Outliers Ozone

Silvia Pineda

Lectura Fichero de datos

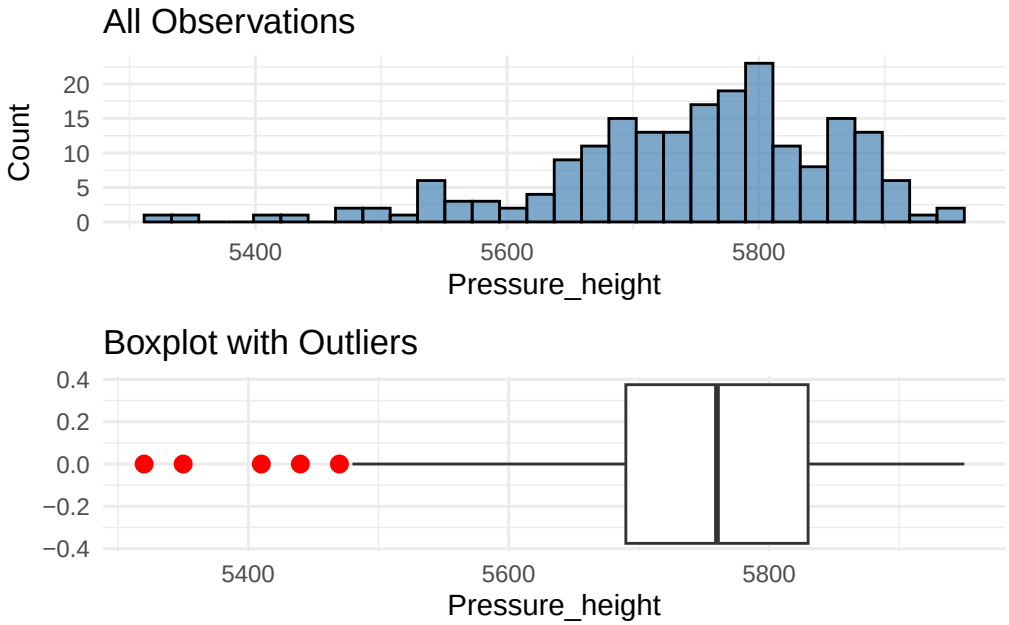
```
data <- read.csv("ozone.csv")  # import data
data$Month<-as.factor(data$Month)
data$Day_of_month<-as.factor(data$Day_of_month)
data$Day_of_week<-as.factor(data$Day_of_week)
```

Uso de la función outliers() y extreme()

```
source("outliers.R")
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

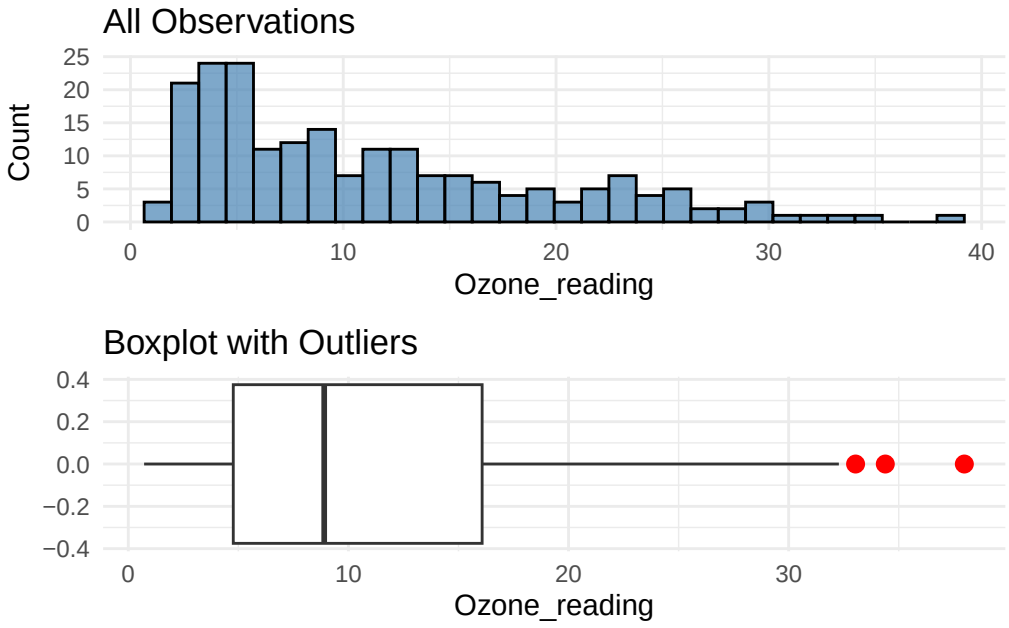
```
outliers(data,"Pressure_height")
```



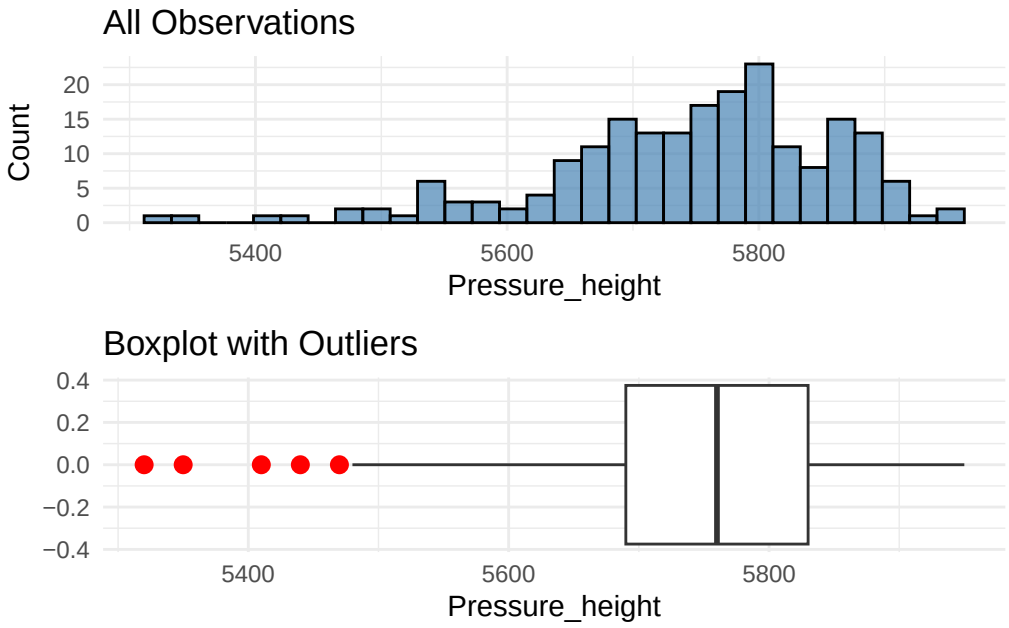

```
 Outliers identified in Pressure_height :  5 outliers
 Proportion (%) of outliers: 2.46 %
```

```
[1] 5410 5350 5470 5320 5440
```

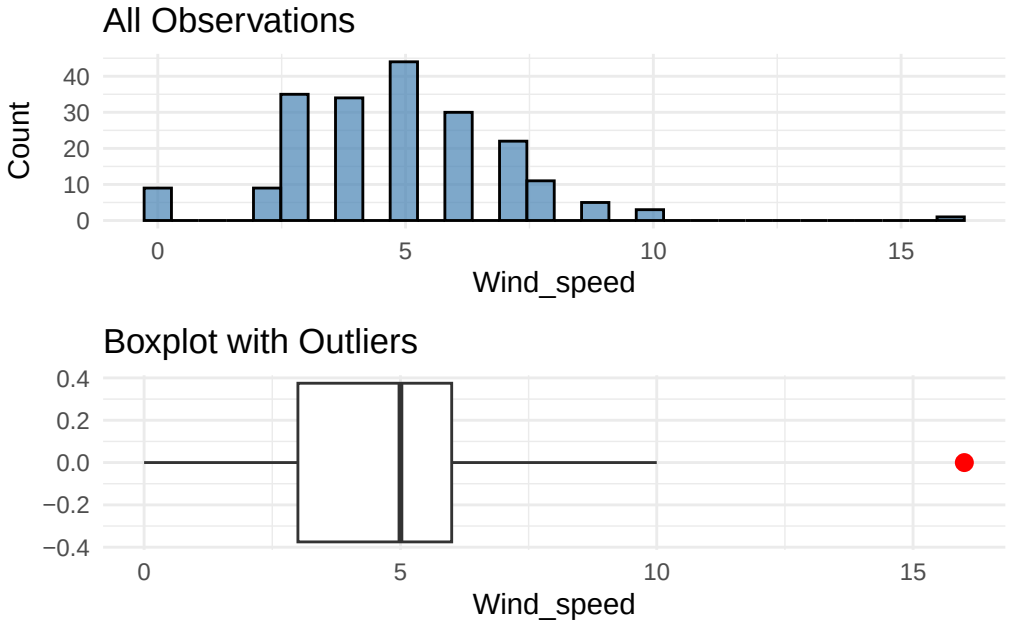
```r
# Aplicar la función a múltiples variables numéricas o enteras
numeric_integer_vars <- names(which(sapply(data, is.numeric) | sapply(data, is.integer)))
# Aplicar la función 'outliers' a cada una de las variables numéricas
outliers_results <- lapply(numeric_integer_vars, function(var) {
  outliers(data, var)  # Llamar a la función pasando el nombre de la variable
})
```



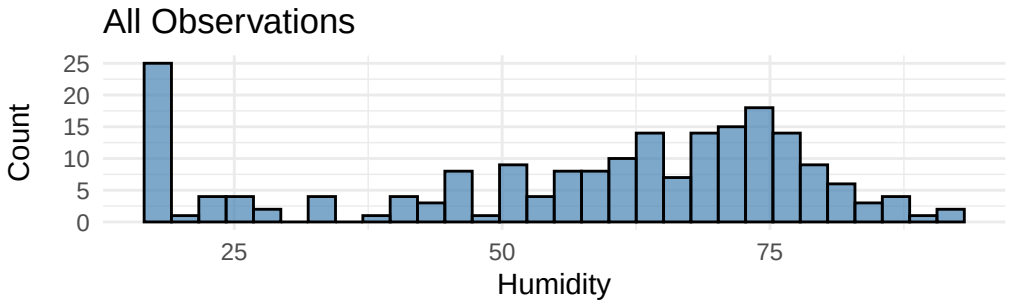

```
Outliers identified in Ozone_reading :  3 outliers
Proportion (%) of outliers: 1.48 %
```

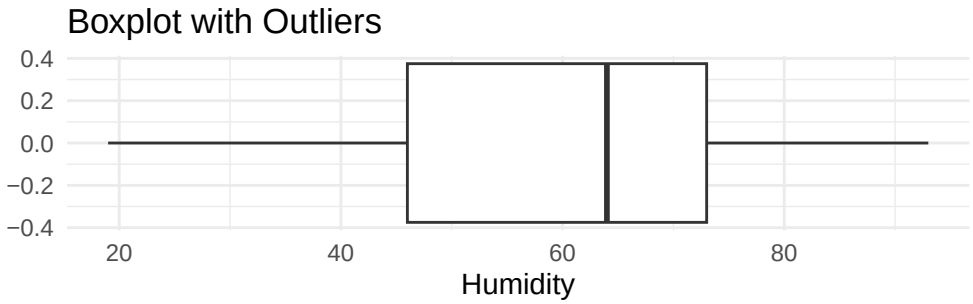


```
Outliers identified in Pressure_height :  5 outliers
Proportion (%) of outliers: 2.46 %
```

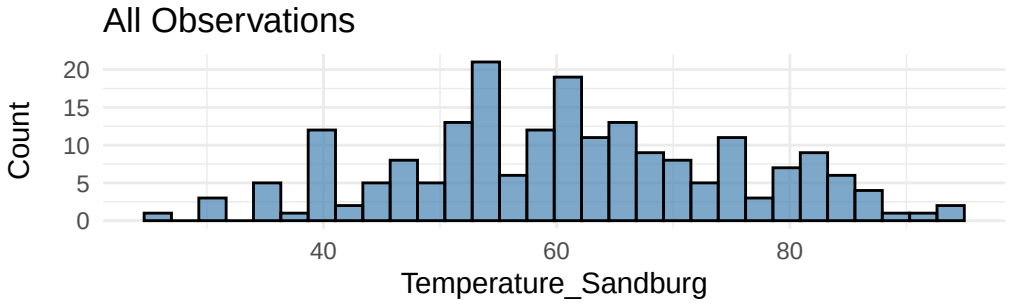


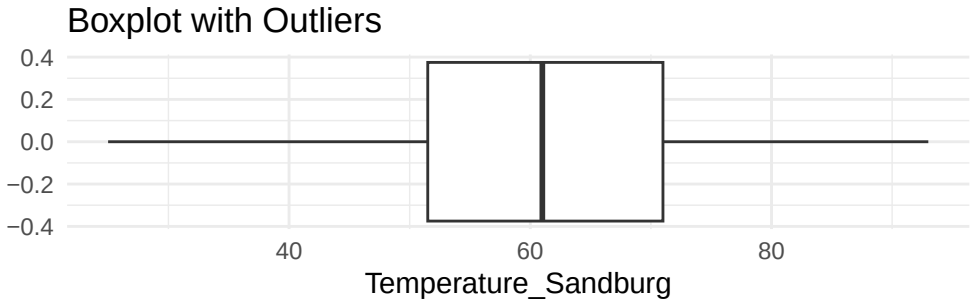

```
Outliers identified in Wind_speed :  1 outliers
Proportion (%) of outliers: 0.49 %
```



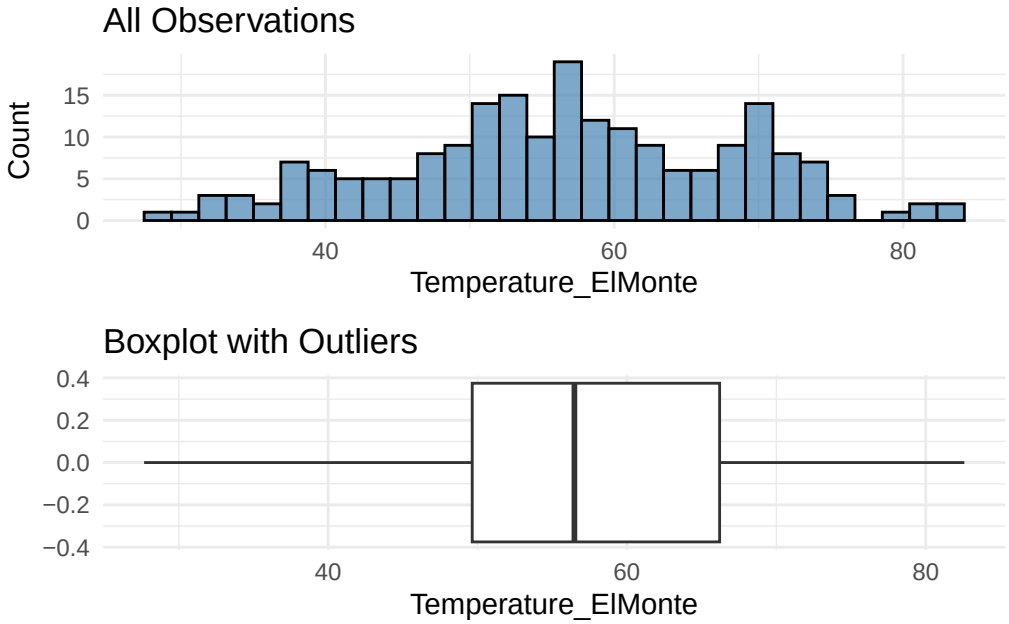



```
Outliers identified in Humidity :  0 outliers
Proportion (%) of outliers: 0 %
```





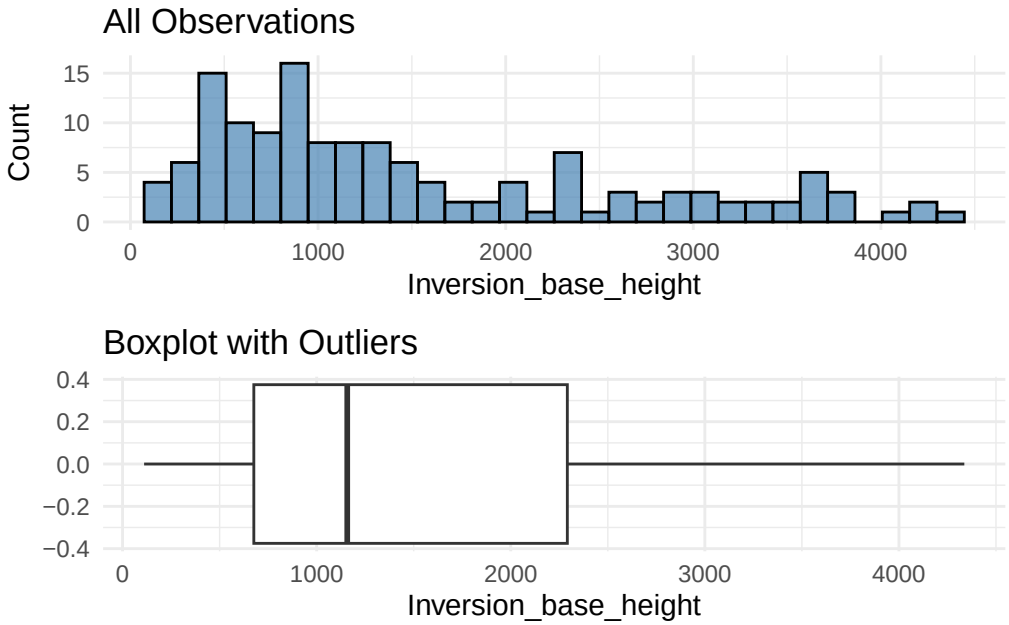
```
Outliers identified in Temperature_Sandburg :  0 outliers
Proportion (%) of outliers: 0 %
```



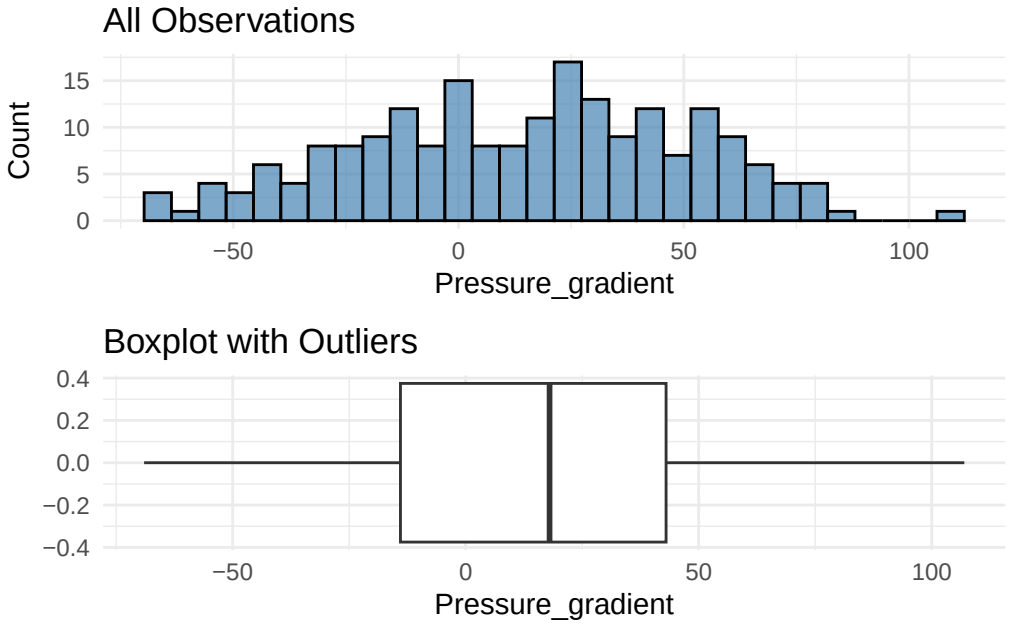

```
Outliers identified in Temperature_ElMonte :  0 outliers
Proportion (%) of outliers: 0 %
```

```
Warning: Removed 63 rows containing non-finite outside the scale range
(`stat_bin()`).
```

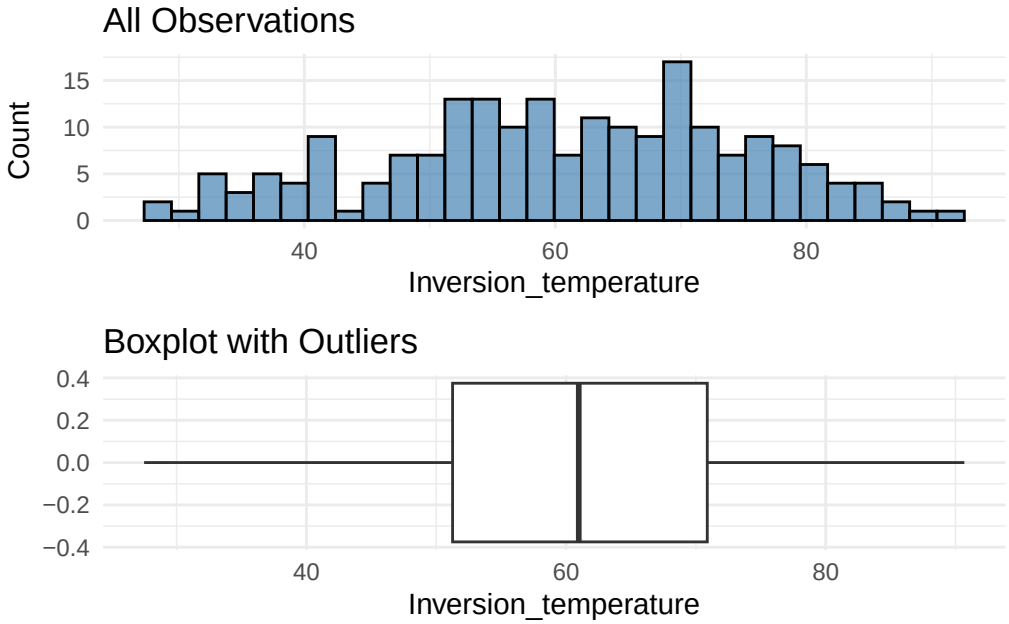
```
Warning: Removed 63 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



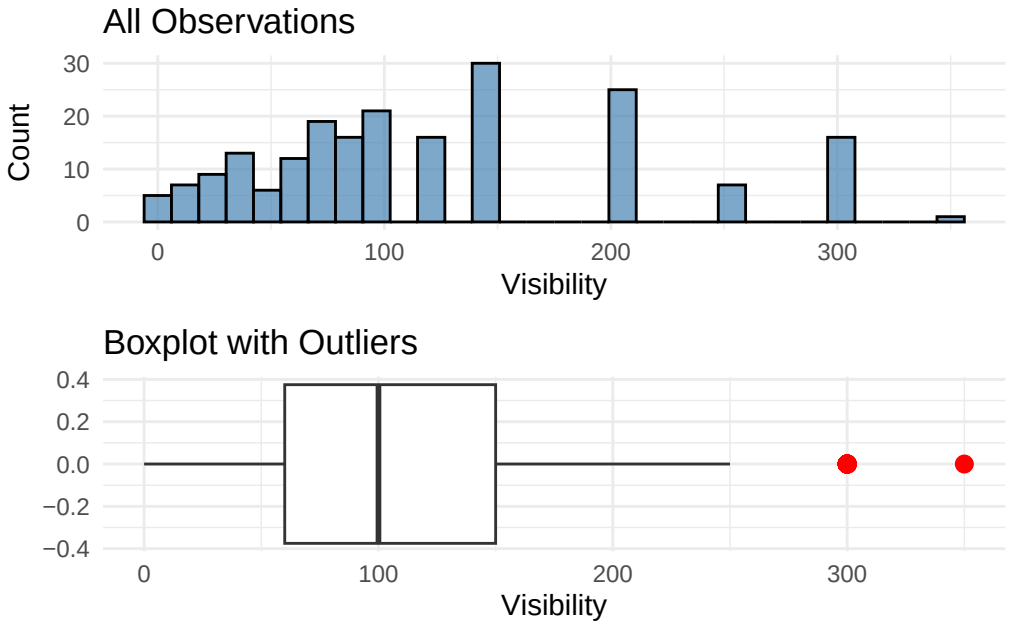

```
Outliers identified in Inversion_base_height :  0 outliers
Proportion (%) of outliers: 0 %
```



```
Outliers identified in Pressure_gradient :  0 outliers
Proportion (%) of outliers: 0 %
```

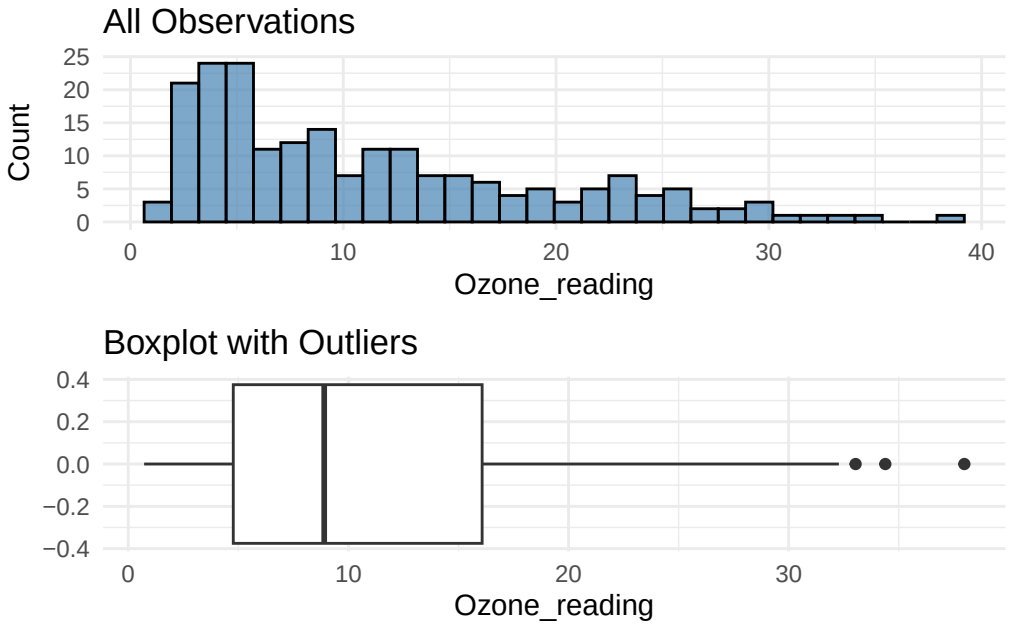



```
Outliers identified in Inversion_temperature :  0 outliers
Proportion (%) of outliers: 0 %
```

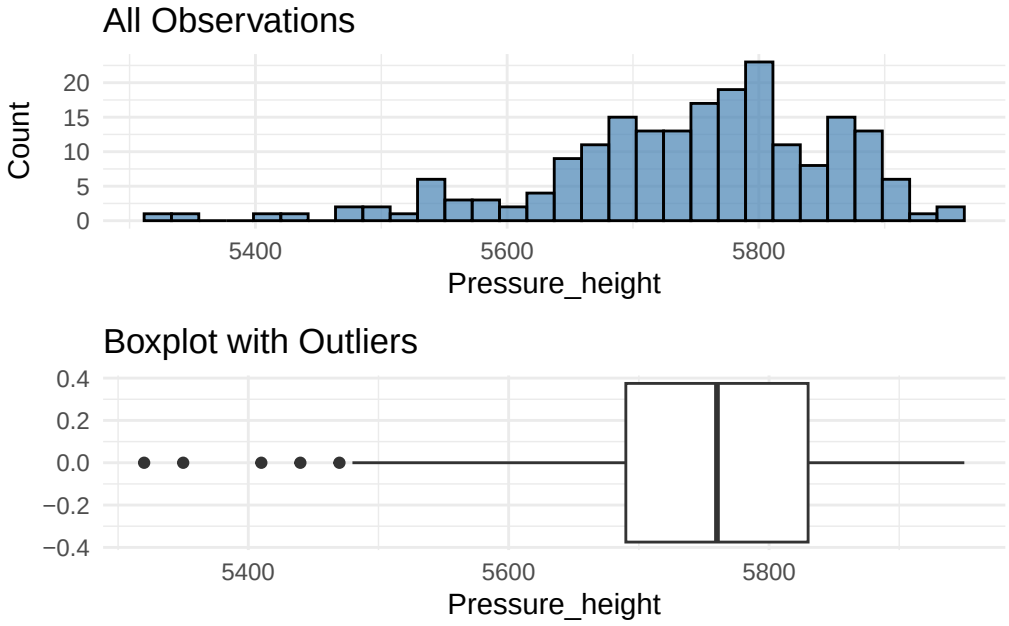



```
Outliers identified in Visibility :  17 outliers
Proportion (%) of outliers: 8.37 %
```

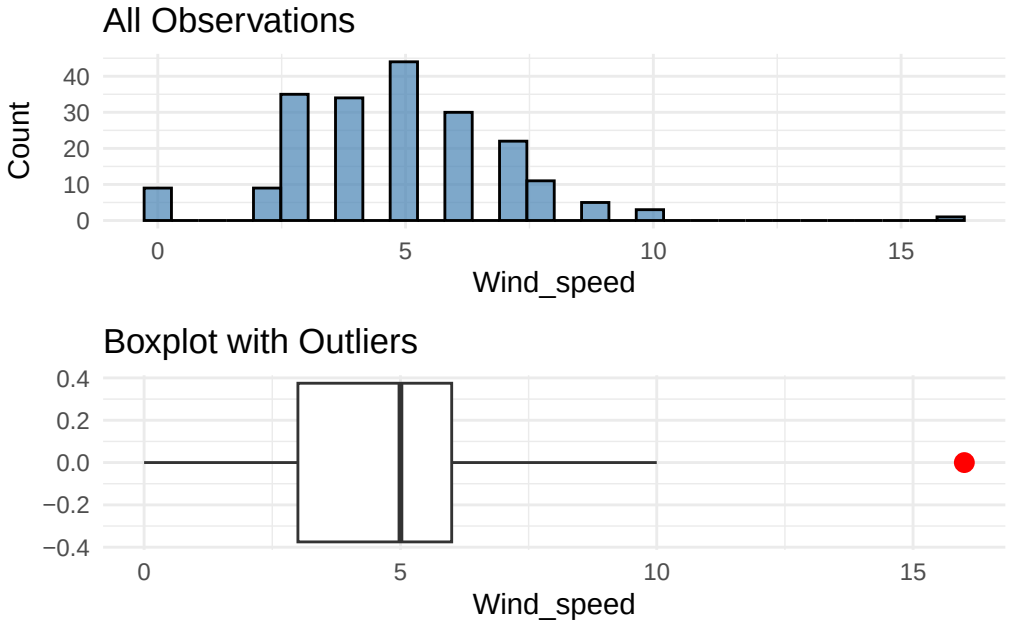
```r
extreme_results <- lapply(numeric_integer_vars, function(var) {
  extreme(data, var)  # Llamar a la función pasando el nombre de la variable
})
```



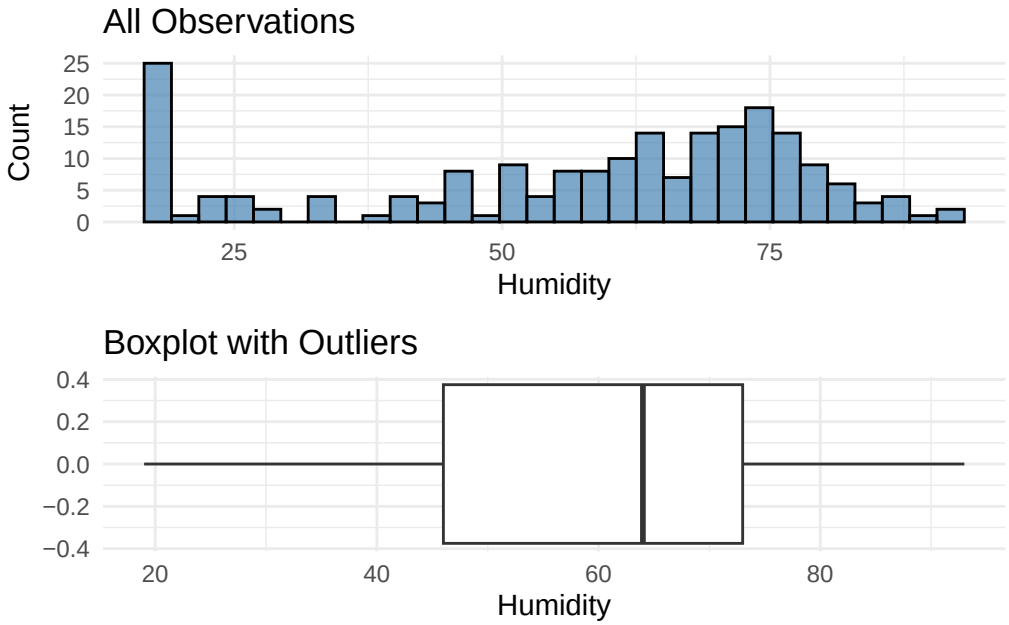

```
Outliers identified in Ozone_reading :  0 outliers
Proportion (%) of outliers: 0 %
```



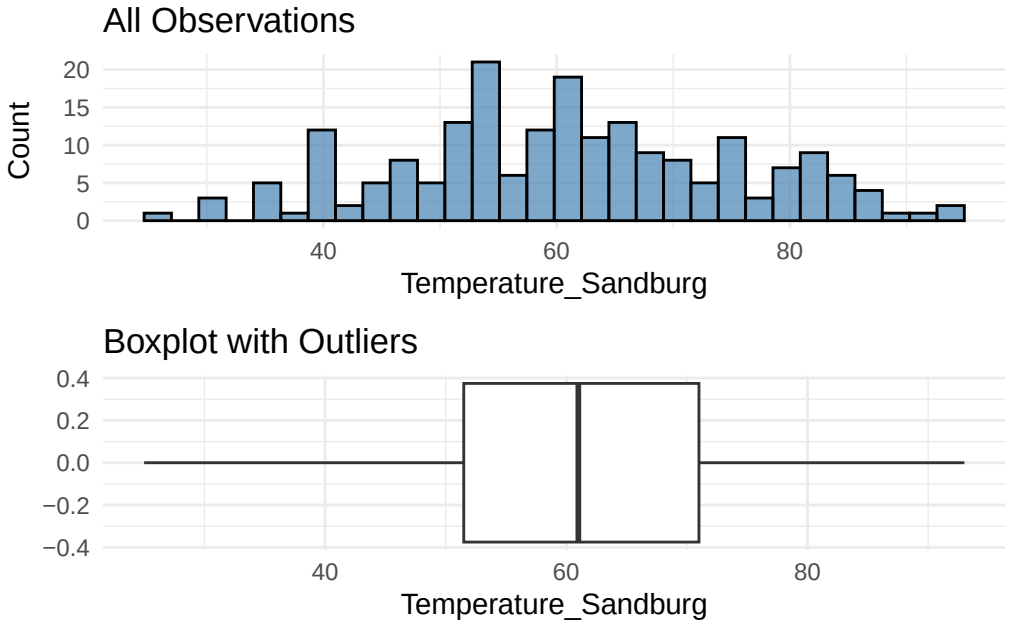
```
Outliers identified in Pressure_height :  0 outliers
Proportion (%) of outliers: 0 %
```



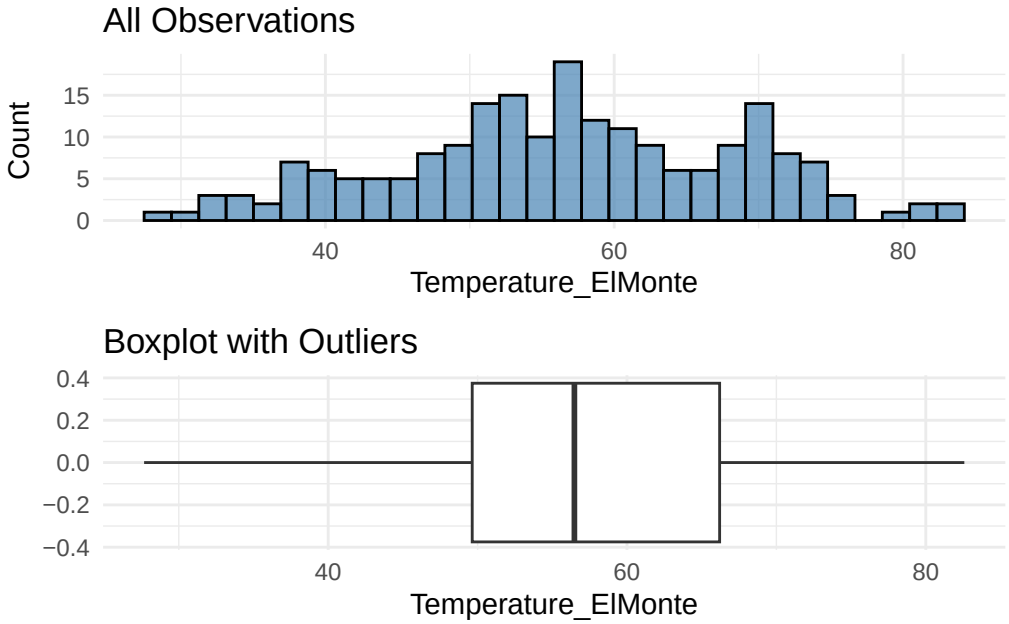

```
Outliers identified in Wind_speed :  1 outliers
Proportion (%) of outliers: 0.49 %
```




```
Outliers identified in Humidity :  0 outliers
Proportion (%) of outliers: 0 %
```

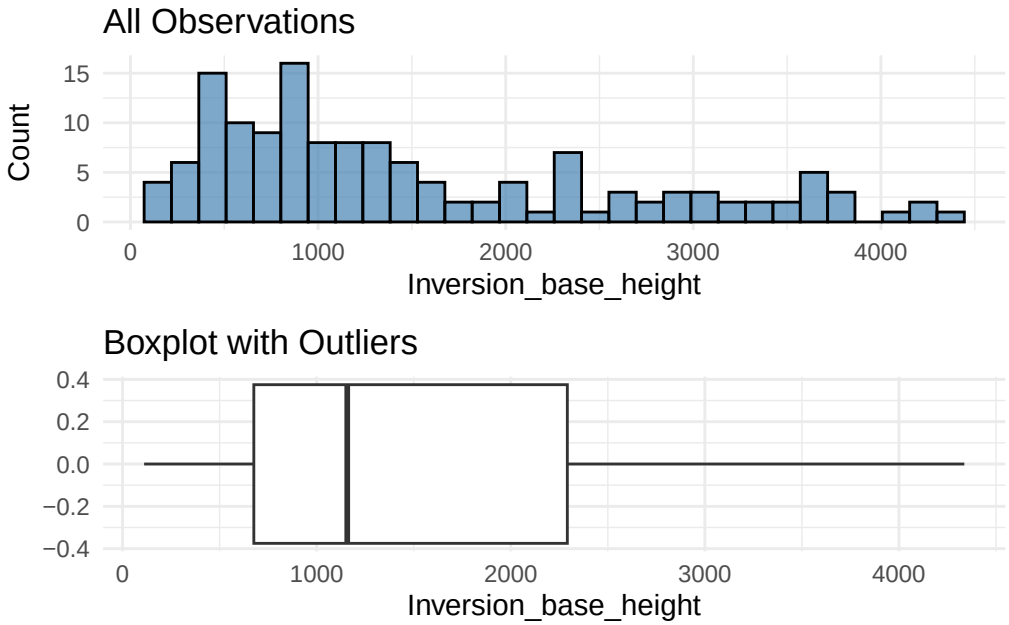


```
Outliers identified in Temperature_Sandburg :  0 outliers
Proportion (%) of outliers: 0 %
```

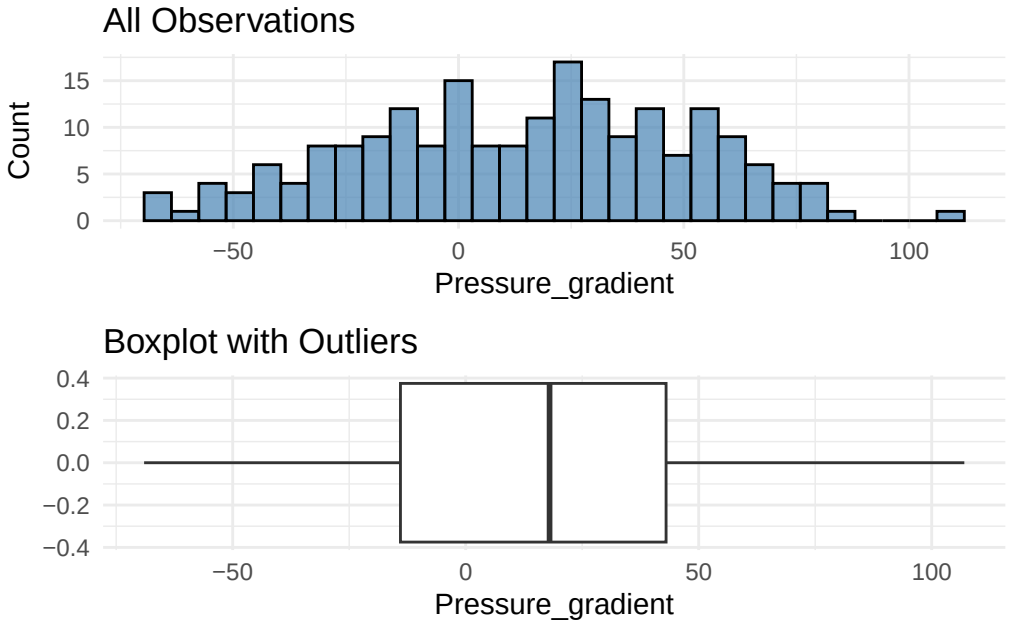



```
Outliers identified in Temperature_ElMonte :  0 outliers
Proportion (%) of outliers: 0 %
```

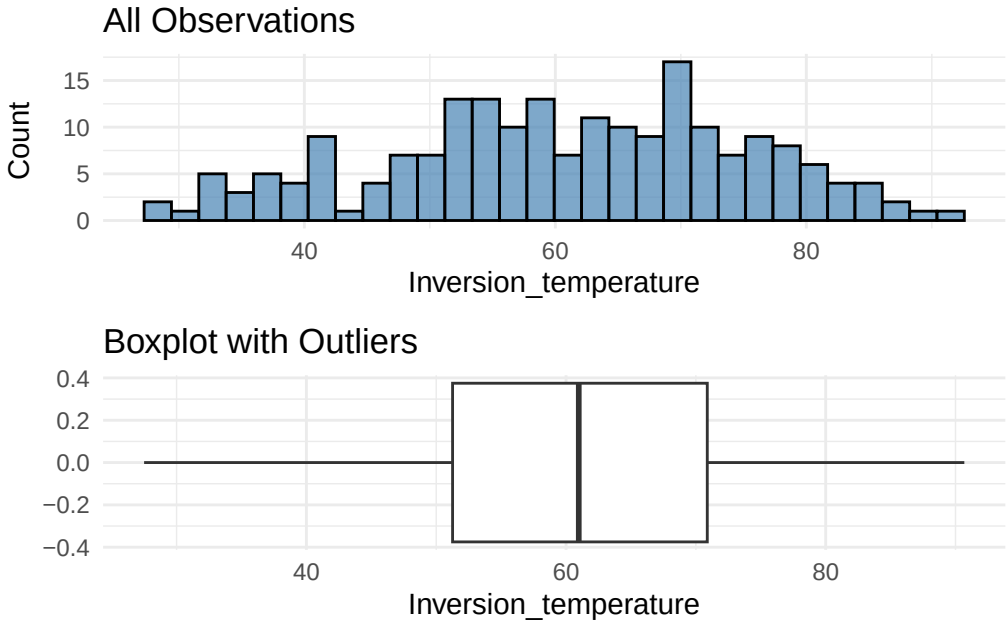
```
Warning: Removed 63 rows containing non-finite outside the scale range (`stat_bin()`).
Removed 63 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



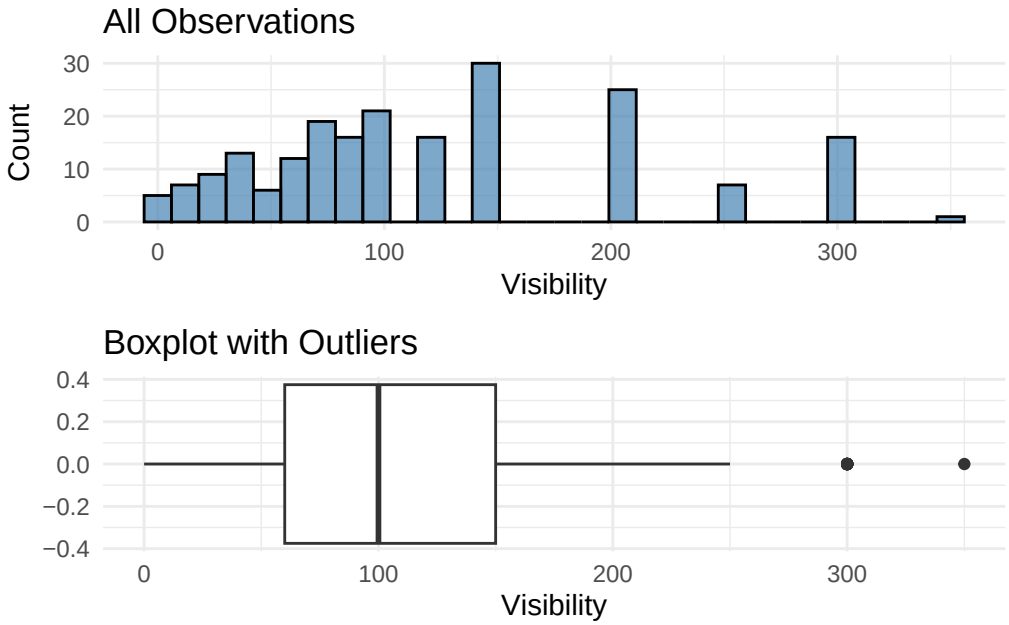

```
Outliers identified in Inversion_base_height :  0 outliers
Proportion (%) of outliers: 0 %
```



```
Outliers identified in Pressure_gradient :  0 outliers
Proportion (%) of outliers: 0 %
```




```
Outliers identified in Inversion_temperature :  0 outliers
Proportion (%) of outliers: 0 %
```




```
Outliers identified in Visibility :  0 outliers
Proportion (%) of outliers: 0 %
```

Las variables con datos atípicos son:

Pressure_heigth (2.46%): valores muy pequeños que parecen parte de una distribución asimétrica

Ozone_reading (1.48%): valores muy grandes que parecen parte de una distribución asimétrica

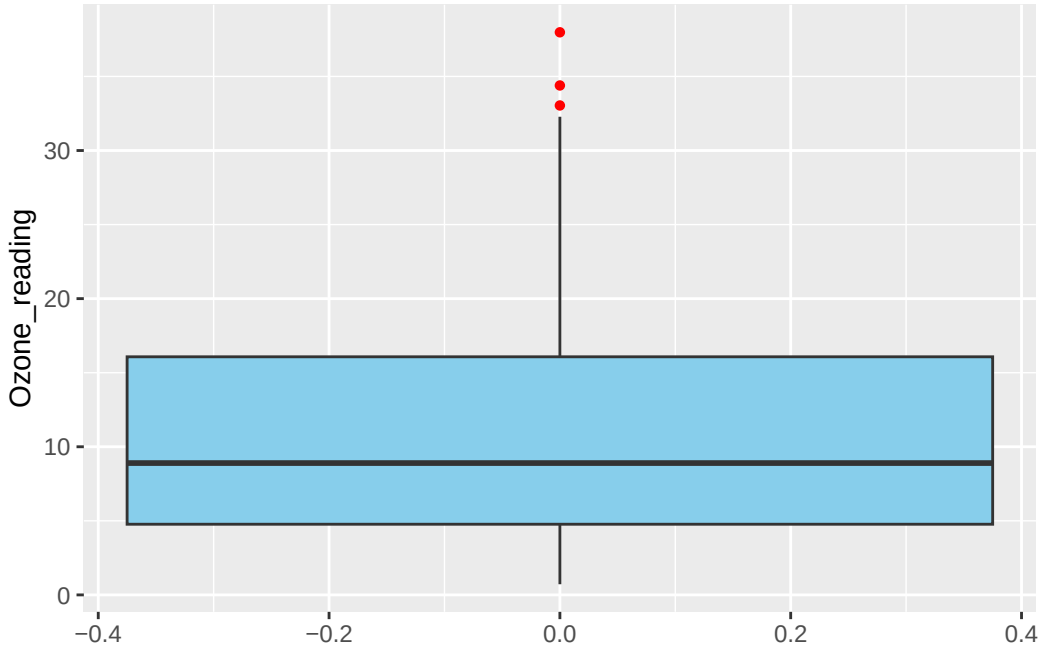
Wind_speed (0.49%): Valor que es también extremo y que se sale completamente de la distrinbución

Visibility (8.37%): Valores que corresponden a los mismos valores de 300 y 350 que parecen claramente parte de la variable. Además es un número muy elevado como para ser dato atípico.

Vamos por tanto a realizar el estudio bivariante de Pressure_heigth, Ozone_reading y Wind_speed

Estudio de la variable Ozone Reading

```
##### OZONE READING #####
ggplot(data, aes(y = Ozone_reading)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16)
```




```
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Ozone_reading)$out  # outlier values.
out_ind <- which(data$Ozone_reading %in% c(outlier_values))
data[out_ind,]
```

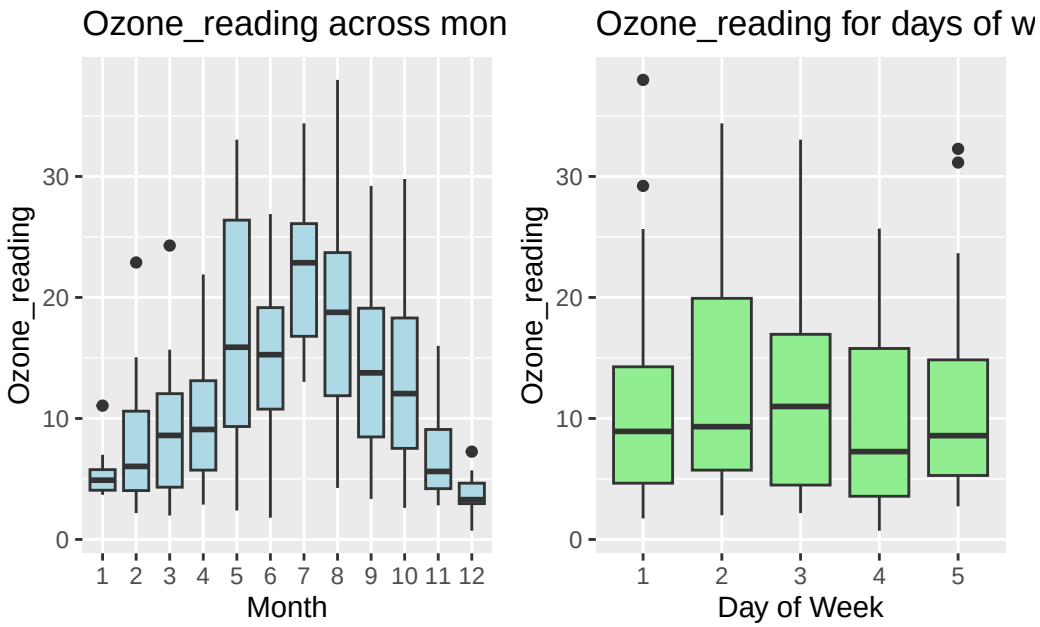
```
    Month Day_of_month Day_of_week Ozone_reading Pressure_height Wind_speed
82      5           12           3         33.04            5880          3
104     7            6           2         34.39            5900          6
130     8           30           1         37.98            5950          5
    Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
82        80                   80               73.04                  436
104       86                   87               81.68                  990
130       62                   92               82.40                  557
    Pressure_gradient Inversion_temperature Visibility
82                  0                86.36         40
```

```
104                22                  85.10          40
130                 0                  90.68          70
```

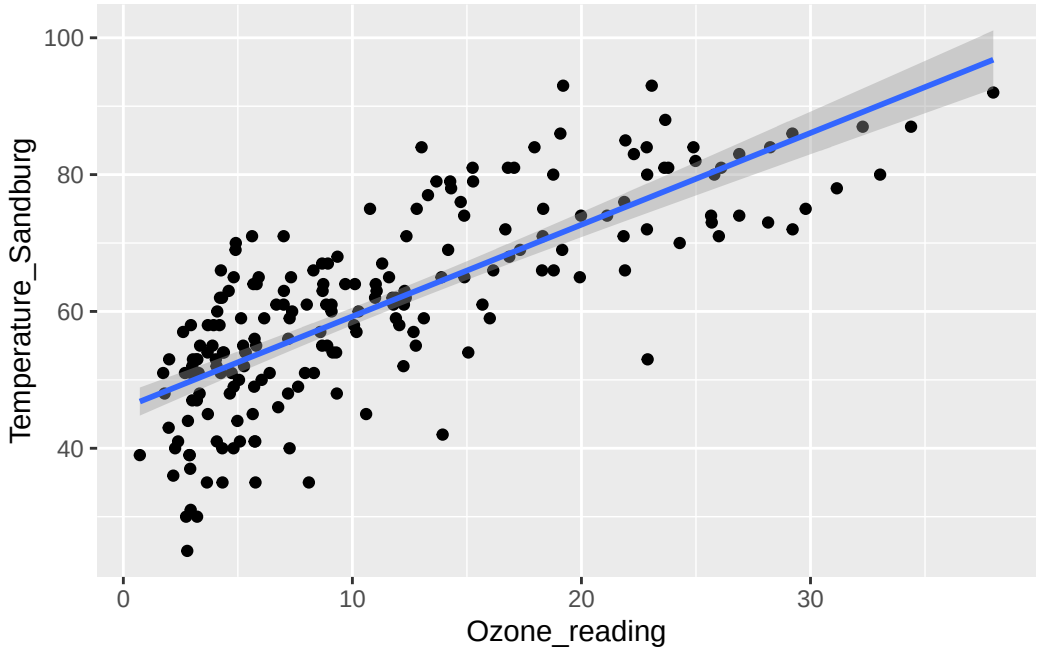
```
###Los valores extremos son:
extreme_values <- boxplot.stats(data$Pressure_height,coef=3)$out  # extreme values.
ext_ind <- which(data$Pressure_height %in% c(extreme_values))
data[ext_ind,]
```

```
 [1] Month                 Day_of_month          Day_of_week
 [4] Ozone_reading         Pressure_height       Wind_speed
 [7] Humidity              Temperature_Sandburg  Temperature_ElMonte
[10] Inversion_base_height Pressure_gradient     Inversion_temperature
[13] Visibility
<0 rows> (or 0-length row.names)
```

```
library(patchwork)  # Para combinar gráficos fácilmente

# Gráfico 1: Pressure Height por mes
p1 <- ggplot(data, aes(x = as.factor(Month), y = Ozone_reading)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Ozone_reading across months", x = "Month", y = "Ozone_reading")

# Gráfico 2: Pressure Height por día de la semana
p2 <- ggplot(data, aes(x = as.factor(Day_of_week), y = Ozone_reading)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Ozone_reading for days of week", x = "Day of Week", y = "Ozone_reading")

# Combinar ambos gráficos en una fila
p1 + p2
```




```
ggp <- ggplot(data,aes(Ozone_reading, Temperature_Sandburg)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```




```
summary(lm(data$Ozone_reading~data$Temperature_Sandburg))
```

```
Call:
lm(formula = data$Ozone_reading ~ data$Temperature_Sandburg)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4273  -3.8316  -0.4737   3.2197  15.1344

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -15.88133    1.61779  -9.817   <2e-16 ***
data$Temperature_Sandburg    0.44598    0.02579  17.294   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.207 on 201 degrees of freedom
Multiple R-squared:  0.5981,    Adjusted R-squared:  0.5961
F-statistic: 299.1 on 1 and 201 DF,  p-value: < 2.2e-16
```

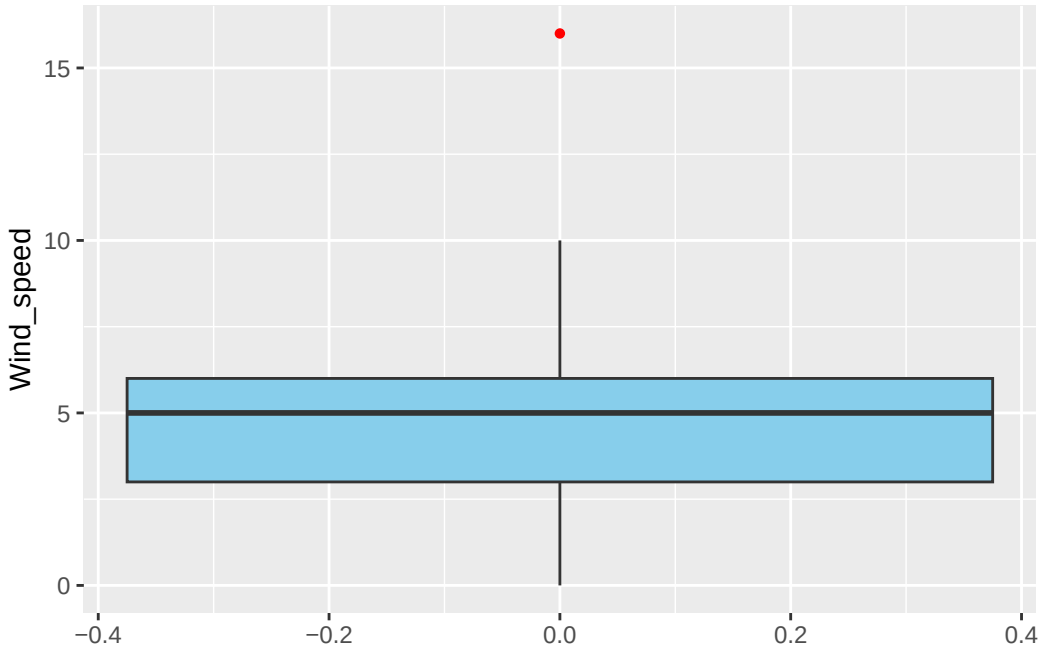
Esta variable está claramente asociada con los meses del año, perteneciendo los valores más

altos de esta variable a los meses de verano. Además vemos una clara asoaciación con la variable de temperatura.

CONCLUSIÓN: No borramos estos valores atípicos porque son parte de una asociación

Estudio de la variable WIND SPEED

```
ggplot(data, aes(y = Wind_speed)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16)
```



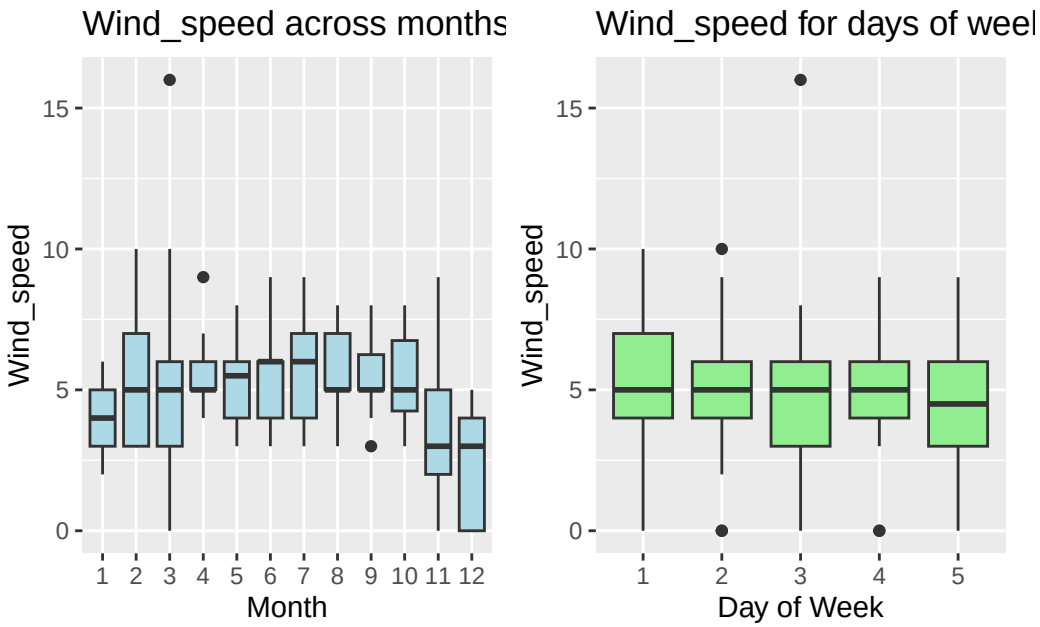

```
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Wind_speed)$out  # outlier values.
out_ind <- which(data$Wind_speed %in% c(outlier_values))
data[out_ind,]
```

```
   Month Day_of_month Day_of_week Ozone_reading Pressure_height Wind_speed
37     3            3           3          2.79            5320         16
   Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
37       45                   25               27.68                    NA
   Pressure_gradient Inversion_temperature Visibility
37                39                  27.5        200
```

```
###Los valores extremos son:
extreme_values <- boxplot.stats(data$Wind_speed,coef=3)$out  # extreme values.
ext_ind <- which(data$Wind_speed %in% c(extreme_values))
data[ext_ind,]
```

```
   Month Day_of_month Day_of_week Ozone_reading Pressure_height Wind_speed
37     3            3           3          2.79            5320         16
   Humidity Temperature_Sandburg Temperature_ElMonte Inversion_base_height
37       45                   25               27.68                    NA
   Pressure_gradient Inversion_temperature Visibility
37                39                  27.5        200
```

```
library(patchwork)  # Para combinar gráficos fácilmente

# Gráfico 1: Pressure Height por mes
p1 <- ggplot(data, aes(x = as.factor(Month), y = Wind_speed)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Wind_speed across months", x = "Month", y = "Wind_speed")

# Gráfico 2: Pressure Height por día de la semana
p2 <- ggplot(data, aes(x = as.factor(Day_of_week), y = Wind_speed)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Wind_speed for days of week", x = "Day of Week", y = "Wind_speed")

# Combinar ambos gráficos en una fila
p1 + p2
```



En este caso vemos que el outlier de wind_speed no está asoiciado con las variables de interés y además es un extremo.

CONCLUSIÓN: Este outlier no tiene ninguna asociación aparente, por tanto este dato missing si lo quitamos

```
outlier_values <- boxplot.stats(data$Wind_speed)$out  # outlier values.
out_ind <- which(data$Wind_speed %in% c(outlier_values))
data[out_ind,"Wind_speed"]<-NA
```

Estudio Multivariante

```
library(dbscan)
```

```
Attaching package: 'dbscan'

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

    as.dendrogram
```

```r
library(class)
library(ggplot2)
library(tidyverse)
```

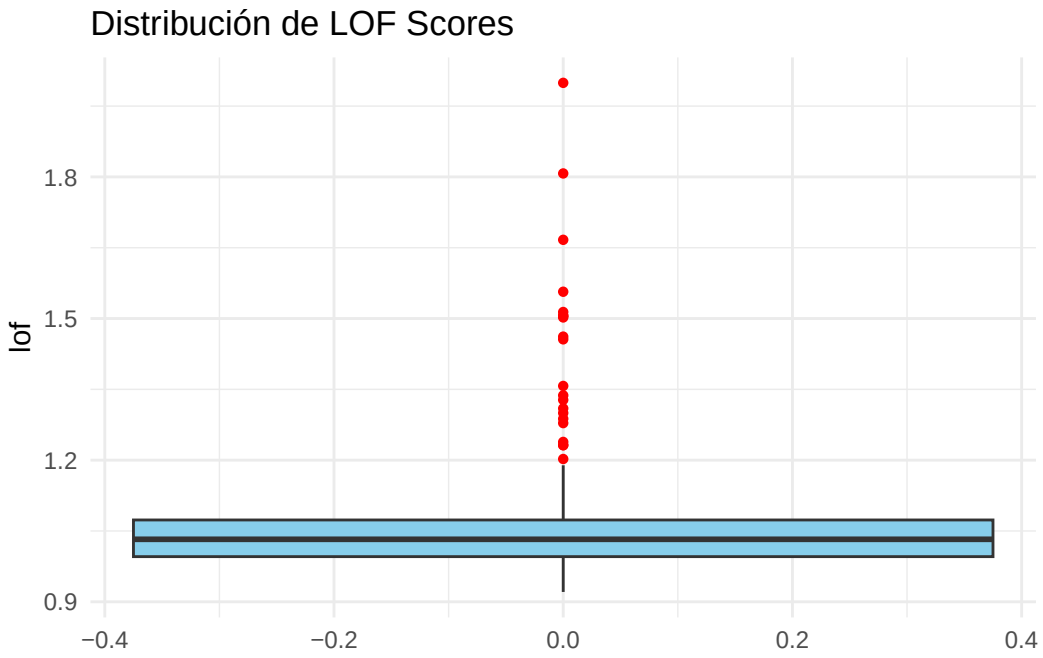
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
v readr     2.1.5

-- 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 becor
```

```r
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)


####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])
```

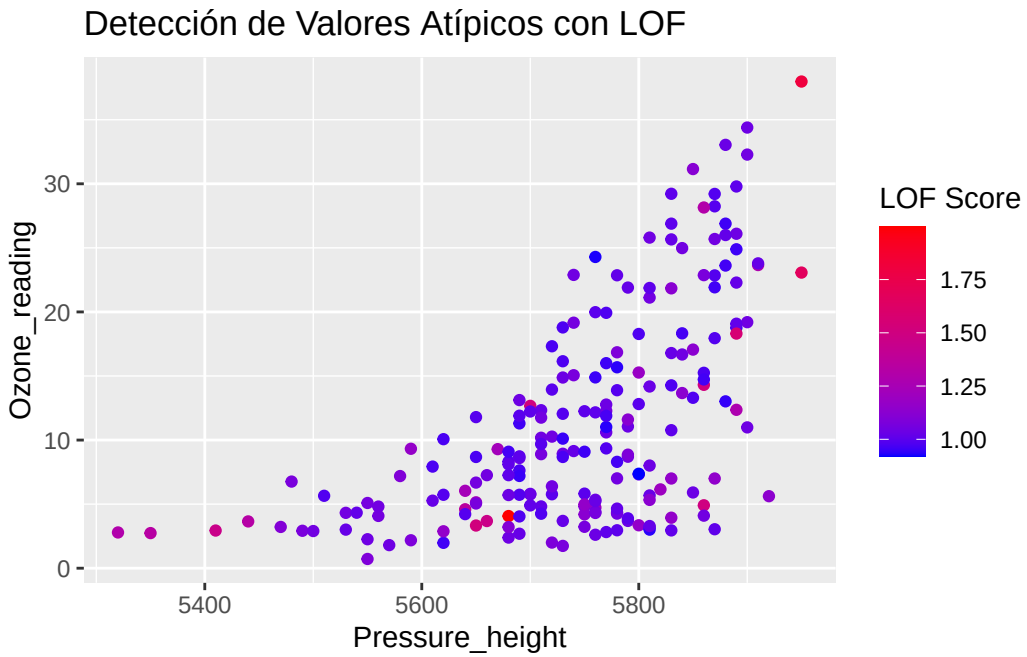
```
    Month Day_of_month Day_of_week Ozone_reading Pressure_height Wind_speed
11      1           19           1          4.07            5680          5
47      3           18           4         12.67            5700          4
97      6           16           3         14.31            5860          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
```

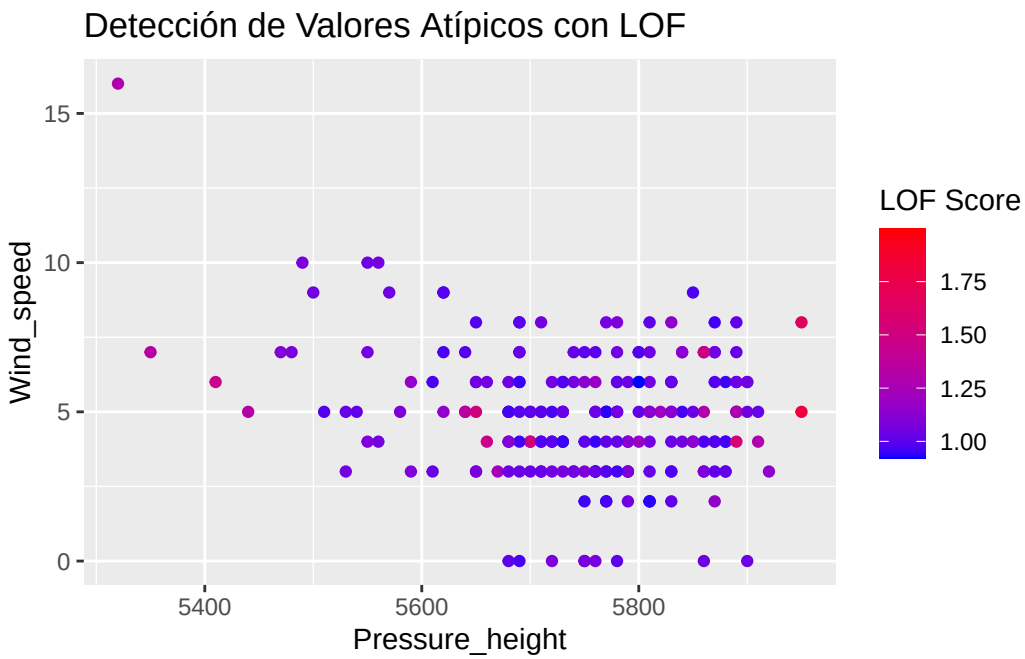
```r
data$lof<-lof

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



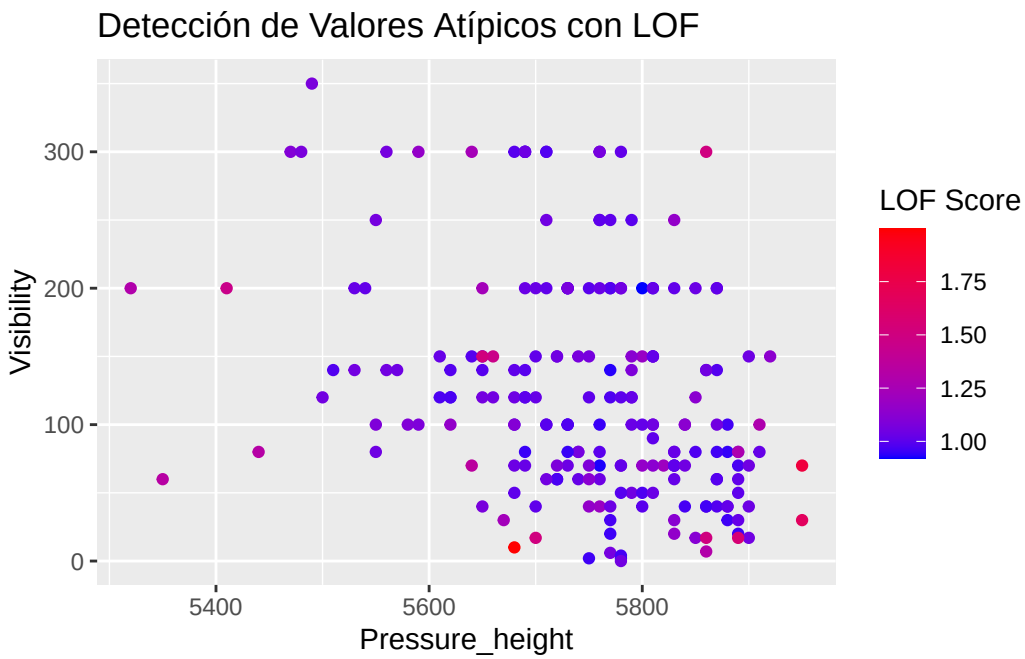
```r
####Comprobamos las cuantitativas
ggplot(data, aes(x = Pressure_height, y = Ozone_reading, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



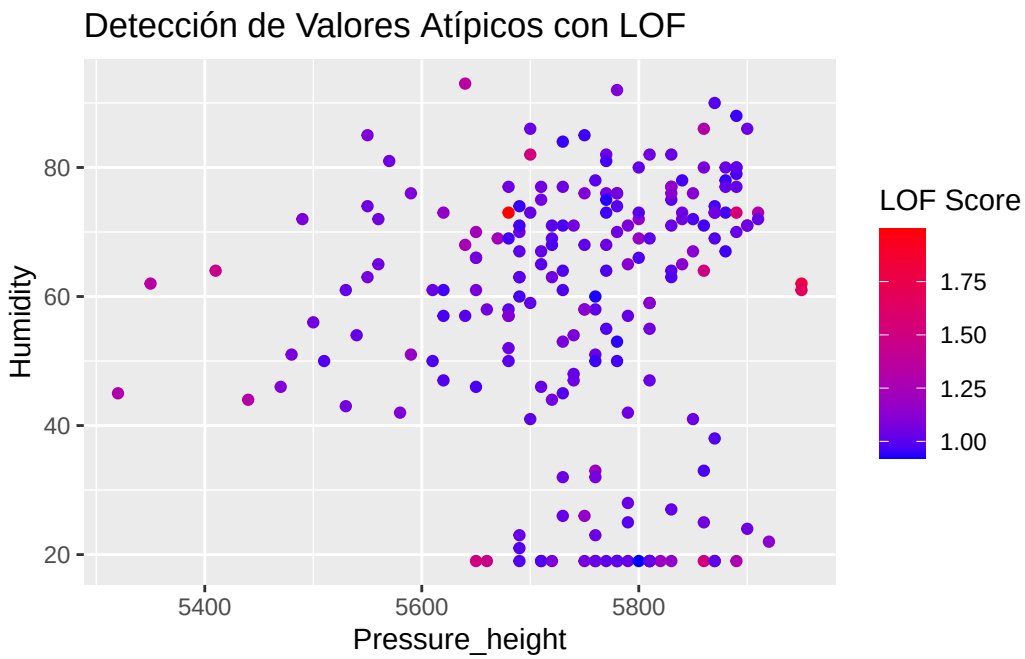

```r
ggplot(data, aes(x = Pressure_height, y = Wind_speed, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



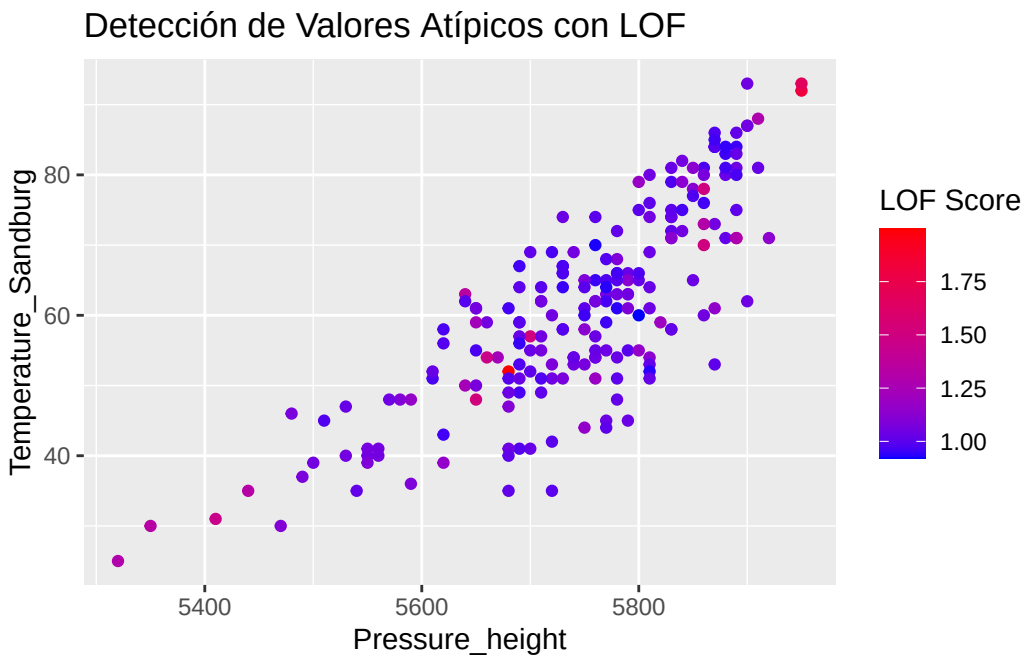

```
ggplot(data, aes(x = Pressure_height, y = Visibility, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```




```
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")
```

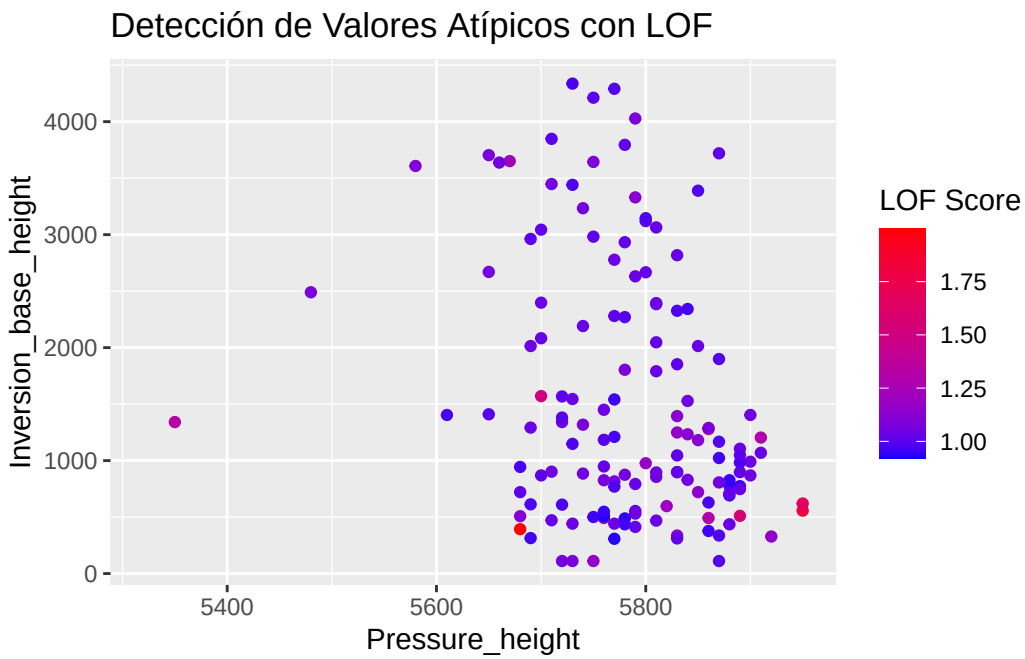



```
ggplot(data, aes(x = Pressure_height, y = Temperature_Sandburg, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



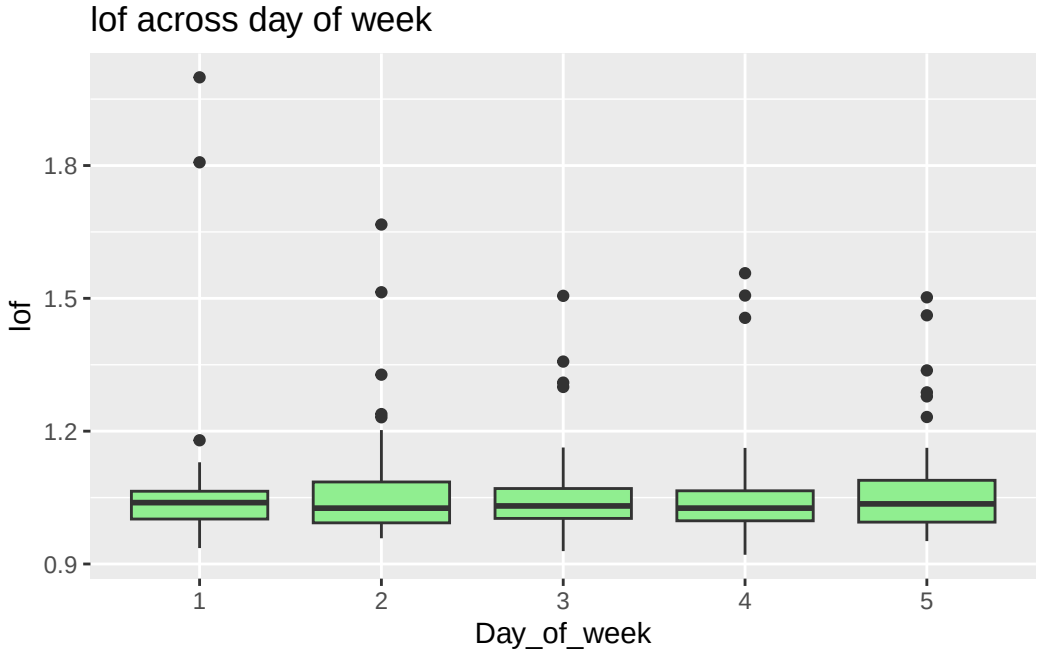

```
ggplot(data, aes(x = Pressure_height, y = Inversion_base_height, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```

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




```
####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")
```


```
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")
```



No observamos ninguna observación que tenga un LOF especialmente grande más allá de que los puntos rojos que son los que mayor LOF tienen siempre corresponden a los puntos más alejados de las nubes de puntos al representar las variables dos a dos que corresponden a los datos atípicos que ya hemos observado en el estudio previo. No borraremos nada más.