Name: Janhavi Vijay Pawar Roll No. TI56

Group B: Assignments based on Data Analytics using Python

Perform the following operations using Python on the Air quality and Heart Diseases data sets

a. Data cleaning

b. Data integration

c. Data transformation

d. Error correcting

e. Data model building

Air Pollution In India

Air pollution in India is a serious issue with the major sources being fuelwood and biomass burning, fuel adulteration, vehicle emission and traffic congestion. In autumn and winter months, large scale crop residue burning in agriculture fields – a low cost alternative to mechanical tilling – is a major source of smoke, smog and particulate pollution. India has a low per capita emissions of greenhouse gases but the country as a whole is the third largest after China and the United States.A 2013 study on non-smokers has found that Indians have 30% lower lung function compared to Europeans.

The Air (Prevention and Control of Pollution) Act was passed in 1981 to regulate air pollution and there have been some measurable improvements. However, the 2016 Environmental Performance Index ranked India 141 out of 180 countries.

Pollution is turning the Taj Mahal yellow, despite efforts by the Indian government to control air contamination around the poignant 17th century monument and keep it shimmering white, a parliamentary committee has said.

In a report to parliament this week, the standing committee on transport, tourism and culture said airborne particles were being deposited on the monument’s white marble, giving it a yellow tinge.

The monument, in the northern city of Agra about four hours drive south of the capital, was built by Mughal emperor Shah Jahan as a mausoleum for his wife Mumtaz Mahal.

Authorities have made various attempts in the past to keep the area around the Taj Mahal pollution free, including setting up an air pollution monitoring station in Agra.

But the committee said that while air pollutants such as sulphur dioxide and nitrous oxide gases were generally within permissible limits, “suspended particulate matter” had been recorded at high levels except during the rainy season.

It suggested a clay pack treatment that is non-corrosive and non-abrasive be carried out to remove deposits on the marble. “The committee recommends that while undertaking any conservation activity at the Taj Mahal, abundant cautions should be taken to retain the original glory of the shimmering white marble used in this.”

Attracting around 20,000 visitors every day, the monument was completed in 1648 after 17 years of construction by 20,000 workers.

In [1]:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.plotly as py
%matplotlib inline
plt.rcParams['figure.figsize'] = (10, 7)

# Warnings
import warnings
warnings.filterwarnings('ignore')

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files i

import os
print(os.listdir("../input"))

# Any results you write to the current directory are saved as output.
```

```
['data.csv']
```

In [2]:

```
data=pd.read_csv('../input/data.csv',encoding = "ISO-8859-1")
data.head()
```

Out[2]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	l
0	150	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	
1	151	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	
2	152	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	
3	150	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	
4	151	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	

In [3]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
stn_code                        291665 non-null object
sampling_date                   435739 non-null object
state                           435742 non-null object
location                        435739 non-null object
agency                          286261 non-null object
type                            430349 non-null object
so2                             401096 non-null float64
no2                             419509 non-null float64
rspm                            395520 non-null float64
spm                             198355 non-null float64
location_monitoring_station     408251 non-null object
pm2_5                           9314 non-null float64
date                            435735 non-null object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

In [4]:

```
replacements = {
   'state': {
      r'Uttaranchal': 'Uttarakhand',
   }
}

data.replace(replacements, regex=True, inplace=True)
```

What is sulfur dioxide?

Sulfur dioxide is a gas. It is invisible and has a nasty, sharp smell. It reacts easily with other substances to form harmful compounds, such as sulfuric acid, sulfurous acid and sulfate particles.

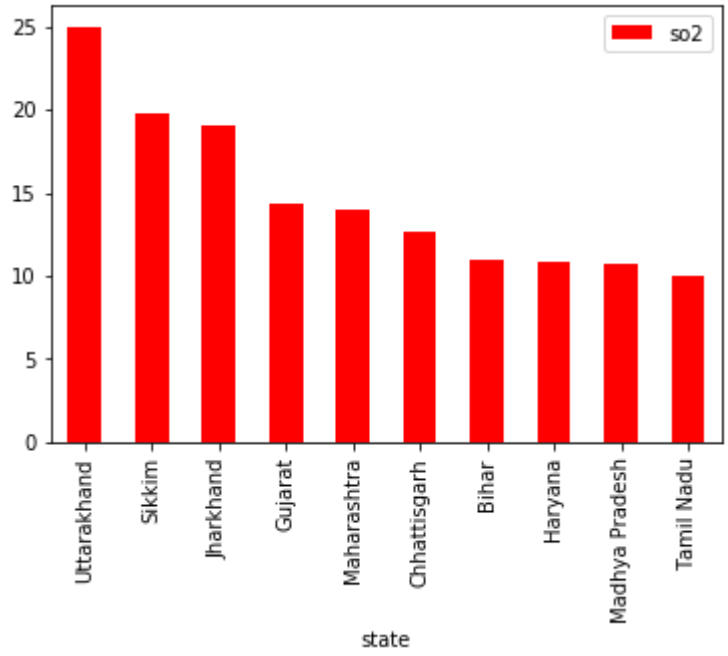
About 99% of the sulfur dioxide in air comes from human sources. The main source of sulfur dioxide in the air is industrial activity that processes materials that contain sulfur, eg the generation of electricity from coal, oil or gas that contains sulfur. Some mineral ores also contain sulfur, and sulfur dioxide is released when they are processed. In addition, industrial activities that burn fossil fuels containing sulfur can be important sources of sulfur dioxide.

Sulfur dioxide is also present in motor vehicle emissions, as the result of fuel combustion. In the past, motor vehicle exhaust was an important, but not the main, source of sulfur dioxide in air. However, this is no longer the case.

TOP 10

In [5]:

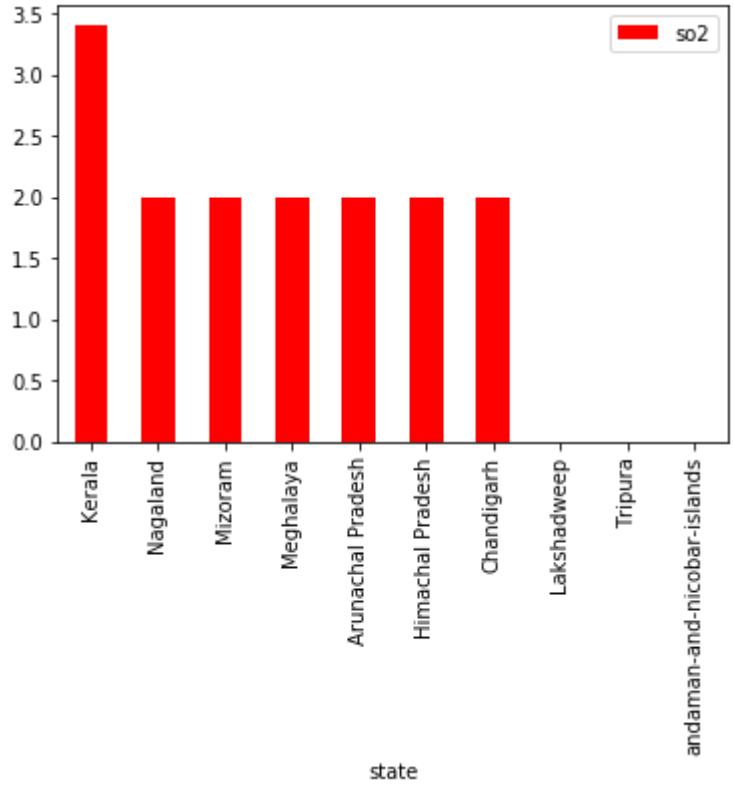
```
data[['so2','state']].groupby(["state"]).median().sort_values(by='so2',ascending=False).hea
plt.show()
```



BOTTOM 10

In [6]:

```
data[['so2','state']].groupby(["state"]).median().sort_values(by='so2',ascending=False).tai
plt.show()
```



How does sulfur dioxide affect human health?

Sulfur dioxide affects human health when it is breathed in. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. The effects of sulfur dioxide are felt very quickly and most people would feel the worst symptoms in 10 or 15 minutes after breathing it in.

Those most at risk of developing problems if they are exposed to sulfur dioxide are people with asthma or similar conditions.

What's being done to manage sulfur dioxide?

Because of the adverse health effects of high levels of sulfur dioxide, the Australian Government has taken steps to manage and reduce the amount of sulfur dioxide produced. These include:

- implementing national fuel quality standards;
- supporting the implementation of tighter vehicle emission standards; and
- promoting alternative fuels.

What is Nitrogen Dioxide (NO2)

As for PM and O3, the evidence on NO2 and health comes from different sources of i nformation, including observational epidemiology, controlled human exposures to po llutants and animal toxicology. The observational data are derived from studies ou tdoors where NO2 is one component of the complex mixture of different pollutants f ound in ambient air and from studies of NO2 exposure indoors where its sources inc lude unvented combustion appliances. Interpretation of evidence on NO2 exposures o utdoors is complicated by the fact that in most urban locations, the nitrogen oxid es that yield NO2 are emitted primarily by motor vehicles, making it a strong indi cator of vehicle emissions (including other unmeasured pollutants emitted by these sources). NO2 (and other nitrogen oxides) is also a precursor for a number of harm ful secondary air pollutants, including nitric acid, the nitrate part of secondary inorganic aerosols and photo oxidants (including ozone). The situation is also com plicated by the fact that photochemical reactions take some time (depending on the composition of the atmosphere and meteorological parameters) and air can travel so me distance before secondary pollutants are generated. These relationships are sho wn schematically in Figure 1.
Figure 1: Simplified relationship of nitrogen oxides emissions with formation of N O2 and other harmful reaction products including O3 and PM Simplified relationship of nitrogen oxides emissions with formation of NO2 and other harmful reaction prod ucts including O3 and PM

Health risks from nitrogen oxides may potentially result from NO2 itself or its re action products including O3 and secondary particles. Epidemiological studies of N O2 exposures from outdoor air are limited in being able to separate these effects. Additionally, NO2 concentrations closely follow vehicle emissions in many situatio ns so that NO2 levels are generally a reasonable marker of exposure to traffic rel ated emissions.

Given these complex relationships, findings of multivariate models that include NO 2 and other pollutants need cautious interpretation. While multi-pollutant models have been routinely applied to various forms of observational data, they may mis-specify underlying relationships. Even models that include only NO2 and PM, NO2 an d O3, or NO2, PM and O3 do not reflect the interrelationships among these pollutan ts. Statistical models considering interactions must be based on a strong a priori hypothesis about the nature of these interactions to allow their interpretation. W ith these constraints in mind, the working group recommended against using regress ion coefficients for NO2 from regression models for the purpose of quantitative ri sk assessment.

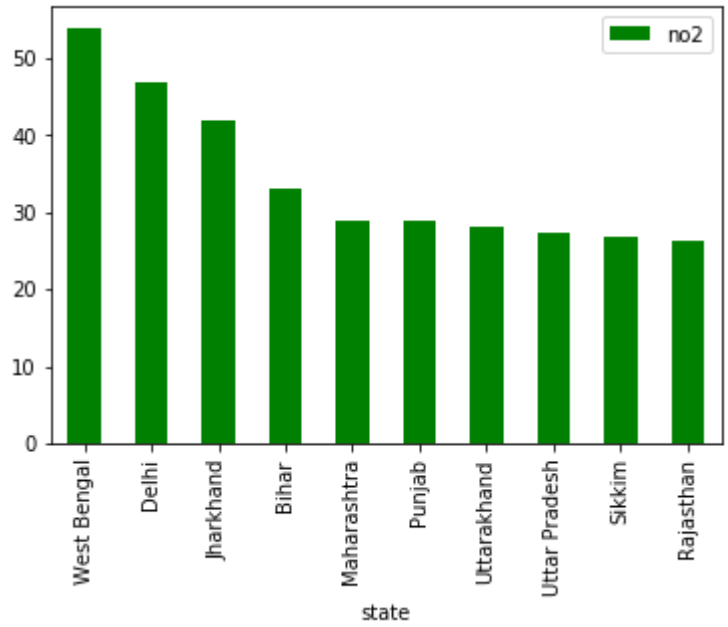
Evidence of the health effects of NO2 by itself thus comes largely from toxicologi cal studies and from observational studies on NO2 exposure indoors. The studies of outdoor NO2 may be most useful under the following circumstances:

- Evidence for NO2 effects assessed at fixed levels of exposure to other pollutants
- Evidence for modification of the effect of PM by NO2, possibly indicating a potential consequence of HNO3 vapour and/or PM nitrate.

TOP 10

In [7]:

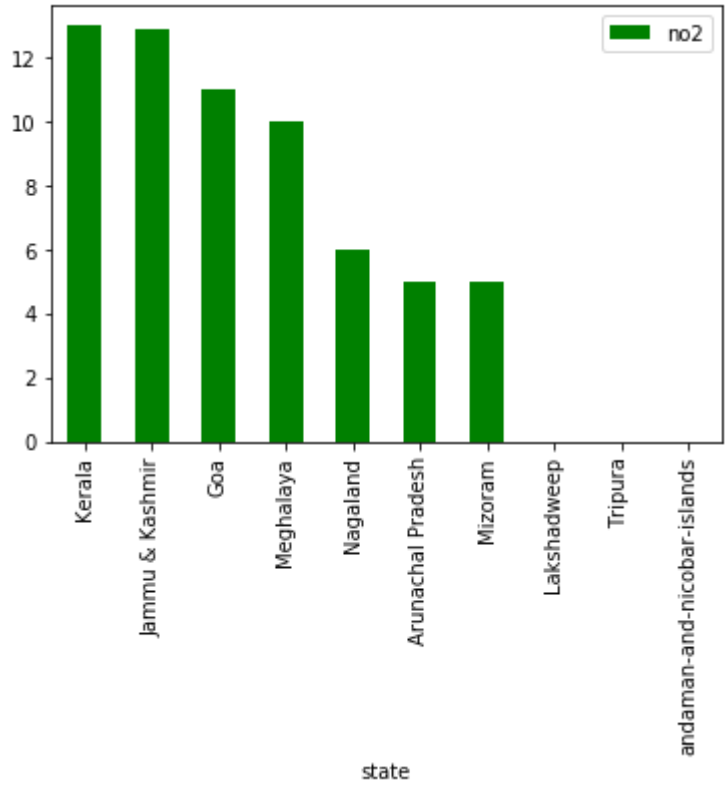
```
data[['no2','state']].groupby(["state"]).median().sort_values(by='no2',ascending=False).hea
plt.show()
```



BOTTOM 10

In [8]:

```
data[['no2','state']].groupby(["state"]).median().sort_values(by='no2',ascending=False).tai
plt.show()
```



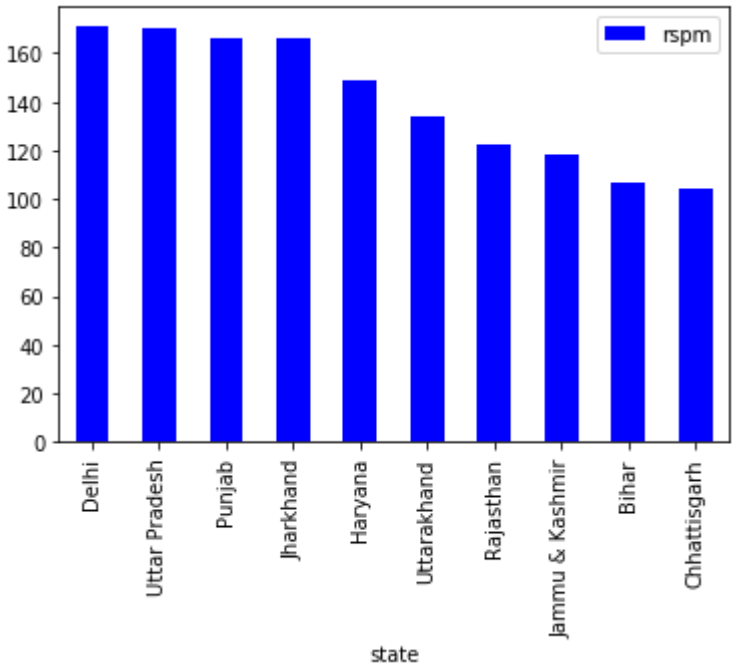
**RSPM **

RSPM is that fraction of TSPM which is readily inhaled by humans through their respiratory system and in general, considered as particulate matter with their diameter (aerodynamic) less than 2.5 micrometers. Larger particles would be filtered in the nasal duct.

TOP 10

In [9]:

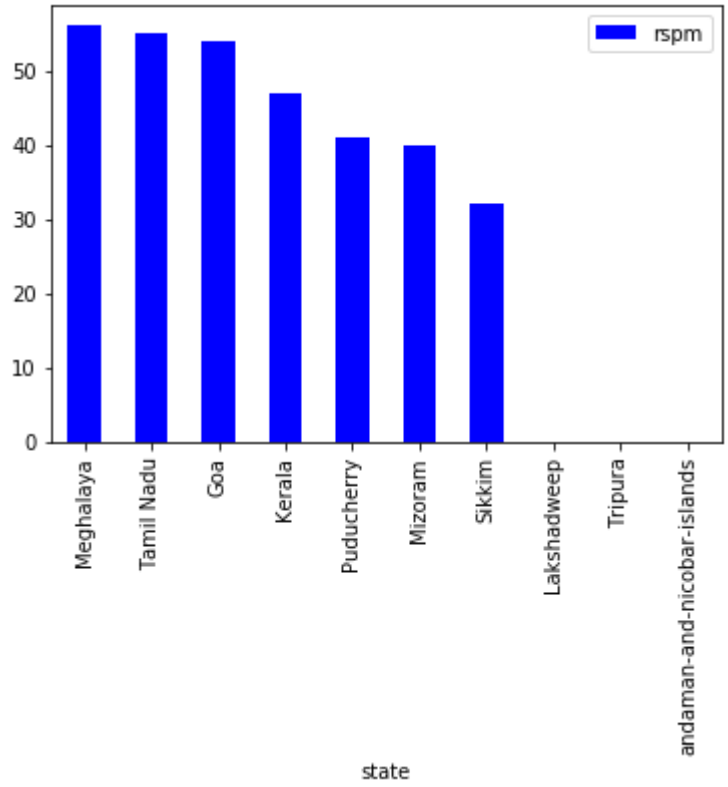
```
data[['rspm','state']].groupby(["state"]).median().sort_values(by='rspm',ascending=False).h
plt.show()
```



BOTTOM 10

In [10]:

```
data[['rspm','state']].groupby(["state"]).median().sort_values(by='rspm',ascending=False).t
plt.show()
```



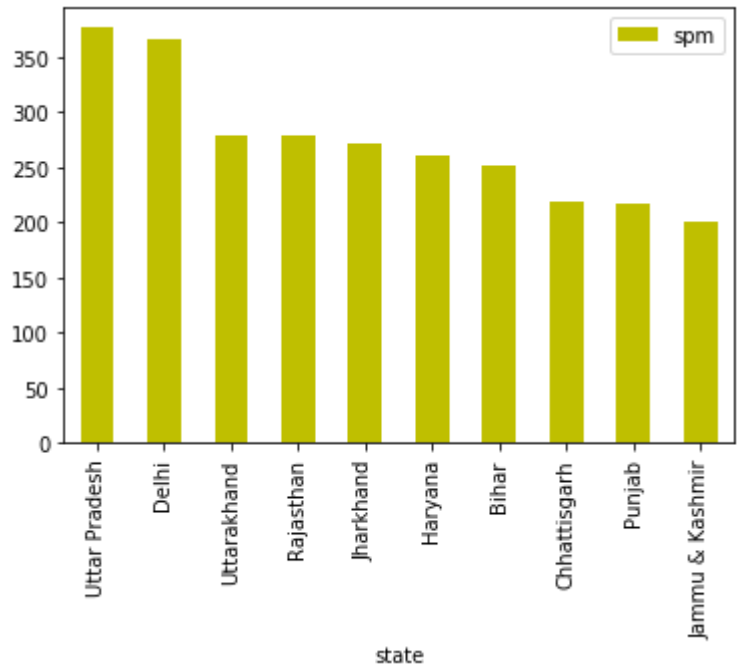
SPM

Suspended Particulate Matter (SPM) are microscopic solid or liquid matter suspended in Earth's atmosphere. The term aerosol commonly refers to the particulate/air mixture, as opposed to the particulate matter alone.[3] Sources of particulate matter can be natural or anthropogenic. They have impacts on climate and precipitation that adversely affect human health.

TOP 10

In [11]:

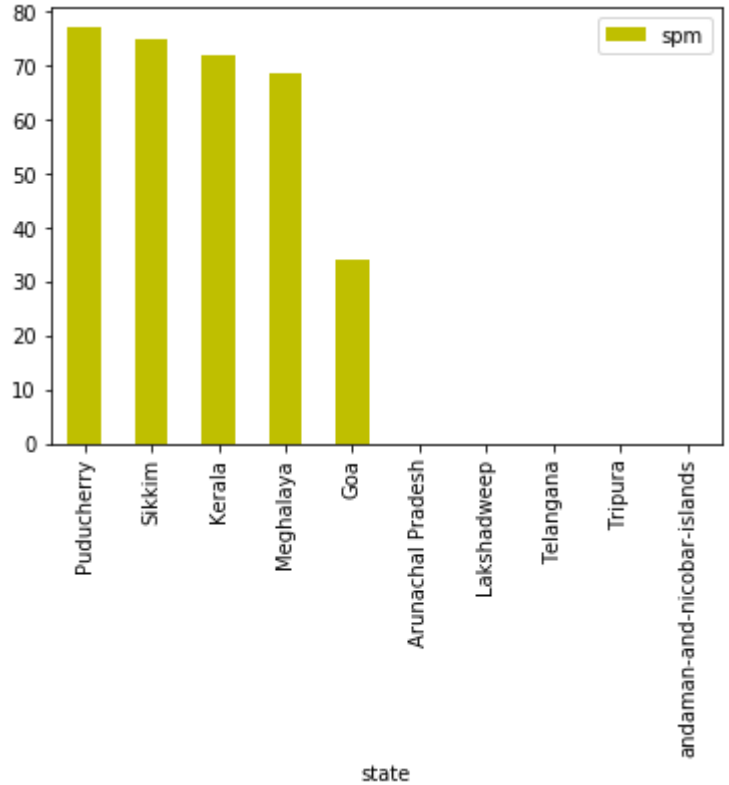
```
data[['spm','state']].groupby(["state"]).median().sort_values(by='spm',ascending=False).hea
plt.show()
```



BOTTAM 10

In [12]:

```
data[['spm','state']].groupby(["state"]).median().sort_values(by='spm',ascending=False).tai
plt.show()
```

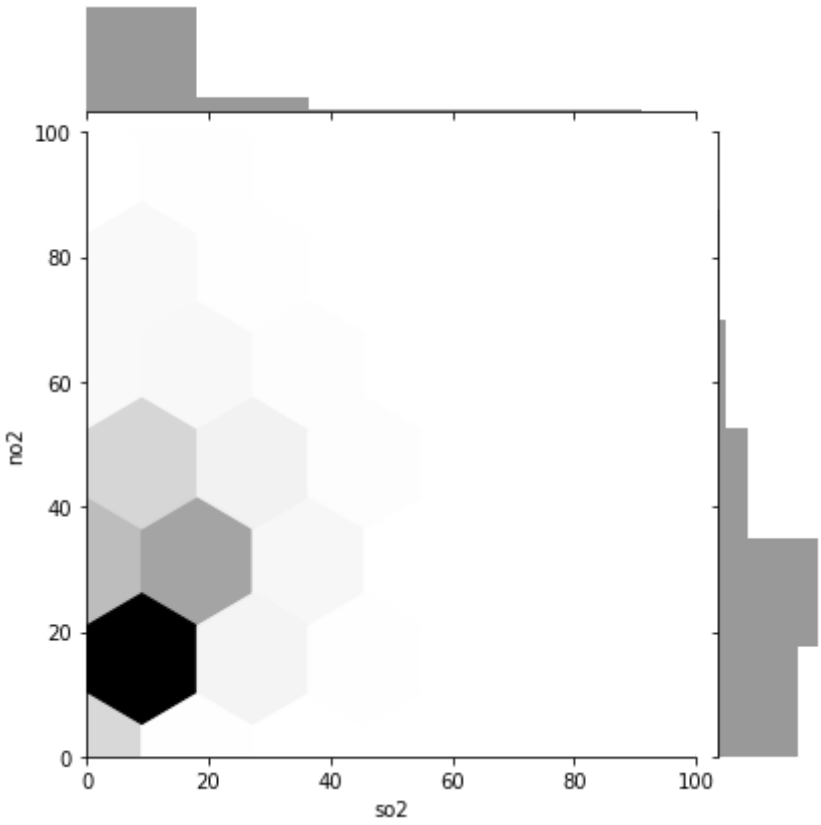


In [13]:

```
#Exploring relationship between proportion of Sulphur dioxide & Nitrogen dioxide
#sns.lmplot(x='so2',y='no2',data=data)
sns.jointplot(x='so2', y='no2', data=data,kind='hex',color='k',xlim={0,100}, ylim={0,100})
```

Out[13]:

```
<seaborn.axisgrid.JointGrid at 0x7f7a999be9b0>
```



In [14]:

```
data['date'] = pd.to_datetime(data['date'],format='%Y-%m-%d') # date parse
data['year'] = data['date'].dt.year # year
data['year'] = data['year'].fillna(0.0).astype(int)
data = data[(data['year']>0)]
```

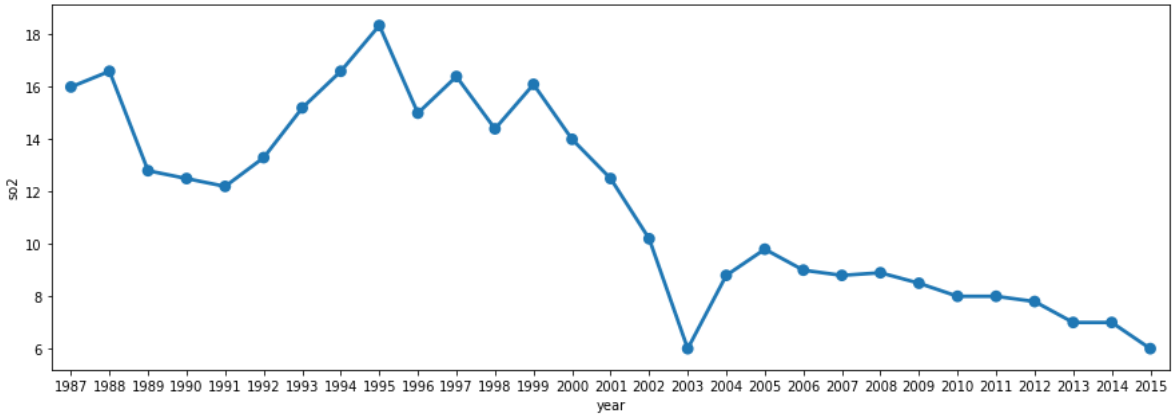
SO2 Analysis

In [15]:

```
df = data[['so2','year','state']].groupby(["year"]).median().reset_index().sort_values(by='
f,ax=plt.subplots(figsize=(15,5))
sns.pointplot(x='year', y='so2', data=df)
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a975db208>
```

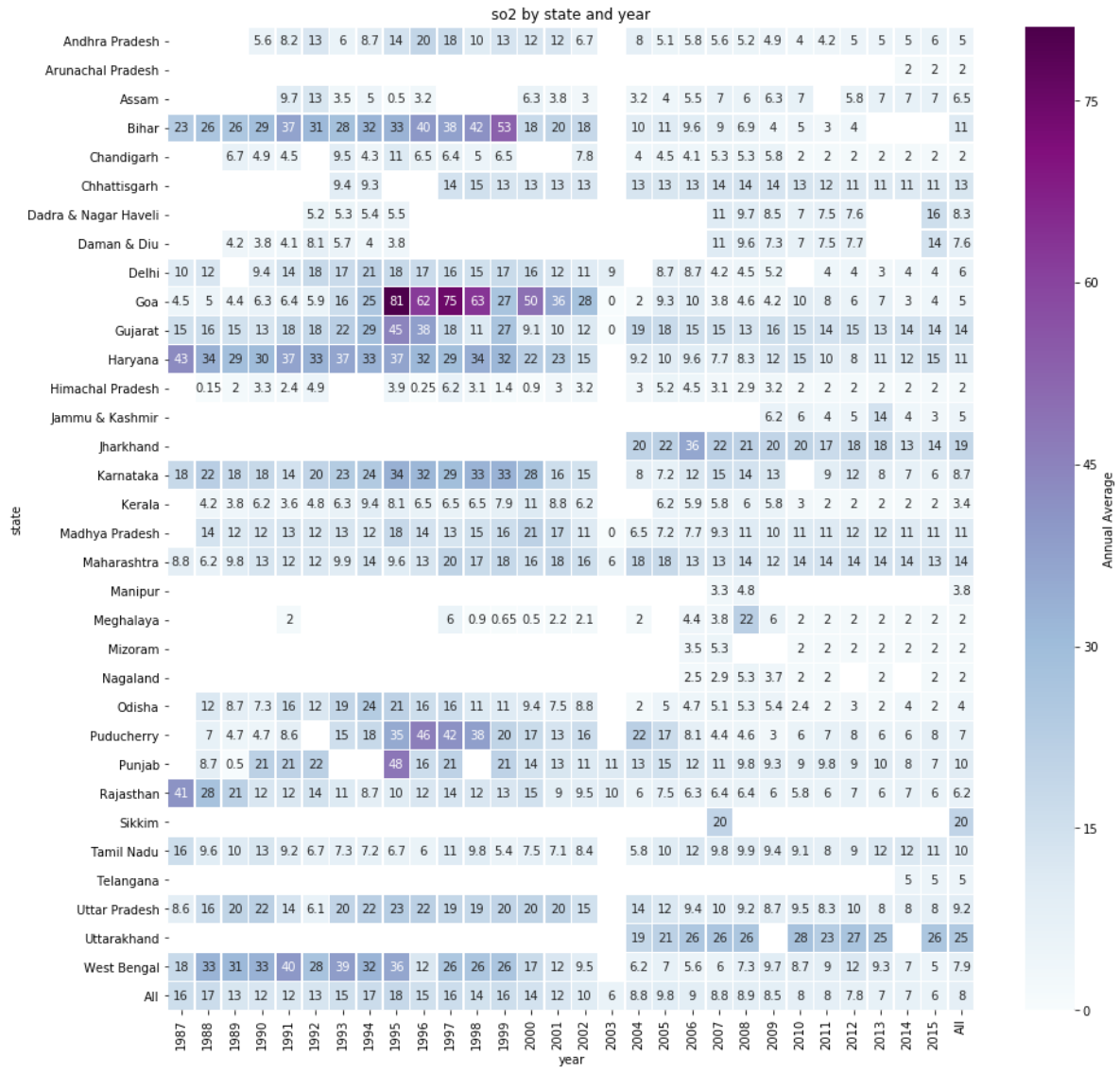


In [16]:

```
#Heatmap Pivot with State as Row, Year as Col, No2 as Value
f, ax = plt.subplots(figsize=(15,15))
ax.set_title('{} by state and year'.format('so2'))
sns.heatmap(data.pivot_table('so2', index='state',
                columns=['year'],aggfunc='median',margins=True),
                annot=True,cmap="BuPu", linewidths=.5, ax=ax,cbar_kws={'label': 'Annual Ave
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a975352b0>
```



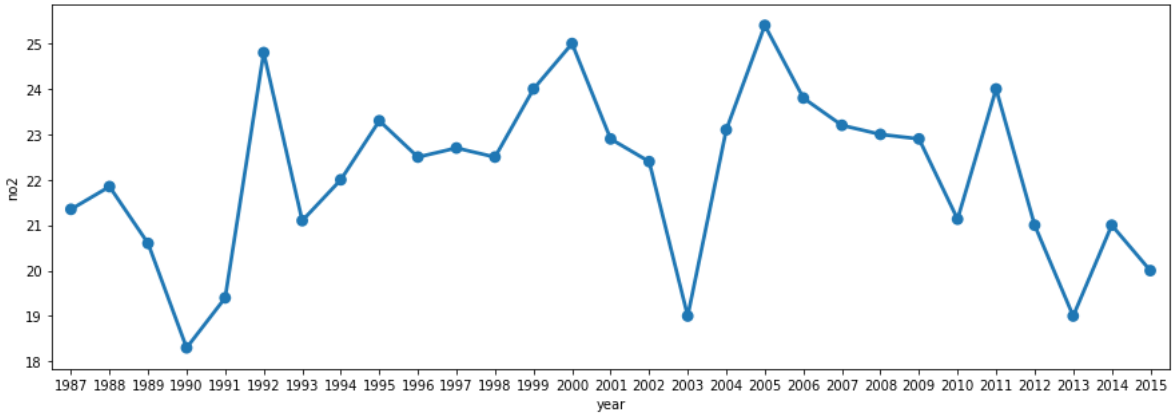
****NO2 Analysis**

In [17]:

```
df = data[['no2','year','state']].groupby(["year"]).median().reset_index().sort_values(by='
f,ax=plt.subplots(figsize=(15,5))
sns.pointplot(x='year', y='no2', data=df)
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a90c69d68>
```

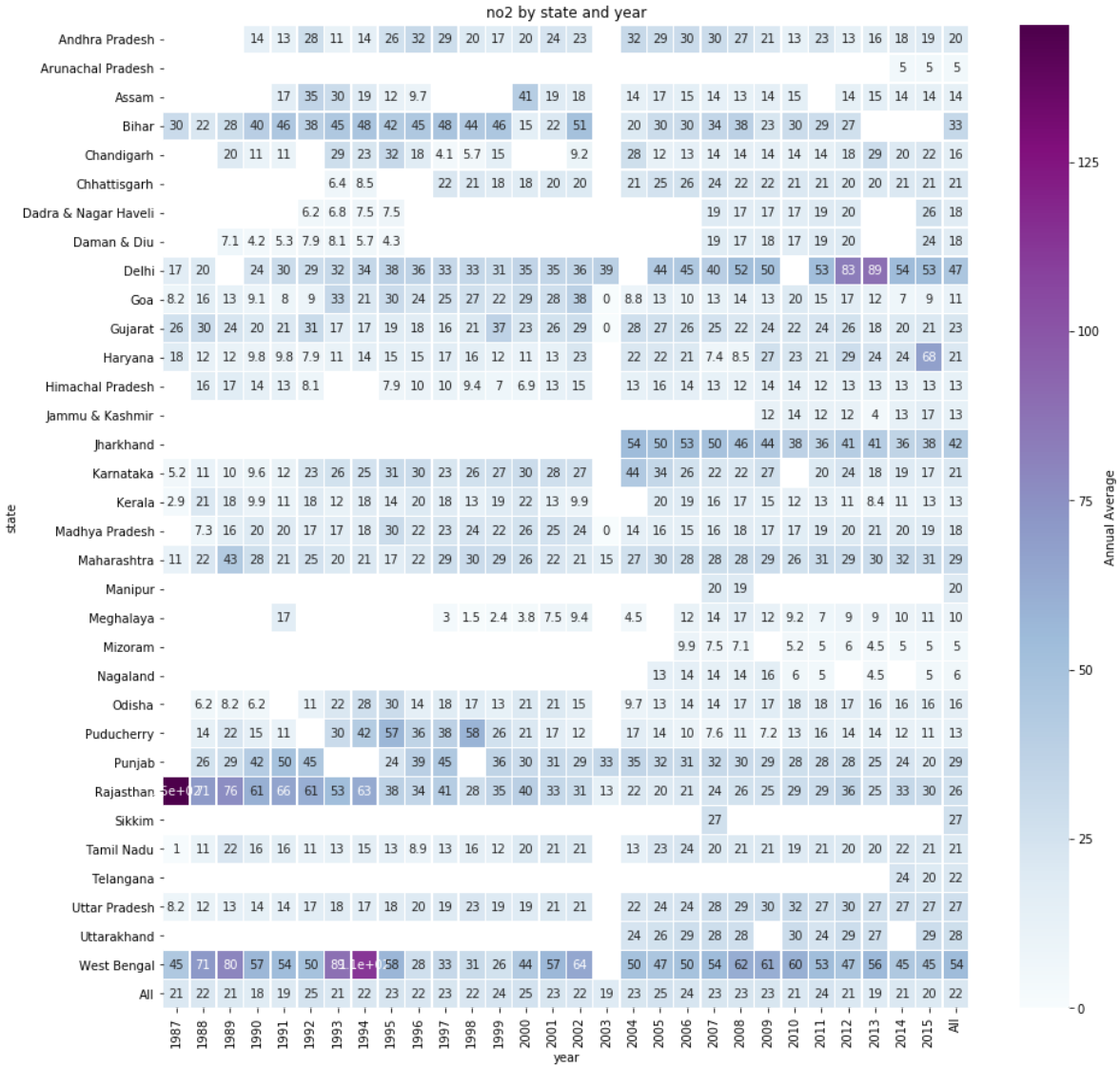


In [18]:

```
#Heatmap Pivot with State as Row, Year as Col, So2 as Value
f, ax = plt.subplots(figsize=(15,15))
ax.set_title('{} by state and year'.format('no2'))
sns.heatmap(data.pivot_table('no2', index='state',
                columns=['year'],aggfunc='median',margins=True),
                annot=True,cmap="BuPu", linewidths=.5, ax=ax,cbar_kws={'label': 'Annual Ave
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7a90bdde48>
```



CONCLUSION

- Mainly Northern states have high air pollution
- South & North East states have less air pollution

**If you have any opinion or any Suggestion Please Comment It **

In [19]: