# Projet Statistiques

Théo Lazzaroni, Gautier Poursin

25 Mai 2020

## Introduction

Dans ce projet, nous allons utiliser les données fournies par "Rain in Australia" accessible sur le site Kaggle : https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. Nous allons plus particulièrementnous intéressés à 3 villes : **Brisbane, Canberra et Adelaide**. Ces villes ont été choisies pour leur localisation. Adelaide est située à 1000km à l'ouest de Sydney, Canberra est elle située à 100km des côtes et à 350km de Sydney. Enfin, Brisbane est à 500km au nord de Sydney sur les côtes également. Ces choix nous permettent d'avoir des situations géographiques assez différentes mais où la pluie devrait être assez présente pour avoir des résultats plus probants. **Le but du projet est de réussir via une étude des données à prédire le temps qu'il fera le lendemain** à partir des informations disponibles sur les jours précédents.

## Phase 1: Prise en main des données

Dans cette première partie, nous allons étudier l'ensemble des données qui nous intéressent sur nos 3 villes sélectionnées. La première remarque que nous avons pu faire est que nous ne disposons pas de l'ensemble des données. Certaines valeurs sont **NA**. Pour remédier à ce problème, nous allons proposer 2 solutions:
- on **utilise une loi normale de mêmes paramètres que les lois étudiées.**
- On remplace les valeurs NA par **la moyenne des valeurs de la variable.**

Malheureusement, ces deux choix tendent à nous faire nous approcher d'une loi normale, puisqu'on surestime la population autour de la moyenne. Cependant, le faible nombre de valeurs manquantes ne va que très peu influencé l'allure des variables.

On va commencer par remplacer nos valeurs **NA** par la moyenne des valeurs **non NA**. Voici quelques résultats graphiques obtenus:

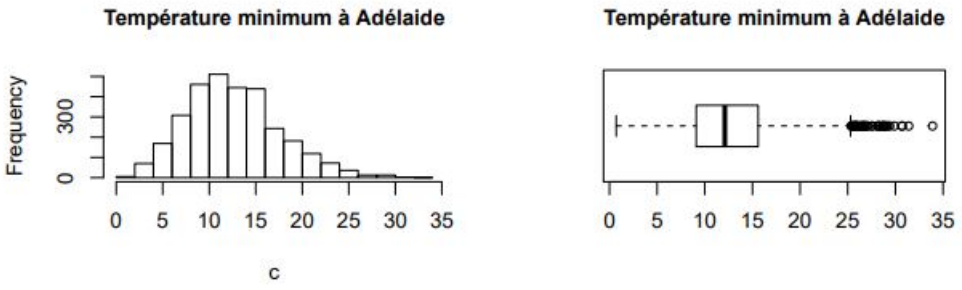


Figure 1: Histogramme des températures à Adélaide

On obtient, avec cette technique, une moyenne de 12.6 degrés et une variance de 24.8.

Avec la seconde technique, la moyenne est de 12.6 degrés et la variance de 25.9. On pourrait penser que les 2 techniques ont un effet similaire sur les résultats. Cependant, il existe certaines variables pour lesquelles la loi normale centrée réduite ne paraît pas être une bonne solution :

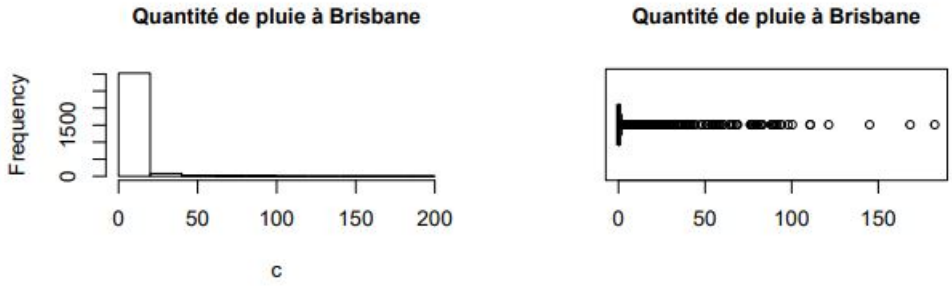


Figure 2: Histogramme des pluie à Brisbane

Le projet comporte des variables catégorielles/qualitatives. Pour traiter ces variables, nous avons décidé de donner une valeur à chaque direction du vent. Par exemple, voici les graphiques obtenus pour la direction du vent à Brisbane lors de jour de pluie ou non.

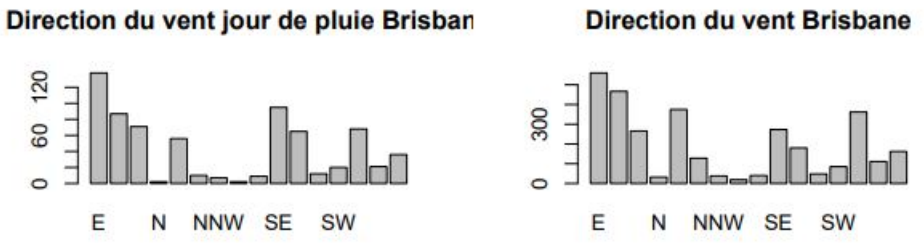


Figure 3: Histogramme des jour de pluie ou non à Brisbane

Par la suite, nous nous sommes interesses à une analyse bivariées des différentes variables. On calcule la covariance et la corrélation puis on trace le Scatter associé. Par exemple, voici le scatter associé à la ville d'Adélaïde entre la température, l'évaporation et la quantité de pluie.

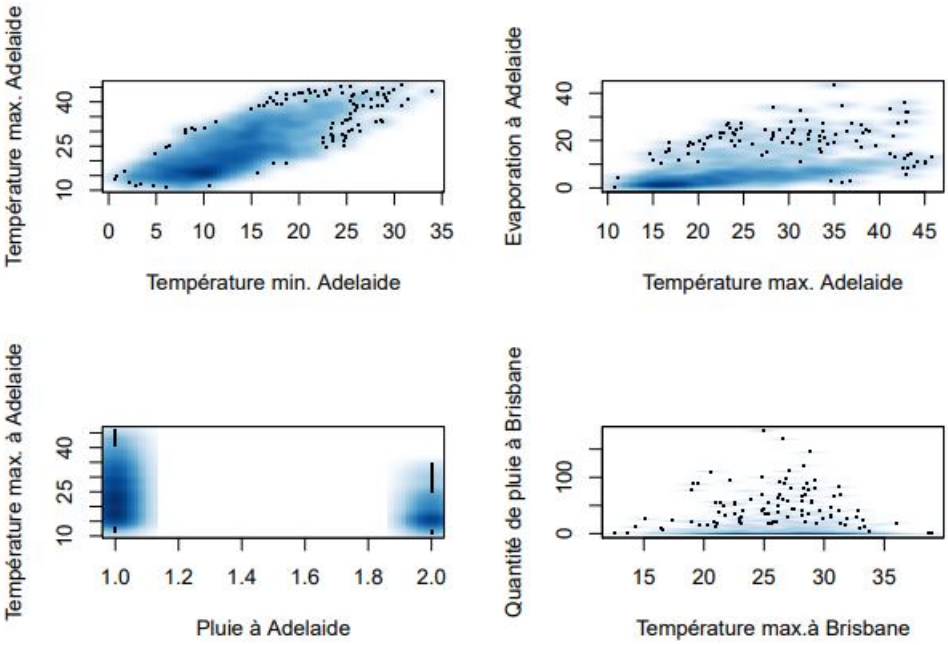


Figure 4: Scatter pour la ville d'Adélaïde

Nous avons pu en conclure que la pluie est fortement influencée par la température, ce qui est logique. Nous avons pu en déduire une autre relation entre température et évaporation: plus la température est élevée, plus l'évaporation est importante. Ces 2 résultats sont en cohérence avec les résultats

attendus.

Il est aussi possible de mettre en évidences les différentes corrélations entre les variables numériques. Par exemple, vous pouvez retrouver ci dessous la graphique des corrélations associé à la ville de Brisbane.

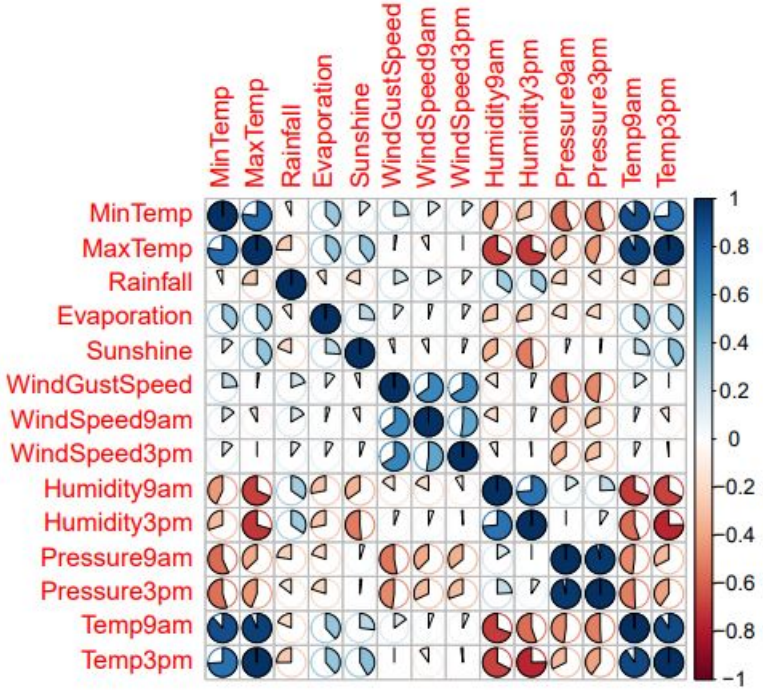


Figure 5: Graphique des corrélations à Brisbane

On peut conclure que température maximal/minimal et température à 3pm/9am sont fortement corrélés positivement. Tandis que humidité et température sont corrélés, mais de manière négative. Si on a une forte température, on a une faible humidité et inversement.

De manière générale, ce graphique est symétrique par rapport à la diagonale et on remarque qu'il est similaire pour les 3 villes. Les remarques faites pour la ville de Brisbane sont valables pour n'importe quelle autre ville.

Pour calculer la covariance entre une variable continue et une variable discrète, il y a 2 possibilités: soit on essaye de rendre continue la variable discrète en appliquant ensuite un test du 2. Si le test s'avère être bon, alors notre approximation n'est pas abérante et donc on peut ensuite calculer la covariance entre 2 variables continues. Sinon, on peut aussi discrétiser notre variable continue. C'est à

dire que l'on sélectionne toutes les valeurs de la variable discrète puis on prend les valeurs de la variable continue pour chaque $x_i$. On peut alors calculer la covariance entre 2 variables discrètes. On préférera ici utiliser la 2ème méthode, les résultats obtenus via R ont été plus concluants.

Concernant la covariance entre 2 variables qualitatives, Lorsqu'on étudie simultanément deux variables qualitatives, il est commode de présenter les données sous forme d'une table de contingence, synthèse des observations selon les modalités des variables qu'elles ont présentées. À partir de cette table, on définit la notion de profil, dont on se sert pour réaliser un diagramme de profils faisant bien apparaître la liaison entre les deux variables, lorsqu'il en existe une. Pour quantifier cette liaison, l'indicateur fondamental est le khideux. Toutefois, comme il n'est pas d'usage commode dans la pratique, on introduit encore les indicateurs phi-deux, T de Tschuprow et C de Cramer, liés au khi-deux. Les deux derniers sont compris entre 0 et 1, et sont d'autant plus grands que la liaison est forte, ce qui facilite leur interprétation.

## Phase 2: Modélisation des lois

Pour la suite du sujet, nous avons décidé de nous passer de quelques variables, en raison d'un nombre trop important de valeurs manquantes.On remarque aisément que les données utilisées sont des séries temporelles et par conséquent les observations sont donc fortement corréler entre elles. De ce fait, on ne peut pas dire que les données forment des échantillons i.i.d. Pour remédier à ce problème, on réalise le sous-échantillonnage suivant : on réalise une distribution mois par mois de chaque ville (par exemple : trouver la loi de la température à 9h à Brisbane, au mois de Septembre). On négligera tout impact du temps sur les données (réchauffement climatique etc.).

Dans un premier temps, nous devions réaliser un sous échantillonage particulier en prenant des jours aléatoirement distant d'au moins 3 jours. Pour cela, on crée une fonction random_day qui nous renvoie une string correspondant à un jour entre m et n. Ensuite, la fonction selection_data s'occupe de faire les tirages (quasi-aléatoires) des jours sélectionnés. En réalité, il choisit un jour entre 1 et 3, 6 et 9, 12 et 15, 18 et 20, 23 et 27. C'est le tirage le plus aléatoire que nous ayons réussi à implémenter pour satisfaire le condition de 3 jours d'écart minimum entre les jours du tirage.

Une fois l'échantillonnage réalisé, on s'intéresse ensuite à trouver une distribution paramétrique pour les différentes variables (normal, gamma . . . ). On

trace d'abord les histogrammes et histogrammes lissés des variables pour ensuite utiliser différents tests (comme Shapiro ou Kolmogorov-Smirnov) afin de trouver la distribution adéquate. On prendra ici $= 10\%$ en valeur seuil minimum de la p-value pour les tests de Shapiro et $= 20\%$ pour le test du $\chi^2$avec une loi gamma, afin de garder l'hypothèse choisie, sinon on rejette l'hypothèse. On écrit donc une fonction etude_variable_mois qui réalise ces opérations et qui renvoie la distribution paramétrique le plus probable. On note qu'ici on remplace les NA values par la moyenne des valeurs non nulles.

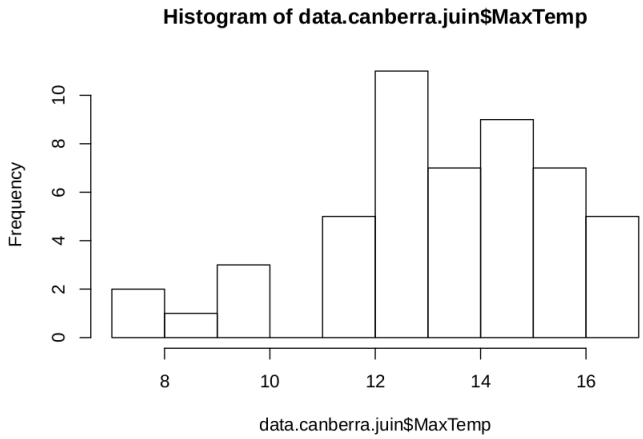


Figure 6: Graphique du maximum de température à Canberra au mois de juin

On remarque après quelques tests que la frontière entre une loi normale et une loi gamma est fine pour certains paramètres. Par exemple, MaxTemp à Brisbane oscille entre loi normale et loi gamma. On va donc implémenter une deuxième fonction qui va renvoyer sur tous les mois dans une ville, la distribution paramétrique la plus plausible. Ainsi, on prendra la même loi pour chaque mois de l'année afin de pouvoir comparer les EMV du maximum de vraisemblance de mêmes paramètres (cf. question 4)).

Après avoir obtenu les résultats via le fichier rmd, voici la répartition choisie: [MinTemp : normale; MaxTemp : normale; Rainfall : gamma; WindGustSpeed : gamma; WindSpeed9am : gamma; WindSpeed3pm : gamma; Humidity9am : normale; Humidity3pm : normale; Pressure9am : normale; Pressure3pm : normale; Temp9am : normale; Temp3pm : normale]
Elle a été choisie en pernant la "moyenne" des trois villes et en supposant que les variables 3pm et 9am d'une même grandeur doivent suivre une loi identique.

On s'intéresse maintenant à l'écriture de la vraisemblance des différentes variables en utilisant les distributions paramétriques supposées à la question 2). Une

loi normale $\mathcal{N}(\mu, \sigma^2)$ a pour fonction de densité : $f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
On en déduit la vraisemblance $L(x_1, ..., x_n|\mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{n/2} exp(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2})$.

On peut donc calculer les EMV de $\mu$ et $\sigma$ en dérivant selon les deux paramètres. On obtient finalement $\bar{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ respectivement les moyennes et variances empiriques.

On sait que les intervalles de confiance des deux paramètres sont les suivants : $\bar{X} - \alpha/2\frac{S}{\sqrt{n-1}} < \mu < \bar{X} + \alpha/2\frac{S}{\sqrt{n-1}}$ avec $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
De même, $\frac{nS^2}{1-\alpha/2} < \sigma^2 < \frac{nS^2}{\alpha/2}$.

Une loi gamme $\Gamma(a, b)$ a pour fonction de densité : $f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$.
On en déduit la log-vraisemblance $\mathcal{L}(x_1, ..., x_n|a, b) = na\log(b) - n\log(\Gamma(a)) + (a-1)\sum_{i=1}^{n}\log(x_i) - b\sum_{i=1}^{n} x_i$.

En dérivant selon b, on obtient un estimateur du paramètre $\bar{b} = \frac{a}{\bar{x}}$ avec $\bar{x}$ la moyenne empirique. De même, on trouve $\bar{a} = \log(b\sum_{i=1}^{n} x_i)$, cependant nous ne voyons pas comment aller plus loin dans le calcul étant donnée la realtion entre a et b même.

On a donc décidé de prendre un simple estimateur pour continuer dans les questions suivantes. On peut estimer $a$ et $b$ grâce à la méthode des moments en sachant que $E(X) = \frac{a}{b}$ et $V(X) = \frac{a}{b^2}$. Finalement, on déduit de la méthode des moments : $a = \frac{E(X)^2}{V(X)}$ et $b = \frac{E(X)}{V(X)}$.

Cependant, ces deux valeurs données grâce à la méthode des moments ne nous donnent pas un EMV, on va donc tenter de calculer directement le maximum de la fonction pour obtenir les estimateurs de $a$ et $b$.

On a juste à tracer leur évolution en fonction des mois dans une même ville.

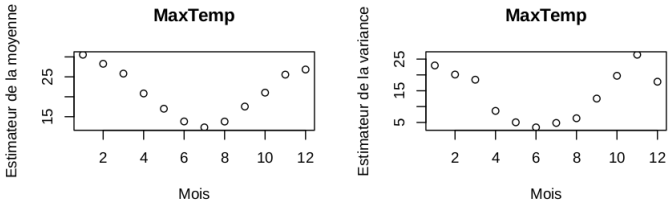


Figure 7: Moyenne et variance température à Brisbane

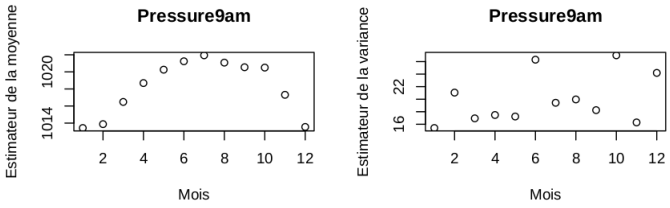


Figure 8: Moyenne et varaince de pression à Brisbane

On remarque un comportement sinusoïdal concernant les variables suivant une distribution normale. C'est évidemment un comportement prévisible et qui met en avant les différences entre été et hiver. Noter aussi pour la variance que l'on a de plus gros écarts l'hiver pour la température min et inversement pour le max. Pour la loi gamma on dirait plutôt du stationnaire avec du bruit.

Enfin, nous devions vérifier les suppositions trouvées lors de la question précédente. Nous avons donc réalisé des tests, que vous pouvez retrouver sur le fichier rmd. On remarque qu'on retrouve toujours une p-value très haute et on valide donc l'hypothèse H0. **Ce résultat nous conforte dans l'idée que les paramètres ne sont pas constants et varient avec les saisons.**

# Phase 3: Prédiction de pluie

On notera que l'on s'est permis d'utiliser la variable RainTomorrow pour éviter d'avoir à utiliser un décalage d'indice avec RainToday pour savoir s'il pleut effectivement le lendemain.

Maintenant que nous avons déterminé les différentes lois suivies par les variables ainsi que l'évolution des différents paramètres statistiques, nous allons désormais nous recentrer sur un modèle prédictif pour Adelaide, Brisbane et Canberra.

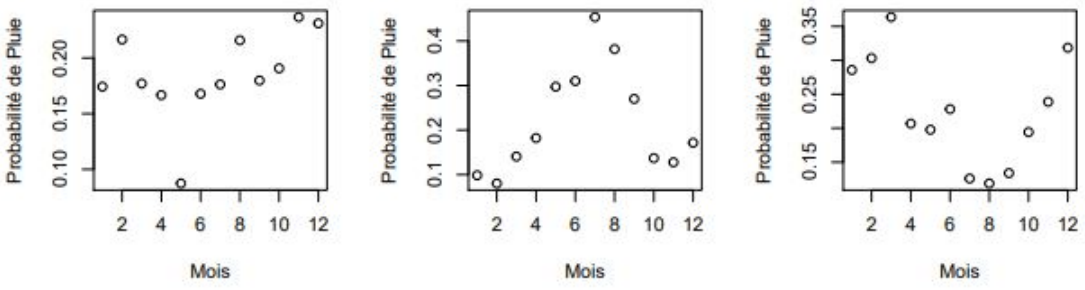


Figure 9: Probabilité de pluie par mois le lendemain à Canberra, Adelaïde et Brisbane

Pour commencer, on s'intéresse à la probabilité de pluie le lendemain mois par mois. Pour cela, on implémente une fonction qui renvoie les données d'une ville sur les 10 ans pour 1 mois et on calcule ensuite la probabilité de pluie le lendemain durant le mois. Nous réalisons en plus des tests de shapiro.

A l'aide des résultats des tests de Shapiro et notamment avec la pvalue, on remarque que la probablité de pluie par mois le lendemain à Brisbane et Adelaïde suivent une loi normale. La pvalue est supérieur à 20%(24% pour Adelaïde et 60% pour Brisbane).

Concernant la ville de Canberra, il est assez difficile d'affirmer qu'elle suit une loi normale. En effet, la pvalue est faible (7% ) et il n'y a pas l'allure d'une loi normale. On se permettra néanmoins de valider l'hypothèse pour la suite.

Maintenant que l'on a régardé comment se comportait RainTomorrow dans les différentes villes et selon les mois, on va tenter de proposer un modèle prédictif basé sur une régression logistique en modélisant la probabilité conditionnelle de pluie le lendemain selon les autres variables.

On pose $Y = f(X_1, ..., X_n)$ avec $X_1, ..., X_n$ les variables explicatives de $Y$. On va estimer un échantillon $Y_1, ..., Y_n$ et on a $L(Y_1, ..., Y_n|B) = \prod_{i=j}^{n} \frac{exp(\sum_{i=1}^{n} B_{i,j} X_{i,j})}{1+exp(\sum_{i=1}^{n} B_{i,j} X_{i,j})}$

On utilise ensuite la fonction glm pour estimer un modèle prédictif.

On remarque que le critère d'information d'Akaike est très élevé dans les trois cas (plus de 300), on peut donc s'attendre à ce que le modèle ne soit pas optimal. En effet, ce critère permet de mesurer la qualité Nous ne sommes pas certains de comment interpréter les différents résultats de la fonction glm. On suppose qu'il pleut environ 45 jours à Canberra et 140 à Adelaide sur la totalité des jours dans les data-frames.

Après quelques recherches, nous pensons avoir compris que glm utilisait à la fois les log-vraisemblances, la matrice Hessienne, approximée via l'information de Fisher par une méthode itérative (d'où le Number of Fisher Scoring Iterations donné dans le summary de glm). On approxime $B$ itérativement par $B_{k+1} = B_k \frac{\nabla B_k}{H_{B_k}}$ où $\nabla B_k$ est la dérivée de la log-vraisemblance par rapport à $B$.

Nous ne sommes pas aller plus loin dans le projet car nous avons eu le plus grand mal à bien comprendre comment utiliser efficacement les résultats fournis par glm. Cependant, on peut supposer que les estimateurs soient assez différents selon les villes et qu'il soit donc important de combiner les données pour parvenir à un résultat adéquat.

# Conclusion

Bien que nous ne soyons pas parvenu à prédire le temps qu'il fera demain, ce projet nous a permis de prendre conscience de la puissance que les statistiques pouvaient avoir, si elles sont bien utilisées.

Après avoir pris en main le data-frame initial en traçant différents graphes et en mettant en valeur les corrélations entre les variables, nous avons ensuite modéliser les lois des variables par des distributions paramétriques normale ou gamma et nous avons estimé les différents paramètres de celles-ci dans diverses conditions.

Enfin, via une régression logistique, nous avons estimer un modèle prédictif pour la pluie le lendemain. Il aurait fallu comparer les différens modèles entre eux et les mélanger pour obtenir un modèle utilisable partout en Australie.