Exploratory Data Analysis

In [1]:

```
!pip install pandas_profiling
!pip install sweetviz
```

```
Requirement already satisfied: pandas_profiling in c:\users\admin\anacon
da3\lib\site-packages (3.6.6)
Requirement already satisfied: ydata-profiling in c:\users\admin\anacond
a3\lib\site-packages (from pandas_profiling) (4.0.0)
Requirement already satisfied: matplotlib<3.7,>=3.2 in c:\users\admin\an
aconda3\lib\site-packages (from ydata-profiling->pandas_profiling) (3.5.
2)
Requirement already satisfied: numpy<1.24,>=1.16.0 in c:\users\admin\ana
conda3\lib\site-packages (from ydata-profiling->pandas_profiling) (1.21.
5)
Requirement already satisfied: requests<2.29,>=2.24.0 in c:\users\admin
\anaconda3\lib\site-packages (from ydata-profiling->pandas_profiling)
(2.28.1)
Requirement already satisfied: typeguard<2.14,>=2.13.2 in c:\users\admin
\anaconda3\lib\site-packages (from ydata-profiling->pandas_profiling)
(2.13.3)
Requirement already satisfied: statsmodels<0.14,>=0.13.2 in c:\users\adm
in\anaconda3\lib\site-packages (from ydata-profiling->pandas_profiling)
(0.13.2)
```

In [2]:

```
#Load the libraries
import pandas as pd
import numpy as np
import pandas_profiling as pp
import sweetviz as sv
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_12044\4092403677.py:4: Depreca
tionWarning: `import pandas_profiling` is going to be deprecated by April
1st. Please use `import ydata_profiling` instead.
  import pandas_profiling as pp
```

In [4]:

```
data1 = pd.read_csv("data_clean (1).csv")
```

In [5]:

```
data1
```

Out[5]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
0	1	41.0	190.0	7.4	67	5	1	2010	67	S
1	2	36.0	118.0	8.0	72	5	2	2010	72	C
2	3	12.0	149.0	12.6	74	5	3	2010	74	PS
3	4	18.0	313.0	11.5	62	5	4	2010	62	S
4	5	NaN	NaN	14.3	56	5	5	2010	56	S
...	...	...	...	...	...	...	...	...	...	...
153	154	41.0	190.0	7.4	67	5	1	2010	67	C
154	155	30.0	193.0	6.9	70	9	26	2010	70	PS
155	156	NaN	145.0	13.2	77	9	27	2010	77	S
156	157	14.0	191.0	14.3	75	9	28	2010	75	S
157	158	18.0	131.0	8.0	76	9	29	2010	76	C

158 rows × 10 columns

In [6]:

```
data1.tail(10)
```

Out[6]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
148	149	30.0	193.0	6.9	70	9	26	2010	70	C
149	150	NaN	145.0	13.2	77	9	27	2010	77	PS
150	151	14.0	191.0	14.3	75	9	28	2010	75	S
151	152	18.0	131.0	8.0	76	9	29	2010	76	PS
152	153	20.0	223.0	11.5	68	9	30	2010	68	S
153	154	41.0	190.0	7.4	67	5	1	2010	67	C
154	155	30.0	193.0	6.9	70	9	26	2010	70	PS
155	156	NaN	145.0	13.2	77	9	27	2010	77	S
156	157	14.0	191.0	14.3	75	9	28	2010	75	S
157	158	18.0	131.0	8.0	76	9	29	2010	76	C

In [7]:

```
data1
```

Out[7]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
0	1	41.0	190.0	7.4	67	5	1	2010	67	S
1	2	36.0	118.0	8.0	72	5	2	2010	72	C
2	3	12.0	149.0	12.6	74	5	3	2010	74	PS
3	4	18.0	313.0	11.5	62	5	4	2010	62	S
4	5	NaN	NaN	14.3	56	5	5	2010	56	S
...	...	...	...	...	...	...	...	...	...	...
153	154	41.0	190.0	7.4	67	5	1	2010	67	C
154	155	30.0	193.0	6.9	70	9	26	2010	70	PS
155	156	NaN	145.0	13.2	77	9	27	2010	77	S
156	157	14.0	191.0	14.3	75	9	28	2010	75	S
157	158	18.0	131.0	8.0	76	9	29	2010	76	C

158 rows × 10 columns

In [8]:

```
#Data Structure
type(data1)
data1.shape
```

Out[8]:

```
(158, 10)
```

In [9]:

```
#data types
data1.dtypes
```

Out[9]:

```
Unnamed: 0      int64
Ozone         float64
Solar.R       float64
Wind          float64
Temp C         object
Month          object
Day             int64
Year            int64
Temp            int64
Weather        object
dtype: object
```

Data type conversion

In [10]:

```
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  158 non-null    int64  
 1   Ozone       120 non-null    float64
 2   Solar.R     151 non-null    float64
 3   Wind        158 non-null    float64
 4   Temp C      158 non-null    object 
 5   Month       158 non-null    object 
 6   Day         158 non-null    int64  
 7   Year        158 non-null    int64  
 8   Temp        158 non-null    int64  
 9   Weather     155 non-null    object 
dtypes: float64(3), int64(4), object(3)
memory usage: 12.5+ KB
```

In [11]:

```
data1
```

Out[11]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
0	1	41.0	190.0	7.4	67	5	1	2010	67	S
1	2	36.0	118.0	8.0	72	5	2	2010	72	C
2	3	12.0	149.0	12.6	74	5	3	2010	74	PS
3	4	18.0	313.0	11.5	62	5	4	2010	62	S
4	5	NaN	NaN	14.3	56	5	5	2010	56	S
...	...	...	...	...	...	...	...	...	...	...
153	154	41.0	190.0	7.4	67	5	1	2010	67	C
154	155	30.0	193.0	6.9	70	9	26	2010	70	PS
155	156	NaN	145.0	13.2	77	9	27	2010	77	S
156	157	14.0	191.0	14.3	75	9	28	2010	75	S
157	158	18.0	131.0	8.0	76	9	29	2010	76	C

158 rows × 10 columns

In [12]:

```
data2=data1.iloc[:,1:]
```

In [13]:

```
data2
```

Out[13]:

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

158 rows × 9 columns

In [14]:

```
#The method .copy() is used here so that any changes made in new DataFrame don't get refl
data=data2.copy()
```

In [15]:

```
data['Month']=pd.to_numeric(data['Month'],errors='coerce')
data['Temp C']=pd.to_numeric(data['Temp C'],errors='coerce')# coerce will introduce NA va
data['Weather']=data['Weather'].astype('category')           #data['Wind']=data['Wind'].a
```

In [16]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 9 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   Temp C   157 non-null    float64 
 4   Month    157 non-null    float64 
 5   Day      158 non-null    int64   
 6   Year     158 non-null    int64   
 7   Temp     158 non-null    int64   
 8   Weather  155 non-null    category
dtypes: category(1), float64(5), int64(3)
memory usage: 10.3 KB
```

Duplicates

In [17]:

```
#Count of duplicated rows
data[data.duplicated()].shape
```

Out[17]:

```
(1, 9)
```

In [18]:

```
data
```

Out[18]:

	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	67.0	5.0	1	2010	67	S
1	36.0	118.0	8.0	72.0	5.0	2	2010	72	C
2	12.0	149.0	12.6	74.0	5.0	3	2010	74	PS
3	18.0	313.0	11.5	62.0	5.0	4	2010	62	S
4	NaN	NaN	14.3	56.0	5.0	5	2010	56	S
...	...	...	...	...	...	...	...	...	...
153	41.0	190.0	7.4	67.0	5.0	1	2010	67	C
154	30.0	193.0	6.9	70.0	9.0	26	2010	70	PS
155	NaN	145.0	13.2	77.0	9.0	27	2010	77	S
156	14.0	191.0	14.3	75.0	9.0	28	2010	75	S
157	18.0	131.0	8.0	76.0	9.0	29	2010	76	C

158 rows × 9 columns

In [19]:

```
#Print the duplicated rows
data[data.duplicated()]
```

Out[19]:

	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
156	14.0	191.0	14.3	75.0	9.0	28	2010	75	S

In [20]:

```
data_cleaned1=data.drop_duplicates()
```

In [21]:

```
data_cleaned1.shape
```

Out[21]:

```
(157, 9)
```

Drop columns

In [22]:

```
data_cleaned2=data_cleaned1.drop('Temp C',axis=1)
```

In [23]:

```
data_cleaned2
```

Out[23]:

	Ozone	Solar.R	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5.0	1	2010	67	S
1	36.0	118.0	8.0	5.0	2	2010	72	C
2	12.0	149.0	12.6	5.0	3	2010	74	PS
3	18.0	313.0	11.5	5.0	4	2010	62	S
4	NaN	NaN	14.3	5.0	5	2010	56	S
...	...	...	...	...	...	...	...	...
152	20.0	223.0	11.5	9.0	30	2010	68	S
153	41.0	190.0	7.4	5.0	1	2010	67	C
154	30.0	193.0	6.9	9.0	26	2010	70	PS
155	NaN	145.0	13.2	9.0	27	2010	77	S
157	18.0	131.0	8.0	9.0	29	2010	76	C

157 rows × 8 columns

Rename the columns

In [24]:

```
#rename the Solar column
data_cleaned3 = data_cleaned2.rename({'Solar.R': 'Solar'}, axis=1)
```

In [25]:

```
data_cleaned3
```

Out[25]:

	Ozone	Solar	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5.0	1	2010	67	S
1	36.0	118.0	8.0	5.0	2	2010	72	C
2	12.0	149.0	12.6	5.0	3	2010	74	PS
3	18.0	313.0	11.5	5.0	4	2010	62	S
4	NaN	NaN	14.3	5.0	5	2010	56	S
...	...	...	...	...	...	...	...	...
152	20.0	223.0	11.5	9.0	30	2010	68	S
153	41.0	190.0	7.4	5.0	1	2010	67	C
154	30.0	193.0	6.9	9.0	26	2010	70	PS
155	NaN	145.0	13.2	9.0	27	2010	77	S
157	18.0	131.0	8.0	9.0	29	2010	76	C

157 rows × 8 columns

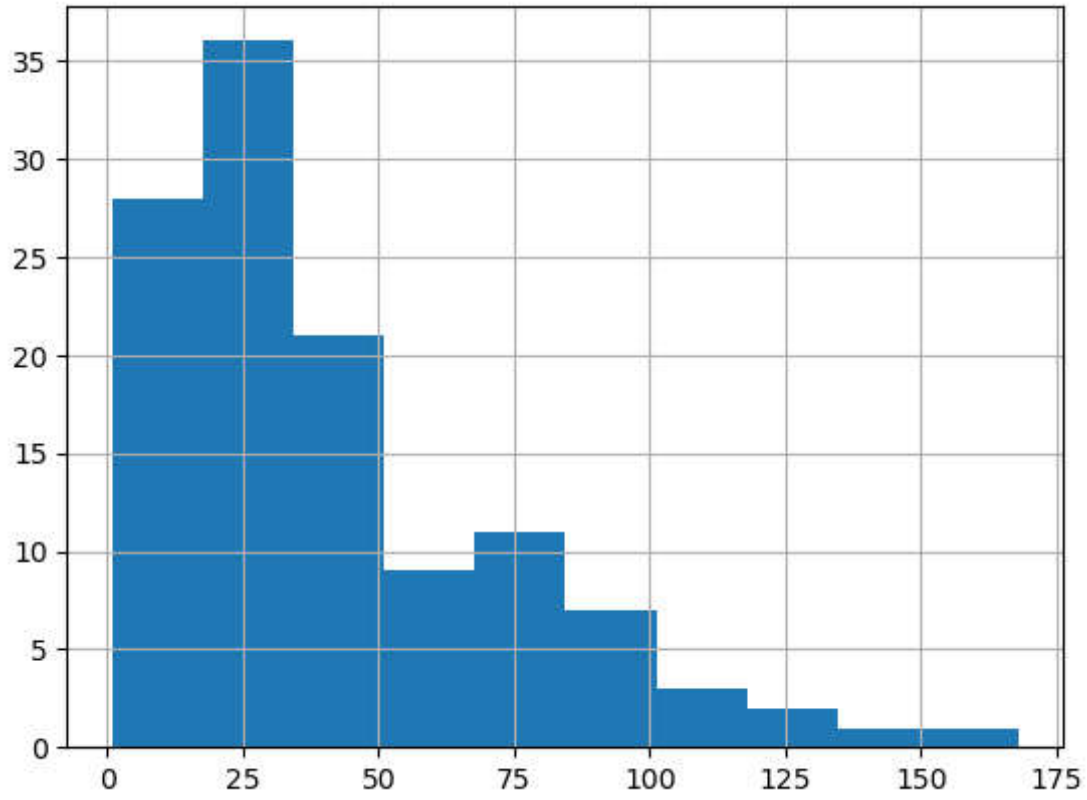
Outlier Detection

In [26]:

```
# histogram of Ozone
data_cleaned3['Ozone'].hist()
```

Out[26]:

```
<AxesSubplot:>
```

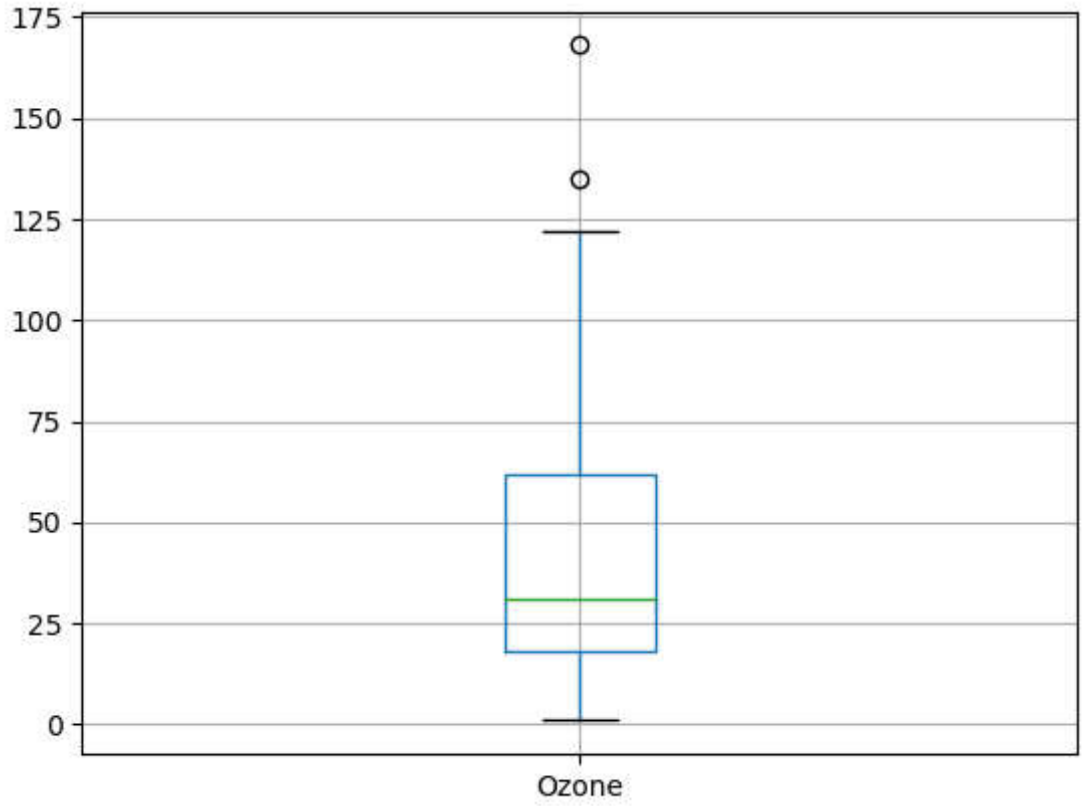


In [27]:

```
#Box plot
data_cleaned3.boxplot(column=['Ozone'])
```

Out[27]:

```
<AxesSubplot:>
```



In [28]:

```
#Descriptive stat
data_cleaned3['Ozone'].describe()
```

Out[28]:

```
count    119.000000
mean      41.815126
std       32.659249
min        1.000000
25%       18.000000
50%       31.000000
75%       62.000000
max      168.000000
Name: Ozone, dtype: float64
```

In [29]:

```
data_cleaned3
```

Out[29]:

	Ozone	Solar	Wind	Month	Day	Year	Temp	Weather
0	41.0	190.0	7.4	5.0	1	2010	67	S
1	36.0	118.0	8.0	5.0	2	2010	72	C
2	12.0	149.0	12.6	5.0	3	2010	74	PS
3	18.0	313.0	11.5	5.0	4	2010	62	S
4	NaN	NaN	14.3	5.0	5	2010	56	S
...	...	...	...	...	...	...	...	...
152	20.0	223.0	11.5	9.0	30	2010	68	S
153	41.0	190.0	7.4	5.0	1	2010	67	C
154	30.0	193.0	6.9	9.0	26	2010	70	PS
155	NaN	145.0	13.2	9.0	27	2010	77	S
157	18.0	131.0	8.0	9.0	29	2010	76	C

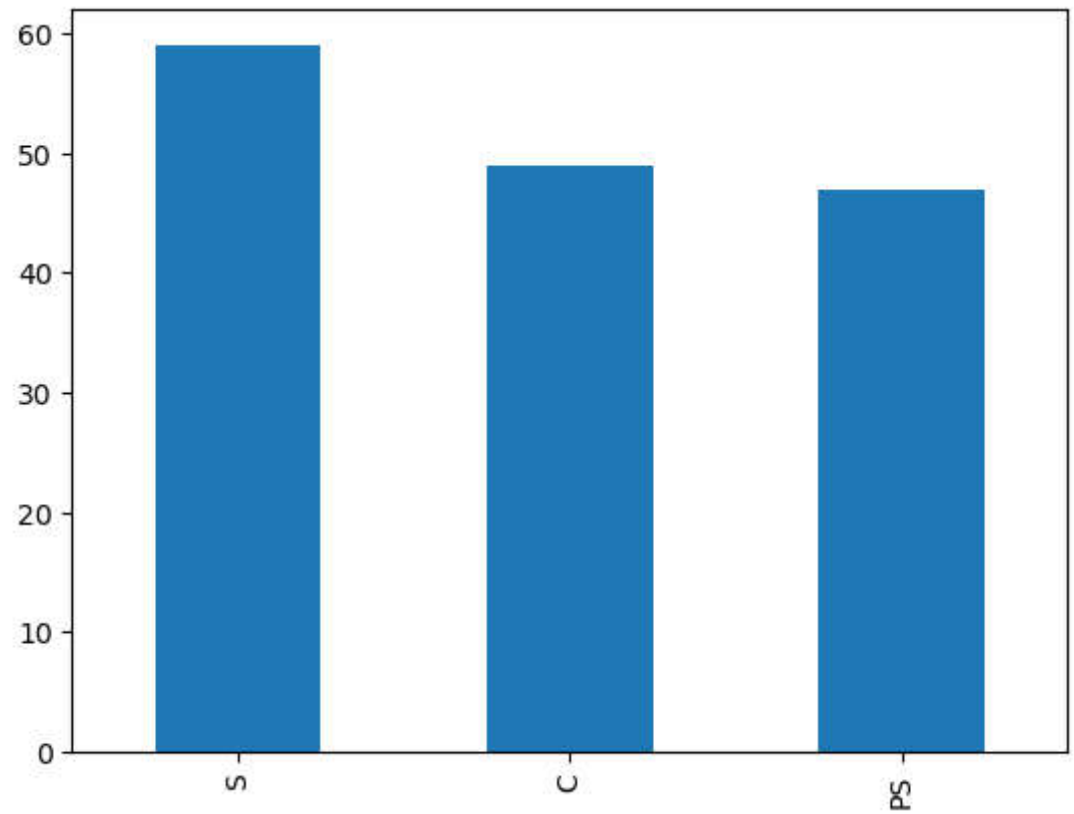
157 rows × 8 columns

In [30]:

```
#Bar plot
data['Weather'].value_counts().plot.bar()
```

Out[30]:

```
<AxesSubplot:>
```



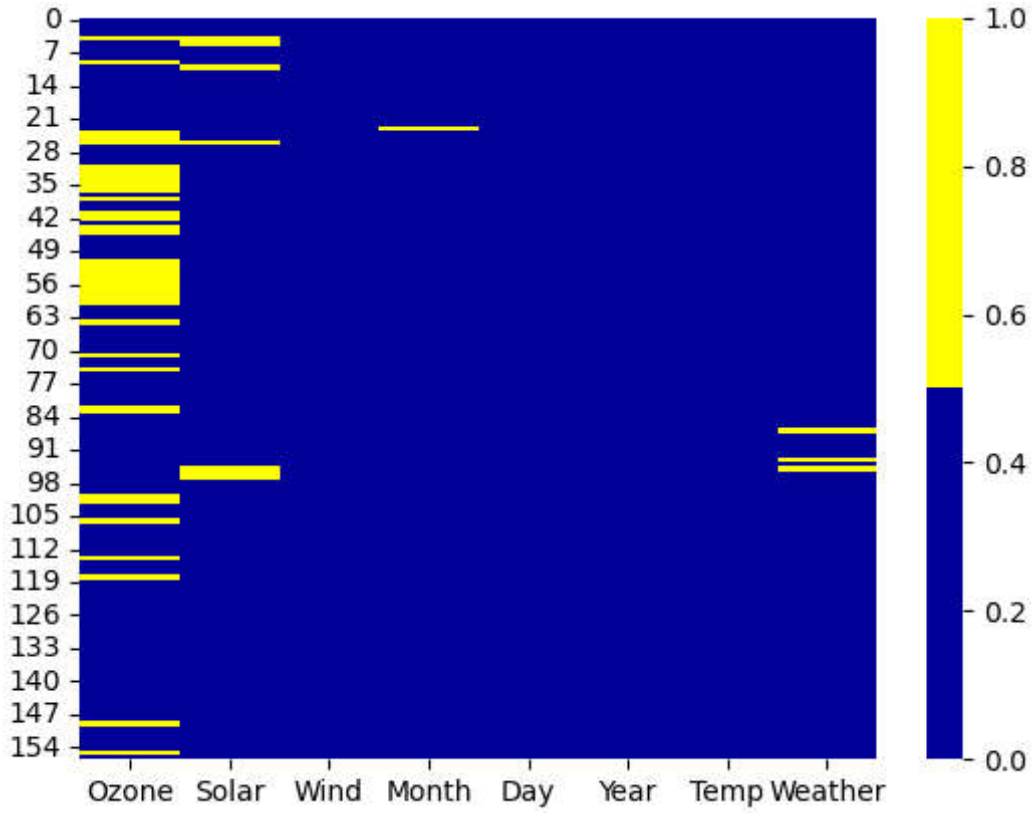
Missing Values and Imputation

In [31]:

```
import seaborn as sns
cols = data_cleaned3.columns
colours = ['#000099', '#ffff00'] # specify the colours - yellow is missing. blue is not m
sns.heatmap(data_cleaned3[cols].isnull(),
            cmap=sns.color_palette(colours))
```

Out[31]:

<AxesSubplot:>



In [32]:

```
data_cleaned3[data_cleaned3.isnull().any(axis=1)].head()
```

Out[32]:

	Ozone	Solar	Wind	Month	Day	Year	Temp	Weather
4	NaN	NaN	14.3	5.0	5	2010	56	S
5	28.0	NaN	14.9	5.0	6	2010	66	C
9	NaN	194.0	8.6	5.0	10	2010	69	S
10	7.0	NaN	6.9	5.0	11	2010	74	C
23	32.0	92.0	12.0	NaN	24	2010	61	C

In [33]:

```
data_cleaned3.isnull().sum()
```

Out[33]:

```
Ozone      38
Solar       7
Wind        0
Month       1
Day         0
Year        0
Temp        0
Weather     3
dtype: int64
```

In [34]:

```
#Mean Imputation
mean = data_cleaned3['Ozone'].mean()
print(mean)
```

```
41.81512605042017
```

In [35]:

```
data_cleaned3['Ozone'] = data_cleaned3['Ozone'].fillna(mean)
```

In [36]:

```
data_cleaned3
```

Out[36]:

	Ozone	Solar	Wind	Month	Day	Year	Temp	Weather
0	41.000000	190.0	7.4	5.0	1	2010	67	S
1	36.000000	118.0	8.0	5.0	2	2010	72	C
2	12.000000	149.0	12.6	5.0	3	2010	74	PS
3	18.000000	313.0	11.5	5.0	4	2010	62	S
4	41.815126	NaN	14.3	5.0	5	2010	56	S
...	...	...	...	...	...	...	...	...
152	20.000000	223.0	11.5	9.0	30	2010	68	S
153	41.000000	190.0	7.4	5.0	1	2010	67	C
154	30.000000	193.0	6.9	9.0	26	2010	70	PS
155	41.815126	145.0	13.2	9.0	27	2010	77	S
157	18.000000	131.0	8.0	9.0	29	2010	76	C

157 rows × 8 columns

In [37]:

```
#Missing value imputation for categorical vlaue
#Get the object columns
obj_columns=data_cleaned3[['Weather']]
```

In [38]:

```
obj_columns.isnull().sum()
```

Out[38]:

```
Weather    3
dtype: int64
```

In [39]:

```
#Missing value imputation for categorical vlaue
obj_columns=obj_columns.fillna(obj_columns.mode().iloc[0])
```

In [40]:

```
obj_columns.isnull().sum()
```

Out[40]:

```
Weather    0
dtype: int64
```

In [41]:

```
data_cleaned3.shape
```

Out[41]:

```
(157, 8)
```

In [42]:

```
obj_columns.shape
```

Out[42]:

```
(157, 1)
```

In [43]:

```
#Join the data set with imputed object dataset
data_cleaned4=pd.concat([data_cleaned3,obj_columns],axis=1)
```

In [44]:

```
data_cleaned4.isnull().sum()
```

Out[44]:

```
Ozone      0
Solar      7
Wind       0
Month      1
Day        0
Year       0
Temp       0
Weather    3
Weather    0
dtype: int64
```

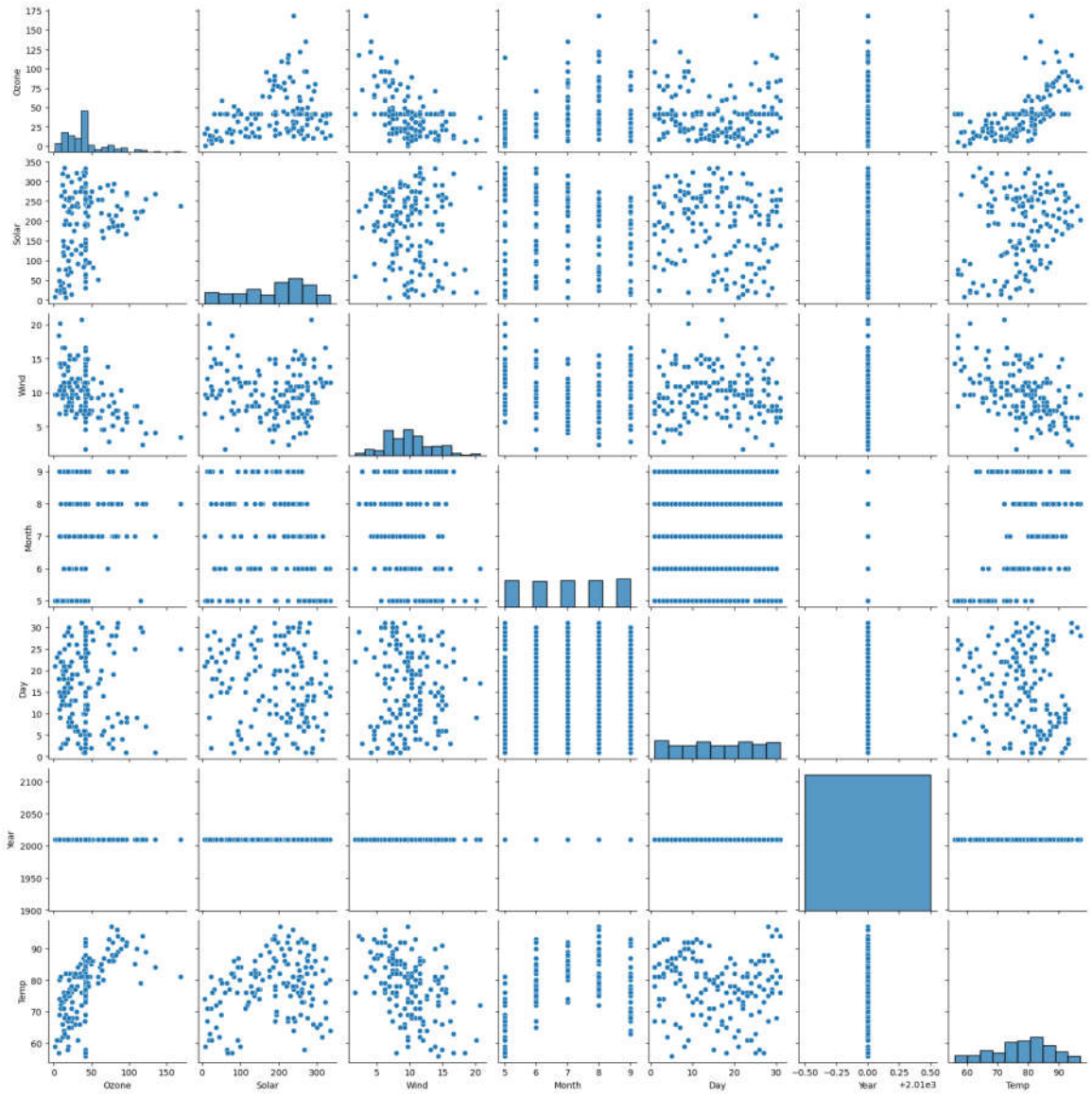
Scatter plot and Correlation analysis

In [45]:

```
# Seaborn visualization library
import seaborn as sns
# Create the default pairplot
sns.pairplot(data_cleaned3)
```

Out[45]:

```
<seaborn.axisgrid.PairGrid at 0x1da1f8a6e50>
```



In [46]:

```
#Correlation
data_cleaned3.corr()
```

Out[46]:

	Ozone	Solar	Wind	Month	Day	Year	Temp
Ozone	1.000000	0.308687	-0.520004	0.132860	-0.021916	NaN	0.606500
Solar	0.308687	1.000000	-0.057407	-0.094012	-0.155663	NaN	0.273558
Wind	-0.520004	-0.057407	1.000000	-0.166216	0.029900	NaN	-0.441228
Month	0.132860	-0.094012	-0.166216	1.000000	0.050055	NaN	0.398516
Day	-0.021916	-0.155663	0.029900	0.050055	1.000000	NaN	-0.122787
Year	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Temp	0.606500	0.273558	-0.441228	0.398516	-0.122787	NaN	1.000000

Transformations

Dummy Variable

In [47]:

```
#Creating dummy variable for Weather column
data_cleaned4=pd.get_dummies(data,columns=['Weather'])
```

In [48]:

```
data_cleaned4
```

Out[48]:

	Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather_C	Weather_PS	Weath
0	41.0	190.0	7.4	67.0	5.0	1	2010	67	0	0	
1	36.0	118.0	8.0	72.0	5.0	2	2010	72	1	0	
2	12.0	149.0	12.6	74.0	5.0	3	2010	74	0	1	
3	18.0	313.0	11.5	62.0	5.0	4	2010	62	0	0	
4	NaN	NaN	14.3	56.0	5.0	5	2010	56	0	0	
...	...	...	...	...	...	...	...	...	...	...	
153	41.0	190.0	7.4	67.0	5.0	1	2010	67	1	0	
154	30.0	193.0	6.9	70.0	9.0	26	2010	70	0	1	
155	NaN	145.0	13.2	77.0	9.0	27	2010	77	0	0	
156	14.0	191.0	14.3	75.0	9.0	28	2010	75	0	0	
157	18.0	131.0	8.0	76.0	9.0	29	2010	76	1	0	

158 rows × 11 columns

In [49]:

```
data_cleaned4=data_cleaned4.dropna()
```

Normalization of the data

In [50]:

```
#Normalization of the data
from numpy import set_printoptions
from sklearn.preprocessing import MinMaxScaler
```

In [51]:

```
data_cleaned4.values
```

Out[51]:

```
array([[ 41. , 190. ,   7.4, ...,   0. ,   0. ,   1. ],
       [ 36. , 118. ,   8. , ...,   1. ,   0. ,   0. ],
       [ 12. , 149. ,  12.6, ...,   0. ,   1. ,   0. ],
       ...,
       [ 30. , 193. ,   6.9, ...,   0. ,   1. ,   0. ],
       [ 14. , 191. ,  14.3, ...,   0. ,   0. ,   1. ],
       [ 18. , 131. ,   8. , ...,   1. ,   0. ,   0. ]])
```

In [52]:

```
array = data_cleaned3.values

scaler = MinMaxScaler(feature_range=(0,1))
rescaledX = scaler.fit_transform(array[:,0:5])

#transformed data
set_printoptions(precision=2)
print(rescaledX[0:5,:])
```

```
[[0.24 0.56 0.3  0.   0.  ]
 [0.21 0.34 0.33 0.   0.03]
 [0.07 0.43 0.57 0.   0.07]
 [0.1  0.94 0.52 0.   0.1 ]
 [0.24  nan 0.66 0.   0.13]]
```

In [53]:

```
# Standardize data (0 mean, 1 stdev)
from sklearn.preprocessing import StandardScaler
```

In [54]:

```
array = data_cleaned4.values
scaler = StandardScaler().fit(array)
rescaledX = scaler.transform(array)

# summarize transformed data
set_printoptions(precision=2)
print(rescaledX[0:5,:])
```

```
[[-0.02  0.05 -0.71 -1.15 -1.53 -1.7   0.   -1.15 -0.64 -0.68  1.28]
 [-0.17 -0.75 -0.54 -0.62 -1.53 -1.59  0.   -0.62  1.57 -0.68 -0.78]
 [-0.9  -0.41  0.77 -0.4  -1.53 -1.48  0.   -0.4  -0.64  1.47 -0.78]
 [-0.72  1.43  0.45 -1.69 -1.53 -1.36  0.   -1.69 -0.64 -0.68  1.28]
 [-0.57  1.27 -0.37 -1.37 -1.53 -1.02  0.   -1.37 -0.64  1.47 -0.78]]
```

Speed up the EDA process ¶

In [55]:

```
EDA_report= pp.ProfileReport(data)
EDA_report.to_file(output_file='report.html')
```

Summarize dataset: 55/55 [00:10<00:00, 4.57it/s, 100% Completed]

Generate report structure: 1/1 [00:05<00:00, 100% 5.05s/it]

Render HTML: 1/1 [00:02<00:00, 100% 2.02s/it]

Export report to file: 1/1 [00:00<00:00, 100% 16.71it/s]

In [56]:

```
sweet_report = sv.analyze(data)
sweet_report.show_html('weather_report.html')
```

Done! Use 'show' commands to display/save. [100%] 00:01 -> (00:00 left)

```
Report weather_report.html was generated! NOTEBOOK/COLAB USERS: the web br
owser MAY not pop up, regardless, the report IS saved in your notebook/col
ab files.
```

In [ ]:

In [ ]:

```
```

In [ ]:

```
```