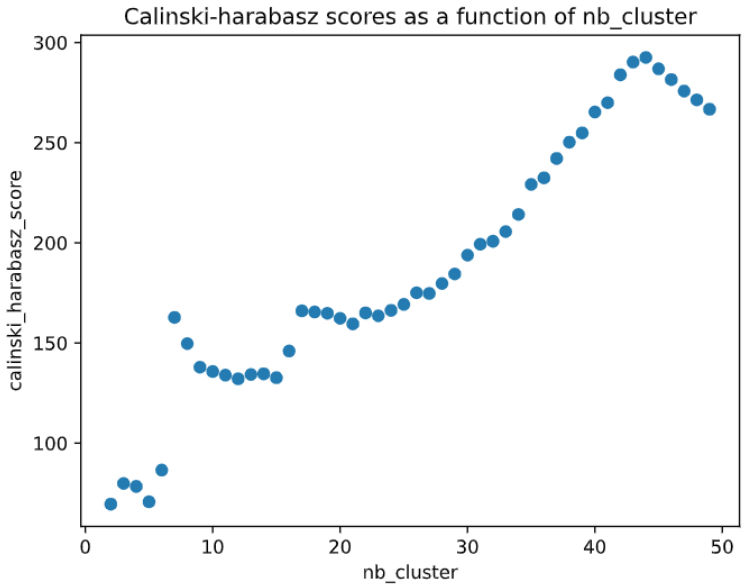
Part 5 : Application of unsupervised learning

For this exercise, I tried a lot of different datasets before to find the one I finally worked on.

FIRST DATASETS EXPERIMENTS:

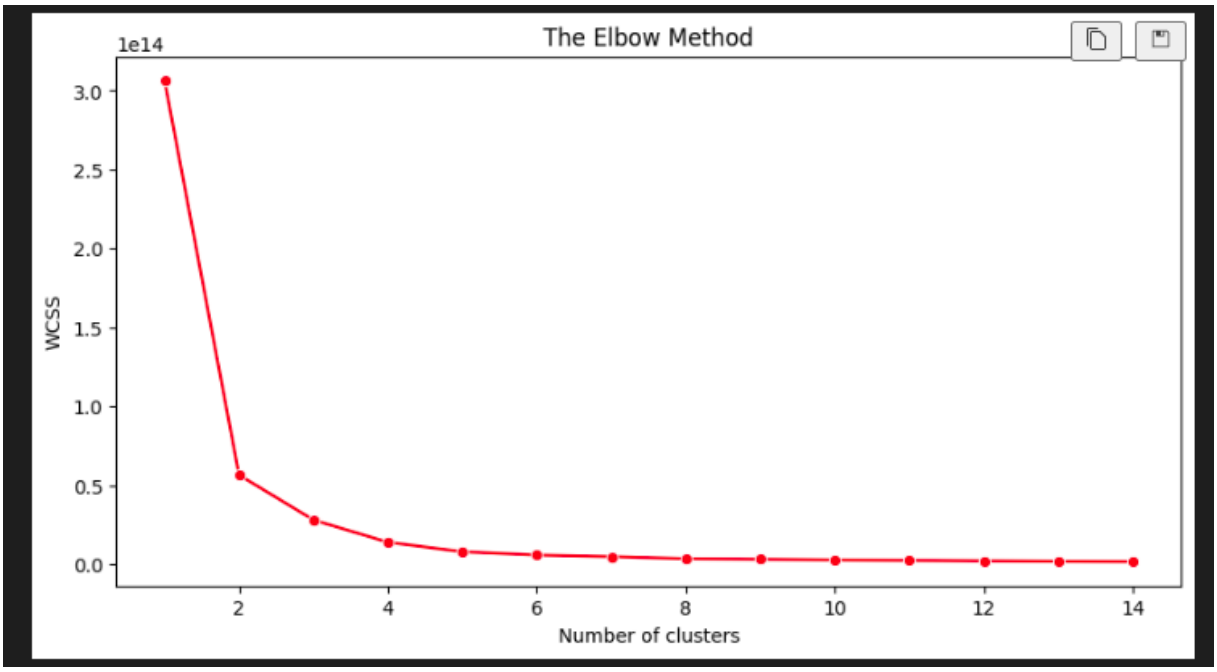
I wanted to do clustering, so I searched in links provided about interesting datasets for experimenting some clustering algorithms. I performed it and spent a lot of time on it before understanding these datas were far too complicated for a few experienced developer like me.

After some experimentations I got some interesting scores for my clustering but I found out that it gaves me as many clusters as my number of people taking part in the survey :



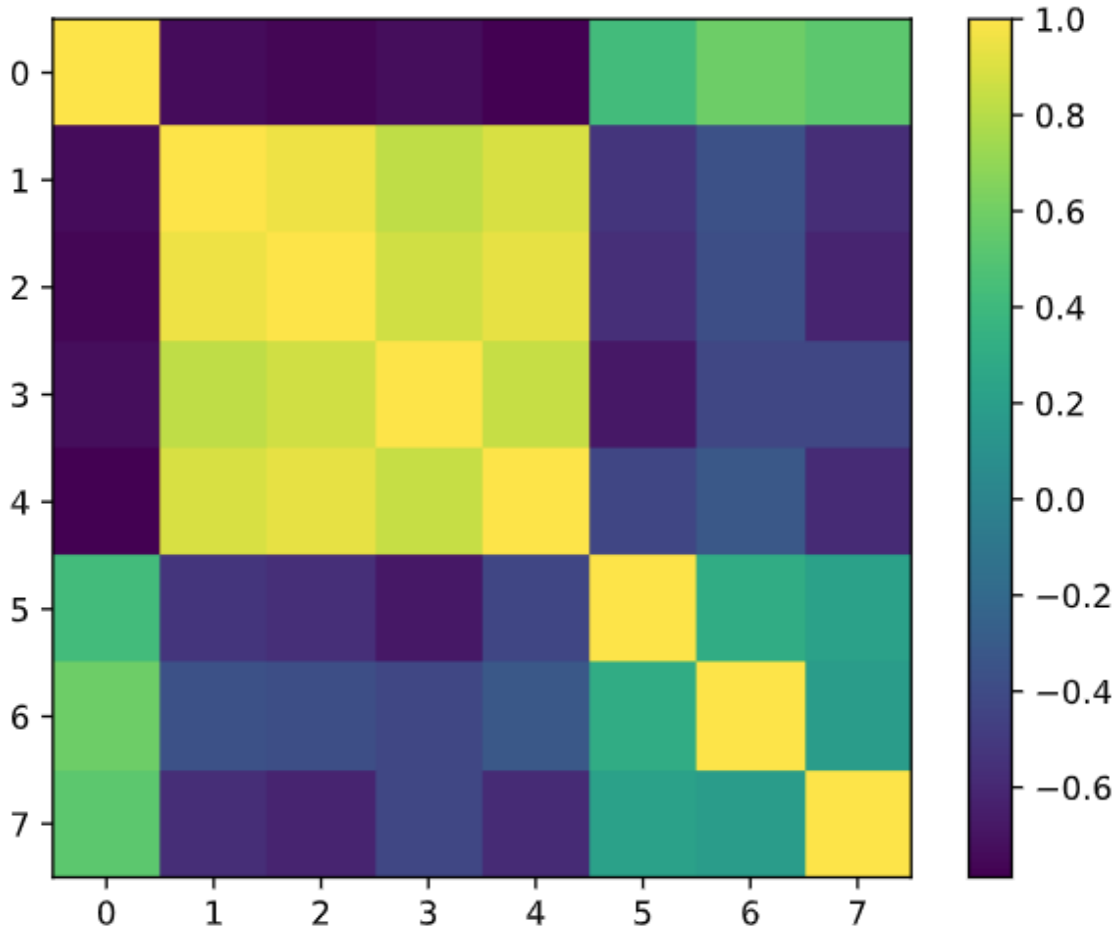
I was really frustrated about this, and I couldn't get how to manage as data was hard to visualize.

I, after, chose some cars characteristics to try to cluster them and fin some vehicle types thanks to the data. But after many tries, I also couldn't get a result interesting enough to do a good clustering.

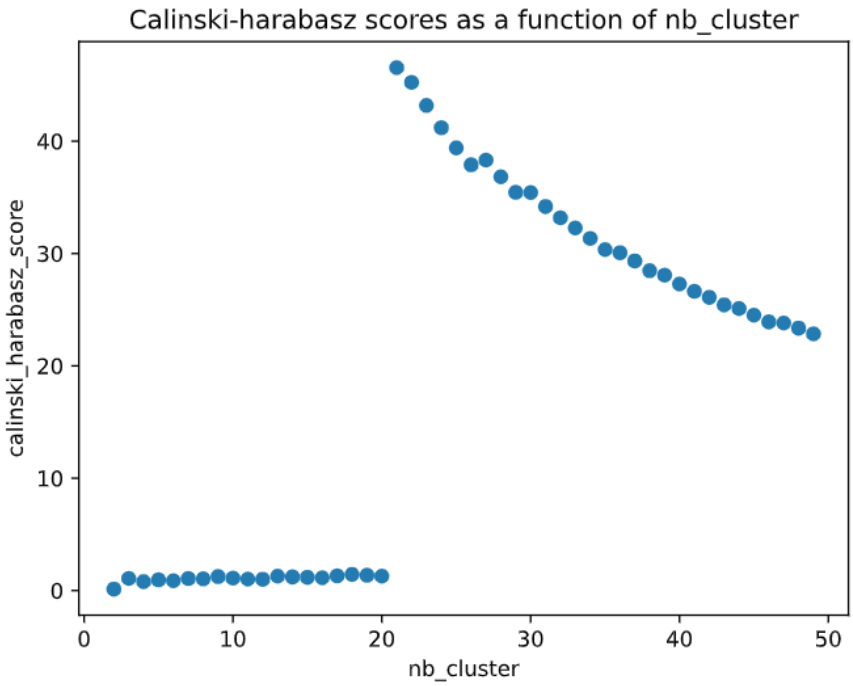


I always got 3 types of vehicles instead of around 10 I should have found for getting a result looking like vehicle types. I so chosen to try dataset with much more works on it to be able to be helped by already existing projects.

On another car dataset I had also redundancy issues and was not sure about best way to handle it so preferred to find another way. After some time, I think now it could have been a really interesting one if I just chose one of these redundant characteristics and try PCA on dataset:



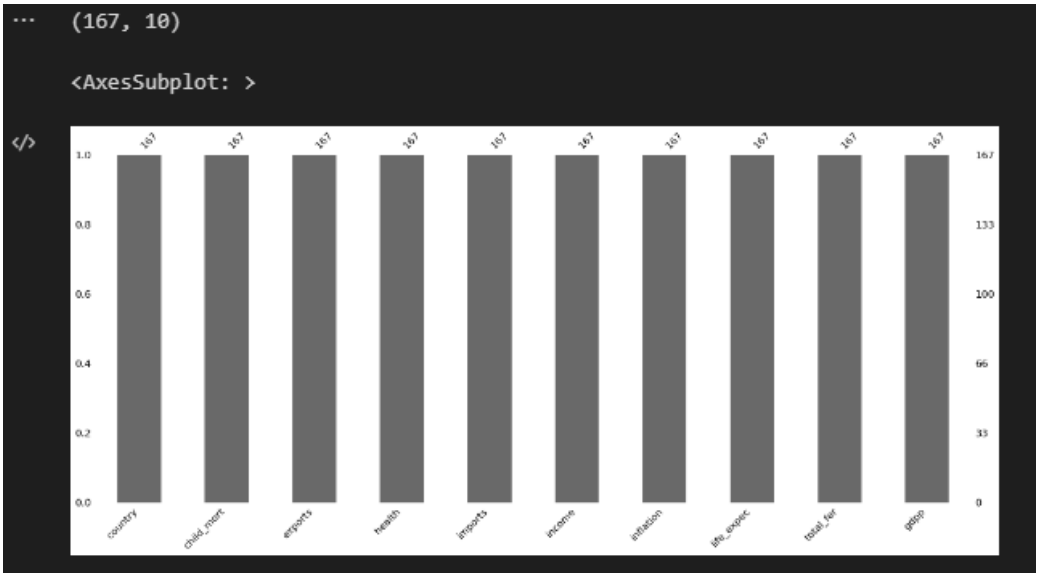
Correlation matrix



I also tried some on Spotify dataset but couldn't find an interesting point of view of clustering so quickly chose to fin another very simple dataset.

FINAL DATASET EXPERIMENTS:

I first started to understand data I got. So I printed the shape and the amount of data in each column to check of utility to remove lines with missing data.



We can here observe that these parameters are giving us some information about important values to detect if a country is well developed or may need help to develop properly and catch biggest ones.

I then printed some data to see the shape and check if there was a magnitude difference between some parameters and if some parameters were categorical. I for example printed means and std for parameters to detect this.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

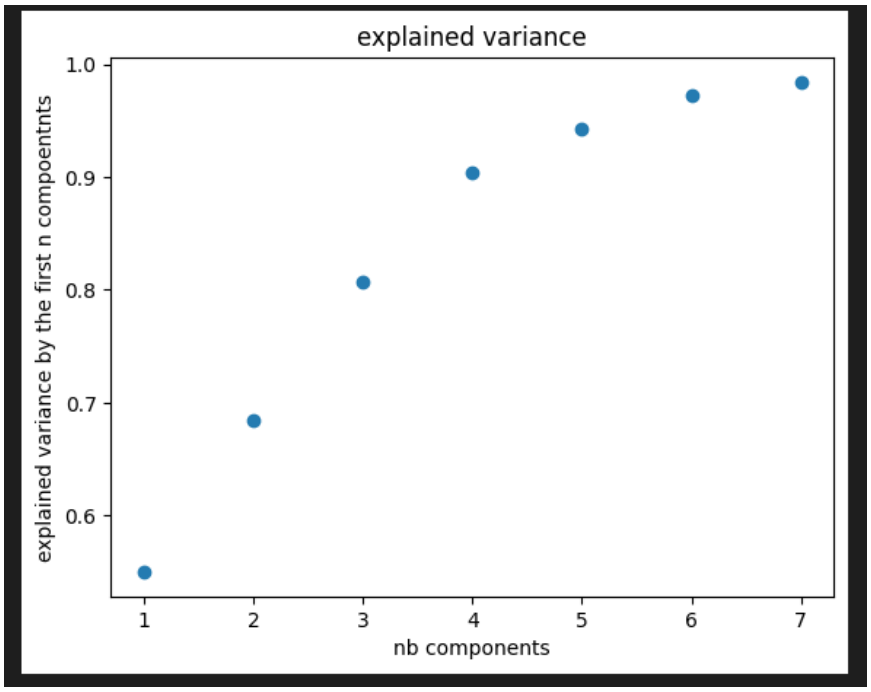
I then checked for correlation matrix to spot some redundancy in parameters.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1.000000	-0.318093	-0.200402	-0.127211	-0.524315	0.288276	-0.886676	0.848478	-0.483032
exports	-0.318093	1.000000	-0.114408	0.737381	0.516784	-0.107294	0.316313	-0.320011	0.418725
health	-0.200402	-0.114408	1.000000	0.095717	0.129579	-0.255376	0.210692	-0.196674	0.345966
imports	-0.127211	0.737381	0.095717	1.000000	0.122406	-0.246994	0.054391	-0.159048	0.115498
income	-0.524315	0.516784	0.129579	0.122406	1.000000	-0.147756	0.611962	-0.501840	0.895571
inflation	0.288276	-0.107294	-0.255376	-0.246994	-0.147756	1.000000	-0.239705	0.316921	-0.221631
life_expec	-0.886676	0.316313	0.210692	0.054391	0.611962	-0.239705	1.000000	-0.760875	0.600089
total_fer	0.848478	-0.320011	-0.196674	-0.159048	-0.501840	0.316921	-0.760875	1.000000	-0.454910
gdpp	-0.483032	0.418725	0.345966	0.115498	0.895571	-0.221631	0.600089	-0.454910	1.000000

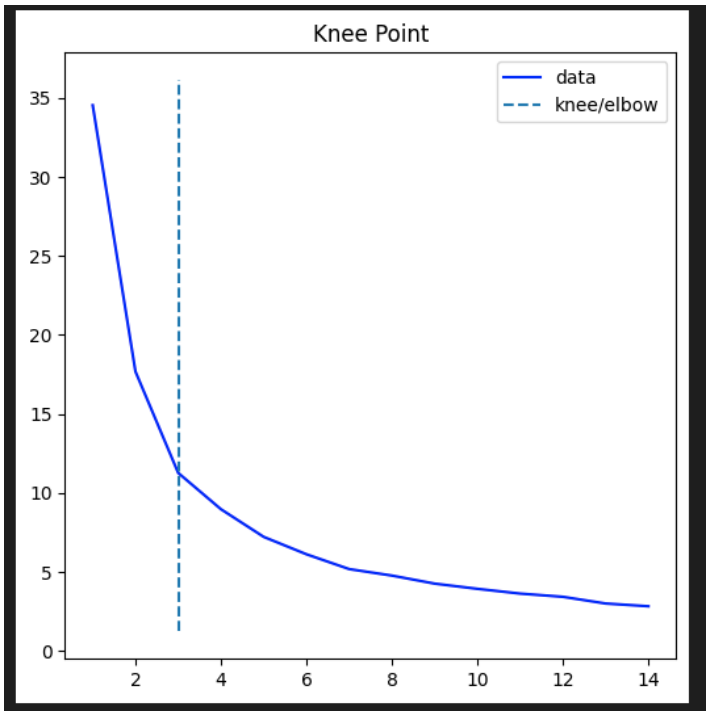
Thanks to correlation matrix we can detect that lot of parameters are relatively correlated (export – import – income…)

This help us to figure out that PCA should be useful to create a lower parametric dataset based on this one and maybe be able to plot this data.

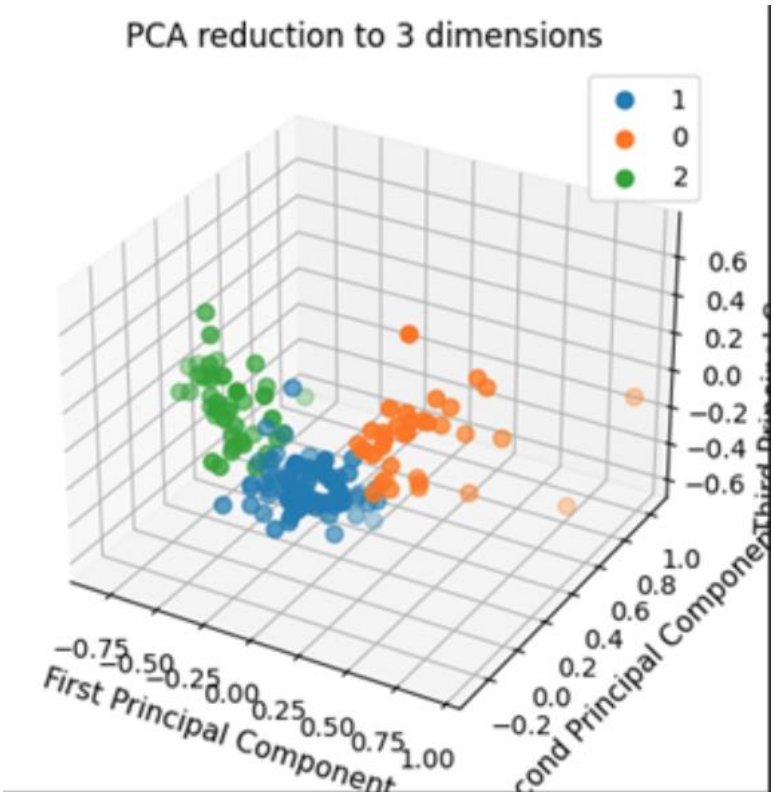
So after dropping country name as it won't be useful for our analysis, I performed PCA with explained variance calculus to get best reduction possible to keep more than 83% of variance in dataset:



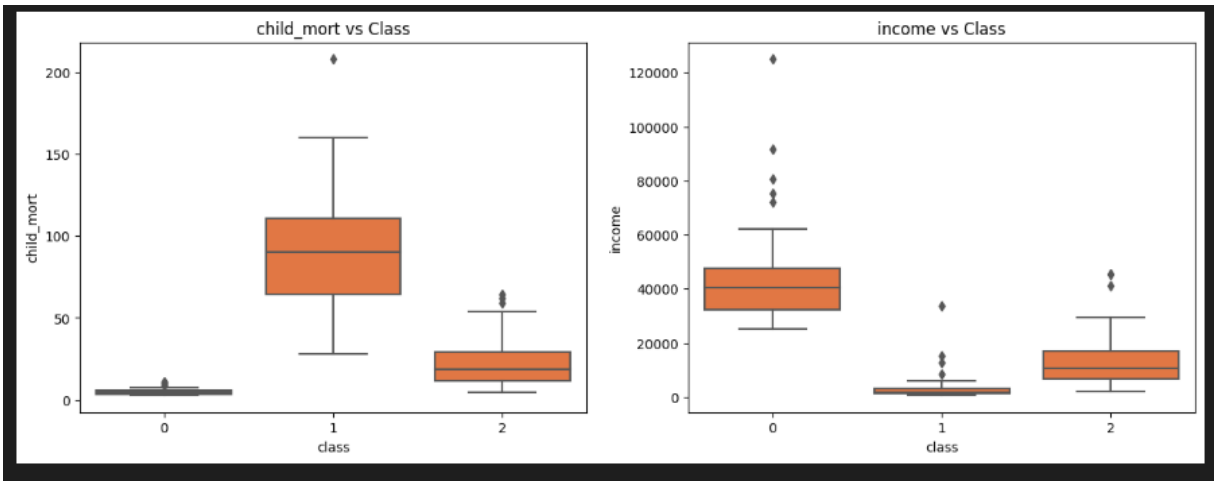
Based on this data, I performed KMean with Kneedle algorithm to get the most interesting number of clusters:



I finally performed KMeans algorithm with 3 clusters and plot it in 3 dimensions to get to this result:



I then transferred these labels obtained thanks to clustering to original dataset and plot some information about values in child mortality and income to detect which cluster is more sensible to need help or not.



I finally labelled it and saved it in result file.

	country	child_mort	exports	health	imports	income	\
0	Afghanistan	90.2	10.0	7.58	44.9	1610	
1	Albania	16.6	28.0	6.55	48.6	9930	
2	Algeria	27.3	38.4	4.17	31.4	12900	
3	Angola	119.0	62.3	2.85	42.9	5900	
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	
..	...	...	...	...	...	...	
162	Vanuatu	29.2	46.6	5.25	52.7	2950	
163	Venezuela	17.1	28.5	4.91	17.6	16500	
164	Vietnam	23.3	72.0	6.84	80.2	4490	
165	Yemen	56.3	30.0	5.18	34.4	4480	
166	Zambia	83.1	37.0	5.89	30.9	3280	

	inflation	life_expec	total_fer	gdpp	class
0	9.44	56.2	5.82	553	need help
1	4.49	76.3	1.65	4090	may need help
2	16.10	76.5	2.89	4460	may need help
3	22.40	60.1	6.16	3530	need help
4	1.44	76.8	2.13	12200	may need help
..	...	...	...	...	...
162	2.62	63.0	3.50	2970	may need help
163	45.90	75.4	2.47	13500	may need help
164	12.10	73.1	1.95	1310	may need help
165	23.60	67.5	4.67	1310	need help
166	14.00	52.0	5.40	1460	need help