Water Quality Prediction

Ashmita Gupta : A20512498
Sumanth Donthula : A20519856

Content

Objective

Challenges

Data Preparation and Processing

Modeling

Year wise Results

Objective

This study uses machine learning algorithms to cluster Indian states based on water quality data. K-means clustering was used to group states into six clusters based on their water quality and air quality parameters. However, infrastructure inadequacy, inadequate enforcement of regulations pose challenges to improving water and air quality.

Challenges

Item 1

Preprocess the Air Quality Data as it is present in multiple excel sheets with each feature a sheet, missing values for same states and cities in the data set.

Item 2

Joining the Air Quality Data with the Water Quality Data.

Item 3

Building a clustering model and analysing the year on performance of state.

Data Preparation and Processing

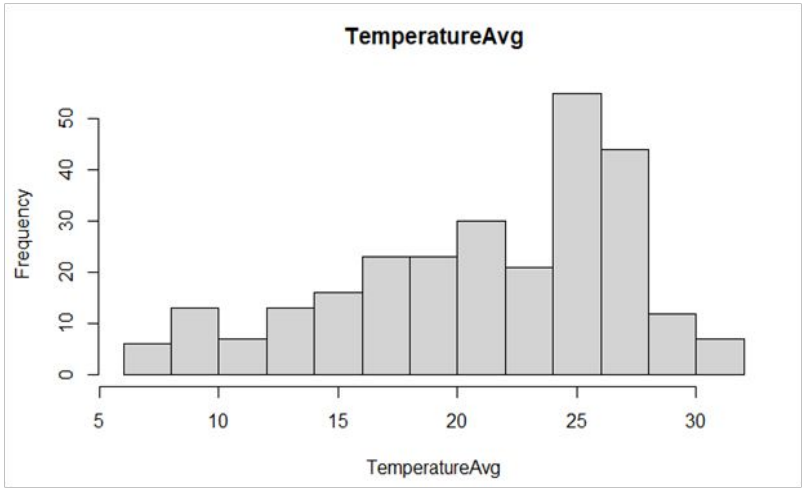
Raw File Water Quality:

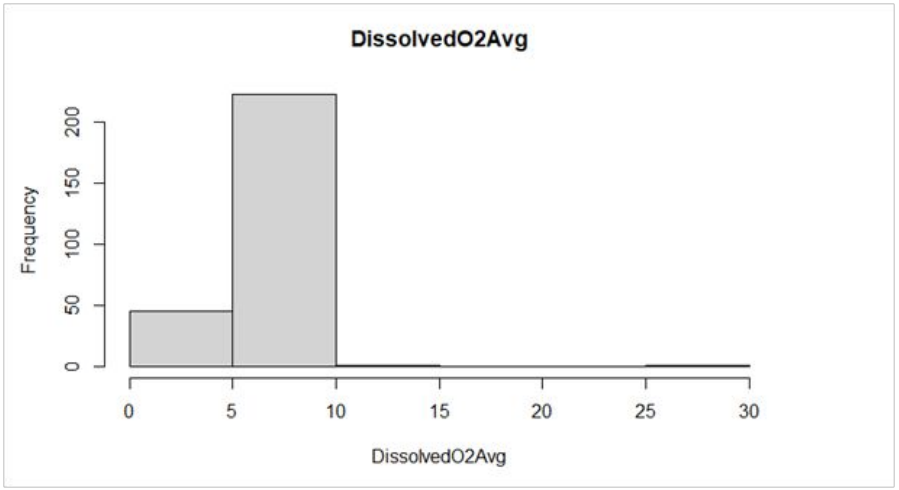
			Temperature		DissolvedO2mg/l		pH
Stationcode	Station Name	State Name	Min	Max	Min	Max	Min
1448	NAGAVALI AT THOTAPALLI REGULATOR,VIZIANAGARAM	ANDHRA PRADESH	26.0	31.0	6.0	8.0	6.6
2352	VAMSADHARA, KALINGAPATNAM,VIZIANAGARAM	ANDHRA PRADESH	26.0	29.0	6.2	8.2	6.2
1393	DAMANGANGA AT D/S OF MADHUBAN,DAMAN	DAMAN 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	7.8	8.6	7.1

Table001 (Page 1-12) Sheet1

Data Preparation and Processing

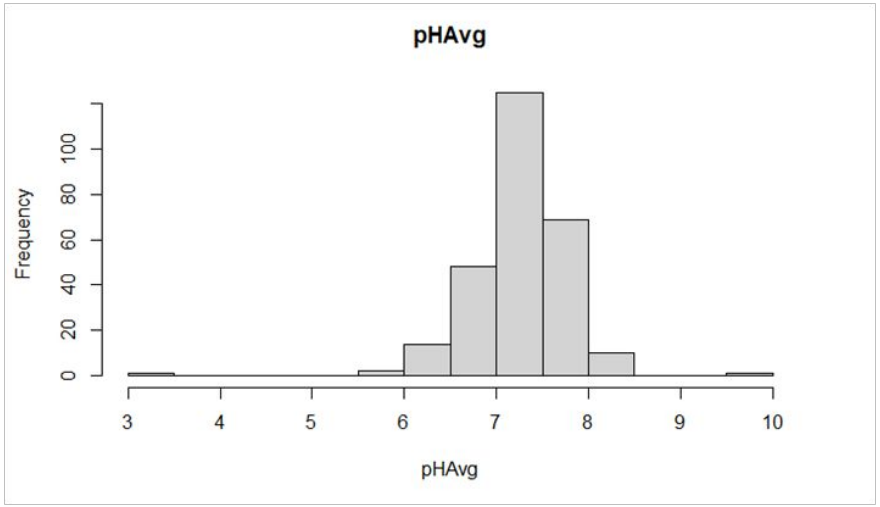
EDA:

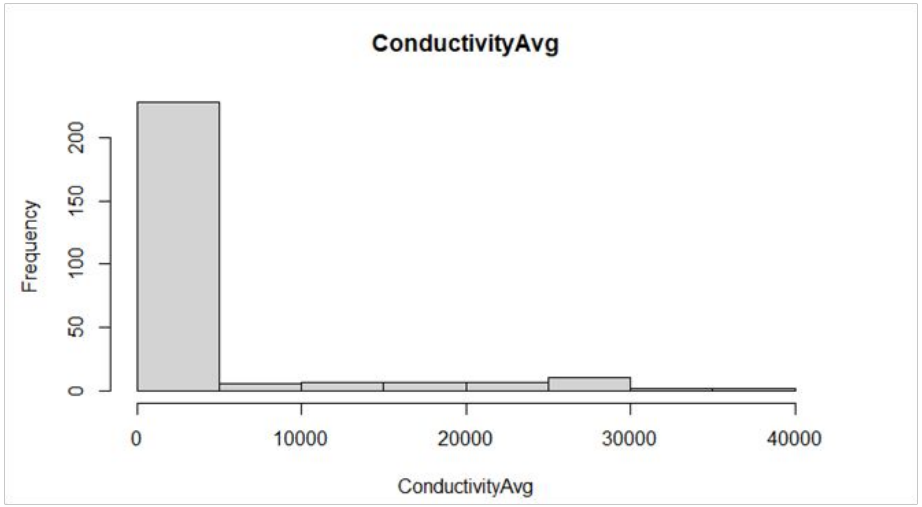




Data Preparation and Processing

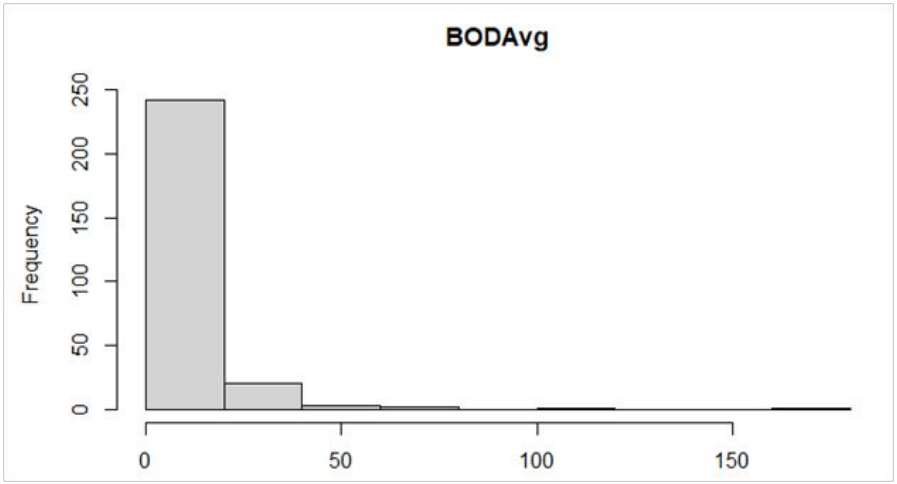
EDA:

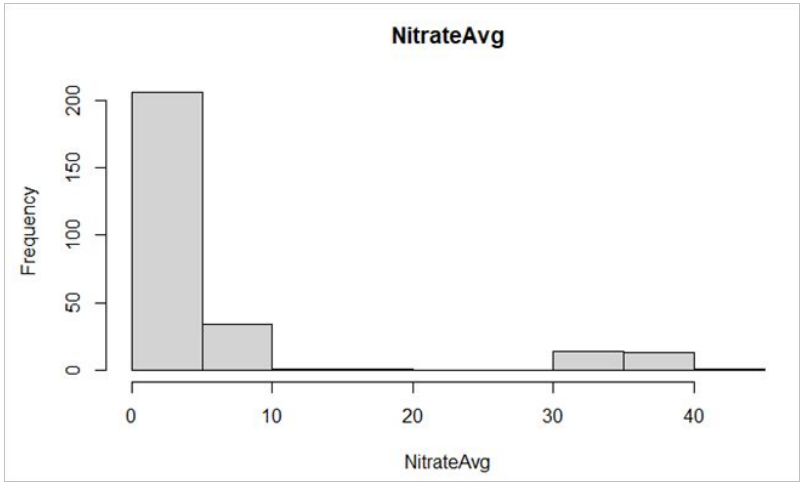




Data Preparation and Processing

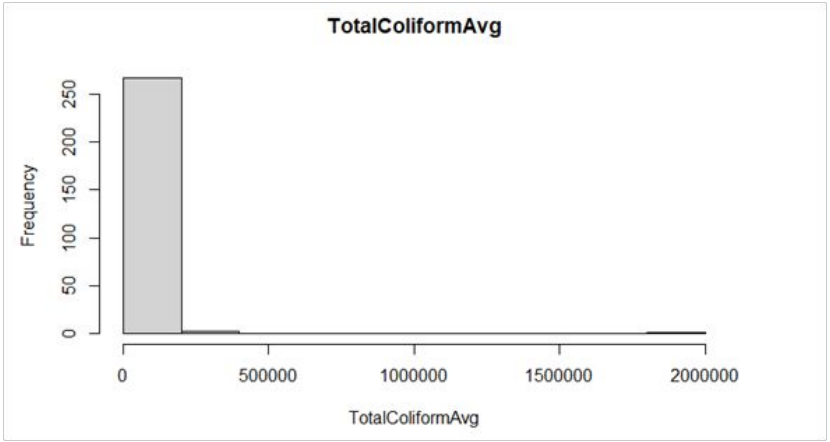
EDA:

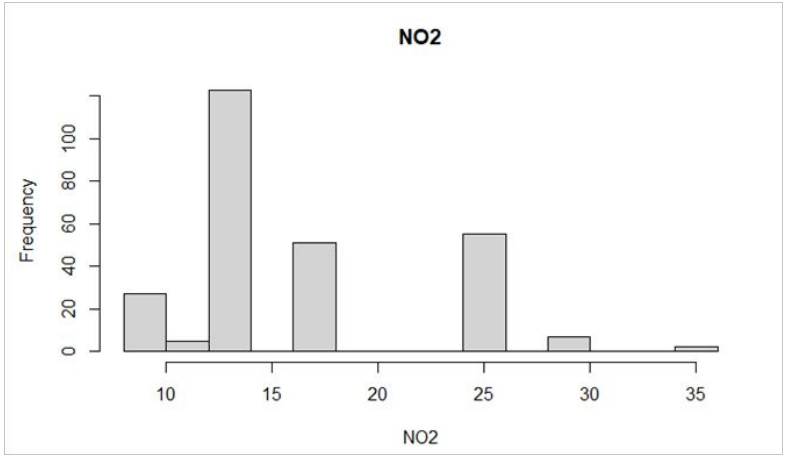




Data Preparation and Processing

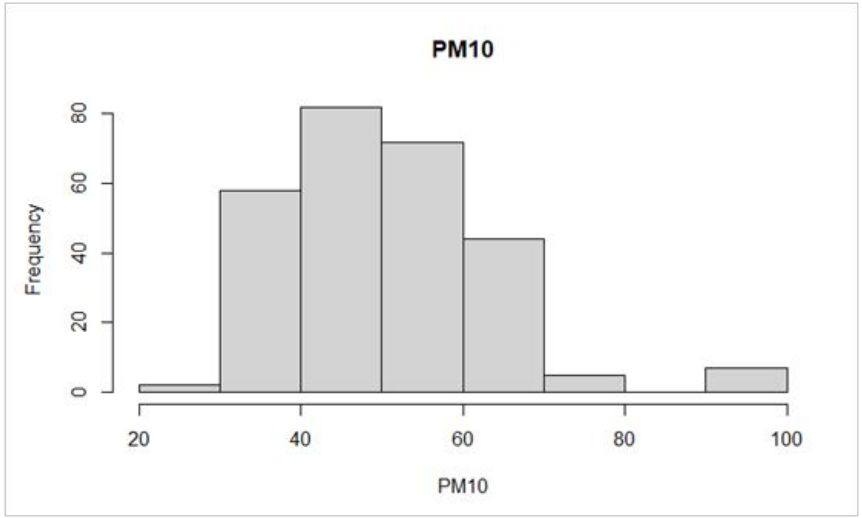
EDA:

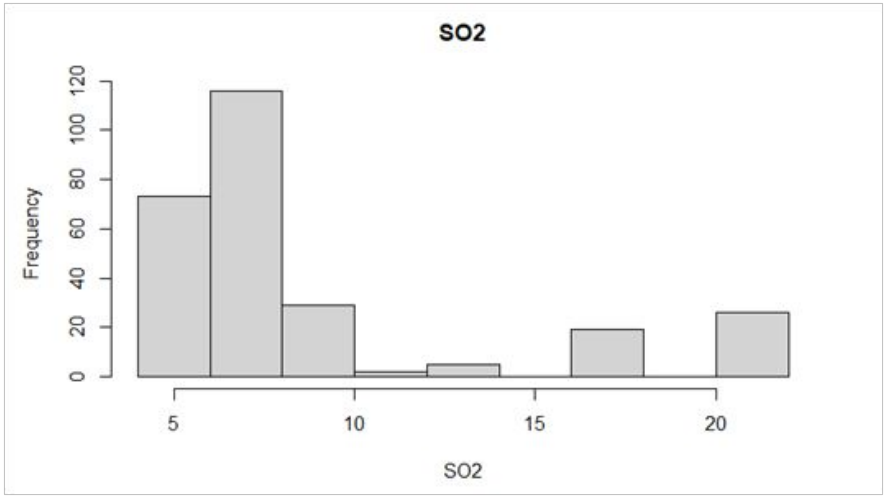




Data Preparation and Processing

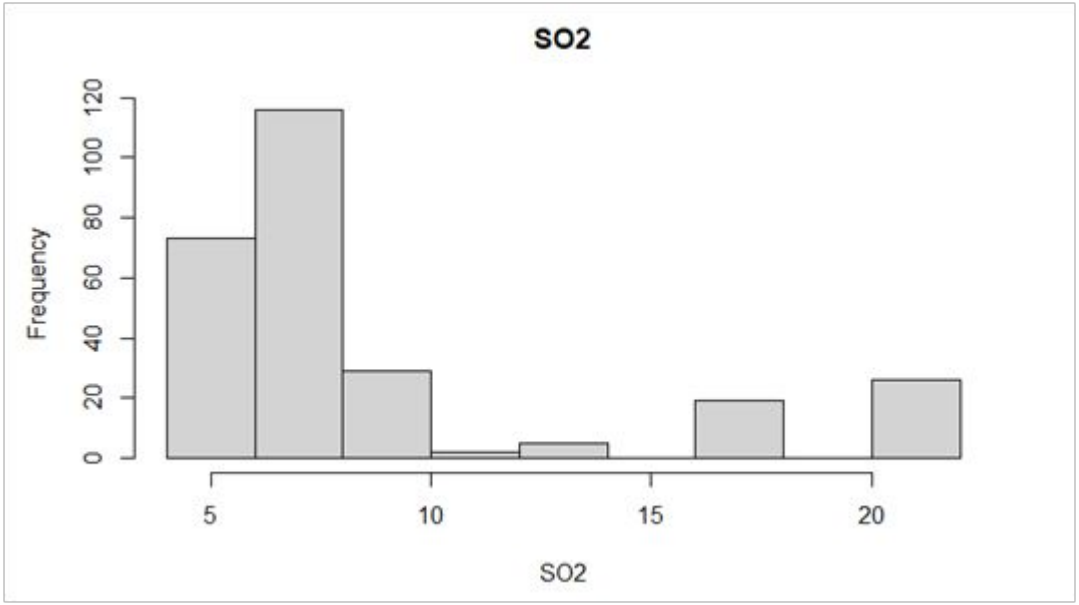
EDA:





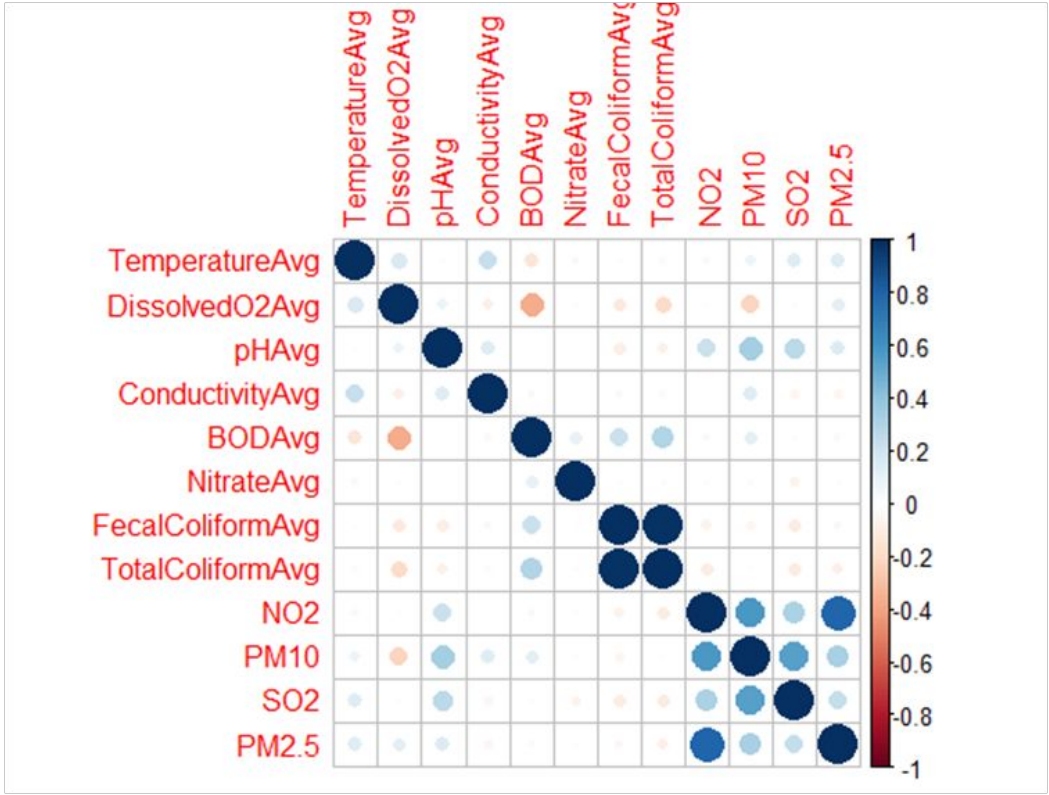
Data Preparation and Processing

EDA:



Data Preparation and Processing

Covariance Matrix:



Data Preparation and Processing

PCA:

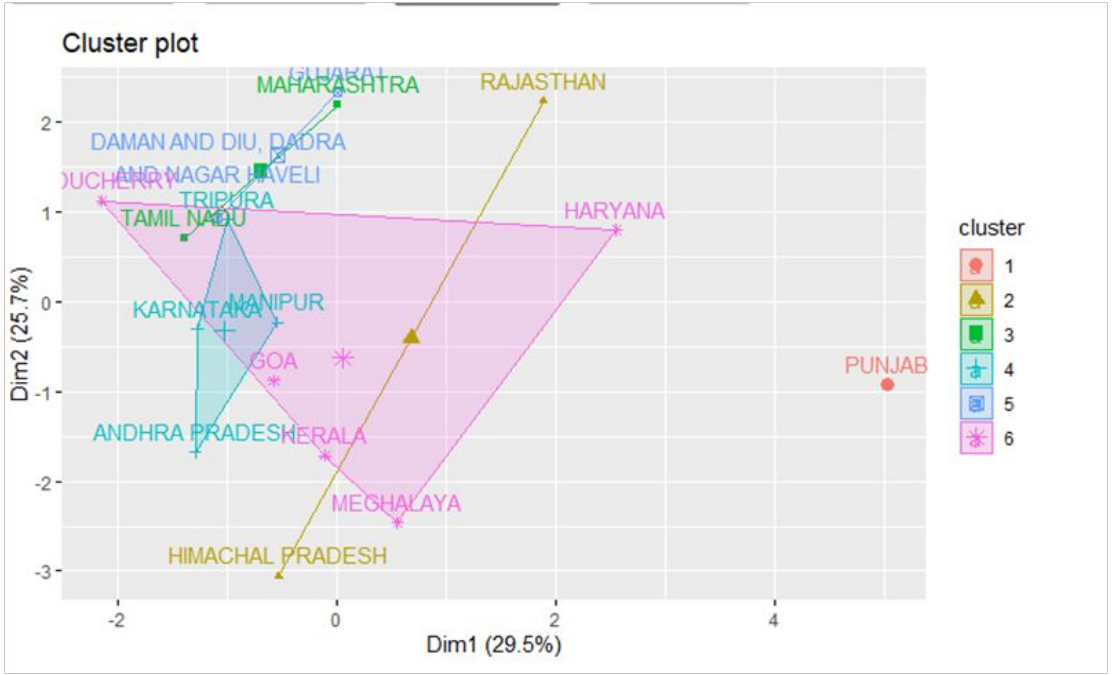
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
TemperatureAvg	0.35247931	-0.16609907	-0.41145789	0.12002435	-0.39078002	-0.28510478	0.22205628
DissolvedO2Avg	0.52797169	0.03667973	0.02672452	0.03326143	0.19101186	-0.14498752	-0.09851537
pHAvg	-0.03466755	-0.37731106	-0.22467152	-0.37981335	0.54072343	0.23271310	-0.03703716
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

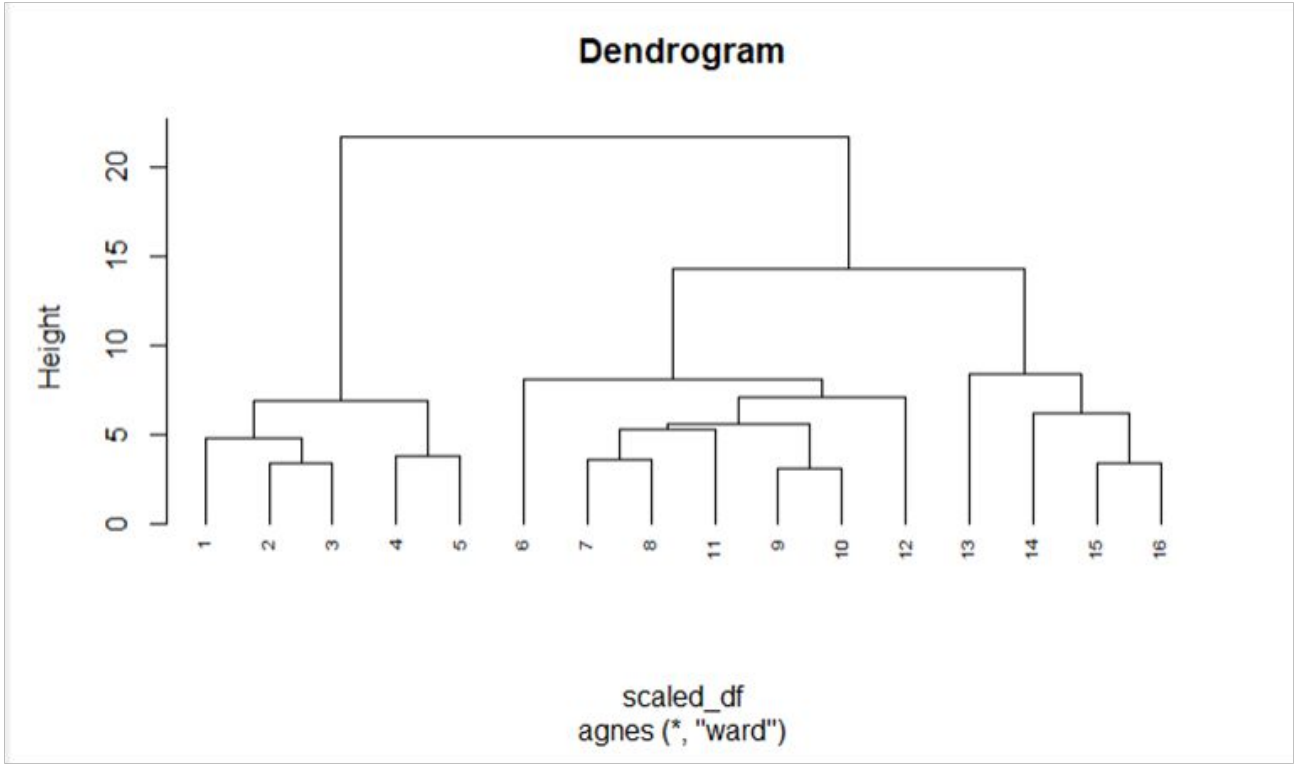
```
 [1] 0.294633176 0.256715245 0.120933762 0.103176527 0.080930646 0.065953880 0.033703160 0.021155724 0.014902656
[10] 0.004966155 0.002929070
```

Modeling

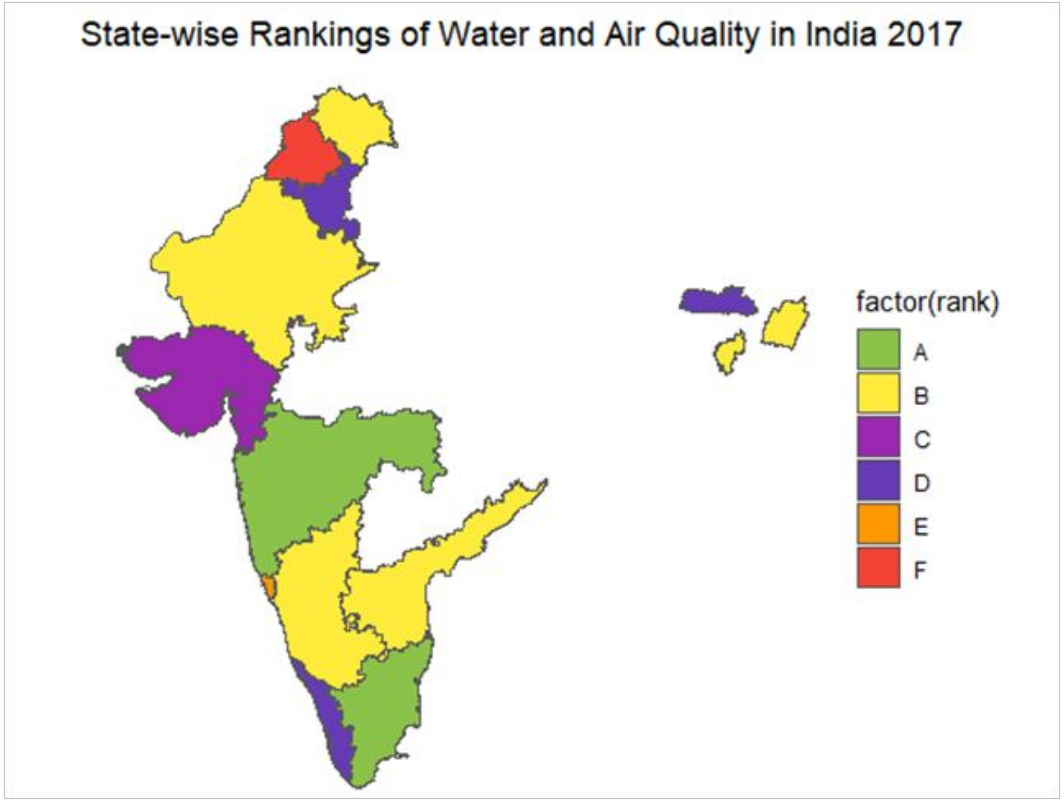
Clustering:



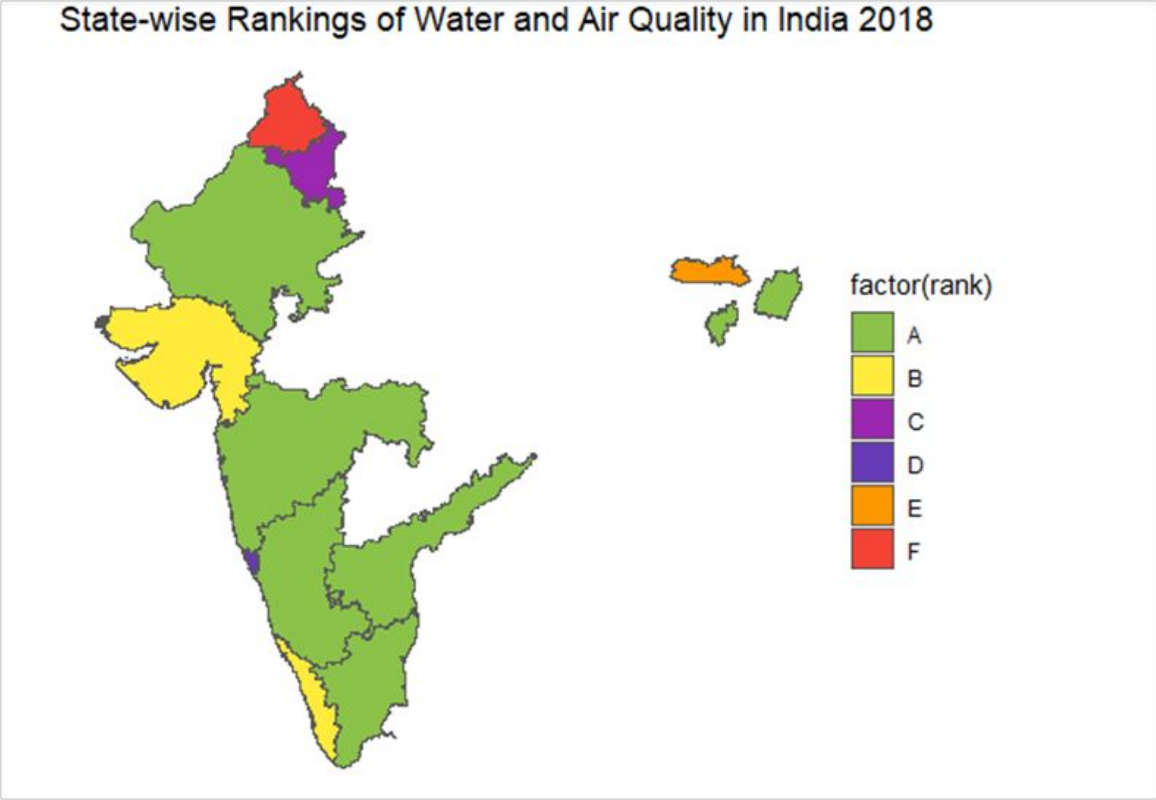
Modeling


Dendrogram
Height
0
5
10
15
20
1
2
3
4
5
6
7
8
11
9
10
12
13
14
15
16
scaled_df
agnes (*, "ward")

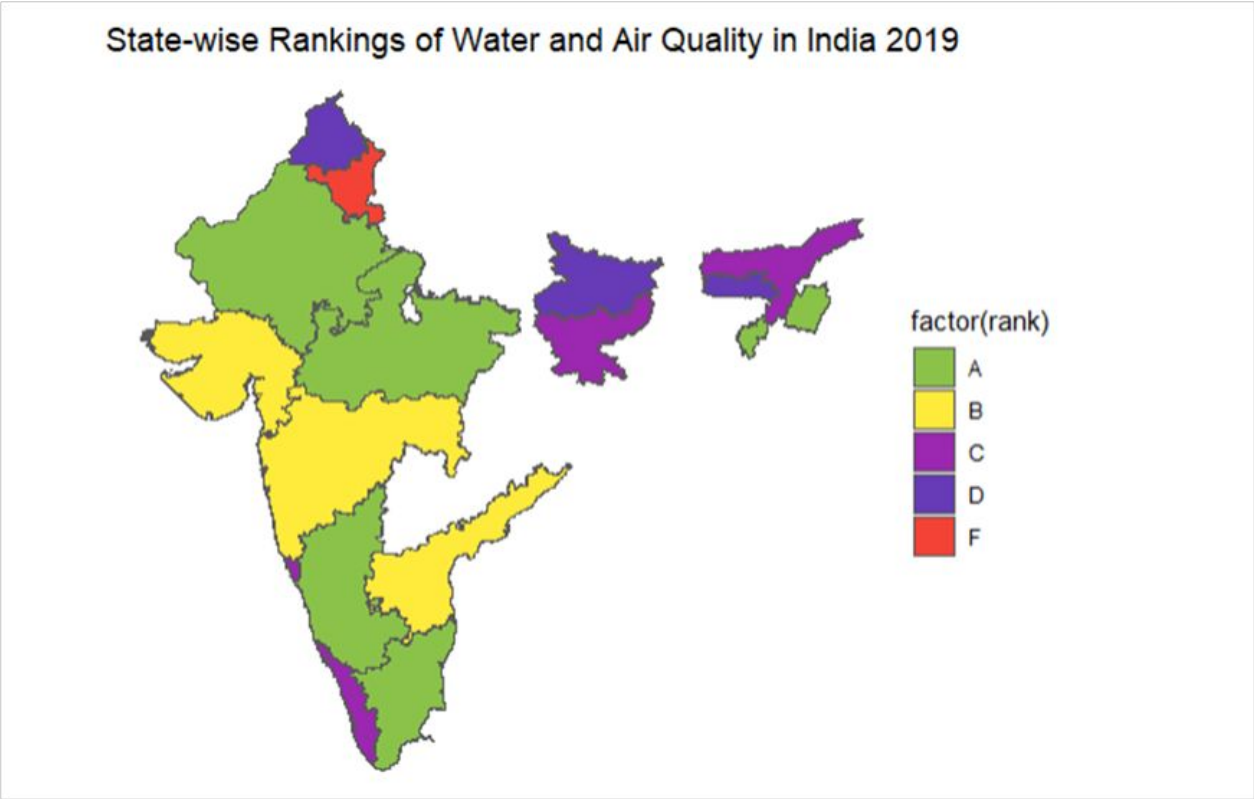
Year On Analysis



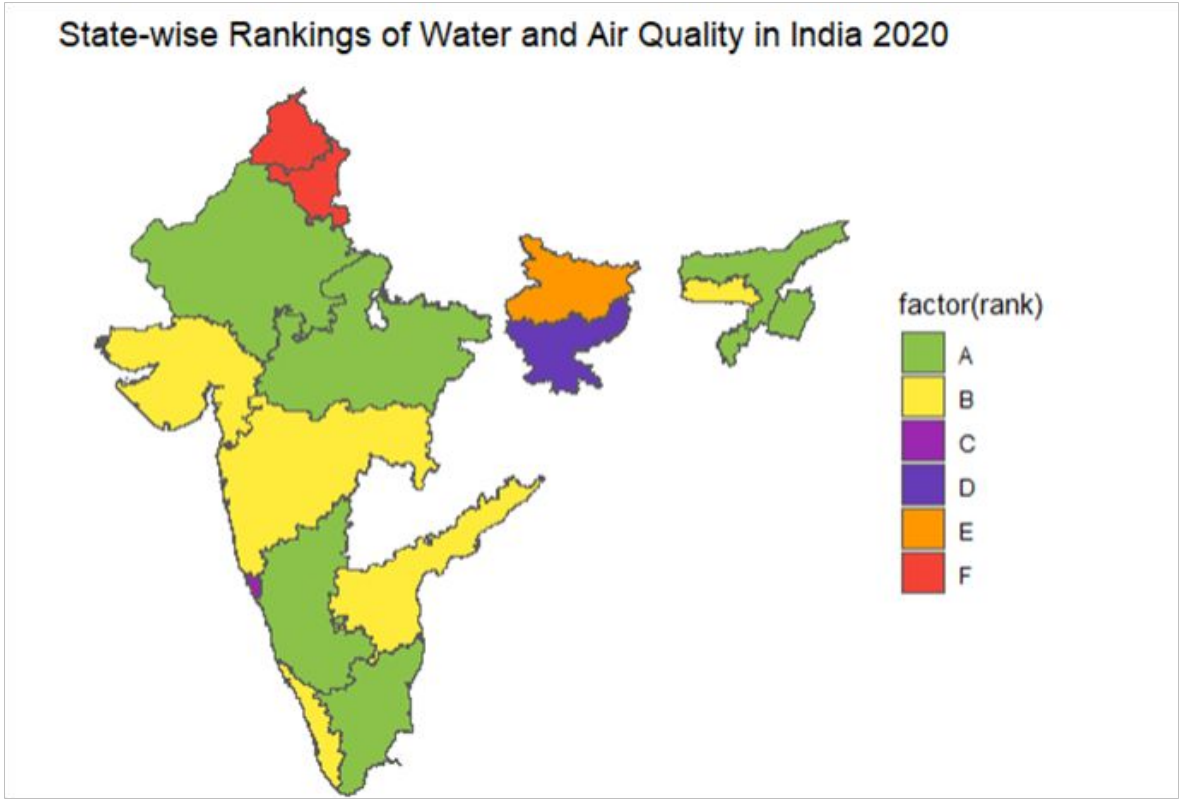
Year On Analysis



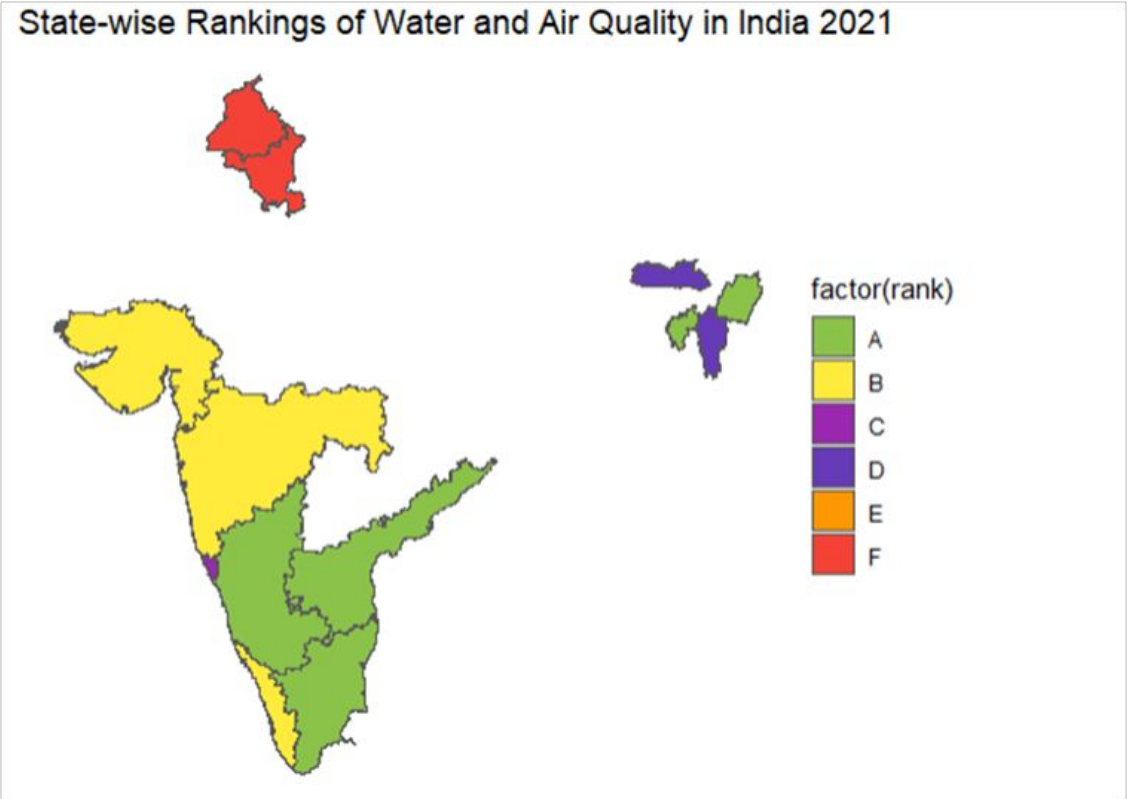
Year On Analysis



Year On Analysis



Year On Analysis



THANK YOU!