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

In [3]:
```
df=pd.read_csv('AirQualityodisha.csv')
df
```

Out[3]:

	Stn Code	Sampling Date	State	City	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM1
0	68	02-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	11	24	14
1	68	06-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	10	23	13
2	68	09-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	8	25	12
3	68	13-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	10	25	13
4	68	16-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	9	26	18
...	...	...	...	...	...	...	...	...	...	
2387	819	15-12-15	Odisha	Kalinga Nagar	Roof of RO OFFICE BUILDING	Odisha State Pollution Control Board	Industrial Area	2	10	9
2388	819	17-12-15	Odisha	Kalinga Nagar	Roof of RO OFFICE BUILDING	Odisha State Pollution Control Board	Industrial Area	2	10	9
2389	819	22-12-15	Odisha	Kalinga Nagar	Roof of RO OFFICE BUILDING	Odisha State Pollution Control Board	Industrial Area	2	10	9
2390	819	26-12-15	Odisha	Kalinga Nagar	Roof of RO OFFICE BUILDING	Odisha State Pollution Control Board	Industrial Area	2	10	9
2391	819	29-12-15	Odisha	Kalinga Nagar	Roof of RO OFFICE BUILDING	Odisha State Pollution Control Board	Industrial Area	2	10	9

2392 rows × 11 columns

In [12]:
```
#--------------DATA CLEANING------------#
```

In [11]:
```
df.head()
```

Out[11]:

	Stn Code	Sampling Date	State	City	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	68	02-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	11	24	143
1	68	06-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	10	23	133
2	68	09-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	8	25	125
3	68	13-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	10	25	137
4	68	16-01-15	Odisha	Talcher	T.T.P.S.Colony, Talcher	Odisha State Pollution Control Board	Industrial Area	9	26	186

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Stn Code                        2392 non-null   int64  
 1   Sampling Date                   2392 non-null   object 
 2   State                           2392 non-null   object 
 3   City                            2392 non-null   object 
 4   Location of Monitoring Station  2392 non-null   object 
 5   Agency                          2392 non-null   object 
 6   Type of Location                2392 non-null   object 
 7   SO2                             2392 non-null   int64  
 8   NO2                             2392 non-null   int64  
 9   RSPM/PM10                       2392 non-null   int64  
 10  PM 2.5                          2060 non-null   float64
dtypes: float64(1), int64(4), object(6)
memory usage: 205.7+ KB
```

In [10]:
```
df.replace(to_replace=-200,value=np.nan,inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Stn Code                        2392 non-null   int64  
 1   Sampling Date                   2392 non-null   object 
 2   State                           2392 non-null   object 
 3   City                            2392 non-null   object 
 4   Location of Monitoring Station  2392 non-null   object 
 5   Agency                          2392 non-null   object 
 6   Type of Location                2392 non-null   object 
 7   SO2                             2392 non-null   int64  
 8   NO2                             2392 non-null   int64  
 9   RSPM/PM10                       2392 non-null   int64  
 10  PM 2.5                          2060 non-null   float64
dtypes: float64(1), int64(4), object(6)
memory usage: 205.7+ KB
```

In [17]:
```
df.drop(['PM 2.5','RSPM/PM10'],axis = 1,inplace = True)
```

In [19]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   Stn Code                        2392 non-null   int64 
 1   Sampling Date                   2392 non-null   object
 2   State                           2392 non-null   object
 3   City                            2392 non-null   object
 4   Location of Monitoring Station  2392 non-null   object
 5   Agency                          2392 non-null   object
 6   Type of Location                2392 non-null   object
 7   SO2                             2392 non-null   int64 
 8   NO2                             2392 non-null   int64 
dtypes: int64(3), object(6)
memory usage: 168.3+ KB
```

In [22]:
```
df.drop('Agency', axis=1, inplace=True) #deleterd due to null value

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 8 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Stn Code                        2392 non-null   int64
 1   Sampling Date                   2392 non-null   object
 2   State                           2392 non-null   object
 3   City                            2392 non-null   object
 4   Location of Monitoring Station  2392 non-null   object
 5   Type of Location                2392 non-null   object
 6   SO2                             2392 non-null   int64
 7   NO2                             2392 non-null   int64
dtypes: int64(3), object(5)
memory usage: 149.6+ KB
```

In [23]:
```
#------------DATA INTEGRATION-------------#
```

In [24]:
```
sns.set_theme(style="whitegrid")
```

In [29]:
```
df.shape
```

Out[29]:
```
(2392, 8)
```

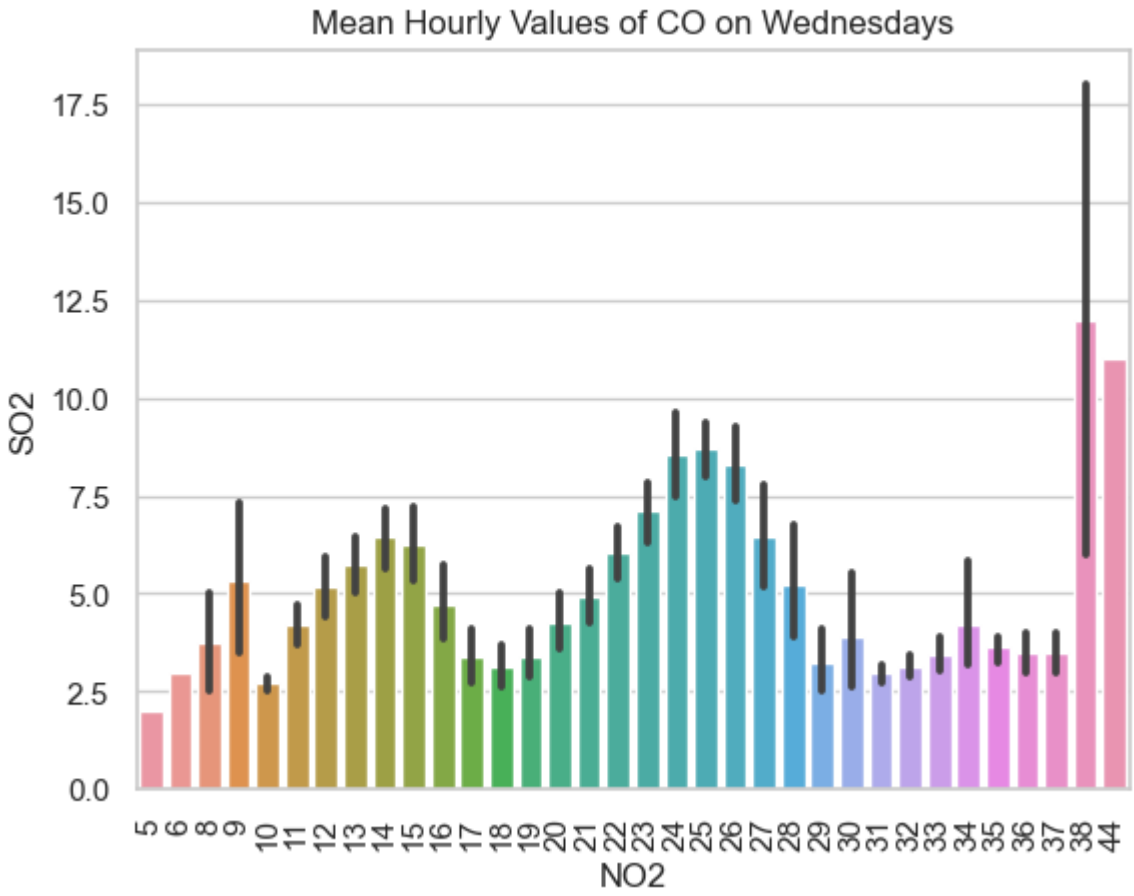
In [ ]:
```
Q1 = df.quantile(0.25)                  #first 25% of the data
Q3 = df.quantile(0.75)                  #first 75% of the data
IQR = Q3 - Q1                           #IQR = InterQuartile Range

scale = 2                              #For Normal Distributions, scale = 1.5
lower_lim = Q1 - scale*IQR
upper_lim = Q3 + scale*IQR

lower_outliers = (df[df.columns[2:13]] < lower_lim)
upper_outliers = (df[df.columns[2:13]] > upper_lim)
```

In [34]:
```
df[df.columns[2:13]][(lower_outliers | upper_outliers)].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 6 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   State                           0 non-null      object
 1   City                            0 non-null      object
 2   Location of Monitoring Station  0 non-null      object
 3   Type of Location                0 non-null      object
 4   SO2                             33 non-null     float64
 5   NO2                             0 non-null      float64
dtypes: float64(2), object(4)
memory usage: 112.2+ KB
```

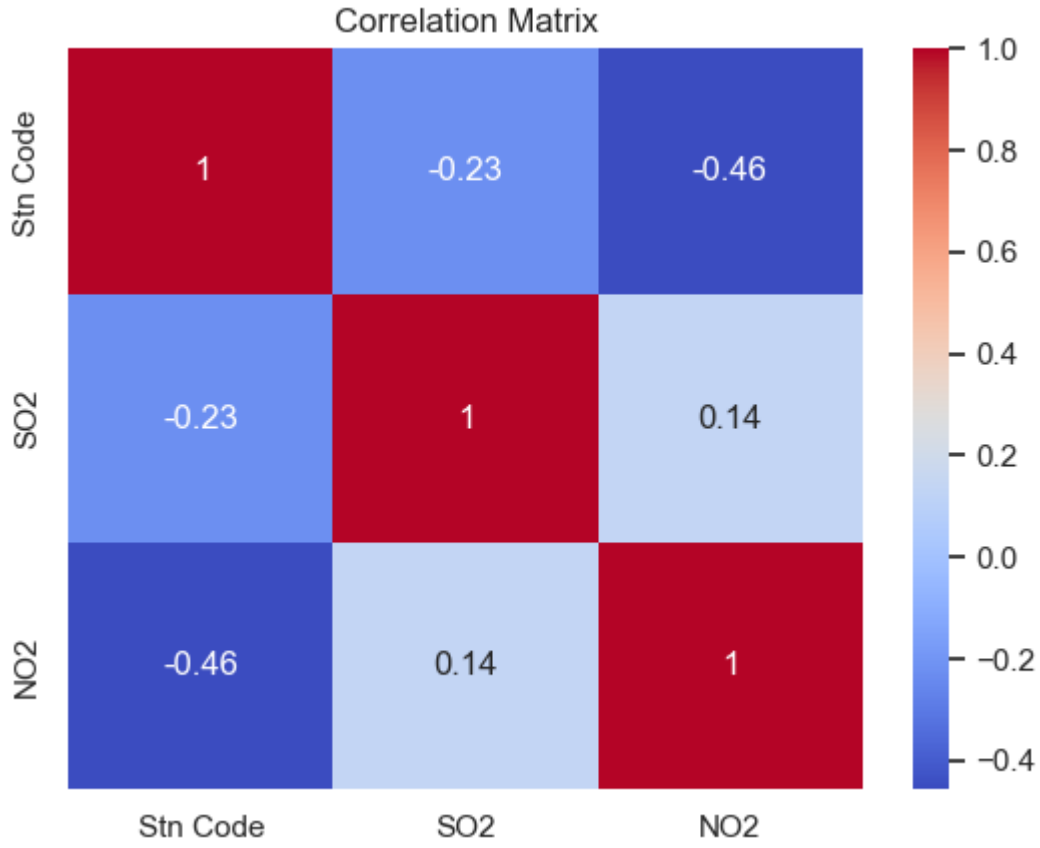
In [35]:
```
#----------------DATA TRANSFORMATION----------------#
```

In [50]:
```
sns.barplot(x='NO2',y='SO2', data=df.sort_values('NO2'))
plt.title('Mean Hourly Values of CO on Wednesdays')
plt.xticks(rotation=90)
plt.show()
```

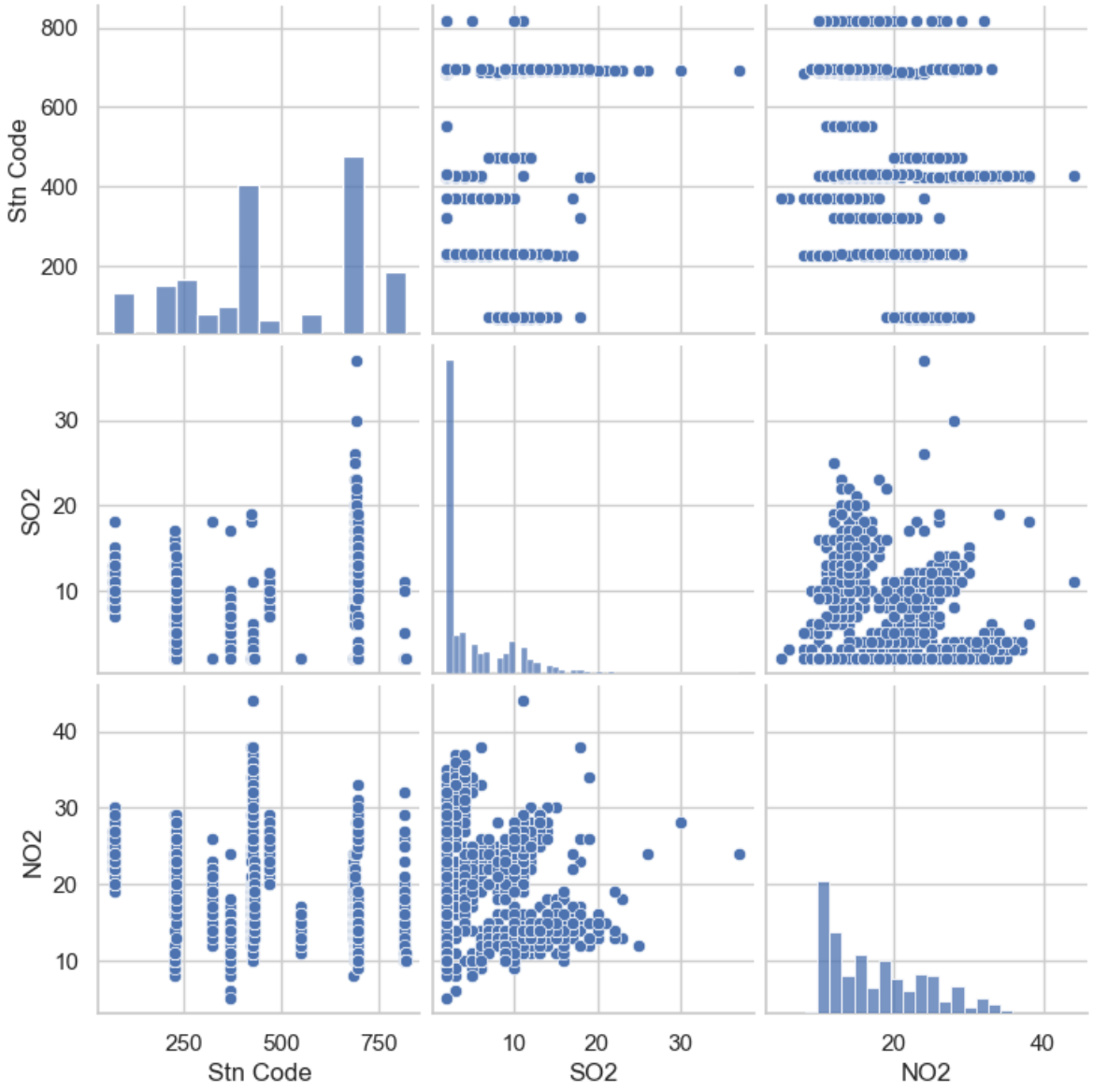


In [51]:
```
sns.heatmap(df.corr(),annot=True,cmap = 'coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
C:\Users\cnnar\AppData\Local\Temp\ipykernel_18168\712187306.py:1: FutureWarni
ng: The default value of numeric_only in DataFrame.corr is deprecated. In a f
uture version, it will default to False. Select only valid columns or specify
the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot=True,cmap = 'coolwarm')
```



In [54]:
```
sns.pairplot(df)
plt.show()
```



In [56]:
```
#-----------------Model BULIDING-----------------#
```

In [62]:
```
df.drop('Location of Monitoring Station', axis=1, inplace=True)
```

In [65]:
```
from sklearn.model_selection import train_test_split

Y = df['NO2'] #variável de predição
X = df.drop(['State','City'], axis=1)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, randc
print(X_train.shape, X_test.shape)
```

```
(1913, 5) (479, 5)
```

In [72]:
```
sns.pairplot(df, hue='Season')
plt.show()
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File C:\ProgramData\anaconda3\lib\site-packages\pandas\core\indexes\base.py:3802, in Index.get_loc(self, key, method, tolerance)
   3801 try:
-> 3802     return self._engine.get_loc(casted_key)
   3803 except KeyError as err:

File C:\ProgramData\anaconda3\lib\site-packages\pandas\_libs\index.pyx:138, in pandas._libs.index.IndexEngine.get_loc()

File C:\ProgramData\anaconda3\lib\site-packages\pandas\_libs\index.pyx:165, in pandas._libs.index.IndexEngine.get_loc()

File pandas\_libs\hashtable_class_helper.pxi:5745, in pandas._libs.hashtable.PyObjectHashTable.get_item()
```

In [ ]: