

Case Calina

Roger Gregório Marcondes

Abaixo as instruções do case:

Em uma agência de Marketing Digital uma das épocas mais importante para o ramo de e-commerce é a “Black Friday”, período sazonal em que muitos dos clientes se planejam com promoções e ações através das mídias pagas para chamar a atenção dos usuários. Pensando nessa data muito especial, um cliente da Calina solicitou uma análise para prevermos qual será a receita da Black Friday de 2020.

O banco de dados enviado contém dados de 3 mídias em que o cliente investe (Mídia A, B e C) e o total da receita gerada no site por semana, desde a primeira semana de 2018 até a última semana de outubro de 2020.

Para responder ao cliente análise o banco de dados, crie um modelo teste e um modelo final que deve prever as próximas 4 semanas, respectivas ao mês de novembro de 2020 (a última semana é a semana da Black Friday).

```
# Leitura dos dados
library(readxl)
CienciadeDados <- read_excel("C:/Users/will_/Documents/roger/CienciadeDados.xlsx",
  col_types = c("date", "numeric", "numeric",
    "numeric", "numeric"))
head(CienciadeDados, 6)
```

```
## # A tibble: 6 x 5
##   Week                A      B      C Receita
##   <dtm>             <dbl> <dbl> <dbl>   <dbl>
## 1 2018-01-01 00:00:00 3338.  174.  8748. 458913.
## 2 2018-01-08 00:00:00 5444.  200.  9345 560616.
## 3 2018-01-15 00:00:00 4703.  227.  9383 452347.
## 4 2018-01-22 00:00:00 6479.  197. 10742 510146.
## 5 2018-01-29 00:00:00 7381.  192. 13849 573287.
## 6 2018-02-05 00:00:00 6627.  214. 13889 392558.
```

Vamos eliminar a variável “Week” e visualizar a estrutura das colunas restantes que utilizaremos para o modelo.

```
cols <- c(colnames(CienciadeDados)[-1])
CienciadeDados <- CienciadeDados[,cols]
str(CienciadeDados)
```

```
## tibble [148 x 4] (S3: tbl_df/tbl/data.frame)
## $ A      : num [1:148] 3338 5444 4703 6479 7381 ...
## $ B      : num [1:148] 174 200 227 197 192 ...
## $ C      : num [1:148] 8748 9345 9383 10742 13849 ...
## $ Receita: num [1:148] 458913 560616 452347 510146 573287 ...
```

```
# Análise descritiva dos dados
library(psych)
headTail(CienciadeDados)
```

```
##           A           B           C  Receita
## 1  3337.71 174.17  8747.98 458912.58
## 2   5443.7 200.34    9345 560615.92
## 3  4703.43 227.28  9382.93 452347.24
## 4  6478.87 196.63 10741.63 510145.61
## 5      ...      ...      ...      ...
## 6 11703.84 478.72 18067.81  443026.8
## 7 11949.66 578.42 19751.82 516503.72
## 8 12172.7 625.48 19538.84 519165.64
## 9 11122.95 571.25 17593.92 531238.89
```

```
summary(CienciadeDados)
```

```
##           A           B           C           Receita
## Min.      : 3338   Min.      : 174.2   Min.      : 1704   Min.      : 246277
## 1st Qu.: 8575   1st Qu.: 356.3   1st Qu.:13540   1st Qu.: 446826
## Median :10632   Median : 485.8   Median :15986   Median : 516138
## Mean    :10254   Mean    : 533.4   Mean    :15942   Mean    : 549261
## 3rd Qu.:11949   3rd Qu.: 611.2   3rd Qu.:18383   3rd Qu.: 597870
## Max.    :21938   Max.    :2967.8   Max.    :31543   Max.    :1311738
```

```
# Lendo os dados do ano 2018
```

```
library(readxl)
```

```
X2018 <- read_excel("C:/Users/will/Documents/roger/2018.xltx", col_types = c("numeric",
  "numeric", "numeric", "numeric"))
```

```
head(X2018, 6)
```

```
## # A tibble: 6 x 4
##       A       B       C Receita
##   <dbl> <dbl> <dbl>   <dbl>
## 1 3338.  174.  8748. 458913.
## 2 5444.  200.  9345  560616.
## 3 4703.  227.  9383. 452347.
## 4 6479.  197. 10742. 510146.
## 5 7381.  192. 13849. 573287.
## 6 6627.  214. 13889. 392558.
```

```
# Análise descritiva dos dados
```

```
library(psych)
```

```
headTail(X2018)
```

```
##           A           B           C  Receita
## 1  3337.71 174.17  8747.98 458912.58
## 2   5443.7 200.34    9345 560615.92
## 3  4703.43 227.28  9382.93 452347.24
## 4  6478.87 196.63 10741.63 510145.61
## 5      ...      ...      ...      ...
## 6  5406.53 1182.78 14261.55 427266.37
## 7 10379.88 1135.69 12502.12 452414.13
## 8  9473.46 1767.9 11495.16 397445.87
## 9  7754.31 2967.76 12521.37 430528.42
```

```
str(X2018)
```

```
## tibble [53 x 4] (S3: tbl_df/tbl/data.frame)
## $ A      : num [1:53] 3338 5444 4703 6479 7381 ...
## $ B      : num [1:53] 174 200 227 197 192 ...
## $ C      : num [1:53] 8748 9345 9383 10742 13849 ...
## $ Receita: num [1:53] 458913 560616 452347 510146 573287 ...
```

```
summary(X2018)
```

##	A	B	C	Receita
## Min.	: 3338	Min. : 174.2	Min. : 8748	Min. : 366691
## 1st Qu.:	5949	1st Qu.: 277.5	1st Qu.:12502	1st Qu.: 429839
## Median :	7659	Median : 315.0	Median :13644	Median : 471429
## Mean :	8256	Mean : 479.4	Mean :13804	Mean : 498007
## 3rd Qu.:	10432	3rd Qu.: 430.4	3rd Qu.:15287	3rd Qu.: 515772
## Max.	:21050	Max. :2967.8	Max. :19958	Max. :1311738

```
par(mfrow = c(1,3))
```

```
# Histograma da variável Mídia A em relação a receita gerada no site por semana em 201
```

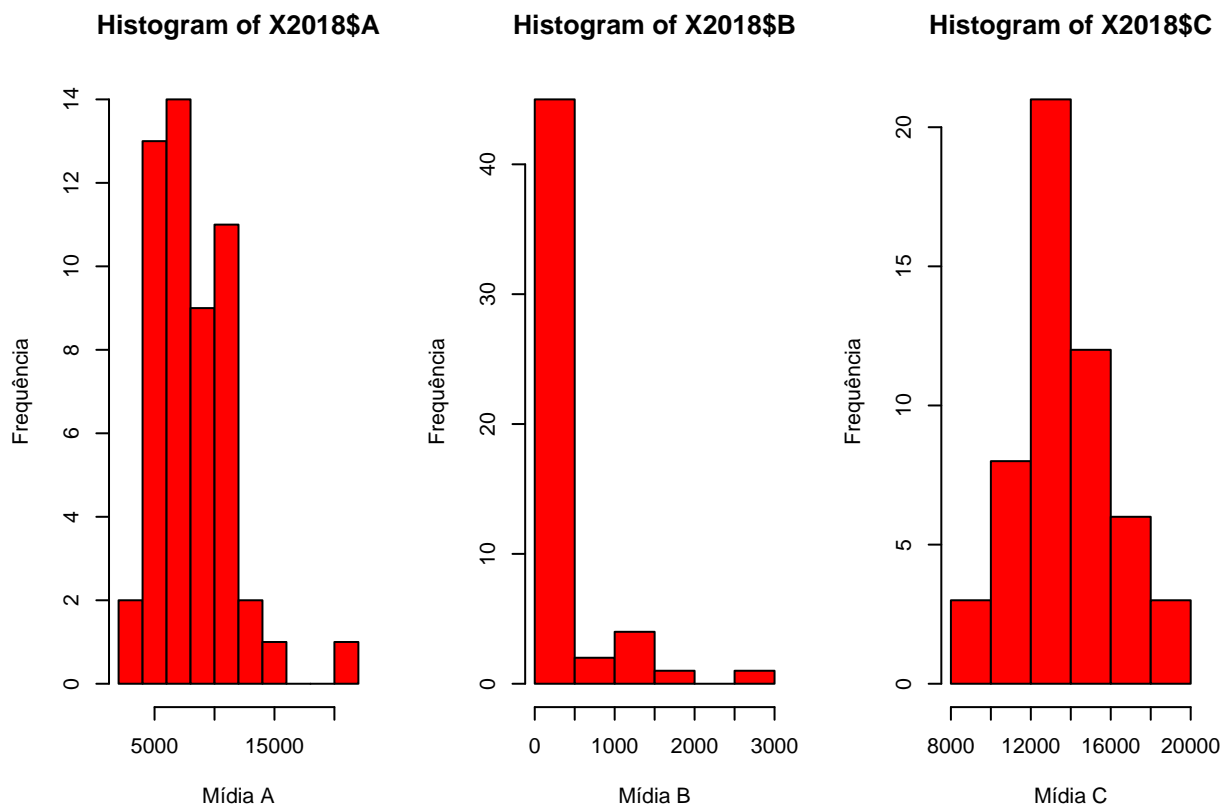
```
hist(x = X2018$A, xlab = "Mídia A", ylab = "Frequência", col = "red")
```

```
# Histograma da variável Mídia B em relação a receita gerada no site por semana
```

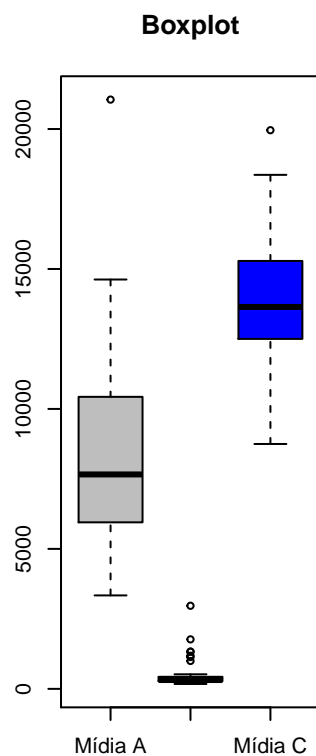
```
hist(x = X2018$B, xlab = "Mídia B", ylab = "Frequência", col = "red")
```

```
# Histograma da variável Mídia C em relação a receita gerada no site por semana
```

```
hist(x = X2018$C, xlab = "Mídia C", ylab = "Frequência", col = "red")
```



```
boxplot(X2018$A, X2018$B, X2018$C, names = c('Mídia A', 'Mídia B', 'Mídia C'),
        col = c('gray', 'red', 'blue'), main = 'Boxplot')
```



```
# Lendo os dados do ano 2019
```

```
X2019 <- read_excel("C:/Users/will_/Documents/roger/2019.xlsx", col_types = c("numeric",  
  "numeric", "numeric", "numeric"))  
head(X2019, 6)
```

```
## # A tibble: 6 x 4  
##       A       B       C Receita  
##   <dbl> <dbl> <dbl>   <dbl>  
## 1  8108.  8108.  8108. 502646.  
## 2 10425.   520. 16827. 532188.  
## 3 11767.   529. 16398. 532090.  
## 4 11933.   587. 10765. 487116.  
## 5  9554.   495. 18376. 468514.  
## 6  8524.   696. 21563. 540494.
```

```
library(psych)  
headTail(X2019)
```

```
##           A           B           C  Receita  
## 1  8107.8 8107.81  8107.82 502645.59  
## 2 10425.1  520.46 16827.48 532188.35  
## 3 11767.04 529.44  16398.2 532089.69  
## 4 11932.67 586.69 10764.78 487116.46  
## 5      ...      ...      ...      ...  
## 6 11687.42  574.12 16387.98 529545.26  
## 7 12222.74  375.24 14165.79 505568.21  
## 8 13777.74  342.49 12863.75 500124.99  
## 9  13257.5  425.33  11896.6  507259.8
```

```
str(X2019)
```

```
## tibble [52 x 4] (S3: tbl_df/tbl/data.frame)  
## $ A      : num [1:52] 8108 10425 11767 11933 9554 ...  
## $ B      : num [1:52] 8108 520 529 587 495 ...  
## $ C      : num [1:52] 8108 16827 16398 10765 18376 ...  
## $ Receita: num [1:52] 502646 532188 532090 487116 468514 ...
```

```
summary(X2019)
```

```
##           A           B           C           Receita  
## Min.      : 5882    Min.      : 241.6    Min.      : 1704    Min.      : 352021  
## 1st Qu.: 9506     1st Qu.: 474.5    1st Qu.:13995    1st Qu.: 431346  
## Median :10553     Median : 516.5    Median :15509    Median : 504107  
## Mean    :10801     Mean    : 677.0    Mean    :15799    Mean    : 529047  
## 3rd Qu.:11808     3rd Qu.: 593.1    3rd Qu.:18053    3rd Qu.: 562554  
## Max.    :21938     Max.    :8107.8    Max.    :31543    Max.    :1192700
```

```
par(mfrow = c(1,3))
```

Histograma da variável Média A em relação a receita gerada no site por semana em 201

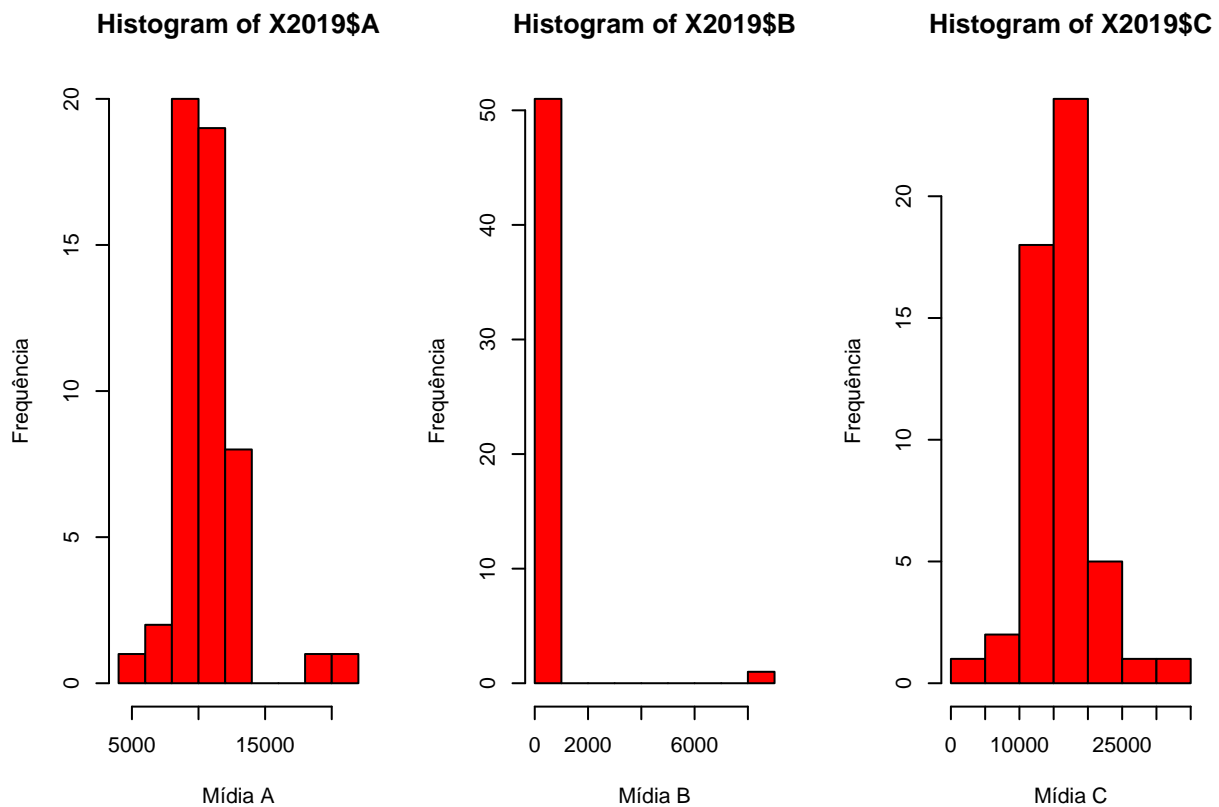
```
hist(x = X2019$A, xlab = "Média A", ylab = "Frequência", col = "red")
```

Histograma da variável Média B em relação a receita gerada no site por semana

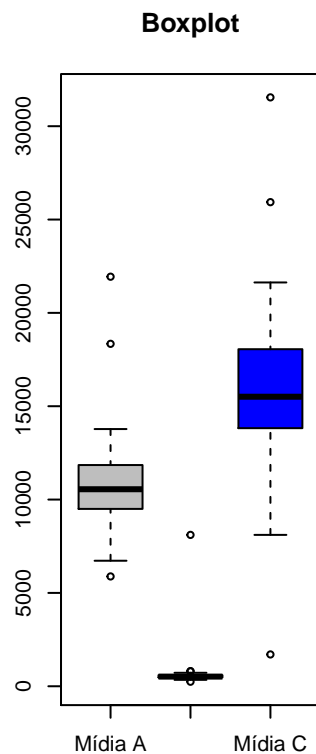
```
hist(x = X2019$B, xlab = "Média B", ylab = "Frequência", col = "red")
```

Histograma da variável Média C em relação a receita gerada no site por semana

```
hist(x = X2019$C, xlab = "Média C", ylab = "Frequência", col = "red")
```



```
boxplot(X2019$A, X2019$B, X2019$C, names = c('Média A', 'Média B', 'Média C'),
        col = c('gray', 'red', 'blue'), main = 'Boxplot')
```



Lendo os dados do ano 2020

```
X2020 <- read_excel("C:/Users/will/Documents/roger/2020.xlsx", col_types = c("numeric",
  "numeric", "numeric", "numeric"))
head(X2020, 6)
```

```
## # A tibble: 6 x 4
##       A      B      C Receita
##   <dbl> <dbl> <dbl>   <dbl>
## 1 11585.  368. 12499. 604771.
## 2 12554.  252. 14753. 597394.
## 3 11729   364. 18537. 556095.
## 4 13164   723. 18073. 601305.
## 5 13813.   724. 19037. 516837.
## 6 14697.   577. 16652. 541131.
```

```
library(psych)
headTail(X2020)
```

```
##           A      B      C  Receita
## 1 11584.61 368.45 12499.45 604770.99
## 2 12554.19 252.46 14752.89 597394.36
## 3   11729 363.94 18537.19 556094.99
## 4   13164 722.95 18072.51 601304.88
## 5      ...    ...    ...      ...
## 6 11703.84 478.72 18067.81  443026.8
## 7 11949.66 578.42 19751.82 516503.72
```



```
## 8 12172.7 625.48 19538.84 519165.64
## 9 11122.95 571.25 17593.92 531238.89
```

```
str(X2020)
```

```
## tibble [43 x 4] (S3: tbl_df/tbl/data.frame)
## $ A      : num [1:43] 11585 12554 11729 13164 13813 ...
## $ B      : num [1:43] 368 252 364 723 724 ...
## $ C      : num [1:43] 12499 14753 18537 18073 19037 ...
## $ Receita: num [1:43] 604771 597394 556095 601305 516837 ...
```

```
summary(X2020)
```

	A	B	C	Receita
## Min.	: 8592	Min. : 252.5	Min. :12499	Min. : 246277
## 1st Qu.:	11396	1st Qu.: 468.6	1st Qu.:17771	1st Qu.: 542628
## Median :	11949	Median : 578.4	Median :18677	Median : 604771
## Mean :	12056	Mean : 604.2	Mean :18584	Mean : 636880
## 3rd Qu.:	12574	3rd Qu.: 710.9	3rd Qu.:19740	3rd Qu.: 731395
## Max.	:15125	Max. :1124.2	Max. :23536	Max. :1170807

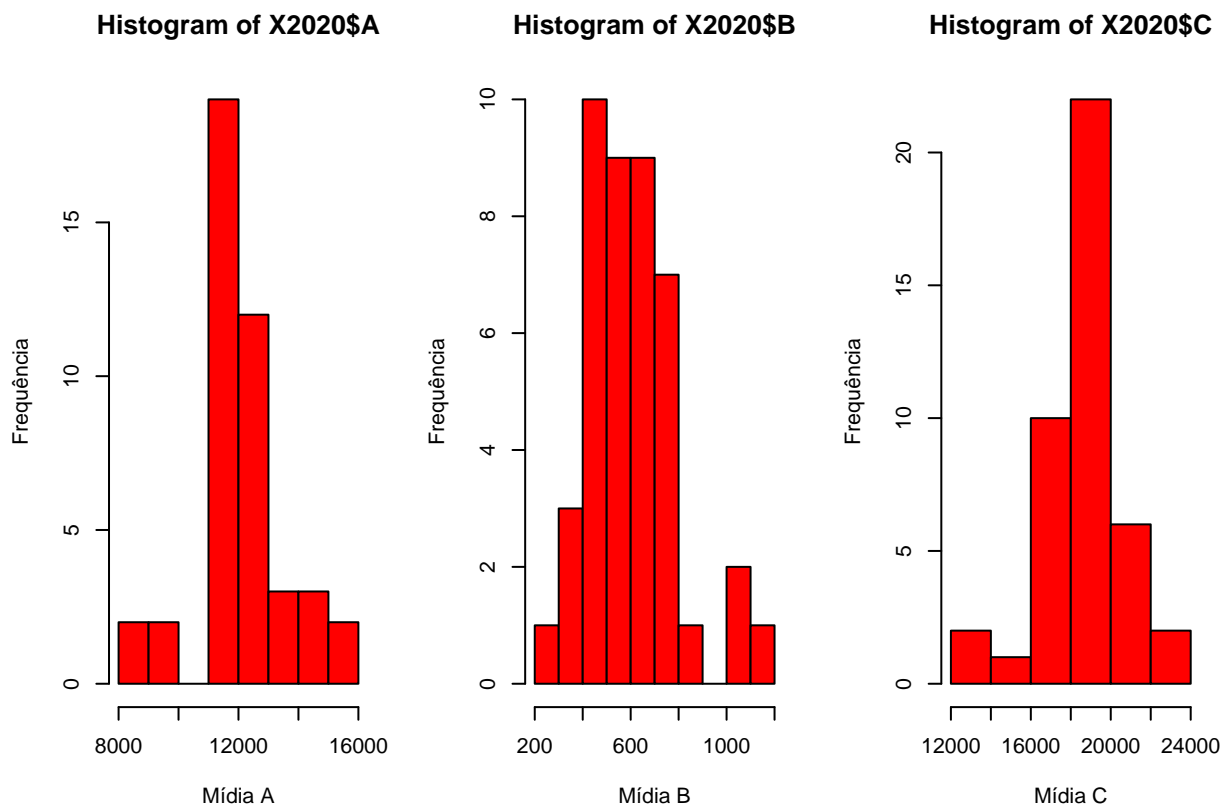
```
par(mfrow = c(1,3))
```

```
# Histograma da variável Mídia A em relação a receita gerada no site por semana em 201
```

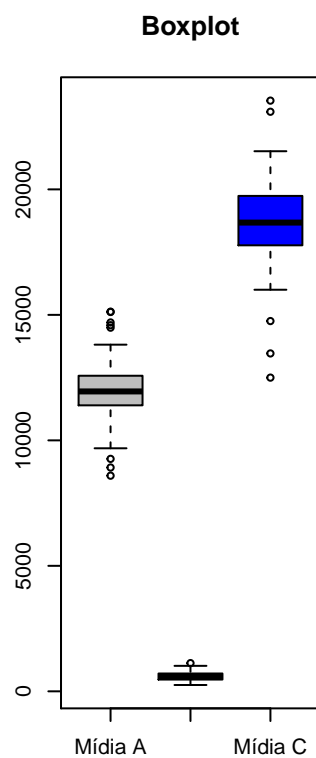
```
hist(x = X2020$A, xlab = "Mídia A", ylab = "Frequência", col = "red")
```

```
# Histograma da variável Mídia B em relação a receita gerada no site por semana
hist(x = X2020$B, xlab = "Mídia B", ylab = "Frequência", col = "red")
```

```
# Histograma da variável Mídia C em relação a receita gerada no site por semana
hist(x = X2020$C, xlab = "Mídia C", ylab = "Frequência", col = "red")
```



```
boxplot(X2020$A, X2020$B, X2020$C, names = c('Mídia A', 'Mídia B', 'Mídia C'),
        col = c('gray', 'red', 'blue'), main = 'Boxplot')
```



Criação do Modelo

```
modelo <- lm(CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B + CienciadeDados$C, data = CienciadeDados)
```

Resumo do modelo:

```
summary(modelo)
```

```
##
```

```
## Call:
```

```
## lm(formula = CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B +  
##     CienciadeDados$C, data = CienciadeDados)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max  
## -367969  -63796  -18133   46282   485200
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   88712.176  44806.754   1.980  0.0496 *  
## CienciadeDados$A    22.604    3.799   5.950 1.95e-08 ***  
## CienciadeDados$B     1.486    31.267   0.048  0.9622  
## CienciadeDados$C    14.300    3.006   4.757 4.72e-06 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 114800 on 144 degrees of freedom
```

```
## Multiple R-squared:  0.4614, Adjusted R-squared:  0.4502
```

```
## F-statistic: 41.13 on 3 and 144 DF,  p-value: < 2.2e-16
```

Método Stepwise

```
step(modelo, direction = "both")
```

```
## Start:  AIC=3452.67
```

```
## CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B +  
##     CienciadeDados$C
```

```
##
```

```
##              Df Sum of Sq      RSS      AIC  
## - CienciadeDados$B  1 2.9786e+07 1.8984e+12 3450.7  
## <none>                                1.8984e+12 3452.7  
## - CienciadeDados$C  1 2.9835e+11 2.1968e+12 3472.3  
## - CienciadeDados$A  1 4.6669e+11 2.3651e+12 3483.2
```

```
##
```

```
## Step:  AIC=3450.68
```

```
## CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C
```

```
##
```

```
##              Df Sum of Sq      RSS      AIC  
## <none>                                1.8984e+12 3450.7  
## + CienciadeDados$B  1 2.9786e+07 1.8984e+12 3452.7  
## - CienciadeDados$C  1 2.9835e+11 2.1968e+12 3470.3  
## - CienciadeDados$A  1 4.9085e+11 2.3893e+12 3482.7
```

```
##
## Call:
## lm(formula = CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C,
##     data = CienciadeDados)
##
## Coefficients:
##      (Intercept)  CienciadeDados$A  CienciadeDados$C
##           89109.84           22.64           14.30
```

```
# Método backward
step(modelo, direction = "backward")
```

```
## Start:  AIC=3452.67
## CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B +
##     CienciadeDados$C
```

```
##
##              Df Sum of Sq      RSS   AIC
## - CienciadeDados$B  1 2.9786e+07 1.8984e+12 3450.7
## <none>                                1.8984e+12 3452.7
## - CienciadeDados$C  1 2.9835e+11 2.1968e+12 3472.3
## - CienciadeDados$A  1 4.6669e+11 2.3651e+12 3483.2
```

```
##
## Step:  AIC=3450.68
## CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C
```

```
##
##              Df Sum of Sq      RSS   AIC
## <none>                                1.8984e+12 3450.7
## - CienciadeDados$C  1 2.9835e+11 2.1968e+12 3470.3
## - CienciadeDados$A  1 4.9085e+11 2.3893e+12 3482.7
```

```
##
## Call:
## lm(formula = CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C,
##     data = CienciadeDados)
##
## Coefficients:
##      (Intercept)  CienciadeDados$A  CienciadeDados$C
##           89109.84           22.64           14.30
```

```
# Método forward
step(modelo, direction = "forward")
```

```
## Start:  AIC=3452.67
## CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B +
##     CienciadeDados$C
```

```
##
## Call:
## lm(formula = CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$B +
```

```
##      CienciadeDados$C, data = CienciadeDados)
##
## Coefficients:
##      (Intercept)  CienciadeDados$A  CienciadeDados$B  CienciadeDados$C
##      88712.176      22.604      1.486      14.300
```

Criação do Modelo

```
modelo1 <- lm(CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C, data = Cien
```

Resumo do modelo:

```
summary(modelo1)
```

```
##
## Call:
## lm(formula = CienciadeDados$Receita ~ CienciadeDados$A + CienciadeDados$C,
##      data = CienciadeDados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -368061  -64046  -18030   46863   485470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    89109.839   43867.076    2.031   0.044 *
## CienciadeDados$A     22.642     3.698    6.123 8.18e-09 ***
## CienciadeDados$C     14.300     2.996    4.774 4.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114400 on 145 degrees of freedom
## Multiple R-squared:  0.4614, Adjusted R-squared:  0.454
## F-statistic: 62.12 on 2 and 145 DF,  p-value: < 2.2e-16
```

Avaliação do Modelo

1 - Significância

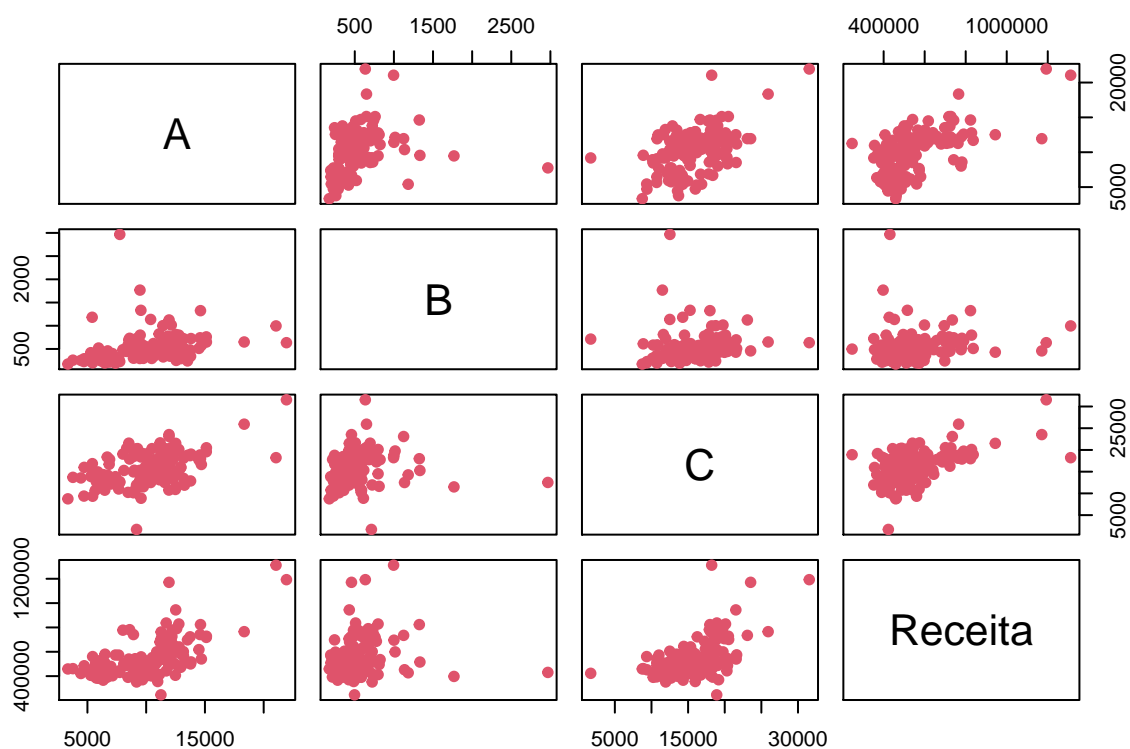
```
anova(modelo1)
```

```
## Analysis of Variance Table
##
## Response: CienciadeDados$Receita
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## CienciadeDados$A    1 1.3282e+12 1.3282e+12 101.448 < 2.2e-16 ***
## CienciadeDados$C    1 2.9835e+11 2.9835e+11  22.787 4.376e-06 ***
## Residuals       145 1.8984e+12 1.3093e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Linearidade
cor(CienciadeDados)
```

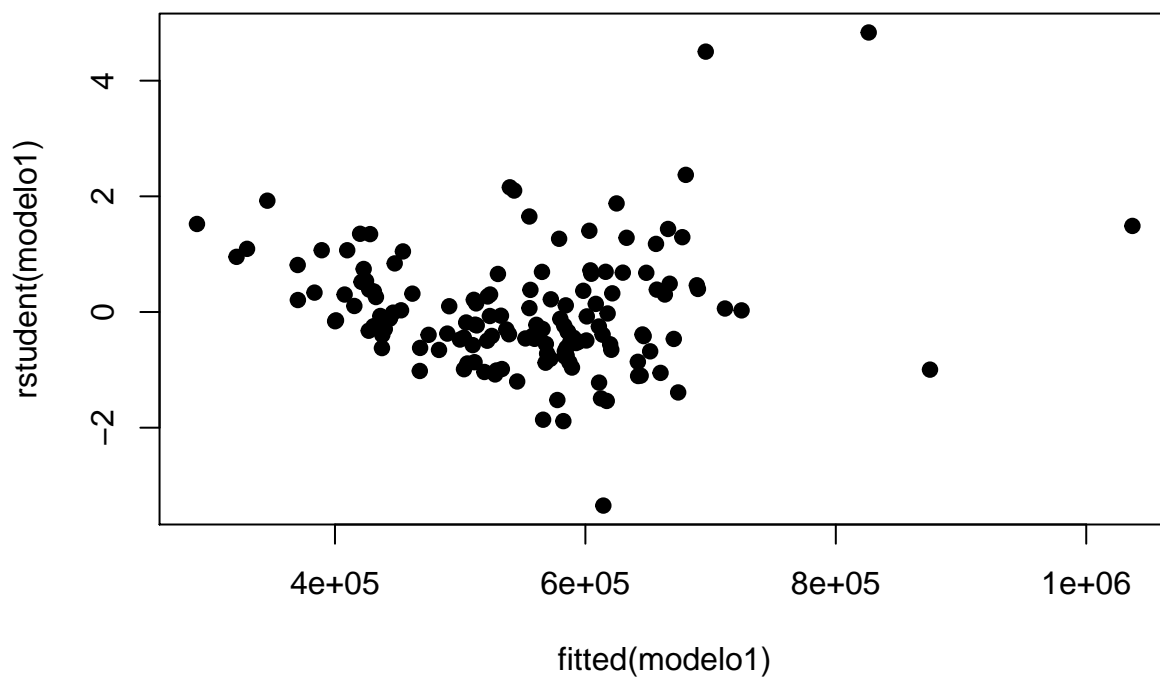
```
##           A           B           C  Receita
## A      1.0000000 0.2483670 0.5198268 0.6138411
## B      0.2483670 1.0000000 0.1284094 0.1550358
## C      0.5198268 0.1284094 1.0000000 0.5676200
## Receita 0.6138411 0.1550358 0.5676200 1.0000000
```

```
pairs(CienciadeDados, col = 2, pch = 19)
```



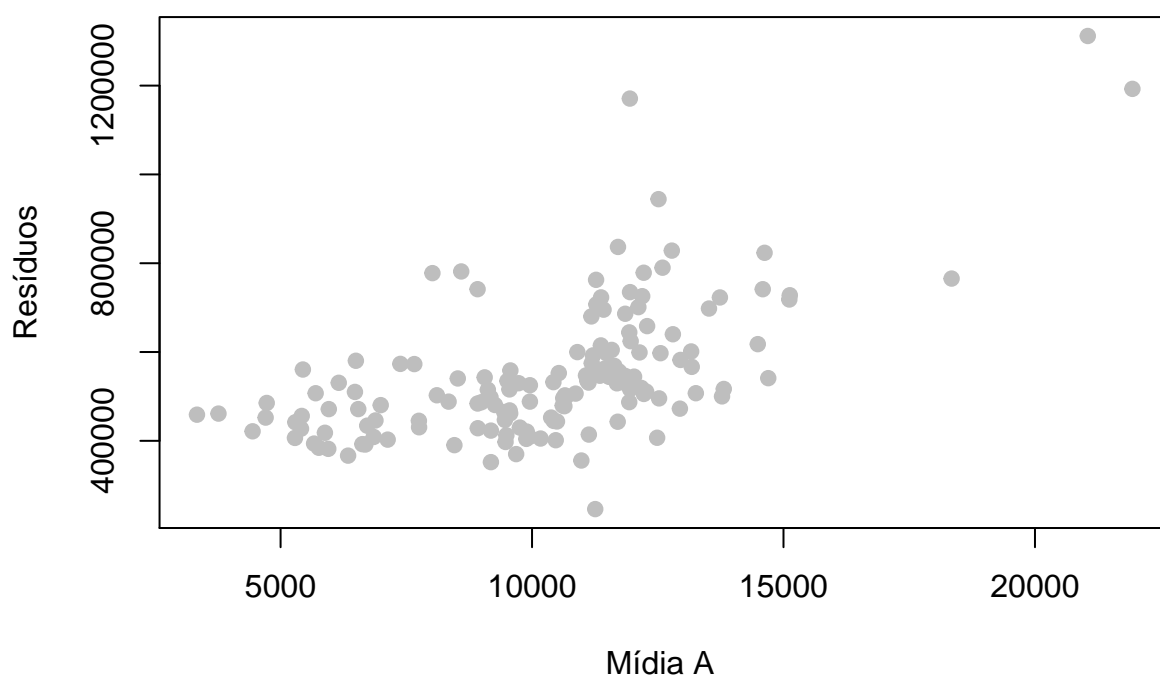
```
#Homocedasticidade dos resíduos
```

```
plot(rstudent(modelo1) ~ fitted(modelo1), pch = 19)
```

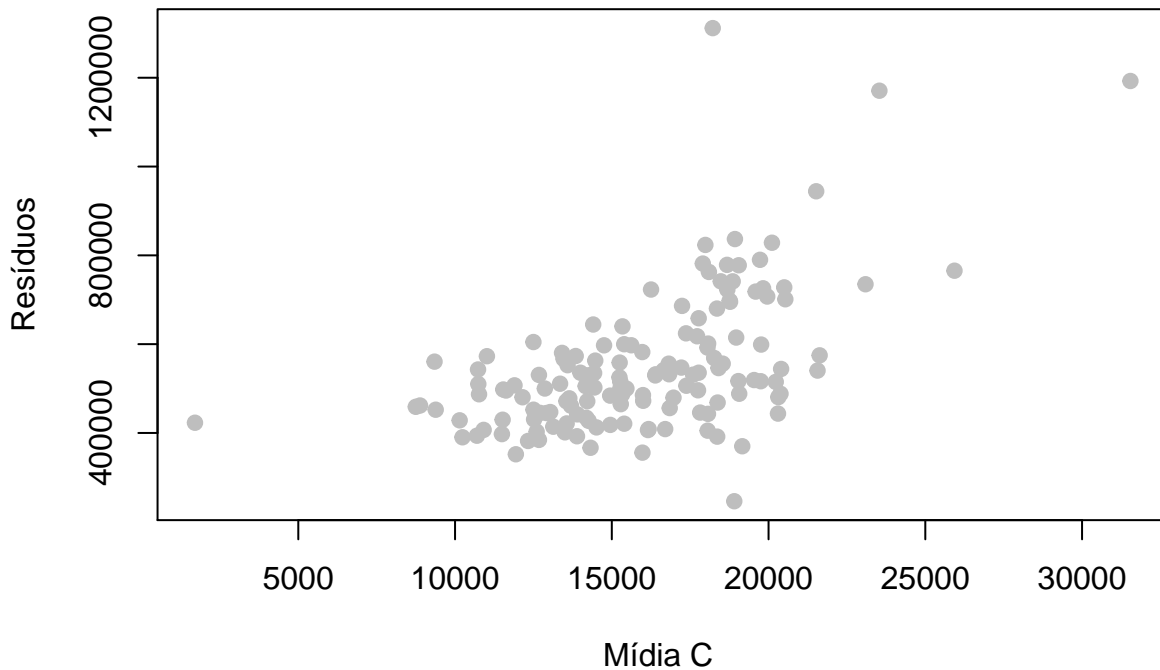


```
# abline(h = 0, lty = 2, col = "red")
```

```
plot(x = CienciadeDados$A, y = CienciadeDados$Receita, col = 'gray', pch = 19,
     xlab = 'Média A ', ylab = "Resíduos")
```



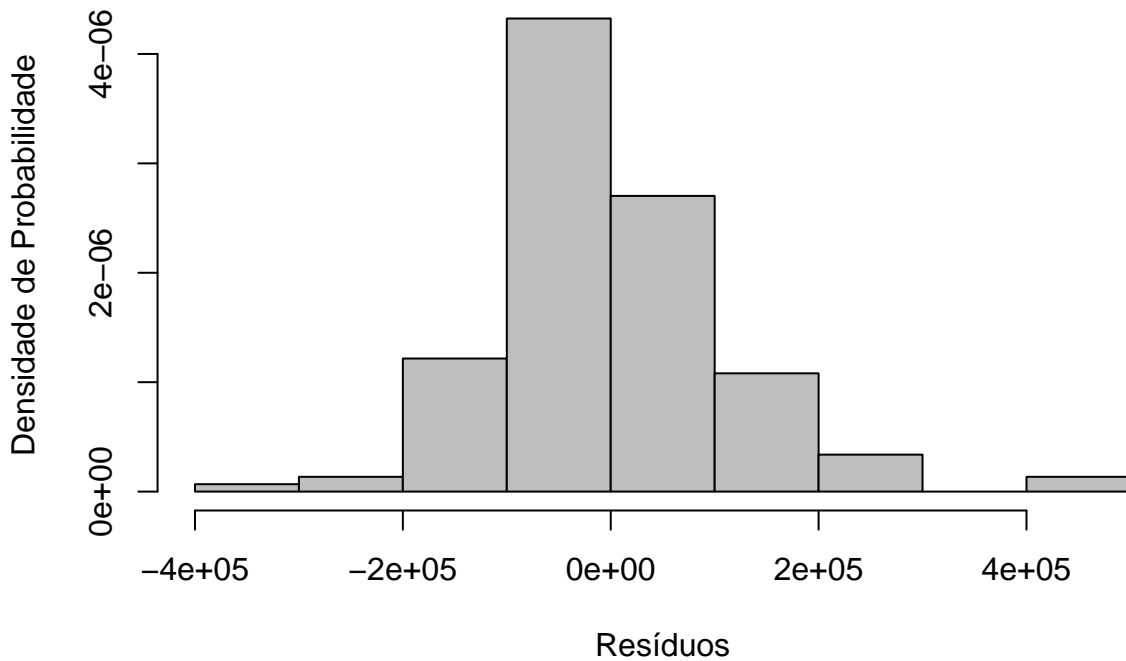
```
plot(x = CienciadeDados$C,y = CienciadeDados$Receita, col = 'gray',pch = 19,
     xlab = 'Média C ', ylab = "Resíduos")
```



```
# Normalidade dos Resíduos
```

```
hist(x = modelo1$residuals, col = 'gray', xlab = 'Resíduos', ylab = 'Densidade de Probab
     probability = TRUE)
```


Histogram of modelo1\$residuals



```
#lines(density(modelo1$residuals))
```

```
shapiro.test(modelo1$residuals)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  modelo1$residuals  
## W = 0.93175, p-value = 1.52e-06
```

Séries Temporais

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':  
##   method           from  
##   as.zoo.data.frame zoo
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##  
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':  
##  
##    as.Date, as.Date.numeric
```

```
library(nortest)  
CienciadeDadosR<- ts(CienciadeDados[,4])  
plot.ts(CienciadeDadosR,main="Receita")
```

