informe

Adrián Calzadilla González

12 de diciembre de 2016

Regresión data set Wizmir

Objetivo

Este fichero contiene la información del tiempo de la ciudad de *Esmira* (*Izmir*), que es el segundo mayor puerto de Turquía tras Estambul y la tercera ciudad del país en población con 4.168.415 habitantes en 2015. El fichero posee la información del tiempo correspondiente al periodo *01/01/1994* hasta el *31/12/1997*. A partir de estas características el objetivo es predecir la temperatura media.

Análisis de datos

Hipótesisi inicial

La temperatura media se trata de promedios estadísticos cogidos a lo largo del día, por lo tanto es muy posible que la relación entre la temperatura media y las temperaturas máximas y mínimas sea importante para realizar un buen modelo.

Análisis general

Para analizar el conjunto de datos proporcionado. En primer lugar se lee y exporta el fichero con los datos de *wizmir*, se introducen los nombres de las variables a mano y se comprueba si es un *data.frame*.

```
wizmir <- read.csv2("./wizmir/wizmir.dat", header = F, sep = ",", comment.char = "@", dec = ".")
names(wizmir) <- c("Max_temperature", "Min_temperature", "Dewpoint", "Precipitation", "Sea_level_pressu
is.data.frame(wizmir)
```

```
## [1] TRUE
```

Seguidamente, se analiza el *data.frame* y el tipo de variables que contiene:

```
## Características generales de wizmir
attach(wizmir)
is.data.frame(wizmir)
```

```
## [1] TRUE
```

```
dim(wizmir)
```

```
## [1] 1461   10
```

```
str(wizmir)
```

```
## 'data.frame':    1461 obs. of  10 variables:
##  $ Max_temperature   : num  88.2 88 91.6 64.4 94.1 81.3 62.6 53.6 53.4 91.4 ...
##  $ Min_temperature   : num  57.2 58.6 62.1 42.8 72.3 62.6 53.6 35.6 44.1 59.9 ...
##  $ Dewpoint          : num  53.6 54.9 60.4 37.4 46.8 37.4 51.4 30.3 45.4 54.1 ...
##  $ Precipitation     : num  0 0 0 0.2 0 0 0.2 0 0 0 ...
##  $ Sea_level_pressure: num  30 29.8 29.8 30.1 29.9 ...
##  $ Standard_pressure : num  7.3 7.3 7.2 7.8 7.2 6.8 7.9 7.6 7.7 7.3 ...
##  $ Visibility        : num  9.09 10.7 8.29 21.1 17.2 21.2 17 16.1 9.09 7.37 ...
##  $ Wind_speed        : num  16.1 18.3 18.3 27.5 25.3 27.5 25.3 20.8 16.1 20.8 ...
##  $ Max_wind_speed    : num  34.3 34.3 34.3 34.3 34.3 ...
##  $ Mean_temperature  : num  74.3 75.2 76.1 47.1 83.9 70.8 59.7 44.9 47.8 77.8 ...
```

```
### Existen valores "NA"
table(is.na(wizmir))
```

```
##
## FALSE
## 14610
```

```
### Tipos de datos de las variables
is.double(Max_temperature)
```

```
## [1] TRUE
```

```
is.double(Min_temperature)
```

```
## [1] TRUE
```

```
is.double(Dewpoint)
```

```
## [1] TRUE
```

```
is.double(Precipitation)
```

```
## [1] TRUE
```

```
is.double(Sea_level_pressure)
```

```
## [1] TRUE
```

```
is.double(Standard_pressure)
```

```
## [1] TRUE
```

```
is.double(Visibility)
```

```
## [1] TRUE
```

```r
is.double(Wind_speed)
```

```
## [1] TRUE
```

```r
is.double(Max_wind_speed)
```

```
## [1] TRUE
```

```r
is.double(Mean_temperature)
```

```
## [1] TRUE
```

Resumen del *data.frame* en el que se nos muestra el valor máximo y mínimo registrado, el primer cuartil, la mediana, el tercer cuartil y la media de cada variable.

```r
summary(wizmir)
```

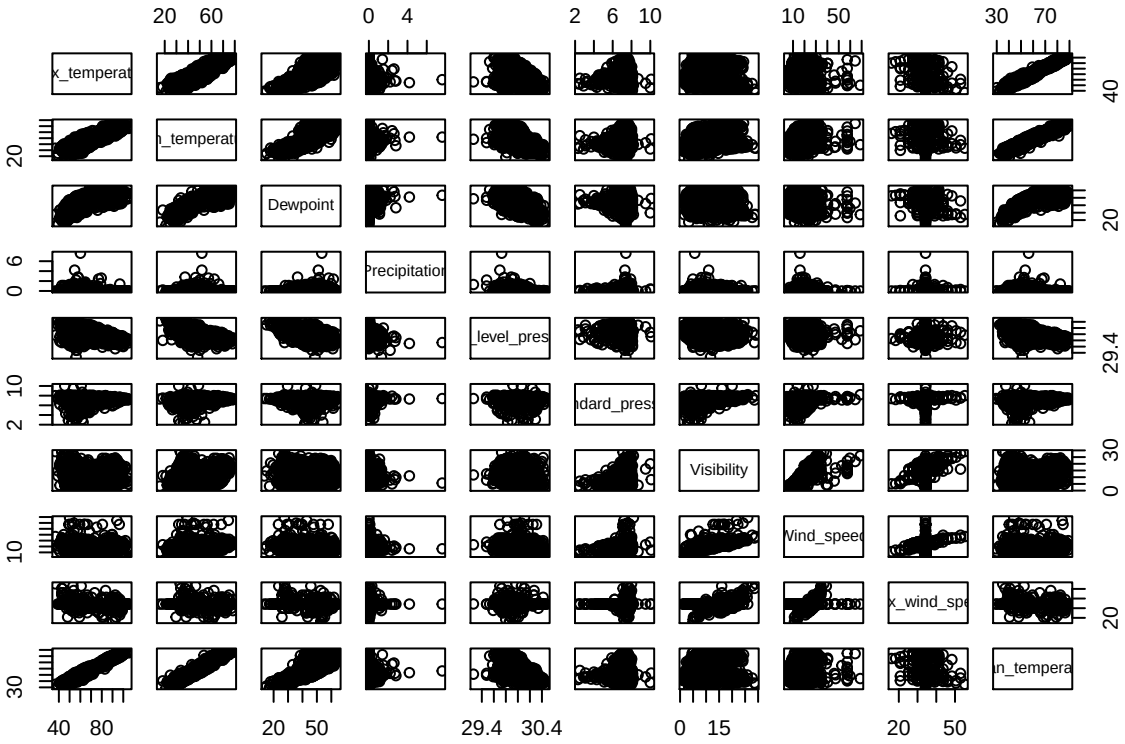
```
##  Max_temperature  Min_temperature    Dewpoint     Precipitation
##  Min.   : 36.70   Min.   :15.80   Min.   :13.60   Min.   :0.00000
##  1st Qu.: 59.00   1st Qu.:40.10   1st Qu.:41.30   1st Qu.:0.00000
##  Median : 70.70   Median :50.00   Median :48.20   Median :0.00000
##  Mean   : 72.22   Mean   :50.74   Mean   :46.62   Mean   :0.09257
##  3rd Qu.: 87.10   3rd Qu.:62.20   3rd Qu.:53.60   3rd Qu.:0.00000
##  Max.   :105.00   Max.   :78.60   Max.   :64.40   Max.   :7.60000
##  Sea_level_pressure Standard_pressure   Visibility      Wind_speed
##  Min.   :29.26      Min.   : 2.300    Min.   : 0.92   Min.   : 4.72
##  1st Qu.:29.85      1st Qu.: 7.100    1st Qu.: 6.56   1st Qu.:16.10
##  Median :29.95      Median : 7.300    Median :10.50   Median :19.81
##  Mean   :29.97      Mean   : 7.197    Mean   :11.16   Mean   :19.81
##  3rd Qu.:30.08      3rd Qu.: 7.600    3rd Qu.:15.40   3rd Qu.:23.00
##  Max.   :30.48      Max.   :10.100    Max.   :29.10   Max.   :68.80
##  Max_wind_speed  Mean_temperature
##  Min.   :16.11   Min.   :29.40
##  1st Qu.:34.28   1st Qu.:49.60
##  Median :34.28   Median :60.00
##  Mean   :34.28   Mean   :61.51
##  3rd Qu.:34.28   3rd Qu.:75.20
##  Max.   :55.24   Max.   :89.90
```

A continuación, se calculan algunas métricas que no salen en *summary*:

```r
## Desviación estandar para cada una de las variables
allSd <- apply(wizmir, 2, sd)
## Mediana de cada una de las variables
allMedian <- apply(wizmir, 2, median)
## Rango intercuartílico
allIQR <- apply(wizmir, 2, IQR)
```

Por último, se grafican todas las variables todas con todas.

```r
# Gráfica de todos con todos
plot(wizmir)
```



Variables

Conjunto de datos explicado uno a uno.

Se ha implementado una función para el calculo de la *moda* esta es:

```r
## Función para calcular la moda

getMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Max_temperature

Esta variable se utiliza para guardar la temperatura máxima obtenida en cada observación. Su unidad es el grado *Farenheit*.

Medidas de centralidad

```r
mean(Max_temperature)
```

```
## [1] 72.22416
```

```r
median(Max_temperature)
```

```
## [1] 70.7
```

```r
getMode(Max_temperature)
```

```
## [1] 64.4
```

Medidas de dispersión

```r
var(Max_temperature)
```

```
## [1] 253.6602
```

```r
sd(Max_temperature)
```

```
## [1] 15.92671
```

```r
max(Max_temperature)
```

```
## [1] 105
```

```r
min(Max_temperature)
```

```
## [1] 36.7
```

```r
range(Max_temperature)
```

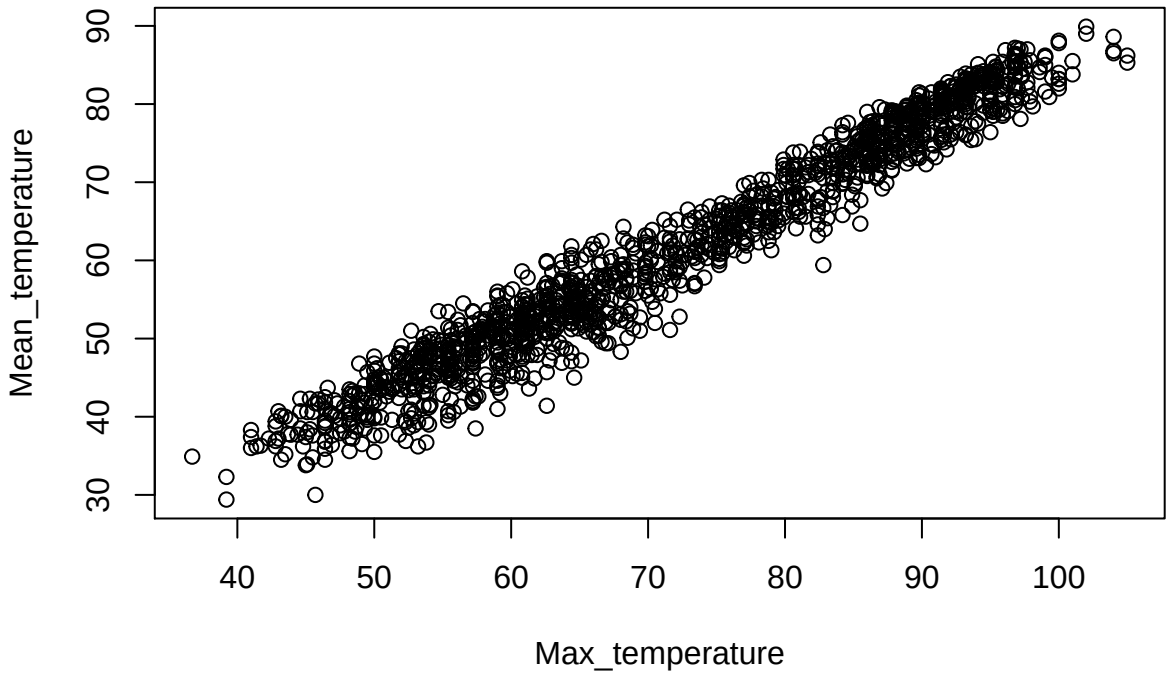
```
## [1]  36.7 105.0
```

```r
quantile(Max_temperature)
```

```
##    0%   25%   50%   75%  100%
##  36.7  59.0  70.7  87.1 105.0
```

Gráfica de Max_temperature respecto a Mean_temperature, la salida.

```r
plot(Max_temperature, Mean_temperature)
```



Min_temperature

Medidas de centralidad

```
mean(Min_temperature)
```

```
## [1] 50.74025
```

```
median(Min_temperature)
```

```
## [1] 50
```

```
getMode(Min_temperature)
```

```
## [1] 42.8
```

Medidas de dispersión

```
var(Min_temperature)
```

```
## [1] 174.9292
```

```
sd(Min_temperature)
```

```
## [1] 13.22608
```

```r
max(Min_temperature)
```

```
## [1] 78.6
```

```r
min(Min_temperature)
```

```
## [1] 15.8
```

```r
range(Min_temperature)
```

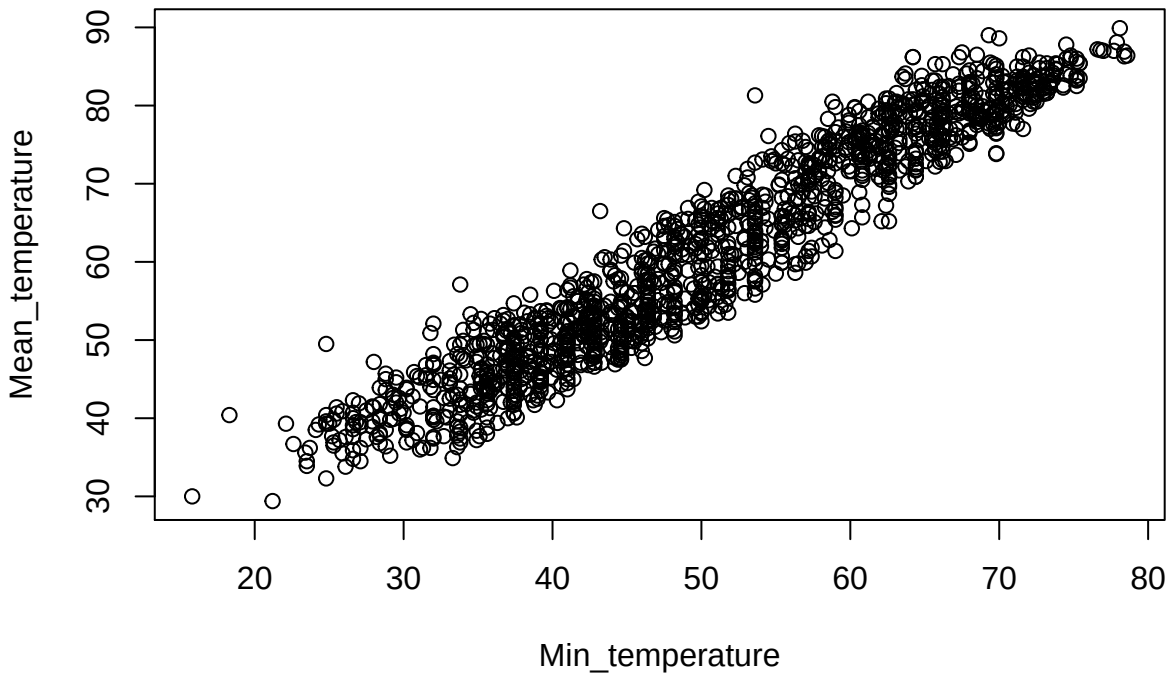
```
## [1] 15.8 78.6
```

```r
quantile(Min_temperature)
```

```
##   0%  25%  50%  75% 100%
## 15.8 40.1 50.0 62.2 78.6
```

Gráfica de Max_temperature respecto a Mean_temperature, la salida.

```r
plot(Min_temperature, Mean_temperature)
```



Dewpoint

Medidas de centralidad

```r
mean(Dewpoint)
```

```
## [1] 46.62356
```

```r
median(Dewpoint)
```

```
## [1] 48.2
```

```r
getMode(Dewpoint)
```

```
## [1] 49
```

Medidas de dispersión

```r
var(Dewpoint)
```

```
## [1] 87.32051
```

```r
sd(Dewpoint)
```

```
## [1] 9.344545
```

```r
max(Dewpoint)
```

```
## [1] 64.4
```

```r
min(Dewpoint)
```

```
## [1] 13.6
```

```r
range(Dewpoint)
```

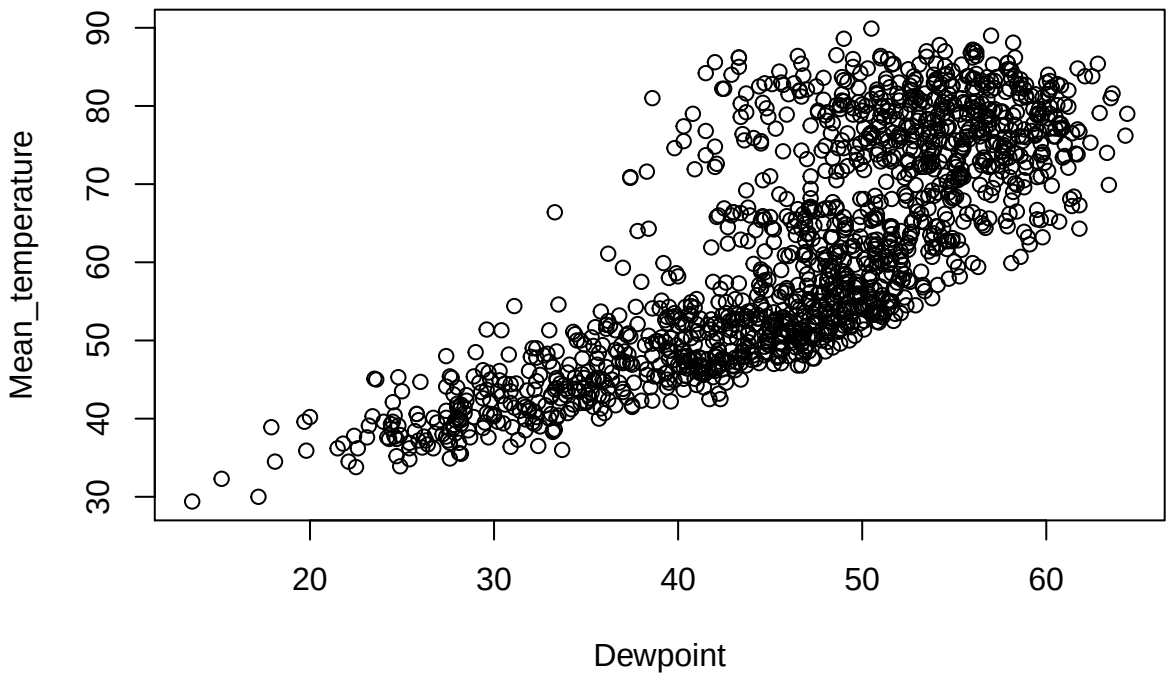
```
## [1] 13.6 64.4
```

```r
quantile(Dewpoint)
```

```
##   0%  25%  50%  75% 100%
## 13.6 41.3 48.2 53.6 64.4
```

Gráfica de Dewpoint respecto a Mean_temperature, la salida.

```
plot(Dewpoint, Mean_temperature)
```



Precipitation

Medidas de centralidad

```
mean(Precipitation)
```

```
## [1] 0.09256674
```

```
median(Precipitation)
```

```
## [1] 0
```

```
getMode(Precipitation)
```

```
## [1] 0
```

Medidas de dispersión

```
var(Precipitation)
```

```
## [1] 0.1244684
```

```r
sd(Precipitation)
```

```
## [1] 0.3528008
```

```r
max(Precipitation)
```

```
## [1] 7.6
```

```r
min(Precipitation)
```

```
## [1] 0
```

```r
range(Precipitation)
```

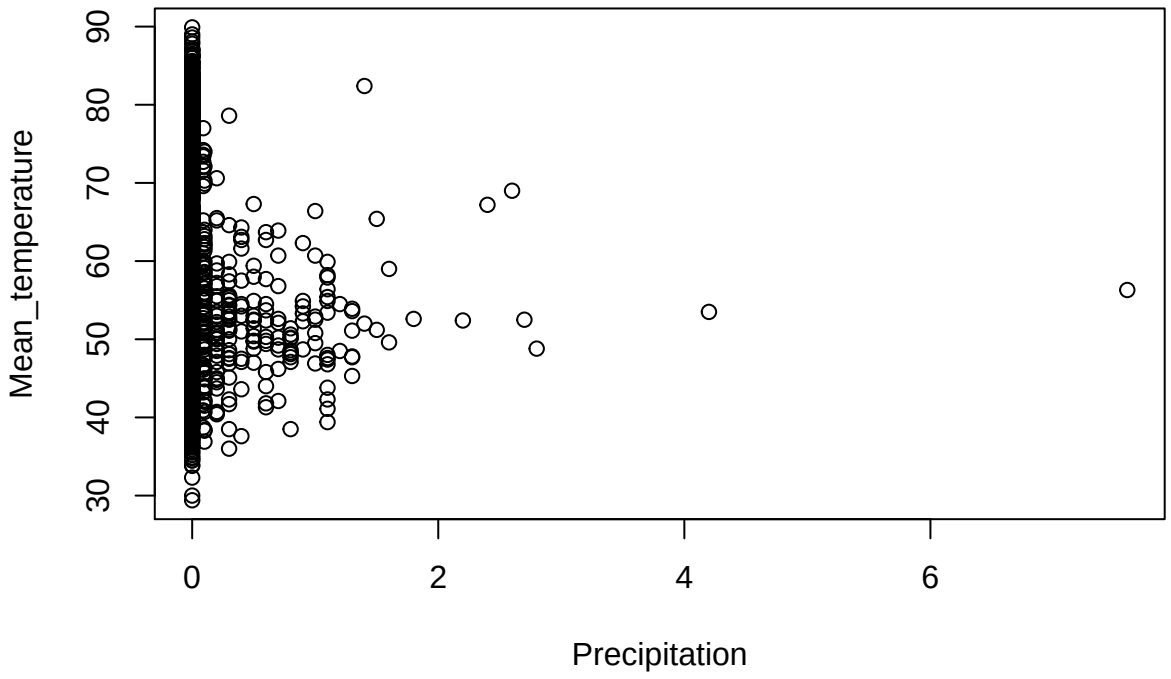
```
## [1] 0.0 7.6
```

```r
quantile(Precipitation)
```

```
##   0%  25%  50%  75% 100%
##  0.0  0.0  0.0  0.0  7.6
```

Gráfica de Precipitation respecto a Mean_temperature, la salida.

```r
plot(Precipitation, Mean_temperature)
```



Sea_level_pressure

Medidas de centralidad

```
mean(Sea_level_pressure)
```

```
## [1] 29.97111
```

```
median(Sea_level_pressure)
```

```
## [1] 29.95
```

```
getMode(Sea_level_pressure)
```

```
## [1] 29.94
```

Medidas de dispersión

```
var(Sea_level_pressure)
```

```
## [1] 0.02811959
```

```
sd(Sea_level_pressure)
```

```
## [1] 0.167689
```

```
max(Sea_level_pressure)
```

```
## [1] 30.48
```

```
min(Sea_level_pressure)
```

```
## [1] 29.26
```

```
range(Sea_level_pressure)
```

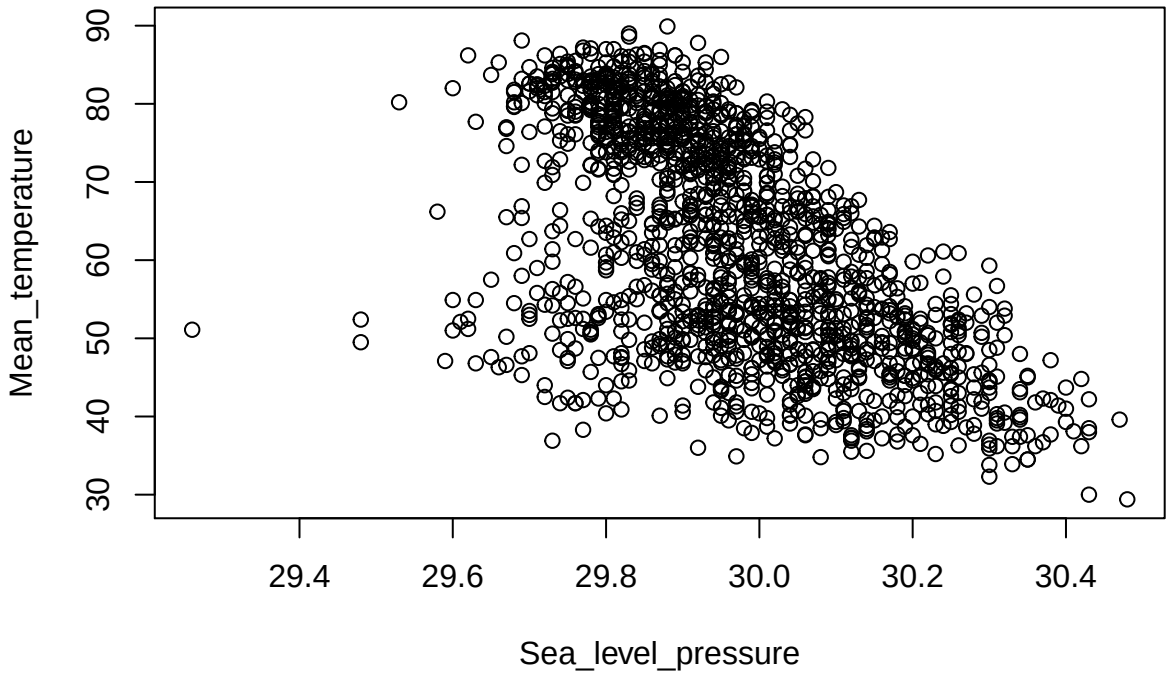
```
## [1] 29.26 30.48
```

```
quantile(Sea_level_pressure)
```

```
##    0%   25%   50%   75%  100%
## 29.26 29.85 29.95 30.08 30.48
```

Gráfica de Sea_level_pressure respecto a Mean_temperature, la salida.

```
plot(Sea_level_pressure, Mean_temperature)
```



Standard_pressure

Medidas de centralidad

```
mean(Standard_pressure)
```

```
## [1] 7.196783
```

```
median(Standard_pressure)
```

```
## [1] 7.3
```

```
getMode(Standard_pressure)
```

```
## [1] 7.2
```

Medidas de dispersión

```
var(Standard_pressure)
```

```
## [1] 0.4691609
```

```r
sd(Standard_pressure)
```

```
## [1] 0.6849532
```

```r
max(Standard_pressure)
```

```
## [1] 10.1
```

```r
min(Standard_pressure)
```

```
## [1] 2.3
```

```r
range(Standard_pressure)
```

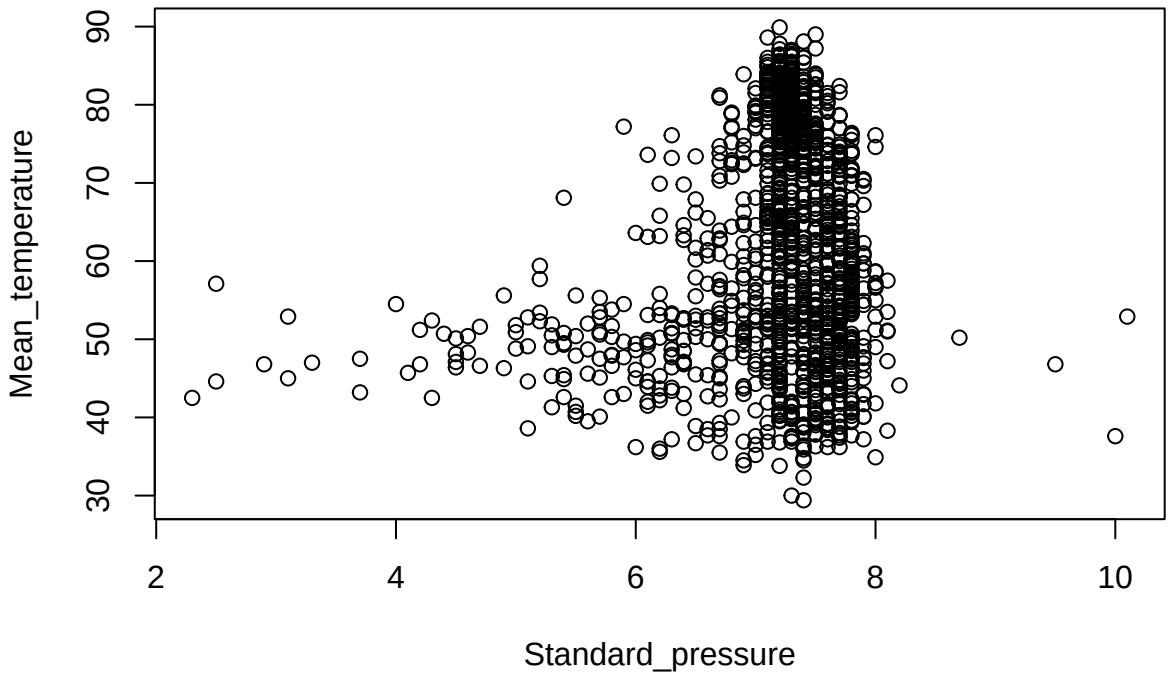
```
## [1]  2.3 10.1
```

```r
quantile(Standard_pressure)
```

```
##   0%  25%  50%  75% 100%
##  2.3  7.1  7.3  7.6 10.1
```

Gráfica de Standard_pressure respecto a Mean_temperature, la salida.

```r
plot(Standard_pressure, Mean_temperature)
```



Visibility

Medidas de centralidad

```
mean(Visibility)
```

```
## [1] 11.15803
```

```
median(Visibility)
```

```
## [1] 10.5
```

```
getMode(Visibility)
```

```
## [1] 7.14
```

Medidas de dispersión

```
var(Visibility)
```

```
## [1] 29.23202
```

```
sd(Visibility)
```

```
## [1] 5.406665
```

```
max(Visibility)
```

```
## [1] 29.1
```

```
min(Visibility)
```

```
## [1] 0.92
```

```
range(Visibility)
```

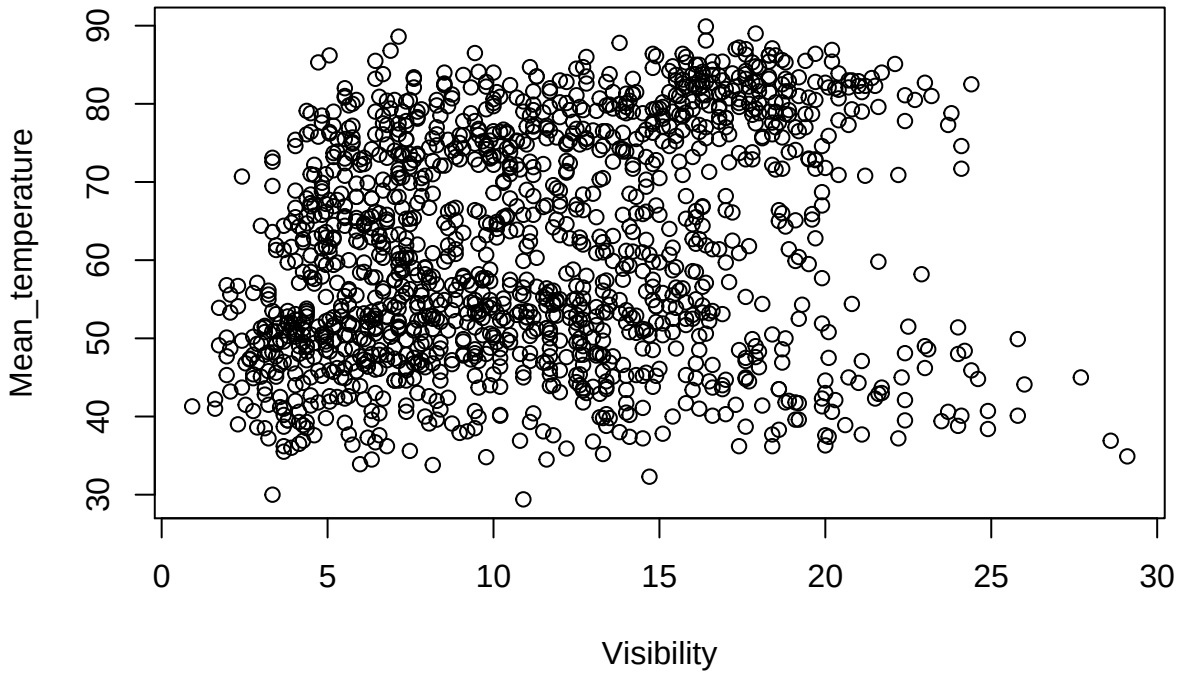
```
## [1]  0.92 29.10
```

```
quantile(Visibility)
```

```
##    0%   25%   50%   75%  100%
##  0.92  6.56 10.50 15.40 29.10
```

Gráfica de Visibility respecto a Mean_temperature, la salida.

```
plot(Visibility, Mean_temperature)
```



Wind_speed

Medidas de centralidad

```
mean(Wind_speed)
```

```
## [1] 19.81176
```

```
median(Wind_speed)
```

```
## [1] 19.81
```

```
getMode(Wind_speed)
```

```
## [1] 20.8
```

Medidas de dispersión

```
var(Wind_speed)
```

```
## [1] 50.9118
```

```r
sd(Wind_speed)
```

```
## [1] 7.13525
```

```r
max(Wind_speed)
```

```
## [1] 68.8
```

```r
min(Wind_speed)
```

```
## [1] 4.72
```

```r
range(Wind_speed)
```

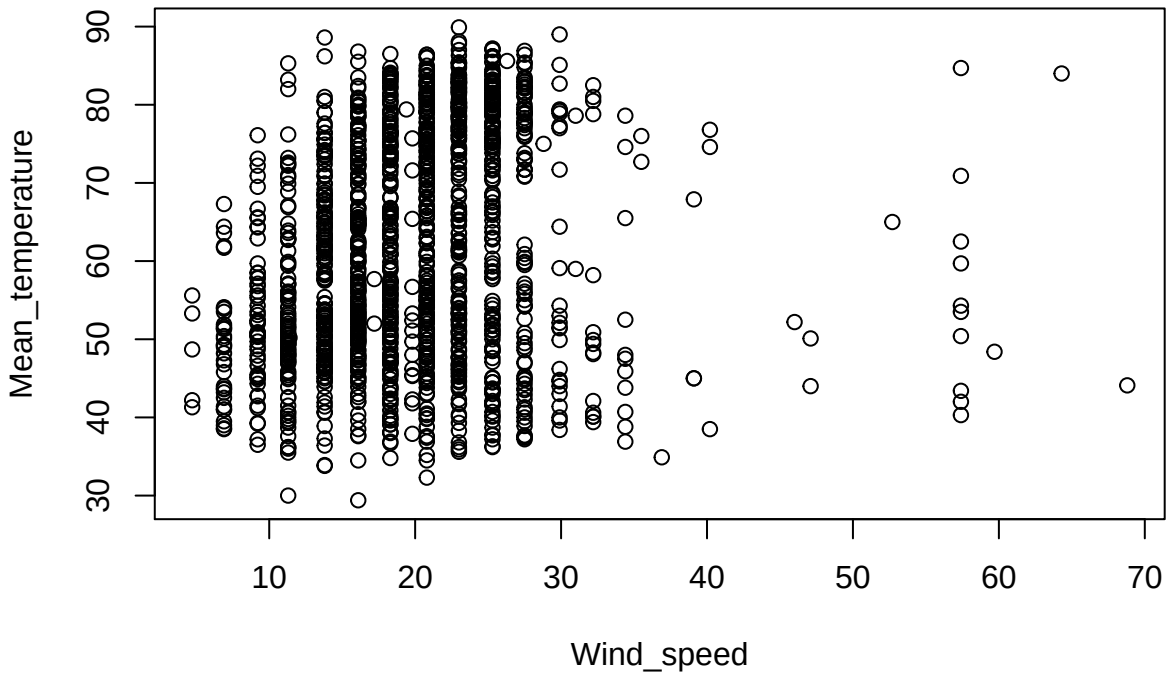
```
## [1]  4.72 68.80
```

```r
quantile(Wind_speed)
```

```
##    0%   25%   50%   75%  100%
##  4.72 16.10 19.81 23.00 68.80
```

Gráfica de Wind_speed respecto a Mean_temperature, la salida.

```r
plot(Wind_speed, Mean_temperature)
```



Max_wind_speed

Medidas de centralidad

```
mean(Max_wind_speed)
```

```
## [1] 34.28037
```

```
median(Max_wind_speed)
```

```
## [1] 34.28
```

```
getMode(Max_wind_speed)
```

```
## [1] 34.28
```

Medidas de dispersión

```
var(Max_wind_speed)
```

```
## [1] 5.962512
```

```
sd(Max_wind_speed)
```

```
## [1] 2.441826
```

```
max(Max_wind_speed)
```

```
## [1] 55.24
```

```
min(Max_wind_speed)
```

```
## [1] 16.11
```

```
range(Max_wind_speed)
```

```
## [1] 16.11 55.24
```

```
quantile(Max_wind_speed)
```

```
##    0%   25%   50%   75%  100%
## 16.11 34.28 34.28 34.28 55.24
```

Gráfica de Max_wind_speed respecto a Mean_temperature, la salida.

```r
plot(Max_wind_speed, Mean_temperature)
```