

Báo cáo Lab 1 – Preprocessing - Data mining

Tên: Phan Trung Hiếu

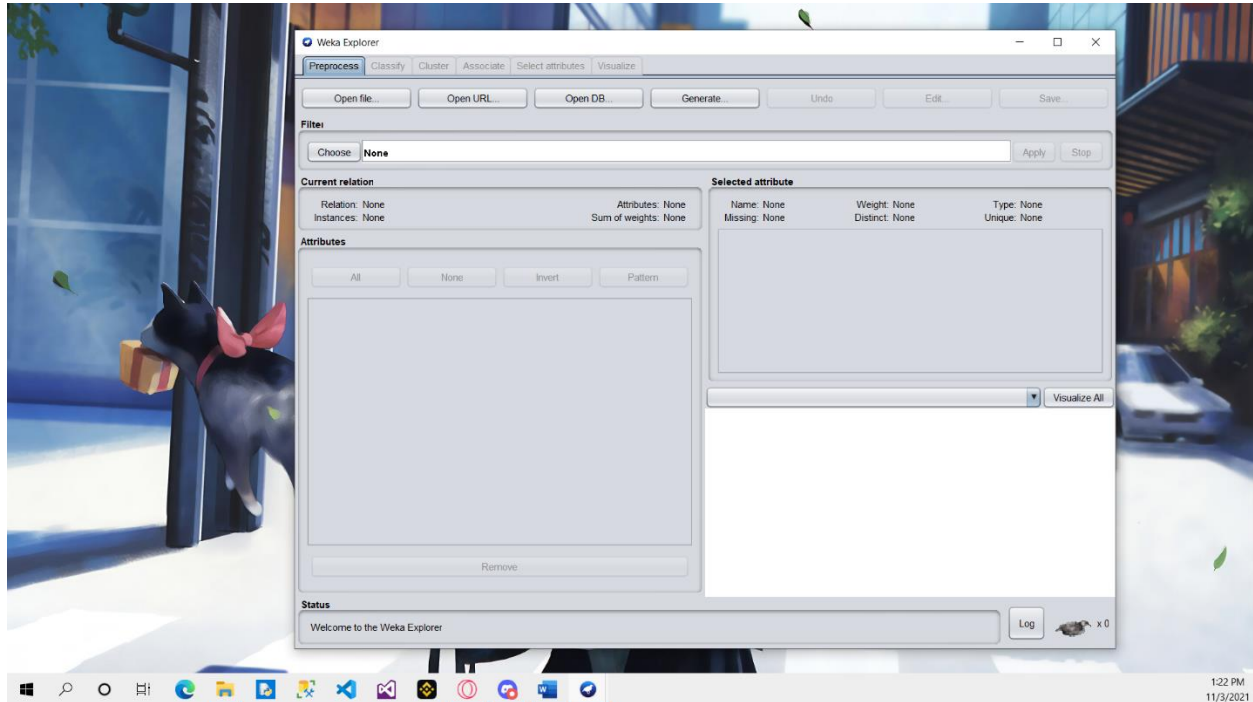
MSSV: 19127404

Bảng tiến độ

| STT | Công việc | | Tỉ lệ hoàn thành |
|-----|--------------|--|------------------|
| 1 | Yêu cầu 1 | Cài đặt weka | 100% |
| 2 | Yêu cầu 2 | Đọc dữ liệu Weka | 100% |
| | | Khám phá tập dữ liệu Weather | 100% |
| | | Khám phá tập dữ liệu tín dụng Đức | 100% |
| 3 | Yêu cầu 3 | Đếm số dòng thiếu dữ liệu | 100% |
| 4 | | Đếm số cột thiếu dữ liệu | 100% |
| 5 | | Điền giá trị vào chỗ thiếu dữ liệu | 100% |
| 6 | | Xoá các dòng bị thiếu với ngưỡng cho trước | 100% |
| 7 | | Xoá các cột bị thiếu với ngưỡng cho trước | 100% |
| 8 | | Chuẩn hoá thuộc tính | 100% |
| 9 | | Tính giá trị biểu thức thuộc tính | 100% |
| 10 | Viết báo cáo | | 100% |

Yêu cầu 1:

Ảnh chụp:



Tab Preprocess:

+ *Current relation*: thông tin về mối quan hệ của database đang xét, gồm có tên của relation, số lượng mẫu và số lượng các thuộc tính

+ *Attributes*: danh sách các thuộc tính của relation

+ *Selected attribute*: hiện ra thông tin cách mà dữ liệu được phân phối theo thuộc tính này, ngoài ra bên dưới còn có cả đồ thị dạng cột để biểu diễn thông tin một cách trực quan hơn

Tab Classify: tab này dùng để build model phân lớp, sử dụng thuật toán theo ý của user như linear regression, support vector machine, bayes ...

Tab Cluster: tab này dùng để gom nhóm các phần tử tương tự gần nhau, sử dụng các thuật toán như EM, FilteredClusterer, HierarchicalClusterer, ...

Tab Associate: tab này dùng để tìm ra các luật kết hợp từ các thuộc tính của dữ liệu, sử dụng các thuật toán quen thuộc tìm tập phổ biến như Apriori, FP-Tree...

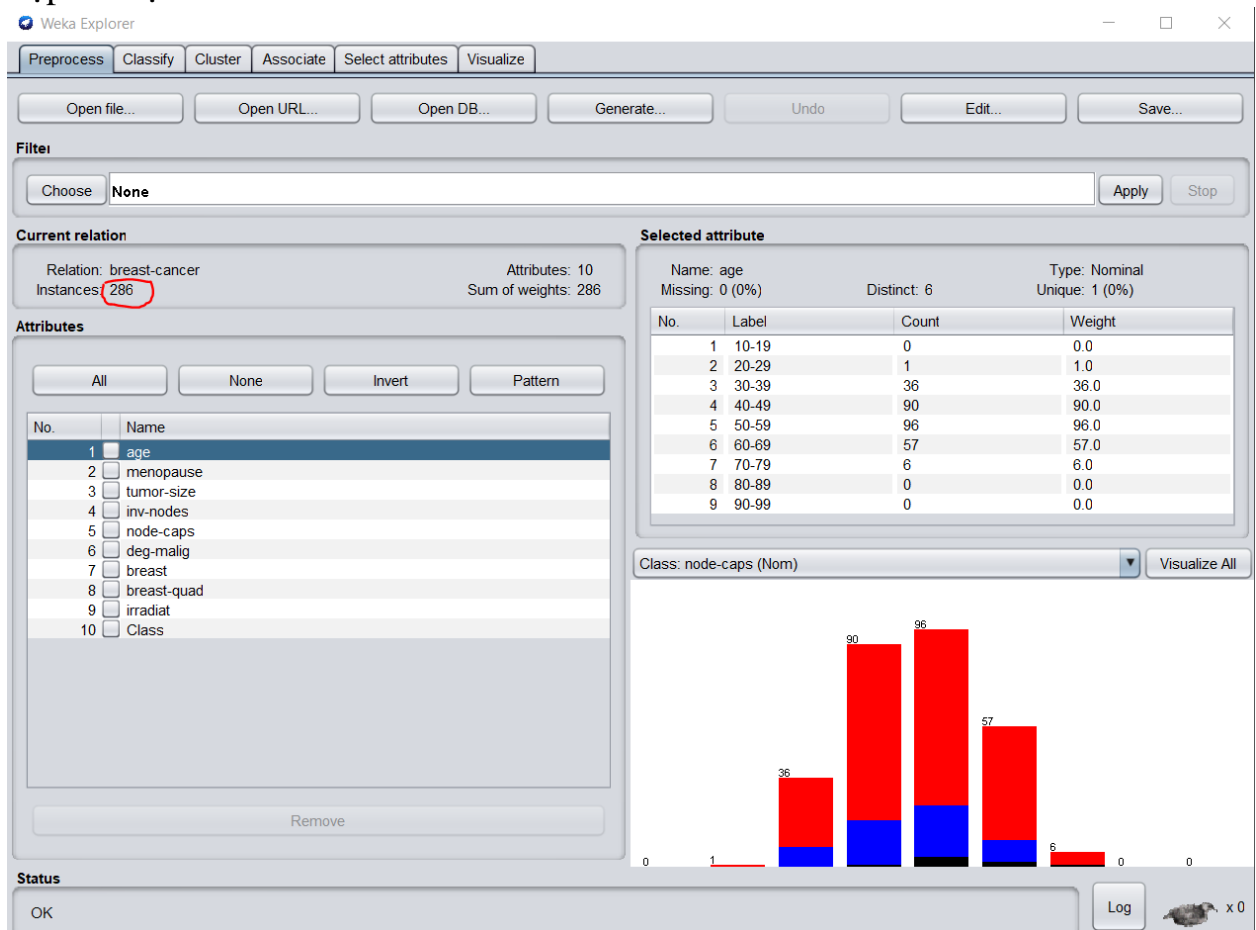
Tab Select attribute: tab này dùng để loại bỏ các attribute không ảnh hưởng hoặc có ít ảnh hưởng đến kết quả, làm cho model chạy tốt hơn và ít bị nhiễu hơn

Tab Visualize: tab này cung cấp biểu đồ dữ liệu để dễ dàng phân tích thêm

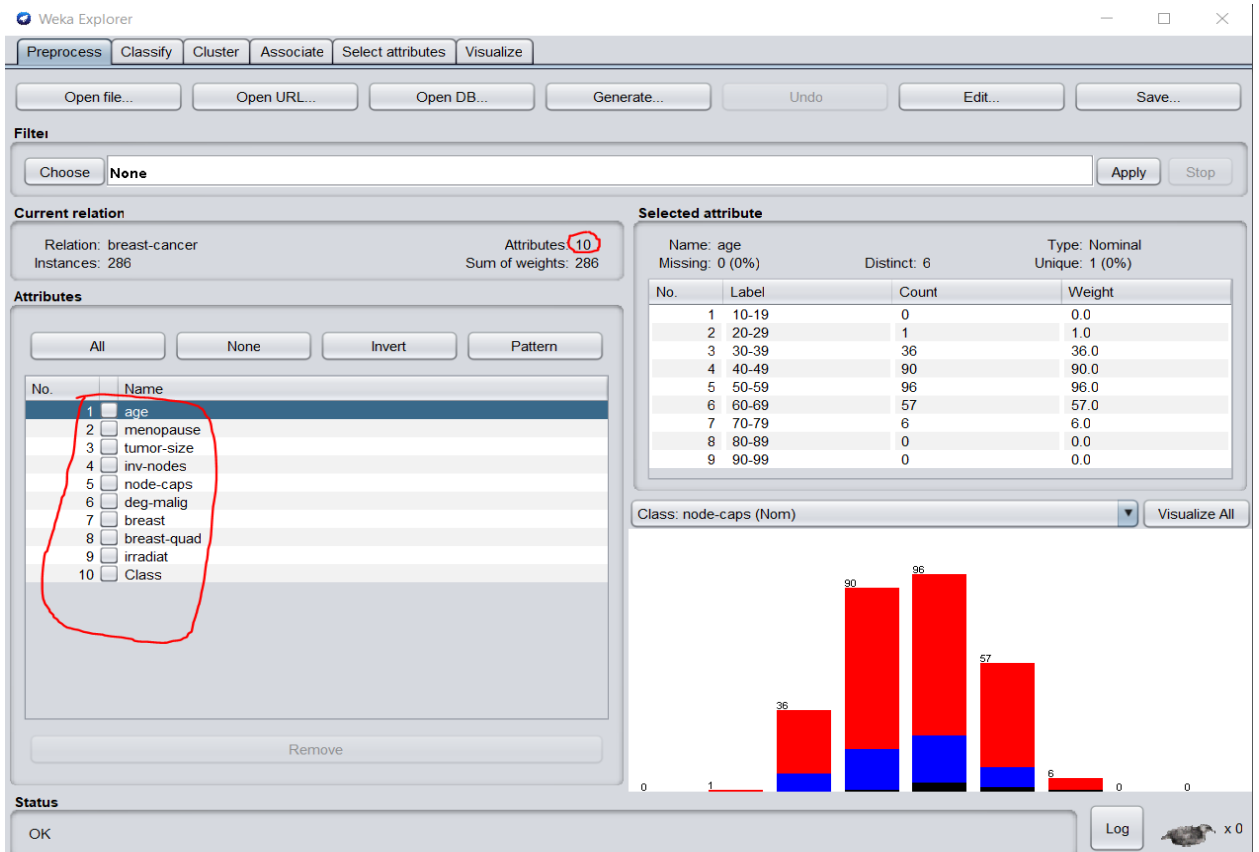
Yêu cầu 2:

2.1 Đọc dữ liệu vào Weka

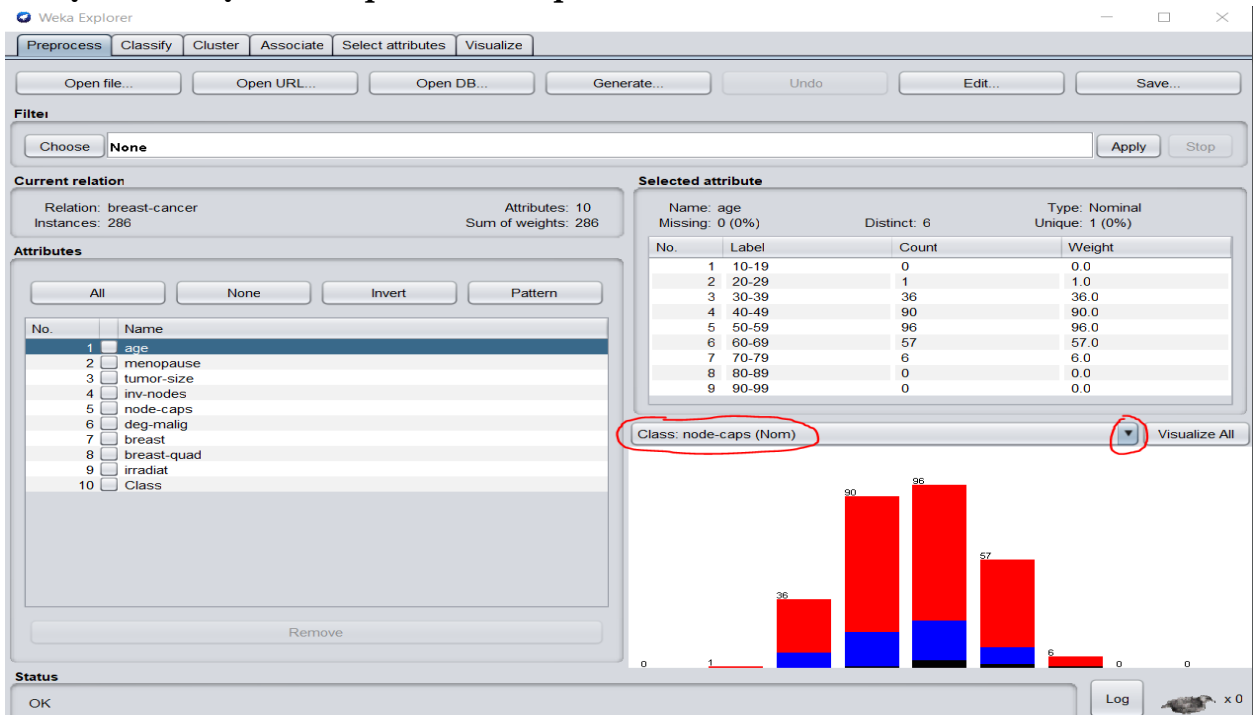
1. Tập dữ liệu có 286 mẫu



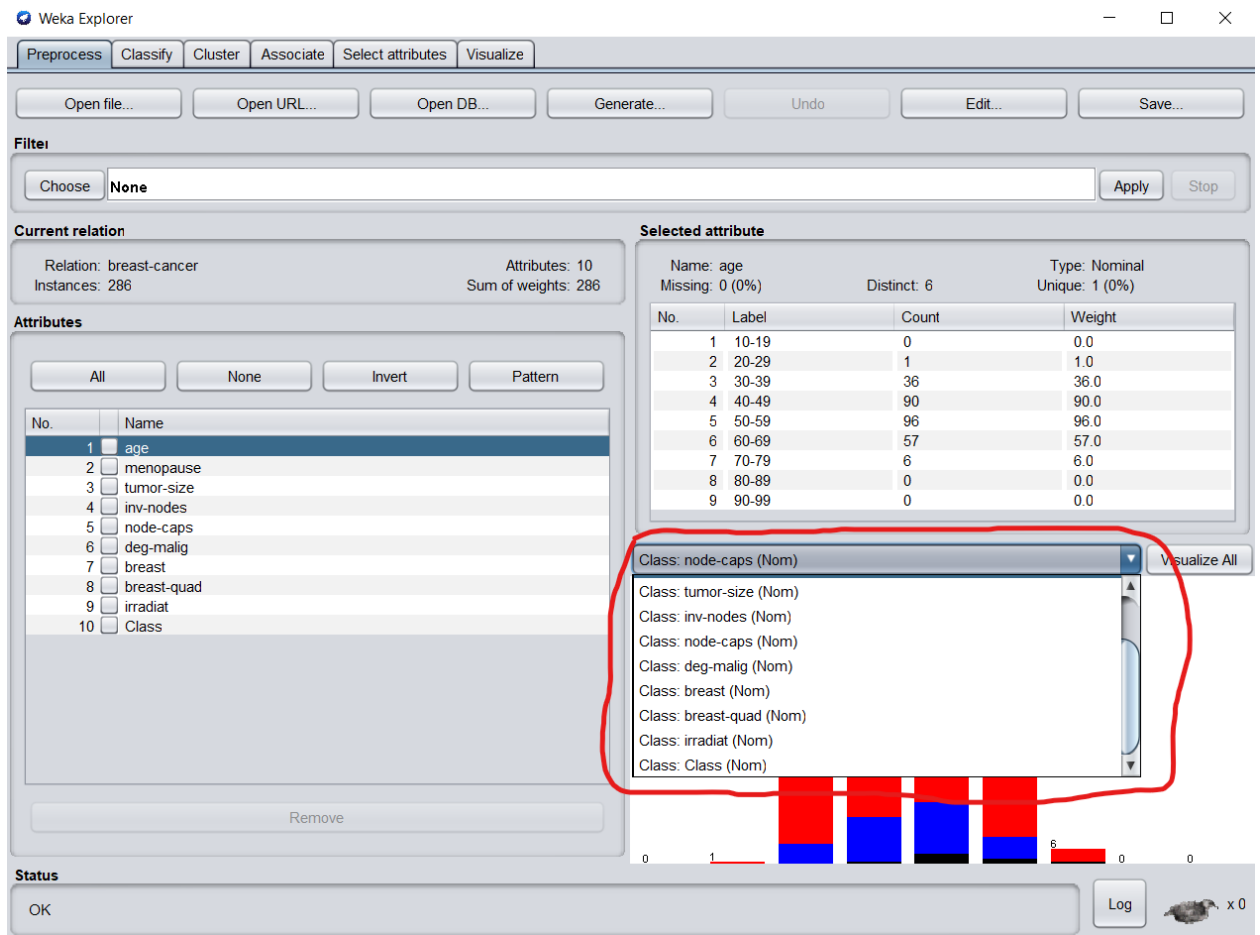
2. Tập dữ liệu có 10 thuộc tính



3. Thuộc tính chọn làm lớp là node-caps



Chúng ta hoàn toàn có thể thay đổi thuộc tính làm lớp bằng cách nhấn vào dấu mũi tên bên cạnh và chọn thuộc tính mình muốn:



4. Ta có bảng sau

| Thuộc tính | Số dòng bị mất dữ liệu |
|-------------|------------------------|
| Age | 0 |
| Menopause | 0 |
| Tumor-size | 0 |
| Inv-nodes | 0 |
| Node-caps | 8 |
| Deg-mailg | 0 |
| Breast | 0 |
| Breast-quad | 1 |
| Irradiat | 0 |
| Class | 0 |

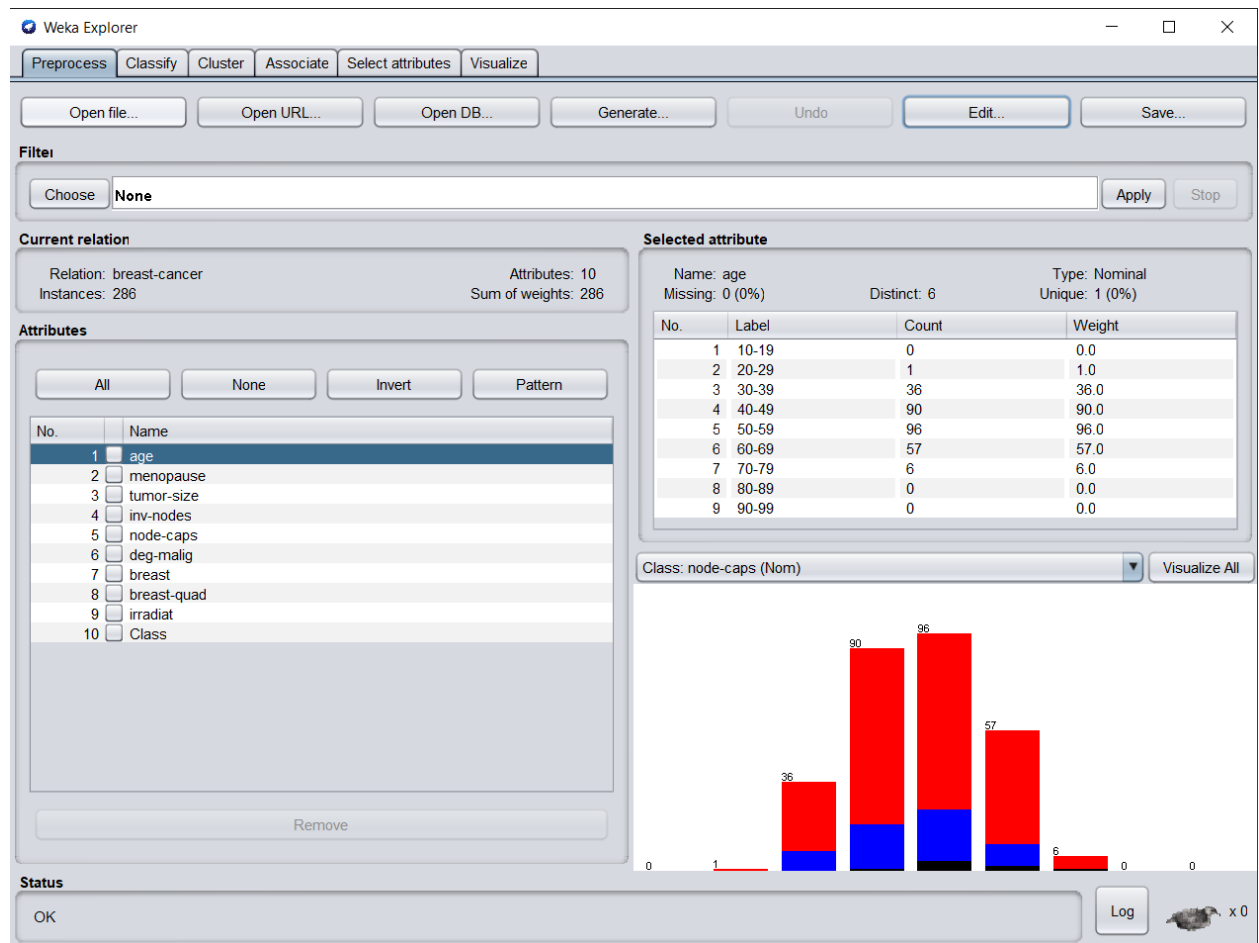
(Số liệu được lấy trong mục EDIT, sort theo từng lớp của thuộc tính sau đó đếm các dòng trống)

Ta dễ thấy thuộc tính thiếu nhiều dữ liệu nhất là Node-caps (8 dòng), ít nhất là các thuộc tính Age, menopause, ... (0 dòng)

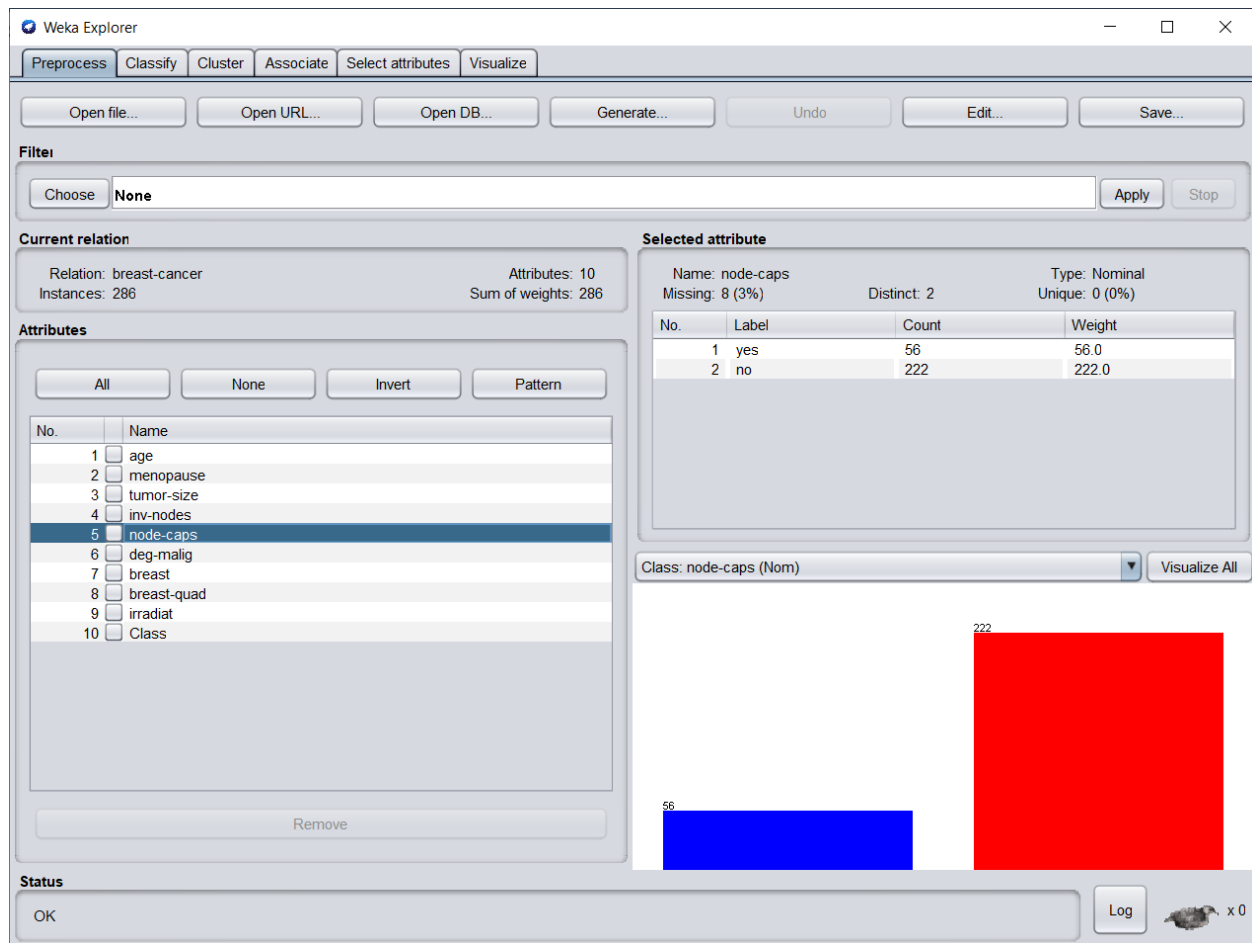
Các cách giải quyết khi gặp vấn đề missing value:

- Xoá dòng đó
- Thay thế thuộc tính bị thiếu bởi các giá trị mean/mode/median của cột đó
- Dùng một thuật toán, mô hình nào đó để dự đoán giá trị bị mất, ví dụ như LinearRegression
-

5. Ta có 2 hình sau :



Hình 1



Hình 2

Ở hình 2, ta đang chọn thuộc tính hiển thị đồ thị là node-caps và thuộc tính phân lớp cũng là node-caps.

Label “no” màu đỏ còn label “yes” thì có màu xanh, đồ thị có 2 cột tương ứng với 2 label

Trở lại hình 1, thì đồ thị có 9 cột tương ứng với 9 label, tuy nhiên thì lần này lại có đến 3 màu là xanh, đỏ và đen. Tính chất của các cột vẫn không đổi, tức là nếu không xét màu sắc, thì đồ thị này chính là phân phối của các mẫu theo thuộc tính age ứng với cái label tương ứng. Còn màu sắc thì chính là phân loại class theo thuộc tính node-caps. Màu đỏ tương ứng với label “no” và màu xanh thì có label “yes”, màu đen chính là các dòng bị missing value.

Vậy ta đặt tên cho đồ thị ở hình 1 là “Đồ thị phân loại thuộc tính node-caps theo phân phối thuộc tính age”

Tổng quát thì ta đặt tên cho loại đồ thị này là “Đồ thị phân bố lớp”

2.2 Khám phá tập dữ liệu Weather

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

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** [Apply] [Stop]

Current relation
Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes
[All] [None] [Invert] [Pattern]

| No. | Name |
|-----|---|
| 1 | <input checked="" type="checkbox"/> outlook |
| 2 | <input type="checkbox"/> temperature |
| 3 | <input type="checkbox"/> humidity |
| 4 | <input type="checkbox"/> windy |
| 5 | <input type="checkbox"/> play |

[Remove]

Selected attribute
Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|----------|-------|--------|
| 1 | sunny | 5 | 5.0 |
| 2 | overcast | 4 | 4.0 |
| 3 | rainy | 5 | 5.0 |

Class: play (Nom) [Visualize All]

Status
OK [Log] x 0

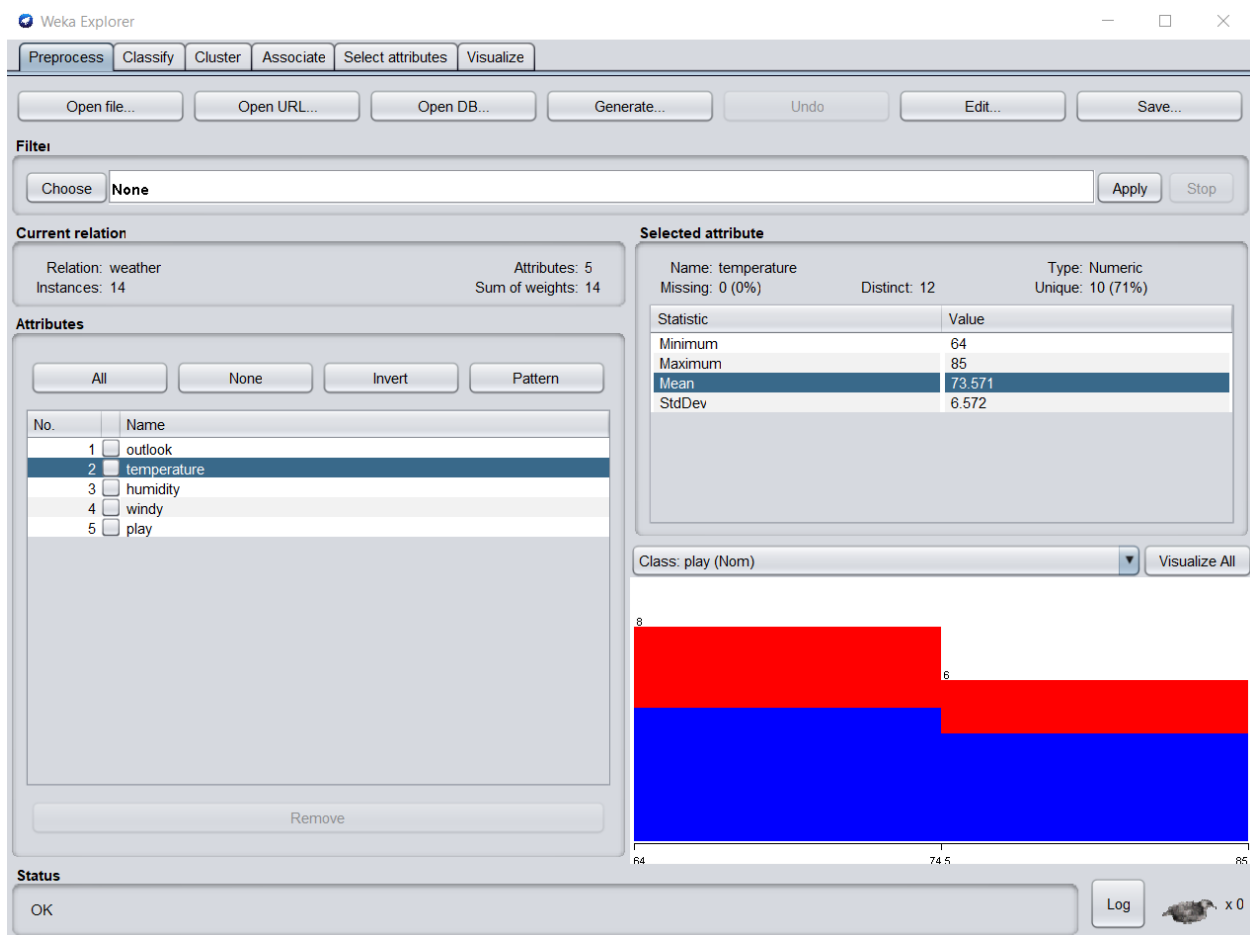
Tập dữ liệu có 5 thuộc tính, 14 mẫu

Phân loại kiểu dữ liệu:

- Nominal: outlook, windy, play
- Numeric: temperature, humidity

Thuộc tính làm lớp là “play”

2.

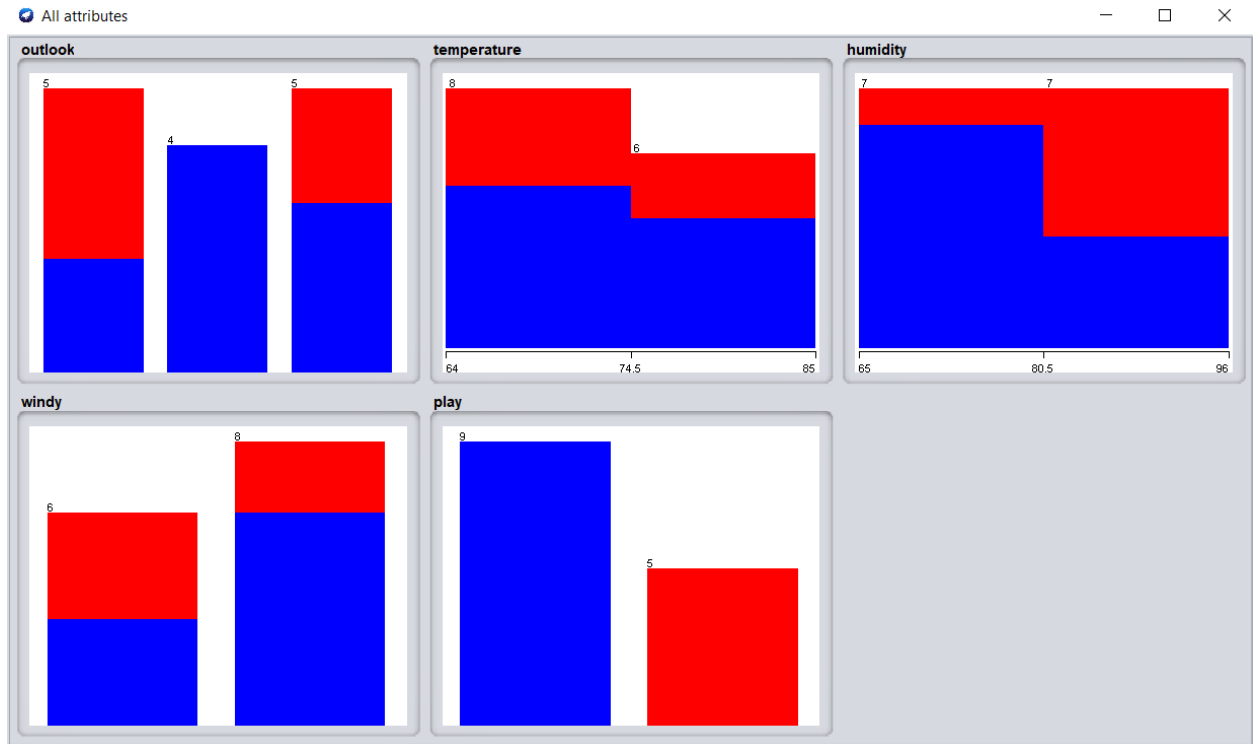


Five number summary của thuộc tính temperature và humidity là :

| | Min | Q1 | Mean | Q3 | Max |
|-------------|-----|----|--------|----|-----|
| Temperature | 64 | 69 | 73.571 | 80 | 85 |
| Humidity | 65 | 70 | 81.643 | 90 | 96 |

Như hình trên, thì weka chỉ cung cấp cho chúng ta 2 số liệu trong Five number summary, đó là minimum và maximum. Ngoài ra thì không có 3 số liệu còn lại, thay vào đó weka cung cấp cho chúng ta giá trị trung bình và độ lệch chuẩn(mean, standard deviation)

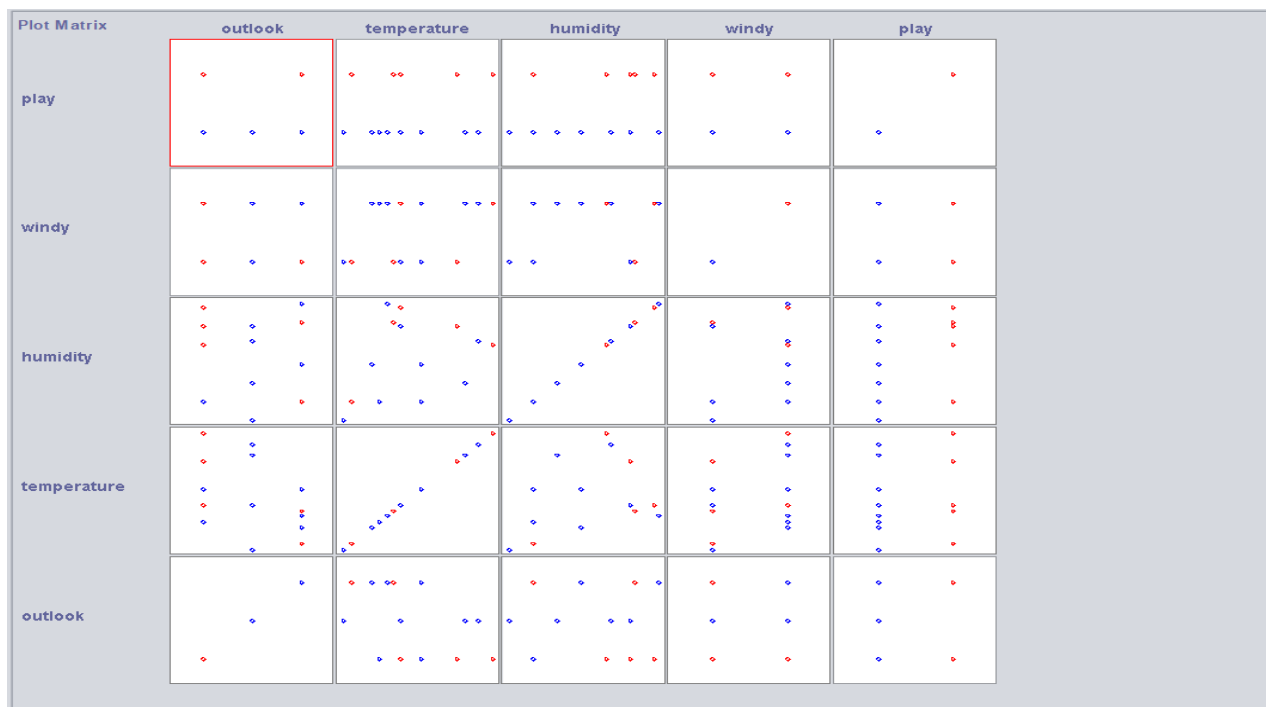
3.



Đồ thị biểu diễn các thuộc tính của tập dữ liệu

4.

Thuật ngữ sử dụng cho các đồ thị ở tab Visualize là đồ thị phân tán (scatter plot)



2.3 Khám phá tập dữ liệu Tín dụng Đức

1.

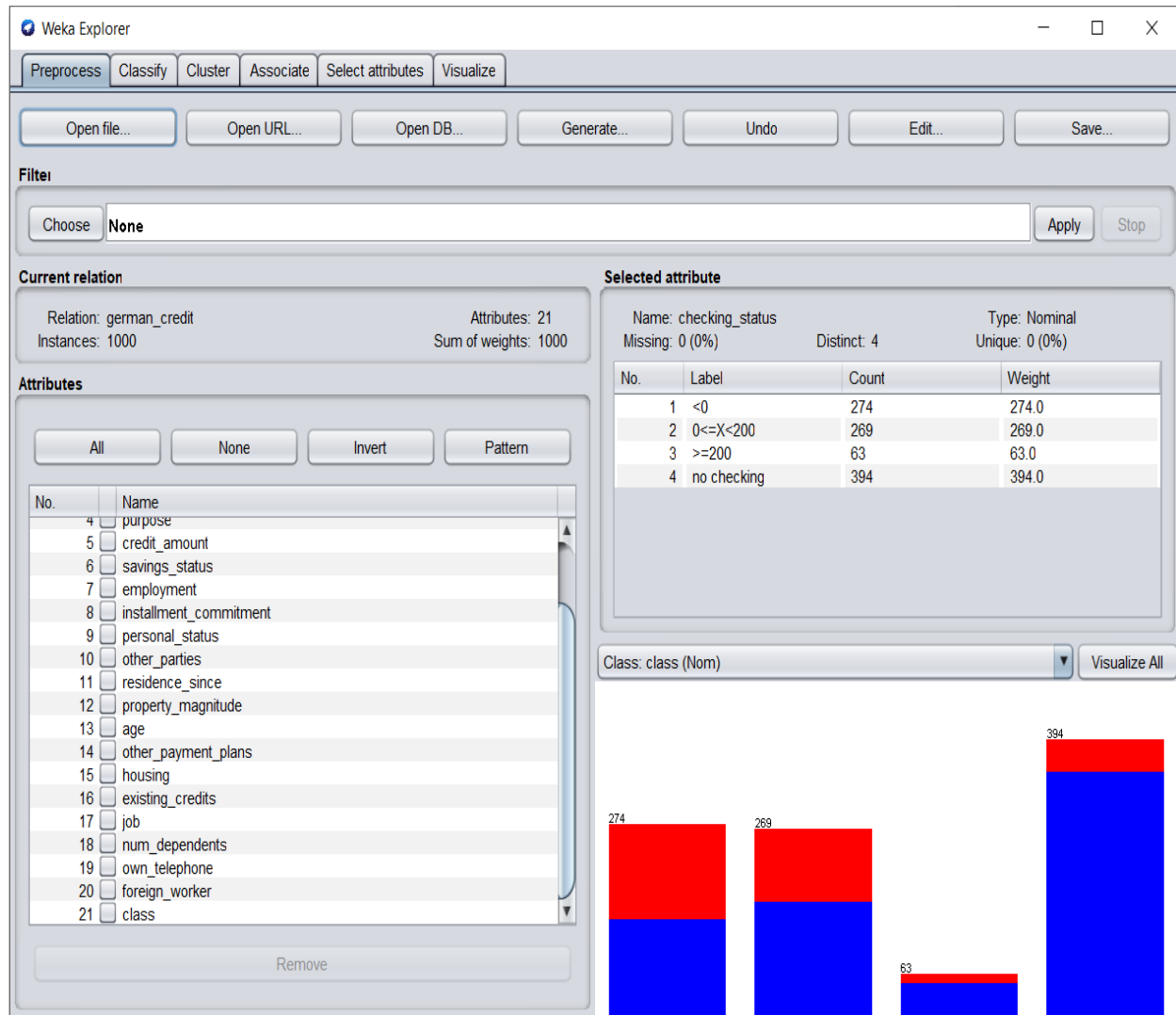
```

credit-g.arff - Notepad
File Edit Format View Help
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut für Statistik und "Ökonometrie
% Universität 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 attributes and
% is in the file "german.data".
%
% For algorithms that need numerical attributes, Strathclyde 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.numeric: 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)

```

Mở file credit-g.arff bằng notepad

Nội dung chú thích bao gồm tiêu đề, thông tin nguồn gốc cũng như chủ sở hữu nghiên cứu tập tin, số lượng mẫu, số lượng thuộc tính và còn cung cấp thêm kiểu dữ liệu của thuộc tính, mô tả chi tiết về thuộc tính đó. Ở phía dưới thì có thêm phần chi phí ma trận.

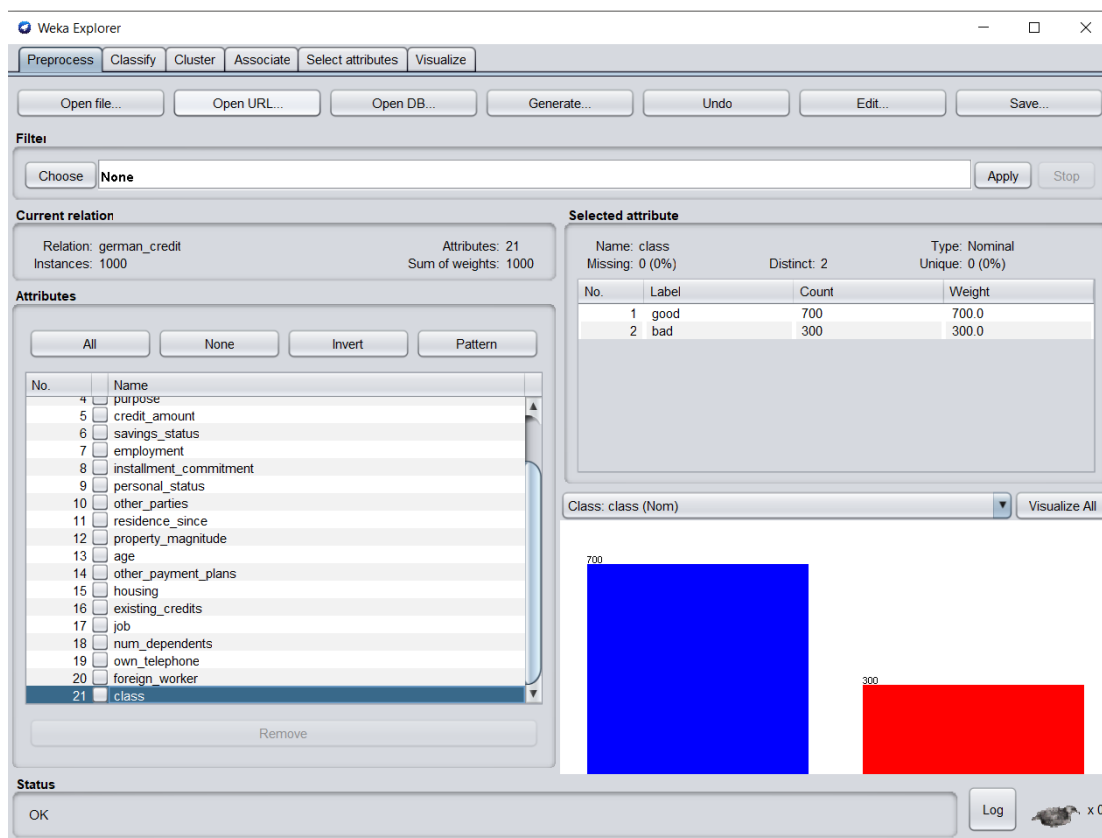


Tập thuộc tính bao gồm 1000 mẫu và 21 thuộc tính

Thông tin về 5 thuộc tính điển hình trong tập dữ liệu:

- Duration (rời rạc): thời hạn vay tín dụng (month)
- Saving_status(liên tục): lượng tiền có trong tài khoản tiết kiệm, được chia thành các mốc như sau : (<100), $[100, 500]$, $[500, 1000]$, (>1000), no known savings
- Housing (rời rạc) : trạng thái về nhà ở hiện tại : thuê, sở hữu, ở miễn phí
- Purpose (rời rạc) : mục đích của việc vay tín dụng : mua xe mới, mua xe cũ, thiết bị, đồ nội thất , nghỉ dưỡng, giáo dục
- Employment (liên tục): liệt kê số lượng nhân viên mà người vay tín dụng có, được chia thành các mốc như sau : unemployed, (<1), $[1, 4]$, $[4, 7]$, (>7)

2.



Tên thuộc tính lớp: class(good hoặc bad). Như hình trên thì dữ liệu bị lệch về phía good (màu xanh)

3.

Trong bộ lọc thuộc tính, có 2 phần là Đánh giá thuộc tính (**Attribute evaluator**) và Phương pháp tìm kiếm (**Search method**) dùng để hỗ trợ cho thuật toán Đánh giá thuộc tính.

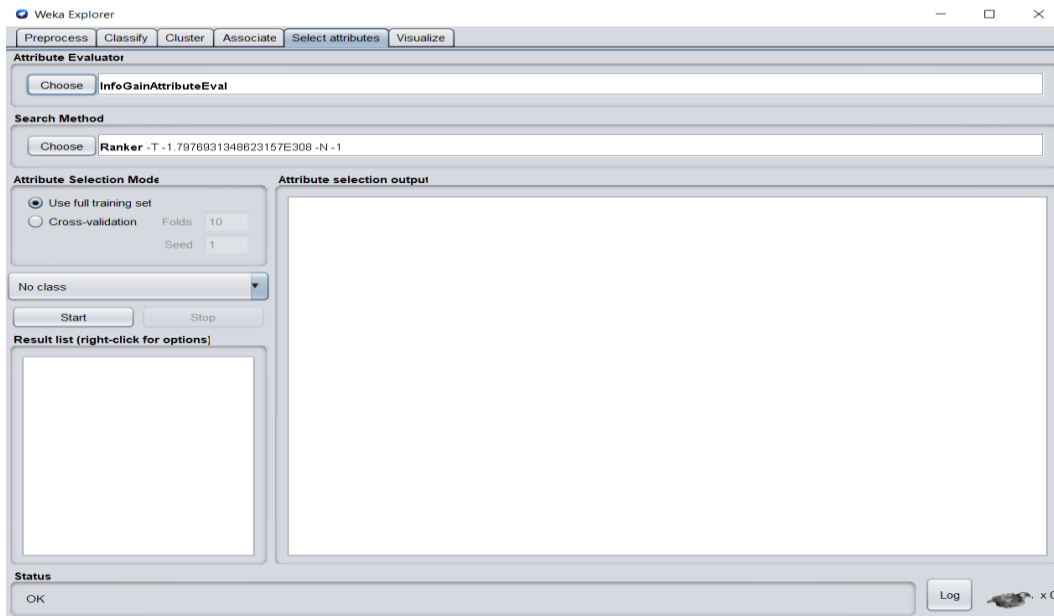
Đánh giá thuộc tính bao gồm:

- **CfsSubsetEval** : đánh giá giá trị của một tập con(gồm một vài thuộc tính) bằng cách xem xét khả năng dự đoán riêng của từng đối tượng với mức độ dư thừa giữa chúng
- **ClassifierAttributeEval**: đánh giá giá trị của thuộc tính bằng cách sử dụng bộ phân loại do người dùng chỉ định
- **ClassifierSubsetEval**: đánh giá các tập hợp con thuộc tính trên dữ liệu training. Sử dụng bộ phân loại để ước lượng giá trị của một tập hợp con các thuộc tính
- **CorrelationAttributeEval**: đánh giá thuộc tính bằng cách xét sự tương quan của nó so với lớp. Đối với các thuộc tính có kiểu dữ liệu nominal, các giá trị sẽ được xem xét như một chỉ số và mối tương quan của nó được thể hiện qua giá trị trung bình.
- **GainRatioAttributeEval**: đánh giá giá trị của thuộc tính bằng cách tính toán Gain ratio so với lớp
- **InfoGainAttributeEval**: đánh giá giá trị của thuộc tính bằng cách tính toán Information gain so với lớp
- **OneRAttributeEval**: đánh giá giá trị thuộc tính thông qua bộ phân loại OneR
- **PrincipalComponents**: giảm chiều dữ liệu và tìm ra tập con có các thuộc tính có độ tương quan cao nhất so với lớp
- **ReliefAttributeEval**: đánh giá giá trị của thuộc tính bằng cách liên tục lấy mẫu một cá thể và xem xét giá trị của thuộc tính đã cho với phiên bản gần nhất của cùng một lớp hoặc khác lớp
- **SymmetricalUncertAttributeEval**: đánh giá giá trị của thuộc tính bằng cách ước lượng sự đối xứng không cụ thể với lớp
- **WrapperSubsetEval**: đánh giá thuộc tính bằng cách dùng một thuật toán học máy nào đó và ước lượng độ chính xác của thuật toán bằng phương pháp Cross-validation

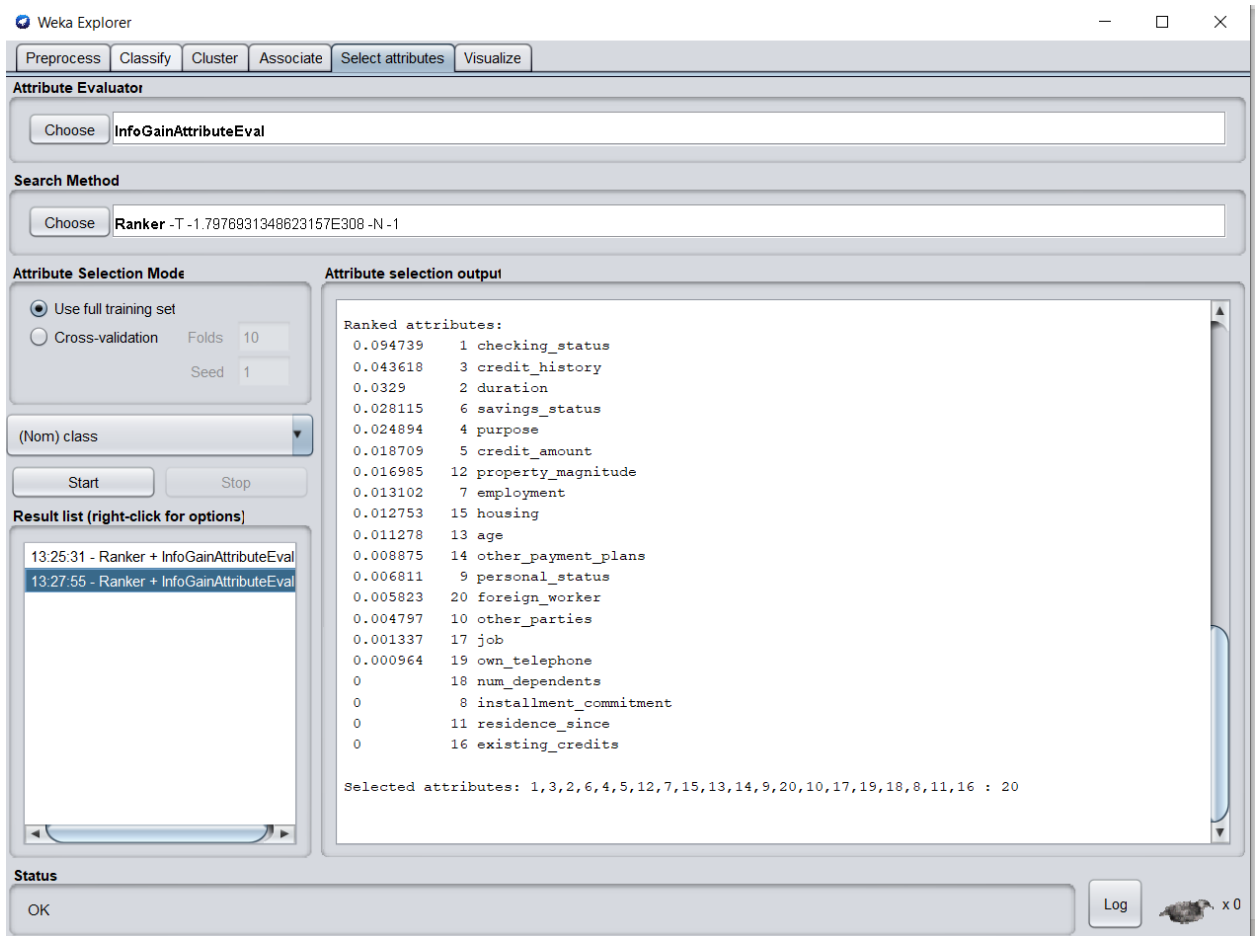
4.

Dùng InfoGainAttributeEval để chọn 5 thuộc tính có độ tương quan cao nhất đối với thuộc tính lớp

- Bước 1: vào tab Select attribute. Tại mục Attribute Evaluator, chọn InfoGainAttributeEval. Phần mềm sẽ tự chọn Search method (Cụ thể là Ranker)



- Bước 2: chọn (Nom) class và nhấn Start

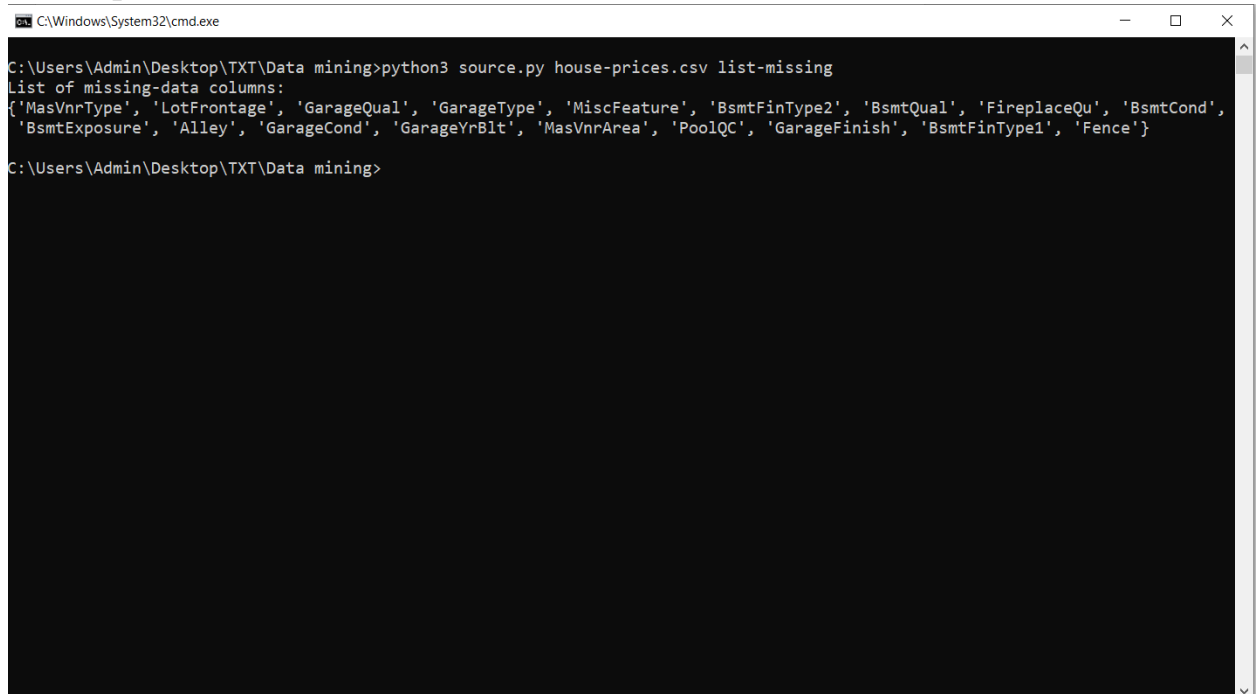


Từ kết quả, dễ thấy 5 thuộc tính có độ tương quan cao nhất so với lớp là checking_status, credit_history, duration, savings_status, purpose và credit_amount

Yêu cầu 3:

1. Liệt kê các cột bị thiếu dữ liệu:

- Cú pháp: `python3 <source_name> <data_file_name> list-missing`
- Kết quả:



```
C:\Windows\System32\cmd.exe

C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv list-missing
List of missing-data columns:
{'MasVnrType', 'LotFrontage', 'GarageQual', 'GarageType', 'MiscFeature', 'BsmtFinType2', 'BsmtQual', 'FireplaceQu', 'BsmtCond',
'BsmtExposure', 'Alley', 'GarageCond', 'GarageYrBlt', 'MasVnrArea', 'PoolQC', 'GarageFinish', 'BsmtFinType1', 'Fence'}
```

2. Đếm số dòng bị thiếu dữ liệu:

- Cú pháp: `python3 <source_name> <data_file_name> count-missing`
- Kết quả:

C:\Windows\System32\cmd.exe

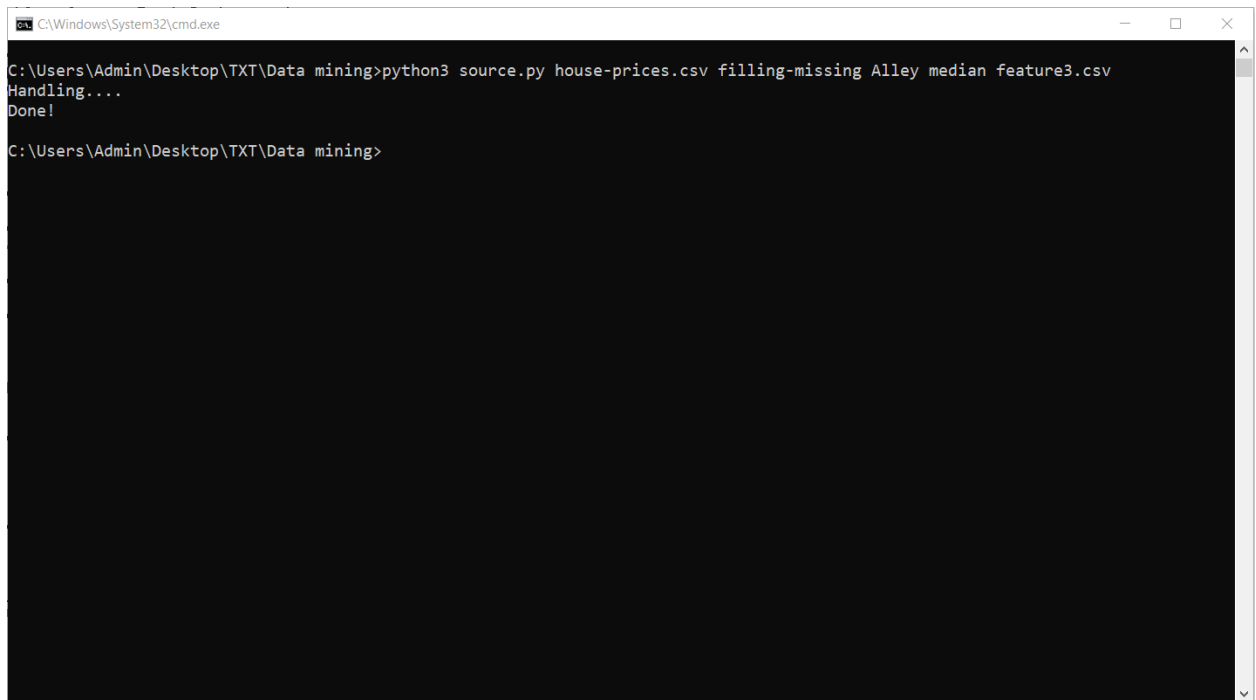
```
C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv count-missing
```

```
The number of missing-data rows:      1000
```

```
C:\Users\Admin\Desktop\TXT\Data mining>
```

3. Điền giá trị bị thiếu

- Cú pháp: `python3 <source_name> <data_file_name> filling-missing <attribute> <method> <file_out>`
- Kết quả: (file kết quả feature3.csv được đính kèm trong thư mục bài làm)



```
C:\Windows\System32\cmd.exe
C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv filling-missing Alley median feature3.csv
Handling....
Done!
C:\Users\Admin\Desktop\TXT\Data mining>
```

4. Xóa các dòng thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước:

- Cú pháp: `python3 <source_name> <data_file_name> deleting-rows <rate> <file_out>`
- Kết quả: (file kết quả feature4.csv được đính kèm trong thư mục bài làm)

```
C:\Windows\System32\cmd.exe

C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv deleting-rows 0.1 feature4.csv
Handling...
Done!

C:\Users\Admin\Desktop\TXT\Data mining>
```

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

- Cú pháp: `python3 <source_name> <data_file_name> deleting-cols <rate> <file_out>`
- Kết quả: (file `feature5.csv` được đính kèm trong thư mục bài làm)

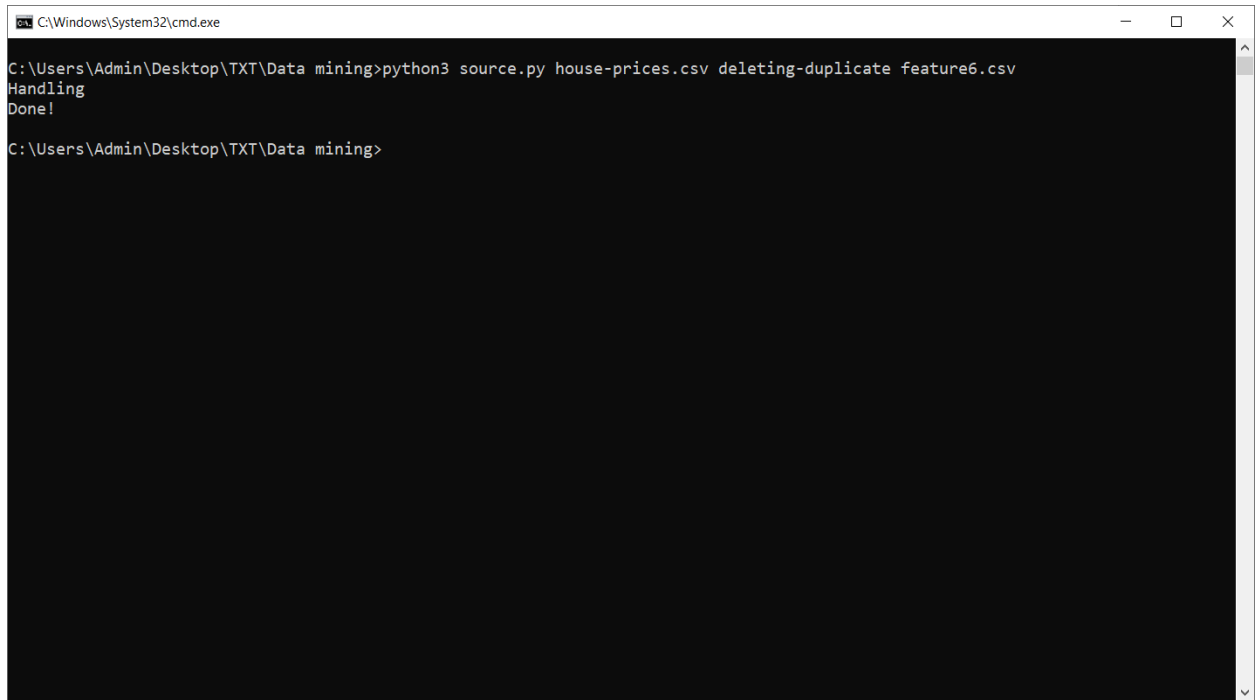
```
C:\Windows\System32\cmd.exe

C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv deleting-cols 0.1 feature5.csv
Handling...
Done!

C:\Users\Admin\Desktop\TXT\Data mining>
```

6. Xóa các mẫu bị trùng lặp

- Cú pháp: `python3 <source_name> <data_file_name> deleting-duplicate <file_out>`
- Kết quả: (file `feature6.csv` được đính kèm trong thư mục bài làm)



```
C:\Windows\System32\cmd.exe

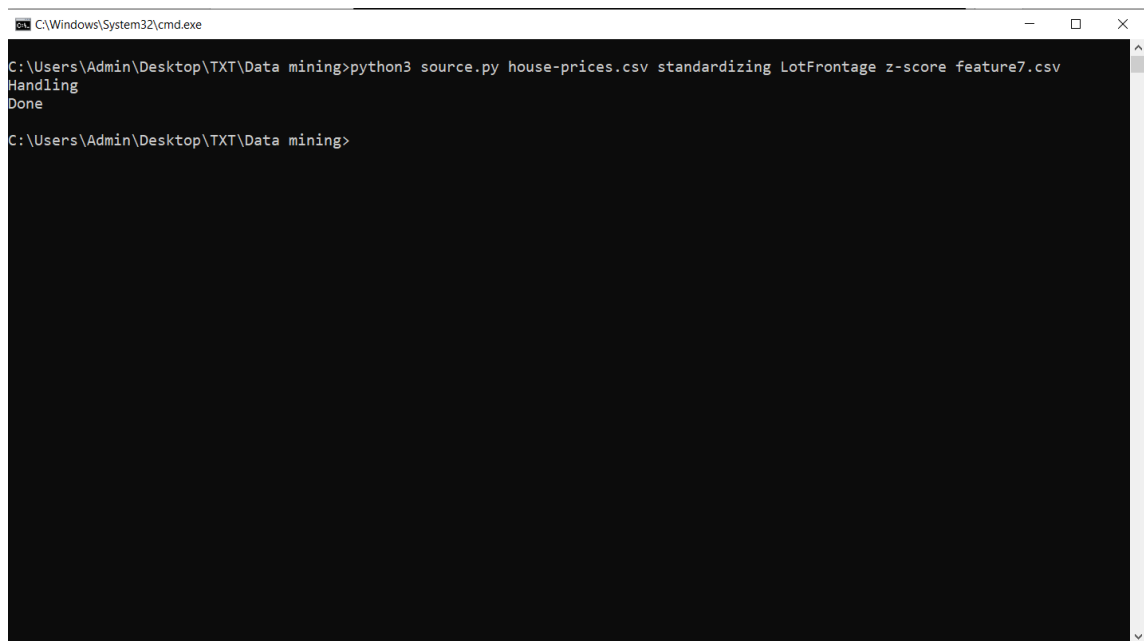
C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv deleting-duplicate feature6.csv
Handling
Done!

C:\Users\Admin\Desktop\TXT\Data mining>
```

7. Chuẩn hoá dữ liệu theo phương pháp Z-score và min-max

- Cú pháp: `python3 <source_name> <data_file_name> standardizing <attribute> <method> <file_out>`

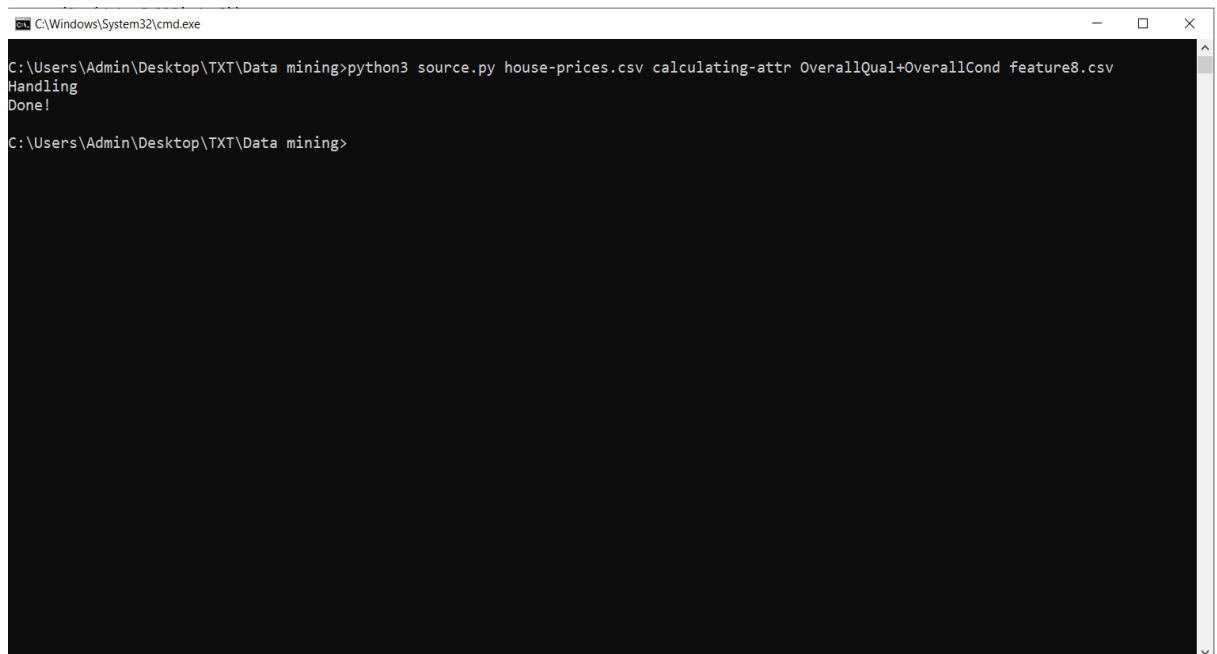
- Kết quả: (file feature7.csv được đính kèm trong thư mục bài làm)



```
C:\Windows\System32\cmd.exe
C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv standardizing LotFrontage z-score feature7.csv
Handling
Done
C:\Users\Admin\Desktop\TXT\Data mining>
```

8. Tính giá trị thuộc tính biểu thức

- Cú pháp: python3 <source_name> <data_file_name> calculating-attr <request> <file_out>
- Kết quả: (file feature8.csv được đính kèm trong thư mục bài làm)



```
C:\Windows\System32\cmd.exe
C:\Users\Admin\Desktop\TXT\Data mining>python3 source.py house-prices.csv calculating-attr OverallQual+OverallCond feature8.csv
Handling
Done!
C:\Users\Admin\Desktop\TXT\Data mining>
```