Análise de Regressão Linear

Nome: Ananda Almeida de Sá
Matrícula: 150117345
Data: 04/05/2022

Introdução

Leitura de dados

O programa utilizado para analisar os dados disponibilizados em Excel será o R Studio, versão 4.2.0, importados como um data frame (planilha), onde as colunas representam as variáveis de estudo e cada linha representa um hospital dos Estados Unidos no período de 1975-1976.

Análise Exploratória

Estatísticas Descritivas

Uma maneira fácil de obter algumas estatísticas descritivas das variáveis em estudo é através do comando *summary()*, que retorna as estatísticas *mínimo*, *quartis*, *média* e *máximo*. Para medir a variabilidade, utilize as funções *var()* e *sd()* para obter a *variância* e o *desvio padrão* outra forma de avaliar essas estatísticas é através do gráfico *boxplot*. Aplicando para cada variável, tem-se

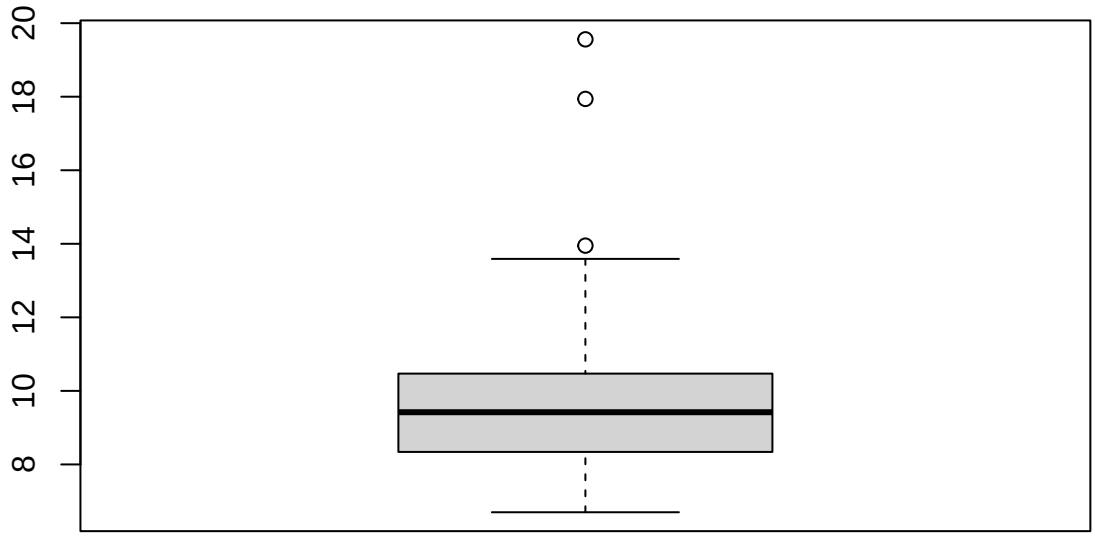
- **Duração da Internação**

A duração de internação é uma variável quantitativa contínua que representa a duração média da internação de todos os pacientes no hospital (em dias). Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Duração da Internação
Mínimo	6.700000
1º Quartil	8.340000
Mediana	9.420000
Média	9.648319
3º Quartil	10.470000
Máximo	19.560000
Variância	3.653664
Desvio Padrão	1.911456

E o *boxplot* da variável é dado por:

```
boxplot(data$`Duração da Internação`)
```



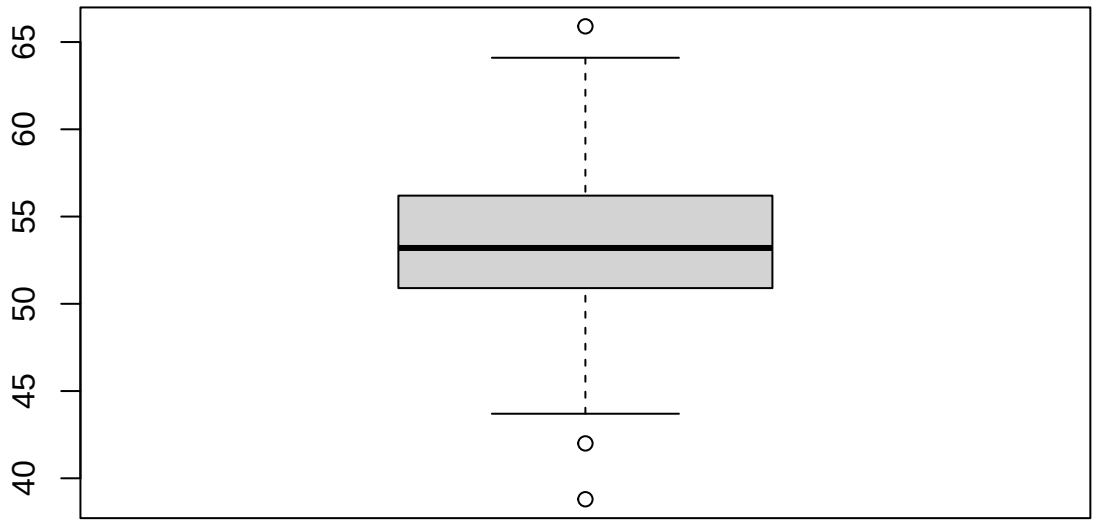
- **Idade**

A idade é uma variável quantitativa contínua que representa a idade média dos pacientes de cada hospital. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Idade
Mínimo	38.800000
1º Quartil	50.900000
Mediana	53.200000
Média	53.231858
3º Quartil	56.200000
Máximo	65.900000
Variância	19.905940
Desvio Padrão	4.461607

E o *boxplot* da variável é dado por:

```
boxplot(data$Idade)
```



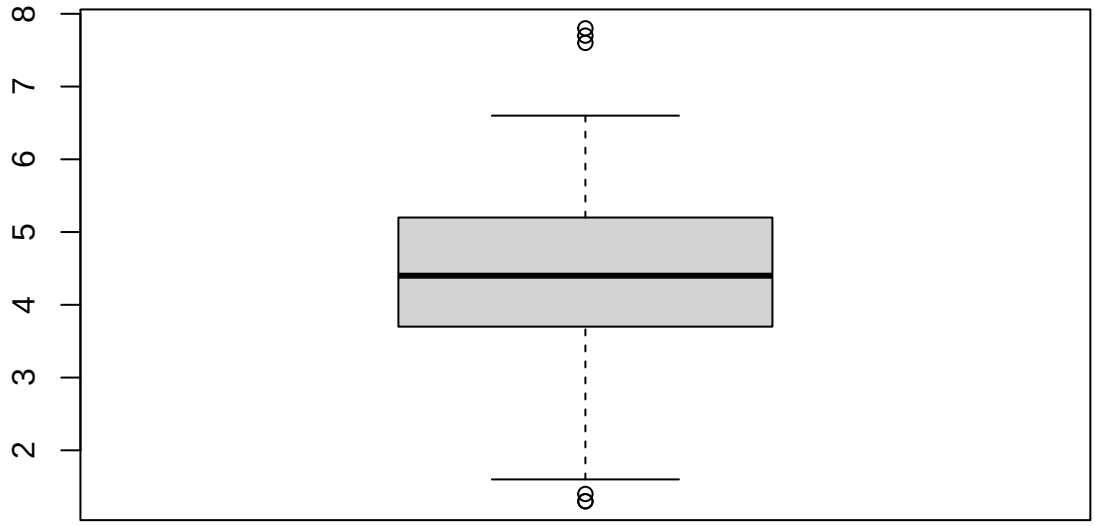
- **Risco de Infecção**

O risco de infecção é uma variável quantitativa contínua que representa a probabilidade média estimada de adquirir infecção no hospital (em %). Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Risco de Infecção
Mínimo	1.300000
1º Quartil	3.700000
Mediana	4.400000
Média	4.354867
3º Quartil	5.200000
Máximo	7.800000
Variância	1.798034
Desvio Padrão	1.340908

E o *boxplot* da variável é dado por:

```
boxplot(data$`Risco de Infecção`)
```



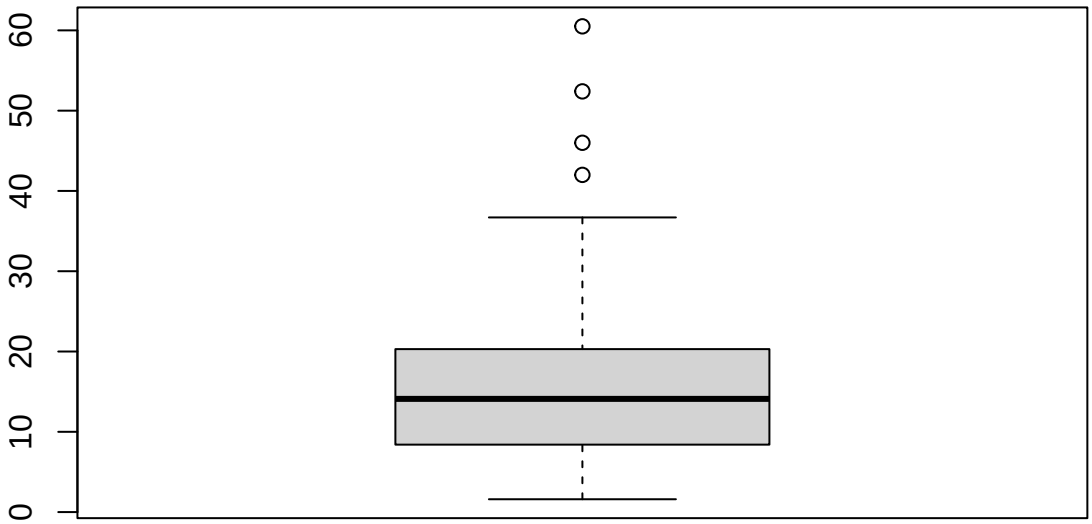
- **Proporção de Culturas de Rotina**

A proporção de culturas de rotina é uma variável quantitativa contínua que representa a razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infeção adquirida no hospital, vezes 100. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Proporção de Culturas de Rotina
Mínimo	1.60000
1º Quartil	8.40000
Mediana	14.10000
Média	15.79292
3º Quartil	20.30000
Máximo	60.50000
Variância	104.74924
Desvio Padrão	10.23471

E o *boxplot* da variável é dado por:

```
boxplot(data$`Proporção de Culturas de Rotina`)
```



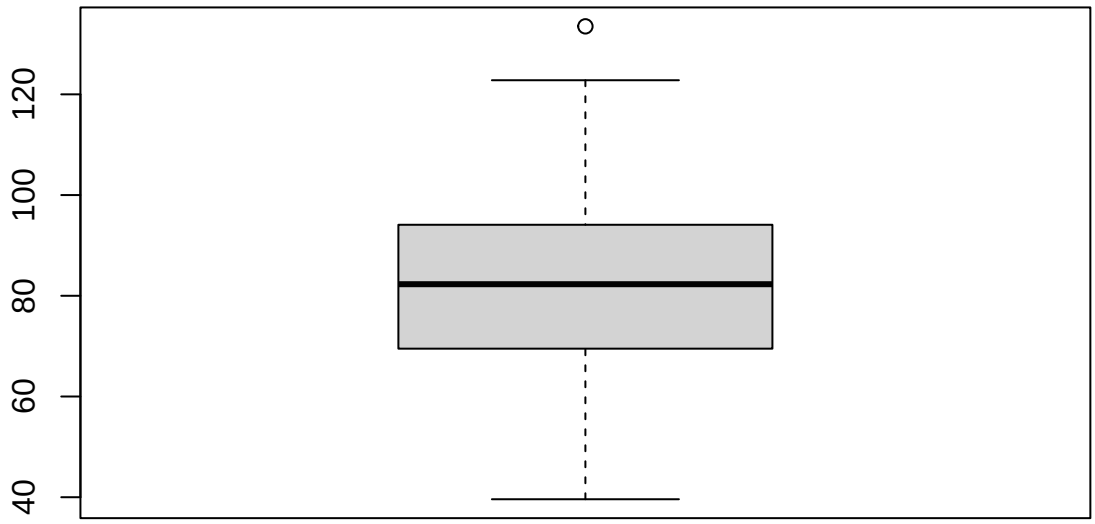
* **Proporção de Raio-X de Tórax de Rotina**

A proporção de raio-X de tórax de rotina é uma variável quantitativa contínua que representa a razão do número de raio-X de tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Proporção de Raio-X de Tórax de Rotina
Mínimo	39.60000
1º Quartil	69.50000
Mediana	82.30000
Média	81.62832
3º Quartil	94.10000
Máximo	133.50000
Variância	374.95776
Desvio Padrão	19.36383

E o *boxplot* da variável é dado por:

```
boxplot(data$`Proporção de Raio-X de Tórax de Rotina`)
```



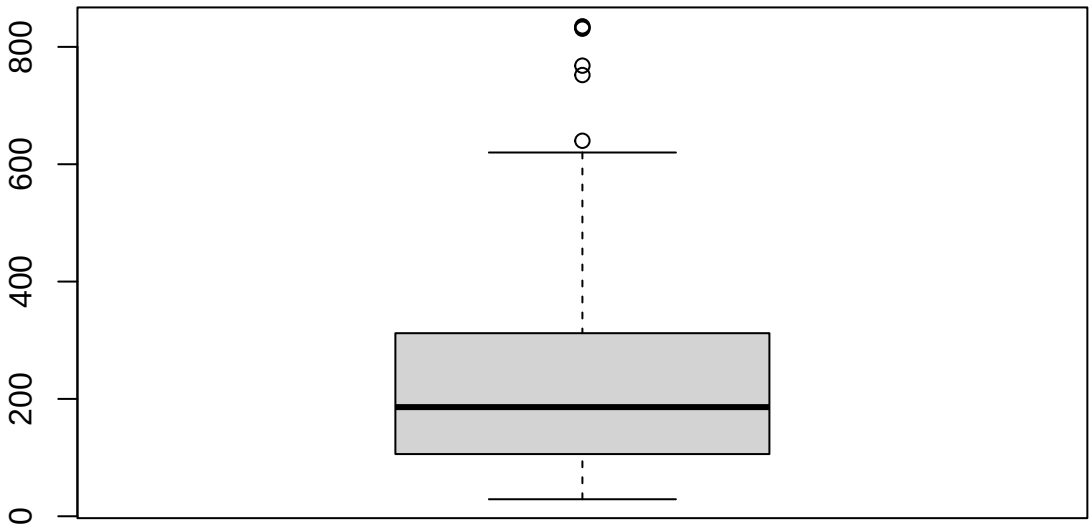
- **Número de leitos**

O número de leitos é uma variável quantitativa discreta que representa o número médio de leitos no hospital durante o período de estudo. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Número de leitos
Mínimo	29.0000
1º Quartil	106.0000
Mediana	186.0000
Média	252.1681
3º Quartil	312.0000
Máximo	835.0000
Variância	37188.3018
Desvio Padrão	192.8427

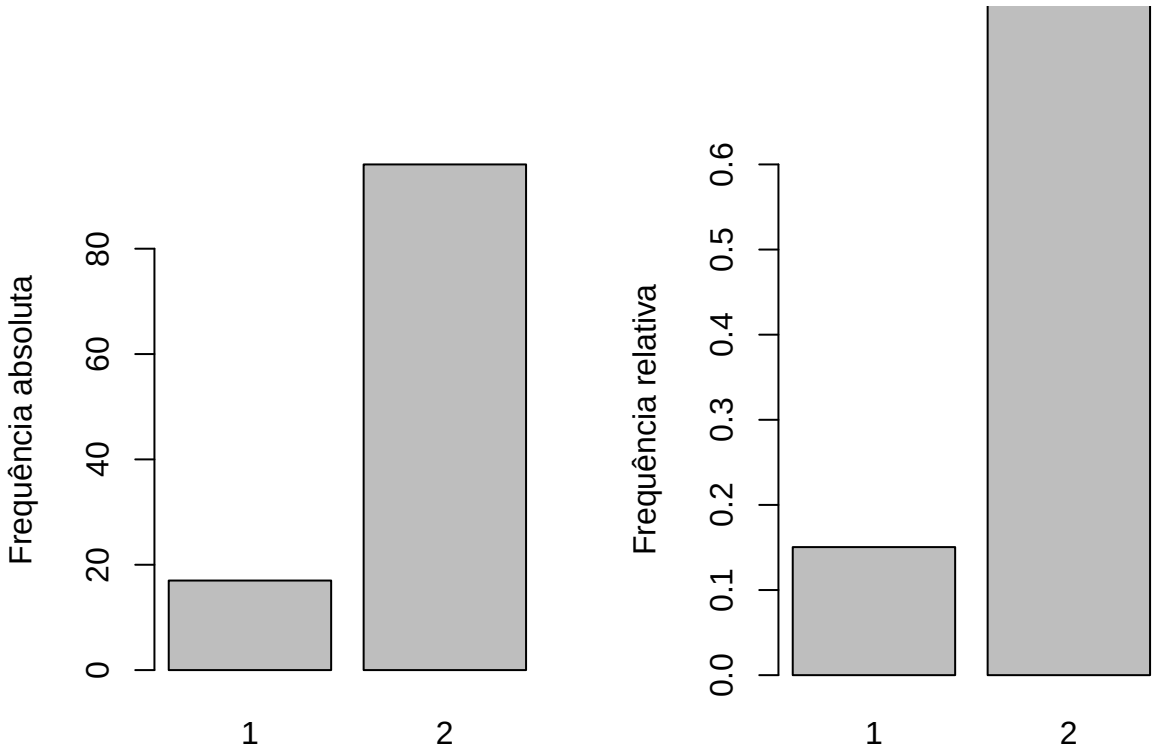
E o *boxplot* da variável é dado por:

```
boxplot(data$`Número de leitos`)
```



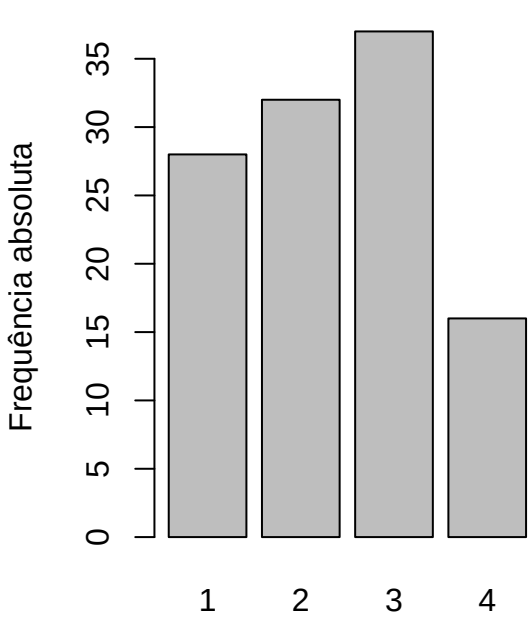
- **Filiação a Escola de Medicina**

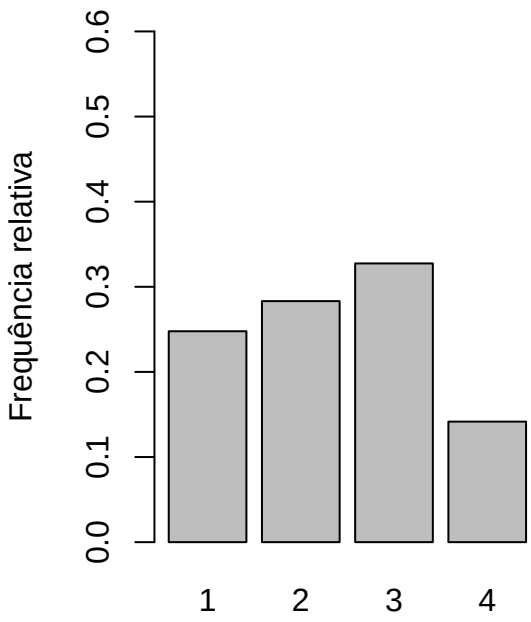
A filiação a escola de medicina é uma variável qualitativa ordinal onde o 1 significa que a escola tem filiação, e 2 que não tem. Os resultados das funções mencionadas anteriormente estão descritos a seguir.



- **Região**

A região é uma variáveis qualitativas ordinais onde “NE” se refere ao Nordeste, “W” ao Oeste, “S” ao Sul e “NC” ao ?????. Os resultados das funções mencionadas anteriormente estão descritos a seguir.





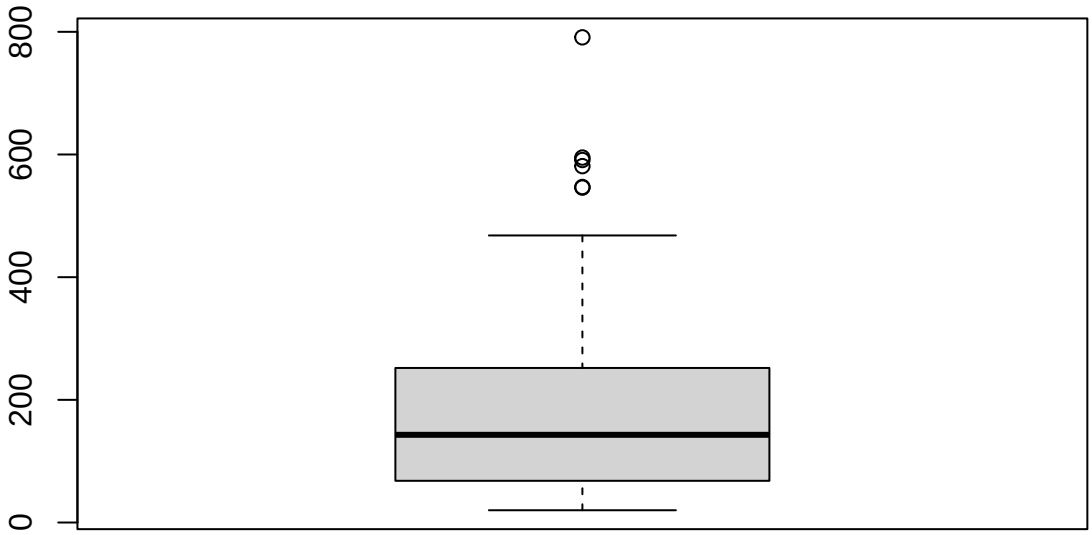
- **Média diária de pacientes**

O média diária de pacientes é uma variável quantitativa discreta que representa o número médio de pacientes no hospital por dia durante o período do estudo. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Média diária de pacientes
Mínimo	20.0000
1º Quartil	68.0000
Mediana	143.0000
Média	191.3717
3º Quartil	252.0000
Máximo	791.0000
Variância	23642.0035
Desvio Padrão	153.7596

E o *boxplot* da variável é dado por:

```
boxplot(data$`Média diária de pacientes`)
```



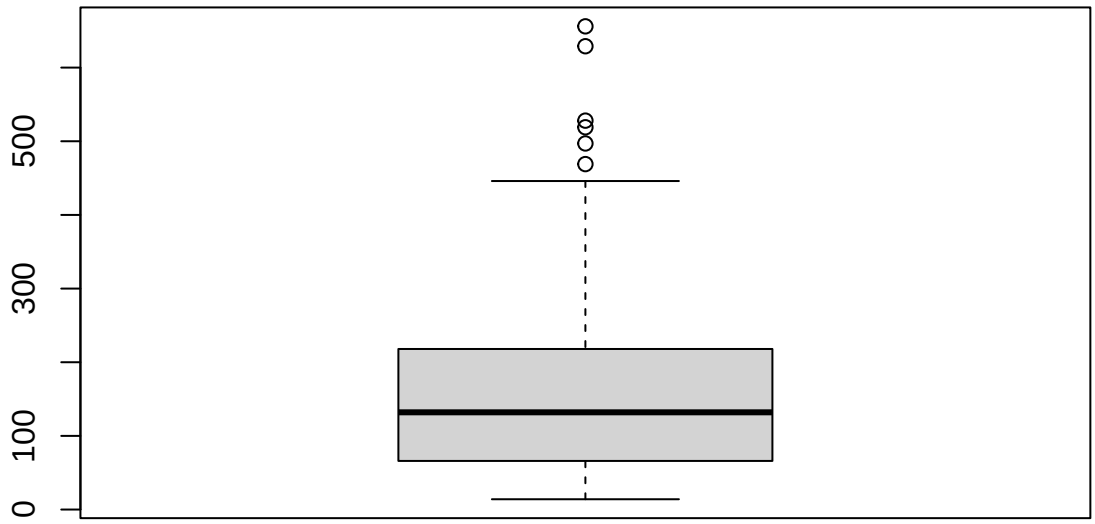
- **Número de enfermeiro(s)**

O número de enfermeiro(s) é uma variável quantitativa discreta que representa o Número médio de enfermeiros(as) de tempo-integral ou equivalente registrados e licenciados durante o período de estudo (número de tempos integrais+metade do número de tempo parcial). Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Número de enfermeiro(s)
Mínimo	14.0000
1º Quartil	66.0000
Mediana	132.0000
Média	173.2478
3º Quartil	218.0000
Máximo	656.0000
Variância	19394.8488
Desvio Padrão	139.2654

E o *boxplot* da variável é dado por:

```
boxplot(data$`Número de enfermeiro(s)`)
```



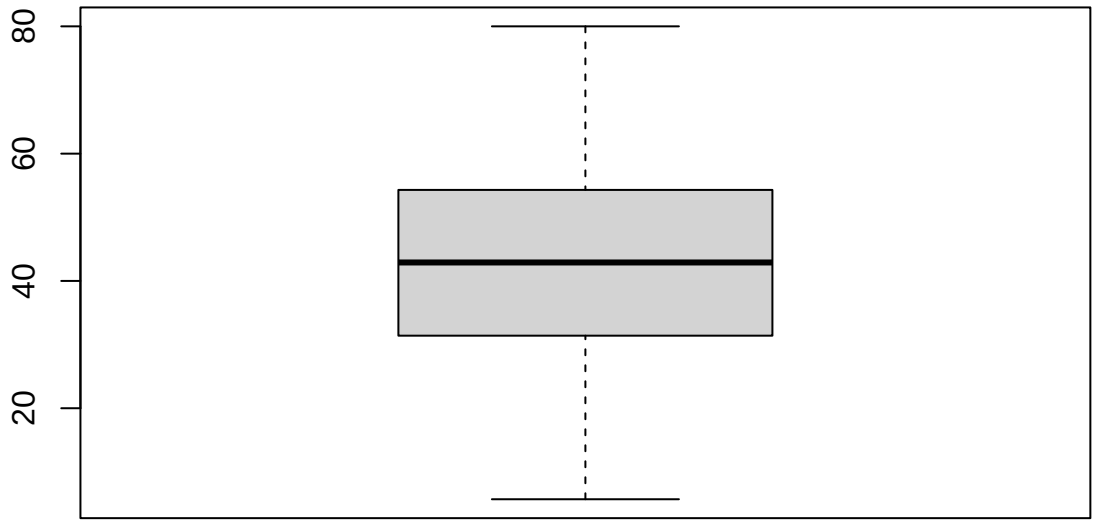
- **Facilidades e serviços disponíveis**

A facilidades e serviços disponíveis é uma variável quantitativa contínua que representa a porcentagem de 35 potenciais facilidades e serviços que são fornecidos pelo hospital. Os resultados das funções mencionadas anteriormente estão descritos a seguir.

	Facilidades e serviços disponíveis
Mínimo	5.70000
1º Quartil	31.40000
Mediana	42.90000
Média	43.15929
3º Quartil	54.30000
Máximo	80.00000
Variância	231.06619
Desvio Padrão	15.20086

E o *boxplot* da variável é dado por:

```
boxplot(data$`Facilidades e serviços disponíveis`)
```



Correlação entre as variáveis

Para verificar a natureza e a força da relação entre as variáveis e identificar lacunas e pontos discrepantes no conjunto de dados, utiliza-se a matriz de correlação aplicado no script a seguir.

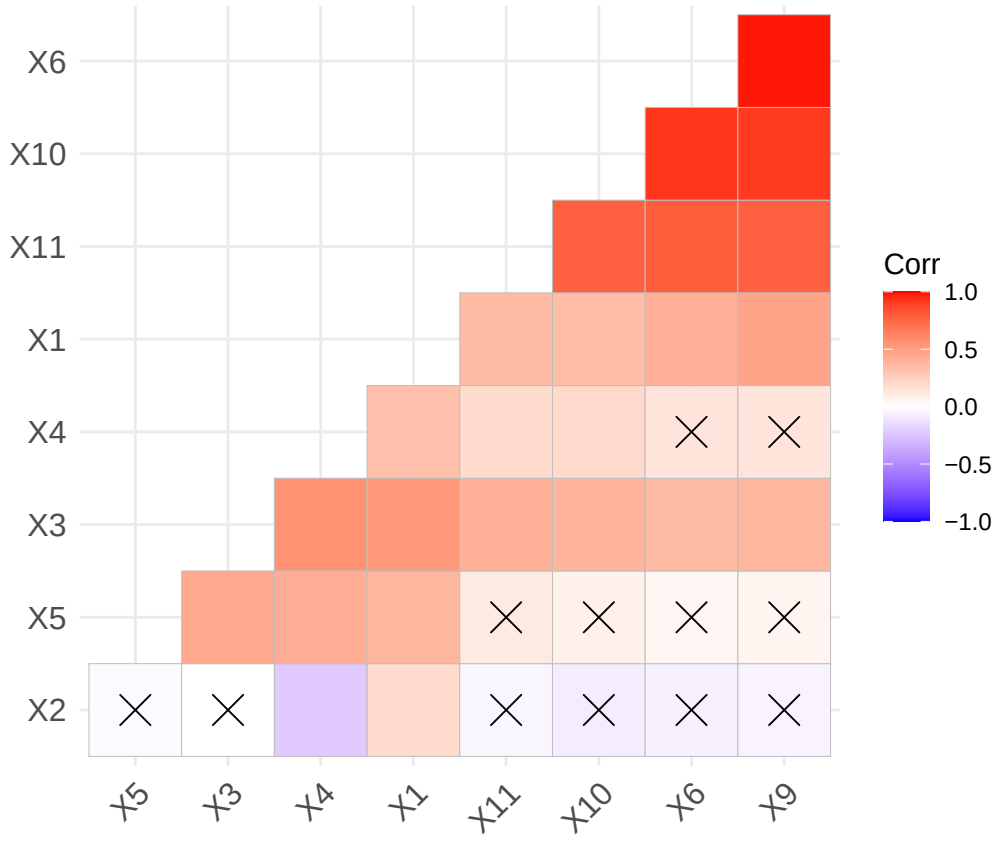
```
library(ggcorrplot)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
library(dplyr)
```

```
##
## 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
```

```
pmat = datax %>% select_if(is.numeric) %>%cor_pmat()

datax %>% select_if(is.numeric) %>% cor(.) %>%
  ggcorrplot( type = "lower", p.mat = pmat, hc.order = TRUE)
```




```
knitr::kable(cbind(names(data),names(datax)))
```

Número de Identificação	ID
Duração da Internação	X1
Idade	X2
Risco de Infecção	X3
Proporção de Culturas de Rotina	X4
Proporção de Raio-X de Tórax de Rotina	X5
Número de leitos	X6
Filiação a Escola de Medicina	X7
Região	X8
Média diária de pacientes	X9
Número de enfermeiro(s)	X10
Facilidades e serviços disponíveis	X11

nesta matriz, percebe-se que quanto mais vermelho, maior é a correlação positiva, no qual as variável de facilidades e serviços disponiveis está positivamente correlacionada com o número de enfermeiro(s), número de leitos e a média diaria de pacientes. Para o número de leitos, há uma alta correlação positiva com a média diária depacientes.

Para verificar a ausência significativa de correlação, o teste de ausência de correlação com nível de significância de 5%, no gráfico acima, temos que as variáveis que são marcadas com X, o teste não foi significante, indicando independência entre estas variáveis, entre elas, a idade do paciente, não tem correlação com o risco de infeção do hospital.

Objetivo

Número de enfermeira(o)s

Deseja-se estudar se o número de enfermeira(o)s está relacionado às instalações, ou seja, os números de leitos do hospital, e se há diferenças entre os serviços disponíveis pelos hospitais. Neste caso, a variável resposta é o número de enfermeira(o)s e as duas outras variáveis são explicativas. Para isso, faz-se necessário a aplicação da regressão linear múltipla realizada no script a seguir:

```
# teste de ausencia de regresao (significativo)
summary(aov(`Número de enfermeiro(s)` ~ `Número de leitos`+ `Facilidades e serviços disponíveis`, data=
```

```
##                                      Df  Sum Sq Mean Sq F value Pr(>F)
## `Número de leitos`                    1 1820644 1820644 601.362 <2e-16 ***
## `Facilidades e serviços disponíveis`  1   18550   18550   6.127 0.0148 *
## Residuals                           110  333029    3028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
modcomp<-lm(`Número de enfermeiro(s)` ~ `Número de leitos` +
              I(`Número de leitos`^2)+
              `Facilidades e serviços disponíveis`+
              I(`Facilidades e serviços disponíveis`^2)+
              `Facilidades e serviços disponíveis`*`Número de leitos`
              , data = data)

summary(modcomp)
```

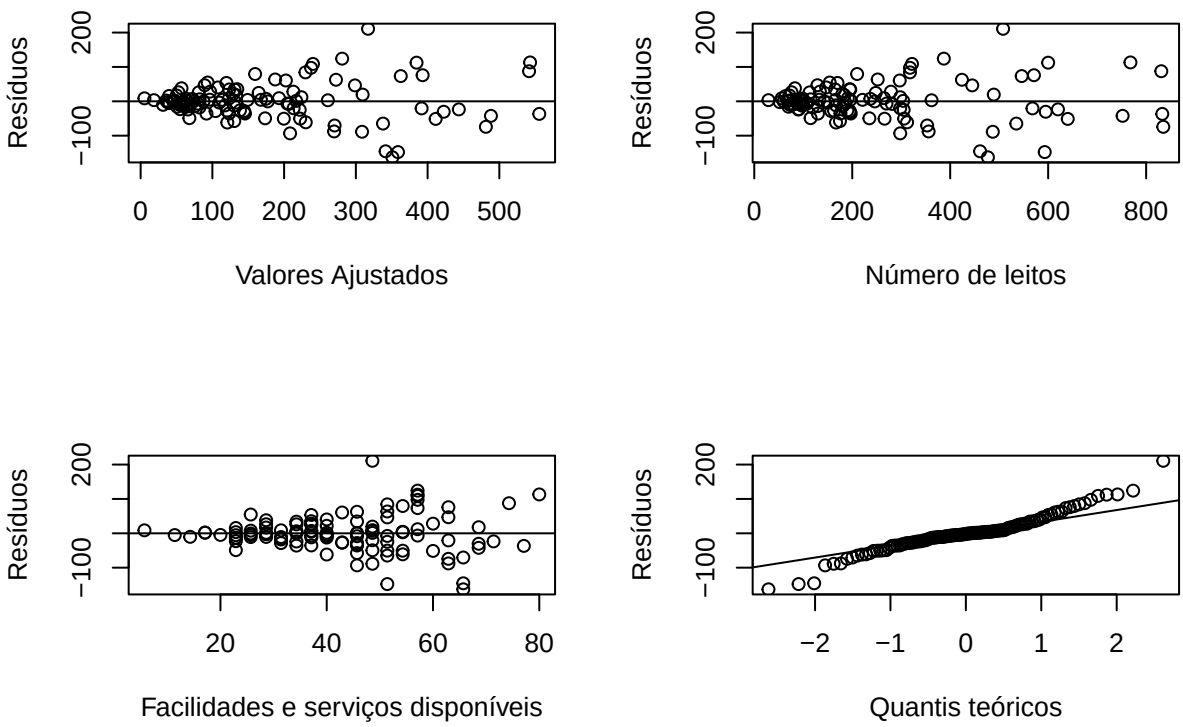
```
##
## Call:
## lm(formula = `Número de enfermeiro(s)` ~ `Número de leitos` +
##     I(`Número de leitos`^2) + `Facilidades e serviços disponíveis` +
##     I(`Facilidades e serviços disponíveis`^2) + `Facilidades e serviços disponíveis` *
##     `Número de leitos`, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.973  -24.758   -2.339   21.880  210.964
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -3.125e+01  3.698e+01
## `Número de leitos`                                4.217e-01  2.023e-01
## I(`Número de leitos`^2)                          -5.006e-04  2.686e-04
## `Facilidades e serviços disponíveis`              2.212e+00  2.350e+00
## I(`Facilidades e serviços disponíveis`^2)        -3.646e-02  3.770e-02
```

```
## `Número de leitos`:`Facilidades e serviços disponíveis`  9.797e-03  5.920e-03
##                                                        t value Pr(>|t|)
## (Intercept)                                             -0.845   0.3999
## `Número de leitos`                                       2.084   0.0395 *
## I(`Número de leitos`^2)                                 -1.864   0.0651 .
## `Facilidades e serviços disponíveis`                     0.941   0.3486
## I(`Facilidades e serviços disponíveis`^2)               -0.967   0.3356
## `Número de leitos`:`Facilidades e serviços disponíveis`  1.655   0.1009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.89 on 107 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8447
## F-statistic: 122.8 on 5 and 107 DF,  p-value: < 2.2e-16
```

```
shapiro.test(residuals(modcomp))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modcomp)
## W = 0.94595, p-value = 0.0001791
```

```
# windows()
par(mfrow = c(2, 2))
plot(fitted(modcomp), residuals(modcomp),xlab="Valores Ajustados",ylab="Resíduos")
abline(h=0)
plot(data$`Número de leitos`, residuals(modcomp),xlab="Número de leitos",ylab="Resíduos")
abline(h=0)

plot(data$`Facilidades e serviços disponíveis`, residuals(modcomp),xlab="Facilidades e serviços disponí
abline(h=0)

qqnorm(residuals(modcomp), ylab="Resíduos",xlab="Quantis teóricos",main="")
qqline(residuals(modcomp))
```



Número de enfermeira(o)s

```
# Teste de ausencia de regressao
summary(aov(`Duração da Internação` ~
              `Proporção de Raio-X de Tórax de Rotina`+
              `Proporção de Culturas de Rotina`,
            data = data))
```

```
##                                        Df Sum Sq Mean Sq F value   Pr(>F)
## `Proporção de Raio-X de Tórax de Rotina`   1   59.9   59.86  19.605 2.26e-05
## `Proporção de Culturas de Rotina`          1   13.5   13.46   4.406   0.0381
## Residuals                                110  335.9    3.05
##
## `Proporção de Raio-X de Tórax de Rotina` ***
## `Proporção de Culturas de Rotina`        *
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Regressao
modcomp<-lm(`Duração da Internação` ~
              `Proporção de Raio-X de Tórax de Rotina`+
              `Proporção de Culturas de Rotina`,
```

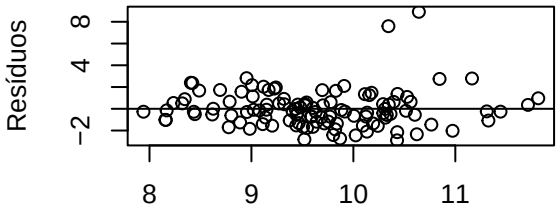
```
            data = data)
summary(modcomp)
```

```
##
## Call:
## lm(formula = `Duração da Internação` ~ `Proporção de Raio-X de Tórax de Rotina` +
##     `Proporção de Culturas de Rotina`, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8987 -1.0846 -0.2387  0.6384  8.9177
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            6.66147    0.71664   9.295 1.61e-15
## `Proporção de Raio-X de Tórax de Rotina`  0.02935    0.00942   3.116  0.00234
## `Proporção de Culturas de Rotina`         0.03741    0.01782   2.099  0.03809
##
## (Intercept)                           ***
## `Proporção de Raio-X de Tórax de Rotina` **
## `Proporção de Culturas de Rotina`        *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.747 on 110 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1642
## F-statistic: 12.01 on 2 and 110 DF,  p-value: 1.922e-05
```

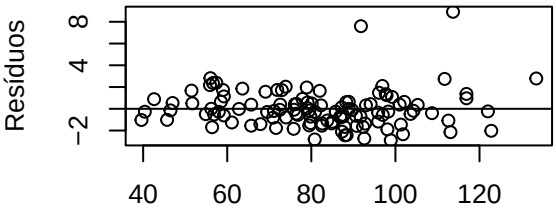
```
# Teste de normalidade
shapiro.test(residuals(modcomp))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modcomp)
## W = 0.86649, p-value = 1.146e-08
```

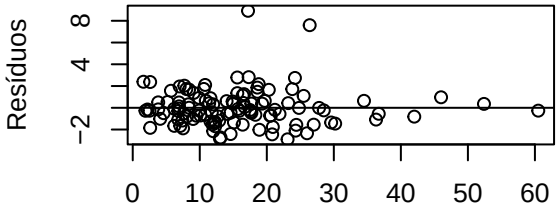
```
# bartlett.test(residuals(modcomp))



# windows()
par(mfrow = c(2, 2))
plot(fitted(modcomp), residuals(modcomp),xlab="Valores Ajustados",ylab="Resíduos")
abline(h=0)
plot(data$`Proporção de Raio-X de Tórax de Rotina`, residuals(modcomp),xlab="Proporção de Raio-X de Tóra
abline(h=0)

plot(data$`Proporção de Culturas de Rotina`, residuals(modcomp),xlab="Proporção de Culturas de Rotina",
abline(h=0)

qqnorm(residuals(modcomp), ylab="Resíduos",xlab="Quantis teóricos",main="")
qqline(residuals(modcomp))
```



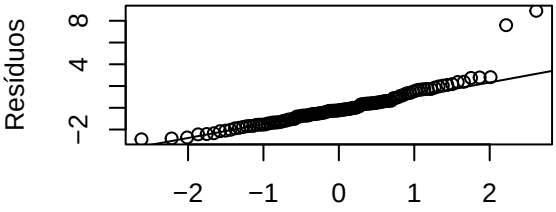
Valores Ajustados



Proporção de Raio−X de Tórax de Rotina



Proporção de Culturas de Rotina



Quantis teóricos