Here is a summary of the results about the different parts of this work. The method and interpretations of the results will be explained.

The whole project was subclassed in 5 different categories depending on the kind of work that was done. Thus :

* ***1\_Creation\_df*** corresponds to the importation and manipulation of the data which will be used throughout the entire project.
* ***2\_Visulization*** deals with exploratory graphical visualization to have an insight of our data.
* ***3\_Precipitation\_understanding*** handles the staticial issues to figure the links between precipitation, temperature, location and seasonnality out.
* ***4\_Temperature*** tries to define how the temperature evolved these 45 last years in France by statistical methods.
* ***5\_Prediction\_Temperature*** tries to forecast mean temperature in the next 36 months using temperature data.

**1\_Creation\_df :**

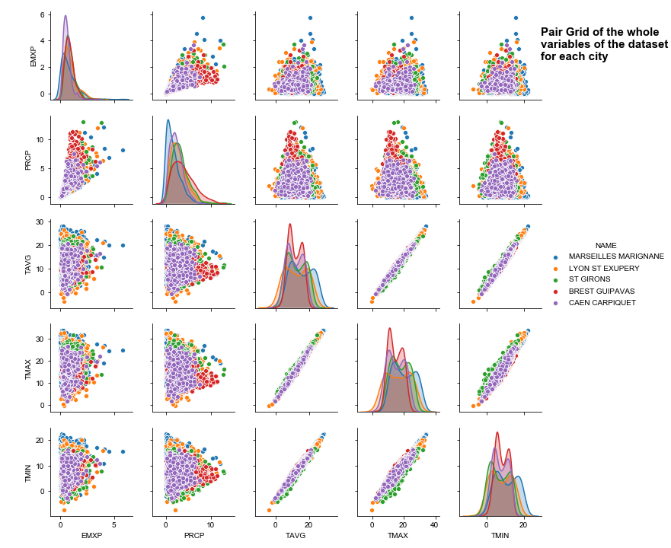
In this part, climatic data from the NOAA were imported as csv file. The following manipulations needed to be achieved.

First of all, we dealt with missing values (NA’s). Rows of cities and year 2020 containing too many NA’s were removed from the dataset. For the rest of our data, the few remaining NA’s were transformed into mean values for each column.

Secondly, the temperature had to be converted from Fahrenheit to Celsius.

Finally, after some minor checking we simply saved the new data base as a csv. Thus, we got monthly temperature and precipitation variables between 1975 and 2019 and indexed by 5 cities in France.

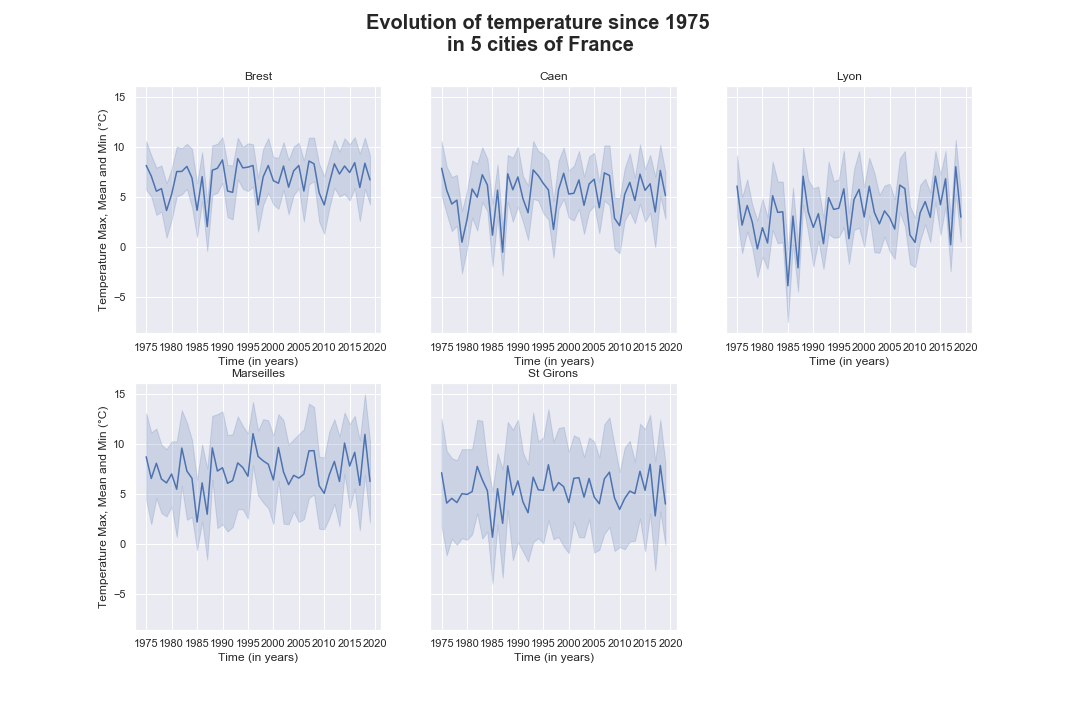
**2\_Visualization :**

Before doing any research about climatic questions we needed to briefly observe our data. This second part entierly aimed at doing that. Our data were visualized in order to explore them.

We began by a Pairplot to get every relation between each variable.

Some inferences can be done according to this plot :

* All the temperature variables are linearly correlated.
* The more the temperature increases, the more the variation of PRCP and EMXP increase.
* PRCP variation is more important in the north than south.

What we are the most interested in this project is the evolution of the temperature in France, which is the point of part 4 and 5. We found interesting to plot the distinct temperature for each studied city. Thus, we subploted some lineplots of seaborn library. This enabled me to practice with Matplotlib and Seaborn.

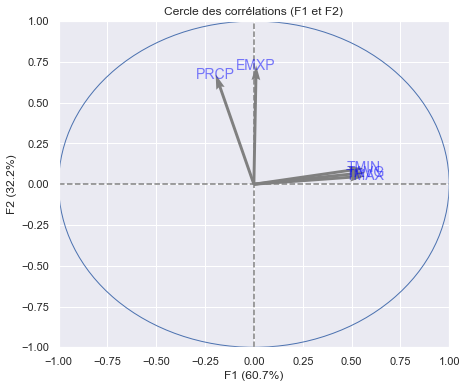
We can notice that there is a slight but noteworthy increase of temperature with time for each of the cities.

Also, the temperatures of Marseilles and Brest look notably higher than the others.

**3\_Precipitation\_understanding:**

This third part concerns the first scientifical question we tried to answer : Is the precipitation (PRCP) influenced by other climatic variables?

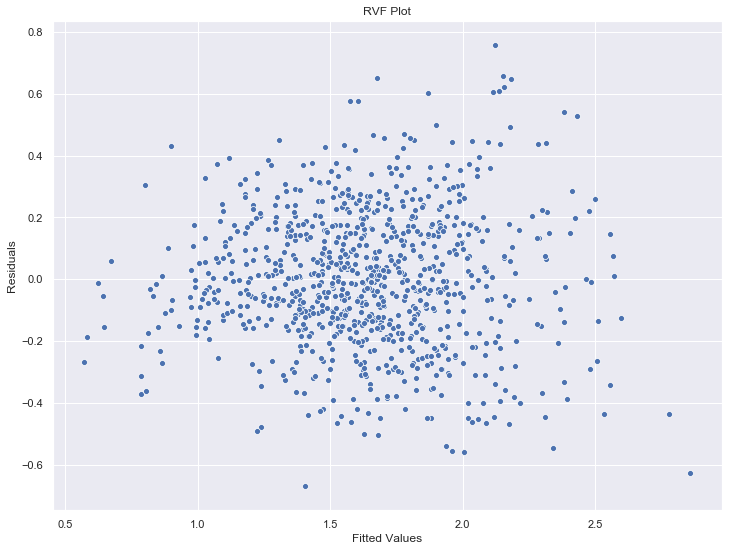
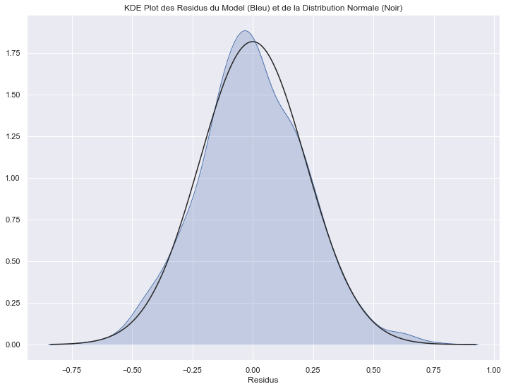
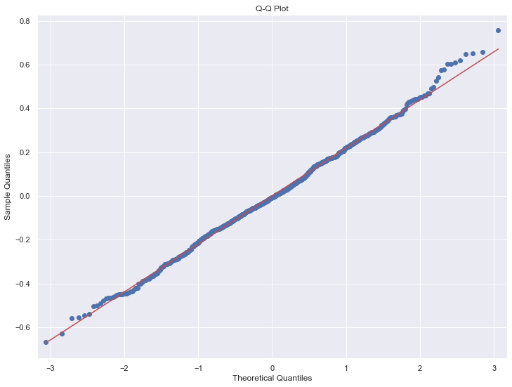
To answer this question, we began with finding which variable could be interesting to test. Of course, we picked temperature variables, extreme daily precipitation and the location but we also added the saisonnality. To do so, we transformed the data to get saisonnal rows.

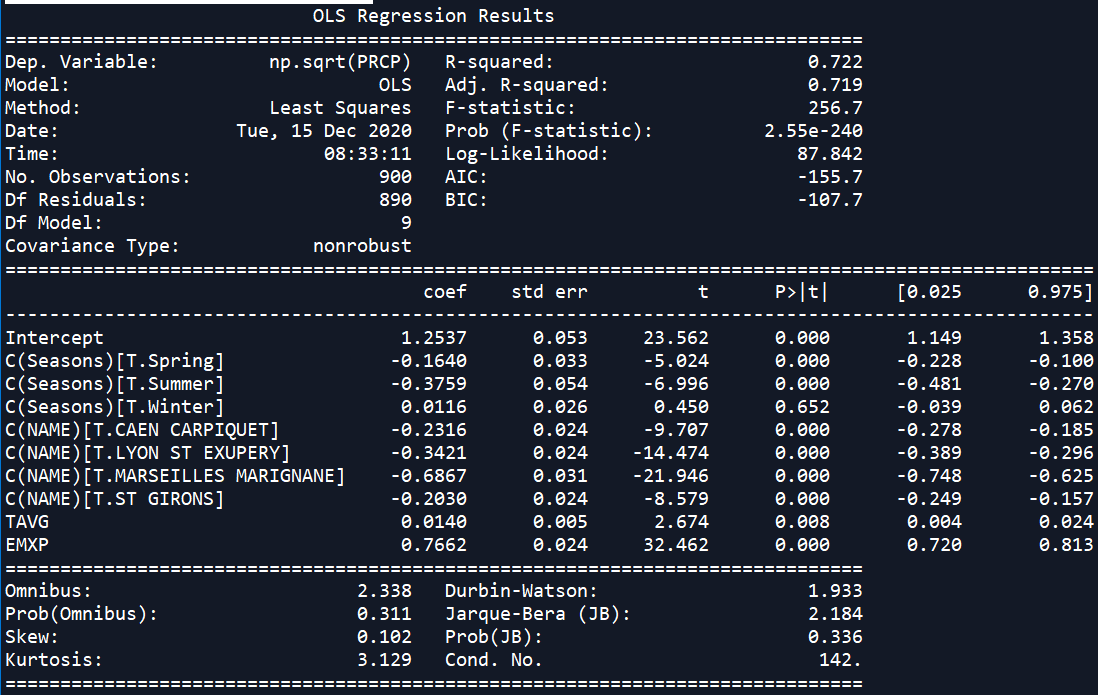
Before doing any modeling, we made sure that no autocorrelation was present between our explanatory variables. We could have used our previous pairplot to observe linear relations but I definitly wanted to train my conception of ACPs in this project. What we are really interested in the ACP is not the reduction of dimensions but only to check using the correlation circle if some variables are close together.

Thus, we can say that as predicted TAVG, TMIN and TMAX are totally confounding. We see that PRCP and EMXP are closely correlated too !

Now we can create our model. We know that our response variable (PRCP) is continuous, so that it will follow a normal distribution. Thus, we need to do a Linear Regression. But before any computation we must check the distribution of our residuals !

This way, we initiated a Q-Q plot, plot of normality and a plot of homoscedasticity but the homoscedasticty was definitely not good. A strong rise of the variance indicated that our residuals were not aleatory distributed. The cause of this issue appeared to be the nonlinear relation between our response variable and our explanatory variables. To resolve this problem we had to linearize this relation with a link function : root square. After this modification, all of the residuals plots occured to be fine. The Q-Q plot shows that the distribution of the residuals fitted with the distribution of the model, the plot of normality shows a close normality of the residuals and the homoscedasticity plot shows no structure and homoscedastical residuals.



We can finally compute the Linear Regression. To precise, our null hypothesis is the absence of effet for any of our explanatory variables on the variation of the response variable.

On this summary we can firstly check if the model is validated :

* R-squared (0.722) : which means that 72% of the distribution of our points is explained by our model. That is not bad !
* Prob of F test (2.55e-240) : indicates that we can reject the null hypothesis.
* Each t test shows us a significant correlation of each variable with PRCP excepted for the Winter parameter which exceeded a probability of 0.05.

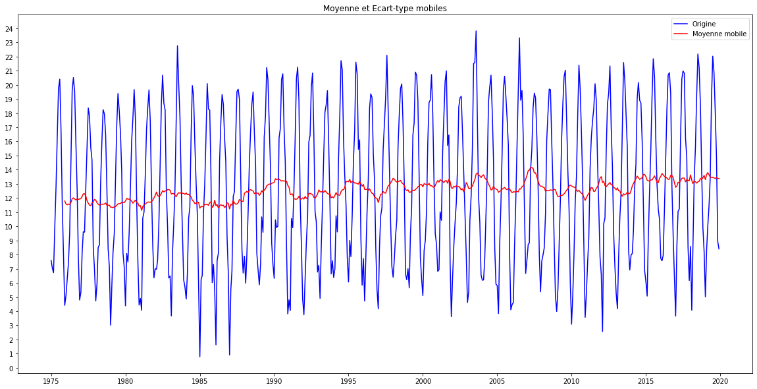
Finally, we can interpret :

* The monthly precipitation looks to follow a gradient according to the seasonality. Thus, autumn is the moister season followed by spring and finally by summer (winter is excluded of the interpretations).
* The variation of the location looks to be interestingly correlated with the precipitation. Cities can be classified from moister to drier with :

Brest > St Girons > Caen > Lyon > Marseilles

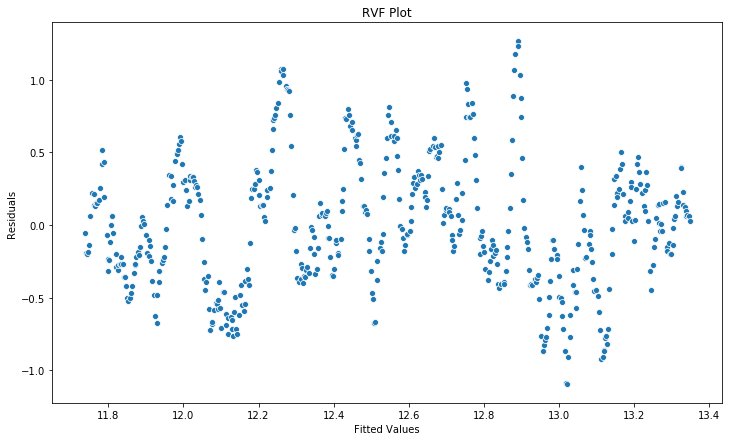
* The mean temperature (TAVG) has a weak positive impact on PRCP. When the temperature increases by 1, PRCP increases by 0.0140.
* The EMXP is strongly positivly correlated with PRCP like it was expected on the ACP. When EMXP increases by 1, PRCP increases by 0.7662.

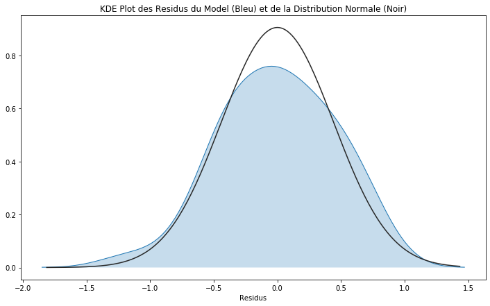
**4\_Temperature :**

In this part, our question was : How has the mean temperature evolved since 1975 ?

We began to get a visual idea of our TAVG by plotting moving average. It looks like it has a real positive trend with time.

Now, it’s time to test it but there is still the effect of seasonality. We must deseasonalize with the function seasonal\_decompose of statsmodel and only get the trend.

This way, we can build our linear regression where the response variable is TAVG and the explanatory variable is the time that ran from year 1975 (created on the base of the variable DATE). Here, TAVG follows a normal distribution. Before computing, like allways, we need to test the distribution of our residuals and it doesn’t look good ! In fact, there is a clear pattern on the homoscedasticity test due to a remaining effect of the season. Because seasonality is too disabling we can’t work with it. We must annualize.

Now that we annualized our data we can test the residuals. They look homoscedastics but now the problem is about their normality. Indeed, residuals don’t really follow the normal distribution. But to be sure we will do a proper test by computing the Shapiro test. It appears that the p-value asssociated is 0.715 which is greatly higher than our 0.05 threshold. Hence, the null hypothesis is rejeceted which means that we truly have a normal distribution on our residuals.

Une image contenant texte

Description générée automatiquementFinally, we can compute the linear regression where the null hypothesis is the absence of correlation between the time and TAVG.

We can check the validity of the model with :

* R-squared (0.534) : it is not especially high but not suprising because we only tested one variable.
* Prob of F-test (1.18e-08) : it indicates that we can reject the null hypothesis.
* Prob of t-test (0.000) : it indicates that time is significativly correlated with TAVG.

Hence, we can observe that, in average, each year the mean temperature increases by 0.0364°C. That doesn’t look really important but when you calculate, since the last 45 years the temperature in France has increased by 1.638°C !

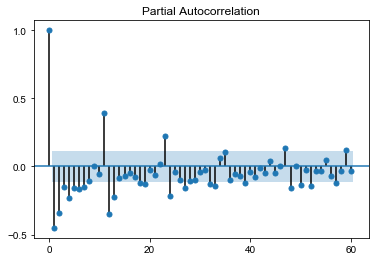
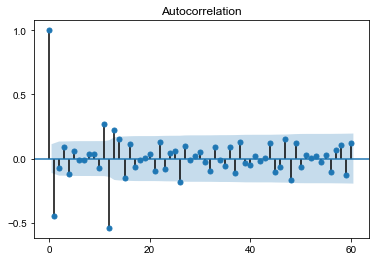
**5\_Prediction\_Temperature :**

The final project was about forecasting the mean temperature for the next few years. We chose randomly 36 months. To do so we built a SARIMA model on a new dataframe containing only TAVG and time.

Une image contenant texte

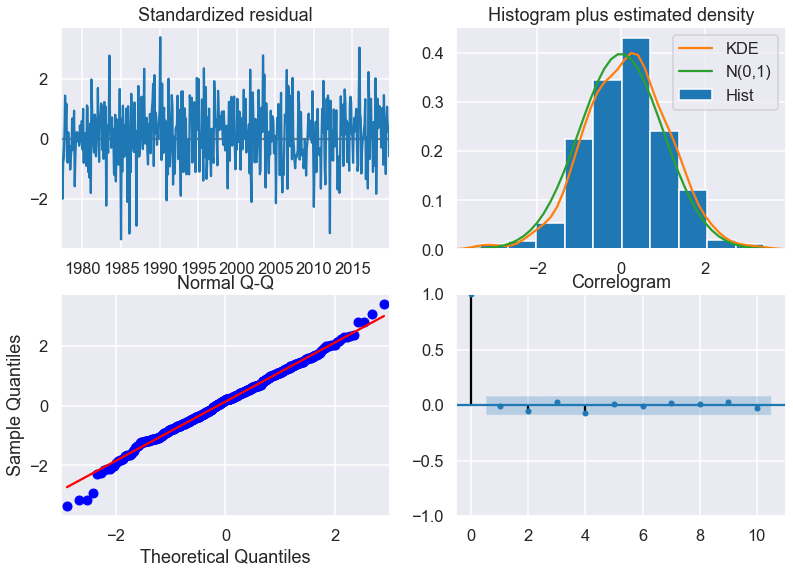
Description générée automatiquementBefore any computing we must know if our time series is stationary or not because only stationary data can be used to model forecastings. The moving average from the previous project showed us something not stationary but we need to be sure with an ADF test. The ADF value and the p-value indicate that we can say at 95% that our data are stationary.

The next step is to find the SARIMA terms from the ACF and PACF plots. In each one we indicated a saisonnality of 12 months :



Hence, we can attribute the value for each term :

* Auto Regressive (p) = 1
* Order of Differencing (d) = 0
* Moving Average (q) = 1
* Seasonal AR (P) = 1
* Differencing Order (D) = 1
* Seasonal AM (Q) = 1

Then, we can build the SARIMA model and check the distribution of the residuals with the very usefull function plot\_diagnostics. This summarise all the assumptions about the residuals we must do. From left to right and top to bottom :

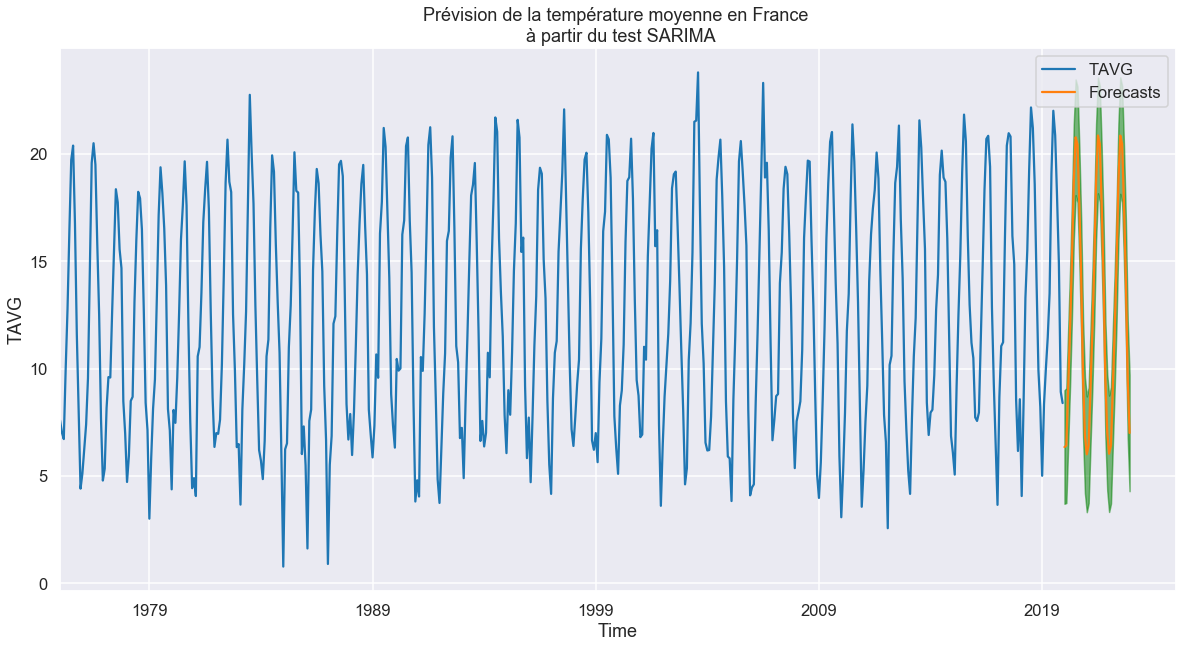
Residuals are independantly distributed around 0, they follow a normal distribution, they fit with the model distribution and there is no autocorrelation which is not explained by the model.

Une image contenant texte, noir, tableau de points

Description générée automatiquementAnd then we can compute the SARIMA model and checks its summary :

This summary gives us some precious information. We can see that none of the p-value of the t tests are higher than 0.05 except for the ar.S.L12 which is 0.068 (it depends on your jugement whether to keep it or not). This means that the lag of one month that we initiated (ar.L1) is a good predictor and that the seasonal lag of one month (ar.S.L12) is fine but not statistically good. The number of forecast errors of one month (ma.L1) and the seasonnal forecast error of one month (ma.S.12) look like pretty good preditors in this model. Finally, the model estimated the variance of the error with sigma2.

Now, let’s forecast !



The plot shows the 36 months forcastings in orange. Because the TAVG dosen’t move in a large range the forecast doesn’t show a clear pattern but the aim of the exercise was not the results, it was only for training new skills.

Thank you for your interest in this modeste project. If you have any questions, comments or suggestions about it don’t hesitate to contact me !