Análise descritiva

Francisco Rubens

Carregando a base

```
data <- read.csv2("usairpollution.csv")
```

Análise

Os dados são de poluição do ar de 41 cidades dos Estados Unidos. Relativos a base, as variáveis são:

Variável	Descrição
X	Cidade dos Estados Unidos
SO2	Teor de SO2 (dióxido de enxofre) do ar em microgramas por metro cúbico.
temp	Temperatura média anual em Fahrenheit.
manu	Número de empresas manufatureiras que empregam 20 ou mais trabalhadores.
popul	Tamanho da população (censo de 1970) em milhares.
wind	Velocidade média anual do vento em milhas por hora.
precip	Precipitação média anual em polegadas.
predays	Número médio de dias com precipitação por ano.

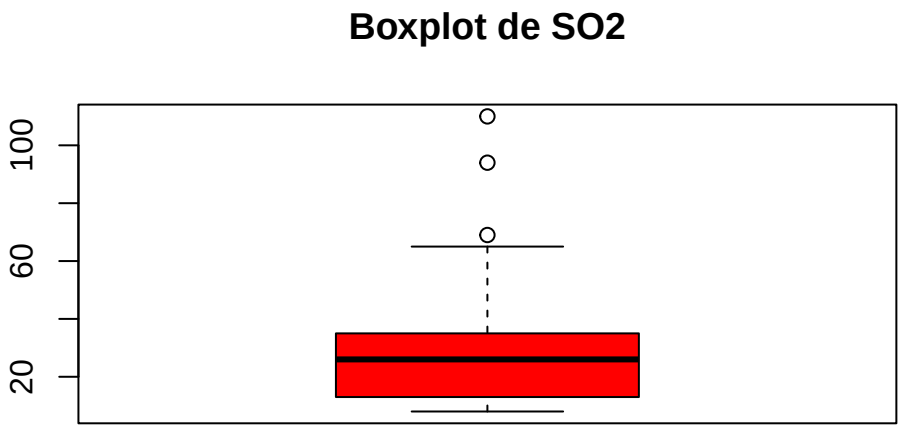
```
summary(data[,-1])
```

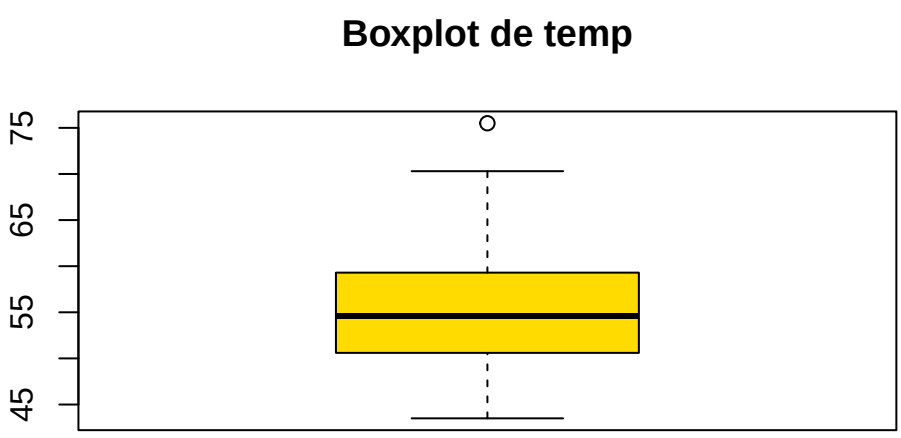
```
      SO2              temp            manu            popul       
 Min.   :  8.00   Min.   :43.50   Min.   :  35.0   Min.   :  71.0  
 1st Qu.: 13.00   1st Qu.:50.60   1st Qu.: 181.0   1st Qu.: 299.0  
 Median : 26.00   Median :54.60   Median : 347.0   Median : 515.0  
 Mean   : 30.05   Mean   :55.76   Mean   : 463.1   Mean   : 608.6  
 3rd Qu.: 35.00   3rd Qu.:59.30   3rd Qu.: 462.0   3rd Qu.: 717.0  
```

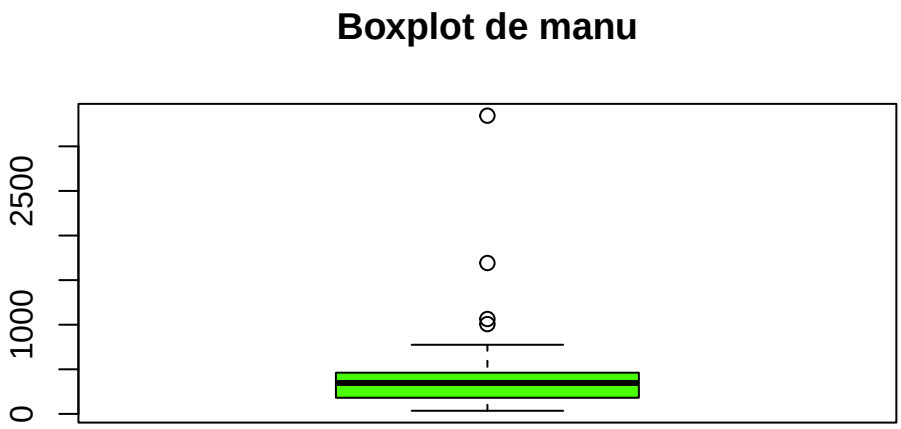
```
Max.   :110.00   Max.   :75.50   Max.   :3344.0   Max.   :3369.0
     wind            precip          predays
Min.   : 6.000   Min.   : 7.05   Min.   : 36.0
1st Qu.: 8.700   1st Qu.:30.96   1st Qu.:103.0
Median : 9.300   Median :38.74   Median :115.0
Mean   : 9.444   Mean   :36.77   Mean   :113.9
3rd Qu.:10.600   3rd Qu.:43.11   3rd Qu.:128.0
Max.   :12.700   Max.   :59.80   Max.   :166.0
```

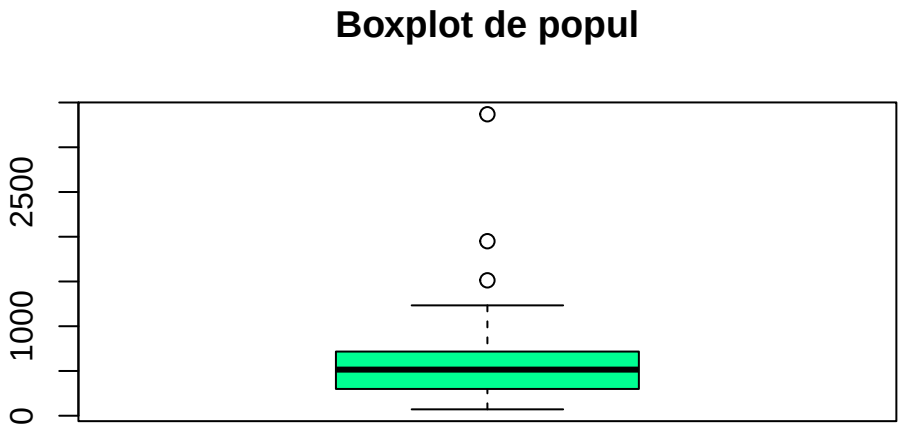
Pelo `summary` é possível ver os mínimos, máximos, médias e os quartis. É possível porceber as variáveis que são bem assimétricas (SO2, manu, popul). Nota-se também que algumas variáveis devem possuir outliers (SO2, manu, popul).

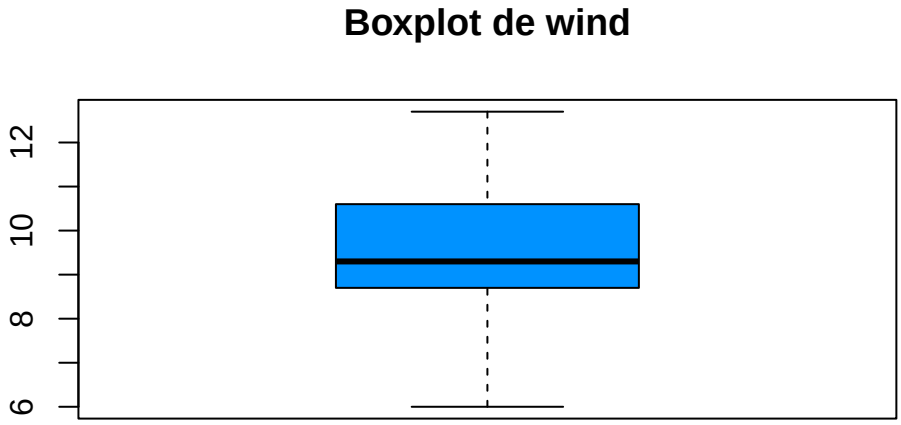
```r
for(i in 2:8){
  boxplot(data[,i], main=paste("Boxplot de",names(data)[i]), col=rainbow(7)[i-1])
}
```

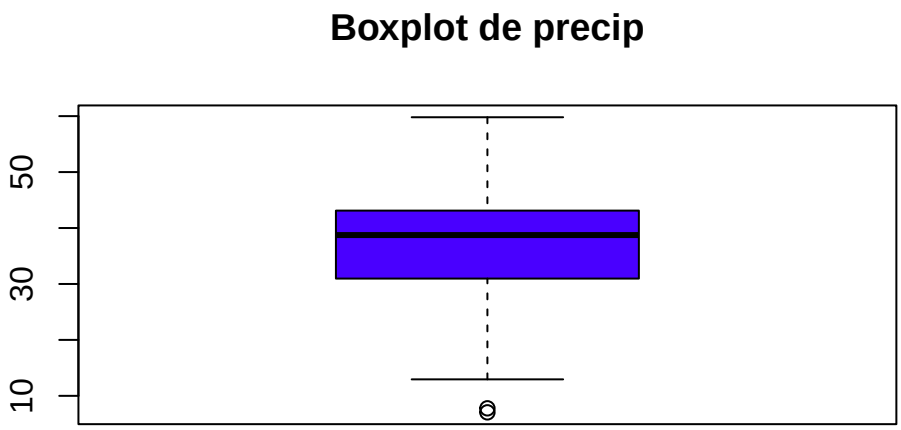


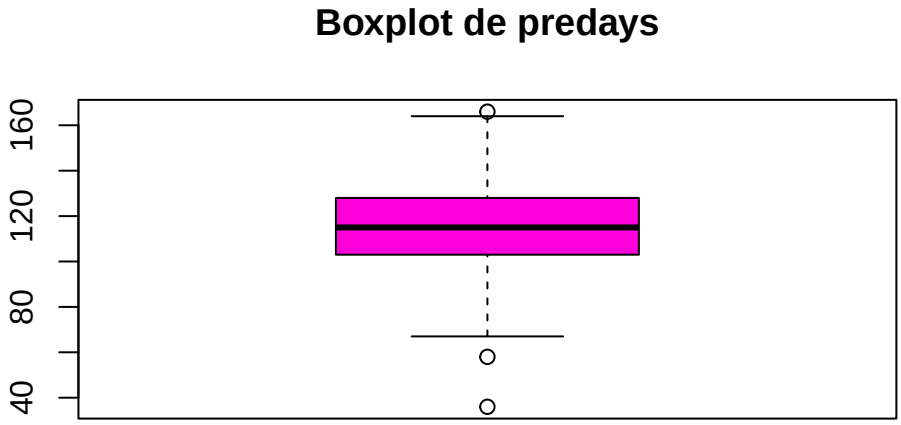

Boxplot de temp
75
65
55
45


Boxplot de manu
2500
1000
0


Boxplot de popul
0
1000
2500


Boxplot de wind
6
8
10
12



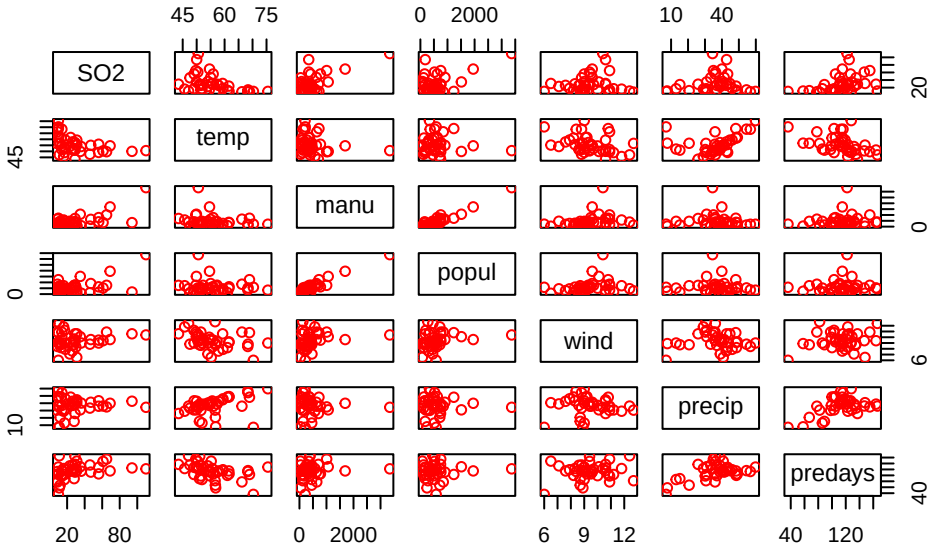


Pelos boxplots é possível perceber as assimetrias e os outliers.

Relação entre as variáveis

A concentração média anual de dióxido de enxofre, em microgramas por metro cúbico, é uma medida da poluição do ar em cidades. A questão de interesse aqui é quais ou como os aspectos do clima e da ecologia humana medidos pelas outras seis variáveis influenciam a poluição?

```
plot(data[,-1], col="red")
```



Por esse gráfico é possível perceber as relação entre as variáveis, e nota-se uma correlação linear forte entre `manu` e `popul`. Já SO2 não parece ter nenhuma correlação forte com outras variáveis, mas é possível perceber alguma correlação com `manu` e `popul`, principalmente com `manu`, em que os valores estão menos dispersos.

Para verificar a influência das variáveis na poluição foram ajustados alguns modelos, o modelo que mais explicou a poluição foi:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Sendo, X_{i1} a variável `manu` e X_{i2} a variável `popul`, além disso, foram removidos os outliers (valores acima de 70) de `SO2`.

```r
df <- data[data$SO2 <= 70,]

fit <- lm(df[,2]~., data = df[,c(-1,-2, -3, -6, -7, -8)])
summary(fit)
```

```
Call:
lm(formula = df[, 2] ~ ., data = df[, c(-1, -2, -3, -6, -7, -8)])

Residuals:
    Min      1Q  Median      3Q     Max
-22.107 -10.828  -2.621   8.853  36.551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.35918    3.82078   5.852 1.10e-06 ***
manu         0.06886    0.01476   4.666 4.15e-05 ***
popul       -0.04193    0.01292  -3.246  0.00253 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.3 on 36 degrees of freedom
Multiple R-squared:  0.4227,    Adjusted R-squared:  0.3907
F-statistic: 13.18 on 2 and 36 DF,  p-value: 5.069e-05
```

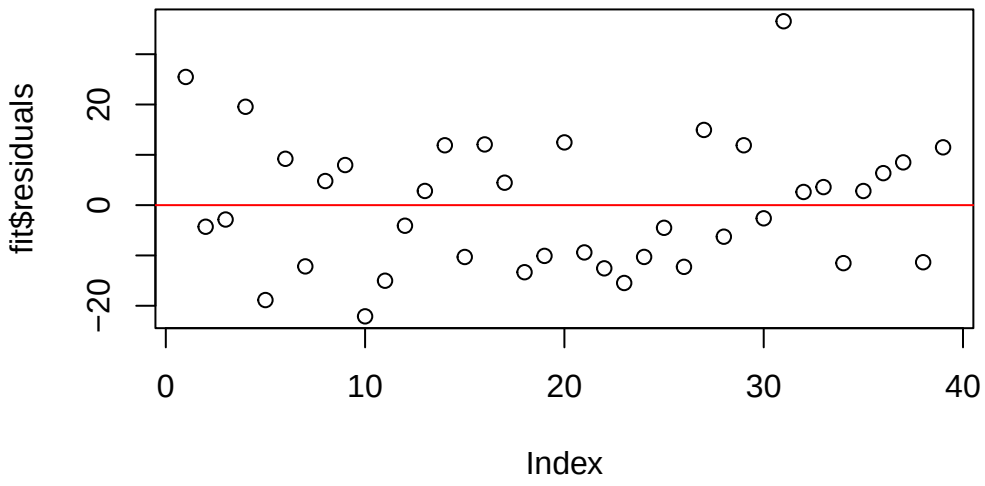
Sendo assim, o modelo ajustado foi:

$$Y_i = 22,36 + 0,07X_{i1} - 0,04X_{i2} + \epsilon_i$$

Modelo que explica 39% da variação dos dados, com todos os coeficiente significativos a 5% de significância.

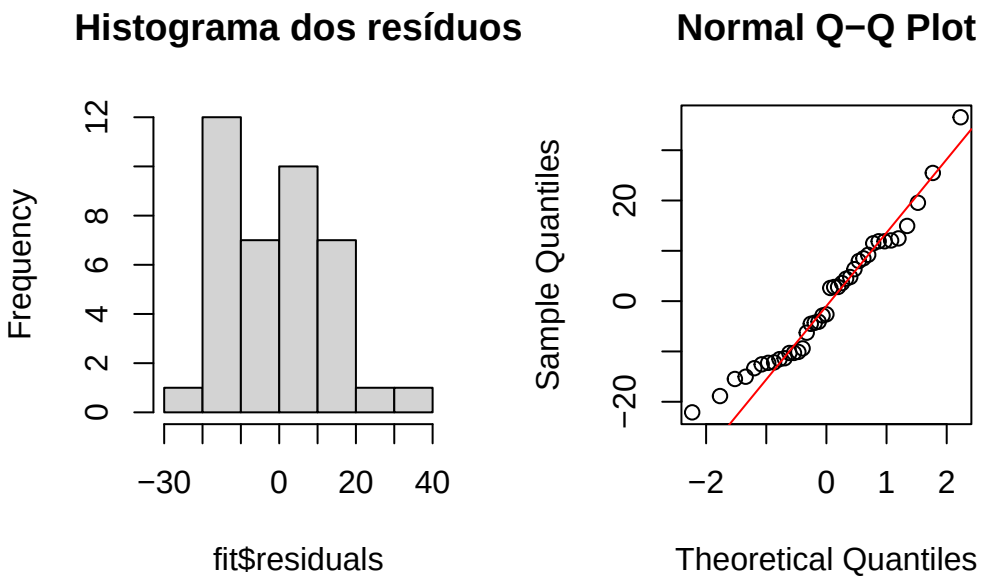
Verificando as suposições

```r
plot(fit$residuals)
abline(h=0, col="red")
```



No gráfico de dispersão não se percebe nenhuma tendência, logo os resíduos parecem estar aleatoriamente distribuidos em torno do zero.

```
par(mfrow=c(1,2))
hist(fit$residuals, main="Histograma dos resíduos")
qqnorm(fit$residuals)
qqline(fit$residuals, col="red")
```



Verificando a suposição de normalidade, vemos uma pequena assimetria no histograma dos resíduos, porém um gráfico os valores concentrados em torno do zero com as caudas mais leves. Já no gráfico Quantil-Quantil da normal, nota-se que apenas as caudas não estão em conformidade com a linha da normal teórica, o que não é um problema.

```
shapiro.test(fit$residuals)
```

```
    Shapiro-Wilk normality test

data:  fit$residuals
W = 0.95846, p-value = 0.1584
```

```
manu <- df[,4]
popul <- df[,5]
bptest(formula(fit))
```

```
    studentized Breusch-Pagan test
```

```
data:  formula(fit)
BP = 0.009341, df = 2, p-value = 0.9953
```

Para complementar a análise gráfica foi feito o teste de Shapiro-Wilk, que não rejeita a hipótese de normalidade (p-valor de 0,1584) e o teste de Breusch-Pagan que não rejeita a hipótese de homogeneidade da variância (p-valor de 0,9953).

Conclusão do modelo

Dado que o modelo parece não violar as suposições do modelo, conclui-se que o número de empresas manufatureiras que empregam 20 ou mais trabalhadores influenciam no aumento do teor de SO2 (dióxido de enxofre) do ar em microgramas por metro cúbico, enquanto o tamanho da população (censo de 1970) em milhares influencia na diminuição.