Ejercicio 2.2: Detección y tratamiento de datos atípicos univariante y bivariante

Silvia Pineda

Lectura Fichero de datos

```
library(ggplot2)
library(patchwork)
data <- read.csv("ozone.csv")  # import data
str(data)
```

```
'data.frame':   203 obs. of  13 variables:
 $ Month                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Day_of_month          : int  5 6 7 8 9 12 13 14 15 16 ...
 $ Day_of_week           : int  1 2 3 4 5 1 2 3 4 5 ...
 $ Ozone_reading         : num  5.34 5.77 3.69 3.89 5.76 6.39 4.73 4.35 3.94 7 ...
 $ Pressure_height       : int  5760 5720 5790 5790 5700 5720 5760 5780 5830 5870 ...
 $ Wind_speed            : int  3 4 6 3 3 3 6 6 3 2 ...
 $ Humidity              : int  51 69 19 25 73 44 33 19 19 19 ...
 $ Temperature_Sandburg  : int  54 35 45 55 41 51 51 54 58 61 ...
 $ Temperature_ElMonte   : num  45.3 49.6 46.4 52.7 48 ...
 $ Inversion_base_height: int  1450 1568 2631 554 2083 111 492 NA 1249 NA ...
 $ Pressure_gradient     : int  25 15 -33 -28 23 9 -44 -44 -53 -67 ...
 $ Inversion_temperature: num  57 53.8 54.1 64.8 52.5 ...
 $ Visibility            : int  60 60 100 250 120 150 40 200 250 200 ...
```

```
summary(data)
```

```
     Month          Day_of_month    Day_of_week     Ozone_reading
 Min.   : 1.000   Min.   : 1.0   Min.   :1.000   Min.   : 0.72
 1st Qu.: 3.000   1st Qu.: 9.0   1st Qu.:2.000   1st Qu.: 4.77
```

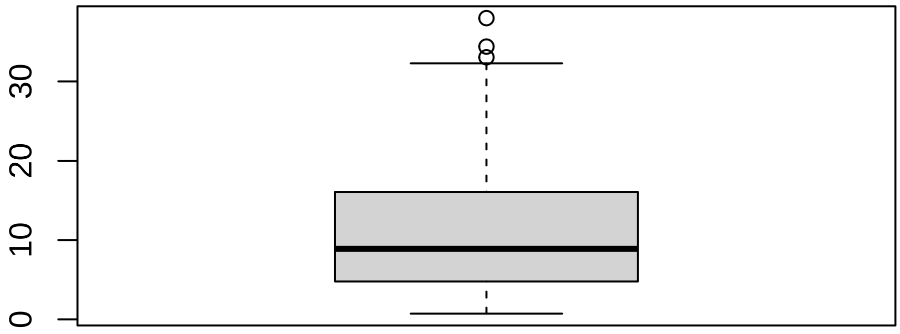
```
Median : 6.000   Median :15.0   Median :3.000   Median : 8.90
Mean   : 6.522   Mean   :15.7   Mean   :3.005   Mean   :11.37
3rd Qu.:10.000   3rd Qu.:23.0   3rd Qu.:4.000   3rd Qu.:16.07
Max.   :12.000   Max.   :31.0   Max.   :5.000   Max.   :37.98

Pressure_height   Wind_speed        Humidity     Temperature_Sandburg
Min.   :5320    Min.   : 0.000   Min.   :19.00   Min.   :25.00
1st Qu.:5690    1st Qu.: 3.000   1st Qu.:46.00   1st Qu.:51.50
Median :5760    Median : 5.000   Median :64.00   Median :61.00
Mean   :5746    Mean   : 4.887   Mean   :57.61   Mean   :61.11
3rd Qu.:5830    3rd Qu.: 6.000   3rd Qu.:73.00   3rd Qu.:71.00
Max.   :5950    Max.   :16.000   Max.   :93.00   Max.   :93.00

Temperature_ElMonte Inversion_base_height Pressure_gradient
Min.   :27.68       Min.   : 111.0        Min.   :-69.00
1st Qu.:49.64       1st Qu.: 676.2        1st Qu.:-14.00
Median :56.48       Median :1157.5        Median : 18.00
Mean   :56.54       Mean   :1522.5        Mean   : 14.43
3rd Qu.:66.20       3rd Qu.:2291.5        3rd Qu.: 43.00
Max.   :82.58       Max.   :4337.0        Max.   :107.00
                    NA's   :63
Inversion_temperature   Visibility
Min.   :27.50         Min.   :  0.0
1st Qu.:51.26         1st Qu.: 60.0
Median :60.98         Median :100.0
Mean   :60.69         Mean   :122.2
3rd Qu.:70.88         3rd Qu.:150.0
Max.   :90.68         Max.   :350.0
```

```r
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)
```

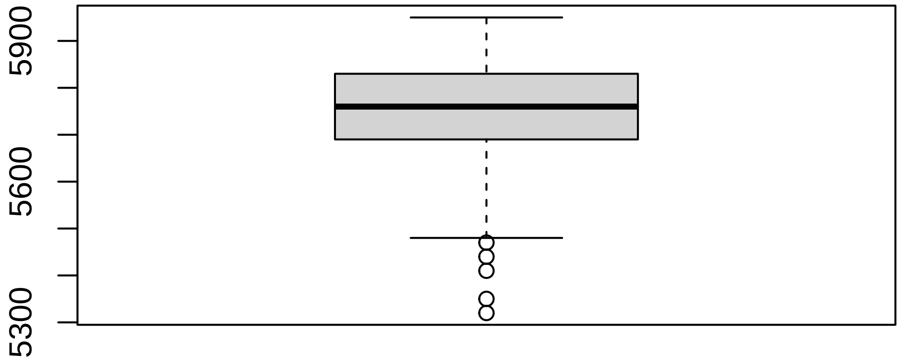
Estudio Univariante

Visualización

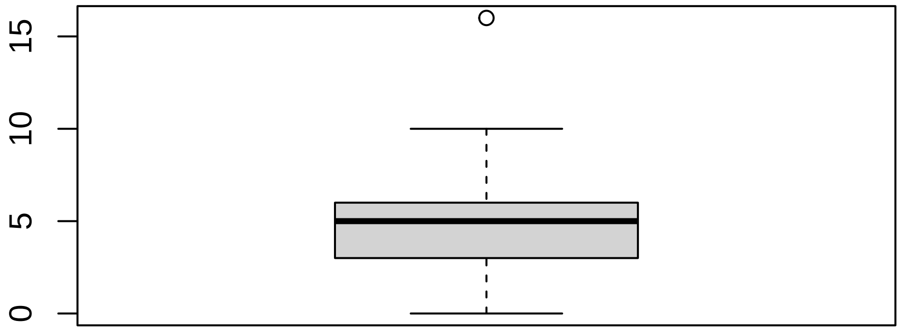
```r
boxplot(data$Ozone_reading)
```



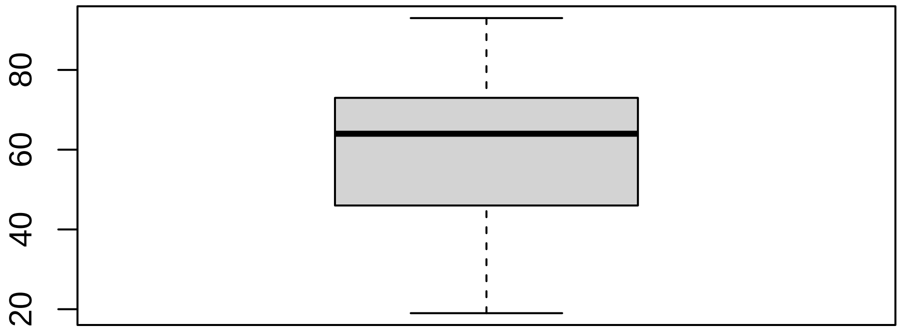
```
boxplot(data$Pressure_height)
```



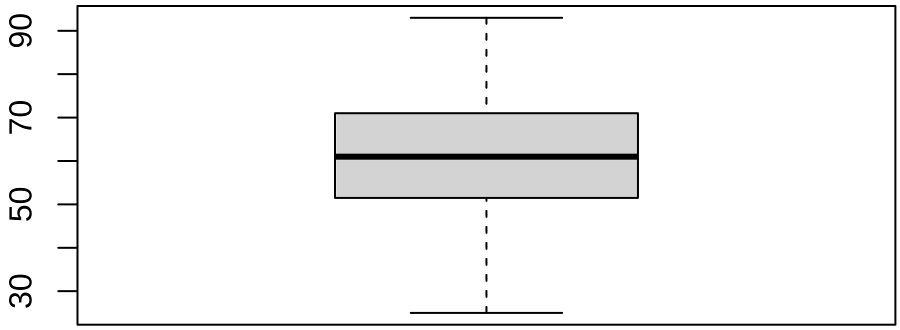
```
boxplot(data$Wind_speed)
```



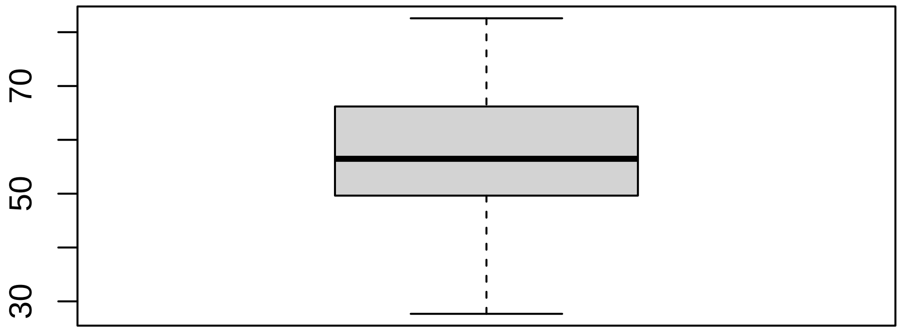
```
boxplot(data$Humidity)
```



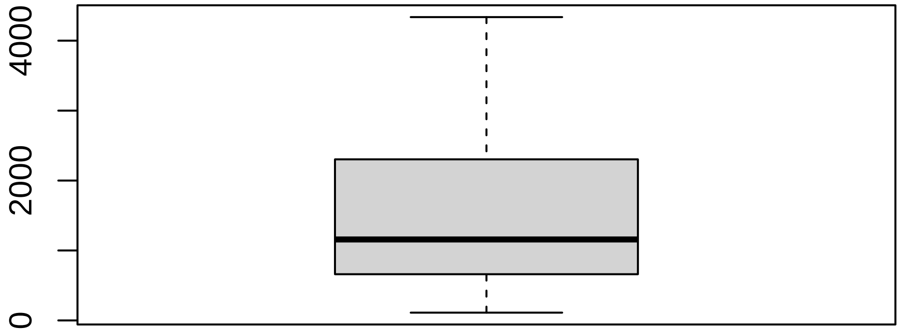
```r
boxplot(data$Temperature_Sandburg)
```



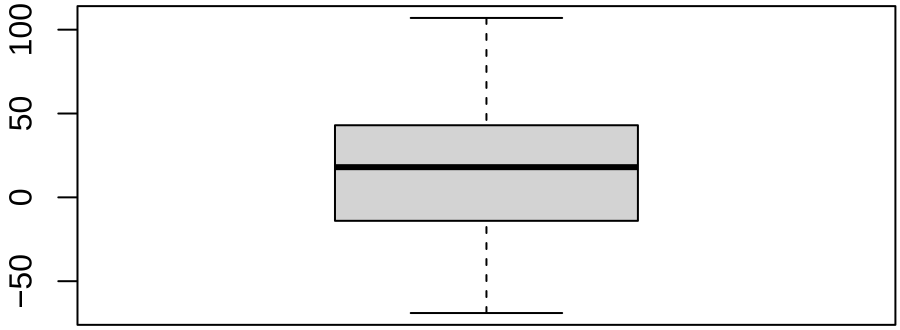
```r
boxplot(data$Temperature_ElMonte)
```



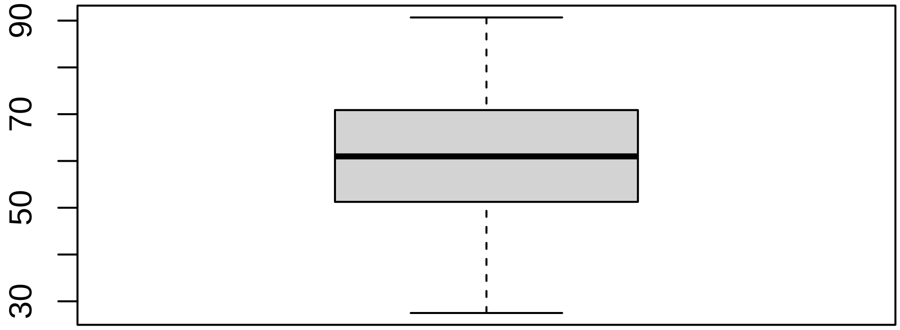
```r
boxplot(data$Inversion_base_height)
```



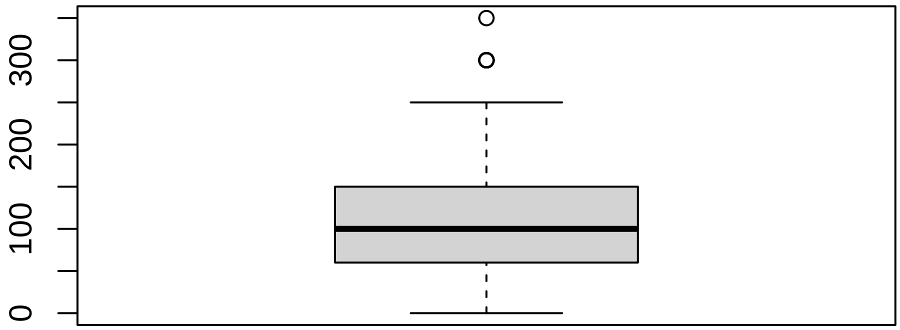
```
boxplot(data$Pressure_gradient)
```



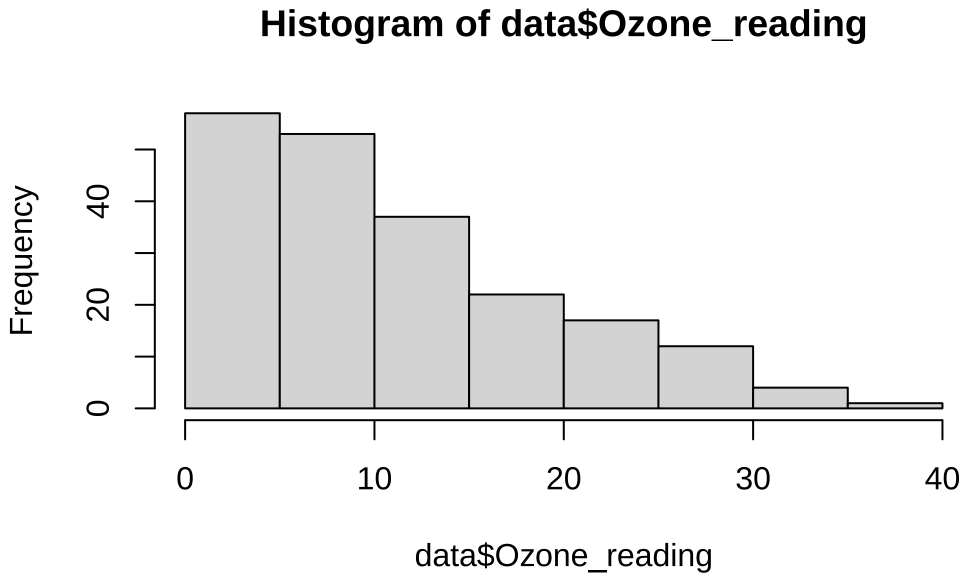
```
boxplot(data$Inversion_temperature)
```



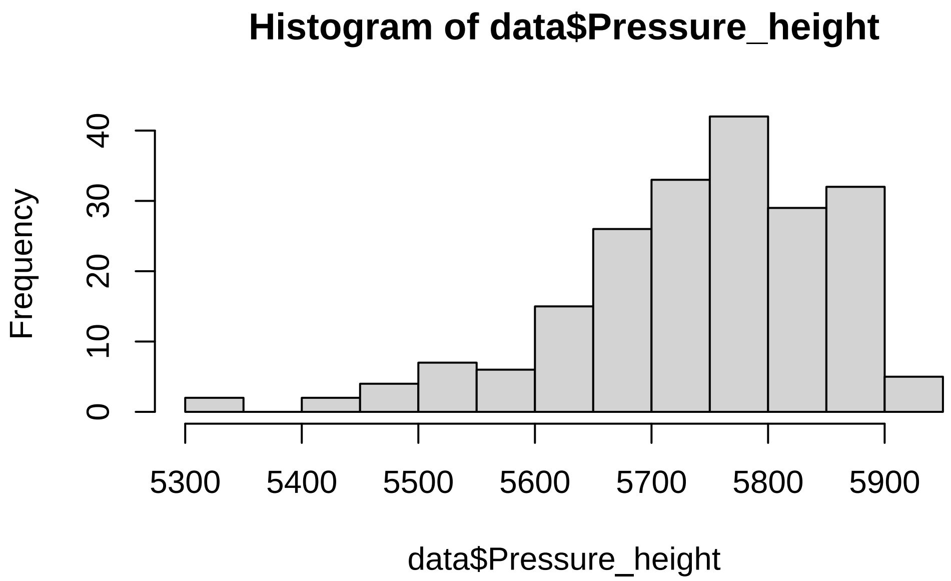
```
boxplot(data$Visibility)
```



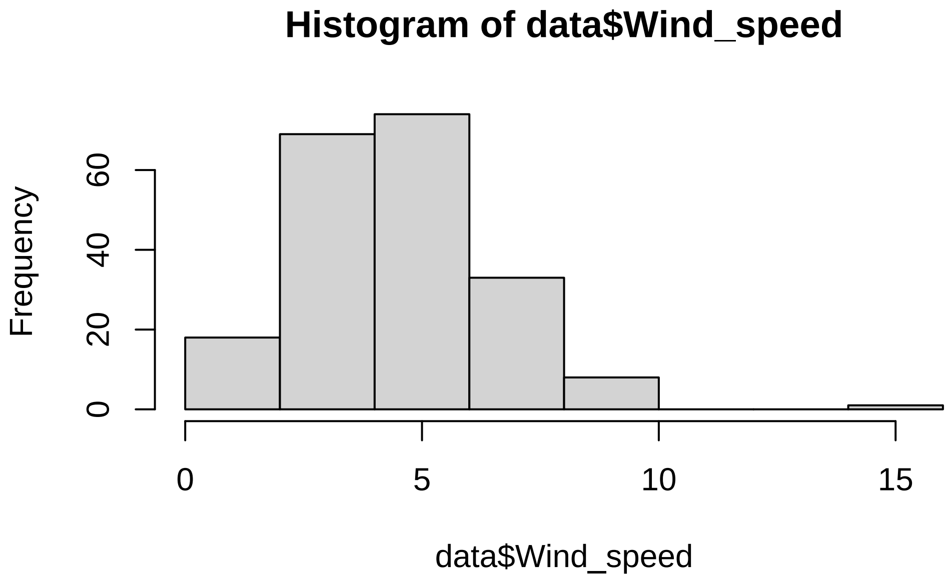
```
#Solo hago histogramas para las variables con atípicos

hist(data$Ozone_reading)
```



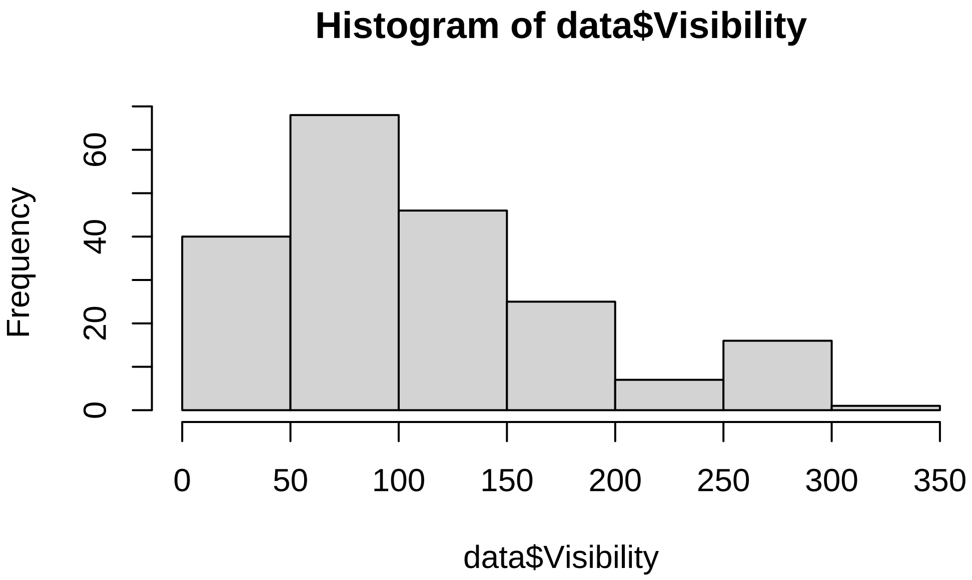
```
hist(data$Pressure_height)
```



```r
hist(data$Wind_speed)
```



```r
hist(data$Visibility)
```



Las variables con atípicos son: `Ozone_reading`, `Pressure_height`, `Wind_speed` y `Visibility`

Cuantificación

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

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

####Miramos la proporción de outliers y extremos
p<-length(out_ind)/length(data$Ozone_reading)*100
q<-length(ext_ind)/length(data$Ozone_reading)*100
cat("El % de outliers para la variable Ozone_reading es:", p,"\n")
```

```
El % de outliers para la variable Ozone_reading es: 1.477833
```

```r
cat("El % de extremos para la variable Ozone_readinges", q,"\n")
```

```
El % de extremos para la variable Ozone_readinges 0
```

```r
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Pressure_height)$out  # outlier values.
out_ind <- which(data$Pressure_height %in% c(outlier_values)) # índices

###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)) # índices

####Miramos la proporción de outliers y extremos
p<-length(out_ind)/length(data$Pressure_height)*100
q<-length(ext_ind)/length(data$Pressure_height)*100
cat("El % de outliers para la variable Pressure_height es:", p,"\n")
```

```
El % de outliers para la variable Pressure_height es: 2.463054
```

```r
cat("El % de extremos para la variable Pressure_height", q,"\n")
```

```
El % de extremos para la variable Pressure_height 0
```

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

###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)) # índices

####Miramos la proporción de outliers y extremos
p<-length(out_ind)/length(data$Wind_speed)*100
q<-length(ext_ind)/length(data$Wind_speed)*100
cat("El % de outliers para la variable Wind_speed es:", p,"\n")
```

```
El % de outliers para la variable Wind_speed es: 0.4926108
```

```
cat("El % de extremos para la variable Wind_speed", q,"\n")
```

```
El % de extremos para la variable Wind_speed 0.4926108
```

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

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

####Miramos la proporción de outliers y extremos
p<-length(out_ind)/length(data$Visibility)*100
q<-length(ext_ind)/length(data$Visibility)*100
cat("El % de outliers para la variable Visibility es:", p,"\n")
```

```
El % de outliers para la variable Visibility es: 8.374384
```

```
cat("El % de extremos para la variable Visibility", q,"\n")
```

```
El % de extremos para la variable Visibility 0
```

Las variables con datos atípicos son:

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

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

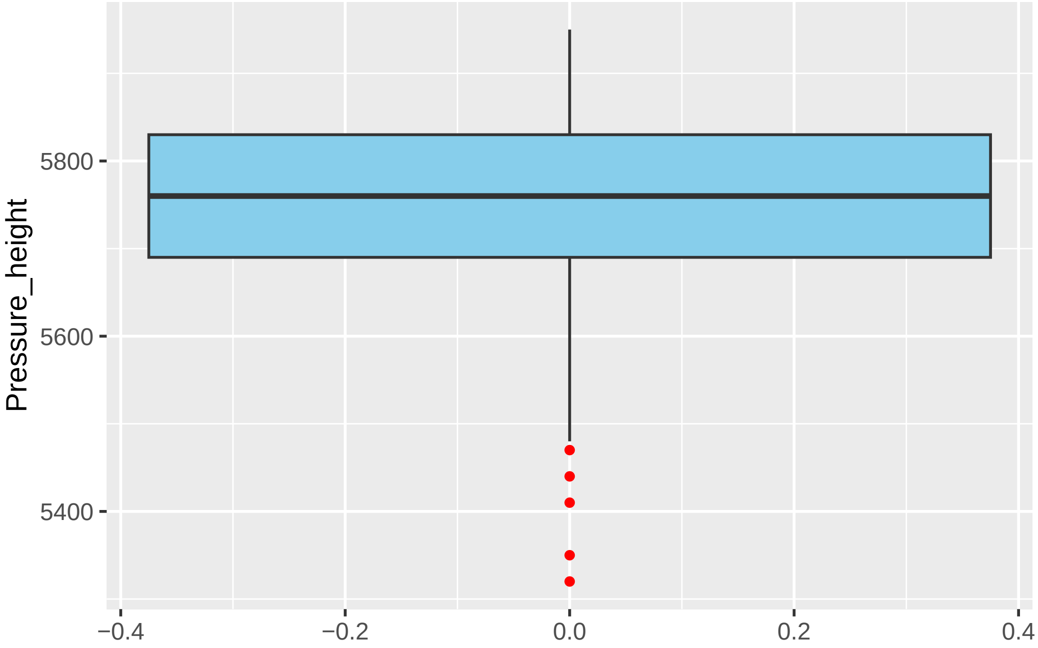
`Wind_speed` (0.49%): Valor que es también extremo y que se sale completamente de la distrinbución

`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_height`, `Ozone_reading` y `Wind_speed`

Estudio de la variable Pressure_height

```r
##### Pressure heigth #####
ggplot(data, aes(y = Pressure_height)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16)
```



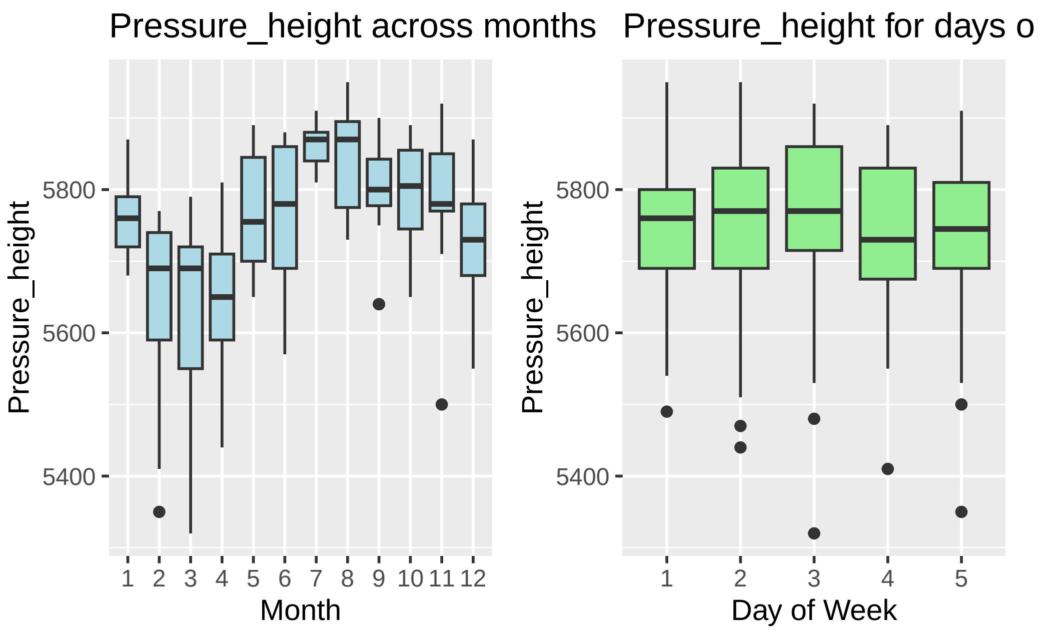
```r
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Pressure_height)$out  # outlier values.
out_ind <- which(data$Ozone_reading %in% c(outlier_values))
data[out_ind,]
```

```
 [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)
```

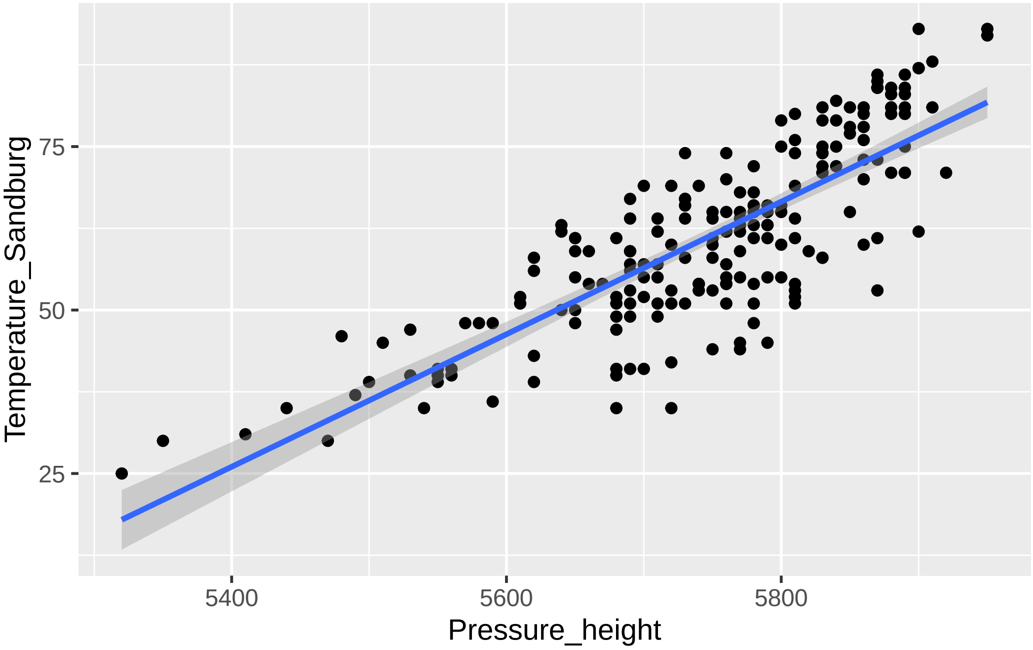
```r
###Los valores extremos son:
extreme_values <- boxplot.stats(data$Pressure_height,coef=3)$out  # extreme values.
ext_ind <- which(data$Pressure_height %in% c(extreme_values))
data[ext_ind,]
```

```
 [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 = Pressure_height)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Pressure_height across months", x = "Month", y = "Pressure_height")

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

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



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



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

```

Call:
lm(formula = data$Pressure_height ~ data$Temperature_Sandburg)

Residuals:
     Min       1Q   Median       3Q      Max
-196.559  -41.846    1.171   39.099  175.891

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5354.1021    20.8228  257.13   <2e-16 ***
data$Temperature_Sandburg    6.4152     0.3319   19.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

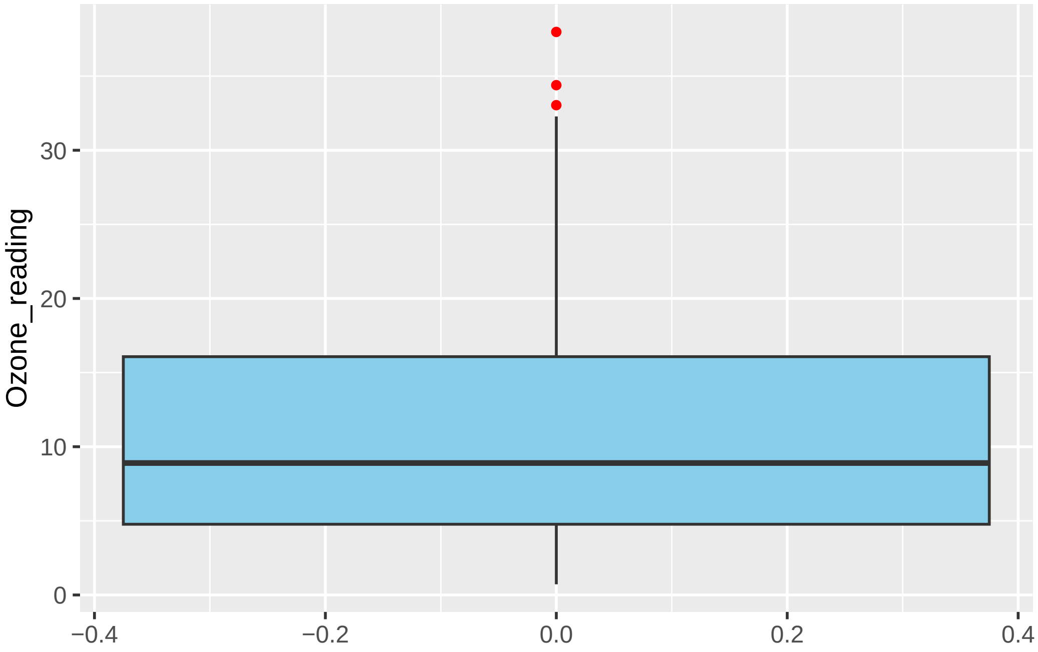
```
Residual standard error: 67.02 on 201 degrees of freedom
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6484
F-statistic: 373.6 on 1 and 201 DF,  p-value: < 2.2e-16
```

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 asociación con la variable de temperatura.

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

Estudio de la variable Ozone Reading

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



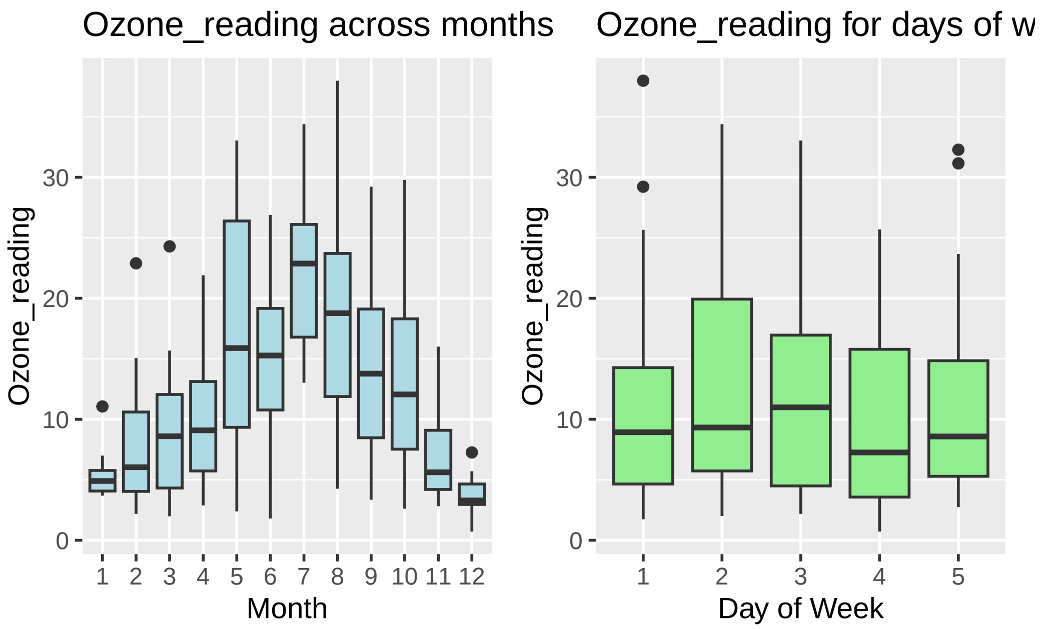
```
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Ozone_reading)$out  # outlier values.
out_ind <- which(data$Ozone_reading %in% c(outlier_values))
data[out_ind,]
```

```
    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
```

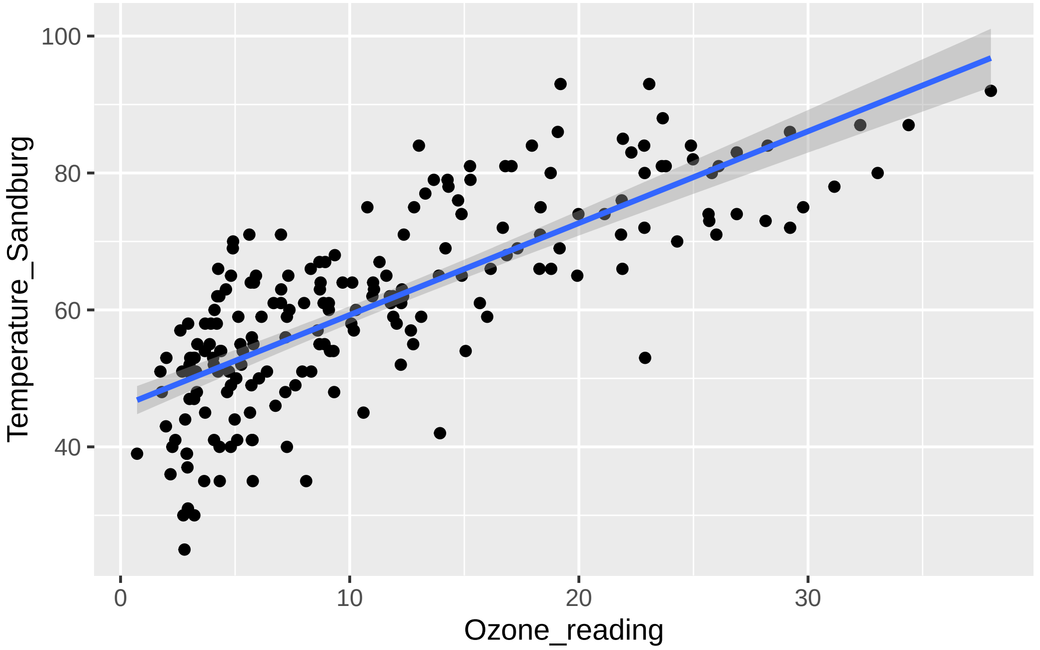
```
###Los valores extremos son:
extreme_values <- boxplot.stats(data$Ozone_reading,coef=3)$out  # extreme values.
ext_ind <- which(data$Ozone_reading %in% c(extreme_values))
data[ext_ind,]
```

```
 [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: Ozone_reading 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: Ozone_reading 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
```



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



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

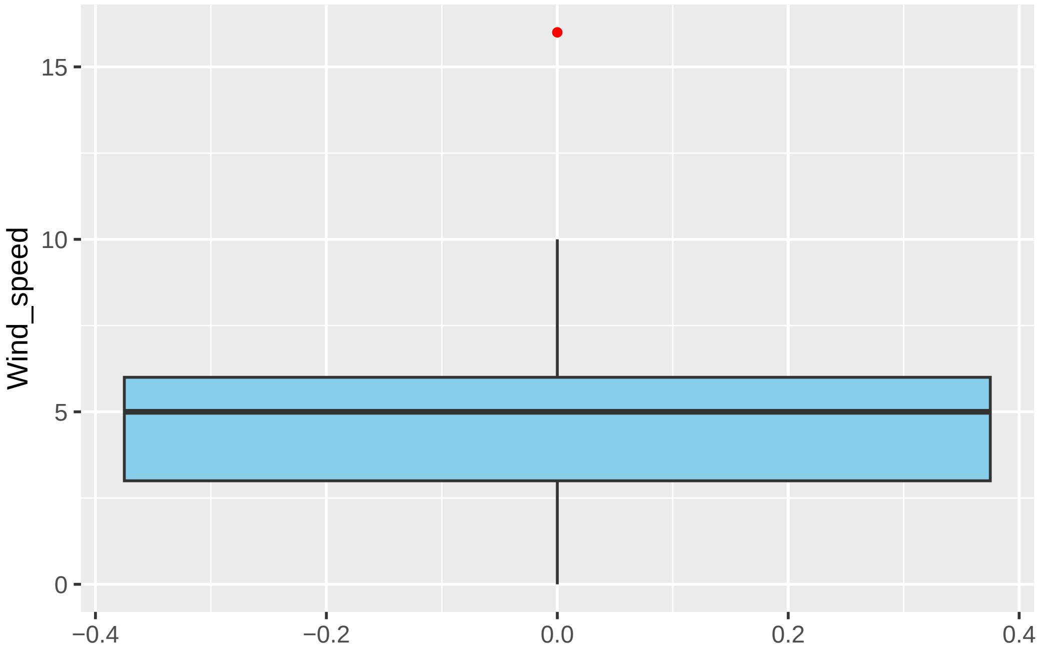
```
Call:
lm(formula = data$Ozone_reading ~ data$Temperature_Sandburg)

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
```

De la misma forma que la variable anterior, 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

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



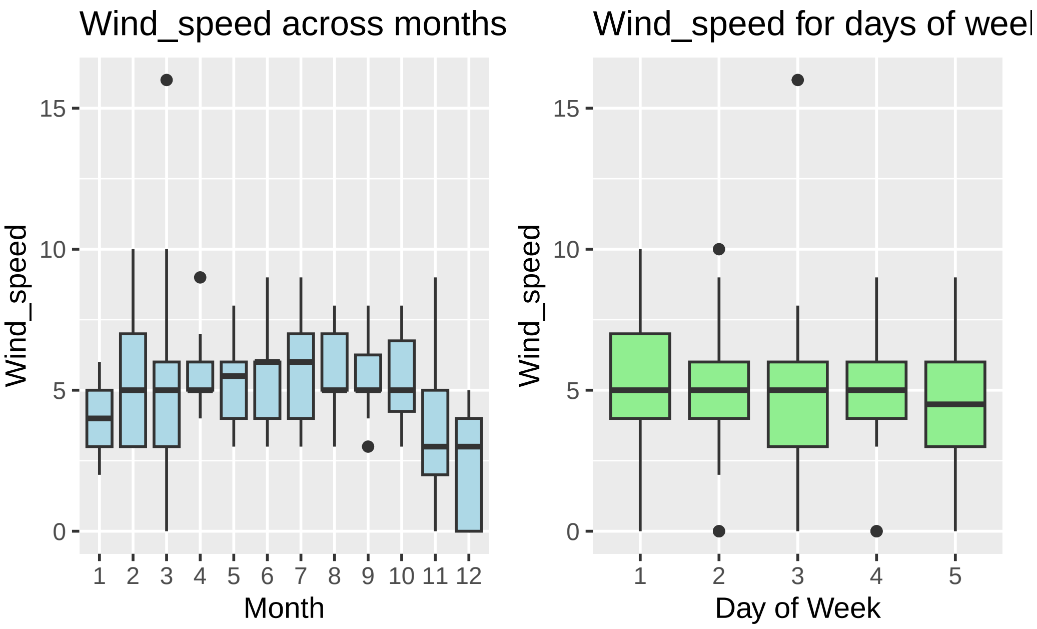
```
###Los valores atípicos son:
outlier_values <- boxplot.stats(data$Wind_speed)$out  # outlier values.
out_ind <- which(data$Wind_speed %in% c(outlier_values))
data[out_ind,]
```

```
   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
```

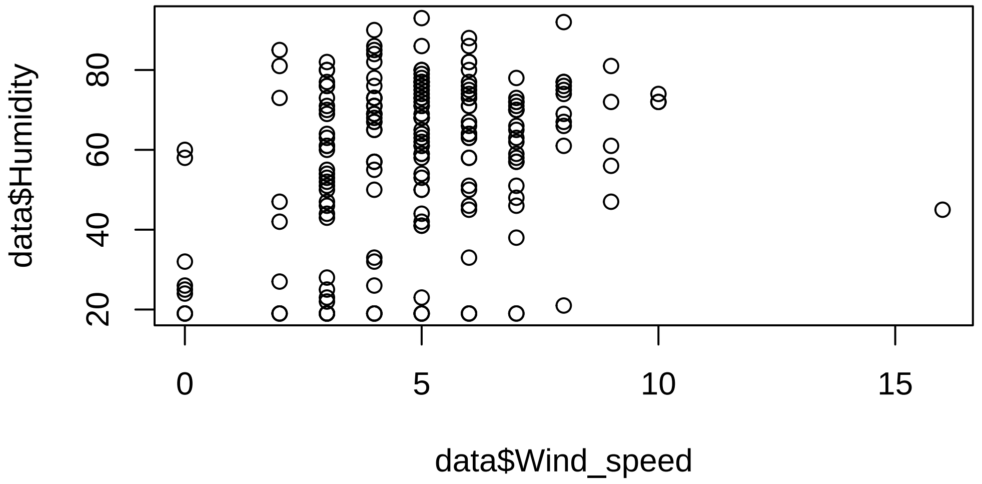
```
###Los valores extremos son:
extreme_values <- boxplot.stats(data$Wind_speed,coef=3)$out  # extreme values.
ext_ind <- which(data$Wind_speed %in% c(extreme_values))
data[ext_ind,]
```

```
   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
```

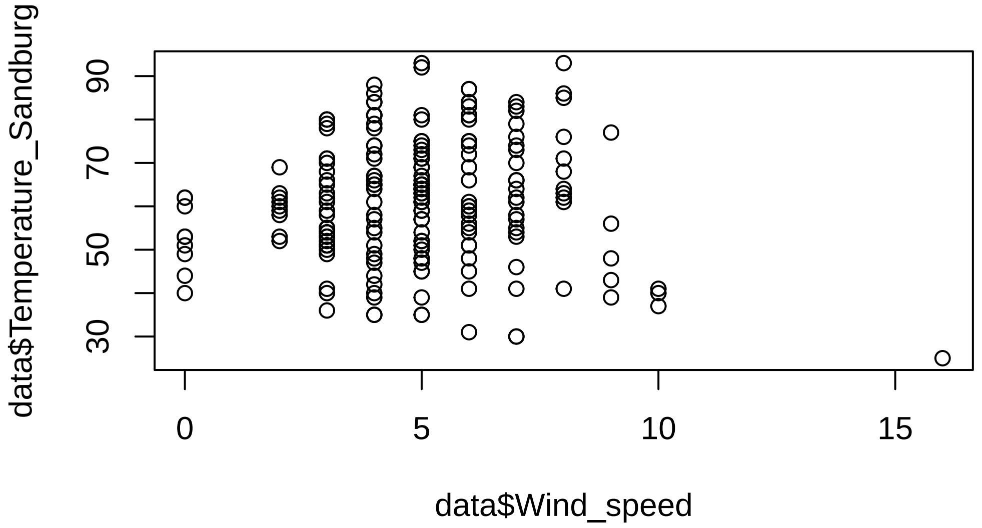
```r
library(patchwork)  # Para combinar gráficos fácilmente

# 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
```



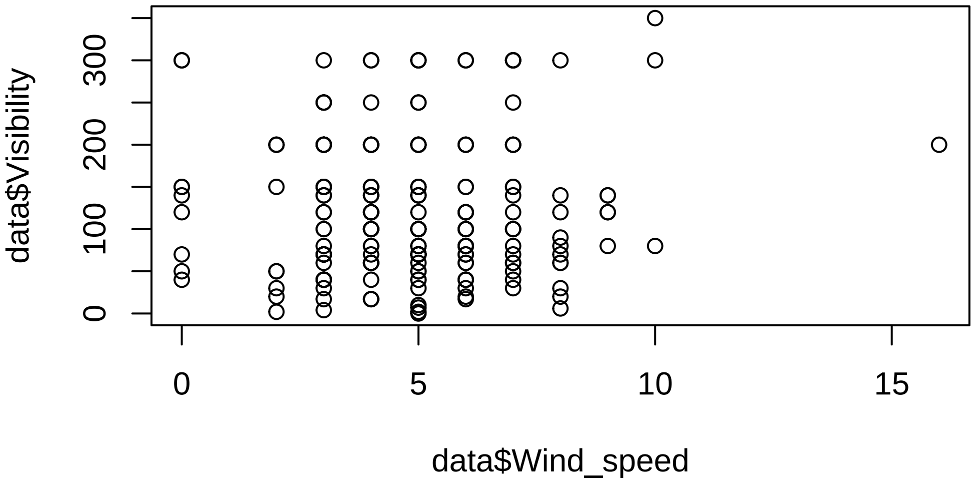
```r
plot(data$Wind_speed,data$Humidity)
```



```
plot(data$Wind_speed,data$Temperature_Sandburg)
```



```
plot(data$Wind_speed,data$Visibility)
```



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
```