Rapport sur les données météorologiques Australiennes

Table des matières

[1 Introduction 3](#_Toc145269229)

[1.1 Objectifs 3](#_Toc145269230)

[1.2 Contexte historique et enjeux 3](#_Toc145269231)

[1.2.1 Contexte géographique et climatique 3](#_Toc145269232)

[2 Exploration des données et visualisation 3](#_Toc145269233)

[2.1 Kaggle 3](#_Toc145269234)

[2.2 Bureau of Meteorology 4](#_Toc145269235)

[2.3 Variables du dataset 4](#_Toc145269236)

[2.4 Statistiques descriptives 6](#_Toc145269237)

[2.4.1 Location 6](#_Toc145269238)

[2.4.2 RainTomorrow 8](#_Toc145269239)

[2.4.3 Directions du vent 9](#_Toc145269240)

[2.4.4 Variables numériques – Distribution et outliers 12](#_Toc145269241)

[2.4.5 Variables numériques – Analyse temporelle et géographique 16](#_Toc145269242)

[2.4.6 Variables numériques – Corrélations 20](#_Toc145269243)

[2.5 Valeurs manquantes 23](#_Toc145269244)

[2.5.1 Vue globale 23](#_Toc145269245)

[2.5.2 Répartition géographique 27](#_Toc145269246)

[2.5.3 Répartition temporelle 28](#_Toc145269247)

[3 Pre-processing et feature engineering 30](#_Toc145269248)

[3.1 Nettoyage des données 30](#_Toc145269249)

[3.1.1 Traitement des valeurs manquantes 30](#_Toc145269250)

[3.1.2 Traitement des valeurs extrêmes 30](#_Toc145269251)

[3.2 Transformation des données 30](#_Toc145269252)

[3.3 Préparation pour la Modélisation 30](#_Toc145269253)

[3.3.1 Séparation du dataset 30](#_Toc145269254)

[3.3.2 Sélection des variables pour la modélisation 30](#_Toc145269255)

[4 Conclusion 30](#_Toc145269256)

# Introduction

## Objectifs

Ce projet consiste à prédire des variables météorologiques à partir d’un jeu de données contenant dix ans de relevés sur de nombreuses stations météo australienne.

Dans un premier temps, nous tenterons de prédire s’il pleuvra le lendemain (variable RainTomorrow). Nous étendrons ensuite nos prévisions à d’autres variables, telle la température maximale, le niveau de précipitations, ou la vitesse du vent, et nous tenterons d’effectuer des prévisions portant sur plusieurs jours.

## Contexte historique et enjeux

La prédiction des conditions météorologique est un domaine particulièrement ancien, qui a été un enjeu pour de nombreuses sociétés au fil des siècles, et ce dès l’invention de l’agriculture à la préhistoire. Initialement prédite au travers de pratiques divinatoires, les méthodes de prédictions se sont enrichies au fil des siècles : l’importance des nuages a été établie par les babyloniens il y a 8.500 ans, celles des relevés météorologiques par la Chine il y a 5.000 ans, et les innombrables dictons populaires en France témoignent d’une part de la place que le domaine revêt auprès de chacun, d’autre part de la diversité des liens constatés (« Noël au balcon, Pâques au tison », par exemple, indiquant qu’une température élevée fin décembre en impliquerait une faible trois mois plus tard).

Aujourd’hui, il s’agit d’un enjeu économique crucial dans de nombreux secteurs, qu’il s’agisse bien entendu toujours de l’agriculture, mais aussi de l’aéronautique, du tourisme, du BTP, des assurances, etc.

Les prévisions météorologiques ont l’avantage d’être tout à la fois un domaine connu par le grand public depuis de nombreuses années et de mobiliser des techniques poussées pour créer des modèles de qualité.

Nous garderons à l’esprit que les modèles les plus puissants actuels ne permettent que difficilement de prédire de façon fiable au-delà de 7 jours.

### Contexte géographique et climatique

une brève description du climat général de l’Australie, ses particularités régionales

# Exploration des données et visualisation

## Sources de données

### Kaggle

Le data set est celui disponible sur Kaggle pour le projet « Rain in Australia » ( <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package> ).

Ce dataset contient presque 10 ans d’observations météorologiques quotidiennes provenant de plusieurs stations météorologiques australiennes. Ces observations sont des observations météorologiques quotidiennes réalisées à 9h et 15h sur une période de 10 ans, du 01/11/2007 au 25/06/2017.

### Bureau of Meteorology

Le site du Bureau of Meteorology du gouvernement australien ( <http://www.bom.gov.au/climate/data/>

) propose de nombreuses données consultables en lignes. Malheureusement, le site ne permet pas d’obtenir directement un jeu comportant toutes les variables du dataset Kaggle. Il ne permet en effet d’obtenir les mêmes features que le dataset Kaggle que sur les 14 derniers mois. Pour la période couverte par le dataset Kaggle, il n’est possible que de télécharger quelques variables (précipitations et témpératures). Dans tous les cas, ce téléchargement doit s’effectuer pour chaque station météorologique, laquelle n’est pas directement indiquée dans le dataset original.

Au final, utiliser le site du Bureau of Meteorology pour enrichir notre dataset ou bien renseigner des données manquantes ne pourra malheureusement pas se faire de façon simple, ni même par une approche de Webscrapping. Elle ne pourra se faire que très ponctuellement en ciblant certaines variables pour des villes particulières.

## Variables du dataset

Le dataset contient les 23 variables suivantes :

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nom de colonne** | **Unité** | **Explication** |
| 1 | Date | Timestamp | Date d’observation |
| 2 | Location | Chaîne de caractères | Nom du lieu de la station météo |
| 3 | MinTemp | Degrés Celsius | Température minimum en 24 heures jusqu’à 9am |
| 4 | MaxTemp | Degrés Celsius | Température maximum en 24 heures jusqu’à 9am |
| 5 | Rainfall | Millimètres | Précipitation en 24 heures jusqu’à 9am |
| 6 | Evaporation | Millimètres | Évaporation en 24 heures jusqu’à 9am |
| 7 | Sunshine | Heure | Soleil radieux en 24 heures jusqu’à minuit |
| 8 | WindGustDir | 16 points cardinaux | Direction de la rafale de vent la plus forte en 24 heures jusqu’à minuit |
| 9 | WindGustSpeed | Kilomètres par heure | Vitesse de la rafale de vent la plus forte en 24 heures jusqu’à minuit |
| 10 | WindDir9am | 16 points cardinaux | Direction de vent à 9am |
| 11 | WindDir3pm | 16 points cardinaux | Direction de vent à 3pm |
| 12 | WindSpeed9am | Kilomètres par heure | Vitesse de vent à 9am |
| 13 | WindSpeed3pm | Kilomètres par heure | Vitesse de vent à 3pm |
| 14 | Humidity9am | Pourcentage | Humidité relative à 9am |
| 15 | Humidity3pm | Pourcentage | Humidité relative à 3pm |
| 16 | Pressure9am | Hectopascals | Pression atmosphérique réduite au niveau moyen de la mer à 9am |
| 17 | Pressure3pm | Hectopascals | Pression atmosphérique réduite au niveau moyen de la mer à 3pm |
| 18 | Cloud9am | Huitièmes | Fraction de ciel obscurcie par les nuages à 9am |
| 19 | Cloud3pm | Huitièmes | Fraction de ciel obscurcie par les nuages à 3pm |
| 20 | Temp9am | Degrés Celsius | Température à 9am |
| 21 | Temp3pm | Degrés Celsius | Température à 3pm |
| 22 | RainToday | Binaire (Yes, No) | La journée en cours a-t-elle reçu des précipitations supérieures à 1 mm en 24 heures jusqu’à 9h ? |
| 23 | RainTomorrow | Binaire (Yes, No) | Le lendemain a-t-il reçu des précipitations dépassant 1 mm en 24 heures jusqu’à 9am ? |

Tableau 1 : Les variables de l’ensemble de données

La figure représente un Overview (généré par la librairie *ydata\_profiling*) du dataframe de 22 colonnes. L’ensemble de données contient 145 460 d’observations dont il y a 21 observations qui sont redondantes.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Figure 2 : Overview du dataset

Les 22 variables se divisent en 3 types dont

* 4 variables catégorielles : *Location, WindGustDir, WindDir9am et WindDir3pm*
* 2 variables booléennes : *RainToday, RainTomorrow*
* 16 variables numériques.

A noter que deux variables sont directement déduites d’autres informations :

* RainToday est indiquée comme True si Rainfall>1
* RainTomorrow d’une date donnée est égale à RainToday de la date du lendemain, pour Location donnée

Nous avons vérifié et confirmé ces deux affirmations.

## Statistiques descriptives

### Location

Il s’agit d’une variable catégorielle indiquant la localisation de la station météorologique.

Il y a 49 Location au total dont la Figure 1 représente la distribution. Les trois stations météorologiques *Uluru, Katherine* et *Nhil* contiennent environ deux fois moins d’observations que les autres. Les deux stations *Sydney* et *Canberra* contiennent plus d’observations que les autres. Le nombre d’observations des autres stations reste à peu près homogène.

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

Remarquons également que le nombre maximal de journées est de 3418 pour Canberra. La plupart des villes ont un nombre de 3 000 journées environ. Or, nous avons vu que la plage de dates couvre la période du 1/11/07 au 25/06/17, soit 3525 journées. Il manque donc l’équivalent d’environ 1 an et demi de mesures pour la plupart des villes, aucune n’est exhaustive sur la plage de dates. Ce point est partiellement important à prendre en compte pour l’analyse par séries temporelles.

Plusieurs Location possèdent le suffixe « Airport », semblant indiquer que certaines stations météorologiques sont assez proches (« Perth » /  « PerthAirport », « Melbourne » / « MelbourneAirport » et « Sydney » / « SydneyAirport »). Il s’agit d’une information à garder en tête pour d’éventuels renseignement de valeurs nulles.

Nous avons recherché les latitudes et longitudes des 49 Location dans l’objectif de pouvoir les situer sur la carte de l’Australie et mieux comprendre certaines données. Pour cela, nous avons croisé le nom de chaque Location avec une liste de villes australiennes de la Australia Cities Database de Kaggle ( <https://www.kaggle.com/datasets/maryamalizadeh/worldcities-australia> ), ainsi qu’avec la liste des stations météorologiques ( <http://www.bom.gov.au/climate/data/lists_by_element/alphaAUS_139.txt> ). Cette opération a nécessité du travail de vérifications manuelles, du fait d’homonymies (Woomera) ou d’orthographes différents (Nhil versus Nhill)

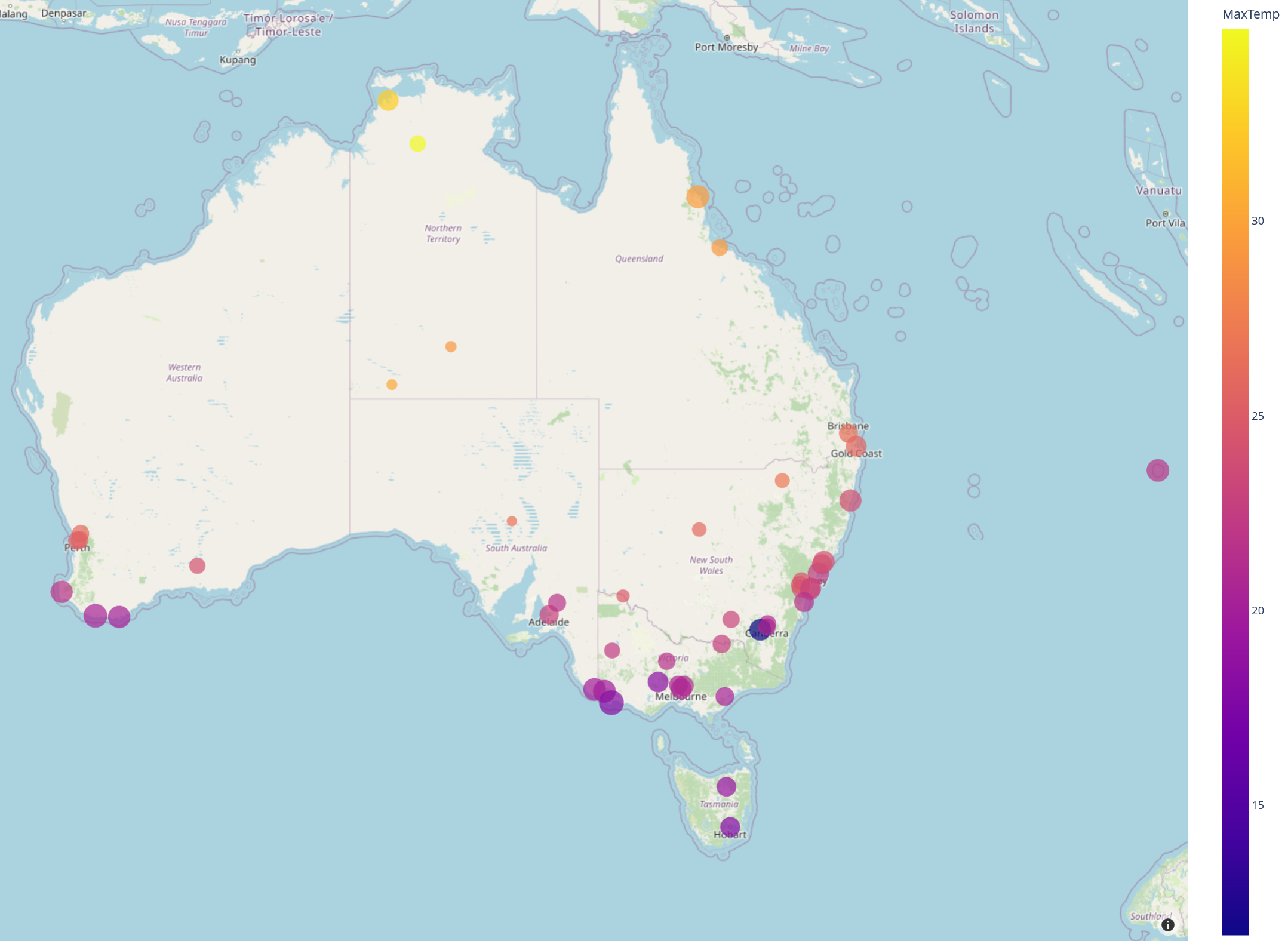


Figure 4 : Location – Couleur : température maximale moyenne – Diamètre : taux de journées pluvieuses

Le graphique ci-dessus montre la répartition des données sur les 49 lieux renseignés. La couleur indique la température maximale 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.

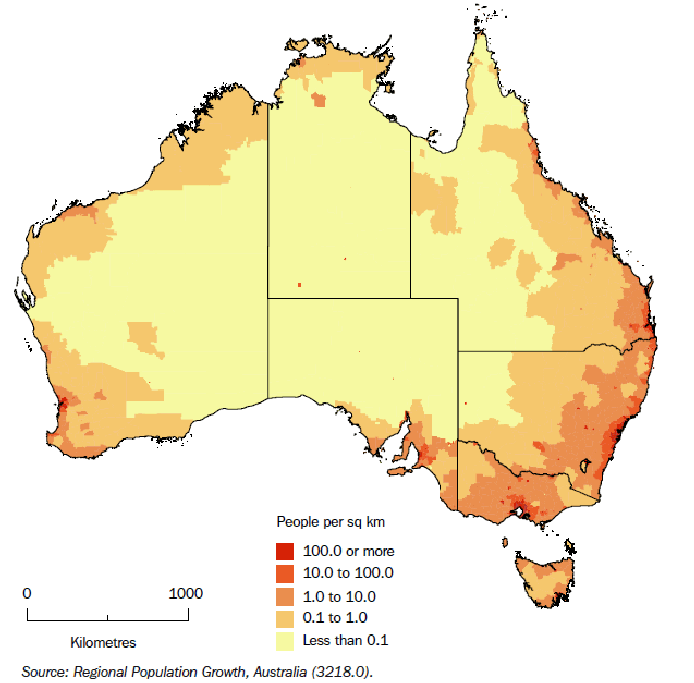
Notons que la ville d’Uluru est située au cœur du désert australien, alors que MountGini est une montagne culminant à 1762m.

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

On remarque également 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 clusters. Enfin, nous voyons que la station de Norfork Island est particulièrement isolée géographiquement, sur une petite île au large oriental.

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, et c’est également conforme à la répartition de la population sur la carte australienne, comme nous pouvons le voir sur cette illustration issue du site de l’Australien Bureau of Statistics :



### RainTomorrow

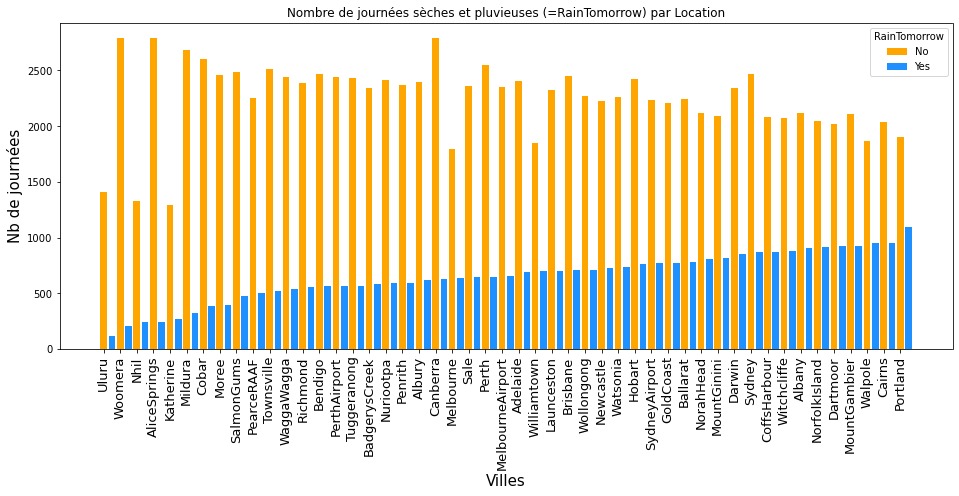
Regardons en premier lieu la variable RainTomorrow, indiquant s’il pleuvra le lendemain. C’est en effet cette variable que nous allons utiliser dans un premier temps comme variable cible pour notre modélisation. Elle revêt donc une importance particulière.

Notons que dans la mesure où RainTomorrow, comme nous l’avons vu plus haut, est égale à RainToday de la veille pour un même lieu, les observations ci-après sont également valables pour RainToday.

Sur l’ensemble du dataset, il y a 2.2% des observations n’ayant pas d’information sur cette variable. On observe 22,4% de journées pluvieuses, donc que le nombre de jours où il ne pleut pas est environ 4 fois plus grand que le nombre de jours où il pleut. Cela nous suggère que dans la phase de modélisation, nous pouvons peut-être mettre en œuvre des techniques d’équilibrage des données pour ne pas confondre notre modèle.

Nous constatons également derrière ce rapport de 1 à 4 sur les modalités de Rainfall se cache une disparité très importante selon les Location. Ainsi, il ne pleut que 6,7% des journées à Woomera, village de 300 habitant situé dans le désert, contre 36,5% à Portland, ville portuaire du sud.

Pas moins de 19 Location sur 49 présentent un taux de journées pluvieuses inférieur à 20%, et 3 ont moins un taux inférieur à 10% (Uluru, Woomera, AliceSprings). Ce déséquilibre très important pour certaines villes va représenter un défi pour notre modèle.



### Directions du vent

Les trois variables WinDir9am, WindDir3pm et WindGustDir sont trois variables catégorielles indiquant la direction du vent, respectivement à 9h00, 15h00, ainsi que pour la rafale de vent la plus forte. Les valeurs possibles sont les 16 directions cardinales (N, NNE, NE, …).

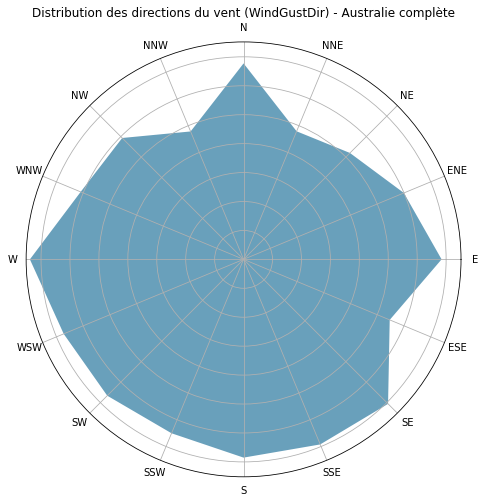
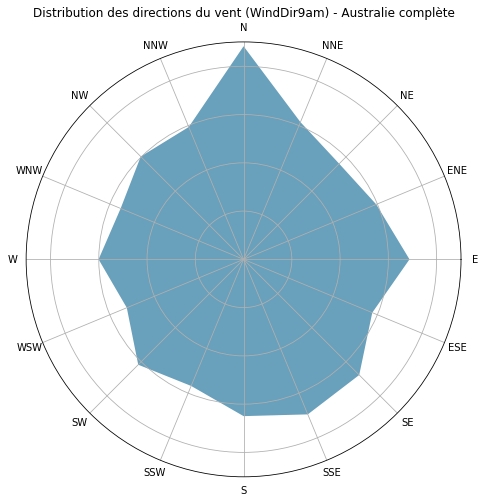
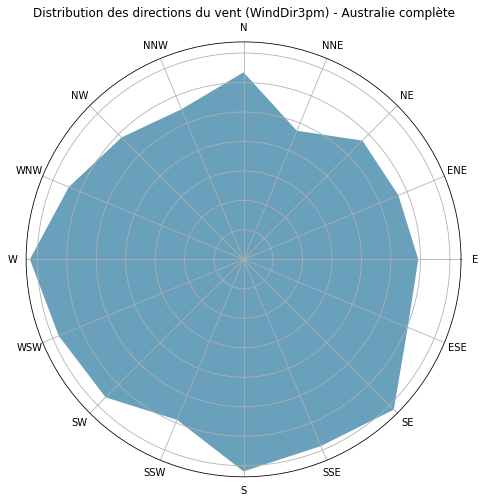
* *WindGustDir* : la direction des rafales de vent est plus fréquemment à l’ouest et moins fréquemment au nord-nord-est
* *WindDir9am*: à 9 heures du matin, la direction du vent est significativement plus fréquemment au nord, et moins fréquemment vers l'ouest-sud-ouest.
* *WindDir3pm* : à 15h, le vent souffle plus fréquemment vers le sud-est et moins fréquemment vers le nord-nord-est (à l’instar de WindGustDir).

Toutefois, la distribution des directions du vent semble assez bien répartie sur l’ensemble du jeu de données.

Une image contenant texte, capture d’écran, Rectangle, ligne

Description générée automatiquement

Figure : Distribution des variables concernant la direction du vent

La distribution de la direction du vent est en revanche radicalement différente selon les Location, comme nous pouvons le voir ci-dessous sur Townsville, Hobart et Bendingo. Il est même étonnant de constater qu’il y a également des distribution très différentes pour une même ville selon la variable observée. Par exemple, on constate que le vent souffle quasiment toujours vers l’ENE à Townsville à 15h00 ainsi que pour les bourrasques, alors qu’il ne souffle que rarement dans cette direction à 9h00 !

Ces trois villes ne sont pas particulières dans le jeu de données. Nous retrouvons le même type de différences dans les distribution du vent pour l’ensemble des Location.

|  |  |  |  |
| --- | --- | --- | --- |
|  | WindGustDir | WindDir9am | WindDir3pm |
| Townsville |  |  |  |
| Hobart |  |  |  |
| Bendigo |  |  |  |

Ces différences peuvent parfois se comprendre en observant simplement la situation géographique du lieu. On voit par exemple que la ville de Hobart se trouve dans un estuaire confiné entre une montagne au nord-est et une autre au sud-ouest, expliquant assez logiquement que les vents ne puissent mécaniquement que circuler vers le nord-ouest ou le sud-est.



Situation géographique de Hobart

Il nous est possible de tester la corrélation de ces 3 variables qualitatives avec RainTomorrow avec un test de χ², avec l’hypothèse nulle supposant qu’il n’existe pas de corrélation.

La méthode « correlation\_vent » calcule la p-value issu du χ² 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.

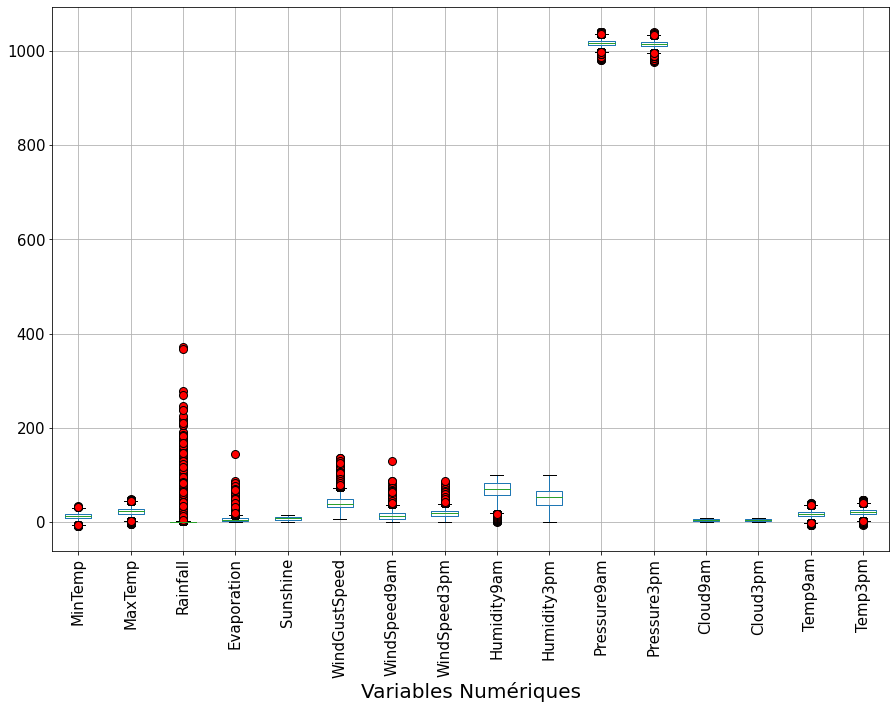
### Variables numériques – Distribution et outliers

La figure suivante représente la distribution de chaque variable numérique. Nous pouvons remarquer que seuls certaines d’entre elles sont distribuées presque normalement, comme *MinTemp*, *Humidity3pm*, *Pressure9am*, *Pressure3pm*, *Temp9am*, *Temp3pm*) tandis que d’autres sont soit asymétriques à droite, soit à gauche.

Une image contenant diagramme, origami, conception

Description générée automatiquement

Distribution des variables numériques



Le diagramme des boîtes à moustaches 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. On voit en particulier que les deux variables de pression atmosphérique ont un ordre de grandeur de 1000 alors que la plupart des autres sont de l’ordre de quelques dizaines.

Une image contenant texte, diagramme, Rectangle, capture d’écran

Description générée automatiquement

Le boxplot des 16 variables quantitatives nous montre que plusieurs variables possèdent un nombre important d’outliers. C’est notamment le cas de Rainfall, dont la boxplot semble montrer que toute valeur non nulle est aberrante. Cela s’explique assez simplement : Rainfall correspond au niveau de précipitations en millimètres. Lorsqu’elle vaut plus de 1, alors RainToday est égale à True. Or, nous avons vu précédemment que seulement 22,4% des lignes ont un RainToday (ou bien un RainTomorrow) à True. Cela implique que dans 77,6% des cas, Rainfall a une valeur inférieure à 1. On comprend alors assez aisément que dès que surgit une averse, le résultat des précipitations enregistré se retrouvera nécessairement en outlier.

En réalité, bien qu’il existe ici de nombreuses valeurs aberrantes d’un point de vue mathématique, il s’agit bel et bien de données réelles, et non de données erronées dans le jeu de données. Nous trouvons par exemple pour les quatre variables concernées des températures comprises entre -7°C et +46°C, ce qui n’a rien d’absurde. Il en va de même pour les autres variables : les outliers de la pression atmosphérique, de la vitesse des vents et des taux d’humidité ont tous des valeurs compatibles avec des données météorologiques correctes.

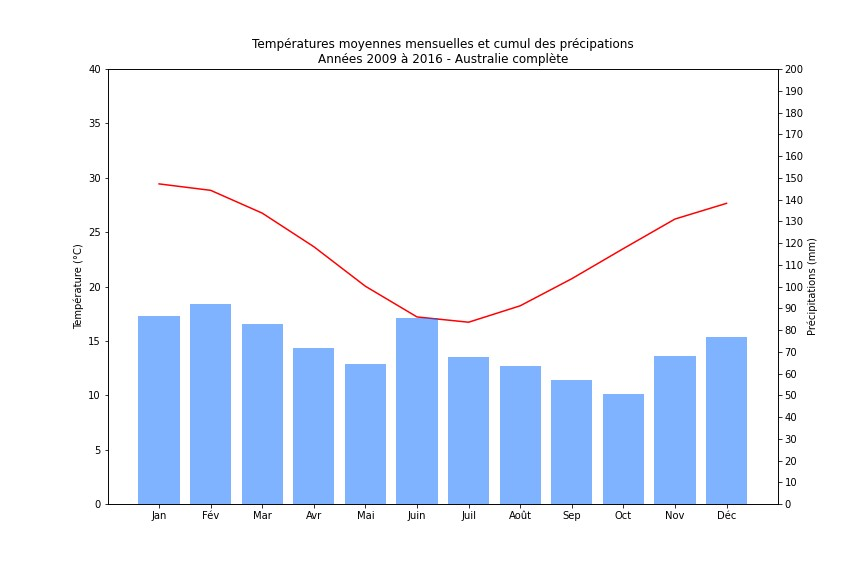
Par conséquent, nous faisons à ce stade le choix de conserver l’intégralité des outliers du jeu de données. Cela impliquera d’être très vigilants sur l’usage de calculs basés sur des moyennes.

### Variables numériques – Analyse temporelle et géographique

Les données sont trop riches pour pouvoir effectuer une visualisation complète de chaque variable selon chaque lieu au fil du temps, aussi nous nous concentrerons ici sur la température maximale et les précipitations. Ce choix n’est pas aléatoire : d’une part il s’agit des deux variables classiquement utilisées pour la représentation de données météorologiques dans le temps, d’autre part le niveau de précipitation permet de déduire logiquement la valeur de RainToday pour un jour donné : prédire Rainfall implique donc de pouvoir prédire RainToday, et possiblement RainTomorrow.

Dans les graphes suivants, nous indiquons la moyenne mensuelle de températures maximales ainsi que le total mensuel des précipitations. Ces données sont effectuées ici uniquement sur huit ans, de 2009 à 2016 inclus, afin de disposer d’années complètes et raisonnablement renseignées.

Il s’agit ici de graphe classique en météorologie. En revanche, l’échelle des précipitations est classiquement graduée avec 2mm pour 1° sur l’échelle des températures : cette proportion d’échelle est un usage habituel pour les pays à climat tempéré, mais n’est pas adapté au climat australien. Nous avons ici choisi plutôt de mettre 5mm pour 1° afin d’avoir des représentations visuelles qui exploitent l’ensemble du schéma pour la majorité des lieux.

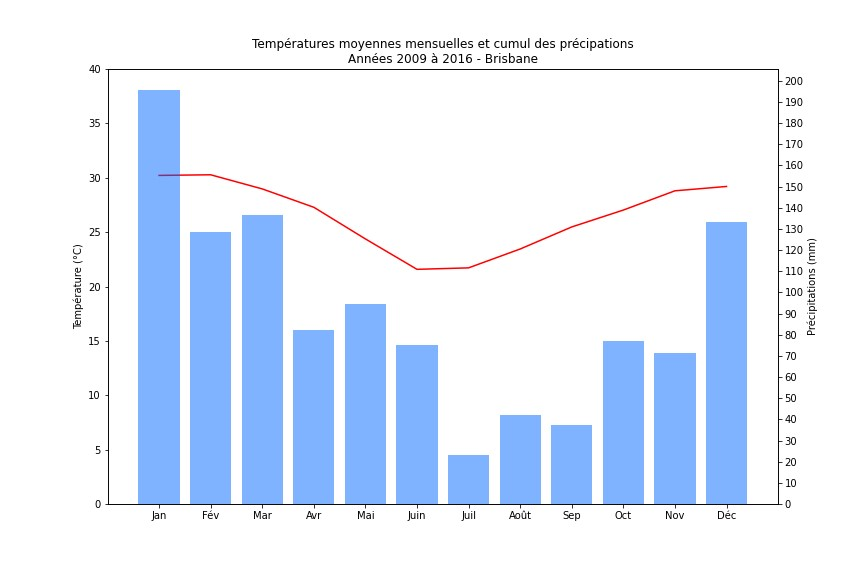
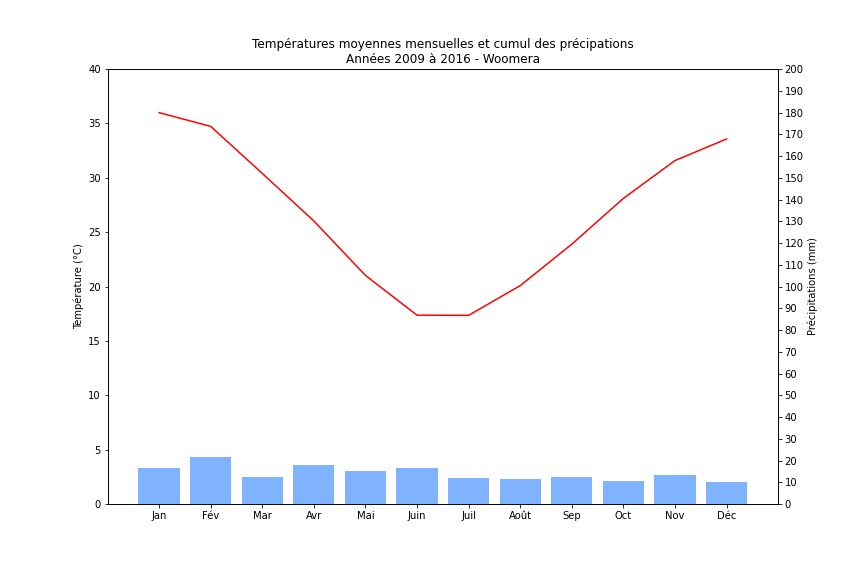


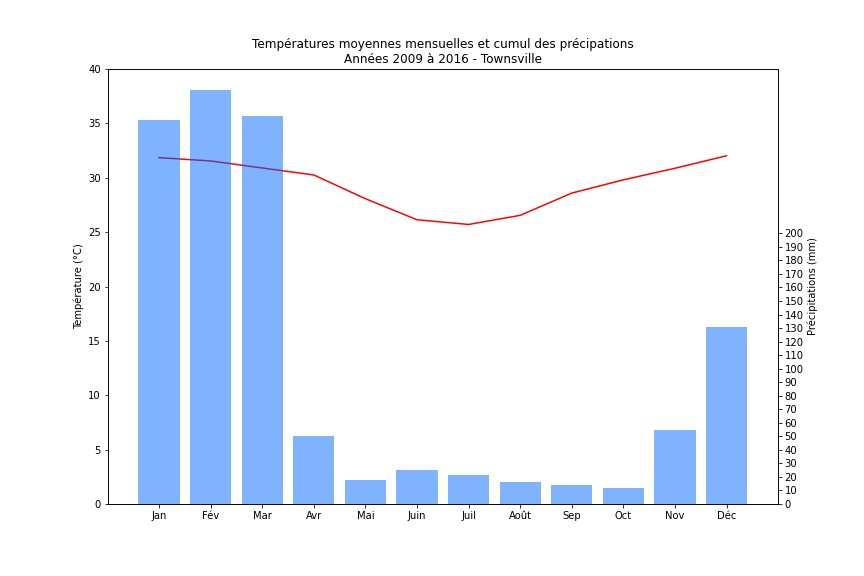
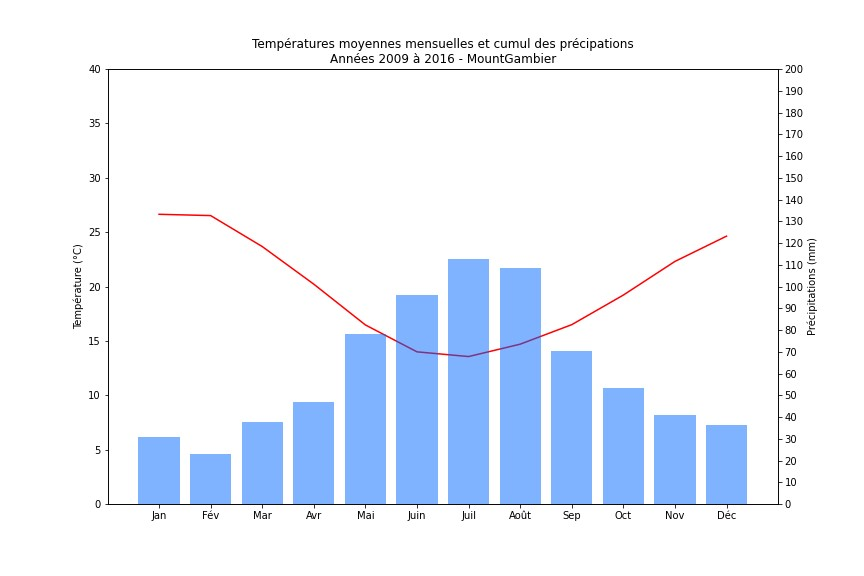
Sur ce premier graphe figurent les données pour l’ensemble des 49 location. On y découvre une température élevée en janvier (30°C) et plus faible en juillet (16°C). Cela s’explique par le fait que l’Australie est située dans l’hémisphère sud. Les saisons sont donc symétriques à celles de la France.

Les précipitations ne semblent visuellement pas corrélées aux températures. Elles varient de 50mm à 90mm en total mensuel, en moyenne par lieu.

Observons maintenant ce même graphique pour 4 lieux très différents. Les échelles allant jusqu’à 40° pour les températures et 200mm pour les précipitations sont identiques pour une meilleure comparaison visuelle des schémas, à l’exception de Townsville qui présente des précipitations exceptionnelles et que nous avons choisie pour cette raison.

Woomera est située dans le désert, Brisbane sur la côte est, Mount Gambier sur le côté sud, Townsville est au nord-est.





Ces quatre graphiques illustrent à quel point il existe plusieurs climats en Australie. Si l’allure de la courbe des températures reste similaire, elle ne présente pas les mêmes valeurs pour chaque ville. Les différences sont tout à fait frappantes concernant les précipitations, avec des niveaux particulièrement faibles à Woomera, une saison pluviale de décembre à mars (été australien) et sèche de juillet à septembre (hiver australien) à Brisbane, des saisons pluviales inversées à Mount Gambier par rapport à Brisbane, des pluies diluviennes à Townsville pendant l’été suivi d’une saison sèche sur le reste de l’année.

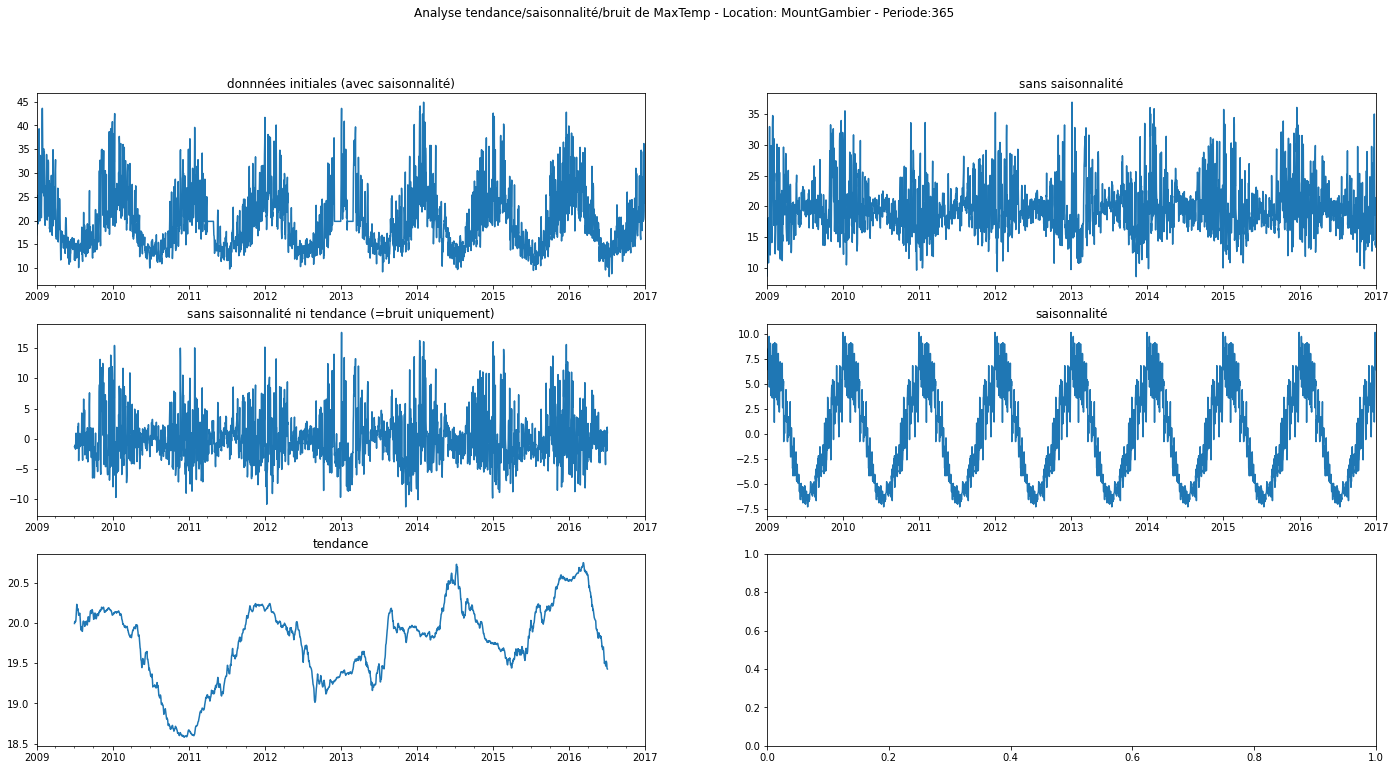
Nous pouvons conclure de cela qu’une modélisation globale sur toute l’Australie pour prédire la météo d’un lieu donné sera probablement trop grossier. En revanche, les villes situées proches géographiquement présentent des caractéristiques communes. Il semblerait donc pertinent d’effectuer des regroupements de villes en fonction de caractéristiques climatiques.

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() )

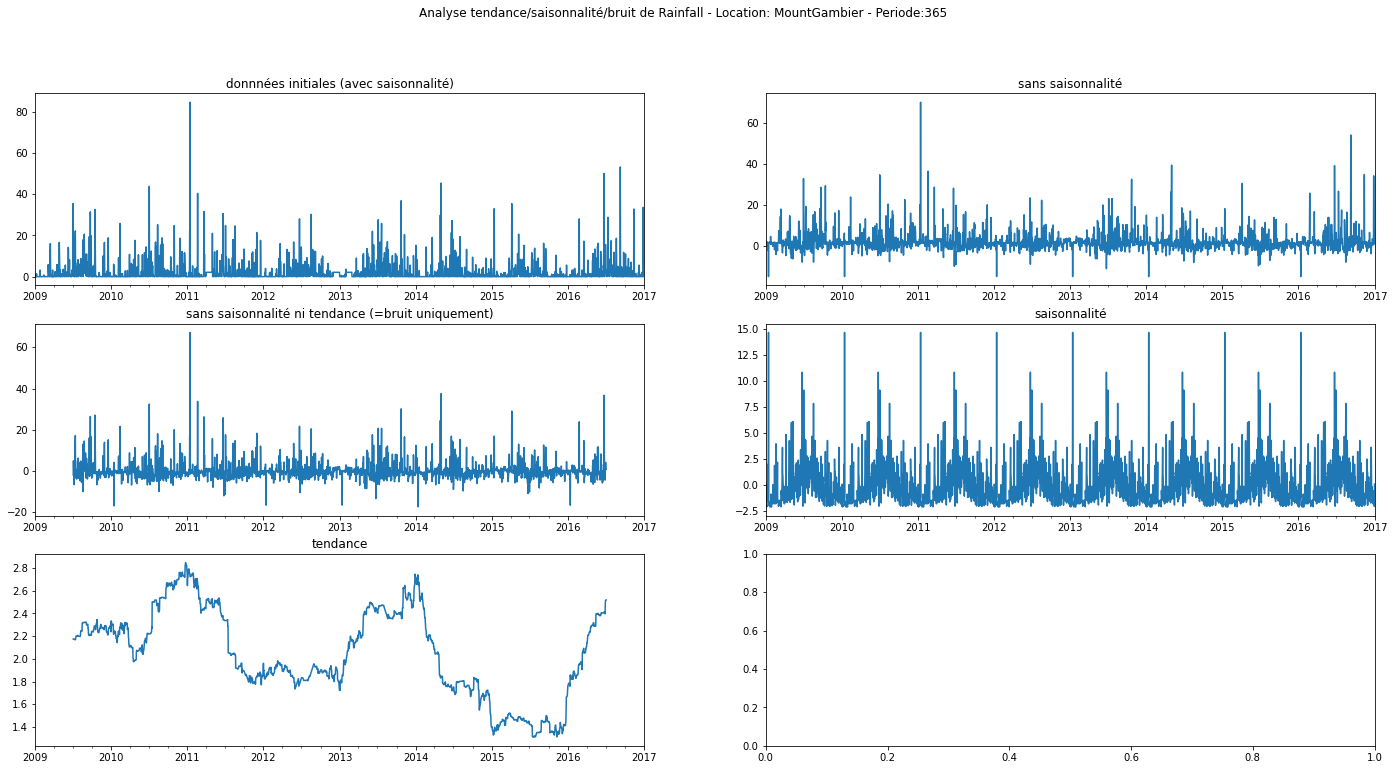
Assez logiquement, la décomposition de la température maximale (MaxTemp) suit un schéma saisonnier de 365 jours. La fonction « test\_max\_saisonalite() » teste plusieurs nombre de journées de saisonalité. Une durée de 365 jour correspond bien à la fois à une variance maximale de la saisonnalité et d’une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

Nous regardons ci-dessous les décompositions saisonnières pour les variables MaxTemp et Rainfall de Mount Gambier. Il ne nous semble pas pertinent de l’effectuer sur l’ensemble de l’Australie, étant donné les constats précédents de différences locales importantes.

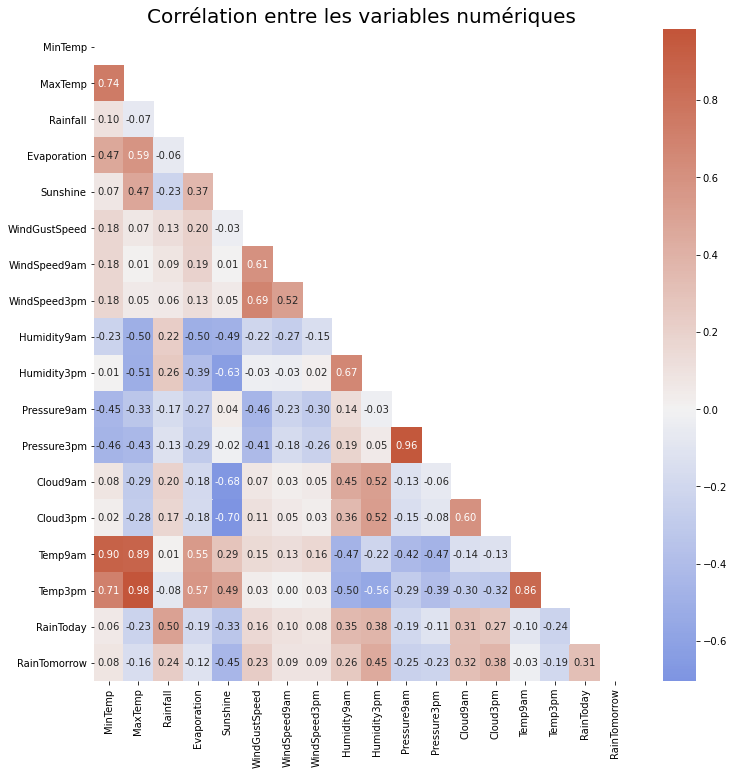


Nous constatons ici que la saisonnalité explique des variations de températures de -7° à +10°, celles-ci variant entre 12° et 40°, soit une amplitude saisonnière de 17° par rapport à une amplitude sur les données initiales de 28°. Ces chiffres encourageants sont à modérer par le bruit restant significatif, puisque variant entre -7° et +12°, soit une amplitude de 19°. Cela implique une probable mauvaise prédiction de cette variable par un modèle de type ARIMA.

Ce même schéma sur Rainfall montre des résultats assez similaires : une saisonnalité réelle, mais un résidu restant significatif, et même plus important en variation que le poids de la saisonnalité. 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 ».



### Variables numériques – Corrélations



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, comme nous le verrons après, il s’agit malheureusement de deux features ayant un taux élevé de données nulles.

A l’inverse, RainTomorrow semble très peu corrélée aux températures (de 0,03 à 0,19).

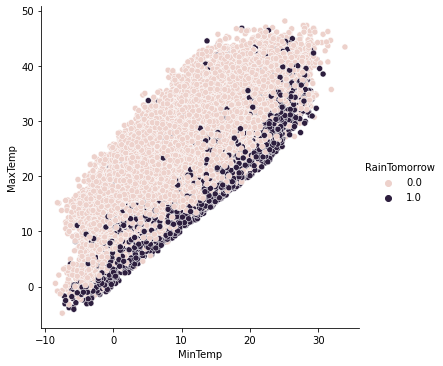
La vitesse du vent à 9h00 et 15h00 est elle aussi très peu corrélée avec RainTomorrow (0,09). La vitesse des rafales l’est en revanche davantage (0,23).

De ces constats il ressort d’une part que la prédiction de RainTomorrow ne pourra se concentrer de prime abord sur une seule feature explicative, et également qu’il devrait être possible de supprimer certaines variables relatives à la vitesse du vent et aux température du fait de leur faible corréaltion.

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 (0,89), ainsi que Temp9am et Temp3pm. Les températures min et max présentent aussi une corrélation intéressante (0,74). Bref, nous constatons que l’ensemble des variables de températures présentent une très forte corrélation entre elles.

C’est aussi le cas de la pression : les variables Pressure3am et Pressure9pm sont très fortement corrélées (0,96).

En regardant le graphique des pairplot entre chaque variable numérique, nous pouvons constater un lien intéressant entre les températures Minimales et Maximales.

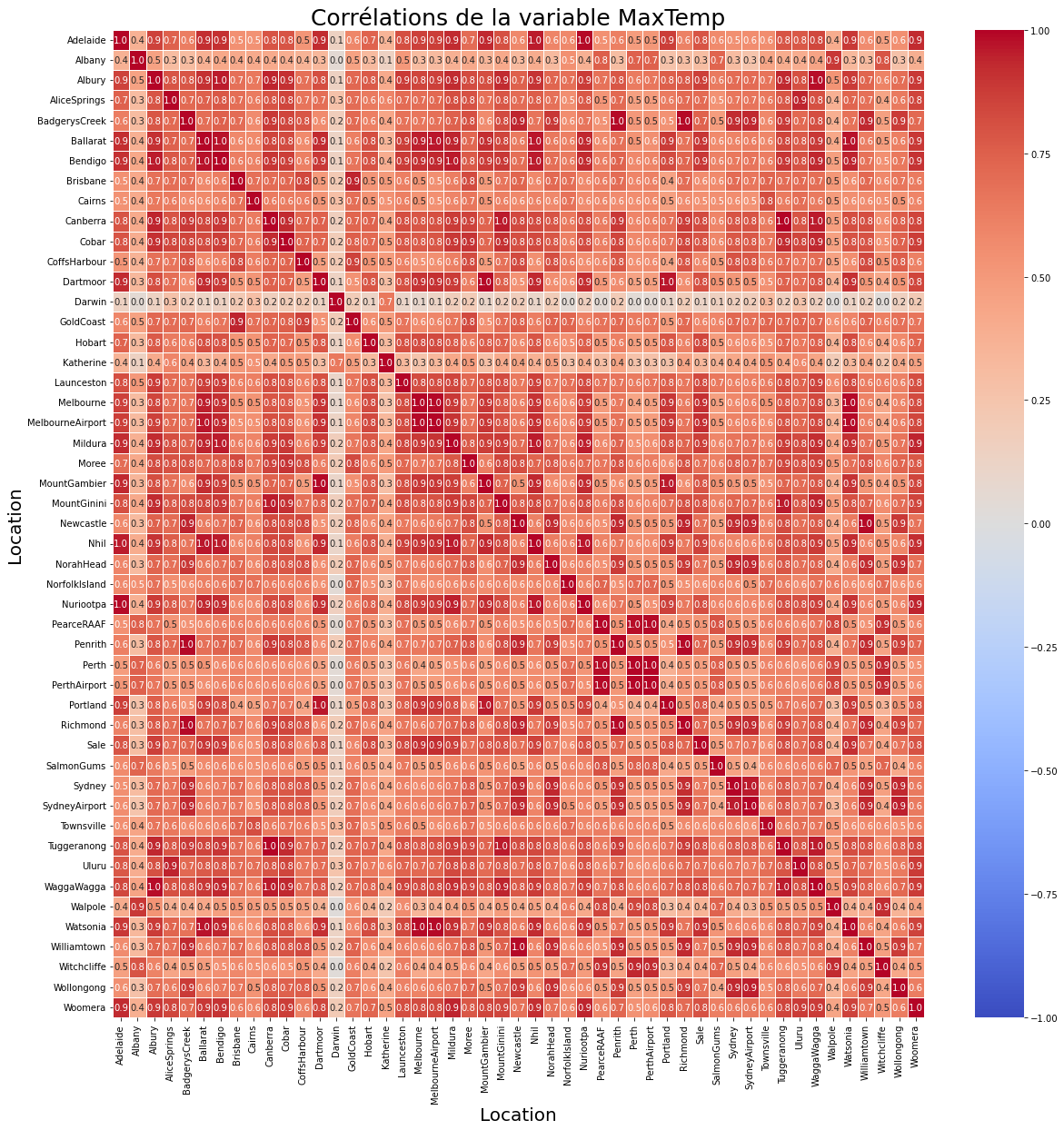


Le graphique ci-dessus trace en bleu les journées avec un RainTomorrow positif, et les positionne sur un graphe aux coordonnées (MinTemps, MaxTemp). Il est frappant de constater que les points bleus sont largement positionnés sur la première bissectrice, ce qui signifie qu’une très faible amplitude thermique est 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 0,19 avec chaque variable de température individuellement, elle l’est à 0,33 avec cette nouvelle variable !

Notons aussi que cette nouvelle variable est corrélée à 0,75 avec Humidity3pm, 0,58 avec Sunshine et 0,53 avec Cloud9am. Là aussi, il s’agit de valeurs supérieures à ce que nous avions avec les variables initiales de température.

Il est également intéressant d’observer les corrélations entre différents lieux pour une même variable, que ce soit pour clusteriser les villes par climats ou pour trouver des villes similaires dans l’objectif de remplacer des données manquantes.

Nous voyons 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.



## Valeurs manquantes

### Vue globale

Après avoir compris l’ensemble de données, nous devons rechercher les valeurs manquantes. Les valeurs manquantes dans un ensemble de données jouent un rôle très important dans un projet. Si elles ne sont pas traitées, les résultats risquent de ne pas être pertinents.

Les données manquantes peuvent être divisées en 3 catégories :

* **Missing Completely at Random (MCAR)**: ce sont des données qui manquent complètement au hasard. C’est-à-dire que les valeurs manquantes n’ont aucune corrélation avec d’autres valeurs de l’ensemble de données observées ou manquantes.
* **Missing at Random (MAR)**:
* **Not Missing at Random (NMAR)**:

La figure suivante représentant le pourcentage de valeurs manquantes pour chaque variable montre que toutes les variables contiennent des valeurs manquantes, sauf *Location* qui est complete. Nous pouvons voir que les quatre variables *Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm* comportent un grand nombre de valeurs manquantes.

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

Figure : Pourcentage de valeurs manquantes pour chaque variable

Ci-après nous voyons une matrice des valeurs manquantes de chaque variable. La couleur de chaque cellule de la matrice est basée sur l’existence ou non des données. Si la couleur est noire, les données existent. Si la couleur est blanche, les données sont manquantes. A partir de ce graphique, nous avons une image de la proportion de données manquantes dans une ligne (observation) ou une colonne (variable).

Comme le montre le graphique résultant, les colonnes *Evaporation*, *Sunshine*, *Cloud9am* et *Cloud3pm* affichent de grandes parties de données manquantes. Cela a été identifié dans le graphique à barres ci-dessus, mais l’avantage supplémentaire est qu’on peut voir comment ces données manquantes sont distribuées dans le dataframe.

Sur le côté droit de la matrice se trouve une sparkline qui va de 0 à gauche au nombre total de colonnes dans le cadre de données à droite. Lorsqu’une ligne a une valeur dans chaque colonne, la ligne sera à la position maximale à droite. A mesure que les valeurs manquantes commencent à augmenter dans cette ligne, la ligne se déplacera vers la gauche. On peut observer qu’il y a des lignes (observations) contenant un grand nombre de valeurs manquantes.

Une image contenant croquis, art, bâtiment, noir et blanc

Description générée automatiquement avec une confiance moyenne

Matrice des valeurs manquantes

Nous pouvons également nous intéresser au taux de variables manquantes non pas par colonnes, mais par ligne, c’est-à-dire par observation. Ce graphe nous permet de constater que 1,92% des lignes possèdent plus de la moitié des informations manquantes, ce qui les rendra difficilement exploitables.

Une image contenant texte, capture d’écran, affichage, diagramme

Description générée automatiquement

Pourcentage de valeurs manquantes pour chaque observation

Au total, 41% des lignes possèdent au moins une variable nulle. Il ne semble donc pas envisageable de supprimer toutes ces lignes, et des solutions de remplacement de valeurs manquantes devront être déployées. Pour cela, avant d’envisager une solution basée sur l’exploitation d’autres features, il nous faut connaitre la corrélation de nullité entre les différentes variables. C’est ce que montre le graphique siuvant.

Une image contenant capture d’écran, diagramme, conception

Description générée automatiquement

Corrélation de nullité entre les variables

* Les valeurs proches de 1 indiquent que la présence de valeurs manquantes dans une variable est corrélée à la présence de valeurs manquantes dans une autre variable.
* Les valeurs proches de -1 indiquent que la présence de valeurs manquantes dans une variable est anti-corrélée à la présence de valeurs manquantes dans une autre variable. Autrement dit, lorsque des valeurs manquantes sont présentes dans une variable, des valeurs de données sont présentes dans l’autre variable et inversement.
* Les valeurs proches de 0 indiquent qu’il y a peu ou pas de relation entre la présence de valeurs manquantes dans une variable et dans une autre.

Nous pouvons voir dans l’ensemble de données que la variable *WindGustSpeed* et la *WindGustDir* ont une corrélation de 1, ce qui souligne que si la valeur de *WindGustSpeed* est manquante, la valeur de *WindGustDir* sera également manquante. On observe le même effet entre *RainToday* et *Rainfall*. Ce dernier point avait été précédemment vérifié lorsque nous avions regardé si RainToday était bien égal à True lorsque Rainfall est supérieure à 1 : il ne sera donc malheureusement pas possible de renseigner RainToday grâce à Rainfall, à moins d’enrichir le jeu de données initial par des données sur Rainfall complémentaires.

A partir de ces corrélations, nous pouvons représenter un regroupement hiérarchique des variables qui ont de fortes corrélations de nullité. Si plusieurs variables sont regroupées au niveau zéro, la présence de valeurs manquantes dans l’une de ces variables est directement liée à la présence ou à l’absence de valeurs manquantes dans les autres colonnes. Plus les variables sont séparées dans l’arbre, moins les valeurs manquantes sont susceptibles d’être corrélées entre les variables. Dans le graphique de dendrogram, nous pouvons voir qu’il y a deux groupes distincts. Le premier se trouve sur le côté gauche (*Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm*) qui ont tous un degré élevé de la valeur manquante. La seconde est à droite, avec le reste des variables qui sont plus complètes.

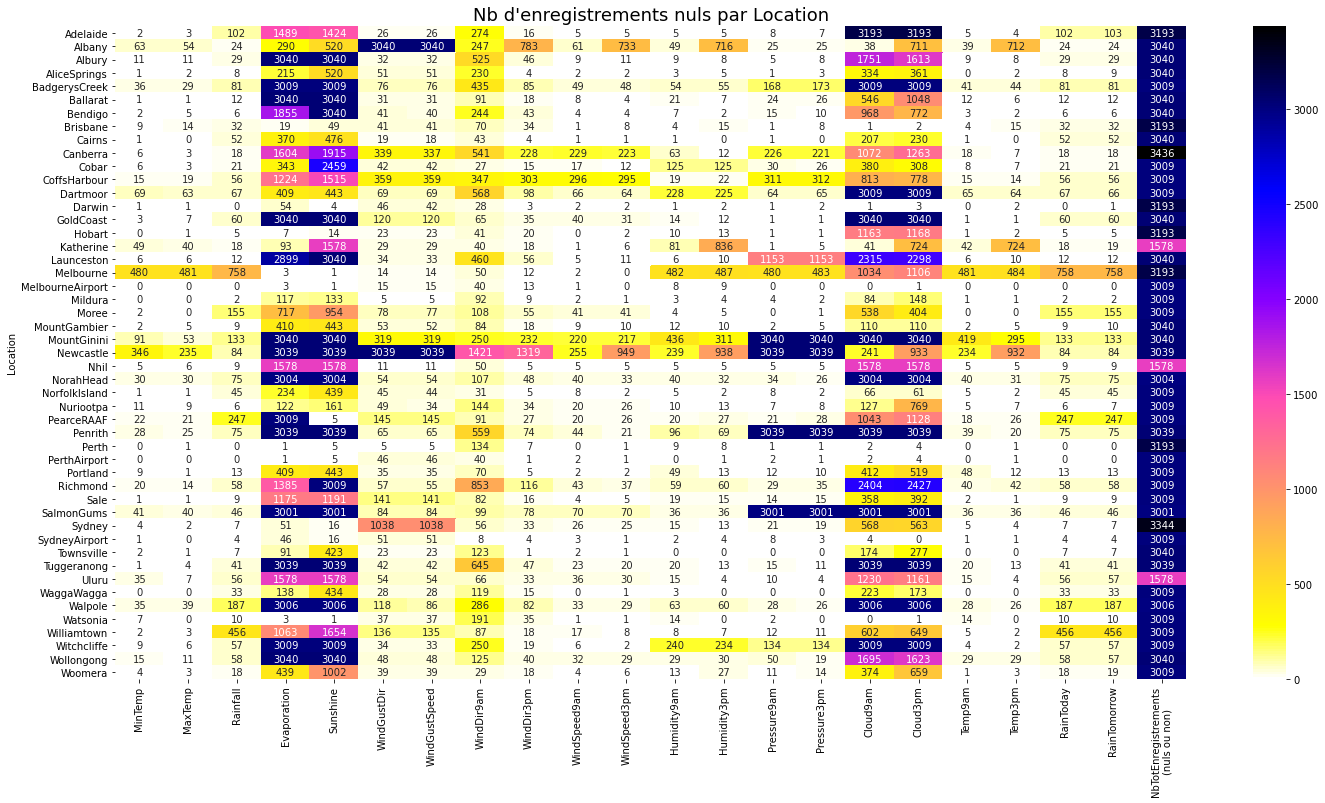
Une image contenant croquis, diagramme, dessin, Dessin technique

Description générée automatiquement

Dendrogramme de nullité entre les variables

### Répartition géographique

Regardons maintenant les données manquantes par lieux. Chaque ligne du prochain graphique indique, pour chaque Location, le nombre d’enregistrements nuls de chaque variable. La dernière colonne indique le nombre d’enregistrements total (nuls et non nuls).



Nous avions déjà identifié précédemment que les quatre variables Evaporation, Sunshine, Cloud9am et Cloud3pm ont un taux élevé de données manquantes. Ce graphe nous montre que cela dépend en réalité énormément des villes. Ainsi, ces variables sont totalement absentes pour certains lieux, et presque toujours renseignées pour d’autres !

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). La question du maintien de ces villes dans le jeu de données se pose pour la modélisation.

Enfin, nous voyons que certaines villes expliquent à elles seules un nombre important de valeurs manquantes sur une variable, comme pour WindGustDir, dont une grande partie s’explique par Albany, Newcastle et Sydney. La ville de Sydney étant par ailleurs plutôt bien renseignée, on privilégiera ici une solution consistant soit à chercher à retrouver cette information dans des relevés météo, soit à tenter de la reconstituer par des approches que nous aborderons un peu plus loin.

### 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.

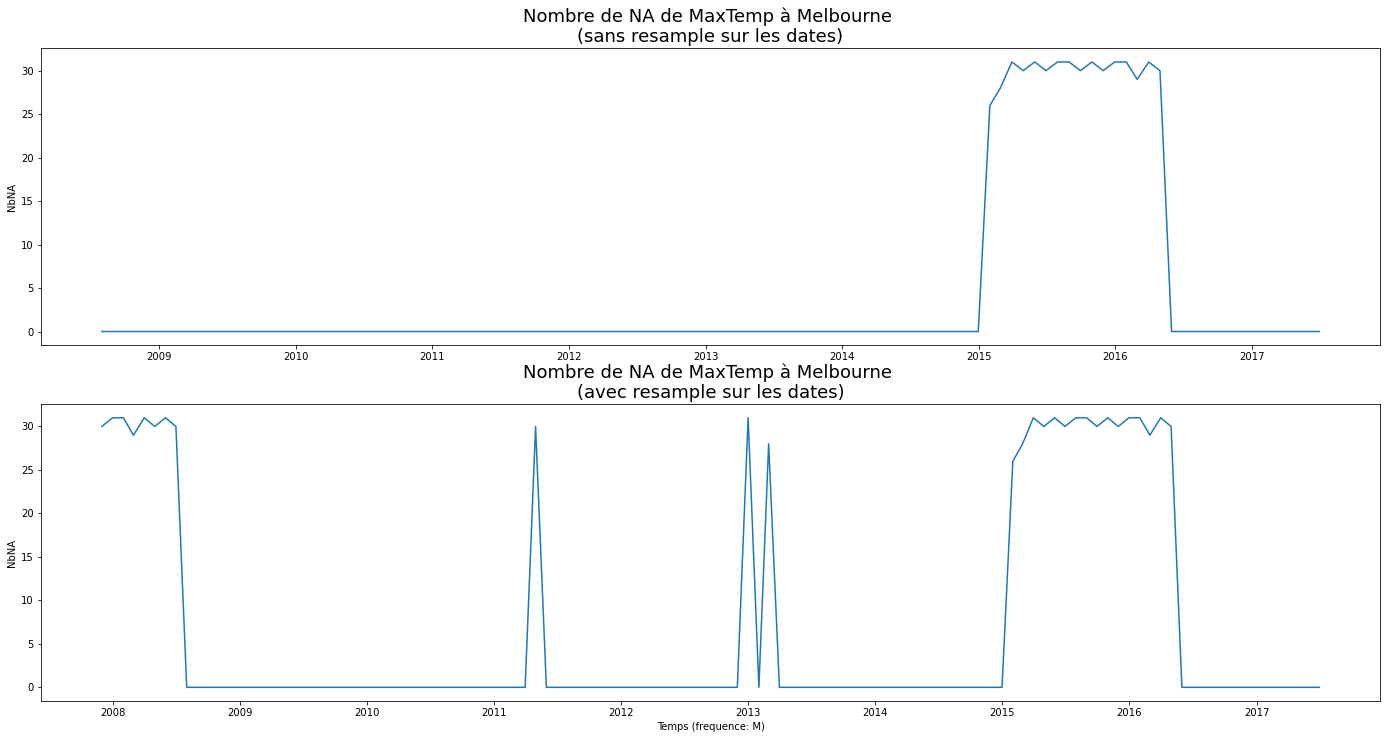
Afin de pouvoir analyser la répartition des données temporellement, il faut préalablement ajouter les journées absentes du jeu de données pour chaque Location. En effet, en l’absence de cette étape, seules les journées connues dans le dataset génèreront un NA pour une variable. Les journées absentes 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. Par conséquent, MaxTemp n’est ici pas seule concernée : si une journée est manquante pour un lieu donné, il va de soi que l’intégralité des variable est manquante pour cette période. Nous pouvons ainsi déduire de ces deux graphes que, pour la ville de Melbourne, il n’existe aucune variable avant mi 2008, ainsi qu’en avril 2011, décembre 2012 et février 2013.

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



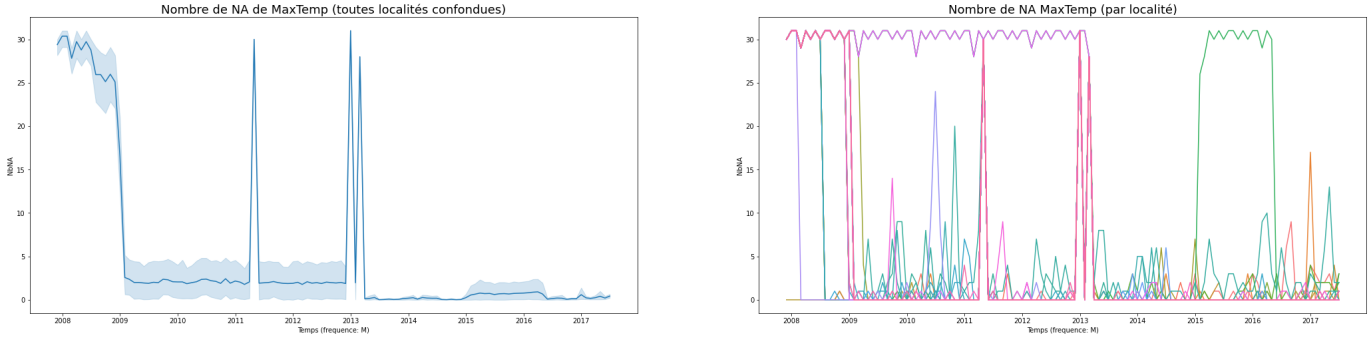
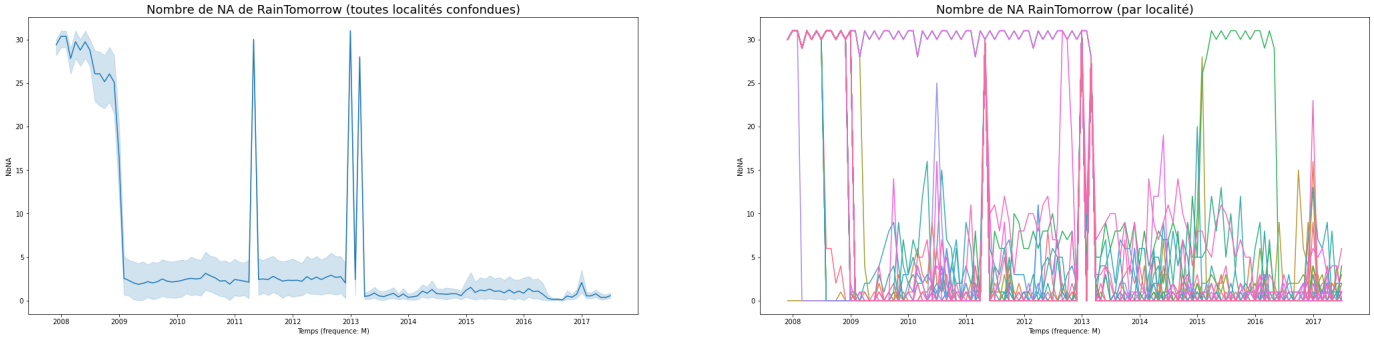
Tentons maintenant de représenter les valeurs manquantes simultanément pour tous les lieux et selon le temps. Le premier graphe illustre les données manquantes pour RainTomorrow, le second pour MaxTemp.

En première colonne, les graphes représentent le nombre de journées pour chaque mois pour lesquels la variable est NA, toutes Location confondues. En seconde colonne, le graphe représente la même chose, mais avec cette fois une courbe par Location. Ces 49 courbes superposées sont évidemment difficillement lisibles en détail mais permettent de montrer quelques tendances intéressantes.

MaxTemp et RainTomorrow sont globalement rarement disponibles avant 2009. Nous voyons aussi qu’il n’existe aucune donnée pour aucune Location pour le mois d’avril 2011. Il en va de même des mois de décembre 2012 et février 2013. Ces remarques sont vraies pour l’intégralité des features.

Sur MaxTemp, il y a une variance de nullité 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. La variance de nullité reste globalement importante sur toutes les périodes pour RainTomorrow.

Nous voyons également qu’il y a des Locations qui ont des mois entiers (voire des années !) sans MaxTemp ni RainTomorrow renseigné. C’est par exemple le cas pour notre tro de villes peu renseignées déjà vu précédemment, Nihl, Katherine et Uluru, qui ne disposent d’aucune donnée avant 2013. Enfin, nous voyons qu’il n’y a jamais de mois pendant lequel MaxTemp ou RainTomorrow serait disponible intégralement pour l’ensemble des Location.



Nous n’allons pas reprendre ici ces graphes pour l’ensemble des variables, mais nous les avons observés (fonction « analyse\_variables\_temps() »). Le graphe de gauche 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 graphe de droite, 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 intégralement absents des données (avril 2011, décembre 2012, février 2013) : si certaines variables présentent 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.

# Pre-processing et feature engineering

## Nettoyage des données

* Remplacer les valeurs de deux variables booléennes (Yes -> 1 et No -> 0)

### Traitement des valeurs manquantes

* Méthodes utilisées
* Résultats et impact

### Traitement des valeurs extrêmes

* Explication des raisons pour lesquelles qu’on a tendance de ne pas les supprimer.
* Correction (convertir les données des

## Transformation des données

* Encodage des variables catégorielles : décrire la méthode dont les variables catégorielles ont été traitées. (Quyen utilise la méthode get\_dummies de pandas pour encoder les variables catégorielles)
* Suppression des variables ou création de nouvelles variables (Sophie a expliqué ce point dans son rapport).
* Normalisation/Standardisation : techniques choisises et raisonnement (Quyen utilise la méthode sklearn.preprocessing.StandardScaler)

## Préparation pour la Modélisation

### Séparation du dataset

* Comment les données ont été divisées en ensembles d’entraînement, de validation et de test (train-test-split, validation croisée…)
* Proportions

### Sélection des variables pour la modélisation

* Quelles variables seront utilisées pour la modélisation
* Justifications

# Conclusion

Résumer des principales découvertes et des difficultés rencontrées