# Rapport d’analyse des données

Table des matières

[Rapport d’analyse des données 1](#_heading=h.gjdgxs)

[General 1](#_heading=h.ctczf2yyz1p3)

[Valeurs nulles 2](#_heading=h.30j0zll)

[En globalité 2](#_heading=h.1fob9te)

[Répartition géographique 4](#_heading=h.3znysh7)

[Répartition temporelle 5](#_heading=h.2et92p0)

[Gestion des valeurs nulles 9](#_heading=h.tyjcwt)

[Corrélations 13](#_heading=h.3dy6vkm)

[Variables quantitatives 13](#_heading=h.1t3h5sf)

[Variable cible “RainTomorrow” 16](#_heading=h.jxwbirqm2o1n)

[Variables qualitatives 17](#_heading=h.4d34og8)

[Géographie 18](#_heading=h.2s8eyo1)

[Répartition géographique 18](#_heading=h.17dp8vu)

[Clusterisation 20](#_heading=h.3rdcrjn)

[Saisonnalité 23](#_heading=)

[KNN Imputer 26](#_heading=)

[Bilan général 27](#_heading=h.lnxbz9)

## 

## General

La base de données se compose de différentes variables météorologiques, dont la variable cible "RainTomorrow" est une boolean indiquant s'il va pleuvoir ou non le jour suivant. La répartition entre jours de pluie et jours sans pluie varie considérablement d'une ville à l'autre, mais dans tous les cas, les jours de pluie représentent une minorité, dans certains cas moins de 10 %. Cela indique que la base de données a une tendance déséquilibrée. Il sera intéressant d'évaluer, lors de la phase de modélisation, la mise en œuvre d'un modèle pour les bases de données déséquilibrées.



En ce qui concerne les variables quantitatives, le diagramme de Mostache montre qu'il existe une grande variation dans la gamme de valeurs de chaque variable, de sorte qu'un processus de mise à l'échelle sera nécessaire avant la phase de modélisation.

## 

## Valeurs nulles

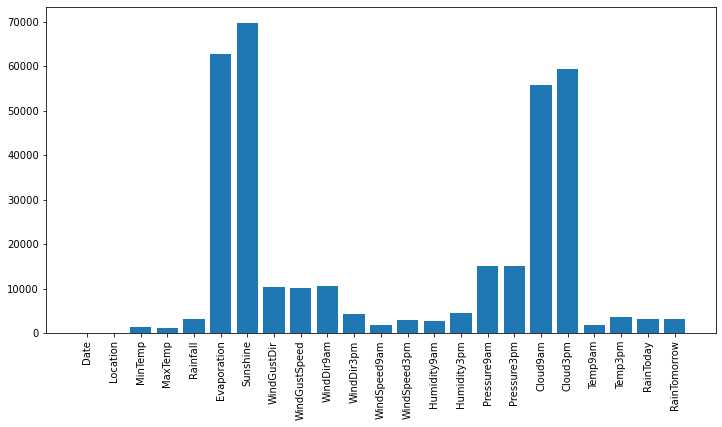
### En globalité

Quatre variables comportent entre un tiers et la moitié de données nulles (Evaporation (43%), Sunshine (48%), Cloud9am (38%), Cloud3pm (41%)).

Les deux variables relatives à la pression (Pressure9am, Pressure3pm), bien que renseignées davantage, ont un taux significatif de données nulles (10%).

Toutes les mesures disposent d’une date et d’une Location.

| Feature | Nb Null | % Null |
| --- | --- | --- |
| Date | 0 | 0% |
| Location | 0 | 0% |
| MinTemp | 1485 | 1% |
| MaxTemp | 1261 | 1% |
| Rainfall | 3261 | 2% |
| Evaporation | 62790 | 43% |
| Sunshine | 69835 | 48% |
| WindGustDir | 10326 | 7% |
| WindGustSpeed | 10263 | 7% |
| WindDir9am | 10566 | 7% |
| WindDir3pm | 4228 | 3% |
| WindSpeed9am | 1767 | 1% |
| WindSpeed3pm | 3062 | 2% |
| Humidity9am | 2654 | 2% |
| Humidity3pm | 4507 | 3% |
| Pressure9am | 15065 | 10% |
| Pressure3pm | 15028 | 10% |
| Cloud9am | 55888 | 38% |
| Cloud3pm | 59358 | 41% |
| Temp9am | 1767 | 1% |
| Temp3pm | 3609 | 2% |
| RainToday | 3261 | 2% |
| RainTomorrow | 3267 | 2% |



### Répartition géographique

En stratifiant les données par ville (Location), nous pouvons voir que les données nulles ne sont pas uniformément réparties géographiquement. En particulier, 3 lieux d’environ moitié moins d’enregistrements que les autres (Katherine, Nhil, Uluru). Il ne s’agit pas de données nulles à proprement parler, mais cela signifie qu’il y a moins de journées pour lesquelles les données météo ont été enregistrées pour ces endroits.

Concernant les variables évoquées plus haut ayant des taux élevés de null, nous voyons qu’il existe en réalité plusieurs villes pour lesquelles ces données sont rarement absentes.

De la même manière, la répartition des données nulles sur les autres features est très disparate. Melbourne concentre notamment une large proportion des données nulles sur nombre de variables. Dans une moindre mesure, c’est également le cas de quelques autres lieux (Albany, Canberra, Coffs Harbour, Mount Gimini, Newcastle, PearceRAAF, Sydney, Williamtown).



### Répartition temporelle

Les données disponibles vont du 1er novembre 2007 jusqu’au 25 juin 2017, ce qui représente 3525 journées. Toutefois, les enregistrements météo ne recouvrent pas l’intégralité de cette plage. On voit sur le graphique précédent que pour la plupart des villes, seules 3000 journées environ sont disponibles. Il manque donc l’équivalent d’environ 1 an et demi de mesures pour la plupart des villes.

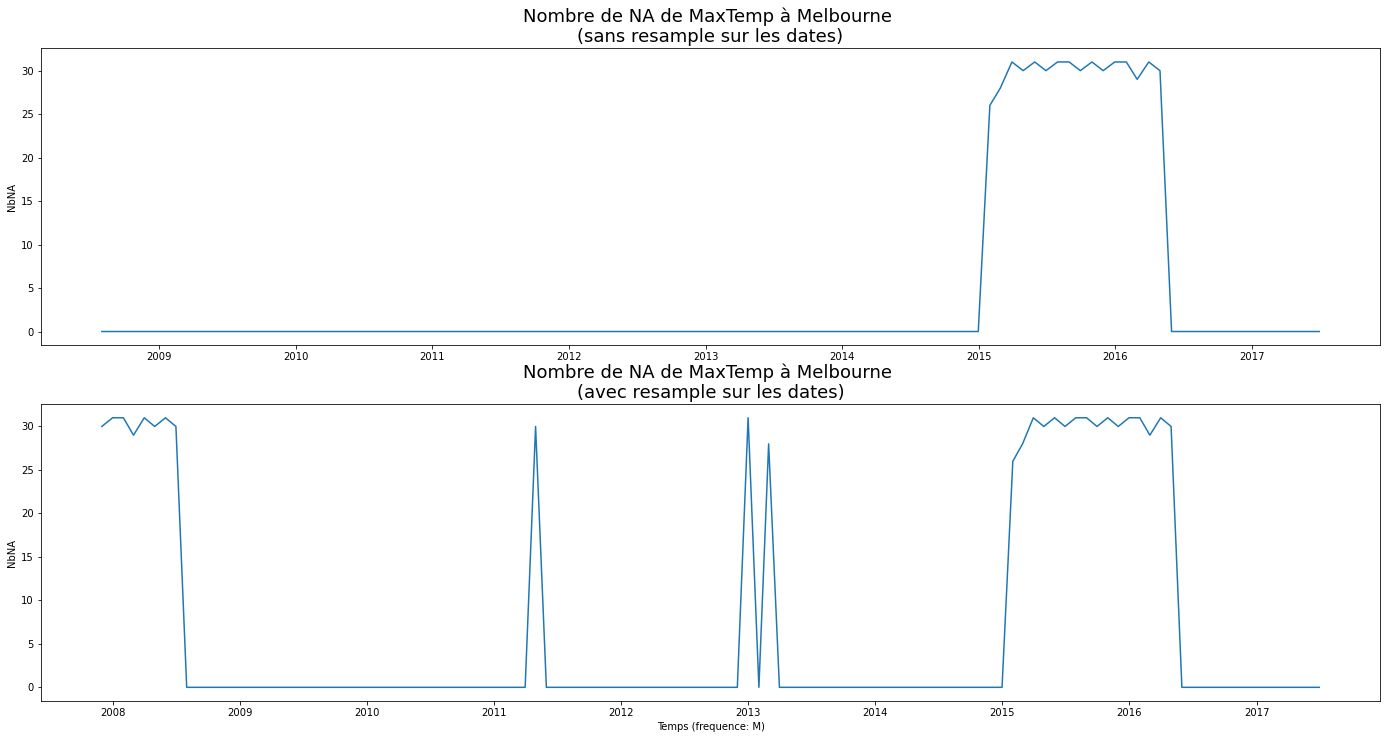
Afin de pouvoir analyser la répartition des données temporellement, il faut préalablement ajouter les journées absentes du jeu de données pour chaque Location. En effet, en l’absence de cette étape, seules les journées connues dans le dataset génèreront un NA pour une variable. Les journées absentes quant à elle n’entraîneront pas de NA. Or, nous avons besoin de savoir s’il y a des journées totalement manquantes.

Exemple sur la variable MaxTemp pour Melbourne : les deux graphiques suivants indiquent le nombre de journées par mois pour lesquelles il y a des NA vus sur la variable MaxTemp à Melbourne.

Le premier graphique, effectué sur le dataset non rééchantilloné, montre qu’il manque la quasi-totalité des données pour chaque mois pour MaxTemp de 2015 à mi 2016 mais ne témoigne pas d’autres données manquantes.

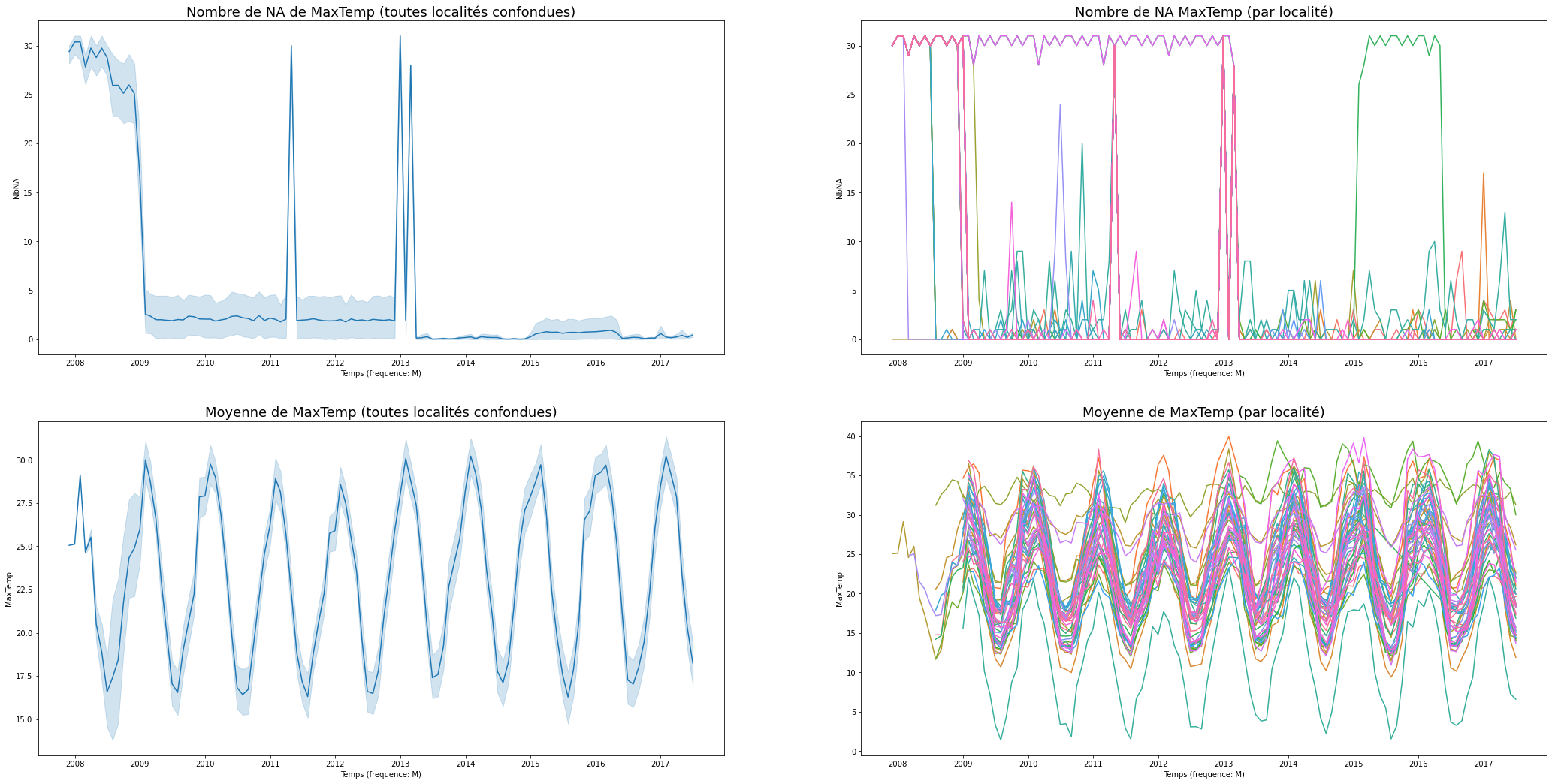
Le second graphique, réalisé sur des données rééchantillonée sur l’intervalle complet des dates met en évidence d’autres périodes pour lesquelles MaxTemp est inconnue. Il s’agit de dates qui étaient totalement absentes des données d’origines.

*(généré avec : comparaison\_avec\_sans\_dates\_reindexees("Melbourne", "MaxTemp", "M") )*



Les quatre graphiques suivants montrent :

* Le nombre de jours par mois pour lesquels MaxTemp est à NA toutes localités confondues
* Idem, avec une courbe par localité
* La moyenne de MaxTemp toutes localités confondues
* Idem, avec une courbe par localité



*(généré avec analyse\_variable\_temps("MaxTemp","M") )*

Du fait du nombre important de localités, nous avons retiré les légendes des deux graphes de droites : l’objectif n’est pas tant ici d’identifier des valeurs pour une localité précisé que d’observer le comportement général des données.

Sur le premier graphe, nous pouvons constater que MaxTemp est globalement rarement disponible avant 2009. Nous voyons aussi qu’il n’existe aucune donné pour aucune Location pour le mois d’avril 2011. Il en va de même des mois de décembre 2012 et février 2013.

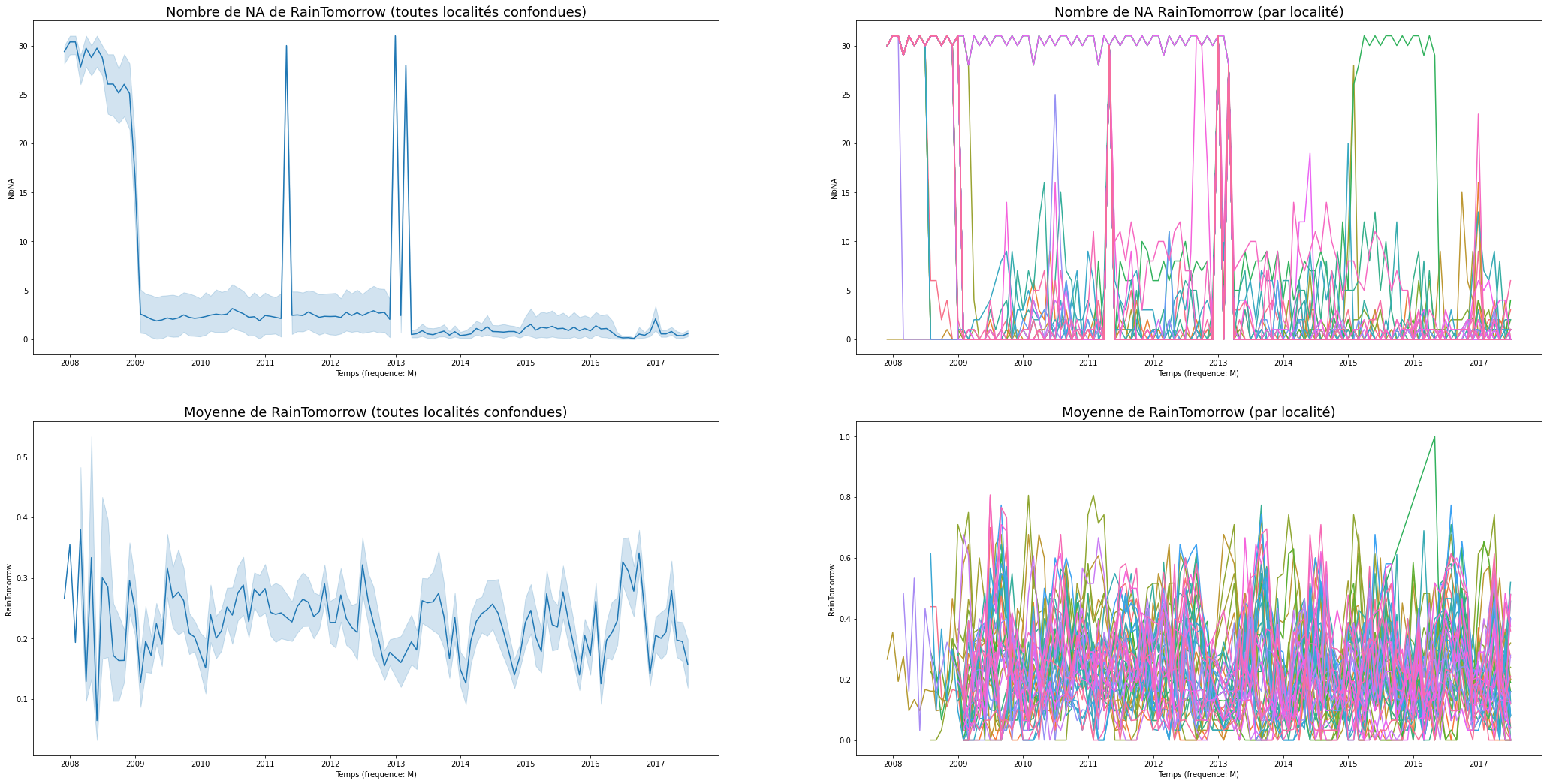
Il y a une variance importante sur le nombre de NA/mois/Location entre 2009 et fin 2012. Celle-ci s’affaiblit grandement ensuite, à l’exception de la période de début 2015 à mi 2016.

Le second graphique permet de mieux comprendre ces constats globaux : on voit qu’il n’existe en effet aucune Location disposant de données sur les 3 mois signalés plus haut (04/11, 12/12 et 02/13). Nous voyons également qu’il y a des Locations qui ont des mois entiers (voire des années !) sans MaxTemp renseigné. Enfin, nous voyons ne semble pas y avoir de période au cours de laquelle MaxTemp est disponible pour l’ensemble des localités.

[analyse de la moyenne plus tard, lorsque j’aurai terminé le module des séries temporelles]

(Cycle annuel, par localité et en global)

Les 4 graphes suivants montrent les mêmes représentations, mais pour la variable cible ‘RainTomorrow’. Il est frappant de voir la similarité des deux premiers graphes, qui indiquent le nombre de jours pour chaque mois pour lequel la variable est indisponible.



Nous n’allons pas reprendre ici ces graphes pour l’ensemble des variables, mais nous les avons observés (fonction « analyse\_variables\_temps() »). Le premier graphe est ainsi quasi identique pour toutes les variables, hormis pour celles dont l’absence est plus fréquente. Le graphe de ces dernières présente logiquement une moyenne mensuelle de NA plus élevée et une variance plus forte. Le second graphe, représentant le nombre de NA mensuel par localités, est en revanche très différent selon les variables témoignant d’une forte disparité de la disponibilité des variables suivant les localités.

La faible disponibilité des données avant 2019 pour les différentes variables tend à faire penser que les données antérieures au 1er janvier 2019 ne sont pas exploitables (hormis pour Canberra, et Sydney).

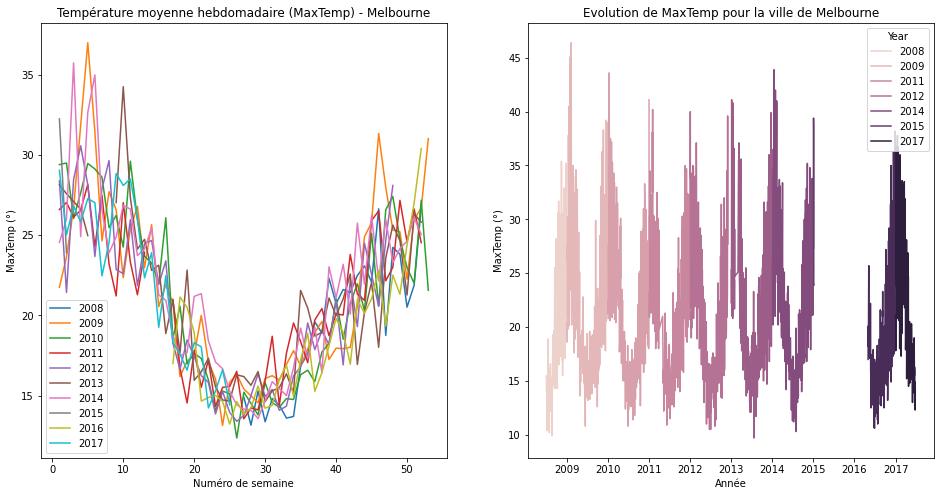
La question se pose de la façon de traiter les trois mois absents des données (avril 2011, décembre 2012, février 2013) : si certaines variables présente un cycle annuel permettant d’envisager une reprise de la valeur à la même date sur d’autres années (MaxTemp par exemple), il n’en va pas de même pour toutes les variables, en particulier la variable cible RainTomorrow.

[à enrichir, non rédigé]

Il y a des mois entiers pour lesquels aucune info n’est disponible sur l’ensemble du jeu de données.

Exemple ci-après (graphe de droite) sur la variable MaxTemp à Melbourne qui est manquante de début 2015 à mi 2016.

Todo : graphe complet des plages de dates manquantes, par variable, par ville



### Gestion des valeurs nulles

#### Metodo KNN\_imputer

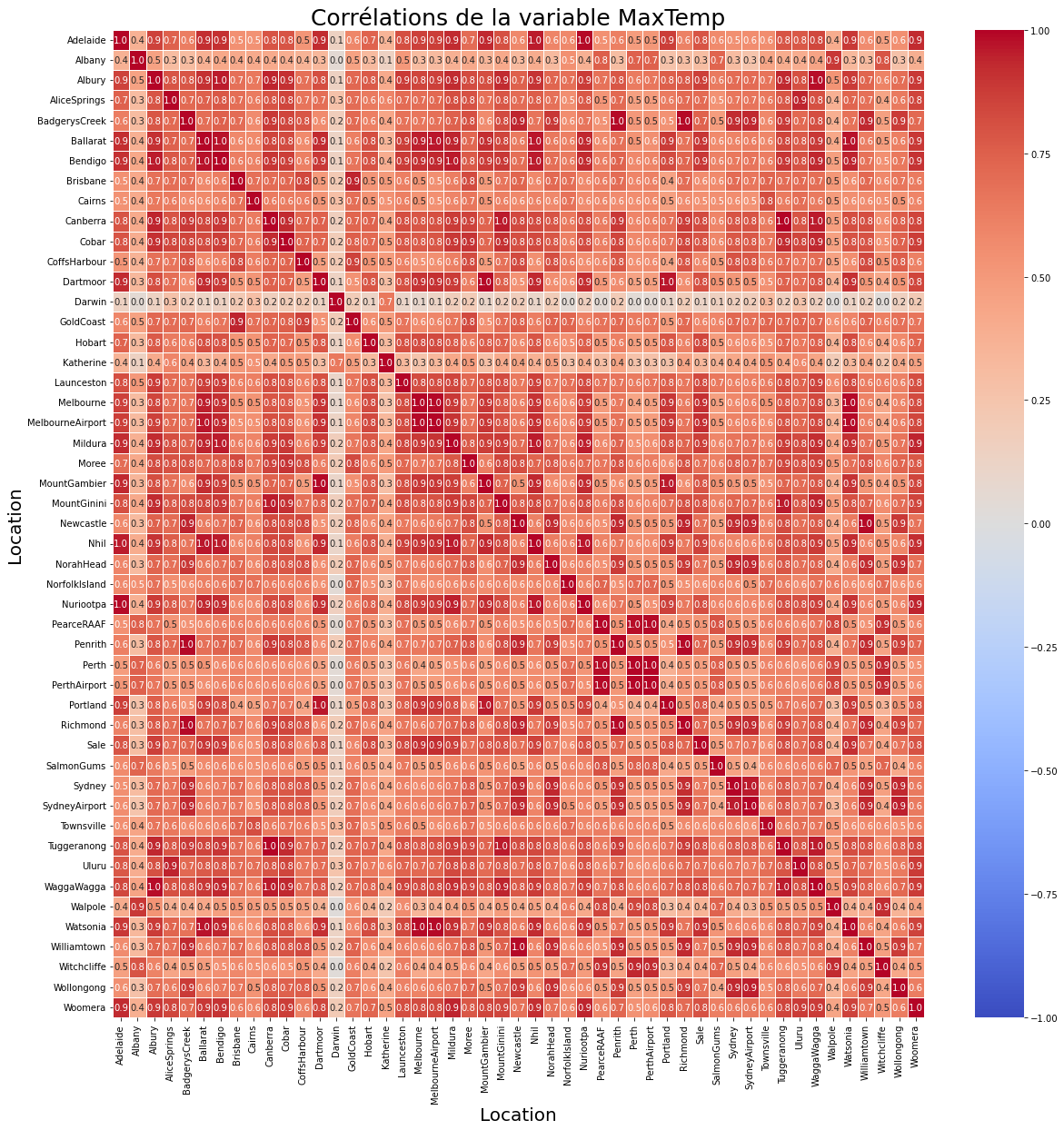
L'une des méthodes les plus couramment utilisées pour traiter les valeurs manquantes est la méthode KNN Imputer, qui estime les valeurs manquantes à l'aide des informations provenant des voisins les plus proches dans l'espace des caractéristiques.

(je pense on peut mettre l'analyse de Quyen ici)

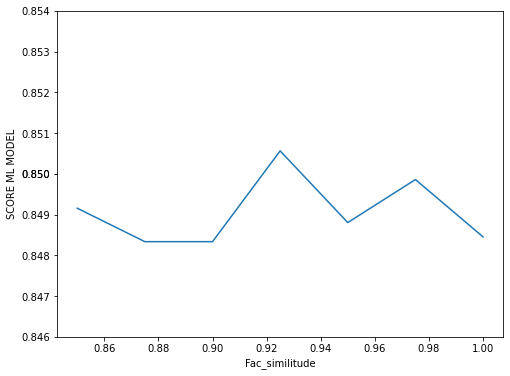
#### Utiliser les données des villes voisines

La méthode KNN remplace les valeurs manquantes en recherchant des similitudes dans l'ensemble de la base de données. Cependant, la base de données ne contient aucune information sur la situation géographique des villes, mais en regardant les emplacements sur la carte de l'Australie, nous observons que de nombreuses villes sont très proches les unes des autres et nous pourrions donc utiliser le voisin le plus proche pour estimer les valeurs manquantes.

A titre d'exemple, nous montrons ci-dessous la corrélation des différentes villes pour la variable "MaxTemp". On constate que certaines villes ont une corrélation très proche de 1, comme c'est le cas de la ville Ballarat avec la ville Bendigo ou la ville de Melbourne. Ceci est logique puisque ces villes sont géographiquement proches les unes des autres. Par contre, pour les villes plus éloignées comme Darwin, la similitude avec le reste des villes est beaucoup plus faible.



Sur cette base, et dans le but d'améliorer les résultats de la méthode KNN\_imputer, nous avons évalué l'utilisation des données du plus proche voisin pour remplacer les valeurs altérantes. Dans ce cas, le plus proche voisin n'a pas été utilisé en termes de géographie mais en termes de données qui, comme nous l'avons vu plus haut, sont souvent géographiquement similaires. Pour cela, les valeurs manquantes d'une ville ont été remplacées par les valeurs de la ville la plus corrélée pour chaque variable, pour autant que cette corrélation soit supérieure à une certaine valeur appelée "Fac\_similitude". Ensuite, le reste des valeurs manquantes a été remplacé par la méthode KNN\_imputer et une méthode RandonForestClassifier a été utilisée pour évaluer le score en fonction des différents types de traitement des valeurs manquantes. Le graphique suivant montre le SCORE pour chaque prédiction en fonction des différentes valeurs du facteur "Fac\_similitude", Fac\_similitude étant égal à 1 le résultat obtenu en traitant toutes les données avec la méthode KNN\_imputer.



Bien que les valeurs de SCORE soient plus élevées que celles obtenues en traitant toutes les données manquantes avec la méthode KNN\_imputer, l'amélioration ne semble pas très significative. Ce point sera étudié plus en profondeur lors de la prochaine étape, lorsque le modèle du machine learning sera ajusté.

Pour simplifier l'analyse, la méthode RandonForestClassifier a été appliquée sur une fraction réduite des données et en traitant les variables qualitatives avec la méthode “get\_dummies”. Dans tous les cas, L'objectif de cette étape n'était pas d'optimiser la méthode de machine learning, mais de faire une première estimation des résultats des méthodes de traitement des valeurs manquantes et de vérifier que le code fonctionne.

#### Cas particulier de RainToday

Cette variable est supposée être à True lorsque Rainfall est supérieure ou égale à 1mm. Un contrôle permet de confirmer cette règle. De plus, lorsque RainToday est à NA, alors Rainfall l’est également et réciproquement. Il n’est donc pas possible d’utiliser Rainfall pour renseigner les valeurs manquantes de RainToday.

#### Cas particulier de RainTomorrow

La variable ‘RainTomorrow’ doit en théorie être égale à la valeur de ‘RainToday’ du lendemain pour la même Location. Après vérification, c’est en effet le cas. Une possibilité pour gérer les NA de cette variable est donc d’affecter la valeur de RainToday de la veille des dates pour lesquelles RainTomorrow vaut NA. Malheureusement, cette piste n’aboutit pas : lorsque RainTomorrow vaut NA, soit RainToday de la veille est à NA également, soit la date n’est pas présente pour la Location.

#### Autres variables

Après ces différents constats, plusieurs pistes s’offrent à nous pour gérer les valeurs nulles.

[à mettre au propre et à discuter/tester]

* Drop des 3 villes avec « seulement » la moitié des dates ?
* Drop des variables avec un taux de NA trop élevé ?
* Reprise des données météo pour les dates/variables manquantes à partir de la ville la plus proche (exemple de Melbourne pour laquelle il manque beaucoup de données, alors que la quasi-totalité des infos est disponible pour MelbourneAirport, situé à quelques kilomètres) ?
* Reprise de la moyenne des valeurs du même jour de l’année à partir des années renseignées ?
* Récupérer les données manquantes à partir d’autres sources de données (site de la météo australienne) ?

## Corrélations

### Variables quantitatives

Aucune variable quantitative n’est fortement corrélée avec RainTomorrow. Il existe toutefois des corrélations intéressantes (comprises entre 0,25 et 0,5) avec Sunshine, Humidity3pm, Humidity9am, Cloud9am, Cloup9pm, RainToday.

Pour Cloud9am et Cloud9pm, il s’agit malheureusement de deux features ayant un taux élevé de données nulles, comme nous l’avons vu en début de rapport.

Il semble donc que la prédiction fine de la variable RainTomorrow nécessitera de mobiliser la plupart des variables.

Nous constatons également de fortes corrélations entre d’autres features. Assez logiquement, c’est le cas de la température maximale (MaxTemp) avec celle relevée à 15h (Temp3pm), et de MinTemp avec Temp9am.

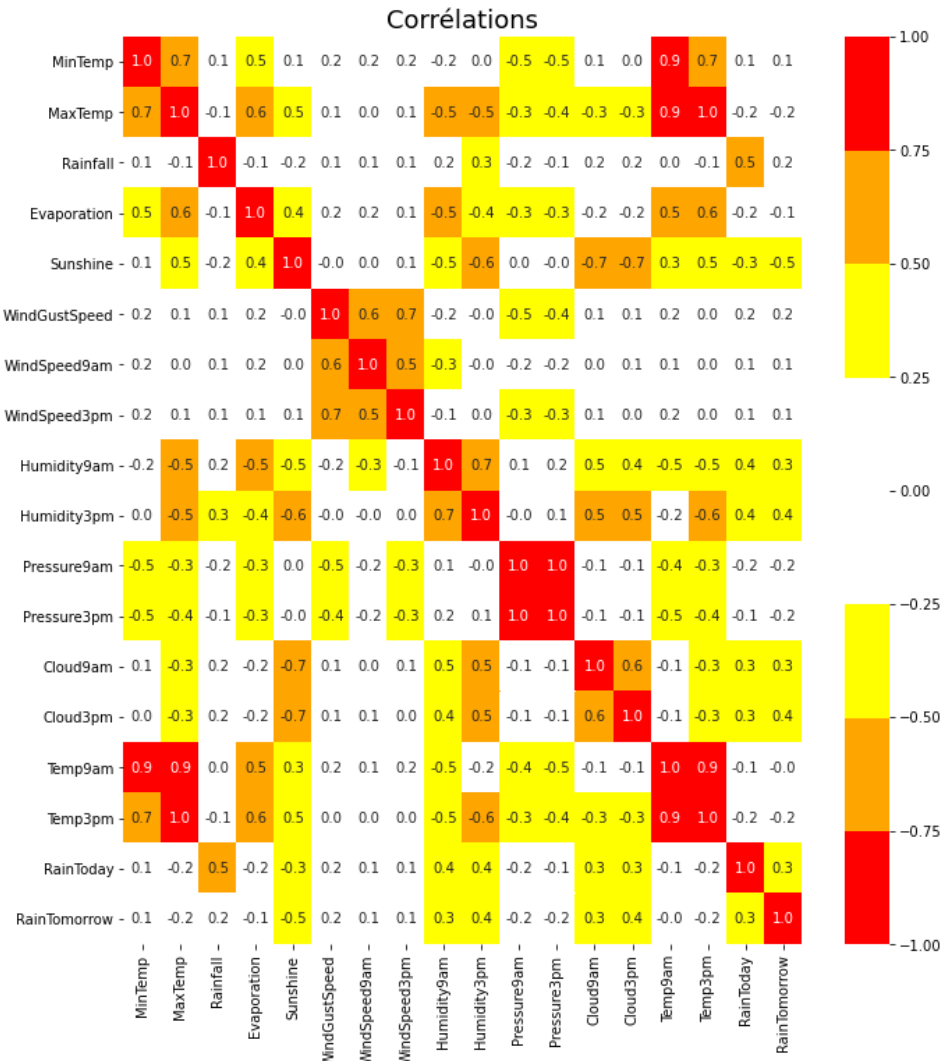
Plus étonnant de prime abord, c’est également le cas de MaxTemp avec Temp9am (0,9), ainsi que Temp9am et Temp3pm

Les températures min et max présentent aussi une corrélation intéressante (0,7).

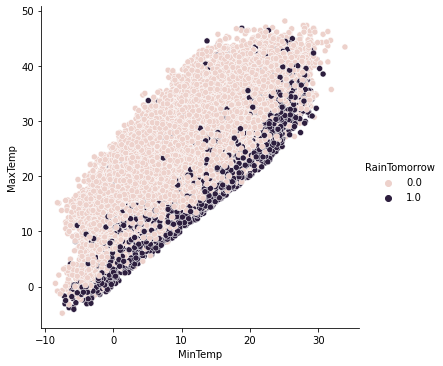
Bref, nous constatons que l’ensemble des variables de températures présentent une très forte corrélation.

C’est aussi le cas de la pression : les variables Pressure3am et Pressure9pm sont très fortement corrélées.

Ces constats peuvent nous permettre d’envisager de supprimer certaines de ces features, afin d’alléger les données traitées par les modèles.



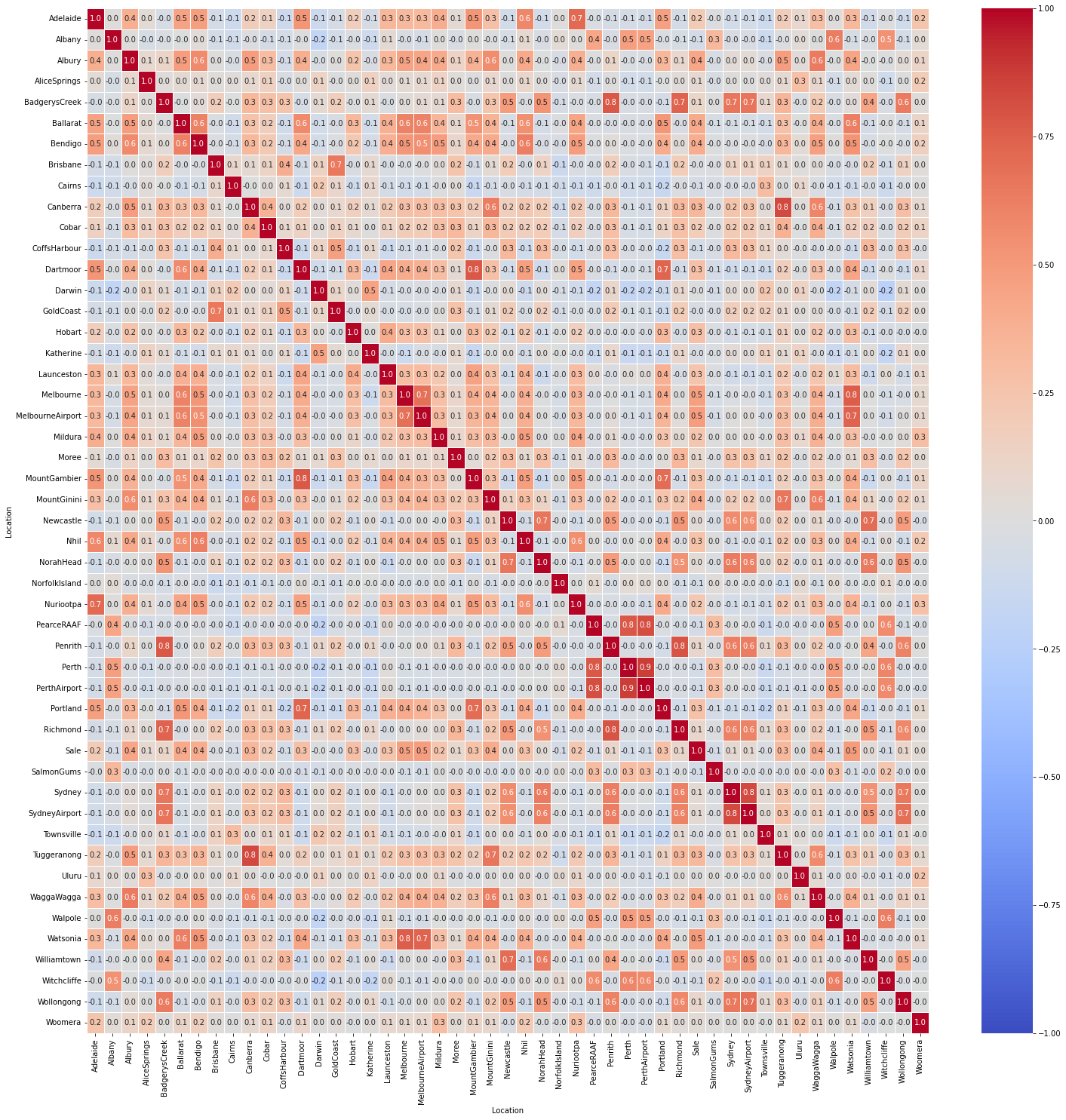
Le graphique ci-après trace en bleu les journées avec un RainTomorrow positif, et les positionne sur un graphe aux coordonnées (MinTemps, MaxTemp). Il est frappant de constater que les points bleus sont largement positionnés sur la première bissectrice, ce qui signifie qu’une très faible amplitude thermique semble fortement associée au fait qu’il pleuve le lendemain. Par conséquent, avant de supprimer les variables de température redondantes, il semble pertinent d’ajouter une nouvelle variable correspondant à l’amplitude thermique, que nous nommerons AmplitudeTemp. Il est d’ailleurs intéressant de constater qu’alors que RainTomorrow n’était corrélé qu’à moins de 0,25 avec chaque variable de température individuellement, elle l’est à 0,3 avec cette nouvelle variable !



### Variable cible “RainTomorrow”

Lorsque nous examinons la corrélation de la variable cible entre les différentes villes, nous observons des cas de fortes corrélations obtenant des valeurs supérieures à 0,8, contrairement au graphique précédent où la relation entre la variable cible et les autres ne dépassait pas une valeur de 0,5. Cela semble à nouveau logique compte tenu de la proximité de certaines villes entre elles.

Sur la base de ces résultats, il sera intéressant d'évaluer dans l'étape suivante l'incorporation de la variable "RainTomorrow" de la ville la plus corrélée. Dans ce cas, étant donné qu'il s'agit de la variable cible, deux modèles de machine learning devraient être appliqués en série.



### Variables qualitatives

Les variables qualitatives, outre la Location, sont celles liées à la direction du vent : WindGustDir (direction des rafales), WinDir9am et WindDir3pm.

Il nous est possible de tester la corrélation de ces 3 variables qualitatives avec RainTomorrow avec un test de Chi2, avec l’hypothèse nulle supposant qu’il n’existe pas de corrélation.

La méthode « correlation\_vent » calcule la p-value issu du Chi2 pour chacune de ces 3 variables afin d’en tester la corrélation avec RainTomorrow. La p-value est très inférieure à 0,05 en globalité, tout comme pour la plupart des villes, ce qui permet de rejeter l’hypothèse nulle et donc d’affirmer l’existence d’une corrélation.

Seules trois villes présentent une p-value pour l’une de ces 3 variable supérieure à 0,05. Cependant, même sur ces villes, il y a chaque fois au moins une variable qualitative avec une variable qualitative ayant une p-value <0,05

Villes ayant au moins une p-value >0,05

|  | WindGustDir | WindDir9am | WindDir3pm |
| --- | --- | --- | --- |
| Canberra | 0,00 | 0,06 | 0,07 |
| Tuggeranong | 0,17 | 0,00 | 0,00 |
| Townsville | 0,00 | 0,07 | 0,00 |
|  |  |  |  |
|  |  |  |  |

Nous pouvons donc déduire que la direction du vent est corrélée à chaque ville par au moins une variable. La force de cette corrélation n’est toutefois pas connue, le Chi2 ne permettant pas de la déterminer.

## Géographie

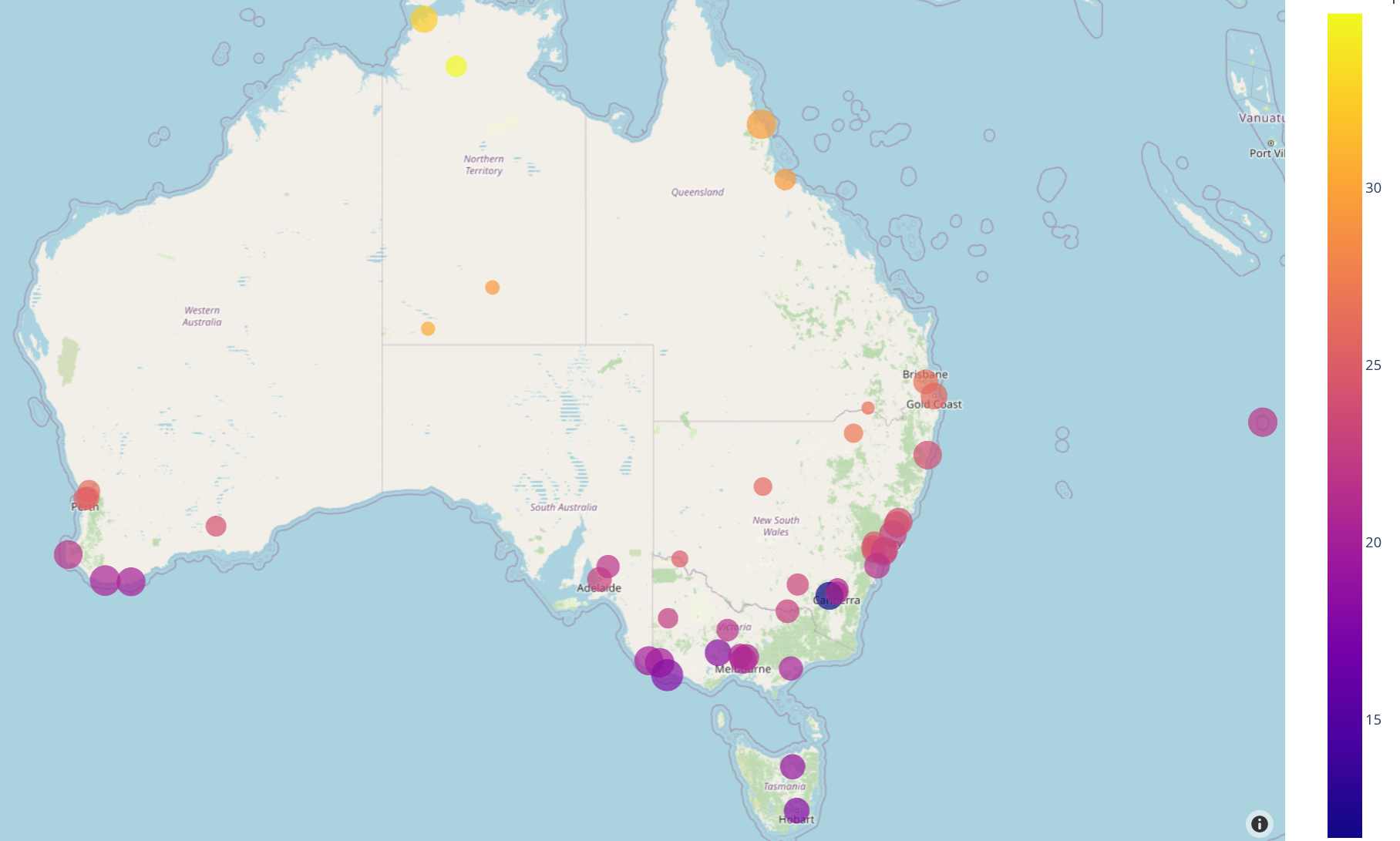
### Répartition géographique

Nous avons enrichi les données avec la latitude et la longitude de chaque Location à partir de plusieurs sources de données [préciser].

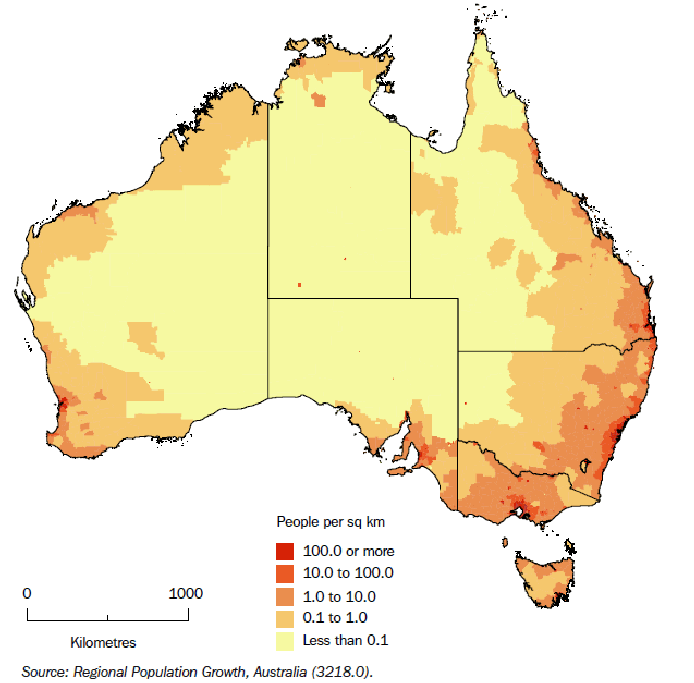
Le graphique ci-dessous montre la répartition des données sur les 49 lieux renseignés. La couleur indique la température maximum moyenne, le diamètre indique la moyenne de la variable RainTomorrow. Ainsi, le petit cercle orange représentant Uluru tout au centre de la carte témoigne qu’il pleut très peu dans cette ville et que les températures maximales sont élevées en moyenne (30°).

A l’inverse, le gros point bleu de MountGinini au sud-est témoigne d’une température maximale très faible (11°) et de précipitations plus importantes.

Notez que la ville d’Uluru est située au cœur du désert australien, alors que MountGini est une montagne culminant à 1762m



Il est frappant de constater que le jeu de données comporte essentiellement des informations sur des villes proches des côtes. C’est un point assez logique du fait de la géographie australienne. Sur le graphique ci-dessous, nous voyons ainsi que la densité de la population est nettement plus forte sur la côte sud-est et sud-ouest, et quasi nulle sur la partie centrale, occupée par le désert.



On remarque un gradient nord-sud assez net pour les températures maximales.

On remarque que la fréquence des RainTomorrow positif se réduit lorsqu’on s’éloigne des côtes.

Hormis quelques outliers tels MountGinini, on peut également constater une certaine homogénéité climatique entre des villes proches géographiquement, ce qui nous amène à envisager la création de cluster.

### Clusterisation

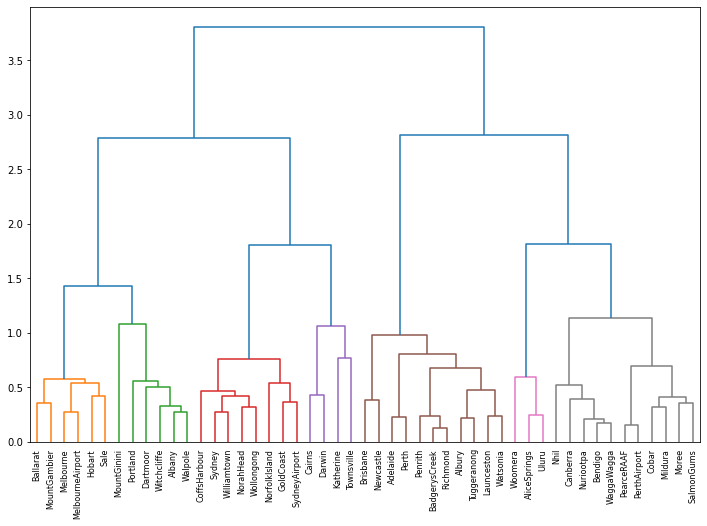
Nous pourrions ajouter une nouvelle variable précisant pour chaque Location le type de climat (désertique, subtropical, mousson, océanique,…) à partir de données complémentaires.

Plutôt que de le faire manuellement (ce qui reste une option, vu le nombre raisonnable de localités), nous pourrions mettre en place un modèle de clusterisation qui se chargerait lui-même de regrouper les Location présentant des similitudes.

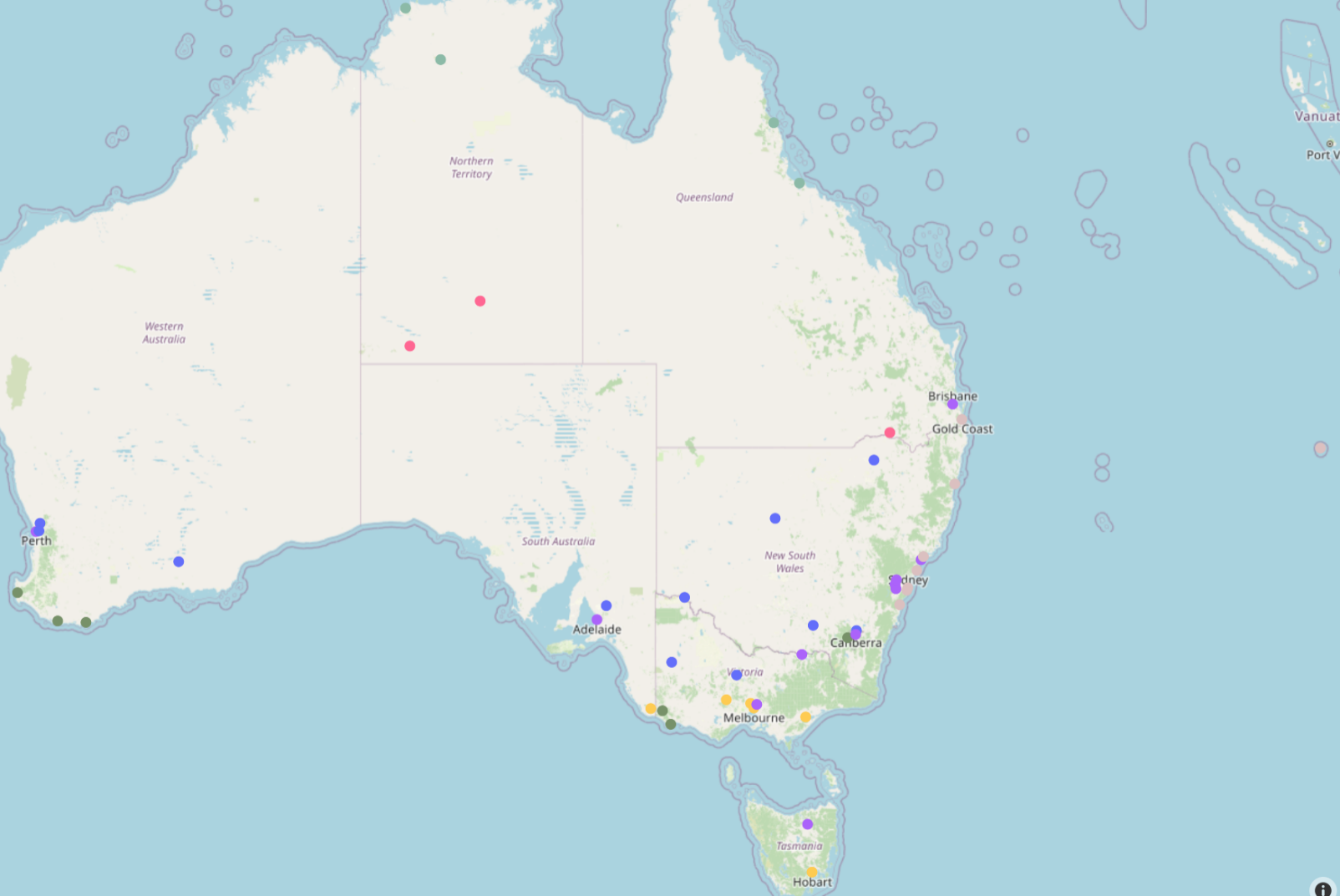
Une approche simple est de calculer la moyenne des valeurs pour chaque ville, puis de tracer le dendrogramme et effectuer une clusterisation par CAH. Nous avons choisi ici de faire 7 clusters (méthode *clusterisation\_groupee()* )

Nous effectuons la clusterisation à partir des données climatiques, et non de la latitude et de la longitude. La proximité géographique des résultats n’est donc pas le fruit de l’exploittaion de ces deux paramètres, mais bien de la cohérence de clusterisation, deux villes proches ayant généralement le même climat.

Le dendrogramme obtenu montre en particulier 2 petits clusters : l’un est composé de Woomera, Alicesprings et Uluru. Il s’agit de 3 villes très arides, deux dont sont au cœur du désert. La second petit cluster est constitué de Cairns, Darwin, Katherine et Townsville. Il s’agit des 4 villes situées le plus au nord.

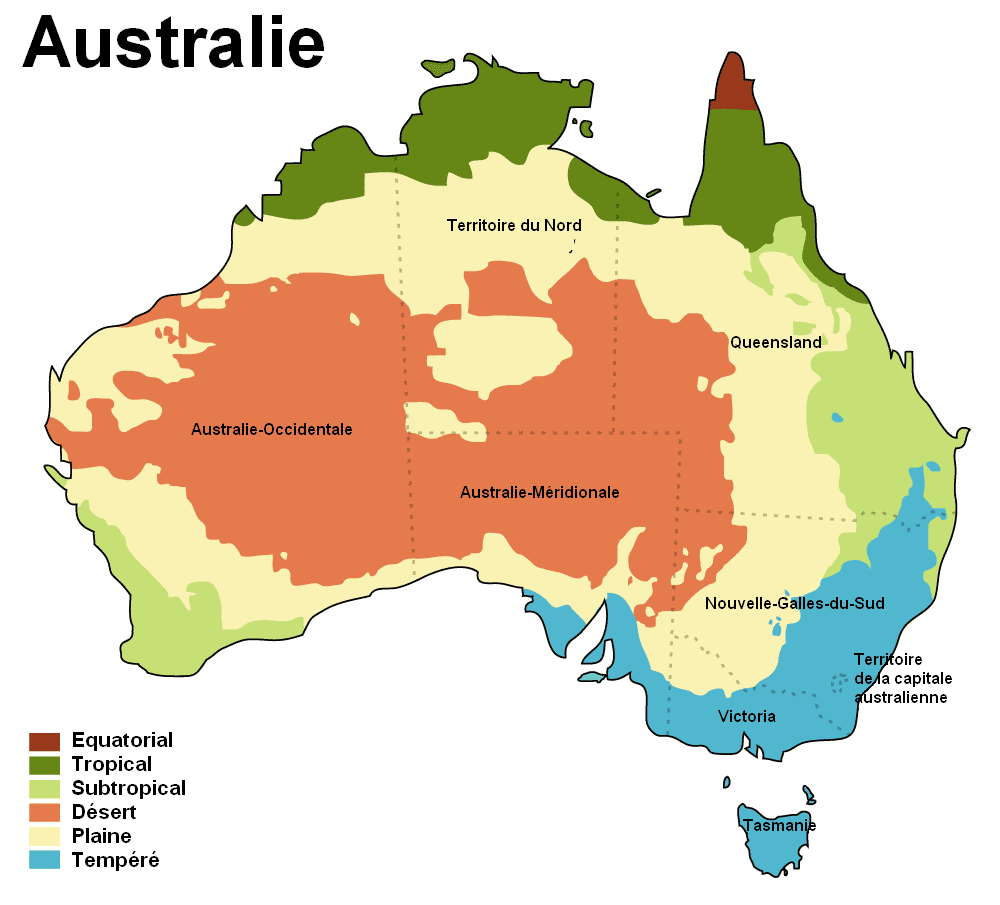


Regardons plus en détail sur la carte comment sont répartis les 7 clusters. Outre les deux exemples cités plus haut, nous retrouvons un groupe sur la côte est, un sur des villes côtières du sud, un autre qui est intermédiaire entre les villes côtières et le désert.



Regardons maintenant une carte représentant les zones climatiques de l’Australie. L’exemple ci-après est issu de la page Wikipédia ‘Climat de l’Australie’.

Zones climatiques australienne (source : Wikipédia : <https://fr.wikipedia.org/wiki/Climat_de_l'Australie> )



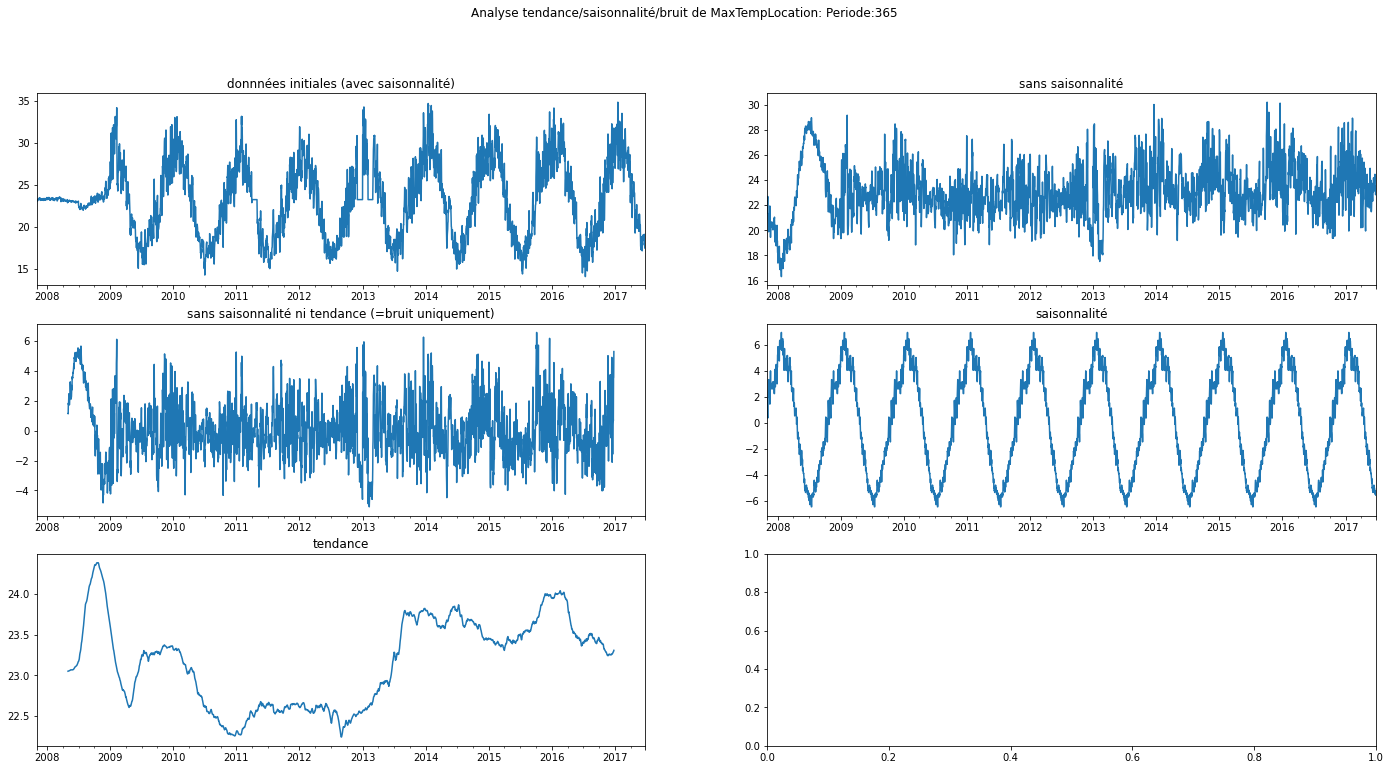
Nous retrouvons bien la cohérence des 4 villes nordiques, correspondant à la zone climatique tropicale. Les villes des plaines forment également un cluster à part entière. Les zones subtropicales et tempérées correspondent aux 4 derniers clusters, mais avec des frontières assez différentes. Le fait qu’un de notre cluster soit spécifique aux villes côtières orientales semble être un signe d’une cohérence intéressante à approfondir.

## Saisonnalité

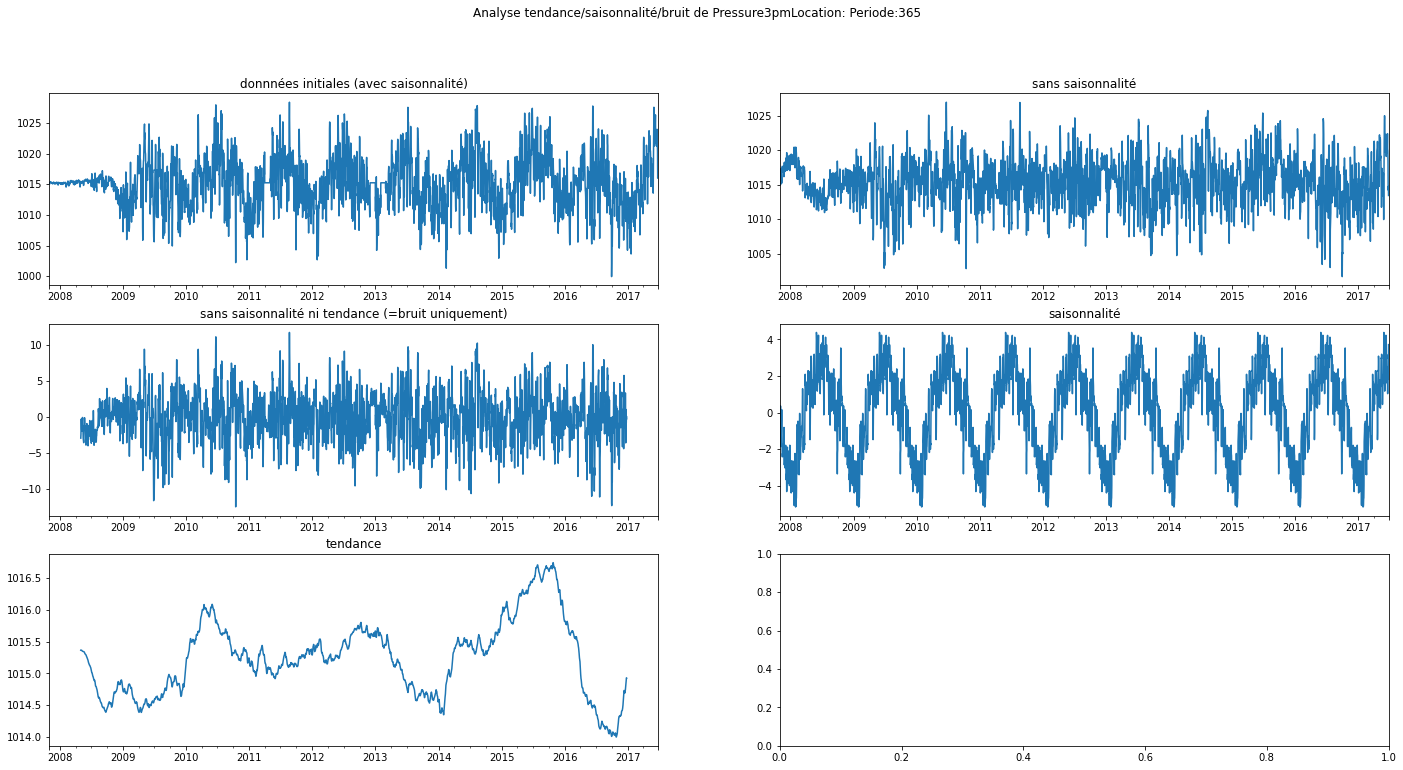
La nature même des données incite à rechercher des saisonnalités dans les différentes variables.

Pour effectuer cette analyse, il convient au préalable de remplir les plages de dates totalement absentes du jeu de données (fonction « reindexation temporelle() )

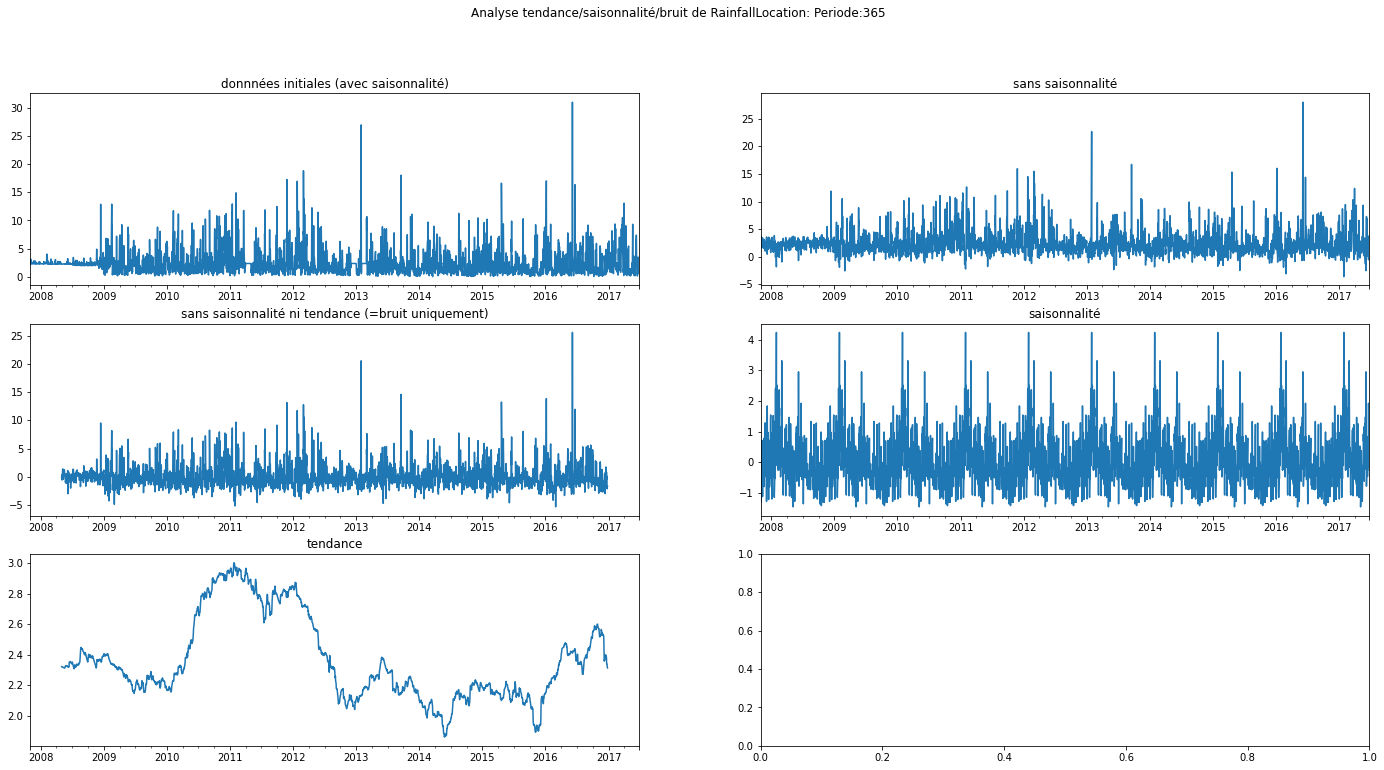
Assez logiquement, la décomposition de la température maximale (MaxTemp) suit un schéma saisonnier de 365 jours. La fonction « test\_max\_saisonalite() » teste plusieurs nombre de journées de saisonalité. Une durée de 365 jour correspond bien à la fois à une variance maximale de la saisonnalité et d’une variation minimale du bruit.



Nous constatons ici que la saisonnalité explique des variations de températures de -6° à +6°, celles-ci variant entre 15° et 35°, soit une amplitude saisonnière de 12° par rapport à une amplitude sur les données initiales de 20°. Ces chiffres encourageants sont à modérer par le bruit restant significatif, puisque variant entre -4° et +6°, soit une amplitude de 10°. Cela implique une probable mauvaise prédiction de cette variable par un modèle de type ARIMA.



Ce même schéma sur la variable de pression atmosphérique à 15h (Pressure3pmLocation) montre des résultats assez similaires : une saisonnalité réelle, mais un résidu restant significatif, et même encore plus important que le poids de la saisonnalité.

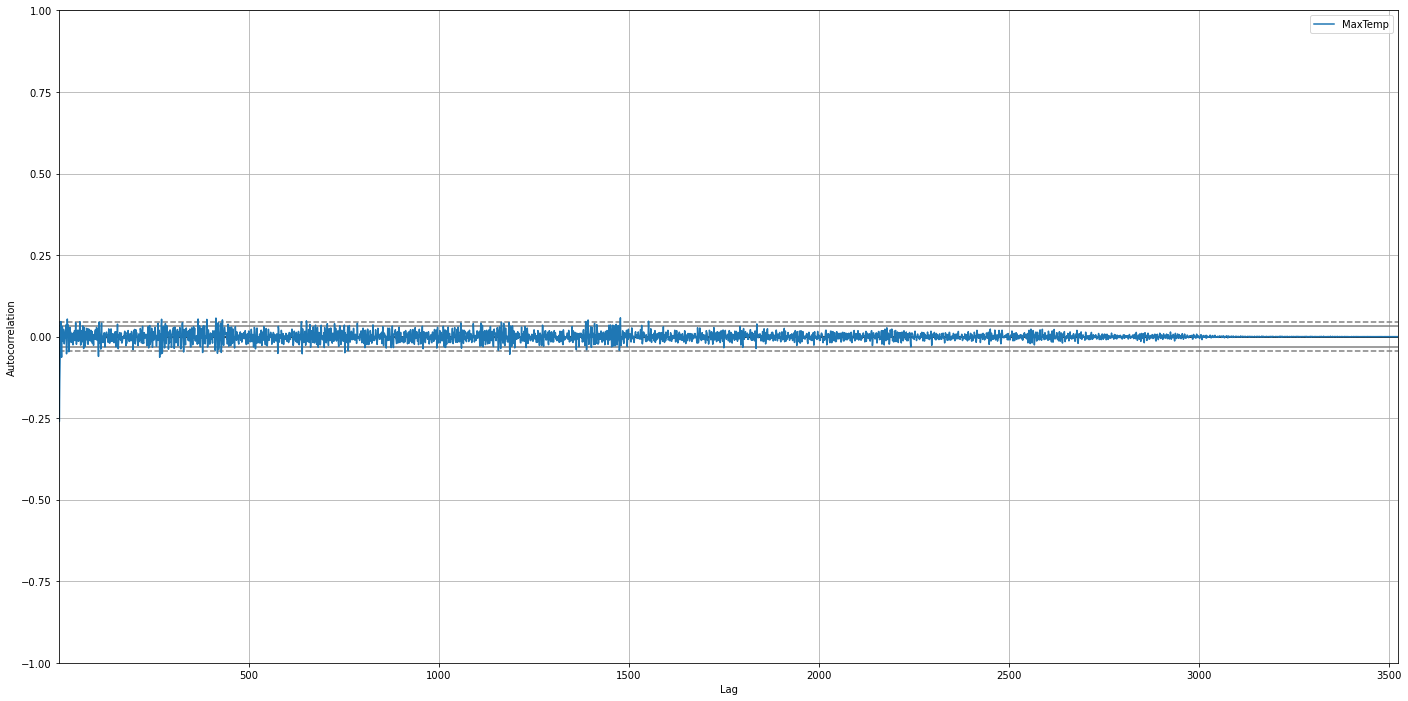


A noter que le niveau de précipitations (Rainfall) semble disposer d’une saisonnalité peu significative au regard du résidu. Cela nous rend plutôt pessimiste sur la qualité de prévisions avec un modèle de série temporelle univariée tel que SARIMA pour notre variable cible « RainTomorrow ».

Autocorrélation

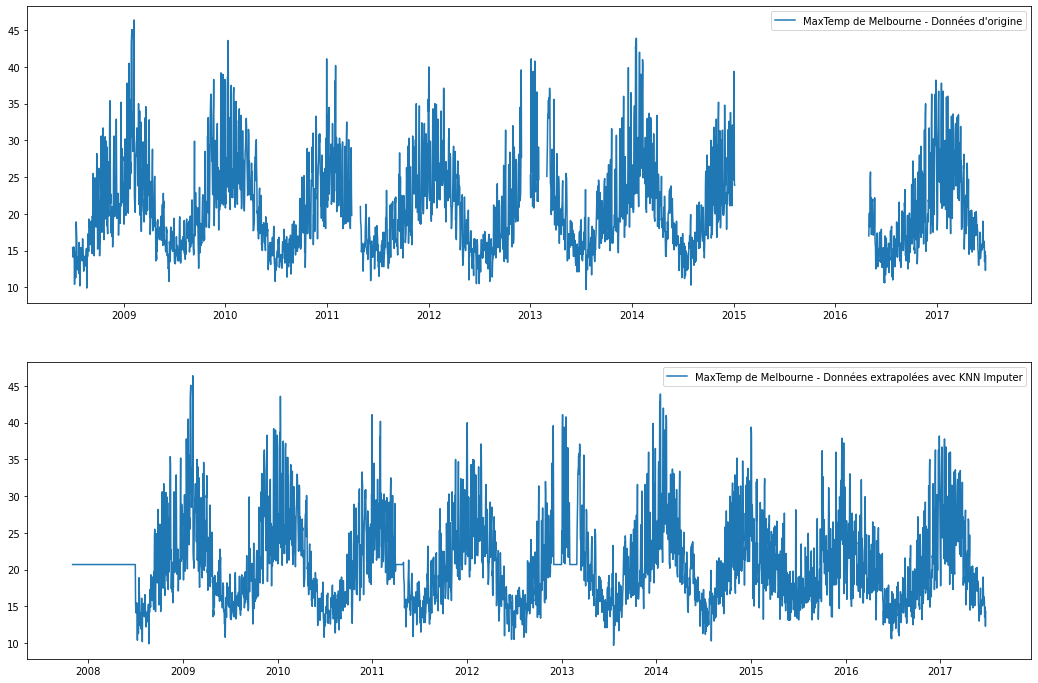
Le graphe d’autocorrélation des variables différenciées (ici MaxTemp) est proche de 0 en tout point.

* Est-ce à dire que la valeur pour une journée ne dépend jamais de celle de la veille ??
* Pourquoi ne voit-on ici aucune saisonnalité ?



## KNN Imputer

L’imputation par KNN nous permet de renseigner les NA d’une façon plus adaptée qu’en reprenant des moyennes. En particulier, cela nous permet de retrouver une saisonnalité et une dispersion des données semblables au jeu de données d’origine. Nous voyons sur le graphe ci-dessous la courbe des température maximales de Melbourne, avec en haut les données d’origines (comprenant plusieurs plages non renseignées) et en bas les données comprenant l’imputation knn. On note que les imputations faites entre 2015 et mi 2016 semblent globalement cohérentes et assez variées. En revanche, celles des petites plages (avril 2011, décembre 2012, février 2013) ont la même valeur. Il en va de même pour les données antérieures à la première valeur renseignée.



## Bilan général

[in progress]

* Enrichissement des données avec AmplitudeTemp, longitude, latitude, climat
* Drop des données redondantes
* Gestion des NA : KNN imputer (évaluer la valeur de knn ainsi que les colonnes pertinentes)