Homework02 : Data Preparation

Student Name : Puriwat Sangrawee
Student ID: 65056071

1. การเตรียมข้อมูลของนักศึกษาจากข้อมูลความสนใจส่วนตัว
จากข้อมูลความสนใจส่วนตัวของนักศึกษาจำนวน 66 คน ซึ่งประกอบไปด้วย 5 แอตทริบิวต์ (ไฟล์: *StudentGender.csv*) ดังต่อไปนี้

Attribute	คำอธิบาย
Favorite Color	สีที่ชอบ
Favorite Music Genre	แนวดนตรีที่ชอบ
Favorite Beverage	เครื่องดื่มแอลกอฮอล์ที่ชอบ
Favorite Soft Drink	เครื่องดื่มที่ไม่มีส่วนผสมของแอลกอฮอล์ที่ชอบ
Gender	เพศ

ให้นักศึกษาทำการจัดเตรียมข้อมูลโดยใช้เครื่องมือใน Azure Machine Learning Studio เพื่อให้พร้อมกับการนำไปใช้งานโดย

1) ทำการจัดการกับข้อมูลที่สูญหาย (Missing Values) ในแต่ละแอตทริบิวต์ด้วยค่าที่เหมาะสม นักศึกษาจะเลือกใช้วิธีใด..... เพราะอะไร .....

Ans>> เลือกใช้วิธี Replace with Mode เนื่องจากแต่ละ Missing Value ถูกจัดอยู่ในกลุ่ม Categorical data ทั้งหมด

2) จงหาว่าแอตทริบิวต์ที่มีความสำคัญต่อตัวแปรตามคือเพศของนักศึกษา จงตอบคำถามต่อไปนี้

- ควรใช้เทคนิคอะไรในการพิจารณาหาแอตทริบิวต์ที่สำคัญดังกล่าว เพราะอะไร

Ans>> เทคนิค Chi Square เนื่องจากทั้ง Feature และ Target Column ต่างก็เป็นข้อมูลประเภท Categorical (ซึ่งข้อมูลลักษณะดังกล่าว หากทำการวิเคราะห์ด้วย Azure ML มีแค่วิธีนี้เท่านั้น)

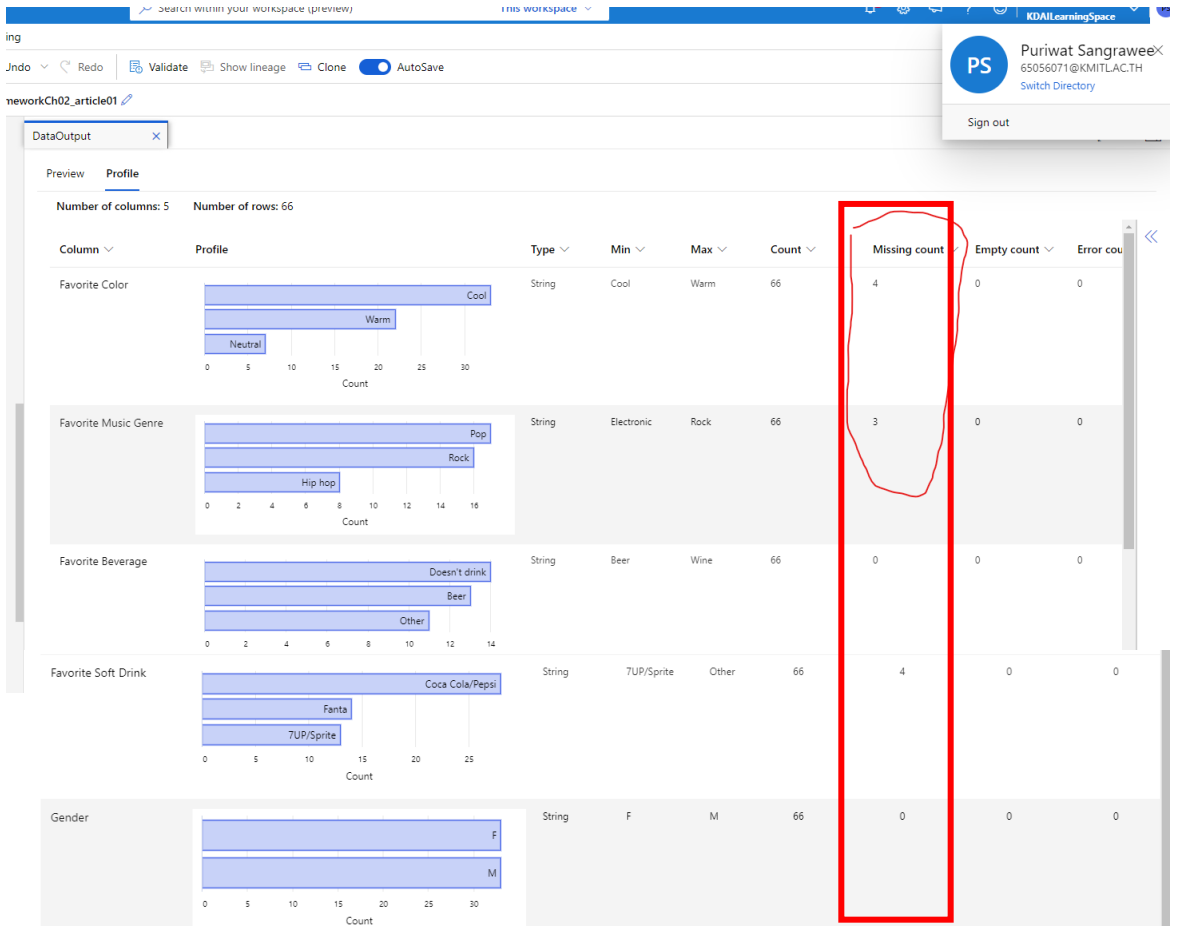
- จำนวน 2 แอตทริบิวต์แรกคืออะไร

Ans>> Favorite Music Genre / Favorite Beverage

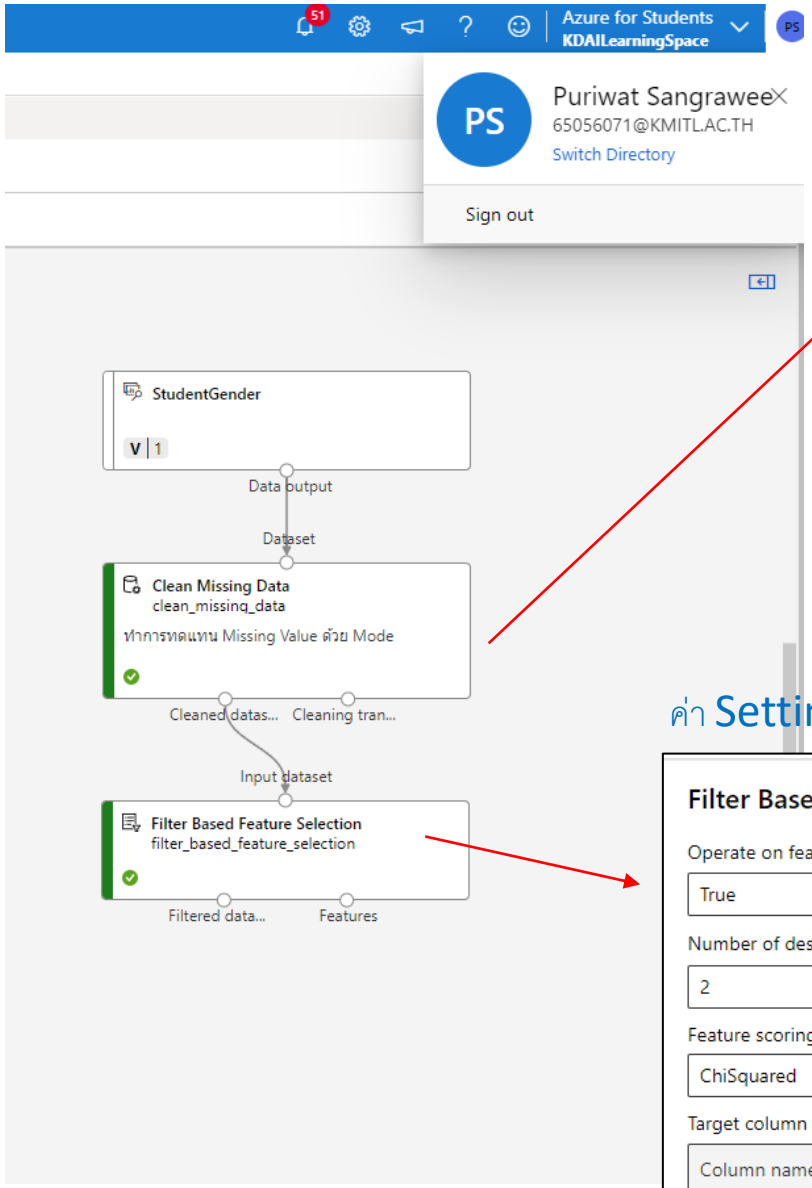
3) แสดงหน้าจอการเตรียมข้อมูล

Ans>> ดู Flowchart ในหน้า 2

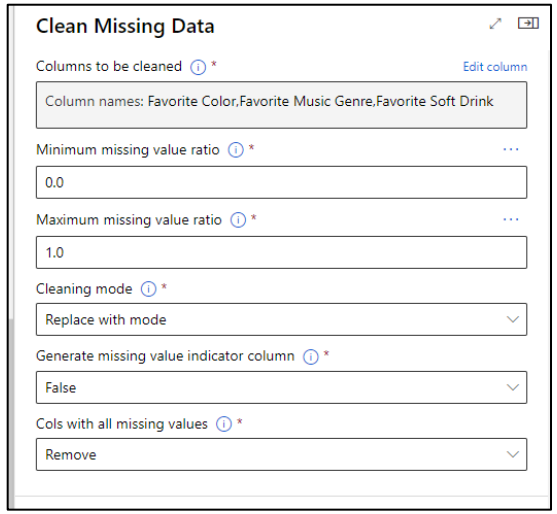
Show Missing Value detected



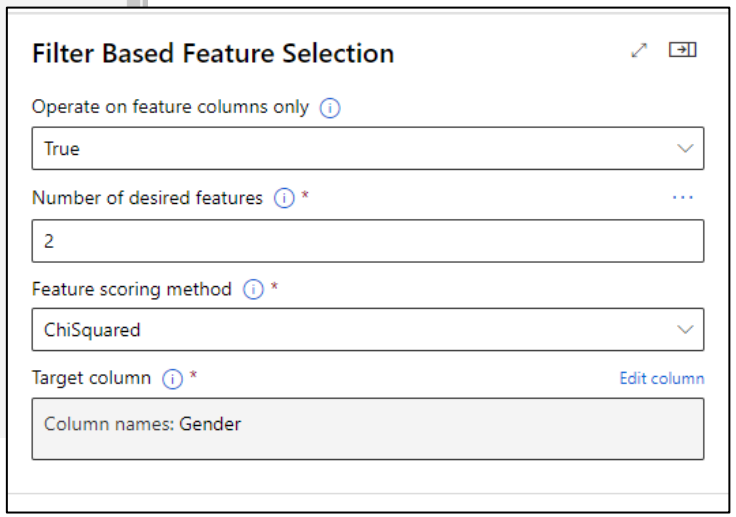
จากภาพ จะเห็นว่าทุกคอลัมน์ที่มี Missing Value (Favorite Color , Favorite Music Genre, และ Favorite Soft Drink ต่างก็เป็น Categorical ทั้งหมด ดังนั้นเราสามารถใช้วิธี Replace with Mode ได้)



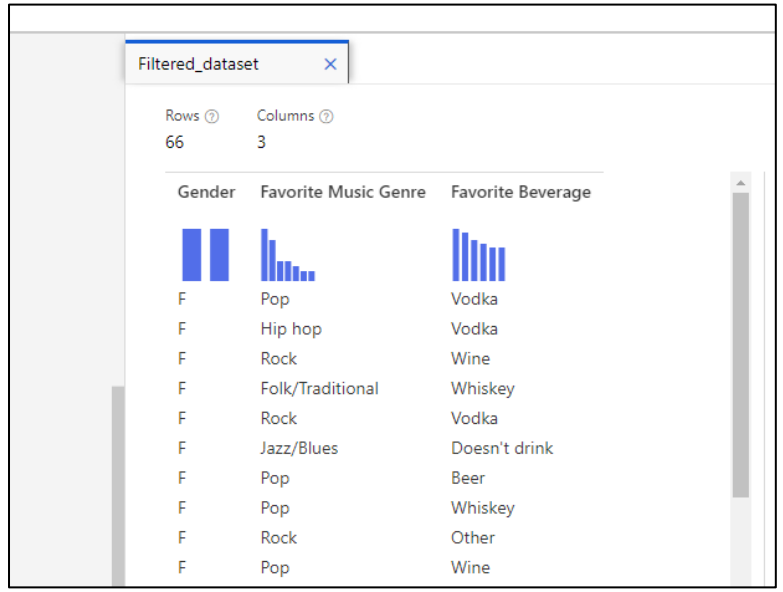
ค่า Setting : Replace Missing Value



ค่า Setting : Feature Selection



ค่า Final Result



2. การเตรียมข้อมูลเพื่อวิเคราะห์คุณภาพของไวน์

จากข้อมูลคุณสมบัติของไวน์จำนวน 6,497 รายการ ซึ่งประกอบไปด้วยแอตทริบิวต์จำนวน 12 แอตทริบิวต์ ทั้ง 12 แอตทริบิวต์เป็นข้อมูลที่บ่งบอกถึงคุณภาพของไวน์ (ไฟล์: *WineQuality_Dataset.csv*) ดังนี้

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality คุณภาพของไวน์ถูกแบ่งออกเป็น 11 ระดับตามคะแนน (quality score 0-10)

ให้นักศึกษาทำการจัดเตรียมข้อมูลเพื่อใช้ในการวิเคราะห์ข้อมูลโดย

1) ทำการจัดการกับข้อมูลที่สูญหาย (Missing Values) ในแต่ละแอตทริบิวต์ด้วยค่าที่เหมาะสม นักศึกษาจะเลือกใช้วิธีใด Ans>> เลือกใช้วิธี Remove Entire Row

2) จากข้อมูลดังกล่าวควรแปลงข้อให้อยู่ในรูปมาตรฐาน (Normalization) หรือไม่? เพราะอะไร?

Ans>> ควร เนื่องจากหากลองพิจารณาใน record เดียวกัน จะพบว่าค่าค่อนข้างกว้างมาก เช่น Data จาก chloride และ total Sulphur ที่ห่างกันมากกว่า 1000 เท่า เป็นต้น ซึ่งจากลักษณะดังกล่าว หากไม่ทำการ Normalize จะทำให้หากนำข้อมูลไปทำนาย จะทำให้ตัวแปรใดตัวแปรหนึ่งมีอิทธิพลมากหรือน้อยกว่าที่ควรมีจริง

3) ทำการเลือกแอตทริบิวต์ที่มีความสำคัญต่อการทำนายคุณภาพของไวน์**จำนวน 4 แอตทริบิวต์แรก** ให้นักศึกษาเลือกวิธีการที่เหมาะสมเองโดยพิจารณาจากข้อมูลในชุดข้อมูล

Ans>>4 Attributes แรกคือ Alcohol , Density , Volatile Acidity และ Chlorides

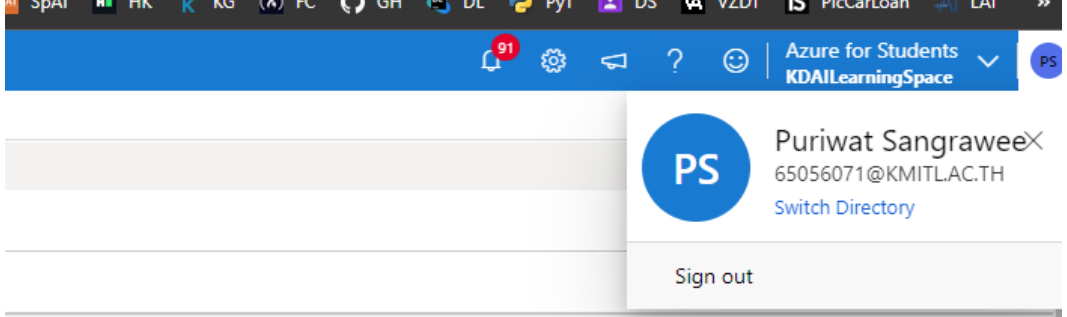
4) แสดงหน้าจอการเตรียมข้อมูล

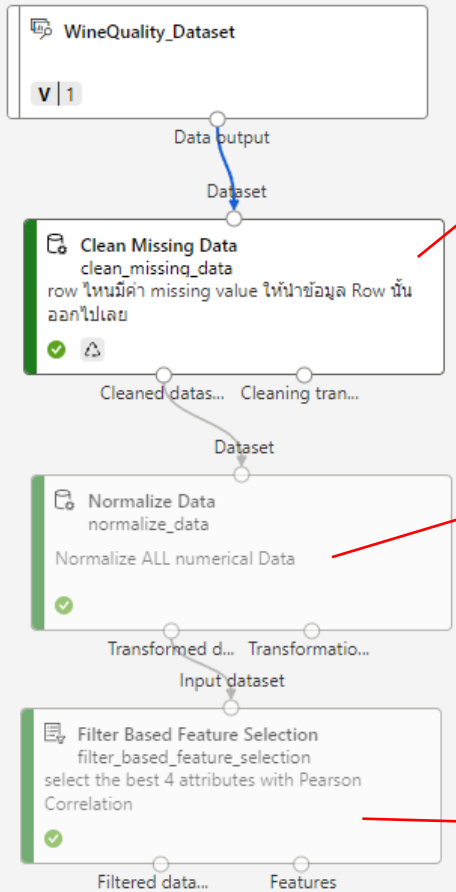
Ans>> ดู Flowchart ในหน้า 4

Show Missing Value detected

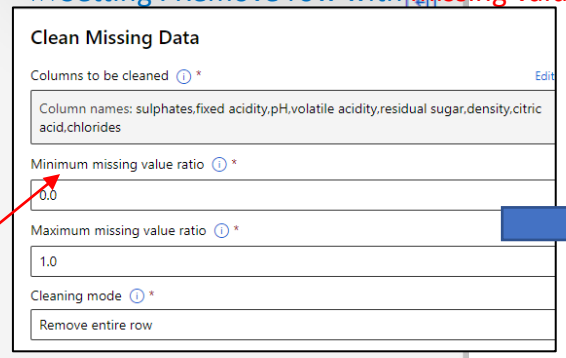
จากภาพ จะเห็นว่าทั้ง 8 **Columns** ที่มี **Missing Value** (sulphates , fixed acidity ,pH , volatile acidity ,residual sugar, density , citric acid , chlorides) เมื่อนับจำนวน records ที่ มี Missing Value จะพบว่า**มี** 89 **records จากจำนวน** total records 6497 ซึ่งตัวที่มี missing มีเพียง 1.37% ซึ่งในทางทฤษฎี หาก missing value น้อยกว่า 20% สามารถนำออกจากการพิจารณาได้เลย

ดังนั้นเราจะเอาออกเลย





ค่า Setting : Remove row with missing value

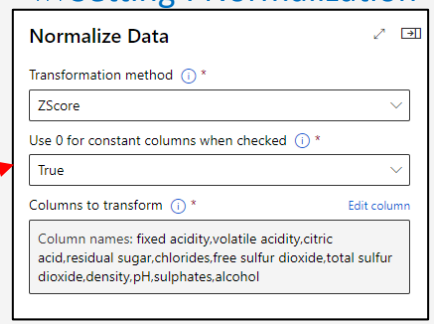


Result : หลังลบ row ที่มี missing value

Rows 6,418 Columns 12

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.3	0.22	0.3	8.2	0.047	42	207	0.9966	3.33	0.46	9.5	6
7.1	0.43	0.61	11.8	0.045	54	155	0.9974	3.11	0.45	8.7	5
7.1	0.44	0.62	11.8	0.044	52	152	0.9975	3.12	0.46	8.7	6
6.8	0.25	0.31	13.3	0.05	69	202	0.9972	3.22	0.48	9.7	6
7.1	0.43	0.61	11.8	0.045	54	155	0.9974	3.11	0.45	8.7	5
7.1	0.44	0.62	11.8	0.044	52	152	0.9975	3.12	0.46	8.7	6
6.9	0.24	0.33	1.7	0.035	47	136	0.99	3.26	0.4	12.6	7

ค่า Setting : Normalization



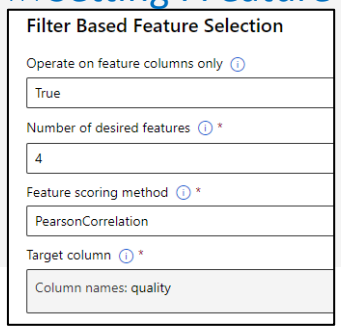
Result : หลัง Normalized

Transformed_dataset

Rows 6,418 Columns 12

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
-0.170008	-0.423523	0.284174	3.213308	-0.316443	0.819921	0.966152	2.098866	-1.35992	-0.546068	-1.422425	6
-0.708166	-0.241671	0.146631	-0.805753	-0.202188	-0.926988	0.293585	-0.23127	0.507651	-0.277676	-0.835781	6
0.675667	-0.362906	0.559262	0.309484	-0.173625	-0.025357	-0.325884	0.134894	0.258641	-0.613166	-0.332944	6
-0.016249	-0.665992	0.009087	0.646159	0.054884	0.932625	1.249337	0.301332	-0.177125	-0.881559	-0.500556	6
-0.016249	-0.665992	0.009087	0.646159	0.054884	0.932625	1.249337	0.301332	-0.177125	-0.881559	-0.500556	6
0.675667	-0.362906	0.559262	0.309484	-0.173625	-0.025357	-0.325884	0.134894	0.258641	-0.613166	-0.332944	6
-0.785045	-0.120436	-1.091264	0.330526	-0.316443	-0.025357	0.364382	0.068319	-0.239378	-0.411872	-0.751975	6

ค่า Setting : Feature Selection



Final Result :

Rows 6,418 Columns 5

quality	alcohol	density	volatile acidity	chlorides
6	-1.422425	2.098866	-0.423523	-0.316443
6	-0.835781	-0.23127	-0.241671	-0.202188
6	-0.332944	0.134894	-0.362906	-0.173625
6	-0.500556	0.301332	-0.665992	0.054884
6	-0.500556	0.301332	-0.665992	0.054884
6	-0.332944	0.134894	-0.362906	-0.173625

3. การเตรียมข้อมูลเพื่อจำแนกลูกค้าผู้ซื้อรถยนต์

จากข้อมูลลูกค้าผู้ซื้อรถยนต์ที่บริษัทแห่งหนึ่งได้ทำการรวบรวมขึ้นจำนวน 3,592 คน ซึ่งประกอบไปด้วยแอตทริบิวต์ จำนวน 10 แอตทริบิวต์ (Automobile_Customer_Segmentation_Binary.csv) โดยแอตทริบิวต์ Segmentation คือ แอตทริบวต์ที่ใช้ในการจำแนกลูกค้าออกเป็น 2 กลุ่ม คือ A และ D ให้นักศึกษาทำการจัดเตรียมข้อมูลเพื่อใช้ในการวิเคราะห์กลุ่มของลูกค้าดังนี้

1) ชุดข้อมูลนี้มีข้อมูลสูญหาย (Missing Value) หรือไม่ ถ้ามี แอตทริบิวต์ใดบ้างที่มีข้อมูลสูญหาย และนักศึกษามีวิธีการดำเนินการกับข้อมูลที่สูญหายนั้นอย่างไร

Ans>> มี รายละเอียด Attribute และวิธีแก้ไขตามตารางสรุปด้านล่าง

NO.	Col Name with Missing Value	How to Solve
1	FamilySize	Replace with Mean
2	Married	Replace with Mode (FALSE)
3	Profession	Replace with Mode (Healthcare)
4	Graduate	Replace with Mode (FALSE)

2) หากผู้บริหารต้องการ**แบ่งช่วงอายุของลูกค้าออกเป็น 4 ช่วงอายุ** โดยให้มีจำนวนลูกค้าในแต่ละช่วงอายุเท่ากันหรือใกล้เคียงกัน นักศึกษาจะมีวิธีการดำเนินการกับข้อมูลอย่างไร

Ans>> ก่อนจะนำข้อมูลไปใช้ต่อ จะต้องทำการ Transform data ด้วยเทคนิค Quantiles ก่อน โดยเลือกคอลัมน์ Age มาทำด้วยเทคนิคดังกล่าว โดยกำหนด bins เป็น 4 ช่วงอายุลูกค้าก็จะถูกแบ่งออกเป็น 4 ช่วง โดยมีลูกค้าในแต่ละช่วงจำนวนเท่า ๆ กันตามต้องการ

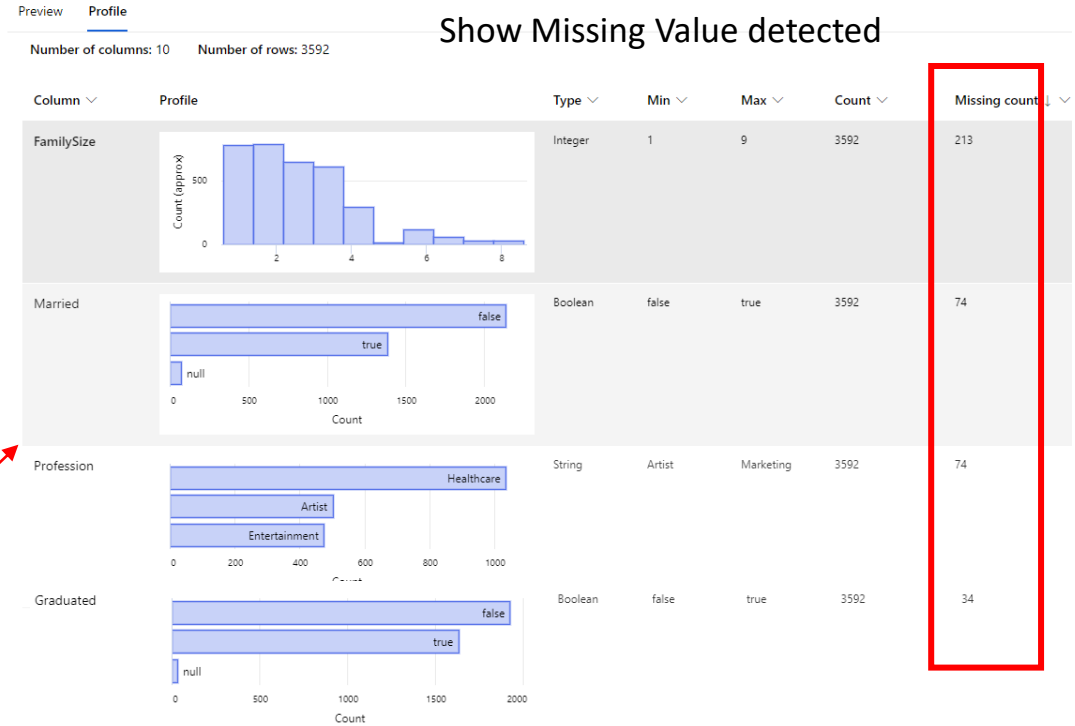
3) ชุดข้อมูลนี้เกิดปัญหา Imbalanced Data หรือไม่ เพราะเหตุใด และนักศึกษาทำการแก้ปัญหา Imbalanced Data อย่างไร

Ans>> มี result ของ segmentation D มีปริมาณมากกว่า A ประมาณ 1.71 เท่า ทำการแก้ไขด้วยการจำลองข้อมูลเพิ่มมา โดยใช้ การ setup SMOTE Percentage = 170% เป็นค่าเริ่มต้น และค่อยๆ ปรับลงมาจนได้จำนวนข้อมูลที่ใกล้เคียงกัน (ค่าสุดท้ายที่ปรับคือ 70%)

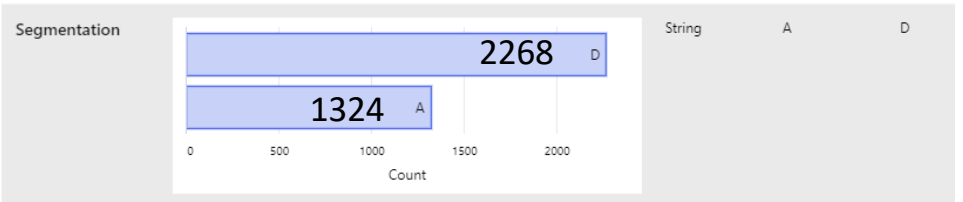
4) แสดงหน้าจอการเตรียมข้อมูล

Ans>> ดู Flowchart ในหน้า 6

Show Missing Value detected



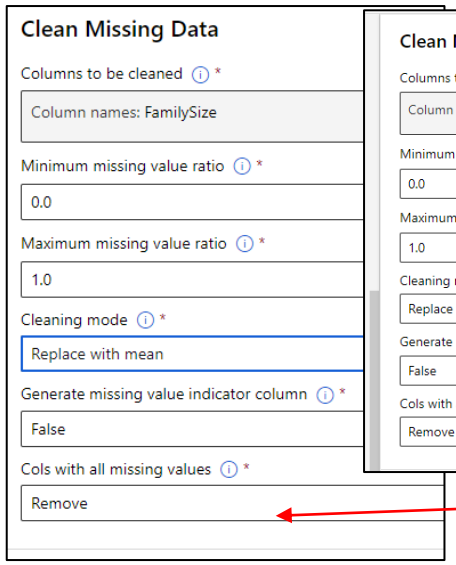
Segmentation

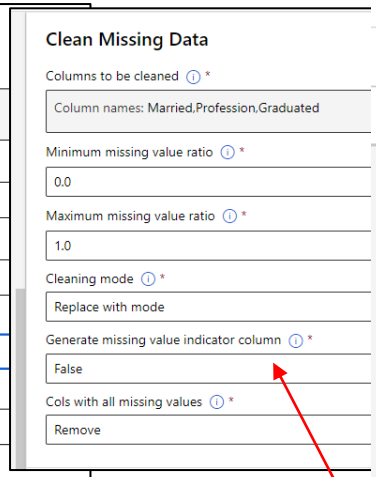


- Equal-depth (frequency) : แต่ละบินจะมีจำนวนข้อมูลเท่า ๆ กัน บางครั้งเรียกวิธีการนี้ว่า “Quantiles”

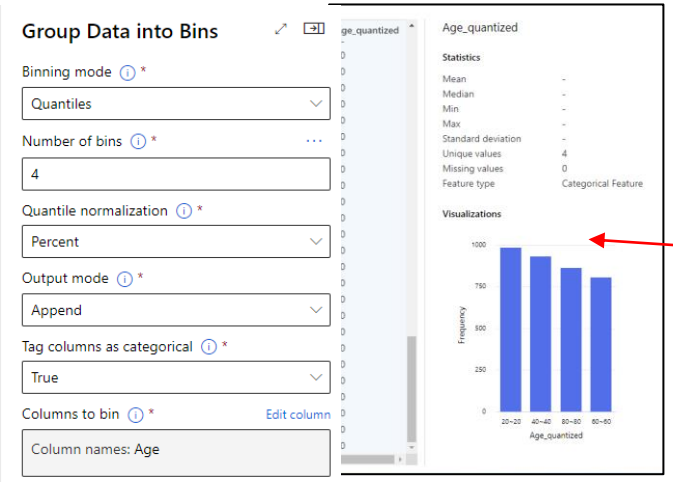
$$m = \frac{n}{N}, \qquad N \le n$$

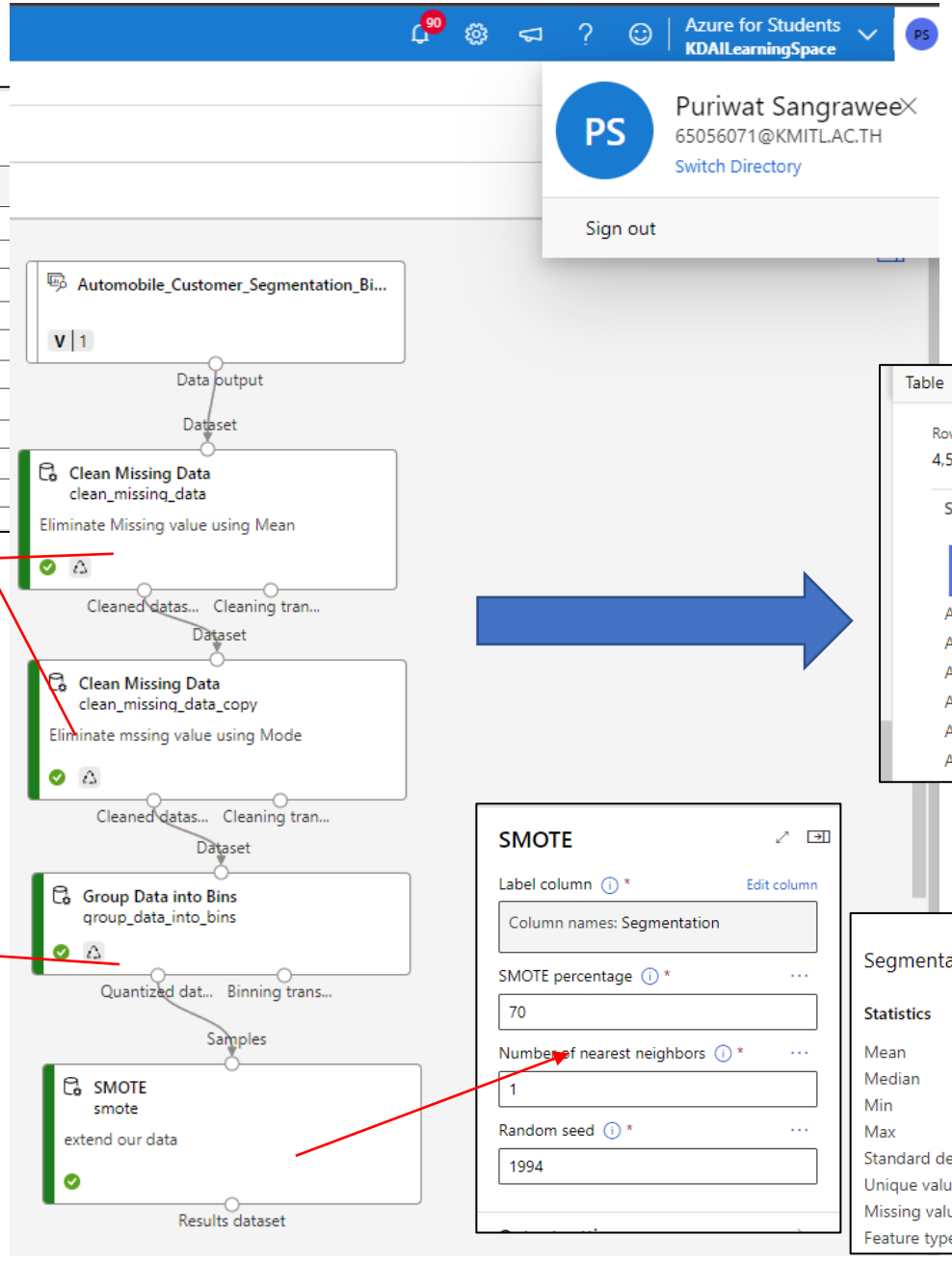
ค่า Setting : Replace Missing value





ค่า Setting : Group data to Bin





Final Result