DHRUV GARG 16 BCE 1190

DATA VISUALIZATION
TIME SERIES ANALYSIS

DATASET : Univariate Time Series Prediction of Five Air Contaminates

Step 1: Import modules and data

```
import pandas as pd
import numpy as np
% matplotlib inline
import matplotlib.pyplot as plt
```

Import dataset

```
data = pd.read_csv('AirQualityUCI.csv', sep=';')
data.head()
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	Unnamed: 15	Unnamed: 16
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13,6	48,9	0,7578	NaN	NaN
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	972.0	13,3	47,7	0,7255	NaN	NaN
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11,9	54,0	0,7502	NaN	NaN

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	Unnamed: 15	Unnamed: 16
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11,0	60,0	0,7867	NaN	NaN
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11,2	59,6	0,7888	NaN	NaN

Step 2: Cleaning the dataset

Removing missing values

We can see that the columns “Unnamed: 15” and “Unnamed: 16” are empty. So we drop them.

```
data = data.drop(['Unnamed: 15', 'Unnamed: 16'], axis=1)
```

Removing not a number values

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

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
9357	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9358	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9359	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9360	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9361	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Drop NaN values and obtain 9357 records

```
data = data.dropna()
```

Step 3: View the modified data

```
data.describe()
```

	PT08.S1(CO)	NMHC(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)
count	8991.000000	914.000000	8991.000000	7718.000000	8991.000000	7715.000000	8991.000000	8991.000000
mean	1099.833166	218.811816	939.153376	246.896735	835.493605	113.091251	1456.264598	1022.906128
std	217.080037	204.459921	266.831429	212.979168	256.817320	48.370108	346.206794	398.484288
min	647.000000	7.000000	383.000000	2.000000	322.000000	2.000000	551.000000	221.000000
25%	937.000000	67.000000	734.500000	98.000000	658.000000	78.000000	1227.000000	731.500000
50%	1063.000000	150.000000	909.000000	180.000000	806.000000	109.000000	1463.000000	963.000000
75%	1231.000000	297.000000	1116.000000	326.000000	969.500000	142.000000	1674.000000	1273.500000
max	2040.000000	1189.000000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000

Step 4: Time series conversion of dataset

```
data.loc[:,'Datetime'] = data['Date'] + ' ' + data['Time']

from datetime import datetime
DateTime = []
for x in data['Datetime']:
   DateTime.append(datetime.strptime(x,'%d/%m/%Y %H.%M.%S'))

datetime = pd.Series(DateTime)
data.index = datetime
```

Step 5: Viewing the modified dataset

```
data.head()
```

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH	Datetime
2004-03-10 18:00:00	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13,6	48,9	0,7578	10/03/2004 18.00.00
2004-03-10 19:00:00	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	972.0	13,3	47,7	0,7255	10/03/2004 19.00.00
2004-03-10 20:00:00	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11,9	54,0	0,7502	10/03/2004 20.00.00
2004-03-10 21:00:00	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11,0	60,0	0,7867	10/03/2004 21.00.00
2004-03-10 22:00:00	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11,2	59,6	0,7888	10/03/2004 22.00.00

EXPLORATORY DATA VISUALIZATION

Step 1: Segregate the 5 gases as S1, S2, S3, S4, and S5.

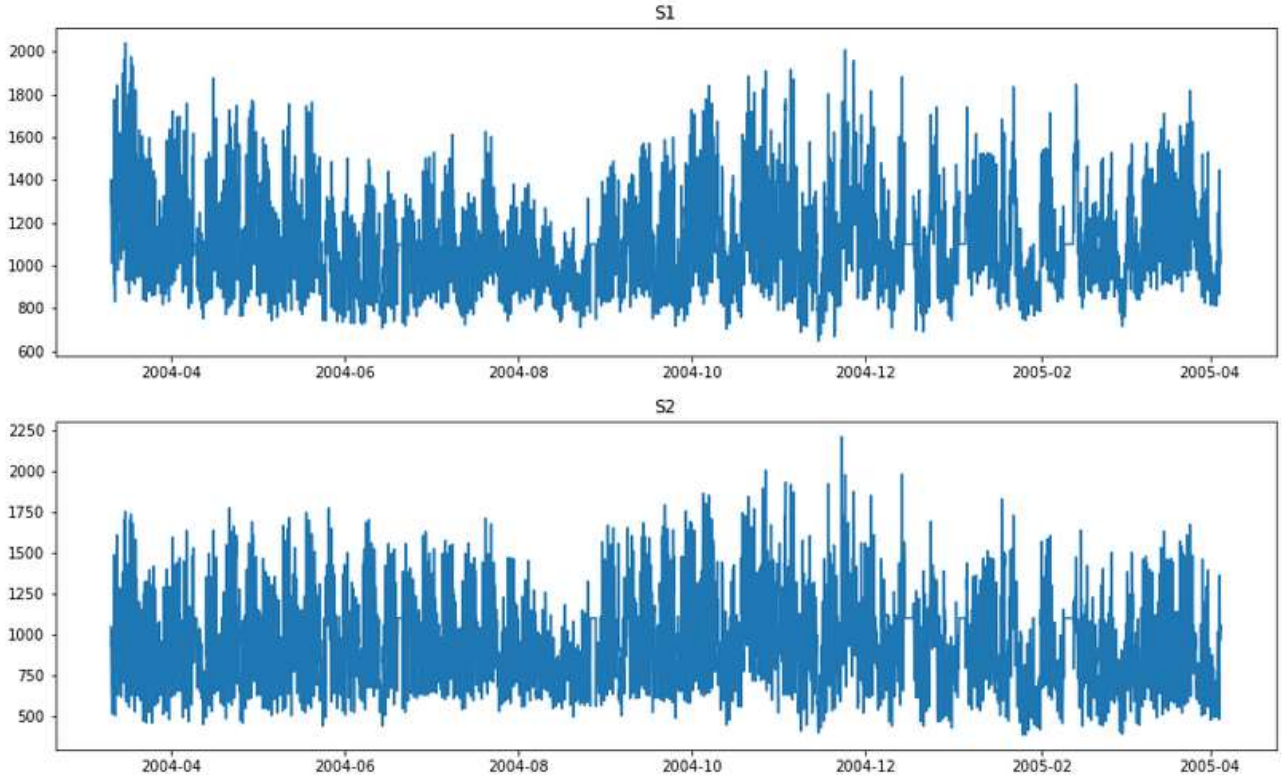
```
S1 = data['PT08.S1(CO)'].fillna(data['PT08.S1(CO)'].mean())
S2 = data['PT08.S2(NMHC)'].fillna(data['PT08.S1(CO)'].mean())
S3 = data['PT08.S3(NOx)'].fillna(data['PT08.S1(CO)'].mean())
S4 = data['PT08.S4(NO2)'].fillna(data['PT08.S1(CO)'].mean())
S5 = data['PT08.S5(O3)'].fillna(data['PT08.S1(CO)'].mean())
```

DHRUV GARG 16 BCE 1190

PLOT 1 : Time series plot for all contaminants over the entire period

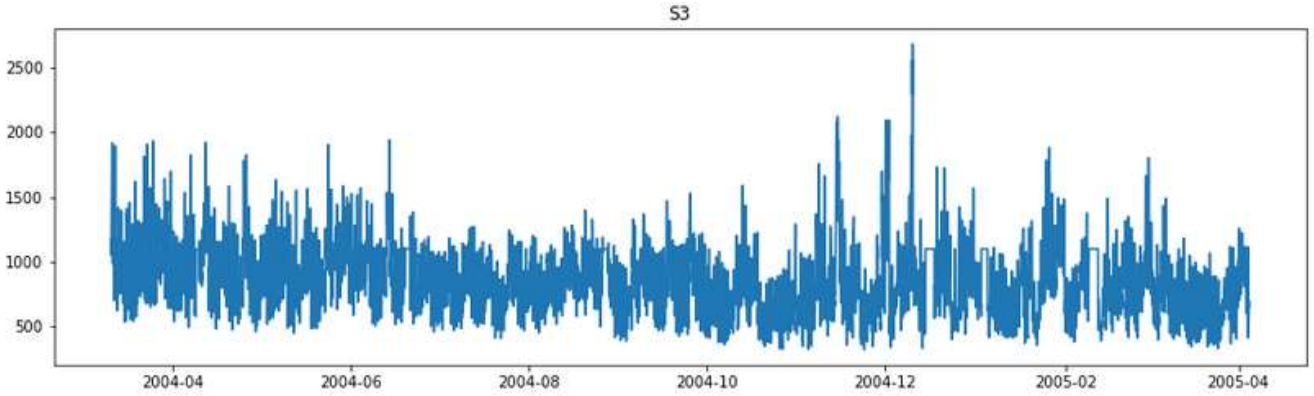
```
fig, axes = plt.subplots(5,1, figsize=(15,24))

axes[0].plot(S1)
axes[0].set_title ('S1')
axes[1].plot(S2)
axes[1].set_title ('S2')
axes[2].plot(S3)
axes[2].set_title ('S3')
axes[3].plot(S4)
axes[3].set_title ('S4')
axes[4].plot(S5)
axes[4].set_title ('S5')
```

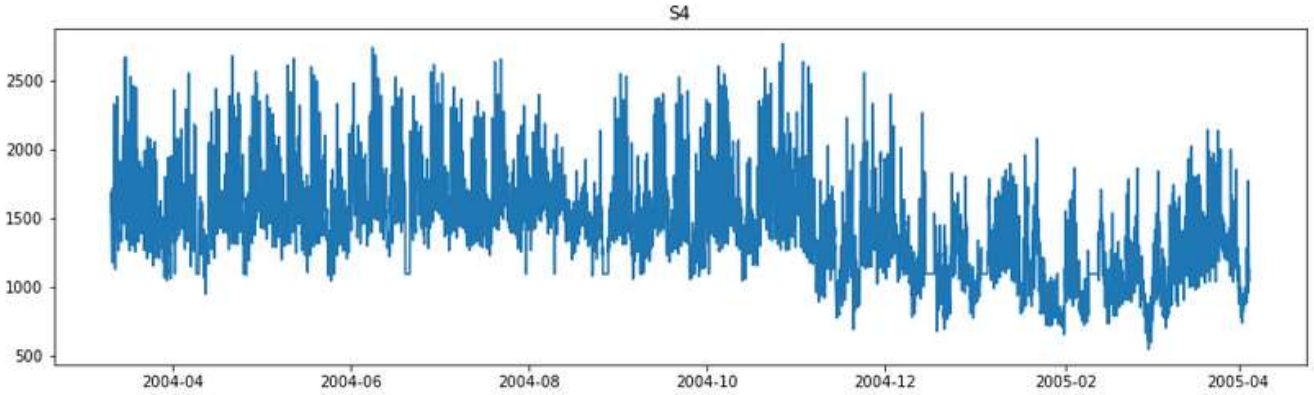


Pollutant “CO” decreases during the summer months, and increases during winter months (Year:2004/05).

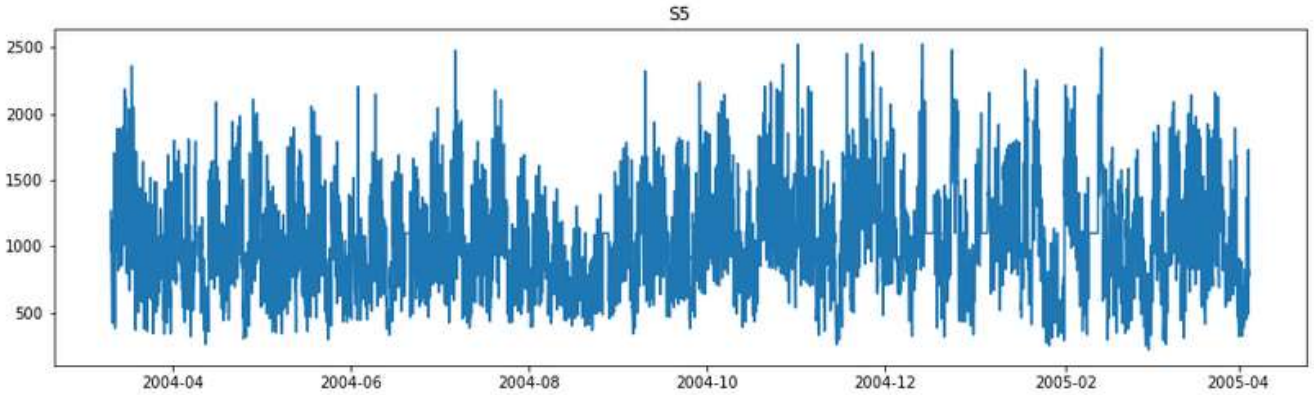
Pollutant “NMHC” remains more or less constant in the summer months, with a sudden drop in Aug-Sept and peaks towards Dec, followed by a gradual decrease (Year:2004/05).



Pollutant “NOx” decreases between Apr-July, remains constant for Aug-Sept and increases from Oct-Jan. It peaks in December (Year:2004/05).



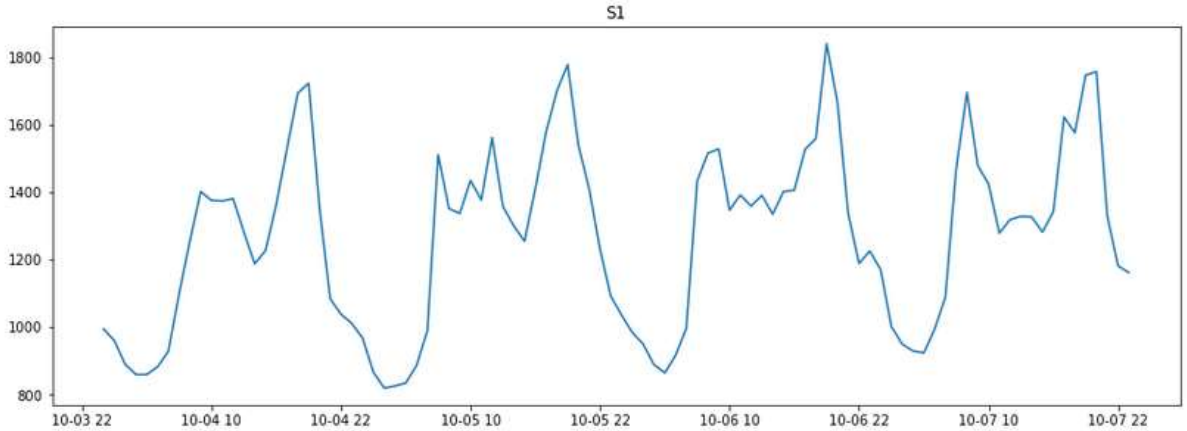
Pollutant “NO2” remains at very high levels(2500) through the year and decreases from Dec-Mar (winter/spring). Then it increases again (Year:2004/05).

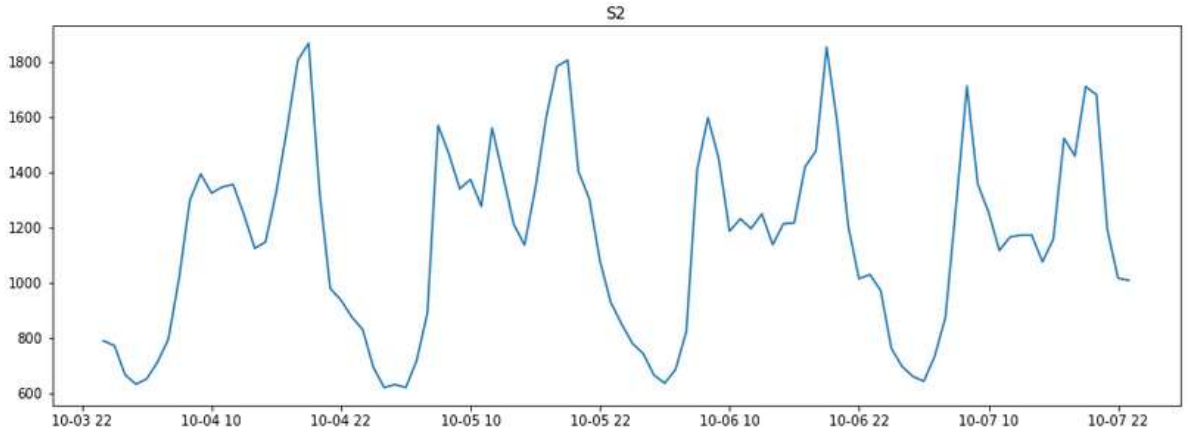


Pollutant “O3” remains at high levels (2000) through the year and decreases in Jul-Sep and Mar-Apr (Year:2004/05).

PLOT 2 : Time series plot for all contaminants over a 4 month period

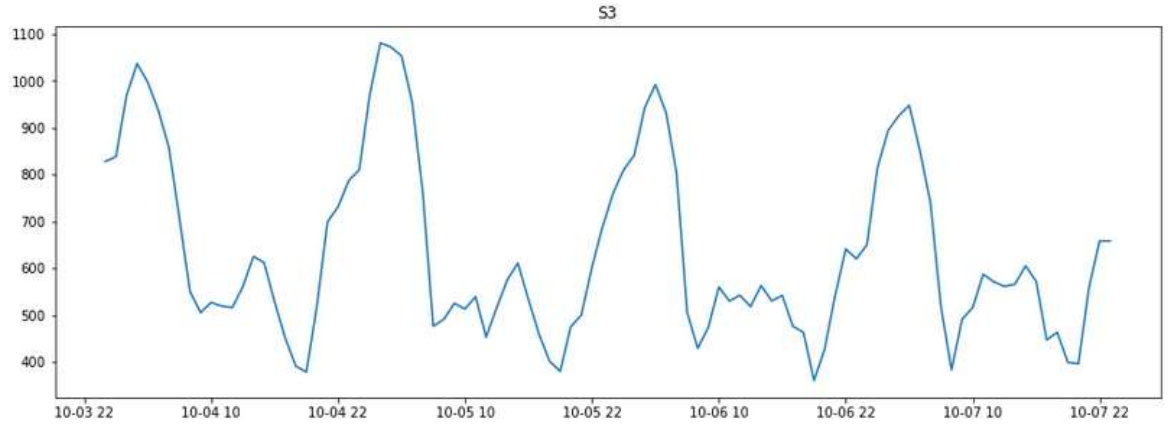
```
fig, axes = plt.subplots(5,1, figsize=(15,30))
axes[0].plot(S1['2004-10-04':'2004-10-07'])
axes[0].set_title ('S1')
axes[1].plot(S2['2004-10-04':'2004-10-07'])
axes[1].set_title ('S2')
axes[2].plot(S3['2004-10-04':'2004-10-07'])
axes[2].set_title ('S3')
axes[3].plot(S4['2004-10-04':'2004-10-07'])
axes[3].set_title ('S4')
axes[4].plot(S5['2004-10-04':'2004-10-07'])
axes[4].set_title ('S5')
```



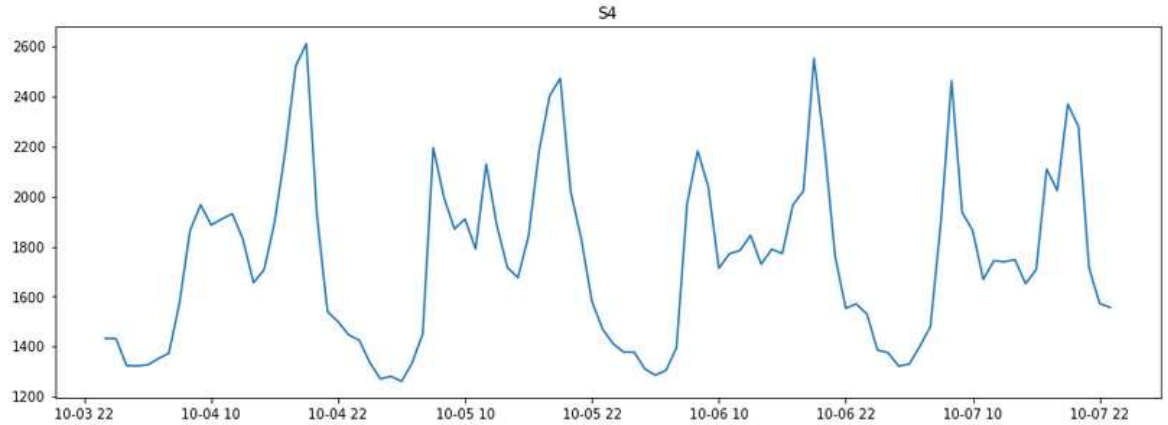


Pollutant “CO” is increasing gradually during the summer months from Mar-Jul (Year:2010).

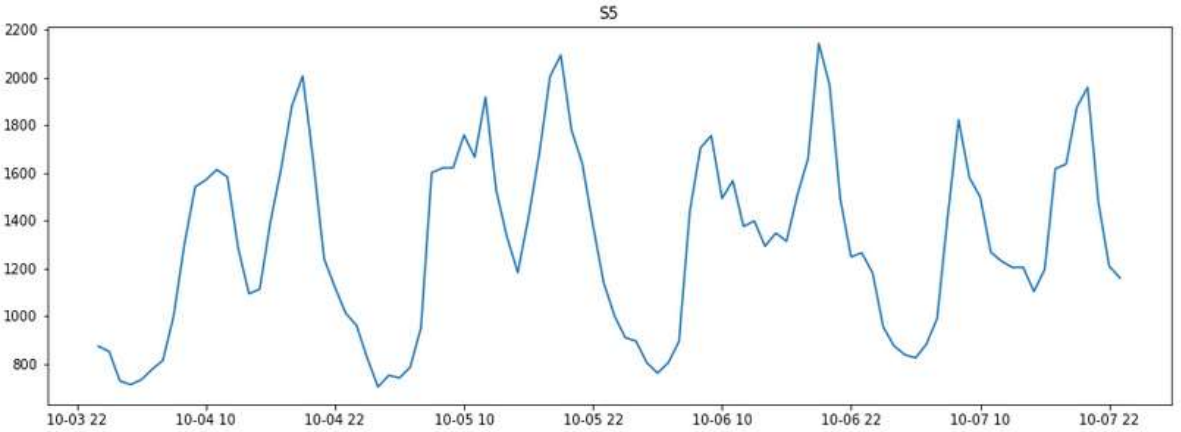
Contaminant “NMHC” is seen to be more or less same during the months Mar-Jul (Year:2010).



Contaminant “NOx” is seen to be decreasing gradually at first between Mar-Jun and then rapidly in July(Year:2010).



Contaminant “NO2” remains more or less at the same high levels (around 2500) during the summer months (Mar-Jul) (Year:2010).



Contaminant “O3” remains at more or less at the same high levels (around 2100) during Mar-Jun, with a slight decrease to (around 1900-1950) during Jul (Year:2010).