การรวบรวมและการได้มาซึ่งข้อมูล 2

ข้อมูลนับเป็นหัวใจสำคัญของการวิเคราะห์ข้อมูลด้วยเครื่องมือทางวิทยาการข้อมูล การรวบรวมและการได้มาของข้อมูลจึงเป็นขั้นตอนที่สำคัญ ข้อมูลที่จะนำมาวิเคราะห์จะต้องเป็นข้อมูลที่สามารถนำไปสู่การตอบปัญหาธุรกิจได้ อีกทั้งจะต้องถูกเตรียมให้อยู่ในรูปแบบที่เหมาะสม และข้อมูลต้องมีคุณภาพเพียงพอสำหรับการนำไปวิเคราะห์ ในบทนี้จะกล่าวถึงแหล่งกำเนิดข้อมูล รูปแบบการแทนข้อมูล การเตรียมข้อมูลเบื้องต้น คุณภาพและการปรับปรุงคุณภาพของข้อมูล

2.1 แหล่งกำเนิดข้อมูล

เราสามารถรวบรวมข้อมูลได้จากแหล่งกำเนิดข้อมูลหลายประเภท บางครั้งข้อมูลอาจได้มาจากการดำเนินการสำรวจหรือการจดบันทึกจากการสังเกต ในขณะที่ข้อมูลบางอย่างสามารถรวบรวมได้จากแหล่งเก็บข้อมูลที่เคยมีการสำรวจและรวบรวมไว้แล้ว แหล่งกำเนิดข้อมูล สามารถแบ่งตามวิธีการได้มาและคุณลักษณะการเข้าถึงข้อมูลได้ ดังนี้

แบบสอบถาม (Questionnaires)

การได้มาซึ่งข้อมูลด้วยการออกแบบสอบถาม เป็นวิธีการที่เป็นที่รู้จักและนำไปใช้อย่างแพร่หลายในการเก็บรวบรวมข้อมูล ข้อมูลที่ได้ถือเป็นข้อมูลชั้นปฐมภูมิ ที่มาจากแหล่งข้อมูลโดยตรง และไม่ผ่านกระบวนการปรุงแต่งข้อมูล จึงทำให้มีความน่าเชื่อถือสูง การบันทึกข้อมูลลงในแบบสอบถามนั้น สามารถทำได้ 2 วิธี คือ

1. **การบันทึกโดยเจ้าของข้อมูล** วิธีการนี้ผู้ออกแบบสอบถามทำหน้าที่กระจายแบบสอบถาม และผู้ให้ข้อมูลเป็นผู้กรอกข้อมูลลงในแบบสอบถาม ข้อมูลจึงมาจากเจ้าของข้อมูลโดยตรง ซึ่งมีความถูกต้องและน่าเชื่อถือสูง อย่างไรก็ตามข้อมูลที่ได้อาจมีความหลากหลายสูง โดยเฉพาะคำถามปลายเปิด ดังนั้น ผู้ออกแบบสอบถามจึงต้องออกแบบฟอร์มสำรวจที่ชัดเจน และอธิบายวิธีการกรอกข้อมูลให้แก่ผู้ให้ข้อมูล เพื่อให้ได้ข้อมูลที่ตรงกับความต้องการ และนำไปสู่การวิเคราะห์ข้อมูลโดยสะดวก
2. **การบันทึกโดยผู้สำรวจ** เป็นการกรอกข้อมูลลงแบบสอบถามโดยผู้ออกแบบสอบถาม ซึ่งข้อมูลที่ได้อาจมาจากการสัมภาษณ์ผู้ให้ข้อมูล หรือการสังเกต

ปรากฏการณ์ที่เกิดขึ้น วิธีการนี้ช่วยให้ข้อมูลที่ได้มามีความเป็นมาตรฐาน และยังสามารถใช้รวบรวมข้อมูลจากผู้ให้ข้อมูลที่ไม่สามารถกรอกข้อมูลได้โดยตรง ซึ่งอาจเป็นคน สัตว์ พืช หรือปรากฏการณ์ทางธรรมชาติ อย่างไรก็ตามผู้ออกแบบสอบถามควรต้องระมัดระวังการนำความเห็นของตนเองบันทึกลงเป็นข้อมูล ซึ่งอาจทำให้ข้อมูลที่ได้มาไม่ถูกต้องและตรงตามความเป็นจริง

ในการสร้างแบบสอบถาม ผู้ออกแบบสอบถามสามารถสร้างแบบสอบถามได้ทั้งในรูปแบบกระดาษ และแบบฟอร์มออนไลน์ ซึ่งปัจจุบันมีการพัฒนาเครื่องมือสร้างแบบสอบถามออนไลน์ เช่น Google Forms[1] และ Microsoft Forms[2] ที่คอยอำนวยความสะดวกในการออกแบบและกระจายแบบสอบถาม ผู้สนใจสามารถเลือกใช้เครื่องมือสร้างแบบสอบถามได้ตามความเหมาะสม

1: Google Forms
https://docs.google.com/forms

2: Microsoft Forms
https://forms.office.com/

เครื่องบริการเว็บ (Web Servers)

เวิลด์ไวด์เว็บ (World Wide Web: WWW) หรือ เว็บ (Web) เป็นระบบสารสนเทศที่เอกสาร และทรัพยากรอื่นๆ เช่น รูปภาพ วิดีโอ และเสียง สามารถเชื่อมโยงถึงกันได้ เว็บเป็นแหล่งข้อมูลขนาดใหญ่ มีพลวัต และความหลากหลายของข้อมูลสูง ข้อมูลของเว็บนั้นถูกเก็บไว้ในเครื่องบริการเว็บ (Web Server) กลุ่มของเอกสารและทรัพยากรเว็บที่มีชื่อโดเมน (Domain Name) ร่วมกันถูกเก็บไว้ในเครื่องบริการเว็บเครื่องเดียวกัน เรียกว่า เว็บไซต์ (Website) โดยปกติแล้วเอกสารและทรัพยากรเว็บจะถูกแสดงผลผ่านเว็บบราวเซอร์ (Web Browser) โดยเข้าถึงผ่านระบบอินเทอร์เน็ต อย่างไรก็ตามเอกสารและทรัพยากรที่ถูกจัดเก็บบนเครื่องบริการเว็บ ยังสามารถเข้าถึงผ่านทางช่องทางอื่นๆ ได้ เช่น โปรโตคอลการโอนย้ายไฟล์ (File Transfer Protocol: FTP) และเข้าถึงโดยตรงจากเครื่องบริการเว็บ ซึ่งจำกัดเฉพาะผู้มีสิทธิเข้าถึงข้อมูลเท่านั้น

บริการผ่านเว็บ (Web Services)

เว็บไซต์บางประเภทถูกพัฒนาขึ้นเพื่อเป็นแหล่งบริการข้อมูลที่มีความจำเพาะ เช่น ข้อมูลสภาพภูมิอากาศ ภาพถ่ายดาวเทียม ตลาดหุ้น และเอกสารวิชาการ เป็นต้น โดยมีช่องทางเฉพาะให้สามารถเข้าถึง และดาวน์โหลดข้อมูลที่ให้บริการได้ เช่น การเข้าถึงและดาวน์โหลดจากหน้าเว็บเพจ (Web Page) และการเข้าถึงผ่านส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) ข้อมูลที่ได้จากบริการผ่านเว็บนี้มักเป็นข้อมูลชั้นทุติยภูมิ ที่มีการจัดระเบียนและเป็นมาตรฐาน ซึ่งง่ายต่อการนำไปใช้ประโยชน์ในการวิเคราะห์ต่อไป

ฐานข้อมูล (Databases)

ข้อมูลที่ถูกสำรวจและรวบรวมไว้มักถูกจัดเก็บในฐานข้อมูล ซึ่งมีการจัดระเบียบข้อมูลเป็นอย่างดี ในอดีตฐานข้อมูลอยู่ในรูปแบบสมุดทะเบียน จนกระทั่งมีการพัฒนาให้อยู่ในรูปแบบอิเล็กทรอนิกส์ ที่ถูกจัดเก็บในระบบคอมพิวเตอร์ มีทั้งแบบออฟไลน์ ซึ่งมักจัดเก็บข้อมูลที่ใช้เฉพาะภายในองค์กรเท่านั้น และแบบออนไลน์ ซึ่งจัดเก็บข้อมูลที่ใช้งานร่วมกันหลายหน่วยงาน อาจให้บริการข้อมูลแก่สาธารณะผ่านบริการผ่านเว็บด้วย การเข้าถึงข้อมูลจากฐานข้อมูลโดยตรงจะต้องใช้ความรู้และเทคนิคขั้นสูง ซึ่งต้องอาศัยนักพัฒนาโปรแกรมฐานข้อมูลในการสร้างการสอบถาม (Query) ข้อมูลที่ต้องการ

คลังเก็บออนไลน์ (Online Repositories)

คลังเก็บออนไลน์มีลักษณะคล้ายกับบริการผ่านเว็บ แต่ต่างกันโดยที่ข้อมูลที่สามารถดาวน์โหลดได้นั้น เป็นข้อมูลที่บุคคลอื่นนำเข้าสู่คลังเก็บข้อมูล ผู้ใช้บริการคลังเก็บออนไลน์ให้บริการเฉพาะพื้นที่เก็บและช่องทางการเข้าถึงข้อมูลเท่านั้น ปัจจุบันมีผู้ให้บริการคลังเก็บออนไลน์จำนวนมาก มีทั้งบริการคลังเก็บออนไลน์ส่วนบุคคล เช่น Google Drive[3] และ Dropbox[4] เป็นต้น ซึ่งการเข้าถึงข้อมูลจะต้องได้รับอนุญาตจากเจ้าของข้อมูล และคลังเก็บออลไน์ข้อมูลเปิด ซึ่งเมื่อเจ้าของข้อมูลนำข้อมูลเข้าสู่คลังเก็บออนไลน์แล้ว ข้อมูลนั้นจะสามารถเข้าถึงได้ทั่วไป ข้อมูลจากคลังเก็บออนไลน์นั้นมักมีรูปแบบไม่เป็นมาตรฐาน ซึ่งขึ้นอยู่กับเจ้าของข้อมูลและข้อกำหนดของผู้ให้บริการ ดังนั้น การนำข้อมูลจากคลังเก็บออนไลน์มาใช้งาน จะต้องศึกษาและทำความเข้าใจโครงสร้างการจัดเก็บข้อมูล เพื่อให้สามารถดึงเอาข้อมูลมาวิเคราะห์ได้อย่างถูกต้อง

3: Google Drive
https://drive.google.com/

4: Dropbox
www.dropbox.com

แนะนำคลังเก็บออนไลน์

UC Irvine Machine Learning Repository:
https://archive.ics.uci.edu/
Kaggle Datasets:
www.kaggle.com/datasets
ศูนย์กลางข้อมูลเปิดภาครัฐ:
https://data.go.th/
ศูนย์กลางข้อมูลเปิดสถาบันสารสนเทศทรัพยากรน้ำ (องค์การมหาชน):
https://opendata.hii.or.th/

แฟ้มลงบันทึกข้อมูลออก (Log Files)

แฟ้มลงบันทึกข้อมูลออก คือ ระเบียนของเหตุการณ์ที่เกิดขึ้นในระบบคอมพิวเตอร์ มักจะมีการบันทึกเหตุการณ์หรือการดำเนินงานที่เกิดขึ้นในระบบปฏิบัติการ ซอฟต์แวร์ หรือการสื่อสารโดยอัตโนมัติ เพื่อประโยชน์ในการวิเคราะห์ระบบและการกู้คืนสถานะเมื่อเกิดข้อผิดพลาด ข้อมูลแฟ้มลงบันทึกข้อมูลออกนี้ ถือเป็นแหล่งข้อมูลหนึ่งที่สามารถนำมาใช้ประโยชน์ได้ โดยเฉพาะงานทางด้านระบบคอมพิวเตอร์และการสื่อสารข้อมูล

2.2 การแทนข้อมูล

เมทริกซ์ข้อมูล (Data Matrix)

ในการเตรียมข้อมูลสำหรับการวิเคราะห์ ข้อมูลสามารถถูกแทนในรูป *เมทริกซ์ข้อมูล (Data Matrix)* ขนาด $n \times d$ เมื่อ n และ d คือ จำนวนแถวและคอลัมน์ ตามลำดับ โดยคอลัมน์แสดงค่าของตัวแปรหรือลักษณะเด่นที่สนใจ ข้อมูลหนึ่งข้อมูลจะถูกบันทึกค่าตัวแปรแต่ละตัวแปรที่สนใจไว้ในแต่ละคอลัมน์ของหนึ่งแถวของเมทริกซ์ เมทริกซ์ข้อมูลสามารถแสดงได้ ดังนี้

$$\mathbf{D} = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_d \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{pmatrix} \end{array}$$

เมื่อ $\mathbf{x}_i$ แสดงถึงข้อมูลลำดับที่ i ซึ่งสามารถเขียนได้ ดังนี้

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$$

และ X_j แสดงถึงค่าตัวแปรที่ j ของข้อมูลทุกข้อมูล ซึ่งสามารถเขียนได้ ดังนี้

$$X_j = (x_{1,j}, x_{2,j}, ..., x_{n,j})$$

ข้อมูลแต่ละแถวอาจถูกเรียกว่า เอนทิตี (Entity) กรณีตัวอย่าง (Instance) ตัวอย่าง (Example) ระเบียน (Record) รายการ (Transaction) วัตถุ (Object) จุดข้อมูล (Data Point) เวกเตอร์ลักษณะเด่น (Feature Vector) ทูเพิล (Tuple) หรืออื่นๆ ในขณะเดียวกันแต่ละคอลัมน์อาจถูกเรียกว่า แอตทริบิวต์ (Attribute) คุณสมบัติ (Property) ลักษณะเด่น (Feature) มิติ (Dimension) ตัวแปร (Variable) เขตข้อมูล (Field) หรืออื่นๆ ทั้งนี้ขึ้นอยู่กับแนวทางการประยุกต์ใช้

เมทริกซ์ข้อมูลยังสามารถแสดงในมุมมองของตารางข้อมูล (Data Table) โดยแถวกับคอลัมน์ของเมทริกซ์ข้อมูลและตารางข้อมูลมีความสอดคล้องกัน ตาราง 2.1 แสดงตัวอย่างข้อมูลตัวละครจากหนังสือการ์ตูน Marvel จากชุดข้อมูล Marvel Wikia [4] โดยแต่ละแถวจัดเก็บค่าคุณลักษณะหรือตัวแปรของตัวละครหนึ่งตัว และแต่ละคอลัมน์แทนตัวแปรที่สนใจ เราสามารถเขียนข้อมูลตัวละคร Spider-Man ในรูปทูเพิล ได้ ดังนี้

[4] ***Comic Characters***. 2015. url: https://github.com/fivethirtyeight/data/tree/master/comic-characters

$\mathbf{x}_2$ = (1678, Spider-Man (Peter Parker), Secret, Good, Hazel, Male, N/A, Living, ..., 1962)

ตาราง 2.1: ตัวอย่างข้อมูลตัวละครจากหนังสือการ์ตูน Marvel จากชุดข้อมูล Marvel Wikia (ที่มา: https://github.com/fivethirtyeight/data/blob/master/comic-characters/marvel-wikia-data.csv)

	Page ID X_1	Name X_2	ID X_3	Align X_4	Eye X_5	Hair X_6	Sex X_7	Gsm X_8	Alive X_9	...	Year X_d
$\mathbf{x}_1$	1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	Brown	Male	N/A	Living	...	1962
$\mathbf{x}_2$	7139	Captain America (Steven Rogers)	Public	Good	Blue	White	Male	N/A	Living	...	1941
...	...	...	...	...	...	...	...	...	...	...	...
$\mathbf{x}_n$	321461	Oyakata (Earth-616)	Secret	Neutral	Brown	Black	Male	N/A	Living	...	1998

ชนิดข้อมูล (Data Types)

ชนิดข้อมูลของตัวแปร เป็นอีกสิ่งหนึ่งที่จะต้องพิจารณา เนื่องการดำเนินการ (Operation) ของข้อมูลแต่ละชนิดมีนิยามที่แตกต่างกัน เช่น การบวกค่าตัวแปรชนิดตัวเลขคือการนำค่าของข้อมูลมารวมกัน แต่การดำเนินการบวกของตัวแปรชนิดข้อความอาจหมายถึงการนำข้อความมาเชื่อมกัน การวิเคราะห์ข้อมูลด้วยเครื่องมือทางวิทยาการข้อมูลล้วนมีพื้นฐานบนการดำเนินการทางคณิตศาสตร์ ดังนั้น เมื่อทำการแทนข้อมูลในรูปแบบเมทริกซ์ข้อมูล หรือตารางข้อมูลแล้ว จึงจำเป็นต้องกำหนดชนิดข้อมูลของแต่ละตัวแปรให้เหมาะสมด้วย เราสามารถจำแนกชนิดข้อมูลออกเป็น 2 ประเภทหลัก ได้ดังนี้

ข้อมูลเชิงคุณภาพ (Qualitative Data)

ข้อมูลเชิงคุณภาพ คือ ข้อมูลที่ไม่สามารถอธิบายได้ด้วยตัวเลข และการดำเนินการทางคณิตศาสตร์พื้นฐาน โดยปกติแล้วมักอยู่ในรูปของภาษาธรรมชาติ (Natural Language) ข้อมูลเชิงคุณภาพสามารถแบ่งออกได้เป็น 2 ประเภท ดังนี้

1) ข้อมูลสายอักขระ (String Data)

สายอักขระ หรือ ข้อความ คือ กลุ่มของอักขระที่เรียงต่อกันเพื่อสื่อความหมาย หรือค่าของข้อมูล ที่เป็นไปได้จำนวนมาก หรือค่าที่เป็นไปได้เป็นเซตอนันต์ (Infinite Set) เช่น ชื่อของบุคคล รหัสนักศึกษา เบอร์โทรศัพท์ บทความ และข้อความคิดเห็น เป็นต้น

2) ข้อมูลเชิงกลุ่ม (Categorical Data)

ข้อมูลเชิงกลุ่ม คือ ข้อความบ่งบอกประเภทหรือหมวดหมู่ ค่าของข้อมูลมีจำนวนจำกัด หรือค่าที่เป็นไปได้เป็นเซตจำกัด (Finite Set) เช่น ชื่อคณะในมหาวิทยาลัยเชียงใหม่ ชื่อประเทศ สี เพศ วุฒิการศึกษา และชนิดของสกุลพืช เป็นต้น ข้อมูลเชิงกลุ่มยังสามารถแบ่งตามมาตรการวัดได้ 2 ประเภท คือ

1. **ข้อมูลนามบัญญัติ (Nominal Data)** เป็นข้อมูลแสดงกลุ่มหรือประเภทที่สามารถนำค่าข้อมูลมาเปรียบเทียบความเหมือนหรือแตกต่างกันได้ แต่ไม่สามารถเรียงลำดับค่าของข้อมูลได้ ตัวอย่างเช่น
   - เพศ = {ชาย, หญิง}
   - สีผม = {ดำ, น้ำตาล, แดง, ขาว}
   - สถานะ = {โสด, สมรส, หย่าร้าง}

   พิจารณาตัวแปรเพศจากตัวอย่างซึ่งมีค่าที่เป็นไปได้ 2 ค่า คือ ชาย และหญิง สมมติให้นักศึกษา 3 คน นักศึกษาคนที่ 1 และ 2 เป็นเพศชาย และนักศึกษาคนที่ 3 เป็นเพศหญิง เราสามารถกล่าวได้ว่า นักศึกษาคนที่ 1 และ 2 มีค่าตัวแปรเพศเหมือนกัน ส่วนนักศึกษาคนที่ 1 และ 3 มีค่าตัวแปรเพศต่างกัน แต่ไม่สามารถกล่าวได้ว่า นักศึกษาคนที่ 1 มีค่าตัวแปรเพศมากกว่าหรือน้อยกว่า นักศึกษาคนที่ 3
2. **ข้อมูลเชิงอันดับ (Ordinal Data)** เป็นข้อมูลแสดงกลุ่มหรือประเภท ที่สามารถนำค่าข้อมูลมาเปรียบเทียบความเหมือนหรือแตกต่าง และเรียงลำดับค่าของข้อมูลได้ แต่ไม่สามารถวัดความแตกต่างระหว่างค่าข้อมูลได้ว่ามีค่ามากน้อยเพียงใด ตัวอย่างเช่น
   - ระดับการศึกษา = {ปฐมศึกษา, มัธยม, อนุปริญญา, ปริญญา}
   - ความพึงพอใจ = {พอใจ, ปานกลาง, ไม่พอใจ}

   พิจารณาตัวแปรความพึงพอใจจากตัวอย่างซึ่งมีค่าที่เป็นไปได้ 3 ค่าคือ พอใจ ปานกลาง และไม่พอใจ สมมติให้ลูกค้าผู้มาใช้บริการธนาคาร 2 คน โดยลูกค้าคนที่ 1 มีความพึงพอใจในการให้บริการระดับปานกลาง ส่วนลูกค้าคนที่ 2 ไม่พึงพอใจในการให้บริการ เราสามารถกล่าวได้ว่าลูกค้าคนที่ 1 มีความพึงพอใจในการให้บริการมากกว่าลูกค้าคนที่ 2 แต่ไม่สามารถระบุได้ว่า คนที่ 1 มีความพึงพอใจในการให้บริการมากกว่าลูกค้าคนที่ 2 เท่าใด

ตัวอย่าง 2.1

ข้อมูลตัวละครจากหนังสือการ์ตูน Marvel ซึ่งแสดงตัวอย่างข้อมูลในตาราง 2.1 **ตัวแปร Name มีชนิดข้อมูลเป็นข้อมูลสายอักขระ** เนื่องจากค่าที่เป็นไปได้ของตัวแปร ซึ่งเป็นชื่อของตัวละครมีจำนวนมาก และหากมีตัวละครใหม่เกิดขึ้น ค่าที่เป็นไปได้ของตัวแปร Name จะมีเพิ่มขึ้นตามไปด้วย จึงทำให้ค่าที่เป็นไปได้ของตัวแปร Name เป็นเซตอนันต์ นอกจากนี้เนื่องจาก

ค่าตัวแปร Page ID เป็นตัวระบุ (Identifier) ของตัวละครแต่ละตัว จึงทำให้ค่าตัวแปร Page ID ของตัวละครแต่ละตัวมีค่าไม่ซ้ำกัน ดังนั้นค่าที่เป็นไปได้ของตัวแปร Page ID จึงสามารถเพิ่มขึ้นตามจำนวนตัวละครที่เพิ่มขึ้น นอกจากนี้ แม้ว่าลักษณะข้อมูลจะเป็นตัวเลข แต่เราไม่สามารถใช้การดำเนินการทางคณิตศาสตร์พื้นฐานกับค่าตัวแปรนี้อย่างมีความหมายได้ จึงสรุปได้ว่า **ตัวแปร Page ID มีชนิดข้อมูลเป็นข้อมูลสายอักขระ**

ส่วน **ตัวแปร ID Sex และ Alive มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ** เนื่องจากตัวแปรทั้ง 3 ตัวแปรมีค่าที่เป็นไปได้อยู่ในเซตจำกัด นั้นคือ

- ID = {Known to Authorities, No Dual, Public, Secret}
- Sex = {Agender, Female, Genderfluid, Male}
- Alive = {Deceased, Living}

และค่าที่เป็นไปได้ของตัวแปรเหล่านี้ไม่สามารถเรียงลำดับอย่างมีความหมายได้

ตัวแปร Align มีชนิดข้อมูลเป็นข้อมูลเชิงอันดับ เนื่องจากมีค่าที่เป็นไปได้อยู่ในเซต {Good, Neutral, Bad} และค่าของตัวแปรที่เป็นไปได้ทั้ง 3 ค่า สามารถเรียงลำดับตามพฤติกรรมได้โดย Good > Neutral > Bad หรือ Good < Neutral < Bad ก็ได้

ข้อมูลเชิงปริมาณ (Quantitative Data)

ข้อมูลเชิงปริมาณ หรือ ข้อมูลเชิงตัวเลข (Numerical Data) คือ ข้อมูลที่สามารถอธิบายได้ด้วยตัวเลข และสามารถใช้การดำเนินการทางคณิตศาสตร์พื้นฐานบนค่าของข้อมูลได้ ข้อมูลเชิงปริมาณ สามารถแบ่งตามมาตรการวัดได้ 2 ประเภท คือ

1. **ข้อมูลระดับอันตรภาค (Interval-scaled Data)** เป็นข้อมูลที่มีช่วงห่างหรือระยะห่างเท่าๆ กัน และสามารถวัดค่าระยะห่างได้ แต่เป็นข้อมูลที่ไม่มีศูนย์แท้ จึงไม่สามารถนำไปเปรียบเทียบในลักษณะสัดส่วนได้ ตัวอย่าง เช่น
   - อุณหภูมิ ในหน่วยวัดองศาเซลเซียส (°C) หรือองศาฟาเรนไฮต์ (°F)
   - มาตรความเป็นกรด-เบส (pH meter)
   - คะแนนสอบ SAT ซึ่งมีค่าอยู่ระหว่าง 200-800

   พิจารณาตัวแปรอุณหภูมิในหน่วยองศาเซลเซียส (°C) สมมติว่าวันนี้มีอุณหภูมิเฉลี่ย 20°C และวันก่อนหน้ามีอุณหภูมิเฉลี่ย 10°C เราสามารถกล่าวได้ว่า อุณหภูมิเฉลี่ยของวันนี้ลดลงจากอุณหภูมิเฉลี่ยวันก่อนหน้า

10°C อย่างไรก็ตาม เราไม่สามารถสรุปได้ว่า วันนี้อากาศหนาวกว่าวันก่อนหน้าเป็น 2 เท่า

2. **ข้อมูลระดับอัตราส่วน (Ratio-scaled Data)** เป็นข้อมูลที่มีคุณสมบัติเช่นเดียวกับข้อมูลระดับอันตรภาค แต่ข้อมูลระดับอัตราส่วนเป็นข้อมูลที่มีศูนย์แท้ (Natural Zero) หรือ ศูนย์สัมบูรณ์ (Absolute Zero) จึงทำให้ข้อมูลประเภทนี้สามารถนำไปเปรียบเทียบในลักษณะสัดส่วนได้ ตัวอย่าง เช่น
   - อุณหภูมิ ในหน่วยเคลวิน (Kelvin: K)
   - อายุ
   - ความสูง
   - รายได้

   พิจารณาตัวแปรอายุ สมมติว่านายเอ มีอายุ 20 ปี และนายบี มีอายุ 10 ปี เราสามารถกล่าวได้ว่า นายเอมีอายุมากกว่านายบี 10 ปี และ นายเอมีอายุเป็น 2 เท่าของนายบี

เมื่อพิจารณาชนิดข้อมูลที่แบ่งตามมาตรการวัดแล้ว สามารถเปรียบเทียบคุณลักษณะ และความสามารถในการจัดการข้อมูลแต่ละชนิดได้ ดังตาราง 2.2

ตาราง 2.2: คุณลักษณะและความสามารถในการดำเนินการของแต่ละชนิดข้อมูล

คุณลักษณะและความสามารถในการจัดการ	ชนิดข้อมูล			
	ข้อมูลเชิงคุณภาพ		ข้อมูลเชิงปริมาณ	
	ข้อมูลนามบัญญัติ	ข้อมูลเชิงลำดับ	ข้อมูลระดับอัตรภาค	ข้อมูลระดับอัตราส่วน
เปรียบเทียบความเท่ากันได้	✓	✓	✓	✓
เรียงลำดับของค่าข้อมูลได้		✓	✓	✓
นับความถี่ได้	✓	✓	✓	✓
หาค่าฐานนิยม (Mode) ได้	✓	✓	✓	✓
หาค่ามัธยฐาน (Median) ได้		✓	✓	✓
หาค่าเฉลี่ย (Mean) ได้			✓	✓
วัดระยะห่างระหว่างค่าข้อมูลได้			✓	✓
สามารถดำเนินการบวกหรือลบค่าข้อมูลได้			✓	✓
สามารถดำเนินการคูณและหารค่าข้อมูลได้				✓
มีค่าศูนย์แท้				✓

ตัวอย่าง 2.2

ข้อมูลดอก Iris จากชุดข้อมูล Fisher Iris [5] ประกอบด้วย 4 ตัวแปร คือ ความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) ความกว้างของกลีบดอก

[5] Penny Analytics. ***Fisher Iris Data***. url: https://pennyanalytics.com/free-trial/

(Petal Width) และ ชนิดพันธุ์ (Species) แสดงตัวอย่างข้อมูล ดังตารางต่อไปนี้

ตาราง 2.3: ตัวอย่าง ข้อมูล ดอก Iris จาก ชุด ข้อมูล Fisher Iris (ที่มา: https://pennyanalytics.com/free-trial/)

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
...	...	...	...	...
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
...	...	...	...	...
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

ตัวแปร ความยาวของกลีบเลี้ยง ความกว้างของกลีบเลี้ยง ความยาวของกลีบดอก และ ความกว้างของกลีบดอก (หน่วย: เซนติเมตร) จัดเป็นข้อมูลระดับอัตราส่วน เนื่องจากเป็นข้อมูลเชิงตัวเลข (Numerical Data) แสดงความขนาดของกลีบเลี้ยงและกลีบดอก จึงมีค่าศูนย์แท้

ส่วน **ตัวแปร ชนิดพันธุ์ (Species) มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ** เพราะมีค่าที่เป็นไปได้อยู่ในเซต {setosa, versicolor, virginica} และค่าที่เป็นไปได้ทั้ง 3 ค่า ไม่สามารถนำมาเรียงลำดับอย่างมีความหมายได้

2.3 การเตรียมข้อมูล

การเข้ารหัสข้อมูลเชิงกลุ่ม

เนื่องจากวิธีการ หรือเครื่องมือในการวิเคราะห์ข้อมูลทางด้านวิทยาการข้อมูลส่วนใหญ่ ถูกนำเสนอขึ้นบนฐานการดำเนินการเชิงตัวเลข ดังนั้น ค่าตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงกลุ่มจึงต้องถูกเข้ารหัส (Encoding) ในที่นี้จะกล่าววิธีการเข้ารหัสพื้นฐานสำหรับข้อมูลนามบัญญัติ และข้อมูลเชิงอันดับ 2 วิธีการ ดังต่อไปนี้

การเข้ารหัสแบบวัน-ฮอท (One-hot Encoding)

การเข้ารหัสแบบวัน-ฮอท เป็นวิธีการเข้ารหัสสำหรับข้อมูลนามบัญญัติ โดยใช้การแทนข้อมูลแบบทวิภาค (Binary) แทนแต่ละค่าที่เป็นไปได้ของตัวแปร ดังนั้น จากตัวแปร 1 ตัวแปร จะถูกขยายจำนวนตัวแปรเท่ากับจำนวนค่าที่เป็นไปได้ของตัวแปรนั้น ค่าที่แทนลงในตัวแปรที่ขยายมานี้ จะมีค่าเป็น 0 ทั้งหมด ยกเว้นในตัวแปรที่สัมพันธ์กับค่าตัวแปรตั้งต้น จะมีค่าเป็น 1 เพื่อให้เข้าใจมากยิ่งขึ้นจะขอยกตัวอย่างดังต่อไปนี้

ตัวอย่าง 2.3

จากตัวแปร Sex ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 4 ค่า คือ Agender Female Genderfluid และ Male สามารถทำการเข้ารหัสแบบวัน-ฮอท โดยขยายตัวแปรออกเป็น 4 ตัวแปร ได้ดังนี้

Sex	isAgender	isFemale	isGenderfluid	isMale
Agender	1	0	0	0
Female	0	1	0	0
Genderfluid	0	0	1	0
Male	0	0	0	1

สังเกตได้ไว้ แต่ละตัวแปรใหม่ที่ขยายออกออกมานั้น มีความสัมพันธ์กับตัวแปรตั้งต้น โดยค่าที่เป็นไปได้ 1 ค่าจะถูกสร้างตัวแปรใหม่ 1 ตัวแปรขึ้นมาแทน และตัวแปรที่ขยายออกมานี้จะมีค่าที่เป็นไปได้ 0 หรือ 1 เท่านั้น และจะต้องมีเพียงตัวแปรเดียวเท่านั้นที่สามารถมีค่าเท่ากับ 1 ได้ โดย

- ถ้าตัวแปรใหม่ isAgender มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Agender
- ถ้าตัวแปรใหม่ isFemale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Female
- ถ้าตัวแปรใหม่ isGenderfluid มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Genderfluid
- ถ้าตัวแปรใหม่ isMale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Male

ดังนั้น ข้อมูล Spider-Man มีค่าตัวแปร Sex เท่ากับ Male เมื่อทำการเข้ารหัสแล้ว ตัวแปร isMale ซึ่งขยายมาจากตัวแปร Sex เดิม จะมีค่าเท่ากับ 1 ส่วนตัวแปร isAgender isFemale และ isGenderfluid จะมีค่าเท่ากับ 0

การเข้ารหัสเชิงลำดับ (Ordinal Encoding)

การเข้ารหัสเชิงลำดับ เป็นวิธีการเข้ารหัสสำหรับข้อมูลเชิงอันดับ โดยการให้ค่าตัวเลขตามลำดับของค่าข้อมูล แทนการใช้ข้อความ ดังนั้น เมื่อข้อมูลถูกเข้ารหัสแล้วความหมายเชิงลำดับของข้อมูลจะยังคงถูกสงวนไว้ ศึกษาวิธีการการเข้ารหัสเชิงลำดับได้ ดังตัวอย่างต่อไปนี้

ตัวอย่าง 2.4

จากตัวแปร Align ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 3 ค่า คือ Good Neutral และ Bad โดย กำหนดให้ลำดับของข้อมูลเป็น Good > Neutral > Bad สามารถทำการเข้าเชิงลำดับ ได้ดังนี้

Align	Encoded Align
Bad	1
Neutral	2
Good	3

สังเกตได้ว่า การเข้ารหัสเชิงอันดับ ไม่ทำให้เกิดการเพิ่มตัวแปรในชุดข้อมูล เพียงแต่เปลี่ยนการแทนค่าข้อมูลเชิงอันดับ จากข้อความเป็นตัวเลขเท่านั้น โดยการให้ค่าตัวเลขสัมพันธ์กับลำดับของค่าข้อมูล

อย่างไรก็ตาม การให้ค่าตัวเลขไม่ได้จำกัดให้เริ่มต้นจากค่า 1 เท่านั้น สามารถเลือกใช้ค่าตัวเลขใดก็ได้ แต่ยังคงต้องรักษาลำดับของค่าข้อมูลให้คงอยู่ ตัวอย่างเช่น

Align	Encoded Align
Bad	-1
Neutral	0
Good	1

การแปลงค่าข้อมูลเชิงปริมาณ

การแปลงค่าข้อมูลเชิงปริมาณ ในบางครั้งสามารถช่วยให้แบบจำลองสำหรับการวิเคราะห์ข้อมูล มีความแม่นยำมากยิ่งขึ้น วิธีการแปลงค่าข้อมูลเชิงตัวเลขที่เป็นที่นิยมใช้กันอย่างแพร่หลาย คือ การทำให้เป็นบรรทัดฐาน (Normalization) และการทำให้เป็นมาตรฐาน (Standardization) มีรายละเอียดของเทคนิควิธีการ ดังต่อไปนี้

การทำให้เป็นบรรทัดฐาน (Normalization)

การทำให้เป็นบรรทัดฐาน คือ การแปลงค่าข้อมูลเชิงตัวเลขให้อยู่ในช่วง $[0, 1]$ โดยไม่บิดเบือนการกระจายของข้อมูล กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, ..., x_{n,j})$ เป็นชุดของค่าตัวแปร X_j ของทุกๆ ข้อมูล เราสามารถแปลงค่า $x_{i,j} \in X_j$ ให้อยู่ในช่วง $[0, 1]$ ได้ ด้วยเทคนิค Min-max Normalization โดยใช้สมการ ต่อไปนี้

$$x_{i,j}^{\text{norm}} = \frac{x_{i,j} - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{2.1}$$

เมื่อ $x_{i,j}^{\text{norm}}$ คือ ค่า $x_{i,j}$ ที่ถูกทำให้เป็นบรรทัดฐานแล้ว $\min(X_j)$ คือ ค่าที่น้อยที่สุดในชุดของค่าตัวแปร X_j และ $\max(X_j)$ คือ ค่าที่มากที่สุดในชุดของค่าตัวแปร X_j ค่าที่น้อยที่ใน X_j จะถูกแปลงค่าให้มีค่าใหม่เท่ากับ 0 ในขณะที่ค่าที่มากที่สุดใน X_j จะมีค่าที่ถูกแปลงค่าแล้วเท่ากับ 1

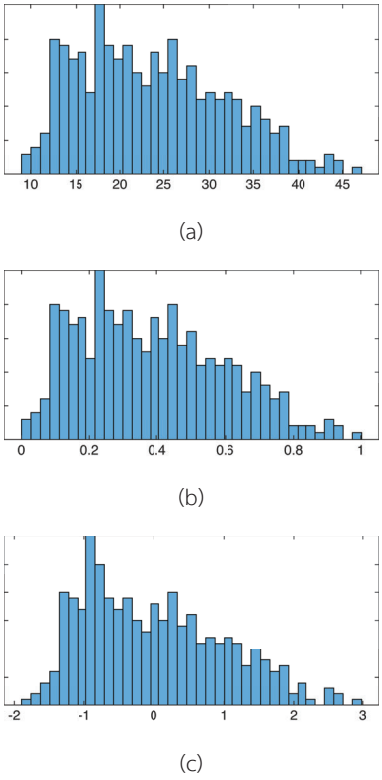
ตัวอย่าง 2.5

สมมติให้ ชุดข้อมูลหนึ่ง ประกอบด้วยข้อมูลทั้งหมด 5 ข้อมูล และแต่ละข้อมูลมีค่าตัวแปรอายุ (Age) ดังนี้

25, 32, 21, 45, 19

สามารถแปลงค่าข้อมูลของตัวแปรอายุ ด้วยเทคนิค Min-max Normalization โดยพิจารณาค่าที่มากที่สุด คือ 45 และค่าที่น้อยที่สุด คือ 19 จะได้ค่าตัวแปรอายุของแต่ละข้อมูลที่ถูกทำให้เป็นบรรทัดฐานแล้ว ดังนี้

Age	Normalized Age
25	$\frac{25-19}{45-19} = 0.23$
32	0.50
21	0.08
45	1.00
19	0.00



ภาพ 2.1: ตัวอย่าง ผล ลัพธ์ จาก การ แปลง ค่า ข้อมูล ด้วย การ ทำให้ เป็น บรรทัดฐาน และ การ ทำให้ เป็น มาตรฐาน (a) ข้อมูล ดั้งเดิม (b) ผลลัพธ์ จาก การ ทำให้ เป็น บรรทัดฐาน ด้วย เทคนิค Min-max Normalization (c) ผลลัพธ์จากการทำให้เป็นมาตรฐาน

การทำให้เป็นมาตรฐาน (Standardization)

การทำให้เป็นมาตรฐาน คือ การแปลงค่าข้อมูลเชิงตัวเลขให้จุดศูนย์กลางของข้อมูลมีค่าเท่ากับ 0 และ ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) มีค่าเท่ากับ 1 เราสามารถแปลงค่า $x_{i,j} \in X_j$ ให้เป็นมาตรฐาน โดยใช้สมการ ต่อไปนี้

$$x_{i,j}^{\text{stand}} = \frac{x_{i,j} - \text{mean}(X_j)}{\text{std}(X_j)} \tag{2.2}$$

เมื่อ $x_{i,j}^{\text{stand}}$ คือ ค่า $x_{i,j}$ ที่ถูกทำให้เป็นมาตรฐานแล้ว ซึ่งมักถูกเรียกว่า คะแนนมาตรฐานซี (Z-score) $\text{mean}(X_j)$ คือ ค่าเฉลี่ยน (Mean) บนชุดของค่าตัวแปร X_j และ $\text{std}(X_j)$ คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูลในชุดของค่าตัวแปร X_j การทำให้เป็นมาตรฐานมีสมมติฐานว่า ค่าข้อมูลมีลักษณะการกระจายแบบปกติ (Normal Distribution) อย่างไรก็ตาม ข้อมูลที่มีลักษณะการกระจายแบบอื่นยังคงสามารถใช้การทำให้เป็นมาตรฐานในการแปลงค่าข้อมูลได้

ตัวอย่าง 2.6

จากข้อมูลอายุ ในตัวอย่าง 2.5 สามารถคำนวณค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานได้ ดังนี้

ค่าเฉลี่ย = 28.4

ส่วนเบี่ยงเบนมาตรฐาน = 10.53

สามารถแปลงค่าข้อมูลของตัวแปรอายุ ด้วยการทำให้เป็นมาตรฐานได้ดังนี้

Age	Standardized Age
25	$\frac{25-28.4}{10.53}$ = -0.32
32	0.34
21	-0.70
45	1.58
19	-0.89

คุณภาพข้อมูล

สิ่งสำคัญประการหนึ่งของการวิเคราะห์ข้อมูล คือ คุณภาพข้อมูล เทคนิควิธีการวิเคราะห์ส่วนใหญ่ถูกนำเสนอขึ้นภายใต้สมมติฐานว่าข้อมูลที่นำมาวิเคราะห์มีคุณภาพดี ข้อมูลที่มีคุณภาพควรมีคุณลักษณะ ต่อไปนี้

1) ความถูกต้องและแม่นยำ (Accuracy and Precision)

ความถูกต้อง หมายถึง ค่าของตัวแปรต่างๆ ที่รวบรวมได้จะต้องเป็นไปตามความจริง ความไม่ถูกต้องของข้อมูลอาจเกิดขึ้นได้จาก ความผิดพลาดของการตรวจวัดค่าของทั้งอุปกรณ์ตรวจวัดหรือผู้บันทึกข้อมูล ความเข้าใจผิดหรือ ความมีอคติผู้บันทึกข้อมูล ส่วนความแม่นยำ หมายถึง ความสามารถในการวัดค่าตัวแปรเดิมซ้ำหลายครั้ง ภายใต้เงื่อนไขการวัดเดิม แล้วยังคงได้ค่าที่ตรวจวัดได้เป็นค่าเดิมเสมอ ข้อมูลที่มีคุณภาพ ควรมีความถูกต้องสูง จึงจะทำให้ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลตรงความเป็นจริง และสามารถนำไปใช้ให้เกิดประโยชน์

ได้ หากข้อมูลมีความถูกต้องต่ำแล้ว ผลการวิเคราะห์ข้อมูลย่อมไม่ตรวจตามความเป็นจริง และการนำไปใช้ประโยชน์อาจเกิดผลเสียหายได้ นอกจากนี้ข้อมูลจะต้องมีความแม่นสูงด้วย จึงจะทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพ และผลลัพธ์ที่ได้มีความถูกต้องและน่าเชื่อถือสูง

2) ความสมบูรณ์ (Completeness)

ข้อมูลที่มีความสมบูรณ์ คือ ข้อมูลที่สามารถรวบรวมค่าของทุกตัวแปรที่สนใจได้ครบถ้วน ไม่มีค่าของตัวแปรใดที่ไม่ทราบค่าหรือมีค่าสูญหาย (Missing Value) หากข้อมูลที่รวบรวมได้มีตัวแปรที่ไม่ทราบค่าหรือค่าสูญหายสูง ย่อมไม่สามารถนำมาวิเคราะห์ได้ แม้ว่าจะมีเทคนิควิธีการเติมค่าข้อมูลสูญหายที่ช่วยเพิ่มความสมบูรณ์ให้ข้อมูลได้ แต่จะต้องมีข้อมูลที่ทราบค่าจำนวนมากเพียงพอที่จะสามารถทำการเติมค่าข้อมูลที่สูญหายให้ถูกต้อง หรือใกล้เคียงความเป็นจริงมากที่สุด

3) ความต้องกัน (Consistency)

ความต้องกันของข้อมูล หมายถึง ค่าข้อมูลที่สื่อถึงสิ่งเดียวกันต้องมีค่าเหมือนกัน บางครั้งการรวบรวมข้อมูลจากหลายแหล่งข้อมูลที่แทนค่าสิ่งเดียวกันด้วยวิธีการแทนค่าต่างกัน เมื่อนำข้อมูลจากแหล่งข้อมูลเหล่านั้นมารวมกัน ทำให้ชุดข้อมูลใหม่ที่ได้เกิดปัญหาความไม่ต้องกันของข้อมูล ดังแสดงตัวอย่างในตาราง 2.4 ซึ่งเป็นชุดข้อมูลที่เกิดจากการรวมชุดข้อมูล Marvel Wikia และ DC Wikia [4] เข้าด้วยกัน พิจารณาค่าตัวแปร First Appear ของข้อมูลตัวละคร Spider-Man และ Bess Lynn ซึ่งมาจากชุดข้อมูล Marvel Wikia และ DC Wikia ตามลำดับ สังเกตได้ว่าค่าตัวแปรของทั้ง 2 ตัวละครสื่อถึงสิ่งเดียวกัน คือ เดือนสิงหาคม (August) ปี ค.ศ. 1962 แต่ค่าตัวแปร First Appear ของตัวละคร Spider-Man และ Bess Lynn มีค่าเท่ากับ Aug-62 และ 1962, August ตามลำดับ ซึ่งเกิดจากการแทนค่าเดือนและปีด้วยรูปแบบที่ต่างกันในแต่ละชุดข้อมูล ดังนั้น ชุดข้อมูลที่เกิดจากการรวมชุดข้อมูล Marvel Wikia และ DC Wikia นี้จึงเกิดปัญหาความไม่ต้องกันของข้อมูล ความไม่ต้องกันของข้อมูลนี้ทำให้ข้อมูลมีความซับซ้อนสูง และส่งผลให้ผลลัพธ์หรือแบบจำลองที่ได้จากการวิเคราะห์ข้อมูลมีความซับซ้อนสูงตามไปด้วย

[4] ***Comic Characters***. 2015. url: https://github.com/fivethirtyeight/data/tree/master/comic-characters

4) ความสมเหตุสมผล (Validity)

ความสมเหตุสมผลของข้อมูล หมายถึง ค่าของตัวแปรเป็นไปตามกฎธุรกิจหรือเงื่อนไขบังคับ ความสมเหตุสมผลสามารถแบ่งได้เป็น 2 ประเภท คือ

1. ความสมเหตุสมผลของค่าข้อมูลตัวแปรเดียว เป็นการพิจารณาความสมเหตุสมผลของค่าตัวแปรใดตัวแปรหนึ่งเท่านั้น ตัวอย่างเช่น

ตาราง 2.4: ตัวอย่างข้อมูลตัวละครที่เกิดจากการรวมชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel และ DC จากชุดข้อมูล Marvel Wikia และ DC Wikia เข้าด้วยกัน (ที่มา: https://github.com/fivethirtyeight/data/tree/master/comic-characters)

Name	ID	Align	Eye	...	First Appear	Year	Publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel	...	Aug-62	1962	Marvel
Captain America (Steven Rogers)	Public	Good	Blue	...	Mar-41	1941	Marvel
...	...	...	...	...	...	...	...
Yologarch (Earth-616)	N/A	Bad	N/A	...	N/A	N/A	Marvel
Batman (Bruce Wayne)	Secret	Good	Blue	...	1939, May	1939	DC
...	...	...	...	...	...	...	...
Bess Lynn (New Earth)	Secret	Good	Blue	...	1962, August	1962	DC
...	...	...	...	...	...	...	...
Mookie (New Earth)	Public	Bad	Blue	...	N/A	N/A	DC

- ค่าของตัวแปรที่เกี่ยวข้องกับเดือน จะต้องเป็น 1 ใน 12 เดือนที่เป็นไปได้เท่านั้น
- ค่าของตัวแปรระยะทาง จะต้องไม่มีเป็นจำนวนลบ
- ค่าตัวแปรคณะของข้อมูลนักศึกษาระดับปริญญาตรีของมหาวิทยาลัยเชียงใหม่ จะต้องเป็นค่าหนึ่ง 20 คณะหรือ 2 วิทยาลัยเท่านั้น

2. ความสมเหตุสมผลของค่าข้อมูลระหว่างหลายตัวแปร เป็นการพิจารณาความสอดคล้องของค่าตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไปประกอบกัน ตัวอย่างเช่น
   - ค่าของตัวแปรชื่ออำเภอ และ ชื่อจังหวัดภูมิลำเนา ของข้อมูลประชากรในประเทศไทย ถ้าค่าข้อมูลตัวแปรชื่อจังหวัดของประชากรคนหนึ่งคือ เชียงใหม่ แล้วค่าข้อมูลตัวแปรชื่ออำเภอ จะต้องเป็น 1 ใน 25 อำเภอของจังหวัดเชียงใหม่ เท่านั้น
   - ค่าตัวแปรคณะและหลักสูตรของข้อมูลนักศึกษามหาวิทยาลัยเชียงใหม่ ถ้าข้อมูลนักศึกษาคนหนึ่ง มีค่าตัวแปรคณะ คือ คณะวิทยาศาสตร์ แล้ว ค่าตัวแปรหลักสูตรของนักศึกษาคนนี้ จะต้องเป็นหลักสูตรที่เปิดสอนในคณะวิทยาศาสตร์ เท่านั้น

5) ความเป็นเอกลักษณ์ (Uniqueness)

ความเป็นเอกลักษณ์ของข้อมูล มีคุณลักษณะคือข้อมูลหนึ่งๆ จะปรากฏเพียงแห่งเดียวหรือระเบียนข้อมูลเดียวในชุดข้อมูลเท่านั้น เนื่องจากในการวิเคราะห์ข้อมูลดำเนินการภายในสมมติฐานว่าข้อมูลแต่ละข้อมูลเป็นอิสระต่อกัน การที่มีข้อมูลเดียวกันปรากฏซ้ำซ้อนในหลายระเบียนในชุดข้อมูล ทำให้ข้อมูลเหล่านั้นไม่เป็นอิสระต่อกัน ในขณะที่เทคนิควิธีการวิเคราะห์ข้อมูลแปลผลว่าข้อมูลเหล่านั้นเป็นข้อมูลคนละตัวกัน จึงทำให้ข้อมูลเหล่านั้นถูกให้ความสำคัญมากว่าข้อมูลอื่นๆ ดังนั้น การวิเคราะห์ข้อมูลจึงเกิดความโน้มเอียง (Bias) สูง แสดงในตัวอย่างในตาราง 2.5 มีการเก็บข้อมูลของตัวละคร Spider-Man (Peter Parker) ซ้ำใน 2 ระเบียน อีกทั้งตัวแปร Hair และ Alive ของระเบียนที่ 2 ยังเกิดค่าสูญหาย

(ในตัวอย่างนี้แทนค่าสูญหายด้วยค่า N/A)

ตาราง 2.5: ตัวอย่างปัญหาความไม่เป็นเอกลักษณ์ของข้อมูลในชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel

Page ID	Name	ID	Align	Eye	Hair	Sex	Gsm	Alive	...	Year
1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	Brown	Male	N/A	Living	...	1962
7139	Captain America (Steven Rogers)	Public	Good	Blue	White	Male	N/A	Living	...	1941
...	...	...	...	...	...	...	...	...	...	...
1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	N/A	Male	N/A	N/A	...	1962
...	...	...	...	...	...	...	...	...	...	...
321461	Oyakata (Earth-616)	Secret	Neutral	Brown	Black	Male	N/A	Living	...	1998

6) ความสม่ำเสมอ (Uniformity)

ความสม่ำเสมอของข้อมูล คือ การมีมาตรฐานเดียวกันสำหรับการเก็บค่าข้อมูลแต่ละตัวแปรในชุดข้อมูลเดียวกัน เช่น การใช้หน่วยวัดเดียวกัน และการใช้รูปแบบการจัดเก็บเดียวกัน เป็นต้น ซึ่งความสม่ำเสมอของข้อมูลนี้มักเกี่ยวข้องกับความต้องกันของข้อมูลด้วย

การทำความสะอาดข้อมูล

เมื่อพบว่าข้อมูลขาดคุณภาพตามคุณลักษณะข้อใดข้อหนึ่ง ต่อมาเราจะต้องทำความสะอาดข้อมูล เพื่อให้ข้อมูลมีคุณภาพมากขึ้นและสามารถนำไปวิเคราะห์ต่อได้ การทำความสะอาดข้อมูลที่ดีจะต้องอาศัยความรู้เฉพาะด้านทางธุรกิจด้วย จึงจะทำให้ข้อมูลที่ถูกทำความสะอาดแล้วใกล้เคียงความเป็นมากที่สุด ในหัวนี้จะกล่าวถึงวิธีการทำความสะอาดข้อมูลด้วยวิธีการคำนวณอย่างง่ายร่วมกับการอาศัยความรู้เฉพาะด้านทางธุรกิจ ดังนี้

การกำจัดความต้องกันของข้อมูล

เมื่อตรวจพบว่าข้อมูลของค่าตัวแปรใดเกิดความไม่ต้องกันของข้อมูล เนื่องด้วยวิธีการแทนค่าข้อมูลที่แตกต่างกัน สามารถแก้ไขได้ด้วยวิธีการสร้างกฎสำหรับแก้ไขค่าข้อมูล โดยยึดรูปแบบการแทนค่าข้อมูลรูปแบบใดรูปแบบหนึ่งที่เหมาะสมไว้เป็นมาตรฐาน และทำการแปลงค่าข้อมูลในรูปแบบอื่นๆ ให้อยู่ในรูปแบบตามมาตรฐานที่กำหนด วิธีการเช่นนี้จะสามารถทำการแก้ไขค่าข้อมูลได้อย่างอัตโนมัติโดยการเขียนโปรแกรมตามกฎที่สร้างขึ้น เพื่อตรวจสอบข้อมูลแต่ละระเบียนในชุดข้อมูล

ตัวอย่าง 2.7

จากตัวอย่างข้อมูลในตาราง 2.4 ซึ่งตรวจพบว่าค่าตัวแปร First Appear เกิดความไม่ต้องกันของข้อมูล โดยการแทนค่าเดือนและปี มีรูปแบบที่แตกต่างกัน 2 รูปแบบ คือ [เดือน(อักษรย่อ)-ปี ค.ศ.(2 หลักท้าย)] และ [ปี ค.ศ., เดือน] ในที่นี้จะยึดรูปแบบ [เดือน(อักษรย่อ)-ปี ค.ศ.(2 หลักท้าย)] เป็นรูปแบบมาตรฐาน ดังนั้น ค่าตัวแปร First Appear ของข้อมูลใดที่อยู่ในรูปแบบ [ปี ค.ศ., เดือน] จะถูกแก้ไขให้อยู่ในรูปแบบมาตรฐาน ได้ผลลัพธ์ดังนี้

Name	...	First Appear	Year	Publisher
Spider-Man (Peter Parker)	...	Aug-62	1962	Marvel
Captain America (Steven Rogers)	...	Mar-41	1941	Marvel
...	...	...	...	...
Yologarch (Earth-616)	...	N/A	N/A	Marvel
Batman (Bruce Wayne)	...	1939, May	1939	DC
...	...	...	...	...
Bess Lynn (New Earth)	...	Aug-62	1962	DC
...	...	...	...	...
Mookie (New Earth)	...	N/A	N/A	DC

การเติมข้อมูลสูญหาย

ปัญหาข้อมูลสูญหาย ถือเป็นปัญหาที่สามารถตรวจจับได้ง่ายกว่าปัญหาคุณภาพข้อมูลอื่นๆ และในปัจจุบันมีงานวิจัยหลายงานที่นำเสนอวิธีการเติมข้อมูลค่าสูญหายด้วยวิธีการคำนวณโดยอาศัยข้อมูลที่มีอยู่ ในหัวข้อนี้จะกล่าวถึงวิธีการเติมค่าข้อมูลสูญหายอย่างง่ายด้วยค่ากลางของข้อมูล วิธีการนี้มีสมมติฐานว่าข้อมูลที่สูญหายมีค่าเท่ากับข้อมูลที่มีอยู่ส่วนใหญ่ ซึ่งสามารถอธิบายได้โดยการใช้ค่ากลางข้อมูล ได้แก่ ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม อย่างไรก็ตาม ตัวแปรแต่ละชนิดข้อมูลมีความสามารถในการหาค่ากลางข้อมูลต่างกัน (ดูตาราง 2.2) ดังนั้น จึงต้องเลือกใช้ค่ากลางให้เหมาะสมกับข้อมูลด้วย วิธีการเติมค่าข้อมูลสูญหายด้วยค่ากลางข้อมูล มีขั้นตอน ดังนี้

ขั้นตอนที่ 1 ตรวจจับตัวแปรที่เกิดปัญหาค่าข้อมูลสูญหาย

ขั้นตอนที่ 2 สำหรับแต่ละตัวแปรที่เกิดปัญหาค่าข้อมูลสูญหาย X_j ทำการระบุชนิดข้อมูล

ขั้นตอนที่ 2.1 คำนวณค่ากลาง (ค่าเฉลี่ย ค่ามัธยฐาน หรือค่าฐานนิยม) ของค่าตัวแปร X_j จากข้อมูลที่มีอยู่

ขั้นตอนที่ 2.2 นำค่ากลางที่คำนวณได้ เติมให้แก่ค่าตัวแปร X_j ของข้อมูลที่มีค่าตัวแปร X_j สูญหาย

การเติมค่าข้อมูลสูญหายนั้น ต้องพึงระลึกไว้เสมอว่าค่าข้อมูลที่เติมลงไปไม่ใช่ค่าข้อมูลจริง ความถูกต้องของข้อมูลขึ้นอยู่กับวิธีการเติมค่าข้อมูลสูญหายที่เลือกใช้ และข้อมูลที่มีอยู่ด้วย จึงต้องดำเนินการอย่างระมัดระวัง ทั้งนี้อาจอาศัยคำแนะนำจากผู้เชี่ยวชาญเฉพาะด้านที่เกี่ยวข้องกับปัญหาธุรกิจ หรือหากตัวแปรใดมีค่าสูญหายจำนวนมากแล้ว อาจต้องดำเนินการเก็บรวบรวมข้อมูลตัวแปรนั้นใหม่ หรือ พิจารณาไม่ใช้ตัวแปรนั้นในการวิเคราะห์ข้อมูล

ตัวอย่าง 2.8

ชุดข้อมูลต่อไปนี้ ประกอบด้วยตัวแปร IQ และ Job Performance ของข้อมูลจำนวน 14 ระเบียน

	IQ	Job Performance
$\mathbf{x}_1$	78	N/A
$\mathbf{x}_2$	84	N/A
$\mathbf{x}_3$	84	N/A
$\mathbf{x}_4$	85	N/A
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	14

ในตัวอย่างนี้จะทำการเติมค่าข้อมูลสูญหายในชุดข้อมูลข้างต้น มีขั้นตอน

ดังนี้

ขั้นตอนที่ 1 ตรวจจับได้ว่าภายในชุดข้อมูล มีค่าตัวแปร Job Performance ของข้อมูล $\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ และ $\mathbf{x}_4$ สูญหาย

ขั้นตอนที่ 2 ตัวแปร Job Performance มีชนิดข้อมูลเป็นข้อมูลระดับอัตราส่วน ดังนั้น สามารถหาค่าเฉลี่ยได้

ขั้นตอนที่ 2.1 คำนวณค่าเฉลี่ยของค่าตัวแปร Job Performance จากข้อมูลที่มีอยู่

ค่าเฉลี่ย = 11.70

ขั้นตอนที่ 2.2 นำค่าเฉลี่ยที่คำนวณได้ เติมให้แก่ค่าตัวแปร Job Performance ของข้อมูล $\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ และ $\mathbf{x}_4$ จะได้ชุดข้อมูลที่ถูกปรับปรุงคุณภาพแล้ว ดังนี้

	IQ	Job Performance
$\mathbf{x}_1$	78	**11.70**
$\mathbf{x}_2$	84	**11.70**
$\mathbf{x}_3$	84	**11.70**
$\mathbf{x}_4$	85	**11.70**
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	14

การกำจัดความซ้ำซ้อนกันของข้อมูล

โดยปกติแล้ว การสำรวจและจัดเก็บข้อมูลนั้น ข้อมูลแต่ตัวจะถูกกำหนดตัวระบุ (Identifier) ที่ไม่ซ้ำกัน สำหรับระบุตัวตนของข้อมูล เช่น รหัสนักศึกษาที่สามารถใช้ระบุตัวนักศึกษา และรหัสประจำตัวผู้ป่วย (Hospital Number) ที่ใช้ระบุตัวผู้ป่วยในแต่ละโรงพยาบาล เป็นต้น ดังนั้น ตัวระบุนี้จึงสามารถนำมาใช้ในการตรวจหาข้อมูลซ้ำซ้อนในชุดข้อมูลเดียวกันได้ หากตรวจพบค่าของตัวระบุเดียวกันในข้อมูลหลายระเบียนแล้ว นั่นหมายถึงข้อมูลนั้นเกิดการซ้ำซ้อนขึ้น การกำจัดความซ้ำซ้อนกันของข้อมูลนี้ทำได้โดย

1. เลือกรักษาระเบียนข้อมูลที่มีค่าตัวระบุเดียวกันไว้เพียงระเบียนข้อมูลเดียวเท่านั้น อาจเลือกข้อมูลโดยอาศัยกฎทางธุรกิจ คำแนะนำจากผู้เชี่ยวชาญเฉพาะด้านที่เกี่ยวข้องกับปัญหาธุรกิจ หรือพิจารณาประเด็นความถูกต้องของข้อมูล ความสมบูรณ์ของข้อมูล ความสมเหตุสมผลของข้อมูล และความเป็นปัจจุบันของข้อมูล
2. ผสานระเบียนข้อมูลที่ซ้ำซ้อนกันให้เป็นเพียง 1 ระเบียนข้อมูล อาจเลือกค่าตัวแปรแต่ละตัวแปรที่มีความถูกต้อง และเป็นปัจจุบันจากต่างระเบียนของข้อมูลที่ซ้ำซ้อนกัน มาสร้างเป็นระเบียนข้อมูลใหม่เพียงระเบียนข้อมูลเดียว

ตัวอย่าง 2.9

จากตาราง 2.5 เมื่อพิจารณาตัวแปร Page ID ซึ่งเป็นตัวระบุของตัวละครแต่ละตัวแล้ว พบว่ามี 2 ระเบียนข้อมูลที่มีค่าตัวแปร Page ID เหมือนกัน คือ 1678 ซึ่งเป็นข้อมูลตัวละคร Spider-Man (Peter Parker) ดังนั้น ชุดข้อมูลนี้จึงมีข้อมูลตัวละคร Spider-Man (Peter Parker) ซ้ำซ้อนกัน

จากระเบียนข้อมูลที่ซ้ำซ้อนกันของข้อมูลตัวละคร Spider-Man (Peter Parker) พบว่าระเบียนข้อมูลแรกมีความสมบูรณ์ของข้อมูลมากกว่าอีกระเบียนข้อมูลหนึ่ง คือ ระเบียนข้อมูลแรกมีค่าสูญหาย 1 ค่าตัวแปร นั้นคือ ตัวแปร Gsm ในขณะที่อีกระเบียนข้อมูลหนึ่งมีค่าสูญหาย 3 ค่าตัวแปร นั่นคือ Hair Gsm และ Alive ดังนั้น จึงเลือกรักษาข้อมูลของตัวละคร Spider-Man (Peter Parker) ในระเบียนข้อมูลแรกไว้ และลบระเบียนข้อมูลที่ซ้ำซ้อนกันอีกระเบียนข้อมูลหนึ่งออกจากชุดข้อมูล จะได้ชุดข้อมูลใหม่ ดังนี้

Page ID	Name	...	Hair	Gsm	Alive	...
1678	Spider-Man (Peter Parker)	...	Brown	N/A	Living	...
7139	Captain America (Steven Rogers)	...	White	N/A	Living	...
...	...	...	...	...	...	...
1678	Spider-Man (Peter Parker)	...	N/A	N/A	N/A	...
...	...	...	...	...	...	...
321461	Oyakata (Earth-616)	...	Black	N/A	Living	...

2.4 แบบฝึกหัดท้ายบท

1. นักวิทยาศาสตร์คนหนึ่ง ต้องการศึกษาสายพันธุ์พืชในบริเวณอุทยานแห่งชาติดอยสุเทพ-ปุย จังหวัดเชียงใหม่ เขาสามารถรวบรวมข้อมูลสายสายพันธุ์พืชที่ต้องการจากแหล่งกำเนิดข้อมูลประเภทใดได้บ้าง พร้อมยกตัวอย่างประกอบ
2. องค์ประกอบของตารางข้อมูล (แถว คอลัมน์ และตารางข้อมูล) ใด มีความสัมพันธ์กับคำสำคัญ ต่อไปนี้
   - ระเบียนข้อมูล (Record)
   - ตัวแปร (Variable)
   - จุดข้อมูล (Data Point)
   - ชุดข้อมูล (Dataset)
   - ลักษณะเด่น (Feature)
   - เวกเตอร์ลักษณะเด่น (Feature Vector)
   - เขตข้อมูล (Field)
   - รายการ (Transaction)
3. จงระบุชนิดข้อมูล (ข้อมูลนามบัญญัติ ข้อมูลเชิงอันดับ ข้อมูลระดับอัตรภาค และข้อมูลระดับอัตราส่วน) ของตัวแปรต่อไปนี้ พร้อมอธิบายเหตุผล
   - ปีเกิด (พ.ศ.) ของนักศึกษา
   - สาขาวิชาของนักศึกษา
   - จำนวนสัตว์เลี้ยงของนักศึกษา
   - น้ำหนักและส่วนสูงของนักศึกษา
   - ชื่อของนักศึกษา
4. จงเข้ารหัสข้อมูลของค่าตัวแปร Align = {Bad, Neutral, Good}
5. จงเข้ารหัสข้อมูลของค่าตัวแปร Hair = {Black, Bronze, Brown, Gold, Gray}
6. จากตารางข้อมูลต่อไปนี้ จงแปลงค่าข้อมูลตัวแปร Job Performance ด้วยเทคนิควิธีต่อไปนี้
   - Min-max Normalization
   - Standardization

IQ	Job Performance
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	14

7. ชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel (ดูชุดข้อมูลที่ https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-data.csv) วิธีการเติมค่าข้อมูลสูญหายวิธีใด เหมาะสมสำหรับจัดการค่าข้อมูลสูญหายของตัวแปร ต่อไปนี้
   - Eye
   - Hair
   - Gsm
   - Appearance (จำนวนครั้งการปรากฏตัวในหนังสือการ์ตูน)

2.5 แหล่งศึกษาเพิ่มเติม

1. Mohammed J. Zaki and Wagner Meira Jr. ***Data Mining and Analysis: Fundamental Concepts and Algorithm***. New York, USA: Cambridge University Press, 2014.
2. Baijayanta Roy. ***All about Categorical Variable Encoding***. 2019. https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02.
3. Craig K. Enders. ***Applied Missing Data Analysis***. New York, USA: Guilford Press, 2010.