

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



DATA MINING
LAB 1: PREPROCESSING

Lớp: KTDL & UD 18_21

Nhóm thực hiện:

18120078 – Ngô Phù Hữu Đại Sơn

18120253 – Mai Ngọc Tú

MỤC LỤC

I.	Thông tin khái quát.....	2
I.	Thông tin nhóm	2
II.	Bảng phân công công việc	2
III.	Github:	2
B.	Nội dung	3
I.	Mục tiêu của đồ án:	3
II.	Yêu cầu:	3
III.	Triển khai.....	4
1.	Cài đặt Weka:.....	4
2.	Làm quen với Weka:.....	5
3.	Cài đặt tiền xử lý dữ liệu:	17
IV.	Dánh giá đồ án.....	31
1.	Mức độ hoàn thành của các thành viên	31
2.	Mức độ hoàn thành đồ án:	31
VII.	Nguồn tham khảo	31



I. THÔNG TIN KHÁI QUÁT

I. Thông tin nhóm

MSSV	Họ tên	Vai trò
18120078	Ngô Phù Hữu Đại Sơn	Nhóm trưởng
18120253	Mai Ngọc Tú	Thành viên

II. Bảng phân công công việc

MSSV	Công việc phụ trách	Thời gian thực hiện
18120078	Tìm hiểu chức năng Explorer	26/10 – 26/10
18120078	Tìm hiểu tab Proprocessing của Explorer	26/10 – 26/10
18120253	Khám phá tập dữ liệu Breast Cancer	26/10 – 26/10
18120253	Khám phá tập dữ liệu Weather	27/10 – 27/10
18120078	Khá phá tập dữ liệu Tín dụng Đức	26/10 – 26/10
18120253	Liệt kê các cột bị thiếu dữ liệu	29/10 – 31/10
18120253	Đếm số dòng bị thiếu dữ liệu	29/10 – 31/10
18120253	Điền giá trị bị thiếu	1/11 – 3/11
18120253	Xóa các dòng bị thiếu dữ liệu	1/11 – 3/11
18120078	Xóa các cột bị thiếu dữ liệu	29/10 – 31/10
18120078	Xóa các mẫu bị trùng lặp	29/10 – 31/10
18120078	Chuẩn hóa một thuộc tính numeric	1/11 – 3/11
18120078	Tính giá trị biểu thức thuộc tính	1/11 – 3/11
18120078	Tổng hợp các chức năng & kiểm thử lần cuối	5/11 – 6/11

III. Github:

<https://github.com/IrisStream/Data-Mining>

B. NỘI DUNG

I. Mục tiêu của đồ án:

- Làm quen với các thao tác cơ bản trong tác vụ tiền xử lý dữ liệu thông qua việc áp dụng các công cụ hỗ trợ được cung cấp bởi phần mềm mã nguồn mở Weka.
- Phát huy kỹ năng lập trình để tự cài đặt các thủ tục tiền xử lý dữ liệu đơn giản.

II. Yêu cầu:

1. Cài đặt Weka.

- Chụp hình giao diện chức năng Exploere cùng màn hình desktop.
- Tìm thư mục data trong thư mục cài đặt của Weka và mở một tập dữ liệu bất kì (có phần mở rộng là arff). Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess. Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.

2. Làm quen với Weka.

- Khám phá bộ dữ liệu Breast Cancer
- Khám phá bộ dữ liệu Weather
- Khám phá bộ dữ liệu Tín dụng Đức

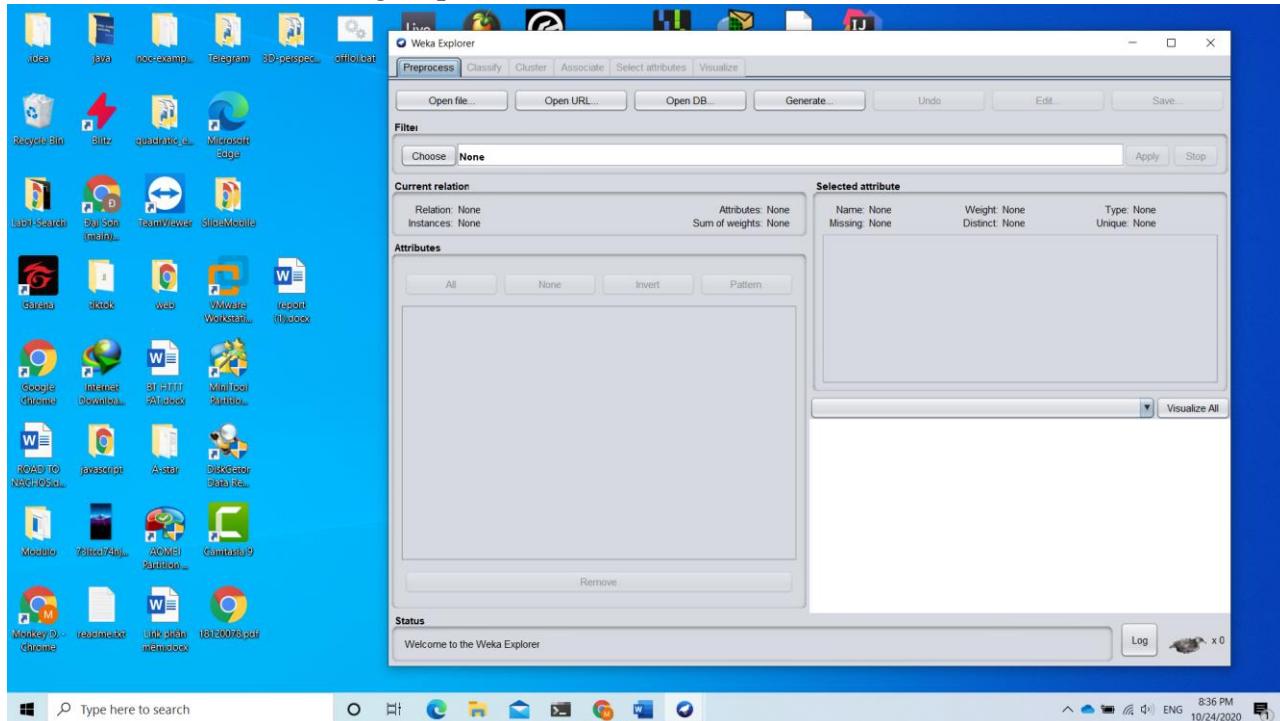
3. Cài đặt tiền xử lý dữ liệu

- Cài đặt chương trình khám phá bộ dữ liệu house_prices gồm có các chức năng:
 - + Liệt kê các cột bị thiếu dữ liệu.
 - + Đếm số dòng bị thiếu dữ liệu.
 - + Diện giá trị bị thiếu.
 - + Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.
 - + Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.
 - + Xóa các mẫu bị trung lặp.
 - + Chuẩn hóa một thuộc tính numeric bằng phương pháp min-mã và Z-score.
 - + Tính giá trị biểu thức thuộc tính.

III. Triển khai

1. Cài đặt Weka:

- Giao diện chức năng Explorer:

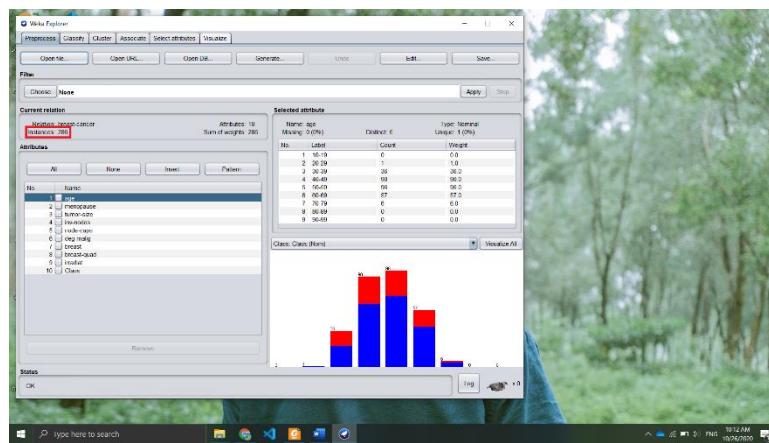


- Giải thích ý nghĩa các nhóm điều khiển
 - Current Relation: Cho biết thông tin về tập dữ liệu đang khảo sát
 - Relation: Tên của quan hệ
 - Instances: Số mẫu
 - Attributes: Số thuộc tính
 - Attributes: Danh sách các thuộc tính được hiển thị theo đúng thứ tự khai báo ở trong file
 - Selected attribute: Liệt kê các giá trị tương ứng với thuộc tính đang được chọn ở nhóm Attributes
- Giải thích ý nghĩa các tab trong giao diện Explorer:
 - Preprocessing: có chức năng lọc dữ liệu, mặc khác cung cấp cho người dùng các thông tin về tập dữ liệu đang xử lý.
 - Classify: có chức năng phân lớp.
 - Cluster: có chức năng gom cụm.
 - Select Attributes: chọn các thuộc tính thích hợp nhất trong dữ liệu.
 - Associate: có chức năng rút ra các luật kết hợp.
 - Visualize: có chức năng trực quan hóa dữ liệu.

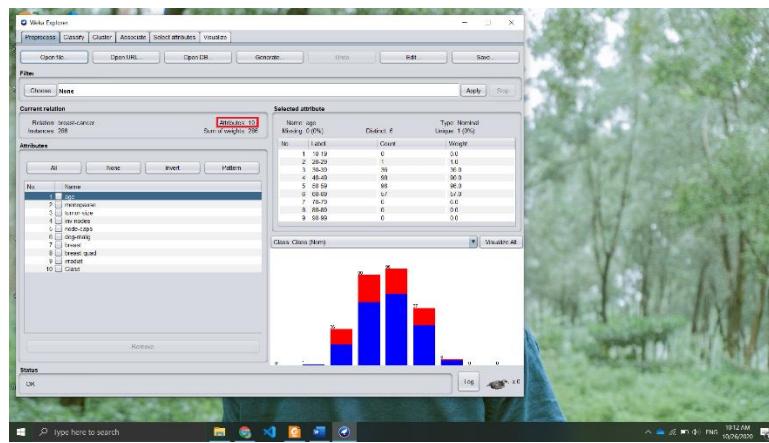
2. Làm quen với Weka:

2.1 Đọc dữ liệu vào Weka:

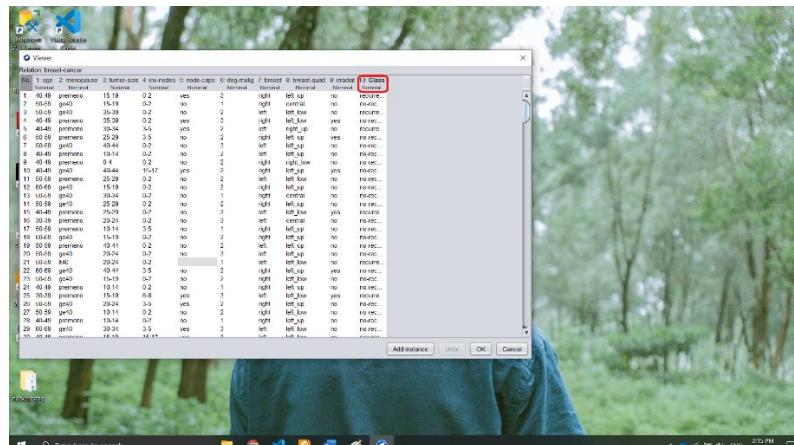
1/ - Tập dữ liệu có 286 mẫu (Instances)



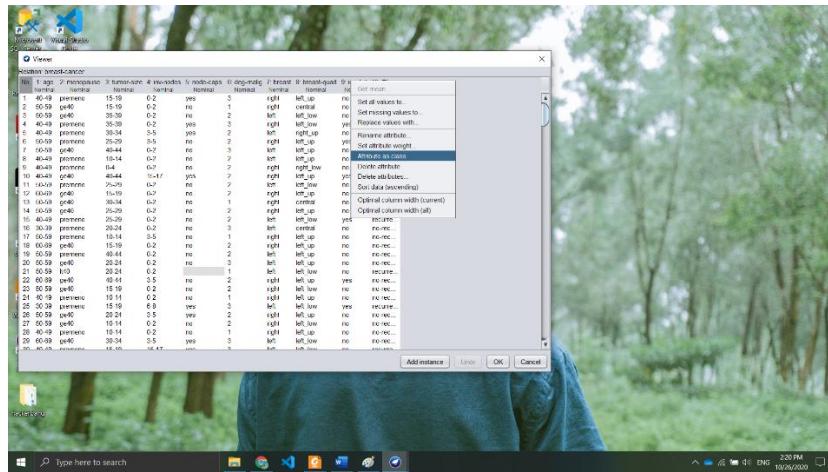
2/ - Tập dữ liệu có 10 thuộc tính (attributes)



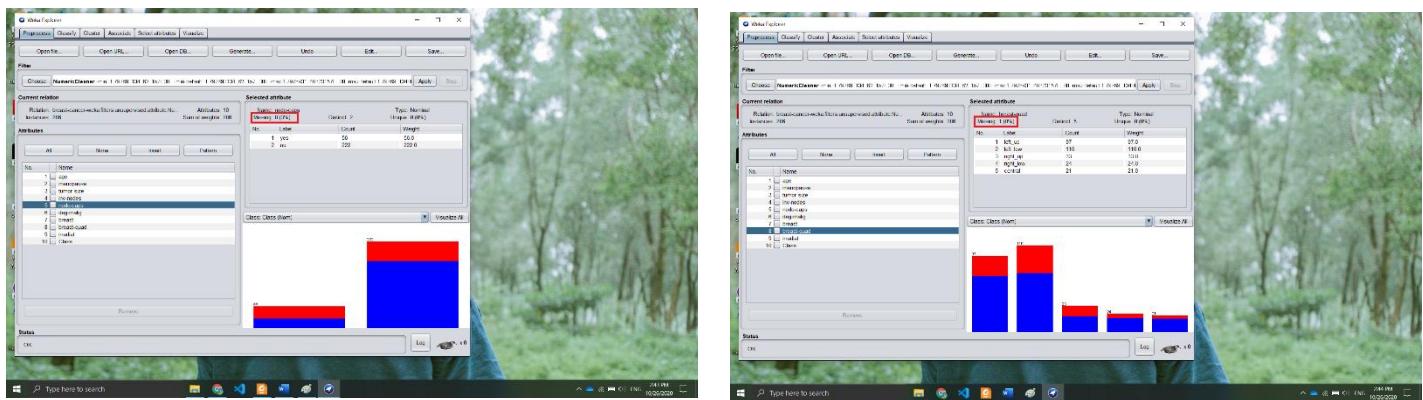
3/ - Thuộc tính Class dùng làm lớp (class)



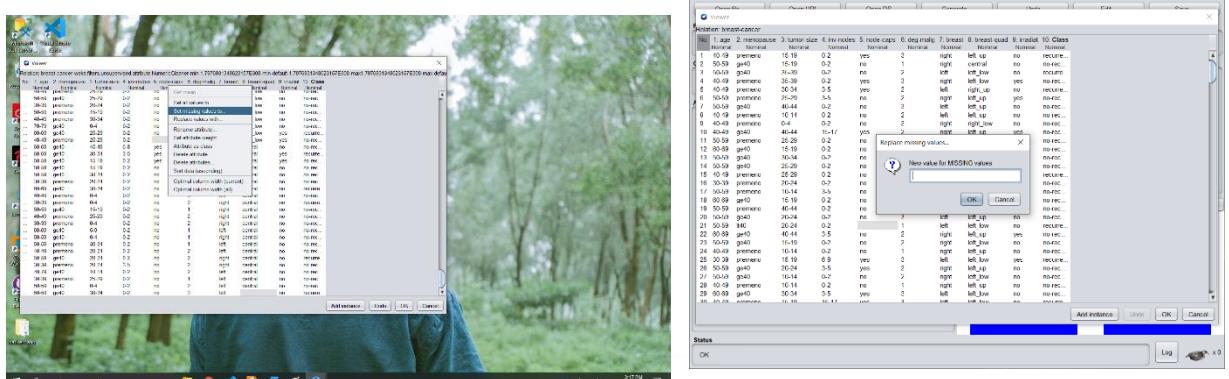
- Có thể thay đổi thuộc tính dùng làm lớp. Ta có thể thay đổi bằng cách chọn một thuộc tính khác, sau đó nhấn chuột phải và chọn **Attribute as class**



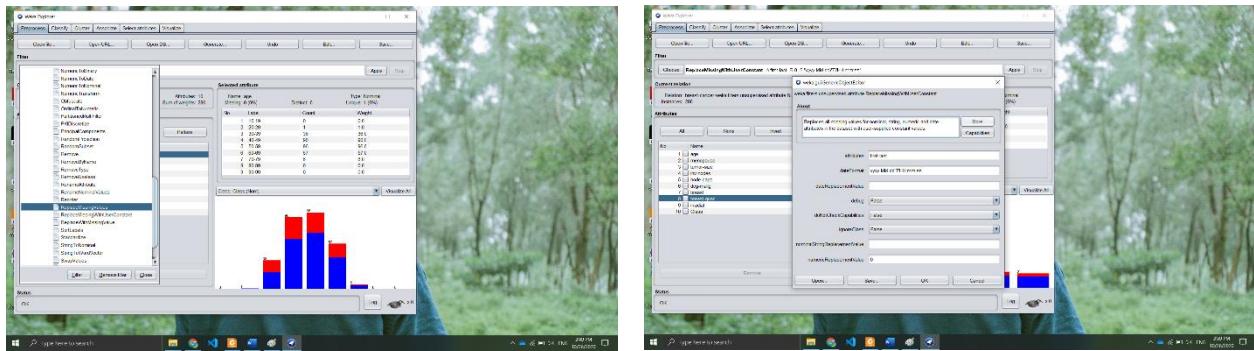
- 4/ - Có 2 thuộc tính bị thiếu dữ liệu: node-caps, breast-quad
 + Thuộc tính thiếu dữ liệu ít nhất: breast-quad (thiếu 1)
 + Thuộc tính thiếu dữ liệu nhiều nhất: node-caps (thiếu 8)



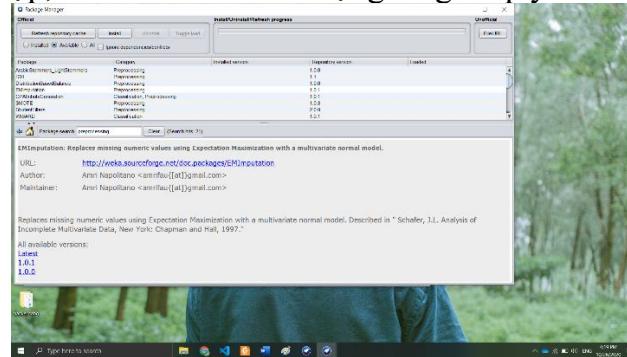
- Các cách giải quyết vấn đề missing values:
 + Cách 1: Chọn Edit, chọn các thuộc tính thiếu dữ liệu, sau đó nhấn chuột phải và chọn **Set missing value to...**. Gán các ô bị thiếu dữ liệu bằng giá trị nhập trong **New value for MISSING values**. Lưu ý giá trị nhập vào trong **New value for MISSING values** là giá trị được cho phép của thuộc tính đó thì mới gán thành công. Ví dụ: thuộc tính node-caps chỉ được gán [yes,no] cho các ô bị thiếu dữ liệu.



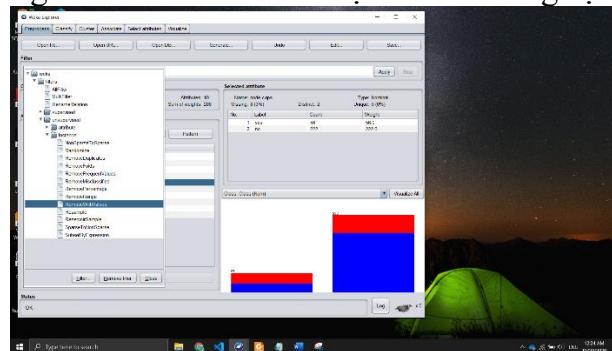
+ Cách 2: dùng filter **ReplaceMissingValues** hoặc **ReplaceMissingWithUserConstant** trong Weka. Filter **ReplaceMissingValues** giúp thêm tự động các ô bị thiếu dữ liệu trong các thuộc tính, dữ liệu được thêm vào là giá trị được cho phép của thuộc tính đó. Còn filter **ReplaceMissingWithUserConstant** sẽ gán cho những ô thiếu dữ liệu của các thuộc tính một giá trị mặc định mà người dùng cho. Ta có thể gán giá trị mặc định dành cho thuộc tính dạng numeric, nominal, date...



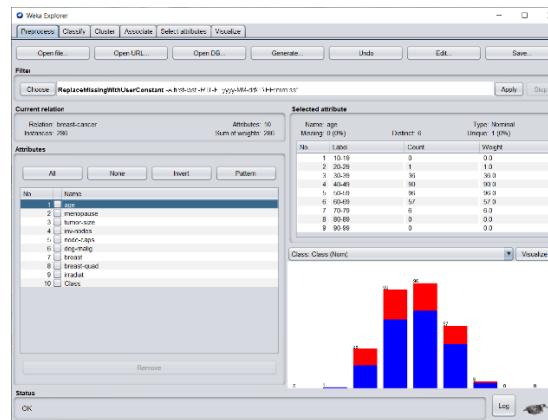
+ Cách 3: Vào Weka GUI chọn **Tools → Package manager**. Trong ô **Package search** nhập “preprocessing” và kết quả là các filter dành cho quá trình tiền xử lý dữ liệu, trong đó có các filter dùng để giải quyết vấn đề missing values. Chọn filter phù hợp, nhấn Install và sử dụng để giải quyết missing values.



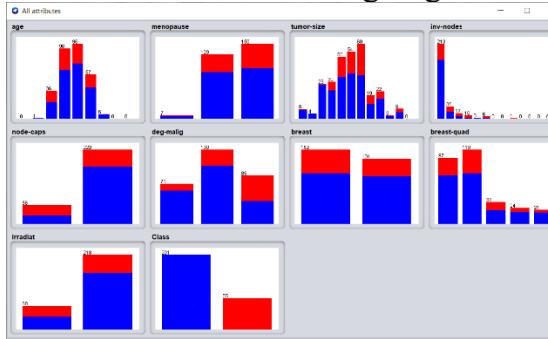
+ Cách 4: Nếu các ô thiếu dữ liệu mang giá trị “Unknow”, ta có thể xóa bỏ những ô có dữ liệu là “Unknow” ra khỏi tập dữ liệu của thuộc tính đó. Ta dùng filter **RemoveWithValues** để xóa các dòng mà ô dữ liệu là “Unknow” và điều này sẽ làm giảm số Instances đi một số k số dòng bị thiếu dữ liệu.



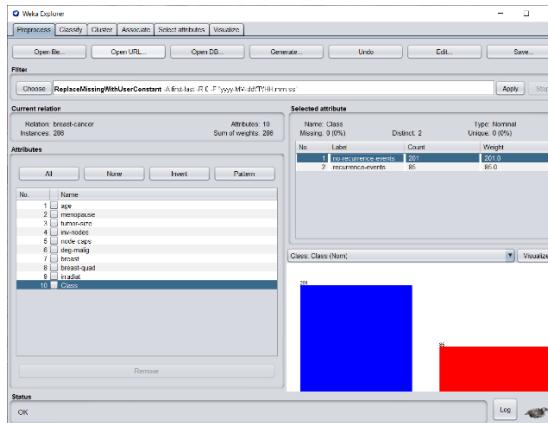
5/ - Ý nghĩa của đồ thị trong của sổ Explorer: đồ thị thể hiện các giá trị Count của từng Label trong cùng một thuộc tính. Các cột sắp xếp theo thứ tự của các Label và chiều cao của cột là giá trị Count tương ứng.



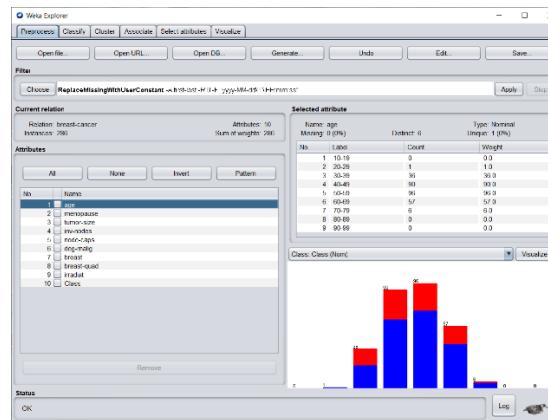
- Đặt tên cho đồ thị là tên của thuộc tính tương ứng với đồ thị đó.



- Màu xanh của đồ thị tương ứng với Label no-recurrence-events của attribute Class (attribute dùng làm lớp). Màu đỏ của đồ thị tương ứng với Label recurrence-events của attribute Class.



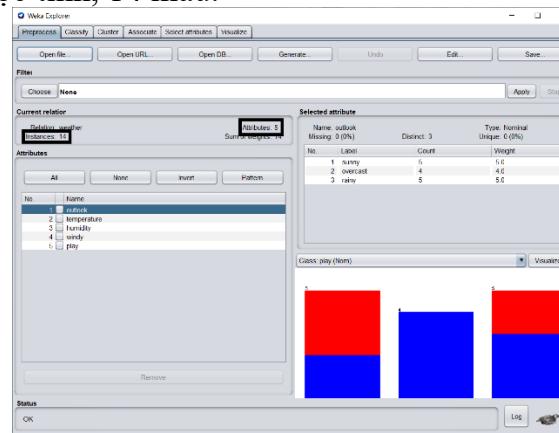
- Đồ thị biểu diễn từng giá trị Count của mỗi Label trong từng thuộc tính theo từng cột đồ thị và chia từng cột theo từng màu khác nhau ứng với các giá trị trong cùng một Label thỏa mãn các giá trị khác nhau của attribute đóng vai trò là lớp (attribute Class). Ví dụ, trong attribute age, các cột ứng với các Label, màu xanh ứng với thỏa mãn no-recurrence-events của attribute Class, màu đỏ ứng với thỏa mãn recurrence-events của attribute Class.



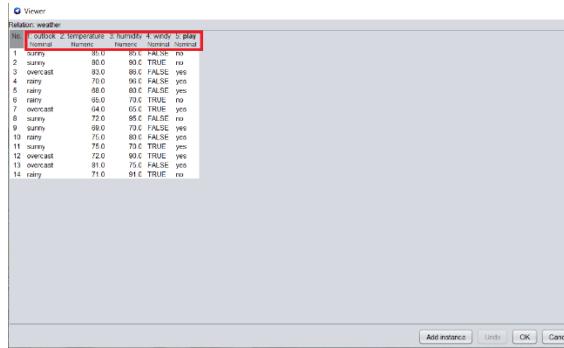
2.2 Khám phá tập dữ liệu Weather:

1/

- Tập dữ liệu có 5 thuộc tính, 14 mẫu.



- Thuộc tính kiểu dữ liệu nominal: outlook, windy, play
- Thuộc tính kiểu dữ liệu numeric: temperature, humidity
- Thuộc tính lớp: play



2/

- Five-number summary của thuộc tính temperature:
 - Minimum: 64.0
 - Maximum: 85.0
 - Median: 72.0
 - Q1: 69.0
 - Q3: 80.0

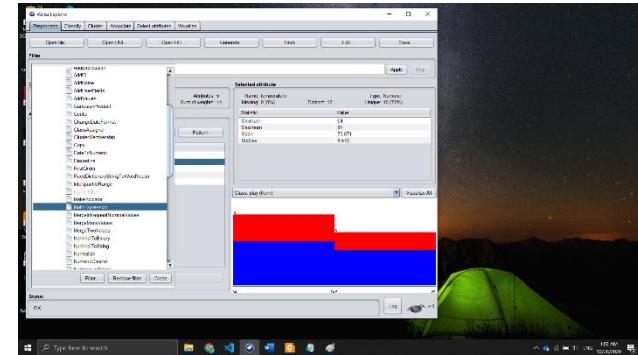
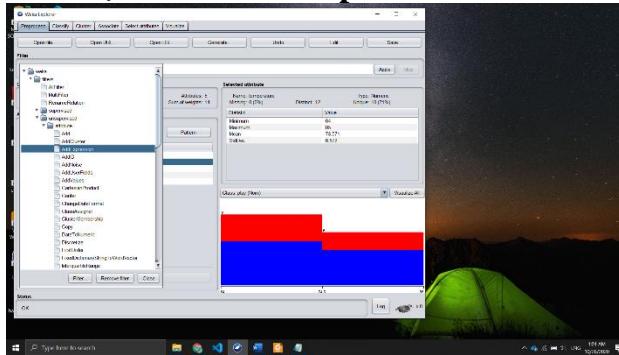
Relation: weather					
No.	outlook	temperature	humidity	windy	play
1	overcast	64.0	65.0	TRUE	yes
2	rainy	65.0	70.0	TRUE	no
3	rainy	69.0	80.0	FALSE	yes
4	sunny	72.0	70.0	FALSE	yes
5	rainy	70.0	90.0	FALSE	yes
6	rainy	71.0	91.0	TRUE	no
7	sunny	72.0	85.0	FALSE	yes
8	overcast	72.0	90.0	TRUE	yes
9	rainy	75.0	80.0	FALSE	yes
10	sunny	75.0	70.0	FALSE	yes
11	sunny	80.0	90.0	TRUE	yes
12	overcast	81.0	75.0	FALSE	yes
13	overcast	83.0	80.0	FALSE	yes
14	sunny	90.0	95.0	FALSE	no

- Five-number summary của thuộc tính :

- + Minimum: 65.0
- + Maximum: 96.0
- + Median: 82.5
- + Q1: 70.0
- + Q2: 90.0

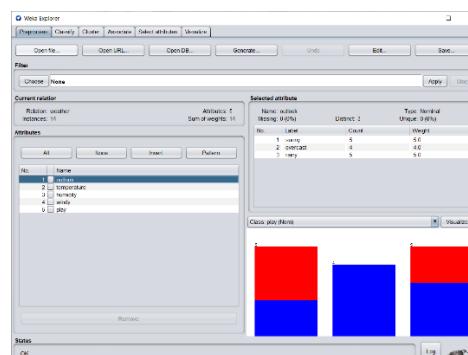
Relation: weather					
No.	outlook	temperature	humidity	windy	play
1	overcast	64.0	65.0	TRUE	yes
2	rainy	65.0	70.0	TRUE	no
3	sunny	69.0	70.0	FALSE	yes
4	sunny	72.0	70.0	FALSE	yes
5	overcast	72.0	75.0	FALSE	yes
6	rainy	68.0	80.0	FALSE	yes
7	rainy	75.0	80.0	FALSE	yes
8	sunny	69.0	80.0	FALSE	yes
9	overcast	83.0	86.0	FALSE	yes
10	sunny	80.0	90.0	TRUE	no
11	overcast	72.0	90.0	TRUE	yes
12	rainy	75.0	90.0	FALSE	yes
13	sunny	72.0	95.0	FALSE	no
14	rainy	70.0	96.0	FALSE	yes

- Weka cung cấp phương thức tính những giá trị này trong filter **AddExpression** hoặc filter **MathExpression**



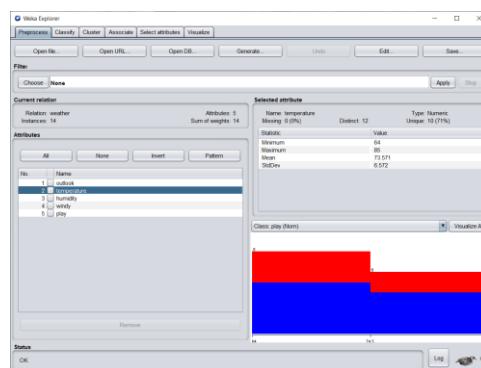
3/

- Thuộc tính outlook:

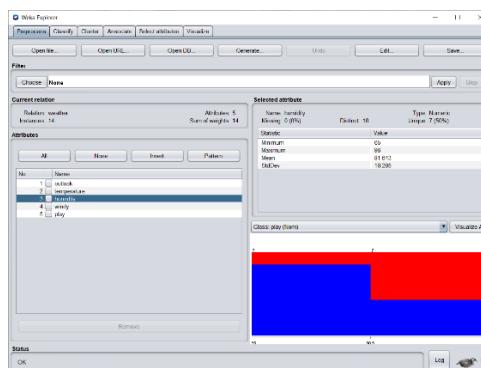


- Thuộc tính temperature:

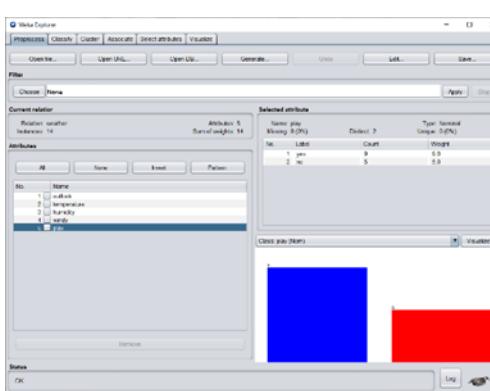
- Thuộc tính humidity:



- Thuộc tính windy:

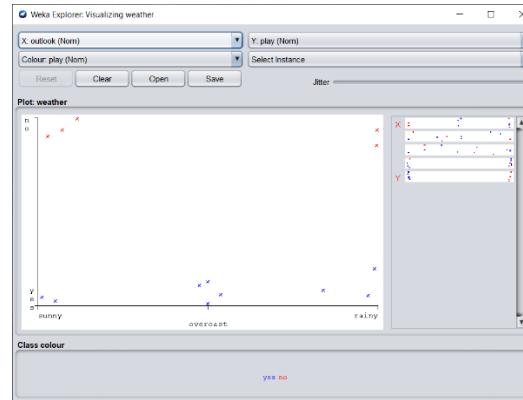


- Thuộc tính play:

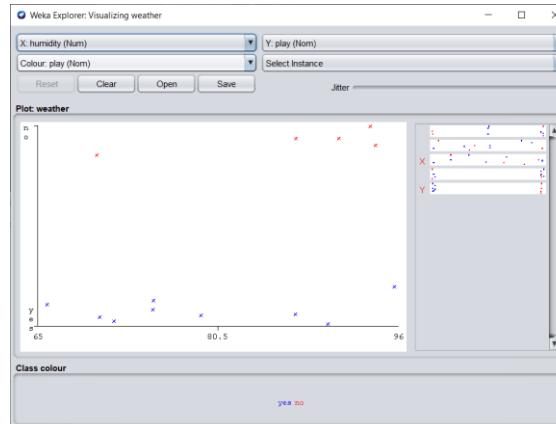


4/

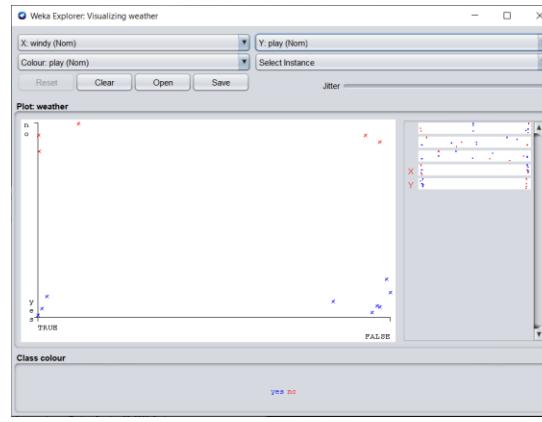
- Thuật ngữ sử dụng là **scatter-plot matrix**.
- Các cặp thuộc tính tương quan:
 - + outlook – play: Nếu giá trị outlook là “overcast” thì thuộc tính play sẽ có giá trị “yes” thể hiện nếu thời tiết âm u thì chắc chắn sẽ đi chơi.



+ humidity – play: humidity có giá trị từ 65 đến 80.5 thì thuộc tính “play” có khả năng cao là “yes”.



+ windy – play: giá trị thuộc tính windy là “false” thì thuộc tính play có khả năng cao là “yes”.



2.3 Khám phá tập dữ liệu Tín dụng Đức:

1. Nội dung của phần ghi chú (Comment) trong credit_g.arff nói về:
 - + Tên của tập dữ liệu: German credit data
 - + Thông tin về tác giả (Professor Dr. Hans Hofmann)
 - + Số lượng mẫu khảo sát: 1000
 - + Có 2 bộ dữ liệu được cung cấp. Bản gốc do Giáo sư Hofmann cung cấp, chứa các thuộc tính phân loại nằm trong file “german.data”
 - + Đối với các thuật toán cần thuộc tính số học, Đại học Strathclyde đã cung cấp tệp "german.data-numeric". Tệp này đã được chỉnh sửa và thêm một số biến chỉ số để phù hợp với các thuật toán không thể đối phó với các biến phân loại. Một số thuộc tính phân loại có thứ tự (chẳng hạn như thuộc tính thứ 17) đã được mã hóa dưới dạng số nguyên.
 - + Số thuộc tính của “german”: 21 (7 số học, 13 phân loại và 1 thuộc tính lớp)
 - + Số thuộc tính của “german.numer”: 24 (24 số học)
 - + Các thuộc tính được mô tả cho người Đức (german)
 - + Ma trận chi phí: (các dòng thể hiện phân lớp trong thực tế và cột thể hiện dự đoán phân lớp).
 - + Gán lại tên các thuộc tính thành các tên có ý nghĩa
 - + Mô tả 5 thuộc tính:
 - *Thuộc tính 1:* tình trạng của các tài khoản thanh toán hiện tại
 - A11: < 0 DM (Deutsche Mark, 1 đơn vị tiền tệ, quy đổi khoảng 90 cents Canada)
 - A12: 0 <= ... < 20 DM
 - A13: ... >= 200 DM
 - A14: không có tài khoản thanh toán
 - *Thuộc tính 2:* Thời gian trong tháng
 - *Thuộc tính 3:* Lịch sử tín dụng:
 - A30: không có tín dụng nào được thực hiện hoặc tất cả các tín dụng đã được chi trả hợp lệ
 - A31: Tất cả các tín dụng tại ngân hàng này đã được chi trả hợp lệ
 - A32: existing credits paid back duly till now
 - A33: delay in paying off in the past
 - A34: critical account/ other credits existing (not at this bank)
 - *Thuộc tính 4:* Mục đích vay tiền:
 - Mua xe mới
 - Mua xe cũ
 - Mua nội thất
 - Mua TV/Radio
 - Mua Đồ gia dụng
 - Sửa chữa
 - Giáo dục
 - Du lịch
 - Huấn luyện lại
 - Kinh doanh
 - Khoản khác
 - *Thuộc tính 13:* Tuổi

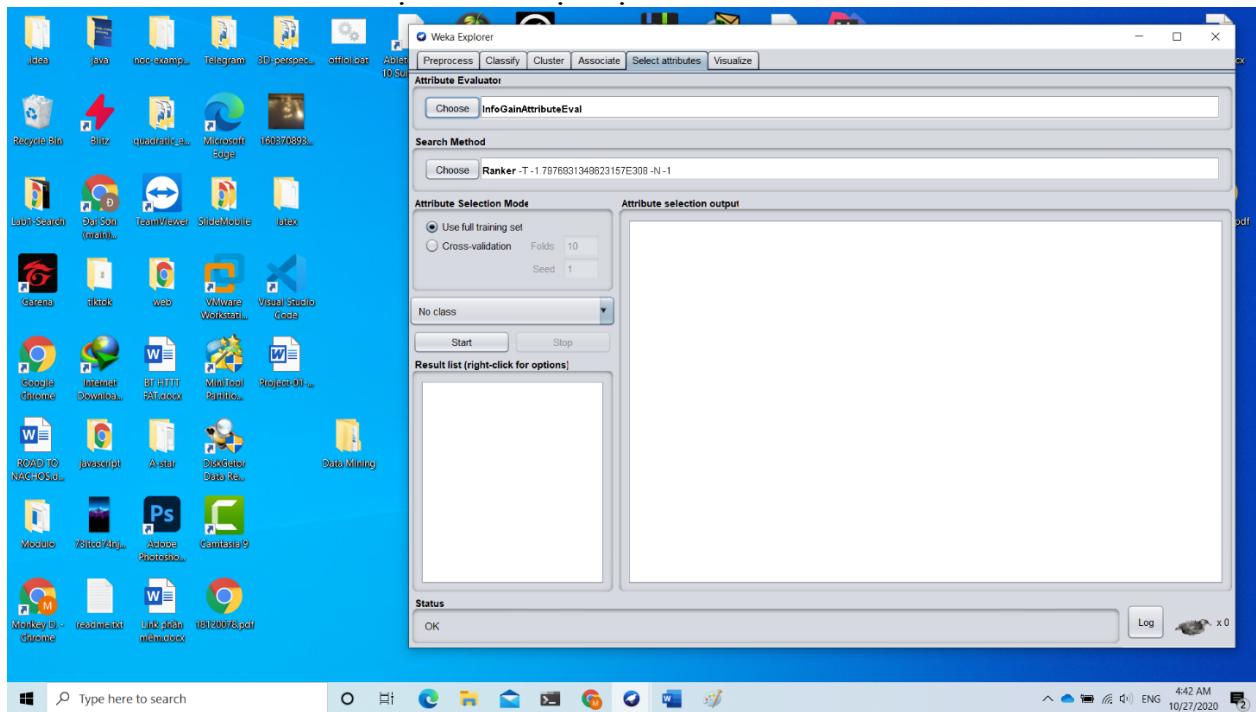
2. *Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?*
 - + Tên thuộc tính lớp là “class”
 - + Phân bố lớp không đồng đều. Các đánh giá tốt nhiều hơn các đánh giá không tốt (7:3)
3. *Sử dụng tab Select attributes. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.*
 - + Để chọn lọc thuộc tính trong weka, cần xác định Attribute Evaluator và Search Method.
 - + *Attribute Evaluator:*
 - *Correlation-based Feature Subset Selection(CfsSubsetEval):* đánh giá giá trị của tập con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng đối tượng cùng với mức độ dựa trên giữa chúng.
 - *Classifier Atribute Evaluator (ClassifierAttributeEval):* Đánh giá giá trị của tập con các thuộc tính bằng cách sử dụng bộ phân loại do người dùng chỉ định.
 - *Classifier Subset Evaluator (ClassifierSubsetEval):* đánh giá tập con các thuộc tính dựa trên tập huấn luyện.
 - *GainRatioAttributeEval:* Chọn các thuộc tính có độ đo Ratio Gain cao.
 - *OneRAttributeEval:* đánh giá tập con các thuộc tính dựa trên bảng tần số giữa tập thuộc tính với thuộc tính lớp.
 - *Principal components Analysis (PCA):* chọn tập thuộc tính không có tương quan với nhau và các thuộc tính trong tập này là tổ hợp tuyến tính của các thuộc tính ban đầu.
 - *ReliefFAttributeEval:* Đánh giá giá trị của một thuộc tính bằng cách liên tục lấy mẫu một đối tượng và xem xét giá trị của thuộc tính đã cho với đối tượng gần nhất của cùng một lớp và khác lớp.
 - *Correlation Based Feature Selection (CorrelationAttributeEval):* Chọn các thuộc tính có tính tương quan với thuộc tính lớp cao. Kết quả trả về số nguyên nằm trong đoạn [-1;1].
 - + Positice Correlation(1): 2 thuộc tính tỉ lệ thuận với nhau
 - + Neutral Correlation(0): 2 thuộc tính không tương quan với nhau
 - + Negative Correlation(-1): 2 thuộc tính tỉ lệ nghịch với nhau
 - *Information Gain Based Feature Selection (InfoGainAttributeEval):* Chọn các thuộc tính có độ đo Information Gain cao. Information Gain có giá trị từ 0 (không có thông tin) đến 1(chứa thông tin tối đa)
 - *Learner Based Feature Selection (WrapperSubsetEval):* Sử dụng 1 thuộc toán học mạnh để áp dụng trên 1 tập các thuộc tính. Độ chính xác của giải thuật học trên tập thuộc tính nào được xấp xỉ nhờ cross-validation. Chọn tập các thuộc tính đưa ra kết quả học tốt nhất
 - *SymmetricalUncertAttributeEval:* Đánh giá giá trị của một tập hợp các thuộc tính bằng cách đo SU (Symmetrical Uncertainty) đối với thuộc tính phân lớp
 - + *Search Method:*
 - *Best First:* Tìm kiếm không gian của các tập con thuộc tính bằng thuật toán Beam search. Việc đặt số lượng nút không cài tiến liên tiếp cho phép

kiểm soát mức độ thực hiện quay lui. Best First có thể bắt đầu với tập thuộc tính trống và tìm kiếm hoặc bắt đầu với tập hợp đầy đủ các thuộc tính và tìm kiếm ngược lại, hoặc bắt đầu tại bất kỳ điểm nào và tìm kiếm theo cả hai hướng (bằng cách xem xét tất cả các lần thêm và xóa thuộc tính đơn lẻ có thể có tại một điểm nhất định).

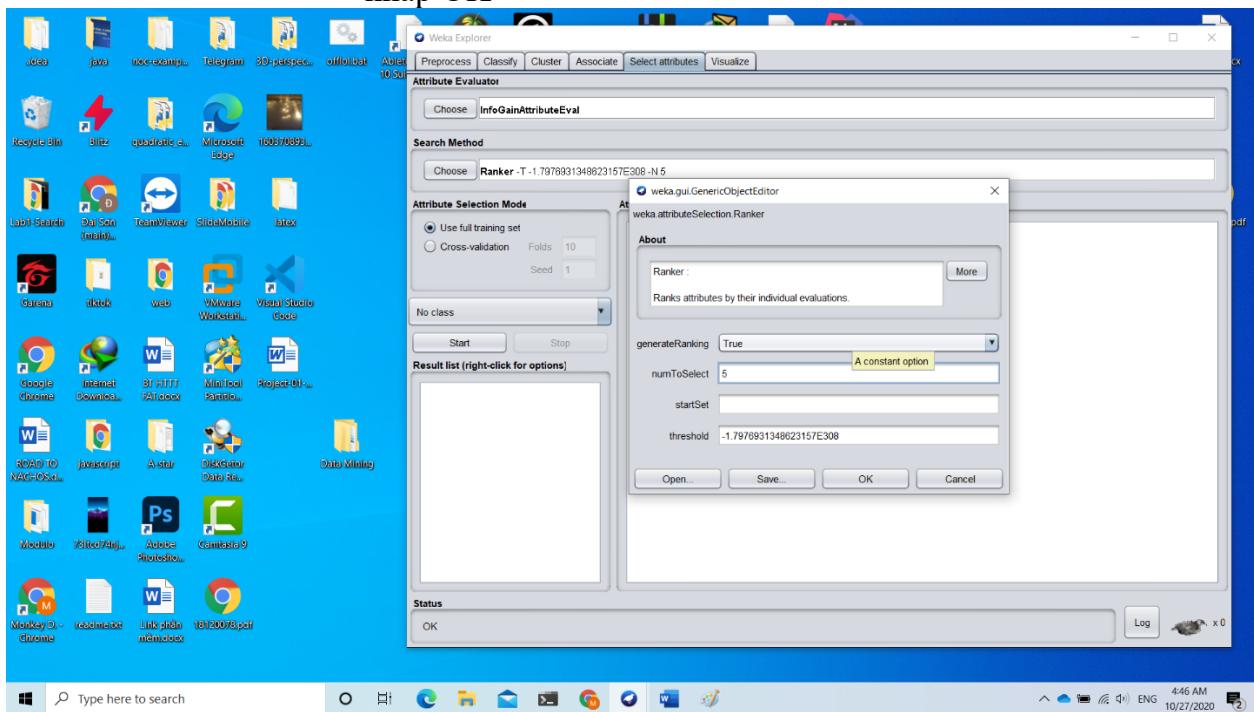
- *GreedyStepwise*: Thực hiện tìm kiếm leo đồi tham lam theo 2 hướng tiến và lùi thông qua không gian của các tập con thuộc tính. Có thể bắt đầu với không có / tất cả các thuộc tính hoặc từ một điểm tùy ý trong không gian. Dừng khi việc thêm / xóa bất kỳ thuộc tính còn lại nào dẫn đến giảm đánh giá.
- *Ranker*: xếp hạng các thuộc tính theo đánh giá cá nhân của nó.

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.*

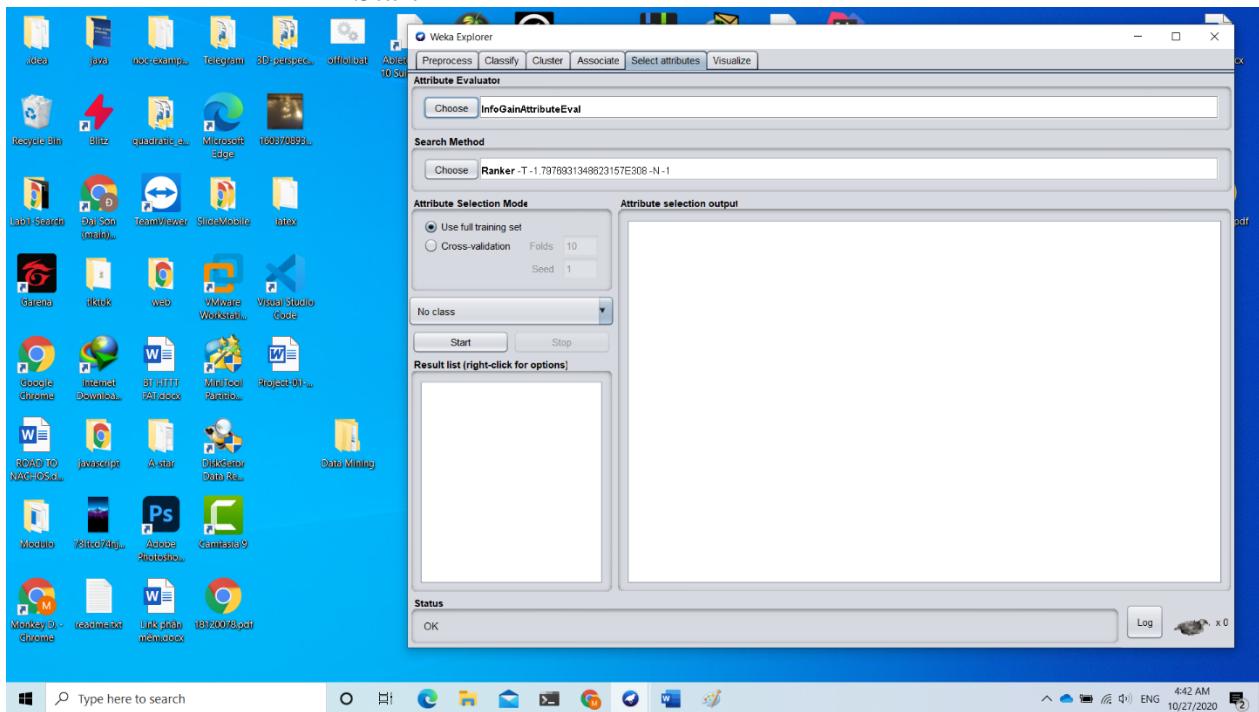
- Cần sử dụng phương pháp Ranker với 1 đồ độ đánh giá đơn thuộc tính (Ví dụ như Information Gain). Đến xếp hạng cho 5 thuộc tính có độ tương quan cao nhất với thuộc tính lớp.
- Các bước thực hiện:
 - Chọn InfoGainAttributeEval tại mục Attribute Evaluator, chọn mục Ranker tại mục Search Method



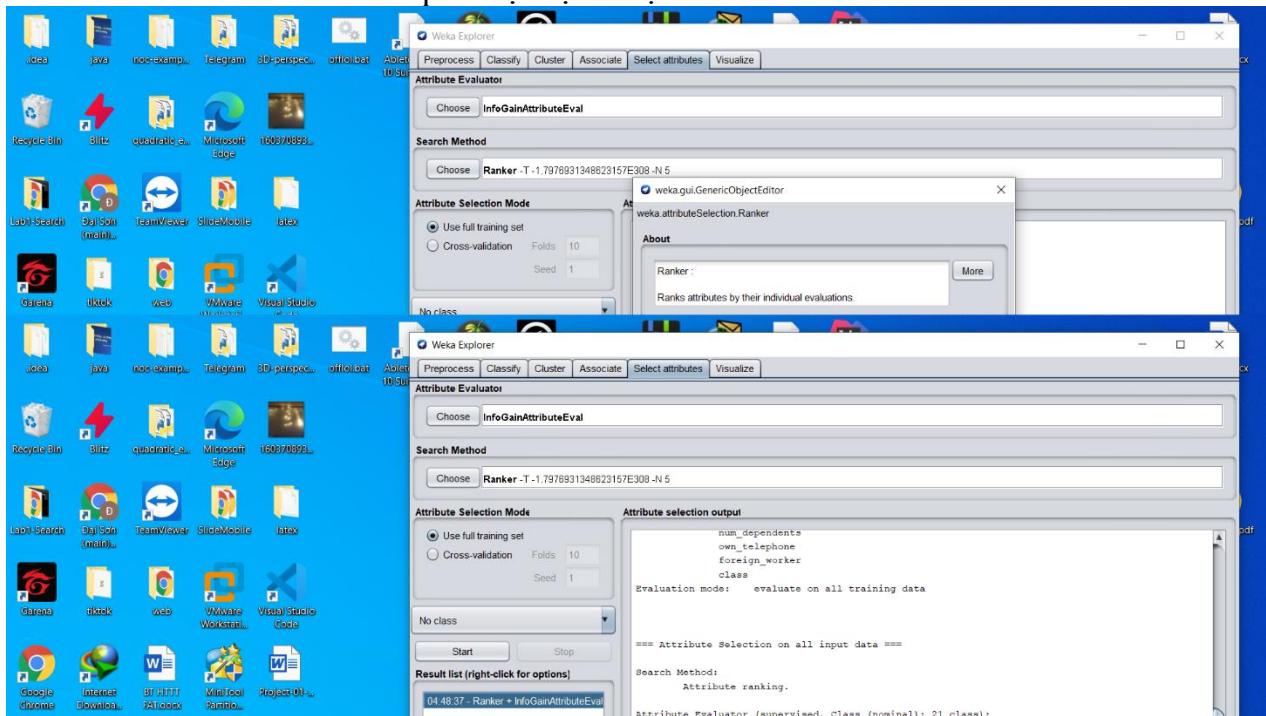
- Nhấp chuột vào ô Search Method, đổi numToSelect thành 5 và nhấp OK



- Tại mục Atribute Selection Mode chọn Use full training set và nhấp Start



- Kết quả chọn lọc thuộc tính



3. Cài đặt tiền xử lý dữ liệu:

3.1 Liệt kê các cột bị thiếu dữ liệu

- Cú pháp: `python missing_list.py data_file_name.csv`
 Với: `data_file_name` là tên file dữ liệu đầu vào
- Test case:
INPUT: `python missing_list.py house-prices.csv`
OUTPUT: số thuộc tính bị thiếu dữ liệu và liệt kê tên các thuộc tính đó

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python missing_list.py house-prices.csv
There are 18 missing attributes!
Alley, FireplaceQu, PoolQC, Fence, MiscFeature, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, LotFrontage, GarageType, GarageYrBlt, GarageFinish,
GarageQual, GarageCond, MasVnrArea
```

3.2 Đếm số dòng bị thiếu dữ liệu

- Cú pháp: `python count_missing_row.py data_file_name.csv`
 Với: `data_file_name` là tên file dữ liệu đầu vào
- Test case:
INPUT: `python count_missing_row.py house-prices.csv`
OUTPUT: số các dòng bị thiếu dữ liệu

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python count_missing_row.py house-prices.csv
There are 1000 rows which missing value
```

3.3 Điền giá trị bị thiếu

- Cú pháp: `python impute.py -i input_file.csv -m method_name -c col1 col2 ... -o output_file.csv`

Với: `input_file` là tên file dữ liệu đầu vào

`method_name` là phương pháp điền dữ liệu bị thiếu bao gồm mean, median cho thuộc tính numeric và mode cho thuộc tính categorical
`col1 col2 ...` là các cột cần điền dữ liệu, có thể chọn nhiều cột
`output_file` là tên tên dữ liệu đầu ra

- Test case 1:

INPUT: `python impute.py -i house-prices.csv -m mean -c LotFrontage -o output.csv`

OUTPUT:

Before

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMatl	Ex
2	1242	20 RL		83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vii	
3	1233	90 RL		70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc	
4	1401	50 RM		50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W	
5	1377	30 RL		52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W	
6	208	20 RL		12493	9844	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W	
7	1392	90 RL		65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Ph	
8	980	20 RL		80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vii	
9	484	120 RM		32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vii	
10	392	60 RL		71	12209	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Reg	Lvl	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vii		
11	730	30 RM		52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Mi	
12	255	20 RL		70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	Mi	
13	1094	20 RL		71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Mi	
14	1021	20 RL		60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vii	
15	1341	20 RL		70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Mi	
16	1025	20 RL		15498	9844	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	St	
17	848	20 RL		36	15523	Pave	IR1	Lvl	AllPub	Inside	Gtl	CollGCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc	
18	457	70 RM		34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As	
19	1266	160 IV		35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	Mi	
20	695	50 RM		51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	W	
21	24	120 RM		44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	Ce	
22	1314	60 RL		108	14774	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	Vii	
23	514	20 RL		71	9187	Pave	Reg	Bnk	AllPub	Inside	Gtl	Somerst	Norm	Norm	Twnhs	1Story	6	5	1983	1983	Gable	CompShg	Vii	
24	1068	60 RL		80	9760	Pave	Reg	Lvl	AllPub	Inside	Gtl	Mod	mes	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg	Hc	
25	1423	120 RM		37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollGCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompShg	Vii	
26	1258	30 RL		56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg	W	
27	620	60 RL		85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vii	
28	1213	30 RL		50	9340	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	6	1941	1950	Hip	CompShg	Mi	
29	71	20 RL		95	13651	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	7	6	1973	1973	Gable	CompShg	Ph	
30	732	80 RL		73	9590	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	SLvl	7	5	2003	2003	Gable	CompShg	Vii		
31	700	120 FV		59	4282	Pave	IR2	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	TwnhsE	1Story	7	5	2004	2004	Gable	CompShg	Mi	
32	532	70 RM		60	6155	Pave	IR1	Lvl	AllPub	FR3	Gtl	BrkSide	RRNn	Feedr	1Fam	2Story	6	8	1920	1999	Gable	CompShg	W	

After

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallStyl	OverallQua	OverallCor	YearBuild	YearRemo	RoofStyle	RoofMatl	Ex
2	0	1242	20 RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vii		
3	1	1233	90 RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc		
4	2	1401	50 RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W		
5	3	1377	30 RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W		
6	4	208	20 RL	12493	9844	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W		
7	5	1392	90 RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Ph		
8	6	980	20 RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vii		
9	7	484	120 RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vii		
10	8	392	60 RL	71	12209	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Reg	Lvl	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vii			
11	9	730	30 RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Mi		
12	10	255	20 RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	Mi		
13	11	1094	20 RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Mi		
14	12	1021	20 RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Cc		
15	13	1341	20 RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Cc		
16	14	1025	20 RL	15498	9844	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	W			
17	15	848	20 RL	36	15523	Pave	IR1	Lvl	AllPub	Inside	Gtl	CollGCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Cc		
18	16	457	70 RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	TwnhsE	2Story	5	5	1916	1950	Gable	CompShg	Cc	
19	17	1266	160 IV	35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	Cc		
20	18	695	50 RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	Cc		
21	19	24	120 RM	44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	8	5	1976	1976	Gable	CompShg	C		

- Test case 2:

INPUT: `python impute.py -i house-prices.csv -m median -c GarageYrBlt -o`

`output.csv`

OUTPUT:

Before

BH1	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	M
1	KitchenQu	TotRmsAb	Functional	Fireplaces	FireplaceQ	GarageTyp	GarageYrBlt	2007	Rfn	3	954	TA	TA	Y	0	56	0	0	0	0	0	0	0	0
2	Gd	7 Typ	0	Attchd						2	462	TA	TA	Y	0	0	0	0	0	0	0	0	0	0
3	TA	6 Typ	0	CarPort	1962	Unf				1	208	TA	TA	Y	0	0	112	0	0	0	0	0	0	0
4	TA	5 Typ	1 Gd	BuiltIn	1929	Rfn				1	160	FA	TA	Y	0	141	0	0	0	0	0	0	0	0
5	TA	4 Typ	0	Detchd	1925	Unf				1	312	TA	TA	Y	355	0	0	0	0	0	0	0	0	0
6	TA	6 Typ	1 Po	Attchd	1960	Rfn				3	792	TA	TA	Y	0	152	0	0	0	0	0	0	0	0
7	TA	8 Mod	0	Detchd	1967	Unf				2	480	TA	TA	Y	0	80	0	0	0	0	0	0	0	0
8	TA	5 Typ	0	Detchd	1963	Unf				2	402	TA	TA	Y	0	125	0	0	0	0	0	0	0	0
9	TA	5 Typ	0	Attchd	1998	Fin				2	560	TA	TA	Y	125	192	0	0	0	0	0	0	0	0
10	Gd	7 Typ	1 TA	BuiltIn	2001	Fin				2	539	TA	TA	Y	0	23	112	0	0	0	0	0	0	0
11	TA	5 Typ	0	Detchd	1962	Unf				1	294	TA	TA	Y	250	0	0	0	0	0	0	0	0	
12	TA	5 Typ	0	Attchd	1957	Rfn				2	884	TA	TA	Y	0	64	0	0	0	0	0	0	0	
13	Gd	6 Typ	0	Detchd	1977	Unf				2	451	TA	TA	Y	252	64	0	0	0	0	0	0	0	
14	Gd	5 Typ	0	Attchd	2005	Fin				4	480	TA	TA	Y	0	0	0	0	0	0	0	0	0	
15	TA	5 Typ	0	Detchd	1974	Unf				2	665	TA	TA	Y	0	72	174	0	0	0	0	0	0	
16	Gd	10 Typ	1 Gd	Attchd	1976	Fin				1	338	TA	TA	Y	120	0	158	0	0	0	0	0	0	
17	TA	5 Typ	1 Fa	Attchd	1972	Unf				2	442	TA	TA	Y	328	128	0	0	189	0	0	0	0	
18	TA	7 Typ	0	Detchd	1916	Unf				2	420	TA	TA	Y	140	0	0	0	0	0	0	0	0	
19	Gd	6 Typ	0	Detchd	1999	Fin				0	0	Y	0	34	0	0	0	0	0	0	0	0	0	
20	TA	5 Typ	0	Detchd	1995	Unf				3	576	TA	TA	Y	112	0	0	0	0	0	0	0	0	
21	TA	6 Typ	1 TA	Attchd	1976	Unf				2	572	TA	TA	Y	100	110	0	0	0	0	0	0	0	
22	Gd	10 Typ	1 TA	BuiltIn	1999	Fin				3	779	TA	TA	Y	668	30	0	0	0	0	0	0	0	
23	TA	5 Typ	0	Attchd	1983	Unf				2	484	TA	TA	Y	120	0	158	0	0	0	0	0	0	
24	TA	7 Typ	0	Attchd	1964	Rfn				2	451	TA	TA	Y	252	64	0	0	0	0	0	0	0	
25	Gd	3 Typ	0	Attchd	2003	Fin				4	480	TA	TA	Y	0	0	0	0	0	0	0	0	0	
26	TA	4 Typ	0	Detchd	1920	Fin				2	502	TA	FA	Y	0	0	84	0	0	0	0	0	0	

After

BI1	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	M	
1	BedroomA	KitchenAbi	KitchenQu	TotRmsAb	Functional	Fireplaces	FireplaceQ	GarageTyp	GarageYrBlt	2007	Rfn	3	954	TA	TA	Y	0	56	0	0	0	0	0	0	
2	3	1 Gd	7 Typ	0	Attchd					2	462	TA	TA	Y	0	0	0	0	0	0	0	0	0	0	
3	2	2 TA	6 Typ	0	CarPort	1962	Unf			1	208	TA	TA	Y	0	0	112	0	0	0	0	0	0	0	
4	3	1 TA	5 Typ	1 Gd	BuiltIn	1929	Rfn			1	160	FA	TA	Y	0	141	0	0	0	0	0	0	0	0	
5	2	1 TA	4 Typ	0	Detchd	1925	Unf			1	312	TA	TA	Y	355	0	0	0	0	0	0	0	0	0	
6	3	1 TA	6 Typ	1 Po	Attchd	1960	Rfn			3	792	TA	TA	Y	0	152	0	0	0	0	0	0	0	0	
7	4	2 TA	8 Mod	0	Detchd	1967	Unf			2	480	TA	TA	Y	0	80	0	0	0	0	0	0	0	0	
8	3	1 TA	5 Typ	0	Detchd	1963	Unf			2	402	TA	TA	Y	0	125	0	0	0	0	0	0	0	0	
9	2	1 TA	5 Typ	0	Attchd	1998	Fin			3	294	TA	TA	Y	250	0	0	0	0	0	0	0	0	0	
10	3	1 Gd	7 Typ	1 TA	BuiltIn	2001	Fin			2	560	TA	TA	Y	125	192	0	0	0	0	0	0	0	0	
11	2	1 TA	5 Typ	0	Detchd	1962	Unf			2	539	TA	TA	Y	0	23	112	0	0	0	0	0	0	0	
12	3	1 TA	5 Typ	0	Attchd	1957	Rfn			1	294	TA	TA	Y	250	0	0	0	0	0	0	0	0	0	
13	1	1 Gd	6 Typ	0	Detchd	1977	Unf			2	884	TA	TA	Y	0	64	0	0	0	0	0	0	0	0	
14	2	1 Gd	5 Typ	0	Attchd	2005	Fin			2	451	TA	TA	Y	252	64	0	0	0	0	0	0	0	0	
15	3	1 TA	5 Typ	0	Detchd	1974	Unf			4	480	TA	TA	Y	0	0	0	0	0	0	0	0	0		
16	2	1 Gd	10 Typ	1 Gd	Attchd	1976	Fin			2	665	TA	TA	Y	0	72	174	0	0	0	0	0	0	0	
17	3	1 TA	5 Typ	1 Fa	Attchd	1972	Unf			1	338	TA	TA	Y	0	0	0	0	0	0	0	0	0		
18	4	1 TA	7 Typ	0	Detchd	1916	Unf			3	513	FA	TA	Y	0	0	96	0	0	0	0	0	0	0	
19	3	1 Gd	6 Typ	0	Detchd	1999	Fin			2	506	TA	TA	Y	0	34	0	0	0	0	0	0	0	0	
20	3	1 TA	5 Typ	0	Detchd	1995	Fin			2	576	TA	TA	Y	112	0	0	0	0	0	0	0	0	0	
21	3	1 TA	6 Typ	1 TA	Attchd	1976	Unf			2	572	TA	TA	Y	100	110	0	0	0	0	0	0	0	0	
22	4	1 Gd	10 Typ	1 TA	BuiltIn	1999	Fin			3	779	TA	TA	Y	668	30	0	0	0	0	0	0	0	0	
23	3	1 TA	5 Typ	0	Attchd	1983	Unf			2	484	TA	TA	Y	120	0	158	0	0	0	0	0	0	0	
24	4	1 TA	7 Typ	0	Attchd	1964	Rfn			2	442	TA	TA	Y	328	128	0	0	189	0	0	0	0	0	
25	1	1 Gd	3 Typ	0	Attchd	2003	Fin			2	420	TA	TA	Y	140	0	0	0	0	0	0	0	0	0	
26	2	1 TA	4 Typ	0	Detchd	1979.5				0	0	Y	0	0	96	0	0	0	0	0	0	0	0	0	
27	4	1 Gd	10 Typ	2 Gd	Attchd	2003	Fin			3	749	TA	TA	Y	168	0	0	0	0	0	0	0	0	0	
28	2	1 TA	4 Typ	0	Attchd	1941	Unf			1	234	TA	N	0	113	0	0	0	0	0	0	0	0	0	
29	3	1 TA	8 Typ	2 Gd	Attchd	1973	Fin			2	516	TA	TA	Y	300	0	0	0	0	0	0	0	0	0	
30	3	1 Gd	6 Typ	1 Gd	Attchd	2003	Fin			2	438	TA	TA	Y	160	22	0	0	0	0	0	0	0	0	
31	2	1 Gd	5 Typ	0	Attchd	2004	Rfn			2	530	TA	TA	Y	156	158	0	0	0	0	0	0	0	0	
32	3	1 TA	6 Typ	0	Detchd	1920	Fin			2	502	TA	FA	Y	0	0	84	0	0	0	0	0	0	0	0

- Test case 3:

INPUT: *python impute.py -i house-prices.csv -m mode -c Alley -o output.csv*

OUTPUT:

Before

house-prices.csv																										
Lưu tự động	Trang đầu	Chèn	Bố trí Trang	Công thức	Dữ liệu	Xem lại	Xem	Trợ giúp	Tim kiếm	Ngọc Tu	N	T	Chia sẻ	Chú thích	x											
CÓ THỂ MẤT DỮ LIỆU Có thể mất một số tính năng nếu bạn lưu số làm việc này ở định dạng phân tách bằng dấu phẩy (.csv). Để duy trì các tính năng này, hãy lưu số làm việc ở định dạng tệp Excel.																										
1		MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	Rooftyle	RoofMat	Ex		
2	1242	20	RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vii			
3	1233	90	RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc			
4	1401	50	RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W			
5	1377	30	RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W			
6	208	20	RL		12493	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W			
7	1392	90	RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Pl			
8	980	20	RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vii			
9	484	120	RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vii			
10	392	60	RL	71	12209	Pave	IR1	Lvl	AllPub	CulDsac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vii			
11	730	30	RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	M		
12	255	20	RL	70	8400	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M			
13	1094	20	RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	M				
14	1021	20	RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vii				
15	1341	20	RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	M				
16	1025	20	RL		15498	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	St			
17	848	20	RL	36	15523	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc			
18	457	70	RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As			
19	1266	160	FV	35	3735	Pave	Grvl	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	1Fam	2Story	7	5	1999	1999	Gable	CompShg	M			
20	695	50	RM	51	6120	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	W			
21	24	120	RM	44	4224	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	1Fam	1Story	5	7	1976	1976	Gable	CompShg	Ce			
22	1314	60	RL	108	14774	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	NoRidge	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	Vii			
23	514	20	RL	71	9187	Pave	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg	Vii				
24	1068	60	RL	80	9760	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg	Hc				
25	1423	120	RM	37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	1Fam	1Story	6	5	2003	2003	Gable	CompShg	Vii				
26	1258	30	RL	56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg	W			
27	620	60	RL	85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vii				
28	213	30	RL	50	9340	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	6	1941	1950	Hip	CompShg	M				
29	71	20	RL	95	13651	Pave	Reg	Lvl	AllPub	Inside	Gtl	SLvl	Norm	1Fam	1Story	7	6	2004	2004	Gable	CompShg	M				
30	732	80	RL	73	9590	Pave	Grvl	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	1Fam	1Story	7	5	2003	2003	Gable	CompShg	Cc			
31	29	700	120	FV	59	4282	Pave	IR2	Lvl	AllPub	Inside	Gtl	Somerst	Norm	1Fam	1Story	7	5	2004	2004	Gable	CompShg	Cc			
32	30	532	70	RM	60	6155	Pave	IR1	Lvl	AllPub	FR3	Gtl	BrkSide	RRNn	Feedr	1Fam	2Story	6	8	1920	1999	Gable	CompShg	W		

After

output.csv																										
Lưu tự động	Trang đầu	Chèn	Bố trí Trang	Công thức	Dữ liệu	Xem lại	Xem	Trợ giúp	Tim kiếm	Ngọc Tu	N	T	Chia sẻ	Chú thích	x											
CÓ THỂ MẤT DỮ LIỆU Có thể mất một số tính năng nếu bạn lưu số làm việc này ở định dạng phân tách bằng dấu phẩy (.csv). Để duy trì các tính năng này, hãy lưu số làm việc ở định dạng tệp Excel.																										
1		MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	Rooftyle	RoofMat	Ex		
2	0	1242	20	RL	83	9849	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vii	
3	1	1233	90	RL	70	9842	Pave	Grvl	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc		
4	2	1401	50	RM	50	6000	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W	
5	3	1377	30	RL	52	6292	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W	
6	4	208	20	RL		12493	Pave	Grvl	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W		
7	5	1392	90	RL	65	8944	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Vii		
8	6	980	20	RL	80	8816	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vii	
9	7	484	120	RM	32	4500	Pave	Grvl	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	W	
10	8	392	60	RL	71	12209	Pave	Grvl	IR1	Lvl	AllPub	CulDsac	Gtl	Mitchel	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Cc		
11	9	730	30	RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	Cc	
12	10	255	20	RL	70	8400	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M		
13	11	1094	20	RL	71	9230	Pave	Grvl	Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	Cc		
14	12	1021	20	RL	60	7024	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Cc		
15	13	1341	20	RL	70	8294	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	Cc		
16	14	1025	20	RL		15498	Pave	Grvl	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	1Fam	1Story	8	6	1976	1976	Hip	W			
17	15	848	20	RL	36	15523	Pave	Grvl	IR1	Lvl	AllPub	FR3	Gtl	Mitchel	Norm	1Fam	2Story	5	6	1972	1972	Gable	CompShg	Cc		
18	16	457	70	RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	Cc		
19	17	1266	160	FV	35	3735	Pave	Grvl	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	1Fam	2Story	7	5	1999	1999	H				

- Test case 4:

INPUT: *python impute.py -i house-prices.csv -m mode -c LotFrontage -o output.csv*

OUTPUT:

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python impute.py -i house-prices.csv -m mode -c LotFrontage -o output.csv
compute mode
Invalid column: LotFrontage
```

3.4 Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước

- Cú pháp: *python del_missing_rows.py -i input_file.csv -m 0.x -o output_file.csv*

Với: *input_file* là tên file dữ liệu đầu vào

0.x là tỉ lệ mất dữ liệu tối đa cho phép

output_file là tên tên dữ liệu đầu ra

- Test case 1:

INPUT: *python del_missing_rows.py -i house-prices.csv -m 0.1 -o output.csv*

OUTPUT: Số hàng dữ liệu trước khi xóa là 1000, sau khi xóa số hàng còn lại là 920

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python del_missing_rows.py -i house-prices.csv -m 0.1 -o output.csv
There are 1000 rows before deleting
There are 920 rows after deleting
```

- Test case 2:

INPUT: *python del_missing_rows.py -i house-prices.csv -m 0.12 -o output.csv*

OUTPUT: Số hàng dữ liệu trước khi xóa là 1000, sau khi xóa số hàng còn lại là 923

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python del_missing_rows.py -i house-prices.csv -m 0.12 -o output.csv
There are 1000 rows before deleting
There are 923 rows after deleting
```

- Test case 3:

INPUT: *python del_missing_rows.py -i house-prices.csv -m 0.2 -o output.csv*

OUTPUT: Số hàng dữ liệu trước khi xóa là 1000, sau khi xóa số hàng còn lại là 1000

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python del_missing_rows.py -i house-prices.csv -m 0.2 -o output.csv
There are 1000 rows before deleting
There are 1000 rows after deleting
```

- Test case 4:

INPUT: *python del_missing_rows.py -i house-prices.csv -m 0.3 -o output.csv*

OUTPUT: Số hàng dữ liệu trước khi xóa là 1000, sau khi xóa số hàng còn lại là 1000

```
D:\Năm 3\KTDL&UD\LAB1\Data-Mining\Lab1-Preprocessing\sources>python del_missing_rows.py -i house-prices.csv -m 0.3 -o output.csv
There are 1000 rows before deleting
There are 1000 rows after deleting
```

3.5 Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước

- Cú pháp: *python del_missing_cols.py -i input_file.csv -m 0.x -o output_file.csv*

Với: *input_file* là tên file dữ liệu đầu vào

0.x là tỉ lệ mất dữ liệu tối đa cho phép

output_file là tên tên dữ liệu đầu ra

MSSubClass	0.0	RoofStyle	0.0	CentralAir	0.0	GarageCars	0.0
MSZoning.....	0.0	RoofMatl.....	0.0	Electrical.....	0.0	GarageArea	0.0
LotFrontage	0.173	Exterior1st.....	0.0	1stFlrSF	0.0	GarageQual	0.06
LotArea	0.0	Exterior2nd.....	0.0	2ndFlrSF	0.0	GarageCond	0.06
Street.....	0.0	MasVnrType	0.593	LowQualFinSF	0.0	PavedDrive	0.0
Alley.....	0.941	MasVnrArea	0.01	GrLivArea	0.0	WoodDeckSF	0.0
LotShape.....	0.0	ExterQual.....	0.0	BsmtFullBath	0.0	OpenPorchSF	0.0
LandContour.....	0.0	ExterCond.....	0.0	BsmtHalfBath	0.0	EnclosedPorch	0.0
Utilities	0.0	Foundation	0.0	FullBath	0.0	3SsnPorch	0.0
LotConfig	0.0	BsmtQual	0.027	HalfBath	0.0	ScreenPorch	0.0
LandSlope	0.0	BsmtCond0.....	0.027	BedroomAbvGr	0.0	PoolArea	0.0
Neighborhood	0.0	BsmtExposure.....	0.028	KitchenAbvGr	0.0	PoolQC	1.0
Condition1	0.0	BsmtFinType1.....	0.027	KitchenQual	0.0	Fence	0.815
Condition2	0.0	BsmtFinSF1	0.0	TotRmsAbvGrd	0.0	MiscFeature	0.963
BldgType.....	0.0	BsmtFinType2.....	0.029	Functional	0.0	MiscVal	0.0
HouseStyle.....	0.0	BsmtFinSF2	0.0	Fireplaces	0.0	MoSold	0.0
OverallQual.....	0.0	BsmtUnfSF	0.0	FireplaceQu	0.501	YrSold	0.0
OverallCond.....	0.0	TotalBsmtSF	0.0	GarageType	0.06	SaleType	0.0
YearBuilt	0.0	Heating	0.0	GarageYrBlt	0.06	SaleCondition	0.0
YearRemodAdd	0.0	HeatingQC	0.0	GarageFinish	0.06	SalePrice	0.0

Bảng thống kê tần số xuất hiện của các thuộc tính

- Test case 1:

INPUT: *python del_missing_cols.py -i house-prices.csv -m 0.9 -o output.csv*

OUTPUT: Những thuộc tính bị xóa

- Alley: 0.941
- PoolQC: 1.0
- MiscFeature: 0.963

- Test case 2:

INPUT: *python del_missing_cols.py -i house-prices.csv -m 0.8 -o output.csv*

OUTPUT:

- Alley: 0.941
- PoolQC: 1.0
- Fence: 0.815
- MiscFeature: 0.963

- Test case 3:

INPUT: *python del_missing_cols.py -i house-prices.csv -m 0.7 -o output.csv*

OUTPUT:

- Alley: 0.941
- PoolQC: 1.0
- Fence: 0.815
- MiscFeature: 0.963

- Test case 4:

INPUT: *python del_missing_cols.py -i house-prices.csv -m 0.6 -o output.csv*
OUTPUT:

- *Alley:* 0.941
- *PoolQC:* 1.0
- *Fence:* 0.815
- *MiscFeature:* 0.963

- Test case 5:

INPUT: *python del_missing_cols.py -i house-prices.csv -m 0.5 -o output.csv*
OUTPUT:

- *Alley:* 0.941
- *MasVnrType:* 0.593
- *FireplaceQu:* 0.501
- *PoolQC:* 1.0
- *Fence:* 0.815
- *MiscFeature:* 0.963

3.6 Xóa các mẫu bị trùng lặp

- Cú pháp: *python del_duplicate_rows.py -i input_file.csv -o output_file.csv*

Với: *input_file* là tên file dữ liệu đầu vào

output_file là tên file dữ liệu đầu ra

- Test case:

INPUT: *python del_duplicate_rows.py -i house-prices.csv -o output.csv*

OUTPUT: Có 284 dòng bị trùng. Một số dòng trùng nhau như:

- 396 253
- 399 166
- 400 201
- 407 92
- 415 414
- 432 5
- 444 74
- 445 334
- 446 420
- 450 403
- 452 344
- 453 306
- 454 89
- 457 361
- 458 354
- 460 302
- 467 50
- 469 391
- 471 411
- 483 110

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

- Cú pháp: `python data_normalization.py -i input_file.csv -m method -c col1 col2 ... -o output_file.csv`

Với: `input_file` là tên của file dữ liệu đầu vào.

`method` là tên phương pháp chuẩn hóa (“min-max” hoặc “z-score”).

`col1 col2 ...` là tên các thuộc tính cần chuẩn hóa.

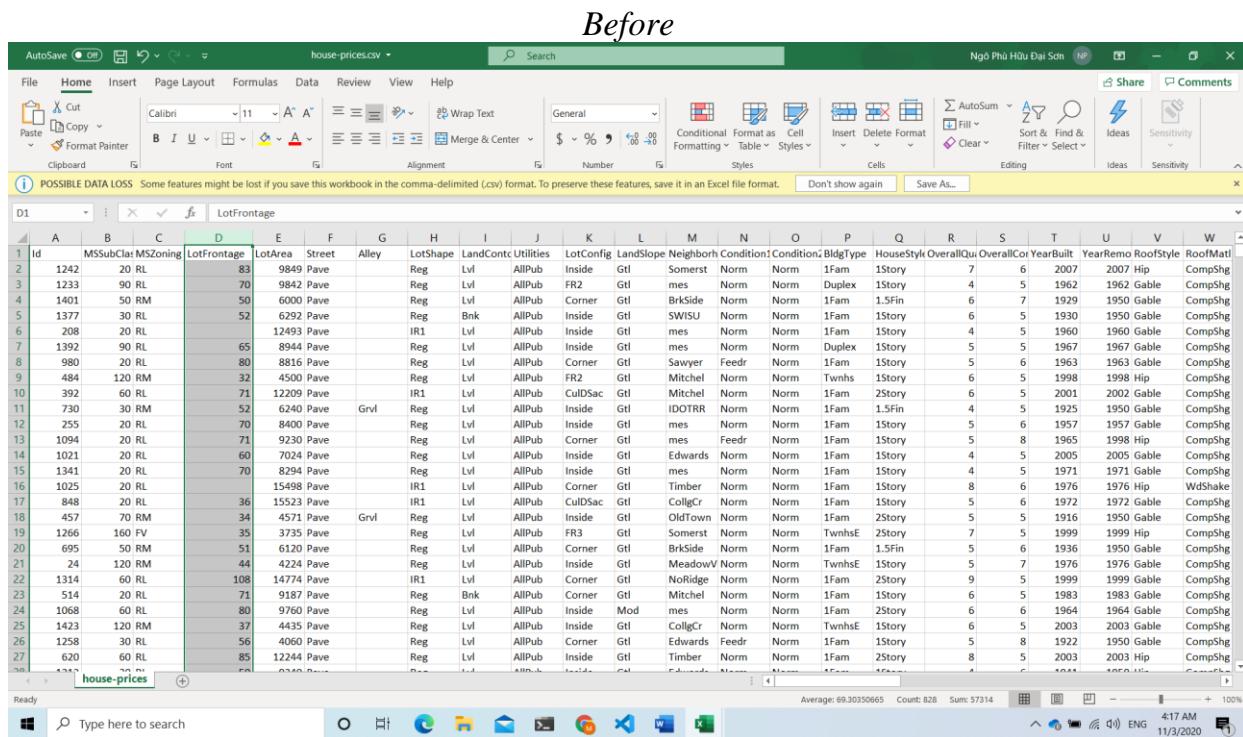
`output_file` là tên file dữ liệu đầu ra.

- Test case 1:

INPUT: `python data_normalization.py -i house-prices.csv -m min-max -c LotFrontage -o output.csv`

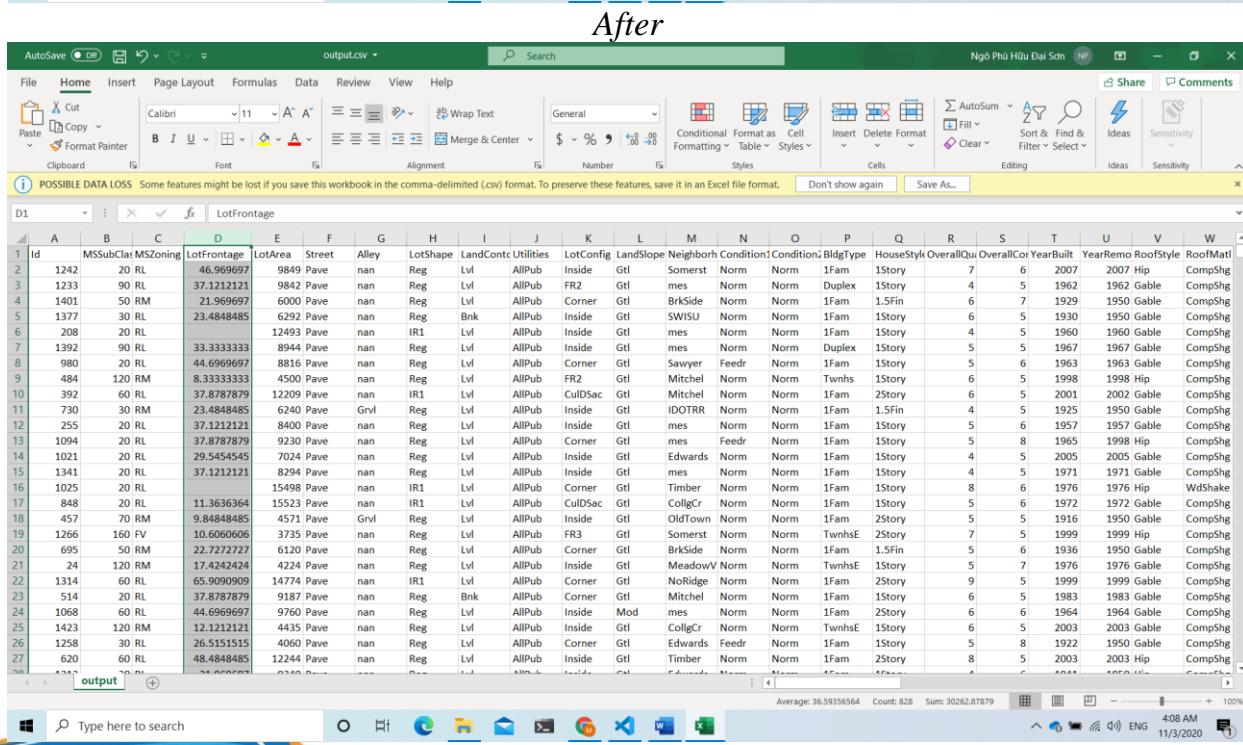
OUTPUT:

Before



Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RooStyle	RoofMatl
2	1242	20 RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	
3	1233	90 RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	
4	1401	50 RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	
5	1377	30 RL	52	6292	Pave	Reg	Brk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	
6	208	20 RL	12493	9844	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	
7	1392	90 RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	
8	98	20 RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	
9	484	120 RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	
10	392	60 RL	71	12209	Pave	IR1	Lvl	AllPub	CulSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	
11	730	30 RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg
12	255	20 RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	
13	1094	20 RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	
14	1021	20 RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	
15	1341	20 RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	
16	1025	20 RL	15498	9844	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	
17	848	20 RL	36	15523	Pave	IR1	Lvl	AllPub	CulSac	Gtl	CollCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	
18	457	70 RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg
19	1266	160 FV	35	3735	Pave	Reg	Lvl	AllPub	FR2	Gtl	Somerst	Norm	Norm	Twnhs	2Story	7	5	1999	1999	Hip	CompShg	
20	695	50 RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	
21	24	120 RM	44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	
22	1314	60 RL	108	14774	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	
23	514	20 RL	71	9187	Pave	Reg	Brk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg	
24	1068	60 RL	80	9760	Pave	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg	
25	1423	120 RM	37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompShg	
26	1258	30 RL	56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg	
27	620	60 RL	85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	
28	125	20 RL	60	6240	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1950	1950	Hip	CompShg	

After



Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RooStyle	RoofMatl
2	1242	20 RL	46.996997	9849	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg
3	1233	90 RL	37.1212121	9842	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg
4	1401	50 RM	21.996997	6000	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg
5	1377	30 RL	23.4848485	6292	Pave	nan	Reg	Brk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg
6	208	20 RL	12493	9844	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	
7	1392	90 RL	33.3333333	8944	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg
8	980	20 RL	44.696997	8816	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg
9	484	120 RM	8.3333333	4500	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg
10	392	60 RL	37.8787879	12209	Pave	nan	IR1	Lvl	AllPub	CulDsc	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg
11	730	30 RM	23.4848485	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1950	1950	Gable	CompShg
12	255	20 RL	37.1212121	8400	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg
13	1094	20 RL	37.8787879	9230	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg
14	1021	20 RL	29.5454545	7024	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg
15	1341	20 RL	37.1212121	8294	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg
16	1025	20 RL	15498	9844	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	
17	848	20 RL	11.3636364	15523	Pave	nan	IR1	Lvl	AllPub	CulDsc	Gtl	CollCr	Norm	Norm	1Fam	2Story	5	6	1972	1972	Gable	CompShg
18	457	70 RM	9.84848485	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	Twnhs	2Story	7	5	1916	1950	Gable	CompShg
19	1266	160 FV	10.6060606	3735	Pave	nan	Reg	Lvl	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	
20	695	50 RM	22.7272727	6120	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg
21	24	120 RM	17.4242424	4224	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	Twnhs	1Story	5	7	1976	1976	Gable	CompShg
22	1314	60 RL	65.9090909	14774	Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg
23	514	20 RL	37.8787879	9187	Pave	nan	Reg	Brk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg
24	1068	60 RL	44.696997	9760	Pave	nan	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg
25	1423	120 RM	12.1212121	4435	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompShg
26	1258	30 RL	26.5151515	4060	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg
27	620	60 RL	48.4848485	12244	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg



- Test case 2:

INPUT: `python data_normalization.py -i house-prices.csv -m min-max -c`

`MSSubClass MSZoning LotFrontage LotArea Street` -o output.csv

OUTPUT:

Before

house-prices.csv		house-prices.csv																									
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help	Calibri	11	A	A	Wrap Text	General	\$ % , .	00.00	Conditional	Format as	Cell	Styles	Insert	Delete	Format	AutoSum	Fill	Sort & Filter	Select
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street																					
2	1242	20 RL	83	9849	Pave		Alley	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	Hip	CompShg	Vrl				
3	1233	90	70	9842	Pave			Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc			
4	1401	50 RM	50	6000	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W			
5	1377	30 RL	52	6292	Pave			Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W			
6	208	20 RL		12493	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W			
7	1392	90 RL	65	8944	Pave			Reg	Lvl	AllPub	Inside	Gtl	Swyer	Feedr	Norm	1Fam	1Story	5	5	1967	1967	Gable	CompShg	Prl			
8	980	20 RL	80	8816	Pave			Reg	Lvl	AllPub	Corner	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1999	Hip	CompShg	Vrl			
9	484	120 RM	32	4500	Pave			Reg	Lvl	AllPub	FR2	Gtl	CollGr	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vrl			
10	392	60 RL	71	12209	Pave			IR1	Lvl	AllPub	CulDsc	Gtl	Mitchel	Norm	Norm	1Fam	2Story	5	6	1971	1971	Gable	CompShg	M			
11	730	30 RM	52	6240	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1925	1950	Gable	CompShg	M			
12	255	20 RL	70	8400	Pave			Reg	Lvl	AllPub	Corner	Gtl	Edwards	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M			
13	1094	20 RL	71	9230	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	M			
14	1021	20 RL	60	7024	Pave			Reg	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vrl			
15	1341	20 RL	70	8294	Pave			IR1	Lvl	AllPub	Corner	Gtl	OldTown	Norm	Norm	TwnhsE	2Story	8	6	1976	1976	Hip	WdShake	St			
16	1025	20 RL		15498	Pave			Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	M			
17	848	20 RL	36	15523	Pave			IR1	Lvl	AllPub	CulDsc	Gtl	CollGr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc			
18	457	70 RM	34	4571	Pave			Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	As			
19	1266	160 FV	35	3735	Pave			Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	M			
20	695	50 RM	51	6120	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1Story	5	6	1993	1993	Hip	CompShg	W			
21	24	120 RM	44	4224	Pave			Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	Ge			
22	1314	60 RL	108	14774	Pave			IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	Vrl			
23	514	20 RL	71	9187	Pave			Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg	Vrl			
24	1068	60 RL	80	9760	Pave			Reg	Lvl	AllPub	Inside	Gtl	CollGr	Norm	Norm	TwnhsE	1Story	6	6	1964	1964	Gable	CompShg	Hc			
25	1423	120 RM	37	4435	Pave			Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	2003	2003	Gable	CompShg	Vrl			
26	1258	30 RL	56	4060	Pave			Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vrl			
27	620	60 RL	85	12244	Pave			Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vrl			

After

output.csv		output.csv																									
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help	Calibri	11	A	A	Wrap Text	General	\$ % , .	00.00	Conditional	Format as	Cell	Styles	Insert	Delete	Format	AutoSum	Fill	Sort & Filter	Select
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street		Alley	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	Hip	CompShg	Vrl			
2	1242	100 RL	46.9697	203.9164	Pave			Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc			
3	1233	141.1765 RM	37.12121	203.9131	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W			
4	1401	117.6471 RM	21.9697	202.1158	Pave			Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W			
5	1377	105.8824 RL	23.48485	202.2524	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W			
6	208	100 RL		205.1533	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1967	1967	Gable	CompShg	Prl			
7	1392	141.1765 RL	33.33333	203.4933	Pave			Reg	Lvl	AllPub	Inside	Gtl	Swyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Hip	CompShg	Vrl			
8	980	100 RL	44.69697	203.4332	Pave			Reg	Lvl	AllPub	Corner	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1999	Hip	CompShg	Vrl			
9	484	158.8235 RM	8.333333	201.4141	Pave			Reg	Lvl	AllPub	FR2	Gtl	CollGr	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vrl			
10	392	123.5294 RL	37.87879	205.0204	Pave			IR1	Lvl	AllPub	CulDsc	Gtl	Mitchel	Norm	Norm	1Fam	2Story	5	6	1972	1972	Gable	CompShg	Hc			
11	730	105.8824 RM	23.48485	202.2281	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1925	1950	Gable	CompShg	M			
12	255	100 RL	37.12121	203.2386	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M			
13	1094	100 RL	37.87879	203.6268	Pave			Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	M			
14	1021	100 RL	29.54545	202.5949	Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vrl			
15	1341	100 RL	37.12121	203.189	Pave			Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	WdShake	St			
16	1025	100 RL	206.559	nan	Pave			IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	CompShg	Hc			
17	848	100 RL	11.36364	206.5707	Pave			Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	As			
18	457	120.4118 RM	9.848485	201.4474	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable</td					



- Test case 3:

INPUT: `python data_normalization.py -i house-prices.csv -m z-score -c`

`LotFrontage -o output.csv`

OUTPUT:

Before

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMat
2	1242	20 RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompSg	
3	1233	90 RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompSg	
4	1401	50 RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompSg	
5	1377	30 RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompSg	
6	208	20 RL		12493	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompSg	
7	1392	90 RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	sawyer	Feedr	Norm	1Fam	1Story	5	5	1967	1967	Gable	CompSg	
8	980	20 RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompSg		
9	484	120 RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompSg	
10	392	60 RL	71	12029	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompSg	
11	730	30 RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompSg
12	255	20 RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompSg	
13	1094	20 RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompSg	
14	1021	20 RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompSg	
15	1341	20 RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompSg	
16	1025	20 RL		15498	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	
17	848	20 RL	36	15523	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	CollCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompSg	
18	457	70 RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompSg
19	1266	160 FV	35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompSg	
20	695	50 RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompSg	
21	24	120 RM	44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompSg	
22	1314	60 RL	108	14774	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompSg	
23	514	20 RL	71	9187	Pave	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompSg	
24	1068	60 RL	80	9760	Pave	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompSg	
25	1423	120 RM	37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompSg	
26	1258	30 RL	56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompSg	
27	620	60 RL	85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompSg	

After

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMat
2	1242	20 RL	2.186525522	9849	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompSg
3	1233	90 RL	0.111189079	9842	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompSg
4	1401	50 RM	-3.081636218	6000	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompSg
5	1377	30 RL	-2.762353688	6292	Pave	nan	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompSg
6	208	20 RL		12493	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompSg	
7	1392	90 RL	-0.687017245	8944	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompSg
8	980	20 RL	1.707601728	8816	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompSg
9	484	120 RM	-5.955178985	4500	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompSg
10	392	60 RL	0.270830344	12029	Pave	nan	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompSg
11	730	30 RM	-2.762353688	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompSg
12	255	20 RL	0.111189079	8400	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompSg
13	1094	20 RL	0.270830344	9230	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompSg
14	1021	20 RL	-1.485223569	7024	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompSg
15	1341	20 RL	0.111189079	8294	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompSg
16	1025	20 RL		15498	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	
17	848	20 RL	-5.316613926	15523	Pave	nan	IR1	Lvl	AllPub	CulDSac	Gtl	CollCr	Norm	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompSg
18	457	70 RM	-5.635896456	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompSg
19	1266	160 FV	-5.476255191	3735	Pave	nan	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompSg
20	695	50 RM	-2.921994953	6120	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompSg
21	24	120 RM	-4.039483807	4224	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompSg
22	1314	60 RL	6.177557144	14774	Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompSg
23	514	20 RL	0.270830344	9187	Pave	nan	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompSg
24	1068	60 RL	1.707601728	9760	Pave	nan	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompSg
25	1423	120 RM	-5.156972661	4435	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompSg
26	1258	30 RL	-2.123788629	4060	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompSg
27	620	60 RL	2.505808052	12244	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompSg

- Test case 4:

INPUT: `python data_normalization.py -i house_prices.csv -m z-score -c`

MSSubClass MSZoning LotFrontage LotArea Street -o output.csv

OUTPUT:

Before

The screenshot shows a Microsoft Excel spreadsheet titled "house-prices.csv". The data consists of 27 rows and 20 columns. The columns represent various features of houses, such as ID, street type, zoning, lot area, condition, and roof style. The data is presented in a tabular format with some cells containing numerical values and others containing descriptive text like "Pave" or "Grlv". The Excel interface includes a ribbon at the top with tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Help. There are also standard toolbar icons for cutting, pasting, and formatting. A status bar at the bottom provides information about the average value (3652.475416), count (4832), and sum (10325548) of the data.

After

AutoSave OFF

output.csv Search

File Home Insert Page Layout Formulas Data Review View Help

Cut

Font Alignment Number Styles Cells Editing

Wrap Text Merge & Center

General \$ % , .

Conditional Format as Table Styles Insert Delete Format

AutoSum Fill Sort & Filter Ideas

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

B1 MSSubClass MSSZoning LotFrontage LotArea Street Alley LotShape LandCont: Utilities LotConfig Landslope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCor YearBuilt YearRemo RoofStyle RoofMat Ex

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Id	MSSubClass	MSSZoning	LotFrontaj	LotArea	Street	Alley	LotShape	LandCont:	Utilities	LotConfig	Landslope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCor	YearBuilt	YearRemo	Rooftyle	RoofMat	Ex
2	1242	-0.27198 RL	2.186526	-0.2568	-0.2568	Street	nan	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vil
3	1233	0.251777 RL	0.111189	-0.26175	Pave		nan	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Hc
4	1401	-0.04751 RM	-3.08164	2.98056	Pave		nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1.5Fin	1Story	6	7	1929	1950	Gable	CompShg	W
5	1377	-0.19716 RL	-2.76235	-2.77393	Pave		nan	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W
6	208	-0.27198 RL	1.614245	Pave			nan	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W
7	1392	0.251777 RL	-0.68702	-0.89732	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Ph
8	980	-0.27198 RL	1.707602	-0.98781	Pave		nan	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vil
9	484	0.476244 RM	-5.95518	-0.40205	Pave		nan	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg	Vil
10	392	0.02731 RM	0.27083	1.413271	Pave		nan	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vil
11	730	-0.19716 RM	-2.76235	-2.81073	Pave	Grlv	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1.5Fin	1Story	4	5	1925	1950	Gable	CompShg	M	
12	255	-0.27198 RL	0.111189	-0.182819	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M
13	1094	-0.27198 RL	0.27083	0.69484	Pave		nan	Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	M	
14	1021	-0.27198 RL	-1.48522	-2.25593	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vil	
15	1341	-0.27198 RL	0.111189	-1.3572	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	M
16	1025	-0.27198 RL	3.740752	Pave			nan	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976	Hip	WDShake	As
17	848	-0.27198 RL	-5.13661	3.758443	Pave		nan	IR1	Lvl	AllPub	CulDSac	Gtl	ColligCr	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Hc	
18	457	0.102132 RM	-5.6359	-3.99181	Pave	Grlv	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As	
19	1266	0.775533 FV	-5.47626	-4.58341	Pave		nan	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg	M
20	695	-0.04751 RM	-2.92199	-2.89565	Pave		nan	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1.5Fin	1Story	5	6	1936	1950	Gable	CompShg	W
21	24	0.476244 RM	-4.03948	-4.23736	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg	Cel
22	1314	0.02731 RM	6.177557	3.228409	Pave		nan	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg	Vil
23	514	-0.27198 RL	0.27083	-0.75257	Pave		nan	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg	Vil
24	1068	0.02731 RM	1.707602	-0.31978	Pave		nan	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg	Hc
25	1423	0.476244 RM	-5.15697	-4.08805	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	ColligCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompShg	Vil
26	1258	-0.19716 RL	-2.12379	-4.35342	Pave		nan	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg	W
27	620	0.02731 RM	2.505808	1.438039	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vil
28	1321	0.27198 RL	-0.20164	0.61609	Pave		nan	Reg	Lvl	AllPub	Inside	Gtl	EdwardV	Norm	Norm	1Fam	2Story	6	5	1954	1954	Gable	CompShg	Vil

3.8 Tính giá trị biểu thức thuộc tính

- Cú pháp: `python expression_calculate.py -i input_file.csv -o output_file.csv`
Với: `input_file` là tên của file dữ liệu đầu vào.
`output_file` là tên file dữ liệu đầu ra.
Sau đó nhập biểu thức cần tính toán và tên của cột mới.
 - Test case 1:

- Test case 1:

Sau đó nhập biểu thức cần tính toán và tên của cột mới.
Tuy nhiên

- Test case 1:

INPUT: $LotFrontage * 40$

OUTPUT:

Before

AutoSave		house-prices.csv														Search													
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help															Share		Comments				
Cut		Copy		Format Painter		Font		Text		Merge & Center		Number		Formatting		Table		Cell Styles		AutoSum		Fill		Sort & Filter		Ideas		Sensitivity	
POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.																													
D1	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			
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMatl	ExterQual	ExterCond	ExterPct	ExterType		
2	1242	20	RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
3	1233	90	RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg	Ht	Ex	Ex	Ex	Ex	Ex	
4	1401	50	RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg	W	Ex	Ex	Ex	Ex	Ex		
5	1377	30	RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg	W	Ex	Ex	Ex	Ex	Ex		
6	208	20	RL	12493	9844	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg	W	Ex	Ex	Ex	Ex	Ex		
7	1392	90	RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg	Pl	Ex	Ex	Ex	Ex	Ex		
8	980	20	RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963	Gable	CompShg	Vir	Ex	Ex	Ex	Ex	Ex	
9	484	120	RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Twrhs	1Story	6	5	1998	1998	Hip	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
10	392	60	RL	71	12209	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
11	730	30	RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg	M	Ex	Ex	Ex	Ex	Ex	
12	255	20	RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg	M	Ex	Ex	Ex	Ex	Ex		
13	1094	20	RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg	M	Ex	Ex	Ex	Ex	Ex	
14	1021	20	RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
15	1341	20	RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	M	Ex	Ex	Ex	Ex	Ex		
16	1025	20	RL	15499	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake	Ex	Ex	Ex	Ex	Ex	Ex			
17	848	20	RL	36	15523	Pave	IR1	Lvl	AllPub	CulDsa	Gtl	CollGr	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg	Ht	Ex	Ex	Ex	Ex	Ex		
18	457	70	RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg	As	Ex	Ex	Ex	Ex	Ex	
19	1266	160	FV	35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	TwrhsE	2Story	7	5	1999	1999	Hip	CompShg	M	Ex	Ex	Ex	Ex	Ex		
20	695	50	RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg	W	Ex	Ex	Ex	Ex	Ex		
21	24	120	RM	44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	TwrhsE	1Story	5	7	1976	1976	Gable	CompShg	Ce	Ex	Ex	Ex	Ex	Ex		
22	1314	60	RL	108	14774	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	1Fam	2Story	9	5	1999	1999	Hip	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
23	514	20	RL	71	9187	Pave	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
24	1068	60	RL	80	9760	Pave	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg	Ht	Ex	Ex	Ex	Ex	Ex		
25	1423	120	RM	37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollGr	Norm	TwrhsE	1Story	6	5	2003	2003	Gable	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		
26	1258	30	RL	56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950	Gable	CompShg	W	Ex	Ex	Ex	Ex	Ex	
27	620	60	RL	85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	Vir	Ex	Ex	Ex	Ex	Ex		

After

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD
1	GarageYrB	GarageFini	GarageCar	GarageQui	GarageCor	PavedDriv	WoodDecln	OpenPorch	EnclosedPorch	3SsnPorch	ScreenPorch	PorlArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCond	PriceSale	Test case 1	
2	2007 Rfn	3	954 TA	TA	Y	0	56	0	0	0	0	0	nan	nan	0	6	2007 New	Partial	248328	3320			
3	1962 Unf	2	462 TA	TA	Y	0	0	0	0	0	0	0	0	0	0	3	2007 WD	Normal	101800	2800			
4	1929 Rfn	1	208 TA	TA	Y	0	0	112	0	0	0	0	0	0	0	7	2008 WD	Normal	120000	2000			
5	1925 Unf	1	160 Fa	TA	Y	0	141	0	0	0	0	0	0	0	0	4	2008 WD	Normal	91000	2080			
6	1960 Rfn	1	312 TA	TA	Y	355	0	0	0	0	0	0	0	GdWo	0	4	2008 WD	Normal	141000				
7	1967 Unf	3	792 TA	TA	Y	0	152	0	0	0	0	0	0	0	0	4	2008 WD	Normal	124000	2600			
8	1963 Unf	2	480 TA	TA	Y	0	80	0	0	0	0	0	0	MnPrv	0	6	2009 WD	Normal	139000	3200			
9	1998 Unf	2	402 TA	TA	Y	0	125	0	0	0	0	0	0	0	0	5	2006 WD	Normal	164000	1280			
10	2001 Fin	2	560 TA	TA	Y	125	192	0	0	0	0	0	0	0	0	6	2009 WD	Normal	215000	2840			
11	1962 Unf	2	539 TA	TA	Y	0	23	112	0	0	0	0	0	0	0	1	2009 WD	Normal	103000	2080			
12	1957 Rfn	1	294 TA	TA	Y	250	0	0	0	0	0	0	0	0	0	6	2010 WD	Normal	145000	2800			
13	1977 Unf	2	884 TA	TA	Y	0	64	0	0	0	0	0	0	MnPrv	0	10	2008 WD	Normal	146000	2840			
14	2005 Fin	2	451 TA	TA	Y	252	64	0	0	0	0	0	0	0	0	6	2008 WD	Normal	176000	2400			
15	1974 Unf	4	480 TA	TA	Y	0	0	0	0	0	0	0	0	GdWo	0	6	2007 WD	Normal	123000	2800			
16	1976 Fin	2	665 TA	TA	Y	0	72	174	0	0	0	0	0	0	0	5	2008 COD	Abnrmnl	287000				
17	1972 Unf	1	338 TA	TA	Y	0	0	0	0	0	0	0	0	0	0	8	2009 WD	Normal	133500	1440			
18	1916 Unf	3	513 Fa	Fa	Y	0	0	96	0	0	0	0	0	0	0	5	2008 COD	Abnrmnl	98000	1360			
19	1999 Unf	2	506 TA	TA	Y	0	34	0	0	0	0	0	0	0	0	3	2006 WD	Normal	183900	1400			
20	1995 Unf	2	576 TA	TA	Y	112	0	0	0	0	0	0	0	MnPrv	0	4	2009 WD	Normal	141500	2040			
21	1976 Unf	2	572 TA	TA	Y	100	110	0	0	0	0	0	0	0	0	6	2007 WD	Normal	129900	1760			
22	1999 Fin	3	779 TA	TA	Y	668	30	0	0	0	0	0	0	0	0	5	2010 WD	Normal	333168	4320			
23	1983 Unf	2	484 TA	TA	Y	120	0	158	0	0	0	0	0	0	0	6	2007 WD	Normal	134000	2840			
24	1964 Rfn	2	442 TA	TA	Y	328	128	0	0	189	0	0	0	0	0	6	2008 WD	Normal	167900	3200			
25	2003 Fin	2	420 TA	TA	Y	140	0	0	0	0	0	0	0	0	0	3	2008 WD	Normal	136500	1480			
26	nan	nan	0	0	nan	nan	0	96	0	0	0	0	0	0	0	7	2009 WD	Normal	99900	2240			
27	2003 Fin	3	749 TA	TA	Y	168	0	0	0	0	0	0	0	0	0	8	2008 WD	Normal	305000	3400			
28	2004 Fin	3	734 TA	TA	Y	0	112	0	0	0	0	0	0	0	0	6	2009 WD	Normal	143000	2600			



- Test case 2:

INPUT: *MSSubClass - LotFrontage * 2*

OUTPUT:

Before

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemo	RoofStyle	RoofMatl
2	1242	20 RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007	Hip	CompShg	Vrl		
3	1233	90 RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1962 Gable	CompShg	Hc		
4	1401	50 RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1950 Gable	CompShg	W		
5	1377	30 RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950 Gable	CompShg	W		
6	208	20 RL		12493	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1960 Gable	CompShg	W		
7	1392	90 RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967	1967 Gable	CompShg	Vrl		
8	980	20 RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1963 Gable	CompShg	Vrl		
9	484	120 RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1998 Hip	CompShg	Vrl		
10	392	60 RL	71	12209	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001	2002 Gable	CompShg	Vrl		
11	730	30 RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1950 Gable	CompShg	M	
12	255	20 RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957	1957 Gable	CompShg	M		
13	1094	20 RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	8	1965	1998 Hip	CompShg	M			
14	1021	20 RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2005 Gable	CompShg	Vrl		
15	1341	20 RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971	1971 Gable	CompShg	M		
16	1025	20 RL		15498	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976 Hip	WdShake	St		
17	848	20 RL	36	15523	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	CollGr	Norm	Norm	1Fam	1Story	5	6	1972	1972 Gable	CompShg	Hc		
18	457	70 RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1950 Gable	CompShg	As	
19	1266	160 FV	35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999	1999 Hip	CompShg	M		
20	695	50 RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1950 Gable	CompShg	W		
21	24	120 RM	44	4224	Pave	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976 Gable	CompShg	Ce		
22	1314	60 RL	108	14774	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	9	5	1999	1999 Gable	CompShg	Vrl		
23	514	20 RL	71	9187	Pave	Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983 Gable	CompShg	Vrl		
24	1068	60 RL	80	9760	Pave	Reg	Lvl	AllPub	Inside	Mod	mes	Norm	Norm	1Fam	2Story	6	6	1964	1964 Gable	CompShg	Hc		
25	1423	120 RM	37	4435	Pave	Reg	Lvl	AllPub	Inside	Gtl	CollGr	Norm	Norm	Twnhs	1Story	6	5	2003	2003 Gable	CompShg	Vrl		
26	1258	30 RL	56	4060	Pave	Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1950 Gable	CompShg	W		
27	620	60 RL	85	12244	Pave	Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	2Story	8	5	2003	2003 Hip	CompShg	Vrl		

After

A	B	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27	CD			
1	GarageYrB	GarageFin	GarageCar	GarageAre	GarageQu	GarageCor	PavedDriv	WoodDeck	OpenPorch	EnclosedP.	3SsnPorch	ScreenPor.	PoolArea	PoolQC	Fence	MiscFeatu	MiscVal	MoSold	YrsSold	SaleType	SaleCondit	SalePrice	Test case 2									
2	2007 Rfn	3	954 TA	TA	Y	0	56	0	0	0	0	0	nan	nan	0	6	2007 New	Partial	248328	-146												
3	1962 Unf	2	462 TA	TA	Y	0	0	0	0	0	0	0	nan	nan	0	3	2007 WD	Normal	101800	-50												
4	1929 Rfn	1	208 TA	TA	Y	0	0	112	0	0	0	0	nan	nan	0	7	2008 WD	Normal	120000	-50												
5	1925 Unf	1	160 Fa	TA	Y	0	141	0	0	0	0	0	nan	nan	0	4	2008 WD	Normal	91000	-74												
6	1960 Rfn	1	312 TA	TA	Y	355	0	0	0	0	0	0	0	0	GdWo	nan	0	4	2008 WD	Normal	141000											
7	1967 Unf	3	792 TA	TA	Y	0	152	0	0	0	0	0	0	0	0	0	4	2009 WD	Normal	124000	-40											
8	1963 Unf	2	480 TA	TA	Y	0	80	0	0	0	0	0	0	0	MnPrv	nan	0	6	2009 WD	Normal	139000	-140										
9	1998 Unf	2	402 TA	TA	Y	0	125	0	0	0	0	0	0	0	0	0	5	2006 WD	Normal	164000	56											
10	2001 Fin	2	560 TA	TA	Y	125	192	0	0	0	0	0	0	0	0	0	6	2009 WD	Normal	215000	-82											
11	1962 Unf	2	539 TA	TA	Y	0	23	112	0	0	0	0	0	0	0	0	1	2009 WD	Normal	103000	-74											
12	1957 Rfn	1	294 TA	TA	Y	250	0	0	0	0	0	0	0	0	0	0	6	2010 WD	Normal	145000	-120											
13	1977 Unf	2	884 TA	TA	Y	0	64	0	0	0	0	0	0	0	MnPrv	nan	0	10	2006 WD	Normal	146000	-122										
14	2005 Fin	2	451 TA	TA	Y	252	64	0	0	0	0	0	0	0	0	0	6	2008 WD	Normal	176000	-100											
15	1974 Unf	4	480 TA	TA	Y	0	0	0	0	0	0	0	0	0	GdWo	nan	0	6	2007 WD	Normal	123000	-120										
16	1976 Fin	2	665 TA	TA	Y	0	72	174	0	0	0	0	0	0	0	0	5	2008 COD	Abnorml	287000												
17	1972 Unf	1	338 TA	TA	Y	0	0	0	0	0	0	0	0	0	0	0	8	2009 WD	Normal	133500	-52											
18	1916 Unf	3	513 Fa	Fa	Y	0	0	96	0	0	0	0	0	0	0	0	5	2008 COD	Abnorml	98000	2											
19	1999 Unf	2	506 TA	TA	Y	0	34	0	0	0	0	0	0	0	0	0	3	2006 WD	Normal	183900	90											
20	1995 Unf	2	576 TA	TA	Y	112	0	0	0	0	0	0	0	0	MnPrv	nan	0	4	2009 WD	Normal	141500	-52										
21	1976 Unf	2	572 TA	TA	Y	100	110	0	0	0	0	0	0	0	0	0	6	2007 WD	Normal	129900	32											
22	1999 Fin	3	779 TA	TA	Y	668	30	0	0	0	0	0	0	0	0	0	5	2010 WD	Normal	333168	-156											
23	1983 Unf	2	484 TA	TA	Y	120	0	158	0	0	0	0	0	0	0	0	6	2007 WD	Normal	134000	-122											
24	1964 Rfn	2	442 TA	TA	Y	328	128	0	0	0	189	0	0	0	0	0	0	6	2008 WD	Normal	167900	-100										
25	2003 Fin	2	420 TA	TA	Y	140	0	0	0	0	0	0	0	0	0	0	3	2008 WD	Normal	136500	46											
26	nan	nan	0	0	nan	Y	0	96	0	0	0	0	0	0	0	0	0	7	2009 WD	Normal	99900	-82										
27	2003 Fin	3	749 TA	TA	Y	168	0	0	0</																							

- Test case 3:

INPUT: $(OverallQual + OverallCond) * YearBuilt / YearRemodAdd$
OUTPUT:

Before

After

- Test case 4:

INPUT: $(MSZoning + LotFrontage) / 2$
OUTPUT:

```
(venv) C:\Users\Iris Stream\Desktop\test>python expression_calculate.py -i house-prices.csv -o output.csv
Enter expression: (MSZoning + LotFrontage) / 2
Enter new column name: Test case 4
Invalid attributes: MSZoning
```

IV. Đánh giá đồ án

1. Mức độ hoàn thành của các thành viên

MSSV	Mức độ hoàn thành công việc	Đóng góp
18120078	100%	60%
18120253	80%	40%

2. Mức độ hoàn thành đồ án:

- Tìm hiểu và thực hiện được các giao diện chức năng của Weka một cách cơ bản.
- Cài đặt được các hàm tiền xử lý dữ liệu với Python.
- Sử dụng được tham số dòng lệnh để gọi hàm
- Xử lý được 1 số lỗi cú pháp thường gặp
- Cài đặt xử lý biểu thức trung tố kết hợp với tính toán trên các thuộc tính dữ liệu

VII. Nguồn tham khảo

- <https://pandas.pydata.org/pandas-docs/stable/index.html>
- <https://waikato.github.io/weka-wiki/documentation/>
- <https://www.youtube.com/watch?v=nVJhPRWJApo>
- https://www.youtube.com/watch?v=TF1yh5PKaqI&feature=emb_logo
- <https://www.youtube.com/watch?v=bhxqV3GK-K8>
- <https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/>
- <https://www.statisticshowto.com/how-to-find-a-five-number-summary-in-statistics/>
- <https://stackoverflow.com/questions/35813934/calculate-variance-quartiles-in-weka-explorer>