In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [4]:
```
df=pd.read_csv(r"D:\Stranger Things\New folder\you_ex_data seience\EXCEL\EDA\EDA
```

In [5]:
```
df
```

Out[5]:

	Ozone	Solar.R	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5	1	2010	67	S
1	36.0	118.0	8.0	5	2	2010	72	C
2	12.0	149.0	12.6	5	3	2010	74	PS
3	18.0	313.0	11.5	5	4	2010	62	S
4	NaN	NaN	14.3	5	5	2010	56	S
...	...	...	...	...	...	...	...	...
153	41.0	190.0	7.4	5	1	2010	67	C
154	30.0	193.0	6.9	9	26	2010	70	PS
155	NaN	145.0	13.2	9	27	2010	77	S
156	14.0	191.0	14.3	9	28	2010	75	S
157	18.0	131.0	8.0	9	29	2010	76	C

158 rows × 8 columns

In [6]:
```
df.head(2)
```

Out[6]:

	Ozone	Solar.R	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5	1	2010	67	S
1	36.0	118.0	8.0	5	2	2010	72	C

In [7]:
```
df.shape
```

Out[7]:
```
(158, 8)
```

In [8]:
```
df.columns
```

Out[8]:
```
Index(['Ozone', 'Solar.R', 'Wind', 'Month', 'Day', 'Year', 'Temp', 'Weather'],
dtype='object')
```

In [9]:
```
df.dtypes
```

Out[9]:
```
Ozone      float64
Solar.R    float64
Wind       float64
Month       object
Day          int64
Year         int64
Temp         int64
Weather     object
dtype: object
```

In [10]:
```
df.describe()
```

Out[10]:

	Ozone	Solar.R	Wind	Day	Year	Temp
count	120.000000	151.000000	158.000000	158.000000	158.0	158.000000
mean	41.583333	185.403974	9.957595	16.006329	2010.0	77.727848
std	32.620709	88.723103	3.511261	8.997166	0.0	9.377877
min	1.000000	7.000000	1.700000	1.000000	2010.0	56.000000
25%	18.000000	119.000000	7.400000	8.000000	2010.0	72.000000
50%	30.500000	197.000000	9.700000	16.000000	2010.0	78.500000
75%	61.500000	257.000000	11.875000	24.000000	2010.0	84.000000
max	168.000000	334.000000	20.700000	31.000000	2010.0	97.000000

In [11]:
```
df.describe(include='object')
```

Out[11]:

	Month	Weather
count	158	155
unique	6	3
top	9	S
freq	34	59

In [12]:
```
df['Month']
```

Out[12]:
```
0      5
1      5
2      5
3      5
4      5
      ..
153    5
154    9
155    9
156    9
157    9
Name: Month, Length: 158, dtype: object
```

In [13]:
```
df['Month'].unique()
```

Out[13]:
```
array(['5', 'May', '6', '7', '8', '9'], dtype=object)
```

In [14]:
```
df['Month'].nunique()
```

Out[14]:
```
6
```

In [15]:
```
df['Month'].value_counts()
```

Out[15]:
```
Month
9      34
5      31
7      31
8      31
6      30
May     1
Name: count, dtype: int64
```

In [16]:
```
df['Month'].replace('May','5',inplace=True) #replace only value not datatype.
```

In [17]:
```
df['Month']
```

Out[17]:
```
0      5
1      5
2      5
3      5
4      5
      ..
153    5
154    9
155    9
156    9
157    9
Name: Month, Length: 158, dtype: object
```

In [18]:
```
df['Month'].unique()
```

Out[18]:
```
array(['5', '6', '7', '8', '9'], dtype=object)
```

In [19]:
```
df['Month'].astype('int64')
```

Out[19]:
```
0      5
1      5
2      5
3      5
4      5
      ..
153    5
154    9
155    9
156    9
157    9
Name: Month, Length: 158, dtype: int64
```

In [20]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 8 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Ozone    120 non-null    float64
 1   Solar.R  151 non-null    float64
 2   Wind     158 non-null    float64
 3   Month    158 non-null    object 
 4   Day      158 non-null    int64  
 5   Year     158 non-null    int64  
 6   Temp     158 non-null    int64  
 7   Weather  155 non-null    object 
dtypes: float64(3), int64(3), object(2)
memory usage: 10.0+ KB
```

In [21]:
```
df['Month']=df['Month'].astype('int64')
```

In [22]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 8 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Ozone    120 non-null    float64
 1   Solar.R  151 non-null    float64
 2   Wind     158 non-null    float64
 3   Month    158 non-null    int64  
 4   Day      158 non-null    int64  
 5   Year     158 non-null    int64  
 6   Temp     158 non-null    int64  
 7   Weather  155 non-null    object 
dtypes: float64(3), int64(4), object(1)
memory usage: 10.0+ KB
```

In [23]:
```
df.duplicated()
```

Out[23]:
```
0      False
1      False
2      False
3      False
4      False
       ...
153    False
154    False
155    False
156     True
157    False
Length: 158, dtype: bool
```

In [24]:
```
df.duplicated().sum()
```

Out[24]:
```
1
```

In [25]:
```
df[df.duplicated()]
```

Out[25]:

	Ozone	Solar.R	Wind	Month	Day	Year	Temp	Weather
156	14.0	191.0	14.3	9	28	2010	75	S

In [26]:
```
df.drop_duplicates(inplace=True)
```

In [27]:
```
df.shape
```

Out[27]:
```
(157, 8)
```

In [28]:
```
df.duplicated().sum()
```

Out[28]:
```
0
```

In [29]:
```
df1=df.copy()
```

In [30]:
```
df1.head()
```

Out[30]:

	Ozone	Solar.R	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5	1	2010	67	S
1	36.0	118.0	8.0	5	2	2010	72	C
2	12.0	149.0	12.6	5	3	2010	74	PS
3	18.0	313.0	11.5	5	4	2010	62	S
4	NaN	NaN	14.3	5	5	2010	56	S

In [31]:
```
df1.drop('Year',axis=1,inplace=True)
```

In [32]:
```
df1.columns
```

Out[32]:
```
Index(['Ozone', 'Solar.R', 'Wind', 'Month', 'Day', 'Temp', 'Weather'], dtype='o
bject')
```

In [33]:
```
df1.drop(['Month','Ozone'],axis=1)
```

Out[33]:

	Solar.R	Wind	Day	Temp	Weather
0	190.0	7.4	1	67	S
1	118.0	8.0	2	72	C
2	149.0	12.6	3	74	PS
3	313.0	11.5	4	62	S
4	NaN	14.3	5	56	S
...	...	...	...	...	...
152	223.0	11.5	30	68	S
153	190.0	7.4	1	67	C
154	193.0	6.9	26	70	PS
155	145.0	13.2	27	77	S
157	131.0	8.0	29	76	C

157 rows × 5 columns

In [34]:
```
df1.head()
```

Out[34]:

	Ozone	Solar.R	Wind	Month	Day	Temp	Weather
0	41.0	190.0	7.4	5	1	67	S
1	36.0	118.0	8.0	5	2	72	C
2	12.0	149.0	12.6	5	3	74	PS
3	18.0	313.0	11.5	5	4	62	S
4	NaN	NaN	14.3	5	5	56	S

In [35]:
```
df.columns
```

Out[35]:
```
Index(['Ozone', 'Solar.R', 'Wind', 'Month', 'Day', 'Year', 'Temp', 'Weather'],
dtype='object')
```

In [36]:
```
df1.rename({'Temp':'Te'},inplace=True,axis=1)
```

In [37]:
```
df1.head()
```

Out[37]:

	Ozone	Solar.R	Wind	Month	Day	Te	Weather
0	41.0	190.0	7.4	5	1	67	S
1	36.0	118.0	8.0	5	2	72	C
2	12.0	149.0	12.6	5	3	74	PS
3	18.0	313.0	11.5	5	4	62	S
4	NaN	NaN	14.3	5	5	56	S

In [38]:
```
df1.rename({'Solar.R':'So.R'},inplace=True,axis=1)
```

In [39]:
```
df1.head()
```

Out[39]:

	Ozone	So.R	Wind	Month	Day	Te	Weather
0	41.0	190.0	7.4	5	1	67	S
1	36.0	118.0	8.0	5	2	72	C
2	12.0	149.0	12.6	5	3	74	PS
3	18.0	313.0	11.5	5	4	62	S
4	NaN	NaN	14.3	5	5	56	S

In [41]:
```
df1.rename({'Te':'Temperature '},inplace=True,axis=1)
```

In [42]:
```
df1.head()
```

Out[42]:

	Ozone	So.R	Wind	Month	Day	Temperature	Weather
0	41.0	190.0	7.4	5	1	67	S
1	36.0	118.0	8.0	5	2	72	C
2	12.0	149.0	12.6	5	3	74	PS
3	18.0	313.0	11.5	5	4	62	S
4	NaN	NaN	14.3	5	5	56	S

In [43]:
```
df1.isnull()
```

Out[43]:

	Ozone	So.R	Wind	Month	Day	Temperature	Weather
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	True	True	False	False	False	False	False
...	...	...	...	...	...	...	...
152	False	False	False	False	False	False	False
153	False	False	False	False	False	False	False
154	False	False	False	False	False	False	False
155	True	False	False	False	False	False	False
157	False	False	False	False	False	False	False

157 rows × 7 columns

In [44]:
```
df1.isnull().sum()
```

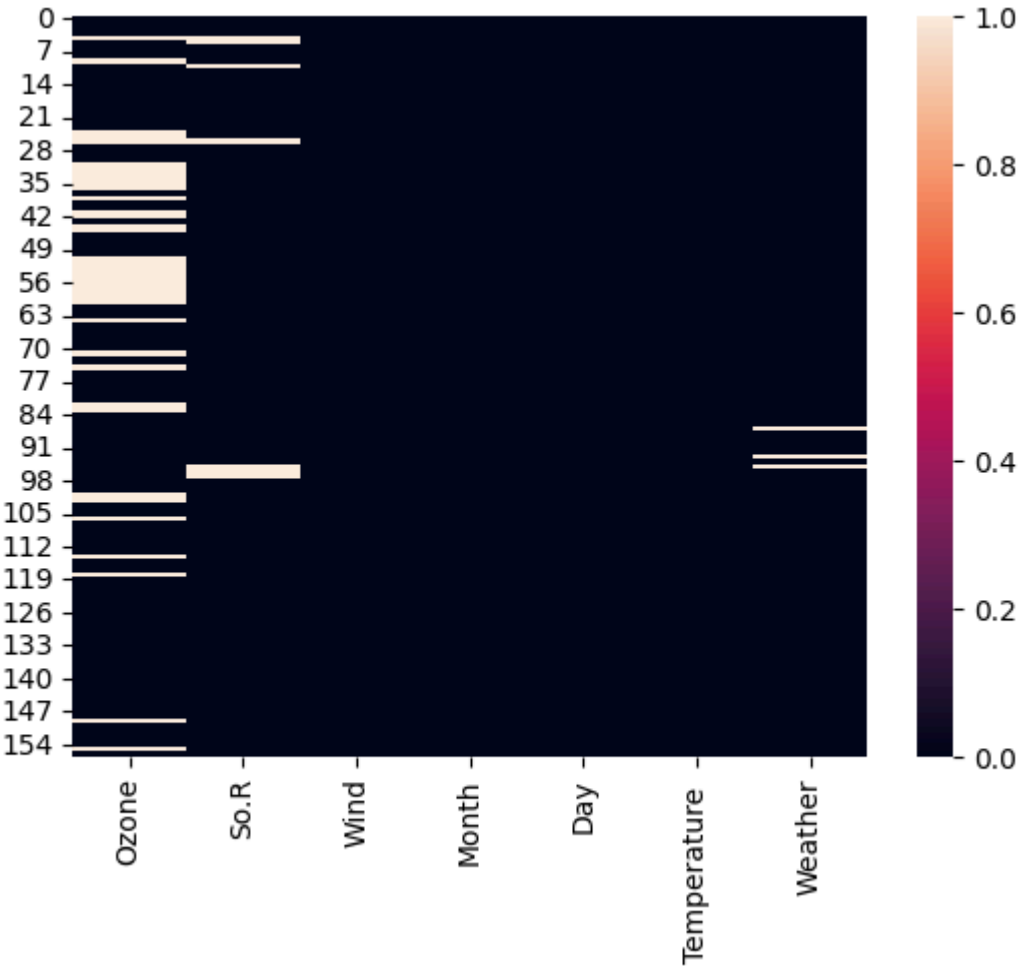
Out[44]:
```
Ozone           38
So.R             7
Wind             0
Month            0
Day              0
Temperature      0
Weather          3
dtype: int64
```

In [45]:
```
df.isna().sum()
```

Out[45]:
```
Ozone      38
Solar.R     7
Wind        0
Month       0
Day         0
Year        0
Temp        0
Weather     3
dtype: int64
```

In [46]:
```
sns.heatmap(df1.isnull())
```

Out[46]: <Axes: >



In [47]:
```
len(df1)
```

Out[47]: 157

In [48]:
```
for i in (df1.isnull()).sum():
  print(i/len(df1)*100)
```

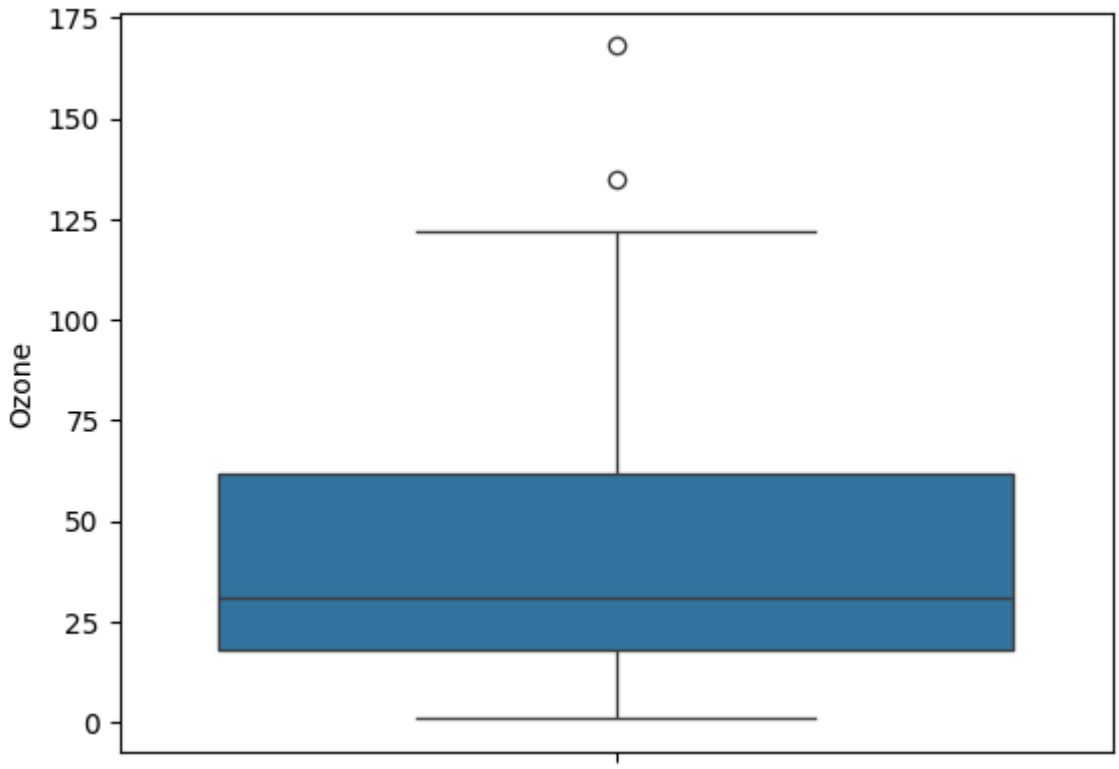
```
24.203821656050955
4.45859872611465
0.0
0.0
0.0
0.0
1.910828025477707
```

In [49]:
```
df1['Ozone'].max()
```

Out[49]: 168.0

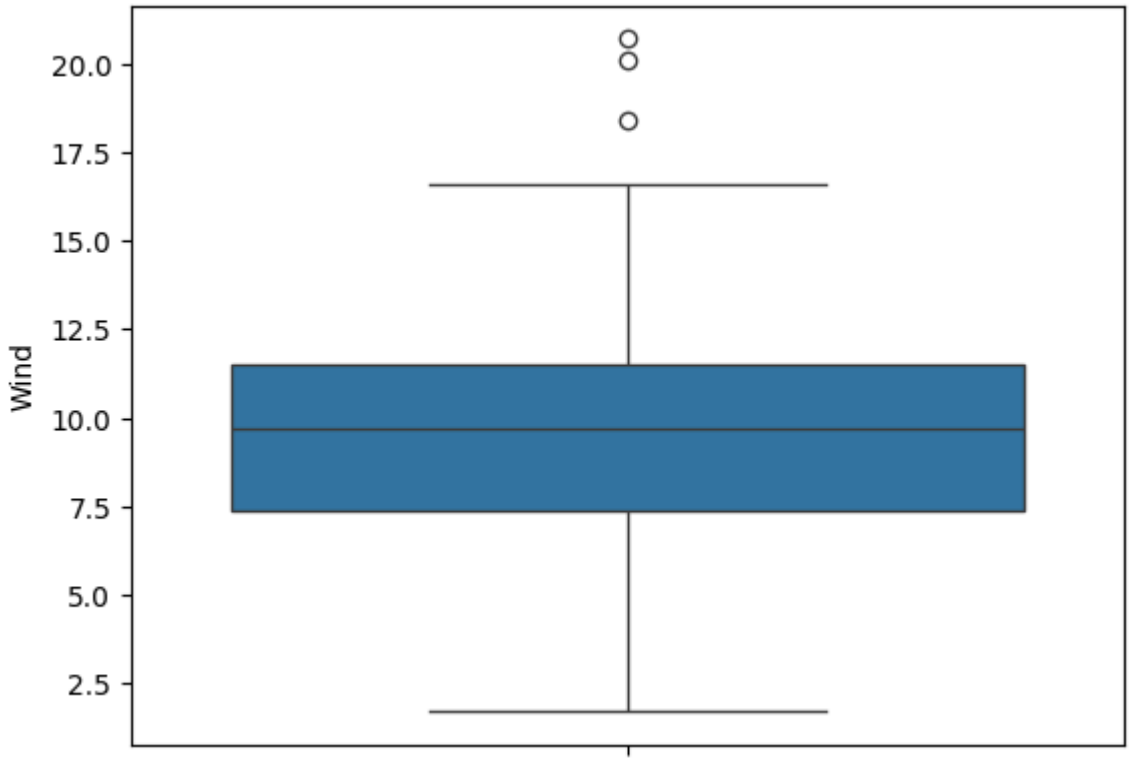
In [50]:
```
sns.boxplot(df1['Ozone'])
```

Out[50]: <Axes: ylabel='Ozone'>



In [51]:
```
sns.boxplot(df1['Wind'])
```

Out[51]: <Axes: ylabel='Wind'>



In [52]:
```
df1_mean=df1['Ozone'].mean()
```

In [53]:
```
df1_mean
```

Out[53]:
```
41.81512605042017
```

In [54]:
```
df1_median=df1['Ozone'].median()
```

In [55]:
```
df1_median
```

Out[55]:
```
31.0
```

In [56]:
```
df1['Ozone'].fillna(df1_median,inplace=True)
```

In [57]:
```
df1['So.R'].fillna(df1['So.R'].mean(),inplace=True)
```

In [58]:
```
df1.isnull().sum()
```

Out[58]:
```
Ozone          0
So.R           0
Wind           0
Month          0
Day            0
Temperature    0
Weather        3
dtype: int64
```

In [59]:
```
df1['Weather'].fillna(df1['Weather'].mode()[0],inplace=True)
```

In [60]:
```
df1.isnull().sum()
```

Out[60]:
```
Ozone          0
So.R           0
Wind           0
Month          0
Day            0
Temperature    0
Weather        0
dtype: int64
```

In [61]:
```
df1['Weather'].value_counts()
```

Out[61]:
```
Weather
S     61
C     49
PS    47
Name: count, dtype: int64
```

In [62]:
```
df1
```

Out[62]:

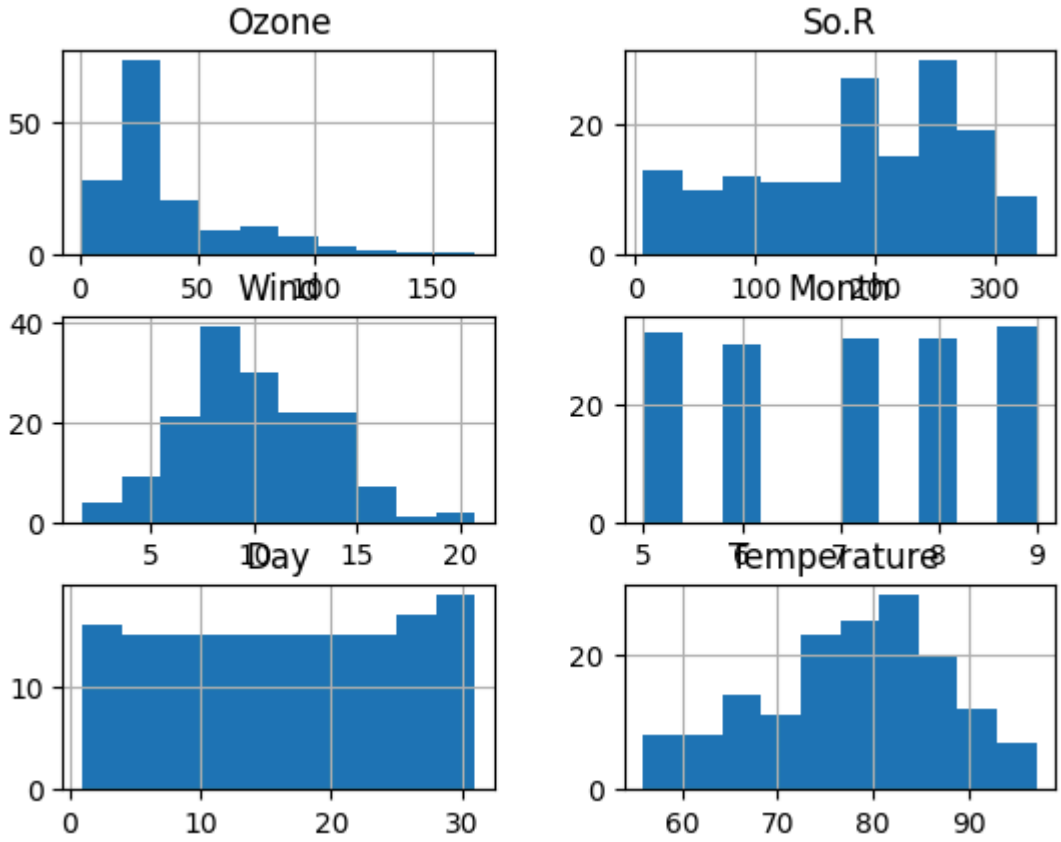
	Ozone	So.R	Wind	Month	Day	Temperature	Weather
0	41.0	190.000000	7.4	5	1	67	S
1	36.0	118.000000	8.0	5	2	72	C
2	12.0	149.000000	12.6	5	3	74	PS
3	18.0	313.000000	11.5	5	4	62	S
4	31.0	185.366667	14.3	5	5	56	S
...	...	...	...	...	...	...	...
152	20.0	223.000000	11.5	9	30	68	S
153	41.0	190.000000	7.4	5	1	67	C
154	30.0	193.000000	6.9	9	26	70	PS
155	31.0	145.000000	13.2	9	27	77	S
157	18.0	131.000000	8.0	9	29	76	C

157 rows × 7 columns

Outlier Detection

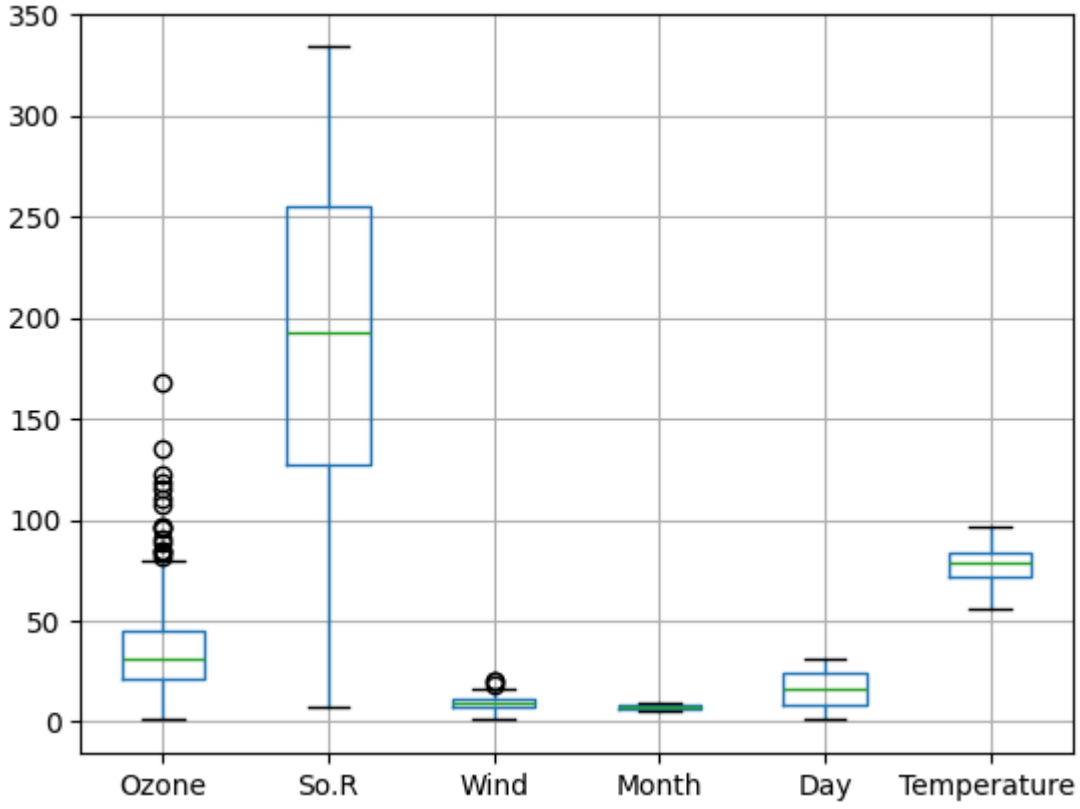
In [63]:
```
df1.hist()
```

Out[63]:
```
array([[<Axes: title={'center': 'Ozone'}>,
        <Axes: title={'center': 'So.R'}>],
       [<Axes: title={'center': 'Wind'}>,
        <Axes: title={'center': 'Month'}>],
       [<Axes: title={'center': 'Day'}>,
        <Axes: title={'center': 'Temperature '}>]], dtype=object)
```



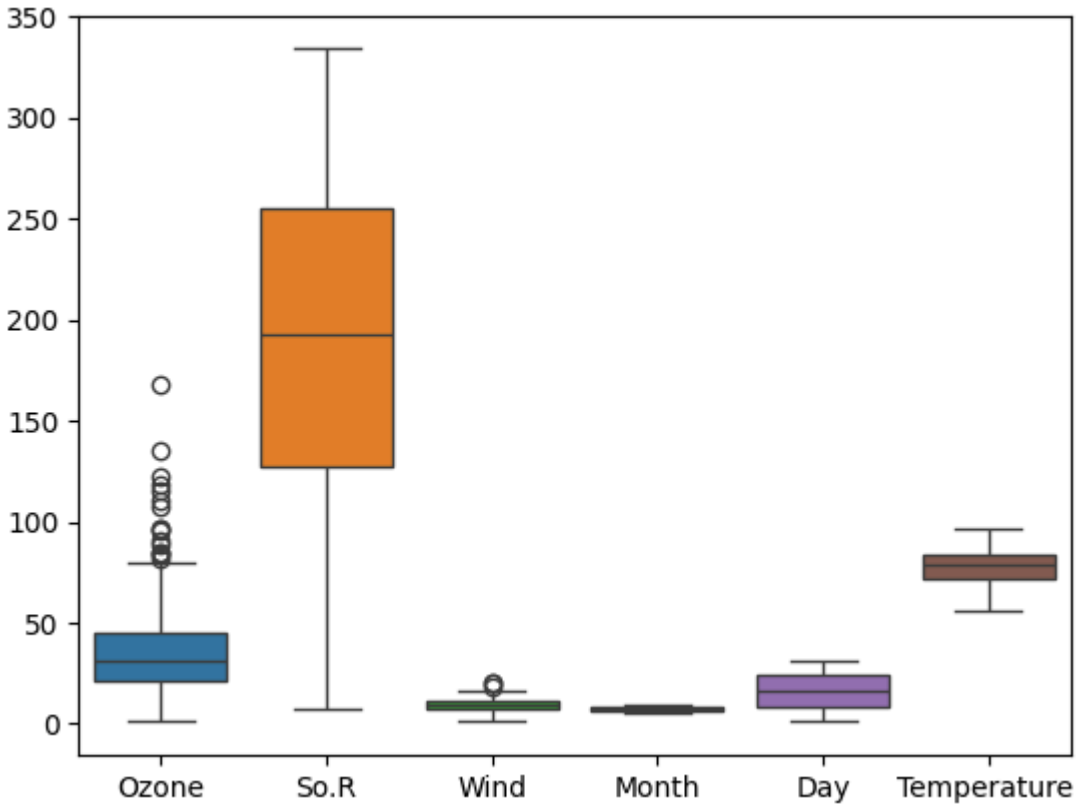
In [64]:
```
df1.boxplot()
```

Out[64]: <Axes: >



In [65]:
```
sns.boxplot(df1)
```

Out[65]: <Axes: >



In [68]:
```
df1.describe()
```

Out[68]:

	Ozone	So.R	Wind	Month	Day	Temperature
count	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000
mean	39.197452	185.366667	9.929936	7.019108	15.929936	77.745223
std	28.781992	86.998999	3.505188	1.434338	8.974404	9.405334
min	1.000000	7.000000	1.700000	5.000000	1.000000	56.000000
25%	21.000000	127.000000	7.400000	6.000000	8.000000	72.000000
50%	31.000000	193.000000	9.700000	7.000000	16.000000	79.000000
75%	45.000000	255.000000	11.500000	8.000000	24.000000	84.000000
max	168.000000	334.000000	20.700000	9.000000	31.000000	97.000000

In [67]:
```
df1['Ozone'].quantile(0.25)
```

Out[67]:
```
21.0
```

In [82]:
```
def out_dec(data,colu):
    q1=data[colu].quantile(0.25)
    q3=data[colu].quantile(0.75)
    iqr=q3-q1

    upper_extreme=q3+(1.5*iqr)
    lower_extreme=q1-(1.5*iqr)

    return lower_extreme,upper_extreme
```

In [83]:
```
out_dec(df1,'Ozone')
```

Out[83]: (-15.0, 81.0)

In [85]:
```
df1[df1['Ozone']>81]
```

Out[85]:

	Ozone	So.R	Wind	Month	Day	Temperature	Weather
29	115.0	223.0	5.7	5	30	79	C
61	135.0	269.0	4.1	7	1	84	S
68	97.0	267.0	6.3	7	8	92	PS
69	97.0	272.0	5.7	7	9	92	C
70	85.0	175.0	7.4	7	10	89	PS
85	108.0	223.0	8.0	7	25	85	PS
88	82.0	213.0	7.4	7	28	88	S
98	122.0	255.0	4.0	8	7	89	C
99	89.0	229.0	10.3	8	8	90	PS
100	110.0	207.0	8.0	8	9	90	C
116	168.0	238.0	3.4	8	25	81	PS
120	118.0	225.0	2.3	8	29	94	S
121	84.0	237.0	6.3	8	30	96	S
122	85.0	188.0	6.3	8	31	94	C
123	96.0	167.0	6.9	9	1	91	C
126	91.0	189.0	4.6	9	4	93	PS

In [87]:
```
df1[df1['Ozone']>81].shape[0]
```

Out[87]: 16

In [92]:
```
df1.loc[df1['Ozone']>81.0,'Ozone']=81.0
```

In [93]:
```
df1[df1['Ozone']>81].shape[0]
```

Out[93]: 0

In [94]:
```
df1[df1['Ozone']==81].shape[0]
```

Out[94]: 16

In [96]:
```
out_dec(df1,'Wind')
```

Out[96]: (1.2500000000000009, 17.65)

In [97]:
```
df1[df1['Wind']>17.65].shape[0]
```

Out[97]: 3

In [99]:
```
df1.loc[df1['Wind']>17.65,'Wind']=17.65
```

In [100…
```
df1[df1['Wind']>17.65].shape[0]
```

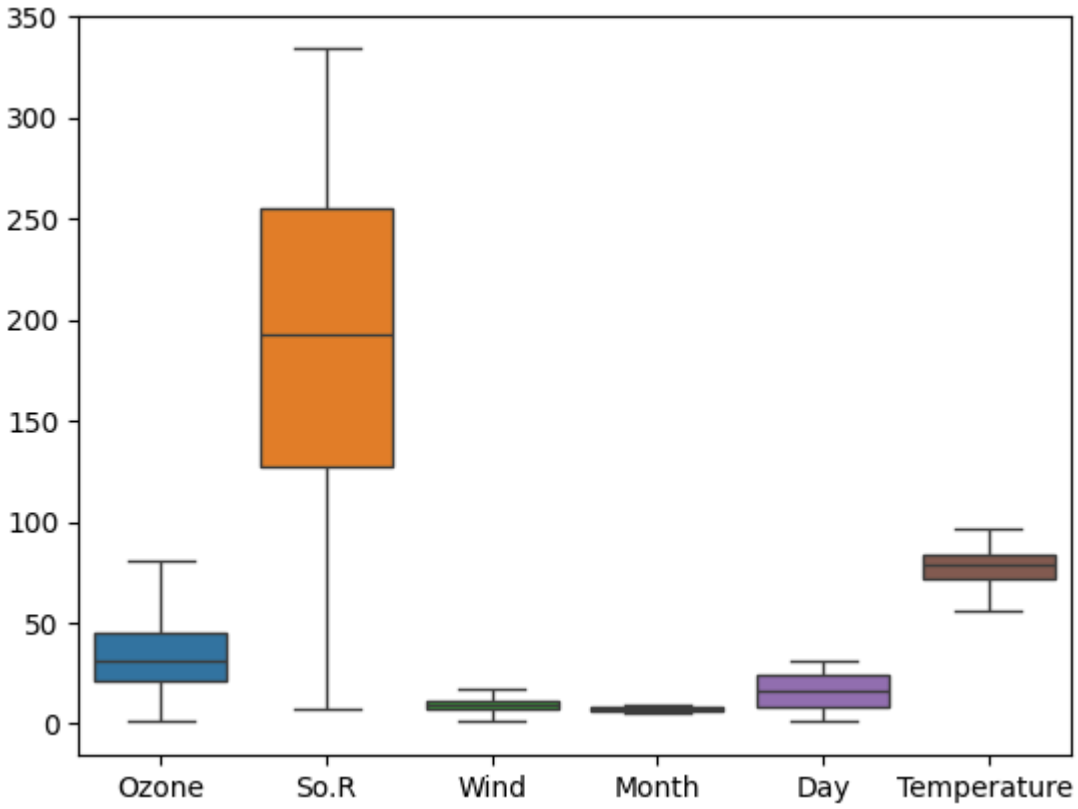
Out[100… 0

In [101…
```
df1[df1['Wind']==17.65].shape[0]
```

Out[101… 3

In [102…
```
sns.boxplot(df1
            )
```

Out[102… <Axes: >



In [103…
```
df1.describe()
```

Out[103…

	Ozone	So.R	Wind	Month	Day	Temperature
count	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000
mean	36.738854	185.366667	9.890127	7.019108	15.929936	77.745223
std	22.475955	86.998999	3.400652	1.434338	8.974404	9.405334
min	1.000000	7.000000	1.700000	5.000000	1.000000	56.000000
25%	21.000000	127.000000	7.400000	6.000000	8.000000	72.000000
50%	31.000000	193.000000	9.700000	7.000000	16.000000	79.000000
75%	45.000000	255.000000	11.500000	8.000000	24.000000	84.000000
max	81.000000	334.000000	17.650000	9.000000	31.000000	97.000000

In [ ]: