ใบความรู้ที่ 2.3 เรื่อง การเตรียมข้อมูลและประมวลผลข้อมูล
หน่วยที่ 2 การประมวลผลข้อมูล แผนการจัดการเรียนรู้ที่ 8 เรื่อง การเตรียมข้อมูล
รายวิชา เทคโนโลยี รหัส ว23103 ชั้นมัธยมศึกษาปีที่ 3

<u>การเตรียมข้อมูล</u>

การเตรียมข้อมูลหมายถึงการดำเนินการกับข้อมูลที่รวบรวมมาเพื่อให้ข้อมูลที่มีคุณภาพพร้อมนำไปประมวลผล เนื่องจากข้อมูลมีบทบาทสำคัญในการแก้ปัญหาจึงจำเป็นอย่างยิ่งที่เราต้องคำนึงถึงคุณภาพของข้อมูลก่อนนำไปประมวลผลอย่างไรก็ตามข้อมูลบางส่วนที่ได้จากการรวบรวมอาจยังไม่สามารถนำไปประมวลผลได้ในทันทีจึงจำเป็นต้องทำความสะอาดข้อมูลก่อน เช่น ข้อมูลมีความซ้ำซ้อนมีค่าหรือลักษณะที่ผิดปกติจากข้อมูลอื่นหรือมีรายการข้อมูลที่ขาดหายไป การทำความสะอาดข้อมูลเป็นกระบวนการการปรับปรุง/แก้ไขข้อมูลโดยคำนึงถึงลักษณะของข้อมูลและวิธีการประมวลผล

<u>แนวทางการตรวจสอบความผิดปกติของข้อมูลเพื่อทำความสะอาดข้อมูลมีดังนี้</u>

1. ความสมบูรณ์ (validity) ข้อมูลที่รวบรวมมีความสุขต้องตามข้อกำหนด เช่น
   - ข้อมูลชนิดของข้อมูลมีความสอดคล้องกันเช่นอายุเป็นข้อมูลชนิดตัวเลขชื่อเป็นข้อมูลชนิดข้อความ
   - ข้อมูลมีค่าสอดคล้องกับความเป็นจริงเช่นน้ำหนักต้องไม่เป็นจำนวนลบวันเกิดต้องเป็นวันที่ในอดีตคะแนนต้องมีค่าอยู่ช่วงศูนย์ถึง 100 จากคะแนนเต็ม 100 คะแนนวันที่ 30 ต้องไม่ใช่ในวันเดือนกุมภาพันธ์
   - ข้อมูลบางอย่างจะมีค่าไม่ซ้ำกันเช่นรหัสประจำตัวนักเรียนในโรงเรียนเดียวกันเลขทะเบียนรถยนต์เลขบัตรประชาชน
   - ข้อมูลบางอย่างต้องไม่เป็นค่ะว่างเช่นชื่อนักเรียนวันเดือนปีเกิด
   - ข้อมูลมีค่าผิดปกติจากข้อมูลขั้นเช่นเก็บรวบรวมอายุของนักเรียนแต่มีข้อมูลอายุเป็น 150 ปี
2. รูปแบบเดียวกัน (uniformity) ข้อมูลเรื่องเดียวกันต้องเก็บอยู่ในรูปแบบเดียวกันเช่นข้อมูลน้ำหนักมีหน่วยเป็นกิโลเหมือนกัน วันที่ในรูปแบบ วว/ดด/ปป/ หรือ ดด/วว/ปป หรือรูปแบบ พ.ศ. หรือ ค.ศ.
3. ความครบถ้วน(completeness) ข้อมูลที่เกี่ยวข้องต้องถูกรวบรวมอย่างครบครบถ้วน
4. ความทันสมัย(timelinrss) ข้อมูลที่มีค่าที่สุดของกับเวลาและสถานการณ์

อ้างอิงข้อมูลจาก :: สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี

ตัวอย่างที่ 1 ข้อมูลสถานการณ์ covid-19 ในจังหวัดประจวบคีรีขันธ์ ระหว่างวันที่ 22 มีนาคม - 18 เมษายน 2563

ลำดับที่	สัญชาติ	อายุ	เพศ
1	จีน	73	หญิง
2	สวิตเซอร์แลนด์	82	ชาย
3	สหราชอาณาจักร	51	ชาย
4	สหราชอาณาจักร	60	ชาย
5	สหราชอาณาจักร	-	ชาย
6	สหราชอาณาจักร	72	ชาย
7	เบลเยี่ยม	58	ชาย
8	เบลเยี่ยม	81	หญิง
9	เยอรมัน	59	ชาย
10	ไทย	45	ชาย
11	ไทย	26	หญิง
12	ไทย	26	หญิง
13	ไทย	48	หญิง
14	ไทย	34	หญิง
15	ไทย	19	ชาย
16	ไทย	57	ชาย
17	ไทย	31	หญิง

อ้างอิงข้อมูลจาก :: กรมควบคุมโรค กระทรวงสาธารณสุข

เมื่อพิจารณาค่าข้อมูลพบว่า-ไม่มีอายุเงินคงเหลือในข้อมูลลำดับที่ 5 6 มีความแตกต่างจากลำดับอื่นๆ มาก เราอาจตั้งข้อสังเกตได้ว่าอาจมีความผิดปกติของข้อมูล ซึ่งอาจเป็นไปได้ 2 กรณี คือ

1. อาจเกิดการผิดพลาดจากการบันทึกข้อมูล
2. ในลำดับที่ 56 อาจเป็นข้อมูลที่ไม่สามารถระบุได้ หรือหลักฐานไม่ชัดเจนเนื่องจากเป็นชาวต่างชาติ เป็นต้น

ก่อนนำข้อมูลในตารางไปประมวลผลเราต้องพิจารณาเพิ่มเติมว่าจะนำข้อมูลของลำดับที่ -56 ไปใช้ในส่วนของการประมวลผลหรือไม่ เช่น ในกรณีที่ต้องการหาอายุเฉลี่ย ถ้าไม่มีข้อมูลในลำดับที่ 5 ต้องตัดข้อมูลนี้ออก ซึ่งอาจแตกต่างกันไปแล้วแต่ละกรณี ทดลองพิจารราสถานการณ์ต่างๆต่อไปนี้และประเมินว่าจะนำลำดับที่ 6 มาใช้หรือไม่ เพราะเหตุใด