Rapport d'étude

Modèle linéaire généralisé et choix de modèles

EMSBD6 - Bruno KUBECZKA

9 Juillet 2023

Abstract

Cette étude a pour objet la mise en pratique de la régression logistique dans le cadre de la prédiction du fait qu'il pleuvra ou non le lendemain.

Contents

1 Démarche de l'étude

1.1 Méthode de sélection de modèles

Nous allons aborder le projet selon une **méthode de sélection de modèles de type Hold-Out**.

Nous avons è disposition 2 jeu de données :

- un jeu de données **meteo.train** de **1180 observations** pour lequel la variable la varibale d'intérêt pluie.demain est donnée
- un jeu de données **meteo.test** de **290 observations** pour lequel la variable la variable d'intérêt pluie.demain n'est donnée

Considérons l'hypothèse que les données des 2 jeux sont issus d'un même jeu de données sur lequel un tirage aléatoire 75%/25% a été réalisé. Etant donnés les index contenus dans les jeu de données fournis, l'hypothèse est raisonnable.

Nous pouvons alors aborder l'étude selon une **méthode de sélection de modèles de type Hold-Out**, à savoir:

- une 1ère **phase d'entraînement et de validation** permettra d'identifier le meilleur modèle dans le cadre d'une prédiction

  Pour la mise en oeuvre, le jeu **meteo.train** dont on connaît les valeurs de la variable d'intérêt **pluie.demain** va être éclaté en 2 jeux de données distincts et tirés au hasard

  - Un **jeu de données d'entraînement des modèles** ; on choisit de prendre 80% des données de **meteo.train**.
  - Un **jeu de données de validation des modèles** ; on choisit de prendre 20% des données du eu train

  A l'issue de cette phase, les critères de sélection (cf. ci-dessous) permettront de conclure sur le modèle le plus à même de prédire les valeurs du jeu **meteo.test**

- Une **phase de validation** basée sur le jeu de données **meteo.test** permettra de tester le "meilleur" modèle entraîné et validé.

  Dans ce projet, cette phase se limitera à la prédiction des valeurs binaires ***pluie.demain***, le résultat de test faisant l'objet de l'évaluation du projet.

1.2 Démarche

Une **Analyse exploratoire** permettra d'étudier les corrélations

- entre les covariables et la variable d'intérêt ***pluie.demain***
- entre les covariables elles-mêmes

L'analyse des covariables se fera par famille (températures, nébulosité, vents).

A partir de l'analyse, il pourra être possible d'identifier les trop fortes corrélations au sein d'une même famille, puis entre variables des différentes familles.

L'**identification des modèles candidats** adoptera plusieurs stratégies :

- une **1ère stratégie** consistera à utiliser toutes les covariables d'origine sans restriction aucune (on ne considère pas les colinéarités identifiées) dans un modèle "complet"
- une **2ème stratégie** consistera à sélectionner les covariables en se basant sur l'analyse exploratoire ; on s'autorise à introduire dans ce modèle des covariables transformées comme l'amplitude

A noter que la sélection step sera appliqué sur le modèle "complet" de différente façon

- **méthode descendante**
- **méthode progressive depuis le modèle complet** (on retire les covariables, avec la possibilité d'en ajouter une déjà retirée précédemment)
- **Méthode ascendante depuis un modèle constant** vers le modèle complet
- **Méthode progressive ascendante depuis un modèle constant** : on ajoute des covariables avec la possibilité à chaque itérations d'en retirer une qui a été ajoutée précédemment)

1.2.1 Identification des modèles candidats

Dans la phase

1.2.2 Critère de sélection du "meilleur" modèle

BKU : justifier le choix de AIC

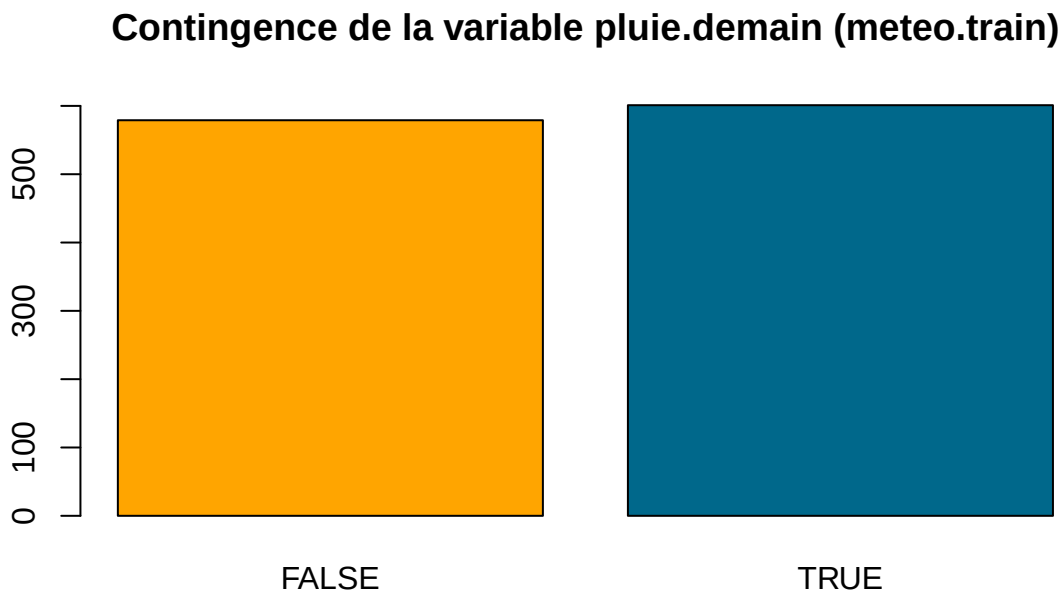
2 Préparation des données

BKU : renommage des variables

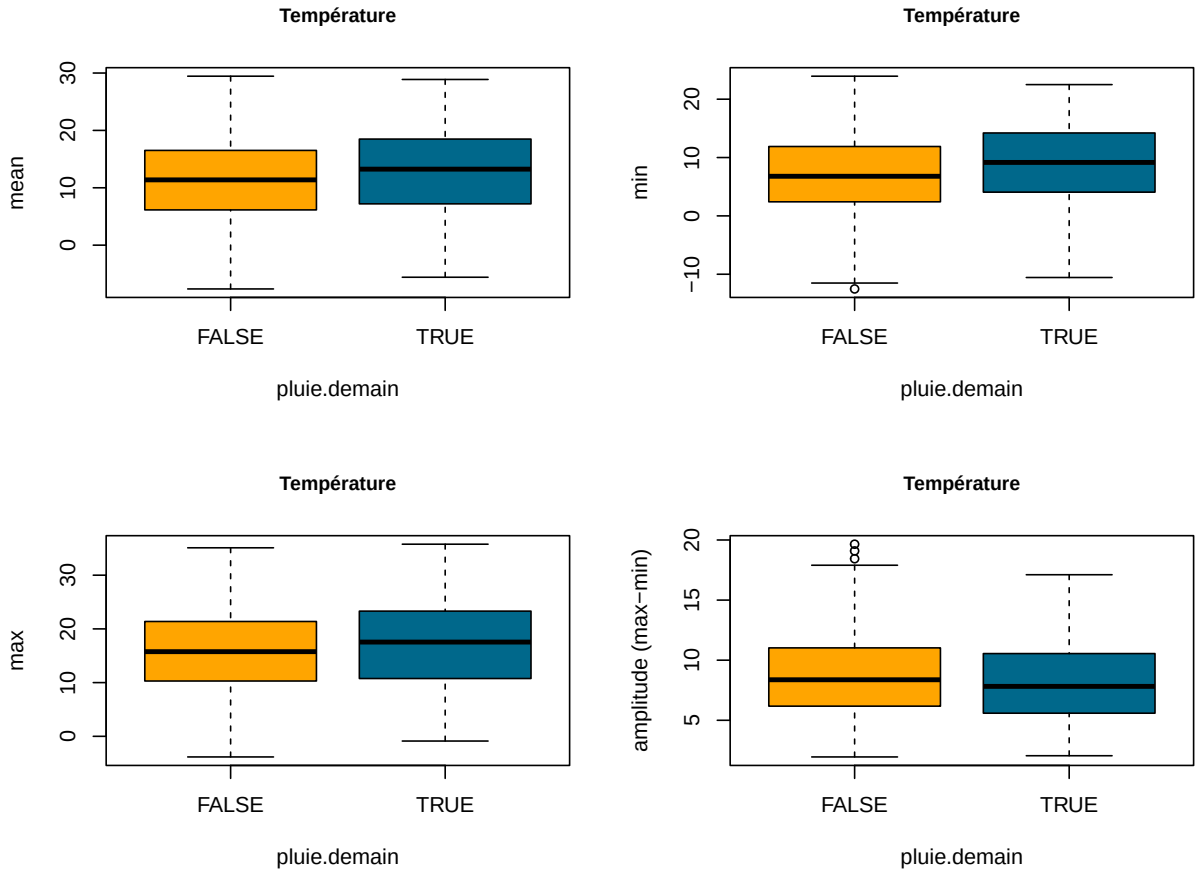
BKU : Ajout de variables amplitudes

3 Analyse exploratoire

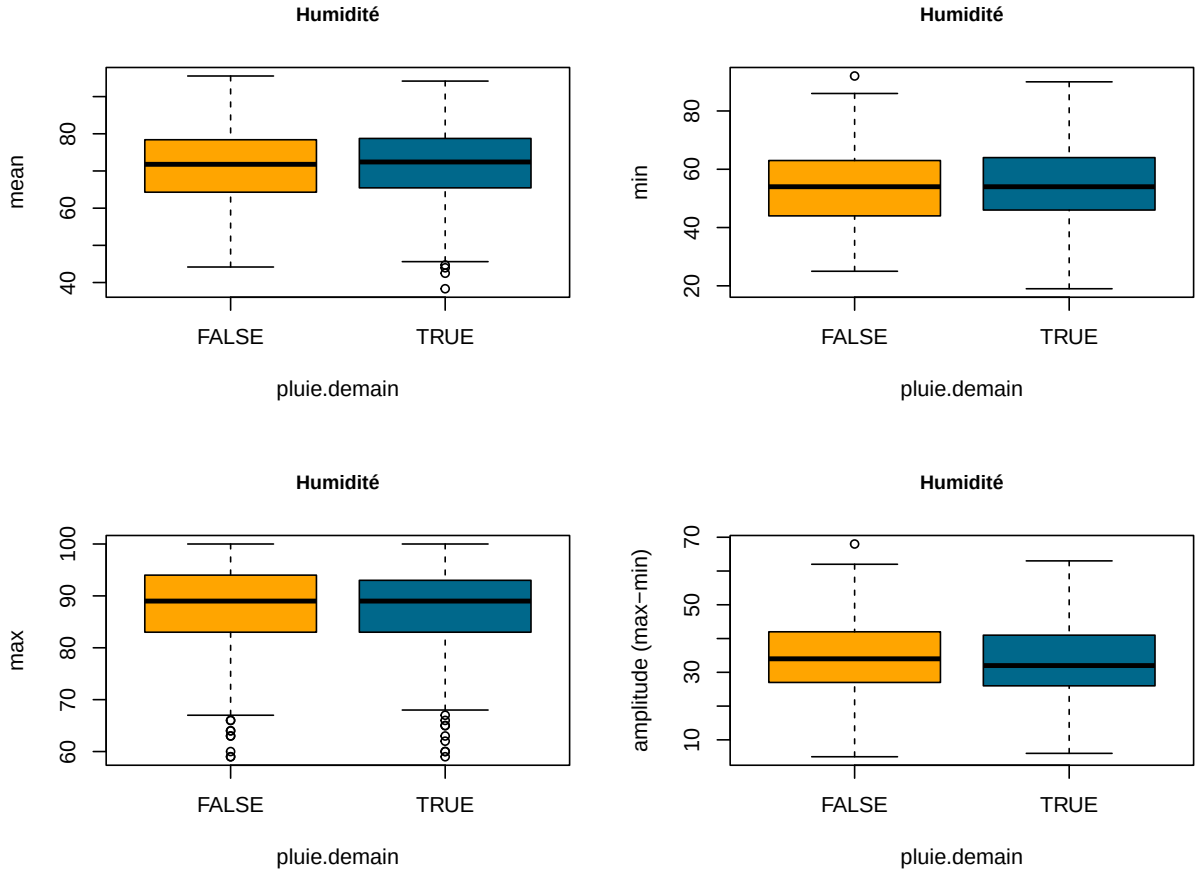
3.1



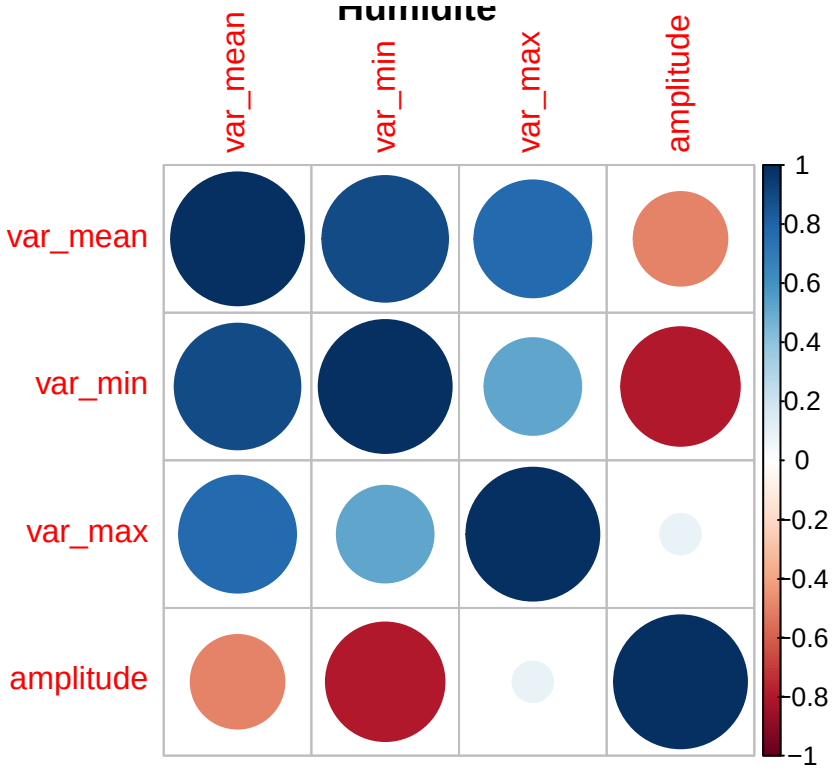
3.2 Température



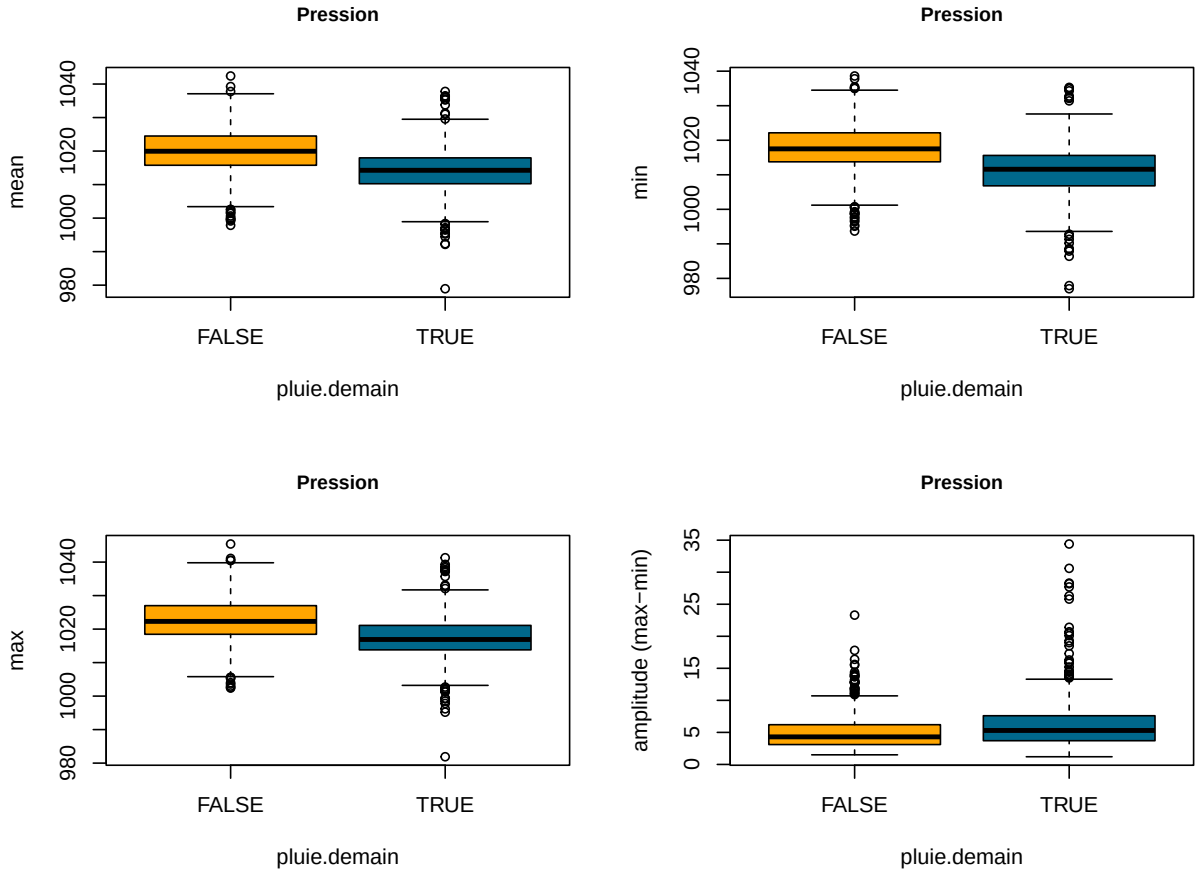
3.3 Humidité relative



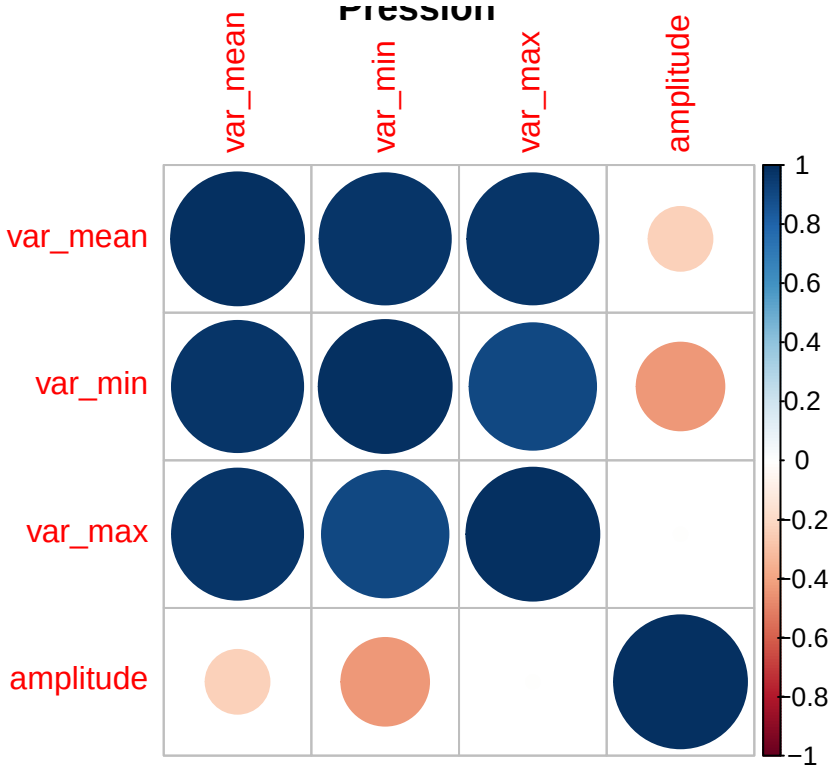

```
##               var_mean    var_min    var_max   amplitude
## var_mean    1.0000000  0.8939695 0.77390562 -0.49621785
## var_min     0.8939695  1.0000000 0.52914005 -0.79544270
## var_max     0.7739056  0.5291400 1.00000000  0.09333577
## amplitude -0.4962178 -0.7954427 0.09333577  1.00000000
```


Humidite
var_mean
var_min
var_max
amplitude
var_mean
var_min
var_max
amplitude
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

3.4 Pression




```
##               var_mean    var_min      var_max    amplitude
## var_mean    1.0000000  0.9735478  0.972200761 -0.230220060
## var_min     0.9735478  1.0000000  0.904746052 -0.434919125
## var_max     0.9722008  0.9047461  1.000000000 -0.009935123
## amplitude -0.2302201 -0.4349191 -0.009935123  1.000000000
```

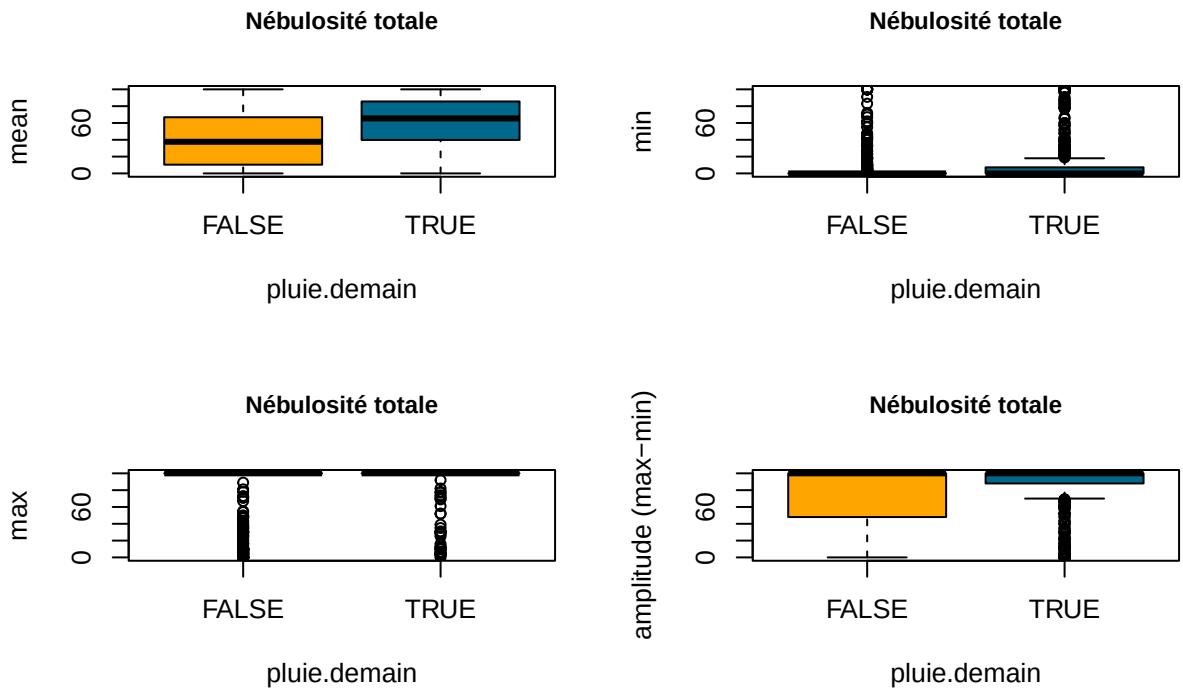


pluie.demain==TRUE : une légère tendance à une humidité moyenne/min/max plus grande pluie.demain==FALSE : une légère tendance à une humidité moyenne/min/max plus grande

=> Les corrélations entre min/max/mean sont positives et relativement fortes (>0.9) => Les corrélations entre min/max/meanet amplitude sont négativeset relativement faibles => La corrélation la plus faible est constatée entre amplitude et max

Idées pour la modélisation . Inclure un unique représentant parmi moyenne/min/max fortement corrélées : max en l'occurrence . inclure l'amplitude . Considérer la covariable produit amplitude*max

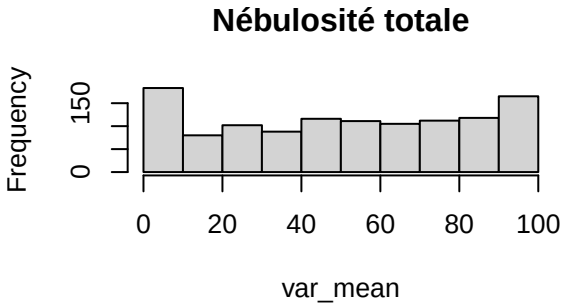
3.5 Nébulosité totale

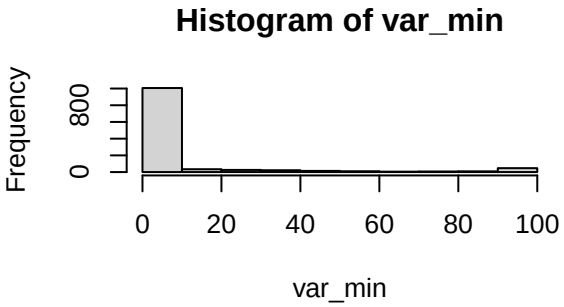


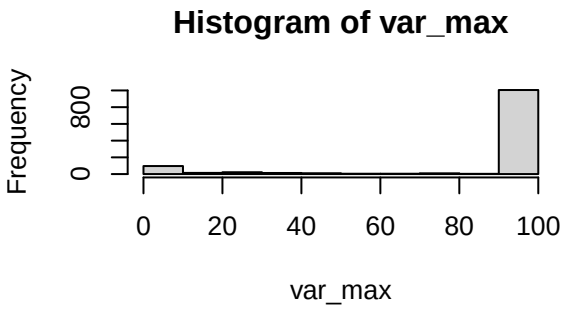

```
##       0%      25%      50%      75%     100%
##   0.0000  23.8050  51.6700  78.5325 100.0000
```

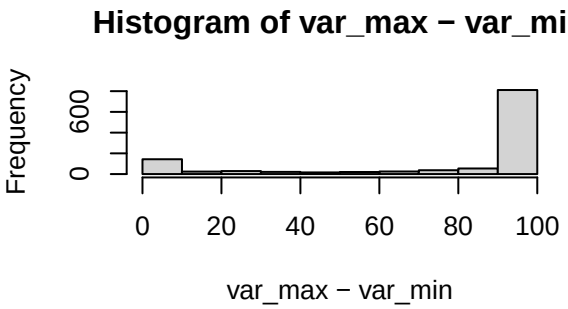
```
##    0%   25%   50%   75%  100%
##   0.0   0.0   0.0   2.4 100.0
```

```
##   0%  25%  50%  75% 100%
##    0  100  100  100  100
```



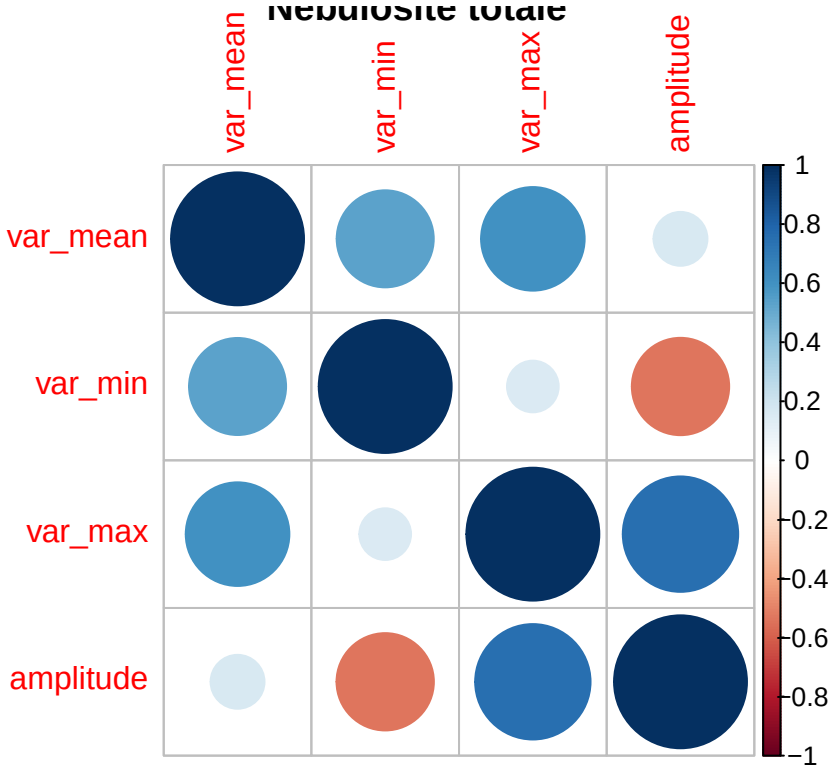







```
##      0%     25%     50%     75%    100%
##   0.000  75.175 100.000 100.000 100.000
```

```
##            var_mean    var_min   var_max  amplitude
## var_mean  1.0000000  0.5329495 0.6074140  0.1642736
## var_min   0.5329495  1.0000000 0.1509833 -0.5360951
## var_max   0.6074140  0.1509833 1.0000000  0.7535390
## amplitude 0.1642736 -0.5360951 0.7535390  1.0000000
```

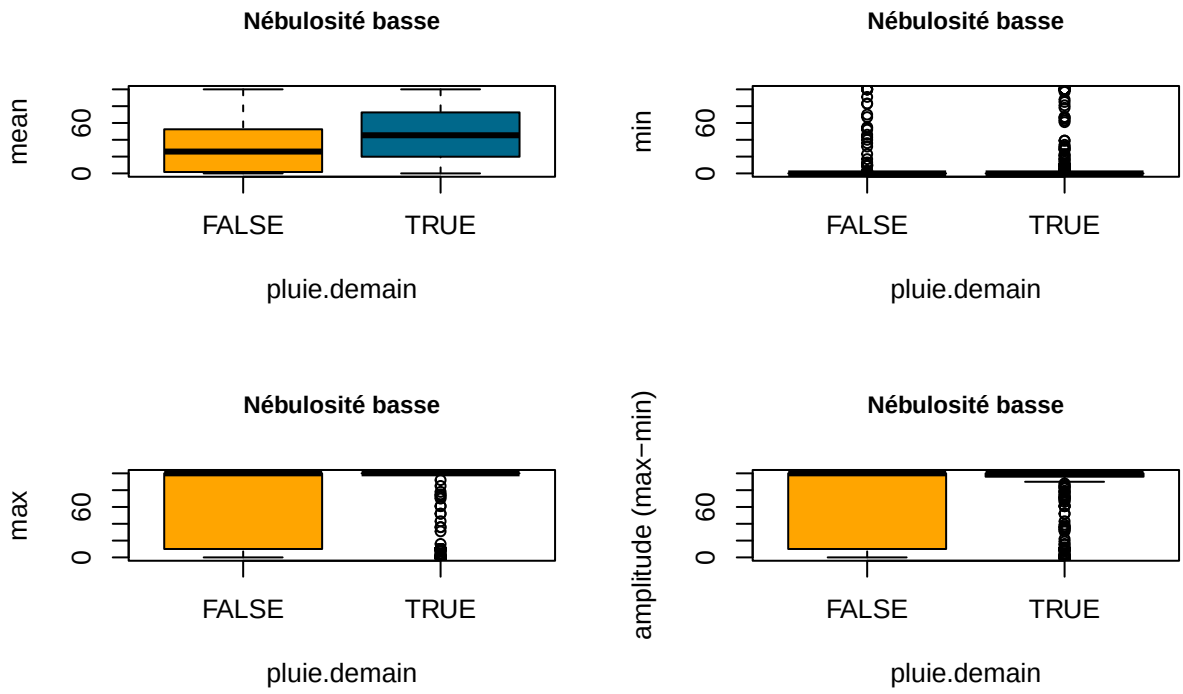


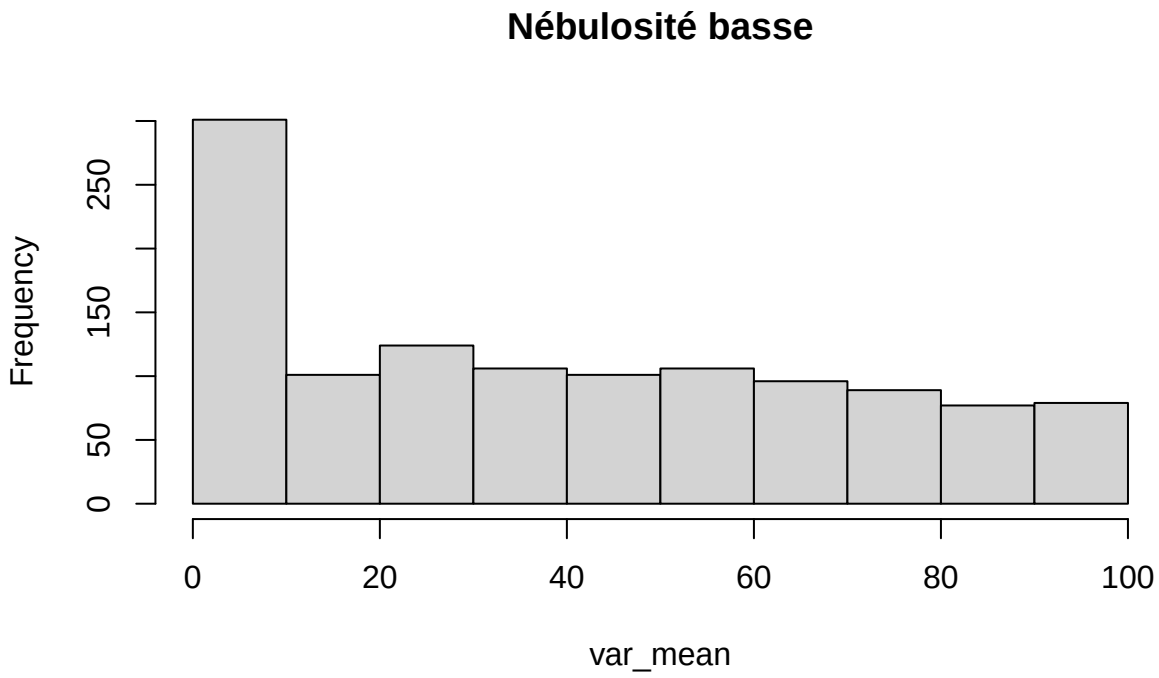
NEBULOSITE TOTALE

. mean/min/max sont relativement corrélés (50/60%) . amplitude est moyennement corrélé avec min/max (50/75%) . amplitude est peu corrélé avec mean (0.16) . amplitude présente les caractéristiques suivantes . amplitude est "bipolarisée" : soit 0% (min et max sont les même valeurs) soit 100% (min et max sont 0/100) . amplitude=100% => pluie.demain==TRUE . amplitude=0% => pluie.demain==FALSE

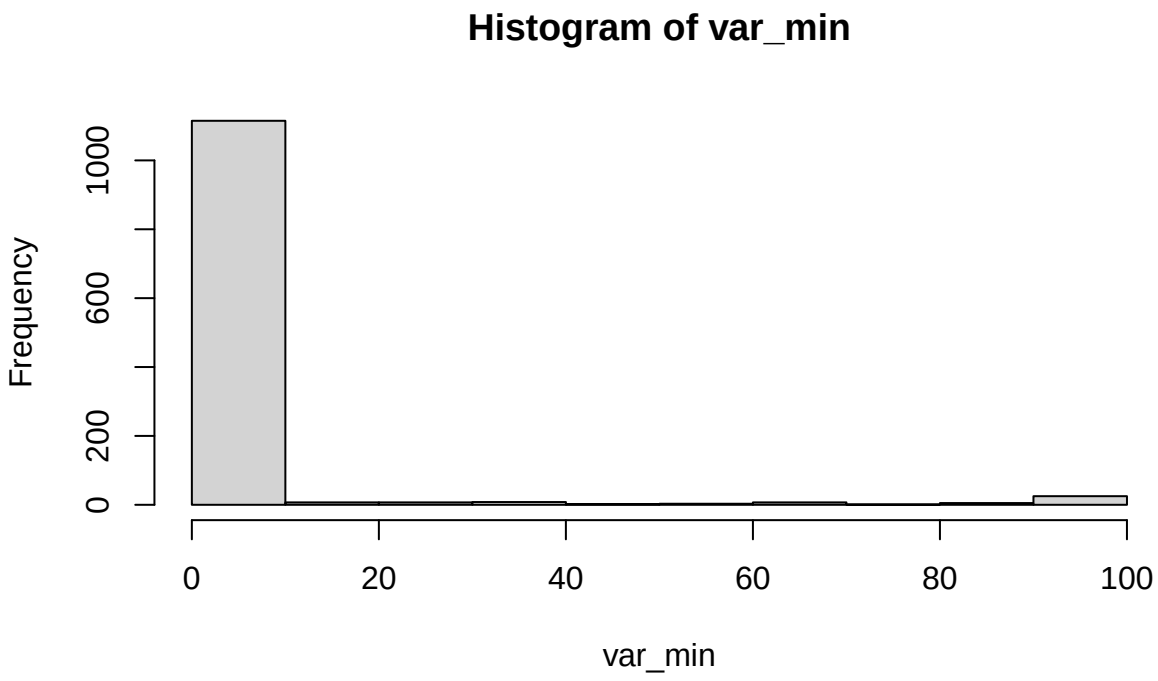
Idées pour la modélisation . inclure mean et amplitude sous forme booleenne . inclure le produit mean*amplitude

3.6 Nébulosité basse

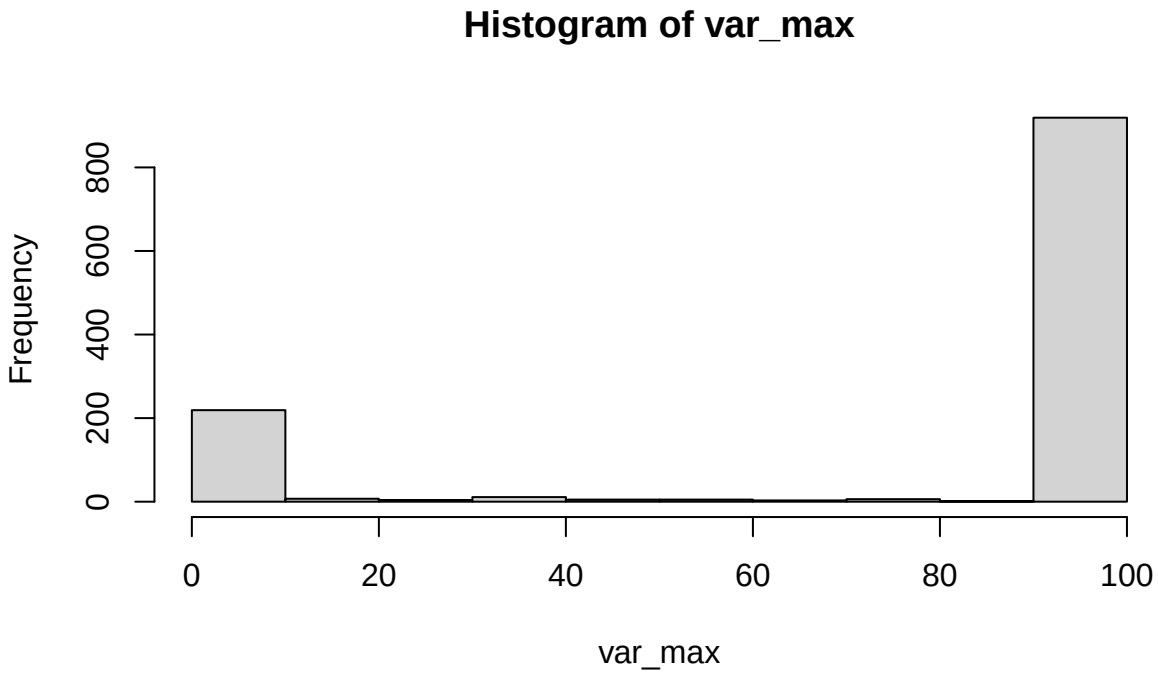





```
##      0%     25%     50%     75%    100%
##   0.000   9.420  36.355  65.760 100.000
```

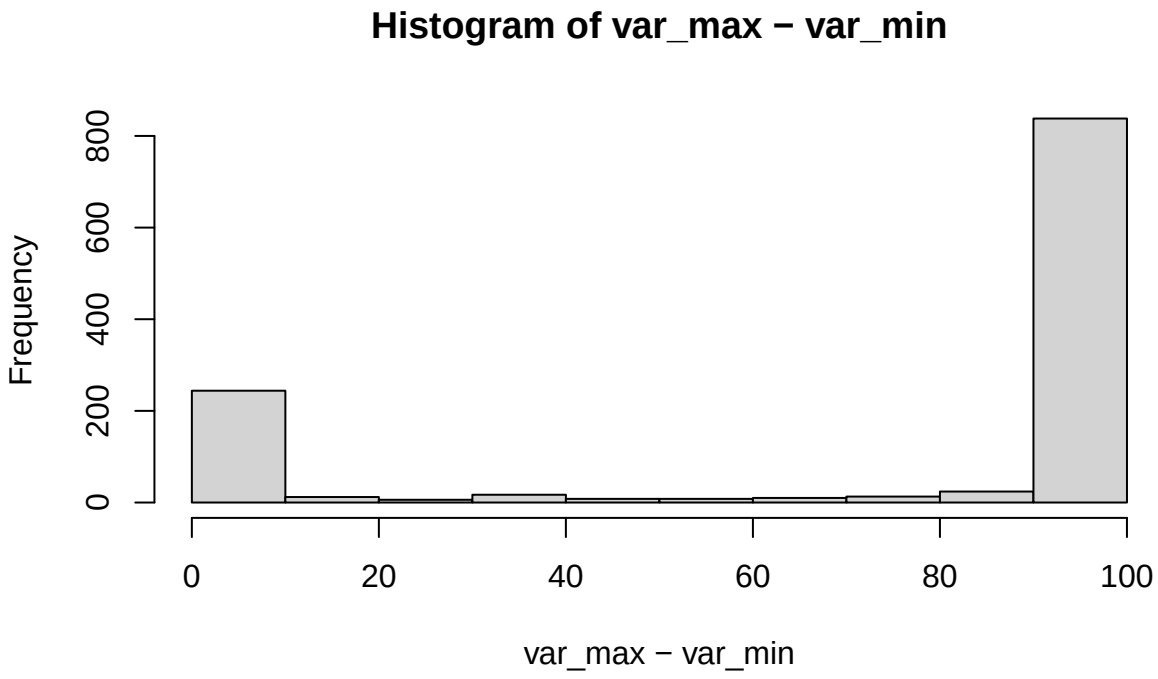



```
##    0%   25%   50%   75%  100%
##     0     0     0     0   100
```



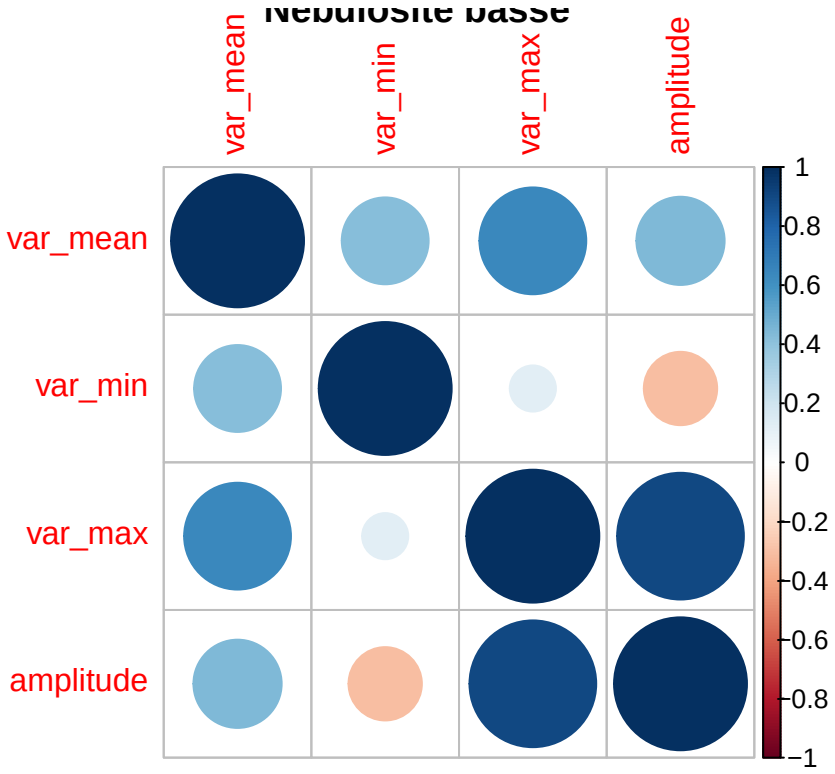

```
##    0%   25%   50%   75%  100%
```

```
##    0  100  100  100  100
```




```
##     0%    25%    50%    75%   100%
##   0.00  60.75 100.00 100.00 100.00
```

```
##            var_mean    var_min   var_max  amplitude
## var_mean  1.0000000  0.4273706 0.6460351  0.4405601
## var_min   0.4273706  1.0000000 0.1203414 -0.3042092
## var_max   0.6460351  0.1203414 1.0000000  0.9090733
## amplitude 0.4405601 -0.3042092 0.9090733  1.0000000
```

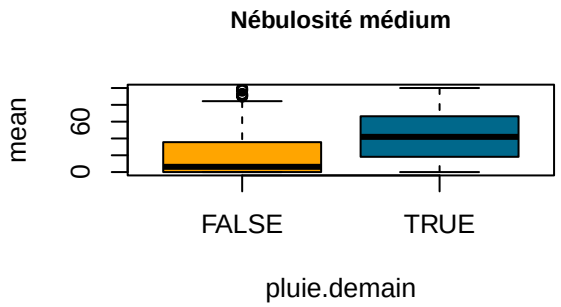


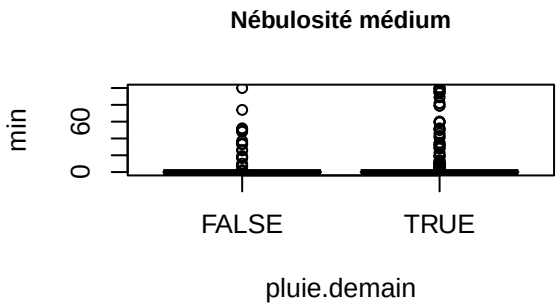
NEBULOSITE BASSE

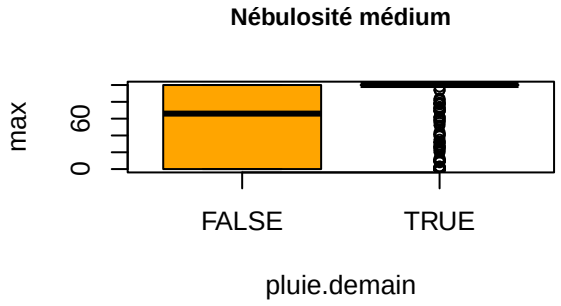
. min/max/mean sont plutôt faiblement corrélées . amplitude présente les caractéristiques suivantes . amplitude est “bipolarisée” : soit 0% (min et max sont les même valeurs) soit 100% (min et max sont 0/100) . amplitude=100% => pluie.demain==TRUE . amplitude=0% => pluie.demain==FALSE

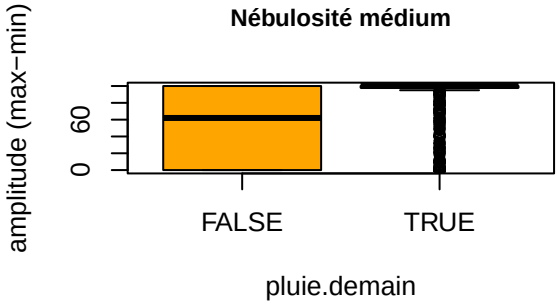
Idées pour la modélisation . inclure mean et amplitude sous forme booleenne . inclure le produit min*amplitude

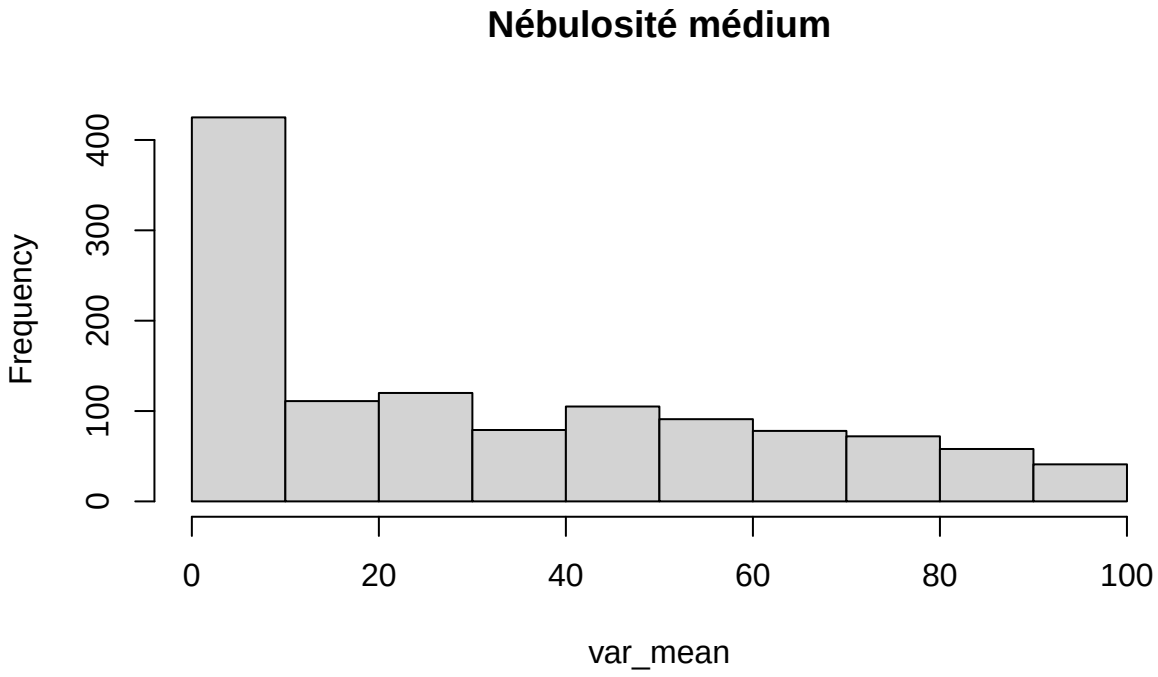
Nébulosité medium



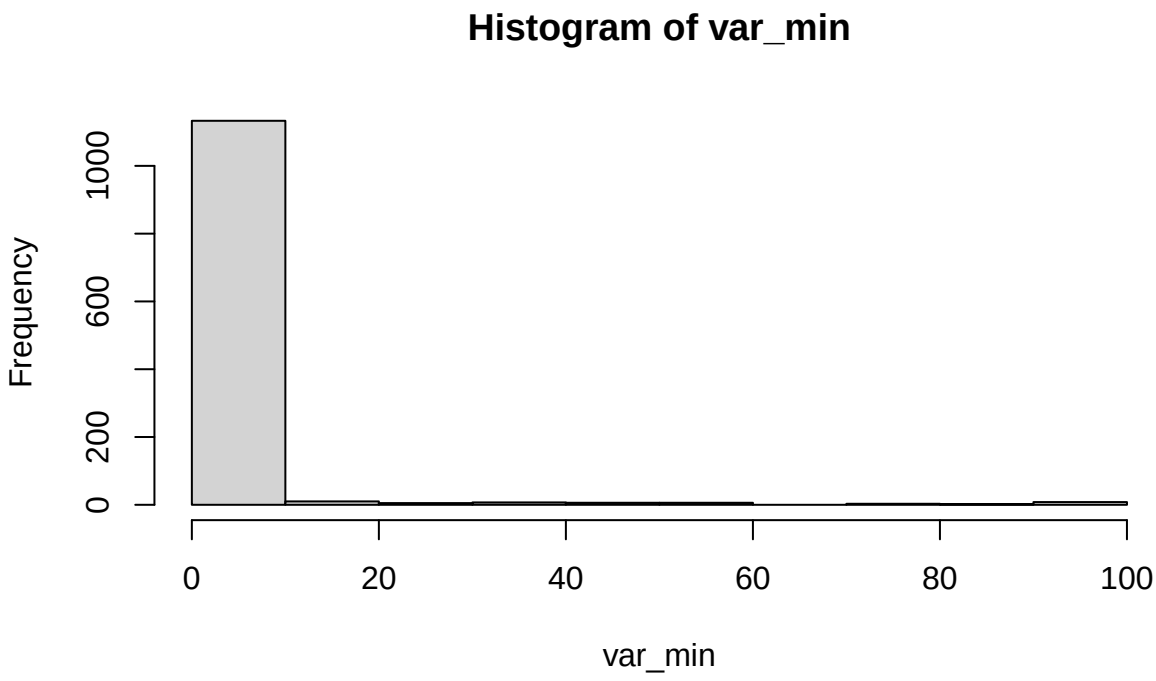




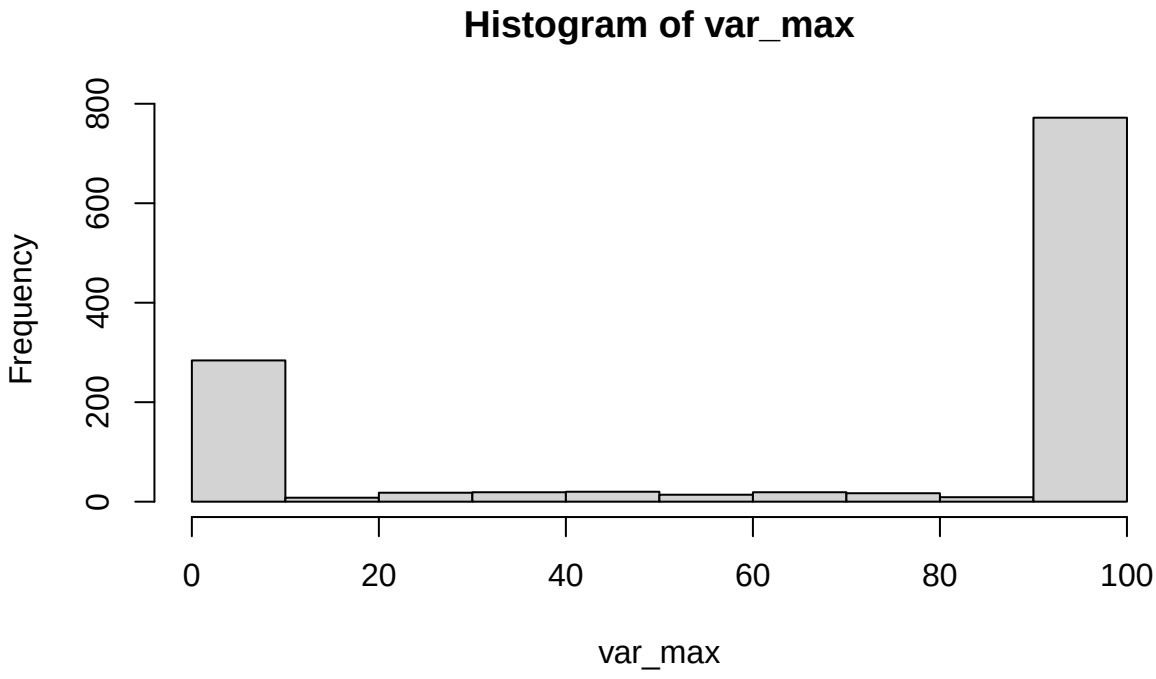





```
##      0%     25%     50%     75%    100%
##    0.00    1.83   24.98   54.21  100.00
```

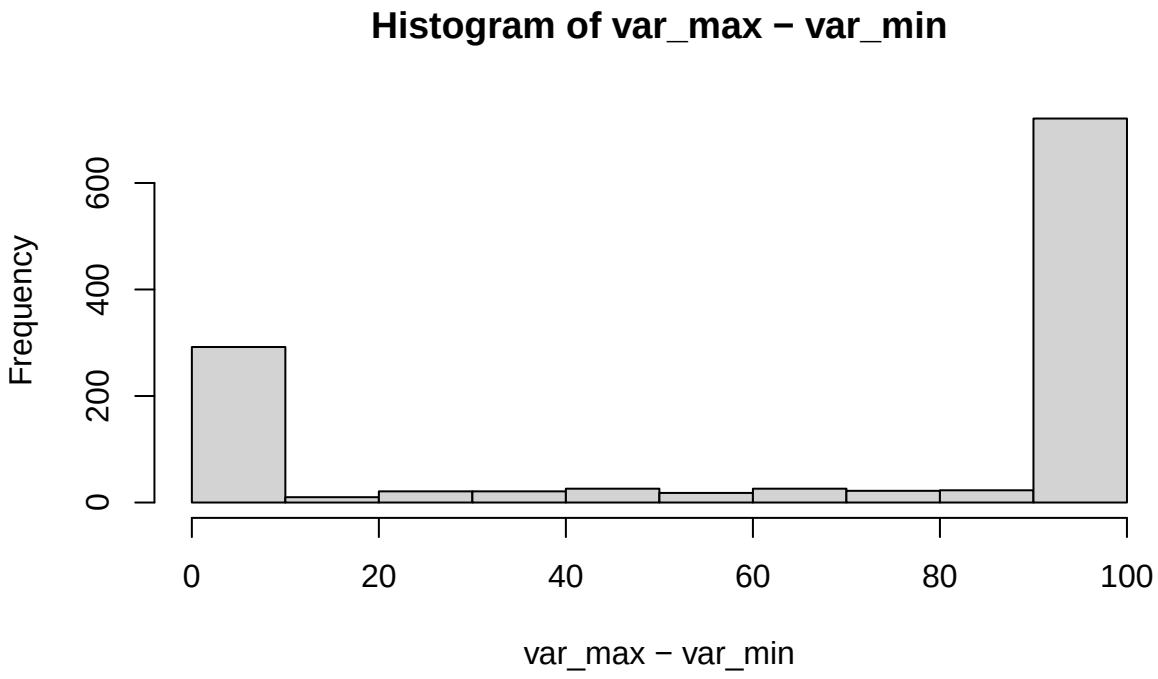



```
##   0%  25%  50%  75% 100%
##    0    0    0    0  100
```



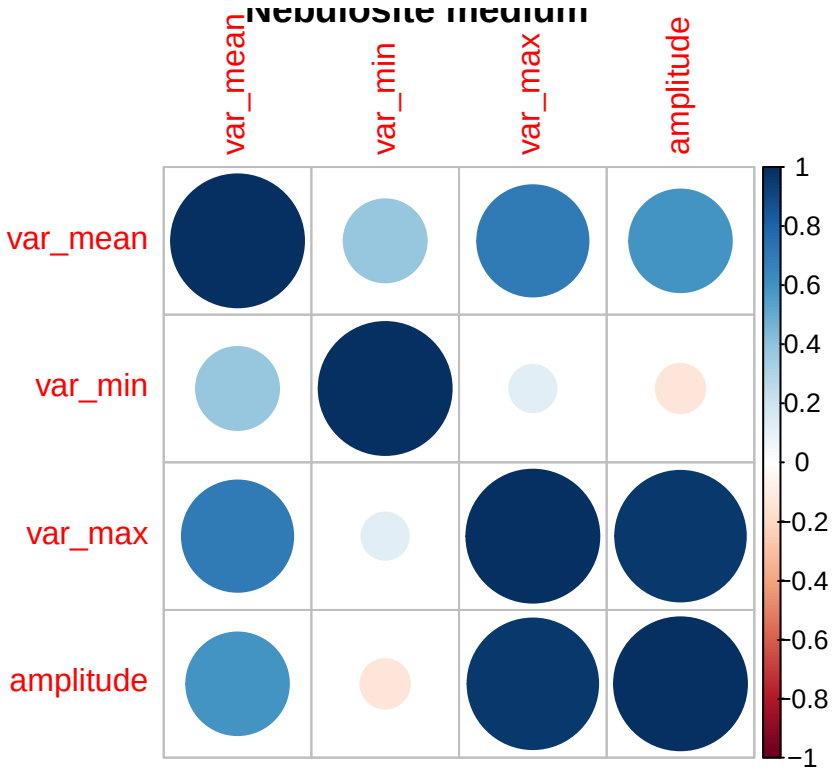

```
##   0%  25%  50%  75% 100%
```

```
##   0.00  22.75 100.00 100.00 100.00
```




```
##   0%  25%  50%  75% 100%
##    0   13  100  100  100
```

```
##            var_mean    var_min   var_max  amplitude
## var_mean  1.0000000  0.3896551 0.7000574  0.5957026
## var_min   0.3896551  1.0000000 0.1265068 -0.1385866
## var_max   0.7000574  0.1265068 1.0000000  0.9648614
## amplitude 0.5957026 -0.1385866 0.9648614  1.0000000
```

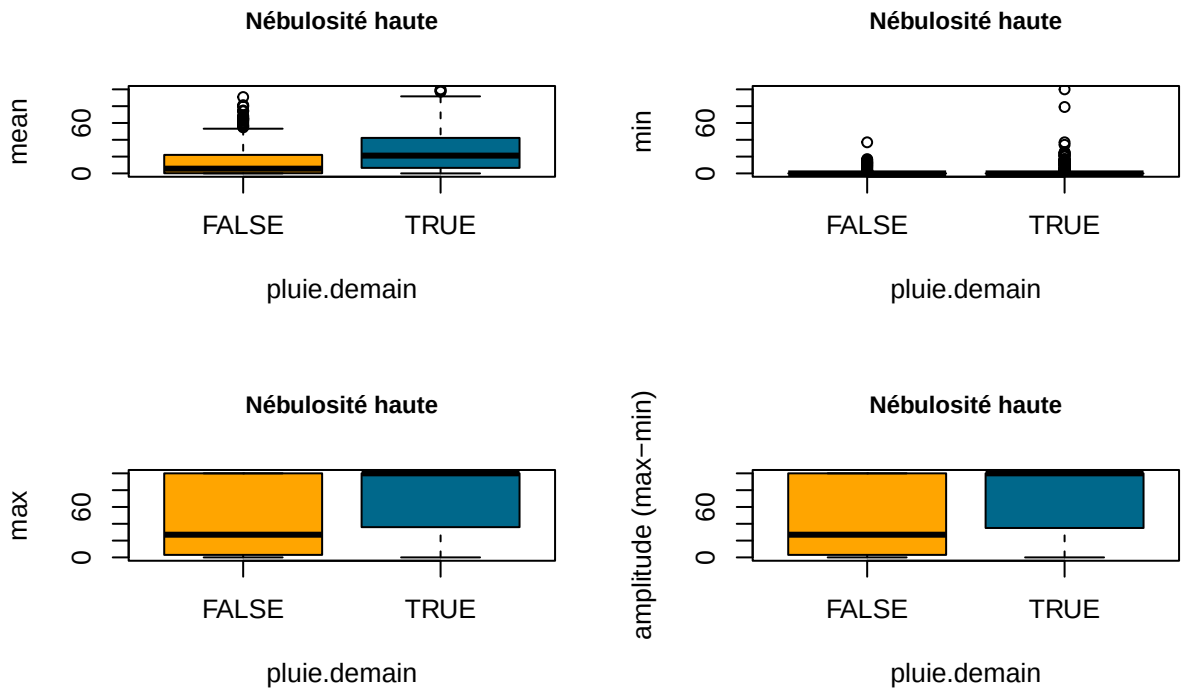


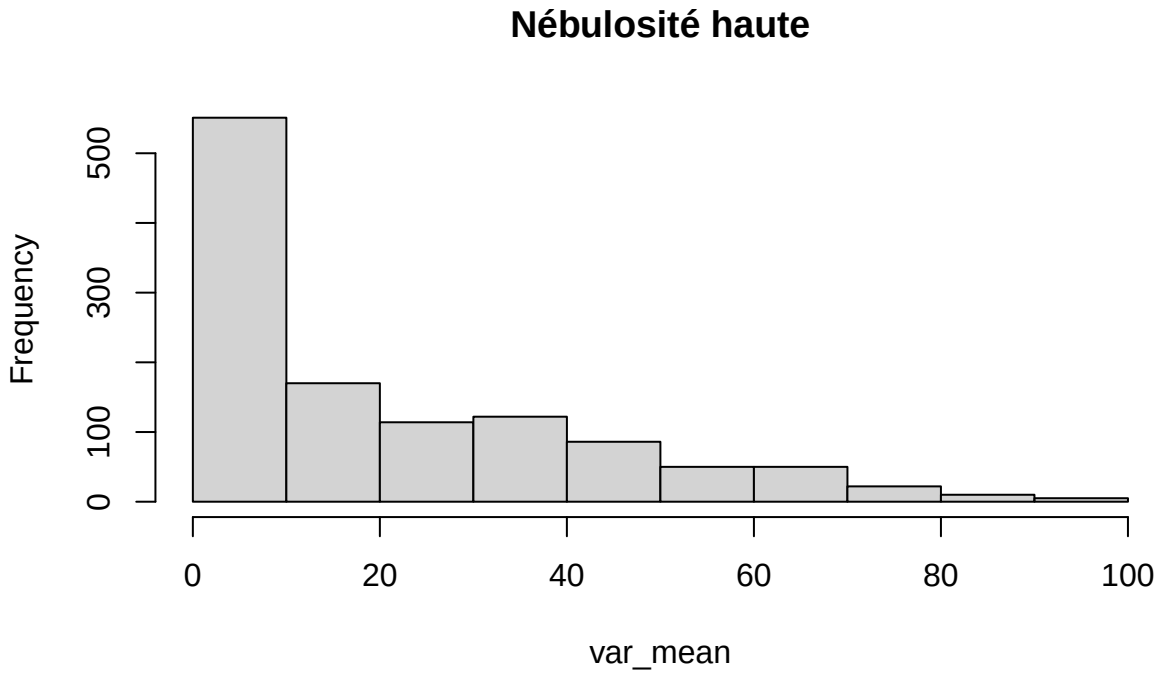
NEBULOSITE MEDIUM

. mean semble lié positivement avec pluie.demain . max/amplitude semblent liés positivement avec pluie.demain . amplitude est “bipolarisée” : soit 0% (min et max sont les même valeurs) soit 100% (min et max sont 0/100) . amplitude=100% => pluie.demain==TRUE . amplitude=0% => pluie.demain==FALSE

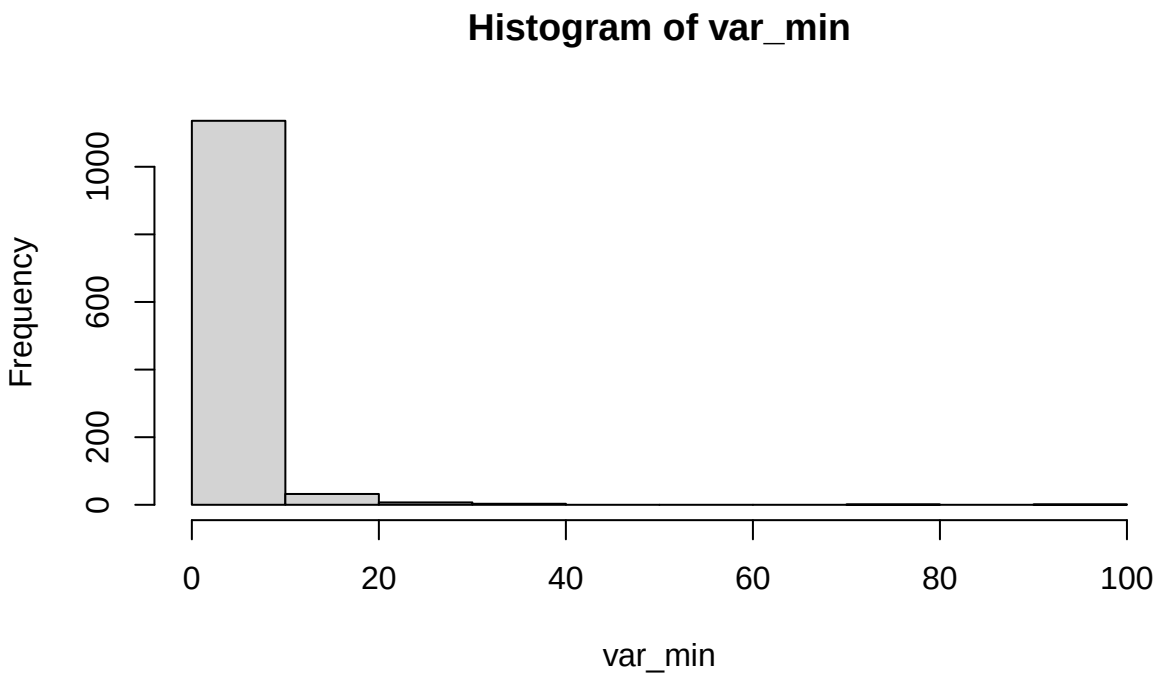
Idées pour la modélisation . inclure mean et amplitude sous forme booleenne . inclure le produit mean*amplitude

3.7 Nébulosité haute

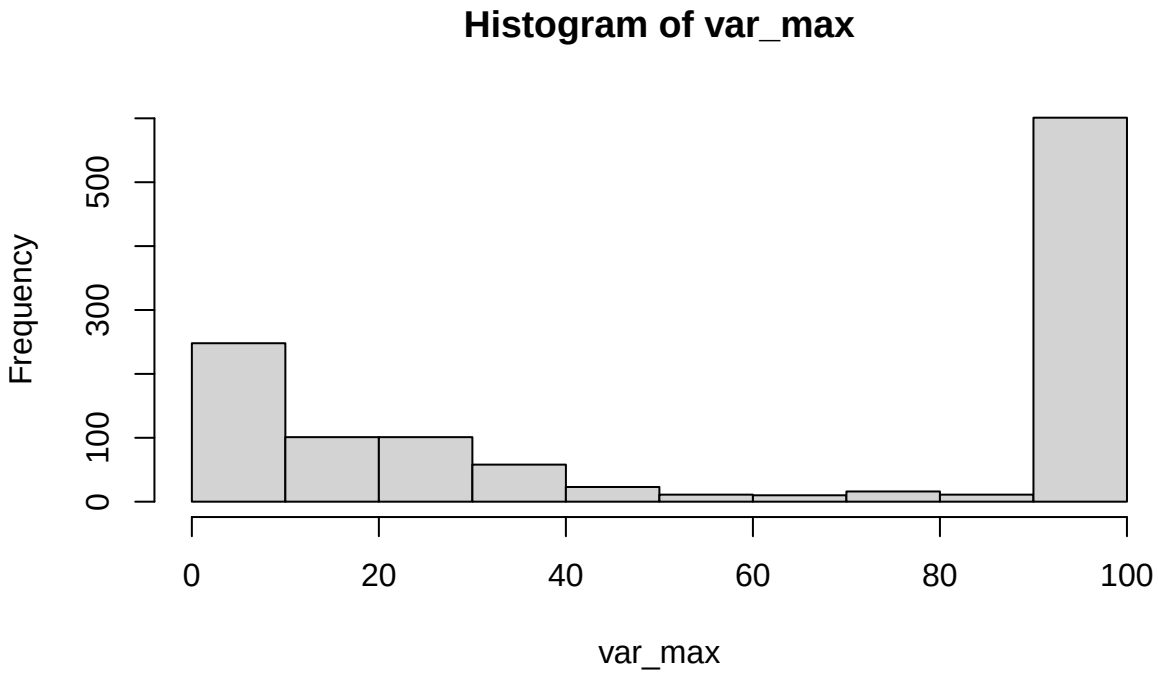





```
##       0%      25%      50%      75%      100%
##   0.0000   1.6575  11.8800  33.2600 100.0000
```

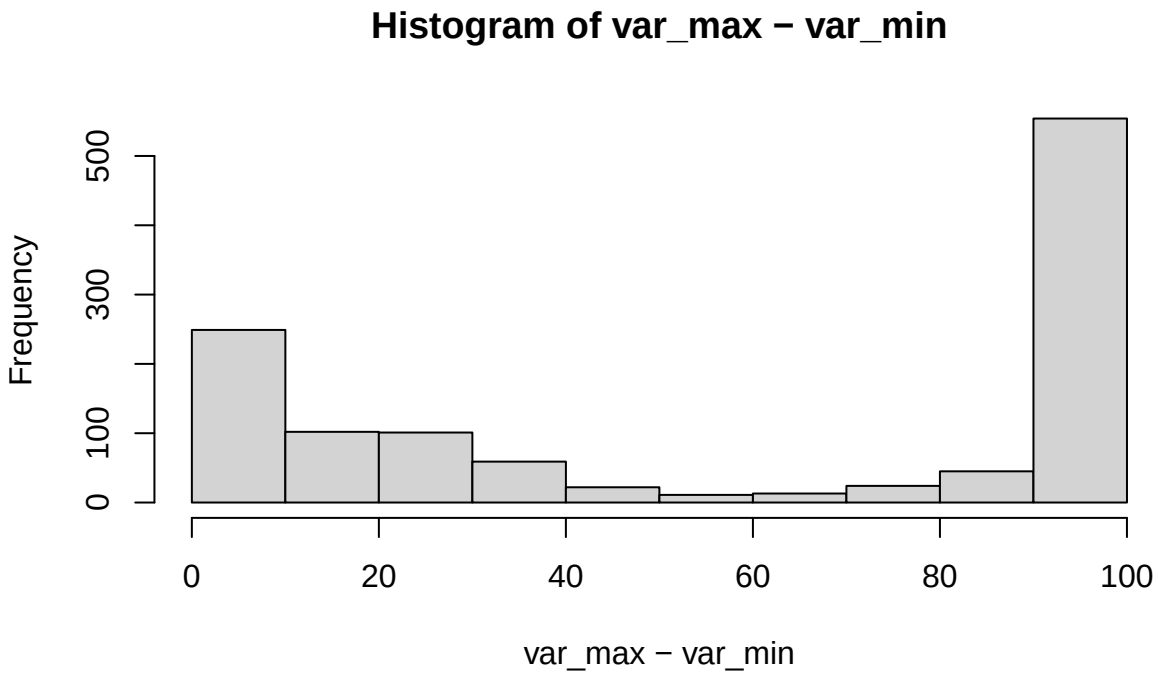



```
##    0%   25%   50%   75%  100%
##     0     0     0     0   100
```

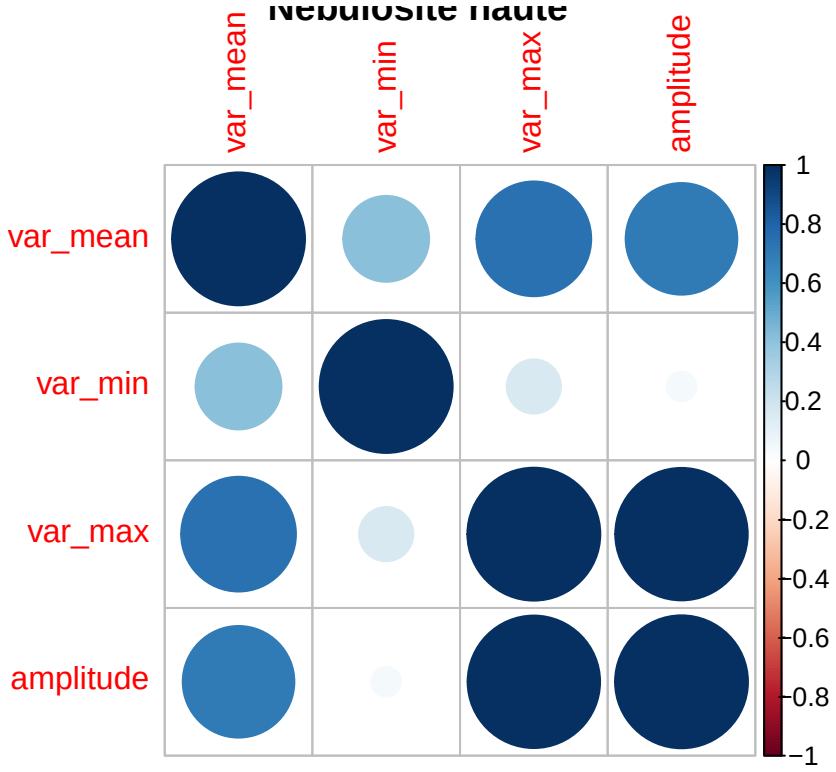



```
##    0%   25%   50%   75%  100%
```

```
##    0   15   97  100  100
```




```
##   0%  25%  50%  75% 100%
##    0   15   84  100  100

##   0%  25%  50%  75% 100%
##    0   35  100  100  100

##   0%  25%  50%  75% 100%
##    0    3   27  100  100

##            var_mean    var_min   var_max  amplitude
## var_mean  1.0000000 0.41745207 0.7453130 0.70505473
## var_min   0.4174521 1.00000000 0.1665816 0.04919213
## var_max   0.7453130 0.16658163 1.0000000 0.99302842
## amplitude 0.7050547 0.04919213 0.9930284 1.00000000
```



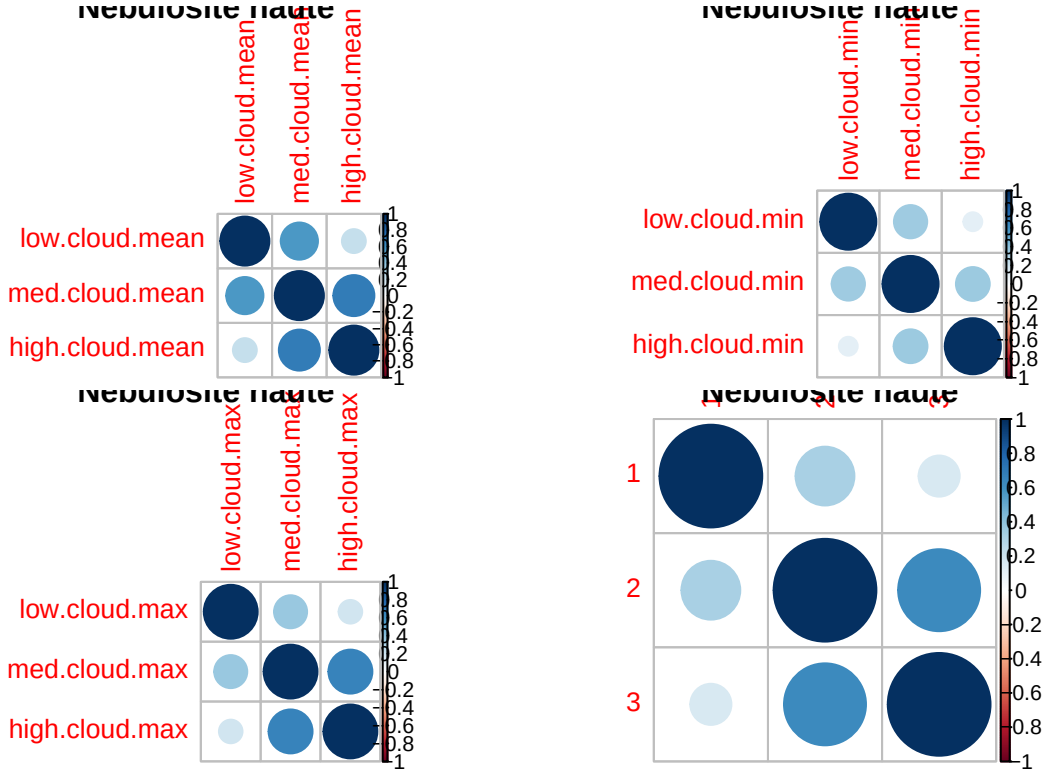
NEBULOSITE HAUTE

. mean semble lié positivement avec pluie.demain . max/amplitude semblent liés positivement avec pluie.demain . amplitude est "bipolarisée" : soit 0% (min et max sont les même valeurs) soit 100% (min et max sont 0/100) . amplitude=100% => pluie.demain==TRUE . amplitude=0% => pluie.demain==FALSE

Idées pour la modélisation . inclure mean et amplitude sous forme booleenne . inclure le produit mean*amplitude

3.8 Corrélation entre nébulosité

```
##                 low.cloud.mean med.cloud.mean high.cloud.mean
## low.cloud.mean       1.0000000      0.5737378       0.2364875
## med.cloud.mean       0.5737378      1.0000000       0.6981242
## high.cloud.mean      0.2364875      0.6981242       1.0000000

##                low.cloud.min med.cloud.min high.cloud.min
## low.cloud.min      1.0000000     0.3522855      0.1127867
## med.cloud.min      0.3522855     1.0000000      0.3636318
## high.cloud.min     0.1127867     0.3636318      1.0000000

##                low.cloud.max med.cloud.max high.cloud.max
## low.cloud.max      1.0000000     0.3759214      0.1913910
## med.cloud.max      0.3759214     1.0000000      0.6654278
## high.cloud.max     0.1913910     0.6654278      1.0000000

##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.3277003 0.1600297
## [2,] 0.3277003 1.0000000 0.6364111
## [3,] 0.1600297 0.6364111 1.0000000
```



CORRELATION entre NEBULOSITES

entre nébulosités basse et haute : corrélations positives et faibles entre mean/min/max/amplitude (~0.1/0.2)

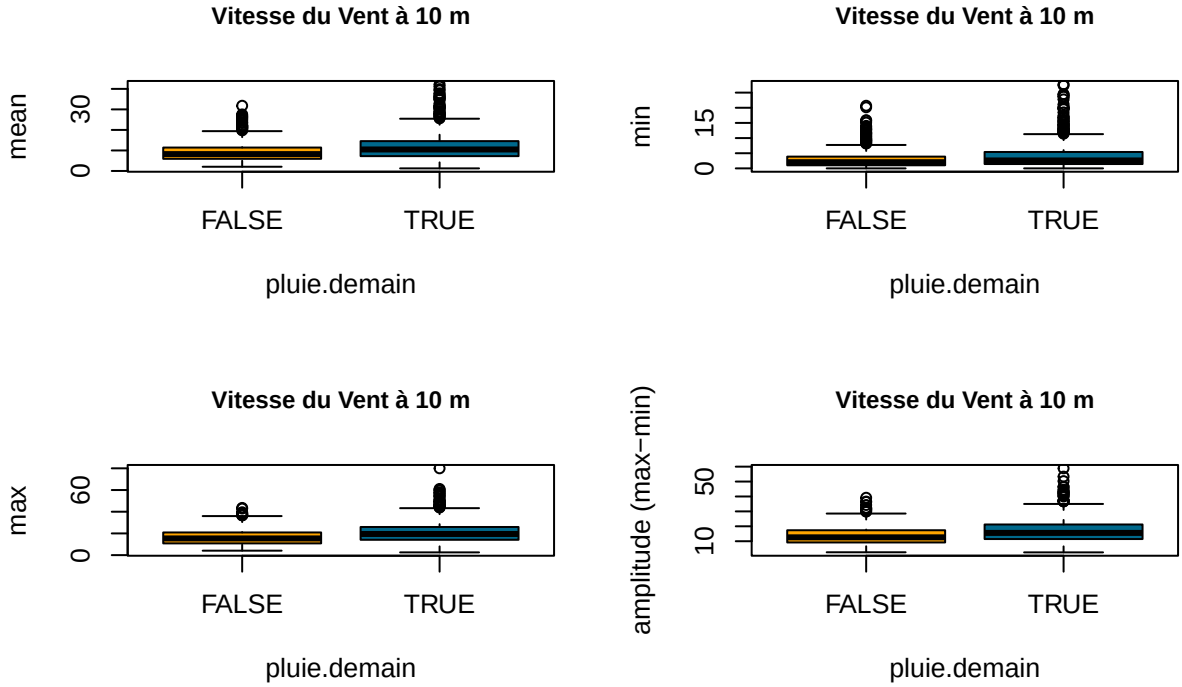
entre nébulosités basse et moyenne : corrélations positives et relativement faibles entre mean/min/max/amplitude (~0.3/0.5)

entre nébulosités moyenne et haute : corrélations positives et relativement forte entre mean/min/max/amplitude (~0.6)

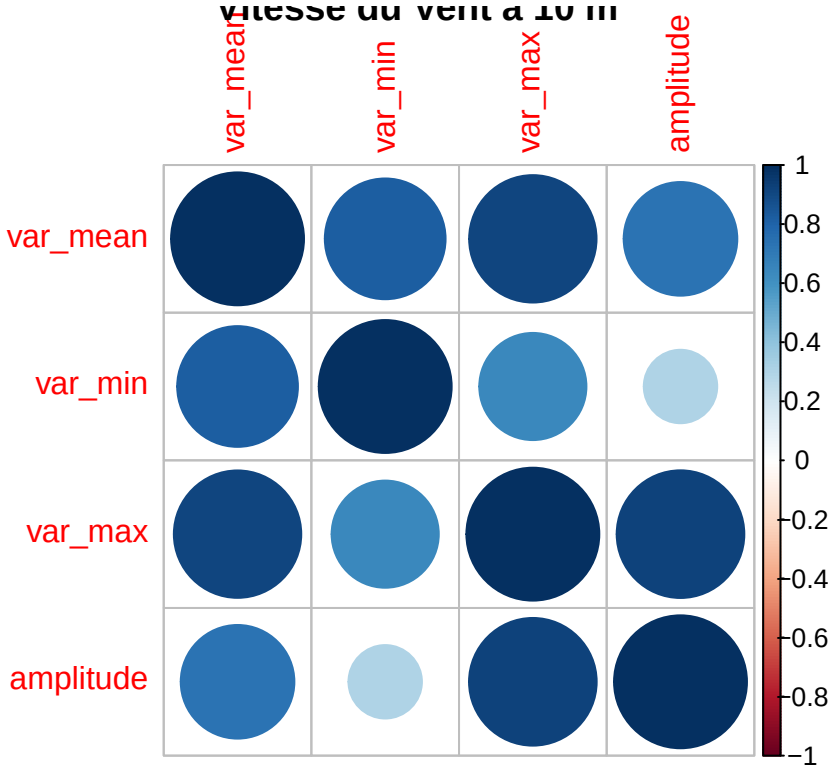
Idées pour la modélisation . inclure le schéma des 3 nébulosités sous la forme imaginée . pour chaque nébulosité :

3.9 Vitesse et sens du vent à 10 m (force et direction)

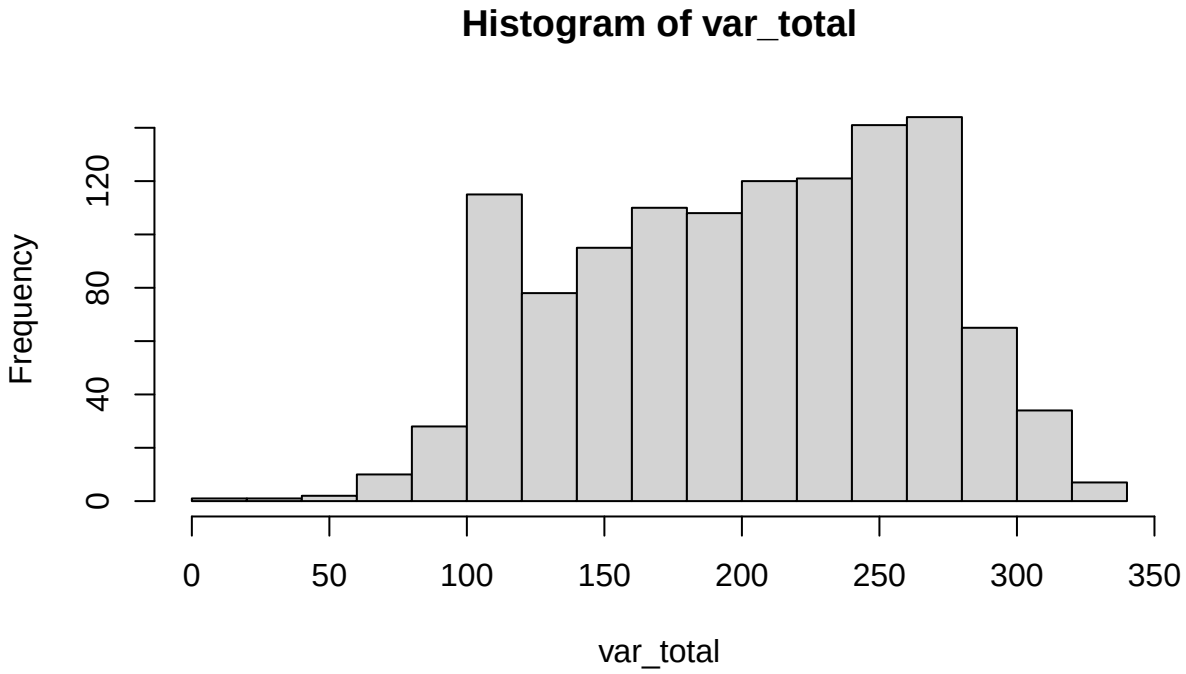
3.9.1 Vitesse du vent

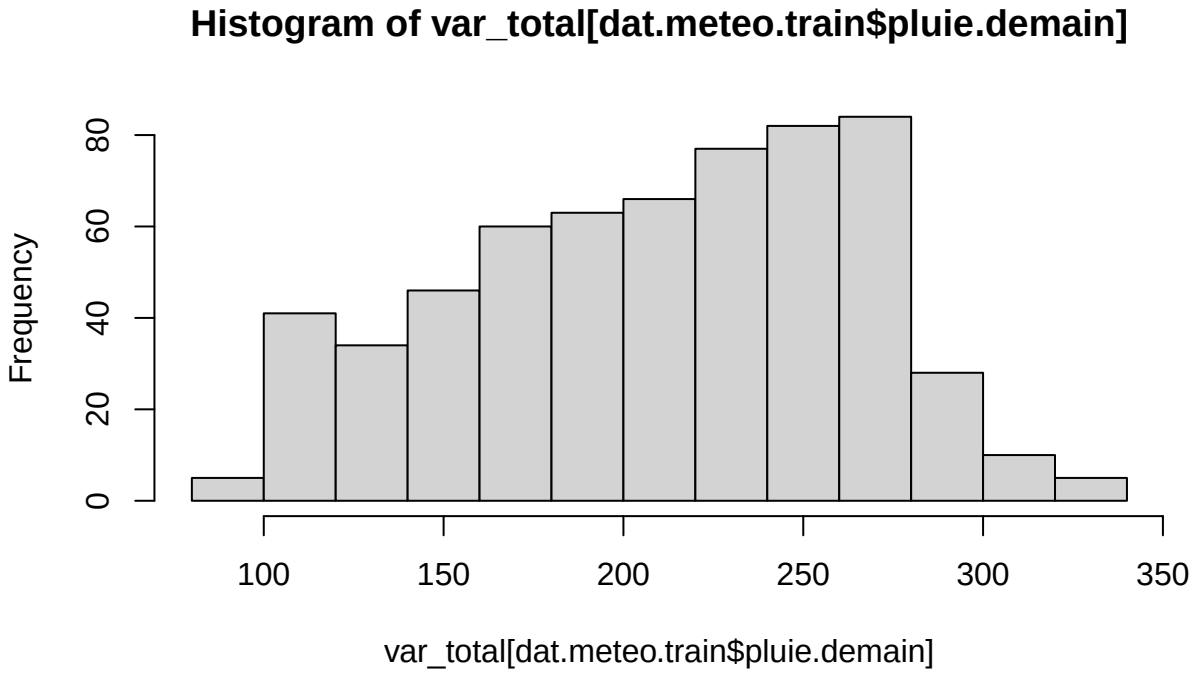


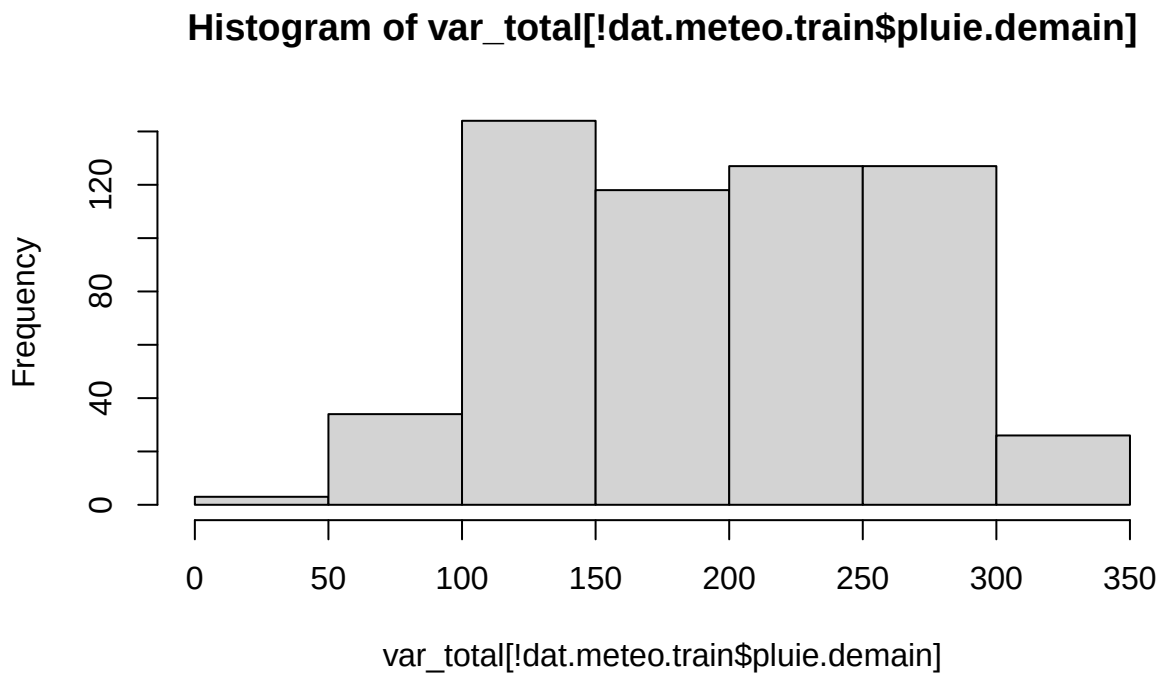

```
##              var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8250362 0.9185578 0.7308883
## var_min    0.8250362 1.0000000 0.6498235 0.3051279
## var_max    0.9185578 0.6498235 1.0000000 0.9221170
## amplitude 0.7308883 0.3051279 0.9221170 1.0000000
```

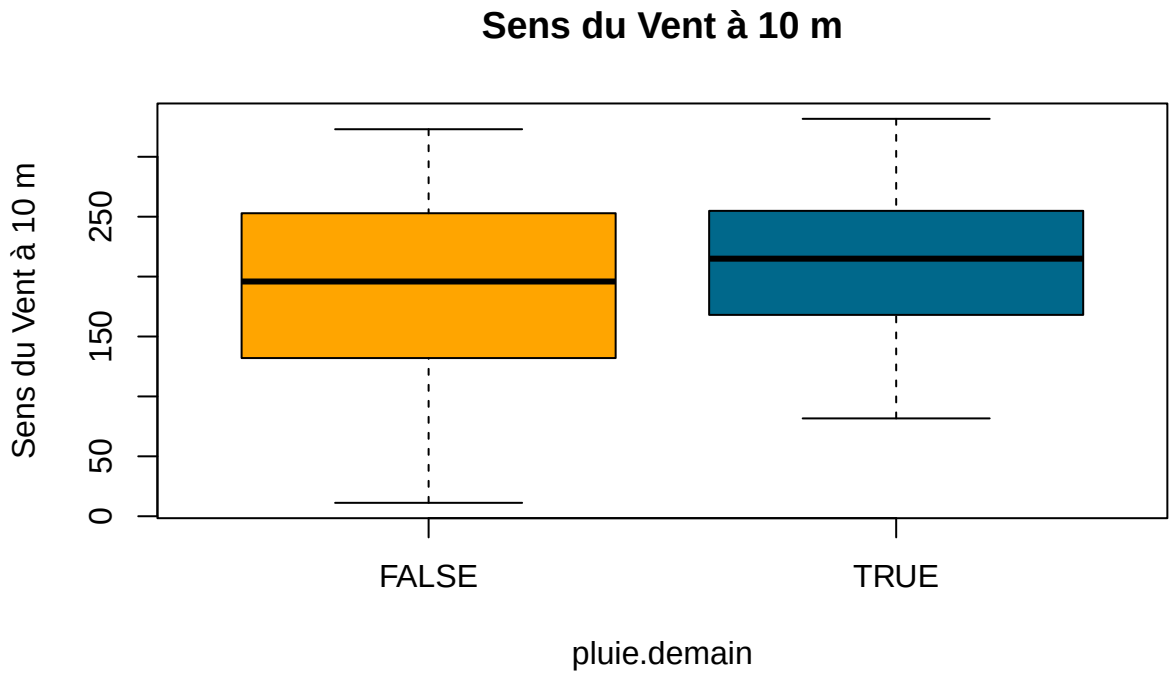

vitesse du vent à 10 m
var_mean
var_min
var_max
amplitude
var_mean
var_min
var_max
amplitude
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

3.9.2 Sens du vent









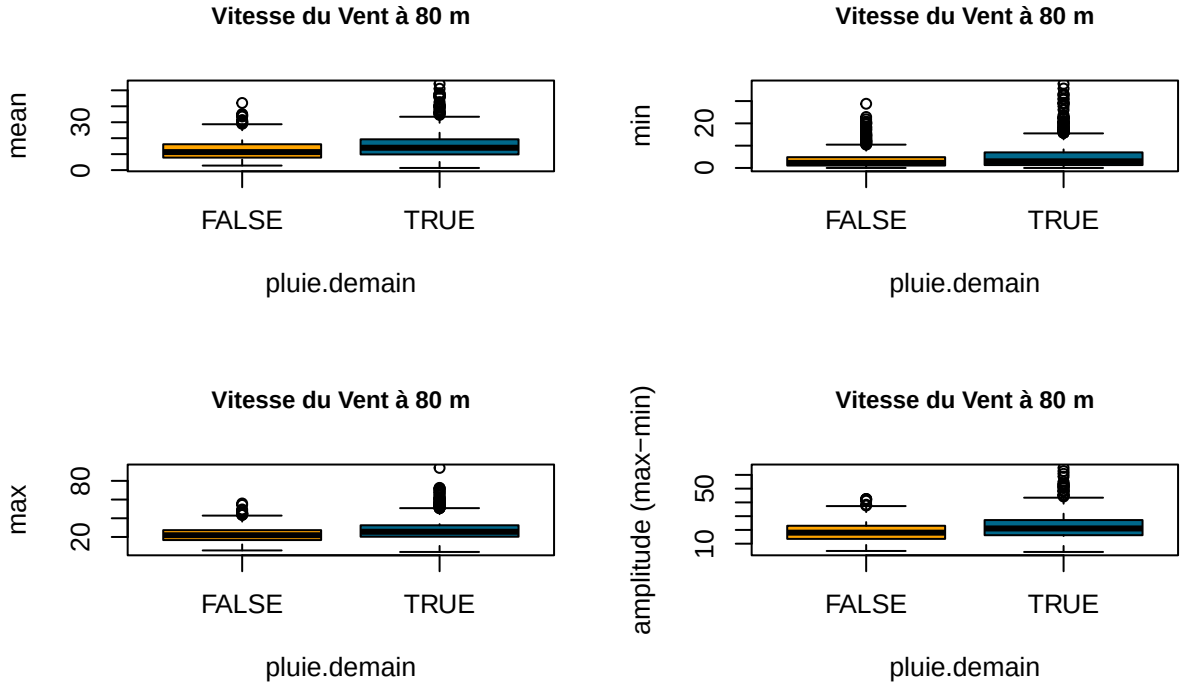

```
## [1] 579

##      0%     25%     50%     75%    100%
##   11.19 132.01 195.91 252.89 323.00
```

```
## [1] 601
```

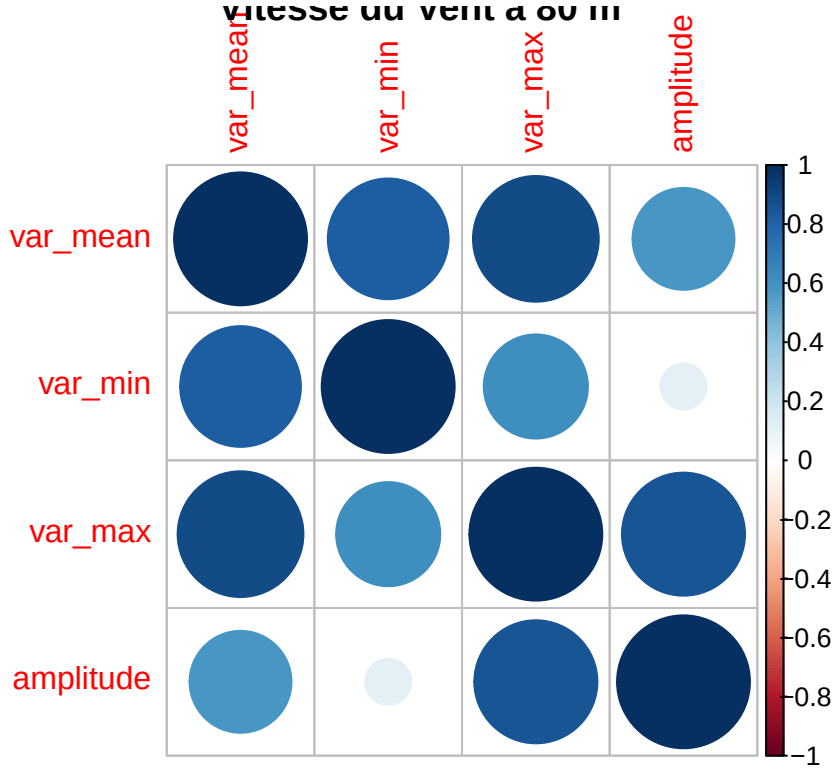
```
##     0%    25%    50%    75%   100%
##  81.64 168.04 214.96 254.86 331.67
```

3.10 Vitesse et sens du vent à 80 m (force et direction)

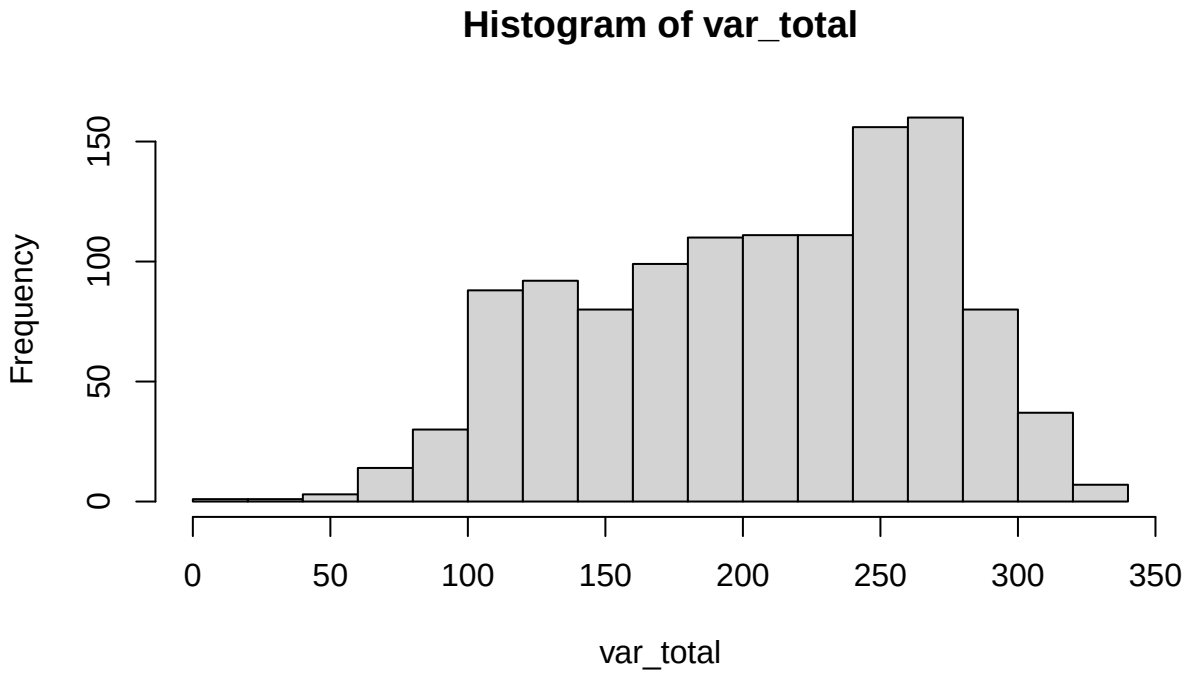
3.10.1 Vitesse du vent

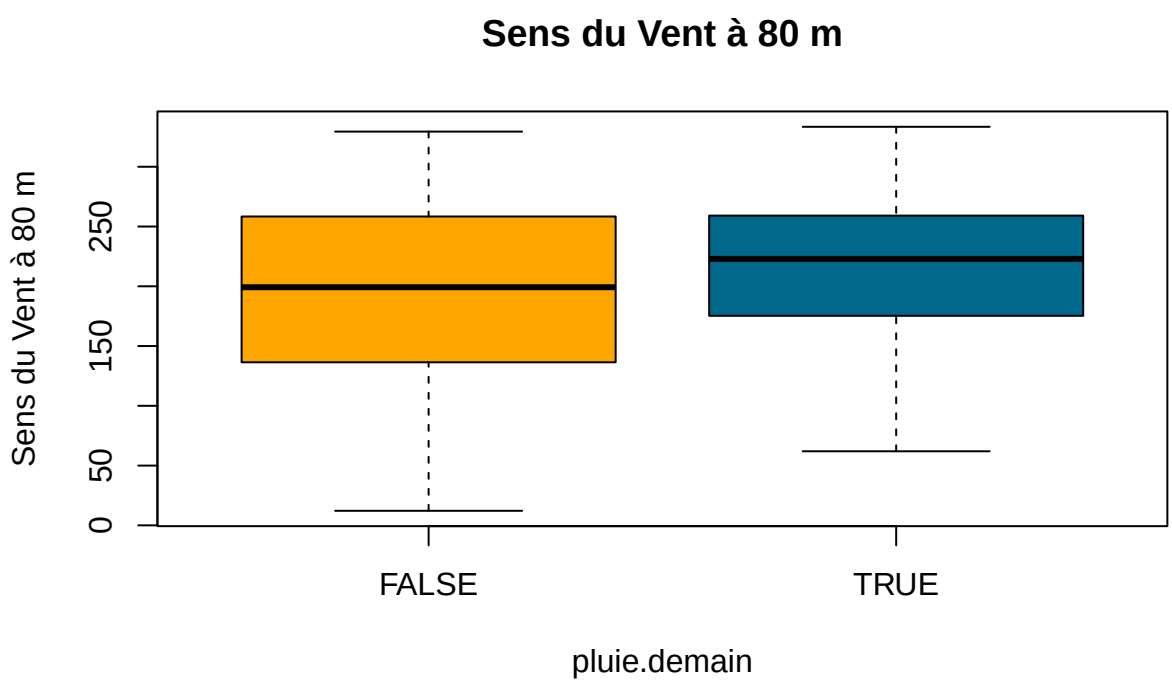



```
##            var_mean   var_min   var_max amplitude
## var_mean  1.0000000 0.8251264 0.8953142 0.5879983
## var_min   0.8251264 1.0000000 0.6137479 0.1199484
## var_max   0.8953142 0.6137479 1.0000000 0.8574200
## amplitude 0.5879983 0.1199484 0.8574200 1.0000000
```



3.10.2 Sens du vent






```
##       0%      25%      50%      75%     100%
##   12.180 136.415 199.230 258.385 329.410
```

```
##      0%    25%    50%    75%   100%
##   61.98 175.31 222.94 259.12 333.43
```

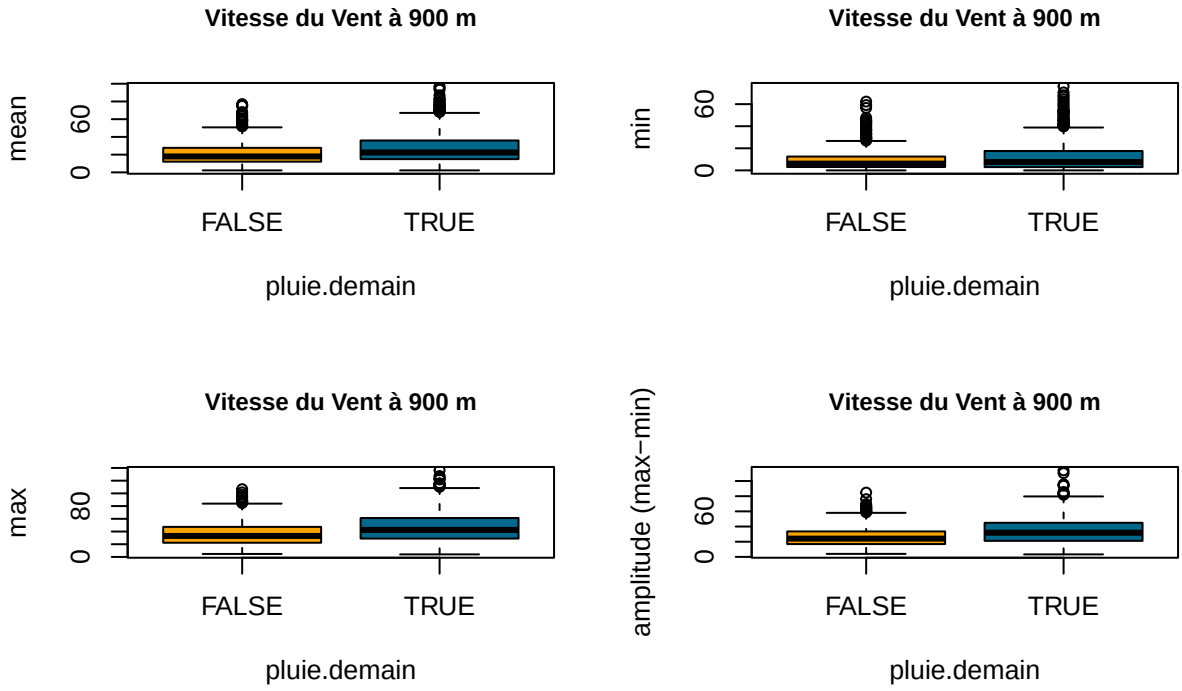
VENT à 80 mètres

=> Les corrélations entre min/max/mean sont positives et relativement fortes (>0.8) => Les corrélations entre min/max/mean et amplitude sont positives et . forte avec mean et max (>0.8) . relativement forte avec mean (~0.6) . relativement faible avec min (~0.6)

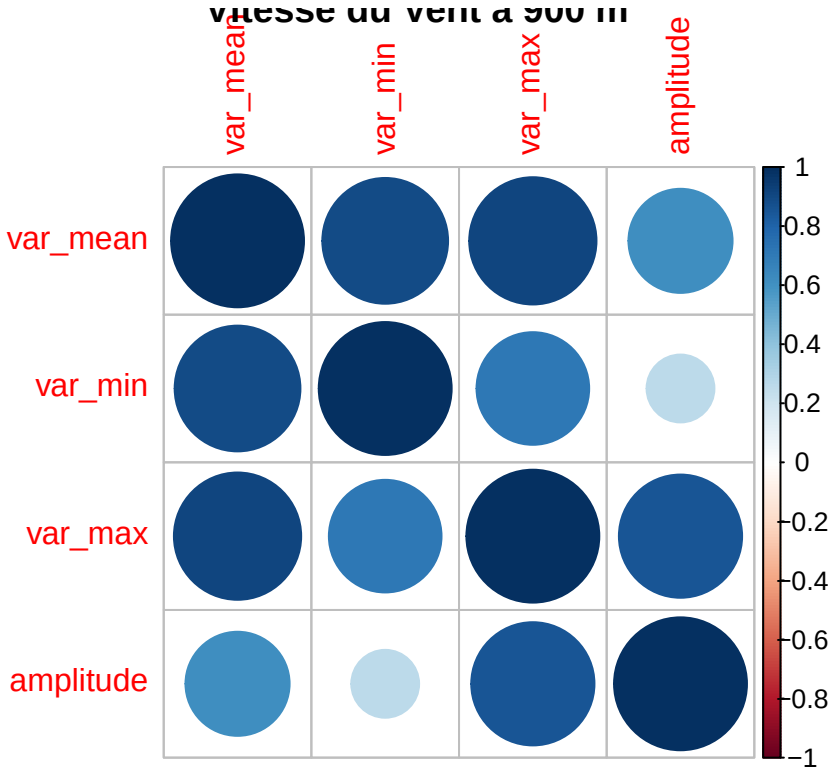
Idées pour la modélisation Inclure un unique représentant parmi moyenne/min/max fortement corrélées : min en l'occurrence . inclure l'amplitude . Considérer la covariable produit amplitude*min

3.11 Vitesse et sens du vent à 900 m (force et direction)

3.11.1 Vitesse du vent




```
##              var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8953039 0.9168068 0.6133562
## var_min    0.8953039 1.0000000 0.7179283 0.2605610
## var_max    0.9168068 0.7179283 1.0000000 0.8591355
## amplitude 0.6133562 0.2605610 0.8591355 1.0000000
```

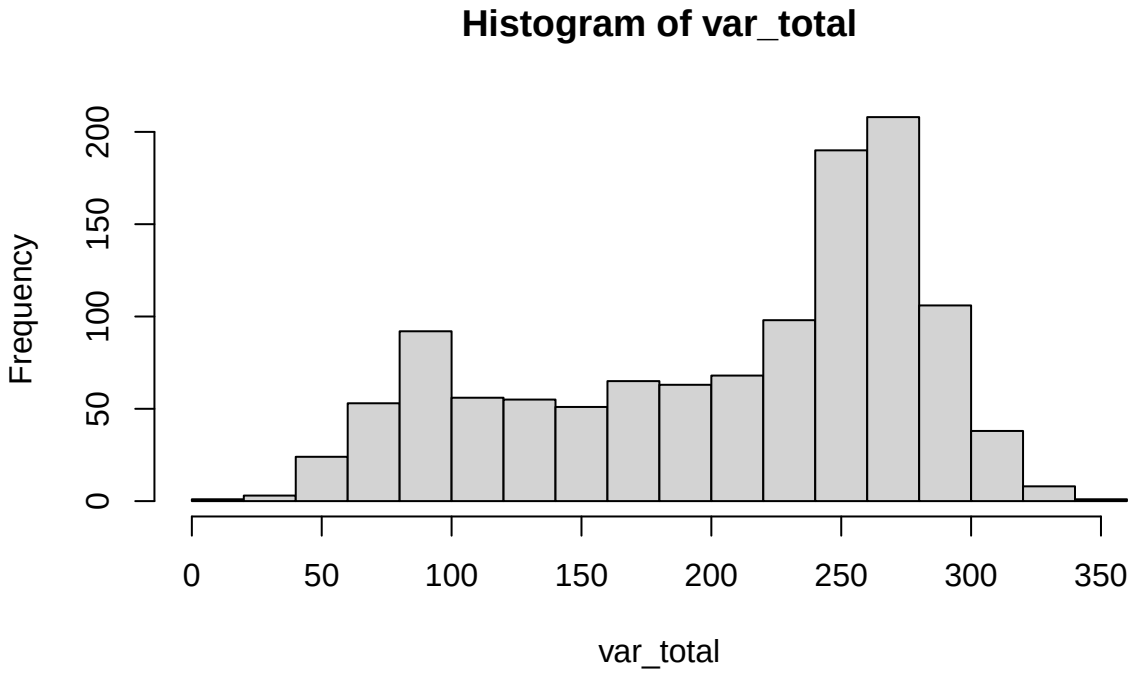


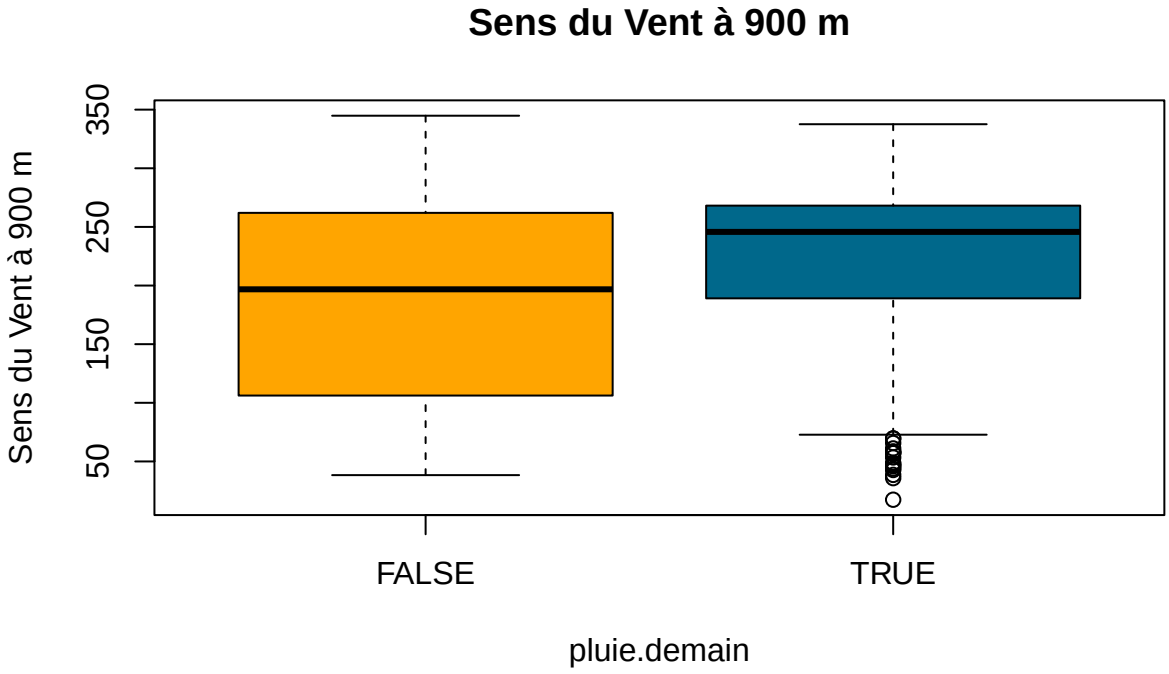
VENT à 900 mètres

=> Les corrélations entre min/max/mean sont positives et relativement fortes (>0.9) => Les corrélations entre min/max/mean et amplitude sont positives et . forte avec max (>0.8) . relativement forte avec mean (~0.6) . relativement faible avec min (~0.26)

Idées pour la modélisation . Inclure un unique représentant parmi moyenne/min/max fortement corrélées : min en l'occurrence . inclure l'amplitude . Considérer la covariable produit amplitude*min

3.11.2 Sens du vent






```
##      0%     25%     50%     75%    100%
##  38.340 106.200 196.790 261.985 344.820
```

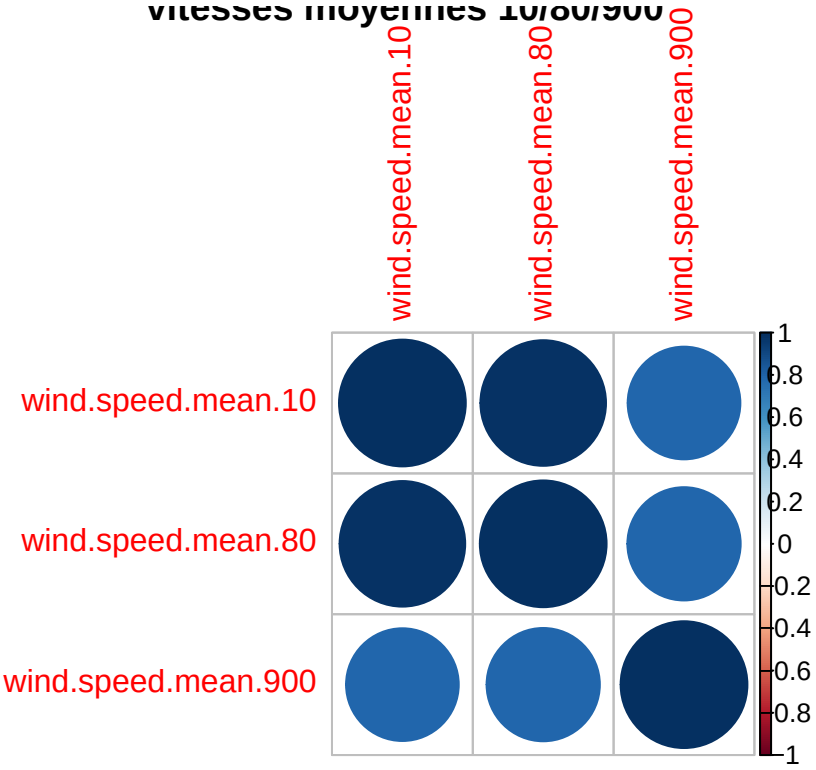
```
##      0%     25%     50%     75%    100%
##   17.37  189.07  245.75  268.14  337.56
```

SENS du VENT à 900 mètres => on note une différence notable dans la distribution des directions du vent en fontction du fait qu'il ait plu le lendemain

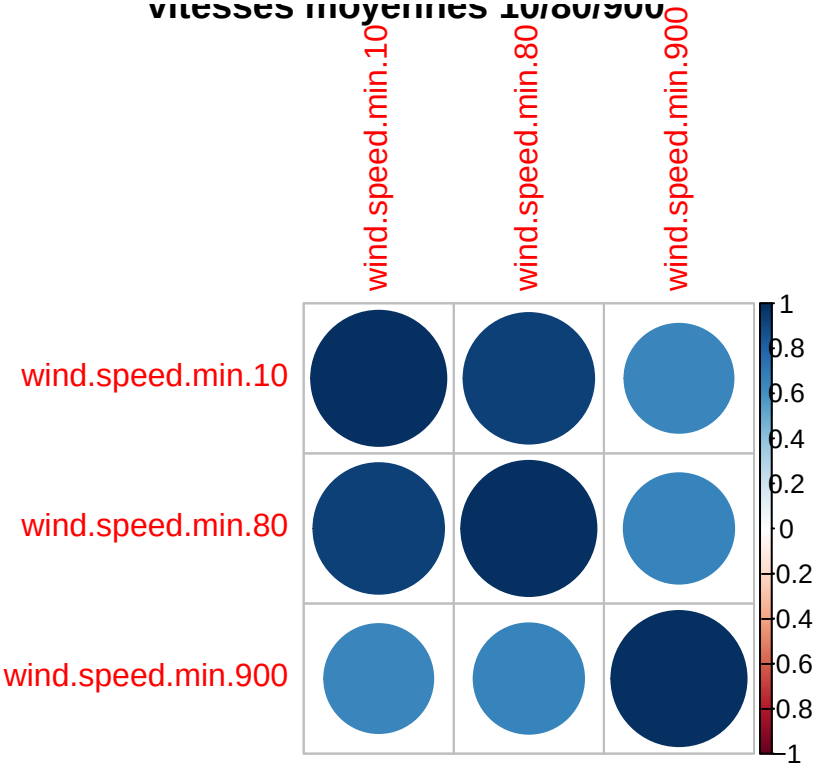
Idées pour la modélisation . inclure la direction du vent . combiner à la donnée vitesse du vent . i.e amplitude*min*dir

3.12 Corrélation entre vitesses et sens du vent

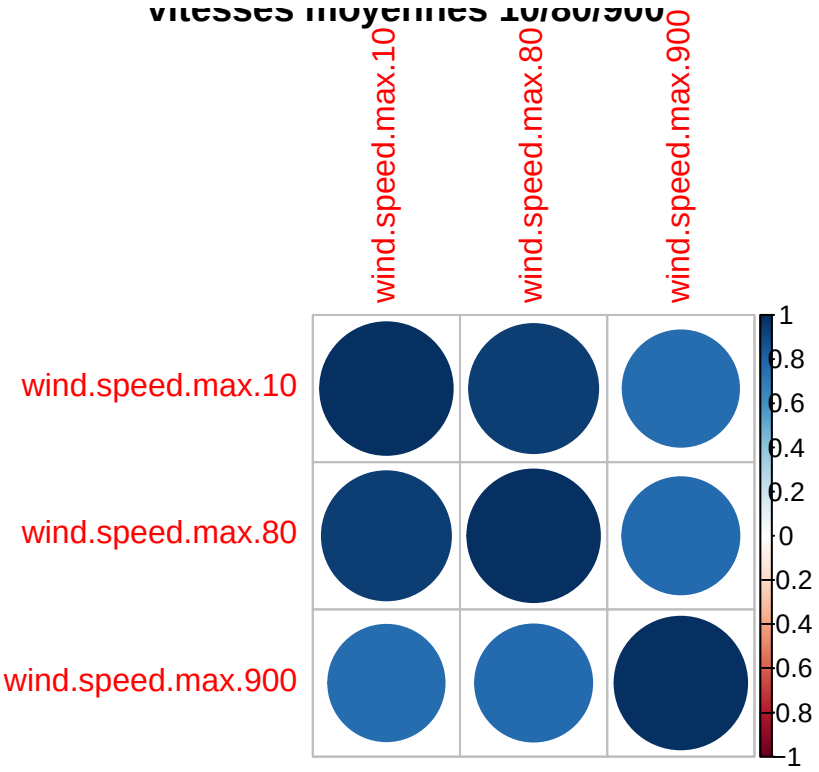
```
##                       wind.speed.mean.10 wind.speed.mean.80 wind.speed.mean.900
## wind.speed.mean.10             1.0000000          0.9816588           0.7905158
## wind.speed.mean.80             0.9816588          1.0000000           0.7985150
## wind.speed.mean.900            0.7905158          0.7985150           1.0000000
```



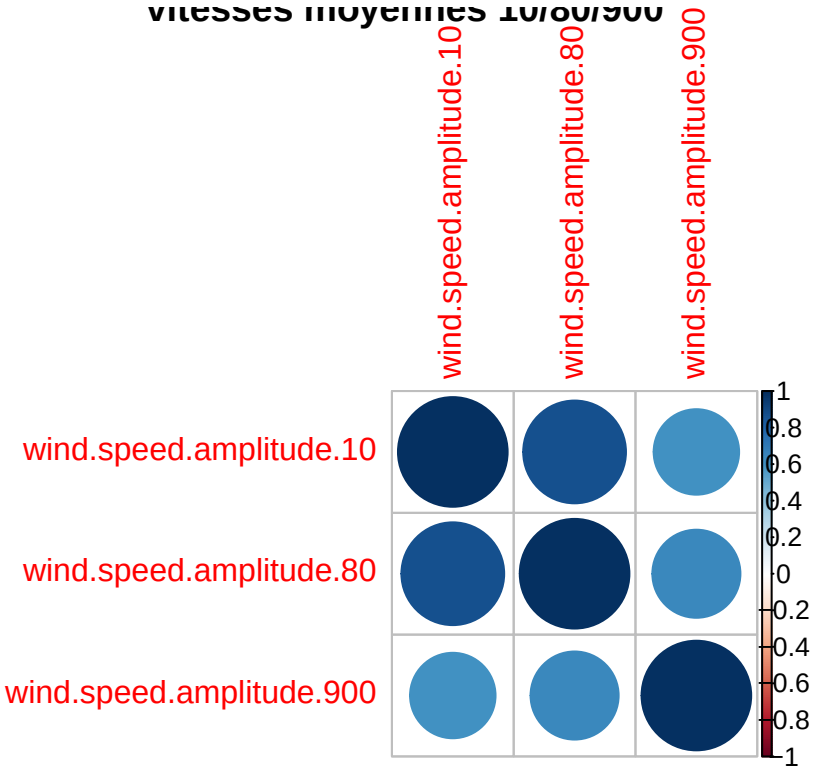

```
##                       wind.speed.min.10 wind.speed.min.80 wind.speed.min.900
## wind.speed.min.10             1.0000000         0.9333039          0.6500372
## wind.speed.min.80             0.9333039         1.0000000          0.6666382
## wind.speed.min.900            0.6500372         0.6666382          1.0000000
```



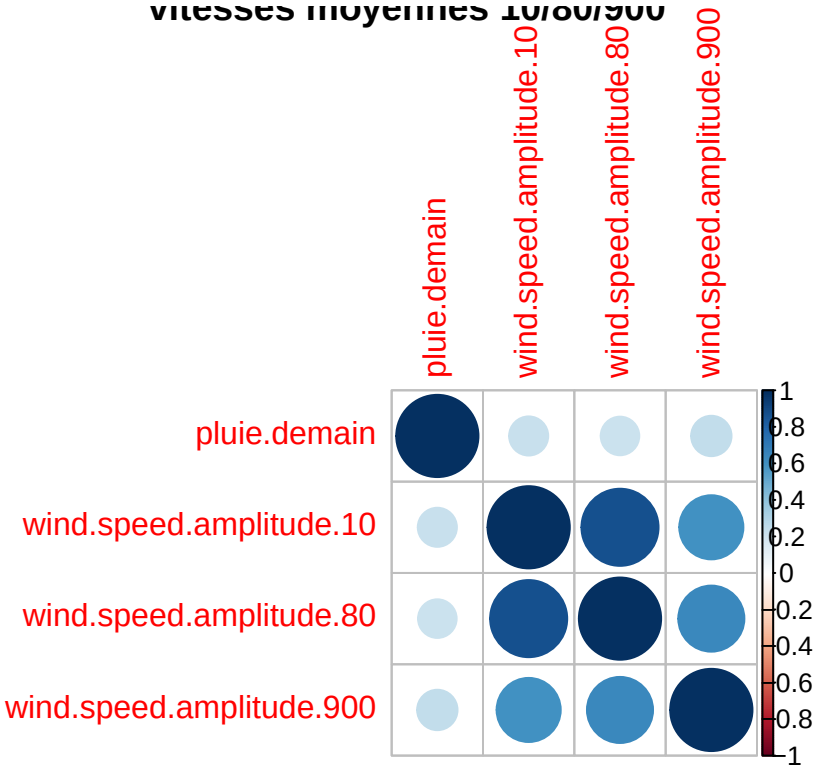

```
##                     wind.speed.max.10 wind.speed.max.80 wind.speed.max.900
## wind.speed.max.10           1.0000000         0.9476338          0.7699367
## wind.speed.max.80           0.9476338         1.0000000          0.7799064
## wind.speed.max.900          0.7699367         0.7799064          1.0000000
```



```
##                           wind.speed.amplitude.10 wind.speed.amplitude.80
## wind.speed.amplitude.10                 1.0000000               0.8787860
## wind.speed.amplitude.80                 0.8787860               1.0000000
## wind.speed.amplitude.900                0.6076737               0.6454319
##                           wind.speed.amplitude.900
## wind.speed.amplitude.10                  0.6076737
## wind.speed.amplitude.80                  0.6454319
## wind.speed.amplitude.900                 1.0000000
```




```
##              pluie.demain  wind.speed.amplitude.10  wind.speed.amplitude.80
##                 1.0000000                0.2257354                0.2199786
## wind.speed.amplitude.900
##                 0.2409870
```



Les corrélations entre les amplitudes 10/80/900 et la variable d'intéret projetée sur [0,1] sont faibles (~25%).

Le vent du jour a peu d'influence sur le risque de pluie du lendemain

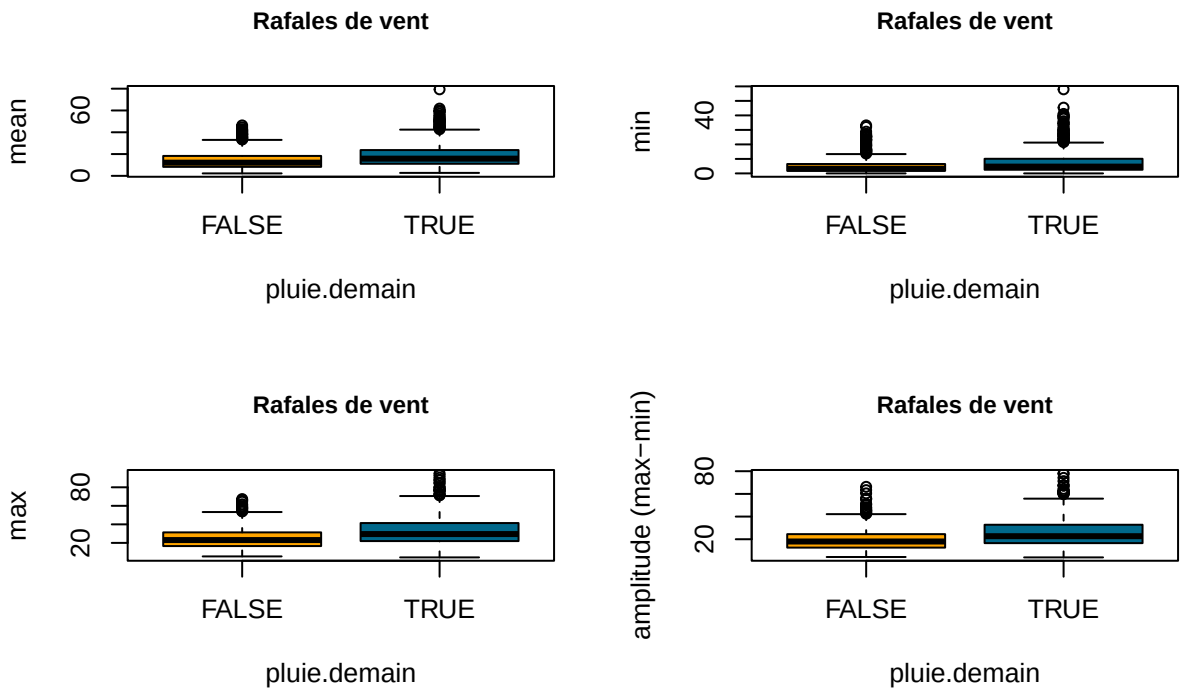
CORRELATION entre VENTS

. Les données à 10, 80, 900 sont fortement corrélées ; on choisira un unique altitude

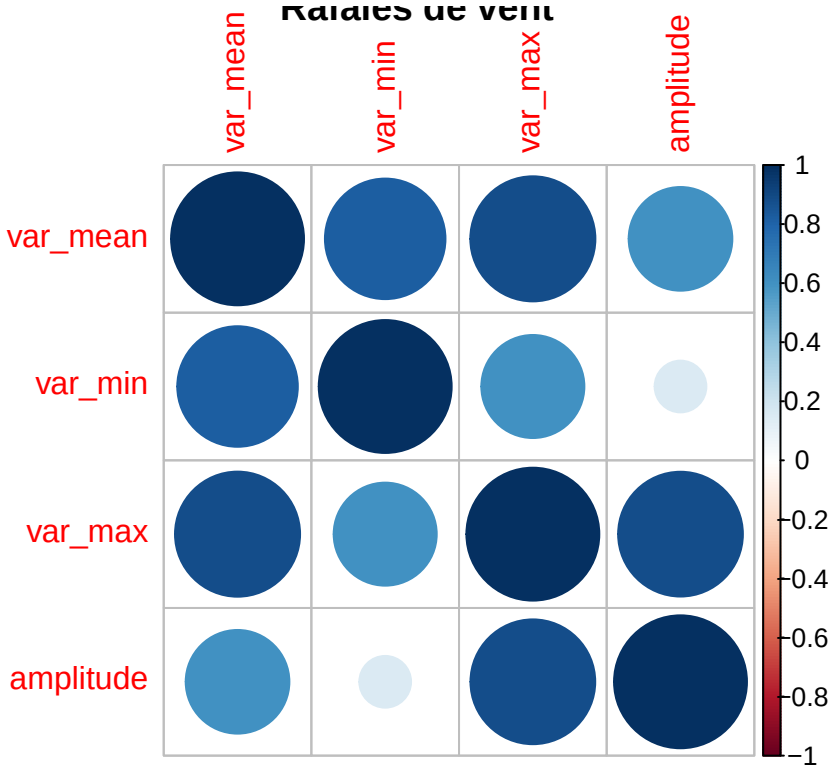
. Les données à 900 sont celles les plus corrélées avec la variable pluie.demain projetée sur [0;1]

Idée pour la modélisation . on ne considère que les données à 900m . on inclura la covariable produit amplitudemin

3.13 Rafales de vent




```
##            var_mean   var_min   var_max amplitude
## var_mean  1.0000000 0.8223337 0.8853670 0.6083197
## var_min   0.8223337 1.0000000 0.6007769 0.1513257
## var_max   0.8853670 0.6007769 1.0000000 0.8811236
## amplitude 0.6083197 0.1513257 0.8811236 1.0000000
```



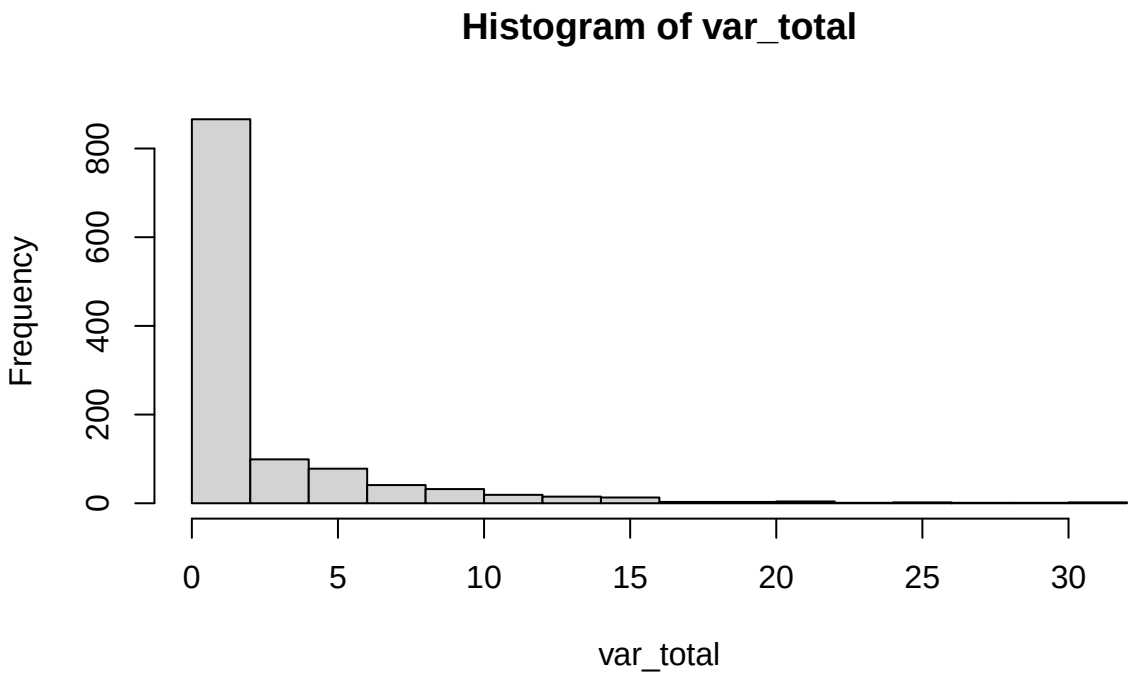
RAFALES de VENT

=> Les corrélations entre min/max/mean sont fortes (>0.8 => Les corrélations entre max/mean et amplitude sont plutôt forte (>0.6) => La corrélation entre min et amplitude est faible (0.15)

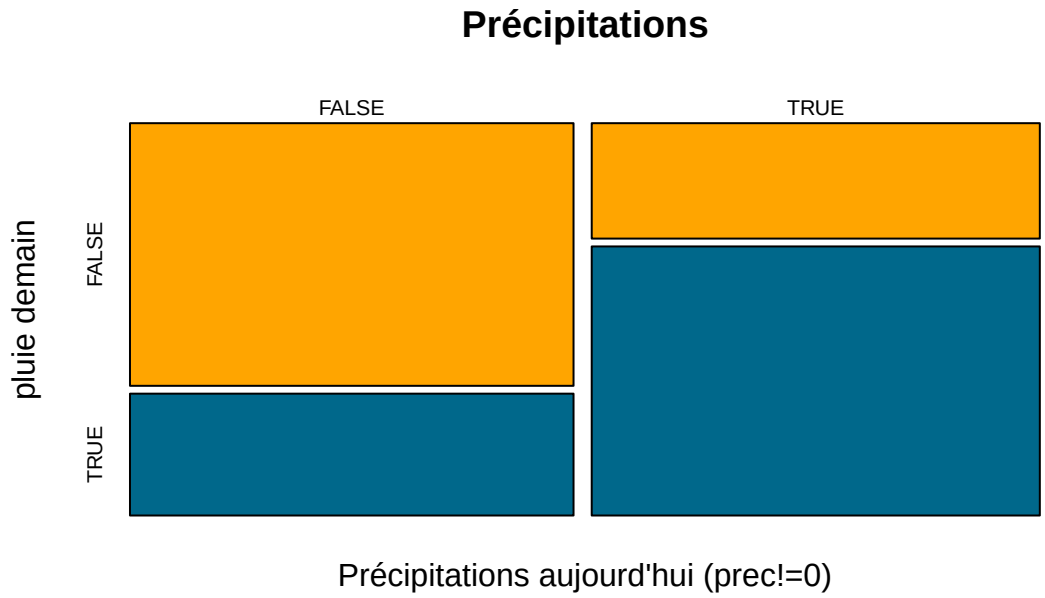
Idées pour la modélisation . Inclure un unique représentant parmi moyenne/min/max fortement corrélées : min en l'occurrence . inclure l'amplitude . Considérer la covariable produit amplitude*min

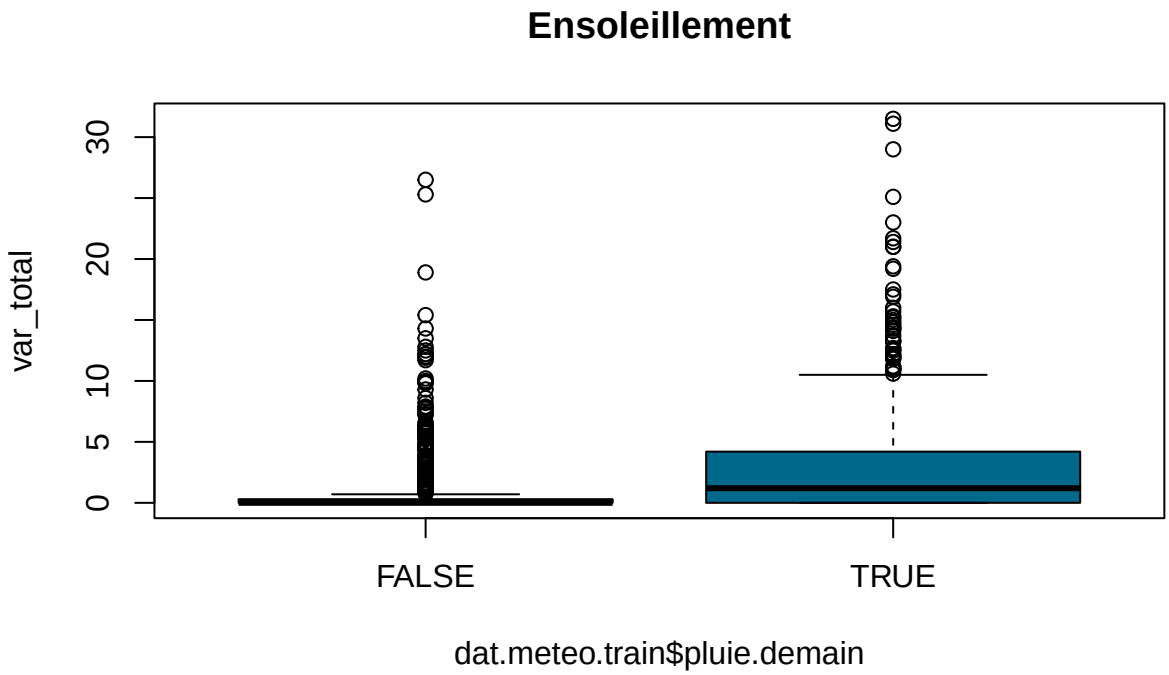
3.14 Covariables simples

3.14.1 Précipitations




```
## [1] 0.4974576
```






```
##   0%  25%  50%  75% 100%
##  0.0  0.0  0.0  0.3 26.5

##   0%  25%  50%  75% 100%
```

```
##  0.0  0.0  1.2  4.2 31.5
```

=> Quand il a plu le lendemain, 75% des valeurs de précipitations sont 0 => Quand il n'a pas plu le lendemain, 50% des valeurs sont 0

PRECIPITATION

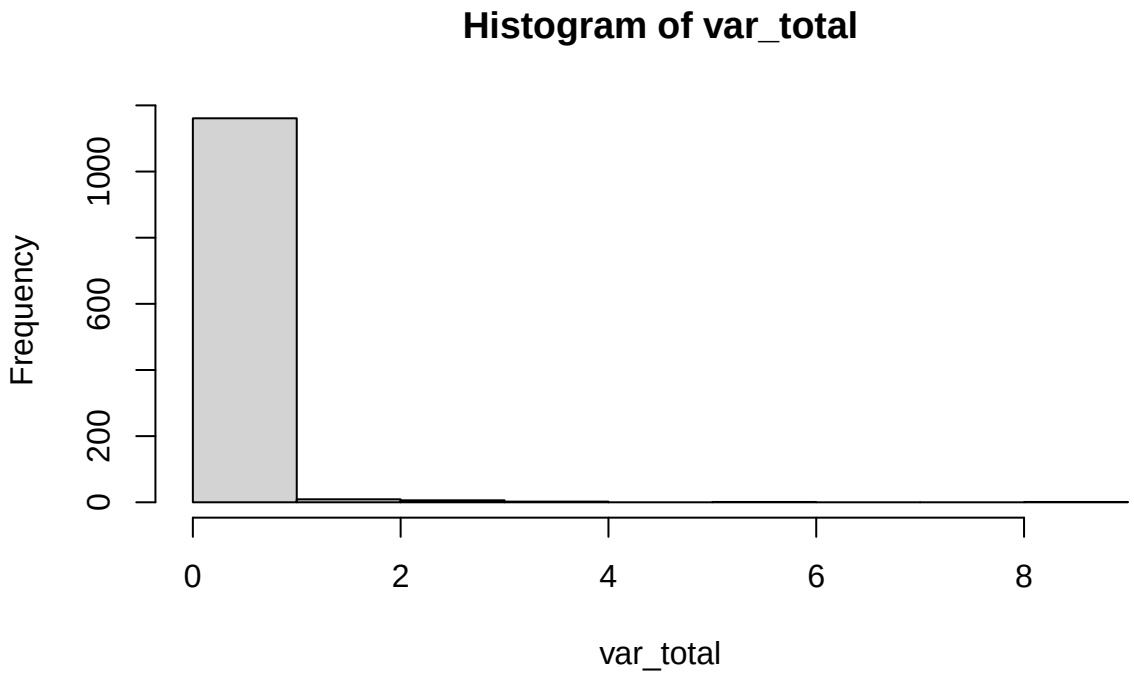
Différence dans les distributions selon pluie.demain

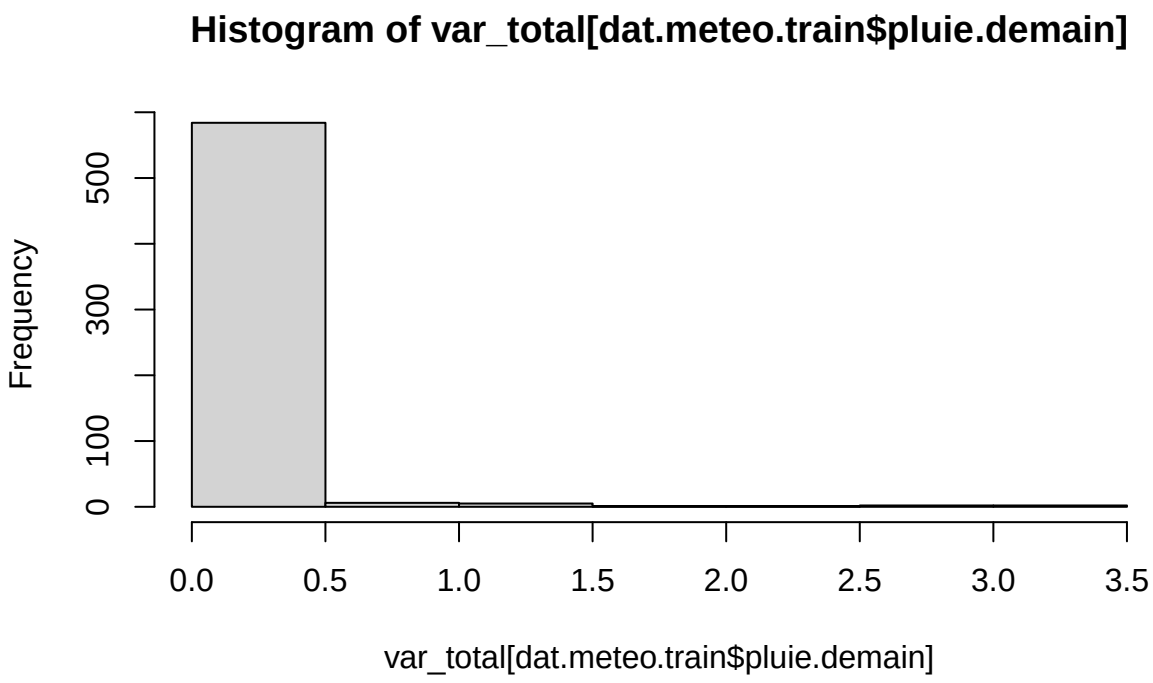
La majorité des valeurs sont nulles . pluie.demain=TRUE : 50% des valeurs sont nulles . pluie.demain=FALSE : 75% des valeurs sont nulles

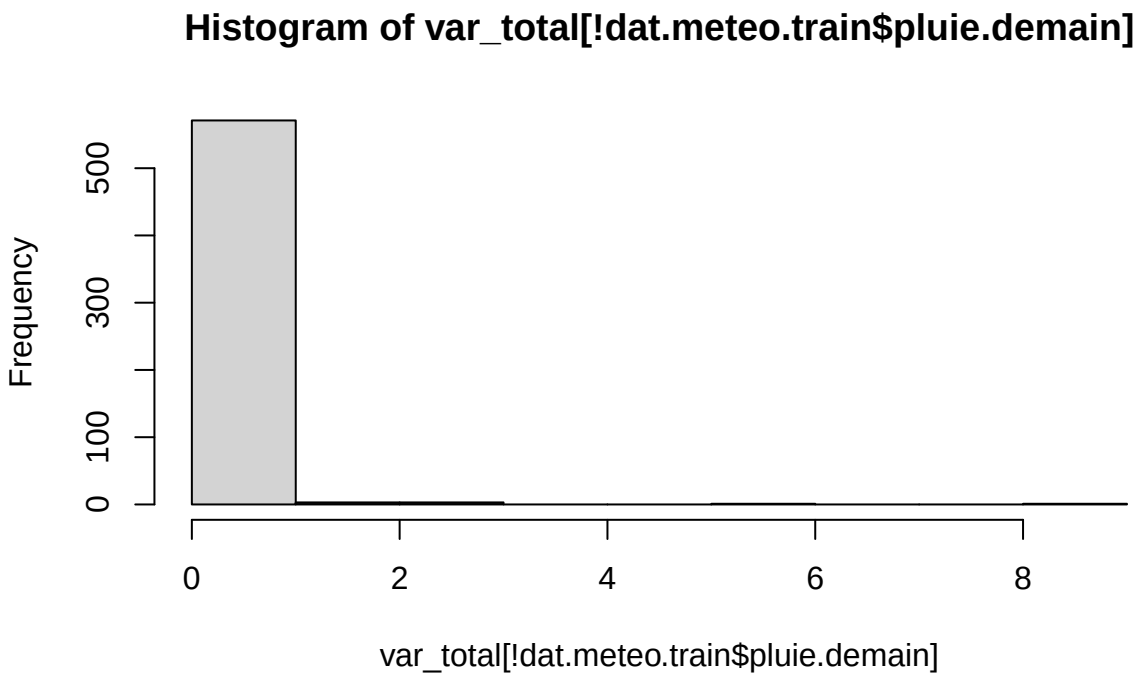
Idée pour la modélisation . Inclure Précipitation . Inclure Précipitation sous forme booléenne . inclure sous la forme d'un produit précipitation*(précipitation>0.5)

3.14.2 Enneigement

BKU : !!! la majeure partie des valeurs est nulle . Considérer une variable booléenne . vérifier la corrélation entre neige==TRUT et pluie demain==TRUE

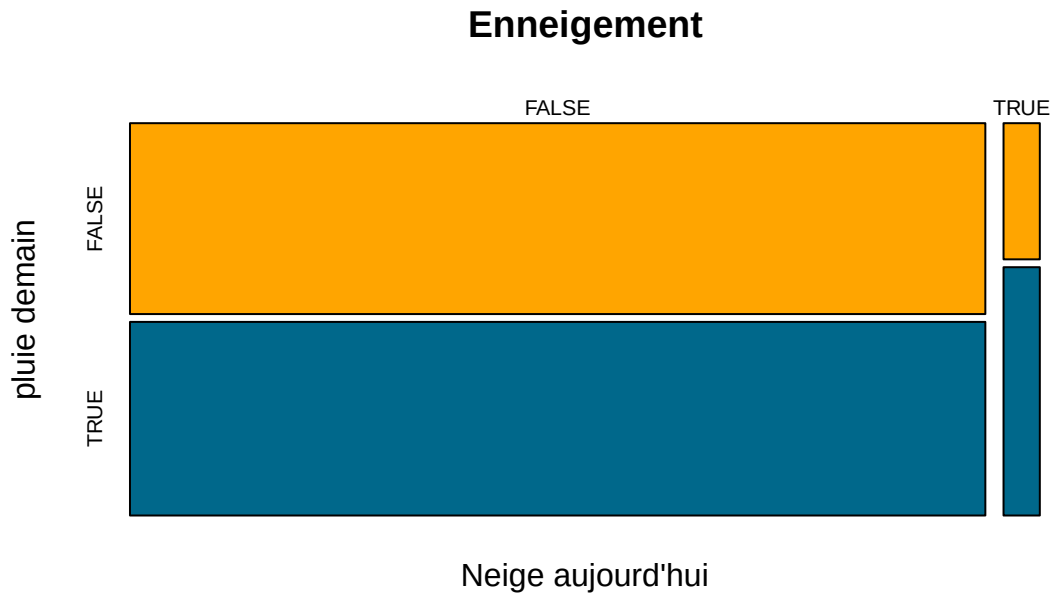




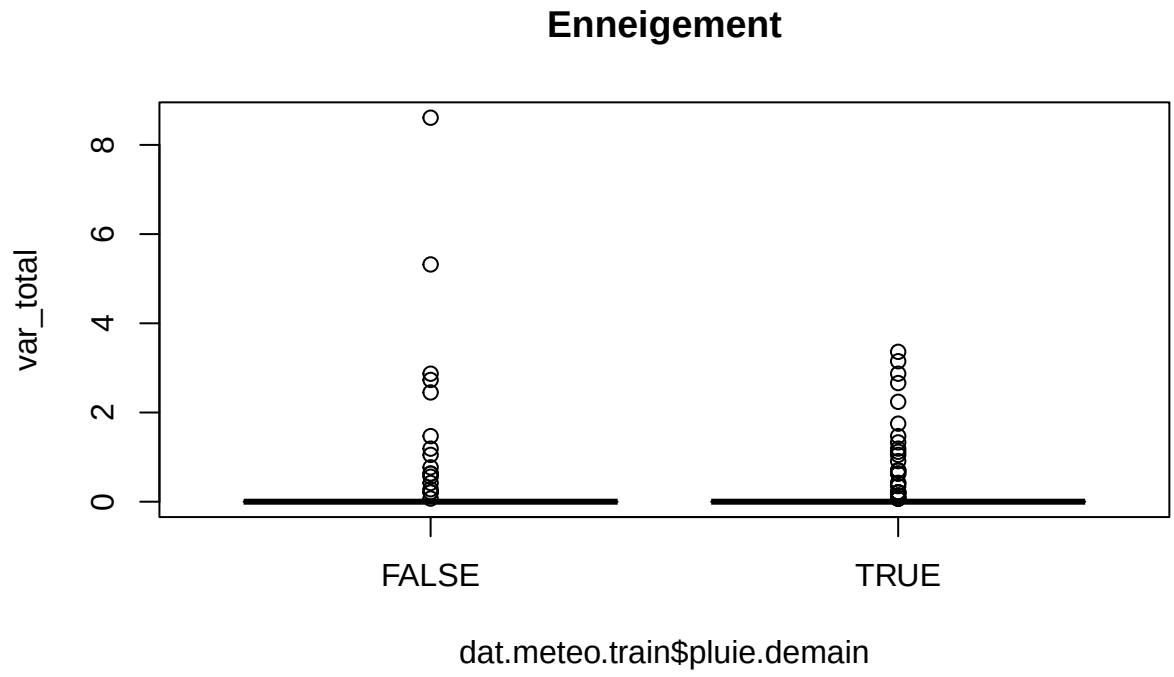



```
## [1] 0.959322
```

=> 95% des valeurs sont nuls : pas de neige



=> légère différence entre le ratio de “pluie.demain” selon l’enneigement du jour.
Quand il a neigé, le risque de pluie est plus fort le lendemain quand il n’a pas neigé, 1 chance sur 2 qu’il pleuve le lendemain




```
##    0%   25%   50%   75%  100%
```

```
## 0.00 0.00 0.00 0.00 8.61
```

```
##   0%  25%  50%  75% 100%
## 0.00 0.00 0.00 0.00 3.36
```

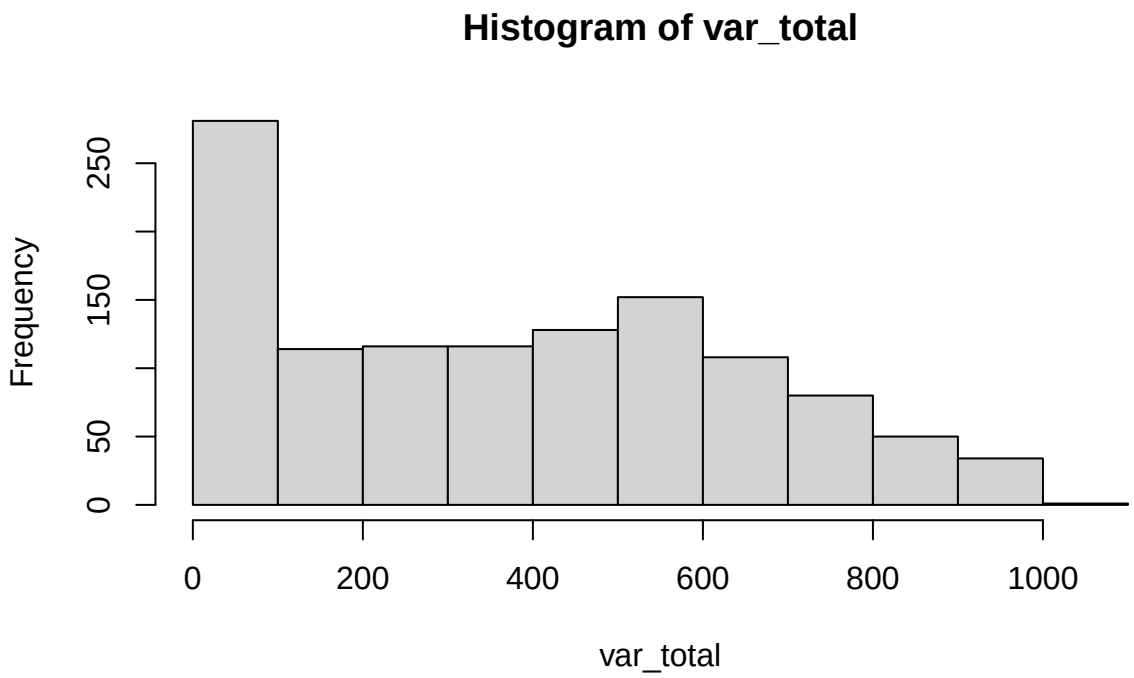
=> Qu'il pleuve ou non le lendemain, 75% des valeurs sont nuls => Quand il a neigé la veille, le risque de pluie est plus fort à mesure que les précipitations de neige sont fortes

NEIGE du JOUR :

. Les jours sans neige représentent 50% des cas. . Qu'il ait plu ou non, 75% des valeurs d'enneigement sont nuls . Quand il a plu, et qu'il a neigé, l'enneigement a été plus fort

Idée pour la modélisation . Inclure Neige . Inclure Neige sous forme booléenne . inclure sous la forme d'un produit neige*(neige>0)

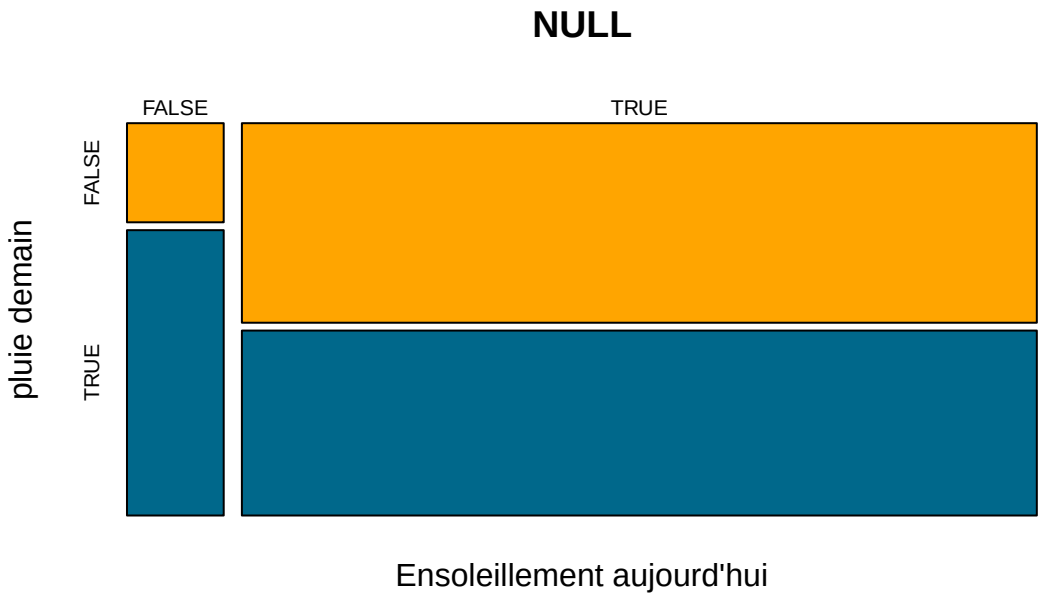
3.14.3 Ensoleillement

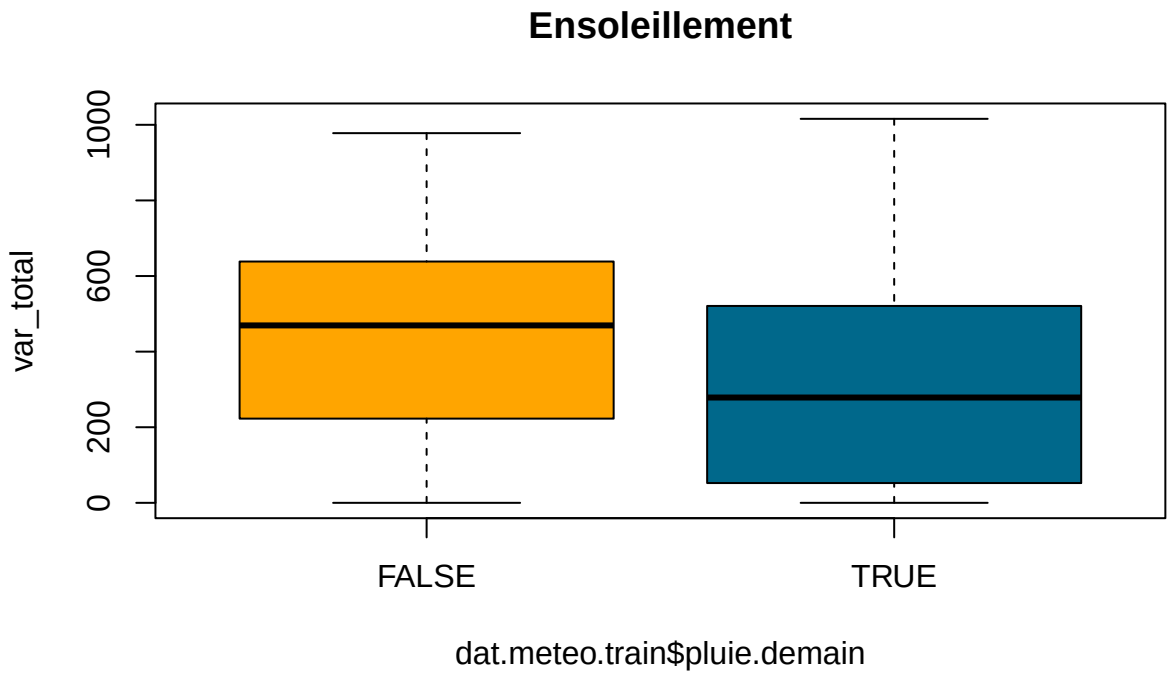



```
## [1] 0.1084746
```

```
## 10%
##   0
```

=> 10% des valeurs sont nuls (10% des journées sans soleil)





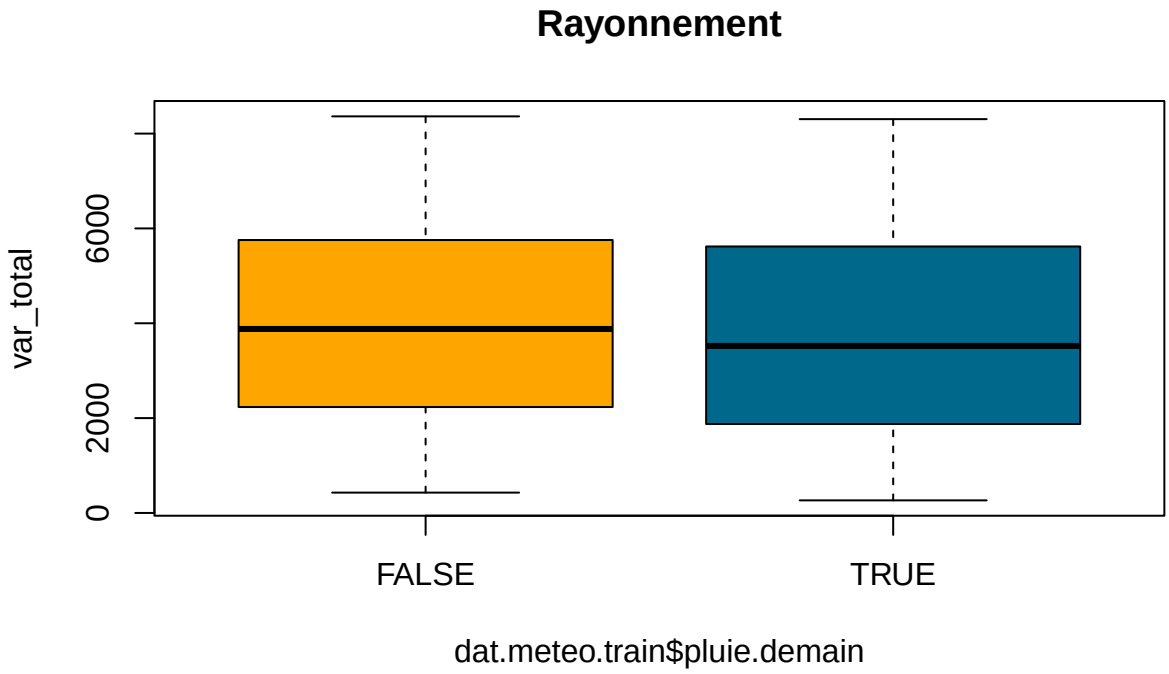

```
##      0%     25%     50%     75%    100%
##   0.000 222.765 469.320 638.340 977.980

##      0%     25%     50%     75%    100%
```

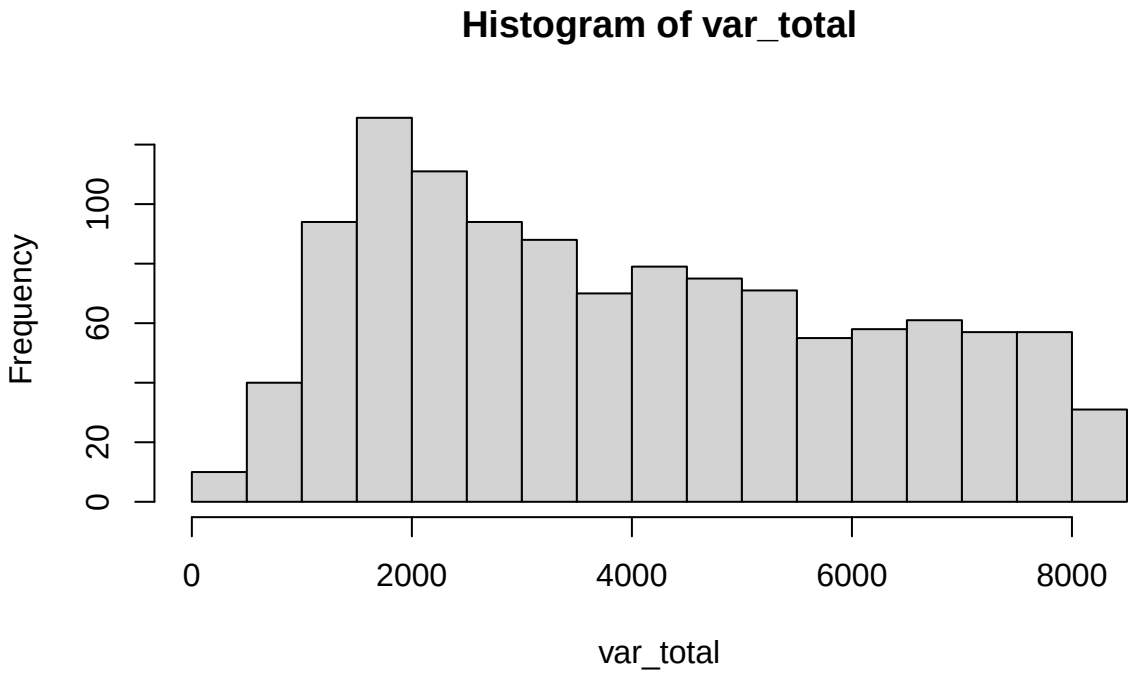
```
##    0.00   52.34  278.57  520.81 1015.83
```

ENSOLEILLEMENT :

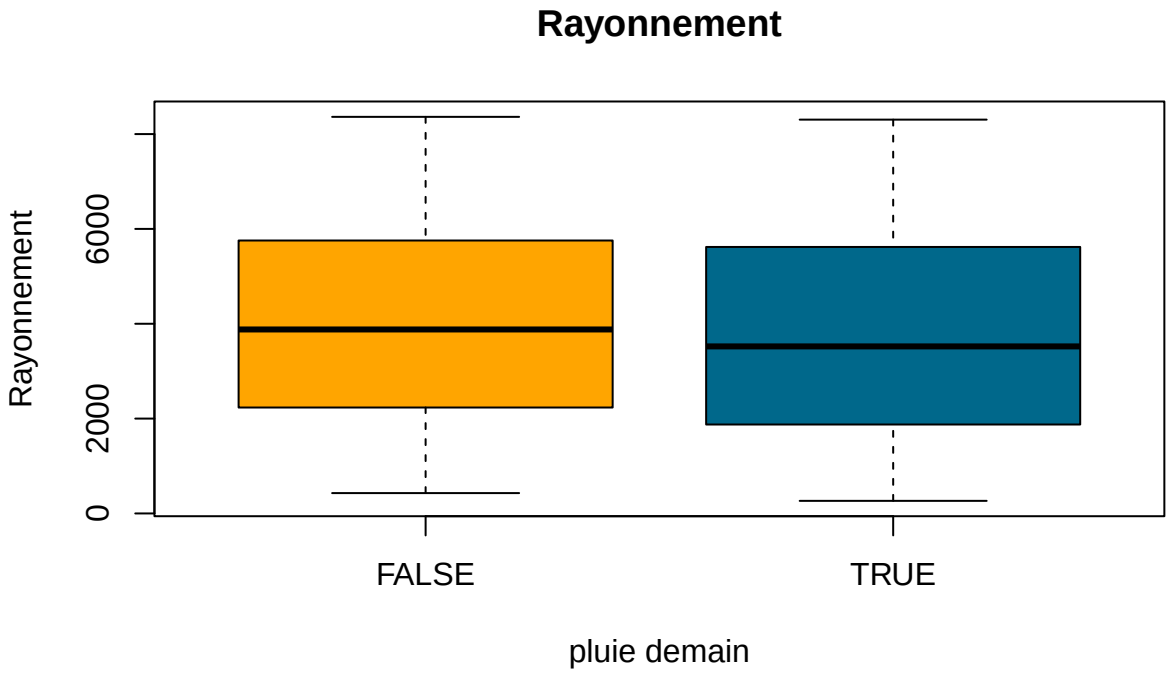
Idée pour la modélisation . Inclure Ensoleillement . Inclure Ensoleillement sous forme booléenne . inclure sous la forme d'un produit

3.14.4 Rayonnement





=> 10% des valeurs sont nuls (10% des journées sans soleil)



=> Légère différence de la distribution de la radiation selon pluie.demain => Un rayonnement plus faible augmente le risque de pluie

```
##       0%      25%      50%      75%     100%
##  428.980 2233.455 3879.510 5755.630 8363.330
```

```
##      0%     25%     50%     75%    100%
##  265.22 1875.23 3522.62 5619.46 8304.59
```

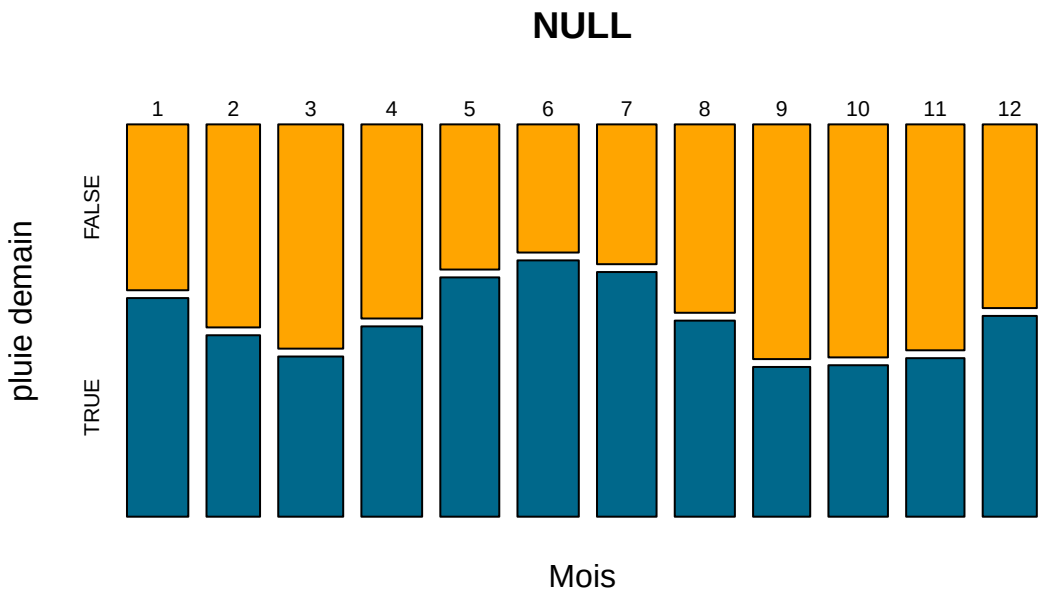
RAYONNEMENT :

=> La variation de la distribution du rayonnement selon pluie.demain est minime

=> pluie.demain==TRUE : une légère tendance à un rayonnement plus faible => pluie.demain==FALSE : une légère tendance à un rayonnement plus fort

Idées pour la modélisation . inclure le rayonnement en l'état

3.14.5 Mois



Mois

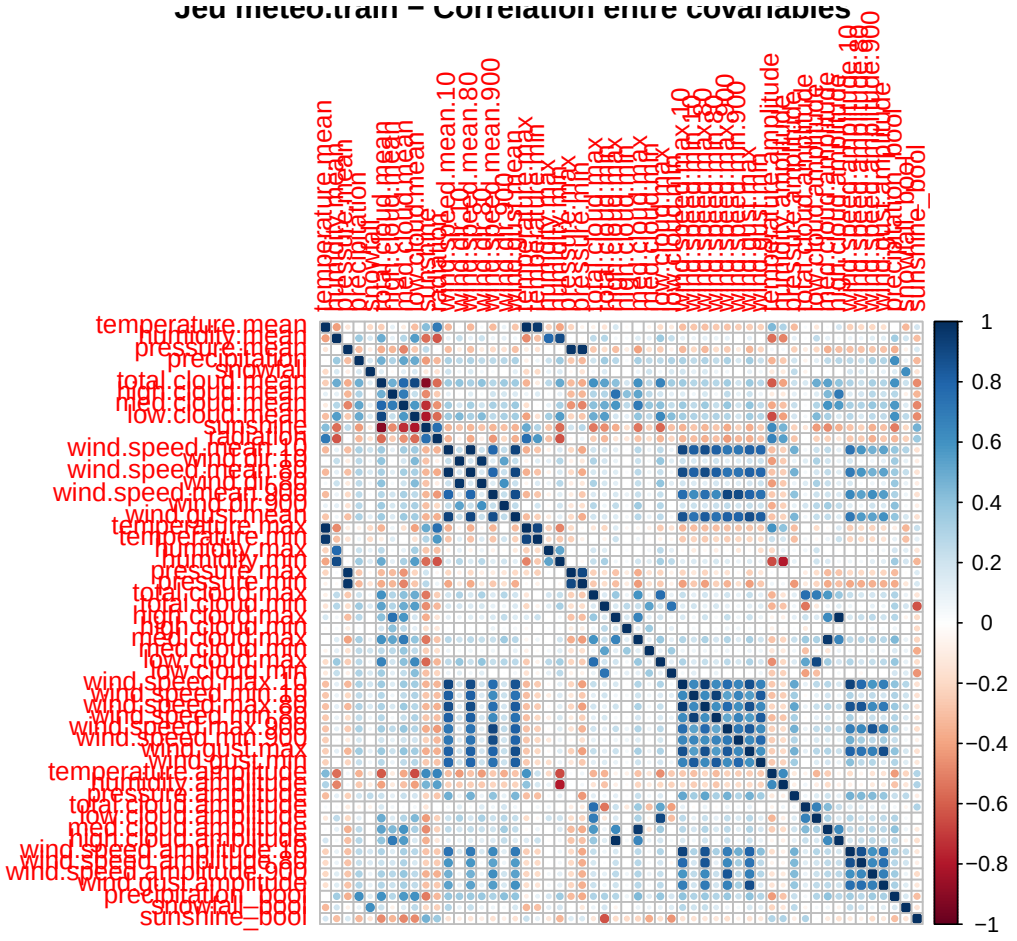
=> Le mois de l'année a une influence sur la possibilité de pluie le lendemain => Le risque de pluie est plus grand d'Avril à Septembre, de décembre à Février => Le risque de pluie est plus faible en Mars, de Septembre à Novembre

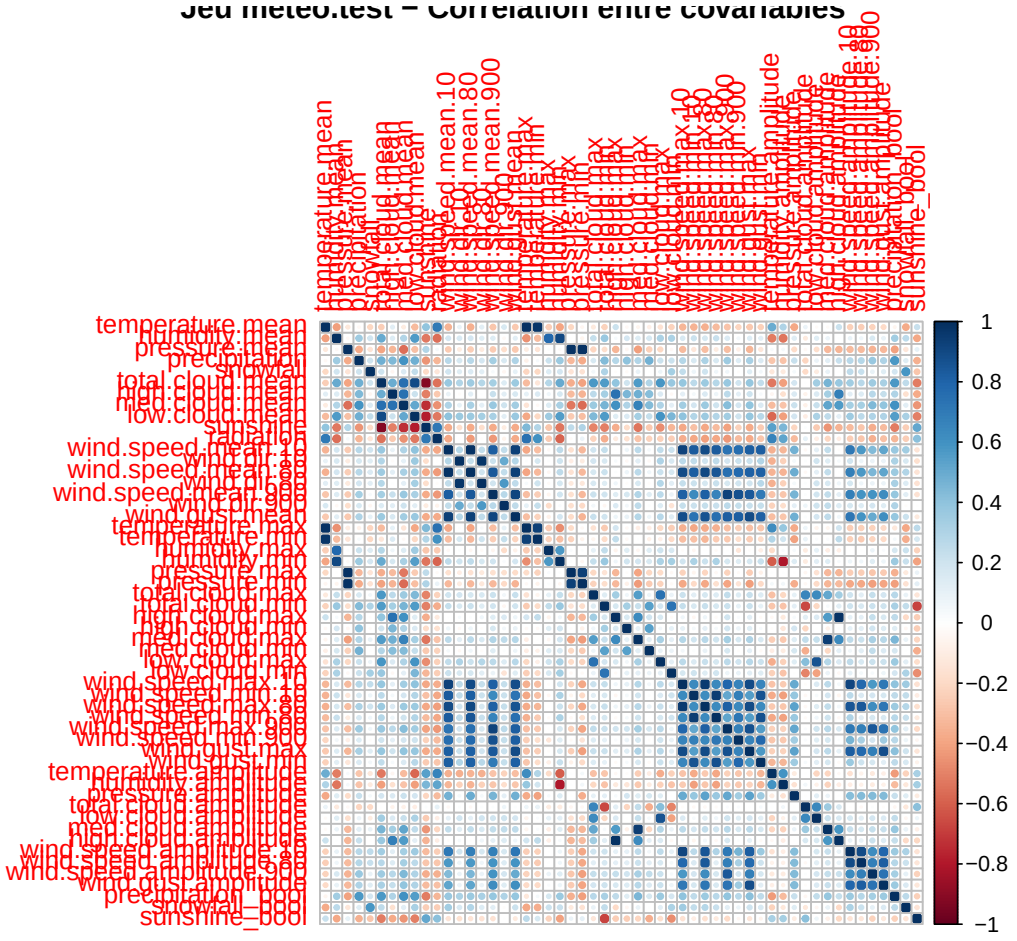
BKU : étonnant ! inversion TRUE / FALSE ? !!

Idée pour la modélisation . Inclure le mois

3.15 Colinéarité des covariables

BKU : vérifier la corrélation de toutes les variables des variables sélectionnées


Jeu meteo.train – Corrélation entre covariables
temperature.mean
humidity.mean
pressure.mean
precipitation
snowfall
total.cloud.mean
high.cloud.mean
med.cloud.mean
low.cloud.mean
sunshine
radiation
wind.speed.mean.10
wind.speed.mean.80
wind.speed.mean.900
wind.dir.mean.10
wind.dir.mean.80
wind.dir.mean.900
wind.gust.mean
temperature.max
temperature.min
humidity.max
humidity.min
pressure.max
pressure.min
total.cloud.max
total.cloud.min
high.cloud.max
high.cloud.min
med.cloud.max
med.cloud.min
low.cloud.max
low.cloud.min
wind.speed.max.10
wind.speed.min.10
wind.speed.max.80
wind.speed.min.80
wind.speed.max.900
wind.speed.min.900
wind.gust.max
wind.gust.min
temperature.amplitude
humidity.amplitude
pressure.amplitude
total.cloud.amplitude
low.cloud.amplitude
med.cloud.amplitude
high.cloud.amplitude
wind.speed.amplitude.10
wind.speed.amplitude.80
wind.speed.amplitude.900
wind.gust.amplitude
precipitation_bool
snowfall_bool
sunshine_bool
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Jeu meteo.test – Corrélation entre covariables
temperature.mean
humidity.mean
pressure.mean
precipitation
snowfall
total.cloud.mean
high.cloud.mean
med.cloud.mean
low.cloud.mean
sunshine
radiation
wind.speed.mean.10
wind.speed.mean.80
wind.dir.mean
wind.speed.mean.900
wind.dir.900
wind.gust.max
temperature.max
temperature.min
humidity.max
humidity.min
pressure.max
pressure.min
total.cloud.max
total.cloud.min
high.cloud.max
high.cloud.min
med.cloud.max
med.cloud.min
low.cloud.max
low.cloud.min
wind.speed.max.10
wind.speed.min.10
wind.speed.max.80
wind.speed.min.80
wind.speed.max.900
wind.speed.min.900
wind.gust.max
wind.gust.min
temperature.amplitude
humidity.amplitude
pressure.amplitude
total.cloud.amplitude
low.cloud.amplitude
med.cloud.amplitude
high.cloud.amplitude
wind.speed.amplitude.10
wind.speed.amplitude.80
wind.speed.amplitude.900
wind.gust.amplitude
precipitation_bool
snowfall_bool
sunshine_bool
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

4 Analyse exploratoire

4.1 Corrélation des covariables avec la variable d'intérêt

La variable d'intérêt ***pluie.demain*** est **binaire**;

Pour les **variables explicatives numériques continues**, on regarde la **distribution de la covariable** selon les modalités de la variable d'intérêt.

boxplot(x ~ y)

Pour les **variables explicatives catégorielles**, on regarde le **lien entre la covariable et la variable d'intérêt** ; en découle une **distribution des modalités de la covariable* fonction des modalités de la variable explicative.

mosaicplot(x ~ y)

4.2 Corrélation des covariables entre elles

5 Modélisation

5.1 Jeu d'entraînement et de validation

Afin d'identifier le meilleur modèle apte à la prédiction du jeu de données **meteo.test**, on va séparer le jeu de données **meteo.train** en 2 jeu de données tiré aléatoirement.

- 80% du jeu de données servira à l'ajustement des modèles : ce sera le **jeu d'entraînement**
- 20% du jeu de données servira à mesurer la capacité prédictive du modèle : ce sera le **jeu de validation**.

Pour déterminer les observations du jeu de données qui serviront à l'entraînement du modèle, on génère un vecteur **scp.train** de valeurs booléennes dont 80% valent ***TRUE*** et 20% valent ***FALSE***.

```
scp.train.size <- 0.8

scp.train = sample(c(TRUE, FALSE),
                   nrow(dat.meteo.train), replace=TRUE,
                   prob=c(scp.train.size, 1-scp.train.size))
```

A noter

Pour permettre la reproductibilité de l'entraînement / validation à chaque execution du code, le vecteur est sauvegardédans un fichier **scp.train.dat**. S'il est présent dans le répertoire d'execution du script, le fichier est chargé et utilisé. Si le fichier n'est pas présent un nouveau tirage aléatoire est effectué.

5.2 Stratégie 1 : approche naïve

L'idée de l'approche est d'appliquer une sélection "experte" des covariables.

Plusieurs constats pour réduire le nombre de covariables

- On peut regrouper les covariables par famille (Température, Vitesse du vent, Nébulosité,. . .)

  Dans ces familles, il est probable que les covariables soient corrélées et qu'on peut réduire leur nombre en identifiant un représentant, et **utiliser l'amplitude**.

- On peut aussi imaginer que des familles de covariables soient corrélées entre elles et représentent une redondance de l'information

  Par ex., il est possible que les minutes d'ensoleillement ou le rayonnement solaire soit négativement corrélés à la nébulosité.

5.2.1 Modèle complet

```
s1.res.glm.0.formula <- formula("pluie.demain ~ .")

s1.res.glm.0 <- glm(s1.res.glm.0.formula,
                    data=dat.meteo.train[scp.train,1:42],
                    family="binomial")
```

```
## pluie.demain ~ .

##
## Call:
## glm(formula = s1.res.glm.0.formula, family = "binomial", data = dat.meteo.train[scp.train,
##     1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5031  -0.7550   0.1961   0.8116   2.9117
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         73.7147110 15.0314058   4.904 9.39e-07 ***
## month2              -0.0132678  0.4429964  -0.030 0.976107
## month3              -0.8387870  0.4684515  -1.791 0.073365 .
## month4              -0.6733874  0.5583258  -1.206 0.227785
## month5              -0.4521766  0.5954122  -0.759 0.447593
## month6               0.1622394  0.6874908   0.236 0.813442
## month7              -0.0336527  0.6974775  -0.048 0.961518
## month8              -0.6828923  0.6364085  -1.073 0.283253
## month9              -1.3230103  0.5609431  -2.359 0.018347 *
## month10             -0.6254191  0.4634727  -1.349 0.177202
## month11             -0.6603483  0.4435160  -1.489 0.136515
## month12              0.3802086  0.4450745   0.854 0.392962
## temperature.mean     0.0254591  0.1971093   0.129 0.897229
## humidity.mean        0.0245098  0.0370180   0.662 0.507903
## pressure.mean        0.6175079  0.1714006   3.603 0.000315 ***
## precipitation        0.0314721  0.0339082   0.928 0.353326
## snowfall            -0.5129122  0.2942913  -1.743 0.081356 .
## total.cloud.mean     0.0063683  0.0140613   0.453 0.650623
## high.cloud.mean      0.0024756  0.0080849   0.306 0.759454
## med.cloud.mean       0.0013417  0.0079571   0.169 0.866103
## low.cloud.mean       0.0015123  0.0095200   0.159 0.873782
## sunshine             0.0002276  0.0010625   0.214 0.830406
## radiation           -0.0001361  0.0001447  -0.940 0.346966
## wind.speed.mean.10   0.0760043  0.1137102   0.668 0.503876
## wind.dir.10          0.0009734  0.0067890   0.143 0.885991
```

```
## wind.speed.mean.80  -0.1748699  0.0823707  -2.123 0.033757 *
## wind.dir.80         -0.0034128  0.0070108  -0.487 0.626405
## wind.speed.mean.900  0.0279499  0.0310723   0.900 0.368380
## wind.dir.900         0.0048660  0.0017219   2.826 0.004715 **
## wind.gust.mean       0.0216781  0.0432794   0.501 0.616451
## temperature.max      0.1759224  0.1185547   1.484 0.137838
## temperature.min     -0.1273574  0.1014711  -1.255 0.209439
## humidity.max        -0.0080884  0.0233485  -0.346 0.729028
## humidity.min        -0.0124955  0.0212603  -0.588 0.556709
## pressure.max        -0.3025472  0.0899722  -3.363 0.000772 ***
## pressure.min        -0.3920926  0.0948061  -4.136 3.54e-05 ***
## total.cloud.max      0.0036821  0.0057556   0.640 0.522338
## total.cloud.min      0.0036454  0.0079201   0.460 0.645323
## high.cloud.max       0.0025760  0.0033209   0.776 0.437931
## high.cloud.min      -0.0031764  0.0209364  -0.152 0.879410
## med.cloud.max        0.0080919  0.0036372   2.225 0.026097 *
## med.cloud.min        0.0001829  0.0107483   0.017 0.986424
## low.cloud.max        0.0037732  0.0039249   0.961 0.336376
## low.cloud.min        0.0030775  0.0084320   0.365 0.715130
## wind.speed.max.10    0.0339002  0.0411628   0.824 0.410187
## wind.speed.min.10    0.1615872  0.0765513   2.111 0.034786 *
## wind.speed.max.80    0.0209400  0.0344184   0.608 0.542924
## wind.speed.min.80   -0.0485034  0.0497587  -0.975 0.329673
## wind.speed.max.900  -0.0182251  0.0142873  -1.276 0.202090
## wind.speed.min.900  -0.0152546  0.0233454  -0.653 0.513477
## wind.gust.max        0.0093969  0.0197554   0.476 0.634316
## wind.gust.min        0.0273250  0.0325643   0.839 0.401407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.39  on 936  degrees of freedom
## Residual deviance:  932.85  on 885  degrees of freedom
## AIC: 1036.8
##
## Number of Fisher Scoring iterations: 5
```

La majorité des covariables ne sont pas identifiées comme significatives. Néanmoins, pour cette approche naïve, aucune précaution concernant la colinéarité des covariavbles n'a été prise.

Ce modèle a de fortes cchances de ne pas être très performant.

Les covariables les plus significatives sont:

- la **pression atmosphérique**
- la **nébulosité**
- la **vitesse du vent**
- Le **mois** notamment **septembre** et **mars**

5.3 STEP forward

Afin d'identifier les variables significatives et éliminées celles qui ne joueraient pas de rôle prépondérant (intrinséqment ou par colinéarité), on procède à une méthodologie Step Forward,

- depuis un modèle constant
- vers le modèle complet

```
# modèle constant initial
s1.model.constant <- glm (pluie.demain~1,
                          data=dat.meteo.train[scp.train,1:42],
                          family=binomial)

s1.model.full <- formula(s1.res.glm.0)
```

```
## pluie.demain ~ 1
```

```
## pluie.demain ~ month + temperature.mean + humidity.mean + pressure.mean +
##     precipitation + snowfall + total.cloud.mean + high.cloud.mean +
##     med.cloud.mean + low.cloud.mean + sunshine + radiation +
##     wind.speed.mean.10 + wind.dir.10 + wind.speed.mean.80 + wind.dir.80 +
##     wind.speed.mean.900 + wind.dir.900 + wind.gust.mean + temperature.max +
##     temperature.min + humidity.max + humidity.min + pressure.max +
##     pressure.min + total.cloud.max + total.cloud.min + high.cloud.max +
##     high.cloud.min + med.cloud.max + med.cloud.min + low.cloud.max +
##     low.cloud.min + wind.speed.max.10 + wind.speed.min.10 + wind.speed.max.80 +
##     wind.speed.min.80 + wind.speed.max.900 + wind.speed.min.900 +
##     wind.gust.max + wind.gust.min
```

```
## glm(formula = pluie.demain ~ med.cloud.max + pressure.min + wind.dir.900 +
##     temperature.max + wind.gust.max + total.cloud.mean + month +
##     temperature.mean + snowfall + wind.speed.min.10 + wind.speed.min.80 +
##     wind.speed.mean.80 + wind.speed.max.10 + precipitation, family = binomial,
##     data = dat.meteo.train[scp.train, 1:42])
```

Le modèle identifié par la méthode Step Forward est le suivant

```
## pluie.demain ~ med.cloud.max + pressure.min + wind.dir.900 +
##     temperature.max + wind.gust.max + total.cloud.mean + month +
##     temperature.mean + snowfall + wind.speed.min.10 + wind.speed.min.80 +
##     wind.speed.mean.80 + wind.speed.max.10 + precipitation
```

Le modèle identifié par la méthode Step Forward depuis un modèle constant présente els caractéristiques suivantes:

```
##
## Call:
## glm(formula = pluie.demain ~ med.cloud.max + pressure.min + wind.dir.900 +
##     temperature.max + wind.gust.max + total.cloud.mean + month +
##     temperature.mean + snowfall + wind.speed.min.10 + wind.speed.min.80 +
##     wind.speed.mean.80 + wind.speed.max.10 + precipitation, family = binomial,
##     data = dat.meteo.train[scp.train, 1:42])
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.5316  -0.8055   0.2699   0.8253   2.7529
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         69.065992  13.109630   5.268 1.38e-07 ***
## med.cloud.max        0.010194   0.002468   4.130 3.63e-05 ***
## pressure.min        -0.072296   0.012782  -5.656 1.55e-08 ***
## wind.dir.900         0.003705   0.001285   2.883 0.003940 **
## temperature.max      0.285948   0.079478   3.598 0.000321 ***
## wind.gust.max        0.016938   0.012362   1.370 0.170637
## total.cloud.mean     0.014694   0.004272   3.439 0.000583 ***
## month2              -0.191483   0.416726  -0.459 0.645880
## month3              -1.032922   0.408546  -2.528 0.011462 *
## month4              -0.986669   0.446120  -2.212 0.026990 *
## month5              -0.795879   0.472379  -1.685 0.092021 .
## month6              -0.370048   0.518381  -0.714 0.475318
## month7              -0.451179   0.556900  -0.810 0.417847
## month8              -1.033511   0.540745  -1.911 0.055970 .
## month9              -1.609385   0.496883  -3.239 0.001200 **
## month10             -0.760179   0.434232  -1.751 0.080010 .
## month11             -0.682064   0.421845  -1.617 0.105909
## month12              0.391751   0.418021   0.937 0.348678
## temperature.mean    -0.213112   0.087997  -2.422 0.015444 *
## snowfall            -0.593192   0.266392  -2.227 0.025963 *
## wind.speed.min.10    0.193304   0.063395   3.049 0.002295 **
## wind.speed.min.80   -0.053175   0.044185  -1.203 0.228794
## wind.speed.mean.80  -0.097397   0.038286  -2.544 0.010961 *
## wind.speed.max.10    0.051304   0.027353   1.876 0.060706 .
## precipitation        0.039186   0.027330   1.434 0.151631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.39  on 936  degrees of freedom
## Residual deviance:  958.07  on 912  degrees of freedom
## AIC: 1008.1
##
## Number of Fisher Scoring iterations: 5
```

5.4 Stratégie 4 : approche exploratoire

5.5 Résultats

Table 1: Résumé des modèles

x
stratégie 1
pluie.demain ~ month + temperature.mean + humidity.mean + pressure.mean + precipitation + snowfall + total.cloud.mean + high.cloud.mean + med.cloud.mean + low.cloud.mean + sunshine + radiation + wind.speed.mean.10 + wind.dir.10 + wind.speed.mean.80 + wind.dir.80 + wind.speed.mean.900 + wind.dir.900 + wind.gust.mean + temperature.max + temperature.min + humidity.max + humidity.min + pressure.max + pressure.min + total.cloud.max + total.cloud.min + high.cloud.max + high.cloud.min + med.cloud.max + med.cloud.min + low.cloud.max + low.cloud.min + wind.speed.max.10 + wind.speed.min.10 + wind.speed.max.80 + wind.speed.min.80 + wind.speed.max.900 + wind.speed.min.900 + wind.gust.max + wind.gust.min
pluie.demain ~ med.cloud.max + pressure.min + wind.dir.900 + temperature.max + wind.gust.max + total.cloud.mean + month + temperature.mean + snowfall + wind.speed.min.10 + wind.speed.min.80 + wind.speed.mean.80 + wind.speed.max.10 + precipitation
pluie.demain ~ med.cloud.max + pressure.min + wind.dir.900 + temperature.max + total.cloud.mean + month + temperature.mean + snowfall + wind.speed.min.10 + wind.speed.mean.80 + wind.speed.max.10 + precipitation
pluie.demain ~ month + pressure.mean + precipitation + snowfall + total.cloud.mean + wind.speed.mean.80 + wind.dir.900 + temperature.max + temperature.min + pressure.max + pressure.min + med.cloud.max + wind.speed.max.10 + wind.speed.min.10
pluie.demain ~ month + pressure.mean + precipitation + snowfall + total.cloud.mean + wind.speed.mean.80 + wind.dir.900 + temperature.max + temperature.min + pressure.max + pressure.min + med.cloud.max + wind.speed.max.10 + wind.speed.min.10 + wind.speed.min.80 + wind.gust.mean
stratégie 4
pluie.demain ~ month + temperature.amplitude * temperature.min + humidity.amplitude * humidity.max + pressure.amplitude * pressure.max + total.cloud.amplitude * total.cloud.mean + low.cloud.amplitude * low.cloud.min + med.cloud.amplitude * med.cloud.min + high.cloud.amplitude * high.cloud.min + wind.speed.amplitude.10 * wind.speed.min.10 * wind.dir.10 + wind.speed.amplitude.80 * wind.speed.min.80 * wind.dir.80 + wind.speed.amplitude.900 * wind.speed.min.900 * wind.dir.900 + wind.gust.amplitude * wind.gust.min + precipitation + precipitation_bool + snowfall + snowfall_bool + sunshine + sunshine_bool + radiation
pluie.demain ~ med.cloud.amplitude + pressure.max + precipitation_bool + high.cloud.amplitude + pressure.amplitude + temperature.min + month + snowfall + wind.speed.min.10 + temperature.amplitude + total.cloud.mean + wind.dir.900 + wind.speed.min.80 + sunshine_bool + snowfall_bool + low.cloud.amplitude + low.cloud.min
pluie.demain ~ med.cloud.amplitude + pressure.max + precipitation_bool + high.cloud.amplitude + pressure.amplitude + temperature.min + month + snowfall + wind.speed.min.10 + temperature.amplitude + wind.dir.900 + wind.speed.min.80 + sunshine_bool + snowfall_bool + low.cloud.amplitude + low.cloud.min + med.cloud.min + pressure.max:pressure.amplitude + med.cloud.amplitude:med.cloud.min
pluie.demain ~ month + temperature.amplitude + temperature.min + pressure.amplitude + pressure.max + low.cloud.amplitude + low.cloud.min + med.cloud.amplitude + med.cloud.min + high.cloud.amplitude + wind.speed.min.10 + wind.dir.10 + wind.speed.min.80 + wind.dir.900 + wind.gust.amplitude + precipitation_bool + snowfall + snowfall_bool + sunshine_bool + pressure.amplitude:pressure.max + med.cloud.amplitude:med.cloud.min + wind.speed.min.10:wind.dir.10
pluie.demain ~ month + temperature.amplitude + temperature.min + pressure.amplitude + pressure.max + low.cloud.amplitude + low.cloud.min + med.cloud.amplitude + med.cloud.min + high.cloud.amplitude + wind.speed.min.10 + wind.dir.10 + wind.speed.min.80 + wind.dir.900 + wind.gust.amplitude + precipitation_bool + snowfall + snowfall_bool + sunshine_bool + pressure.amplitude:pressure.max + med.cloud.amplitude:med.cloud.min + wind.speed.min.10:wind.dir.10

Table 2: Résumé des scores

	nb.covariables	nb.coefficients	aic	deviance.test.mk.vers.m0	deviance.test.m0.vers.mk	seuil.optimal	precision	auc	erreur
stratégie 1									
s1.res.glm.0	41	52	1036.8481	0.1285630	0	0.52	0.7283951	0.7647338	0.3609034
	14	25	1008.0707	0.1408232	0	0.57	0.7283951	0.7713047	0.3638729
s1.res.step__forward									
	12	23	1007.3372	0.1347659	0	0.63	0.7325103	0.7680531	0.3660824
s1.res.step__both__from__constant									
	14	25	999.0086	0.1919723	0	0.55	0.7407407	0.7740821	0.3594378
s1.res.step__backward									
	16	27	998.9853	0.2045818	0	0.52	0.7325103	0.7730660	0.3589099
s1.res.step__both__from__full									
stratégie 4									
s4.res.glm.0	33	64	1055.4534	0.0980230	0	0.46	0.7325103	0.7799079	0.3523774
	17	28	1011.2045	0.1397635	0	0.63	0.7201646	0.7741498	0.3624264
s4.res.step__forward									
	17	30	1010.1197	0.1556754	0	0.69	0.7201646	0.7692047	0.3639715
s4.res.step__both__from__constant									
	19	33	1002.3126	0.2216322	0	0.55	0.7366255	0.7788917	0.3548398
s4.res.step__backward									
	19	33	1002.3126	0.2216322	0	0.55	0.7366255	0.7788917	0.3548398
s4.res.step__both__from__full									