ปฏิบัติการที่ 2
การจัดเตรียมข้อมูลก่อนการวิเคราะห์

วัตถุประสงค์

1. เพื่อให้สามารถแปลงชนิดข้อมูล (Datatype) ของตัวแปรให้อยู่รูปแบบที่เหมาะสมสำหรับการวิเคราะห์ได้
2. เพื่อให้สามารถจัดการกับค่าสูญหาย (Missing Value) โดยวิธีการเติมค่าข้อมูลหรือลบข้อมูลได้

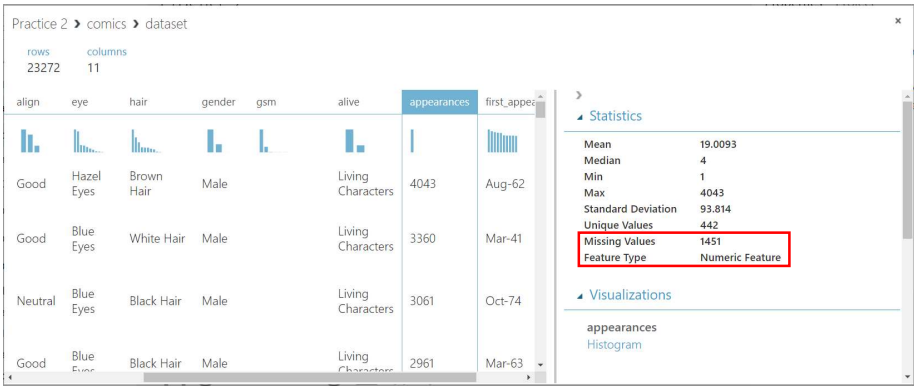
1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ชุดข้อมูล 120 Years of Olympic History athletes and Results (สำหรับการฝึกปฏิบัติการ)

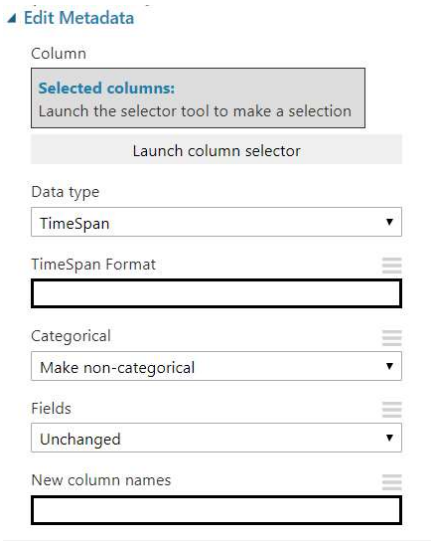
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

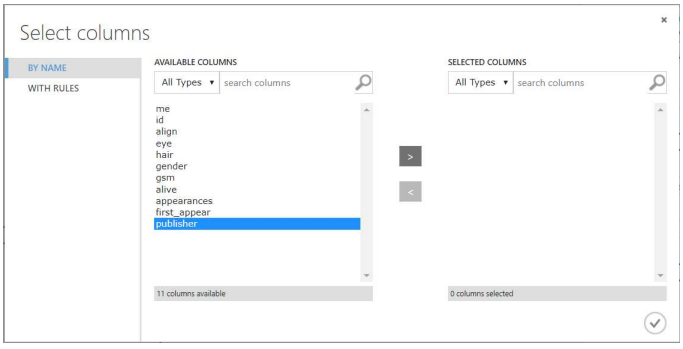
1. นำเข้าชุดข้อมูล Comic Characters จากแฟ้มข้อมูล comics.csv โดยเปลี่ยนค่าในเซลล์ที่มีค่า NA เป็นค่าว่างเปล่าด้วย และตั้งชื่อชุดข้อมูลเป็น comics (ดูหัวข้อ ปฏิบัติการที่ 1)
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 2”
3. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังรูป



5. ตรวจสอบพร้อมจดบันทึกชนิดข้อมูลและจำนวนค่าสูญหาย (Missing Values) ของแต่ละตัวแปร โดยคลิกที่ชื่อคอลัมน์ แล้วสังเกตค่า Feature Type และ Missing Values ในหน้าต่างย่อย Statistics
6. หากต้องการเปลี่ยนแปลงชนิดข้อมูลของตัวแปร สามารถทำได้โดยใช้โมดูล Edit Metadata (ภายใต้ Data Transformation → Manipulation)
7. ลากโมดูล Edit Metadata มาวางบน Workspace แล้วเชื่อมโยงโมดูลชุดข้อมูล comics กับโมดูล Edit Metadata โดยให้ชุดข้อมูล comics เป็นข้อมูลเข้า
8. คลิกที่โมดูล Edit Metadata ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata)

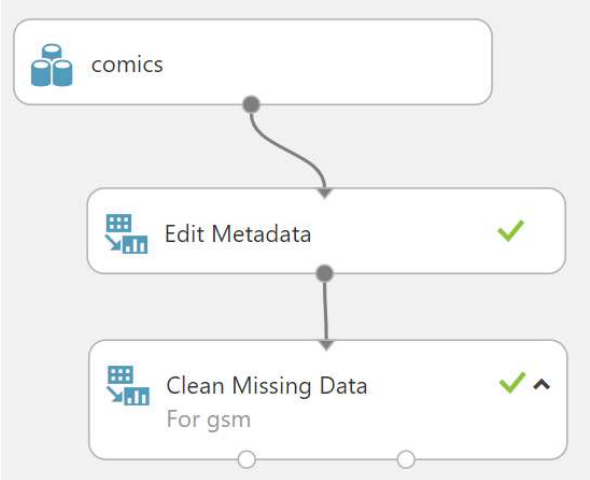


9. คลิก Launch the selector to make a selection เพื่อเลือกตัวแปรที่ต้องการเปลี่ยนชนิดข้อมูล จะปรากฏไดอะล็อก Select columns ดังรูป

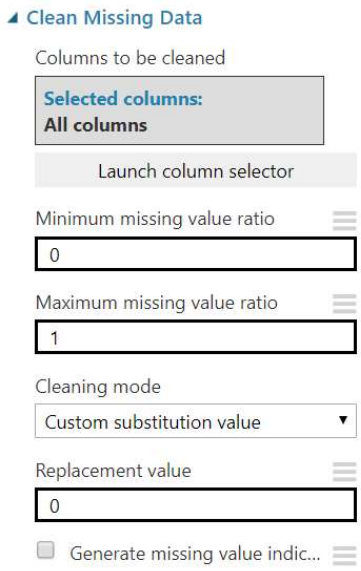


10. ในที่นี้จะแก้ไขชนิดข้อมูลของตัวแปร eye และ hair โดยคลิกเลือกตัวแปร eye จากส่วน AVALIABLE COLUMNS แล้วคลิกปุ่ม > ต่อมาเลือกตัวแปร hair จากส่วน AVALIABLE COLUMNS แล้วคลิกปุ่ม > (สามารถเลือกตัวแปรหลายตัวพร้อมกันกันโดยการกดปุ่ม CTRL พร้อมกับคลิกเลือก) จะสังเกตได้ว่าทั้งตัวแปร eye และ hair ถูกย้ายมายังส่วน SELECTED COLUMNS เสร็จแล้วคลิกปุ่ม ✓
11. ต่อมาเลือกชนิดข้อมูลที่ต้องการจากลิสต์ Data type ในที่นี้เลือกเป็น String และกำหนดให้เป็นข้อมูลที่จัดเป็นกลุ่ม โดยเลือกตัวเลือก Make categorical จากลิสต์ Categorical กรณีนี้จะทำให้ตัวแปร eye และ hair เป็นข้อมูลที่จัดเป็นกลุ่ม (Categorical Data)
12. คลิก **RUN** แล้วดูผลลัพธ์ของโมดูล Edit Metadata เปรียบเทียบกับผลลัพธ์ของโมดูลชุดข้อมูล comics จะพบว่าก่อนการเรียกใช้โมดูล Edit Metadata ชนิดของข้อมูลตัวแปร eye และ hair เป็น String Feature หลังจากการแปรชนิดข้อมูลของตัวแปรทั้งสองแล้ว ชนิดของข้อมูลตัวแปรเป็น Categorical Feature

    ข้อแนะนำ หากต้องการเปลี่ยนแปลงชนิดข้อมูลตัวแปรอื่นๆ เพิ่มเติมสามารถทำได้โดยใช้งานโมดูล Edit Metadata โดยที่ส่งผลลัพธ์ของโมดูลก่อนหน้าเป็นข้อมูลเข้า ทำเช่นนี้เป็นทอดๆ จนกว่าชนิดข้อมูลของทุกตัวแปรเป็นไปตามที่ต้องการ
13. การจัดการค่าสูญหาย (Missing Value) ทำได้โดยใช้โมดูล Clean Missing Data (ภายใต้ Data Transformation → Manipulation)
14. เมื่อลากโมดูล Clean Missing Data วางบน Workspace แล้ว เชื่อมโยงโมดูล Edit Metadata กับโมดูล Clean Missing Data โดยให้ผลลัพธ์จากโมดูล Edit Metadata เป็นข้อมูลเข้าของโมดูล Clean Missing Data และเพิ่มข้อความคิดเห็น (Comment) โดยการคลิกขวาที่โมดูล เลือก Edit Comment เป็น For gsm

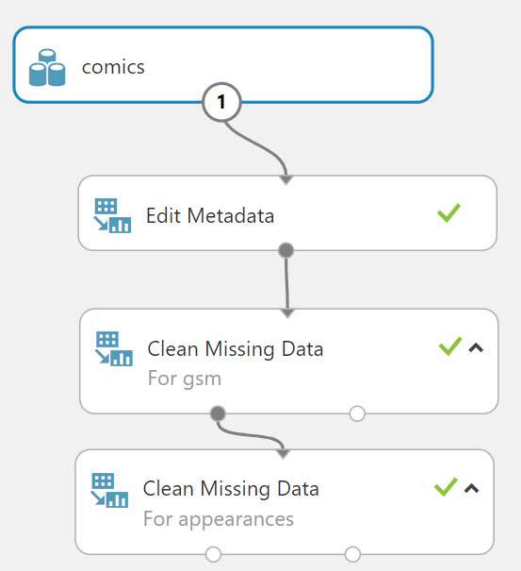


15. คลิกที่โมดูล Clean Missing Data ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data)



16. คลิก Launch the selector to make a selection เพื่อเลือกตัวแปรที่ต้องการจัดการกับค่าสูญหาย จะปรากฏไดอะล็อก Select columns คลิกที่ **BY NAME** เพื่อเลือกจากชื่อคอลัมน์
17. ทำการเลือกตัวแปรที่ต้องการ ในที่นี้จะเลือกตัวแปร gsm แล้วคลิก ปุ่ม ✓
    ข้อแนะนำ สามารถเลือกตัวแปรเพื่อจัดการกับค่าสูญหายได้ครั้งละหลายตัวแปร อย่างไรก็ตาม ตัวแปรเหล่านั้นจะต้องใช้วิธีการจัดการค่าสูญหายด้วยวิธีเดียวกัน หากต้องการจัดการค่าสูญหายของตัวแปรอื่นๆ ด้วยวิธีการไม่เหมือนกัน สามารถทำได้โดยใช้งานโมดูล Clean Missing Data แยกเฉพาะกลุ่มตัวแปรที่ใช้วิธีการจัดการค่าสูญหายเดียวกัน โดยที่ส่งผลลัพธ์ของโมดูลก่อนหน้าเป็นข้อมูลเข้า ทำเช่นนี้เป็นทอดๆ
18. ต่อมาเลือกวิธีการจัดค่าสูญหายจากลิสต์ Cleaning Mode เป็น Remove entire column เนื่องจากสังเกตได้ว่าโดยส่วนมากค่า gsm ของข้อมูลทุกแถวจจะสูญหาย
19. คลิก **RUN** แล้วดูผลลัพธ์ของโมดูล Clean Missing Data โดยโมดูลนี้ให้ผลลัพธ์ 2 อย่าง คือ 1) ชุดข้อมูลมูลที่ถูกจัดการค่าสูญหายแล้ว (Cleaned Dataset) สำหรับนำข้อมูลชุดนี้ไปใช้ในการวิเคราะห์ต่อไป และ 2) ตัวแปลงข้อมูล (Cleaning Transformation) สำหรับนำไปใช้กับข้อมูลชุดใหม่ที่มีโครงสร้างข้อมูลเดียวกันและต้องใช้วิธีการจัดการกับค่าสูญหายด้วยวิธีเดียวกัน
20. คลิกเลือก Visualize จากโหนดส่วนต่อประสานข้อมูลออกโหลดแรกของ โมดูล Clean Missing Data จะพบว่าคอลัมน์ gsm หายไป

21. ต่อมาทำการจัดการค่าสูญหายของตัวแปร appearance โดยการลากโมดูล Clean Missing Data ใหม่มาวางบน workspace และเพิ่มข้อความคิดเห็น (Comment) เป็น For appearance แล้วให้ผลลัพธ์ของโมดูล Clean Missing Data สำหรับตัวแปร gsm เป็นข้อมูลเข้าของโมดูล Clean Missing Data สำหรับตัวแปร appearance



22. ทำการเลือกตัวแปร appearance แล้วเลือกวิธีการจัดค่าสูญหายจากลิสต์ Cleaning Mode เป็น Replace with mean
23. คลิก **RUN** แล้วดูผลลัพธ์ของโมดูล Clean Missing Data สำหรับตัวแปร appearance จะพบว่าจำนวนข้อมูลสูญหายของตัวแปร appearance มีค่าเท่ากับ 0 (จากเดิมมีจำนวนข้อมูลสูญหาย 1,451 ข้อมูล)

3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. นำชุดข้อมูล 120 Years of Olympic History athletes and Results จากแฟ้มข้อมูล athlete_events.csv เข้าสู่โปรแกรม ML Studio (ทำการเปลี่ยนค่าระบุข้อมูลสูญหายจาก NA เป็นค่า ว่างเปล่า ด้วย) กำหนดชื่อชุดข้อมูลเป็น “athlete events”
2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 2” โดยให้นำชุดข้อมูล athlete events เข้าสู่การทดลอง
3. จดบันทึกชนิดข้อมูลและจำนวนค่าสูญหายของแต่ละตัวแปร
4. ทำการเปลี่ยนชนิดข้อมูล และจัดการค่าสูญหายของแต่ละตัวแปร ตามรายละเอียดในตารางด้านล่างนี้

ตัวแปร	ชนิดข้อมูล	วิธีการจัดการค่าสูญหาย
ID	String Feature	(ไม่มี Missing Value)
Name	String Feature	(ไม่มี Missing Value)
Sex	Categorical Feature	(ไม่มี Missing Value)
Age	Numeric Feature (Integer)	Replace with median
Height	Numeric Feature (Floating Point)	Replace with mean
Weight	Numeric Feature (Floating Point)	Replace with mean
Team	Categorical Feature	(ไม่มี Missing Value)
NOC	Categorical Feature	(ไม่มี Missing Value)
Games	String Feature	(ไม่มี Missing Value)
Year	Numeric Feature (Integer)	(ไม่มี Missing Value)
Season	Categorical Feature	(ไม่มี Missing Value)
City	Categorical Feature	(ไม่มี Missing Value)
Sport	Categorical Feature	(ไม่มี Missing Value)
Event	String Feature	(ไม่มี Missing Value)
Medal	Categorical Feature	Custom substitution value (Replacement value = None)

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_02_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th