ABOVE THE NORM

Are US countries polluted?

A supervised approach for linear regression and classification

Our purpose is to find out a model with relevant variables that describes what affects pollution.

Methods of estimation:

- subset selection
- LASSO
- simple OLS

We try to classify each US country according to a categorical variable.

Methods of classification:

- Tree classifiers
- Random Forest
- K-NN

RESPONSE VARIABLE: AIR QUALITY INDEX

From which we create the variable 'polluted' as yellow, orange or red.

REGRESSORS:

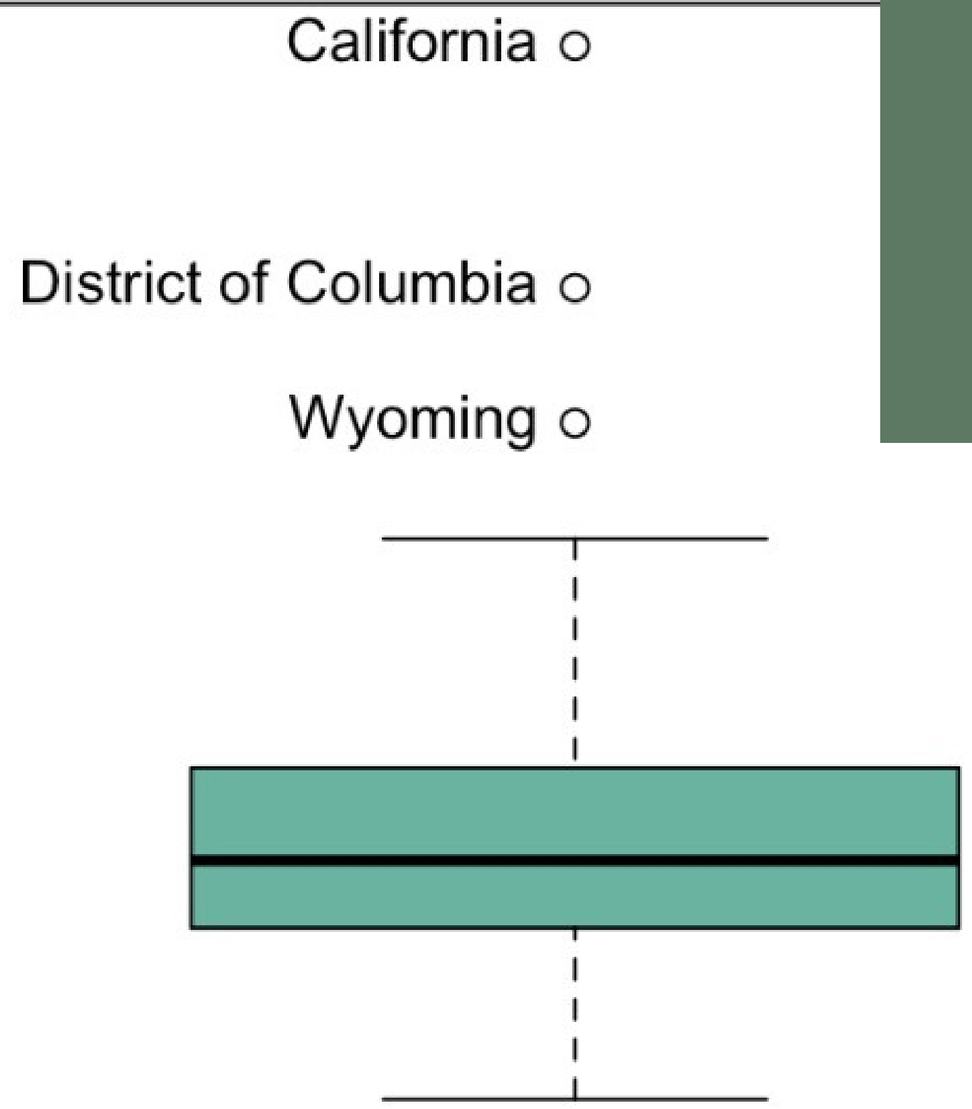
GDP CONTRIBUTION

- accomodation
- construction
- education
- finance
- healtchare
- information
- manufacturing
- mining
- professional
- retail
- transportation
- utilities
- waste

- precipitations
- lockdown
- rural population
- number of factories

DATA UNDERSTANDING

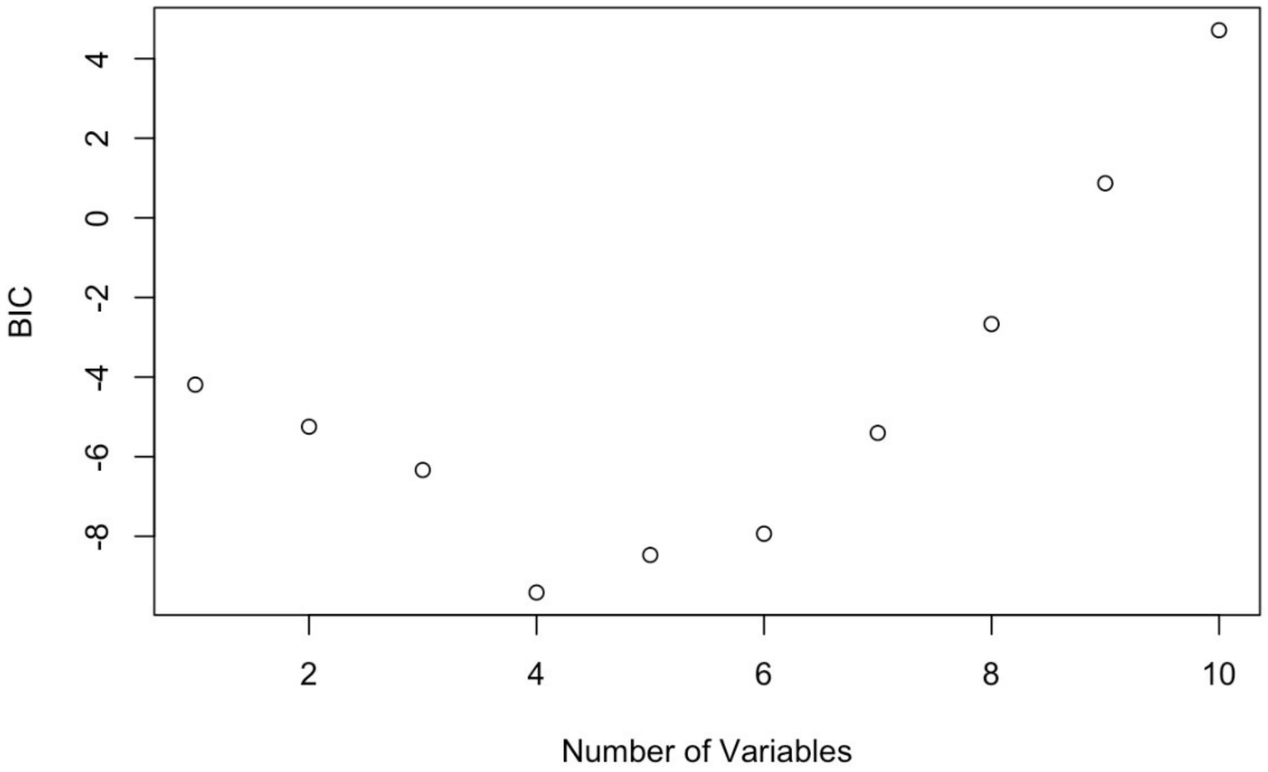
A brief look at the outliers



THREE MOST POLLUTED COUNTRIES: WYOMING, DISTRICT OF COLUMBIA, CALIFORNIA.

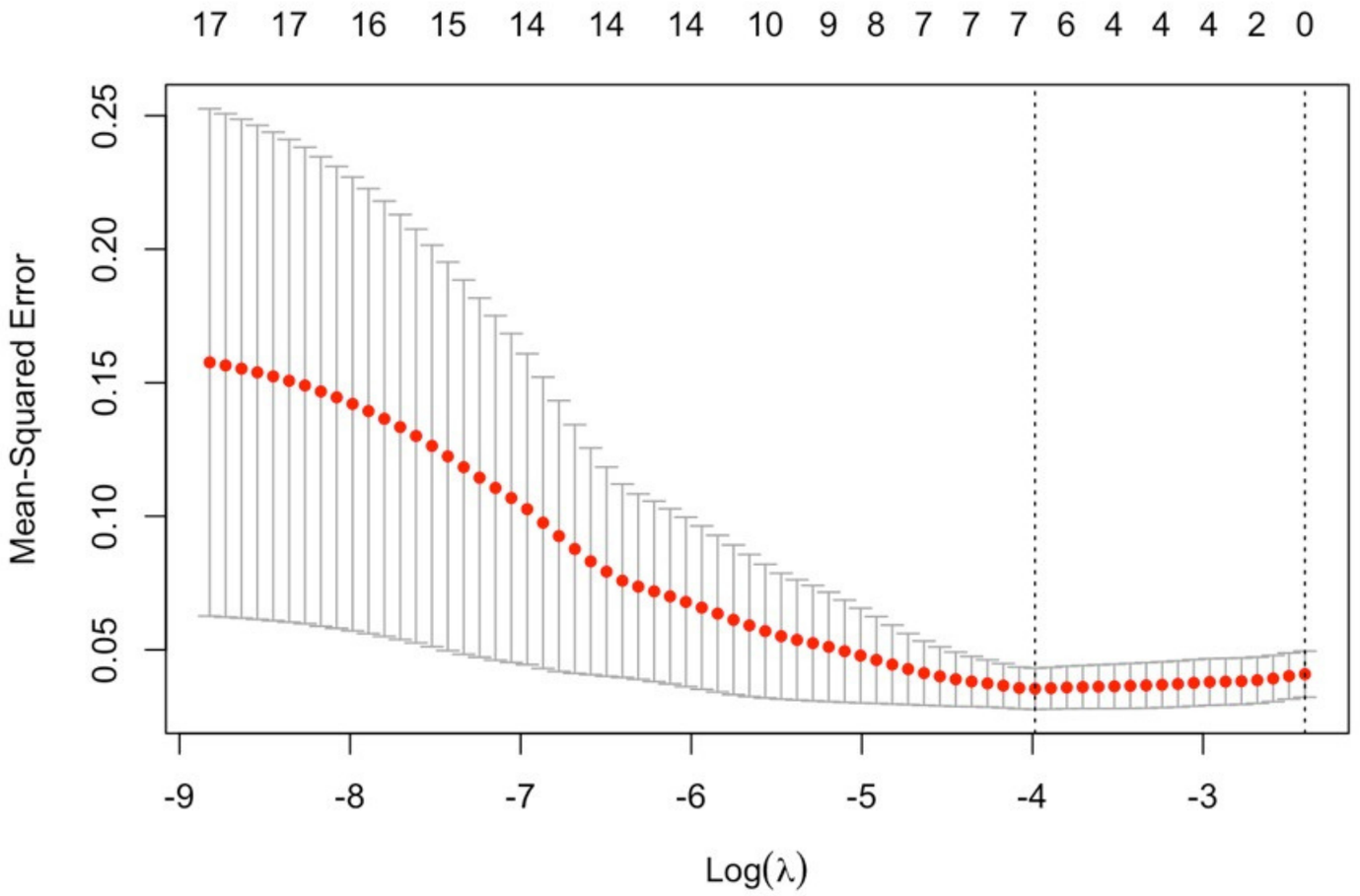
Forward selection

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.47557    0.09556   4.977 2.49e-05 ***
pop_rural       -0.14914    0.15937  -0.936   0.3568
manufacturing    0.40009    0.15580   2.568   0.0155 *
precipitations -0.43971    0.16021  -2.745   0.0101 *
n_factories      0.35065    0.18634   1.882   0.0696 .
```



BIC selects the model with four variables

LASSO MODEL SELECTION



Lasso results

VARIABLES WITH THE HIGHEST RELEVANCE ARE:

1. PRECIPITATIONS
2. POP_RURAL
3. NUMBER OF FACTORIES

```
18 x 1 sparse Matrix of class "dgCMatrix"
                              1
(Intercept)      0.39150936
accommodation    .
construction     .
education        .
finance          .
healthcare       .
information      0.04989436
manufacturing    0.01901585
mining          -0.01451254
professional     .
retail           .
transportation   .
utilities        .
waste            .
precipitations  -0.04991940
lockdown        -0.02005140
pop_rural       -0.03172000
n_factories      0.03915931
```

Post-lasso inference

WITH THE POST-LASSO INFERENCE WE SELECT ONLY TWO SIGNIFICATIVE VARIABLES:

PRECIPITATIONS

NUMBER OF FACTORIES

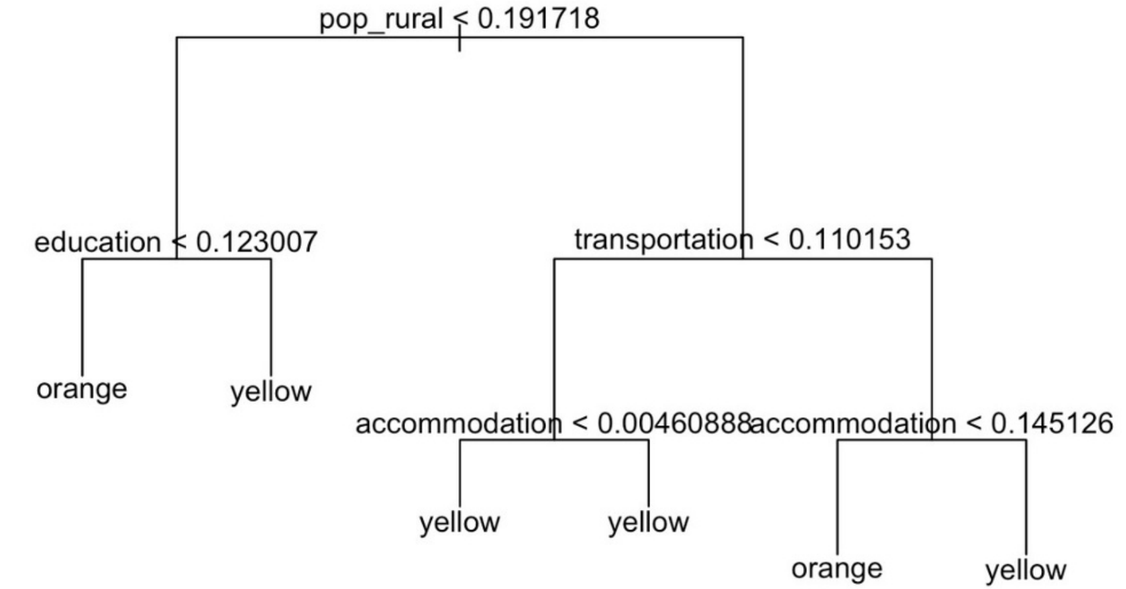
'esting results at lambda = 0.022, with alpha = 0.100

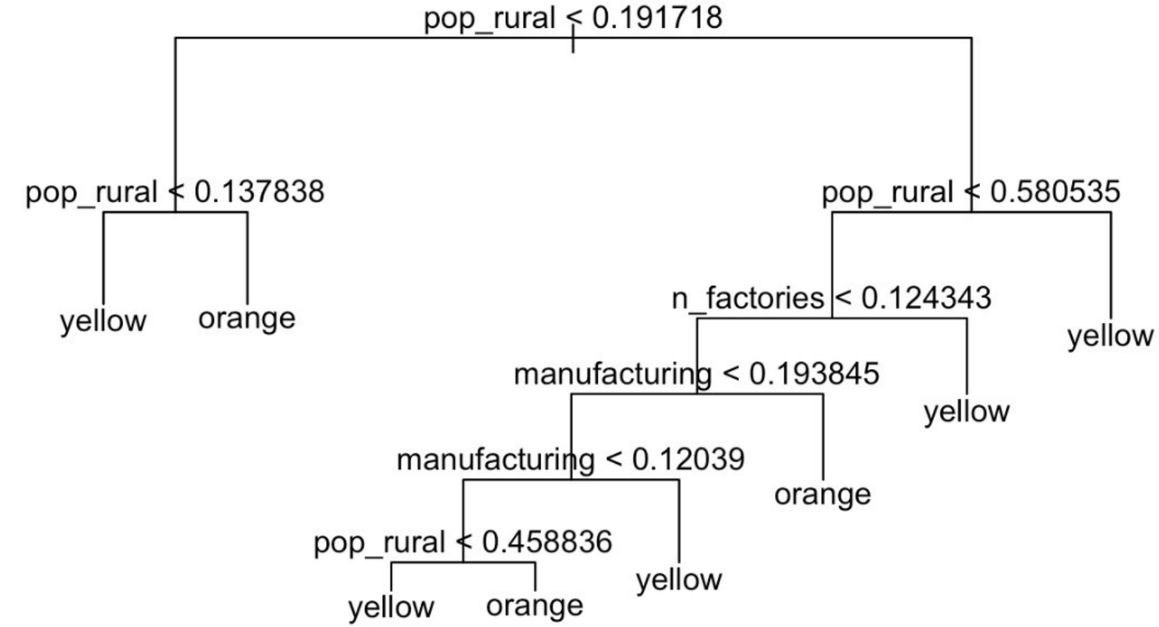
Var	Coef	Z-score	P-value	LowConfPt	UpConfPt	LowTailArea	UpTailArea
1	0.183	1.482	0.096	-0.077	0.889	0.049	0.050
2	-0.155	-0.808	0.427	-0.573	0.635	0.049	0.050
3	-0.098	-0.854	0.625	-0.252	0.855	0.049	0.000
5	0.177	0.765	0.703	-2.269	0.471	0.050	0.049
6	0.179	1.828	0.482	-0.635	0.327	0.000	0.050
7	0.101	1.412	0.152	-0.071	0.283	0.049	0.049
8	-0.057	-1.969	0.055	-0.105	0.002	0.049	0.050
9	-0.466	-1.959	0.565	-0.759	2.326	0.050	0.050
10	-0.352	-1.337	0.586	-0.765	1.833	0.050	0.000
11	0.068	0.491	0.362	-0.614	1.050	0.049	0.050
12	0.181	1.109	0.181	-0.268	1.173	0.050	0.050
13	0.245	0.693	0.788	-5.559	0.776	0.050	0.049
14	-0.076	-2.891	0.004	-0.131	-0.030	0.050	0.048
15	-0.066	-2.227	0.128	-0.114	0.033	0.050	0.049
16	-0.058	-1.680	0.148	-0.115	0.037	0.049	0.050
17	0.099	3.034	0.003	0.044	0.176	0.049	0.050

TREE CLASSIFIERS

The first graph on the right presents a misclassification rate equal to 17%

The second graph on the right, that has four variables, outputs a misclassification rate equal to 16%.





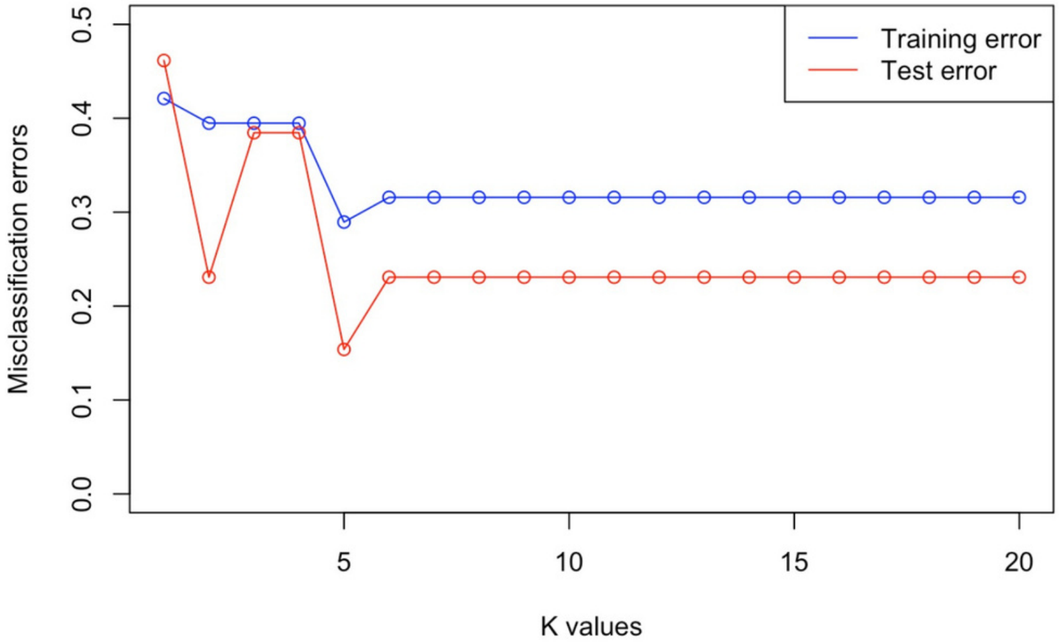
Random Forest

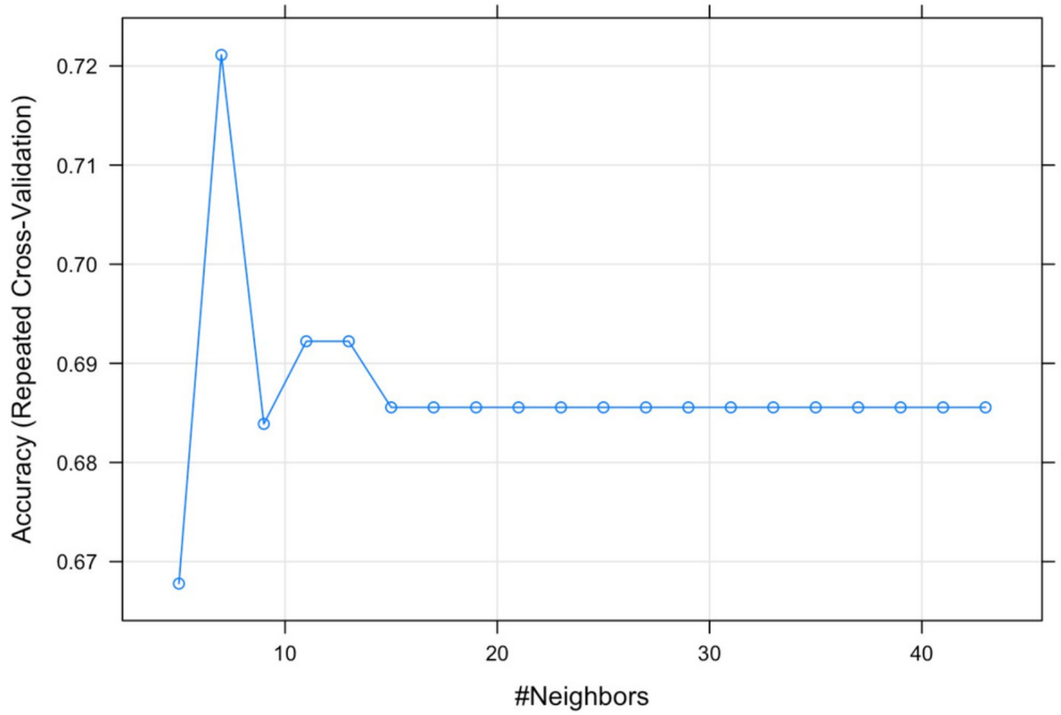
	dt_test_labels		
pred_rf	yellow	orange	red
yellow	10	1	0
orange	3	2	0
red	0	0	0

- 10 / 11 countries are well predicted as yellow
- 2 / 5 countries are well predicted as orange
- No contries are classified as red.

The overall accuracy is 75%

K-Nearest Neighbor





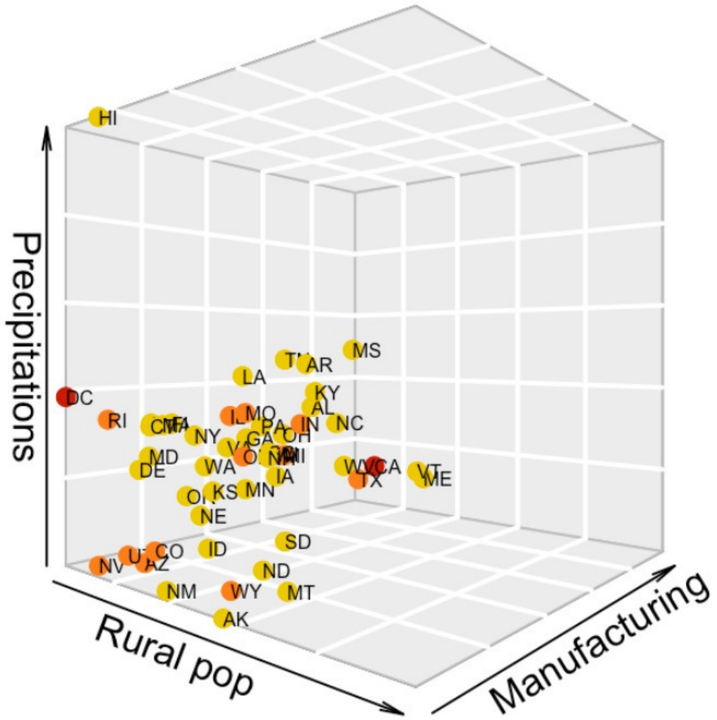
We make the train and test error as a function of hyperparameter k

5 seems the k-value that minimizes the test error while it maximizes the accuracy. (with repeated cross validation)

We try to run the K-NN algorithm on three most important features selected from our dataset:

1) rural population
2) precipitations
3) manufacturing

K-NN WITH THE THREE MOST RELEVANT VARIABLES



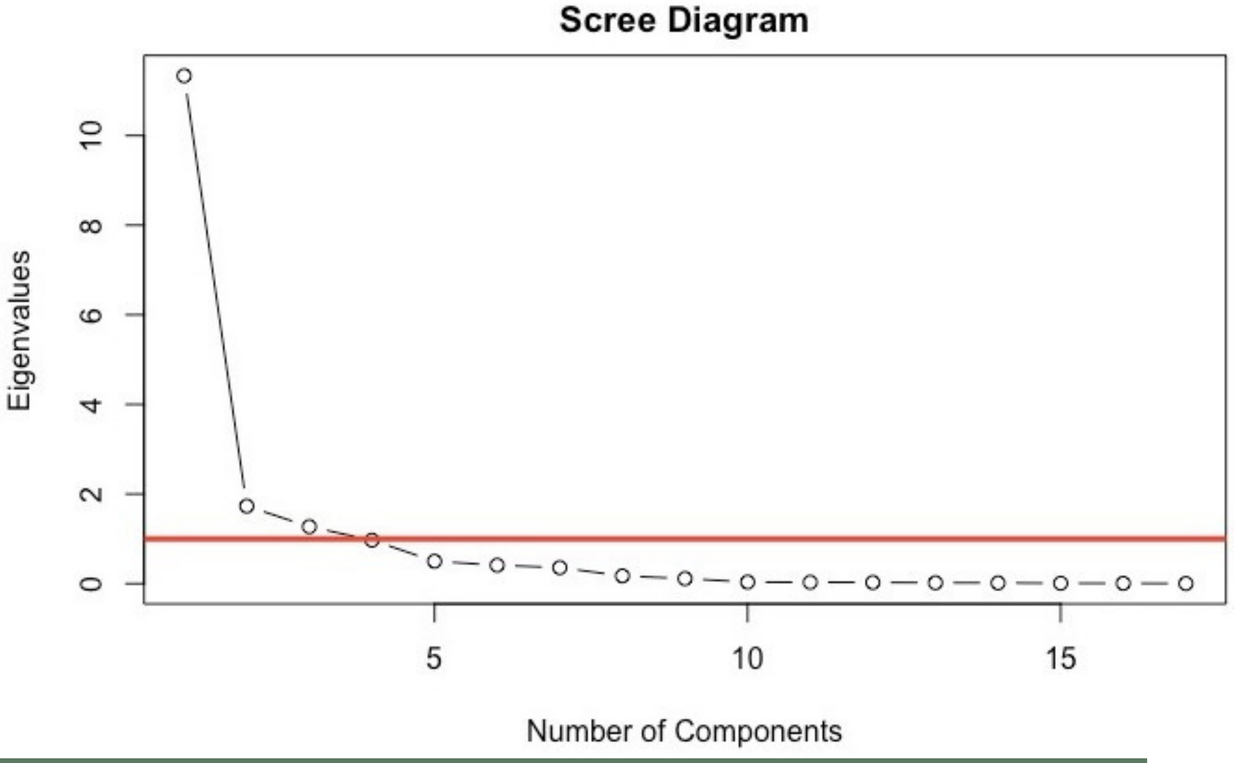
Our desire was to have a more interpretable classification

UNSUPERVISED APPROACH

We use PCA and hierarchical clustering in order to visualize and understand similarities between States.

```
 [1] 0.6667693615 0.1017681849 0.0747747607 0.0570338016 0.0293750682
 [6] 0.0241220129 0.0209919683 0.0101389749 0.0066064876 0.0020078697
[11] 0.0017756346 0.0015292575 0.0011555602 0.0009580432 0.0004755402
[16] 0.0003973726 0.0001201014
```

	eigenval	%var	%cumvar
[1,]	11.335	66.677	66.677
[2,]	1.730	10.177	76.854
[3,]	1.271	7.477	84.331
[4,]	0.970	5.703	90.035
[5,]	0.499	2.938	92.972
[6,]	0.410	2.412	95.384
[7,]	0.357	2.099	97.484
[8,]	0.172	1.014	98.497



Principal Component Analysis

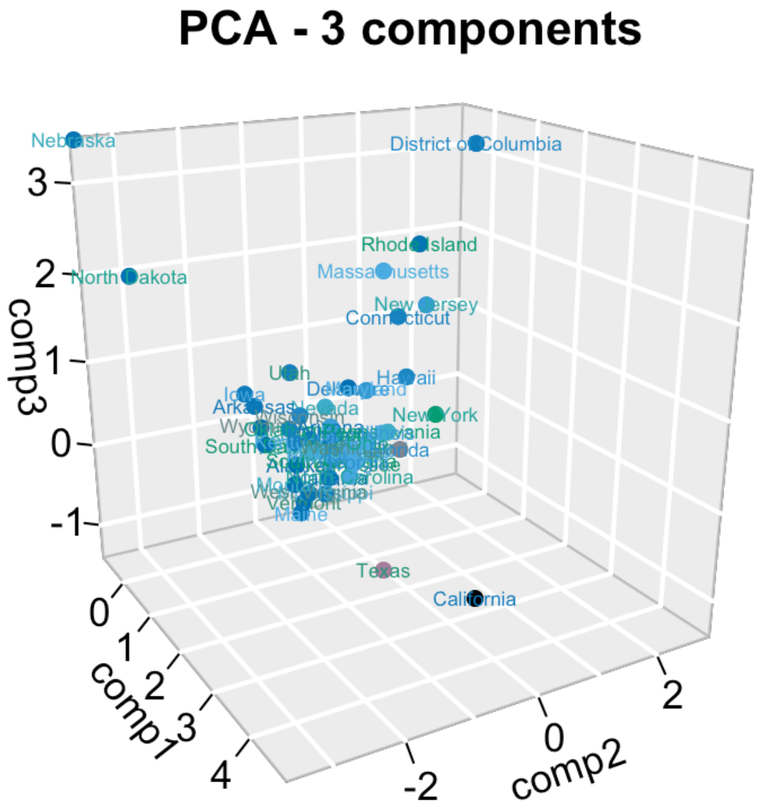
We found that the first component explains 65% of the variance .

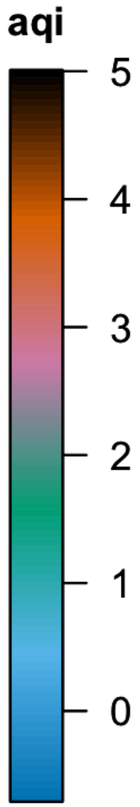
There are other two components deriving from an eigenvalue just above one

	comp1	comp2	comp3	communality
accommodation	0.972	0.039	0.034	0.947461
construction	0.975	0.072	0.060	0.959409
education	0.880	-0.086	-0.152	0.804900
finance	0.959	0.025	-0.044	0.922242
healthcare	0.986	0.014	-0.004	0.972408
information	0.903	0.069	-0.027	0.820899
manufacturing	0.905	0.093	0.118	0.841598
mining	-0.126	0.654	-0.582	0.782316
professional	0.989	0.004	-0.037	0.979506
retail	0.982	0.056	0.059	0.970941
transportation	0.965	0.107	0.048	0.944978
utilities	0.982	0.060	0.031	0.968885
waste	0.989	0.035	0.030	0.980246
precipitations	-0.078	-0.435	-0.002	0.195313
lockdown	0.254	-0.679	0.437	0.716526
pop_rural	-0.480	0.410	0.534	0.683656
n_factories	0.080	-0.662	-0.635	0.847869

Generating the loadings of three components

PCA HAS A HEAVY ROLE IN EXPLAINING ALMOST EVERY VARIABLE.





Plot of the scores of the three components for each State

CLUSTERING

WE MADE THREE TYPE OF CLUSTERING:

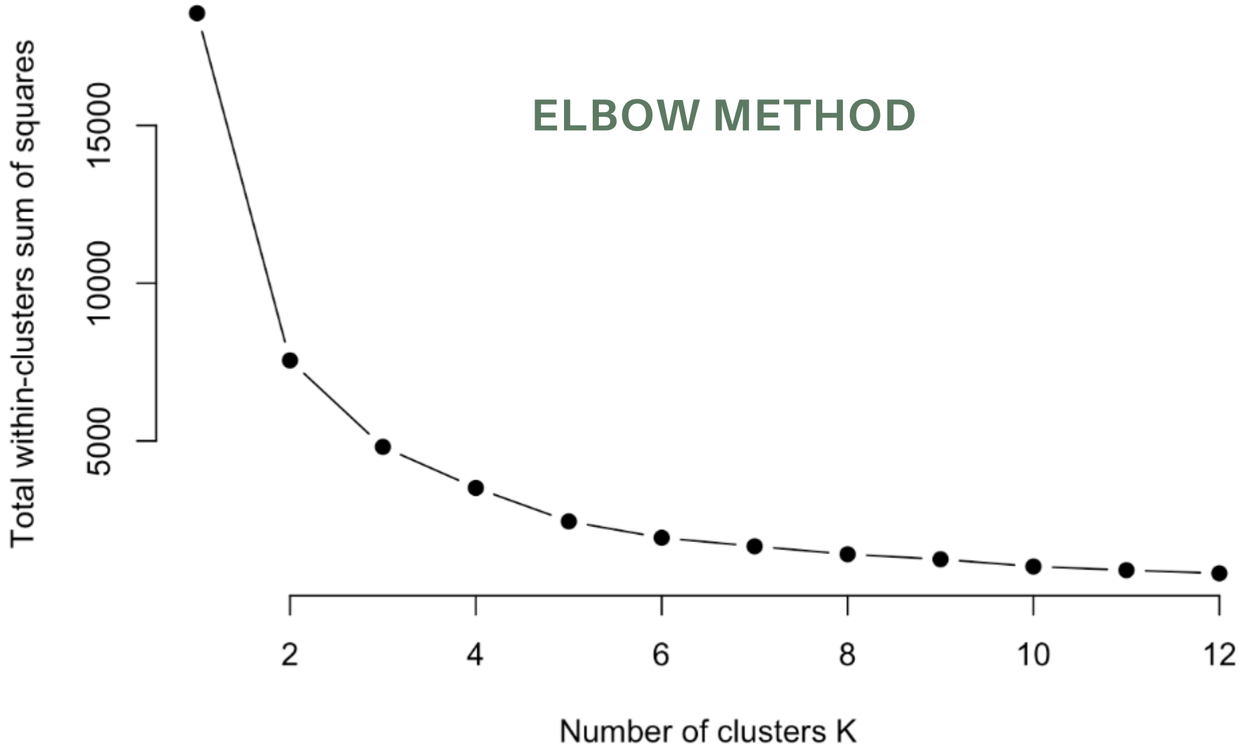
1. AVERAGE LINKAGE
2. COMPLETE LINKAGE
3. WARD LINKAGE

WE CHOOSE THE WARD METHOD BECAUSE:

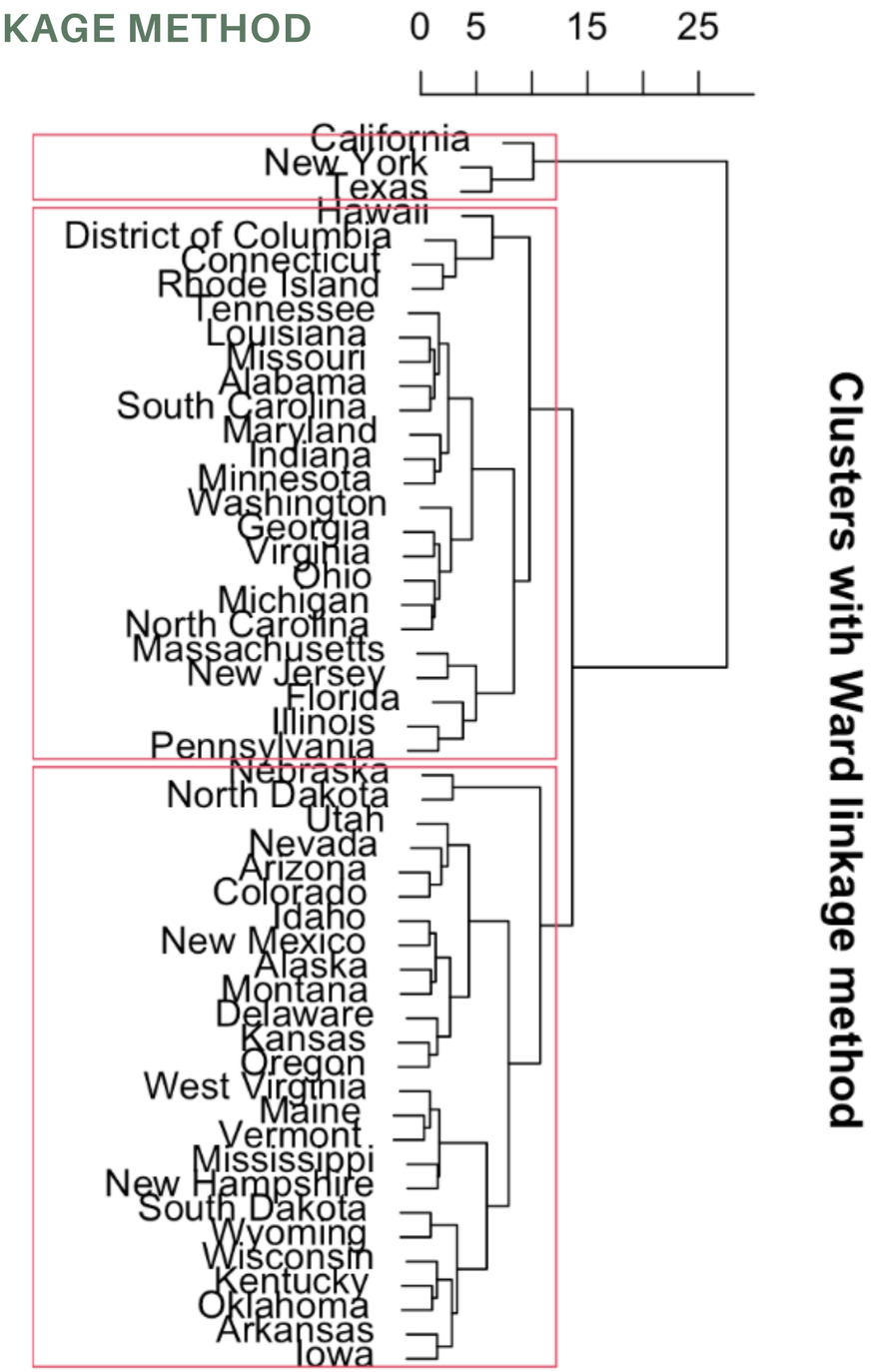
- It reduces the distance of the outliers
- it is the one that divides the States in a clearer way

CLUSTERING

ELBOW METHOD



WARD LINKAGE METHOD



A COMPARISON OF THE CLUSTERING METHOD WITH THE CLASSIFICATION OF POLLUTION

The way in which we cluster the countries does not match the way in which we classified on a discrete scale US countries.

In conclusion, the common factor among countries is not driven by pollution

cluster	*cl* yellow	orange	red	***Total***
1	17 73.9 %	5 21.7 %	1 4.3 %	23 100 %
2	18 72 %	7 28 %	0 0 %	25 100 %
3	1 33.3 %	1 33.3 %	1 33.3 %	3 100 %
Total	36 70.6 %	13 25.5 %	2 3.9 %	51 100 %

χ^2=8.503 · df=4 · Cramer's V=0.289 · Fisher's p=0.136

THANK YOU FOR YOUR ATTENTION!

Andrea, Lorenzo, Margherita, Mathias, Gabriele, Matteo