

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



BÁO CÁO TRẢ LỜI CÁC CÂU HỎI TỰ LUẬN
ĐỒ ÁN 1: Tiền xử lý dữ liệu

Môn học: Khai thác dữ liệu và ứng dụng 19_21

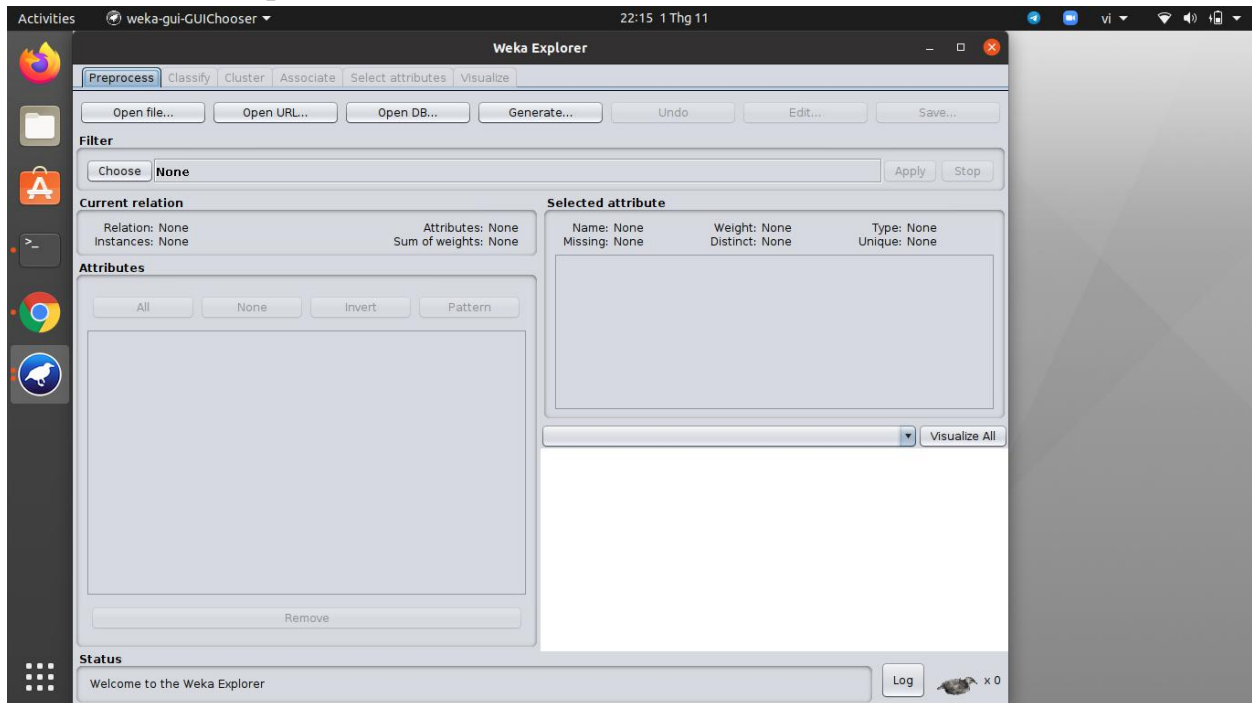
Giáo viên: GS.TS Lê Hoài Bắc

Trợ giảng: Nguyễn Khánh Toàn, Lê Minh Nhật

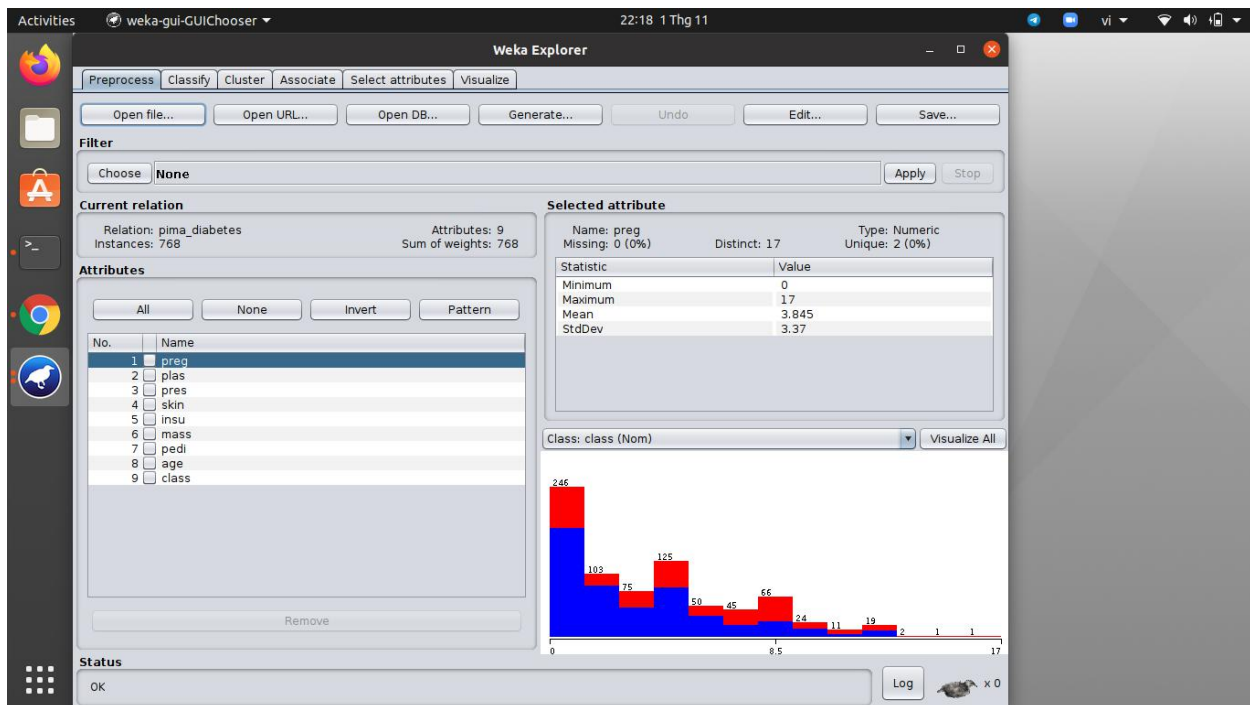
Mã số sinh viên	Họ và tên
19120207	Hồ Hoàng Duy
19120364	Nguyễn Đắc Thắng

Yêu cầu 1: Cài đặt Weka (1đ)

- Chụp hình giao diện chức năng Explorer cùng màn hình desktop và báo cáo lại ảnh chụp.



- Sinh viên tìm thư mục data trong thư mục cài đặt của Weka và mở một tập dữ liệu bất kì (có phần mở rộng là arff). Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess. Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.



? Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess.

- Current relation: Cho biết các thông tin chung liên quan đến tập dữ liệu đang được thực thi, như: tên tập dữ liệu, số mẫu, số tập thuộc tính, ...
- Attributes: Hiển thị danh sách các thuộc tính hiện tại trong tập dữ liệu.
- Selected attribute: Thể hiện các thông tin có liên quan đến thống kê (thường trong trường hợp dữ liệu ở dưới dạng số) như trung bình cộng, giá trị nhỏ nhất, giá trị lớn nhất của thuộc tính đã chọn trước ở phần Attributes. Đối với dữ liệu dạng định danh, phần mềm sẽ cung cấp danh sách định danh và số lượng mỗi định danh.

? Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.

- Preprocess: Tiền xử lý (tab mặc định khi truy cập giao diện Explorer)
- Classify: Phân lớp dữ liệu
- Cluster: Gom nhóm dữ liệu
- Associate: Khai thác luật kết hợp
- Select Attributes: Lựa chọn thuộc tính (để xét sự tương quan giữa các thuộc tính)
- Visualize: Trực quan hóa dữ liệu

Yêu cầu 2: Làm quen với Weka (6đ)

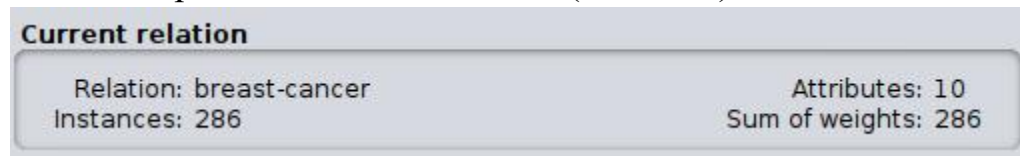
2.1 Đọc dữ liệu vào Weka: đọc tập dữ liệu breast_cancer.arff

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

Trả lời: Tập dữ liệu có 286 mẫu (instances).

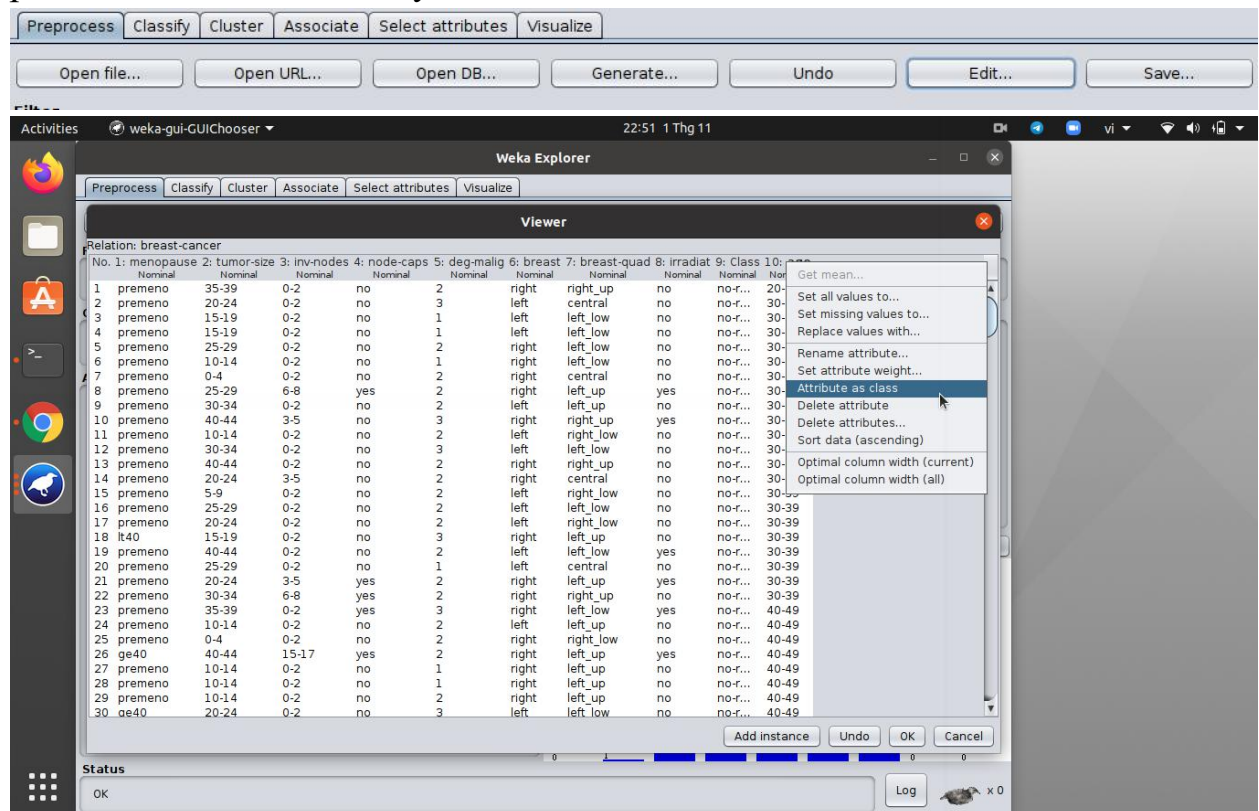
- Tập dữ liệu có bao nhiêu thuộc tính (attributes)?

Trả lời: Tập dữ liệu có 10 thuộc tính (attributes).



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

Trả lời: Thuộc tính “Class” được dùng làm lớp (class). Có thể thay đổi thuộc tính dùng làm lớp, bằng cách chọn Edit trong tab Preprocess, click chuột phải vào thuộc tính cần thay đổi và chọn Attributes as class.



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

Trả lời: Có 2 thuộc tính bị thiếu dữ liệu (missing values): breast-quad và node-caps. Thuộc tính breast-quad thiếu dữ liệu ít nhất (1 mẫu), thuộc tính node-caps thiếu dữ liệu nhiều nhất (8 mẫu).

Selected attribute			
Name: breast-quad		Type: Nominal	
Missing: 1 (0%)		Unique: 0 (0%)	
		Distinct: 5	
No.	Label	Count	Weight
1	left_up	97	97.0
2	left_low	110	110.0
3	right_up	33	33.0
Attributes that match a reg. expression			24.0
5	central	21	21.0

Selected attribute			
Name: node-caps		Type: Nominal	
Missing: 8 (3%)		Unique: 0 (0%)	
		Distinct: 2	
No.	Label	Count	Weight
1	yes	56	56.0
2	no	222	222.0

Để giải quyết vấn đề missing value, ta có thể có các cách sau đây:

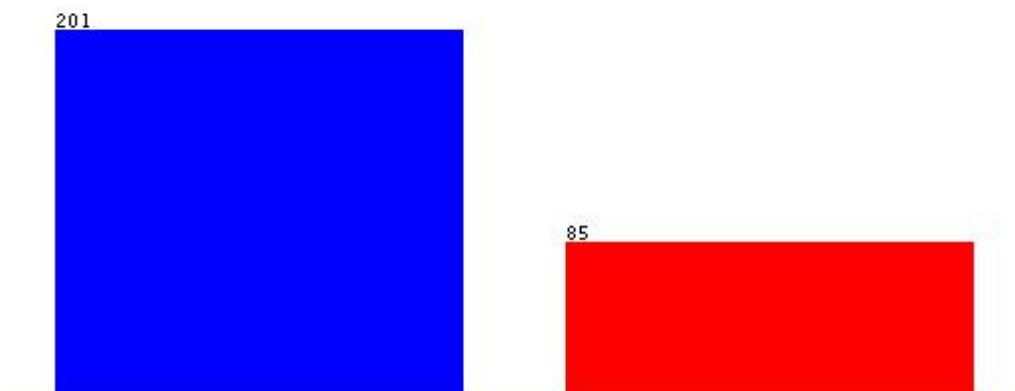
- + Loại bỏ dữ liệu bị thiếu.
- + Bổ sung dữ liệu thiếu (bằng các phương pháp lấy trung bình, mean, median (với thuộc tính numeric), mode (thuộc tính categorical) hoặc kNN)
- Giải thích ý nghĩa của đồ thị trong cửa sổ Explorer. Bạn đặt tên cho đồ thị này là gì? Màu xanh và màu đỏ có nghĩa gì? Đồ thị này biểu diễn cho cái gì? Đồ thị trong cửa sổ Explorer thể hiện các giá trị Count của từng Label trong cùng một thuộc tính. Các cột sắp xếp theo thứ tự của các Label và chiều cao của cột là giá trị Count tương ứng.
Có thể đặt tên cho đồ thị này là biểu đồ phân bố dựa theo lớp. Màu xanh biểu thị tại mỗi khoảng dữ liệu của thuộc tính đang được chọn, có bao nhiêu mẫu có kết quả no-recurrence-events, ngược lại màu đỏ cho kết quả recurrence-events

Selected attribute

Name: Class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	no-recurrence-ev...	201	201.0
2	recurrence-events	85	85.0

Class: Class (Nom) Visualize All



2.2 Khám phá tập dữ liệu Weather: tập dữ liệu weather.numeric.arff

- 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?

Trả lời: Tập dữ liệu có 5 thuộc tính, 14 mẫu

Current relation

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

Phân loại thuộc tính theo kiểu dữ liệu:

- + Thuộc tính số: temperature, humidity
- + Thuộc tính định danh: outlook, windy, play

Thuộc tính “play” là lớp.

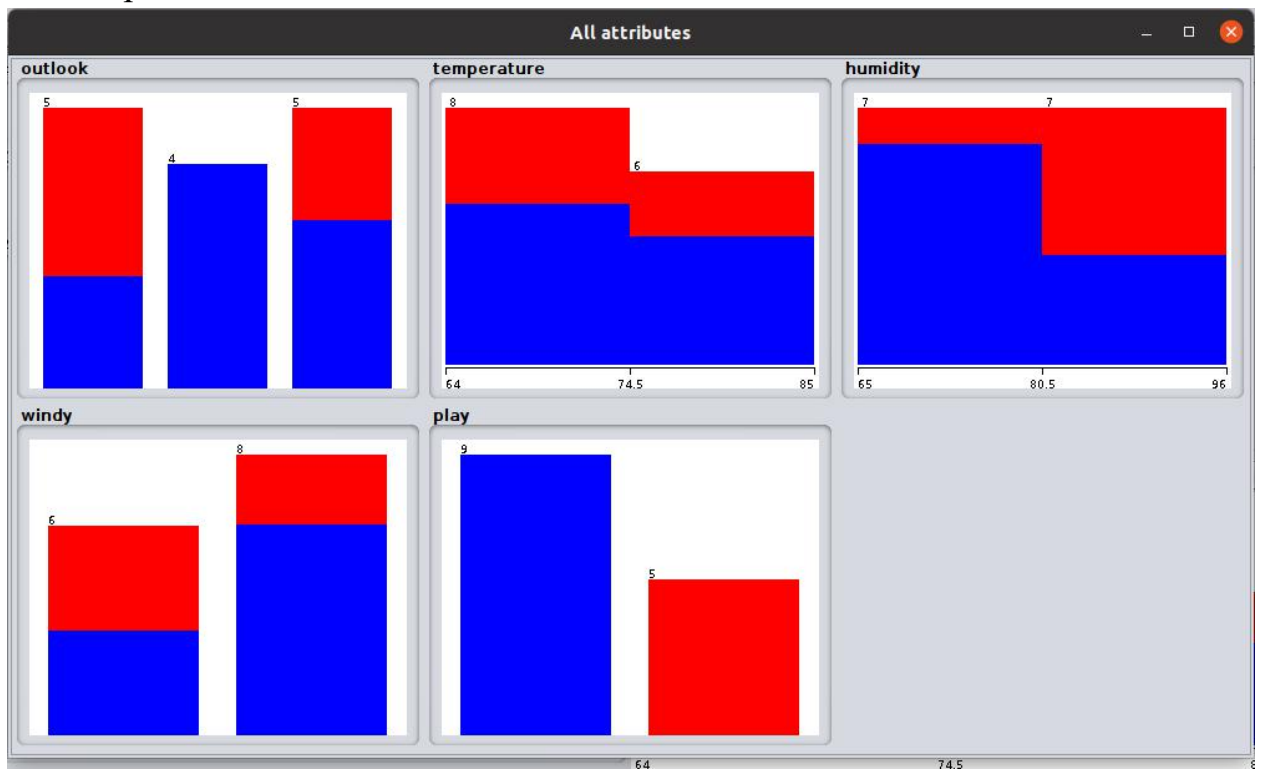
- Liệt kê five-number summary của thuộc tính temperature và humidity. Weka có cung cấp những giá trị này không?

Trả lời:

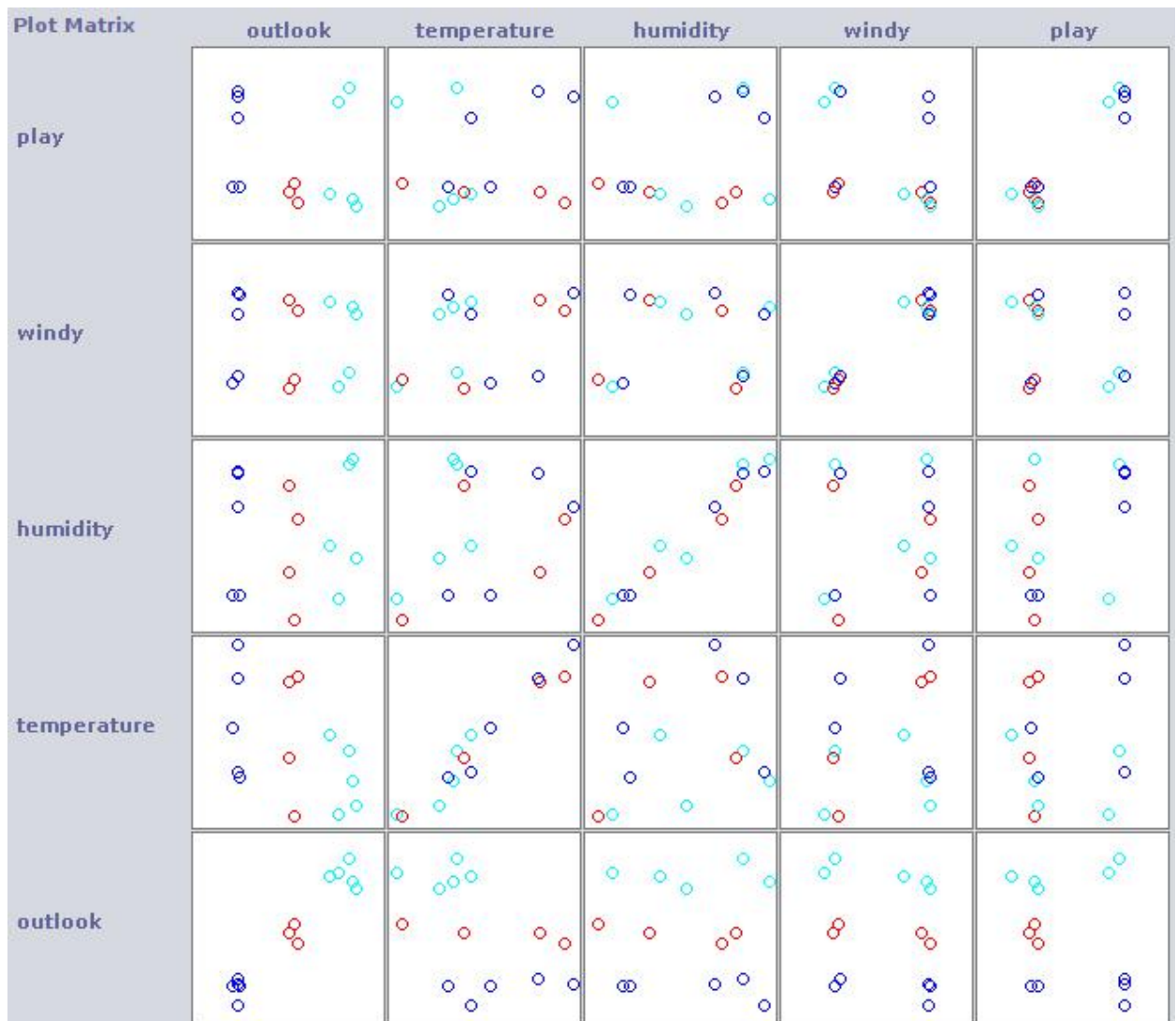
Five-number summary của temperature và humidity:

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

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



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



Theo quan sát của cá nhân thì có các cặp thuộc tính outlook – play, humidity – play, windy - play có vẻ tương quan với nhau.

2.3 Khám phá tập dữ liệu tín dụng Đức: tập dữ liệu credit-g.arff

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

Trả lời: Nội dung chú thích ở đầu tập tin mô tả về tập dữ liệu tín dụng Đức. Phần mô tả bao gồm tiêu đề, thông tin về nguồn gốc của tập tin, số lượng mẫu, số lượng thuộc tính và loại dữ liệu của các thuộc tính và mô tả chi tiết về thuộc tính. Ngoài ra, phần chú thích còn cho biết thêm về ma trận chi phí

và ánh xạ ý nghĩa của giá trị của thuộc tính với ký hiệu hiển thị trên giao diện.

Tập thuộc tính có 1000 mẫu với 21 thuộc tính.

Dưới đây là thông tin về 5/21 thuộc tính có trong tập dữ liệu:

- + duration (thuộc tính rời rạc): thời hạn vay tín dụng (tính theo tháng)
- + purpose (thuộc tính rời rạc): mục đích của việc vay tín dụng
 - Mua xe mới
 - Mua xe đã qua sử dụng
 - Mua nội thất, thiết bị
 - Mua TV/Radio
 - Mua thiết bị gia dụng
 - Sửa chữa
 - Giáo dục
 - Đi nghỉ dưỡng
 - Chi phí đào tạo lại
 - Kinh doanh
 - Khác
- + housing (thuộc tính rời rạc): chỗ ở, bao gồm 3 trạng thái
 - Thuê
 - Sở hữu nhà
 - Tự do
- + saving_status (thuộc tính liên tục): tài khoản tiết kiệm, được chia ra các mức
 - Dưới 100 Mark Đức
 - Từ 100 đến 500 Mark Đức
 - Từ 500 đến 1000 Mark Đức
 - Trên 1000 Mark Đức
 - Không xác định
- + check_status (thuộc tính liên tục): trạng thái của tài khoản séc hiện có
 - Dưới 0
 - Từ 0 đến 200 Mark Đức
 - Trên 200 Mark Đức
 - Không có tài khoản séc
- Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?

Trả lời: Thuộc tính “class” là thuộc tính lớp, phân bố lệch về “good”.

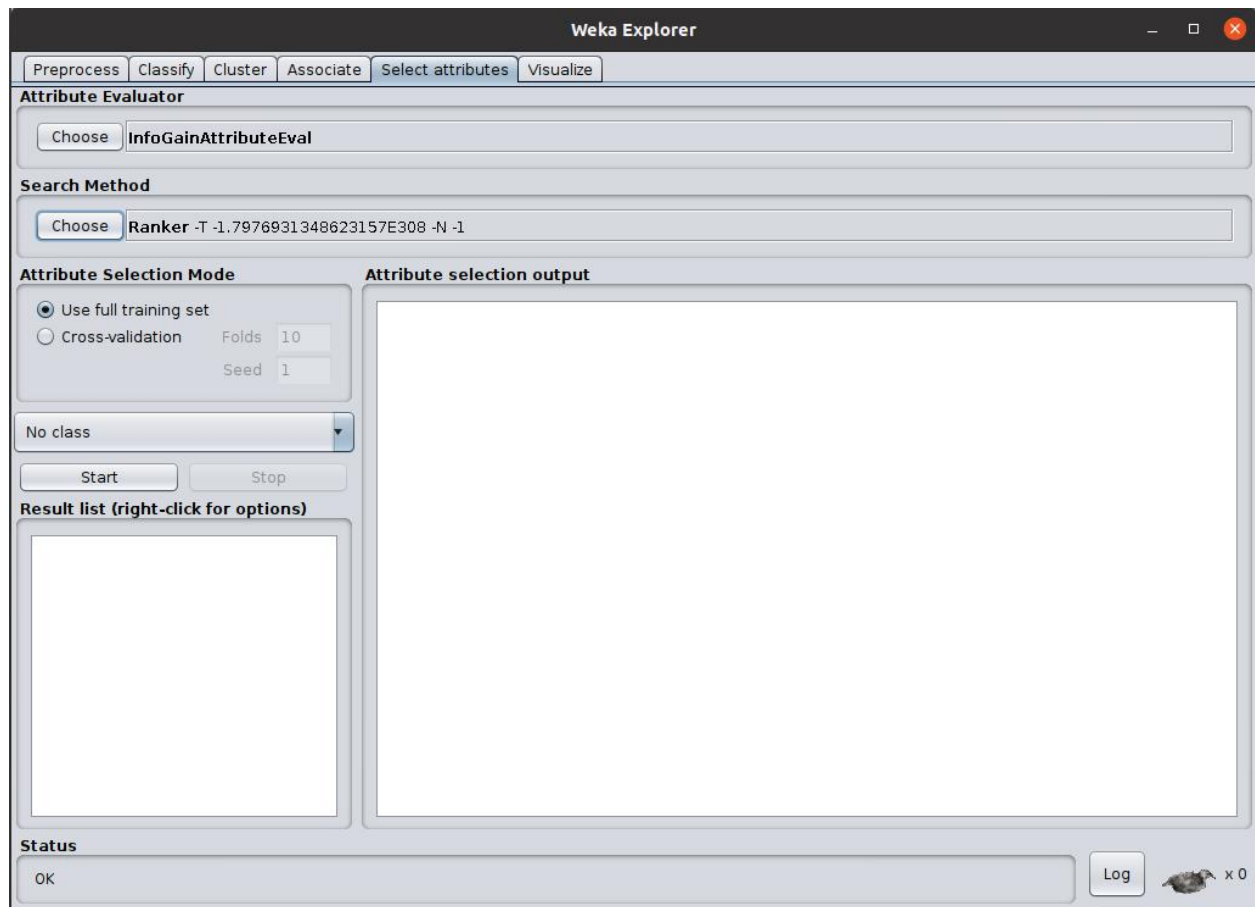
- Sử dụng tab Select attributes. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.

Trả lời:

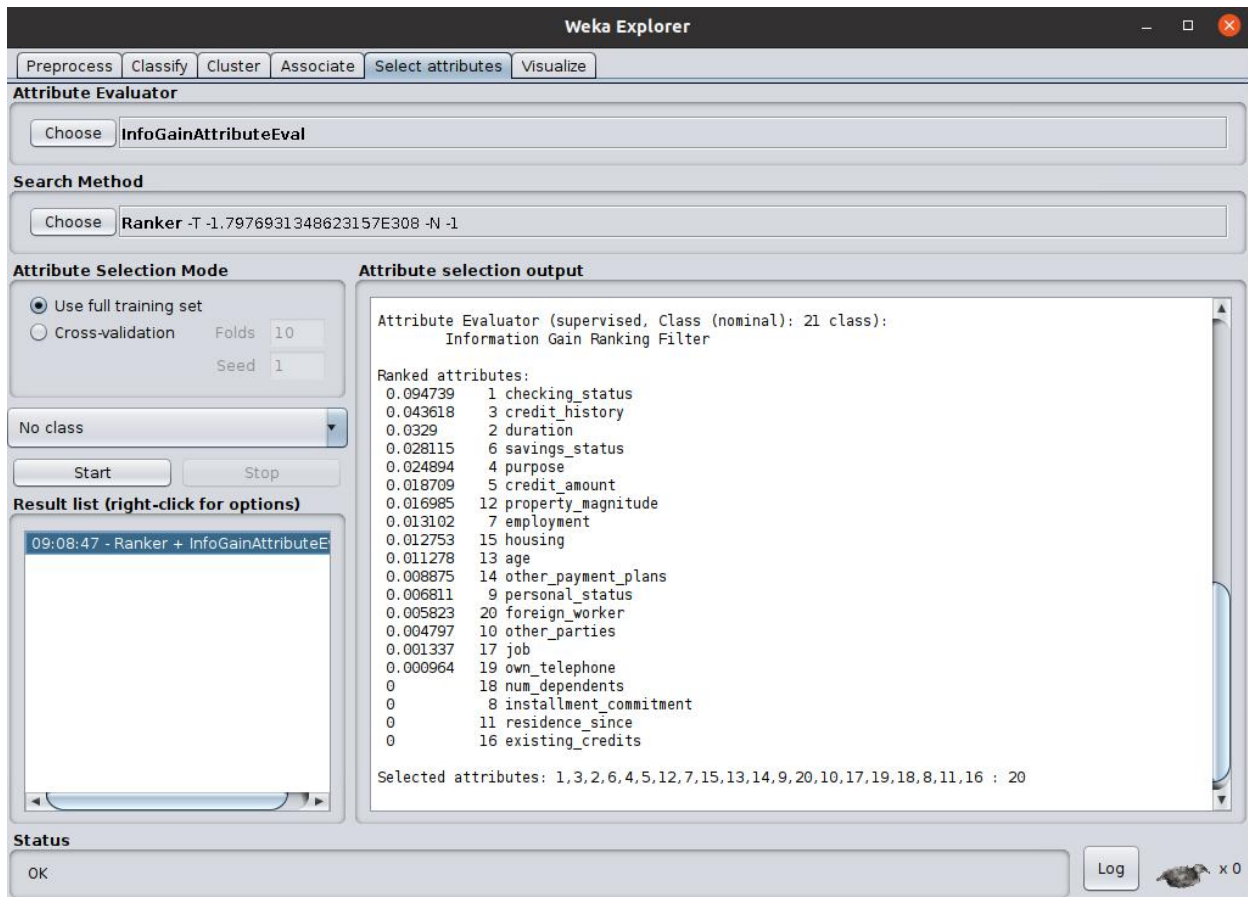
- + GainRatioAttributeEval: được sử dụng trong thuật toán C4.5 do Quinlan đưa ra. Ý tưởng của thuật toán là xét tất cả các phép thử có thể phân chia tập dữ liệu đã cho và chọn 1 phép thử cho GainRatio nhỏ nhất.
- + InfoGainAttributeEval: Đo mức hiệu quả của 1 thuộc tính trong bài toán phân lớp dữ liệu.
- + Ngoài ra còn có nhiều options lọc thuộc tính, phụ thuộc vào dụng ý của người sử dụng
- Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng.

Trả lời: Theo như mô tả trên thì việc chọn GainRatioAttributeEval hay InfoGainAttributeEval đều cho biết về các thuộc tính có tương quan cao nhất đối với thuộc tính lớp. Trong báo cáo, người viết chọn InfoGainAttributeEval để chọn các thuộc tính này.

- + B1: Chọn tab Select attributes. Tại mục Attribute Evaluator, chọn InfoGainAttributeEval. Tại mục Search Method, chọn Ranker.



- + B2: Nhấn nút Start
- + B3: Đọc kết quả



Từ hình trên, ta có thể thấy 5 thuộc tính có tương quan cao nhất với thuộc tính lớp là checking_status, credit_history, duration, saving_status, purpose

Yêu cầu 3: Cài đặt tiền xử lý dữ liệu (5đ)

Source code của nhóm được đặt trong thư mục Source (cùng cấp với file báo cáo này), ở đây nhóm chỉ đề cập cú pháp tham số dòng lệnh khi chạy chương trình và báo cáo kết quả, nhận xét sau khi chạy chương trình.

1. Liệt kê các cột bị thiếu dữ liệu

- Cú pháp: `python3 list-missing.py house-prices.csv`
- Kết quả:

```
(base) hhduy@HP:~/Documents/Lab/Data Mining/Lab01/Source$ python3 list-missing.py house-prices.csv
---Columns that have missing values---
LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature,
There are 18 missing attributes!
```

2. Đếm số dòng bị thiếu dữ liệu

- Cú pháp: `python row-missing.py house-prices.csv`
- Kết quả:

LotFrontage: 173	FireplaceQu: 501
Alley: 941	GarageType: 60
MasVnrType: 593	GarageYrBlt: 60
MasVnrArea: 10	GarageFinish: 60
BsmtQual: 27	GarageQual: 60
BsmtCond: 27	GarageCond: 60
BsmtExposure: 28	PoolQC: 1000
BsmtFinType1: 27	Fence: 815
BsmtFinType2: 29	MiscFeature: 963

```
(base) hhduty@HP:~/Documents/Lab/Data Mining/Lab01/Source$ python3 row-missing.py house-prices.csv
--Rows that have missing values--
LotFrontage: 173
Alley: 941
MasVnrType: 593
MasVnrArea: 10
BsmtQual: 27
BsmtCond: 27
BsmtExposure: 28
BsmtFinType1: 27
BsmtFinType2: 29
FireplaceQu: 501
GarageType: 60
GarageYrBlt: 60
GarageFinish: 60
GarageQual: 60
GarageCond: 60
PoolQC: 1000
Fence: 815
MiscFeature: 963
```

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

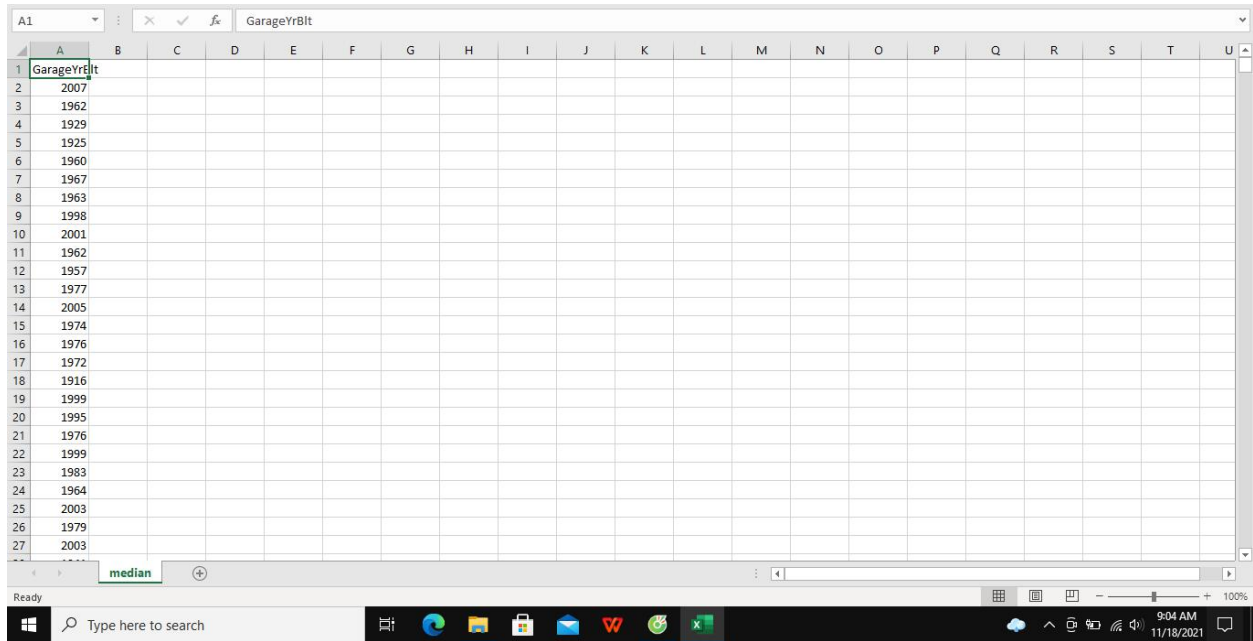
- Cú pháp: `python fill-missing-value.py house-prices.csv --method=... -columns=..... --out = <tên file đã chuẩn hóa>`
(với `--method=...` là tên phương pháp điền vào các giá trị thiếu (mean, median, mode); `--columns=...` là tên của cột cần điền vào)
- Kết quả:
 - ❖ Các cột có thể dùng phương pháp mean, median(cột dữ liệu số):

LotFrontage
MasVnrArea
GarageYrBlt

Test case 1: method = mean, columns = LotFrontage

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	LotFrontage																				
2	83																				
3	70																				
4	50																				
5	52																				
6	69																				
7	65																				
8	80																				
9	32																				
10	71																				
11	52																				
12	70																				
13	71																				
14	60																				
15	70																				
16	69																				
17	36																				
18	34																				
19	35																				
20	51																				
21	44																				
22	108																				
23	71																				
24	80																				
25	37																				
26	56																				
27	85																				

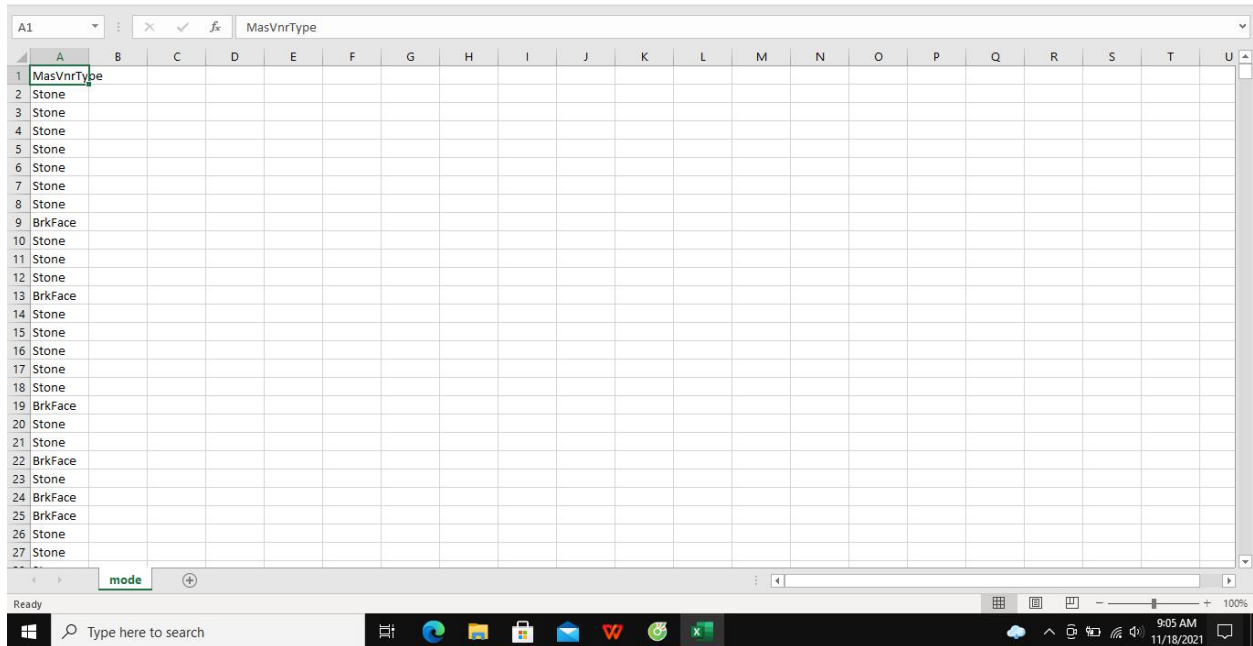
Test case 2: method = median, column = GarageYrBlt



❖ Các cột có thể dùng phương pháp mode (cột dữ liệu phân loại):

Alley	BsmtCond
MasVnrType	BsmtExposure
BsmtQual	BsmtFinType1
BsmtFinType2	FireplaceQu
GarageType	GarageFinish
GarageQual	GarageCond
Fence	MiscFeature

Test case 3: method = mode, columns = MasVnrType



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	MasVnrType																				
2	Stone																				
3	Stone																				
4	Stone																				
5	Stone																				
6	Stone																				
7	Stone																				
8	Stone																				
9	BrkFace																				
10	Stone																				
11	Stone																				
12	Stone																				
13	BrkFace																				
14	Stone																				
15	Stone																				
16	Stone																				
17	Stone																				
18	Stone																				
19	BrkFace																				
20	Stone																				
21	Stone																				
22	BrkFace																				
23	Stone																				
24	BrkFace																				
25	BrkFace																				
26	Stone																				
27	Stone																				

Test case 4: method = mode, columns = Alley

❖ Nhận xét:

- Cột LotFrontage có mean = 69, median = 64 (trước khi điền vào các dòng bị thiếu giá trị).
 - Cột MasVnrArea có mean = 108
 - Cột Alley có mode là 'Grvl', cột MasVnrType có mode là 'BrkFace'.
4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các dòng bị thiếu hơn 50% giá trị các thuộc tính).
- Cú pháp: `python del-missing-row.py house-prices.csv --lim=... --out=del-row.csv`
(với `--lim= ...` là ngưỡng tỉ lệ thiếu người dùng tự xác định trong chương trình, như: 0.3, 0.4, 0.5, 0.6, ... tương ứng 30%, 40%, 50%, 60%, ...; `--out` là tên file ghi kết quả sau khi xử lý.)
 - Kết quả
5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu).
- Cú pháp: `python del-missing-columns.py house-prices.csv --lim=... --out=del-column.csv`

(với --lim= ... là ngưỡng tỉ lệ thiếu người dùng tự xác định trong chương trình, như: 0.3, 0.4, 0.5, 0.6, ... tương ứng 30%, 40%, 50%, 60%, ...; --out là tên file ghi kết quả sau khi xử lí.)

- Kết quả: khi lim = 0.5 thì các cột bị xóa là Alley, MasVnrType, FireplaceQu, PoolQC, Fence, MiscFeature.

6. Xóa các mẫu bị trùng lặp.

- Cú pháp: `python3 del_duplicate.py house-prices.csv <tên file sau khi tiến hành xóa mẫu>`
- Kết quả: từ file dữ liệu ban đầu có 1000 dòng, sau khi xóa mẫu trùng lặp thì ta có file dữ liệu mới có 716 dòng.

```
(base) hhdhuy@HP:~/Documents/Lab/Data Mining/Lab01/Source$ python3 del_duplicate.py house-prices.csv not_duplicate.csv
1001
Id,MSSubClass,MSZoning,LotFrontage,LotArea,Street,Alley,LotShape,LandContour,Utilities,LotConfig,LandSlope,Neighborhood,Condit
ion1,Condition2,BldgType,HouseStyle,OverallQual,OverallCond,YearBuilt,YearRemodAdd,RoofStyle,RoofMatl,Exterior1st,Exterior2nd,
MasVnrType,MasVnrArea,ExterQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinSF1,BsmtFinType2,BsmtF
inSF2,BsmtUnfSF,TotalBsmtSF,Heating,HeatingQC,CentralAir,Electrical,1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea,BsmtFullBath,Bsmt
HalfBath,FullBath,HalfBath,BedroomAbvGr,KitchenAbvGr,KitchenQual,TotRmsAbvGrd,Functional,Fireplaces,FireplaceQu,GarageType,Gar
ageYrBlt,GarageFinish,GarageCars,GarageArea,GarageQual,GarageCond,PavedDrive,WoodDeckSF,OpenPorchSF,EnclosedPorch,3SsnPorch,Sc
reenPorch,PoolArea,PoolQC,Fence,MiscFeature,MiscVal,MoSold,YrSold,SaleType,SaleCondition,SalePrice

1242,20,RL,83.0,9849,Pave,,Reg,Lvl,AllPub,Inside,Gtl,Somerst,Norm,Norm,1Fam,1Story,7,6,2007,2007,Hip,CompShg,VinylSd,VinylSd,S
tone,0.0,Gd,TA,PConc,Gd,TA,Av,Unf,0,Unf,0,1689,1689,GasA,Ex,Y,SBrkr,1689,0,0,1689,0,0,2,0,3,1,Gd,7,Typ,0,,Attchd,2007.0,RFn,3,
954,TA,TA,Y,0,56,0,0,0,0,,,0,6,2007,New,Partial,248328

1233,90,RL,70.0,9842,Pave,,Reg,Lvl,AllPub,FR2,Gtl,mes,Norm,Norm,Duplex,1Story,4,5,1962,1962,Gable,CompShg,HdBoard,HdBoard,,0.0
,TA,TA,Slab,,,,,0,0,0,0,GasA,TA,Y,SBrkr,1224,0,0,1224,0,0,2,0,2,2,TA,6,Typ,0,,CarPort,1962.0,Unf,2,462,TA,TA,Y,0,0,0,0,0,0,,,
,0,3,2007,WD,Normal,101800

-----
717
Id,MSSubClass,MSZoning,LotFrontage,LotArea,Street,Alley,LotShape,LandContour,Utilities,LotConfig,LandSlope,Neighborhood,Condit
ion1,Condition2,BldgType,HouseStyle,OverallQual,OverallCond,YearBuilt,YearRemodAdd,RoofStyle,RoofMatl,Exterior1st,Exterior2nd,
MasVnrType,MasVnrArea,ExterQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinSF1,BsmtFinType2,BsmtF
inSF2,BsmtUnfSF,TotalBsmtSF,Heating,HeatingQC,CentralAir,Electrical,1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea,BsmtFullBath,Bsmt
HalfBath,FullBath,HalfBath,BedroomAbvGr,KitchenAbvGr,KitchenQual,TotRmsAbvGrd,Functional,Fireplaces,FireplaceQu,GarageType,Gar
ageYrBlt,GarageFinish,GarageCars,GarageArea,GarageQual,GarageCond,PavedDrive,WoodDeckSF,OpenPorchSF,EnclosedPorch,3SsnPorch,Sc
reenPorch,PoolArea,PoolQC,Fence,MiscFeature,MiscVal,MoSold,YrSold,SaleType,SaleCondition,SalePrice

1233,90,RL,70.0,9842,Pave,,Reg,Lvl,AllPub,FR2,Gtl,mes,Norm,Norm,Duplex,1Story,4,5,1962,1962,Gable,CompShg,HdBoard,HdBoard,,0.0
,TA,TA,Slab,,,,,0,0,0,0,GasA,TA,Y,SBrkr,1224,0,0,1224,0,0,2,0,2,2,TA,6,Typ,0,,CarPort,1962.0,Unf,2,462,TA,TA,Y,0,0,0,0,0,0,,,
,0,3,2007,WD,Normal,101800

1401,50,RM,50.0,6000,Pave,,Reg,Lvl,AllPub,Corner,Gtl,BrkSide,Norm,Norm,1Fam,1.5Fin,6,7,1929,1950,Gable,CompShg,WdShing,Wd Shng
,,0.0,TA,TA,BrkTil,TA,TA,No,Unf,0,Unf,0,862,862,GasA,TA,Y,SBrkr,950,208,0,1158,0,0,1,0,3,1,TA,5,Typ,1,Gd,BuiltIn,1929.0,RFn,1,
208,TA,TA,Y,0,0,112,0,0,0,,,0,7,2008,WD,Normal,120000
```

AutoSave not_duplicate.csv Search HỒ HOÀNG DUY

File Home Insert Page Layout Formulas Data Review View Help

Share Comments

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	MSSubClas	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighbori	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCo	YearBuilt	YearR
2	1233	90 RL	70	9842	Pave			Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962	1
3	1401	50 RM	50	6000	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929	1
4	1377	30 RL	52	6292	Pave			Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930	1
5	208	20 RL		12493	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960	1
6	980	20 RL	80	8816	Pave			Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963	1
7	484	120 RM	32	4500	Pave			Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998	1
8	730	30 RM	52	6240	Pave			Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925	1
9	1021	20 RL	60	7024	Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005	2
10	1025	20 RL		15498	Pave			IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976	1
11	457	70 RM	34	4571	Pave			Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916	1
12	695	50 RM	51	6120	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936	1
13	24	120 RM	44	4224	Pave			Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1
14	1258	30 RL	56	4060	Pave			Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	8	1922	1
15	1213	30 RL	50	9340	Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	6	1941	1
16	71	20 RL	95	13651	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	7	6	1973	1
17	700	120 FV	59	4282	Pave			IR2	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	TwnhsE	1Story	7	5	2004	2
18	532	70 RM	60	6155	Pave			IR1	Lvl	AllPub	FR3	Gtl	BrkSide	RRNn	Feedr	1Fam	2Story	6	8	1920	1
19	1326	30 RM	40	3636	Pave			Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1Story	4	4	1922	1
20	988	20 RL	83	10159	Pave			IR1	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	1Fam	1Story	9	5	2009	2
21	590	40 RM	50	9100	Pave			Reg	Lvl	AllPub	Inside	Gtl	BrkSide	RRAn	Feedr	1Fam	1Story	5	6	1930	1
22	507	60 RL	80	9554	Pave			IR1	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	2Story	8	5	1993	1
23	919	60 RL	103	13125	Pave			IR1	Lvl	AllPub	Corner	Gtl	SawyerW	Norm	Norm	1Fam	2Story	7	5	1991	1
24	1240	20 RL	64	9037	Pave			IR1	HLS	AllPub	Inside	Gtl	Timber	Norm	Norm	1Fam	1Story	8	5	2006	2
25	1206	20 RL	90	14684	Pave			IR1	Lvl	AllPub	CulDSac	Gtl	SawyerW	Norm	Norm	1Fam	1Story	7	7	1990	1
26	831	20 RL	80	11900	Pave			IR1	Lvl	AllPub	Corner	Gtl	mes	Norm	Norm	1Fam	1Story	6	5	1957	1
27	827	45 RM	50	6130	Pave			Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5Unf	5	6	1924	1

not_duplicate

Ready Type here to search

AutoSave not_duplicate.csv Search HỒ HOÀNG DUY

File Home Insert Page Layout Formulas Data Review View Help

Share Comments

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
695	1454	20 RL	90	17217	Pave			Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	5	5	2006	2
696	174	20 RL	80	10197	Pave			IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	6	5	1961	1
697	1236	70 RL	96	13132	Pave			Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	5	5	1914	1
698	979	20 RL	68	9450	Pave			Reg	Bnk	AllPub	Inside	Mod	Edwards	Norm	Norm	1Fam	1Story	4	5	1954	1
699	213	60 FV	72	8640	Pave			Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5	2009	2
700	458	20 RL		53227	Pave			IR1	Low	AllPub	CulDSac	Mod	ClearCr	Norm	Norm	1Fam	1Story	4	6	1954	1
701	62	75 RM	60	7200	Pave			Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	2.5Unf	5	7	1920	1
702	211	30 RL	67	5604	Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	5	6	1925	1
703	826	20 RL	114	14803	Pave			Reg	Lvl	AllPub	Inside	Gtl	NridgHt	PosN	Norm	1Fam	1Story	10	5	2007	2
704	985	90 RL	75	10125	Pave			Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	Duplex	1.5Fin	5	5	1977	1
705	1253	20 RL	62	9858	Pave			Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	5	6	1968	1
706	1053	60 RL	100	9500	Pave			Reg	Lvl	AllPub	Corner	Gtl	mes	Artery	Norm	1Fam	2Story	6	6	1964	1
707	582	20 RL	98	12704	Pave			Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2008	2
708	1420	20 RL		16381	Pave			IR1	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	1Story	6	5	1969	1
709	1417	190 RM	60	11340	Pave			Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon	2Story	4	6	1885	1
710	668	20 RL	65	8125	Pave			Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	1Story	6	5	1994	1
711	394	30 RL		7446	Pave			Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Feedr	Norm	1Fam	1Story	4	5	1941	1
712	554	20 RL	67	8777	Pave			Reg	Lvl	AllPub	Inside	Gtl	Edwards	Feedr	Norm	1Fam	1Story	4	5	1949	2
713	1190	60 RL	60	7500	Pave			Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5	1999	1
714	192	60 RL		7472	Pave			IR1	Lvl	AllPub	CulDSac	Gtl	mes	Norm	Norm	1Fam	2Story	7	9	1972	2
715	990	60 FV	65	8125	Pave			Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5	2006	2
716	982	60 RL	98	12203	Pave			IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	1998	1
717	862	190 RL	75	11625	Pave			Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	2fmCon	1Story	5	4	1965	1
718																					
719																					
720																					
721																					

not_duplicate

Ready Type here to search

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

- Cú pháp: `python normalize.py house-prices.csv --method=... --column=... --out=...`
 method có thể là: minmax hoặc zscore.
 column chỉ có thể điền vào 1 tên cột (là 1 thuộc tính numeric)
 out là tên file kết quả trả về sau khi chuẩn hóa.

Các thuộc tính numeric có thể chuẩn hóa:

MSSubClass	MasVnrArea	1stFlrSF	BsmtHalfBath
LotFrontage	BsmtFinSF1	2ndFlrSF	FullBath
LotArea	BsmtFinSF2	LowQualFinSF	HalfBath
OverallQual	BsmtUnfSF	GrLivArea	BedroomAbvGr
OverallCond	TotalBsmtSF	BsmtFullBath	KitchenAbvGr
TotRmsAbvGrd	Fireplaces	GarageCars	GarageArea
WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch
ScreenPorch	PoolArea	MiscVal	MoSold
SalePrice			

Với các thuộc tính bị thiếu giá trị như LotFrontage, MasVnrArea, GarageYrBlt thì nhóm có điền vào các giá trị bị thiếu bằng phương pháp mean (như đã trình bày ở chức năng 3).

- Kết quả

	Min-Max	Z-Score
LotFrontage	min = 21 max = 153	mean = 69.251 sd= 19.33
MSSubClass	min = 20 max = 190	mean = 56.35 sd = 42.83
MasVnrArea	Min = 0 Max = 1600	Mean = 108.579 Sd = 188.85
GarageArea	Min = 0 Max = 1390	Mean = 483.783 Sd = 223.601

Hình ảnh demo chuẩn hóa Min-Max và Z-Score với thuộc tính MSSubClass

```
(base) hhdny@HP:~/Documents/Lab/Data Mining/Lab01/Source$ python3 normalize.py house-prices.csv --method=minmax --column=MSSub
Class --out=normalize1.csv
-----MinMax Normalization-----
Range: 170.0
Min: 20.0
Mean: 56.35
```


The image displays a Windows desktop environment with three main components: two Excel spreadsheets and a terminal window.

Excel Spreadsheet 1 (normalize1.csv): This spreadsheet shows a column labeled 'MSSubClass' with 27 rows of data. The values are: 0, 0.412, 0.176, 0.059, 0, 0.412, 0, 0.588, 0.235, 0.059, 0, 0, 0, 0, 0, 0.294, 0.824, 0.176, 0.588, 0.235, 0, 0.235, 0.588, 0.059, 0.235.

Terminal Window: The terminal shows the execution of a Python script to perform z-score normalization. The command is: `python3 normalize.py house-prices.csv --method=zscore --column=MSSubClass --out=normalize2.csv`. The output indicates: `-----Z-score Normalization-----`, `Mean: 56.35`, and `Sd: 42.83138452116635`.

Excel Spreadsheet 2 (normalize2.csv): This spreadsheet shows the same 'MSSubClass' column but with z-score normalized values. The values are: -0.849, 0.786, -0.148, -0.615, -0.849, 0.786, -0.849, 1.486, 0.085, -0.615, -0.849, -0.849, -0.849, -0.849, -0.849, 0.319, 2.42, -0.148, 1.486, 0.085, -0.849, 0.085, 1.486, -0.615, 0.085.

8. Tính giá trị biểu thức thuộc tính: ví dụ đối với một tập dữ liệu có chứa 2 thuộc tính width và height thì biểu thức $\text{width} * \text{height}$ sẽ trả về tập dữ liệu cũ với một thuộc tính mới có giá trị ở mỗi mẫu là tích của thuộc tính width và height trong mẫu tương ứng, với điều kiện cả 2 giá trị width và height đều

không bị thiếu, trong trường hợp bị thiếu thì giá trị biểu thức coi như bị thiếu.
Lưu ý: biểu thức có thể có nhiều thuộc tính và nhiều phép toán bao gồm cộng, trừ, nhân, chia.

- Cú pháp: `python expression_calculate.py -i <input_file.csv> -o <output_file.csv>`

Màn hình console sẽ hiện ra, người dùng chỉ cần nhập tên biểu thức và các toán tử xuất hiện giữa chúng (như trong input của các test case); sau đó nhập tên cho cột mới để lưu lại giá trị trong từng dòng ứng với cột đó.

- Kết quả: xuất hiện cột thuộc tính mới ở cuối cùng so với các cột đã có (xem hình)

Test case 1:

INPUT: $(\text{OverallQual} + \text{OverallCond}) * \text{YearBuilt} / \text{YearRemodAdd}$

OUTPUT:

	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM
1	ScreenPor	PoolArea	PoolQC	Fence	MiscFeatu	MiscVal	MoSold	YrSold	SaleType	SaleCondi	SalePrice	test1									
2	0	0	nan	nan	nan	0	6	2007	New	Partial	248328	13									
3	0	0	nan	nan	nan	0	3	2007	WD	Normal	101800	9									
4	0	0	nan	nan	nan	0	7	2008	WD	Normal	120000	12.86									
5	0	0	nan	nan	nan	0	4	2008	WD	Normal	91000	10.88718									
6	0	0	nan	GdWo	nan	0	4	2008	WD	Normal	141000	9									
7	0	0	nan	nan	nan	0	4	2009	WD	Normal	124000	10									
8	0	0	nan	MnPrv	nan	0	6	2009	WD	Normal	139000	11									
9	0	0	nan	nan	nan	0	5	2006	WD	Normal	164000	11									
10	0	0	nan	nan	nan	0	6	2009	WD	Normal	215000	10.99451									
11	0	0	nan	nan	nan	0	1	2009	WD	Normal	103000	8.884615									
12	0	0	nan	nan	nan	0	6	2010	WD	Normal	145000	11									
13	0	0	nan	MnPrv	nan	0	10	2006	WD	Normal	146000	12.78529									
14	0	0	nan	nan	nan	0	6	2008	WD	Normal	176000	9									
15	0	0	nan	GdWo	nan	0	6	2007	WD	Normal	123000	9									
16	0	0	nan	nan	nan	0	5	2008	COD	Abnorml	287000	14									
17	0	0	nan	nan	nan	0	8	2009	WD	Normal	133500	11									
18	0	0	nan	nan	nan	0	5	2008	COD	Abnorml	98000	9.825641									
19	0	0	nan	nan	nan	0	3	2006	WD	Normal	183900	12									
20	0	0	nan	MnPrv	nan	0	4	2009	WD	Normal	141500	10.92103									
21	0	0	nan	nan	nan	0	6	2007	WD	Normal	129900	12									
22	0	0	nan	nan	nan	0	5	2010	WD	Normal	333168	14									
23	0	0	nan	nan	nan	0	6	2007	WD	Normal	134000	11									
24	189	0	nan	nan	nan	0	6	2008	WD	Normal	167900	12									
25	0	0	nan	nan	nan	0	3	2008	WD	Normal	136500	11									
26	0	0	nan	nan	nan	0	7	2009	WD	Normal	99900	12.81333									
27	0	0	nan	nan	nan	0	8	2008	WD	Normal	305000	13									

Test case 2:

INPUT: $(\text{MSSubClass} + \text{LotFrontage}) / 2$

OUTPUT:

AutoSave out2.csv HỒ HOÀNG DUY

File Home Insert Page Layout Formulas Data Review View Help

CD1 test2

	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP
1	Fence	MiscFeatu	MiscVal	MoSold	YrSold	SaleType	SaleCondi	SalePrice	test2												
2	nan	nan	0	6	2007	New	Partial	248328	51.5												
3	nan	nan	0	3	2007	WD	Normal	101800	80												
4	nan	nan	0	7	2008	WD	Normal	120000	50												
5	nan	nan	0	4	2008	WD	Normal	91000	41												
6	GdWo	nan	0	4	2008	WD	Normal	141000													
7	nan	nan	0	4	2009	WD	Normal	124000	77.5												
8	MnPrv	nan	0	6	2009	WD	Normal	139000	50												
9	nan	nan	0	5	2006	WD	Normal	164000	76												
10	nan	nan	0	6	2009	WD	Normal	215000	65.5												
11	nan	nan	0	1	2009	WD	Normal	103000	41												
12	nan	nan	0	6	2010	WD	Normal	145000	45												
13	MnPrv	nan	0	10	2006	WD	Normal	146000	45.5												
14	nan	nan	0	6	2008	WD	Normal	176000	40												
15	GdWo	nan	0	6	2007	WD	Normal	123000	45												
16	nan	nan	0	5	2008	COD	Abnorml	287000													
17	nan	nan	0	8	2009	WD	Normal	133500	28												
18	nan	nan	0	5	2008	COD	Abnorml	98000	52												
19	nan	nan	0	3	2006	WD	Normal	183900	97.5												
20	MnPrv	nan	0	4	2009	WD	Normal	141500	50.5												
21	nan	nan	0	6	2007	WD	Normal	129900	82												
22	nan	nan	0	5	2010	WD	Normal	333168	84												
23	nan	nan	0	6	2007	WD	Normal	134000	45.5												
24	nan	nan	0	6	2008	WD	Normal	167900	70												
25	nan	nan	0	3	2008	WD	Normal	136500	78.5												
26	nan	nan	0	7	2009	WD	Normal	99900	43												
27	nan	nan	0	8	2008	WD	Normal	305000	72.5												

out2

Ready Average: 62.82285369 Count: 828 Sum: 51954.5

Type here to search

(những ô không có giá trị là những ô trong cột MSSubClass hoặc LotFrontage mà có giá trị rỗng nên chương trình sẽ không tính toán được kết quả)

Test case 3:

INPUT: LotFrontage * 50

OUTPUT:

AutoSave out3.csv HỒ HOÀNG DUY

File Home Insert Page Layout Formulas Data Review View Help

CD1 test3

	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ
1	OpenPorc	EnclosedF	3SsnPorch	ScreenPor	PoolArea	PoolQC	Fence	MiscFeatu	MiscVal	MoSold	YrSold	SaleType	SaleCondi	SalePrice	test3						
2	56	0	0	0	0	nan	nan	nan	0	6	2007	New	Partial	248328	4150						
3	0	0	0	0	0	nan	nan	nan	0	3	2007	WD	Normal	101800	3500						
4	0	112	0	0	0	nan	nan	nan	0	7	2008	WD	Normal	120000	2500						
5	141	0	0	0	0	nan	nan	nan	0	4	2008	WD	Normal	91000	2600						
6	0	0	0	0	0	nan	GdWo	nan	0	4	2008	WD	Normal	141000							
7	152	0	0	0	0	nan	nan	nan	0	4	2009	WD	Normal	124000	3250						
8	80	0	0	0	0	nan	MnPrv	nan	0	6	2009	WD	Normal	139000	4000						
9	125	0	0	0	0	nan	nan	nan	0	5	2006	WD	Normal	164000	1600						
10	192	0	0	0	0	nan	nan	nan	0	6	2009	WD	Normal	215000	3550						
11	23	112	0	0	0	nan	nan	nan	0	1	2009	WD	Normal	103000	2600						
12	0	0	0	0	0	nan	nan	nan	0	6	2010	WD	Normal	145000	3500						
13	64	0	0	0	0	nan	MnPrv	nan	0	10	2006	WD	Normal	146000	3550						
14	64	0	0	0	0	nan	nan	nan	0	6	2008	WD	Normal	176000	3000						
15	0	0	0	0	0	nan	GdWo	nan	0	6	2007	WD	Normal	123000	3500						
16	72	174	0	0	0	nan	nan	nan	0	5	2008	COD	Abnorml	287000							
17	0	0	0	0	0	nan	nan	nan	0	8	2009	WD	Normal	133500	1800						
18	0	96	0	0	0	nan	nan	nan	0	5	2008	COD	Abnorml	98000	1700						
19	34	0	0	0	0	nan	nan	nan	0	3	2006	WD	Normal	183900	1750						
20	0	0	0	0	0	nan	MnPrv	nan	0	4	2009	WD	Normal	141500	2550						
21	110	0	0	0	0	nan	nan	nan	0	6	2007	WD	Normal	129900	2200						
22	30	0	0	0	0	nan	nan	nan	0	5	2010	WD	Normal	333168	5400						
23	0	158	0	0	0	nan	nan	nan	0	6	2007	WD	Normal	134000	3550						
24	128	0	0	189	0	nan	nan	nan	0	6	2008	WD	Normal	167900	4000						
25	0	0	0	0	0	nan	nan	nan	0	3	2008	WD	Normal	136500	1850						
26	96	0	0	0	0	nan	nan	nan	0	7	2009	WD	Normal	99900	2800						
27	0	0	0	0	0	nan	nan	nan	0	8	2008	WD	Normal	305000	4250						

out3

Ready Average: 3465.175333 Count: 828 Sum: 2865700

Type here to search

Đánh giá đồ án:

MSSV	Họ và tên	Tỷ lệ hoàn thành	Tỷ lệ đóng góp
19120207	Hồ Hoàng Duy	100%	50%
19120364	Nguyễn Đức Thắng	100%	50%

Phân công đồ án:

MSSV	Công việc	Mức độ hoàn thành
19120207	Yêu cầu 1 + 2 + Chức năng 8 yêu cầu 3, kiểm tra lại lần cuối bài nộp.	Tốt
19120364	Chức năng 1 → 7 yêu cầu 3.	Tốt