Projet De Statistiques - La pluie en Australie

Loan Godard

Le 25 mai 2020

L'objet de ce projet est de prédire la pluie en Australie en se basant sur un certain nombre de variables quantitatives et qualitatives. Pour cela nous suivrons le sujet décomposé en 3 phases. Dans la première phase, on découvre et on étudie les variables données, on met en évidence certaines corrélation qui les lies et on remarque certaines tendances. Dans la deuxième phase, on cherche à déterminer les lois de certaines variables en effectuant un certain nombre de test. Enfin dans la troisième phase, on tente de prédire la pluie à l'aide d'une régression logistique.

J'ai trouvé la chronologie du sujet logique et assez naturel dans la démarche : on découvre les données puis on les étudies de plus en plus en profondeur. Malgré cela je pense qu'on aurait peu atteindre l'objectif qui est "prédire la pluie" plus rapidement et en supprimant quelques étapes.

J'ai travaillé seul sur ce projet, je n'ai donc pas eu de problèmes de communication ou de répartition du travail, mais le projet était assez long à faire seul.

Dans le compte rendu suivant, j'essaye de suivre la chronologie des questions pour ne pas perdre le file du sujet.

1 Prise en main des données

On importe dans un premier temps les données dans un dataFrame et on les découvre
Avant d'étudier les données, nous devons d'abord les traiter en éliminant, imputant ou en remplaçant les données manquantes. Notre première stratégie est de remplacer les données manquantes par la moyenne du mois pour chaque villes.

Nous étudions 22 variables différentes (Nous laissons de côté "RiskMM"). Toutes les variables sont quantitatives sauf les directions du vent, les fractions du ciels obscurcis par les nuages, et la variable "RainToday".

Pour remplacer les données manquantes, on utilise plusieurs stratégie :

- Dans tous les cas on remplace les données manquantes de "RainToday" par la majorité de chaque Ville: s'il y a plus de jour de pluie que de non pluie à NorfolkIsland alors les données manquantes seront remplacés par ""Yes"
- Dans tous les cas on complète les données manquantes des directions grâce à l'algorithme kNN, on prend un k arbitraire.
- Pour les autres données on remplace les données manquantes soit par la moyenne de chaque ville soit par la médiane de chaque ville, on choisira lesquelles étudiées plus tard.

J'ai choisis de me concentrer sur les données de trois villes choisis arbitrairement de sorte à ce qu'elle soit assez éloignée : Ballarat, Perinth et NorfolkIsland. Malheureusement, je n'avais pas fait attention et

je me suis rendu compte en fin de projet qu'il manquait beaucoup de données à Perinth, c'est pourquoi j'ai décidé de remplacer l'étude de cette ville par l'étude d'Albury.

Toutes les variables sont quantitatives sauf les directions du vent et la variable "RainToday" qui sont qualitatives.

Je vais me concentrer sur les données suivantes pour les premières études de données : Pressure3pm,MinTemp,MaxTemp,Temp3pm,Humidity3pm,Rainfall,Cloud3pm. J'ai choisis ces données par intuition car je pense que ces variables ont un impact direct sur la pluie.

On fait une étude descriptive des données.

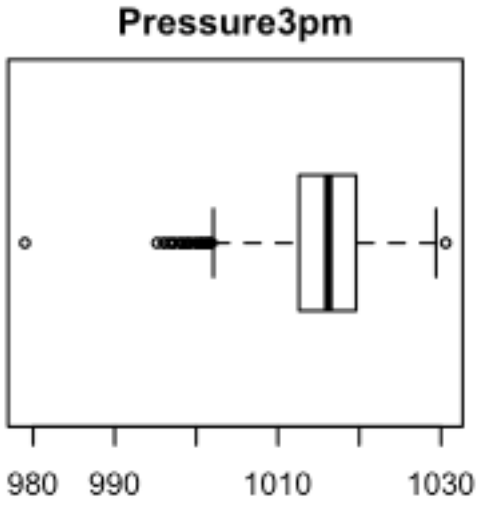

Figure 1: Pression à NorfolkIsland

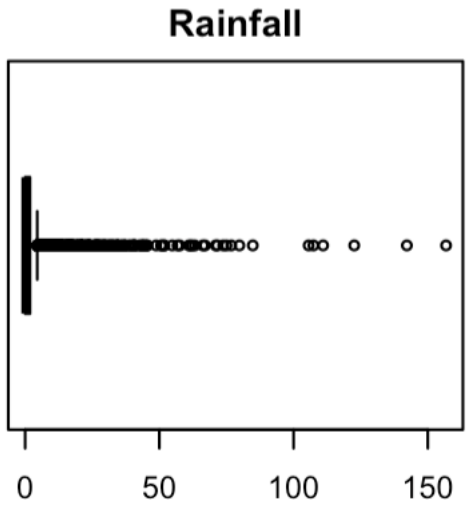

Figure 2: Pluviosité à NorfolkIsland

Dans une premier temps on étudie les différents quartiles des variables choisis.

On remarque que la pression est assez concentré.En effet, la moyenne est égale à la médiane qui est égale à 1016. L'écart est seulement de 5 pour une variable qui prend des valeurs entre 980 et 1030. Pour Rainfall (la pluviosité), la médiane est à 0.2 pour des variables positives qui vont jusqu'à 156 au maximum. On constate donc une variable dont les valeurs sont extrêmement concentré mais qui on des extrêmes importantes.

Les autres variables ne sont pas particulièrement surprenante et sont assez étalés, sans valeurs extrêmes.

Nous avons aussi tracé les histogrammes de chaque variables et on remarque une forme en cloche pour presque tous les histogrammes. On peut alors supposer que certaines variables ont une répartition de loi Normale (à confirmer avec des test). On étudié les données dont les valeurs manquantes on été remplacé par la moyenne ou par la médiane et on ne remarque presque pas de différence. On étudiera ensuite uniquement les valeurs dont les données manquantes ont été remplacé par la médiane.

Après avoir étudié de manière descriptive les variables, nous effectuons une analyse bivariée pour chercher des liens entre certaines variables. On a ci-dessous une représentation de la matrice de corrélation :

Sans trop de surprise la pression le matin est corrélé avec la pression de l'après-midi et la température minimale et corrélé avec la température maximale.

Plus intéressant : L'humidité est corrélé avec le taux de nuage et la pression. Aussi la pluviosité est fortement corrélé avec l'humidité et la pression est corrélé avec l'humidité.

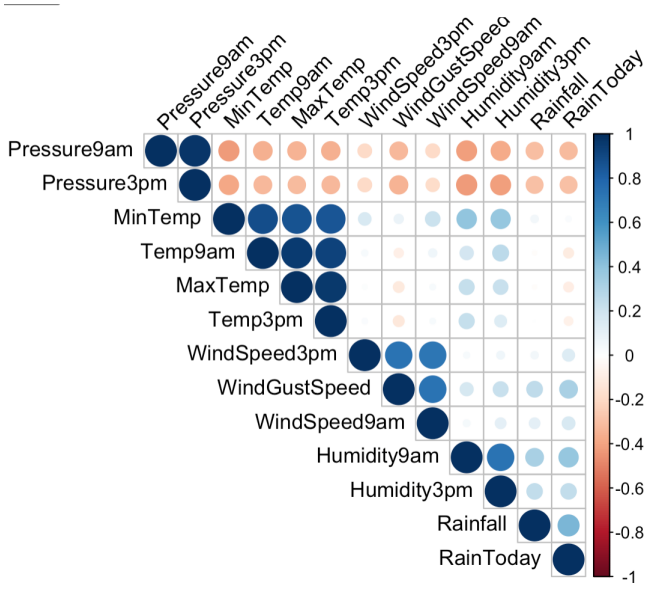

Figure 3: Représentation de la matrice de corrélation

De la même manière on trace la matrice de covariance avec la fonction cov. Pour calculer la covariance entre une variable continue et une variable discrète, on peut essayer de rendre continue la variable discrète pour ensuite faire une test du Khi 2. Si le test a une p-value convenable, on pourra considèrer que notre approximation et bonne. Sinon, on fait l'inverse est on discrétise notre variable continue et on pourra alors calculer la covariance de deux variables discrètes. On remarque que la covariance de l'humidité le matin avec la pression de l'après midi est assez élévée (26.8). On constate également que le taux de nuage le matin varie avec l'humidité de l'après-midi. Enfin, on remarque que L'humidité varie négativement avec la température (-128).

On remarque grâce à la figure 4 que lors des jours de pluie: la pression a tendance a être légèrement plus basse et plus dispersé, au contraire la vitesse du vent est plus élevé et a des valeur maximale plus élevé. On remarque également que la vitesse du vent décroît en fonction de la pression (C'est normale car le vent naît sous l'effet des différences de pressions). Enfin, la pression est décroissante en fonction de l'humidité.

Nous identifions alors trois variables importante qui interviennent dans le fait qu'il pleuve ou non : la pression, l'humidité et la vitesse du vent. Une piste pour prédire la pluie serait alors de prédire ces variables qui semblent lié à la pluie, mais cela serait à vérifier.

La figure 5 montre deux représentations graphiques d'une variable discrète en fonction d'une variable continue. Les boxplot et les violons expriment les mêmes données sous deux manière un peu différentes, elles représentent toutes les deux de quelle manière sont distribuées les données.

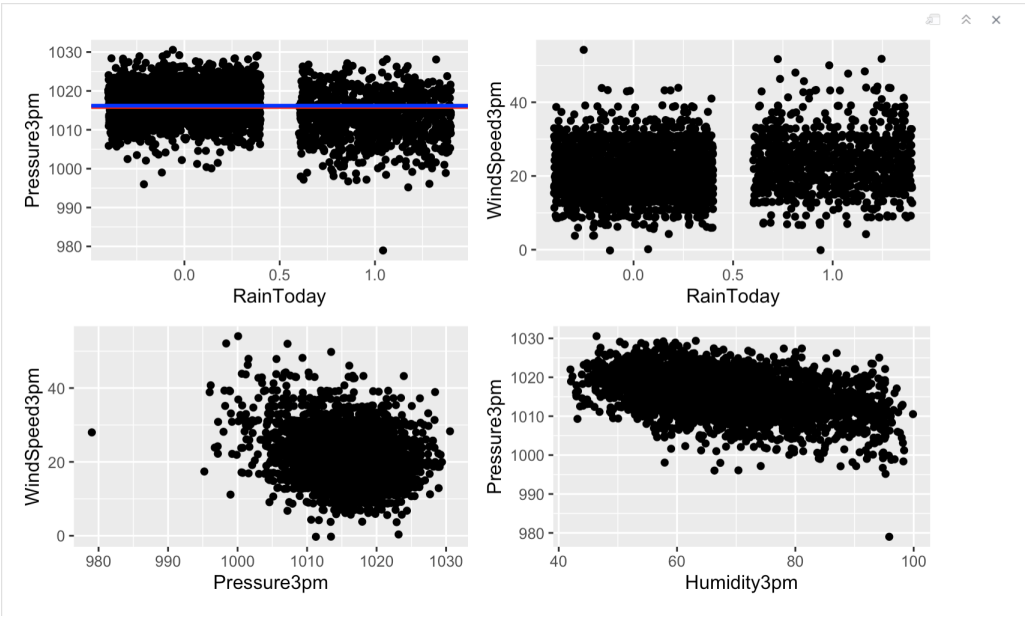

Figure 4: Étude bivarié de certaines variables

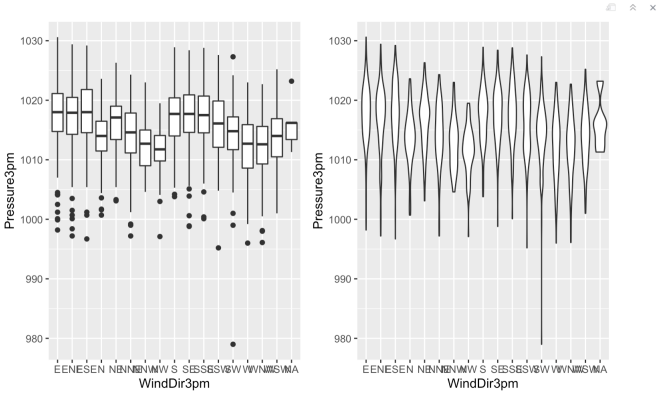

Figure 5: Étude bivarié de certaines variables

2 Modélisation des lois

Dans un premier temps, on effectue plusieurs opérations pour sélectionner des jours aléatoirement de chaque mois étalé sur les 9 ans disponibles, espacés de 3 jours au minimum. Cela permet de limiter la dépendance entre les données de jours successifs.

Pour chaque ville et pour chaque variable, on trace les histogrammes et densité pour les douze mois de l'année.

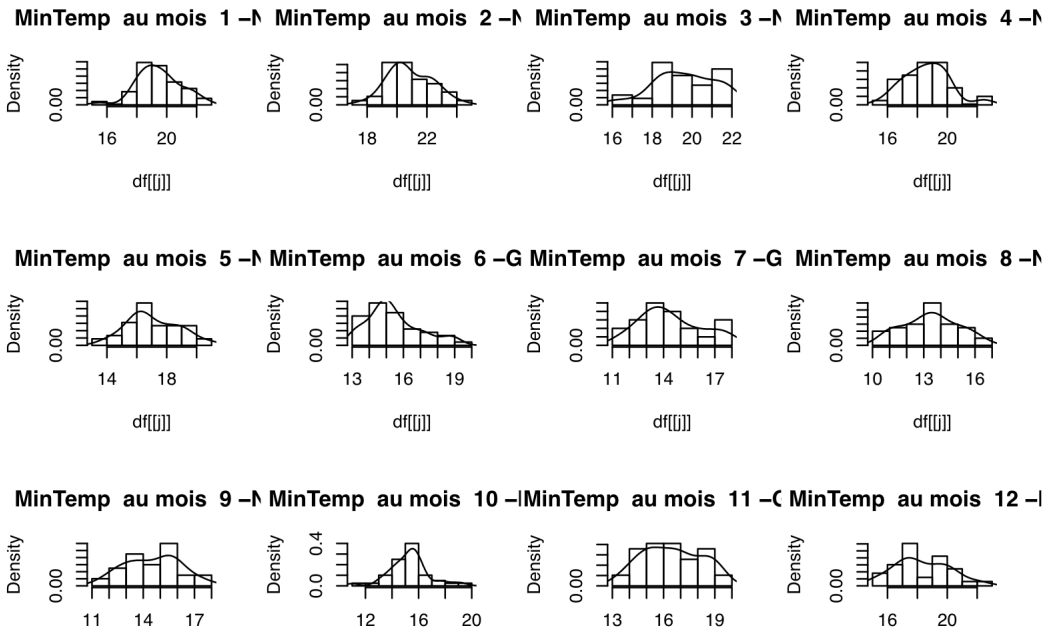

Figure 6: Densité de minTemp sur les 12 mois

On remarque que la plupart des densités ressemblent à une densité de loi normale, on effectue alors un test de shapiro pour tester si les variables suivent vraiment une loi normale. Les résultats sont données par la matrice suivante :

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	1	1	1	1	1	1	1	1	1	1	1	1
[2,]	1	1	1	0	1	1	0	1	0	1	1	1
[3,]	0	1	1	0	1	1	1	1	0	1	1	0
[4,]	1	0	1	1	1	1	1	1	1	1	1	0
[5,]	1	1	0	1	1	0	1	1	0	1	1	0
[6,]	0	1	1	1	1	0	1	1	1	1	1	1
[7,]	0	1	1	0	1	0	0	1	1	1	0	1
[8,]	0	1	0	0	1	0	1	1	1	1	1	1
[9,]	0	1	0	0	1	0	1	1	1	1	1	1
[10,]	1	1	0	1	1	1	1	1	1	1	1	1
[11,]	0	1	1	0	1	1	1	1	1	1	1	1

Figure 7: Résultat des test shapiro

les lignes représentent l'indice de la variable dans intCaracteristique =c(3,4,7,10,11,12,13,14,15,16,17) dont les coordonnées représentent l'indice des variables dans les dataframes contenant toutes les variables. Les colonnes représentent les mois.

Ainsi, si resultat[i,j]==1 c'est que le test de shapiro du caractéristique indexé i au mois j a une p-value supérieur à 10% et donc qu'on ne rejette pas H_0 : La variable suit une loi Normale. On acceptera par la suite H_0. On a 101 variables dont la p-value a été supérieur à 10% et 31 variables qui n'ont pas la p-value supérieur à 10%. Au vu des courbes de densité obtenues, on supposera que les 31 variables restantes suivent une loi gamma. Par exemple, resultat[4,3] = 1 donc la 10e variable de nos données au

3e moi suit une loi normale : C'est WindSpeed9am.

Soit $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$ La vraisemblance de X est donné par :

$$L_{(\mu,\sigma)}(x_1 \dots x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

On passe à la log-Vraisemblance pour maximiser plus simplement cette fonction et ainsi trouver l'estimateur du maximum de vraisemblance des paramètres...

$$\mathcal{L}_{(\mu,\sigma)}(x_1 \dots x_n) = -n(\ln(\sigma) + ln(\sqrt{2\pi})) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

On maximise cette dernière fonction en $\theta = (\mu, \sigma)$.

$$\nabla\mathcal{L}_{(\mu,\sigma)}(x_1 \dots x_n) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\boxed{\Leftrightarrow \hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\tfrac{1}{n}\sum_{i=1}^{n} X_i, \tfrac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2)}$$

Les intervalles de confiance des estimateurs sont :

$$\hat{\mu} \in [\bar{X} \pm q_{\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}]$$

Où $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ (Se prouve avec loi de Student) et

$$\hat{\sigma} \in [\frac{nS^2}{1-q_{\frac{\alpha}{2}}}; \frac{nS^2}{q_{\frac{\alpha}{2}}}]$$

Soit $Y \hookrightarrow \Gamma(a, b)$

La log-vraisemblance est donné par

$$\mathcal{L}(a,b) = na\ln(b) - n\ln(\Gamma(a)) + (a-1)\sum_{i=1}^{n}\ln(x_i) - b\sum_{i=1}^{n}x_i$$

Les EMV sont :

$$\boxed{\hat{a} = \bar{X}_n \text{ et } \hat{b} = \frac{a}{\bar{X}_N}}$$

Ensuite on trace l'évolution des moyennes de chaque variables en fonction du temps. On remarque que les température sont sinusoïdales et minimum en hiver et maximum en été, ce qui est logique car elle évolue en fonction des saison. Pour la pression on vois une évolution parabolique (cf Figure 8) : la pression semble en moyenne maximale en été et minimale en hiver. Mais aussi la variance est maximum en été donc il faut faire attention aux valeurs extrêmes qui pourraient augmenter fortement la moyenne en été.

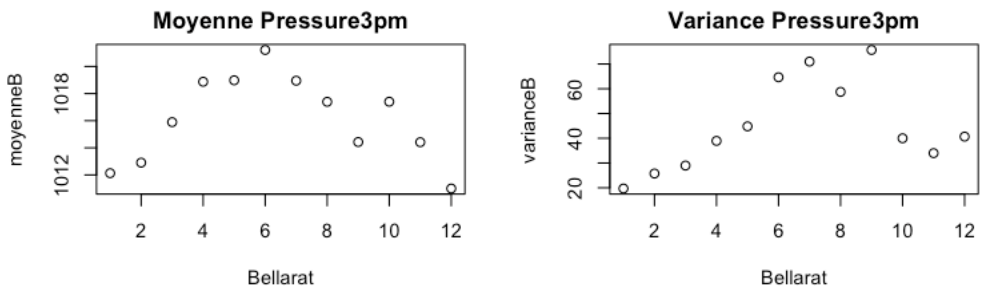

Figure 8: Évolution de la moyenne au cours du temps

3 Prédiction de la pluie

On étudie la probabilité qu'il pleuve demain chaque mois, simplement en divisant le nombre de jours de pluie par le nombre de jours chaque mois. La figure 9 nous montre la densité de la probabilité qu'il pleuve demain. Et on effectue un test de shapiro, dont la p-value est presque toujours supérieur à 0.10. Donc on vas supposer que la probabilité qu'il pleuve demain suit une loi normale.
Ensuite on montre à l'aide d'un test Chisquare que la probabilité de pluie en hiver est différente de la probabilité de pluie en été (cf Figure 10).

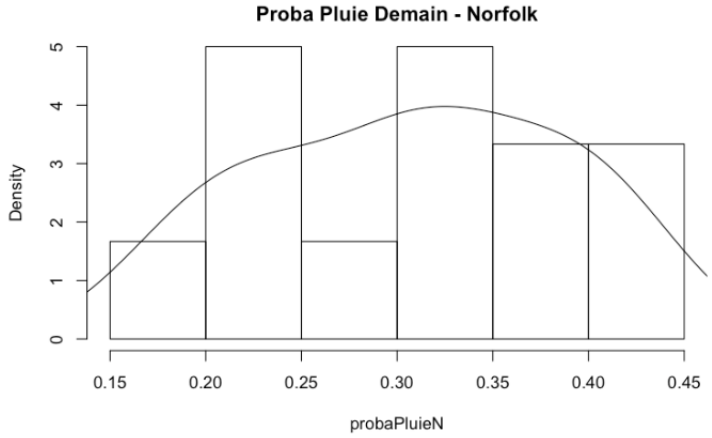

Figure 9: Probabilité qu'il pleuve demain à NorfolkIsland

```
          Été Hiver
Pluie      27   194
Non Pluie 102   146

        Pearson's Chi-squared test with Yates' continuity correction

data:  tabTest
X-squared = 47.55, df = 1, p-value = 5.362e-12
```

Figure 10: Probabilité de pluie en hiver vs probabilité de pluie en été à Ballarat

On a $H_0 : p_1 = p_2$ et $H_1 : p_1 \neq p_2$. La p-Value est très faible, on rejette H_0, la proba de pluie en hiver est différente de la probabilité de pluie en été.
Maintenant, nous allons tenter de prédire la pluie demain à l'aide du régression logistique.

$Y = k|X$ est une VA discrète, car k prend 0 ou 1. On nous donne la fonction suivante :

$$f(X) = P(Y = 1|X) = \frac{e^{\sum_i} B_i X_i}{1 + e^{\sum_i} B_i X_i}$$

On en déduit la fonction de densité pour $Y = k|X$ noté g :

$$g(Y|X) = P(Y = 1|X)^y P(Y = 0|X)^{1-y}$$
$$g(Y|X) = P(Y = 1|X)^y (1 - P(Y = 1|X))^{1-y}$$
$$g(Y|X) = f(X)^y (1 - f(X))^{1-y}$$

La log-vraisemblance est donné par:
$\mathcal{L}(y_1 \dots y_n|B_i) = \sum_i \ln(f(X)^{y_i}(1 - f(X))^{1-y_i}) = \sum_{i=1}^n (y_i(\sum_{j=1}^n B_j X_j) - \ln(1 + e^{\sum_{l=1}^n B_l X_l}))$
l'idée est en faite d'estimer les B_i pour prédire Y avec les X_i qui sont des variables connues.
On effectue la régression avec la commande glm sur R et voici les B_i estimés (cf Figure 11).

```
Call:  glm(formula = dataAlbury$RainTomorrow ~ dataAlbury[, 3] + dataAlbury[,
    4] + dataAlbury[, 7] + dataAlbury[, 10] + dataAlbury[, 11] +
    dataAlbury[, 12] + dataAlbury[, 13] + dataAlbury[, 14] +
    dataAlbury[, 15] + dataAlbury[, 16] + dataAlbury[, 17], family = "binomial",
    data = dataAlbury)

Coefficients:
     (Intercept)   dataAlbury[, 3]   dataAlbury[, 4]   dataAlbury[, 7]  dataAlbury[, 10]  dataAlbury[, 11]
       87.495523          0.121990         -0.134381          0.055246         -0.002787         -0.007478
dataAlbury[, 12]  dataAlbury[, 13]  dataAlbury[, 14]  dataAlbury[, 15]  dataAlbury[, 16]  dataAlbury[, 17]
       -0.004519          0.086402          0.433297         -0.527152          0.036482          0.019868

Degrees of Freedom: 3010 Total (i.e. Null);  2999 Residual
Null Deviance:      3057
Residual Deviance: 1720         AIC: 1744

Call:  glm(formula = dataBallarat$RainTomorrow ~ dataBallarat[, 3] +
    dataBallarat[, 4] + dataBallarat[, 7] + dataBallarat[, 10] +
    dataBallarat[, 11] + dataBallarat[, 12] + dataBallarat[,
    13] + dataBallarat[, 14] + dataBallarat[, 15] + dataBallarat[,
    16] + dataBallarat[, 17], family = "binomial", data = dataBallarat)

Coefficients:
       (Intercept)   dataBallarat[, 3]   dataBallarat[, 4]   dataBallarat[, 7]  dataBallarat[, 10]
        109.401185           0.013973           -0.057543            0.055215           -0.024986
dataBallarat[, 11]  dataBallarat[, 12]  dataBallarat[, 13]  dataBallarat[, 14]  dataBallarat[, 15]
         -0.008068           -0.009782            0.066986            0.168822           -0.282904
dataBallarat[, 16]  dataBallarat[, 17]
          0.101863           -0.030281

Degrees of Freedom: 3027 Total (i.e. Null);  3016 Residual
Null Deviance:      3457
Residual Deviance: 2147         AIC: 2171
```

Figure 11: Régression linéaire pour Ballarat et Albury

Malheureusement, j'ai un problème de données. Le Lundi mon programme fonctionnait avec mes données et après avoir relancé R, une erreur persiste au moment d'exécuter la commande glm :"type (list) incorrect pour la variable 'dataAlbury[, 3]'".

Je prevoyais d'utiliser cette fonction pour chaque mois pour rejoindre l'idée qu'on avait précédement d'affiner l'étude...

Conclusion

Pour conclure ce projet et cette UE, on comprend l'intérêt et le potentiel des statistiques grâce aux études menées lors des TP et de ce projet, les statistiques permettent de visualiser, comprendre et prédire les données. Il faut cependant parfois faire attention à l'interprétation à avoir envers ce que nous montre les données, les test sont alors de bon "garde-fous".
Pour conclure ce projet, on a mené une étude complète et assez naturelle au niveau de la démarche scientifique : on découvre les données pour allez de plus en plus en profondeur. Cependant je regrette ne pas avoir atteint l'objectif qui était de prédire la pluie et je pense ne pas être très loin de réussir, j'ai malheureusement manqué de temps...