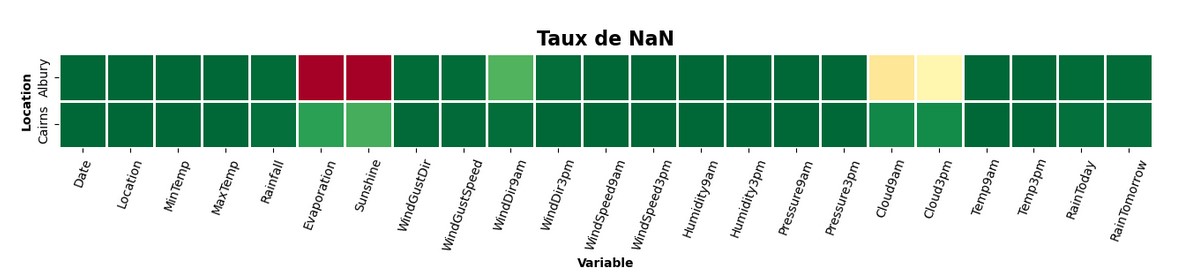
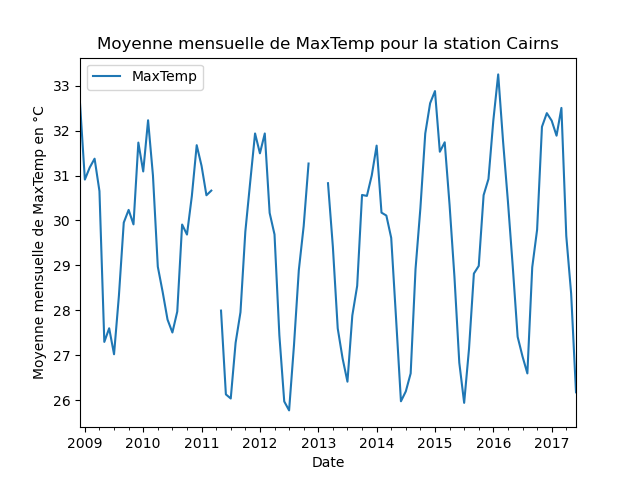
## 1.5 Trous chronologiques

1.5.1 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 dataset original contenant encore des 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.

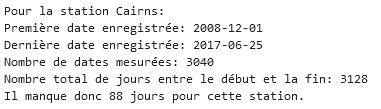
Comme il a beaucoup de dates (3040) pour cette grandeur, nous allons faire une moyenne mensuelle pour MaxTemp, et afficher cette moyenne mensuelle au cours du temps.

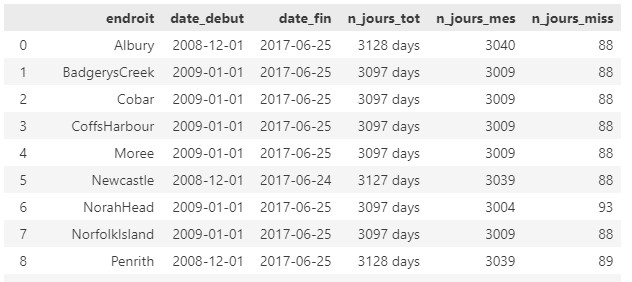
Nous voyons apparaitre des « trous » sur ce graphique, autour du premier trimestre 2011, et à la frontière entre 2012 et 2013. Ceci est très surprenant : MaxTemp ne contient pas de nan pour la station Cairns.



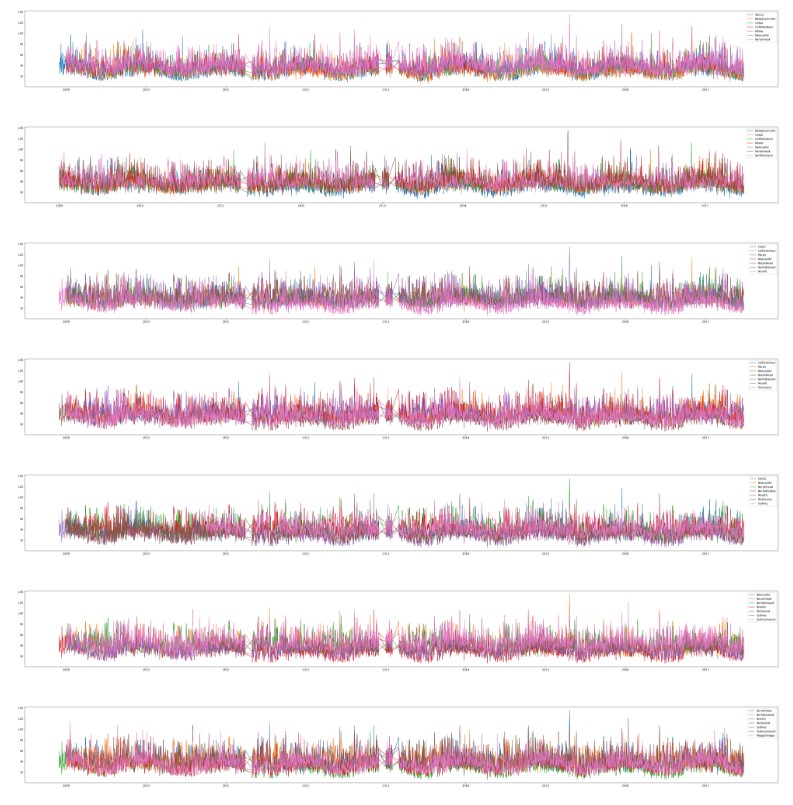
Vérifions dans le tableau des valeurs moyennées. Il y a 3 nan en avril 2011, en décembre 2012 et en février 2013 qui ont été créés lors du processus de moyenne mensuelle. Cela signifie que poru ces dates, i n’y a aucune entrée dans le dataset original. C’est préoccupant, et cela nécessite des investigations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

1.5.2 Analyse des trous chronologiques :

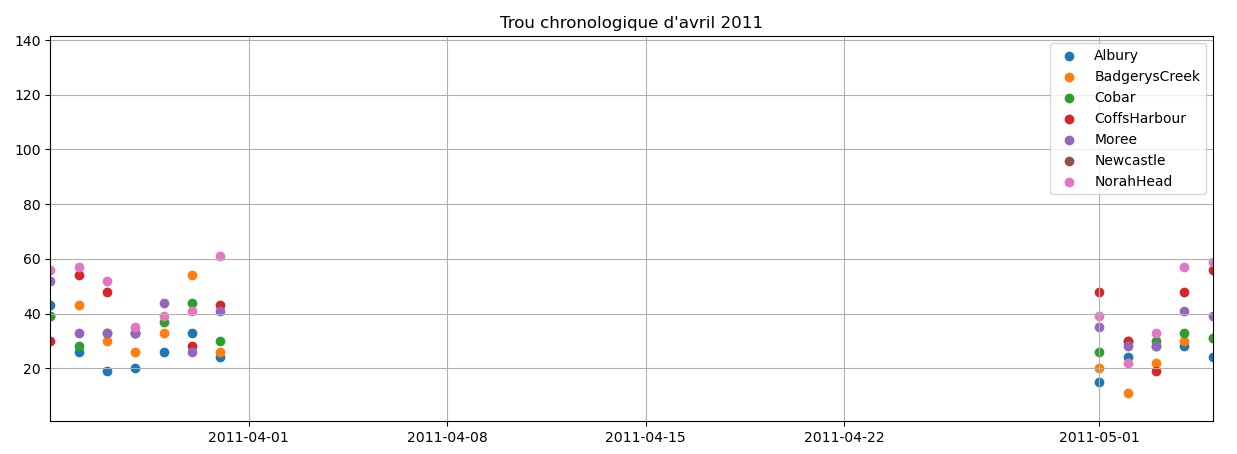
Un décompte pour la station Cairns effectué sur la base des dates enregistrées donne le résultat suivant :

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 ?

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 ?

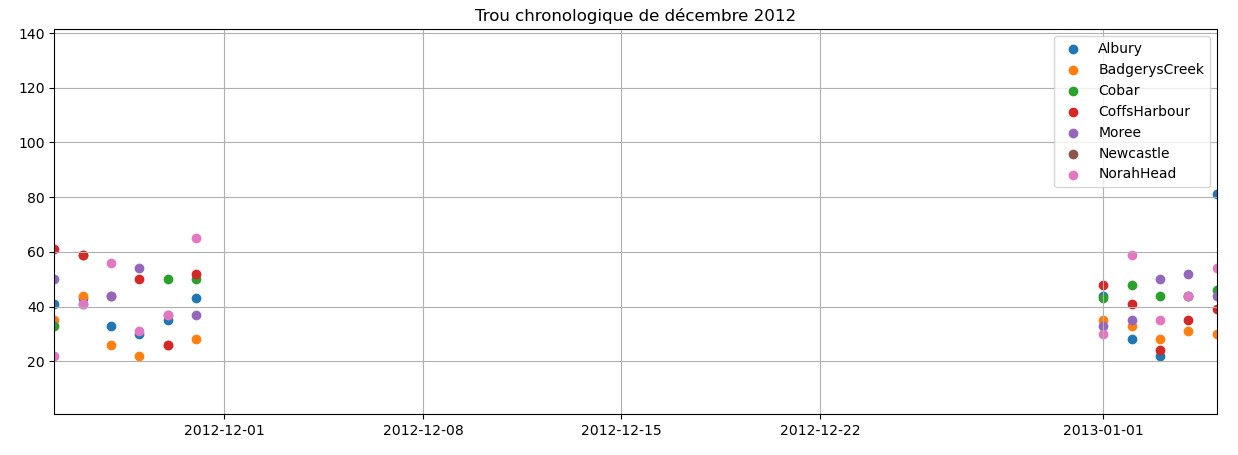
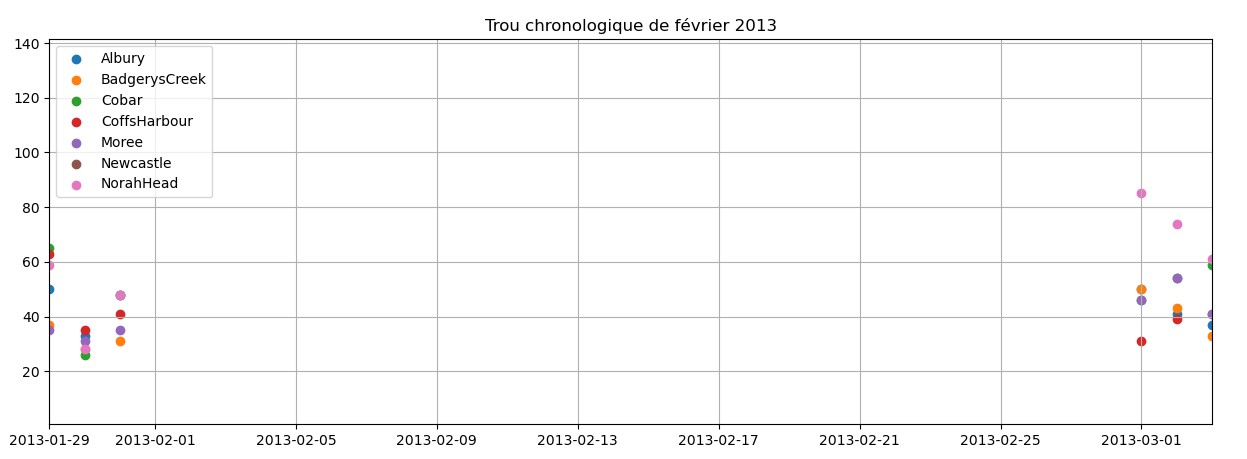
On trace l’évolution d’une grandeur (ici WindGustSpeed) au cours du temps pour les 49 stations, en superposant 7 courbes sur 7 graphiques différents.

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.

1.5.3 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.

# 2. 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.

## 2.1 Quels choix effectuer ?

2.1.1 Variables numériques ou 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 ploc 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, valeurs maximales et minimales qui possèdent un sens physique, 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.

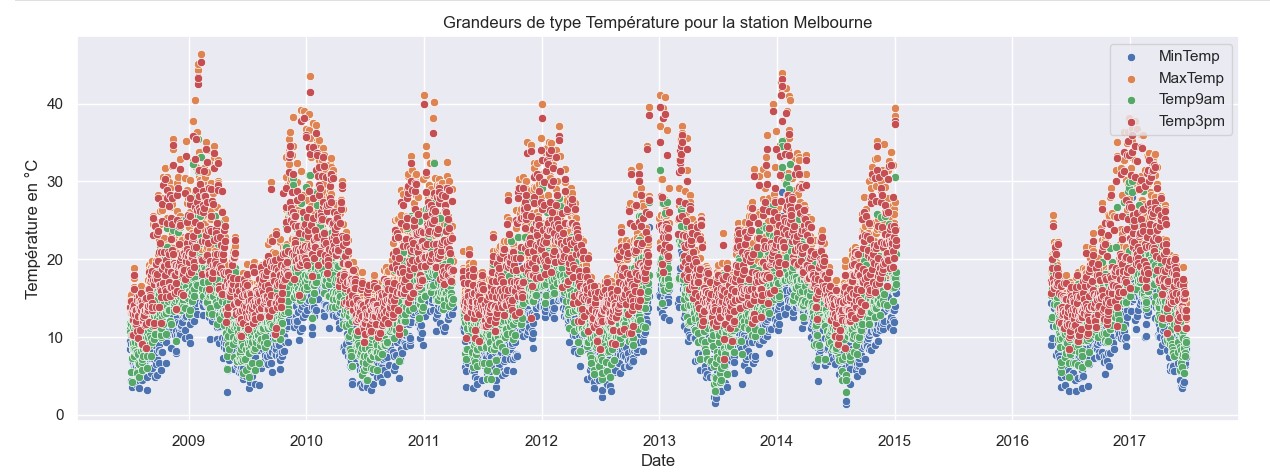
A ce stade de notre réflexion, les informations dont nous disposons ne nous permettent pas de trancher nettement en faveur de l’une ou l’autre. Nous prévoyons de tester les deux lors de la phase de modélisation. Voici un résumé des avantages et inconvénients de chaque méthode :

* **Méthode n°1** : Replacement par la moyenne par station. Conserve des données en remplaçant les nan par des valeurs robustes et cohérentes, mais peut créer des biais.
* **Méthode n°2** : Suppression des NaN . Données finales de qualités, mais peut être en nombre insuffisant.

Détaillons maintenant notre analyse sur la méthode n°1.

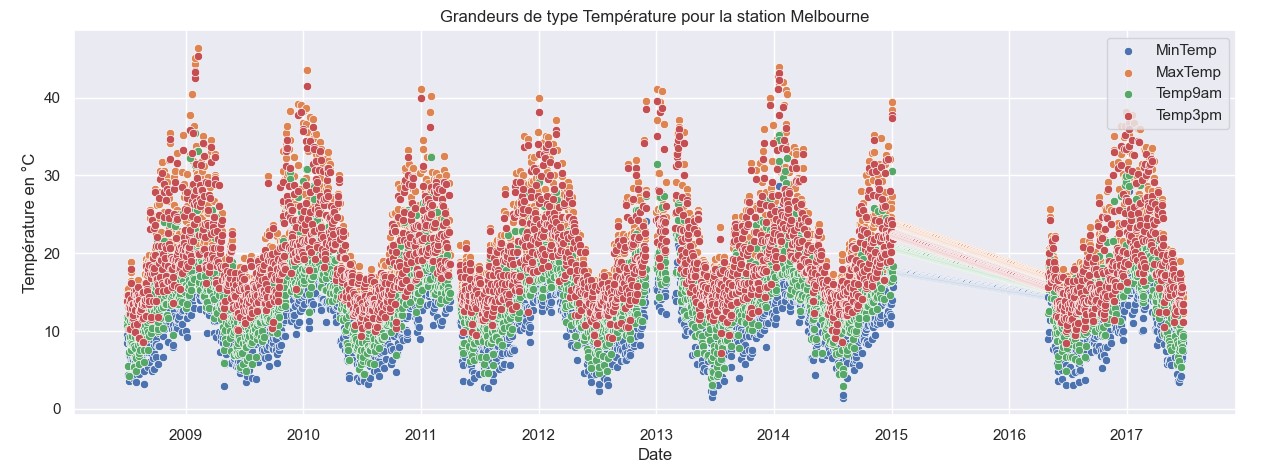
Comme nous l’avons vu, il existe deux grandes catégories pour la distribution des nan :

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

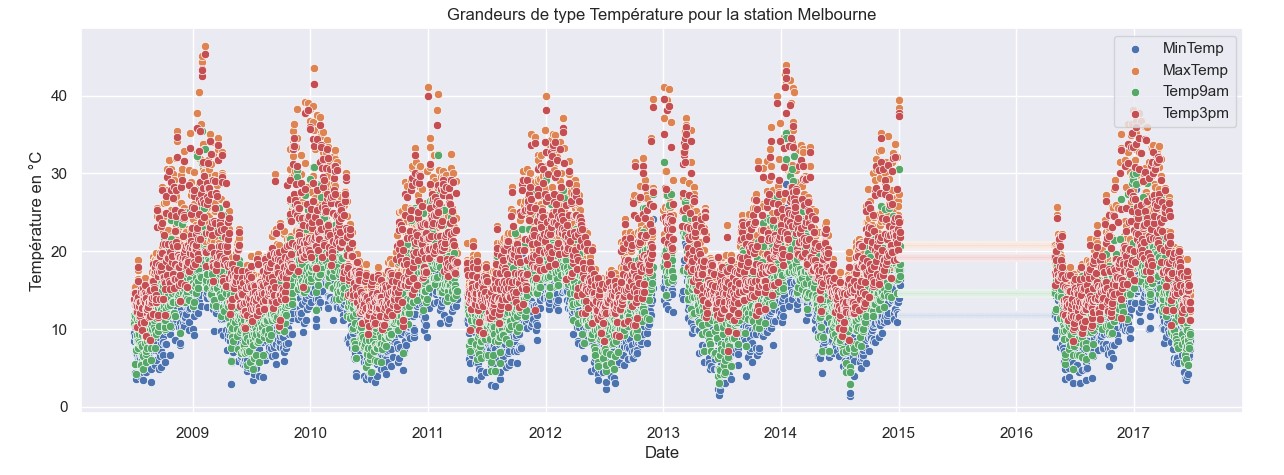
Exemple sur le bloc température mesurées à Melbourne :

Nous pouvons repérer trois choses :

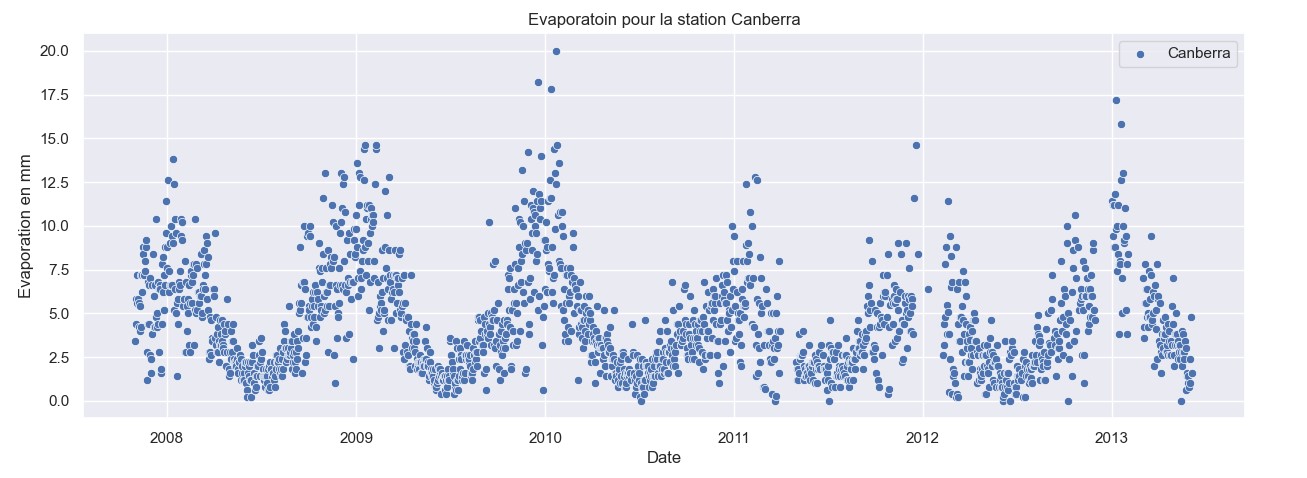
* Les trois trous chronologiques apparaissent sous forme de fines bandes. Ce ne sont pas des nan.
* Une large bande de nan pendant un an et demi environ, de janvier 2015 à avril 2016. Sans doute l’appareil unique de mesure des températures servant à alimenter en données nos 4 variable du bloc température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

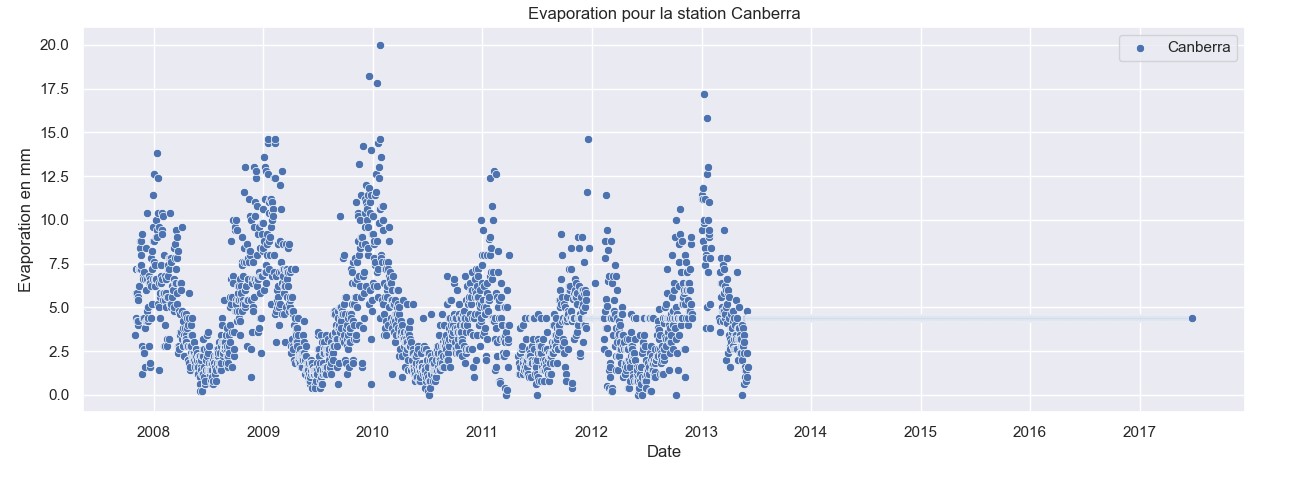
**Nous avons tout d’abord pensé à remplacer les nan par interpolation, au vu de la régularité des mesure, notamment leur périodicité saisonnière. Cela donnerait :

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, 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 (nb : 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. Conserve des données en remplaçant les nan par des valeurs robustes et cohérentes, mais peut créer des biais.
* Méthode n°2 : Suppression des NaN . Données finales de qualités, mais peut 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.

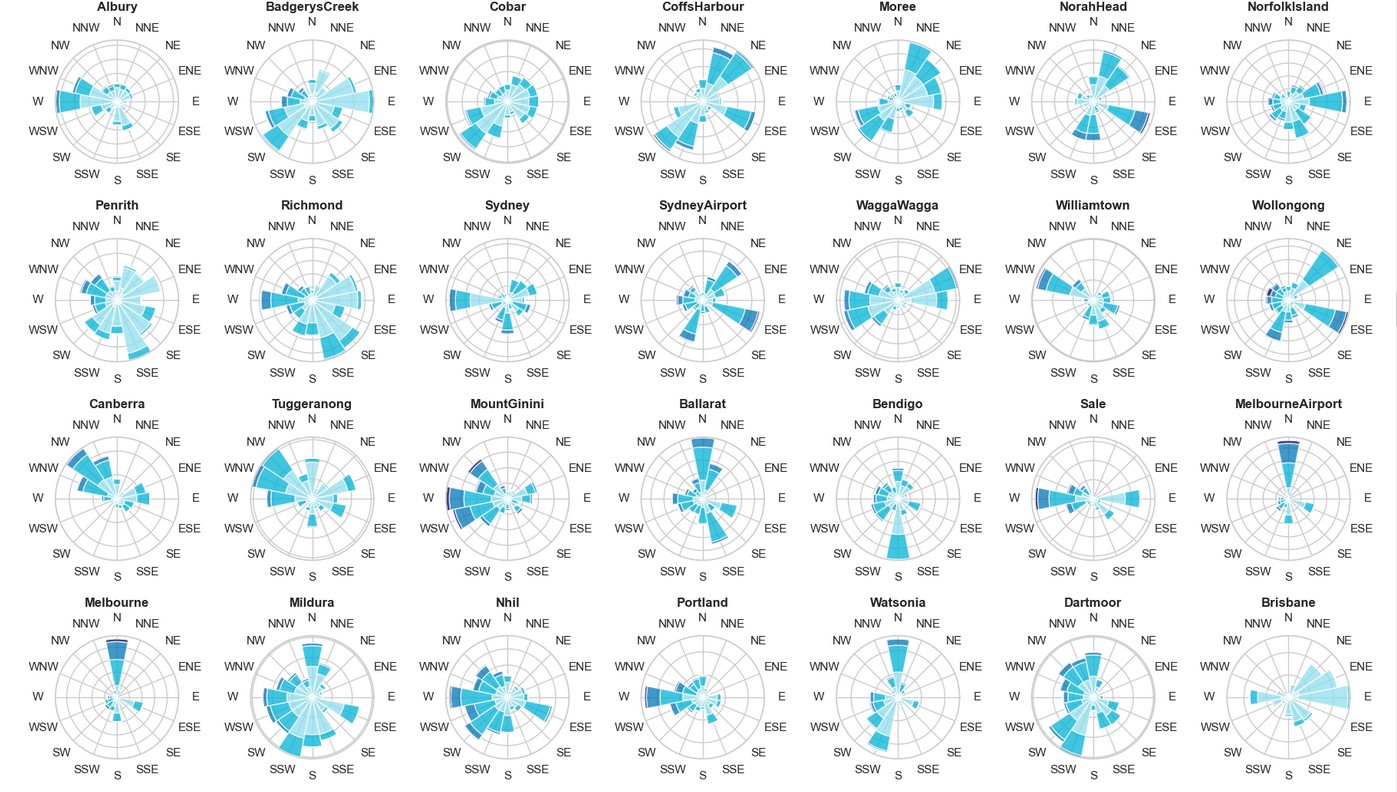
2.1.2 Variables catégorielles ou qualitatives

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 correllées 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 préprocessing).

**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).

## 2.2 Pre processing

2.2.1 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 qu peuvent être faibles (0.5 mm).

2.2.2 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 corellées 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 preprocessing
  + 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 et Feature Hashing Encoding?
  + Ou Supprimer la colonne. On perd la dépendance avec la géographie, mais on garde quand même l’info pertinente sur les grandeurs physiques.

## 2.3 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.