Importing Libraries

In [43]:
```
import matplotlib.pyplot as plt
import pandas as pd
import pylab as pl
import numpy as np
import sklearn
%matplotlib inline
```

Loading the Data Set

In [44]:
```
df = pd.read_csv("cpcb_dly_aq_tamil_nadu-2014.csv")
df = df.rename(columns={'RSPM/PM10': 'RSPMorPM10'})
df
```

Out[44]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPMorPM10	PM 2.5
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	12-10-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

2879 rows × 11 columns

Exploring the Data Set

In [45]:
```
print(df.head())
```

```
  Stn Code Sampling Date       State City/Town/Village/Area  \
0       38    01-02-2014  Tamil Nadu                Chennai
1       38    01-07-2014  Tamil Nadu                Chennai
2       38    21-01-2014  Tamil Nadu                Chennai
3       38    23-01-2014  Tamil Nadu                Chennai
4       38    28-01-2014  Tamil Nadu                Chennai

                                Location of Monitoring Station  \
0  Kathivakkam, Municipal Kalyana Mandapam, Chennai
1  Kathivakkam, Municipal Kalyana Mandapam, Chennai
2  Kathivakkam, Municipal Kalyana Mandapam, Chennai
3  Kathivakkam, Municipal Kalyana Mandapam, Chennai
4  Kathivakkam, Municipal Kalyana Mandapam, Chennai

                                  Agency Type of Location   SO2   NO2  \
0  Tamilnadu State Pollution Control Board  Industrial Area  11.0  17.0
1  Tamilnadu State Pollution Control Board  Industrial Area  13.0  17.0
2  Tamilnadu State Pollution Control Board  Industrial Area  12.0  18.0
3  Tamilnadu State Pollution Control Board  Industrial Area  15.0  16.0
4  Tamilnadu State Pollution Control Board  Industrial Area  13.0  14.0

   RSPMorPM10  PM 2.5
0        55.0     NaN
1        45.0     NaN
2        50.0     NaN
3        46.0     NaN
4        42.0     NaN
```

In [46]:
```
print(df.tail())
```

```
     Stn Code Sampling Date       State City/Town/Village/Area  \
2874      773    12-03-2014  Tamil Nadu                Trichy
2875      773    12-10-2014  Tamil Nadu                Trichy
2876      773    17-12-2014  Tamil Nadu                Trichy
2877      773    24-12-2014  Tamil Nadu                Trichy
2878      773    31-12-2014  Tamil Nadu                Trichy

                                Location of Monitoring Station  \
2874  Central Bus Stand, Trichy  Tamilnadu State Pollution Control Board
2875  Central Bus Stand, Trichy  Tamilnadu State Pollution Control Board
2876  Central Bus Stand, Trichy  Tamilnadu State Pollution Control Board
2877  Central Bus Stand, Trichy  Tamilnadu State Pollution Control Board
2878  Central Bus Stand, Trichy  Tamilnadu State Pollution Control Board

                                  Agency  \
2874  Residential, Rural and other Areas
2875  Residential, Rural and other Areas
2876  Residential, Rural and other Areas
2877  Residential, Rural and other Areas
2878  Residential, Rural and other Areas

                                  Type of Location   SO2   NO2  RSPMorPM10  PM 2.5
2874  Residential, Rural and other Areas  15.0  18.0       102.0     NaN
2875  Residential, Rural and other Areas  12.0  14.0        91.0     NaN
2876  Residential, Rural and other Areas  19.0  22.0       100.0     NaN
2877  Residential, Rural and other Areas  15.0  17.0        95.0     NaN
2878  Residential, Rural and other Areas  14.0  16.0        94.0     NaN
```

In [47]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Stn Code                    2879 non-null   int64
 1   Sampling Date               2879 non-null   object
 2   State                       2879 non-null   object
 3   City/Town/Village/Area      2879 non-null   object
 4   Location of Monitoring Station  2879 non-null  object
 5   Agency                      2879 non-null   object
 6   Type of Location            2879 non-null   object
 7   SO2                         2866 non-null   float64
 8   NO2                         2866 non-null   float64
 9   RSPMorPM10                  2875 non-null   float64
 10  PM 2.5                      0 non-null      float64
dtypes: float64(4), int64(1), object(6)
memory usage: 247.5+ KB
None
```

In [48]:
```
print(df.describe())
```

```
          Stn Code          SO2          NO2  RSPMorPM10  PM 2.5
count  2879.000000  2866.000000  2866.000000  2875.000000     0.0
mean    475.756201    11.503138    22.136776    62.494261     NaN
std     277.675577     5.951702     7.128694    31.368745     NaN
min      38.000000     2.000000     5.000000    12.000000     NaN
25%     238.000000     8.000000    17.000000    41.000000     NaN
50%     366.000000    12.000000    22.000000    55.000000     NaN
75%     764.000000    15.000000    25.000000    78.000000     NaN
max     773.000000    49.000000    71.000000   269.000000     NaN
```

Identifying null Values

In [49]:
```
print(df.isnull())
```

```
      Stn Code  Sampling Date  State  City/Town/Village/Area  \
0        False          False  False                   False
1        False          False  False                   False
2        False          False  False                   False
3        False          False  False                   False
4        False          False  False                   False
...        ...            ...    ...                     ...
2874     False          False  False                   False
2875     False          False  False                   False
2876     False          False  False                   False
2877     False          False  False                   False
2878     False          False  False                   False

      Location of Monitoring Station  Agency  Type of Location    SO2   NO2  \
0                              False   False             False  False  False
1                              False   False             False  False  False
2                              False   False             False  False  False
3                              False   False             False  False  False
4                              False   False             False  False  False
...                              ...     ...               ...    ...    ...
2874                           False   False             False  False  False
2875                           False   False             False  False  False
2876                           False   False             False  False  False
2877                           False   False             False  False  False
2878                           False   False             False  False  False

      RSPMorPM10  PM 2.5
0          False    True
1          False    True
2          False    True
3          False    True
4          False    True
...          ...     ...
2874       False    True
2875       False    True
2876       False    True
2877       False    True
2878       False    True

[2879 rows x 11 columns]
```

In [50]:
```
c = df.isnull().sum()
print(c)
```

```
Stn Code                        0
Sampling Date                   0
State                           0
City/Town/Village/Area          0
Location of Monitoring Station  0
Agency                          0
Type of Location                0
SO2                            11
NO2                            13
RSPMorPM10                      4
PM 2.5                       2879
dtype: int64
```

In [51]:
```
print('Total Sum of null values in the Data set = ',c.sum())
```

```
Total Sum of null values in the Data set =  2907
```

In [52]:
```
print(df['RSPMorPM10'].value_counts()) #frequency of values
```

```
47.0    64
41.0    62
43.0    59
51.0    58
40.0    58
        ..
163.0    1
138.0    1
211.0    1
202.0    1
238.0    1
Name: RSPMorPM10, Length: 169, dtype: int64
```

Data Preprocessing - Replacing the null values

In [53]:
```
df.drop_duplicates()
```

Out[53]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPMorPM10	PM 2.5
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	12-10-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

2879 rows × 11 columns

In [54]:
```
df.fillna(0)
```

Out[54]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPMorPM10	PM 2.5
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	0.0
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	0.0
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	0.0
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	0.0
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	0.0
2875	773	12-10-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	0.0
2876	773	17-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	0.0
2877	773	24-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	0.0
2878	773	31-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	0.0

2879 rows × 11 columns

In [55]:
```
df.duplicated()
```

Out[55]:
```
0       False
1       False
2       False
3       False
4       False
        ...
2874    False
2875    False
2876    False
2877    False
2878    False
Length: 2879, dtype: bool
```

Data Normalization

In [57]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df['Values_standardized'] = scaler.fit_transform(df[['RSPMorPM10']])
```

In [62]:
```
scaler = StandardScaler()
df['Values_standardized'] = scaler.fit_transform(df[['RSPMorPM10']])
df
```

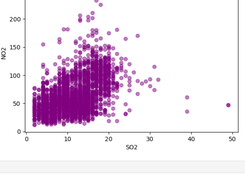
Out[62]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPMorPM10	PM 2.5	Values_standardized
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN	-0.238950
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN	-0.557794
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN	-0.398372
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN	-0.525910
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN	-0.653447
...	...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN	1.259617
2875	773	12-10-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN	0.908889
2876	773	17-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN	1.195848
2877	773	24-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN	1.036426
2878	773	31-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN	1.004542

2879 rows × 12 columns

Data Visualisation

In [63]:
```
# Scatter Plot:
plt.scatter(df['SO2'], df['RSPMorPM10'], alpha=0.5, color='purple')
plt.title('RSPM/PM10')
plt.xlabel('SO2')
plt.ylabel('NO2')
plt.show()
```



In [ ]:
```
```