이상치 처리

- 이상치 :극단적인 값 또는 존재할 수 없는 값
- 이상치 처리

1) 행제거 또는 2) 값을 채워넣음(대표값 또는 예측값으로)

In [78]:

```
import pandas as pd
```

In [79]:

```
students = pd.read_csv("data/students.csv")
students
```

Out[79]:

	english	math	class
0	100	999	1
1	90	90	1
2	80	80	1
3	70	20	1
4	20	90	2
5	90	100	2
6	80	80	2
7	90	99	A

In [80]:

```
students.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8 entries, 0 to 7
Data columns (total 3 columns):
english    8 non-null int64
math       8 non-null int64
class      8 non-null object
dtypes: int64(2), object(1)
memory usage: 272.0+ bytes
```

1) 범주형 변수의 이상치 확인 예

1반과 2반만 존재하는 학교라고 예를들자. 1과 2외의 데이터는 이상치

In [81]:

```
# 클래스 열의 값이 1 또는 2인 행만
students[ students["class"].isin(['1','2'])  ]
```

Out[81]:

	english	math	class
0	100	999	1
1	90	90	1
2	80	80	1
3	70	20	1
4	20	90	2
5	90	100	2
6	80	80	2

In [82]:

```
# 클래스 열의 값이 1 또는 2가 아닌 행만 <--- 이상치
students[ ~students["class"].isin(['1','2'])  ]
```

Out[82]:

	english	math	class
7	90	99	A

이상치를 가진 행 삭제 예

In [83]:

```
students = students[ students["class"].isin(['1','2'])  ]
```

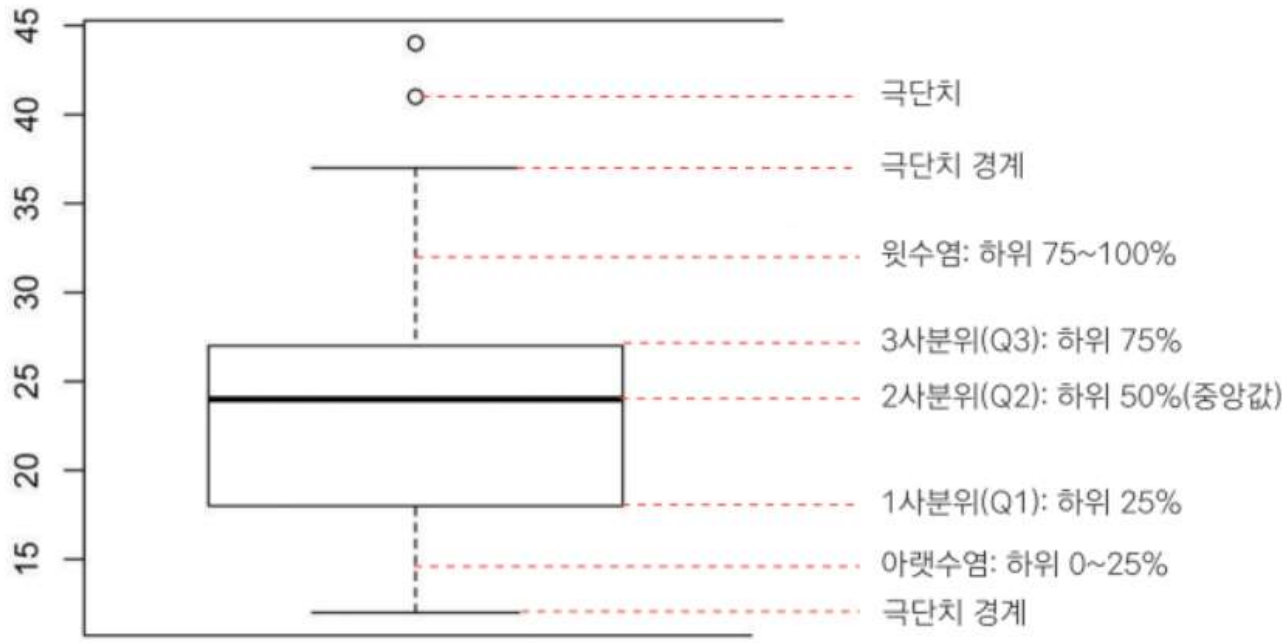
In [84]:

```
students
```

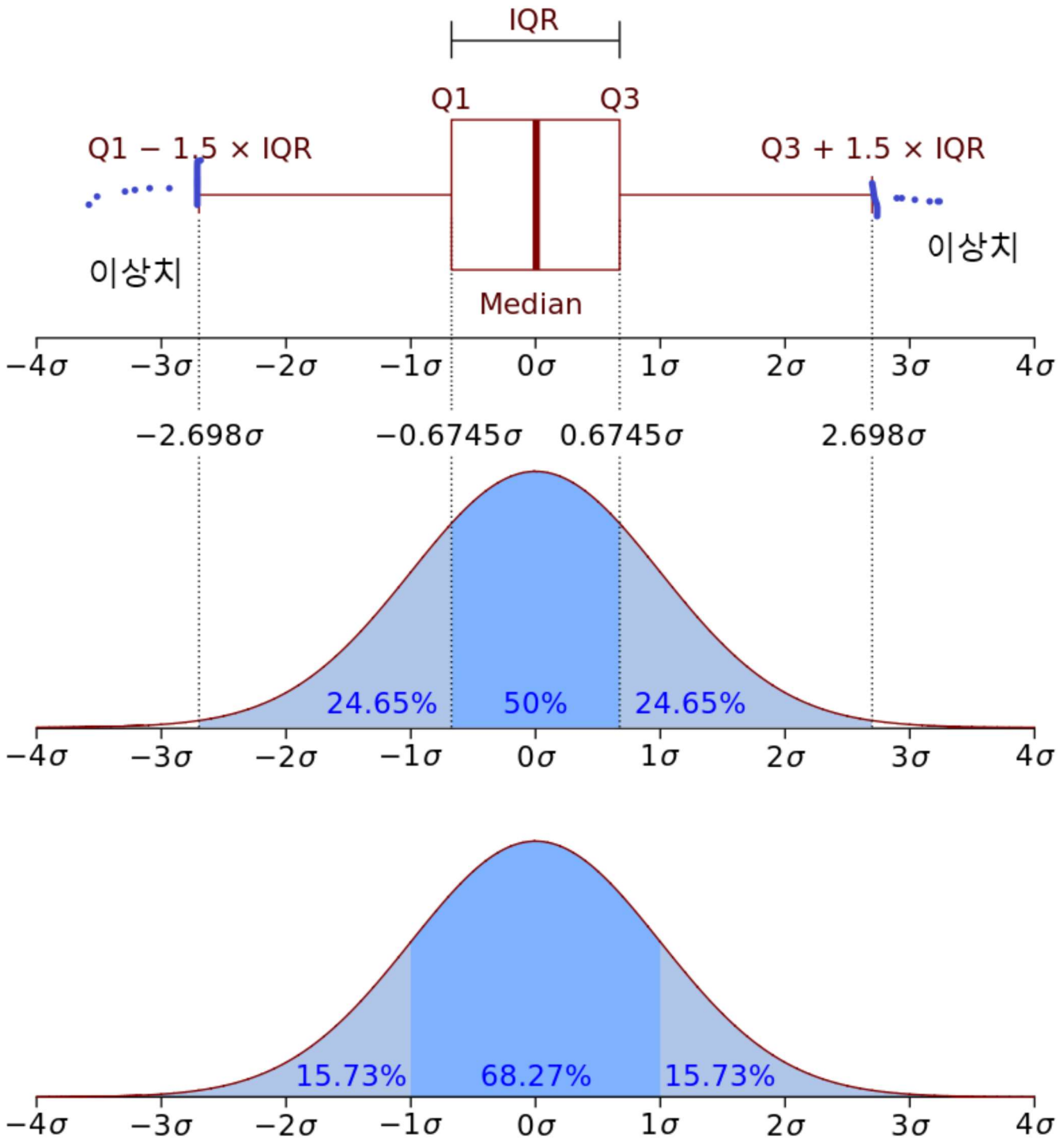
Out[84]:

	english	math	class
0	100	999	1
1	90	90	1
2	80	80	1
3	70	20	1
4	20	90	2
5	90	100	2
6	80	80	2

2) 연속형 변수의 이상치 확인 예



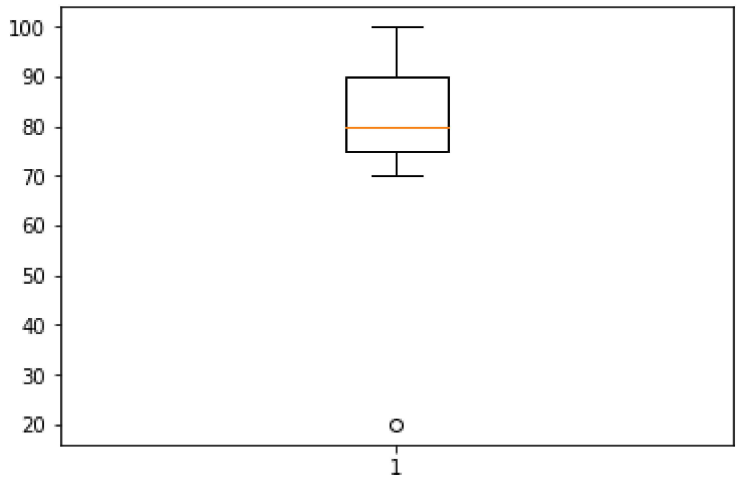
상자 그림	값	설명
상자 아래 세로 점선	아래수염	하위 0~25% 내에 해당하는 값
상자 밑면	1사분위수(Q1)	하위 25% 위치 값
상자 내 굵은 선	2사분위수(Q2)	하위 50% 위치 값(중앙값)
상자 윗면	3사분위수(Q3)	하위 75% 위치 값
상자 위 세로 점선	윗수염	하위 75~100% 내에 해당하는 값
상자 밖 가로선	극단치 경계	Q1, Q3 밖 1.5 IQR 내 최대값
상자 밖 점 표식	극단치	Q1, Q3 밖 1.5 IQR을 벗어난 값


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
이상치
이상치
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

In [85]:

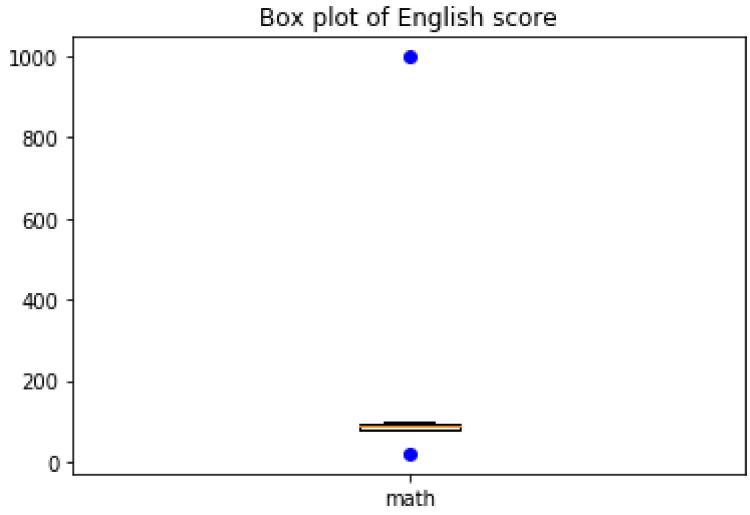
```
# Basic box plot
import matplotlib.pyplot as plt
%matplotlib inline


a =plt.boxplot(students['english'])
plt.show()
```



In [86]:

```
plt.boxplot(students['math'], sym="bo")
plt.title('Box plot of English score')
plt.xticks([1], ['math'])
plt.show()
```



연속적 변수 이상치 구하기 예

In [92]:

```
import numpy as np
Q1 = np.percentile(students["math"], 25)
Q3 = np.percentile(students["math"], 75)
IQR = Q3 - Q1
outlier_step = 1.5 * IQR

outlier_step
```

Out[92]:

22.5

In [95]:

```
# 연속적 변수 이상치 출력
students[(students["math"] < Q1 - outlier_step) | (students["math"] > Q3 + outlier_step)]
```

Out[95]:

	english	math	class
0	100	999	1
3	70	20	1