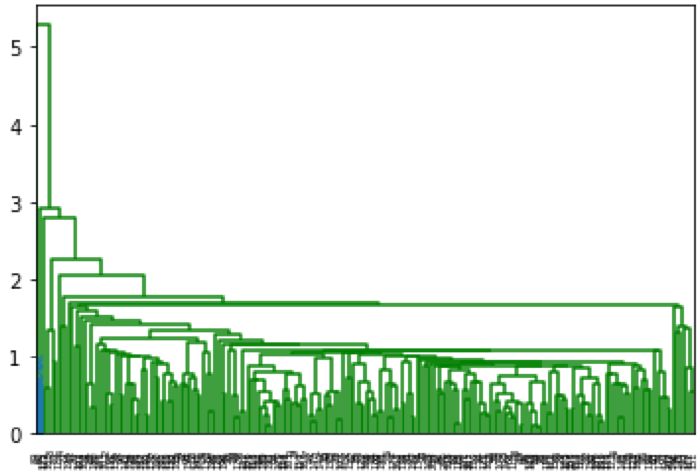
Task 2 for Machine Learning, by Eduardo Sánchez López and José Alejandro Libreros Montaño

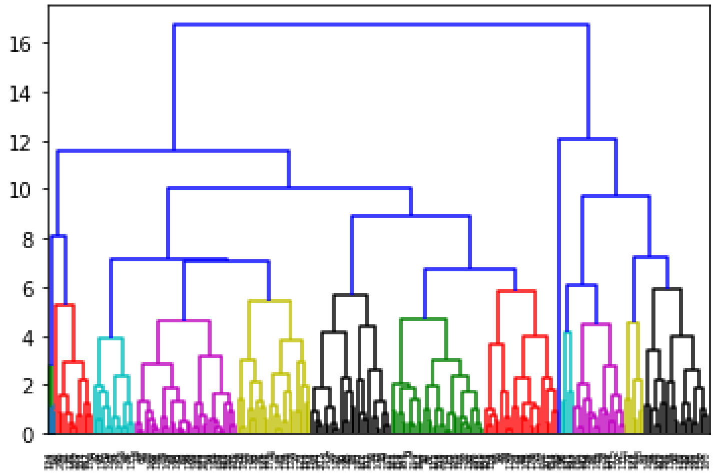
Clustering with the elements.

First we tried the three types of linkage studied in class:

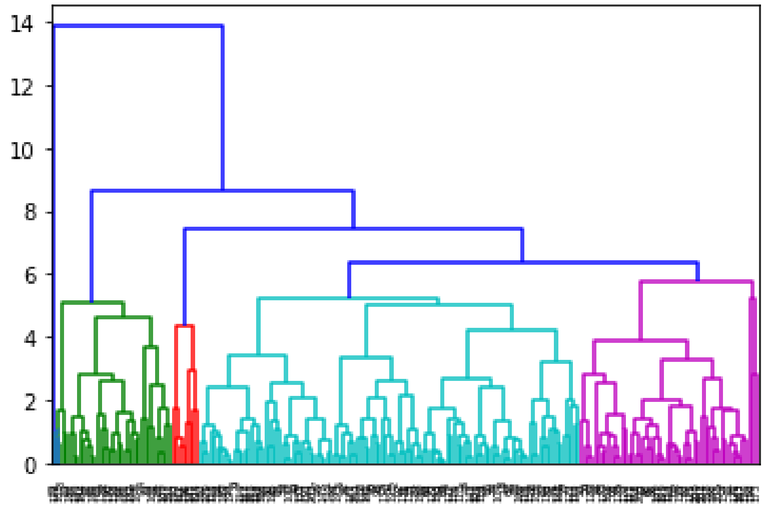
<u>Single:</u>



<u>Complete:</u>



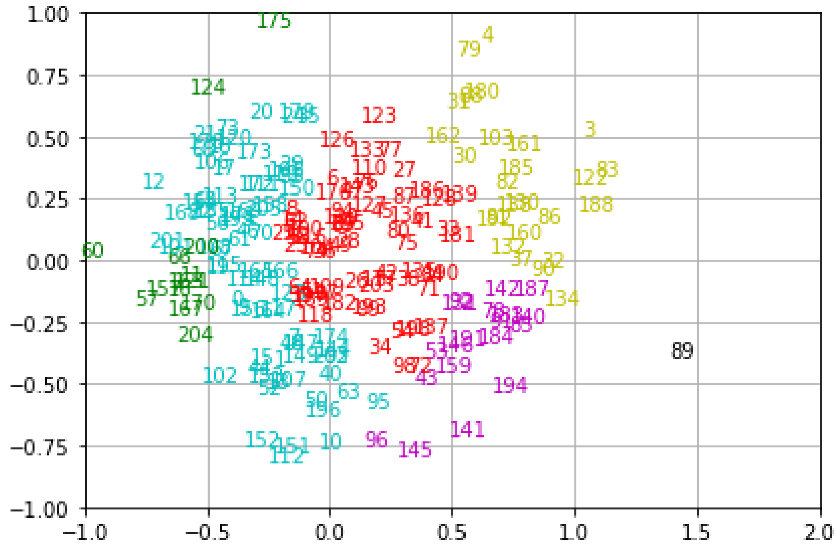
Average:



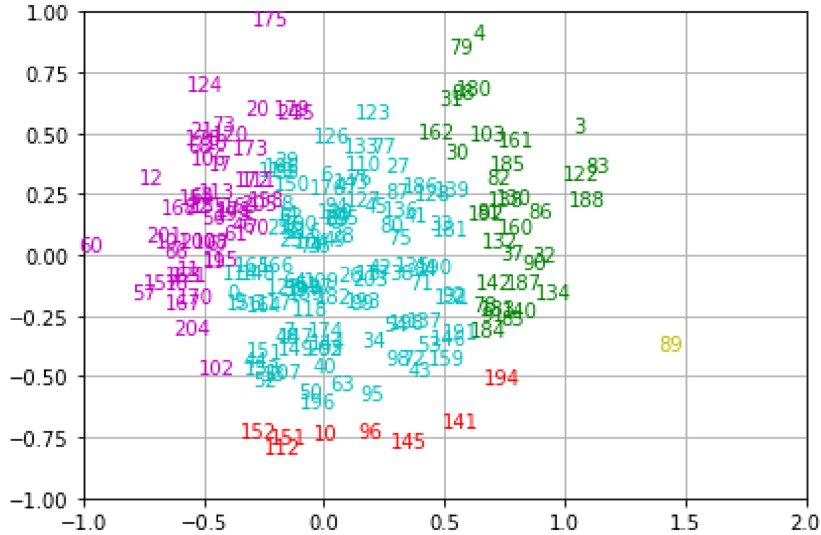
The complete and average linkage clusters are quite similar, so we decided to check the PCA and decide from that.

For the complete linkage we used 9 as the cut, and for the average linkage we used 6.

Complete:



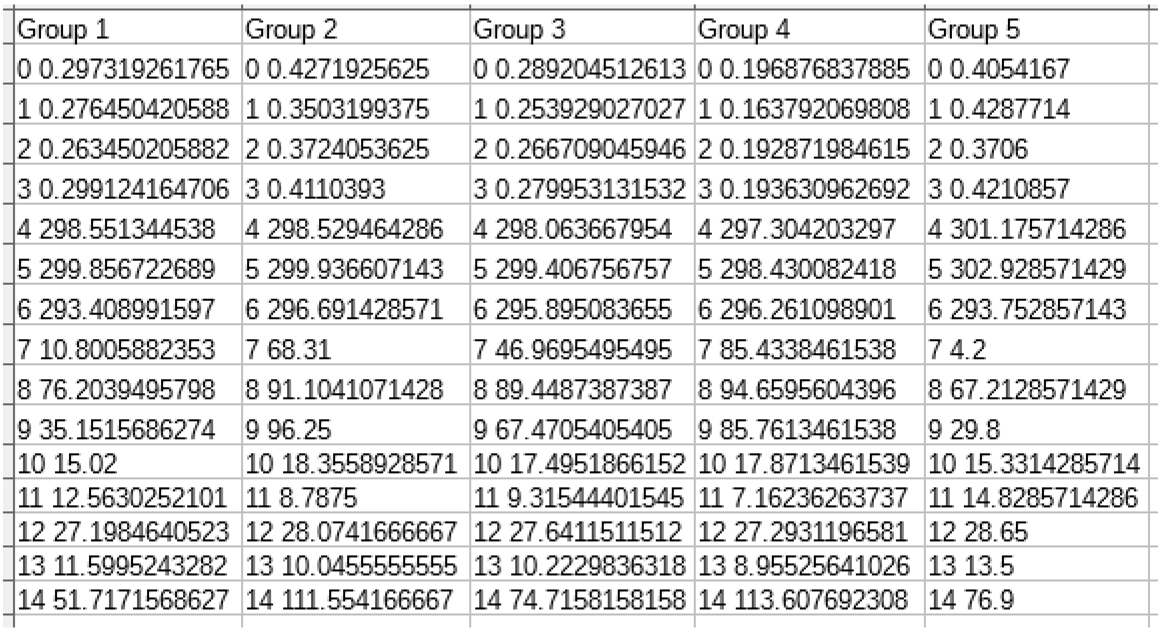
Average:



At the end we choose to use the average cluster, since it’s more compact and the outliers are way more defined. (89, and maybe 60 and 175)

Colors associated with groups:
Green → 1
Red → 2
Cian → 3
Purple → 4
Yellow → 5

Means of the groups:



Group 1	Group 2	Group 3	Group 4	Group 5
0 0.297319261765	0 0.4271925625	0 0.289204512613	0 0.196876837885	0 0.4054167
1 0.276450420588	1 0.3503199375	1 0.253929027027	1 0.163792069808	1 0.4287714
2 0.263450205882	2 0.3724053625	2 0.266709045946	2 0.192871984615	2 0.3706
3 0.299124164706	3 0.4110393	3 0.279953131532	3 0.193630962692	3 0.4210857
4 298.551344538	4 298.529464286	4 298.063667954	4 297.304203297	4 301.175714286
5 299.856722689	5 299.936607143	5 299.406756757	5 298.430082418	5 302.928571429
6 293.408991597	6 296.691428571	6 295.895083655	6 296.261098901	6 293.752857143
7 10.8005882353	7 68.31	7 46.9695495495	7 85.4338461538	7 4.2
8 76.2039495798	8 91.1041071428	8 89.4487387387	8 94.6595604396	8 67.2128571429
9 35.1515686274	9 96.25	9 67.4705405405	9 85.7613461538	9 29.8
10 15.02	10 18.3558928571	10 17.4951866152	10 17.8713461539	10 15.3314285714
11 12.5630252101	11 8.7875	11 9.31544401545	11 7.16236263737	11 14.8285714286
12 27.1984640523	12 28.0741666667	12 27.6411511512	12 27.2931196581	12 28.65
13 11.5995243282	13 10.0455555555	13 10.2229836318	13 8.95525641026	13 13.5
14 51.7171568627	14 111.554166667	14 74.7158158158	14 113.607692308	14 76.9

Meaning of each characteristic:
0: ndvi_ne: Pixel northeast of city centroid
1: ndvi_nw: Pixel nortwest of city centroid
2: ndvi_se: Pixel southeast of city centroid
3: ndvi_sw: Pixel southwest of city centroid

Data from NOAA’S NCEP Climate Forecast System Reanalysis.
4: reanalysis_air_temp_k: Mean air temperature.
5: reanalysis_avg_temp_k: Average air temperature.
6: reanalysis_dew_point_temp_k: Mean dew point temperature.
7: reanalysis_precip_amt_kg_per_m2: Total precipitation.
8: reanalysis_relative_humidity_percent: Mean relative humidity.
9: reanalysis_sat_precip_amt_mm: Total precipitation.
10: reanalysis_specific_humidity_g_per_kg: Mean specific humidity.
11: reanalysis_tdtr_k: Diurnal temperature range.

Data from NOAA’S GHCN daily climate data weather station measurements
12: station_avg_temp_c: Average temperature.
13: station_diur_temp_rng_c: Diurnal temperature range.
14: station_precip_mm: Total precipitation.

Labels:

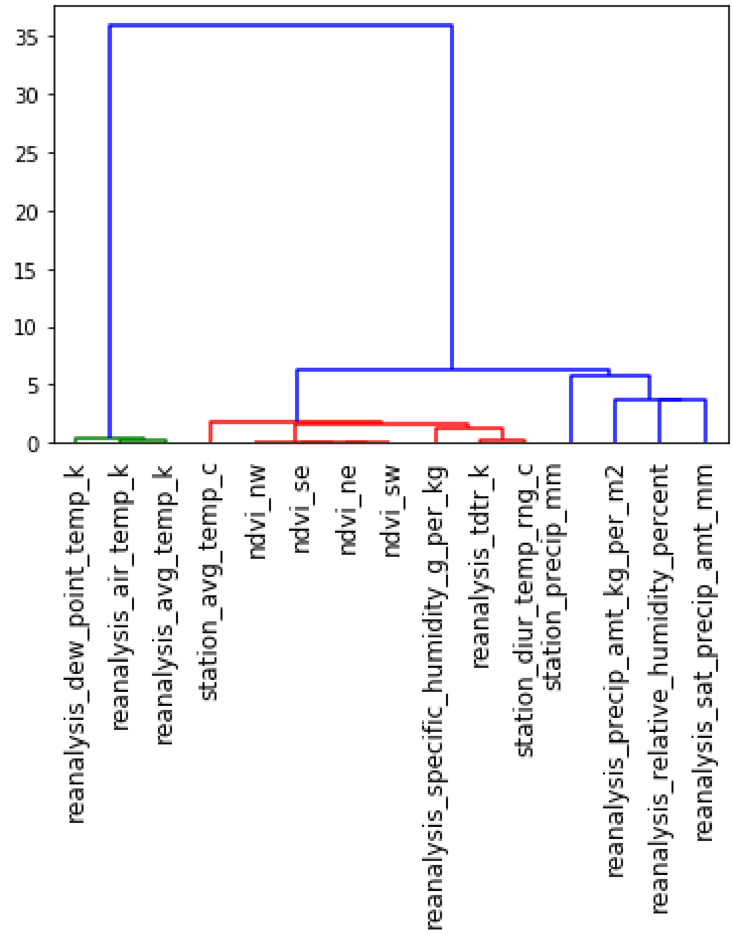
We label them in terms of the precipitation/humidity, since the mean temperature doesn't change too much between the groups.

Group 1: Low precipitations/humidity
Group 2: High precipitations/humidity
Group 3: Medium precipitations/humidity
Group 4: Maximum precipation/humidity.

The group 5 seems like an outlier that should belong in the group 2, specially since a value seems abnormally low (4.2 in row 7, precipitations in that week) which can indicate a really strange week or an error in the measure of the data. Either way, for the k-means algorithm that week (and maybe the week 60) will be consider outliers and deleted.

The group 2 seems very similar to the group 3, but the difference is that in group 2 there's a conflict in the data of the humidity and precipitations between the two sources (Reanalysis data and the daily climate data) That data should be similar, but it's not, so maybe one of the sources committed an error in the measurement.

Clustering with the characteristics.



The data correlated is described as follows:

Green

Mean dew point temperature (°K)
Mean air temperature (°K)

Average air temperature (°K)

Is important to highlight that the dew point temperature is related to the air temperature, that explains the first and second label in that graph.

Red
Average temperature (°C)
ndvi_se – Pixel southeast of city centroid

ndvi_sw – Pixel southwest of city centroid

ndvi_ne – Pixel northeast of city centroid

ndvi_nw – Pixel northwest of city centroid

Mean specific humidity

Diurnal temperature range (°K)

Diurnal temperature range (°C)

We can see the corelation between the temperature and specific humidity.

Cian

Total precipitation
Total precipitation (kg per m2)
Mean relative humidity (percent)
Total precipitation (mm)

From the dataset, the precipitation in that location is related to the relative humidity, which are related with the weather tropical conditions.

Problems encountered and solved:

1. Matrix error.

We had to redone the matrix, because Python had shown the error
```python
ValueError: setting an array element with a sequence.
```
We had copied the matrix `states` into a new `temp` matrix.

This has to be related with the way the .csv are converted from .ods through libreOffice.