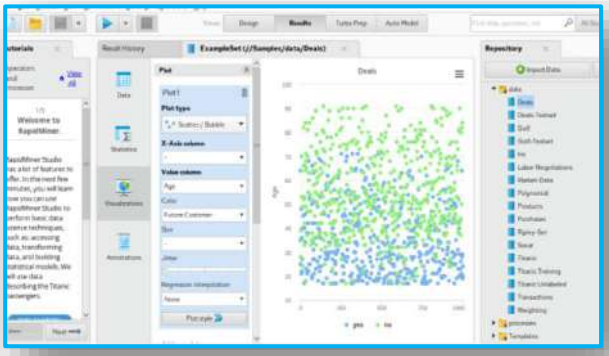


สรุปบทเรียน

การใช้โปรแกรมดิจิทัลเพื่อการวิเคราะห์ข้อมูล

(Data Analysis and RapidMiner)

โดย นางสาวศศิรินทร์ ศรีสมเขียว

นักวิชาการเกษตรชำนาญการ

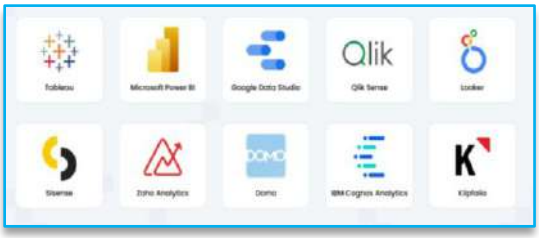
Data Analysis คือ กระบวนการที่ใช้ในการพิจารณา ค้นคว้า หาความหมาย ข้างในข้อมูล (inside) หรือรูปแบบ (pattern) ที่ซ่อนอยู่ในข้อมูลที่เป็นประโยชน์ ซึ่งข้อมูลเป็นส่วนสำคัญขององค์กรที่สามารถช่วยในการตัดสินใจ นอกจากนี้ การวิเคราะห์ข้อมูลนั้นสามารถนำมาใช้ในงานวิจัยต่างๆ ได้อีกด้วย ในการวิเคราะห์ข้อมูลมีขั้นตอนที่สำคัญ ได้แก่ 1) การระบุปัญหา (define the problem) 2) การเก็บรวรวมข้อมูลที่เกี่ยวข้อง (collect the data) 3) การเตรียมข้อมูลเบื้องต้น (pre-Process the data) 4) การวิเคราะห์ข้อมูล (analysis the data) และ 5) การตีความหมายของผลลัพธ์ (interpret the results)

เครื่องมือประเภทที่ 1 ที่นิยมนำมาใช้ในการวิเคราะห์ข้อมูลเรียกว่า Spreadsheets คือ โปรแกรมพวก Microsoft Excel และ Google sheets เครื่องมือประเภทนี้เหมาะสำหรับข้อมูลขนาดไม่ใหญ่มาก แต่จะเหมาะสมกับข้อมูลขนาดเล็กถึงขนาดกลาง ซึ่งสามารถทำให้จัดการข้อมูล (Manipulate) วิเคราะห์ข้อมูลทางสถิติ สามารถเอาข้อมูลต่างๆ มาสร้างเป็นกราฟได้

เครื่องมือประเภทที่ 2 คือ ซอฟต์แวร์ที่เป็นทางด้านสถิติโดยตรง (statistical software) เช่น SPSS, SAS, และ Strata จะมีเมนูการวิเคราะห์ทางสถิติเชิงลึก เช่น การทำ regression analysis การทำ time series analysis รวมถึงการทำ data mining เป็นต้น

เครื่องมือประเภทที่ 3 คือ data mining tools เป็นเครื่องมือด้านการทำเหมืองข้อมูล เช่น ซอฟต์แวร์ RapidMiner, KNIME, และ Weka เป็นโปรแกรมด้านวิทยาการข้อมูล รวมถึงการเรียนรู้ของเครื่อง (machine learning: ML) ที่สามารถวิเคราะห์ข้อมูลแล้วสร้างเป็นโมเดลรูปแบบต่างๆ ได้

เครื่องมือประเภทที่ 4 คือ (business intelligence tools: BI tools) เช่น Tableau, Power BI, และ QlikView เป็นเครื่องมือที่ไว้สร้างกราฟที่มีลักษณะแดชบอร์ด สามารถ Drag และ Drop เพื่อให้ผู้ใช้ที่ไม่ได้เป็นผู้เชี่ยวชาญ (non-technical) สามารถสร้างกราฟรูปแบบต่างๆ ได้อย่างสวยงามได้



เครื่องมือประเภทที่ 5 คือ การเขียนโค้ดคำสั่งการทำงานด้วยตนเอง กลุ่มเครื่องมือ (programming languages) เช่น โปรแกรม Python และ โปรแกรม R เพื่อใช้ในการวิเคราะห์ข้อมูลตามที่ต้องการ ซึ่งโปรแกรมเหล่านี้เป็น open source จึงสามารถใช้งานได้ฟรีผ่านเว็บไซต์

สำหรับโปรแกรม RapidMiner เป็นโปรแกรมที่นำมาใช้ในการวิเคราะห์ข้อมูลเน้นทางด้านวิชาการข้อมูล และการเรียนรู้ของเครื่อง (machine learning: ML) ลักษณะของโปรแกรมจะมีลักษณะ drag และ drop คือ คำสั่งแต่ละคำสั่งของโรแกรมจะอยู่ในรูปแบบของกล่อง โดยไม่จำเป็นต้องเขียนโค้ด

- **การทำความเข้าใจข้อมูล**
  - **ชุดข้อมูล (dataset)** คือ รายการที่อธิบายวัตถุตั้งแต่ 1 วัตถุขึ้นไป โดยวัตถุ คือ สิ่ง/เหตุการณ์ที่เราสนใจ
  - **ตัวแปร (variable)** คือ คุณลักษณะที่อธิบายวัตถุ ตัวแปรอธิบายหตุการณ์ " ประเภทของตัวแปรเป็นตัวกำหนดค่าที่เป็นไปได้ทั้งหมด (ขอบเขต) ของตัวแปร แบ่งออกเป็น 1) ตัวแปรเชิงคุณภาพ (categorical) คือ ตัวแปรที่ค่าที่เป็นไปได้ใช้บอกลักษณะ การแบ่งเป็นประเภท เช่น เชื้อชาติ : ไทย จีน ญี่ปุ่น พม่า ลาว อาชีพ : ข้าราชการ ค้าขาย นักธุรกิจ เป็นต้น และ 2) ตัวแปรเชิงปริมาณ (numerical) คือ ตัวแปรที่ค่าเป็นตัวเลข แสดงปริมาณความมากน้อยนำไปคำนวณได้ (บอกความแตกต่างเป็นตัวเลขได้ ค่าเฉลี่ยมีความหมาย) เช่น จำนวนบุตร : 0, 1, 2,... รายได้ : >0 เป็นต้น
  - **สถิติเชิงพรรณนา (descriptive statistics)** คือ สถิติที่นำมาใช้ในการอธิบายคุณลักษณะเบื้องต้นของข้อมูล ซึ่งนำมาใช้เพื่อ 1) วัดแนวโน้มเข้าสู่ส่วนกลาง (central tendency) เพื่อการหาค่าของข้อมูลเพื่อเป็นตัวแทนของข้อมูลทั้งหมด ค่าแนวโน้มเข้าสู่ส่วนกลางเป็นค่ากลางที่อยู่กึ่งกลางระหว่างข้อมูลทั้งหมด ใช้เป็นตัวแทนของชุดข้อมูลสำหรับสำหรับตัวแปรแต่ละตัว ประเภทของค่าแนวโน้มเข้าสู่ส่วนกลางที่สำคัญ ได้แก่ ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) มัธยฐาน (Median) ฐานนิยม (Mode)

2) การวัดการกระจายตัวของข้อมูล (data dispersion) เพื่อการวัดว่าข้อมูลแต่ละตัวมีค่าแตกต่างกันมากน้อยเพียงใด เนื่องจากการวัดแนวโน้มเข้าสู่ส่วนกลางเพียงอย่างเดียว เป็นการอธิบายข้อมูลไม่สมบูรณ์ และมีโอกาสคลาดเคลื่อนได้ ดังนั้นควรนำลักษณะการกระจายตัวของกลุ่มข้อมูลมาพิจารณาควบคู่ด้วย ซึ่งข้อมูลการกระจาย คือ การที่ข้อมูลแต่ละชุดมีค่าไม่เท่ากัน ถ้าชัดข้อมูลประกอบด้วย

- ค่าข้อมูลที่แตกต่างกันมาก เรียกว่า ข้อมูลมีการกระจายมาก
- ค่าข้อมูลที่แตกต่างกันน้อย เรียกว่า ข้อมูลมีการกระจายน้อย
- ถ้าข้อมูลมีค่าเท่ากันทั้งหมด เรียกว่า ข้อมูลไม่มีการกระจาย

สำหรับสถิติที่ใช้ในการวัดการกระจายตัวของข้อมูล ได้แก่ พิสัย (range) คำนวณได้จากการหาความแตกต่างระหว่างค่าสูงสุดและค่าต่ำสุด ดังนั้น พิสัยจึงบอกถึงความกว้างของข้อมูลชุดนั้นๆ และส่วนเบี่ยงเบนมาตรฐาน (standard deviation) สามารถใช้วัดค่าต่างๆ ในชุดข้อมูลกระจายตัวห่างจากค่าเฉลี่ยมากน้อยเพียงใด

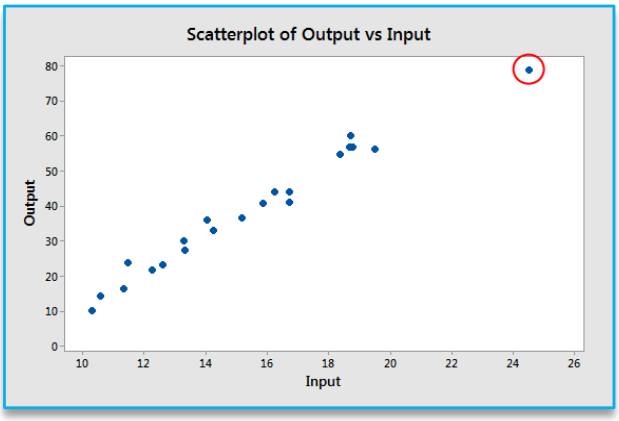
- **การเตรียมข้อมูล**

การเตรียมข้อมูลมีวัตถุประสงค์เพื่อลดปริมาณข้อมูลขยะ เนื่องจากเมื่อนำข้อมูลมาวิเคราะห์อาจทำให้เกิดการตัดสินใจผิดพลาดได้ แต่หากข้อมูลมีความถูกต้องเมื่อนำมาวิเคราะห์ จะช่วยให้สามารถตัดสินใจได้ถูกต้องมากยิ่งขึ้น ลักษณะของข้อมูลที่มีคุณภาพควรที่จะมี 5 คุณลักษณะ ได้แก่ 1) ความสอดคล้องกัน (consistency) ค่าของข้อมูลมีความสอดคล้องกันในทุกแหล่งข้อมูล 2) ความถูกต้อง (accuracy) ค่าของข้อมูลจัดเก็บตามจริง และรูปแบบข้อมูล

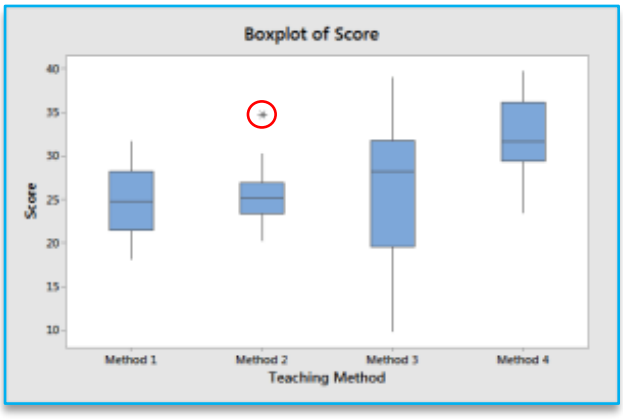
ถูกต้องตามที่กำหนดไว้ 3) ความถูกต้องสมบูรณ์ (completeness) ไม่มีข้อมูลสูญหาย มีข้อมูลครบถ้วน 4) สามารถตรวจสอบได้ (auditability) สามารถทำการตรวจสอบข้อมูลได้ 5) ความทันสมัย (timeliness) ความทันเวลา ข้อมูลมีความพร้อมใช้เมื่อต้องการ การเตรียมข้อมูลทำให้เห็นภาพรวมทั้งหมดได้ดีขึ้น หากข้อมูลมีขนาดใหญ่เกินไปเราควรทำการลดขนาดข้อมูล (data reduction) ให้เหลือข้อมูลขนาดที่จำเป็นเท่านั้น การทำความสะอาดข้อมูล (data cleaning) ข้อมูลไม่สะอาดส่งผลต่อความถูกต้องของการวิเคราะห์ข้อมูล ข้อมูลที่ไม่สะอาด ผู้ใช้จะลังเลที่จะเชื่อผลที่ได้จากการวิเคราะห์ข้อมูลดังกล่าว นอกจานี้ ควรมีการเติมค่าที่สูญหาย รวมถึงการจัดการค่าสุดโต่ง ซึ่งค่าสุดโต่งคือ ค่าของข้อมูลที่สูงหรือต่ำจากข้อมูลตัวอื่นๆ สูงอย่างผิดปกติ

- **การจัดการค่าสุดโต่ง (Outliers)** การพิจารณาค่าสุดโต่งโดยใช้วิธีต่างๆ ดังนี้

1) โดยใช้แผนภาพกระจายของข้อมูล (scatter plot)



2) โดยใช้แผนภาพกล่อง (box plot)

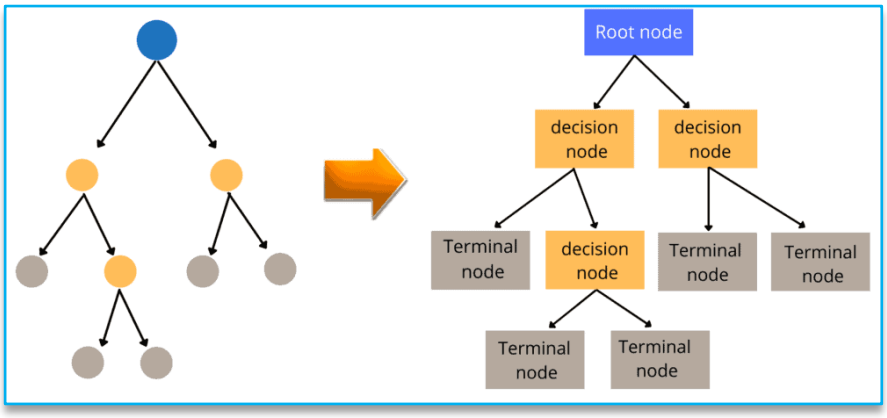


การจัดการค่าสุดโต่ง ได้แก่ 1) การลบข้อมูลรายการที่มีค่าสุดโต่ง ข้อดีคือ สามารถรักษาข้อมูลเดิมของรายอื่นๆ ที่มีค่าปกติ ข้อเสียคือ จำนวนรายการข้อมูลน้อยลง 2) แก้ไขค่าผิดปกติ โดยผู้เชี่ยวชาญ ข้อดีคือ รักษาข้อมูลเดิม และข้อมูลที่รบกวนถูกแก้ไข ข้อเสียคือ ผู้เชี่ยวชาญหายาก เสี่ยงต่อการแก้ไขและข้อมูลผิดพลาด การลดขนาดข้อมูล (data reduction) การลดขนาดข้อมูลช่วยลดพื้นที่ในการเก็บข้อมูล ยังช่วยให้อัลกอริทึมวิเคราะห์ข้อมูลทำงานได้เร็วขึ้น บริหารจัดการข้อมูลง่ายขึ้น การลดขนาดข้อมูลอาจทำได้โดยการลดจำนวนตัวแปร (dimensionality reduction) การแทนด้วยค่าที่เล็กลง (numerosity reduction) การลดขนาดข้อมูลจะช่วยให้โปรแกรมหรืออัลกอริทึมที่ใช้ในการประมวลผลวิเคราะห์ข้อมูลทำงานได้เร็วขึ้น

- **การสร้างโมเดลเพื่องานจำแนกประเภท**

การเรียนรู้ของเครื่อง (Machine learning) คือ กระบวนการสอนคอมพิวเตอร์โดยใช้ข้อมูลตัวอย่างเพื่อให้คอมพิวเตอร์เรียนรู้ได้ด้วยตนเอง และสกัดเอาสาระออกมาจากข้อมูลได้ การจำแนกประเภทเป็นการเรียนรู้แบบมีผู้สอน เป็นการมีชุดข้อมมูลและมีชุดคำตอบที่ถูกต้องให้กับคอมพิวเตอร์ได้เรียนรู้ โดยการจำแนกประเภทของข้อมูลจะเป็นคำตอบเชิงคุณภาพ คอมพิวเตอร์จะทำการพิจารณาคุณลักษณะของวัตถุ (feature) ที่ต้องการจะจำแนก

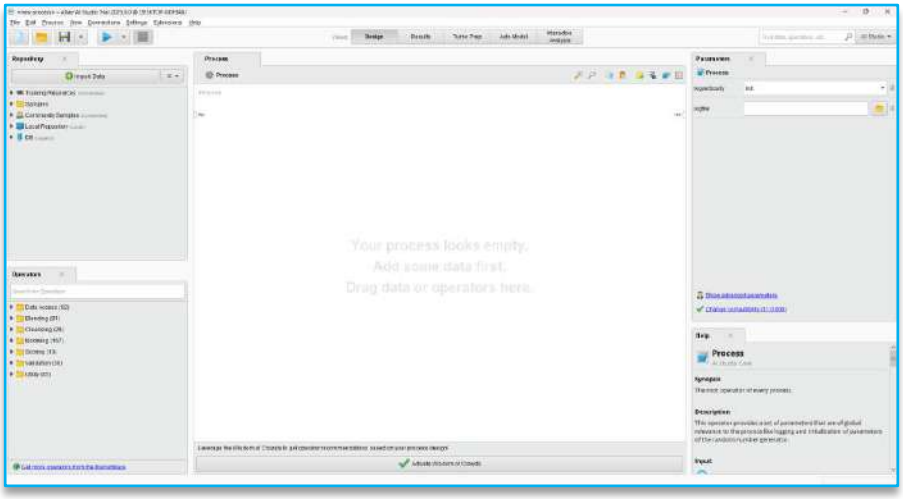
- **การจำแนกแบบไบนารี (binary classification)** คือ การจำแนกประเภทโดยคอมพิวเตอร์ ค่าที่เป็นไปได้จะมีคำตอบ 2 ค่า ส่วนใหญ่แล้วจะใช้ค่าแทนด้วย 1-0
- **ต้นไม้ตัดสินใจ (decision tree)** ฟีเจอร์ที่อยู่ด้านบนของต้นไม้มีความสำคัญ (ช่วยจำแนกข้อมูลได้ดีกว่า) มากกว่าฟีเจอร์ที่อยู่ด้านล่าง



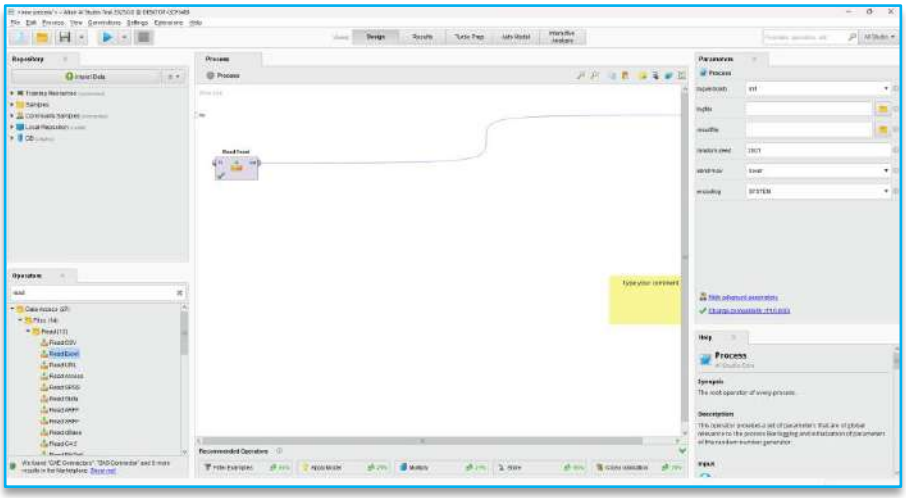
- การเลือกฟีเจอร์ที่สำคัญ (informative feature) คือ ฟีเจอร์ที่นำมาใช้จำแนกข้อมูลแล้วได้กลุ่มที่วัตถุอยู่ในคลาสเดียวกันหมดหรือเกือบหมด เกณฑ์ที่ใช้วัดความสำคัญของฟีเจอร์ Information gain คือ เลือกฟีเจอร์ที่นำมาแบ่งข้อมูลแล้วได้กลุ่มที่คล้ายกันมากกว่า (ยิ่งสูงยิ่งดี) Chi-square คือ เลือกฟีเจอร์ที่มีความสัมพันธ์กับตัวแปรเป้าหมายมากกว่า (ยิ่งสูงยิ่งดี)

โปรแกรม RapidMiner

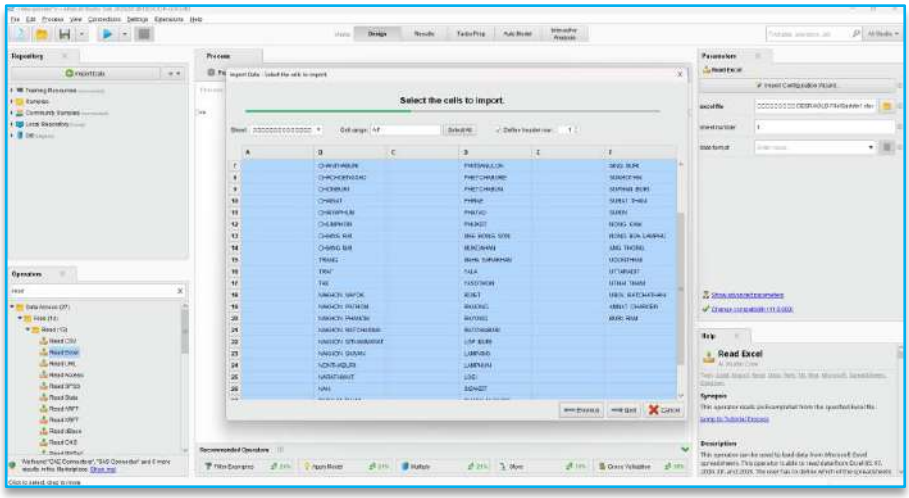
- ตัวอย่างหน้าจอแสดงผลโปรแกรม RapidMiner → blank process



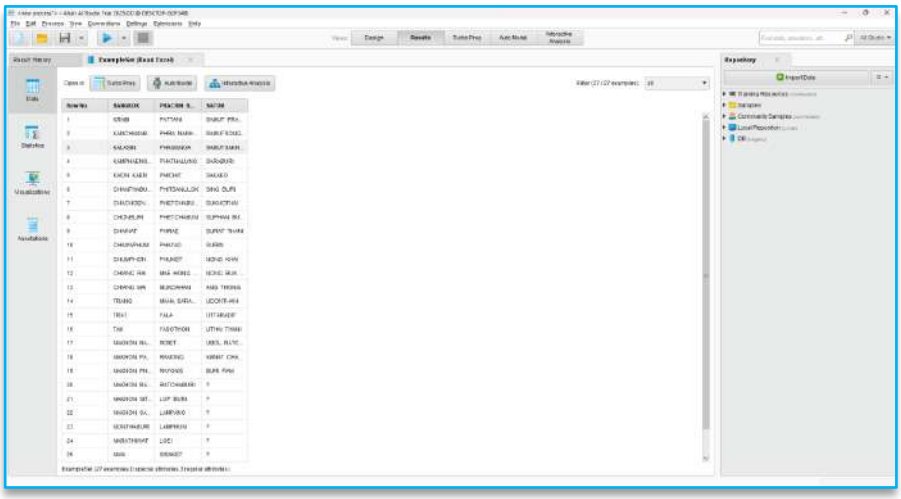
- การเลือกคำสั่งเพื่อแสดงผล Read excel เพื่ออ่านไฟล์ข้อมูล excel ที่จะนำมาใช้ในการวิเคราะห์



- การดึงไฟล์ข้อมูล excel ที่นำมาวิเคราะห์ โดยใช้คำสั่ง import configuration wizard



- การลากกล่อง out หรือ output เพื่อ run process เพื่อการแสดงผลลัพธ์ โดยผลลัพธ์ที่ได้จะแสดงในหน้า Results



- การตรวจสอบ statistics ของข้อมูล ลายละเอียด details ของข้อมูลก่อนที่จะใช้คำสั่งวิเคราะห์อื่นๆ