seminarioCD4_graph_pandas

August 27, 2025

[6]:
```
import pandas as pd
```

[7]:
```
url="/Users/marcobarragan/Documents/Seminario-de-Ciencia-de-datos/salary.csv"
data = pd.read_csv(url)
```

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

[8]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

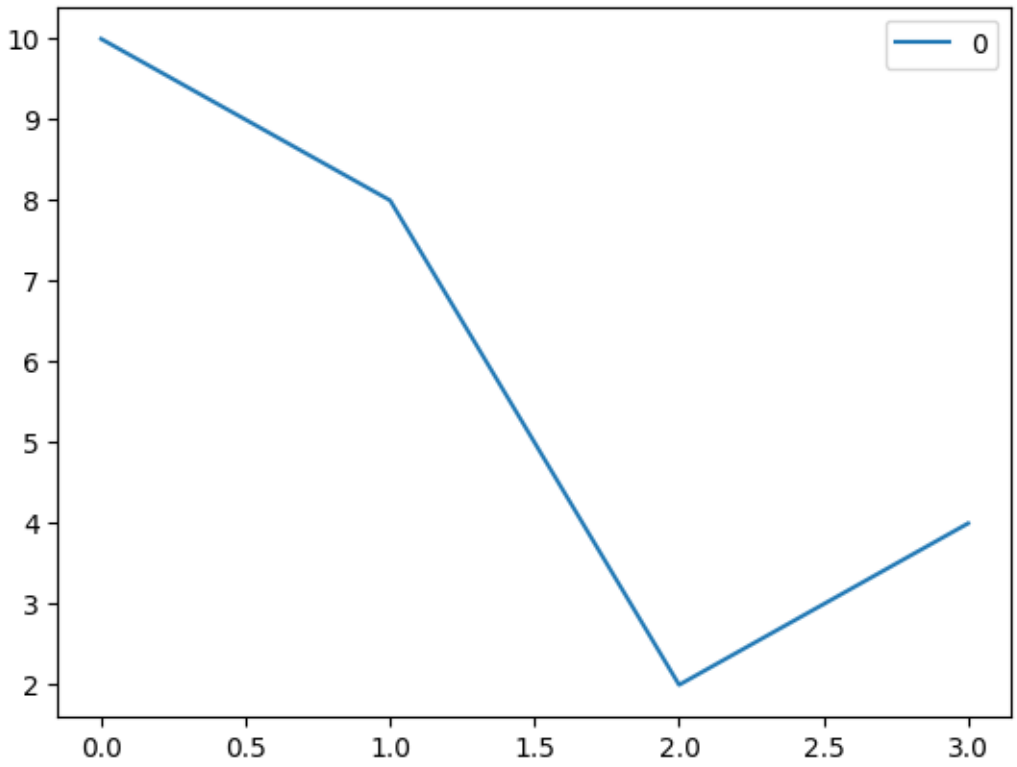
[9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
```

```
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

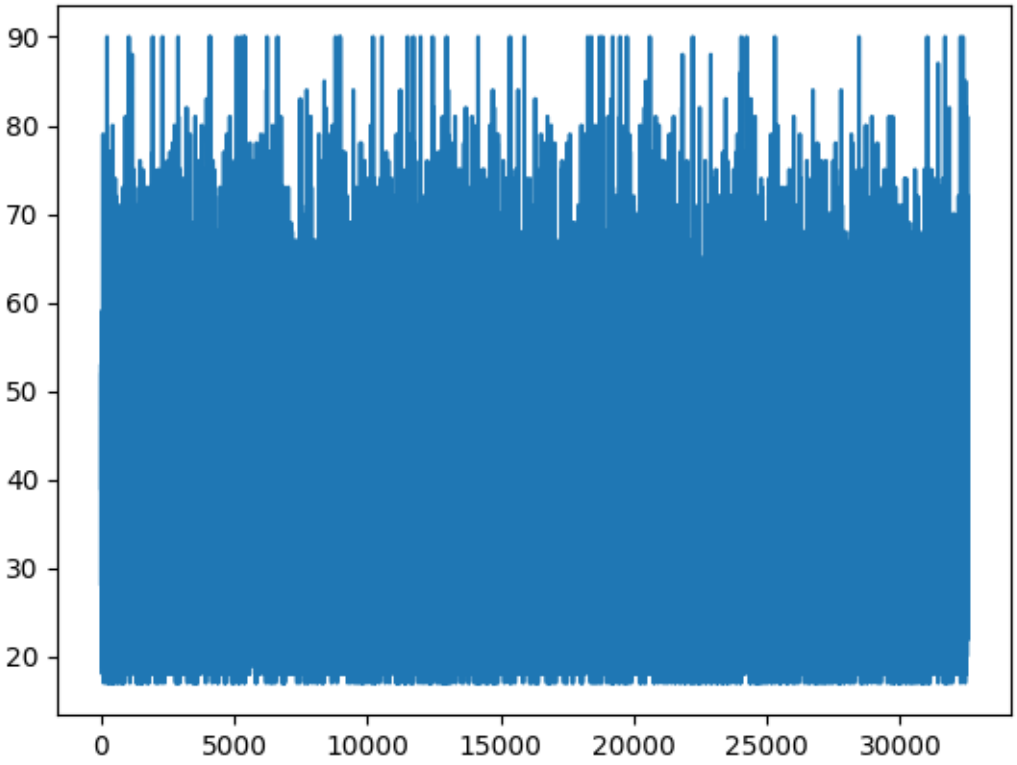
[10]:
```
datos=pd.DataFrame([10,8,2,4])
datos.plot()
```

[10]: <Axes: >



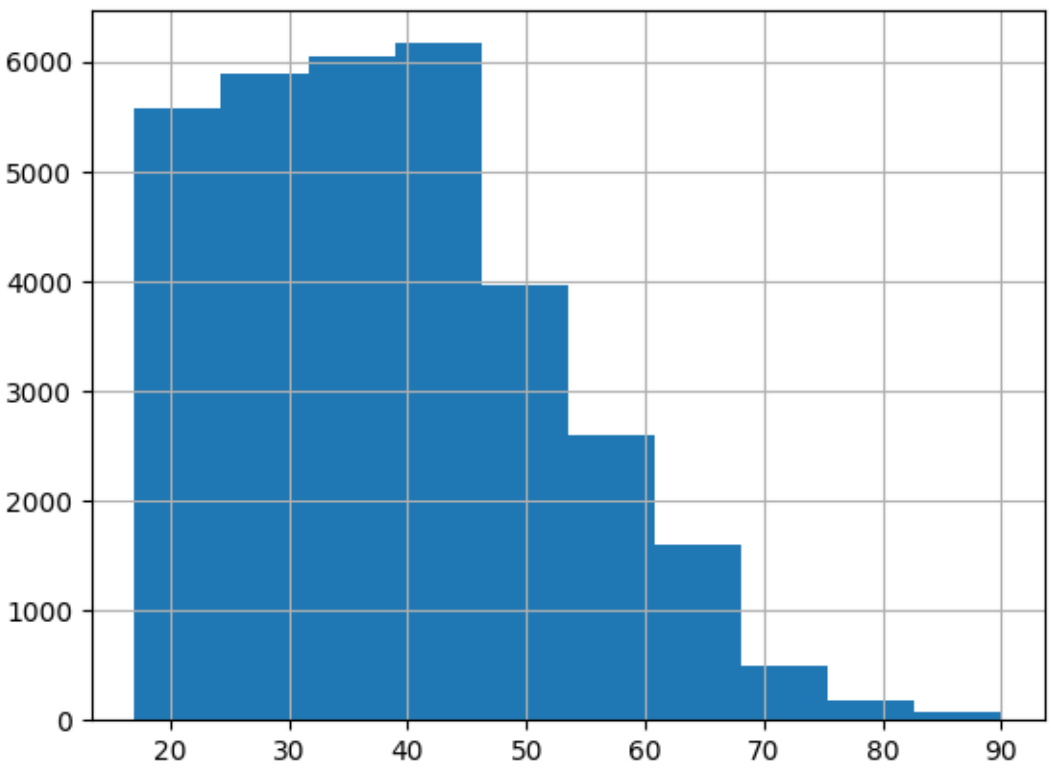
[11]:
```
data["age"].plot()
```

[11]: <Axes: >



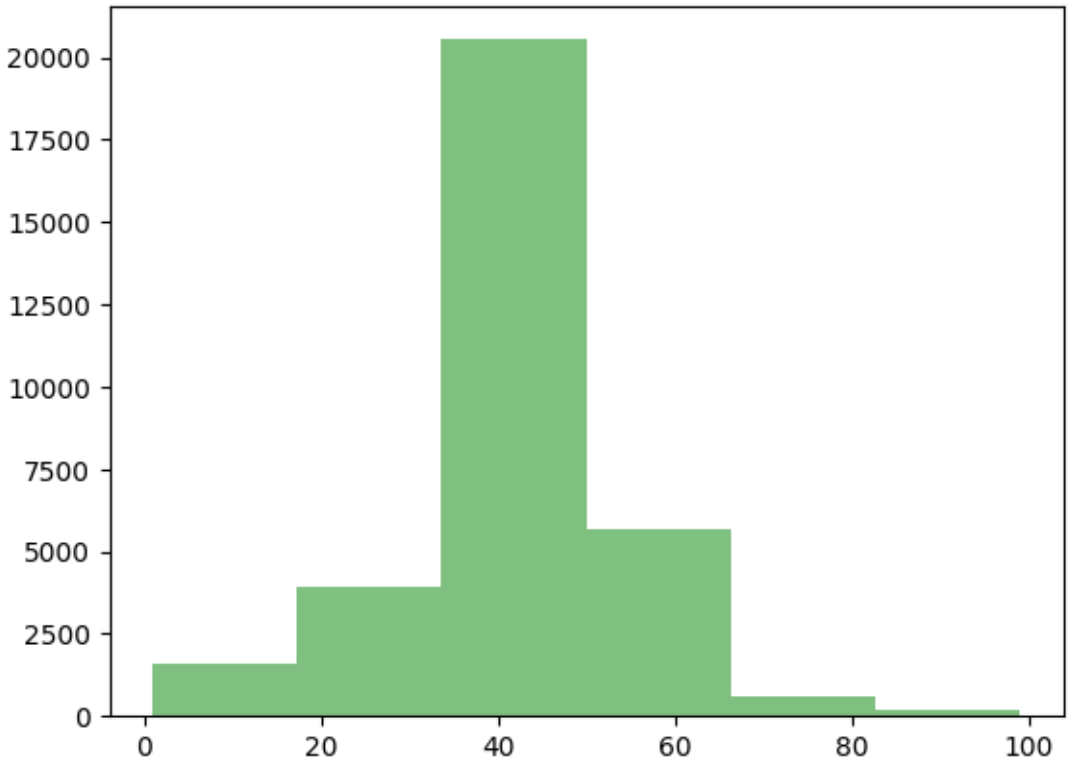
[12]:
```
data["age"].hist()
```

[12]: <Axes: >



[13]:
```
data["hours-per-week"].hist(bins=6, grid=False, color="Green", alpha=0.5)
```

[13]: <Axes: >



0.1 Box plot

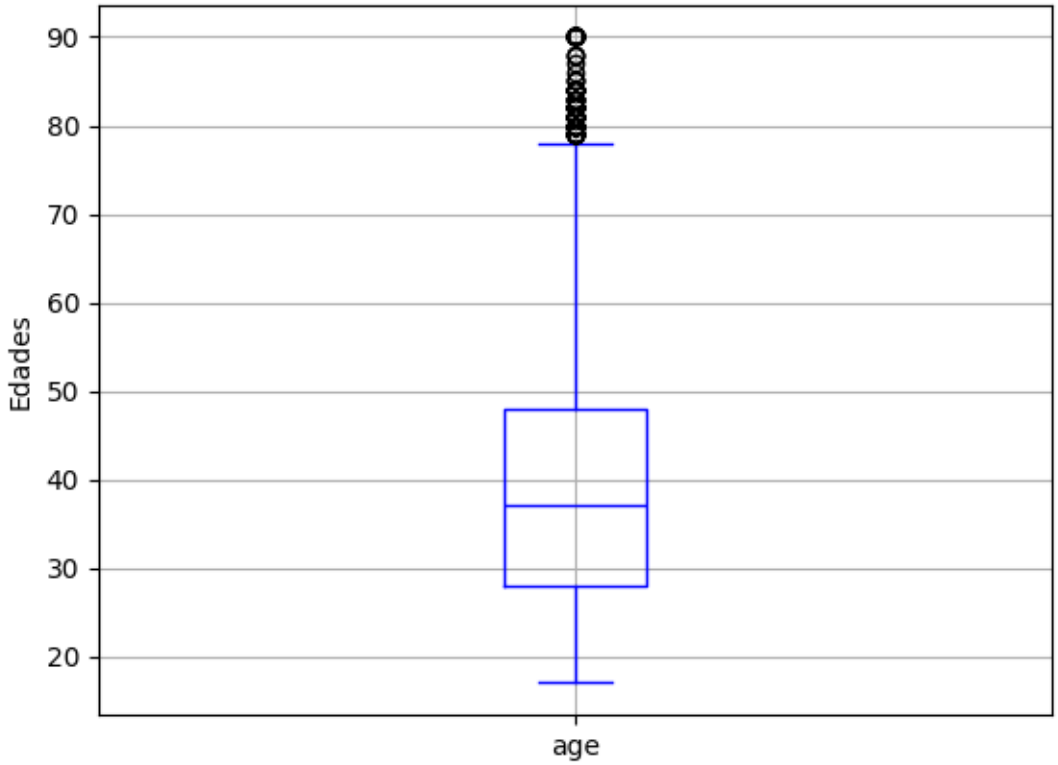
[14]:
```
from scipy import stats
```

[15]:
```
p75ages = stats.scoreatpercentile(data["age"], 75)
print("Percentil 75 = {}".format(p75ages))

p10ages = stats.scoreatpercentile(data["age"], 10)

print("Percentil 10 = {}".format(p10ages))

q2ages = stats.scoreatpercentile(data["age"], 50)

print("Cuartil 2 = {}".format(q2ages))
```

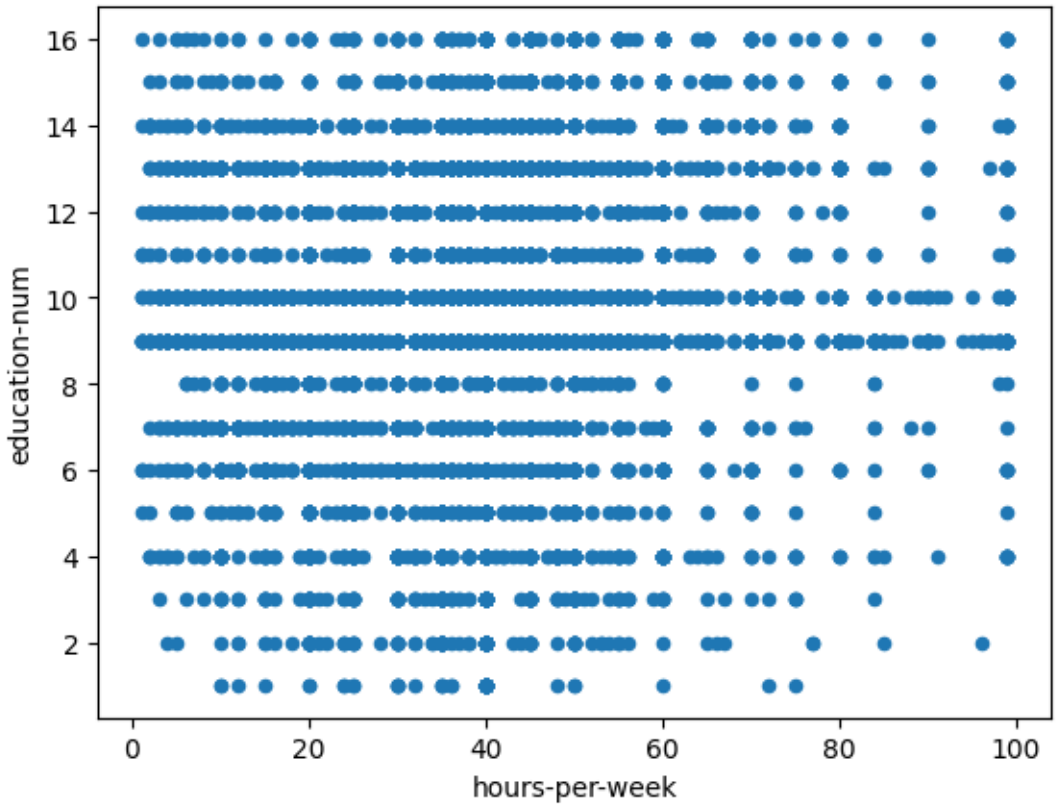
```
Percentil 75 = 48.0
Percentil 10 = 22.0
Cuartil 2 = 37.0
```

[36]:
```
data["age"].plot.box(grid=True, ylabel="Edades", color="blue")
```

[36]: <Axes: ylabel='Edades'>



[37]:
```
data.plot.scatter(x='hours-per-week', y='education-num');
```



[18]:
```
data.head()
```

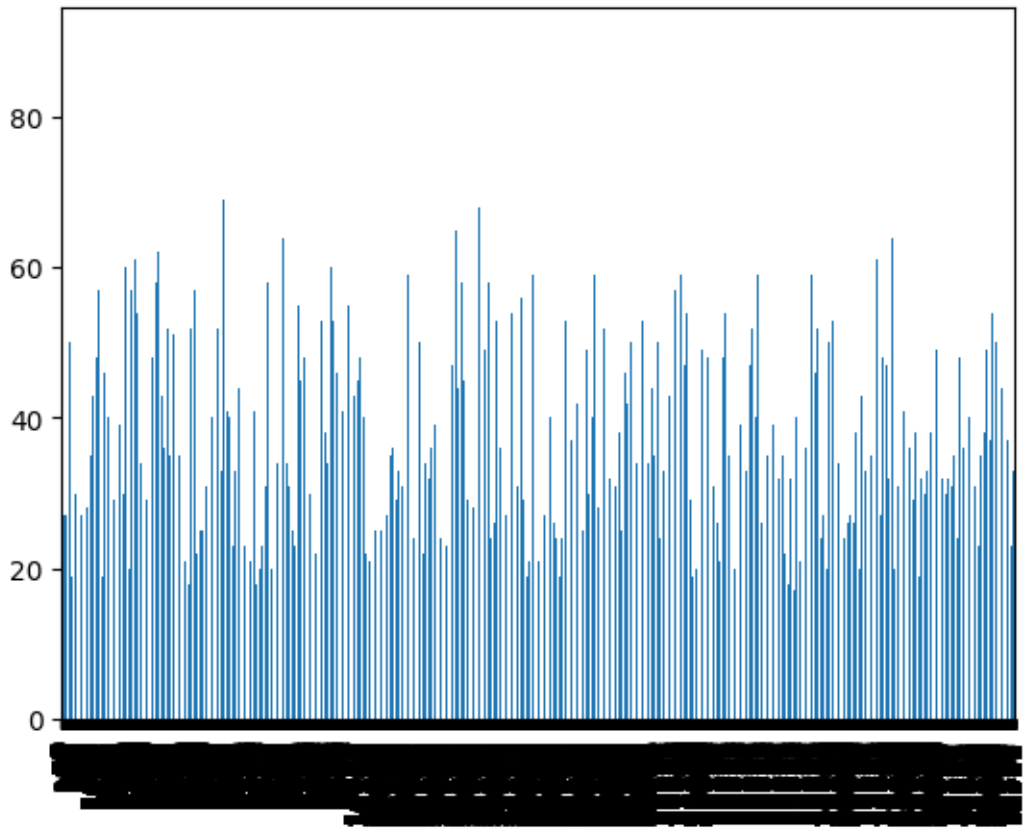
[18]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female					

```
4               0               0               40              Cuba    <=50K
```

[19]:
```
data["age"].plot.bar()
```

[19]:
```
<Axes: >
```



[20]:
```
tab_fr_edad=data['age'].value_counts()
```

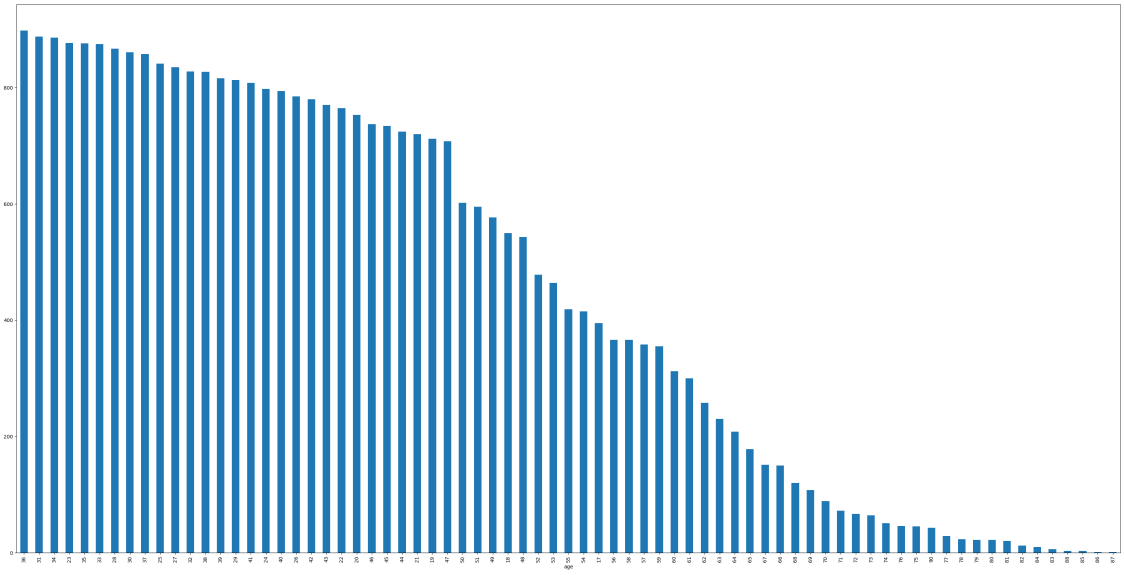
[21]:
```
tab_fr_edad
```

[21]:
```
age
36    898
31    888
34    886
23    877
35    876
     ...
83      6
88      3
85      3
```

```
86      1
87      1
Name: count, Length: 73, dtype: int64
```

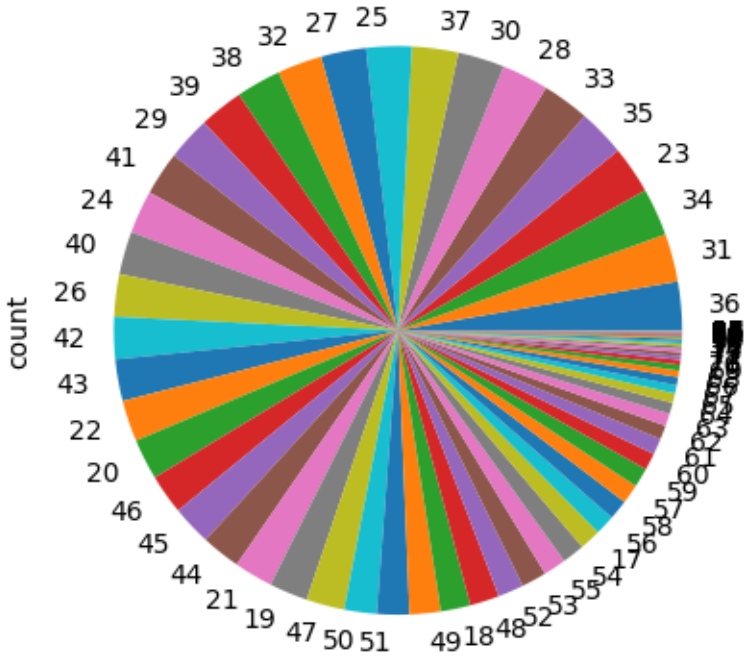
[22]:
```
tab_fr_edad.plot.bar(figsize=(40,20))
```

[22]:
```
<Axes: xlabel='age'>
```



[23]:
```
tab_fr_edad.plot.pie()
```

[23]:
```
<Axes: ylabel='count'>
```

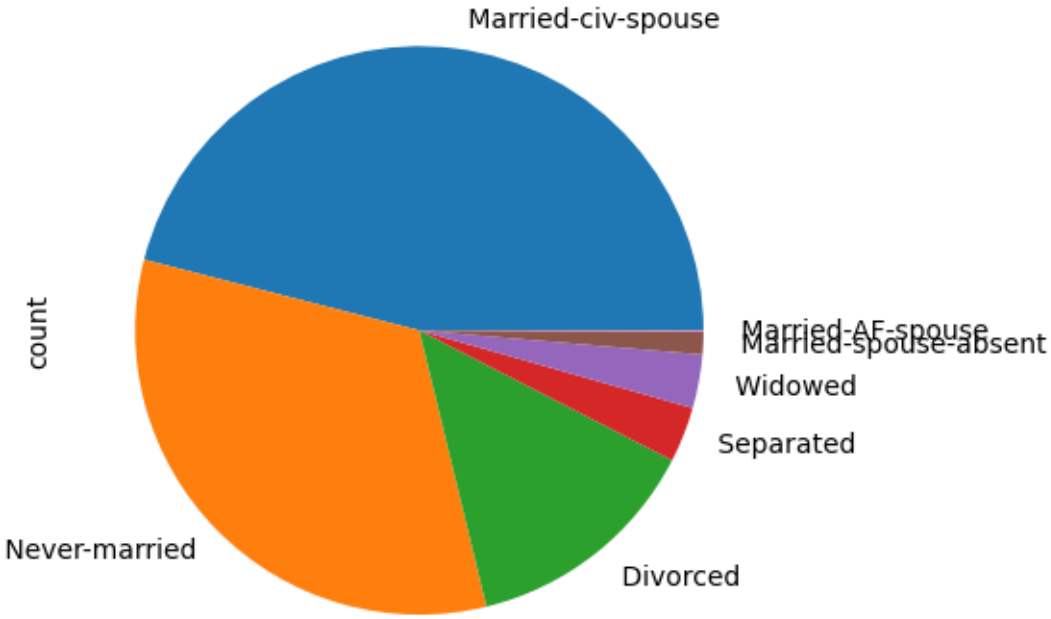


[24]:
```
tab_fr_mar_est=data['marital-status'].value_counts()
tab_fr_mar_est
```

[24]:
```
marital-status
Married-civ-spouse       14976
Never-married            10683
Divorced                  4443
Separated                 1025
Widowed                    993
Married-spouse-absent      418
Married-AF-spouse           23
Name: count, dtype: int64
```

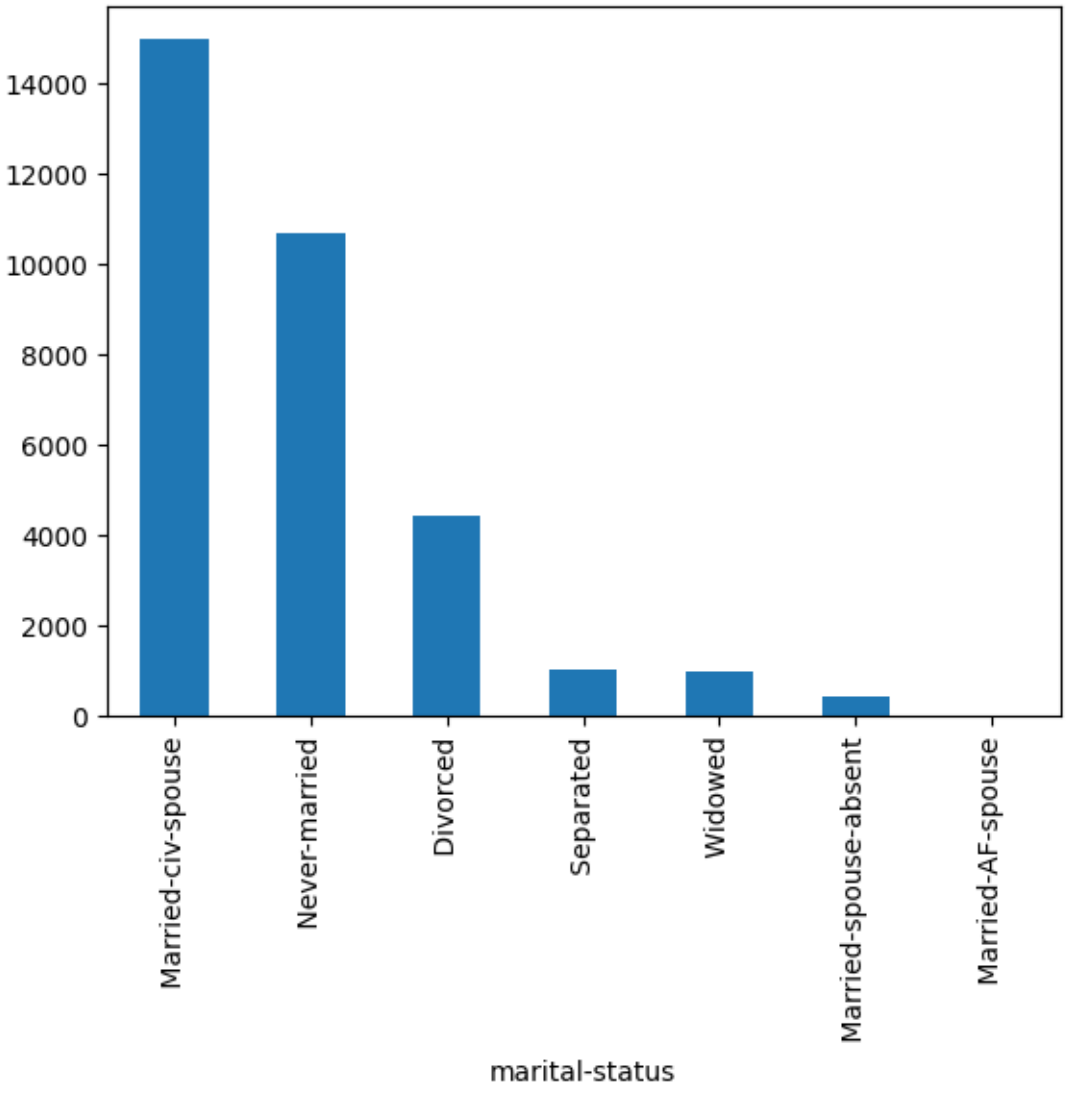
[25]:
```
tab_fr_mar_est.plot.pie()
```

[25]:
```
<Axes: ylabel='count'>
```



[26]:
```
tab_fr_mar_est.plot.bar()
```

[26]: <Axes: xlabel='marital-status'>



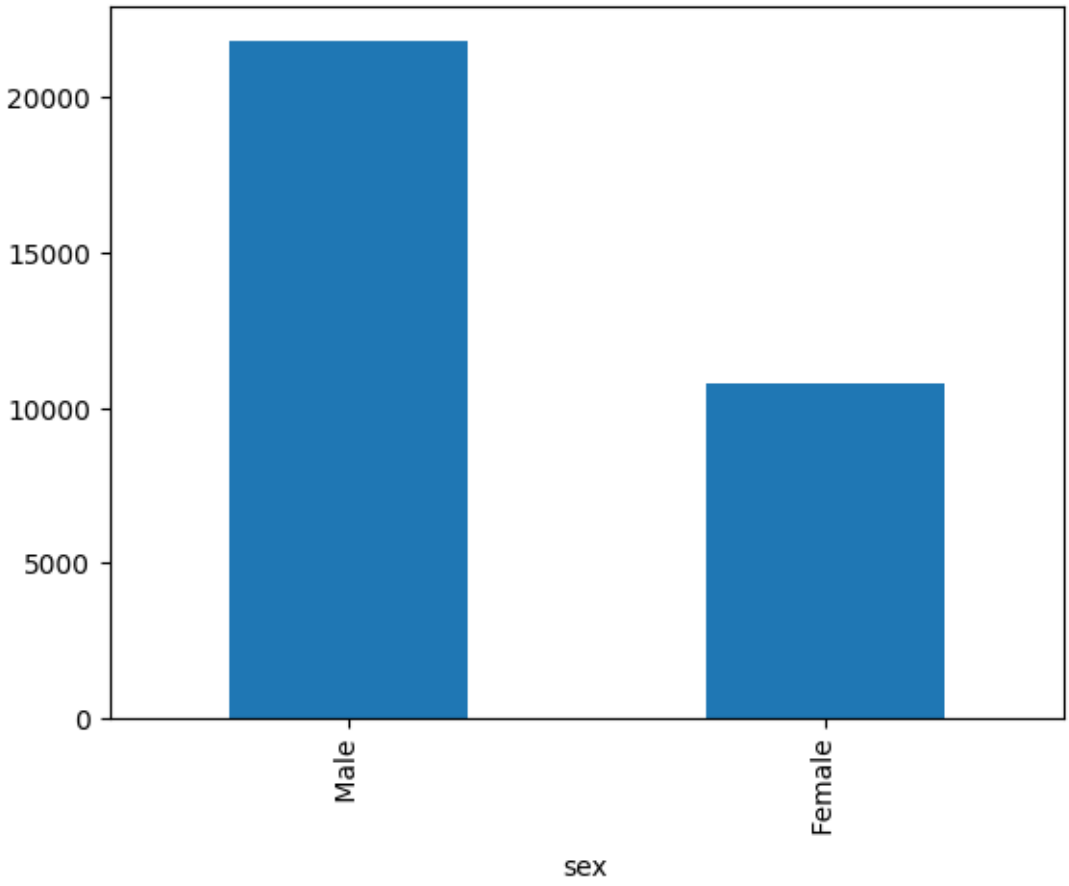
[27]:
```
tab_fr_sex=data["sex"].value_counts()
```

[28]:
```
tab_fr_sex
```

[28]:
```
sex
Male      21790
Female    10771
Name: count, dtype: int64
```

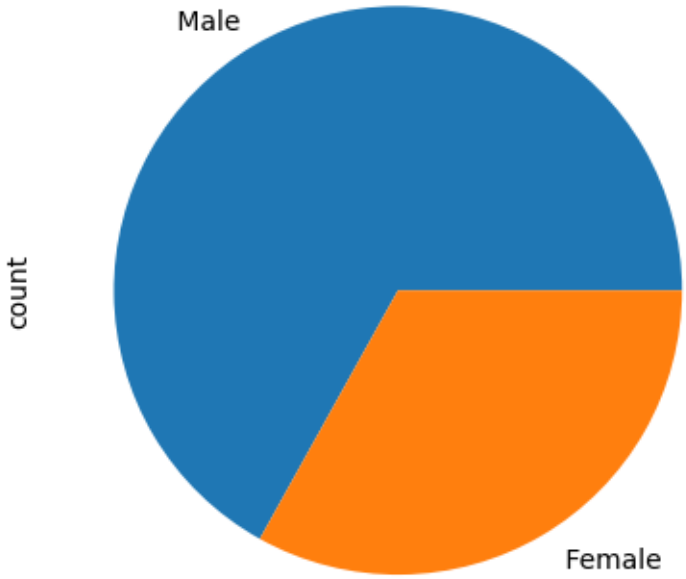
[29]:
```
tab_fr_sex.plot.bar()
```

[29]:
```
<Axes: xlabel='sex'>
```



[30]:
```
tab_fr_sex.plot.pie()
```

[30]:
```
<Axes: ylabel='count'>
```



Filtrar a los empleados con sueldo mayor a 50k para crear nuevos gráficos

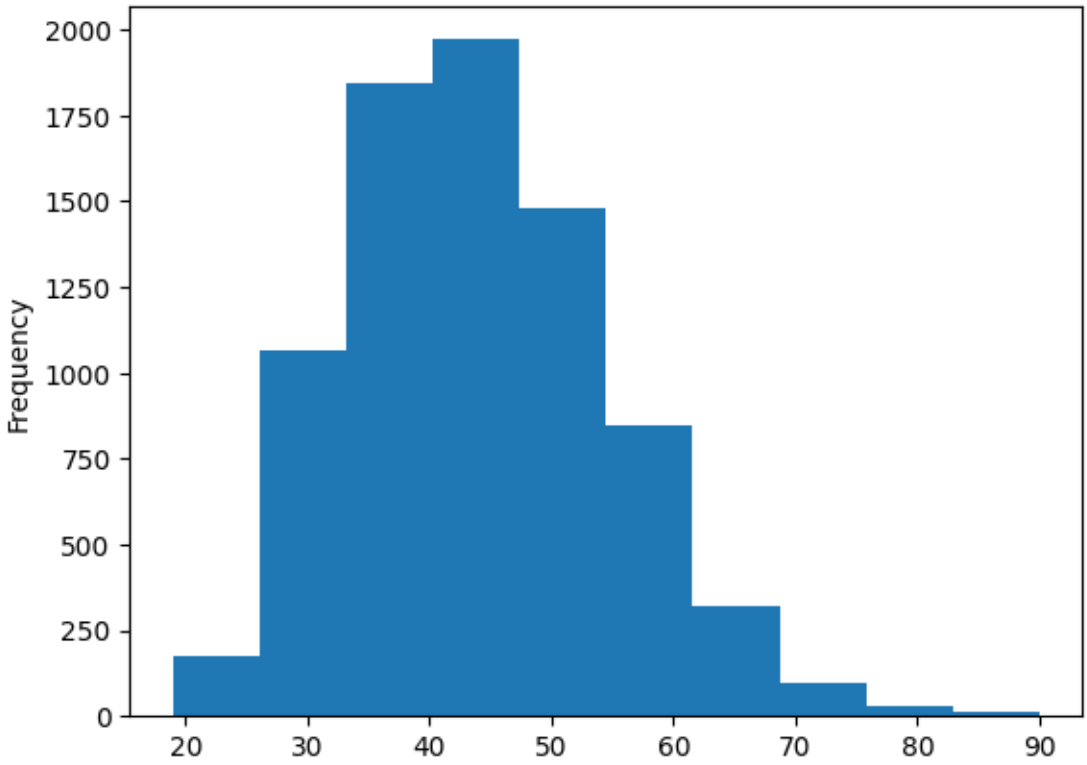
[31]:
```
sal_mayor50=data[data["salary"]==" >50K"]
```

[32]:
```
sal_mayor50.shape
```

[32]:
```
(7841, 15)
```

[33]:
```
sal_mayor50["age"].plot.hist()
```

[33]:
```
<Axes: ylabel='Frequency'>
```

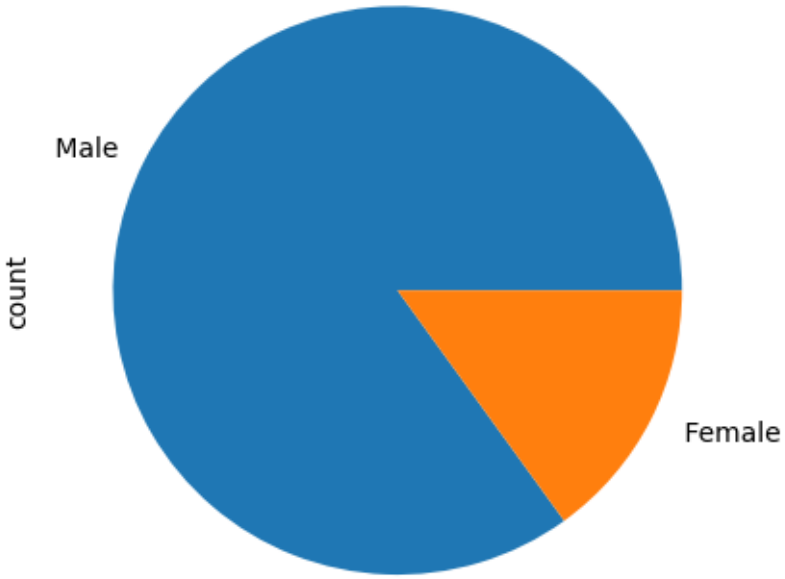


[34]:
```
tab_fr_sex_mayor50=sal_mayor50["sex"].value_counts()
tab_fr_sex_mayor50
```

[34]:
```
sex
Male      6662
Female    1179
Name: count, dtype: int64
```

[35]:
```
tab_fr_sex_mayor50.plot.pie()
```

[35]:
```
<Axes: ylabel='count'>
```


Male
count
Female