ปฏิบัติการที่ 1
การเตรียมแฟ้มข้อมูลและการนำเข้าข้อมูลสู่เครื่องมือวิเคราะห์ข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปวิเคราะห์ด้วยเครื่องมือวิเคราะห์ข้อมูลทางวิทยาการข้อมูล
2. เพื่อให้สามารถนำข้อมูลเข้าสู่เครื่องมือวิเคราะห์ข้อมูลเพื่อประมวลผลต่อไปได้

1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ชุดข้อมูล 2004 New Car and Truck (สำหรับการทำปฏิบัติ)

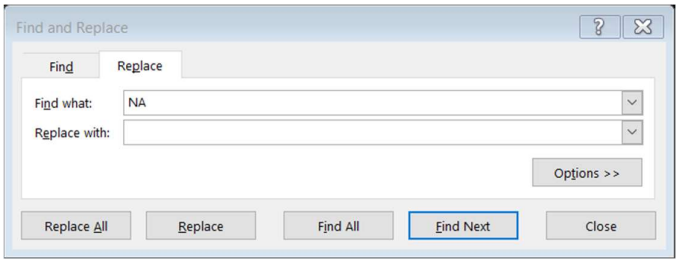
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

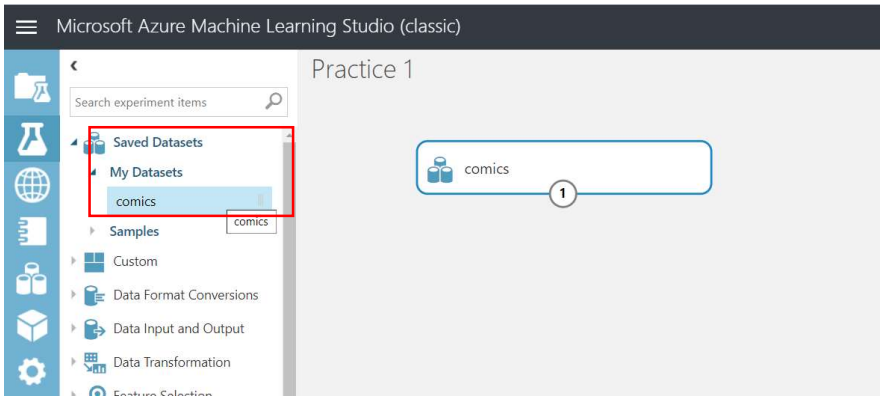
1. เปิดแฟ้มข้อมูล comics.csv ด้วยโปรแกรม Microsoft Excel แถวข้อมูลแรกระบุชื่อคอลัมน์ (ตัวแปร) แถวข้อมูลที่ 2 ถึง 23,273 เป็นระเบียนข้อมูลของตัวละครแต่ละตัว

	A	B	C	D	E	F	G	H	I	J	K
1	name	id	align	eye	hair	gender	gsm	alive	appearanc	first_appe	publisher
2	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hai	Male	NA	Living Cha	4043	Aug-62	marvel
3	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hai	Male	NA	Living Cha	3360	Mar-41	marvel
4	Wolverine (James \"Logan\" Howl	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Cha	3061	Oct-74	marvel
5	Iron Man (Anthony \"Tony\" Stark	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Cha	2961	Mar-63	marvel
6	Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Cha	2258	Nov-50	marvel
7	Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Cha	2255	Nov-61	marvel
8	Reed Richards (Earth-616)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Cha	2072	Nov-61	marvel
9	Hulk (Robert Bruce Banner)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Cha	2017	May-62	marvel

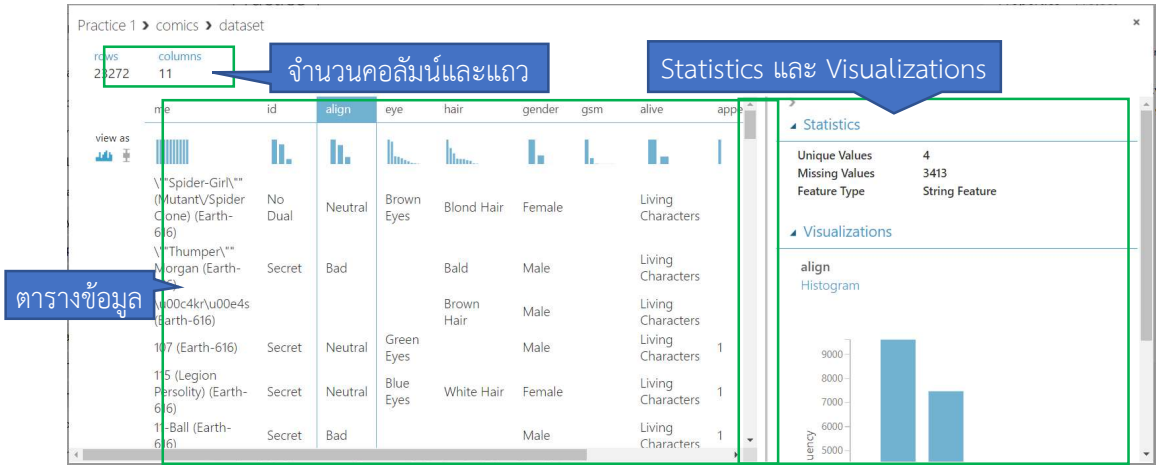
2. ทำการแทนที่ข้อมูลในเซลล์ข้อมูลที่มีค่า NA ด้วยค่า ว่างเปล่า <u>เนื่องจาก ML Studio กำหนดให้เซลล์ข้อมูลใดที่เป็นข้อมูลสูญหาย ต้องเป็นเซลล์ว่างเปล่า</u>



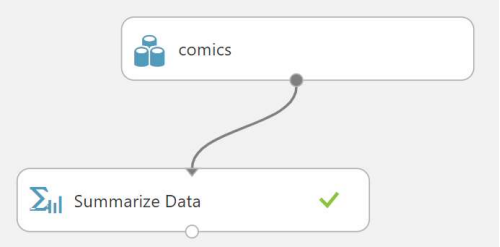
3. บันทึกแฟ้มข้อมูล comics.csv
4. เปิดโปรแกรม ML Studio (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การเข้าใช้งานเบื้องต้น)
5. นำชุดข้อมูลจากแฟ้มข้อมูล comics.csv เข้าสู่ ML Studio (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การนำชุดข้อมูลจากคอมพิวเตอร์ส่วนตัวเข้าสู่ ML Studio) โดยกำหนดชื่อชุดข้อมูลเป็น “comics” และเลือกชนิดของชุดข้อมูลเป็น “Generic CSV File with a header (.csv)”
6. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 1” (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การสร้างการทดลอง)
7. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace



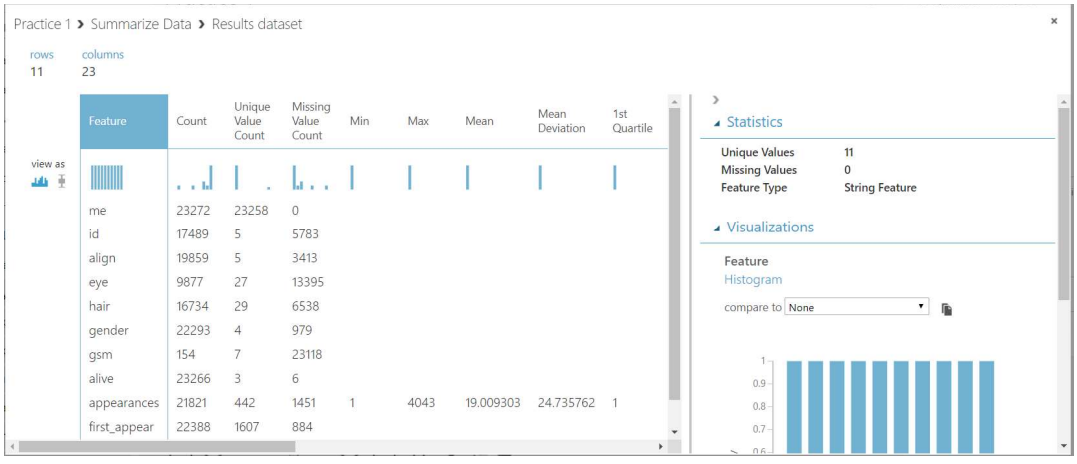
8. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังรูป



9. ในหน้าต่างแสดงข้อมูล ประกอบด้วยส่วนต่างๆ ดังนี้
    - จำนวนคอลัมน์และแถว แสดงจำนวนตัวแปรและระเบียนข้อมูลตามลำดับ
    - ตารางข้อมูล ประกอบด้วย
        - ชื่อคอลัมน์ เมื่อคลิกที่ชื่อคอลัมน์ข้อมูลในส่วน Statistics และ Visualizations ซึ่งอธิบายข้อมูลทางสถิติของตัวแปรนั้นในชุดข้อมูล
        - การกระจายของข้อมูลในแต่ละตัวแปร แสดงใต้ชื่อคอลัมน์
        - ข้อมูลแต่ละระเบียน (อาจไม่แสดงครบทุกข้อมูล)
    - Statistics และ Visualizations แสดงค่าทางสถิติและการกระจายของข้อมูล
10. สรุปข้อมูลในชุดข้อมูล comics โดยลากโมดูล Summarize Data ที่อยู่ภายใต้ Statistical Functions → ในหน้าต่างย่อย Modules มาวางบน Workspace (ไว้ด้านล่างกล่องโมดูลชุดข้อมูล comics)
11. คลิกที่ โหนดส่วนต่อประสานข้อมูลออก ของกล่องโมดูลชุดข้อมูล comics แล้วลากวางยัง โหนดส่วนต่อประสานข้อมูลเข้าของกล่องโมดูลชุดข้อมูล Summarize Data



12. คลิกคำสั่ง **RUN**
13. เมื่อโปรแกรมประมวลผลเรียบร้อย ดูผลลัพธ์ของโมดูล Summarize Data โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป



3. แบบฝึกปฏิบัติการ

ให้นักศึกษานำชุดข้อมูล 2004 New Car and Truck จากแฟ้มข้อมูล cars04.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “Cars04” และสร้างการทดลอง กำหนดชื่อเป็น “Lab 1” โดยให้นำชุดข้อมูล Cars04 เข้าสู่การทดลอง และใช้โมดูล Summarize Data ในการสรุปข้อมูลในชุดข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_01_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th