



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN

-----&&-----

# BÁO CÁO BÀI TẬP LAB 1 KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

**TÊN BÀI TẬP:**  
**DATA PREPROCESS AND DATA  
EXPLORATION**

**LỚP: 21KHMT1**

**Thông tin nhóm:**

21127099 Nguyễn Tấn Lộc

21127411 Trần Thanh Quý

**Giảng viên hướng dẫn:**

Thầy Lê Hoài Bắc

Thầy Nguyễn Bảo Long

Thầy Nguyễn Ngọc Đức

## MỤC LỤC

<b>THÔNG TIN NHÓM .....</b>	<b>2</b>
<b>THÔNG TIN BÀI TẬP .....</b>	<b>2</b>
<b>BẢNG PHÂN CHIA CÔNG VIỆC.....</b>	<b>3</b>
<b>‘KẾT QUẢ BÀI TẬP .....</b>	<b>4</b>
<b>1. Cài đặt Weka .....</b>	<b>4</b>
1.1. Yêu cầu 1 .....	4
1.2. Yêu cầu 2 .....	4
<b>2. Làm quen với Weka .....</b>	<b>7</b>
2.1. Khám phá bộ dữ liệu “Breast Cancer” .....	7
2.2. Khám phá bộ dữ liệu “Weather” .....	15
2.3. Khám phá bộ dữ liệu “Credit in Germany” .....	23
<b>3. Tiền xử lý dữ liệu trong Python .....</b>	<b>33</b>
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>45</b>

## THÔNG TIN NHÓM

MSSV	HỌ TÊN	Email	TỶ LỆ HOÀN THÀNH CÔNG VIỆC
21127099	Nguyễn Tân Lộc	21127099@student.hcmus.edu.vn	100%
21127411	Trần Thanh Quý	21127411@student.hcmus.edu.vn	100%

## THÔNG TIN BÀI TẬP

**Mã học phần:** CSC14004

**Tên học phần:** Khai thác dữ liệu và ứng dụng

**Tên đồ án:** Data Preprocess and Data Exploration

**Lớp:** 21KHMT1

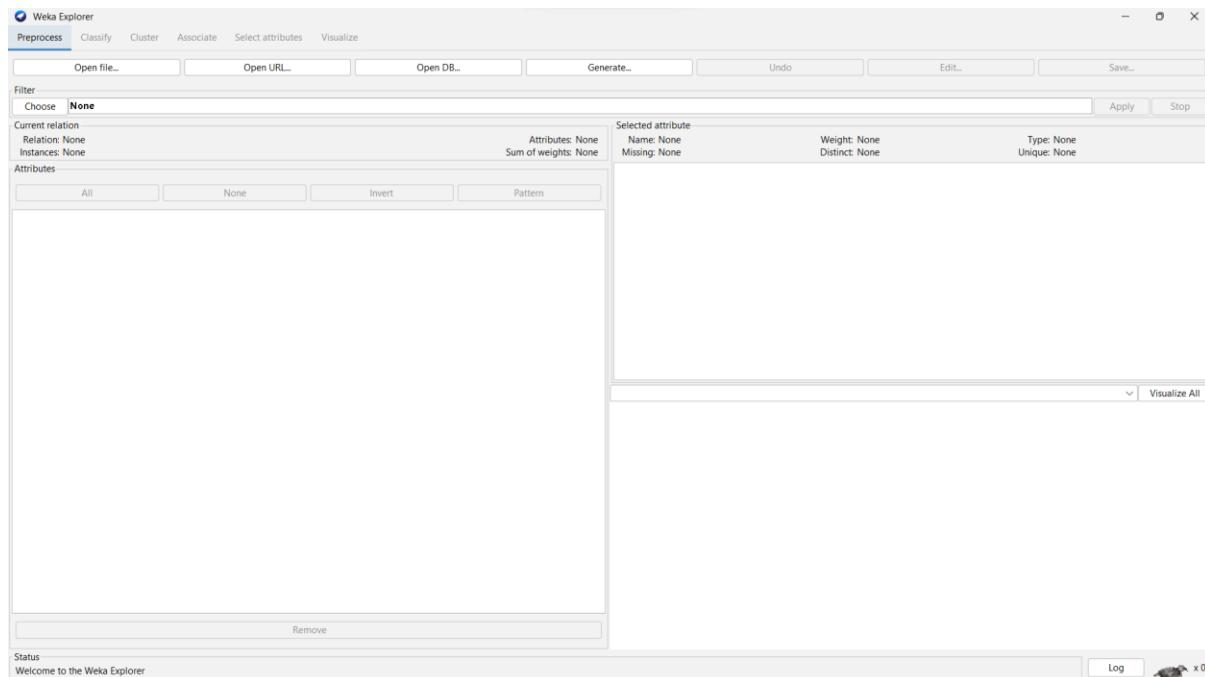
## BẢNG PHÂN CHIA CÔNG VIỆC

Phần	Công việc	Phân công
3.1. Cài đặt WEKA	Yêu cầu 1	Nguyễn Tân Lộc
	Yêu cầu 2	Nguyễn Tân Lộc
3.2. Làm quen với Weka	3.2.1. Khám phá bộ dữ liệu ‘Bread Cancer’.	Nguyễn Tân Lộc
	3.2.2. Khám phá bộ dữ liệu ‘Weather’.	Trần Thanh Quý
	3.2.3. Khám phá bộ dữ liệu ‘Credit in Germany’.	Trần Thanh Quý
3.3. Tiền xử lý dữ liệu trong Python	1. Trích xuất các cột có giá trị bị thiếu.	Nguyễn Tân Lộc
	2. Đếm số dòng thiếu dữ liệu.	Trần Thanh Quý
	3. Điền giá trị còn thiếu bằng giá trị trung bình, trung vị (đối với thuộc tính numerical) và mode (đối với thuộc tính categorical).	Nguyễn Tân Lộc
	4. Xóa các hàng chứa nhiều hơn một số giá trị bị thiếu cụ thể.	Trần Thanh Quý
	5. Xóa các cột chứa nhiều hơn một số giá trị bị thiếu cụ thể	Nguyễn Tân Lộc
	6. Xóa các mẫu trùng lặp	Trần Thanh Quý
	7. Chuẩn hóa thuộc tính số bằng phương pháp min-max và Z-score	Nguyễn Tân Lộc
	8. Thực hiện cộng, trừ, nhân, chia giữa hai thuộc tính numerical	Trần Thanh Quý
Tổng kết	Viết Report	Nguyễn Tân Lộc
	Debug và tổng hợp File Source Code	Trần Thanh Quý

# KẾT QUẢ BÀI TẬP

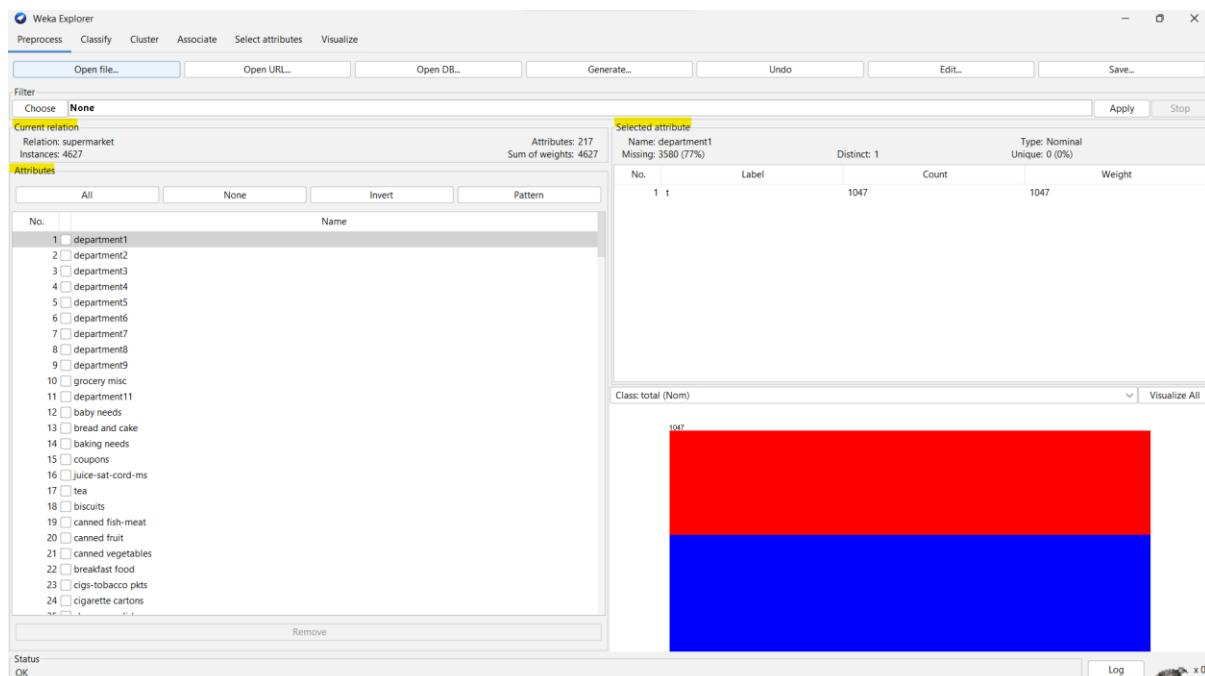
## 1. Cài đặt Weka

### 1.1. Yêu cầu 1



Giao diện Weka Explorer

### 1.2. Yêu cầu 2



Bộ dữ liệu supermarket.arff

Trong bộ dữ liệu <b>supermarket.arff</b>	
<b>Current Relation</b>	<ul style="list-style-type: none"> <li>- Relation: supermarket.</li> <li>- Instances: 4627.</li> <li>- Attributes: 217.</li> <li>- Sum of weight: 4627.</li> </ul>
<b>Attributes</b>	Có 217 thuộc tính, đây là một số thuộc tính: department1, department2, department3, ...
<b>Selected Attribute</b>	<p>Thuộc tính đang chọn là department1 bao gồm các thông tin:</p> <ul style="list-style-type: none"> <li>- Name: department1.</li> <li>- Type: Nominal.</li> <li>- Missing: 3580 (77%).</li> <li>- Distinct: 1.</li> <li>- Unique: 0 (0%).</li> </ul>
Trong Tag Preprocess	
<b>Current Relation</b>	<p>Thông tin chung của bộ dữ liệu, bao gồm:</p> <ul style="list-style-type: none"> <li>- Relation: Tên của bộ dữ liệu.</li> <li>- Instances: số lượng hàng trong bảng.</li> <li>- Attributes: số lượng thuộc tính (trường dữ liệu).</li> <li>- Sum of weight: tổng trọng số.</li> </ul>
<b>Attributes</b>	Những thuộc tính trong bộ dữ liệu, khi click vào 1 thuộc tính thì thông tin chi tiết của thuộc tính sẽ được hiển thị ở phần bên phải ( <b>Selected Attribute</b> )
<b>Selected Attribute</b>	<p>Thông tin chi tiết của thuộc tính được chọn, bao gồm:</p> <ul style="list-style-type: none"> <li>- Name: Tên của thuộc tính.</li> <li>- Type: Kiểu, loại của thuộc tính.</li> <li>- Missing: Số lượng giá trị thiếu.</li> <li>- Distinct: Số lượng giá trị khác biệt.</li> </ul>

	<ul style="list-style-type: none"><li>- Unique: Số lượng giá trị độc nhất, dựa trên tỷ lệ % các giá trị của thuộc tính đang chọn so với giá trị của các thuộc tính khác.</li><li>- Bảng các giá trị của thuộc tính, hiển thị thêm số lượng và trọng lượng theo tỷ lệ của giá trị.</li></ul> <p>Bên dưới phần này có biểu đồ các giá trị.</p>
--	--

Trong 1 bộ dữ liệu, các thuộc tính sẽ có các loại khác nhau, dưới đây là 5 loại thuộc tính:

<b>Numeric</b>	Cho biết giá trị nhỏ nhất, lớn nhất, trung bình và độ lệch chuẩn của dữ liệu số. Loại này đại diện cho 1 số thực (floating-point number)
<b>Nominal</b>	Cho biết số lượng giá trị định danh khác nhau và số lượng mẫu thuộc mỗi giá trị định danh. Loại này đại diện cho một tập hợp cố định các giá trị định danh. (nominal values)
<b>String</b>	Cho biết số lượng giá trị chuỗi khác nhau và số lượng mẫu thuộc mỗi giá trị chuỗi. Loại này đại diện cho một tập hợp đang mở rộng các giá trị định danh. Thông thường được sử dụng trong phân loại văn bản.
<b>Date</b>	Cho biết giá trị nhỏ nhất, lớn nhất, trung bình và độ lệch chuẩn của dữ liệu ngày tháng. Loại này đại diện cho một ngày.
<b>Relational</b>	Cho biết số lượng thuộc tính con và số lượng mẫu trong mỗi quan hệ. Loại này có thể chứa các thuộc tính khác và được sử dụng để biểu diễn dữ liệu Multi-Instance.

Bước đầu tiên trong Machine Learning (học máy), chính là Tiền xử lý dữ liệu nên Tab đầu tiên của Weka là Preprocess. Weka sẽ có thêm 5 Tag khác với các chức năng khác nhau.

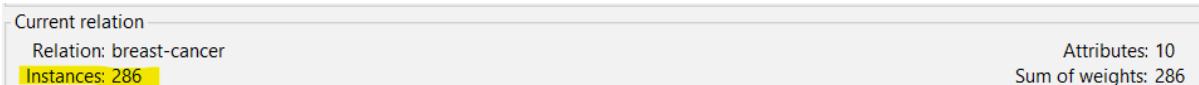
<b>Classify</b>	Tag này cung cấp một số thuật toán học máy để phân loại dữ liệu như Hồi quy tuyến tính, Hồi quy logistic, Decision Tree, RandomTree, RandomForest, NaiveBayes, v.v.
<b>Cluster</b>	Tag này có một số thuật toán phân cụm được cung như SimpleKMeans, FilteredClusterer, HierarchicalClusterer, v.v.
<b>Associate</b>	Tag này có một số công cụ hỗ trợ khai thác tập phổ biến trên dữ liệu như Apriori, FilteredAssociator và FP-Growth.
<b>Select Attributes</b>	Tag này cho phép lựa chọn tính năng dựa trên một số thuật toán như ClassifierSubsetEval, PrincipalComponents, v.v.
<b>Visualize</b>	Tag này cho phép trực quan hóa dữ liệu đã xử lý để phân tích chuyên sâu.

## 2. Làm quen với Weka

### 2.1. Khám phá bộ dữ liệu “Breast Cancer”

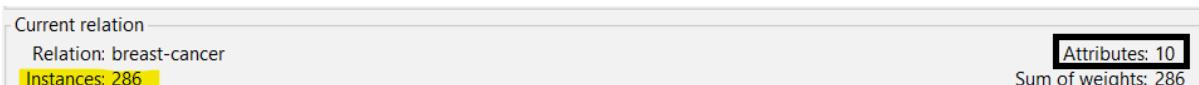
#### a. How many instances does this data set have?

- Bộ dữ liệu này có 286 mẫu.



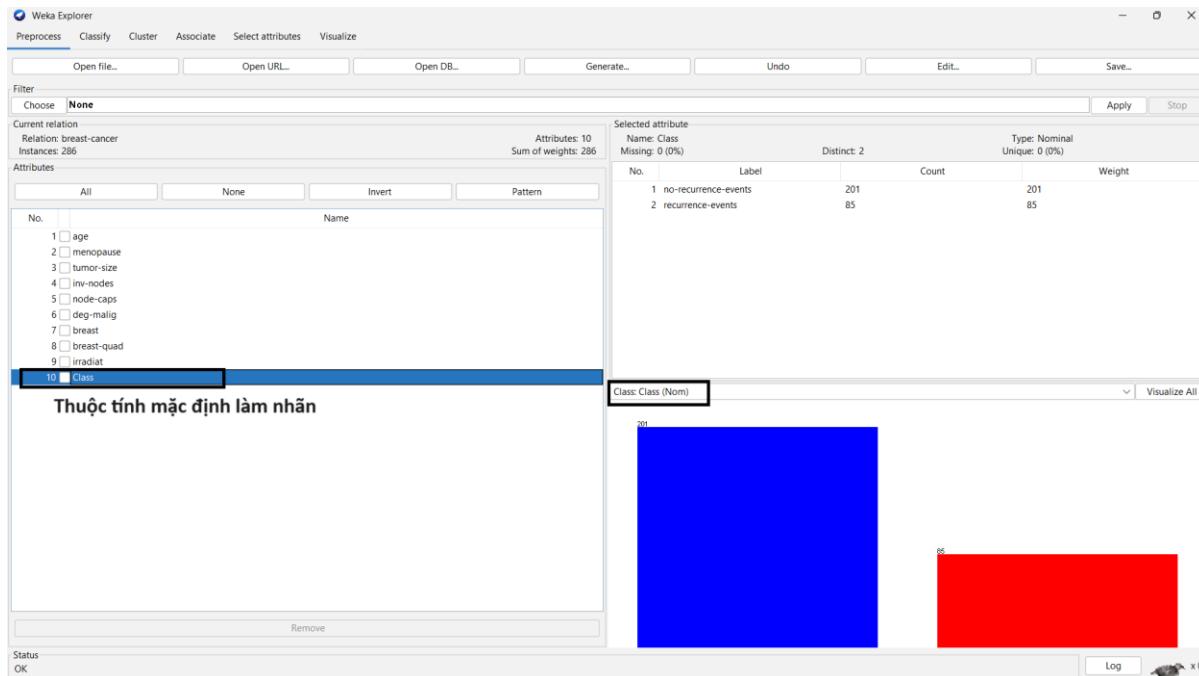
#### b. How many attributes does this data set have?

- Bộ dữ liệu này có 10 thuộc tính khác nhau.

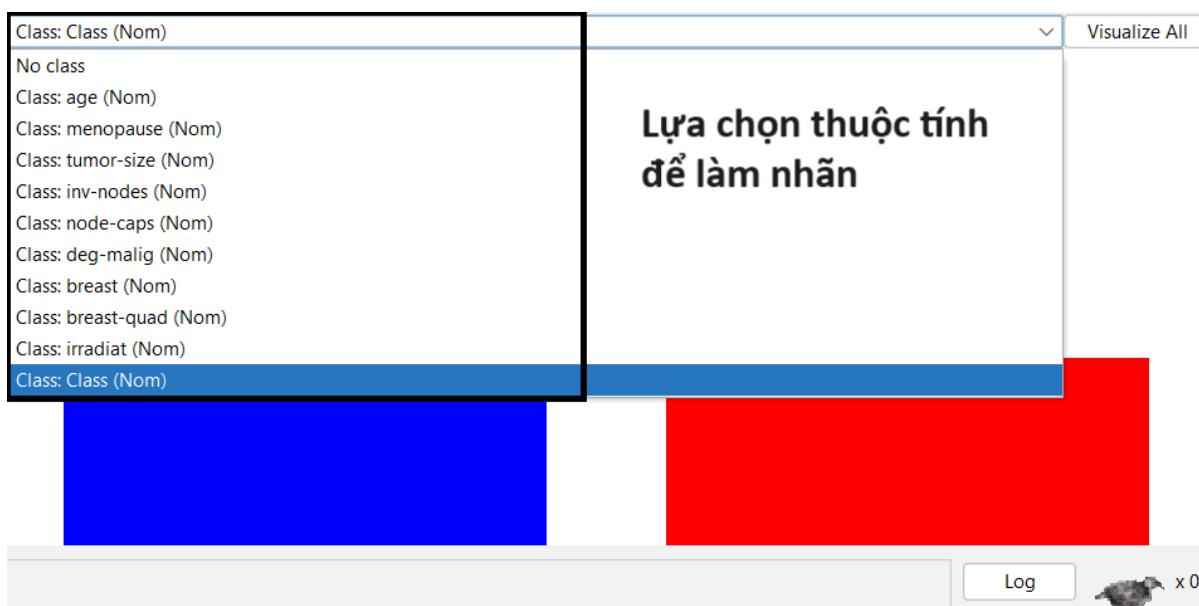


**c. Which attribute is used for the label? Can it be changed? How?**

- Trong 1 bộ dữ liệu, thông thường thuộc tính cuối cùng là thuộc tính mặc định cho làm nhãn (mặc định là Class) như hình bên dưới.



- Mặc dù đây là thuộc tính mặc định của 1 bộ dữ liệu, nhưng ta có thể thay đổi thuộc định làm nhãn bằng 2 cách sau:
- **Cách 1:** Nhấp chuột vào dòng chữ **Class: Class** bên trên biểu đồ trực quan hóa của thuộc tính và chọn thuộc tính khác làm thuộc tính nhãn.



- **Cách 2:** Chọn **Edit**, sau đó muốn chọn thuộc tính nào làm nhãn thì đê chuột vào tên thuộc tính. Tiếp tục chuột phải và chọn **Attribute as class**.

Thuộc tính làm nhăn sẽ in đậm tên thuộc tính và sẽ bị đẩy xuống cột cuối cùng trong bộ dữ liệu.

The screenshot shows the Weka Explorer interface with the 'breast-cancer' dataset loaded. A context menu is open over the 'Class' column, with the 'Attribute as class' option highlighted. A tooltip 'Thuộc tính làm nhăn có in đậm tên thuộc tính.' is displayed. The 'Type: Nominal Unique: 0 (0%)' panel shows 'Weight' values of 201 and 85. The 'Status' panel shows 'OK'.

#### d. What is the meaning of each attribute?

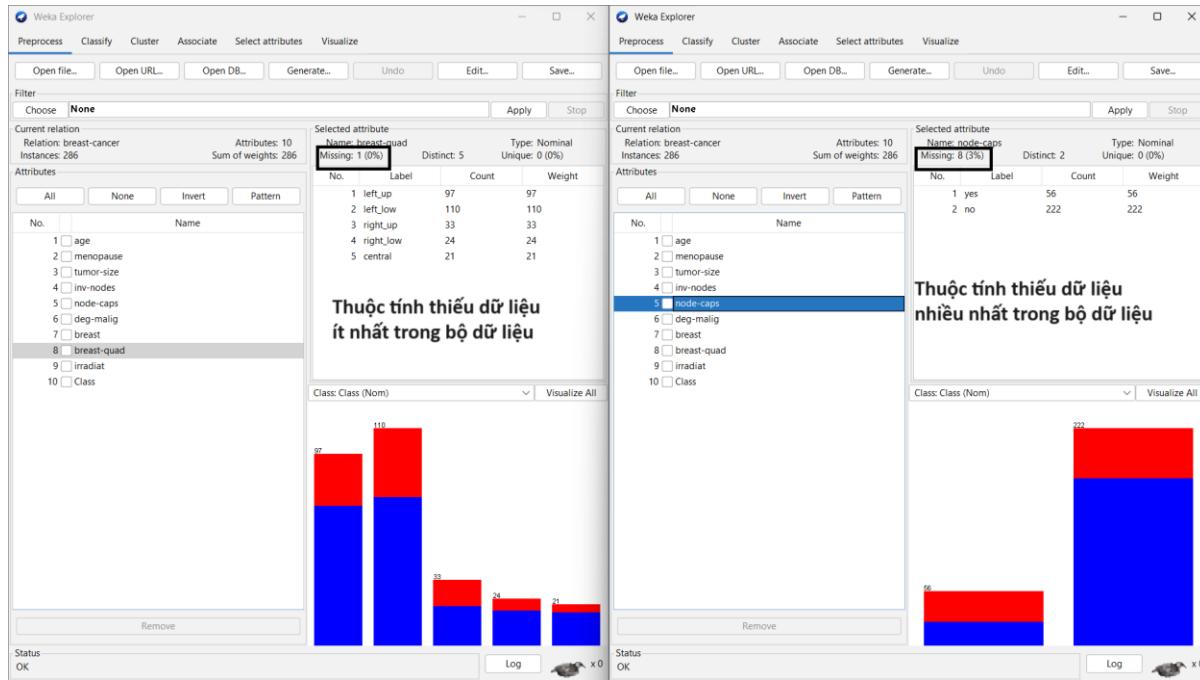
- Đây là bộ dữ liệu về việc ung thư vú nên các thuộc tính liên quan tới bệnh nhân.

No.	Name
1 <input type="checkbox"/> age	
2 <input type="checkbox"/> menopause	
3 <input type="checkbox"/> tumor-size	
4 <input type="checkbox"/> inv-nodes	
5 <input type="checkbox"/> node-caps	
6 <input type="checkbox"/> deg-malig	
7 <input type="checkbox"/> breast	
8 <input type="checkbox"/> breast-quad	
9 <input type="checkbox"/> irradiat	
10 <input type="checkbox"/> Class	

Tên thuộc tính	Ý nghĩa
age	Các độ tuổi của bệnh nhân được chẩn đoán bệnh. Đây là yếu tố quan trọng để đánh giá nguy cơ và chẩn đoán.
menopause	Các giai đoạn mãn kinh của bệnh nhân. Nó được phân loại là tiền mãn kinh hoặc hậu mãn kinh.
tumor-size	Kích thước của khối u.

inv-nodes	Số nút bạch huyết liên quan thường đại diện cho số lượng nút bạch huyết cánh cổ (nách) có liên quan tới di căn của việc ung thư.
node-caps	Nút bạch huyết cánh cổ có thể đề cập đến việc khối u đã lan tới nút bạch huyết cánh cổ hay chưa.
deg-malig	Độ ác (cấp độ mô học) của khối u có thể viết tắt là "Degree of Malignancy", đại diện cho mức độ của sự ác tính của khối u, đó là một đo lường về mức độ hung dữ của ung thư. Cấp độ tăng dần từ 1, 2, 3 là mức độ từ thông thường tới bất thường.
breast	Thuộc tính này có thể chỉ định vùng của vùng bị ảnh hưởng, có thể là vùng bên trái hoặc bên phải của vùng bị ảnh hưởng.
breast-quad	Góc phần tư của vùng bị ảnh hưởng có thể chỉ ra phần tư của vùng bị ảnh hưởng nơi khối u đặt. Vùng bị ảnh hưởng thường được chia thành bốn phần để cung cấp thông tin cụ thể về vị trí.
irradiat	Đề cập đến việc bệnh nhân đã nhận liệu pháp Xạ Trị hay chưa là một phần của quá trình điều trị.
Class	Chỉ ra liệu tình trạng của bệnh nhân có tính lành tính (không phải ung thư) hay ác tính (có ung thư). Bệnh nhân có tái phát bệnh sau khi điều trị không.

**e. Let's investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.**

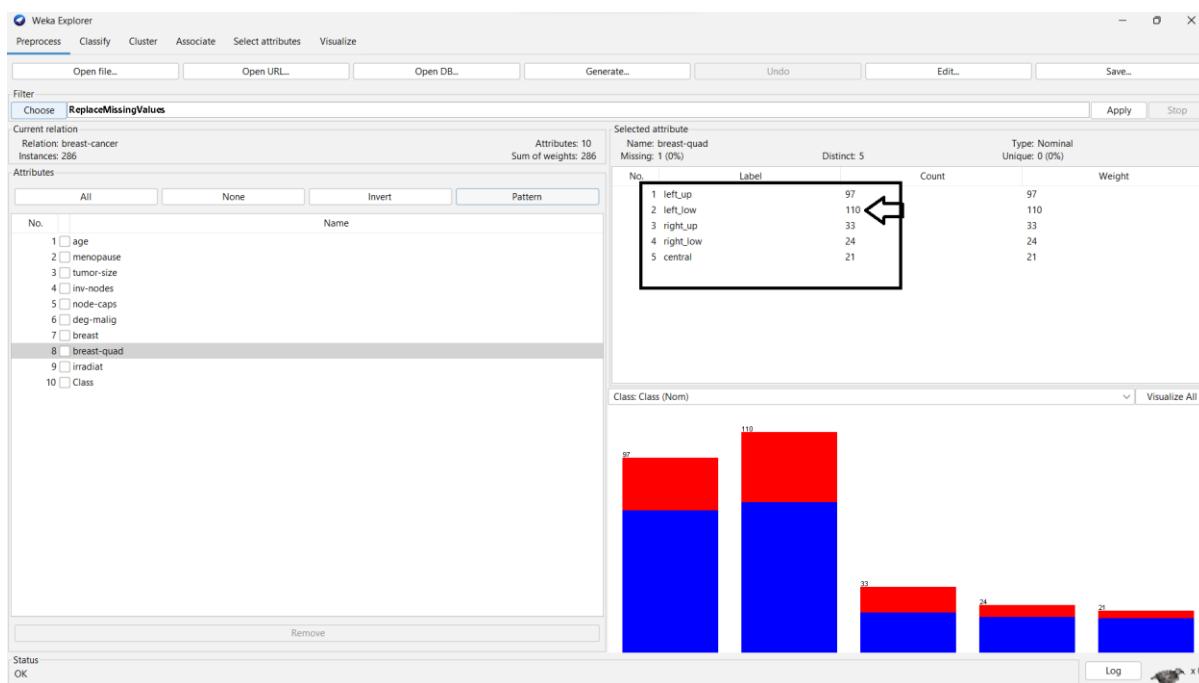


- Trong bộ dữ liệu “Breast Cancer”, có hai thuộc tính bị thiếu dữ liệu, nghĩa là có giá trị thiếu:
  - **breast-quad:** 1 missing value (0%) → thuộc tính có ít dữ liệu thiếu nhất.
  - **node-caps:** 8 missing values (3%) → thuộc tính có nhiều dữ liệu thiếu nhất.
- Một số cách thông thường để giải quyết vấn đề thiếu hụt dữ liệu:
  - Cách 1: Loại bỏ đi những dòng dữ liệu có giá trị thiếu, nghĩa là xóa đi dòng có chứa giá trị thiếu đó.
  - Cách 2: Loại bỏ đi những thuộc tính có chứa nhiều giá trị thiếu, nghĩa là xóa đi cột thuộc tính trong bộ dữ liệu.
  - ➔ Hai cách trên chỉ là những kỹ thuật đơn giản để giải quyết vấn đề bị thiếu dữ liệu. Song, các phương pháp đó lại không được đánh giá cao vì khi ta loại bỏ đi những dòng dữ liệu thiếu giá trị có thể dẫn đến việc mất đi các thuộc tính hoặc dòng dữ liệu quan trọng, làm giảm số lượng của thuộc tính hay dòng dữ liệu có nguy cơ ảnh hưởng đến các giai đoạn sử dụng sau này. Vì thế nên cân nhắc thật kỹ xem xét dòng dữ liệu hay thuộc tính đó có quan trọng không để loại bỏ.
  - Cách 3: Tùy vào đặc tính, tính chất và phân bố dữ liệu trong bộ dữ liệu của thuộc tính có giá trị thiếu đó, ta có thể thêm vào các giá trị thiếu bằng

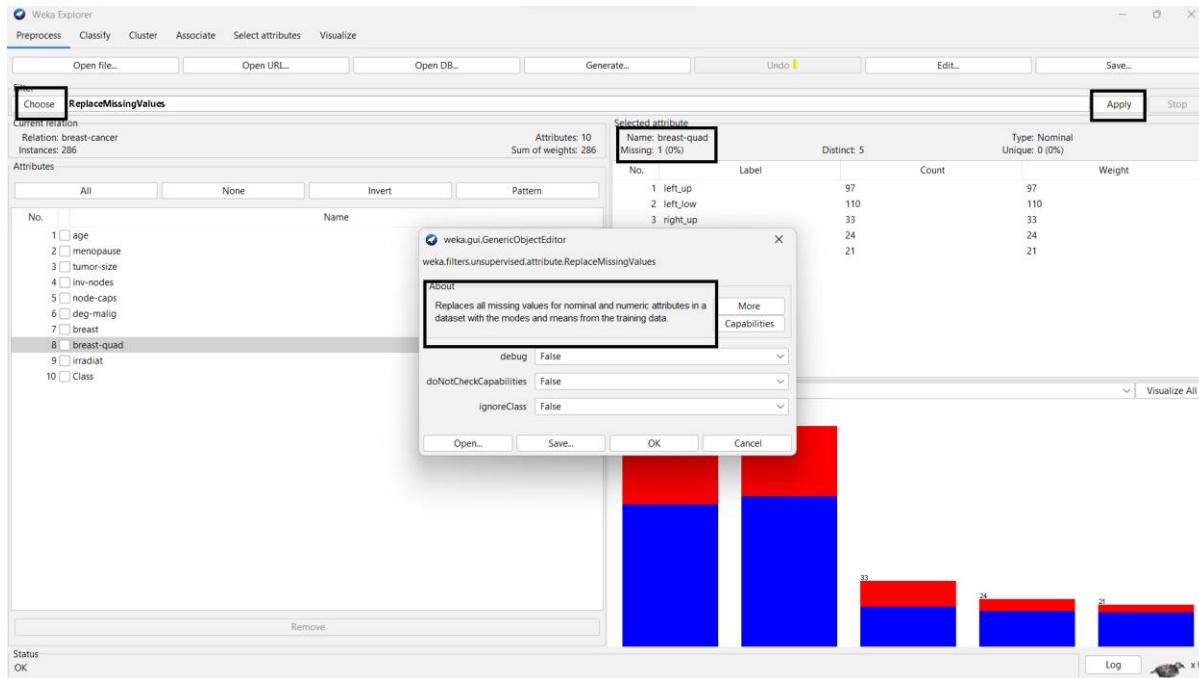
nhiều phương pháp khác nhau như thêm vào các giá trị thiếu là giá trị mean (trung bình), median (trung vị), mode (giá trị có tần số xuất hiện nhiều nhất), hoặc là thay thế bằng một giá trị random trong khoảng từ giá trị thấp nhất tới giá trị cao nhất của thuộc tính, ... Ta có thể sử dụng các thuật toán Machine Learning để tính giá trị thiếu.

**f. Let's propose solutions to the problem of missing values in the specific attribute.**

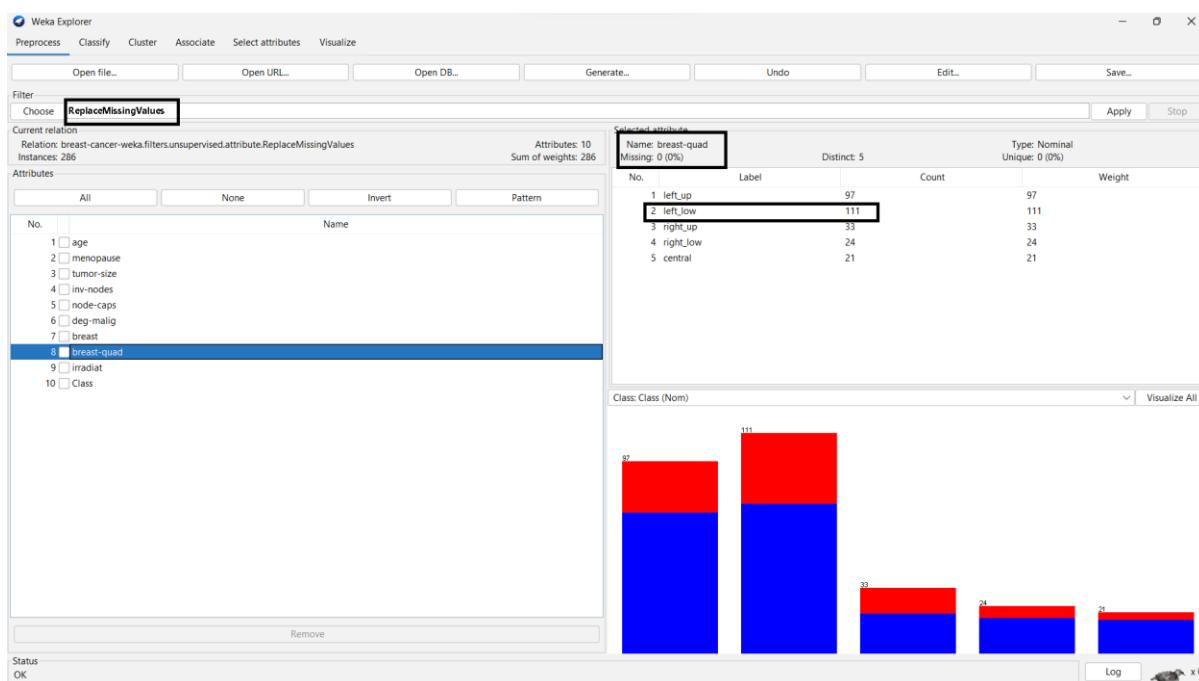
- Khi chọn thuộc tính **breast-quad** để xử lý việc bị thiếu dữ liệu, ta thấy được số lượng từng giá trị bị chênh lệch nhau rất rõ ràng, nên ta có thể sử dụng hàm **Mode**, nghĩa là hàm sẽ lấy giá trị có tần số xuất hiện cao nhất trong bộ các giá trị.



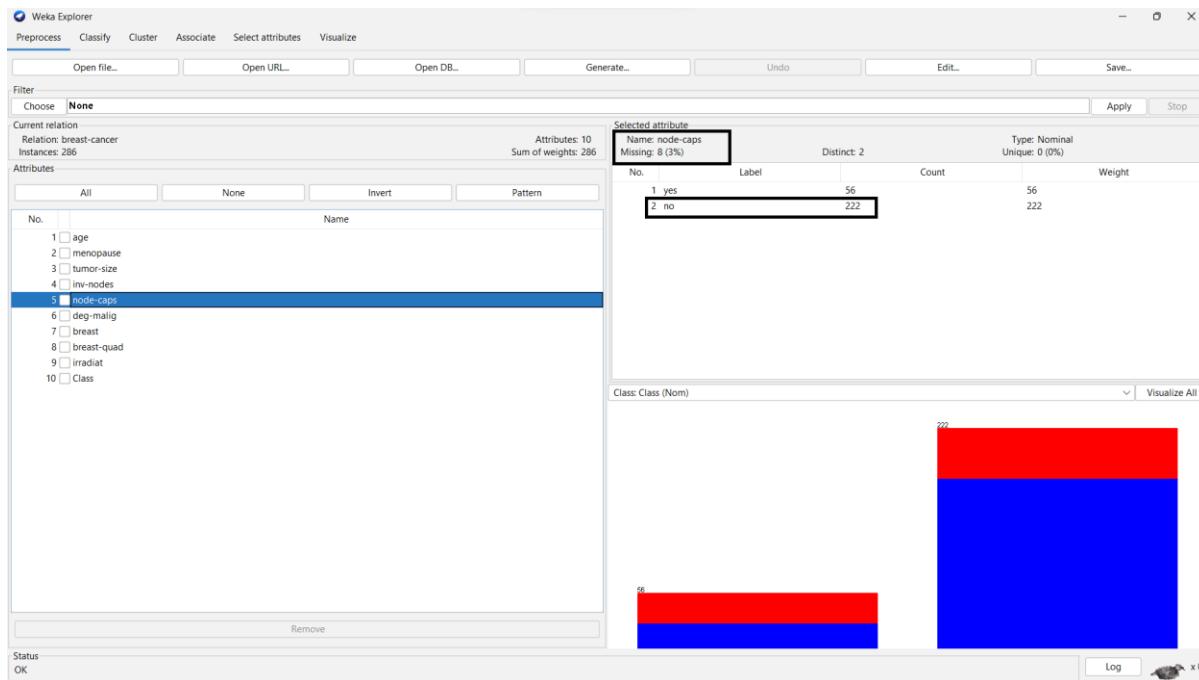
- Cách để sử dụng hàm Mode áp dụng Weka như sau:
  - o Bước 1: Ở Filter, ta chọn Choose.
  - o Bước 2: Chọn theo thứ tự weka → filters → unsupervised → attribute → ReplaceMissingValues.
  - o Bước 3: Nhấn Apply ở góc phải màn hình.
- Note: Có 3 loại Replace Missing Values trong filters nhưng vì ta dùng hàm Mode nên phải sử dụng **ReplaceMissingValues**.



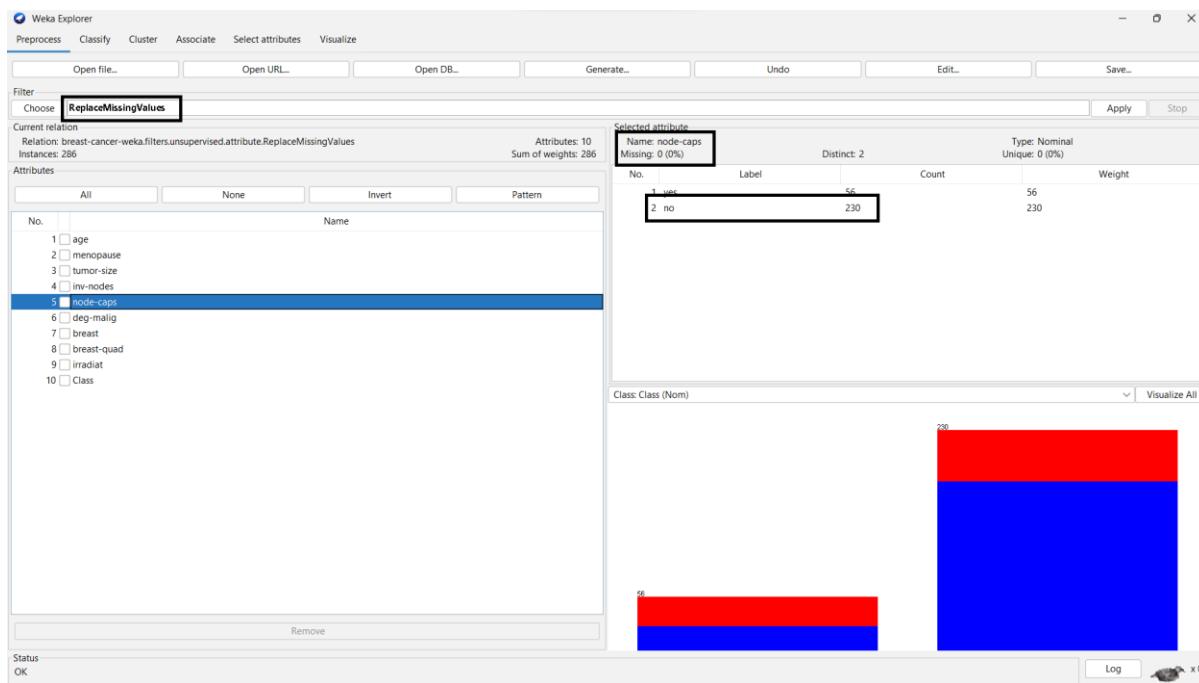
- Sau khi nhấn Apply, số lượng Missing Values sẽ là 0 và số lượng left\_low là 111.



- Khi chọn thuộc tính **node-caps** để xử lý việc bị thiếu dữ liệu, ta chỉ thấy 2 giá trị là **yes** và **no**, chênh lệch số lượng là rất lớn nên ta có thể sử dụng hàm **Mode** như đã sử dụng đối với thuộc tính **breast-quad**.
- Khi áp dụng **ReplaceMissingValues** cho thuộc tính **breast-quad** thì tất cả thuộc tính cũng được áp dụng theo.
- Dưới đây là hình ảnh trước và sau khi áp dụng **ReplaceMissingValues**.



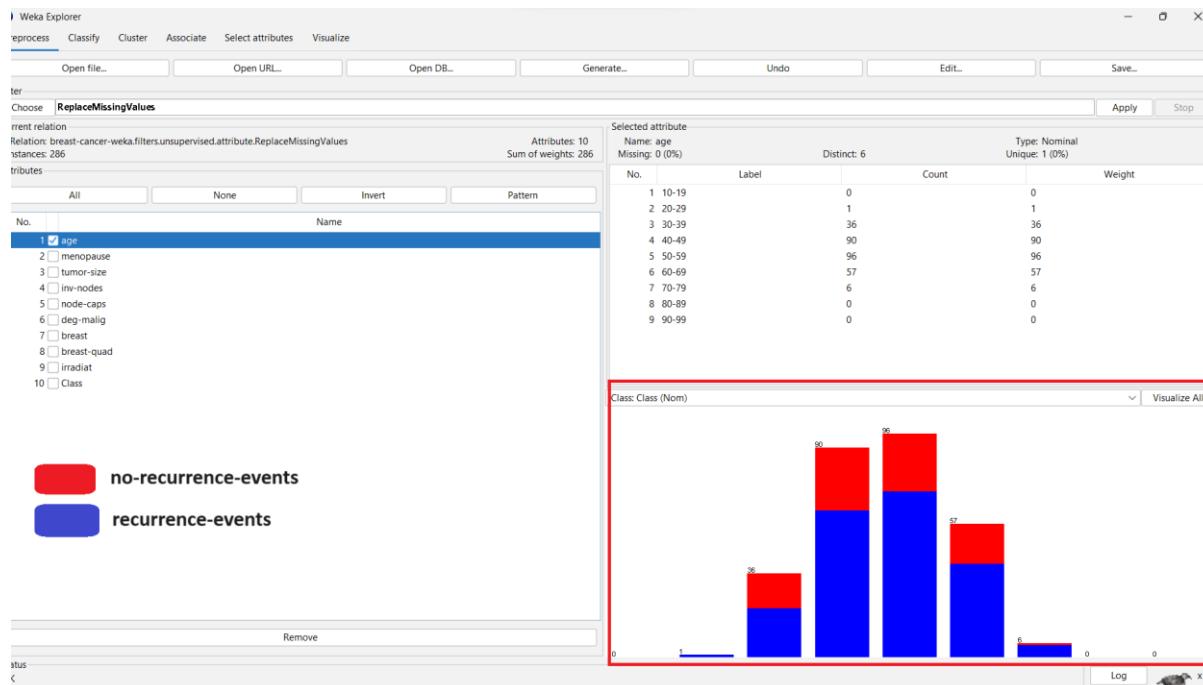
Trước khi áp dụng **ReplaceMissingValues**



Sau khi áp dụng **ReplaceMissingValues**

**g. Let's explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.**

- Ý nghĩa của biểu đồ trong Weka Explorer: thể hiện sự tương quan giữa thuộc tính đang chọn và thuộc tính làm nhãn (thuộc tính Class). Dựa vào số lượng của từng giá trị mà biểu đồ sẽ thể hiện.



- Hình bên trên là ta chọn thuộc tính **age** để so sánh với thuộc tính làm nhãn là **Class**. Ta có thể hiểu ngắn gọn như sau: biểu đồ trên thể hiện số lượng bệnh nhân có chẩn đoán mắc bệnh ung thư vú với phân loại là có tái phát bệnh hay không. Ví dụ ở cột thứ 3 (30-39) với số lượng là 36 người thì sẽ biết được màu xanh của cột là số lượng người tái phát bệnh và màu đỏ số lượng người không tái phát.
- Nhìn vào biểu đồ, ta cài đặt tiêu đề và chú thích như sau:
  - Tên biểu đồ: Số lượng bệnh nhân tái phát bệnh trong từng độ tuổi.
  - Chú thích:
    - Màu đỏ: số lượng bệnh nhân không tái phát bệnh.
    - Màu xanh: số lượng bệnh nhân tái phát bệnh.

## 2.2. Khám phá bộ dữ liệu “Weather”

a. *How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?*

- Sau khi mở bộ dữ liệu **Weather**, các câu trả lời lần lượt là:
  - Có 5 thuộc tính và 14 mẫu dữ liệu.
  - Có 3 thuộc tính thuộc loại **categorical**: outlook, windy, play.
  - Có 2 thuộc tính thuộc loại **numerical**: temperature, humidity.
  - Thuộc tính được sử dụng làm nhãn: play. (Có giá trị yes or no, tương tự như bộ dữ liệu **Breast Cancer**).

No.	1: outlook	2: temperature	3: humidity	4: wind	5: play
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

b. Let's list *five-number summary* of two attributes *temperature* and *humidity*. Does WEKA provide these values?

- **Five-number summary:** là một kỹ thuật tóm tắt dữ liệu phi tham số. Nó có thể được sử dụng để mô tả việc phân phối các mẫu dữ liệu cho dữ liệu với bất kỳ phân phối nào. **Five-number summary** liên quan đến việc tính toán 5 đại lượng thống kê tóm tắt, cụ thể là:
  - **Median:** Giá trị ở giữa trong mẫu, còn được gọi là phân vị thứ 50.
  - **1<sup>st</sup> Quartile:** Phân vị thứ 25.
  - **3<sup>rd</sup> Quartile:** Phân vị thứ 75.
  - **Minimum:** Giá trị nhỏ nhất trong mẫu.
  - **Maximum:** Giá trị lớn nhất trong mẫu.
- Sau đây là bảng thể hiện các giá trị thống kê ở trên của 2 thuộc tính **temperature** và **humidity**:

Statistic	Temperature	Humidity
<b>Minimum</b>	64	65
<b>1<sup>st</sup> Quartile</b>	69.25	71.25
<b>Median</b>	72	82.5
<b>3<sup>rd</sup> Quartile</b>	78.75	90
<b>Maximum</b>	85	96

```

import pandas as pd
import numpy as np
from numpy import percentile
#Temperature's data
temperature = np.array([85.0, 80.0, 83.0, 70.0, 68.0, 65.0,
                       64.0, 72.0, 69.0, 75.0, 75.0, 72.0, 81.0, 71.0])
humidity = np.array([85.0, 90.0, 86.0, 96.0, 80.0, 70.0, 65.0, 95.0, 70.0, 80.0, 70.0, 90.0, 75.0, 91.0])
# calculate quartiles
quartiles_1 = percentile(temperature, [25, 50, 75])
quartiles_2 = percentile(humidity, [25, 50, 75])
# calculate min/max
data_min_1, data_max_1 = temperature.min(), temperature.max()
data_min_2, data_max_2 = humidity.min(), humidity.max()
summary_df = pd.DataFrame({
    'Statistic': ['Minimum', '1st Quartile', 'Median', '3rd Quartile', 'Maximum'],
    'Temperature': [data_min_1, quartiles_1[0], quartiles_1[1], quartiles_1[2], data_max_1],
    'Humidity': [data_min_2, quartiles_2[0], quartiles_2[1], quartiles_2[2], data_max_2]
})
summary_df

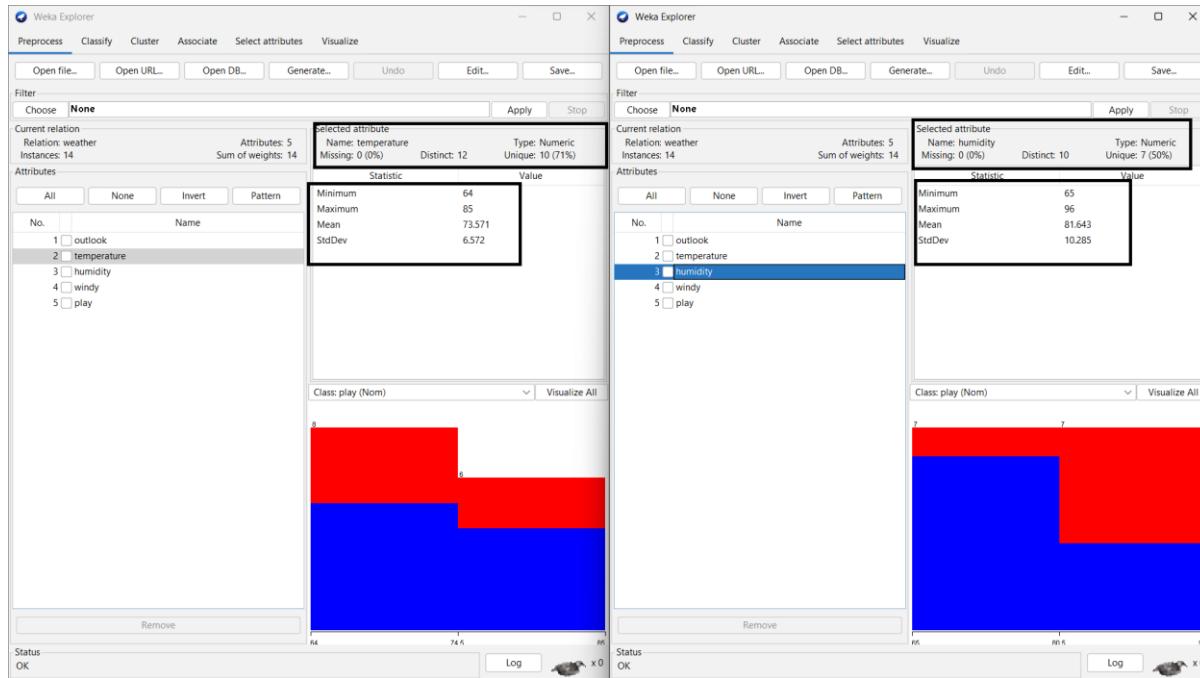
```

0.0s

	Statistic	Temperature	Humidity
0	Minimum	64.00	65.00
1	1st Quartile	69.25	71.25
2	Median	72.00	82.50
3	3rd Quartile	78.75	90.00
4	Maximum	85.00	96.00

### *Đoạn Code Python tính toán Five-Number Summary*

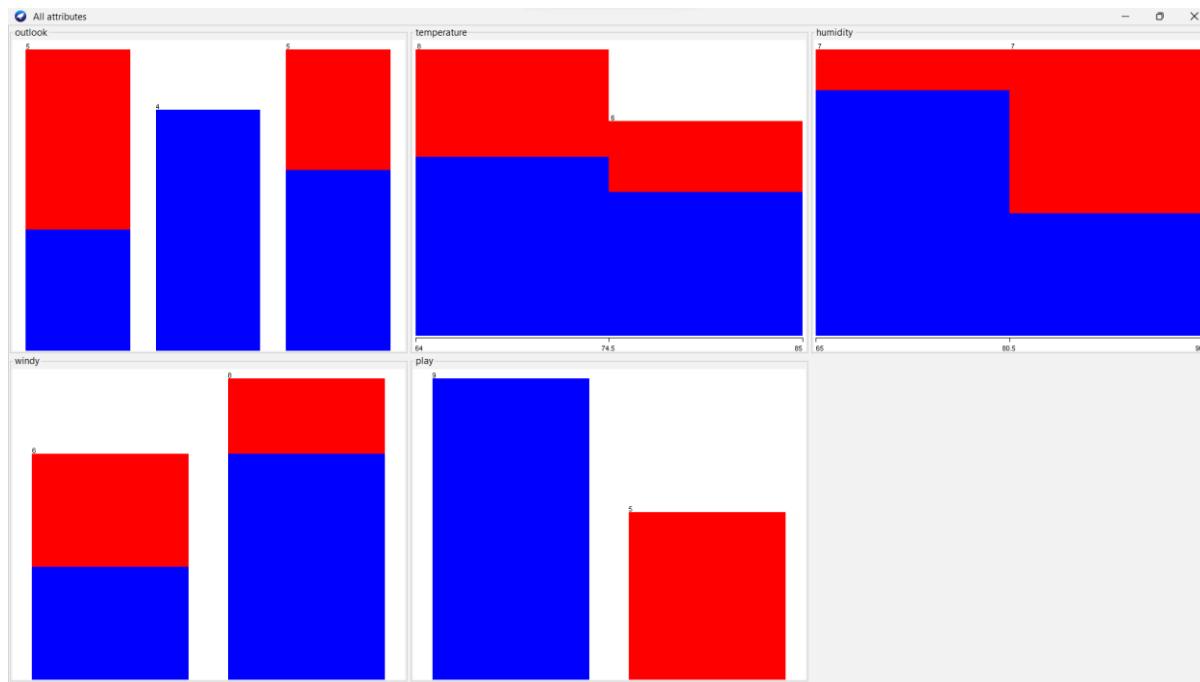
- Trong Weka chỉ thể hiện 2 trong số 5 giá trị thống kê trên, đó là: **Minimum** và **Maximum**. Bên cạnh đó, Weka còn thể hiện giá trị **Mean** (trung bình) và độ lệch chuẩn của mẫu (StdDev).



### *Thông tin các giá trị của temperature và humidity*

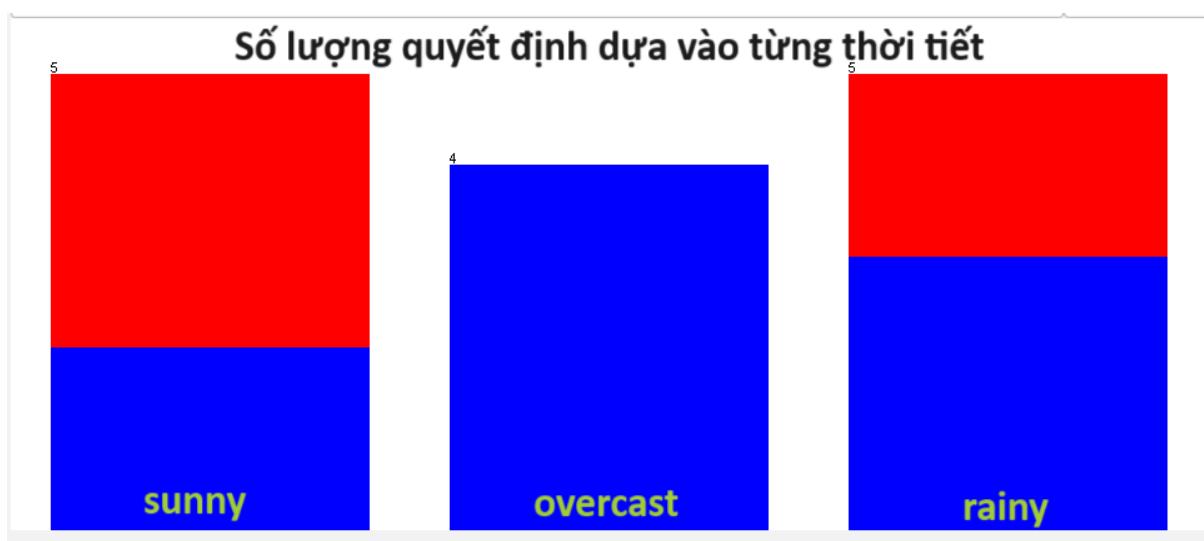
- Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.*

- Để có thể xem tất cả biểu đồ các thuộc tính của bộ dữ liệu, ta nhấn vào **Visualize All** bên trên biểu đồ của thuộc tính đang chọn.



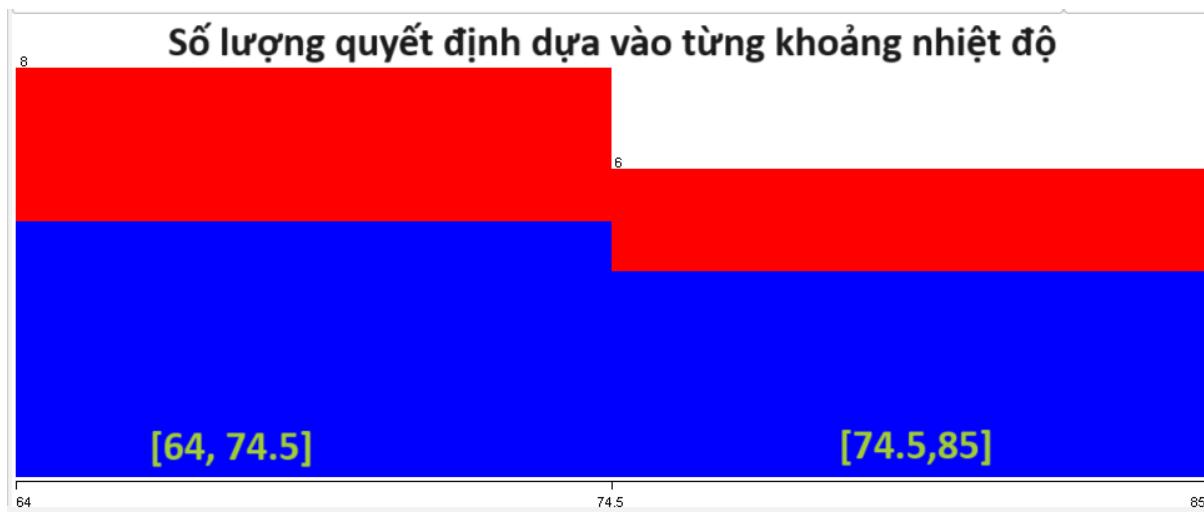
*Biểu đồ của tất cả thuộc tính trong bộ dữ liệu Weather*

- 5 biểu đồ từ trái qua phải tương ứng cho các thuộc tính: outlook, temperature, himidity, windy, play. Như đã giải thích ở phần “Khám phá bộ dữ liệu Breast Cancer”, các biểu đồ thể hiện sự tương quan giữa thuộc tính đang chọn và thuộc tính làm nhãn (thuộc tính Class). Trong thuộc tính Nominal, trục x sẽ là các cột giá trị và trục y sẽ là số lượng của từng cột giá trị. Trong thuộc tính Numeric, trục x sẽ là khoảng giá trị từ **Minimum** đến **Maximum** và trục y sẽ là tần số xuất hiện của các quyết định.
  - Việc sử dụng màu sắc để biểu thị tỷ lệ số lượng một thuộc tính cụ thể (có giá trị **yes** hoặc **no**) trong ngũ cành của thuộc tính **play**. Màu xanh thường được sử dụng đại diện cho giá trị **yes**, màu đỏ được sử dụng đại diện cho giá trị **no**.
  - Dưới đây là thuộc tính **outlook**, nghĩa là quyết định đi chơi hay không sẽ dựa vào thời tiết.
    - Tên biểu đồ: Số lượng quyết định dựa vào từng thời tiết.
    - Chú thích:
      - Màu xanh: quyết định **yes**.
      - Màu đỏ: quyết định **no**.
  - Đối với **sunny** với số lượng là 5 quyết định thì sẽ có 1 số quyết định **yes** (màu xanh) và một số quyết định **no** (màu đỏ). Đối với **overcast** sẽ có 4 quyết định. Đối với **rainy** với số lượng là 5 quyết định thì sẽ có 1 số quyết định **yes** (màu xanh) và một số quyết định **no** (màu đỏ).
- ➔ Tổng cộng:  $5+4+5=14$  mẫu dữ liệu.



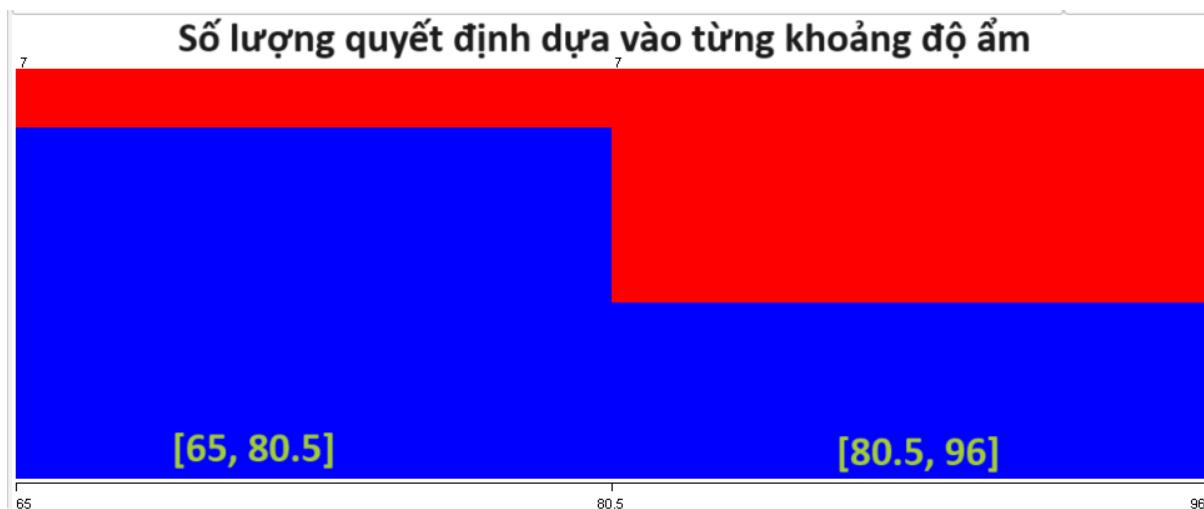
Biểu đồ số lượng quyết định dựa vào từng thời tiết

- Dưới đây là thuộc tính **temperature**, nghĩa là quyết định đi chơi hay không sẽ dựa vào nhiệt độ.
  - Tên biểu đồ: Số lượng quyết định dựa vào từng khoảng nhiệt độ.
  - Chú thích:
    - Màu xanh: quyết định yes.
    - Màu đỏ: quyết định no.
- Quyết định sẽ được chia thành 2 khoảng [Minimum, Mean] và [Mean, Maximum], tức là khoảng 1 [64, 74.5] và khoảng 2 [74.5, 85]. Ở khoảng 1 sẽ có 8 quyết định, khoảng 2 có 6 quyết định trong đó có một số quyết yes (màu xanh) và quyết định no (màu đỏ).  
→ Tổng cộng:  $8+6=14$  mẫu dữ liệu.



Biểu đồ số lượng quyết định dựa vào từng khoảng nhiệt độ

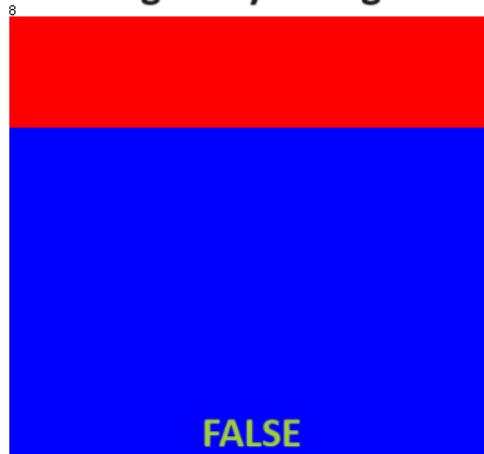
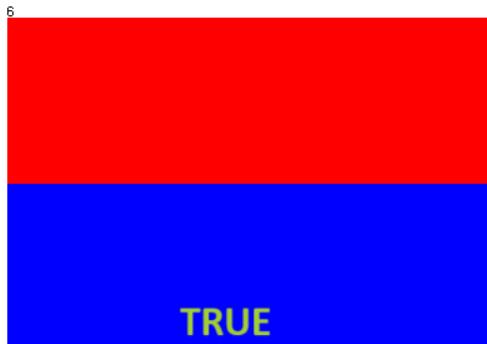
- Dưới đây là thuộc tính **humidity**, nghĩa là quyết định đi chơi hay không sẽ dựa vào từng khoảng độ ẩm.
  - o Tên biểu đồ: Số lượng quyết định dựa vào từng khoảng độ ẩm.
  - o Chú thích:
    - Màu xanh: quyết định **yes**.
    - Màu đỏ: quyết định **no**.
- Quyết định sẽ được chia thành 2 khoảng [Minimum, Mean] và [Mean, Maximum], tức là khoảng 1 [65, 80.5] và khoảng 2 [80.5, 96]. Ở khoảng 1 sẽ có 7 quyết định, khoảng 2 có 7 quyết định trong đó có một số quyết **yes** (màu xanh) và quyết định **no** (màu đỏ).
  - ➔ Tổng cộng:  $7+7=14$  mẫu dữ liệu.



*Biểu đồ số lượng quyết định dựa vào từng khoảng độ ẩm*

- Dưới đây là thuộc tính **windy**, nghĩa là quyết định đi chơi hay không sẽ dựa vào có gió hay không.
  - o Tên biểu đồ: Số lượng quyết định dựa vào có gió hay không.
  - o Chú thích:
    - Màu xanh: quyết định **yes**.
    - Màu đỏ: quyết định **no**.
- Đối với **TRUE** với số lượng là 6 quyết định thì sẽ có 1 số quyết định **yes** (màu xanh) và một số quyết định **no** (màu đỏ). Đối với **FALSE** với số lượng là 8 quyết định thì sẽ có 1 số quyết định **yes** (màu xanh) và một số quyết định **no** (màu đỏ).
  - ➔ Tổng cộng:  $8+6=14$  mẫu dữ liệu.

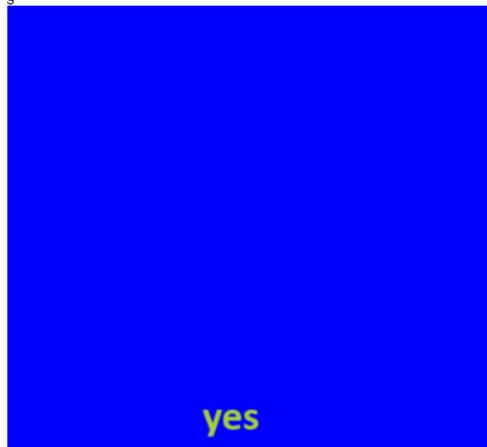
### Số lượng quyết định dựa vào có gió hay không



Biểu đồ số lượng quyết định dựa vào có gió hay không

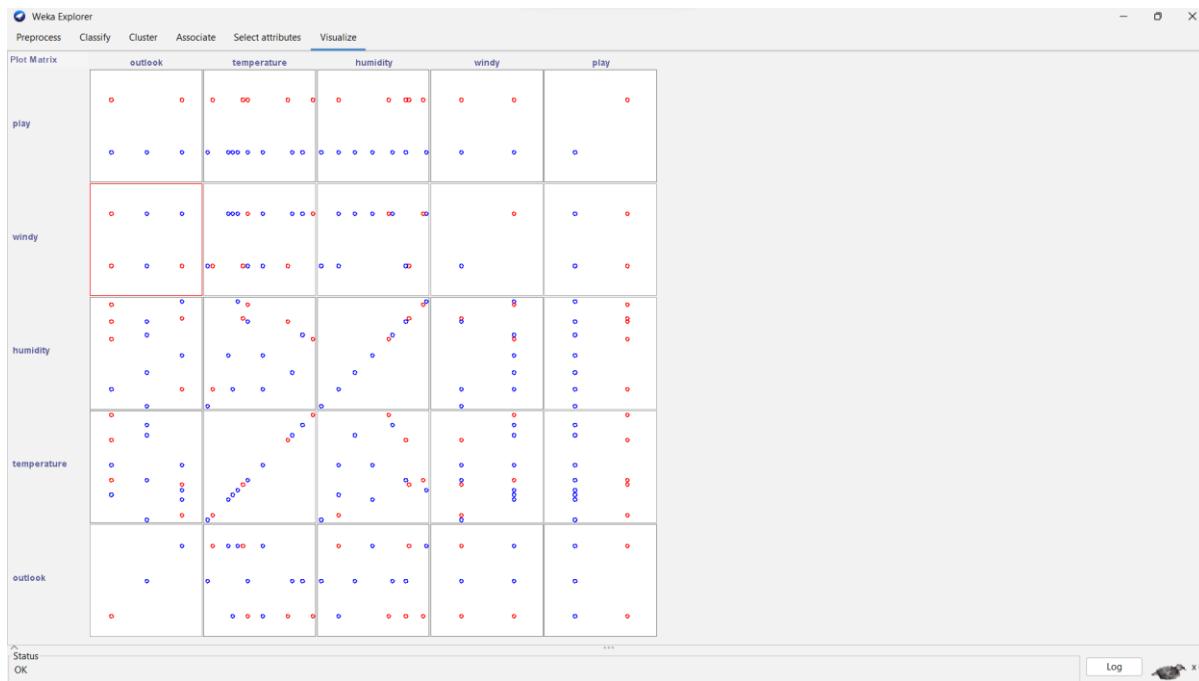
- Dưới đây là thuộc tính **play**, là thuộc tính được dùng làm nhãn (thuộc tính **Class**), nghĩa là quyết định đi chơi hay không.
  - Tên biểu đồ: Số lượng quyết định đi chơi hay không.
  - Chú thích:
    - Màu xanh: quyết định yes.
    - Màu đỏ: quyết định no.
- Đối với **yes** với số lượng là 9 quyết định (màu xanh). Đối với **no** với số lượng là 5 quyết định (màu đỏ)
- Tổng cộng:  $9-5=14$  mẫu dữ liệu.

### Số lượng quyết định đi chơi hay không



Biểu đồ số lượng quyết định đi chơi hay không

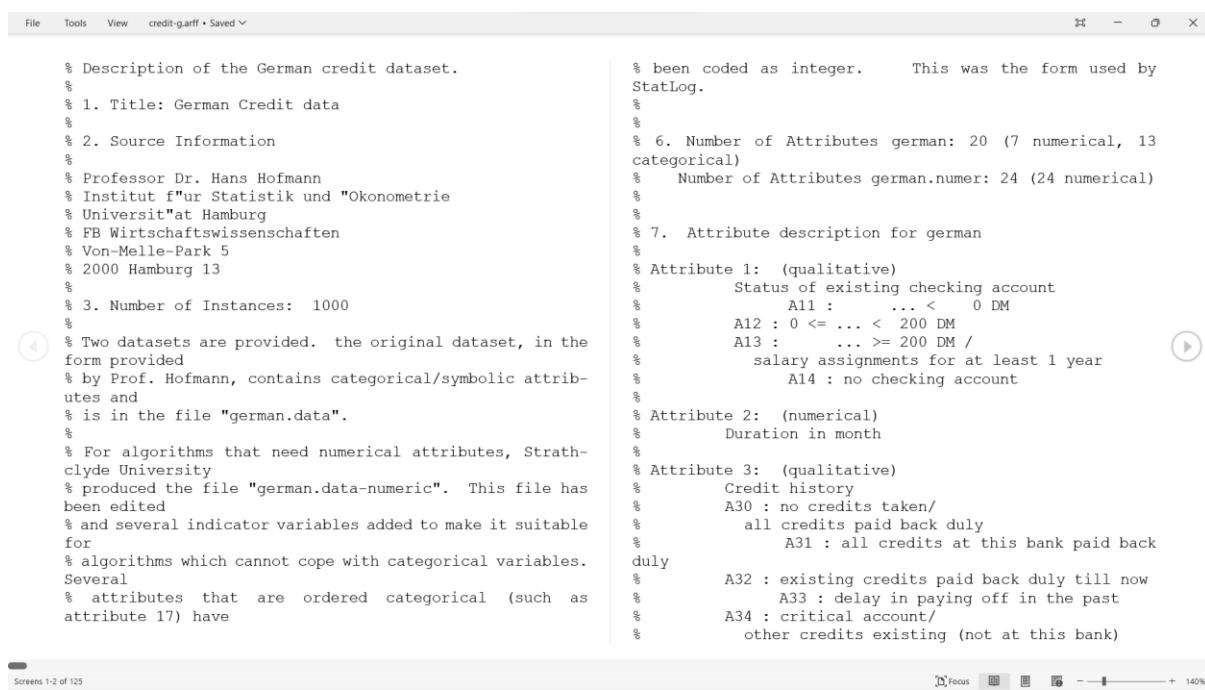
*d. Let's move to the Visualize tag. What's the name of this chart? Do you think there are any pairs of different attributes that have correlated?*



- Biểu đồ trên có tên là **Scatter Plots Matrix/ Scatter Diagram**. Biểu đồ phân tán là một loại biểu đồ sử dụng tọa độ Descartes để hiển thị giá trị và mối quan hệ giữa hai biến định lượng cho một tập dữ liệu. Dữ liệu được hiển thị dưới dạng tập hợp các điểm, mỗi điểm có giá trị của một biến xác định vị trí trên trục hoành và giá trị của biến khác xác định vị trí trên trục tung.
- Có 3 mối tương quan chính:
  - Tương quan cao (Positive Correlation): các điểm dữ liệu sẽ gần nhau và theo hướng đi lên.
  - Không tương quan (No Correlation): các điểm dữ liệu sẽ gần nhau và theo hướng đi xuống.
  - Tương quan thấp (Negative Correlation): các điểm dữ liệu sẽ gần nhau và không có chiều hướng lên hoặc xuống mà là phân tán đều.
- Nhìn vào biểu đồ trên, ta có thể thấy ngoại trừ 3 cặp thuộc tính giống nhau mới có mối tương quan cao, còn các cặp thuộc tính khác nhau thì các điểm dữ liệu được phân bố không theo hướng lên hay xuống nên vì vậy không có thuộc tính nào có mối tương quan cả.

### 2.3. Khám phá bộ dữ liệu “Credit in Germany”

- a. *What is the content of the comments section in credit-g.arff (when opened with any text editor) about? How many samples does the data set have? How many attributes? Describe any five attributes (must have both discrete and continuous attributes).*



```
% Description of the German credit dataset.  
%  
% 1. Title: German Credit data  
%  
% 2. Source Information  
%  
% Professor Dr. Hans Hofmann  
% Institut f"ur Statistik und "Okonometrie  
% Universit"at Hamburg  
% FB Wirtschaftswissenschaften  
% Von-Melle-Park 5  
% 2000 Hamburg 13  
%  
% 3. Number of Instances: 1000  
%  
% Two datasets are provided. the original dataset, in the  
% form provided  
% by Prof. Hofmann, contains categorical/symbolic attrib-  
utes and  
% is in the file "german.data".  
%  
% For algorithms that need numerical attributes, Strath-  
clyde University  
% produced the file "german.data-numeric". This file has  
been edited  
% and several indicator variables added to make it suitable  
for  
% algorithms which cannot cope with categorical variables.  
Several  
% attributes that are ordered categorical (such as  
attribute 17) have  
% been coded as integer. This was the form used by  
StatLog.  
%  
% 6. Number of Attributes german: 20 (7 numerical, 13  
categorical)  
%   Number of Attributes german.numer: 24 (24 numerical)  
%  
%  
% 7. Attribute description for german  
%  
% Attribute 1: (qualitative)  
%   Status of existing checking account  
%     A11 : ... < 0 DM  
%     A12 : 0 <= ... < 200 DM  
%     A13 : ... >= 200 DM /  
%           salary assignments for at least 1 year  
%     A14 : no checking account  
%  
% Attribute 2: (numerical)  
%   Duration in month  
%  
% Attribute 3: (qualitative)  
%   Credit history  
%     A30 : no credits taken/  
%           all credits paid back duly  
%     A31 : all credits at this bank paid back  
duly  
%     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)
```

Tệp dữ liệu **credit-g.arff** được mở bằng Word

- Nội dung của ghi chú liên quan đến việc mô tả tập dữ liệu, bao gồm các thông tin sau:
  - Tiêu đề của bộ dữ liệu.
  - Nguồn thông tin của bộ dữ liệu.
  - Số lượng mẫu trong bộ dữ liệu.
  - Số lượng thuộc tính trong bộ dữ liệu.
  - Mô tả từng thuộc tính trong bộ dữ liệu.
- ➔ Theo ghi chú, có hai bộ dữ liệu được cung cấp. Bộ dữ liệu gốc được cung cấp bởi giáo sư Hofmann và chứa các thuộc tính phân loại/biểu tượng trong tệp german.data. Đối với thuật toán cần thuộc tính số, Đại học Strathclyde cung cấp tệp german.datanumeric. Tệp này được chỉnh sửa và thêm một số biến của bộ chỉ báo để phù hợp với các thuật toán không thể sử dụng các biến phân loại. Một số thuộc tính được sắp xếp theo thứ tự phân loại, (ví dụ như thuộc tính 17), đã được mã hóa thành số nguyên. Đây là hình thức được StatLog sử dụng.

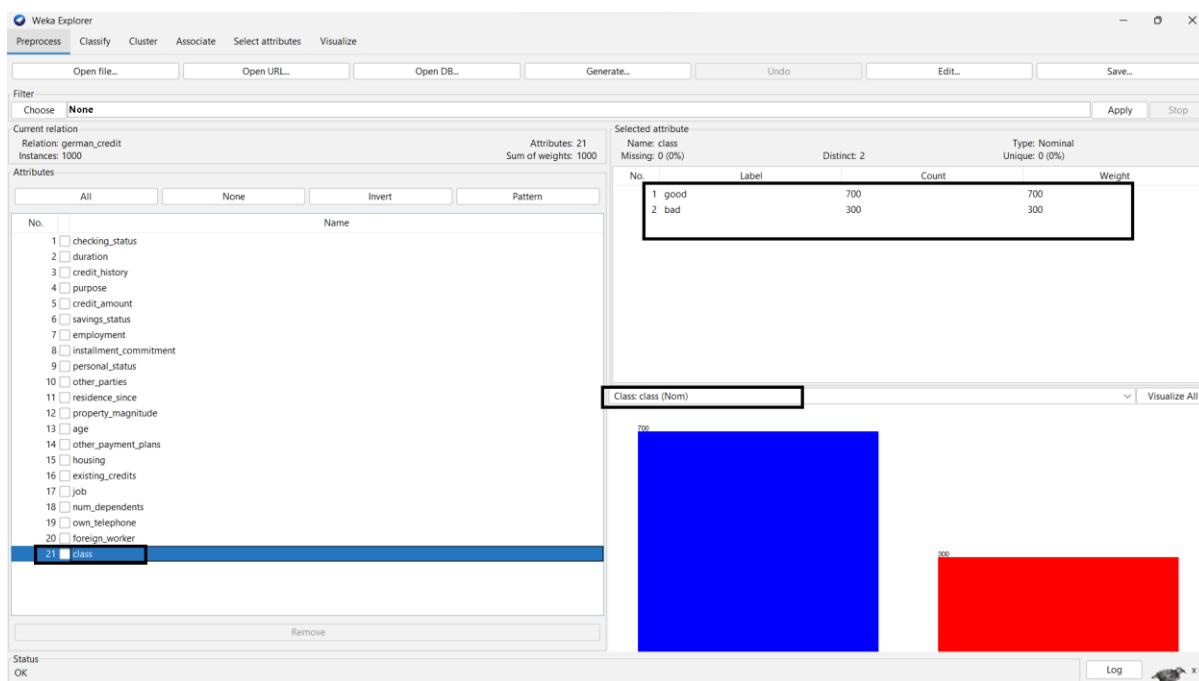
Current relation		Attributes: 21
Relation: german_credit		Sum of weights: 1000
Instances: 1000		

- Bộ dữ liệu này có 1000 mẫu dữ liệu và tổng cộng 21 thuộc tính.
- Dưới đây là 5 thuộc tính đầu tiên của bộ dữ liệu:

Tên thuộc tính	Kiểu dữ liệu	Loại thuộc tính	Ý nghĩa
checking_status	Nominal	Rời rạc	Trạng thái tại thời điểm hiện tại của tài khoản khách hàng.
duration	Numeric	Liên tục	Thời hạn tín dụng của tài khoản. (tính theo tháng)
credit_history	Nominal	Rời rạc	Lịch sử tín dụng của tài khoản.
purpose	Nominal	Rời rạc	Mục đích của việc vay mượn tín dụng.
credit_amount	Numeric	Liên tục	Số dư trong thẻ tín dụng.

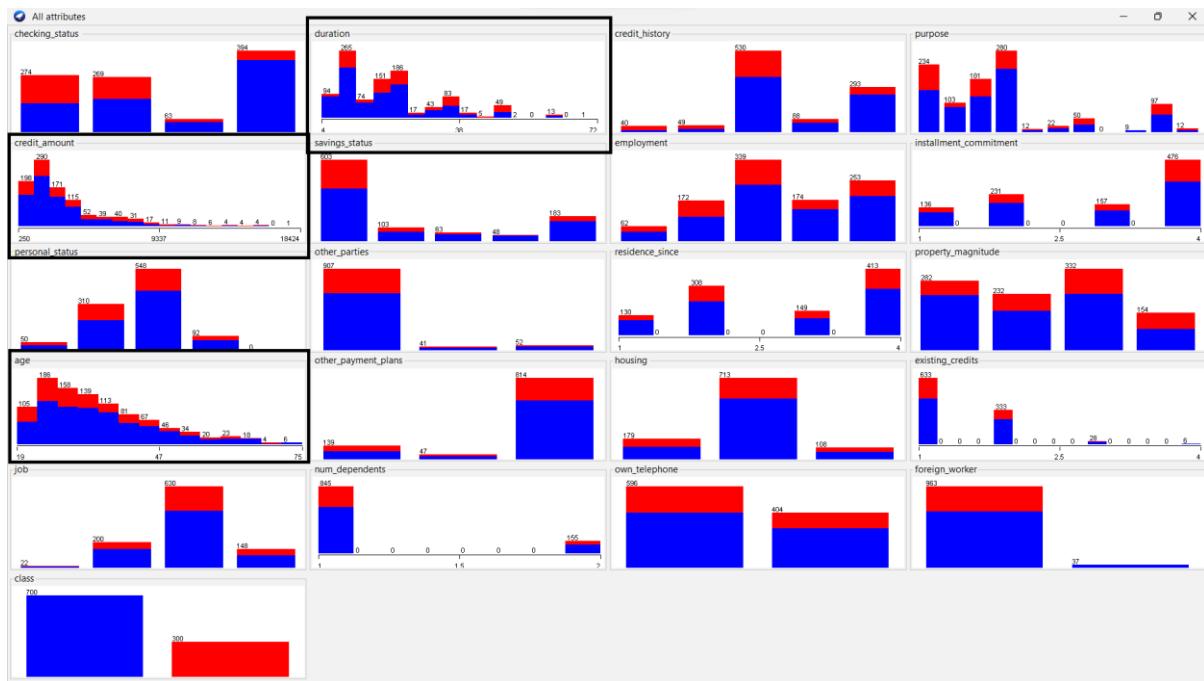
### b. Which attribute is used for the label?

- Thuộc tính được dùng làm nhãn là thuộc tính cuối cùng của bộ dữ liệu, có tên là **class** và có ý nghĩa là để xác định rủi ro tín dụng của khách hàng đó là tốt hay xấu. Nhìn vào hình ảnh bên dưới có thể thấy phân bố nghiêng hẳn về giá trị **good**.



Thuộc tính làm nhãn trong bộ dữ liệu **credit-g.arff**

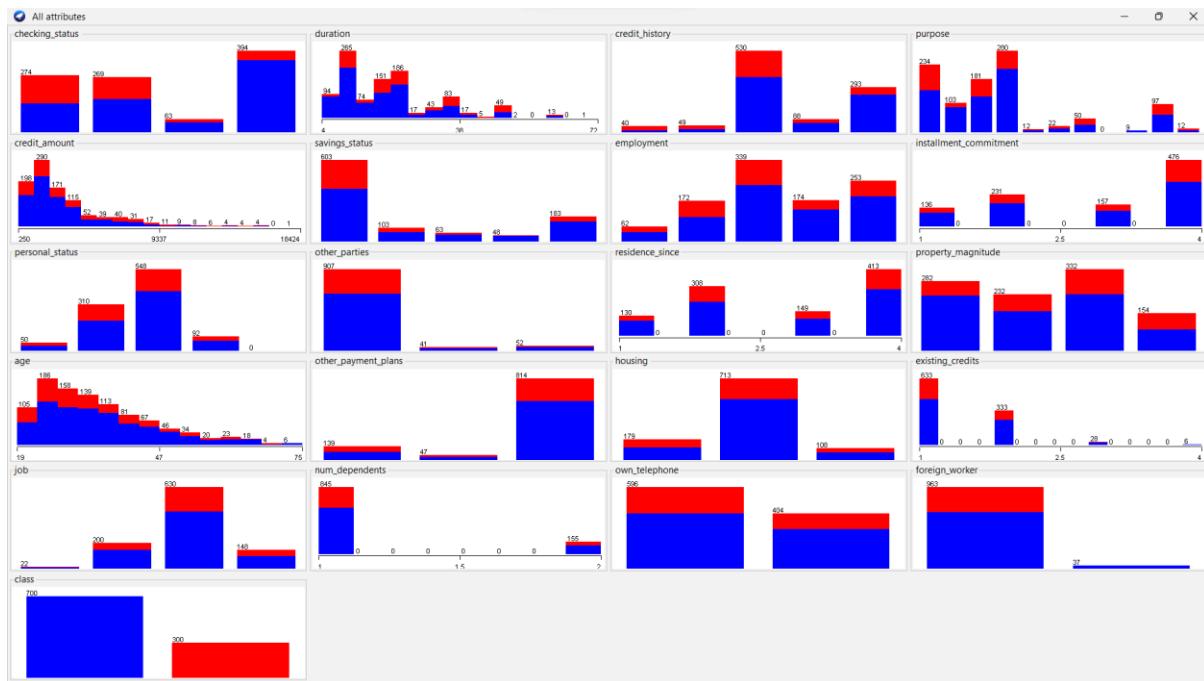
c. Let's describe the distribution of continuous attributes? (Left skewed or right skewed ?)



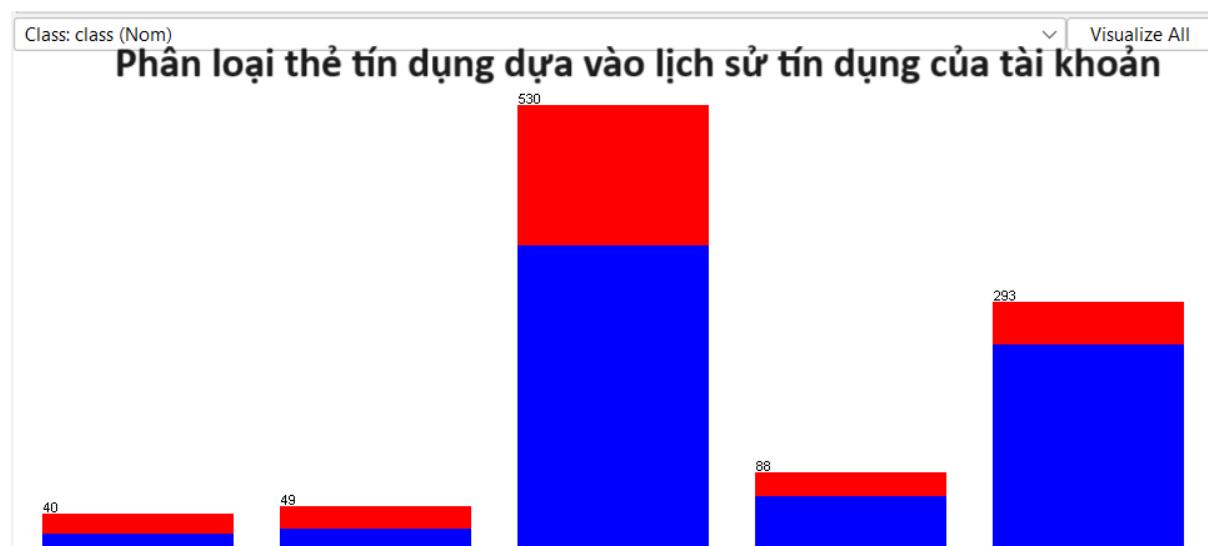
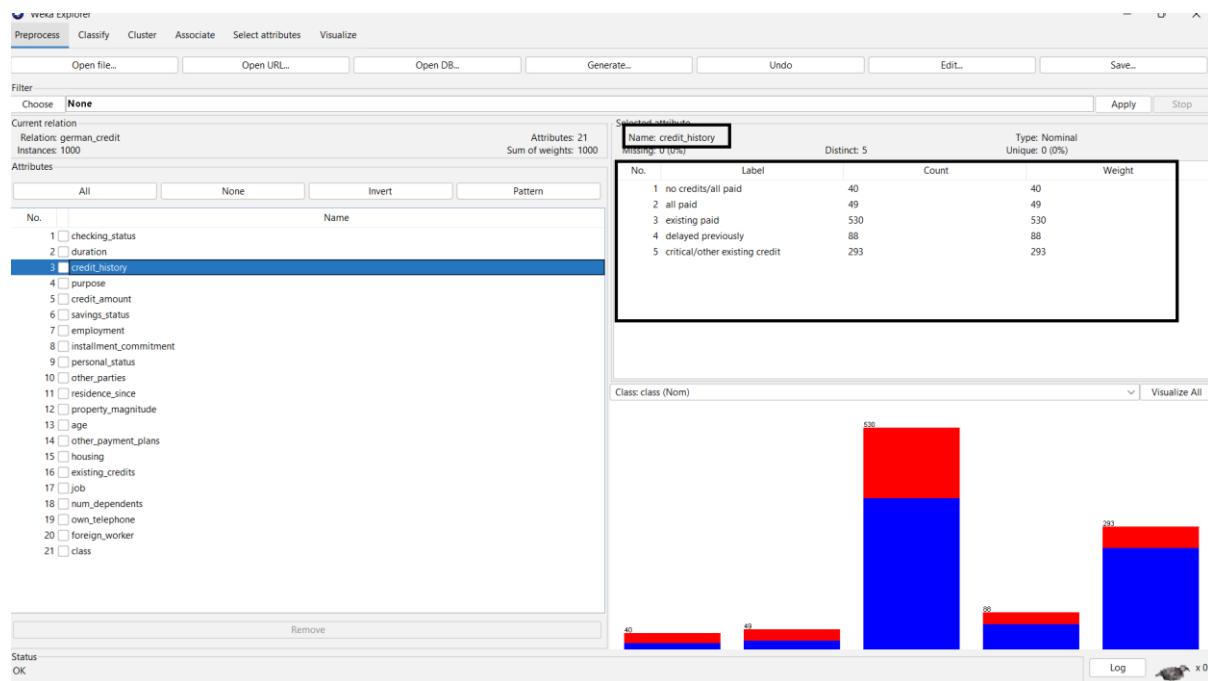
- Nhìn vào hình ảnh bên trên, có 3 thuộc tính thuộc kiểu thuộc tính liên tục là: duration, credit\_amount và age.

<b>duration</b>	Dữ liệu phân bố lệch trái
<b>credit_amount</b>	Dữ liệu phân bố lệch trái
<b>age</b>	Dữ liệu phân bố lệch trái

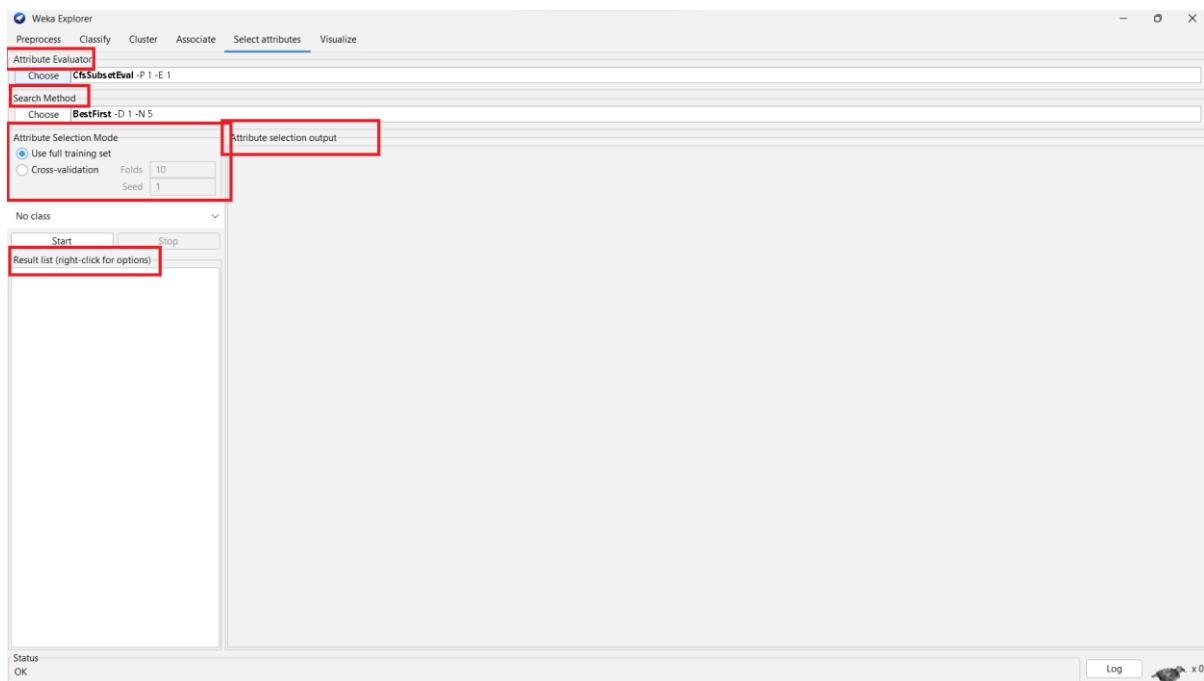
*d. Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.*



- Bộ dữ liệu này có 21 biểu đồ từ trái qua phải tương ứng cho các thuộc tính: checking\_status, duration, credit\_history, purpose, credit\_amount, .... Như đã giải thích ở các phần trên, các biểu đồ thể hiện sự tương quan giữa thuộc tính đang chọn và thuộc tính làm nhãn (thuộc tính Class). Trong thuộc tính Nominal, trục x sẽ là các cột giá trị và trục y sẽ là số lượng của từng cột giá trị. Trong thuộc tính Numeric, trục x sẽ là khoảng giá trị từ **Minimum** đến **Maximum** và trục y sẽ là tần số xuất hiện của các quyết định.
- Việc sử dụng màu sắc để biểu thị tỷ lệ số lượng một thuộc tính cụ thể (có giá trị **good** hoặc **bad**) trong ngữ cảnh của thuộc tính "class." Màu xanh thường sử dụng đại diện cho giá trị **good**, màu đỏ sử dụng đại diện cho giá trị **bad**.
- Dưới đây là thuộc tính **credit\_history**, nghĩa là phân loại thẻ tín dụng tốt hay xấu dựa vào lịch sử tín dụng của tài khoản.
- Từ trái qua phải, lần lượt là các giá trị phân biệt của thuộc tính **credit\_history**.
- Độ cao của từng cột giá trị trong thuộc tính là số lượng của giá trị đó.
- Màu sắc phân lớp cho giá trị.
  - o Tên biểu đồ: Phân loại thẻ tín dụng dựa vào lịch sử tín dụng của tài khoản.
  - o Chú thích:
    - Màu xanh: tín dụng **good**.
    - Màu đỏ: tín dụng **bad**.

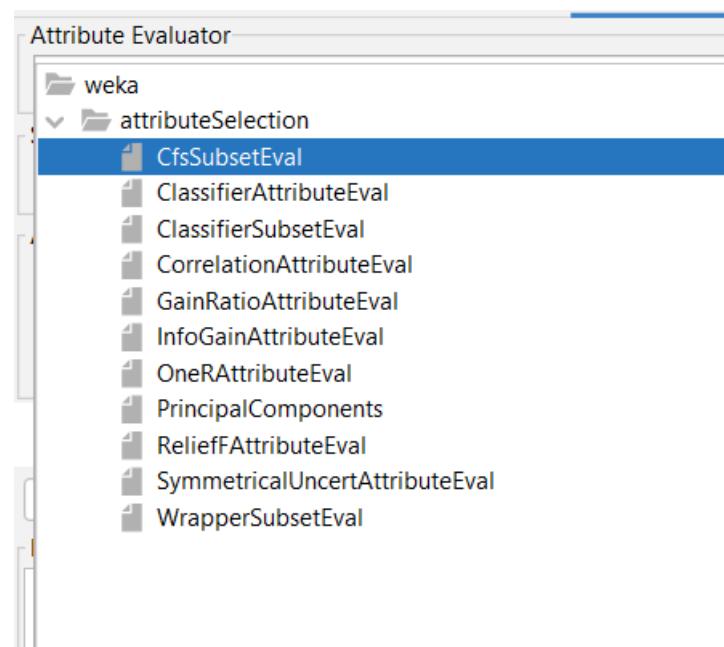


e. Let's move to the Select attributes tag. Describe all of the options for attribute selection.



*Selected attributes tag*

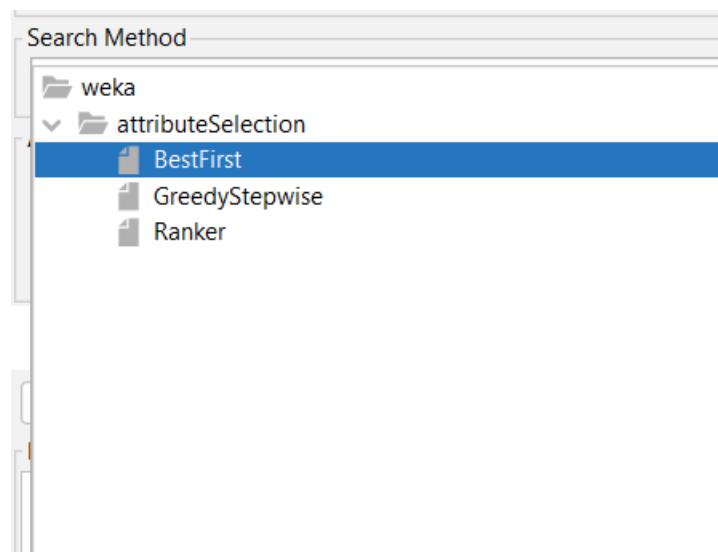
- Trình đánh giá thuộc tính (**Attribute Evaluator**) có 11 phương pháp dùng để đánh giá thuộc tính của bộ dữ liệu.



*Phương pháp đánh giá Attribute Evaluator*

Tên phương pháp	Mô tả
<b>CfsSubsetEval</b>	Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính độc lập và đo đặc sự dữ thừa giữa chúng.
<b>ClassifierSubsetEval</b>	Đánh giá các tập thuộc tính con trong tập huấn luyện hoặc tập kiểm tra một cách độc lập.
<b>ClassifierAttributeEval</b>	Đánh giá thuộc tính bằng cách sử dụng bộ phân loại được lựa chọn bởi người dùng.
<b>CorrelationAttributeEval</b>	Đánh giá một thuộc tính dựa trên mức độ tương quan với lớp nhãn.
<b>GainRatioAttributeEval</b>	Đánh giá một thuộc tính dựa trên tỷ lệ gia tăng.
<b>InfoGainAttributeEval</b>	Đánh giá một thuộc tính dựa trên thông tin thu thập.
<b>OneRAttributeEval</b>	Đánh giá một thuộc tính bằng cách sử dụng bộ phân loại OneR.
<b>PrincipalComponents</b>	Thực hiện phân tích thành phần chính và biến đổi dữ liệu.
<b>ReliefFAttributeEval</b>	Đánh giá một thuộc tính dựa trên các thể hiện.
<b>SymmetricalUncertAttributeEval</b>	Đánh giá một thuộc tính dựa trên tính bất đối xứng
<b>WrapperSubsetEval</b>	Đánh giá tập thuộc tính dựa trên một bộ phân loại và xác thực chéo.

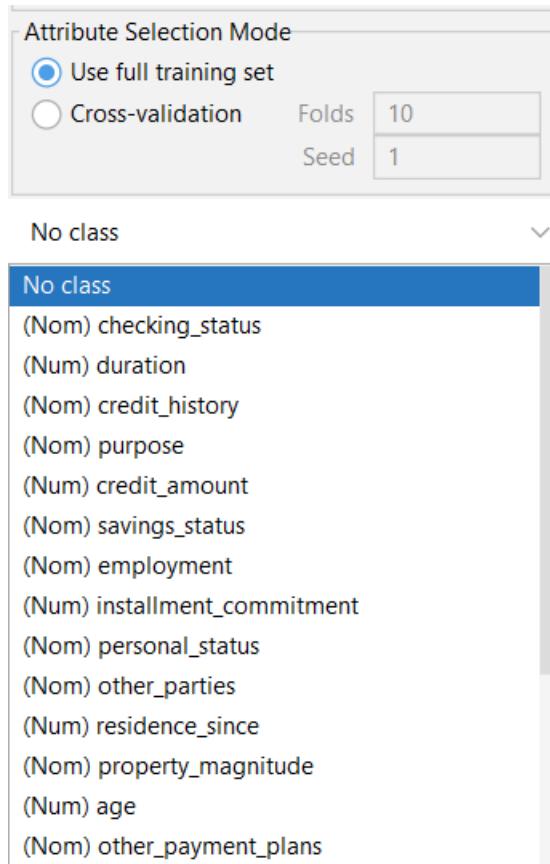
- Các phương pháp tìm kiếm (**Search Method**) có 3 phương pháp tìm kiếm.



*Phương pháp tìm kiếm Search Method*

Tên phương pháp	Mô tả
<b>BestFirst</b>	Thực hiện thuật toán leo đồi tham lam (GHC – Greedy Hill Climbing) kết hợp với quay lui (backtracking) có thể được thực hiện theo một số cách khác nhau. Chúng ta có thể tiến hành tìm kiếm "tiến" (forward) từ một tập hợp rỗng của các thuộc tính, hoặc "lui" (backward) từ tập hợp chứa toàn bộ các thuộc tính. Ngoài ra, cũng có thể bắt đầu từ một trạng thái cụ thể và thực hiện tìm kiếm theo hai hướng khác nhau.
<b>GreedyStepwise</b>	Thuật toán tìm kiếm tham lam trong không gian các tập thuộc tính có khả năng tiến và lui. Tuy nhiên, nó khác với quay lui ở điểm không tiếp tục tìm kiếm khi thêm hoặc xoá thuộc tính tốt nhất dẫn đến sự giảm của giá trị đánh giá.
<b>Ranker</b>	Đánh giá bằng xếp hạng của các thuộc tính và lựa chọn các thuộc tính bằng cách loại bỏ những thuộc tính có xếp hạng thấp.

- Chế độ chọn thuộc tính (Attribute Selection Mode) để lựa chọn thuộc tính làm lớp phân loại hoặc dự đoán.



### *Chế độ chọn thuộc tính Attribute Selection Mode*

- Sau khi lựa chọn phương pháp đánh giá, phương pháp tìm kiếm và thuộc tính làm nhãn thì nhấn vào phím **Start** bên dưới **Attribute Selection Mode**. Sau đó thông tin thực hiện sẽ hiện ở bên phải màn hình **Attribute selection output**.

```

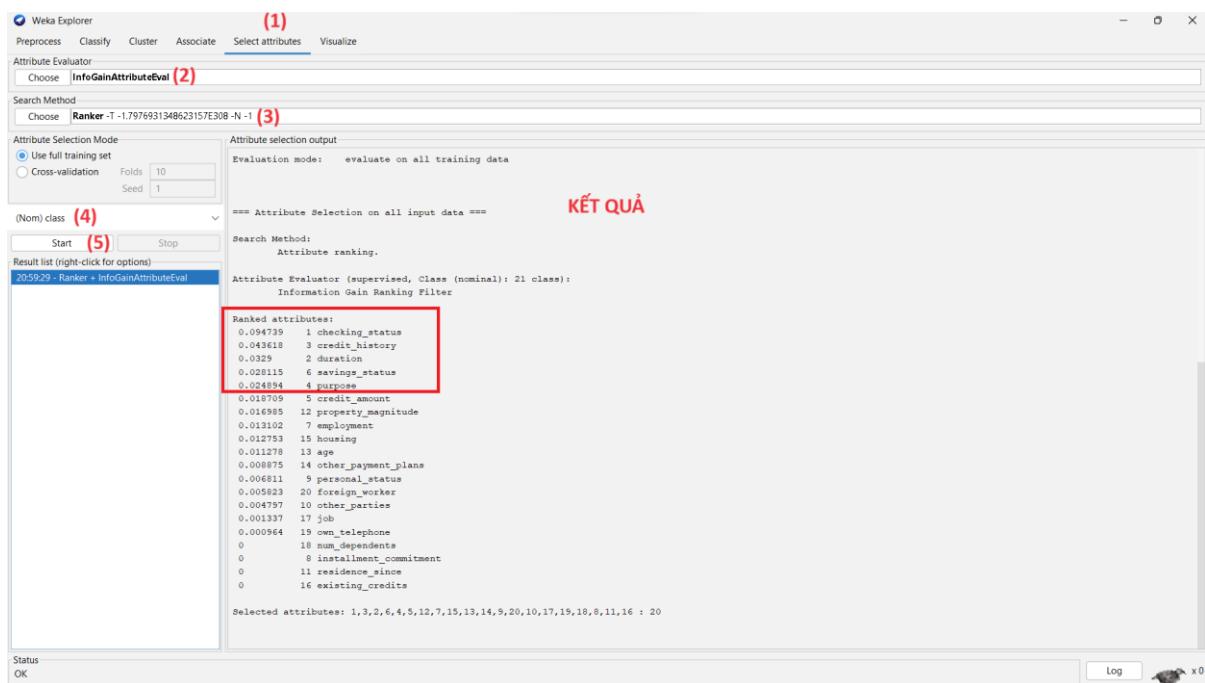
Attribute selection output
==== Run information ====
Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: german_credit
Instances: 1000
Attributes: 21
    checking_status
    duration
    credit_history
    purpose
    credit_amount
    savings_status
    employment
    installment_commitment
    personal_status
    other_parties
    residence_since
    property_magnitude
    age
    other_payment_plans
    housing
    existing_credits
    job
    num_dependents
    own_telephone
    foreign_worker
    class
Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ====
Search Method:
    Best first.
    Start set: no attributes

```

**f. Which options should be used to select the 5 attributes with the highest correlation? (Step-by-step description, with step-by-step photos and final results)**

- Nên sử dụng phương pháp đánh giá **InfoGainAttributeEval** và phương pháp tìm kiếm là **Ranker** để chọn ra 5 thuộc tính có độ tương quan cao nhất so với thuộc tính làm nhãn.
- Giải thích lí do: Phương pháp đánh giá **InfoGainAttributeEval** đo lường giá trị của một thuộc tính bằng cách tính toán thông tin liên quan đến lớp khi sử dụng phương pháp tìm kiếm **Ranker**. Kết quả sẽ là một danh sách xếp hạng độ tương quan của các thuộc tính với thuộc tính lớp, được sắp xếp từ cao đến thấp.



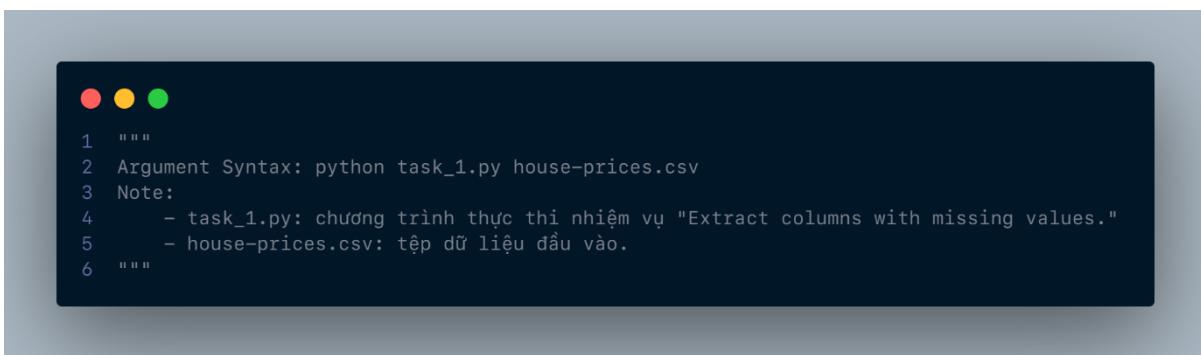
- Các bước tiến hành sẽ hiện trong hình ảnh theo thứ tự như sau:
  - Bước 1: Chọn tag **Select attributes**.
  - Bước 2: Chọn phương pháp đánh giá **InfoGainAttributeEval** ở **Attribute Evaluator**.
  - Bước 3: Chọn phương pháp tìm kiếm **Ranker** ở **Search Method**.
  - Bước 4: Chọn thuộc tính làm nhãn là **class** ở bên dưới **Attribute Selection Mode**.
  - Bước 5: Nhấn vào **Start** ở bên dưới **Attribute Selection Mode**.
  - Bước 6: Xem thông tin kết quả ở **Attribute selection output**.
- ➔ Như vậy, sử dụng phương pháp trên, chúng ta có thể xác định 5 thuộc tính có mối tương quan cao nhất với thuộc tính nhãn **class** theo thứ tự là: checking\_status, credit\_history, duration, savings\_status, và purpose.

### 3. Tiền xử lý dữ liệu trong Python

- Thông tin đặc tả quan trọng trước khi sử dụng các chương trình:
  - o Các chương trình sử dụng các tham số của dòng lệnh để xử lí và hoạt động trên Console/Terminal.
- Tổng cộng sẽ có 9 file thực thi các yêu cầu và 1 file dữ liệu đầu vào:
  - o house-prices.csv: File CSV dữ liệu đầu vào.
  - o processFile.py: Đọc và ghi file.
  - o task\_1.py: File thực thi yêu cầu 1.
  - o task\_2.py: File thực thi yêu cầu 2.
  - o task\_3.py: File thực thi yêu cầu 3.
  - o task\_4.py: File thực thi yêu cầu 4.
  - o task\_5.py: File thực thi yêu cầu 5.
  - o task\_6.py: File thực thi yêu cầu 6.
  - o task\_7.py: File thực thi yêu cầu 7.
  - o task\_8.py: File thực thi yêu cầu 8.
- File test các yêu cầu (dữ liệu đầu ra) thì luôn luôn có đuôi csv.
- Mỗi file sẽ yêu cầu số lượng tham số khác nhau.
- Trong bộ dữ liệu này, thuộc tính **PoolQC** thiếu 100% giá trị nên các yêu cầu của câu 3, câu 7, câu 8 không có cơ sở để thực hiện các yêu cầu này.

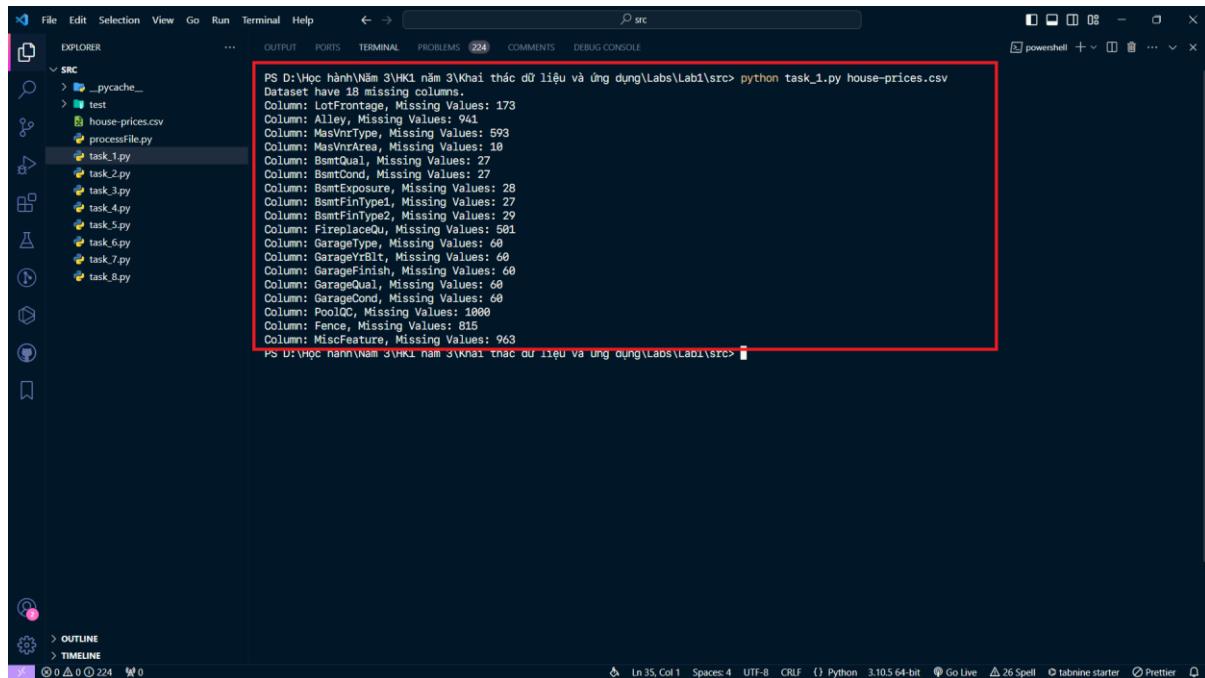
#### a. Extract columns with missing values

- Cú pháp chạy chương trình:



```
1 """
2 Argument Syntax: python task_1.py house-prices.csv
3 Note:
4     - task_1.py: chương trình thực thi nhiệm vụ "Extract columns with missing values."
5     - house-prices.csv: tệp dữ liệu đầu vào.
6 """
```

- Kết quả chạy chương trình:

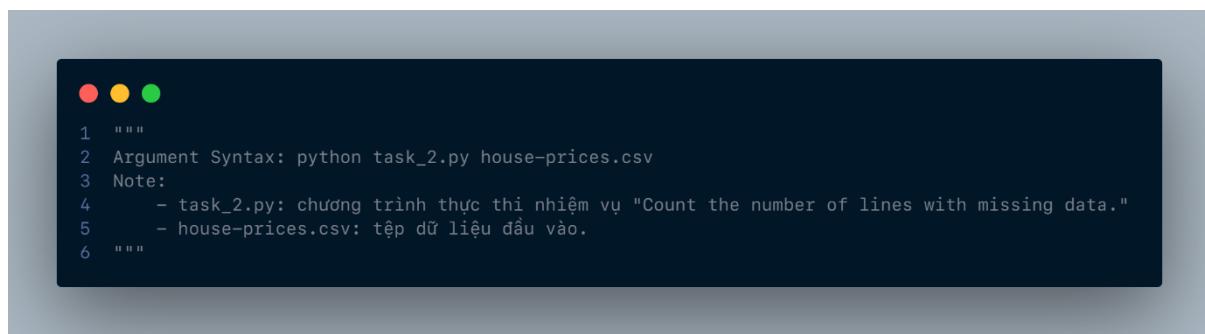


```
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\src> python task_1.py house-prices.csv
Dataset have 18 missing columns.
Column: LotFrontage, Missing Values: 173
Column: Alley, Missing Values: 941
Column: MasVnrType, Missing Values: 593
Column: MasVnrArea, Missing Values: 10
Column: BsmtQual, Missing Values: 27
Column: BsmtCond, Missing Values: 27
Column: BsmtExposure, Missing Values: 28
Column: BsmtFinType1, Missing Values: 27
Column: BsmtFinType2, Missing Values: 29
Column: FireplaceQu, Missing Values: 501
Column: GarageType, Missing Values: 60
Column: GarageYRBlt, Missing Values: 60
Column: GarageFinish, Missing Values: 60
Column: GarageQual, Missing Values: 60
Column: GarageCond, Missing Values: 60
Column: PoolQC, Missing Values: 1000
Column: Fence, Missing Values: 815
Column: MiscFeature, Missing Values: 963
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\src>
```

*Kết quả có 18 thuộc tính bị thiếu dữ liệu*

### **b. Count the number of lines with missing data.**

- Cú pháp chạy chương trình:



```
1 """
2 Argument Syntax: python task_2.py house-prices.csv
3 Note:
4     - task_2.py: chương trình thực thi nhiệm vụ "Count the number of lines with missing data."
5     - house-prices.csv: tệp dữ liệu đầu vào.
6 """
```

- Kết quả chạy chương trình:

```
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\src> python task_2.py house-prices.csv
The number of lines with missing data: 1000
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\src>
```

Kết quả có 1000 dòng bị thiếu dữ liệu

*c. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).*

- Áp dụng cột **LotFrontage** cho mean, median.
- Áp dụng cột **Alley** cho mode.
- Cú pháp chạy chương trình:

```
1 """
2 Argument Syntax: python task_3.py house-prices.csv --method=<method> --columns <col1> <col2> <col3> ... --out=<output file name>
3 Note:
4     - task_3.py: chương trình thực thi nhiệm vụ "Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute)."
5     - house-prices.csv: tệp dữ liệu đầu vào.
6     - --method: phương pháp áp dụng xử lý bài toán.
7         + mean: trung bình. (dành cho thuộc tính số)
8         + median: trung vị. (dành cho thuộc tính số)
9         + mode: giá trị có tần số xuất hiện nhiều nhất (dành cho thuộc tính định dạng, phân loại)
10    - --columns: các cột thuộc tính thiếu dữ liệu cần thêm vào.
11    - --out: tệp dữ liệu đầu ra. (để lưu dữ liệu mới)
12 """
```

- Kết quả chạy chương trình:

AutoSave off task\_3\_mean.csv • Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQua	OverallCor	YearBuilt	YearRemo	RoofStyle	RoomMat	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	53	6292	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950 Gable	CompShg	W	
6	208	20 RL	69.303506	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	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	CulSac	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	IDOTRB	Norm	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	69.303506	15498	Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976 Hip	WdShake	Sti	
17	848	20 RL	76	15523	Pave		IR1	Lvl	AllPub	CulSac	Gtl	ColligCr	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	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	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983 Gable	CompShg	Vii	
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	ColligCr	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	

### Kết quả sử dụng hàm MEAN

AutoSave off task\_3\_median.csv • Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQua	OverallCor	YearBuilt	YearRemo	RoofStyle	RoomMat	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	53	6292	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1950 Gable	CompShg	W	
6	208	20 RL	68	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	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	CulSac	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	IDOTRB	Norm	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	Edward	Norm	Norm	1Fam	1Story	4	5	2005	2006 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	68	15498	Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1976 Hip	WdShake	Sti	
17	848	20 RL	36	15523	Pave		IR1	Lvl	AllPub	CulSac	Gtl	ColligCr	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	TwnhsE	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	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	Corner	Gtl	Mitchel	Norm	Norm	1Fam	1Story	6	5	1983	1983 Gable	CompShg	Vii	
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	ColligCr	Norm	Norm	TwnhsE	1Story	6	5	2003	2006 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	

### Kết quả sử dụng hàm MEDIAN

### Kết quả sử dụng hàm MODE

- d. Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).

- Cú pháp chạy chương trình:

```

1 """
2 Argument Syntax: python task_4.py house-prices.csv --threshold=values_in [0,1]> --out=output_file_name
3 Note:
4     task_4.py: chương trình thực hiện việc "Deleting rows containing more than a particular number of missing values (example: delete rows with the number of missing values is more than 50% of the number of attributes)."
5     - house-prices.csv: tập dữ liệu đầu vào.
6     - --threshold: tỷ lệ (ngưỡng) để xóa giá trị. (ĐÁT MỤC NẮM TRONG KHÔNG [0,1])
7     --out: tập dữ liệu đầu ra. (để lưu dữ liệu mới)
8 """

```

- Kết quả chạy chương trình:

- o Ban đầu có 1000 dòng bị thiếu dữ liệu (ở câu b), sau khi xóa các dòng thiếu dữ liệu thì còn lại 920 dòng.

```
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\21127099_21127411\Source> python task_4.py house-prices.csv --threshold=0.1 -oout=task_4_test_10%.csv
The number of lines with missing data: 920
PS D:\Học hành\Năm 3\HK1 năm 3\Khai thác dữ liệu và ứng dụng\Labs\Lab1\21127099_21127411\Source>
```

Kết quả còn lại 920 dòng bị thiếu dữ liệu

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
900	333	20 RL		85	10655 Pave		IR1	Lvl	AllPub	Inside	Gtl	NridgeHt	Norm	Norm	1Fam	1Story	8	5	2003	2004 Gable	CompShg Vi	
901	414	30 RM		56	8960 Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1Story	5	6	1927	1950 Gable	CompShg W	
902	254	80 RL		85	9350 Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	SLvl	6	7	1964	1991 Hip	CompShg Hc	
903	314	20 RL		150	215245 Pave		IR3	Low	AllPub	Inside	Sev	Timber	Norm	Norm	1Fam	1Story	7	5	1965	1965 Hip	CompShg Br	
904	174	20 RL		80	10197 Pave		IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	6	5	1961	1961 Gable	CompShg W	
905	1236	70 RL		96	13132 Pave		Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	5	5	1914	1950 Gable	CompShg W	
906	979	20 RL		68	9450 Pave		Reg	Bnk	AllPub	Inside	Mod	Edward	Norm	Norm	1Fam	1Story	4	5	1954	1954 Gable	CompShg M	
907	213	60 FV		72	8640 Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5	2009	2009 Gable	CompShg Vi	
908	458	20 RL			53227 Pave		IR1	Low	AllPub	CulDSac	Mod	ClearCr	Norm	Norm	1Fam	1Story	4	6	1954	1994 Flat	Tar&Grv Ph	
909	62	75 RM		60	7200 Pave		Reg	Lvl	AllPub	Inside	Gtl	IDOTTR	Norm	Norm	1Fam	2.5Unf	5	7	1920	1996 Gable	CompShg M	
910	826	20 RL		114	14803 Pave		Reg	Lvl	AllPub	Inside	Gtl	NridgeHt	PosN	PosN	1Fam	1Story	10	5	2007	2008 Hip	CompShg Ce	
911	1253	20 RL		62	9858 Pave		Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	5	6	1968	1968 Gable	CompShg Hc	
912	1053	60 RL		100	9500 Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Artery	Norm	1Fam	2Story	6	6	1964	1978 Gable	CompShg Vi	
913	582	20 RL		98	12704 Pave		Reg	Lvl	AllPub	Inside	Gtl	NridgeHt	Norm	Norm	1Fam	1Story	8	5	2008	2009 Hip	CompShg Vi	
914	1420	20 RL			16381 Pave		IR1	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	1Story	6	5	1969	1969 Gable	CompShg Ph	
915	1417	190 RM		60	11340 Pave		Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon	2Story	4	6	1885	1950 Gable	CompShg Vi	
916	668	20 RL		65	8125 Pave		Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	1Story	6	5	1994	1998 Gable	CompShg Hc	
917	1190	60 RL		60	7500 Pave		Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	1999	1999 Gable	CompShg Vi	
918	192	60 RL			7472 Pave		IR1	Lvl	AllPub	CulDSac	Gtl	mes	Norm	Norm	1Fam	2Story	7	9	1972	2004 Gable	CompShg Hc	
919	990	60 FV		65	8125 Pave		Reg	Lvl	AllPub	Inside	Gtl	Somers	Norm	Norm	1Fam	2Story	7	5	2006	2006 Gable	CompShg Vi	
920	982	60 RL		98	12703 Pave		IR1	Lvl	AllPub	Corner	Gtl	NoBridge	Norm	Norm	1Fam	2Story	9	5	1999	1999 Hip	CompShg Vi	
921	862	190 RL		75	11625 Pave		Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	2fmCon	1Story	5	4	1965	1965 Hip	CompShg Ph	

Minh chứng bằng file CSV kết quả (920 dòng thiếu dữ liệu + 1 dòng tên thuộc tính)

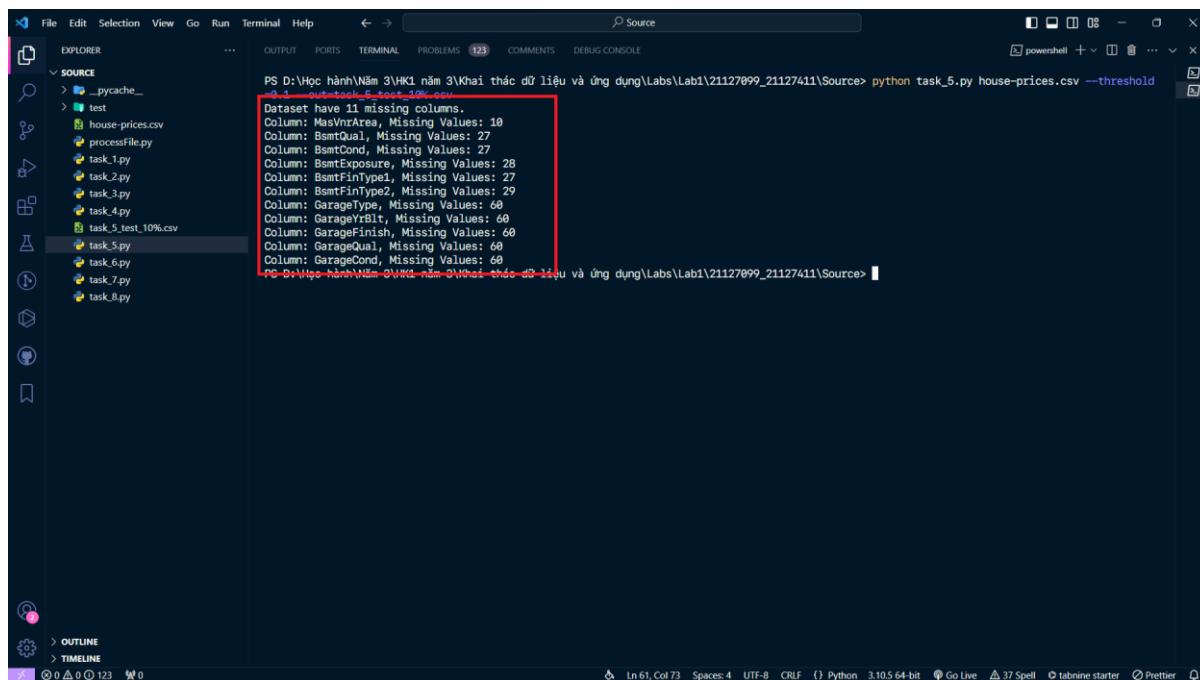
e. Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).

- Cú pháp chạy chương trình:

```
1 ...
2 Argument Syntax: python task 5.py house-prices.csv --threshold=<value in [0,1]> --out=<output file name>
3 Hint:
4   - task_5.py: chương trình thực thi nhiệm vụ "Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples)."
5   - house-prices.csv: tệp dữ liệu dữ liệu về giá nhà
6   - house-prices-test.csv: tệp dữ liệu dữ liệu để kiểm tra (mất nước nằm trong khoảng [0,1])
7   - --out: tệp dữ liệu đầu ra (không dữ liệu null)
8 ...
9
```

- #### - Kết quả chạy chương trình:

- Ban đầu có 18 thuộc tính bị thiếu dữ liệu (ở câu a), sau khi xóa các cột thiếu dữ liệu thì còn lại 11 thuộc tính bị thiếu dữ liệu.

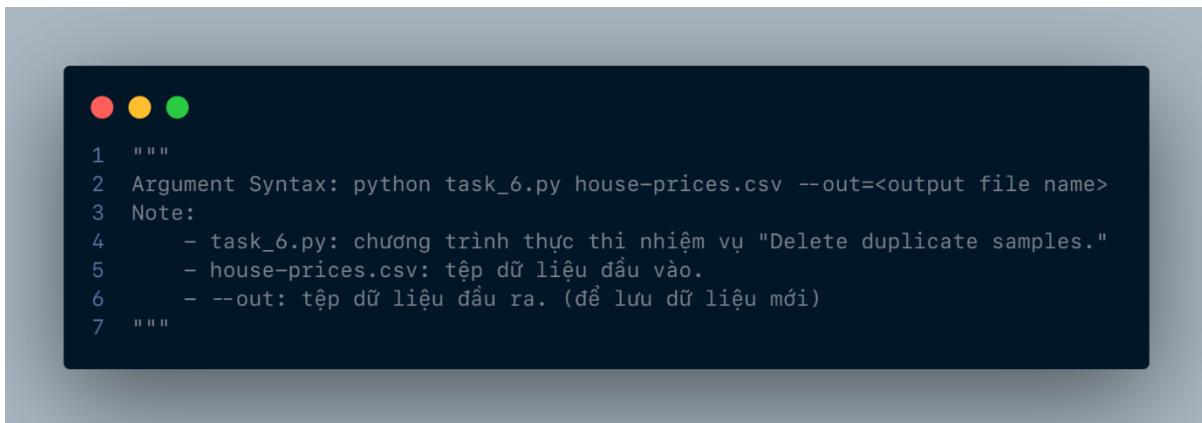


Kết quả còn lại 11 thuộc tính bị thiếu dữ liệu

*Minh chứng bằng file CSV kết quả*

### f. Delete duplicate samples.

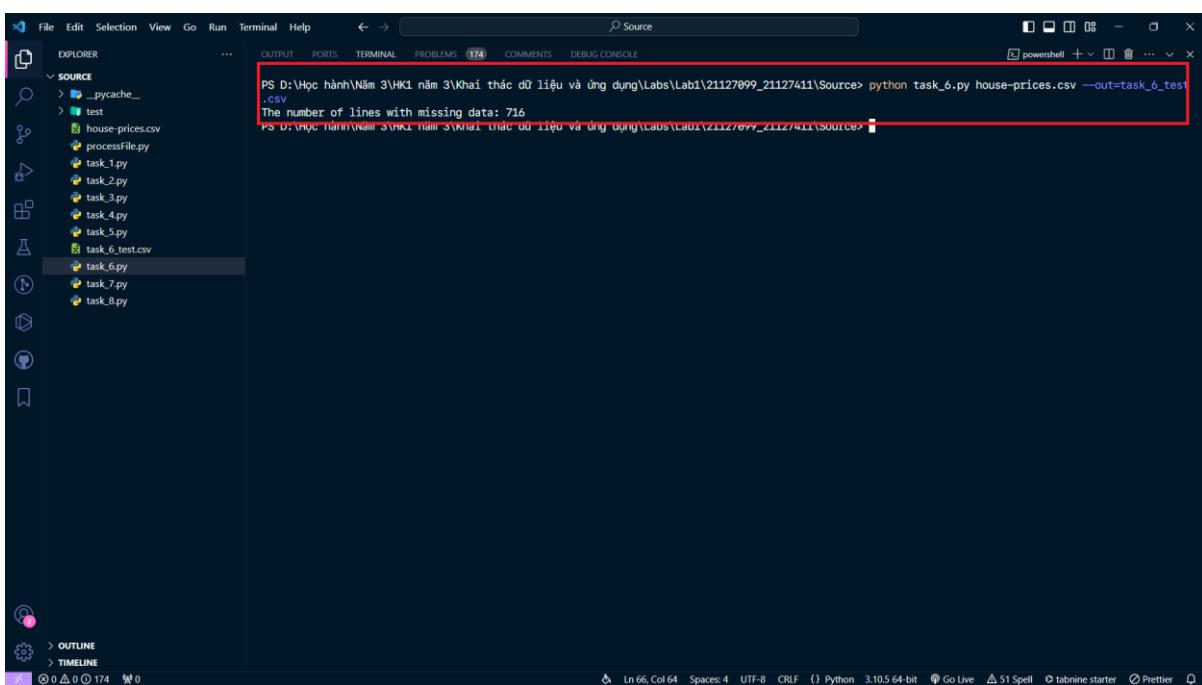
- Cú pháp chạy chương trình:



```
1 """
2 Argument Syntax: python task_6.py house-prices.csv --out=<output file name>
3 Note:
4     - task_6.py: chương trình thực thi nhiệm vụ "Delete duplicate samples."
5     - house-prices.csv: tệp dữ liệu đầu vào.
6     - --out: tệp dữ liệu đầu ra. (để lưu dữ liệu mới)
7 """
```

- Kết quả chạy chương trình:

- o Ban đầu có 1000 dòng bị thiếu dữ liệu (ở câu b), sau khi xóa các dòng trùng lặp thì còn lại 716 dòng.



The number of lines with missing data: 716

Kết quả còn lại 716 dòng bị thiếu dữ liệu

The screenshot shows an Excel spreadsheet titled 'task\_6\_test.csv'. The first row contains column headers: Id, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W. The subsequent 716 rows contain data for houses, with columns including LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, KitchenQual, TotRmsAbvGrd, FullBath, HalfBath, Fireplaces, GarageCars, GarageYrBlt, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, WoodDeckSF, OpenPorchSF, EnclosedPorch, PorchArea, ScreenPorch, MiscVal, SaleType, SaleCondition, and SalePrice.

*Minh chứng bằng file CSV kết quả (716 dòng thiếu dữ liệu + 1 dòng tên thuộc tính)*

#### **g. Normalize a numeric attribute using min-max and Z-score methods.**

- Áp dụng cột **Id**, **LotFrontage**, **MSSubClass**, **LotArea** cho minmax và zscore.
- Cú pháp chạy chương trình:

```

1 """
2 Argument Syntax: python task_7.py house-prices.csv --method=<operation> --columns <col1> <col2> ... --out=<output file name>
3 Note:
4     - task_7.py: chương trình thực thi nhiệm vụ "Normalize a numeric attribute using min-max and Z-score methods."
5     - house-prices.csv: tập dữ liệu đầu vào.
6     - --method: phương pháp áp dụng xử lý bài toán.
7         + minmax.
8         + zscore.
9     - --columns: các cột thuộc tính số cần thực hiện bài toán.
10    - --out: tập dữ liệu đầu ra. (để lưu dữ liệu mới)
11 """

```

- Kết quả chạy chương trình:

AutoSave off task\_7\_test\_minmax... Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A1 A B C D E F G H I J K L M N O P Q R S T U V W

**Grvl**

ID	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCor	YearBuilt	YearRemo	RoofStyle	RoomMat	Ex
0.85048	0	R NL	0.469669	0.0391639	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	1Fam	1Story	7	6	2007	Hip	CompShg	Vrl	
0.8443072	0	0.4117647	0.0311211	0.0391131	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	1Fam	Duplex	4	5	1962	1962	Gable	CompShg Hc	
0.9595336	0.1764705	RM	0.219669	0.021158	Pave		Reg	Bnk	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg W	
0.9430727	0.0588235	RM	0.2348484	0.0225244	Pave		Reg	Lvl	AllPub	Inside	Gtl	SWISU	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg W	
0.1412894			0.0515325	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1960	1961	Gable	CompShg W	
0.953361	0.4117647	RM	0.3333333	0.0349303	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg Ph	
0.6707818	0	0 RL	0.4469696	0.0343316	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	1Fam	1Story	5	6	1963	1963	Gable	CompShg Vil	
0.330598	0.5882352	RM	0.083333	0.0141414	Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg Vil	
0.2674897	0.3529411	RM	0.3787878	0.0502639	Pave		IR1	Lvl	AllPub	CubSdC	Gtl	Mitchel	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg Vil	
0.4993141	0.0588235	RM	0.2348484	0.022281	Pave		Reg	Lvl	AllPub	Inside	Gtl	IDOTRB	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg M	
0.1735253	0 RL	0.3712121	0.032386	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg M		
0.7489711	0 RL	0.3787878	0.0362682	Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	1Fam	1Story	5	8	1965	1998	Hip	CompShg M		
0.6990926	0 RL	0.2954545	0.0259486	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	5	2005	2005	Gable	CompShg Vil		
0.9183813	0 RL	0.3712121	0.03189	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg M		
0.7016460	0 RL	0.065599					IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake St	
0.5802469	0 RL	0.1136363	0.065707	Pave		IR1	Lvl	AllPub	CubSdC	Gtl	CollCr	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg Hc		
0.3120713	0.2941176	RM	0.098485	0.0144736	Pave		Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	1Fam	2Story	5	5	1916	1950	Gable	CompShg As	
0.8669410	0.8235294	FV	0.1060606	0.0105628	Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg M	
0.4753086	0.1764705	RM	0.2272727	0.021172	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg W	
0.0105891	0.5882352	RM	0.1747424	0.0128503	Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg Ce	
0.8998628	0.3529411	RM	0.659096	0.062203	Pave		IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg Vil	
0.3511659			0.3787878	0.0360671	Pave		Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg Vil	
0.7311385	0.3529411	RM	0.4469696	0.038748	Pave		Reg	Lvl	AllPub	Inside	Mod	mes	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg Hc	
0.9746227	0.5882352	RM	0.1212121	0.0138374	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	TwnhsE	1Story	6	5	2003	2003	Gable	CompShg Vil	
0.8614540	0.0588235	RM	0.2651515	0.0120883	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	1Fam	1Story	5	8	1922	1950	Gable	CompShg W	
0.4238683	0.3529411	RL	0.4848484	0.050368	Pave		Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg Vil	
0.830598	0.0588235	RL	0.2196696	0.036783	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	6	1941	1950	Hip	CompShg M	

task\_7 test\_minmax

### Kết quả sử dụng phương thức MINMAX

AutoSave off task\_7\_test\_zscore... Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A1 A B C D E F G H I J K L M N O P Q R S T U V W

**Grvl**

ID	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContac	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCor	YearBuilt	YearRemo	RoofStyle	RoomMat	Ex
1.1962637	-0.84868	RL	0.6442436	-0.03953	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	1Fam	1Story	7	6	2007	2007	Hip	CompShg Vil	
1.1743494	0.7856388	RL	0.0327610	-0.04029	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Duplex	1Story	4	5	1962	1962	Gable	CompShg Hc	
1.5834159	-0.48286	RM	-0.90798	-0.45877	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	6	7	1929	1950	Gable	CompShg W	
1.5249778	-0.6152	RL	-0.81391	-0.42697	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	1Fam	1Story	6	5	1930	1950	Gable	CompShg W	
1.321244	-0.84868	RL	0.2484680		Pave		IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1960	1960	Gable	CompShg W	
1.5615016	0.7856388	RL	-0.20242	-0.1381	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Duplex	1Story	5	5	1967	1967	Gable	CompShg Ph	
0.5583148	-0.84868	RL	0.5031322	-0.15205	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	1Fam	1Story	5	6	1963	1963	Gable	CompShg Vil	
0.-0.64941	1.8860598	RM	-1.75465	-0.62216	Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Twnhs	1Story	6	5	1998	1998	Hip	CompShg Vil	
0.-0.87342	0.85218	RM	0.07978	0.2175336	Pave		IR1	Lvl	AllPub	CubSdC	Gtl	Mitchel	Norm	1Fam	2Story	6	5	2001	2002	Gable	CompShg Vil	
1.-0.05042	-0.6152	RM	0.81391	-0.43263	Pave		Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRB	Norm	1Fam	1.5Fin	4	5	1925	1950	Gable	CompShg M
2.-1.207	-0.84868	RL	0.0327610	-0.19736	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	5	6	1957	1957	Gable	CompShg M	
3.0835895	-0.84868	RL	0.07978	-0.10695	Pave		Reg	Lvl	AllPub	Corner	Gtl	Feedr	Norm	1Fam	1Story	5	8	1965	1998	Hip	CompShg M	
4.0581465	-0.84868	RL	-0.43761	-0.34724	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edward	Norm	1Fam	1Story	4	5	2005	2006	Gable	CompShg Vil	
5.04373207	-0.84868	RL	0.0327610	-0.20895	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg M	
6.06678862	-0.84868	RL	0.5757840		Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	1Fam	1Story	8	6	1976	1976	Hip	WdShake St	
7.0369054	-0.84868	RL	-1.56665	0.5785075	Pave		IR1	Lvl	AllPub	CubSdC	Gtl	CollCr	Norm	1Fam	1Story	5	6	1972	1972	Gable	CompShg Hc	
8.-0.71515	0.1169615	RM	-1.66058	-0.61443	Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg Ce	
9.02547018	0.4199544	VL	-1.61354	-0.70549	Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	TwnhsE	2Story	7	5	1999	1999	Hip	CompShg As	
0.-0.13564	-0.14826	RM	-0.86094	-0.44547	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	1Fam	1.5Fin	5	6	1936	1950	Gable	CompShg W	
1.176947	0.4860598	RM	-1.1902	0.65322	Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	TwnhsE	1Story	5	7	1976	1976	Gable	CompShg Ce	
2.3715779	0.085218	RM	1.8201710	0.4966235	Pave		IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	1Fam	2Story	9	5	1999	1999	Gable	CompShg Vil	
3.-0.57636	-0.84868	RL	0.07978	-0.11163	Pave		Reg	Bnk	AllPub	Corner	Gtl	Mitchel	Norm	1Fam	1Story	6	5	1983	1983	Gable	CompShg Vil	
4.07725877	0.085218	RM	0.5031322	-0.04922	Pave		Reg	Lvl	AllPub	Inside	Mod	mes	Norm	1Fam	2Story	6	6	1964	1964	Gable	CompShg Hc	
5.13639841	0.4860598	RM	-1.51946	-0.62924	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollCr	Norm	TwnhsE	1Story	6	5	2003	2006	Gable	CompShg Vil	
6.1235224	-0.6152	RM	-0.62576	-0.67009	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edward	Feedr	1Fam	1Story	5	8	1922	1950	Gable	CompShg W	
7.-0.31826	0.085218	RL	0.7383179	0.2113460	Pave		Reg	Lvl	AllPub	Inside	Gtl	Timber	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg Vil	
8.1256510	-0.6152	RM	-0.90795	-0.09497	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	1Fam	1Story	4	6	1941	1950	Hip	CompShg M	

task\_7 test\_zscore

### Kết quả sử dụng phương thức ZSCORE

#### h. Performing addition, subtraction, multiplication, and division between two numerical attributes.

- Áp dụng cột Id, LotFrontage cho +, -, \*, /.
- Các dòng thiếu dữ liệu khi tính toán sẽ ra kết quả là giá trị nan.
- Cột cuối cùng result là cột kết quả sau khi thực hiện phép tính.

- Cú pháp chạy chương trình:

```

1 """
2 Argument Syntax: python task_8.py house-prices.csv --method=<operation> --columns <col1> <col2> --out=<output file name>
3 Note:
4     - task_8.py: chương trình thực thi nhiệm vụ "Performing addition, subtraction, multiplication, and division between two numerical attributes."
5     - house-prices.csv: tập dữ liệu đầu vào.
6     - --method: phương pháp áp dụng xử lý bài toán.
7         + add or +: phép cộng.
8         + sub or -: phép trừ.
9         + mul or *: phép nhân.
10        + div or /: phép chia.
11    - --columns: 2 cột thuộc tính số cần thực hiện bài toán.
12    - --out: tập dữ liệu đầu ra. (để lưu dữ liệu mới)
13 """

```

- Kết quả chạy chương trình:

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	
1	Yc	GarageYB	GarageFin	GarageCar	GarageAre	GarageQui	GarageCor	PavedDriv	WoodDeck	OpenPorch	EnclosedP	35snPorch	ScreenPori	PoolArea	PoolQC	Fence	MiscFeatu	MiscVal	MoSold	YrsSold	SaleType	SaleCond	SalePrice	result
2	2007	Rfn	3	954	TA	TA	Y	0	56	0	0	0	0	0	0	0	0	0	6	2007	New	Partial	248323	1325
3	1962	Unf	2	462	TA	TA	Y	0	0	0	0	0	0	0	0	0	0	0	3	2007	WD	Normal	101800	1303
4	1929	Rfn	1	208	TA	TA	Y	0	0	112	0	0	0	0	0	0	0	0	7	2008	WD	Normal	120000	1451
5	1925	Unf	1	160	Fa	TA	Y	0	141	0	0	0	0	0	0	0	0	0	4	2008	WD	Normal	91000	1429
6	1960	Rfn	1	312	TA	TA	Y	355	0	0	0	0	0	0	0	GdWo	0	4	2008	WD	Normal	141000	nan	
7	1967	Unf	3	792	TA	TA	Y	0	152	0	0	0	0	0	0	0	0	0	4	2009	WD	Normal	124000	1457
8	1963	Unf	2	480	TA	TA	Y	0	80	0	0	0	0	0	0	MnPrv	0	6	2009	WD	Normal	139000	1060	
9	1998	Unf	2	402	TA	TA	Y	0	125	0	0	0	0	0	0	0	0	5	2006	WD	Normal	164000	516	
10	2001	Fin	2	560	TA	TA	Y	125	192	0	0	0	0	0	0	0	0	6	2009	WD	Normal	215000	463	
11	1962	Unf	2	539	TA	TA	Y	0	23	112	0	0	0	0	0	0	0	0	1	2009	WD	Normal	103000	782
12	1957	Rfn	1	294	TA	TA	Y	250	0	0	0	0	0	0	0	0	0	6	2010	WD	Normal	145000	325	
13	1977	Unf	2	884	TA	TA	Y	0	64	0	0	0	0	0	0	MnPrv	0	10	2006	WD	Normal	146000	1165	
14	2005	Fin	2	451	TA	TA	Y	252	64	0	0	0	0	0	0	0	0	6	2008	WD	Normal	176000	1081	
15	1974	Unf	4	480	TA	TA	Y	0	0	0	0	0	0	0	0	GdWo	0	6	2007	WD	Normal	123000	1411	
16	1976	Fin	2	665	TA	TA	Y	0	72	174	0	0	0	0	0	0	0	5	2008	COD	Abnorml	287000	nan	
17	1972	Unf	1	338	TA	TA	Y	0	0	0	0	0	0	0	0	0	0	8	2009	WD	Normal	133500	884	
18	1916	Unf	3	513	Fa	Y	0	0	96	0	0	0	0	0	0	0	0	5	2008	COD	Abnorml	98000	491	
19	1999	Unf	2	506	TA	TA	Y	0	34	0	0	0	0	0	0	0	0	3	2008	WD	Normal	183900	1301	
20	1995	Unf	2	576	TA	TA	Y	112	0	0	0	0	0	0	0	MnPrv	0	4	2009	WD	Normal	141500	746	
21	1976	Unf	2	572	TA	TA	Y	100	110	0	0	0	0	0	0	0	0	6	2007	WD	Normal	129900	68	
22	1999	Fin	3	779	TA	TA	Y	668	30	0	0	0	0	0	0	0	0	5	2010	WD	Normal	333168	1422	
23	1983	Unf	2	484	TA	TA	Y	120	0	158	0	0	0	0	0	0	0	6	2007	WD	Normal	134000	585	
24	1964	Rfn	2	442	TA	TA	Y	328	128	0	0	189	0	0	0	0	0	6	2008	WD	Normal	167900	1148	
25	2003	Fin	2	420	TA	TA	Y	140	0	0	0	0	0	0	0	0	0	3	2008	WD	Normal	136500	1460	
26	0	0	0	0	Y	0	96	0	0	0	0	0	0	0	0	0	0	7	2009	WD	Normal	99900	1314	
27	2003	Fin	3	749	TA	TA	Y	168	0	0	0	0	0	0	0	0	0	8	2008	WD	Normal	305000	705	
28	1941	Unf	1	234	TA	TA	N	0	113	0	0	0	0	0	0	0	0	8	2009	WD	Normal	113000	1263	

Kết quả sử dụng phép cộng

AutoSave off task\_8\_test\_subtraction... Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A1 BH BI BJ BK BL BM BN BO BP BQ BR BS BT BU BV BW BX BY BZ CA CB CC CD result

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	result
1	Yr: GarageYrB GarageFini GarageCar GarageAre GarageCor PavedDriv WoodDeck OpenPorch EnclosedPorch ScreenPorch PorchArea PoolQC Fence MiscFeatu MiscVal MoSold YrsOld SaleType SaleCondit SalePrice																							
2	2007 RFn	3	954 TA	TA	Y	0	56	0	0	0	0	0	0					0	6	2007 New	Partial	248328	1159	
3	1962 Unf	2	462 TA	TA	Y	0	0	0	0	0	0	0	0					0	3	2007 WD	Normal	101800	1163	
4	1929 RFn	1	208 TA	TA	Y	0	0	112	0	0	0	0	0					0	7	2008 WD	Normal	120000	1351	
5	1925 Unf	1	160 Fa	TA	Y	0	141	0	0	0	0	0	0					0	4	2008 WD	Normal	91000	1325	
6	1960 RFn	1	312 TA	TA	Y	355	0	0	0	0	0	0	0	GdWo				0	4	2008 WD	Normal	141000	nan	
7	1967 Unf	3	792 TA	TA	Y	0	152	0	0	0	0	0	0					0	4	2009 WD	Normal	124000	1327	
8	1963 Unf	2	480 TA	TA	Y	0	80	0	0	0	0	0	0	MnPrv				0	6	2009 WD	Normal	139000	900	
9	1998 Unf	2	402 TA	TA	Y	0	125	0	0	0	0	0	0					0	5	2006 WD	Normal	164000	452	
10	2001 Fin	2	560 TA	TA	Y	125	192	0	0	0	0	0	0					0	6	2009 WD	Normal	215000	321	
11	1962 Unf	2	539 TA	TA	Y	0	23	112	0	0	0	0	0					0	1	2009 WD	Normal	103000	678	
12	1957 RFn	1	294 TA	TA	Y	250	0	0	0	0	0	0	0					0	6	2010 WD	Normal	145000	185	
13	1977 Unf	2	884 TA	TA	Y	0	64	0	0	0	0	0	0	MnPrv				0	10	2006 WD	Normal	146000	1023	
14	2005 Fin	2	451 TA	TA	Y	252	64	0	0	0	0	0	0					0	6	2008 WD	Normal	176000	961	
15	1974 Unf	4	480 TA	TA	Y	0	0	0	0	0	0	0	0	GdWo				0	6	2007 WD	Normal	123000	1271	
16	1976 Unf	2	665 TA	TA	Y	0	72	174	0	0	0	0	0					0	5	2008 COD	Abnorml	287000	nan	
17	1972 Unf	1	338 TA	TA	Y	0	0	0	0	0	0	0	0					0	8	2009 WD	Normal	133500	812	
18	1916 Unf	3	513 Fa	Fa	Y	0	0	96	0	0	0	0	0					0	5	2008 COD	Abnorml	98000	423	
19	1999 Unf	2	506 TA	TA	Y	0	34	0	0	0	0	0	0					0	3	2008 WD	Normal	183900	1231	
20	1995 Unf	2	576 TA	TA	Y	112	0	0	0	0	0	0	0	MnPrv				0	4	2009 WD	Normal	141500	644	
21	1976 Unf	2	572 TA	TA	Y	100	110	0	0	0	0	0	0					0	6	2007 WD	Normal	129900	-20	
22	1999 Fin	3	779 TA	TA	Y	668	30	0	0	0	0	0	0					0	5	2010 WD	Normal	333168	1206	
23	1983 Unf	2	484 TA	TA	Y	120	0	158	0	0	0	0	0					0	6	2007 WD	Normal	134000	443	
24	1964 RFn	2	442 TA	TA	Y	328	128	0	0	0	0	189	0					0	6	2008 WD	Normal	167900	988	
25	2003 Fin	2	420 TA	TA	Y	140	0	0	0	0	0	0	0					0	3	2008 WD	Normal	136500	1386	
26	0	0	0	Y	0	96	0	0	0	0	0	0	0					0	7	2009 WD	Normal	99900	1202	
27	2003 Fin	3	749 TA	TA	Y	168	0	0	0	0	0	0	0					0	8	2008 WD	Normal	305000	535	
28	1941 Unf	1	234 TA	TA	N	0	113	0	0	0	0	0	0					0	8	2009 WD	Normal	113000	1163	

*Kết quả sử dụng phép trừ*

AutoSave off task\_8\_test\_multiplication... Saved to this PC Search NGUYỄN TÂN LỘC

File Home Insert Page Layout Formulas Data Review View Automate Help

Font Alignment Number Styles Cells Editing Add-ins Analyze Data

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. Don't show again Save As...

A1 BH BI BJ BK BL BM BN BO BP BQ BR BS BT BU BV BW BX BY BZ CA CB CC CD result

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	result
1	GarageYrB GarageFini GarageCar GarageAre GarageCor PavedDriv WoodDeck OpenPorch EnclosedPorch ScreenPorch PorchArea PoolQC Fence MiscFeatu MiscVal MoSold YrsOld SaleType SaleCondit SalePrice																							
2	2007 RFn	3	954 TA	TA	Y	0	56	0	0	0	0	0	0				0	6	2007 New	Partial	248328	103086		
3	1962 Unf	2	462 TA	TA	Y	0	0	0	0	0	0	0	0				0	3	2007 WD	Normal	101800	86310		
4	1929 RFn	1	208 TA	TA	Y	0	0	112	0	0	0	0	0				0	7	2008 WD	Normal	120000	70050		
5	1925 Unf	1	160 Fa	TA	Y	0	141	0	0	0	0	0	0	GdWo			0	4	2008 WD	Normal	91000	71604		
6	1960 RFn	1	312 TA	TA	Y	355	0	0	0	0	0	0	0				0	4	2008 WD	Normal	141000	nan		
7	1967 Unf	3	792 TA	TA	Y	0	152	0	0	0	0	0	0				0	4	2009 WD	Normal	124000	90480		
8	1963 Unf	2	480 TA	TA	Y	0	80	0	0	0	0	0	0	MnPrv			0	6	2009 WD	Normal	139000	78400		
9	1998 Unf	2	402 TA	TA	Y	0	125	0	0	0	0	0	0				0	5	2006 WD	Normal	164000	15488		
10	2001 Fin	2	560 TA	TA	Y	125	192	0	0	0	0	0	0				0	6	2009 WD	Normal	215000	27832		
11	1962 Unf	2	539 TA	TA	Y	0	23	112	0	0	0	0	0				0	1	2009 WD	Normal	103000	37960		
12	1957 RFn	1	294 TA	TA	Y	250	0	0	0	0	0	0	0				0	6	2010 WD	Normal	145000	17850		
13	1977 Unf	2	884 TA	TA	Y	0	64	0	0	0	0	0	0	MnPrv			0	10	2006 WD	Normal	146000	77674		
14	2005 Fin	2	451 TA	TA	Y	252	64	0	0	0	0	0	0	GdWo			0	6	2008 WD	Normal	176000	61260		
15	1974 Unf	4	480 TA	TA	Y	0	0	0	0	0	0	0	0				0	5	2007 COD	Abnorml	287000	nan		
16	1976 Fin	2	665 TA	TA	Y	0	72	174	0	0	0	0	0				0	6	2007 WD	Normal	123000	93870		
17	1972 Unf	1	338 TA	TA	Y	0	0	0	0	0	0	0	0				0	8	2009 WD	Normal	133500	30528		
18	1916 Unf	3	513 Fa	Fa	Y	0	0	96	0	0	0	0	0				0	5	2008 COD	Abnorml	98000	15538		
19	1999 Unf	2	506 TA	TA	Y	0	34	0	0	0	0	0	0				0	3	2006 WD	Normal	183900	44310		
20	1995 Unf	2	576 TA	TA	Y	112	0	0	0	0	0	0	0				0	4	2009 WD	Normal	141500	35445		
21	1976 Unf	2	572 TA	TA	Y	100	110	0	0	0	0	0	0				0	6	2007 WD	Normal	129900	1056		
22	1999 Fin	3	779 TA	TA	Y	668	30	0	0	0	0	0	0				0	5	2010 WD	Normal	333168	141912		
23	1983 Unf	2	484 TA	TA	Y	120	0	158	0	0	0	0	0				0	6	2007 WD	Normal	134000	36494		
24	1964 RFn	2	442 TA	TA	Y	328	128	0	0	0	0	189	0				0	6	2008 WD	Normal	167900	85440		
25	2003 Fin	2	420 TA	TA	Y	140	0	0	0	0	0	0	0				0	7	2009 WD	Normal	136500	52651		
26	0	0	0	Y	0	96	0	0	0	0	0	0	0				0	8	2008 WD	Normal	99900	70448		
27	2003 Fin	3	749 TA	TA	Y	168	0	0	0	0	0	0	0				0	8	2008 WD	Normal	305000	52700		
28	1941 Unf	1	234 TA	TA	N	0	113	0	0	0	0	0	0				0	8	2009 WD	Normal	113000	60650		

*Kết quả sử dụng phép nhân*

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CL
1	Yr: GarageYrB GarageFin GarageCar GarageAre GarageQu GarageCoi PavedDriv WoodDeck OpenPorcl EnclosedP35snPorch ScreenPorl PoolArea PoolQC Fence MiscFeatu MiscVal MoSold YrSold SaleType SaleCondit SalePrice result																						
2	2007 RFn	3	954 TA	TA	Y	0	56	0	0	0	0	0	0				0	6	2007 New	Partial	248326	14.963855	
3	1962 Unf	2	462 TA	TA	Y	0	0	0	0	0	0	0	0				0	3	2007 WD	Normal	101806	17.614285	
4	1929 RFn	1	208 TA	TA	Y	0	0	112	0	0	0	0	0				0	7	2008 WD	Normal	120000	28.02	
5	1925 Unf	1	160 Fa	TA	Y	0	141	0	0	0	0	0	0				0	4	2008 WD	Normal	91000	26.480769	
6	1960 RFn	1	312 TA	TA	Y	355	0	0	0	0	0	0	0				0	4	2008 WD	Normal	141000	nan	
7	1967 Unf	3	792 TA	TA	Y	0	152	0	0	0	0	0	0				0	4	2009 WD	Normal	124000	21.415384	
8	1963 Unf	2	480 TA	TA	Y	0	80	0	0	0	0	0	0				0	6	2009 WD	Normal	139000	12.25	
9	1998 Unf	2	402 TA	TA	Y	0	125	0	0	0	0	0	0				0	5	2006 WD	Normal	164000	15.125	
10	2001 Fin	2	560 TA	TA	Y	125	192	0	0	0	0	0	0				0	6	2009 WD	Normal	215000	5.521127	
11	1962 Unf	2	539 TA	TA	Y	0	23	112	0	0	0	0	0				0	1	2009 WD	Normal	103000	14.038461	
12	1957 RFn	1	294 TA	TA	Y	250	0	0	0	0	0	0	0				0	6	2010 WD	Normal	145000	3.6428571	
13	1977 Unf	2	884 TA	TA	Y	0	64	0	0	0	0	0	0				0	10	2006 WD	Normal	146000	15.408450	
14	2005 Fin	2	451 TA	TA	Y	252	64	0	0	0	0	0	0				0	6	2008 WD	Normal	176000	17.016666	
15	1974 Unf	4	480 TA	TA	Y	0	0	0	0	0	0	0	0				0	6	2007 WD	Normal	123000	19.157142	
16	1976 Fin	2	665 TA	TA	Y	0	72	174	0	0	0	0	0				0	5	2008 COD	Abnorml	287000	nan	
17	1972 Unf	1	338 TA	TA	Y	0	0	0	0	0	0	0	0				0	8	2009 WD	Normal	133500	23.555555	
18	1916 Unf	3	513 Fa	Fa	Y	0	0	96	0	0	0	0	0				0	5	2008 COD	Abnorml	98000	13.441176	
19	1999 Unf	2	506 TA	TA	Y	0	34	0	0	0	0	0	0				0	3	2008 WD	Normal	183900	36.717428	
20	1995 Unf	2	576 TA	TA	Y	112	0	0	0	0	0	0	0				0	4	2009 WD	Normal	141500	13.627450	
21	1976 Unf	2	572 TA	TA	Y	100	110	0	0	0	0	0	0				0	6	2007 WD	Normal	129900	0.5454545	
22	1999 Fin	3	779 TA	TA	Y	668	30	0	0	0	0	0	0				0	5	2010 WD	Normal	333168	12.166666	
23	1983 Unf	2	484 TA	TA	Y	120	0	158	0	0	0	0	0				0	6	2007 WD	Normal	134000	7.239437	
24	1964 RFn	2	442 TA	TA	Y	328	128	0	0	189	0	0	0				0	6	2008 WD	Normal	167900	13.35	
25	2003 Fin	2	420 TA	TA	Y	140	0	0	0	0	0	0	0				0	3	2008 WD	Normal	136500	38.459459	
26	0	0	0	Y	0	96	0	0	0	0	0	0	0				0	7	2009 WD	Normal	99900	22.646285	
27	2003 Fin	3	749 TA	TA	Y	168	0	0	0	0	0	0	0				0	8	2008 WD	Normal	305000	7.2941176	
28	1941 Unf	1	234 TA	TA	N	0	113	0	0	0	0	0	0				0	8	2009 WD	Normal	113000	24.26	

Kết quả sử dụng phép chia

## TÀI LIỆU THAM KHẢO

- Tài liệu của giáo viên
- How to Calculate the 5-Number Summary for Your Data in Python by Jason Brownlee on August 8, 2019: <https://machinelearningmastery.com/how-to-calculate-the-5-number-summary-for-your-data-in-python/>
- How to Better Understand Your Machine Learning Data in Weka by Jason Brownlee on August 22, 2019: <https://machinelearningmastery.com/better-understand-machine-learning-data-weka/>
- Pattern recognition of data by scatter plot by California State University, Sacramento: <https://www.csus.edu/indiv/m/mirzaagham/math1/S4.pdf>
- Trực quan hóa dữ liệu – Phần 6: Các dạng biểu đồ thể hiện sự tương quan của dữ liệu: <http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/1006-truc-quan-hoa-du-lieu-phan-6cac-dang-bieu-do-the-hien-su-tuong-quan-cua-du-lieu>
- Weka Documentation: <https://waikato.github.io/weka-wiki/documentation/>
- Argparse Documentation: <https://docs.python.org/3/library/argparse.html>
- Sys Documentation: <https://docs.python.org/3/library/sys.html>
- Data Preprocessing for Data Mining:  
[https://www.youtube.com/watch?v=ifGJk2S3Y4U&list=PL4gu8xQu0\\_5Le\\_OyCHx-fhTOli-WDHjuy](https://www.youtube.com/watch?v=ifGJk2S3Y4U&list=PL4gu8xQu0_5Le_OyCHx-fhTOli-WDHjuy)

---HẾT---