7장. 데이터 전처리			
학번	32233421	이름	이은지

※ airquailty.csv 는 대기 오염에 관한 데이터를 저장하고 있다. 읽어서 df에 저장한 뒤 물음에 답하시오 (1~7)

1. 컬럼별 결측값의 개수를 보이시오.

```
>>> pd.isna(df).sum()
Ozone      37
Solar.R     7
Wind        0
Temp        0
Month       0
Day         0
dtype: int64
```

2. 결측값이 있는 행을 제거한 후 df2에 저장하시오. df2.head()의 결과를 보이시오

```
>>> df2=df.dropna()
>>> df2.head()
   Ozone  Solar.R  Wind  Temp  Month  Day
0   41.0    190.0   7.4    67      5    1
1   36.0    118.0   8.0    72      5    2
2   12.0    149.0  12.6    74      5    3
3   18.0    313.0  11.5    62      5    4
6   23.0    299.0   8.6    65      5    7
```

3. 결측값을 추정하여 df3에 저장하시오. df3.head()의 결과를 보이시오.

```
>>> df2=df.iloc[:, :].to_numpy()
>>> imputer=KNNImputer(n_neighbors=5)
>>> df2=imputer.fit_transform(df2)
>>> df.iloc[:,:]=df2
>>> df3=df.iloc[:,:]
>>> df3.head()
   Ozone  Solar.R  Wind  Temp  Month  Day
0   41.0    190.0   7.4    67      5    1
1   36.0    118.0   8.0    72      5    2
2   12.0    149.0  12.6    74      5    3
3   18.0    313.0  11.5    62      5    4
4   18.2    159.0  14.3    56      5    5
```

4. Wind 컬럼의 특이값(Z-score의 절대값이 2보다 큰 경우)을 보이시오

```
>>> print(outliers)
8      20.1
17     18.4
47     20.7
52      1.7
120     2.3
125     2.8
```

5. Temp 컬럼의 값을 기준으로 내림차순으로 정렬하여 df4에 저장하시오.

df4.hed()의 결과를 보이시오

```
>>> df4=df.sort_values('Temp', ascending=False)
>>> df4.head()
     Ozone  Solar.R  Wind  Temp  Month  Day
119   76.0    203.0   9.7    97      8   28
121   84.0    237.0   6.3    96      8   30
120  118.0    225.0   2.3    94      8   29
122   85.0    188.0   6.3    94      8   31
41     NaN    259.0  10.9    93      6   11
```

6. df에서 10개의 행을 표본 추출하여 df5에 저장하시오 df5의 내용을 보이시오

```
>>> df5=df.sample(n=10, random_state=123)
>>> df5
     Ozone  Solar.R  Wind  Temp  Month  Day
41     NaN    259.0  10.9    93      6   11
114    NaN    255.0  12.6    75      8   23
130   23.0    220.0  10.3    78      9    8
24     NaN     66.0  16.6    57      5   25
87    52.0     82.0  12.0    86      7   27
79    79.0    187.0   5.1    87      7   19
124   78.0    197.0   5.1    92      9    2
128   32.0     92.0  15.5    84      9    6
125   73.0    183.0   2.8    93      9    3
42     NaN    250.0   9.2    92      6   12
```

7. 결측값이 없는 df2에 대해 월별로 Wind, Temp 컬럼의 최대값을 집계하여

보이시오

```
>>> df3=df2.groupby('Month')[['Wind','Temp']].max()
>>> df3
       Wind  Temp
Month
5      20.1    81
6      20.7    90
7      14.9    92
8      15.5    97
9      16.6    93
```

8. mtcats.csv 데이터셋에서 gear 와 carb 별로 평균연비(mpg)를 집계하여 보이시오. (피벗 테이블 이용)

```
>>> df3=df.pivot_table(index='gear', columns='carb',
...                    values='mpg', aggfunc='mean')
>>>
>>> df3
carb          1      2     3      4     6     8
gear
3     20.333333  17.15  16.3  12.62   NaN   NaN
4     29.100000  24.75   NaN  19.75   NaN   NaN
5           NaN  28.20   NaN  15.80  19.7  15.0
```