Tipologia i cicle de vida de les dades
Pràctica 2. Neteja i anàlisi de dades
Solució

Aitor Ferrus Blasco [aferrus]
Alonso López i Vicente [alopezvic]

05/01/2021

Contents

1. Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

Resposta

El dataset que hem escollit és *Rain in Australia* (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package).

Conté 10 anys de dades d'observacions diàries del clima en diferents llocs d'Australia. Conté una variable objectiu (RainTomorrow) per predir el temps del dia següent. Si la variable és Yes indica que ha plogut el dia següent 1mm o més. Amb aquesta variable podem entrenar models per tal de predir si plourà el dia següent.

Les variables que inclou el dataset són les següents:

Variable	Descripció
Date	La data de l'observació
Location	El nom de la localització de l'estació metereològica.
MinTemp	La temperatura mínima en graus Celsius
MaxTemp	La temperatura màxima en graus Celsius
Rainfall	La quantitat de pluja registrada durant el dia en mm
Evaporation	La denominada Class A pan evaporation (mm) durant 24 hores a les 9am
Sunshine	El nombre d'hores de sol durant el dia.
WindGustDir	La direcció de la ratxa de vent més forta en les 24 hores fins la mitjanit
WindGustSpeed	La velocitat (km/h) de la ratxa de vent més forta en les 24 hores fins a mitjanit
WindDir9am	Direcció del vent a les 9am
WindDir3pm	Direcció del vent a les 3pm
WindSpeed9am	Mitjana de la Velocitat del vent (km/hr) 10 minuts abans de les 9am
WindSpeed3pm	Mitjana de la Velocitat del vent (km/hr) 10 minuts abans de les 3pm
Humidity9am	Humitat (percentatge) a les 9am
Humidity3pm	Humitat (percentatge) a les 3pm
Pressure9am	Pressió atmosfèrica (hpa) reduïda al nivell mitjà del mar a les 9am
Pressure3pm	Pressió atmosfèrica (hpa) reduïda al nivell mitjà del mar a les 3pm
Cloud9am	Fracció del cel enfosquida pels núvols a les 9am. Es mesura en "oktas", els quals són una unitat de vuitens. Registre quants hi ha
Cloud3pm	Fracció del cel enfosquida pels núvols a les 3pm. Es mesura en "oktas", els quals són una unitat de vuits. Registre quants hi ha
Temp9am	Temperatura (graus Celsius) a les 9am
Temp3pm	Temperatura (graus Celsius) a les 3pm
RainToday	Booleà: 1 si la precipitació (mm) en les 24 hours anteriors a les 9am és superior a 1mm, sinó 0
RainTomorrow	La quantitat de pluja al dia següent en mm. Utilitzada per crear la variable resposta RainTomorrow. Un tipus de mesura del "risc".

2. Integració i selecció de les dades d'interès a analitzar.

Resposta

Hem seleccionat les dades de Melbourne, ja que tenen poques NA. Creiem que l'anàlisi que es pot realitzar en aquesta localització és pot adapatar ràpidament a qualsevol de les altres estacions que inclou el dataset.

```
library(readr)
weatherAUS <- read_csv("../data/weatherAUS.csv",
#weatherAUS <- read_csv("weatherAUS.csv",
    col_types = cols(Date = col_date(format = "%Y-%m-%d"),
```

```
        Evaporation = col_double(), Sunshine = col_double()))
weatherMelb <- weatherAUS[weatherAUS$Location == "Melbourne",]
summary(weatherMelb)
```

```
##       Date               Location            MinTemp         MaxTemp
##  Min.   :2008-07-01   Length:3193        Min.   : 1.40   Min.   : 9.70
##  1st Qu.:2010-09-07   Class :character   1st Qu.: 8.70   1st Qu.:16.10
##  Median :2013-01-13   Mode  :character   Median :11.40   Median :19.50
##  Mean   :2013-01-02                      Mean   :11.78   Mean   :20.77
##  3rd Qu.:2015-04-19                      3rd Qu.:14.60   3rd Qu.:24.20
##  Max.   :2017-06-25                      Max.   :28.60   Max.   :46.40
##                                          NA's   :480     NA's   :481
##     Rainfall      Evaporation       Sunshine      WindGustDir
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Length:3193
##  1st Qu.: 0.00   1st Qu.: 2.20   1st Qu.: 3.100   Class :character
##  Median : 0.00   Median : 4.00   Median : 6.500   Mode  :character
##  Mean   : 1.87   Mean   : 4.65   Mean   : 6.385
##  3rd Qu.: 1.20   3rd Qu.: 6.40   3rd Qu.: 9.600
##  Max.   :82.20   Max.   :23.80   Max.   :13.900
##  NA's   :758     NA's   :3       NA's   :1
##  WindGustSpeed     WindDir9am         WindDir3pm         WindSpeed9am
##  Min.   : 11.00   Length:3193        Length:3193        Min.   : 0.00
##  1st Qu.: 33.00   Class :character   Class :character   1st Qu.:11.00
##  Median : 43.00   Mode  :character   Mode  :character   Median :17.00
##  Mean   : 45.61                                        Mean   :19.13
##  3rd Qu.: 56.00                                        3rd Qu.:26.00
##  Max.   :122.00                                        Max.   :67.00
##  NA's   :14                                            NA's   :2
##   WindSpeed3pm   Humidity9am      Humidity3pm      Pressure9am
##  Min.   : 0.0   Min.   : 14.00   Min.   :  6.00   Min.   : 988.9
##  1st Qu.:15.0   1st Qu.: 58.00   1st Qu.: 41.00   1st Qu.:1012.6
##  Median :20.0   Median : 68.00   Median : 51.00   Median :1017.9
##  Mean   :22.1   Mean   : 67.55   Mean   : 51.18   Mean   :1017.6
##  3rd Qu.:28.0   3rd Qu.: 78.00   3rd Qu.: 61.00   3rd Qu.:1023.0
##  Max.   :76.0   Max.   :100.00   Max.   :100.00   Max.   :1039.0
##                 NA's   :482      NA's   :487      NA's   :480
##   Pressure3pm        Cloud9am        Cloud3pm        Temp9am
##  Min.   : 988.3   Min.   :0.000   Min.   :0.000   Min.   : 2.90
##  1st Qu.:1010.7   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:11.28
##  Median :1016.1   Median :7.000   Median :6.000   Median :14.10
##  Mean   :1015.8   Mean   :5.314   Mean   :5.336   Mean   :14.60
##  3rd Qu.:1021.1   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:17.40
##  Max.   :1035.8   Max.   :8.000   Max.   :8.000   Max.   :35.50
##  NA's   :483      NA's   :1034    NA's   :1106    NA's   :481
##     Temp3pm       RainToday         RainTomorrow
##  Min.   : 7.20   Length:3193        Length:3193
##  1st Qu.:14.90   Class :character   Class :character
##  Median :18.20   Mode  :character   Mode  :character
##  Mean   :19.26
##  3rd Qu.:22.50
##  Max.   :45.40
##  NA's   :484
```

3. Neteja de les dades

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Resposta

```
# Verifiquem si les dades no tenen valors nulls
sort(colMeans(is.na(weatherMelb)), decreasing = TRUE)
```

```
##      Cloud3pm      Cloud9am      Rainfall     RainToday  RainTomorrow
##  0.3463827122  0.3238333855  0.2373943000  0.2373943000  0.2373943000
##   Humidity3pm       Temp3pm   Pressure3pm   Humidity9am       MaxTemp
##  0.1525211400  0.1515815847  0.1512683996  0.1509552145  0.1506420294
##       Temp9am       MinTemp   Pressure9am    WindDir9am   WindGustDir
##  0.1506420294  0.1503288443  0.1503288443  0.0156592546  0.0043845913
## WindGustSpeed    WindDir3pm   Evaporation  WindSpeed9am      Sunshine
##  0.0043845913  0.0037582211  0.0009395553  0.0006263702  0.0003131851
##          Date      Location  WindSpeed3pm
##  0.0000000000  0.0000000000  0.0000000000
```

Les dades contenen elements buits en totes les columnes excepte Date i Location. Les columnes Cloud3pm , Cloud9am tenen mes de un 30% de valors nulls. Aixi que hem decidit que el nombre es molt gran i exclourem aquestes columnes del nostre dataset.

```
# Eliminem les  Columnes Cloud3pm i Cloud9am
weatherMelb <- subset( weatherMelb, select = -c(Cloud3pm, Cloud9am ) )
```

```
# Imputem valors, utilitzem package VIM i funció kNN.
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
weatherMelb_complet <- kNN(weatherMelb)
weatherMelb <- weatherMelb_complet[0:21]
```

Hem utilitzat kNN per a imputar els valors perduts aixi que les nostres dades no deurien de tenir cap valor null. Ho confirmem:

```
# Verifiquem que les dades no tenen valors nulls
sort(colMeans(is.na(weatherMelb)), decreasing = TRUE)
```

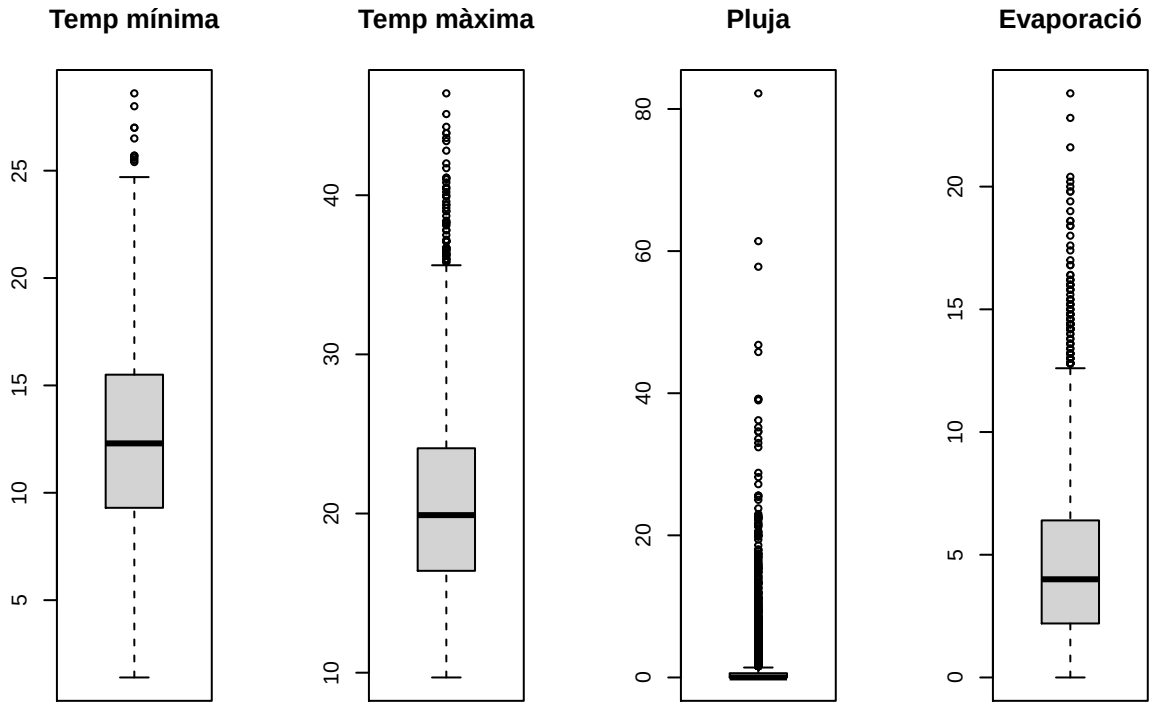
```
##          Date      Location       MinTemp       MaxTemp      Rainfall
##             0             0             0             0             0
##   Evaporation      Sunshine   WindGustDir WindGustSpeed    WindDir9am
##             0             0             0             0             0
##    WindDir3pm  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
##             0             0             0             0             0
##   Pressure9am   Pressure3pm       Temp9am       Temp3pm     RainToday
```

```
##             0             0             0             0             0
##  RainTomorrow
##             0
```

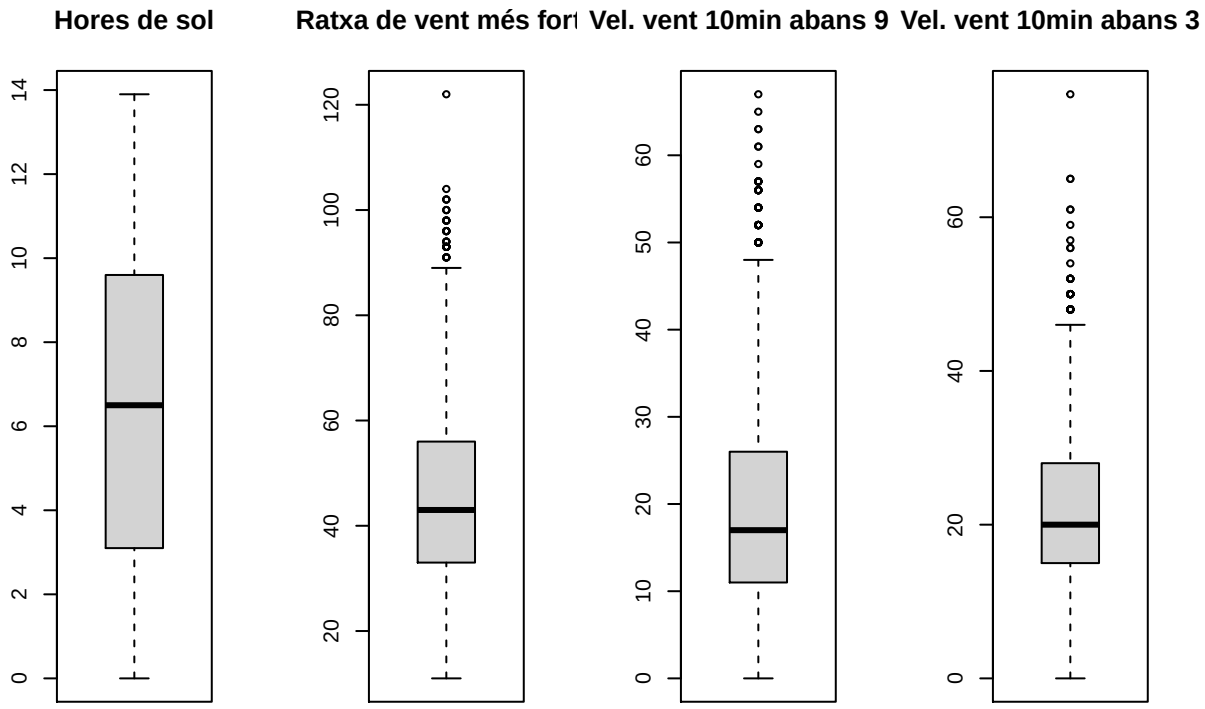
3.2. Identificació i tractament de valors extrems.

Resposta

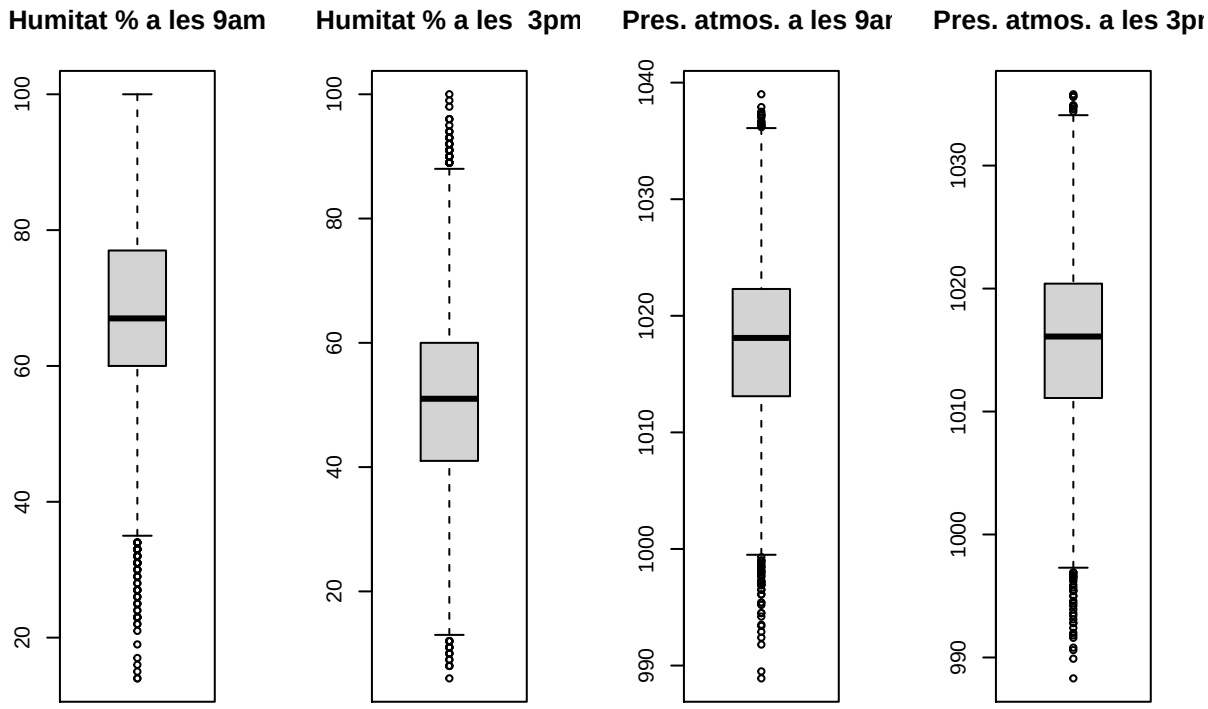
```
par(mfrow=c(1,4))
boxplot(weatherMelb$MinTemp, na.rm=TRUE, main="Temp mínima")
boxplot(weatherMelb$MaxTemp, na.rm=TRUE, main="Temp màxima")
boxplot(weatherMelb$Rainfall, na.rm=TRUE, main="Pluja")
boxplot(weatherMelb$Evaporation, na.rm=TRUE, main="Evaporació")
```



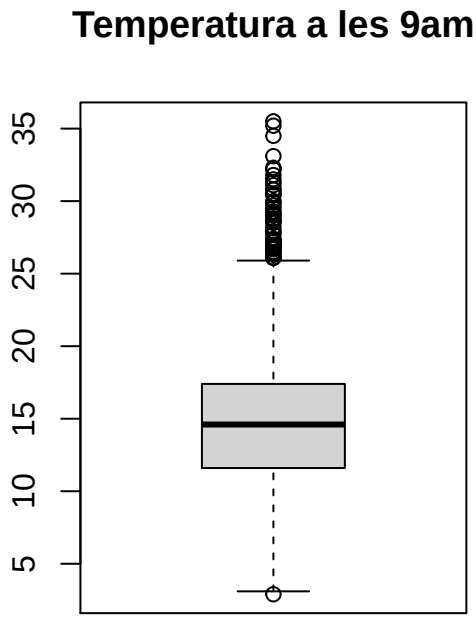

```
par(mfrow=c(1,4))
boxplot(weatherMelb$Sunshine, na.rm=TRUE, main="Hores de sol")
boxplot(weatherMelb$WindGustSpeed, na.rm=TRUE, main="Ratxa de vent més forta")
boxplot(weatherMelb$WindSpeed9am, na.rm=TRUE, main="Vel. vent 10min abans 9am")
boxplot(weatherMelb$WindSpeed3pm, na.rm=TRUE, main="Vel. vent 10min abans 3pm")
```

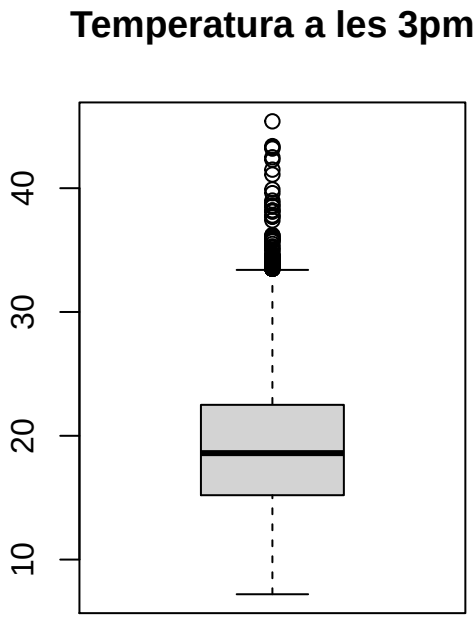



```
par(mfrow=c(1,4))
boxplot(weatherMelb$Humidity9am, na.rm=TRUE, main="Humitat % a les 9am")
boxplot(weatherMelb$Humidity3pm, na.rm=TRUE, main="Humitat % a les  3pm")
boxplot(weatherMelb$Pressure9am, na.rm=TRUE, main=" Pres. atmos. a les 9am")
boxplot(weatherMelb$Pressure3pm, na.rm=TRUE, main=" Pres. atmos. a les 3pm")
```




```
par(mfrow=c(1,2))
boxplot(weatherMelb$Temp9am, na.rm=TRUE, main="Temperatura a les 9am")
boxplot(weatherMelb$Temp3pm, na.rm=TRUE, main="Temperatura a les 3pm")
```





Correcció valors atípics de les columnes MinTemp, MaxTemp , Temp9am i Temp3pm: Les temperatures màximes d'Austràlia en Melbourne, rarament passen de 30 graus Celsius i les temperatures mínimes rarament passen de 20 graus Celsius. Per aquesta raó hem decidit corregir els valors atípics de les columnes MinTemp, MaxTemp , Temp9am i Temp3pm.

També corregim els valors atípics de la columna Evaporation. Al cap de l'any Australia té una mitja de 1200 mm així que si dividim entre 365 ens ix a 3.2.. Els nombres solen ser majors en estiu i primavera i menors en la tardor i l'hivern. Així que observant el boxplot les dades superiors al 12mm semblen ser dades errònies i per tant les hem de corregir.

Pel que fa a les variables WindGustSpeed, WindSpeed9am i WindSpeed3pm. És veritat que podem observar certs outliers però, no crec que siguin dades errònies. Australia és un país que sofreix de tornados cada any sobretot en les àrees amb gran població com Melbourne així que entenc que aquestes dades foren extretes durant eixos dies puntuals.

Pel que fa a les variables Humidity9am i Humidity3pm. És veritat que podem observar certs outliers, però, després d'investigar semblem dades que es poden donar Australia i en cap moment són dades errònies.

Pel que fa a les variables Pressure9am i Pressure3pm. Com anteriorment, no tinc evidències de què aquest outliers siguin dades errònies per tant crec que no faria falta tractar-les.

```
# Apliquem una simple funció per a substituir tots els valors superiors per NA
# MinTemp, MaxTemp , Temp9am i Temp3pm.
weatherMelb$MinTemp <- sapply(weatherMelb$MinTemp, function(x) ifelse(x>25, NA, x))
weatherMelb$MaxTemp <- sapply(weatherMelb$MaxTemp, function(x) ifelse(x>35, NA, x))
weatherMelb$Temp9am <- sapply(weatherMelb$Temp9am, function(x) ifelse(x>25, NA, x))
weatherMelb$Temp3pm <- sapply(weatherMelb$Temp3pm, function(x) ifelse(x>32, NA, x))
```

```
# Evaporation
weatherMelb$Evaporation <- sapply(weatherMelb$Evaporation, function(x) ifelse(x>12, NA, x))
```

```
# Verifiquem percentaje de valors nulls despres de tractar els outliers
sort(colMeans(is.na(weatherMelb)), decreasing = TRUE)
```

```
##   Evaporation       Temp3pm       Temp9am       MaxTemp       MinTemp
##   0.036329471   0.031318509   0.027247103   0.023802067   0.003131851
##          Date      Location      Rainfall      Sunshine   WindGustDir
##   0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
## WindGustSpeed    WindDir9am    WindDir3pm  WindSpeed9am  WindSpeed3pm
##   0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
##   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm     RainToday
##   0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
##  RainTomorrow
##   0.000000000
```

```
# Imputem valors, utilitzem package VIM i funció kNN.
library(VIM)
weatherMelb_complet <- kNN(weatherMelb)
weatherMelb <- weatherMelb_complet[0:21]
weatherMelb_complet <- weatherMelb_complet[0:21]
```

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades.

Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Resposta

Cal establir quines dades utilitzarem per predir la pluja. Cerquem si una (o més) variables ens serveixen per contruir un model que predigui la pluja al dia següent. L'objectiu és determinar si hi ha una relació entre les variables.

```
# Creació variable Month
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
weatherMelb <- weatherMelb %>% mutate(Month = Date)
weatherMelb$Month<- months(weatherMelb$Month)
```

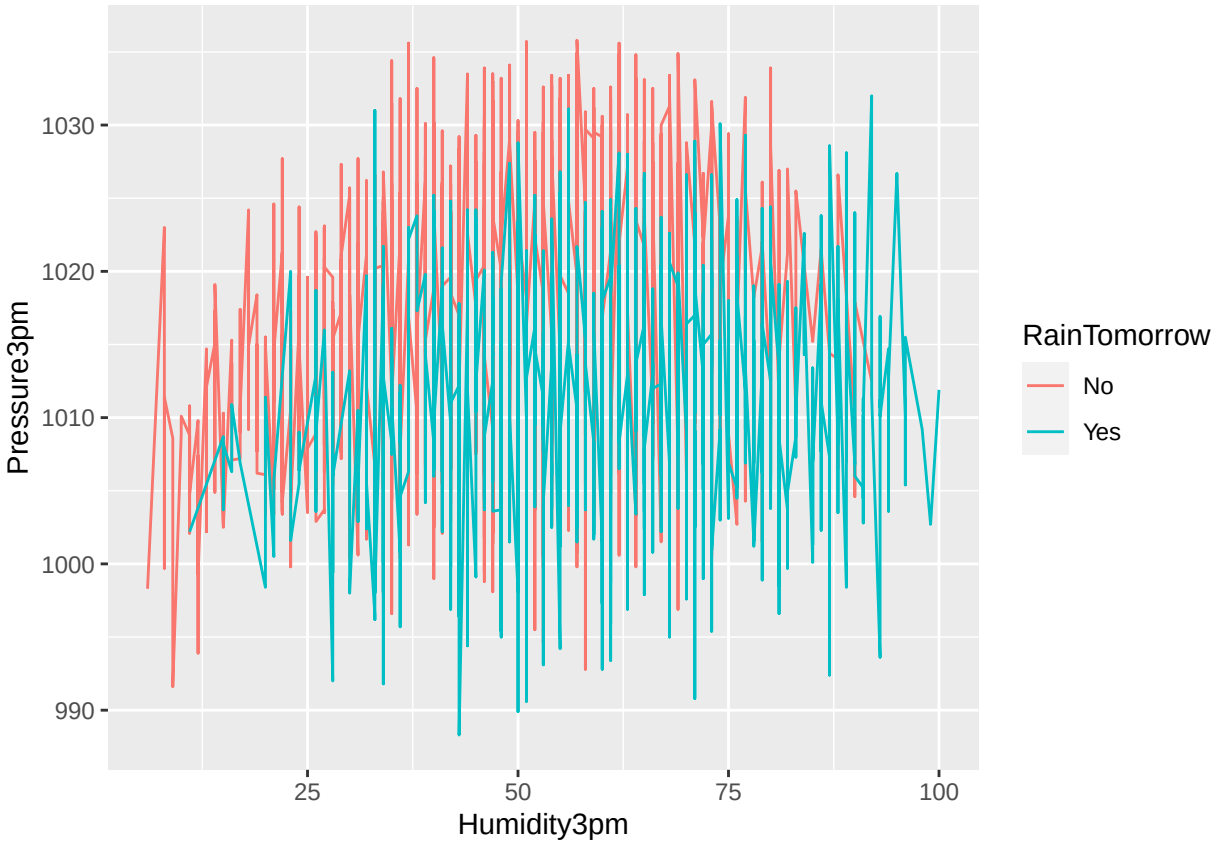
```
#  Creacio variable mesos
col <- c(6,7,12,13,14,15,16,17,21,22)
weatherMelb <- weatherMelb[col]
```

Hem seleccionat les següents variables:

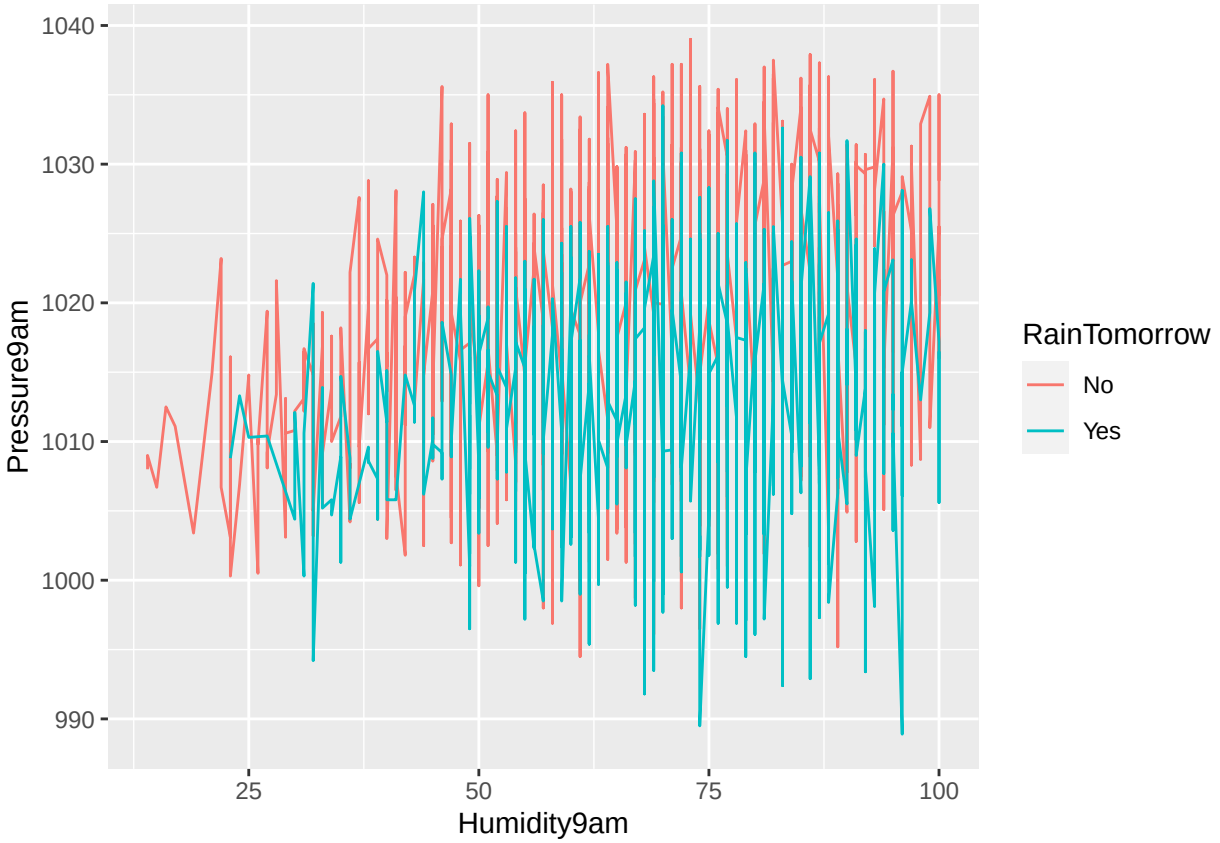
1. Humidity

2. Presure
3. WindSpeed
4. Evaporation
5. Sunshine
6. Month

```
# Podem veure com afecten Humidity + Presure amb la probabilitat de pluja
library(ggplot2)
ggplot(weatherMelb, mapping = aes(x = Humidity3pm , y = Pressure3pm, color = RainTomorrow) ) +
  geom_line()
```

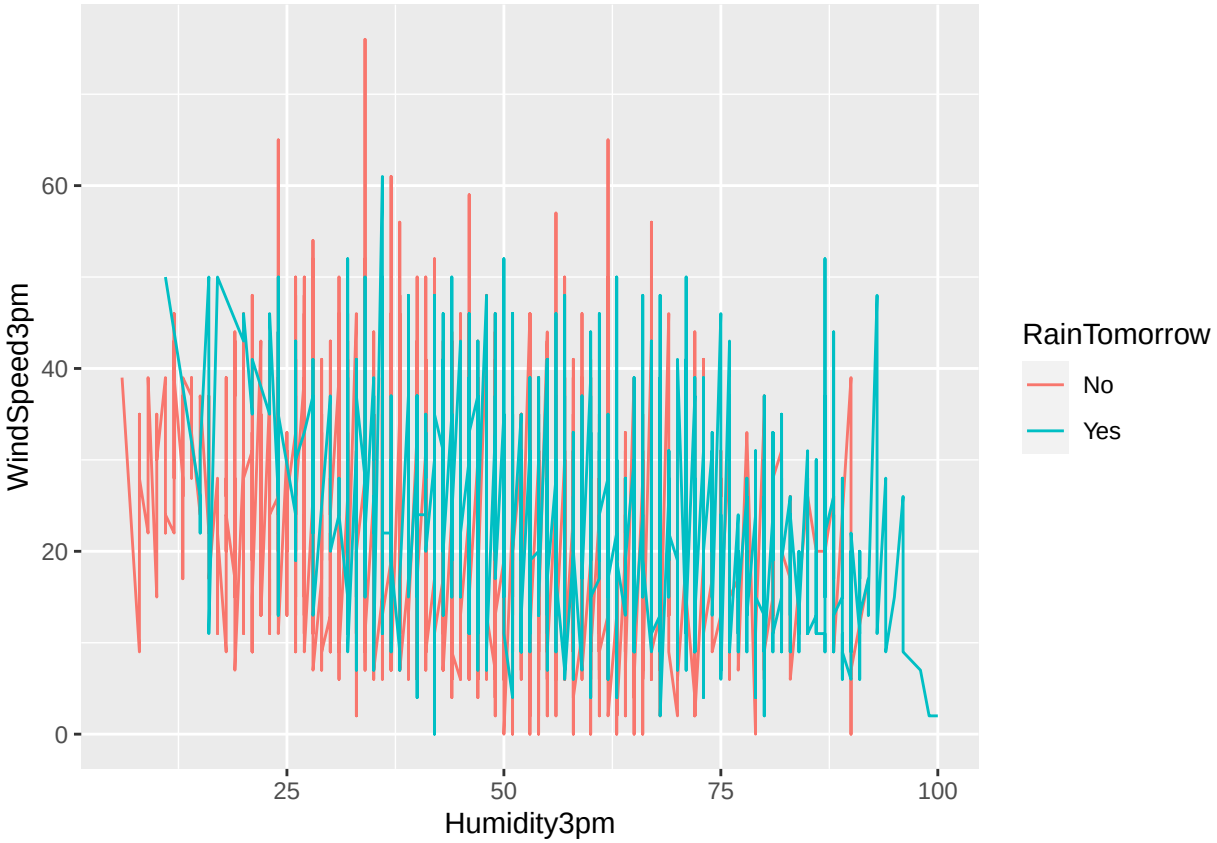



```
ggplot(weatherMelb, mapping = aes(x = Humidity9am , y = Pressure9am, color = RainTomorrow) ) +
  geom_line()
```

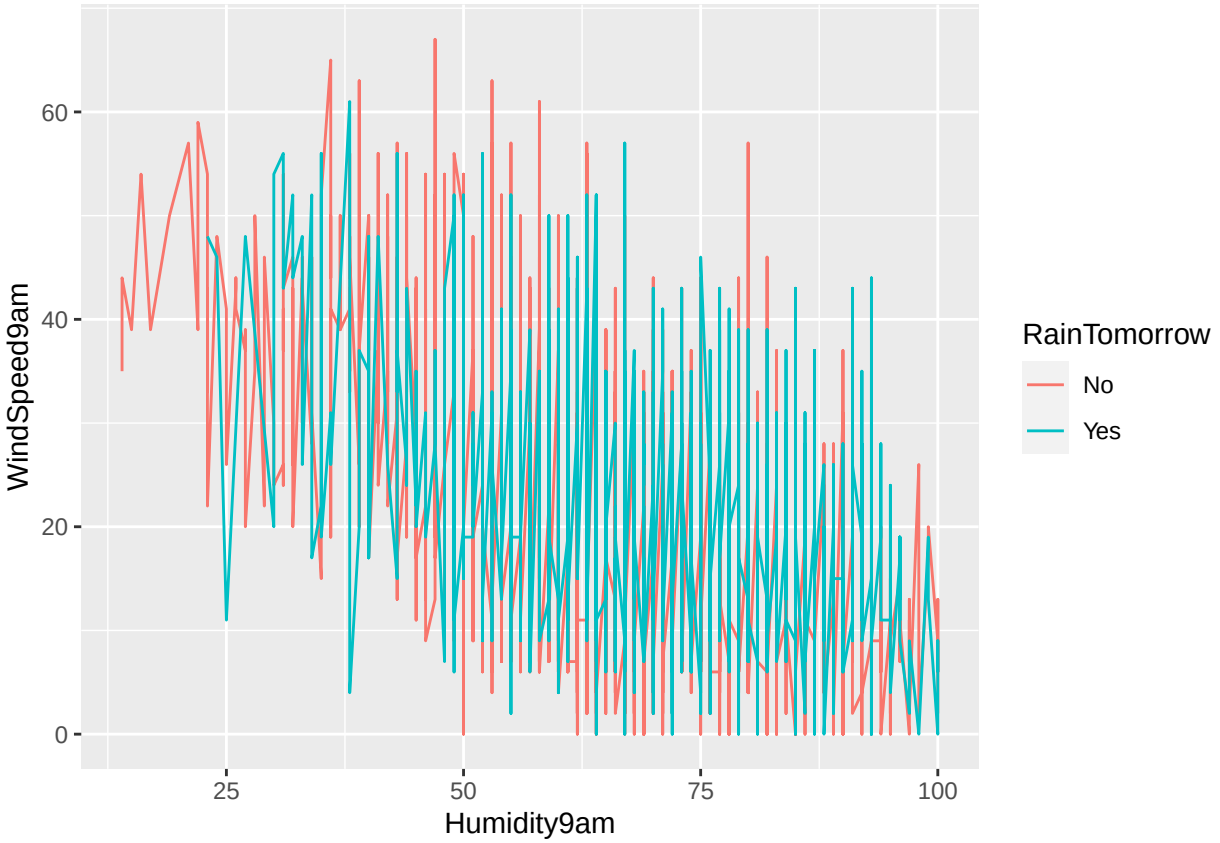


Si la pressió es baixa i la humitat és alta, és un fet clar que una massa d'aire humida s'acosta i pot estar associada a un front de pluges. Observant els gràfics podem assumir que el que hem dit abans és correcte.

```
# Podem veure com afecten Humidity + WindSpeed amb la probabilitat de pluja
ggplot(weatherMelb, mapping = aes(x = Humidity3pm , y = WindSpeed3pm, color = RainTomorrow) ) +
  geom_line()
```

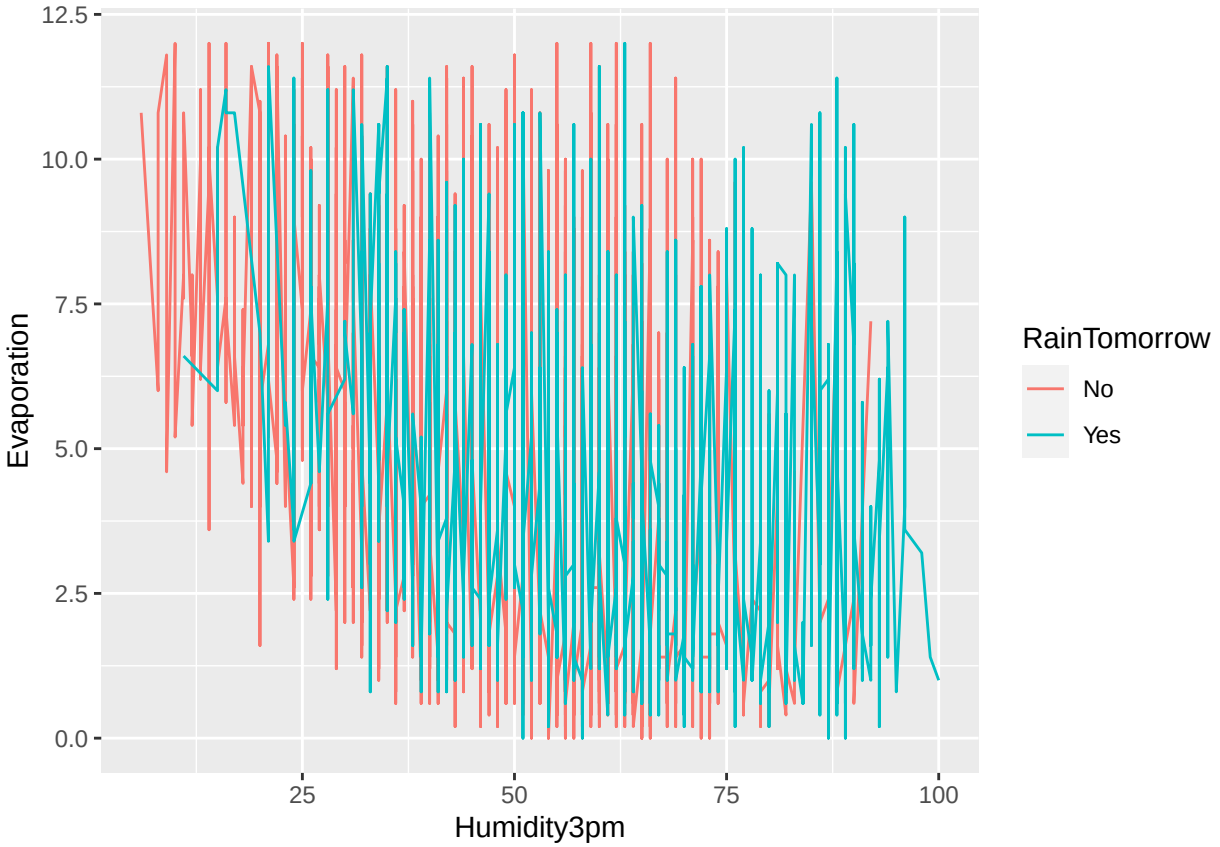



```
ggplot(weatherMelb, mapping = aes(x = Humidity9am , y = WindSpeed9am, color = RainTomorrow) ) +
  geom_line()
```

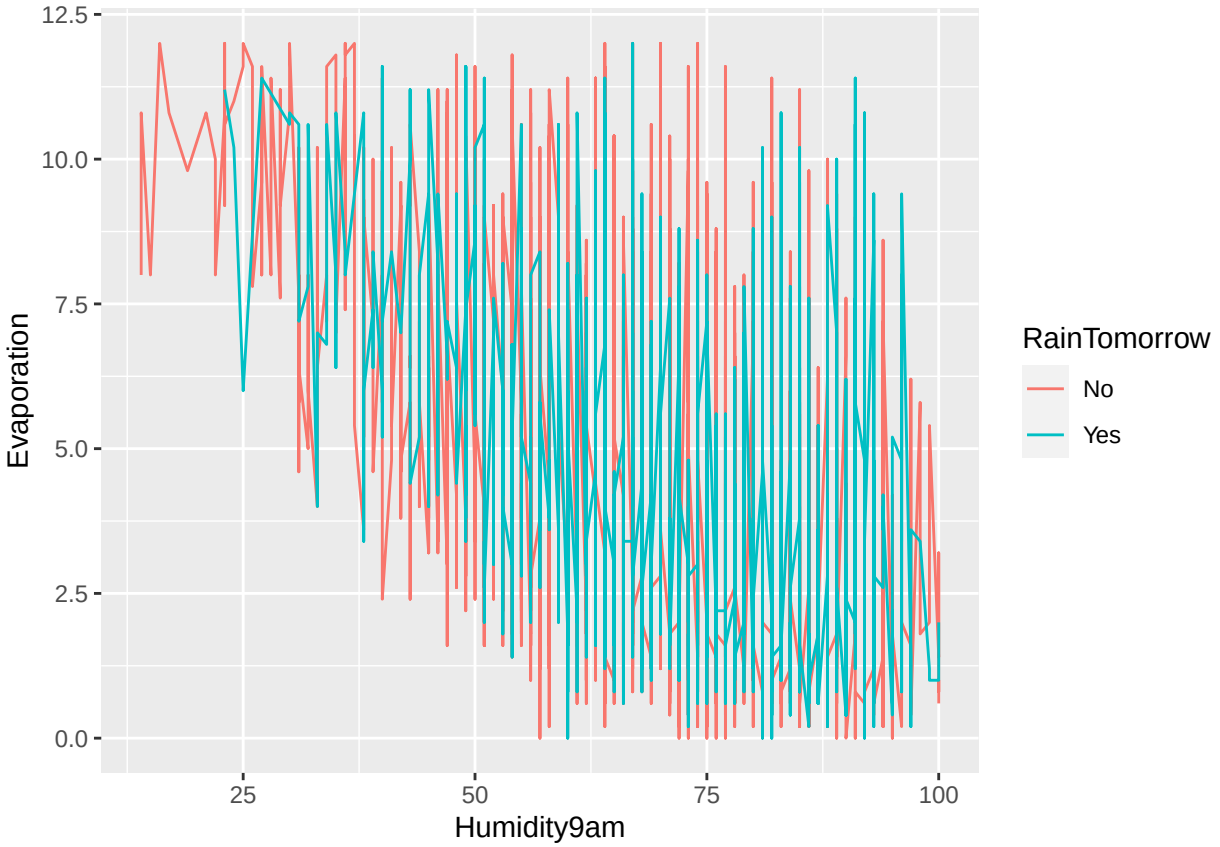


Un altre factor important és el vent, ja que l'aire fred no saturat absorbeix la humitat amb molta eficàcia. Als gràfics mostrats prèviament, podem veure una tendència en la qual velocitats de vent petites amb mesures d'humitats petites donen lloc a la no pluja a l'endemà mentre que com més incrementem aquestes dues variables la probabilitat de pluja sembla augmentar considerablement.

```
# Podem veure com afecten Humidity + Evaporation amb la probabilitat de pluja
ggplot(weatherMelb, mapping = aes(x = Humidity3pm , y = Evaporation, color = RainTomorrow) ) +
  geom_line()
```

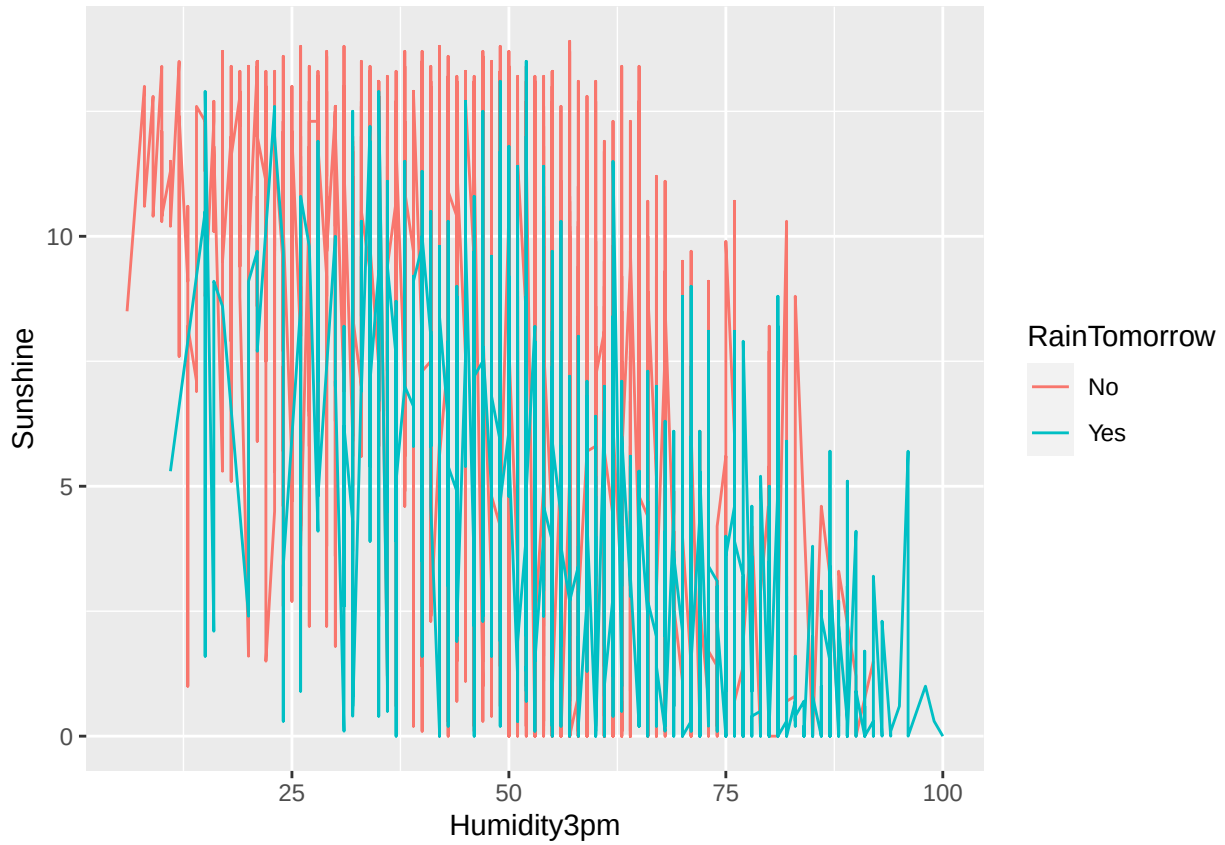



```
ggplot(weatherMelb, mapping = aes(x = Humidity9am , y = Evaporation, color = RainTomorrow) ) +
  geom_line()
```

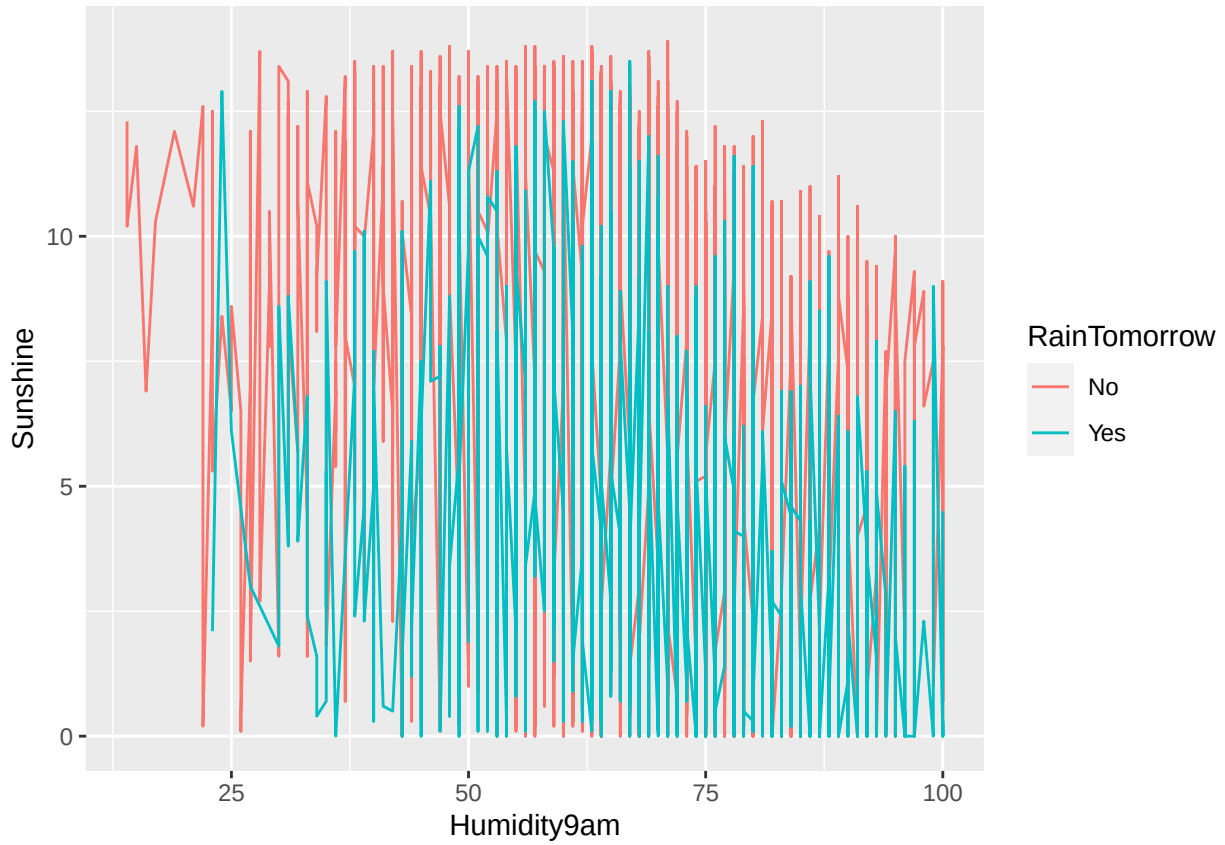


Quan la massa d'aire no està saturada (humitat relativa del 100%) la quantitat d'aigua evaporada es compensa amb una quantitat d'aigua igual condensada. En les nostres dades desafortunadament no tenim la humitat relativa i per tant no podem fer una comparació correcta. A més si observem la gràfica no hi ha cap relació entre la Evaporacio i humitat que provoqui l'augmente de les probabilitats de pluja a l'endemà tan sols podem observar que a més humitat registrada les probabilitats augmenten.

```
# Podem veure com afecten Humidity + Sunshine amb la probabilitat de pluja
ggplot(weatherMelb, mapping = aes(x = Humidity3pm , y = Sunshine, color = RainTomorrow) ) +
  geom_line()
```

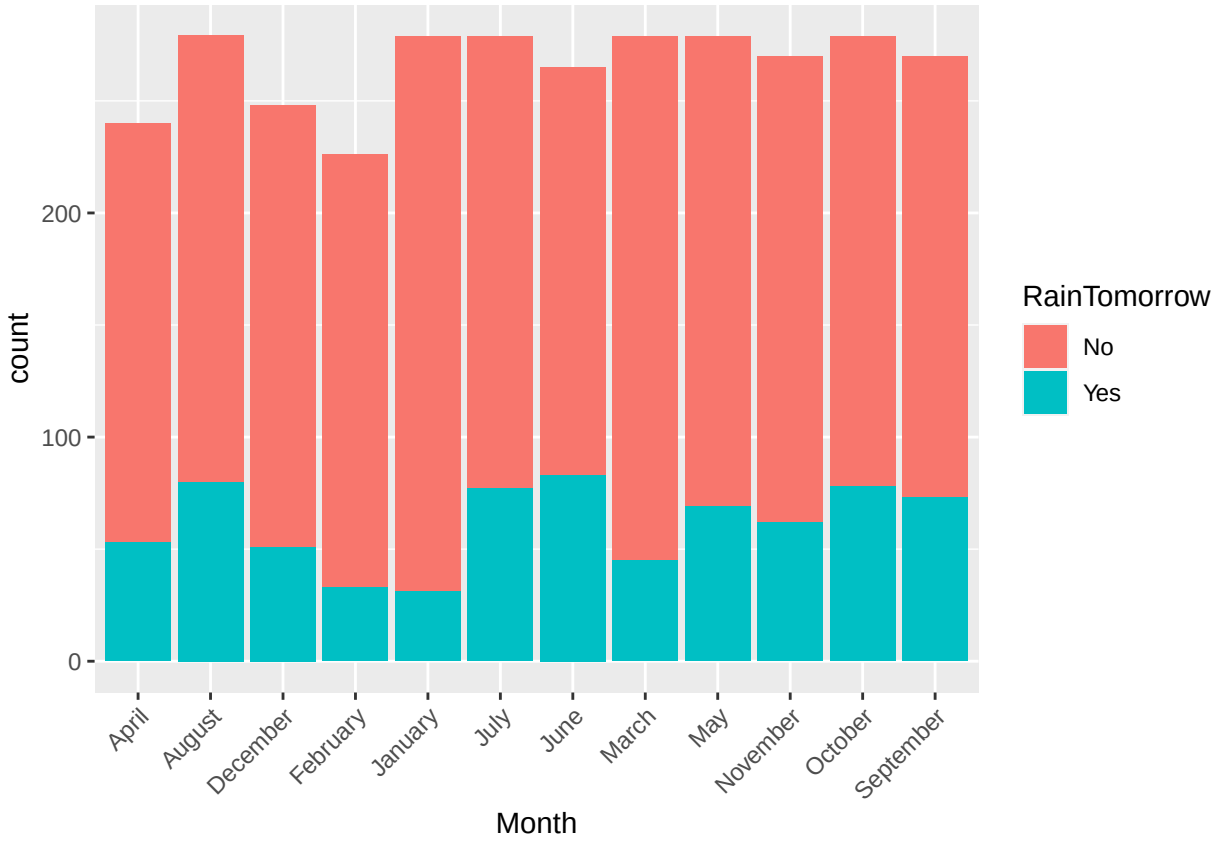



```
ggplot(weatherMelb, mapping = aes(x = Humidity9am , y = Sunshine, color = RainTomorrow) ) +
  geom_line()
```



En aquest grafic podem observar que a mennys hores de sol i mes humitat, les probabiliyats de plutja al sandema augmenten, mentre que a mes hores de sol i menys humitat es probabiliyats de plutja al sandema disminuixen.

```
p <- ggplot(data=weatherMelb,aes(x=Month,fill=RainTomorrow))+geom_bar()
p + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Amb el gràfic anterior podem concloure que en Melbourne no hi ha uns mesos específics de pluja, sembla que les precipitacions són similars al llarg de l'any durant els diferents mesos.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Resposta

Utilitzem el test de Shapiro-Wilk per comprovar la normalitat. Si el pvalor és inferior a 0.05, el nivell de significació, podrem rebutjar la hipòtesi nul · la i concloure que les dades no tenen una distribució normal. En cas contrari, si el pvalor és major que 0.05 podrem concloure que les dades segueixen una distribució normal.

```
# Utilitzem el test de Shapiro-Wilk per comprovar la normalitat de totes les variables cuantitatives
shapiro.test(weatherMelb_complet$MinTemp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$MinTemp
## W = 0.99356, p-value = 1.061e-10
```

```
shapiro.test(weatherMelb_complet$MaxTemp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$MaxTemp
## W = 0.95622, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$Rainfall)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Rainfall
## W = 0.36727, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$Evaporation)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Evaporation
## W = 0.95253, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$Sunshine)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Sunshine
## W = 0.95924, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$WindGustSpeed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$WindGustSpeed
## W = 0.96669, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$WindSpeed9am)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$WindSpeed9am
## W = 0.93527, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$WindSpeed3pm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$WindSpeed3pm
## W = 0.97805, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$Humidity9am)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Humidity9am
## W = 0.98611, p-value < 2.2e-16
```

```r
shapiro.test(weatherMelb_complet$Humidity3pm)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Humidity3pm
## W = 0.99306, p-value = 2.918e-11
```

```
shapiro.test(weatherMelb_complet$Pressure9am)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Pressure9am
## W = 0.99459, p-value = 1.786e-09
```

```
shapiro.test(weatherMelb_complet$Pressure3pm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Pressure3pm
## W = 0.99646, p-value = 7.544e-07
```

```
shapiro.test(weatherMelb_complet$Temp9am)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Temp9am
## W = 0.99431, p-value = 8.043e-10
```

```
shapiro.test(weatherMelb_complet$Temp3pm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weatherMelb_complet$Temp3pm
## W = 0.96951, p-value < 2.2e-16
```

Cap de les variables segueix una distribució normal. El pvalor calculat és inferior a 0.05, el nivell de significació, així que podem rebutjar la hipòtesi nul · la i concloure que les dades no tenen una distribució normal.

Per comprovar la homoscedasticitat, és a dir, la igualtat de variàncies, podem utilitzar el test de Levene si les dades segueixen una distribució normal, o el de Fligner-Killen si les dades no segueixen una distribució normal.

```
# Utilitzem Fligner-Killen perquè les dades no són normals.
fligner.test(MinTemp ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  MinTemp by RainTomorrow
## Fligner-Killeen:med chi-squared = 0.0091404, df = 1, p-value = 0.9238
```

```
fligner.test(MaxTemp ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  MaxTemp by RainTomorrow
```

```
## Fligner-Killeen:med chi-squared = 1.3223, df = 1, p-value = 0.2502
```

```
fligner.test(Rainfall ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Rainfall by RainTomorrow
## Fligner-Killeen:med chi-squared = 466.58, df = 1, p-value < 2.2e-16
```

```
fligner.test(Evaporation  ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Evaporation by RainTomorrow
## Fligner-Killeen:med chi-squared = 1.1947, df = 1, p-value = 0.2744
```

```
fligner.test(Sunshine ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Sunshine by RainTomorrow
## Fligner-Killeen:med chi-squared = 17.025, df = 1, p-value = 3.689e-05
```

```
fligner.test(WindGustSpeed ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  WindGustSpeed by RainTomorrow
## Fligner-Killeen:med chi-squared = 17.77, df = 1, p-value = 2.492e-05
```

```
fligner.test(WindSpeed9am ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  WindSpeed9am by RainTomorrow
## Fligner-Killeen:med chi-squared = 4.8131, df = 1, p-value = 0.02824
```

```
fligner.test(WindSpeed3pm ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  WindSpeed3pm by RainTomorrow
## Fligner-Killeen:med chi-squared = 11.404, df = 1, p-value = 0.0007329
```

```
fligner.test(Humidity9am ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Humidity9am by RainTomorrow
## Fligner-Killeen:med chi-squared = 25.604, df = 1, p-value = 4.192e-07
```

```r
fligner.test(Humidity3pm ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Humidity3pm by RainTomorrow
## Fligner-Killeen:med chi-squared = 124.51, df = 1, p-value < 2.2e-16
```

```r
fligner.test(Pressure9am ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Pressure9am by RainTomorrow
## Fligner-Killeen:med chi-squared = 18.768, df = 1, p-value = 1.476e-05
```

```r
fligner.test(Pressure3pm ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Pressure3pm by RainTomorrow
## Fligner-Killeen:med chi-squared = 13.633, df = 1, p-value = 0.0002223
```

```r
fligner.test(Temp9am ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Temp9am by RainTomorrow
## Fligner-Killeen:med chi-squared = 2.1401, df = 1, p-value = 0.1435
```

```r
fligner.test(Temp3pm ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Temp3pm by RainTomorrow
## Fligner-Killeen:med chi-squared = 0.040323, df = 1, p-value = 0.8409
```

Podem veure que les proves amb les variables MinTemp, MaxTemp, Evaporation, Temp9am i Temp3pm respecte de RainTomorrow resulten amb un p-valor superior al nivell de significació. Per tant, per aquestes variables, podem rebutjar la hipòtesi nul · la d'homoscedasticitat i concloure que aquestes variables presenten variàncies iguals pels grups de RainTomorrow. La resta de variables, en canvi, presenten variàncies estadísticament diferents per als grups de RainTomorrow.

4.3. Aplicació de proves estadístiques.

Per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Resposta

Contrast d'hipòtesi

Aplicarem una prova per contrast d'hipòtesi de tipus paràmetric, la t de Student, a les variables en que hem comprovat l'homoscedasticitat. En el cas de la distribució de la mitjana d'aquestes variables, segons el

teorema central del límit i donat que la mida de la nostra mostra és gran, es pot considerar que segueixen una distribució normal.

```r
t.test(MinTemp ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Welch Two Sample t-test
##
## data:  MinTemp by RainTomorrow
## t = 1.0418, df = 1186.7, p-value = 0.2977
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1608195  0.5249779
## sample estimates:
##  mean in group No mean in group Yes
##          12.32575          12.14367
```

```r
t.test(MaxTemp ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Welch Two Sample t-test
##
## data:  MaxTemp by RainTomorrow
## t = 9.2473, df = 1209.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.651231 2.540576
## sample estimates:
##  mean in group No mean in group Yes
##          21.24203          19.14612
```

```r
t.test(Evaporation ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Welch Two Sample t-test
##
## data:  Evaporation by RainTomorrow
## t = 2.6668, df = 1212.1, p-value = 0.007761
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07993605 0.52494298
## sample estimates:
##  mean in group No mean in group Yes
##          4.440399          4.137959
```

```r
t.test(Temp9am ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Welch Two Sample t-test
##
## data:  Temp9am by RainTomorrow
## t = 2.3033, df = 1169.9, p-value = 0.02144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06182317 0.77268694
## sample estimates:
##  mean in group No mean in group Yes
```

```
##          14.76147          14.34422
```

```
t.test(Temp3pm ~ RainTomorrow, data = weatherMelb_complet)
```

```
##
##  Welch Two Sample t-test
##
## data:  Temp3pm by RainTomorrow
## t = 11.188, df = 1181.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.952774 2.783343
## sample estimates:
##  mean in group No mean in group Yes
##          19.78629          17.41823
```

Podem veure que la variable MinTemp presenta un p-valor superior al nivell de signifcació, cosa que implica que no hi han diferències estadísticament significatves entre el grup NO i Yes de RainTomorrow. La resta de variables sí que presenten diferències estadísticament significatives.

Regressió logística

Al seguent apartat anem a realitzar un model de Regressio logística.

```
weatherMelb$RainTomorrow <- as.factor(weatherMelb$RainTomorrow)
levels(weatherMelb$RainTomorrow)[levels(weatherMelb$RainTomorrow)=="No"] <- "0"
levels(weatherMelb$RainTomorrow)[levels(weatherMelb$RainTomorrow)=="Yes"] <- "1"

# Model Regressio logística
model.logist=glm(formula=RainTomorrow ~ Evaporation+Sunshine+WindSpeed9am+
                   WindSpeed3pm+Humidity9am+Humidity3pm+Pressure9am+
                   Pressure3pm,family=binomial(link=logit),data=weatherMelb)
summary(model.logist)
```

```
##
## Call:
## glm(formula = RainTomorrow ~ Evaporation + Sunshine + WindSpeed9am +
##     WindSpeed3pm + Humidity9am + Humidity3pm + Pressure9am +
##     Pressure3pm, family = binomial(link = logit), data = weatherMelb)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0556  -0.6494  -0.4107  -0.2049   2.9436
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.130e+02  8.167e+00  13.839  < 2e-16 ***
## Evaporation  -5.119e-02  2.167e-02  -2.363 0.018147 *
## Sunshine     -1.395e-01  1.621e-02  -8.610  < 2e-16 ***
## WindSpeed9am  2.049e-02  5.609e-03   3.654 0.000259 ***
## WindSpeed3pm  9.396e-04  6.106e-03   0.154 0.877690
## Humidity9am   7.421e-04  4.669e-03   0.159 0.873714
## Humidity3pm   3.454e-02  4.330e-03   7.977 1.50e-15 ***
## Pressure9am  -2.292e-02  2.091e-02  -1.096 0.273066
## Pressure3pm  -9.096e-02  2.118e-02  -4.294 1.76e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3445.3  on 3192  degrees of freedom
## Residual deviance: 2690.9  on 3184  degrees of freedom
## AIC: 2708.9
##
## Number of Fisher Scoring iterations: 5
```

```
# Pred
prediction<-predict(model.logist, newdata=weatherMelb)
# taula de confusió
table(weatherMelb$RainTomorrow,prediction >= 0.5)
```

```
##
##     FALSE TRUE
##   0  2390   68
##   1   561  174
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(as.numeric(weatherMelb$RainTomorrow),fitted(model.logist))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  as.numeric(weatherMelb$RainTomorrow), fitted(model.logist)
## X-squared = 37341, df = 8, p-value < 2.2e-16
```

Segons Hosmer-Lemeshow Goodness of Fit (GOF) Test. El nostre model sembla que no encaixa bé perquè tenim una diferència significativa entre el model i les dades observades (és a dir, el valor p és inferior a 0,05)

Segons la taula de confusió amb valors TN = 2389, TP = 177, FN = 559, FP = 68.

Observant la taula de confusió el nostre model prediu tan sols és capaç de predir el 24% dels casos en el que l'endemà plourà.

"Specificity" o L'especificitat és la proporció de casos correctament classificats entre les respostes negatives. A la taula de confusió, l'especificitat és $68/(68 + 2389) = 2.7$ %. "Sensitivity" o La sensibilitat és la proporció dels classificats correctament entre els veritables participants que han donat una resposta afirmativa. A la taula de confusió, la sensibilitat és $177/(177 + 559) = 24$ %.

Inclús sense haver creat dues dades les de test i d'entrenament, sinó que hem utilitzat les mateixes per a crear el model i per a la predicció. El percentatge de vegades que prediu correctament és molt baix. Per tant, no podem predir si l'endemà plourà mitjançant aquest model.

Pel que fa a les variables que són rellevants. Observant els z-statitic p-values, podem observar que les variables Sunshine, WindSpeed9am, Humidity3pm i Pressure3pm tenen efecte amb el resultat de la variable Raintomorrow.

Els coeficients de les variables significatives suggereixen que:

Sunshine = -1.397e-01 . La variable Sunshine afecta negativament, és a dir quan aquest variable està present la probabilitat de pluja a l'endemà disminueix.

WindSpeed9am = 2.043e-02. La variable WindSpeed afecta positivament, és a dir quan aquest variable està present la probabilitat de pluja a l'endemà augmenta.

Humidity3pm = 3.449e-02. La variable WindSpeed afecta positivament, és a dir quan aquest variable està present la probabilitat de pluja a l'endemà augmenta.

Pressure3pm -9.146e-02. La variable WindSpeed afecta negativament, és a dir quan aquest variable està present la probabilitat de pluja a l'endemà disminueix.

Aquestes dades coincideixen amb el que em conclòs anteriorment mitjançant l'observació de les gràfiques.

Random Forest

A continuació executarem un mètode de classificació, el random forest. Posteriorment farem una predicció del resultat de lel model i ho validarem mitjançant una matriu de confusió.

```
library(rminer)
wmc<-weatherMelb_complet
wmc$RainTomorrow<-as.factor(wmc$RainTomorrow)
# Cal treure la variable Location ja que té un únic valor i genera un error en train.
wmc <- subset(wmc, select = - c(Location))
h<-holdout(wmc$RainTomorrow,ratio=2/3, mode="stratified")
data_train<-wmc[h$tr,]
data_test<-wmc[h$ts,]

# Farem una validació creuada amb 4 folds.
library(caret)
```

```
## Loading required package: lattice
```

```
train_control<-trainControl(method="cv",number=4)
mod<-train(RainTomorrow~.,data=data_train,method="rf",trControl=train_control)

pred<-predict(mod, newdata=data_test)
confusionMatrix(pred,data_test$RainTomorrow,positive="Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  774 136
##        Yes  45 109
##
##                Accuracy : 0.8299
##                  95% CI : (0.8059, 0.852)
##     No Information Rate : 0.7697
##     P-Value [Acc > NIR] : 8.843e-07
##
##                   Kappa : 0.4483
##
##  Mcnemar's Test P-Value : 2.237e-11
##
##             Sensitivity : 0.4449
##             Specificity : 0.9451
##          Pos Pred Value : 0.7078
##          Neg Pred Value : 0.8505
##              Prevalence : 0.2303
##          Detection Rate : 0.1024
##    Detection Prevalence : 0.1447
##       Balanced Accuracy : 0.6950
```

```
##
##        'Positive' Class : Yes
##
```

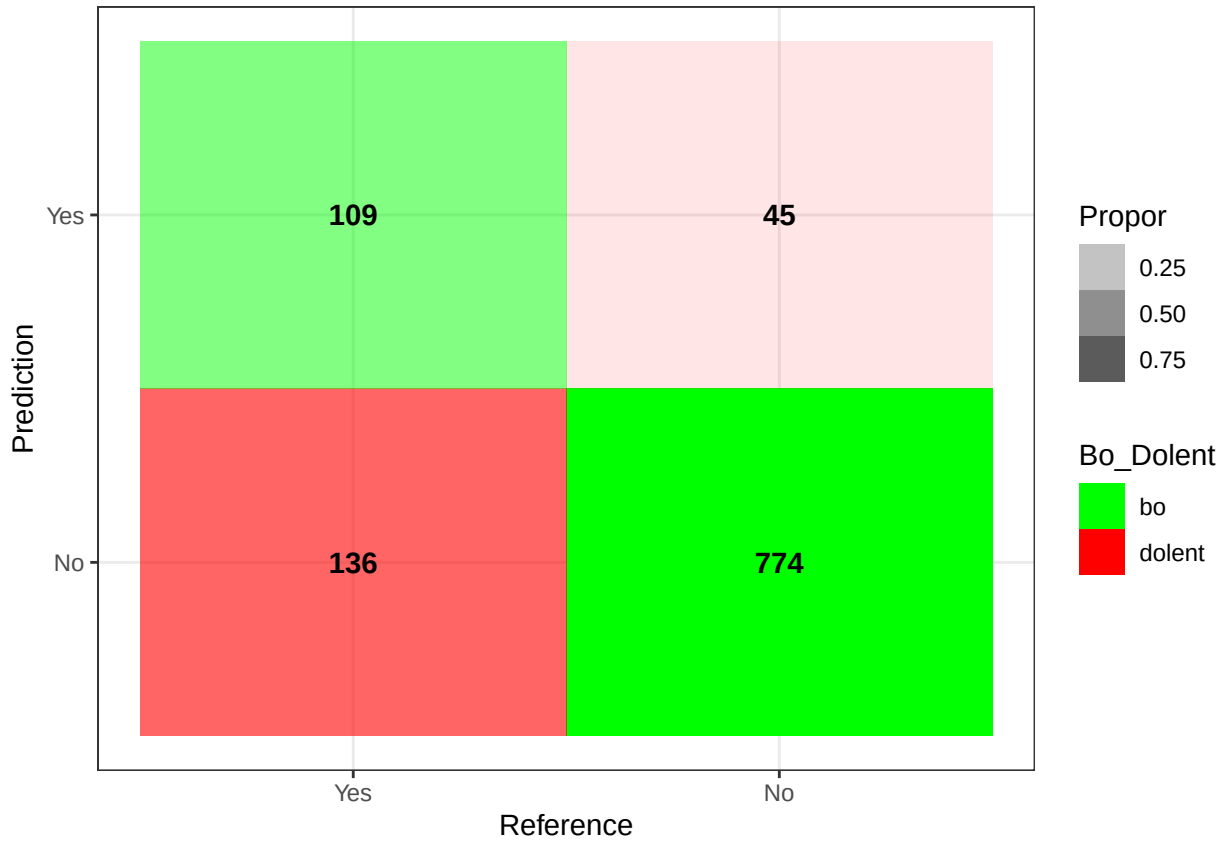
```r
# Visualització de la matriu de confusió

library(ggplot2)
library(dplyr)

table <- data.frame(confusionMatrix(pred,data_test$RainTomorrow,positive="Yes")$table)

plotTable <- table %>%
  mutate(Bo_Dolent = ifelse(table$Prediction == table$Reference, "bo", "dolent")) %>%
  group_by(Reference) %>%
  mutate(Propor = Freq/sum(Freq))

ggplot(data = plotTable, mapping = aes(x = Reference, y = Prediction, fill = Bo_Dolent, alpha = Propor)) +
  geom_tile() +
  geom_text(aes(label = Freq), vjust = .5, fontface  = "bold", alpha = 1) +
  scale_fill_manual(values = c(bo = "green", dolent = "red")) +
  theme_bw() +
  xlim(rev(levels(table$Reference)))
```



Podem veure que el nostre model no és gaire útil, ja que la sensitivitat és inferior a 0.5, cosa que indica la proporció de casos positius detectats respecte el total de casos positius. Per tant, no podem predir si al dia següent plourà mitjançant aquest model.

5. Representació dels resultats.

A partir de taules i gràfiques.

Resposta

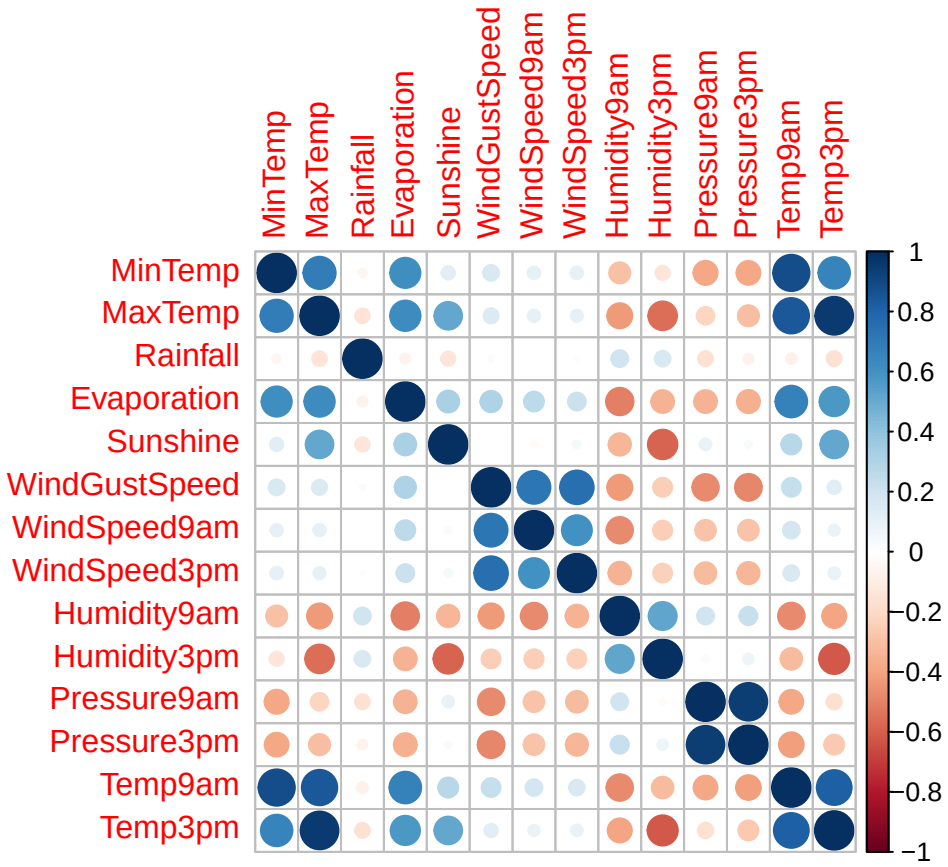
Les taules i les gràfiques estan incloses en cada apartat reforçant els resultats obtinguts i incloent conclusions extretes a partir d'aquestes. Evidentment, podríem afegir molts gràfics relacionats amb aquestes dades, com ara els gràfics de correlacions següents que mostren les correlacions entre cada parell de variables nimèriques del conjunt de dades o clusteritzar el conjunt de dades jeràrquicament.

```
library(corrplot)
```

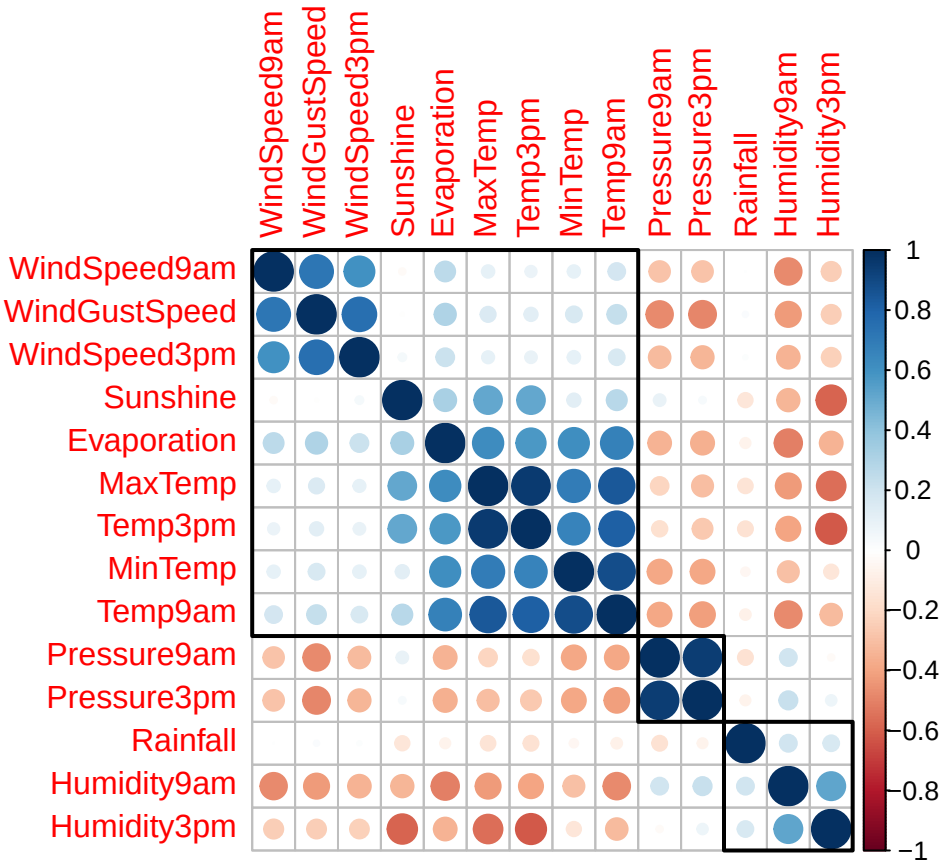
```
## corrplot 0.84 loaded
```

```
# Gràfic de correlacions
wmcorr<-weatherMelb_complet
#Treiem les variables no numèriques
wmcorr<-subset(wmcorr, select = - c(Date,Location,WindGustDir,WindDir9am,
                                    WindDir3pm,RainToday,RainTomorrow))

corr.wmcorr<-cor(wmcorr)
corrplot(corr.wmcorr,method="circle")
```




```
corrplot(corr.wmcorr, order = "hclust", addrect = 3)
```



6. Resolució del problema.

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Resposta

Dels resultats obtinguts podem treure diverses conclusions.

El model de Regressió Logística no és capaç de respondre al problema plantejat. El model creat és molt pobre i tan sols podem extreure quines variables com Sunshine, WindSpeed9am, Humidity3pm i Pressure3pm tenen efecte amb el resultat de la variable Raintomorrow. Desafortunadament, aquestes variables no són prou significatives per a crear un model que pugui predir la variable Raintomorrow amb suficient qualitat. Aquestes variables significatives foren prèviament observades durant el anàlisis gràfic de les relacions de les variables amb la variable Raintomorrow.

Cap dels models creats té la suficient qualitat per a respondre al problema. Inclòs utilitzar totes les variables del dataset no som capaços de crear un model que respongui satisfactòriament el problema.

Prèviament, s'han sotmès les dades a un preprocessament per a manejar els casos de zeros o elements buits i valors extrems (outliers). Per al cas del primer, s'ha fet ús d'un mètode d'imputació de valors (Knn) de tal forma que no hàgim d'eliminar registres del conjunt de dades inicial i que l'absència de valors no impliqui arribar a resultats poc precisos en les anàlisis. Per al cas del segon, s'ha optat per incloure els valors extrems de les variables Humidity9am, Humidity3pm, Pressure9am i Pressure3pm, ja que no hi havia cap indicació de què foren valors erronis i la resta de variables hem optat per convertir en NA els valors extrems per a més tard realitzar un altre mètode d'imputació de valors (Knn).

```
# El fitxer CSV amb les dades finals analitzades.
write.csv(weatherMelb,"../data/weatherMelb.csv", row.names = FALSE)
```

7. Codi.

Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

Resposta

El codi es troba incrustat en cada apartat de la pràctica.

8. Contribucions

Contribucions	Firma
Investigació prèvia	Aitor Ferrus Blasco, Alonso López i Vicente
Redacció de les respostes	Aitor Ferrus Blasco, Alonso López i Vicente
Desenvolupament codi	Aitor Ferrus Blasco, Alonso López i Vicente