Pràctica 2. Data Cleaning

Eduard Ruiz Sole

Gener 2020

Table of Contents

Detalls de la pràctica

Descripció

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Resolució

Descripció del dataset.

El conjunt de dades a analitzar s'ha obtingut a partir del següent enllaç kaggle, està format per 24 variables i un total de 142193 registres a analitzar. A continuació es detallaran les característiques de les variables :

- **Date :** Data de l'observació.
- **Location :** Ubicació de l'estació meteorològica que recull les dades.
- **MinTemp :** Temperatura mínima en graus Celsius.
- **MaxTemp :** Temperatura màxima en graus Celsius.
- **Rainfall :** Quantitat d'aigua registrada en mm (1L/m2).
- **Evaporation :** Quantitat d'evaporació en mm (1L/m2) en les últimes 24 hores fins les 9 del matí.
- **Sunshine :** Nombre d'hores de sol al dia.
- **WindGustDir :** La direcció de la ratxa de vent més forta en les últimes 24 hores fins a la mitjanit.
- **WindGustSpeed :** La velocitat (km/h) de la ràfega més forta de vent en les últimes 24 hores fins a la mitjanit.
- **WindDir9am :** Direcció del vent a les 9h00 del matí.
- **WindDir3pm :** Direcció del vent a les 15h00 del migdia.
- **WindSpeed9am :** Mitjana de la velocitat del vent (km/h) dels 10 minuts abans de les 9h00 del matí.
- **WindSpeed3pm :** Mitjana de la velocitat del vent (km/h) dels 10 minuts abans de les 15h00 del migdia.
- **Humidity9am :** Percentatge d'humitat a les 9h00 del matí.
- **Humidity3pm :** Percentatge d'humitat a les 15h00 del migdia.
- **Pressure9am :** Pressió atmosfèrica (hpa), reduida al nivell mitjà del mar, a les 9h00 del matí.

- **Pressure3pm :** Pressió atmosfèrica (hpa), reduida al nivell mitjà del mar, a les 15h00 del migdia.
- **Cloud9am :** Fracció del cel enfosquit pels núvols a les 9h00 del matí (mesura en "oktas"). Es registra quantes oktas del cel estan enfosquides pels núvols. Una mesura de 0 indica un cel completament clar mentre que un 8 indica que està completament ennuvolat.
- **Cloud3pm :** Fracció del cel enfosquit pels núvols a les 15h00 del migdia.
- **Temp9am :** Temperatura en graus Celsius a les 9h00 del matí.
- **Temp3pm :** Temperatura en graus Celsius a les 15h00 del migdia.
- **RainToday :** Boolean: 1, si la precipitació en les últimes 24 hores fins a les 9h00 del matí supera 1 mm ((1L/m2)), pel contrari, pren el valor 0.
- **RISK_MM :** Quantitat de pluja del dia següent en mm. S'utilitza per crear la variable "RainTomorrow".
- **RainTomorrow :** Representa la variable objectiu. Ha plogut demà?

Perquè és important i quina pregunta/problema pretén respondre?

Tal i com es pot comprovar en la definició de les variables, explicada en l'apartat anterior, l'objectiu precís d'aquesta pràctica és trobar un model que sigui capaç de determinar si el dia següent, al present, plourà o no a Australia. Els models de regressió permetran preedir si plou o no, en funció de les característiques i propietats del conjunt de variables i registres.

Un cop definit l'objectiu principal, hi han altres incògnites a resoldre, com per exemple, determinar quines són les variables que més influeixen en la hipòtesi de pluja. Aquests anàlisis poden ser de gran rellevància en qualsevol sector relacionat amb la meteorologia. Com per exemple les múltiples apps de previsió de temperatures i pluges que sovint s'equivoquen i ens fan agafar el paraigües quan no és necessàri o situacions contràries.

Integració i selecció de les dades d'interès a analitzar

La font de dades correspont a un fitxer CSV descarregat des del lloc web Kaggle. La funció read.csv() extraurà les dades i crearà un objecte de tipus data.frame :

```
# Càrrega i breu visualització del fitxer weatherAUS.csv
setwd("/Users/eduardruiz/Desktop/EDU/Data Science/M2.951 Tipologia i
cicle de vida de les dades/Neteja i anàlisi de dades/Rain in Australia/")
data <- read.csv("weatherAUS.csv", header = TRUE)
head(data)

##         Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine
WindGustDir
## 1 2008-12-01   Albury    13.4    22.9      0.6          NA       NA
W
## 2 2008-12-02   Albury     7.4    25.1      0.0          NA       NA
WNW
## 3 2008-12-03   Albury    12.9    25.7      0.0          NA       NA
WSW
## 4 2008-12-04   Albury     9.2    28.0      0.0          NA       NA
NE
## 5 2008-12-05   Albury    17.5    32.3      1.0          NA       NA
W
## 6 2008-12-06   Albury    14.6    29.7      0.2          NA       NA
WNW
##   WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm
Humidity9am
## 1            44          W        WNW           20           24
71
## 2            44        NNW        WSW            4           22
44
## 3            46          W        WSW           19           26
38
## 4            24         SE          E           11            9
45
## 5            41        ENE         NW            7           20
82
## 6            56          W          W           19           24
55
##   Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am
Temp3pm
## 1          22      1007.7      1007.1        8       NA    16.9
21.8
## 2          25      1010.6      1007.8       NA       NA    17.2
24.3
## 3          30      1007.6      1008.7       NA        2    21.0
23.2
## 4          16      1017.6      1012.8       NA       NA    18.1
```

```
26.5
## 5          33      1010.8      1006.0         7         8    17.8
29.7
## 6          23      1009.2      1005.4        NA        NA    20.6
28.9
##   RainToday RISK_MM RainTomorrow
## 1        No     0.0           No
## 2        No     0.0           No
## 3        No     0.0           No
## 4        No     1.0           No
## 5        No     0.2           No
## 6        No     0.0           No
```

Un cop carregat el conjunt de dades, el primer pas consisteix en determinar la tipologia de les variables que han de ser analitzades i decidir quines no representen cap interès en l'estudi, per poder procedir a eliminar-les. La funció summary ajuda a tenir una perspectiva més global del conjunt de dades i permet extreure les primeres conclusions.

```
# Visualització del tipus de variable
sapply(data, function(x) class(x))
```

```
##          Date      Location       MinTemp       MaxTemp      Rainfall
##      "factor"      "factor"     "numeric"     "numeric"     "numeric"
##   Evaporation      Sunshine   WindGustDir WindGustSpeed    WindDir9am
##     "numeric"     "numeric"      "factor"     "integer"      "factor"
##    WindDir3pm  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
##      "factor"     "integer"     "integer"     "integer"     "integer"
##   Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
##     "numeric"     "numeric"     "integer"     "integer"     "numeric"
##       Temp3pm     RainToday       RISK_MM  RainTomorrow
##     "numeric"      "factor"     "numeric"      "factor"
```

```
summary(data)
```

```
##          Date            Location        MinTemp         MaxTemp
##  2013-03-02:    49   Canberra:  3418   Min.   :-8.50   Min.   :-4.80
##  2013-03-03:    49   Sydney  :  3337   1st Qu.: 7.60   1st Qu.:17.90
##  2013-03-04:    49   Perth   :  3193   Median :12.00   Median :22.60
##  2013-03-06:    49   Darwin  :  3192   Mean   :12.19   Mean   :23.23
##  2013-03-07:    49   Hobart  :  3188   3rd Qu.:16.80   3rd Qu.:28.20
##  2013-03-10:    49   Brisbane:  3161   Max.   :33.90   Max.   :48.10
##  (Other)   :141899   (Other) :122704   NA's   :637     NA's   :322
##     Rainfall       Evaporation        Sunshine      WindGustDir
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.00   W      : 9780
##  1st Qu.:  0.00   1st Qu.:  2.60   1st Qu.: 4.90   SE     : 9309
##  Median :  0.00   Median :  4.80   Median : 8.50   E      : 9071
##  Mean   :  2.35   Mean   :  5.47   Mean   : 7.62   N      : 9033
##  3rd Qu.:  0.80   3rd Qu.:  7.40   3rd Qu.:10.60   SSE    : 8993
##  Max.   :371.00   Max.   :145.00   Max.   :14.50   (Other):86677
##  NA's   :1406     NA's   :60843    NA's   :67816   NA's   : 9330
```

```
##  WindGustSpeed      WindDir9am      WindDir3pm     WindSpeed9am
##  Min.   :  6.00   N      :11393   SE     :10663   Min.   :  0
##  1st Qu.: 31.00   SE     : 9162   W      : 9911   1st Qu.:  7
##  Median : 39.00   E      : 9024   S      : 9598   Median : 13
##  Mean   : 39.98   SSE    : 8966   WSW    : 9329   Mean   : 14
##  3rd Qu.: 48.00   NW     : 8552   SW     : 9182   3rd Qu.: 19
##  Max.   :135.00   (Other):85083   (Other):89732   Max.   :130
##  NA's   :9270     NA's   :10013   NA's   : 3778   NA's   :1348
##   WindSpeed3pm    Humidity9am      Humidity3pm      Pressure9am
##  Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 980.5
##  1st Qu.:13.00   1st Qu.: 57.00   1st Qu.: 37.00   1st Qu.:1012.9
##  Median :19.00   Median : 70.00   Median : 52.00   Median :1017.6
##  Mean   :18.64   Mean   : 68.84   Mean   : 51.48   Mean   :1017.7
##  3rd Qu.:24.00   3rd Qu.: 83.00   3rd Qu.: 66.00   3rd Qu.:1022.4
##  Max.   :87.00   Max.   :100.00   Max.   :100.00   Max.   :1041.0
##  NA's   :2630    NA's   :1774     NA's   :3610     NA's   :14014
##   Pressure3pm        Cloud9am        Cloud3pm        Temp9am
##  Min.   : 977.1   Min.   :0.00    Min.   :0.0     Min.   :-7.20
##  1st Qu.:1010.4   1st Qu.:1.00    1st Qu.:2.0     1st Qu.:12.30
##  Median :1015.2   Median :5.00    Median :5.0     Median :16.70
##  Mean   :1015.3   Mean   :4.44    Mean   :4.5     Mean   :16.99
##  3rd Qu.:1020.0   3rd Qu.:7.00    3rd Qu.:7.0     3rd Qu.:21.60
##  Max.   :1039.6   Max.   :9.00    Max.   :9.0     Max.   :40.20
##  NA's   :13981    NA's   :53657   NA's   :57094   NA's   :904
##     Temp3pm      RainToday        RISK_MM        RainTomorrow
##  Min.   :-5.40   No  :109332   Min.   :  0.000   No :110316
##  1st Qu.:16.60   Yes : 31455   1st Qu.:  0.000   Yes: 31877
##  Median :21.10   NA's:  1406   Median :  0.000
##  Mean   :21.69                 Mean   :  2.361
##  3rd Qu.:26.40                 3rd Qu.:  0.800
##  Max.   :46.70                 Max.   :371.000
##  NA's   :2726
```

La primera conclusió que s'extreu és la gran quantitat de valor perduts (NA's) i l'existència de valors extrems. Totes aquestes anomalies seran tractades en els pròxims apartats.

Una vegada observades totes les dades disponibles, seleccionarem aquelles que siguin d'interès i aportin valor al anàlisi que es vol realitzar. En aquest cas, la variable "RISK_MM" no aporta un valor significatiu a la mostra i per tant, pot ser eliminada. Per altra banda, la resta de variables aporten positivament valor a la mostra i per tant, les deixem intactes.

```
# S'elimina la columna/variable "RISK_MM"
data <- data[, -(23)]
```

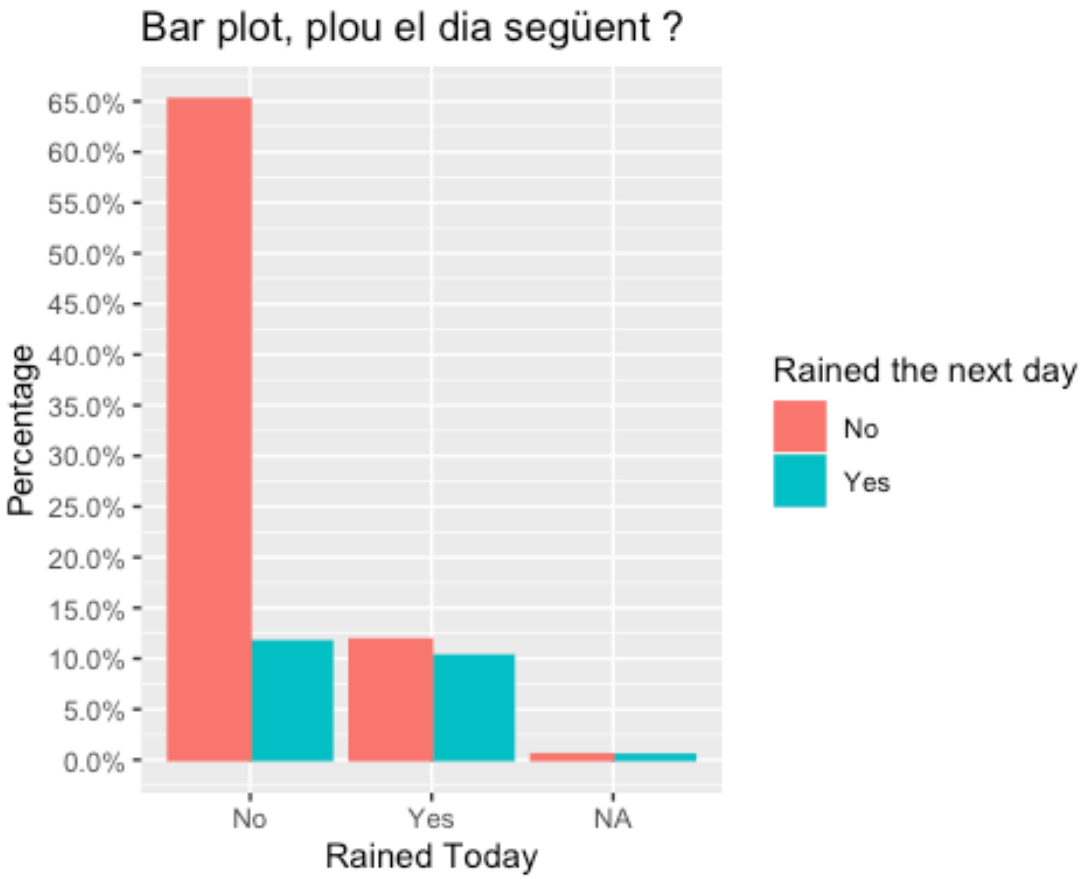
Visualització i representació de dades

La visualització de dades ajuda a interpretar moltes hipòtesis i trobar relacions entre variables que passen desapercebudes. En aquest exemple, es mostrarà mitjançant una representació visual la incògnita següent :

- Quan plou un dia, sol ploure el dia següent ?

```
library(tidyverse)

data1 <- data

data1 = data1 %>%
    mutate_at(vars(Location, WindGustDir, WindDir9am, WindDir3pm,
RainToday, RainTomorrow), as.factor)

data1 %>% summarise_each(list(~ sum(is.na(.)) / length(.) * 100))

##   Date Location   MinTemp   MaxTemp  Rainfall Evaporation Sunshine
WindGustDir
## 1    0        0 0.4479827 0.2264528 0.9887969    42.78903 47.69292
6.561504
##   WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm
Humidity9am
## 1      6.519308   7.041838   2.656952    0.9480073     1.849599
1.2476
##   Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm   Temp9am
Temp3pm
## 1    2.538803    9.855619    9.832411 37.73533 40.15247 0.6357556
1.917113
##   RainToday RainTomorrow
## 1 0.9887969            0

data1 %>%
    group_by(RainTomorrow) %>%
    summarise_each(list(~ sum(is.na(.)) / length(.) * 100))

## # A tibble: 2 x 23
##   RainTomorrow  Date Location MinTemp MaxTemp Rainfall Evaporation
Sunshine
##   <fct>        <dbl>    <dbl>   <dbl>   <dbl>    <dbl>       <dbl>
<dbl>
## 1 No               0        0   0.419   0.242    0.662        42.4
47.8
## 2 Yes              0        0   0.549   0.173    2.12         44.0
47.2
## # … with 15 more variables: WindGustDir <dbl>, WindGustSpeed <dbl>,
## #   WindDir9am <dbl>, WindDir3pm <dbl>, WindSpeed9am <dbl>,
WindSpeed3pm <dbl>,
```

```
## #   Humidity9am <dbl>, Humidity3pm <dbl>, Pressure9am <dbl>,
Pressure3pm <dbl>,
## #   Cloud9am <dbl>, Cloud3pm <dbl>, Temp9am <dbl>, Temp3pm <dbl>,
## #   RainToday <dbl>

data1 %>%
    ggplot(aes(x = RainToday, fill = RainTomorrow, color = RainTomorrow))
+
    geom_bar(aes(y = ((..count..) / sum(..count..))), position = "dodge")
+
    scale_y_continuous(breaks = seq(0, 1, by = 0.05),
                       labels = scales::percent) +
    labs(x = "Rained Today",
         y = "Percentage",
         title = "Bar plot, plou el dia següent ?",
         color = "Rained the next day",
         fill = "Rained the next day"
         )
```



Aquest gràfic ens permet determinar que quan plou un dia, acostuma a ploure el dia següent i per tant, no es solen produir pluges puntuals, sino més aviat de llarga durada. La visualització ens mostra un conjunt de dades buides (NA), les qual s'han d'analitzar per definir uns resultats més acurats.

Neteja de dades

En aquest apartat es començarà a moldejar la mostra amb la finalitat de trobar el conjunt de dades més eficient i fàcil d'analitzar. En primer lloc, s'ajustarà la variable "Date", passant de tipus "factor" a "Date" i transformant les dates en un format més òptim, com són els mesos, tanmateix es convertiran les varibles binàries RainToday i RainTomorrow en numèriques (de ["Yes", "No"] a [1, 0]). Aquesta modificació permetrà una millor resolució del problema plantejat.

```
library(tidyverse)

# Conversió a tipus "Date"
data$Date <- as.Date(as.character(data$Date))
# Classificació de les dates per mesos (Jan, Feb, Mar..)
data$Date = month.abb[lubridate::month(data$Date)]

# Conversió de la variable "RainTomorrow"
data$RainTomorrow <- str_replace_all(data$RainTomorrow, "No", "0")
data$RainTomorrow <- str_replace_all(data$RainTomorrow, "Yes", "1")
data$RainTomorrow <- as.integer(data$RainTomorrow)
```

En segon lloc, s'analitzaran els possibles valors errònis, ja que poden comportar confusions i resultats desviats si es tenen en compte.

- Zeros i elements buits

Com ja s'ha comentat anteriorment, existeixen una gran quantitat de valors buits que representen una pèrdua d'informació i per tant, cal tractar-los de la millor manera. Existeix la possibilitat que aquests valors representin certs valors sentinella, els quals s'haurien d'analitzar, però no és el cas. A continuació es pot observar el nombre total de valors NA per cada variable i el percentatge corresponent [0-1] :

```
library(purrr)

# Visualització del nombre total de valors NA
sapply(data, function(x) sum(is.na(x)))
```

```
##          Date      Location       MinTemp       MaxTemp      Rainfall
##             0             0           637           322          1406
##   Evaporation      Sunshine   WindGustDir WindGustSpeed    WindDir9am
##         60843         67816          9330          9270         10013
##    WindDir3pm  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm
##          3778          1348          2630          1774          3610
##   Pressure9am   Pressure3pm      Cloud9am      Cloud3pm       Temp9am
##         14014         13981         53657         57094           904
##       Temp3pm     RainToday  RainTomorrow
##          2726          1406             0
```

```r
# Percentatge de valors NA per variable
map(data, ~mean(is.na(.)))
```

```
## $Date
## [1] 0
##
## $Location
## [1] 0
##
## $MinTemp
## [1] 0.004479827
##
## $MaxTemp
## [1] 0.002264528
##
## $Rainfall
## [1] 0.009887969
##
## $Evaporation
## [1] 0.4278903
##
## $Sunshine
## [1] 0.4769292
##
## $WindGustDir
## [1] 0.06561504
##
## $WindGustSpeed
## [1] 0.06519308
##
## $WindDir9am
## [1] 0.07041838
##
## $WindDir3pm
## [1] 0.02656952
##
## $WindSpeed9am
## [1] 0.009480073
##
## $WindSpeed3pm
## [1] 0.01849599
##
## $Humidity9am
## [1] 0.012476
##
## $Humidity3pm
## [1] 0.02538803
##
## $Pressure9am
## [1] 0.09855619
```

```
##
## $Pressure3pm
## [1] 0.09832411
##
## $Cloud9am
## [1] 0.3773533
##
## $Cloud3pm
## [1] 0.4015247
##
## $Temp9am
## [1] 0.006357556
##
## $Temp3pm
## [1] 0.01917113
##
## $RainToday
## [1] 0.009887969
##
## $RainTomorrow
## [1] 0
```

Amb aquests resultats es pot confirmar l'elevat nombre de valors buits i més concretament, en les variables "Cloud9am", "Cloud3pm", "Sunshine" i "Evaporation" el percentatge de NA's supera el 35%. Per resoldre aquest conflicte, es pot optar per :

- Eliminar els registres que continguin un valor NA.
- Eliminar les variables amb un percentatge de NA superior.
- Imputar els valors NA.

En aquest cas s'opta per eliminar els registres que contene un valor NA, ja que el total de registres (142.193) és molt elevat i es creu que no afectarà de forma significativa el resultat final de l'anàlisi.

```
# Supressió dels valors NA
data <- na.omit(data)
# Nombre de registres finals
nrow(data)
```

```
## [1] 56420
```

```
# Visualització del nombre total de valors NA
sapply(data, function(x) sum(is.na(x)))
```

```
##          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      Cloud9am      Cloud3pm       Temp9am
```

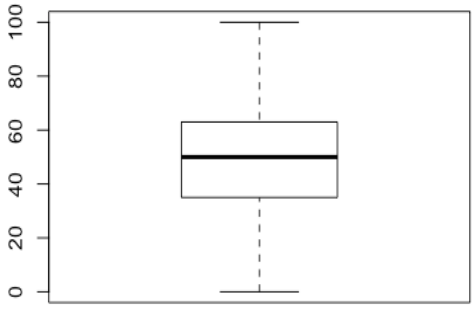
```
##             0             0             0             0             0
##       Temp3pm     RainToday  RainTomorrow
##             0             0             0
```

Tot i que s'hagui decidit per eliminar registres i per tant, perdre certa informació, el nombre de registres totals (56.420) continua sent prou elevat per aplicar mètodes de regressió i predicció. Per tant, es considera l'opció escollida com a satisfactoria. L'última funció valida la supressió de valors NA.

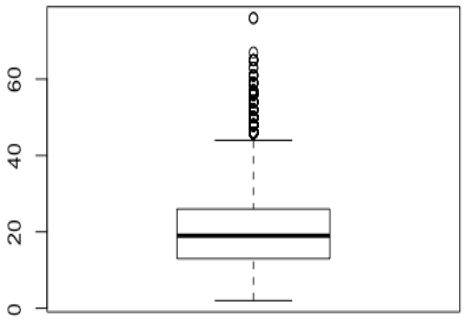
- Valors extrems

Un cop els valors buits ja han estat gestionats, cal tractar els anomenats valors extrems (outliers). Aquests valors s'identifiquen per ser relativament diferents a la majoria, els valors atípics s'han d'analitzar per poder afirmar si formen valors errònis o poden ser incorporats al conjunt de dades. Per trobar-los es pot utilitzar la tècnica boxplot, ja sigui de forma automàtica amb una funció present a R o amb la representació d'un diagrama de caixa. A continuació es pot veure l'aplicació d'ambdós mètodes, en aquelles variables numèriques :

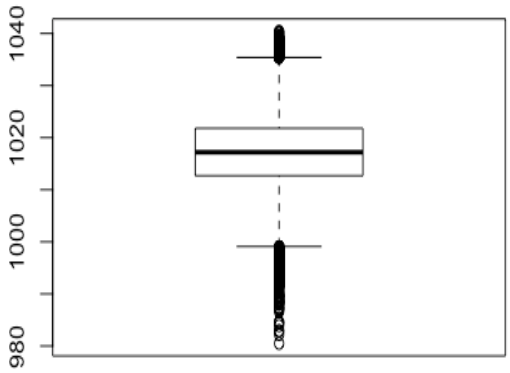
```
# Representació de valors extrems vía funció directa
boxplot.stats(data$MinTemp)$out # Valors extrems en la variable
"Temperatura mínima".

## [1] -6.7

boxplot.stats(data$MaxTemp)$out # Valors extrems en la variable
"Temperatura màxima".

## [1] 47.3 46.4 46.8 46.4 46.7 46.3 46.7 48.1 46.8

boxplot.stats(data$Sunshine)$out # No existeixen valors extrems en la
variable "Sunshine".

## numeric(0)

boxplot.stats(data$Temp9am)$out # Valors extrems en la variable
"Temperatura 9h00".

## [1] 39.4 39.0 38.9 39.1

boxplot.stats(data$Temp3pm)$out # Valors extrems en la variable
"Temperatura 15h00".

##  [1] 45.8 44.7 44.9 43.9 44.1 46.1 45.4 46.1 44.8 44.5 45.2 44.1 43.7
45.3 46.1
## [16] 45.8

# Representació de valors extrems vía diagrama de caixa
boxplot(data$Humidity3pm) # No existeixen valors extrems en la variable
"Humitat 15h00".
```



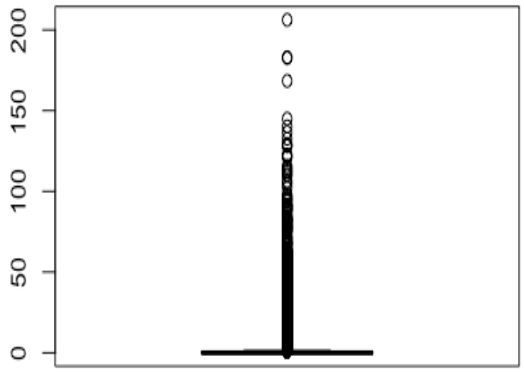

```r
boxplot(data$WindSpeed3pm) # Valors extrems en la variable "WindSpeed3pm" representat per punts.
```




```r
boxplot(data$Pressure9am) # Valors extrems en la variable "Pressure9am" representat per punts.
```




```r
boxplot(data$Rainfall) # Valors extrems en la variable "Rainfall"
representat per punts.
```




```r
# No es representaran les variables següents :
# boxplot(data$Evaporation) # boxplot(data$WindSpeed9am) #
boxplot(data$Humidity3pm)
# boxplot(data$WindGustSpeed) # boxplot(data$WindSpeed3pm) #
boxplot(data$Pressure9am)
# boxplot(data$Humidity9am) # boxplot(data$Cloud9am)
# boxplot(data$Pressure3pm) # boxplot(data$Cloud3pm)
```

Un cop analitzats els valors extrems obtinguts, s'ha determinat que tots els registres comprenen una part llògica i per tant són correctes. Per exemple, la variable amb més valors extrems és "Rainfall", amb valors pròxims als 200 mm. Tot i que pugui semblar una quantitat molt elevada, l'arribada d'un cicló podria comportar tals valors (més de 200 mm de pluja.) i en els últims 13 anys, a Australia, n'hi han hagut. No s'ha adjuntat l'anàlisi de valors extrems de totes les variables, però s'han estudiat i determinat que són correctes i per tant, es deixaran tal com estan, sense modificacions.

- Variables categòriques

Per no perdre cap tipus d'informació, les variables categòriques s'han de tractar de forma correcta. És per aquest motiu que es realitzarà una transformació cap a variables numèriques, amb l'objectiu de millorar els resultats a l'hora d'aplicar qualsevol mètode d'anàlisi.

La tècnica escollida per realitzar d'una manera ràpida i senzilla aquesta operació pertany a la funció dummyVars(). Aquesta dividirà les variables com tants múltiples factors tingui, com per exemple, en la variable Date, aquesta és transformarà en DateJan, DateFeb, DateMar, etc .. La funció s'aplicarà per totes les variables menys "RainTomorrow", ja que simbolitza la variable objectiu.

```
library(caret)

# Anàlisi de les variables categòriques
dummy <- dummyVars(" ~ .", data[,1:22])
# Nou data.frame amb les noves variables numèriques
dataDum <- data.frame(predict(dummy, data[,1:22]))
# Nou data.frame final
dataFinal <- cbind(dataDum[,1:127], RainTomorrow = data$RainTomorrow)
```

El data frame (dataFinal) conté les dades preprocessades que estan llestes per ser utilitzades en la pròxima fase d'anàlisi.

Nota: Aquest data frame s'utilitzarà en cas de voler analitzar les variables categòriques. Contràriament s'utilitzarà el data frame (data).

- Exportació de les dades preprocessades

Tal i com s'ha comentat anteriorment, el data frame amb les dades preprocessades ja ha estat creat i ara toca exportar-lo per a poder ser utilitzat en la fase d'anàlisi.

```
# Nou fitxer amb les dades preprocessades
write.csv(dataFinal, "weatherAUS_clean.csv")
```

Anàlisi de les dades

- Selecció dels grups de dades que es volen analitzar/comparar

L'objectiu de la selecció es basa en dividir la mostra en diferents sub-conjunts amb el fi de poder aplicar hipòtesis i comparar resultats d'interès, des d'un punt de vista analític. És per això que en aquest apartat es prepararan certs grups de dades, que posteriorment seran analitzats i comparats per mètodes analítics (no la totalitat).

Per realitzar el filtratge, primer de tot es realitzarà una exploració de les dades i s'acotarà del total d'aquestes, aquelles que més interessen, tot això permetrà prescindir d'informació redundant.

```
# Per realitzar comparacions segons les estacions de l'any, es pot
dividir el conjunt de la següent manera :
dataFinal.tardor <- dataFinal[dataFinal$DateMar == 1 | dataFinal$DateApr
== 1 | dataFinal$DateMay == 1,] # Conjunt de dades a la tardor

dataFinal.estiu <- dataFinal[dataFinal$DateDec == 1 | dataFinal$DateJan
== 1 | dataFinal$DateFeb == 1,] # Conjunt de dades a l'estiu

dataFinal.hivern <- dataFinal[dataFinal$DateJun == 1 | dataFinal$DateJul
== 1 | dataFinal$DateAug == 1,] # Conjunt de dades a l'hivern

dataFinal.primavera <- dataFinal[dataFinal$DateSep == 1 |
dataFinal$DateOct == 1 | dataFinal$DateNov == 1,] # Conjunt de dades a la
primavera

# Una altra composició de conjunts de dades poden ser les ciutats
d'Australia. En el cas que es vulgui analitzar la ciutat de Sydney més
detalladament, es podria constituir un conjunt de dades de la següent
manera :
dataFinal.sydney <- dataFinal[dataFinal$Location.Sydney == 1,] # Conjunt
de dades de Sydney
```

En cas de voler comparar mètodes estadístics sobre conjunts de dades diferents, la tècnica emprada anteriorment representa la millor forma de segmentar i dividir la mostra original en parts desitjades. En el nostre cas, es posa per exemple que ens agradaria analitzar si les pluges a la primavera són més pronunciades que a la tardor i per això s'ha dividit la mostra en "dataFinal.primavera" i "dataFinal.tardor".

Nota: Les estacions meteorològiques a Autràlia no són les mateixes que a Europa.

- Comprovació de la normalitat i homogeneïtat de la variància

L'objectiu d'aquest apartat, en primer lloc, és el de verificar la suposició de normalitat de les variables quantitatives que formen la mostra. En segons lloc es comprovarà la igualtat de variàncies entre els grups que s'han de comparar, és a dir, l'anàlisi d'homoscedasticitat.

Les proves més habituals de normalitat, són els tests de Kolmogorov-Smirnov i de Shapiro-Wilk, tot i que en el nostre cas, s'utilitzarà la prova de normalitat de Anderson- Darling. En canvi, en les proves d'homogeneïtat de variàncies és habitual utilitzar la proba de Levene (si es segueix una distribució normal) i la de Fligner-Killeen (No paramètrica).

En la comprovació de normalitat no es prenen en compte les variables categòriques, per tant, s'utilitzarà el data.frame "data" per determinar si les variables numèriques segueixen una distribució normal.

Les variables seguiran una distribució normal si el p-valor obtingut és superior al nivell de significació establert de $\alpha = 0{,}05$.

```
library(nortest)

# Test de normalitat
nv = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
      p_val = ad.test(data[,i])$p.value
      cat("Variable : ", col.names[i]," ", "p-value : ", p_val, "\n")
    }
  }
```

```
## Variable :  MinTemp   p-value :  3.7e-24
## Variable :  MaxTemp   p-value :  3.7e-24
## Variable :  Rainfall   p-value :  3.7e-24
## Variable :  Evaporation   p-value :  3.7e-24
## Variable :  Sunshine   p-value :  3.7e-24
## Variable :  WindGustSpeed   p-value :  3.7e-24
## Variable :  WindSpeed9am   p-value :  3.7e-24
## Variable :  WindSpeed3pm   p-value :  3.7e-24
## Variable :  Humidity9am   p-value :  3.7e-24
## Variable :  Humidity3pm   p-value :  3.7e-24
## Variable :  Pressure9am   p-value :  5.399032e-23
## Variable :  Pressure3pm   p-value :  3.7e-24
## Variable :  Cloud9am   p-value :  3.7e-24
## Variable :  Cloud3pm   p-value :  3.7e-24
## Variable :  Temp9am   p-value :  3.7e-24
## Variable :  Temp3pm   p-value :  3.7e-24
## Variable :  RainTomorrow   p-value :  3.7e-24
```

No s'ha pogut utilitzar la prova de Shapiro (shapiro.test) per qué el màxim número de registres no pot ser superior a 5000 i el data.frame "data" en conté més. Tot i això, com es pot comprovar amb els resultats obtinguts, tots els valors p-value són inferiors al nivell de significació, fet que manifesta la no distribució normal del conjunt de variables.

Degut als resultats no paramètrics de la prova de normalitat, es procedirà amb la comprovació d'homoscedasticitat amb l'aplicació del test Fligner-Killeen.

```
# Test d'homoscedasticitat
fligner.test(Sunshine ~ RainToday, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Sunshine by RainToday
## Fligner-Killeen:med chi-squared = 174.49, df = 1, p-value < 2.2e-16

fligner.test(Temp9am ~ RainToday, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Temp9am by RainToday
## Fligner-Killeen:med chi-squared = 20.386, df = 1, p-value = 6.329e-06

fligner.test(Humidity3pm ~ RainToday, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Humidity3pm by RainToday
## Fligner-Killeen:med chi-squared = 333.75, df = 1, p-value < 2.2e-16
```

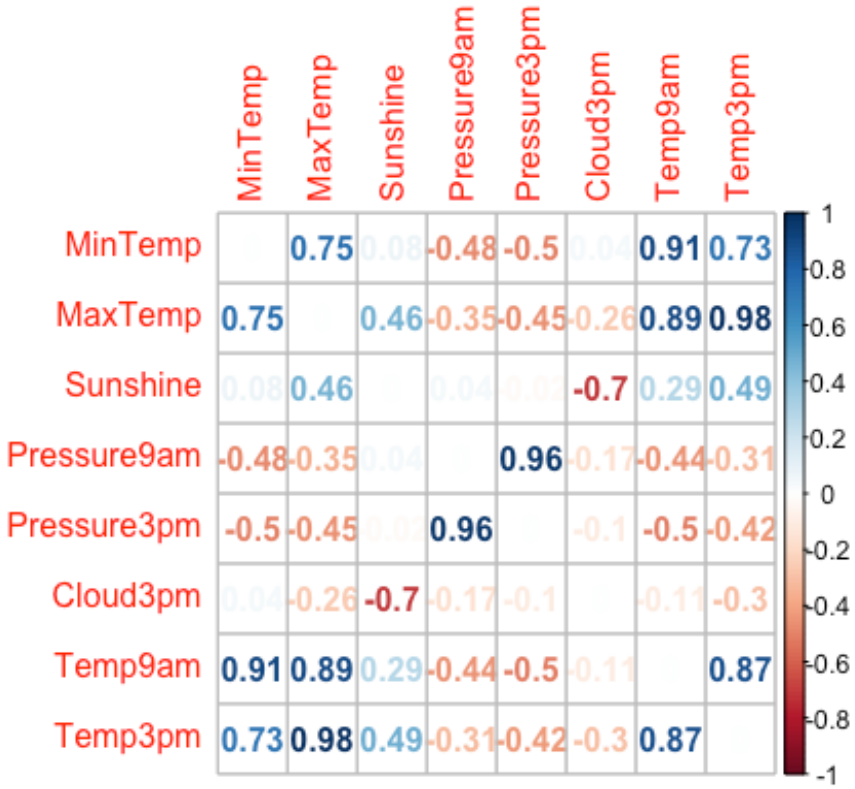
En aquest apartat s'intenta trobar una variable que tingui homoscedasticitat amb el fet de que plogui avui o no. Tal i com es pot comprovar, s'obtenen p-valors molt baixos i per tant, no es pot acceptar la hipòtesi nul·la de que les variàncies d'ambdues mostres són homogènies. Tot i això, la variable que indica més concordància és la "Temp9am".

Anàlisi estadístic

- Correlació de variables

La correlació és un atribut força rellevant a l'hora de buscar influències entre variables d'un conjunt. En aquest cas, s'han estudiat dos tipus de correlacions, la primera d'elles, entre totes les variables de la mostra. La segona, es centra més en les correlacions de les variables respecte si plou o no (RainToday). El test escollit per aquesta segona és el coeficient de correlació de Spearman, ja que la mostra no segueix una distribució normal.

```
library(corrplot)

# Correlació entre variables de la mostra
numeric <- map_lgl(data, is.numeric) # Selecció de les variables numèriques
correlations <- cor(data[,numeric]) # Correlació de les variables numèriques
diag(correlations) <- 0
high <- apply(abs(correlations) >= 0.7, 2, any) # Correlacions superior a 0.8
corrplot(correlations[high, high], method = "number") # Gràfic de correlacions
```



El primer test ens mostra les variables amb una correlació superior a 0.7 (alta). Com es pot observar, hi ha una forta cohesió entre les variables : MinTemp, MaxTemp, Sunshine, Preassure9am, Preassure3pm, Cloud3pm, Temp9am i Temp3pm.

```r
# Correlació Spearman respecte la variable "RainToday"
data$RainToday <- as.character(data$RainToday)
data$RainToday <- replace(data$RainToday, data$RainToday == "No", 0)
data$RainToday <- replace(data$RainToday, data$RainToday == "Yes", 1)
data$RainToday <- as.integer(data$RainToday)

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

for (i in 2:(ncol(data))) {
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    spearman_test = cor.test(data[,i], data[,22], method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value

    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(data)[i]
  }
}
print(corr_matrix)
```

```
##                   estimate       p-value
## MinTemp        0.04105337  1.751508e-22
## MaxTemp       -0.21876123  0.000000e+00
## Rainfall       0.84422686  0.000000e+00
## Evaporation   -0.25028123  0.000000e+00
## Sunshine      -0.33018130  0.000000e+00
## WindGustSpeed  0.13786850 1.909673e-237
## WindSpeed9am   0.08320390  3.145157e-87
## WindSpeed3pm   0.08151760  8.564401e-84
## Humidity9am    0.39614379  0.000000e+00
## Humidity3pm    0.38345629  0.000000e+00
## Pressure9am   -0.17603954  0.000000e+00
## Pressure3pm   -0.09551415 1.849129e-114
## Cloud9am       0.30628621  0.000000e+00
## Cloud3pm       0.27451060  0.000000e+00
## Temp9am       -0.10246475 1.626548e-131
## Temp3pm       -0.22503854  0.000000e+00
## RainToday      1.00000000  0.000000e+00
## RainTomorrow   0.30909823  0.000000e+00
```

Amb els resultats obtinguts d'aquest segon test, es pot observar com la variable més rellevant, en la determinació de si plou o no (RainToday), és "Rainfall", seguit de "Humidity9am" i "Sunshine", cal remarcar que els valors obtingut de R^2 no són gaire elevats.

- Contrast d'hipòtesis

En aquest apartat, es decideix determinar si hi ha una fracció del cel enfosquit pels núvols a les 9h00 del matí igual a l'estiu que a l'hivern. Per respondre aquesta hipòtesi, es divideix la mostra en dos conjunts segons la estació meteorològica desitjada (realitzar en l'apartat "Selecció dels grups de dades que es volen analitzar/comparar").

El contrast que s'aplicarà és sobre dues mostres i per diferència de mitjanes, a continuació es formulen les hipòtesis nul·la i alternativa :

$H0 : \mu1 - \mu2 = 0$ H1 : μ1 − μ2 < 0

*μ1 -> Mitjana del conjunt dataFinal.hivern$Cloud9am * \mu2 -> Mitjanadelconjuntdata Final.estiu$Cloud9am

α = 0, 05.

```
# Contrast d'hipòtesis
t.test(dataFinal.hivern$Cloud9am, dataFinal.estiu$Cloud9am, alternative =
"less")

##
##  Welch Two Sample t-test
##
## data:  dataFinal.hivern$Cloud9am and dataFinal.estiu$Cloud9am
## t = -5.4958, df = 27758, p-value = 1.962e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1297701
## sample estimates:
## mean of x mean of y
##  4.166833  4.352035
```

Com era d'esperar, s'obté un p-valor inferior al valor de significació i per tant, es rebutja la hipòtesi nul·la que mantenia la igualitat de núvols en les estacions d'hivern i estiu. Obviament, a l'hivern hi haurà mes núvols i períodes de cel "engrisat".

- Model de regresió lineal

Per obtenir un model de regressió que s'ajusti amb un grau considerable de precisió, s'han de definir aquelles variables que estan més correlacionades amb la variable que es vol obtenir/estudiar. L'anàlisi ja ha estat efectuat en els altres apartats i les variables que es pendran en compte són les següents :

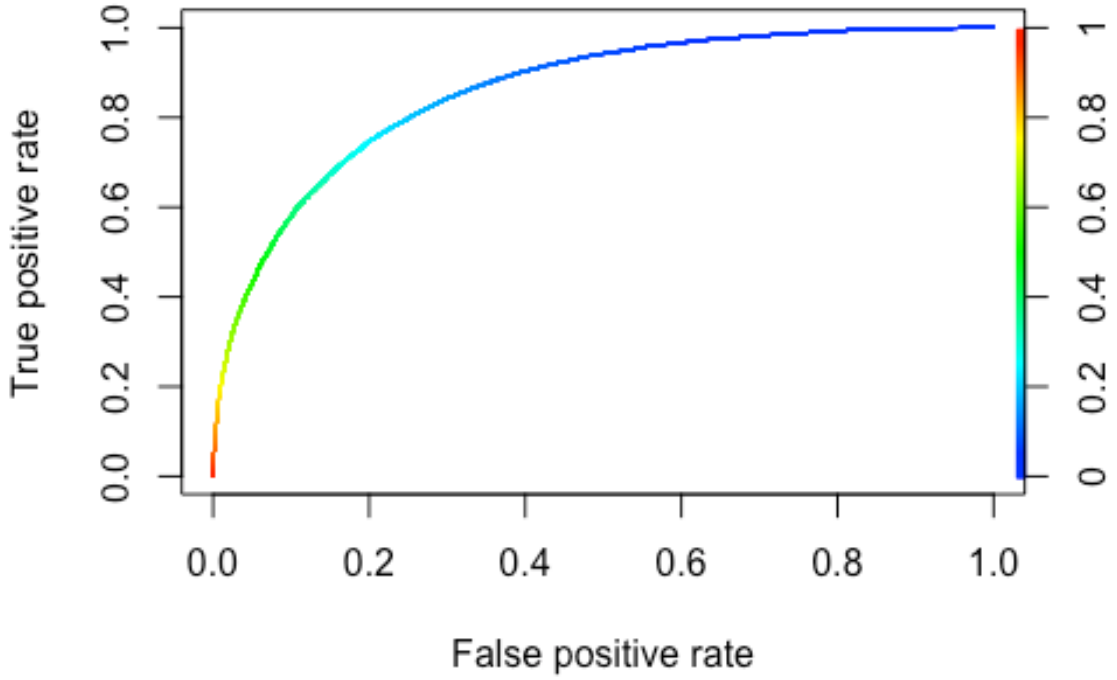
Rainfall, Humidity9am, Sunshine, MinTemp, MaxTemp, Pressure3pm, Temp3pm i Cloud3pm.

A continuació es crearan diferents models de regressió amb l'objectiu de triar aquell que millor interpreti les dades. Cal tenir en compte que la variable a preedir és binomial i per tant, els processos poden presentar certs dubtes que no apareixerien si es tingués una variable numérica. Com per exemple passaria en un hipotètic estudi de la temperatura a la ciutat de Sydney, si es vol predir la temperatura dels pròxims mesos. El problema que ha estat escollit en aquest projecte correspon a predir la variable "RainTomorrow" i per tant, saber si demà plourà.

```
library(ROCR)

# Regresors quantitatius
Rainfall = data$Rainfall
Humidity9am = data$Humidity9am
Sunshine = data$Sunshine
MinTemp = data$MinTemp
MaxTemp = data$MaxTemp
Pressure3pm = data$Pressure3pm
Temp3pm = data$Temp3pm
Cloud3pm = data$Cloud3pm

# Variable a predir
RainTomorrow = data$RainTomorrow

# Model de regressió lineal 1
model <- lm(RainTomorrow ~ Rainfall + Humidity9am + Sunshine + MinTemp +
MaxTemp + Pressure3pm + Cloud3pm + Temp3pm, data = data)
summary(model)

##
## Call:
## lm(formula = RainTomorrow ~ Rainfall + Humidity9am + Sunshine +
##     MinTemp + MaxTemp + Pressure3pm + Cloud3pm + Temp3pm, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30838 -0.22127 -0.06645  0.10578  1.13987
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.3654431  0.2610084   55.04   <2e-16 ***
## Rainfall     0.0059645  0.0002226   26.80   <2e-16 ***
## Humidity9am  0.0017561  0.0001014   17.32   <2e-16 ***
## Sunshine    -0.0321967  0.0006420  -50.15   <2e-16 ***
## MinTemp      0.0045501  0.0004067   11.19   <2e-16 ***
## MaxTemp      0.0341142  0.0013165   25.91   <2e-16 ***
## Pressure3pm -0.0137894  0.0002545  -54.19   <2e-16 ***
## Cloud3pm     0.0081859  0.0007983   10.25   <2e-16 ***
```

```
## Temp3pm     -0.0420072  0.0012836  -32.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3465 on 56411 degrees of freedom
## Multiple R-squared:  0.3011, Adjusted R-squared:  0.301
## F-statistic:  3038 on 8 and 56411 DF,  p-value: < 2.2e-16

# Model de regressió lineal 2
model2 = glm(formula = RainTomorrow ~ Rainfall + Humidity9am + Sunshine +
               MinTemp + MaxTemp + Pressure3pm + Cloud3pm + Temp3pm,
               family = binomial,
               data = data)
summary(model2)

##
## Call:
## glm(formula = RainTomorrow ~ Rainfall + Humidity9am + Sunshine +
##     MinTemp + MaxTemp + Pressure3pm + Cloud3pm + Temp3pm, family =
binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9134  -0.5700  -0.3134  -0.1425   3.1924
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 95.9038992  2.1391998   44.83   <2e-16 ***
## Rainfall     0.0286514  0.0018740   15.29   <2e-16 ***
## Humidity9am  0.0186854  0.0009108   20.52   <2e-16 ***
## Sunshine    -0.1627201  0.0050974  -31.92   <2e-16 ***
## MinTemp      0.0711709  0.0038395   18.54   <2e-16 ***
## MaxTemp      0.1786198  0.0100587   17.76   <2e-16 ***
## Pressure3pm -0.0962231  0.0020919  -46.00   <2e-16 ***
## Cloud3pm     0.1475387  0.0076438   19.30   <2e-16 ***
## Temp3pm     -0.2675254  0.0098211  -27.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 59493  on 56419  degrees of freedom
## Residual deviance: 41370  on 56411  degrees of freedom
## AIC: 41388
##
## Number of Fisher Scoring iterations: 5

# Predicció i ROC amb el model 2
prob_pred = predict(model2, type = 'response')
ROCRpred <- prediction(prob_pred, data$RainTomorrow)
```

```
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```




```
# Model de regressió lineal 3 & ANOVA
model3 = glm(formula = RainTomorrow ~ Rainfall + Humidity9am + Sunshine +
               MinTemp + MaxTemp + Pressure3pm + Cloud3pm + Temp3pm,
               family = binomial,
               data = data)
caret::varImp(model3)

##              Overall
## Rainfall    15.28883
## Humidity9am 20.51478
## Sunshine    31.92245
## MinTemp     18.53642
## MaxTemp     17.75778
## Pressure3pm 45.99815
## Cloud3pm    19.30163
## Temp3pm     27.23989

anova(model3, test = "Chisq")

## Analysis of Deviance Table
##
```

```
## Model: binomial, link: logit
##
## Response: RainTomorrow
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        56419      59493
## Rainfall     1   3378.3     56418      56115 < 2.2e-16 ***
## Humidity9am  1   2789.1     56417      53326 < 2.2e-16 ***
## Sunshine     1   6737.1     56416      46589 < 2.2e-16 ***
## MinTemp      1    756.0     56415      45833 < 2.2e-16 ***
## MaxTemp      1    142.3     56414      45691 < 2.2e-16 ***
## Pressure3pm  1   2900.7     56413      42790 < 2.2e-16 ***
## Cloud3pm     1    639.3     56412      42151 < 2.2e-16 ***
## Temp3pm      1    780.3     56411      41370 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions

L'estudi presentat anteriorment és fruit d'una llarga llista d'etapes, començant en primer lloc per la captura de les dades, que en aquest cas ha estat la localització d'un dataset (Kaggle) que s'acotés a les necessitats del problema. Un cop obtingut el dataset, s'ha procedit a la seva inserció al programa analític per un posterior preprocessat. Al llarg del projecte s'ha pogut observar la importància de la fase de visulització i representació de gràfics, tant per fer hipòtesis, com comparacions.

L'etapa més dificultosa i mandrosa, sempre resulta ser la neteja de dades, amb l'anàlisi de valors buits i extrems. La metodologia aplicada ha estat força directa, degut al nombre elevat de dades s'ha decidit per suprimir aquells registres que no coincidien amb un valor estàndard o llògic.

A continuació s'ha arribat a l'etapa d'anàlisi i és aquí on s'han trobat les conclusions més importants. Primer de tot, s'han detallat les variables amb una correlació més elevada, respecte l'hipòtesi de pluja, i la guanyadora ha estat "Rainfall", com és força evident. Amb els models de regressió s'han pogut comprovar quines variables tenen més influència i quines són menys rellevants. S'han vist també les proves de ROC i el model ANOVA en certa mesura.

Personalment, hem quedo amb les ganes d'analitzar en més profunditat el dataset escollit, però estic satisfet de l'anàlisi realitzat. Encara es podria extreure molt de coneixement amb aquestes dades, com un model de predicció sobre la temperatura a Austràlia, o la humitat per poblacions, etc ..

Recursos

- Vegas, E. (2017). Preprocesamiento de datos. Material UOC.
- Gibergans, J. (2017). Regresión lineal múltiple. Material UOC.
- Rovira, C. (2008). Contraste de hipótesis. Material UOC.
- http://www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test
- https://www.r-graph-gallery.com/
- https://www.kaggle.com/jsphyg/weather-dataset-rattle-package (Kernels)