

Ejercicio 2.3: Detección y tratamiento de datos atípicos multivariante

Silvia Pineda

Lectura Fichero de datos y carga de librerías

```
library(dbscan)
```

```
Warning: package 'dbscan' was built under R version 4.5.2
```

```
Attaching package: 'dbscan'
```

```
The following object is masked from 'package:stats':
```

```
as.dendrogram
```

```
library(class)
library(ggplot2)
library(patchwork)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
vforcats   1.0.1     v stringr   1.6.0
v lubridate 1.9.4     v tibble    3.3.0
v purrr    1.2.0     v tidyverse 1.3.1
```

```
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become non-conflicting
```

```
data <- read.csv("ozone.csv") # import data
data$Month<-as.factor(data$Month)
data$Day_of_month<-as.factor(data$Day_of_month)
data$Day_of_week<-as.factor(data$Day_of_week)
```

Estudio Multivariante

```
#####Aplicamos LOF
k<-round(log(nrow(data)))
datos_lof<-scale(select(data,-Month,-Day_of_month,-Day_of_week,-Inversion_base_height))
lof<-lof(datos_lof,minPts = k)

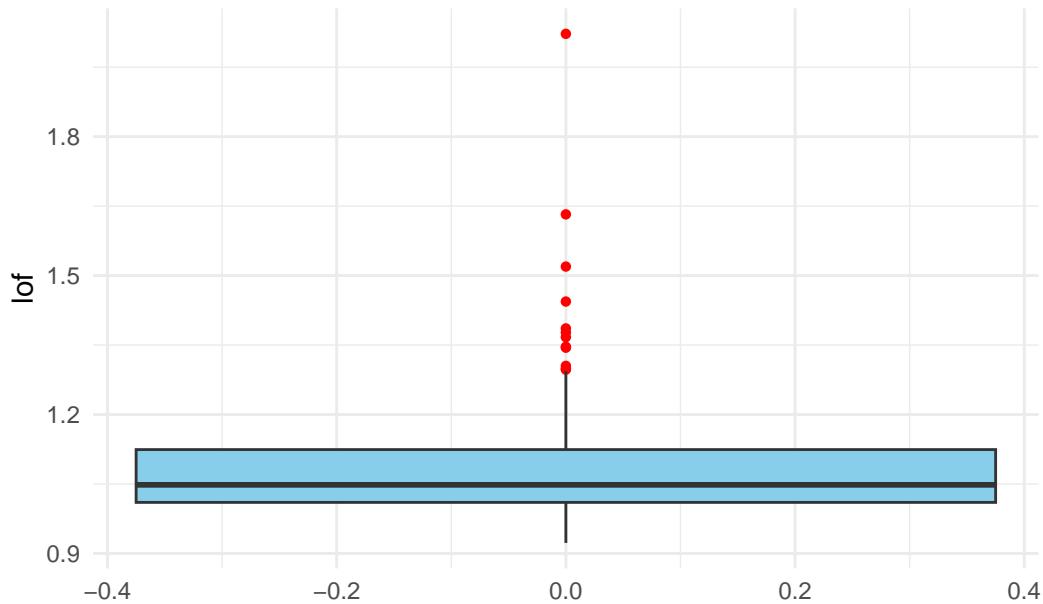
data$lof<-lof

#Imprimimos los que son >1.5
data[which(data$lof>1.5),]
```

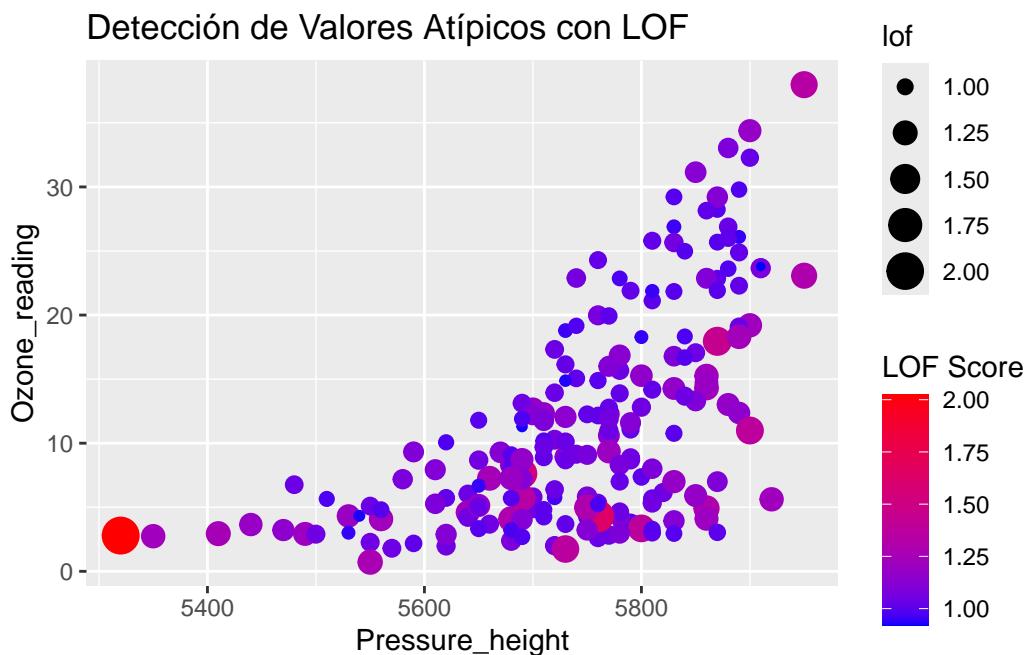
	Month	Day_of_month	Day_of_week	Ozone_reading	Pressure_height	Wind_speed
37	3	3	3	2.79	5320	16
43	3	12	5	7.63	5690	0
188	12	8	3	4.31	5760	0
	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height		
37	45	25	27.68	NA		
43	60	49	46.04	613		
188	32	62	56.12	826		
	Pressure_gradient	Inversion_temperature	Visibility	lof		
37	39	27.50	200	2.021929		
43	-27	59.72	300	1.519404		
188	-16	64.76	300	1.632160		

```
ggplot(data, aes(y = lof)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16) +
  theme_minimal() +
  labs(title = "Distribución de LOF Scores")
```

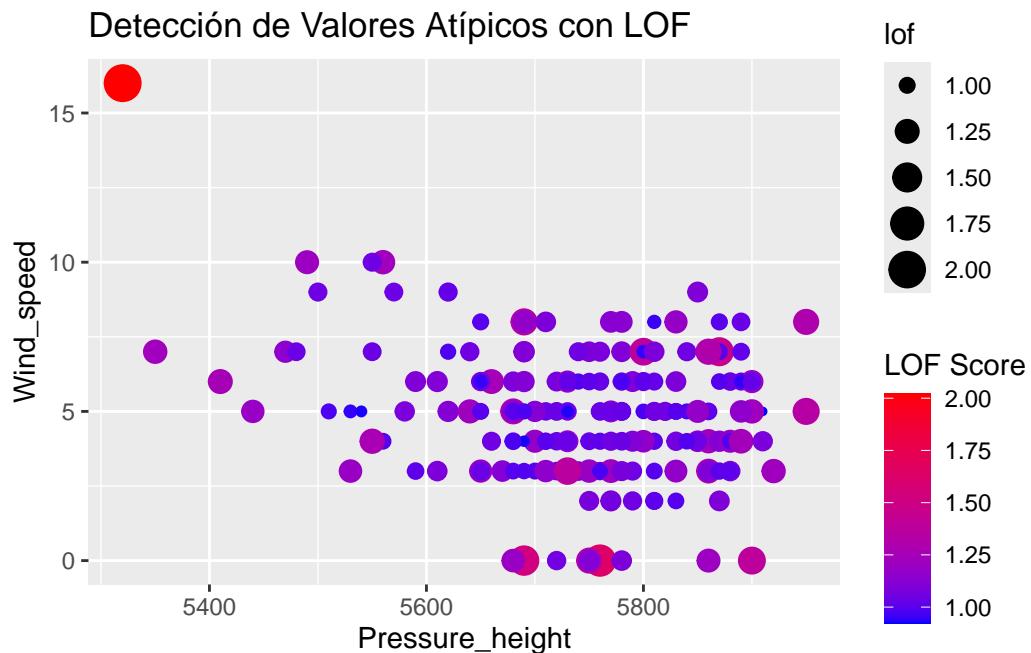
Distribución de LOF Scores



```
#####Comprobamos las cuantitativas
ggplot(data, aes(x = Pressure_height, y = Ozone_reading, colour = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```

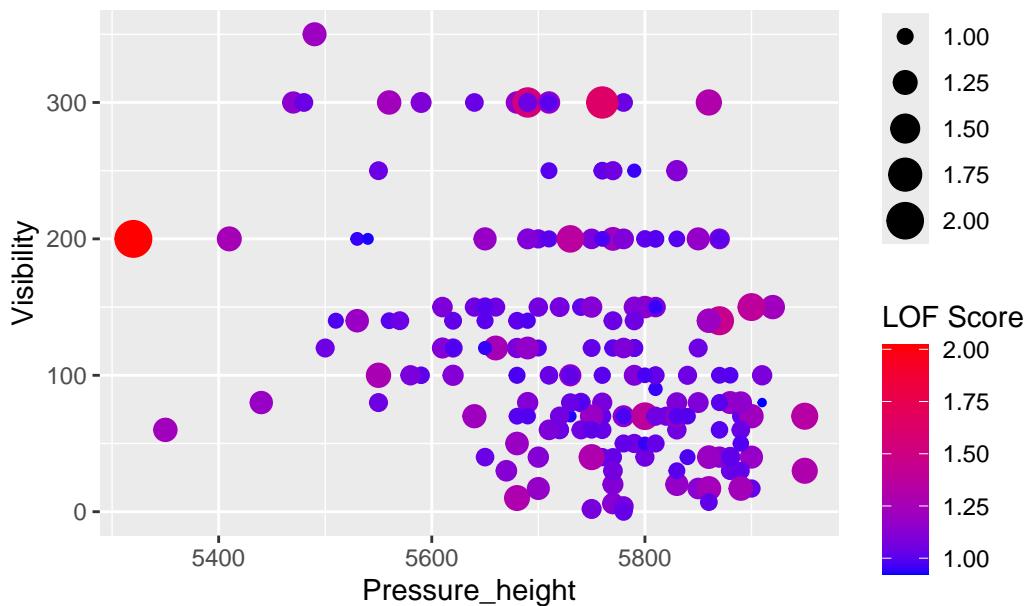


```
ggplot(data, aes(x = Pressure_height, y = Wind_speed, colour = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



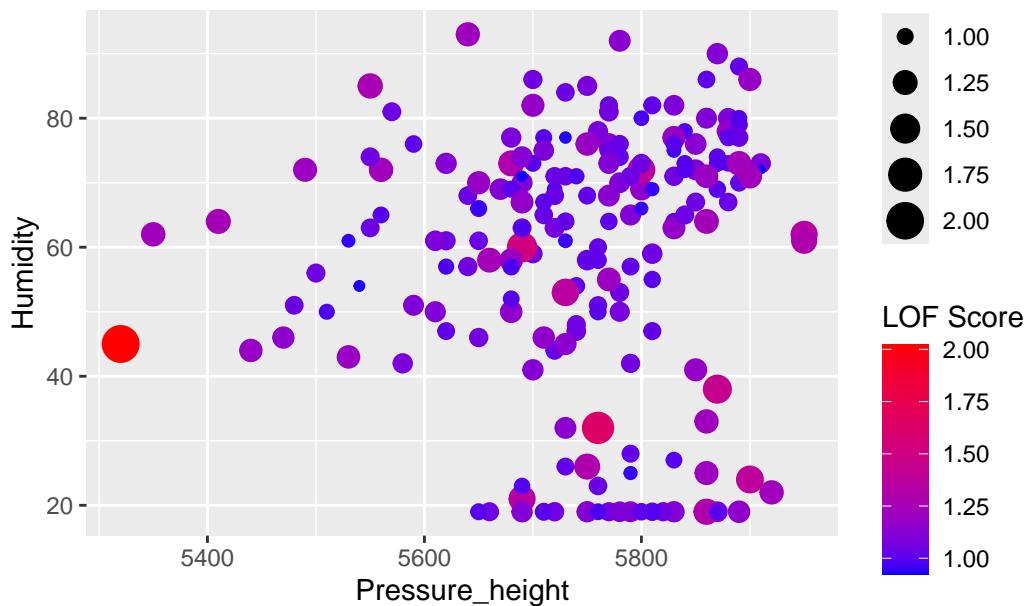
```
ggplot(data, aes(x = Pressure_height, y = Visibility, colour = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```

Detección de Valores Atípicos con LOF

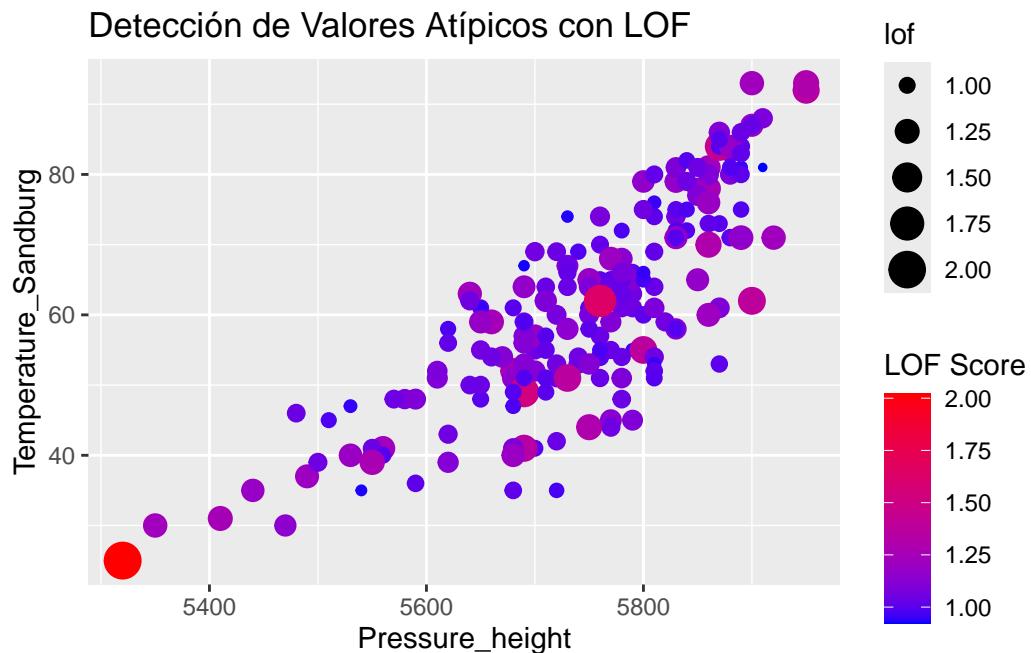


```
ggplot(data, aes(x = Pressure_height, y = Humidity, colour = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```

Detección de Valores Atípicos con LOF

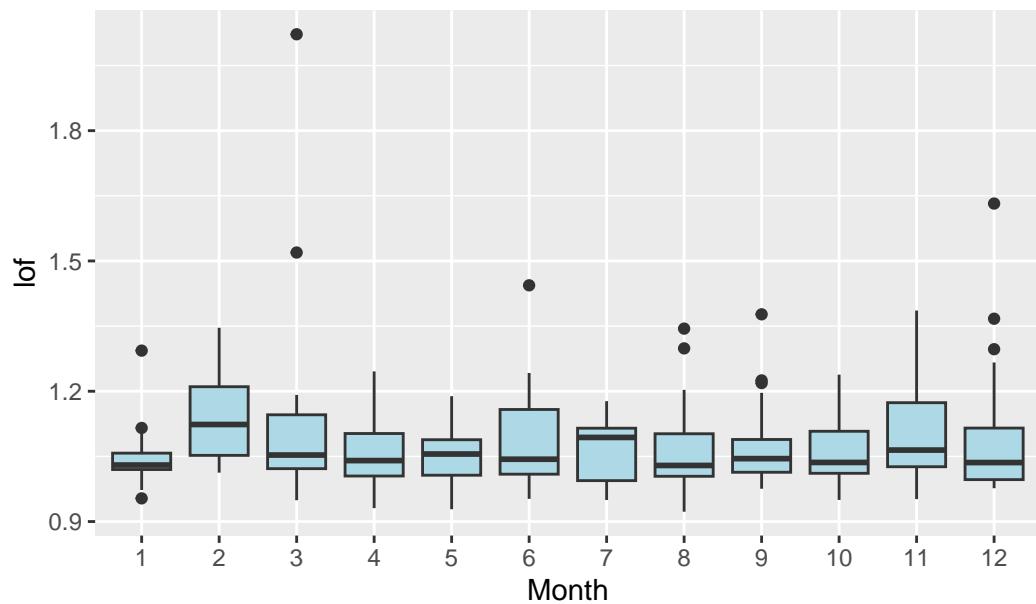


```
ggplot(data, aes(x = Pressure_height, y = Temperature_Sandburg, colour = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



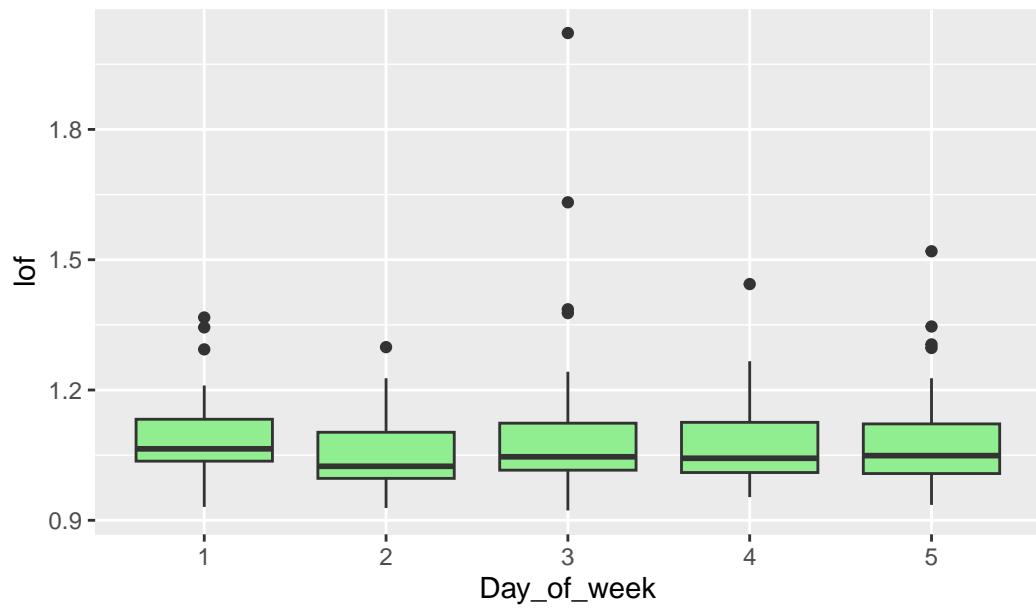
```
####Comprobamos las cualitativas
ggplot(data, aes(x = as.factor(Month), y = lof)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "lof across months", x = "Month", y = "lof")
```

lof across months



```
ggplot(data, aes(x = as.factor(Day_of_week), y = lof)) +  
  geom_boxplot(fill = "lightgreen") +  
  labs(title = "lof across day of week", x = "Day_of_week", y = "lof")
```

lof across day of week



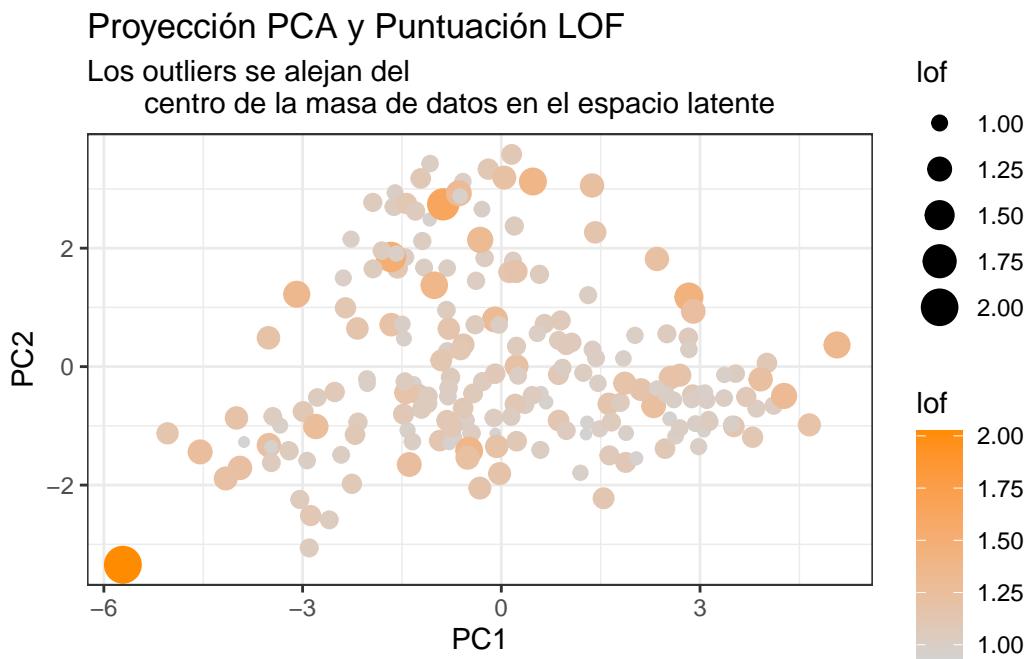
```

# Ejecutamos un PCA rápido sobre las variables numéricas
pca_res <- prcomp(scale(select(data, -Month, -Day_of_month,
                                -Day_of_week, -Inversion_base_height, -lof)), center = TRUE)

# Creamos un dataframe para graficar
df_pca <- as.data.frame(pca_res$x)
df_pca$lof <- data$lof

ggplot(df_pca, aes(x = PC1, y = PC2, color = lof)) +
  geom_point(aes(size = lof)) +
  scale_color_gradient(low = "lightgrey", high = "darkorange") +
  labs(title = "Proyección PCA y Puntuación LOF",
       subtitle = "Los outliers se alejan del
                  centro de la masa de datos en el espacio latente") +
  theme_bw()

```



Tras aplicar el algoritmo LOF vemos que hay 3 observaciones con $\text{LOF} > 1.5$ pero solo una. que tiene un $\text{LOF} = 3.08$ y que sale de la distribución. Al representar las variables dos a dos coloreadas por el LOF, vemos como hay un punto que corresponde a varios de los Outliers que hemos detectado en el estudio uni y bivariante. Vemos que esa observación corresponde al extremos de Wind_Speed y que además también es un outlier de Pressure_height. Vemos que siempre está alejada de la nube de puntos y que en el PCA además también está completamente fuera.

En base a esto podría tener mucho sentido borrar la observación entera.