Data Science Project: Rainfall in Barcelona, analysis and recommendations

Authors: Josep Manel Marí

Mentor: Kevin Mamaqi

February 2022

Index

1. Introduction

1.2 Objective of this study

The objective of this study is to make an analysis and recommendation on the rainfall in the city of Barcelona based on its historical and present data.

1.3 Working plan

The working plan is as follow:

- Gather the information of public sources.
- Engineering of data for analysis with Pandas.
- Load information into an SQL format.
- Elaboration of tables and graphics and short description of them.
- Descriptive analysis. Long term and medium/short term tendencies.
- Machine Learning models: regression and clustering.
- Conclusions.
- Presentation to the client.

2. Data gathering

2.1 Data gathering methods

The information needed for the study has been downloaded from:

- “Dades meteorològiques de la XEMA”, available in the open data portal of the la Generalitat [1].

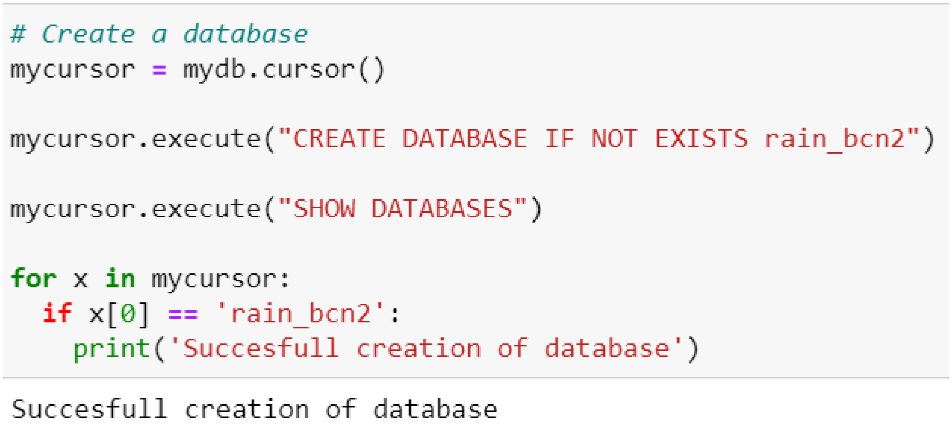
- “Precipitacions acumulades mensuals de la ciutat de Barcelona des de 1786” and “Temperatures mitjanes mensuals de l’aire de la ciutat de Barcelona des de 1780”, available in the open data portal of the Ajuntament de Barcelona [1 and 2].

The data downloaded from XEMA is the meteorological installation X4 located in Ciutat Vella. The years available are from 2009.

2.2 Data processing

The datasets have been loaded to a csv files, which have been engineered with Pandas and it’s libraries.

Table 1. Creation of the database “rain_bcn2”



```
# Create a database
mycursor = mydb.cursor()

mycursor.execute("CREATE DATABASE IF NOT EXISTS rain_bcn2")

mycursor.execute("SHOW DATABASES")

for x in mycursor:
  if x[0] == 'rain_bcn2':
    print('Succesfull creation of database')
```

```
Succesfull creation of database
```

This information has been loaded into a SQL file.

3. Data analysis

3.1 Long term historical rainfall tendencies

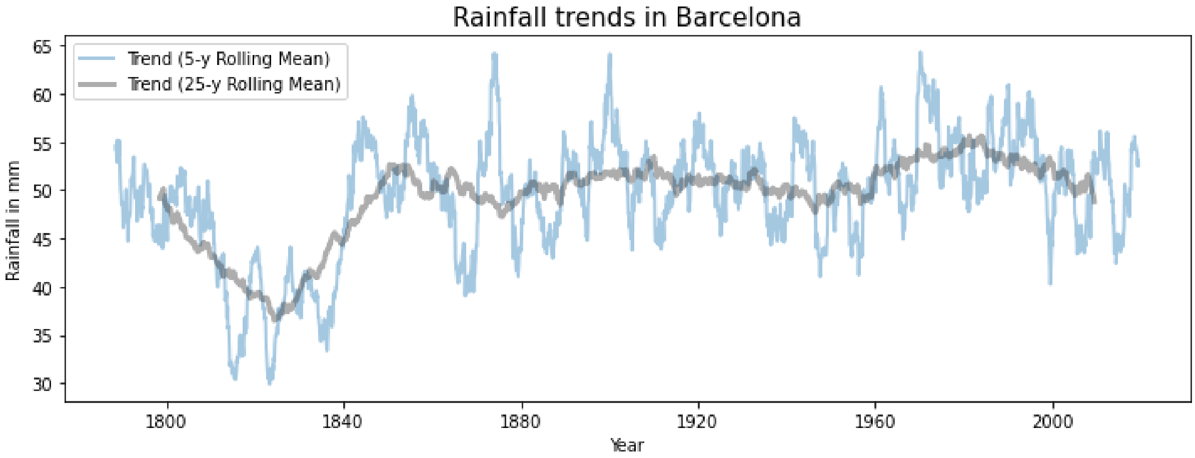


Figure 1

The long term rainfall trend in Barcelona shows a reduction of precipitation but in the last few years has increased again. The lowest point was in the decade of 1830.

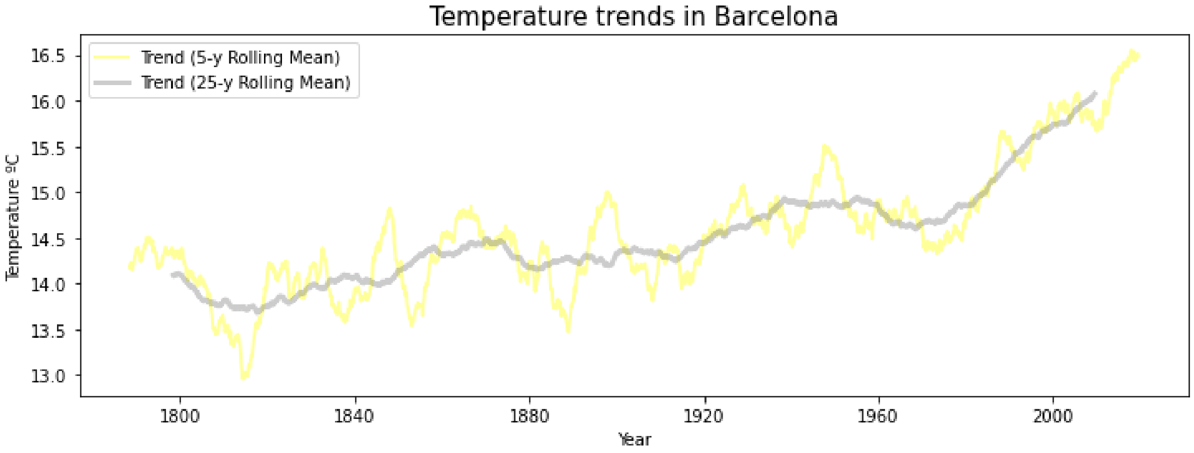


Figure 2.

On the other hand, temperatures have risen constantly and with an exponential rate during the last fifty years.

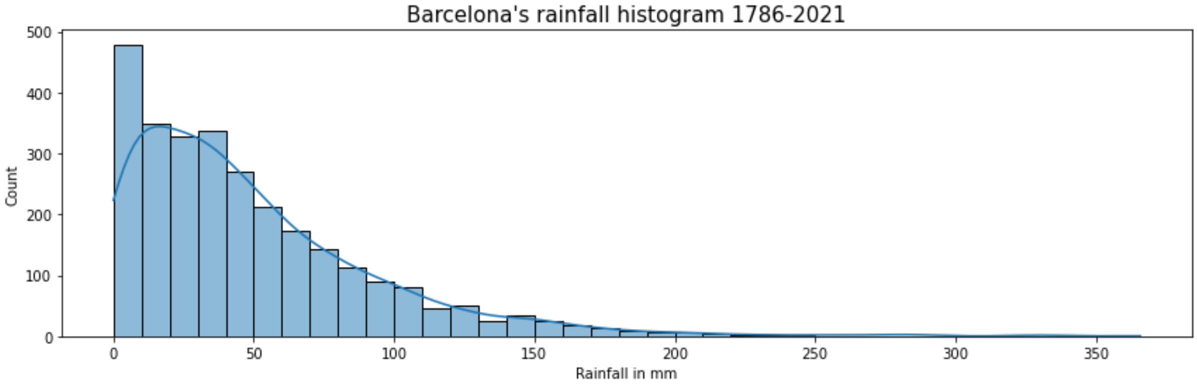


Figure 3.

The distribution of rainfall is asymmetric because Barcelona being a Mediterranean city has periods of drought with short episodes of intense precipitations.

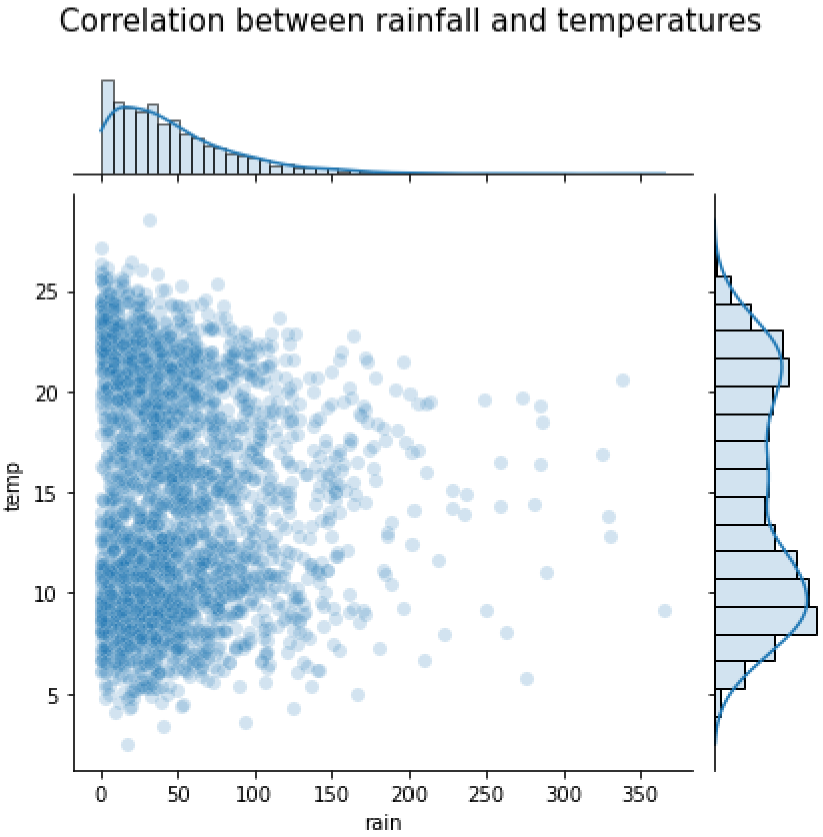


Figure 4.

There is not a correlation between temperature and rainfall.

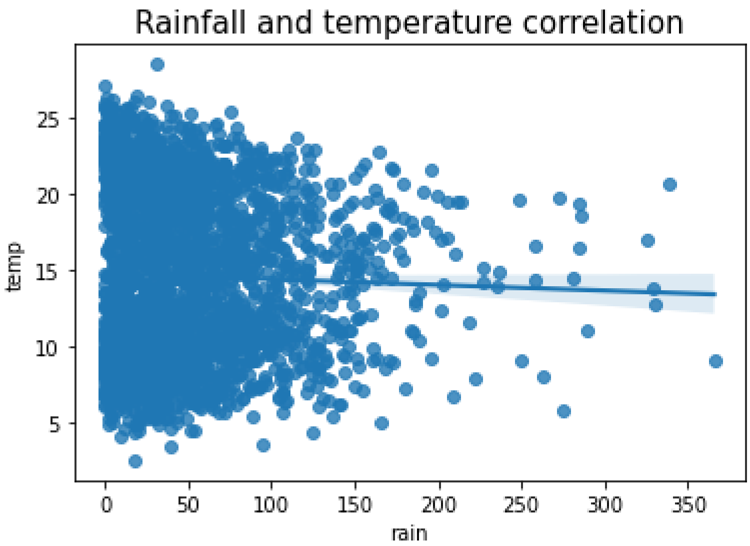


Figure 4.

This graphic shows clearly that rain is independent of temperature, thus, they are not correlated.

The pluviometry of Barcelona is typical of a Mediterranean city.

3.2 Medium/short term rainfall tendencies

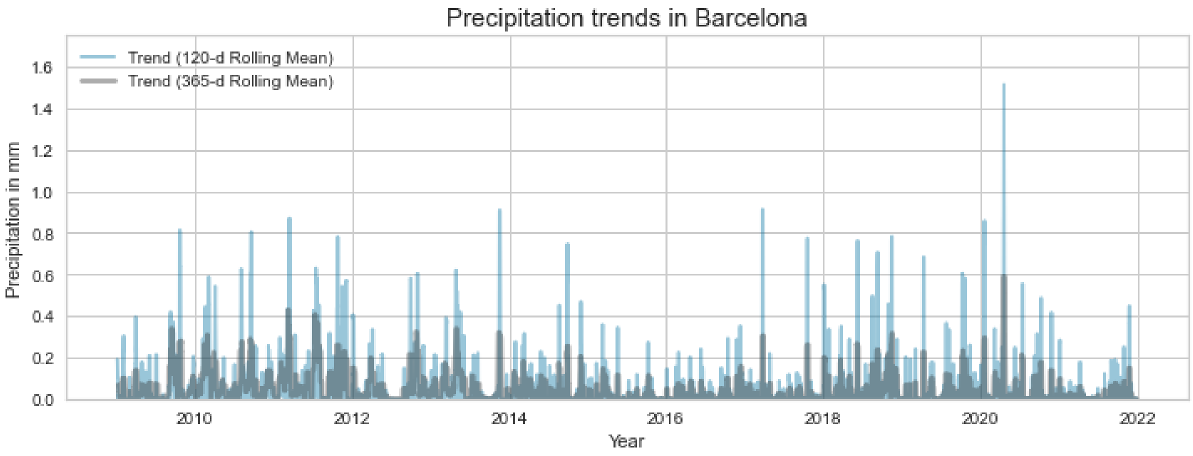


Figure 6.

From 2019 to 2020 the precipitation in Barcelona increased, compared to the previous period. From 2021 till today, the city has experienced a disminution of rainfall.

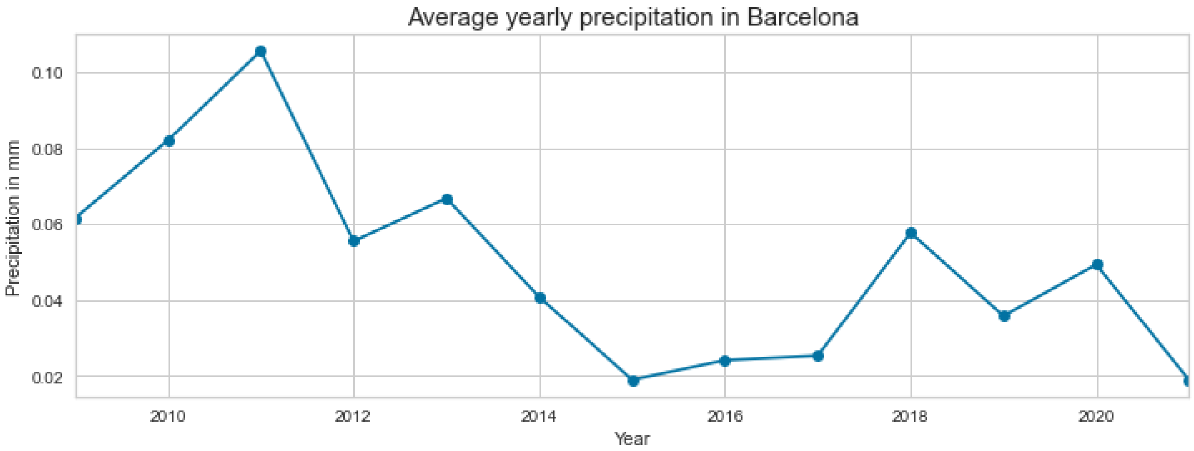


Figure 7.

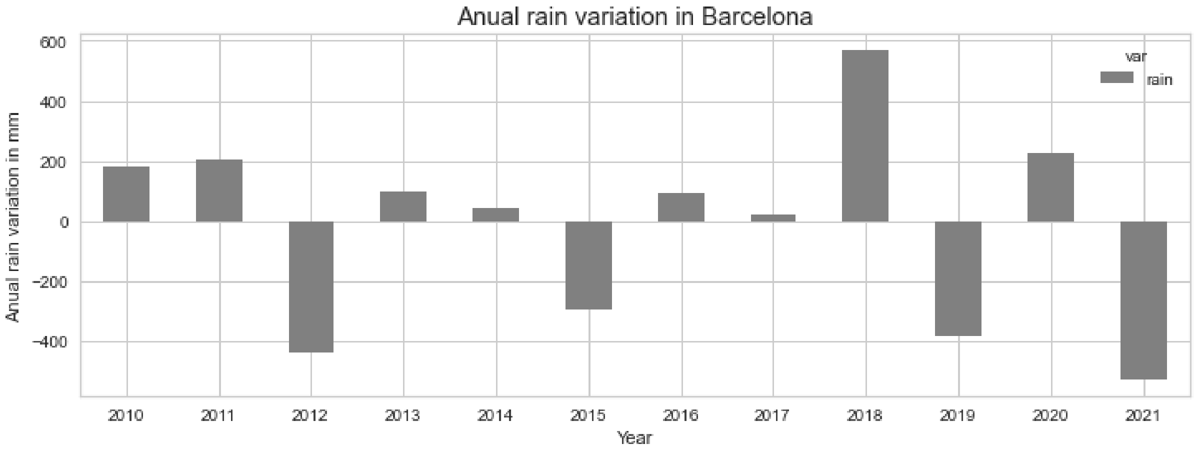


Figure 8.

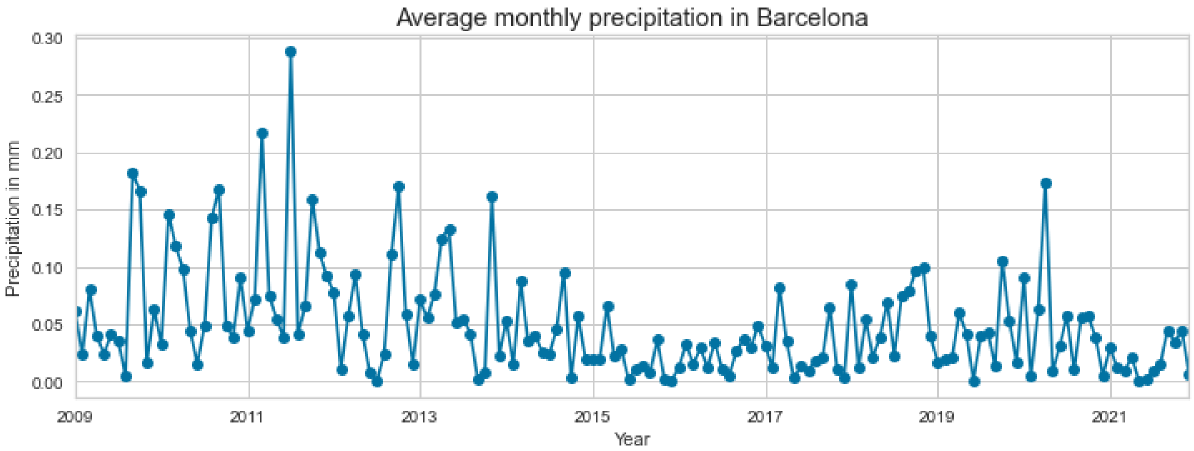


Figure 9.

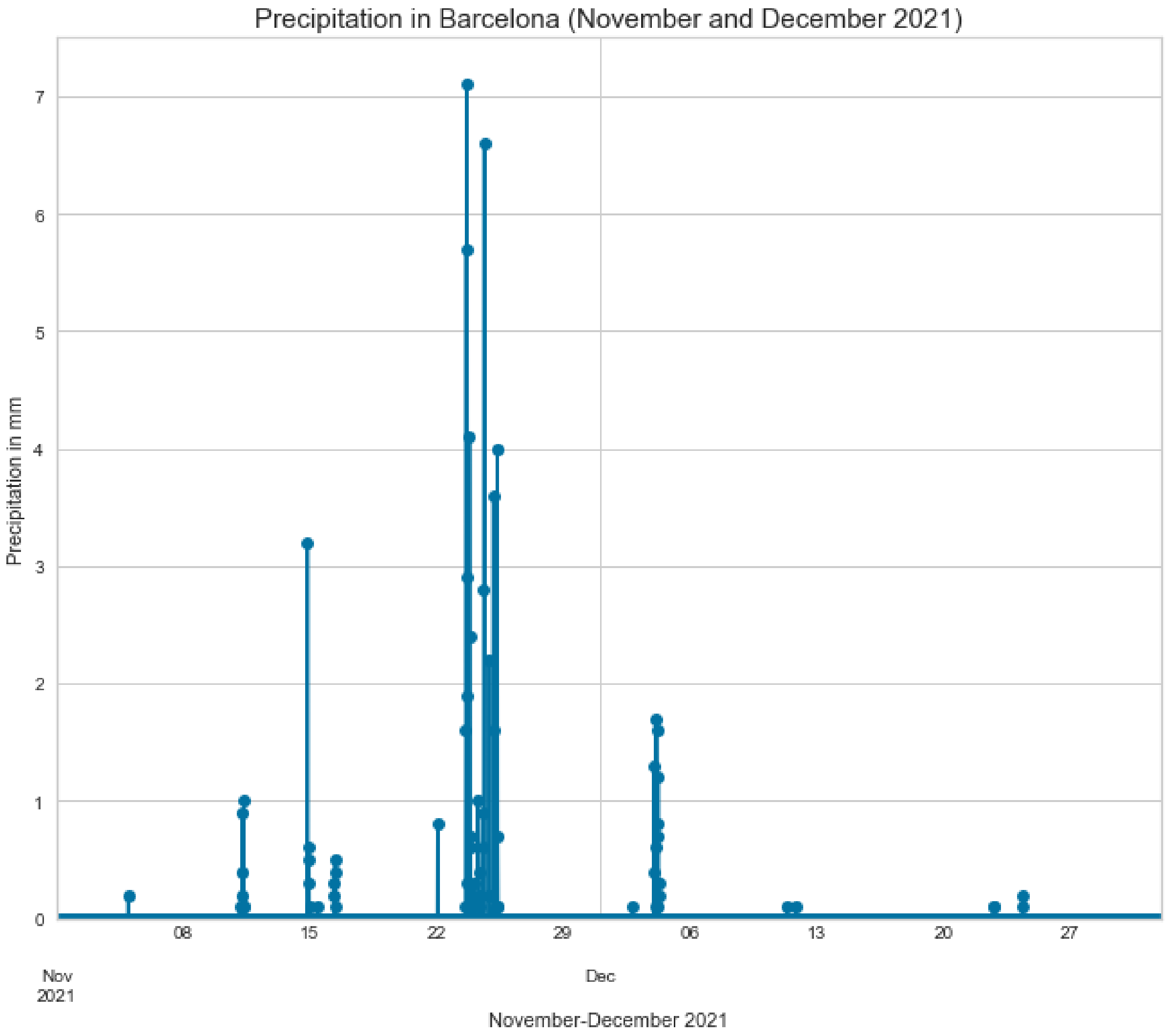


Figure 10.

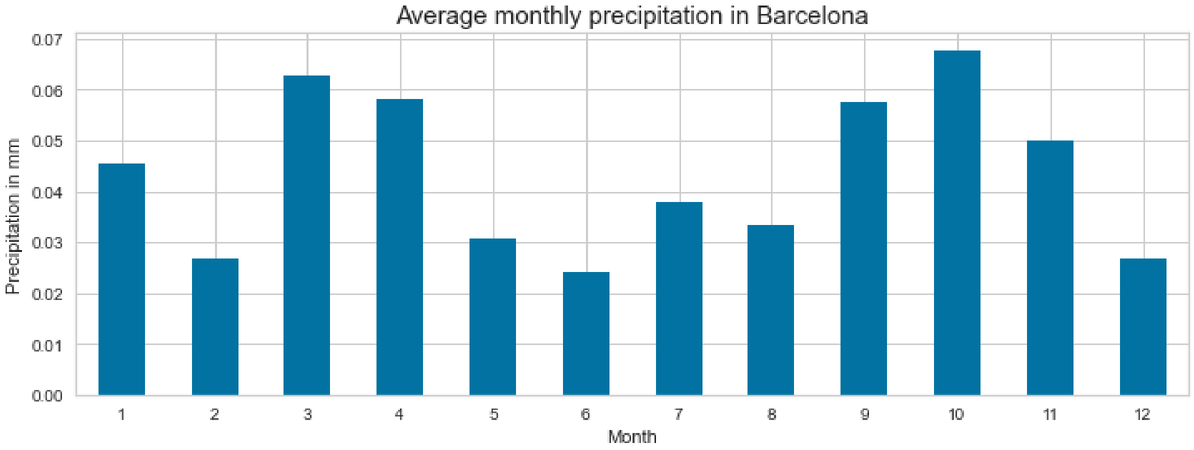


Figure 11.

In November and March rainfall is higher, and in June and February, the lowest.

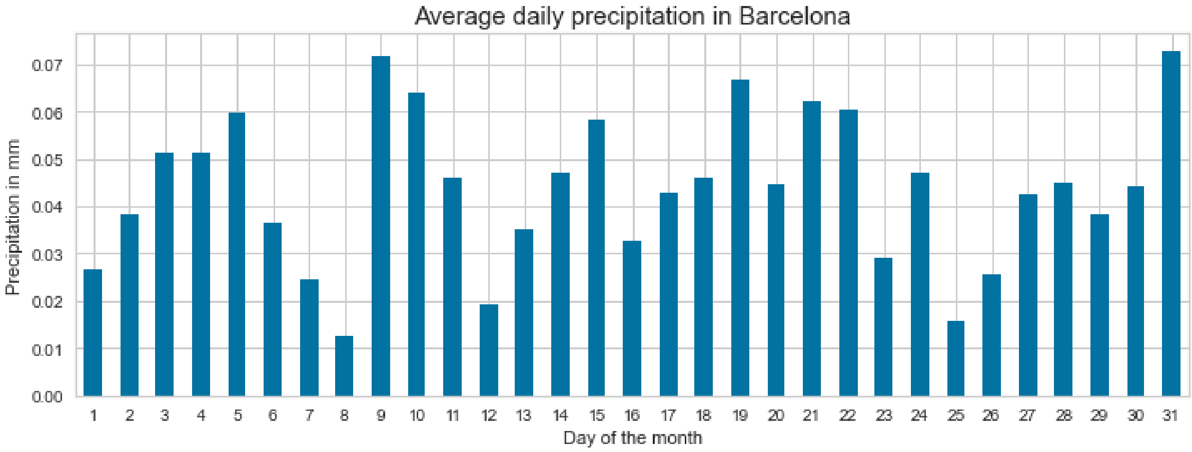


Figure 12.

As a curiosity, it rained more the days of the month 31 and 9.

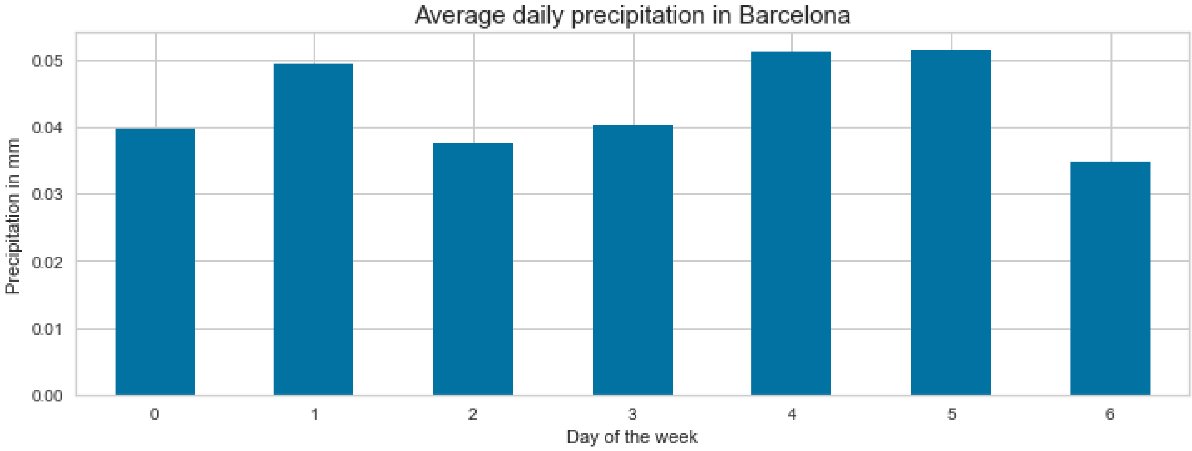


Figure 13.

Saturday and Tuesday are the days with more rain. Sunday and Wednesday, the less.

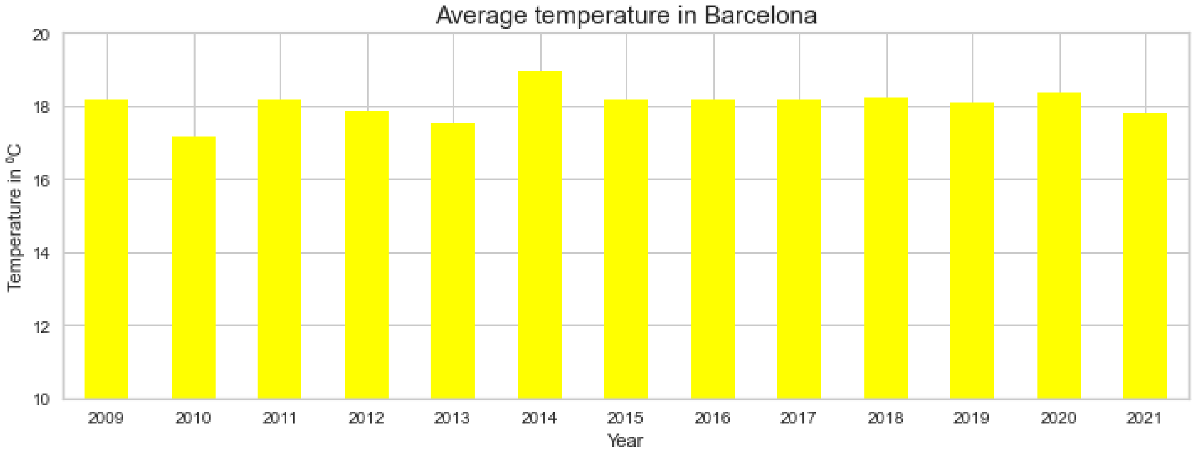


Figure 14.

In 2021 the temperature in Barcelona was the lowest since 2013.

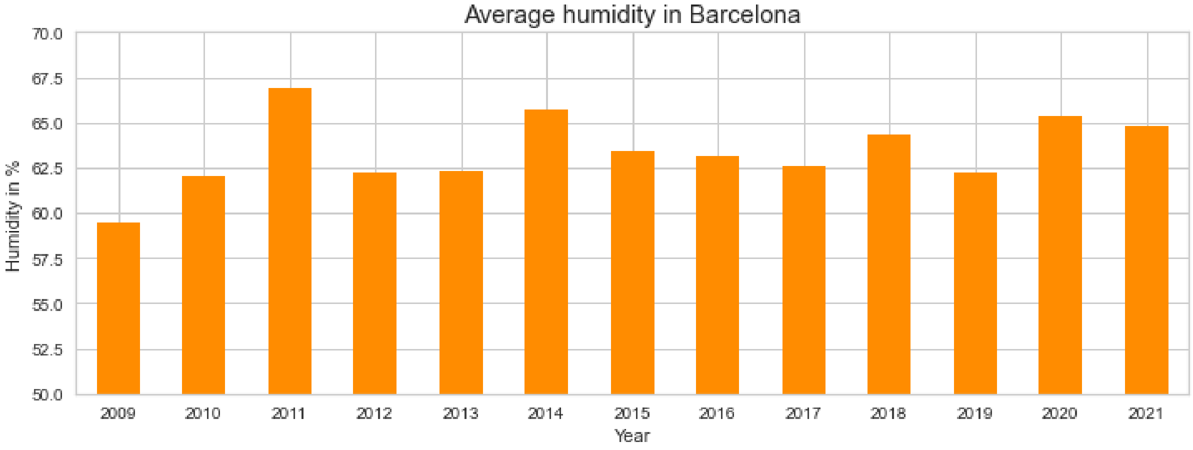


Figure 15.

2011 and 2014 were the most humids years.

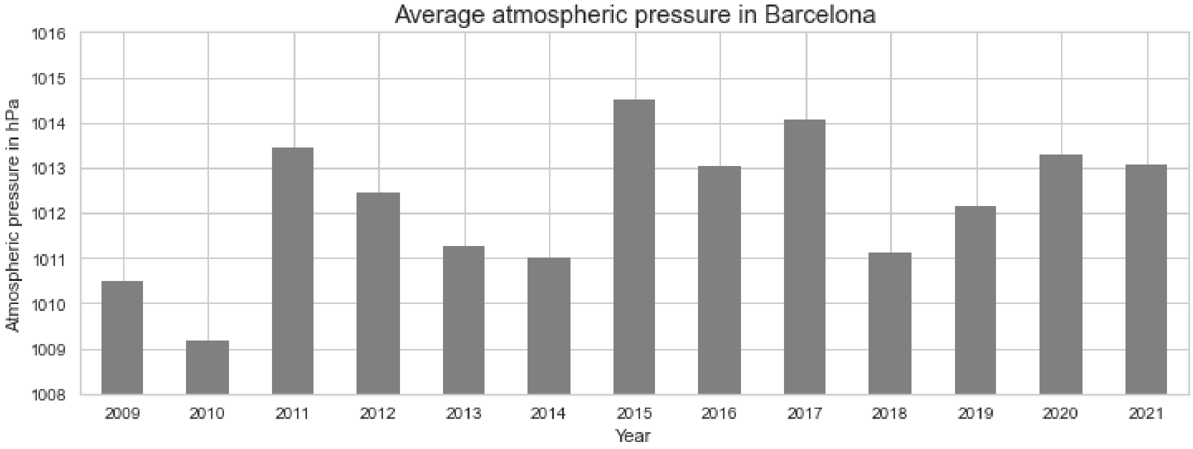


Figure 16.

We can also observe the slight changes of atmospheric pressure.

As mentioned before there is no correlation between rainfall, temperature or atmospheric pressure, but a slight correlation between rain and humidity.

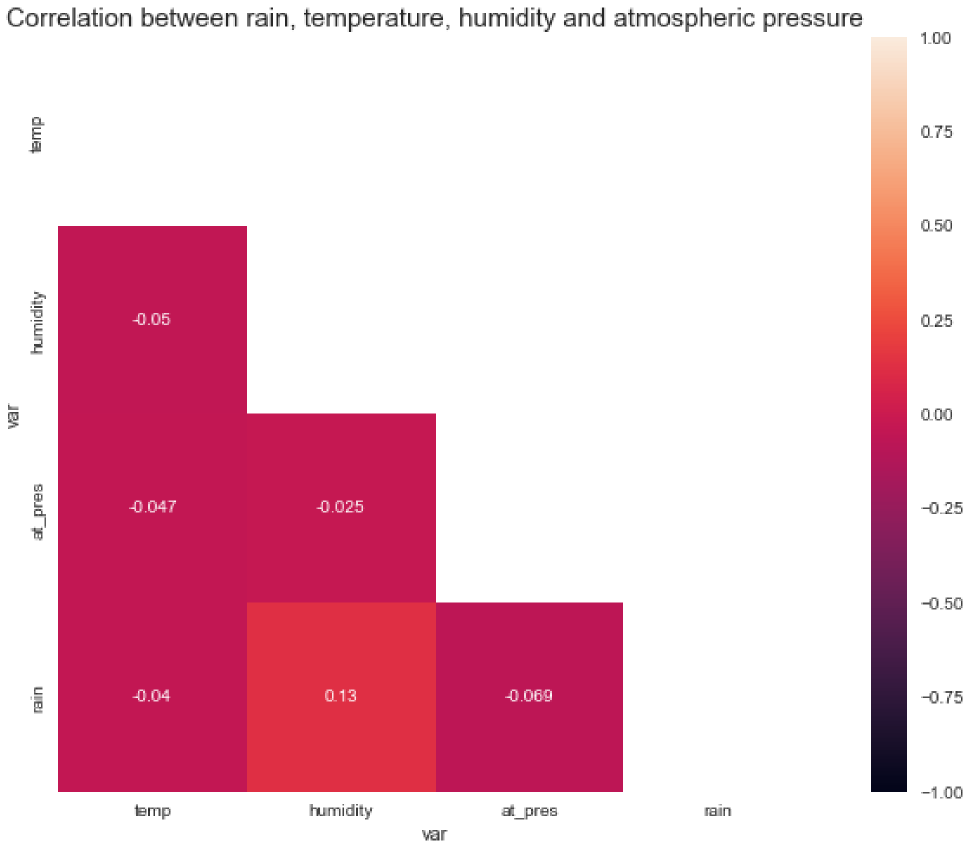
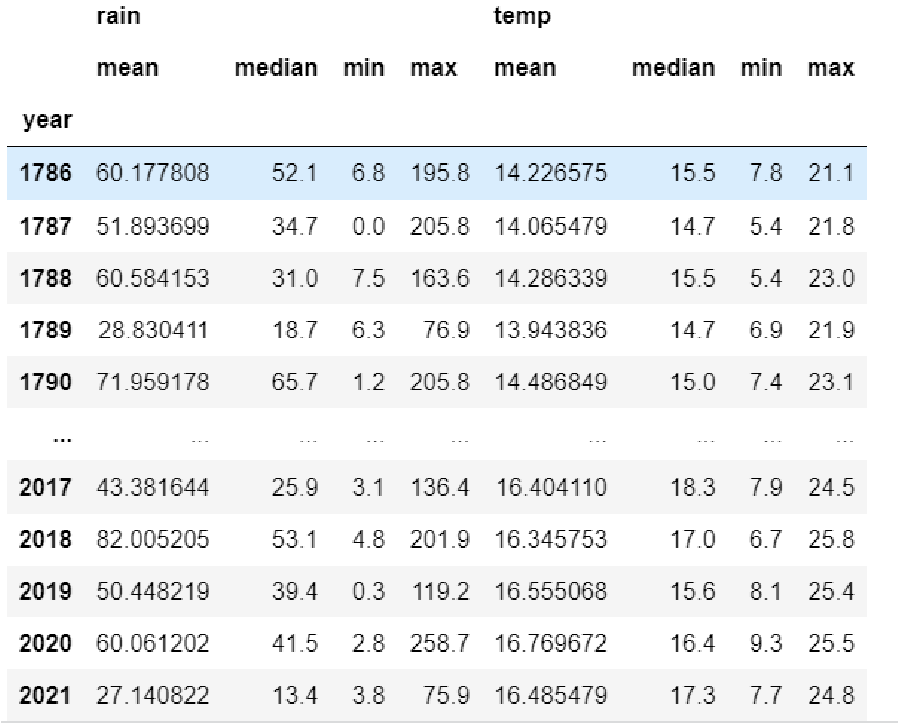


Figure 17.

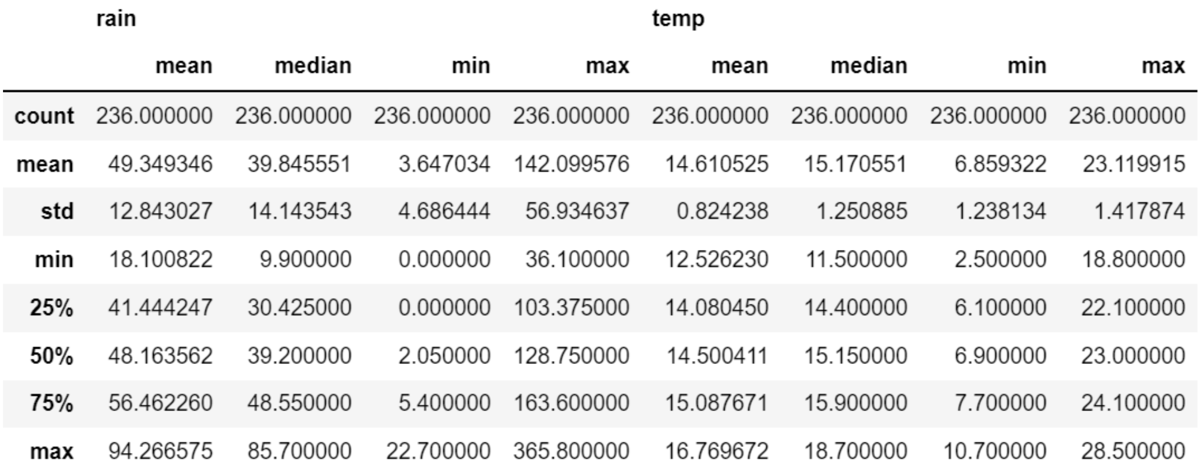
3.1 Rainfall statistics

Table 2. Rainfall and temperature statistics 1786-2021.



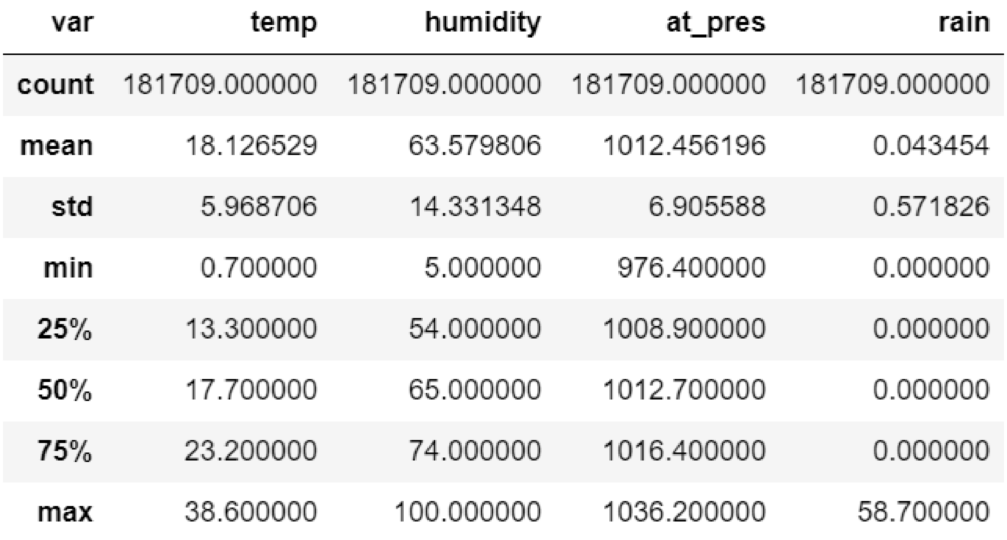
	rain				temp			
	mean	median	min	max	mean	median	min	max
year								
1786	60.177808	52.1	6.8	195.8	14.226575	15.5	7.8	21.1
1787	51.893699	34.7	0.0	205.8	14.065479	14.7	5.4	21.8
1788	60.584153	31.0	7.5	163.6	14.286339	15.5	5.4	23.0
1789	28.830411	18.7	6.3	76.9	13.943836	14.7	6.9	21.9
1790	71.959178	65.7	1.2	205.8	14.486849	15.0	7.4	23.1
...	...	...	...	...	...	...	...	...
2017	43.381644	25.9	3.1	136.4	16.404110	18.3	7.9	24.5
2018	82.005205	53.1	4.8	201.9	16.345753	17.0	6.7	25.8
2019	50.448219	39.4	0.3	119.2	16.555068	15.6	8.1	25.4
2020	60.061202	41.5	2.8	258.7	16.769672	16.4	9.3	25.5
2021	27.140822	13.4	3.8	75.9	16.485479	17.3	7.7	24.8

Table 3. Rainfall and temperature statistics 1786-2021 (II).



	rain				temp			
	mean	median	min	max	mean	median	min	max
count	236.000000	236.000000	236.000000	236.000000	236.000000	236.000000	236.000000	236.000000
mean	49.349346	39.845551	3.647034	142.099576	14.610525	15.170551	6.859322	23.119915
std	12.843027	14.143543	4.686444	56.934637	0.824238	1.250885	1.238134	1.417874
min	18.100822	9.900000	0.000000	36.100000	12.526230	11.500000	2.500000	18.800000
25%	41.444247	30.425000	0.000000	103.375000	14.080450	14.400000	6.100000	22.100000
50%	48.163562	39.200000	2.050000	128.750000	14.500411	15.150000	6.900000	23.000000
75%	56.462260	48.550000	5.400000	163.600000	15.087671	15.900000	7.700000	24.100000
max	94.266575	85.700000	22.700000	365.800000	16.769672	18.700000	10.700000	28.500000

Table 4. Rainfall and temperature statistics 2009-2021.



var	temp	humidity	at_pres	rain
count	181709.000000	181709.000000	181709.000000	181709.000000
mean	18.126529	63.579806	1012.456196	0.043454
std	5.968706	14.331348	6.905588	0.571826
min	0.700000	5.000000	976.400000	0.000000
25%	13.300000	54.000000	1008.900000	0.000000
50%	17.700000	65.000000	1012.700000	0.000000
75%	23.200000	74.000000	1016.400000	0.000000
max	38.600000	100.000000	1036.200000	58.700000

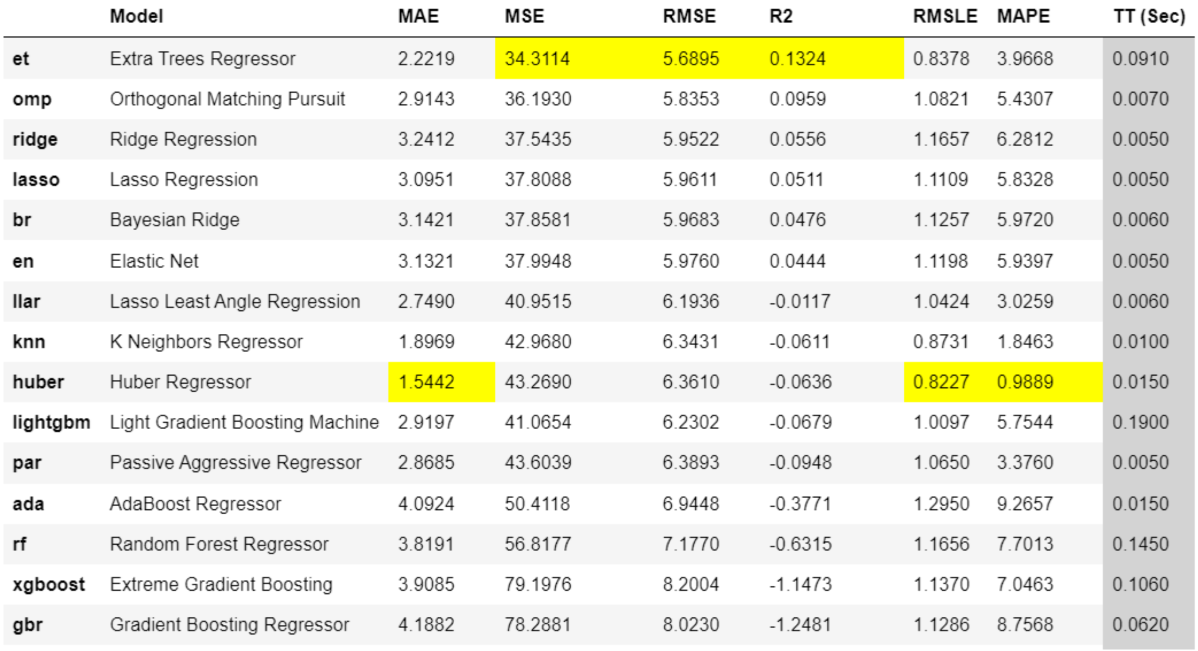
Rainfall depends clearly from other factors that are not present in the dataset analyzed, like hot and cold masses of air, anticyclones, etc.

4. Machine Learning method selection

4.1 Regression

With the help of PyCaret library and Sklearn we have created a Machine Learning model to predict rainfall.

Table 5. Regression models tests.



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	2.2219	34.3114	5.6895	0.1324	0.8378	3.9668	0.0910
omp	Orthogonal Matching Pursuit	2.9143	36.1930	5.8353	0.0959	1.0821	5.4307	0.0070
ridge	Ridge Regression	3.2412	37.5435	5.9522	0.0556	1.1657	6.2812	0.0050
lasso	Lasso Regression	3.0951	37.8088	5.9611	0.0511	1.1109	5.8328	0.0050
br	Bayesian Ridge	3.1421	37.8581	5.9683	0.0476	1.1257	5.9720	0.0060
en	Elastic Net	3.1321	37.9948	5.9760	0.0444	1.1198	5.9397	0.0050
llar	Lasso Least Angle Regression	2.7490	40.9515	6.1936	-0.0117	1.0424	3.0259	0.0060
knn	K Neighbors Regressor	1.8969	42.9680	6.3431	-0.0611	0.8731	1.8463	0.0100
huber	Huber Regressor	1.5442	43.2690	6.3610	-0.0636	0.8227	0.9889	0.0150
lightgbm	Light Gradient Boosting Machine	2.9197	41.0654	6.2302	-0.0679	1.0097	5.7544	0.1900
par	Passive Aggressive Regressor	2.8685	43.6039	6.3893	-0.0948	1.0650	3.3760	0.0050
ada	AdaBoost Regressor	4.0924	50.4118	6.9448	-0.3771	1.2950	9.2657	0.0150
rf	Random Forest Regressor	3.8191	56.8177	7.1770	-0.6315	1.1656	7.7013	0.1450
xgboost	Extreme Gradient Boosting	3.9085	79.1976	8.2004	-1.1473	1.1370	7.0463	0.1060
gbr	Gradient Boosting Regressor	4.1882	78.2881	8.0230	-1.2481	1.1286	8.7568	0.0620

The best model for our dataset is Extra Trees Regressor. In terms of Mean Absolute Error (MAE) and Coefficient of determination (R2) the results are modest since, as we have mentioned early, there are many other factors that affect pluviometry and, for this reason, it is difficult to predict.

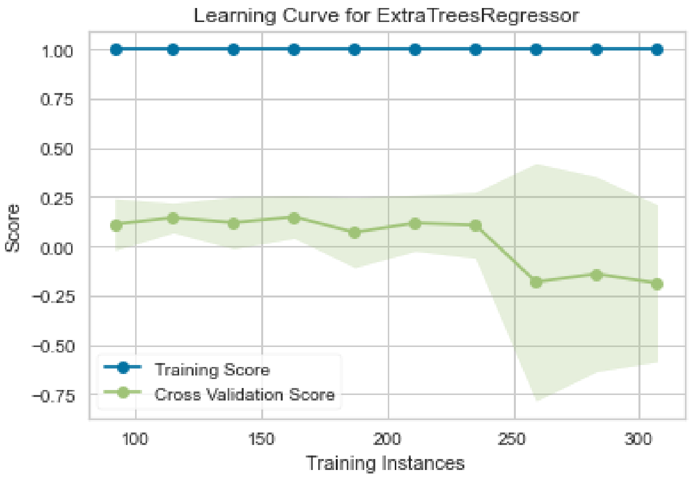
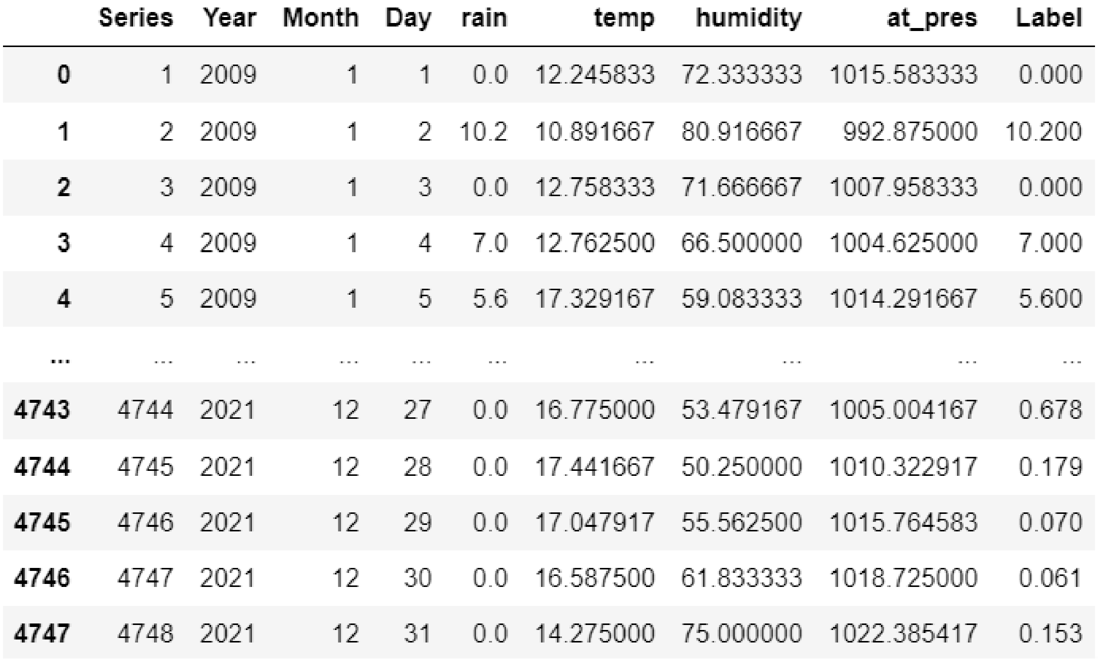


Figure 18.

Table 6. Regression predicted results.



	Series	Year	Month	Day	rain	temp	humidity	at_pres	Label
0	1	2009	1	1	0.0	12.245833	72.333333	1015.583333	0.000
1	2	2009	1	2	10.2	10.891667	80.916667	992.875000	10.200
2	3	2009	1	3	0.0	12.758333	71.666667	1007.958333	0.000
3	4	2009	1	4	7.0	12.762500	66.500000	1004.625000	7.000
4	5	2009	1	5	5.6	17.329167	59.083333	1014.291667	5.600
...	...	...	...	...	...	...	...	...	...
4743	4744	2021	12	27	0.0	16.775000	53.479167	1005.004167	0.678
4744	4745	2021	12	28	0.0	17.441667	50.250000	1010.322917	0.179
4745	4746	2021	12	29	0.0	17.047917	55.562500	1015.764583	0.070
4746	4747	2021	12	30	0.0	16.587500	61.833333	1018.725000	0.061
4747	4748	2021	12	31	0.0	14.275000	75.000000	1022.385417	0.153

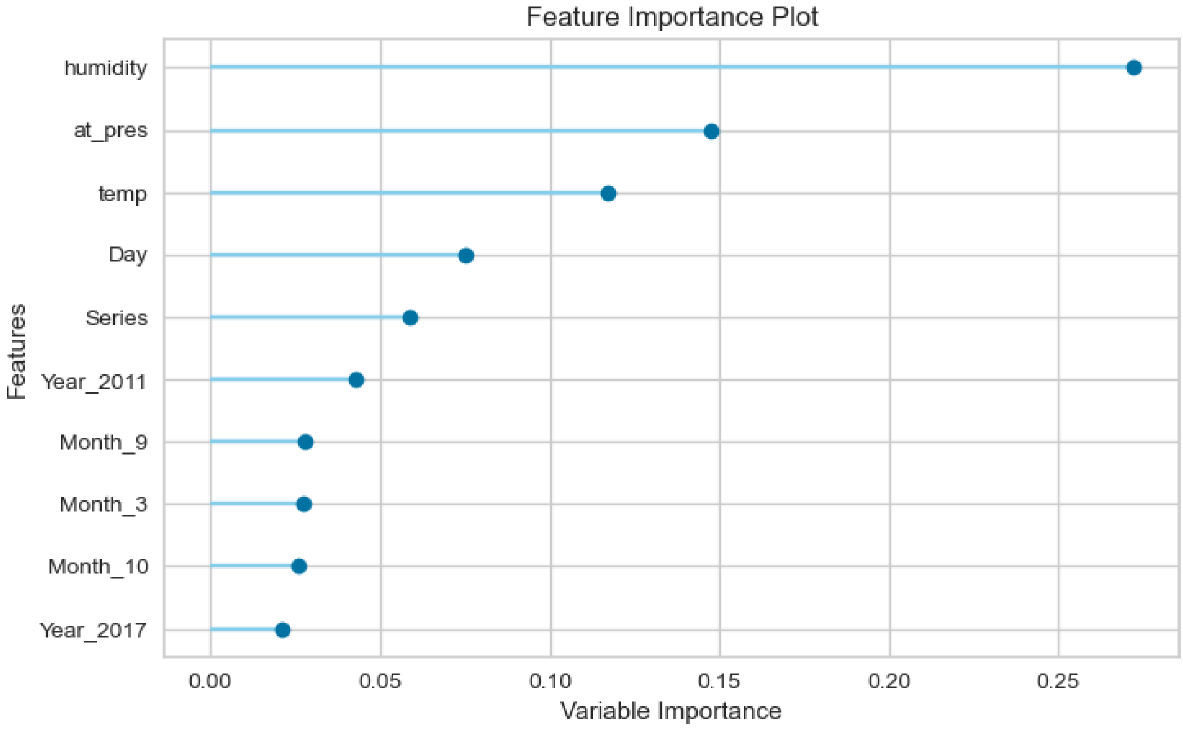


Figure 19.

The most important factors in the model are humidity and atmospheric pressure.

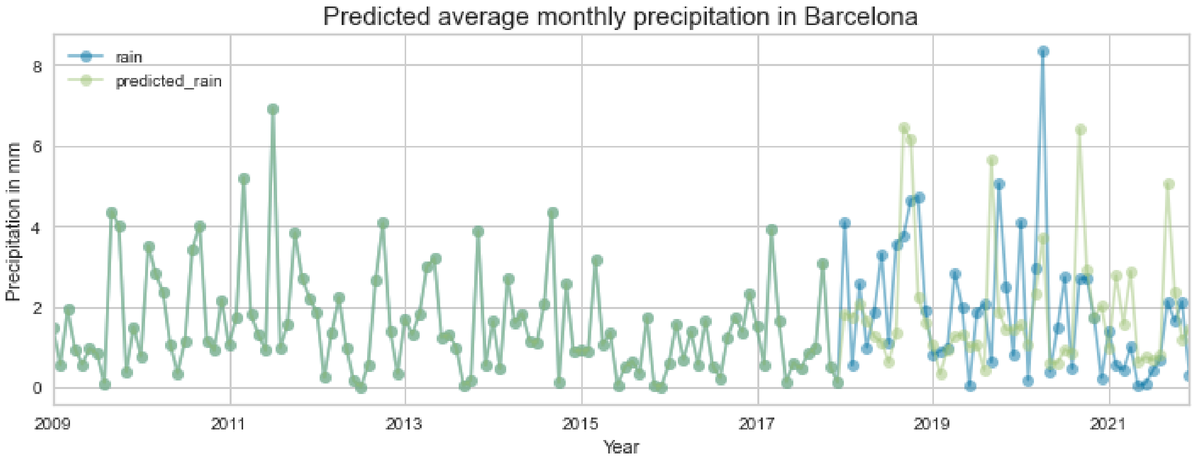
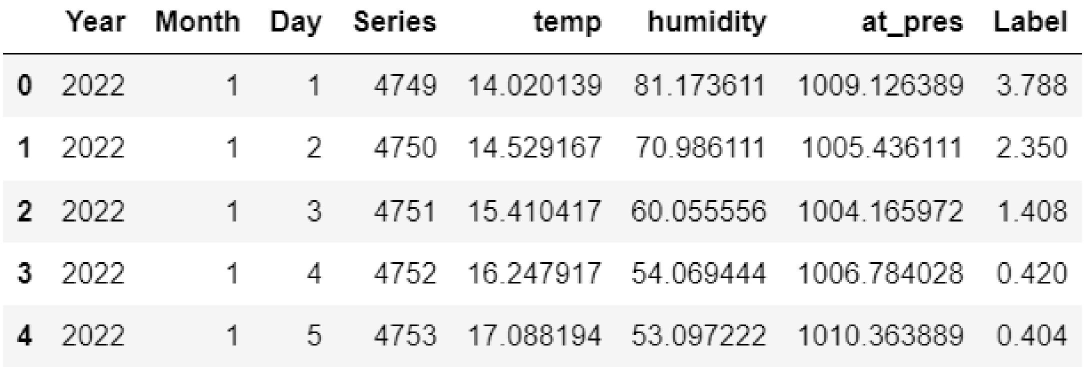


Figure 20.

The predicted data follow the tendency but it’s not very accurate.

Table 7. Rainfall forecast results from 1 of January 2022 to 5 of January.



	Year	Month	Day	Series	temp	humidity	at_pres	Label
0	2022	1	1	4749	14.020139	81.173611	1009.126389	3.788
1	2022	1	2	4750	14.529167	70.986111	1005.436111	2.350
2	2022	1	3	4751	15.410417	60.055556	1004.165972	1.408
3	2022	1	4	4752	16.247917	54.069444	1006.784028	0.420
4	2022	1	5	4753	17.088194	53.097222	1010.363889	0.404

4.2 Clustering

Through the Agglomerative Clustering model we have analyzed the existence of clusters in the rainfall data.

In the dendrogram we can see clearly the clusters generated.

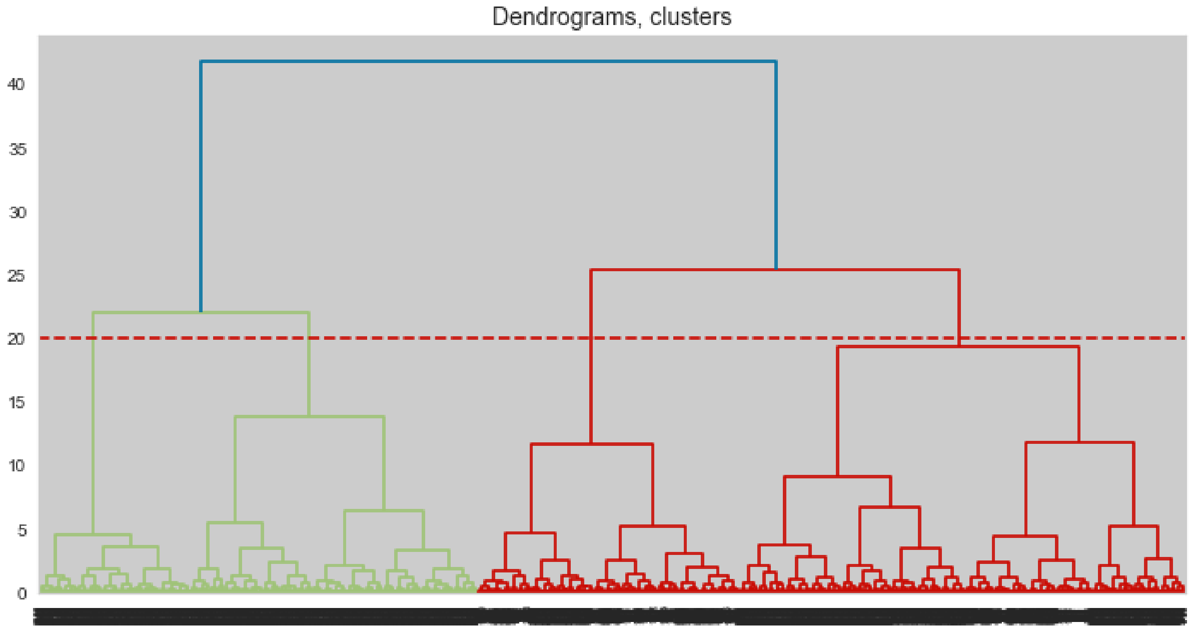


Figure 21.

According to Silhouette scores of the different models, we should analise 2 or 5 clusters.

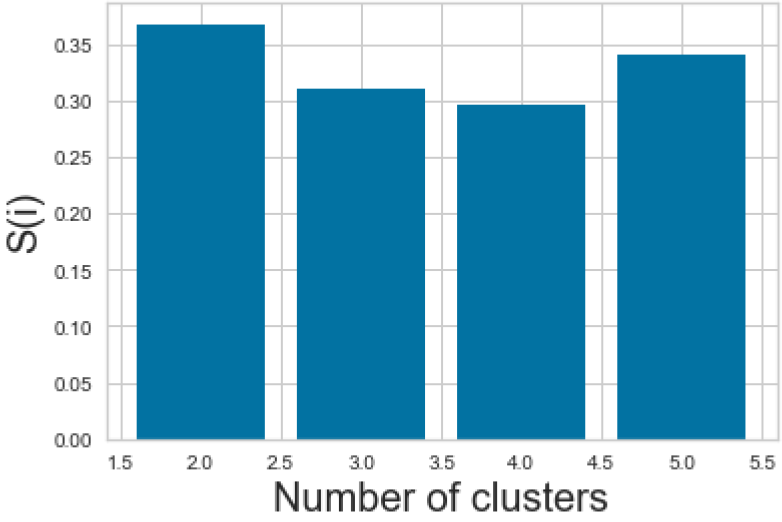


Figure 22

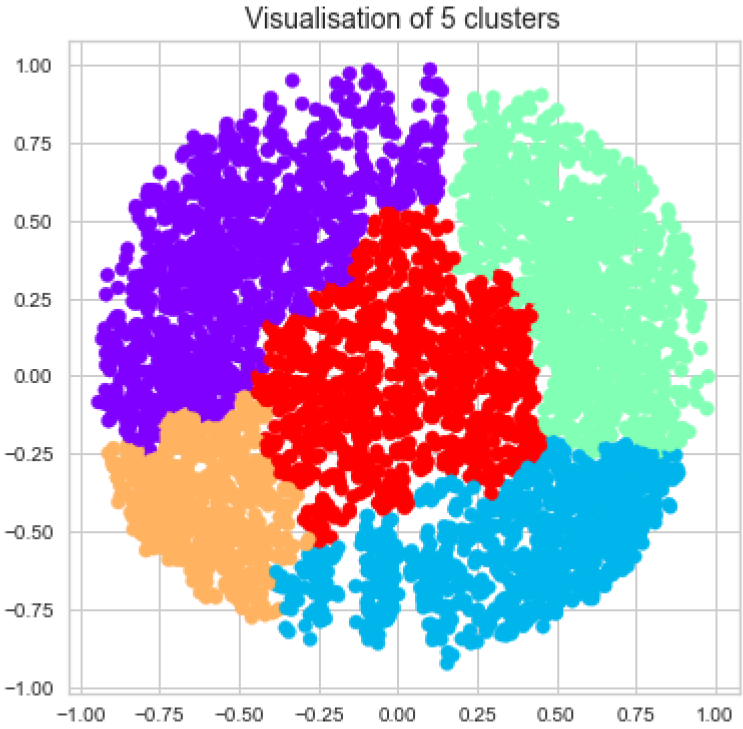


Figure 23.

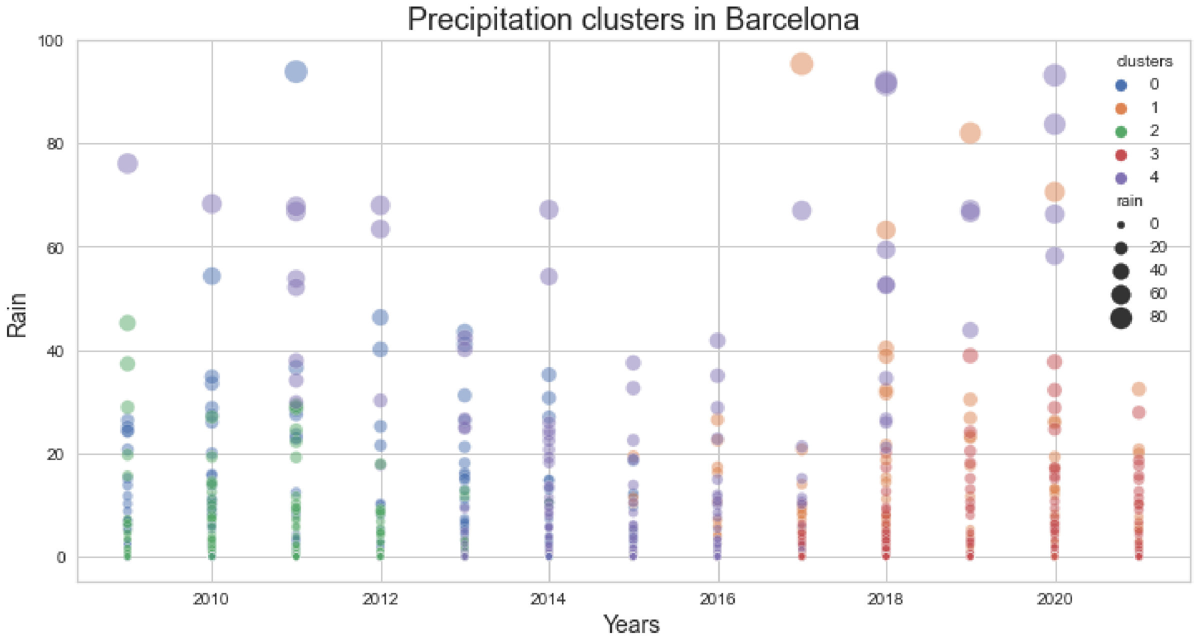


Figure 24.

It appears the algorithm is influenced by years. Being cluster 2 especially between 2009 and 2012, and 3 and 1 between 2017 and 2021. Cluster 4 is very present when there is a lot of rainfall.

5. Final conclusions

Barcelona has a Mediterranean climate in which rain is scarce and downpors could become torrential. Summers are hot and winters, temperate.

Water used in Barcelona comes mainly from Llobregat and Besos rivers and from desalted water from the sea. Human water consumption seems assured in the medium term, but due to climate change this could be otherwise in the future.

The recents yearly average rain seems similar to the last years or indeed superior. So, by now climate change seems that it hasn’t affected pluviometry yet.

What is a worrying concern is the rise of temperature which is developing at an exponential increase.

Our proposal consists of saving water from other sources and consuming more water from aquifers and wells. Some uses are in place right now, like watering parks and streets. Others are traditional uses, like fontains. We believe that a mitigating solution for heat could be spraying water in the street, plazas, parks and in the public transport. Reducing traffic intensity and prioritizing electric motors could also help to reduce the temperature during summer.

6. References

1. “Dades meteorològiques de la XEMA | Dades obertes de Catalunya.” *Dades obertes de Catalunya*, https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee. Accessed 9 February 2022.

2.“Precipitacions acumulades mensuals de la ciutat de Barcelona des de 1786 - Open Data Barcelona.” *Open Data BCN*, 22 October 2019, https://opendata-ajuntament.barcelona.cat/data/ca/dataset/precipitacio-hist-bcn. Accessed 9 February 2022.

3.“Temperatures mitjanes mensuals de l'aire de la ciutat de Barcelona des de 1780 - Open Data Barcelona.” *Open Data BCN*, 22 October 2019, https://opendata-ajuntament.barcelona.cat/data/ca/dataset/temperatures-hist-bcn.

Accessed 9 February 2022.