

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

```
path = "/Users/zhangpuchang/Downloads/archive/clinical_data.csv"
df = pd.read_csv(path, header=0)
```

数据摘要

```
df["sex"].value_counts() # 标称属性 例: sex
```

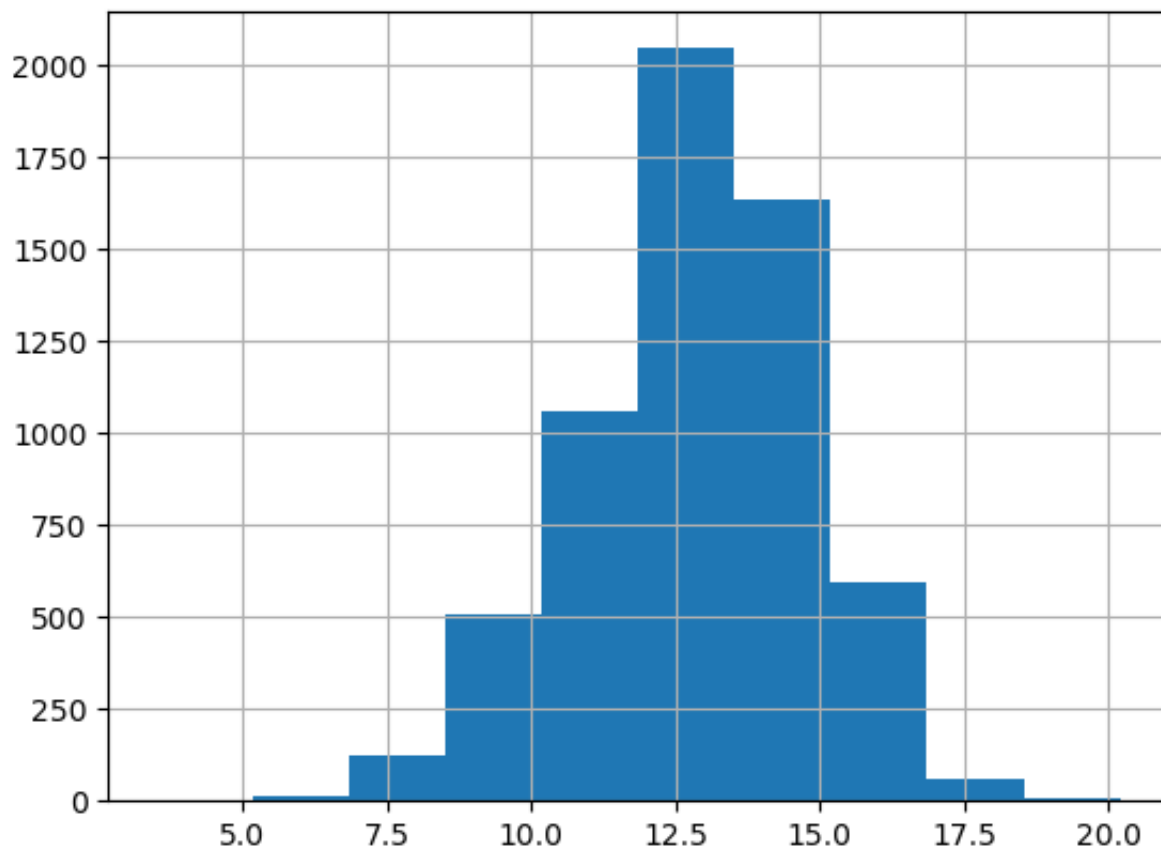
```
M      3243
F      3145
Name: sex, dtype: int64
```

```
preop_hb = df["preop_hb"] # 数值属性 5数概括及缺失值的个数 例: IMDb-rating
null_cnt = preop_hb.isnull().sum()
preop_hb = preop_hb.dropna(axis = 0)
min_preop_hb = min(preop_hb)
max_preop_hb = max(preop_hb)
pct_25 = np.percentile(preop_hb, 25)
Median = np.median(preop_hb)
pct_75 = np.percentile(preop_hb, 75)
print("缺失值个数: {}".format(null_cnt))
print("最小值: {}".format(min_preop_hb))
print("Q1: {}".format(pct_25))
print("中位数: {}".format(Median))
print("Q3: {}".format(pct_75))
print("最大值: {}".format(max_preop_hb))
```

```
缺失值个数: 341
最小值: 3.5
Q1: 11.6
中位数: 13.0
Q3: 14.2
最大值: 20.2
```

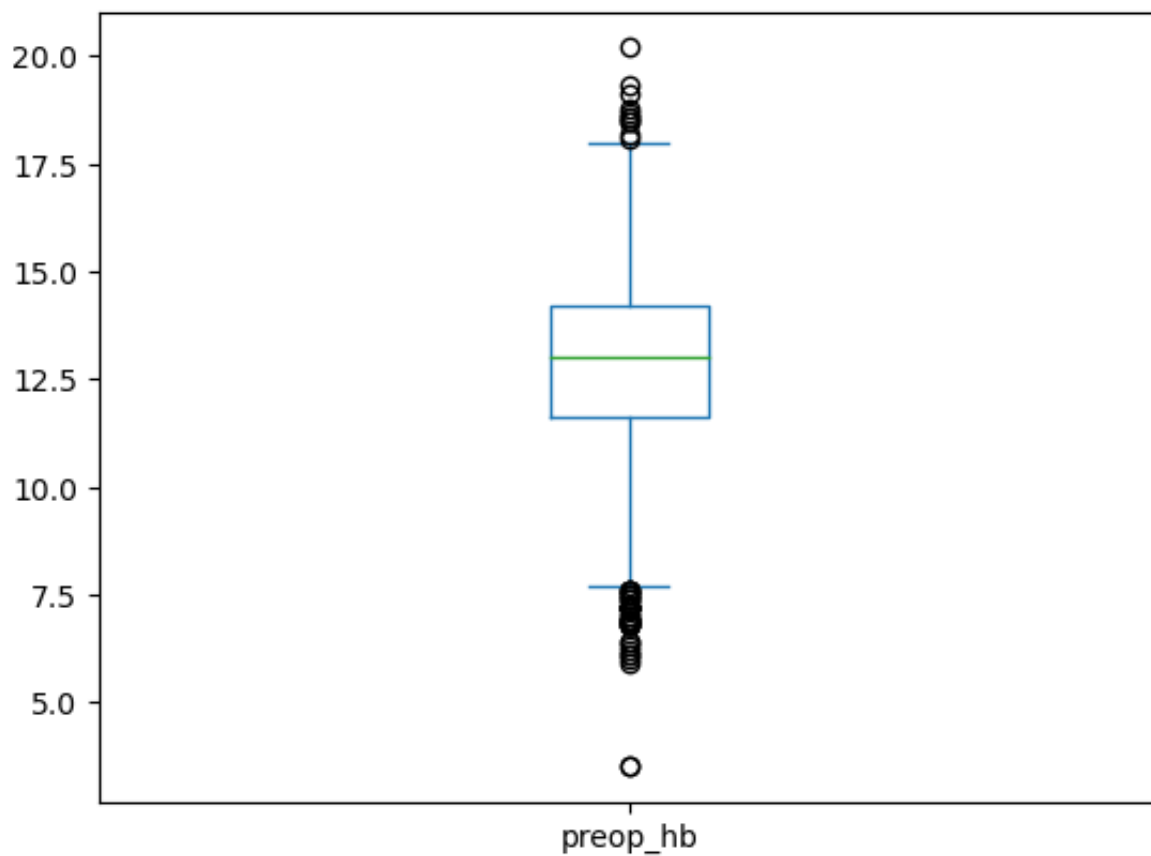
```
# 数据可视化
import matplotlib.pyplot as plt
df["preop_hb"].hist() # 直方图
```

<Axes: >



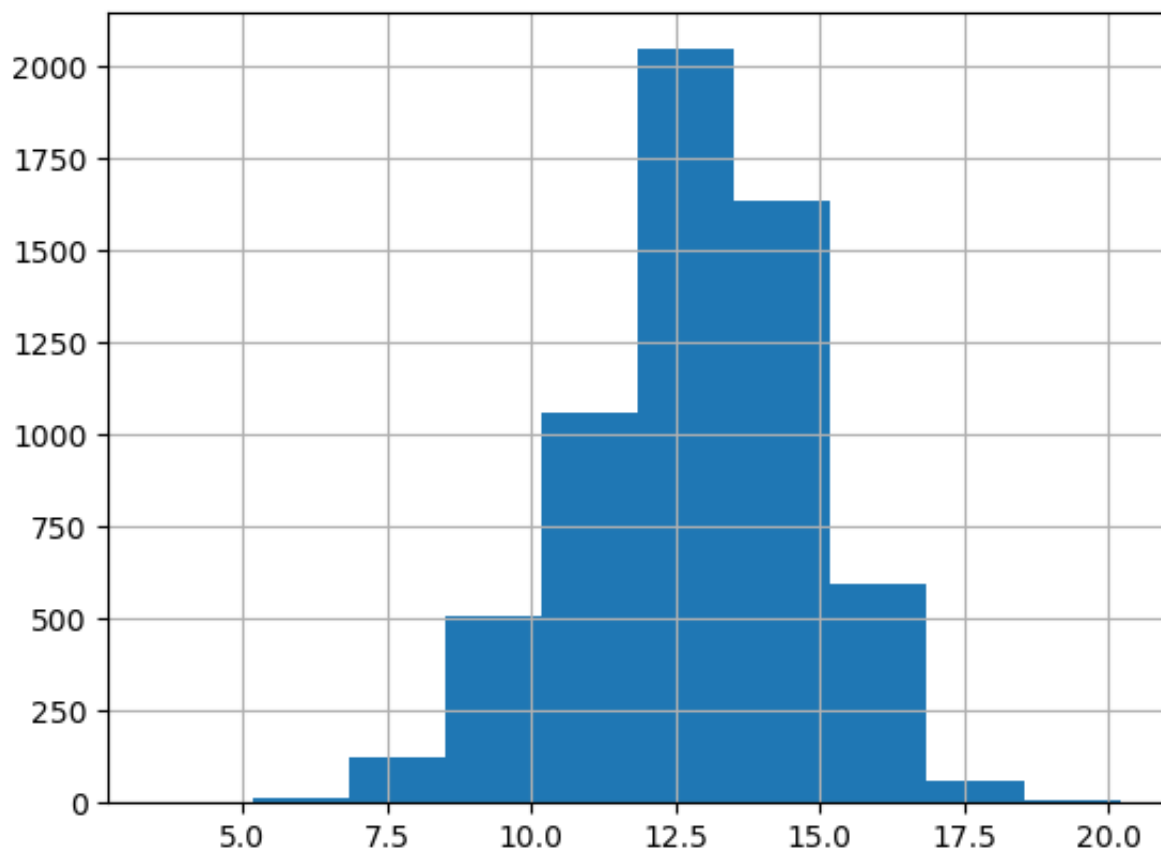
```
df["preop_hb"].plot.box() # 盒图及离群点
```

<Axes: >



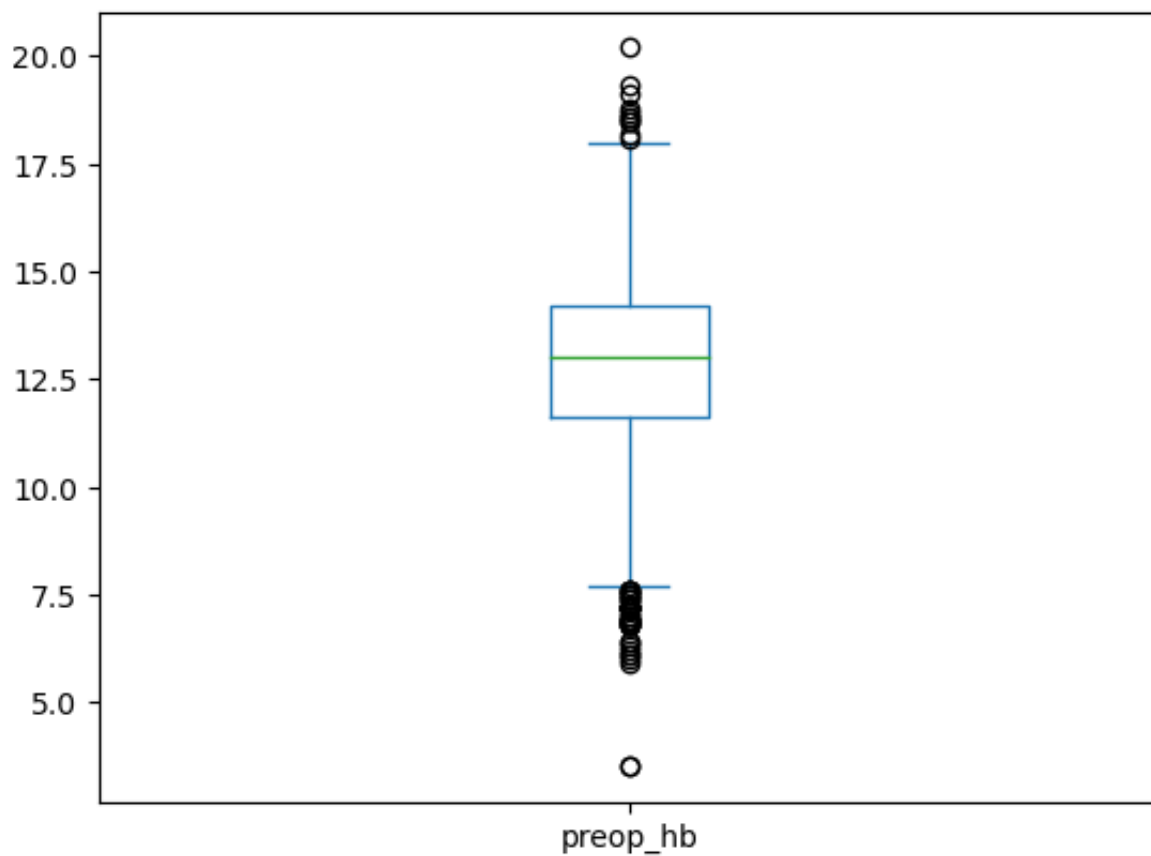
```
# 缺失值处理  
# 剔除缺失值  
data_dropna = df["preop_hb"].dropna(axis = 0)  
df["preop_hb"].hist() #直方图
```

<Axes: >



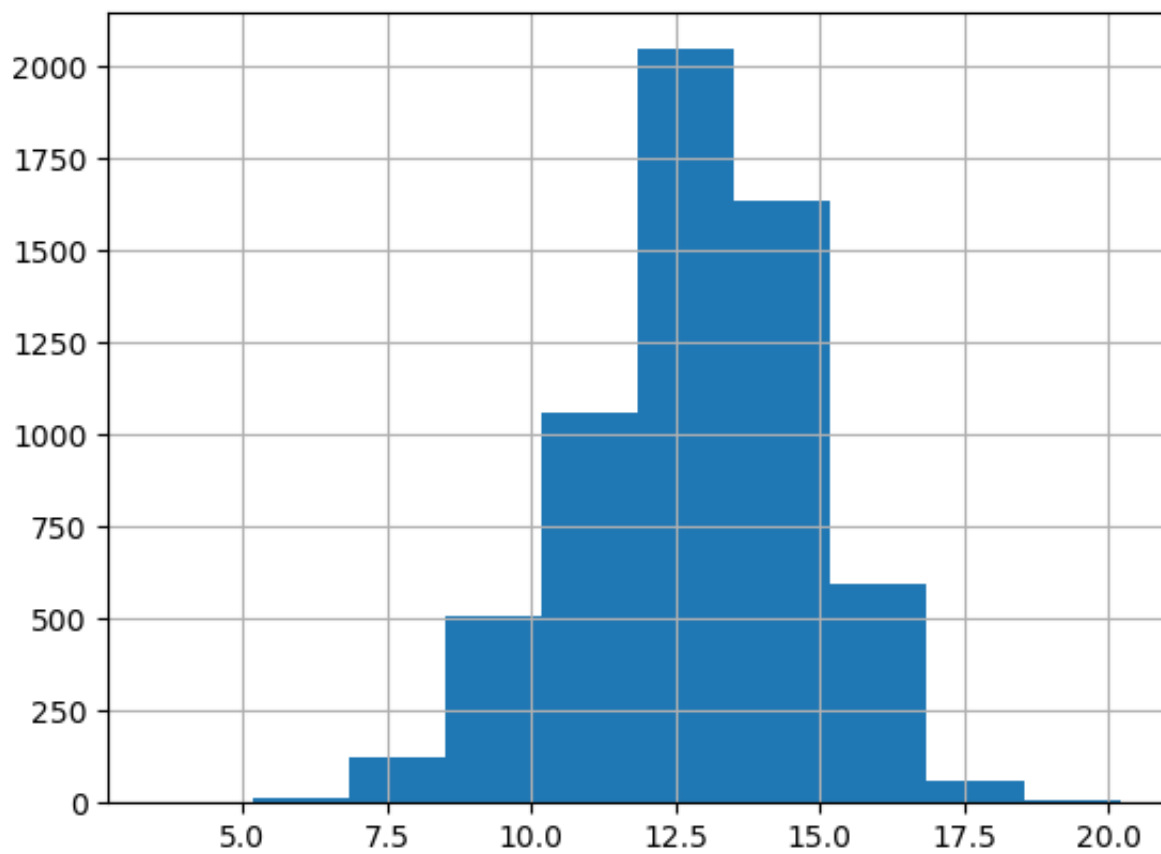
```
df["preop_hb"].plot.box() # 盒图及离群点
```

<Axes: >



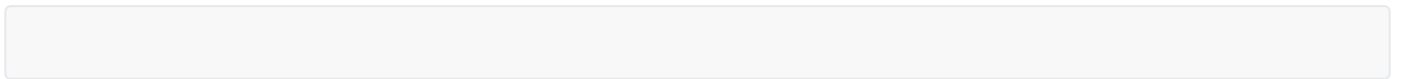
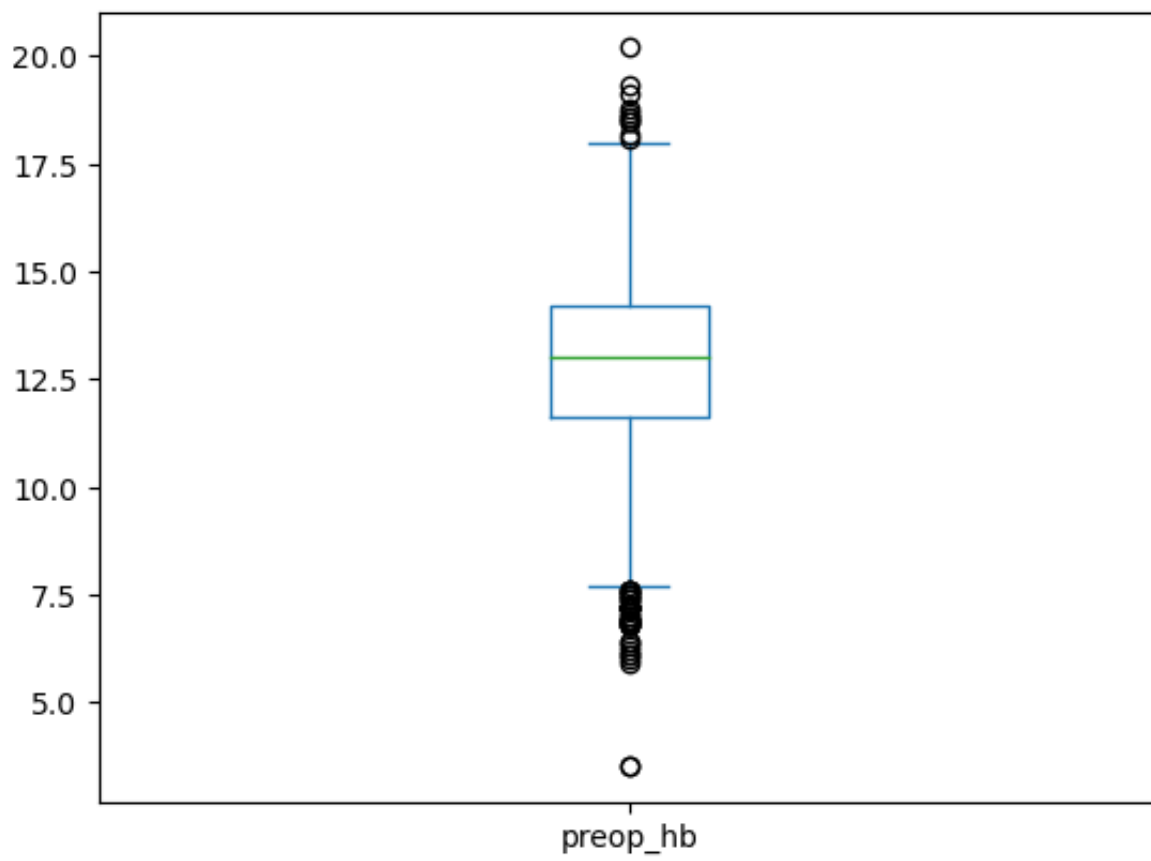
```
# 用最高频率值来填补缺失值
data_fillna=df["preop_hb"].fillna(df["preop_hb"].mode())
data_fillna.hist() # 直方图
```

<Axes: >



```
data_fillna.plot.box()
```

<Axes: >



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

```
path = "/Users/zhangpuchang/Downloads/movies_dataset.csv"
df = pd.read_csv(path, header=0)
```

数据摘要

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

```
Hollywood / English      14649
Bollywood / Indian       2645
Tollywood                 1172
Anime / Kids              1049
Wrestling                 433
Punjabi                   332
Stage shows               129
Pakistani                  92
Dub / Dual Audio          45
3D Movies                  1
Name: industry, dtype: int64
```

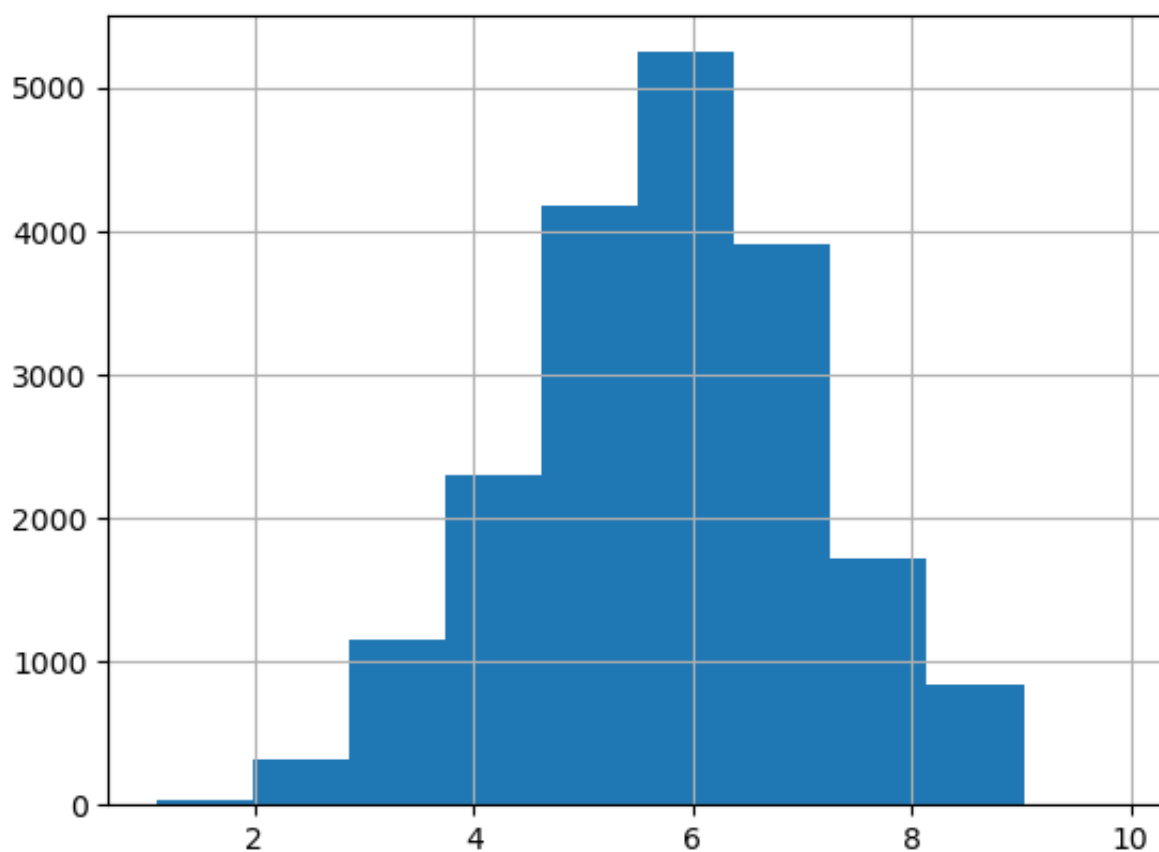
```
ranks = df["IMDb-rating"] # 数值属性 5数概括及缺失值的个数 例: IMDb-rating
null_cnt = ranks.isnull().sum()
ranks = ranks.dropna(axis = 0)
min_ranks = min(ranks)
max_ranks = max(ranks)
pct_25 = np.percentile(ranks, 25)
Median = np.median(ranks)
pct_75 = np.percentile(ranks, 75)
print("缺失值个数: {}".format(null_cnt))
print("最小值: {}".format(min_ranks))
print("Q1: {}".format(pct_25))
print("中位数: {}".format(Median))
print("Q3: {}".format(pct_75))
print("最大值: {}".format(max_ranks))
```


缺失值个数: 841
最小值: 1.1
Q1: 4.8
中位数: 5.7
Q3: 6.6
最大值: 9.9

数据可视化

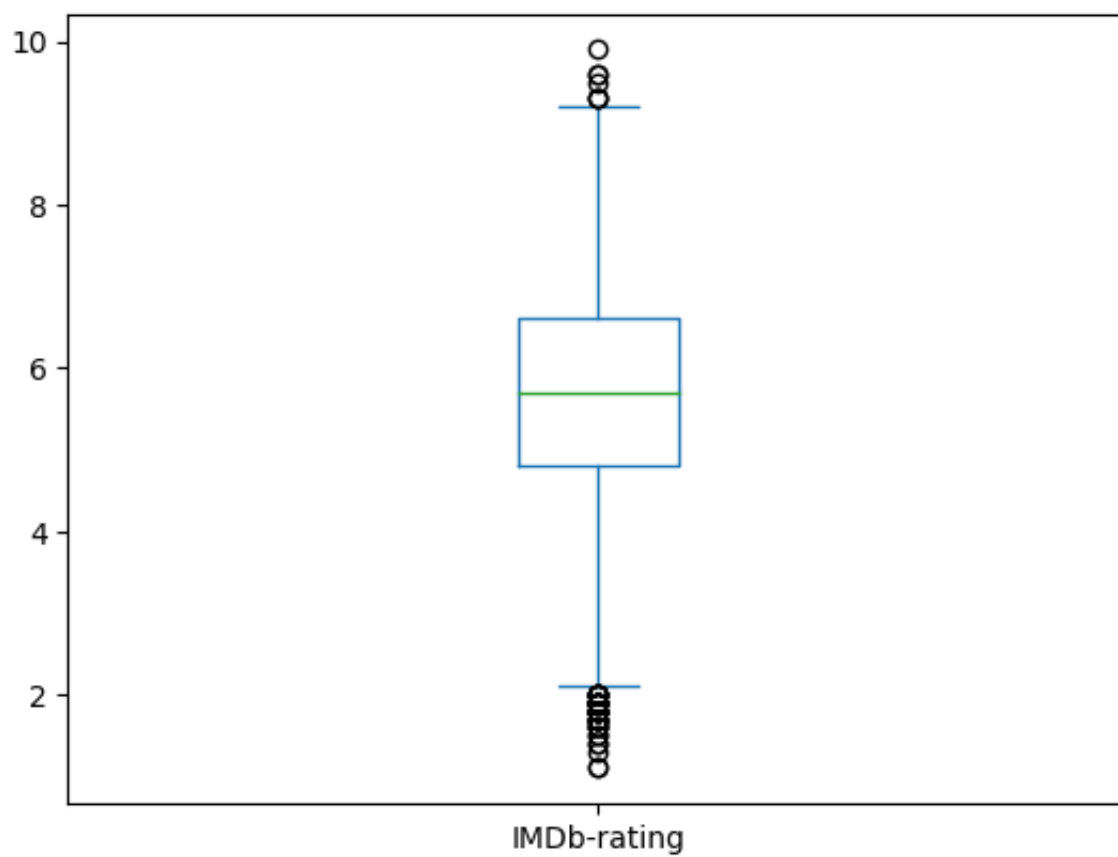
```
import matplotlib.pyplot as plt  
df["IMDb-rating"].hist() # 直方图
```

<Axes: >



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

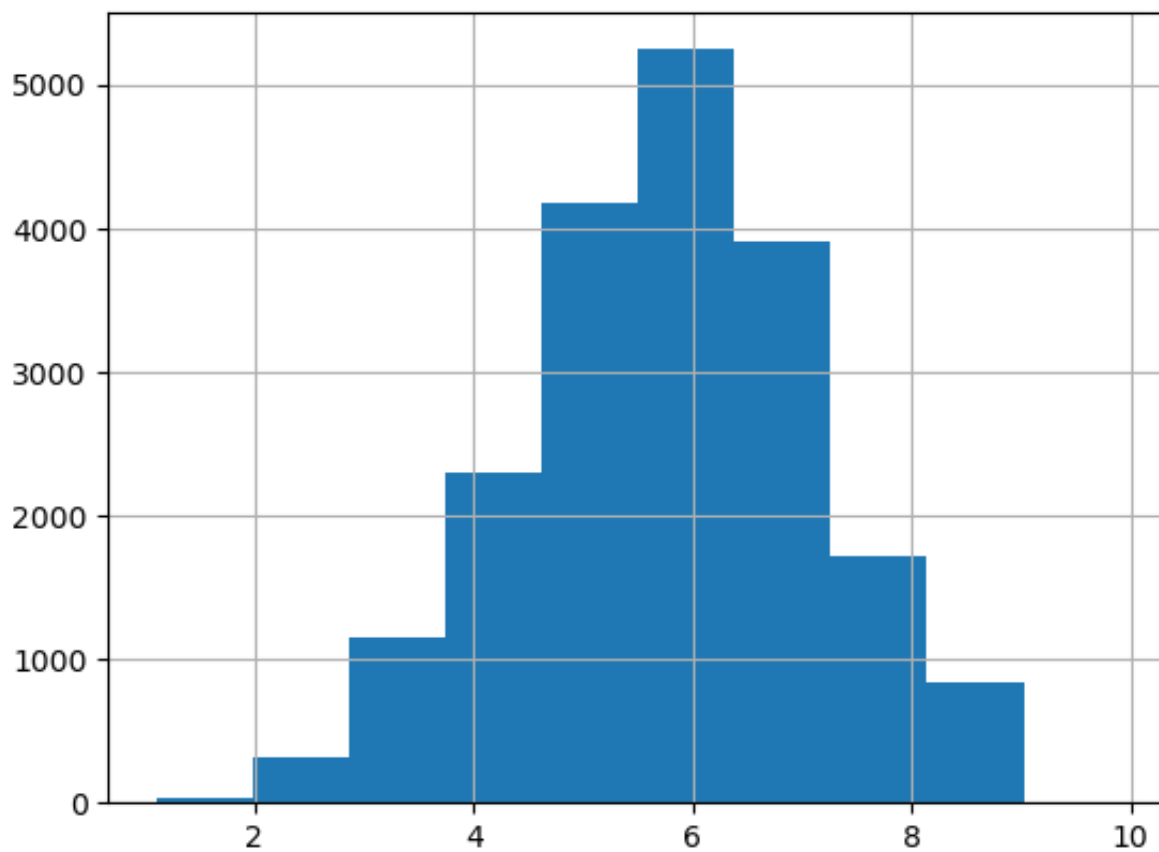
<Axes: >



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

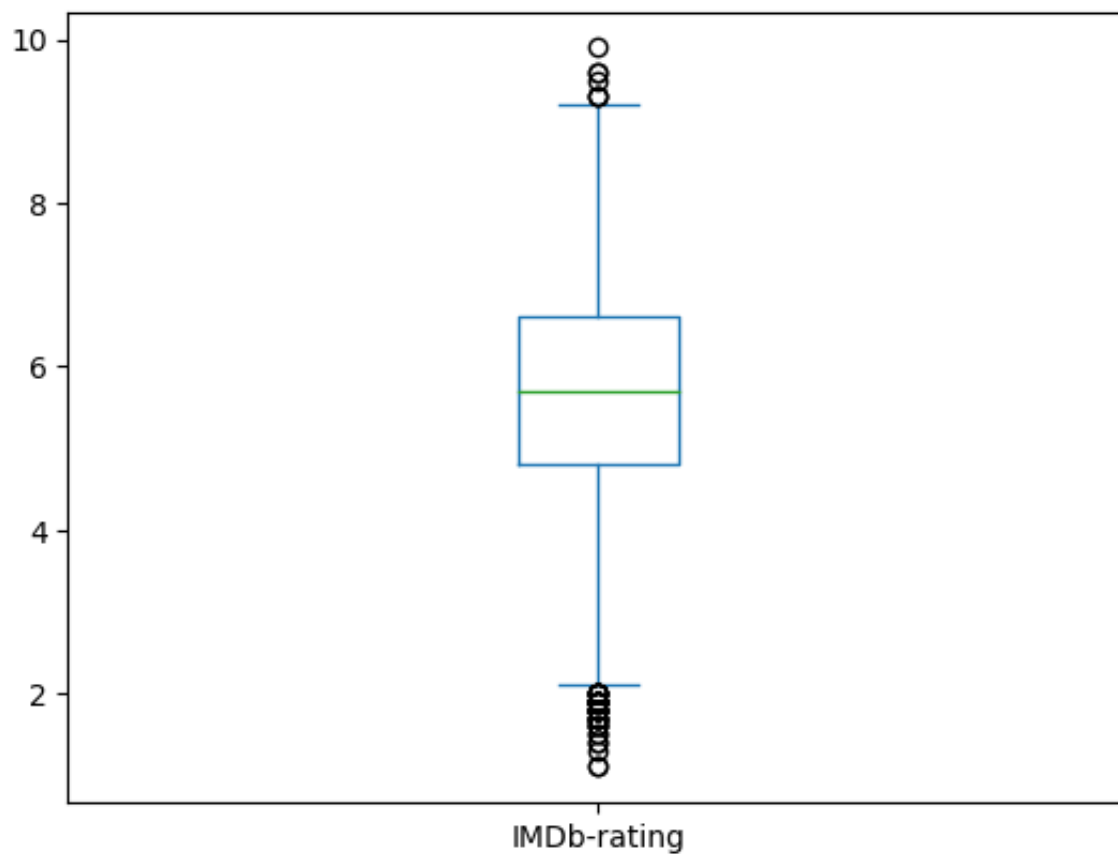
```
df["IMDb-rating"].hist() #直方图
```

<Axes: >



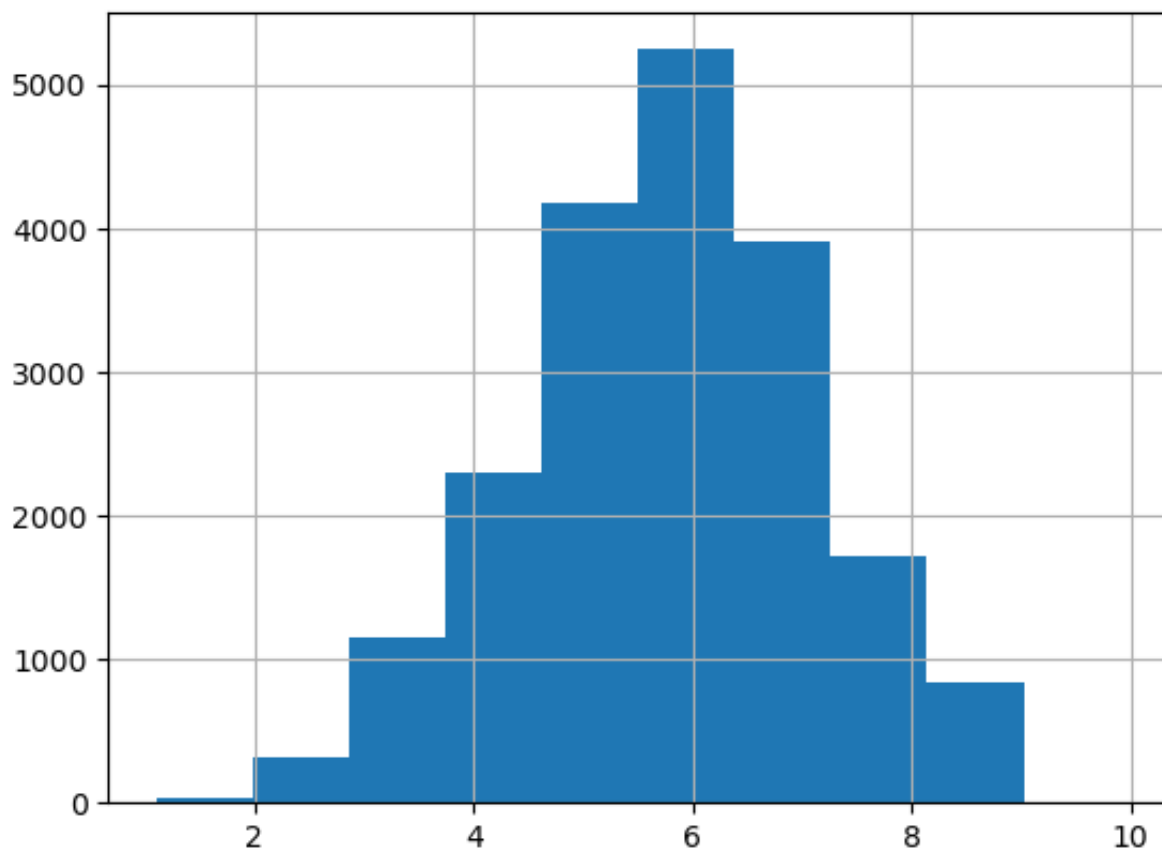
```
df["IMDb-rating"].plot.box() # 盒图及离群点
```

<Axes: >



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

<Axes: >



```
data_fillna.plot.box()
```

<Axes: >

