TP Clasificación Estadística

Camilo D'Aloisio

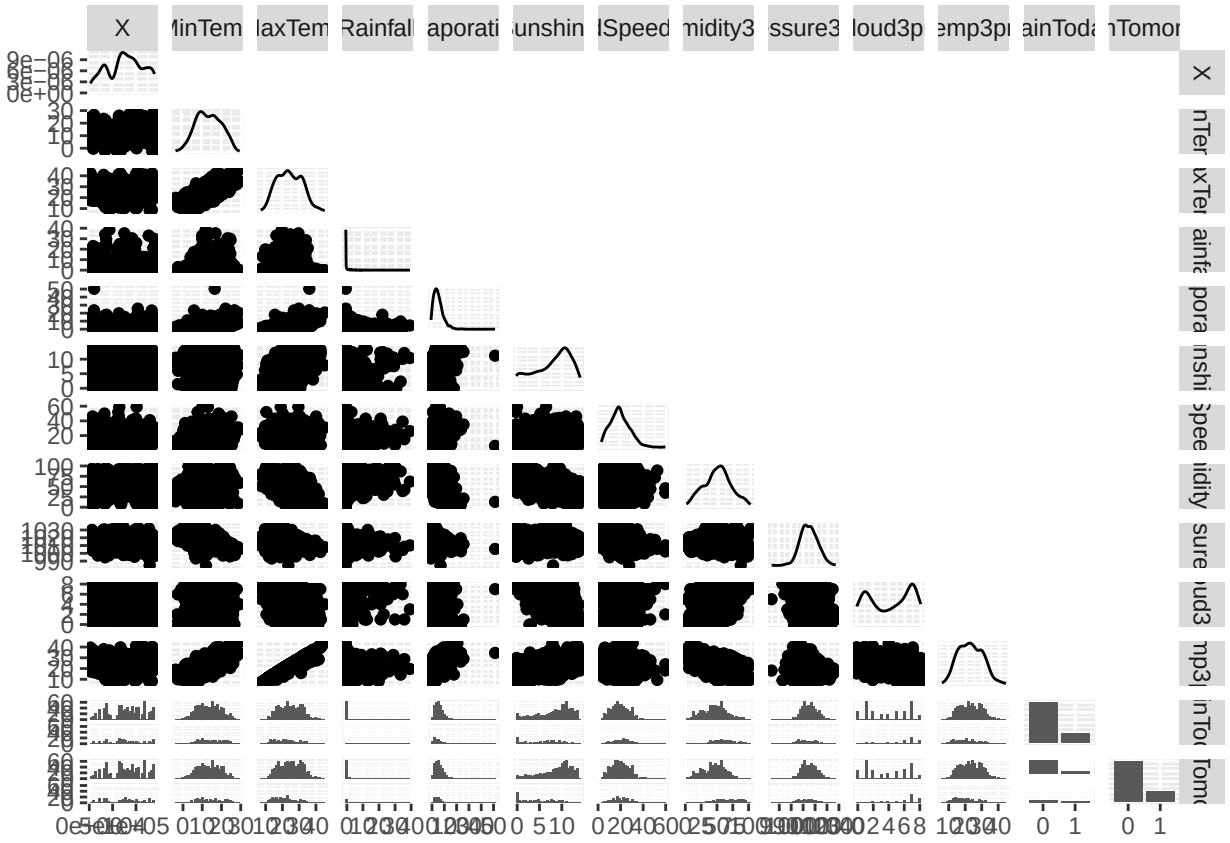
2022-12-08

Cargamos los datos y modificamos las variables RainTomorrow y RainToday para que queden como categóricas y con las opciones 1 (si llovió) y 0 (si no llovió). Las otras variables son numéricas.

```
current_directory <- getwd()
setwd(current_directory)

datos <- read.csv("lluviaAus.csv")

datos$RainTomorrow <-  ifelse(datos$RainTomorrow=="Yes", 1, 0)
datos$RainTomorrow <- as.factor(datos$RainTomorrow)
datos$RainToday <-  ifelse(datos$RainToday=="Yes", 1, 0)
datos$RainToday <- as.factor(datos$RainToday)
```

Ejercicio 1

```
library(ggplot2)
library(GGally)

#ggpairs(datos)
ggpairs(datos, upper = "blank")
```

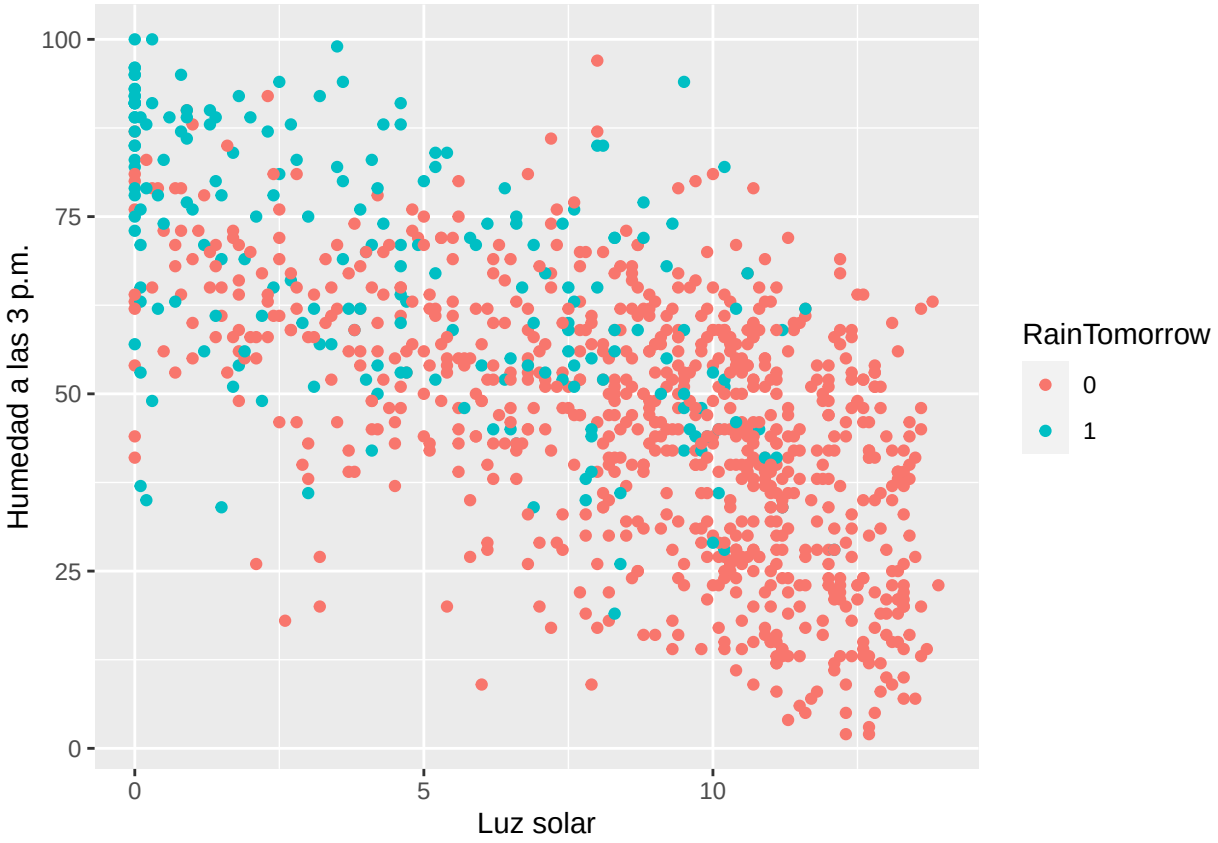



```
#como usamos la funcion con tantas variables?
```

Las varaibles temperatura maxima, temperatura minima y temperatura a las 3 de la tarde parecen estar bastante correlacionadas.

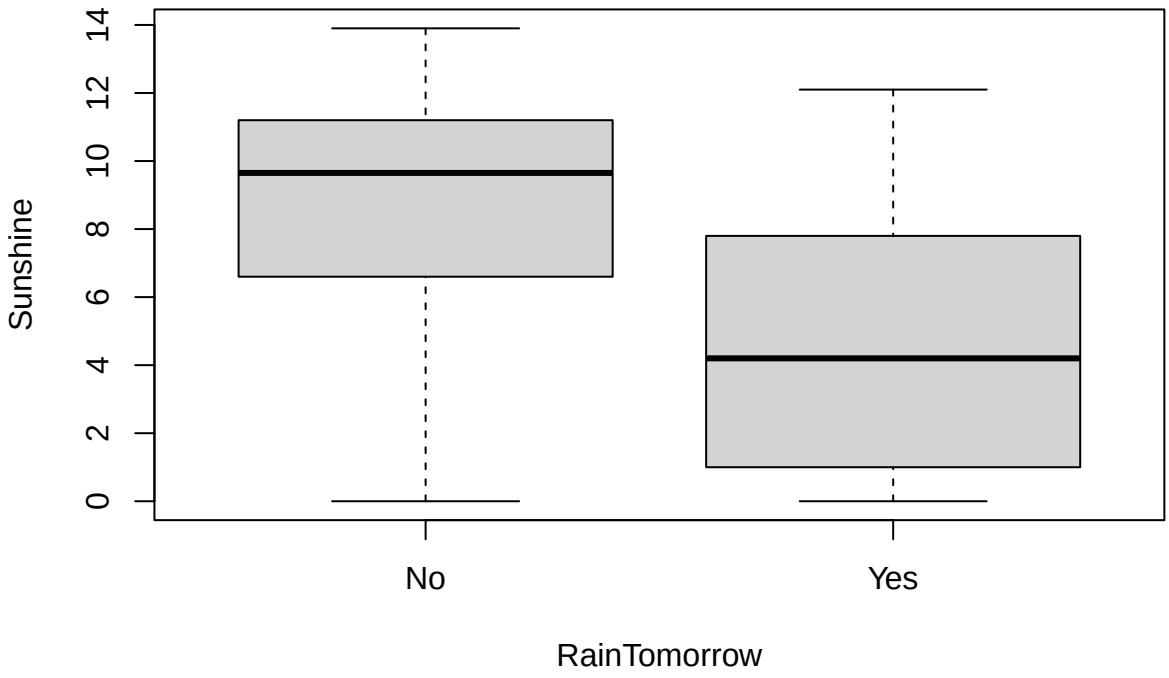
Ejercicio 2

```
datos %>% ggplot(mapping=aes(x=Sunshine, y=Humidity3pm, color=RainTomorrow)) +
  geom_point() +
  labs(x="Luz solar", y="Humedad a las 3 p.m.")
```



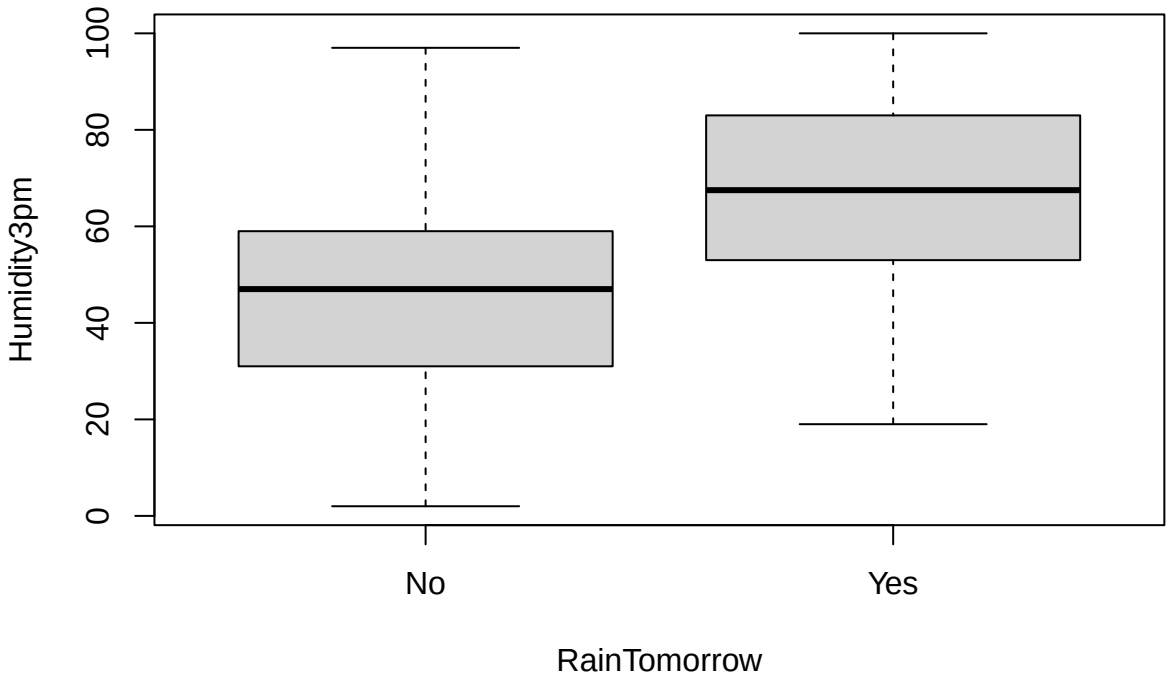
Ejercicio 3

```
boxplot(
  datos$Sunshine[datos$RainTomorrow==0],
  datos$Sunshine[datos$RainTomorrow==1],
  names=c("No", "Yes"),
  xlab="RainTomorrow",
  ylab="Sunshine"
)
```



Observamos que, en lineas generales, cuando no llueve hay más luz solar que cuando llueve ya que todos los cuantiles son mayores. Sin embargo, el bigote inferior se encuentra al mismo nivel (en 0) por lo que puede haber días que no llueva donde haya poca luz solar.

```
boxplot(
  datos$Humidity3pm[datos$RainTomorrow==0],
  datos$Humidity3pm[datos$RainTomorrow==1],
  names=c("No", "Yes"),
  xlab="RainTomorrow",
  ylab="Humidity3pm"
)
```



Acá, al contrario, observamos que los días que llueve suele haber más humedad a las 3 de la tarde que los días que no llueve.

Ejercicio 4

```
clasificador.movil <-  function(datos, etiquetas, h, x0) {
  #etiquetas_en_ventana <-  etiquetas[norm(datos - x0, type="2") <= h]
  etiquetas_en_ventana <-  etiquetas[abs(datos - x0) <= h]
  res <- ifelse(sum(etiquetas_en_ventana==1) / length(etiquetas_en_ventana) >= 0.5, 1, 0)
  return(res)
}

#solo funciona para x y x0 en R. Para Rn (con norma dos) sería:

clasificador.movil2 <- function(datos, etiquetas, h, x0) {
  etiquetas_en_ventana <-  etiquetas[sqrt(sum((datos - x0)**2)) <= h]
  res <- ifelse((sum(etiquetas_en_ventana) / length(etiquetas_en_ventana)) >= 0.5, 1, 0)
  return(res)
}

# Nota:
# podemos reemplazar todo por una sola funcion haciendo:
# etiquetas_en_ventana <-  etiquetas[norm(datos - x0, type="2") <= h]
# ejemplo:
```

```r
# norm(-2, type="2") -> 2
# norm(c(3,4), type="2") -> 5
```

Ejercicio 5

```r
ventana_optima <-  function(datos, etiquetas, ventanas) {
  ECM_ventanas <-  c()
  etiquetas <- as.numeric(etiquetas==1)
  for (h in 1:length(ventanas)) {
    clasificaciones <-  c()
    for (j in 1:length(etiquetas)) {
      clasificaciones[j] <-  clasificador.movil(datos[-j], etiquetas[-j], ventanas[h], datos[j])
    }
    ECM_ventanas[h] <-  mean((etiquetas - clasificaciones)^2)
  }
  index_res <- order(ECM_ventanas)[1]
  res <-  ventanas[index_res]
  return(res)
}
```

Ejercicio 6

Primero defimos la funcion

```r
error_de_clasificacion <-  function(datos, etiquetas, ventana) {
  errores <-  c()
  etiquetas <- as.numeric(etiquetas==1)
  for (i in 1:length(etiquetas)) {
    clasificacion_i <-  clasificador.movil(datos[-i], etiquetas[-i], ventana, datos[i])
    error <-  abs(etiquetas[i] - clasificacion_i)
    errores[i] <-  error
  }
  return(mean(errores))
}
```

Ahora vamos a evaluarla

```r
ventana <- ventana_optima(datos$Sunshine, datos$RainTomorrow, seq(0.01, 2, length.out=200))

error_de_clasificacion(datos$Sunshine, datos$RainTomorrow, ventana)
```

```
## [1] 0.178
```

Luego el error es 0.178