# Rapport d’analyse des données

Table des matières

[Rapport d’analyse des données 1](#_Toc141131691)

[Valeurs nulles 1](#_Toc141131692)

[En globalité 1](#_Toc141131693)

[Répartition géographique 2](#_Toc141131694)

[Répartition temporelle 3](#_Toc141131695)

[Gestion des valeurs nulles 4](#_Toc141131696)

[Corrélations 4](#_Toc141131697)

[Variables quantitatives 4](#_Toc141131698)

[Variables qualitatives 7](#_Toc141131699)

[Géographie 8](#_Toc141131700)

[Répartition géographique 8](#_Toc141131701)

[Pistes 9](#_Toc141131702)

[Bilan général 9](#_Toc141131703)

[TODO / A réfléchir 10](#_Toc141131704)

## Valeurs nulles

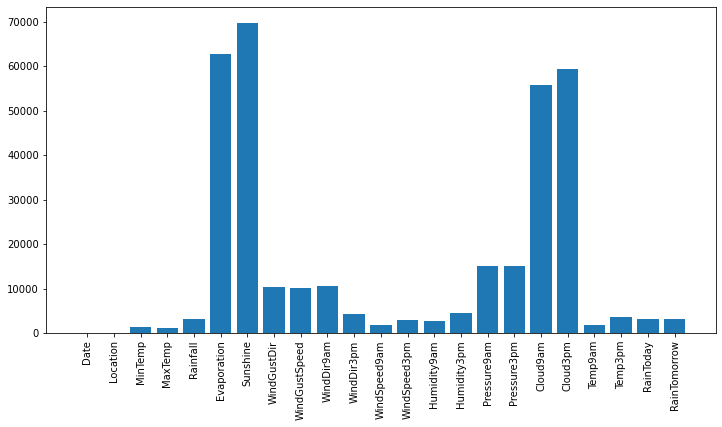
### En globalité

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

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

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

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

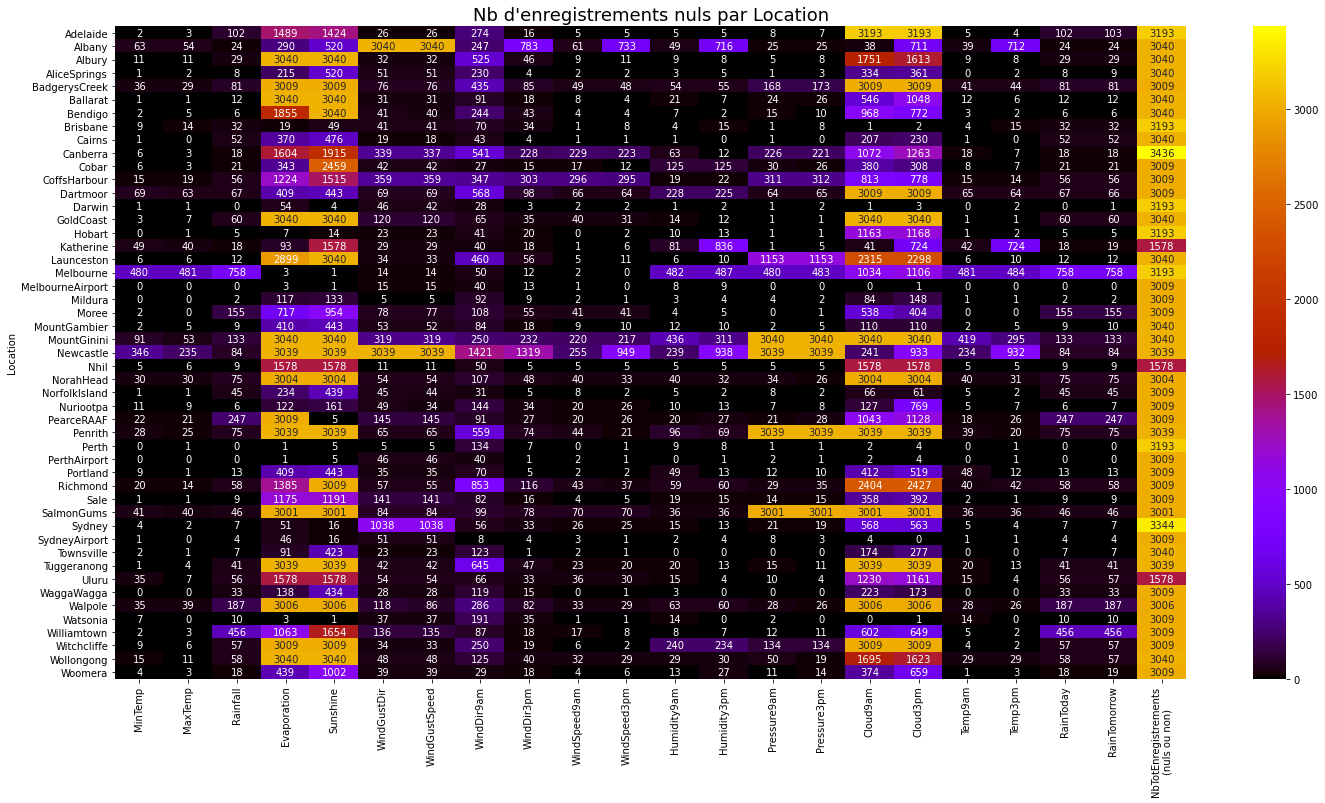


### Répartition géographique

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

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

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



### Répartition temporelle

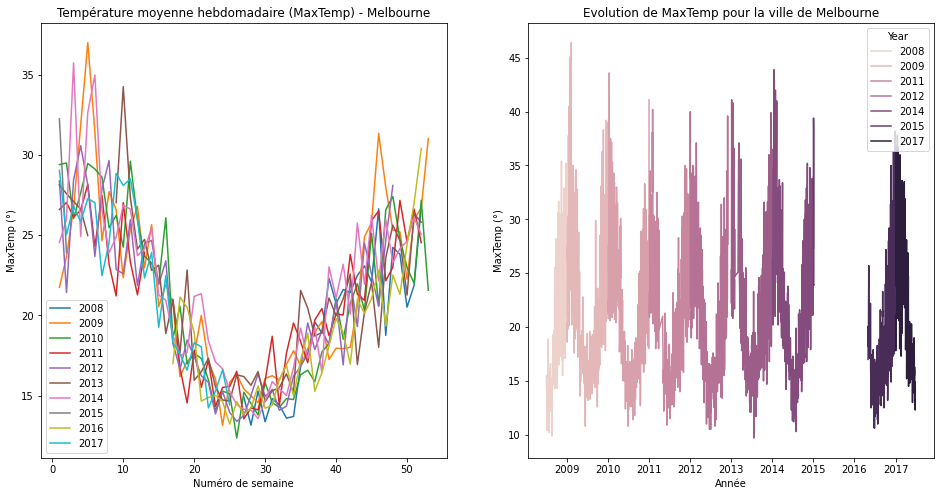
Les données disponibles vont du 1er novembre 2007 jusqu’au 25 juin 2017, ce qui représente 3525 journées. Toutefois, les enregistrements météo ne recouvrent pas l’intégralité de cette plage. On voir sur le graphique précédent que pour la plupart des villes, seules 3000 journées environ sont disponibles.

[à enrichir, non rédigé]

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

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

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



### Gestion des valeurs nulles

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

[à mettre au propre et à discuter/tester]

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

## Corrélations

### Variables quantitatives

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

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

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

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

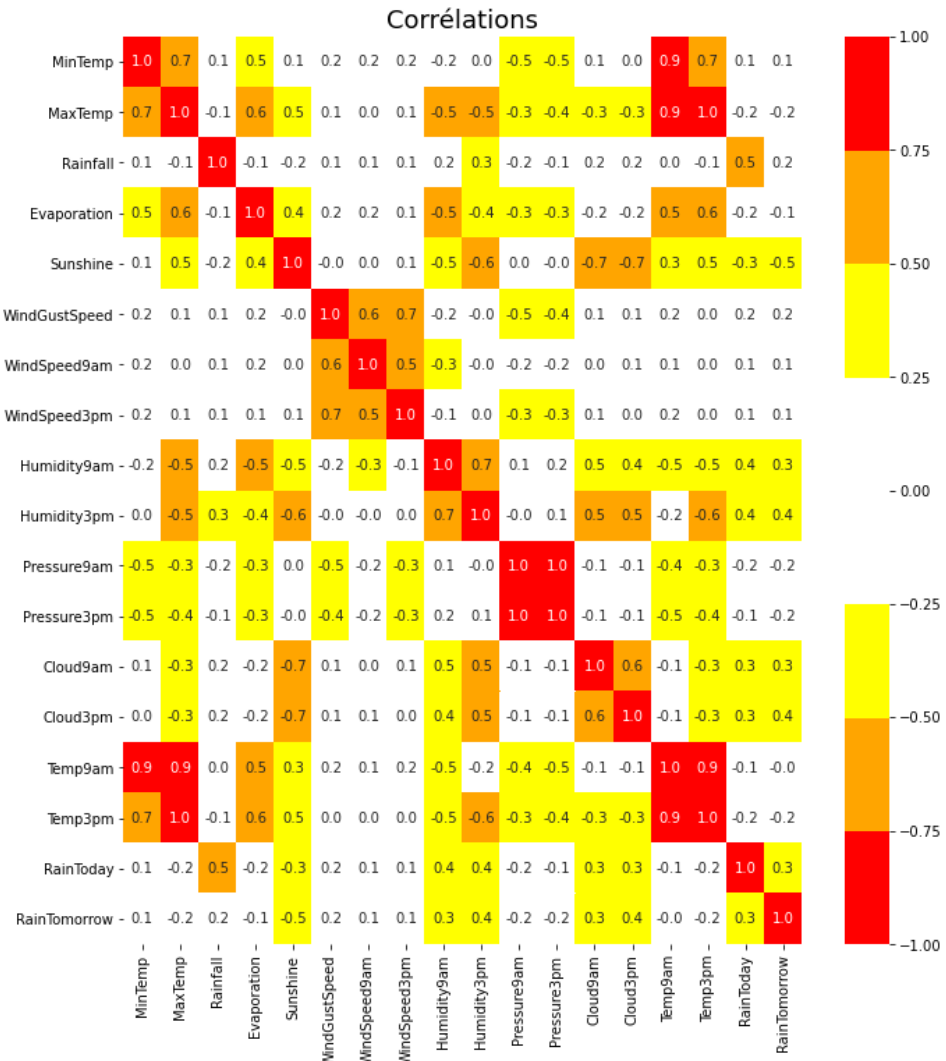
Plus étonnant de prime abord, c’est également le cas de MaxTemp avec Temp9am (0,9), ainsi que Temp9am et Temp3pm

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

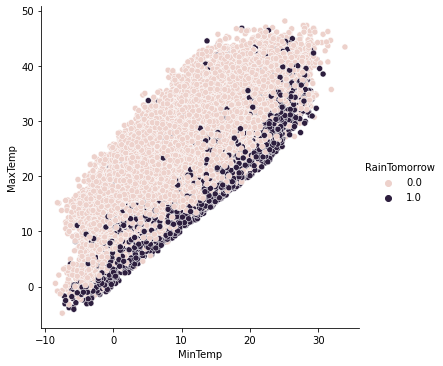
Bref, nous constatons que l’ensemble des variables de températures présentent une très forte corrélation.

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

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



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



### Variables qualitatives

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

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

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

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

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

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

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

## Géographie

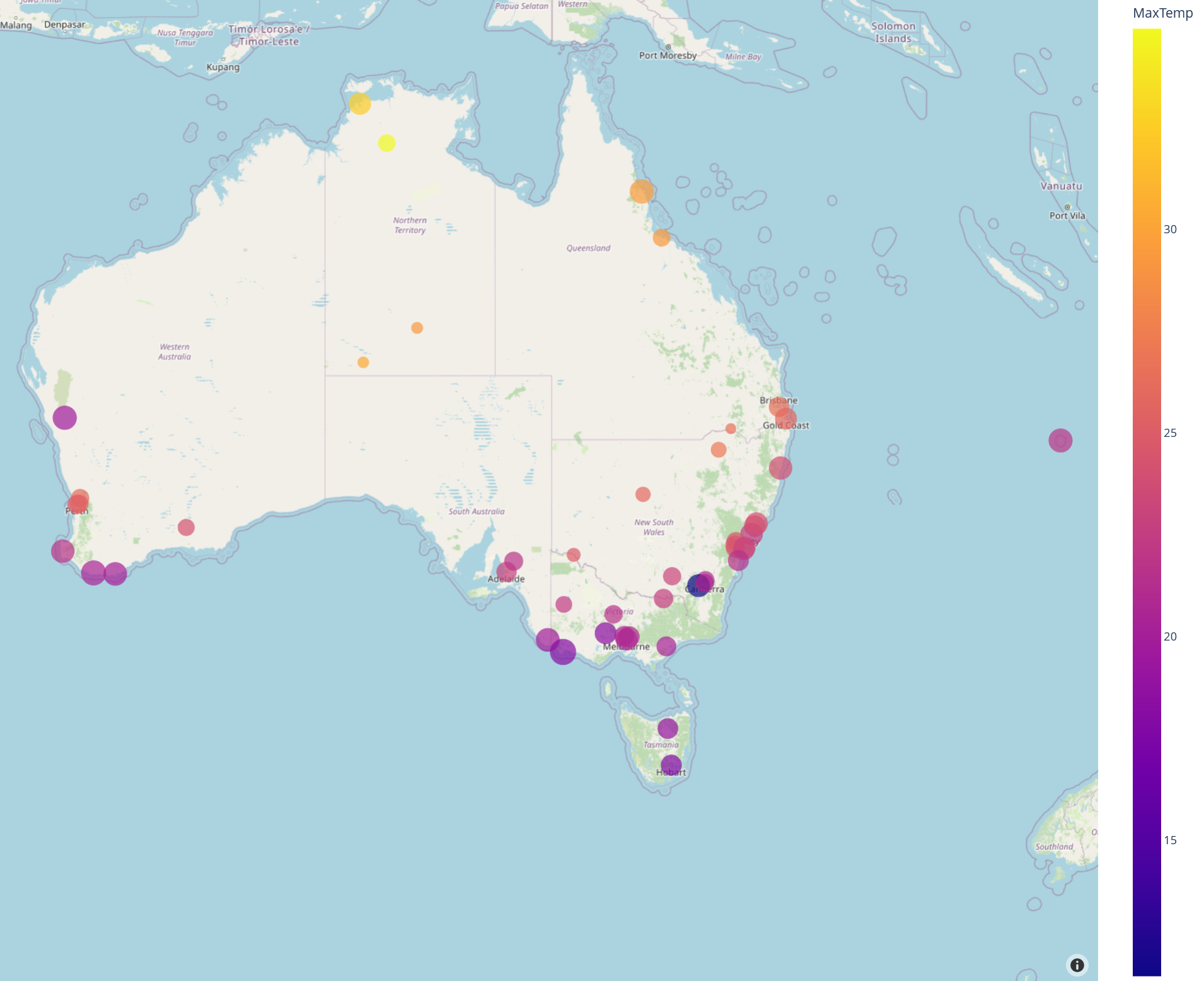
### Répartition géographique

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

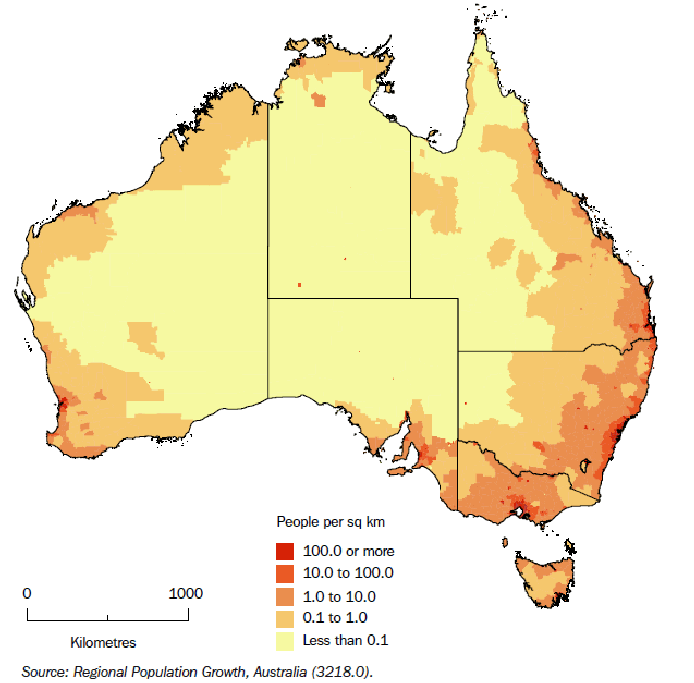
Le graphique ci-dessous montre la répartition des données sur les 49 lieux renseignés. La couleur indique la température maximum moyenne, le diamètre indique la moyenne de la variable RainTomorrow. Ainsi, le petit cercle orange représentant Uluru tout au centre de la carte témoigne qu’il pleut très peu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

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

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



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



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

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

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

### Pistes

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

Nous pourrions également plutôt mettre en place un modèle de clusterisation qui se chargerait lui-même de regrouper les Location présentant des similitudes.

## Bilan général

[in progress]

* Enrichissement des données avec AmplitudeTemp, longitude, lattitude, climat
* Drop des données redondantes
* Gestion des NA ?

## TODO / A réfléchir

Comment faire de la clusterisation temporelle pour chaque ville?

Comment gérer les données nulles (cf. plus haut) ?

Effectuer des analyses temporelles + détaillées