Projet de Data Science : Prévision météo en Australie

Rendu 1 : Introduire, décrire, visualiser et manipuler le jeu de données

Alexandre Winger, Frédéric Vincent, Omar Choa

Octobre 2023

Table des matières

1. Introduction au projet

1.1. Contexte

Notre projet fait partie du cursus de notre formation Data Scientist avec DataScientest.

Notre projet est issu d'une compétition Kaggle, *Rain in Australia*, qui fournit un ensemble de données contenant environ 10 ans d'observations météorologiques quotidiennes provenant de nombreux endroits en Australie. L'objectif de cette compétition Kaggle était de prédire s'il va pleuvoir le jour suivant.

La météorologie est une science avec des applications dans des domaines très divers comme les besoins militaires, la production d'énergie, les transports (aériens, maritimes et terrestres), l'agriculture, la médecine, la construction, la photographie aérienne ou le cinéma. Elle est également appliquée pour la prévision de la qualité de l'air ou de plusieurs risques naturels d'origine atmosphérique.

La météorologie étudie les phénomènes atmosphériques tels que les nuages, les précipitations ou le vent dans le but de comprendre comment ils se forment et évoluent en fonction des paramètres mesurés tels que la pression, la température et l'humidité. Cette discipline scientifique s'appuie sur notamment sur la mécanique des fluides, la thermodynamique, la chimie et les mathématiques.
Grâce à l'informatique et aux simulations numériques, la météorologie moderne permet d'établir des prévisions de l'évolution du temps en s'appuyant sur des modèles mathématiques à court comme à long terme qui assimilent des données de nombreuses sources dont les stations, les satellites et les radars météorologiques.

1.2. Objectifs

Notre projet reprend l'objectif initial de la compétition Kaggle qui l'a inspiré : prédire s'il va pleuvoir le jour suivant. Cet objectif est enrichi de deux autre objectifs : les objectifs secondaires sont de prédire le vent et la température pour le jour suivant.

Chacun de nous trois a un bagage scientifique, mais aucun de nous n'a d'expérience professionnelle dans le domaine de la météorologie.

Notre projet consiste à résoudre un problème d'apprentissage automatique (*machine learning*), qui comprend des problèmes de classification (prédire s'il va pleuvoir le jour suivant, prédire la direction du vent) ainsi que des problèmes de régression (prédire la vitesse du vent, prédire la température).

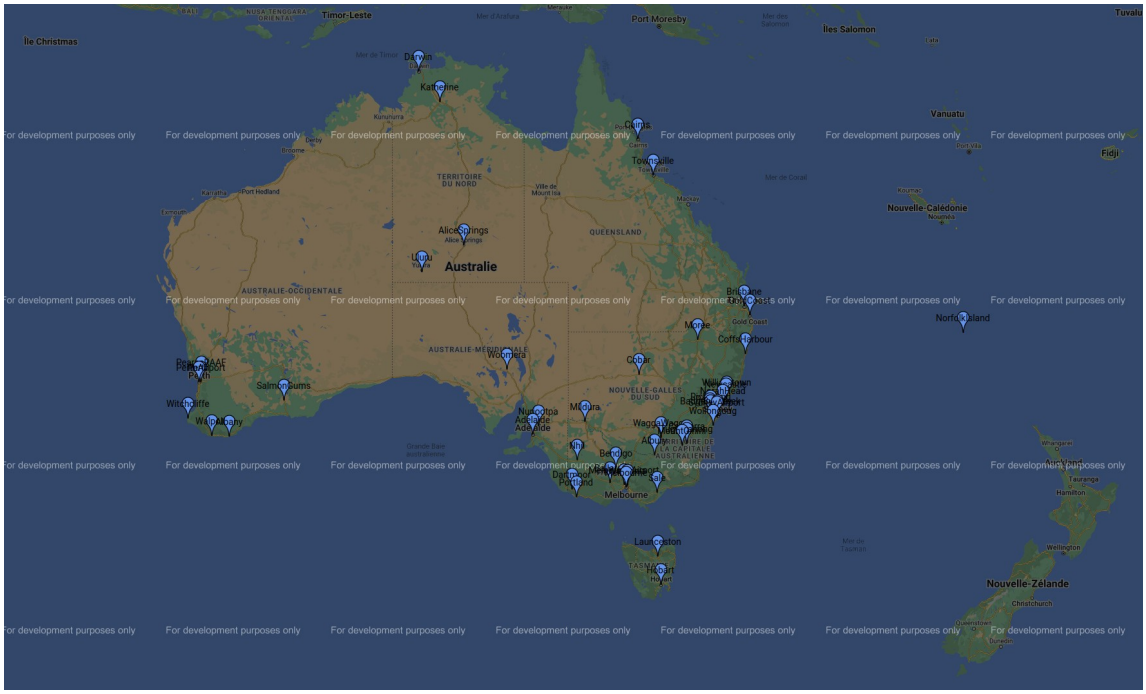
Utiliser des techniques d'apprentissage automatique à la météorologie permet de créer un outil qui répond à un besoin quotidien (connaître les prévisions météo) et qui vient compléter les outils existants (simulations numériques des phénomènes atmosphériques). Un objectif supplémentaire serait d'avoir un outil qui consomme moins de ressources informatiques (temps de calcul, capacité de stockage) que les simulations numériques qui peuvent être lourdes et gourmandes en ressources.

2. Compréhension et manipulation des données

2.1. Présentation globale du jeu de données

2.1.1. Etude géographique

Le jeu de données initiales contient environ 10 années d'observations quotidiennes de la météo pour 49 stations en Australie. L'Australie est un pays sec, il y a donc de grandes étendues sans aucune station météo, les stations sont réparties dans les zones fertiles du pays, comme le montre la carte ci-dessous.



Nous regroupons les stations sur 7 différents territoires qui ont une cohérence géographique. Chacun de ces territoires peut être subdivisé en différentes zones. Ces différentes zones regroupent des stations qui sont géographiquement proches et donc corrélées spatialement. Ci-dessous, la liste des territoires et des zones associées.

- Territoire Sud-Est (33) :
  - zone de Brisbane (4) : Brisbane, Coffs Harbour, Gold Coast, Moree
  - zone de Sidney (9) : Badgerys Creek, Newcastle, Norah Head, Pernth, Richmond, Sidney, Sidney Airport, William Town, Wollongong
  - zone de Canberra (5) : Albury, Canberra, Mount Ginini, Tuggeranong, Wagga Wagga
  - zone de Melbourne (6) : Ballarat, Bendigo, Melbourne, Melbourne Airport, Sale, Watsonia
  - zone de Portland (3) : Dartmoor, Mount Gambier, Nhil, Portland
  - zone d'Adelaïde (2) : Adelaide, Nuriootpa
  - Cobar (1)
  - Mildura (1)
  - Woomera (1)
- Territoire Sud-Ouest (7) :
  - zone de Perth (3) : Perth, Perth Airport, Pearce RAAF
  - Pointe Sud-Ouest (3) : Albany, Walpole, Wichcliffe
  - Salmon Gums
- Territoire Centre (2) : Alice Spring, Uluru
- Territoire Plein Nord (2) : Darwin, Katherine
- Territoire Nord-Est (2) : Cairns, Townsville
- Territoire de l'Ile de Tasmanie (2) : Hobart, Launceston

Ces corrélations spatiales peuvent induire des corrélations dans les observations météo. Nous prévoyons de tester cette hypothèse.

2.1.2. Contenu du jeu de données

Le jeu de données contient 145 460 lignes et 23 colonnes. La volumétrie est plutôt légère (14,1 Mo).

Les deux premières colonnes sont particulières. La première colonne est la date du jour. La deuxième colonne est le nom de la station où sont faites les mesures. Les 21 autres colonnes sont des variables, numériques (quantitatives) ou catégorielles (qualitatives), dont la variable cible.

Chaque ligne du jeu de données correspond à une observation des 21 variables pour une station donnée et un jour donné.

Dans ces 23 variables, nous avons 16 variables numériques et 7 variables catégorielles.

Liste des 16 variables numériques :

- *Cloud3pm* : fraction du ciel obscurcie par les nuages à 9h (octas)
- *Cloud9am* : fraction du ciel obscurcie par les nuages à 15h (octas)
- *Evaporation* : quantité d'eau évaporée dans un bac de classe A sur les dernières 24h, mesurée à 9h00 (mm)
- *Humidity3pm* : humidité mesurée à 15h (%)
- *Humidity9am* : humidité mesurée à 9h (%)
- *MaxTemp* : température maximale enregistrée ce jour (°C)
- *MinTemp* : température minimale enregistrée ce jour (°C)
- *Pressure3pm* : pression atmosphérique mesurée à 15h et corrigée au niveau de la mer (hPa)
- *Pressure9am* : pression atmosphérique mesurée à 9h et corrigée au niveau de la mer (hPa)
- *Rainfall* : quantité d'eau de pluie enregistrée ce jour (mm)
- *Sunshine* : nombre d'heures d'ensoleillement dans la journée (hrs)
- *Temp3pm* : température à 15h (°C)
- *Temp9am* : température à 9h (°C)
- *WindGustSpeed* : vitesse de la plus forte rafale de vent sur les dernières 24h (km/h)
- *WindSpeed3pm* : vitesse moyenne du vent de 14h50 à 15h (km/h)
- *WindSpeed9am* : vitesse moyenne du vent de 8h50 à 9h (km/h)

L’humidité mesurée est l’humidité relative : c’est un pourcentage qui représente le rapport de la pression partielle de la vapeur d'eau contenue dans l'air sur la pression de vapeur saturante (ou tension de vapeur) à la même température. Elle est donc une mesure du rapport entre le contenu en vapeur d'eau de l'air et sa capacité maximale à en contenir dans ces conditions.

Liste des 7 variables catégorielles :

- *Date* : date du jour des mesures
- *Location* : nom de la station où sont faites les mesures
- *WindDir3pm* : direction du vent à 15h (rose des vents à 16 directions)
- *WindDir9am* : direction du vent à 9h (rose des vents à 16 directions)
- *WindGustDir* : direction de la plus forte rafale de vent enregistrée sur les dernières 24h (rose des vents à 16 directions)
- *RainToday* : réponse à la question : « a-t-il plu ce jour ? » (booléen)
- *RainTomorrow* : réponse à la question : « a-t-il plu le lendemain ? » (booléen). C’est notre variable cible.

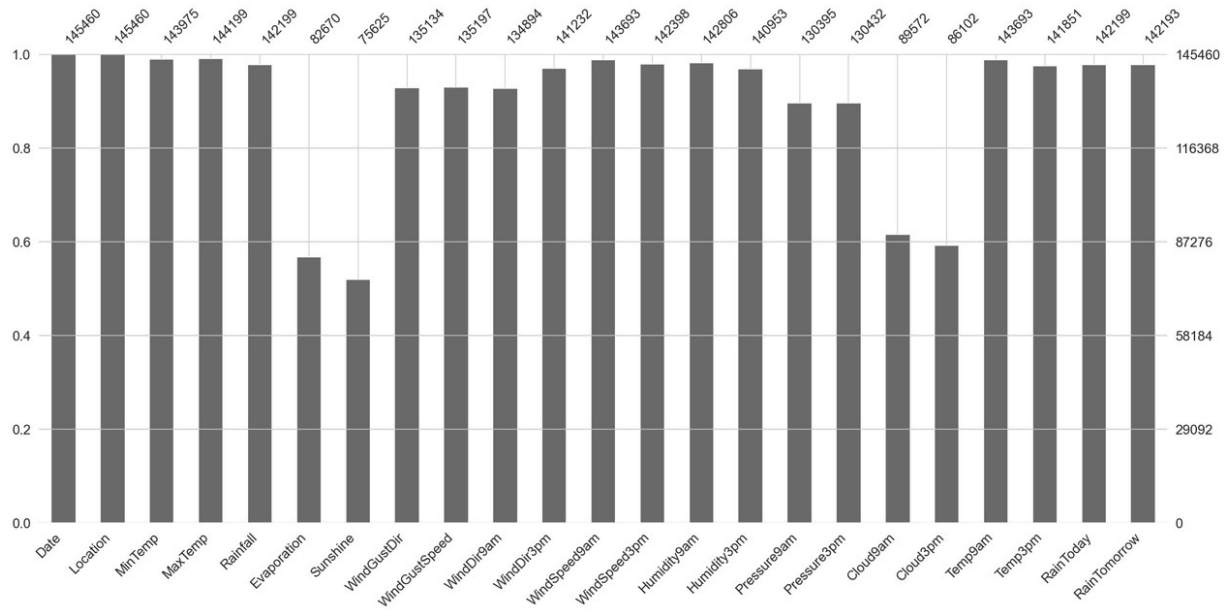
Nous nous sommes demandés si le problème pouvait être subdivisé en 49 problèmes, avec un sous-ensemble de données par station. Mais cette stratégie aurait pour conséquence de faire perdre toutes les informations sur les corrélations spatiales entre les différents sous-ensembles de données. Par exemple, le sous-ensemble « Melbourne » serait complèment scindé du sous-ensemble « Melbourne Airport», alors qu’il est tout à fait raisonnable de penser que les observations de ces deux sous-ensembles soient corrélées, puisque ces deux stations sont séparées par une distance d’une quinzaine de kilomètres.

Une stratégie plus fine pourrait être d’avoir des sous-ensembles qui regroupent les stations qui sont géographiquement proches. Il y aura probablement des corrélations sur les observation entre Melbourne et Melbourne Airport qui sont séparés par une distance d’une quinzaine de kilomètres, par contre il serait étonnant d’avoir des corrélations entre Melbourne et Perth qui sont séparés par une distance supérieure à 3 000 km.

2.2. Étude des valeurs manquantes

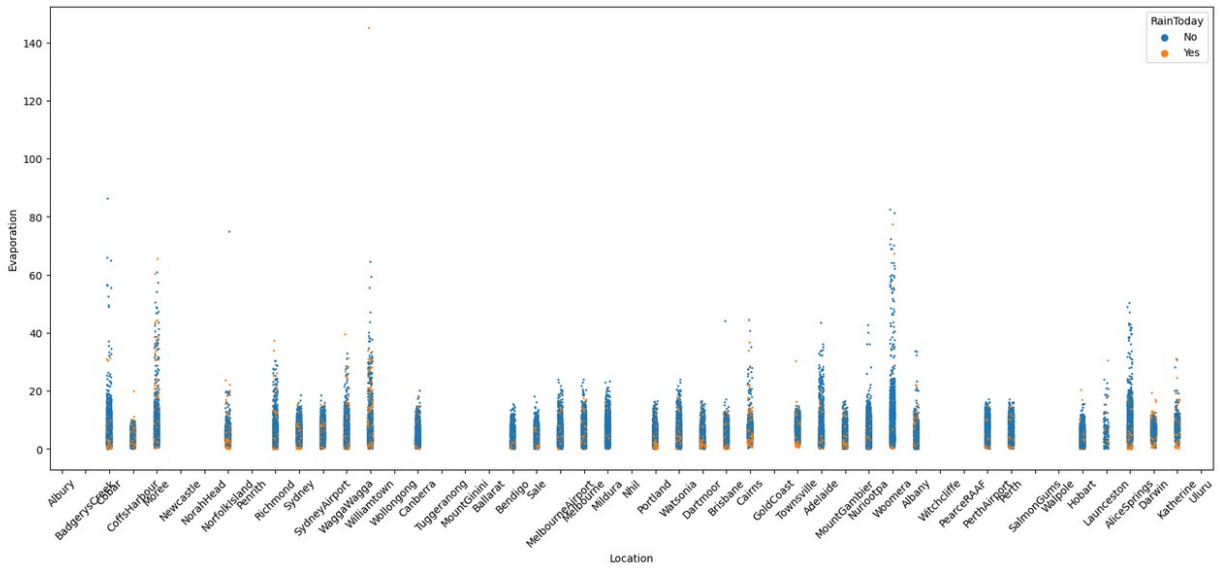
2.2.1. Mise en évidence des valeurs manquantes

Les premières visualisations du jeu de données montrent qu'il y a un certain nombre de valeurs manquantes. Le graphique ci-dessous représente le pourcentage de valeurs manquantes pour chaque variable.



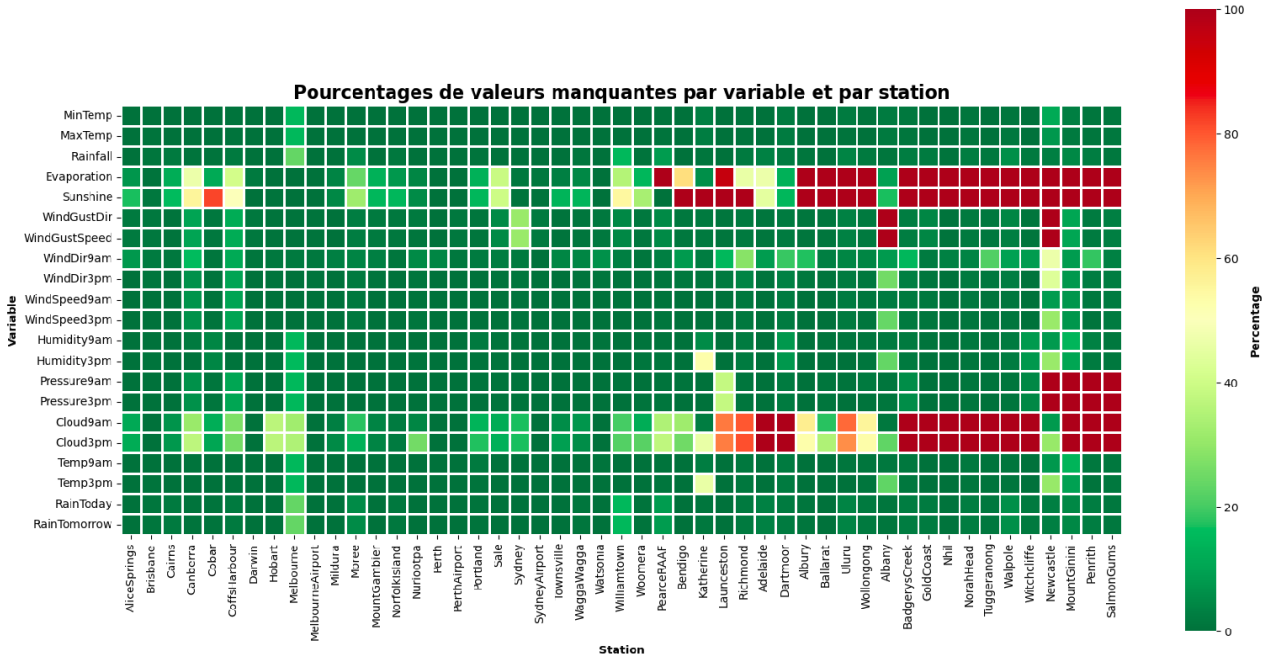
Nous constatons que 4 variables (*Evaporation*, *Sunshine*, *Cloud9am*, *Cloud3pm*) sont très peu mesurées, avec environ 40 % de valeurs mesurées.

La figure ci-dessous montre que la variable *Evaporation* est systématiquement absente d'un certain nombre de stations.



2.2.2. Représentation des valeurs manquantes par variable et par station

Pour systématiser l'étude, nous avons représenté une « *heatmap* » des valeurs manquantes par variable et par station, dans la figure suivante.



Dans certaines stations, des variables sont systématiquement non mesurées, pour des raisons que nous ignorons. Il est probable que ce soit dû à un manque de moyens matériels : des stations peuvent ne pas disposer des instruments de mesure nécessaires.

Nous avons distingué 9 groupes de stations en fonction des variables mesurées. Cette distinction donne lieu à une division du jeu de données en 9 sous-ensembles de données.

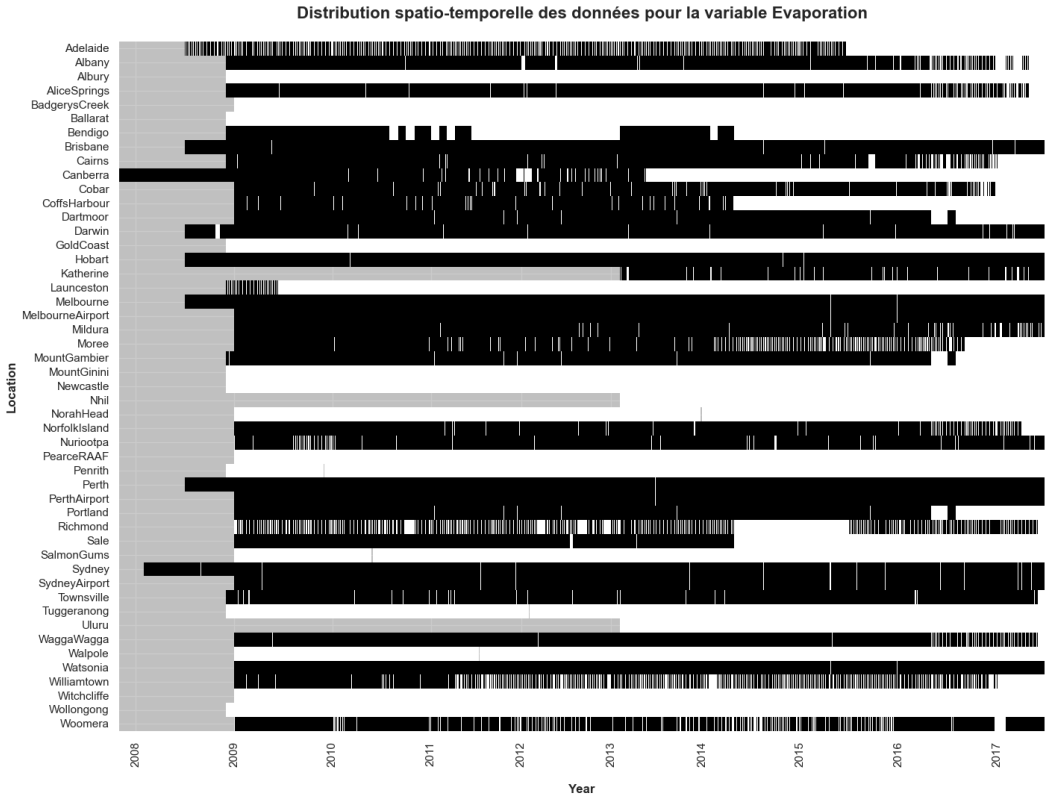
- Groupe 1 : toutes les variables sont mesurées. 26 stations.
- Groupe 2 : seule la variable *Evaporation* n'est pas mesurée. 1 station.
- Groupe 3 : seule la variable *Sunshine* n'est pas mesurée. 4 stations.
- Groupe 4 : seules les variables *Cloud3pm* et *Cloud9am* ne sont pas mesurées. 2 stations.
- Groupe 5 : seules les variables *Evaporation* et *Sunshine* ne sont pas mesurées. 4 stations.
- Groupe 6 : seules les variables *WindGustDir* et *WindGustSpeed* ne sont pas mesurées. 1 station.
- Groupe 7 : les variables *Cloud3pm*, *Cloud9am*, *Evaporation*, *Sunshine* ne sont pas mesurées. 7 stations.
- Groupe 8 : les variables *Evaporation*, *Pressure3pm*, *Pressure9am*, *Sunshine*, *WindGustDir*, *WindGustSpeed* ne sont pas mesurées. 1 station.
- Groupe 9 : les variables *Cloud3pm*, *Cloud9am*, *Evaporation*, *Pressure3pm*, *Pressure9am*, *Sunshine* ne sont pas mesurées. 3 stations.

Ces 9 groupes peuvent donner lieu à 9 familles de modèles pour le travail de modélisation et de prédiction.

2.2.3. Représentation de la distribution spatio-temporelle des valeurs manquantes

Nous avons aussi investigué la distribution des valeurs manquantes dans l'espace et le temps.

La figure suivante montre la distribution spatio-temporelle des valeurs manquantes pour la variable *Evaporation*. Ces valeurs manquantes sont représentées par les bandes blanches.



Les stations qui ne mesurent pas du tout la variable *Evaporation* sont bien visibles avec des bandes blanches sur toute la longueur.

Nous remarquons que des stations comme Bendigo et CoffsHarbour arrêtent de mesurer la variable *Evaporation* à partir de 2014. Il n'y a aucune explication sur l'arrêt de ces mesures.

2.3. Description statistique

2.3.1. Variables numériques

Comme indiqué en introduction, le jeu de données contient **16 variables numériques,** ainsi que 7 **variables catégorielles** que nous choisissons de traiter comme des variables numériques pour cette étude.

Table 1 présente des statistiques descriptives pour l'ensemble de ces **16 variables.**

Table 1. Statistiques descriptives pour les 16 variables numériques.

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine
count	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000
mean	12.194034	23.221348	2.360918	5.468232	7.611178
std	6.398495	7.119049	8.478060	4.193704	3.785483
min	-8.500000	-4.800000	0.000000	0.000000	0.000000
25%	7.600000	17.900000	0.000000	2.600000	4.800000
50%	12.000000	22.600000	0.000000	4.800000	8.400000
75%	16.900000	28.200000	0.800000	7.400000	10.600000
max	33.900000	48.100000	371.000000	145.000000	14.500000

	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
count	135197.000000	143693.000000	142398.000000	142806.000000	140953.000000
mean	40.035230	14.043426	18.662657	68.880831	51.539116
std	13.607062	8.915375	8.809800	19.029164	20.795902
min	6.000000	0.000000	0.000000	0.000000	0.000000
25%	31.000000	7.000000	13.000000	57.000000	37.000000
50%	39.000000	13.000000	19.000000	70.000000	52.000000
75%	48.000000	19.000000	24.000000	83.000000	66.000000
max	135.000000	130.000000	87.000000	100.000000	100.000000

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
count	130395.00000	130432.000000	89572.000000	86102.000000	143693.000000	141851.00000
mean	1017.64994	1015.255889	4.447461	4.509930	16.990631	21.68339
std	7.10653	7.037414	2.887159	2.720357	6.488753	6.93665
min	980.50000	977.100000	0.000000	0.000000	-7.200000	-5.40000
25%	1012.90000	1010.400000	1.000000	2.000000	12.300000	16.60000
50%	1017.60000	1015.200000	5.000000	5.000000	16.700000	21.10000
75%	1022.40000	1020.000000	7.000000	7.000000	21.600000	26.40000
max	1041.00000	1039.600000	9.000000	9.000000	40.200000	46.70000

À première vue, tout paraît cohérent. Par exemple, les valeurs minimales et maximales semblent plausibles.

Toutefois, en regardant de plus près, nous décelons **quelques irrégularités.** Par exemple, pour la variable Rainfall, 75% des valeurs se trouvent en dessous de 0.8, alors que le maximum est de 371. Son écart-type (8.48) est également plusieurs fois plus grand que sa moyenne (2.36). Nous pouvons constater un phénomène similaire pour la variable Evaporation, dont 75% des valeurs se trouvent en dessous de 7.4, alors que le maximum est de 145.

Ces observations semblent **difficilement imputables aux valeurs manquantes.** À titre d'exemple, la variable Sunshine, dont le taux de NaN est de 48.01%, présente une distribution tout à fait cohérente.

Afin de mieux appréhender ces tendances, il est utile de les **visualiser** après suppression des valeurs manquantes et mise à l'échelle par standardisation :

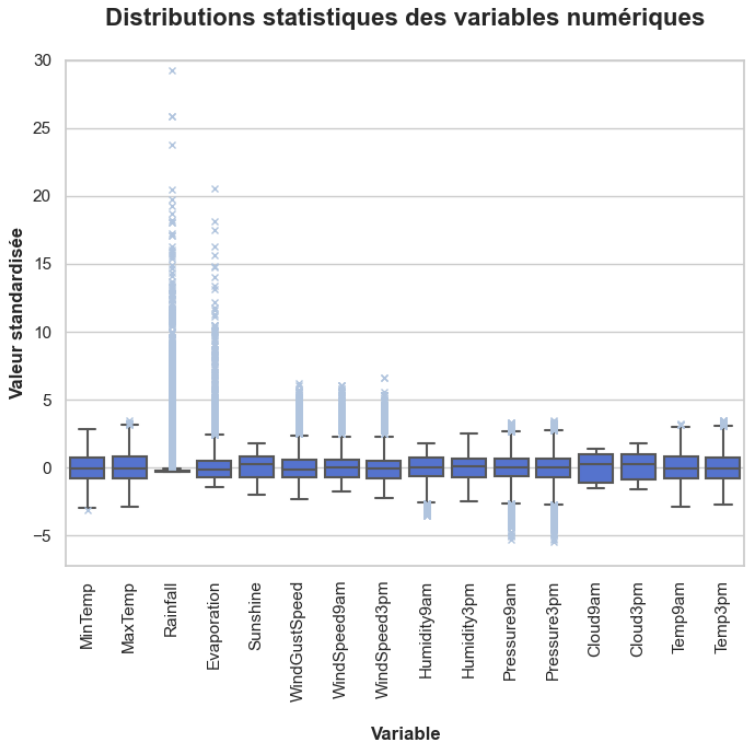


Figure 1. Les distributions statistiques des 16 variables numériques.

Figure 1 met en évidence la présence de **plusieurs distributions** aux **schémas inhabituels,** caractérisés par de **nombreuses valeurs extrêmes,** notamment dans le cas des variables suivantes :

- Rainfall
- Evaporation
- WindGustSpeed
- WindSpeed9am
- WindSpeed3pm
- Humidity9am
- Pressure9am
- Pressure3pm

Regardons une de ces variables de plus près.
Figure 2 présente la distribution de Evaporation (en valeurs absolues, et non plus standardisées).

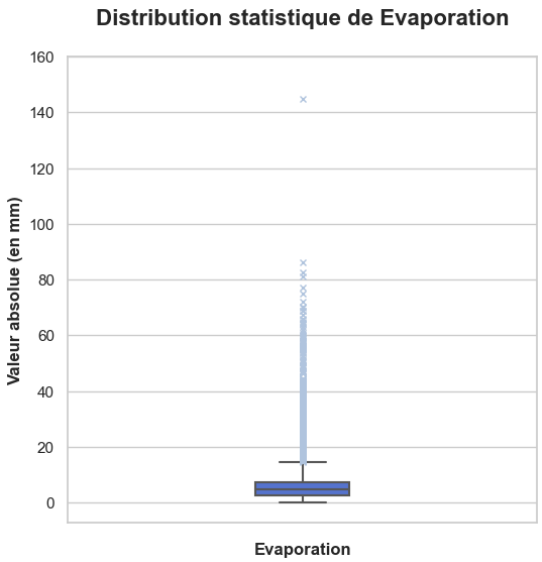


Figure 2. La distribution statistique de Evaporation.

Ce graphique rejoint les premières observations effectuées à partir des statistiques descriptives, à savoir, qu'une grande majorité des valeurs (75%) se trouvent en dessous d'un seuil très faible (7.4).

Les variables sous étude étant des **paramètres météorologiques,** nous pourrions supposer qu'elles soient **fonction de la date et du lieu** d'enregistrement des données ; une décomposition sur ces axes pourrait donc se révéler instructive.

Figure 3 présente la distribution de Evaporation selon les stations météorologiques. Elle montre que **les valeurs extrêmes**—presque toutes supérieures—**varient effectivement en fonction des stations.** Plus important, elle permet aussi de constater que **certaines stations ne disposent d'aucune entrée** pour Evaporation dans le jeu de données.

Ces informations sont importantes à prendre en compte dans la gestion des valeurs manquantes, les choix d'algorithmes et de modèles d'apprentissage automatique et l'interprétation des résultats.

Concernant les valeurs manquantes en particulier, **deux stratégies** se dessinent :

- l'imputation d'un critère statistique de position (moyenne / médiane / mode), ou
- la suppression.

Nous tenterons d'explorer ces deux pistes en parallèle dans la suite du projet.

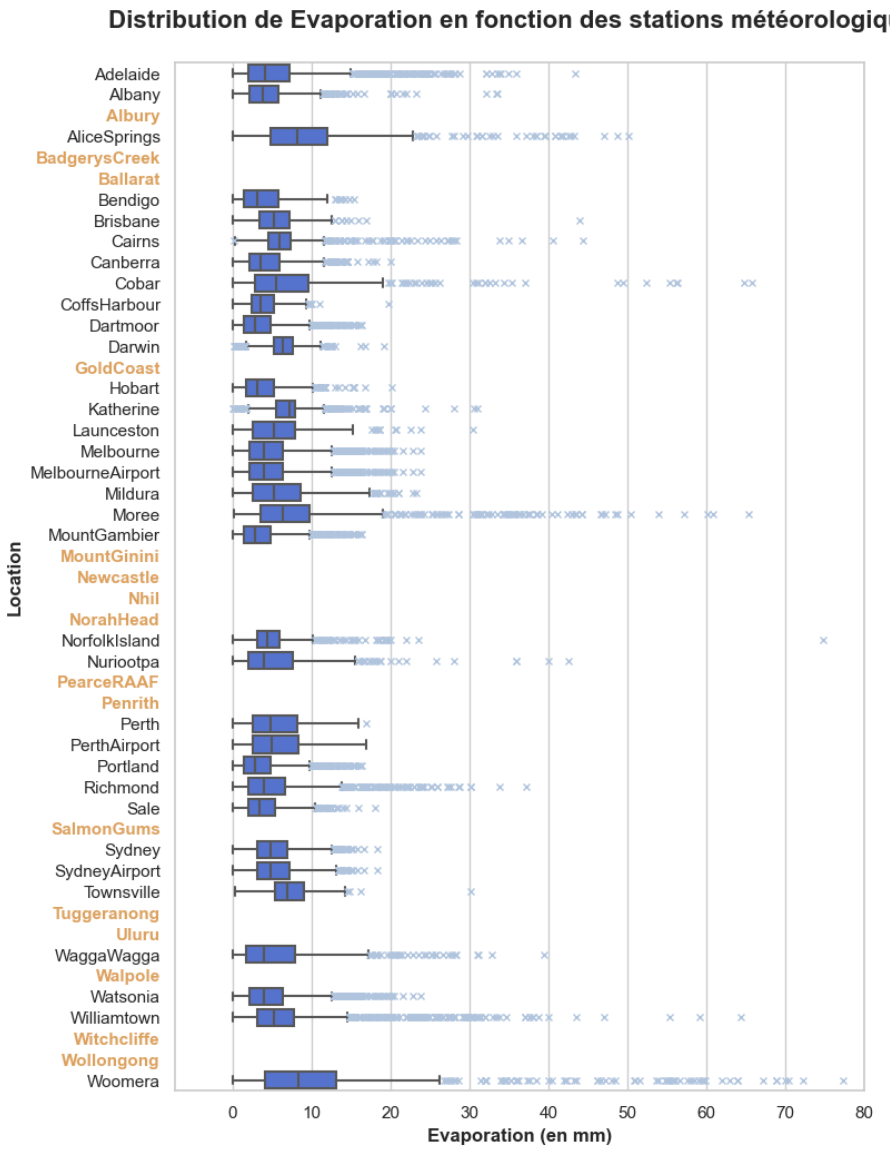


Figure 3. La distribution de Evaporation en fonction des stations météorologiques.
Les noms des stations ne disposant d'aucune donnée pour cette grandeur sont surlignés.
Afin d'améliorer la lisibilité du graphique, l'abscisse a été arrêtée à 80,
excluant 4 valeurs extrêmes supérieures.

2.3.2. Variables catégorielles

Les variables catégorielles sont les suivantes :

- WindGustDir
- WindDir9am
- WindDir3pm
- RainToday
- RainTomorrow

Il n'y a rien d'inattendu dans la distribution des 16 modalités (correspondant aux 16 points cardinaux) des 3 variables liées au vent (WindGustDir, WindDir9am, WindDir3pm).

En revanche, les 2 variables liées à la **pluie** (RainToday, RainTomorrow) présentent un **fort déséquilibre,** comme illustré dans Figure 4**.**

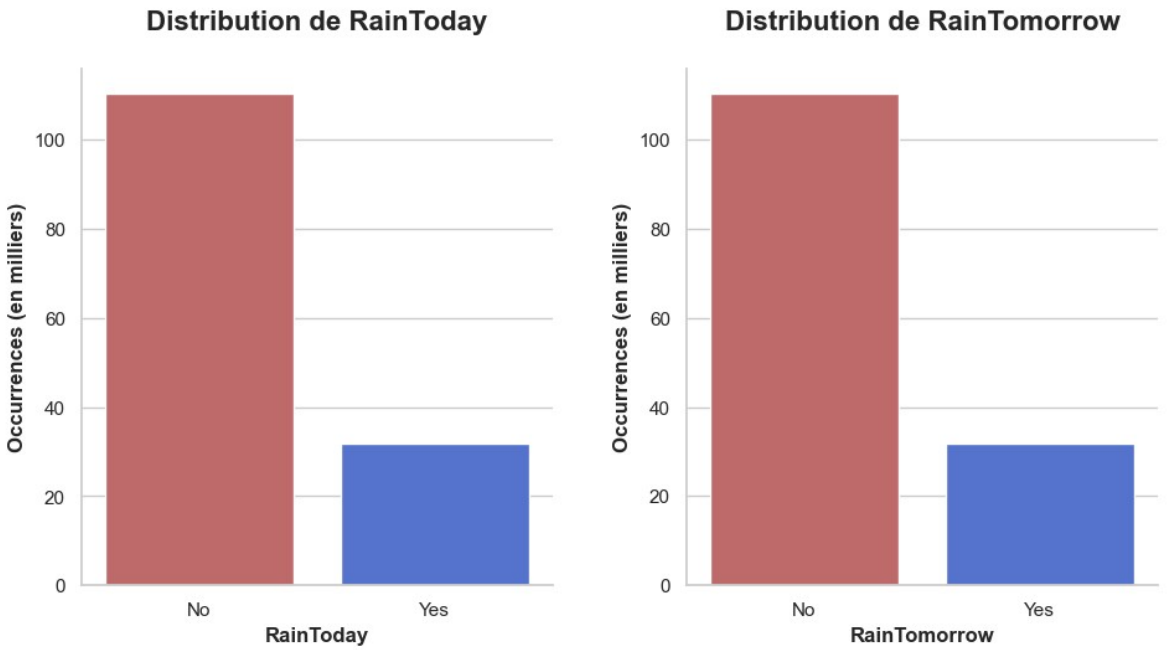


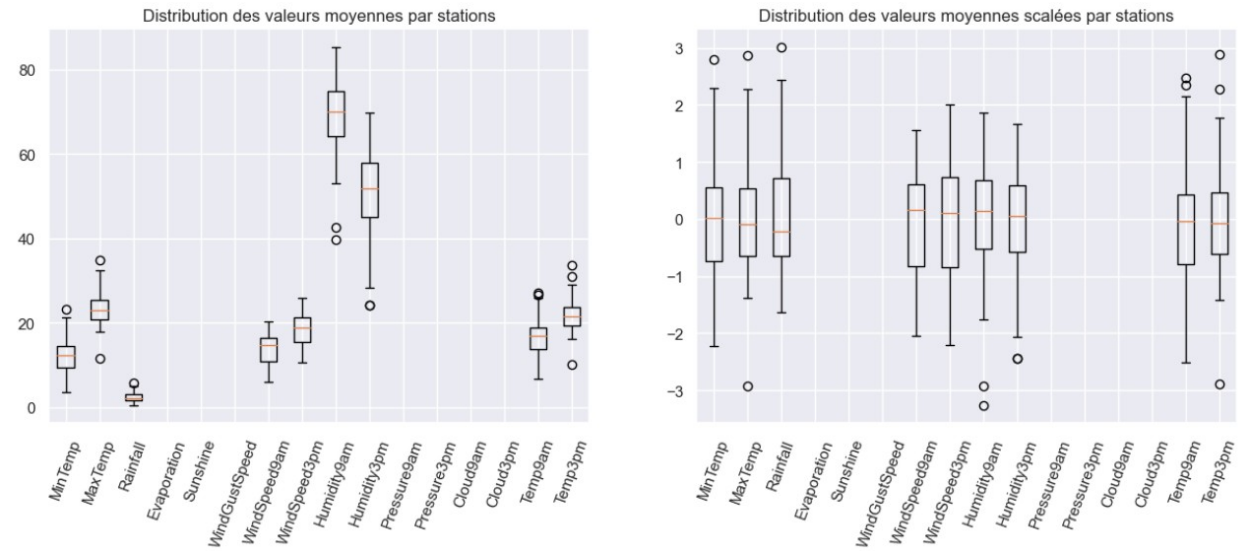
Figure 4. Distributions de RainToday et de RainTomorrow.

Comme RainToday est la variable cible principale de l'étude, il sera nécessaire de mobiliser des méthodes conçues pour le traitement des échantillons contenant une forte disparité entre les classes à prédire (par exemple : rééchantillonnage, techniques de classification avancée).

En outre, ces deux variables catégorielles sont intimement liées à la variable numérique Rainfall (dont la distribution reste à clarifier) : les 3 sont toujours présentes ou absentes ensemble.

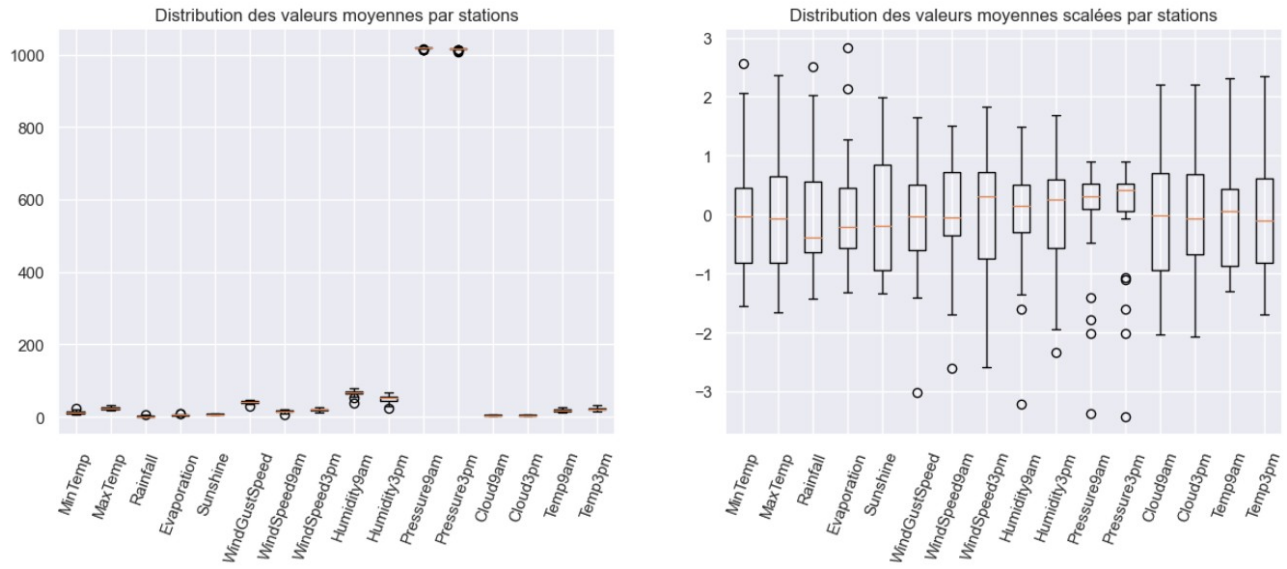
2.3.2. Répartition des valeurs moyennes par station

Pour toutes les variables numériques, nous avons regardé la répartition des valeurs moyennes par station, et nous avons fait le même exercice après standardisation (valeurs moyennes « scalées »), comme c'est illustré par la figure ci-dessous.



Nous constatons qu'il y a une grande dispersion des valeurs moyennes. Les mesures d'humidité ont des valeurs nettement supérieures aux autres variables mesurées. Comme attendu, la standardisation permet d'avoir des valeurs comparables.

Nous avons répété cet exercice sur le jeu de données après avoir enlevé toutes les lignes contenant au moins une valeur manquante. Ce qui permet d'avoir la figure ci-dessous.



Cette analyse complémentaire permet de constater que les mesures de pression sont bien supérieures aux autres mesures, s'il n'y a pas de normalisation.

2.4. Corrélations

2.4.1. Variables numériques

Figure 5 présente une carte de corrélation des variables numériques avec RainToday et RainTomorrow établie en première approche, avec :

- La binarisation de ces deux variables catégorielles sous les noms RainTodayNum et RainTomorrowNum ;
- La simple suppression des valeurs manquantes ; et
- L'utilisation du coefficient de corrélation de Pearson ρ comme métrique.

La ligne et la colonne qui correspondent à RainTomorrowNum, laquelle représente la **variable cible** RainTomorrow, ne montre pas de corrélation particulièrement importante : la plus forte est celle avec Sunshine (–0.45), suggérant la nécessité d'apporter un soin particulier à la stratégie de gestion des valeurs manquantes de cette dernière.

En considérant uniquement les variables **explicatives,** nous pouvons formuler les observations suivantes :

- Les variables liées à la température (MinTemp, MaxTemp, Temp9am, Temp3pm) semblent être relativement corrélées entre elles ($\rho > 0.7$).
- L'ensoleillement (Sunshine) et la couverture nuageuse (Cloud9am, Cloud3pm) semblent être négativement corrélés ($\rho \cong -0.7$).
- L'ensoleillement (Sunshine) et le taux d'humidité à 15 h (Humidity3pm) semblent également être négativement corrélés ($\rho \cong -0.7$).

Ces observations pourraient informer la stratégie de simplification du jeu de données à travers la gestion des valeurs manquantes (en justifiant la suppression de certaines variables si elles peuvent être « représentées » par d'autres variables avec lesquelles elles sont fortement corrélées), la réduction de dimensions et d'autres méthodes.

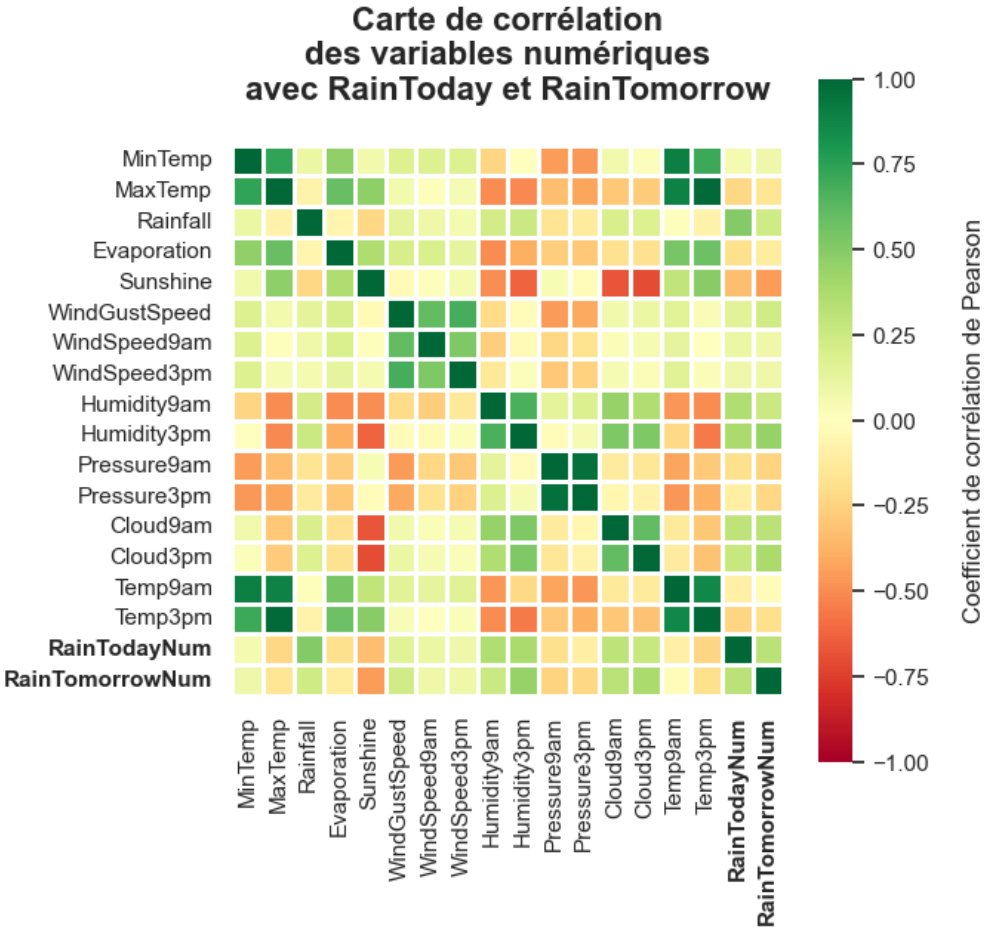


Figure 5. Carte de corrélation des variables numériques avec RainToday et de RainTomorrow.

2.4.2. Variables catégorielles

Table 2 présente les résultats du test de χ^2 d'indépendance visant à déterminer si un lien statistique existe entre chacune des variables catégorielles explicatives et la variable cible RainTomorrow, elle aussi catégorielle.

Les hypothèses formulées sont les suivantes :

$$\begin{cases} H_0 : la\ variable\ explicative\ est\ indépendante\ de\ la\ variable\ cible \\ H_1 : la\ variable\ explicative\ n'est\ pas\ indépendante\ de\ la\ variable\ cible \end{cases}$$

Le seuil de la valeur-p fixé pour le rejet de l'hypothèse nulle est de 0.05.

Comme les valeurs-p sont toutes en-dessous de ce seuil, nous rejetons l'hypothèse H_0 et concluons à l'hypothèse H_1 : les variables explicatives ne sont pas indépendantes de RainTomorrow.

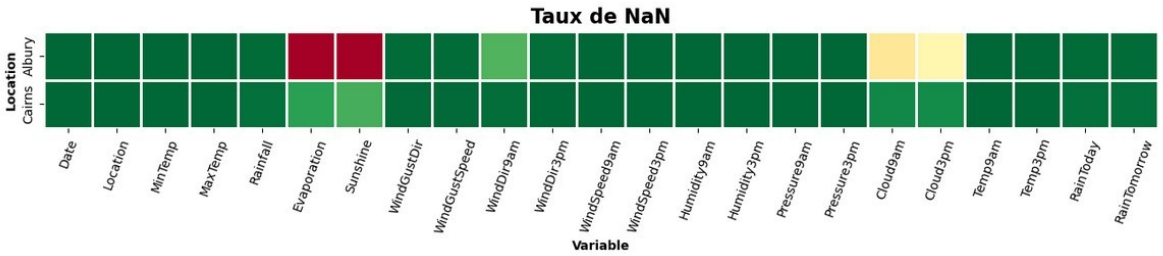
Table 2. Résultats du test de χ^2 d'indépendance entre les variables catégorielles explicatives et *RainTomorrow*.

Variable	Statistique du test χ^2	Valeur-p
Date	16735	0.0
Location	3563	0.0
WindGustDir	1517	0.0
WindDir9am	2178	0.0
WindDir3pm	1283	0.0
RainToday	13799	0.0

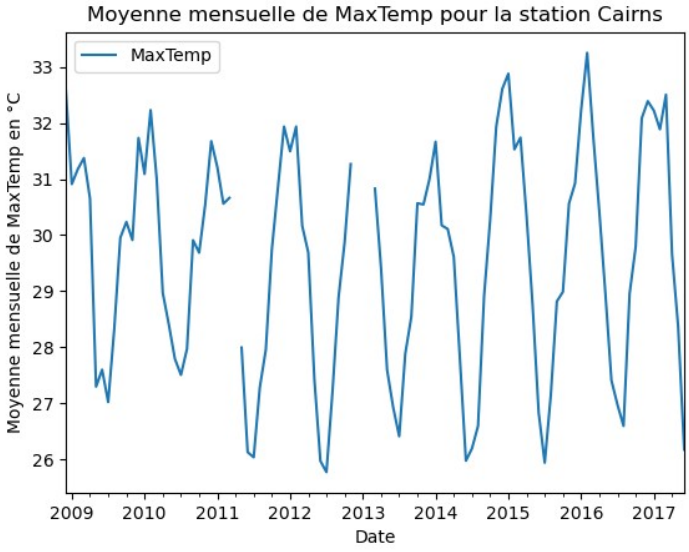
2.5. Trous chronologiques

2.5.1. Mise en évidence

Lors de l'exploration de notre jeu de données, nous avons souhaité effectuer une visualisation de l'évolution de certaines grandeurs au cours du temps pour une station. Le jeu de données original contenant encore des valeurs manquantes (NaN), il a fallu choisir une grandeur qui ne contienne aucun NaN pour une station donnée. Un rapide coup d'oeil à la « *heatmap* » des NaN nous fait porter notre choix sur la grandeur *MaxTemp* pour la station Cairns, qui a l'air très propre. Une analyse plus fine confirme ce choix : aucun NaN n'est présent dans cette colonne.



Comme il y a beaucoup de dates (3040) pour cette grandeur, nous allons faire une moyenne mensuelle pour *MaxTemp*, et afficher cette moyenne mensuelle au cours du temps.



Nous voyons apparaitre des « trous » sur ce graphique, autour du premier trimestre 2011, et à la frontière entre 2012 et 2013. Ceci est très surprenant : *MaxTemp* ne contient pas de NaN pour la station Cairns. Vérifions dans le tableau des valeurs moyennées.

2011-03-31	30.664516
2011-04-30	NaN
2011-05-31	27.996774

2012-12-31	NaN
2013-01-31	32.677419
2013-02-28	NaN

Il y a 3 intervalles de NaN en avril 2011, en décembre 2012 et en février 2013 qui ont été créés lors du processus de moyenne mensuelle. Cela signifie que pour ces dates, il n'y a aucune entrée dans le dataset original. C'est préoccupant, et cela nécessite des investigations supplémentaires autour de ce que nous appellerons désormais des trous chronologiques. Combien y en a-t-il ? Comment sont-ils distribués ?

2.5.2. Analyse des trous chronologiques :

Un décompte pour la station Cairns effectué sur la base des dates enregistrées donne le résultat suivant :

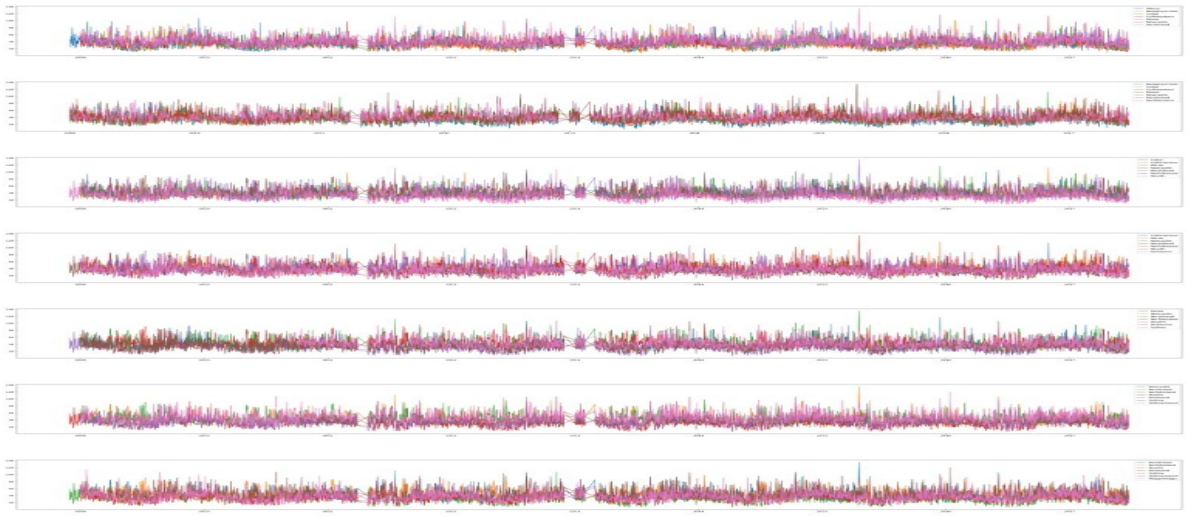
```
Pour la station Cairns:
Première date enregistrée: 2008-12-01
Dernière date enregistrée: 2017-06-25
Nombre de dates mesurées: 3040
Nombre total de jours entre le début et la fin: 3128
Il manque donc 88 jours pour cette station.
```

Ce résultat est indépendant de la nature de la grandeur mesurée puisqu'il se base uniquement sur les dates du tableau. Comme la série des dates ne contient pas de NaN, on en déduit que certaines dates sont totalement absentes du tableau. Ainsi, cela signifie que la station Cairns n'a pas enregistré de mesures pendant 88 jours. Qu'en est-il des autres stations ?

	endroit	date_debut	date_fin	n_jours_tot	n_jours_mes	n_jours_miss
0	Albury	2008-12-01	2017-06-25	3128 days	3040	88
1	BadgerysCreek	2009-01-01	2017-06-25	3097 days	3009	88
2	Cobar	2009-01-01	2017-06-25	3097 days	3009	88
3	CoffsHarbour	2009-01-01	2017-06-25	3097 days	3009	88
4	Moree	2009-01-01	2017-06-25	3097 days	3009	88
5	Newcastle	2008-12-01	2017-06-24	3127 days	3039	88
6	NorahHead	2009-01-01	2017-06-25	3097 days	3004	93
7	NorfolkIsland	2009-01-01	2017-06-25	3097 days	3009	88
8	Penrith	2008-12-01	2017-06-25	3128 days	3039	89

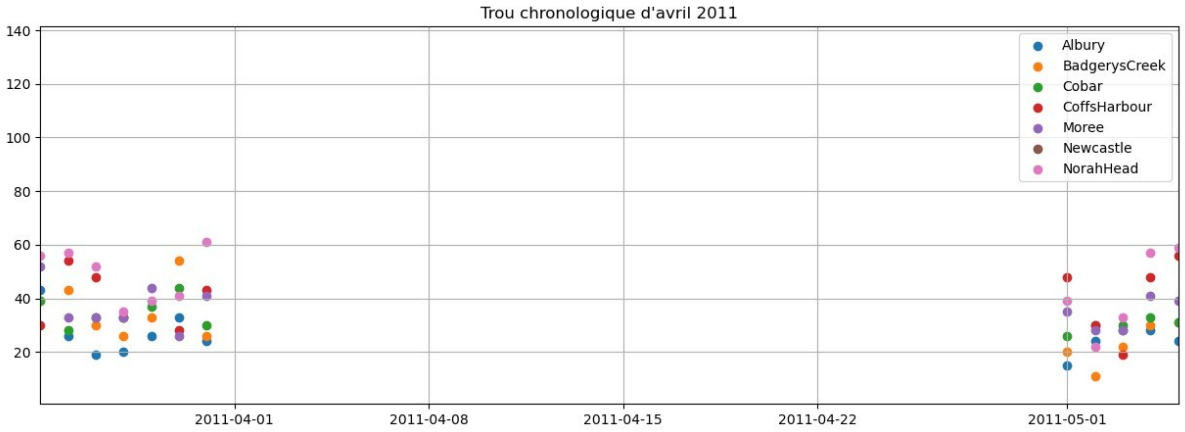
Etonnamment, on retrouve exactement le même nombre de jours manquants, à quelques exceptions près. Il y en a en général 88. Mais où sont-ils distribués ?

On trace l'évolution d'une grandeur (ici *WindGustSpeed*) au cours du temps pour les 49 stations, en superposant 7 courbes sur 7 graphiques différents.



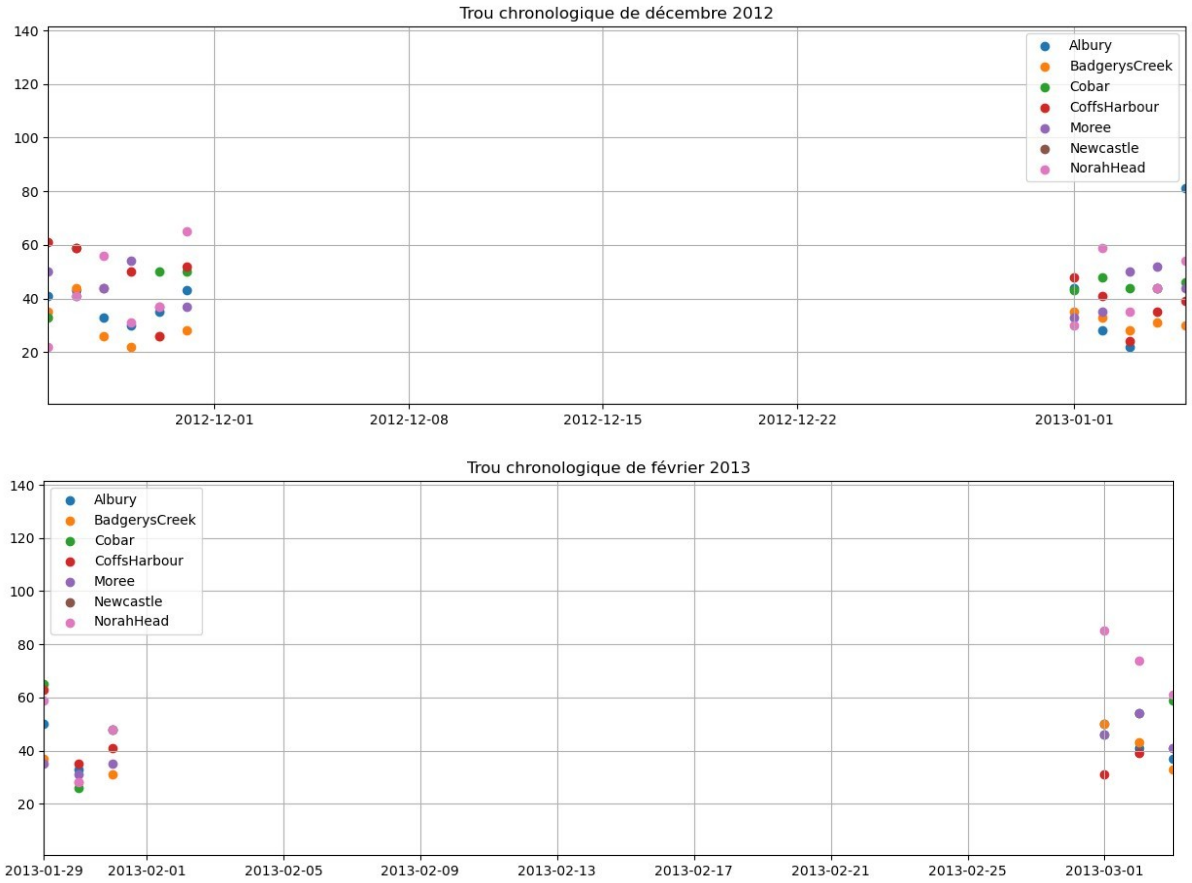
On voit clairement 3 zones de trous, qui ont l'air de se superposer : les dates manquantes ne sont pas distribuées aléatoirement selon les stations, elles se concentrent toutes aux mêmes endroits. Nous cherchons à déterminer les dates de ces trous.

Premier trou : avril 2011



C'est très clair : le début et la fin des trous sont situés exactement aux mêmes dates. Les mesures s'arrêtent au premier avril 2011 pour reprendre le premier mai 2011. Au final, c'est l'intégralité du mois d'avril 2011 qui est absent du tableau.

Autres trous :



On retrouve le même schéma : ce sont des mois pleins et entiers qui sont absents, ici décembre 2012 et février 2013.

2.5.3. Conclusion

La parfaite synchronicité des trous laisse à penser que des périodes d'interruptions de mesure ont été décidées pour l'ensemble des stations, en avril 2011, décembre 2012 et février 2013. La raison de ces interruptions est inconnue. Il est intéressant d'avoir ces informations dans le cas d'une éventuelle modélisation de séries temporelles en vue d'effectuer des projections. Faudra-t-il trouver un moyen de gérer ces valeurs manquantes ? En tout cas, nous en connaissons le nombre et la distribution.

Une question demeure : avril compte 30 jours, décembre 31, et février 2013, 28, ce qui fait un total de 89 jours manquants, et pas 88. Est-ce un problème dans le code comptant les jours manquants ? Certaines stations affichent un nombre de jours manquants égal à -1, ce qui est surprenant : on devrait avoir zéro. En tenant compte d'un éventuel problème de décalage, on retrouverait bien 89 jours.

Ces stations sont au nombre de 3 : Nhill, Katherine et Uluru, et leur mise en service à toutes les trois date précisément du 1er mars 2013, soit après la dernière période d'interruption enregistrée. Était-ce là la raison de l'interruption générale de toutes les autres stations ?

Les données enregistrées démarrent à des dates différentes, mais se terminent toutes à la même date, ou presque : soit le 24 juin 2017, soit le 25 juin 2017.

3. Nettoyage des données

Il est temps de se préoccuper de la gestion des NaN, pour ensuite préparer les 9 tableaux déjà crées à la modélisation via les procédés de pre-processing et de feature engineering.

3.1. Quels choix effectuer ?

3.1.1. Variables numériques (quantitatives)

Nous avons un total de 16 variables numériques à gérer :

- Un bloc de température : *MinTemp*, *MaxTemp*, *Temp9am*, *Temp3pm*
- Un bloc de vitesses de vent : *WindGustSpeed*, *WindSpeed9am*, *WindSpeed3pm*
- Un bloc Humidité: *Humidity9am*, *Humidity3pm*
- Un bloc Pression : *Pression* 9am, *Pression3pm*
- Un bloc couverture nuageuse : *Cloud9am*, *Cloud3pm*.
- Des mesures plus hétéroclites : *Rainfall*, *Evaporation*, *Sunshine*.

Toutes ces variables présentent une statistique cohérente : peu d'outliers, des valeurs maximales et minimales qui possèdent un sens physique, des distributions plausibles, nous l'avons vu précédemment. Seule la variable Rainfall pose problème à cause de sa distribution très hétéroclite, nous verrons plus tard que le problème se règlera de lui-même lorsque nous évoquerons le traitement des variables catégorielles *RainToday* et *RainTomorrow*.

Nous voudrons travailler sur **deux modalités différentes de gestion des NaN** pour ces grandeurs :

- Soit le **remplacement des NaN par la moyenne par station**, qui sera le meilleur estimateur des valeurs manquantes, au sens mathématique et statistique du terme, et permettra de conserver des données.
- Soit leur **suppression pure et simple**, si notre jeu de données est suffisant. Cette suppression s'effectuera à l'échelle des 9 tableaux correspondant au regroupement par différentes modalités de mesure. Alors qu'une suppression massive et aveugle sur le tableau d'origine aurait conduit mécaniquement à la disparition pure et simple de toutes les stations qui ne mesurent pas au moins une grandeur, cette méthode est plus sélective et permet néanmoins de conserver de l'information.

A ce stade de notre réflexion, les informations dont nous disposons ne nous permettent pas de trancher nettement en faveur de l'une ou l'autre. Nous prévoyons de tester les deux lors de la phase de modélisation. Voici un résumé des avantages et inconvénients de chaque méthode :

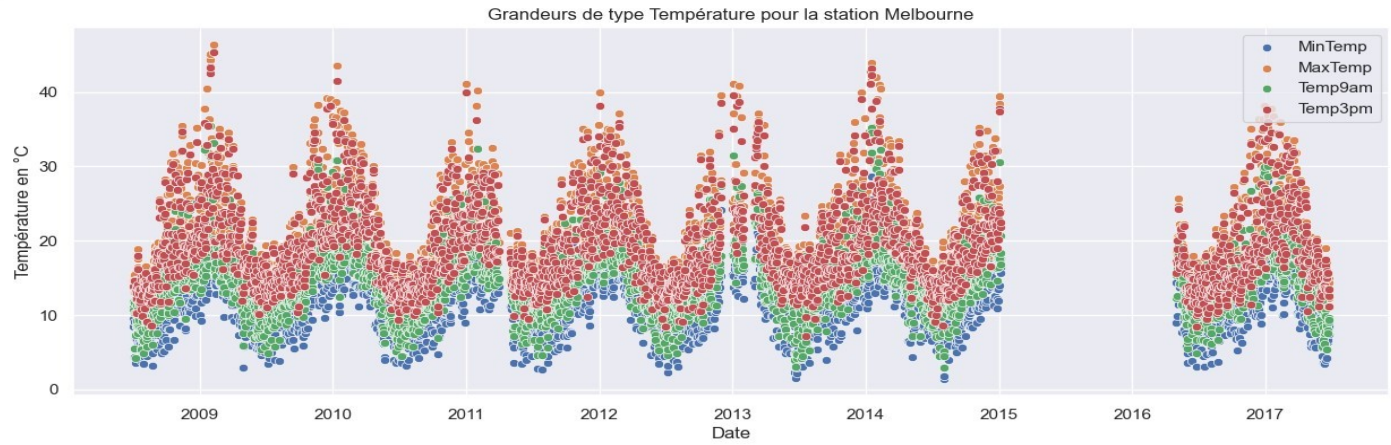
- **Méthode n°1** : Replacement par la moyenne par station. Cette méthode conserve des données en remplaçant les NaN par des valeurs robustes et cohérentes, mais peut créer des biais.
- **Méthode n°2** : Suppression des NaN . Les données finales sont de qualités, mais peuvent être en nombre insuffisant.

Détaillons maintenant notre analyse sur la méthode n°1.

Comme nous l'avons vu, il existe **deux grandes catégories pour la distribution des NaN** :

- Une **aléatoire**, où certaines valeurs enregistrées sont absentes ponctuellement ici où là, pour des raisons inconnues ;
- Une **systématique**, lorsque la station cesse de mesurer une grandeur pendant une période plus ou moins longue.

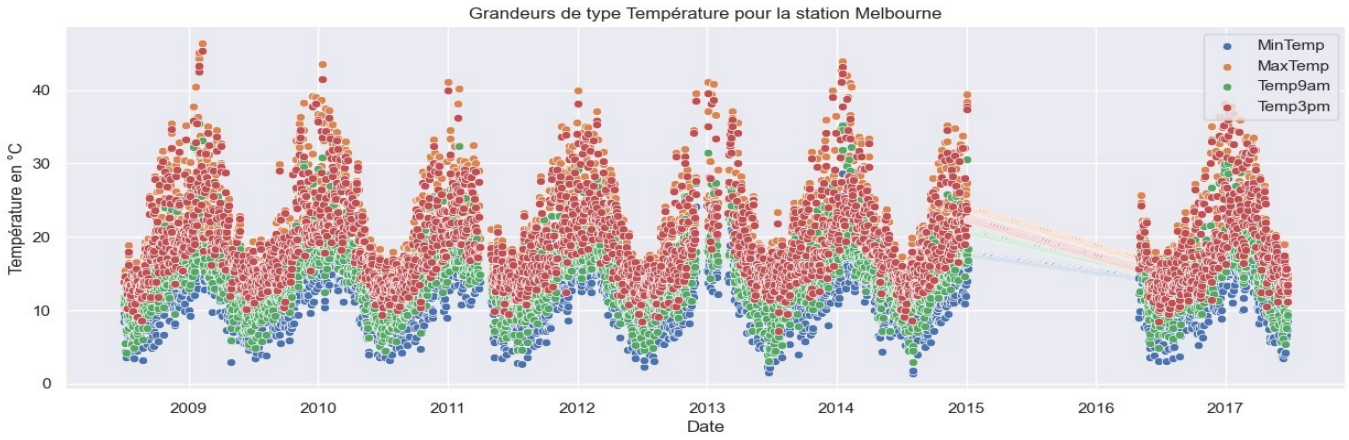
Exemple sur le bloc température mesurées à Melbourne :



Nous pouvons repérer trois choses :

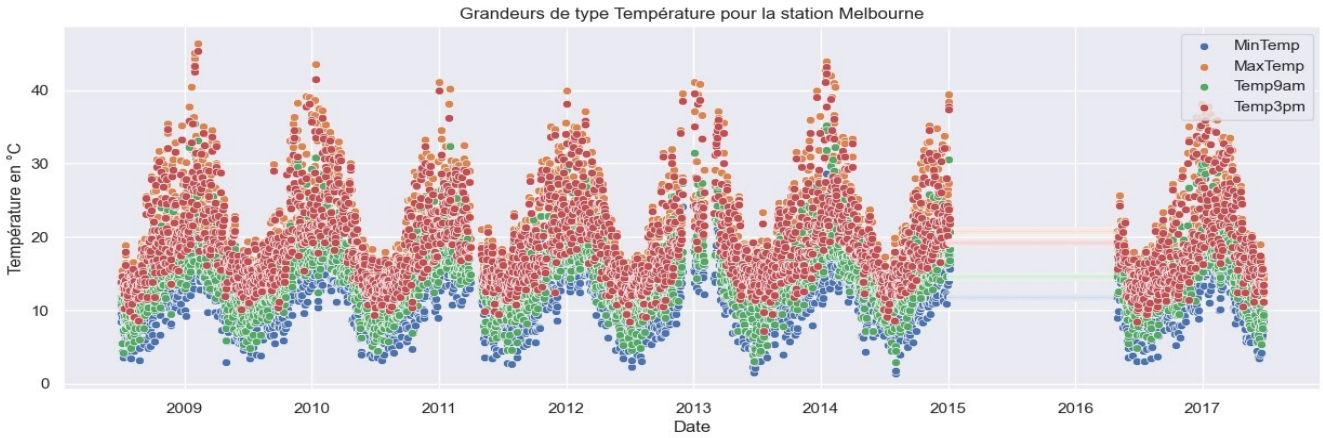
- Les trois trous chronologiques apparaissent sous forme de fines bandes. Ce ne sont pas des NaN.
- Une large bande de NaN pendant un an et demi environ, de janvier 2015 à avril 2016. L'appareil unique de mesure des températures servant à alimenter en données nos 4 variable du bloc température était probablement défectueux ou désactivé pendant la période.
- Les NaN ponctuels ne sont pas visibles sur cette représentation.

Nous avons tout d'abord pensé à remplacer les NaN par interpolation, au vu de la régularité des mesure, notamment leur périodicité saisonnière. Cela donnerait :



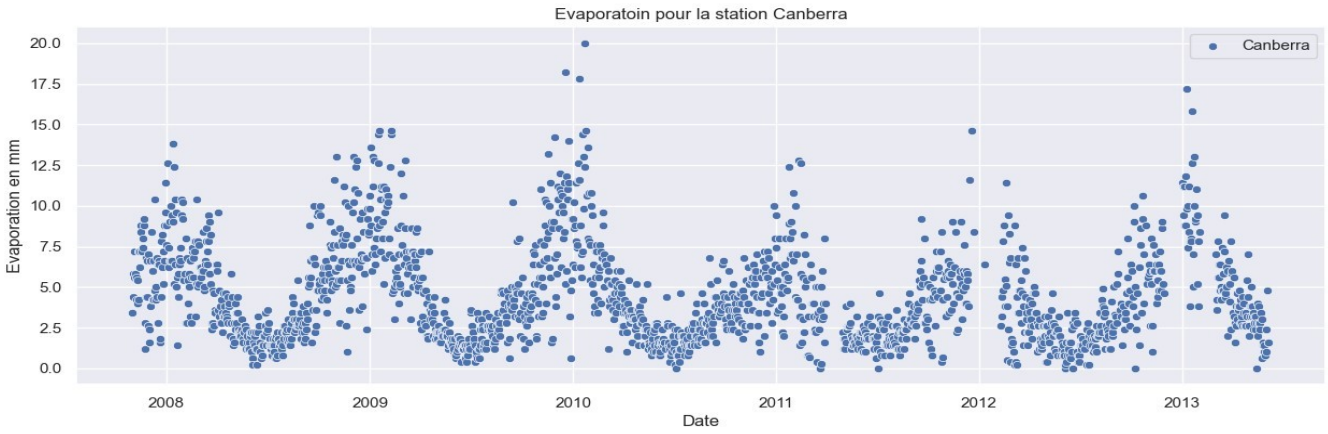
Comme prévu, les trous chronologiques ne sont pas concernés, car ce ne sont pas des NaN. En revanche, la méthode *interpolate* se contente, sans plus d'arguments, de relier la première date non mesurée à la dernière en utilisant un modèle linéaire. On voit bien sur le graphique que cette méthode n'est pas satisfaisante : l'interpolation n'est pas fidèle à ce que l'on aurait pu observer. Par manque de temps et de de connaissances, nous éliminons cette méthode au profit du remplacement par la moyenne.

Voici le résultat, toujours pour Melbourne :

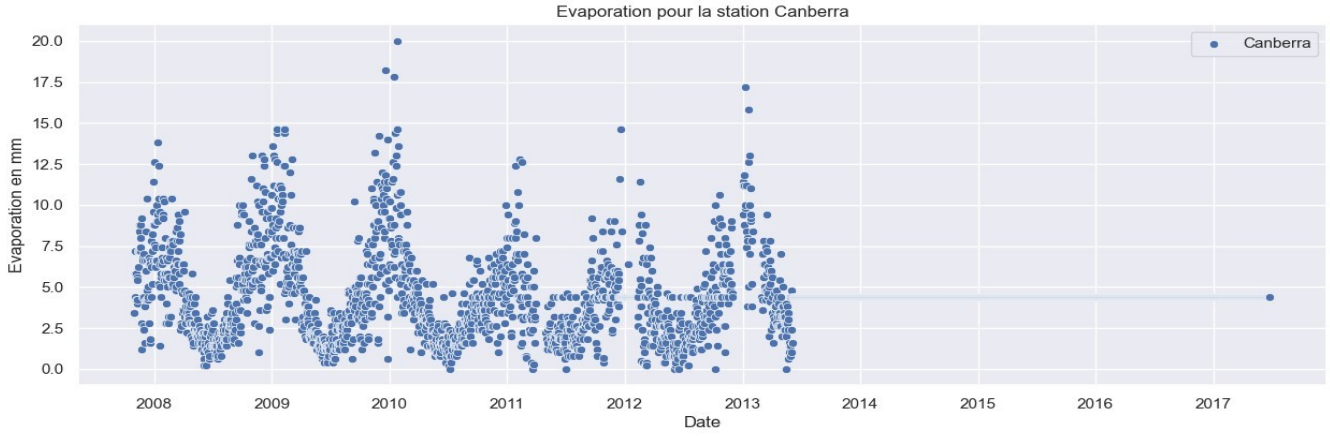


Comme prévu, ce n'est pas l'idéal, mais cela permet de conserver toutes les autres données des lignes contenant des NaN dans les colonnes température pour cette station.

Regardons ce que ce choix implique pour une autre station et pour une autre grandeur. Nous choisissons Canberra, et l'évaporation, pour laquelle le taux de NaN est important (environ 47%).



Ici les NaN aléatoires sont davantage visibles. Remplaçons-les par la moyenne :



Surprise ! Nous n'avions pas pris en compte la nécessité d'intégrer à l'affichage des dates fixes de début de de fin. On voit ici le résultat du remplacement. Canberra a cessé de mesurer l'évaporation à partir de mi 2013. On voit aussi

« apparaitre » les NaN aléatoires entre 2011 et 2013 (N.B. : un test d'interpolation illustre graphiquement le fait que l'interpolation semble mieux fonctionner pour les NaN aléatoires que le remplacement par la moyenne, comme on le voit ici).

On comprend bien le problème désormais :

- La méthode n°1 peut conduire à de vastes remplacements comme ici dont on peut se poser la question de la pertinence.
- Le choix de la méthode pourrait donc dépendre du taux de NaN par mesure :
  - Taux faible : on remplace par la moyenne ou on supprime, les deux ayant leurs avantages et inconvénients déjà mentionnés
  - Taux élevés : on supprime les données pour éviter de trop biaiser. Le problème, c'est que les fenêtres de non-mesures de grandeurs ne sont pas simultanées et on pourrait potentiellement perdre beaucoup de données à fonctionner ainsi.

Conclusion :

Par manque de temps et d'expérience, nous ne pouvons pas pour l'instant nous lancer dans des modes de gestions plus sophistiqués. Nous devons avancer et faire des choix, quitte à y revenir plus tard après une itération par la phase modélisation. Nous retenons donc les deux méthodes 1 et 2, à savoir :

- Méthode n°1 : Replacement par la moyenne par station. Cette méthode conserve des données en remplaçant les NaN par des valeurs robustes et cohérentes, mais peut créer des biais.
- Méthode n°2 : Suppression des NaN. Les données finales sont de qualité, mais peuvent être en nombre insuffisant.

Remarque : les mesures de couvertures nuageuse sont en réalité des catégories numériques de 0 à 9. Nous remplaçons les NaN de cette variable par la moyenne par station arrondie à l'entier le plus proche.

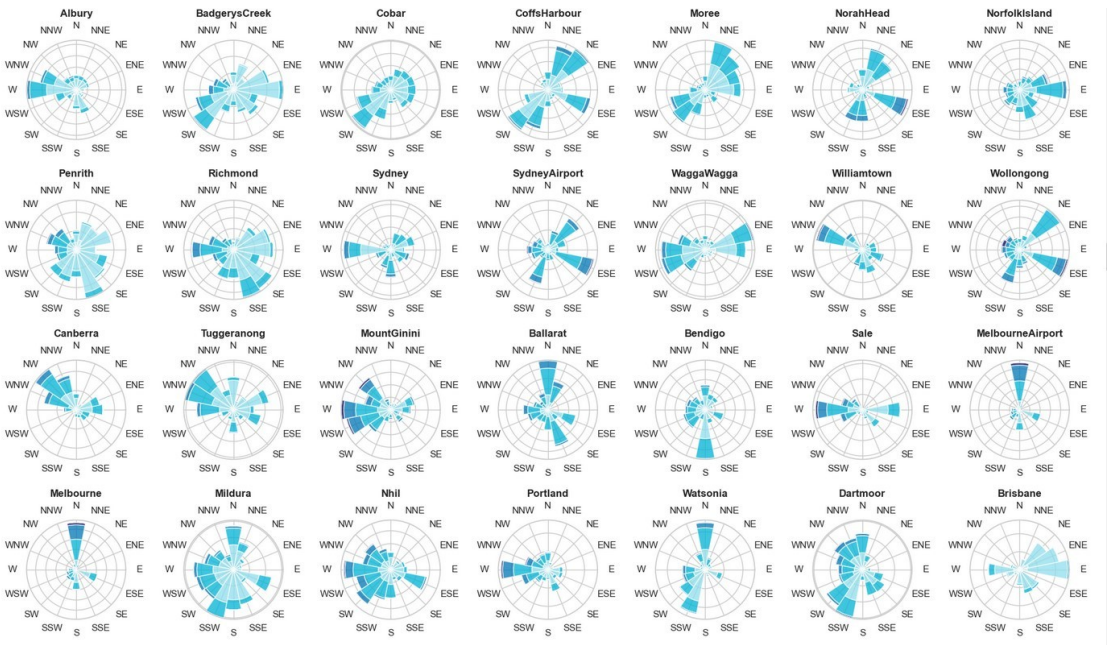
3.1.2. Variables catégorielles (qualitatives)

Nous avons un total de 7 variables catégorielles à gérer :

- Date et Location, qui ne contiennent aucun NaN : aucune décision particulière n'est à prendre.
- Un bloc concernant la direction vent : *WindGustDir*, *WindDir9am*, *WindDir3pm*
- Deux variables binaires concernant la pluie : *RainToday* et *RainTommorow*.

Pour le bloc des directions des vents :

L'analyse concernant les variables numériques précédentes reste valable : on doit choisir entre la suppression pure et simple ou bien le remplacement par la modalité la plus fréquente, avec toujours les mêmes enjeux.
Nous avons tracé pour chaque station la rose des vents, qui est un histogramme en diagramme polaire, dont voici un extrait :



Dans certains cas, la méthode 1 a du sens : la modalité la plus fréquente correspond bien à l'idée que l'on se fait d'une moyenne dans le cas spécifique de la direction des vents (exemples : Albury, Melbourne, Cobar, Richmond). Dans d'autres, on aboutira à des résultats plus problématiques (exemples : CoffsHarbour, WaggaWagga).

Une autre information supplémentaire aidant à prendre une décision pour ce bloc se trouve dans l'analyse des corrélations avec la variable cible *RainTomorrow*. Les directions du vent semblent peu corrélées à la cible, ainsi, il peut être judicieux de remplacer les NaN par la modalité la plus fréquente - quitte à mettre des valeurs problématiques - afin de préserver les autres valeurs (qui corrèlent davantage que le vent) des conséquences de la suppression de la ligne entière. Nous pourrons éventuellement finir par supprimer ce bloc de vents de nos tableaux nettoyés (nous reviendrons sur le sujet spécifique du bloc des vents dans la partie pre-processing).

Pour le bloc Pluie :

Ici, c'est plus facile : l'une des variable est notre variable cible. On supprime les NaN correspondants, on ne peut pas se permettre de biaiser notre modèle à ce niveau là. D'autant que le taux de NaN pour *RainTomorrow* et relativement faible, on ne perd ainsi pas beaucoup de données.

On s'aperçoit que ce choix impacte directement les colonnes *RainToday* et *Rainfall*. En effet, *RainTomorrow* contient la valeur stockée dans *RainToday* du lendemain. Si l'instrument de mesure de la quantité de pluie tombée ne fonctionne pas, il n'y aura rien dans *Rainfall*. Et rien non plus dans *RainToday*, qui se remplit automatiquement (selon la règle : *Yes* si les précipitations dépassent 1 mm sur la journée, *No* dans le cas contraire), et donc rien non plus dans le *RainTomorrow* correspondant. Ainsi, en cas de période prolongée de non mesure de la pluie, supprimer les NaN dans *RainTomorrow* règle le problème des NaN dans les autres colonnes du bloc.

Parfois, cela n'est pas suffisant, il reste en effet le cas des NaN aléatoires et des effets de bords des NaN systématiques. Pour bien comprendre, imaginons n'avoir qu'une seule ligne avec un NaN pour *RainTomorrow*. On la supprime. Mais s'il y avait un NaN ici, c'est parce qu'il y avait deux NaN dans la ligne précédente (la veille, sur *Rainfall* et *RainToday*), et eux n'ont pas été supprimés. On règle ce problème en effectuant une suppression sur *RainToday* aussi.

A l'issue du processus, il n'y a plus de NaN dans la colonne *Rainfall*, ce qui règle le problème de sa distribution compliquée dont nous parlions plus tôt.

Conclusion :

Nos tableaux sont prêts, nettoyés selon la méthode suivante :

- Il n'y a rien à faire pour *Date* et *Location* ;
- Toutes les variables numériques sauf *Rainfall* sont complétées par la moyenne par station ;
- Les variables catégorielles de la direction du vent sont remplaçées par la modalité la plus fréquente ;
- On supprime les NaN sur *RainTomorrow* et *RainToday*
- Cela règle le problème de la gestion de la colonne *Rainfall*
- On garde la possibilité de travailler en amont sur chacun des 9 tableaux pour supprimer directement tous les NaN, ce qui est radical, mais moins que de le faire sur le tableau original, ce qui nous priverait des données de stations entières.

A titre d'exemple, sur le tableau regroupant les stations qui mesures toutes les grandeurs, l'application des méthodes donne :

- *Pour la méthode n°1 : on passe de 80 040 lignes à 77 728, soit une perte de 2,9% des données.*
- *En appliquant la méthode n°2, on conserve 56 420 lignes, soit une perte de 29 % des données.*
- *En faisant un* dropna *en amont sur le tableau originel, on passe de 145 460 lignes à 56 420, soit une perte de 61% des données (il est logique d'avoir le même nombre de lignes restantes entre les points 2 et 3 car un* dropna *sur tout le tableau élimine automatiquement toutes les stations qui ne mesurent pas l'intégralité des grandeurs).*

3.2. Pre-processing

3.2.1. Variables quantitatives

Comme vu précédemment dans la section décrivant les statistiques de ces valeurs, il sera nécessaire de procéder à un standardisation des données quantitatives, les valeurs numériques s'échelonnant sur des intervalles très différents. La pression, notamment, aboutit à des valeurs très élevées (autour de 1000), tandis que la vitesse du vent se situe typiquement autour de la dizaine ou centaine de km/h, ou encore les précipitations qui peuvent être faibles (0.5 mm).

3.2.2. Variables qualitatives

Afin de pouvoir utiliser les variables catégorielles, il va falloir les gérer :

- **Pour le bloc pluie :** ***RainToday, RainTomorrow*****.** Ce sont des variables binaires, un simple remplacement par 0 ou 1 convient.
- **Pour le bloc des vents** (3 colonnes) : chaque grandeur peut prendre 16 modalités différentes, correspondant chacune à une direction de la rose des vents. Une dichotomisation créerait donc 16 x 3 = 48 colonnes supplémentaires. A ce stade, c'est la seule méthode que nous connaissons, et nous nous interrogeons sur le poids que cette surcharge de colonne entrainerait sur la recherche de modèles. D'autant plus que les analyses préliminaires montrent que ces variables semblent peu corrélées à la variable cible. Ainsi se posent deux questions :
  - La suppression pure et simple des colonnes
  - La recherche d'une autre méthode de gestion. Un *flow chart* sur un article traitant de cette question nous propose 3 méthodes : *One Hot Encoding*, *Binary Encoding* et *Feature Hashing Encoding*. Mais à ce stade, nous n'avons pas eu le temps de creuser les investigations.

- **Pour la colonne des dates :** voir en feature enginerring (création de colonnes dates)
- **Pour la colonne *Location***, qui contient la ville où est implantée la station, plusieurs options s'offrent à nous :
  - Entrainer un modèle par station, ce qui évacue la question du pre-processing
  - Il y a 49 stations différentes, la dichotomisation via *get_dummies* créera 49 colonnes supplémentaires. On rencontre le même problème que pour la colonne des vents. Gestion par *One Hot Encoding*, *Binary Encoding* ou *Feature Hashing Encoding* ?
  - Ou Supprimer la colonne. On perd la dépendance avec la géographie, mais on garde quand même l'information pertinente sur les grandeurs physiques.

3.3. Feature engineering

En l'état actuel de nos connaissances, nous n'envisageons que la création de 3 colonnes supplémentaires : l'année, le mois et le jour. D'autres investigations sont à réaliser pour savoir si une métrique spécifique à la météorologie pourrait être créée.