

Ejercicio 2.4: Detección y tratamiento de datos atípicos multivariante

Silvia Pineda

Lectura Fichero de datos y carga de librerías

```
library(dbscan)
```

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

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

```
as.dendrogram
```

```
library(class)
library(ggplot2)
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 tidyr    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
```

```

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)))
lof<-lof(select(data,-Month,-Day_of_month,-Day_of_week,-Inversion_base_height),minPts = k)
cbind(data[lof>1.5,],lof[lof>1.5])

```

	Month	Day_of_month	Day_of_week	Ozone_reading	Pressure_height	Wind_speed
11	1	19	1	4.07	5680	5
47	3	18	4	12.67	5700	4
97	6	16	3	14.31	5860	3
130	8	30	1	37.98	5950	5
131	8	31	2	23.07	5950	8
152	10	7	4	18.31	5890	4
167	11	5	5	4.91	5860	7
197	12	21	2	3.33	5650	5
	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height		
11	73	52	56.48	393		
47	82	57	50.36	1571		
97	64	78	68.72	1279		
130	62	92	82.40	557		
131	61	93	81.68	620		
152	73	71	70.88	511		
167	19	70	62.78	NA		
197	19	48	47.12	NA		
	Pressure_gradient	Inversion_temperature	Visibility	lof[lof > 1.5]		
11	-68	69.80	10	1.999201		
47	68	56.30	17	1.506363		
97	75	71.60	17	1.505662		
130	0	90.68	70	1.807314		
131	27	85.64	30	1.666720		
152	-39	83.84	17	1.556773		
167	-29	61.70	300	1.502327		
197	-28	45.32	150	1.513810		

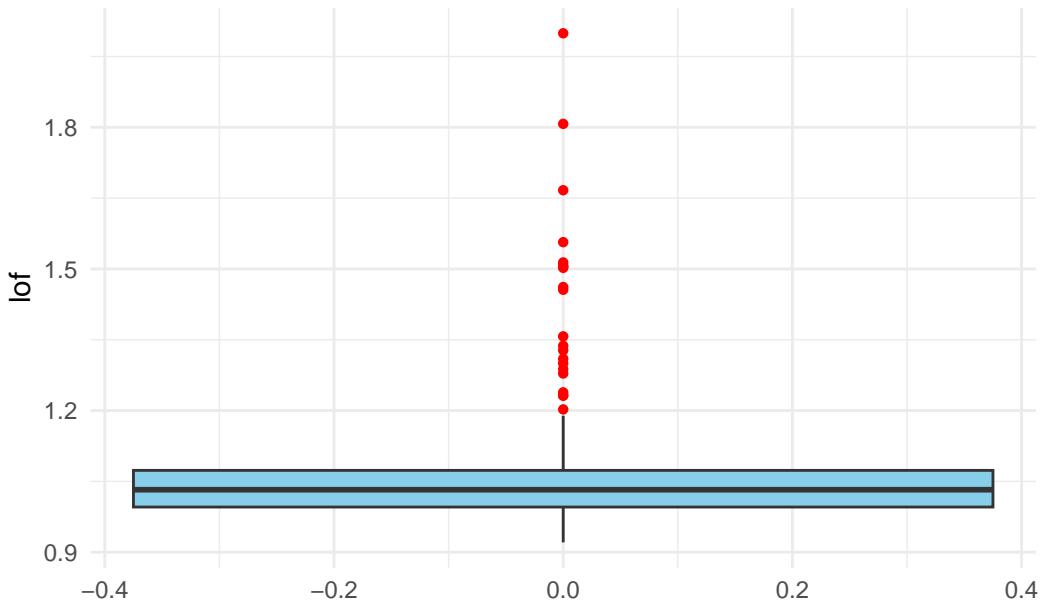
```

data$lof<-lof

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

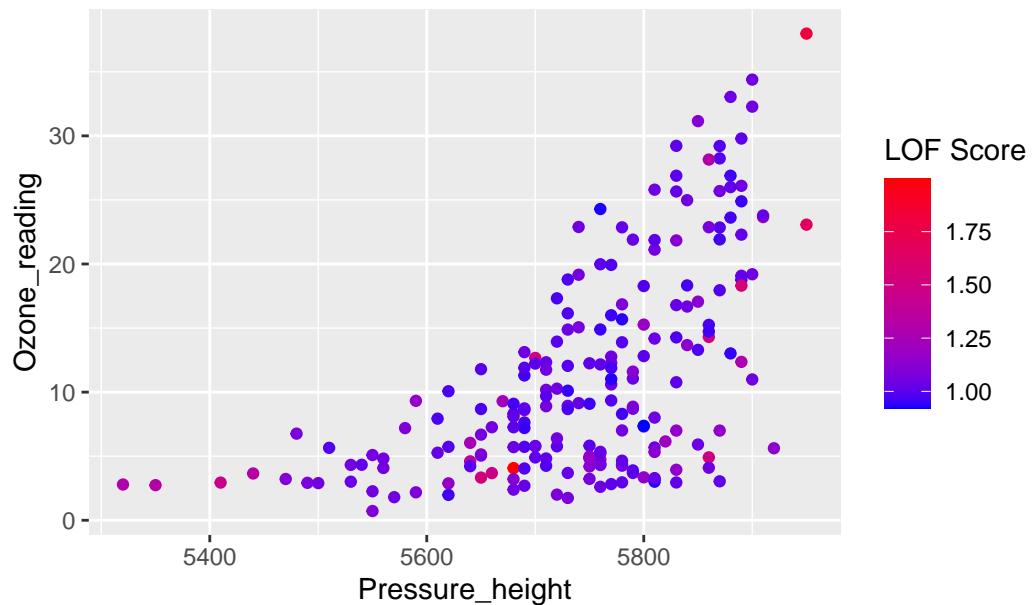


```

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

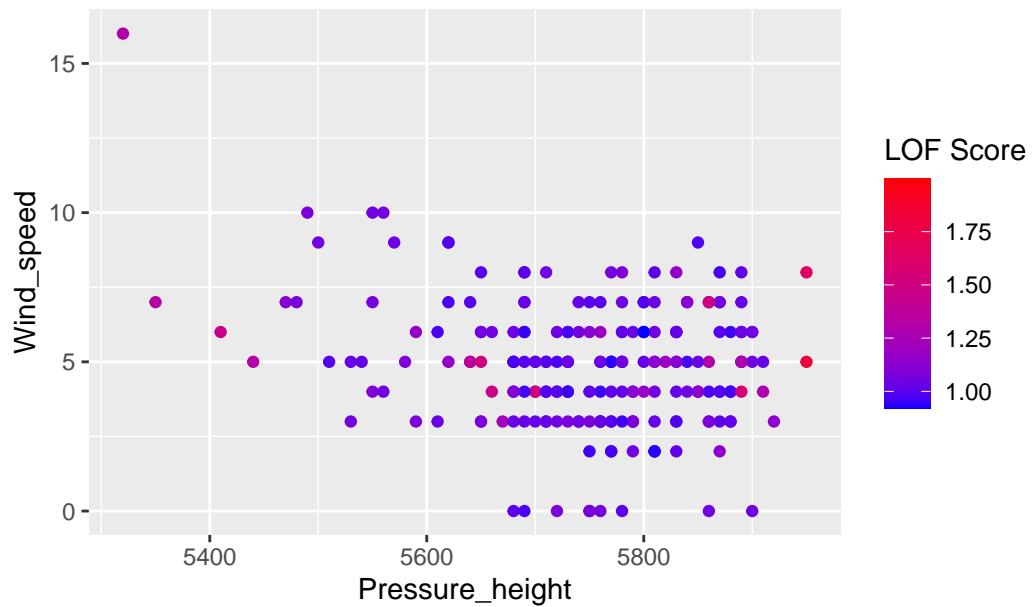
```

Detección de Valores Atípicos con LOF



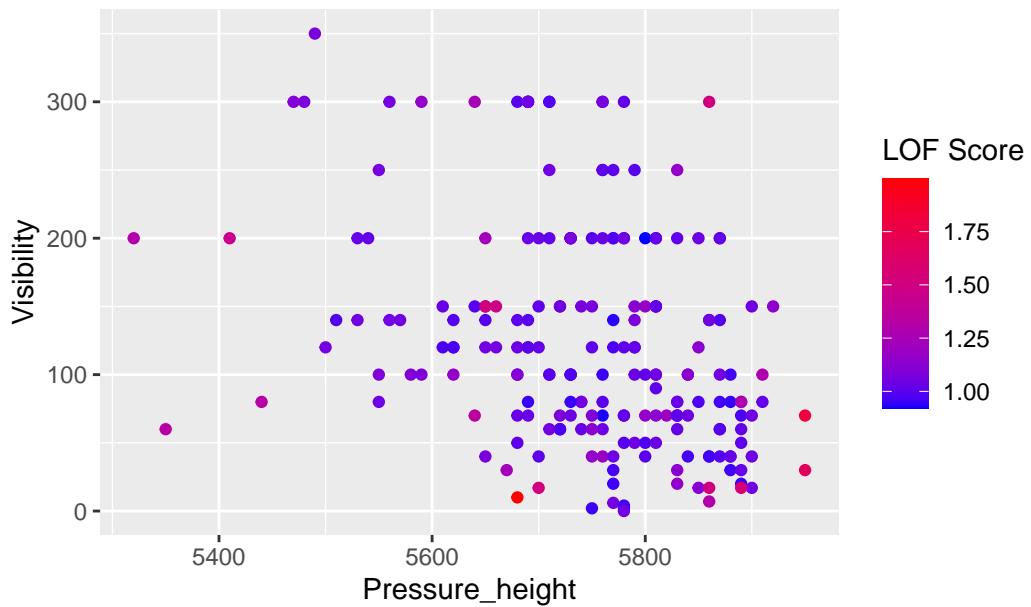
```
ggplot(data, aes(x = Pressure_height, y = Wind_speed, colour = lof)) +  
  geom_point() +  
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +  
  labs(title = "Detección de Valores Atípicos con LOF")
```

Detección de Valores Atípicos con LOF



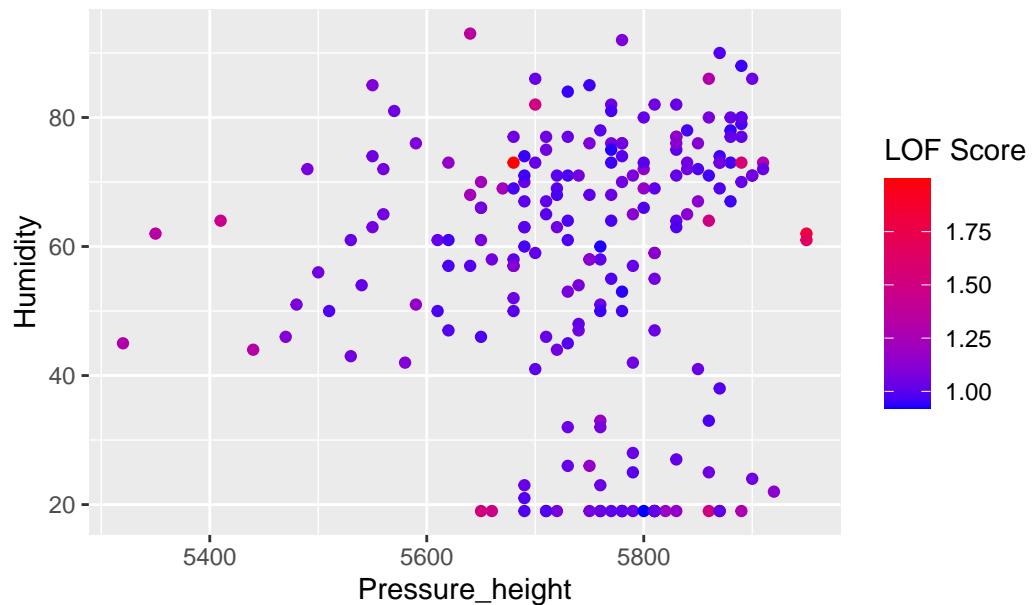
```
ggplot(data, aes(x = Pressure_height, y = Visibility, colour = lof)) +  
  geom_point() +  
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +  
  labs(title = "Detección de Valores Atípicos con LOF")
```

Detección de Valores Atípicos con LOF



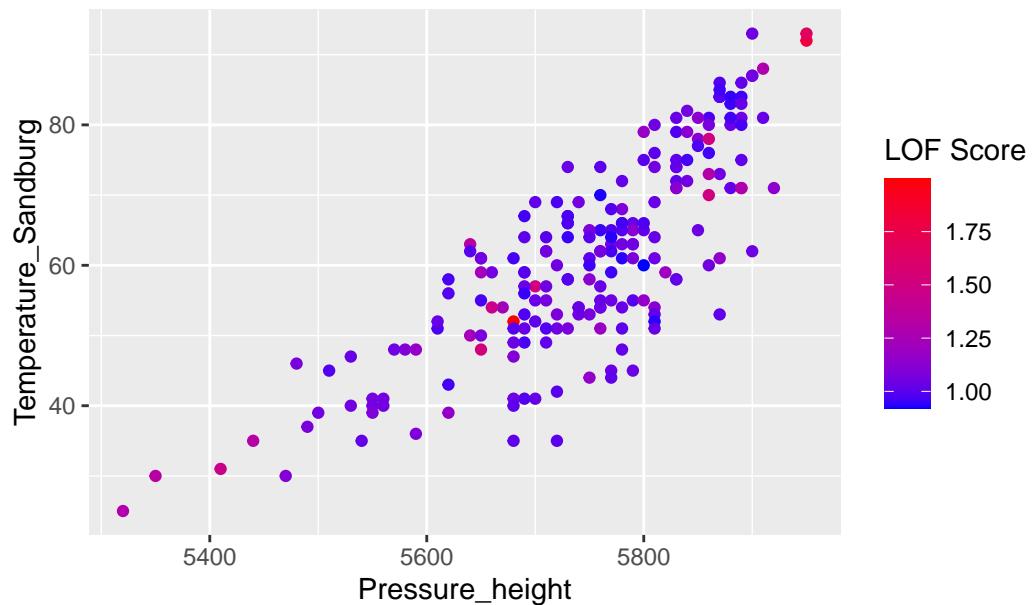
```
ggplot(data, aes(x = Pressure_height, y = Humidity, colour = lof)) +  
  geom_point() +  
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +  
  labs(title = "Detección de Valores Atípicos con LOF")
```

Detección de Valores Atípicos con LOF



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

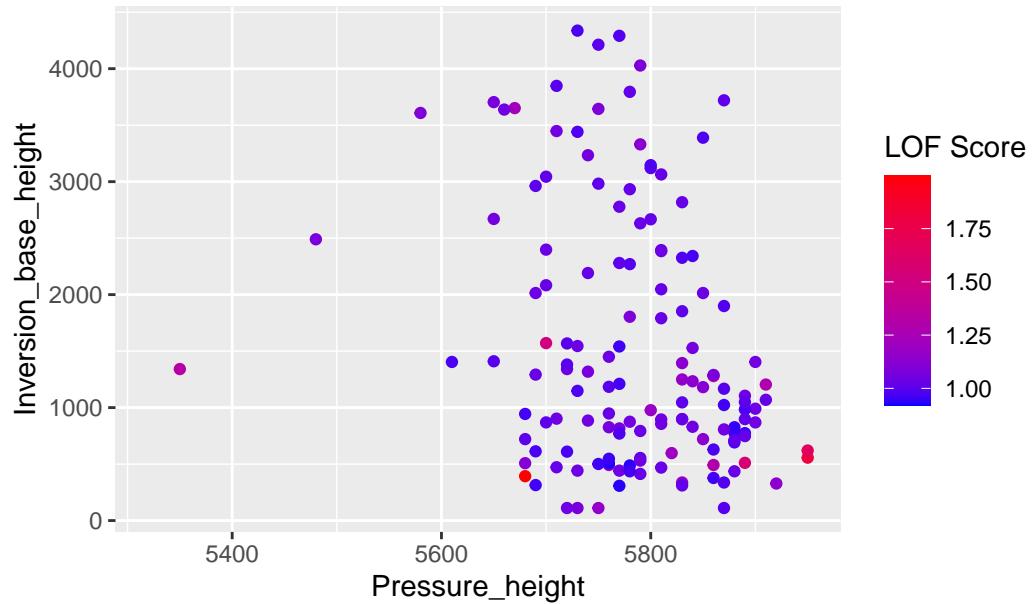
Detección de Valores Atípicos con LOF



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

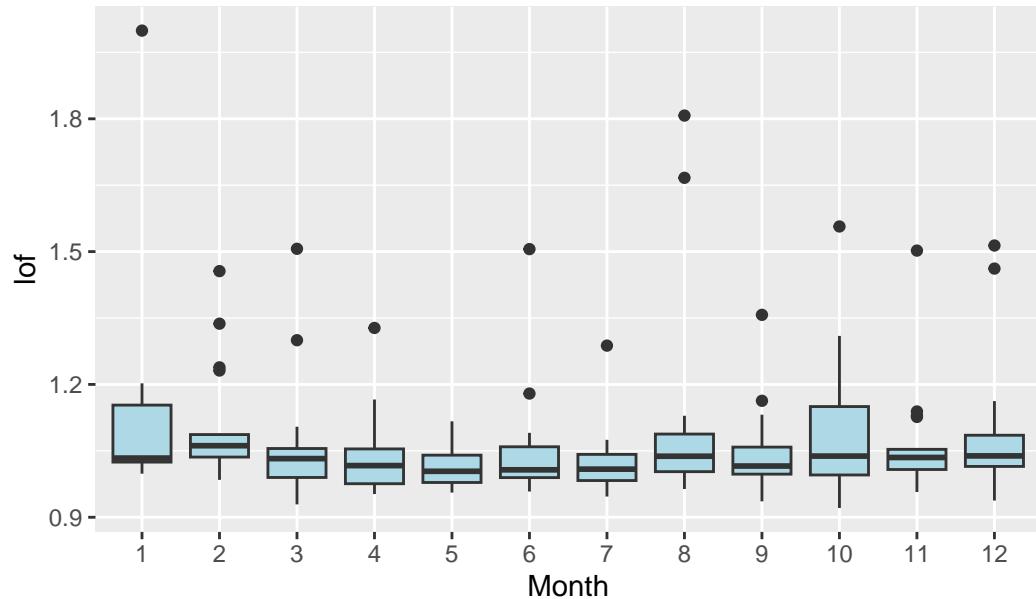
Warning: Removed 63 rows containing missing values or values outside the scale range
(`geom_point()`).

Detección de Valores Atípicos con LOF



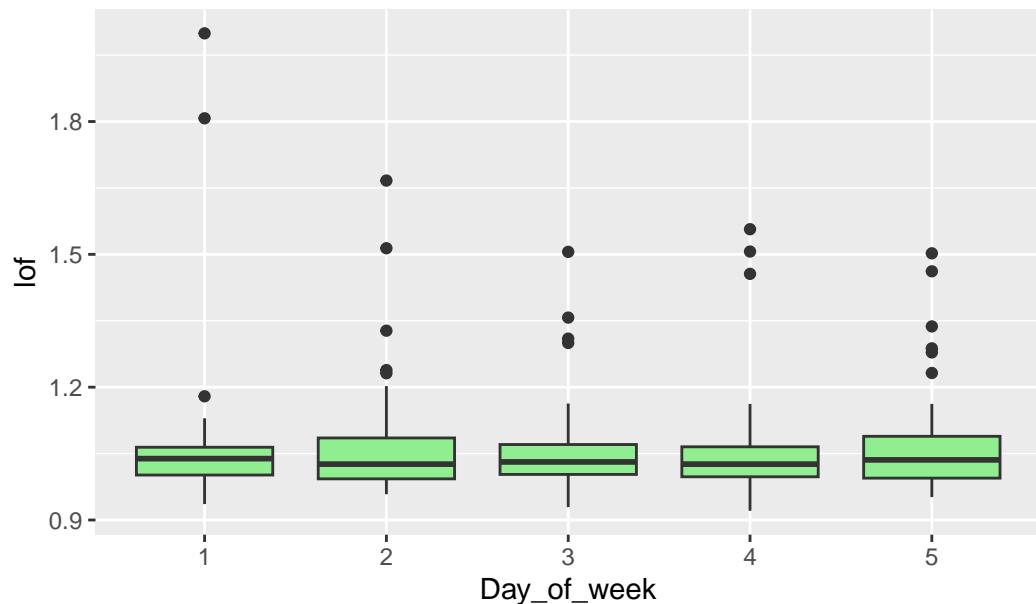
```
####Comprobamos las cualitativas
ggplot(data, aes(x = as.factor(Month), y = lof)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "lof across months", x = "Month", y = "lof")
```

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



Tras aplicar el algoritmo LOF vemos que hay 8 observaciones con $\text{LOF} > 1.5$ pero no pasan del 2. Al representar las variables dos a dos coloreadas por el LOF, los puntos con mayor LOF (en rojo) aparecen sistemáticamente en los bordes de las nubes de puntos, lo que podría sugerir que son atípicos por su combinación multivariante más que por una única variable extrema, pero no se observa ningún LOF especialmente grande y estos casos coinciden con atípicos detectados previamente que ya hemos visto que no eran atípicos porque eran parte de una asociación, por lo que no se eliminarán más observaciones.