In [3]:

```
import numpy as np
import pandas as pd
from pandas import DataFrame
```

In [4]:

```
file_path = "./winemag-data_first150k.csv"
df = pd.read_csv(file_path, header=0)
```

In [5]:

```
# 数据摘要
df["country"].value_counts() # 标称属性 例：country频数
```

Out[5]:

```
US                        62397
Italy                     23478
France                    21098
Spain                      8268
Chile                      5816
Argentina                  5631
Portugal                   5322
Australia                  4957
New Zealand                3320
Austria                    3057
Germany                    2452
South Africa               2258
Greece                      884
Israel                      630
Hungary                     231
Canada                      196
Romania                     139
Slovenia                     94
Uruguay                      92
Croatia                      89
Bulgaria                     77
Moldova                      71
Mexico                       63
Turkey                       52
Georgia                      43
Lebanon                      37
Cyprus                       31
Brazil                       25
Macedonia                    16
Serbia                       14
Morocco                      12
England                       9
Luxembourg                    9
Lithuania                     8
India                         8
Czech Republic                6
Ukraine                       5
Switzerland                   4
South Korea                   4
Bosnia and Herzegovina        4
China                         3
Egypt                         3
Slovakia                      3
Tunisia                       2
Albania                       2
Montenegro                    2
Japan                         2
US-France                     1
Name: country, dtype: int64
```

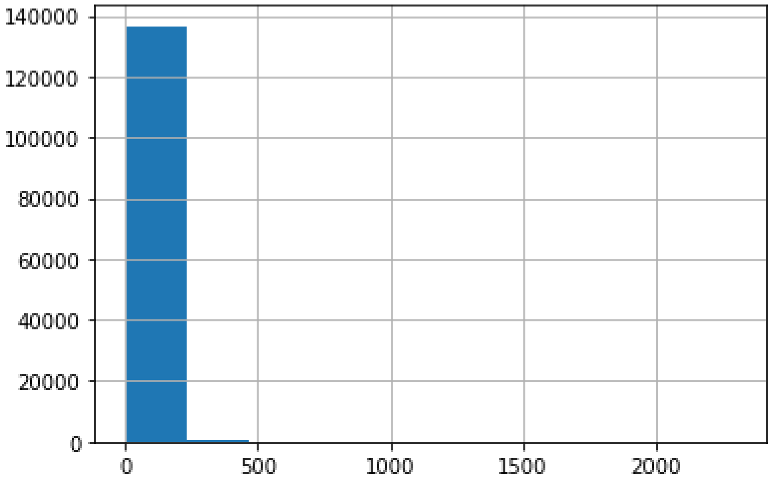
In [9]:

```
nums = df["price"] # 数值属性 5数概括及缺失值的个数 例：price
nullnum = nums.isnull().sum()
nums = nums.dropna(axis = 0)
Minimum = min(nums)
Maximum = max(nums)
Q1 = np.percentile(nums, 25)
Median = np.median(nums)
Q3 = np.percentile(nums, 75)
print("缺失值个数：{}".format(nullnum))
print("最小值：{}".format(Minimum))
print("Q1：{}".format(Q1))
print("中位数：{}".format(Median))
print("Q3：{}".format(Q3))
print("最大值：{}".format(Maximum))
```

```
缺失值个数：13695
最小值：4.0
Q1：16.0
中位数：24.0
Q3：40.0
最大值：2300.0
```

In [5]:

```
# 数据可视化
import matplotlib.pyplot as plt
hist = df["price"].hist() # 直方图
```

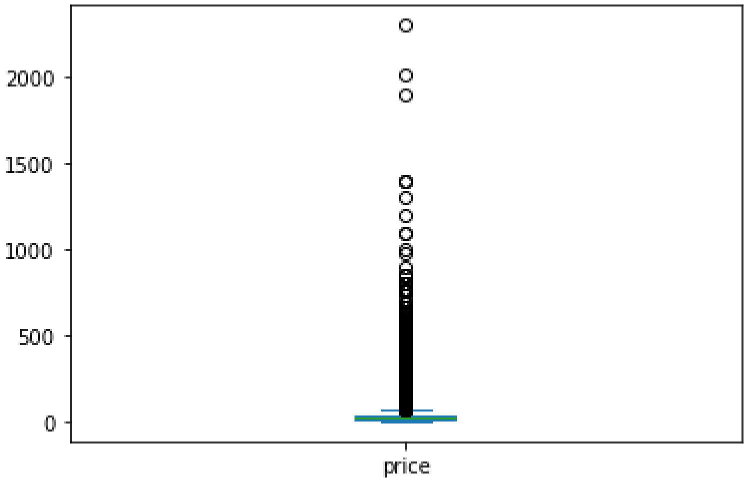


In [15]:

```
df["price"].plot.box() # 盒图及离群点
```

Out[15]:

```
<AxesSubplot:>
```



In [16]:

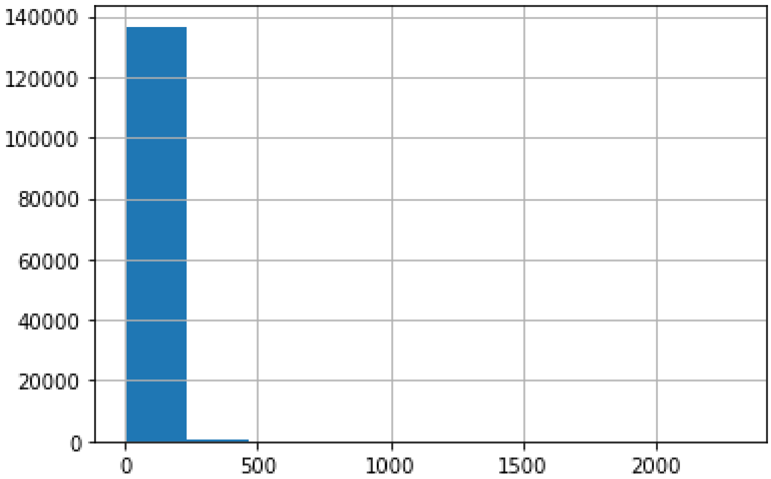
```
# 缺失值处理
# 剔除缺失值
data_dropna = df["price"].dropna(axis = 0)
```

In [17]:

```
data_dropna.hist() # 直方图
```

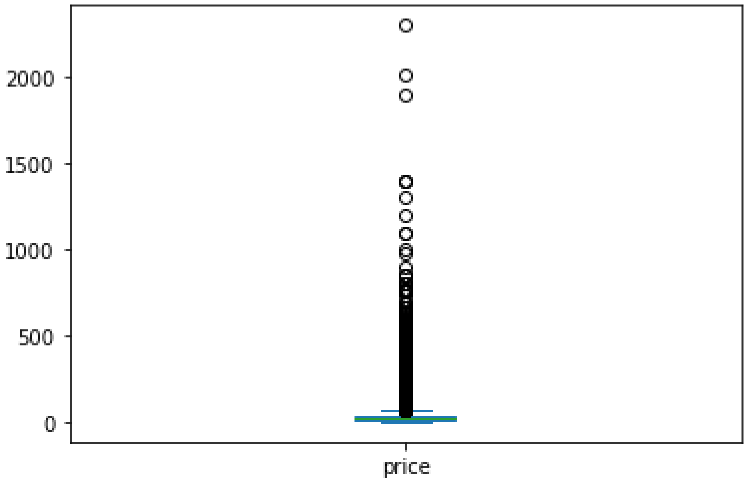
Out[17]:

```
<AxesSubplot:>
```



In [18]:

```
data_dropna.plot.box()
plt.show() # 盒图
```

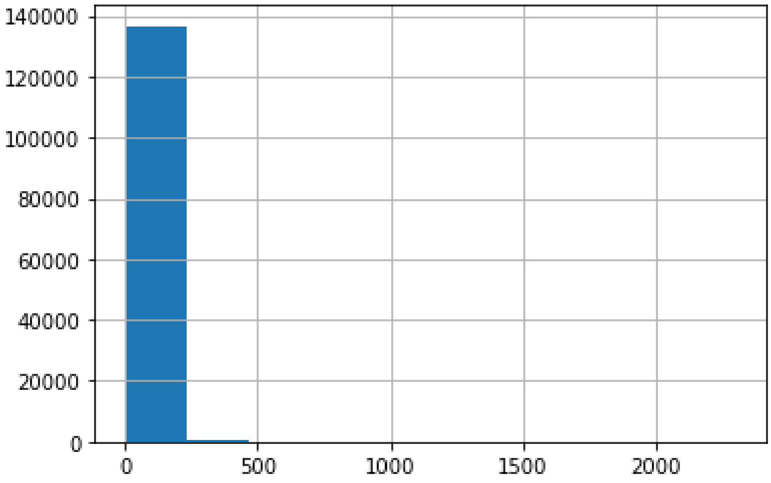


In [6]:

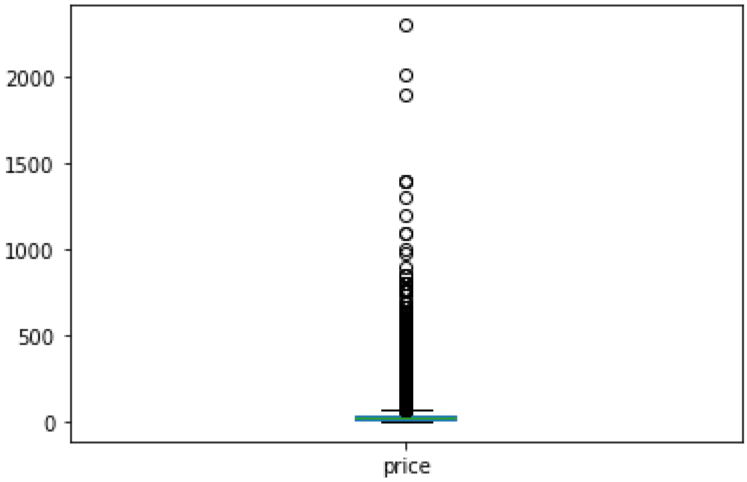
```
# 用最高频率值来填补缺失值
data_fillna=df["price"].fillna(df["price"].mode())
data_fillna.hist() # 直方图
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1171710e8c8>
```



```
In [7]:
data_fillna.plot.box()
plt.show() # 盒图
```



```
In [ ]:
```