Projet

Modèle linéaire généralise et Choix de modèles

Réalisé par Asma GHARIANI

09/01/2022

Table des matières

Introduction

L'objectif principal de cette étude est de choisir et valider un modèle. Ensuite nous réalisons des prédictions afin de choisir le modèle le plus adéquat. Nous nous basons sur une base de données contenant 1180 observations,47 variables quantitative et qualitatives (Pluie.demain=f (Tempmean, Humimean, MeanPressuremean Totalprecipitation, etc.).

Nous mobilisons plusieurs démarches que nous exposons en ce qui suit :

1-Analyse exploratoire

2-Vérification des liens entre les variables

3-Identification des prédicteurs les plus importants afin de construire un modèle valide pour faire une bonne prédiction

4-Estimation de modèle sur un échantillon d'apprentissage

5-Validation de modèle sur un échantillon test

6-Comparaison entre plusieurs modèles et choix du modèle le plus adéquat

7-Proposition d'une prédiction pour l'échantillon test

Nous utilisons les abréviations suivantes :

Tempmean =Temperature.daily.mean..2.m.above.gnd.

Humimean =Relative.Humidity.daily.mean..2.m.above.gnd.

MeanPressuremean = Mean.Sea.Level.Pressure.daily.mean..MSL.
#Totalprecipitation = Total.Precipitation.daily.sum..sfc.
#Snowfall =Snowfall.amount.raw.daily.sum..sfc.
#Totalcloudmean = Total.Cloud.Cover.daily.mean..sfc.
#Highcloudmean = High.Cloud.Cover.daily.mean..high.cld.lay.
Mediumcloudmean = Medium.Cloud.Cover.daily.mean..mid.cld.lay. #Lowcloudmean =Low.Cloud.Cover.daily.mean..low.cld.lay.
#Sunshine = Sunshine.Duration.daily.sum..sfc.
#Waveradia = Shortwave.Radiation.daily.sum..sfc.
Windspdmean10m = Wind.Speed.daily.mean..10.m.above.gnd.
Winddirecmean10m =Wind.Direction.daily.mean..10.m.above.gnd.
#Windspdmean80m = Wind.Speed.daily.mean..80.m.above.gnd.

Winddirectmean80m =Wind.Direction.daily.mean..80.m.above.gnd.
Windspdmean900mb = Wind.Speed.daily.mean..900.mb.
Winddirectmean900mb =Wind.Direction.daily.mean..900.mb.
#Windgustmean = Wind.Gust.daily.mean..sfc.
Tempmax =Temperature.daily.max..2.m.above.gnd.
#Tempmin =Temperature.daily.min..2.m.above.gnd.
#Humimax =Relative.Humidity.daily.max..2.m.above.gnd. #Humimin =
Relative.Humidity.daily.min..2.m.above.gnd. # Meanpressuremax
=Mean.Sea.Level.Pressure.daily.max..MSL.
#Meanpressuremin =Mean.Sea.Level.Pressure.daily.min..MSL.
Totalcloudmax =Total.Cloud.Cover.daily.max..sfc.
#Totalcloudmin =Total.Cloud.Cover.daily.min..sfc.
#Highcloudmax= High.Cloud.Cover.daily.max..high.cld.lay.
Highcloudmin = High.Cloud.Cover.daily.min..high.cld.lay.
#Mediumcloudmax =Medium.Cloud.Cover.daily.max..mid.cld.lay.
#Mediumcloudmin= Medium.Cloud.Cover.daily.min..mid.cld.lay.
# Lowcloudmax=Low.Cloud.Cover.daily.max..low.cld.lay.
Lowcloudmin= Low.Cloud.Cover.daily.min..low.cld.lay.
#Windspdmax10m=Wind.Speed.daily.max..10.m.above.gnd.
#Windspdmin10m= Wind.Speed.daily.min..10.m.above.gnd.
Windspdmax80m= Wind.Speed.daily.max..80.m.above.gnd.
#Windspdmin80m=Wind.Speed.daily.min..80.m.above.gnd.
#Windspdmax900mb= Wind.Speed.daily.max..900.mb.
#Windspdmin900mb= Wind.Speed.daily.min..900.mb.
#Windgustmax= Wind.Gust.daily.max..sfc.
#Windgustmin= Wind.Gust.daily.min..sfc.

1. Analyse exploratoire des données

1.1 Chargement des données

```
data=read.csv("meteo.train.csv",na.strings = "" )
data1=read.csv("meteo.test.csv" )
summary(data1) ## "meteo.test.csv"
```

```
       X              Year          Month             Day              Hour       Minute
 Min.   :  36   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :0   Min.   :0
 1st Qu.: 825   1st Qu.:2012   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:0   1st Qu.:0
 Median :1554   Median :2014   Median : 7.000   Median :16.00   Median :0   Median :0
 Mean   :1517   Mean   :2014   Mean   : 6.859   Mean   :15.26   Mean   :0   Mean   :0
 3rd Qu.:2252   3rd Qu.:2016   3rd Qu.:10.000   3rd Qu.:22.00   3rd Qu.:0   3rd Qu.:0
 Max.   :2936   Max.   :2018   Max.   :12.000   Max.   :31.00   Max.   :0   Max.   :0
 Temperature.daily.mean..2.m.above.gnd. Relative.Humidity.daily.mean..2.m.above.gnd.
 Min.   :-5.040                         Min.   :36.42
 1st Qu.: 5.625                         1st Qu.:64.35
 Median :13.085                         Median :71.64
 Mean   :12.062                         Mean   :71.08
 3rd Qu.:18.170                         3rd Qu.:78.06
 Max.   :26.600                         Max.   :91.46
 Mean.Sea.Level.Pressure.daily.mean..MSL. Total.Precipitation.daily.sum..sfc. Snowfall.amount.raw.daily.sum..sfc.
 Min.   : 983                             Min.   : 0.000                      Min.   :0.00000
 1st Qu.:1012                             1st Qu.: 0.000                      1st Qu.:0.00000
 Median :1017                             Median : 0.100                      Median :0.00000
 Mean   :1017                             Mean   : 2.202                      Mean   :0.08255
 3rd Qu.:1022                             3rd Qu.: 2.475                      3rd Qu.:0.00000
 Max.   :1046                             Max.   :39.300                      Max.   :6.65000
```

```
 Total.Cloud.Cover.daily.mean..sfc. High.Cloud.Cover.daily.mean..high.cld.lay.
 Min.   :  0.00                     Min.   : 0.000
 1st Qu.: 30.20                     1st Qu.: 1.603
 Median : 52.30                     Median :12.210
 Mean   : 53.62                     Mean   :20.451
 3rd Qu.: 79.60                     3rd Qu.:34.358
 Max.   :100.00                     Max.   :99.330
 Medium.Cloud.Cover.daily.mean..mid.cld.lay. Low.Cloud.Cover.daily.mean..low.cld.lay.
 Min.   :  0.00                              Min.   :  0.00
 1st Qu.:  3.23                              1st Qu.: 12.55
 Median : 26.09                              Median : 39.65
 Mean   : 32.66                              Mean   : 41.75
 3rd Qu.: 51.34                              3rd Qu.: 66.50
 Max.   :100.00                              Max.   :100.00
 Sunshine.Duration.daily.sum..sfc. Shortwave.Radiation.daily.sum..sfc. Wind.Speed.daily.mean..10.m.above.gnd.
 Min.   :   0.0                    Min.   : 121.9                      Min.   : 3.03
 1st Qu.: 116.7                    1st Qu.:1946.0                      1st Qu.: 6.35
 Median : 347.6                    Median :3391.8                      Median : 9.35
 Mean   : 353.0                    Mean   :3813.1                      Mean   :11.23
 3rd Qu.: 560.1                    3rd Qu.:5509.5                      3rd Qu.:13.47
 Max.   :1021.1                    Max.   :8259.2                      Max.   :49.14
 Wind.Direction.daily.mean..10.m.above.gnd. Wind.Speed.daily.mean..80.m.above.gnd.
 Min.   : 45.08                             Min.   : 2.990
 1st Qu.:151.67                             1st Qu.: 8.082
 Median :206.38                             Median :12.565
 Mean   :201.27                             Mean   :14.725
 3rd Qu.:250.45                             3rd Qu.:18.340
 Max.   :326.87                             Max.   :63.100
```

```
Wind.Direction.daily.mean..80.m.above.gnd. Wind.Speed.daily.mean..900.mb. Wind.Direction.daily.mean..900.mb.
Min.   : 37.18                             Min.   :  2.16                 Min.   : 32.09
1st Qu.:159.47                             1st Qu.: 12.58                 1st Qu.:165.43
Median :217.01                             Median : 21.76                 Median :239.06
Mean   :207.55                             Mean   : 25.57                 Mean   :211.65
3rd Qu.:256.15                             3rd Qu.: 34.66                 3rd Qu.:265.90
Max.   :330.47                             Max.   :104.89                 Max.   :333.05
Wind.Gust.daily.mean..sfc. Temperature.daily.max..2.m.above.gnd. Temperature.daily.min..2.m.above.gnd.
Min.   : 3.290             Min.   :-2.130                        Min.   :-8.870
1st Qu.: 9.293             1st Qu.: 9.295                        1st Qu.: 2.458
Median :14.445             Median :17.145                        Median : 8.595
Mean   :17.591             Mean   :16.231                        Mean   : 8.022
3rd Qu.:22.065             3rd Qu.:22.823                        3rd Qu.:13.367
Max.   :90.750             Max.   :33.120                        Max.   :21.890
Relative.Humidity.daily.max..2.m.above.gnd. Relative.Humidity.daily.min..2.m.above.gnd.
Min.   : 63.00                              Min.   :17.00
1st Qu.: 83.00                              1st Qu.:45.00
Median : 88.00                              Median :54.00
Mean   : 87.47                              Mean   :53.73
3rd Qu.: 94.00                              3rd Qu.:62.00
Max.   :100.00                              Max.   :88.00
Mean.Sea.Level.Pressure.daily.max..MSL. Mean.Sea.Level.Pressure.daily.min..MSL. Total.Cloud.Cover.daily.max..sfc.
Min.   : 987.3                          Min.   : 980.6                          Min.   :  0.00
1st Qu.:1015.1                          1st Qu.:1009.1                          1st Qu.:100.00
Median :1019.3                          Median :1014.9                          Median :100.00
Mean   :1019.7                          Mean   :1014.1                          Mean   : 91.36
3rd Qu.:1024.5                          3rd Qu.:1019.7                          3rd Qu.:100.00
Max.   :1049.0                          Max.   :1040.5                          Max.   :100.00
```

```
Total.Cloud.Cover.daily.min..sfc. High.Cloud.Cover.daily.max..high.cld.lay. High.Cloud.Cover.daily.min..high.cld.lay.
Min.   :  0.000                   Min.   :  0.00                           Min.   : 0.000
1st Qu.:  0.000                   1st Qu.: 16.25                           1st Qu.: 0.000
Median :  0.000                   Median : 71.00                           Median : 0.000
Mean   : 10.103                   Mean   : 58.67                           Mean   : 1.169
3rd Qu.:  2.925                   3rd Qu.:100.00                           3rd Qu.: 0.000
Max.   :100.000                   Max.   :100.00                           Max.   :84.000
Medium.Cloud.Cover.daily.max..mid.cld.lay. Medium.Cloud.Cover.daily.min..mid.cld.lay.
Min.   :  0.00                             Min.   :  0.000
1st Qu.: 41.25                             1st Qu.:  0.000
Median :100.00                             Median :  0.000
Mean   : 74.00                             Mean   :  2.779
3rd Qu.:100.00                             3rd Qu.:  0.000
Max.   :100.00                             Max.   :100.000
Low.Cloud.Cover.daily.max..low.cld.lay. Low.Cloud.Cover.daily.min..low.cld.lay. Wind.Speed.daily.max..10.m.above.gnd.
Min.   :  0.00                          Min.   :  0.000                         Min.   : 4.35
1st Qu.:100.00                          1st Qu.:  0.000                         1st Qu.:12.44
Median :100.00                          Median :  0.000                         Median :17.31
Mean   : 84.48                          Mean   :  4.745                         Mean   :19.67
3rd Qu.:100.00                          3rd Qu.:  0.000                         3rd Qu.:23.90
Max.   :100.00                          Max.   :100.000                         Max.   :74.34
Wind.Speed.daily.min..10.m.above.gnd. Wind.Speed.daily.max..80.m.above.gnd. Wind.Speed.daily.min..80.m.above.gnd.
Min.   : 0.000                        Min.   : 5.82                         Min.   : 0.000
1st Qu.: 1.080                        1st Qu.:17.93                         1st Qu.: 1.095
Median : 2.465                        Median :24.14                         Median : 2.520
Mean   : 3.906                        Mean   :26.08                         Mean   : 5.068
3rd Qu.: 5.008                        3rd Qu.:30.95                         3rd Qu.: 6.145
Max.   :33.840                        Max.   :94.46                         Max.   :43.210
```

```
Wind.Speed.daily.max..900.mb. Wind.Speed.daily.min..900.mb. Wind.Gust.daily.max..sfc. Wind.Gust.daily.min..sfc.
Min.   :  6.92                Min.   : 0.00                 Min.   :  5.76            Min.   : 0.000
1st Qu.: 25.36                1st Qu.: 3.34                 1st Qu.: 18.00            1st Qu.: 1.800
Median : 39.55                Median : 6.92                 Median : 26.64            Median : 3.960
Mean   : 43.20                Mean   :11.66                 Mean   : 30.30            Mean   : 6.825
3rd Qu.: 55.47                3rd Qu.:15.58                 3rd Qu.: 38.43            3rd Qu.: 8.190
Max.   :148.95                Max.   :81.12                 Max.   :133.20            Max.   :54.720
```

summary(data) "meteo.train.csv"

```
       X               Year          Month             Day             Hour       Minute
 Min.   :   2.0   Min.   :2010   Min.   : 1.000   Min.   : 1.0   Min.   :0   Min.   :0
 1st Qu.: 721.5   1st Qu.:2012   1st Qu.: 3.000   1st Qu.: 8.0   1st Qu.:0   1st Qu.:0
 Median :1451.0   Median :2014   Median : 6.000   Median :16.0   Median :0   Median :0
 Mean   :1459.8   Mean   :2014   Mean   : 6.436   Mean   :15.8   Mean   :0   Mean   :0
 3rd Qu.:2189.0   3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:23.0   3rd Qu.:0   3rd Qu.:0
 Max.   :2940.0   Max.   :2018   Max.   :12.000   Max.   :31.0   Max.   :0   Max.   :0
 Temperature.daily.mean..2.m.above.gnd. Relative.Humidity.daily.mean..2.m.above.gnd.
 Min.   :-7.63                          Min.   :38.33
 1st Qu.: 6.71                          1st Qu.:64.82
 Median :12.08                          Median :72.21
 Mean   :12.23                          Mean   :71.40
 3rd Qu.:17.54                          3rd Qu.:78.63
 Max.   :29.45                          Max.   :95.54
 Mean.Sea.Level.Pressure.daily.mean..MSL. Total.Precipitation.daily.sum..sfc. Snowfall.amount.raw.daily.sum..sfc.
 Min.   : 978.9                           Min.   : 0.000                      Min.   :0.00000
 1st Qu.:1012.4                           1st Qu.: 0.000                      1st Qu.:0.00000
 Median :1017.0                           Median : 0.100                      Median :0.00000
 Mean   :1017.0                           Mean   : 2.085                      Mean   :0.04965
 3rd Qu.:1022.0                           3rd Qu.: 2.300                      3rd Qu.:0.00000
 Max.   :1042.4                           Max.   :31.500                      Max.   :8.61000
```

```
Total.Cloud.Cover.daily.mean..sfc. High.Cloud.Cover.daily.mean..high.cld.lay.
Min.   :  0.00                     Min.   :  0.000
1st Qu.: 23.80                     1st Qu.:  1.657
Median : 51.67                     Median : 11.880
Mean   : 50.76                     Mean   : 20.284
3rd Qu.: 78.53                     3rd Qu.: 33.260
Max.   :100.00                     Max.   :100.000
Medium.Cloud.Cover.daily.mean..mid.cld.lay. Low.Cloud.Cover.daily.mean..low.cld.lay.
Min.   :  0.00                               Min.   :  0.00
1st Qu.:  1.83                               1st Qu.:  9.42
Median : 24.98                               Median : 36.35
Mean   : 31.50                               Mean   : 39.34
3rd Qu.: 54.21                               3rd Qu.: 65.76
Max.   :100.00                               Max.   :100.00
Sunshine.Duration.daily.sum..sfc. Shortwave.Radiation.daily.sum..sfc. Wind.Speed.daily.mean..10.m.above.gnd.
Min.   :   0.0                    Min.   : 265.2                     Min.   : 1.260
1st Qu.: 114.3                    1st Qu.:2096.2                     1st Qu.: 6.428
Median : 366.8                    Median :3675.3                     Median : 9.195
Mean   : 373.1                    Mean   :3984.6                     Mean   :10.707
3rd Qu.: 587.7                    3rd Qu.:5723.6                     3rd Qu.:12.977
Max.   :1015.8                    Max.   :8363.3                     Max.   :42.210
Wind.Direction.daily.mean..10.m.above.gnd. Wind.Speed.daily.mean..80.m.above.gnd.
Min.   : 11.19                             Min.   : 1.34
1st Qu.:152.40                             1st Qu.: 8.68
Median :206.36                             Median :12.41
Mean   :201.82                             Mean   :14.28
3rd Qu.:254.19                             3rd Qu.:17.61
Max.   :331.67                             Max.   :54.03
```

```
Wind.Direction.daily.mean..80.m.above.gnd. Wind.Speed.daily.mean..900.mb. Wind.Direction.daily.mean..900.mb.
Min.   : 12.18                             Min.   : 2.25                  Min.   : 17.37
1st Qu.:157.42                             1st Qu.:13.02                  1st Qu.:144.02
Median :213.78                             Median :19.57                  Median :233.47
Mean   :206.23                             Mean   :24.57                  Mean   :206.22
3rd Qu.:259.06                             3rd Qu.:32.10                  3rd Qu.:265.93
Max.   :333.43                             Max.   :97.06                  Max.   :344.82
Wind.Gust.daily.mean..sfc. Temperature.daily.max..2.m.above.gnd. Temperature.daily.min..2.m.above.gnd.
Min.   : 2.25              Min.   :-3.84                         Min.   :-12.520
1st Qu.: 9.48              1st Qu.:10.58                         1st Qu.:  3.350
Median :14.06              Median :16.54                         Median :  8.005
Mean   :16.69              Mean   :16.54                         Mean   :  8.062
3rd Qu.:21.15              3rd Qu.:22.36                         3rd Qu.: 13.092
Max.   :79.38              Max.   :35.77                         Max.   : 23.940
Relative.Humidity.daily.max..2.m.above.gnd. Relative.Humidity.daily.min..2.m.above.gnd.
Min.   : 59.00                              Min.   :19.00
1st Qu.: 83.00                              1st Qu.:45.00
Median : 89.00                              Median :54.00
Mean   : 87.69                              Mean   :54.04
3rd Qu.: 94.00                              3rd Qu.:63.00
Max.   :100.00                              Max.   :92.00
Mean.Sea.Level.Pressure.daily.max..MSL. Mean.Sea.Level.Pressure.daily.min..MSL. Total.Cloud.Cover.daily.max..sfc.
Min.   : 981.9                          Min.   : 977                            Min.   :  0.00
1st Qu.:1015.4                          1st Qu.:1009                            1st Qu.:100.00
Median :1019.5                          Median :1015                            Median :100.00
Mean   :1019.9                          Mean   :1014                            Mean   : 88.23
3rd Qu.:1024.7                          3rd Qu.:1019                            3rd Qu.:100.00
Max.   :1045.4                          Max.   :1039                            Max.   :100.00
```

```
 Total.Cloud.Cover.daily.min..sfc. High.Cloud.Cover.daily.max..high.cld.lay. High.Cloud.Cover.daily.min..high.cld.lay.
 Min.   :  0.000                  Min.   :  0.00                            Min.   :  0.0000
 1st Qu.:  0.000                  1st Qu.: 15.00                            1st Qu.:  0.0000
 Median :  0.000                  Median : 97.00                            Median :  0.0000
 Mean   :  8.692                  Mean   : 60.17                            Mean   :  0.9432
 3rd Qu.:  2.400                  3rd Qu.:100.00                            3rd Qu.:  0.0000
 Max.   :100.000                  Max.   :100.00                            Max.   :100.0000
 Medium.Cloud.Cover.daily.max..mid.cld.lay. Medium.Cloud.Cover.daily.min..mid.cld.lay.
 Min.   :  0.00                             Min.   :  0.000
 1st Qu.: 22.75                             1st Qu.:  0.000
 Median :100.00                             Median :  0.000
 Mean   : 70.94                             Mean   :  2.097
 3rd Qu.:100.00                             3rd Qu.:  0.000
 Max.   :100.00                             Max.   :100.000
 Low.Cloud.Cover.daily.max..low.cld.lay. Low.Cloud.Cover.daily.min..low.cld.lay. Wind.Speed.daily.max..10.m.above.gnd.
 Min.   :  0                             Min.   :  0.000                        Min.   : 2.52
 1st Qu.:100                             1st Qu.:  0.000                        1st Qu.:12.32
 Median :100                             Median :  0.000                        Median :17.36
 Mean   : 80                             Mean   :  3.879                        Mean   :19.06
 3rd Qu.:100                             3rd Qu.:  0.000                        3rd Qu.:23.44
 Max.   :100                             Max.   :100.000                        Max.   :79.99
 Wind.Speed.daily.min..10.m.above.gnd. Wind.Speed.daily.max..80.m.above.gnd. Wind.Speed.daily.min..80.m.above.gnd.
 Min.   : 0.00                         Min.   : 3.98                         Min.   : 0.000
 1st Qu.: 1.14                         1st Qu.:18.27                         1st Qu.: 1.140
 Median : 2.41                         Median :23.85                         Median : 2.600
 Mean   : 3.57                         Mean   :25.35                         Mean   : 4.727
 3rd Qu.: 4.45                         3rd Qu.:29.92                         3rd Qu.: 5.830
 Max.   :27.73                         Max.   :93.84                         Max.   :37.700
```

```
 Wind.Speed.daily.max..900.mb. Wind.Speed.daily.min..900.mb. Wind.Gust.daily.max..sfc. Wind.Gust.daily.min..sfc.
 Min.   :  4.02                Min.   : 0.00                 Min.   : 4.32            Min.   : 0.000
 1st Qu.: 24.54                1st Qu.: 3.05                 1st Qu.:19.08            1st Qu.: 2.160
 Median : 37.12                Median : 6.73                 Median :26.10            Median : 3.960
 Mean   : 41.82                Mean   :11.09                 Mean   :29.31            Mean   : 6.502
 3rd Qu.: 54.37                3rd Qu.:15.31                 3rd Qu.:37.08            3rd Qu.: 8.280
 Max.   :136.25                Max.   :76.13                 Max.   :95.04            Max.   :57.960
 pluie.demain
 Mode :logical
 FALSE:579
 TRUE :601
```

```
str(data)

## 'data.frame':    1180 obs. of  47 variables:
##  $ X                                           : int  2 4 6 8 10 12 14 1
6 18 20 ...
##  $ Year                                        : int  2010 2010 2010 201
0 2010 2010 2010 2010 2010 2010 ...
##  $ Month                                       : int  6 6 6 6 6 6 6 6 6
6 ...
##  $ Day                                         : int  2 4 6 8 10 12 14 1
6 18 20 ...
##  $ Hour                                        : int  0 0 0 0 0 0 0 0 0
0 ...
##  $ Minute                                      : int  0 0 0 0 0 0 0 0 0
0 ...
##  $ Temperature.daily.mean..2.m.above.gnd.      : num  15 17.3 21.6 20.2
22.6 ...
##  $ Relative.Humidity.daily.mean..2.m.above.gnd.: num  76.5 77.6 69.5 75.
1 73.5 ...
##  $ Mean.Sea.Level.Pressure.daily.mean..MSL.    : num  1015 1017 1015 100
8 1004 ...
##  $ Total.Precipitation.daily.sum..sfc.         : num  1 0 3.7 0.2 0 2.2
1.8 1.8 17.5 1.2 ...
##  $ Snowfall.amount.raw.daily.sum..sfc.         : num  0 0 0 0 0 0 0 0 0
0 ...
##  $ Total.Cloud.Cover.daily.mean..sfc.          : num  79.8 4.7 42.1 67.5
56.3 ...
##  $ High.Cloud.Cover.daily.mean..high.cld.lay.  : num  3 0.67 21.21 54.71
```

```
50.25 ...
##  $ Medium.Cloud.Cover.daily.mean..mid.cld.lay. : num  31.6 0 25.9 65.8 5
5.3 ...
##  $ Low.Cloud.Cover.daily.mean..low.cld.lay.    : num  79.2 4.5 35.3 18.9
34.2 ...
##  $ Sunshine.Duration.daily.sum..sfc.           : num  287.2 821.4 441.3
41.9 473.2 ...
##  $ Shortwave.Radiation.daily.sum..sfc.         : num  6710 7974 4834 539
0 7216 ...
##  $ Wind.Speed.daily.mean..10.m.above.gnd.      : num  11.64 6.34 8.4 5.4
9.16 ...
##  $ Wind.Direction.daily.mean..10.m.above.gnd.  : num  275 230 215 205 17
9 ...
##  $ Wind.Speed.daily.mean..80.m.above.gnd.      : num  14.99 8.92 10.38 6
.53 11.91 ...
##  $ Wind.Direction.daily.mean..80.m.above.gnd.  : num  268 199 208 206 18
6 ...
##  $ Wind.Speed.daily.mean..900.mb.              : num  20.6 27.9 18.9 10.
4 21.9 ...
##  $ Wind.Direction.daily.mean..900.mb.          : num  180.4 93.7 250.1 2
38.6 153 ...
##  $ Wind.Gust.daily.mean..sfc.                  : num  14.88 9.48 13.5 5.
31 12.21 ...
##  $ Temperature.daily.max..2.m.above.gnd.       : num  18.5 25 26.2 24.2
30.7 ...
##  $ Temperature.daily.min..2.m.above.gnd.       : num  11.1 10.4 17.7 14.
7 16.9 ...
##  $ Relative.Humidity.daily.max..2.m.above.gnd. : int  94 92 91 89 97 92
96 96 97 95 ...
##  $ Relative.Humidity.daily.min..2.m.above.gnd. : int  59 54 57 62 39 65
69 64 74 61 ...
##  $ Mean.Sea.Level.Pressure.daily.max..MSL.     : num  1017 1019 1016 101
0 1006 ...
##  $ Mean.Sea.Level.Pressure.daily.min..MSL.     : num  1014 1016 1013 100
6 1001 ...
##  $ Total.Cloud.Cover.daily.max..sfc.           : num  100 28 100 100 100
100 100 100 100 100 ...
##  $ Total.Cloud.Cover.daily.min..sfc.           : num  0 0 0 0 0 0 0 100
0 0 ...
##  $ High.Cloud.Cover.daily.max..high.cld.lay.   : int  16 11 100 100 100
28 100 100 100 24 ...
##  $ High.Cloud.Cover.daily.min..high.cld.lay.   : int  0 0 0 0 0 0 0 0 0
0 ...
##  $ Medium.Cloud.Cover.daily.max..mid.cld.lay.  : int  100 0 100 100 100
100 100 100 100 41 ...
##  $ Medium.Cloud.Cover.daily.min..mid.cld.lay.  : int  0 0 0 0 0 0 0 0 0
0 ...
##  $ Low.Cloud.Cover.daily.max..low.cld.lay.     : int  100 28 100 100 100
100 100 100 100 100 ...
##  $ Low.Cloud.Cover.daily.min..low.cld.lay.     : int  0 0 0 0 0 0 0 29 0
0 ...
##  $ Wind.Speed.daily.max..10.m.above.gnd.       : num  22 15.5 22.7 10.7
20.5 ...
##  $ Wind.Speed.daily.min..10.m.above.gnd.       : num  5.62 1.08 2.41 0 2
```

```
.52 2.28 1.3 4.32 7.2 8.05 ...
##  $ Wind.Speed.daily.max..80.m.above.gnd.       : num  23.8 18.7 32 10.2
23.4 ...
##  $ Wind.Speed.daily.min..80.m.above.gnd.       : num  8.65 0 0.51 1.44 2
.97 ...
##  $ Wind.Speed.daily.max..900.mb.               : num  32.1 48.1 44 22.2
40.8 ...
##  $ Wind.Speed.daily.min..900.mb.               : num  12.25 6.62 5.48 4.
69 4.68 ...
##  $ Wind.Gust.daily.max..sfc.                   : num  25.2 20.2 41.8 11.
2 24.1 ...
##  $ Wind.Gust.daily.min..sfc.                   : num  6.48 2.16 1.08 0.3
6 1.44 ...
##  $ pluie.demain                                : logi  FALSE FALSE TRUE
TRUE TRUE TRUE ...
```

Nous allons renommer les variables des fichiers " meteo.train " & " meteo.test " :

```
```{r Redéfinition DES VARIABLES meteo.train  }

library(dplyr)

data2=rename(Tempmean=Temperature.daily.mean..2.m.above.gnd.,Humimean=Relat
ive.Humidity.daily.mean..2.m.above.gnd.,MeanPressuremean= Mean.Sea.Level.Pr
essure.daily.mean..MSL.   , Totalprecipitation= Total.Precipitation.daily.s
um..sfc.         ,Snowfall=Snowfall.amount.raw.daily.sum..sfc.     ,Totalcl
oudmean= Total.Cloud.Cover.daily.mean..sfc.          ,Highcloudmean= High.C
loud.Cover.daily.mean..high.cld.lay.  , Mediumcloudmean= Medium.Cloud.Cover
.daily.mean..mid.cld.lay.  ,Lowcloudmean=Low.Cloud.Cover.daily.mean..low.cl
d.lay.    ,Sunshine= Sunshine.Duration.daily.sum..sfc.           ,Waveradia
= Shortwave.Radiation.daily.sum..sfc.         , Windspdmean10m= Wind.Speed.
daily.mean..10.m.above.gnd.      , Winddirecmean10m=Wind.Direction.daily.me
an..10.m.above.gnd.  , Windspdmean80m= Wind.Speed.daily.mean..80.m.above.gn
d.      , Winddirectmean80m=Wind.Direction.daily.mean..80.m.above.gnd.  , W
indspdmean900mb= Wind.Speed.daily.mean..900.mb.             ,  Winddirectm
ean900mb =Wind.Direction.daily.mean..900.mb.          , Windgustmean= Wind.
Gust.daily.mean..sfc.                  , Tempmax=Temperature.daily.max..2.m
.above.gnd.      ,Tempmin=Temperature.daily.min..2.m.above.gnd.     ,Humima
x=Relative.Humidity.daily.max..2.m.above.gnd. ,Humimin= Relative.Humidity.d
aily.min..2.m.above.gnd. , Meanpressuremax=Mean.Sea.Level.Pressure.daily.ma
x..MSL.     ,Meanpressuremin =Mean.Sea.Level.Pressure.daily.min..MSL.     ,
Totalcloudmax =Total.Cloud.Cover.daily.max..sfc.          ,Totalcloudmin =T
otal.Cloud.Cover.daily.min..sfc.           ,Highcloudmax= High.Cloud.Cover.
daily.max..high.cld.lay.   , Highcloudmin= High.Cloud.Cover.daily.min..high
.cld.lay.   ,Mediumcloudmax =Medium.Cloud.Cover.daily.max..mid.cld.lay.  ,M
ediumcloudmin= Medium.Cloud.Cover.daily.min..mid.cld.lay.  , Lowcloudmax=Lo
w.Cloud.Cover.daily.max..low.cld.lay.
               ,Lowcloudmin= Low.Cloud.Cover.daily.min..low.cld.lay.    ,Wi
ndspdmax10m=Wind.Speed.daily.max..10.m.above.gnd.       ,Windspdmin10m= Win
d.Speed.daily.min..10.m.above.gnd.       , Windspdmax80m= Wind.Speed.daily.
max..80.m.above.gnd.      ,Windspdmin80m=Wind.Speed.daily.min..80.m.above.g
nd.      ,Windspdmax900mb= Wind.Speed.daily.max..900.mb.              ,Win
dspdmin900mb= Wind.Speed.daily.min..900.mb.              ,Windgustmax= Win
d.Gust.daily.max..sfc.                  ,Windgustmin= Wind.Gust.daily.min..
```
```

```
sfc.
               ,data)

names(data2) “meteo.train”
```

```
##  [1] "X"                    "Year"                 "Month"
##  [4] "Day"                  "Hour"                 "Minute"
##  [7] "Tempmean"             "Humimean"             "MeanPressuremean"
## [10] "Totalprecipitation"   "Snowfall"             "Totalcloudmean"
## [13] "Highcloudmean"        "Mediumcloudmean"      "Lowcloudmean"
## [16] "Sunshine"             "Waveradia"            "Windspdmean10m"
## [19] "Winddirecmean10m"     "Windspdmean80m"       "Winddirectmean80m"
## [22] "Windspdmean900mb"     "Winddirectmean900mb" "Windgustmean"
## [25] "Tempmax"              "Tempmin"              "Humimax"
## [28] "Humimin"              "Meanpressuremax"      "Meanpressuremin"
## [31] "Totalcloudmax"        "Totalcloudmin"        "Highcloudmax"
## [34] "Highcloudmin"         "Mediumcloudmax"       "Mediumcloudmin"
## [37] "Lowcloudmax"          "Lowcloudmin"          "Windspdmax10m"
## [40] "Windspdmin10m"        "Windspdmax80m"        "Windspdmin80m"
## [43] "Windspdmax900mb"      "Windspdmin900mb"      "Windgustmax"
## [46] "Windgustmin"          "pluie.demain"
```

```
{r Redéfinition DES VARIABLES meteo.test  }

library(dplyr)

data3=rename(Tempmean=Temperature.daily.mean..2.m.above.gnd.,Humimean=Relat
ive.Humidity.daily.mean..2.m.above.gnd.,MeanPressuremean= Mean.Sea.Level.Pr
essure.daily.mean..MSL.   , Totalprecipitation= Total.Precipitation.daily.s
um..sfc.         ,Snowfall=Snowfall.amount.raw.daily.sum..sfc.     ,Totalcl
oudmean= Total.Cloud.Cover.daily.mean..sfc.          ,Highcloudmean= High.C
loud.Cover.daily.mean..high.cld.lay.  , Mediumcloudmean= Medium.Cloud.Cover
.daily.mean..mid.cld.lay.  ,Lowcloudmean=Low.Cloud.Cover.daily.mean..low.cl
d.lay.    ,Sunshine= Sunshine.Duration.daily.sum..sfc.          ,Waveradia
= Shortwave.Radiation.daily.sum..sfc.          , Windspdmean10m= Wind.Speed.
daily.mean..10.m.above.gnd.      , Winddirecmean10m=Wind.Direction.daily.me
an..10.m.above.gnd.  , Windspdmean80m= Wind.Speed.daily.mean..80.m.above.gn
d.      , Winddirectmean80m=Wind.Direction.daily.mean..80.m.above.gnd.  , W
indspdmean900mb= Wind.Speed.daily.mean..900.mb.              ,  Winddirectm
ean900mb =Wind.Direction.daily.mean..900.mb.          , Windgustmean= Wind.
Gust.daily.mean..sfc.                  , Tempmax=Temperature.daily.max..2.m
.above.gnd.      ,Tempmin=Temperature.daily.min..2.m.above.gnd.     ,Humima
x=Relative.Humidity.daily.max..2.m.above.gnd. ,Humimin= Relative.Humidity.d
aily.min..2.m.above.gnd. , Meanpressuremax=Mean.Sea.Level.Pressure.daily.ma
x..MSL.     ,Meanpressuremin =Mean.Sea.Level.Pressure.daily.min..MSL.     ,
Totalcloudmax =Total.Cloud.Cover.daily.max..sfc.          ,Totalcloudmin =T
otal.Cloud.Cover.daily.min..sfc.          ,Highcloudmax= High.Cloud.Cover.
daily.max..high.cld.lay.   , Highcloudmin= High.Cloud.Cover.daily.min..high
.cld.lay.   ,Mediumcloudmax =Medium.Cloud.Cover.daily.max..mid.cld.lay.  ,M
ediumcloudmin= Medium.Cloud.Cover.daily.min..mid.cld.lay.  , Lowcloudmax=Lo
w.Cloud.Cover.daily.max..low.cld.lay.
               ,Lowcloudmin= Low.Cloud.Cover.daily.min..low.cld.lay.    ,Wi
ndspdmax10m=Wind.Speed.daily.max..10.m.above.gnd.       ,Windspdmin10m= Win
d.Speed.daily.min..10.m.above.gnd.       , Windspdmax80m= Wind.Speed.daily.
max..80.m.above.gnd.      ,Windspdmin80m=Wind.Speed.daily.min..80.m.above.g
```

```
nd.      ,Windspdmax900mb= Wind.Speed.daily.max..900.mb.                ,Win
dspdmin900mb= Wind.Speed.daily.min..900.mb.                ,Windgustmax= Win
d.Gust.daily.max..sfc.                  ,Windgustmin= Wind.Gust.daily.min..
sfc.
               ,data1)

names(data3) "meteo.test"
```

```
##  [1] "X"                   "Year"                "Month"
##  [4] "Day"                 "Hour"                "Minute"
##  [7] "Tempmean"            "Humimean"            "MeanPressuremean"
## [10] "Totalprecipitation"  "Snowfall"            "Totalcloudmean"
## [13] "Highcloudmean"       "Mediumcloudmean"     "Lowcloudmean"
## [16] "Sunshine"            "Waveradia"           "Windspdmean10m"
## [19] "Winddirecmean10m"    "Windspdmean80m"      "Winddirectmean80m"
## [22] "Windspdmean900mb"    "Winddirectmean900mb" "Windgustmean"
## [25] "Tempmax"             "Tempmin"             "Humimax"
## [28] "Humimin"             "Meanpressuremax"     "Meanpressuremin"
## [31] "Totalcloudmax"       "Totalcloudmin"       "Highcloudmax"
## [34] "Highcloudmin"        "Mediumcloudmax"      "Mediumcloudmin"
## [37] "Lowcloudmax"         "Lowcloudmin"         "Windspdmax10m"
## [40] "Windspdmin10m"       "Windspdmax80m"       "Windspdmin80m"
## [43] "Windspdmax900mb"     "Windspdmin900mb"     "Windgustmax"
## [46] "Windgustmin"
```

Nous allons supprimer toutes les valeurs manquantes possibles du jeu de données (meteo.train.csv) & (meteo.test.csv)

```
data2 <- na.omit(data2)
attach(data2)

data3 <- na.omit(data3)
attach(data3)
```

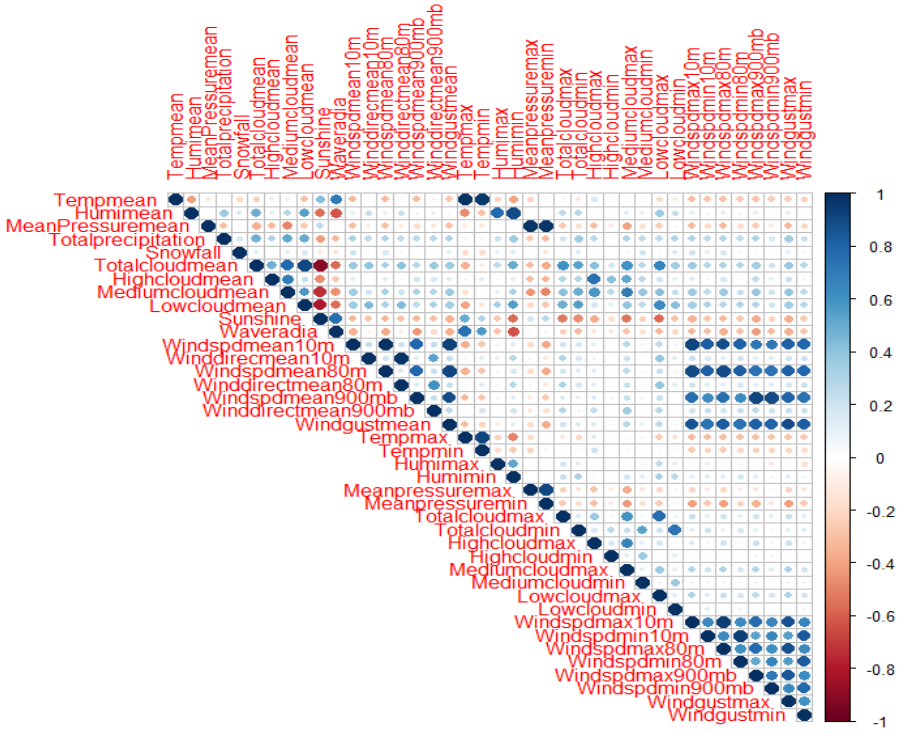
1.2 Vérifier les liens entre les variables

1.2.1 Premières observations sur les corrélations entre les variables : meteo.train.csv(Data2)

Nous utilisons la matrice de corrélation pour évaluer la dépendance entre plusieurs variables en même temps.

```
library(corrplot)
data2.quanti<-data2[,c(7:46)]
cor.data2.quanti<-cor(data2.quanti,use = "complete")
```

```
corrplot(cor.data2.quanti,type="upper")
library(Hmisc)
```




```
rcorr(as.matrix(data2[,7:46]))
```

	Tempmean	Humimean	MeanPressuremean	Totalprecipitation	Snowfall	Totalcloudmean	Highcloudmean
Tempmean	1.00	-0.42	-0.14	-0.01	-0.20	-0.24	0.11
Humimean	-0.42	1.00	-0.01	0.36	0.16	0.49	0.11
MeanPressuremean	-0.14	-0.01	1.00	-0.31	-0.10	-0.36	-0.34
Totalprecipitation	-0.01	0.36	-0.31	1.00	0.17	0.48	0.30
Snowfall	-0.20	0.16	-0.10	0.17	1.00	0.13	-0.02
Totalcloudmean	-0.24	0.49	-0.36	0.48	0.13	1.00	0.48
Highcloudmean	0.11	0.11	-0.34	0.30	-0.02	0.48	1.00
Mediumcloudmean	-0.10	0.29	-0.49	0.52	0.13	0.79	0.70
Lowcloudmean	-0.30	0.56	-0.26	0.51	0.17	0.90	0.24
Sunshine	0.42	-0.55	0.25	-0.42	-0.12	-0.91	-0.47
Waveradia	0.70	-0.62	0.03	-0.32	-0.15	-0.57	-0.25
Windspdmean10m	-0.32	0.11	-0.30	0.32	0.21	0.35	0.11
Winddirecmean10m	0.03	0.24	-0.11	0.26	0.06	0.40	0.03
Windspdmean80m	-0.33	0.11	-0.29	0.31	0.16	0.33	0.13
Winddirectmean80m	0.04	0.22	-0.11	0.26	0.06	0.40	0.05
Windspdmean900mb	-0.31	0.09	-0.19	0.26	0.04	0.29	0.20
Winddirectmean900mb	0.09	0.11	-0.15	0.28	-0.01	0.36	0.17
Windgustmean	-0.28	0.06	-0.29	0.32	0.13	0.35	0.16
Tempmax	0.98	-0.48	-0.10	-0.08	-0.20	-0.34	0.08
Tempmin	0.97	-0.30	-0.15	0.07	-0.18	-0.10	0.15
Humimax	-0.22	0.77	0.02	0.22	0.10	0.27	0.04
Humimin	-0.40	0.89	-0.02	0.36	0.16	0.50	0.12
Meanpressuremax	-0.22	0.02	0.97	-0.26	-0.04	-0.32	-0.31
Meanpressuremin	-0.06	-0.03	0.97	-0.32	-0.15	-0.38	-0.33
Totalcloudmax	-0.08	0.26	-0.26	0.20	0.05	0.61	0.30

	Mediumcloudmean	Lowcloudmean	Sunshine	Waveradia	Windspdmean10m	Winddirecmean10m	Windspdmean80m
Tempmean	-0.10	-0.30	0.42	0.70	-0.32	0.03	-0.33
Humimean	0.29	0.56	-0.55	-0.62	0.11	0.24	0.11
MeanPressuremean	-0.49	-0.26	0.25	0.03	-0.30	-0.11	-0.29
Totalprecipitation	0.52	0.51	-0.42	-0.32	0.32	0.26	0.31
Snowfall	0.13	0.17	-0.12	-0.15	0.21	0.06	0.16
Totalcloudmean	0.79	0.90	-0.91	-0.57	0.35	0.40	0.33
Highcloudmean	0.70	0.24	-0.47	-0.25	0.11	0.03	0.13
Mediumcloudmean	1.00	0.57	-0.74	-0.45	0.34	0.18	0.34
Lowcloudmean	0.57	1.00	-0.80	-0.56	0.36	0.44	0.32
Sunshine	-0.74	-0.80	1.00	0.75	-0.34	-0.28	-0.34
Waveradia	-0.45	-0.56	0.75	1.00	-0.34	-0.09	-0.38
Windspdmean10m	0.34	0.36	-0.34	-0.34	1.00	0.23	0.98
Winddirecmean10m	0.18	0.44	-0.28	-0.09	0.23	1.00	0.18
Windspdmean80m	0.34	0.32	-0.34	-0.38	0.98	0.18	1.00
Winddirectmean80m	0.20	0.44	-0.29	-0.10	0.21	0.97	0.16
Windspdmean900mb	0.33	0.27	-0.34	-0.43	0.79	0.13	0.80
Winddirectmean900mb	0.28	0.33	-0.30	-0.21	0.24	0.53	0.23
Windgustmean	0.36	0.34	-0.35	-0.36	0.92	0.23	0.92
Tempmax	-0.17	-0.41	0.50	0.75	-0.35	-0.07	-0.35
Tempmin	-0.01	-0.16	0.29	0.58	-0.27	0.14	-0.28
Humimax	0.11	0.33	-0.28	-0.29	-0.07	0.11	-0.07
Humimin	0.32	0.56	-0.57	-0.64	0.16	0.28	0.16
Meanpressuremax	-0.45	-0.22	0.20	-0.04	-0.20	-0.09	-0.19
Meanpressuremin	-0.50	-0.28	0.28	0.08	-0.39	-0.12	-0.37
Totalcloudmax	0.41	0.50	-0.52	-0.24	0.20	0.30	0.17

	Winddirectmean80m	Windspdmean900mb	Winddirectmean900mb	Windgustmean	Tempmax	Tempmin	Humimax
Tempmean	0.04	-0.31	0.09	-0.28	0.98	0.97	-0.22
Humimean	0.22	0.09	0.11	0.06	-0.48	-0.30	0.77
MeanPressuremean	-0.11	-0.19	-0.15	-0.29	-0.10	-0.15	0.02
Totalprecipitation	0.26	0.26	0.28	0.32	-0.08	0.07	0.22
Snowfall	0.06	0.04	-0.01	0.13	-0.20	-0.18	0.10
Totalcloudmean	0.40	0.29	0.36	0.35	-0.34	-0.10	0.27
Highcloudmean	0.05	0.20	0.17	0.16	0.08	0.15	0.04
Mediumcloudmean	0.20	0.33	0.28	0.36	-0.17	-0.01	0.11
Lowcloudmean	0.44	0.27	0.33	0.34	-0.41	-0.16	0.33
Sunshine	-0.29	-0.34	-0.30	-0.35	0.50	0.29	-0.28
Waveradia	-0.10	-0.43	-0.21	-0.36	0.75	0.58	-0.29
Windspdmean10m	0.21	0.79	0.24	0.92	-0.35	-0.27	-0.07
Winddirecmean10m	0.97	0.13	0.53	0.23	-0.07	0.14	0.11
Windspdmean80m	0.16	0.80	0.23	0.92	-0.35	-0.28	-0.07
Winddirectmean80m	1.00	0.14	0.61	0.22	-0.05	0.15	0.08
Windspdmean900mb	0.14	1.00	0.23	0.89	-0.33	-0.27	-0.08
Winddirectmean900mb	0.61	0.23	1.00	0.26	0.00	0.19	-0.02
Windgustmean	0.22	0.89	0.26	1.00	-0.31	-0.23	-0.11
Tempmax	-0.05	-0.33	0.00	-0.31	1.00	0.91	-0.24
Tempmin	0.15	-0.27	0.19	-0.23	0.91	1.00	-0.19
Humimax	0.08	-0.08	-0.02	-0.11	-0.24	-0.19	1.00
Humimin	0.27	0.14	0.19	0.13	-0.49	-0.26	0.53
Meanpressuremax	-0.10	-0.10	-0.13	-0.20	-0.18	-0.23	0.03
Meanpressuremin	-0.12	-0.27	-0.16	-0.37	-0.04	-0.08	0.01
Totalcloudmax	0.31	0.16	0.28	0.19	-0.14	0.01	0.18

	Humimin	Meanpressuremax	Meanpressuremin	Totalcloudmax	Totalcloudmin	Highcloudmax	Highcloudmin
Tempmean	-0.40	-0.22	-0.06	-0.08	-0.15	0.18	-0.03
Humimean	0.89	0.02	-0.03	0.26	0.29	0.05	0.04
MeanPressuremean	-0.02	0.97	0.97	-0.26	-0.18	-0.30	-0.10
Totalprecipitation	0.36	-0.26	-0.32	0.20	0.35	0.27	0.04
Snowfall	0.16	-0.04	-0.15	0.05	0.16	-0.01	0.00
Totalcloudmean	0.50	-0.32	-0.38	0.61	0.53	0.41	0.18
Highcloudmean	0.12	-0.31	-0.33	0.30	0.30	0.75	0.42
Mediumcloudmean	0.32	-0.45	-0.50	0.41	0.49	0.59	0.26
Lowcloudmean	0.56	-0.22	-0.28	0.50	0.55	0.18	0.11
Sunshine	-0.57	0.20	0.28	-0.52	-0.47	-0.37	-0.17
Waveradia	-0.64	-0.04	0.08	-0.24	-0.33	-0.15	-0.13
Windspdmean10m	0.16	-0.20	-0.39	0.20	0.17	0.12	0.02
Winddirecmean10m	0.28	-0.09	-0.12	0.30	0.18	0.10	0.00
Windspdmean80m	0.16	-0.19	-0.37	0.17	0.15	0.13	0.03
Winddirectmean80m	0.27	-0.10	-0.12	0.31	0.17	0.12	0.00
Windspdmean900mb	0.14	-0.10	-0.27	0.16	0.11	0.16	0.09
Winddirectmean900mb	0.19	-0.13	-0.16	0.28	0.08	0.25	0.04
Windgustmean	0.13	-0.20	-0.37	0.19	0.16	0.15	0.06
Tempmax	-0.49	-0.18	-0.04	-0.14	-0.20	0.15	-0.04
Tempmin	-0.26	-0.23	-0.08	0.01	-0.07	0.22	-0.01
Humimax	0.53	0.03	0.01	0.18	0.13	-0.01	0.03
Humimin	1.00	0.00	-0.04	0.23	0.32	0.06	0.04
Meanpressuremax	0.00	1.00	0.90	-0.24	-0.16	-0.28	-0.10
Meanpressuremin	-0.04	0.90	1.00	-0.27	-0.17	-0.31	-0.10
Totalcloudmax	0.23	-0.24	-0.27	1.00	0.15	0.38	0.07

	Mediumcloudmax	Mediumcloudmin	Lowcloudmax	Lowcloudmin	Windspdmax10m	Windspdmin10m	Windspdmax80m
Tempmean	0.00	-0.06	-0.18	-0.12	-0.30	-0.27	-0.29
Humimean	0.15	0.12	0.37	0.24	0.11	0.09	0.11
MeanPressuremean	-0.40	-0.16	-0.25	-0.08	-0.32	-0.21	-0.32
Totalprecipitation	0.33	0.25	0.26	0.31	0.31	0.25	0.30
Snowfall	0.08	0.20	0.07	0.16	0.20	0.12	0.17
Totalcloudmean	0.60	0.27	0.67	0.34	0.36	0.28	0.32
Highcloudmean	0.52	0.27	0.20	0.11	0.14	0.06	0.18
Mediumcloudmean	0.70	0.39	0.41	0.24	0.36	0.25	0.36
Lowcloudmean	0.39	0.25	0.65	0.43	0.35	0.31	0.29
Sunshine	-0.54	-0.24	-0.57	-0.30	-0.34	-0.28	-0.33
Waveradia	-0.28	-0.17	-0.34	-0.22	-0.34	-0.30	-0.36
Windspdmean10m	0.28	0.08	0.26	0.08	0.92	0.83	0.89
Winddirecmean10m	0.21	0.05	0.39	0.13	0.21	0.17	0.14
Windspdmean80m	0.27	0.08	0.22	0.06	0.90	0.82	0.90
Winddirectmean80m	0.23	0.04	0.39	0.12	0.19	0.16	0.12
Windspdmean900mb	0.22	0.07	0.23	0.01	0.77	0.63	0.75
Winddirectmean900mb	0.36	0.02	0.29	0.02	0.20	0.22	0.18
Windgustmean	0.27	0.10	0.27	0.06	0.87	0.74	0.84
Tempmax	-0.06	-0.08	-0.26	-0.16	-0.32	-0.31	-0.30
Tempmin	0.08	-0.03	-0.07	-0.06	-0.26	-0.22	-0.26
Humimax	0.01	0.06	0.26	0.12	-0.03	-0.10	-0.04
Humimin	0.19	0.13	0.33	0.26	0.12	0.15	0.13
Meanpressuremax	-0.36	-0.15	-0.21	-0.08	-0.21	-0.13	-0.21
Meanpressuremin	-0.42	-0.17	-0.26	-0.07	-0.41	-0.26	-0.42
Totalcloudmax	0.60	0.07	0.77	0.09	0.23	0.12	0.19

	Windspdmin80m	Windspdmax900mb	Windspdmin900mb	Windgustmax	Windgustmin
Tempmean	-0.26	-0.28	-0.25	-0.25	-0.24
Humimean	0.10	0.11	0.02	0.12	0.02
MeanPressuremean	-0.19	-0.26	-0.11	-0.34	-0.21
Totalprecipitation	0.25	0.32	0.17	0.35	0.23
Snowfall	0.09	0.06	0.00	0.12	0.07
Totalcloudmean	0.28	0.34	0.20	0.37	0.27
Highcloudmean	0.05	0.25	0.10	0.21	0.09
Mediumcloudmean	0.23	0.39	0.21	0.39	0.26
Lowcloudmean	0.31	0.29	0.20	0.34	0.28
Sunshine	-0.28	-0.38	-0.24	-0.37	-0.27
Waveradia	-0.29	-0.43	-0.32	-0.35	-0.28
Windspdmean10m	0.81	0.73	0.73	0.81	0.80
Winddirecmean10m	0.20	0.13	0.15	0.20	0.22
Windspdmean80m	0.83	0.73	0.73	0.81	0.79
Winddirectmean80m	0.19	0.13	0.15	0.19	0.21
Windspdmean900mb	0.63	0.92	0.90	0.80	0.74
Winddirectmean900mb	0.23	0.25	0.20	0.23	0.22
Windgustmean	0.75	0.82	0.81	0.89	0.82
Tempmax	-0.30	-0.30	-0.27	-0.27	-0.27
Tempmin	-0.21	-0.24	-0.21	-0.21	-0.19
Humimax	-0.10	-0.03	-0.13	-0.04	-0.17
Humimin	0.17	0.14	0.08	0.14	0.09
Meanpressuremax	-0.11	-0.16	-0.04	-0.24	-0.13
Meanpressuremin	-0.24	-0.34	-0.16	-0.42	-0.27
Totalcloudmax	0.11	0.19	0.11	0.24	0.13

1.2.2 Interprétation de la matrice de corrélation

Nous nous sommes basés sur la matrice de corrélation pour analyser les relations entre les variables quantitatives. Les résultats font apparaitre les corrélations suivantes :

Forte corrélation positive entre Tempmean, Tempmin et Tempmax

Forte corrélation positive entre Humimean et Humimin

Forte corrélation positive entre Meanpressuremax, Meanpressuremin et MeanPressuremean

Forte corrélation positive entre Totalcloudmean et Lowcloudmean

Forte corrélation négatives entre Totalcloudmean et Sunshine

Forte corrélation positive entre Windspdmean10m , Windspdmean80m, windgustmean ,windspedmax10m et windspedmax80m

Forte corrélation positive entre Winddirecmean10m et Winddirectmean80m

Forte corrélation positive entre Windspdmean80m , windgustmean ,windspedmax10m et windspedmax80m

Forte corrélation positive entre windspedmean900mb ,windgustmean,windspedmin900mb,windspedmax900mb

Forte corrélation positive entre windgustmean et windgustmax

Forte corrélation positive entre Tempmax et Tempmin

Forte corrélation positive entre windspdmax10m et windspdmax80m

Forte corrélation positive entre windspdmin10m et windspdmin80m

1.2.3 Analyse de colinéarités entre Les variables

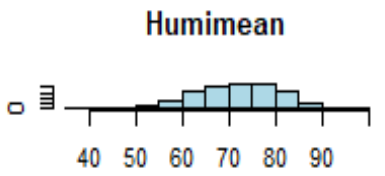
Nous utilisons les histogrammes pour détecter une éventuelle colinéarité entre Les variables.

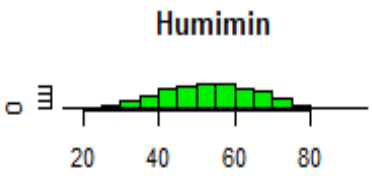
```
attach(data2)
par(mfrow = c(3, 2))
hist(x = Humimean, col = "lightblue", main = "Humimean", xlab = "", ylab =
"")
hist(x = Humimin, col = "green2", main = "Humimin", xlab = "", ylab = "")
hist(x = Tempmax, col = "orange", main = "Tempmax", xlab = "",
     ylab = "")
hist(x = Tempmin, col = "red", main = "Tempmin", xlab = "",
     ylab = "")
hist(x = Tempmean, col = "orange", main = "Tempmean", xlab = "",
     ylab = "")
hist(x = MeanPressuremean, col = "red", main = "MeanPressuremean", xlab = "
",
     ylab = "")

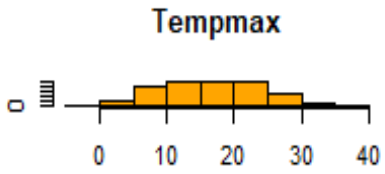
hist(x = Meanpressuremax, col = "slategray", main = "Meanpressuremax", xlab
= "", ylab = "")
hist(x = Meanpressuremin, col = "slategray", main = "Meanpressuremin", xlab
= "", ylab = "")
hist(x = Windspdmean80m, col = "red", main = "Windspdmean80m", xlab = "",
ylab = "")
hist(x = Windspdmean900mb, col = "green2", main = "Windspdmean900mb", xlab
= "", ylab = "")

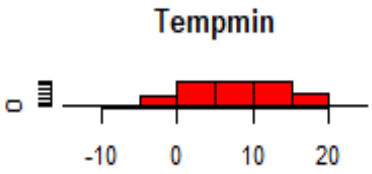
hist(x = Windspdmax10m, col = "red", main = "Windspdmax10m", xlab = "",   y
lab = "")
```

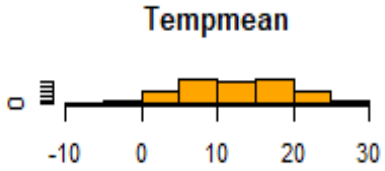
```
hist(x = Winddirectmean80m, col = "violet", main = "Winddirectmean80m", xla
b = "",  ylab = "")
```

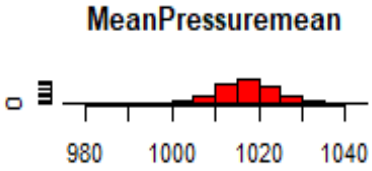


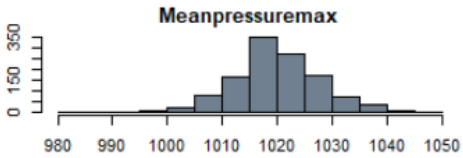


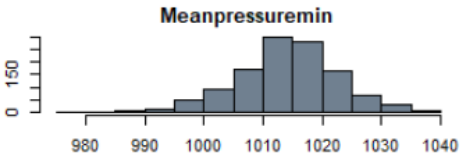


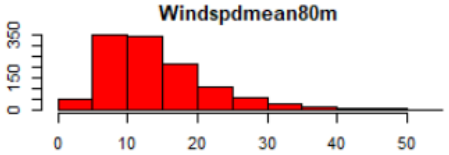


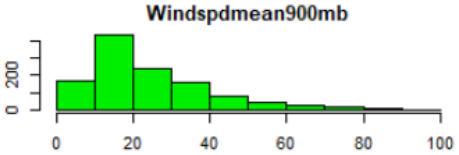


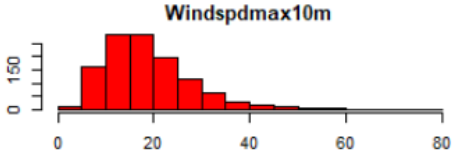


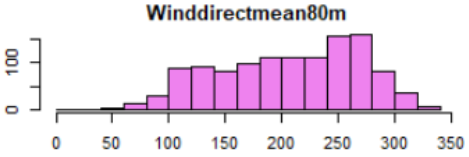


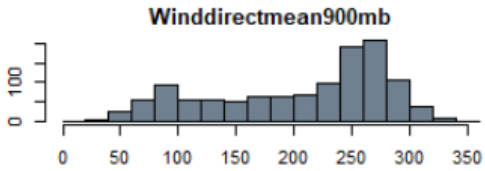


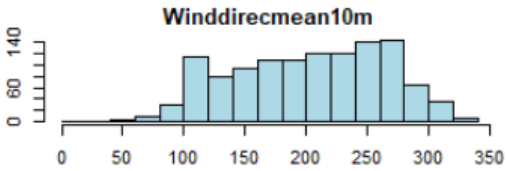












1.2.4 Interprétation des histogrammes

D'après les graphiques des histogrammes ci-dessus nous avons détecté les colinéarités entre les variables suivantes:

Humimean et Humimin

Tempmin,Tempmin,Tempmean

Winddirectmean80m,Winddirecmean900mb

Windspdmean80m, Windspdmean900mb,Windspdmax10m

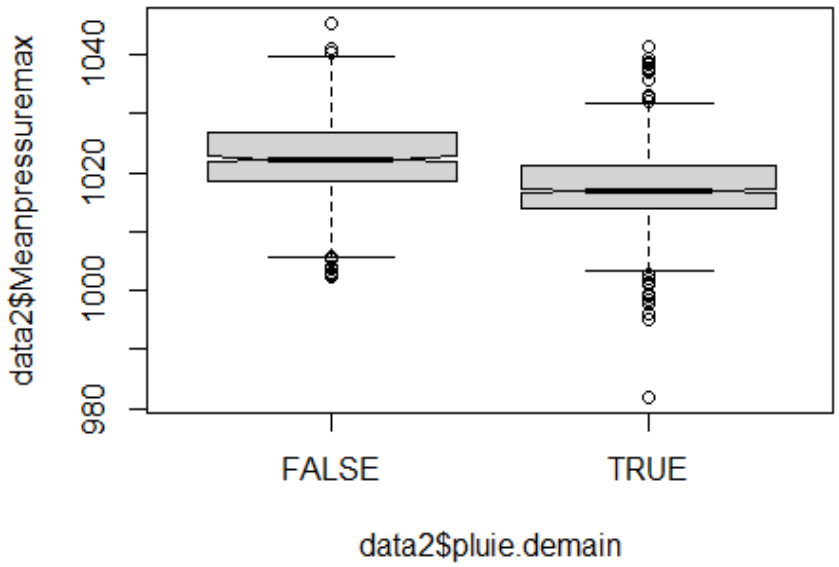
MeanPressuremean,Meanpressuremax,Meanpressuremin,

1.2.5 Recherche des variables les plus pertinentes

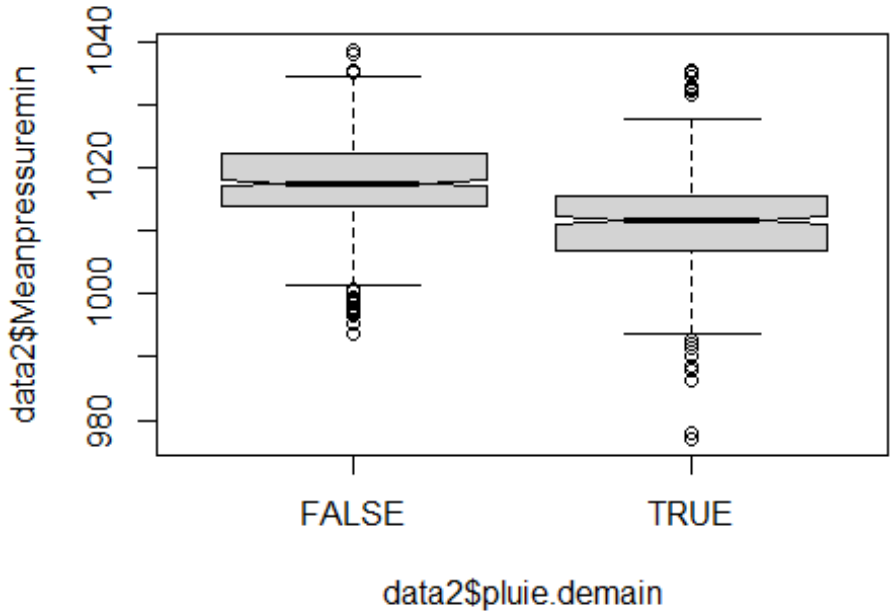
Nous utilisons en ce qui suit les boxplots et les diagrammes de densité qui permettent de comprendre visuellement la significativité d'un prédicteur en examinant le degré de chevauchement des valeurs prédictives fixées en fonction de la variable à prédire (pluie.demain).

-Analyse avec les boxplots

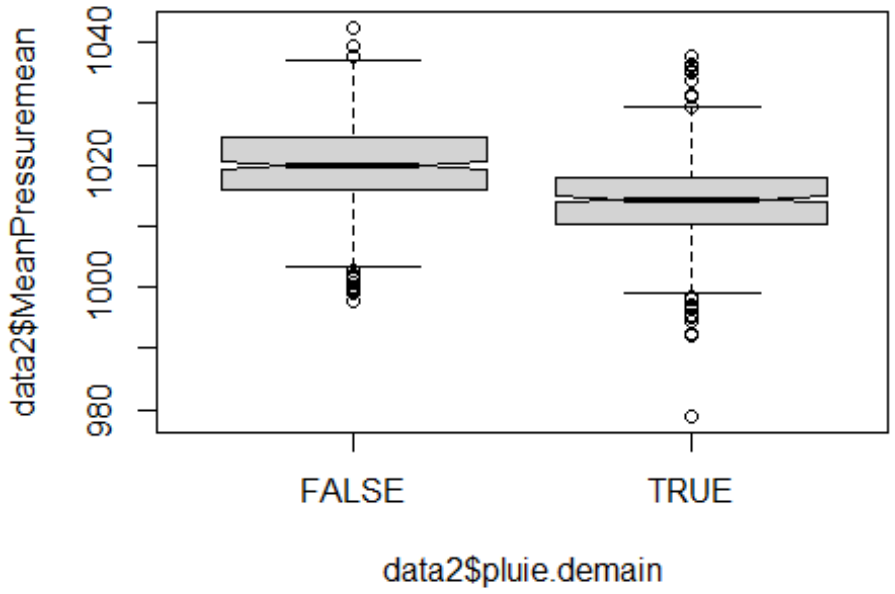
```
boxplot(data2$Meanpressuremax~data2$pluie.demain,varwidth = TRUE, notch = T
RUE, outline = TRUE)
```



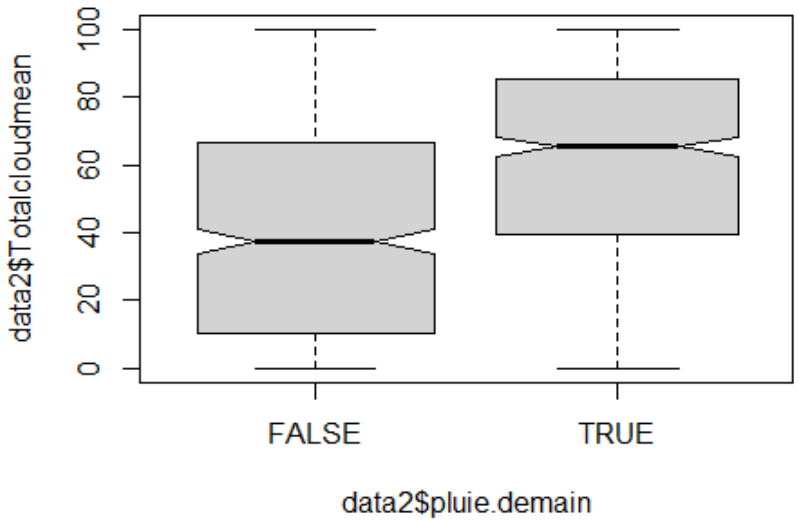

```
boxplot(data2$Meanpressuremin~data2$pluie.demain,varwidth = TRUE, notch = T
RUE, outline = TRUE)
```



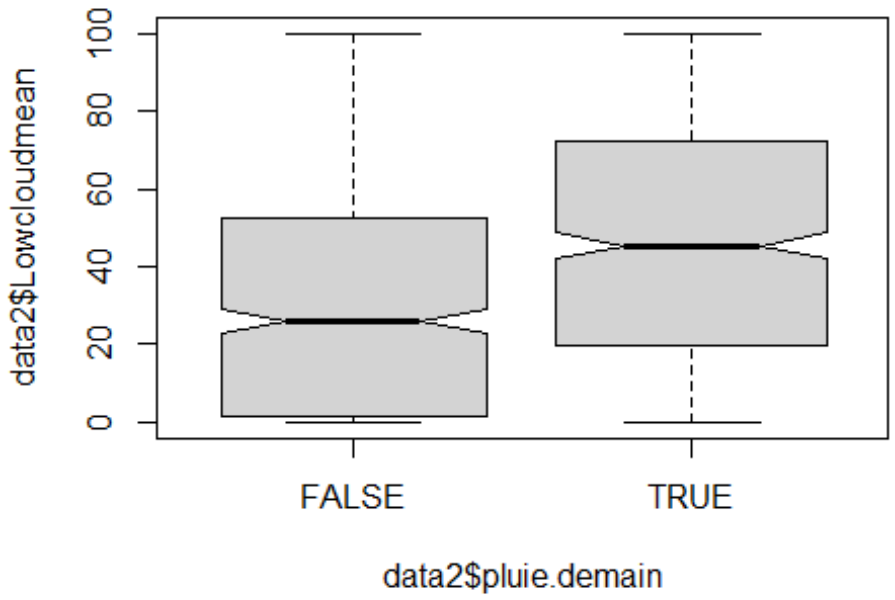

```
boxplot(data2$MeanPressuremean~data2$pluie.demain,varwidth = TRUE, notch =
TRUE, outline = TRUE)
```



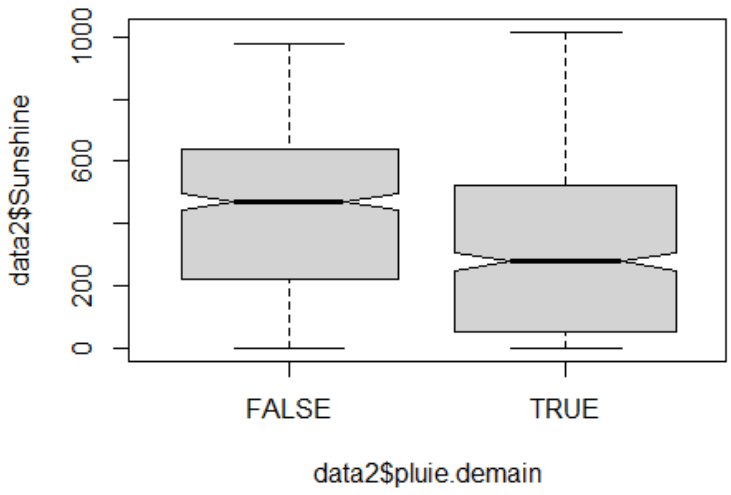

```
boxplot(data2$Totalcloudmean~data2$pluie.demain,varwidth = TRUE, notch = TR
UE, outline = TRUE)
```



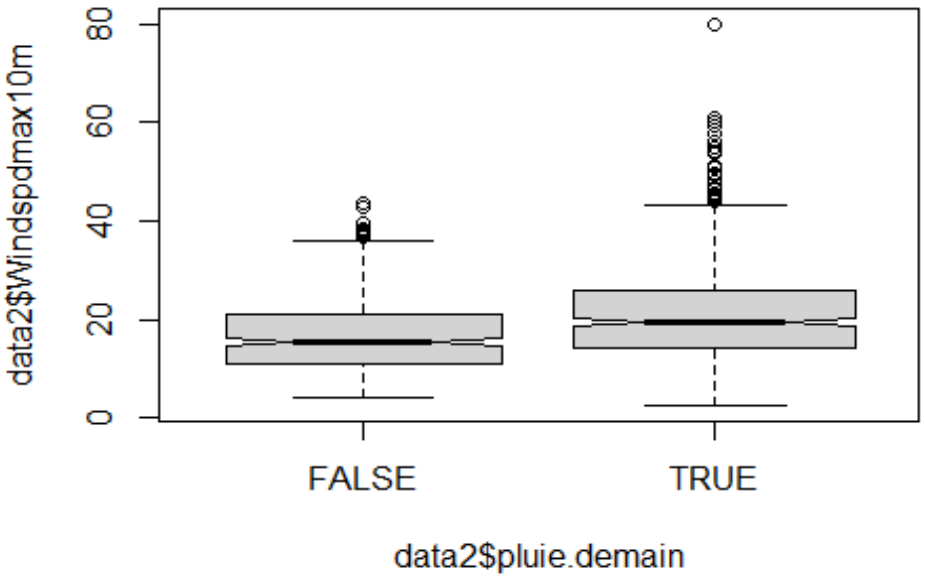

```
boxplot(data2$Lowcloudmean~data2$pluie.demain,varwidth = TRUE, notch = TRUE
, outline = TRUE)
```



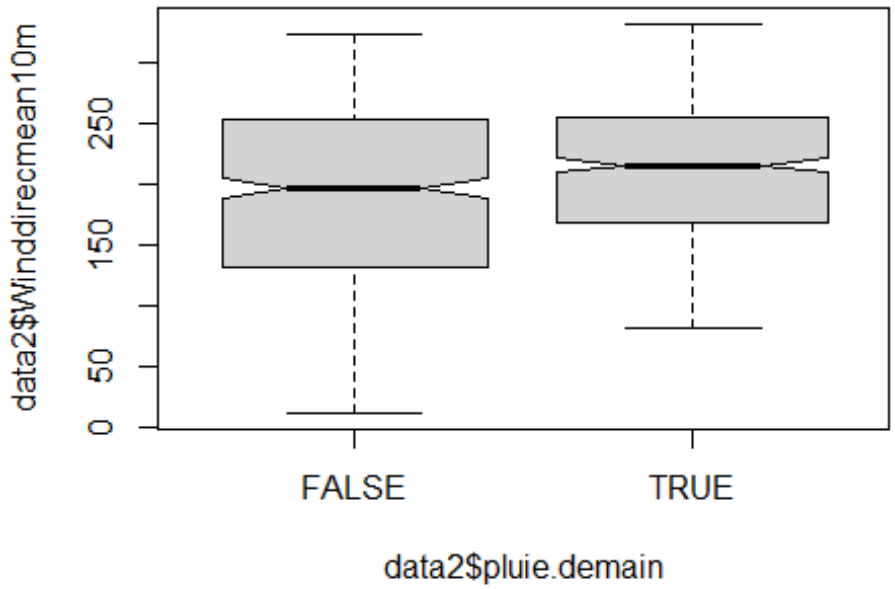

```
boxplot(data2$Sunshine~data2$pluie.demain,varwidth = TRUE, notch = TRUE, ou
tline = TRUE)
```



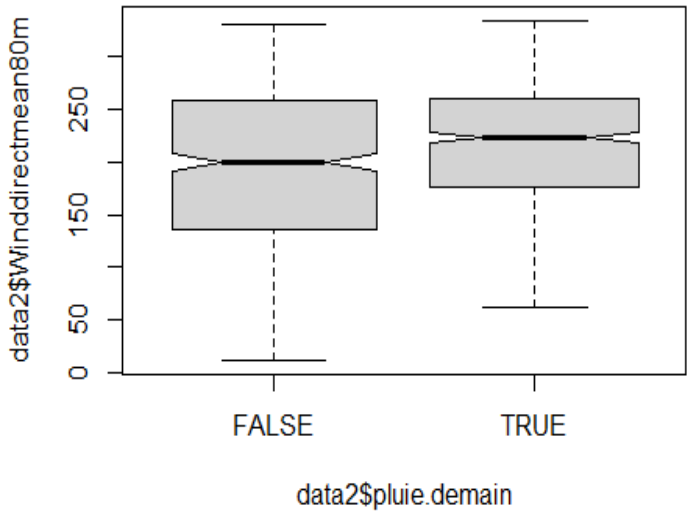

```
boxplot(data2$Windspdmax10m~data2$pluie.demain,varwidth = TRUE, notch = TRU
E, outline = TRUE)
```




```
boxplot(data2$Winddirecmean10m~data2$pluie.demain,varwidth = TRUE, notch =
TRUE, outline = TRUE)
```




```
boxplot(data2$Winddirectmean80m~data2$pluie.demain,varwidth = TRUE, notch =
TRUE, outline = TRUE)
```



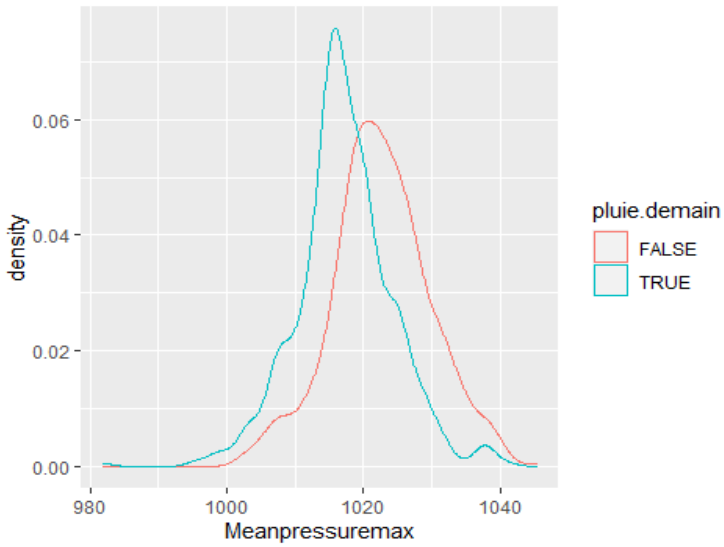
Interprétation des résultats de boxplot

Nous constatons qu'une valeur élevée de chacune de variables suivantes (Totalcloudmean,Lowcloudmean,Sunshine,Winddirectmean80m,Winddirectmean10m,) est associée avec une forte la probabilité de pleuvoir vs Une faible valeur de la variable est associé avec la probabilité de ne pas pleuvoir.

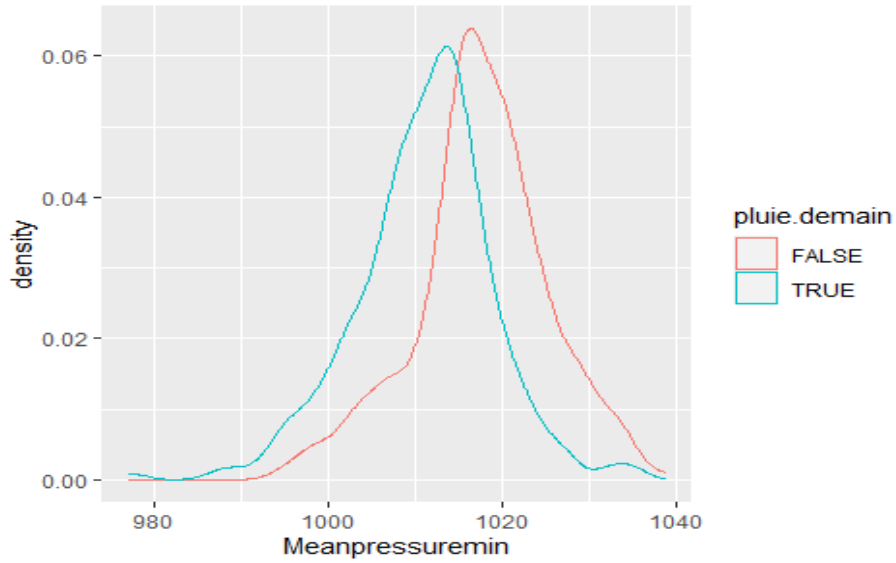
En outre certaines variables comme la tempmean Windspmean10m,Windspmean80m et Windgustmean ayant un faible impact sur la probabilité de pleuvoir .

-Analyse avec les diagrammes de densités

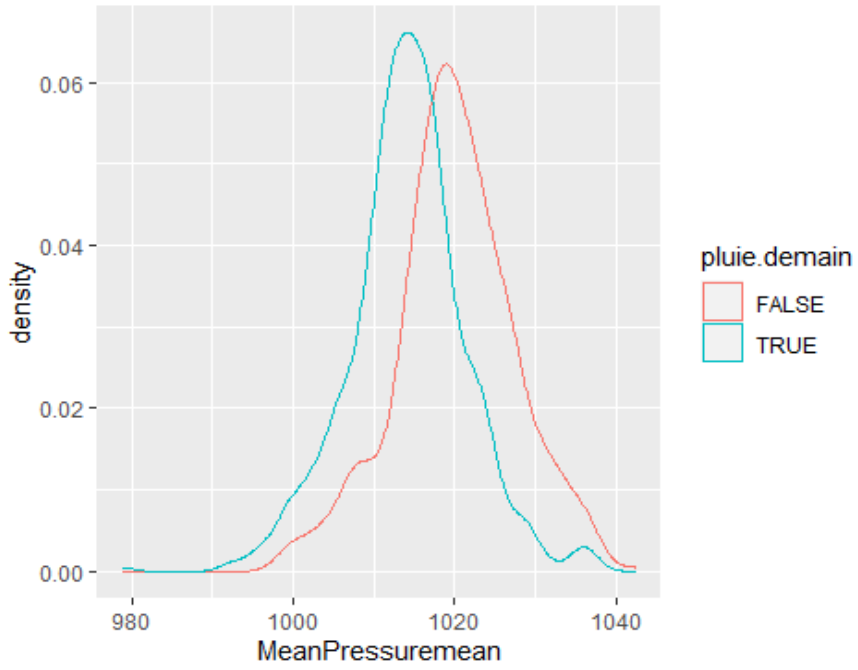
```
library(ggplot2)
# Changer la couleur des traits par groupe
ggplot(data2, aes(x=Meanpressuremax, color=pluie.demain)) +
  geom_density()
```



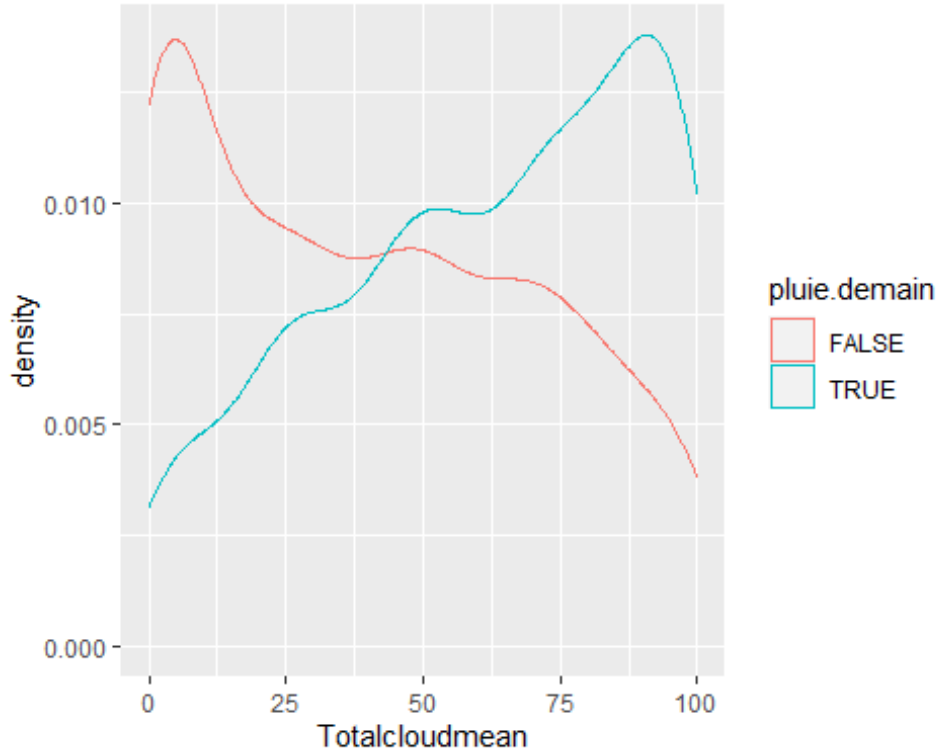
```
ggplot(data2, aes(x=Meanpressuremin, color=pluie.demain)) +
  geom_density()
```



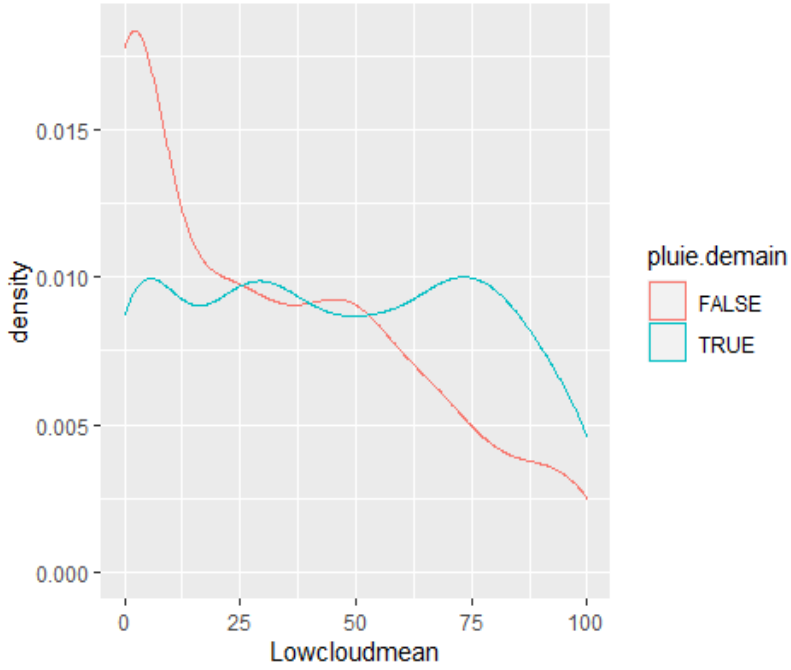

```
ggplot(data2, aes(x=MeanPressuremean, color=pluie.demain)) +
  geom_density()
```



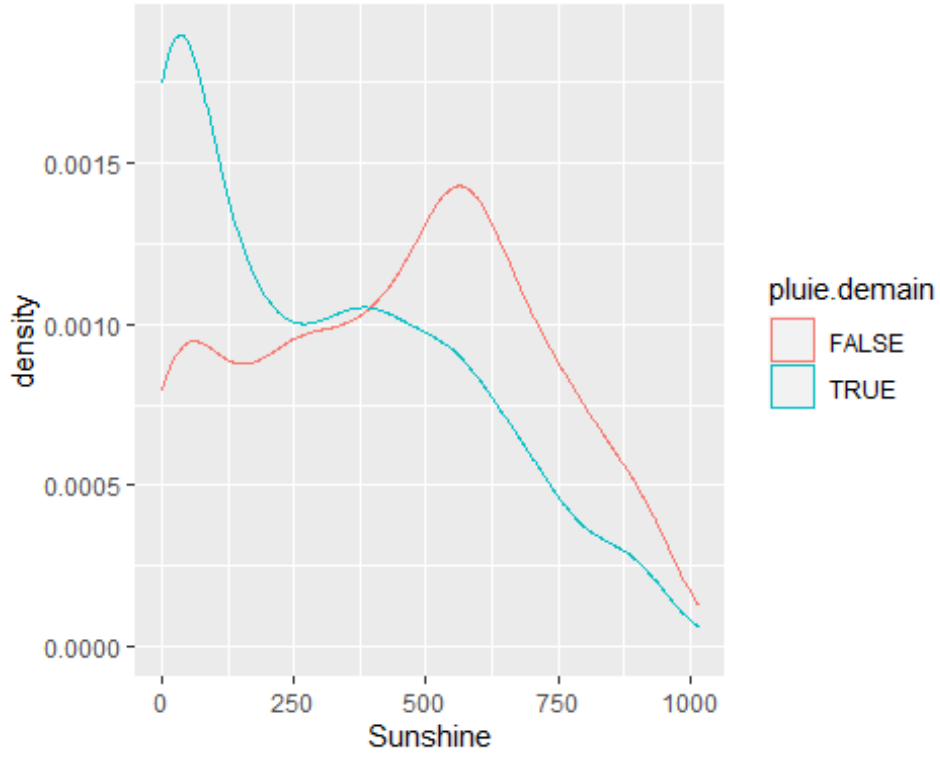

```
ggplot(data2, aes(x=Totalcloudmean, color=pluie.demain)) +
  geom_density()
```



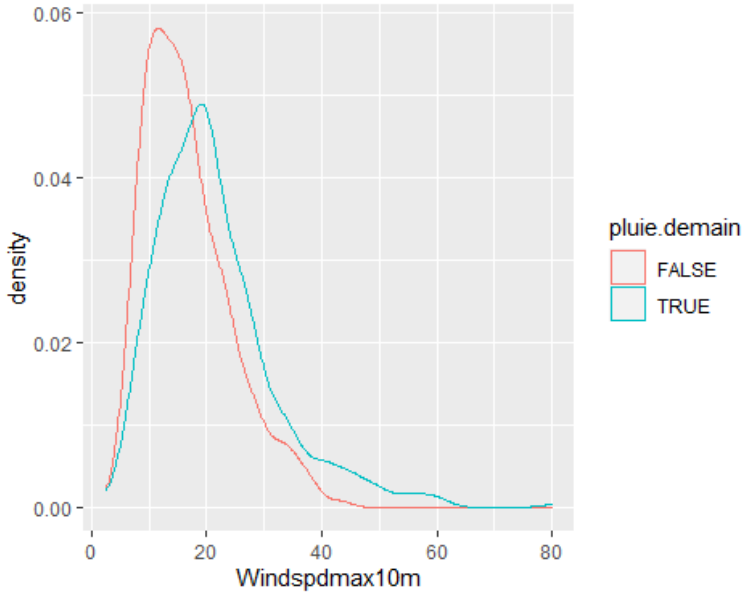

```
ggplot(data2, aes(x=Lowcloudmean, color=pluie.demain)) +
  geom_density()
```



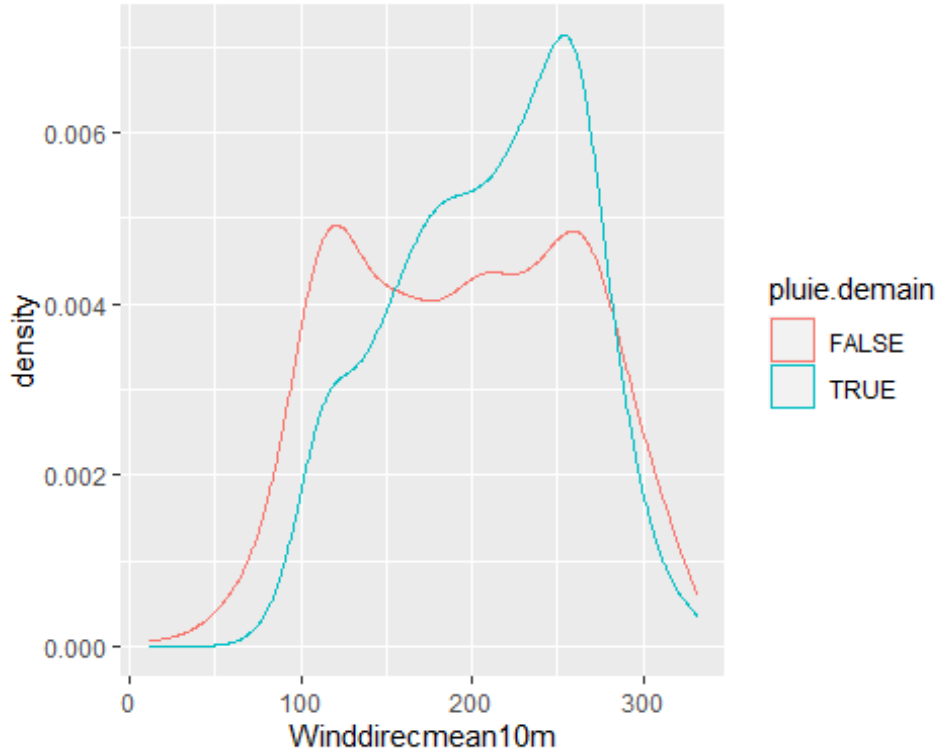

```
ggplot(data2, aes(x=Sunshine, color=pluie.demain)) +
  geom_density()
```



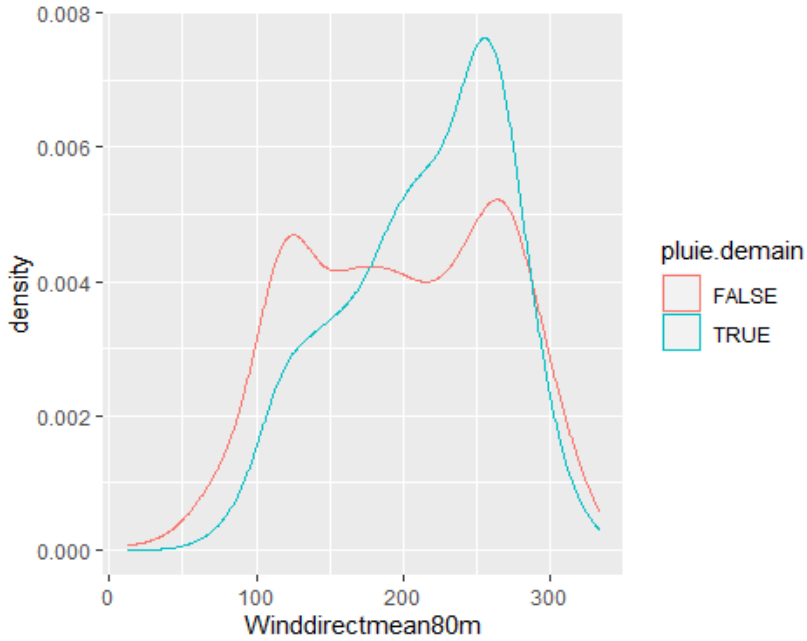

```
ggplot(data2, aes(x=Windspdmax10m, color=pluie.demain)) +
  geom_density()
```




```
ggplot(data2, aes(x=Winddirecmean10m, color=pluie.demain)) +
  geom_density()
```




```
ggplot(data2, aes(x=Winddirectmean80m, color=pluie.demain)) +
  geom_density()
```



#Interprétation de diagramme de densité

Les résultats obtenus par cette méthode nous confirme la pertinence des variables déjà sélectionnées par la méthode précédente .

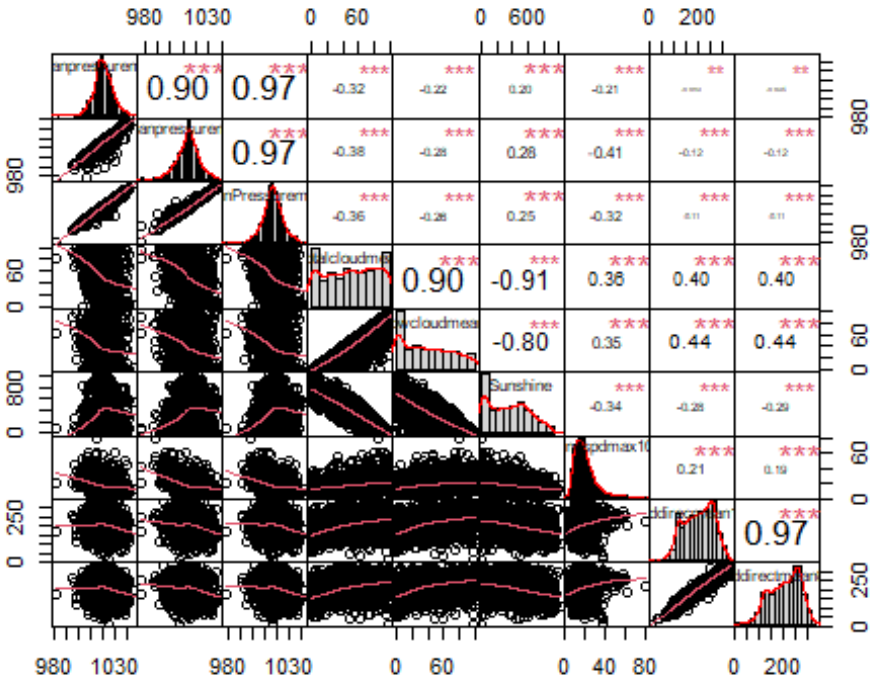
#Meanpressuremax,Meanpressuremin ,Meanpressuremean,Totalcloudmean,
#Lowcloudmean,Sunshine,Winspdmax10m,Windirecmean10m,Winddirectmean80m

1.3 Analyse complémentaire

Nous avons réalisé une analyse supplémentaire afin de s'assurer de la pertinence des variables présélectionnées

#Meanpressuremax,Meanpressuremin ,Meanpressuremean,Totalcloudmean,
#Lowcloudmean,Sunshine,Winspdmax10m,Windirecmean10m,Winddirectmean80m,

```
library(PerformanceAnalytics)
datavp <- data2[, c(29,30,9,12,15,16,39,19,21)]
chart.Correlation(datavp, histogram=TRUE, pch=19)
```



-Interprétation de la charte de corrélation

En haut de la diagonale : On a la valeur de la corrélation plus le niveau de signification en tant qu'étoiles :

En bas de la diagonale : les nuages de points bivariés avec une ligne ajustée sont affichés qui présente la linéarité ou non entre les variables

Dans notre charte nous constatons une forte corrélation entre les variables :

(Meanpressuremax,Meanpressuremin , Meanpressuremean): la valeur du coefficient de corrélation de Pearson correspondante : 0.90 et 0.97, avec une significativité élevée ($p < 0.001$) et enfin une corrélation linéaire positive

Totalcloudmean,Lowcloudmean: la valeur du coefficient de corrélation de Pearson correspondante : 0.90, avec une significativité élevée ($p < 0.001$) et enfin corrélation linéaire positive

Windirecmean10m,Winddirectmean80m : la valeur du coefficient de corrélation de Pearson correspondante : 0.97, avec une significativité élevée ($p < 0.001$) et enfin corrélation linéaire positive.

1.4 Détection des valeurs aberrantes

Les valeurs aberrantes dépendent de la distribution. Nous regardons encore une fois les statistiques de la base de données.

```
summary(data2)

##        X                Year          Month              Day             Ho
ur
##  Min.   :   2.0   Min.   :2010   Min.   : 1.000   Min.   : 1.0   Min.
:0
##  1st Qu.: 721.5   1st Qu.:2012   1st Qu.: 3.000   1st Qu.: 8.0   1st Qu.
:0
##  Median :1451.0   Median :2014   Median : 6.000   Median :16.0   Median
:0
##  Mean   :1459.8   Mean   :2014   Mean   : 6.436   Mean   :15.8   Mean
:0
##  3rd Qu.:2189.0   3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:23.0   3rd Qu.
:0
##  Max.   :2940.0   Max.   :2018   Max.   :12.000   Max.   :31.0   Max.
:0
##      Minute     Tempmean         Humimean     MeanPressuremean
##  Min.   :0   Min.   :-7.63   Min.   :38.33   Min.   : 978.9
##  1st Qu.:0   1st Qu.: 6.71   1st Qu.:64.82   1st Qu.:1012.4
##  Median :0   Median :12.08   Median :72.21   Median :1017.0
##  Mean   :0   Mean   :12.23   Mean   :71.40   Mean   :1017.0
##  3rd Qu.:0   3rd Qu.:17.54   3rd Qu.:78.63   3rd Qu.:1022.0
##  Max.   :0   Max.   :29.45   Max.   :95.54   Max.   :1042.4
##  Totalprecipitation    Snowfall        Totalcloudmean   Highcloudmean
##  Min.   : 0.000     Min.   :0.00000   Min.   :  0.00   Min.   :  0.000
##  1st Qu.: 0.000     1st Qu.:0.00000   1st Qu.: 23.80   1st Qu.:  1.657
##  Median : 0.100     Median :0.00000   Median : 51.67   Median : 11.880
##  Mean   : 2.085     Mean   :0.04965   Mean   : 50.76   Mean   : 20.284
##  3rd Qu.: 2.300     3rd Qu.:0.00000   3rd Qu.: 78.53   3rd Qu.: 33.260
##  Max.   :31.500     Max.   :8.61000   Max.   :100.00   Max.   :100.000
##  Mediumcloudmean   Lowcloudmean       Sunshine        Waveradia
##  Min.   :  0.00   Min.   :  0.00   Min.   :   0.0   Min.   : 265.2
##  1st Qu.:  1.83   1st Qu.:  9.42   1st Qu.: 114.3   1st Qu.:2096.2
##  Median : 24.98   Median : 36.35   Median : 366.8   Median :3675.3
##  Mean   : 31.50   Mean   : 39.34   Mean   : 373.1   Mean   :3984.6
```

```
##  3rd Qu.: 54.21   3rd Qu.: 65.76   3rd Qu.: 587.7   3rd Qu.:5723.6
##  Max.   :100.00   Max.   :100.00   Max.   :1015.8   Max.   :8363.3
##  Windspdmean10m   Winddirecmean10m Windspdmean80m  Winddirectmean80m
##  Min.   : 1.260   Min.   : 11.19   Min.   : 1.34   Min.   : 12.18
##  1st Qu.: 6.428   1st Qu.:152.40   1st Qu.: 8.68   1st Qu.:157.42
##  Median : 9.195   Median :206.36   Median :12.41   Median :213.78
##  Mean   :10.707   Mean   :201.82   Mean   :14.28   Mean   :206.23
##  3rd Qu.:12.977   3rd Qu.:254.19   3rd Qu.:17.61   3rd Qu.:259.06
##  Max.   :42.210   Max.   :331.67   Max.   :54.03   Max.   :333.43
##  Windspdmean900mb Winddirectmean900mb  Windgustmean      Tempmax
##  Min.   : 2.25    Min.   : 17.37      Min.   : 2.25   Min.   :-3.84
##  1st Qu.:13.02    1st Qu.:144.02      1st Qu.: 9.48   1st Qu.:10.58
##  Median :19.57    Median :233.47      Median :14.06   Median :16.54
##  Mean   :24.57    Mean   :206.22      Mean   :16.69   Mean   :16.54
##  3rd Qu.:32.10    3rd Qu.:265.93      3rd Qu.:21.15   3rd Qu.:22.36
##  Max.   :97.06    Max.   :344.82      Max.   :79.38   Max.   :35.77
##     Tempmin           Humimax          Humimin      Meanpressuremax
##  Min.   :-12.520   Min.   : 59.00   Min.   :19.00   Min.   : 981.9
##  1st Qu.:  3.350   1st Qu.: 83.00   1st Qu.:45.00   1st Qu.:1015.4
##  Median :  8.005   Median : 89.00   Median :54.00   Median :1019.5
##  Mean   :  8.062   Mean   : 87.69   Mean   :54.04   Mean   :1019.9
##  3rd Qu.: 13.092   3rd Qu.: 94.00   3rd Qu.:63.00   3rd Qu.:1024.7
##  Max.   : 23.940   Max.   :100.00   Max.   :92.00   Max.   :1045.4
##  Meanpressuremin Totalcloudmax    Totalcloudmin     Highcloudmax
##  Min.   : 977    Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:1009    1st Qu.:100.00   1st Qu.:  0.000   1st Qu.: 15.00
##  Median :1015    Median :100.00   Median :  0.000   Median : 97.00
##  Mean   :1014    Mean   : 88.23   Mean   :  8.692   Mean   : 60.17
##  3rd Qu.:1019    3rd Qu.:100.00   3rd Qu.:  2.400   3rd Qu.:100.00
##  Max.   :1039    Max.   :100.00   Max.   :100.000   Max.   :100.00
##   Highcloudmin      Mediumcloudmax   Mediumcloudmin     Lowcloudmax
##  Min.   :  0.0000   Min.   :  0.00   Min.   :  0.000   Min.   :  0
##  1st Qu.:  0.0000   1st Qu.: 22.75   1st Qu.:  0.000   1st Qu.:100
##  Median :  0.0000   Median :100.00   Median :  0.000   Median :100
##  Mean   :  0.9432   Mean   : 70.94   Mean   :  2.097   Mean   : 80
##  3rd Qu.:  0.0000   3rd Qu.:100.00   3rd Qu.:  0.000   3rd Qu.:100
##  Max.   :100.0000   Max.   :100.00   Max.   :100.000   Max.   :100
##   Lowcloudmin      Windspdmax10m   Windspdmin10m   Windspdmax80m
##  Min.   :  0.000   Min.   : 2.52   Min.   : 0.00   Min.   : 3.98
##  1st Qu.:  0.000   1st Qu.:12.32   1st Qu.: 1.14   1st Qu.:18.27
##  Median :  0.000   Median :17.36   Median : 2.41   Median :23.85
##  Mean   :  3.879   Mean   :19.06   Mean   : 3.57   Mean   :25.35
##  3rd Qu.:  0.000   3rd Qu.:23.44   3rd Qu.: 4.45   3rd Qu.:29.92
##  Max.   :100.000   Max.   :79.99   Max.   :27.73   Max.   :93.84
##  Windspdmin80m    Windspdmax900mb  Windspdmin900mb  Windgustmax
##  Min.   : 0.000   Min.   :  4.02   Min.   : 0.00   Min.   : 4.32
##  1st Qu.: 1.140   1st Qu.: 24.54   1st Qu.: 3.05   1st Qu.:19.08
##  Median : 2.600   Median : 37.12   Median : 6.73   Median :26.10
##  Mean   : 4.727   Mean   : 41.82   Mean   :11.09   Mean   :29.31
##  3rd Qu.: 5.830   3rd Qu.: 54.37   3rd Qu.:15.31   3rd Qu.:37.08
##  Max.   :37.700   Max.   :136.25   Max.   :76.13   Max.   :95.04
##   Windgustmin     pluie.demain
##  Min.   : 0.000   Mode :logical
##  1st Qu.: 2.160   FALSE:579
```

```
##  Median : 3.960   TRUE :601
##  Mean   : 6.502
##  3rd Qu.: 8.280
##  Max.   :57.960
```

Nous observons que les distributions des variables suivantes sont très étalées : #Totalprecipitation,Snowfall,Totalcloudmin,Highcloudmin,Mediumcloudmin,Lowcloudmin,Windspdmin10m,Windspdmin80m,Windgustmin #Windspdmin900mb. Les autres distributions semblent plutot cohérentes.

Pour le moment nous les considérons à priori comme aberrantes compte tenue de la distribution.

Afin de vérifier si ces variables sont aberrantes nous utilisons la méthode de discrétisation qui consiste à découper les variables en faisant une distinction entre min, médiane et max.

```
#discrétisation de variable Totalprécipitation
Breaksprec = c(0, 2 , max(Totalprecipitation))
Totalprecipitation.d = cut(Totalprecipitation, breaks = Breaksprec , include.lowest = TRUE)
summary(Totalprecipitation.d)
```

```
##    [0,2] (2,31.5]
##      866      314
```

```
#discrétisation de variable Snowfall
BreaksSnow = c(0, 0.04, max(Snowfall))
Snowfall.d = cut(Snowfall, breaks = BreaksSnow , include.lowest = TRUE)
summary(Snowfall.d)
```

```
##    [0,0.04] (0.04,8.61]
##        1132          48
```

```
#discrétisation de variable Totalcloudmin
BreaksTTcloumin = c(0, 8, max(Totalcloudmin))
Totalcloudmin.d = cut(Totalcloudmin, breaks = BreaksTTcloumin , include.lowest = TRUE)
summary(Totalcloudmin.d)
```

```
##   [0,8] (8,100]
##     972     208
```

```
#discrétisation de variable Highcloudmin
Breakshigcloumin = c(0, 0.9,  max(Highcloudmin))
Highcloudmin.d = cut(Highcloudmin, breaks = Breakshigcloumin , include.lowest = TRUE)
summary(Highcloudmin.d)
```

```
##   [0,0.9] (0.9,100]
##      1084        96
```

```
#discrétisation de variable Mediumcloudmin
Breaksmedcloumin = c(0, 2, max(Mediumcloudmin))
Mediumcloudmin.d = cut(Mediumcloudmin, breaks = Breaksmedcloumin , include.
```

```
lowest = TRUE)
summary(Mediumcloudmin.d)
```

```
##   [0,2] (2,100]
##    1114      66
```

```
#discrétisation de variable Lowcloudmin
Breakslowcloumin = c(0, 3, max(Lowcloudmin))
Lowcloudmin.d = cut(Lowcloudmin, breaks = Breakslowcloumin , include.lowest
= TRUE)
summary(Lowcloudmin.d)
```

```
##   [0,3] (3,100]
##    1081      99
```

```
#discrétisation de variable Windspdmin10m
Breakswindspmin10 = c(0, 2, max(Windspdmin10m))
Windspdmin10m.d = cut(Windspdmin10m, breaks = Breakswindspmin10 , include.l
owest = TRUE)
summary(Windspdmin10m.d)
```

```
##    [0,2] (2,27.7]
##      503      677
```

```
#discrétisation de variable Windspdmin80m
BreaksWindspdmin80 = c(0, 4, max(Windspdmin80m))
Windspdmin80m.d = cut(Windspdmin80m, breaks = BreaksWindspdmin80 , include.
lowest = TRUE)
summary(Windspdmin80m.d)
```

```
##    [0,4] (4,37.7]
##      760      420
```

```
#discrétisation de variable Windspdmin900mb
Breakswindspdmin900mb = c(0, 11, max(Windspdmin900mb))
Windspdmin900mb.d = cut(Windspdmin900mb, breaks = Breakswindspdmin900mb , i
nclude.lowest = TRUE)
summary(Windspdmin900mb.d)
```

```
##    [0,11] (11,76.1]
##       789       391
```

```
#discrétisation de variable Windgustmin
Breakswindgtmin = c(0, 6, max(Windgustmin))
Windgustmin.d = cut(Windgustmin, breaks = Breakswindgtmin , include.lowest
= TRUE)
summary(Windgustmin.d)
```

```
##  [0,6] (6,58]
##    778    402
```

La discrétisation nous a permis de purifier nos variables. En effet nous avons supprimé celles qui représentent une forte déviation par apport à la moyenne. La comparaison entre la médiane, min et max fait apparaitre que les variables suivantes sont étalées: #Totalprecipitation,Snowfall,Totalcloudmin,Highcloudmin,Mediumcloudmin,Lowcloudmin,Windspdmin10m,Windspdmin80m #et Windgustmin

1.5 Proposition d'un premier modèle

Nous nous sommes basés sur les tests analysés précédemment pour comparer entre plusieurs modèles et en choisir le plus significatif.

```
# Premier étape nous supprimons les variables aberrantes

g1=glm(pluie.demain~.-Totalprecipitation-Snowfall-Totalcloudmin-Highcloudmi
n-Mediumcloudmin-Lowcloudmin-Windspdmin10m
       -Windspdmin80m - Windgustmin -Hour-Minute-X-Day-Year-Month
       ,family = binomial, data = data2)
summary(g1)

##
## Call:
## glm(formula = pluie.demain ~ . - Totalprecipitation - Snowfall -
##     Totalcloudmin - Highcloudmin - Mediumcloudmin - Lowcloudmin -
##     Windspdmin10m - Windspdmin80m - Windgustmin - Hour - Minute -
##     X - Day - Year - Month, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3704  -0.8458   0.2806   0.8631   2.9720
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.153e+01  1.220e+01   5.042 4.61e-07 ***
## Tempmean             1.438e-01  1.608e-01   0.894 0.371119
## Humimean             2.040e-02  3.162e-02   0.645 0.518732
## MeanPressuremean     5.029e-01  1.367e-01   3.680 0.000234 ***
## Totalcloudmean       9.295e-03  1.146e-02   0.811 0.417312
## Highcloudmean       -1.807e-03  6.122e-03  -0.295 0.767944
## Mediumcloudmean      8.326e-03  5.786e-03   1.439 0.150146
## Lowcloudmean         4.030e-03  7.219e-03   0.558 0.576649
## Sunshine             5.177e-04  8.605e-04   0.602 0.547415
## Waveradia            2.467e-05  9.178e-05   0.269 0.788128
## Windspdmean10m       4.510e-02  8.384e-02   0.538 0.590575
## Winddirecmean10m     4.731e-03  5.603e-03   0.844 0.398441
## Windspdmean80m      -1.081e-01  6.164e-02  -1.753 0.079536 .
## Winddirectmean80m   -8.541e-03  5.809e-03  -1.470 0.141498
## Windspdmean900mb     1.159e-02  2.463e-02   0.471 0.637880
## Winddirectmean900mb  5.243e-03  1.407e-03   3.726 0.000194 ***
## Windgustmean         2.907e-02  3.213e-02   0.905 0.365526
## Tempmax              2.458e-03  9.416e-02   0.026 0.979175
## Tempmin             -1.068e-01  8.492e-02  -1.257 0.208688
## Humimax             -4.430e-03  2.013e-02  -0.220 0.825797
## Humimin             -1.334e-02  1.799e-02  -0.742 0.458158
## Meanpressuremax     -2.527e-01  7.124e-02  -3.547 0.000389 ***
```

```
## Meanpressuremin     -3.150e-01  7.611e-02  -4.139 3.48e-05 ***
## Totalcloudmax        3.899e-03  4.787e-03   0.815 0.415339
## Highcloudmax         3.550e-03  2.787e-03   1.274 0.202799
## Mediumcloudmax       6.201e-03  3.069e-03   2.021 0.043294 *
## Lowcloudmax          8.179e-04  3.255e-03   0.251 0.801624
## Windspdmax10m        3.071e-02  3.260e-02   0.942 0.346248
## Windspdmax80m        1.419e-02  2.724e-02   0.521 0.602431
## Windspdmax900mb     -7.957e-03  1.166e-02  -0.682 0.495090
## Windspdmin900mb      1.279e-03  1.749e-02   0.073 0.941679
## Windgustmax          5.716e-03  1.525e-02   0.375 0.707891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4 on 1179  degrees of freedom
## Residual deviance: 1258.1  on 1148  degrees of freedom
## AIC: 1322.1
##
## Number of Fisher Scoring iterations: 4

# Nous constatons que le modèle g1 présente plusieurs variables ayant une v
aleur p-value non significatif avec un AIC=1322.1



# Nous essayons d'améliorer sa qualité dans l'étape suivantes (g2)

# Deuxième étape suppression de certaines variables qui présentent

une forte corrélation

#modèle g2


g2=glm(pluie.demain~.-Totalprecipitation-Snowfall-Totalcloudmin-Highcloudmi
n-Mediumcloudmin-Lowcloudmin-Windspdmin10m
       -Windspdmin80m - Windgustmin -Hour-Minute-X-Day-Year-Month
     -Tempmax-Tempmin -Humimean -Windspdmean10m-Windgustmean-Windspdmax80m
     -Windspdmin900mb-Windspdmax900mb,family = binomial, data = data2)
summary(g2)

##
## Call:
## glm(formula = pluie.demain ~ . - Totalprecipitation - Snowfall -
##     Totalcloudmin - Highcloudmin - Mediumcloudmin - Lowcloudmin -
##     Windspdmin10m - Windspdmin80m - Windgustmin - Hour - Minute -
##     X - Day - Year - Month - Tempmax - Tempmin - Humimean - Windspdmean1
0m -
##     Windgustmean - Windspdmax80m - Windspdmin900mb - Windspdmax900mb,
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4170  -0.8434   0.3069   0.8742   2.9450
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.132e+01  1.209e+01   5.072 3.94e-07 ***
## Tempmean             3.910e-02  1.736e-02   2.253 0.024283 *
## MeanPressuremean     4.738e-01  1.326e-01   3.572 0.000354 ***
## Totalcloudmean       6.728e-03  1.132e-02   0.594 0.552337
## Highcloudmean       -1.766e-03  6.072e-03  -0.291 0.771227
## Mediumcloudmean      8.692e-03  5.741e-03   1.514 0.129998
## Lowcloudmean         4.854e-03  7.104e-03   0.683 0.494470
## Sunshine             3.078e-04  8.456e-04   0.364 0.715844
## Waveradia            7.937e-05  8.045e-05   0.987 0.323822
## Winddirecmean10m     4.120e-03  5.542e-03   0.743 0.457300
## Windspdmean80m      -5.082e-02  2.356e-02  -2.157 0.031015 *
## Winddirectmean80m   -8.042e-03  5.782e-03  -1.391 0.164247
## Windspdmean900mb     9.529e-03  8.922e-03   1.068 0.285530
## Winddirectmean900mb  4.914e-03  1.375e-03   3.574 0.000351 ***
## Humimax              7.608e-03  1.153e-02   0.660 0.509454
## Humimin             -7.970e-03  9.072e-03  -0.878 0.379702
## Meanpressuremax     -2.365e-01  6.954e-02  -3.401 0.000671 ***
## Meanpressuremin     -3.014e-01  7.337e-02  -4.108 3.99e-05 ***
## Totalcloudmax        3.913e-03  4.762e-03   0.822 0.411282
## Highcloudmax         3.480e-03  2.762e-03   1.260 0.207775
## Mediumcloudmax       6.179e-03  3.037e-03   2.035 0.041888 *
## Lowcloudmax          1.511e-04  3.211e-03   0.047 0.962450
## Windspdmax10m        4.106e-02  2.185e-02   1.879 0.060215 .
## Windgustmax          1.052e-02  1.123e-02   0.937 0.348922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1262.7  on 1156  degrees of freedom
## AIC: 1310.7
##
## Number of Fisher Scoring iterations: 4
```

 # Dans cette étape nous observons que l'AIC du modèle g2 a diminué.

Nous appliquons par la suite l'Anova sur ce modèle (g2) pour voir quelle co variable ayant un p-value non sgnificatif

```
anova(g2,test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: pluie.demain
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 1179     1635.4
```

```
## Tempmean             1   16.508      1178     1618.9 4.845e-05 ***
## MeanPressuremean     1  167.501      1177     1451.4 < 2.2e-16 ***
## Totalcloudmean       1   76.024      1176     1375.4 < 2.2e-16 ***
## Highcloudmean        1    8.694      1175     1366.7 0.0031921 **
## Mediumcloudmean      1    9.004      1174     1357.7 0.0026934 **
## Lowcloudmean         1    0.378      1173     1357.3 0.5386986
## Sunshine             1    1.483      1172     1355.8 0.2232984
## Waveradia            1    0.021      1171     1355.8 0.8838512
## Winddirecmean10m     1    0.538      1170     1355.3 0.4633002
## Windspdmean80m       1   10.982      1169     1344.3 0.0009198 ***
## Winddirectmean80m    1    0.298      1168     1344.0 0.5849784
## Windspdmean900mb     1    4.906      1167     1339.1 0.0267566 *
## Winddirectmean900mb  1   18.478      1166     1320.6 1.718e-05 ***
## Humimax              1    0.449      1165     1320.2 0.5025771
## Humimin              1    2.196      1164     1318.0 0.1383818
## Meanpressuremax      1    0.588      1163     1317.4 0.4433019
## Meanpressuremin      1   27.140      1162     1290.2 1.893e-07 ***
## Totalcloudmax        1    8.975      1161     1281.2 0.0027375 **
## Highcloudmax         1    6.091      1160     1275.2 0.0135878 *
## Mediumcloudmax       1    4.361      1159     1270.8 0.0367698 *
## Lowcloudmax          1    0.127      1158     1270.7 0.7215241
## Windspdmax10m        1    7.062      1157     1263.6 0.0078738 **
## Windgustmax          1    0.881      1156     1262.7 0.3479278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suite à cette étape nous avons supprimé les variables : Lowcloudmean ,Lowcloudmax,Humimax

#modèle g3

```
g3=glm(pluie.demain~.-Totalprecipitation-Snowfall-Totalcloudmin-Highcloudmin-Mediumcloudmin-Lowcloudmin-Windspdmin10m
       -Windspdmin80m - Windgustmin -Hour-Minute-X-Day-Year-Month
     -Tempmax-Tempmin -Humimean -Windspdmean10m-Windgustmean-Windspdmax80m
     -Windspdmin900mb-Windspdmax900mb
     -Humimax-Lowcloudmax- Highcloudmean
,family = binomial, data = data2)
summary(g3)
```

```
##
## Call:
## glm(formula = pluie.demain ~ . - Totalprecipitation - Snowfall -
##     Totalcloudmin - Highcloudmin - Mediumcloudmin - Lowcloudmin -
##     Windspdmin10m - Windspdmin80m - Windgustmin - Hour - Minute -
##     X - Day - Year - Month - Tempmax - Tempmin - Humimean - Windspdmean10m -
##     Windgustmean - Windspdmax80m - Windspdmin900mb - Windspdmax900mb -
##     Humimax - Lowcloudmax - Highloudmean, family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4104  -0.8478   0.3063   0.8714   2.9314
```

```
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           6.174e+01  1.205e+01   5.125 2.98e-07 ***
## Tempmean              3.744e-02  1.705e-02   2.196 0.028125 *
## MeanPressuremean      4.787e-01  1.325e-01   3.613 0.000303 ***
## Totalcloudmean        6.010e-03  1.117e-02   0.538 0.590514
## Mediumcloudmean       8.316e-03  5.442e-03   1.528 0.126468
## Lowcloudmean          5.848e-03  6.833e-03   0.856 0.392070
## Sunshine              3.335e-04  8.423e-04   0.396 0.692188
## Waveradia             8.415e-05  7.979e-05   1.055 0.291600
## Winddirecmean10m      4.438e-03  5.520e-03   0.804 0.421469
## Windspdmean80m       -5.295e-02  2.313e-02  -2.289 0.022062 *
## Winddirectmean80m    -8.370e-03  5.751e-03  -1.455 0.145561
## Windspdmean900mb      9.334e-03  8.875e-03   1.052 0.292944
## Winddirectmean900mb   4.899e-03  1.375e-03   3.563 0.000367 ***
## Humimin              -5.237e-03  8.145e-03  -0.643 0.520241
## Meanpressuremax      -2.378e-01  6.952e-02  -3.421 0.000625 ***
## Meanpressuremin      -3.049e-01  7.321e-02  -4.165 3.11e-05 ***
## Totalcloudmax         4.303e-03  3.823e-03   1.125 0.260408
## Highcloudmax          3.103e-03  2.266e-03   1.369 0.170998
## Mediumcloudmax        6.138e-03  2.851e-03   2.153 0.031310 *
## Windspdmax10m         4.249e-02  2.167e-02   1.961 0.049908 *
## Windgustmax           1.025e-02  1.122e-02   0.914 0.360703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1263.2  on 1159  degrees of freedom
## AIC: 1305.2
##
## Number of Fisher Scoring iterations: 4

anova(g3,g2,test = "LRT") # g3 sgnificatif

# Effectuer une comparaison entre les modèles g2 et g3
## Analysis of Deviance Table)
#g3# 1      1159     1263.2
#g2# 2      1156     1262.7  3  0.51143   0.9164
```

le modèle g2 possède un p-value non significatif 0.9164 *donc on choisit le modèle g3*

```
# Effectuer un comparaison entre les modèles g1 et g3
anova(g3,g1,test = "LRT") # g3 sgnificatif
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
# g3# 1      1159     1263.2
# g1# 2      1148     1258.1 11   5.1075   0.9258
```

le modele g1 possède un p-value non significatif 0.9258 *donc on choisit le modèle g3*

```r
# Appliquer les Critères AIC et BIC pour comparer g1 et g3
c(BIC(g3),BIC(g1))
```

```
## [1] 1411.780 1484.478
```

```r
c(AIC(g3),AIC(g1))
```

```
## [1] 1305.241 1322.134
```

```r
#Après cette comparaison nous choisissons le modèle g3

#  choisir le modele g3
step(g3)
```

```
## Call:  glm(formula pluie.demain ~ Tempmean + MeanPressuremean + Mediumcl
oudmean +
##     Lowcloudmean + Waveradia + Windspdmean80m + Winddirectmean80m +
##     Windspdmean900mb + Winddirectmean900mb + Meanpressuremax +
##     Meanpressuremin + Highcloudmax + Mediumcloudmax + Windspdmax10m,
##     family = binomial, data = data2)
##
## Coefficients:
##         (Intercept)             Tempmean     MeanPressuremean
##           63.2364460            0.0348834            0.5060734
##     Mediumcloudmean         Lowcloudmean            Waveradia
##           0.0086752            0.0098179            0.0001235
##      Windspdmean80m    Winddirectmean80m     Windspdmean900mb
##          -0.0543200           -0.0038980            0.0127039
## Winddirectmean900mb      Meanpressuremax      Meanpressuremin
##           0.0046515           -0.2509566           -0.3204917
##        Highcloudmax       Mediumcloudmax        Windspdmax10m
##           0.0035623            0.0079133            0.0541016
##
## Degrees of Freedom: 1179 Total (i.e. Null);  1165 Residual
## Null Deviance:       1635
## Residual Deviance: 1267  AIC: 1297
```

```r
#modèle g4=step(g3)

g4=glm( pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean +
    Lowcloudmean + Waveradia + Windspdmean80m + Winddirectmean80m +
    Windspdmean900mb + Winddirectmean900mb + Meanpressuremax +
    Meanpressuremin + Highcloudmax + Mediumcloudmax + Windspdmax10m,
    family = binomial, data = data2)
summary(g4)
```

```
##
## Call:
## glm(formula = pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudme
an +
##     Lowcloudmean + Waveradia + Windspdmean80m + Winddirectmean80m +
##     Windspdmean900mb + Winddirectmean900mb + Meanpressuremax +
##     Meanpressuremin + Highcloudmax + Mediumcloudmax + Windspdmax10m,
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3607  -0.8449   0.2870   0.8730   2.8411
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.324e+01  1.189e+01   5.320 1.04e-07 ***
## Tempmean             3.488e-02  1.674e-02   2.084 0.037146 *
## MeanPressuremean     5.061e-01  1.314e-01   3.853 0.000117 ***
## Mediumcloudmean      8.675e-03  4.159e-03   2.086 0.037011 *
## Lowcloudmean         9.818e-03  3.449e-03   2.847 0.004419 **
## Waveradia            1.235e-04  6.162e-05   2.004 0.045038 *
## Windspdmean80m      -5.432e-02  2.267e-02  -2.396 0.016581 *
## Winddirectmean80m   -3.898e-03  1.549e-03  -2.517 0.011851 *
## Windspdmean900mb     1.270e-02  7.949e-03   1.598 0.109984
## Winddirectmean900mb  4.651e-03  1.295e-03   3.592 0.000328 ***
## Meanpressuremax     -2.510e-01  6.905e-02  -3.634 0.000279 ***
## Meanpressuremin     -3.205e-01  7.244e-02  -4.424 9.67e-06 ***
## Highcloudmax         3.562e-03  2.232e-03   1.596 0.110545
## Mediumcloudmax       7.913e-03  2.572e-03   3.077 0.002091 **
## Windspdmax10m        5.410e-02  1.892e-02   2.859 0.004244 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4 on 1179  degrees of freedom
## Residual deviance: 1267.1  on 1165  degrees of freedom
## AIC: 1297.1
##
## Number of Fisher Scoring iterations: 4
```

1.6 Interprétation de modèle choisi

Nous nous intéressons plus précisément aux estimateurs et au p-value.

On prend les exemples de deux covariables suivantes.

MeanPressuremean est significative avec un impact positif sur la probabilité que la pluie tombe.

exp(5.061e-01)=1.658809 x Probabilité

Cependant Windspdmean80m est significative mais son impact sur la probabilité que la pluie tombe est négative.

exp(-5.061e-01)=0.947129xProbabilité

Evaluation de la qualité de modèle choisi (g4=step(g3))

Nous nous focalisons sur la déviance du modèle. Les tests de rapport des vraisemblances et le calcul de la p_value à l'écart de degré de liberté entre le modèle Null à la cste et le modèle retenu qui nous donne la significativité globale du modèle.

La sortie nous indique :

Null deviance: 1635.4 on 1179 degrees of freedom

Residual deviance: 1267.1 on 1165 degrees of freedom

```
pchisq(1635.4 - 1267.1 , 1179 - 1165 , lower = F)
```

```
## [1] 5.924801e-70
```

On remarque que P-value est très faible donc on rejette le modèle sans covariable et on garde notre modèle=Notre modèle est utile

Dans le sommaire du résultat de glm, la déviance du modèle ajusté est indiquée comme Residual Deviance. Le sommaire inclut aussi une autre valeur, Null Deviance, qui correspond à la déviance du modèle nul ne comptant aucun prédicteur. Ces deux valeurs jouent un rôle semblable à la somme des écarts carrés résiduels et la somme des écarts carrés totaux. On peut donc définir le pseudo R2 (ou R2 de McFadden) comme la fraction de la déviance du modèle nul expliquée par le modèle incluant les prédicteurs.

```
#### Extraction des coefficients du modele
coef(g4)
```

```
##         (Intercept)            Tempmean    MeanPressuremean     Mediumcl
oudmean
##       63.2364460146        0.0348834336        0.5060733975        0.008
6752062
##        Lowcloudmean           Waveradia      Windspdmean80m   Winddirect
mean80m
##        0.0098178921        0.0001235098       -0.0543199698       -0.003
8979783
##    Windspdmean900mb Winddirectmean900mb     Meanpressuremax     Meanpres
suremin
##        0.0127039416        0.0046514695       -0.2509565565       -0.320
4916864
##        Highcloudmax      Mediumcloudmax       Windspdmax10m
##        0.0035623035        0.0079132535        0.0541015551
```

```
#### Extraction des résidus
resid(g4)
```

```
#### pseudo_R2
library(DescTools)
PseudoR2(g4)
```

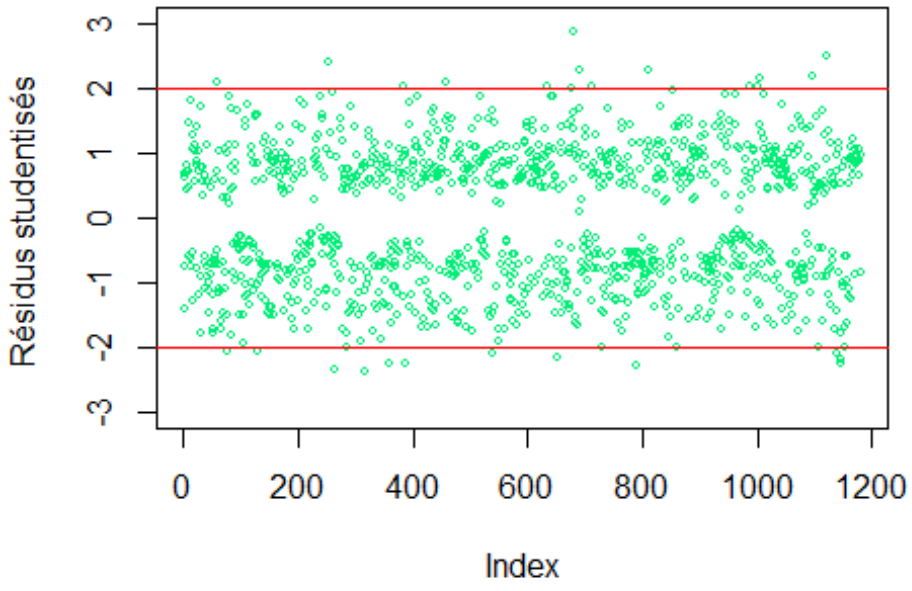
```
##  McFadden
## 0.2251859
```

```
PseudoR2(g4,"all")

##         McFadden     McFaddenAdj        CoxSnell      Nagelkerke   Aldric
hNelson
##        0.2251859       0.2068419       0.2680884       0.3574927       0.
2378604
## VeallZimmermann           Efron McKelveyZavoina            Tjur
AIC
##        0.4094835       0.2861717       0.3752691       0.2829294    1297.
1443390
##              BIC          logLik         logLik0              G2
##    1373.2433847    -633.5721695    -817.7085764     368.2728139
```

Pseudo R2 fait apparaitre plusieurs critères avec des valeurs différentes. Nous ne pouvons pas avoir des informations claires permettant de valider notre modèle. Nous utilisons donc AIC et BIC pour l'évaluation de modèle

Après avoir obtenu un modèle, nous diagnostiquons la régression afin de valider ou non le modèle. L'analyse des résidus est de ce point de vue très importante. Il est important de noter qu'en régression logistique, on s'intéresse la plupart du temps aux résidus de déviance. On construit généralement un index plot pour détecter les valeurs aberrantes (en dehors des lignes).

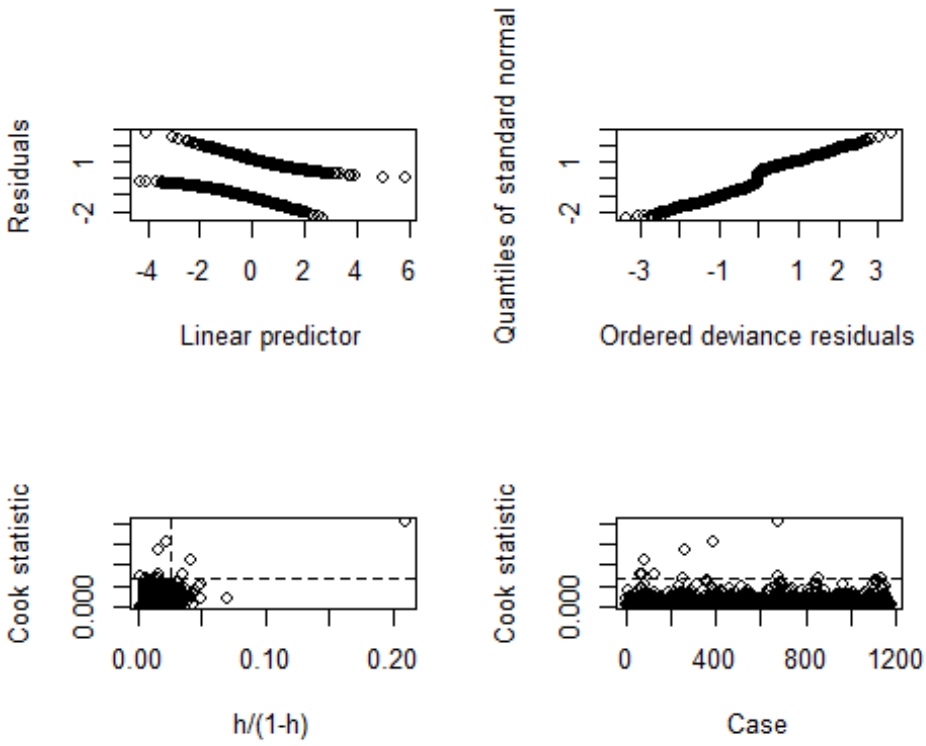
```
par(mfrow = c(1, 1))
plot(rstudent(g4), type = "p", cex = 0.5, ylab = "Résidus studentisés ",
    col = "springgreen2", ylim = c(-3, 3))
abline(h = c(-2, 2), col = "red")
```



Le graphique des résidus affiche une répartition relativement homogène des résidus, on constate alors que la distribution des résidus est symétrique autour de 0, la symétrie de résidu est un signe que leur distribution suit la loi normale, On remarque aussi la présence de quelques points aberrants(en dehors de lignes).

Analyse de la distribution des résidus suivant la loi normale

```
library(boot)
glm.diag.plots(g4)
```



Interprétation

Le diagramme **Quantile-Quantile** Q-Q plot (en haut à droite)montre que les queues droite et gauche sont petites et les valeurs extrêmes du graphique tombent près du centre sauf une petite déviation au milieu . Nous avons donc une distribution uniforme des données.

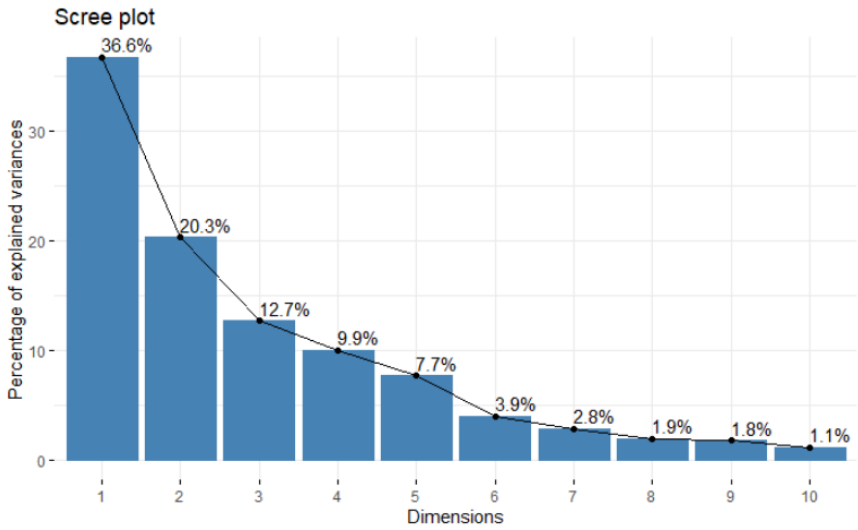
Analyse ACP (modèle g4)

Nous appliquons sur les 14 variables de g4 une analyse ACP afin de les synthétiser en quelles nouvelles variables appeler composantes principalement qui peuvent être visualiser graphiquement.

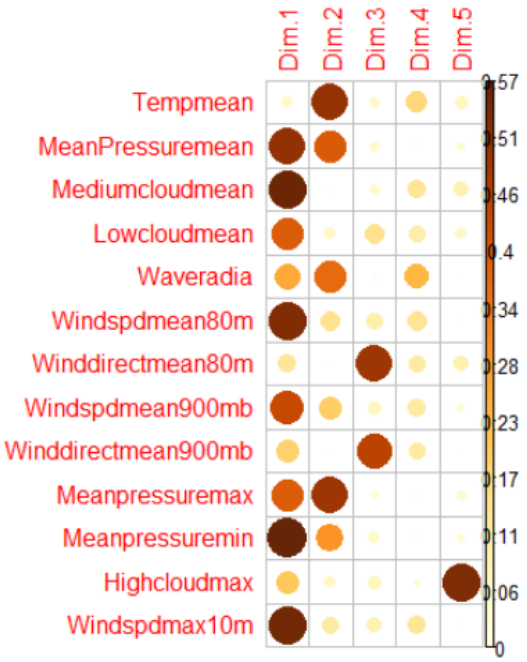
```r
{r}
rary("FactoMineR")
rary("factoextra")
rary("yarrr")

a2.global=data2[,c(7,9,14,15,17,20:23,29,30,33,39,47)]
mary(data2.global)

lection du nombre de composantes: Nous utilisons le critère du coude pour le choix des axes

z_eig(pca.global,addlabels = TRUE)

z_cos2(pca.global, choice = "var", axes = 1)

z_cos2(pca.global, choice = "var", axes = 2)

proportion d'inertie expliquée par les 3 premiers axes est de de 69 %. Cela reste acceptable pour 14 variables.:

mmary(mkt.pca.global)Description des axes selon les variables
.global=PCA(data2.global,quali.sup=14,graph=FALSE)
t(pca.global,choix = "var",cex=0.75)

orrélation des varibales avec les différentes dimensions
rplot(var$cos2, is.corr=FALSE)
```

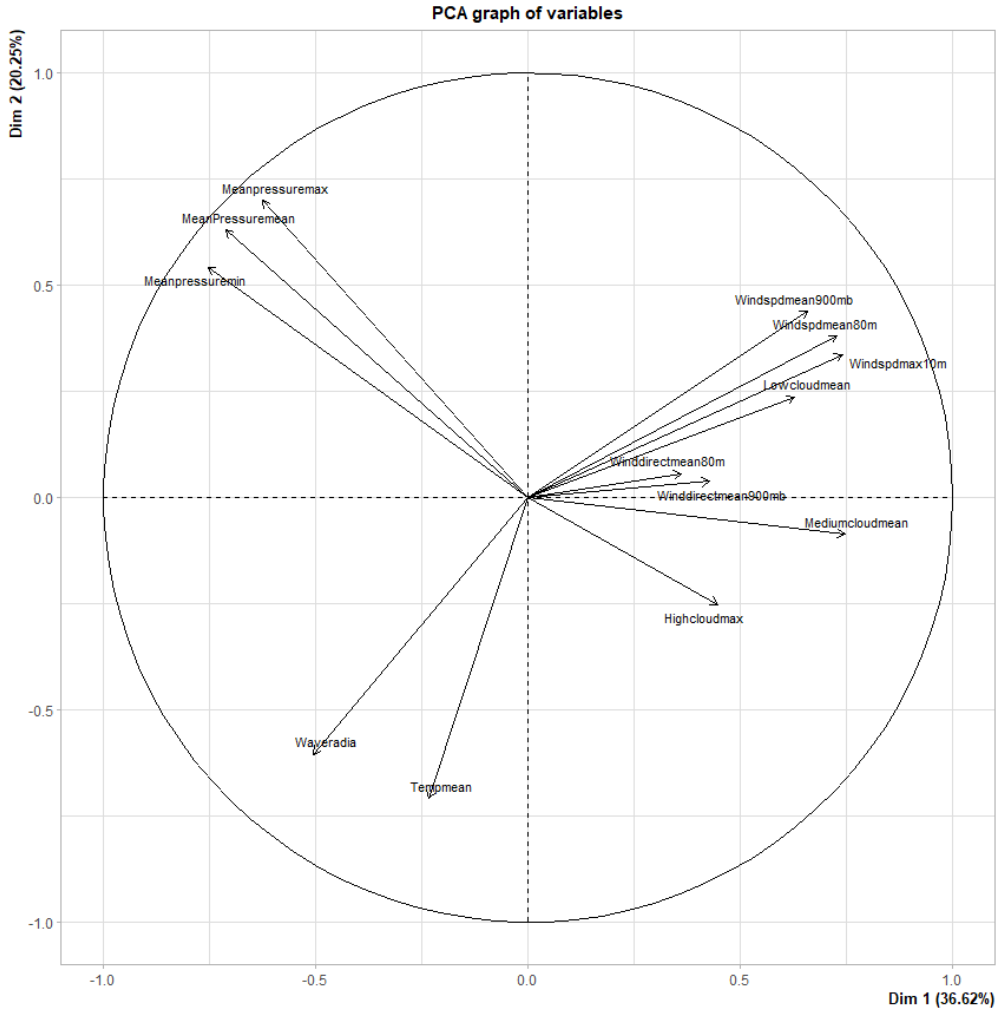
Sélection du nombre de composantes : Nous utilisons le critère du coude pour le choix des axes

La proportion d'inertie expliquée par les 3 premiers axes est de de 69 %. Cela reste acceptable pour 14 variables.



Présentation de la corrélation des variables avec les différentes dimensions





#Le graphique ci-dessus est également connu sous le nom de graphique de corrélation des variables. Il montre les relations #entre toutes les variables. Il peut être interprété comme suit :

Les variables positivement corrélées sont regroupées.

Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique.

La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP

1.7 Régression par Recherche Exhaustive

Autre méthode de sélection d'un meilleur modèle est de celle Best subset. Il s'agit d'une technique de construction de modèle permettant de trouver le meilleur groupe (sous-ensemble) de variables prédictives qui prévoient le mieux les réponses d'une variable dépendante. La sélection de modèle peut être vue comme la recherche du modèle optimal, au sens d'un critère choisi, parmi toutes les possibilités. On voudrait chercher le modèle de régression qui explique le mieux si la pluie tombe le lendemain ou non.

La principale fonction pour faire de la sélection de variables est regsubsets

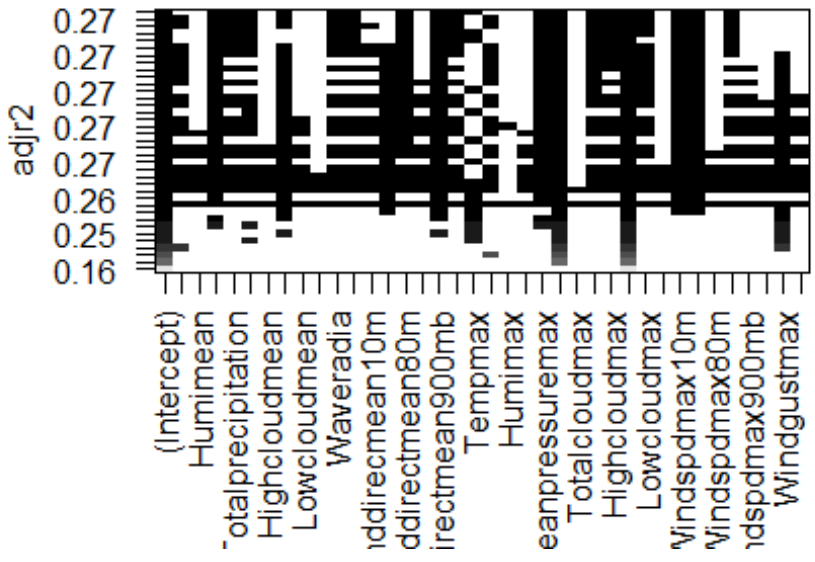
```
#Utiliser la fonction regsubsets() (du package leaps pour effectuer une sél
ection de variables via l'approche exhaustive Best Subset.
library(leaps)
 res1 = regsubsets(pluie.demain~.-X-Year-Month-Day-Hour-Minute-Snowfall-Hig
hcloudmin-Mediumcloudmin,data=data2,
                   nbest = 1,       # 1 seul meilleur pour chaque nombre d
e variables
                   nvmax = NULL,    # NULL pas de limites pour le
                     force.in = NULL,  # pas de variables à inclure de for
ce
                     force.out = NULL,  # pas de variables à exclure de fo
rce.
            method = "exhaustive")  # choix de la méthode exhaustive)
names(res1)
```

```
##  [1] "np"        "nrbar"     "d"         "rbar"      "thetab"    "first"
##  [7] "last"      "vorder"    "tol"       "rss"       "bound"     "nvmax"
## [13] "ress"      "ir"        "nbest"     "lopt"      "il"        "ier"
## [19] "xnames"    "method"    "force.in"  "force.out" "sserr"     "interc
ept"
## [25] "lindep"    "nullrss"   "nn"        "call"
```

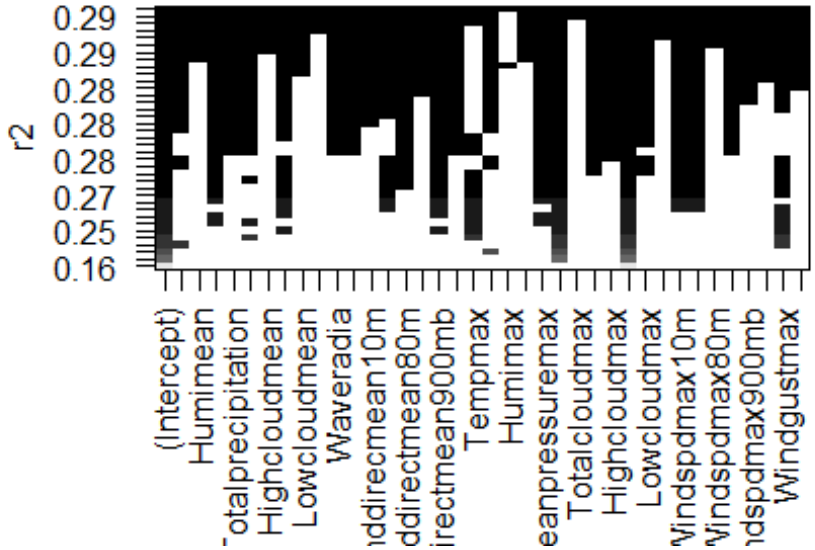
```
reg.summary <- summary(res1)
reg.summary
```

Nous réalisons des plots suivants selon les critères : `adjr2`, BIC et R2

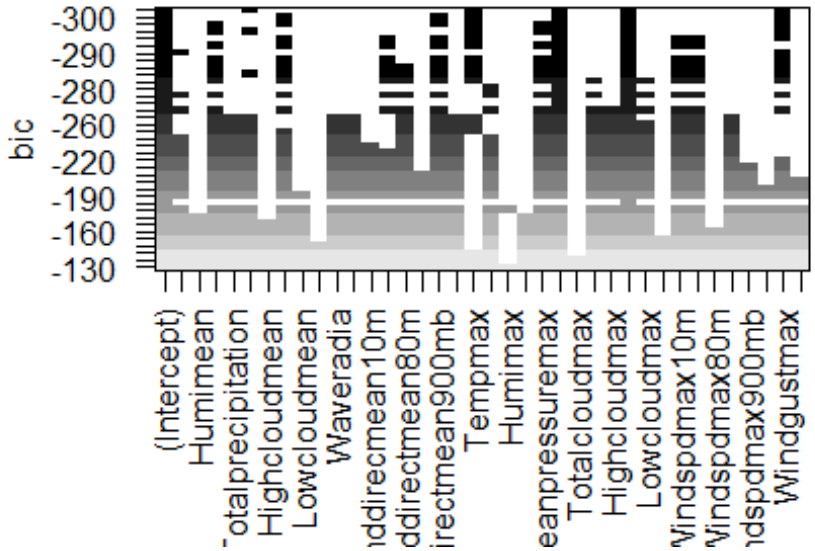
```
plot(res1,scale="adjr2")
```




```
 plot(res1,scale="r2")
```




```
plot(res1,scale ="bic")
```



Pour choisir le modèle à sélectionner, nous identifions l'emplacement du point maximum / minimum pour chaque critère : RSS, R2 ajusté, Cp et BIC. Dans chaque cas, afficher les variables sélectionnées.

```
min.rss <- which.min(reg.summary$rss)
max.adjr2 <- which.max(reg.summary$adjr2)
```

```
min.cp <- which.min(reg.summary$cp)
min.bic <- which.min(reg.summary$bic)
#
min.rss
```

```
## [1] 37
```

```
#
max.adjr2
```

```
## [1] 19
```

```
#
min.cp
```

```
## [1] 16
```

```
#
min.bic
```

```
## [1] 5
```

La liste des variables sélectionnées en se basant chaque fois sur les critères : RSS, BIC, CP et adjr2

```
names(which(reg.summary$which[min.rss,]==TRUE))
```

```
##  [1] "(Intercept)"         "Tempmean"            "Humimean"
##  [4] "MeanPressuremean"    "Totalprecipitation"  "Totalcloudmean"
##  [7] "Highcloudmean"       "Mediumcloudmean"     "Lowcloudmean"
## [10] "Sunshine"            "Waveradia"           "Windspdmean10m"
## [13] "Winddirecmean10m"    "Windspdmean80m"      "Winddirectmean80m"
## [16] "Windspdmean900mb"    "Winddirectmean900mb" "Windgustmean"
## [19] "Tempmax"             "Tempmin"             "Humimax"
## [22] "Humimin"             "Meanpressuremax"     "Meanpressuremin"
## [25] "Totalcloudmax"       "Totalcloudmin"       "Highcloudmax"
## [28] "Mediumcloudmax"      "Lowcloudmax"         "Lowcloudmin"
## [31] "Windspdmax10m"       "Windspdmin10m"       "Windspdmax80m"
## [34] "Windspdmin80m"       "Windspdmax900mb"     "Windspdmin900mb"
## [37] "Windgustmax"         "Windgustmin"
```

```
names(which(reg.summary$which[max.adjr2,]==TRUE))
```

```
##  [1] "(Intercept)"         "MeanPressuremean"    "Totalprecipitation"
##  [4] "Totalcloudmean"      "Mediumcloudmean"     "Waveradia"
##  [7] "Windspdmean10m"      "Winddirectmean80m"   "Winddirectmean900mb"
## [10] "Windgustmean"        "Tempmax"             "Meanpressuremax"
## [13] "Meanpressuremin"     "Totalcloudmin"       "Highcloudmax"
## [16] "Mediumcloudmax"      "Lowcloudmax"         "Windspdmax10m"
## [19] "Windspdmin10m"       "Windspdmin80m"
```

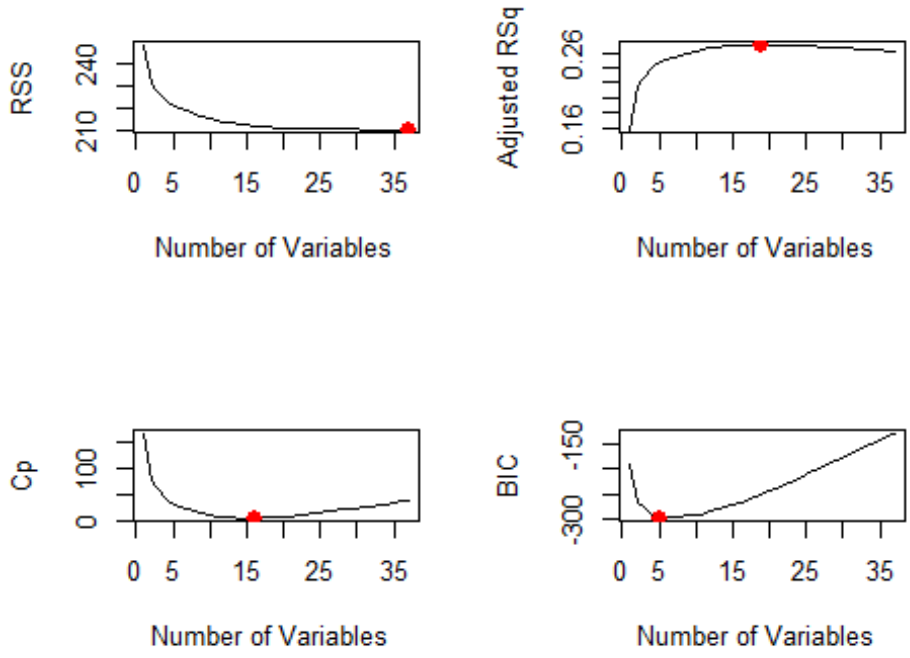
```
names(which(reg.summary$which[min.cp,]==TRUE))
```

```
##  [1] "(Intercept)"         "Tempmean"            "MeanPressuremean"
##  [4] "Mediumcloudmean"     "Windspdmean80m"      "Winddirectmean80m"
##  [7] "Winddirectmean900mb" "Tempmin"             "Meanpressuremax"
## [10] "Meanpressuremin"     "Totalcloudmin"       "Highcloudmax"
```

```
## [13] "Mediumcloudmax"      "Lowcloudmax"         "Windspdmax10m"
## [16] "Windspdmin10m"       "Windgustmax"

names(which(reg.summary$which[min.bic,]==TRUE))

## [1] "(Intercept)"     "Totalcloudmean"  "Tempmax"         "Meanpressurem
in"
## [5] "Mediumcloudmax"  "Windgustmax"
```

Sur une même fenêtre graphique nous représentons les courbes des différents critères. Nous ajoutons sur chaque courbe, le maximum/minimum correspondant.

```
par(mfrow =c(2,2))
plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")
points(min.rss,reg.summary$rss[min.rss],col ="red",cex =2, pch =20)
plot(reg.summary$adjr2,xlab="Number of Variables ",ylab="Adjusted RSq",type
="l")
points(max.adjr2,reg.summary$adjr2[max.adjr2],col ="red",cex =2, pch =20)
plot(reg.summary$cp,xlab="Number of Variables ",ylab="Cp",type="l")
points(min.cp,reg.summary$cp[min.cp],col ="red",cex =2, pch =20)
plot(reg.summary$bic,xlab="Number of Variables ",ylab="BIC",type="l")
points(min.bic,reg.summary$bic[min.bic],col ="red",cex =2, pch =20)
```



Nous réalisons une régression avec le meilleur modèle selon la statistique BIC

```
var.bic <- names(which(reg.summary$which[min.bic,]==TRUE))
var.bic.formula <- paste("pluie.demain", "~", paste(var.bic[-1], collapse="
+ "))
var.bic.formula

## [1] "pluie.demain ~ Totalcloudmean + Tempmax + Meanpressuremin + Mediumc
loudmax + Windgustmax"
```

```
best.model <- glm(var.bic.formula,family=binomial, data=data2)
summary(best.model)
```

```
##
## Call:
## glm(formula = var.bic.formula, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2825  -0.8975   0.3940   0.8631   2.6940
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     59.236757  10.376489   5.709 1.14e-08 ***
## Totalcloudmean   0.012263   0.002987   4.105 4.04e-05 ***
## Tempmax          0.064181   0.010523   6.099 1.07e-09 ***
## Meanpressuremin -0.061488   0.010128  -6.071 1.27e-09 ***
## Mediumcloudmax   0.010884   0.002060   5.284 1.27e-07 ***
## Windgustmax      0.022965   0.005630   4.079 4.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1300.9  on 1174  degrees of freedom
## AIC: 1312.9
##
## Number of Fisher Scoring iterations: 4
```

Le résultat de glm montre que les 5 variables ayant tous un p-value significatif avec un AIC de 1312. Ces variables, malgré leur importance, n'explique qu'une partie des informations. D 'autres variables pourraient compléter notre modèle.

Choisir le meilleur modèle suivant le critère RSS

```
var.rss <- names(which(reg.summary$which[min.rss,]==TRUE))
var.rss.formula <- paste("pluie.demain", "~", paste(var.rss[-1], collapse="
+ "))
var.rss.formula
```

```
## [1] "pluie.demain ~ Tempmean + Humimean + MeanPressuremean + Totalprecip
itation + Totalcloudmean + Highcloudmean + Mediumcloudmean + Lowcloudmean +
Sunshine + Waveradia + Windspdmean10m + Winddirecmean10m + Windspdmean80m +
Winddirectmean80m + Windspdmean900mb + Winddirectmean900mb + Windgustmean +
Tempmax + Tempmin + Humimax + Humimin + Meanpressuremax + Meanpressuremin +
Totalcloudmax + Totalcloudmin + Highcloudmax + Mediumcloudmax + Lowcloudmax
+ Lowcloudmin + Windspdmax10m + Windspdmin10m + Windspdmax80m + Windspdmin8
0m + Windspdmax900mb + Windspdmin900mb + Windgustmax + Windgustmin"
```

```
best.model1 <- glm(var.rss.formula,family=binomial, data=data2)
summary(best.model1)
```

```
##
## Call:
```

```
## glm(formula = var.rss.formula, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5712  -0.8298   0.2753   0.8398   2.9424
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.186e+01  1.230e+01   5.029 4.93e-07 ***
## Tempmean             1.655e-01  1.624e-01   1.019 0.308034
## Humimean             1.711e-02  3.210e-02   0.533 0.593876
## MeanPressuremean     5.256e-01  1.389e-01   3.784 0.000154 ***
## Totalprecipitation   2.515e-02  2.720e-02   0.925 0.355146
## Totalcloudmean       1.315e-02  1.185e-02   1.109 0.267442
## Highcloudmean       -2.233e-03  6.262e-03  -0.357 0.721424
## Mediumcloudmean      4.730e-03  6.426e-03   0.736 0.461699
## Lowcloudmean        -3.427e-03  8.029e-03  -0.427 0.669467
## Sunshine             4.337e-04  8.718e-04   0.497 0.618848
## Waveradia            4.585e-05  9.341e-05   0.491 0.623528
## Windspdmean10m      -8.886e-02  9.376e-02  -0.948 0.343268
## Winddirecmean10m     5.198e-03  5.681e-03   0.915 0.360232
## Windspdmean80m      -7.431e-02  6.815e-02  -1.090 0.275557
## Winddirectmean80m   -8.863e-03  5.869e-03  -1.510 0.130984
## Windspdmean900mb     2.076e-02  2.561e-02   0.811 0.417558
## Winddirectmean900mb  5.462e-03  1.444e-03   3.783 0.000155 ***
## Windgustmean         2.143e-02  3.634e-02   0.590 0.555408
## Tempmax             -9.598e-03  9.511e-02  -0.101 0.919616
## Tempmin             -1.199e-01  8.557e-02  -1.401 0.161226
## Humimax             -1.422e-03  2.040e-02  -0.070 0.944442
## Humimin             -1.106e-02  1.827e-02  -0.605 0.544855
## Meanpressuremax     -2.681e-01  7.391e-02  -3.628 0.000286 ***
## Meanpressuremin     -3.228e-01  7.607e-02  -4.244 2.20e-05 ***
## Totalcloudmax        3.460e-03  4.821e-03   0.718 0.473032
## Totalcloudmin        6.602e-03  5.978e-03   1.104 0.269452
## Highcloudmax         3.195e-03  2.833e-03   1.128 0.259521
## Mediumcloudmax       6.443e-03  3.125e-03   2.062 0.039214 *
## Lowcloudmax          2.500e-03  3.360e-03   0.744 0.456923
## Lowcloudmin         -2.868e-04  6.967e-03  -0.041 0.967170
## Windspdmax10m        6.221e-02  3.432e-02   1.812 0.069926 .
## Windspdmin10m        1.731e-01  6.363e-02   2.720 0.006532 **
## Windspdmax80m        9.165e-03  2.821e-02   0.325 0.745252
## Windspdmin80m       -5.994e-02  4.179e-02  -1.434 0.151504
## Windspdmax900mb     -1.279e-02  1.203e-02  -1.063 0.287645
## Windspdmin900mb     -6.592e-03  1.890e-02  -0.349 0.727225
## Windgustmax          1.355e-02  1.658e-02   0.818 0.413549
## Windgustmin          1.746e-02  2.726e-02   0.640 0.521865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1242.8  on 1142  degrees of freedom
## AIC: 1318.8
```

```
##
## Number of Fisher Scoring iterations: 4
```

Avec le critère RSS on a un modèle qui présente un glm avec plusieurs variables dont 6 seulement qui sont significatives. Les résultats font apparaitre plusieurs variables ont risque d'avoir un problème de sur-dispersion.

Choisir le meilleur modèle suivant le critère adjr2

```
var.adjr2 <- names(which(reg.summary$which[max.adjr2,]==TRUE))
var.adjr2.formula <- paste("pluie.demain", "~", paste(var.adjr2[-1], collap
se=" + "))
var.adjr2.formula
```

```
## [1] "pluie.demain ~ MeanPressuremean + Totalprecipitation + Totalcloudme
an + Mediumcloudmean + Waveradia + Windspdmean10m + Winddirectmean80m + Win
ddirectmean900mb + Windgustmean + Tempmax + Meanpressuremax + Meanpressurem
in + Totalcloudmin + Highcloudmax + Mediumcloudmax + Lowcloudmax + Windspdm
ax10m + Windspdmin10m + Windspdmin80m"
```

```
best.model2 <- glm(var.adjr2.formula,family=binomial, data=data2)
summary(best.model2)
```

```
##
## Call:
## glm(formula = var.adjr2.formula, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5048  -0.8264   0.2896   0.8504   2.9005
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.118e+01  1.191e+01   5.138 2.78e-07 ***
## MeanPressuremean     5.138e-01  1.334e-01   3.852 0.000117 ***
## Totalprecipitation   2.298e-02  2.356e-02   0.976 0.329237
## Totalcloudmean       8.000e-03  5.326e-03   1.502 0.133128
## Mediumcloudmean      3.504e-03  5.061e-03   0.692 0.488676
## Waveradia            1.310e-04  6.346e-05   2.065 0.038962 *
## Windspdmean10m      -1.778e-01  5.493e-02  -3.237 0.001208 **
## Winddirectmean80m   -3.805e-03  1.611e-03  -2.362 0.018178 *
## Winddirectmean900mb  4.833e-03  1.320e-03   3.662 0.000251 ***
## Windgustmean         4.201e-02  1.957e-02   2.146 0.031868 *
## Tempmax              3.299e-02  1.688e-02   1.955 0.050640 .
## Meanpressuremax     -2.590e-01  7.186e-02  -3.604 0.000313 ***
## Meanpressuremin     -3.185e-01  7.197e-02  -4.426 9.60e-06 ***
## Totalcloudmin        6.175e-03  4.043e-03   1.527 0.126673
## Highcloudmax         2.978e-03  2.248e-03   1.325 0.185221
## Mediumcloudmax       7.696e-03  2.641e-03   2.914 0.003566 **
## Lowcloudmax          3.624e-03  2.629e-03   1.378 0.168113
## Windspdmax10m        7.654e-02  2.491e-02   3.073 0.002119 **
## Windspdmin10m        1.926e-01  5.713e-02   3.372 0.000746 ***
## Windspdmin80m       -7.386e-02  3.666e-02  -2.015 0.043946 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1249.7  on 1160  degrees of freedom
## AIC: 1289.7
##
## Number of Fisher Scoring iterations: 4
```

Le modèle choisi selon le critère adjr2 comprend plusieurs variables significatives. De même la valeur de AIC est nettement meilleure (1289.7) des autres modèles.

Choisir le meilleur modèle suivant le critère Cp

```
var.cp <- names(which(reg.summary$which[min.cp,]==TRUE))
var.cp.formula <- paste("pluie.demain", "~", paste(var.cp[-1], collapse=" +
"))
var.cp.formula
```

```
## [1] "pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean + Wind
spdmean80m + Winddirectmean80m + Winddirectmean900mb + Tempmin + Meanpressu
remax + Meanpressuremin + Totalcloudmin + Highcloudmax + Mediumcloudmax + L
owcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax"
```

```
best.model3 <- glm(var.cp.formula,family=binomial, data=data2)
summary(best.model3)
```

```
##
## Call:
## glm(formula = var.cp.formula, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4811  -0.8140   0.2683   0.8580   2.8732
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         62.606221  11.852733   5.282 1.28e-07 ***
## Tempmean             0.159670   0.051476   3.102 0.001923 **
## MeanPressuremean     0.507009   0.133265   3.805 0.000142 ***
## Mediumcloudmean      0.007686   0.004221   1.821 0.068635 .
## Windspdmean80m      -0.108050   0.029949  -3.608 0.000309 ***
## Winddirectmean80m   -0.002899   0.001547  -1.875 0.060826 .
## Winddirectmean900mb  0.004599   0.001287   3.573 0.000353 ***
## Tempmin             -0.116496   0.055214  -2.110 0.034867 *
## Meanpressuremax     -0.259860   0.070539  -3.684 0.000230 ***
## Meanpressuremin     -0.312160   0.073240  -4.262 2.02e-05 ***
## Totalcloudmin        0.007285   0.003824   1.905 0.056758 .
## Highcloudmax         0.003027   0.002218   1.365 0.172251
## Mediumcloudmax       0.007611   0.002610   2.916 0.003545 **
## Lowcloudmax          0.005544   0.002342   2.367 0.017913 *
## Windspdmax10m        0.061028   0.022388   2.726 0.006412 **
## Windspdmin10m        0.114462   0.036023   3.177 0.001486 **
## Windgustmax          0.018079   0.010463   1.728 0.084016 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1253.2  on 1163  degrees of freedom
## AIC: 1287.2
##
## Number of Fisher Scoring iterations: 4
```

Le modèle choisi selon le critère adjr2 comprend plusieurs variables significatives. De même la valeur de AIC est nettement meilleure (1287.2) des autres modèles.

La performance du modèle issu d'une méthode d'apprentissage s'évalue par sa capacité de prévision. La mesure de cette performance est très importante puisque, d'une part, elle permet d'opérer une sélection de modèle dans une famille associée à la méthode d'apprentissage utilisée et, d'autre part, elle guide le choix de la méthode en comparant chacun des modèles optimisés à l'étape précédente. Enfin, elle fournit une mesure de la qualité ou encore de la confiance que l'on peut accorder à la prévision.

Partage de l'échantillon en un ensemble d'apprentissage et un ensemble test (par exemple en prenant 2/3:1/3).

```
set.seed(10)
train <- sample(1:nrow(data2),2*nrow(data2)/3)
test <- (-train)
test
```

Utiliser regsubsets() sur l'ensemble d'apprentissage à l'aide de la méthode exhaustive.

```
regfit.best <- regsubsets(pluie.demain~.-X-Year-Month-Day-Hour-Minute,data=
data2[train,],nvmax=10)
```

Calculer l'erreur de test pour le meilleur modèle de chaque taille.

```
test.mat <- model.matrix(pluie.demain~.-X-Year-Month-Day-Hour-Minute,data=d
ata2[test,])

# initialisation de l'erreur de prediction
val.errors <- rep(NA ,10)
for(i in 1:10){
  # extraction des estimateurs des coefs
  coefi <- coef(regfit.best,id=i)
  # calcul de la prediction
  pred5 <- test.mat[,names(coefi)]%*%coefi
  # calcul de l'erreur de prediction
  val.errors[i] <- mean((data2$pluie.demain[test]-pred5)^2)
}
val.errors

##  [1] 0.2220517 0.2077234 0.2034660 0.2000028 0.1937633 0.1969085 0.20079
75
##  [8] 0.1962813 0.1964854 0.1952214
```

Sur l'ensemble des données, effectuer une sélection de variables par la méthode exhaustive et sélectionner le meilleur modèle.

```
regfit.best <- regsubsets(pluie.demain~.-X-Year-Month-Day-Hour-Minute
                          ,data=data2,nvmax=10)
coef(regfit.best,which.min(val.errors))

##     (Intercept)  Totalcloudmean         Tempmax Meanpressuremin  Mediumc
loudmax
##    11.217987387     0.002253715     0.011545514    -0.011150040     0.00
2358918
##     Windgustmax
##     0.004321003

best.model <- glm(var.bic.formula,family=binomial, data=data2)
summary(best.model)

##
## Call:
## glm(formula = var.bic.formula, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2825  -0.8975   0.3940   0.8631   2.6940
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     59.236757  10.376489   5.709 1.14e-08 ***
## Totalcloudmean   0.012263   0.002987   4.105 4.04e-05 ***
## Tempmax          0.064181   0.010523   6.099 1.07e-09 ***
## Meanpressuremin -0.061488   0.010128  -6.071 1.27e-09 ***
## Mediumcloudmax   0.010884   0.002060   5.284 1.27e-07 ***
## Windgustmax      0.022965   0.005630   4.079 4.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1300.9  on 1174  degrees of freedom
## AIC: 1312.9
##
## Number of Fisher Scoring iterations: 4
```

Le meilleur modèle généré par cette méthode contient des informations limitées. Nous ne retenons pas pour le moment ce modèle.

-→ Suite aux différents méthodes mobilisées nous constatons que le modèle g4 est le plus significatif (sur la base de l'AIC et le BIC) De même, les variables sélectionnées représentent une bonne quantité d'informations. Pour confirmer ce choix nous mobilisons la méthode de validation croisée pour confirmer notre choix

En ce qui suit nous présentons la méthode :

On calcule une matrice de confusion et donc on mesure un taux d'erreur on évalue l'air sous la courbe ROC sur l'échantillon d'apprentissage et sur l'échantillon test.

2. Validation croisée avec le modèle g4

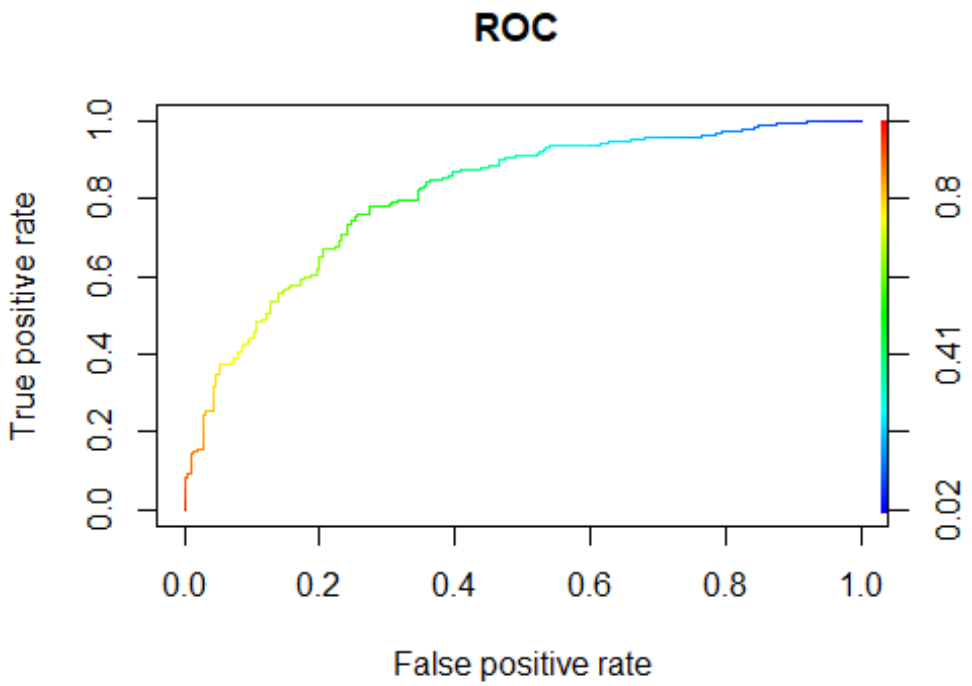
```
train = sample(c(T, F), nrow(data2), replace = T, prob = c(.6, .4))
# nous utilisons uniquement la base d'entrainement

#g4 step(g3)
g4 =
  glm( pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean +
    Lowcloudmean + Waveradia + Windspdmean80m + Winddirectmean80m +
    Windspdmean900mb + Winddirectmean900mb + Meanpressuremax +
    Meanpressuremin + Highcloudmax + Mediumcloudmax + Windspdmax10m,
    family = binomial, data = data2[train, ])
summary(g4)

##
## Call:
## glm(formula = pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudme
an +
##     Lowcloudmean + Waveradia + Windspdmean80m + Winddirectmean80m +
##     Windspdmean900mb + Winddirectmean900mb + Meanpressuremax +
##     Meanpressuremin + Highcloudmax + Mediumcloudmax + Windspdmax10m,
##     family = binomial, data = data2[train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4418  -0.8520  -0.2979   0.8865   2.7611
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          4.830e+01  1.463e+01   3.303 0.000958 ***
## Tempmean             2.769e-02  2.158e-02   1.283 0.199585
## MeanPressuremean     3.550e-01  1.685e-01   2.107 0.035146 *
## Mediumcloudmean      1.206e-02  5.445e-03   2.215 0.026756 *
## Lowcloudmean         1.043e-02  4.333e-03   2.407 0.016069 *
## Waveradia            1.656e-04  7.932e-05   2.088 0.036803 *
## Windspdmean80m      -9.123e-02  2.930e-02  -3.114 0.001848 **
## Winddirectmean80m   -5.352e-03  1.920e-03  -2.788 0.005304 **
```

```
## Windspdmean900mb     1.342e-02  1.001e-02   1.340 0.180228
## Winddirectmean900mb  5.343e-03  1.635e-03   3.269 0.001081 **
## Meanpressuremax     -1.646e-01  8.846e-02  -1.861 0.062780 .
## Meanpressuremin     -2.411e-01  9.184e-02  -2.625 0.008667 **
## Highcloudmax         2.122e-03  2.913e-03   0.729 0.466258
## Mediumcloudmax       7.985e-03  3.253e-03   2.455 0.014105 *
## Windspdmax10m        7.753e-02  2.465e-02   3.145 0.001659 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.80  on 716  degrees of freedom
## Residual deviance: 785.75  on 702  degrees of freedom
## AIC: 815.75
##
## Number of Fisher Scoring iterations: 4

#Nous éffectuons la prédiction uniquement sur la base de test
pred1 = predict(g4, data2[!train, ], type = "response")
# Nous évaluons l'erreur de prédiction
mean(abs(pred1 - data2[!train, "pluie.demain"]), na.rm = T)

## [1] 0.3566688

#Matrice de confusion
table(data2[!train, "pluie.demain"], pred1>.5)

##
##         FALSE TRUE
##   FALSE   147   68
##   TRUE     52  196

mean(data2[!train, "pluie.demain"] == (pred1>.5), na.rm=T)

## [1] 0.7408207

# change de seuil
table(data2[!train, "pluie.demain"], pred1>.7)

##
##         FALSE TRUE
##   FALSE   188   27
##   TRUE    123  125

mean(data2[!train, "pluie.demain"] == (pred1>.7), na.rm=T)

## [1] 0.6760259

#  Nous avons comparé deux seuils différents qui sont respectivement 0.5 et

 0.7. En utilisant le seuil de 0.7 nous constatons l'exitance de plus de

«Faux positifs » et de « vrais négatifs »
```

3. Vérifier la qualité de prédiction (modèle g4)

Nous allons en ce qui suit étudier la courbe ROC et mesurer l'AUC

```
library(ROCR)
library(ggplot2)

p = prediction(pred1, data2[!train, ]$pluie.demain)
Perf = performance(p, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC ")
```



Interprétation de la courbe ROC (modèle g4):

Les taux de vrais positifs augmentent rapidement plus que les faux positifs. Nous observons également que la courbe est au dessus de la diagonale.

```
table(data2[!train, "pluie.demain"], pred1>.5)

##
##          FALSE TRUE
##   FALSE    147   68
##   TRUE      52  196

performance(p, "auc")@y.values[[1]]

## [1] 0.8111965
```

L'air sous la courbe est de 0.81 ce qui signifie que le modèle est de bonne qualité. Plus AUC augmente plus le modèle présente une bonne qualité de prédiction

-→ Nous essayons maintenant de comparer les résultats de prédiction de deux modèles g4 et celui qui est sélectionné par la méthode de régression par recherche exhaustive.

Nous représentons ci-dessous la courbe de modèle sélectionné par la méthode de régression par recherche exhaustive.

```
table(data2$pluie.demain[test], pred5>.5)
```

```
##
##         FALSE TRUE
##   FALSE   128   62
##   TRUE     46  158
```

```
performance(p1, "auc")@y.values[[1]]
```

```
## [1] 0.7834881
```

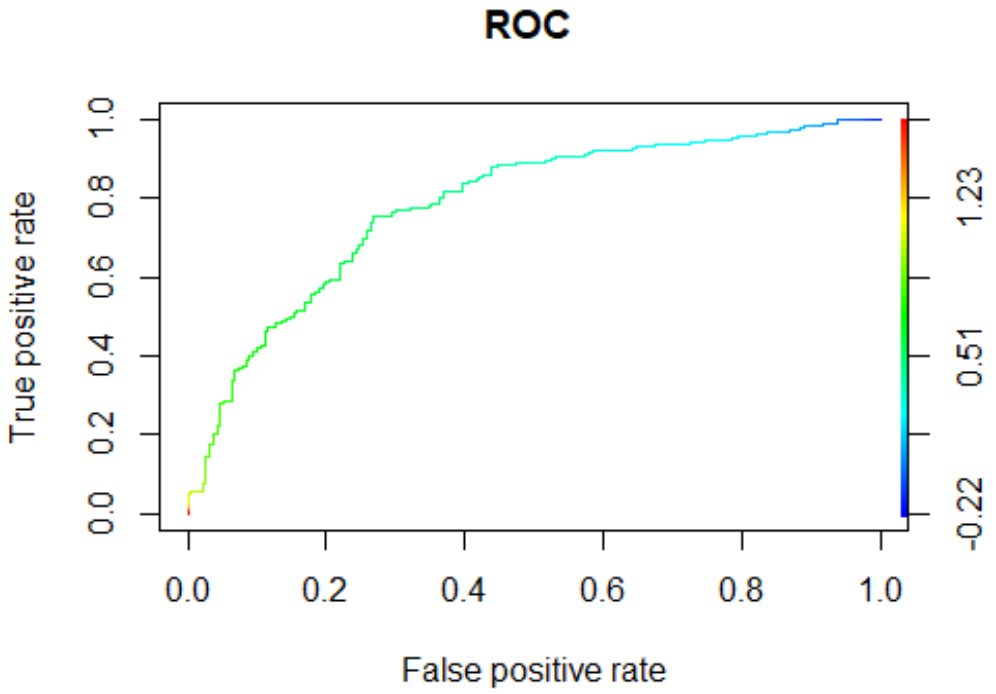
```
mean(data2$pluie.demain[test] == (pred5>.5), na.rm=T)
```

```
## [1] 0.7258883
```

```
mean(abs(pred5 - data2$pluie.demain[test]), na.rm = T)
```

```
## [1] 0.3866845
```

```
p = prediction(pred1, data2[!train, ]$pluie.demain)
Perf = performance(p, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC ")
```



Courbe ROC du modèle sélectionné par la méthode de recherche exhaustive

La courbe ROC n'augmente d'une façon significative par rapport à celle du modèle g4

4. Etude comparative entre les deux modèles (g4 et bestmodel3)

```
table(data2[!train, ]$pluie.demain, pred1>.5)

##
##         FALSE TRUE
##   FALSE   147   68
##   TRUE     52  196

performance(p, "auc")@y.values[[1]]

## [1] 0.8111965

mean(data2[!train, ]$pluie.demain == (pred1>.5), na.rm=T)

## [1] 0.7408207

mean(abs(pred1- data2[!train, ]$pluie.demain), na.rm = T)

## [1] 0.3566688

library(ROCR)

p1 = prediction(pred5, data2$pluie.demain[test])
Perf1 = performance(p1, "tpr", "fpr")
p = prediction(pred1, data2[!train, ]$pluie.demain)
Perf = performance(p, "tpr", "fpr")
library(ROCR)
data(ROCR.simple)
preds <- cbind(p = ROCR.simple$predictions,
               p1= abs(ROCR.simple$predictions +
               rnorm(length(ROCR.simple$predictions), 0, 0.1)))

pred.mat <- prediction(preds, labels = matrix(ROCR.simple$labels,
               nrow = length(ROCR.simple$labels), ncol = 2) )

perf.mat <- performance(pred.mat, "tpr", "fpr")
plot(perf.mat, colorize = TRUE)
```

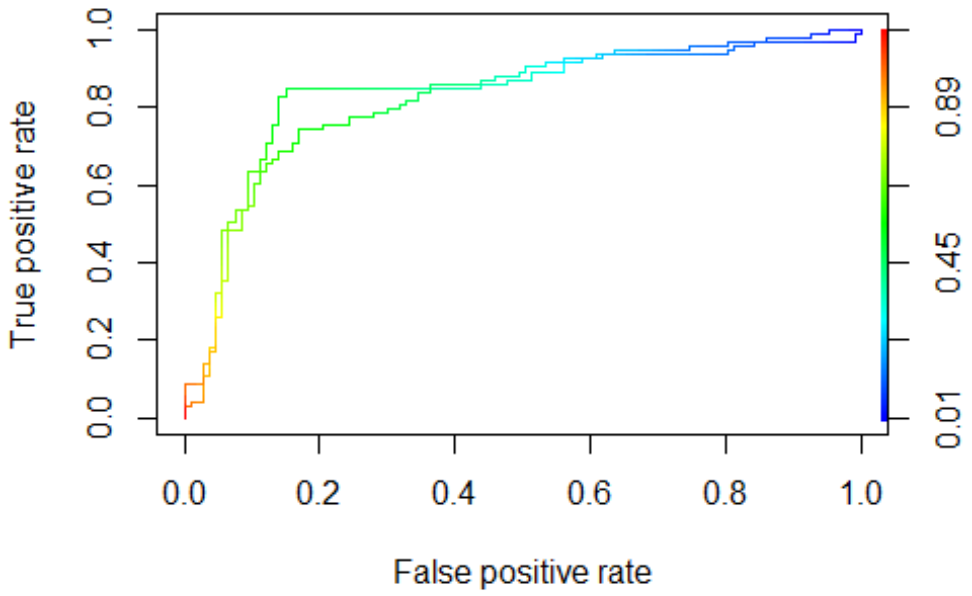


Tableau récapitulatif des valeurs qui présentent la qualité de prédiction de chaque modèle :

	Erreur de prédiction	Taux de bonne prédiction	Taux de vrai negatif	Taux de faux positif	AUC
Modele g4	0.35	0.74	68	52	0.81
Best model3	0.38	0.72	62	46	0.78

1-Le taux d'erreur de prédiction du modèle choisi g4 est plus petit que l'erreur de prédiction du modèle validé par la régression de recherche exhaustive (Best model3)

2-Le modèle g4 présente un AUC plus élevé q que Best model3 c'est un bon indicateur pour comparer les deux classifieurs

3- Le taux de bonne prédiction du modèle g4 est supérieur de taux de prédiction de modèle (Best model3)

D'après tous ces interprétations nous constatons que le modèle g4 donne une bonne qualité de prédiction par apport au modèle (Best model3)

5. Etude comparative entre les deux modèles (reg. Logistique et reg.probit)

Afin de s'assurer de la qualité de modèle choisi(g4),nous comparons les résultats à travers une régression Probit et une régression logistique :

```
#régression probit
g5= glm( pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean +
    Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
    Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
    Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,
    family = binomial(link="probit"),data=data2)

summary(g5)

##
## Call:
## glm(formula = pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudme
an +
##     Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
##     Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
##     Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,
##     family = binomial(link = "probit"), data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5253  -0.8618   0.2521   0.8711   3.0444
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         37.0015519  6.7981026   5.443 5.24e-08 ***
## Tempmean             0.0331172  0.0070277   4.712 2.45e-06 ***
## MeanPressuremean     0.2754342  0.0764586   3.602 0.000315 ***
## Mediumcloudmean      0.0054791  0.0023467   2.335 0.019554 *
## Windspdmean80m      -0.0649103  0.0172922  -3.754 0.000174 ***
## Winddirectmean80m   -0.0019819  0.0008639  -2.294 0.021781 *
## Winddirectmean900mb  0.0025329  0.0007430   3.409 0.000652 ***
## Meanpressuremax     -0.1391889  0.0409342  -3.400 0.000673 ***
## Meanpressuremin     -0.1743853  0.0415941  -4.193 2.76e-05 ***
## Totalcloudmax        0.0039129  0.0019597   1.997 0.045861 *
## Totalcloudmin        0.0034653  0.0021947   1.579 0.114348
## Mediumcloudmax       0.0042112  0.0015692   2.684 0.007283 **
## Windspdmax10m        0.0363833  0.0128739   2.826 0.004711 **
## Windspdmin10m        0.0618921  0.0207655   2.981 0.002878 **
## Windgustmax          0.0120676  0.0060510   1.994 0.046119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
```

```
## Residual deviance: 1261.9  on 1165  degrees of freedom
## AIC: 1291.9
##
## Number of Fisher Scoring iterations: 5
```

```
# Validation croisée en probit
# Nous  utilison uniquement la base d'entrainement
g6 =
 glm( pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean +
    Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
    Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
    Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,
    family = binomial(link="probit"), data = data2[train, ])
summary(g6)
```

```
##
## Call:
## glm(formula = pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudme
an +
##     Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
##     Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
##     Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,
##     family = binomial(link = "probit"), data = data2[train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5414  -0.8890  -0.2581   0.9014   2.8606
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         28.5020999  8.4191783   3.385 0.000711 ***
## Tempmean             0.0307286  0.0091078   3.374 0.000741 ***
## MeanPressuremean     0.1745038  0.0990687   1.761 0.078164 .
## Mediumcloudmean      0.0077786  0.0030479   2.552 0.010706 *
## Windspdmean80m      -0.0897952  0.0230703  -3.892 9.93e-05 ***
## Winddirectmean80m   -0.0025473  0.0010722  -2.376 0.017518 *
## Winddirectmean900mb  0.0027287  0.0009292   2.937 0.003318 **
## Meanpressuremax     -0.0838883  0.0525955  -1.595 0.110719
## Meanpressuremin     -0.1202834  0.0535368  -2.247 0.024657 *
## Totalcloudmax        0.0038737  0.0023946   1.618 0.105738
## Totalcloudmin        0.0007677  0.0028307   0.271 0.786223
## Mediumcloudmax       0.0035468  0.0019677   1.802 0.071468 .
## Windspdmax10m        0.0494915  0.0167190   2.960 0.003074 **
## Windspdmin10m        0.0719744  0.0294111   2.447 0.014398 *
## Windgustmax          0.0133923  0.0075442   1.775 0.075869 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.80  on 716  degrees of freedom
## Residual deviance: 783.95  on 702  degrees of freedom
## AIC: 813.95
```

```
##
## Number of Fisher Scoring iterations: 4

# Nous effectuons une prédiction, uniquement sur la base de test
pred4 = predict(g6, data2[!train, ], type = "response")
# Nous évaluons l'erreur de prédiction
mean(abs(pred4 - data2[!train, "pluie.demain"]), na.rm = T)

## [1] 0.3561464

# Matrice de confusion
table(data2[!train, "pluie.demain"], pred4>.5)

##
##         FALSE TRUE
##   FALSE   149   66
##   TRUE     48  200

mean(data2[!train, "pluie.demain"] == (pred4>.5), na.rm=T)

## [1] 0.7537797

# Validation croisée k-fold

k = 10
index = sample(1:k, nrow(data2), replace=T)
res.logistique = rep(NA, k)
res.probit = rep(NA, k)

for(i in 1:k){
  reg.logistique = glm( pluie.demain ~ Tempmean + MeanPressuremean + Medium
cloudmean +
    Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
    Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
    Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,
    family = binomial,
    data = data2[index != i, ]
  )

  reg.probit =
   glm( pluie.demain ~ Tempmean + MeanPressuremean + Mediumcloudmean +
    Windspdmean80m + Winddirectmean80m + Winddirectmean900mb +
    Meanpressuremax + Meanpressuremin + Totalcloudmax + Totalcloudmin +
    Mediumcloudmax + Windspdmax10m + Windspdmin10m + Windgustmax,

    family = binomial(link="probit"),
    data = data2[index != i, ]
  )

  pred.logistique = predict(reg.logistique, newdata=data2[index == i, ],
                            type="response")
  pred.probit = predict(reg.probit, newdata=data2[index == i, ],
                            type="response")

  res.logistique[i] = mean(data2[index==i, "Pluie.demain"] == (pred.logisti
```

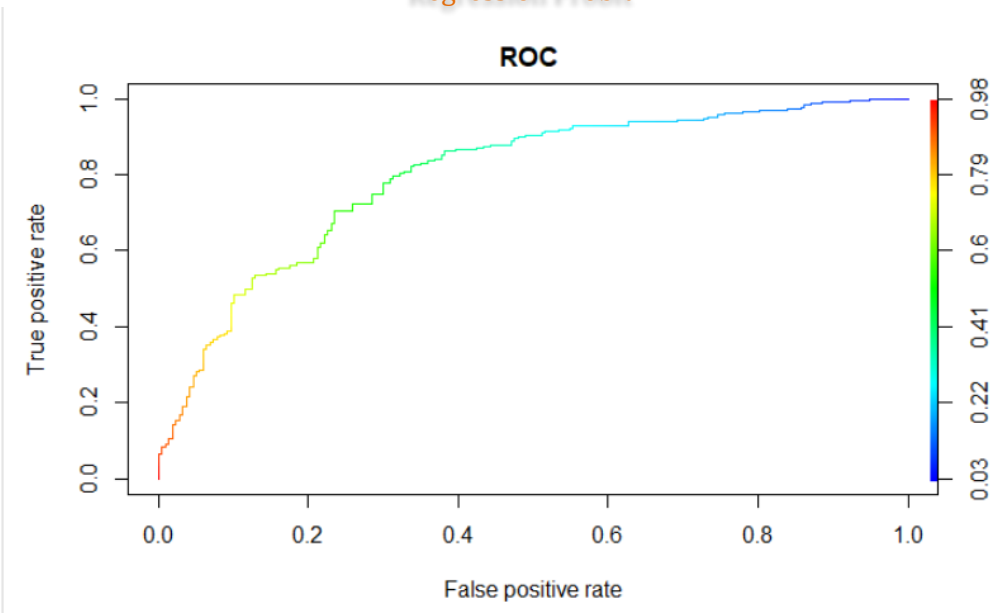
```
que >.5), na.rm = T)
  res.probit[i] = mean(data2[index==i, "Pluie.demain"] == (pred.probit >.5)
, na.rm = T)

}

mean(res.logistique)

mean(res.probit)


AIC(reg.probit)

## [1] 1151.267

AIC(reg.logistique)

## [1] 1148.672

# Nous avons un AIC de reg.Probit plus grande que AIC reg.logitique

Le modèle reg.logistique  et mieux que le modèle reg.probit suivant le

critère de AIC
```

Nous allons faire maintenant une étude comparative entre les deux types de regression

```{r}
Regression Probit
library(ROCR)
p2 = prediction(pred4, data2[!train, ]$pluie.demain)
Perf2 = performance(p2, "tpr", "fpr")
plot(Perf2, colorize = TRUE, main = "ROC ")
table(data2[!train, ]$pluie.demain, pred4>.5)
performance(p2, "auc")@y.values[[1]]
mean(data2[!train, ]$pluie.demain == (pred4>.5), na.rm=T)
mean(abs(pred4 - data2[!train, ]$pluie.demain), na.rm = T)

Regression Logistique
p = prediction(pred1, data2[!train, ]$pluie.demain)
Perf = performance(p, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC ")
table(data2[!train, ]$pluie.demain, pred1>.5)
performance(p, "auc")@y.values[[1]]
mean(data2[!train, ]$pluie.demain == (pred1>.5), na.rm=T)
mean(abs(pred1- data2[!train, ]$pluie.demain), na.rm = T)
```

```
library(ROCR)

p = prediction(pred1, data2[!train, ]$pluie.demain)
Perf = performance(p, "tpr", "fpr")
p2 = prediction(pred4, data2[!train, ]$pluie.demain)
Perf2 = performance(p2, "tpr", "fpr")
library(ROCR)
data(ROCR.simple)
preds <- cbind(p = ROCR.simple$predictions,
               p2= abs(ROCR.simple$predictions +
               rnorm(length(ROCR.simple$predictions), 0, 0.1)))

pred.mat <- prediction(preds, labels = matrix(ROCR.simple$labels,
               nrow = length(ROCR.simple$labels), ncol = 2) )

perf.mat <- performance(pred.mat, "tpr", "fpr")
plot(perf.mat, colorize = TRUE)

```

Regression Probit



Regression Logistique

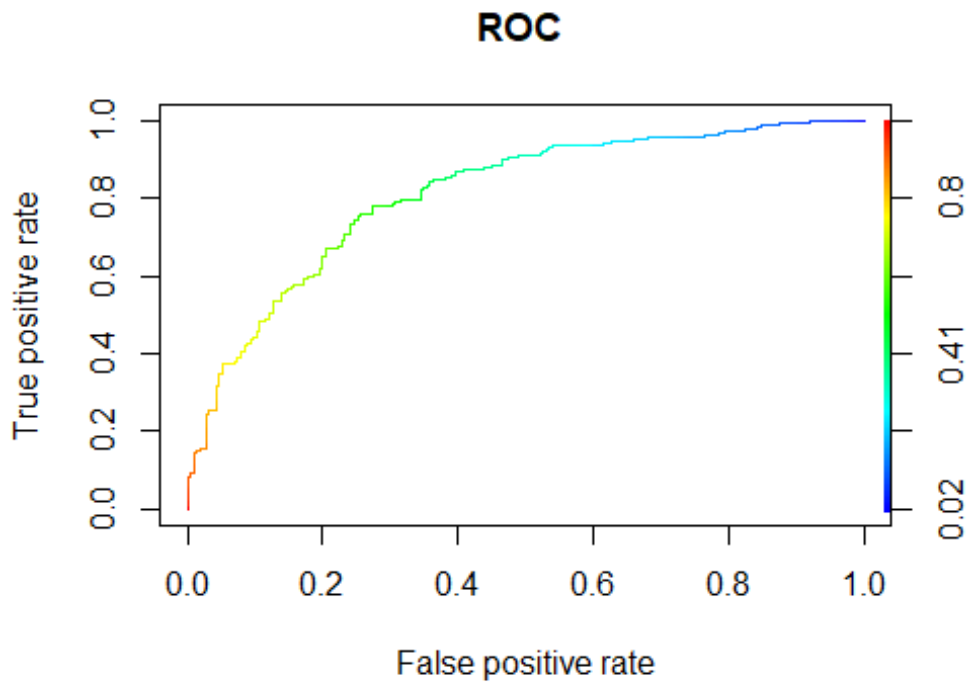


Tableau récapitulatif des valeurs qui présentent la qualité de prédiction de chaque modèle :

	Erreur de prédiction	Taux de bonne prédiction	Taux de vrai negative	Taux de faux positive	AUC
Reg.logistique (model g4)	0.35	0.74	68	52	0.81
Reg.probit	0.36	0.74	58	57	0.80

1-Le taux d'erreur de prédiction du modèle choisi g4 est plus petit que l'erreur de prédiction du modèle validé par la régression logistique

2-Le modèle g4 présente un AUC plus élevé que le modèle de la régression logistique c'est un bon indicateur pour comparer les deux classifieurs

3- Le taux de bonne prédiction du modèle g4 est supérieur de taux de prédiction de modèle de la régression logistique

--->Suites à ces différentes étapes d'analyse et d'étude comparatives, nous choisissons le modèle g4

Sur la base de modèle choisi, nous allons proposer une prédiction pour le fichier météo-test

```
pred= predict(g4, newdata = data3, type = "response")
pred2 = (pred >= 0.5)
pred2
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
13
## FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
TRUE
##    14    15    16    17    18    19    20    21    22    23    24    25
26
##  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
TRUE
##    27    28    29    30    31    32    33    34    35    36    37    38
39
## FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
TRUE
##    40    41    42    43    44    45    46    47    48    49    50    51
52
##  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE
TRUE
##    53    54    55    56    57    58    59    60    61    62    63    64
65
##  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
TRUE
##    66    67    68    69    70    71    72    73    74    75    76    77
78
##  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
TRUE
##    79    80    81    82    83    84    85    86    87    88    89    90
91
##  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
FALSE
##    92    93    94    95    96    97    98    99   100   101   102   103
104
##  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
TRUE
##   105   106   107   108   109   110   111   112   113   114   115   116
117
## FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE
TRUE
##   118   119   120   121   122   123   124   125   126   127   128   129
130
```

```
##  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
TRUE
##   131   132   133   134   135   136   137   138   139   140   141   142
143
##  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE
FALSE
##   144   145   146   147   148   149   150   151   152   153   154   155
156
##  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE
FALSE
##   157   158   159   160   161   162   163   164   165   166   167   168
169
## FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
TRUE
##   170   171   172   173   174   175   176   177   178   179   180   181
182
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
FALSE
##   183   184   185   186   187   188   189   190   191   192   193   194
195
##  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##   196   197   198   199   200   201   202   203   204   205   206   207
208
## FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
TRUE
##   209   210   211   212   213   214   215   216   217   218   219   220
221
##  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
TRUE
##   222   223   224   225   226   227   228   229   230   231   232   233
234
## FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
FALSE
##   235   236   237   238   239   240   241   242   243   244   245   246
247
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
FALSE
##   248   249   250   251   252   253   254   255   256   257   258   259
260
##  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE
TRUE
##   261   262   263   264   265   266   267   268   269   270   271   272
273
##  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE
FALSE
##   274   275   276   277   278   279   280   281   282   283   284   285
286
## FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
TRUE
##   287   288   289   290
## FALSE FALSE  TRUE  TRUE
```

```r
# enregistrer pred2
write.csv(pred2,file = "pred2.csv",row.names = FALSE)
summary(pred2)
```

```
##    Mode   FALSE    TRUE
## logical     133     157
```

```r
# ajouter colonne prediction
data4=cbind(data3,pred2)
# renomer la colonne
library(dplyr)
data4=rename(pluie.demain=pred2,data4)
```

Conclusion

Le modèle développé dans le cadre de ce travail apparait comme performant dans la prédiction s'il pleuvra le lendemain. Disposant d'un échantillon d'apprentissage et de test destiné à nous informer sur les facteurs qui peuvent déterminer cette prévision,

Nous avons pu réaliser plusieurs tests et analyses statistiques. Dans un premier temps nous avons réalisé une étude exploratoire qui vise à purifier nos données.

Nous avons vérifié les liens entre les variables à traves la matrice de corrélation qui nous a permis d'évaluer la dépendance entre plusieurs variables en même temps.

Dans un second temps nous avons utilisé les diagrammes de densité et les boxplots qui nous ont permis d'avoir la représentation graphique de données statistiques et visualiser la répartition des observations des variables quantitatives. Cette étape nous a aidé à identifier les valeurs extrêmes et de comprendre la répartition des observations. Par la suite nous avons mobilisé la méthode de discrétisation des variables afin d'identifier les variables aberrantes. Ce travail préliminaire nous a permis de supprimer les variables inutiles à notre analyse

Nous avons mené une étude comparative entre plusieurs modèles moyennant plusieurs critères tel ques l'anova, le p-value, l'AIC et le BIC . En se basant sur ces critères nous avons choisi le modèle g4 . Nous appliquons sur les 14 variables de g4 une analyse ACP afin de les synthétiser en quelles nouvelles variables appeler composantes principalement qui peuvent être visualiser graphiquement.

Après le choix de ce modèle nous avons également utilisé la méthode de recherche exhaustive qui consiste principalement à essayer toutes les solutions possibles et de générer le modèle le plus adéquat pour faire une prédiction.

Ensuite nous avons testé notre modèle g4 par la validation croisée pour s'assurer de sa qualité de prédiction. Les résultats font apparaitre une bonne qualité de modèle. En effet les indicateurs AUC et taux de prédiction sont satisfaisants.

Ces résultats nous ont également permis de comparer ce modèle g4 avec celui généré par la méthode recherche exhaustive. Nous concluons à travers cette comparaison que le

modèle g4 est meilleur pour le moment. Afin de s'assurer de ce choix nous mobilisons dans un dernier temps la régression Probit et comparer sa qualité de prédiction avec notre modèle g4. Les résultats nous confirment que ce modèle est le meilleur pour une prédiction. Enfin ce modèle a été utilisé pour faire la prédiction.