Task 1 of Machine Learning, by Eduardo Sánchez López and José Alejandro Libreros Montaño.

1. Preprocessing

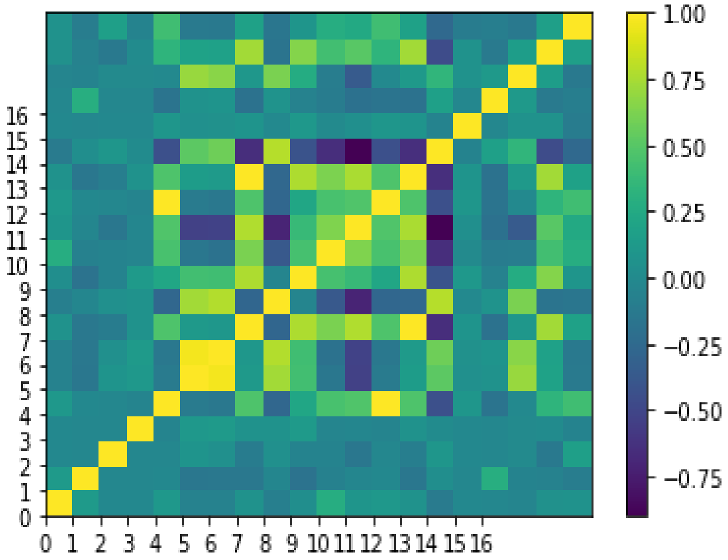
We deleted all the data about identifiers in the dataset (The week and the year).

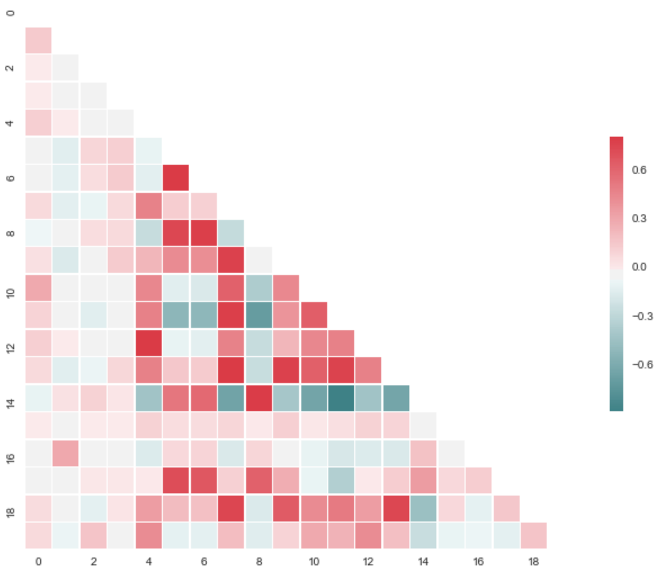
To avoid the corruption of data, in the blank spaces we put the mean of the 3 above and 3 below rows instead of a zero (mirror).

We considered to make a column with the mean of the min and max temperatures in a week, but at the end we decided not to do it yet.

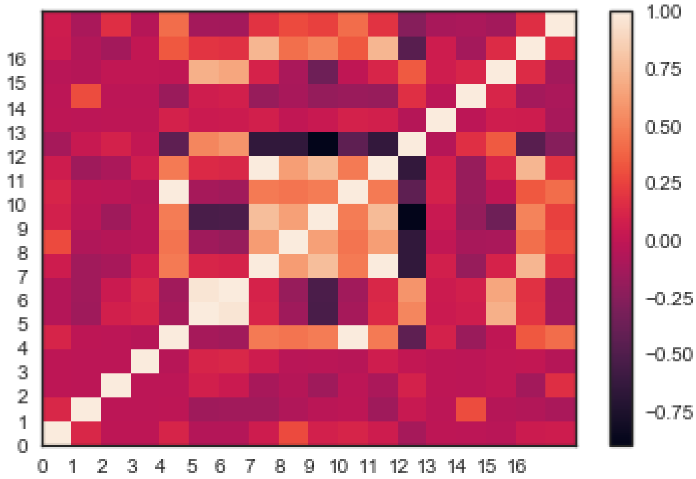
2. Correlation

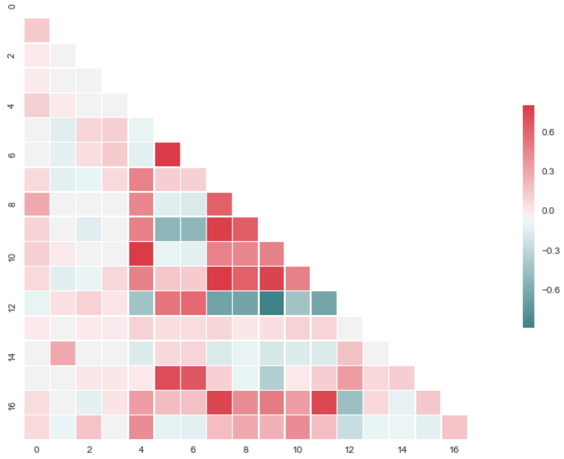
For the study of the correlation and PCA we execute the algorithms, decide if there's something that we should fix, and then execute the algorithm again. We do this procedure until we are done.



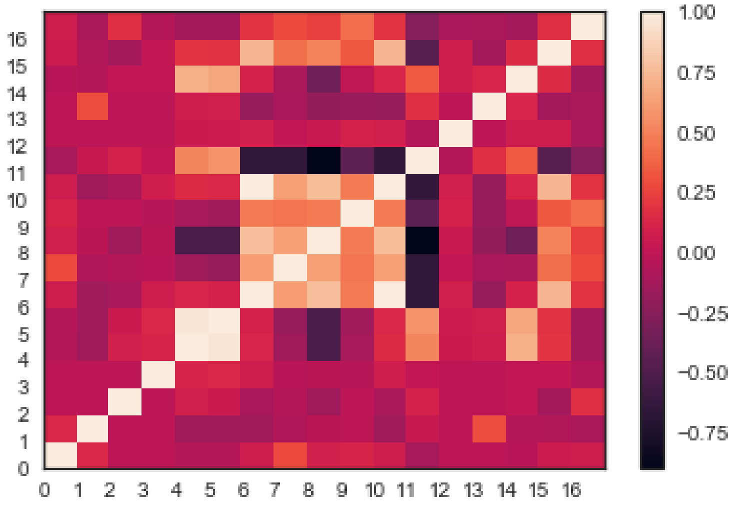


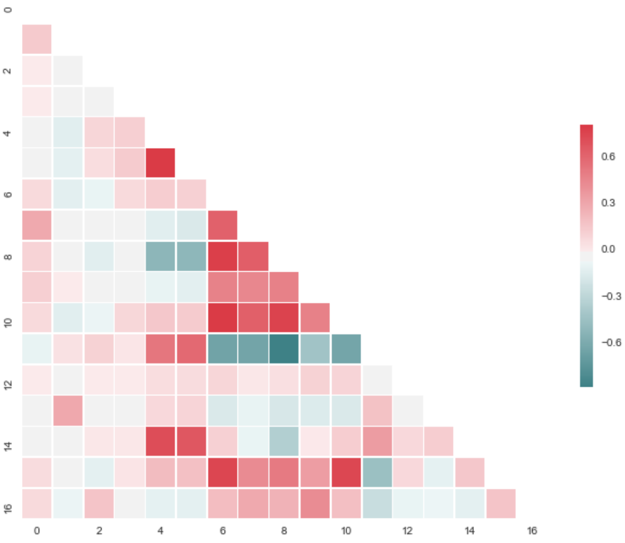
There are some labels that are very correlated:
The average temperature is very related to the maximum and minimum temperature by obvious reasons, because it's the mean between those two values, so we deleted the maximum and minimum temperatures from the dataset since it's repeated information.



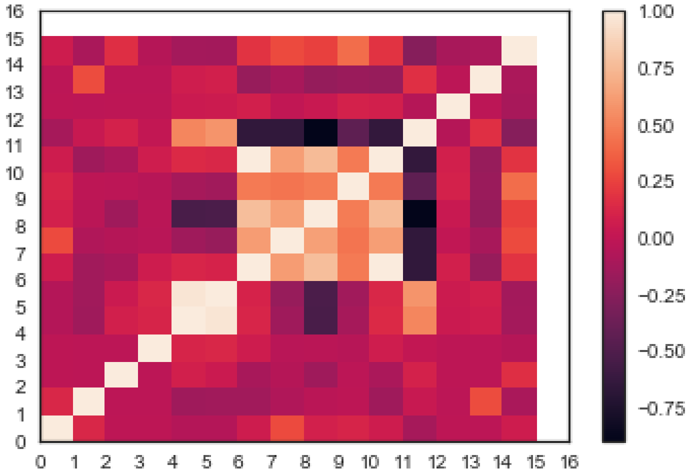


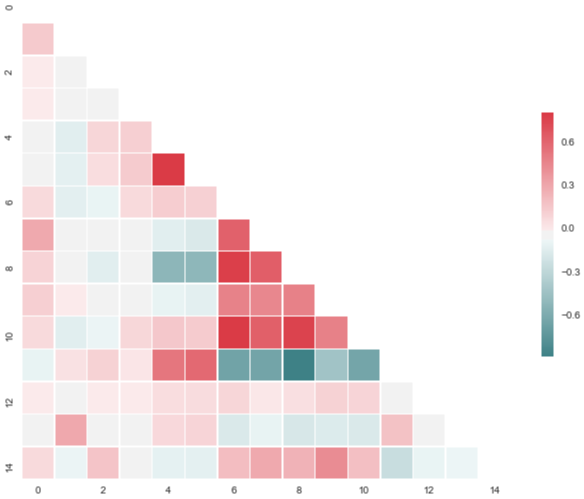
4 and 10 are very correlated as well.
Once we get to the dataset we see that they are almost the same columns. That makes sense because 4 is the total precipitation by a satellite and 10 is the total precipitation reanalysed and measured by a climate forecast system. They are correlated because they are the same data, so we proceed to delete the column 4 because it's the only column with data extracted from the satellite method.





Here we can finally say that we have a clear correlated group, from the columns 6 to 10, which are the columns related with the humidity and precipitation. We can obtain a rule that says that the precipitations and the dew point are related with the humidity. That makes a lot of sense because it is only normal that the humidity goes up once there's more water. There might be a case to delete one or more columns of this group, but for the time being we will keep them as it is.
The biggest negative correlation is between the column 11 (The diurnal temperature range) with the columns 6 through 10 (The columns that talk about humidity, as said before). We concluded that when the difference between the maximum and minimum temperature is high, there are less possibilities that it rains.
The last group that we observed that had a high correlation is between the columns that have data about the temperatures in Kelvin (4) and the columns that have data about the same topic but in celsius (14 and 15). The only difference is the measure, so we considered that they are the same data and proceeded to delete the columns 14 and 15. The reason for deleting this ones and not the ones that are in Kelvin is because we have more data of this problem in Kelvins rather than Celsius.
We didn't delete the column 16 because although is data about precipitations and we already have it from another source, it's not so strongly related so it might have relevant data.



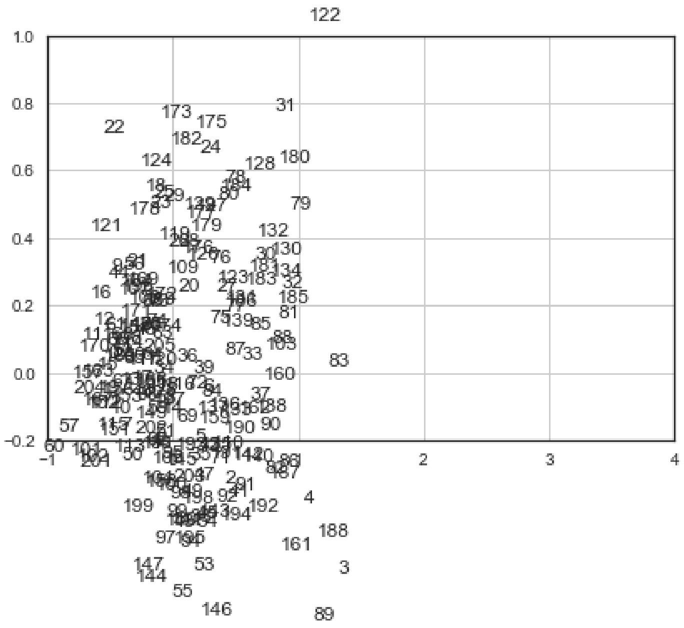


Finally we see a cleaner dataset with the two correlated groups that we talked about previously.

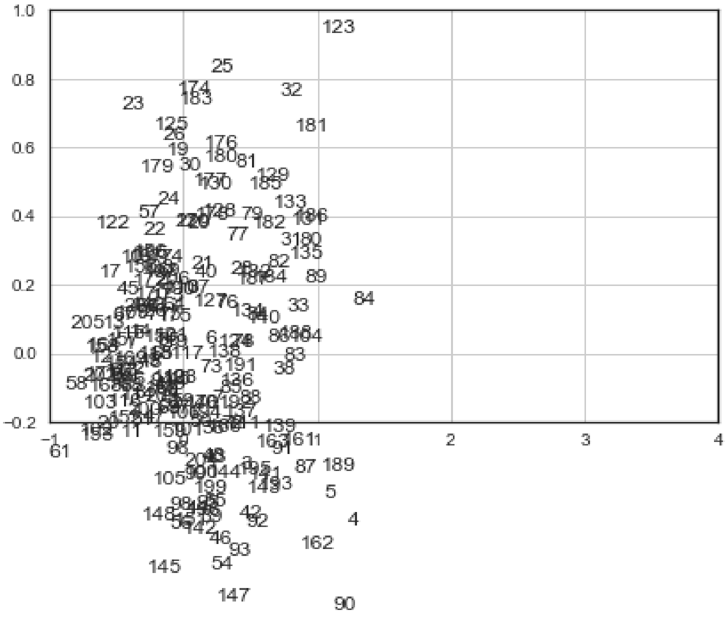
As a final anotation, the first 4 columns (0 through 3) are not correlated at all with the rest of the database. It corresponds with the normalized difference vegetation in the 4 cardinal points, so we can say that the climate does not have a strong relation with the vegetation of the place.

PCA:
We can see that after each iteration in which we delete redundant data the elements are getting more grouped into 1 group, with only a few outliers, specially the week 123. We can say that the climate is very similar through the years and the enviromental behaviour doesn't change much.

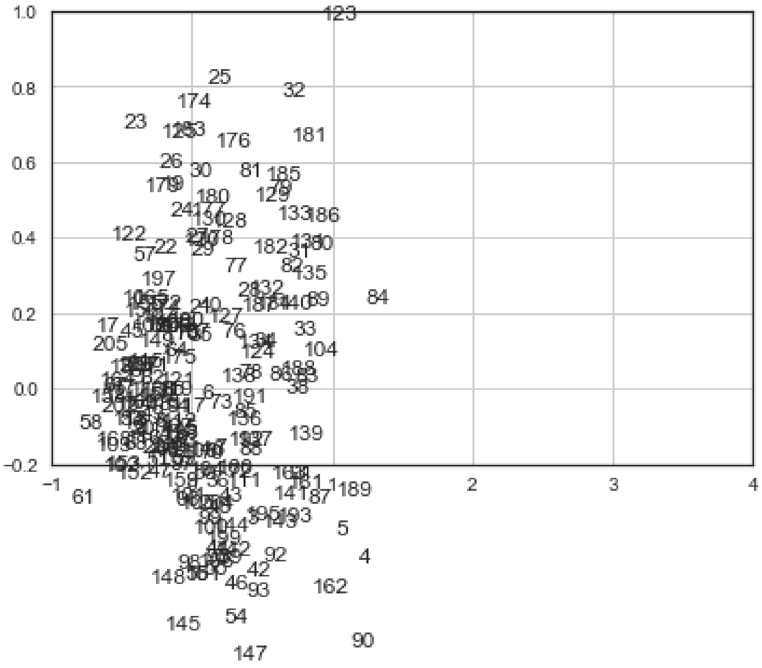
First iteration



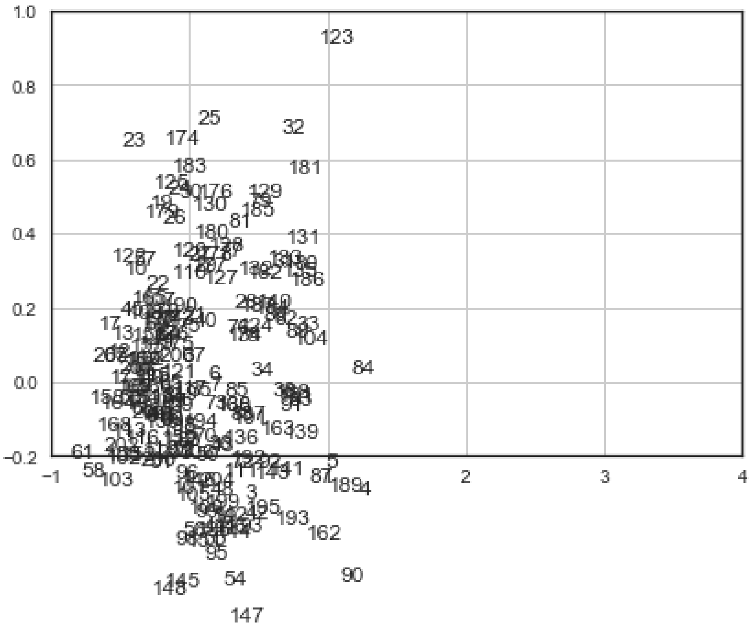
Second iteration



Third iteration



Fourth iteration


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−1
2
3
4
123
25
32
23
174
181
183
176
129
130
185
81
180
131
186
84
104
33
34
6
85
139
163
61
58
103
5
87
189
4
193
195
3
162
95
145
148
54
90
147