Energy Consumption in France - Project

August 24th 2021

Florence D'AMORE

1- Introduction

1-1. Background

The purpose of this study is to analyze data on electricity consumption in France, from 2013 until March 2021.

This project follows on the project which carried out during my DataScientest training : AnEnPy. It was written in a team.
The complete database analysis has already been carried out : cleaning, visualization, statistical analysis.

This study is a summary of AnEnPy Project.

How evolute the energy consumption in France?

1-2. Problem

Is it possible to predict the energy consumption in France?

2- Data acquisition

2-1. Data sources

The data sourced from :

- Information relating to the production and consumption of renewable and non-renewable energies, in France, by region from 2013/01/01 to 2021/03/31 : éCO2mix

- Information relating to meteorological observations, in France, by municipality, from 2013/01/01 to 2021/01/04 : synop

- Legal population in 2018, in France : population

By combining the datasets, we can better understand the evolution of consumption.

2-2. Data cleaning

Concerning the file “éCO2mix” :

- There are 1752191 rows and 65 columns.

- Only the columns interesting for this project were keeped : “Code INSEE region”, “Région”, “Date”, “Heure”, “Consommation (MW)”, “Thermique (MW)”, “Nucléaire (MW)”, “Eolien (MW)”, “Solaire (MW)”, “Hydraulique (MW)”, “Pompage (MW)”, “Bioénergies (MW)”.

- There aren’t duplicate data.

- The missing values were deleted, and for “Nucléaire (MW)” and “Pompage (MW)”, there were replaced by 0.

Concerning the file “synop” :

- There are 1048574 rows and 15 columns.

- Only the columns interesting for this project were keeped : “Date”, “Région”, “Température (°C)”, “Humidité”, “Visibilité horizontale”, “Précipitations dans les 24 dernières heures”.

- There aren’t duplicate data.

- The missing values were delated.

Concerning the file “population” :

- There are 16 rows and 7 columns.

- Only the columns interesting for this project were keeped : “CODREG”, “REG”, “PTOT”.

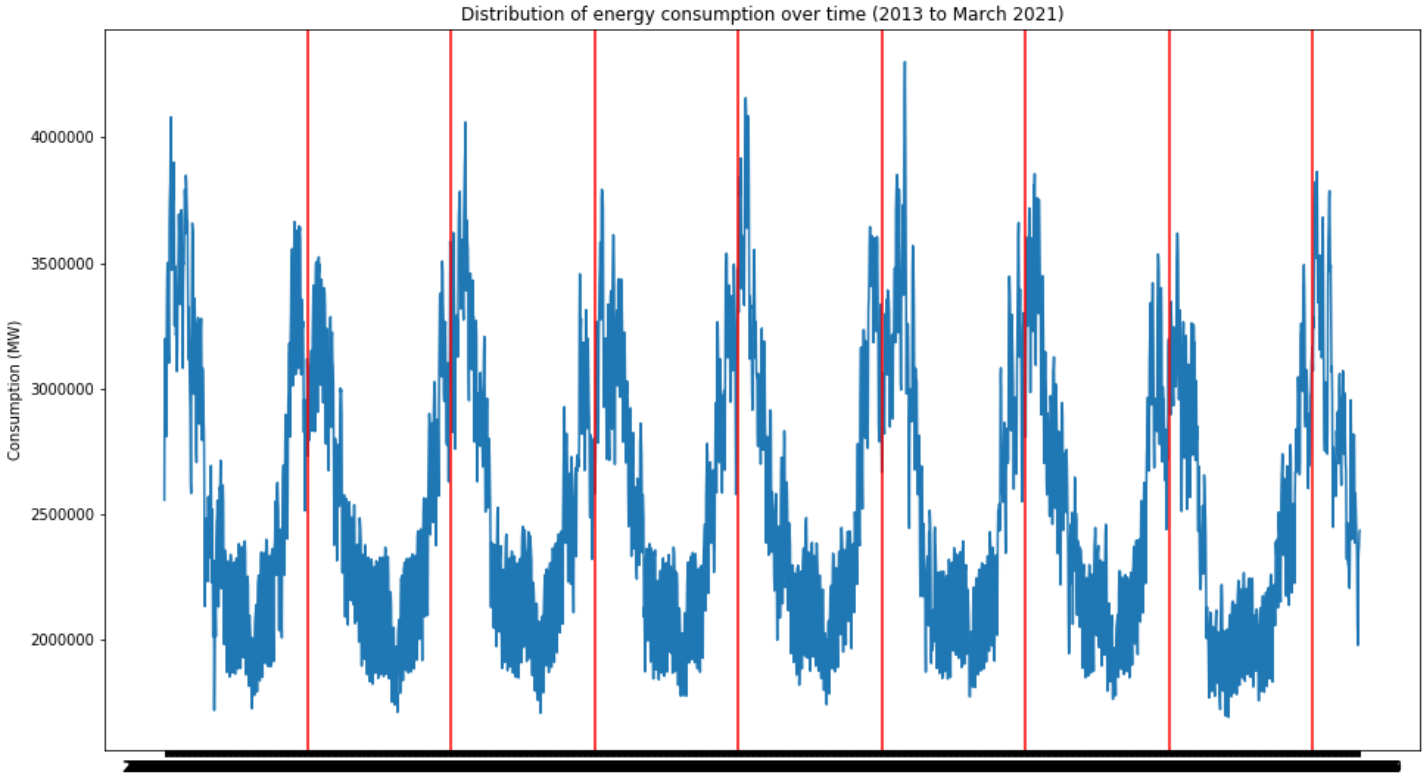
- There aren’t duplicate data.

- There aren’t missing values.

3- Exploratory Data Analysis

3-1. Distribution of consumption over time

Figure 1 :



The red line indicates the year.

According to figure 1, the consumption decreases around the middle of the year. When the temperatures are higher, in France.
We can see that the energy consumption is cyclical over time.

3-2. Consumption depending on the weather

As energy consumption decreases in summer, let's look at the influence of weather conditions on consumption.

Figure 2 :

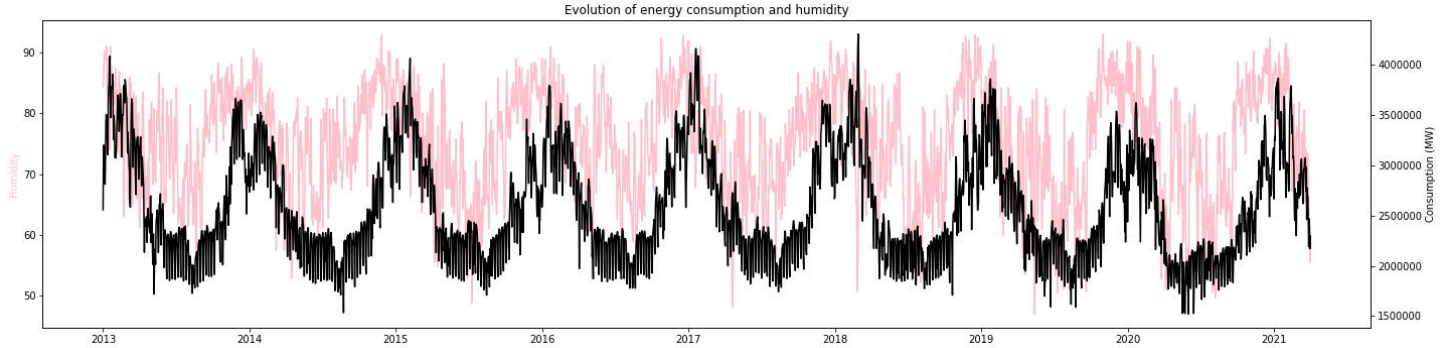


Figure 3 :

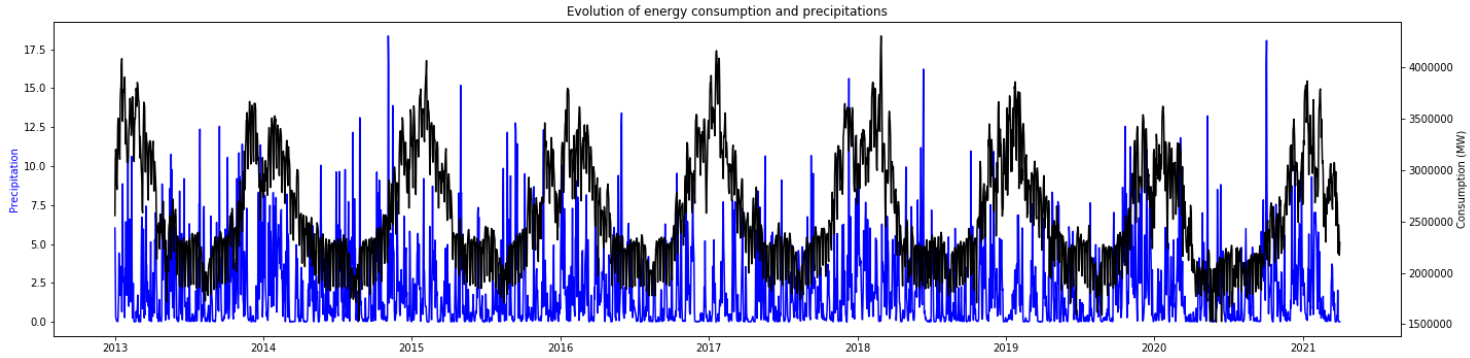


Figure 4 :

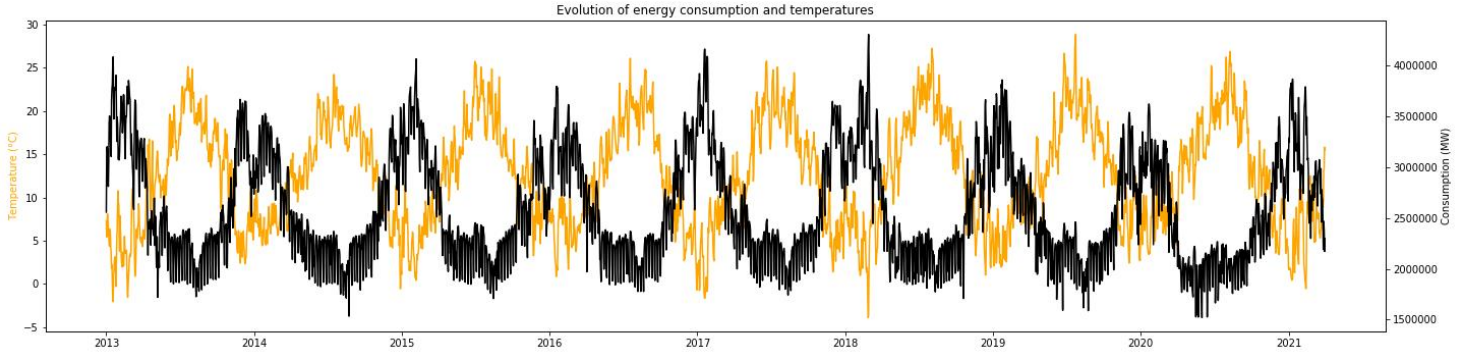
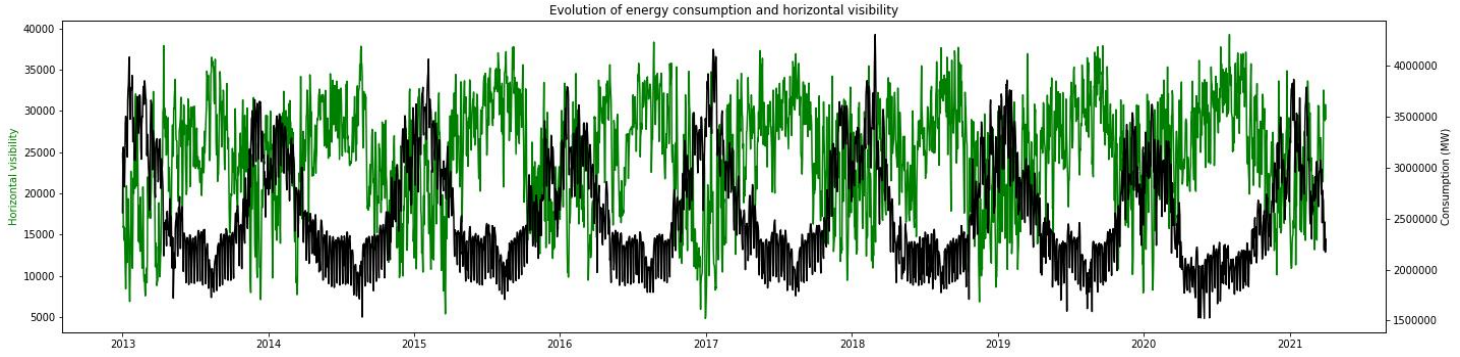


Figure 5 :

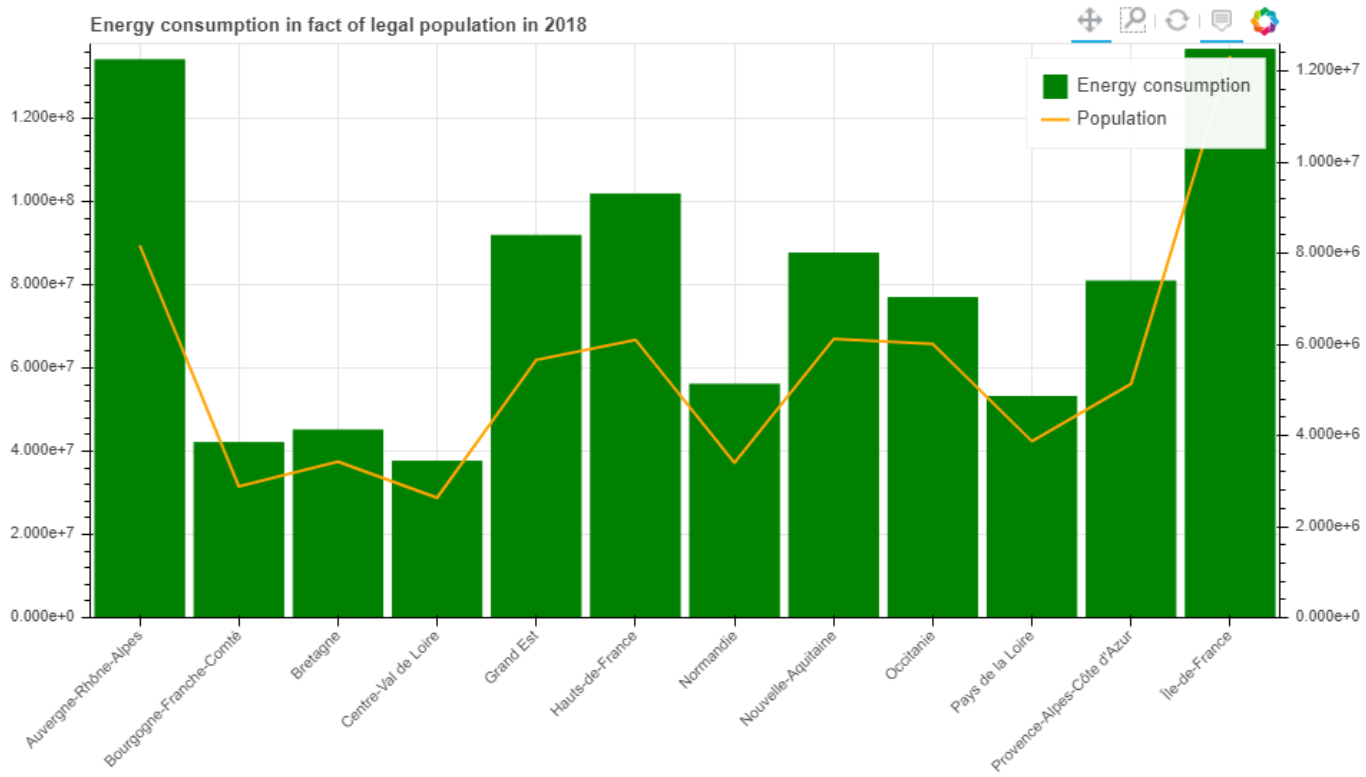


According to figure 2, 4 and 5, humidity, temperatures and horizontal visibility seems to influence energy consumption.
The consumption increases when the humidity in the air is high, and it decreases when the temperature and the horizontal visibility are high.

The precipitations don't seem to influence the energy consumption.

3-3. Relationship between energy consumption and legal population

Figure 6 :



The figure 6 indicates that the most populated regions are the regions which consume the most.

3-4. Correlation between the datasets "éCO2mix" and "synop"

Figure 7 : Heatmap

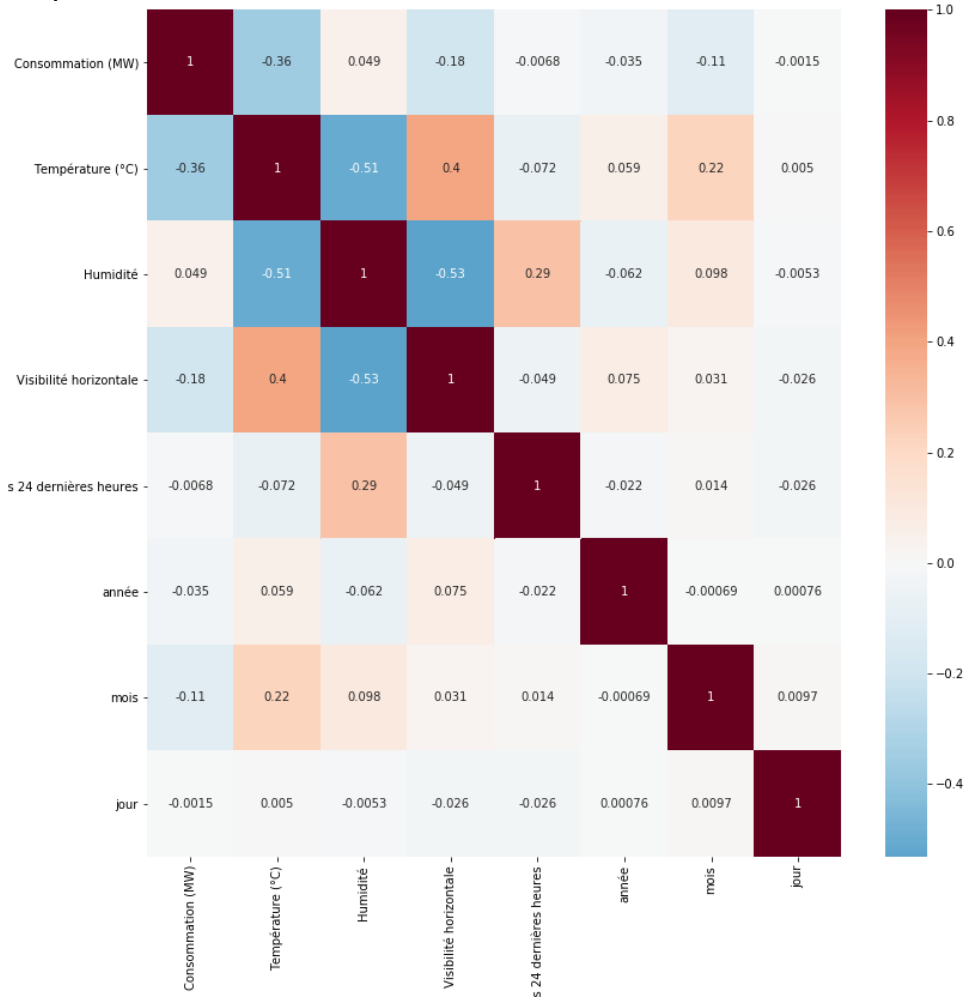


Figure 8 : Correlation values

Data	Correlation values
Consommation (MW)	1.000000
Température (°C)	0.356824
Visibilité horizontale	0.184401
Mois	0.109626
Humidité	0.049148
Année	0.035428
Précipitations dans les 24 dernières heures	0.006838
jour	0.001540

The correlation values indicate that temperatures, horizontal visibility and humidity are the meteorological parameters best correlated with consumption.

4- Predictive modeling

We have chosen to remove the data dating from 2021, because we have observations until the month of March, which explains the increase in consumption in 2021.

We split of dataset : “Consommation (MW)” in target variable and the rest in “data”
And, we split of training data (2013 to 2017) and the data test (2018 to 2020)

4-1. Model SGDRegressor

Figure 9 :

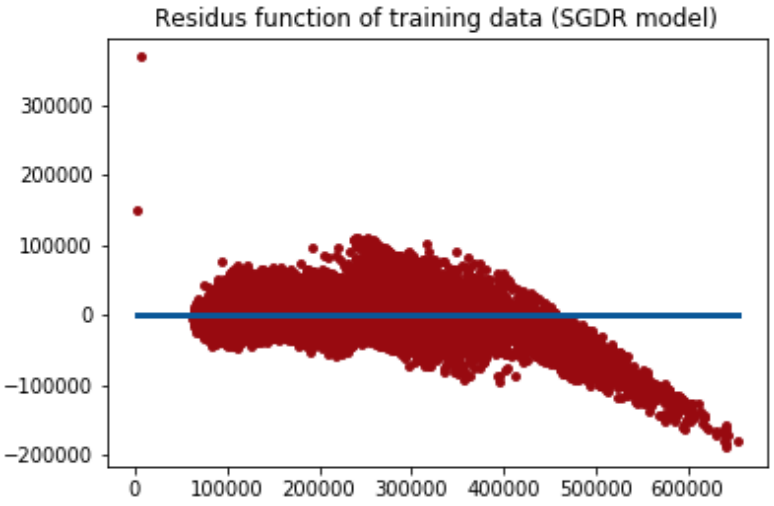
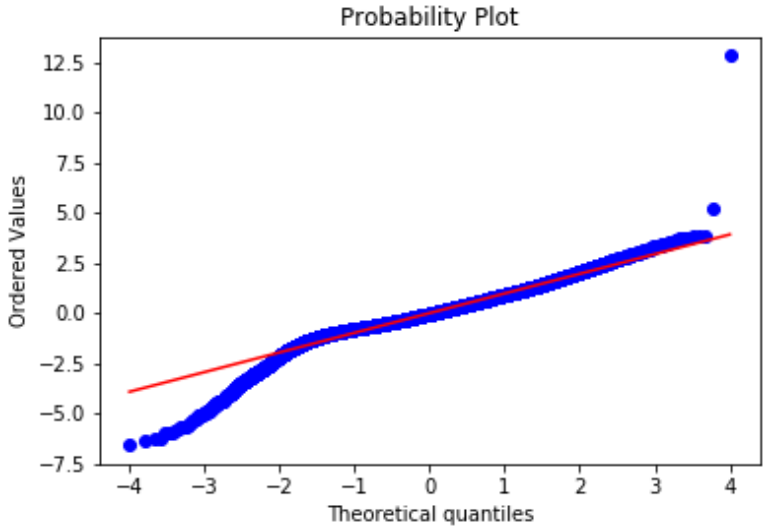
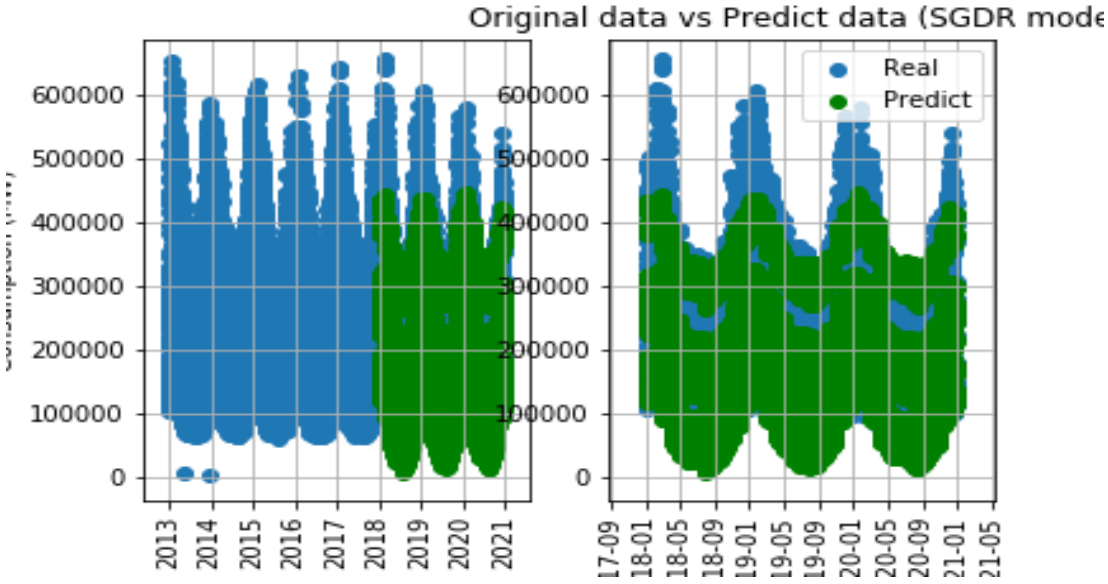


Figure 10 :



The mean of residues is : -7.3489

Figure 11 :



Results :

Score Train : 0.9229232943475061
Score Test : 0.8356436905202062

Mean Squared Error Train : 827232485.6118164
Mean Squared Error Test : 1580360271.5835862

Mean Absolute Error test : 32480.04067602805
Mean Absolute Error train : 21330.105001421703

4-2. Model Lasso

Figure 12 :

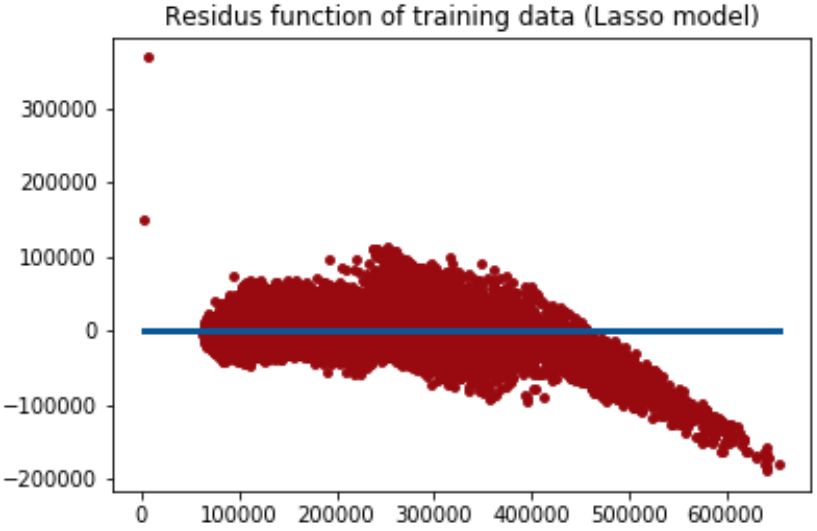
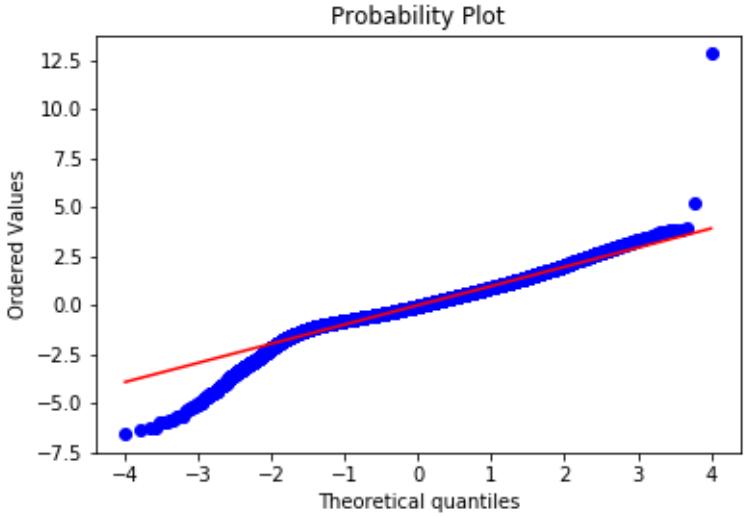
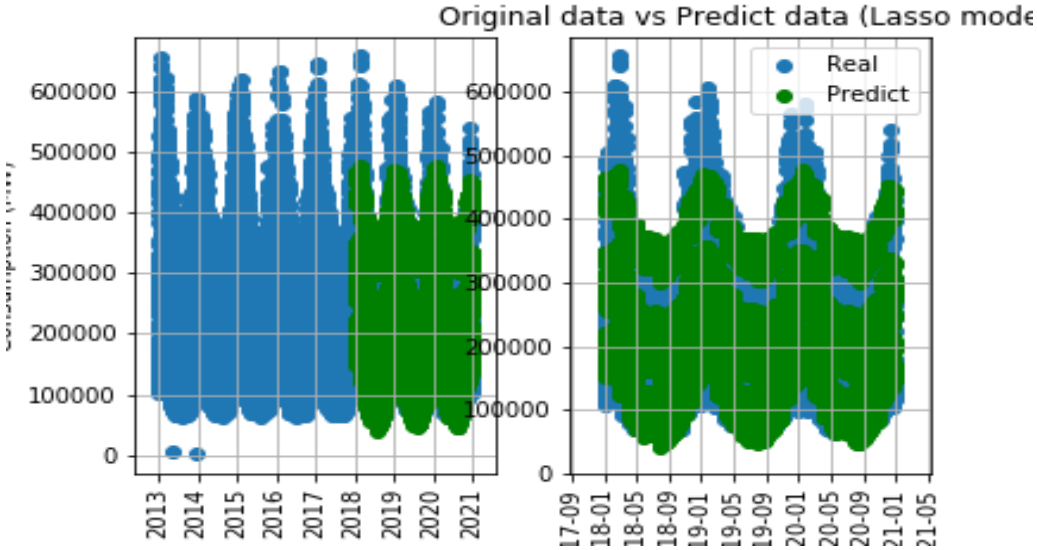


Figure 13 :



The mean of residues is : -8.7034641615559l8e-12

Figure 14 :



Results :

Score train: 0.9229835534939903
Score test : 0.900409871856391

Mean Squared Error test : 957604137.3661265
Mean Squared Error train : 826585749.0510747

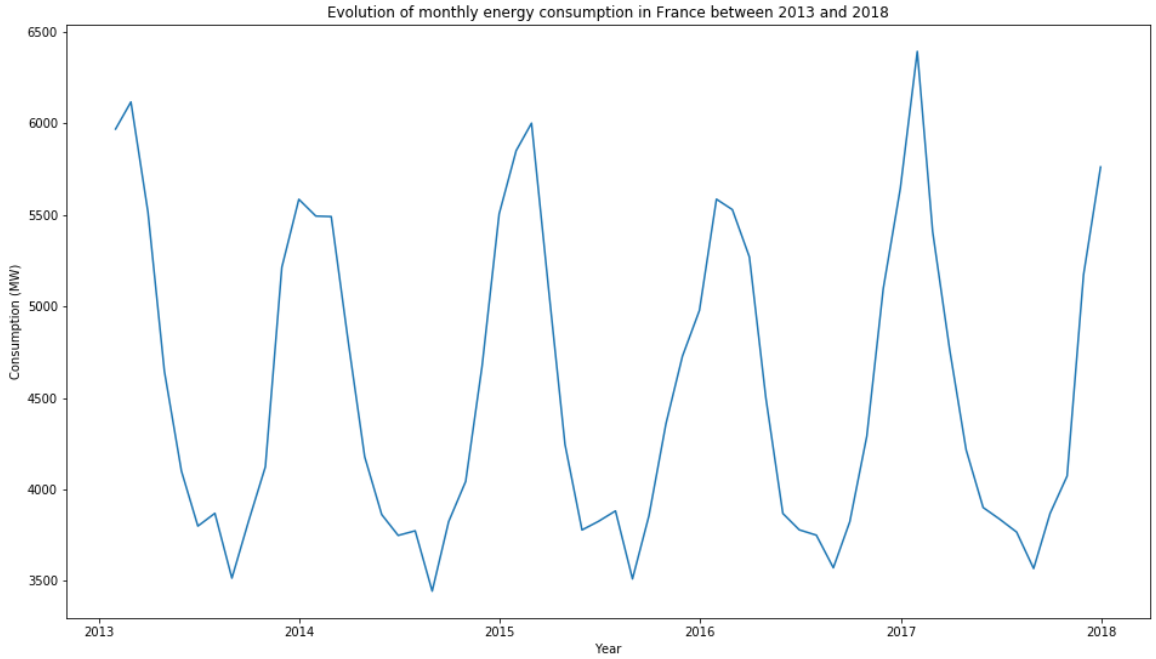
Mean Absolute Error test : 23075.070830936842
Mean Absolute Error train : 21237.978005079258

These two linear regression models indicate correct predictions.

4-3. Time series

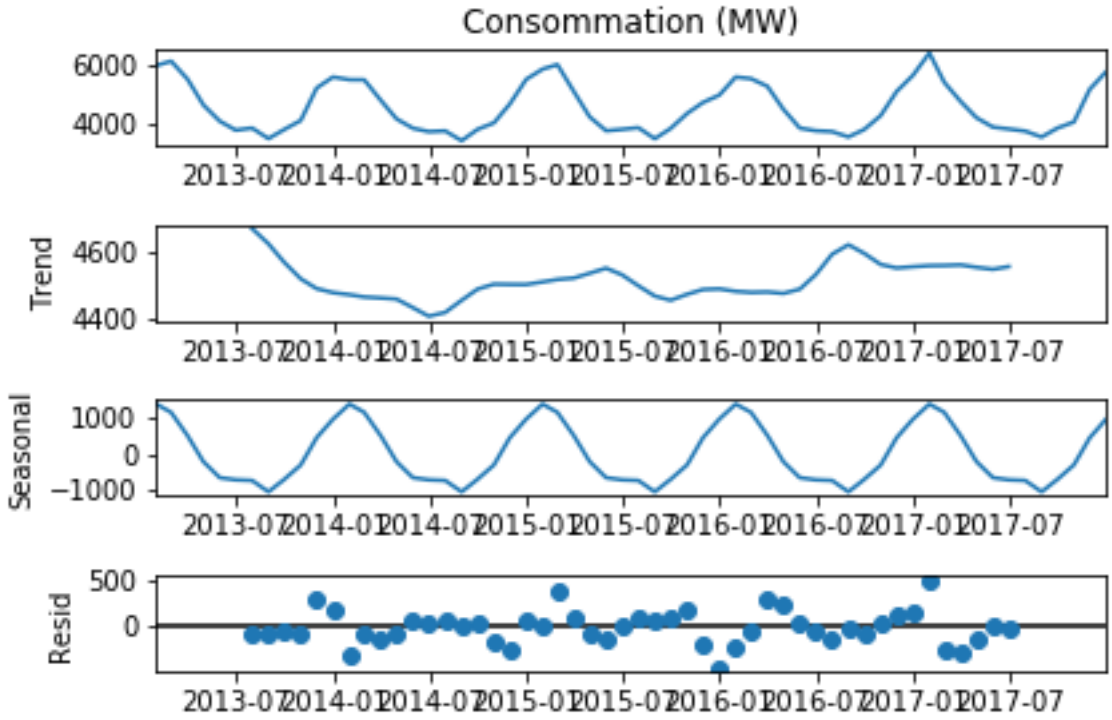
We split of training data (2013 à 2017) and the data test (2018 à 2021).

Figure 15 :



As seen previously, the figure 15 shows that the consumption is cyclical.

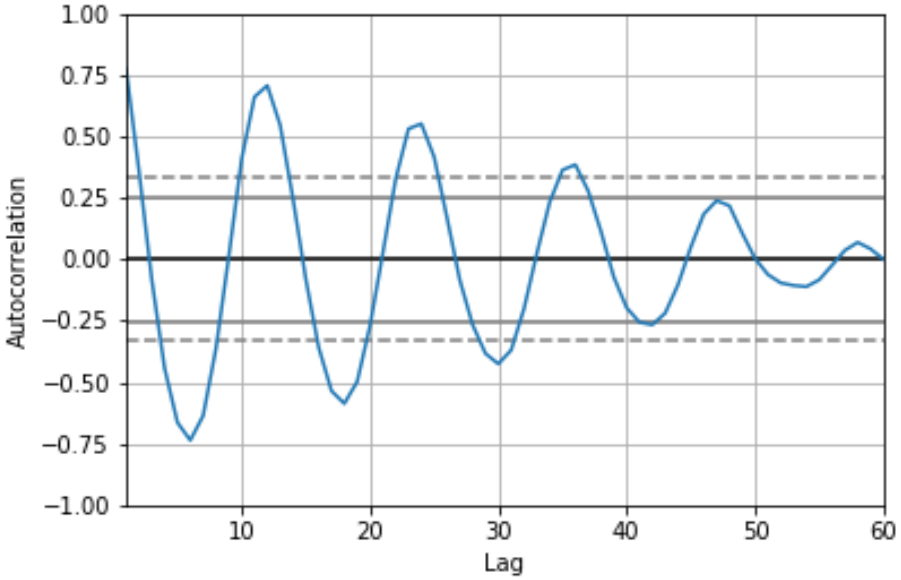
Figure 16 : Decomposition of the series



According the figure 16, the residue doesn't appear to have large variations over time. The decomposition is successful.
The seasonality is period 12 (annual).

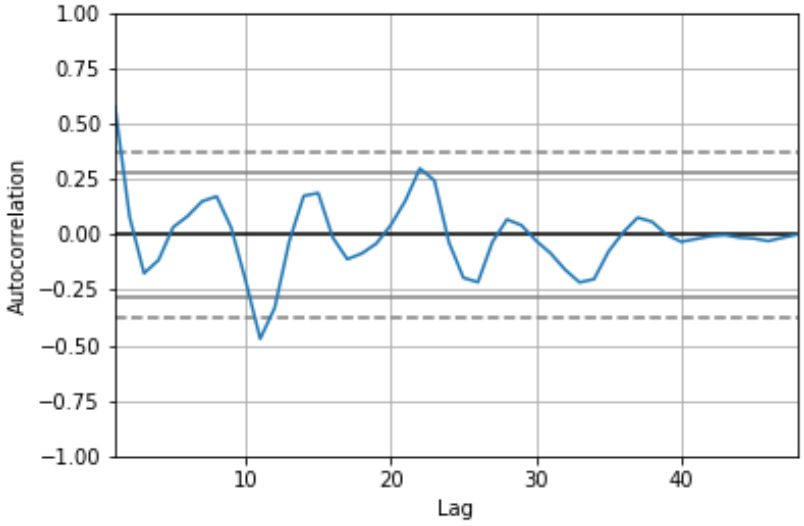
This series is modeled by an additive model. We can seasonally adjust it using linear regression.

Figure 17 : Autocorrelation function on the data series



The simple autocorrelation decreases rapidly towards 0, the process is stationary.

Figure 18 : Apply a seasonality differentiation and autocorrelation function on the differentiated series



The result is fairly satisfactory. We have an estimator of the simple autocorrelogram of our process.

To ensure the stationarity of the differentiated series, we carried out the statistical test of Dickey-Fuller.

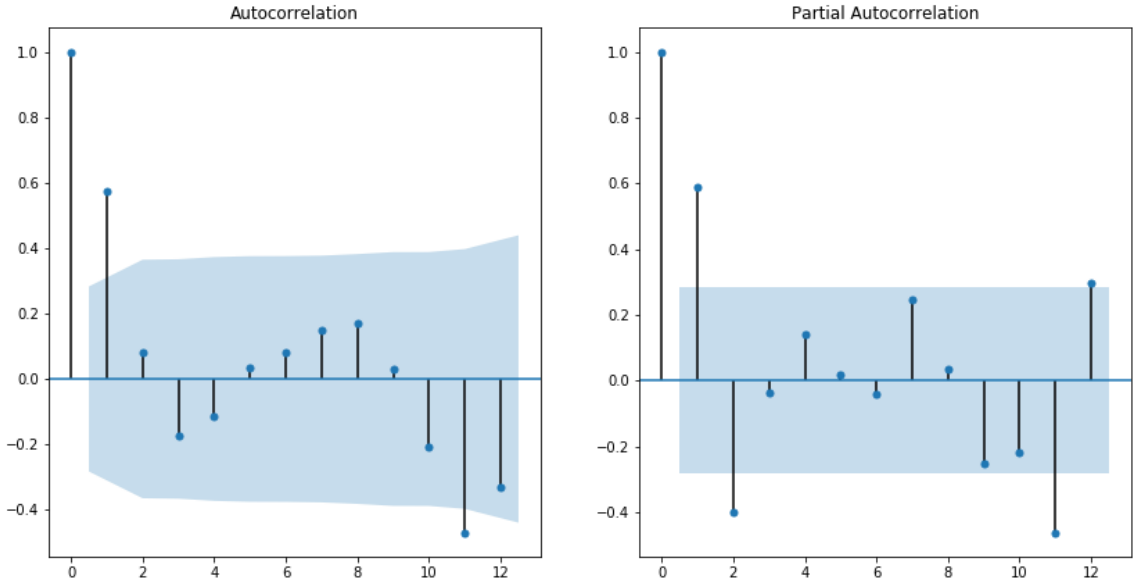
Figure 19 : Statistical test of Dickey-Fuller

Statistiques ADF : -4.333832414670628
p-value : 0.0003879570766472639
Valeurs Critiques :
1%: -3.6209175221605827
5%: -2.9435394610388332
10%: -2.6104002410518627

The p-value is less than 0.05, so H0 is rejected and the series is stationary.

We search the parameters P, Q, p, q

Figure 20 : Plot the simple autocorrelation function and the partial autocorrelation function



Seem that the simple and partial autocorrelograms cancel each over out from rank 2.

We use linear regression by trying to estimate the best model.

Figure 21 : SARIMAX Results

SARIMAX Results

Dep. Variable:	Consommation (MW)	No. Observations:	60
Model:	SARIMAX(1, 1, 0)x(1, 1, [1], 12)	Log Likelihood	-328.518
Date:	Tue, 24 Aug 2021	AIC	665.036
Time:	16:08:17	BIC	672.437
Sample:	01-31-2013	HQIC	667.821
	- 12-31-2017		
Covariance Type:	opg		

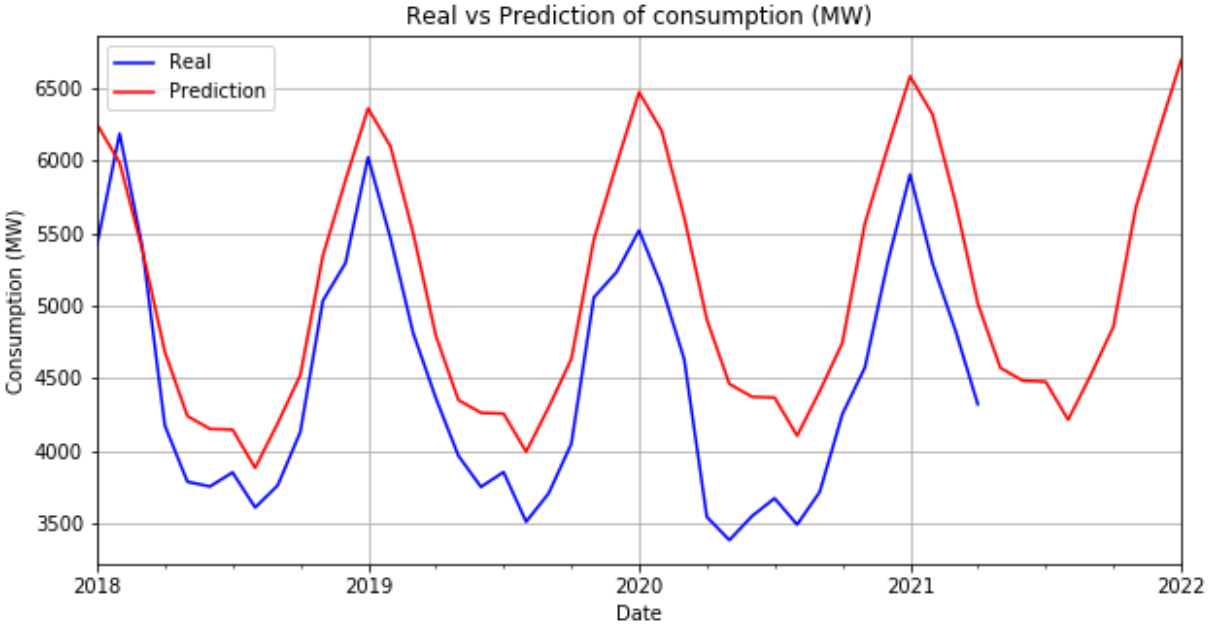
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0635	0.227	-0.280	0.780	-0.508	0.381
ar.S.L12	0.1930	0.318	0.608	0.543	-0.430	0.816
ma.S.L12	-0.9966	0.189	-5.263	0.000	-1.368	-0.625
sigma2	5.335e+04	3.6e-06	1.48e+10	0.000	5.33e+04	5.33e+04

Ljung-Box (Q):	51.71	Jarque-Bera (JB):	50.20
Prob(Q):	0.10	Prob(JB):	0.00
Heteroskedasticity (H):	2.57	Skew:	-1.39
Prob(H) (two-sided):	0.07	Kurtosis:	7.23

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

After having redone this process by changing the regression parameters, and after analysis of the results, the SARIMA model SARIMA(1,1,0)(1,1,1,12) is satisfactory.

Figure 22 : Prediction between Jan 2018 and Dec 2021 to compare with the test dataset



The Mean Absolute Prediction Error is 14.27%.

According the figure 22 and the result of MAPE, the prediction of energy consumption seems correct until the end of 2019 – beginning of 2020.
After 2020, there seems to be a significant gap between actual consumption and the prediction.

5- Conclusion

The energy consumption in France seems depend of the weather and the population density.
The COVID 19 has changed our consumption habits, making it harder to predict.

It will be interesting to redo this analysis with the full year of 2021.

It will also be interesting to analyze the energy production to put them in parallel with consumption.
This will be the subject of an other study.

6- Sources

- "AnEnPy", DataScientest.com
- Energy Consumption Forecast, by jrhea : Energy