Communication de résultat Hackathon

Data Analyst - Xuefei ZHANG
20 avril 2022

SOMMAIRE

1. Contexte du projet

Dans le but d'améliorer la résilience du modèle agricole face aux **aléas climatiques**, le Varenne agricole de l'eau et de l'adaptation au changement climatique a organisé un hackathon pour faire émerger les solutions de demain et **doter les agriculteurs d'outils d'anticipation et d'adaptation aux effets du changement climatique.**

Le Hackathon s'est déroulé du 3 au 5 décembre dans la région Drôme, avec au total 8 équipes composées d'informaticiens, codeurs, data scientists, étudiants, chercheurs, ingénieurs agronomes, météorologues, agriculteurs et vignerons,.

En tant que data analyst qui voudrais bien **mettre en pratique les compétences acquises** et de les **mettre au service des problématiques du monde réel**, je me suis intégrée dans le projet FLORAL qui noue le lien entre les données météo et données agricoles.

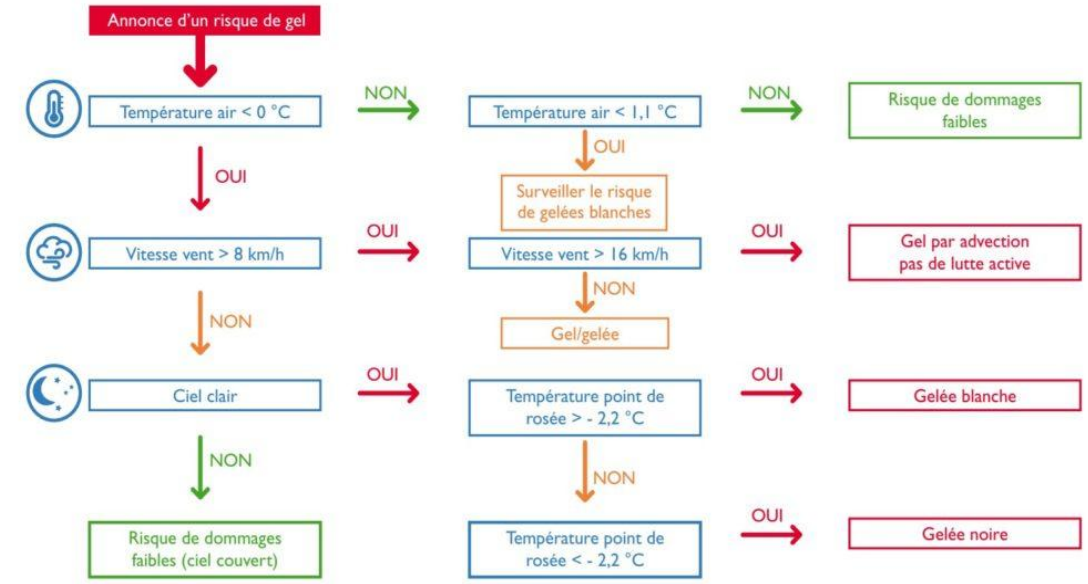
2. Projet FLORAL

2.1 Projet FLORAL, c'est quoi?

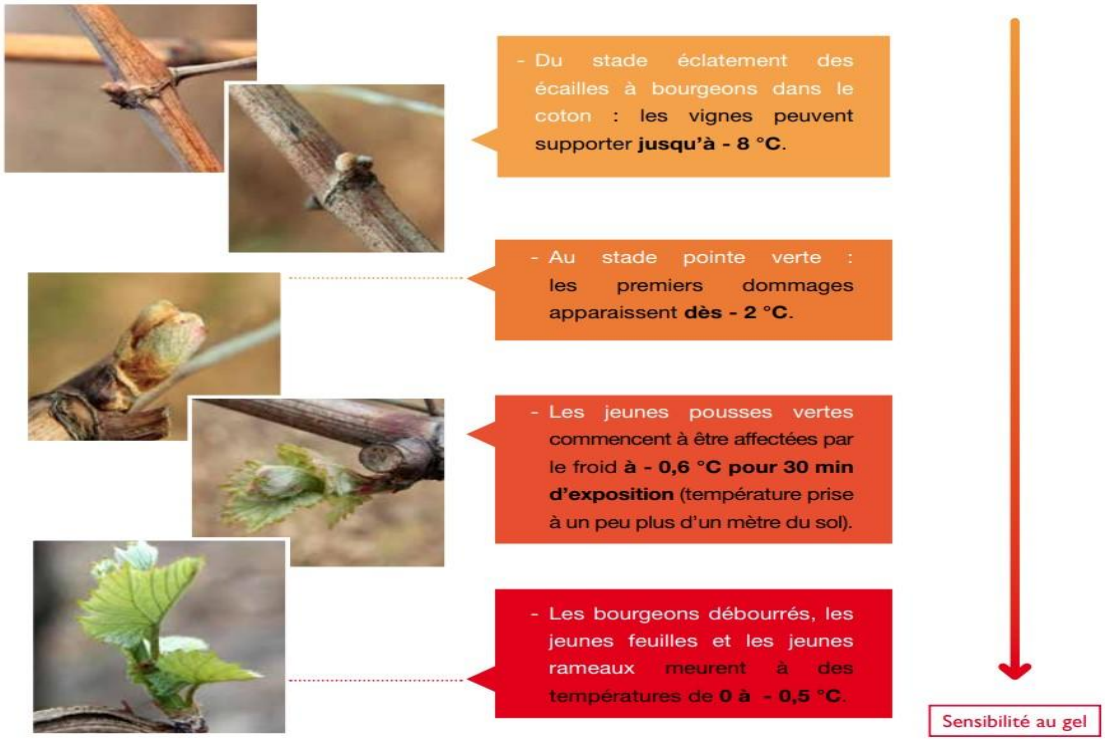
Projet FLORAL consiste à construire un outil permettant d'évaluer l'exposition au gel ainsi que le risque lié à cette exposition en fonction du stade phénologique observé sur la parcelle.

Floral permet à l'agriculteur d'obtenir un niveau de risques liés aux aléas climatiques pour les jours à venir, à partir du stade phénologique observé sur sa parcelle **géolocalisée** et des prévisions **météorologiques.**

Le fonctionnement de l'outil se résume en l'établissement d'un scoring permettant de prédire un niveau de risque pour la culture, en fonction du stade phénologique observé et des prévisions météo.



Clé de diagnostic de l'apparition et du type de gel – © Barclay Poling E, 2008



Sensibilité des organes végétaux de la vigne au gel

2.2 Constitution de l'équipe

Pour sortir une application ou siteweb à un certain usage, il necéssite des compétences de plusieurs disciplines: développement web et/ou application, designer UI UX etc.

Mais pour réaliser la prédiction - doter les agriculteurs des outils d'anticipation et d'adaptation aux effets du changement climatique, on a surement un pôle data solide. C'est aussi l'enjeu primaire qui assure la qualité de prédiction.

Le fait de sortir une solution opérationnelle pour un certain domaine ne se fait pas dans les nuages, il nous faut donc une connaissance du terrain, pour que la solution soit correpondante au besoin métier et conforme à la spécificité de vigne.

De plus mais pas moins important, un concert ne s'harmonise pas sans le chef d'orchestre. Donc un product manager est du besoin pour piloter le projet.

En résumé, dans l'équipe FLORAL, on a des compétences complètes visant à sortir une solution opérationnelle au bout de 48h.

Pôle data
A pour la mission de récupérer-nettoyer-procéder-analyser-modéliser des jeux de données de tout azimut: météo, vigne…. Et en sortir un algorithme opérationnel dédié au besoin spécifique du projet

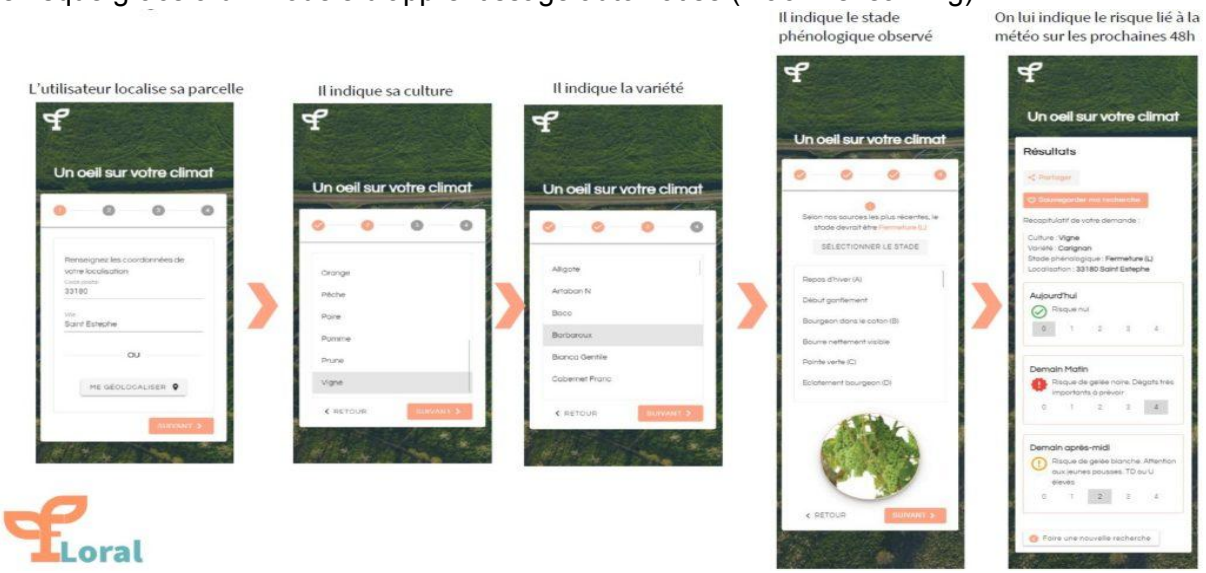
Pôle dev
consiste à développer tout ce qui concerne le siteweb, l'application et l'intégration de l'algorithme dans la solution

Pôle product management
a pour responsabilité de piloter le projet et orienter le développement du produit, y compris la schématisation et hiérarchisation de fonctionnalités, synchronisation de l'équipe etc.

Pôle Agri-consultant
est l'expert agronome qui donne ses conseils tout au long du projet pour que la solution soit conforme au besoin agricole.

2.3 Projet FLORAL - prototype de solution

A partir des coordonnées géographiques renseignées par l'utilisateur, une requête est adressée pour obtenir les données de prévision météorologique sur la zone afin de prédire le risque grâce à un modèle d'apprentissage automatisé (machine learning).



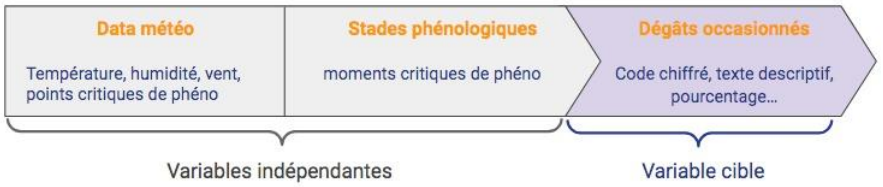
3. Réflexion sur les démarches et méthodologie

3.1 Réflexion: variables et sources de données

Les data qu'on doit prendre en compte dans la prise de décision - sources éventuelles:

- Météologique: température, pluie, force et vitesse de vent, humidité - MétéoFrance
- Agricole: culture, stade, variété de la culture, type de terre - Chambre d'agriculture de la Drôme
- Géographique: géolocalisation de parcelle, montagne, plateau, coline… - utilisateurs et système GPS

Suite à la réflexion et discussion en équipe, on a décidé de récupérer les données auprès de la Météo France et Chambre d'agriculture de la Drôme respectivement pour les données météologiques et agricoles.



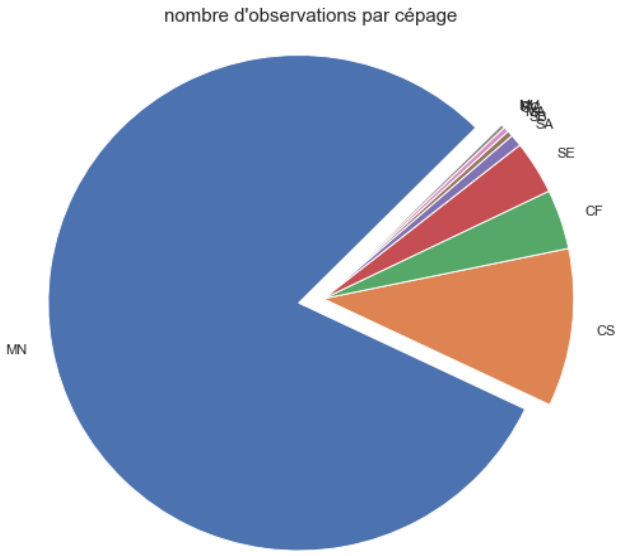
3.2 Réflexion: les data qu'on a et qu'on a besoin

On a des données météo et stades phénologiques, mais pour effectuer des prédictions, il faut encore des mesures de risques. A la rigueur, on cherche tout d'abord dans la base de données des vignes pour savoir si déjà les vignerons ont leurs mesures du métier qui pourraient être utilisées pour mesurer quantitativement les dégâts (variable cible).....

Finalement, on s'est rendu compte qu'il n'y a pas à nos jours une mesure conventionnée du métier. Autrement dit on n'a pas des statistiques qui pourraient servir à la variable cible.

4.1 Nettoyage et traitement de jeux de données

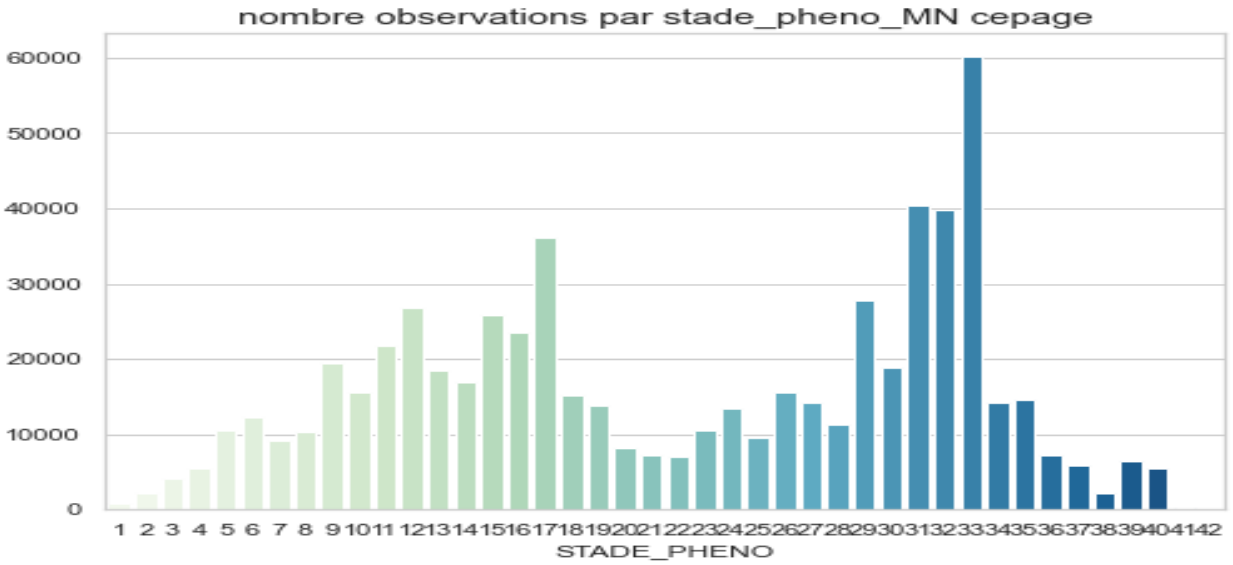
Data vigne



Cépage: variable qualitative nominale
Nombre d'observations: quantitative

Piechart:
Pour illustrer la proportion de divers types de cépages dans la région.

Le cépage MN représente environ 80% des observations. Autrement dit, pour la vigne de la région Drôme, le cépage MN est majoritaire.



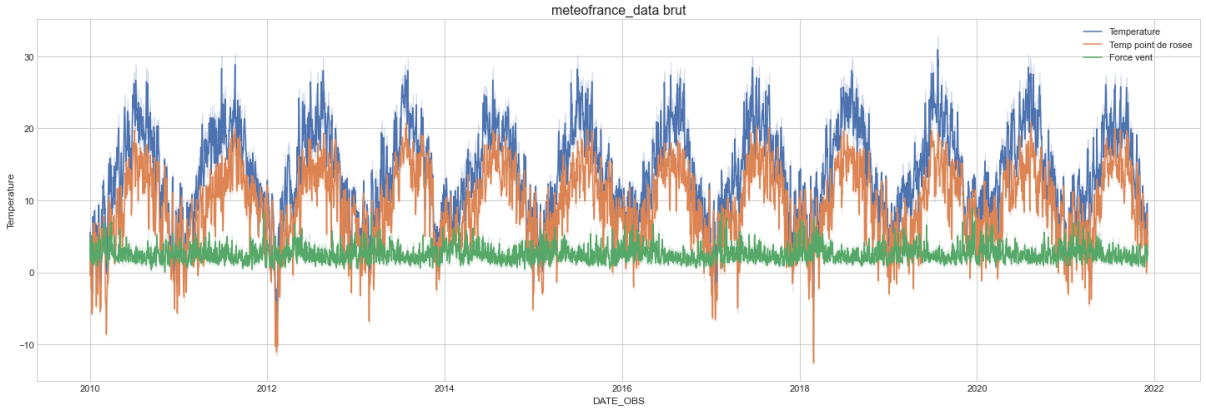
Stade_pheno: variable qualitative ordinale catégorielle
nombre d'observations: quantitative

Barchart pour illustrer le nombre d'observations par stade phénologique.

Pour le cépage MN, on constate d'après cet barchart que les nombres d'observations par stade se manifeste un caractère bi-modale: un pic se concentre entre les stades 11 et 17, l'autre est entre 31 et 33.

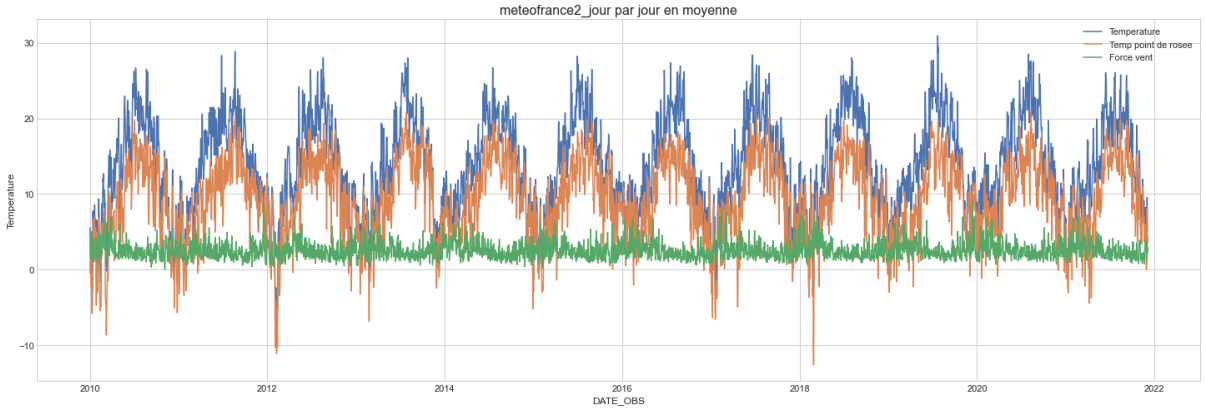
Data météo brut (meteofrance)

Linechart :
Évolution des 3 variables quantitatives en fonction du temps (séries temporelles)



On a récupéré un grand volume de data météo datant de 2010. Cela dit en terme de quantité et de couverture de la durée, on est sur le bon chemin.
On constate de ce lineplot que toutes ces 3 variables présentent leur saisonnalité à l'intervalle 12 mois.

Data météo en moyenne par jour (meteofrance2)



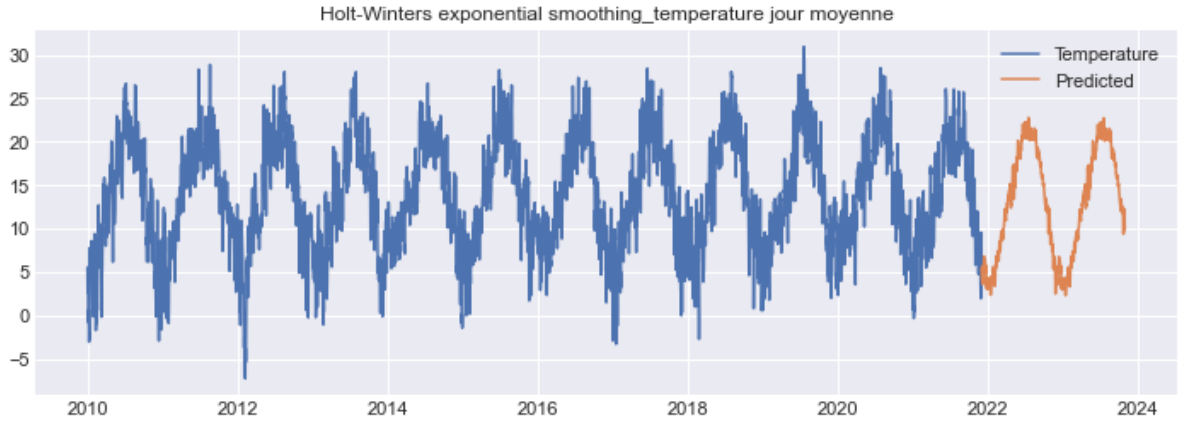
Meteofrance2 a de mêmes indicateurs: temperature, temp point de rosee et force vent. Et elle a pour chaque indicateur en espace d'un jour de multiples observations. D'ailleurs, le nombre d'observations varie selon les jours. Dans ce contexte, pour faciliter nos aproches, on a besoin, pour un jour, d'un seul point de données pour chaque indicateur. C'est ainsi qu'on a fait la moyenne des observations de chaque jour pour chaque indicateur. Ça donne un seul donnée pour chaque indicateur en espace d'un jour.

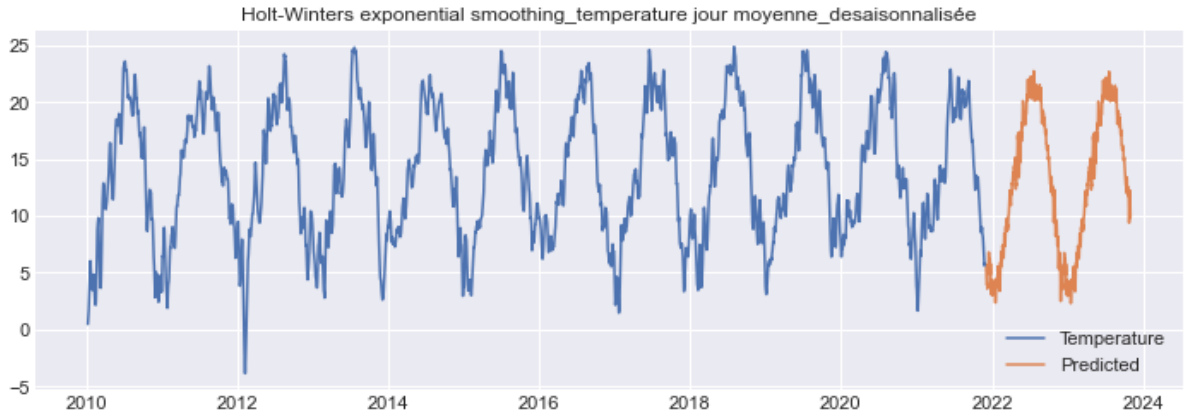
De plus, cette opération va faciliter ensuite la désaisonnalisation et la prédiction.

Ici on prendre la variable Temperature pour example et emploie la méthode de lissage exponentiel Holt-Winters pour effectuer la prédiction.

Prédiction (l'exemple Temperature moyenne/jour)

Ici on va prendre pour l'exemple la variable Temperature, et effectuer la prédiction repectivement sur les séries brutes avant désaisonnalisation et les séries désaisonnalisées.





Pour quelles raisons qu'on utilise cette méthode ?

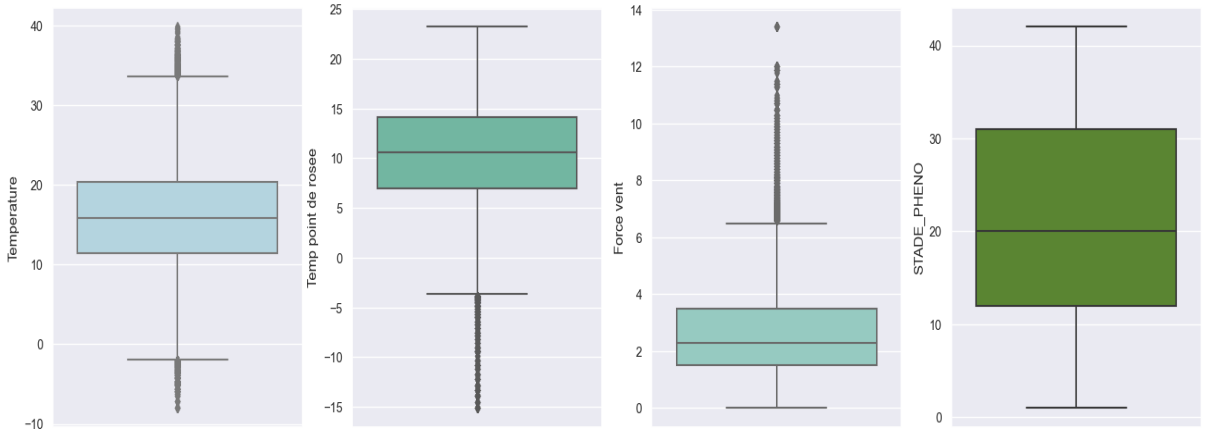
- Lissage exponentiel (Exponential Smoothing) attache plus d'importance sur les données récentes, ctd. elle est plus adaptée à à prédiction pour les changent vite et au fil du temps (météo, CA des start-up, un nouveau produit…) C'est ainsi qu'on utilise lissage exponentiel pour la prédiction de certains indicateurs en météo.
- D'ailleurs, on a dans notre cas les séries temporelles brutes et désaisonnalisées. Sachant que lissage exponentiel peut être appliquer à la fois à ces deux types de séries, on a bien raison d'opter pour cetter méthode bien inclusive.

Mais qu'est ce que c'est Holt-Winters?

Holt-Winters permet de prendre en compte dans sa méthode de prédiction le côté saisonnalité et côté tendance. Ça permet d'avoir une prédiction plus appropriée.

4.2 Analyse descriptive - univariée

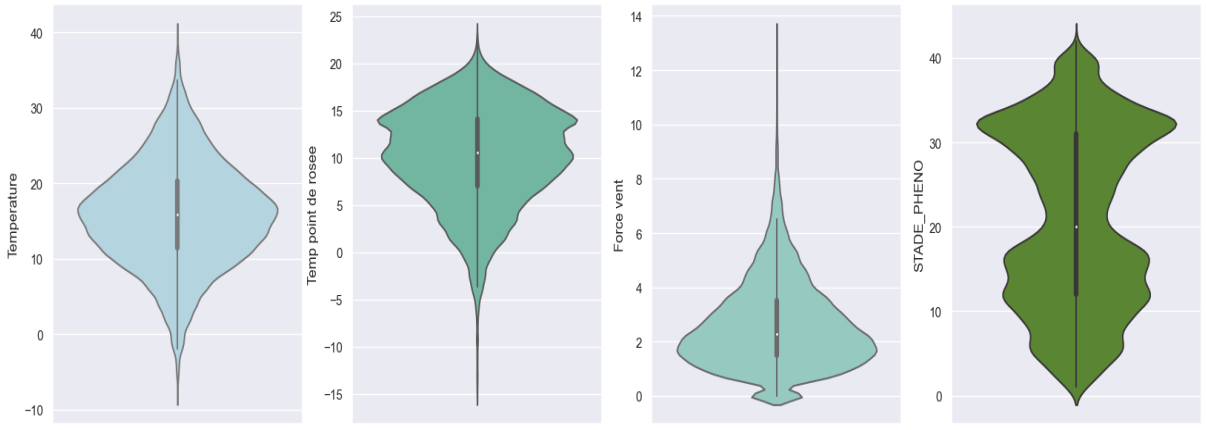
Boxplot:
mesure de dispersion (quantiles) des variables quantitatives.



Selon **boxplot**, on constate que:

- Les observations de la variable Temperature a la moitié de ses observations entre 11-21, avec ses outliers sur les deux extrêmes (-2 vers -10, 33 - 40)
- Temp point de rosee se concentre entre 7 et 14, avec ses outliers unilatéraux (-15 vers -4)
- Force vent se concentre entre 2 et 4, ses outliers sont également unilatéraux (6 - 14).
- Les observations de STADE_PHENO n'a pas d'outliers et la moitié se concentre dans la fourchette 12 et 32.

Violinplot
combine les fonctionnalités de **histogramme** (distribution de densité) et celles de **boxplot** (les quantiles).

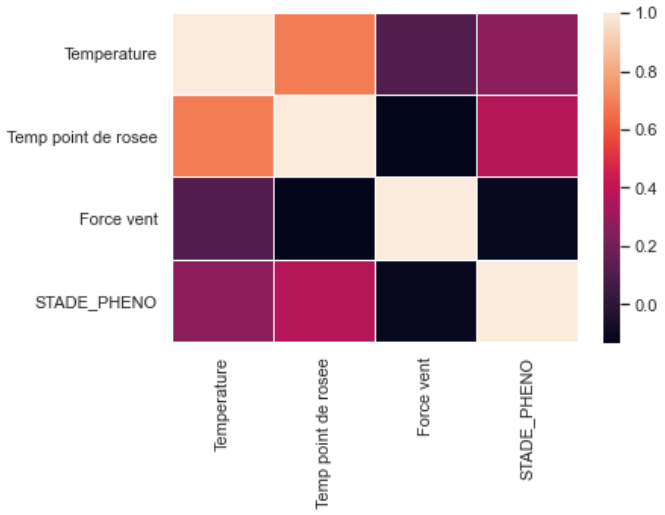


Selon ces graphiques, on observe que les variables météo connaissent leur distribution uni-modale, alors que la variable phénologique stade_pheno a sa distribution bi-modale.

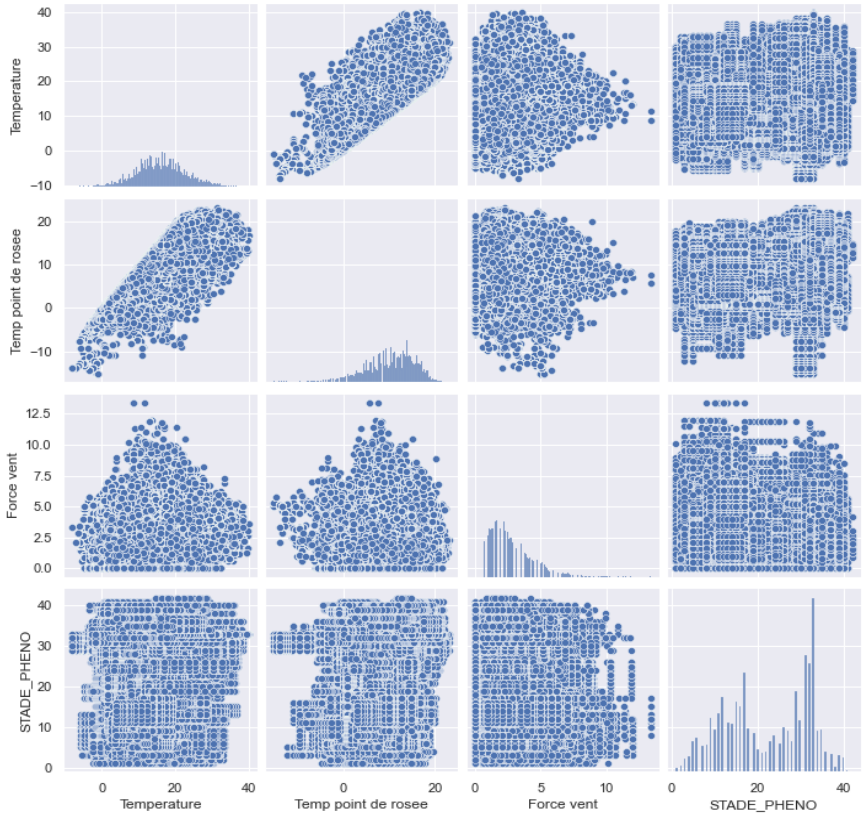
4.2 Analyse descriptive - bivariée

Étude de corrélations entre les variables:
Les températures et températures point de rosée se corréllent positivement à 68%.
A partir de cet **heatmap**, on n'en constate pas de corrélations supérieures à 30%.

Heatmap:
Illustrer la nuance des chiffres (coeff de corrélation) à travers les couleurs nuancées



Il est alors nécessaire d'étudier la possibilité de régression entre ces deux variables fortement corrélées.

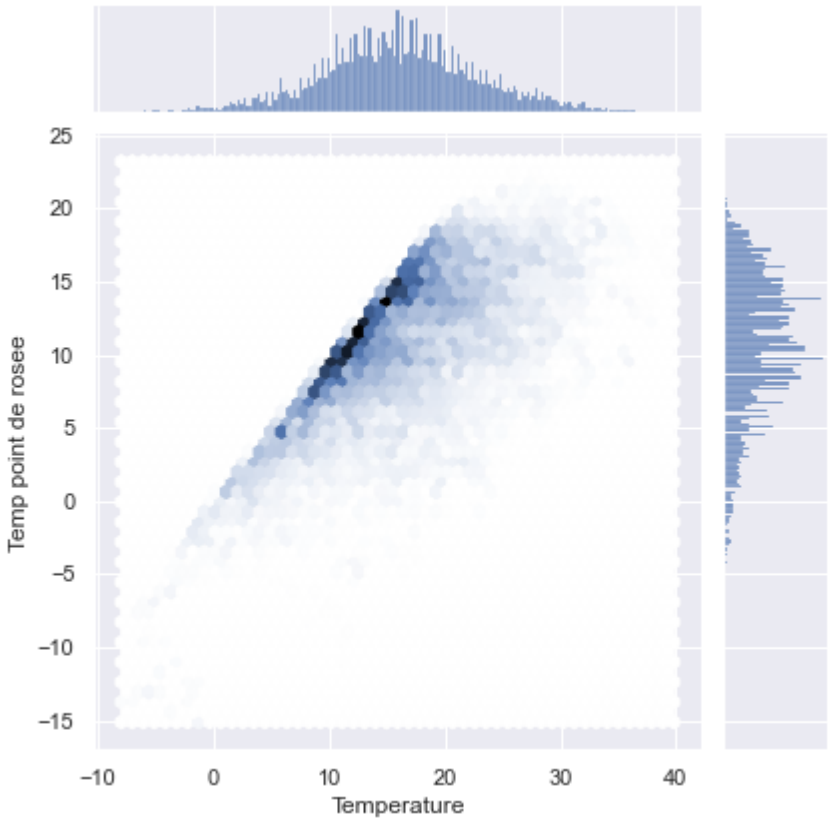


Pairplot (scatterplot + histplot)

Scatterplot :
possibilité de régression entre les variables quantitatives

Histplot / histogram univarié (en diagonal):

- Pour illustrer la distribution de densité de chaque variable quantitative.
- À la différence de barchart, histogram n'a pas de trous entre les bars, c'est toujours continu.



Jointplot (hexbin + histplot)

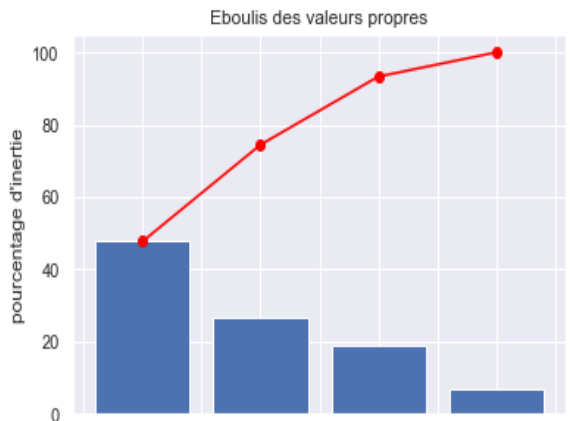
Hexbin:

- Relation entre deux variables quantitatives
- On a énormément de données
- Le plus la couleur est foncée, le plus le nombre de points superposés.

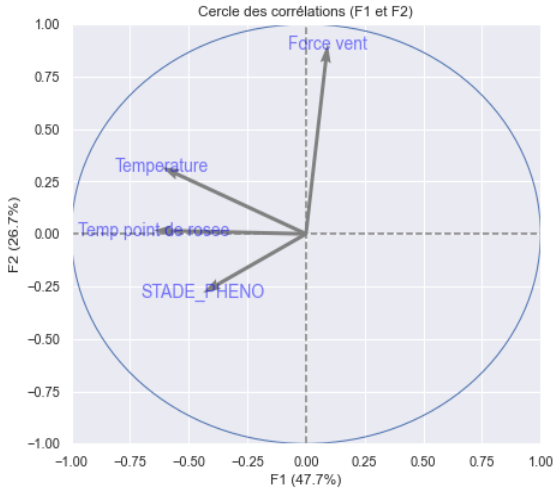
Histplot / histogram (en haut et à droite):
Densité distribution de chaque variable quantitative.

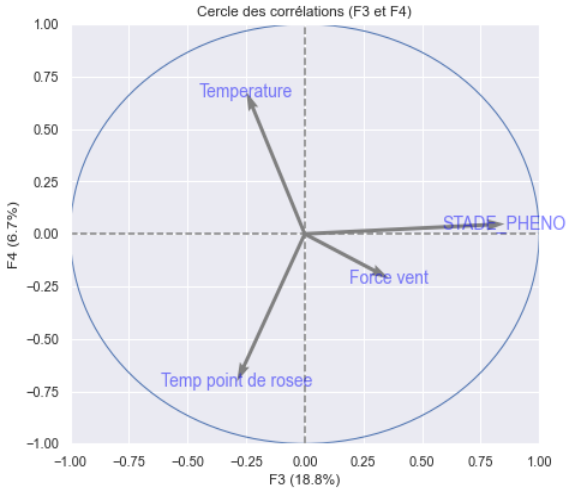
Selon **jointplot** qui regroupe hexbin et histogram, Temperature a sa densité autour de 10 et 20, temp point de rosée connaît sa densité entre 7 et 17. Par contre selon **scatterplot**, aucune de ces variables ont la colinéarité.
C'est ainsi qu'on a besoin de l'analyse exploratoire et inférentielle pour en décortiquer plus d'information.

4.3 Analyse exploratoire - ACP



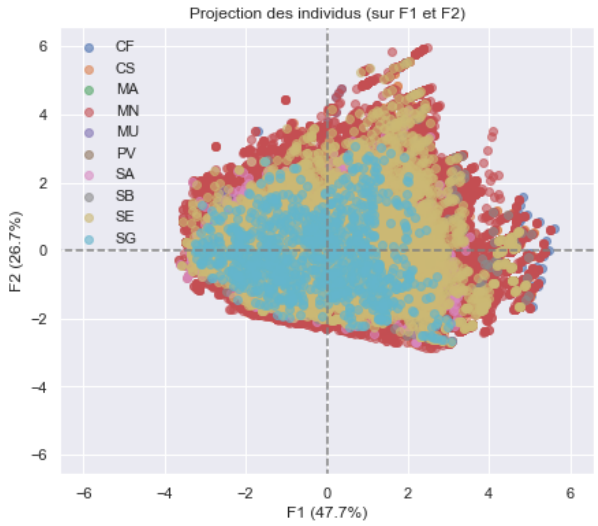
- Facteurs 1 et 2 représentent en somme quasi 75% de l'inertie totale.
- Facteurs 1+2+3 représentent au total vers 93% de l'inertie totale.

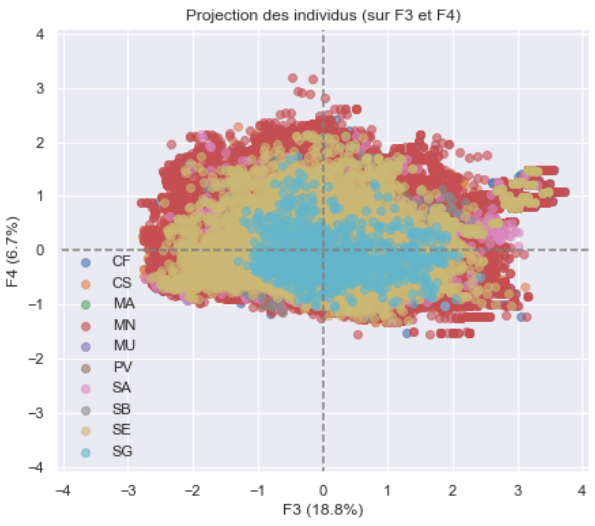




- Temp point de rosee se superpose avec F1 au sens négatif
- Force_vent se superpose quasiment avec F2 au sens postif et a une flèche longue
- Temperature est aussi pas mal représenté par F1
- STADE_PHENO se superpose avec F3 au sens positif et a une longue flèche. Donc elle est bien représentée par F3 qui joue 18.8% d'inertie totale.

4.3 Analyse exploratoire - ACP - projection des individus





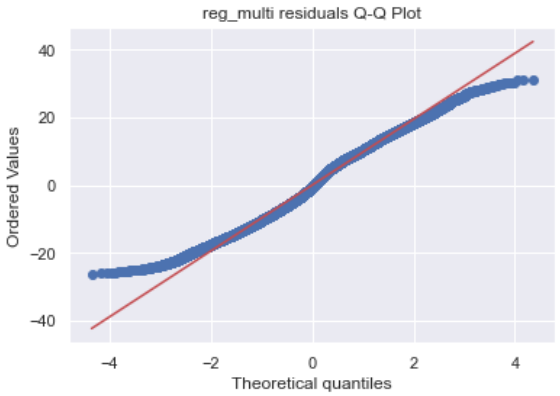
On en constate que sur la dimension de ces 4 facteurs, les cépages ne se partitionnent pas et se superposent énormément, pour les cépages différents, ces 4 facteurs qui regroupent toutes les variables en question n'ont pas d'impact différencié, autrement dit tous les cépages partagent les caractéristiques pareilles sur toutes les variables.

Donc **il n'y a pas grande raison de prendre en compte le cépage comme variable dans notre algorithme.**

4.4 Statistique inférentielle_multi régression linéaire

Pour savoir si il existe une corrélation entre les **variables météo** (temperature, temp point de rosee, force vent) et **variable vigne** (STADE_PHENO), on fait ici une régression linéaire multiple.

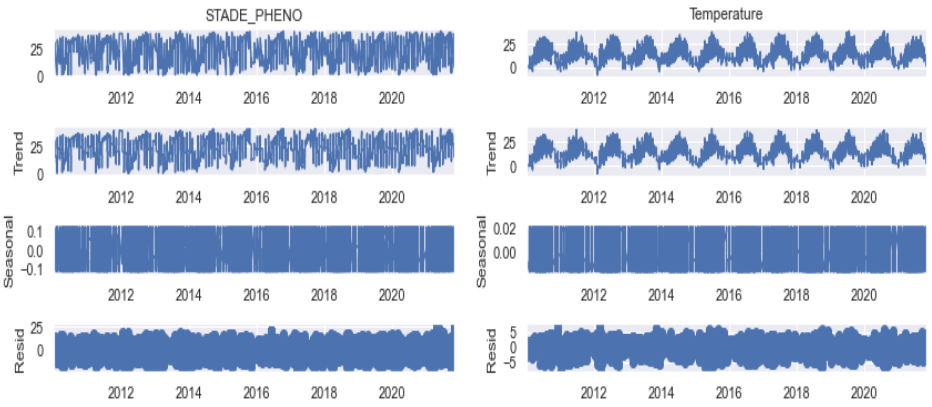
Selon la régression, on constate que : P-values < 5%
Donc on en conclut que les variables météo (incluant Temperature,Temp point de rosee, Force vent) ont de l'impact sur la variable vigne (stade_pheno,) mais au total la variance expliquée en stade_pheno est seulement 14% - traduit par le R-squared (coefficient de détermination).
Ainsi on peut en conclure qu'**il n'existe pas de corrélation considérables entre variables météo et variable vigne.**

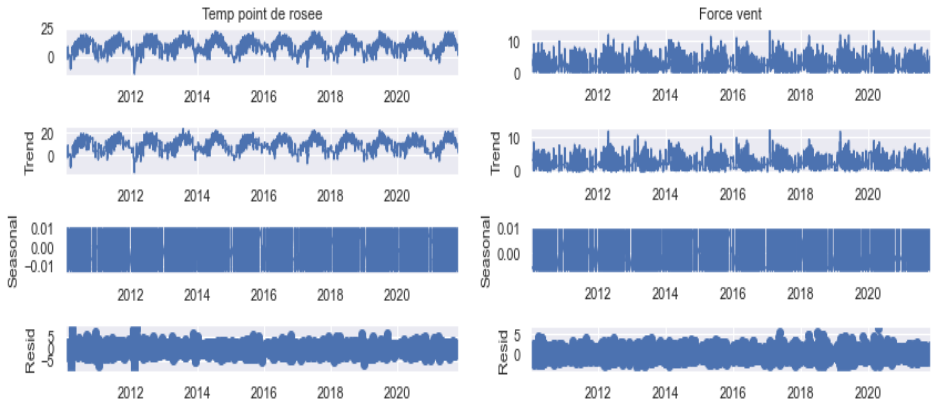


Normalité de résidu.

4.5 Saisonnalité, est-elle impactante?

Mais on a bien constaté que ces 4 variables sont toutes les séries temporelles.
Donc à la rigueur, on va étudier: d'une part, si la saisonnalité est suffisamment importante pour ne pas être ignorer, d'autre part, si c'est correcte de prendre en compte le côté saisonnier dans l'algorithme.





Selon ces 4 plot de seasonal_decompose, on en conclut que déjà les **Trend** présentent la fluctuation quasiment à la même échelle avec les séries brutes, et que les **Seasonal** se cantonnent dans les fourchettes super bridées (-0.1 - 0.1), donc on pourrait ignorer l'impact de la saisonnalité.

4.5 la saisonnalité est une facteur directe dans la prédiction?

Selon ce lineplot, on observe que toutes ces 4 variables se manifestent un comportement saisonnier à l'intervalle de 12 mois. Mais à noter que ici on ne cherche à prédire aucune de ces 4 variables pour lesquelles les data proviendront de Météo France et d'autres instituts publics.
Ce qu'on a pour but de prédiction, c'est le **niveau de risques** et ses impacts à l'agriculture en fonction de là où se trouve la parcelle (dans ce cas précis, on prend la vigne pour example). De plus, même si la saisonnalité a de l'influence (très petite) sur la prédiction, elle **n'a pas de rapport directe avec la valeur qu'on va prédire - risques**, vu qu'elle est déjà incluse dans ces 4 variables explicatives là.

C'est ainsi qu'on avance dans l'étape suivante **sans devoir prendre en compte le côté de saisonnalité des séries temporelles.**

5. Comment les données se tournent-elles?
- Schéma de flux de données chez projet FLORAL

Faute de jeux de données existantes conventionnés pour mesurer les dégâts, on n'a qu'à faire avec en référant la littérature du domaine vigne.

Suite à des recherches dans la littérature du vignoble, en combinaison avec les expériences des vignerons, on en sort une méthode de traitement qui noue le lien au maximum possible avec la pratique afin d'effectuer la prédiction sur de divers niveaux de risque.

DATE_OBS	Temperature	Temp point de rosee	Force vent	CEPAGE
2010-01-04	3.8	3.2	2.1	MN
2010-01-04	3.9	3.2	1.5	MN
2010-01-04	4.0	3.6	1.5	MN
2010-01-04	3.8	3.4	1.8	MN
2010-01-04	3.7	3.2	3.3	MN

```
data_youpi['SCORE'] = data_youpi.SCORE_TEMP.map(str) \
                    + data_youpi.SCORE_VENT.map(str) \
                    + data_youpi.SCORE_TEMP_ROSEE.map(str)\
                    + data_youpi.STADE.map(str)
```

4 STADE	2 SCORE_VENT	3 SCORE_TEMP_ROSEE	1 SCORE_TEMP	SCORE
1	1	0	0	0101
1	1	0	0	0101
1	1	0	0	0101
1	1	0	0	0101
1	2	0	0	0201
...	...	...	...	...
3	1	0	0	0103
3	1	0	0	0103
3	1	0	0	0103
3	0	0	0	0003
3	0	0	0	0003

On classe les risques en 5 niveaux:
risque nul - risque modéré - risque moyen - risque fort - risque très fort.
Représenté respectivement par chiffre 0 - 1 - 2 - 3 - 4.

```
liste0 =['0000','0001','0002','0003','0111', '0222','0011','0012','0013','0101','0102','0103','0112','0113','0121','0122','0123','0201','0202','0203','0211','0212','0213','0221','0223','0301',
         '0302','0303','0311','0312','0313','0321','0322','0323','1001','1002','1011','1012','1101','1102','1111','1121','1122','1201','1202','1211','1321','1322','2111','2112','2121','2122',
         '2201','2202','2211','2212','2221','2222','2301','2302','2321','2322','3011','3012','3021','3022','3111','3112','3121','3122','3201','3202','3212','3221','3222','3301','3302','3321',
         '3322']
liste1 = ['1003','1013','1023','1103','1113','1123','1203','1213','1223','1303','1313','1323']
liste2 =['2013','2113','2203','2213','3013','3113','3203','3213']
liste3 =['2123','2223','2303','2323','3303']
liste4 =['3023','3123','3223','3323','4001','4002','4003','4023','4121','4122','4123','4221','4222','4223','4321','4322','4323']
```

6. Conclusion

J'avais pour un des mes objectives de m'intégrer dans une équipe pour savoir comment se passe, dans la pratique, la collaboration inter-disciplinaire. Et on sait que pour tous types de projets, l'enjeu de succès c'est qu'on puisse optimiser la collaboration et la coopération de toute expertise qu'on a. Ce projet me permet d'avoir tout d'abord, une première expérience sur comment les différents rôles se collaborent autour d'une mission dans le cadre de création de solution opérationnelle. Il m'a donnée aussi une connaissance globale de tous les métiers y relatifs tel que data scientist, développeur web, product manager etc.

D'ailleurs, j'ai appris à travers ce projet des méthodologies en matière de gestion de projet technologique, en data science et des connaissances en vigne. Ce qui m'assure que pour pouvoir faire du bien la profession de Data Analyst, une connaissance des enjeux du domaine où se trouvent les données est indispensable. C'est ainsi que pour tirer la meilleure partie de l'analyse, on doit s'assurer que l'analyse se base sur un fond solide. Cela résonne aussi à ce que la data scientist - Mme Christelle DROUSSARD a affirmé: "la première chose d'être un bon data scientist c'est d'avoir la curiosité pour de divers domaines". Contrairement à ce qu'on a pour cliché contre ce métier. Et on est d'accord que la profondeur ne pénalise pas la vision/la largeur.

De plus mais pas moins important, ce projet hackathon me sert une occasion de faire l'évaluation de mes compétences en data science et ensuite l'auto-estimation sur comment puis-je améliorer pour progresser jusqu'au niveau souhaité. Et cela me donne une vision plus claire sur mon ambition d'avancer dans le métier jusqu'au data scientist.

Du côté pratique, on avait un bloçage disant le manque de mesures quantitatives en dégâts du domaine vigne. Ainsi on était dans l'obligation de créer de notre façon les mesures de risques. C'était pas idéale, mais en terme d'approches, on a fait le nécessaire avec toute rigueur, pour que la méthode soit alignée avec le besoin réel au maximum possible. C'est ainsi que le problème est levé. Et c'est pas étonnant que à part le nettoyage et le traitement, l'on affronte en amont de milliers de challenges dans la pratique tel que le manque de data, la mauvaise source de data, data non-daté etc., ce qui embête même pénalise la prédiction. la data science nous apprend pas simplement les théorèmes statistiques ou/et programmation, mais aussi la méthodologie et la rigueur d'aborder les problématiques. C'est ainsi qu'on a pu s'en sortir avec une méthode de scoring et d'évaluation de toute rigueur possible et convient au besoin métier à la fois.

Spécialement remercie à :

Christelle TROUSSARD, Jean-Philippe TROUSSARD, Damien AMAUDRY, Pierre CARTIGN, Manon MARRON, Chloé RENAULT, Alexandre THEBAULT et toute l'équipe d'organisation de cet évènement hackathon Varenne de l'eau.