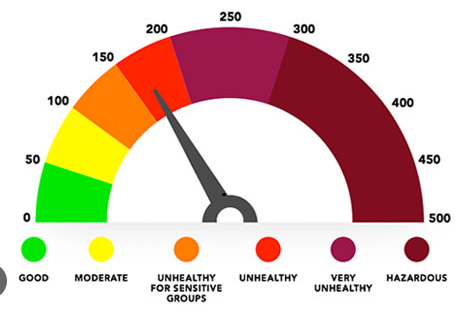
Examining air quality has become one of the most essential activities for the peoples in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by electricity, transportation, fuel uses etc. The deposition of harmful gases is creating a serious problem for the quality of life in smart cities. With this increasing air pollution, we need to implement efficient air quality monitoring and prediction models which collect information about the concentration of air pollutants and provide assessment of air pollution in area.



**Introduction:**

In the developing countries like India, the rapid increase in population and economic upswing in cities have lead to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has direct impact on humans health. There has been increased public awareness about the same in our country. Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution. Précised air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society. The main sources of air pollution are burning of fossil fuels, emission of noxious gases and solid substances from vehicles and industries. Such materials are Sulphur oxides, nitrogen dioxides, particulate matter, carbon monoxide. Currently supervising and scrutinizing air quality is a very crucial issue to have a healthful life, and it also very important. By applying data mining techniques air pollution can be analysed, so that apt actions can be taken for reduction of air pollution.

**Data Accessing:**

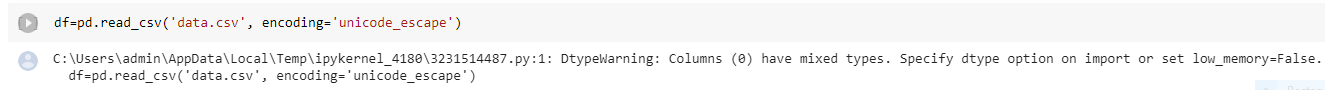
* **Asking Questions:**

The questions that were asked are:

1. **In which state the amount of SO2 is highest?**
2. **In which state the amount of SO2 is lowest?**
3. **In which state the amount of NO2 is highest?**
4. **In which state the amount of NO2 is lowest?**
5. **In which state the amount of SPM is highest?**
6. **In which state the amount of SPM is lowest?**
7. **In which state the amount of RSPM is highest?**
8. **In which state the amount of RSPM is lowest?**
9. **In which year the amount of SO2 is highest?**
10. **In which year the amount of So2 is lowest?**
11. **In which year the amount of NO2 is highest?**
12. **In which year the amount of No2 is lowest?**
13. **In which year the amount of SPM is highest?**
14. **In which year the amount of SPM is lowest?**
15. **In which year the amount of RSPM is highest?**
16. **In which year the amount of RSPM is lowest?**
17. **Which area has highest value in the air pollution?**
18. **What is the range of air quality index (AQI)**

* **Data Wrangling:**
* **Gathering Data:**

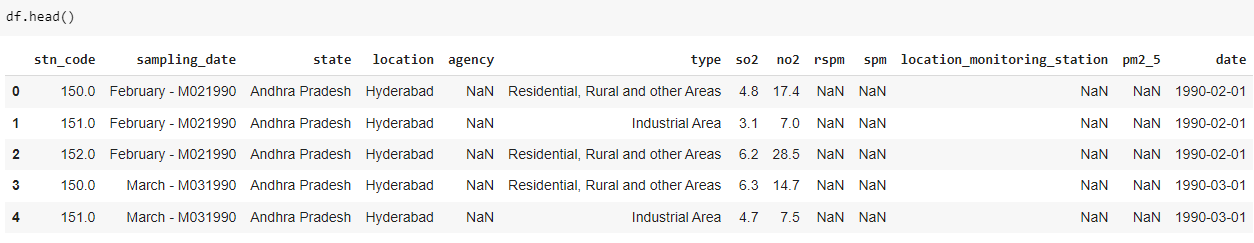
The data is captured using pandas “csv” command.



* **Assessing data:**

Both dirty and messy data is handled in this stage.

The sample of our data is shown below:



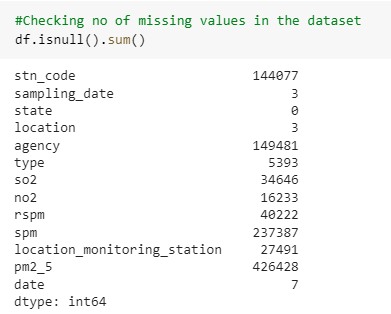
Dirty data:

There are so many missing values in the dataset. The values in the dataset are valid, accurate and also consistent.

Messy data:

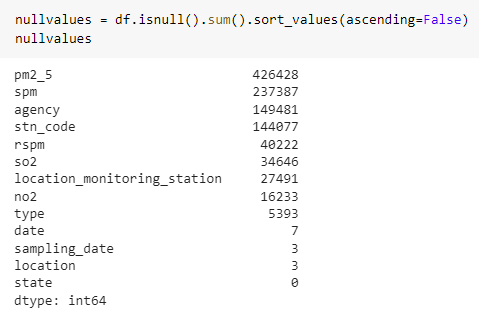
There is no tidiness issues in the data.

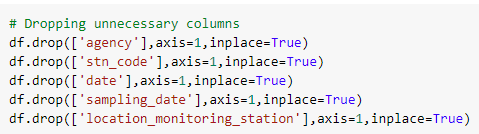
The missing values in our dataset are:



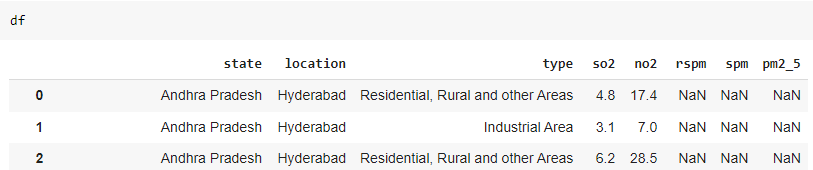
* **Cleaning data:**

Because there were so many missing values in the data. So, we treated it in the following way:

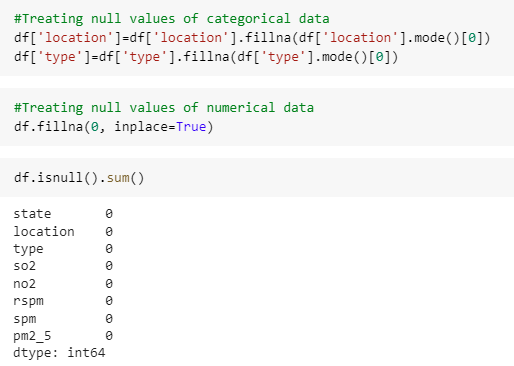


Agency, stn\_code, data, sampling\_data and location\_monitoring\_station has so many missing values. Also, these columns does not helps us in the visualization process. So, we dropped these unnecessary colums. 

After dropping, our data frame looks like this.



The columns location and type has little null values and these two columns were used in the visualization process. So, we replaced the null values with the most occurring ones. Now, our dataset has no missing values.

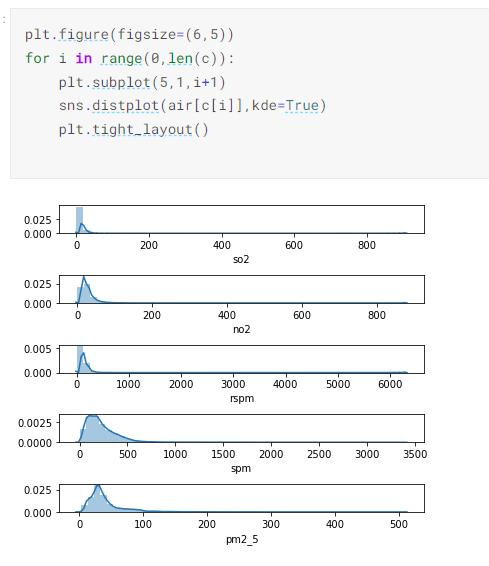


* **Exploratory data analysis (EDA):**
* **Data Analysis:**

In data analysis, we performed univariate and multivariate analysis:

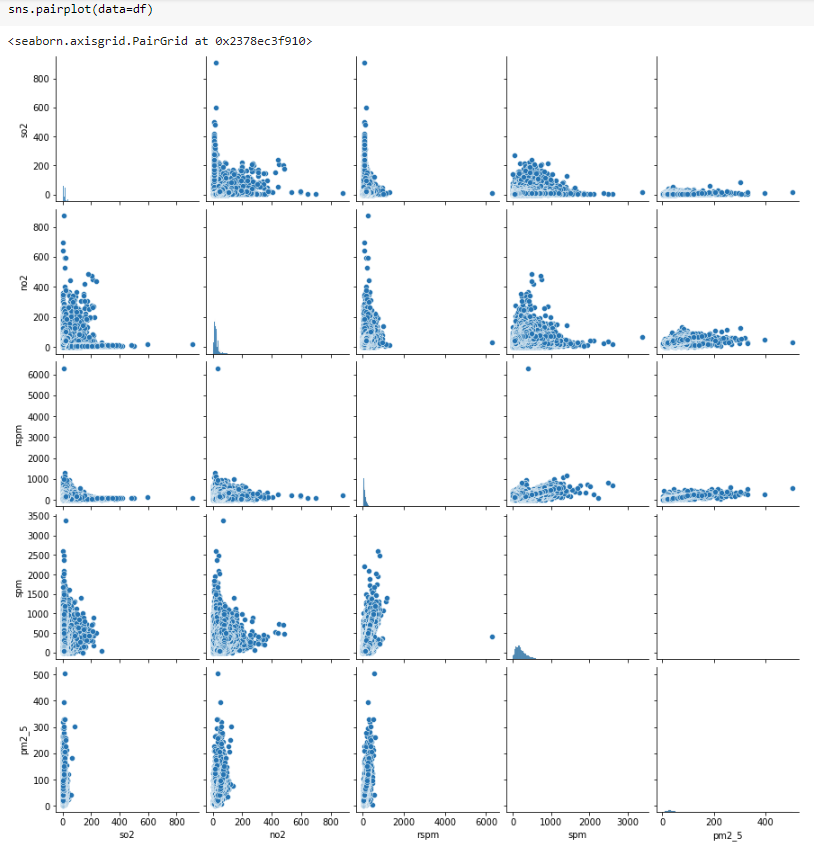
Univariate Analysis:

For, univariate analysis, we created a distplot of some feature columns i-e. so2, no2, rspm, spm, pm2\_5.

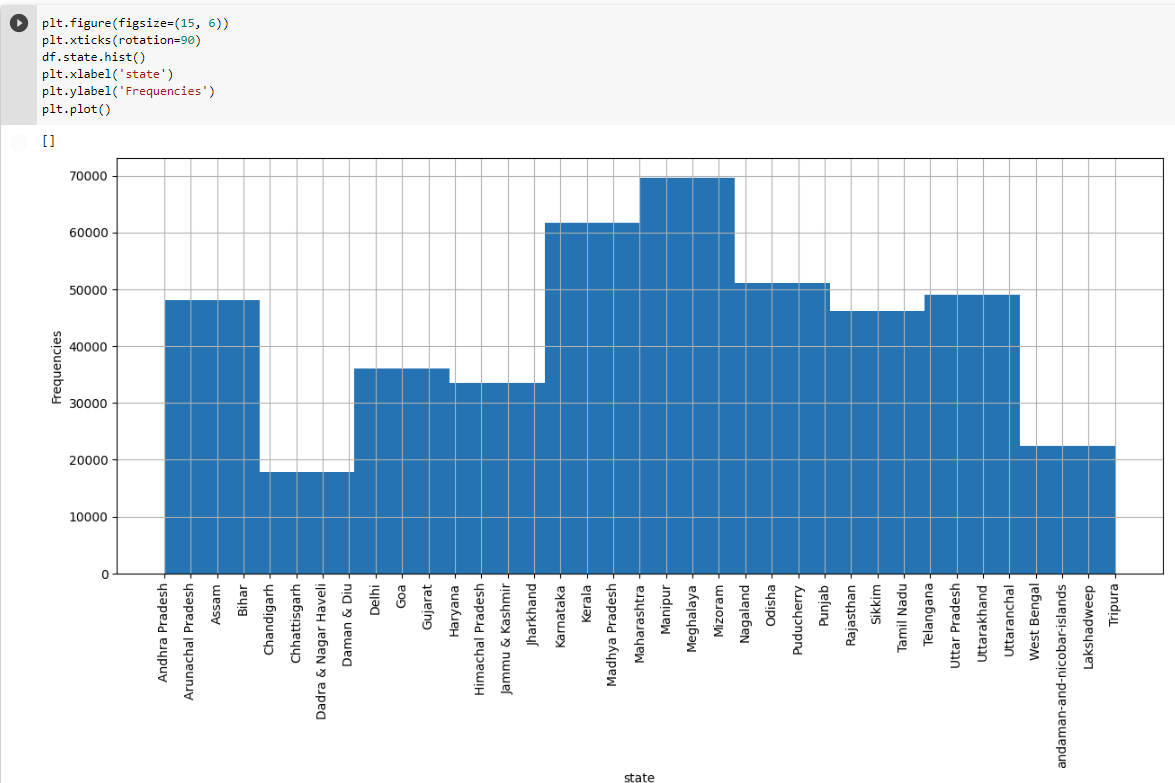


Multivariate Analysis:

Pair plot of whole dataset:

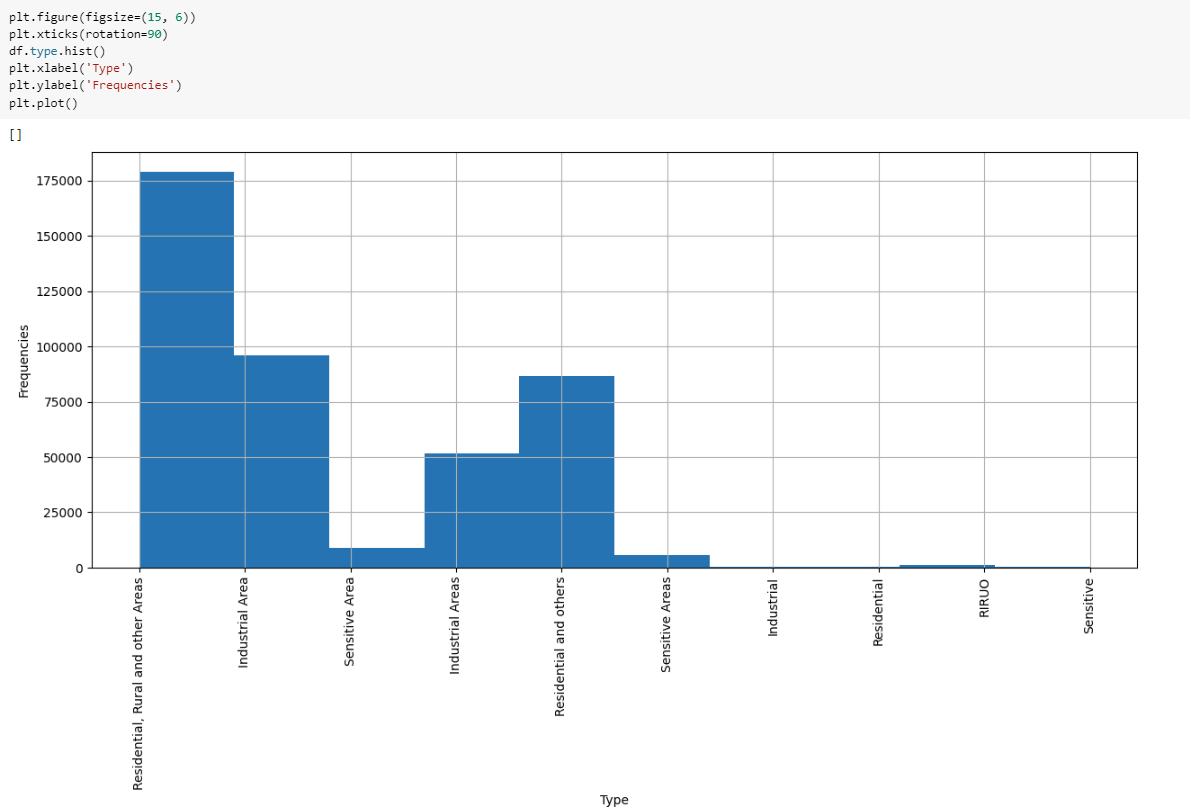


Histogram of frequency and state is shown below:



The highest frequency is of Maharashtra state, Manipur state, Meghalaya and Mizoram state.

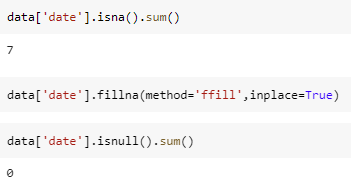
The histogram of area type and frequency is shown below:

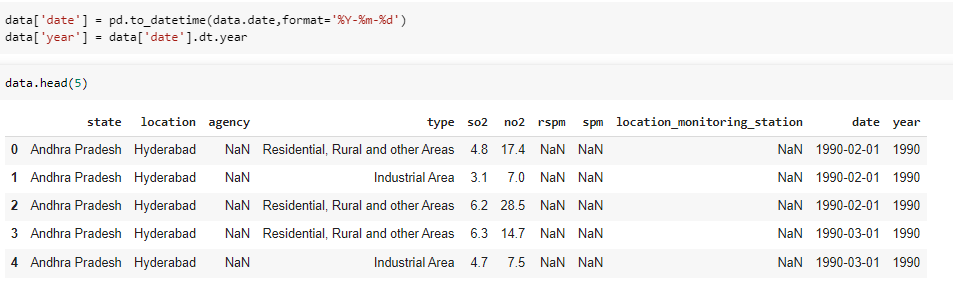


The histogram shows that is of area “Residential, rural and other areas”.

* **Feature Engineering:**

In feature engineering, we created of “year” by modifying an existing date column. Because we wanted to visualize the data according to years.





We don’t need date column. So, we dropped out.

* **Handling outliers:**

Detect outliers:

We detected outliers using Boxplot.



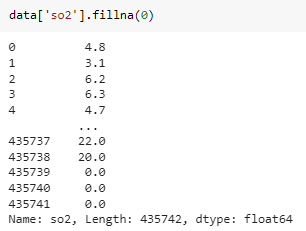
Here, we can see that so2, no2, rspm, spm and pm2\_5 contain some outliers.

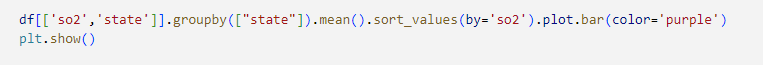
Remove outliers:

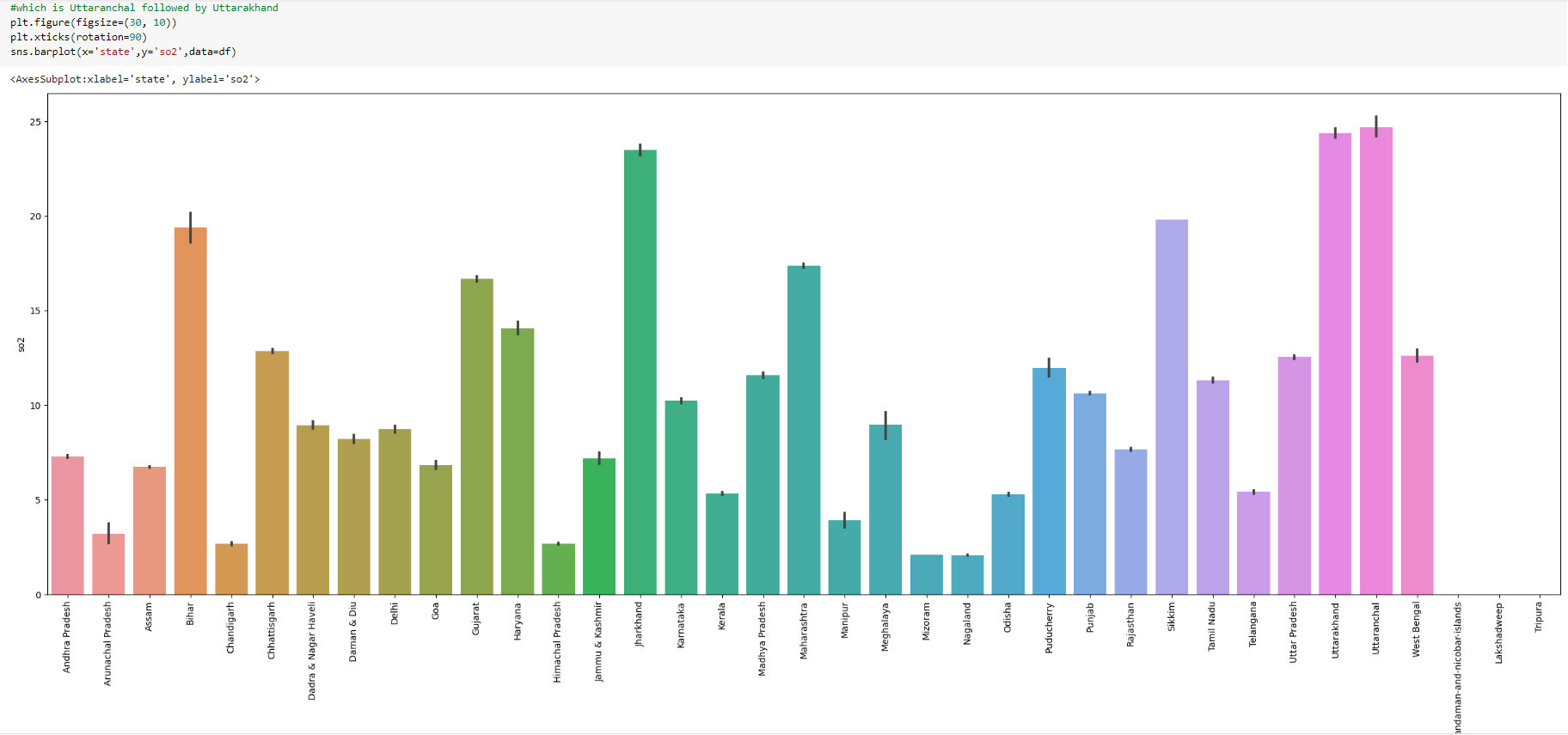
Because the above outliers are not effecting our data a lot. So, removing them is not necessary.

* **Asking Questions:**

1. **In which state the amount of SO2 is highest?**





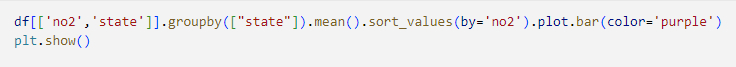


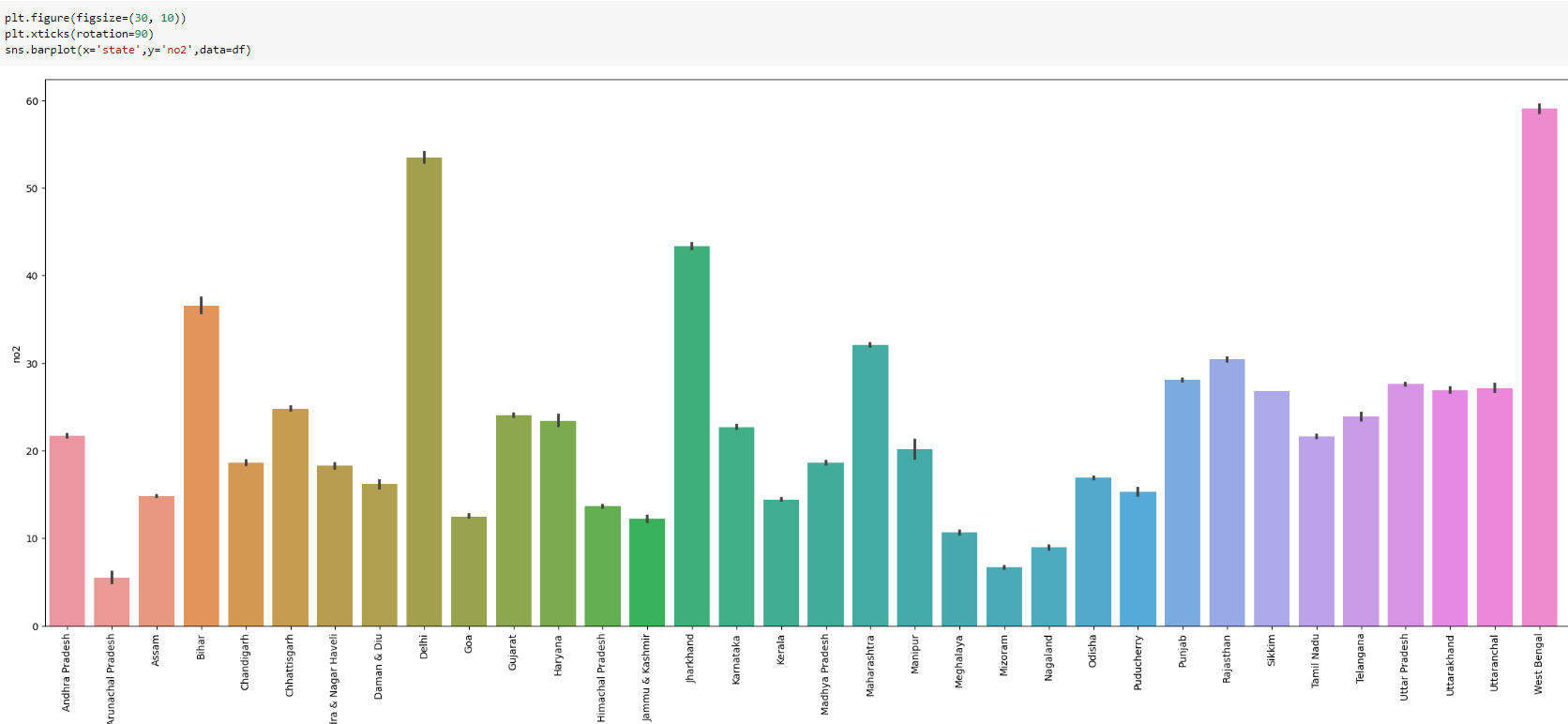
So, Utarkhand state has highest amout of s02.

1. **In which state the amount of SO2 is lowest?**

Mizoram and Nagaland states have very less amount of so2 in the air.

1. **In which state the amount of NO2 is highest?**



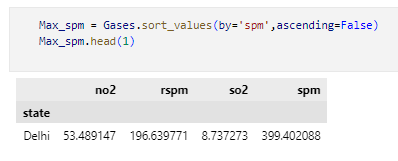


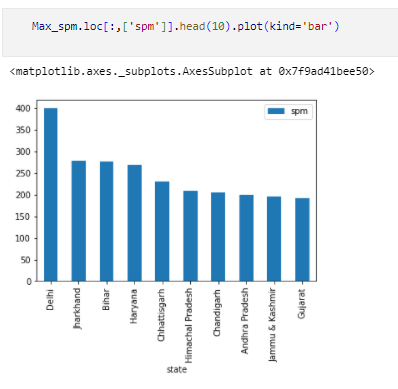
The west Bengal state has highest amount of NO2 gas.

1. **In which state the amount of NO2 is lowest?**

The Arunalchal Pradesh has least amount of NO2 gas.

1. **In which state the amount of SPM is highest?**



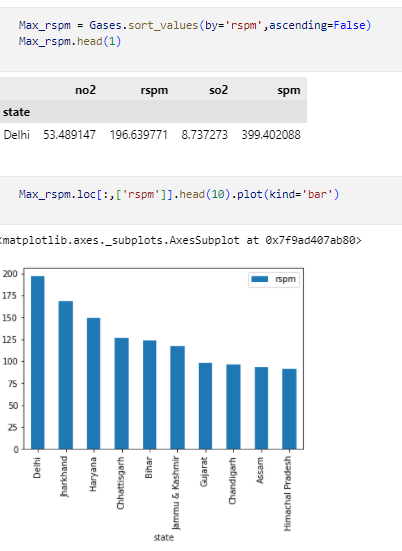


**From above we can see that Delhi has highest spm level.**

1. **In which state the amount of SPM is lowest?**

**Also, Gujrat has lowest SPM.**

1. **In which state the amount of RSPM is highest?**



**From above we concluded, Delhi has highest RSPM.**

1. **In which state the amount of RSPM is lowest?**

**Also, Himachal Pradesh has lowest RSPM.**

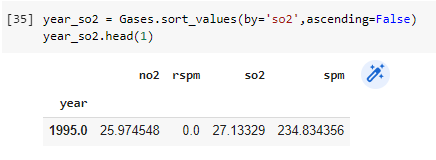
**Year Wise Analysis:**

We created a pivot table of all gases and now checking which year has the highest and lowest amount of these gases

https://lh6.googleusercontent.com/s6DkjON8wXkjJlrUfsWBpYrPEtFwMmaEg0lnB0zazqW90o1dbWX1vh8v6Skf0oGWkO-U8Accd7xtEr6UhK72--ly43vKTG8vITxT41Fh4Xhc2zPc1i1SOqyAFb6xUsPPSA8OE6-z0a1RdvS2mRbHbUG-psM6kCaV8YUmZNnu7pbgPKr87qVpg1RitF944Q

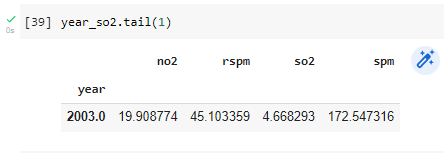
1. **In which year the amount of SO2 is highest?**

In 1995, the amount of SO2 is the highest.



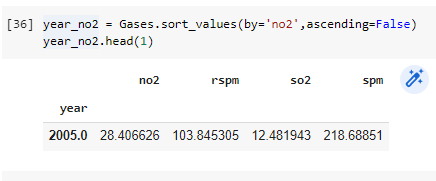
1. **In which year the amount of So2 is lowest?**

In 2003, the amount of SO2 is the lowest.



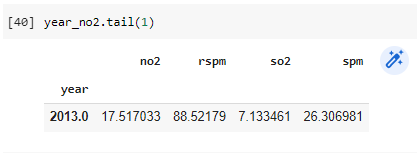
1. **In which year the amount of NO2 is highest?**

In 2005, the amount of NO2 is the highest.



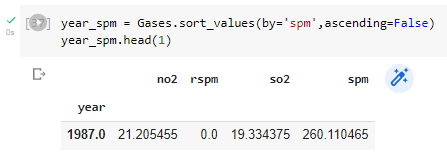
1. **In which year the amount of No2 is lowest?**

In 2013, the amount of NO2 is the lowest.



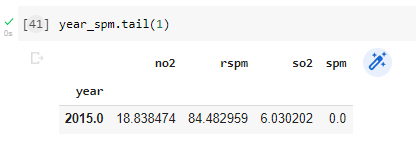
1. **In which year the amount of SPM is highest?**

In 1987, the amount of SPM is the highest.



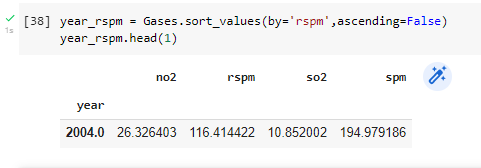
1. **In which year the amount of SPM is lowest?**

In 2015, the amount of SPM is the lowest.



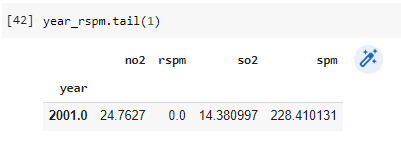
1. **In which year the amount of RSPM is highest?**

In 2004, the amount of RSPM is the highest.

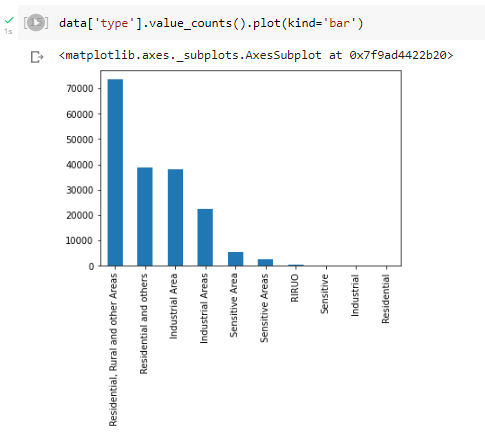


1. **In which year the amount of RSPM is lowest?**

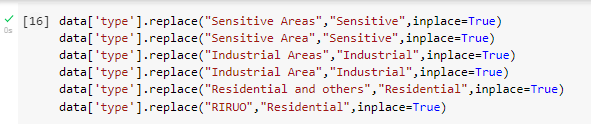
In 2001, the amount of RSPM is the lowest

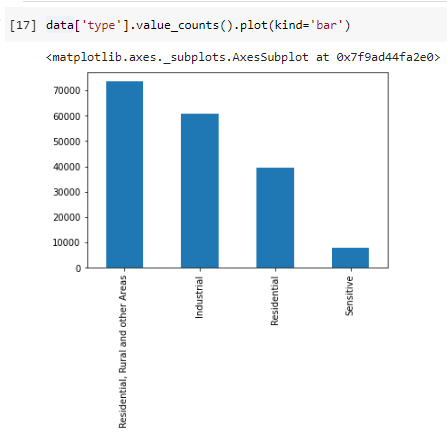


1. **Which area has highest value in the air pollution?**



Sensitive area and sensitive areas are similar area types so, we are merging them. Similarly for industrial and others.

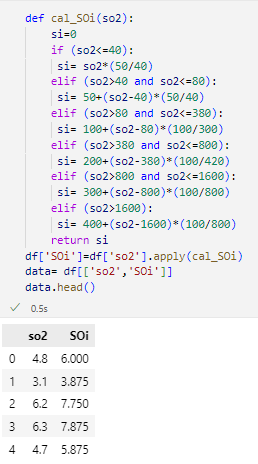




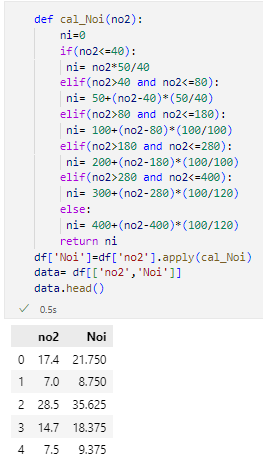
1. **What is the range of air quality index (AQI)?**

**For this we will first calculate the individual pollutant index of SO2, NO2, SPM and RSPM.**

**For SO2:**



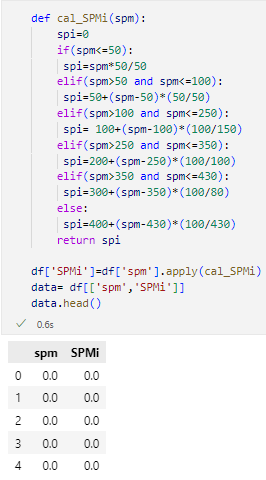
**NO2:**



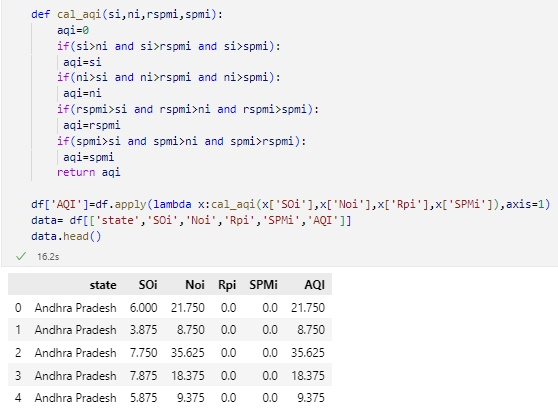
**RSPM:**



**SPM:**

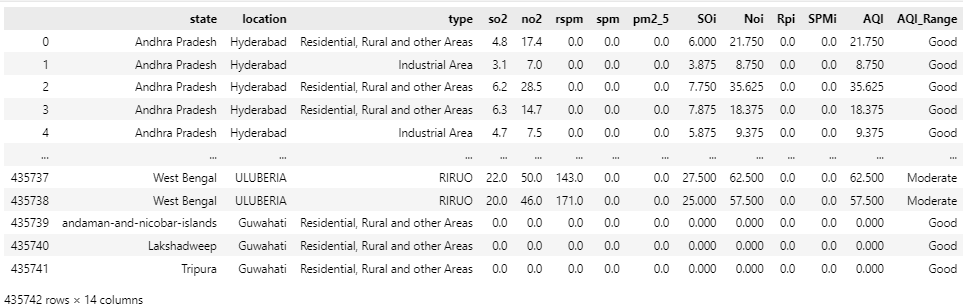


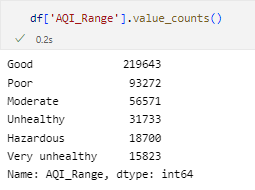
**Now we will calculate the Air Quality Index:**



**After that we will find the range of Air Quality Index:**

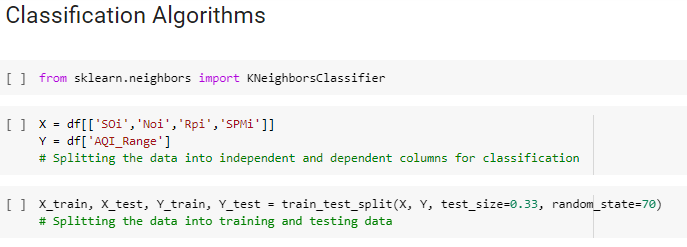


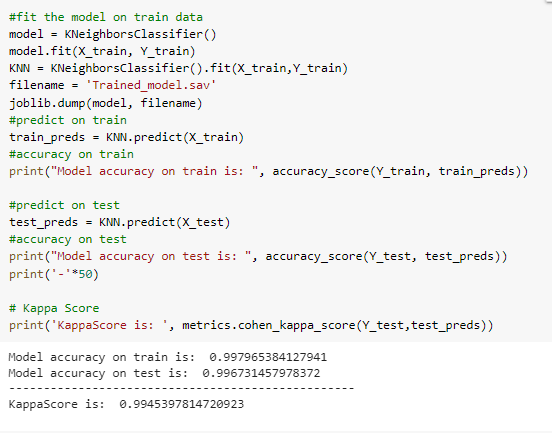


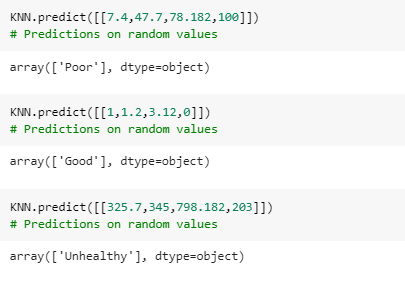


**Machine Learning Model:**

**Dataframe we selected in the ML model:**

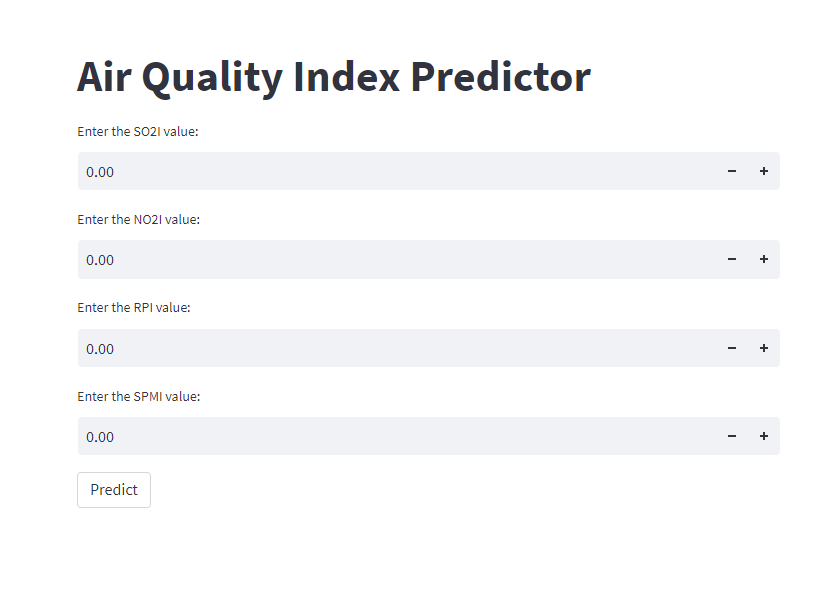




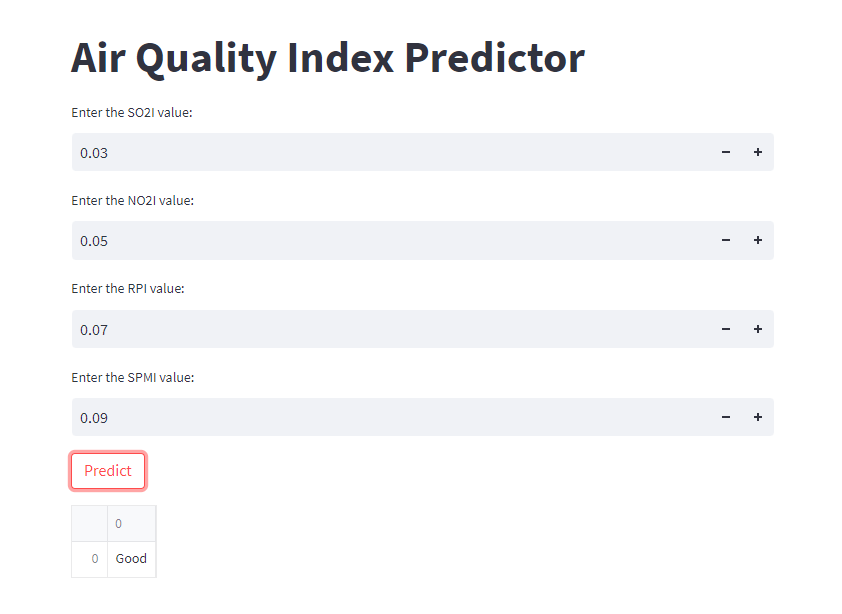


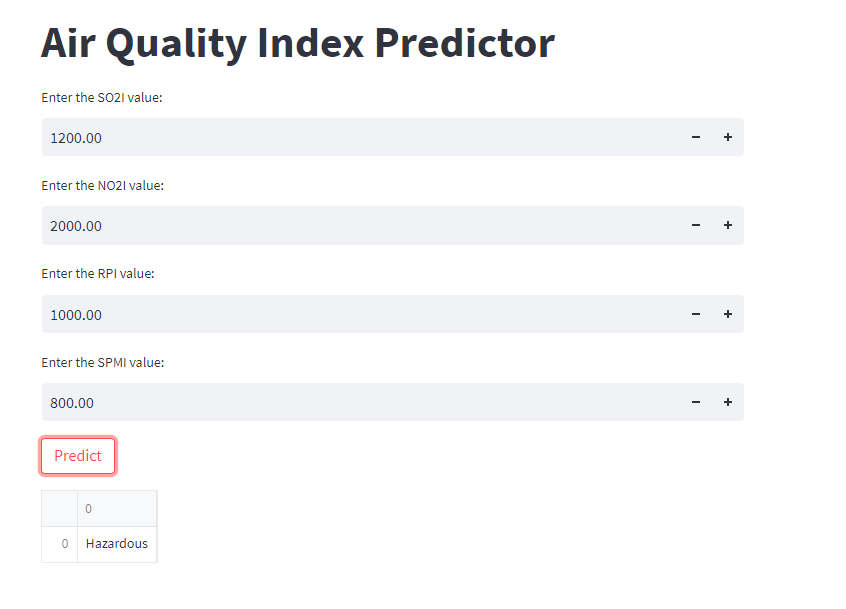
**Front end:**

Our interface looks like this:



**After selecting values:**





**Conclusion**

The above code is a combination of several steps for data preprocessing, data visualization, data modeling and model persistence. The code first loads the dataset from a csv file into a Pandas DataFrame, then it performs some data cleaning and preprocessing operations such as removing the unnecessary columns, replacing missing values, and selecting specific columns to use as input and output variables. Then it uses visualization techniques to explore the data and check for missing values. After that, the code uses the K neighbor classifier the joblib library is used to store and load the model.