

Ejercicio 2.3: Detección y tratamiento de datos atípicos univariante y bivariante (automatización)

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("Funciones_propias.R")
```

Attaching package: 'dplyr'

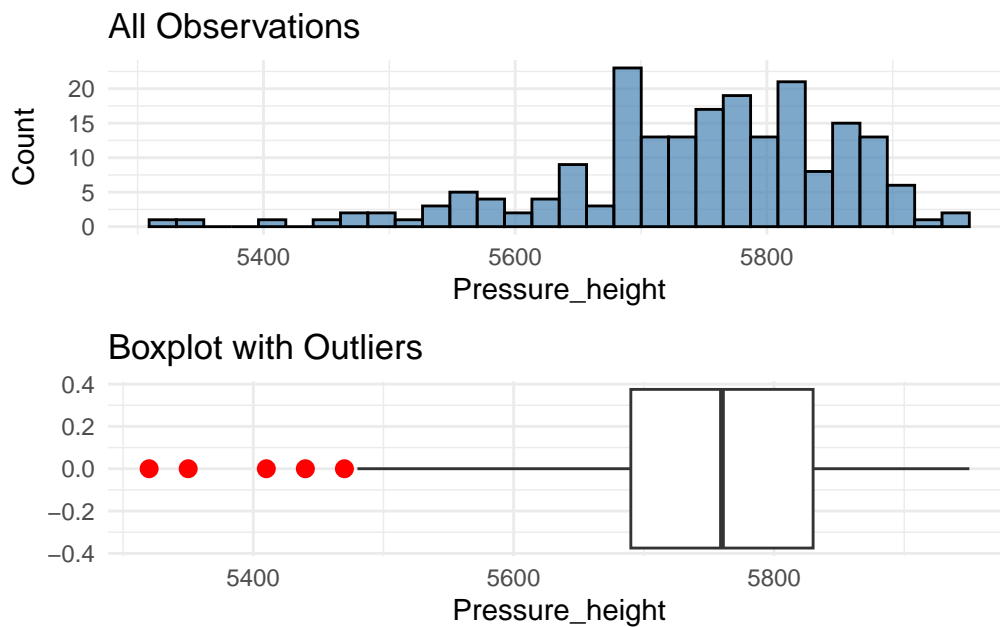
The following objects are masked from 'package:stats':

filter, lag

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

intersect, setdiff, setequal, union

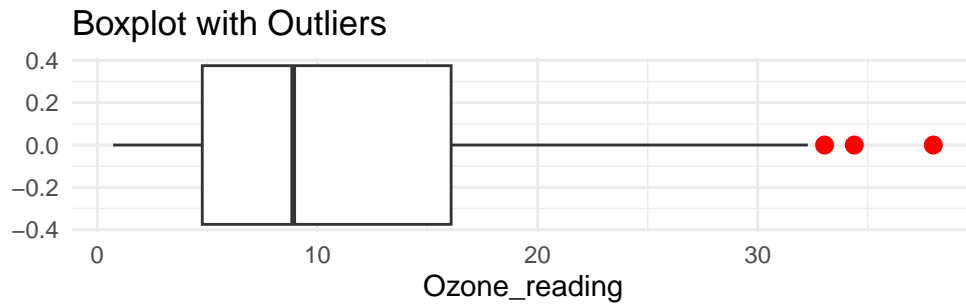
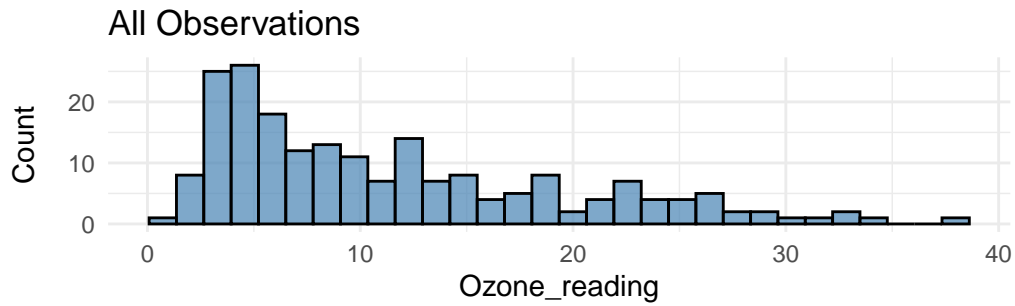
```
outliers(data,"Pressure_height")
```



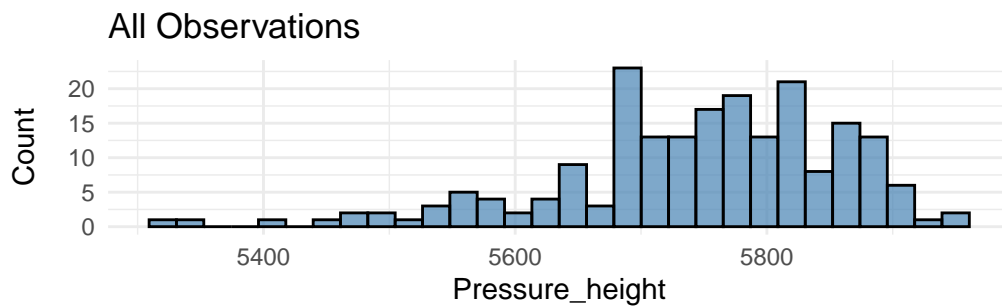
Outliers identified in Pressure_height : 5 outliers
Proportion (%) of outliers: 2.46 %

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

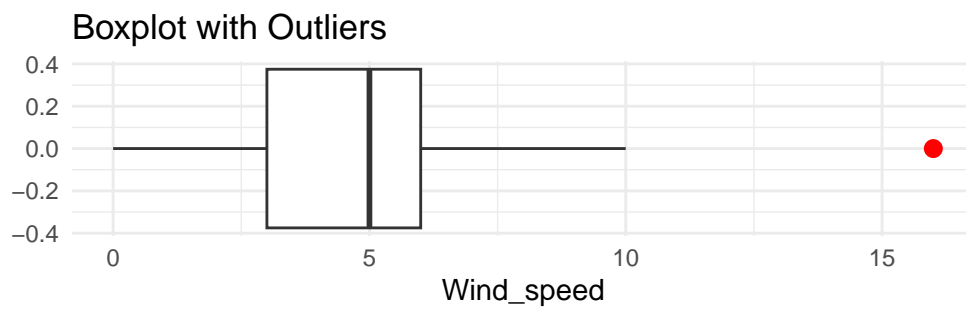
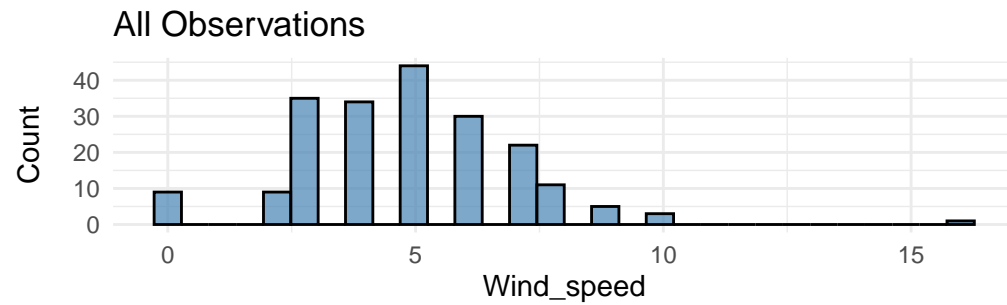
```
# 1. Obtenemos los nombres de las columnas numéricas  
numeric_vars <- names(data)[sapply(data, is.numeric)]  
  
# 2. Usamos una función anónima para pasar 'data' y el 'nombre'  
outliers_results <- lapply(numeric_vars, function(v) outliers(data, v))
```



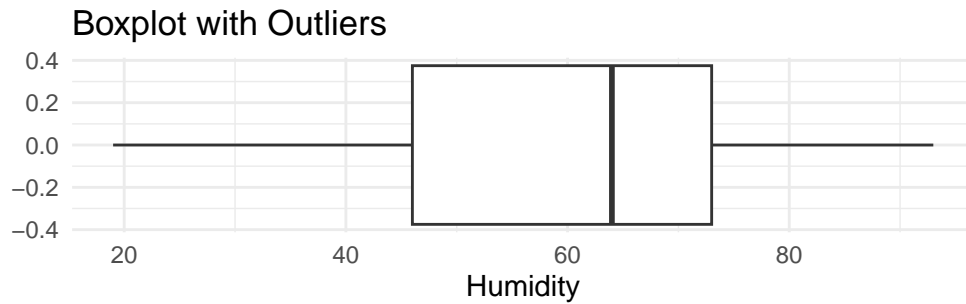
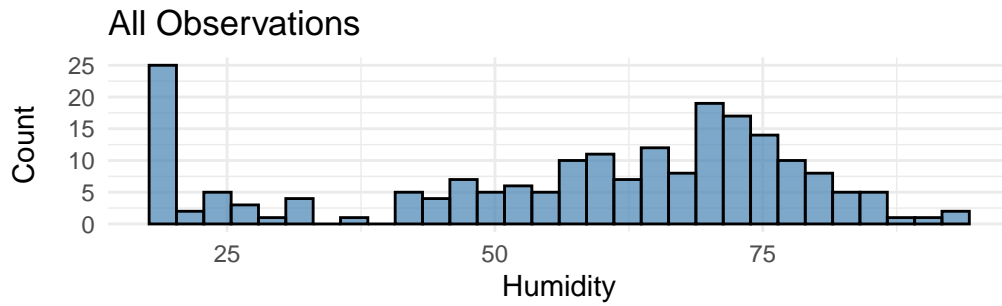
Outliers identified in Ozone_reading : 3 outliers
 Proportion (%) of outliers: 1.48 %



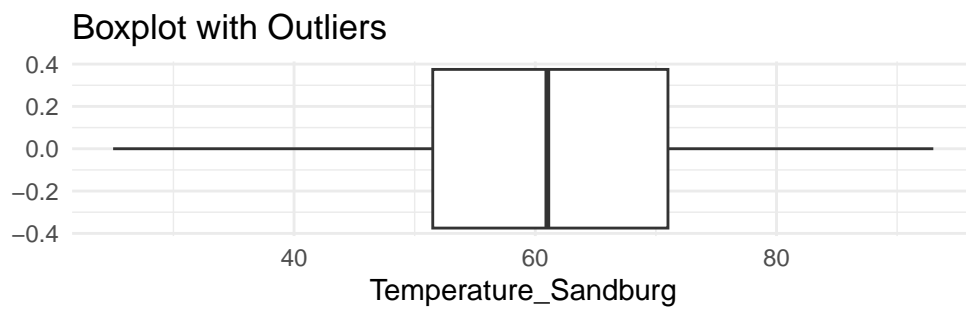
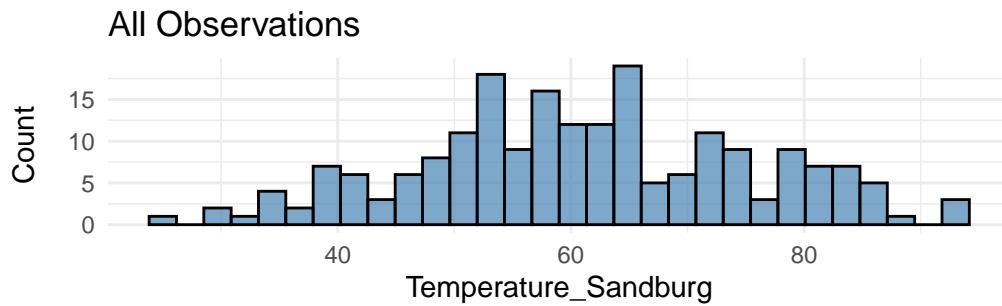
Outliers identified in Pressure_height : 5 outliers
Proportion (%) of outliers: 2.46 %



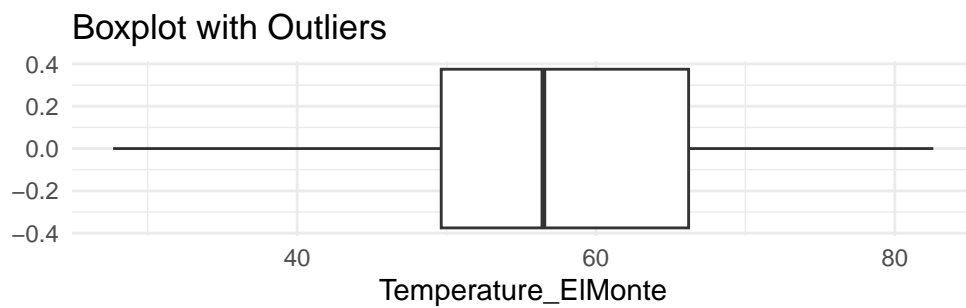
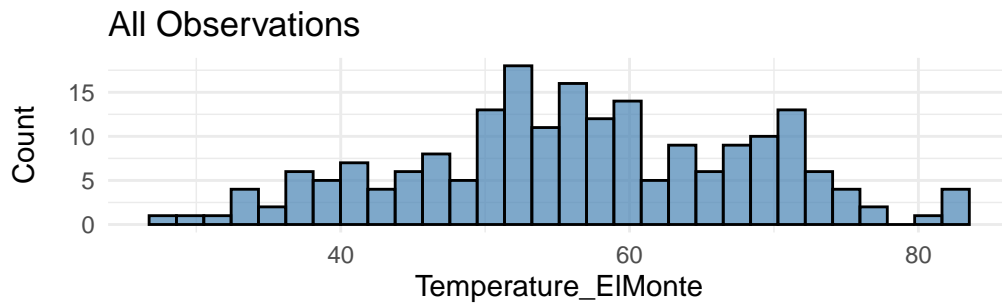
Outliers identified in Wind_speed : 1 outliers
Proportion (%) of outliers: 0.49 %



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



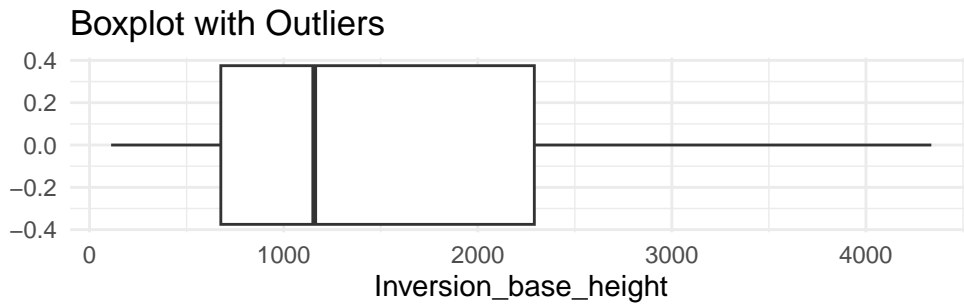
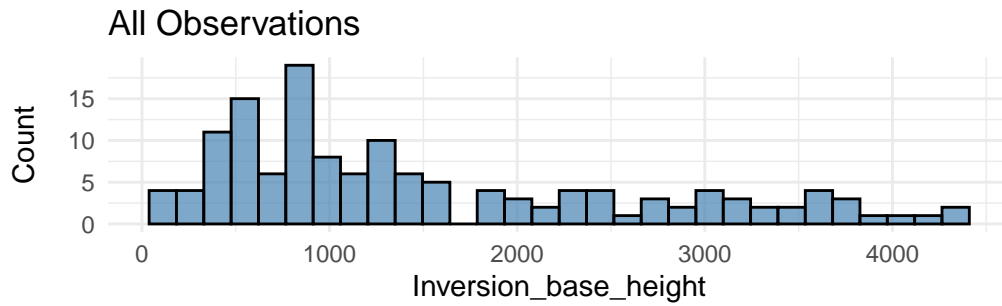
Outliers identified in Temperature_Sandburg : 0 outliers
Proportion (%) of outliers: 0 %



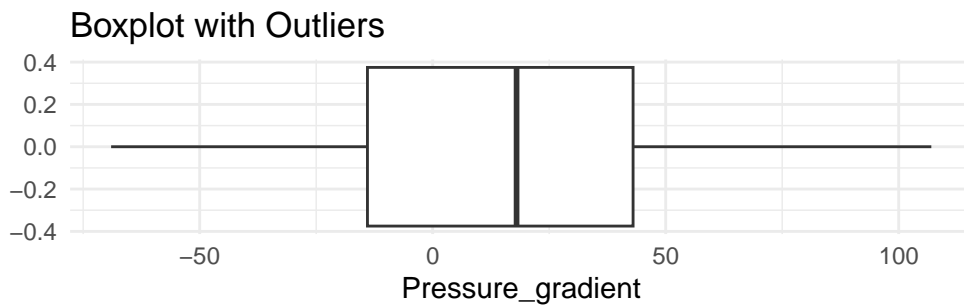
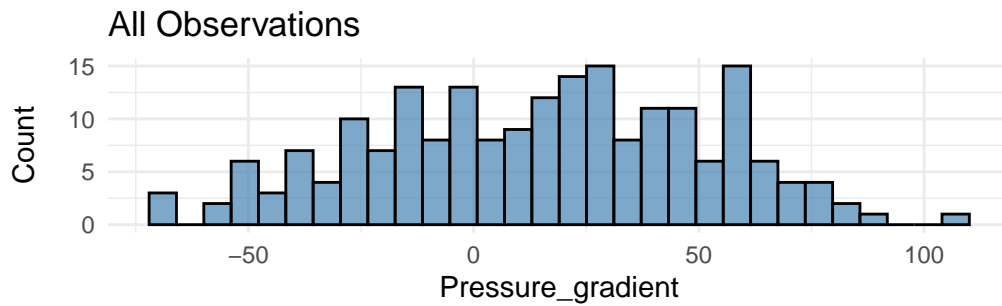
Outliers identified in Temperature_ElMonte : 0 outliers
Proportion (%) of outliers: 0 %

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

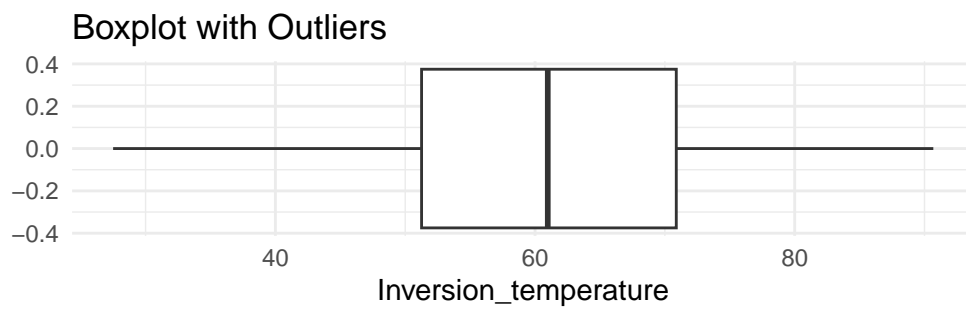
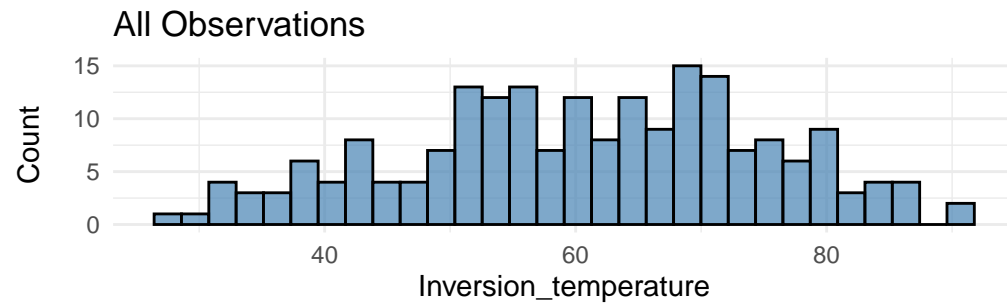
Warning: Removed 63 rows containing non-finite outside the scale range
(`stat_boxplot()`).



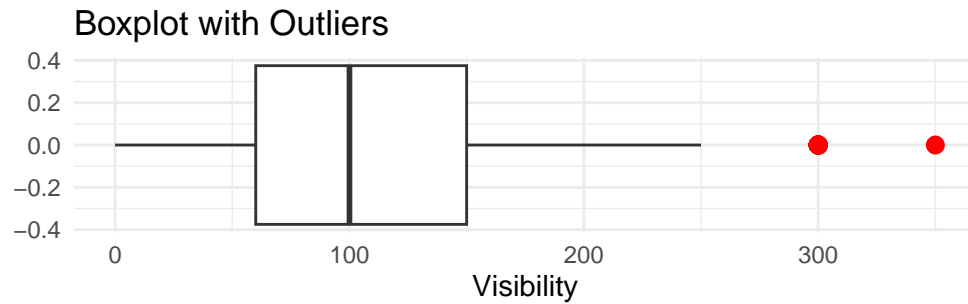
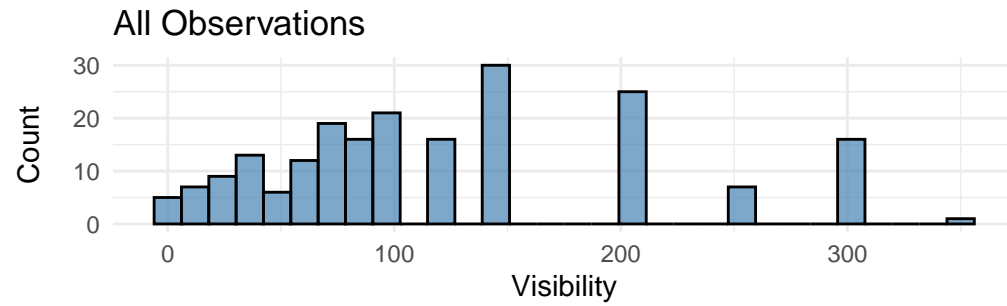
Outliers identified in Inversion_base_height : 0 outliers
Proportion (%) of outliers: 0 %



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

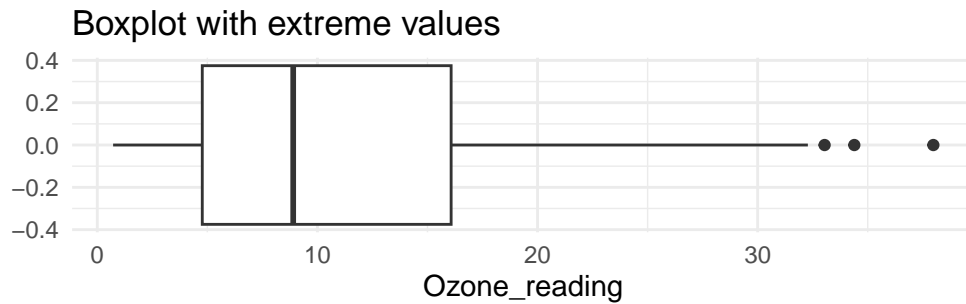
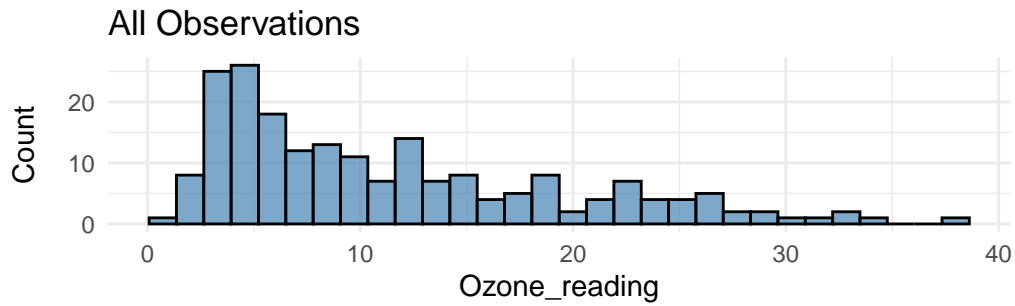


Outliers identified in Inversion_temperature : 0 outliers
Proportion (%) of outliers: 0 %

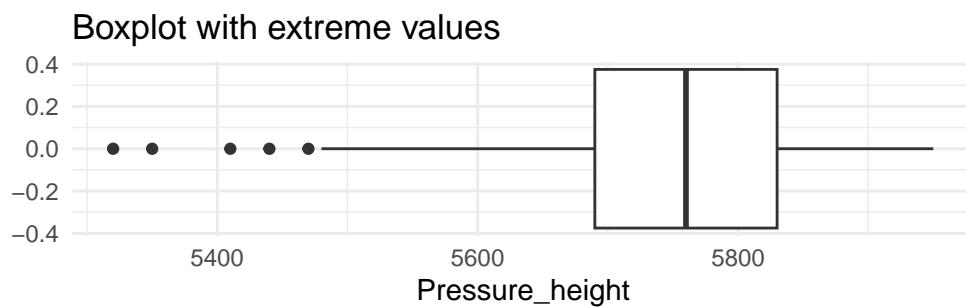
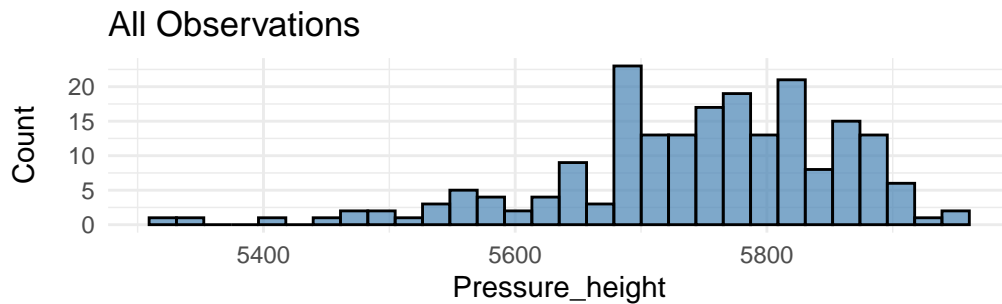


Outliers identified in Visibility : 17 outliers
 Proportion (%) of outliers: 8.37 %

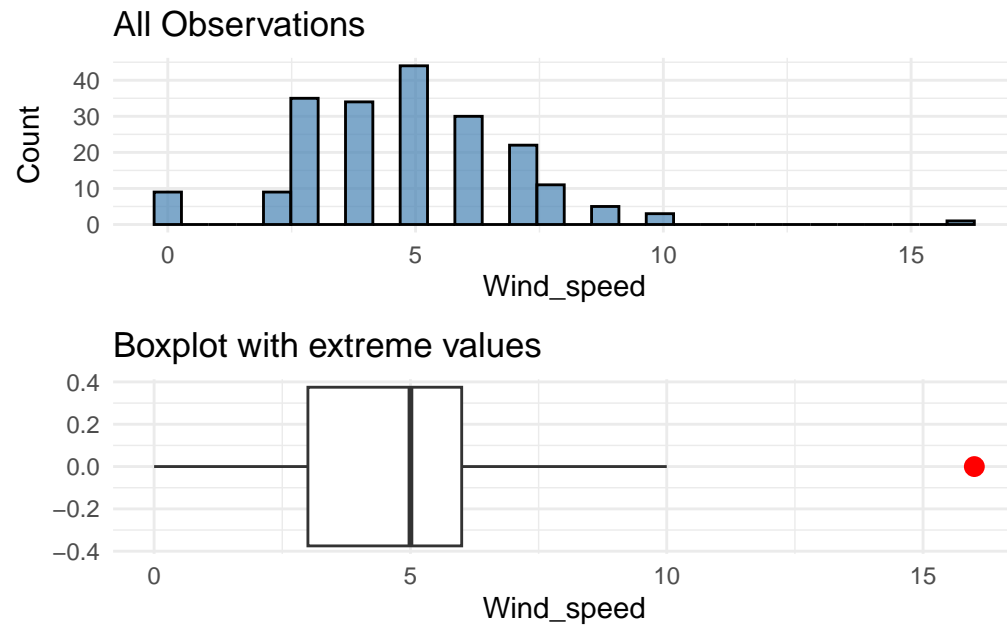
```
extreme_results <- lapply(numeric_vars, function(v) extreme(data, v))
```



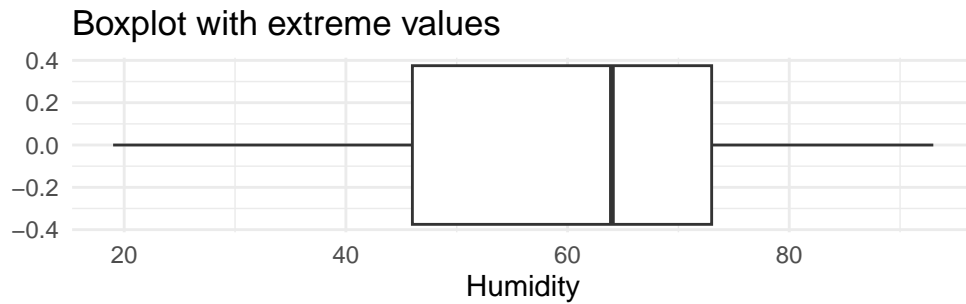
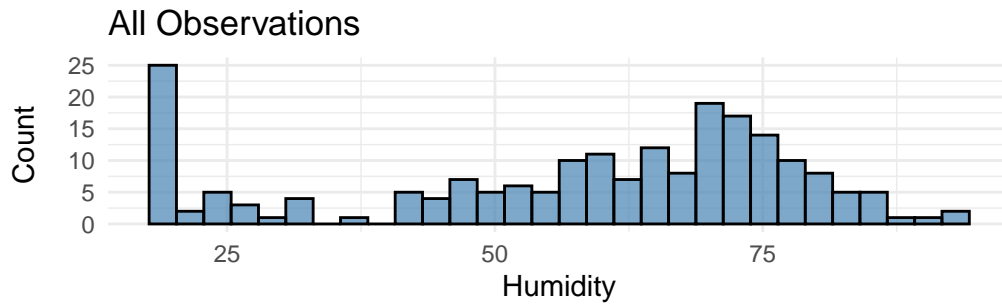
Extremes identified in Ozone_reading : 0 extreme values
 Proportion (%) of extreme values: 0 %



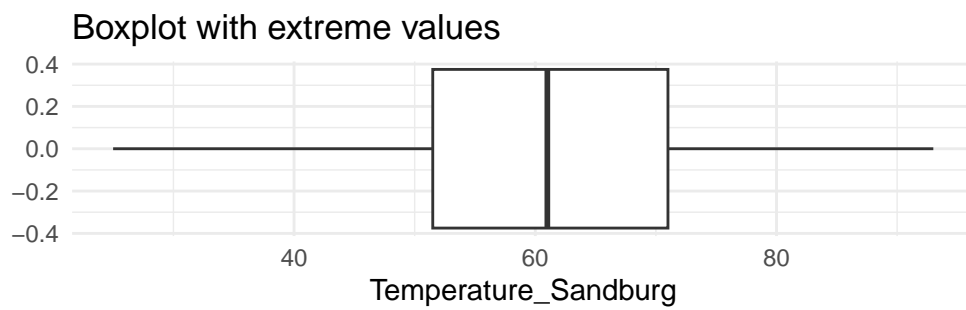
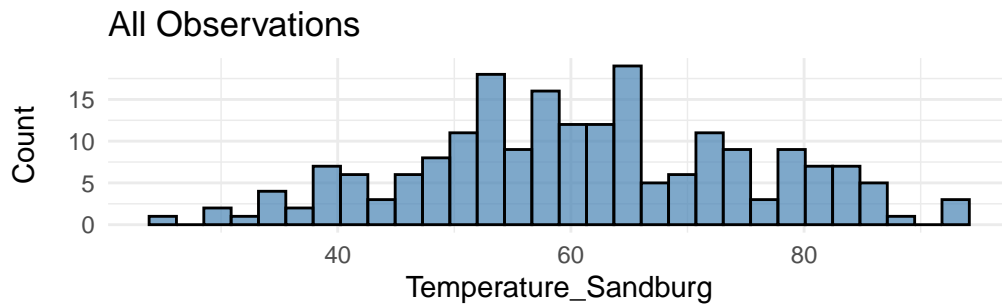
Extremes identified in Pressure_height : 0 extreme values
Proportion (%) of extreme values: 0 %



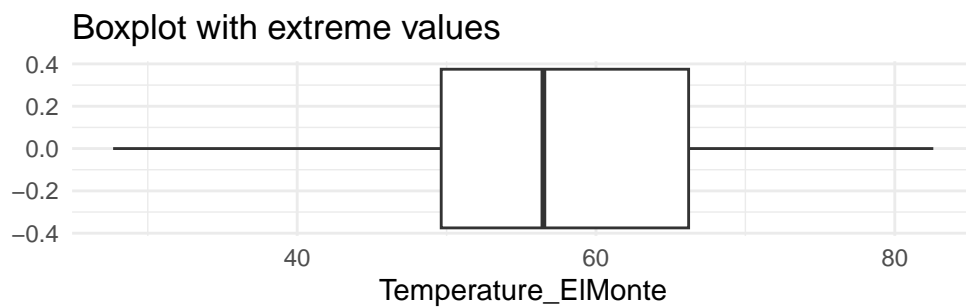
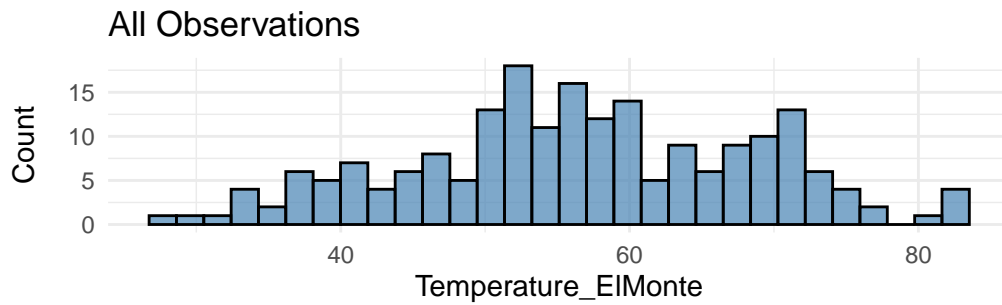
Extremes identified in Wind_speed : 1 extreme values
Proportion (%) of extreme values: 0.49 %



Extremes identified in Humidity : 0 extreme values
 Proportion (%) of extreme values: 0 %

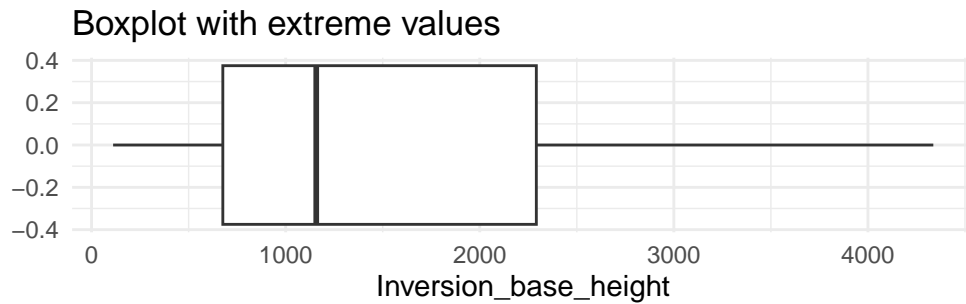
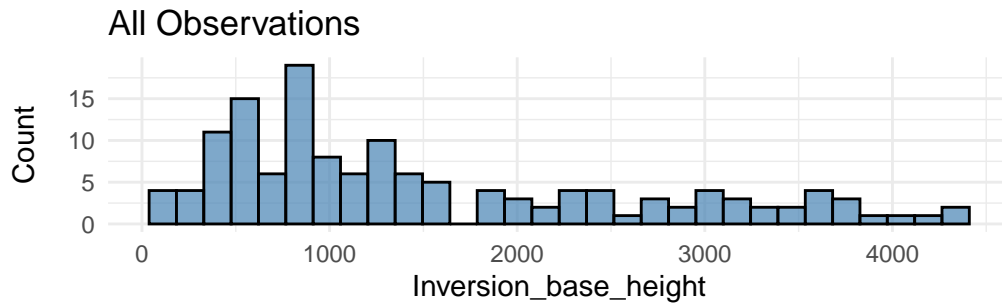


Extremes identified in Temperature_Sandburg : 0 extreme values
Proportion (%) of extreme values: 0 %

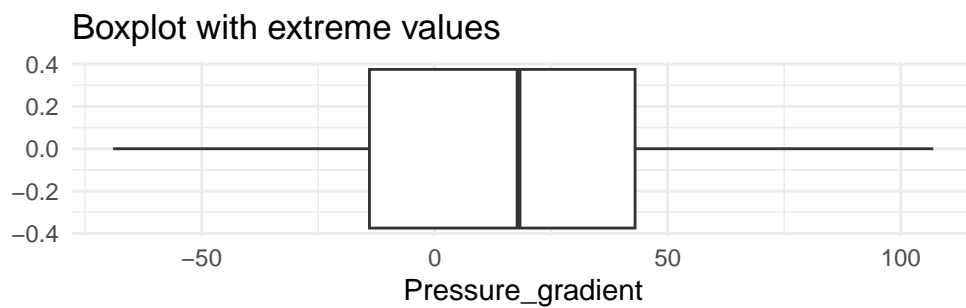
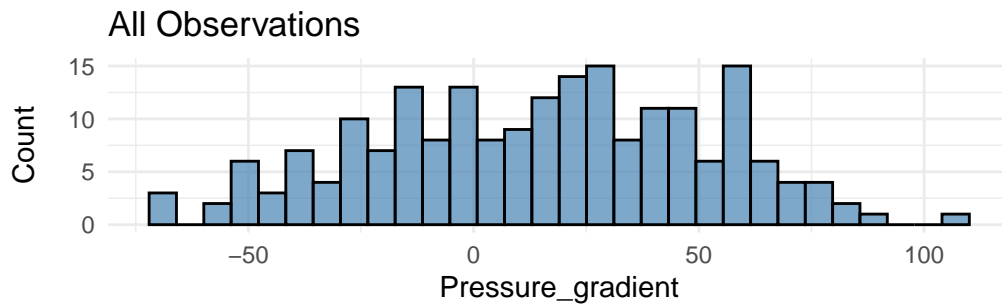


Extremes identified in Temperature_ElMonte : 0 extreme values
Proportion (%) of extreme values: 0 %

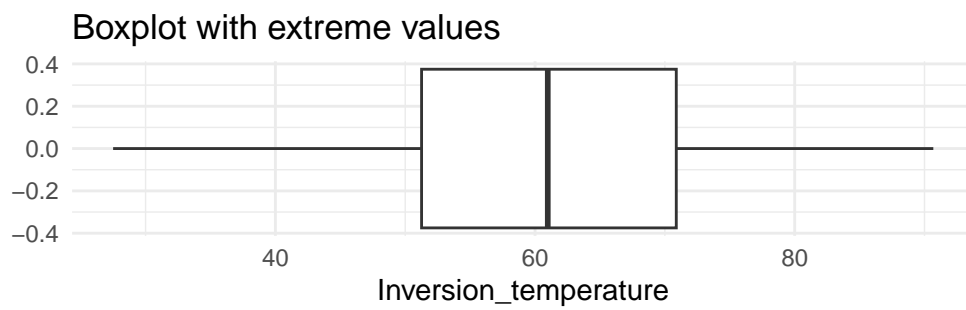
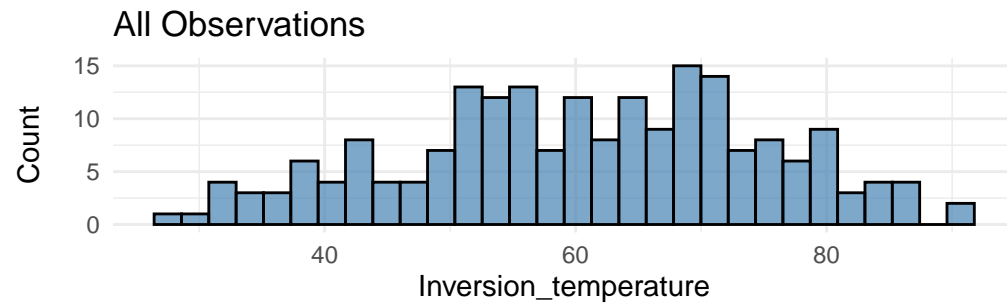
Warning: Removed 63 rows containing non-finite outside the scale range (``stat_bin()``).
Removed 63 rows containing non-finite outside the scale range
(``stat_boxplot()``).



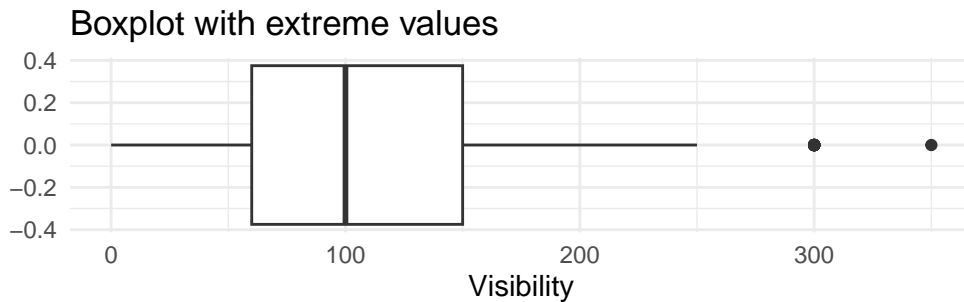
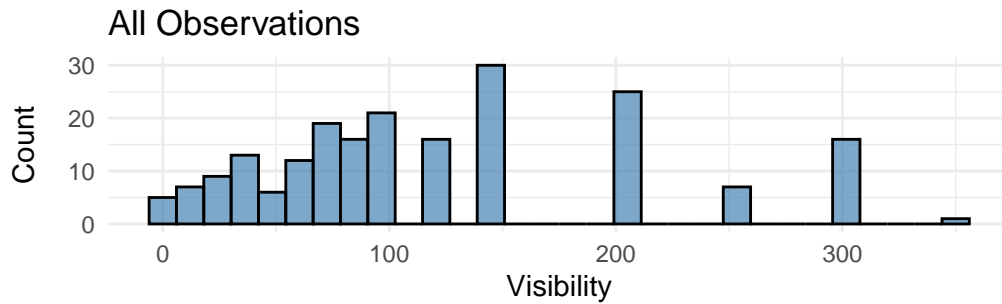
Extremes identified in Inversion_base_height : 0 extreme values
Proportion (%) of extreme values: 0 %



Extremes identified in Pressure_gradient : 0 extreme values
Proportion (%) of extreme values: 0 %



Extremes identified in Inversion_temperature : 0 extreme values
Proportion (%) of extreme values: 0 %



Extremes identified in Visibility : 0 extreme values
 Proportion (%) of extreme values: 0 %

Las variables con datos atípicos son:

Pressure_height (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 distribución

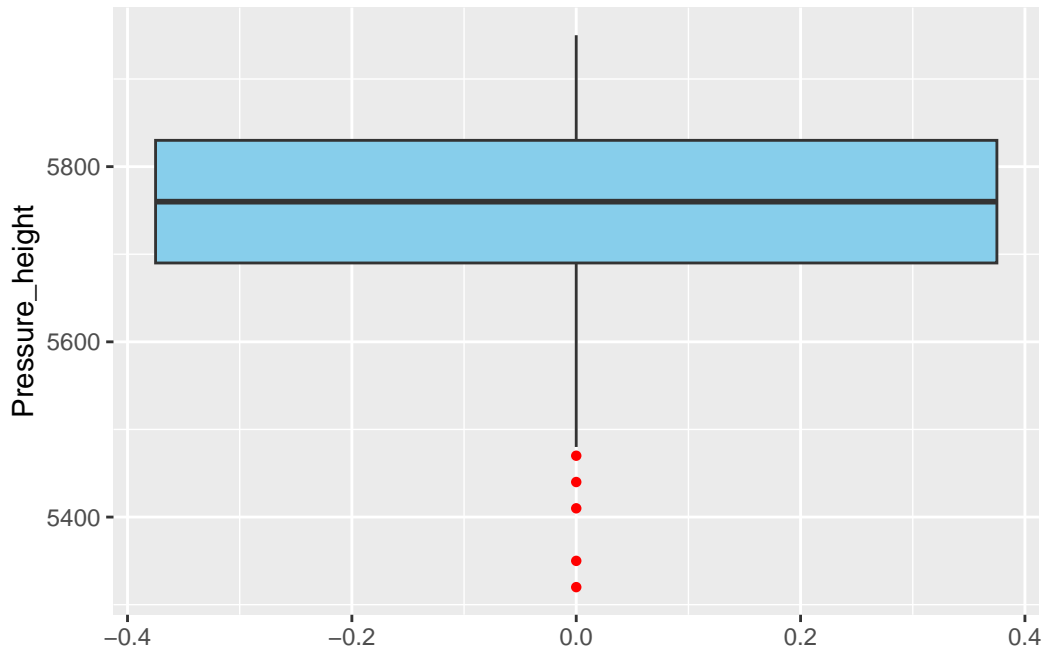
Visibility (8.37%): Valores que corresponden a los mismos valores de 300 y 350 que parecen claramente parte de la variable. Además es un número muy elevado como para ser dato atípico.

De todos los outliers el único que se ve como extremo es un valor en **Wind_speed**

Vamos por tanto a realizar el estudio bivalente de **Pressure_height**, **Ozone_reading** y **Wind_speed**

Estudio de la variable Pressure_height

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



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

	Month	Day_of_month	Day_of_week	Ozone_reading	Pressure_height	Wind_speed
21	2	5	4	2.94	5410	6
22	2	6	5	2.74	5350	7
36	3	2	2	3.22	5470	7
37	3	3	3	2.79	5320	16
64	4	13	2	3.65	5440	5

	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height
21	64	31	32.18	NA
22	62	30	32.54	1341
36	46	30	29.66	NA

37	45	25	27.68	NA
64	44	35	33.08	NA
	Pressure_gradient	Inversion_temperature	Visibility	
21	28	32.36	200	
22	18	45.86	60	
36	44	29.30	300	
37	39	27.50	200	
64	24	32.54	80	

```
### 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)
```

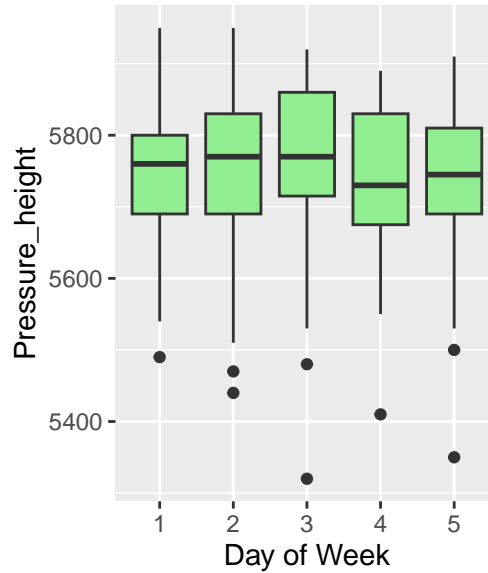
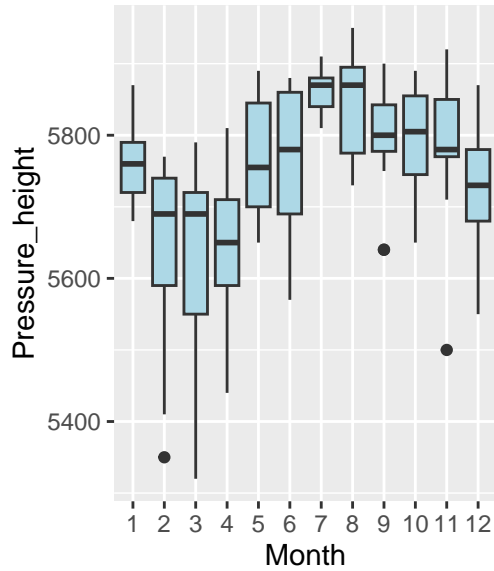
```
library(patchwork) # Para combinar gráficos fácilmente

# 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")

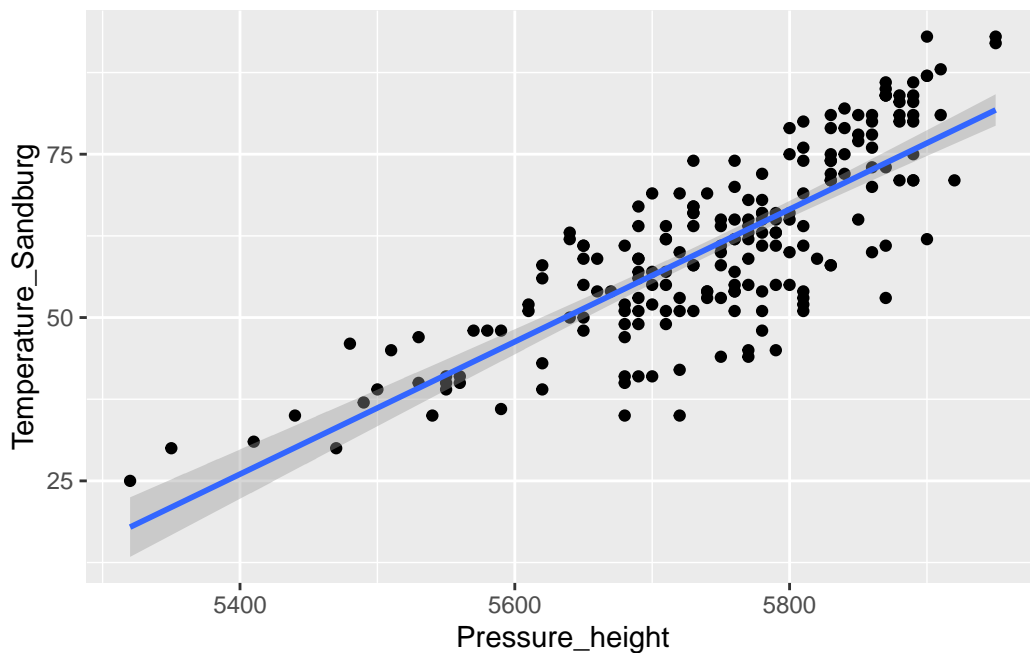
# Gráfico 2: Pressure Height por día de la semana
p2 <- ggplot(data, aes(x = as.factor(Day_of_week), y = Pressure_height)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Pressure_height for days of week", x = "Day of Week", y = "Pressure_height")

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

Pressure_height across months Pressure_height for days o



```
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

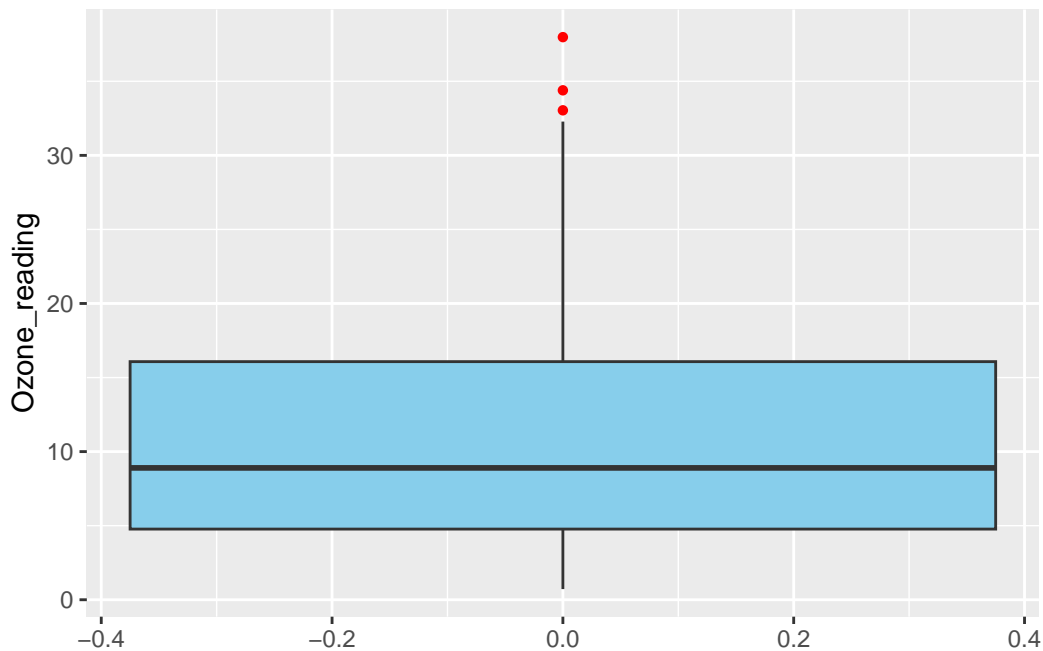
Esta variable está claramente asociada con los meses del año, perteneciendo los valores más

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,]
```

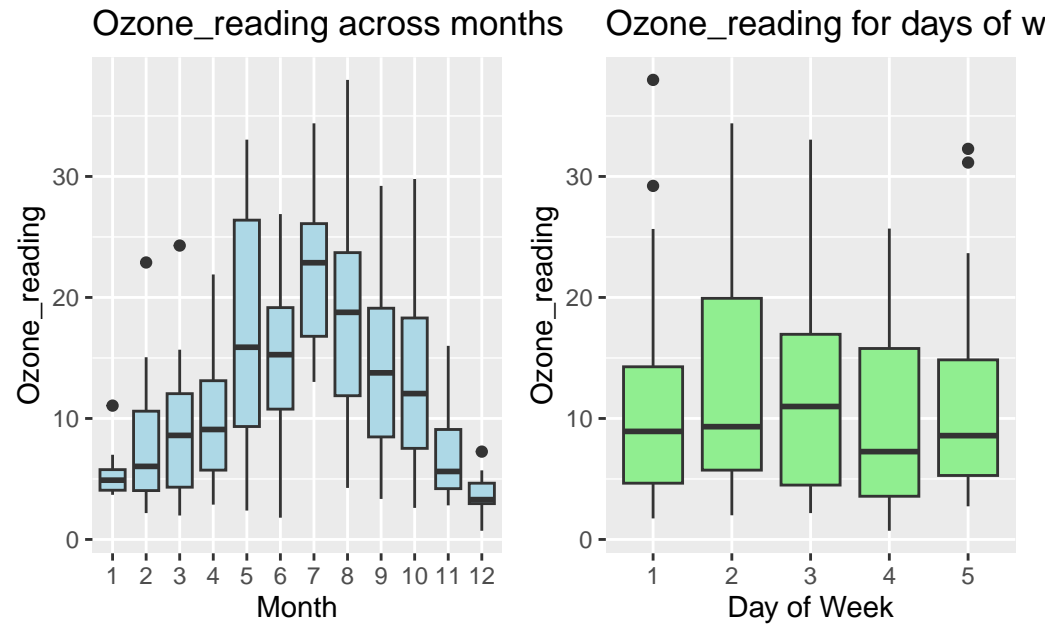
```
[1] Month           Day_of_month      Day_of_week
[4] Ozone_reading    Pressure_height    Wind_speed
[7] Humidity         Temperature_Sandburg Temperature_ElMonte
[10] Inversion_base_height Pressure_gradient Inversion_temperature
[13] Visibility
<0 rows> (or 0-length row.names)
```

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

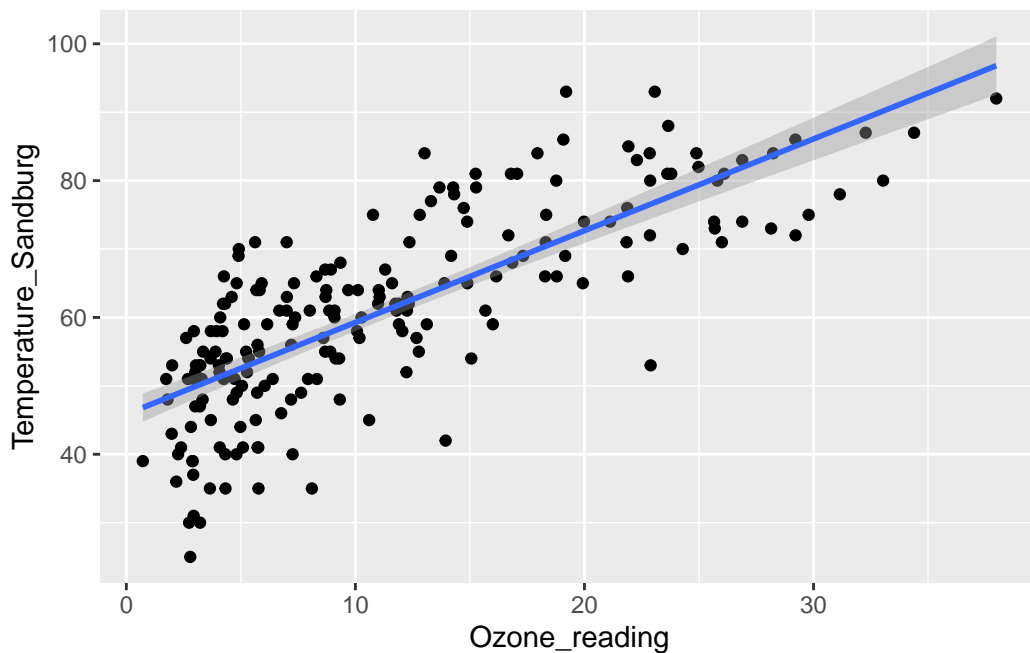
# Gráfico 1: Ozone_reading por mes
p1 <- ggplot(data, aes(x = as.factor(Month), y = Ozone_reading)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Ozone_reading across months", x = "Month", y = "Ozone_reading")

# 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
```



```
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

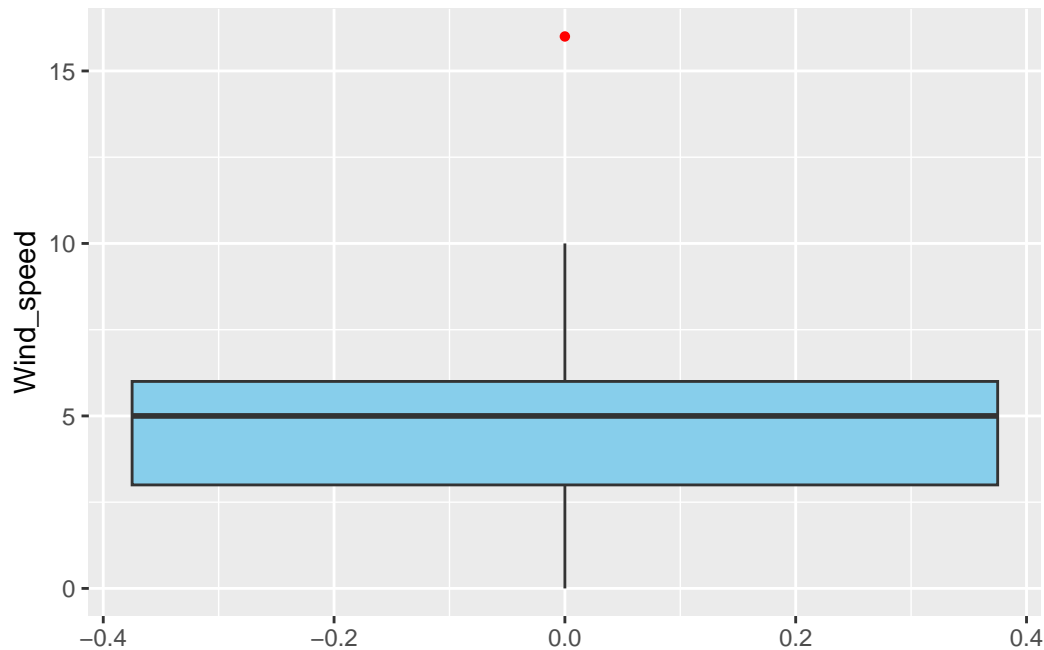
De la misma forma que la variable anterior, esta variable está claramente asociada con los

meses del año, perteneciendo los valores más altos de esta variable a los meses de verano. Además vemos una clara asociación 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,]
```

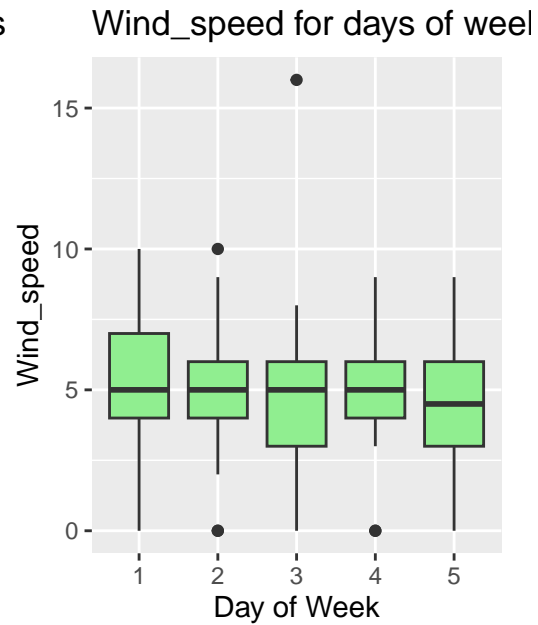
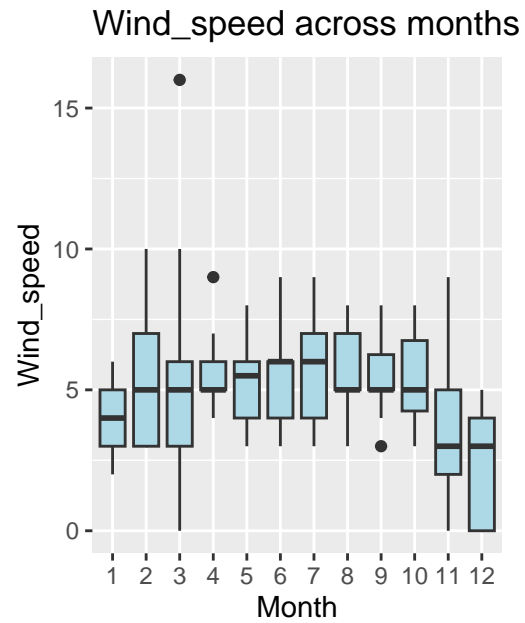
	Month	Day_of_month	Day_of_week	Ozone_reading	Pressure_height	Wind_speed
37	3	3	3	2.79	5320	16
	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height		
37	45		25	27.68		NA
	Pressure_gradient	Inversion_temperature	Visibility			
37		39		27.5		200

```
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")

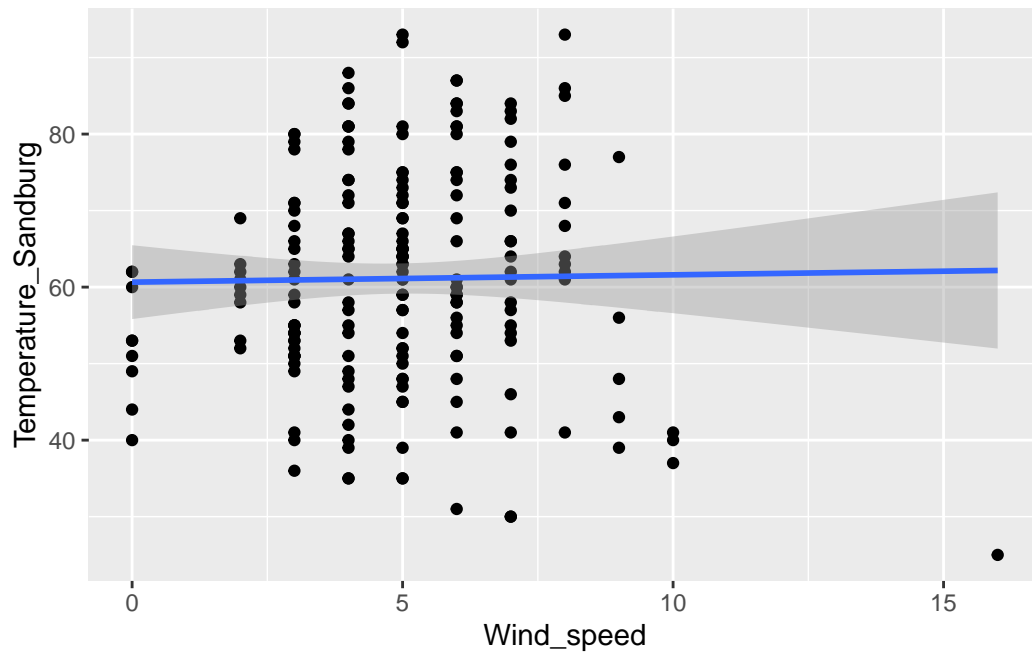
# Gráfico 2: Pressure Height por día de la semana
p2 <- ggplot(data, aes(x = as.factor(Day_of_week), y = Wind_speed)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Wind_speed for days of week", x = "Day of Week", y = "Wind_speed")

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

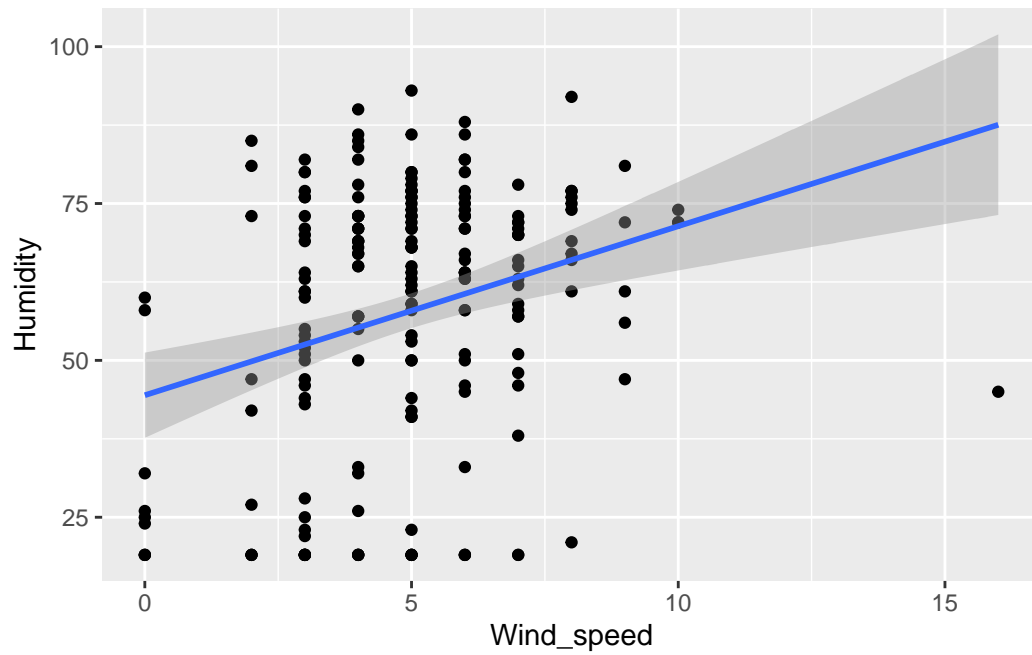


Comprobamos con alguna cuantitativa

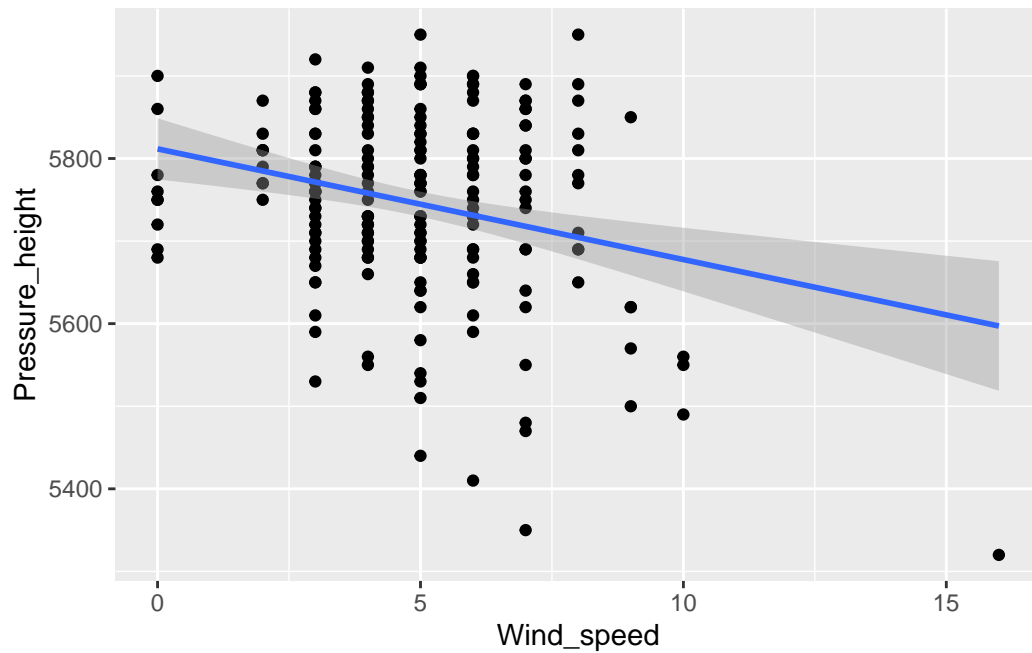
```
ggp <- ggplot(data, aes(Wind_speed, Temperature_Sandburg)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



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



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



En este caso vemos que el outlier de `wind_speed` no está asociado 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
```