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",
    col_types = cols(Date = col_date(format = "%Y-%m-%d"),
        Evaporation = col_double(), Sunshine = col_double()))
```

```r
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)
```

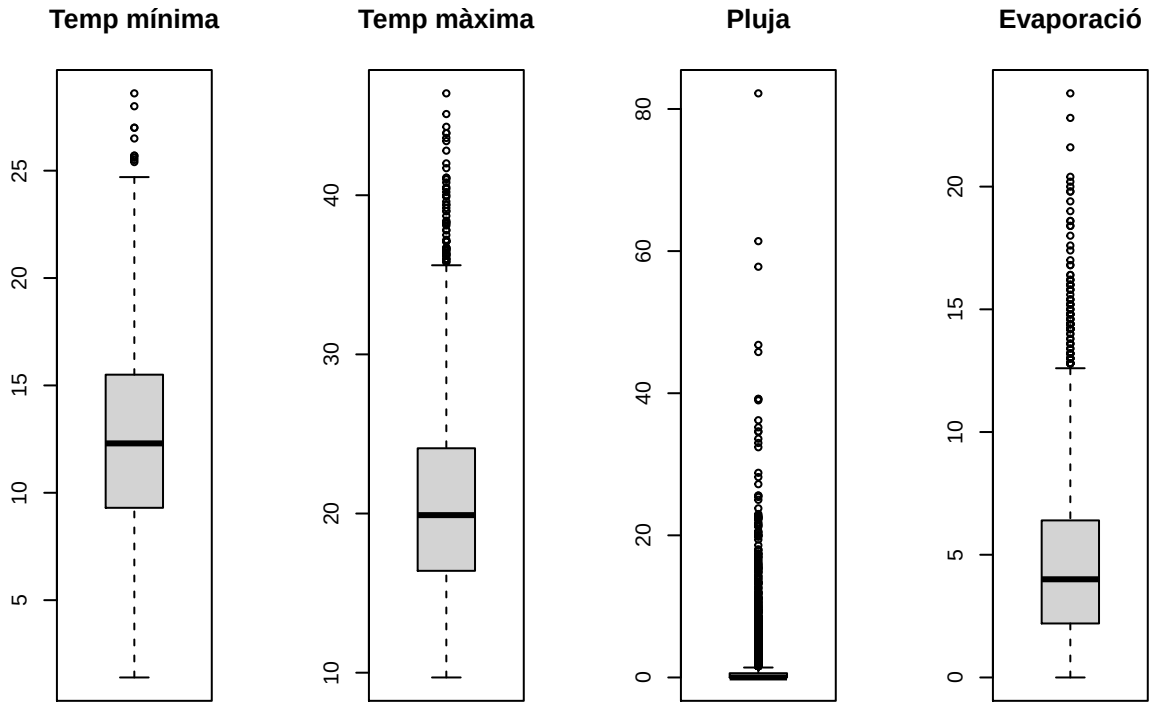
```
##          Date      Location       MinTemp       MaxTemp      Rainfall
##             0             0             0             0             0
##   Evaporation      Sunshine   WindGustDir WindGustSpeed    WindDir9am
##             0             0             0             0             0
##  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

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

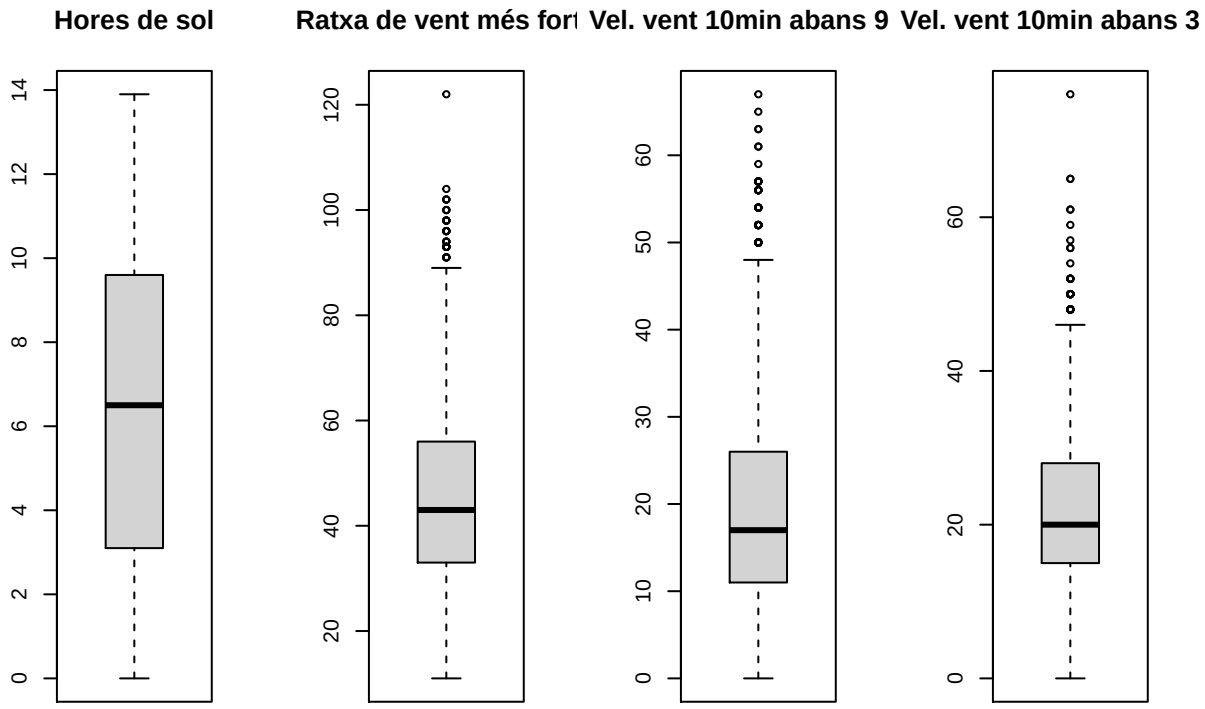
3.2. Identificació i tractament de valors extrems.

Resposta

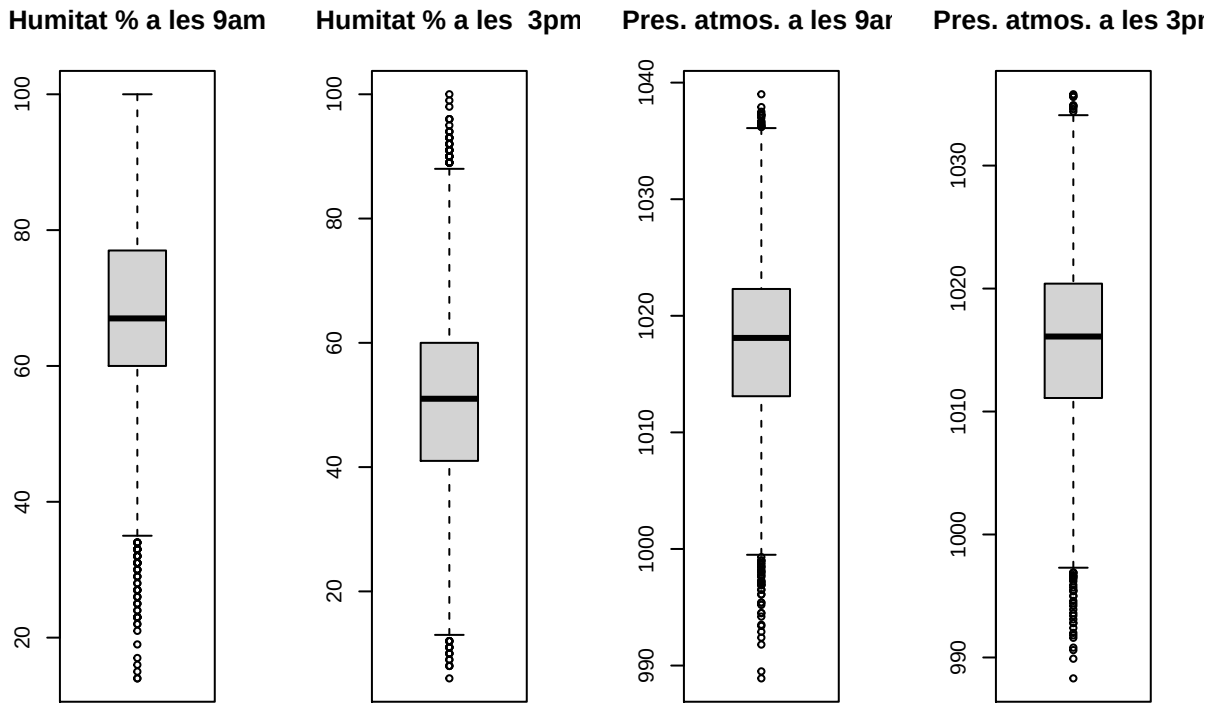
```
par(mfrow=c(1,4))
boxplot(weatherMelb$MinTemp, na.rm=TRUE, main="Temp mínima")
boxplot(weatherMelb$MaxTemp, na.rm=TRUE, main="Temp màxima")
boxplot(weatherMelb$Rainfall, na.rm=TRUE, main="Pluja")
boxplot(weatherMelb$Evaporation, na.rm=TRUE, main="Evaporació")
```



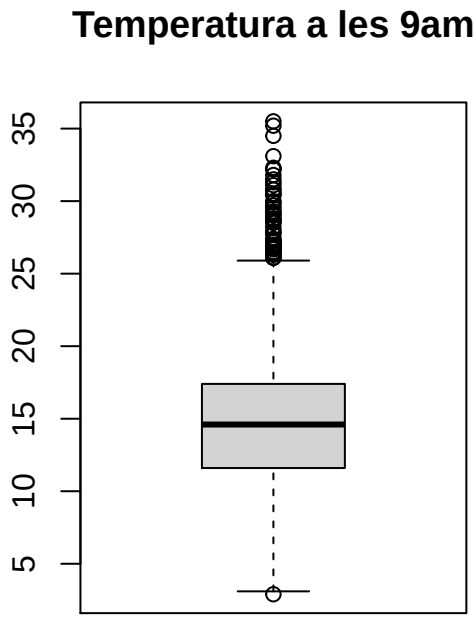

```
par(mfrow=c(1,4))
boxplot(weatherMelb$Sunshine, na.rm=TRUE, main="Hores de sol")
boxplot(weatherMelb$WindGustSpeed, na.rm=TRUE, main="Ratxa de vent més forta")
boxplot(weatherMelb$WindSpeed9am, na.rm=TRUE, main="Vel. vent 10min abans 9am")
boxplot(weatherMelb$WindSpeed3pm, na.rm=TRUE, main="Vel.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

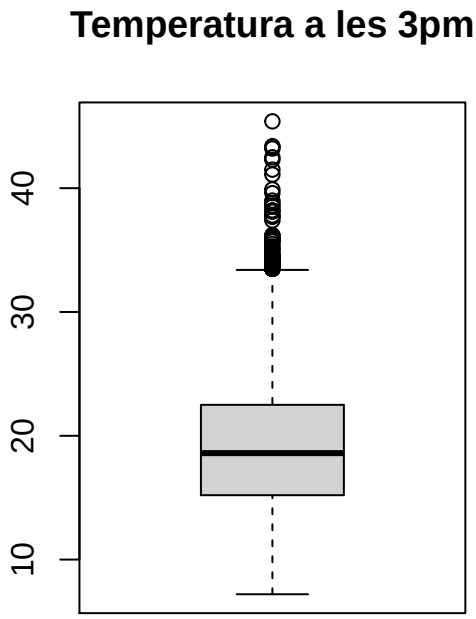



```
par(mfrow=c(1,4))
boxplot(weatherMelb$Humidity9am, na.rm=TRUE, main="Humitat % a les 9am")
boxplot(weatherMelb$Humidity3pm, na.rm=TRUE, main="Humitat % a les  3pm")
boxplot(weatherMelb$Pressure9am, na.rm=TRUE, main=" Pres. atmos. a les 9am")
boxplot(weatherMelb$Pressure3pm, na.rm=TRUE, main=" Pres.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

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

```r
# 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
```

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

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades.

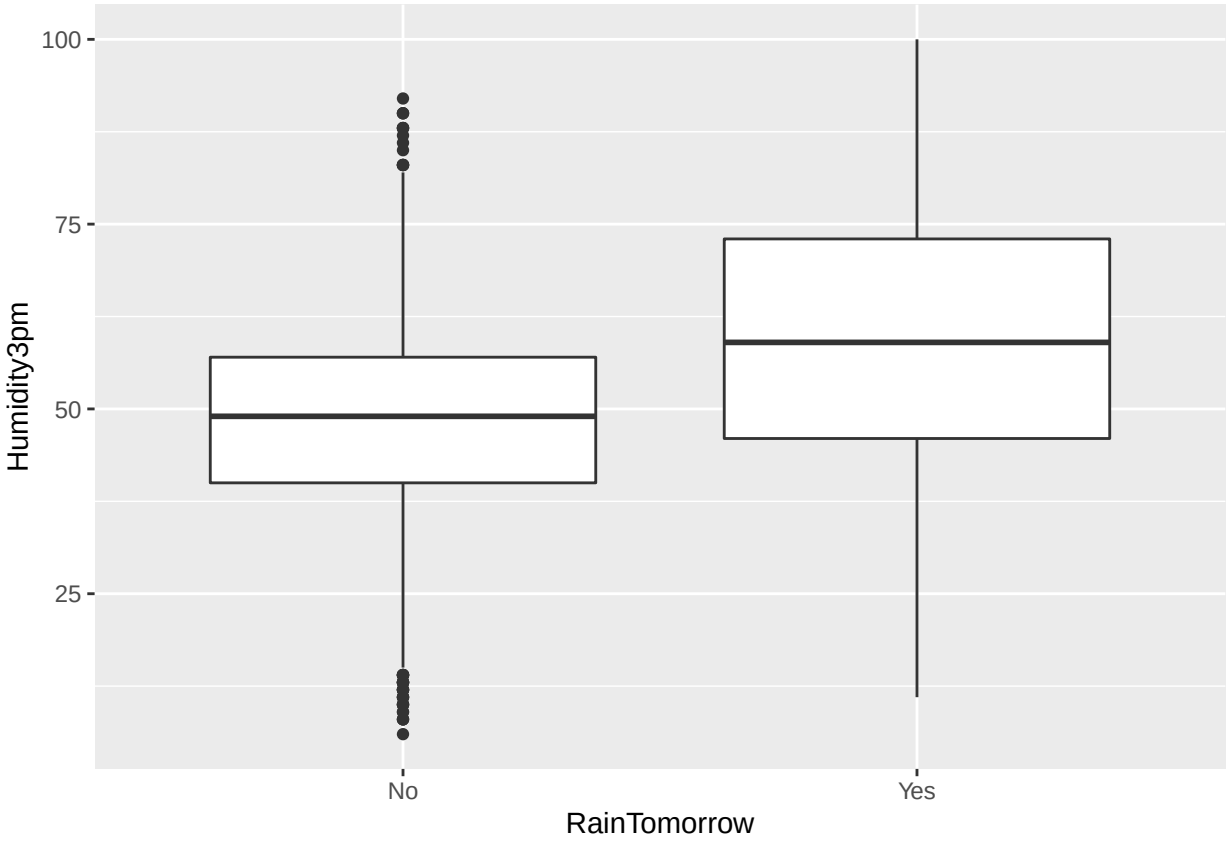
Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Resposta

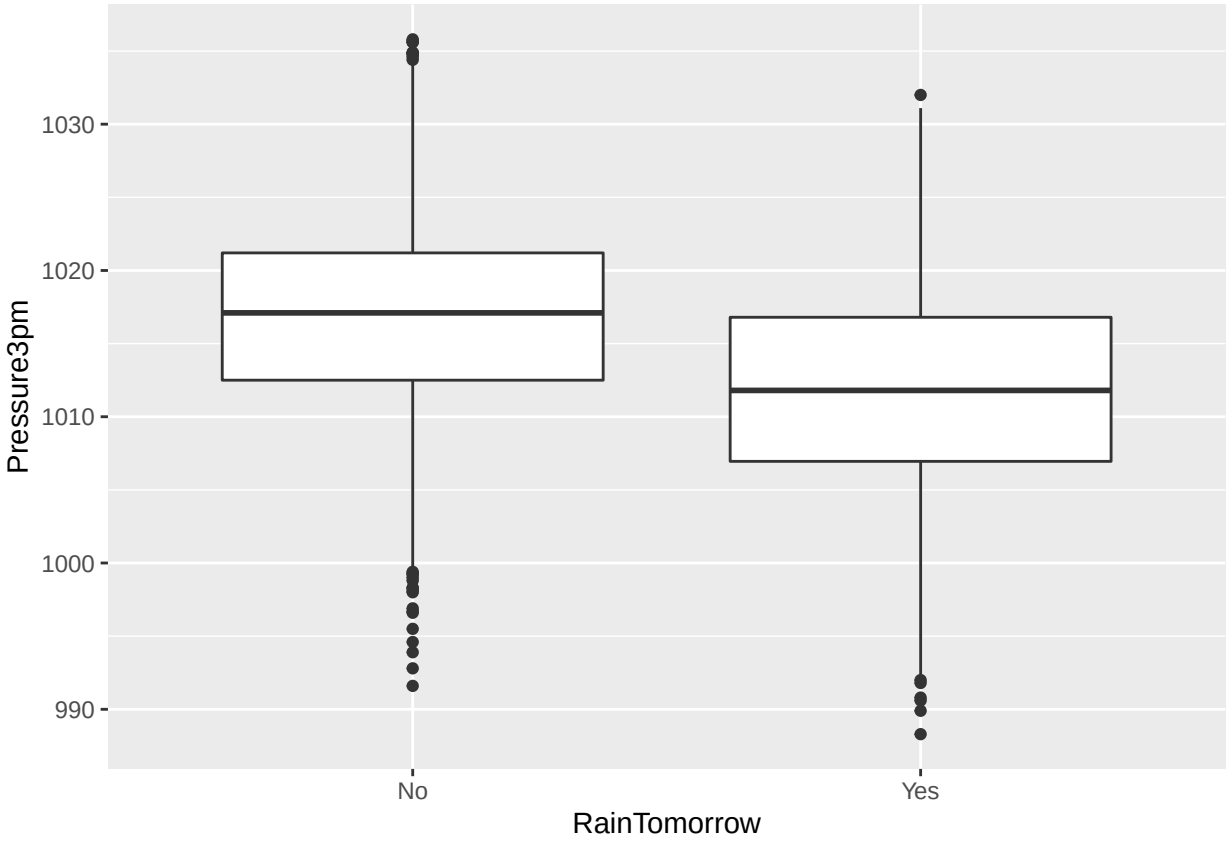
Pendent. Cal establir quines dades utilitzarem per predir la pluja. Cerquem si una (o més) variables ens serveixen per contruir un model que predigui la pluja al dia següent. L'objectiu és determinar si hi ha una relació entre les variables.

Aitor sugeriment: Podríem utilitzar la variable date per a crear una variable mesos. Aquesta variable mesos tindrà una gran correlació amb prediuré la pluja el dia seguent, ja que els mesos de pluja la probabilitat augmenta. La pressió atmosfèrica + humitat també ens ajudaran. Si la pressió baixa + la humitat augmenta, és un fet clar que una massa d'aire humida s'acosta i pot estar associada a un front de pluges. La evaporacio. 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. Un altre factor important és el vent, ja que l'aire fred no saturat absorbeix la humitat amb molta eficàcia.

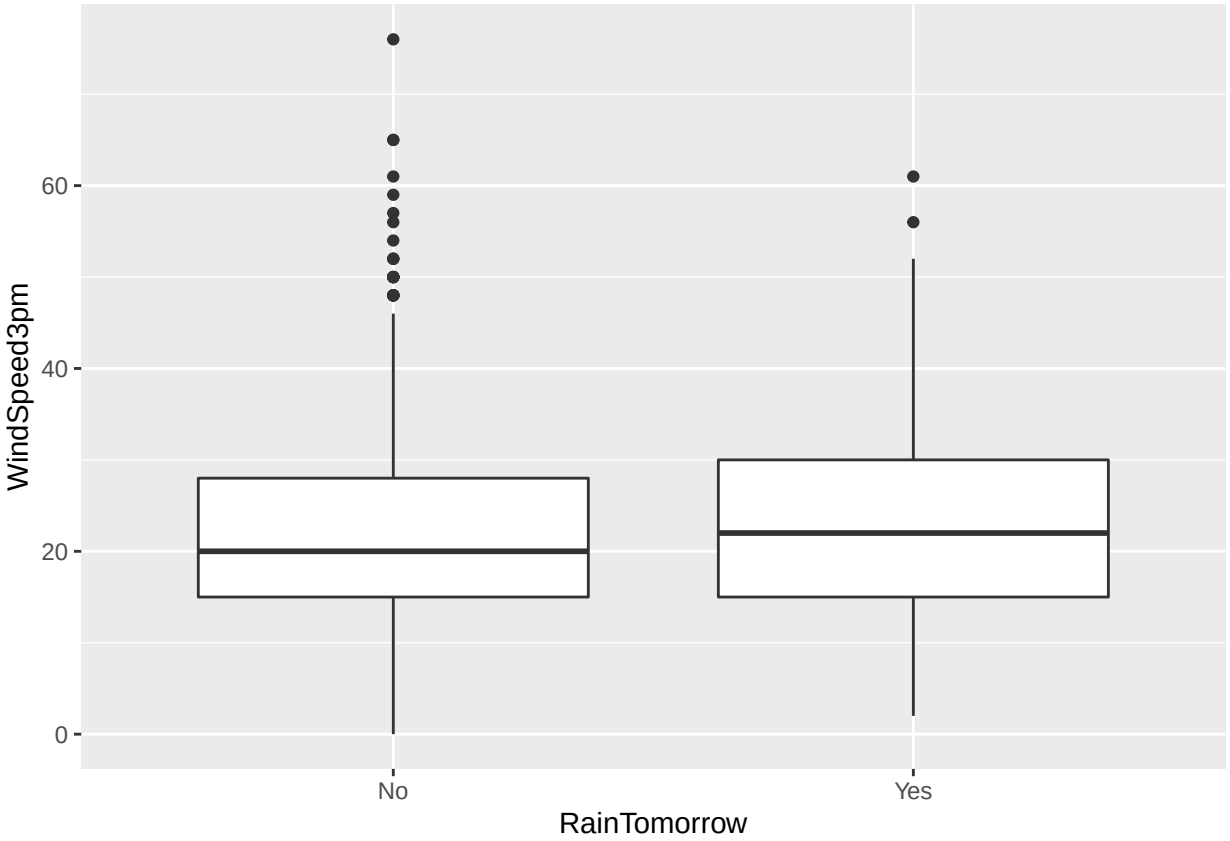
```r
# Podem veure quines variables tenen diferències significatives respecte RainTomorrow
library(ggplot2)
p <- ggplot(weatherMelb, aes(RainTomorrow,Humidity3pm))
p + geom_boxplot()
```



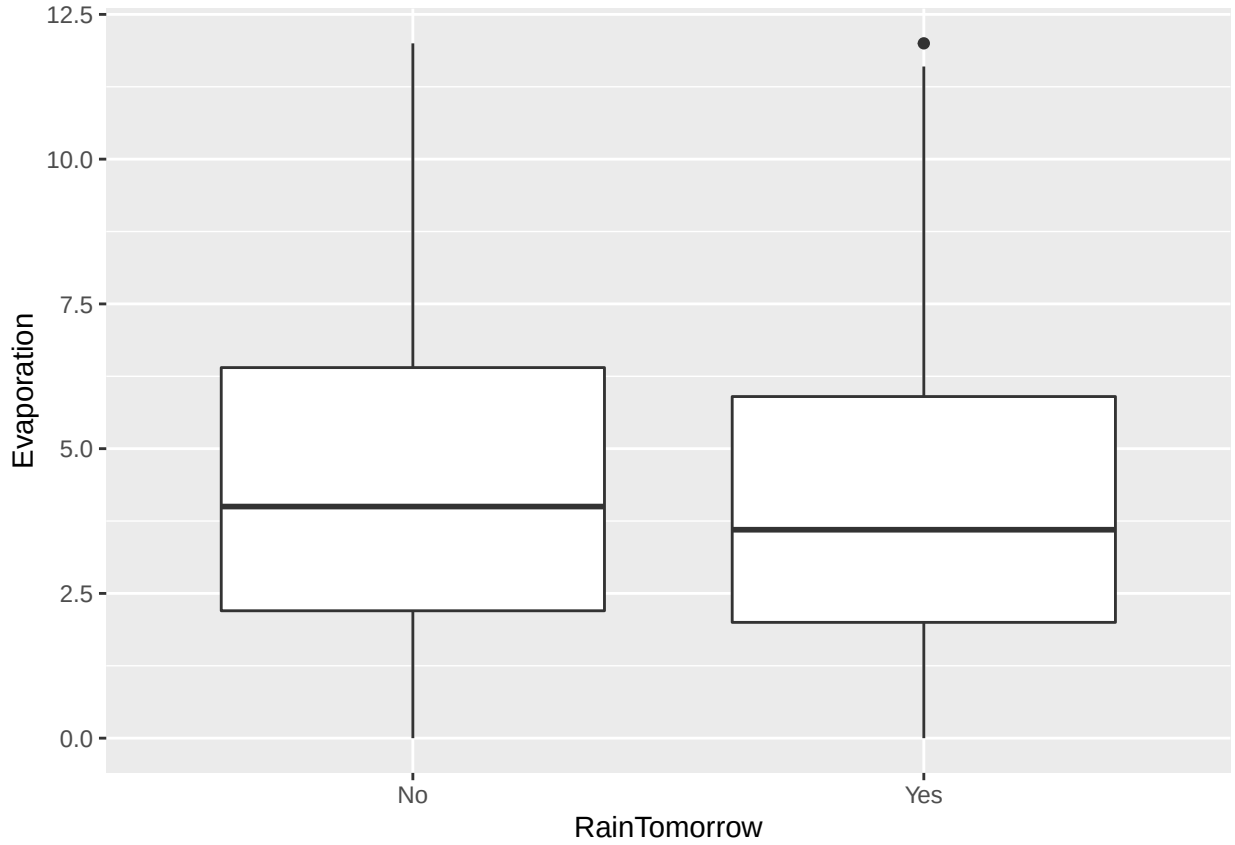

```
# Podem veure quines variables tenen diferències significatives respecte RainTomorrow
p <- ggplot(weatherMelb, aes(RainTomorrow,Pressure3pm))
p + geom_boxplot()
```



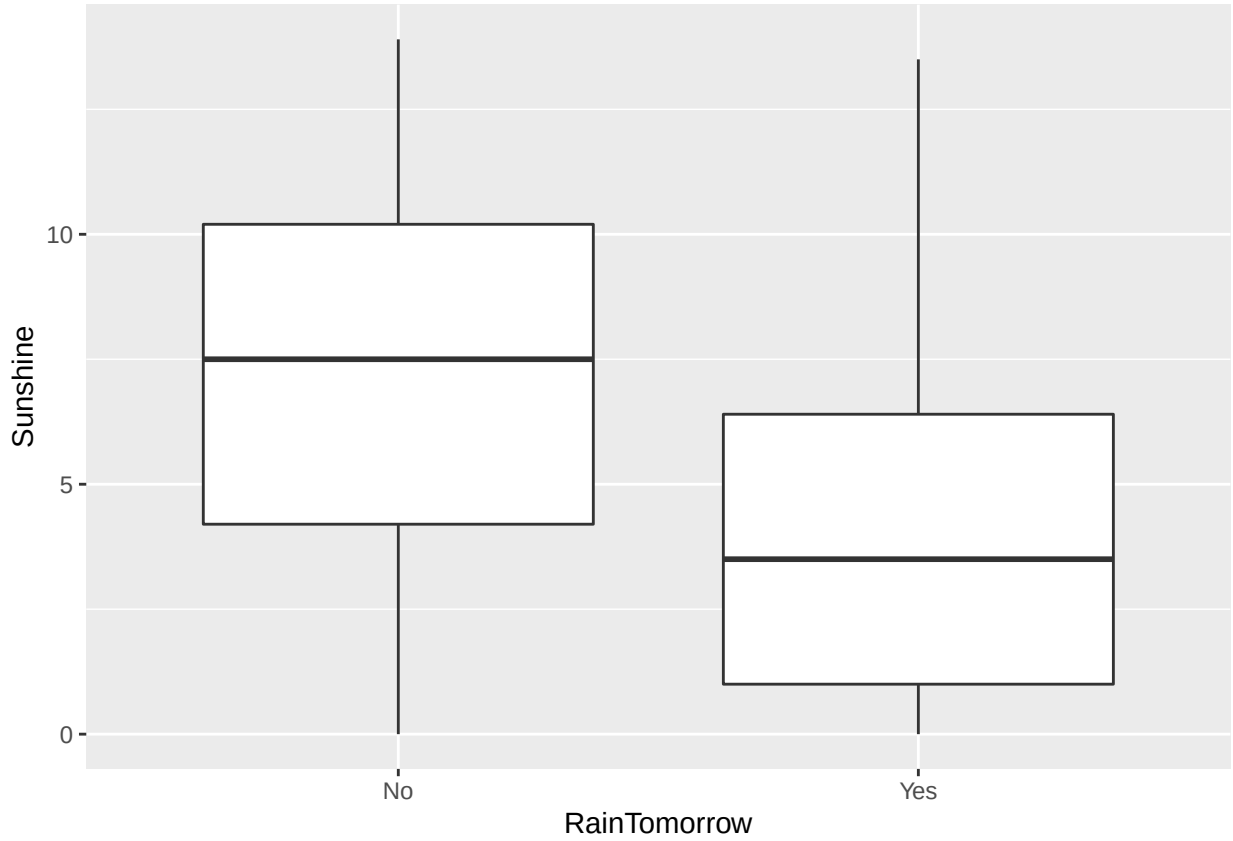

```
# Podem veure quines variables tenen diferències significatives respecte RainTomorrow
p <- ggplot(weatherMelb, aes(RainTomorrow,WindSpeed3pm))
p + geom_boxplot()
```




```
# Podem veure quines variables tenen diferències significatives respecte RainTomorrow
p <- ggplot(weatherMelb, aes(RainTomorrow,Evaporation))
p + geom_boxplot()
```




```
# Podem veure quines variables tenen diferències significatives respecte RainTomorrow
p <- ggplot(weatherMelb, aes(RainTomorrow,Sunshine))
p + geom_boxplot()
```



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.

```
# Suposem que hem escollit la variable
shapiro.test(weatherMelb$WindSpeed9am)
```

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

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(Rainfall ~ WindSpeed9am, data = weatherMelb)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Rainfall by WindSpeed9am
## Fligner-Killeen:med chi-squared = 106.03, df = 36, p-value = 7.973e-09
```

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

5. Representació dels resultats.

A partir de taules i gràfiques.

Resposta

6. Resolució del problema.

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

Resposta

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

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