hp-04

July 3, 2023

[1]:
```
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

[17]:
```
features = ["Age", "Workclass", "fnlwgt", "Education", "Education-Num",␣
 ↪"Martial Status", "Occupation", "Relationship",
            "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week",␣
 ↪"Country", "Target"]

df = pd.read_csv('adult.data', names=features)
df
```

[17]:

	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	
…	…	…	…	…	…	
32556	27	Private	257302	Assoc-acdm	12	
32557	40	Private	154374	HS-grad	9	
32558	58	Private	151910	HS-grad	9	
32559	22	Private	201490	HS-grad	9	
32560	52	Self-emp-inc	287927	HS-grad	9	

	Martial Status	Occupation	Relationship	Race	\
0	Never-married	Adm-clerical	Not-in-family	White	
1	Married-civ-spouse	Exec-managerial	Husband	White	
2	Divorced	Handlers-cleaners	Not-in-family	White	
3	Married-civ-spouse	Handlers-cleaners	Husband	Black	
4	Married-civ-spouse	Prof-specialty	Wife	Black	
…	…	…	…	…	
32556	Married-civ-spouse	Tech-support	Wife	White	
32557	Married-civ-spouse	Machine-op-inspct	Husband	White	
32558	Widowed	Adm-clerical	Unmarried	White	
32559	Never-married	Adm-clerical	Own-child	White	

```
32560   Married-civ-spouse     Exec-managerial            Wife   White

           Sex  Capital Gain  Capital Loss  Hours per week         Country  \
0         Male          2174             0              40   United-States
1         Male             0             0              13   United-States
2         Male             0             0              40   United-States
3         Male             0             0              40   United-States
4       Female             0             0              40            Cuba
…         …            …             …              …             …
32556   Female             0             0              38   United-States
32557     Male             0             0              40   United-States
32558   Female             0             0              40   United-States
32559     Male             0             0              20   United-States
32560   Female         15024             0              40   United-States

       Target
0       <=50K
1       <=50K
2       <=50K
3       <=50K
4       <=50K
…       …
32556   <=50K
32557    >50K
32558   <=50K
32559   <=50K
32560    >50K

[32561 rows x 15 columns]
```

[3]:
```
df.isnull().sum()
```

[3]:
```
Age               0
Workclass         0
fnlwgt            0
Education         0
Education-Num     0
Martial Status    0
Occupation        0
Relationship      0
Race              0
Sex               0
Capital Gain      0
Capital Loss      0
Hours per week    0
Country           0
Target            0
```

```
dtype: int64
```

[4]:
```
df['Sex'].value_counts()
```

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

Male count - 21,790 and Female count - 10,771

[16]:
```
df['Sex'] == 'Female'
```

[16]:
```
0        False
1        False
2        False
3        False
4        False
         …
32556    False
32557    False
32558    False
32559    False
32560    False
Name: Sex, Length: 32561, dtype: bool
```

[6]:
```
Female_ = df[df['Sex'] == 'Female']
```

[22]:
```
Female_['Age'].mean()
```

[22]:
```
nan
```

[8]:
```
df.Country.value_counts(normalize = True)
```

[8]:
```
 United-States                 0.895857
 Mexico                        0.019748
 ?                             0.017905
 Philippines                   0.006081
 Germany                       0.004207
 Canada                        0.003716
 Puerto-Rico                   0.003501
 El-Salvador                   0.003255
 India                         0.003071
 Cuba                          0.002918
 England                       0.002764
 Jamaica                       0.002488
 South                         0.002457
 China                         0.002303
```

```
 Italy                          0.002242
 Dominican-Republic             0.002150
 Vietnam                        0.002058
 Guatemala                      0.001966
 Japan                          0.001904
 Poland                         0.001843
 Columbia                       0.001812
 Taiwan                         0.001566
 Haiti                          0.001351
 Iran                           0.001321
 Portugal                       0.001136
 Nicaragua                      0.001044
 Peru                           0.000952
 France                         0.000891
 Greece                         0.000891
 Ecuador                        0.000860
 Ireland                        0.000737
 Hong                           0.000614
 Cambodia                       0.000584
 Trinadad&Tobago                0.000584
 Laos                           0.000553
 Thailand                       0.000553
 Yugoslavia                     0.000491
 Outlying-US(Guam-USVI-etc)     0.000430
 Honduras                       0.000399
 Hungary                        0.000399
 Scotland                       0.000369
 Holand-Netherlands             0.000031
Name: Country, dtype: float64
```

The proportion of Germany country is 0.004207

[9]:
```
df['Target'].value_counts()
```

[9]:
```
 <=50K    24720
 >50K      7841
Name: Target, dtype: int64
```

[10]:
```
df['Target'] == '>50K'
```

[10]:
```
0        False
1        False
2        False
3        False
4        False
         …
32556    False
```

```
32557    False
32558    False
32559    False
32560    False
Name: Target, Length: 32561, dtype: bool
```

[14]:
```
df['Target'] == '<=50K'
```

[14]:
```
0        False
1        False
2        False
3        False
4        False
         …
32556    False
32557    False
32558    False
32559    False
32560    False
Name: Target, Length: 32561, dtype: bool
```

[12]:
```
df['Age'][df['Target'] == '>50K']
```

[12]:
```
Series([], Name: Age, dtype: int64)
```

[13]:
```
df['Age'][df['Target'] == '<=50K']
```

[13]:
```
Series([], Name: Age, dtype: int64)
```

[19]:
```
MaxSal = df[df['Target'] == '>50K']
MinSal = df[df['Target'] == '<=50K']
```

[21]:
```
MaxSal = df['Age'].mean()
```

[20]:
```
MaxSal = df['Age'].std()
```

[26]:
```
MinSal = df['Age'].mean()
```

[27]:
```
MinSal = df['Age'].std()
```

[32]:
```
df['Education'].value_counts()
```

[32]:
```
HS-grad         10501
Some-college     7291
Bachelors        5355
Masters          1723
Assoc-voc        1382
```

```
 11th             1175
 Assoc-acdm       1067
 10th              933
 7th-8th           646
 Prof-school       576
 9th               514
 12th              433
 Doctorate         413
 5th-6th           333
 1st-4th           168
 Preschool          51
Name: Education, dtype: int64
```

[ ]: