ปฏิบัติการที่ 2
การจัดเตรียมข้อมูลก่อนการวิเคราะห์

วัตถุประสงค์

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

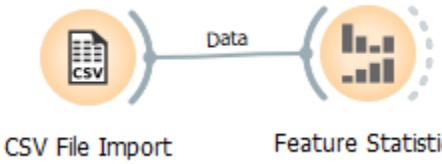
ชุดข้อมูลปฏิบัติการ

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

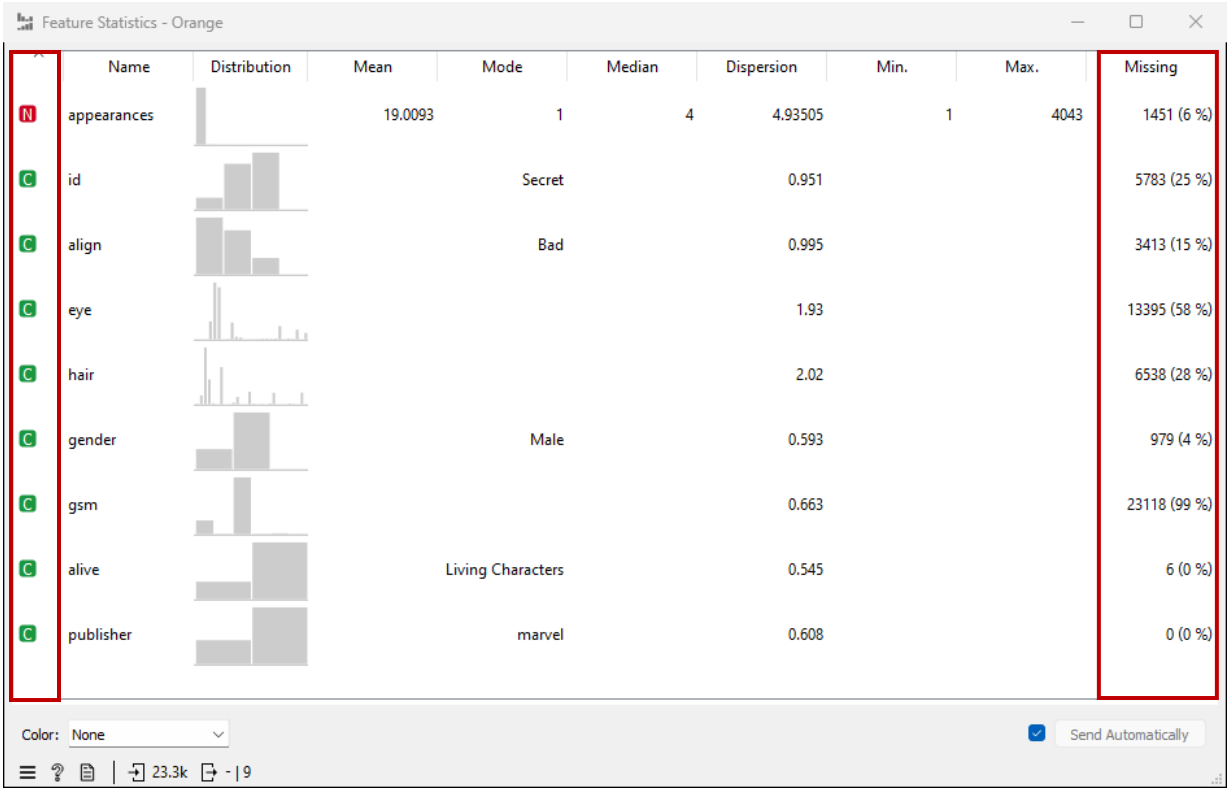
ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

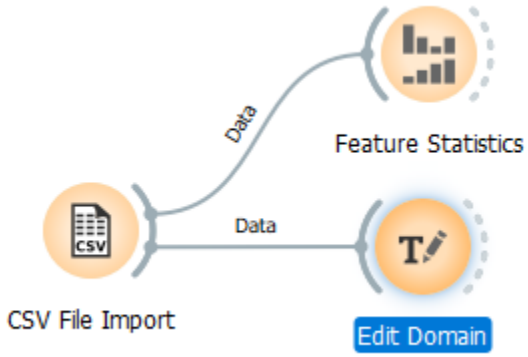
1. เปลี่ยนค่าในเซลล์ที่มีค่า NA ข้อมูล Comic Characters จากแฟ้มข้อมูล comics.csv เป็นค่าว่างเปล่า (ดูหัวข้อ ปฏิบัติการที่ 1)
2. เปิดโปรแกรม Orange
3. ทำการบันทึก workspace โดยไปที่เมนู File เลือก Save จากนั้นทำการตั้งชื่อไฟล์ในรูปแบบ Practice_02_id.ows โดยแทน id ด้วยรหัสนักศึกษา แล้วกดปุ่ม Save
4. นำชุดข้อมูลจากแฟ้มข้อมูล comics.csv เข้าสู่โปรแกรม Orange โดยใช้โมดูล CSV File Import
5. คลิกเลือกโมดูล Feature Statistics จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล CSV File Import จากด้าน output เข้าสู่โมดูล Feature Statistics ด้าน input ดังรูป



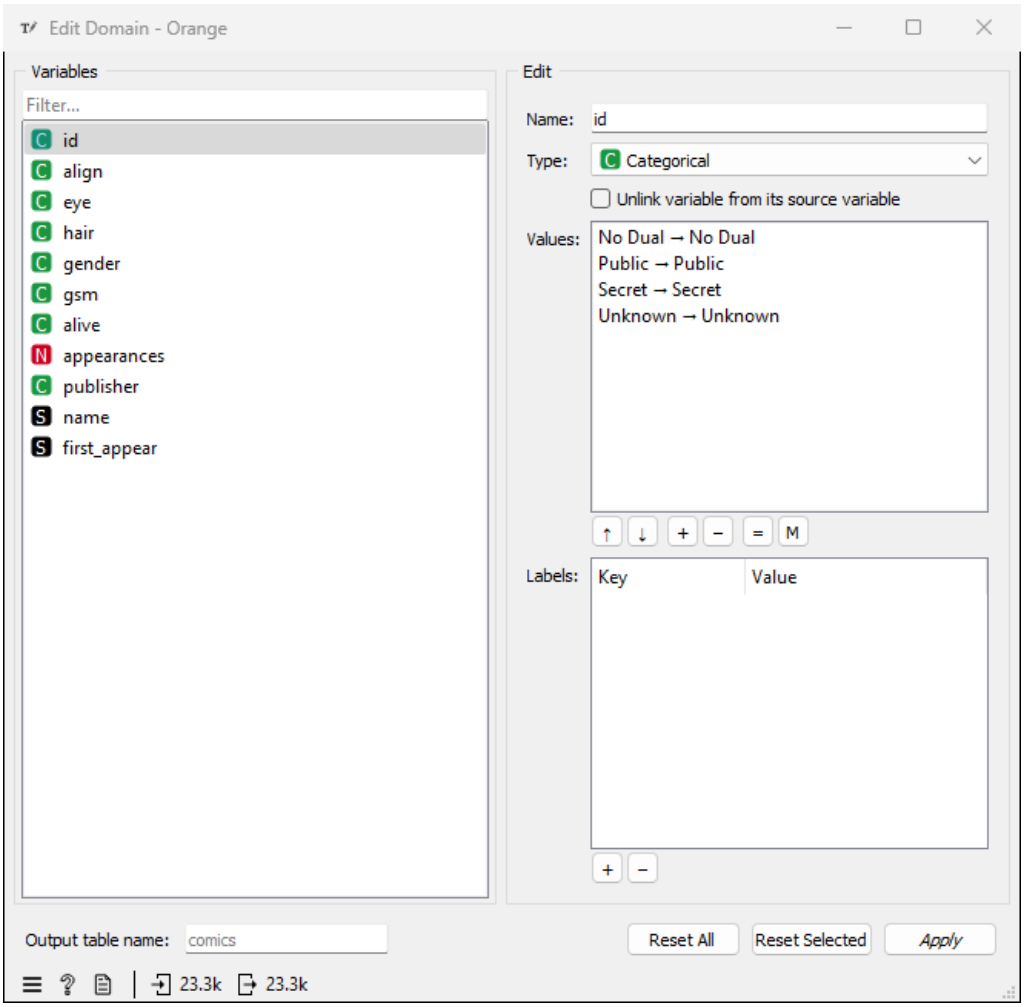
6. ดับเบิลคลิกที่โมดูล Feature Statistics จากปรากฏหน้าต่างแสดงผลลัพธ์ค่าสถิติเชิงพรรณาของแต่ละตัวแปรในชุดข้อมูล ดังรูป



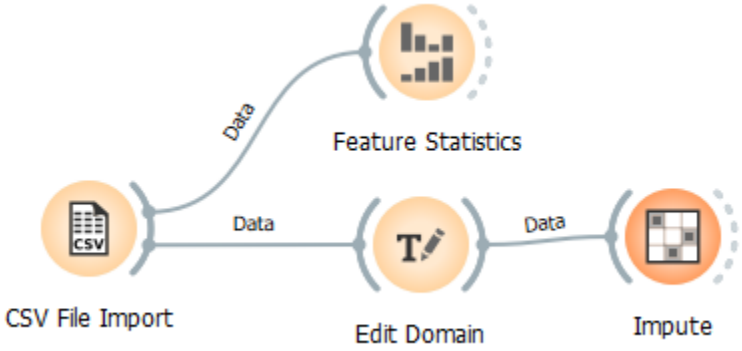
7. สังเกตสัญลักษณ์ N และ C หน้าชื่อตัวแปร เมื่อ N แทน Numerical C แทน Categorical T แทน Time และ S แทน String ซึ่งแทนชนิดข้อมูลของแต่ละตัวแปรที่โปรแกรมกำหนดให้อัตโนมัติ
8. สังเกตจำนวนข้อมูลสูญหาย (Missing Data) จากคอลัมน์ Missing
9. หากต้องการเปลี่ยนแปลงชนิดข้อมูลของตัวแปร สามารถทำได้โดยใช้โมดูล Edit Domain
10. คลิกเลือกโมดูล Edit Domain จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล CSV File Import จากด้าน output เข้าสู่โมดูล Edit Domain ด้าน input ดังรูป



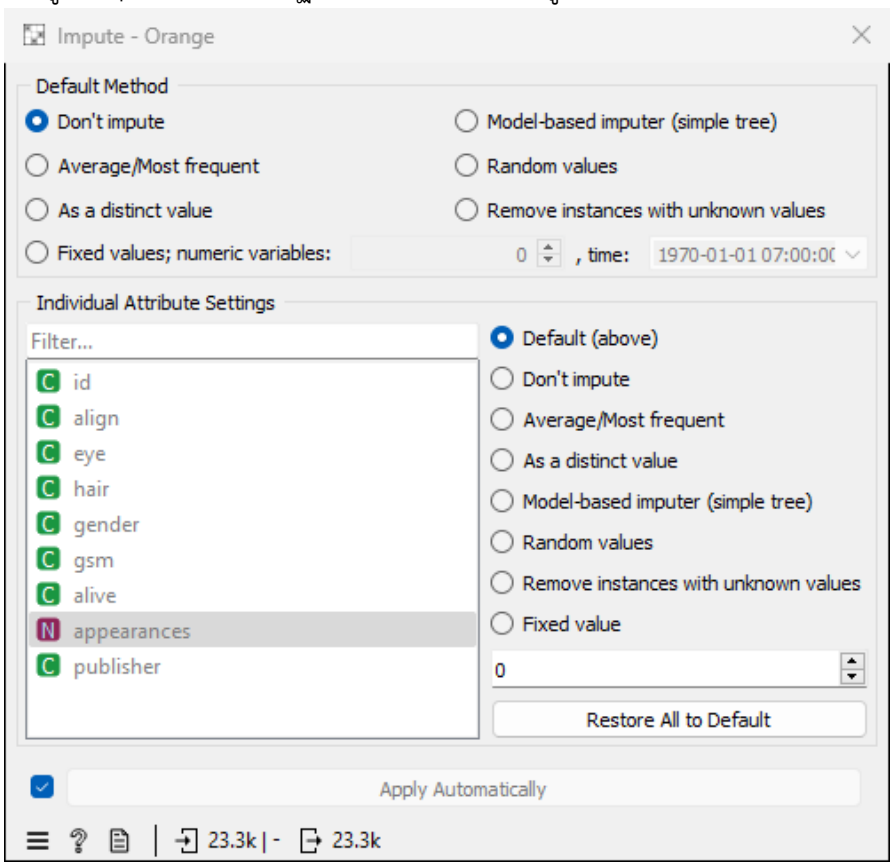
11. ดับเบิลคลิกที่โมดูล Edit Domain จากปรากฏหน้าต่างการตั้งค่า ดังรูป



12. คลิกเลือกตัวแปรที่ต้องการเปลี่ยนแปลงชนิดข้อมูล แล้วเลือกชนิดตัวแปรที่ต้องการที่ช่อง Type
13. จากนั้นคลิกปุ่ม Apply ผลลัพธ์ของโมดูล Edit Domain คือชุดข้อมูลที่เปลี่ยนแปลงชนิดข้อมูลแล้ว
14. การจัดการค่าสูญหาย (Missing Value) ทำได้โดยใช้โมดูล Impute คลิกเลือกโมดูล Impute จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Edit Domain (นำผลลัพธ์จากโมดูล Edit Domain ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Impute ด้าน input ดังรูป



15. ดับเบิลคลิกที่โมดูล Impute จากปรากฏหน้าต่างการตั้งค่า ดังรูป



16. สามารถเลือกวิธีการเติมข้อมูลสูญหายแบบค่าเริ่มต้น (Default) ได้จากส่วน Default Method ในส่วนนี้ วิธีการที่เลือกจะถูกใช้กับทุก ๆ ตัวแปร
17. นอกจากนี้ยังสามารถกำหนดวิธีการเฉพาะบางตัวแปรได้ จากส่วน Individual Attribute Settings โดยคลิกเลือกชื่อตัวแปร แล้วเลือกวิธีการที่จะใช้ด้านขวา ในที่นี้จะกำหนด ดังนี้

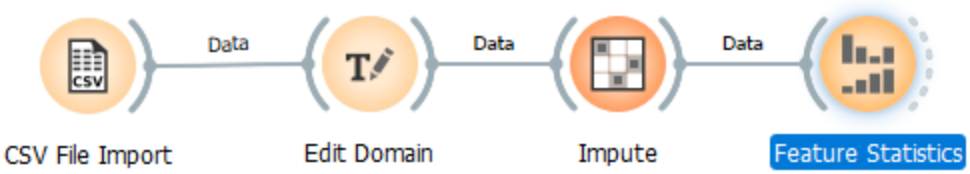
    ตัวแปร appearance เลือกวิธีการ Average/Most frequent

    ตัวแปร align เลือกวิธีการ Average/Most frequent

    ตัวแปร eye เลือกวิธีการ Average/Most frequent

    ตัวแปร hair เลือกวิธีการ Average/Most frequent

    ตัวแปร alive เลือกวิธีการ Fixed value ด้วยค่า Living Characters
18. จากนั้นปิดหน้าต่างตั้งค่านี้ ผลลัพธ์ของโมดูล Impute คือชุดข้อมูลที่ผ่านการจัดการข้อมูลสูญหายตามที่ได้กำหนดไว้แล้ว
19. หากนำโมดูล Feature Statistics มาต่อเข้ากับโมดูล Impute ดังรูป จะพบว่าจำนวนข้อมูลสูญหายของตัวแปร appearance align eye hair และ alive เท่ากับ 0



แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. นำชุดข้อมูล 120 Years of Olympic History athletes and Results จากแฟ้มข้อมูล athlete_events.csv เข้าสู่โปรแกรม Orenge (ทำการเปลี่ยนค่าระบุข้อมูลสูญหายจาก NA เป็นค่า ว่างเปล่า ด้วย)
2. 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	Replace with average
Height	Numeric Feature	Replace with average
Weight	Numeric Feature	Replace with average
Team	Categorical Feature	(ไม่มี Missing Value)
NOC	Categorical Feature	(ไม่มี Missing Value)
Games	String Feature	(ไม่มี Missing Value)
Year	Numeric Feature	(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	Fixed value ด้วยค่า None

สิ่งที่ต้องส่งเป็นการบ้าน ทำการบันทึกไฟล์ workspace ของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_02_id.ows โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th