Week 5: Project

This dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Adult).

It was extracted by Barry Becker from the 1994 Census database. This database contains age, workclass, fnlwgt, education level, the years for education, marital status, occupation, relationship, race, sex, capital gain, capital loss, working hours per week, native-country and other information to know whether the annual income of a person exceeds 50K

There are a total of 1643623 records in this database, and some parts of record are missing.

I need to clean up the missing values in the database first, then sort the data into a DataFrame to facilitate the next data analysis and visualization. Firstly, initialize the required list.

In [5]:

```
path = "adult.data"
f = open(path,'r',encoding = 'utf8')
read = str(f.read())
people = read.split('\n')
age=[]
workclass=[]
fnlwgt=[]
education = []
educationnum =[]
maritalstatus =[]
occupation =[]
relationship =[]
race =[]
sex =[]
capitalgain =[]
capitalloss =[]
hoursperweek =[]
nativecountry =[]
income =[]
```

Second, Process data and generate a DataFrame

In [10]:

```
import pandas as pd
import matplotlib.pyplot as plt
for d in people:
    num = len(d.split(', '))
    if num == 15:
        age.append(int(d.split(', ')[0]))
        workclass.append(d.split(', ')[1])
        fnlwgt.append(d.split(', ')[2])
        education.append(d.split(', ')[3])
        educationnum.append(int(d.split(', ')[4]))
        maritalstatus.append(d.split(', ')[5])
        occupation.append(d.split(', ')[6])
        relationship.append(d.split(', ')[7])
        race.append(d.split(', ')[8])
        sex.append(d.split(', ')[9])
        capitalgain.append(d.split(', ')[10])
        capitalloss.append(d.split(', ')[11])
        hoursperweek.append(int(d.split(', ')[12]))
        nativecountry.append(d.split(', ')[13])
        incomes=d.split(', ')[14]
        if incomes == '<=50K':
            income.append(0)
        else:
            income.append(1)
df = pd.DataFrame({'age':age,'workclass':workclass,'fnlwgt':fnlwgt,'education':educatio
n,'educationnum':educationnum,'maritalstatus':maritalstatus,'occupation':occupation,'re
lationship':relationship,'race':race,'sex':sex,'capitalgain':capitalgain,'capitalloss':
capitalloss,'hoursperweek':hoursperweek,'nativecountry':nativecountry,'income':income})
```

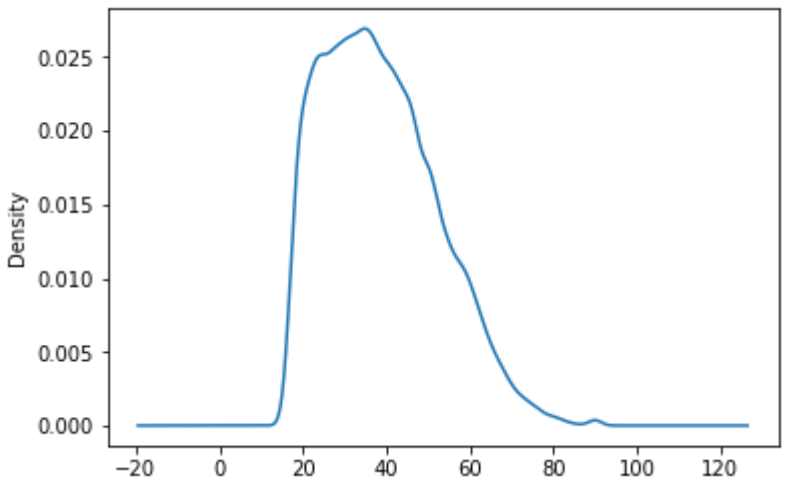
Then, Visualize and analyze data

In [11]:

```
df['age'].plot(kind='kde')
```

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20ead068240>
```



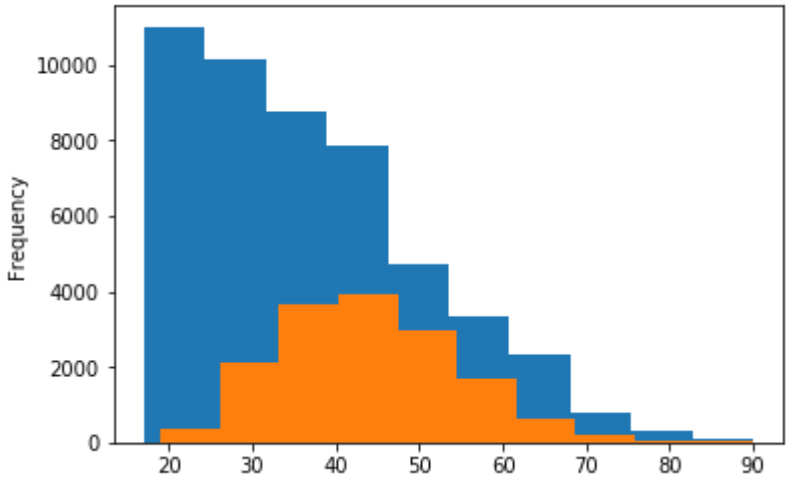
We can observe the age distribution of the Kernel Density Estimate Plot.
It can be seen from the image that the data is mainly concentrated in the age of 20 to 60 years old. In the case of older than 60 years, errors may occur due to insufficient sample size.

In [12]:

```
df[['age','income']].groupby(by='income')['age'].plot(kind='hist', stacked=True)
```

Out[12]:

```
income
0    AxesSubplot(0.125,0.125;0.775x0.755)
1    AxesSubplot(0.125,0.125;0.775x0.755)
Name: age, dtype: object
```

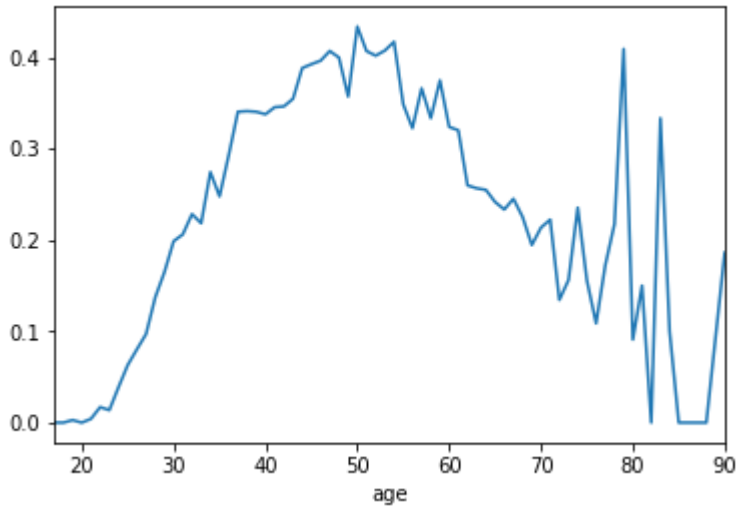


In [13]:

```
df[['age','income']].groupby(by='age')['income'].mean().plot(kind='line')
```

Out[13]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20ead008710>
```



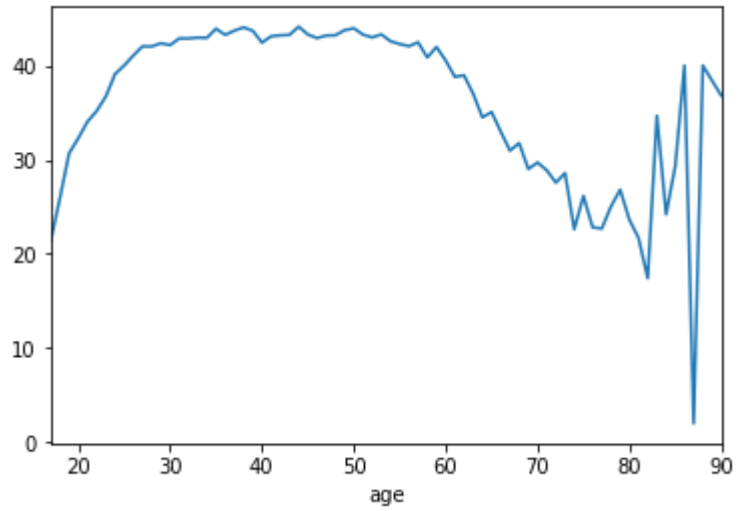
The first diagram is Stacked Histogram Plot and the second is Line plot.
We can compare the proportion of higher income and lower income in different age groups.
As can be seen from the above figures, people around the age of 50 have the highest probability of earning high income. The proportion of high-income people from 20 to 50 years old is gradually increasing, and gradually decreasing after 50 years old.

In [14]:

```
df[['age','hoursperweek']].groupby(by='age')['hoursperweek'].mean().plot(kind='line')
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eab6fd630>
```



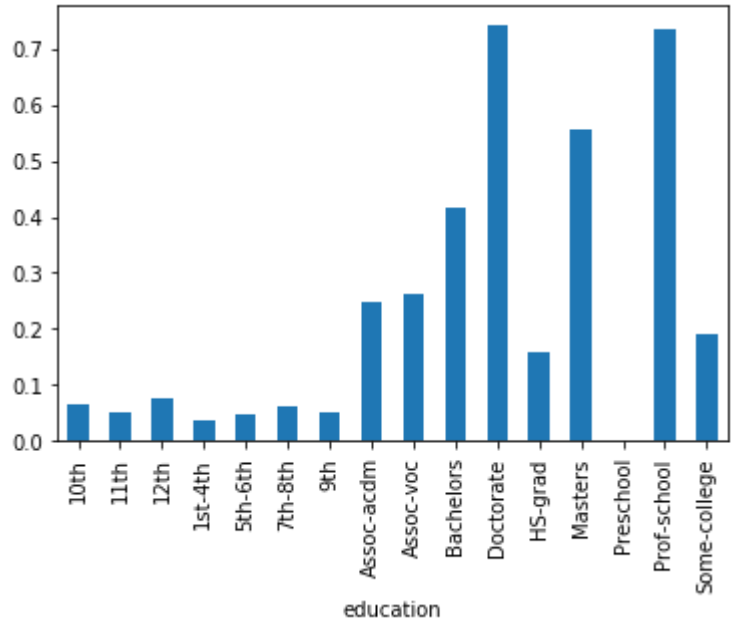
This is a Line plot.
We can analyze changes and trends in average weekly working hours as the age.
From the age of seventeen to twenty-seven, the working hours per week gradually increase. From 27 to 60 years old, working hours are about 43 hours, and working hours begain to decline after the age of 60. After the age of 70, there is a large fluctuation due to insufficient sample size.

In [15]:

```
df[['education','income']].groupby(by='education')['income'].mean().plot(kind='bar')
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eab887550>
```



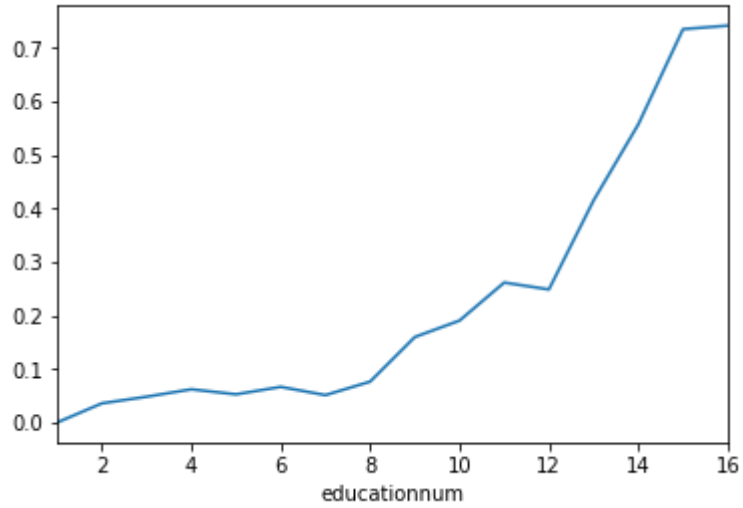
This is Bar Chart.
We can know the income status of people with different academic qualifications.
Obviously, the higher education level, the greater the probability of becoming a high-income people.

In [16]:

```
df[['educationnum','income']].groupby(by='educationnum')['income'].mean().plot(kind='line')
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eab96d6d8>
```



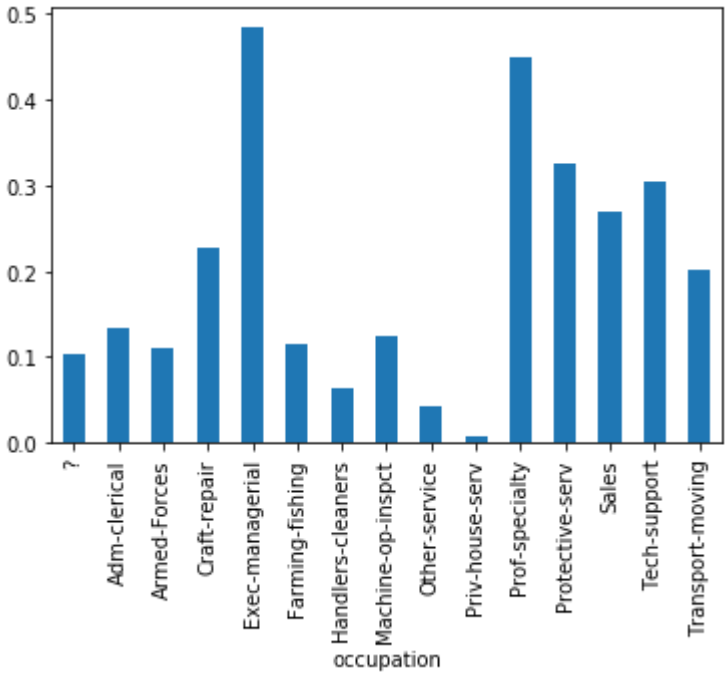
This is a Line Plot.
It shows the income status of people with different education years.
Generally, the longer the education time, the higher income.

In [17]:

```
df[['occupation','income']].groupby(by='occupation')['income'].mean().plot(kind='bar')
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eab7f5128>
```



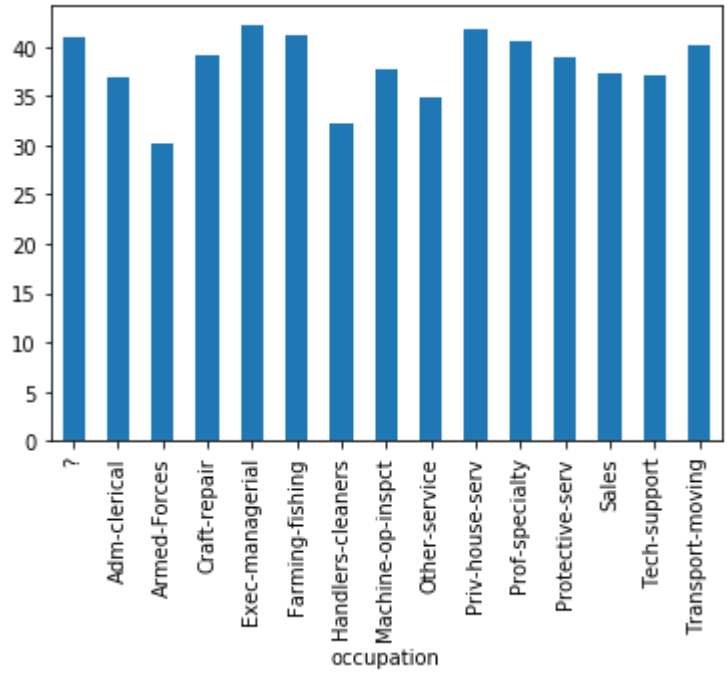
This is a Bar Chart.
It shows income status of people with different occupation.
Prof-specialty, Exec-managerial and Protective-serv have higher salary.

In [18]:

```
df[['occupation','age']].groupby(by='occupation')['age'].mean().plot(kind='bar')
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eafb19748>
```



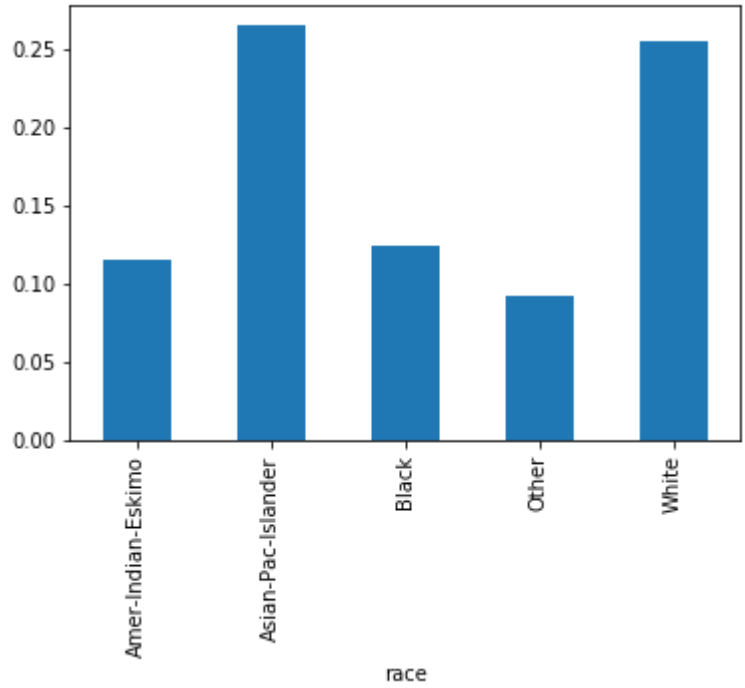
This is a Bar Chart.
Higher income positions do not necessarily require a higher age.

In [19]:

```
df[['race','income']].groupby(by='race')['income'].mean().plot(kind='bar')
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eafba5780>
```



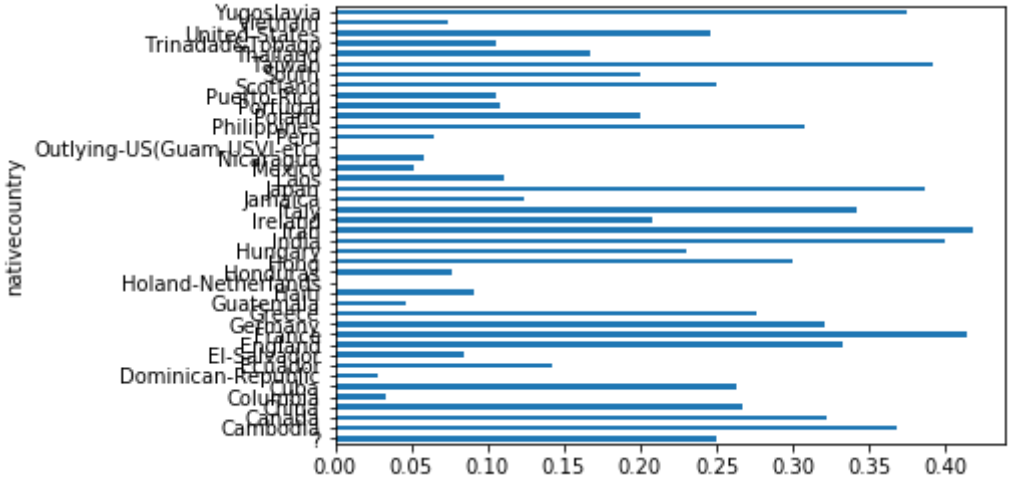
This is a Bar Chart.
It shows Income status of people with different races.
White and Asian people maybe have higher salary.

In [21]:

```
df[['nativecountry','income']].groupby(by='nativecountry')['income'].mean().plot(kind=
'barh')
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20eafd08a90>
```



This is a Bar Chart.
It describes the income status of people with different nativecountries.
People from Cambodia, France, India, Iran, Japan, Taiwan and Yugoslavia have a higher probability of getting a higher salary.

Thank you!