Analisi New York Air Quality Measurements

David Marabottini

2025-10-16

```
# install.packages('lubridate')
# install.packages("VIM")
# install.packages("psych")
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(lubridate)
```

```
##
## Caricamento pacchetto: 'lubridate'
```

```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(VIM)
```

```
## Caricamento del pacchetto richiesto: colorspace
```

```
## Caricamento del pacchetto richiesto: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Caricamento pacchetto: 'VIM'
```

```
## Il seguente oggetto è mascherato da 'package:datasets':
##
##     sleep
```

```
library(psych)
```

Esercizio di EDA – Analisi del dataset “AirQuality”

Utilizza il dataset airquality integrato in R, che contiene misurazioni giornaliere della qualità dell’aria a New York nel 1973. Obiettivo: esplorare e descrivere il dataset per individuare pattern, relazioni e anomalie nei dati. Consegna Caricamento e panoramica dei dati Carica il dataset e visualizza le sue prime righe, struttura e sommario statistico. Identifica le variabili quantitative e qualitative. Gestione dei valori mancanti Conta i valori mancanti per colonna. Proponi almeno due strategie diverse per gestirli e applicane una. Statistiche descrittive Calcola media, mediana, deviazione standard e range per le variabili numeriche. Visualizza una tabella riassuntiva pulita con questi indicatori. Visualizzazione dei dati Crea almeno: Un boxplot per individuare outlier. Un histogramma o density plot per la distribuzione di una variabile. Un grafico di correlazione tra variabili quantitative (es. Ozone vs Temp). Analisi di correlazione Calcola e visualizza la matrice di correlazione. Commenta almeno una relazione significativa. Conclusioni Riassumi brevemente le principali scoperte dell’analisi (max 5 righe).

New york air quality

Daily air quality measurements in New York, May to September 1973.

```
data("airquality")
```

```
help(airquality)
```

```
## avvio in corso del server httpd per la guida ... fatto
```

A data frame with 153 observations on 6 variables.

variables:

- [,1] **Ozone**: numeric (ppb)
- [,2] **Solar.R**: numeric (lang)
- [,3] **Wind**: numeric (mph)
- [,4] **Temp**: numeric (degrees F)
- [,5] **Month**: numeric (1–12)
- [,6] **Day**: numeric (1–31)

Benchè tutte le variabli siano numeriche ordinali, le variabili `ozone`, `Solar.R`, `Wind` e `Temp` sono quantitative, mentre `Month` e `Day` sono qualitative

```
colSums(is.na(airquality))
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

Dalla tabella sopra emerge che ci sono 37 misurazioni mancanti per l’ozono e 7 per i raggi solari

```r
length(which(is.na(airquality)))
```

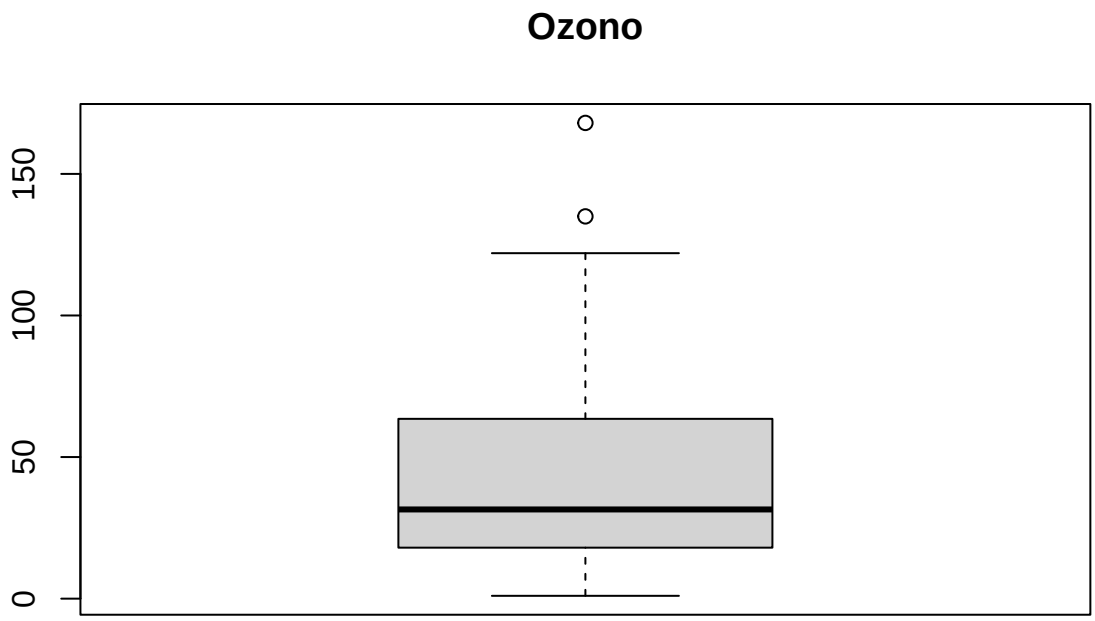
```
## [1] 44
```

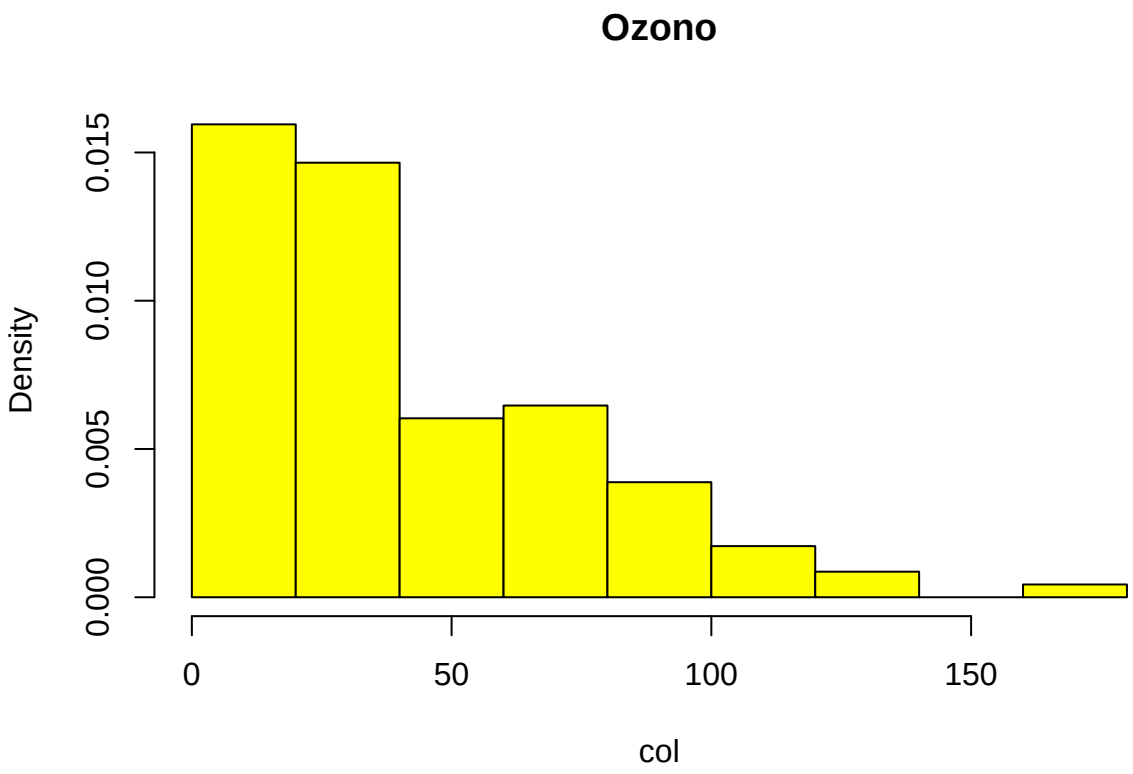
considerando che la tabella sospra evidenziava 37 casi in cui non erano presenti misurazioni di ozono e 7 in cui non erano presenti di raggi solari, e il numero di celle che hanno dati mancanti è 44, possiamo concludere che non ci sono righe in cui mancano contemporaneamente ozono e raggi solari.

Cerchiamo di capire la distribuzione dei dati per comprendere come gestire queste mancanze

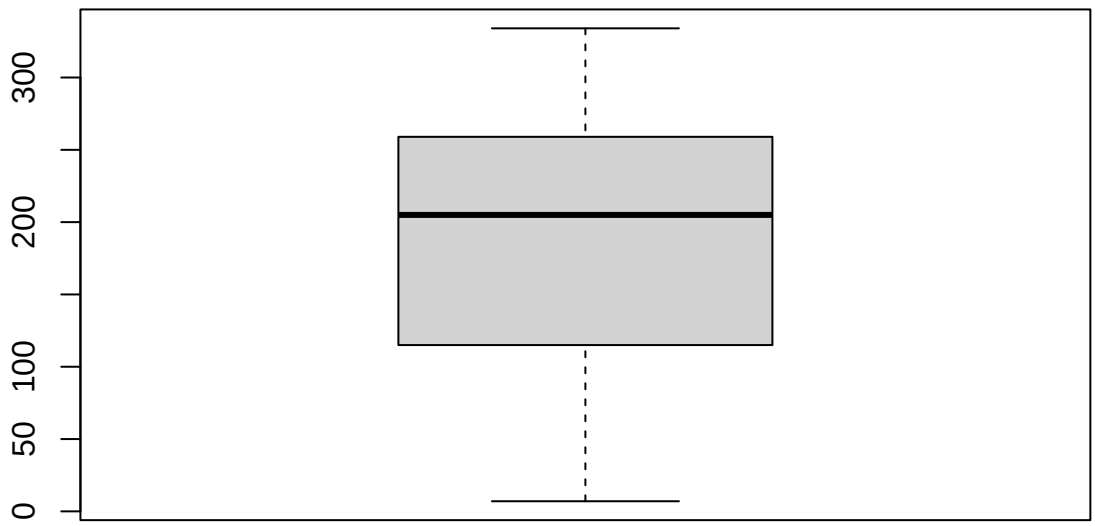
```r
analyze_col <- function(col, name) {
  # par(mfrow=c(1,2))
  boxplot(col, main=name)

  hist(col, col='yellow', freq=F, main=name)
  # lines(density(col), lwd=2, col="red")
  return(summary(col))
}
```

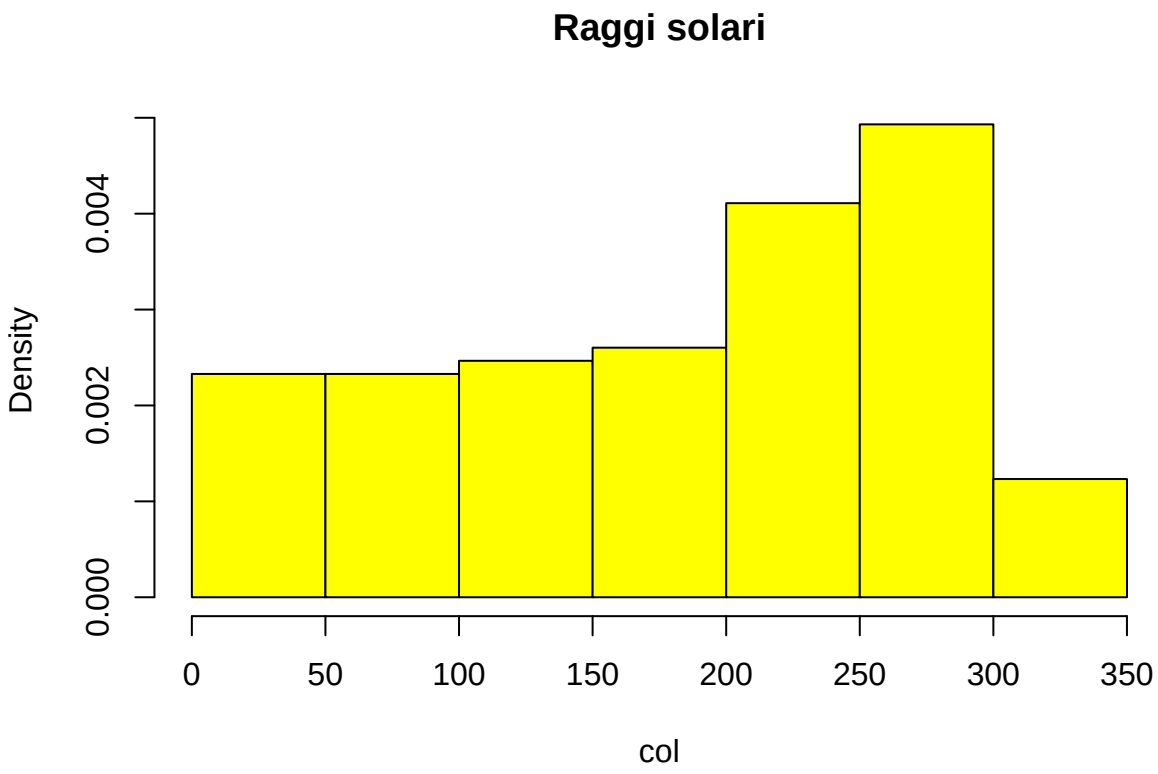
```r
summary_ozono = analyze_col(airquality$Ozone, 'Ozono')
```






```
summary_raggi_solari = analyze_col(airquality$Solar.R, 'Raggi solari')
```


Raggi solari
0
50
100
200
300




```
summary_ozono
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   18.00   31.50   42.13   63.25  168.00      37
```

```
print((summary_ozono[[5]] - summary_ozono[[2]])/(summary_ozono[[6]] - summary_ozono[[1]]))
```

```
## [1] 0.2709581
```

```
(summary_raggi_solari[[5]] - summary_raggi_solari[[2]])/(summary_raggi_solari[[6]] - summary_raggi_sola
```

```
## [1] 0.4373089
```

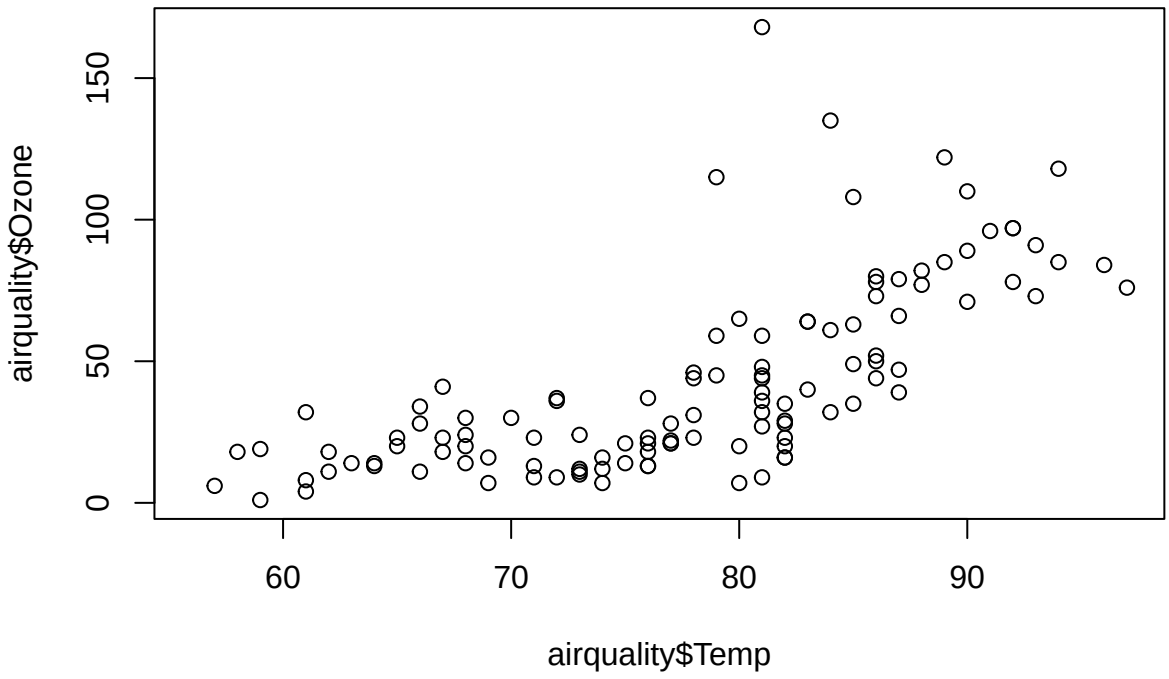
```
matrice_cor_completa <- cor(airquality, use = "complete.obs")
correlazioni_ozone <- matrice_cor_completa[, "Ozone"]
round(correlazioni_ozone, 4)
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##  1.0000  0.3483 -0.6125  0.6985  0.1429 -0.0052
```

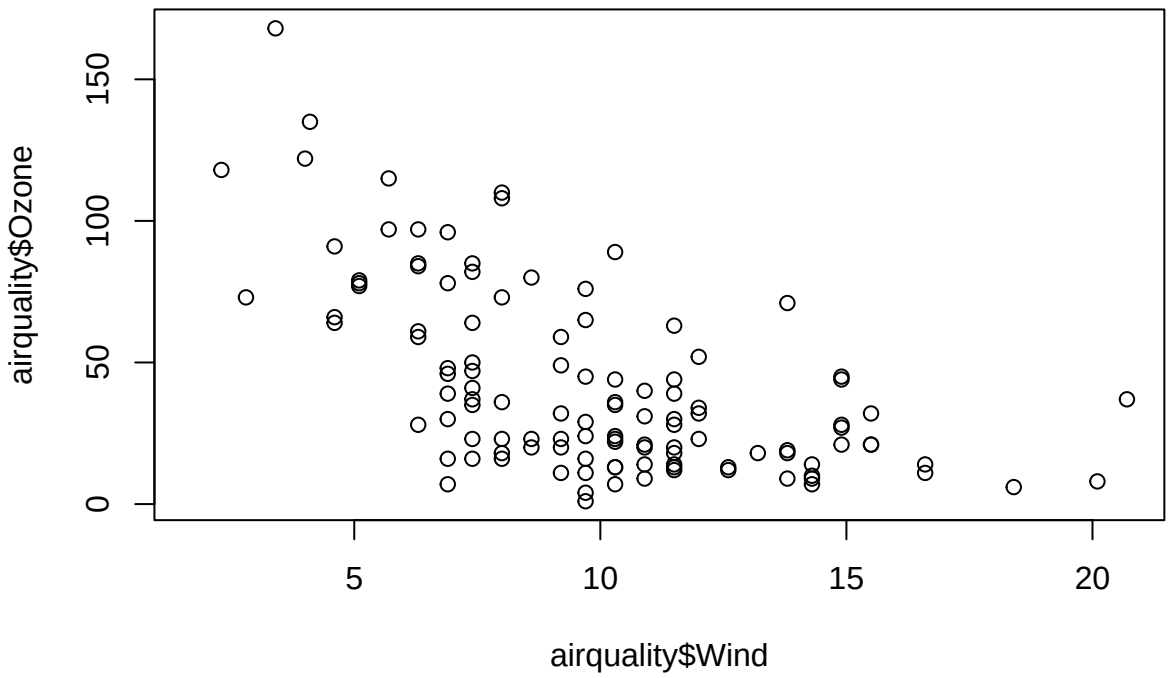
Sembra esserci una correlazione lineare inversa discreta con Temp, mentre una correlazione lineare diretta forte con Temp.

possiamo imputare l'ozono considerando una di queste 2 variabili

```r
plot(airquality$Temp, airquality$Ozone)
```

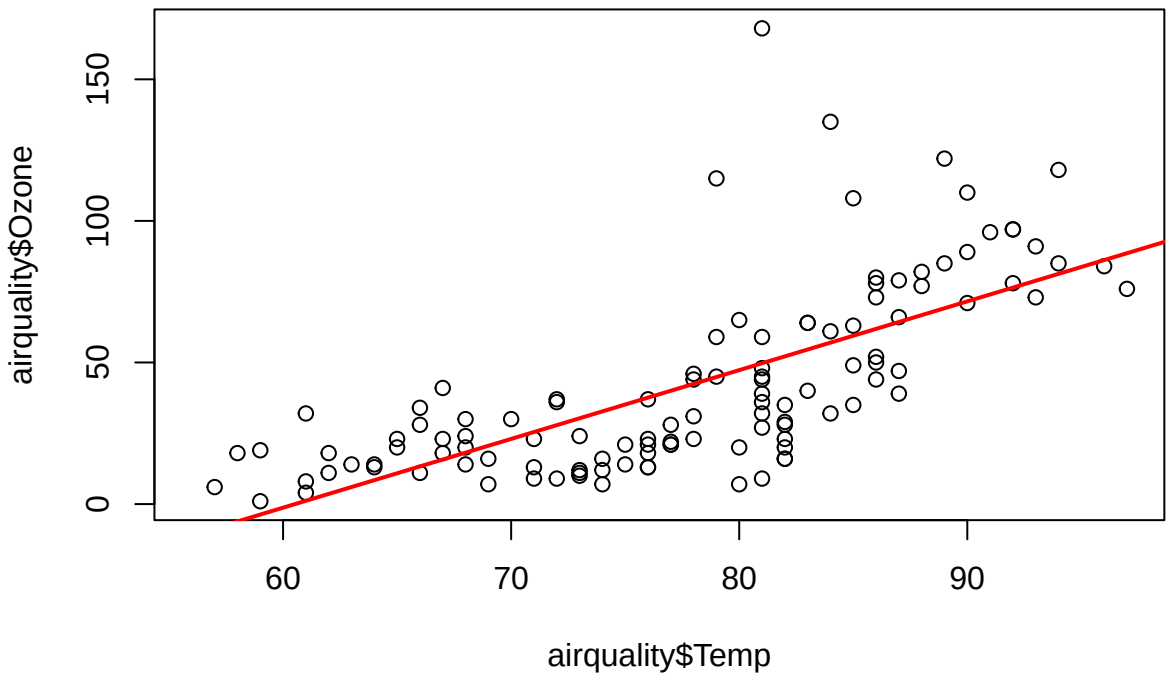



```r
plot(airquality$Wind, airquality$Ozone)
```



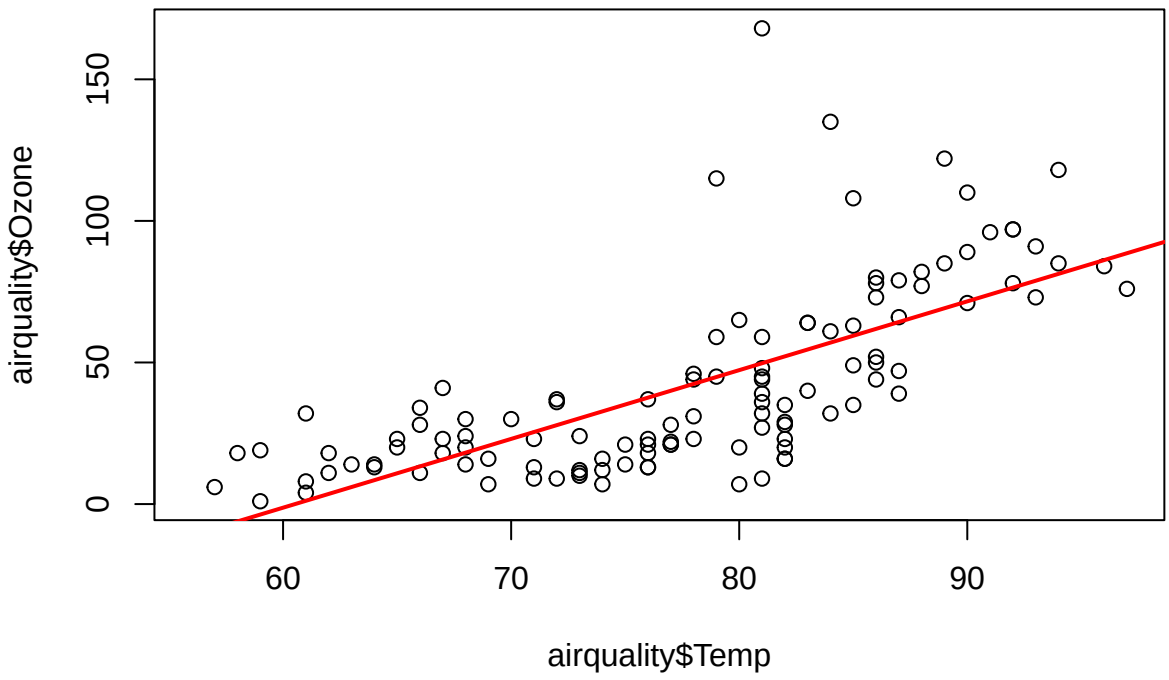
La relazione tra temperatura e ozono mi pare piu esponenziale, ma forse puo essere approssimata da una retta

```
modello <- lm(Ozone ~ Temp , data = airquality)

summary_model = summary(modello)
m = summary_model$coefficients[, 1][['Temp']]
q = summary_model$coefficients[, 1][['(Intercept)']]

plot(airquality$Temp, airquality$Ozone)
abline(a = q, b = m, col = "red", lwd = 2)
```



Da questo grafico sembra evidente che la relazione sia più che altro logaritmica

```
modello <- lm(Ozone ~ Temp , data = airquality)

summary_model = summary(modello)
m = summary_model$coefficients[, 1][['Temp']]
q = summary_model$coefficients[, 1][['(Intercept)']]

plot(airquality$Temp, airquality$Ozone)
abline(a = q, b = m, col = "red", lwd = 2)
```

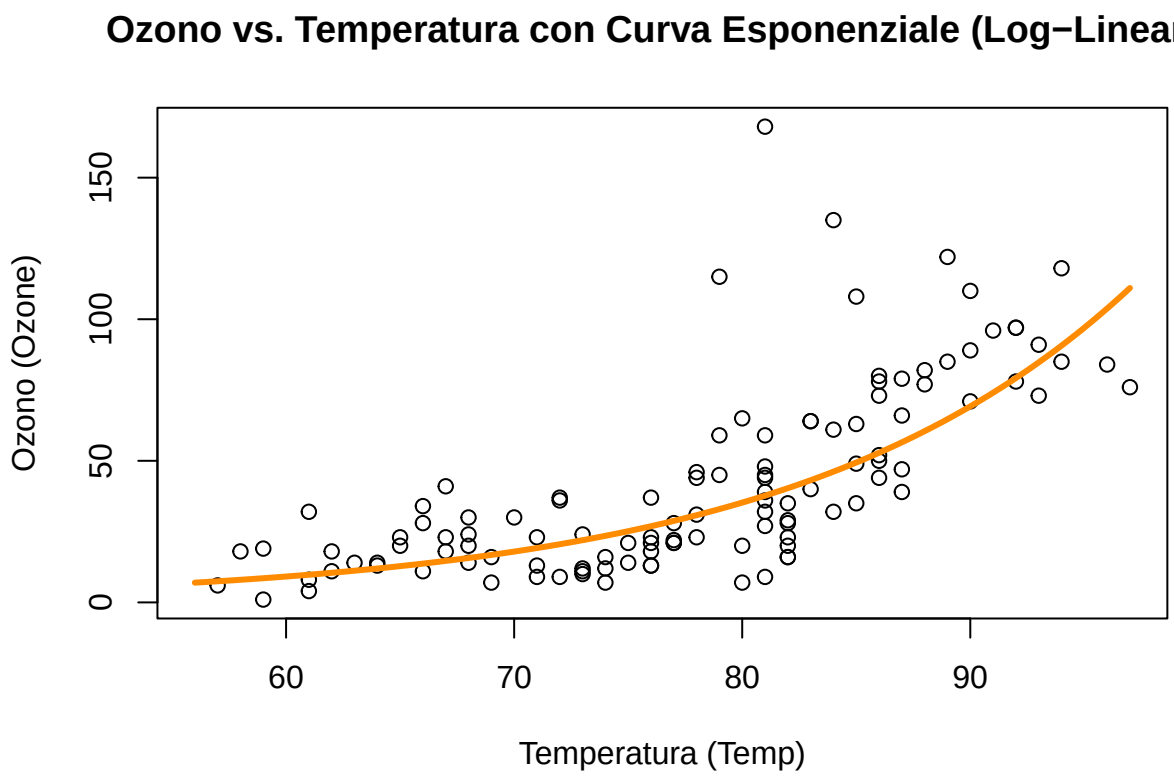



```r
modello_log <- lm(log(Ozone) ~ Temp, data = airquality)
summary(modello_log)
```

```
##
## Call:
## lm(formula = log(Ozone) ~ Temp, data = airquality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14469 -0.33095  0.02961  0.36507  1.49421
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.83797    0.45100  -4.075 8.53e-05 ***
## Temp         0.06750    0.00575  11.741  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5848 on 114 degrees of freedom
##   (37 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.5473, Adjusted R-squared:  0.5434
## F-statistic: 137.8 on 1 and 114 DF,  p-value: < 2.2e-16
```

```r
coeff_log <- summary(modello_log)$coefficients
q_log <- coeff_log[, "Estimate"]["(Intercept)"]
```

```r
m_log <- coeff_log[, "Estimate"]["Temp"]

plot(airquality$Temp, airquality$Ozone,
     xlab = "Temperatura (Temp)",
     ylab = "Ozono (Ozone)",
     main = "Ozono vs. Temperatura con Curva Esponenziale (Log-Lineare)")

curve(expr = exp(q_log + m_log * x), add = T, col = "darkorange", lwd = 3)
```



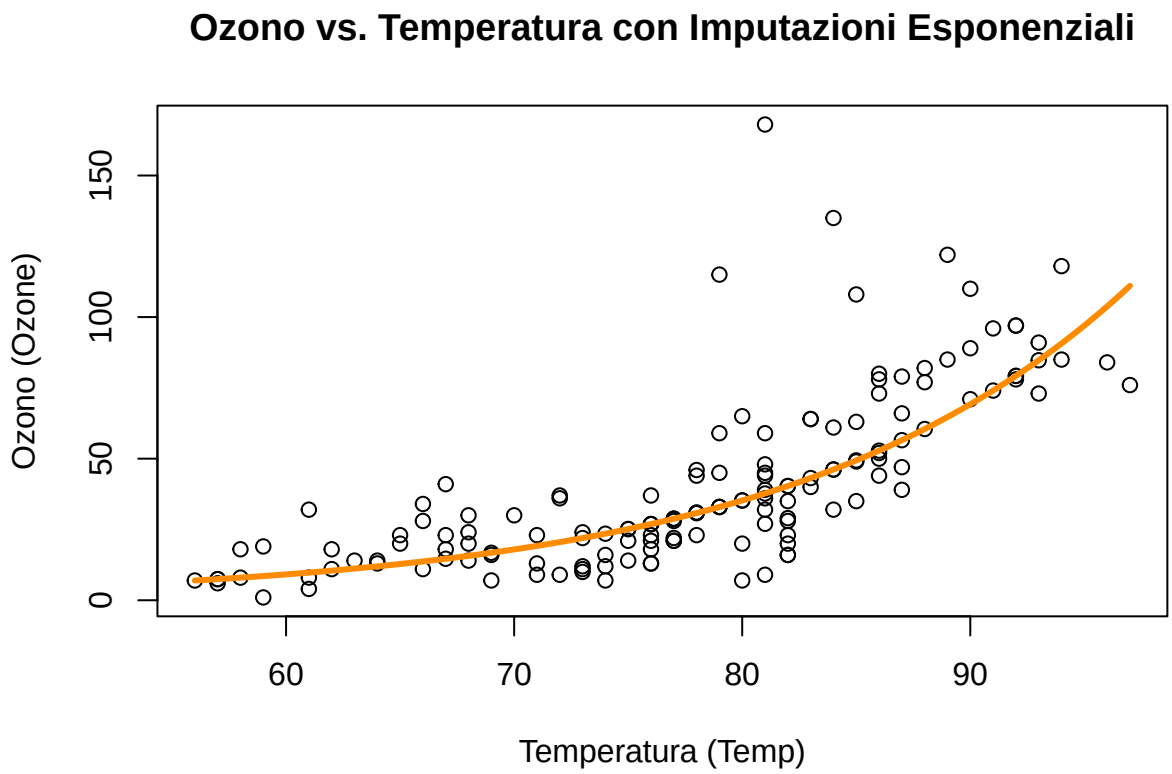
L'esponenziale sembra avere una presa migliore rispetto alla retta

```r
indexes_na_ozone = which(is.na(airquality$Ozone))
temp_imputazione = airquality$Temp[indexes_na_ozone]
valori_imputati_esponenziali = exp(q_log + m_log * temp_imputazione)
airquality$Ozone[indexes_na_ozone] <- valori_imputati_esponenziali
cat("Valori NA in Ozone dopo l'imputazione esponenziale:", sum(is.na(airquality$Ozone)))
```

```
## Valori NA in Ozone dopo l'imputazione esponenziale: 0
```

```r
# 1. Trova gli indici dei valori NA originali in Ozone (Assicurati che airquality sia il dataframe con
# Se hai già imputato airquality in precedenza, ricarica i dati o usa una copia.
# Per chiarezza, assumiamo che tu stia lavorando sull'originale con NA:
indici_na_ozone <- which(is.na(airquality$Ozone))

# 2. Estrai la temperatura dei giorni con NA
```

```r
temp_na <- airquality$Temp[indici_na_ozone]

# 3. Calcola i valori di OZONE imputati con la curva esponenziale per quei giorni
# Formula: Ozone = exp(q_log + m_log * Temp)
ozone_imputato <- exp(q_log + m_log * temp_na)

# 4. Crea il grafico originale e la curva (il tuo codice)
plot(airquality$Temp, airquality$Ozone,
     xlab = "Temperatura (Temp)",
     ylab = "Ozono (Ozone)",
     main = "Ozono vs. Temperatura con Imputazioni Esponenziali")

curve(expr = exp(q_log + m_log * x), add = TRUE, col = "darkorange", lwd = 3)
```



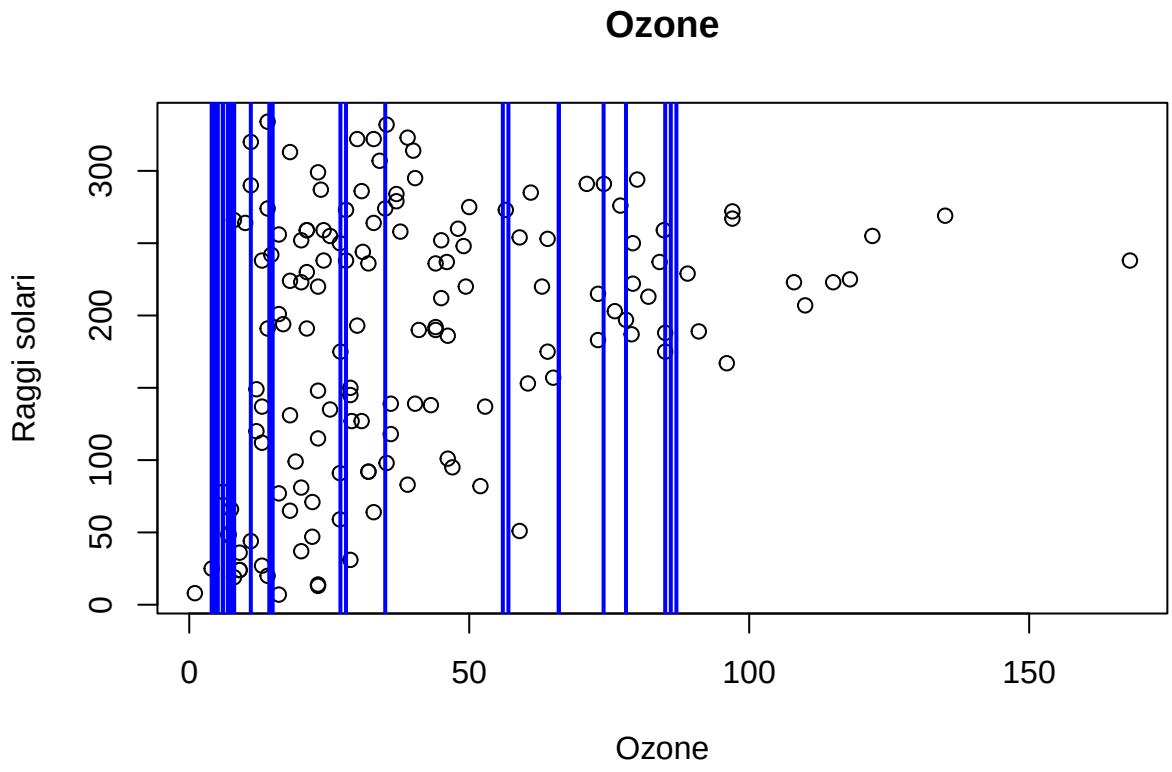

```r
matrice_cor_completa <- cor(airquality, use = "complete.obs")
correlazioni_ozone <- matrice_cor_completa[, "Solar.R"]
round(correlazioni_ozone, 4)
```

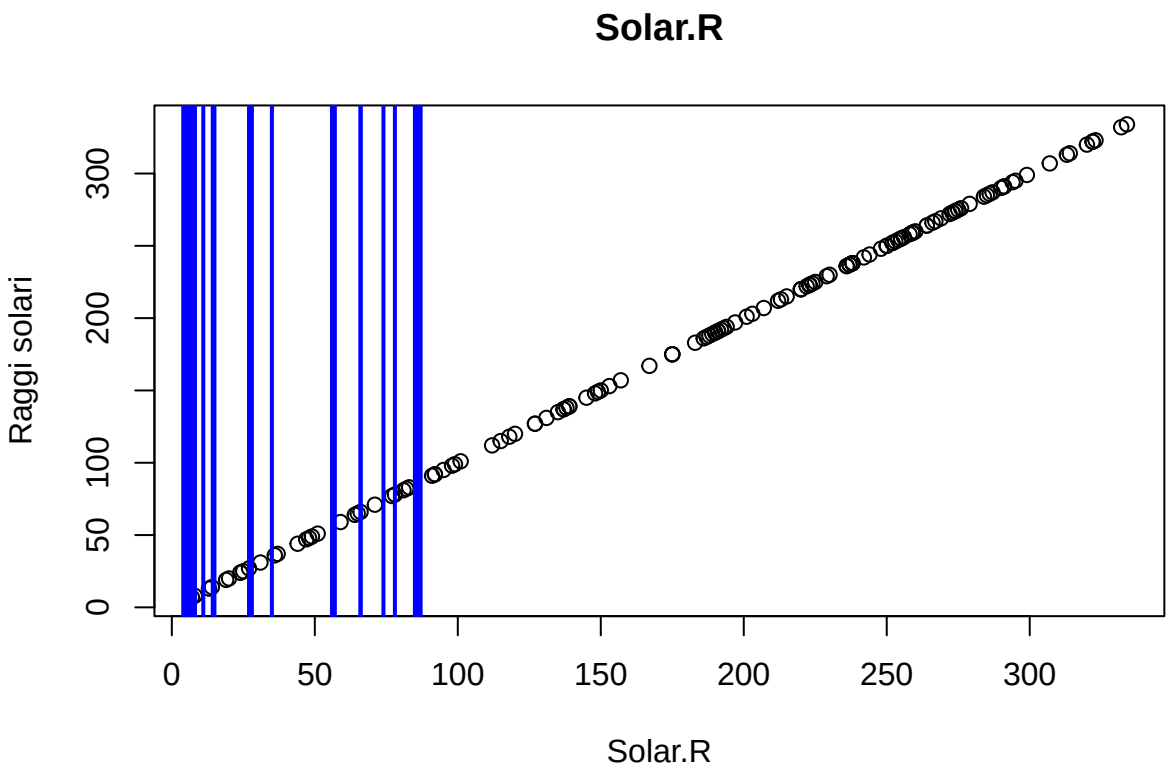
```
##   Ozone Solar.R    Wind    Temp   Month     Day
##  0.3264  1.0000 -0.0568  0.2758 -0.0753 -0.1503
```

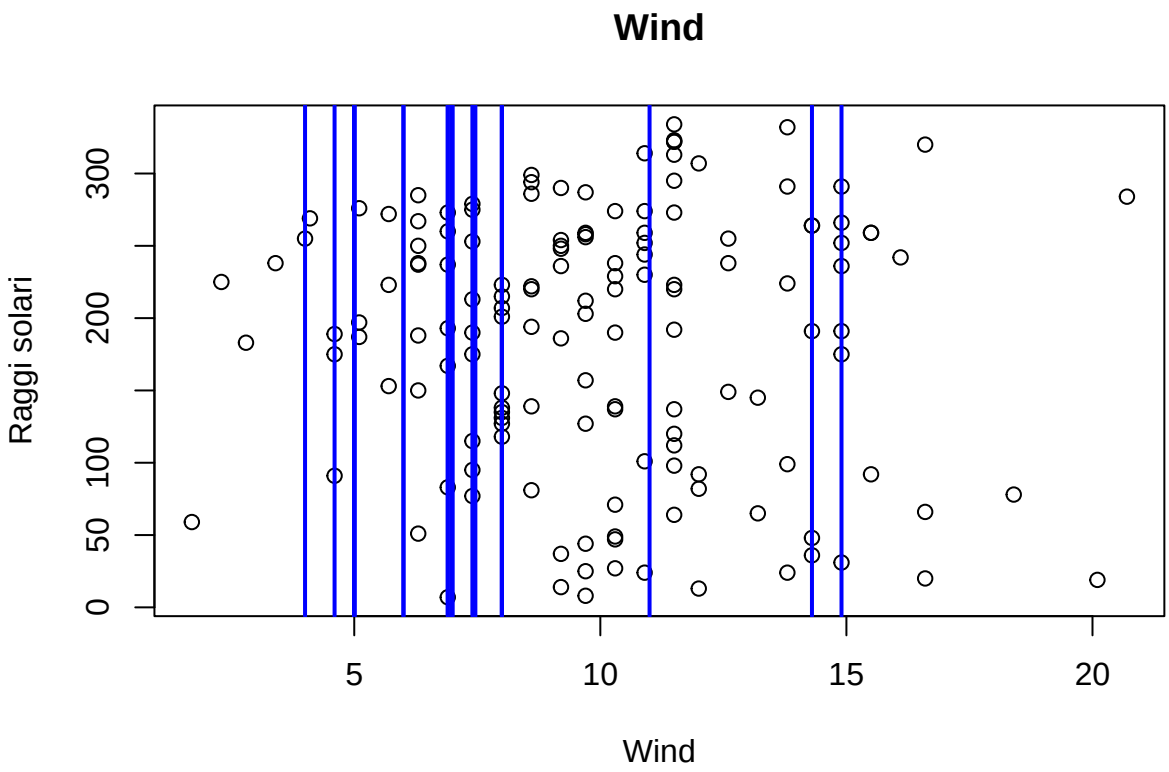
si può notare una correlazione lineare abbastanza lieve tra raggi solari e ozono o raggi solari e temperatura, ma troppo bassa per utilizzarla direttamente come modello di imputazione.

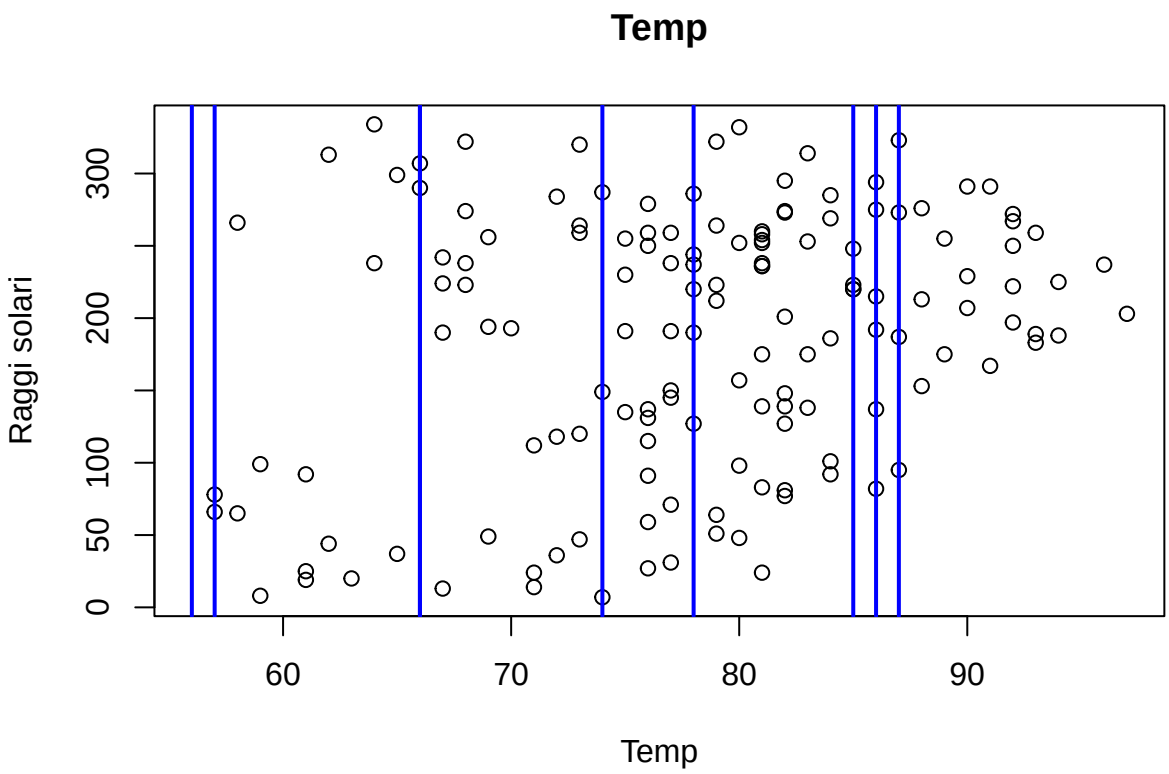
Stampo degli scatterplot per verificare eventuali correlazioni non lineari

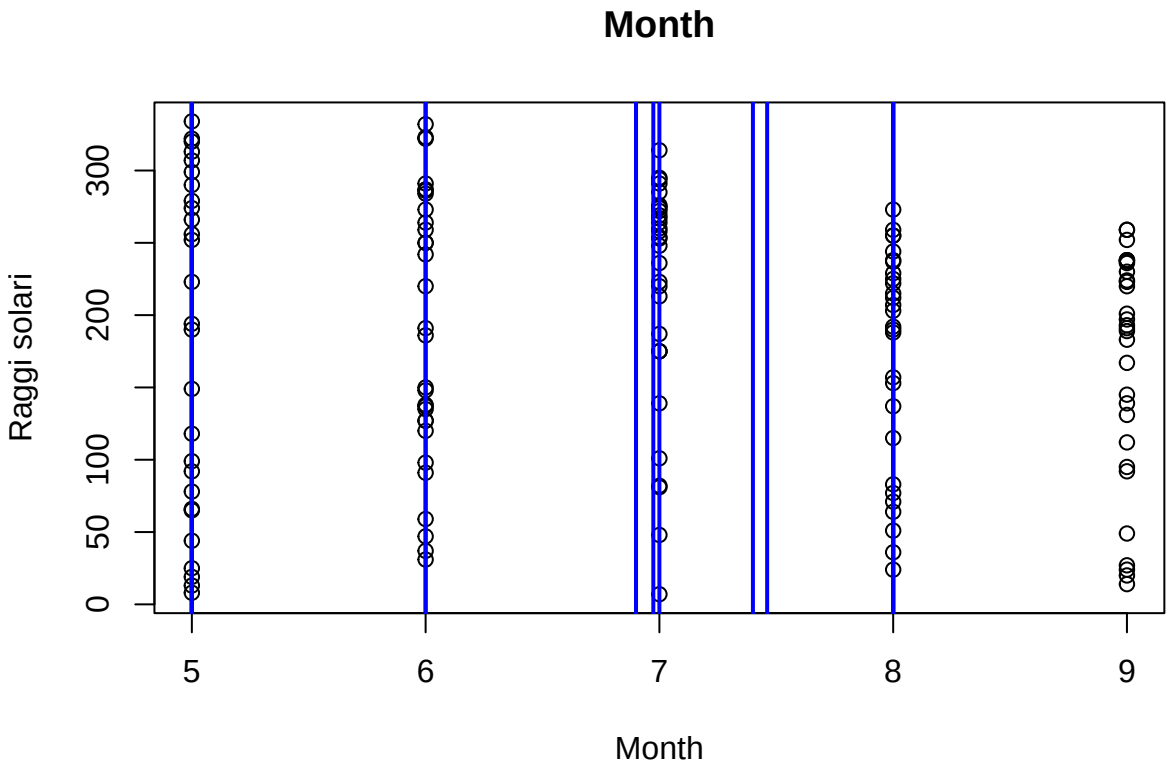
```r
solar_r_na = airquality[is.na(airquality$Solar.R), ]
for(n in names(airquality)) {
  plot(airquality[[n]], airquality$Solar.R, main=n, ylab="Raggi solari", xlab=n)
  for(i in solar_r_na) {
    abline(v=i, col="blue", lwd = 2)
  }
}
```

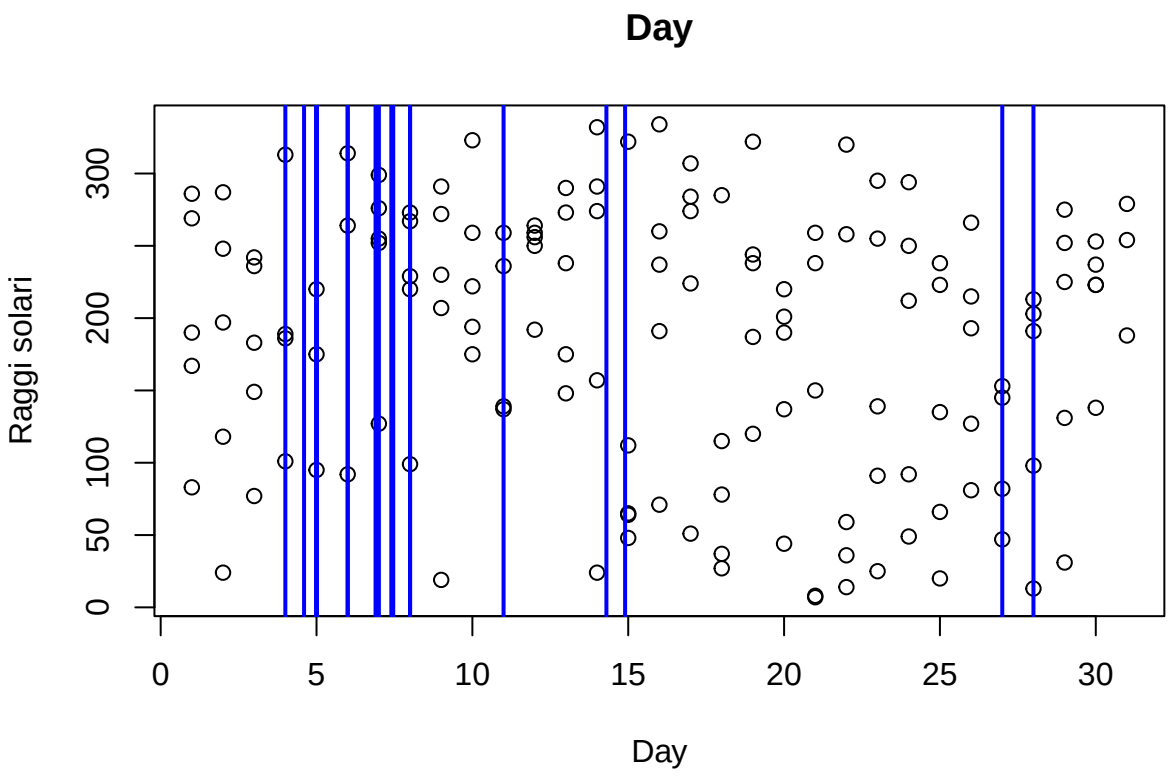



Solar.R
Raggi solari
0
50
100
200
300
0
50
100
150
200
250
300
Solar.R


Wind
Raggi solari
0
50
100
200
300
5
10
15
20
Wind


Temp
Raggi solari
0
50
100
200
300
60
70
80
90
Temp


Month
Raggi solari
0
50
100
200
300
5
6
7
8
9
Month




```
Data_Stringa = paste("1973", airquality$Month, airquality$Day, sep = "-")
airquality$Date_to_nr = yday(ymd(Data_Stringa))
```

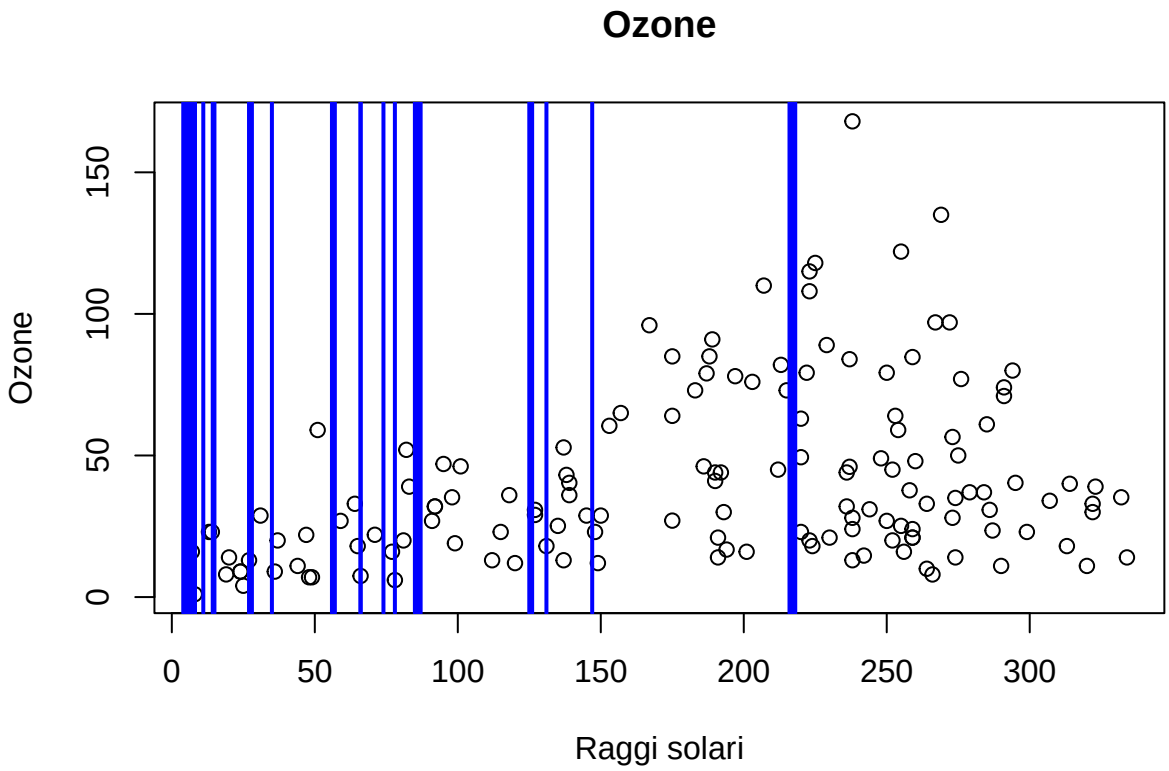
```
matrice_cor_completa <- cor(airquality, use = "complete.obs")
correlazioni_ozone <- matrice_cor_completa[, "Solar.R"]
round(correlazioni_ozone, 4)
```

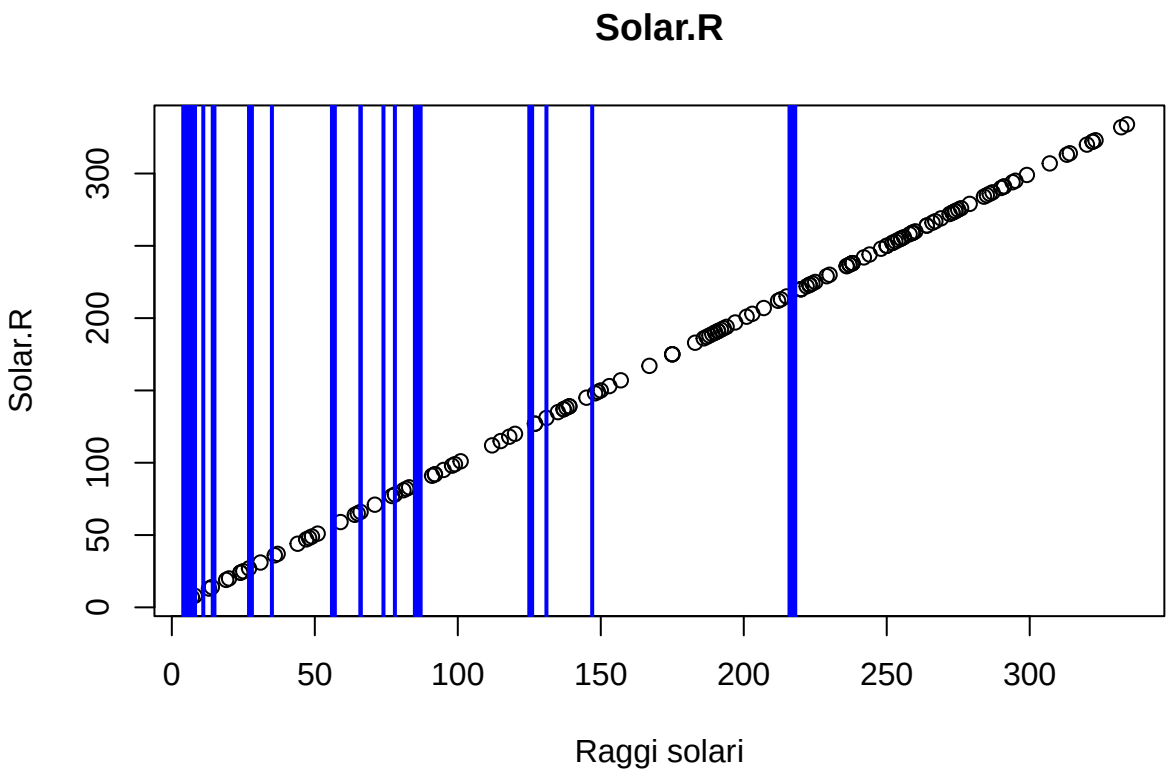
```
##      Ozone    Solar.R       Wind       Temp      Month        Day Date_to_nr
##     0.3264     1.0000    -0.0568     0.2758    -0.0753    -0.1503    -0.1047
```

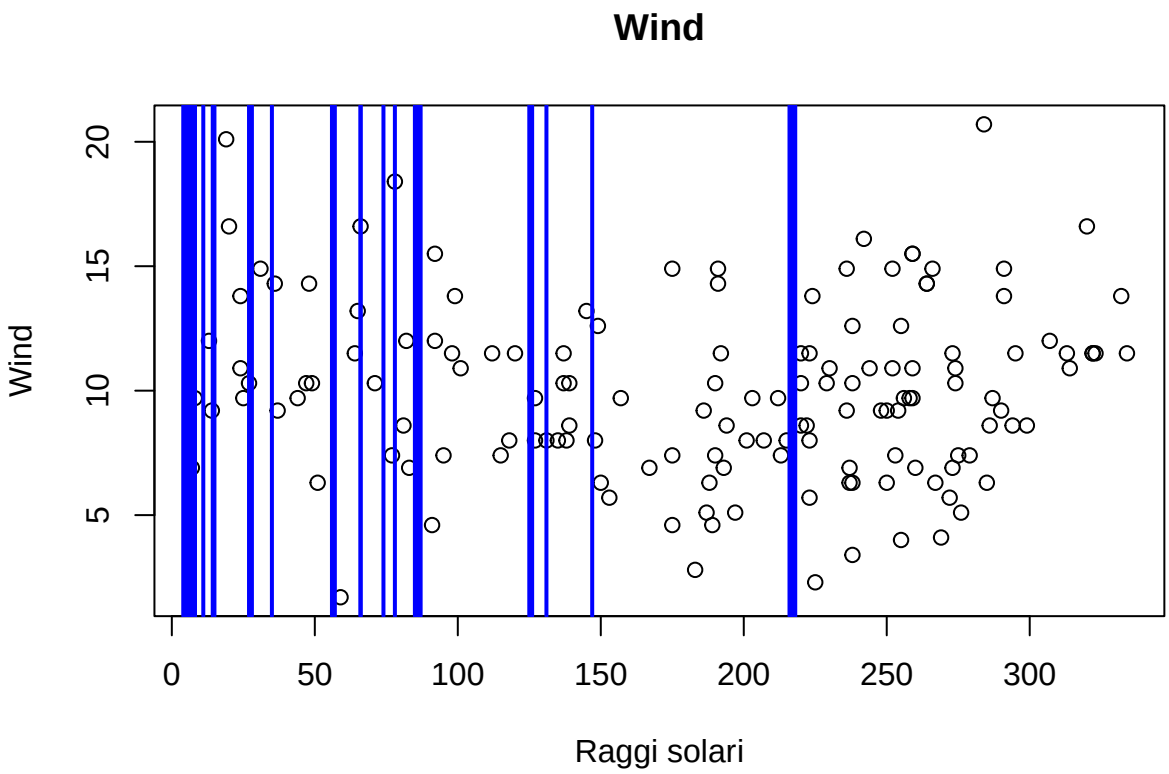
Non sembrano esserci correlazioni lineari con i raggi solari

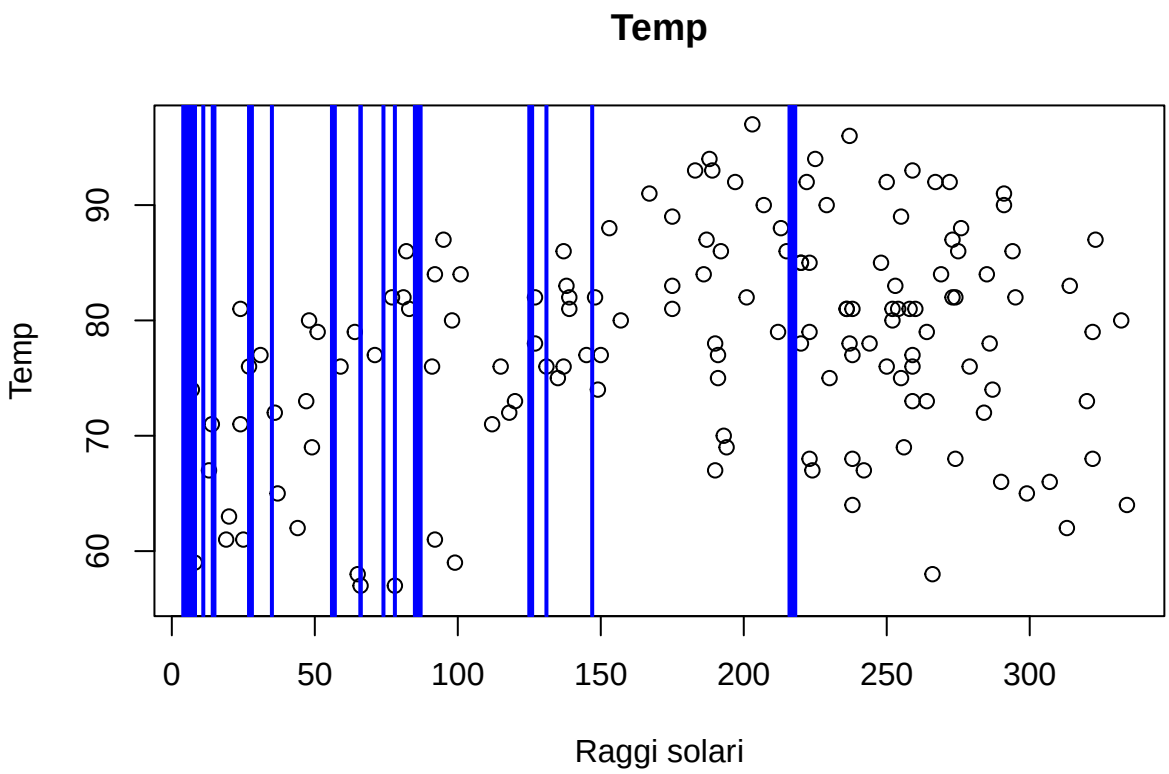
Disegno uno scatterplot per verificare non lineari

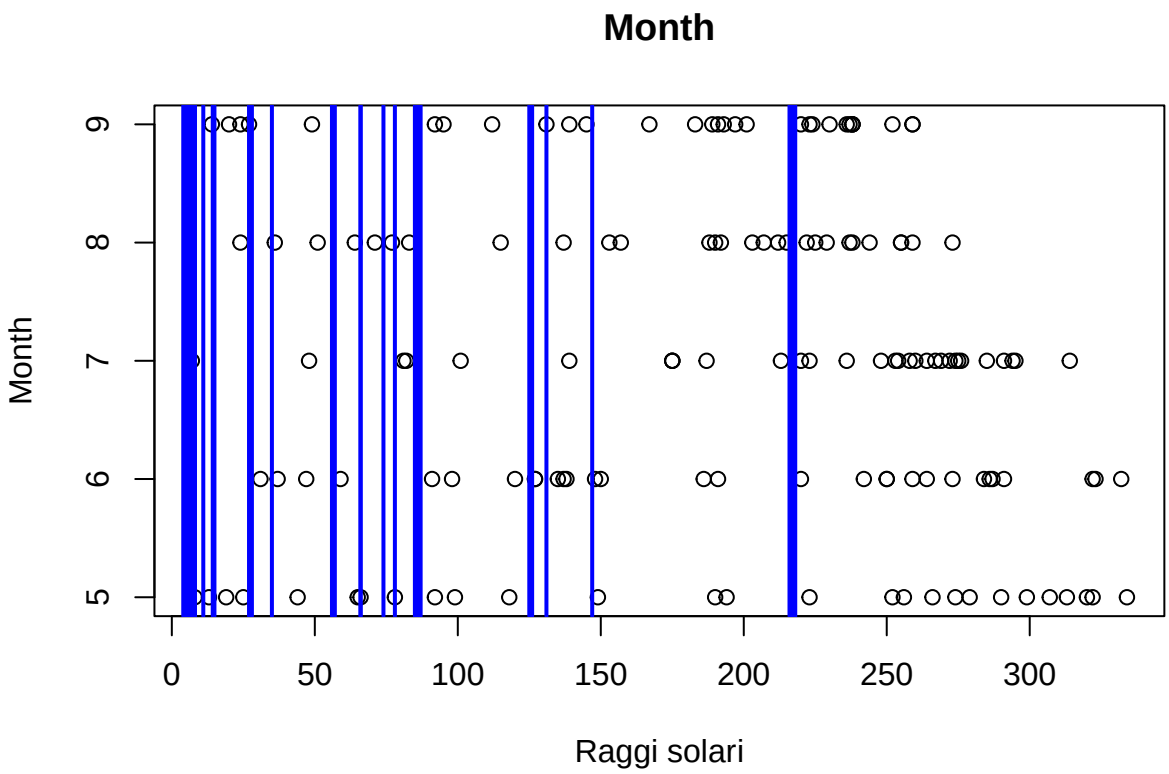
```
solar_r_na = airquality[is.na(airquality$Solar.R), ]
for(name in names(airquality)) {
  plot(airquality$Solar.R, airquality[[name]], xlab='Raggi solari', ylab=name, main=name)

  for(i in solar_r_na) {
    abline(v=i, col="blue", lwd = 2)
  }
}
```

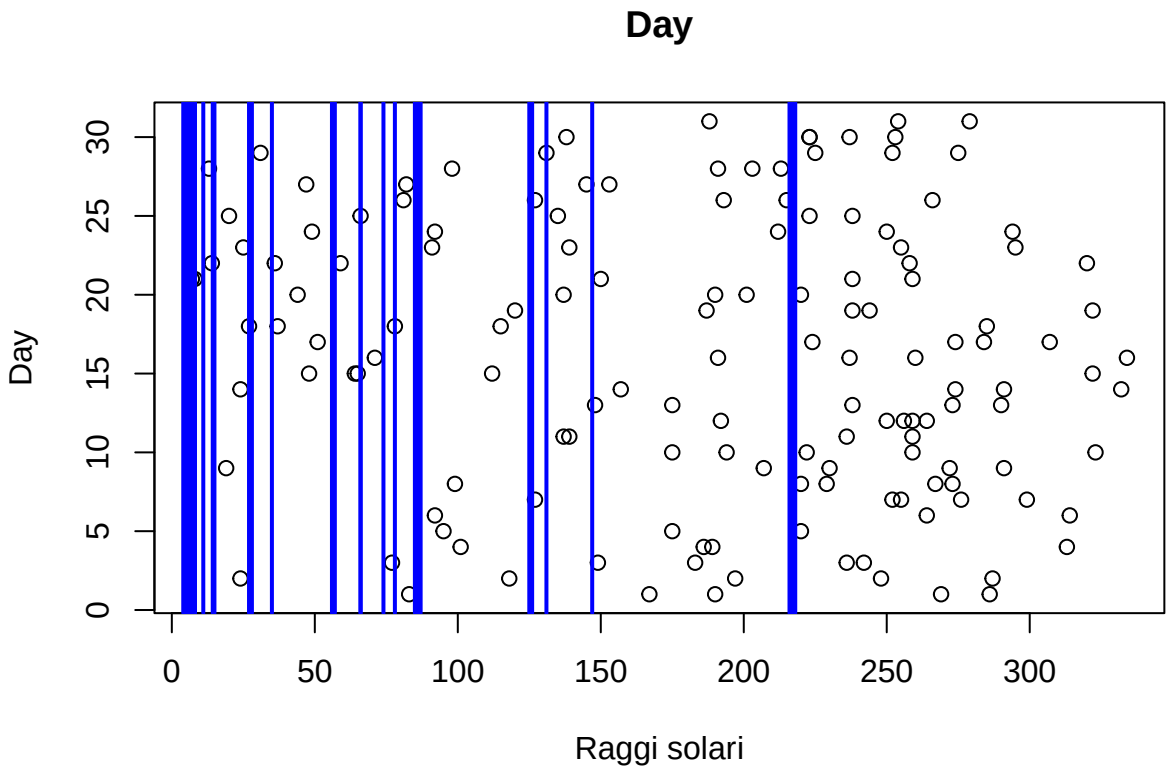

Ozone
Ozone
0
50
100
150
0
50
100
150
200
250
300
Raggi solari

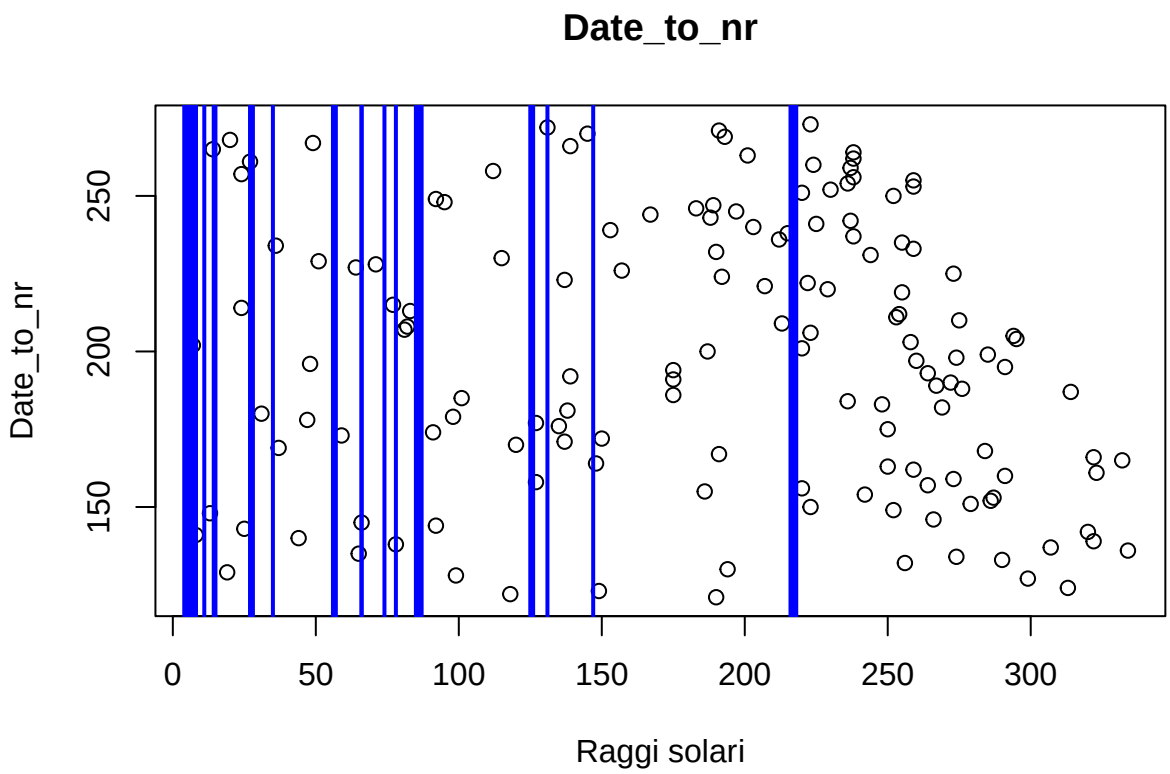

Solar.R
Solar.R
0
50
100
200
300
0
50
100
150
200
250
300
Raggi solari


Wind
Wind
5
10
15
20
0
50
100
150
200
250
300
Raggi solari

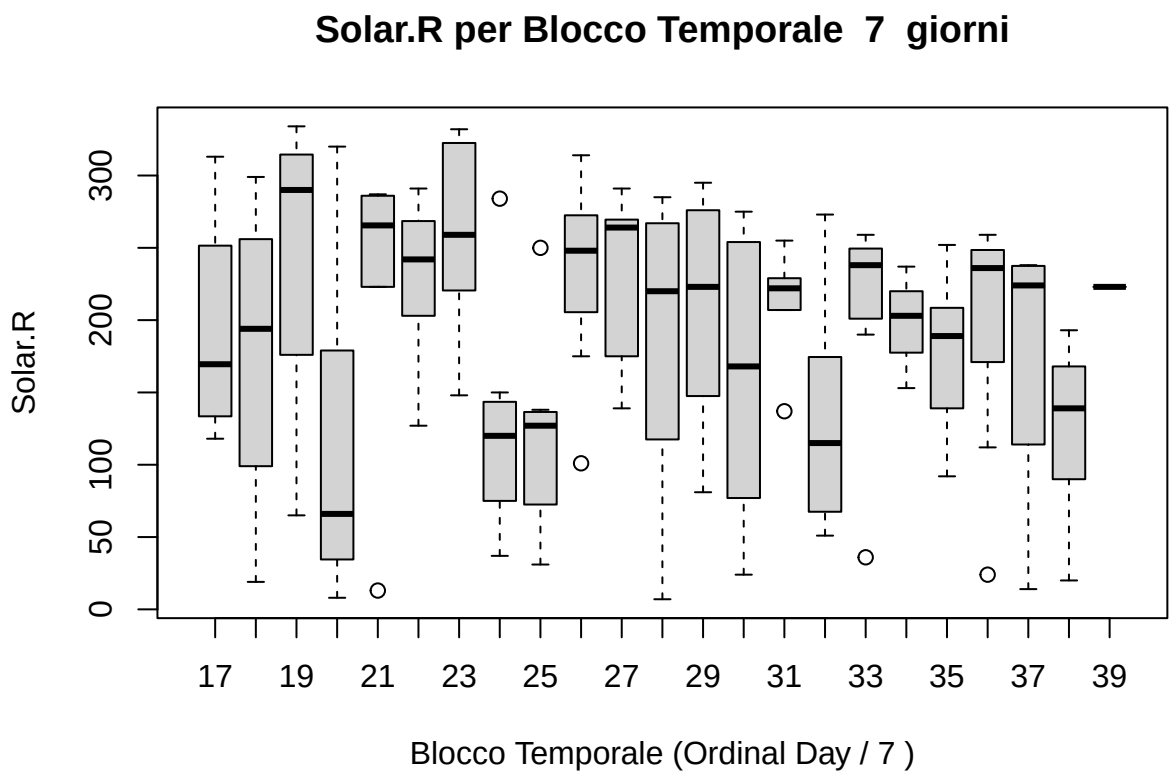

Temp
Temp
60
70
80
90
0
50
100
150
200
250
300
Raggi solari


Month
Month
9
8
7
6
5
0
50
100
150
200
250
300
Raggi solari


Day
Day
0
5
10
15
20
25
30
0
50
100
150
200
250
300
Raggi solari




```
nr_periods = 7
boxplot(airquality$Solar.R ~ as.integer(airquality$Date_to_nr / nr_periods),
        main = paste("Solar.R per Blocco Temporale ", nr_periods, " giorni"),
        xlab = paste("Blocco Temporale (Ordinal Day /",nr_periods,")"),
        ylab = "Solar.R")
```



Non riesco a trovare alcuna relazione solida, piuttosto preferisco eliminare le righe e continuare

```
which(is.na(airquality$Solar.R))
```

```
## [1]  5  6 11 27 96 97 98
```

```
airquality_filtered = airquality[!is.na(airquality$Solar.R), ]
```

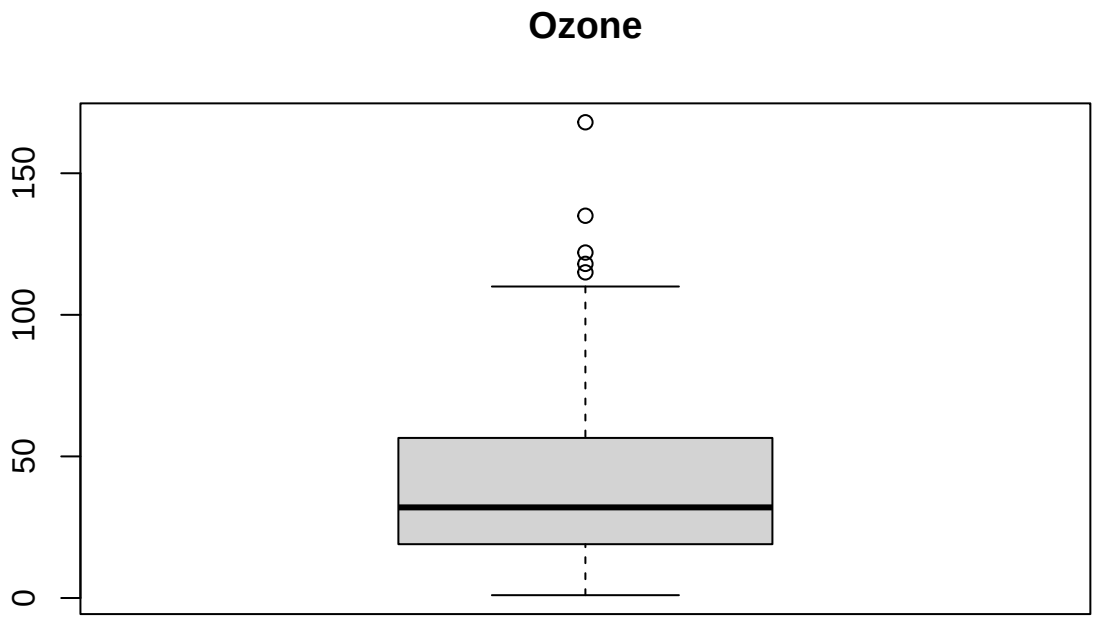
```
calculate_rapresentants <- function(x) c(mean(x), median(x), sd(x), var(x), min(x), max(x), summary(x))
```

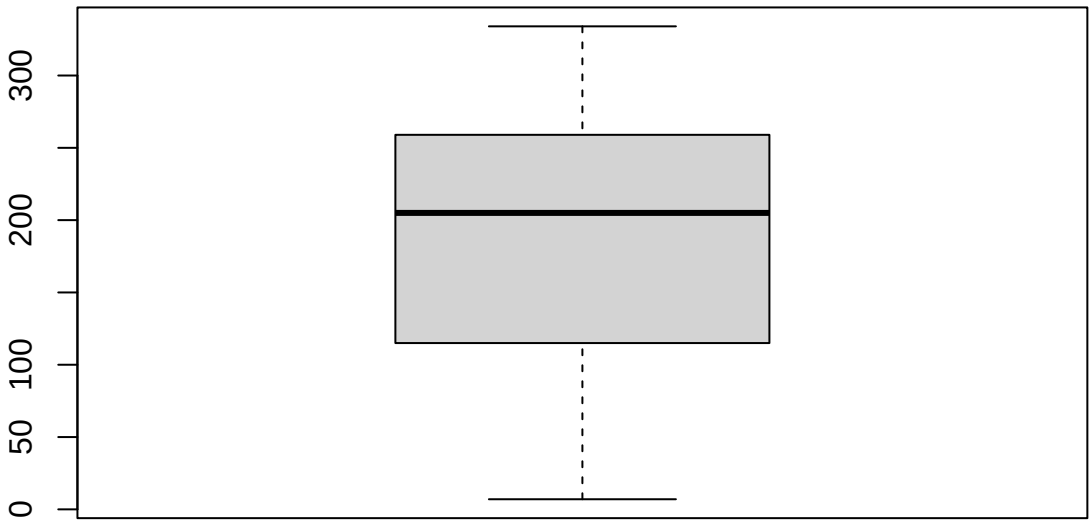
```
describe(airquality_filtered)
```

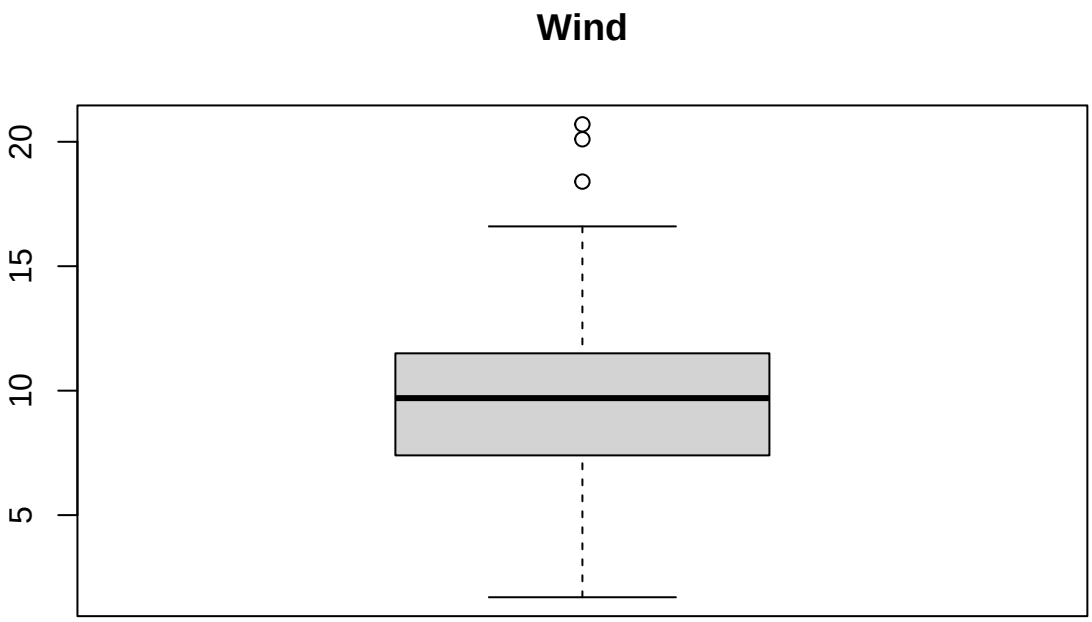
```
##            vars   n   mean    sd median trimmed   mad   min   max range  skew
## Ozone         1 146  41.12 30.51   32.0   37.12 23.15   1.0 168.0   167  1.32
## Solar.R       2 146 185.93 90.06  205.0  190.34 98.59   7.0 334.0   327 -0.42
## Wind          3 146  10.00  3.51    9.7    9.92  3.41   1.7  20.7    19  0.33
## Temp          4 146  78.12  9.22   79.0   78.45  8.90  57.0  97.0    40 -0.32
## Month         5 146   7.03  1.40    7.0    7.03  1.48   5.0   9.0     4  0.00
## Day           6 146  16.12  8.79   16.0   16.19 10.38   1.0  31.0    30 -0.06
## Date_to_nr    7 146 198.36 43.94  197.5  198.47 56.34 121.0 273.0   152  0.00
##            kurtosis   se
## Ozone          1.71 2.52
## Solar.R       -1.00 7.45
## Wind           0.12 0.29
```

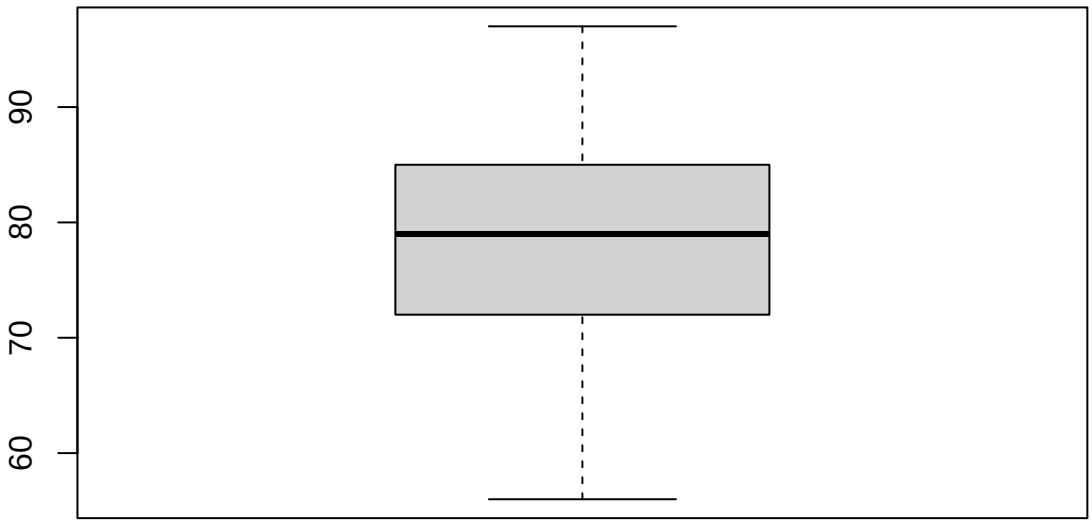
```
## Temp          -0.45 0.76
## Month         -1.29 0.12
## Day           -1.17 0.73
## Date_to_nr    -1.22 3.64
```

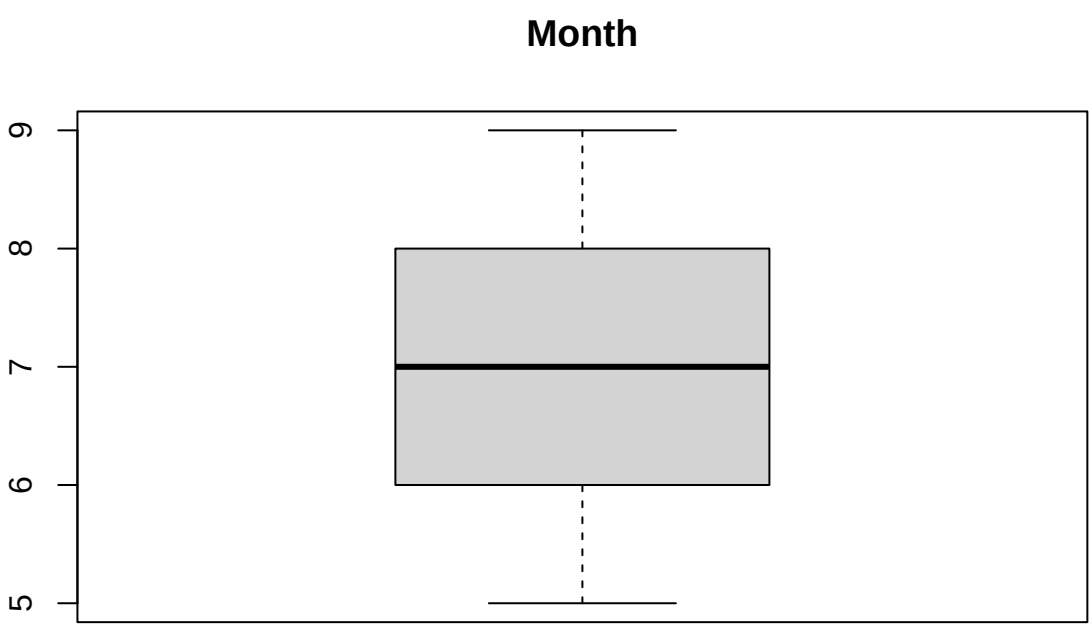
```
for(name in names(airquality)) {
  boxplot(airquality[[name]], main=name)
}
```

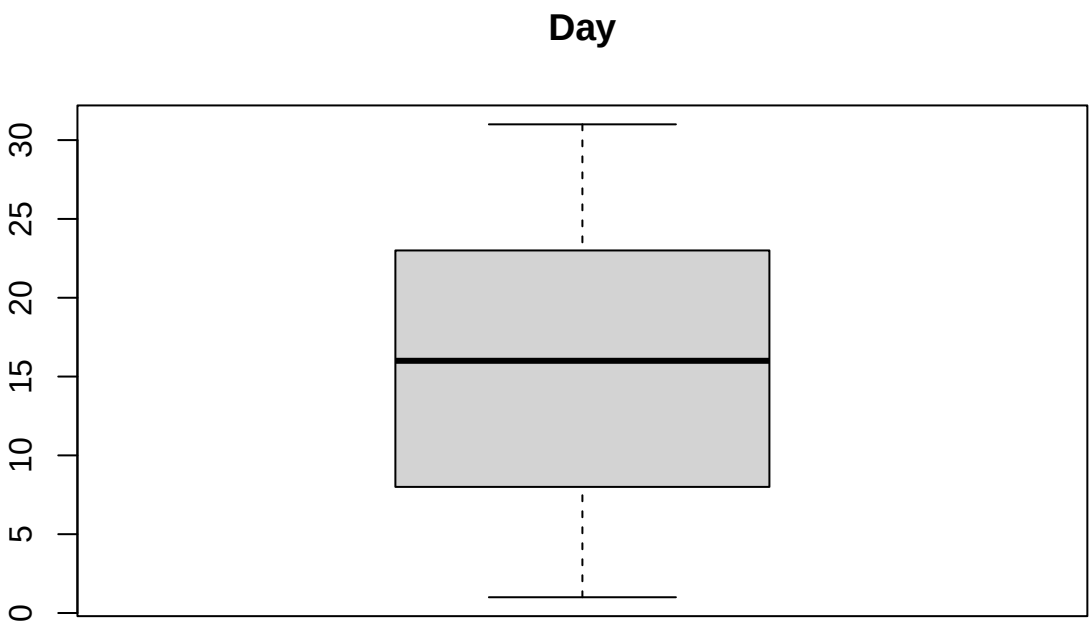


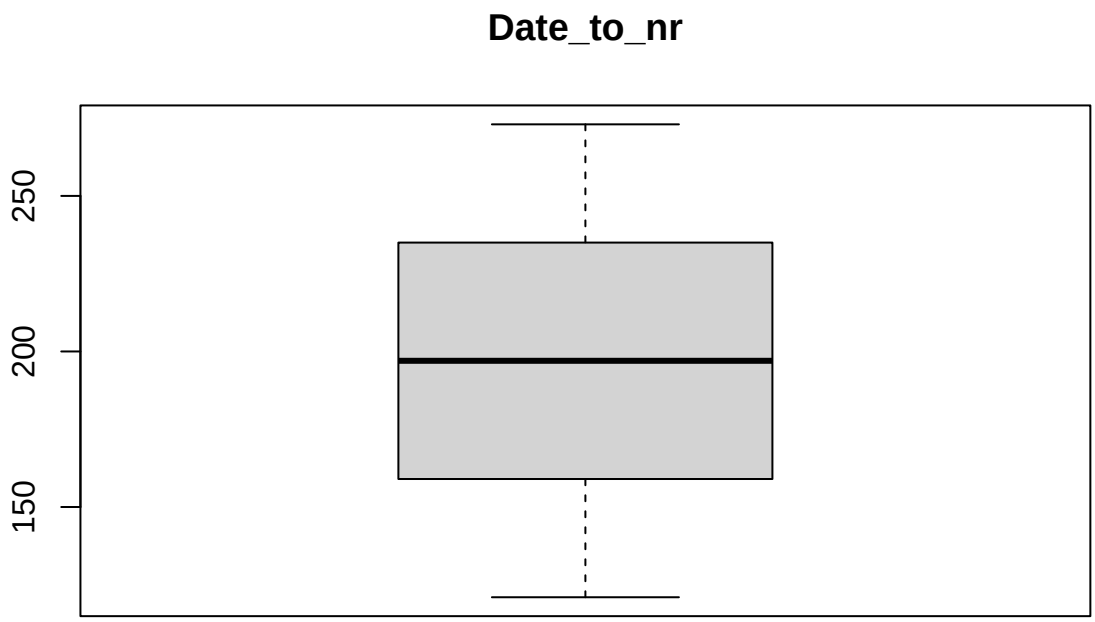

Solar.R
0
50
100
200
300


Wind
20
15
10
5


Temp
90
80
70
60

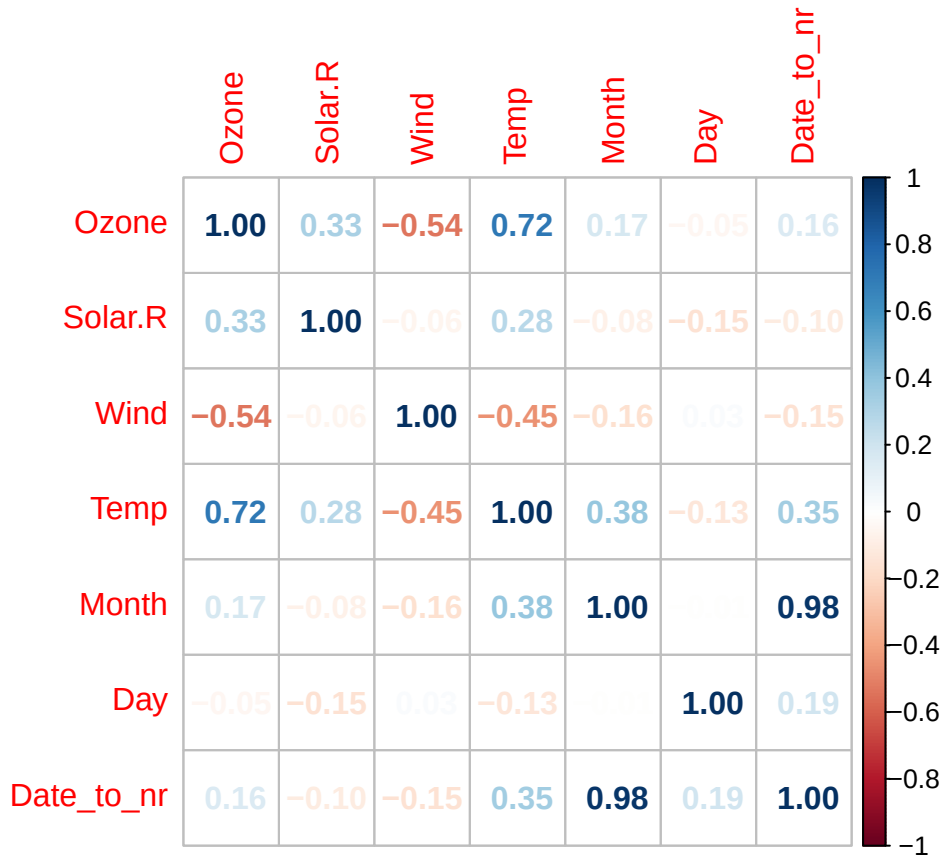

Month
9
8
7
6
5


Day
0
5
10
15
20
25
30



- ozone ha degli outlier alti
- i raggi solari non sembrano avere mai degli outlier
- ci sono stati dei giorni in cui il vento è stato particolarmente forte
- la temperatura non ha mai raggiunto estremi
- Sui valori relativi alle date questa analisi non ha senso

```
corrplot(cor(airquality, use = "complete.obs"), method = 'number')
```



Dalla matrice di correlazione è possibile evidenziare:

- Una correlazione diretta molto forte tra ozono e temperatura
- una correlazione inversa discretamente forte tra temperatura e vento
- altre correlazioni minori

Da questo ho escluso la correlazione tra mese e la variabile artificiale date_to_nr in quanto si tratta di una correlazione artificiale

In questo esercizio abbiamo analizzato il dataset di dati metereologici nella città di New york.

Abbiamo scoperto che la temperatura influenza la presenza di ozono nell'area in modo esponenziale (anche se ben approssimabile a livello lineare per queste temperature) e abbiamo il sospetto che ci sia una correlazione a campana tra ozono e raggi solari.

Inoltre abbiamo scoperto che in quei giorni l'ozono e il vento hanno raggiunto qualche valore anomalo anche se nulla da far pensare ad anomalie preoccupanti