In [2]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

data= pd.read_csv("adult.csv")
data
```

Out[2]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

48842 rows × 15 columns

DISPLAY TOP 10 ROWS OF THE DATASET

In [3]:
```
data.head(10)
```

Out[3]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K
5	34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0	30	United-States	<=50K
6	29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	Male	0	0	40	United-States	<=50K
7	63	Self-emp-not-inc	104626	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	3103	0	32	United-States	>50K
8	24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	Female	0	0	40	United-States	<=50K
9	55	Private	104996	7th-8th	4	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	10	United-States	<=50K

CHECK LAST 10 ROWS OF THE DATASET

In [4]:
```
data.tail(10)
```

Out[4]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48832	32	Private	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0	0	40	United-States	<=50K
48833	43	Private	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	<=50K
48834	32	Private	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0	0	11	Taiwan	<=50K
48835	53	Private	321865	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>50K
48836	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0	40	United-States	<=50K
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

FIND SHAPE OF OUR DATASET(NUMBER OF ROWS AND NUMBER OF COLUMNS)

In [5]:
```
data.shape
```

Out[5]:
```
(48842, 15)
```

In [6]:
```
print("Number of Rows", data.shape[0])
print("Number of Columns", data.shape[1])
```

```
Number of Rows 48842
Number of Columns 15
```

GETTING INFORMATION NUMBER OF ROWS, TOTAL NUMBER OF COLUMNS, DATATYPES OF EACH COLUMN AND MEMORY REQUIREMENT

In [7]:
```
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
```

```
 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 SAMPLE FROM THE DATASET(50%)

In [8]:
```
data.sample(frac=0.50, random_state=111)
```

Out[8]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
31652	54	Local-gov	172991	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	40	United-States	<=50K
20931	31	Private	73514	HS-grad	9	Never-married	Sales	Own-child	Asian-Pac-Islander	Female	0	0	40	United-States	<=50K
38653	31	Private	416415	HS-grad	9	Separated	Adm-clerical	Not-in-family	White	Male	0	0	45	United-States	<=50K
32939	22	Private	203182	Bachelors	13	Never-married	Exec-managerial	Unmarried	White	Female	0	0	30	United-States	<=50K
17673	59	Private	226922	HS-grad	9	Divorced	Sales	Unmarried	White	Female	0	1762	30	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
47522	60	Self-emp-not-inc	33717	11th	7	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
38727	45	Private	88061	11th	7	Married-spouse-absent	Machine-op-inspct	Unmarried	Asian-Pac-Islander	Female	0	0	40	South	<=50K
26925	43	Private	174325	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Black	Male	0	0	40	United-States	<=50K
19894	24	Local-gov	117109	Bachelors	13	Never-married	Adm-clerical	Own-child	Black	Female	0	0	27	United-States	<=50K
43332	20	?	183083	Some-college	10	Never-married	?	Own-child	White	Female	0	0	20	United-States	<=50K

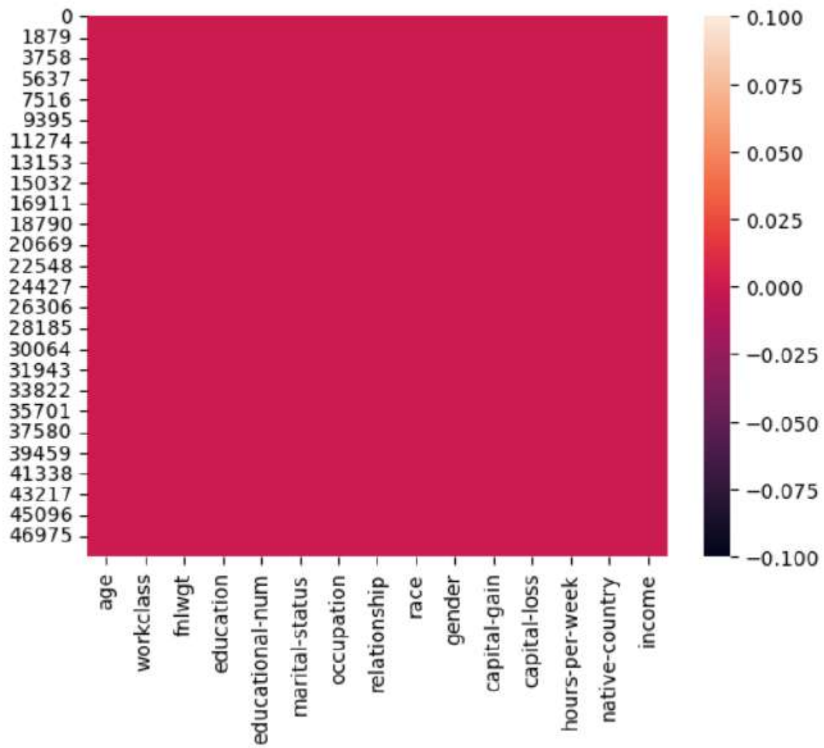
CHECK NULL VALUES IN THE DATASET

In [9]:
```
data.isnull().sum(axis=0)
```

Out[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
```

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

Out[10]: <Axes: >



PERFORM DATA CLEANING[REPLACE "?" WITH NaN]

In [11]:
```
data.tail(20)
```

Out[11]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
48822	41	?	202822	HS-grad	9	Separated	?	Not-in-family	Black	Female	0	0	32	United-States	<=50K
48823	72	?	129912	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	0	0	25	United-States	<=50K
48824	45	Local-gov	119199	Assoc-acdm	12	Divorced	Prof-specialty	Unmarried	White	Female	0	0	48	United-States	<=50K
48825	31	Private	199655	Masters	14	Divorced	Other-service	Not-in-family	Other	Female	0	0	30	United-States	<=50K
48826	39	Local-gov	111499	Assoc-acdm	12	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	20	United-States	>50K
48827	37	Private	198216	Assoc-acdm	12	Divorced	Tech-support	Not-in-family	White	Female	0	0	40	United-States	<=50K
48828	43	Private	260761	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	Mexico	<=50K
48829	65	Self-emp-not-inc	99359	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	1086	0	60	United-States	<=50K
48830	43	State-gov	255835	Some-college	10	Divorced	Adm-clerical	Other-relative	White	Female	0	0	40	United-States	<=50K
48831	43	Self-emp-not-inc	27242	Some-college	10	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	50	United-States	<=50K
48832	32	Private	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0	0	40	United-States	<=50K
48833	43	Private	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	<=50K
48834	32	Private	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0	0	11	Taiwan	<=50K
48835	53	Private	321865	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>50K
48836	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0	40	United-States	<=50K
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

In [12]:
```
data.isin(["?"]).sum()
```
Out[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
```
In [13]:
```
data.columns
```
Out[13]:
```
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')
```

```
      dtype='object')
```

In [14]:
```
data["workclass"]=data["workclass"].replace("?", np.NaN)
data["occupation"]=data["occupation"].replace("?", np.NaN)
data["native-country"]=data["native-country"].replace("?", np.NaN)
```

In [15]:
```
data.isin(["?"]).sum()
```

Out[15]:
```
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
```

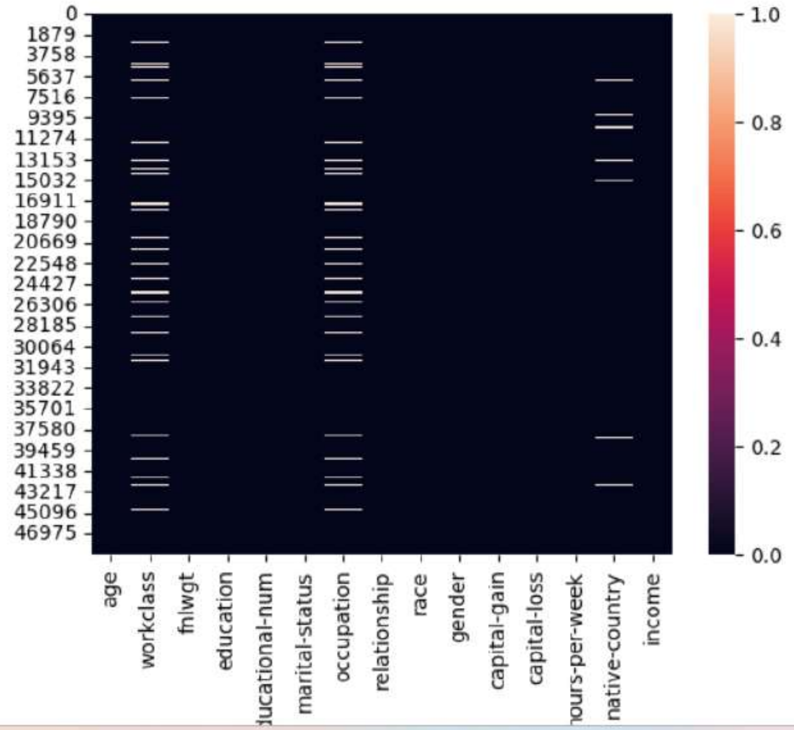
In [16]:
```
data.isnull().sum()
```

Out[16]:
```
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
```

```
gender                0
capital-gain          0
capital-loss          0
hours-per-week        0
native-country      857
income                0
dtype: int64
```

In [17]:
```
sns.heatmap(data.isnull())
```

Out[17]: <Axes: >



DROP ALL THE MISSING VALUES

In [18]:
```
per_missing= data.isnull().sum()*100/len(data)
```

In [19]:
```
per_missing
```

Out[19]:
```
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
```

In [20]:
```
data.dropna(how="any", inplace=True)
```

In [21]:
```
data.shape
```

Out[21]:
```
(45222, 15)
```

CHECK FOR DUPLICATE DATA AND DROP THEN

In [22]:
```
data= data.drop_duplicates()
```

In [23]:
```
data.shape
```

Out[23]:
```
(45175, 15)
```

GET OVERALL STATISTICS ABOUT THE DATAFRAME

In [24]:
```
data.describe(include="all")
```

Out[24]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
count	45175.000000	45175	4.517500e+04	45175	45175.000000	45175	45175	45175	45175	45175	45175.000000	45175.000000	45175.000000	45175	45175
unique	NaN	7	NaN	16	NaN	7	14	6	5	2	NaN	NaN	NaN	41	2
top	NaN	Private	NaN	HS-grad	NaN	Married-civ-spouse	Craft-repair	Husband	White	Male	NaN	NaN	NaN	United-States	<=50K
freq	NaN	33262	NaN	14770	NaN	21042	6010	18653	38859	30495	NaN	NaN	NaN	41256	33973
mean	38.556170	NaN	1.897388e+05	NaN	10.119314	NaN	NaN	NaN	NaN	NaN	1102.576270	88.687593	40.942512	NaN	NaN
std	13.215349	NaN	1.056524e+05	NaN	2.551740	NaN	NaN	NaN	NaN	NaN	7510.249876	405.156611	12.007730	NaN	NaN
min	17.000000	NaN	1.349200e+04	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	1.000000	NaN	NaN
25%	28.000000	NaN	1.173925e+05	NaN	9.000000	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	40.000000	NaN	NaN
50%	37.000000	NaN	1.783120e+05	NaN	10.000000	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	40.000000	NaN	NaN
75%	47.000000	NaN	2.379030e+05	NaN	13.000000	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	45.000000	NaN	NaN
max	90.000000	NaN	1.490400e+06	NaN	16.000000	NaN	NaN	NaN	NaN	NaN	99999.000000	4356.000000	99.000000	NaN	NaN

In [25]:
```
data.columns
```
Out[25]:
```
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')
```
In [26]:
```
data["education"].unique()
```
Out[26]:
```
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)
```
In [27]:
```
data["educational-num"].unique()
```
Out[27]:
```
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-GAIN, AND CAPITAL-LOSS

In [28]:
```
data.columns
```
Out[28]:
```
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')
```
In [29]:
```
data= data.drop(['educational-num','capital-gain','capital-loss'], axis=1)
```

In [29]:
```python
data= data.drop(['educational-num','capital-gain','capital-loss'], axis=1)
```

In [30]:
```python
data.columns
```

Out[30]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

UNIVARIATE ANALYSIS

WHAT IS THE DISTRIBUTION OF AGE COLUMN?

In [31]:
```python
data.columns
```

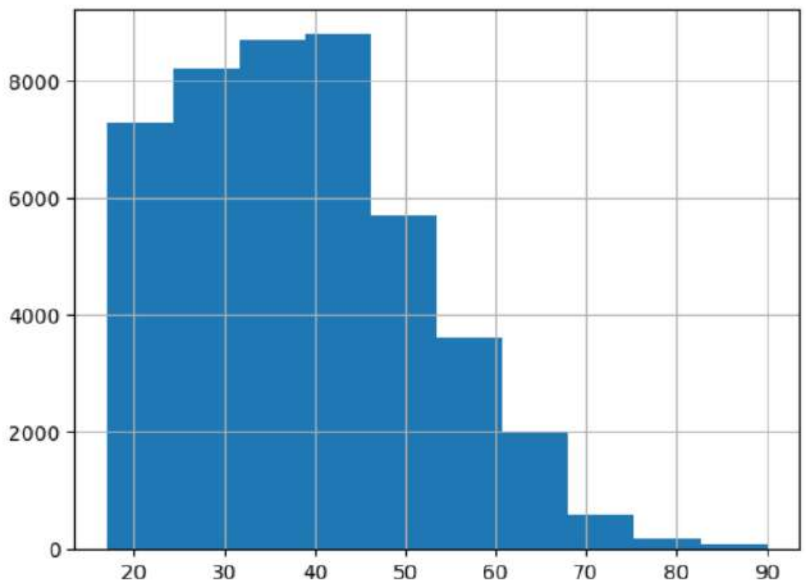
Out[31]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

In [32]:
```python
data["age"].describe()
```

Out[32]:
```
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
```

```
In [33]: data["age"].hist()
```

Out[33]: <Axes: >



FIND TOTAL NUMBER OF PERSONS HAVING AGE BETWEEN 17 TO 48(INCLUSIVE) USING BETWEEN METHOD

In [34]:
```
sum((data["age"]>=17) & (data["age"]<=48))
```

Out[34]:
```
34858
```

In [35]:
```
#OR
sum(data["age"].between(17,48))
```

Out[35]:
```
34858
```

WHAT IS THE DISTRIBUTION OF WORKCLASS COLUMN?

In [36]:
```
data.columns
```

Out[36]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

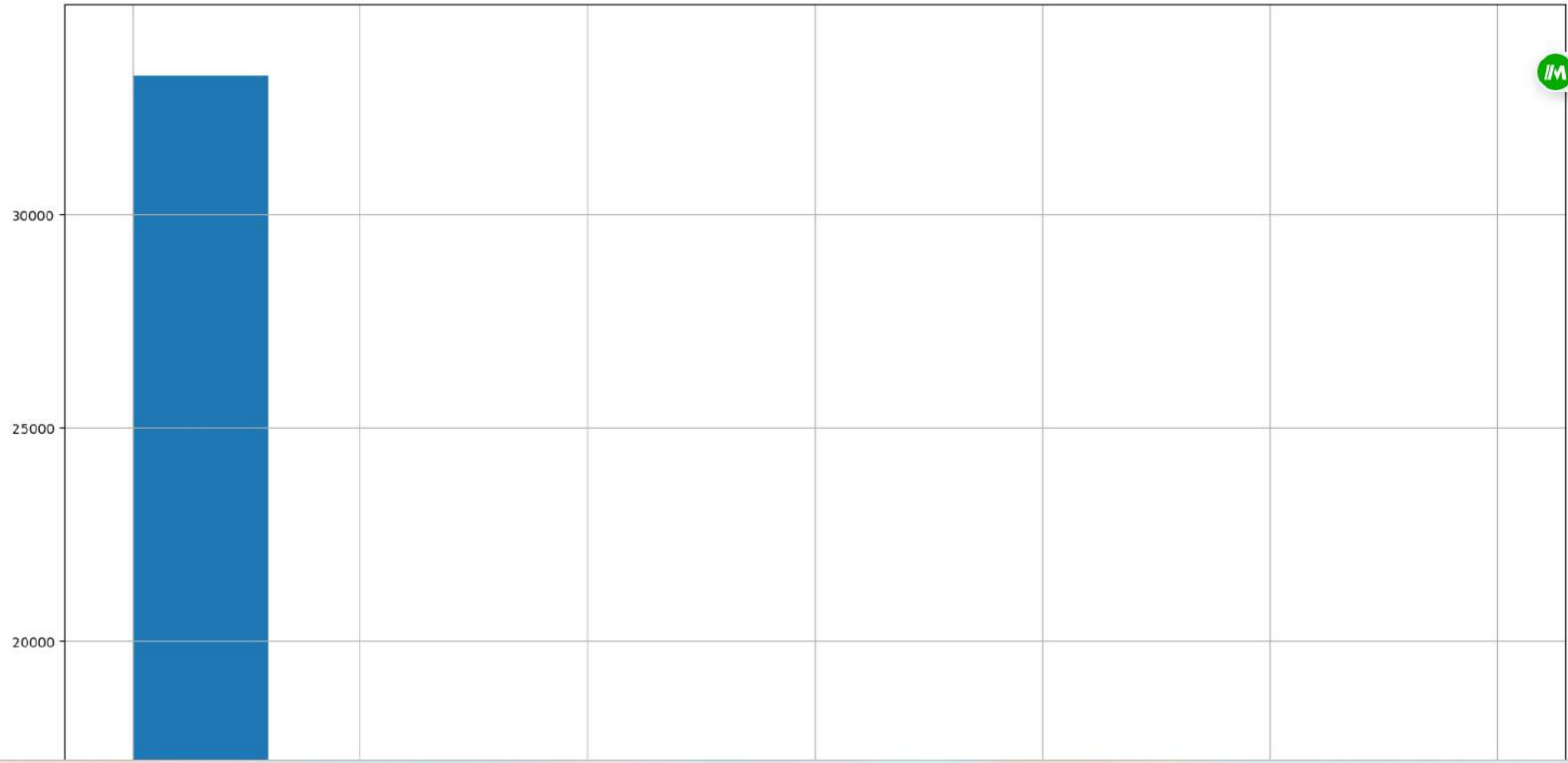
In [37]:
```
data["workclass"].describe()
```

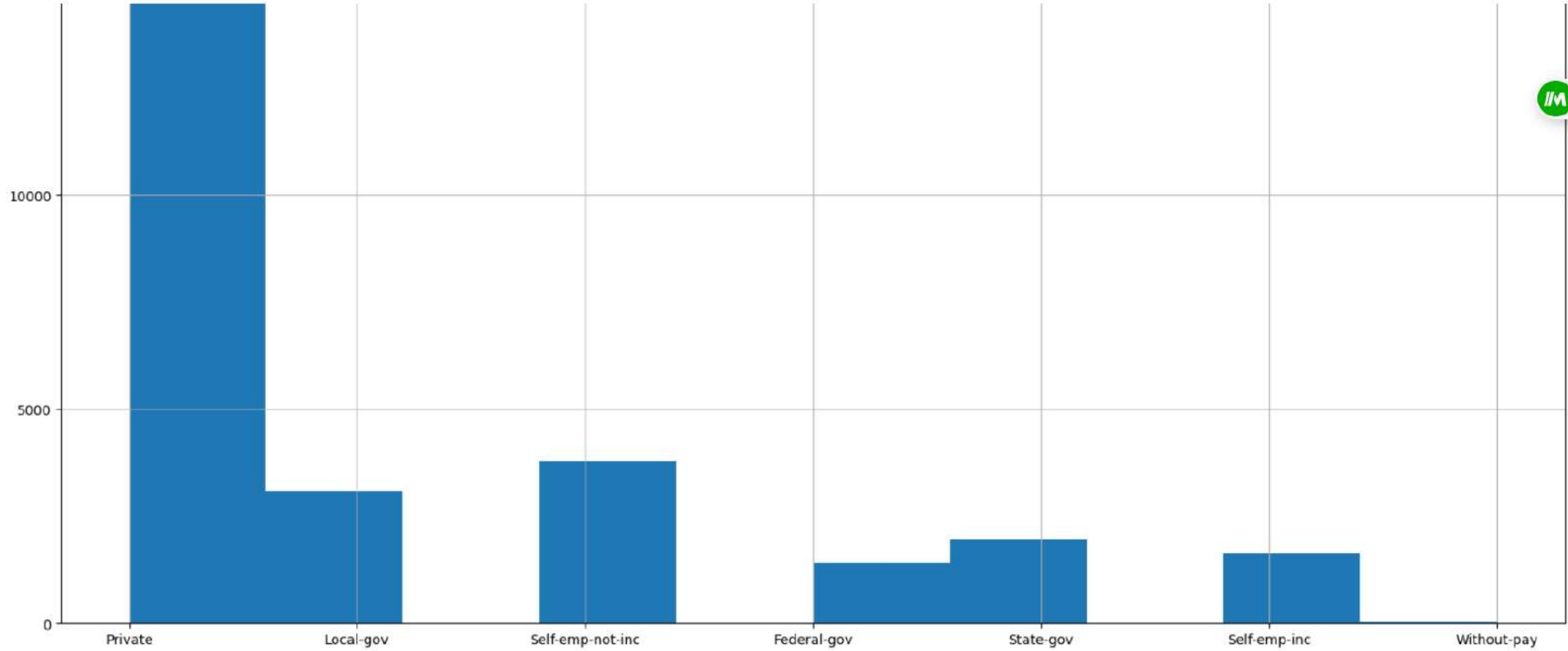
Out[37]:
```
count       45175
unique          7
top       Private
freq        33262
Name: workclass, dtype: object
```

In [44]:
```
plt.figure(figsize=(20,20))

data["workclass"].hist()
```

Out[44]:




10000
5000
0
Private
Local-gov
Self-emp-not-inc
Federal-gov
State-gov
Self-emp-inc
Without-pay

HOW MANY PERSONS HAVING BACHELORS AND MASTERS DEGREE

In [45]:
```
data.columns
```

Out[45]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

In [47]:
```
data["education"]
```

Out[47]:
```
0                 11th
1              HS-grad
2           Assoc-acdm
3         Some-college
5                 10th
             ...
48837       Assoc-acdm
48838          HS-grad
48839          HS-grad
48840          HS-grad
48841          HS-grad
Name: education, Length: 45175, dtype: object
```

In [53]:
```
len(data[filter1 | filter2])
```

Out[53]:
```
7559
```

In [54]:
```
#or
sum(data["education"].isin(["Bachelors", "MASTERS"]))
```

Out[54]:
```
7559
```

BIVARIATE ANALYSIS