ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA: CÔNG NGHỆ THÔNG TIN

BÁO CÁO
LAB01 – Preprocessing

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG – CQ2017

Sinh viên thực hiện:
Nguyễn Văn Thìn – 1712787
Lê Bá Quyền – 1712713
Nguyễn Lê Trường Thành - 1712775

Q5, ngày 6 tháng 10 năm 2019

I. THÔNG TIN NHÓM

Member	MSSV	Email	Tự đánh giá cá nhân
Nguyễn Văn Thìn	1712787	vanthin7111999@gmail.com	90%
Lê Bá Quyền	1712713	quyenleba2291@gmail.com	90%
Nguyễn Lê Trường Thành	1712775	provip218@gmail.com	90%

(Điểm thành viên chia đều)

Quá trình làm việc nhóm- Phân chia task cho từng member	
Phần trả lời câu hỏi	**Thành:** deadline 26/9 đã completed
Phần code	**Quyền:** Code BT1 **Thìn:** Code BT2 21/9 -> 27/9 >>> Tìm hiểu cú pháp python 28/9 -> 30/9 >>> Code 1 số yêu cầu 1/10 -> 5/10 >>> Hoàn thiện các phần còn lại 6/10 chỉnh sửa phần còn thiếu – summit 23h trên moodle

Những phần đã làm được:

Trả lời câu hỏi tự luận

Câu hỏi từ 1->9

Nội dung thực hiện cài đặt(25 điểm)

1. **Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản (15 điểm)**

a) Chuẩn hóa min-max trên danh sách thuộc tính chỉ định.
b) Chuẩn hóa Z-scores trên danh sách thuộc tính chỉ định.
c) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ rộng trên danh sách thuộc tính chỉ định.
d) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ sâu trên danh sách thuộc tính chỉ định.
e) Xóa các mẫu dữ liệu thiếu giá trị trên danh sách thuộc tính chỉ định.
f) Điền giá trị bị thiếu trên danh sách thuộc tính chỉ định, giá trị được điền là giá trị trung bình (mean) của thuộc tính nếu đó là thuộc tính số hoặc điền giá trị có tần số xuất hiện cao nhất (mode) nếu là thuộc tính rời rạc.

2. Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước (10 điểm)

1.Xóa các mẫu rỗng.
2. Xóa các mẫu bị trùng lắp
3. Chuyển diện tích về km2
4. Sử dụng chương trình đã cài đặt ở phần B-1. để xóa các mẫu bị thiếu diện tích

Những phần chưa làm được:

Câu 10 phần trả lời câu hỏi

II. TRẢ LỜI CÂU HỎI TỰ LUẬN

YÊU CẦU 1 – TẠO TẬP TIN ARFF TỪ TẬP DỮ LIỆU COURSE RATINGS

1. Tập dữ liệu được đọc vào Weka thành công hay không? Nếu có, trả lời câu hỏi tiếp theo. Nếu không, cho biết lỗi gặp phải và bạn đã sửa lỗi đó như thế nào?

- Tập dữ liệu được đọc vào weka thành công.

2. Sau khi đọc dữ liệu thành công, quan sát thông tin thể hiện trên giao diện Explorer và trả lời những câu hỏi sau đây

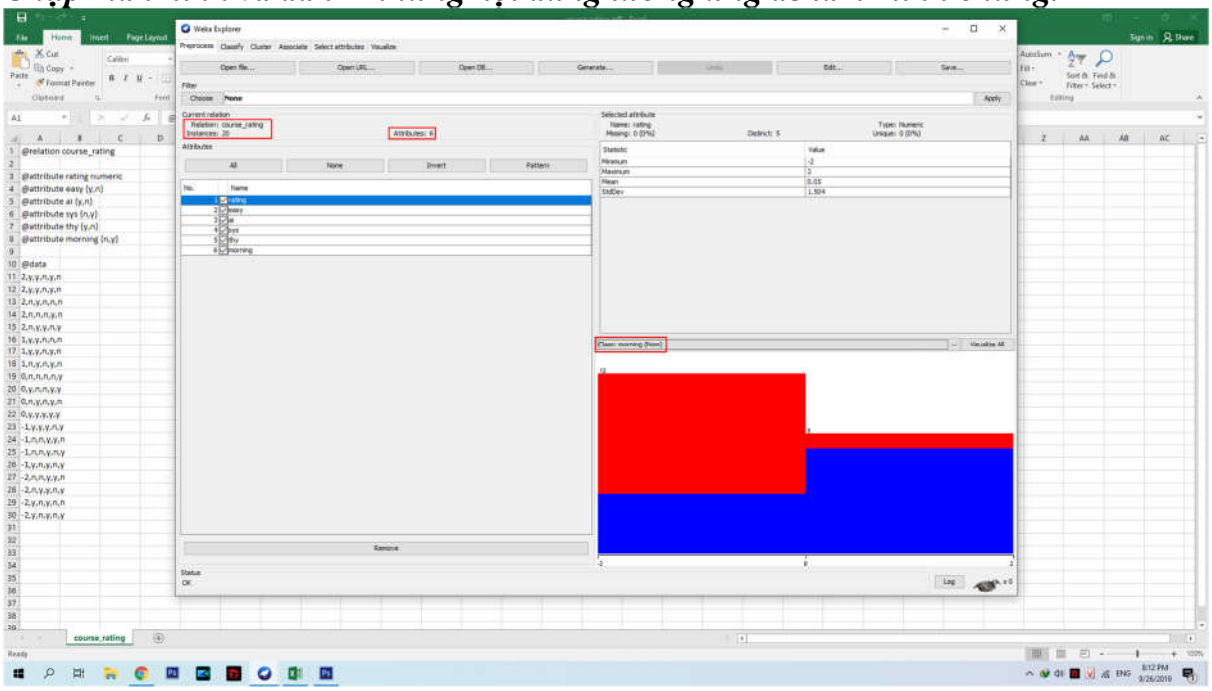
- Tên của mối quan hệ (relation) trong dữ liệu là gì?
course_rating

- Tập dữ liệu có bao nhiêu mẫu (instances)?
instances: 20

- Tập dữ liệu có bao nhiêu thuộc tính (attributes)?
attributes: 6

- Thuộc tính nào trong tập dữ liệu là thuộc tính lớp (class)?
rating

Chụp màn hình và đán những nội dung tương ứng để làm minh chứng.



3. Bạn có nhận thấy điều gì đáng chú ý khi quan sát các thông tin thống kê và đồ thị trình diễn?

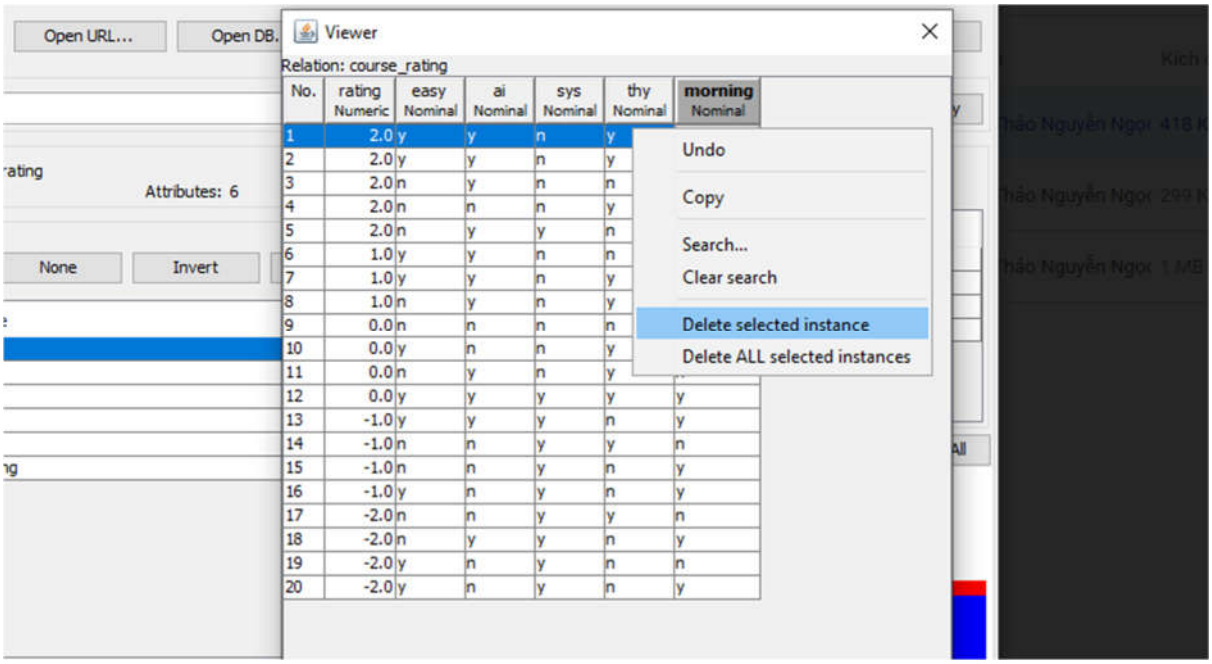
Đồ thị trình diễn ngoài chia ra theo giá trị thì còn được chia thành các màu (đỏ, xanh) trong từng giá trị cụ thể

4. Điều gì xảy ra nếu thuộc tính lớp là thuộc tính đầu tiên bên trái? So sánh nội dung các đồ thị trình diễn do chức năng Visualize All (nằm góc dưới bên phải trong tab Preprocess) cung cấp, trong trường hợp thuộc tính lớp là cột đầu tiên và trong trường hợp thuộc tính lớp là cột cuối cùng. Lưu ý rằng tại bước này ta chỉ quan sát thông tin trình diễn dữ liệu chứ không làm bất kỳ thao tác gì ảnh hưởng đến nội dung dữ liệu.

Nếu để thuộc tính lớp là cột đầu tiên thì thì ***Visualize All*** sẽ ra được đồ thị thì theo thuộc tính lớp morning. Còn nếu để ở cuối thì sẽ ko ra được đồ thị

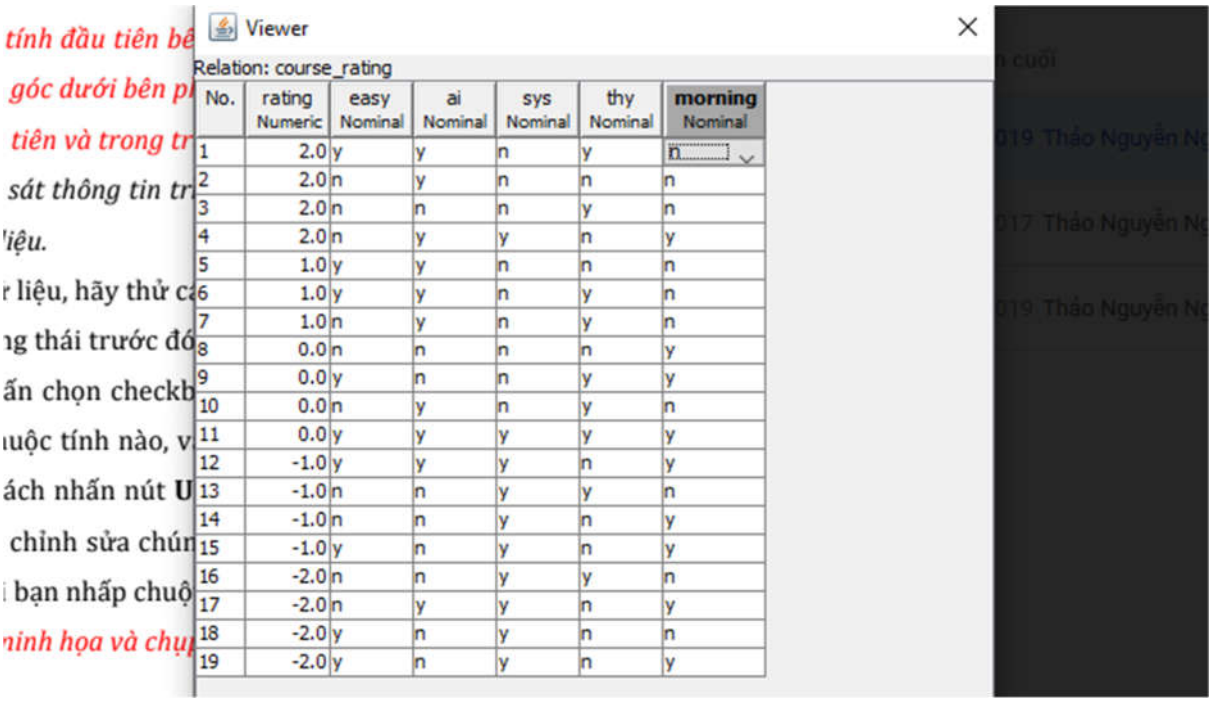
5. Với mỗi thao tác nêu trên, tùy ý thực hiện minh họa và chụp màn hình làm minh chứng

Xóa:



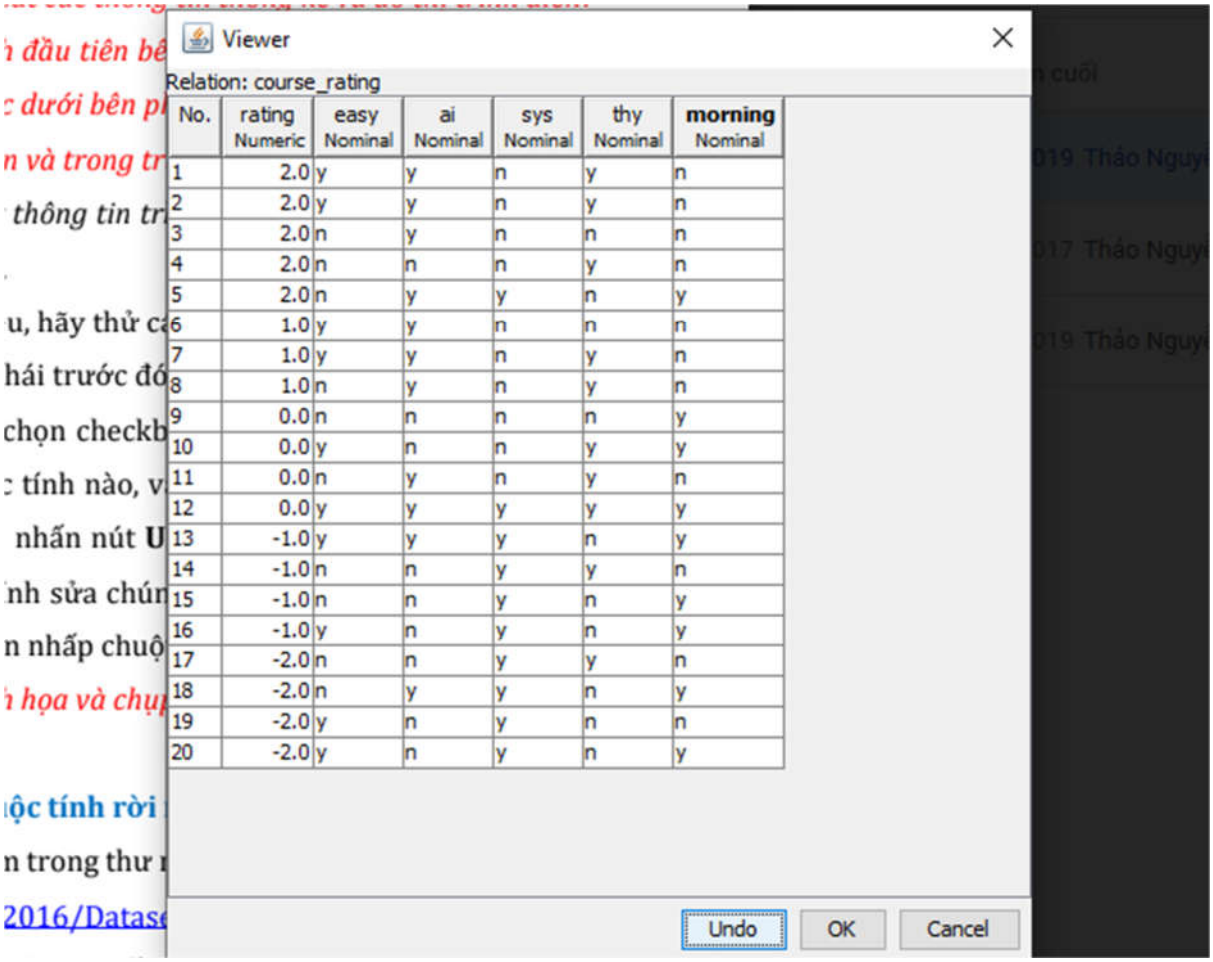
No.	rating Numeric	easy Nominal	ai Nominal	sys Nominal	thy Nominal	morning Nominal
1	2.0	y	y	n	y	
2	2.0	y	y	n	y	
3	2.0	n	y	n	n	
4	2.0	n	n	n	y	
5	2.0	n	y	y	n	
6	1.0	y	y	n	n	
7	1.0	y	y	n	y	
8	1.0	n	y	n	y	
9	0.0	n	n	n	n	
10	0.0	y	n	n	y	
11	0.0	n	y	n	y	
12	0.0	y	y	y	y	y
13	-1.0	y	y	y	n	y
14	-1.0	n	n	y	y	n
15	-1.0	n	n	y	n	y
16	-1.0	y	n	y	n	y
17	-2.0	n	n	y	y	n
18	-2.0	n	y	y	n	y
19	-2.0	y	n	y	n	n
20	-2.0	y	n	y	n	y

Sau khi xóa:

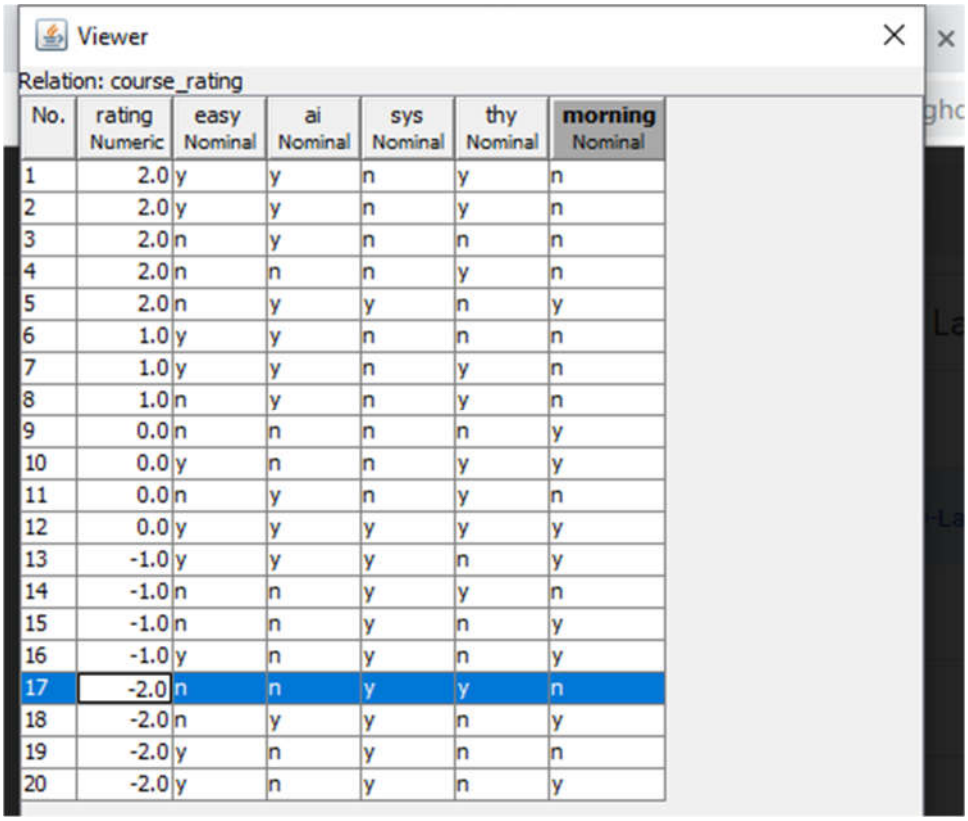


No.	rating Numeric	easy Nominal	ai Nominal	sys Nominal	thy Nominal	morning Nominal
1	2.0	y	y	n	y	n
2	2.0	n	y	n	n	n
3	2.0	n	n	n	y	n
4	2.0	n	y	y	n	y
5	1.0	y	y	n	n	n
6	1.0	y	y	n	y	n
7	1.0	n	y	n	y	n
8	0.0	n	n	n	n	y
9	0.0	y	n	n	y	y
10	0.0	n	y	n	y	n
11	0.0	y	y	y	y	y
12	-1.0	y	y	y	n	y
13	-1.0	n	n	y	y	n
14	-1.0	n	n	y	n	y
15	-1.0	y	n	y	n	y
16	-2.0	n	n	y	y	n
17	-2.0	n	y	y	n	y
18	-2.0	y	n	y	n	n
19	-2.0	y	n	y	n	y

Undo:


Viewer
Relation: course_rating
No. | rating Numeric | easy Nominal | ai Nominal | sys Nominal | thy Nominal | morning Nominal
1 | 2.0 | y | y | n | y | n
2 | 2.0 | y | y | n | y | n
3 | 2.0 | n | y | n | n | n
4 | 2.0 | n | n | n | y | n
5 | 2.0 | n | y | y | n | y
6 | 1.0 | y | y | n | n | n
7 | 1.0 | y | y | n | y | n
8 | 1.0 | n | y | n | y | n
9 | 0.0 | n | n | n | n | y
10 | 0.0 | y | n | n | y | y
11 | 0.0 | n | y | n | y | n
12 | 0.0 | y | y | y | y | y
13 | -1.0 | y | y | y | n | y
14 | -1.0 | n | n | y | y | n
15 | -1.0 | n | n | y | n | y
16 | -1.0 | y | n | y | n | y
17 | -2.0 | n | n | y | y | n
18 | -2.0 | n | y | y | n | y
19 | -2.0 | y | n | y | n | n
20 | -2.0 | y | n | y | n | y
Undo
OK
Cancel

Chỉnh sửa:

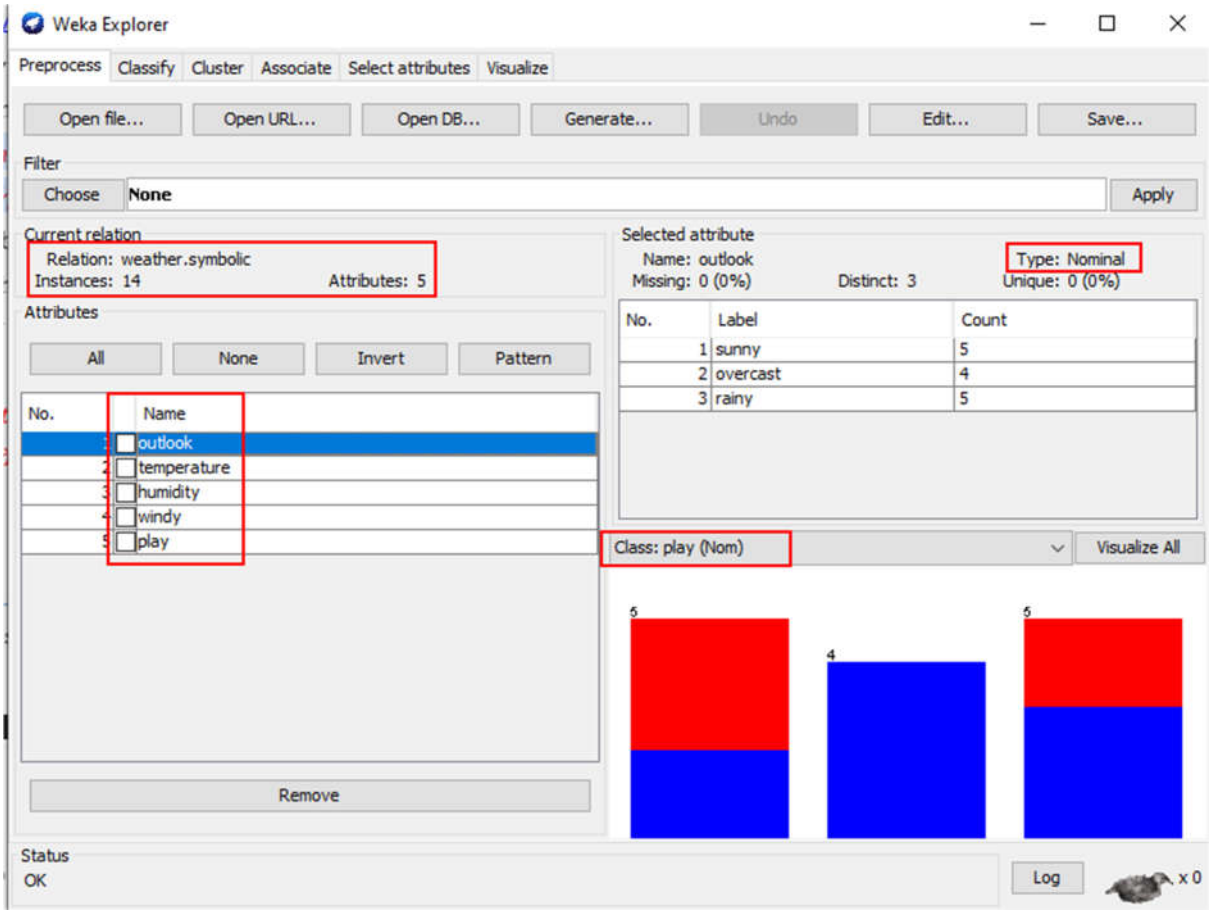


No.	rating Numeric	easy Nominal	ai Nominal	sys Nominal	thy Nominal	morning Nominal
1	2.0	y	y	n	y	n
2	2.0	y	y	n	y	n
3	2.0	n	y	n	n	n
4	2.0	n	n	n	y	n
5	2.0	n	y	y	n	y
6	1.0	y	y	n	n	n
7	1.0	y	y	n	y	n
8	1.0	n	y	n	y	n
9	0.0	n	n	n	n	y
10	0.0	y	n	n	y	y
11	0.0	n	y	n	y	n
12	0.0	y	y	y	y	y
13	-1.0	y	y	y	n	y
14	-1.0	n	n	y	y	n
15	-1.0	n	n	y	n	y
16	-1.0	y	n	y	n	y
17	-2.0	n	n	y	y	n
18	-2.0	n	y	y	n	y
19	-2.0	y	n	y	n	n
20	-2.0	y	n	y	n	y

Yêu cầu 2 – Khảo sát tập dữ liệu Weather (thuộc tính rời rạc)

6. Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Tên của các thuộc tính này là gì? Các thuộc tính này có loại gì? Thuộc tính nào là lớp?

Tập dữ liệu có 14 mẫu, 5 thuộc tính: outlook, temperature, humidity, windy và play. Các thuộc tính này thuộc loại nominal. Thuộc tính lớp là play

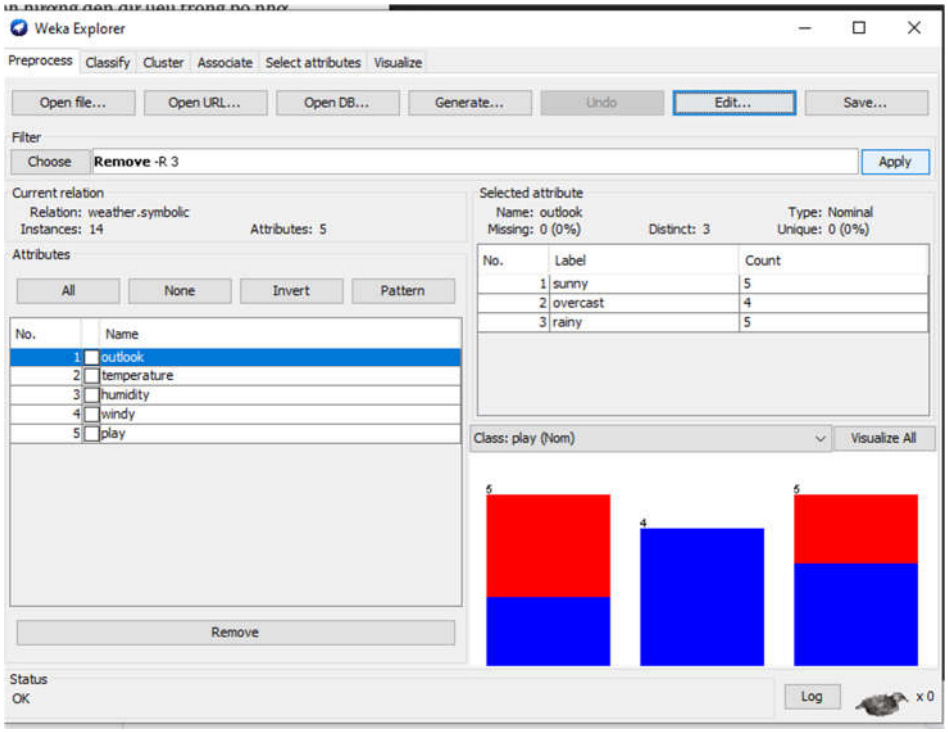


7. Chức năng của cột đầu tiên ở cửa sổ Viewer là gì?

Để đánh dấu số thứ tự cho các giá trị thuộc tính ở dưới

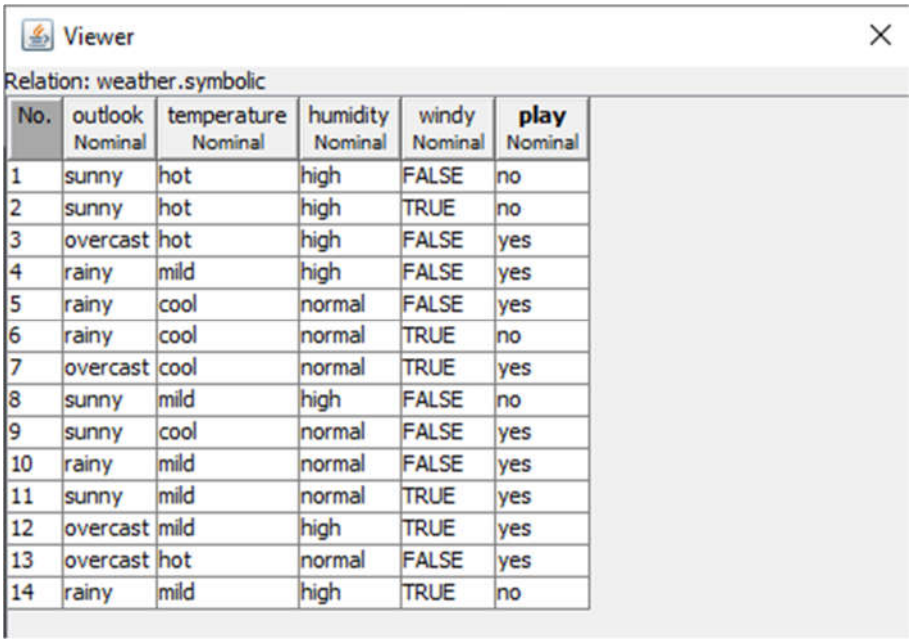
Lớp của mẫu thứ 8 trong tập dữ liệu là gì? No

Xóa thuộc tính 3



Sử dụng bộ lọc weka.unsupervised.instance.RemoveWithValues để loại bỏ mọi mẫu có giá trị thuộc tính humidity là high.

Trước khi áp dụng bộ lọc



No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

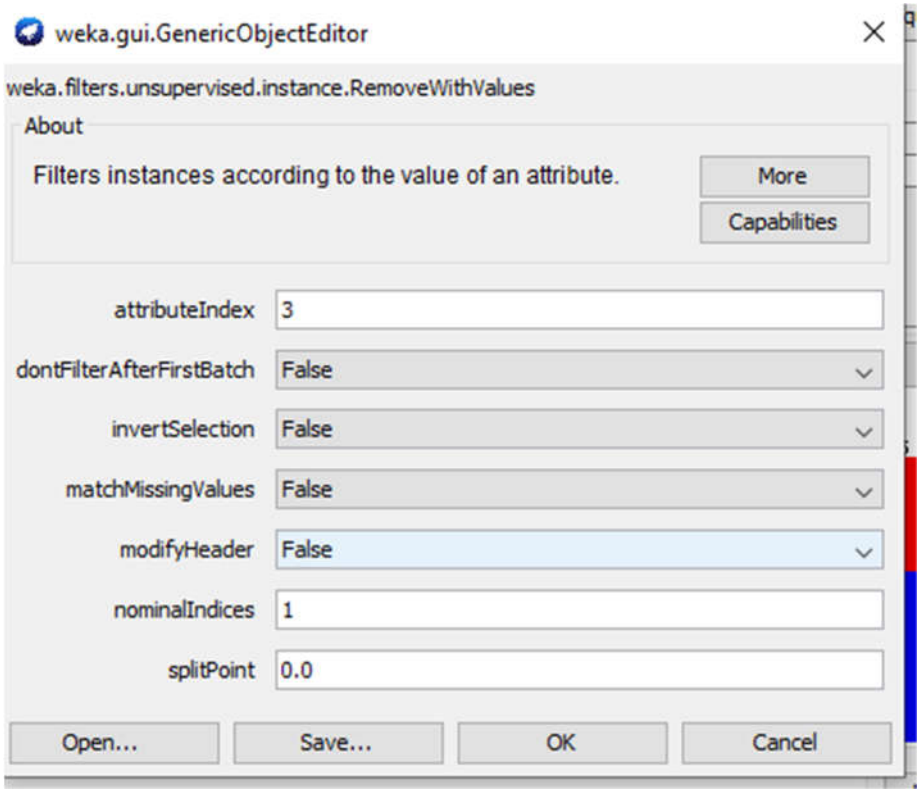
Sau khi áp dụng bộ lọc

Viewer

Relation: weather.symbolic-weka.filters.unsupervised.instance.RemoveWi

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	rainy	cool	normal	FALSE	yes
2	rainy	cool	normal	TRUE	no
3	overcast	cool	normal	TRUE	yes
4	sunny	cool	normal	FALSE	yes
5	rainy	mild	normal	FALSE	yes
6	sunny	mild	normal	TRUE	yes
7	overcast	hot	normal	FALSE	yes

Tham số áp dụng



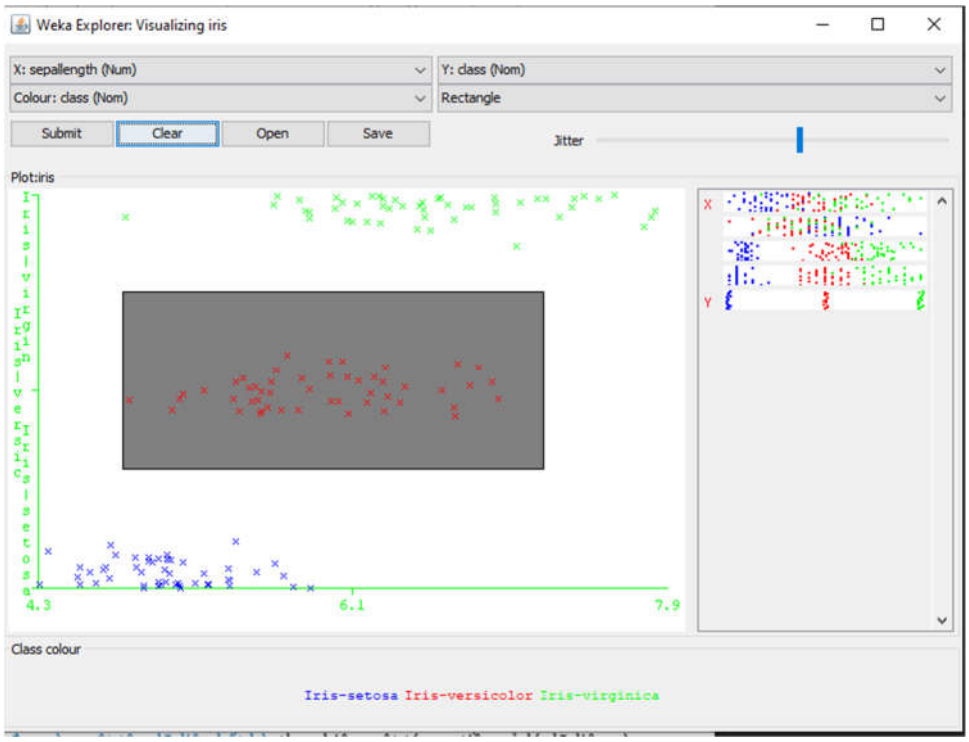
9. Mô tả tập dữ liệu. Nội dung của phần ghi chú (comment) nói về điều gì?

- **Tập dữ liệu có bao nhiêu mẫu?** 150 (50 thuộc tính mỗi 3 lớp)
- **Bao nhiêu thuộc tính?** 5 thuộc tính
- **Miền giá trị của thuộc tính petallength là gì?** 1.0 đến 6.9
- **Tập dữ liệu có bao nhiêu thuộc tính số và bao nhiêu thuộc tính rời rạc?** 4 thuộc tính số và 1 thuộc tính rời rạc

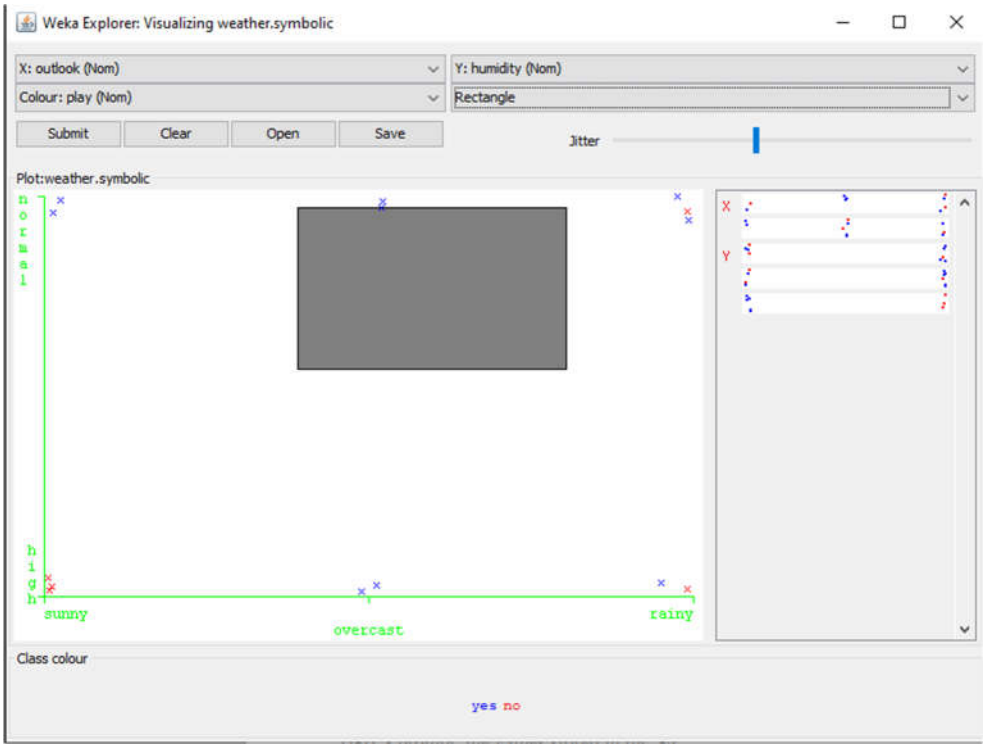
- **Tên của thuộc tính lớp là gì?** Iris Setosa, Iris Versicolour, Iris Virginica
- **Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?** Cân bằng (33% cho mỗi lớp)

Đổi trục x thành petalwidth và trục y thành petallengh**.**

Tùy ý minh họa

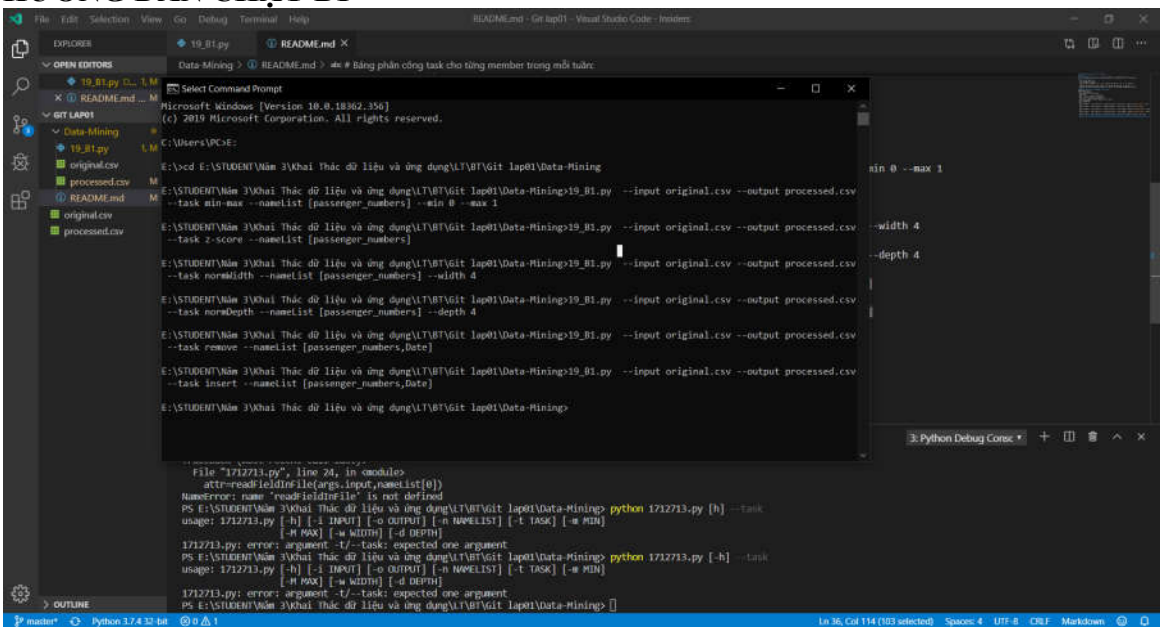


10. Thực hiện lại những câu hỏi trong Yêu cầu 3.

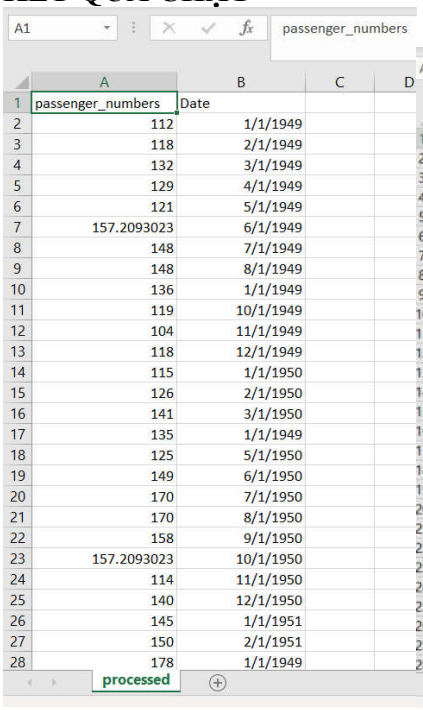


B – Nội dung thực hiện cài đặt (25 điểm)

HƯỚNG DẪN CHẠY B1

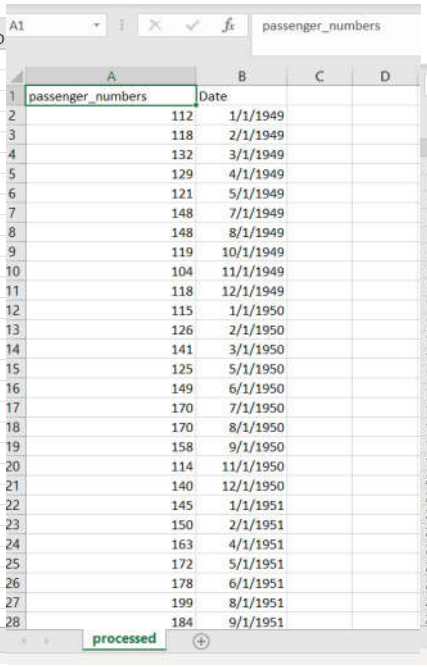


KẾT QUẢ CHẠY



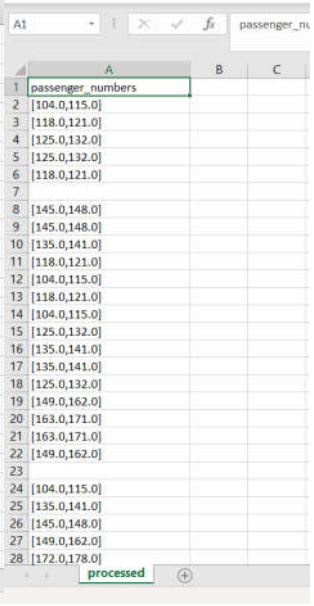
	A	B
1	passenger_numbers	Date
2	112	1/1/1949
3	118	2/1/1949
4	132	3/1/1949
5	129	4/1/1949
6	121	5/1/1949
7	157.2093023	6/1/1949
8	148	7/1/1949
9	148	8/1/1949
10	136	1/1/1949
11	119	10/1/1949
12	104	11/1/1949
13	118	12/1/1949
14	115	1/1/1950
15	126	2/1/1950
16	141	3/1/1950
17	135	1/1/1949
18	125	5/1/1950
19	149	6/1/1950
20	170	7/1/1950
21	170	8/1/1950
22	158	9/1/1950
23	157.2093023	10/1/1950
24	114	11/1/1950
25	140	12/1/1950
26	145	1/1/1951
27	150	2/1/1951
28	178	1/1/1949

f)



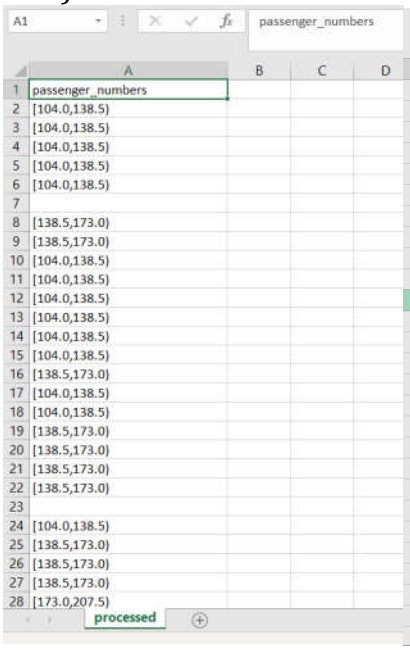
	A	B
1	passenger_numbers	Date
2	112	1/1/1949
3	118	2/1/1949
4	132	3/1/1949
5	129	4/1/1949
6	121	5/1/1949
7	148	7/1/1949
8	148	8/1/1949
9	119	10/1/1949
10	104	11/1/1949
11	118	12/1/1949
12	115	1/1/1950
13	126	2/1/1950
14	141	3/1/1950
15	125	5/1/1950
16	149	6/1/1950
17	170	7/1/1950
18	170	8/1/1950
19	158	9/1/1950
20	114	11/1/1950
21	140	12/1/1950
22	145	1/1/1951
23	150	2/1/1951
24	163	4/1/1951
25	172	5/1/1951
26	178	6/1/1951
27	199	8/1/1951
28	184	9/1/1951

e)



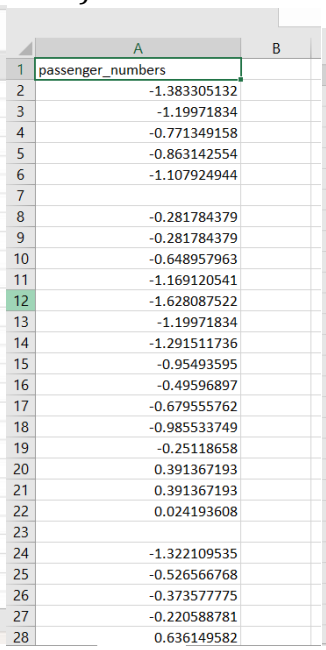
	A
1	passenger_numbers
2	[104.0,115.0]
3	[118.0,121.0]
4	[125.0,132.0]
5	[125.0,132.0]
6	[118.0,121.0]
7	
8	[145.0,148.0]
9	[145.0,148.0]
10	[135.0,141.0]
11	[118.0,121.0]
12	[104.0,115.0]
13	[118.0,121.0]
14	[104.0,115.0]
15	[125.0,132.0]
16	[135.0,141.0]
17	[135.0,141.0]
18	[125.0,132.0]
19	[149.0,162.0]
20	[163.0,171.0]
21	[163.0,171.0]
22	[149.0,162.0]
23	
24	[104.0,115.0]
25	[135.0,141.0]
26	[145.0,148.0]
27	[149.0,162.0]
28	[172.0,178.0]

d)



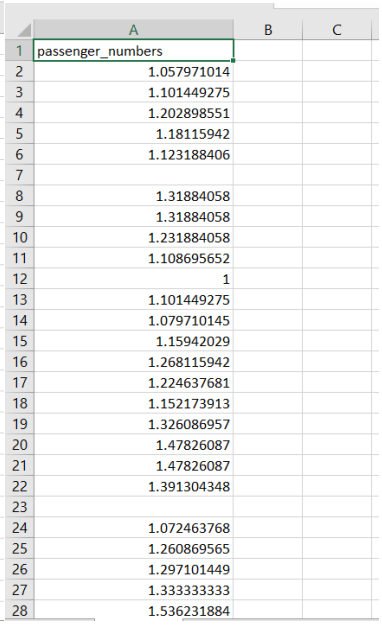
	A
1	passenger_numbers
2	[104.0,138.5)
3	[104.0,138.5)
4	[104.0,138.5)
5	[104.0,138.5)
6	[104.0,138.5)
7	
8	[138.5,173.0)
9	[138.5,173.0)
10	[104.0,138.5)
11	[104.0,138.5)
12	[104.0,138.5)
13	[104.0,138.5)
14	[104.0,138.5)
15	[104.0,138.5)
16	[138.5,173.0)
17	[104.0,138.5)
18	[104.0,138.5)
19	[138.5,173.0)
20	[138.5,173.0)
21	[138.5,173.0)
22	[138.5,173.0)
23	
24	[104.0,138.5)
25	[138.5,173.0)
26	[138.5,173.0)
27	[138.5,173.0)
28	[173.0,207.5)

c)



	A
1	passenger_numbers
2	-1.383305132
3	-1.19971834
4	-0.771349158
5	-0.863142554
6	-1.107924944
7	
8	-0.281784379
9	-0.281784379
10	-0.648957963
11	-1.169120541
12	-1.628087522
13	-1.19971834
14	-1.291511736
15	-0.95493595
16	-0.49596897
17	-0.679555762
18	-0.985533749
19	-0.25118658
20	0.391367193
21	0.391367193
22	0.024193608
23	
24	-1.322109535
25	-0.526566768
26	-0.373577775
27	-0.220588781
28	0.636149582

b)



	A
1	passenger_numbers
2	1.057971014
3	1.101449275
4	1.202898551
5	1.18115942
6	1.123188406
7	
8	1.31884058
9	1.31884058
10	1.231884058
11	1.108695652
12	1
13	1.101449275
14	1.079710145
15	1.15942029
16	1.268115942
17	1.224637681
18	1.152173913
19	1.326086957
20	1.47826087
21	1.47826087
22	1.391304348
23	
24	1.072463768
25	1.260869565
26	1.297101449
27	1.333333333
28	1.536231884

a)

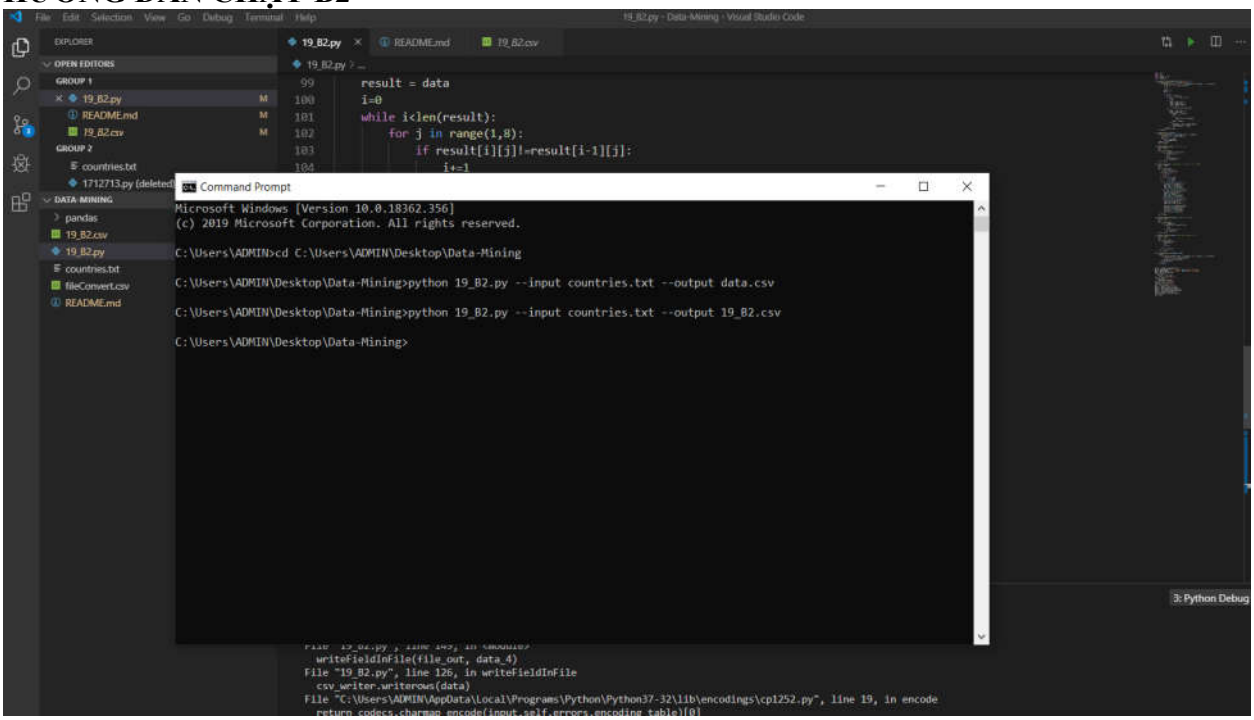
CÁC LỆNH COMMAND LINE

a) python 19_B1.py --input original.csv --output processed.csv --task min-max --nameList [passenger_numbers] --min 0 --max 1

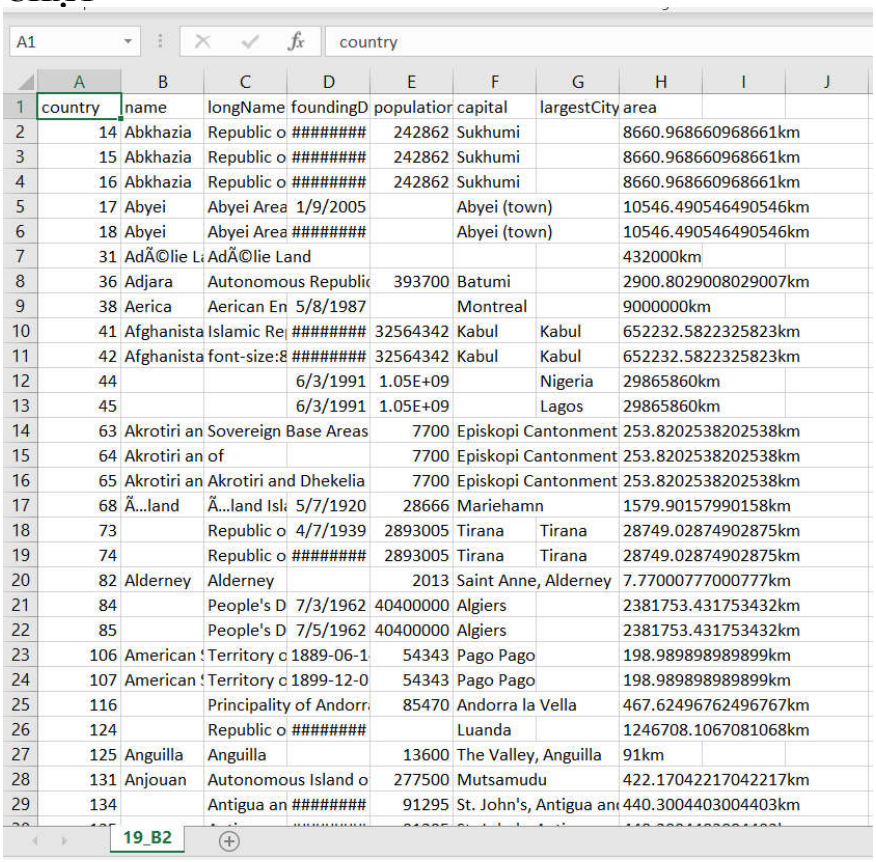
b) python 19_B1.py --input original.csv --output processed.csv --task z-score --nameList [passenger_numbers]

c) python 19_B1.py --input original.csv --output processed.csv --task normWidth --nameList [passenger_numbers] --width 4
d) python 19_B1.py --input original.csv --output processed.csv --task normDepth --nameList [passenger_numbers] --depth 4
e) python 19_B1.py --input original.csv --output processed.csv --task remove --nameList [passenger_numbers,Date]
f) python 19_B1.py --input original.csv --output processed.csv --task insert --nameList [passenger_numbers,Date]

HƯỚNG DẪN CHẠY B2



KẾT QUẢ CHẠY



A1 | country

	A	B	C	D	E	F	G	H	I	J
1	country	name	longName	foundingD	populatior	capital	largestCity	area		
2	14	Abkhazia	Republic o	########	242862	Sukhumi		8660.968660968661km		
3	15	Abkhazia	Republic o	########	242862	Sukhumi		8660.968660968661km		
4	16	Abkhazia	Republic o	########	242862	Sukhumi		8660.968660968661km		
5	17	Abyei	Abyei Area	1/9/2005		Abyei (town)		10546.490546490546km		
6	18	Abyei	Abyei Area	########		Abyei (town)		10546.490546490546km		
7	31	AdÃ©lie L	AdÃ©lie Land					432000km		
8	36	Adjara	Autonomous Republic		393700	Batumi		2900.8029008029007km		
9	38	Aerica	Aerican En	5/8/1987		Montreal		9000000km		
10	41	Afghanista	Islamic Re	########	32564342	Kabul	Kabul	652232.5822325823km		
11	42	Afghanista	font-size:8	########	32564342	Kabul	Kabul	652232.5822325823km		
12	44			6/3/1991	1.05E+09		Nigeria	29865860km		
13	45			6/3/1991	1.05E+09		Lagos	29865860km		
14	63	Akrotiri an	Sovereign Base Areas		7700	Episkopi Cantonment		253.8202538202538km		
15	64	Akrotiri an	of		7700	Episkopi Cantonment		253.8202538202538km		
16	65	Akrotiri an	Akrotiri and Dhekelia		7700	Episkopi Cantonment		253.8202538202538km		
17	68	Ã…land	Ã…land Isl	5/7/1920	28666	Mariehamn		1579.90157990158km		
18	73		Republic o	4/7/1939	2893005	Tirana	Tirana	28749.02874902875km		
19	74		Republic o	########	2893005	Tirana	Tirana	28749.02874902875km		
20	82	Alderney	Alderney		2013	Saint Anne, Alderney		7.77000777000777km		
21	84		People's D	7/3/1962	40400000	Algiers		2381753.431753432km		
22	85		People's D	7/5/1962	40400000	Algiers		2381753.431753432km		
23	106	American S	Territory o	1889-06-1	54343	Pago Pago		198.989898989899km		
24	107	American S	Territory o	1899-12-0	54343	Pago Pago		198.989898989899km		
25	116		Principality of Andorr		85470	Andorra la Vella		467.62496762496767km		
26	124		Republic o	########		Luanda		1246708.1067081068km		
27	125	Anguilla	Anguilla		13600	The Valley, Anguilla		91km		
28	131	Anjouan	Autonomous Island o		277500	Mutsamudu		422.17042217042217km		
29	134		Antigua an	########	91295	St. John's, Antigua an		440.3004403004403km		

19_B2

CÁC LỆNH COMMAND LINE

python 19_B2.py --input countries.txt --output 19_B2.csv