ก (น้อยกว่า 15 คน)             5.183099         2.816901
  ขนาดใหญ่ (มากกว่า 50 คน)              33.690141        18.309859
```

Messages

```
[5] WARNING:
2 expected frequencies are less than 5
```

หมายเหตุ ให้สังเกต ใน WARNING: สำหรับการทดสอบไค-สแคว์ ค่าความถี่ควรต้องมีจำนวน > 5

การทดสอบสมมติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H_0 และ H_1

   H_0: Letter และ ORG_SIZE เป็นอิสระต่อกัน (คือไม่มีความสัมพันธ์กัน)

   H_1: Letter และ ORG_SIZE ไม่เป็นอิสระต่อกัน (คือมีความสัมพันธ์กัน)

2. กำหนดระดับนัยสำคัญของการทดสอบ $\alpha = 0.05$
3. สถิติทดสอบ: Chi-square test for independence

   $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$; $E_i = np_i$ มีการแจกแจงแบบ χ^2 ที่ df = 2 เมื่อ H_0 เป็นจริง
4. กฎการตัดสินใจ ปฏิเสธ H_0 ถ้า p-value < α
5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง

   ค่าสถิติ X-squared = 6.9529 p-value = 0.03092
6. ตัดสินใจ ตีความและให้ความหมาย

   ปฏิเสธ H_0

   การออกใบรับรองการฝึกงาน (Letter) ไม่เป็นอิสระหรือมีสัมพันธ์กับขนาดขององค์กรที่ฝึกงาน (ORG_SIZE)

2. Save active data set ไว้ใช้งานต่อไป

คำสั่ง: Data > Active data set > **Save active data set**

ทดสอบความสัมพันธ์ของตัวแปรเชิงปริมาณ

ความรู้และทักษะในการปฏิบัติงาน (WSK), อัตลักษณ์นิสิต “ทักษะการสื่อสาร” (USK), GPA มีความสัมพันธ์กันหรือไม่ (Correlation)

Pearson correlation สำหรับข้อมูลที่มีการแจกแจงปกติ

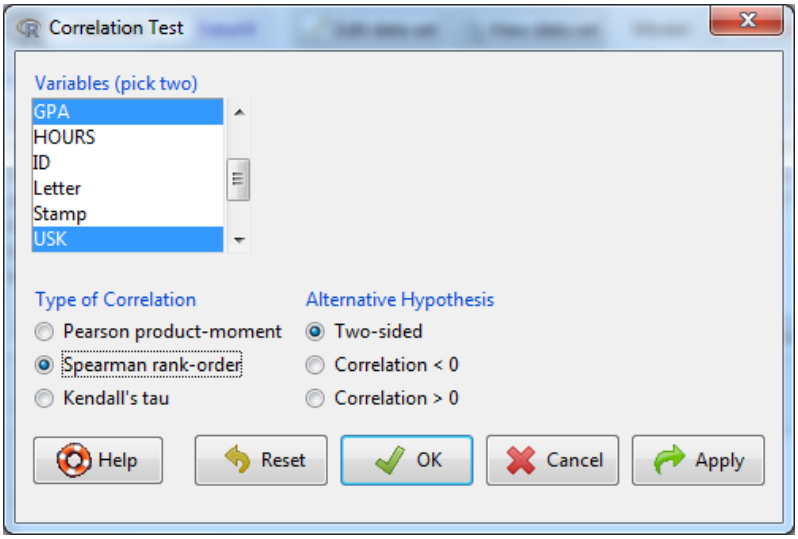
Spearman rank correlation สำหรับข้อมูลที่ไม่มีการแจกแจงปกติ

Workshop 10:

1. Load data set Training2 มาใช้งาน (มีตัวแปร WSK และ USK แล้ว)
2. ทดสอบว่า WSK, USK, GPA มีการแจกแจงปกติหรือไม่

   ผลการทดสอบ: WSK, USK ไม่มีการแจกแจงปกติ, GPA มีการแจกแจงปกติ
3. หาความสัมพันธ์

คำสั่ง: Statistics > Summaries > **Correlation test**



Script & Output

```
> with(DataAll, cor.test(GPA, USK, alternative="two.sided", method="spearman"))

        Spearman's rank correlation rho

data:  GPA and USK
S = 51671.69, p-value = 0.4295
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0959376
```

```
> with(DataAll, cor.test(GPA, WSK, alternative="two.sided", method="spearman"))

        Spearman's rank correlation rho

data:  GPA and WSK
S = 42469.79, p-value = 0.1219
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.189415
```

```
> with(DataAll, cor.test(USK, WSK, alternative="two.sided", method="spearman"))

        Spearman's rank correlation rho

data:  USK and WSK
S = 11932.41, p-value = 1.246e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7722563
```

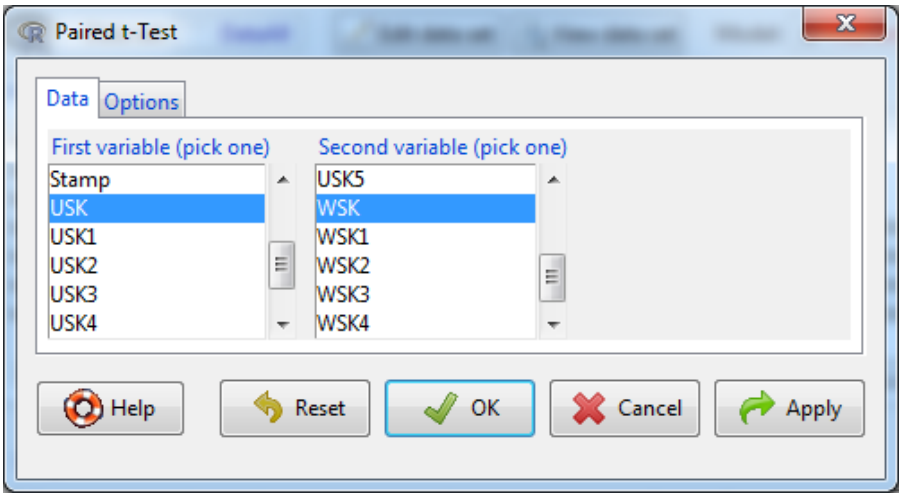
สรุป: ตัวแปรที่สัมพันธ์กันมีเพียงคู่เดียวคือ WSK และ USK

สัมประสิทธ์สหสัมพันธ์ (Correlation coefficient) = 0.7722563

เมื่อทดสอบสมมุติฐานทางสถิติจะสรุปได้ว่า ความสัมพันธ์ไม่เป็นศูนย์

4. ทดสอบว่าค่าเฉลี่ย WSK กับ USK แตกต่างกันหรือไม่ (Paired t-test)

คำสั่ง: Statistics > Means > **Paired t-test**



Script & Output

```
> with(DataAll, (t.test(USK, WSK, alternative='two.sided', conf.level=.95, paired=TRUE)))

        Paired t-test

data:  USK and WSK
t = -0.7573, df = 67, p-value = 0.4515
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.10693308  0.04810956
sample estimates:
mean of the differences
            -0.02941176
```

สรุป ทักษะไม่แตกต่างกัน

หมายเหตุ **Paired t-test** เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องต้นว่าผลต่างของข้อมูล (paired) ของประชากร ต้องมีการแจกแจงปกติ หากข้อตกลงเบื้องต้นนี้ไม่เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริกที่คู่กันด้วยคำสั่ง Statistics > Nonparametric tests > **Paired-samples Wilcoxon test**

การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

การพยากรณ์หรือการวิเคราะห์การถดถอย (Regression)

การวิเคราะห์การถดถอยเป็นการพยากรณ์ตัวแปรที่สนใจ (ตัวแปรตาม/dependent variable 1 ตัว) หรือเป็นการอธิบายความแปรปรวนของตัวแปรที่สนใจ ด้วยตัวพยากรณ์หรือตัวอธิบาย (predictor/explainatory/independent variables หลายตัว) โดยที่ตัวแปรต้นและตัวแปรตามต้องมีความสัมพันธ์กันแบบเชิงเส้นอย่างมีนัยสำคัญ ตัวแปรทั้งหมดต้องมีระดับการวัดเป็น scale หากเป็นตัวแปร nominal ต้องมี 2 ค่าเท่านั้น หากมีมากกว่า 2 ค่าต้องมีการทำเป็น dummy variables วิธีการนี้เรียกว่า Multuple Linear Regression และมีข้อตกลงเบื้องต้นว่าตัวแปรตามที่ระดับต่าง ๆ ของตัวแปรต้นต้องมีการแจกแจงแบบปกติ และมีความแปรปรวนเท่ากัน หากตรวจสอบข้อตกลงเบื้องต้นนี้ไม่เป็นจริงอาจจะต้องทำการ Transformation ข้อมูล

หากตัวแปรตามเป็น nominal จะต้องใช้วิธี Logistic Regression

Workshop 11: หา Linear Regression Model ของ CPU data set (จาก Weka data set)

1. อ่านไฟล์ cpu.xls เข้ามาด้วยคำสั่ง .....

2. View data set จะได้

CPU

	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	class
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
6	26	8000	32000	64	8	32	318
7	23	16000	32000	64	16	32	367
8	23	16000	32000	64	16	32	489
9	23	16000	64000	64	16	32	636
10	23	32000	64000	128	32	64	1144
11	400	1000	3000	0	1	2	38
12	400	512	3500	4	1	6	40
13	60	2000	8000	65	1	8	92
14	50	4000	16000	65	1	8	138
15	350	64	64	0	1	4	10
16	200	512	16000	0	4	32	35
17	167	524	2000	8	4	15	19
18	143	512	5000	0	7	32	28
19	143	1000	2000	0	5	16	31
20	110	5000	5000	142	8	64	120
21	143	1500	6300	0	5	32	30
22	143	3100	6200	0	5	20	33
23	143	2300	6200	0	6	64	61
24	110	3100	6200	0	6	64	76
25	320	128	6000	0	1	12	23
26	320	512	2000	4	1	3	69
27	320	256	6000	0	1	6	33
28	320	256	3000	4	1	3	27
29	320	512	5000	4	1	5	77
30	320	256	5000	4	1	6	27

โจทย์ สร้างโมเดลหรือสมการพยากรณ์ประสิทธิภาพของการประมวลผลของ CPU ด้วยคุณลักษณะของคอมพิวเตอร์

Depdendent variable คือ class (ประสิทธิภาพ)

Predictor variables: MYCT, MMIN, MMAX,CACH, CHMIN, CHMAX ทั้งหมด 6 ตัว

มีข้อมูลทั้งหมด 209 รายการ

ตัวแปรทั้งหมดเป็น scale

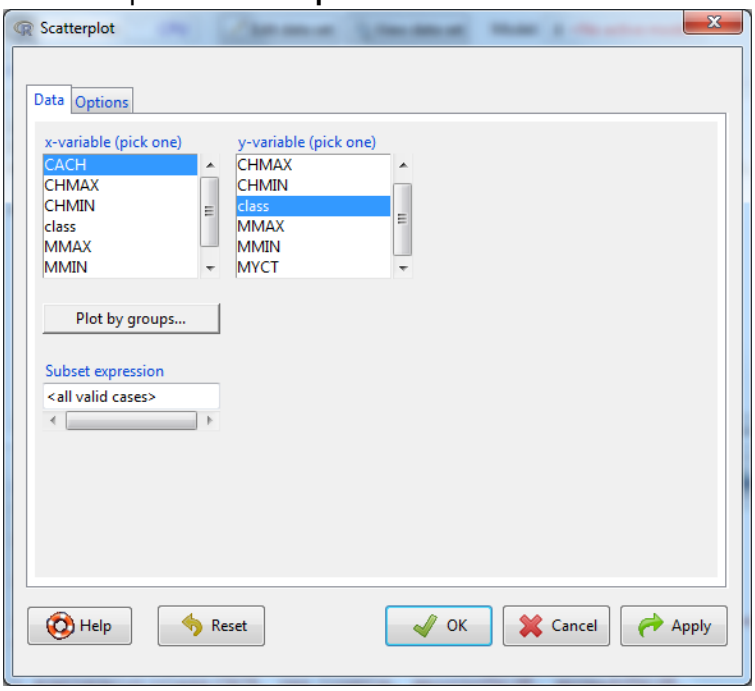
คำถาม: ตัวแปรทุกตัวมีการแจกแจงแบบปกติหรือไม่?

ตัว predictor แต่ละตัวมีความสัมพันธ์เชิงเส้นกับตัวแปร class หรือไม่?

3. ทดสอบการแจกแจงปกติ สมมติว่าทุกตัวมีการแจกแจงแบบปกติ

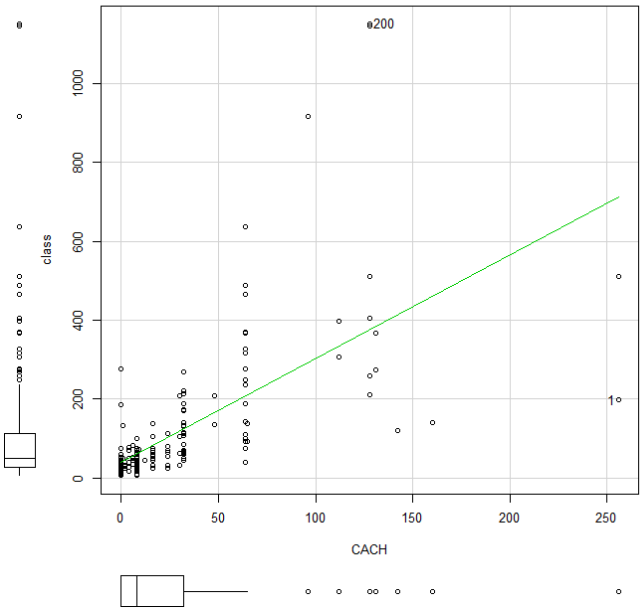
4. ทดสอบว่าตัวแปร class มีความสัมพันธ์กับตัว predictor แต่ละตัวแบบเชิงเส้น

คำสั่ง: Graphs > **Scratterplot**



Script & Output

```
> showData(CPU, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80,
+   maxheight=30)

> scatterplot(class~CACH, reg.line=lm, smooth=FALSE, spread=FALSE,
+   id.method='mahal', id.n = 2, boxplots='xy', span=0.5, data=CPU)
  1 200
  1 200
```



5. สมมติตัวแปร predictor ทุกตัวมีความสัมพันธ์แบบเชิงเส้นกับ class

หาความสัมพันธ์ระหว่าง predictors แต่ละตัวกับ class

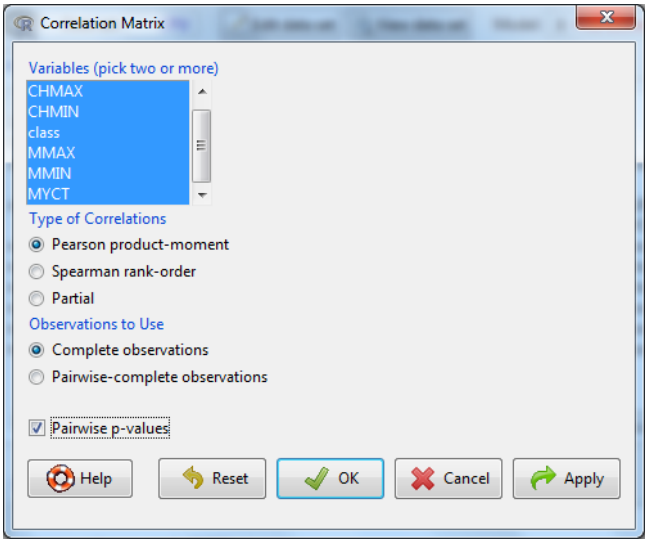
Predictor	r	t, df	p-value
CASH	0.6626414	12.7297, 207	< 2.2e-16
CHMAX	0.6052093	10.9381, 207	< 2.2e-16
CHMIN	0.6089033	11.044, 207	< 2.2e-16
MMAX	0.8630041	24.5775, 207	< 2.2e-16
MMIN	0.7949313	18.8513, 207	< 2.2e-16
MYCT	-0.3070994	-4.6427, 207	= 6.102e-06

หมายเหตุ $t = \frac{r * \sqrt{N-2}}{1-r^2}$

โดยปกติแล้วตัว predictors ก็จะมีความสัมพันธ์กันเอง

ลองหาความสัมพันธ์ของตัวแปรทั้งหมดด้วย

คำสั่ง: Statistics > Summaries > **Correlation matrix** เลือก **Pairwise p-value**

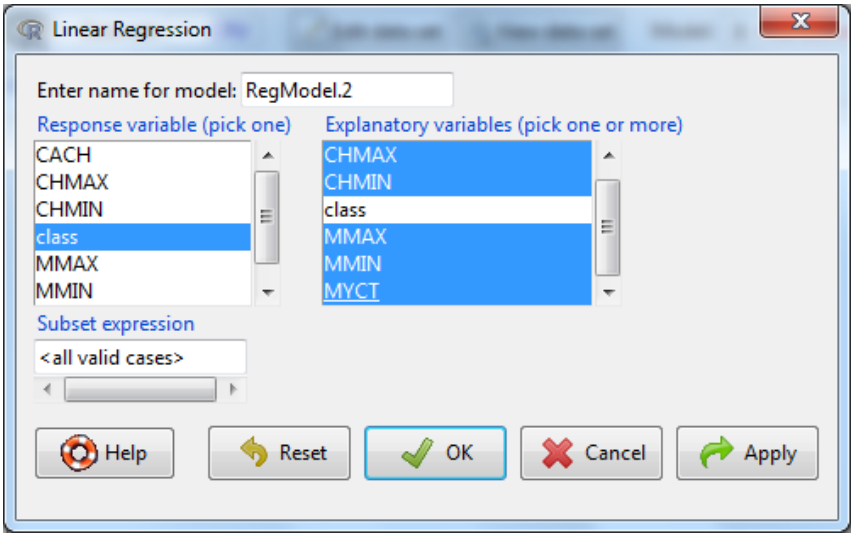


Script & Output

```
> rcorr.adjust(CPU[,c("CACH","CHMAX","CHMIN","class","MMAX","M
+   type="pearson", use="complete")

 Pearson correlations:
         CACH   CHMAX   CHMIN   class    MMAX    MMIN    MYCT
CACH   1.0000  0.4878  0.5822  0.6626  0.5380  0.5347 -0.3210
CHMAX  0.4878  1.0000  0.5483  0.6052  0.5272  0.2669 -0.2505
CHMIN  0.5822  0.5483  1.0000  0.6089  0.5605  0.5172 -0.3011
class  0.6626  0.6052  0.6089  1.0000  0.8630  0.7949 -0.3071
MMAX   0.5380  0.5272  0.5605  0.8630  1.0000  0.7582 -0.3786
MMIN   0.5347  0.2669  0.5172  0.7949  0.7582  1.0000 -0.3356
MYCT  -0.3210 -0.2505 -0.3011 -0.3071 -0.3786 -0.3356  1.0000

 Number of observations: 209

 Pairwise two-sided p-values:
      CACH   CHMAX  CHMIN  class  MMAX   MMIN   MYCT
CACH         <.0001 <.0001 <.0001 <.0001 <.0001 <.0001
CHMAX <.0001        <.0001 <.0001 <.0001 <.0001 0.0003
CHMIN <.0001 <.0001        <.0001 <.0001 <.0001 <.0001
class <.0001 <.0001 <.0001        <.0001 <.0001 <.0001
MMAX  <.0001 <.0001 <.0001 <.0001        <.0001 <.0001
MMIN  <.0001 <.0001 <.0001 <.0001 <.0001        <.0001
MYCT  <.0001 0.0003 <.0001 <.0001 <.0001 <.0001
```

6. หา Regression Model (Explanatory variables ทุกตัว ยกเว้น class ซึ่งเป็น Response variable)

คำสั่ง: Statistics > Fits models > **Linear Regression**

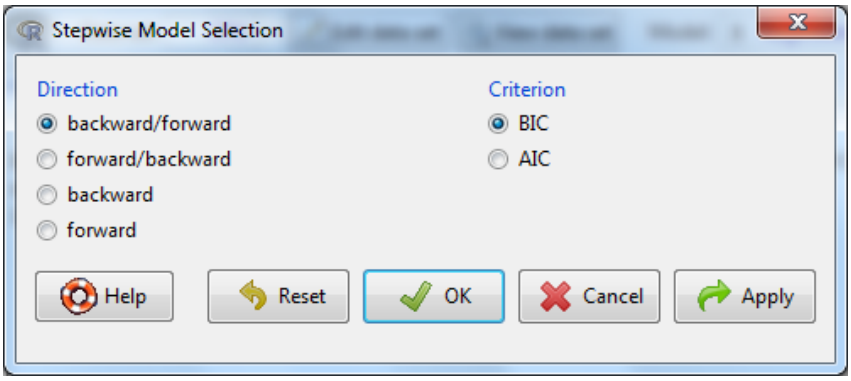


Script & Output

```
> RegModel.2 <- lm(class~CACH+CHMAX+CHMIN+MMAX+MMIN+MYCT, data=CPU)

> summary(RegModel.2)
```

```
Call:
lm(formula = class ~ CACH + CHMAX + CHMIN + MMAX + MMIN + MYCT,
    data = CPU)

Residuals:
    Min      1Q  Median      3Q     Max
-195.82  -25.17    5.40   26.52  385.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.589e+01  8.045e+00  -6.948 5.00e-11 ***
CACH         6.414e-01  1.396e-01   4.596 7.59e-06 ***
CHMAX        1.482e+00  2.200e-01   6.737 1.65e-10 ***
CHMIN       -2.704e-01  8.557e-01  -0.316   0.7524
MMAX         5.571e-03  6.418e-04   8.681 1.32e-15 ***
MMIN         1.529e-02  1.827e-03   8.371 9.42e-15 ***
MYCT         4.885e-02  1.752e-02   2.789   0.0058 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649, Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

คำสั่ง: Models > **Stepwise model selection**



Script & Output

```
> stepwise(RegModel.2, direction='backward/forward', criterion='BIC')

Direction:  backward/forward
Criterion:  BIC

Start:  AIC=1741.63
class ~ CACH + CHMAX + CHMIN + MMAX + MMIN + MYCT

        Df Sum of Sq    RSS    AIC
- CHMIN  1       359 727279 1736.4
<none>               726920 1741.6
- MYCT   1     27985 754905 1744.2
- CACH   1     76009 802929 1757.1
- CHMAX  1    163347 890267 1778.7
- MMIN   1    252179 979099 1798.5
- MMAX   1    271177 998097 1802.5
```

```
Step:  AIC=1736.39
class ~ CACH + CHMAX + MMAX + MMIN + MYCT

       Df Sum of Sq    RSS    AIC
<none>              727279 1736.4
- MYCT   1     28343 755623 1739.0
+ CHMIN  1       359 726920 1741.6
- CACH   1     78715 805995 1752.5
- CHMAX  1    177114 904393 1776.6
- MMIN   1    258252 985531 1794.6
- MMAX   1    270856 998135 1797.2

Call:
lm(formula = class ~ CACH + CHMAX + MMAX + MMIN + MYCT, data = CPU)

Coefficients:
(Intercept)         CACH        CHMAX         MMAX         MMIN         MYCT
 -56.075018     0.629824     1.459877     0.005562     0.015180     0.049113
```

Regression Model ที่ได้คือ

class = -56.075 + 0.630*CASH + 1.460*CHMAX + 0.006*MMAX + 0.015*MMIN + 0.049*MYCT

Model นี้มีความแม่นยำเท่าใด? Multuple R-squared = 0.8649?

ลองทดสอบ Model ในส่วนอื่น

สรุปคำสั่ง R (version 3.1.2)	
การจัดการข้อมูล	
การอ่านข้อมูลจากไฟล์ Excel	Data > **Import data > from Excel**
การสร้างตัวแปรตัวใหม่	Data > Manage variables in active data set > **Compute new variable**
การเปลี่ยนแปลงค่าตัวแปร	Data > Manages variables in active data set > **Recode variables**
การแปลงตัวแปร numeric ให้เป็น factor และการให้ชื่อค่าของตัวแปร (value label)	Data > Manages variables in active data set > **Convert numeric variables to factors**
การจัดเก็บไฟล์ข้อมูล R	Data > Active data set > **Save active data set**
การเปิดไฟล์ข้อมูล R	Data > **Load data set**
สถิติสำหรับตัวแปรเดียว	
ตัวแปรเชิงคุณลักษณะ	
การแจกแจงความถี่	Statistics > Summaries > **Frequency distributions**
กราฟแท่งและกราฟพาย	Graphs > **Bar graph / Pie Chart**
ตัวแปรเชิงปริมาณ	
ค่าสถิติเบื้องต้น (N, Mean และ SD)	Statistics > Summaries > **Numerical summaries**
กราฟฮีสโตแกรม	Graphs > **Histogram**
การแจกแจงความน่าจะเป็นแบบต่าง ๆ เช่น	Distributtion > Continuous distributions > Normal distribution >
การแจกแจงปกติ (Normal distribution)	**Plot normal distribution**
การหาค่าวิกฤต เมื่อกำหนดระดับนัยสำคัญ	**Normal quantiles**
การหาค่าพี (p-value) ของค่าสถิติ	**Normal probabilities**
การทดสอบการแจกแจงปกติ	Statistics > Summaries > **Shapio-Wilk test for normality**
การทดสอบค่าเฉลี่ย	Statistics > Mean > **Single-Sample t-test**
สถิติสำหรับสองตัวแปร	
การเปรียบเทียบค่าความแปรปรวน 2 ค่า	Statistics > Variances > **Two-variances F-test**
การเปรียบเทียบค่าความแปรปรวน > 2 ค่า	Statistics > Variances > **Bartlett's test** / **Levene's test**
การเปรียบเทียบค่าเฉลี่ย 2 ค่า (Independent Samples)	Statistics > Means > **Independent samples t-tes**
	Statistics > **Nonparametric tests > Two-sample Wilcoxon test**
การเปรียบเทียบค่าเฉลี่ย 2 ค่า (Paired Samples)	Statistics > **Means > Paired t-test**
	Statistics > **Nonparametric tests > Paired-samples Wilcoxon test**
การเปรียบเทียบค่าเฉลี่ย 2 ค่าหรือมากกว่า (Independent Samples)	Statistics > **Means > One-way ANOVA**
	Statistics > **Nonparametric tests** > **Kruskal-Wallis test**
การแจกแจงความถี่สองทางและการทดสอบความเป็นอิสระของตัวแปร 2 ตัว	Statistics > **Contingency tables > Two-way tables**
การหากราฟความสัมพันธ์ของ 2 ตัวแปร	Graphs > **Scratterplot**
การหาความสัมพันธ์ของ 2 ตัวแปร	Statistics > Summaries > **Correlation test** / **Correlation matrix**
การวิเคราะห์การถดถอย	Statistics > Fits models > **Linear Regression**
	Models > **Stepwise model selection**

การทำเหมืองข้อมูลด้วย Weka

(Data Mining with Weka)

สุณี รักษาเกียรติศักดิ์

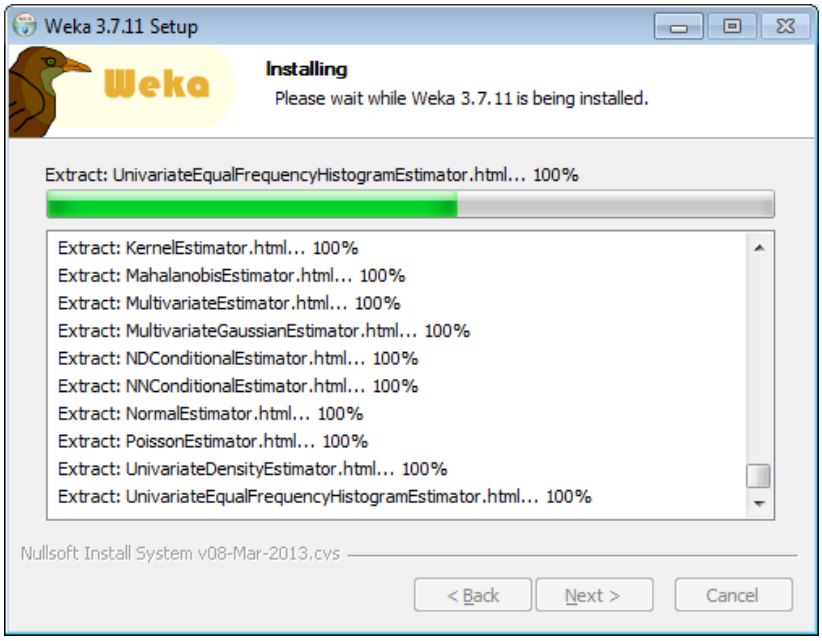
การติดตั้ง Weka

1. Download โปรแกรมที่ http://www.cs.waikato.ac.nz/ml/weka/downloading.html
2. ดับเบิลคลิกเพื่อติดตั้ง

   ในที่นี้จะใช้ version 3.7.11 (Book version 3.6)

   ติดตั้งตาม default และ Next

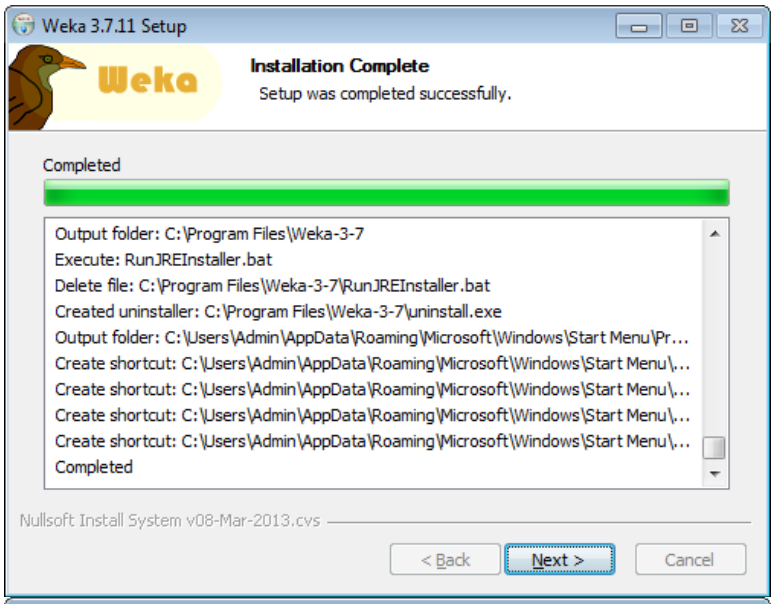
   จนถึง

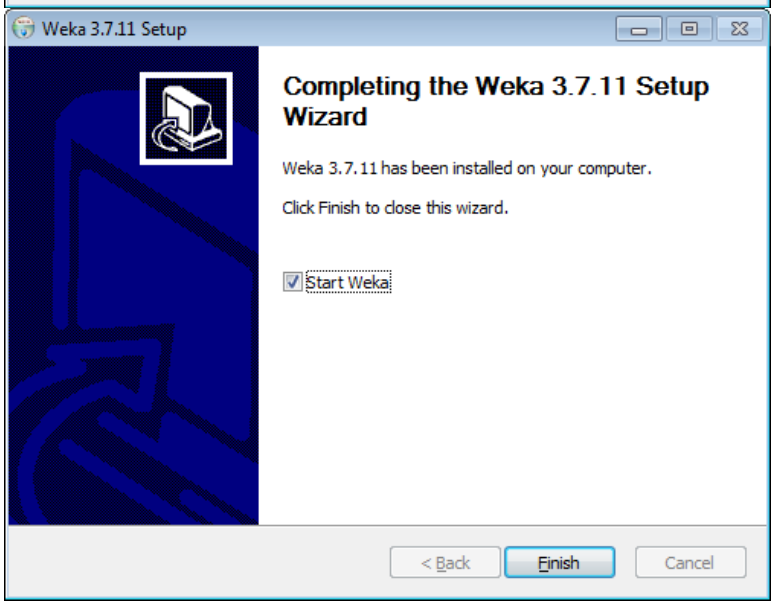


3. Install JRE

กดปุ่ม Install

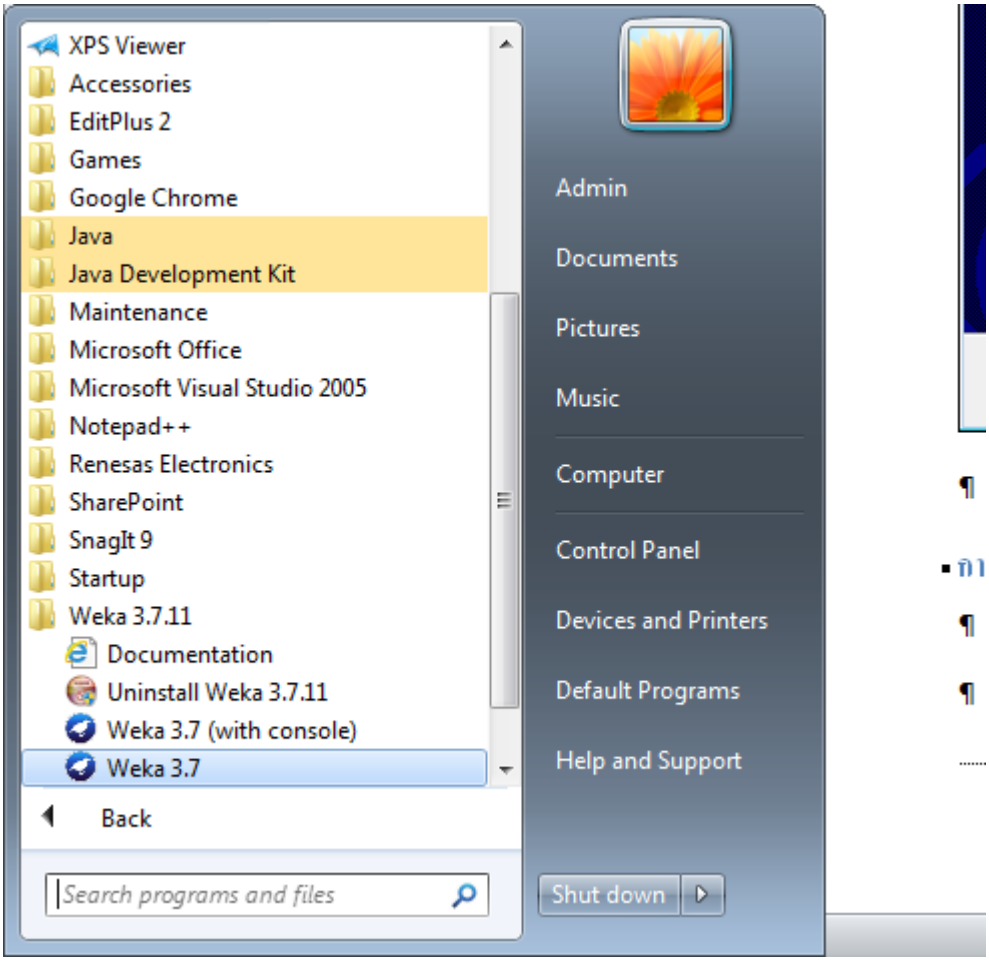
4. ติดตั้งเรียบร้อย





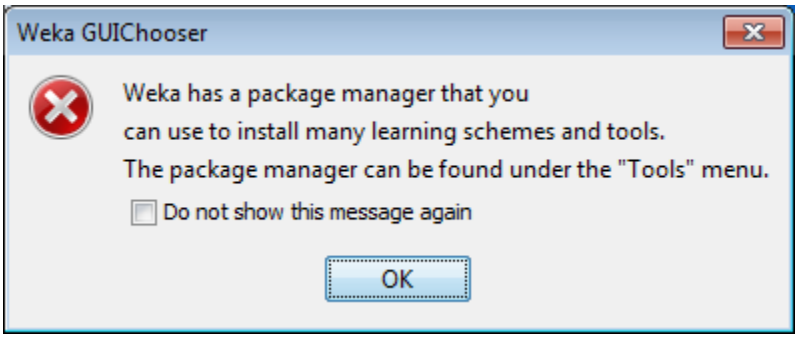
การเรียกใช้ Weka

1. All Programs > Weka 3.7



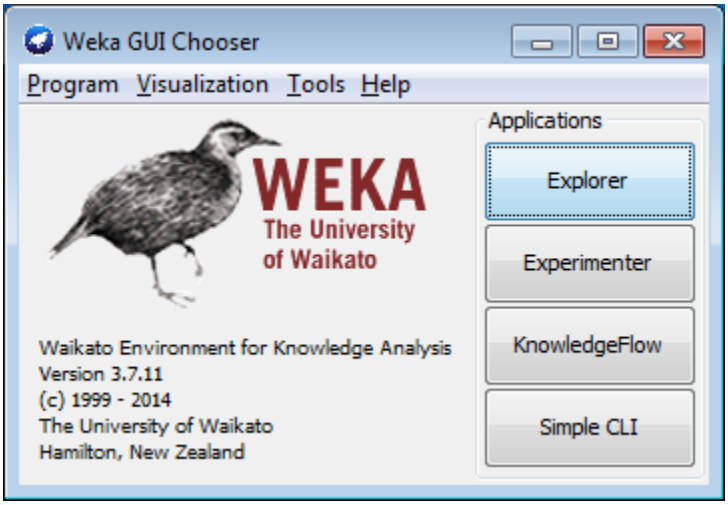
หรือ สำเนา short cut Weka 3.7 มาไว้บน desktop เพื่อสะดวกในการเรียกใช้

ดับเบิลคลิก Weka ไอคอน มาทำงาน จะเกิด



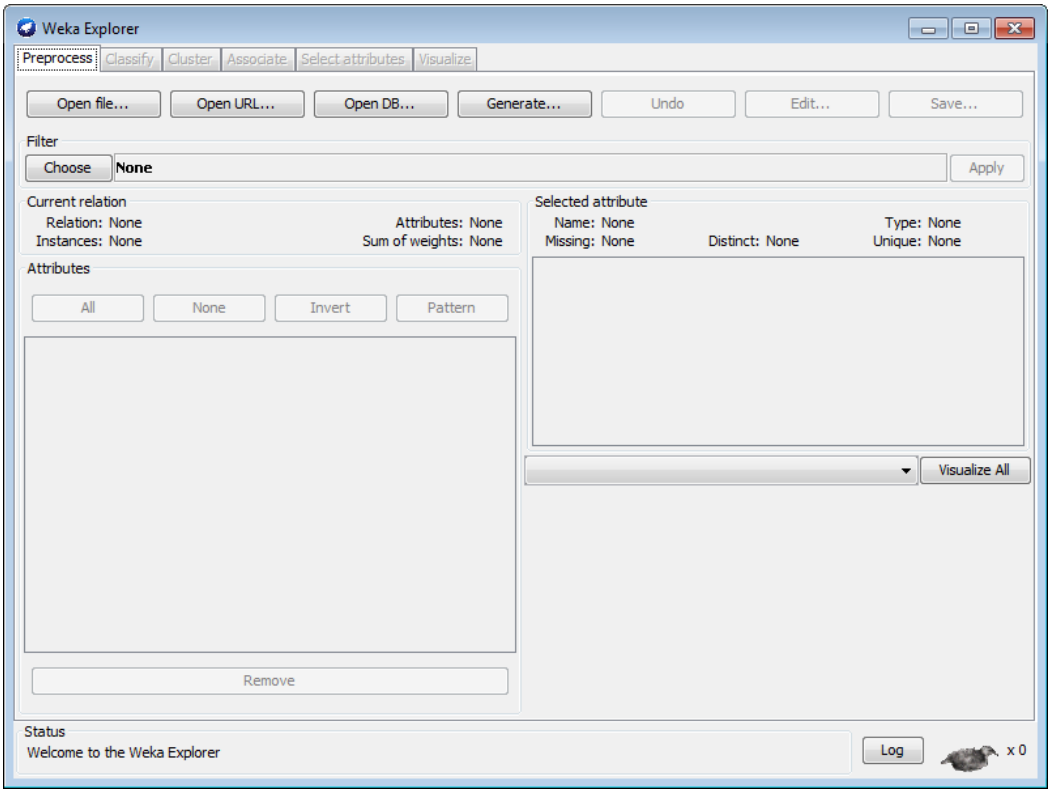
คลิก Do not show this message again แล้วคลิก OK

Weka GUI Chooser

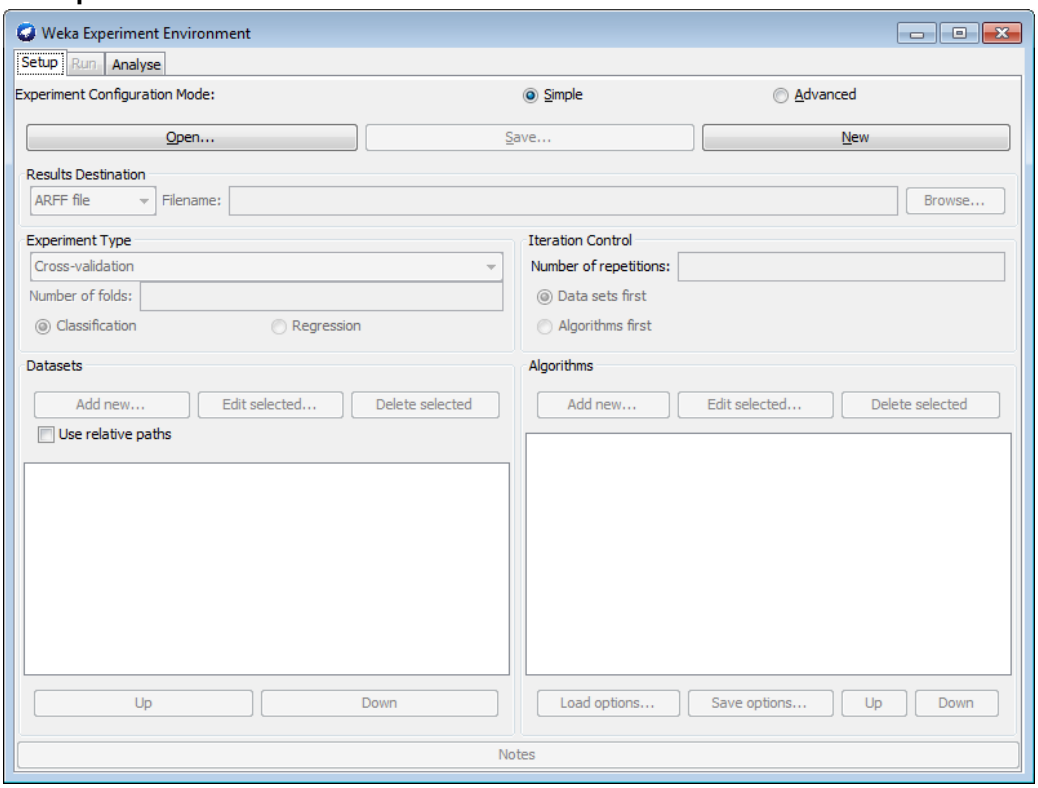


มี Interface การทำงาน 4 แบบ

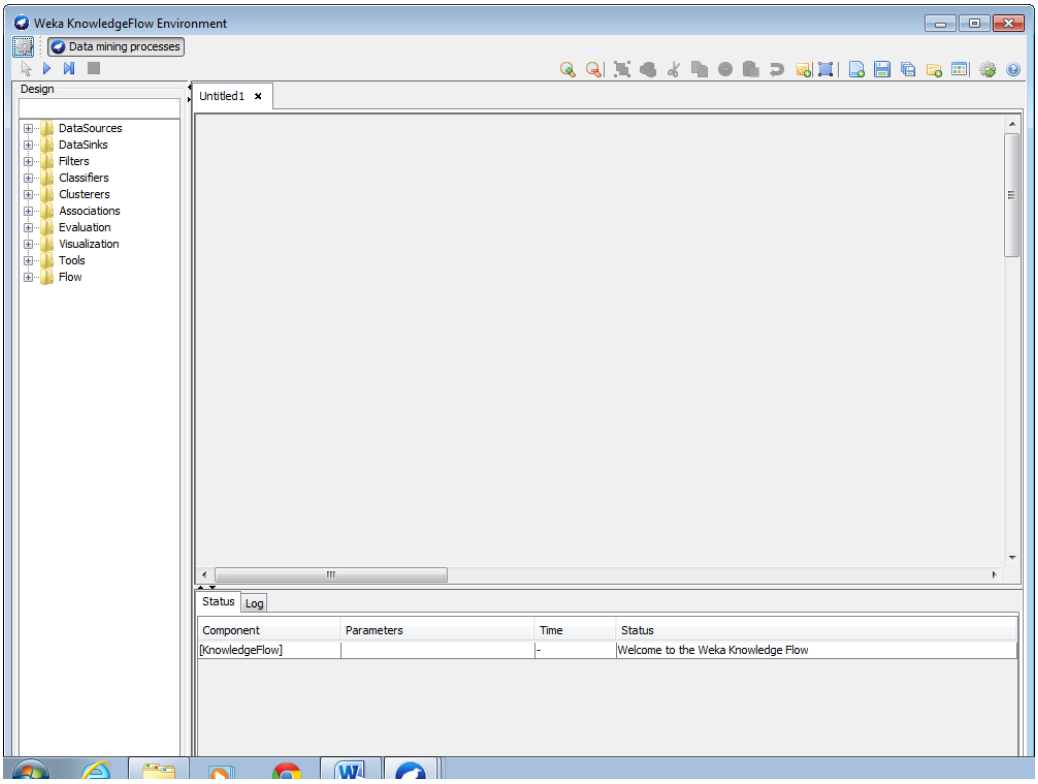
1. Explorer



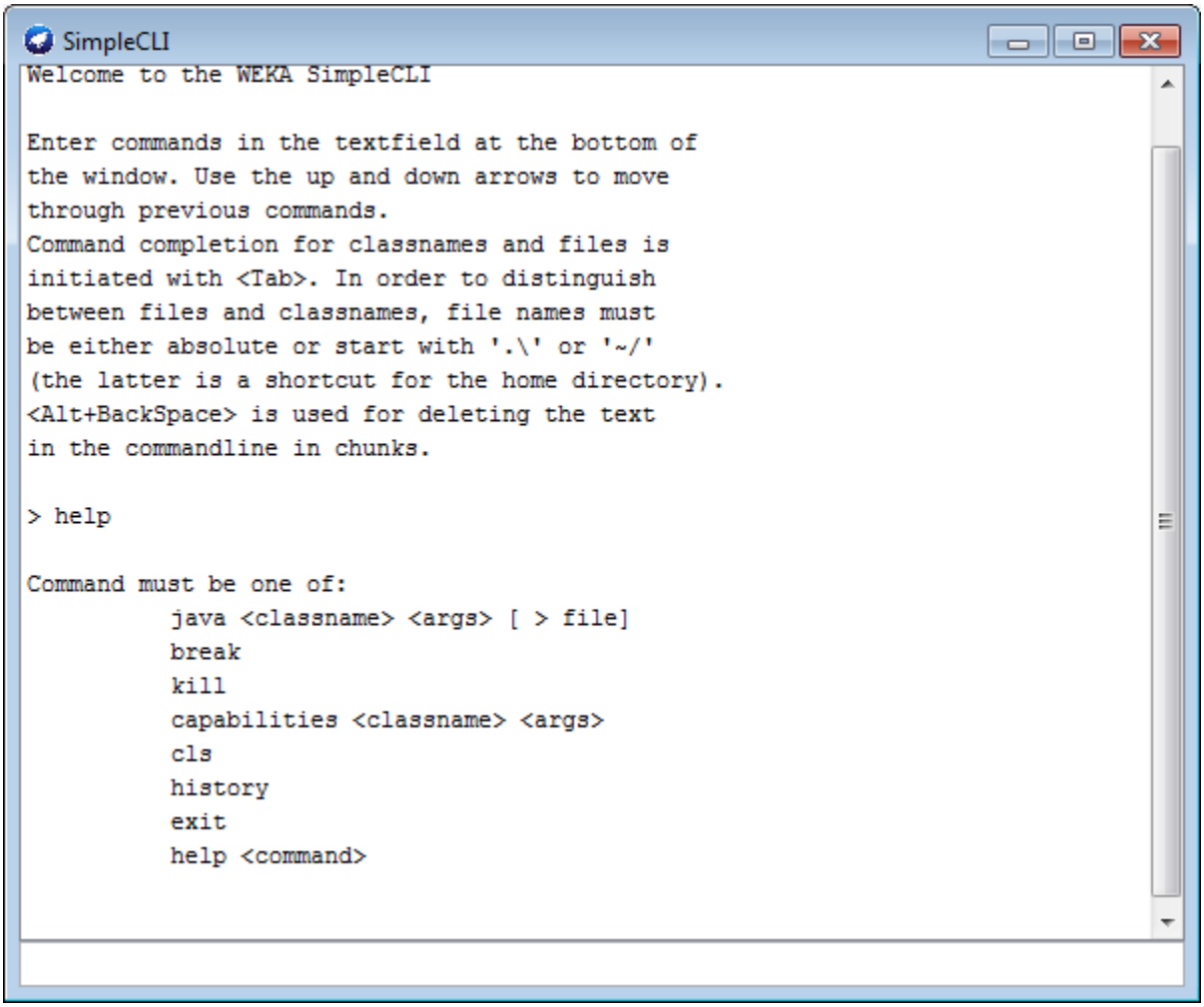
2. Experimenter



3. KnowledgeFlow



4. Simple CLI


SimpleCLI
Welcome to the WEKA SimpleCLI
Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.
> help
Command must be one of:
java <classname> <args> [> file]
break
kill
capabilities <classname> <args>
cls
history
exit
help <command>

Weka_Workshops

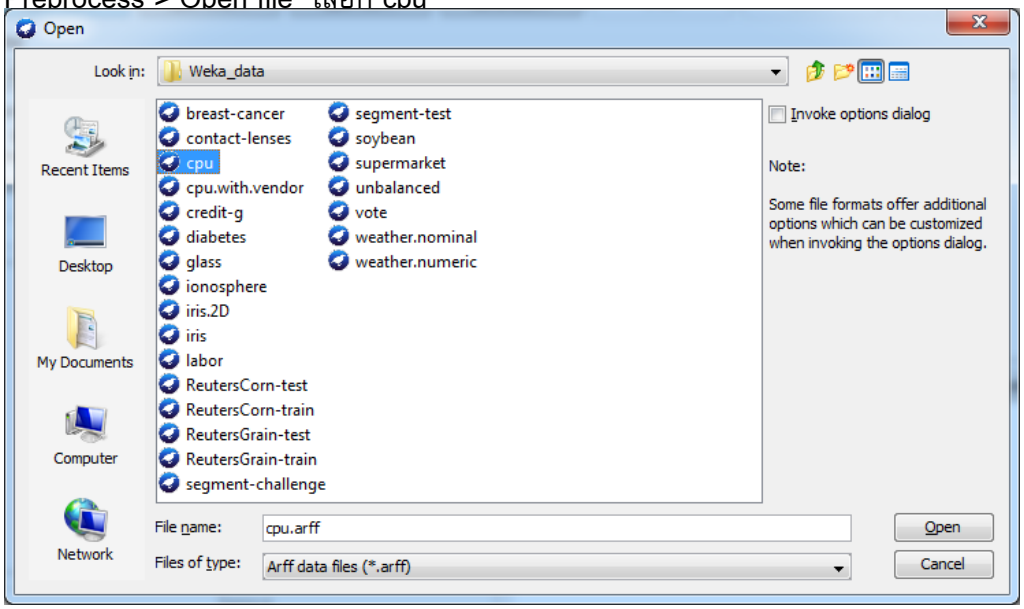
สุณี รักษาเกียรติศักดิ์

WS#1: Numeric Prediction

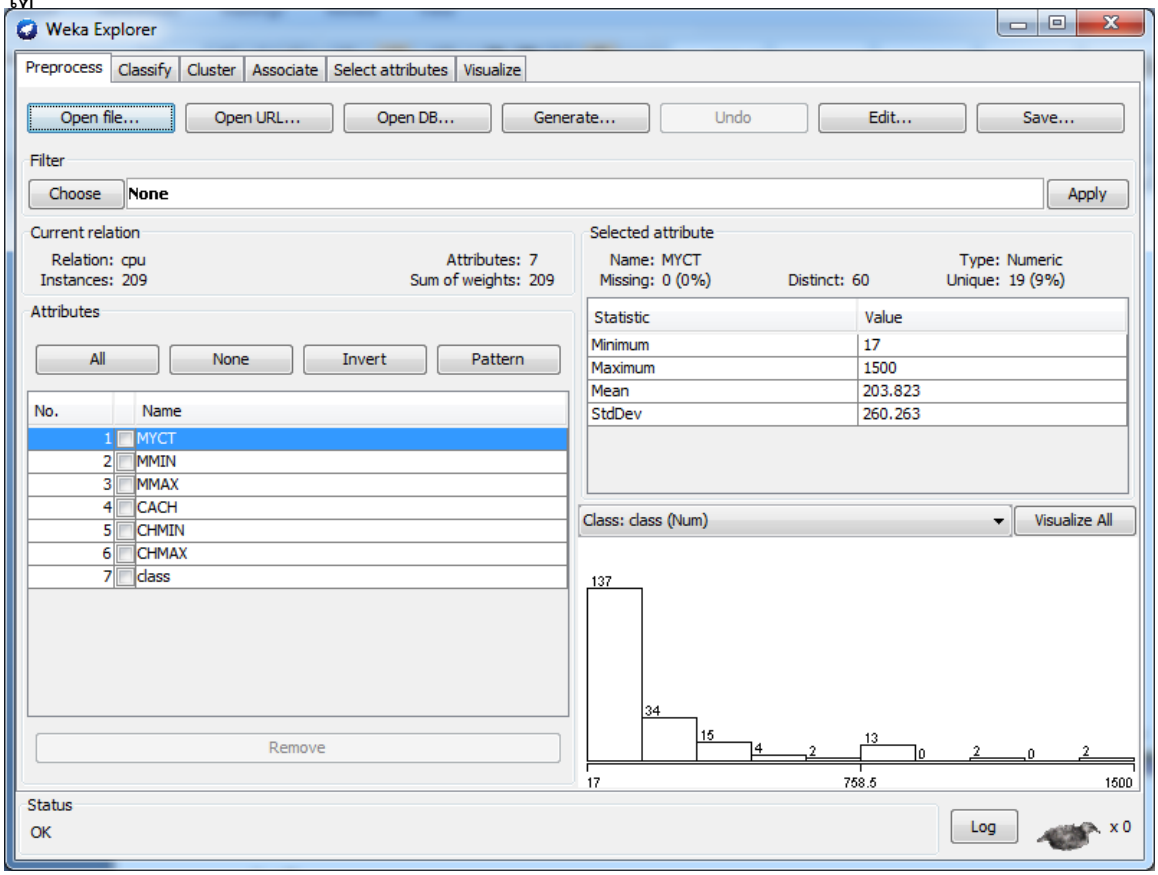
Data set: cpu (Weka_data จะมีไฟล์ format เป็น .arff)

```
%
% As used by Kilpatrick, D. & Cameron-Jones, M. (1998). Numeric prediction
% using instance-based learning with encoding length selection. In Progress
% in Connectionist-Based Information Systems. Singapore: Springer-Verlag.
%
% Deleted "vendor" attribute to make data consistent with with what we
% used in the data mining book.
%
@relation 'cpu'
@attribute MYCT numeric
@attribute MMIN numeric
@attribute MMAX numeric
@attribute CACH numeric
@attribute CHMIN numeric
@attribute CHMAX numeric
@attribute class numeric
@data
125,256,6000,256,16,128,198
29,8000,32000,32,8,32,269
29,8000,32000,32,8,32,220
29,8000,32000,32,8,32,172
29,8000,16000,32,8,16,132
26,8000,32000,64,8,32,318
```

1. Preprocess > Open file เลือก cpu



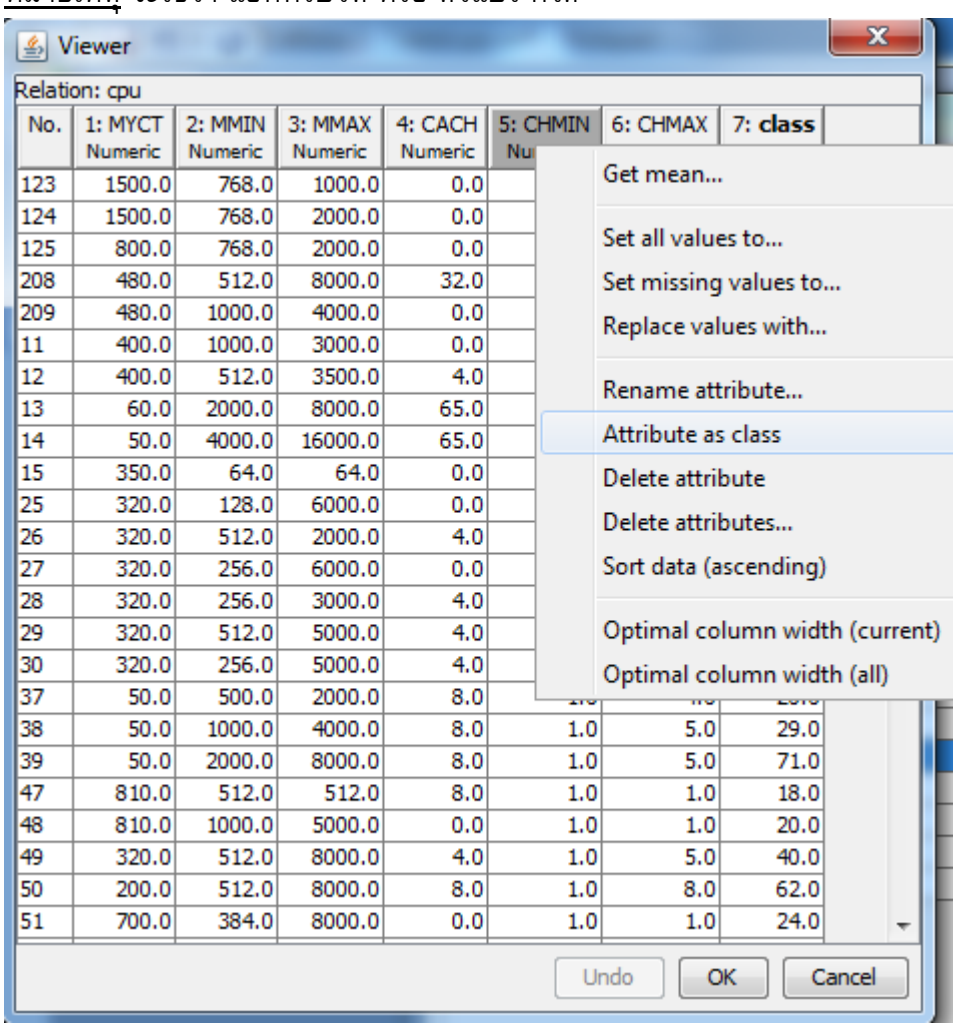
ได้



2. แอททริบิวต์ตัวสุดท้ายจะเป็นตัวแปร class หากต้องการให้ตัวแปรอื่นเป็นตัวแปร class ให้ Edit แล้วคลิกขวาตัวแปรที่ต้องการ เลือก Attribute as class ตัวแปรนั้นก็จะมาอยู่เป็นตัวสุดท้าย และแสดงเป็น

ตัวหนา

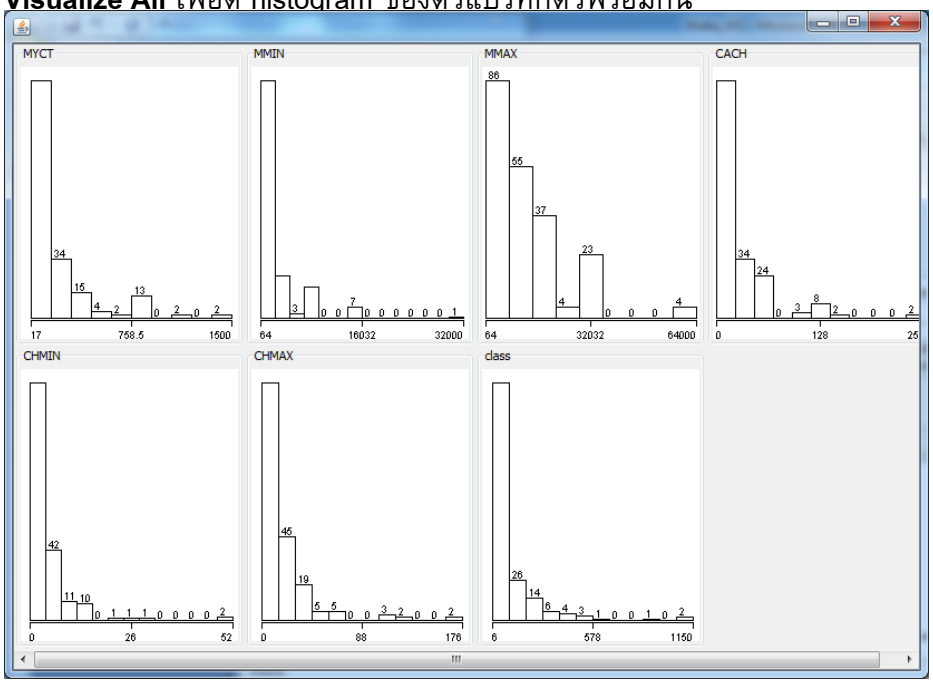
หมายเหตุ จะใช้ว่า แอททริบิวต์ หรือ ตัวแปร ก็ได้



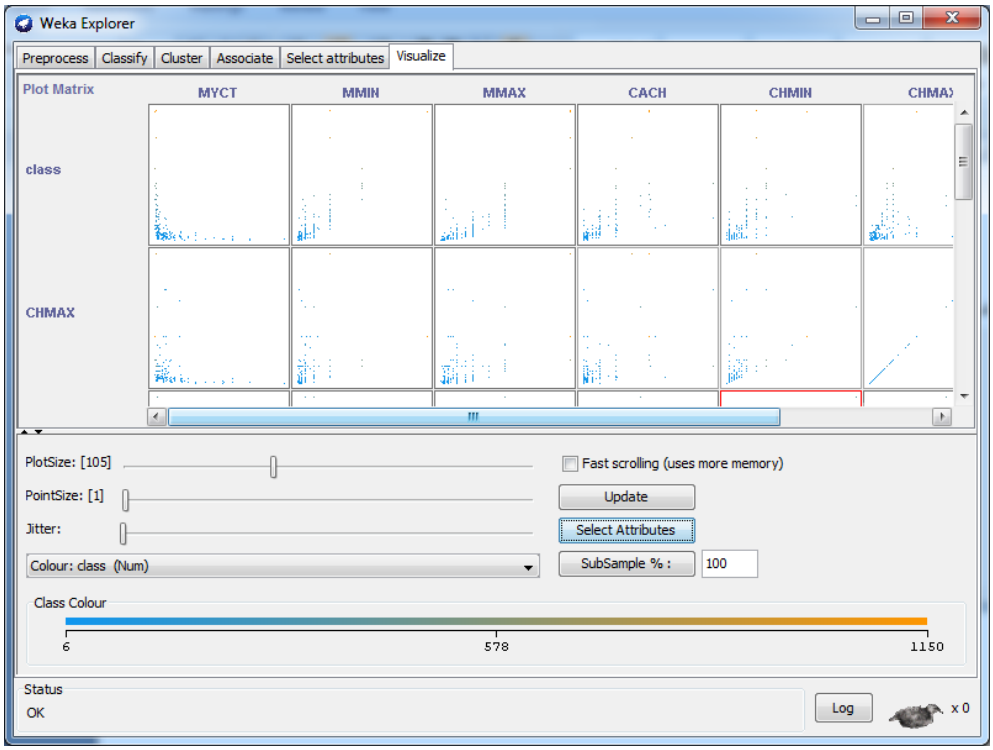
No.	1: MYCT Numeric	2: MMIN Numeric	3: MMAX Numeric	4: CACH Numeric	5: CHMIN Nu	6: CHMAX	7: class
123	1500.0	768.0	1000.0	0.0			
124	1500.0	768.0	2000.0	0.0			
125	800.0	768.0	2000.0	0.0			
208	480.0	512.0	8000.0	32.0			
209	480.0	1000.0	4000.0	0.0			
11	400.0	1000.0	3000.0	0.0			
12	400.0	512.0	3500.0	4.0			
13	60.0	2000.0	8000.0	65.0			
14	50.0	4000.0	16000.0	65.0			
15	350.0	64.0	64.0	0.0			
25	320.0	128.0	6000.0	0.0			
26	320.0	512.0	2000.0	4.0			
27	320.0	256.0	6000.0	0.0			
28	320.0	256.0	3000.0	4.0			
29	320.0	512.0	5000.0	4.0			
30	320.0	256.0	5000.0	4.0			
37	50.0	500.0	2000.0	8.0			
38	50.0	1000.0	4000.0	8.0	1.0	5.0	29.0
39	50.0	2000.0	8000.0	8.0	1.0	5.0	71.0
47	810.0	512.0	512.0	8.0	1.0	1.0	18.0
48	810.0	1000.0	5000.0	0.0	1.0	1.0	20.0
49	320.0	512.0	8000.0	4.0	1.0	5.0	40.0
50	200.0	512.0	8000.0	8.0	1.0	8.0	62.0
51	700.0	384.0	8000.0	0.0	1.0	1.0	24.0

ถ้าต้องการได้ไฟล์ลง Excel ก็ ctrl+A (Select All) แล้วไป paste ลงใน Excel แล้วเพิ่มบรรทัดแรกเป็นชื่อตัวแปร

3. Weka จะสรุปค่า descriptive statistics ของแต่ละตัวแปรให้ พร้อมแสดงกราฟ (histogram) หรือคลิก **Visualize All** เพื่อดู histogram ของตัวแปรทกตัวพร้อมกัน



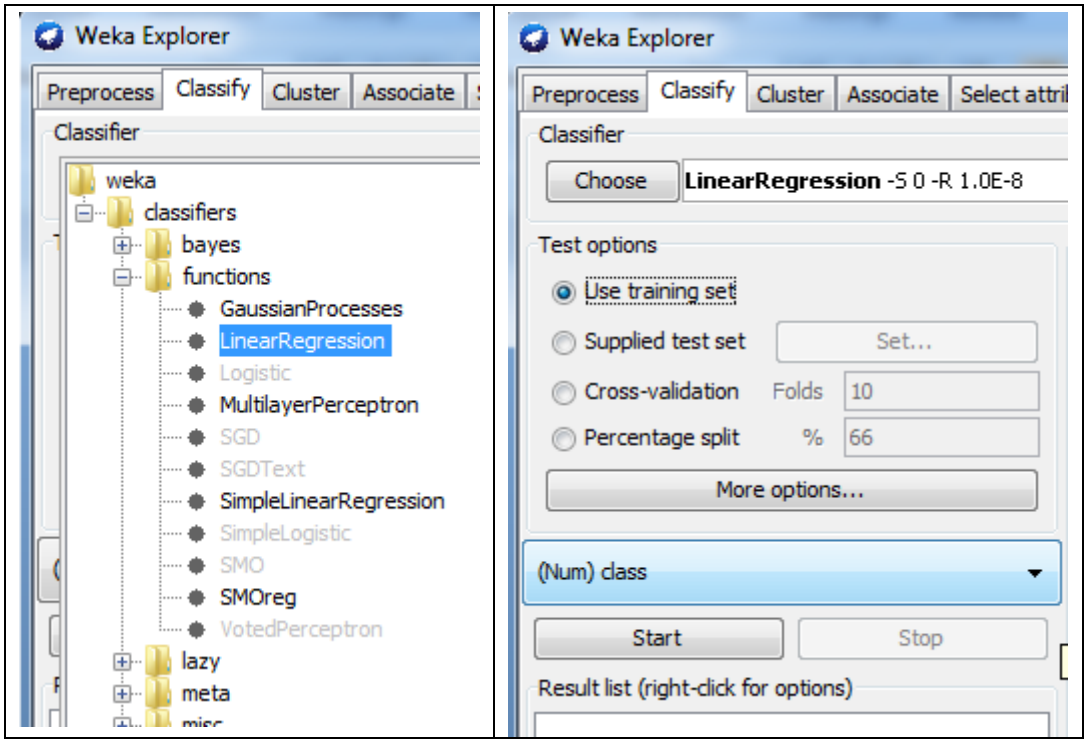
4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ **Visualize**



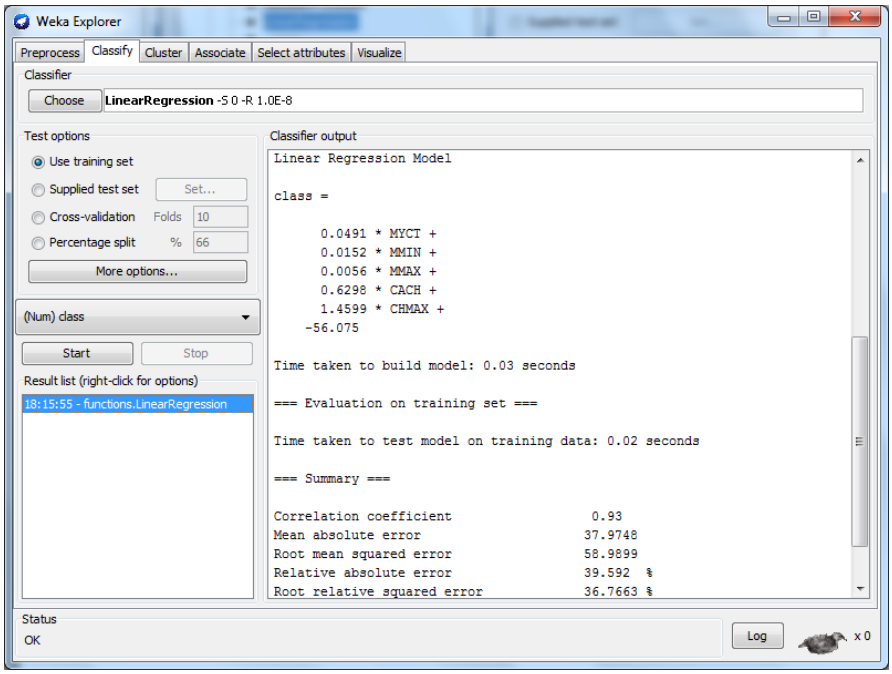
5. Fit linear regressiom model

Classify > Choose Classifier = functions > Linear Regression

Test options = Use training ser (default)



กด Start จะได้ผลลัพธ์



6. การแปลผล

6.1 Model: Class/PRP = -56.075 + 0.0491*MYCT + 0.0152*MMIN + 0.0056 *MMAX + 0.6298*CACH + 1.4599*CHMAX

ตัวแปรใดที่ไม่เข้าในโมเดล?

```
Linear Regression Model

class =

      0.0491 * MYCT +
      0.0152 * MMIN +
      0.0056 * MMAX +
      0.6298 * CACH +
      1.4599 * CHMAX +
    -56.075
```

6.2 Evaluation: Correlation coefficient R=0.93 or R^2=0.8649

ตัวแปร predictor 5 ตัว คือ MYCT, MMIN, MMAX, CACH, CHMAX สามารถอธิบายความแปรปรวนของตัวแปร Class/PRP ได้ประมาณ 81%

```
Correlation coefficient                  0.93
```

6.3 Deployment: เครื่องคอมพิวเตอร์เครื่องใหม่ 1 เครื่อง มีคุณสมบัติดังนี้ MYCT=200, MMIN=1000, MMAX=2000, CASH=0, CHMAX=64 จะมีประสิทธิภาพ (PRP) = 73.58

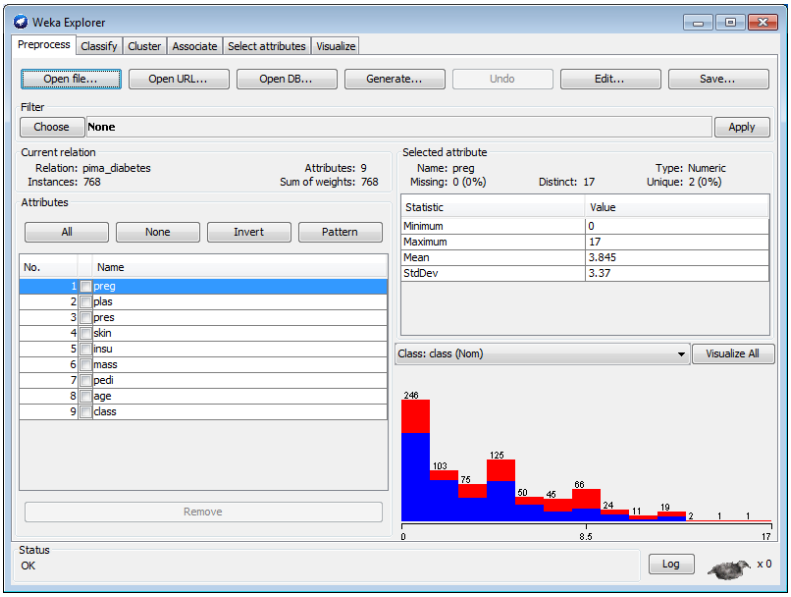
คลิกขวาที่ Result แล้ว Save model

7. ให้ทดสอบทั้ง 3 test options จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของโมเดลจะเหมือนกัน ซึ่งสอดคล้องกับ Stepwise Regression ใน SPSS และใน R แต่ผลลัพธ์ของ Evaluation จาก Correlation coefficient ของทั้ง 3 test mode คือ 1) Use training set, 2) Cross-validation (10 folds), 3) Percentage split (66%) จะไม่เท่ากัน

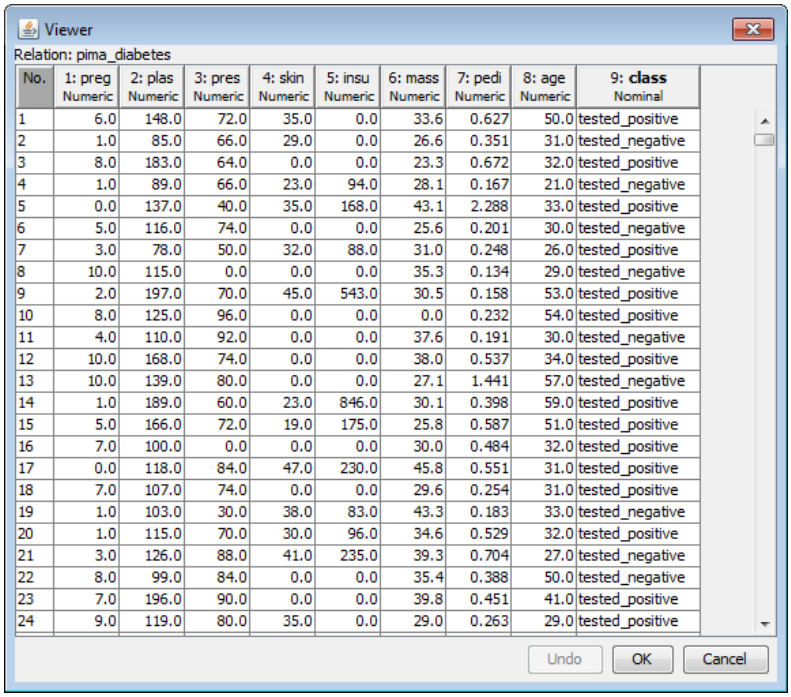
WS#2: Classification by Logistic Regression

Data set: diabetes

1. Preprocess > Open file เลือก diabetes



2. Edit เพื่อดู type และ ค่าของแต่ละตัวแปร

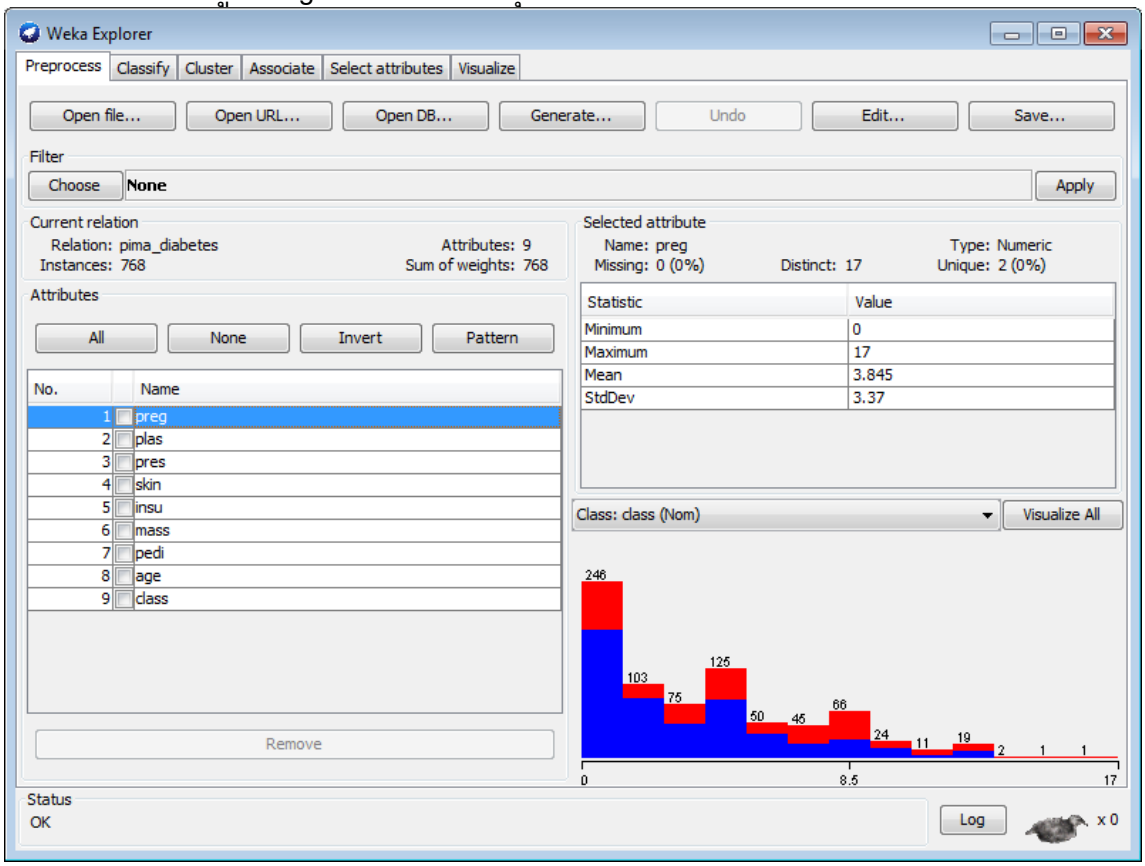


No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_negative
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_positive
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive

ตัวแปร class (dependent/target) มี type เป็น nominal มี 2 ค่า คือ tested negative/tested positive

ตัวแปร classifier (independent/predictor) มี 8 ตัว มี type เป็น numeric ทั้งหมด

3. ดู descriptive statistics ของแต่ละตัวแปรที่ Weka สรุปให้ พร้อมแสดงกราฟ (histogram) หรือคลิก Visualize All เพื่อดู histogram ของตัวแปรทุกตัวพร้อมกัน



ข้อมูลนี้มีทั้งหมด 768 ตัว (Instances/Cases/N)

ตัวแปร 1 ถึง 8 เป็นตัวแปร classifier มี type เป็น numeric ดังนั้น descriptive statistics จึงเป็น Minimum, Maximum, Mean, Standard Deviation (StdDev)

ตัวแปร class จะเป็นตัวแปรสุดท้ายเสมอ มี type เป็น nominal

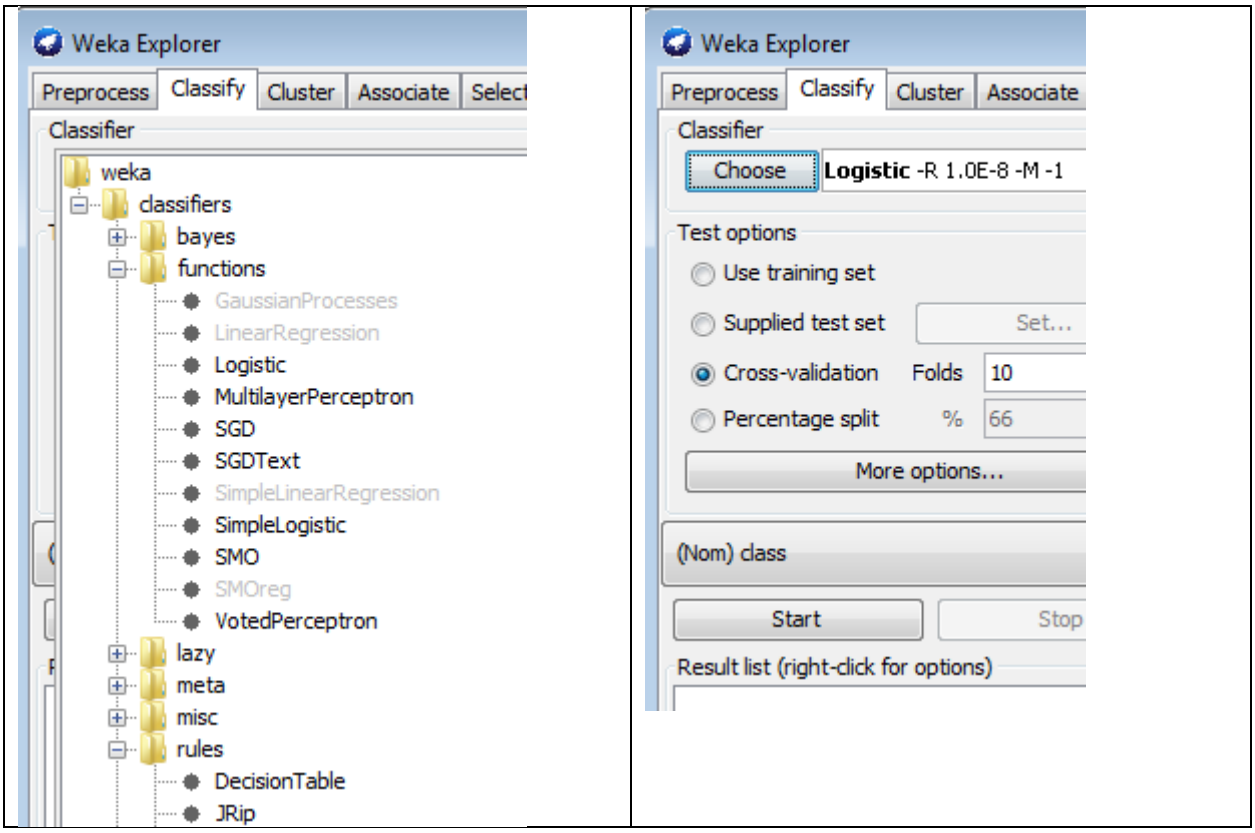
ลองคลิก Visualize All จะเห็นกราฟความสัมพันธ์ของตัวแปร class กับตัวแปร classifier แต่ละตัว

4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ Visualize

5. Fit logistic regressiom model

เลือกแท็บ Classify > Choose Classifier = functions >Logistic (Regression)

Default Test option = Cross validation Fold 10



กด Start จะได้ผลลัพธ์

6. การแปลผล

6.1 **Model:** Class = $e^{-0.1232}$*preg + $e^{-0.0352}$*plas + $e^{0.0133}$*pres + $e^{-0.0006}$*skin + $e^{0.0012}$*insu +
$e^{-0.0897}$*mass + $e^{-0.9452}$*pedi + $e^{-0.0149}$*age + 8.4047 หรือ

Class = 0.8841*preg + 0.9654*plas + 1.0134*pres + 0.9994*skin + 1.0012*insu +
0.9142*mass + 0.3886*pedi + 0.9852*age

Classifier output

```
Logistic Regression with ridge
Coefficients...
                          Class
Variable     tested_negative
============================
preg                   -0.1232
plas                   -0.0352
pres                    0.0133
skin                   -0.0006
insu                    0.0012
mass                   -0.0897
pedi                   -0.9452
age                    -0.0149
Intercept               8.4047
```

```
Odds Ratios...
                          Class
Variable     tested_negative
============================
preg                   0.8841
plas                   0.9654
pres                   1.0134
skin                   0.9994
insu                   1.0012
mass                   0.9142
pedi                   0.3886
age                    0.9852
```

6.2 Evaluation:

Correctly Classified Instances	593	77.2135 %
Incorrectly Classified Instances	175	22.7865 %

6.3 **Deployment:** แทนค่าลงในสมการข้อ 6.1

คลิกขวาที่ Result แล้ว Save model

7. ให้ทดสอบทั้ง 2 test options (Use training set, Cross-validation Folds 10 จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของโมเดลจะเหมือนกัน แต่ accuracy จะไม่เหมือนกัน ในวิธีการ Cross-validation Folds 10 Weka จะแบ่งข้อมูลทั้งหมดออกเป็น 10 ส่วน และรัน 10 ครั้ง โดยครั้งที่ 1 ใช้ ส่วนที่ 1 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set
   ครั้งที่ 2 ใช้ ส่วนที่ 2 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set
   จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set
   โดย evaluation accuracy จะเป็นค่าเฉลี่ยของทั้ง 10 ครั้ง
   จากนั้นจะรันโมเดลอีกครั้งโดยใช้ full training set ทั้งหมดในการรันโมเดล

WS#3: Classification by ML algorithm

Data set: diabetes

1. Preprocess > Open file เลือก diabetes
2. Fit classification model Classifier = rules > **ZeroR** (baseline accuracy)

   การแปลผล

   Model: Class = tested_negative

```
=== Classifier model (full training set) ===

ZeroR predicts class value: tested_negative
```

   Evaluation:

```
Correctly Classified Instances         500               65.1042 %
Incorrectly Classified Instances       268               34.8958 %
=== Confusion Matrix ===

   a   b   <-- classified as
 500   0 |   a = tested_negative
 268   0 |   b = tested_positive
```

   Deployment: tested negative

3. Fit classification model Classifier = rules > OneR

   การแปลผล

   Model: ไม่ make sence

```
=== Classifier model (full training set) ===

plas:
        < 114.5 -> tested_negative
        < 115.5 -> tested_positive
        < 127.5 -> tested_negative
        < 128.5 -> tested_positive
        < 133.5 -> tested_negative
        < 135.5 -> tested_positive
        < 143.5 -> tested_negative
        < 152.5 -> tested_positive
        < 154.5 -> tested_negative
        >= 154.5        -> tested_positive
(587/768 instances correct)
```

   Evaluation: (cross-validation)

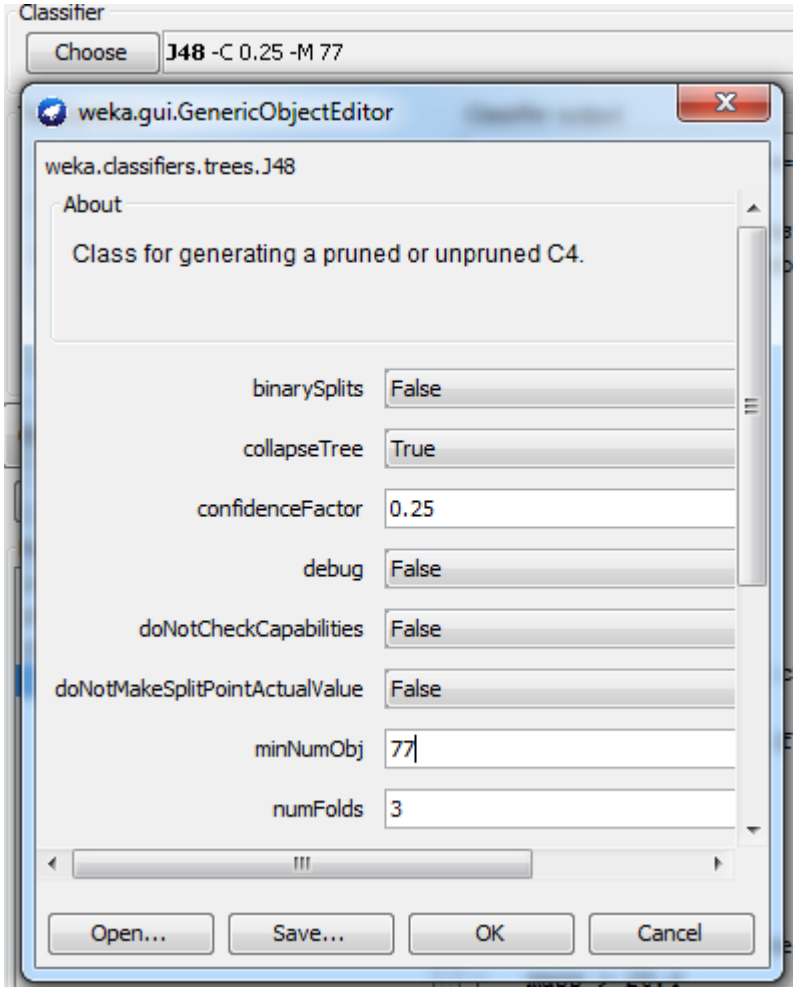
```
Correctly Classified Instances         549               71.4844 %
Incorrectly Classified Instances       219               28.5156 %
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 433  67 |   a = tested_negative
 152 116 |   b = tested_positive
```

Deployment: ใช้กฎ (apply rule)

4. Fit classification model Classifier = trees > J48

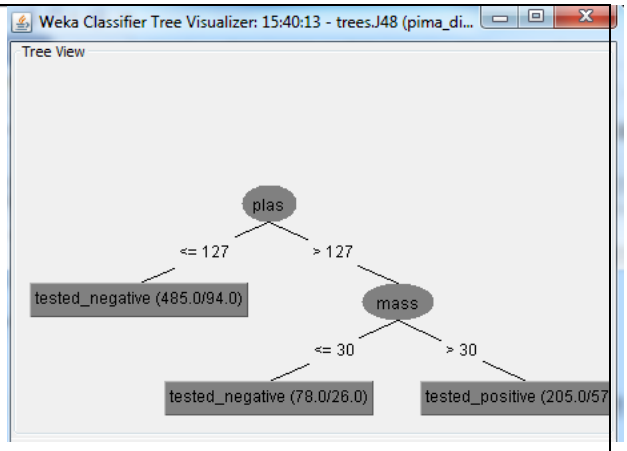
   <u>หมายเหตุ</u> ตาม default parameter จะได้จะได้ 20 กฎ ซึ่งบางกฎ ก็เป็นจริงสำหรับข้อมูลจำนวนน้อย ดังนั้นควรจะปรับ parameter “minNumObj” เช่นถ้าต้องการให้กฎเป็นจริงสำหรับข้อมูล 10% ก็ปรับ minNumObj = 77 โดยคลิกที่ **J48** แก้ minNumObj



การแปลผล

Model: คลิกขวาที่ Result เลือก Visualize tree

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

plas <= 127: tested_negative (485.0/94.0)
plas > 127
|   mass <= 30: tested_negative (78.0/26.0)
|   mass > 30: tested_positive (205.0/57.0)

Number of Leaves  :     3

Size of the tree :      5
```



Evaluation: (cross-validation)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         570               74.2188 %
Incorrectly Classified Instances       198               25.7813 %
=== Confusion Matrix ===

   a   b   <-- classified as
 445  55 |   a = tested_negative
 143 125 |   b = tested_positive
```

Deployment: คลิกขวาที่ Result แล้ว Save model

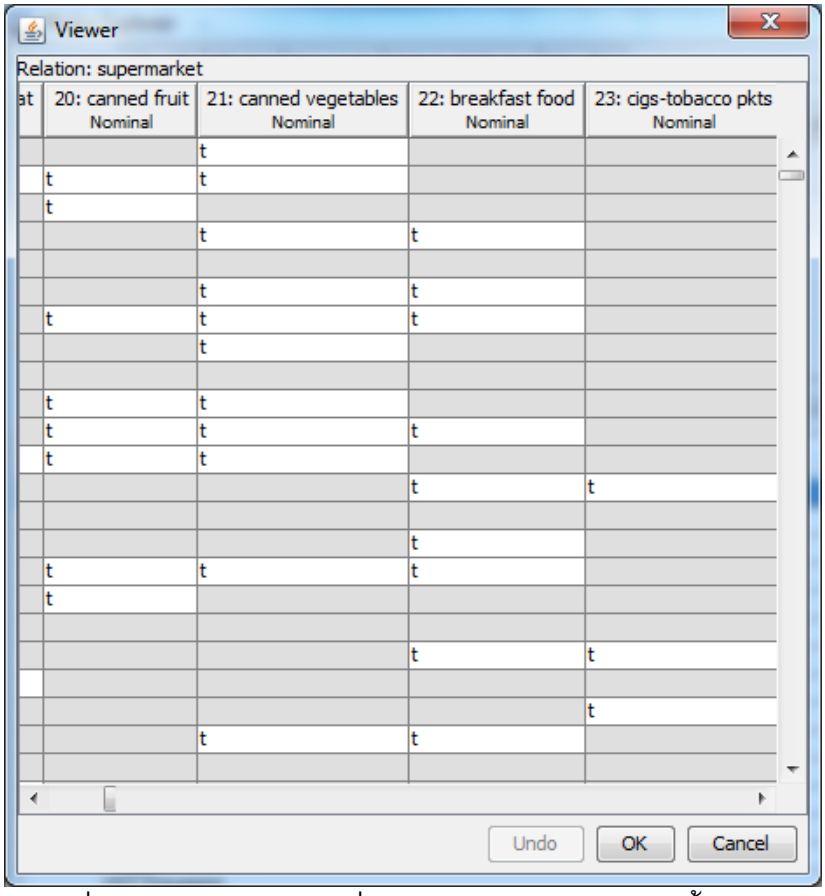
7. ให้เปรียบเทียบผลการทดลองด้วย Classifier ตัวอื่น ๆ เช่น rules > PART, rules > JRIP (จะดำเนินการใน Experimenter Interface)

WS#4: Association Problem Type

Data set: supermarket

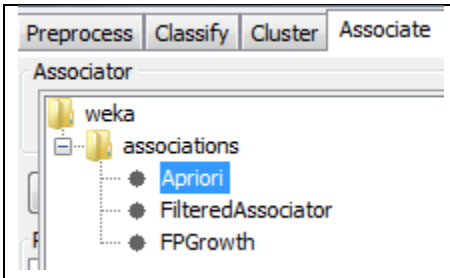
1. Preprocess > Open file เลือก supermarket

   Edit เพื่อดูการจัดเก็บข้อมูล



แถวหนึ่ง ๆ คือตะกร้าสินค้าหนึ่ง ๆ t (=true) มีสินค้าชนิดนั้น

2. Fit Association model โดยเลือกแท็บ Associate



เลือก Associator algorithm เป็น Apriori

คลิกที่ Apriori (รูปทางขวา)

ปรับ parameter ที่สำคัญ 3 ตัวคือ

-lowerBoundMinSupport = 0.2

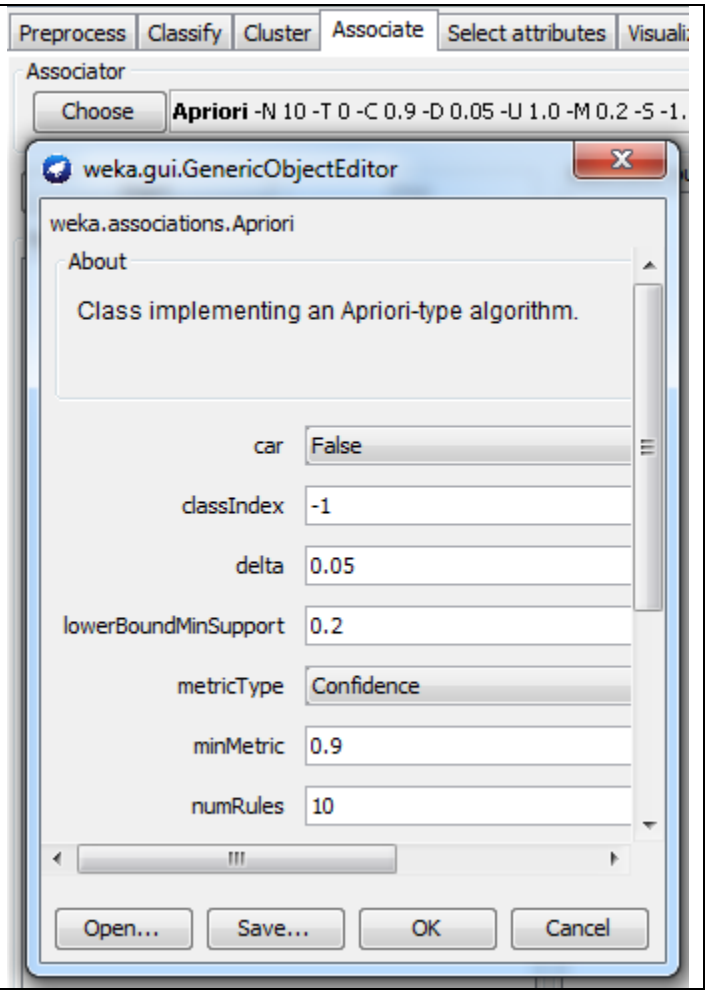
-minMetric (min Confidence) = 0.9

-numRules = 10

ผลลัพธ์: ไม่พบกฎใด ๆ เลย

ลองปรับ

-minMetric (min Confidence) = 0.85



การแปลผล

Model:

```
Best rules found:

 1. biscuits=t frozen foods=t fruit=t vegetables=t 1039 ==> bread and cake=t 929    <conf:(0.89)> lift:(1.24) lev:(0.
 2. fruit=t vegetables=t total=high 1050 ==> bread and cake=t 938    <conf:(0.89)> lift:(1.24) lev:(0.04) [182] conv:
 3. fruit=t total=high 1243 ==> bread and cake=t 1104    <conf:(0.89)> lift:(1.23) lev:(0.05) [209] conv:(2.49)
 4. biscuits=t total=high 1228 ==> bread and cake=t 1082    <conf:(0.88)> lift:(1.22) lev:(0.04) [198] conv:(2.34)
 5. milk-cream=t total=high 1217 ==> bread and cake=t 1071    <conf:(0.88)> lift:(1.22) lev:(0.04) [195] conv:(2.32)
 6. biscuits=t margarine=t vegetables=t 1054 ==> bread and cake=t 925    <conf:(0.88)> lift:(1.22) lev:(0.04) [166] c
 7. frozen foods=t total=high 1273 ==> bread and cake=t 1117    <conf:(0.88)> lift:(1.22) lev:(0.04) [200] conv:(2.27
 8. biscuits=t margarine=t fruit=t 1073 ==> bread and cake=t 938    <conf:(0.87)> lift:(1.21) lev:(0.04) [165] conv:
 9. party snack foods=t total=high 1120 ==> bread and cake=t 979    <conf:(0.87)> lift:(1.21) lev:(0.04) [172] conv:
10. vegetables=t total=high 1270 ==> bread and cake=t 1110    <conf:(0.87)> lift:(1.21) lev:(0.04) [195] conv:(2.21)
```

Evaluation: Min Support และ Confidence

Deployment: เอา กฎ ไปใช้

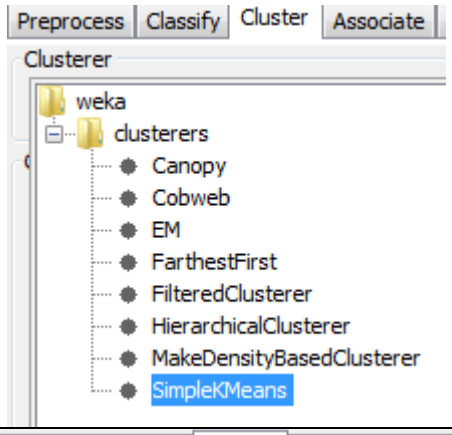
WS#5: Clustering

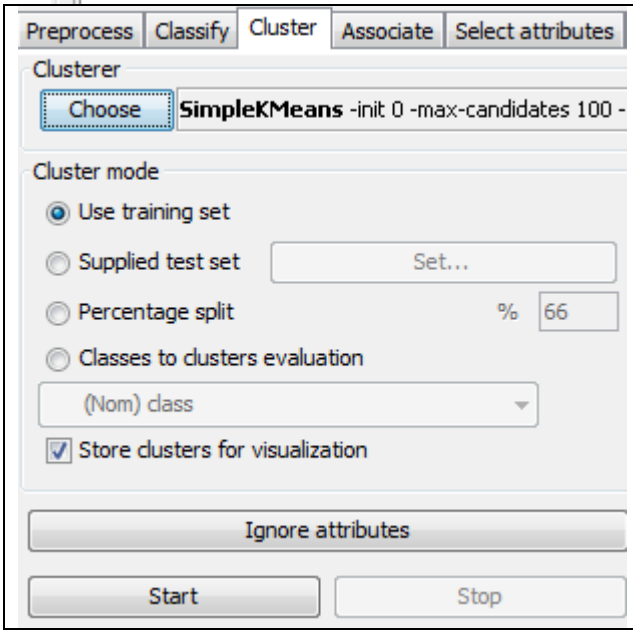
Data set: iris

1. Preprocess > Open file เลือก **iris**

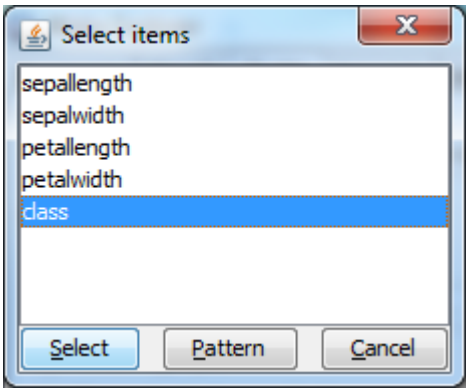
   ในที่นี้ iris มีตัวแปร class แต่เราจะไม่ใช้ในการ clustering แต่จะใช้ในการเปรียบเทียบกับค่าจากโมเดล

2. Fit Clustering model โดยเลือกแท็บ Cluster

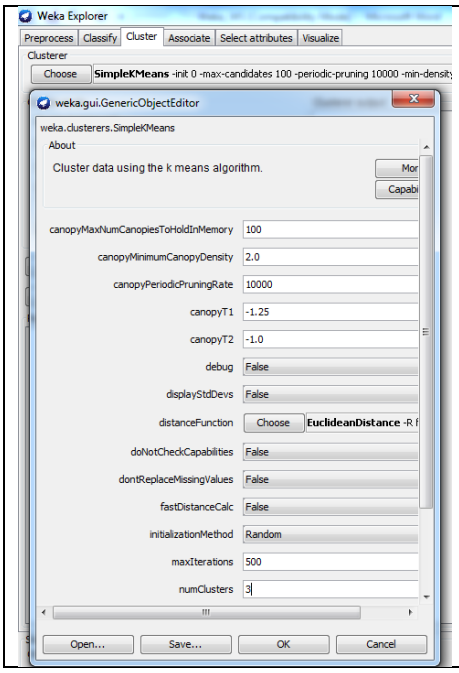




กด Ignore attributes



เลือก class แล้วกดปุ่ม Select



คลิก SimpleKMeans

ปรับ numCluster = 3

การแปลผล

Model:

```
Final cluster centroids:
                        Cluster#
Attribute      Full Data          0          1          2
                   (150)       (61)       (50)       (39)
=============================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795
```

Evaluation:

```
Within cluster sum of squared errors: 6.998114004826762
```

Deployment: คลิกขวาที่ Result แล้ว Save model

Witten: “Data mining is an experimental science”

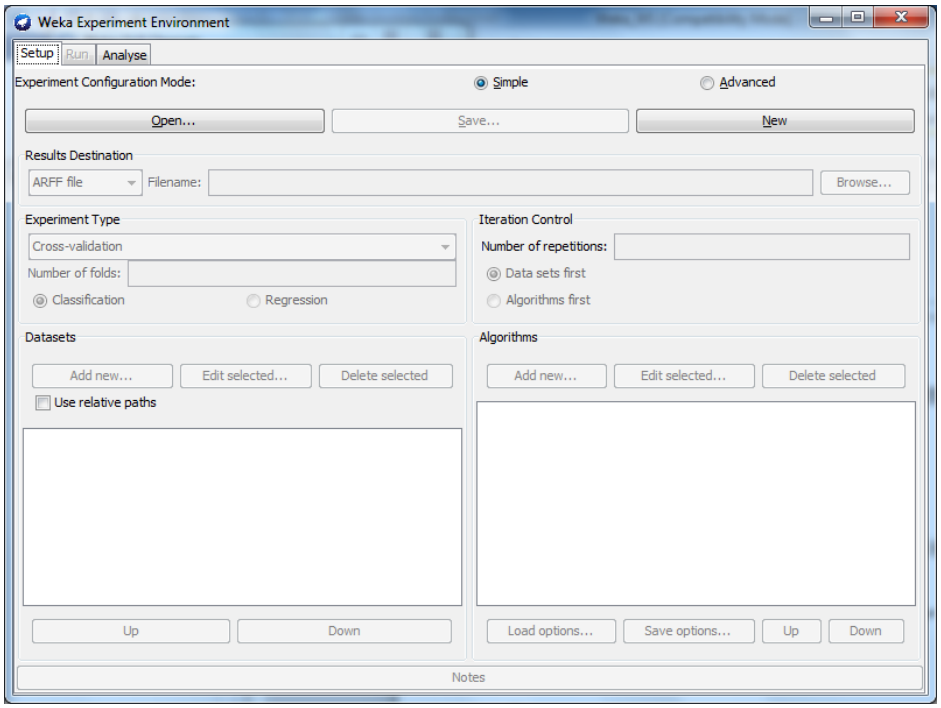
WS#6 จงเปรียบเทียบการวิเคราะห์โจทย์ Classification โดยใช้หลาย algoithms

Experiment 1

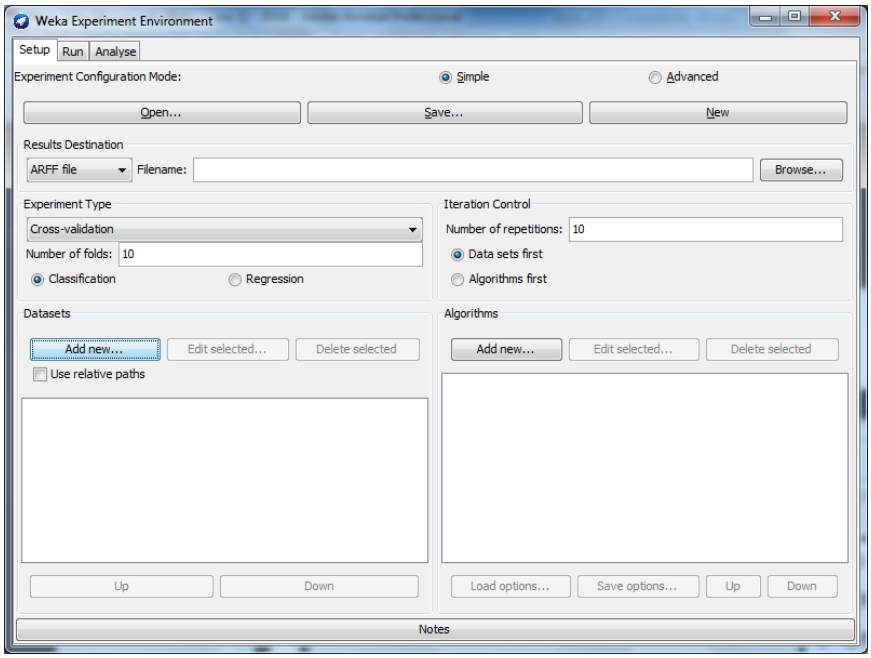
Data set: iris

Algoritms: trees > J48, rules > OneR, rules > ZeroR

1. Weka GUI Chooser เลือก Experimenter

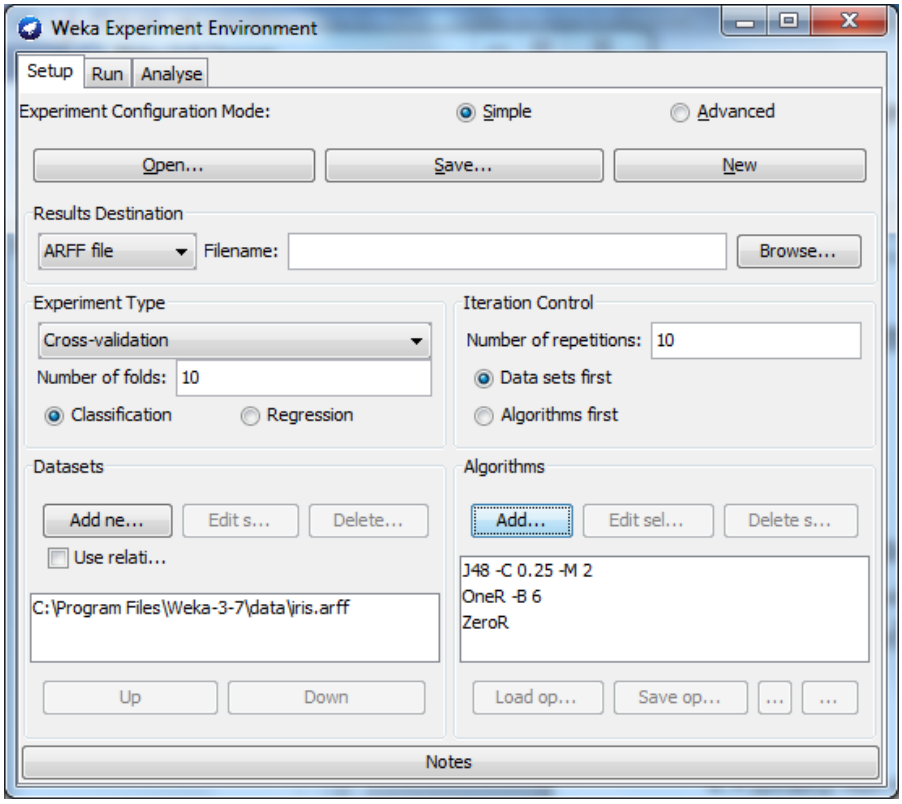


2. คลิก **New**

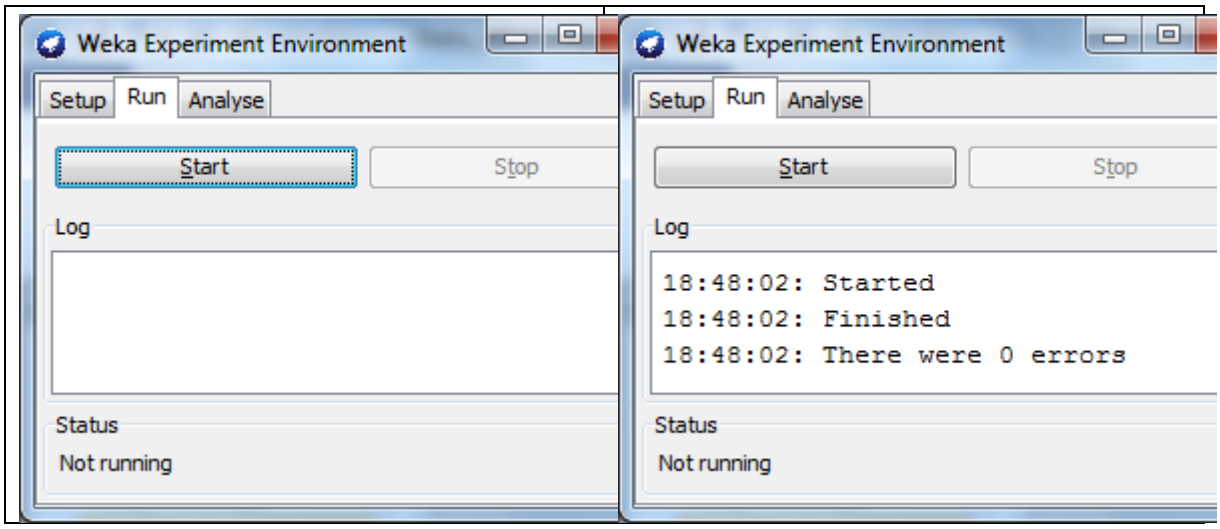


3. ที่ Datasets คลิก Add new แล้วเปิดเลือก iris

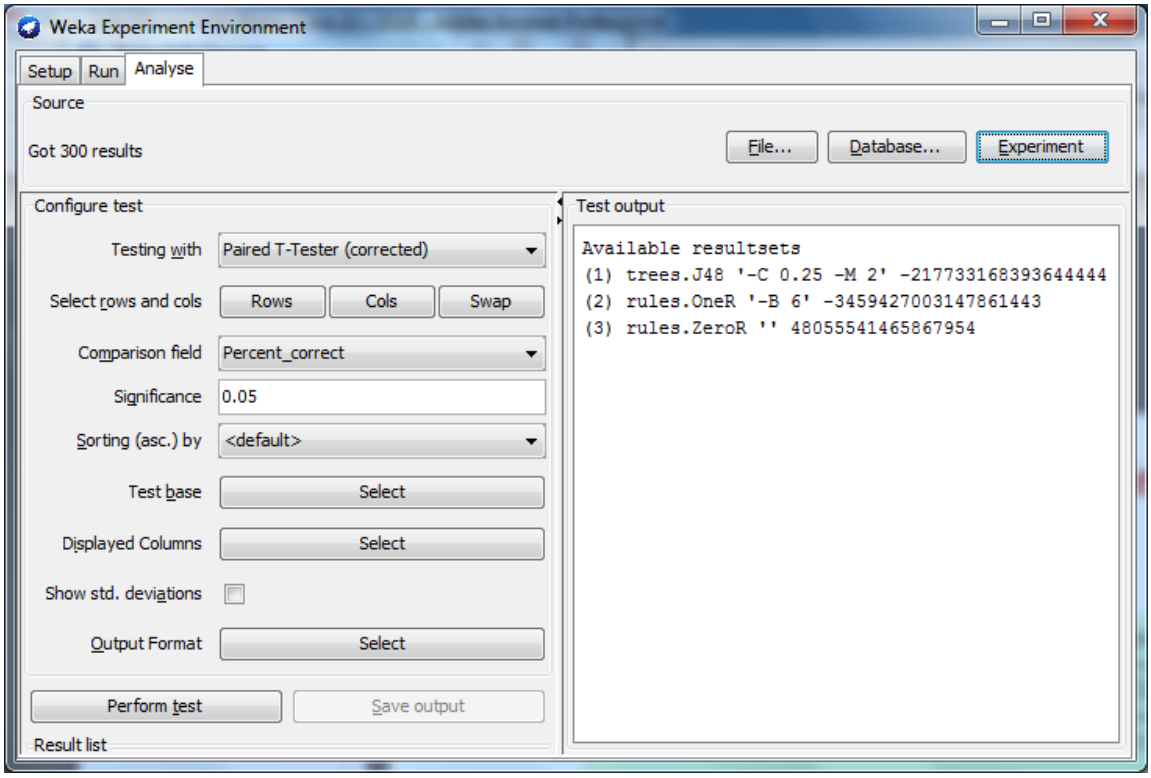
   ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)



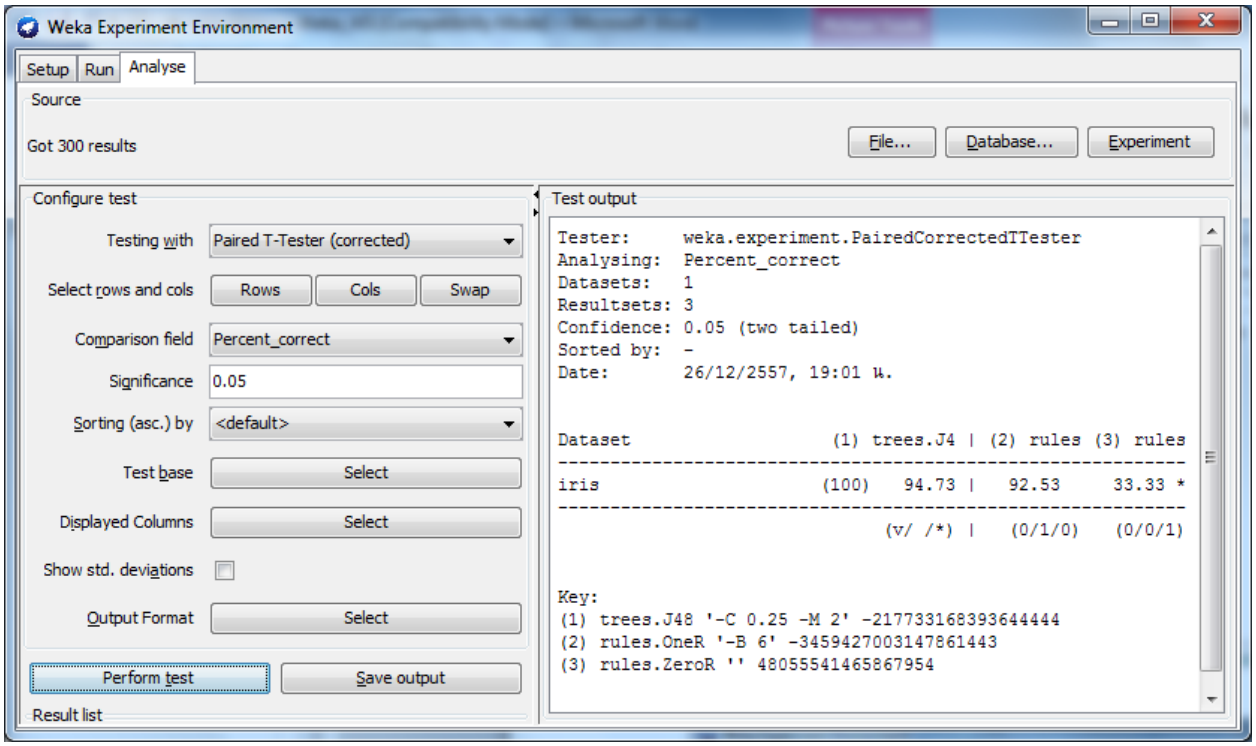
4. ไปที่แท็บ Run คลิก Start



5. ไปที่แท็บ Analyze คลิก Experiment



6. คลิก Perform test



7. การแปลผลจาก Witten

```
Dataset                 (1) trees.J4 | (2) rules (3) rules
----------------------------------------------------------
iris                   (100)   94.73 |   92.53     33.33 *
----------------------------------------------------------
                              (v/ /*) |   (0/1/0)   (0/0/1)


Key:
(1) trees.J48
(2) rules.OneR
(3) rules.ZeroR
```

v significantly better

* significantly worse

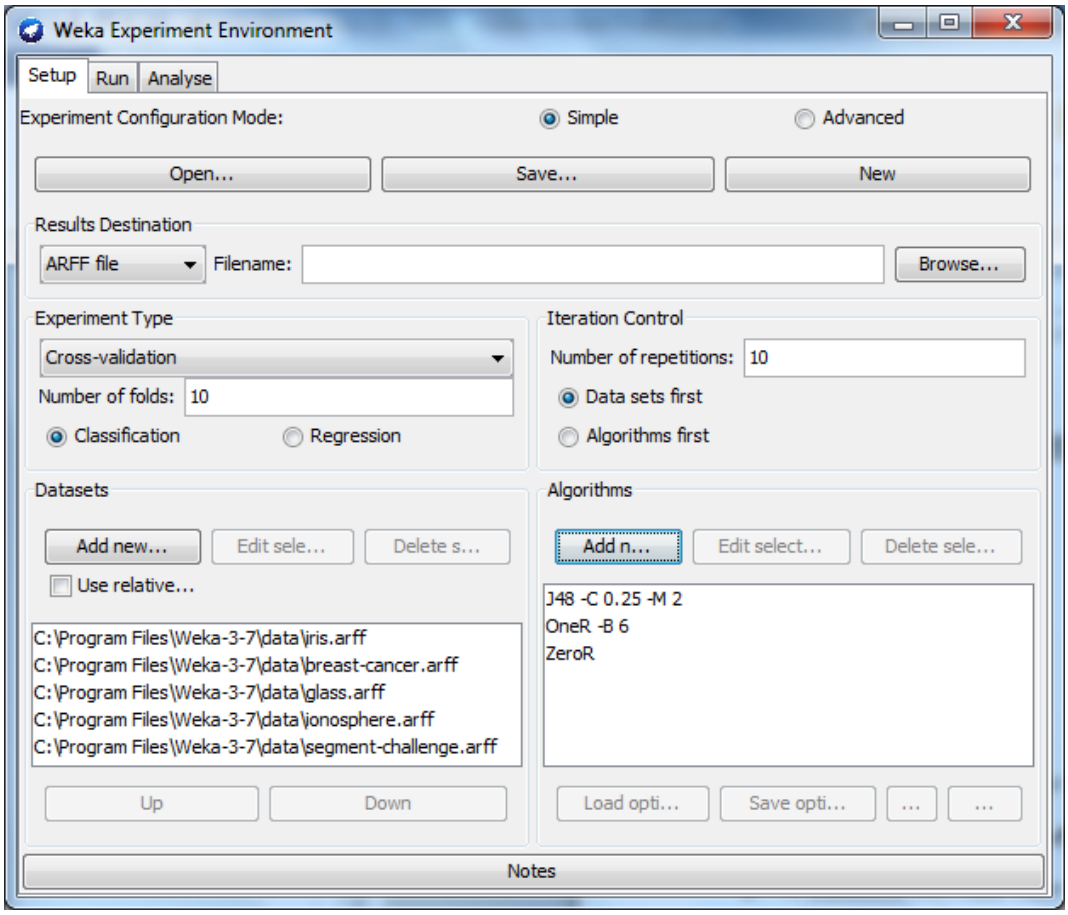
- ZeroR (33.3%) is significantly worse than J48 (94.7%)
- Cannot be sure that OneR (92.5%) is significantly worse than J48
- ... at the 5% level of statistical significance
- J48 seems better than ZeroR: pretty sure (5% level) that this is not due to chance

Experiment 2:

Data set: iris**,** breast-cancer, glass, ionosphere, segment-challenge

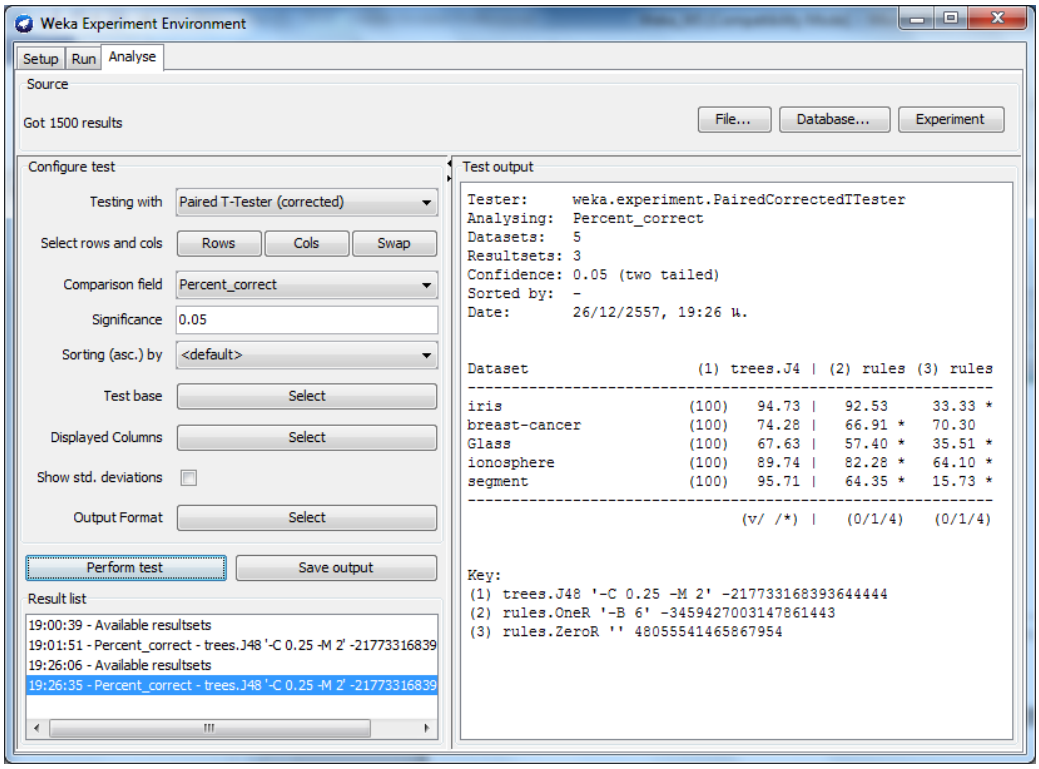
Algoritms: trees > J48, rules > OneR, rules > ZeroR

1. ไปที่แท็บ Set up คลิก New
2. ที่ Datasets คลิก Add new แล้วเปิดเลือก iris**,** breast-cancer, glass, ionosphere, segment-challenge (Add new 5 ครั้ง)
   ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)



3. ไปที่แท็บ Run คลิก Start
4. ไปที่แท็บ Analyze คลิก Experiment

5. คลิก Perform test



6. เปลี่ยน **Test base** Select เป็น rules.OneR