In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:

```
df = pd.read_csv('dataset.csv')
df1 = df.copy()
```

```
C:\Users\jaswa\AppData\Local\Temp\ipykernel_18956\4111772239.py:1: DtypeWarning: Columns (0) have mixed types. Spec
ify dtype option on import or set low_memory=False.
  df = pd.read_csv('dataset.csv')
```

In [5]:

```
df1.head()
```

Out[5]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	2/1/1990
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	2/1/1990
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	2/1/1990
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	3/1/1990
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	3/1/1990

In [6]:

```
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  float64
 7   no2                          419509 non-null  float64
 8   rspm                         395520 non-null  float64
 9   spm                          198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

In [7]:

```
df1.isnull().sum()
```

Out[7]:

```
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
```

In [10]:

```
df['location']
```

Out[10]:

```
0         Hyderabad
1         Hyderabad
2         Hyderabad
3         Hyderabad
4         Hyderabad
            ...
435737     ULUBERIA
435738     ULUBERIA
435739          NaN
435740          NaN
435741          NaN
Name: location, Length: 435742, dtype: object
```

In [11]:

```
rep = {'location': {r'Vishakhapatnam': 'Visakhapatnam', }}
df1.replace(rep, inplace = True)
```

In [12]:

```
df1
```

Out[12]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
435737	SAMP	24-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	22.0	50.0	143.0	NaN	Inside Rampal Industries,ULUBERIA	NaN	1:
435738	SAMP	29-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	20.0	46.0	171.0	NaN	Inside Rampal Industries,ULUBERIA	NaN	1:
435739	NaN	NaN	andaman-and-nicobar-islands	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
435740	NaN	NaN	Lakshadweep	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
435741	NaN	NaN	Tripura	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

435742 rows × 13 columns

In [13]:

```
df1['agency'].value_counts()
df1['type'].value_counts()
```

Out[13]:

```
Residential, Rural and other Areas    179014
Industrial Area                        96091
Residential and others                 86791
Industrial Areas                       51747
Sensitive Area                          8980
Sensitive Areas                         5536
RIRUO                                   1304
Sensitive                                495
Industrial                               233
Residential                              158
Name: type, dtype: int64
```

In [20]:

```
#Dropping null values
df1 = df1.dropna(axis = 0, subset = ['type'])
```

In [21]:

```
df1 = df1.dropna(axis = 0, subset = ['location'])
```

In [22]:

```
df1 = df1.dropna(axis = 0, subset = ['so2'])
```

In [23]:

```
df1 = df1.dropna(axis = 0, subset = ['no2'])
```

In [43]:

```
df1 = df1.dropna(axis = 0, subset = ['spm'])
```

In [44]:

```
df1 = df1.dropna(axis = 0, subset = ['rspm'])
```

In [45]:

```
df1.isnull().sum()
```

Out[45]:

```
stn_code                     102511
sampling_date                     0
state                             0
location                          0
agency                       102511
type                              0
so2                               0
no2                               0
rspm                              0
spm                               0
location_monitoring_station     909
pm2_5                        141732
date                              0
dtype: int64
```

In [46]:

```
df1.head()
```

Out[46]:

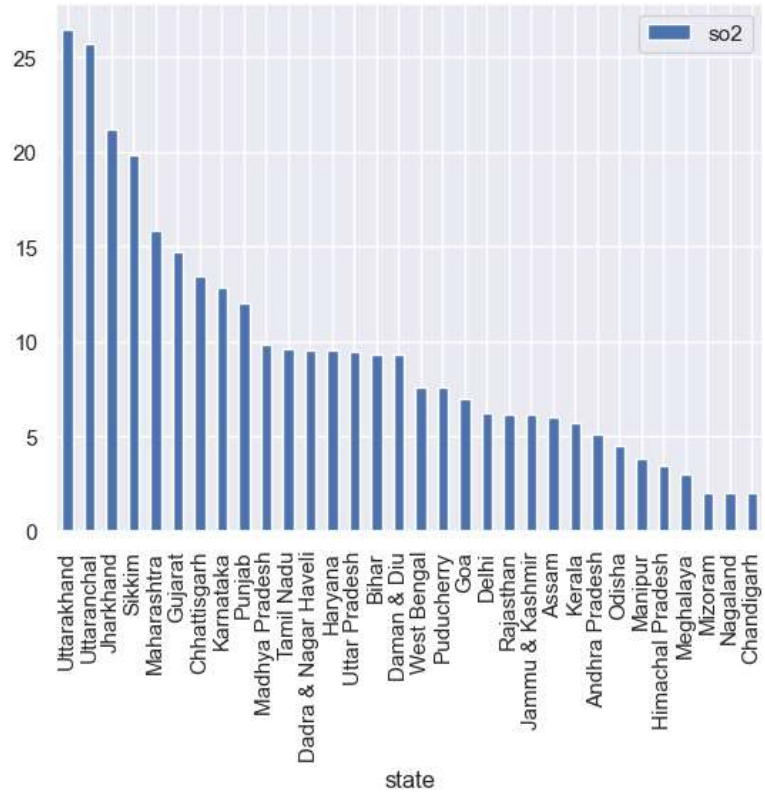
	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
1036	95.0	1/2/2004	Andhra Pradesh	Hyderabad	Andhra Pradesh State Pollution Control Board	Industrial Area	12.9	55.8	143.7	370.7	C.I.T.D., Balanagar, Plot no. A1 to A8, IDA, H...	NaN	2/1/2004
1037	95.0	1/5/2004	Andhra Pradesh	Hyderabad	Andhra Pradesh State Pollution Control Board	Industrial Area	10.4	48.9	124.7	285.7	C.I.T.D., Balanagar, Plot no. A1 to A8, IDA, H...	NaN	5/1/2004
1038	95.0	1/9/2004	Andhra Pradesh	Hyderabad	Andhra Pradesh State Pollution Control Board	Industrial Area	7.6	50.1	88.0	221.3	C.I.T.D., Balanagar, Plot no. A1 to A8, IDA, H...	NaN	9/1/2004
1039	95.0	1/12/2004	Andhra Pradesh	Hyderabad	Andhra Pradesh State Pollution Control Board	Industrial Area	7.3	48.5	82.7	186.7	C.I.T.D., Balanagar, Plot no. A1 to A8, IDA, H...	NaN	12/1/2004
1040	95.0	16-01-04	Andhra Pradesh	Hyderabad	Andhra Pradesh State Pollution Control Board	Industrial Area	6.8	110.3	122.3	270.7	C.I.T.D., Balanagar, Plot no. A1 to A8, IDA, H...	NaN	1/16/2004

In [62]:

```
df1[['so2', 'state']].groupby(['state']).median().sort_values("so2", ascending = False).plot.bar()
```

Out[62]:

```
<AxesSubplot: xlabel='state'>
```

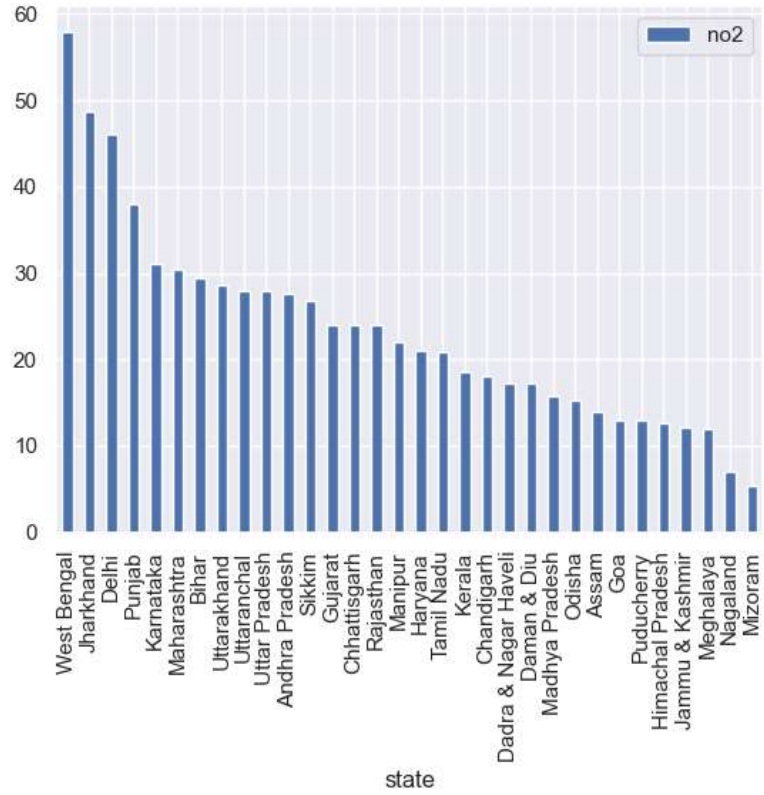


In [63]:

```
df1[['no2', 'state']].groupby(['state']).median().sort_values("no2", ascending = False).plot.bar()
```

Out[63]:

```
<AxesSubplot: xlabel='state'>
```

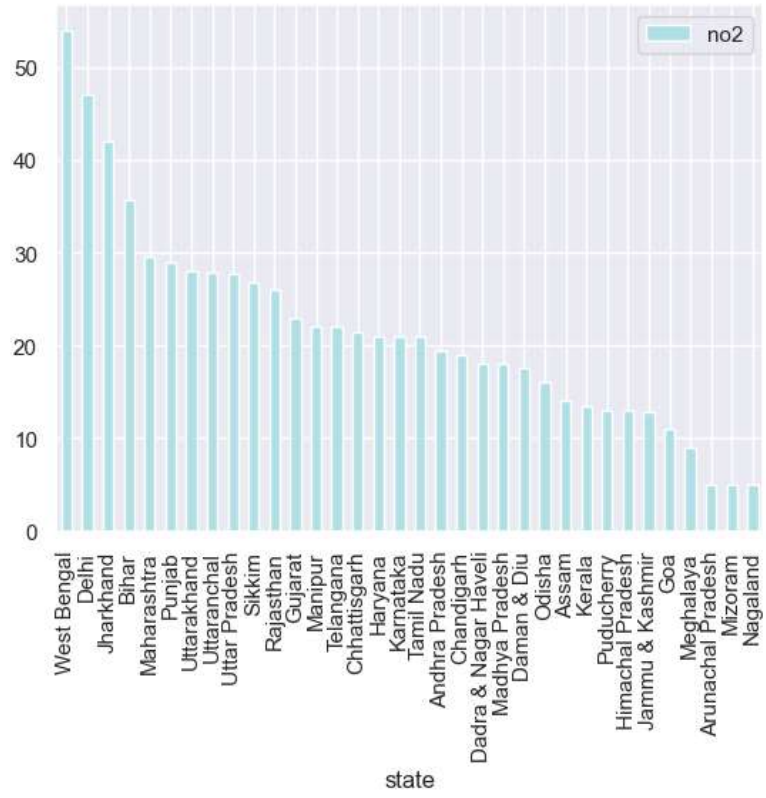


In [42]:

```
df1[['no2', 'state']].groupby(['state']).median().sort_values("no2", ascending = False).plot.bar(color = 'powderblue')
```

Out[42]:

```
<AxesSubplot: xlabel='state'>
```

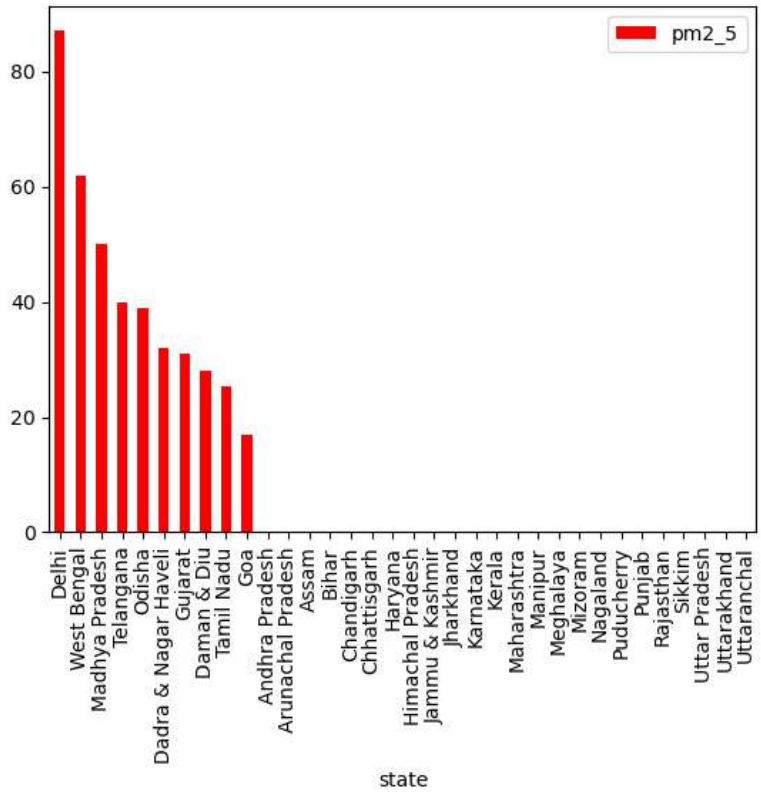


In [35]:

```
df1[['pm2_5', 'state']].groupby(['state']).median().sort_values("pm2_5", ascending = False).plot.bar(color = 'r')
```

Out[35]:

```
<AxesSubplot: xlabel='state'>
```

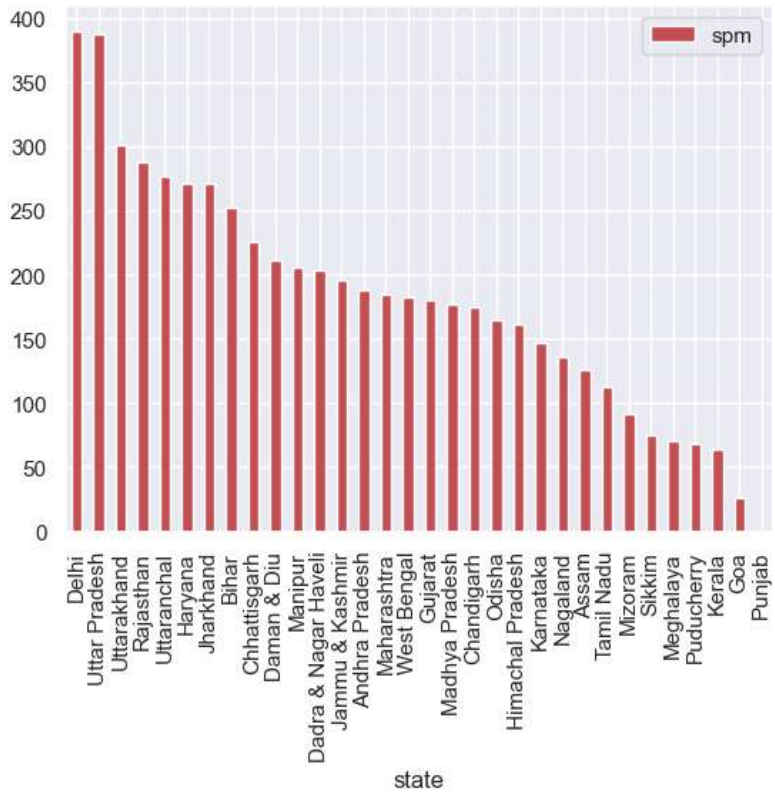


In [47]:

```
df1[['spm', 'state']].groupby(['state']).median().sort_values("spm", ascending = False).plot.bar(color = 'r')
```

Out[47]:

```
<AxesSubplot: xlabel='state'>
```

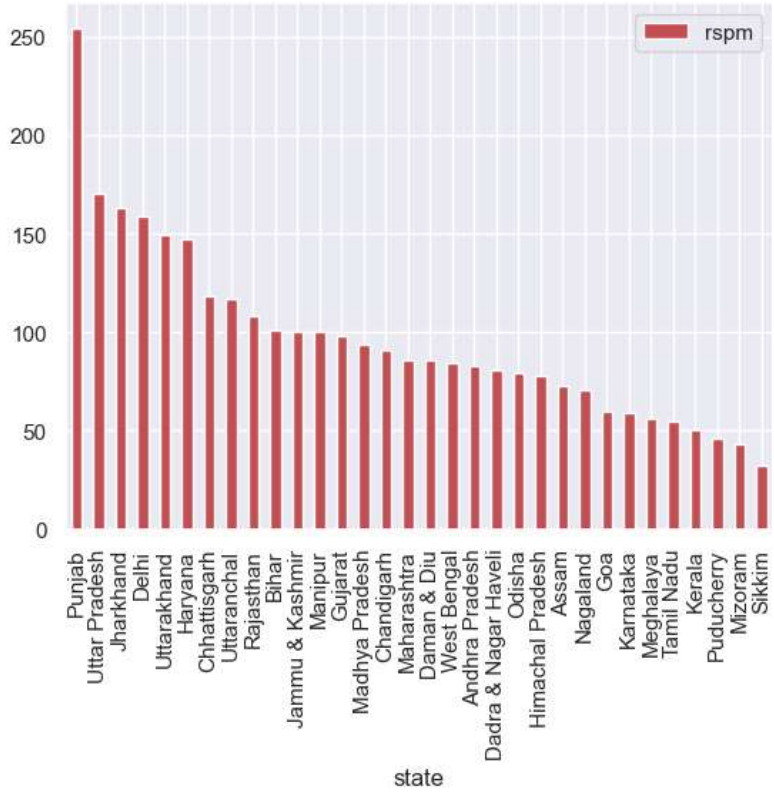


In [52]:

```
df1[['rspm', 'state']].groupby(['state']).median().sort_values("rspm", ascending = False).plot.bar(color = 'r')
```

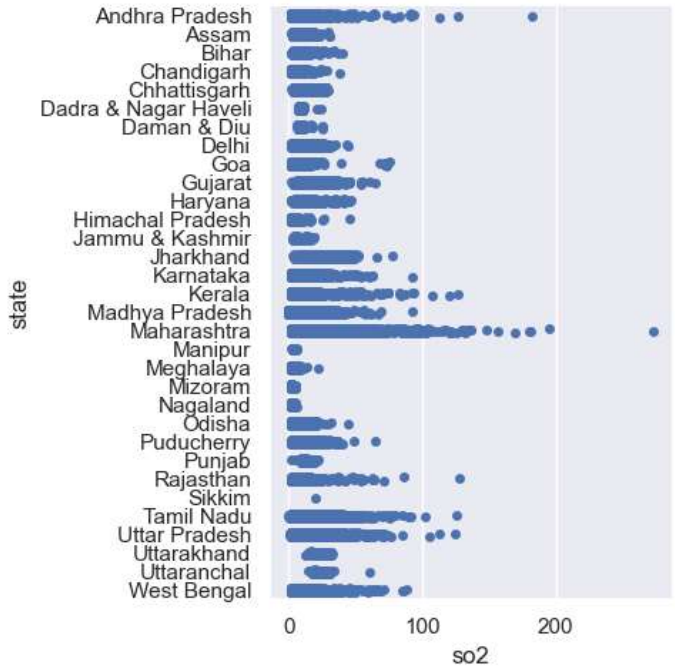
Out[52]:

```
<AxesSubplot: xlabel='state'>
```



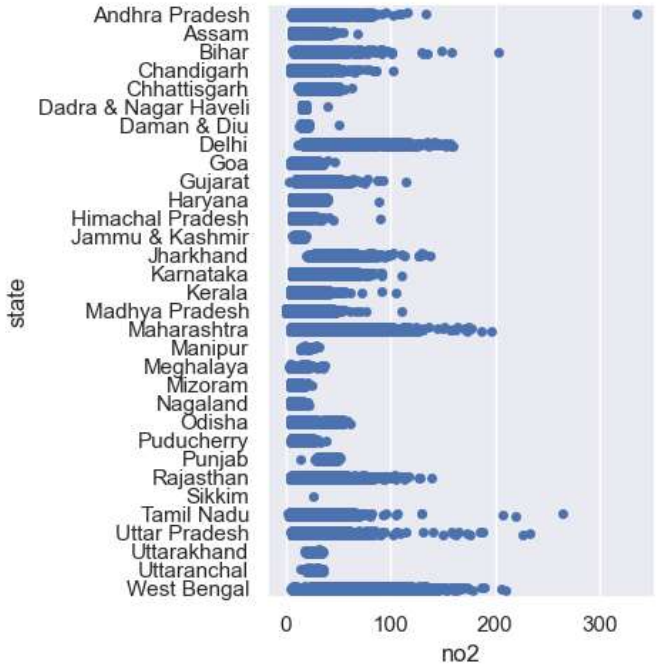
In [57]:

```
sns.catplot(x = 'so2',y = 'state',data = df1)
plt.show()
```



In [58]:

```
sns.catplot(x = 'no2',y = 'state',data = df1)
plt.show()
```

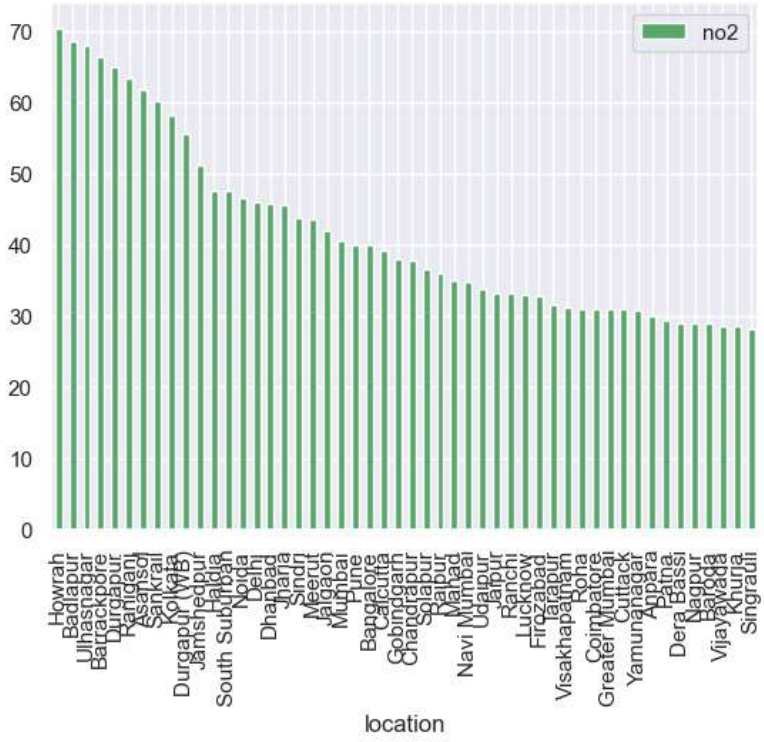


In [59]:

```
df1[['no2', 'location']].groupby(['location']).median().sort_values("no2", ascending = False).head(50).plot.bar(color = 'g')
```

Out[59]:

```
<AxesSubplot: xlabel='location'>
```

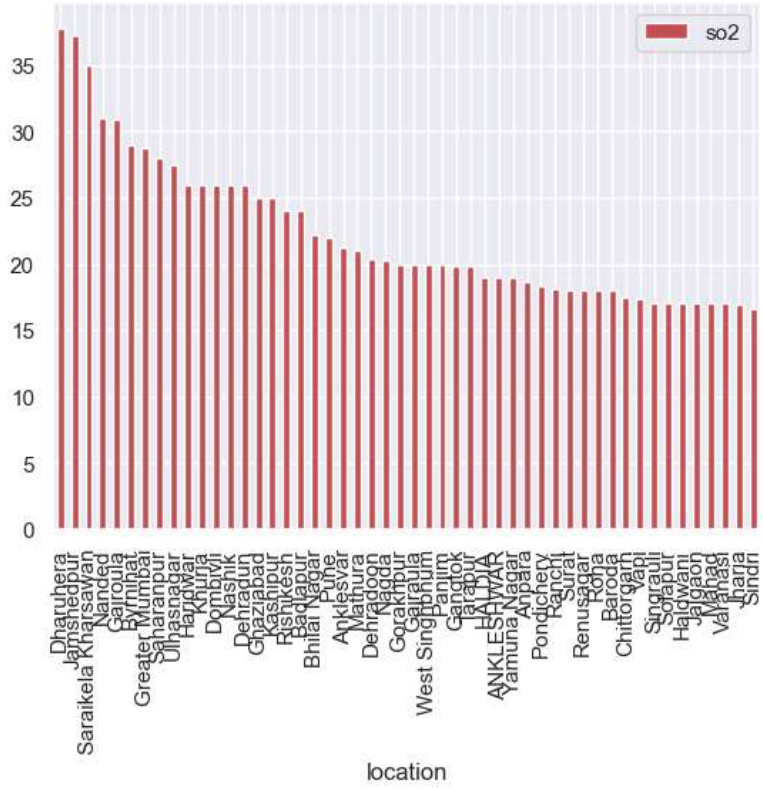


In [61]:

```
df[['so2', 'location']].groupby(['location']).median().sort_values("so2", ascending = False).head(50).plot.bar(color = 'r')
```

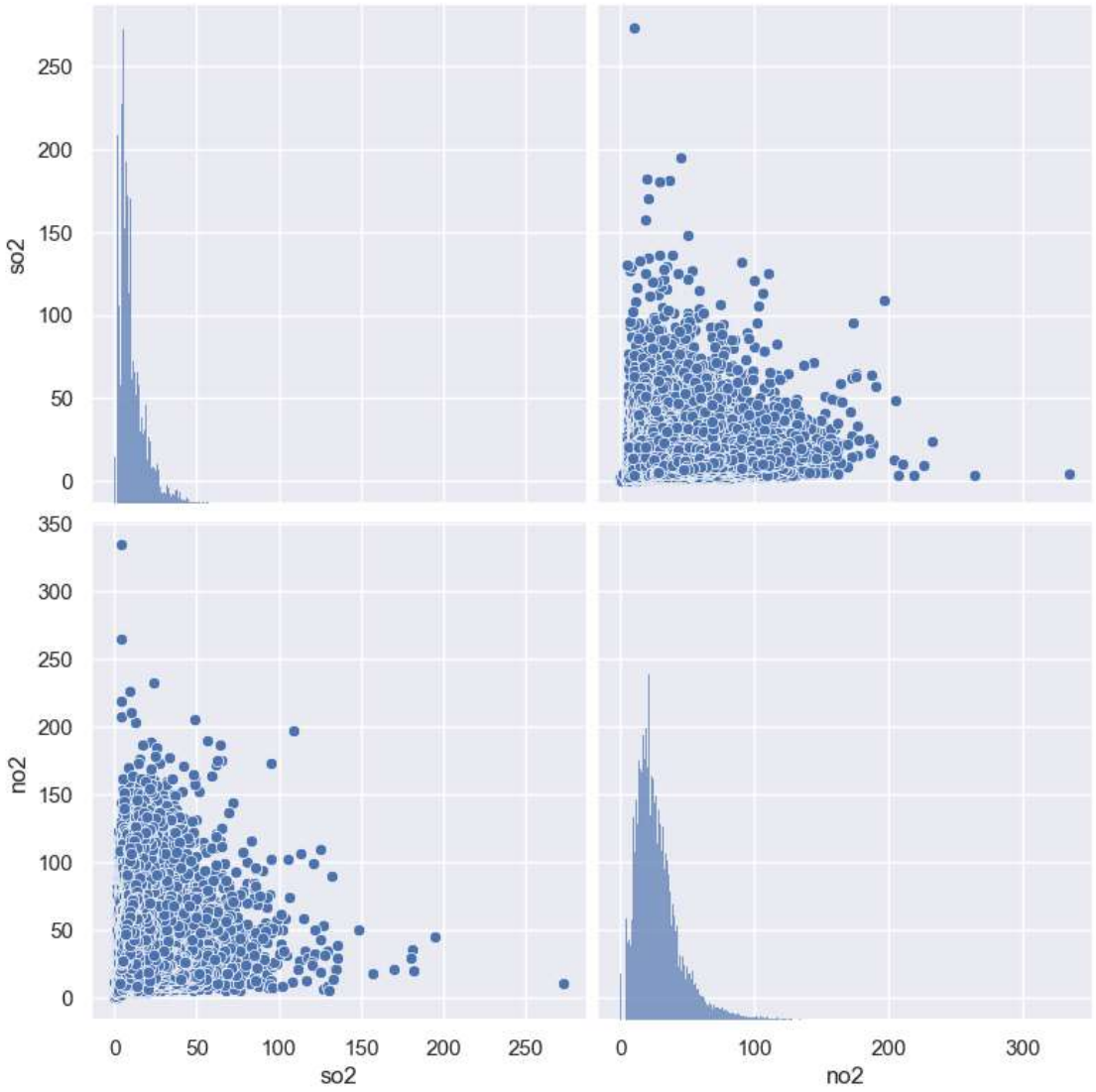
Out[61]:

```
<AxesSubplot: xlabel='location'>
```



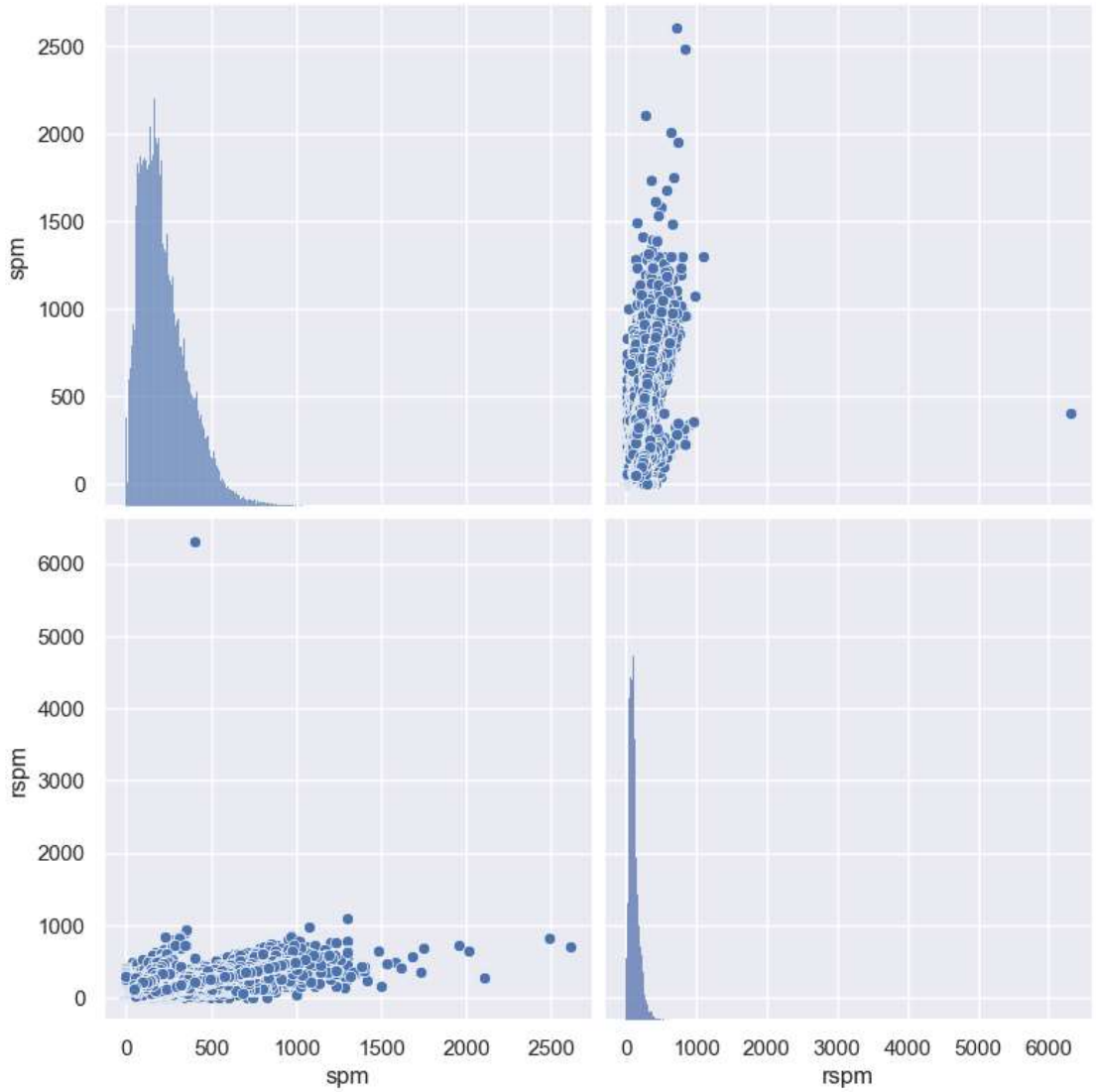
In [68]:

```
sns.set()
cols = ['so2', 'no2']
sns.pairplot(df1[cols], height = 4)
plt.show()
```



In [73]:

```
sns.set()
cols = ['spm', 'rspm']
sns.pairplot(df1[cols], height = 4)
plt.show()
```



In [ ]: