

Bài 102: Xử lý dữ liệu

Đọc dữ liệu đã thu thập

```
import pandas as pd

df = pd.read_csv('https://thachln.github.io/datasets/bank/bank-
additional-full-draw.csv')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   age                   41188 non-null  int64
1   job                   41188 non-null  object
2   marital               41188 non-null  object
3   education             41188 non-null  object
4   default               41188 non-null  object
5   housing               41188 non-null  object
6   loan                  41188 non-null  object
7   contact               41188 non-null  object
8   month                 41188 non-null  object
9   day_of_week           41188 non-null  object
10  duration              41188 non-null  int64
11  campaign              41188 non-null  int64
12  pdays                 41185 non-null  float64
13  previous              41188 non-null  int64
14  poutcome              41188 non-null  object
15  emp.var.rate          41188 non-null  float64
16  cons.price.idx         41188 non-null  float64
17  cons.conf.idx         41188 non-null  float64
18  euribor3m             41188 non-null  float64
19  nr.employed           41188 non-null  float64
20  y                     41188 non-null  object
dtypes: float64(6), int64(4), object(11)
memory usage: 6.6+ MB
```

Lấy danh sách tên cột

```
print(df.columns.tolist())
```

```
['age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact',
'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome',
, 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed',
, 'y']
```

Xem danh sách các giá trị trong từng cột dữ liệu

Tuổi

```
df['age'].unique()
```

```
array([56, 57, 37, 40, 45, 59, 41, 24, 25, 29, 35, 54, 46, 50, 39, 30, 55,
      49, 34, 52, 58, 32, 38, 44, 42, 60, 53, 47, 51, 48, 33, 31, 43, 36,
      28, 27, 26, 22, 23, 20, 21, 61, 19, 18, 70, 66, 76, 67, 73, 88, 95,
      77, 68, 75, 63, 80, 62, 65, 72, 82, 64, 71, 69, 78, 85, 79, 83, 81,
      74, 17, 87, 91, 86, 98, 94, 84, 92, 89], dtype=int64)
```

Công việc

```
df['job'].unique()
```

```
array(['housemaid', 'services', 'admin.', 'blue-collar', 'technician',  
      'retired', 'management', 'unemployed', 'self-employed', 'unknown',  
      'entrepreneur', 'student'], dtype=object)
```

Cứ tiếp tục dùng lệnh `df['tên cột'].unique()` để xem dữ liệu các cột còn lại. Một cách khác là dùng vòng lặp để duyệt từng tên cột rồi hiển thị thông tin tự động.

```
for col in df.columns:  
    print('=====')  
    print('Dữ liệu cột ', col)  
    print(df[col].unique())
```

Kết quả từng cột được trình bày bên dưới.

Dữ liệu cột job

```
['housemaid' 'services' 'admin.' 'blue-collar' 'technician' 'retired'  
 'management' 'unemployed' 'self-employed' 'unknown' 'entrepreneur'  
 'student']
```

Dữ liệu cột marital

```
['married' 'single' 'divorced' 'unknown']
```

Dữ liệu cột job


```
['housemaid' 'services' 'admin.' 'blue-collar' 'technician' 'retired'  
 'management' 'unemployed' 'self-employed' 'unknown' 'entrepreneur'  
 'student']
```

Dữ liệu cột marital

```
['married' 'single' 'divorced' 'unknown']
```

Dữ liệu cột education

```
['basic.4y' 'high.school' 'basic.6y' 'basic.9y' 'professional.course'  
 'unknown' 'high school' 'High School' 'university.degree' 'illiterate']
```

 Bạn để ý các giá trị trong cột education có 3 loại dữ liệu mà tôi bôi đỏ cho thấy do nhập liệu bị sai. Cần phải thay đổi 2 giá trị 'high school' 'High School' thành 'high.school'.

Dữ liệu cột default

```
['no' 'unknown' 'yes']
```

Dữ liệu cột housing

```
['no' 'yes' 'unknown']
```

Dữ liệu cột loan


```
['no' 'yes' 'unknown']
```

Dữ liệu cột contact

```
['telephone' 'cellular']
```

Dữ liệu cột month

```
['may' 'May' 'jun' 'ju' 'aug' 'oct' 'nov' 'dec' 'mar' 'apr' 'sep']
```

 Tương tự tình huống trong cột education, có dữ liệu trong cột month nhập lúc lúc **may**, **May** (tháng 5). Cần phải thống nhất lại.

Dữ liệu cột day_of_week

```
['mon' 'Monday' 'tue' 'wed' 'thu' 'fri']
```

 Giá trị **'mon'** **'Monday'** cần thống nhất lại.

Dữ liệu cột duration


```
[ 261 149 226 ... 1246 1556 1868]
```

Dữ liệu cột campaign

```
[ 1 2 3 4 5 6 7 8 9 10 11 12 13 19 18 23 14 22 25 16 17 15 20 56
 39 35 42 28 26 27 32 21 24 29 31 30 41 37 40 33 34 43]
```

Dữ liệu cột pdays

```
[999. nan 6. 4. 3. 5. 1. 0. 10. 7. 8. 9. 11. 2.
 12. 13. 14. 15. 16. 21. 17. 18. 22. 25. 26. 19. 27. 20.]
```

 Có dữ liệu nan có nghĩa là không nhập liệu (dữ liệu trống). Cần phải xử lý. Trong trường hợp này là cần thay thế nan bằng 999.

Dữ liệu cột previous

```
[0 1 2 3 4 5 6 7]
```

Dữ liệu cột poutcome

```
['nonexistent' 'failure' 'success']
```

Dữ liệu cột emp.var.rate

```
[ 1.1 1.4 -0.1 -0.2 -1.8 -2.9 -3.4 -3. -1.7 -1.1]
```

Dữ liệu cột cons.price.idx

```
[93.994 94.465 93.918 93.444 93.798 93.2 92.756 92.843 93.075 92.893
 92.963 92.469 92.201 92.379 92.431 92.649 92.713 93.369 93.749 93.876
 94.055 94.215 94.027 94.199 94.601 94.767]
```

Dữ liệu cột cons.conf.idx

```
[-36.4 -41.8 -42.7 -36.1 -40.4 -42. -45.9 -50. -47.1 -46.2 -40.8 -33.6
 -31.4 -29.8 -26.9 -30.1 -33. -34.8 -34.6 -40. -39.8 -40.3 -38.3 -37.5
 -49.5 -50.8]
```

Dữ liệu cột euribor3m

```
[4.857 4.856 4.855 4.859 4.86 4.858 4.864 4.865 4.866 4.967 4.961 4.959
 4.958 4.96 4.962 4.955 4.947 4.956 4.966 4.963 4.957 4.968 4.97 4.965
 1.035 1.03 1.031 1.028]
```

Dữ liệu cột nr.employed

```
[5191.  5228.1 5195.8 5176.3 5099.1 5076.2 5017.5 5023.5 5008.7 4991.6
 4963.6]
```

Dữ liệu cột y

```
['no' 'yes']
```

Thống nhất giá trị dữ liệu sai do nhập tay

Cột education

```
df['education'].unique()
```

```
array(['basic.4y', 'high.school', 'basic.6y', 'basic.9y',
       'professional.course', 'unknown', 'high school', 'High School',
       'university.degree', 'illiterate'], dtype=object)
```

```
df['education'] = df['education'].replace('high school',
                                           'high.school')
```

```
df['education'] = df['education'].replace('High School',
                                           'high.school')
```

```
df['education'].unique()
```

```
array(['basic.4y', 'high.school', 'basic.6y', 'basic.9y',
       'professional.course', 'unknown', 'university.degree',
       'illiterate'], dtype=object)
```

Cột month

```
df['month'].unique()
```

```
array(['may', 'May', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'mar',
       'apr', 'sep'], dtype=object)
```

```
df['month'] = df['month'].replace('May', 'may')
```

```
df['month'].unique()
```

```
array(['may', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'mar', 'apr',
       'sep'], dtype=object)
```

Cột day_of_week

```
df['day_of_week'].unique()
```

```
array(['mon', 'Monday', 'tue', 'wed', 'thu', 'fri'], dtype=object)
```

```
df['day_of_week'] = df['day_of_week'].replace('Monday', 'mon')
```

```
df['day_of_week'].unique()
```

```
array(['mon', 'tue', 'wed', 'thu', 'fri'], dtype=object)
```

Cột pdays

```
df['pdays'].unique()
```

```
array([999., nan, 6., 4., 3., 5., 1., 0., 10., 7., 8.,  
       9., 11., 2., 12., 13., 14., 15., 16., 21., 17., 18.,  
       22., 25., 26., 19., 27., 20.])
```

Sử dụng hàm `.fillna(giá trị)` đối với cột dữ liệu:

```
df['pdays'] = df['pdays'].fillna(999)  
df['pdays'].unique()
```

```
array([999., 6., 4., 3., 5., 1., 0., 10., 7., 8., 9.,  
       11., 2., 12., 13., 14., 15., 16., 21., 17., 18., 22.,  
       25., 26., 19., 27., 20.])
```

Lưu dữ liệu

```
df.to_csv('bank-additional-full-processed.csv', index=False)
```

File kết quả này có lưu tại:

```
http://thachln.github.io/datasets/bank/bank-additional-full-processed.csv
```