TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC
KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Lab 01 – Preprocessing Data

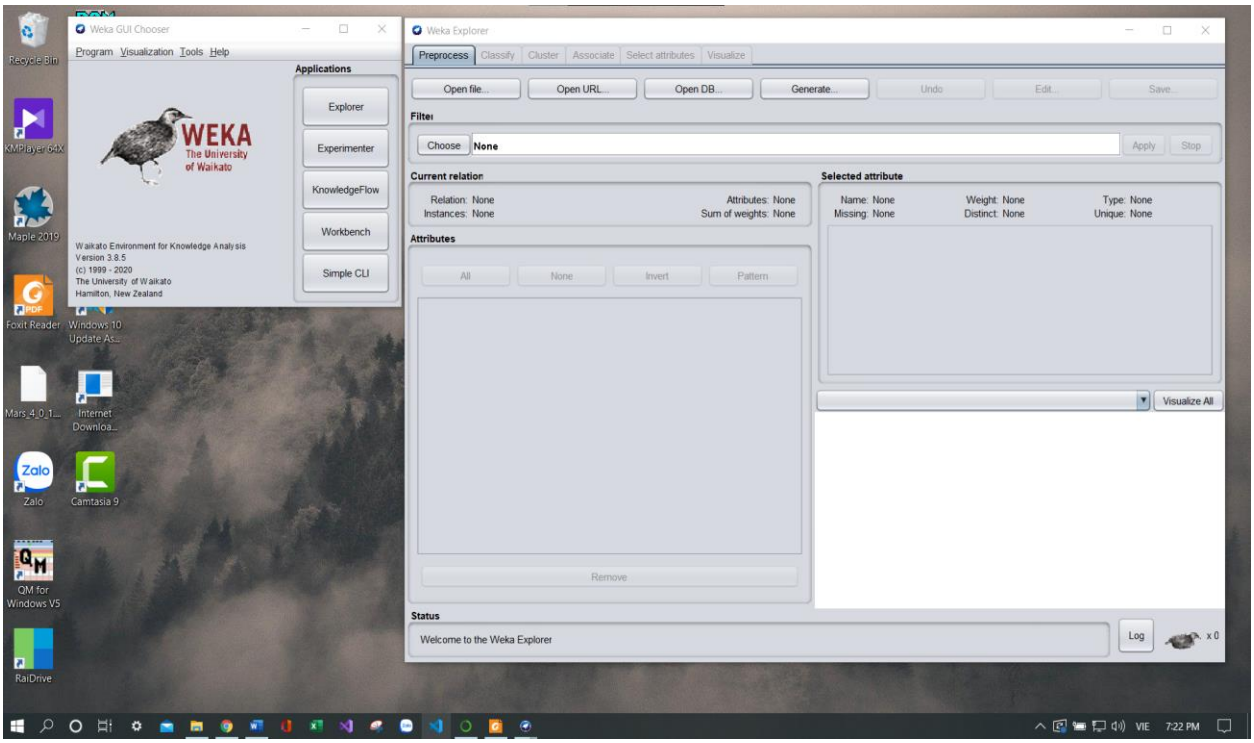
Thành viên:

Nguyễn Minh Lương – 19120571 - Phần trăm công việc: 60%

Dương Thanh Hiệp – 19120505 - Phần trăm công việc: 40%

Phần trăm hoàn thành: 100%

1 Yêu cầu 1: Cài đặt Weka (1 điểm)



- Ý nghĩa các nhóm điều khiển:
  + Current relation: Bảng tóm tắt các thông tin chi tiết về các tập dữ liệu đã tải lên
  + Attributes: liệt kê các thuộc tính của dữ liệu để lựa chọn
  + Selected attributes: Liệt kê thông tin về thuộc tính đã chọn, chẳng hạn như: tên thuộc tính, tỉ lệ dữ liệu bị thiếu, kiểu dữ liệu, các giá trị khác nhau trong tập dữ liệu
- Ý nghĩa các tap trong giao diện Explorer:
  + Preprocess: chọn và thay đổi (xử lí) dữ liệu làm việc
  + Classify: huấn luyện và kiểm tra các mô hình học máy (phân loại hoặc hồi quy/dự đoán)
  + Cluster: để học các nhóm từ dữ liệu (phân cụm)
  + Associate: khám phá các luật kết hợp từ dữ liệu
  + Selected attributes: để xác định và lựa chọn các thuộc tính liên quan(quan trọng) nhất của dữ liệu
  + Visualize: xem, hiển thị biểu đồ tương tác hai chiều đối với dữ liệu

2 Yêu cầu 2: Làm quen với Weka (6 điểm)

2.1 Đọc dữ liệu vào Weka (2 điểm)

1.Tập dữ liệu có bao nhiêu mẫu (instances)?

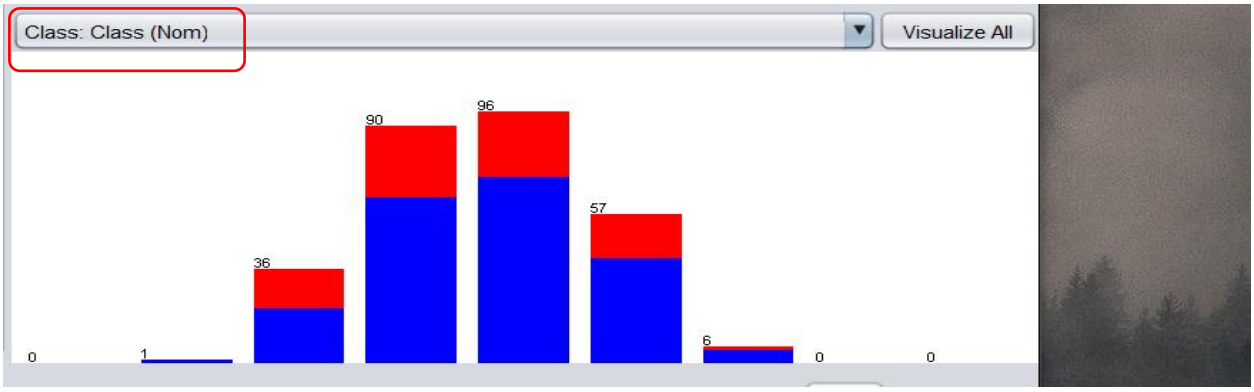
- Tập dữ liệu có 286 mẫu

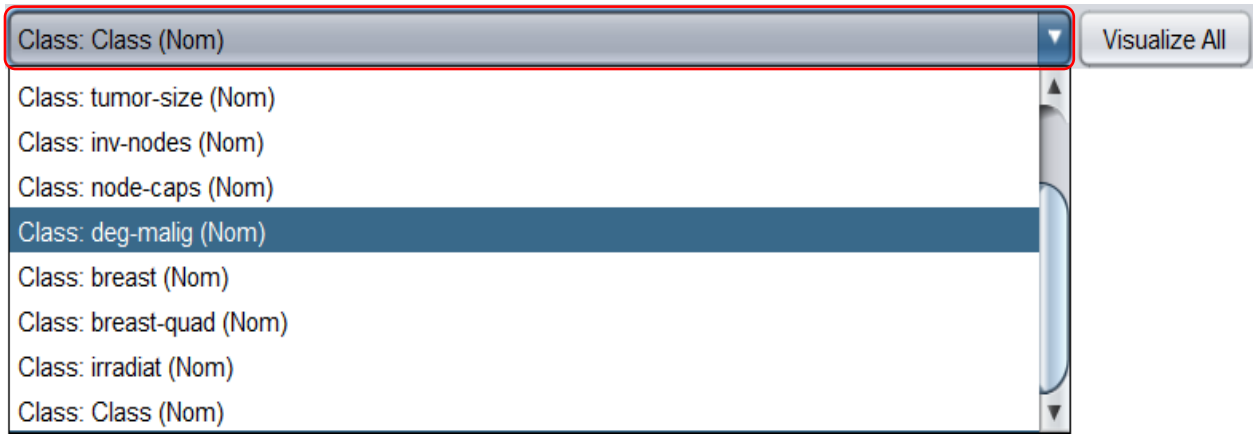
2.Tập dữ liệu có bao nhiêu thuộc tính (attributes)?

- Tập dữ liệu có 10 thuộc tính (9 + thuộc tính lớp)

3.Thuộc tính nào được dùng làm lớp (class)? Có thể thay đổi thuộc tính dùng làm lớp hay không? Nếu có thì bằng cách nào?

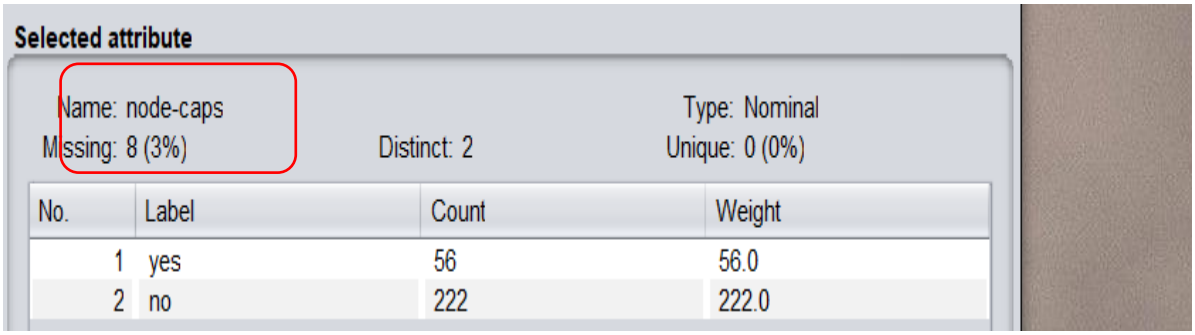
- Thuộc tính Class được sử dụng làm lớp. Có thể thay đổi được thuộc tính lớp bằng cách chọn tại dropdown .

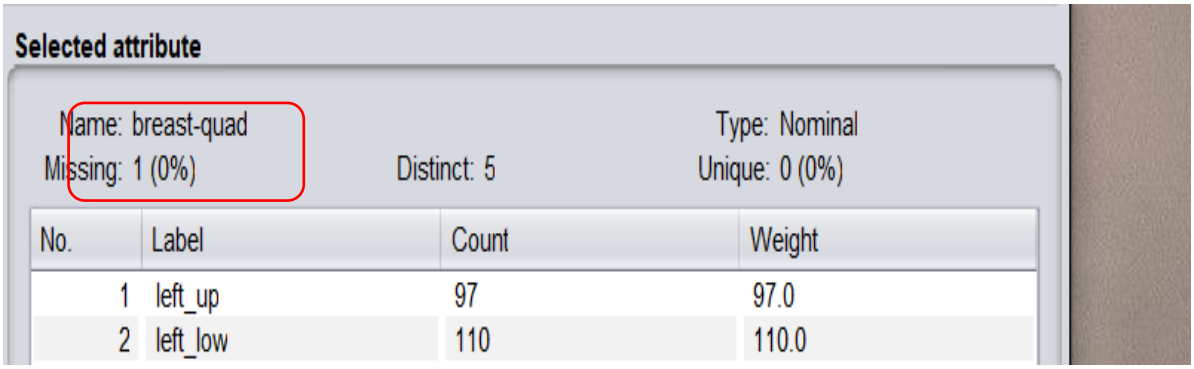




4. Tìm hiểu chi tiết từng thuộc tính trong khung Attributes và cho biết: có bao nhiêu thuộc tính bị thiếu dữ liệu (missing values)? Thuộc tính nào thiếu dữ liệu ít nhất/nhiều nhất? Trình bày tổng quát các cách để giải quyết vấn đề missing values.

- Có 2 thuộc tính bị thiếu dữ liệu node-caps và breast-quad.



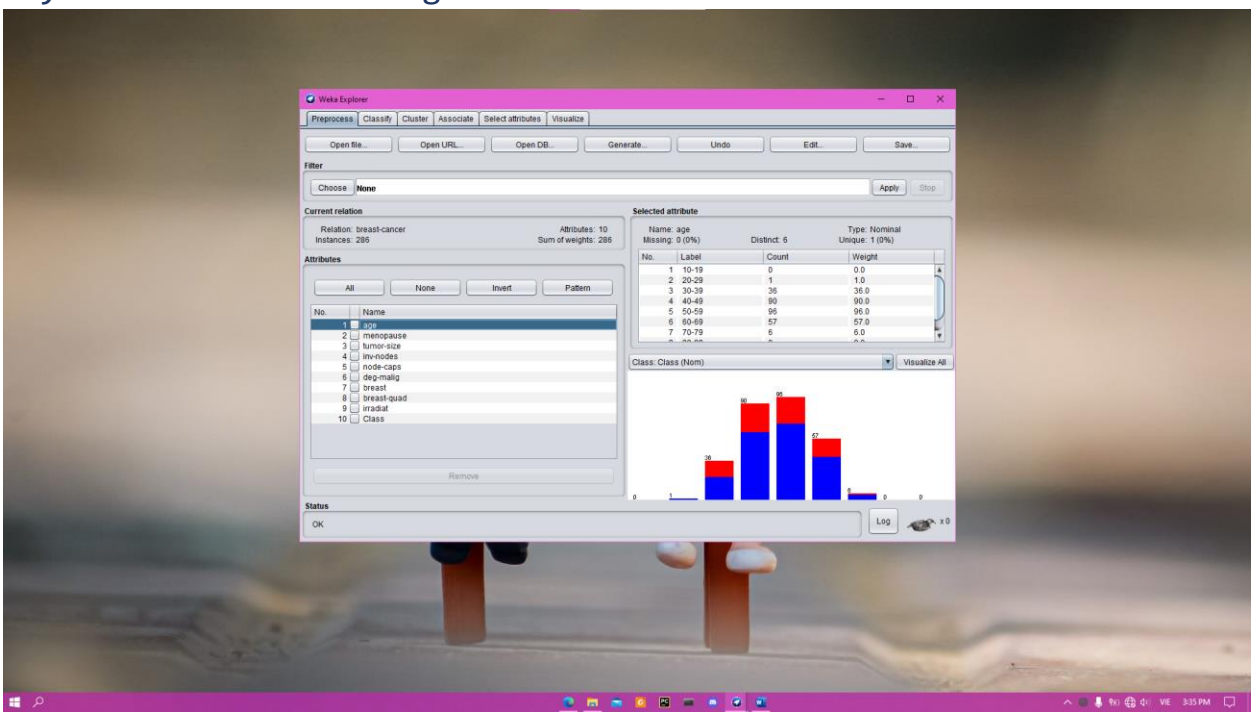


- Thuộc tính node-caps bị thiếu dữ liệu nhiều nhất 8%
- Thuộc tính breast-quad bị thiếu dữ liệu ít nhất 1%
- Các cách xử lí khi bị thiếu dữ liệu:

+Xóa mẫu có dữ liệu bị thiếu: đối với dữ liệu không bị thiếu nhiều.

+Thay dữ liệu bị thiếu bằng các giá trị khác: mean, mode của thuộc tính, giá trị có thể xãy ra nhất (suy luận dựa vào cây quyết định, công thức Bayes,...)

5. Giải thích ý nghĩa của đồ thị trong cửa sổ Explorer. Bạn đặt tên cho đồ thị này là gì? Màu xanh và màu đỏ có nghĩa gì? Đồ thị này biểu diễn cho cái gì?



- Đặt tên là biểu đồ thanh.
- Màu đỏ thể hiện cho số mẫu sẽ bị tái phát, màu xanh thể hiện cho số mẫu không bị tái phát.
- Thể hiện sự phân khúc của lớp đối với mỗi khoảng giá trị trên 1 thuộc tính nhất định.

2.2 Khám phá tập dữ liệu Weather (2 điểm)

1.Tập dữ liệu có bao nhiêu thuộc tính? Bao nhiêu mẫu? Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric). Thuộc tính nào là lớp?

- Tập dữ liệu có 14 mẫu

- Tập dữ liệu có 5 thuộc tính

- Thuộc tính lớp: play

- Phân loại thuộc tính theo kiểu dữ liệu:

+ Numeric: temperature, humidity

+ Categorical: outlook, windy, play

2. Liệt kê five-number summary của thuộc tính temperature và humidity. Weka có cung cấp những giá trị này không?

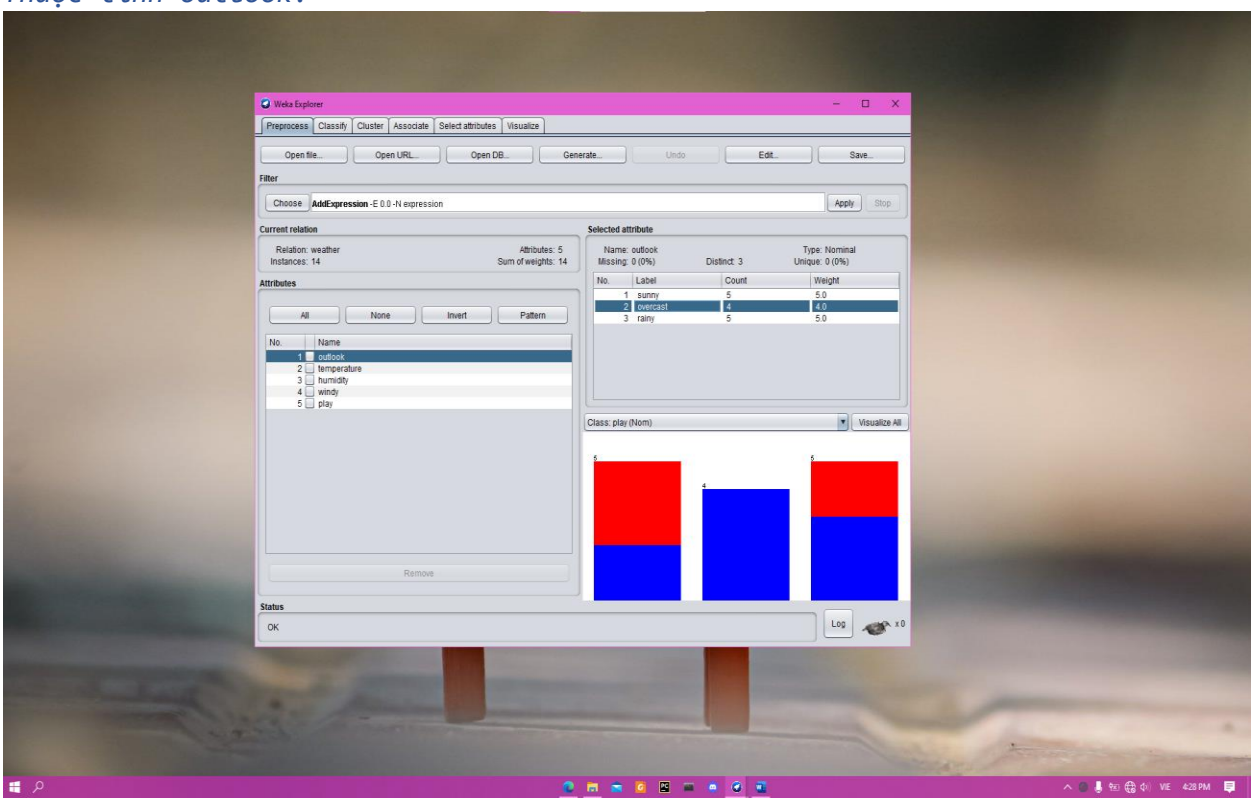
- **temperature:** min=64, max=85, Q1=?, Q2=?, Q3=?

- **humidity:** min=65, max=96, Q1=?, Q2=?, Q3=?

- Weka không cung cấp đủ các giá trị này chỉ cung cấp **min, max, mean, std**.

3. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.

Thuộc tính outlook:



Nhận xét biểu đồ:

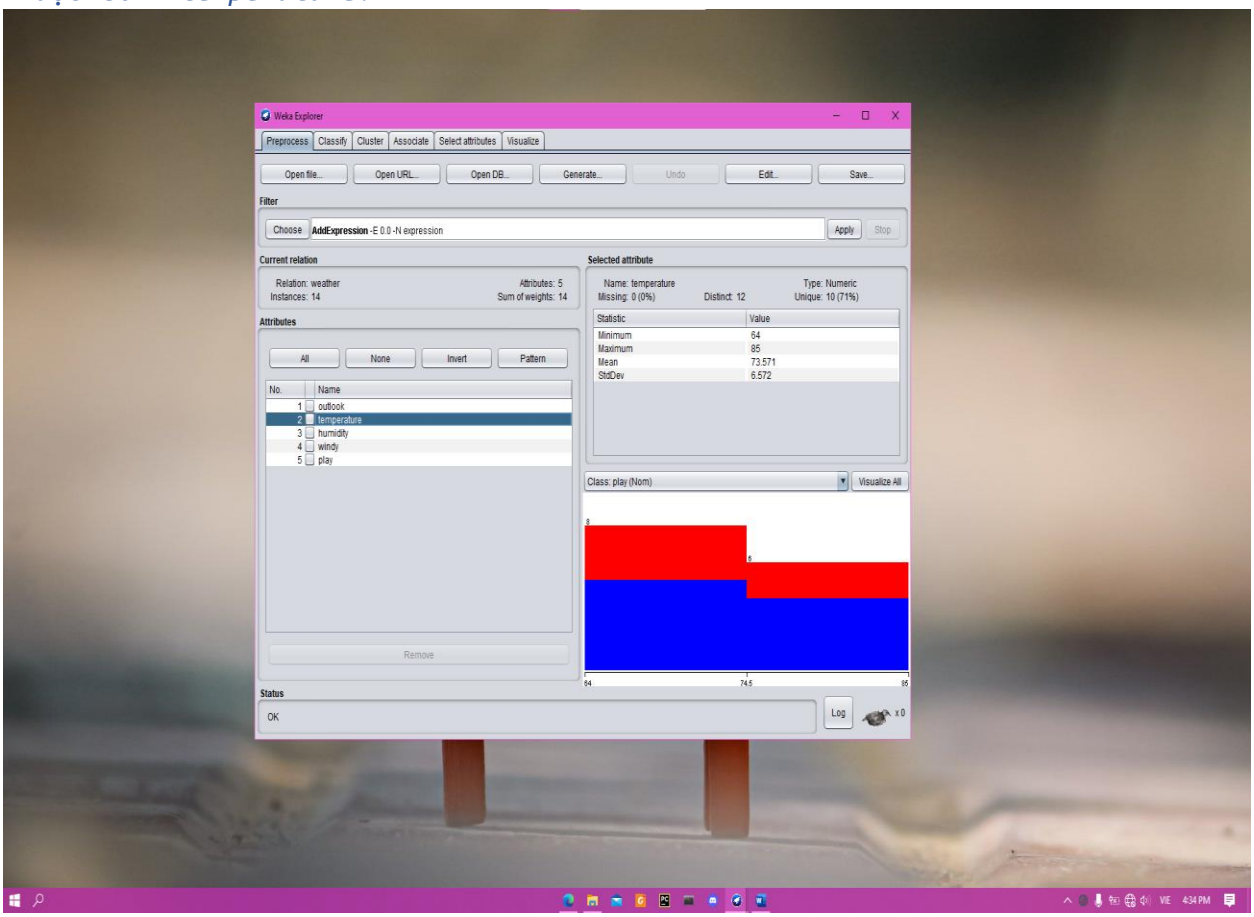
-Khi quang cảnh là **có mây** có 4 quyết định có tỷ lệ sẽ ra ngoài chơi (play) là 100% ***(cả 4 quyết định là có)***.

-Khi quang cảnh là **có mưa** có 5 quyết định có tỷ lệ sẽ ra ngoài chơi (play) là lớn hơn 50% ***(3 quyết định là có)***.

-Khi quang cảnh là **có nắng** có 5 quyết định có tỷ lệ sẽ ra ngoài chơi (play) là bé hơn 50% ***(2 quyết định là có)***.

⇒ *khi quang cảnh là có mây thì mọi người có xu hướng ra ngoài chơi nhiều.*

Thuộc tính temperature:



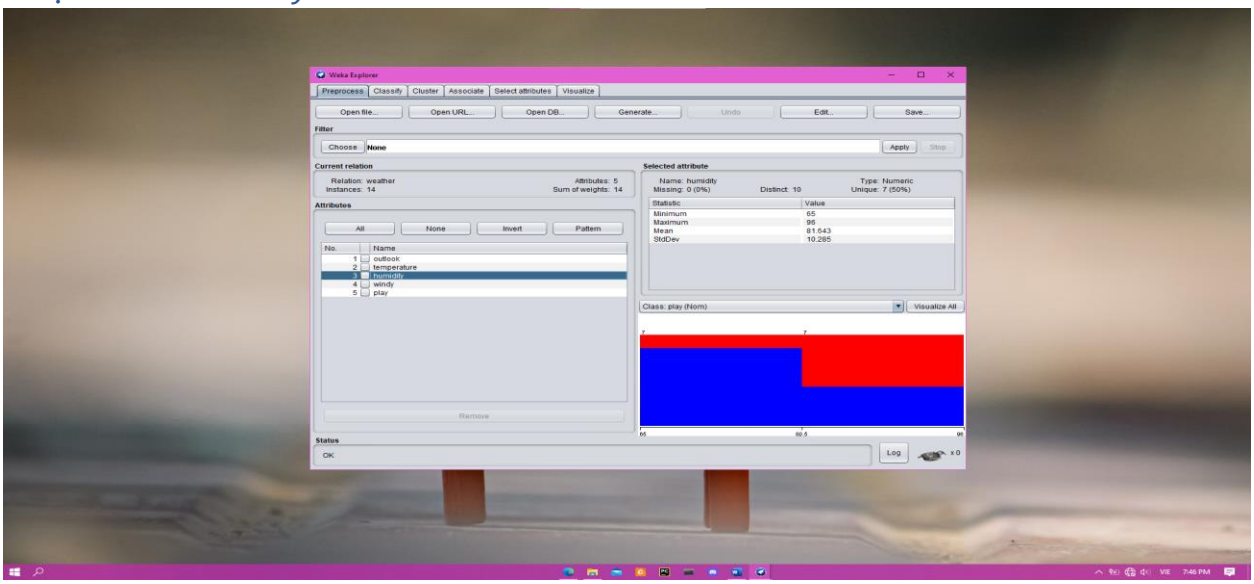
Nhận xét biểu đồ:

-Khi nhiệt độ ở trong ngưỡng **[64, 74.5]** có 8 quyết định với tỷ lệ xảy ra quyết định có đi chơi là hơn 50% (5 quyết định là có).

- Khi nhiệt độ ở trong ngưỡng **(74.5, 85]** có 6 quyết định với tỷ lệ xảy ra quyết định có đi chơi là hơn 50% (4 quyết định là có).

⇒ *Ở đây chúng ta vẫn có thể kết luận là khi nhiệt độ ở ngưỡng* $(74.5; 85]$ *thì tỷ lệ xảy ra quyết định có đi chơi lớn hơn so với ở ngưỡng* $[64; 74.5]$.

Thuộc tính humidity:



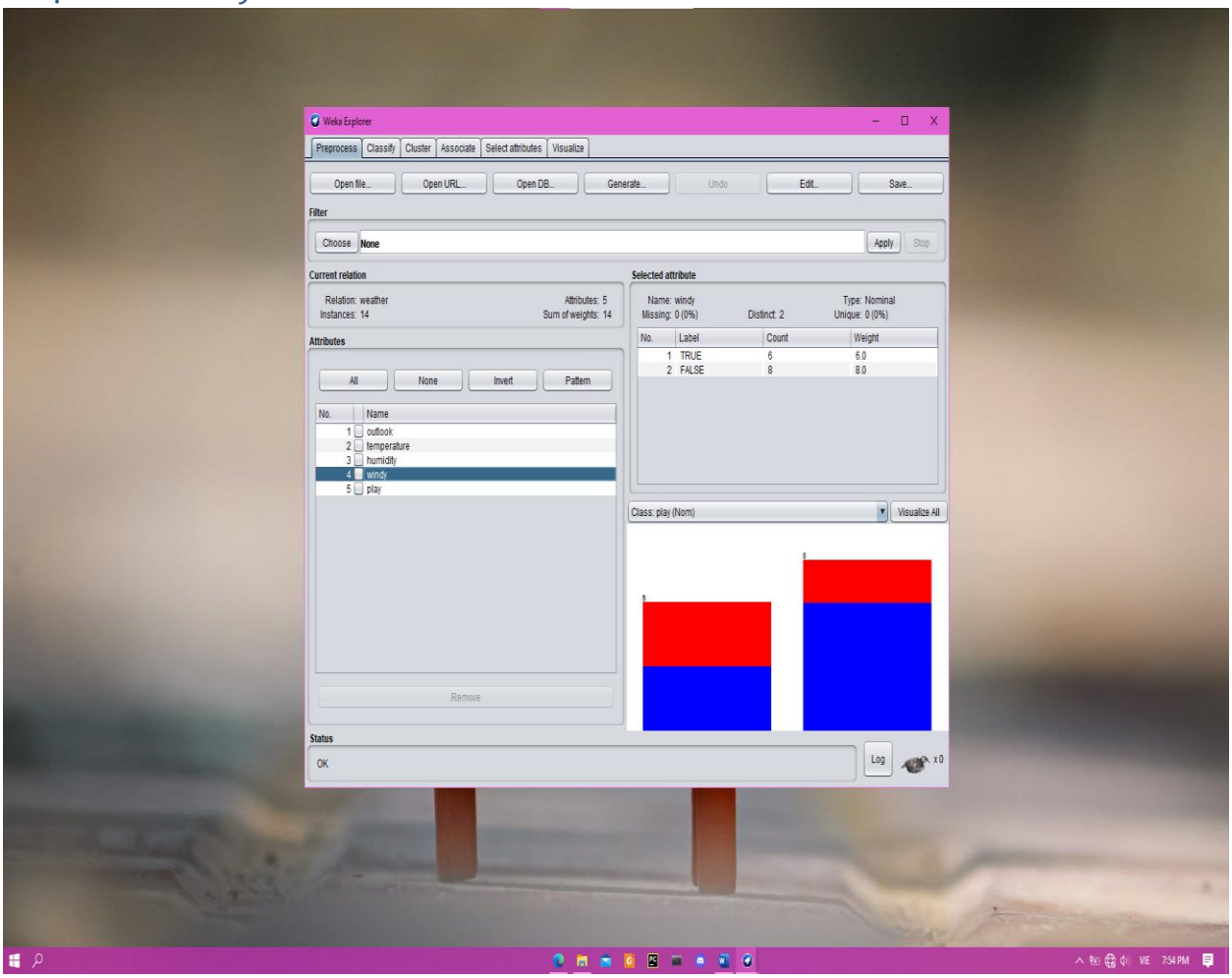
Nhận xét biểu đồ:

- Khi độ ẩm ở trong ngưỡng **[65, 80.5]** có 7 quyết định với tỷ lệ xảy ra quyết định có đi chơi là hơn 50% (6 quyết định là có).

- Khi độ ẩm ở trong ngưỡng **(80.5, 96]** có 7 quyết định với tỷ lệ xảy ra quyết định có đi chơi là ít hơn 50% (3 quyết định là có).

⇒ *vậy khi độ ẩm ở trong ngưỡng* $[65, 80.5]$ *mọi người có xu hướng ra quyết định có đi chơi nhiều.*

Thuộc tính windy:

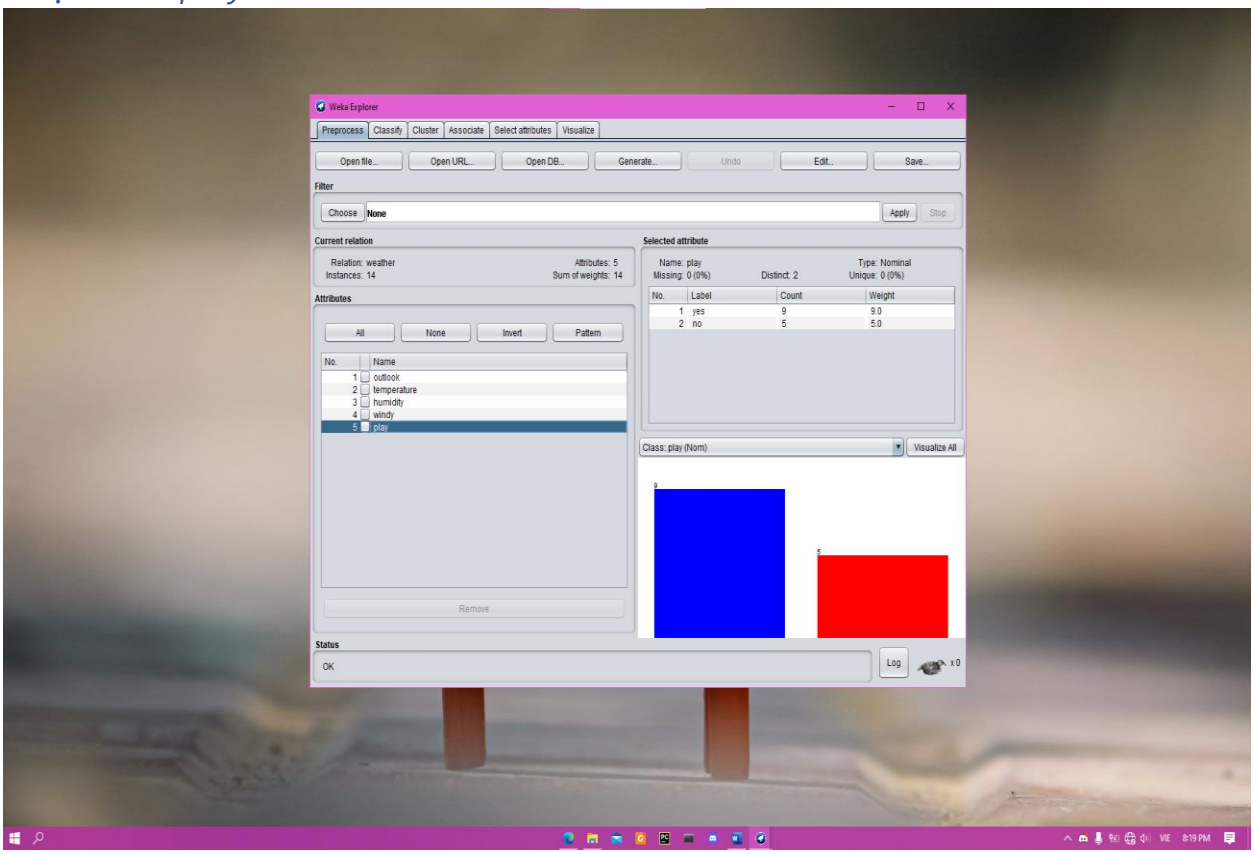


-Khi trời ***có gió*** thì có 6 quyết định với tỷ lệ xảy ra quyết định có đi chơi là 50% ***(3 quyết định là có)***.

-Khi trời ***không có gió*** thì có 6 quyết định với tỷ lệ xảy ra quyết định có đi chời là hơn 50% ***(6 quyết định là có)***.

⇒ *vậy khi trời không có gió thì mọi người có xu hướng ra quyết định có đi chơi nhiều.*

Thuộc tính play:



Nhận xét biểu đồ:

-Có 9 mẫu có quyết định là có đi chơi và 5 mẫu là có quyết định không đi chơi.

4. Chuyển sang tab Visualize. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị ở đây là gì? Chọn jitter tối đa để thấy tổng quan hơn về phân bố dữ liệu. Theo bạn có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

-Scatter Plot.

-(temperature, outlook); (humidity, temperature); (humidity, play).

2.3 Khám phá tập dữ liệu Tín dụng Đức (2 điểm)

1. Nội dung của phần ghi chú (comment) trong credit-g.arff (khi mở bằng 1 text editor bất kì) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Mô tả 5 thuộc tính bất kì (phải vừa có cả thuộc tính rời rạc và thuộc tính liên tục).

- Nội dung phần ghi chú của file credit-g.arff nói về:
  + Tiêu đề của tập dữ liệu:
  + Nguồn của tập dữ liệu.

+ Mô tả chi tiết các thuộc tính của quan hệ.
+ Tên của mối quan hệ.
+ Các quy ước gán nhãn cho giá trị của các thuộc tính.

- Tập dữ liệu có 1000 mẫu
- Tập dữ liệu có 21 thuộc tính
- Mô tả 5 thuộc tính:
  + Thuộc tính liên tục:
    - Age: tuổi của các user
    - Existing_credits : số lượng tín dụng đang có tại ngân hàng
  + Thuộc tính rời rạc:
    - Housing: tình trạng sở nhà của user. Tập giá trị: { rent, own, 'for free'}

      Quy ước gắn nhãn:
      A151 : rent
      A152 : own
      A153 : for free
    - Own_telephone: có sở hữu điện thoại hay không. Tập giá trị: { none, yes}
    - Job: tình trạng công việc hiện tại. Tập giá trị: { 'unemp/unskilled non res', 'unskilled resident', skilled, 'high qualif/self emp/mgmt'}

      Quy ước gắn nhãn:
      A171 : unemployed/ unskilled - non-resident
      A172 : unskilled - resident
      A173 : skilled employee / official
      A174 : management/ self-employed/highly qualified employee/officer

2. Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?

-Tên của thuộc tính lớp là **class** (gồm 2 giá trị good và bad).

-Phân bố của các lớp bị lệch về bên lớp good nhiều hơn với số mẫu nằm trong lớp good là 700 còn lại là 300.

3. Sử dụng tab Select attributes. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.

- *Bộ đánh giá thuộc tính (Attribute Evaluator):* Để đánh giá tập các thuộc tính của tập dữ liệu.

+ CfsSubsetEval: Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính riêng lẻ và mức độ dư thừa giữa chúng

+ ClassifierAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại do người dùng chỉ định.

+ ClassifierSubsetEval: Đánh giá các tập hợp con thuộc tính trên dữ liệu đào tạo hoặc một tập hợp thử nghiệm tạm dừng riêng biệt.

+ CorrelationAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan (Pearson) giữa nó và lớp.

+ GainRatioAttributeEval: Đánh giá một thuộc tính dựa trên tỷ lệ gia tăng.

+ InfoGainAttributeEval: : Đánh giá một thuộc tính dựa trên thông tin thu được.

+ OneRAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại OneR.

+ PrincipalComponents: Thực hiện phân tích và chuyển đổi các thành phần chính của dữ liệu.

+ ReliefFAttributeEval: Đánh giá thuộc tính dựa trên các thể hiện.

+ SymmetricalUncertAttributeEval: Đánh giá giá trị của một thuộc tính bằng cách đo độ không đối xứng.

+ WrapperSubsetEval: Đánh giá các tập thuộc tính bằng cách sử dụng một lược đồ học tập.

- ***Phương thức tìm kiếm (Search Method):*** Để xác định phương pháp tìm kiếm được thực hiện. WEKA cung cấp 3 phương thức tìm kiếm, gồm:

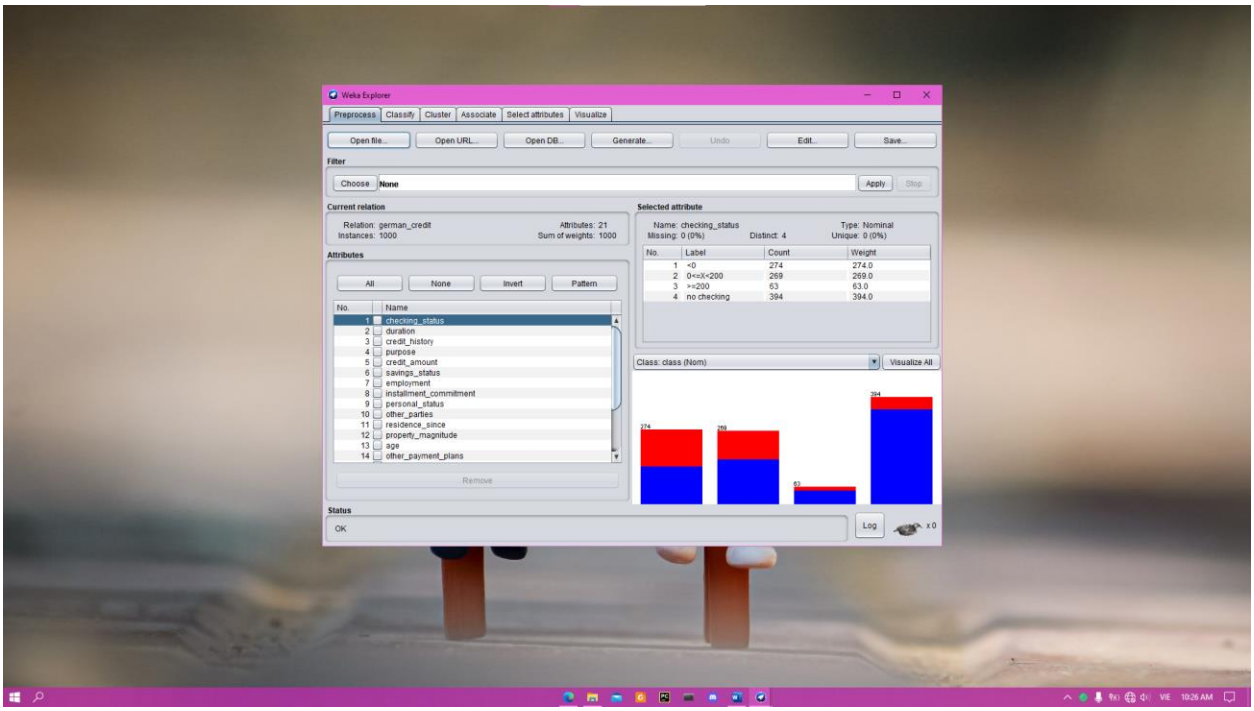
+ BestFirst: Tiến hành kỹ thuật leo đồi tham lam kết hợp với quay lui.

+ GreedyStepwise: Thực hiện tìm kiếm tham lam về phía trước hoặc phía sau thông qua không gian các tập con thuộc tính.

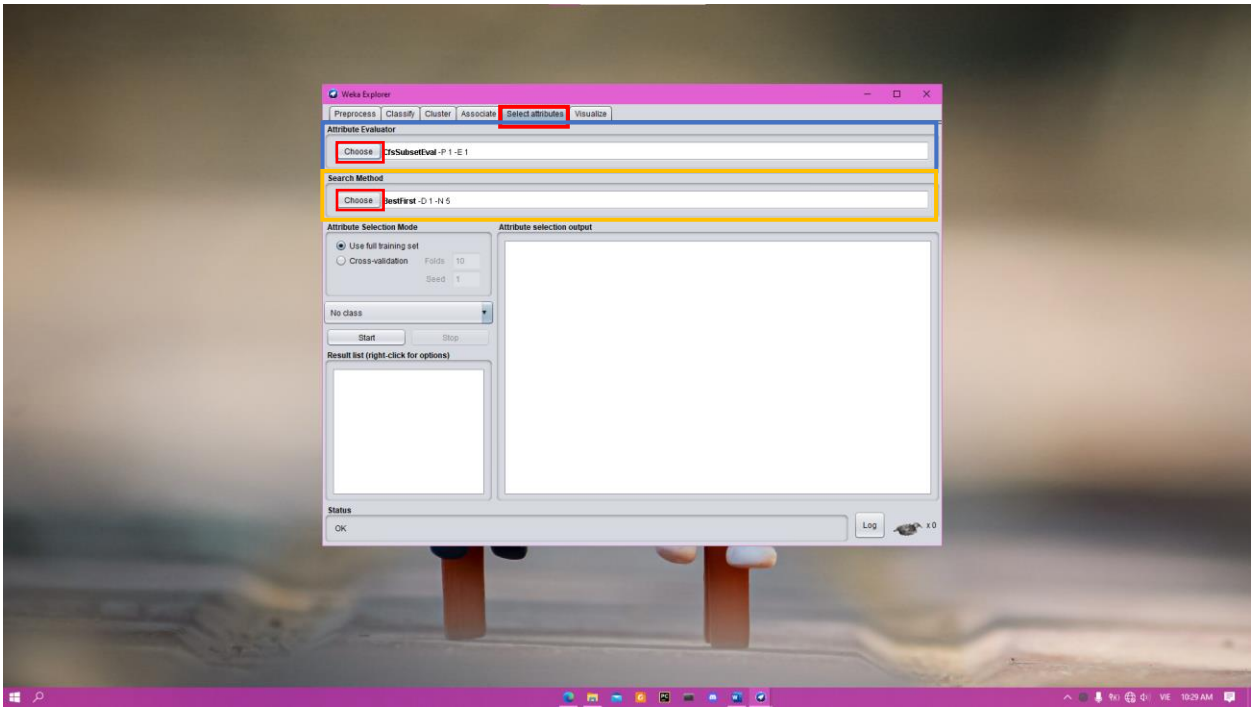
+ Ranker: Xếp hạng các thuộc tính theo đánh giá trọng số của từng thuộc tính. Sử dụng kết hợp với các bộ đánh giá thuộc tính (ReliefF, GainRatio,…).

4. Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng.

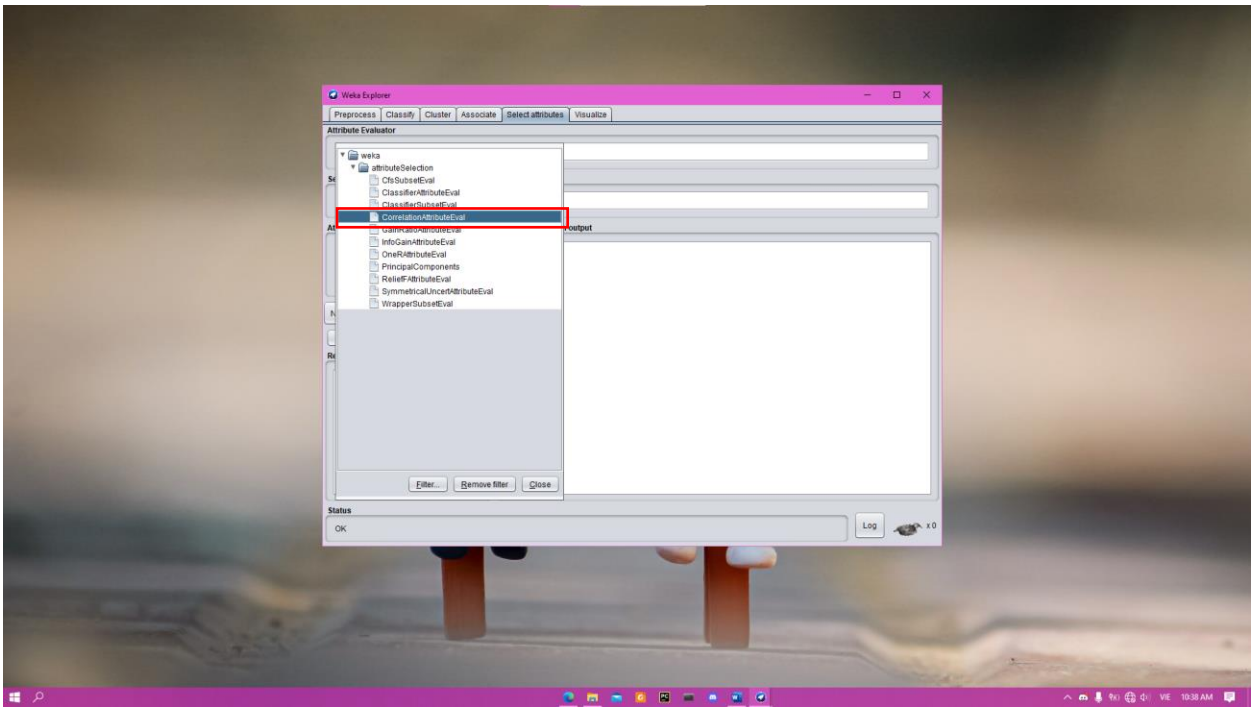
Mở tập dữ liệu “credit-g.arff”



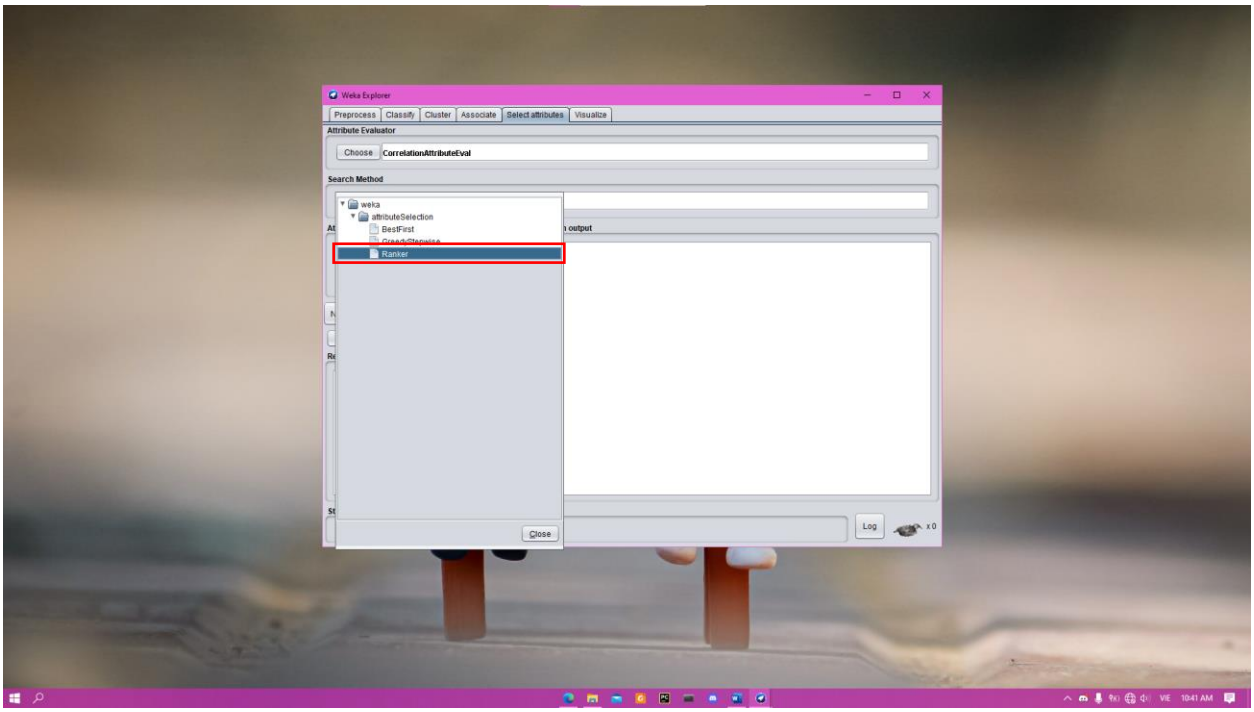
Chọn tab “Select attributes”



Chọn vào **<u>khung “Attribute Evalutor”</u>** chọn vào nút ***choose,*** Chọn **<u>filter “CorrelationAttributeEval”</u>**:

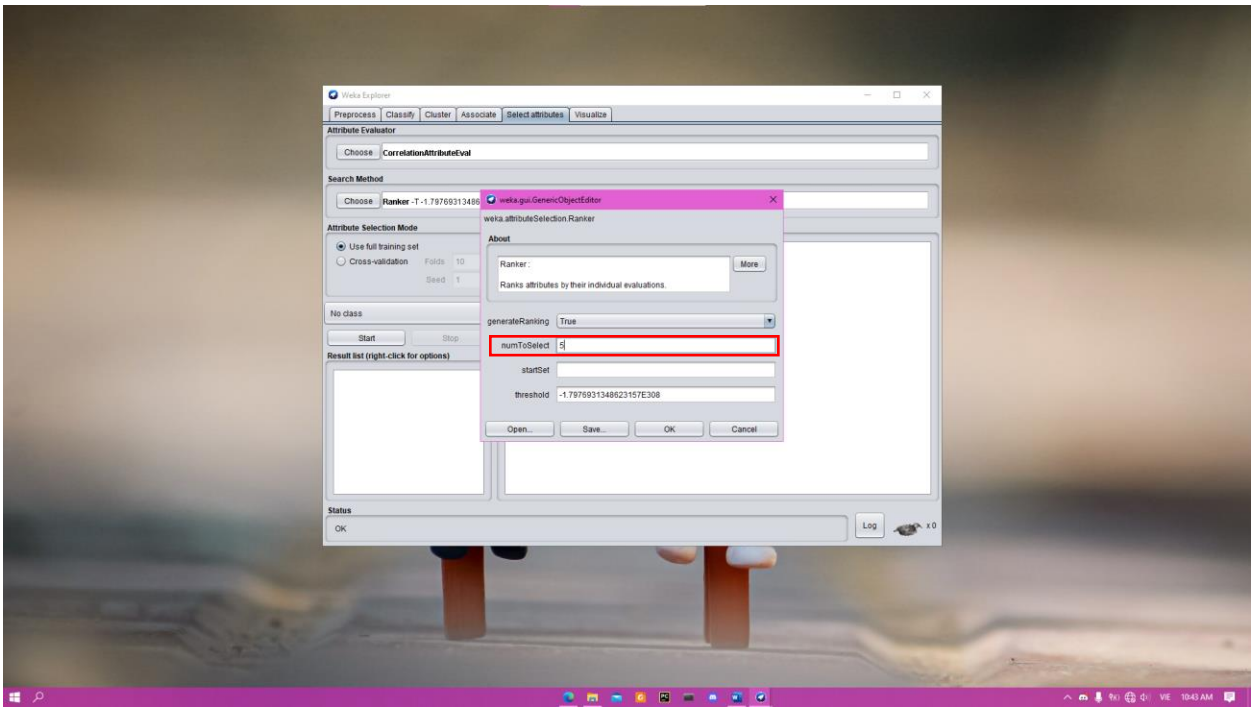


Chọn vào **<u>khung "Search Method"</u>** chọn vào nút ***choose***, chọn phương pháp **<u>search "Ranker"</u>**:



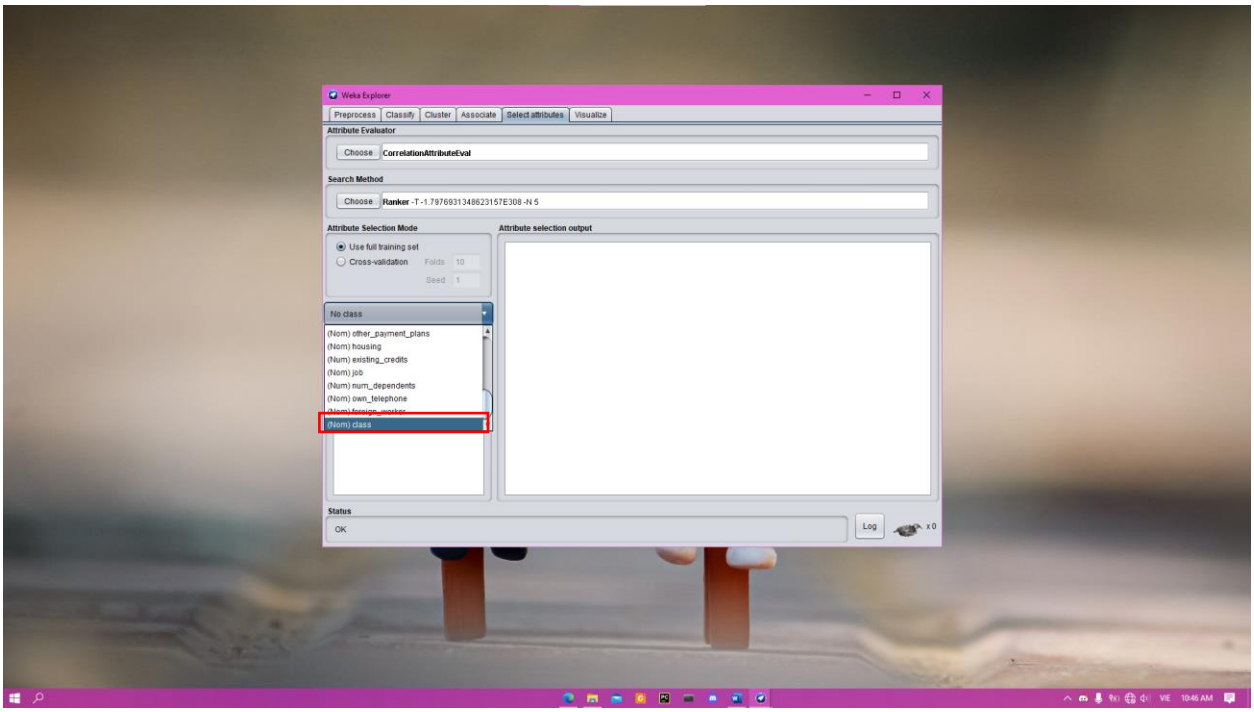
Tùy chỉnh cài đặt của phương pháp Search Ranker:

- Chỉnh phần **<u>"numToSelect"</u>** thành **<u>5</u>** ***(để chọn ra 5 thuộc tính)***.



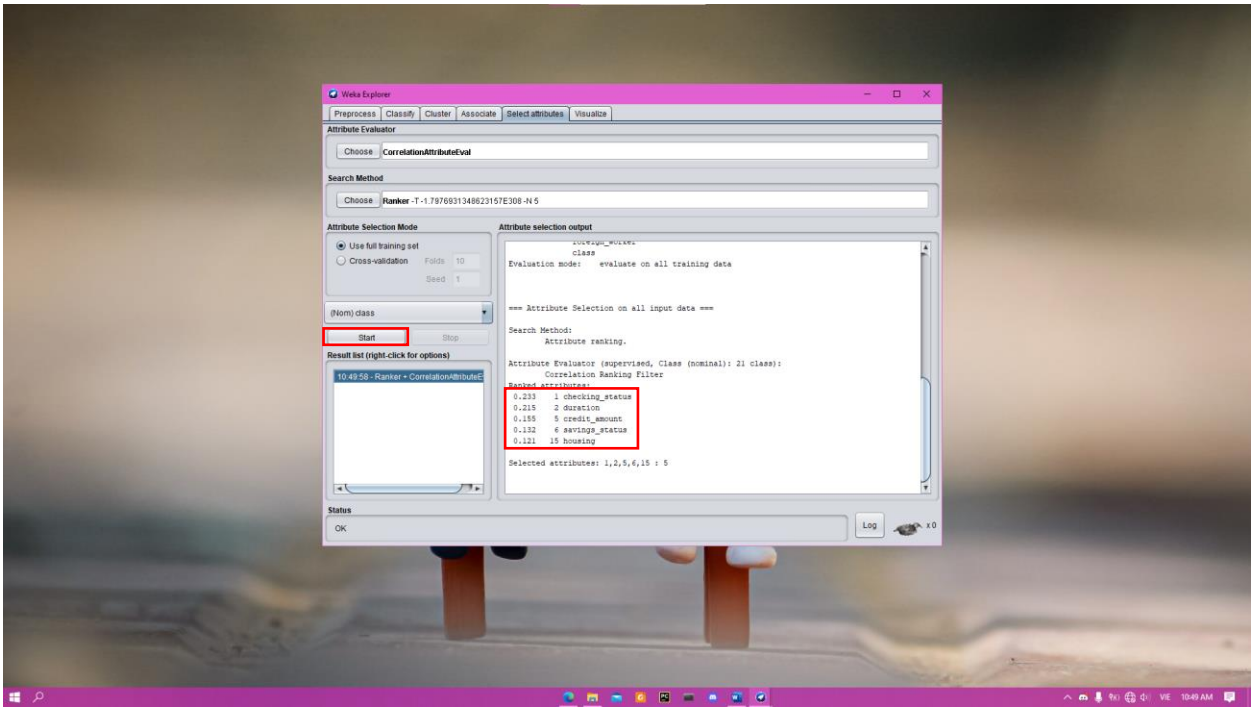
Cuối cùng chọn thuộc tính để tìm sự tương quan giữa thuộc tính đó với các thuộc tính khác:

- Chọn thuộc tính phân lớp “(Nom) class”.



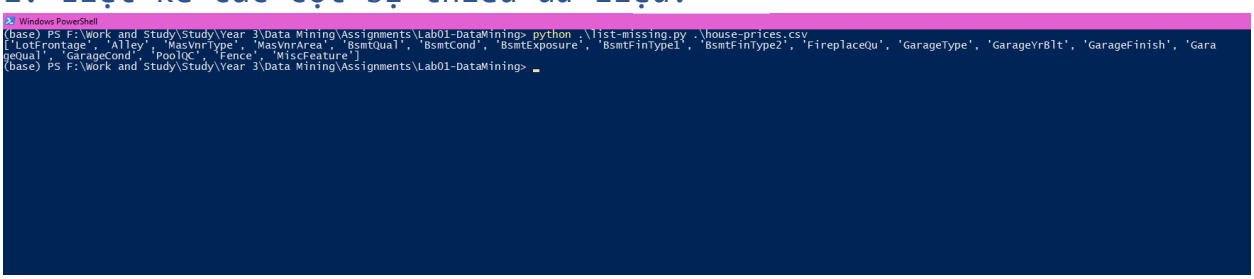
Chọn vào nút start để áp dụng filter vào tập dữ liệu:

- Kết quả là các ***dòng có chứa hệ số tương quan và tên của thuộc tính tương ứng*** có hệ số tương quan cao nhất trong các thuộc tính so với thuộc tính phân lớp **"class"**.



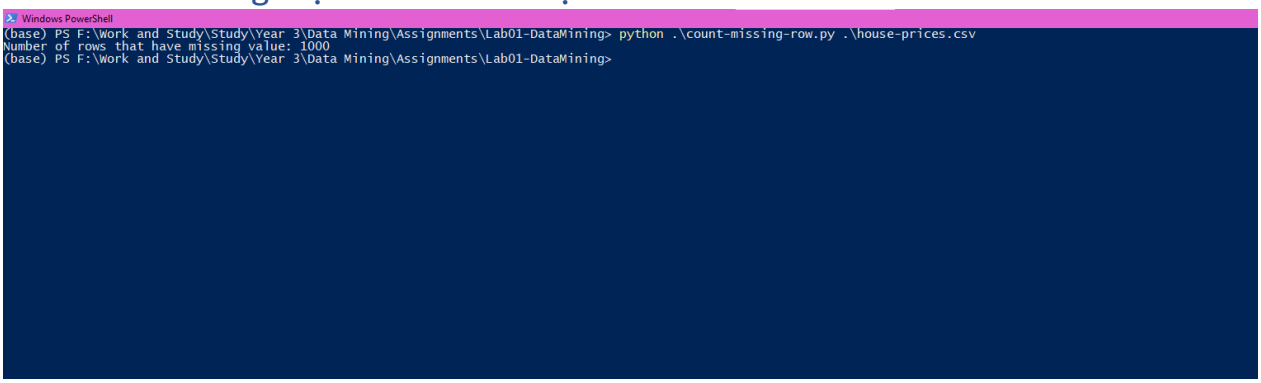
3 Cài đặt tiền xử lý dữ liệu (5 điểm)

1. Liệt kê các cột bị thiếu dữ liệu.



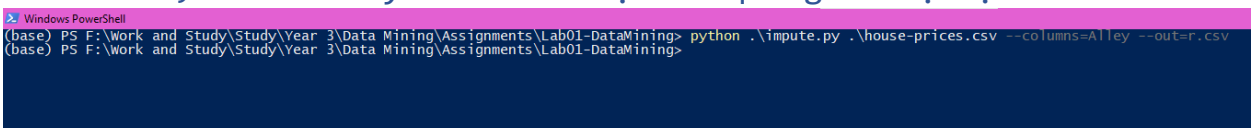
Kết quả cho ra đúng với tên những cột có giá trị bị thiếu trong dataset.

2. Đếm số dòng bị thiếu dữ liệu.



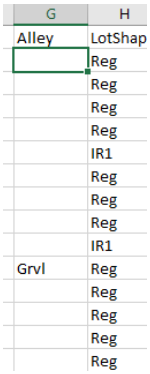
Kết quả cho thấy tất cả các dòng trong dataset đều có 1 hoặc nhiều hơn 1 thuộc tính bị thiếu.

3. Điền giá trị bị thiếu bằng phương pháp mean, median (cho thuộc tính numeric) và mode (cho thuộc tính categorical). Lưu ý: khi tính mean, median hay mode các bạn bỏ qua giá trị bị thiếu.



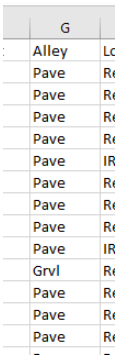
Tiến hành điền thêm dữ liệu vào cột Alley với phương thức Mode.

Trước khi impute

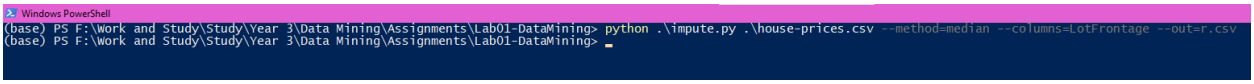


G	H
Alley	LotShap
	Reg
	Reg
	Reg
	Reg
	IR1
	Reg
	Reg
	Reg
	IR1
Grvl	Reg
	Reg
	Reg
	Reg
	Reg

Sau khi impute của cột Alley.

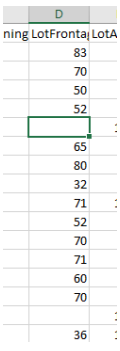


G
Alley
Pave
Pave
Pave
Pave
Pave
Pave
Pave
Pave
Pave
Grvl
Pave
Pave
Pave



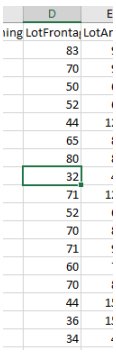
Tiến hành điền thêm dữ liệu vào cột LotFrontage với phương thức là Median

Trước khi impute.



D	
ning	LotFronta	LotA
	83	
	70	
	50	
	52	
	65	
	80	
	32	
	71	
	52	
	70	
	71	
	60	
	70	
	36	

Sau khi impute.



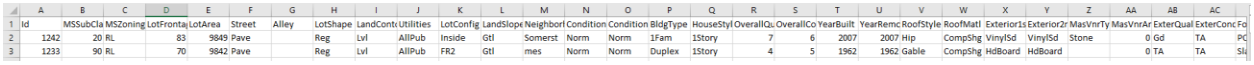
D	E
ing	LotFronta	LotAr
	83	
	70	
	50	
	52	
	44	
	65	
	80	
	32	
	71	
	52	
	70	
	71	
	60	
	70	
	44	
	36	
	34	

4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các dòng bị thiếu hơn 50% giá trị các thuộc tính).

```
Windows PowerShell
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining> python .\remove-missing-row.py .\house-prices.csv --threshold=4 --out=r.csv
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining>
```

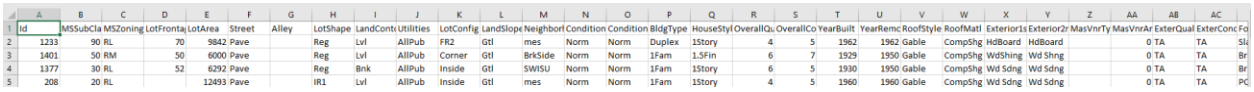
Xóa các dòng bị thiếu với ngưỡng tỷ lệ là 4%

Dữ liệu ban đầu:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Id	MSSubCla	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborl	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCo	YearBuilt	YearRemc	RoofStyle	RoofMatl	Exterior1s	Exterior2r	MasVnrTy	MasVnrAr	ExterQual	ExterConc
2	1242	20	RL	83	9849	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	VinylSd	VinylSd	Stone	0	Gd	TA
3	1233	90	RL	70	9842	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	HdBoard	HdBoard		0	TA	TA

Dữ liệu lúc sau khi xóa:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Id	MSSubCla	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborl	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCo	YearBuilt	YearRemc	RoofStyle	RoofMatl	Exterior1s	Exterior2r	MasVnrTy	MasVnrAr	ExterQual	ExterConc
2	1233	90	RL	70	9842	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	HdBoard	HdBoard		0	TA	TA
3	1401	50	RM	50	6000	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	WdShing	Wd Shng		0	TA	TA
4	1377	30	RL	52	6292	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	Wd Sdng	Wd Sdng		0	TA	TA
5	208	20	RL		12493	Pave		IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	Wd Sdng	Wd Sdng		0	TA	TA

⇒ 1 dòng bị xóa khỏi dataset.

5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu).

```
  -o , --out          output file's name
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining> python .\remove-missing-col.py .\house-prices.csv --threshold=4 --out=r.csv
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining>
```

Thực hiện xóa các cột bị thiếu dữ liệu với ngưỡng là 4%.

```
len(df.columns)
✓ 0.5s
81

len(df2.columns)
✓ 0.4s
69
```

Có 12 cột bị xóa ra khỏi dataset.

6. Xóa các mẫu bị trùng lặp.

```
  -o , --out    output file's name
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining> python .\del-duplicate.py .\house-prices.csv --out=r.csv
```

Không có mẫu nào bị xóa vì không có mẫu nào trùng nhau.

7. Chuẩn hóa một thuộc tính numeric bằng phương pháp min-max và Z-score.

```
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining> python .\standardize.py .\house-prices.csv --method=min-max --column=MSSubClass --out=r.csv
(base) PS F:\Work and Study\Study\Year 3\Data Mining\Assignments\Lab01-DataMining>
```

Tiến hành chuẩn hóa cột MSSubClass dùng phương pháp min-max.

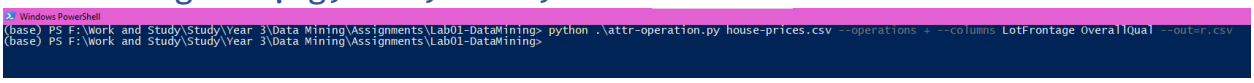
Tước khi chuẩn hóa.

Id	MSSubCla	MS
1242	20	RL
1233	90	RL
1401	50	RM
1377	30	RL
208	20	RL
1392	90	RL
980	20	RL
484	120	RM
392	60	RL
730	30	RM
255	20	RL
1094	20	RL
1021	20	RL
1341	20	RL
1025	20	RL
848	20	RL
457	70	RM

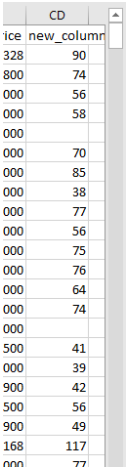
Sau khi chuẩn hóa.

	MSSubCla	MS
2	0	RL
3	0.411765	RL
1	0.176471	RM
7	0.058824	RL
8	0	RL
2	0.411765	RL
0	0	RL
4	0.588235	RM
2	0.235294	RL
0	0.058824	RM
5	0	RL
4	0	RL
1	0	RL
1	0	RL
5	0	RL
8	0	RL
7	0.294118	RM

8. Tính giá trị biểu thức thuộc tính: ví dụ đối với một tập dữ liệu có chứa 2 thuộc tính width và height thì biểu thức width * height sẽ trả về tập dữ liệu cũ với một thuộc tính mới có giá trị ở mỗi mẫu là tích của thuộc tính width và height trong mẫu tương ứng, với điều kiện cả 2 giá trị width và height đều không bị thiếu, trong trường hợp bị thiếu thì giá trị biểu thức coi như bị thiếu. Lưu ý: biểu thức có thể có nhiều thuộc tính và nhiều phép toán bao gồm cộng, trừ, nhân, chia.

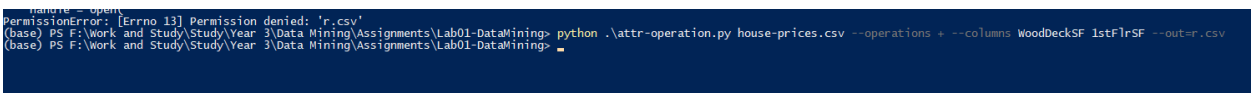


Tiến hành thêm 1 cột mới vào dữ liệu dùng dữ liệu cột **LotFrontage** cộng với **OverallQual**.

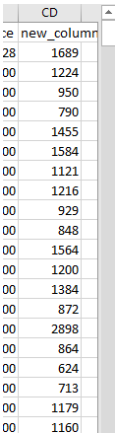


CD	
ice	new_column
328	90
800	74
000	56
000	58
000	
000	70
000	85
000	38
000	77
000	56
000	75
000	76
000	64
000	74
000	
500	41
000	39
900	42
500	56
900	49
168	117
000	77

Cột dữ liệu mới thêm vào có dạng.



Thực hiện thêm 1 cột mới với công thức là **WookDeckSF** cộng với **1stFlrSF**.



CD	
:e	new_column
28	1689
00	1224
00	950
00	790
00	1455
00	1584
00	1121
00	1216
00	929
00	848
00	1564
00	1200
00	1384
00	872
00	2898
00	864
00	624
00	713
00	1179
00	1160

Cột dữ liệu mới thêm vào có dạng.