find the errors drop duplicates

March 12, 2024

[1]:
```
## import the all libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
## import the data set
data = pd.read_csv("C:/Users/Vikas/Desktop/adult.csv")
```

[3]:
```
data.head()
```

[3]:
```
   age  workclass  fnlwgt     education  educational-num      marital-status  \
0   25    Private  226802          11th                7       Never-married
1   38    Private   89814       HS-grad                9  Married-civ-spouse
2   28  Local-gov  336951    Assoc-acdm               12  Married-civ-spouse
3   44    Private  160323  Some-college               10  Married-civ-spouse
4   18          ?  103497  Some-college               10       Never-married

          occupation relationship   race  gender  capital-gain  capital-loss  \
0  Machine-op-inspct    Own-child  Black    Male             0             0
1    Farming-fishing      Husband  White    Male             0             0
2    Protective-serv      Husband  White    Male             0             0
3  Machine-op-inspct      Husband  Black    Male          7688             0
4                  ?    Own-child  White  Female             0             0

   hours-per-week native-country income
0              40  United-States  <=50K
1              50  United-States  <=50K
2              40  United-States   >50K
3              40  United-States   >50K
4              30  United-States  <=50K
```

[4]:
```
data.tail()
```

[4]:
```
       age     workclass  fnlwgt   education  educational-num  \
48837   27       Private  257302  Assoc-acdm               12
48838   40       Private  154374     HS-grad                9
48839   58       Private  151910     HS-grad                9
```

```
48840   22       Private  201490     HS-grad                9
48841   52  Self-emp-inc  287927     HS-grad                9

          marital-status         occupation relationship   race  gender  \
48837  Married-civ-spouse       Tech-support         Wife  White  Female
48838  Married-civ-spouse  Machine-op-inspct      Husband  White    Male
48839             Widowed       Adm-clerical    Unmarried  White  Female
48840       Never-married       Adm-clerical    Own-child  White    Male
48841  Married-civ-spouse    Exec-managerial         Wife  White  Female

       capital-gain  capital-loss  hours-per-week native-country income
48837             0             0              38  United-States  <=50K
48838             0             0              40  United-States   >50K
48839             0             0              40  United-States  <=50K
48840             0             0              20  United-States  <=50K
48841         15024             0              40  United-States   >50K
```

[5]:
```
data.columns
```

[5]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'educational-num',
       'marital-status', 'occupation', 'relationship', 'race', 'gender',
       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',
       'income'],
      dtype='object')
```

[7]:
```
data.describe()
```

[7]:
```
                age        fnlwgt  educational-num  capital-gain  \
count  48842.000000  4.884200e+04     48842.000000  48842.000000
mean      38.643585  1.896641e+05        10.078089   1079.067626
std       13.710510  1.056040e+05         2.570973   7452.019058
min       17.000000  1.228500e+04         1.000000      0.000000
25%       28.000000  1.175505e+05         9.000000      0.000000
50%       37.000000  1.781445e+05        10.000000      0.000000
75%       48.000000  2.376420e+05        12.000000      0.000000
max       90.000000  1.490400e+06        16.000000  99999.000000

       capital-loss  hours-per-week
count  48842.000000    48842.000000
mean      87.502314       40.422382
std      403.004552       12.391444
min        0.000000        1.000000
25%        0.000000       40.000000
50%        0.000000       40.000000
75%        0.000000       45.000000
max     4356.000000       99.000000
```

[8]:
```
data.info()
```

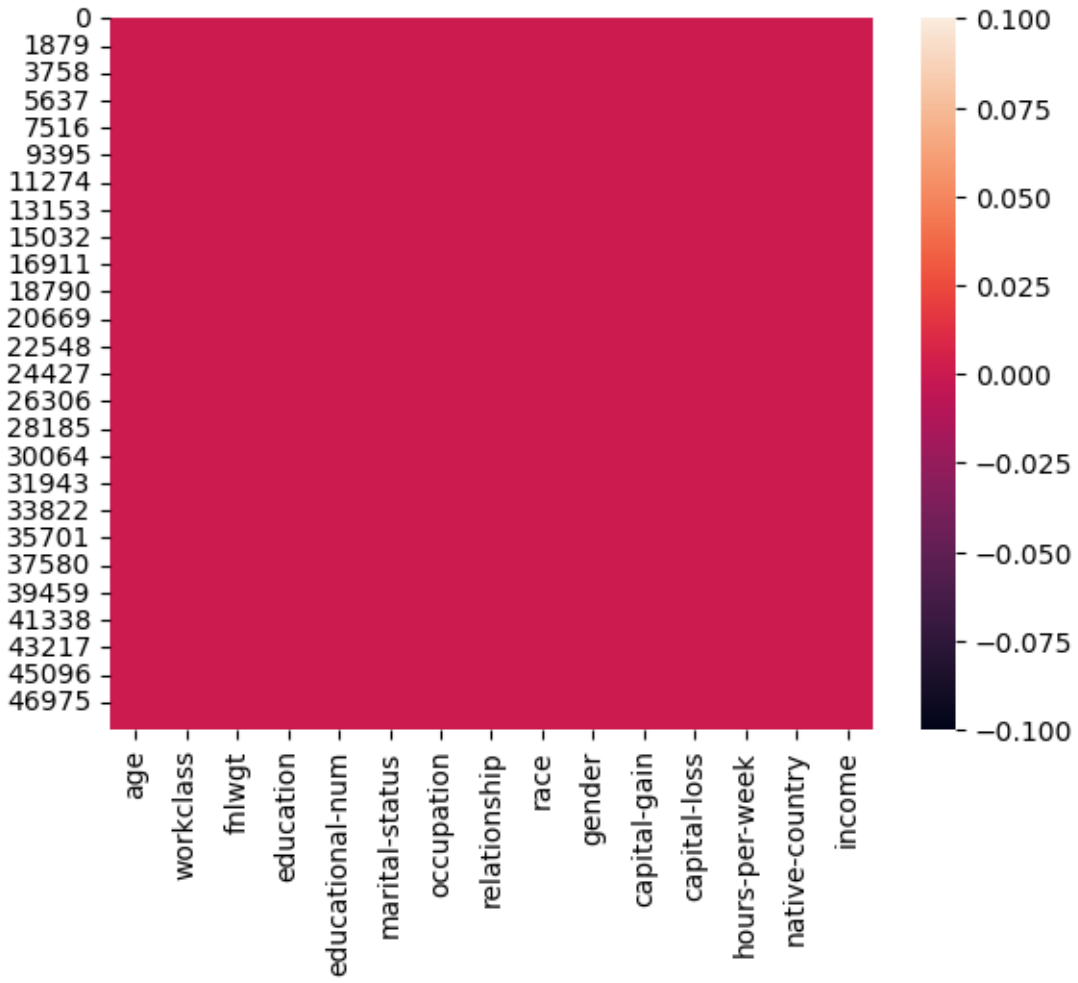
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              48842 non-null  int64
 1   workclass        48842 non-null  object
 2   fnlwgt           48842 non-null  int64
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64
 5   marital-status   48842 non-null  object
 6   occupation       48842 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64
 11  capital-loss     48842 non-null  int64
 12  hours-per-week   48842 non-null  int64
 13  native-country   48842 non-null  object
 14  income           48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

[9]:
```
data.isnull().sum()
```

[9]:
```
age               0
workclass         0
fnlwgt            0
education         0
educational-num   0
marital-status    0
occupation        0
relationship      0
race              0
gender            0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
income            0
dtype: int64
```

[10]:
```
sns.heatmap(data.isnull())
```

[10]:
```
<Axes: >
```



[13]:
```
data.tail(20)
```

[13]:

	age	workclass	fnlwgt	education	educational-num	\
48822	41	?	202822	HS-grad	9	
48823	72	?	129912	HS-grad	9	
48824	45	Local-gov	119199	Assoc-acdm	12	
48825	31	Private	199655	Masters	14	
48826	39	Local-gov	111499	Assoc-acdm	12	
48827	37	Private	198216	Assoc-acdm	12	
48828	43	Private	260761	HS-grad	9	
48829	65	Self-emp-not-inc	99359	Prof-school	15	
48830	43	State-gov	255835	Some-college	10	
48831	43	Self-emp-not-inc	27242	Some-college	10	
48832	32	Private	34066	10th	6	
48833	43	Private	84661	Assoc-voc	11	
48834	32	Private	116138	Masters	14	

```
48835   53           Private  321865       Masters              14
48836   22           Private  310152  Some-college              10
48837   27           Private  257302    Assoc-acdm              12
48838   40           Private  154374       HS-grad               9
48839   58           Private  151910       HS-grad               9
48840   22           Private  201490       HS-grad               9
48841   52      Self-emp-inc  287927       HS-grad               9

          marital-status         occupation    relationship  \
48822          Separated                  ?   Not-in-family
48823  Married-civ-spouse                 ?         Husband
48824           Divorced     Prof-specialty       Unmarried
48825           Divorced      Other-service   Not-in-family
48826  Married-civ-spouse       Adm-clerical            Wife
48827           Divorced       Tech-support   Not-in-family
48828  Married-civ-spouse  Machine-op-inspct         Husband
48829       Never-married     Prof-specialty   Not-in-family
48830           Divorced       Adm-clerical  Other-relative
48831  Married-civ-spouse       Craft-repair         Husband
48832  Married-civ-spouse  Handlers-cleaners         Husband
48833  Married-civ-spouse              Sales         Husband
48834       Never-married       Tech-support   Not-in-family
48835  Married-civ-spouse    Exec-managerial         Husband
48836       Never-married    Protective-serv   Not-in-family
48837  Married-civ-spouse       Tech-support            Wife
48838  Married-civ-spouse  Machine-op-inspct         Husband
48839             Widowed       Adm-clerical       Unmarried
48840       Never-married       Adm-clerical       Own-child
48841  Married-civ-spouse    Exec-managerial            Wife

                    race  gender  capital-gain  capital-loss  hours-per-week  \
48822               Black  Female             0             0              32
48823               White    Male             0             0              25
48824               White  Female             0             0              48
48825               Other  Female             0             0              30
48826               White  Female             0             0              20
48827               White  Female             0             0              40
48828               White    Male             0             0              40
48829               White    Male          1086             0              60
48830               White  Female             0             0              40
48831               White    Male             0             0              50
48832  Amer-Indian-Eskimo    Male             0             0              40
48833               White    Male             0             0              45
48834  Asian-Pac-Islander    Male             0             0              11
48835               White    Male             0             0              40
48836               White    Male             0             0              40
48837               White  Female             0             0              38
```

```
48838               White    Male             0             0              40
48839               White  Female             0             0              40
48840               White    Male             0             0              20
48841               White  Female         15024             0              40

      native-country income
48822  United-States  <=50K
48823  United-States  <=50K
48824  United-States  <=50K
48825  United-States  <=50K
48826  United-States   >50K
48827  United-States  <=50K
48828         Mexico  <=50K
48829  United-States  <=50K
48830  United-States  <=50K
48831  United-States  <=50K
48832  United-States  <=50K
48833  United-States  <=50K
48834         Taiwan  <=50K
48835  United-States   >50K
48836  United-States  <=50K
48837  United-States  <=50K
48838  United-States   >50K
48839  United-States  <=50K
48840  United-States  <=50K
48841  United-States   >50K
```

[12]:
```
## i want to check the null values
data.isin(['?']).sum()
```

[12]:
```
age                  0
workclass         2799
fnlwgt               0
education            0
educational-num      0
marital-status       0
occupation        2809
relationship         0
race                 0
gender               0
capital-gain         0
capital-loss         0
hours-per-week       0
native-country     857
income               0
dtype: int64
```

[14]:
```
data.columns
```

[14]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'educational-num',
       'marital-status', 'occupation', 'relationship', 'race', 'gender',
       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',
       'income'],
      dtype='object')
```

[17]:
```
data['workclass'] = data['workclass'].replace('?',np.nan)
data['occupation'] = data['occupation'].replace('?',np.nan)
data['native-country'] = data['native-country'].replace('?',np.nan)
```

[18]:
```
data.isin(['?']).sum()
```

[18]:
```
age               0
workclass         0
fnlwgt            0
education         0
educational-num   0
marital-status    0
occupation        0
relationship      0
race              0
gender            0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
income            0
dtype: int64
```

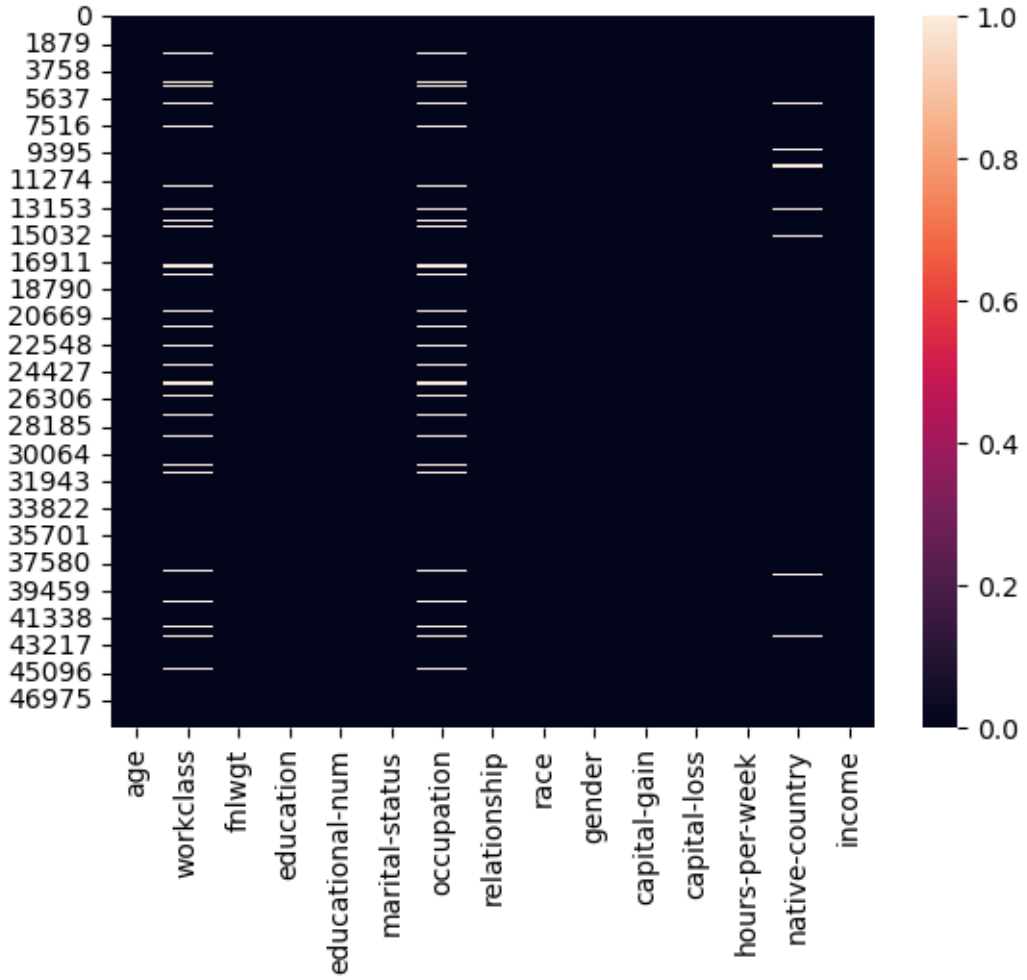
[19]:
```
data.isnull().sum()
```

[19]:
```
age                  0
workclass         2799
fnlwgt               0
education            0
educational-num      0
marital-status       0
occupation        2809
relationship         0
race                 0
gender               0
capital-gain         0
capital-loss         0
hours-per-week       0
native-country     857
```

```
income                0
dtype: int64
```

[20]:
```
sns.heatmap(data.isnull())
```

[20]: <Axes: >



[23]:
```
##  allthe missing values,%
per_mising = data.isnull().sum()*100/len(data)
per_mising
```

[23]:
```
age               0.000000
workclass         5.730724
fnlwgt            0.000000
education         0.000000
educational-num   0.000000
```

```
marital-status     0.000000
occupation         5.751198
relationship       0.000000
race               0.000000
gender             0.000000
capital-gain       0.000000
capital-loss       0.000000
hours-per-week     0.000000
native-country     1.754637
income             0.000000
dtype: float64
```

[26]:
```
## drop the missing rows,
data.dropna(how='any',inplace=True)
```

[27]:
```
data.shape
```

[27]:
```
(45222, 15)
```

[29]:
```
## let check duplicate data
dup = data.duplicated().any()
```

[30]:
```
print("are thre any duplicated values in data",dup)
```

```
are thre any duplicated values in data True
```

[31]:
```
data=data.drop_duplicates()
```

[33]:
```
data.shape
```

[33]:
```
(45175, 15)
```

[ ]:
```

```