Chapter 4
Pre - processing

Data

- ข้อมูลแสดงในรูปแบบของตาราง
  - แถว เรียกว่า **ตัวอย่าง** (example)
  - คอลัมน์ เรียกว่า **แอตทริบิวต์** (attribute) ซึ่มีหน้าที่ (role) ที่ใช้งานบ่อย 3 แบบ
    - **ไอดี** (ID) เป็นแอตทริบิวต์ที่แสดงหมายเลขของข้อมูล หรือ primary key ในฐานข้อมูล
    - **แอตทริบิวต์ทั่วไป** (attribute) เป็นแอตทริบิวต์ปกติที่จะใช้ในการสร้างโมเดลหรือเรียกว่าเป็น ฟีเจอร์ (feature) หรือตัวแปรต้น (independent variable)
    - **ลาเบล** (label) เป็นแอตทริบิวต์ขนิดพิเศษที่มักจะใช้แสดงคำตอบของสิ่งที่เราต้องการจะสร้างโมเดลมาทำนาย หรือ เรียกว่า คลาส (class) หรือตัวแปรตาม (dependent

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

Value Type

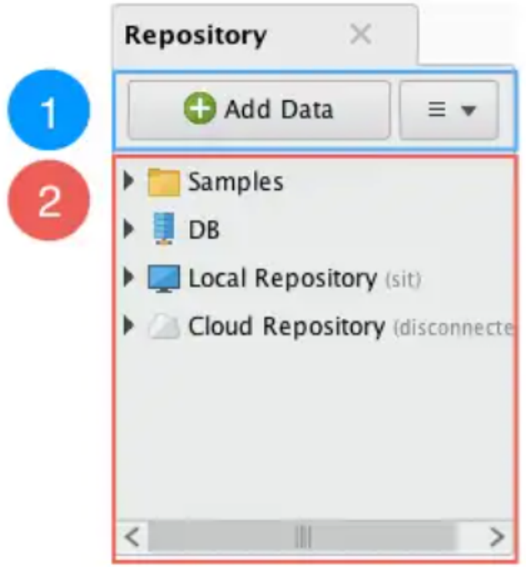
- ค่าของข้อมูลที่เก็บในแต่ละ**แอตทริบิวต์**
  - Polynominal ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่ามากกว่า 2 ค่าขึ้นไป
  - Binominal ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าเพียง 2 ค่าเท่านั้น
  - Numeric หรือ Integer ข้อมูลประเภทตัวเลข
  - Text ข้อมูลประเภทข้อความ

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

polynominal | numeric | binominal | numeric | binominal | numeric | binominal

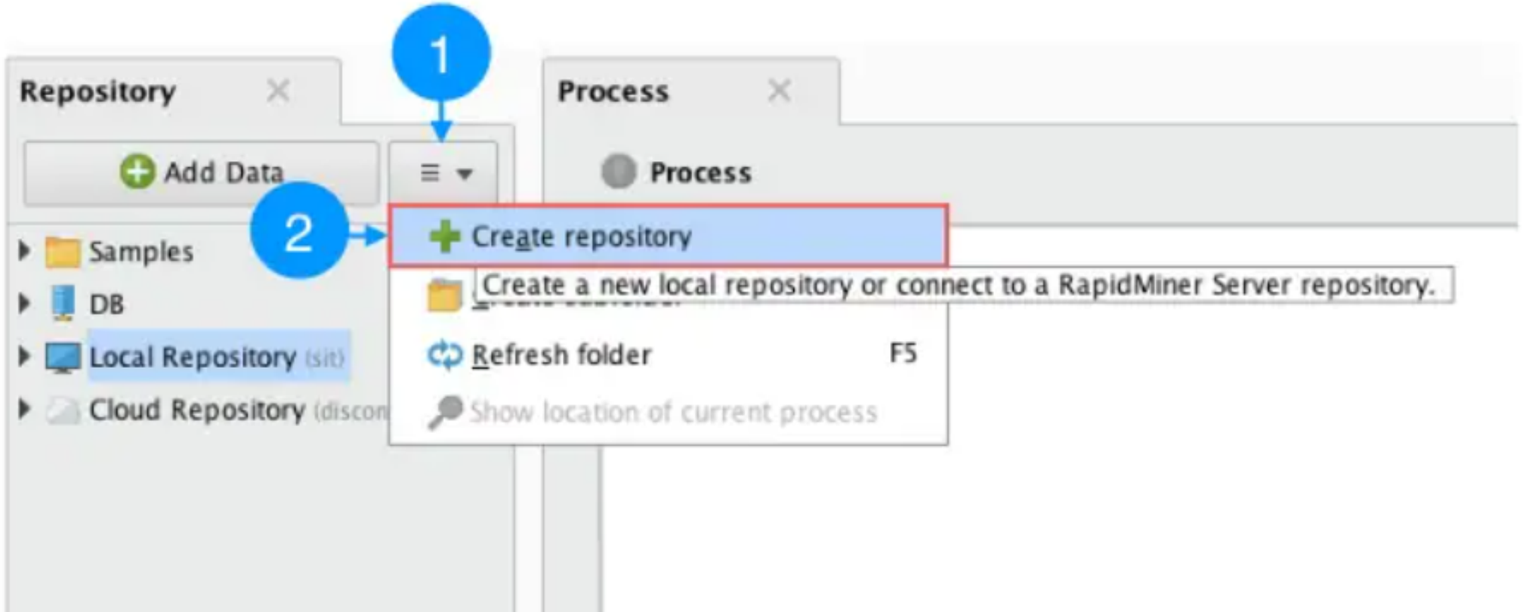
Data Management

- Repository
  - เป็นที่เก็บข้อมูลและ process เพื่อใช้งานใน RapidMiner Studio 7
  - ทำให้ไม่ต้องโหลดข้อมูลจากไฟล์ใหม่ทุกครั้ง
- องค์ประกอบในส่วน Repository
  - ส่วนที่ 1
    - สำหรับสร้าง Repository ใหม่
    - โหลดไฟล์ประเภทต่างๆ เข้าไปไว้ใน Repository
    - สร้างโฟลเดอร์ใหม่
  - ส่วนที่ 2
    - ข้อมูลและ process Sample ที่ RapidMiner Studio 7 เตรียมไว้ให้
    - ข้อมูลที่เก็บอยู่ในแต่ละ Repository



Data Management

- สร้าง Repository ใหม่
  - คลิกที่ ≡ ▾
  - หลังจากนั้นเลือกเมนู Create repository

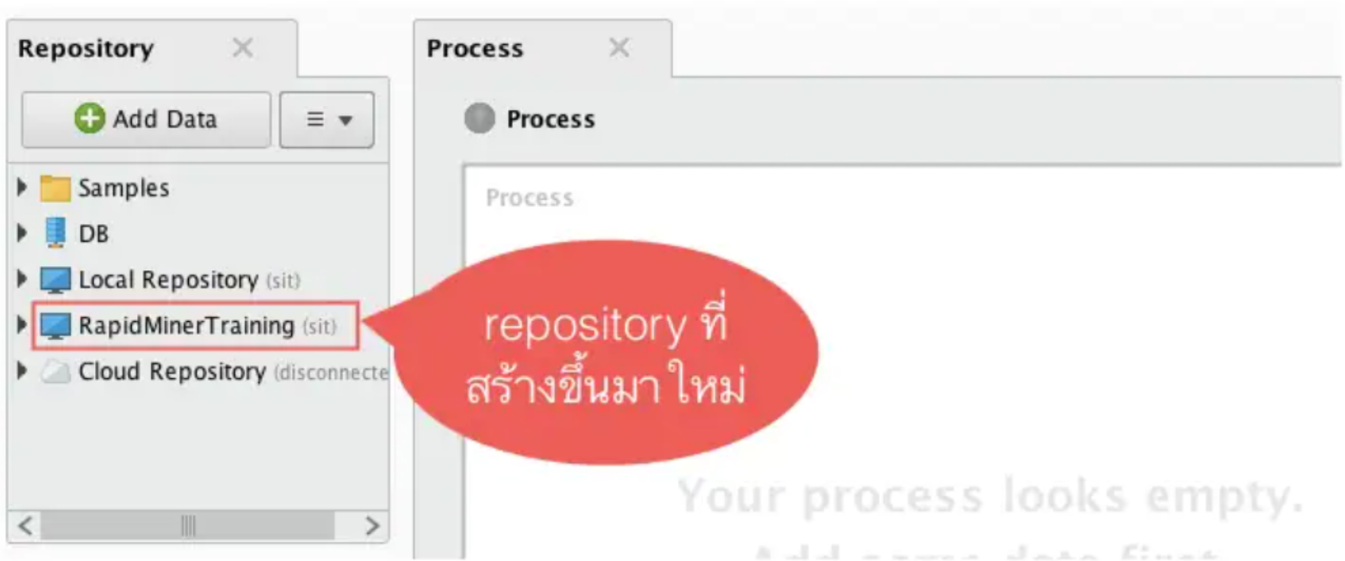


Data Management

- สร้าง Repository ใหม่ (ต่อ)
  - เลือก New local repository
  - กดปุ่ม Next

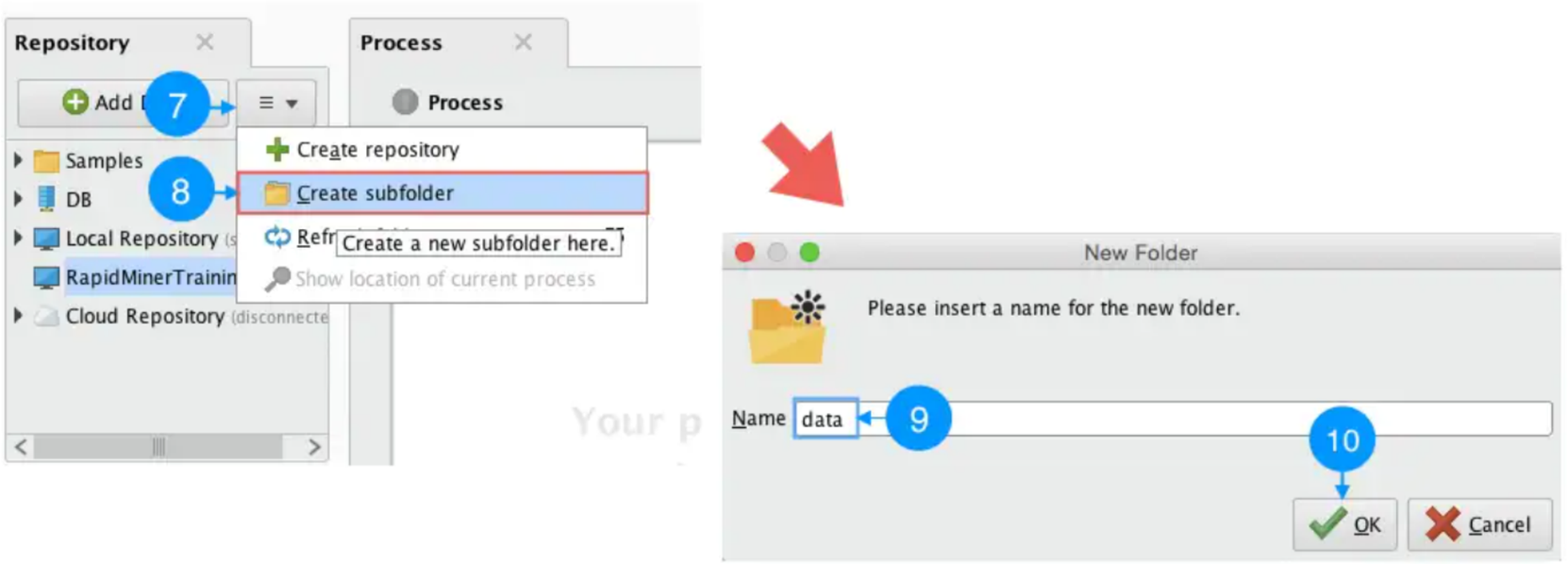
Data Management

- จะปรากฏ repository RapidMinerTraining แสดงขึ้นมาในส่วนของ Repository



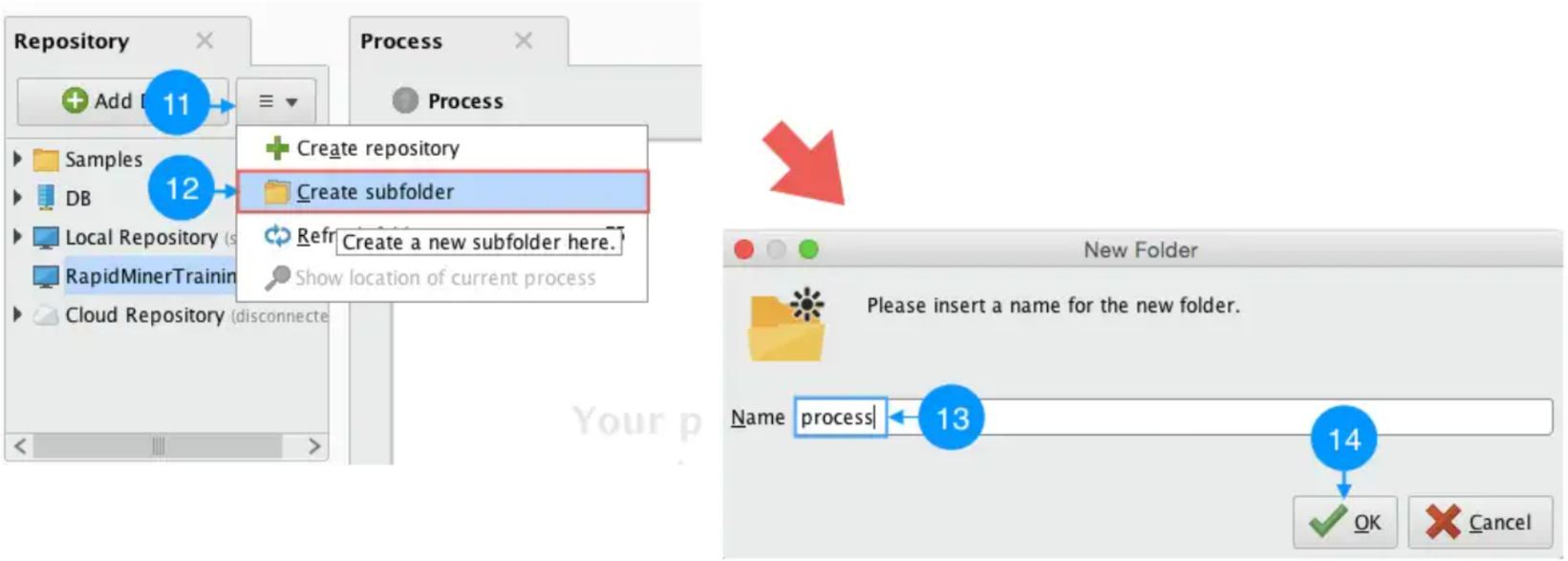
Data Management

- สร้างโฟลเดอร์ใน RapidMinerTraining
  - data สำหรับเก็บข้อมูล



Data Management

- สร้างโฟลเดอร์ใน RapidMinerTraining
  - data สำหรับเก็บข้อมูล
  - process สำหรับเก็บโพรเซสที่สร้างขึ้น



CSV File

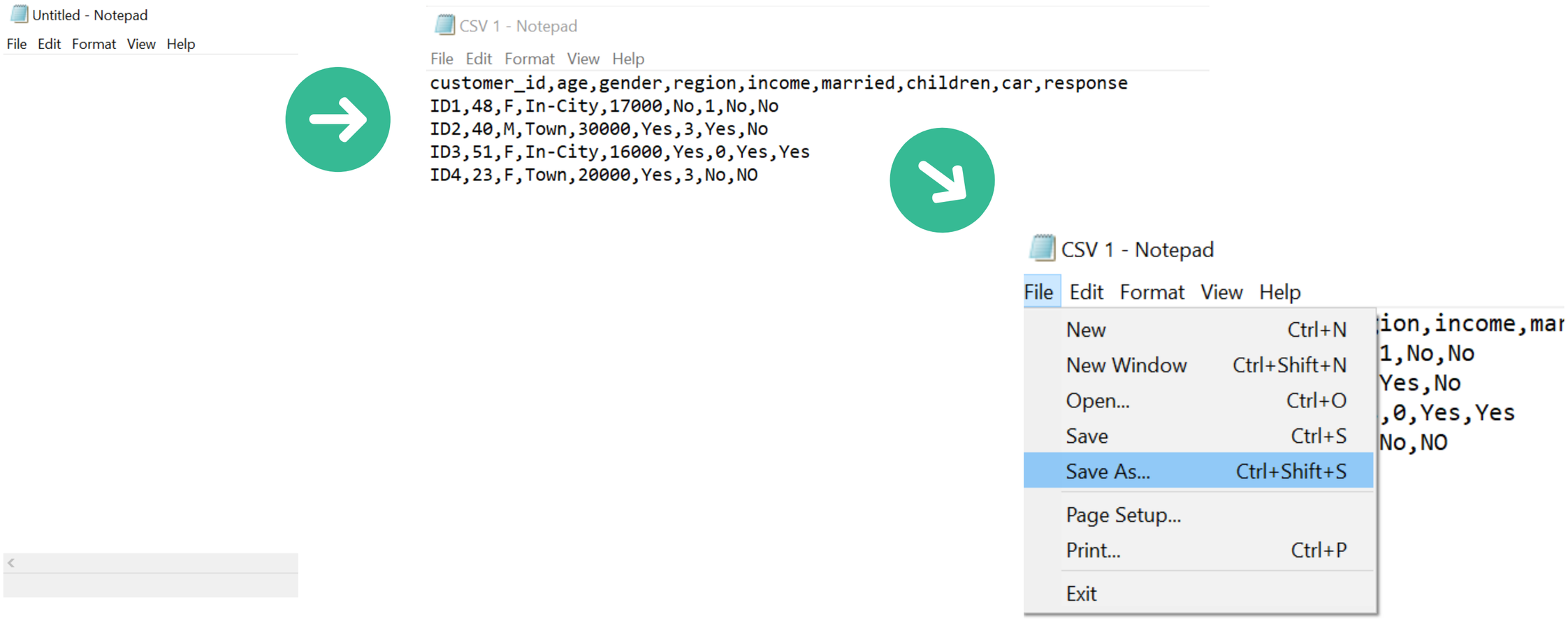
- ไฟล์ประเภท csv ย่อมาจาก Comma Separated Value
- ใช้เครื่องหมาย , (comma) คั่นระหว่างแอตทริบิวต์

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

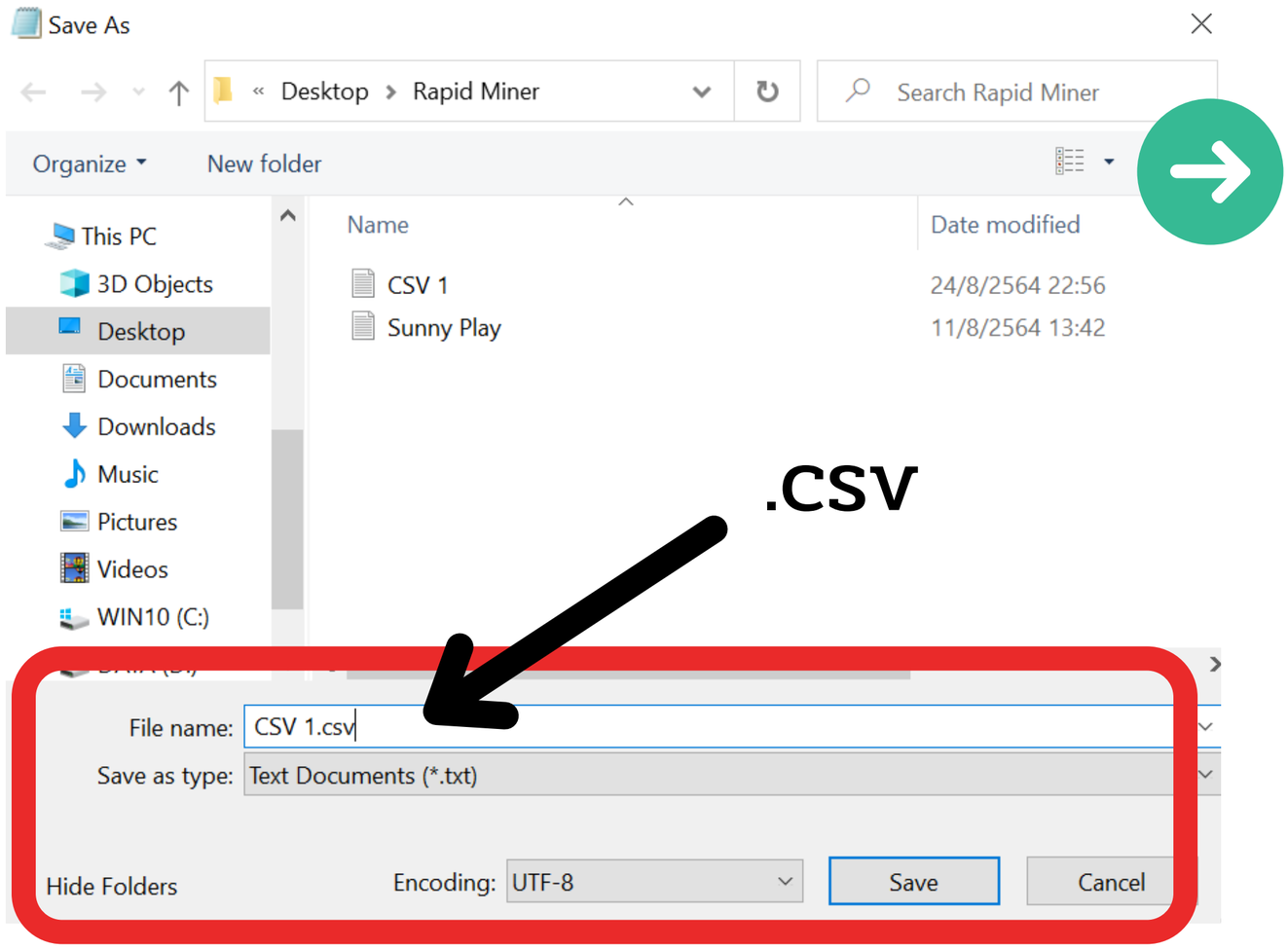
แถวแรกคือ header

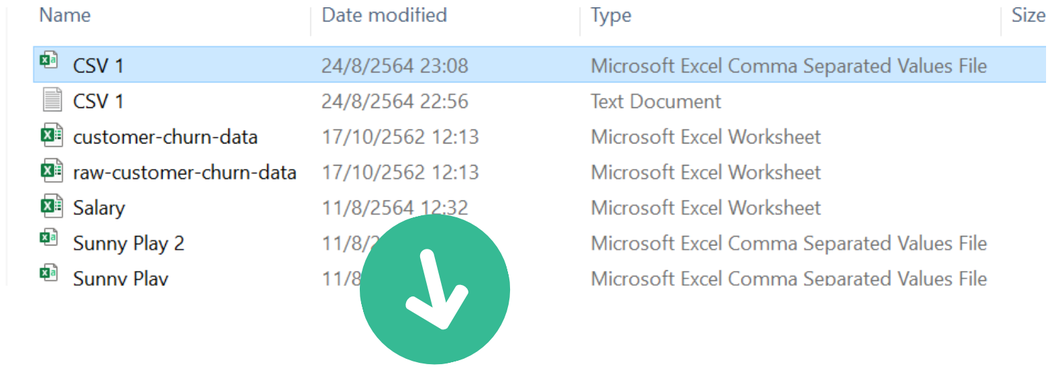
```
customer_id,age,gender,region,income,married,children,ca,response
ID12101,48,FEMALE,INNER_CITY,17546.0,NO,1,NO,NO
ID12102,40,MALE,TOWN,30085.1,YES,3,YES,NO
ID12103,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES
ID12104,23,FEMALE,TOWN,20375.4,YES,3,NO,NO
```

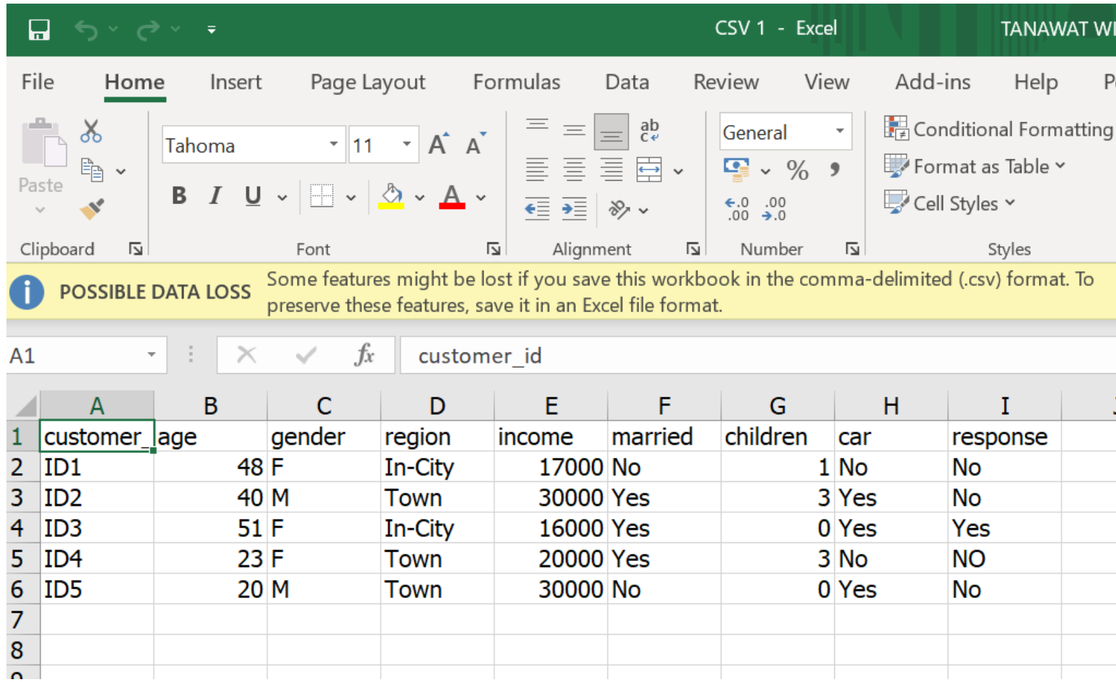
สร้างไฟล์ CSV


Untitled - Notepad
File Edit Format View Help
CSV 1 - Notepad
File Edit Format View Help
customer_id,age,gender,region,income,married,children,car,response
ID1,48,F,In-City,17000,No,1,No,No
ID2,40,M,Town,30000,Yes,3,Yes,No
ID3,51,F,In-City,16000,Yes,0,Yes,Yes
ID4,23,F,Town,20000,Yes,3,No,NO
CSV 1 - Notepad
File Edit Format View Help
New Ctrl+N
New Window Ctrl+Shift+N
Open... Ctrl+O
Save Ctrl+S
Save As... Ctrl+Shift+S
Page Setup...
Print... Ctrl+P
Exit

สร้างไฟล์ CSV

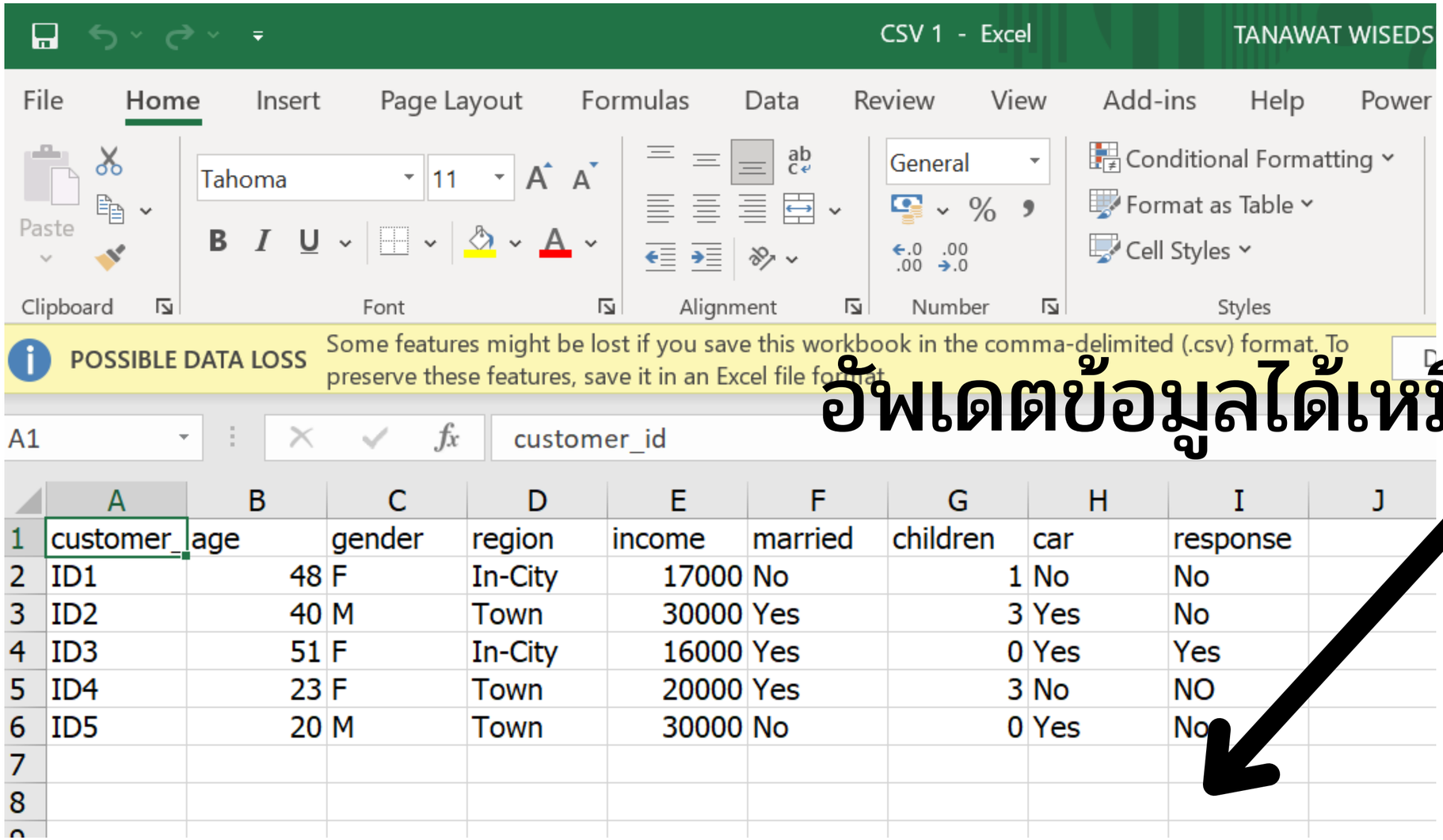






	A	B	C	D	E	F	G	H	I
1	customer_	age	gender	region	income	married	children	car	response
2	ID1	48	F	In-City	17000	No	1	No	No
3	ID2	40	M	Town	30000	Yes	3	Yes	No
4	ID3	51	F	In-City	16000	Yes	0	Yes	Yes
5	ID4	23	F	Town	20000	Yes	3	No	NO
6	ID5	20	M	Town	30000	No	0	Yes	No
7									
8									

สร้างไฟล์ CSV

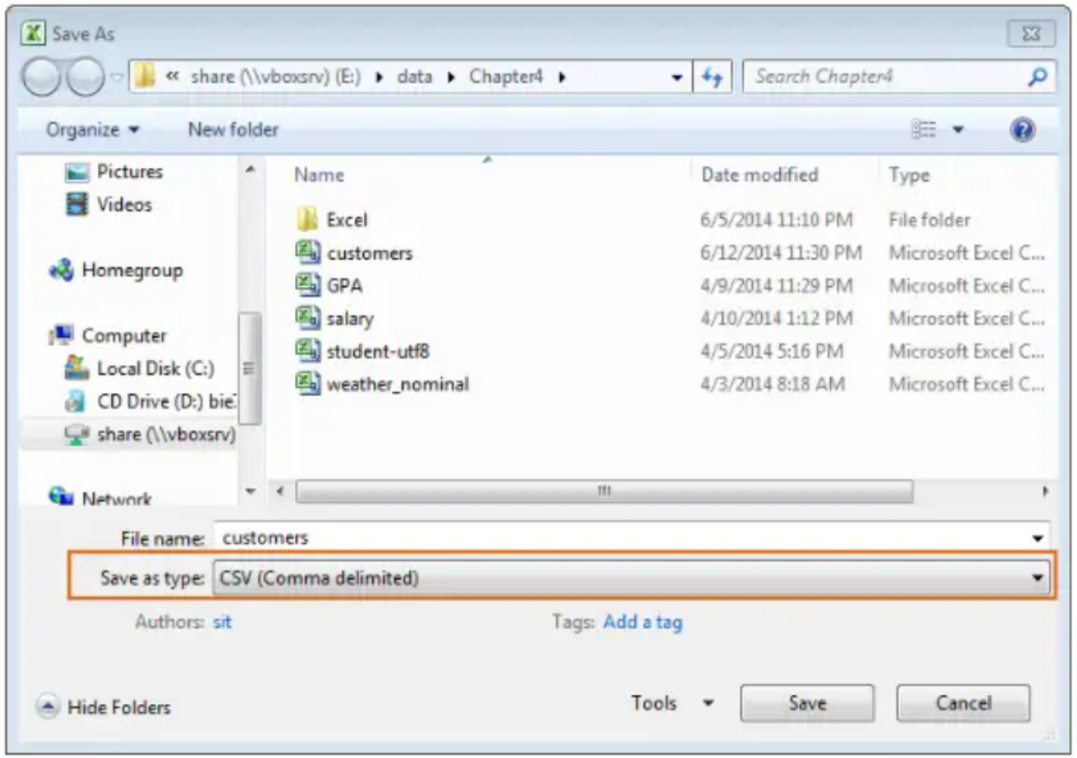


	A	B	C	D	E	F	G	H	I	J
1	customer_	age	gender	region	income	married	children	car	response	
2	ID1	48	F	In-City	17000	No	1	No	No	
3	ID2	40	M	Town	30000	Yes	3	Yes	No	
4	ID3	51	F	In-City	16000	Yes	0	Yes	Yes	
5	ID4	23	F	Town	20000	Yes	3	No	NO	
6	ID5	20	M	Town	30000	No	0	Yes	No	
7										
8										

อัพเดตข้อมูลได้เหมือน Excel

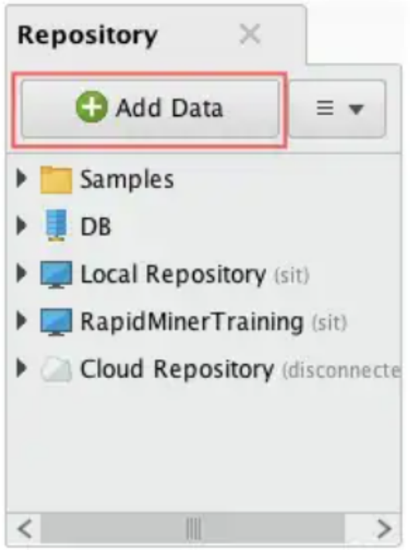
CSV File

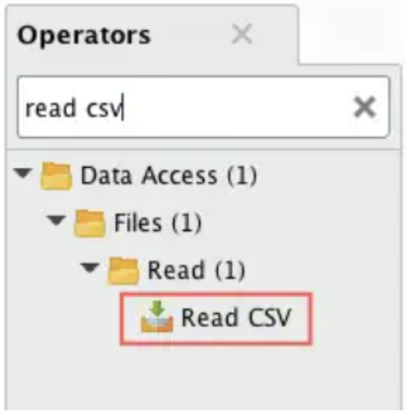
- ไฟล์ csv สามารถ export ได้จาก Excel หรือ database ต่างๆ
- Export จาก Excel
  - เลือก File > Save As > CSV (Comma delimited)



Load CSV File to RapidMiner

- การโหลดไฟล์ csv เข้าไปใช้ใน RapidMiner Studio 7 ทำได้ 2 แบบ
  - ใช้การ import ในส่วนของ Repositories
    - โหลดมาเก็บไว้ใน Repository และใช้งานได้ตลอด
    - ถ้าข้อมูลในไฟล์ csv มีการเปลี่ยนแปลงจะไม่ update ต้องทำการโหลดใหม่
  - ใช้โอเปอเรเตอร์ Read CSV
    - โหลดเข้ามาใช้งานโดยการอ่านจากไฟล์ csv ทุกครั้ง เมื่อไฟล์ update ข้อมูลจะเปลี่ยนตาม





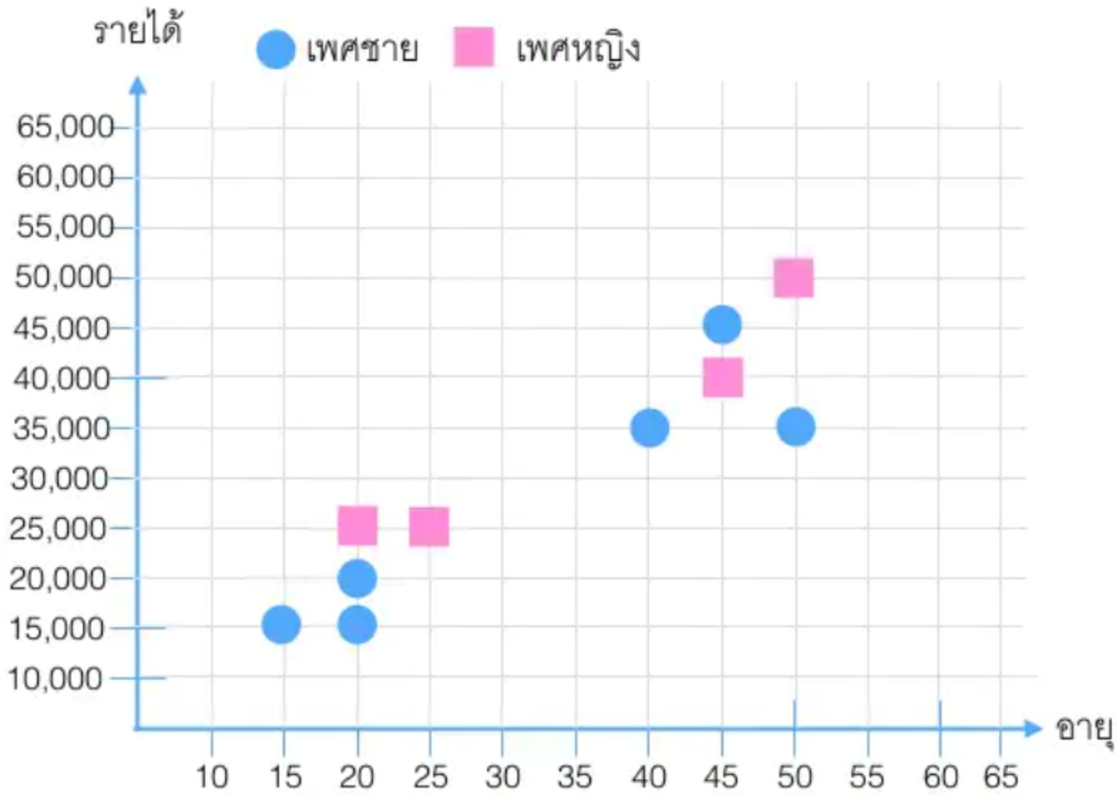
Data preparation

- การจัดการข้อมูล (preprocessing)
  - การเลือกแอตทริบิวต์
    - เลือกตามประเภทของแอตทริบิวต์
    - เลือกบางแอตทริบิวต์
  - การเลือกข้อมูล (example) ตามเงื่อนไข
  - การ join ข้อมูลจาก 2 ชุด
  - มีความผิดพลาดในชุดข้อมูล เช่น
    - ข้อมูลมีค่าไม่ตรงกัน
    - ข้อมูลขาดขายไป
    - ข้อมูลแปลกแยก (outlier)
  - การแปลงข้อมูล
    - Discretization แปลงข้อมูล numeric ให้เป็น nominal
      - แบ่งตามเงื่อนไขที่กำหนด (user defined)
      - แบ่งตามความถี่ที่เท่ากัน (equal frequency)

Outlier

หมายเลข	อายุ	รายได้	เพศ
1	15	15,000	ชาย
2	20	15,000	ชาย
3	20	20,000	ชาย
4	20	25,000	หญิง
5	25	25,000	หญิง
6	40	35,000	ชาย
7	45	40,000	หญิง
8	45	45,000	ชาย
9	50	35,000	ชาย
10	50	50,000	หญิง

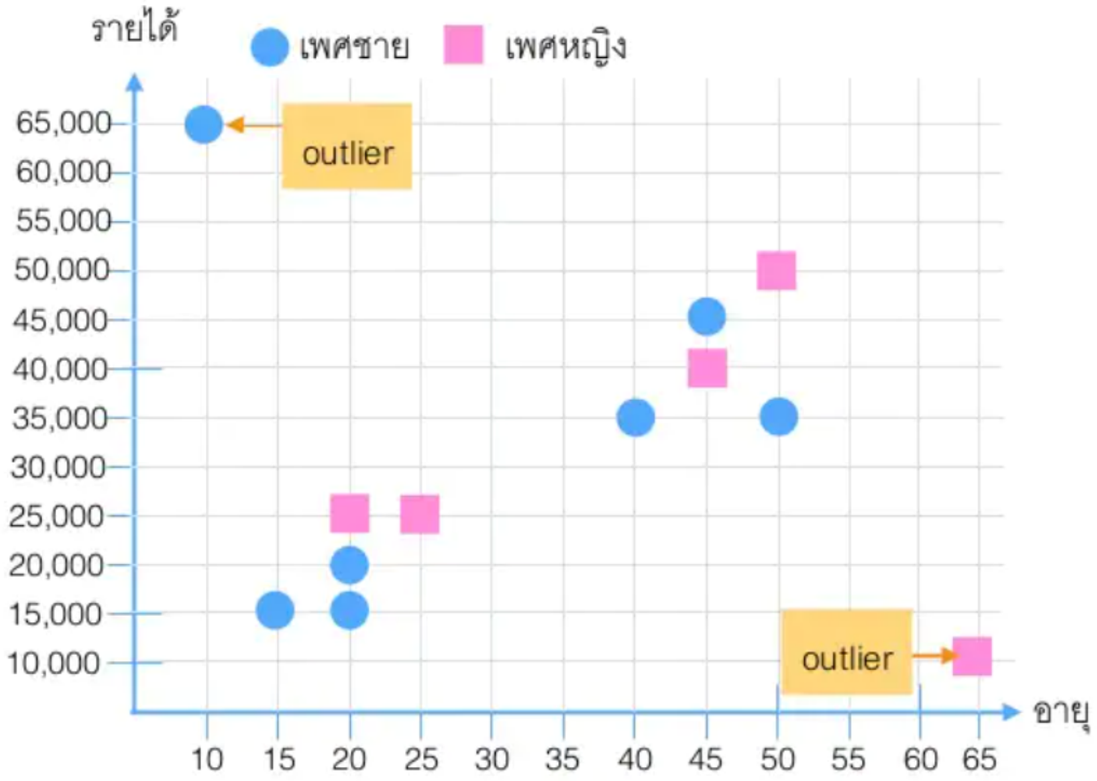
ข้อมูลลูกค้า



scatter plot ของข้อมูลลูกค้า

Outlier

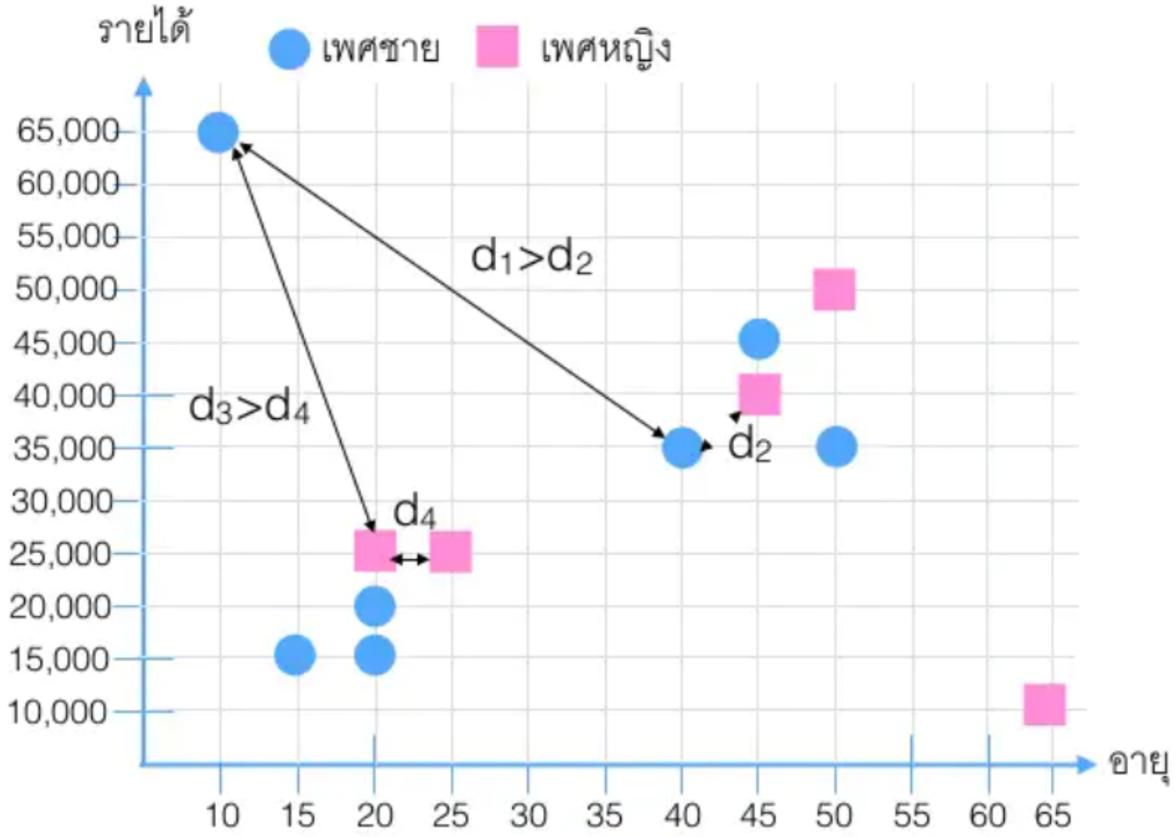
หมายเลข	อายุ	รายได้	เพศ
1	15	15,000	ชาย
2	20	15,000	ชาย
3	20	20,000	ชาย
4	20	25,000	หญิง
5	25	25,000	หญิง
6	40	35,000	ชาย
7	45	40,000	หญิง
8	45	45,000	ชาย
9	50	35,000	ชาย
10	50	50,000	หญิง
11	10	65,000	ชาย
12	65	10,000	หญิง



scatter plot ของข้อมูลลูกค้า

Detect outlier by distance

- วัดระยะห่างระหว่างข้อมูล (example) แต่ละตัวกับข้อมูลใกล้เคียง
- outlier คือ ข้อมูลที่มีระยะห่างกับข้อมูลอื่นๆ มากๆ



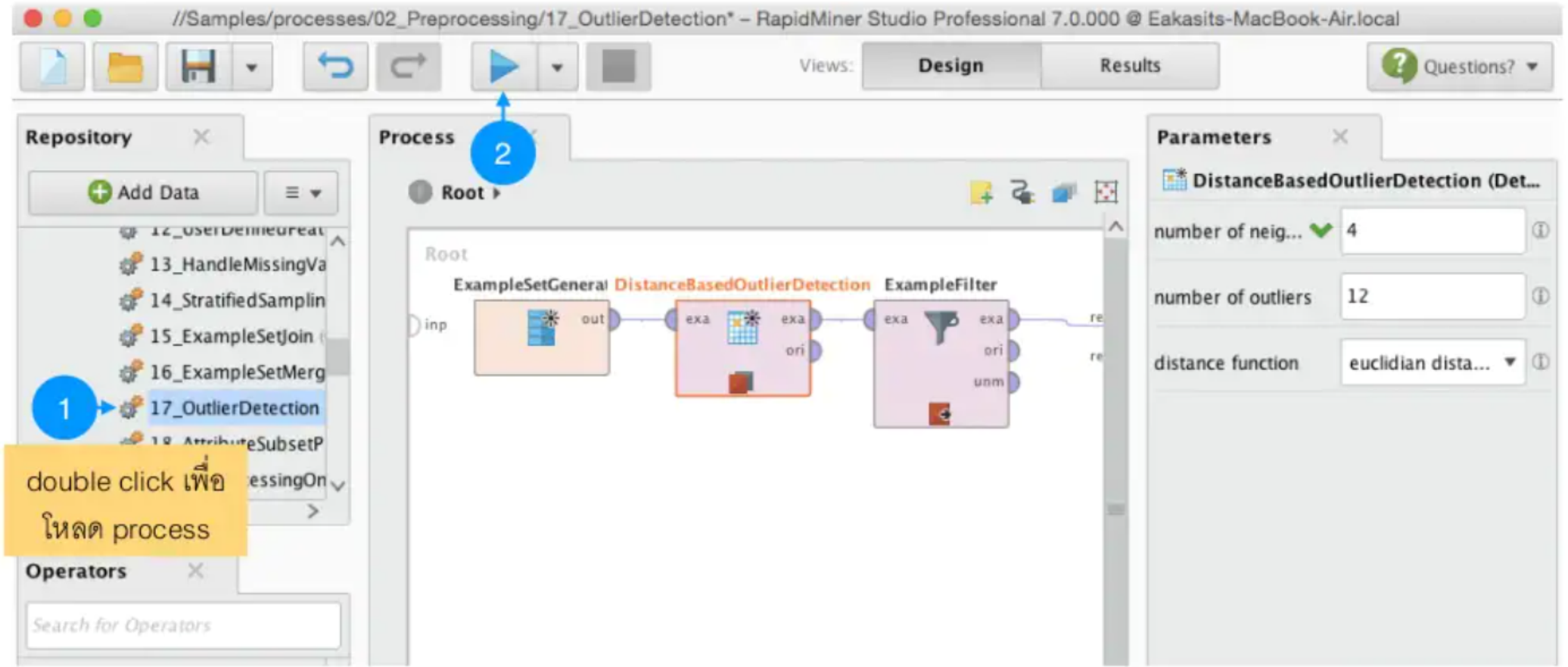
Detect outlier by distance

- โอเปอเรเตอร์ที่เกี่ยวข้อง

โอเปอเรเตอร์		คำอธิบาย
	Generate Data	ใช้สำหรับสร้างข้อมูล
	Detect Outlier (Distances)	ใช้ในการหา outlier ด้วยวิธีดูจากระยะห่างระหว่างข้อมูล
	Filter Examples	ลบข้อมูลตามเงื่อนไขที่กำหนด

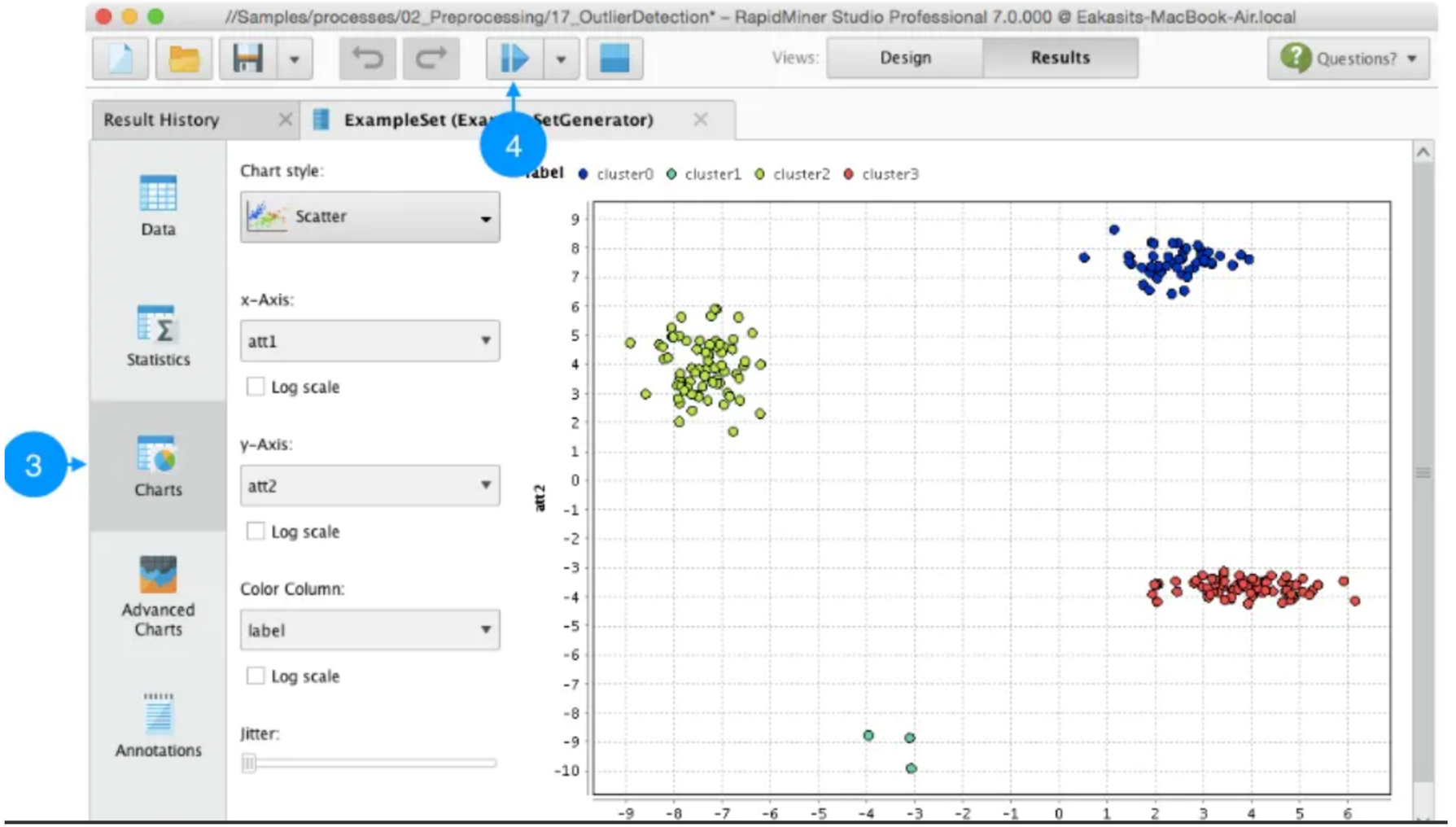
Detect outlier by distance

- เลือกตัวอย่าง Process ที่ RapidMiner Studio 7 เตรียมไว้ให้
  - เลือก Samples > Processes > 02_Preprocessing > 17_OutlierDetection



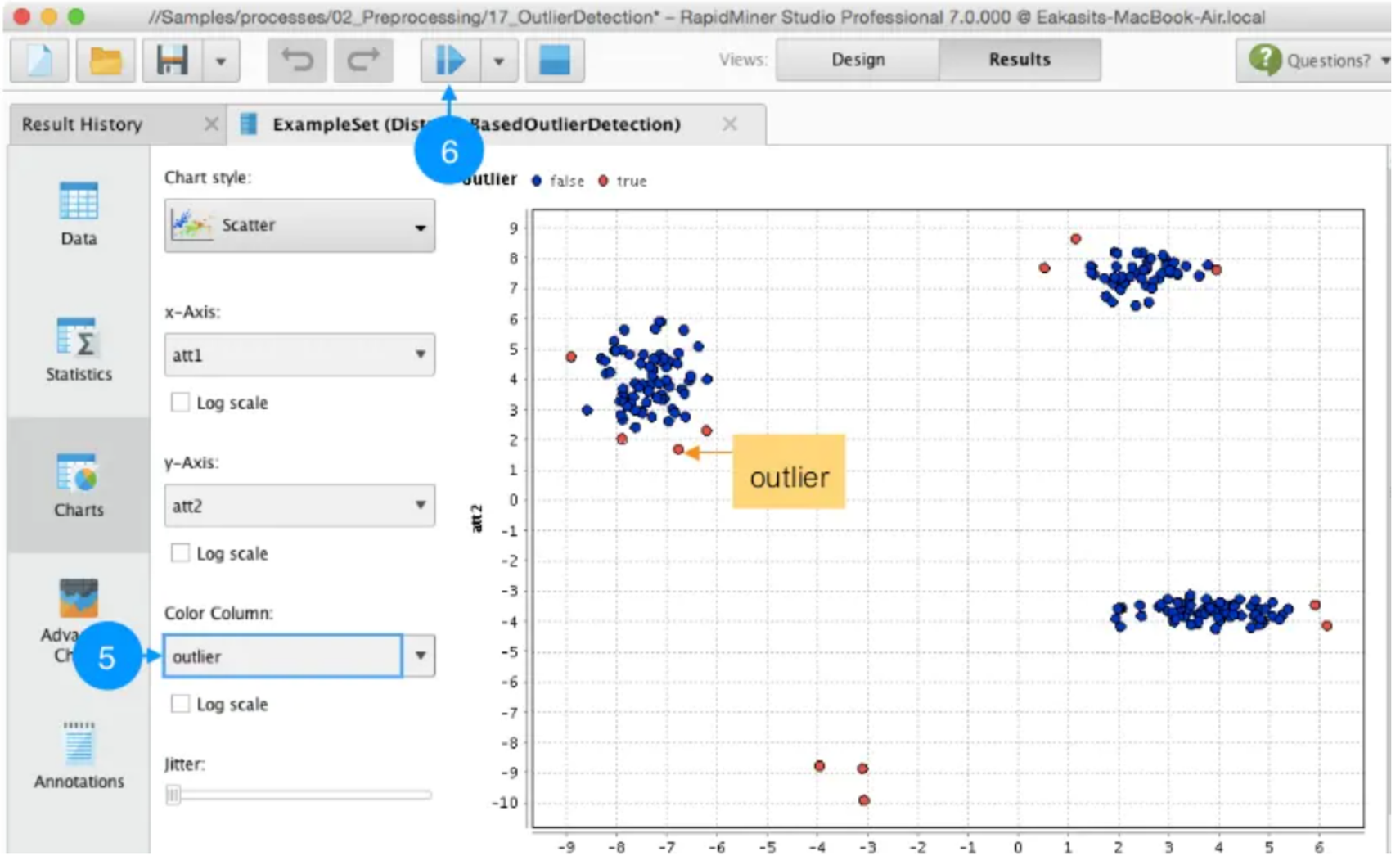
Detect outlier by distance

- แสดงข้อมูลแบบ scatter plot



Detect outlier by distance

- แสดงข้อมูล outlier ใน scatter plot



Detect outlier by distance

- ลบข้อมูลที่เป็น outlier ทิ้ง