adult-income-analysis

February 2, 2024

Import library

[3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading dataset

[4]:
```
data = pd.read_csv("E:\\PYTHON\\Adult Income Analysis\\Dataset\\adult.csv")
```

[5]:
```
data
```

[5]:

	age	workclass	fnlwgt	education	educational-num	\
0	25	Private	226802	11th	7	
1	38	Private	89814	HS-grad	9	
2	28	Local-gov	336951	Assoc-acdm	12	
3	44	Private	160323	Some-college	10	
4	18	?	103497	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	race	gender	\
0	Never-married	Machine-op-inspct	Own-child	Black	Male	
1	Married-civ-spouse	Farming-fishing	Husband	White	Male	
2	Married-civ-spouse	Protective-serv	Husband	White	Male	
3	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	
4	Never-married	?	Own-child	White	Female	
…	…	…	…	…	…	
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

```
0                0             0              40  United-States  <=50K
1                0             0              50  United-States  <=50K
2                0             0              40  United-States   >50K
3             7688             0              40  United-States   >50K
4                0             0              30  United-States  <=50K
…              …           …             …          …    …
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

[48842 rows x 15 columns]
```

Display top10 rows

[6]:
```
data.head(10)
```

[6]:
```
   age         workclass  fnlwgt     education  educational-num  \
0   25           Private  226802          11th                7
1   38           Private   89814       HS-grad                9
2   28         Local-gov  336951    Assoc-acdm               12
3   44           Private  160323  Some-college               10
4   18                 ?  103497  Some-college               10
5   34           Private  198693          10th                6
6   29                 ?  227026       HS-grad                9
7   63  Self-emp-not-inc  104626   Prof-school               15
8   24           Private  369667  Some-college               10
9   55           Private  104996       7th-8th                4

       marital-status         occupation   relationship   race  gender  \
0       Never-married  Machine-op-inspct      Own-child  Black    Male
1  Married-civ-spouse    Farming-fishing        Husband  White    Male
2  Married-civ-spouse    Protective-serv        Husband  White    Male
3  Married-civ-spouse  Machine-op-inspct        Husband  Black    Male
4       Never-married                  ?      Own-child  White  Female
5       Never-married      Other-service  Not-in-family  White    Male
6       Never-married                  ?      Unmarried  Black    Male
7  Married-civ-spouse     Prof-specialty        Husband  White    Male
8       Never-married      Other-service      Unmarried  White  Female
9  Married-civ-spouse       Craft-repair        Husband  White    Male

   capital-gain  capital-loss  hours-per-week native-country income
0             0             0              40  United-States  <=50K
1             0             0              50  United-States  <=50K
2             0             0              40  United-States   >50K
3          7688             0              40  United-States   >50K
```

```
4              0             0              30  United-States  <=50K
5              0             0              30  United-States  <=50K
6              0             0              40  United-States  <=50K
7           3103             0              32  United-States   >50K
8              0             0              40  United-States  <=50K
9              0             0              10  United-States  <=50K
```

Display last10 rows

[7]:
```
data.tail(10)
```

[7]:
```
       age     workclass  fnlwgt     education  educational-num  \
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  \
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  \
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
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
```

Find the shape of the dataset

[8]:
```
data.shape
```

[8]:
```
(48842, 15)
```

[9]:
```
print("Number of rows", data.shape[0])
print("Number of columns", data.shape[1])
```

```
Number of rows 48842
Number of columns 15
```

Getting Information About Our Dataset LikeTotal Number Rows, Total Number of Columns, Datatypes of Each Column And Memory Requirement

[10]:
```
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
```

Fetch random samples from dataset (50%)

[11]:
```
data1 = data.sample(frac=0.50, random_state=1)
data1
```

[11]:

	age	workclass	fnlwgt	education	educational-num	\
391	31	Private	224234	HS-grad	9	
1899	25	Private	149486	HS-grad	9	
24506	36	Self-emp-not-inc	343721	Doctorate	16	
32816	26	?	131777	Bachelors	13	
47892	30	Local-gov	44566	Bachelors	13	
…	…	…	…	…	…	
31987	56	Self-emp-not-inc	50791	Masters	14	
8518	38	Local-gov	51240	Bachelors	13	
1350	18	Private	70021	HS-grad	9	
23734	31	Private	185528	Some-college	10	
39491	30	Local-gov	327825	HS-grad	9	

	marital-status	occupation	relationship	race	gender	\
391	Never-married	Transport-moving	Own-child	Black	Male	
1899	Never-married	Machine-op-inspct	Unmarried	Black	Male	
24506	Never-married	Prof-specialty	Not-in-family	White	Male	
32816	Married-civ-spouse	?	Husband	White	Male	
47892	Married-civ-spouse	Prof-specialty	Husband	White	Male	
…	…	…	…	…	…	
31987	Divorced	Sales	Not-in-family	White	Male	
8518	Married-civ-spouse	Prof-specialty	Husband	White	Male	
1350	Never-married	Handlers-cleaners	Own-child	White	Male	
23734	Divorced	Sales	Own-child	White	Female	
39491	Divorced	Protective-serv	Own-child	White	Female	

	capital-gain	capital-loss	hours-per-week	native-country	income
391	0	0	40	United-States	<=50K
1899	0	0	40	United-States	<=50K
24506	0	0	30	?	>50K
32816	0	2002	40	United-States	<=50K
47892	0	0	40	United-States	<=50K
…	…	…	…	…	…
31987	0	1876	60	United-States	<=50K
8518	0	0	45	United-States	<=50K
1350	0	0	12	United-States	<=50K
23734	0	0	35	United-States	<=50K
39491	0	0	32	United-States	<=50K

```
[24421 rows x 15 columns]
```

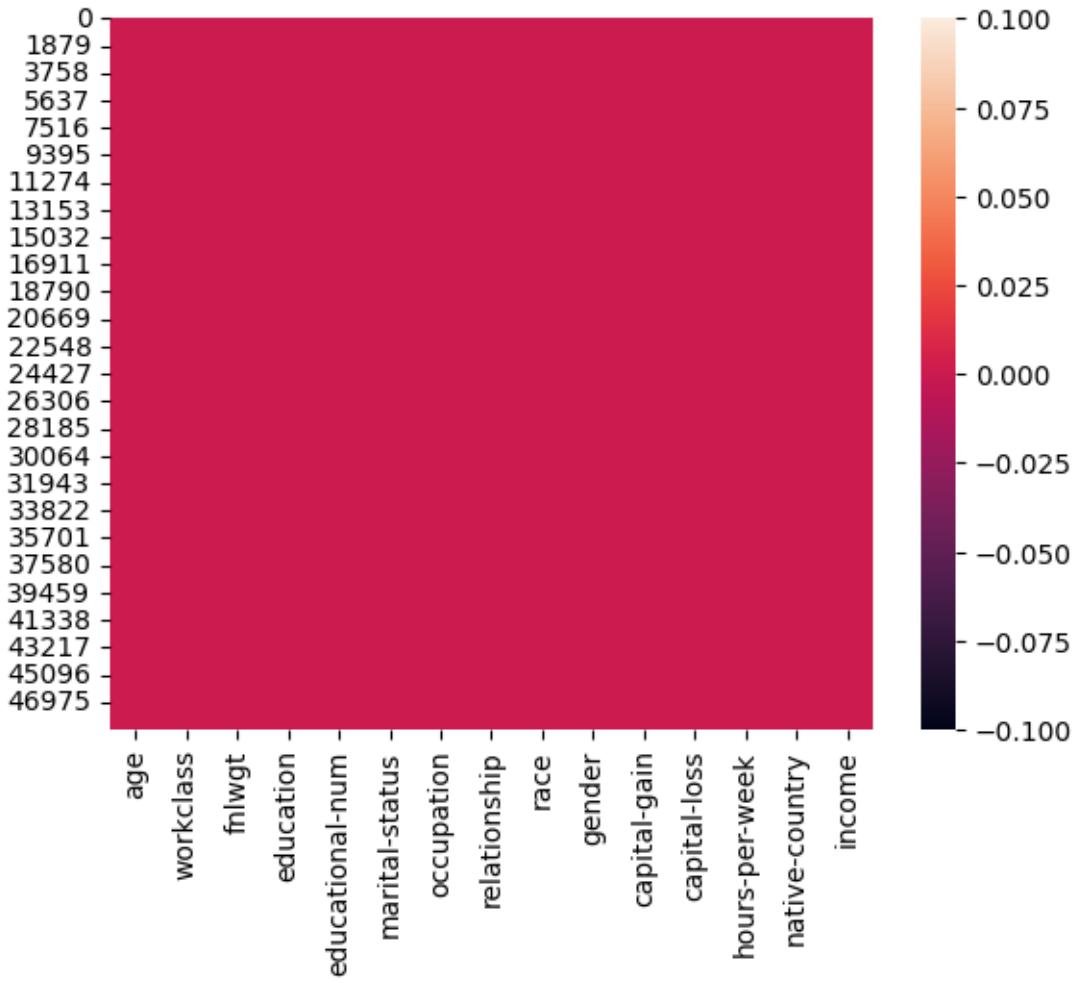
Check null values in dataset

[12]:
```
data.isna().sum(axis=0)
```

[12]:
```
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
```

[13]:
```
sns.heatmap(data.isnull())
```

[13]:
```
<Axes: >
```



####Replace "?"with NaN

[14]:
```
data.isin(["?"]).sum()
```

[14]:
```
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
```

[15]:
```
import numpy as np
```

[16]:
```
data.columns
```

[16]:
```
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
```

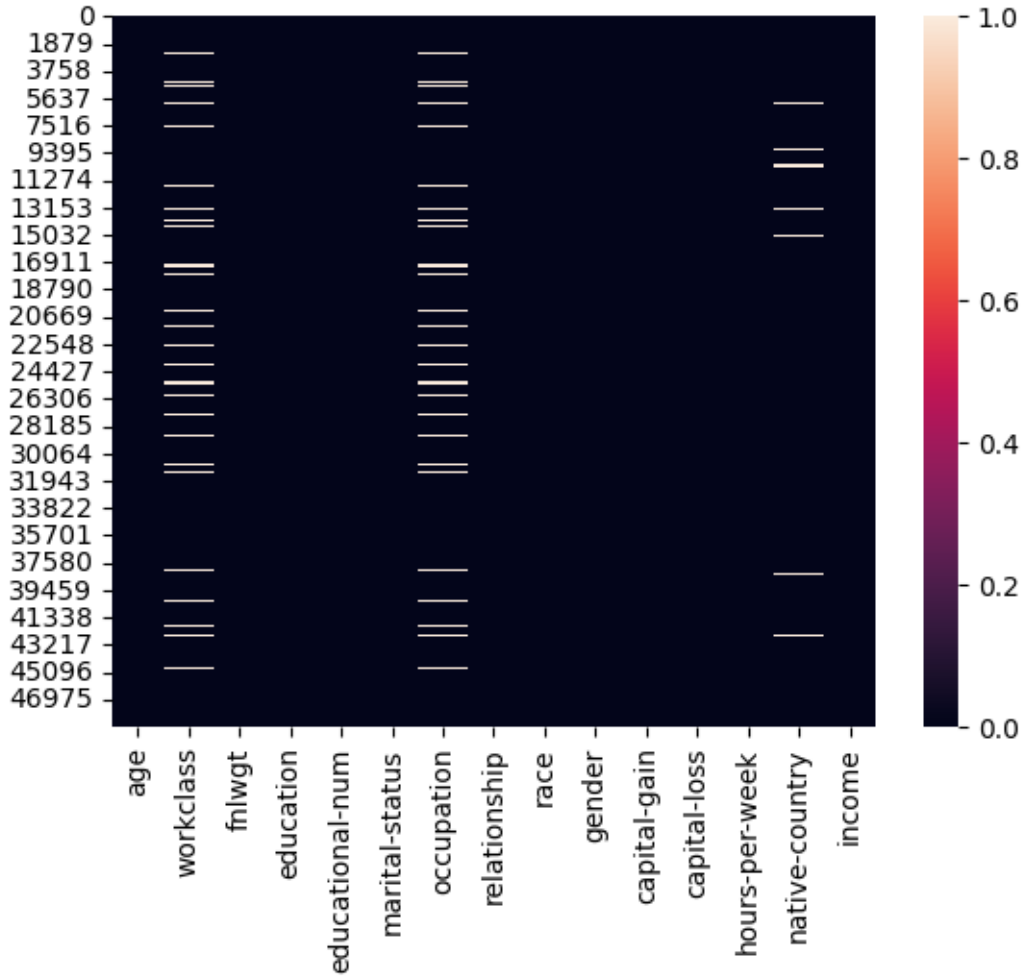
[19]:
```
data.isna().sum()
```

[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: >
```



Drop missing values

[21]:
```
per_missing = data.isnull().sum()*100/len(data)
```

[22]:
```
per_missing
```

[22]:
```
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
```

[23]:
```
data.dropna(how='any',inplace=True)
data.shape
```

[23]:
```
(45222, 15)
```

[24]:
```
48842 - 45222
```

[24]:
```
3620
```

Check for duplicates and drop them

[25]:
```
dup = data.duplicated().any()
print("Are there any duplicate values in data:", dup)
```

```
Are there any duplicate values in data: True
```

[26]:
```
data = data.drop_duplicates()
```

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

[27]:
```
(45175, 15)
```

[28]:
```
print("Duplicates drop:",45222 - 45175)
```

```
Duplicates drop: 47
```

Get overall statistics

[29]:
```
data.describe(include='all')
```

[29]:

	age	workclass	fnlwgt	education	educational-num	\
count	45175.000000	45175	4.517500e+04	45175	45175.000000	
unique	NaN	7	NaN	16	NaN	
top	NaN	Private	NaN	HS-grad	NaN	
freq	NaN	33262	NaN	14770	NaN	
mean	38.556170	NaN	1.897388e+05	NaN	10.119314	
std	13.215349	NaN	1.056524e+05	NaN	2.551740	
min	17.000000	NaN	1.349200e+04	NaN	1.000000	
25%	28.000000	NaN	1.173925e+05	NaN	9.000000	
50%	37.000000	NaN	1.783120e+05	NaN	10.000000	
75%	47.000000	NaN	2.379030e+05	NaN	13.000000	
max	90.000000	NaN	1.490400e+06	NaN	16.000000	

	marital-status	occupation	relationship	race	gender	\
count	45175	45175	45175	45175	45175	
unique	7	14	6	5	2	
top	Married-civ-spouse	Craft-repair	Husband	White	Male	
freq	21042	6010	18653	38859	30495	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	

	capital-gain	capital-loss	hours-per-week	native-country	income
count	45175.000000	45175.000000	45175.000000	45175	45175
unique	NaN	NaN	NaN	41	2
top	NaN	NaN	NaN	United-States	<=50K
freq	NaN	NaN	NaN	41256	33973
mean	1102.576270	88.687593	40.942512	NaN	NaN
std	7510.249876	405.156611	12.007730	NaN	NaN
min	0.000000	0.000000	1.000000	NaN	NaN
25%	0.000000	0.000000	40.000000	NaN	NaN
50%	0.000000	0.000000	40.000000	NaN	NaN
75%	0.000000	0.000000	45.000000	NaN	NaN
max	99999.000000	4356.000000	99.000000	NaN	NaN

[30]:
```
data['education'].unique()
```

[30]:
```
array(['11th', 'HS-grad', 'Assoc-acdm', 'Some-college', '10th',
       'Prof-school', '7th-8th', 'Bachelors', 'Masters', '5th-6th',
       'Assoc-voc', '9th', 'Doctorate', '12th', '1st-4th', 'Preschool'],
      dtype=object)
```

[31]:
```
data['educational-num'].unique()
```

[31]:
```
array([ 7,  9, 12, 10,  6, 15,  4, 13, 14,  3, 11,  5, 16,  8,  2,  1],
      dtype=int64)
```

Drop The Columns education-num, capital-gainandcapital-loss

[34]:
```
data.columns
```

[34]:
```
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')
```

[39]:
```
data = data.drop(['educational-num','capital-gain','capital-loss'],axis=1)
```

[40]:
```
data.columns
```

[40]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

0.0.1 Univariate Analysis

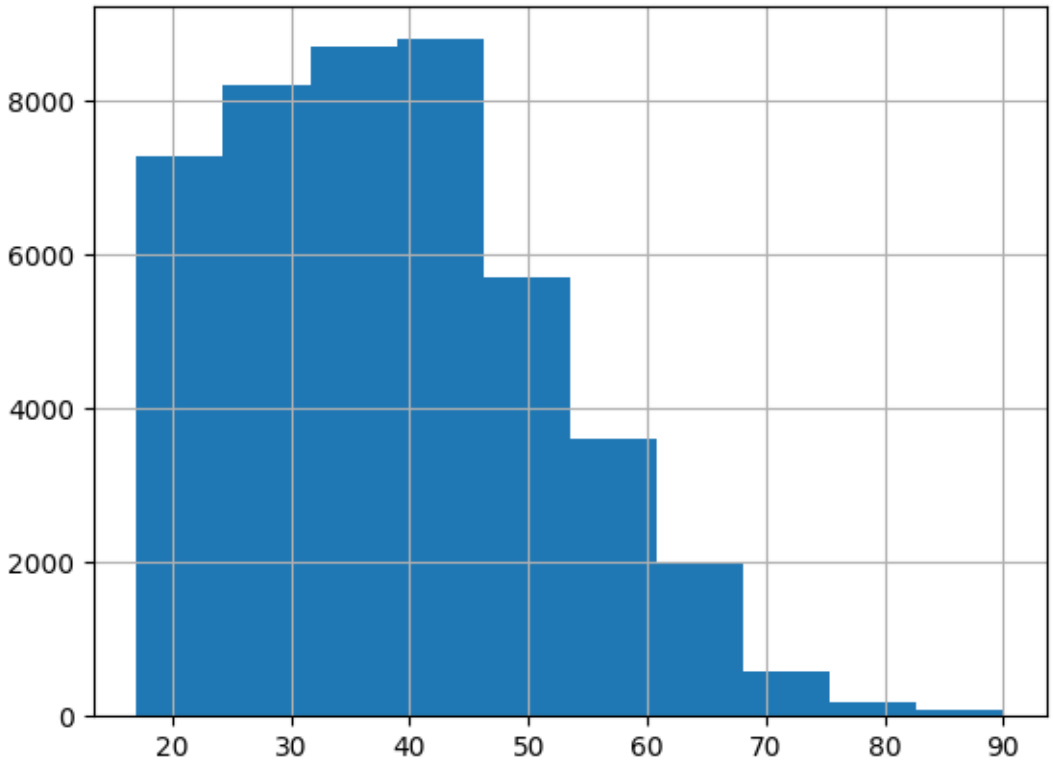
Distribution of age column

[43]:
```
data['age'].describe()
```

[43]:
```
count    45175.000000
mean        38.556170
std         13.215349
min         17.000000
25%         28.000000
50%         37.000000
75%         47.000000
max         90.000000
Name: age, dtype: float64
```

[44]:
```
data['age'].hist()
```

[44]:
```
<Axes: >
```



Find Total Number of Persons Having Age Between 17 To 48 (Inclusive) Using Between Method

[51]:
```
sum((data['age']>=17) & (data['age']<=48))
```

[51]:
```
34858
```

[48]:
```
sum(data['age'].between(17, 48))
```

[48]:
```
34858
```

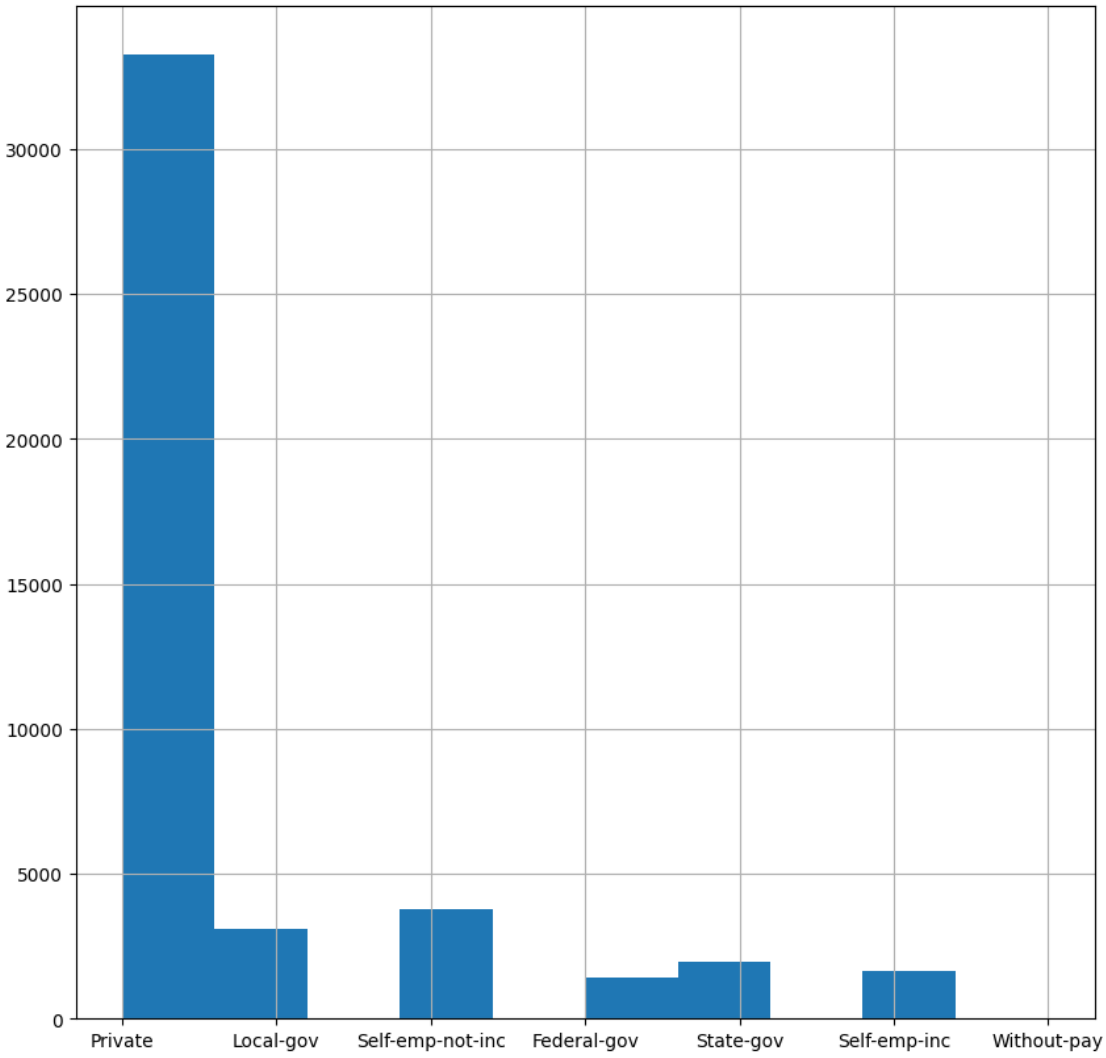
What is The Distribution of Workclass Column?

[52]:
```
data['workclass'].describe()
```

[52]:
```
count       45175
unique          7
top       Private
freq        33262
Name: workclass, dtype: object
```

[54]:
```
plt.figure(figsize=(10,10))
data['workclass'].hist()
```

[54]: <Axes: >



Persons have bachelors and master degrees

[62]:
```
filter1 = data['education']=='Bachelors'
filter2 = data['education']=='Masters'
len(data[filter1 | filter2])
```

[62]: 10072

[64]:
```
sum(data['education'].isin(['Bachelors','Masters']))
```

[64]: 10072

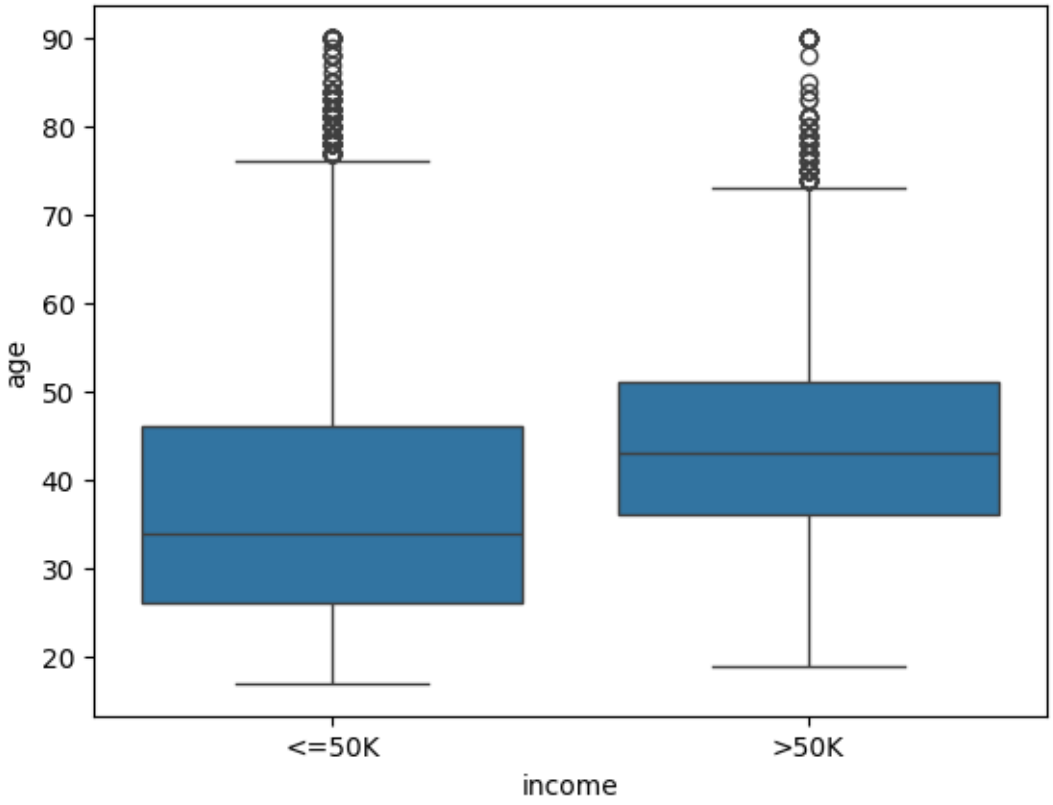
Bivariate analysis

[65]:
```
data.columns
```

[65]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

[67]:
```
sns.boxplot(x='income',y='age', data=data)
```

[67]:
```
<Axes: xlabel='income', ylabel='age'>
```



Replace Salary Values ['<=50K', '>50K'] With 0 and 1

[68]:
```
data['income'].unique()
```

[68]:
```
array(['<=50K', '>50K'], dtype=object)
```

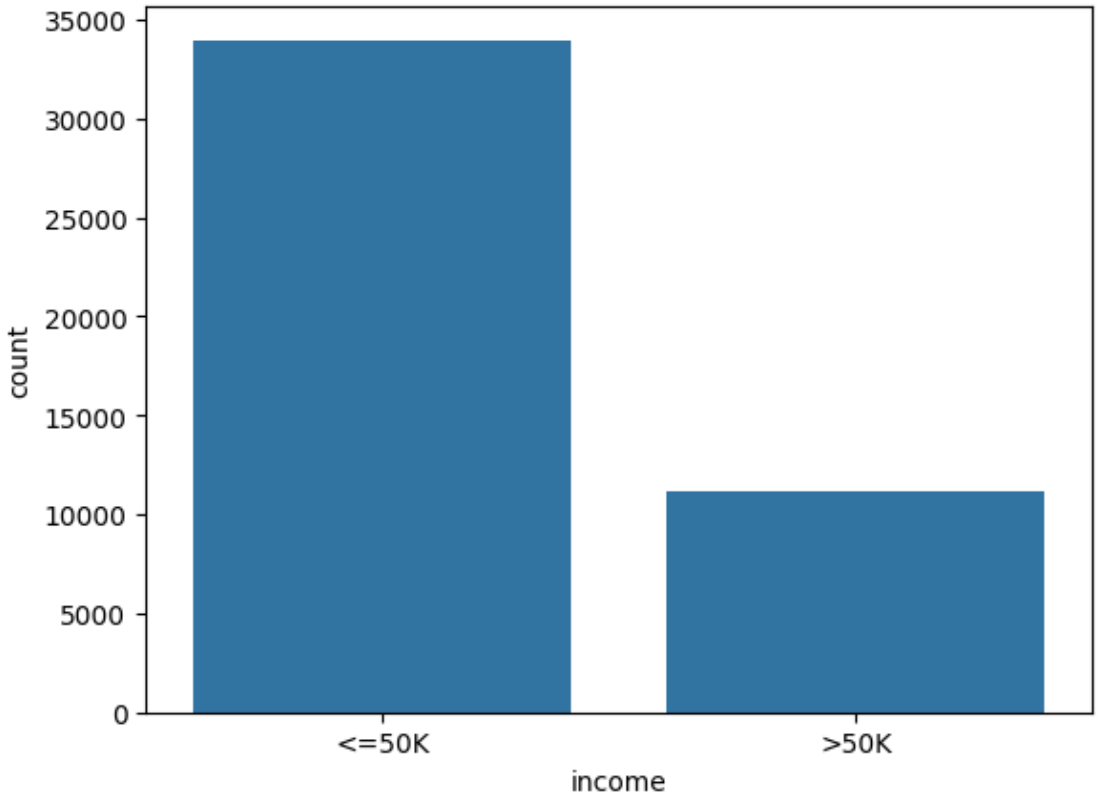
[69]:
```
data['income'].value_counts()
```

[69]:
```
income
<=50K    33973
```

```
>50K     11202
Name: count, dtype: int64
```

[70]:
```
sns.countplot(x='income',data=data)
```

[70]:
```
<Axes: xlabel='income', ylabel='count'>
```



[99]:
```
#Using Map method
income_mapping = {'<=50K': 0, '>50K': 1}
data['encoded_income'] = data['income'].map(income_mapping)
```

[95]:
```
#Using lambda method
data['encoded_income'] = data['income'].apply(lambda x: 0 if x == '<=50K' else␣
 ↪1)
```

[102]:
```
data.head(5)
```

[102]:
```
   age  workclass  fnlwgt     education      marital-status  \
0   25    Private  226802          11th       Never-married
1   38    Private   89814       HS-grad  Married-civ-spouse
2   28  Local-gov  336951    Assoc-acdm  Married-civ-spouse
```

```
3   44    Private  160323  Some-college  Married-civ-spouse
5   34    Private  198693          10th       Never-married

          occupation   relationship   race gender  hours-per-week  \
0  Machine-op-inspct      Own-child  Black   Male              40
1    Farming-fishing        Husband  White   Male              50
2    Protective-serv        Husband  White   Male              40
3  Machine-op-inspct        Husband  Black   Male              40
5      Other-service  Not-in-family  White   Male              30

  native-country income  encoded_income
0  United-States  <=50K               0
1  United-States  <=50K               0
2  United-States   >50K               1
3  United-States   >50K               1
5  United-States  <=50K               0
```

[103]:
```
data.replace(to_replace=['<=50K','>50K'], value=[0,1],inplace=True)
```

[105]:
```
data.head(5)
```

[105]:
```
   age  workclass  fnlwgt     education      marital-status  \
0   25    Private  226802          11th       Never-married
1   38    Private   89814       HS-grad  Married-civ-spouse
2   28  Local-gov  336951    Assoc-acdm  Married-civ-spouse
3   44    Private  160323  Some-college  Married-civ-spouse
5   34    Private  198693          10th       Never-married

          occupation   relationship   race gender  hours-per-week  \
0  Machine-op-inspct      Own-child  Black   Male              40
1    Farming-fishing        Husband  White   Male              50
2    Protective-serv        Husband  White   Male              40
3  Machine-op-inspct        Husband  Black   Male              40
5      Other-service  Not-in-family  White   Male              30

  native-country  income  encoded_income
0  United-States       0               0
1  United-States       0               0
2  United-States       1               1
3  United-States       1               1
5  United-States       0               0
```

Which Workclass Getting The Highest Salary?

[109]:
```
data.groupby('workclass')['income'].mean().sort_values(ascending=False)
```

[109]:
```
workclass
Self-emp-inc        0.554407
Federal-gov         0.390469
Local-gov           0.295161
Self-emp-not-inc    0.279051
State-gov           0.267215
Private             0.217816
Without-pay         0.095238
Name: income, dtype: float64
```

How Has Better Chance To Get Salary >50K Male or Female ?

[111]:
```
data.groupby('gender')['income'].mean().sort_values(ascending=False)
```

[111]:
```
gender
Male      0.312609
Female    0.113692
Name: income, dtype: float64
```

Covert workclass Columns Datatype To Category Datatype

[112]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 45175 entries, 0 to 48841
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             45175 non-null  int64
 1   workclass       45175 non-null  object
 2   fnlwgt          45175 non-null  int64
 3   education       45175 non-null  object
 4   marital-status  45175 non-null  object
 5   occupation      45175 non-null  object
 6   relationship    45175 non-null  object
 7   race            45175 non-null  object
 8   gender          45175 non-null  object
 9   hours-per-week  45175 non-null  int64
 10  native-country  45175 non-null  object
 11  income          45175 non-null  int64
 12  encoded_income  45175 non-null  int64
dtypes: int64(5), object(8)
memory usage: 5.8+ MB
```

[113]:
```
data['workclass'] = data['workclass'].astype('category')
```

[114]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 45175 entries, 0 to 48841
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             45175 non-null  int64
 1   workclass       45175 non-null  category
 2   fnlwgt          45175 non-null  int64
 3   education       45175 non-null  object
 4   marital-status  45175 non-null  object
 5   occupation      45175 non-null  object
 6   relationship    45175 non-null  object
 7   race            45175 non-null  object
 8   gender          45175 non-null  object
 9   hours-per-week  45175 non-null  int64
 10  native-country  45175 non-null  object
 11  income          45175 non-null  int64
 12  encoded_income  45175 non-null  int64
dtypes: category(1), int64(5), object(7)
memory usage: 5.5+ MB
```