MOREL Tom

MAURY Florian

Projet énergie

Utilisation de la base de données Eco2mix du gestionnaire du Réseau de Transport d'Electricité (RTE)

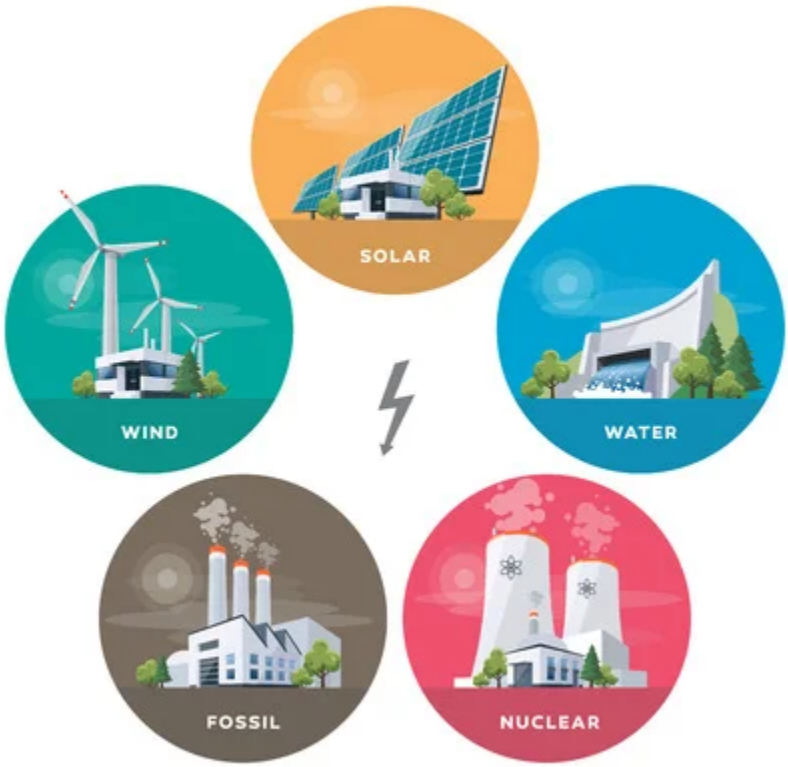


Table des matières

Objectifs :

1) Constater le phasage entre la consommation et la production énergétique au niveau national et au niveau régional.
2) Analyse au niveau national afin d'en déduire une prévision de consommation.
3) Analyse par filière de production : énergies renouvelables / nucléaire.
4) Focus sur les énergies renouvelables, à l'échelle nationale et régionale.

Problématique :

Observation entre consommation et production, en distinguant la part du nucléaire et la part des ENR dans le mix énergétique, au niveau régional et national, dans le temps depuis 2013.

Présentation de données :

Ce projet comporte 3 jeux de données différents: le principal nommé "éco2mix" qui est issu de l'application éco2mix, disponible sur le site internet du gestionnaire du réseau et transport de l'électricité en France. Et deux autres "régions" et "DJU" qui viennent respectivement des bases de données eurostat et du simulateur de calcul des degrés du site Cegibat.

- Base de données éco2mix

Ce jeu de données comporte des données régionales consolidées depuis janvier 2020 et définitives de janvier 2013 à décembre 2019 issues de l'application éCO2mix. Il est à noter que nous ne disposons pas des données pour l'année 2020. Le jeu de données contient différentes informations régionales françaises concernant la consommation d'énergie électrique, la production en fonction du type d'énergie selon le mix énergétique, les échanges physiques d'énergie entre régions ainsi que les taux de charge et de couverture pour chaque demi-heure depuis 2013 et en MégaWatt.

Informations relatives au jeu de données :

Pour répondre à nos objectifs et à notre problématique, nous avons décidé de ne pas utiliser certaines informations qui ne nous sont pas utiles, ce qui nous permet de réduire la très grande base de données proposée.

Nous avons retiré du jeu de données:

- Les indicateurs concernant le taux de charges et de couverture qui contiennent aucunes valeurs avant 2021
- La colonne “colonne 26” qui est vide
- Le code Insee région qui représente la même valeur que la variable Région
- La nature des données, car nous estimons les données toutes fiables.
- Les données de chaque demi-heure, ce qui réduit notre base de données de moitié (cela n’apporte pas d’informations en plus dans nos analyses que de garder le pas temporel ½ heure) .

De plus, nous avons créé deux variables “Mois” et “Année”, à partir de “Date” afin de faciliter l’observation de l’évolution des données par mois et années et depuis 2013, nous avons ainsi pu retirer la variable Date - Heure qui ne nous apporte pas plus que les variables Date et Heure séparées. Nous avons aussi mis en index la variable Date, puis changer le type de l’index en datetime pour pouvoir manipuler nos données en fonction du temps.

Nous avons choisi de remplacer les valeurs manquantes pour “Nucléaire (MW)” par 0, celles ci représentent les régions qui n’ont pas de production nucléaire, donc la production est égale à 0.
Nous avons créé trois variables concernant la production, la production totale “Production Total (MW)” comportant l’intégralité des moyens de production électrique en France, ainsi que deux variables concernant l’intégralité de la production électrique par des d'énergies renouvelables “Production Total ENR (MW)”, et l’intégralité de la production électrique par des énergies non-renouvelables “Production Total ENNR (MW)”.

site: Application éco2mix

- Base de données régions

Ce jeu de données contient les valeurs pour chaque région, la population, la superficie (Km²) et la densité (Population/Km²) par année, ainsi que la consommation et la production totale (nucléaire et d'énergies renouvelables) moyenne par an.

site: https://ec.europa.eu/eurostat/fr/web/regions/data/database

- Base de données DJU

Contient les DJU (degrés jour unifié) de chaque région entre 2013 et 2019. Le degré jour unifié (DJU) est la différence entre la température extérieure et une température de référence qui permet de réaliser des estimations de consommations d'énergie thermique pour maintenir un bâtiment confortable en proportion de la rigueur de l'hiver ou de la chaleur de l'été . La référence habituelle de 18 °C fut définie en considérant que la température intérieure des locaux est à 19 °C et que les apports gratuits internes (occupants, éclairage, équipements, etc.) et externes (rayonnement solaire…) couvrent l'équivalent de 1 °C de déperditions thermiques.

Un degré Jour est calculé à partir des températures météorologiques extrêmes du lieu de du jour J :

Tn: température minimale du jour J

Tx : température maximale du jour J

S: seuil de température de référence choisie (ici 18°)

Moy = (Tn + Tx)/2 : Température moyenne de la journée

Ainsi on a :

Si S<= Moy : DJ=0

Si S>Moy : DJ = S-Moy

Le DJU par régions est donc la somme de tous les DJ du mois

site : https://cegibat.grdf.fr/simulateur/calcul-dju

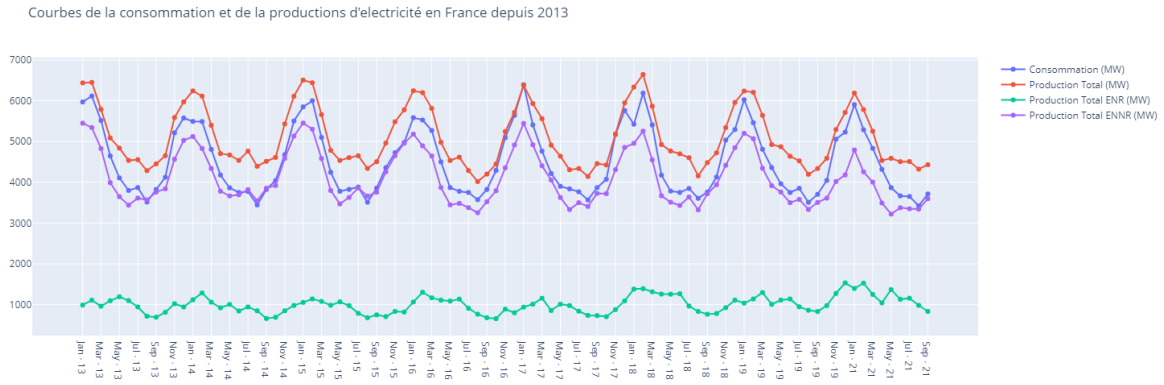
Bon à savoir concernant la distribution de l'électricité sur le réseau et la production électrique en France

- La production et la consommation d'électricité, dans un circuit électrique, doit être en « phase » ce qui signifie qu'il faut éviter la sous-tension et la surtension, afin de réduire le risque de « black_out » électrique, ce qui conduit à la panne sur tout le réseau électrique. En France ce type de panne peut être évité grâce à la coupure volontaire d'une zone en particulier, pendant quelque heure, le temps que d'autres centrales produisent de l'électricité pour que la production coïncide avec la consommation.
- Une énergie renouvelable est une énergie dont la source est inépuisable à l'échelle de l'humain.
- Une énergie non-renouvelable est une énergie qui dépend d'un stock à l'échelle de l'humain.
- Les énergies non-renouvelables que nous utilisons principalement (Nucléaire, gaz, pétrole, charbon) ont comme avantage, le fait qu'elles sont pilotables, ce qui signifie, que nous pouvons les actionner selon notre gré, selon la demande.
- Certaines énergies renouvelables que nous utilisons sont intermittentes, cela se traduit par l'absence de constance dans la production, le solaire dépend de l'ensoleillement, l'éolien dépend de l'intensité du vent.

Visualisations :

Les visualisations sont disponibles dans l'ordre dans le notebook Google Colab

- Courbes de la consommation et courbes de la production à l'échelle nationale, depuis 2013.

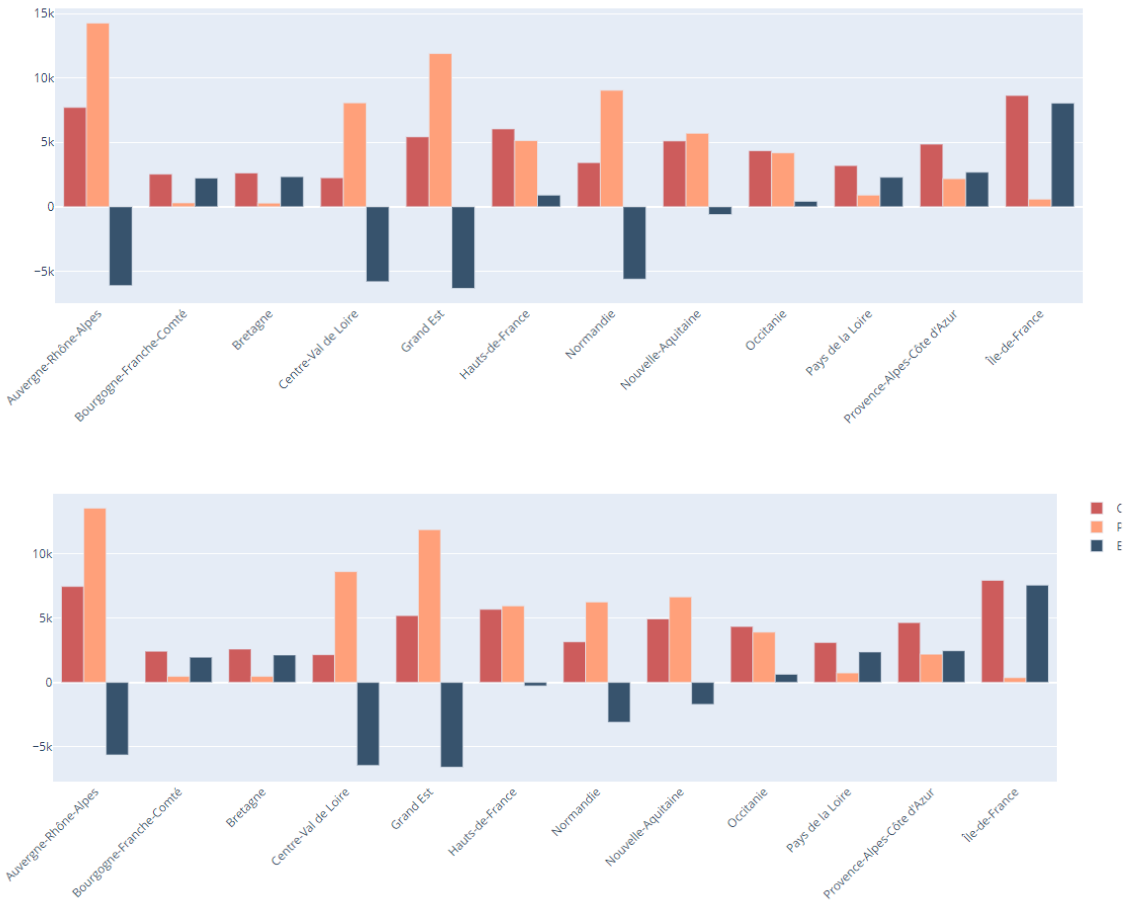


Grâce à cette visualisation de courbes à l'échelle nationale plusieurs éléments sont à souligner :
Les courbes de la consommations et de productions sont périodiques (saisonnalité de 12 mois), sinusoïdales, et synchronisées. On remarque ainsi une hausse de la consommation et production en hiver. La tendance paraît constante.
La courbe de consommation et de production totale sont phasées et de même échelle, la production se doit d'être toujours légèrement supérieure pour éviter un risque de sous production.
Les courbes de production d'électricité ENR et ENNR semblent disposer des mêmes caractéristiques, la part d'énergie non renouvelables est en revanche beaucoup plus importante et suit plus la saisonnalité des consommation et production totales.
L'observation de ces courbes permet d'avancer plusieurs conclusions. La France dispose d'une stratégie énergétique autonome et autosuffisante, dans le circuit électrique européen, elle peut se permettre d'être dans une posture d'exportation avec les pays voisins grâce à l'excédent de sa production.

La France dispose d'un mix énergétique qui n'a que très peu changé depuis 2013, avec une composante ENNR extrêmement élevée notamment lors des périodes de

forte consommation, en période hivernale. Le modèle de gouvernance énergétique français connaît une régularité et une stabilité.

- Graphe en barres de la consommation et courbe de la production à l'échelle régionale, depuis 2013. (1)



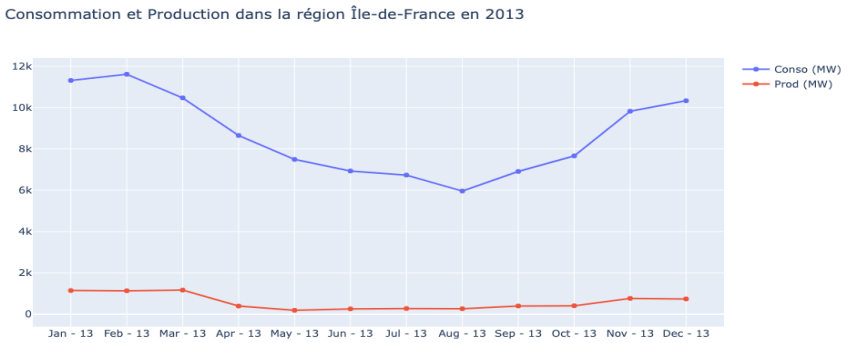
Grâce à la visualisation des deux graphiques diagramme à barres, plusieurs phénomènes sont visibles et interprétables (ici nous gardons seulement les graphes pour 2013 et 2021 pour faire une comparaisons à travers le temps):

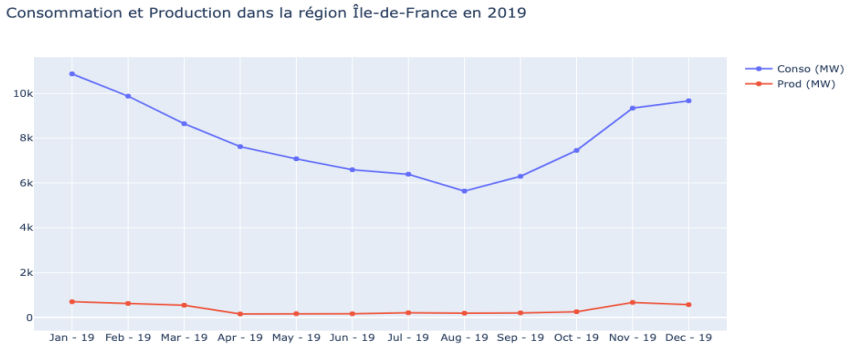
Les régions ne connaissent pas de phasage entre leur consommation et leur production en 2013. Certaines régions consomment plus qu'elles produisent ; l'Île de France, Provence-Alpes-Côte d'Azur, Pays de la Loire, Bourgogne Franche-Comté, et la Bretagne.

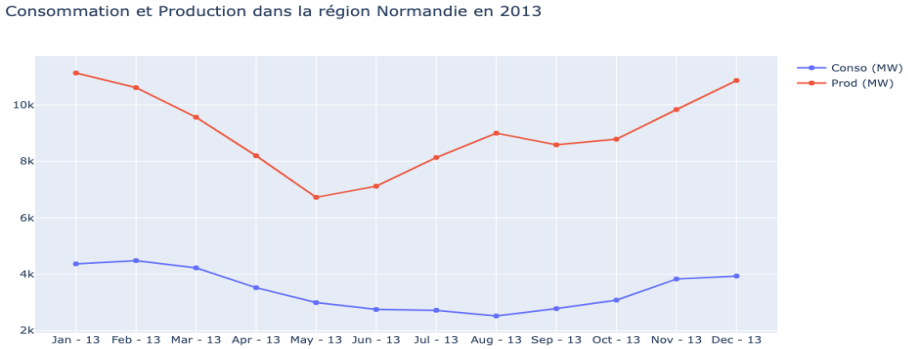
Pour d’autres régions il s’agit du phénomène inverse, une production plus élevée que la consommation ; Auvergne Rhône-Alpes, Centre Val de Loire, Grand-Est, Normandie.
Il existe une dernière catégorie de régions, celles qui ont une production et une consommation relativement équivalentes ; Hauts-de-France, Nouvelle-Aquitaine, et Occitanie.

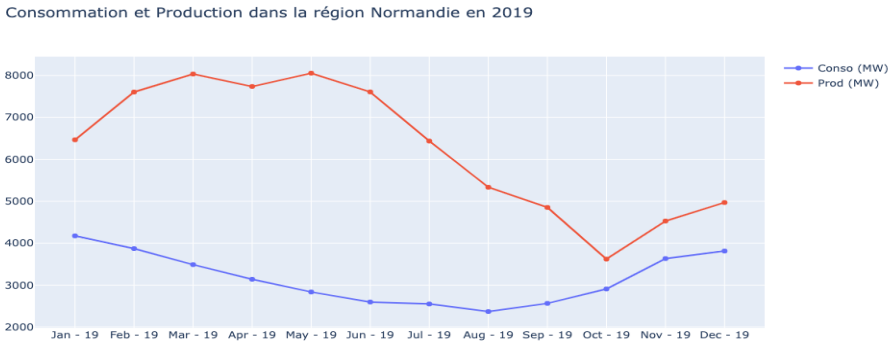
On peut remarquer que les régions, dans leur ensemble, ont presque toujours le même écart de consommation et de production entre 2013 et 2021, à l'exception de de la Normandie qui dispose d’un écart plus faible entre sa consommation et sa production, cela veut dire que les politique de distribution d'énergie dans les régions françaises n’a pas beaucoup changé depuis 2013. La répartition d'énergie en France se fait principalement entre régions limitrophes. Il existe, aussi, des échanges entre certaines régions françaises et étrangères (pays limitrophes à la France), la France produisant plus qu’elle ne consomme, l’exportation est plus forte que l'importation.

- Courbe de la production et de la consommation de chaque région en 2013 et 2019







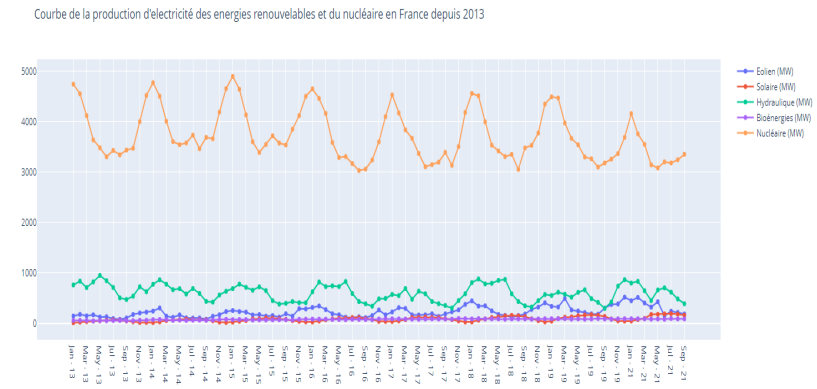


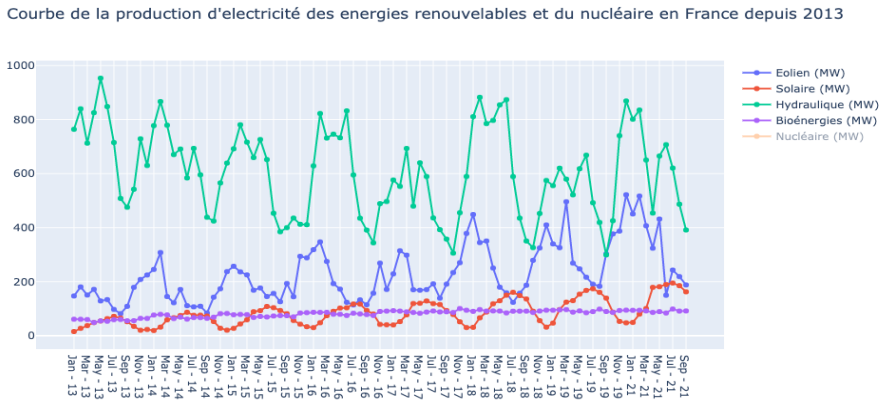
L'observation des quatre graphiques à l'échelle régionale (Île de France et Normandie) permet de voir les variations possibles qu'il peut y avoir à travers le temps. On peut constater le phénomène saisonnier de la consommation domestique régionale, qui semble réellement correspondre aux périodes hivernales. Cependant il est fort de constater que la production domestique régionale ne correspond en rien à un phénomène saisonnier contrairement à la production nationale. En effet, la

production va varier selon si la région produit ou non et le type énergie (si solaire nous aurons un pic en Été, voir région Nouvelle Aquitaine).

Le modèle de gouvernance énergétique français à l'échelle régionale révèle que les régions sont interconnectées au niveau de la plateforme électronique nationale. Certaines régions produisent plus que ce que leur consommation domestique nécessite pour fournir en électricité d'autres régions. Cela s'explique pour des raisons géographiques, démographiques, juridiques, et normes administratives.

- Courbes de production de l'énergie nucléaire et des différentes ENR, par région.



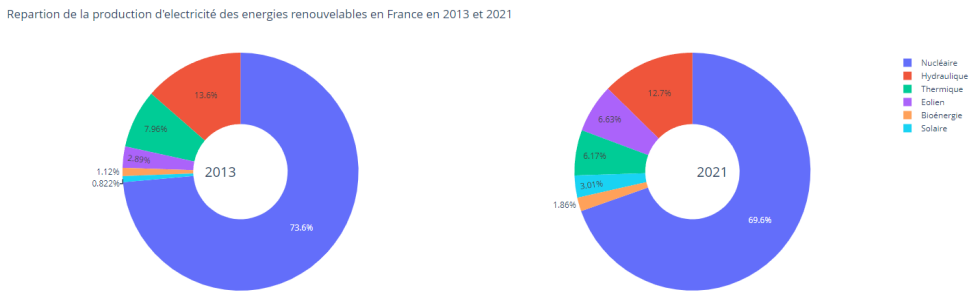


Grâce à la visualisation des courbes et des diagrammes représentant la production électrique du mix énergétique français depuis 2013, certaines composantes du mix énergétique connaissent des évolutions significatives.

Concernant la courbe de répartition de la production d'électricité en France, il apparaît que la production d'électricité d'origine nucléaire supplante les autres moyens de production d'électricité. Elle représente une puissance moyenne annuelle d'environ 3750 MégaWatts tandis que les autres moyens de production d'électricité, pris individuellement, ne dépassent pas le seuil des 1000 MégaWatts de puissance moyenne annuelle.

Les courbes de production moyenne annuelle concernant l'hydraulique, les bioénergies, le thermique semblent relativement stables. Celle concernant le nucléaire semble tendre vers une très faible décroissance, tandis que les courbes de production moyenne annuelle du solaire et de l'éolien semblent croître de manière plus significative. Les différentes productions suivent la même saisonnalité que la consommation, cela signifie que ce sont des énergies pilotables (même l'éolien et les bioénergies dans une moindre mesure) qui produisent lorsque l'on en a besoin. Seul le solaire est produit en grande majorité l'été.

- Camemberts de la production de l'énergie nucléaire et des différentes ENR en 2013 et 2021, en France

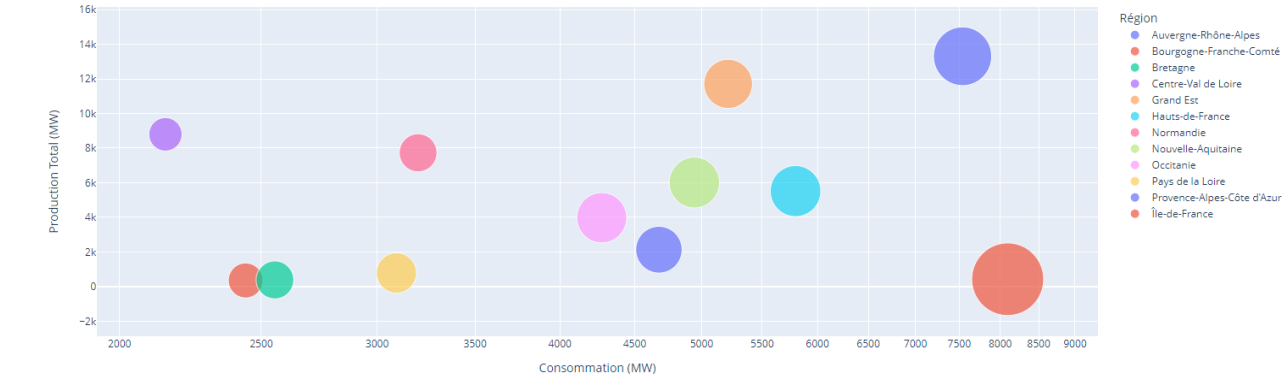


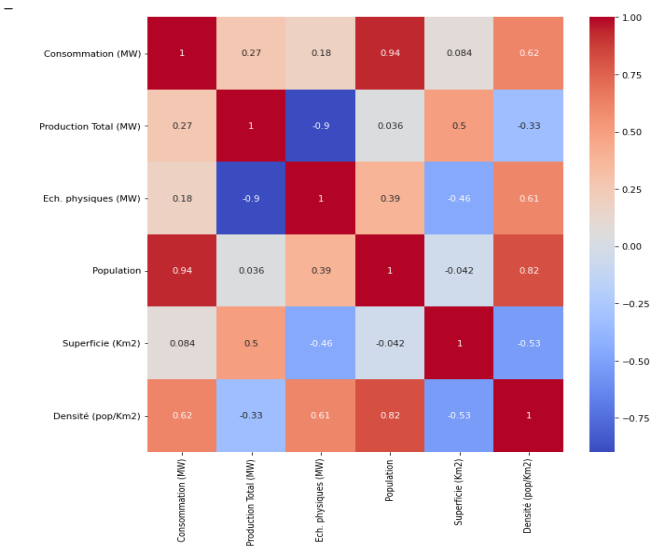
La visualisation des diagrammes représentant les parts de la production moyenne annuelle par type d'énergie, entre 2013 et 2021, permet de vérifier les phénomènes qui étaient remarquables dans le graphique précédent.

La part du nucléaire est passée de 73.56% en 2013 à 70.81% en 2021, la part du thermique est passée de 7.82% à 7.67%, la part des bioénergies est passée de 1.12% à 1.81%, la part de l'hydraulique est passée de 13.77% à 11.13%, la part du solaire est passée de 0.84% à 2.26%, et la part de l'éolien est passée de 2.89% à 6.32%.

Le solaire et l'éolien ont connu une croissance très élevée contrairement aux autres types de moyen de production électrique en France, cela s'explique sûrement par le développement de la politique tournée vers le développement des ENR, ceci dit, il est nécessaire de relativiser car les parts à l'échelle de la production nationale restent minimes notamment en comparaison avec la part de la production électrique d'origine nucléaire.

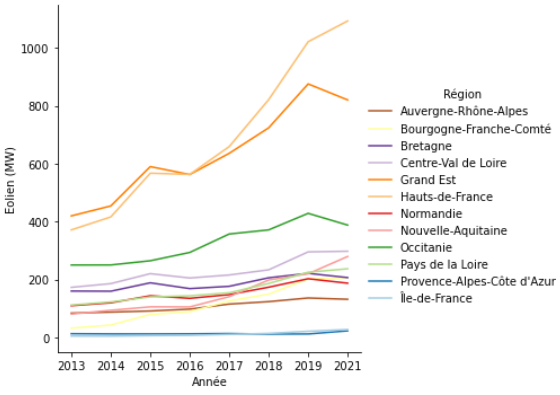
- Nuage de point de la production et la consommation électrique moyenne et de la population par région en France - heatmap de corrélation de variables du dataframe régions.

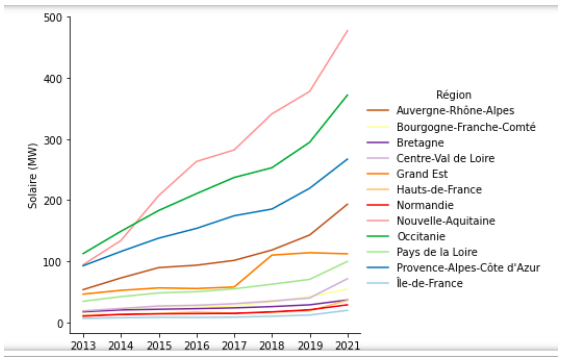




La visualisation du graphique et de la heatmap permettent de remarquer que les régions les plus peuplées sont celles qui consomment le plus et inversement. La taille de la population d'une région n'a aucune incidence sur la production de cette même région.

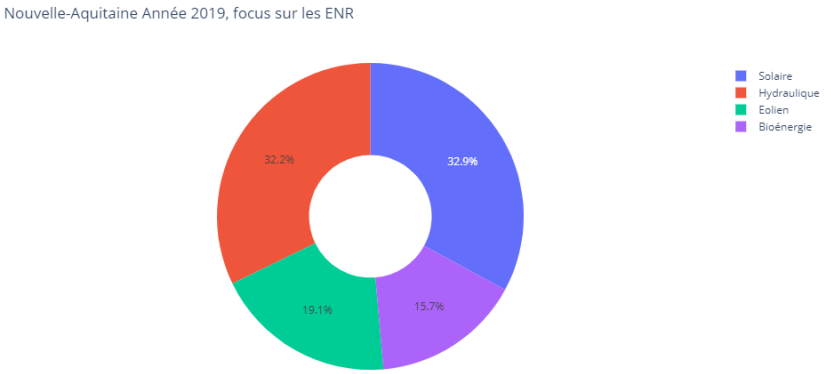
- Courbes de l'évolution la production de l'énergie éolienne/solaire/hydraulique pour chaque région depuis 2013

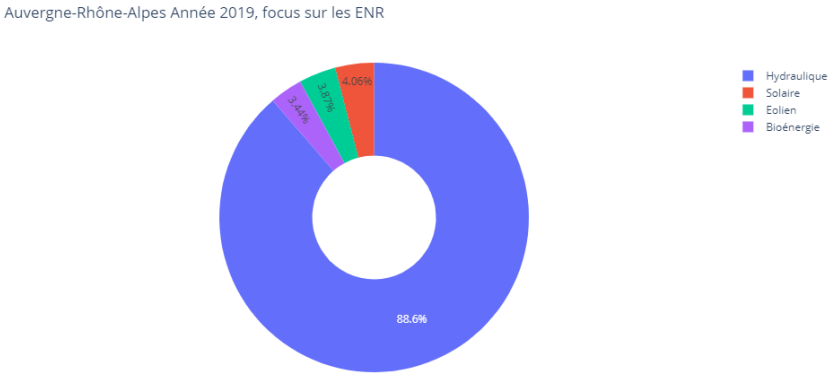


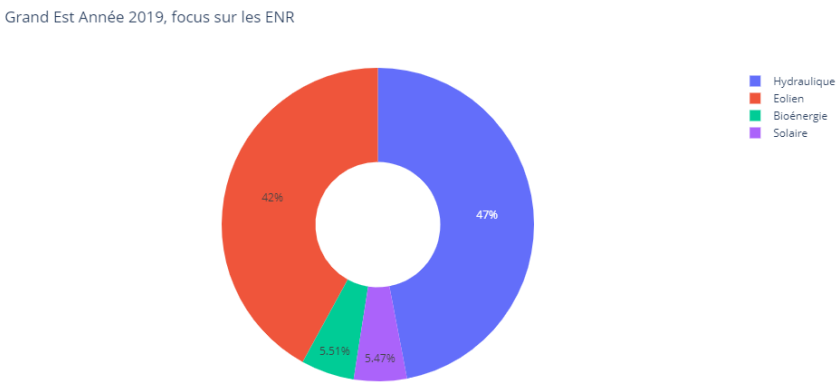


Chaque graphique représente une énergie, chaque courbe représente une région. Nous pouvons observer les différentes régions leader dans la production électrique éolienne, hydraulique, solaire, bioénergies et nucléaire. Les régions Grand-Est et Hauts de France sont celles qui produisent le plus d'énergies éoliennes par exemple.

- Camemberts sur la répartition des parts de la production des énergies renouvelables dans les régions française en 2019 :







Grâce à ces diagrammes camembert, nous pouvons remarquer que les régions ne disposent pas du même mix énergétique concernant les ENR. Certaines régions ont

des prédispositions favorables à l’accueil de certaines infrastructures pour des raisons géographiques, géomorphologiques, climatiques, d’aménagement du territoire, administratives, et politiques. De l'observation que nous pouvons faire de ces camemberts, il est remarquable de constater que la région Auvergne-Rhône-Alpes dispose principalement d’infrastructures hydrauliques. Cela s’explique sûrement par sa localisation au sein de deux massifs.

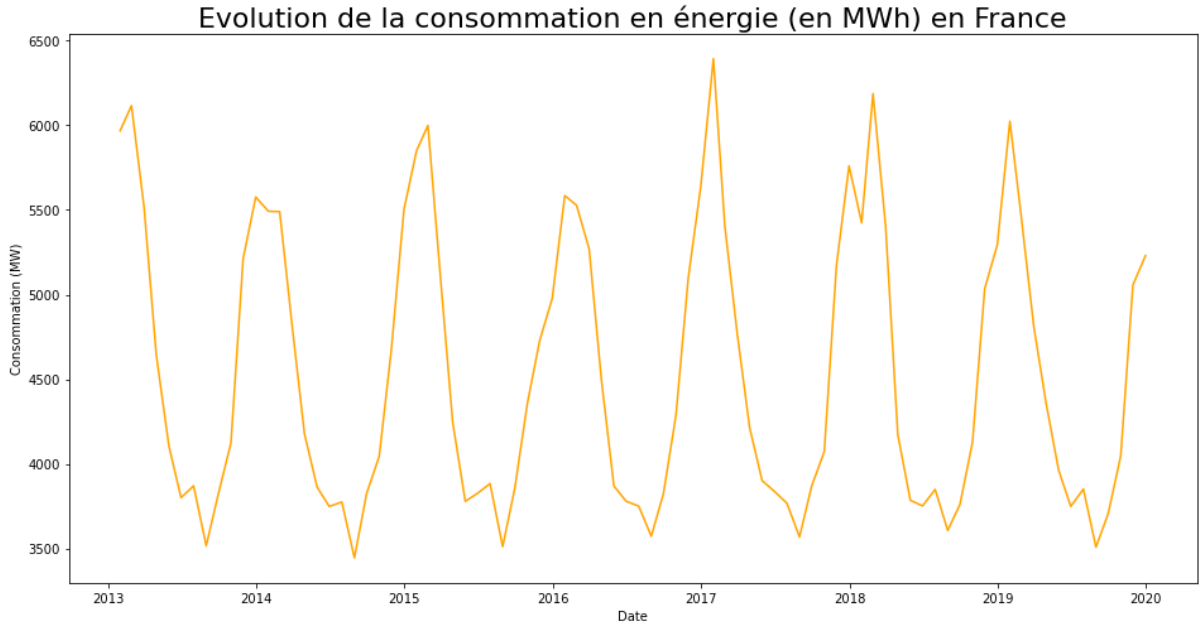
Modélisations :

Nous allons à présent tenter d'effectuer une prédiction de la consommation de l'année 2020 à l'échelle nationale grâce au modèle de séries temporelles Holt-Winters et SARIMA.

Nous allons ainsi effectuer plusieurs approches de traitement des Timeseries : description rapide des données utilisées, ajustement des effets de température sur la consommation en énergie via régression linéaire simple, désaisonnalisation par les moyennes mobiles pour obtenir les consommations corrigées des variations saisonnières (CVS), Holt-Winters et SARIMA avec la comparaison des calculs des métriques principales (MSE, MAE, RMSE, MAPE, R2), la prédiction de la consommation.

Tout d'abord, les dataframes que nous allons utiliser s'appellent *df_dju_france et df_conso_france*, ils comprennent respectivement les valeurs moyennes des DJU et la moyenne de consommation de chaque région par mois entre 2013 et 2019.

Regardons dans un premier temps la consommation nationale depuis 2013.



Comme nous l'avons vu dans la partie des visualisations, les saisonnalités sont très marquées sur ces données de consommation. Dans ce cas de modélisation, nous allons supposer que les consommations hivernales sont fortement impactées par

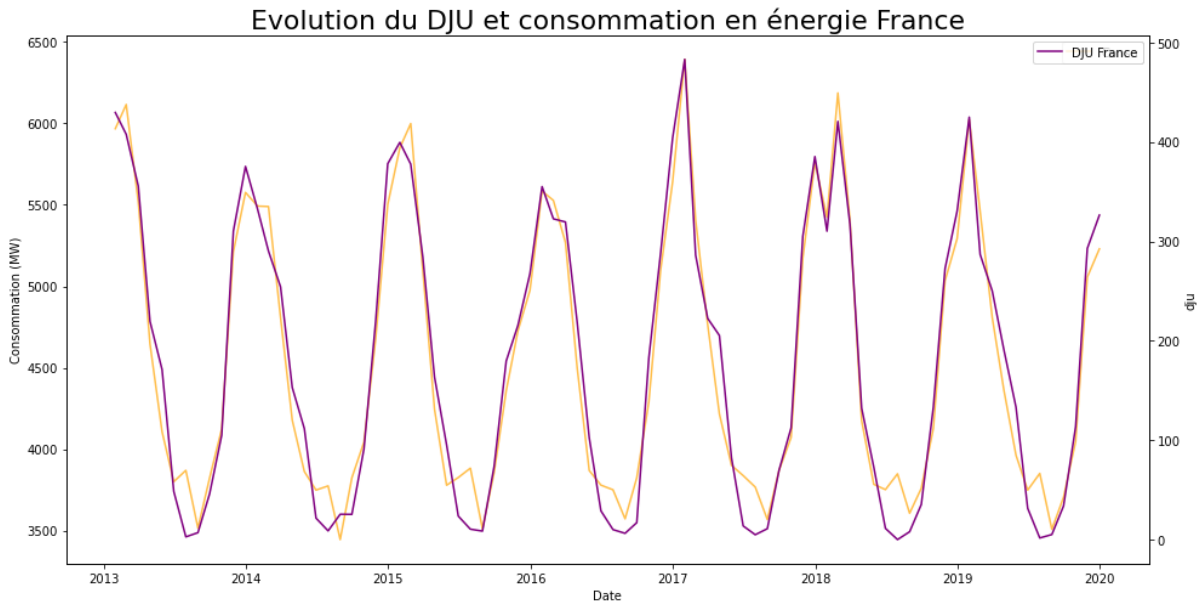
l'effet des chauffages électriques et d'un éclairage plus important. Afin de gommer cet effet, nous allons devoir corriger les consommations grâce aux DJU (Degrés Jour Unifiés).

Données météo utilisées pour corriger les données de l'effet température

On a cherché la moyenne des DJU par mois pour chaque chef lieux de régions (on a estimé que le moyenne DJU de la région pouvait être représentée par la valeur DJU de son chef lieu et que cela amènerait à des analyses fiables).

Nous allons ensuite réaliser un groupage par la moyenne pour obtenir les données lissées en France. Nous pourrions également enrichir encore ce jeu de données avec d'autres stations pour avoir un estimateur encore plus précis ou traiter les algorithmes au niveau régional. Ici nous avons choisi de traiter les données nationales.

En projetant l'évolution de ces données DJU comparativement aux données de consommation d'énergie, les 2 courbes semblent suivre la même saisonnalité ce qui indique bien que les variations de la courbe de consommation d'énergie sont effectivement globalement liées aux écarts de températures. Il est temps de supprimer cet effet.

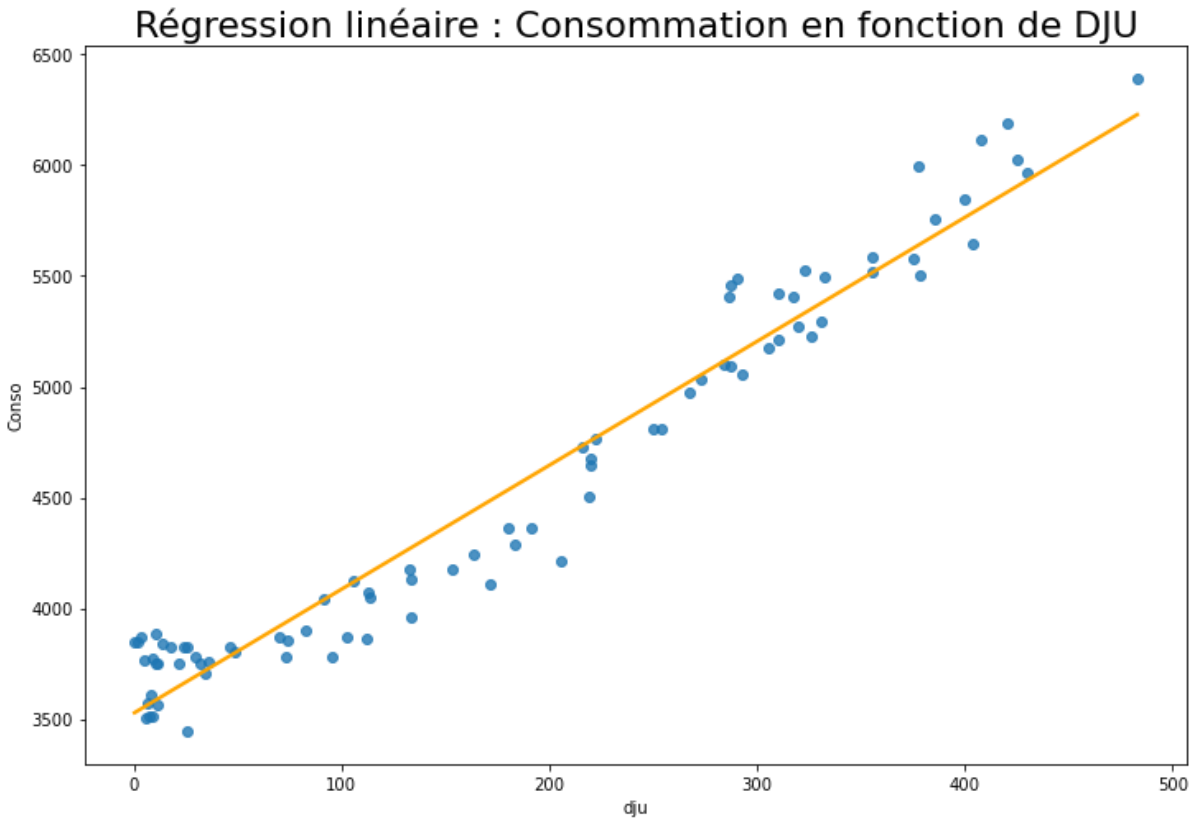


Pour traiter les séries temporelles, nous allons travailler sur les données France. Nous créons donc un dataframe global pour analyse nommé “datas” qui regroupe les deux précédents dataframe.
Pour corriger les données de consommation mensuelles de l'effet température *(dues au chauffage électrique notamment),* nous allons utiliser une régression linéaire. Regardons dans un premier temps la régression linéaire correspondant à :

- X = dju
- Y = consommation totale

Dans ce modèle, nous supposons qu'il existe une relation linéaire entre la variable à expliquer et la variable explicative : $Y=a+bX+\varepsilon$

Nous cherchons donc les paramètres inconnus a et b pour corriger les consommations mensuelles de l'effet de température.



On constate clairement une relation linéaire des variables DJU et Consommation.

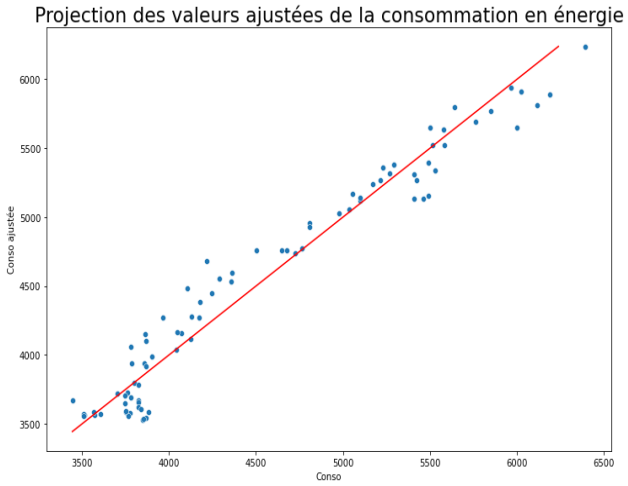
Réalisons à présent la régression linéaire simple grâce à la librairie Statsmodels :

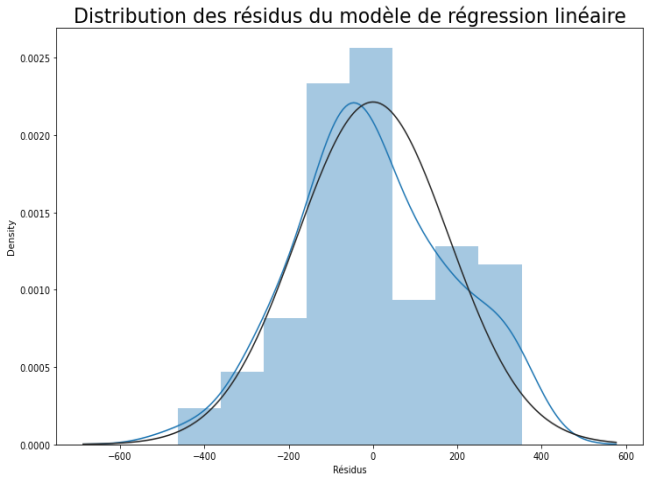
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Conso   R-squared:                       0.952
Model:                            OLS   Adj. R-squared:                  0.951
Method:                 Least Squares   F-statistic:                     1618.
Date:                Wed, 05 Jan 2022   Prob (F-statistic):           9.42e-56
Time:                        15:05:19   Log-Likelihood:                -555.48
No. Observations:                  84   AIC:                             1115.
Df Residuals:                      82   BIC:                             1120.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3526.4586     31.984    110.256      0.000    3462.832    3590.086
dju            5.6140      0.140     40.222      0.000       5.336       5.892
==============================================================================
Omnibus:                        0.452   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.798   Jarque-Bera (JB):                0.603
Skew:                           0.065   Prob(JB):                        0.740
Kurtosis:                       2.606   Cond. No.                         368.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
a = 3526.458613645082 | b = 5.614041533356094
```

Ici, le R2 *(coefficient de détermination)* est de l'ordre de 0.95, ce qui est relativement élevé et au vu de la représentation graphique de notre droite de régression, cela nous indique que le modèle est très bon.
La variable "dju" est également statistiquement significative au niveau de test 5%, sa P-value étant de 0.

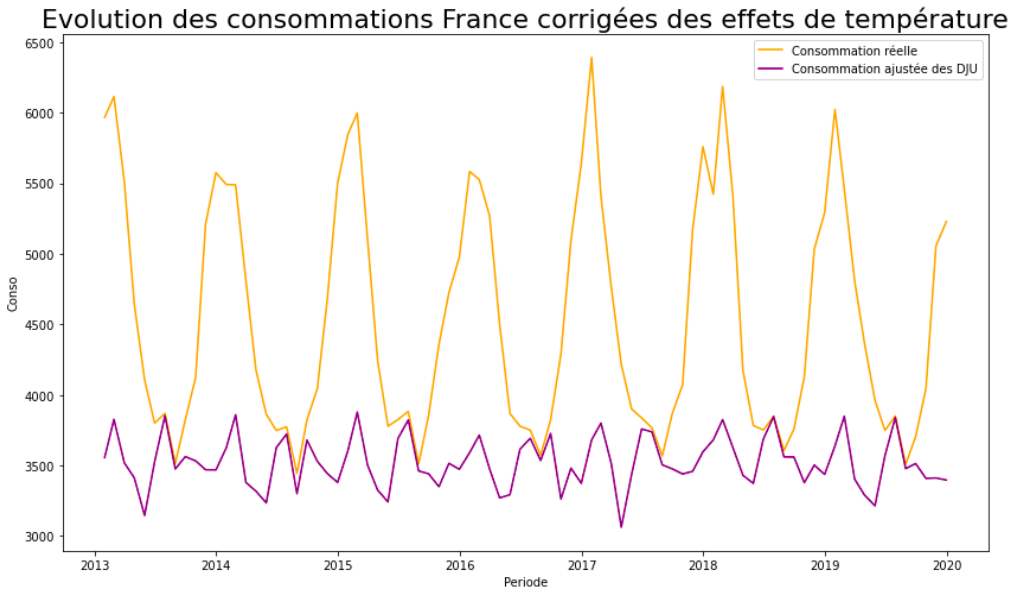
Pour vérifier la performance de notre modèle, nous allons regarder les valeurs ajustées en fonction des valeurs observées :
On remarque que les points projetés sont proches de la première bissectrice, ce qui nous indique que les valeurs ajustées sont proches des valeurs réelles. Notre modèle semble correct.
Enfin, nous allons représenter les résidus, ε étant un paramètre de notre modèle. Nous allons vérifier qu'ils sont bien centrés et de variance constante :
On remarque donc que la distribution des résidus semble centrée et suivent une loi gaussienne.





La P-value est ici non significative au niveau de test de shapiro(> 5%). Les résidus suivent donc une loi normale comme nous le constatons également sur le QQPlot.

Les tests sur notre modèles de régression linéaire étant significatifs, nous allons pouvoir utiliser les coefficients obtenus pour corriger notre consommation d'énergie des effets de la température : Conso Ajusté=$Conso-(dju*b)$

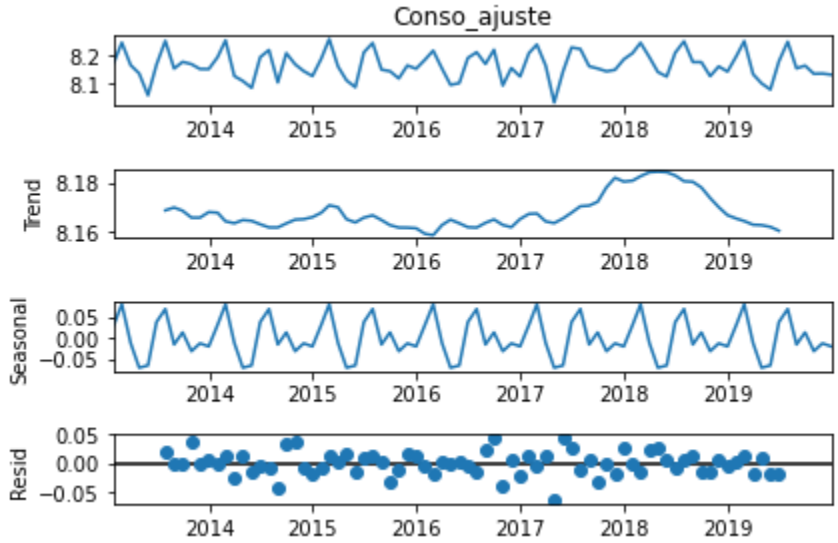


On voit, ici, clairement l'impact des températures sur les consommations d'énergie au niveau national. Il existe encore sur cette série temporelle corrigée un fort impact saisonnier.

Nous allons donc effectuer une désaisonnalisation de la consommation corrigée des effets de température par la méthode des moyennes mobiles.

Désaisonnalisation par moyenne mobiles

Une moyenne mobile est une combinaison linéaire d'instants passés et futurs de notre série temporelle. L'enjeu est de trouver une moyenne mobile qui laisse la tendance invariante, qui absorbe la saisonnalité et qui réduit le résidu :



On peut voir sur ce graphique combiné la décomposition de la tendance et de la saisonnalité ainsi que les résidus. Nous allons placer les valeurs de la décomposition dans un dataframe Pandas :
Les données CVS calculées et projetées nous permettent de voir également que nous n'avons pas d'outliers francs. Nous avons également remarqué 2 pics positifs et 1 pic négatif par période de 12 mois sur le graph de la saisonnalité.

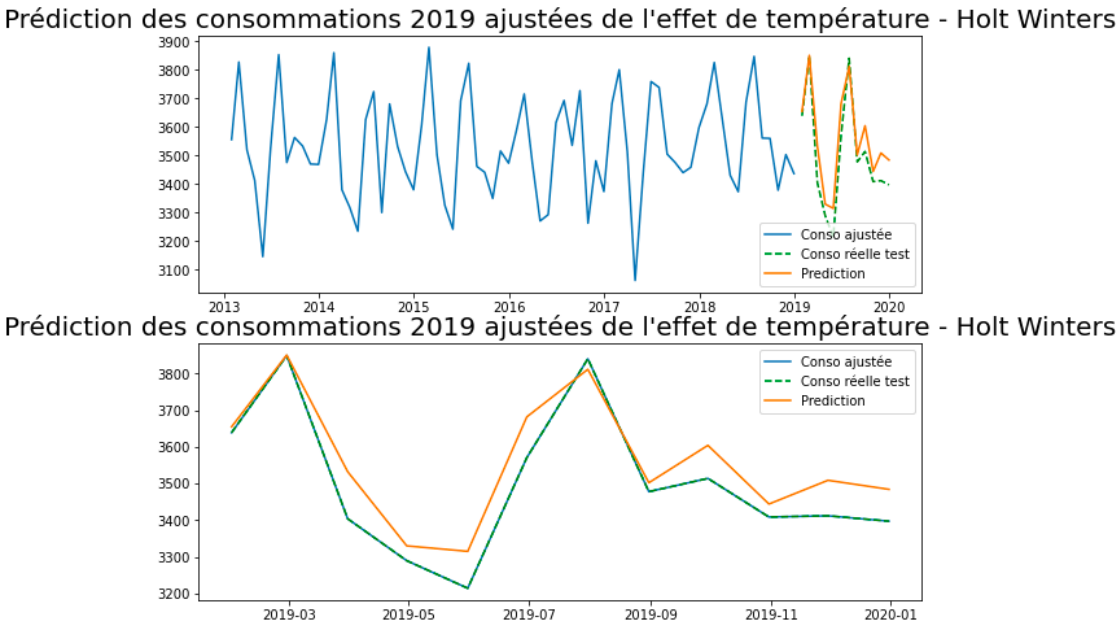
Toutes ces données vont pouvoir donner des points importants par exemple pour notre futur modèle SARIMA et ses paramètres p,d,q.

MODÉLISATION

Méthode de Holt-Winters (lissage exponentiel)¶

La méthode de Holt-Winters est une méthode de lissage exponentiel, basé sur les instants passés. Elle consiste à supposer que Xt est approximable au voisinage de T par aT+(t−T)bT+ST. En désignant par s la période du cycle saisonnier de la série temporelle.

Pour cette prévision, nous allons utiliser la série corrigée des effets de température, nous allons travailler sur la série de 2012 à 2018 *(split Train)* afin de tenter de prévoir l'année 2019 *(split Test)* pour comparer la prévision au données réelles. Nous utiliserons la fonction ExponentialSmoothing de Statsmodels.



La prédiction pour 2019 avec la méthode de Holt Winters est relativement fidèle aux données réelles. La saisonnalité et la tendance sont représentatives.

PRÉVISION SARIMA

Les modèles SARIMA *(Seasonnal AutoRegressive Integrated Moving Average)* permettent de modéliser des séries qui présentent une saisonnalité, comme c'est le cas pour notre dataset. Nous allons donc le comparer au modèle Holt-Winters

Dans un premier temps, nous allons tester la stationnarité de notre série avec le test ADF *(Augmented Dickey-Fuller)*. L'hypothèse nulle du test est que la série temporelle n'est pas stationnaire.

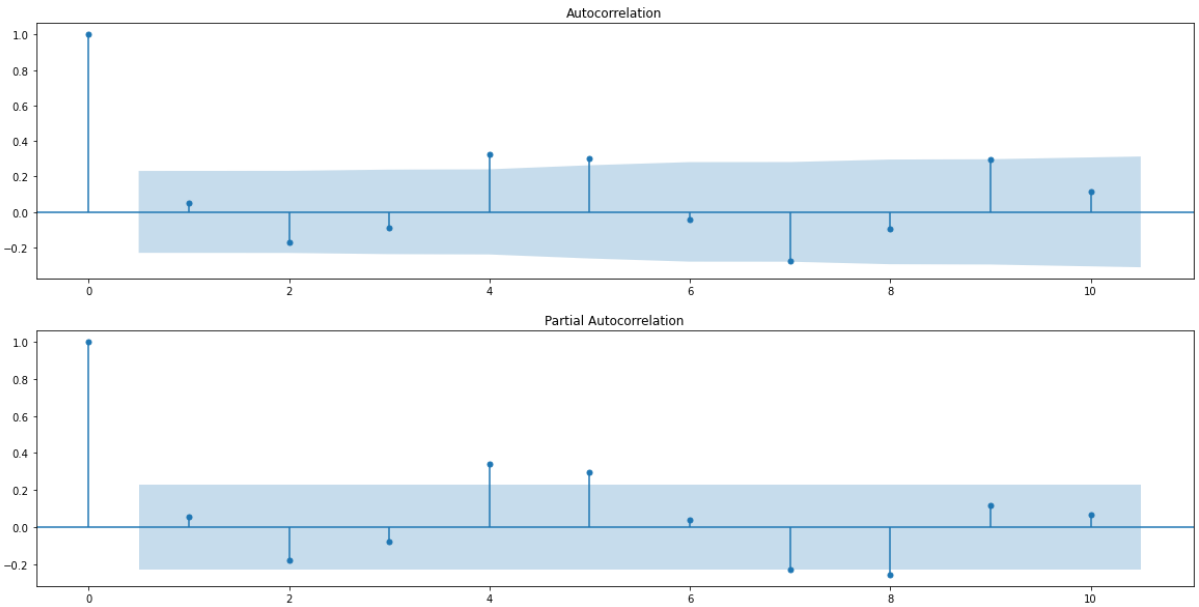
```
p-value ADF série originelle :0.49446144898449884
p-value ADF série différenciée :0.00036065069437072337
```

Au niveau de test 5%, on ne rejette pas l'hypothèse de non-stationnarité de la série, contrairement à la série différenciée. Nous allons donc devoir effectuer une stationnarisation.

Stationnarisation de la série

Nous allons réaliser cette stationnarisation par différenciation de manière itérative si besoin. Nous allons travailler sur le logarithme de la série.

Il existe un problème de stationnarité pour les multiples de 12 selon les autocorrélogrammes. Nous allons donc effectuer une différenciation d'ordre 12 sur cette série différenciée.



Estimation et validation des modèles SARIMA¶

Nous allons créer notre premier modèle SARIMA avec les paramètres estimés ci-dessus. Nous testerons la blancheur des résidus grâce au test de Jung-Box :

```
ABNORMAL_TERMINATION_IN_LNSRCH
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                   84
Model:             SARIMAX(1, 0, 1)x(1, 0, 1, 12)   Log Likelihood                 135.586
Date:                            Tue, 04 Jan 2022   AIC                           -261.171
Time:                                    12:07:57   BIC                           -249.017
Sample:                                         0   HQIC                          -256.286
                                              - 84
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9999   8.12e-05   1.23e+04      0.000       1.000       1.000
ma.L1         -0.1750      0.000  -1554.306      0.000      -0.175      -0.175
ar.S.L12       1.0000   2.72e-05   3.67e+04      0.000       1.000       1.000
ma.S.L12      -0.9630      0.001  -1354.962      0.000      -0.964      -0.962
sigma2         0.0009   5.44e-05     15.934      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                  12.29   Jarque-Bera (JB):                30.10
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.17   Skew:                             0.30
Prob(H) (two-sided):                  0.68   Kurtosis:                         5.87
===================================================================================
```

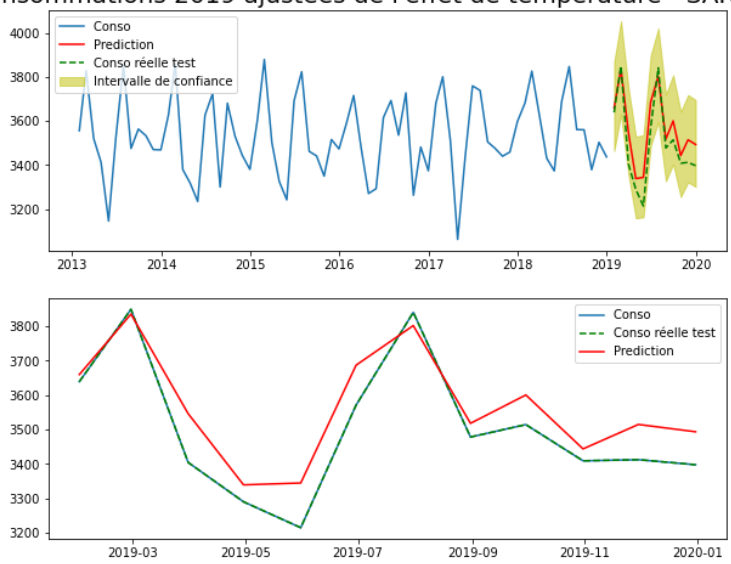
Les tests de significativité des paramètres et de blancheur du résidu sont validés au niveau 5%. De plus, sur les graphiques ACF et PACF des résidus, il n'y a pas de pics fortement significatifs mais il existe tout de même des variations

On rermarque donc ici que les résidus sont bien un bruit blanc et leur normalité est également validée. Le modèle semble donc plutôt performant.

Analyse

Nous allons à nouveau utiliser le split de notre série temporelle pour obtenir un jeu de test et un jeu d'entraînement. Nous pourrons ensuite estimer notre prévision comparativement aux données réelles.

Prédiction des consommations 2019 ajustées de l'effet de température - SARIMA(1,0,1)(1,0,1,12)



Performance des modèles:

A présent, regardons les métriques principales des modèles qui seront des indicateurs de performance et serviront de comparaison entre eux :

MAE (Mean Absolute Error) : Mesure la déviation absolue moyenne entre une estimation prévue et les données réelles.

MSE (Mean Squared Error): La distance, entre la prévision et l'observation, est ici élevée au carré. La sensibilité à l'erreur est meilleure.

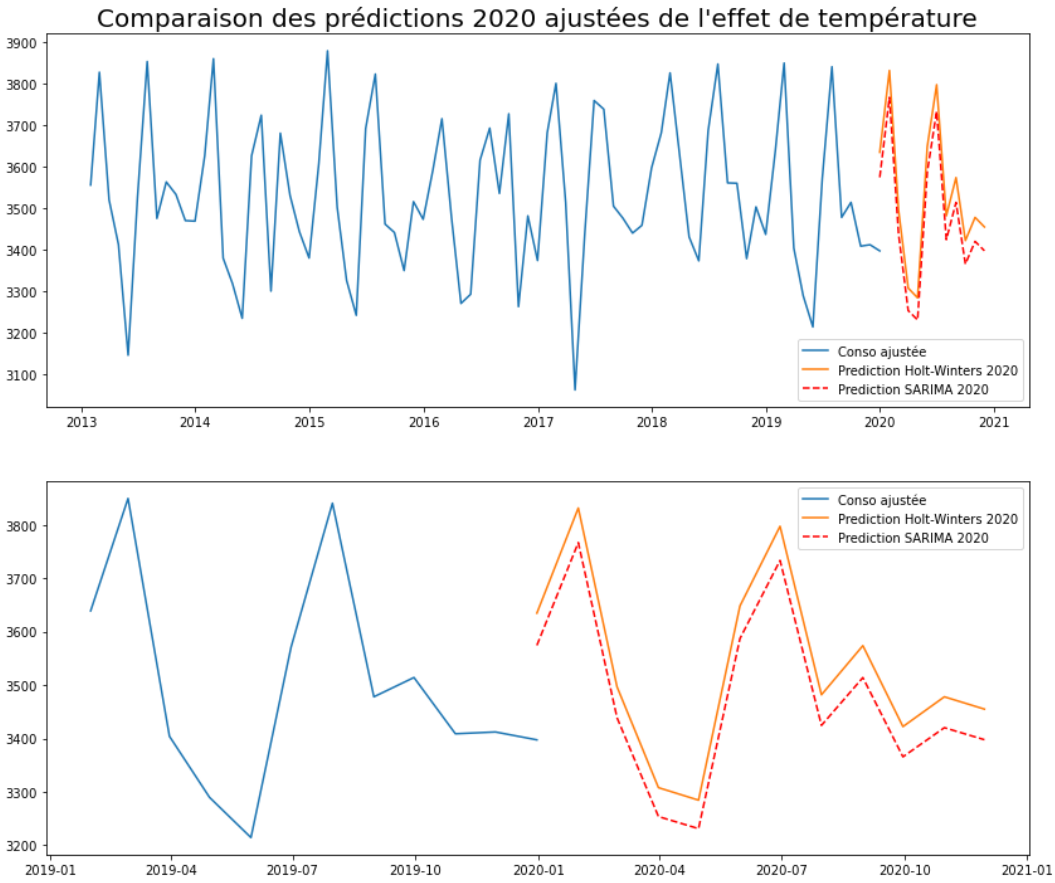
RMSE (Root Mean Squared Error) : C'est la racine carrée du MSE, c'est une métrique largement utilisée.

MAPE (Mean Absolute Percentage Error) : Moyenne des écarts en valeur absolue par rapport aux valeurs observées exprimée en pourcentageComparons les métriques du modèle SARIMA et Holt-Winter :

	Métrique	Résultats HW	Résultats SARIMA
0	MAE	63.397716	72.446820
1	MSE	5724.987532	7078.513574
2	RMSE	75.663647	84.133903
3	MAPE	1.848335	2.111035
4	R^2	0.838343	0.800124

Le modèle Holt-Winters est légèrement mieux que le SARIMA sur ce cas de modélisation, un R2 légèrement plus important et des écarts de variance (RMSE) plus faibles.

Prévision 2020:



Modèle de prédictions avec les DJU (régression linéaire):

Ici nous allons faire un modèle de régression linéaire avec comme variable à expliquer la consommation et comme variable explicative le DJU. Après avoir départager les jeux d'entraînement et de test, nous effectuons une régression sur ces variables, et cela nous amène à pouvoir prédire pour un DJU donnée, une estimation de la consommation moyenne.

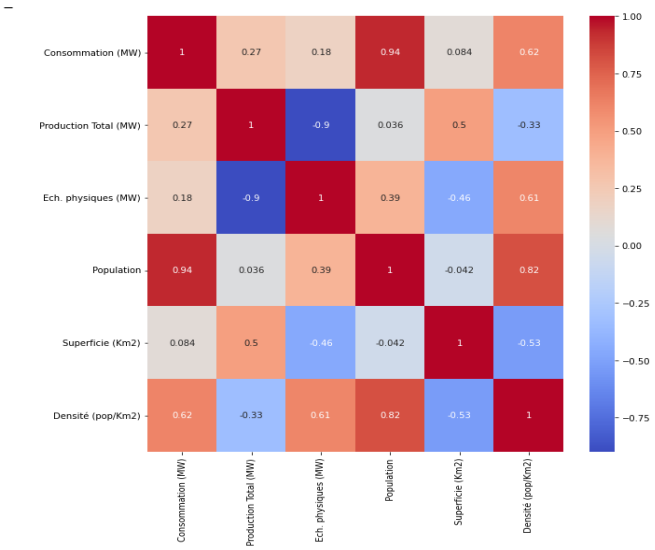
Les scores R2 et de validation des modèles d'entraînement sont très bons (0,95 et 0,92) et ne perdent que très peu sur les modèles de test (0,94 et 0,85). Bien sûr ce modèle ne prend en compte qu'une seule variable explicative ce qui le rend assez trivial.

On peut en conclure que la hausse de consommation de l'énergie en France en période hivernale est dû grandement au froid (hausse du chauffage, plus de lumière

car plus de nuits etc). En effet, on voit très clairement la forte corrélation et significativité des DJU sur la consommation. Le reste de la consommation varie entre 3200 et 3800 MW en moyenne et ceci est donc la partie non expliquée par les DJU. Cette estimation de la consommation est importante car elle permet entre autres de savoir ce que doit produire la France en énergie et notamment en hiver lors des pics de froid, pour éviter tout risque de blackout par exemple.

Modèle prédictif Population et consommation

Nous avions remarqué qu'il existait une forte corrélation entre la population et la consommation sur la heatmap :

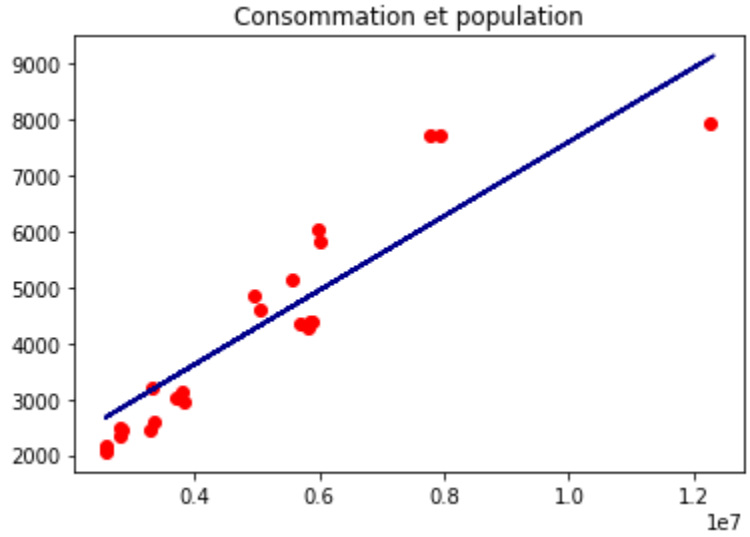


Nous allons réaliser un modèle de prédiction concernant la population et la consommation moyenne annuelle dans une région en France. En utilisant un modèle de régression linéaire avec scikit-learn.

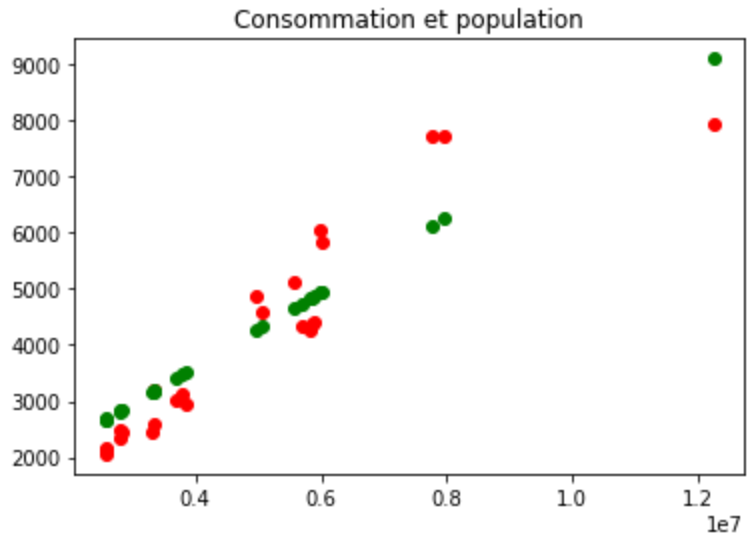
Le modèle habituel de régression linéaire simple est : $y=\beta 0+\beta 1x+\varepsilon y=\beta 0+\beta 1x+\varepsilon$

Nous allons dans un premier temps récupérer le data frame df région, afin de *nettoyer* pour garder et mieux visualiser les colonnes qui nous intéresse qui sont Population et Consommation (MW).

Nous réalisons un test_train_split où x est la valeur explicative (Population) et y la valeur cible (Consommation (MW). Puis nous réalisons une courbe correspondant au modèle prédictif.



Ici nous pouvons observer les différences entre les points prédits (en vert) et les points du test (en rouge).



Afin de vérifier la qualité du modèle nous réalisons un cross validation et un R^2, les résultats sont plutôt satisfaisants. Le modèle semble performant :

```
Coefficient de détermination du modèle d'entrinement : 0.8795738831029518
Coefficient de détermination obtenu par Cv d'entrainement : 0.8419353522525649
Coefficient de détermination du modèle de test: 0.8490592040629823
Coefficient de détermination obtenu par Cv : 0.7058724921693971
```

Nous pouvons dès à présent insérer un nombre concernant la population que l'on veut tester afin d'avoir une estimation de la consommation annuelle moyenne. Ce qui peut être utile pour la RTE pour l'ajustement de ses interfaces de distribution de l'électricité entre les régions de France en suivant l'évolution démographique de celles-ci.

```
#test valeur, à combien estimer la consommation moyenne annuelle pour une population de 9 000 000 dans une région

reg = reg.fit(x,y)
reg.predict(np.array(9000000).reshape(1,-1))
```

Conclusion générale

A travers ce projet, nous avons remarqué et expliqué le phasage entre la consommation et la production à l'échelle nationale. Nous avons analysé et observé la production électrique en fonction des différentes énergies à l'échelle nationale et régionale. Nous avons constaté l'évolution du mix énergétique français depuis 2013, ainsi que le mix énergétique des régions françaises en 2019. Nous avons réalisé deux modèles prédictifs ; l'un concernant la consommation l'année 2020, l'autre concernant la relation population-consommation.