

Rapport sur les données météorologiques Australiennes

Formation : Data Scientist (05/2023 – 04/2024)

Encadrant : Francesco MADRISOTTI

Réaliseurs : Sophie BERTHIER

Luciano LANGHI

Quyen THIEU MARCAUD

Le 14/09/2023 - Définitif

Table des matières

1	Introduction	4
1.1	Objectifs	4
1.2	Contexte historique et enjeux.....	4
1.2.1	Contexte géographique et climatique	4
2	Exploration des données et visualisation.....	5
2.1	Sources de données	5
2.1.1	Kaggle.....	5
2.1.2	Bureau of Meteorology	6
2.2	Variables du dataset	6
2.3	Statistiques descriptives	7
2.3.1	Variables catégorielles	8
2.3.2	Variables numériques	11
2.4	Corrélation	14
2.5	Analyse détaillée des variables	16
2.5.1	RainTomorrow.....	16
2.5.2	Location, MaxTemp et RainTomorrow	16
2.5.3	Analyse temporelle et géographique	18
2.6	Valeurs manquantes	21
2.6.1	Vue globale	21
2.6.2	Répartition géographique.....	25
2.6.3	Répartition temporelle	26
3	Pre-processing et feature engineering.....	28
3.1	Nettoyage des données.....	28
3.1.1	Doublons.....	28
3.1.2	Traitement des valeurs extrêmes	29
3.1.3	Suppression de variables.....	29
3.1.4	Suppression des observations	29
3.1.5	Complétion des données manquantes à l'aide d'autre source de données complémentaire..	30
3.1.6	Imputation des données manquantes	32
3.2	Transformation des données	37
3.2.1	Booléens	37
3.2.2	Directions du vent.....	37
3.3	Ajout de variables	37

3.3.1	Coordonnées des villes	37
3.3.2	Amplitude thermique	37
3.3.3	Information climatique	38
3.3.4	Corrélations des nouvelles variables	41
3.3.5	Normalisation et standardisation	43
4	Conclusion	43

1 Introduction

1.1 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.

1.2 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 prédictive 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.

1.2.1 Contexte géographique et climatique

L'Australie est une immense île située entre l'Océan Pacifique et Indien. Elle se trouve dans l'hémisphère sud, ce qui implique que les saisons sont décalées de 6 mois par rapport à celles de l'hémisphère nord.

Au centre du pays se trouvent d'immenses déserts, occupant 18% du territoire. Leur climat est aride.

La Cordillère australienne (Great Dividing Range) est une immense chaîne de montagne longeant toute la côte est. Sa partie méridionale se nomme les Alpes australiennes. Elle comprend son point culminant (2 228m) et est enneigée.

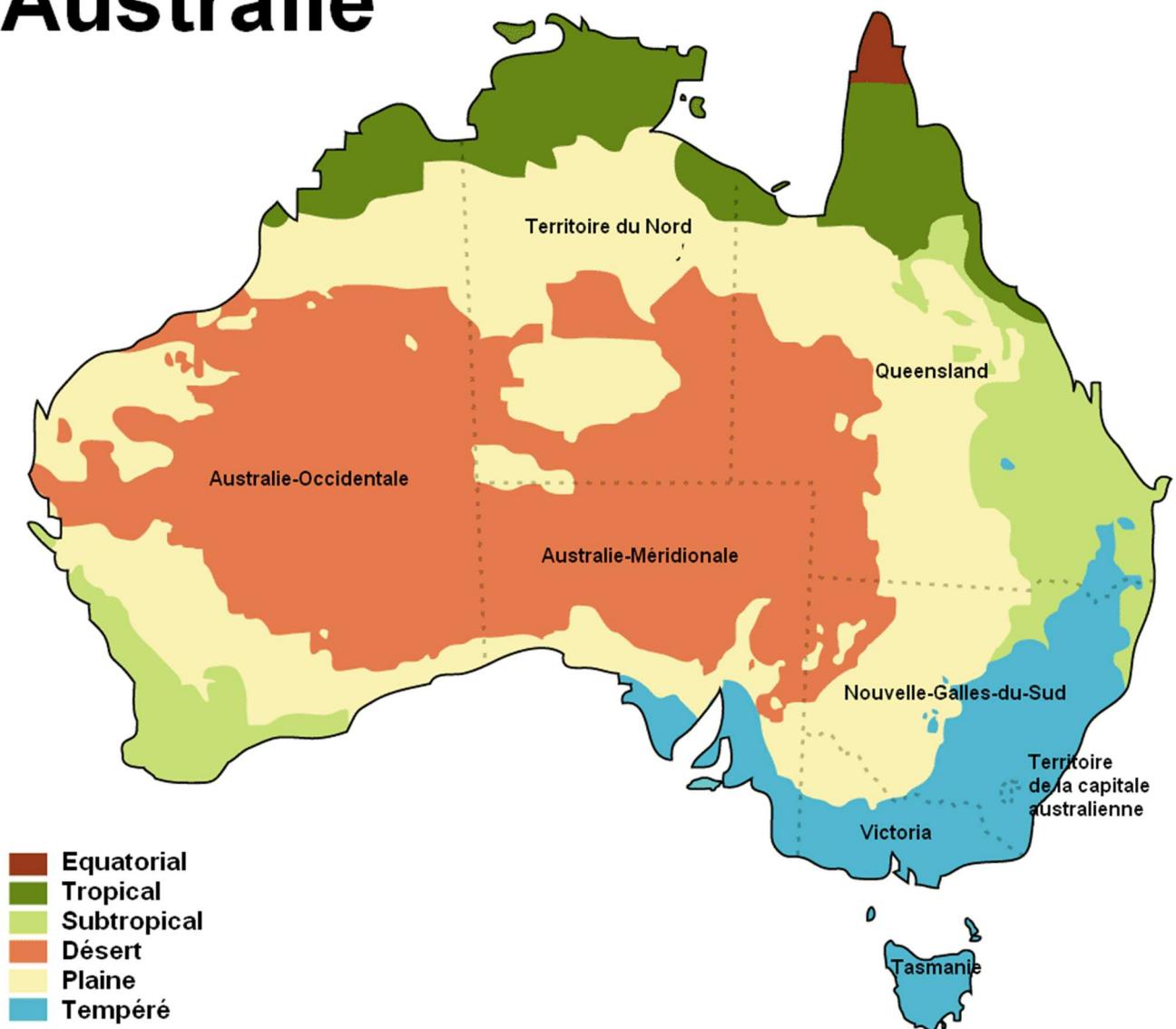
De vastes forêts tropicales longent également la côte est, en particulier sur la partie nord-est, entre l'océan Pacifique et la Cordillère australienne. Il s'agit de zones humides.

Le reste du pays est constitué de plaines de basse altitude, avec une végétation de savane tropicale au nord et de forêt de type méditerranéenne au sud.

De nombreuses îles entourent l'île principale, telles l'île de Tasmanie, au sud-est, et l'île de Norfolk, à plus de 1400km à l'est.

L'Australie possède des climats variés, tropical au nord avec des précipitations particulièrement importantes du fait de la mousson, jusqu'à un climat désertique au centre avec des températures élevées et peu de précipitations, en passant par un climat tempéré au sud-est. Plusieurs cartes de climats existent, avec des répartitions qui divergent sensiblement et des stratifications plus ou moins riches. En voici une :

Australie



Carte des climats australiens

Src : Wikipédia (source : Wikipédia : https://fr.wikipedia.org/wiki/Climat_de_l'Australie)

2 Exploration des données et visualisation

2.1 Sources de données

2.1.1 Kaggle

Le data set est celui disponible sur Kaggle pour le projet « Rain in Australia » (<https://www.kaggle.com/datasets/jphyg/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.

2.1.2 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 températures). 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 Webscraping. Elle ne pourra se faire que très ponctuellement en ciblant certaines variables pour des villes particulières.

2.2 Variables du dataset

Le dataset contient les 23 variables présentées dans le Tableau 1

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

Le Tableau 2 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.

Overview

The screenshot shows the 'Overview' tab selected in a navigation bar with 'Alerts 26' and 'Reproduction' options. Below is a table of dataset statistics:

Dataset statistics	
Number of variables	22
Number of observations	145460
Missing cells	343248
Missing cells (%)	10.7%
Duplicate rows	21
Duplicate rows (%)	< 0.1%
Total size in memory	25.5 MiB
Average record size in memory	184.0 B

Variable types	
Categorical	4
Numeric	16
Boolean	2

Tableau 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.

2.3 Statistiques descriptives

Dans cette section, nous présentons une vue globale sur les distributions des variables du dataframe.

2.3.1 Variables catégorielles

La Figure 1 représente la distribution de la variable *Location* qui contient 49 stations météorologiques. 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.

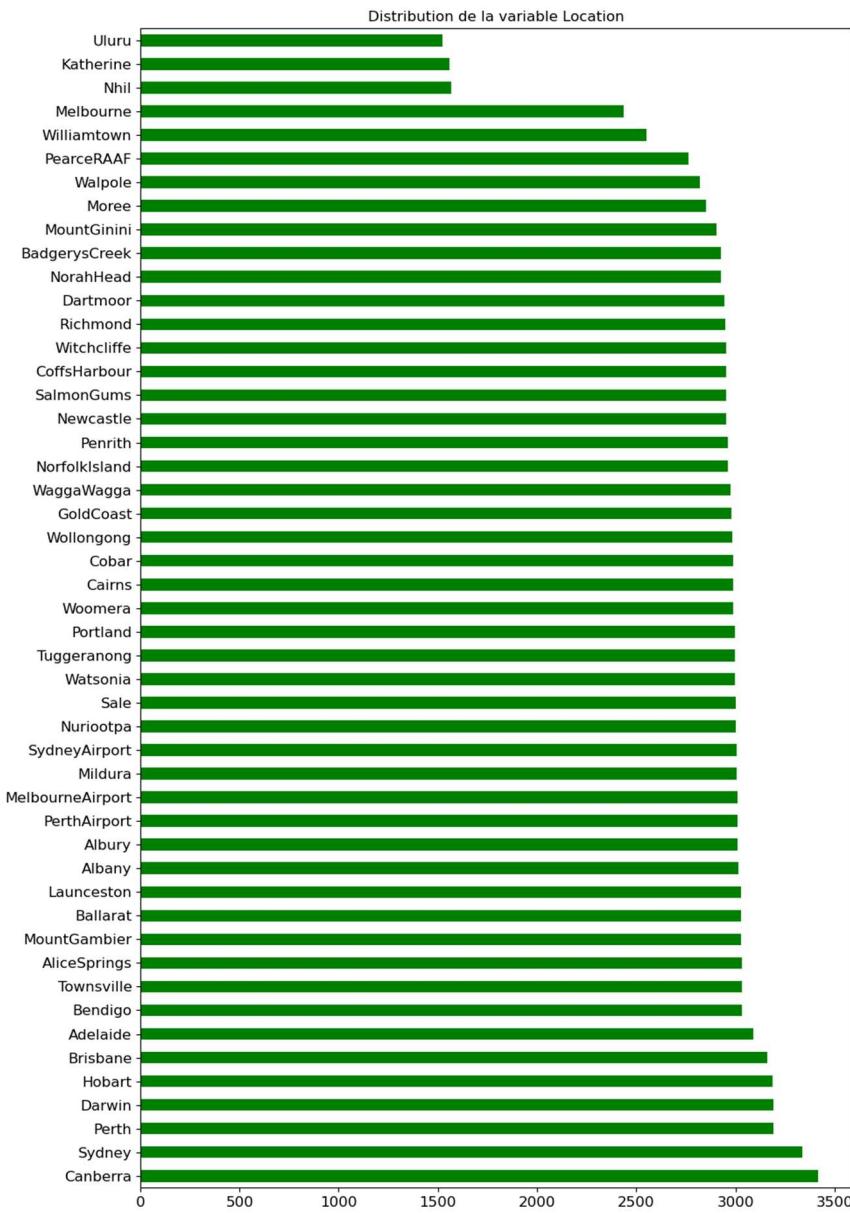


Figure 1 : Distribution de *Location*

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.

La Figure 2 représentent la distribution des trois autres variables catégorielles, *WindDir9am*, *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.

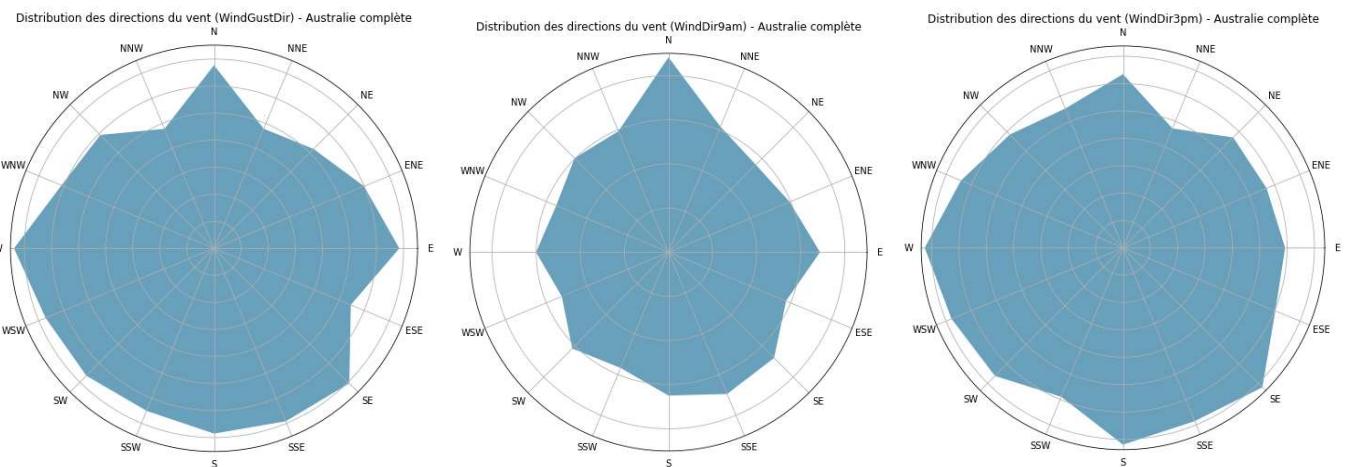


Figure 2: Distribution des variables concernant la direction du vent

La distribution de la direction du vent est en revanche radicalement différente selon les *Location*, comme nous pouvons le voir dans la Figure 3 à Townsville, Hobart et Bendingo. Il est même étonnant de constater qu'il y a également des distributions 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 distributions du vent pour l'ensemble des Location.

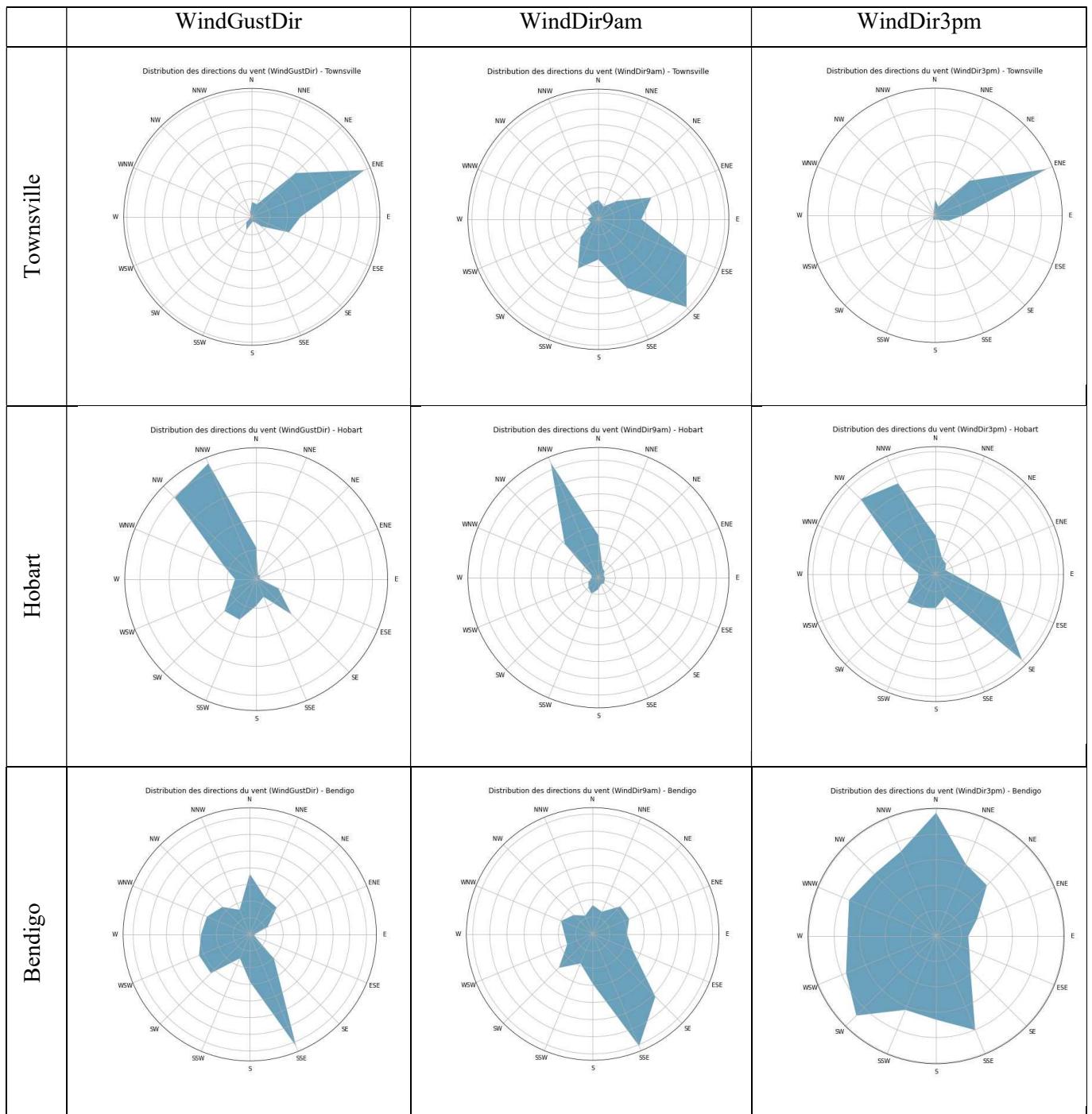


Figure 3 : Distribution des variables concernant la direction du vent à Townsville, Hobart et Bendigo

Ces différences peuvent parfois se comprendre en observant simplement la situation géographique du lieu. On voit dans la Figure 4 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.

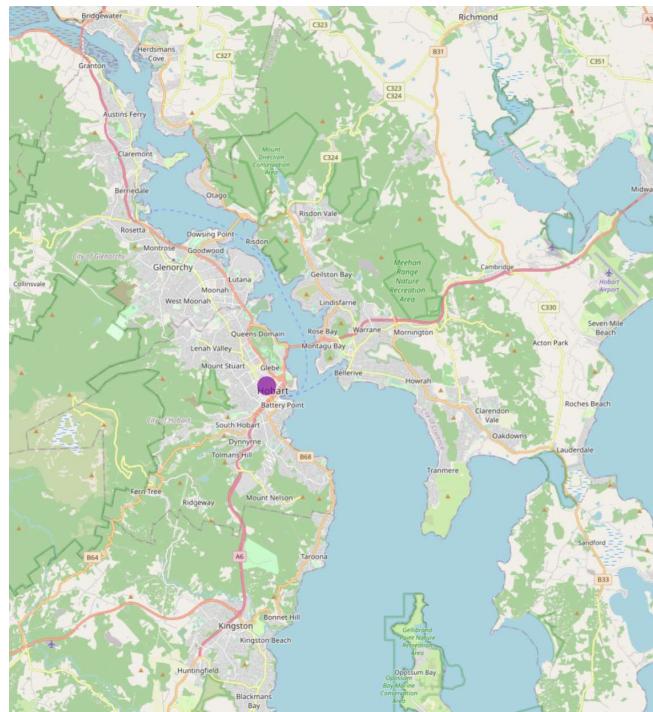


Figure 4: Situation géographique de Hobart

Il nous est possible de tester la corrélation de ces 3 variables qualitatives avec RainTomorrow avec un test de χ^2 , avec l'hypothèse nulle supposant qu'il n'existe pas de corrélation.

La méthode « correlation_vent » calcule la p-value issu du χ^2 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.

2.3.2 Variables numériques

La Figure 5 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.

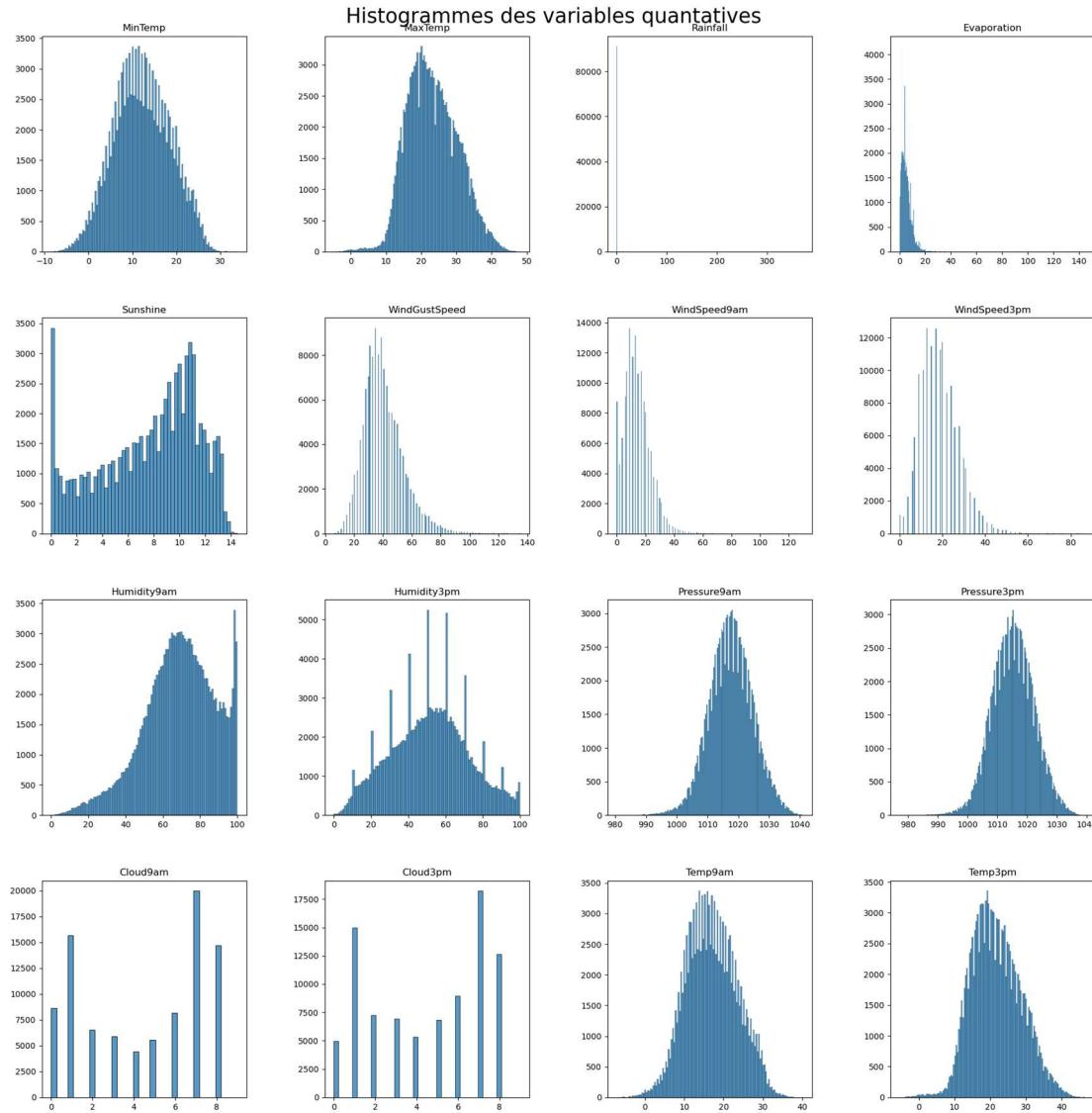


Figure 5: Histogrammes des variables numériques

La Figure 6 représente les boxplots des variables numériques continues. On constate 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.

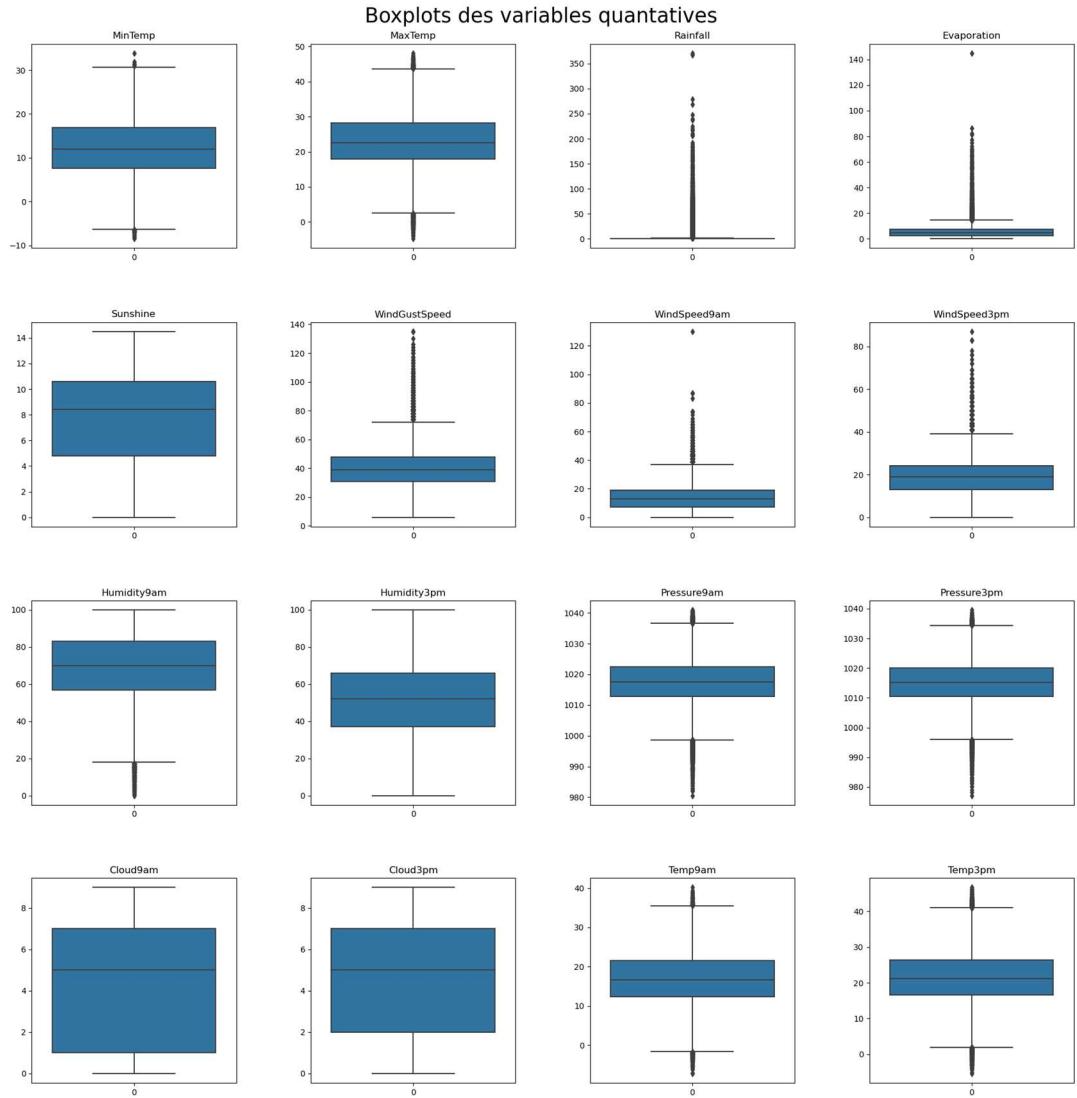


Figure 6: Boxplots des variables quantitatives

Les boxplots des 16 variables quantitatives nous montrent aussi 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 survit 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.

2.4 Corrélation

La Figure 7 représente la corrélation entre les variables numériques. On peut constater qu'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*, *Cloud9pm*, *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).

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).

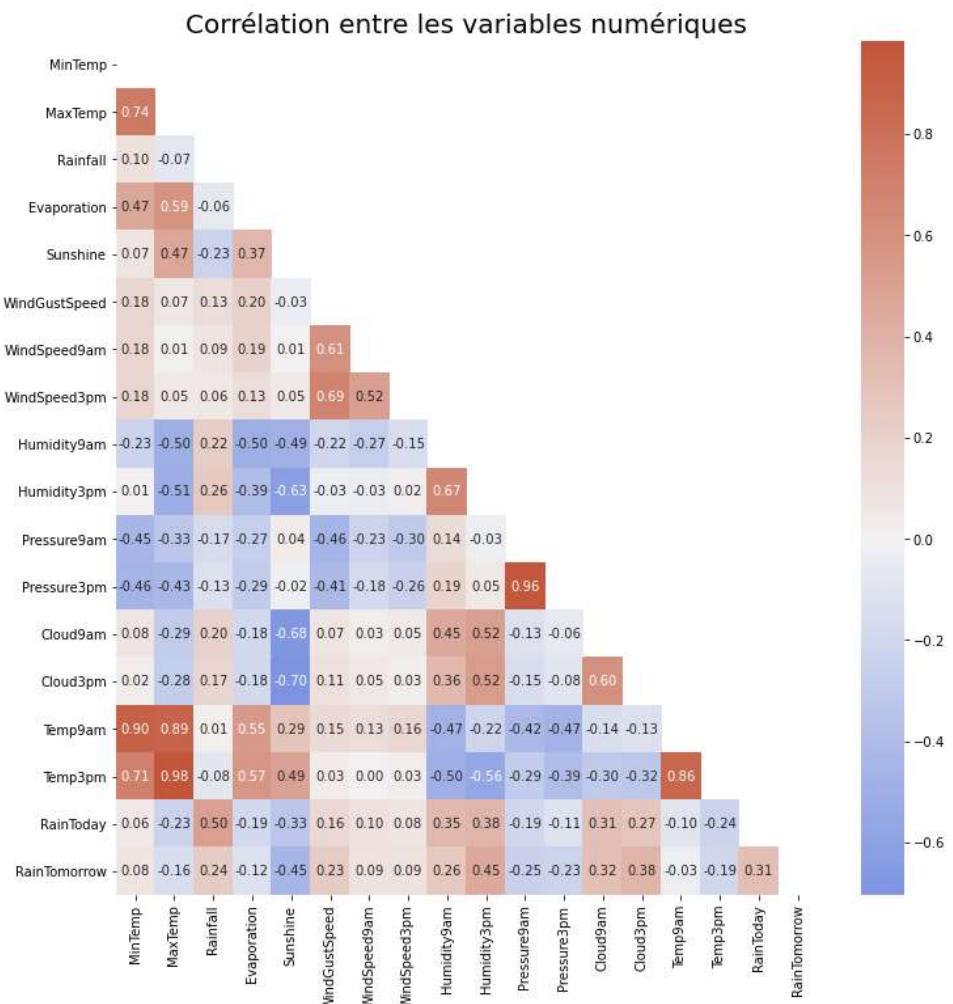
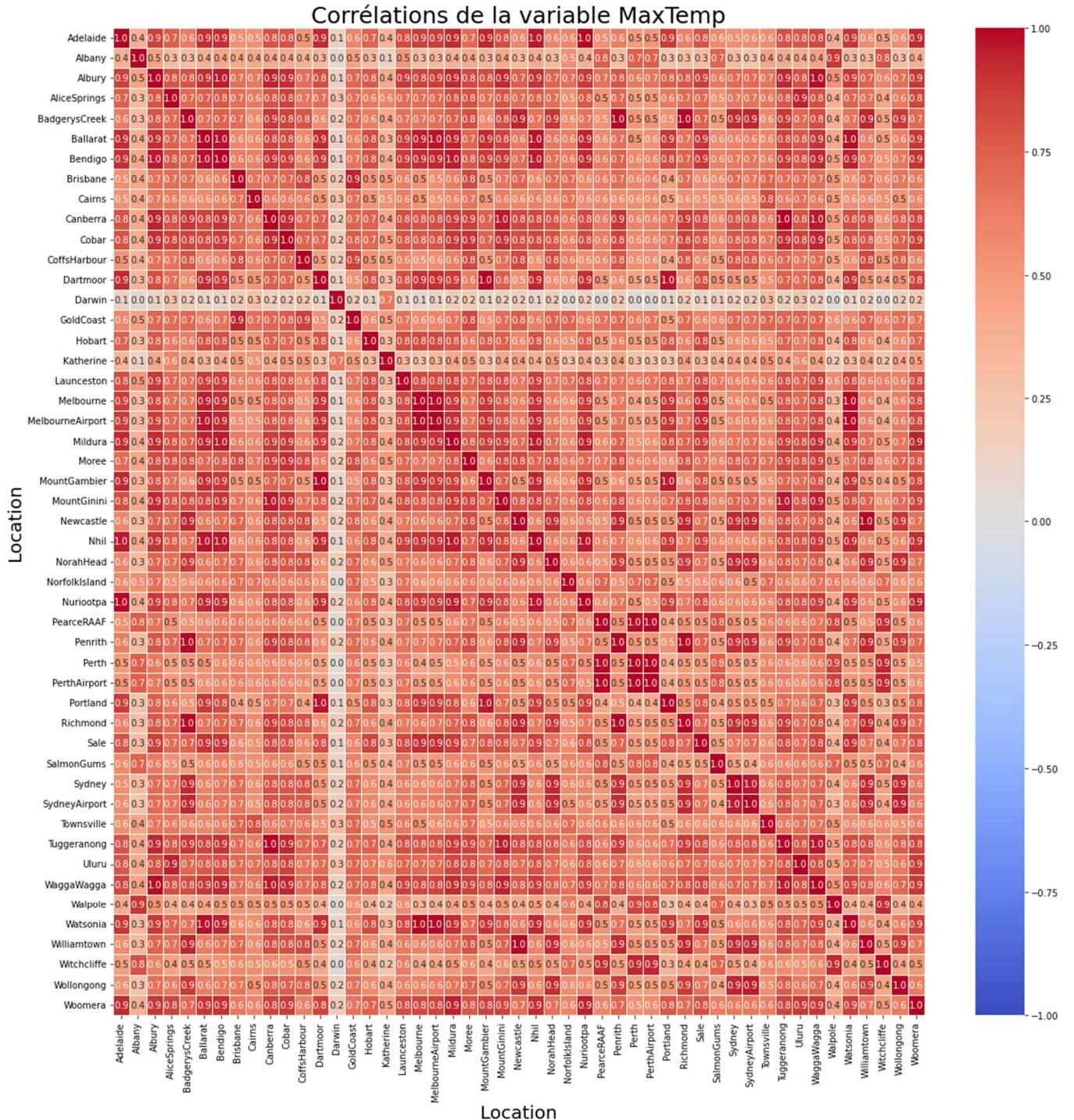


Figure 7: Corrélation des variables numériques

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 corrélées pour une variable donnée.

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.



2.5 Analyse détaillée des variables

Dans cette section, nous allons analyser en détail certaines variables ainsi que les relations entre certaines variables.

2.5.1 RainTomorrow

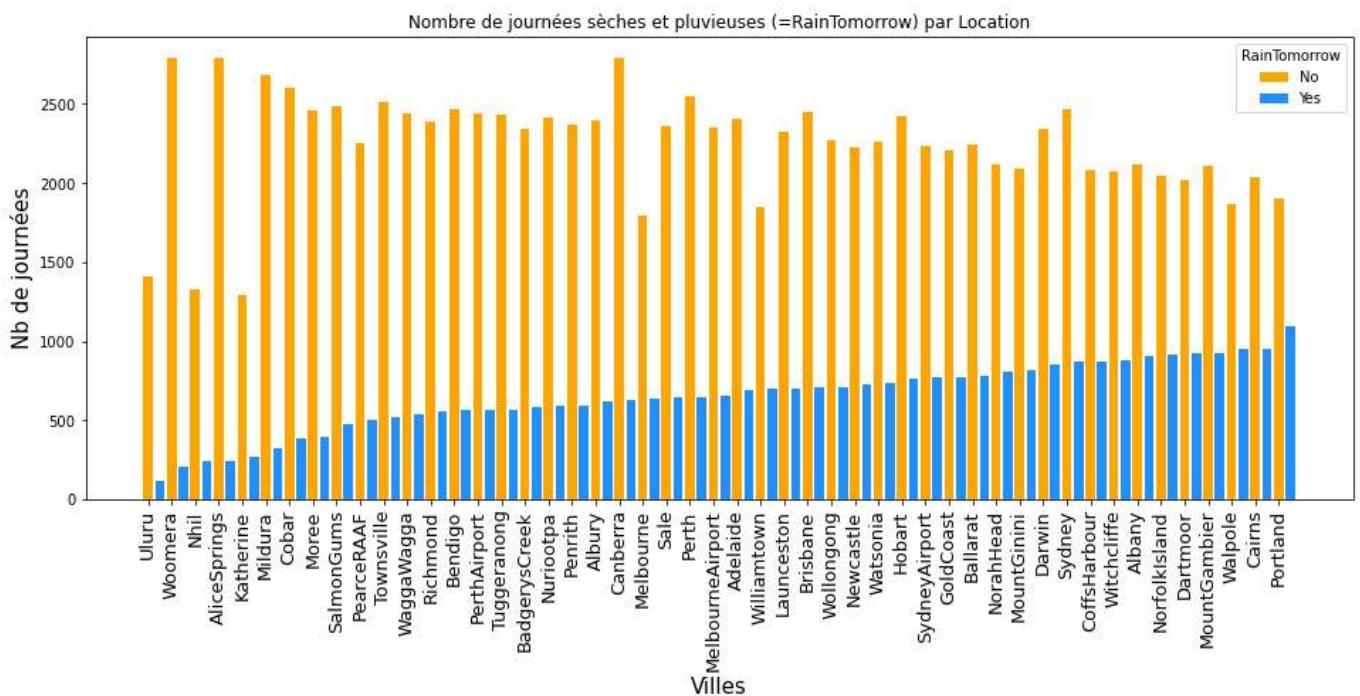
Regardons en premier lieu la variable *RainTomorrow*, indiquant s'il pleuvra le lendemain. C'est en effet cette variable que nous allons utiliser dans un premier temps comme variable cible pour notre modélisation. Elle revêt donc une importance particulière.

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 *RainTomorrow* 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.



2.5.2 Location, MaxTemp et RainTomorrow

Nous avons, en suite, recherché les latitudes et longitudes des 49 stations météorologiques 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)

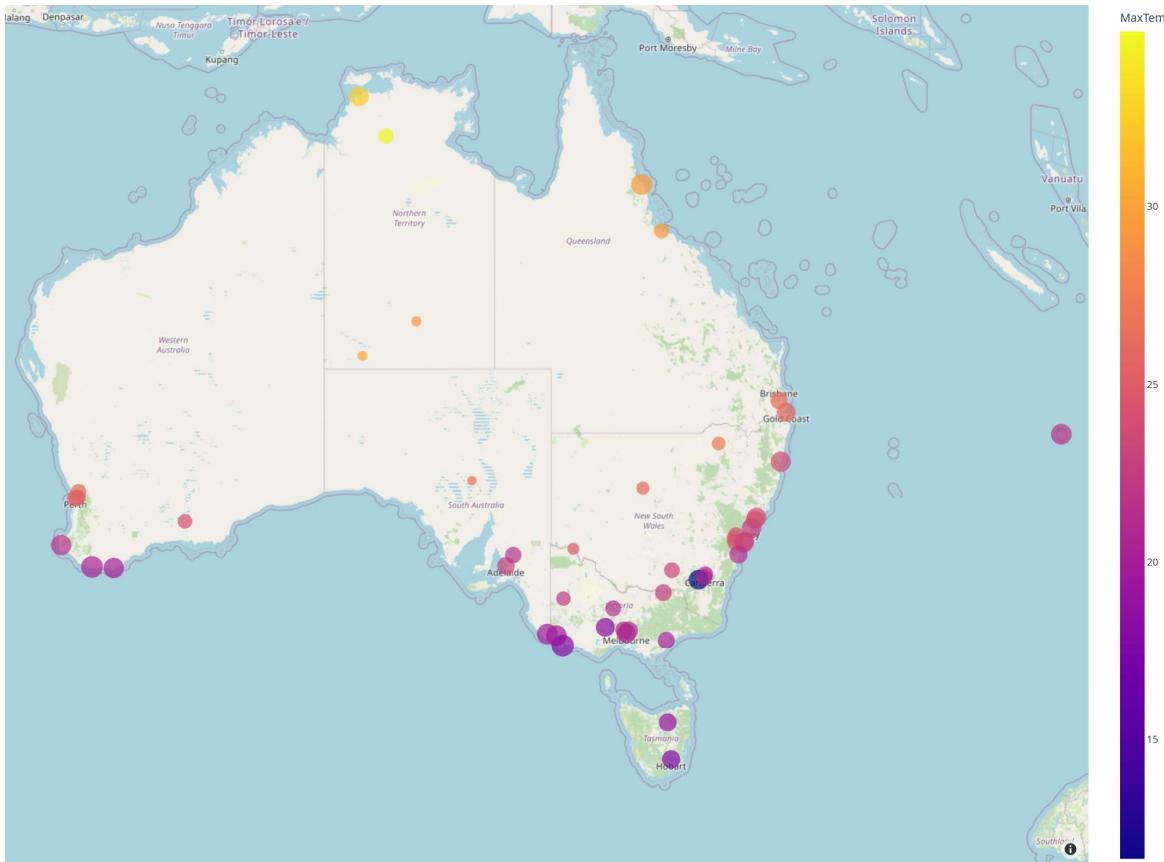


Figure 8: Location – Couleur : température maximale moyenne – Diamètre : taux de journées pluvieuses

La Figure 8 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 la Figure 9 issue du site de l'Australien Bureau of Statistics :

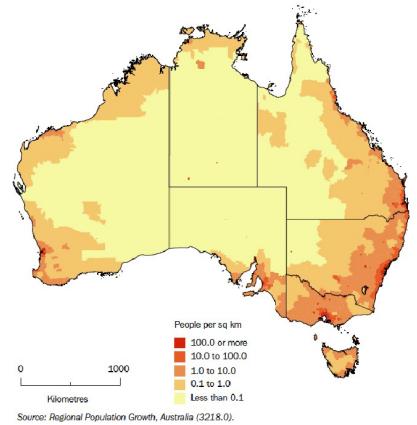


Figure 9 : Répartition de la population Australienne

2.5.3 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 Figure 10 et Figure 11, 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.

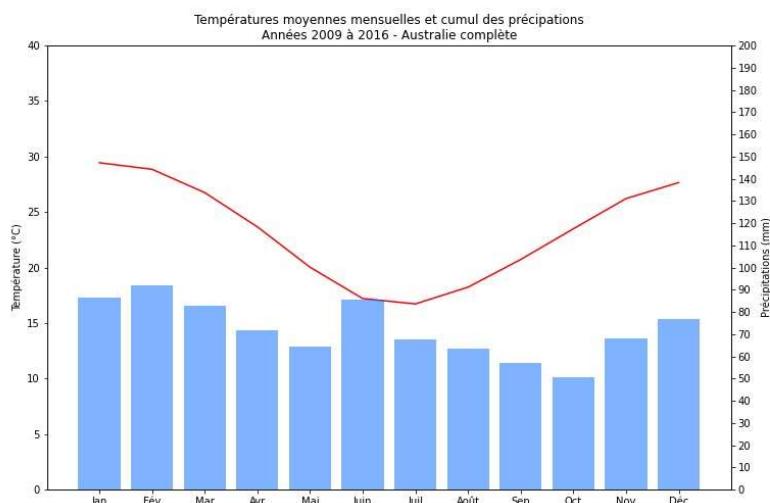


Figure 10 : Températures moyennes mensuelles et cumul des précipitations de 2009 à 2016 – Australie complète

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.

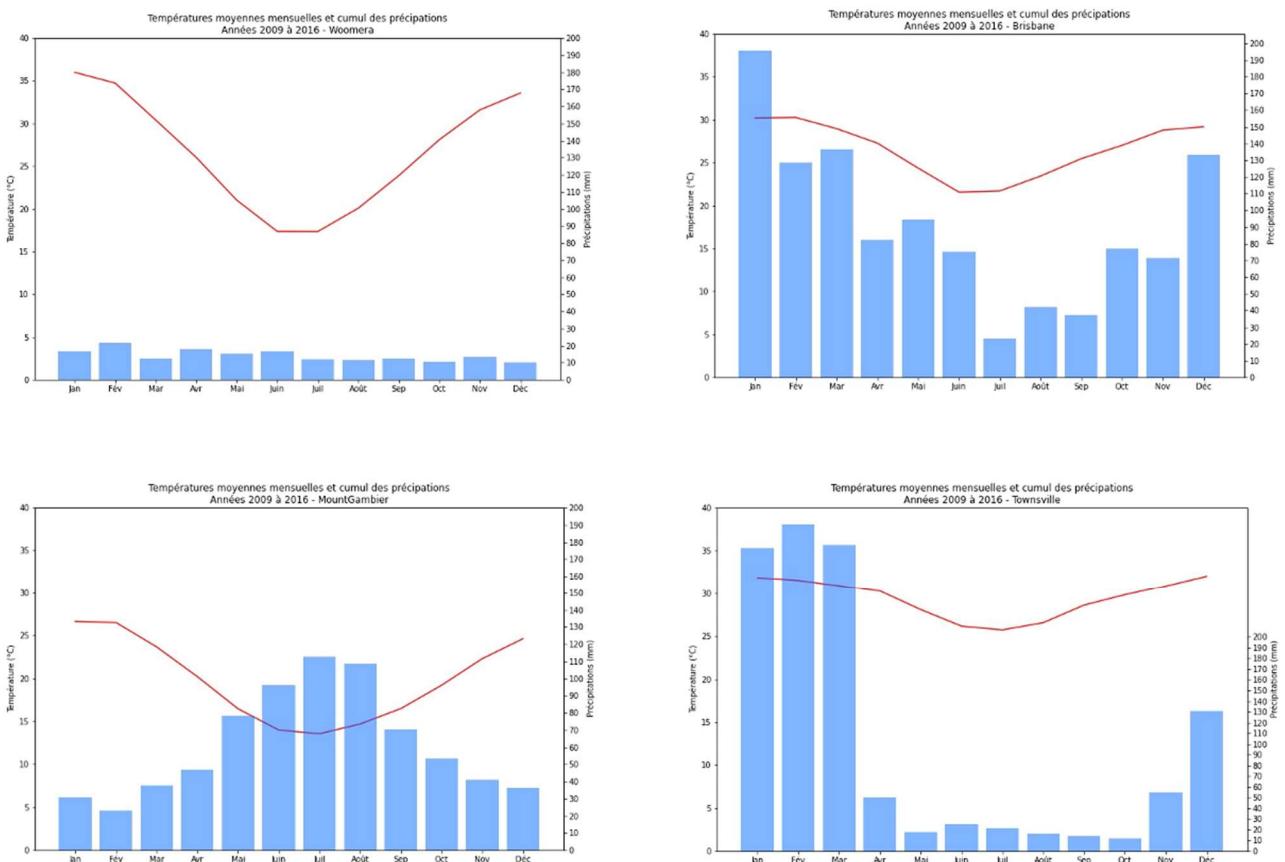


Figure 11: Températures moyennes mensuelles et cumul des précipitations de 2009 à 2016 – Woomera, Brisbane, Mount Gambier et Townsville

Ces quatre graphiques dans la Figure 11 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'il existe des différences significatives concernant les précipitations et la température maximale entre l'Australie dans son ensemble d'une part et chaque ville d'autre part. En

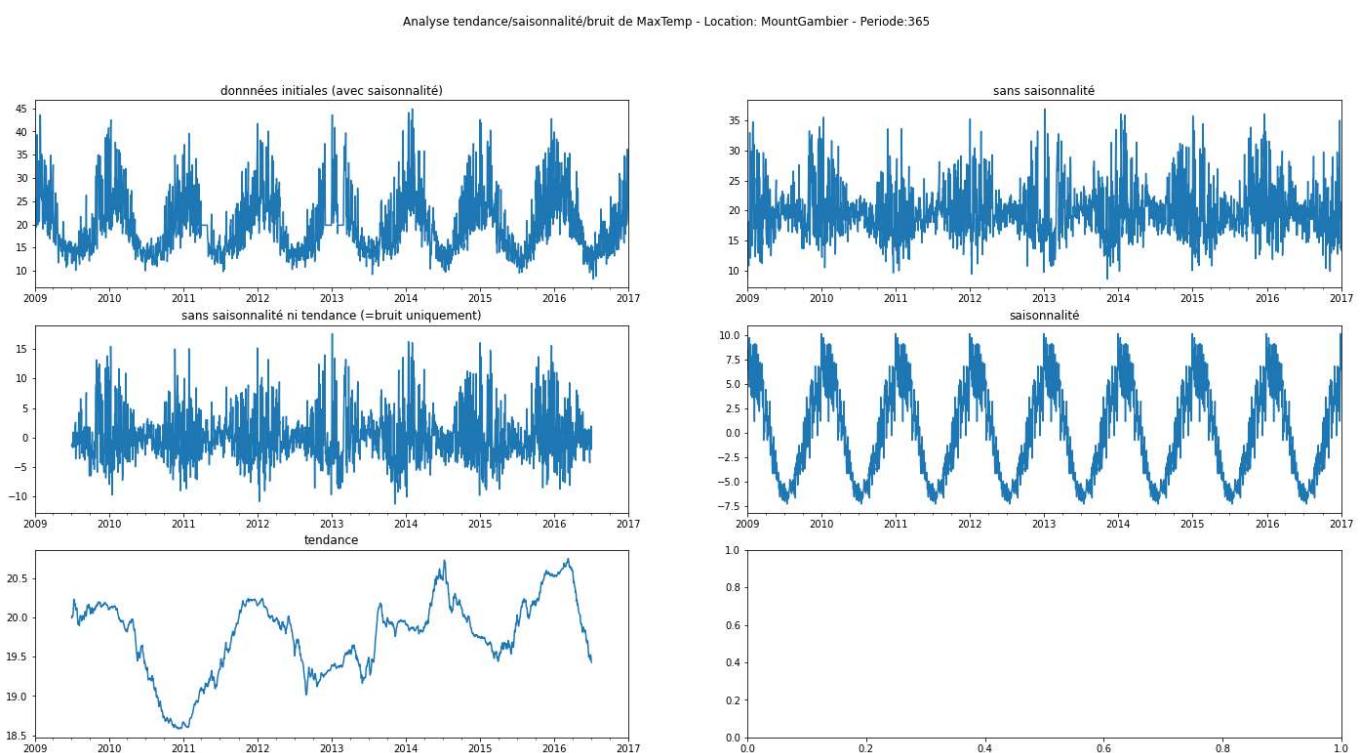
revanche, les villes situées proches géographiquement présentent des caractéristiques communes. Il semblerait donc pertinent d'effectuer des regroupements de villes en fonction de caractéristiques climatiques.

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 saisonnalité. 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 dans la Figure 12 les décompositions saisonnières pour les variables *MaxTemp* et *Rainfall* de Mount Gambier.



Nous constatons ici que la saisonnalité explique des variations de températures de -7° à $+10^{\circ}$, 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 $+12^{\circ}$, soit une amplitude de 19° . Nous verrons dans la partie modélisation si un modèle ARIMA arrive malgré ce constat à prédire de façon pertinente les températures.

Ce même schéma sur *Rainfall*, voir la Figure 13, 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* ».

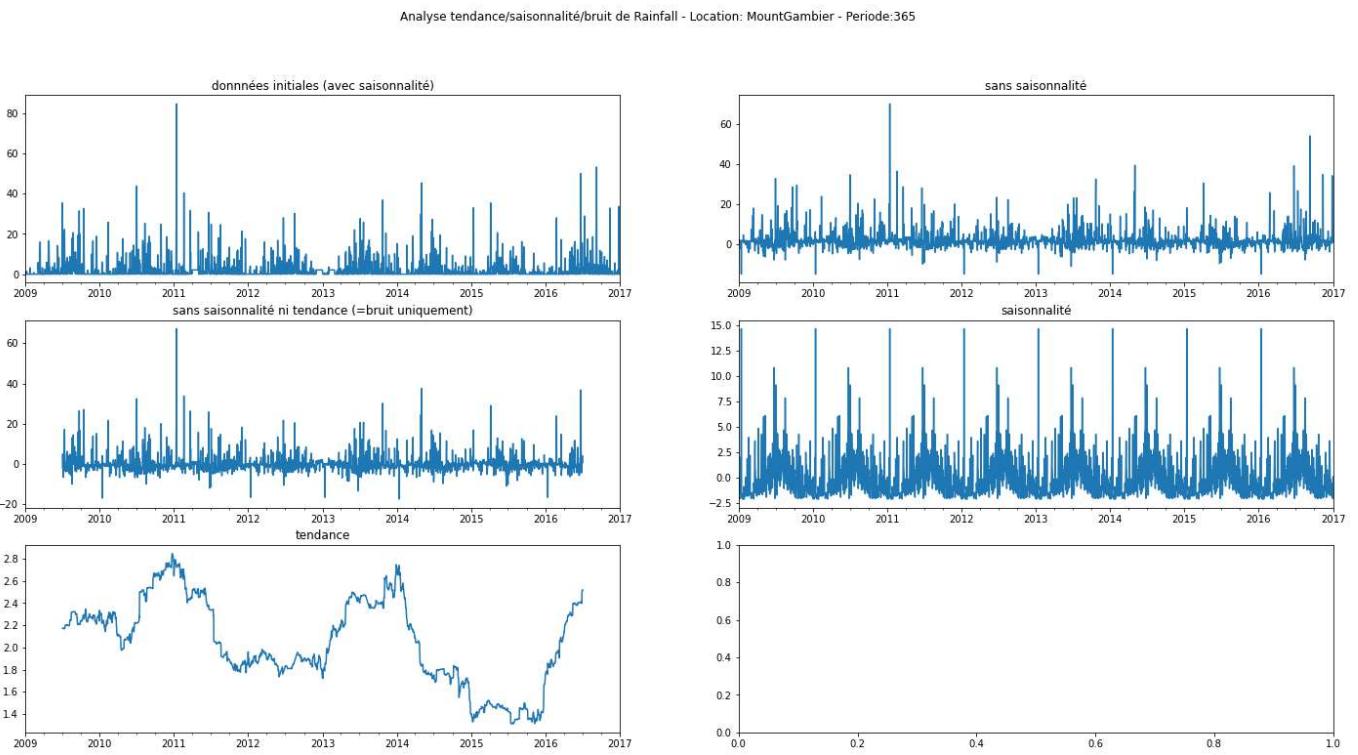


Figure 13 : Analyse tendance/saisonnalité/bruit de *MaxTemp* - MountGambier

2.6 Valeurs manquantes

2.6.1 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 14 représentant le pourcentage de valeurs manquantes pour chaque variable montre que toutes les variables contiennent des valeurs manquantes, sauf *Location* qui est complète. Nous pouvons voir que les quatre variables *Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm* comportent un grand nombre de valeurs manquantes.

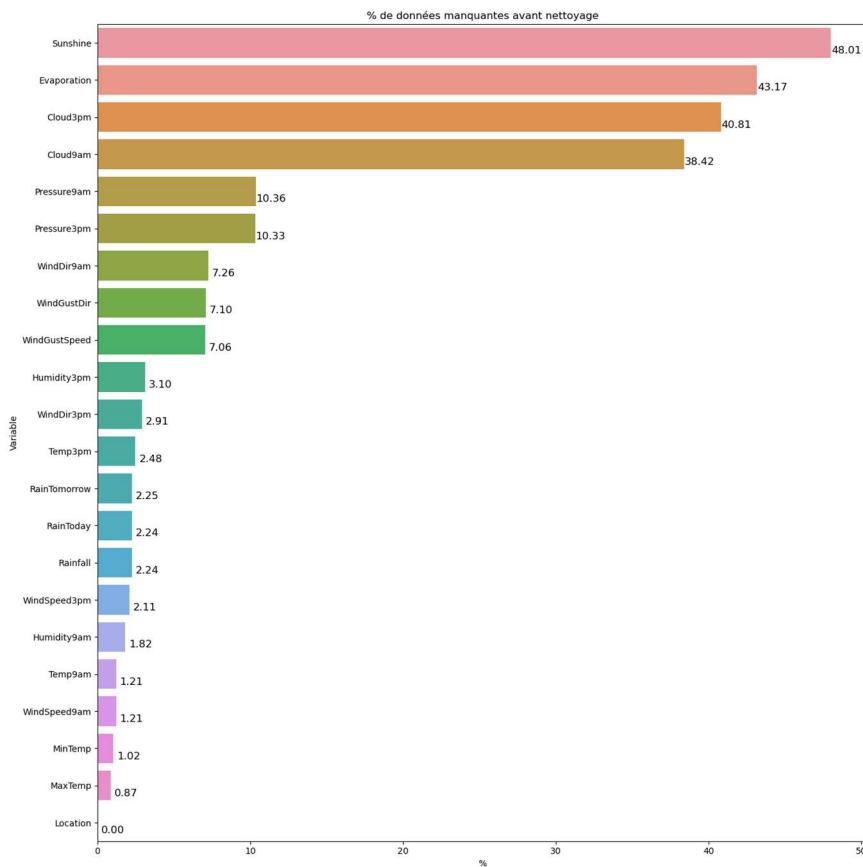


Figure 14: Pourcentage de valeurs manquantes pour chaque variable

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 la Figure 15 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.

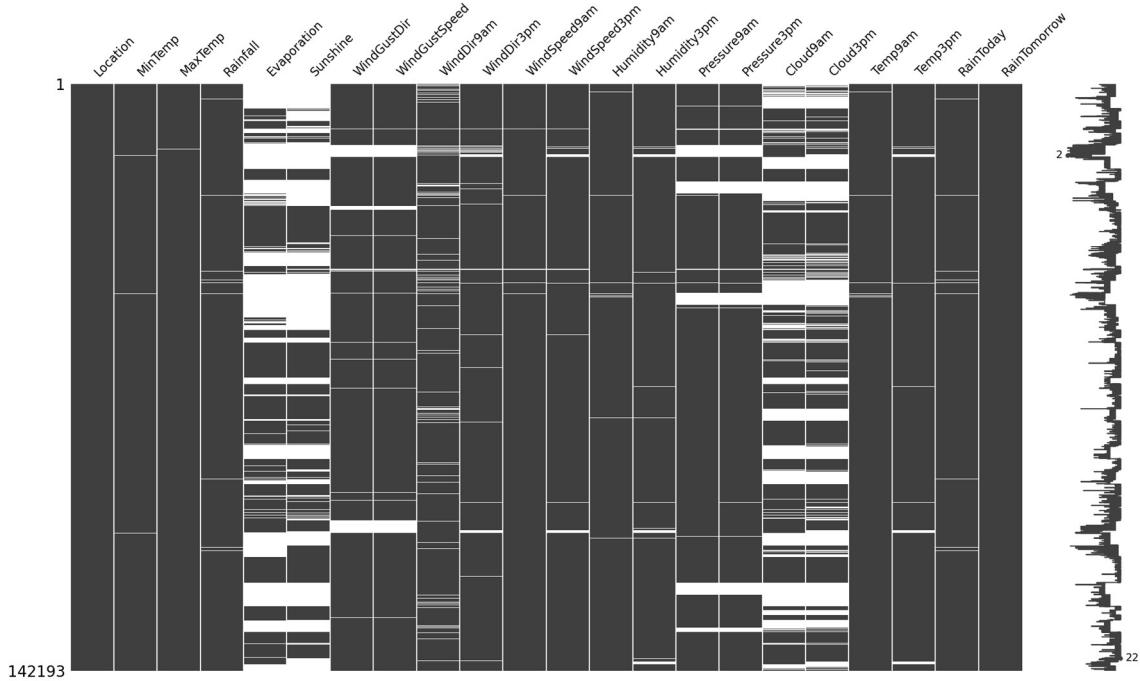


Figure 15 : Matrice des valeurs manquantes

Nous pouvons également nous intéresser au taux de variables manquantes non pas par colonnes, mais par ligne, c'est-à-dire par observation. La Figure 16 nous permet de constater que 1,92% des lignes possèdent plus de la moitié des informations manquantes, ce qui les rendra difficilement exploitables.

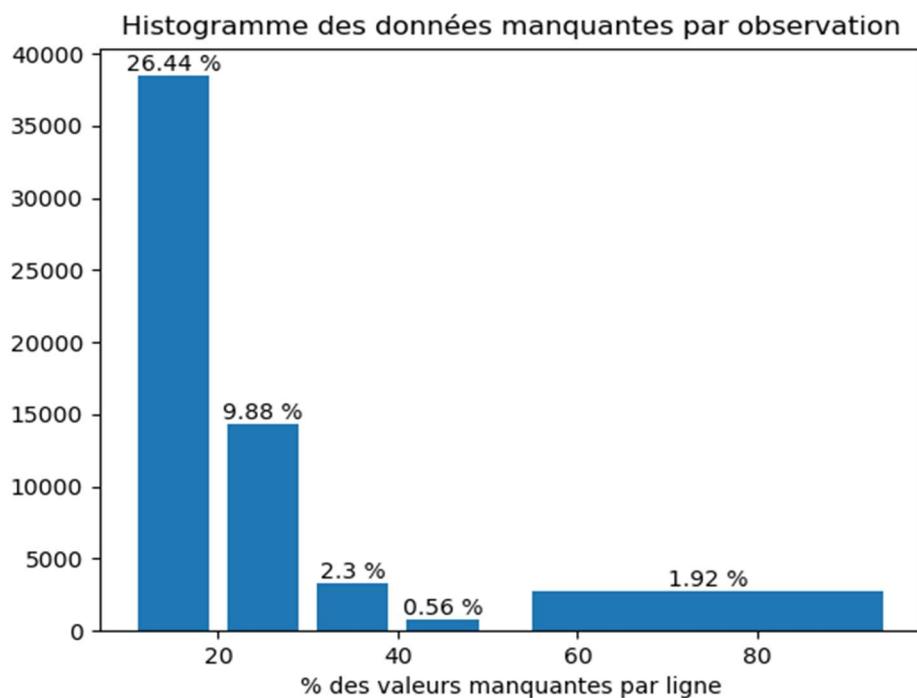


Figure 16 : Pourcentage de valeurs manquantes pour chaque observation

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 connaître la corrélation de nullité entre les différentes variables. C'est ce que montre la Figure 17.

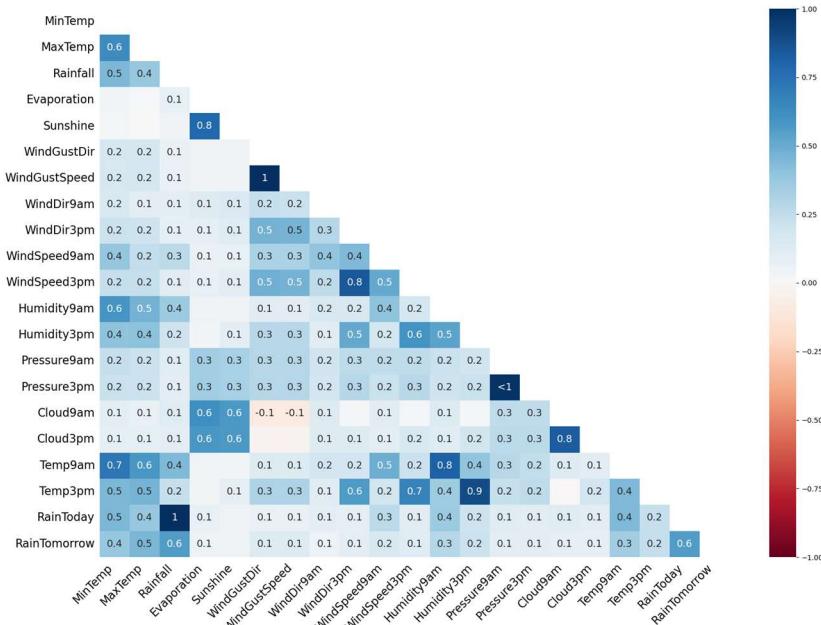


Figure 17 : Corrélation de nullité entre les variables

- 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 comme dans la Figure 18 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.

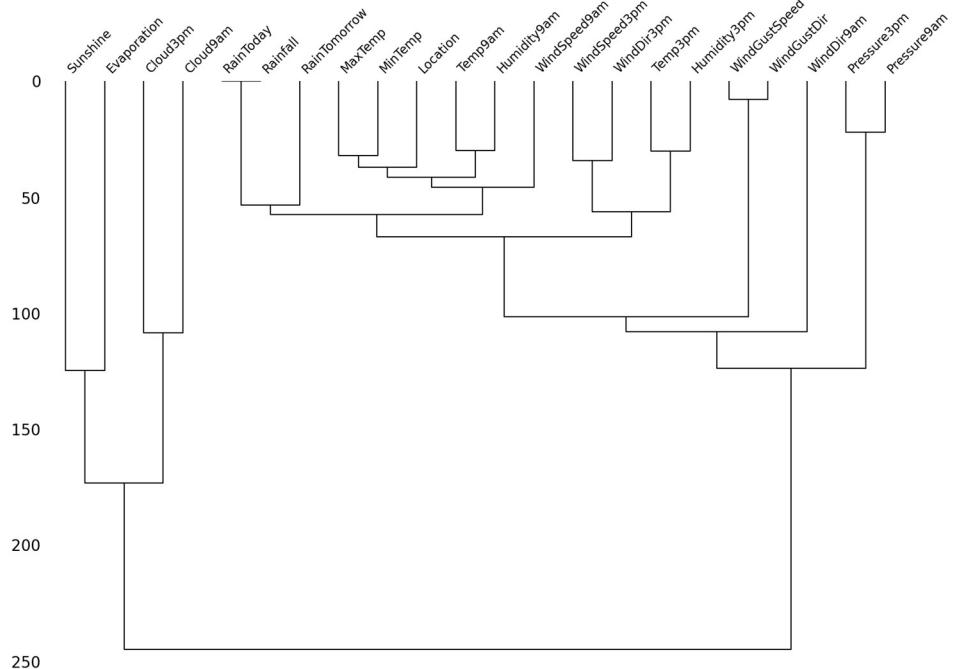


Figure 18 : Dendrogramme de nullité entre les variables

2.6.2 Répartition géographique

Regardons maintenant les données manquantes par lieux représentées dans le Tableau 3. 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 !

Cette représentation nous permet de nous rendre compte que 15% des données *Rainfall* de Williamtown sont manquantes, ce qui est une proportion beaucoup plus élevée que pour les autres villes, alors même que Williamtown est globalement bien renseignée. Nous avons donc téléchargé les données pour *Rainfall* de Williamtown sur le site du Bureau of Meteorology.

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 Ginini, Newcastle, PearceRAAF, Sydney, Williamtown). La question du maintien de ces villes dans le jeu de données pourrait se poser pour la modélisation, mais, étant donné notre problématique, cela nous priverait de la possibilité de prédire la météo pour ces villes, ce qui constituerait une perte importante de qualité.

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, nous aurions souhaité retrouver cette information dans des relevés météo. Malheureusement, seuls les 14 derniers mois sont disponibles pour les variables relatives au vent sur le site du Bureau of Meteorology. Nous devrons donc tenter de reconstituer les données manquantes par des approches que nous aborderons un peu plus loin.

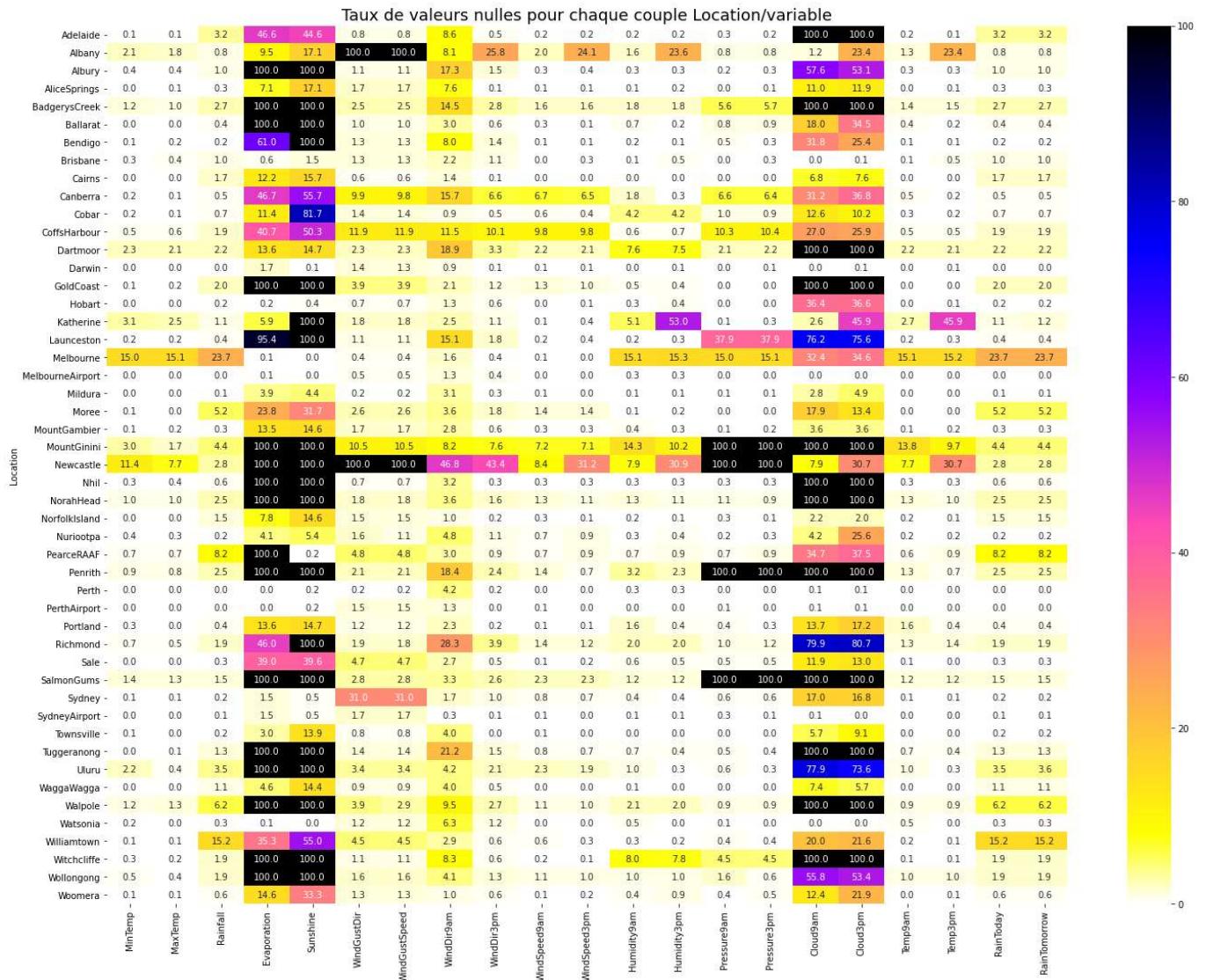


Tableau 3 : Pourcentage des valeurs manquantes pour chaque couple Location/variable

2.6.3 Répartition temporelle

Les données disponibles vont du 1^{er} 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 généreront 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 présentés dans la Figure 19 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 variables 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")*)

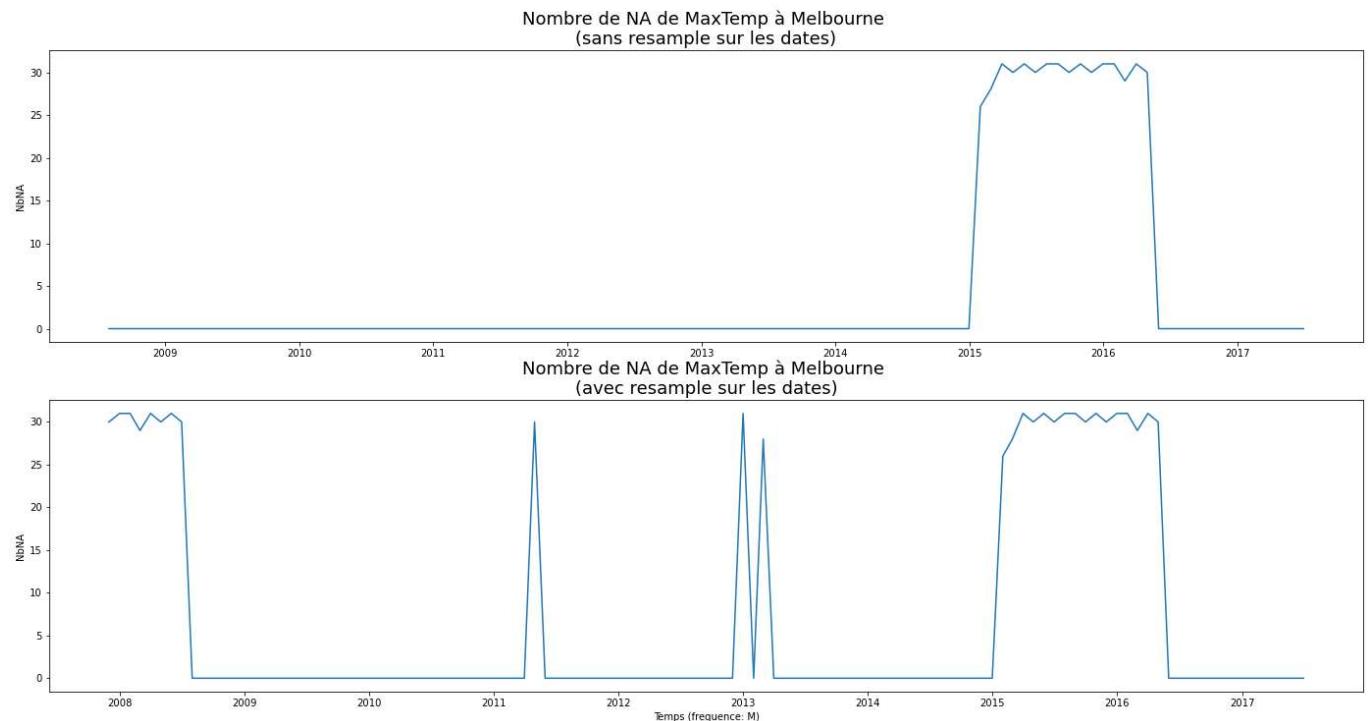


Figure 19 : Nombre de valeurs manquantes de MaxTemps à Melbourne

Tentons maintenant de représenter les valeurs manquantes simultanément pour tous les lieux et selon le temps. La Figure 20 illustre les données manquantes pour *RainTomorrow*, la Figure 21 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 difficilement 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*.

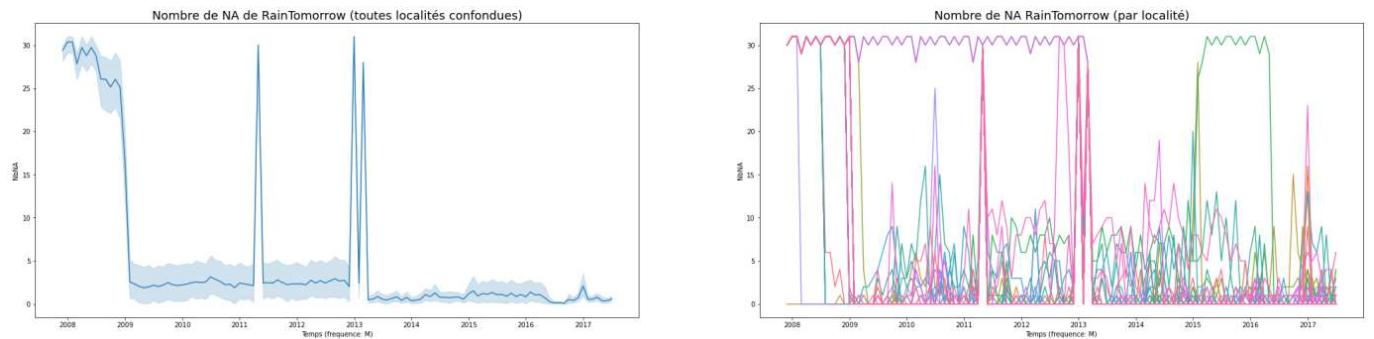


Figure 20 : Nombre de valeurs manquantes de *RainTomorrow*

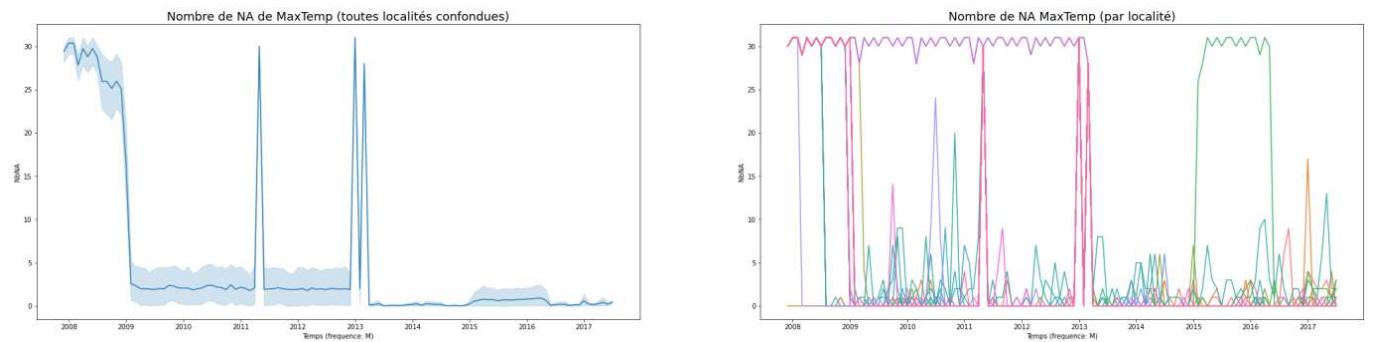


Figure 21 : Nombre de valeurs manquantes de *MaxTemp*

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 1^{er} 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*.

3 Pre-processing et feature engineering

3.1 Nettoyage des données

3.1.1 Doublons

Le nettoyage des données est un point essentiel à effectuer avant toute modélisation.

Parmi les premiers éléments constatés, nous avons vu qu'il y avait 22 lignes dupliquées. Chaque enregistrement correspondant à une date pour un lieu précis, les doublons sont donc dans notre jeu de données de véritables données redondantes, et non pas des informations complémentaires pouvant coïncider dont le doublonnage pourrait être une information pertinente. Un simple appel à `drop_duplicates()` permet d'évacuer ces lignes.

3.1.2 Traitement des valeurs extrêmes

Comme vu plus haut, les valeurs extrêmes de notre jeu de données sont certes des outliers d'un point de vue mathématiques, mais ne sont pas des données aberrantes au regard des échelles de valeurs et des types de données météorologiques. Nous faisons donc le choix de conserver l'intégralité des outliers après les avoir analysés.

3.1.3 Suppression de variables

L'analyse des corrélations à mis en évidence un lien fort entre les quatre variables de température d'une part, et les deux variables de pression d'autre part. La colinéarité de ces variables pouvant nuire à la bonne qualité des prédictions du modèle, nous pourrons supprimer plusieurs features : MinTemp, Temp9am, Temp3pm, Pressure9am. Nous conserverons cependant la possibilité de les conserver ou non selon les modèles utilisés.

3.1.4 Suppression des observations

Le traitement de données manquantes est une étape essentielle de la préparation des données pour l'analyse et la modélisation. Dans notre ensemble de données, plusieurs stratégies ont été adoptées pour traiter ces valeurs manquantes.

Il s'agit là d'un point particulièrement complexe à gérer dans notre jeu de données, car, comme nous l'avons vu plus haut, des données sont manquantes sur toutes les plages de dates, pour tous les lieux et pour toutes les variables, dans des proportions différentes.

Cependant, quatre variables ressortent particulièrement, avec environ 40% de données manquantes. Nous porterons donc une attention particulière à l'impact de ces variables sur la qualité des prédictions, et à la cohérence des imputations qui leur auront été faites.

Sur un plan temporel, très peu de données sont disponibles avant le 1^{er} janvier 2009. Nous pouvons donc supprimer les données antérieures pour les modélisations qui prendront en compte les dates. Nous conserverons en revanche bien ces observations pour les autres types de modèles.

Trois villes (Katherine, Nhil, Uluru) disposent de moitié moins d'enregistrement que les autres, leurs relevés ne débutant qu'en 2013, soit plus de quatre ans après les premiers relevés des autres stations.

Nous avons adopté une approche progressive pour la suppression des lignes contenant des données manquantes.

- **Suppression des lignes avec des données manquantes pour la variable cible :** La variable cible *RainTomorrow* est essentielle pour notre modèle prédictif. Puisque c'est ce que nous cherchons à prédire, les lignes contenant des valeurs manquantes pour cette variable ont été supprimées, représentant 2.2% de l'ensemble de données. Cette suppression est justifiée car imputer la variable cible pourrait introduire un biais ou une inexactitude dans nos prédictions.
- **Suppression des lignes avec une forte proportion de données manquantes :** Avant de procéder à l'imputation des données manquantes, il est essentiel d'évaluer si certaines lignes ont une proportion exorbitante de valeurs manquantes, au point que leur utilité est mise en question. Dans notre jeu de données, toutes les lignes contenant plus de 50% de valeurs manquantes ont été jugées comme n'ayant pas suffisamment d'information pour être utiles et ont donc été supprimées.

Pour autant, ces suppressions vont dépendre du modèle utilisé : pour une modélisation par série temporelle, il nous sera indispensable de disposer des données sur toute la plage de dates. Les lignes avec des données manquantes seront alors conservées et les données seront déduites par l'une des approches listées plus bas.

3.1.5 Complémentation des données manquantes à l'aide d'autre source de données complémentaire

Le site du Bureau of Meteorology nous permet de consulter les relevés des stations. Malheureusement, ce téléchargement ne peut se faire que par station météo, et non en globalité, et seuls les 14 derniers mois permettent de disposer de toutes les variables. En pratique, seules les variables *Rainfall* (dont on peut déduire *RainToday* et *RainTomorrow*), *MaxTemp* et *Sunshine* peuvent être téléchargées. Ce téléchargement doit se faire par *variable* et par station météo, sachant que notre jeu de données n'indique qu'un nom de ville, et non le nom précis de la station. Pour Sydney, par exemple, il n'existe pas moins de 8 stations possibles dont il faudrait donc analyser les données pour déterminer laquelle est la plus proche de notre dataset. Ce travail ne peut donc pas simplement s'effectuer par webscrapping et représenterait un temps considérable d'analyse manuelle. Nous faisons donc le choix de ne télécharger que les données de *MaxTemp* et *Rainfall* pour les Locations qui ont taux de NA élevé. C'est le cas de Melbourne, PearceRAAF et Williamtown pour *Rainfall*, et de Melbourne et Newcastle pour *MaxTemp*. L'exploitation de *Sunshine* nécessiterait de la télécharger pour une quarantaine de Location, ce qui serait trop chronophage en analyse : nous restons sur notre choix d'abandonner cette donnée.

L'obtention de données complémentaires est particulièrement précieuse puisque cela nous permet de déduire logiquement la valeur de *RainToday* manquants, mais également de notre variable cible *RainTomorrow* !

La variable *Rainfall* dispose d'une particularité : contrairement aux autres variables, elle n'indique pas forcément uniquement la valeur pour le jour donné (à savoir le niveau de précipitations en mm), mais cumule parfois les valeurs des jours précédents lorsqu'ils ne sont pas renseignés. En d'autres termes, il est probable que le relevé des pluviomètres ne se faisait pas chaque jour sur la période analysée.

Show in table... ▾

Key: Units = mm 12.3 = Not quality controlled. ↓ = Part of accumulated total
Move mouse over rainfall total to view the period of accumulation.

Graph 2014

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Graph	[Graph]											
1st	0	0	↓	0	6.2	↓	0	↓	0.2	0	↓	0
2nd	0.4	0	34.2	0	↓	2.2	0	↓	0	0	↓	3.2
3rd	0	0	4.6	0	↓	0	0	2.0	3.6	0	1.0	0
4th	0	0	0	↓	↓	0	0	0.2	0	0	0	8.6
5th	0	0	0	↓	1.8	1.8	0	0	↓	0	0	↓
6th	0	↓	↓	0	0	↓	0	0	↓	0	17.8	↓
7th	0	↓	30.2	14.4	↓	0	0	0	↓	0	0	25.2
8th	0	↓	0	1.2	36.2	0	↓	35.6	0.4	0	0	4.2
9th	0	↓	0	↓	1.0	0	↓	0.2	7.2	0	0	6.0
10th	0	1.0	0	↓	21.2	0	↓	6.0	0	0	0	0
11th	0	0	↓	21.0	2.0	↓	0.2	0	0	0	0	17.2
12th	0	0	↓	6.8	0	↓	↓	↓	0	0.4	↓	↓
13th	0	0	8.2	12.4	↓	↓	↓	↓	0	0	0	↓
14th	0	0	6.8	0.2	↓	8.6	↓	5.4	12.8	↓	↓	↓
15th	0	0	9.2	0.2	↓	2.4	↓	0	↓	↓	↓	9.4
16th	0	0	20.0	0	0.8	1.0	↓	0	↓	↓	↓	0
17th	0	0.8	0.8	0	0	2.6	33.4	0.2	↓	0.8	0	0
18th	0	27.0	0	0	0	↓	↓	0.4	0	↓	0	1.2
19th	0	0	0	0	0.2	1.0	↓	21.8	↓	↓	0	0
20th	0	17.0	0	0	0	↓	5.0	21.4	↓	18.0	0	0
21st	0.6	0	↓	0	0	↓	2.0	1.2	↓	1.0	0	0
22nd	1.0	0	↓	0	0	6.2	0.2	↓	0.4	0	0	0
23rd	0.6	0	↓	0	↓	0	0.2	↓	0	0	0	2.4
24th	0.6	1.6	2.6	0	↓	0	0	↓	0	↓	12.4	2.4
25th	↓	0	0.4	↓	0.4	0	↓	17.2	0	↓	16.8	0
26th	7.0	0	2.0	↓	0	0	↓	7.0	↓	1.2	0	↓
27th	0	4.4	5.4	23.0	0.2	↓	↓	21.4	↓	0	0	↓
28th	0	17.4	↓	7.2	0	↓	12.8	12.4	3.6	0	0	8.2
29th	0	↓	0	0	0	↓	0	↓	0	0	0	28.0
30th	0	32.4	1.0	↓	0.6	0	↓	0	0	0	0	0.4
31st	0	11.0	10.0	↓	0	6.8	↓	0	↓	0	0	0
Highest Daily	1.0	17.4	11.0	20.0	14.4	21.2	2.6	21.8	6.0	12.8	17.8	17.2
Monthly Total	10.2	67.4	94.4	106.4	75.0	73.0	34.8	145.4	55.2	40.6	57.4	108.2

Tableau 4 : Rainfall de Williamtown, année 2014 – site du Bureau of Meteorology (conforme à notre dataset)

Cela a deux conséquences :

- à l'issue d'une plage de dates non renseignées, si Rainfall est inférieure à 1mm, alors elle est également inférieure à 1 pour chaque journée de la plage concernée. C'est le cas du 27 au 30 juin.
- Lorsque Rainfall est correctement renseignée pour une date donnée, sa valeur ne porte en réalité pas sur ce jour, mais est à répartir sur les jours qui précédent. Dans l'exemple ci-dessus, il n'a en réalité pas plus 27mm le 18 février : il a plu 27mm au total entre le 5 et le 27 février. Il est même possible qu'il n'ait pas plu du tout le 27 février.

Ce constat nous permet de savoir que lorsque il a plu moins d'un millimètre une certaine journée, alors il a également plu moins d'un millimètre les jours précédent ayant une valeur non renseignée pour Rainfall.

Cela nous permet donc de savoir que dans ce type de situation, les NA de RainToday peuvent être remplacés par False. De même, RainTomorrow de la veille pourra être affecté à False.

En revanche, lorsque la valeur de Rainfall qui suit une séquence de NA est supérieure à un millimètre, il est impossible de déterminer la répartition de la pluviométrie sur la plage de dates. Nous laisserons donc en NA les Rainfall dans cette seconde situation.

3.1.6 Imputation des données manquantes

Une fois les étapes initiales de suppression terminées, trois méthodes d'imputation distinctes ont été envisagées pour traiter les données manquantes restants dans les variables numériques.

Imputation par la Moyenne : cette méthode remplace les valeurs manquantes par la moyenne de la colonne correspondante.

Imputation par la Médiane : les valeurs manquantes sont remplacées par la médiane de la colonne.

Les graphiques Figure 22 et Figure 23 illustrent les distributions des variables avant et après l'imputation. Sur chaque graphique :

- La courbe noire dépeint la distribution de la variable en présence des données manquantes.
- Les courbes rouge et bleue illustrent respectivement les distributions après l'imputation par moyenne et par médiane.

Les méthodes d'imputation par la moyenne ou par la médiane sont certes simples à mettre en œuvre. Toutefois, comme le montrent les graphiques elles peuvent altérer considérablement la distribution des variables, en particulier lorsque celles-ci présentent un taux élevé de données manquantes. De plus, rappelons que nous avons fait le choix de conserver les outliers, ce qui implique une instabilité de la moyenne.

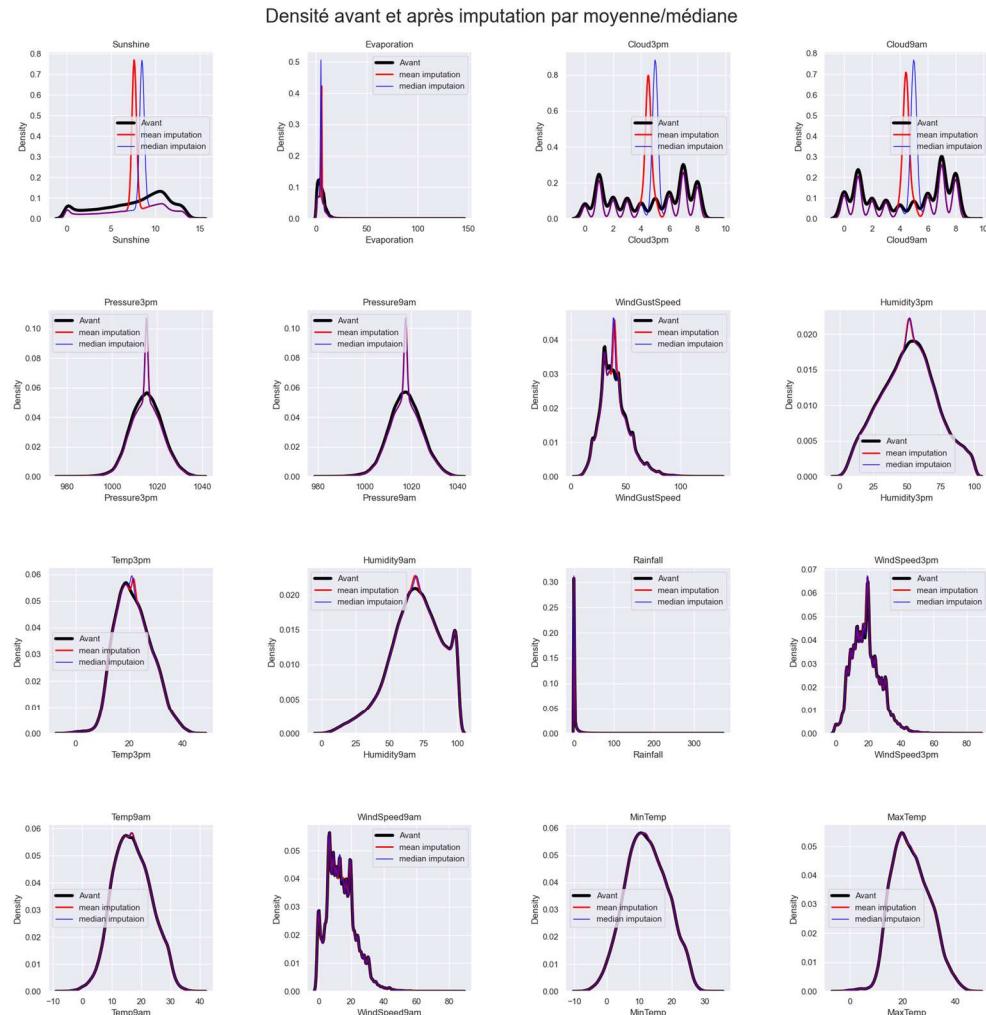


Figure 22: Distribution des variables avant et après l'imputation par moyenne/médiane

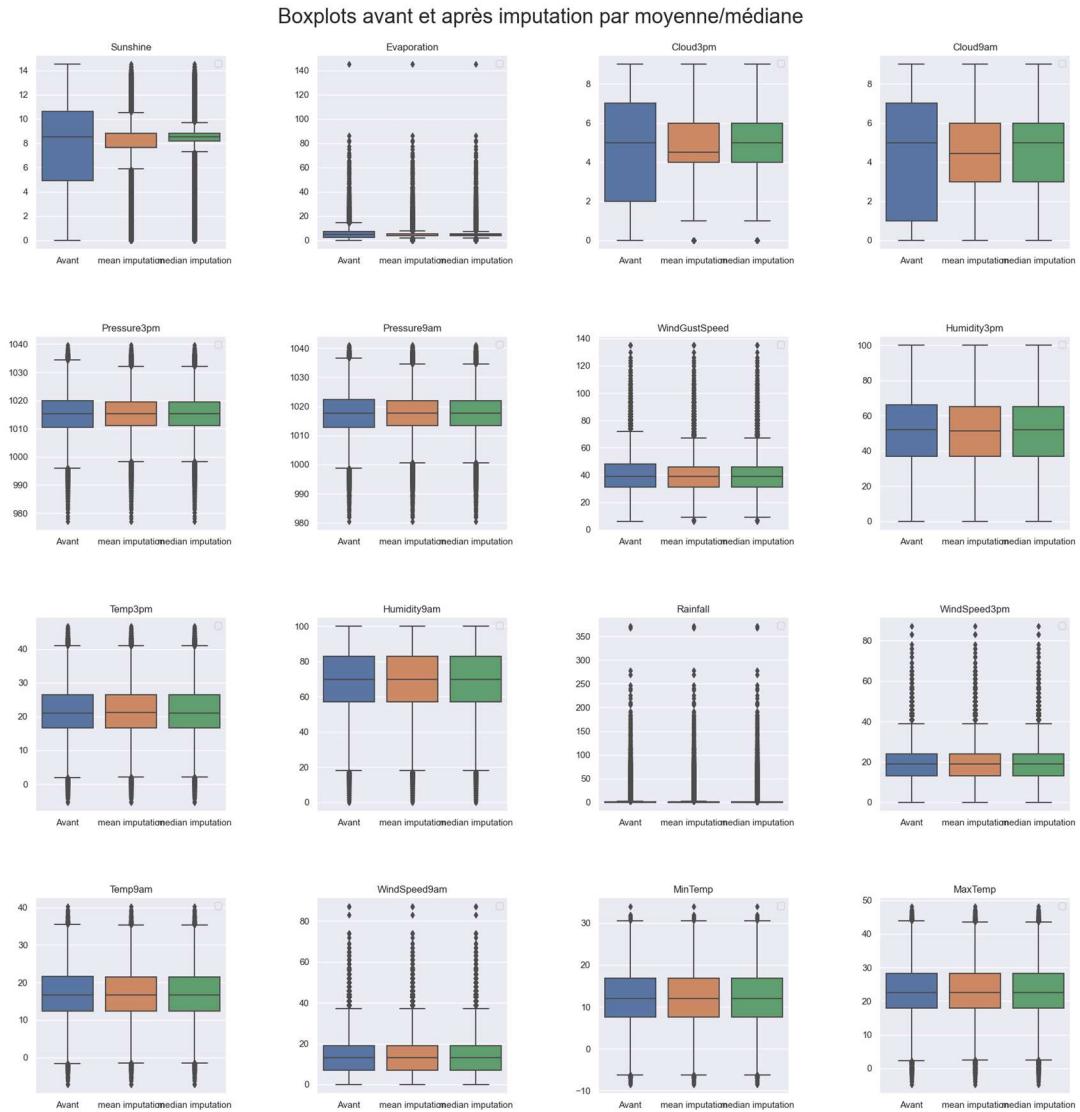


Figure 23 : Boxplots des variables avant et après l'imputation par moyenne/médiane

Imputation KNN : L'imputation basée sur les k plus proches voisins est une méthode plus sophistiquée qui prend en compte les similarités entre les observations pour imputer les données manquantes. Au lieu de remplir avec une valeur unique (comme la moyenne ou la médiane), elle utilise les k observations les plus similaires pour estimer la valeur manquante. Bien que cette méthode puisse être plus précise, elle peut être assez gourmande en temps et en ressources, en particulier pour de grands ensembles de données. La raison est en effet, pour chaque point avec une valeur manquante, l'algorithme KNN essaie de trouver les « k » voisins les plus proches en calculant la distance entre le point cible et tous les autres points. Ensuite, il utilise ces « k » voisins pour imputer la valeur manquante. Pour un ensemble de données de taille 145 460 lignes dont il y a environ 60% de lignes contenant au moins une valeur manquante, cela signifie potentiellement des milliards de calculs de distance.

Pour déterminer la valeur optimale du paramètre k dans l'imputation KNN, nous avons testé différentes valeurs pour k , allant de 2 à 4. Les graphiques ci-dessous illustrent les distributions des variables avant et après l'imputation. Sur chaque graphique :

- La courbe noire dépeint la distribution de la variable en présence des données manquantes.
- Les courbes verte, rouge et bleue illustrent respectivement les distributions après l'imputation KNN pour $k = 2, 3, 4$.

Il est à noter que pour les variables, *Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm*, qui présentent un pourcentage élevé de valeurs manquantes, l'imputation KNN impacte plus sensiblement leurs. En revanche, pour les autres variables avec moins de 10% de valeurs manquantes, l'imputation KNN ne perturbe pas de manière significative leur distribution initiale.

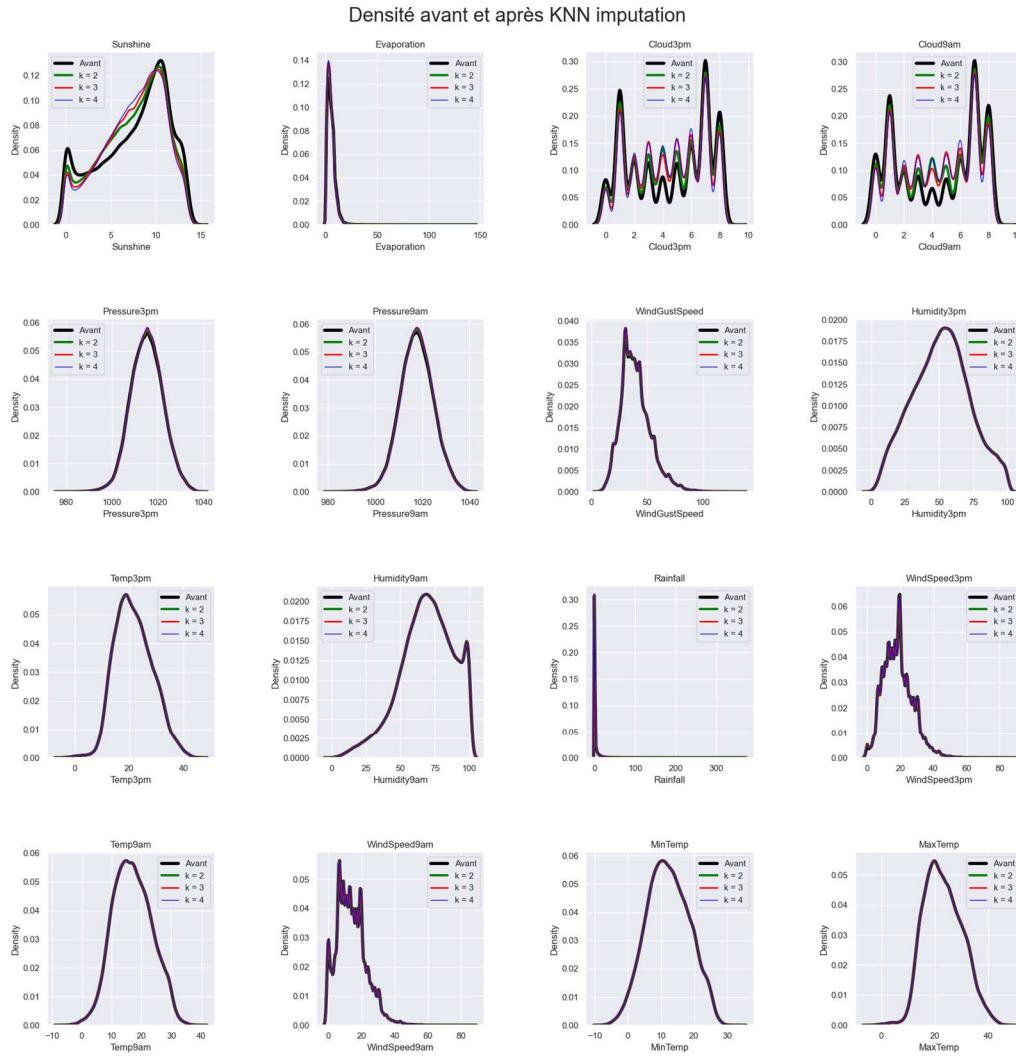


Figure 24 : Densités des variables avant et après l'imputaion KNN

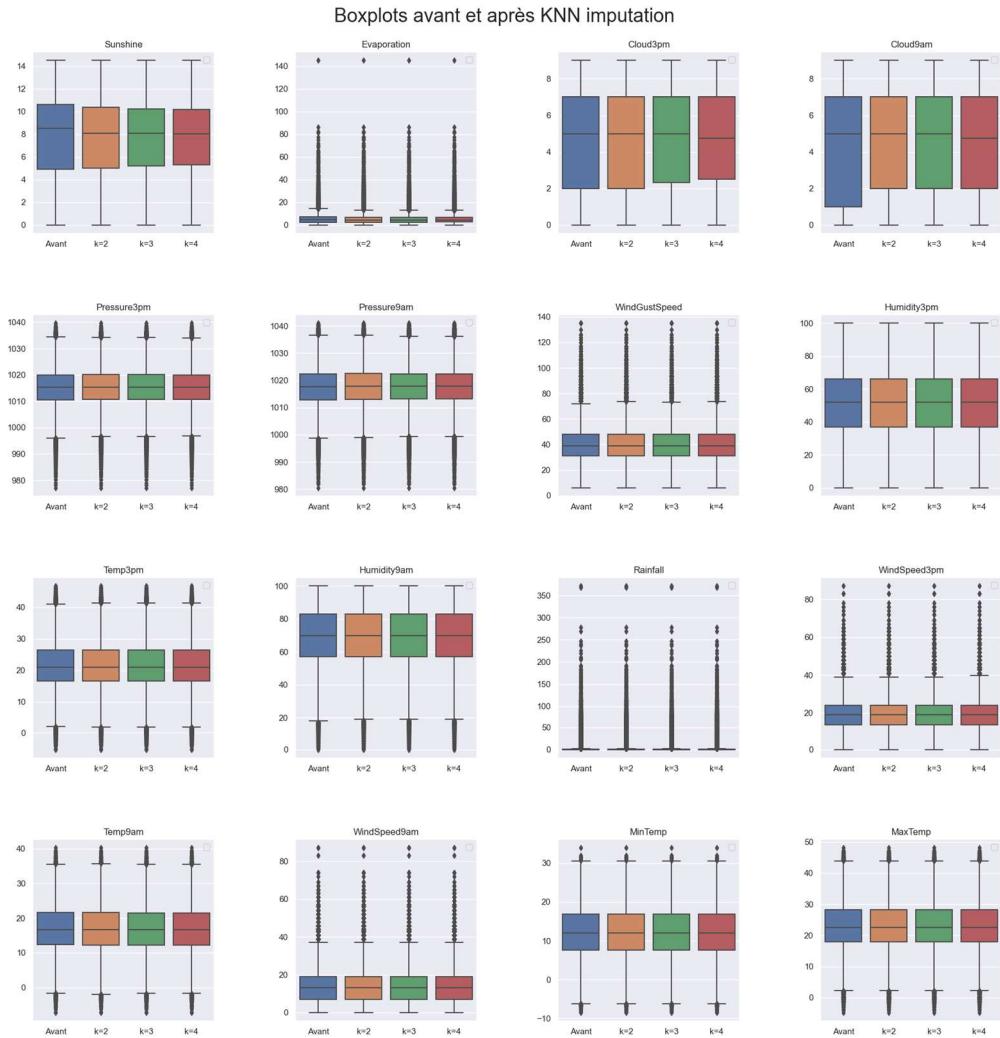


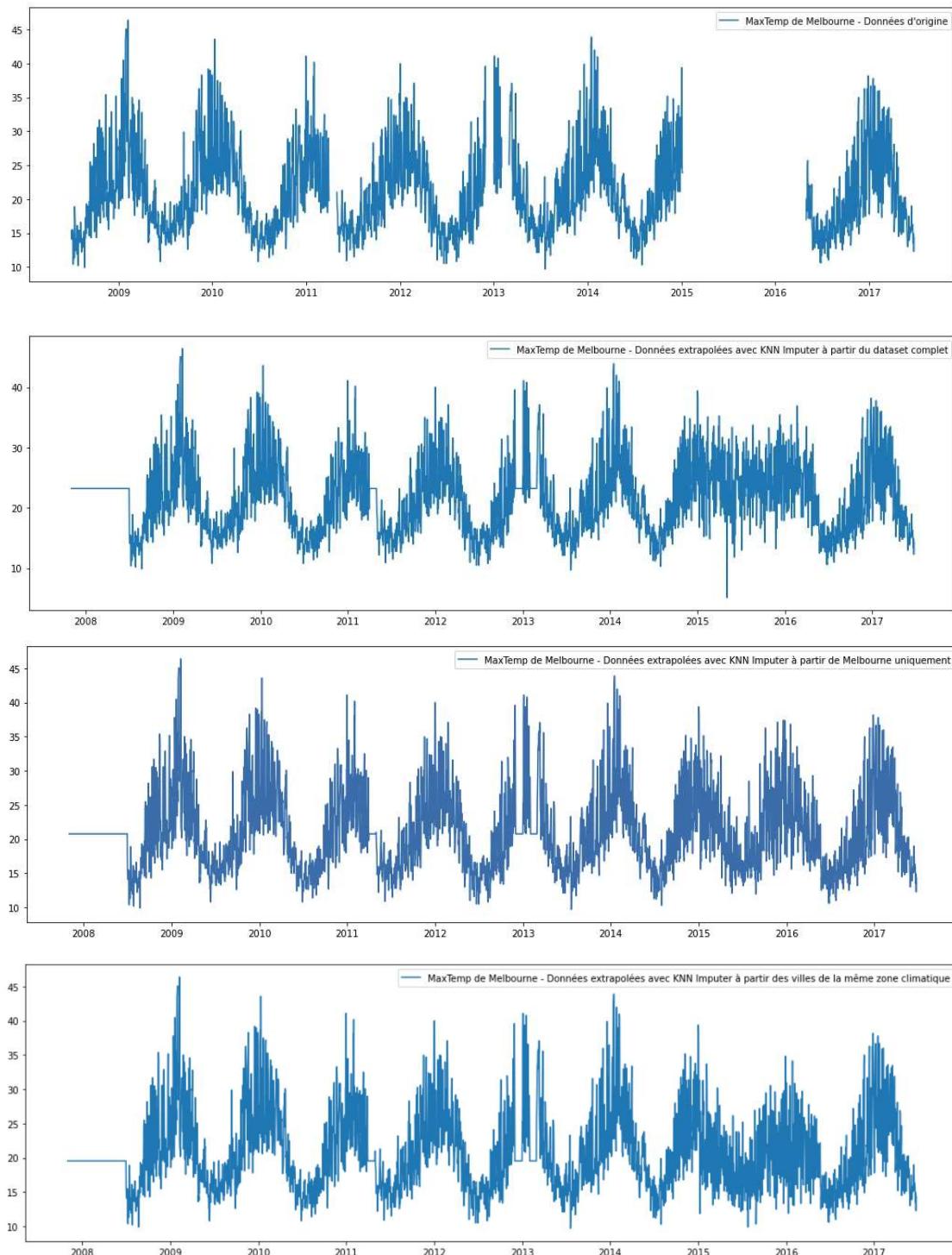
Figure 25: Boxplot des variables avant et après l'imputation KNN, $k = 2, 3, 4$

Après avoir analysé les différentes méthodes d'imputation, il est clair que l'imputation KNN offre des résultats plus fidèles et cohérents comparativement aux méthodes d'imputation par la moyenne ou la médiane. Les distorsions introduites par ces deux dernières techniques, particulièrement visibles dans notre contexte, rendent l'imputation KNN nettement supérieure en matière de préservation de la structure initiale des données. Par conséquent, nous avons décidé d'adopter l'imputation KNN comme méthode privilégiée pour traiter les données manquantes dans cet ensemble de données.

Du fait des diversités climatiques et des spécificités locales vu dans la première partie, nous avons observé les résultats de la KNN imputation effectuée dans un premier temps sur l'ensemble du dataset, puis dans un second temps uniquement pour une ville donnée. Nous voyons ici le graphique de MaxTemp de Melbourne :

- Le premier graphe indique les données originales du dataset, débutant mi 2008, comportant un trou de début 2015 à mi 2016, ainsi que des trous pour les mois d'avril 2011, décembre 2012 et février 2013
- Le second graphe montre le résultat de la KNN imputation ($k=3$) à partir de l'ensemble du dataset : les données de 2015 à mi 2016 sont renseignées d'une façon non satisfaisante, les autres plages manquantes sont remplacées par une valeur unique (temps d'exécution d'environ 30 minutes)

- Le troisième graphe a été réalisé sur un KNN imputer ($k=3$) uniquement sur les données de Melbourne : la plage de 2015 à mi 2016 est renseignée de façon plus satisfaisante, mais les autres plages manquantes sont également remplies par une valeur unique (temps d'exécution instantané)
- Le quatrième graphe a réalisé l'imputation à partir des données des Location de la même zone climatique (temps d'exécution d'environ 30 secondes)



3.2 Transformation des données

3.2.1 Booléens

Nous avons deux variables booléennes : *RainToday* et *RainTomorrow*. De façon assez classique, nous allons remplacer les True par des 1 et les False par des 0.

3.2.2 Directions du vent

Nous avons trois variables (*WindGustDir*, *WindSpeed9am*, *WindSpeed3pm*) donnant la direction du vent selon 16 modalités. Une possibilité est d'effectuer un encodage OneHot, ce qui aboutira au remplacement de ces 3 variables par 45 nouvelles. C'est évidemment considérable.

Une autre approche consiste à considérer le vent selon à partir d'une approche trigonométrique. Un vent ENE pourra ainsi être vu comme un vent d'angle $\pi/8$, un vent sud (S) pourra être considéré comme un vent d'angle $3/2\pi$, etc. L'angle ne permettra cependant pas au modèle de percevoir qu'un angle de $15/8\pi$ est très proche d'un angle de 0. Plutôt que l'angle, nous allons donc considérer les composantes directionnelles X et Y, grâce respectivement au cosinus et au sinus de l'angle du vent.

Ce faisant, nous substituerons seulement 6 nouvelles variables numériques aux trois variables qualitatives.

De plus, allons multiplier ces nouvelles variables par les trois variables de vitesse de vent. Nous disposerons ainsi dans nos trois nouvelles variables de la vitesse du vent pour chaque composante directionnelle, et nous pourrons éventuellement supprimer les trois variables numériques de vitesse initiales.

3.3 Ajout de variables

3.3.1 Coordonnées des villes

Comme vu en début de rapport, nous avons rapidement recherché les coordonnées de chaque ville afin de disposer d'une représentation graphique. Outre leur intérêt visuel, ces variables peuvent avoir un véritable intérêt pour le modèle. Nous avons en particulier identifié plus haut qu'il semblait y avoir un lien direct entre la latitude et la température maximale des villes.

Ces variables peuvent également permettre de supprimer la variable qualitative *Location*, qui dispose de 49 modalités, ce qui crée 49 variables une fois encodées en OneHot.

3.3.2 Amplitude thermique

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

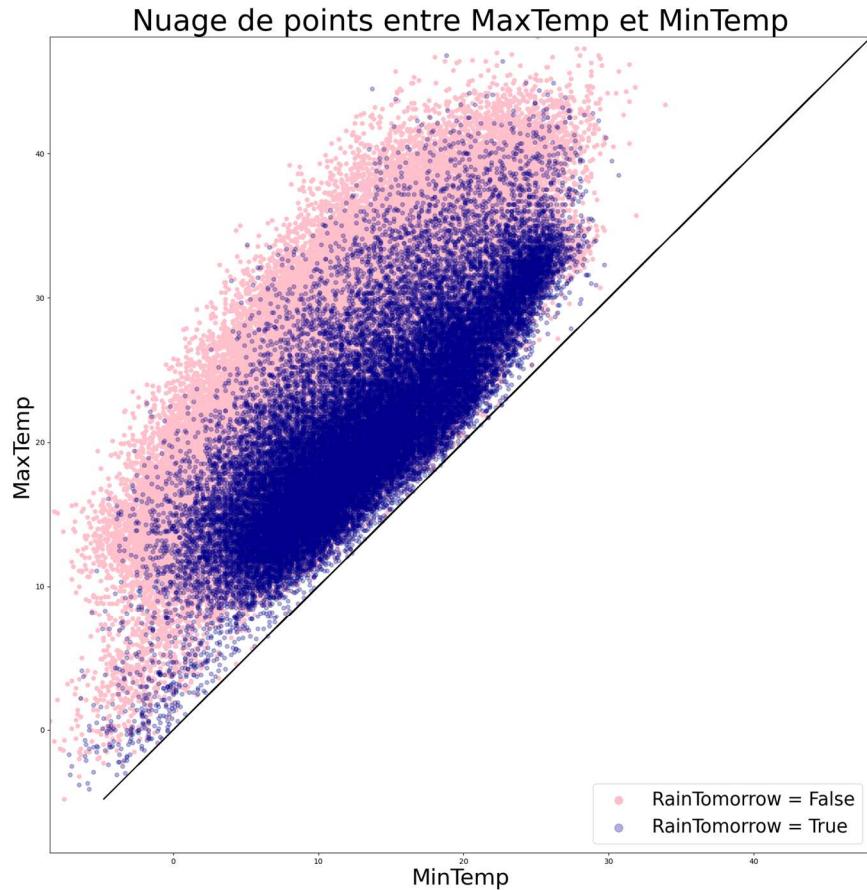


Figure 26 : Nuage de points entre MaxTemp et MinTemp

La Figure 26 trace en bleu les journées avec un RainTomorrow positif, et les positionne sur un graphe aux coordonnées (MinTemps, MaxTemp). Les points bleus semblent largement positionnés un peu au-dessus de la première bissectrice, ce qui signifie qu'une faible amplitude thermique pourrait être fortement associée au fait qu'il pleuve le lendemain. Ce n'est pour autant pas systématique, car il existe aussi de nombreux points bleus très au-dessus de cette droite. Il est donc peut-être 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 nous semble donc particulièrement intéressant d'ajouter cette feature.

3.3.3 Information climatique

Les différents climats australiens, qui se traduisent par des niveaux de températures et de précipitation très différents selon les lieux nous incitent à créer une variable catégorielle indiquant le type de climat pour chaque lieu.

Une approche simple serait de recherche sur Internet cette information, par exemple sur Wikipédia, ou bien par repérage géographique à partir de cartes de zones climatiques.

Nous allons plutôt opter par une approche par clusterisation, qui nous semble tout à la fois plus pertinente à mettre en œuvre afin de profiter des spécificités de notre jeu de données.

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'exploitation 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. Le second petit cluster est constitué de Cairns, Darwin, Katherine et Townsville. Il s'agit des 4 villes situées le plus au nord.

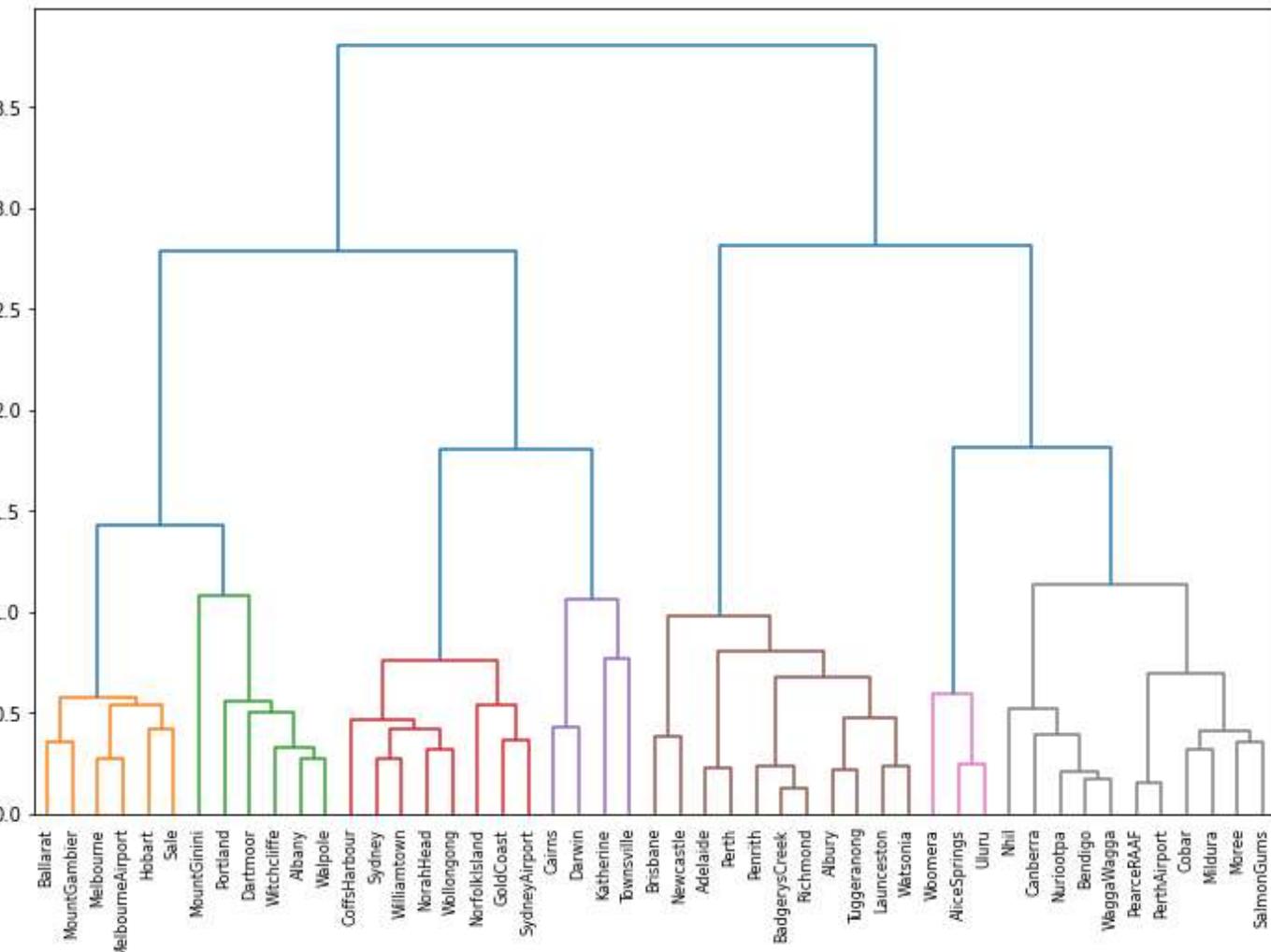


Figure 27 : Résultats de clusterisation par CAH

Regardons plus en détail sur la carte de la Figure 28 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.

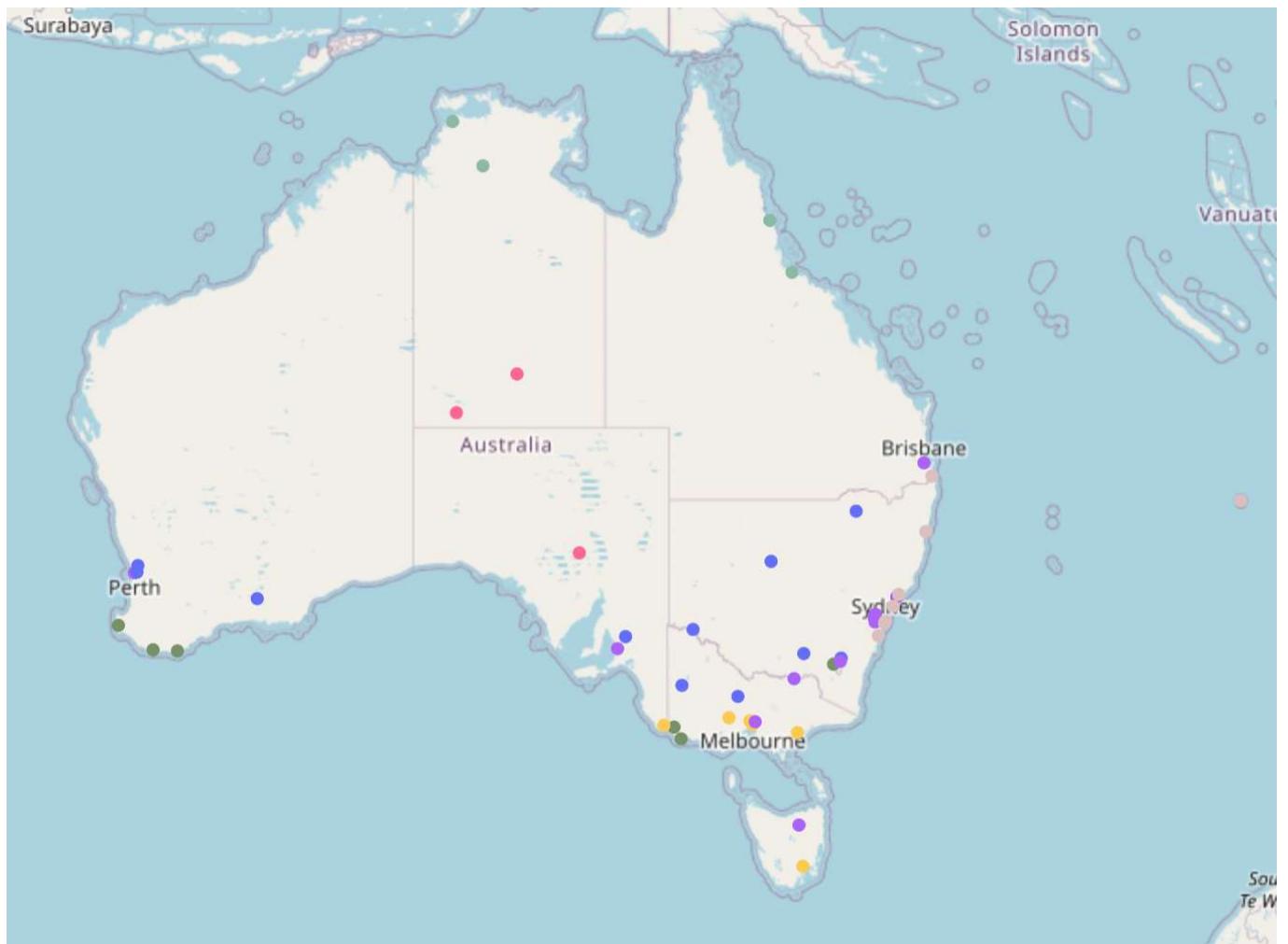
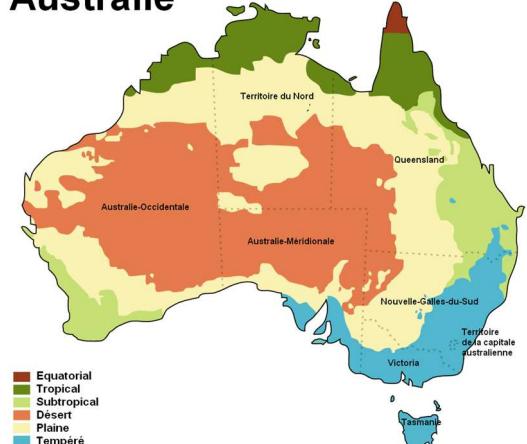


Figure 28 : La répartition des 7 clusters

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)

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.

3.3.4 Corrélation des nouvelles variables

Après ajout des nouvelles features et retrait des variables évoquées, voici la nouvelle matrice de corrélation :

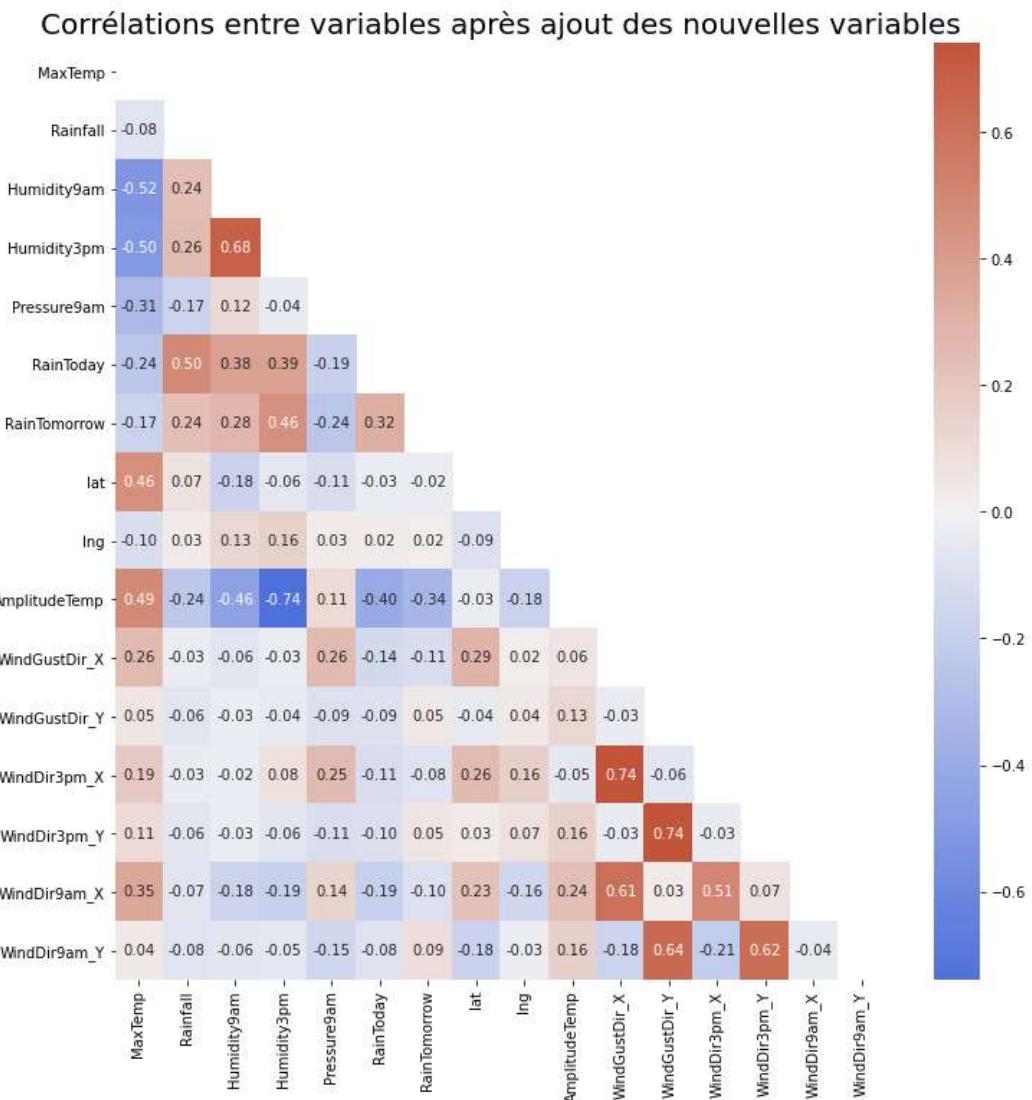


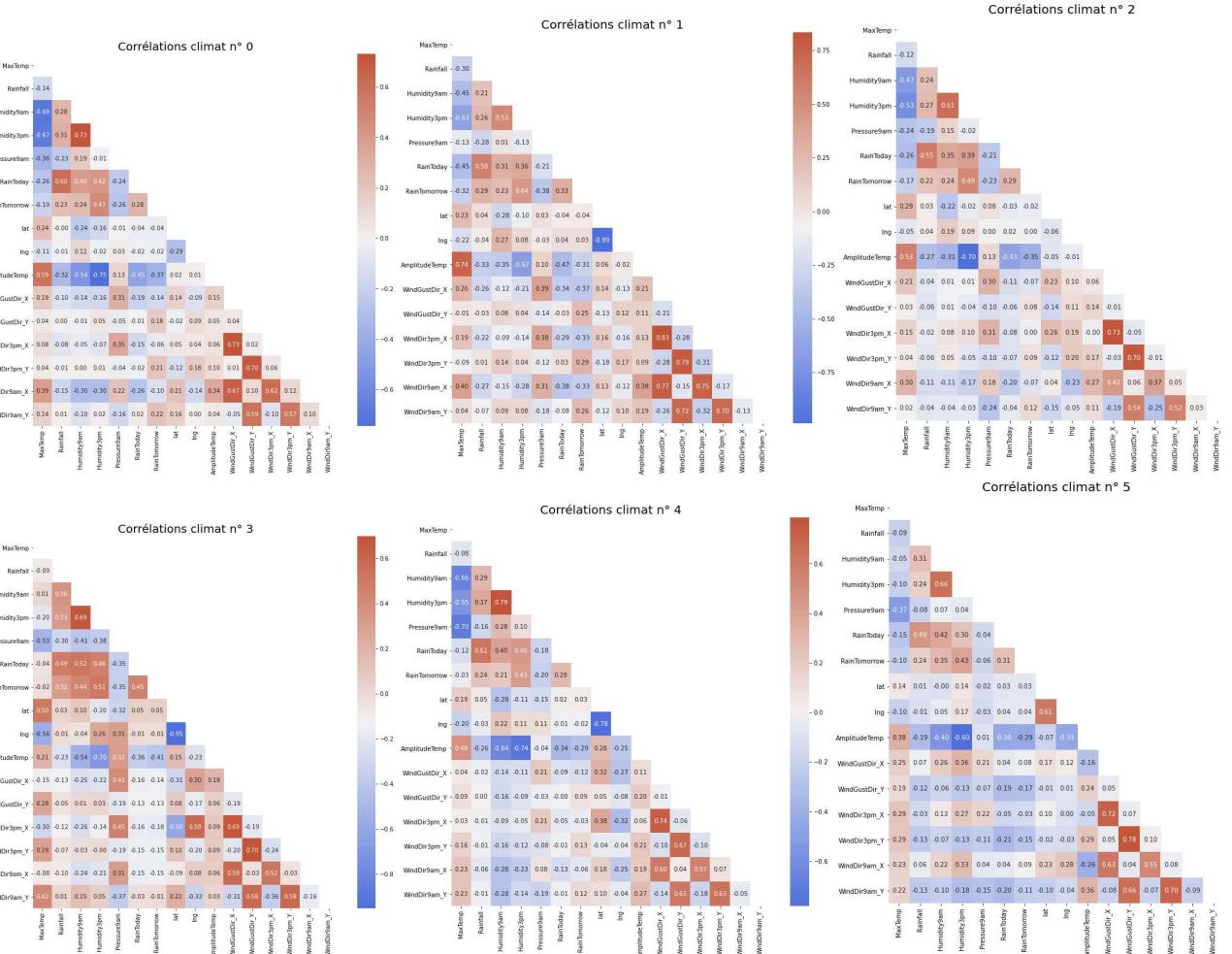
Figure 29 : Corrélations entre variables après ajout des nouvelles variables

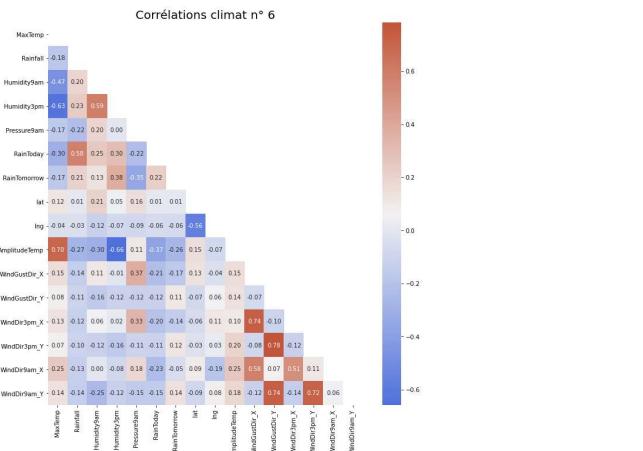
RainTomorrow et *Rainfall* présentent désormais respectivement une corrélation de -0,34 et -0,24 avec la nouvelle variable d'amplitude thermique (*AmplitudeTemp*). Leurs corrélations avec les composantes des vents sont très faibles, de même qu'avec les latitude et longitude des villes.

MaxTemp présente des corrélations bien plus élevées avec les nouvelles variables : 0,49 pour l'amplitude thermique, entre 0,19 et 0,35 pour les composantes X du vent.

Notons une forte corrélation (-0,74) entre le taux d'humidité à 15h00 (*Humidity3pm*) l'amplitude thermique.

Par ailleurs, les corrélations varient suivant les climats que nous avons obtenus par clusterisation. Avec le climat n°1, les composantes X et Y des vents sont corrélées entre 0,25 et 0,37 en valeurs absolues. Dans le climat n°0, c'est la composante X des vents qui est corrélée avec *RainTomorrow*.





3.3.5 Normalisation et standardisation

Nous avons vu que les ordres de grandeur des variables étaient très différents. Il sera donc tout à fait indispensable de standardiser et normaliser les données servant à prédire la variable cible. Notez que cette opération était déjà nécessaire pour la KNN imputation faite plus haut.

4 Conclusion

Le jeu de données dispose d'informations très variées, avec des taux de données non renseignées assez hétérogènes selon les variables, les dates et les lieux.

La principale difficulté a consisté à analyser les données selon plusieurs axes, notamment temporel et géographique, et à trouver des représentations qui permettent de rendre compte de l'état des données pour les différentes variables selon des spécificités locales.

Des approches assez diverses ont ainsi été mises en œuvre pour proposer des visualisations variées, en prenant également en compte le type de représentation visuelle utilisées par des météorologues (diagramme climatique et représentation polaire des directions des vents), la prise en compte des habitudes professionnelles des interlocuteurs des data scientists étant un élément important dans la bonne communication et compréhension des résultats.

Nos analyses nous ont permis d'arriver à une conclusion sur une méthodologie de gestion des valeurs manquantes, à savoir :

- Télécharger des données complémentaires pour des variables et des lieux ciblés
- Remplacer par des 0 les NA de la variable Rainfall d'une plage de date précédent immédiatement une date pour laquelle Rainfall est inférieure à 1mm, et mettre à False RainToday sur cette même plage et à False RainTomorrow la plage un jour plus tôt
- Supprimer les lignes pour lesquelles la variable cible RainTomorrow reste non renseignée
- Enlever les lignes sur lesquelles plus de 50% des variables sont non renseignées
- Imputer les NA restants grâce à KNN Imputer

Nous avons également pu enrichir le jeu de données par des variables complémentaires (latitude et longitude des villes, zones climatiques, amplitude thermique, composantes directionnelles du vent) afin d'aider les modèles de prédiction.

Nos données sont désormais nettoyées, enrichies, et prêtes à alimenter nos modèles de prédiction. Pour autant, nous avons conscience que, selon les résultats de nos modélisations, des ajustements seront certainement à apporter ces prochaines semaines sur le traitement des données que nous avons effectué.