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

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

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

48842 rows × 15 columns

Next steps: Generate code with data | View recommended plots | New interactive sheet

1. Display Top 10 Rows of The Dataset

```
data.head(10)
```

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

Next steps: Generate code with data | View recommended plots | New interactive sheet

2. Check Last 10 Rows of The Dataset

```
data.tail(10)
```

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

3. Find Shape of Our Dataset (Number of Rows And Number of Columns)

```
data.shape
```

```
(48842, 15)
```

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

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

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

```
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
```

5. Fetch Random Sample From the Dataset (50%)

```
data1=data.sample(frac=0.50,random_state=100)
data1
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	na co
12393	37	Private	110331	Prof-school	15	Married-civ-spouse	Other-service	Wife	White	Female	0	0	60	U S
48701	23	Private	45834	Bachelors	13	Never-married	Exec-managerial	Not-in-family	White	Female	0	0	50	U S
17918	28	Private	89718	HS-grad	9	Never-married	Sales	Not-in-family	White	Female	2202	0	48	U S
11352	30	Private	351770	9th	5	Divorced	Other-service	Unmarried	White	Female	0	0	38	U S
36198	31	Private	164190	10th	6	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	40	U S
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
48573	41	Private	318046	Some-college	10	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	48	U S
47252	41	Local-gov	33658	Some-college	10	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	45	U S
33142	69	Private	312653	Some-college	10	Married-civ-spouse	Sales	Husband	White	Male	0	0	25	U S
2965	21	?	334593	Some-college	10	Never-married	?	Not-in-family	White	Male	0	0	40	U S
32089	34	Private	186269	HS-grad	9	Divorced	Adm-clerical	Own-child	White	Male	0	0	40	U S

24421 rows × 15 columns

Next steps: Generate code with data1 | View recommended plots | New interactive sheet

6. Check Null Values In The Dataset

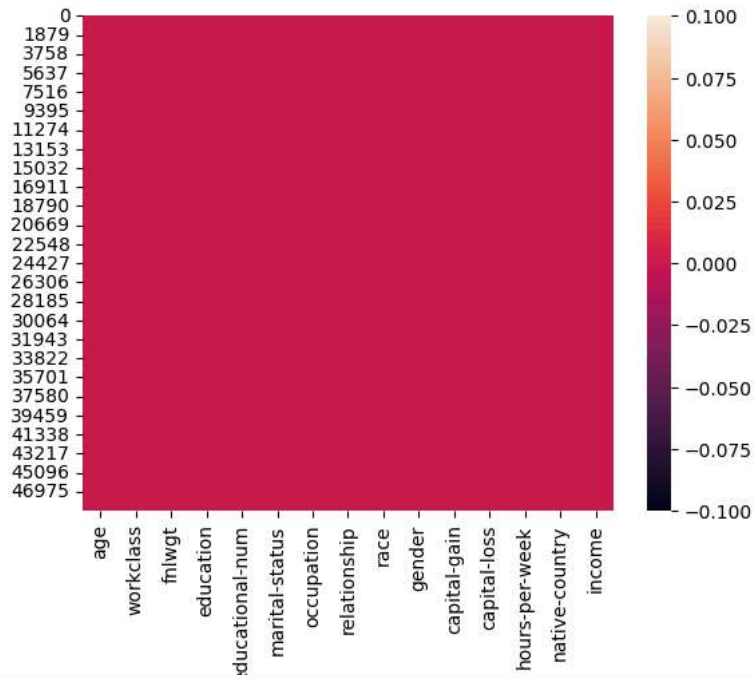
```
data.isnull().sum()
```

	0
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

Generated code may be subject to a license | HimanshuDS14/Exploratory-data-Analysis

```
sns.heatmap(data.isnull())
```

```
<Axes: >
```



7. Perform Data Cleaning [Replace '?' with NaN]

```
data.tail(20)
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours per wee
48822	41	?	202822	HS-grad	9	Separated	?	Not-in-family	Black	Female	0	0	3
48823	72	?	129912	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	0	0	2
48824	45	Local-gov	119199	Assoc-acdm	12	Divorced	Prof-specialty	Unmarried	White	Female	0	0	4
48825	31	Private	199655	Masters	14	Divorced	Other-service	Not-in-family	Other	Female	0	0	3
48826	39	Local-gov	111499	Assoc-acdm	12	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	2
48827	37	Private	198216	Assoc-acdm	12	Divorced	Tech-support	Not-in-family	White	Female	0	0	4
48828	43	Private	260761	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	4
48829	65	Self-emp-not-inc	99359	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	1086	0	6
48830	43	State-gov	255835	Some-college	10	Divorced	Adm-clerical	Other-relative	White	Female	0	0	4
48831	43	Self-emp-not-inc	27242	Some-college	10	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	5
48832	32	Private	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0	0	4
48833	43	Private	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0	0	4
48834	32	Private	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0	0	1
48835	53	Private	321865	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	4
48836	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0	4
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	3
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	4
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	4
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	2
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	4

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

	0
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

```
data.columns
```

```
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')
```

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

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

	0
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

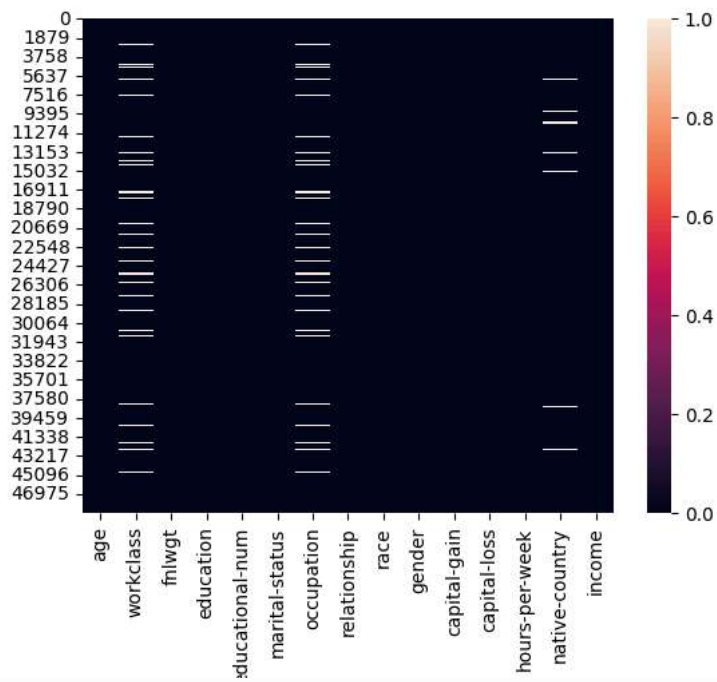

```
data.isnull().sum()
```

	0
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

Generated code may be subject to a license |

```
sns.heatmap(data.isnull())
```

```
<Axes: >
```



8. Drop all The Missing Values

```
per_missing=data.isnull().sum()*100/len(data)
per_missing
```

	0
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

```
data.dropna(how='any',inplace=True)
data.shape
```

```
(45222, 15)
```

9. Check For Duplicate Data and Drop Them

```
dup=data.duplicated().any()
```

```
print("are there any duplicates",dup)
```

```
are there any duplicates True
```

```
data=data.drop_duplicates()
```

10. Get overall Statistics About The DataFrame

```
data.describe()
```

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	45175.000000	4.517500e+04	45175.000000	45175.000000	45175.000000	45175.000000
mean	38.556170	1.897388e+05	10.119314	1102.576270	88.687593	40.942512
std	13.215349	1.056524e+05	2.551740	7510.249876	405.156611	12.007730
min	17.000000	1.349200e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.173925e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783120e+05	10.000000	0.000000	0.000000	40.000000
75%	47.000000	2.379030e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

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

```
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)
```

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

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

11. Drop The Columns education-num, capital-gain, and capital-loss

```
data.columns
```

```
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')
```

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

```
data.columns
```

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

12. What Is The Distribution of Age Column?

```
data.columns
```

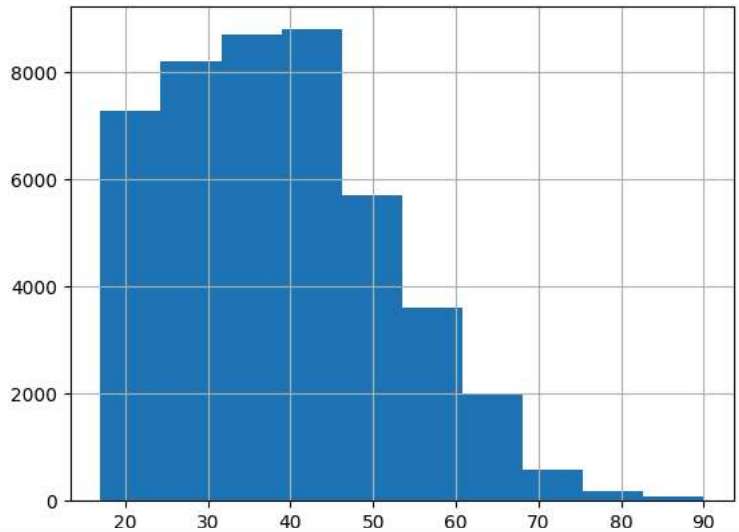
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'marital-status',
       'occupation', 'relationship', 'race', 'gender', 'hours-per-week',
       'native-country', 'income'],
      dtype='object')
```

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

	age
count	45175.000000
mean	38.556170
std	13.215349
min	17.000000
25%	28.000000
50%	37.000000
75%	47.000000
max	90.000000

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

```
<Axes: >
```



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

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

```
34858
```

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

```
34858
```

14. What is The Distribution of Workclass Column?

```
data.columns
```

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

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

	workclass
count	45175
unique	7
top	Private
freq	33262

```
plt.figure(figsize=(10, 10))  # Set figure size properly
data['workclass'].hist()
plt.xlabel("Workclass")
plt.ylabel("Count")
plt.title("Distribution of Workclass")
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.show()
```


Distribution of Workclass
30000
25000
20000
15000
Count