Water Quality Prediction

CSP 571 Final Report

Sumanth Donthula
A20519856

Ashmita Gupta
A20512498

Abstract

Water quality is a significant concern in India, with numerous states facing challenges in managing and maintaining clean water sources. To tackle this issue, machine learning algorithms are being employed to cluster states based on their water quality parameters. This study aims to cluster Indian states based on water quality data using machine learning algorithms. The dataset includes several water quality parameters such as pH, total dissolved solids, and hardness, among others. The data was preprocessed and scaled, and then k-means clustering was applied to the dataset. We have divided states into six clusters based on their water quality and air quality parameters. This study provides valuable insights into water quality in India and can aid policymakers in developing strategies to improve water quality in states with poor water quality.

1. Overview

Water and air pollution are significant challenges in India, affecting the health and well-being of its citizens. According to the Central Pollution Control Board (CPCB), more than half of India's groundwater is contaminated with pollutants such as arsenic, fluoride, and iron, with states like Bihar, Uttar Pradesh, West Bengal, and Rajasthan having the worst water quality. India is also known for having some of the most polluted air in the world, with high levels of particulate matter, nitrogen dioxide, and sulfur dioxide. States like Delhi, Uttar Pradesh, Bihar, and West Bengal are among the worst affected.

The sources of water and air pollution in India are diverse and include industrial effluents, domestic sewage, agricultural runoff, and solid waste disposal for water pollution and vehicular emissions, industrial emissions, construction activities, and biomass burning for air pollution. Exposure to polluted water and air can cause severe health impacts, including respiratory diseases, cancer, neurological disorders, and developmental problems.

To address these issues, the Indian government has launched various initiatives, including the National Clean Air Programme (NCAP) and the Jal Jeevan Mission (JJM), to improve water and air quality. However, infrastructure inadequacy, inadequate enforcement of regulations, and a growing population continue to pose significant challenges to improving water and air quality in India.

2. Objectives:

- **Early detection of water contamination:** The model can be trained to detect contaminants in water at an early stage, which can help prevent widespread contamination and protect public health
- **Water quality monitoring**: This model can be used to monitor water quality in real-time, allowing water treatment plants to take corrective action quickly if the water quality falls below acceptable levels
- **Air quality monitoring:** This model can be used to monitor the air quality and provides information on the level of air pollution across different states in India

3. Data Collection & Processing

Data Collection: We collected water quality data that includes the values of environmental parameters (temperature, dissolved oxygen, pH, conductivity, BOD, Nitrate N + Nitrite N, Fecal Coliform, Total Coliform) from http://www.cpcbenvis.nic.in/water_quality_data.html

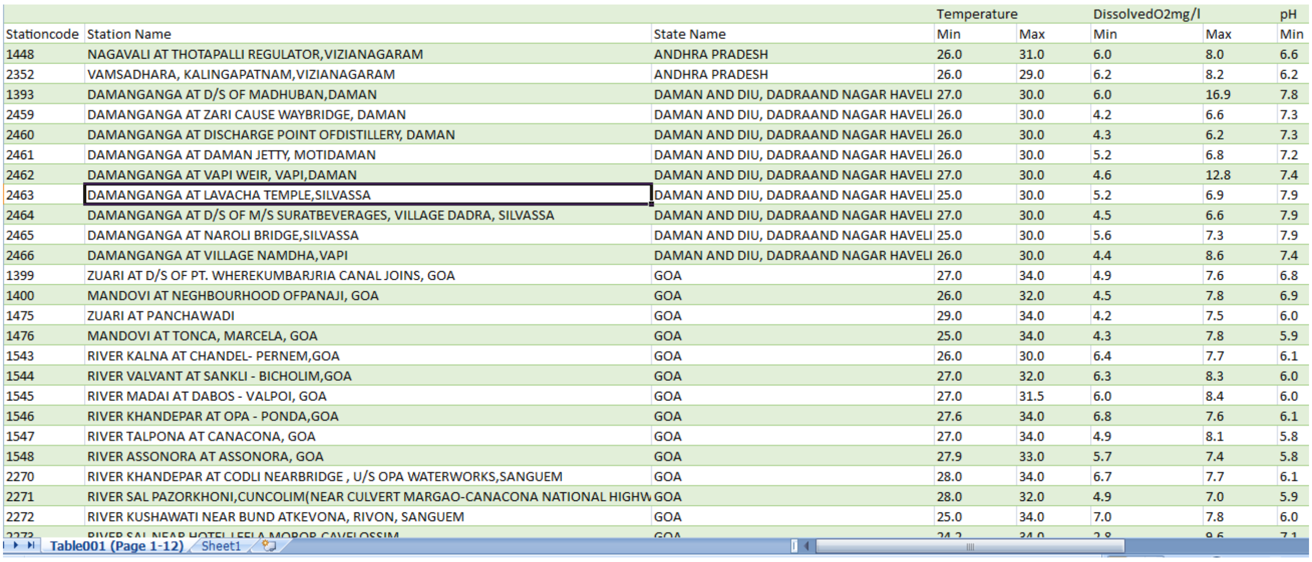
- To Support the features of water quality we have extracted air quality data also to see if it adds some purpose while clustering from http://www.cpcbenvis.nic.in/air_quality_data.html

- In total, we have **Water and Air Quality data** for the years 2017 to 2021 in the form of pdf files for each year

- We collated all the pdfs and combined them into .xlsx files so that they can be processed in R.

Data Processing:

Snapshot of Air Quality Data:

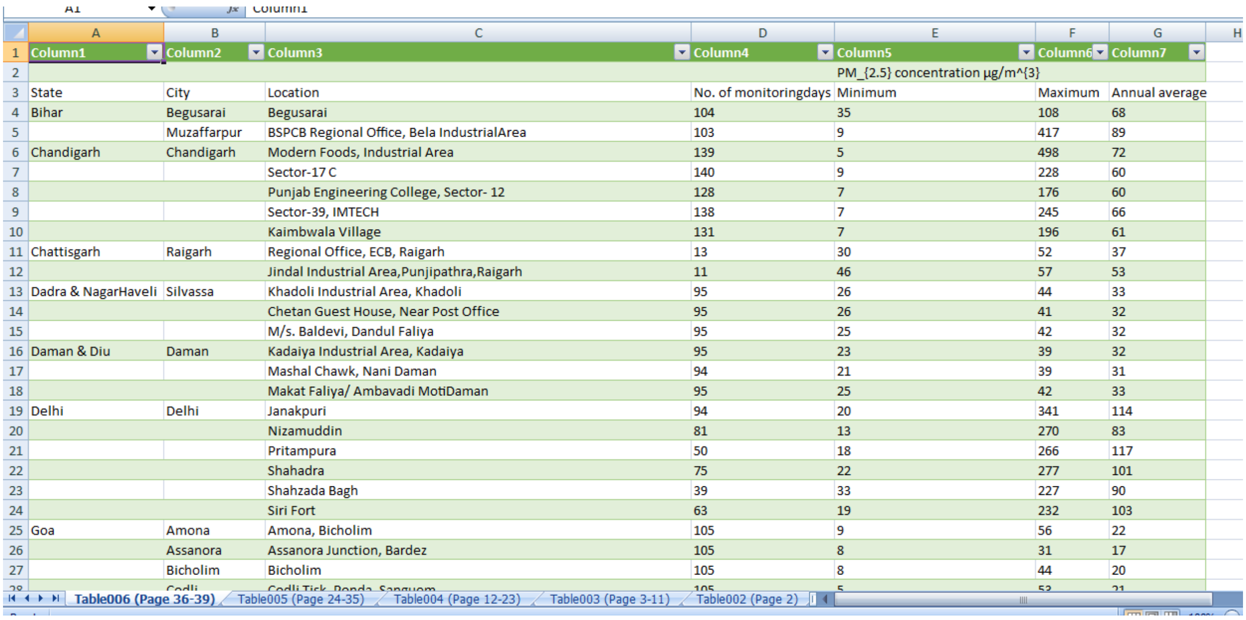


			Temperature		DissolvedO2mg/l		pH
Stationcode	Station Name	State Name	Min	Max	Min	Max	Min
1448	NAGAVALI AT THOTAPALLI REGULATOR,VIZIANAGARAM	ANDHRA PRADESH	26.0	31.0	6.0	8.0	6.6
2352	VAMSADHARA, KALINGAPATNAM,VIZIANAGARAM	ANDHRA PRADESH	26.0	29.0	6.2	8.2	6.2
1393	DAMANGANGA AT D/S OF MADHUBAN,DAMAN	DAMAN AND DIU, DADRAAND NAGAR HAVELI	27.0	30.0	6.0	16.9	7.8
2459	DAMANGANGA AT ZARI CAUSE WAYBRIDGE, DAMAN	DAMAN AND DIU, DADRAAND NAGAR HAVELI	26.0	30.0	4.2	6.6	7.3
2460	DAMANGANGA AT DISCHARGE POINT OFDISTILLERY, DAMAN	DAMAN AND DIU, DADRAAND NAGAR HAVELI	26.0	30.0	4.3	6.2	7.3
2461	DAMANGANGA AT DAMAN JETTY, MOTIDAMAN	DAMAN AND DIU, DADRAAND NAGAR HAVELI	26.0	30.0	5.2	6.8	7.2
2462	DAMANGANGA AT VAPI WEIR, VAPI,DAMAN	DAMAN AND DIU, DADRAAND NAGAR HAVELI	27.0	30.0	4.6	12.8	7.4
2463	DAMANGANGA AT LAVACHA TEMPLE,SILVASSA	DAMAN AND DIU, DADRAAND NAGAR HAVELI	25.0	30.0	5.2	6.9	7.9
2464	DAMANGANGA AT D/S OF M/S SURATBEVERAGES, VILLAGE DADRA, SILVASSA	DAMAN AND DIU, DADRAAND NAGAR HAVELI	27.0	30.0	4.5	6.6	7.9
2465	DAMANGANGA AT NAROLI BRIDGE,SILVASSA	DAMAN AND DIU, DADRAAND NAGAR HAVELI	25.0	30.0	5.6	7.3	7.9
2466	DAMANGANGA AT VILLAGE NAMDHA,VAPI	DAMAN AND DIU, DADRAAND NAGAR HAVELI	26.0	30.0	4.4	8.6	7.4
1399	ZUARI AT D/S OF PT. WHEREKUMBARJRIA CANAL JOINS, GOA	GOA	27.0	34.0	4.9	7.6	6.8
1400	MANDOVI AT NEGHBOURHOOD OFPANAJI, GOA	GOA	26.0	32.0	4.5	7.8	6.9
1475	ZUARI AT PANCHAWADI	GOA	29.0	34.0	4.2	7.5	6.0
1476	MANDOVI AT TONCA, MARCELA, GOA	GOA	25.0	34.0	4.3	7.8	5.9
1543	RIVER KALNA AT CHANDEL- PERNEM,GOA	GOA	26.0	30.0	6.4	7.7	6.1
1544	RIVER VALVANT AT SANKLI - BICHOLIM,GOA	GOA	27.0	32.0	6.3	8.3	6.0
1545	RIVER MADAI AT DABOS - VALPOI, GOA	GOA	27.0	31.5	6.0	8.4	6.0
1546	RIVER KHANDEPAR AT OPA - PONDA,GOA	GOA	27.6	34.0	6.8	7.6	6.1
1547	RIVER TALPONA AT CANACONA, GOA	GOA	27.0	34.0	4.9	8.1	5.8
1548	RIVER ASSONORA AT ASSONORA, GOA	GOA	27.9	33.0	5.7	7.4	5.8
2270	RIVER KHANDEPAR AT CODLI NEARBRIDGE , U/S OPA WATERWORKS,SANGUEM	GOA	28.0	34.0	6.7	7.7	6.1
2271	RIVER SAL PAZORKHONI,CUNCOLIM(NEAR CULVERT MARGAO-CANACONA NATIONAL HIGHW	GOA	28.0	32.0	4.9	7.0	5.9
2272	RIVER KUSHAWATI NEAR BUND ATKEVONA, RIVON, SANGUEM	GOA	25.0	34.0	7.0	7.8	6.0
2273	RIVER SAL NEAR HOTEL LEELA MOBOR CAVELOSSIM	GOA	24.2	34.0	2.8	9.6	7.1

Data cleaning: The data was in different formats for each year. We converted and processed the data in the same format so it’s easy to process this file.

Snapshot of Air Quality Data:

Most of our efforts in the project went into processing this file because its not in a proper format with multiple sheets in excel. More over each sheet have particular feature for example in sheet1 PM2.5 values are there and in sheet2 PM10 values are there.



	Column1	Column2	Column3	Column4	Column5	Column6	Column7
2					PM_{2.5} concentration μg/m^{3}		
3	State	City	Location	No. of monitoringdays	Minimum	Maximum	Annual average
4	Bihar	Begusarai	Begusarai	104	35	108	68
5		Muzaffarpur	BSPCB Regional Office, Bela IndustrialArea	103	9	417	89
6	Chandigarh	Chandigarh	Modern Foods, Industrial Area	139	5	498	72
7			Sector-17 C	140	9	228	60
8			Punjab Engineering College, Sector- 12	128	7	176	60
9			Sector-39, IMTECH	138	7	245	66
10			Kaimbwala Village	131	7	196	61
11	Chattisgarh	Raigarh	Regional Office, ECB, Raigarh	13	30	52	37
12			Jindal Industrial Area,Punjipathra,Raigarh	11	46	57	53
13	Dadra & NagarHaveli	Silvassa	Khadoli Industrial Area, Khadoli	95	26	44	33
14			Chetan Guest House, Near Post Office	95	26	41	32
15			M/s. Baldevi, Dandul Faliya	95	25	42	32
16	Daman & Diu	Daman	Kadaiya Industrial Area, Kadaiya	95	23	39	32
17			Mashal Chawk, Nani Daman	94	21	39	31
18			Makat Faliya/ Ambavadi MotiDaman	95	25	42	33
19	Delhi	Delhi	Janakpuri	94	20	341	114
20			Nizamuddin	81	13	270	83
21			Pritampura	50	18	266	117
22			Shahadra	75	22	277	101
23			Shahzada Bagh	39	33	227	90
24			Siri Fort	63	19	232	103
25	Goa	Amona	Amona, Bicholim	105	9	56	22
26		Assanora	Assanora Junction, Bardez	105	8	31	17
27		Bicholim	Bicholim	105	8	44	20
28		Codli	Codli Tick, Ponda, Sanguem	105	5	53	21

Moreover, if we see there are blank values in state and city because for example Bihar state is repeated twice in 4 and 5 rows so, the 5 row has state value as blank. We need to fill this and for each record we are flagging it's measure which is present in column 5 and pivoting it via mean of measures. Once this part is done we are taking means and filling blank values in our data frame.

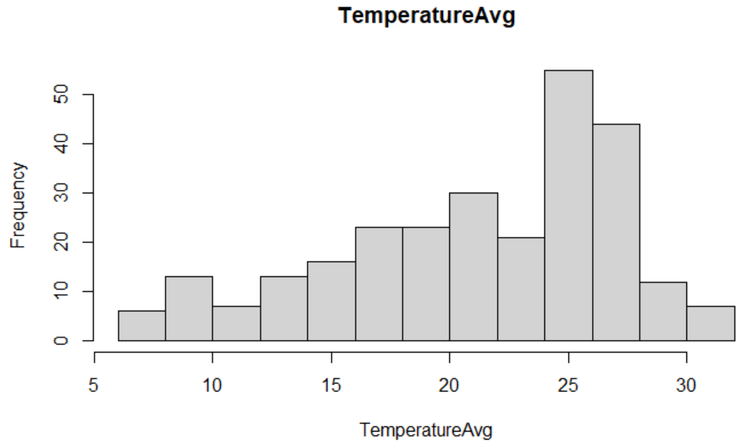
We have merged Water Quality of Data with Air Quality data based on state, city and year.

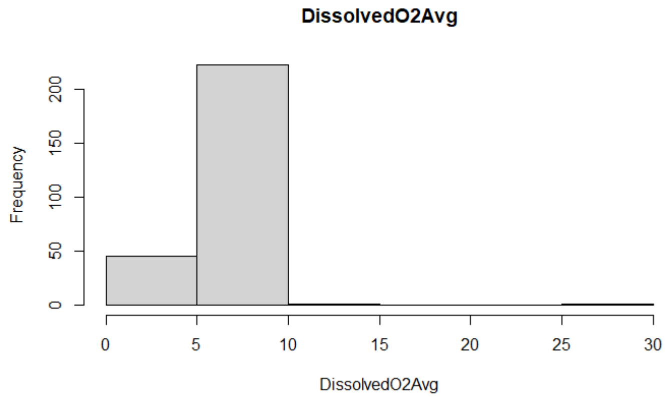
4. Data Analysis

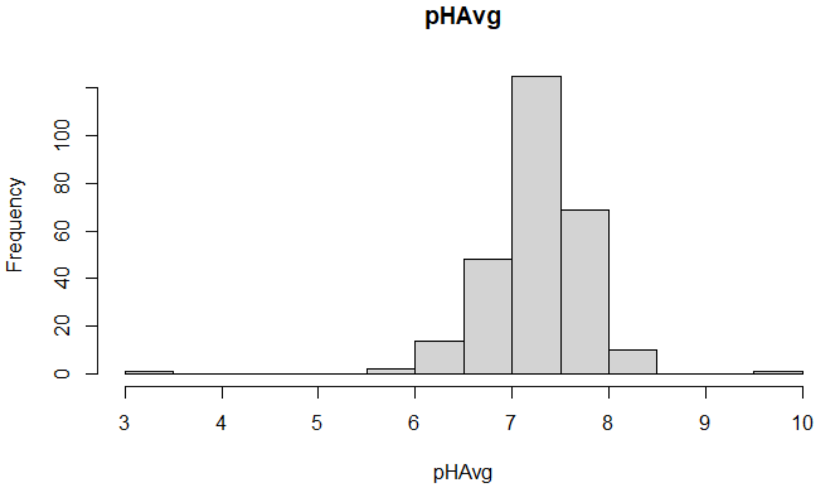
We will present Analysis for 2017 Data and we have done the same analysis for rest of the years. We have got Water quality features like minimum and maximum values, we have averaged them and used while using clustering.

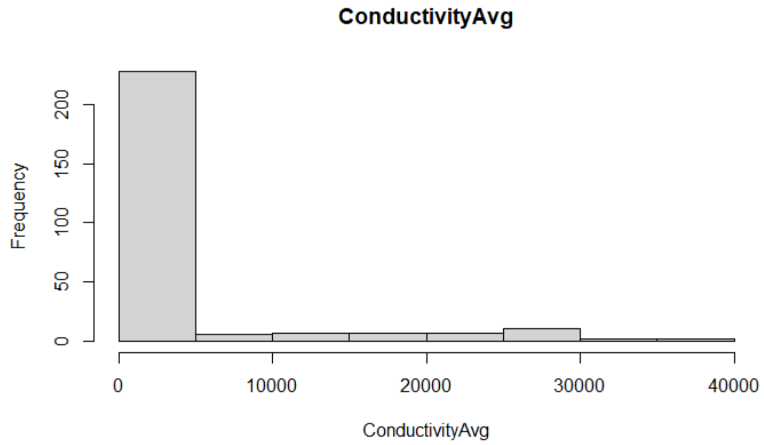
Exploratory Data Analysis (EDA):

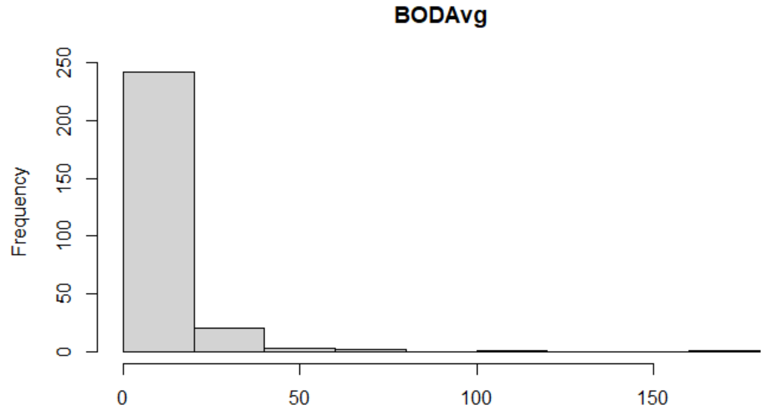
Visualizing the data for all features in the dataset including Water and Air Quality Features.

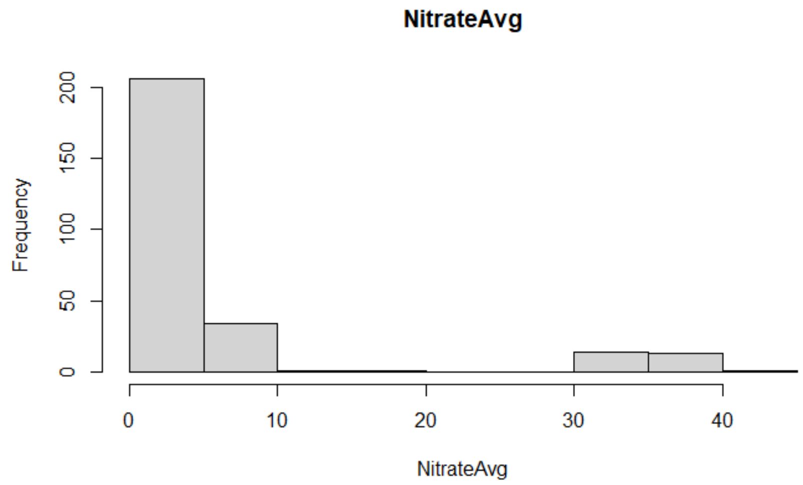


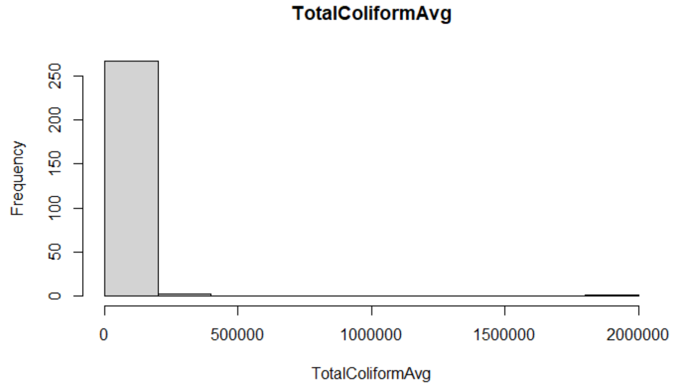

DissolvedO2Avg
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
DissolvedO2Avg

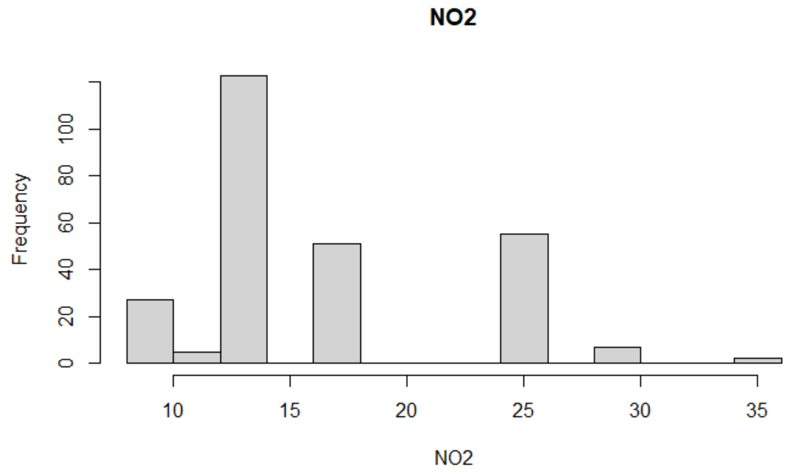

pHAvg
Frequency
0
20
40
60
80
100
3
4
5
6
7
8
9
10
pHAvg

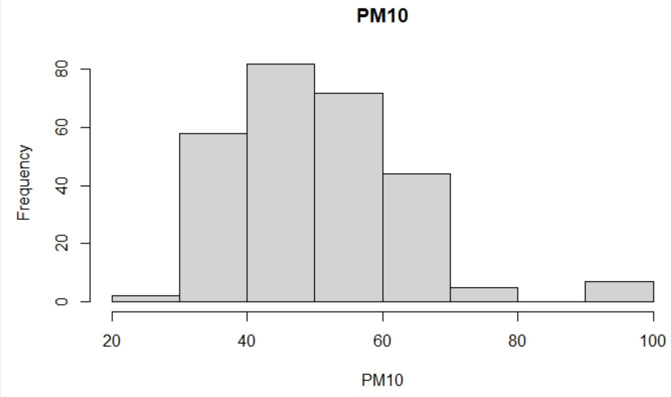

ConductivityAvg
Frequency
0
50
100
150
200
0
10000
20000
30000
40000
ConductivityAvg

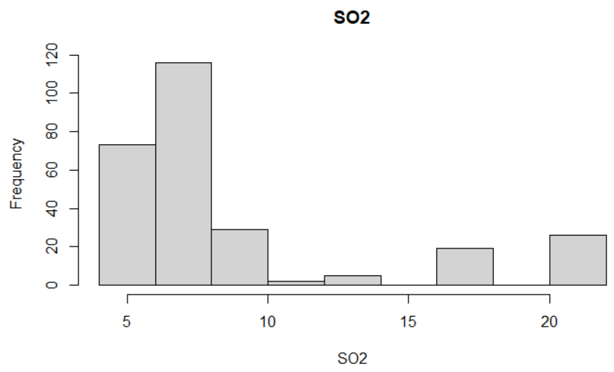

BODAvg
Frequency
0
50
100
150
200
250
0
50
100
150

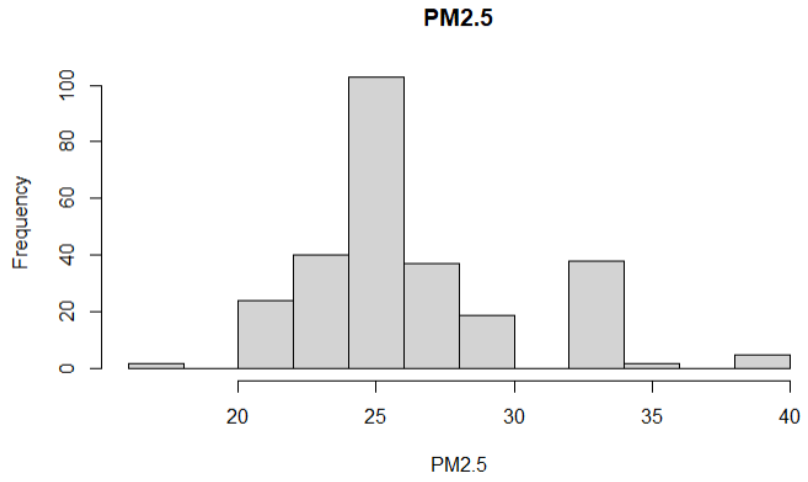

NitrateAvg
Frequency
0
50
100
150
200
0
10
20
30
40
NitrateAvg


TotalColiformAvg
Frequency
0
50
100
150
200
250
0
500000
1000000
1500000
2000000
TotalColiformAvg


NO2
Frequency
0
20
40
60
80
100
10
15
20
25
30
35
NO2

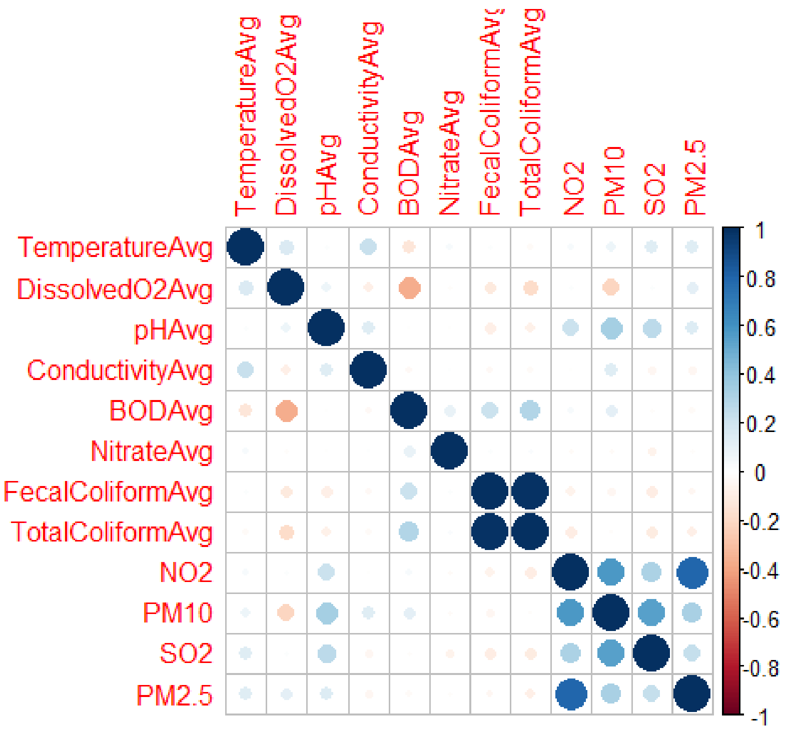

PM10
Frequency
0
20
40
60
80
20
40
60
80
100
PM10





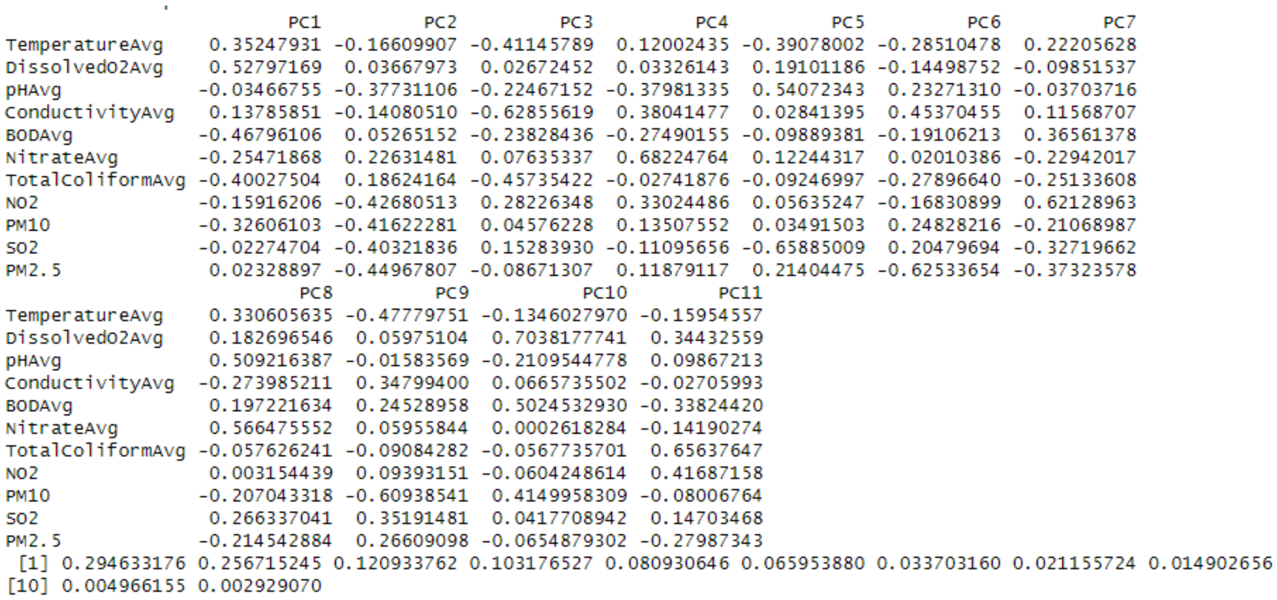
We can observe that most of the attributes are normally distributed, and some attributes are skewed. Anyways, we will perform scaling before using clustering models.

Obtaining the covariance matrix:



We can see that Total Coliform Avg and Fecal Coliform Avg are same and PM2.5, NO2 are highly correlated. So, we will remove those columns for further processing.

Performing PCA for 2017 Data:



```
                        PC1         PC2         PC3         PC4         PC5         PC6         PC7
TemperatureAvg    0.35247931 -0.16609907 -0.41145789  0.12002435 -0.39078002 -0.28510478  0.22205628
DissolvedO2Avg    0.52797169  0.03667973  0.02672452  0.03326143  0.19101186 -0.14498752 -0.09851537
pHAvg            -0.03466755 -0.37731106 -0.22467152 -0.37981335  0.54072343  0.23271310 -0.03703716
ConductivityAvg   0.13785851 -0.14080510 -0.62855619  0.38041477  0.02841395  0.45370455  0.11568707
BODAvg           -0.46796106  0.05265152 -0.23828436 -0.27490155 -0.09889381 -0.19106213  0.36561378
NitrateAvg       -0.25471868  0.22631481  0.07635337  0.68224764  0.12244317  0.02010386 -0.22942017
TotalColiformAvg -0.40027504  0.18624164 -0.45735422 -0.02741876 -0.09246997 -0.27896640 -0.25133608
NO2              -0.15916206 -0.42680513  0.28226348  0.33024486  0.05635247 -0.16830899  0.62128963
PM10             -0.32606103 -0.41622281  0.04576228  0.13507552  0.03491503  0.24828216 -0.21068987
SO2              -0.02274704 -0.40321836  0.15283930 -0.11095656 -0.65885009  0.20479694 -0.32719662
PM2.5             0.02328897 -0.44967807 -0.08671307  0.11879117  0.21404475 -0.62533654 -0.37323578
                         PC8         PC9          PC10        PC11
TemperatureAvg    0.330605635 -0.47779751 -0.1346027970 -0.15954557
DissolvedO2Avg    0.182696546  0.05975104  0.7038177741  0.34432559
pHAvg             0.509216387 -0.01583569 -0.2109544778  0.09867213
ConductivityAvg  -0.273985211  0.34799400  0.0665735502 -0.02705993
BODAvg            0.197221634  0.24528958  0.5024532930 -0.33824420
NitrateAvg        0.566475552  0.05955844  0.0002618284 -0.14190274
TotalColiformAvg -0.057626241 -0.09084282 -0.0567735701  0.65637647
NO2               0.003154439  0.09393151 -0.0604248614  0.41687158
PM10             -0.207043318 -0.60938541  0.4149958309 -0.08006764
SO2               0.266337041  0.35191481  0.0417708942  0.14703468
PM2.5            -0.214542884  0.26609098 -0.0654879302 -0.27987343
 [1] 0.294633176 0.256715245 0.120933762 0.103176527 0.080930646 0.065953880 0.033703160 0.021155724 0.014902656
[10] 0.004966155 0.002929070
```

The first 4 principal components explain 77% of the Variance in Data.

5. Clustering

For water and air quality analysis in Indian states, we used use **k-means clustering** to group states based on their water quality parameters.

Step 1: We started by selecting the relevant water and air quality parameters, such as pH, total dissolved solids, and concentration of various contaminants.

Step 2: We then used k-means clustering for grouping states based on air & water parameters in datsets

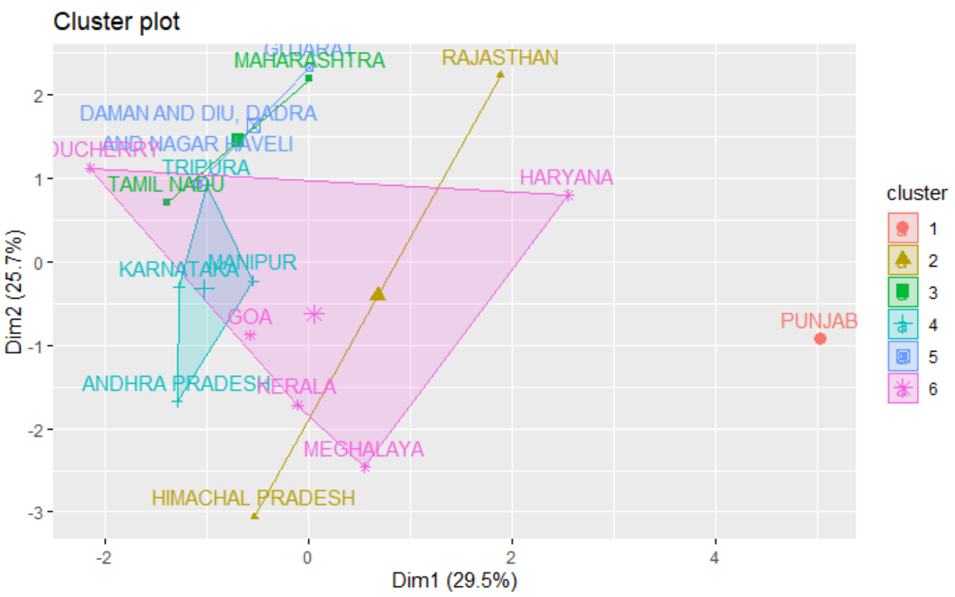
Step 3: We got with 6 clusters on our filtered dataset using K means clustering

Step 4: Then once the clusters were formed, we assigned ranks to the clusters based on their features like Fecal Coliform, BOD, Dissolved O2, PM10, PM2.5 and NO2 to give a meaning to these clusters. We took a reference of below criteria that is ideal for checking water quality:

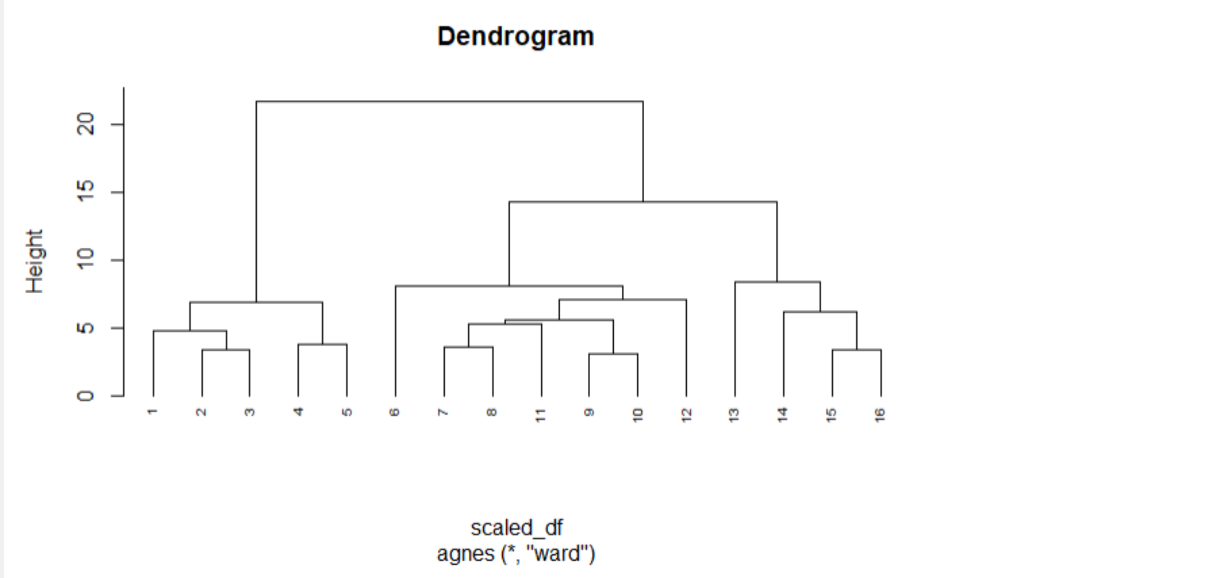
Designated-Best-Use	Class of water	Criteria
Drinking WaterSource without conventional treatment but after disinfection	A	Total Coliforms Organism MPN/100ml shall be 50 or less
		pH between 6.5 and 8.5
		Dissolved Oxygen 6mg/l or more
		Biochemical Oxygen Demand 5 days 20C 2mg/l or less
Outdoor bathing (Organised)	B	Total Coliforms Organism MPN/100ml shall be 500 or less pH between 6.5 and 8.5 Dissolved Oxygen 5mg/l or more
		Biochemical Oxygen Demand 5 days 20C 3mg/l or less
Drinking water source after conventional treatment and disinfection	C	Total Coliforms Organism MPN/100ml shall be 5000 or less pH between 6 to 9 Dissolved Oxygen 4mg/l or more
		Biochemical Oxygen Demand 5 days 20C 3mg/l or less
Propagation of Wild life and Fisheries	D	pH between 6.5 to 8.5 Dissolved Oxygen 4mg/l or more
		Free Ammonia (as N) 1.2 mg/l or less
Irrigation, Industrial Cooling, Controlled Waste disposal	E	pH betwwn 6.0 to 8.5
		Electrical Conductivity at 25C micro mhos/cm Max.2250
Water not suitable for any purpose	F	Total Coliforms Organism greater than 5000 and pH is high

K-means clustering can provide valuable insights into the water quality status of Indian states, such as identifying states with similar water quality profiles or identifying states with poor water quality that require immediate attention. However, like any clustering technique, the interpretation of results should be done with caution, and additional domain knowledge and contextual information should be taken into account.We have performed K Means clustering with 6 clusters on our filtered dataset. Then once the cluster are formed bas we have assigned ranks to the clusters based on their features like Fecal Coliform,BOD, Dissolved O2, PM10, PM2.5 and NO2.

We have obtained the following clusters:

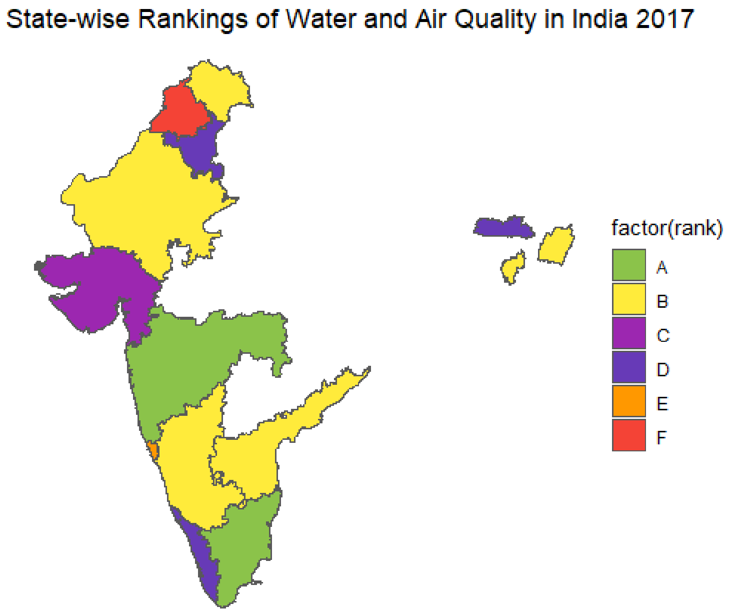


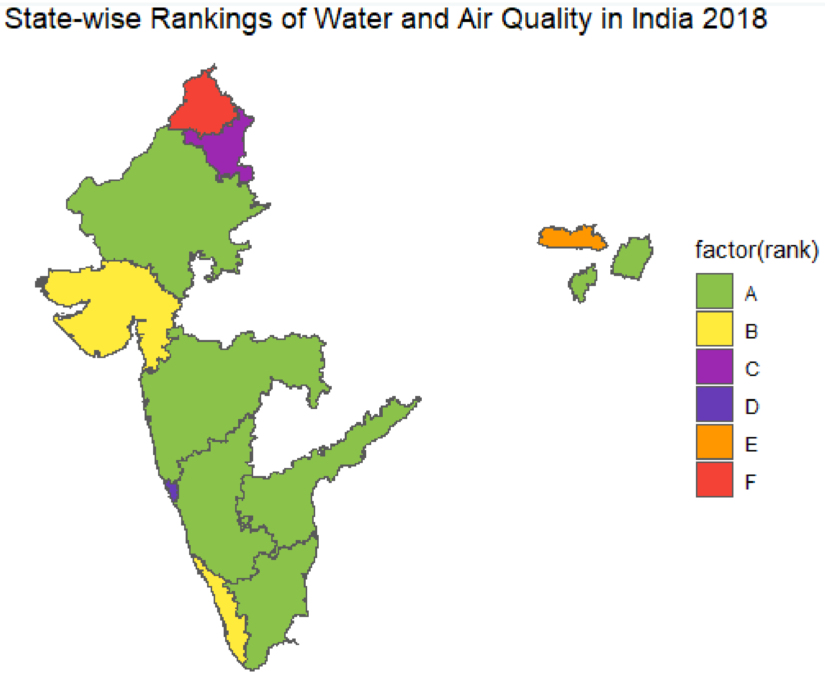
We have performed hierarchical clustering using ward's distance and we obtained same results with 6 clusters.



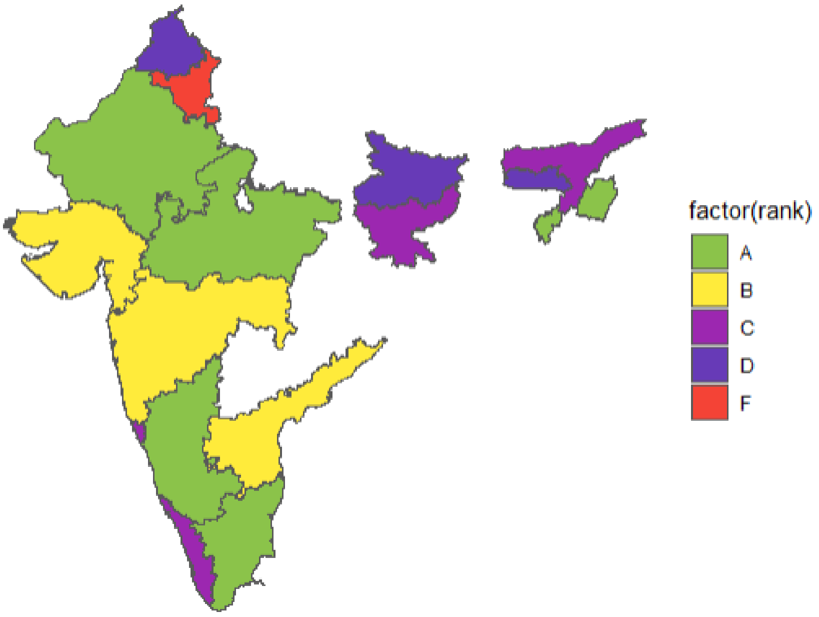
Year On Year Analysis:

We have did the same procedure for 5 years of data and obtained the following results via clustering.

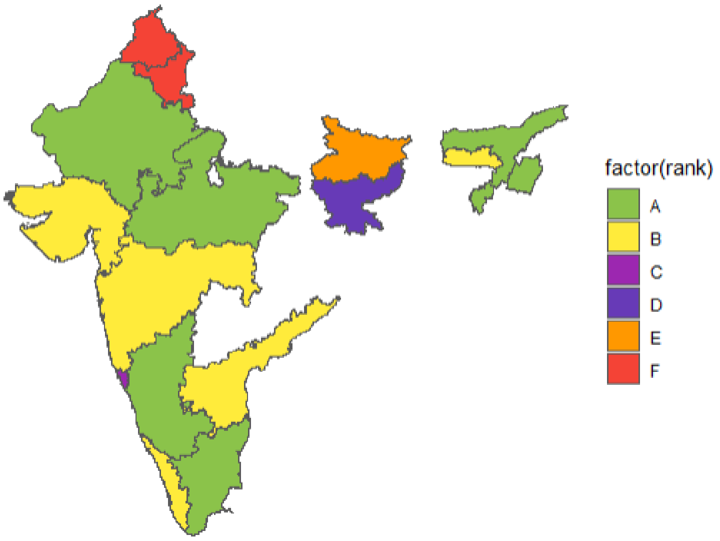

State-wise Rankings of Water and Air Quality in India 2017

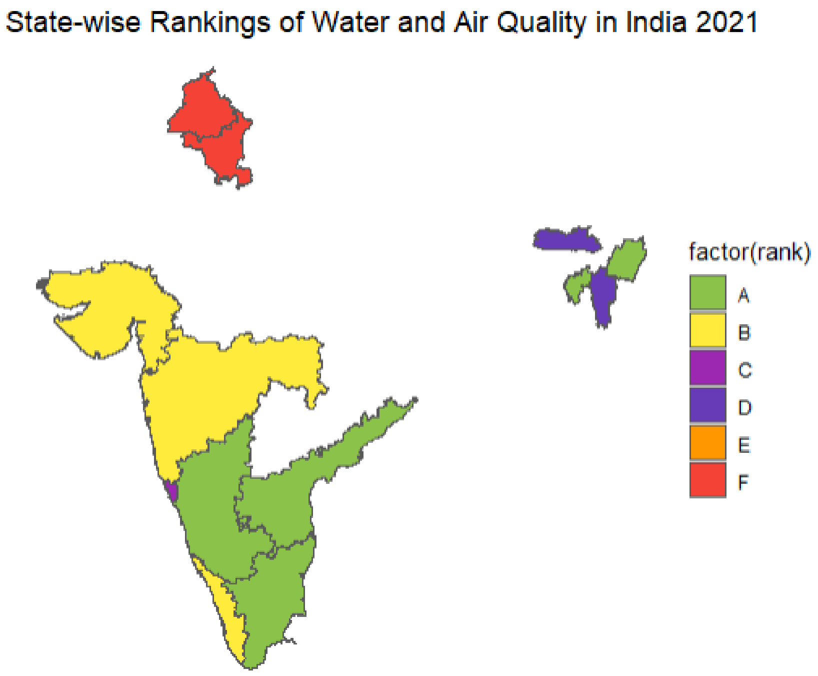

State-wise Rankings of Water and Air Quality in India 2018

State-wise Rankings of Water and Air Quality in India 2019



State-wise Rankings of Water and Air Quality in India 2020





State-wise Rankings of Water and Air Quality in India 2021

Conclusion:

We had data from 16 states available. Further, we have clustered states based on their water and air quality. Two key observations:

- We can see that from 2017 most of the states are performing better in terms of water and air quality.
- We observed that Chandigarh and Haryana are performing very poor in terms of water and air quality and some action should be taken by government authorities to clean up the water

References

Academic Papers:

1. https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1002&context=reu_reports

2. https://www.researchgate.net/publication/351077205_Efficient_Water_Quality_Prediction_for_Indian_Rivers_Using_Machine_Learning

3. https://www.mdpi.com/2306-5338/9/5/92

Articles:

1. https://www.datascience2000.in/2021/10/water-quality-prediction-using-machine.html
2. https://www.researchgate.net/publication/361118196_The_Quality_of_Drinkable_Water_using

_Machine_Learning_Techniques

- https://www.sciencedirect.com/science/article/abs/pii/S0022169419308194

- Existing projects:
- https://www.kaggle.com/code/maujmishra/water-quality-index-prediction
- https://www.kaggle.com/code/imakash3011/water-quality-prediction-7-model/notebook
- Reference Books:
- "Machine Learning Techniques for Water Quality Monitoring" by Yuanyuan Liu, Yongping Li, and
- Junfeng Ji
- "Machine Learning for Environmental Monitoring" by Michael G. Schrlau and Kalyanmoy Deb.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical
- Learning : with Applications in R. New York :Springer, 2013. Print