TI35_DSBDA_5th_AirQuality

January 30, 2025

[33]:
```
import pandas as pd
```

[35]:
```
df1=pd.read_csv("/home/bcl07/air_quality.csv",encoding="ISO-8859-1")
```

```
/tmp/ipykernel_6496/2504295896.py:1: DtypeWarning: Columns (0) have mixed types.
Specify dtype option on import or set low_memory=False.
  df1=pd.read_csv("/home/bcl07/air_quality.csv",encoding="ISO-8859-1")
```

[37]:
```
df1
```

[37]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN					
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN					
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN					
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN					
…	…	…	…	…	…					
435737	SAMP	24-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board					
435738	SAMP	29-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board					
435739	NaN	NaN	andaman-and-nicobar-islands	NaN	NaN					
435740	NaN	NaN	Lakshadweep	NaN	NaN					
435741	NaN	NaN	Tripura	NaN	NaN					

```
1                          Industrial Area   3.1   7.0    NaN  NaN
2       Residential, Rural and other Areas   6.2  28.5    NaN  NaN
3       Residential, Rural and other Areas   6.3  14.7    NaN  NaN
4                          Industrial Area   4.7   7.5    NaN  NaN
…                                        …   …    …    …  …
435737                                RIRUO  22.0  50.0  143.0  NaN
435738                                RIRUO  20.0  46.0  171.0  NaN
435739                                  NaN   NaN   NaN    NaN  NaN
435740                                  NaN   NaN   NaN    NaN  NaN
435741                                  NaN   NaN   NaN    NaN  NaN

              location_monitoring_station  pm2_5        date
0                                     NaN    NaN  1990-02-01
1                                     NaN    NaN  1990-02-01
2                                     NaN    NaN  1990-02-01
3                                     NaN    NaN  1990-03-01
4                                     NaN    NaN  1990-03-01
…                                       …    …          …
435737  Inside Rampal Industries,ULUBERIA    NaN  2015-12-24
435738  Inside Rampal Industries,ULUBERIA    NaN  2015-12-29
435739                                NaN    NaN         NaN
435740                                NaN    NaN         NaN
435741                                NaN    NaN         NaN

[435742 rows x 13 columns]
```

[39]:
```
df1.columns
```

[39]:
```
Index(['stn_code', 'sampling_date', 'state', 'location', 'agency', 'type',
       'so2', 'no2', 'rspm', 'spm', 'location_monitoring_station', 'pm2_5',
       'date'],
      dtype='object')
```

[42]:
```
df1.isnull().sum()
```

[42]:
```
stn_code                       144077
sampling_date                       3
state                               0
location                            3
agency                         149481
type                             5393
so2                             34646
no2                             16233
rspm                            40222
spm                            237387
location_monitoring_station     27491
pm2_5                          426428
```

```
date                                7
dtype: int64
```

[47]:
```
df1['so2'] = df1['so2'].astype('float32')
df1['no2'] = df1['no2'].astype('float32')
df1['rspm'] = df1['rspm'].astype('float32')
df1['spm'] = df1['spm'].astype('float32')
df1['date'] = df1['date'].astype('string')

df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   stn_code                      291665 non-null  object
 1   sampling_date                 435739 non-null  object
 2   state                         435742 non-null  object
 3   location                      435739 non-null  object
 4   agency                        286261 non-null  object
 5   type                          430349 non-null  object
 6   so2                           401096 non-null  float32
 7   no2                           419509 non-null  float32
 8   rspm                          395520 non-null  float32
 9   spm                           198355 non-null  float32
 10  location_monitoring_station   408251 non-null  object
 11  pm2_5                         9314 non-null    float64
 12  date                          435735 non-null  string
dtypes: float32(4), float64(1), object(7), string(1)
memory usage: 36.6+ MB
```

[48]:
```
df1=df1.drop_duplicates()
```

[49]:
```
df1.isna().sum()
```

[49]:
```
stn_code                      144077
sampling_date                      3
state                              0
location                           3
agency                        149466
type                            5357
so2                            34632
no2                            16222
rspm                           40035
spm                           236908
location_monitoring_station    27303
```

```
pm2_5                          425754
date                                7
dtype: int64
```

[51]:
```
percent_missing = df1.isnull().sum() * 100 / len(df1)
```

[52]:
```
percent_missing.sort_values(ascending=False)
```

[52]:
```
pm2_5                          97.859185
spm                            54.453097
agency                         34.354630
stn_code                       33.115973
rspm                            9.202010
so2                             7.960135
location_monitoring_station     6.275571
no2                             3.728613
type                            1.231302
date                            0.001609
sampling_date                   0.000690
location                        0.000690
state                           0.000000
dtype: float64
```

[54]:
```
df1=df1.drop(['stn_code',␣
 ↪'agency','sampling_date','location_monitoring_station','pm2_5'], axis=1)
```

[55]:
```
df1.head()
```

[55]:

	state	location	type	so2	no2	rspm	spm	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	1990-03-01

[56]:
```
df1.columns
```

[56]:
```
Index(['state', 'location', 'type', 'so2', 'no2', 'rspm', 'spm', 'date'],
dtype='object')
```

[57]:
```
col_var = ['state', 'location', 'type','date']
col_num = ['so2','no2','rspm','spm']
```

[58]:
```
for col in df1.columns:
    if df1[col].dtype == 'object' or df1[col].dtype == 'string':
        df1[col] = df1[col].fillna(df1[col].mode()[0])
    else:
        df1[col] = df1[col].fillna(df1[col].mean())
```

[59]:
```
df1.isna().sum()
```

[59]:
```
state       0
location    0
type        0
so2         0
no2         0
rspm        0
spm         0
date        0
dtype: int64
```

[60]:
```
df1
```

[60]:

	state	location	type	so2	no2	rspm
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.800000	17.400000	108.871712
1	Andhra Pradesh	Hyderabad	Industrial Area	3.100000	7.000000	108.871712
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.200000	28.500000	108.871712
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.300000	14.700000	108.871712
4	Andhra Pradesh	Hyderabad	Industrial Area	4.700000	7.500000	108.871712
…	…	…	…	…	…	…
435737	West Bengal	ULUBERIA	RIRUO	22.000000	50.000000	143.000000
435738	West Bengal	ULUBERIA	RIRUO	20.000000	46.000000	171.000000
435739	andaman-and-nicobar-islands	Guwahati	Residential, Rural and other Areas	10.830467	25.823299	108.871712
435740	Lakshadweep	Guwahati	Residential, Rural and other Areas	10.830467	25.823299	108.871712
435741	Tripura	Guwahati				

```
435741  Residential, Rural and other Areas  10.830467  25.823299  108.871712

                spm        date
0        220.774796  1990-02-01
1        220.774796  1990-02-01
2        220.774796  1990-02-01
3        220.774796  1990-03-01
4        220.774796  1990-03-01
…               …           …
435737   220.774796  2015-12-24
435738   220.774796  2015-12-29
435739   220.774796  2015-03-19
435740   220.774796  2015-03-19
435741   220.774796  2015-03-19

[435068 rows x 8 columns]
```

[61]:
```
df1.isna().sum()
```

[61]:
```
state       0
location    0
type        0
so2         0
no2         0
rspm        0
spm         0
date        0
dtype: int64
```

[63]:
```
subSet1 = df1[['state', 'type']]
subSet2 = df1[['state','location']]
```

[64]:
```
subSet1.head()
```

[64]:
```
             state                                type
0  Andhra Pradesh  Residential, Rural and other Areas
1  Andhra Pradesh                 Industrial Area
2  Andhra Pradesh  Residential, Rural and other Areas
3  Andhra Pradesh  Residential, Rural and other Areas
4  Andhra Pradesh                 Industrial Area
```

[65]:
```
subSet2.head()
```

[65]:
```
             state   location
0  Andhra Pradesh  Hyderabad
1  Andhra Pradesh  Hyderabad
2  Andhra Pradesh  Hyderabad
```

```
3  Andhra Pradesh  Hyderabad
4  Andhra Pradesh  Hyderabad
```

[66]:
```
concatenated_df=pd.concat([subSet1,subSet2],axis=1)
```

[67]:
```
concatenated_df
```

[67]:
```
                            state                                type  \
0                   Andhra Pradesh  Residential, Rural and other Areas
1                   Andhra Pradesh                      Industrial Area
2                   Andhra Pradesh  Residential, Rural and other Areas
3                   Andhra Pradesh  Residential, Rural and other Areas
4                   Andhra Pradesh                      Industrial Area
…                              …                                   …
435737                 West Bengal                               RIRUO
435738                 West Bengal                               RIRUO
435739  andaman-and-nicobar-islands  Residential, Rural and other Areas
435740                 Lakshadweep  Residential, Rural and other Areas
435741                     Tripura  Residential, Rural and other Areas

                            state   location
0                   Andhra Pradesh  Hyderabad
1                   Andhra Pradesh  Hyderabad
2                   Andhra Pradesh  Hyderabad
3                   Andhra Pradesh  Hyderabad
4                   Andhra Pradesh  Hyderabad
…                              …         …
435737                 West Bengal   ULUBERIA
435738                 West Bengal   ULUBERIA
435739  andaman-and-nicobar-islands   Guwahati
435740                 Lakshadweep   Guwahati
435741                     Tripura   Guwahati

[435068 rows x 4 columns]
```

[68]:
```
def remove_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    threshold = 1.5 * IQR
    outlier_mask = (column < Q1 - threshold) | (column > Q3 + threshold)
    return column[~outlier_mask]
```

[69]:
```
df1.columns
```

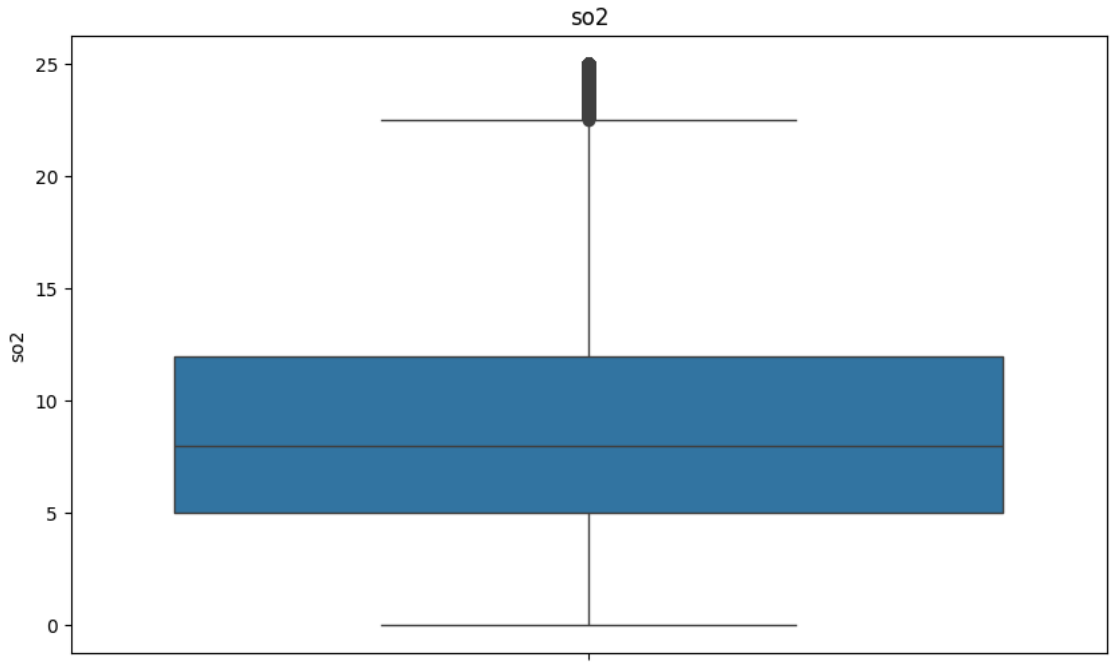
[69]:
```
Index(['state', 'location', 'type', 'so2', 'no2', 'rspm', 'spm', 'date'],
dtype='object')
```

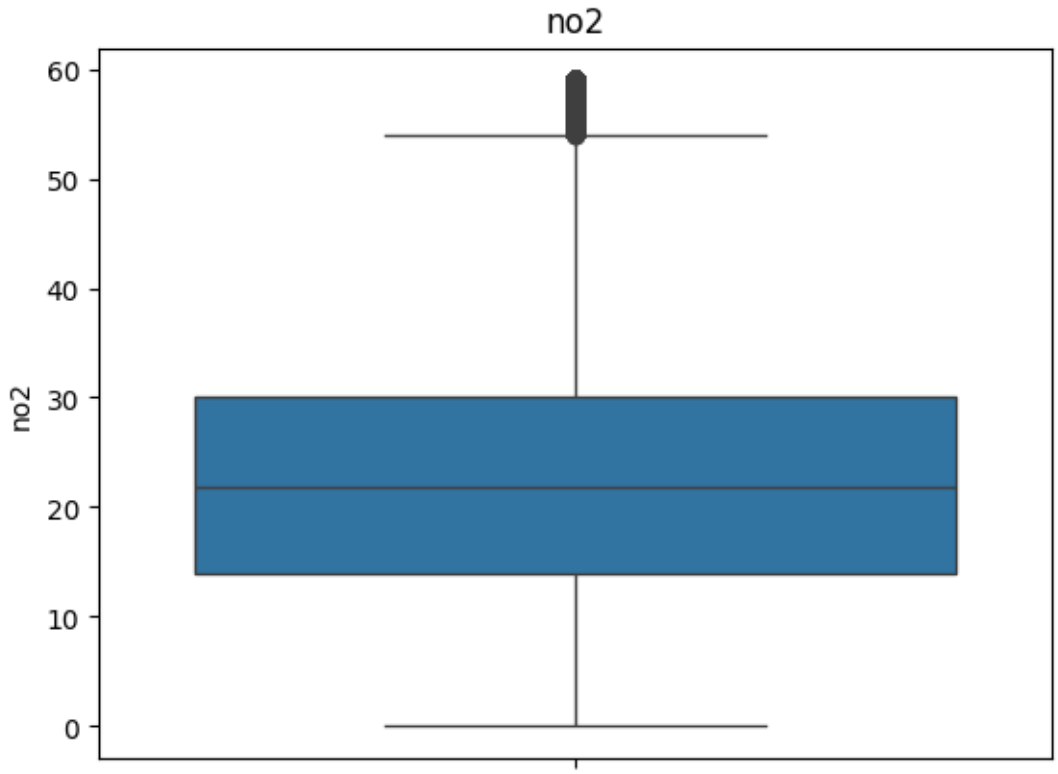
[70]:
```
col_name = ['so2', 'no2', 'rspm', 'spm']
for col in col_name:
    df1[col] = remove_outliers(df1[col])
```

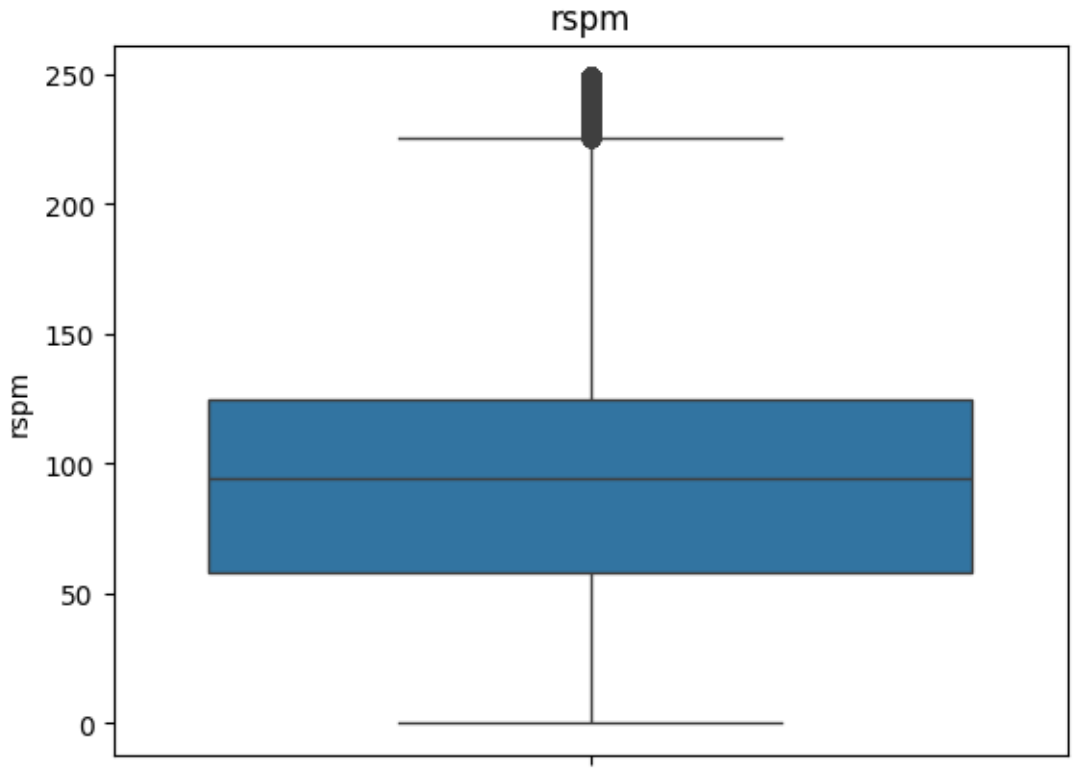
[71]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

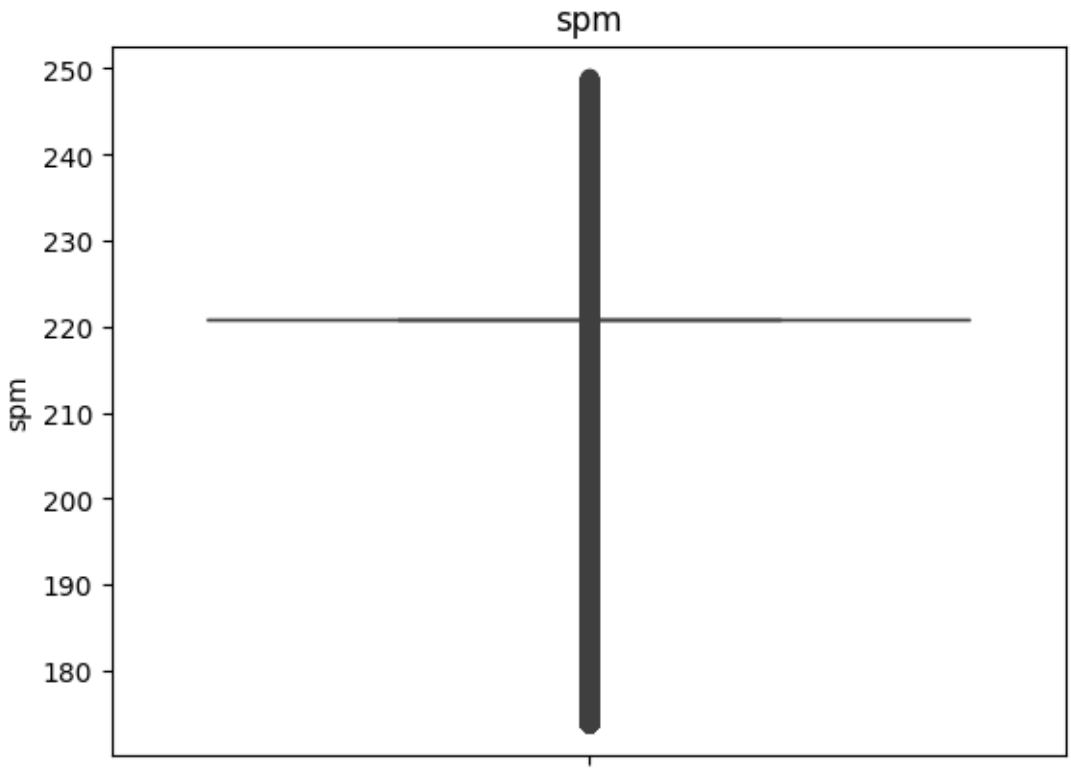
[73]:
```
plt.figure(figsize=(10, 6))  # Adjust the figure size if needed

for col in col_name:
    sns.boxplot(data=df1[col])
    plt.title(col)
    plt.show()
```




no2
60
50
40
30
20
10
0
no2


rspm
250
200
150
100
50
0
rspm



[77]:
```
from sklearn.preprocessing import LabelEncoder

col_label= ['state','location','type']
# Initialize LabelEncoder

encoder = LabelEncoder()
# Iterate over columns
for col in df1.columns:
        # Fit and transform the column
        df1[col] = encoder.fit_transform(df1[col])
```

[78]:
```
df1
```

[78]:

	state	location	type	so2	no2	rspm	spm	date
0	0	114	6	446	1489	2030	464	213
1	0	114	1	197	250	2030	464	213
2	0	114	6	790	3096	2030	464	213
3	0	114	6	823	1144	2030	464	214
4	0	114	1	427	301	2030	464	214

```
…        …        …    …    …    …    …   …    …
435737     35       282     3  2888  5307  2534  464  5059
435738     35       282     3  2809  5113  3098  464  5064
435739     36       100     6  1638  2696  2030  464  4779
435740     17       100     6  1638  2696  2030  464  4779
435741     31       100     6  1638  2696  2030  464  4779

[435068 rows x 8 columns]
```

[ ]: