SRM INSTITUTE OF SCIENCE AND TECHNOLOGY SCHOOL OF COMPUTING

DEPARTMENT OF COMPUTING TECHNOLOGIES 18CSP107L / 18CSP108L - MINOR PROJECT / INTERNSHIP

Air Quality Analysis Using Machine learning

Guide name: Dr . Padmini S

Designation: Associate professor

Department: CTECH

Batch ID:B593

Student 1 Reg. No: RA2011003010887
Student 1 Name: M Raghava Varma

Student 2 Reg. No: RA2011003010893
Student 2 Name: K Sai Manikanta Pitchaiah

ABSTRACT

- Air pollution alludes to the issue of toxins into the air that are harmful to human well being and the entire planet. It can be described as one of the most dangerous threats that the humanity ever faced.

- It causes damage to animals, crops, forests etc. To prevent this problem in transport sectors have to predict air quality from pollutants using machine learning techniques. Subsequently, air quality assessment and prediction has turned into a significant research zone.

- The aim is to investigate machine learning based techniques for air quality prediction. The air quality dataset is preprocessed with respect to univariate analysis, bi-variate and multi-variate analysis, missing value treatments, data validation, data cleaning/preparing.

- Then, air quality is predicted using Auto Regression Model. This application can help the meteorological Department in predicting air quality. In future, this work can be optimized by applying Artificial Intelligence techniques.

Objective :

1.Forecasting Accuracy: Develop machine learning models that can accurately predict concentrations of key air pollutants (e.g., PM2.5, NO2, CO) over various time intervals, ranging from hours to days.

2.Real-time Monitoring: Create models capable of providing real-time or near-real-time air quality predictions.

3.**Spatial Resolution:** Design models that can predict air quality not only at regional levels but also at finer spatial resolutions, helping to identify localized pollution hotspots and facilitating targeted interventions.

4Policy Support: Provide accurate air quality predictions to policymakers and environmental agencies to assist in formulating effective air quality management policies and regulations.

Objective :

- They provide limited accuracy as they are unable to predict the extreme points i.e. the pollution maximum and minimum
- Cut-offs cannot be determined using such approach
- They use inefficient approach for better output prediction.
- The existence of complex mathematical calculations
- Equal treatment to the old data and new data

INTRODUCTION

- Machine learning is to predict the future from past data. Computer studying (ML) is a style of artificial intelligence (AI) that delivers computers the capability to gain knowledge of without being explicitly programmed. Machine finding out makes a speciality of the progress of pc applications that can alternate when exposed to new information and the basics of laptop studying, implementation of a easy laptop finding out algorithm utilizing python. Process of coaching and prediction involves use of specialised algorithm. It feed the training data to an algorithm, and the algorithm uses this training knowledge to offer predictions on a brand new test information.

EXSITING SYSTEM

- Urban air pollutant attention forecast is coping with a surge of large ecological monitoring data and intricate alterations in air pollution. This necessitates effective estimating methods to strengthen prediction accuracy and avoid grave contamination episodes, thereby improving ecological administration resolution-making capacity.

- A brand new contaminant concentration estimation process is established on sizeable amounts of ecological knowledge and deep learning approaches. This integrates colossal data using two forms of deep networks.

DISADVANTAGES OF EXISTING SYSTEM

- They provide limited accuracy as they are unable to predict the extreme points i.e. the pollution maximum and minimum
- Cut-offs cannot be determined using such approach
- They use inefficient approach for better output prediction

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

PROPOSED SYSTEM

- One versions of auto regression model prediction system will be implemented.
- We will test and evaluate both the systems with same test data to find their prediction accuracy.
- The proposed system is developed to predict air quality using air database and then predict the air quality of the region.

SOFTWARE

• Operating system	: Windows 10
• IDE	: anaconda navigator
• Coding Language	: python

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
1.	Ye Liu, Weipeng Cao, Yiwen Liu, Dachuan Li, Qiang Wang	Online Sequential Extreme Learning Machine (OS-ELM) has been confirmed by numerous studies to be an effective algorithm for online learning scenarios. However, we found that some parameters of OS-ELM are randomly assigned and remain unchanged in the subsequent learning process, which leads to great instability in the model performance in practice.	Air quality prediction problems show that EOS-ELM-R is effective	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
2.	M Sitha Ram, Chintamreddy Reshmasri, Shaik Shahila, Juluru Venkata Pavan Saketh	Identification of fresh air by predicting air quality Index is very important for providing better healthy environment to the society. Air pollution causes a severe health issues for the humans as well as threat to the environment.	XGboost helps to predict the air quality with high accuracy rate.	It shows low accuracy

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
3.	Chenchen Li, Yan Li, Yubin Bao	The prevention and control of environmental pollution attracted much attention, and the haze weather directly affects people's travel health. In order to effectively prevent and control air pollution, optimize the air quality evaluation system. In this paper, PM 2.5 , PM 10 , SO 2 , NO 2 , CO and O 3 _8h are used as characteristic factors, and air quality index is used as a decision factor.	The Gradient Boosting Regression algorithm can effectively predict the Air Quality Index (AQI) and the air quality level.	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
4.	Hairong Qu, Runnan Zhang	After entering the new era, people's living standard has been significantly improved, the concept of environmental protection has been deeply rooted. People pursue a greener and healthier lifestyle, and the concern for air quality has become more and more intense.	All the prediction components are superimposed to obtain the final results.	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
5.	Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth,	The air quality monitoring system measures various air pollutants in various locations to maintain good air quality. It is the burning issue in the present scenario. Air is contaminated by the arrival of dangerous gases into the climate from the industries, vehicular emissions, etc..	Algorithm based on the machine learning to predict the future data of pollutants.	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
6.	Senlin Li, Xiaowu Deng, Bo Tang	Air quality is a closely relation to people' life and agricultural operations, due to the steep mountain and slippery road in a rainy day in the Wuling mountain area. In this study, machine learning methods is adopted to predict air quality of the Wuling mountain area, in order to give a better model for the air quality prediction of Wuling mountain area.	Random forest obtained the better performance than decision tree and deep back propagation neural network.	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
7.	N S Aruna Kumari, K S Ananda Kumar, S Hitesh Vardhan Raju	Air quality monitoring and prediction in many industrial and urban areas, it has become one of the most important activities. Owing to different types of pollution, air quality is heavily affected. With increasing air pollution, efficient air quality monitoring models is to be implemented; these models gather data on the concentration of air pollutants.	To solve three problems- prediction, interpolation and feature analysis, previously these problems were solved using three different models	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
8.	Timothy M. Amado Jennifer C. Dela Cruz	One of the biggest environmental problems right now is air pollution. Air quality is needed to be consistently monitored and assessed to ensure better living conditions. The U.S. Environmental Protection Agency (EPA) uses the air quality index (AQI) to standardize the air quality.	Having the neural network to be the best performing model.	It shows low accuracy

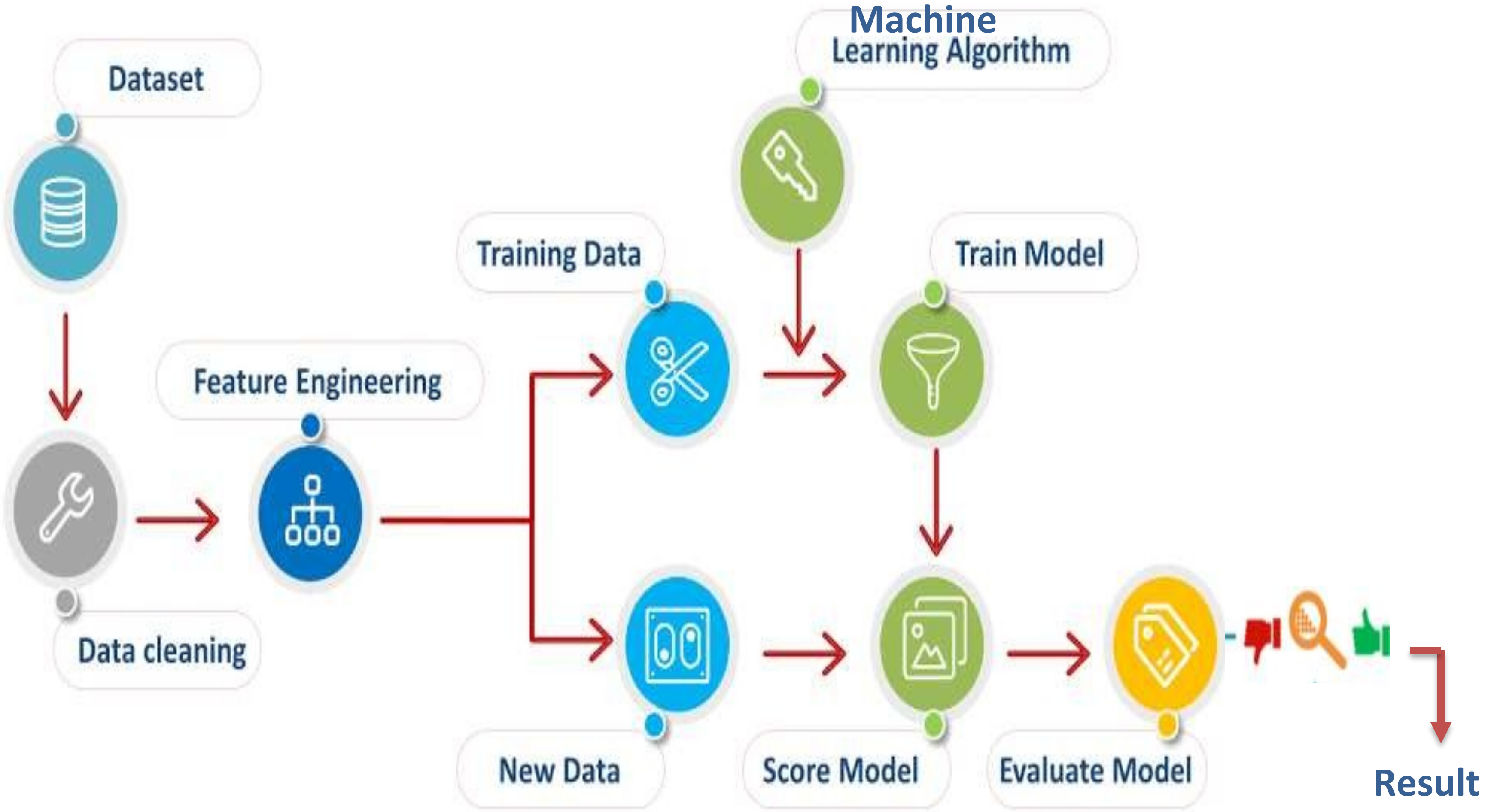
Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
9.	G. Kalaivani, P. Mayilvahanan	Air Pollution (AP) is one of the serious and major environmental problem worldwide. Many researchers have drawn attention and have focused about these problems keeping in mind human health. Air quality prediction information is one of the better ways through which people can be informed to be more vigilant about serious health issues and protect human health caused by air pollution.	The prediction of AQ can be improved by deploying Internet of Things (IoT) based sensor	It shows low accuracy

Literature survey

S.NO	AUTHOR	METHOD	STRENGTH	WEAKNESS
10.	B D Parameshachari , G M Siddesh, V. Sridhar, M Latha	Pollution is the most indispensable and upsetting issues faced in today's day to day life in the world. Over 5000 individuals will lose their life daily due to the various infections of pollution. Air contamination has been perceptible as one of the main issues of metropolitan regions all over the globe, solely in Delhi, Beijing and Tehran and so on.	Algorithms have been proposed for anticipating air contamination.	It shows low accuracy

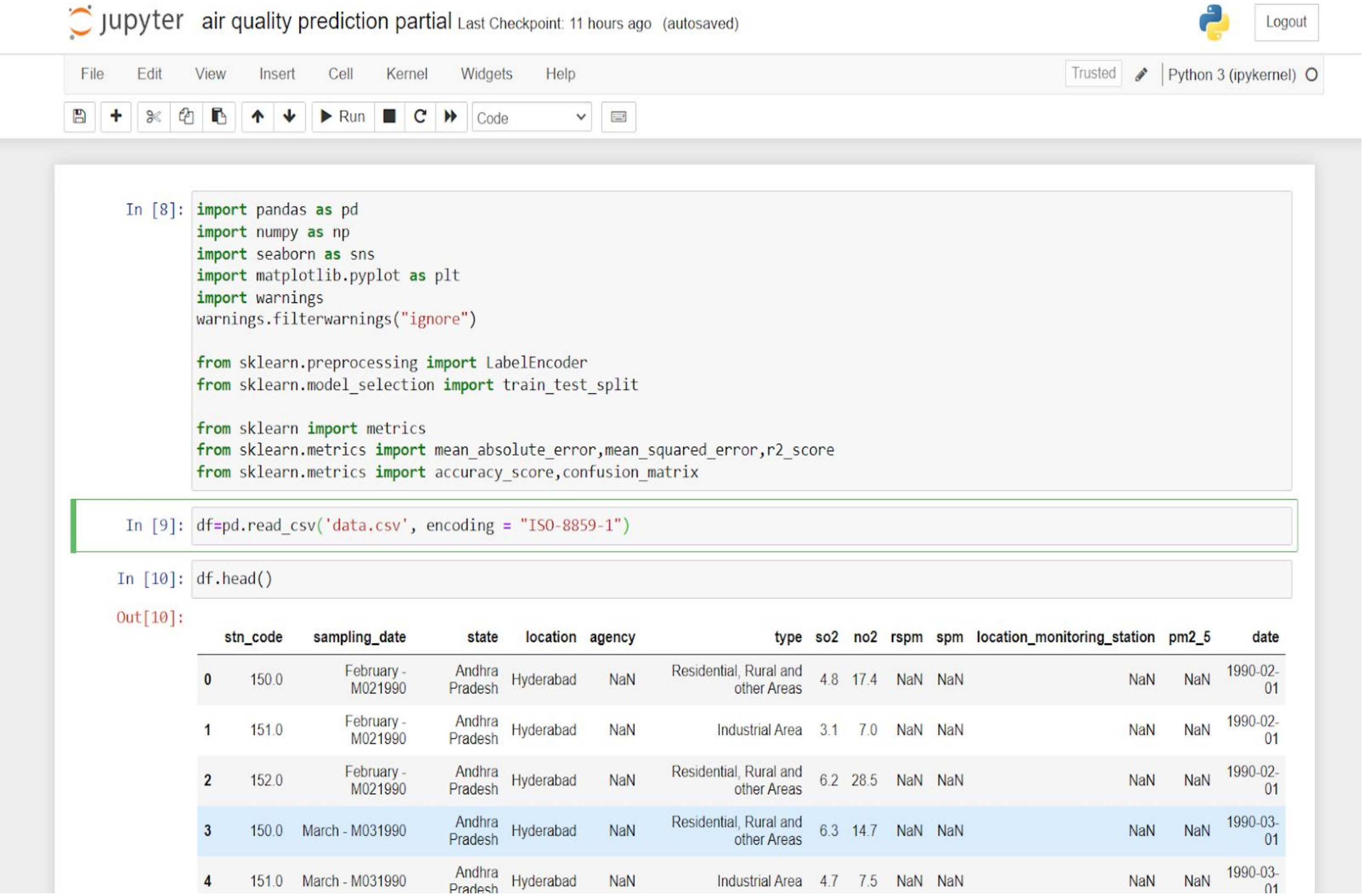
ARCHITECTURE



RESULT

- Based on that dataset we can get the result used our random forest algorithm to predict the result.
- Here we can also find out the accuracy rate of the prediction.
- It will be helpful for finding spam website.

Partial output:



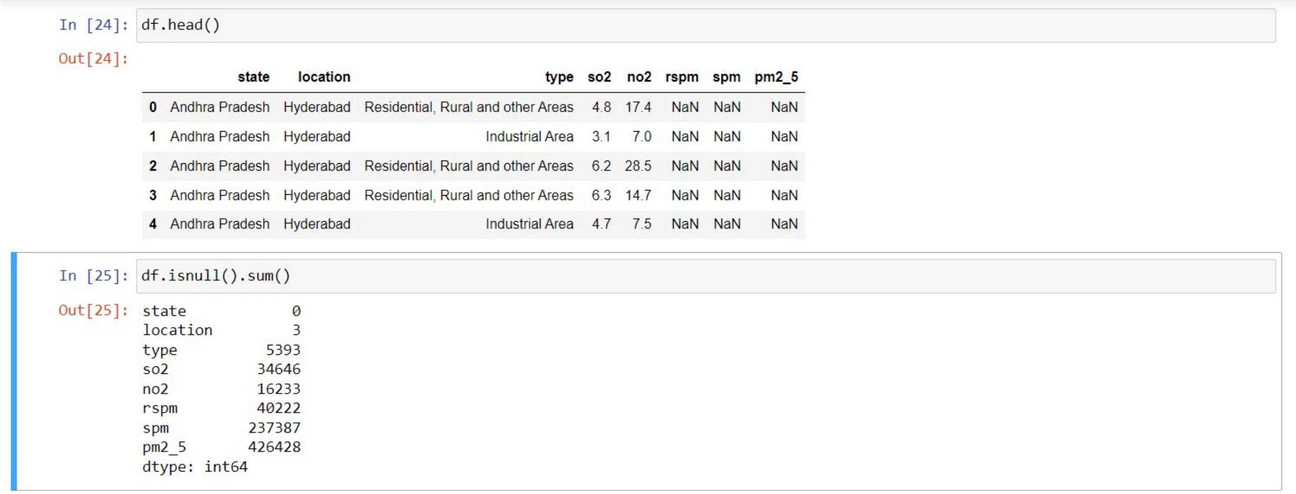
```
In [8]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        import warnings
        warnings.filterwarnings("ignore")

        from sklearn.preprocessing import LabelEncoder
        from sklearn.model_selection import train_test_split

        from sklearn import metrics
        from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
        from sklearn.metrics import accuracy_score,confusion_matrix
```

```
In [9]: df=pd.read_csv('data.csv', encoding = "ISO-8859-1")
```

```
In [10]: df.head()
```

Out[10]:

	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	1990-02-01
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01



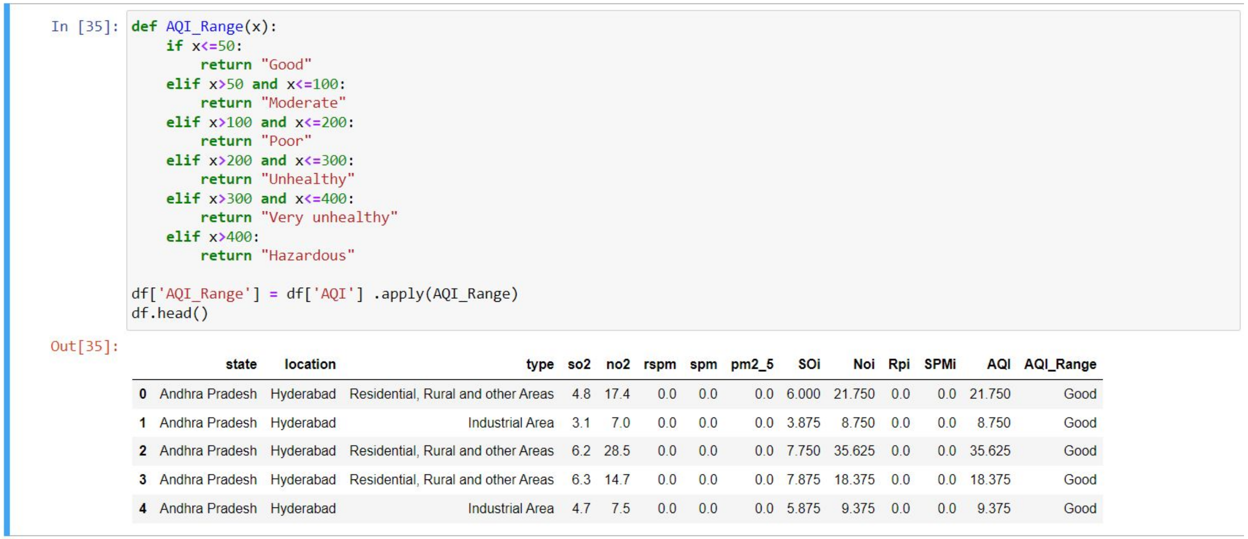
In [24]:
```
df.head()
```

Out[24]:

	state	location	type	so2	no2	rspm	spm	pm2_5
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	NaN
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	NaN

In [25]:
```
df.isnull().sum()
```

Out[25]:
```
state            0
location         3
type          5393
so2          34646
no2          16233
rspm         40222
spm         237387
pm2_5       426428
dtype: int64
```

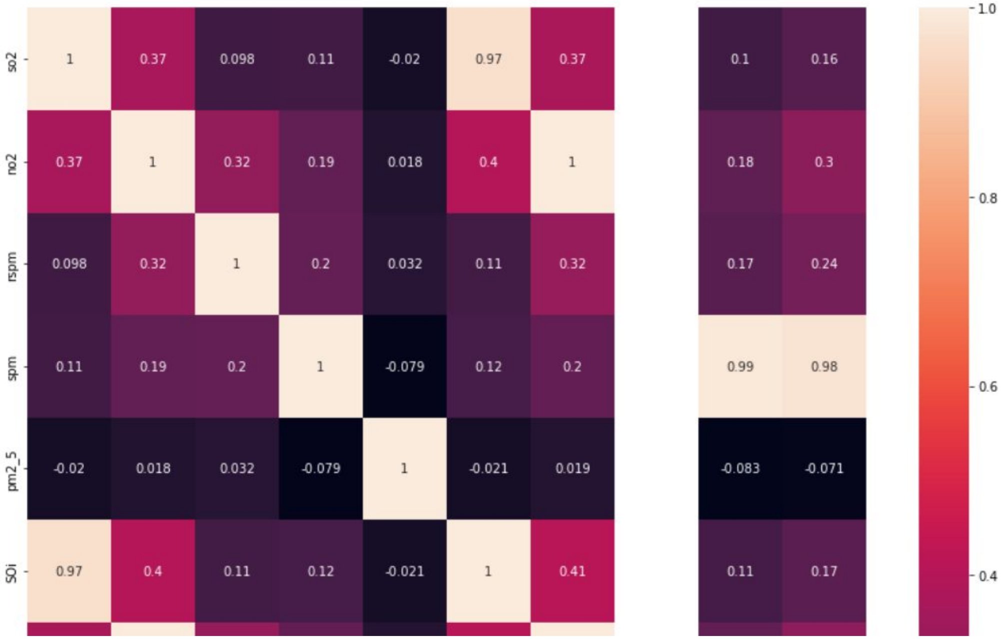


In [35]:
```
def AQI_Range(x):
    if x<=50:
        return "Good"
    elif x>50 and x<=100:
        return "Moderate"
    elif x>100 and x<=200:
        return "Poor"
    elif x>200 and x<=300:
        return "Unhealthy"
    elif x>300 and x<=400:
        return "Very unhealthy"
    elif x>400:
        return "Hazardous"

df['AQI_Range'] = df['AQI'] .apply(AQI_Range)
df.head()
```

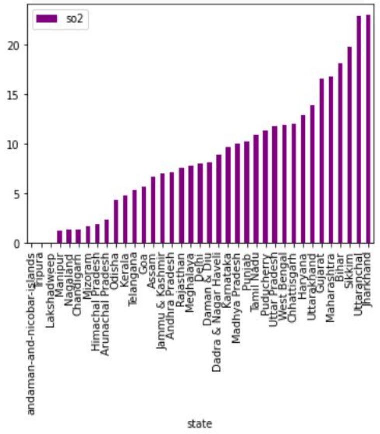
Out[35]:

	state	location	type	so2	no2	rspm	spm	pm2_5	SOi	Noi	Rpi	SPMi	AQI	AQI_Range
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	0.0	0.0	0.0	6.000	21.750	0.0	0.0	21.750	Good
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	0.0	0.0	0.0	3.875	8.750	0.0	0.0	8.750	Good
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	0.0	0.0	0.0	7.750	35.625	0.0	0.0	35.625	Good
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	0.0	0.0	0.0	7.875	18.375	0.0	0.0	18.375	Good
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	0.0	0.0	0.0	5.875	9.375	0.0	0.0	9.375	Good

```
In [36]: plt.figure(figsize=(15,15))
         sns.heatmap(df.corr(), annot=True)
         plt.show()
```



```
In [37]: df[['so2','state']].groupby(["state"]).mean().sort_values(by='so2').plot.bar(color='purple')
         plt.show()
```



Referenc
e

- [1] (2016). PhishMe Q1 2016 Malware Review. [Online]. Available: https://phishme.com/project/phishme-q1-2016-malware-review/
- [2] A. Belabed, E. Aimeur, and A. Chikh, ''A personalized whitelist approach for phishing webpage detection,'' in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.
- [3] Y. Cao, W. Han, and Y. Le, ''Anti-phishing based on automated individual white-list,'' in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.
- [4] T.-C. Chen, S. Dick, and J. Miller, ''Detecting visually similar Web pages: Application to phishing detection,'' ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.
- [5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, ''Clientside defense against Web-based identity theft,'' in Proc. 11th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), 2004, pp. 1–16
- [6] C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar
- [7] J. Corbetta, L. Invernizzi, C. Kruegel, and G. Vigna, ''Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection,'' in Proceedings of Research in Attacks, Intrusions and Defenses (RAID). Gothenburg, Sweden: Springer, 2014.
- [8] X. Deng, G. Huang, and A. Y. Fu, ''An antiphishing strategy based on visual similarity assessment,'' Internet Comput., vol. 10, no. 2, pp. 58–65, 2006.
- [9] Z. Dong, K. Kane, and L. J. Camp, ''Phishing in smooth waters: The state of banking certificates