assignment01

September 10, 2022

1

2

3 Change Kernel

conda install ipykernel python -m ipykernel install –user –name –display-name “py15130”

---

[1]:
```
import warnings
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
warnings.filterwarnings('ignore')
```

[2]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[3]:
```
data = pd.read_csv(
    'adult.csv')
data.head()
```

[3]:
```
   age         workclass  fnlwgt  education  education-num  \
0   39         State-gov   77516  Bachelors             13
1   50  Self-emp-not-inc   83311  Bachelors             13
2   38           Private  215646    HS-grad              9
3   53           Private  234721       11th              7
4   28           Private  338409  Bachelors             13

       marital-status         occupation   relationship   race     sex  \
0       Never-married       Adm-clerical  Not-in-family  White    Male
1  Married-civ-spouse    Exec-managerial        Husband  White    Male
2            Divorced  Handlers-cleaners  Not-in-family  White    Male
3  Married-civ-spouse  Handlers-cleaners        Husband  Black    Male
```

```
4  Married-civ-spouse     Prof-specialty           Wife  Black  Female

   capital-gain  capital-loss  hours-per-week native-country salary
0          2174             0              40  United-States  <=50K
1             0             0              13  United-States  <=50K
2             0             0              40  United-States  <=50K
3             0             0              40  United-States  <=50K
4             0             0              40           Cuba  <=50K
```

DataFrame salary

[4]:
```
data['sex'].value_counts()
```

[4]:
```
Male      21790
Female    10771
Name: sex, dtype: int64
```

[5]:
```
data[data.sex=='Female']['age'].mean()
```

[5]:
```
36.85823043357163
```

[6]:
```
len(data[data['native-country']=='Germany'])/len(data)
```

[6]:
```
0.004207487485028101
```

50K 50K

[7]:
```
data_over50 = data[data['salary']=='>50K']
data_under50 = data[data['salary']=='<=50K']
print('    50K  {}   {}'.format(data_over50['age'].mean(),data_over50['age'].
 ↪std()))
print('    50K  {}   {}'.format(data_under50['age'].mean(),data_under50['age'].
 ↪std()))
```

```
   50K  44.24984058155847  10.51902771985177
   50K  36.78373786407767  14.020088490824813
```

groupby describe

[8]:
```
data.groupby(['race','sex'])['age'].describe()
```

[8]:
```
                              count       mean        std   min   25%   50%  \
race               sex
Amer-Indian-Eskimo Female     119.0  37.117647  13.114991  17.0  27.0  36.0
```

```
                   Male      192.0  37.208333  12.049563  17.0  28.0  35.0
Asian-Pac-Islander Female    346.0  35.089595  12.300845  17.0  25.0  33.0
                   Male      693.0  39.073593  12.883944  18.0  29.0  37.0
Black              Female   1555.0  37.854019  12.637197  17.0  28.0  37.0
                   Male     1569.0  37.682600  12.882612  17.0  27.0  36.0
Other              Female    109.0  31.678899  11.631599  17.0  23.0  29.0
                   Male      162.0  34.654321  11.355531  17.0  26.0  32.0
White              Female   8642.0  36.811618  14.329093  17.0  25.0  35.0
                   Male    19174.0  39.652498  13.436029  17.0  29.0  38.0

                             75%   max
race               sex
Amer-Indian-Eskimo Female  46.00  80.0
                   Male    45.00  82.0
Asian-Pac-Islander Female  43.75  75.0
                   Male    46.00  90.0
Black              Female  46.00  90.0
                   Male    46.00  90.0
Other              Female  39.00  74.0
                   Male    42.00  77.0
White              Female  46.00  90.0
                   Male    49.00  90.0
```

[9]:

```python
data['marital-status'].value_counts()
```

[9]:

```
Married-civ-spouse       14976
Never-married            10683
Divorced                  4443
Separated                 1025
Widowed                    993
Married-spouse-absent      418
Married-AF-spouse           23
Name: marital-status, dtype: int64
```

[10]:

```python
len(data[(data['sex'] == 'Male')&
     (data['salary'] == '>50K') &
     data['marital-status'].str.startswith('Married')])
```

[10]:

```
5965
```

[11]:

```python
len(data[(data['salary'] == '>50K')&
     (data['sex'] == 'Male') &
     (data['marital-status'].isin(['Never-married','Separated', 'Divorced']))])
```

[11]:

```
658
```

50K

[12]:

```
Max_weekworkTime = data['hours-per-week'].max()
data_weekworkTime = data[data['hours-per-week'] == Max_weekworkTime]
ratio = len(data_weekworkTime[data_weekworkTime['salary'] == '>50K'])/
 ↪len(data_weekworkTime)
print('     {}    {}       50K   {}'.
 ↪format(Max_weekworkTime,len(data_weekworkTime),ratio))
```

```
    99    85       50K   0.29411764705882354
```

50K

[13]:

```
data.groupby(['native-country','salary'])['hours-per-week'].mean()
```

[13]:

```
native-country  salary
?               <=50K     40.164760
                >50K      45.547945
Cambodia        <=50K     41.416667
                >50K      40.000000
Canada          <=50K     37.914634
                            …
United-States   >50K      45.505369
Vietnam         <=50K     37.193548
                >50K      39.200000
Yugoslavia      <=50K     41.600000
                >50K      49.500000
Name: hours-per-week, Length: 82, dtype: float64
```

[14]:

```
from sklearn.model_selection import train_test_split

train_valid,test = train_test_split(data, test_size=0.2)
train,valid = train_test_split(data, test_size=0.25)
```

10   10

[15]:

```
from sklearn.model_selection import KFold

train_valid,test = train_test_split(data, test_size=0.2)
kf = KFold(n_splits = 10, shuffle=True, random_state=2022)
for train, valid in kf.split(train_valid):
    print('train:%s , valid: %s ' %(train,valid))
print('test:%s'%(test))
```

```
train:[    0     1     2 … 26045 26046 26047] , valid: [   14    19    40 …
26022 26024 26039]
train:[    0     1     2 … 26045 26046 26047] , valid: [   23    37    49 …
26042 26043 26044]
train:[    0     1     3 … 26045 26046 26047] , valid: [    2    22    29 …
26020 26023 26035]
```

```
train:[    0     1     2 … 26045 26046 26047] , valid: [    4    12    28 …
26034 26036 26038]
train:[    0     1     2 … 26045 26046 26047] , valid: [   17    25    27 …
25975 26013 26031]
train:[    0     1     2 … 26044 26046 26047] , valid: [    3     8    10 …
26033 26037 26045]
train:[    0     2     3 … 26044 26045 26047] , valid: [    1    15    34 …
26003 26011 26046]
train:[    0     1     2 … 26044 26045 26046] , valid: [    6     7    20 …
26015 26030 26047]
train:[    1     2     3 … 26045 26046 26047] , valid: [    0     9    13 …
25990 25996 26041]
train:[    0     1     2 … 26045 26046 26047] , valid: [    5    11    16 …
25967 26012 26029]
test:       age         workclass  fnlwgt     education  education-num  \
17988   54  Self-emp-not-inc  124865  Some-college             10
22812   45           Private  144579     Bachelors             13
17288   47           Private  145290       HS-grad              9
3157    21           Private  305874  Some-college             10
12941   18           Private  123856          11th              7
…    …               …     …            …             …
9118    23           Private  218782          10th              6
4485    27           Private  219371       HS-grad              9
11043   34           Private   32528     Assoc-voc             11
25607   55           Private  225365       HS-grad              9
15029   28         Local-gov  197932  Some-college             10

              marital-status         occupation    relationship   race  \
17988               Divorced              Sales   Not-in-family  White
22812     Married-civ-spouse     Prof-specialty         Husband  White
17288     Married-civ-spouse  Machine-op-inspct         Husband  White
3157      Married-civ-spouse      Other-service         Husband  White
12941          Never-married              Sales       Own-child  White
…                        …                  …              …    …
9118           Never-married  Handlers-cleaners  Other-relative  Other
4485   Married-spouse-absent       Adm-clerical       Unmarried  White
11043  Married-spouse-absent       Adm-clerical       Unmarried  White
25607                Widowed      Other-service       Unmarried  White
15029          Never-married       Adm-clerical       Own-child  White

           sex  capital-gain  capital-loss  hours-per-week native-country  \
17988   Female             0             0              35  United-States
22812     Male             0             0              40  United-States
17288     Male             0             0              45  United-States
3157      Male             0             0              40  United-States
12941   Female             0             0              49  United-States
…        …           …             …             …              …
9118      Male             0             0              40  United-States
```

```
4485   Female             0            0             40        Jamaica
11043  Female             0          974             40  United-States
25607  Female             0            0             30  United-States
15029  Female             0            0             16  United-States

      salary
17988  <=50K
22812   >50K
17288  <=50K
3157   <=50K
12941  <=50K
…      …
9118   <=50K
4485   <=50K
11043  <=50K
25607  <=50K
15029  <=50K

[6513 rows x 15 columns]
```