Trabalho de Machine Learning

Rodrigo de Miranda Videira

11/03/2022

Contexto:

A indústria XPTO fabrica cerveja artesanais e durante o ano ela abre sua fábrica para visitas. O gestor da fábrica pretende usar a ciência de dados para explicar a quantidade de turistas e prever quantos turistas/visitas terá no mês de janeiro/2021. Para isso, a indústria contratou uma consultoria para resolver o problema de negócio.

Bibliotecas utilizadas

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggcorrplot)
library(readxl)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

Carregando a base de dados para análise

```
df <- read_excel("cervejaria.xlsx")
view(df)
```

Análisando as variáveis presentes no dataset

```
names(df)
```

```
## [1] "visitas"   "excursoes" "preco"     "ano"       "trimestre" "data"
```

```
str(df)
```

```
## tibble [28 x 6] (S3: tbl_df/tbl/data.frame)
##  $ visitas  : num [1:28] 86947 134868 143617 102210 93407 ...
##  $ excursoes: num [1:28] 115 135 155 157 110 ...
##  $ preco    : num [1:28] 4.6 5.1 5.3 4.6 4.5 5.6 6.1 6.35 3.6 3.7 ...
##  $ ano      : num [1:28] 2014 2014 2014 2014 2015 ...
##  $ trimestre: num [1:28] 1 2 3 4 1 2 3 4 1 2 ...
##  $ data     : chr [1:28] "Q1 2014" "Q2 2014" "Q3 2014" "Q4 2014" ...
```

Tipos de variáveis

```
#visitas -> Quantitativa discreta
#excursoes -> Quantitativa contínua
#preco -> Quantitativa contínua
#ano -> Categórica ordinal
#trimestre -> Categórica ordinal
#data -> Categórica ordinal
```

Realizando a correção dos tipos categoricos

```
df$trimestre = as.factor(df$trimestre)
df$data = as.factor(df$data)
str(df)
```

```
## tibble [28 x 6] (S3: tbl_df/tbl/data.frame)
##  $ visitas  : num [1:28] 86947 134868 143617 102210 93407 ...
##  $ excursoes: num [1:28] 115 135 155 157 110 ...
##  $ preco    : num [1:28] 4.6 5.1 5.3 4.6 4.5 5.6 6.1 6.35 3.6 3.7 ...
##  $ ano      : num [1:28] 2014 2014 2014 2014 2015 ...
##  $ trimestre: Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
##  $ data     : Factor w/ 28 levels "Q1 2014","Q1 2015",..: 1 8 15 22 2 9 16 23 3 10 ...
```

Realizando análises estátisticas das variáveis:

- Visitas (Quantitativa)

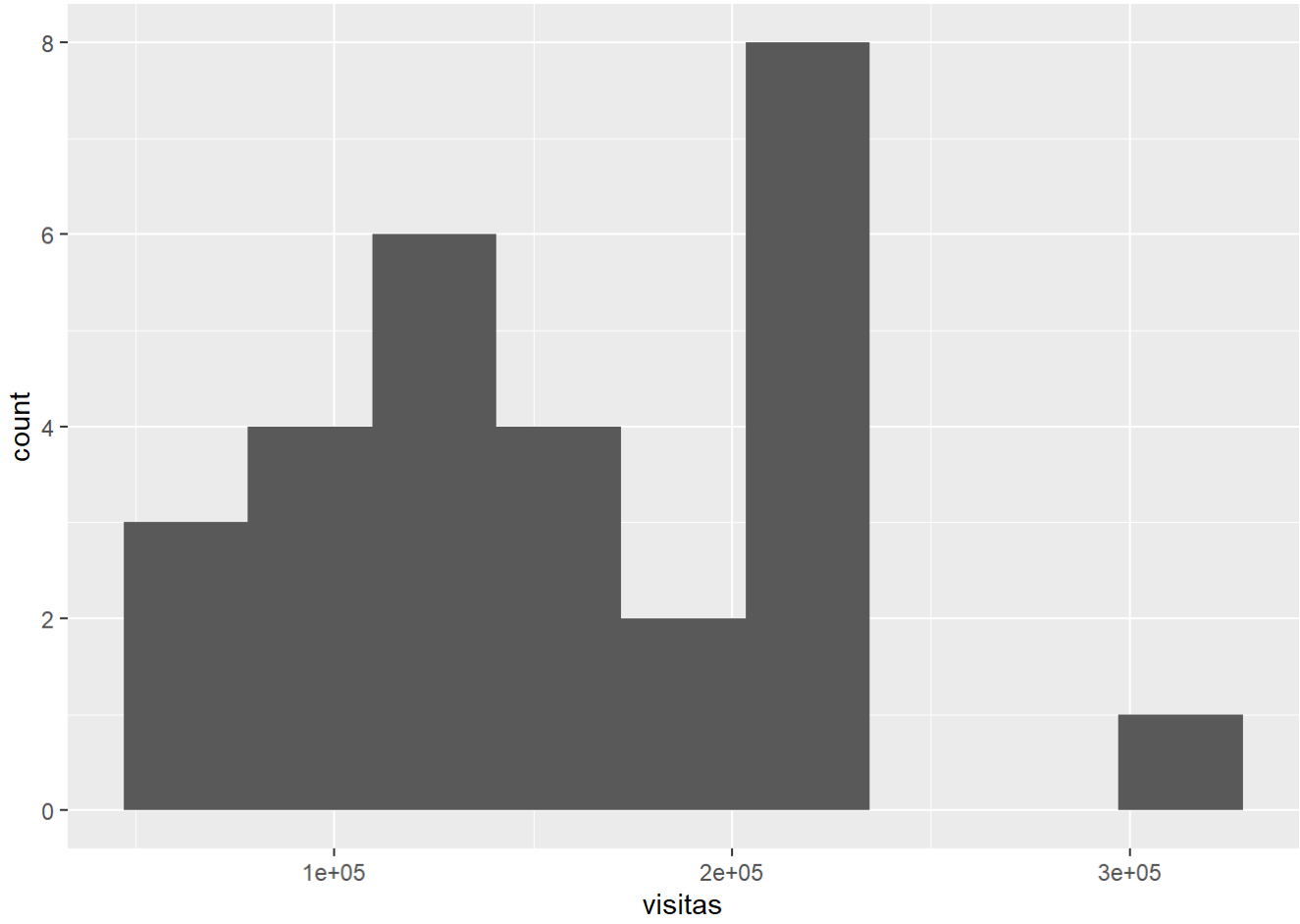
```
summary(df$visitas)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   59924  114479  147154  159206  219222  310199
```

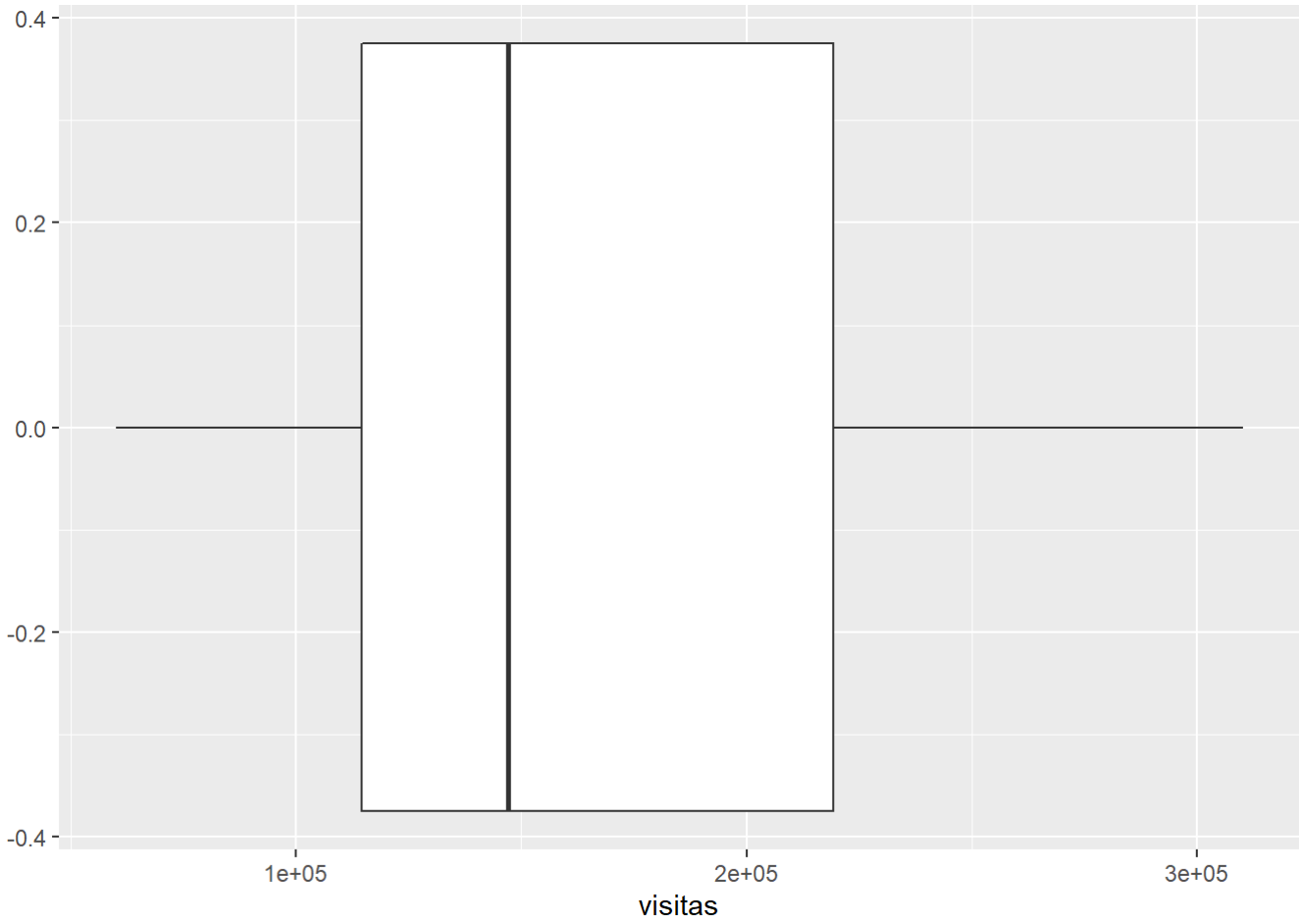
```
# Tirando os quartils Q1 e Q3 para análise de outliers
visitas_Q1 = quantile(df$visitas  , 0.25)
visitas_Q3 = quantile(df$visitas  , 0.75)
visitas_IQR = visitas_Q3 - visitas_Q1
```

Gráficos Visitas

```
ggplot(df, mapping = aes(x = `visitas`)) +
  geom_histogram(bins = 9)
```




```
ggplot(df, mapping = aes(x = `visitas`)) +
  geom_boxplot()
```



Pelos gráficos e valores apurados, a variável “Visitas” possui:

Média: 159206 Mediana: 147154

Como a média é maior que a mediana, e também pelo histograma os dados possuem assimetria a direita
Também pelo gráfico de boxplot não verificamos outliers.

- Excursoes (Quantitativa)

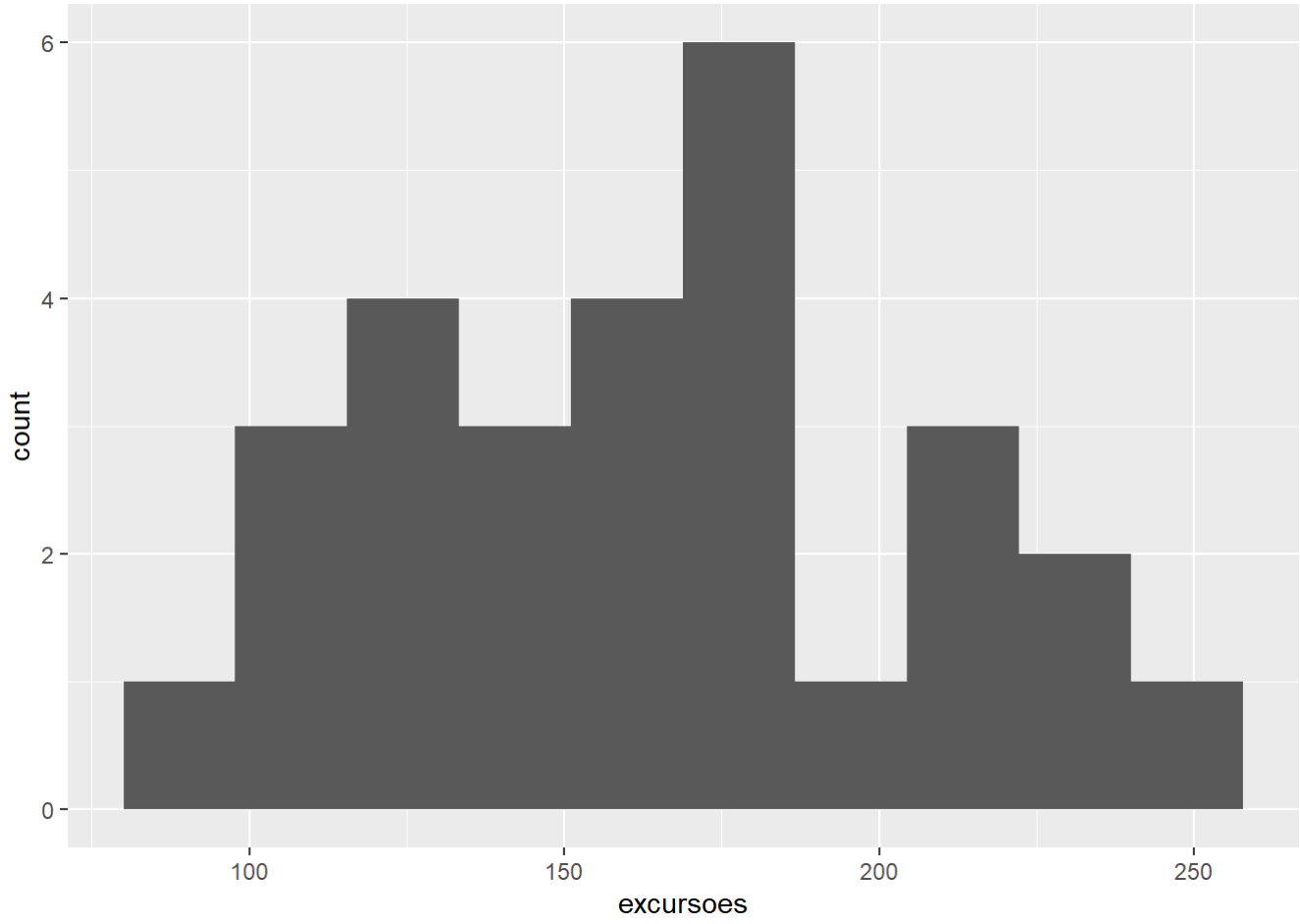
```
summary(df$excursoes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    85.0   130.6   167.0   164.2   187.5   245.0
```

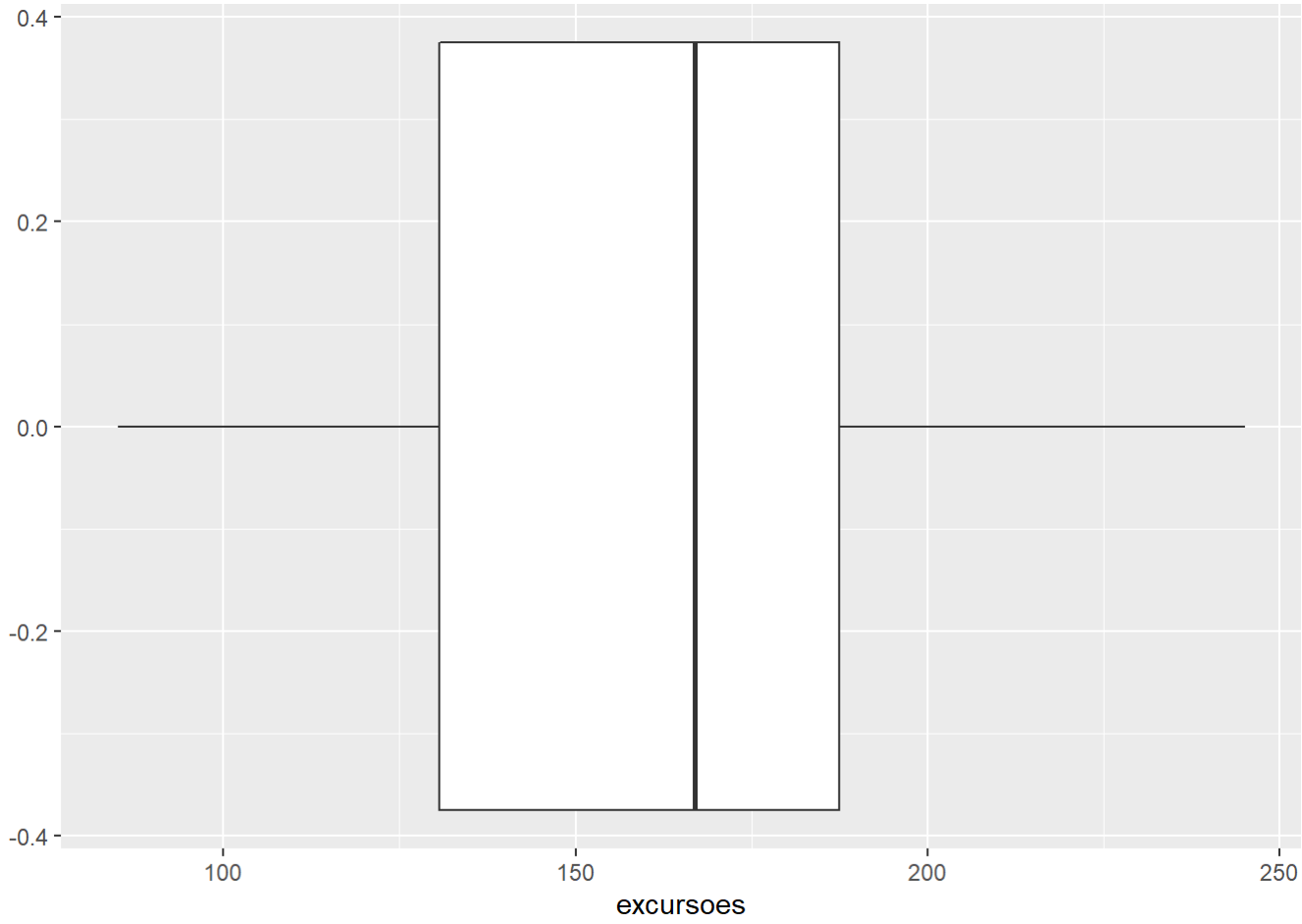
```
# Tirando os quartils Q1 e Q3 para análise de outliers
excursoes_Q1 = quantile(df$excursoes  , 0.25)
excursoes_Q3 = quantile(df$excursoes  , 0.75)
excursoes_IQR = excursoes_Q3 - excursoes_Q1
```

Gráficos excursões

```
ggplot(df, mapping = aes(x = `excursoes`)) +
  geom_histogram(bins = 10)
```




```
ggplot(df, mapping = aes(x = `excursoes`)) +
  geom_boxplot()
```



Pelos gráficos e valores apurados, a variável “Excursoes” possui:

Média: 164.2 Mediana: 167.0

Como a média é menor que a mediana, e também pelo histograma os dados possuem assimetria a esquerda
Também pelo gráfico de boxplot não verificamos outliers.

- Preço (Quantitativa)

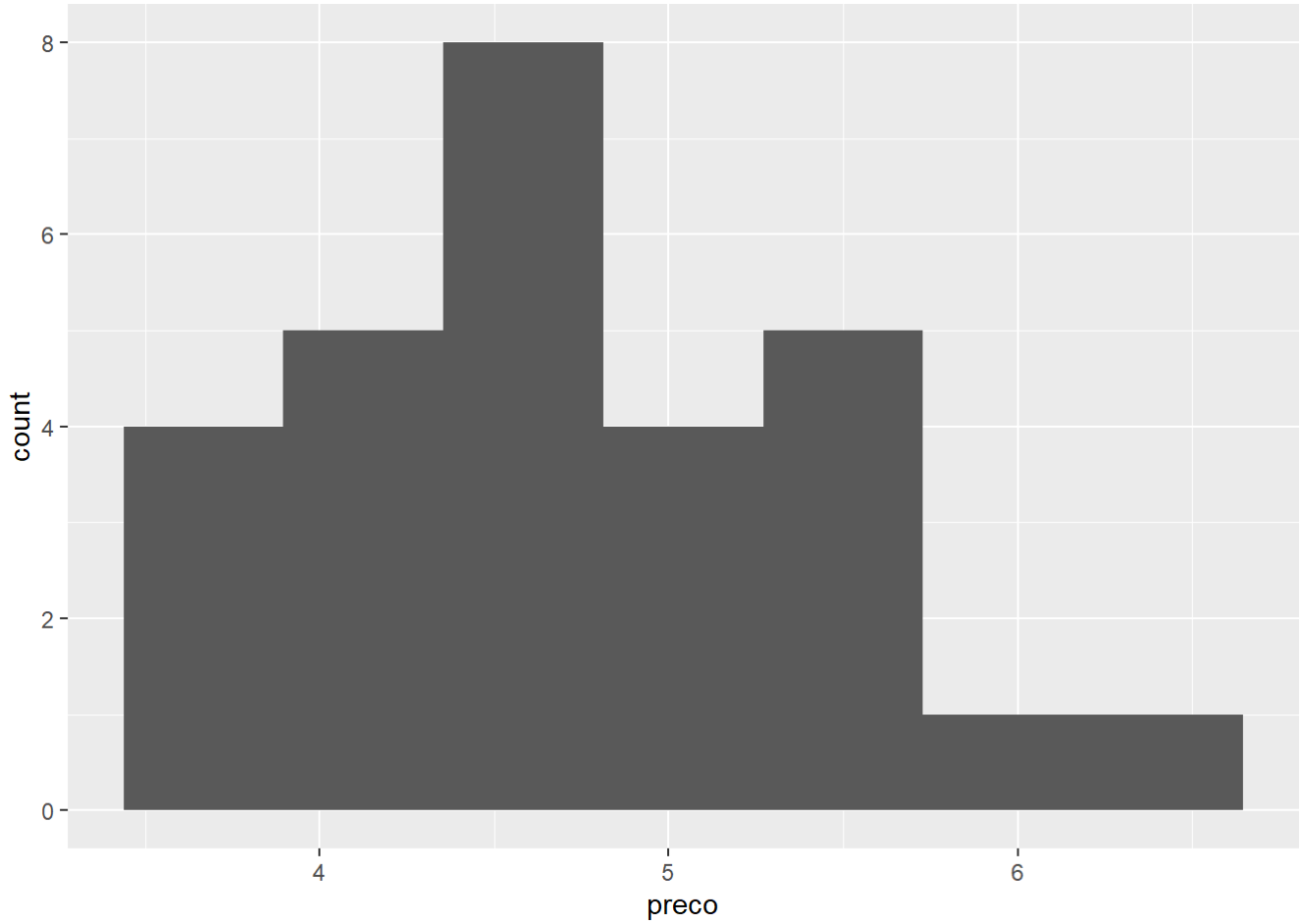
```
summary(df$preco)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.600   4.338   4.600   4.741   5.150   6.350
```

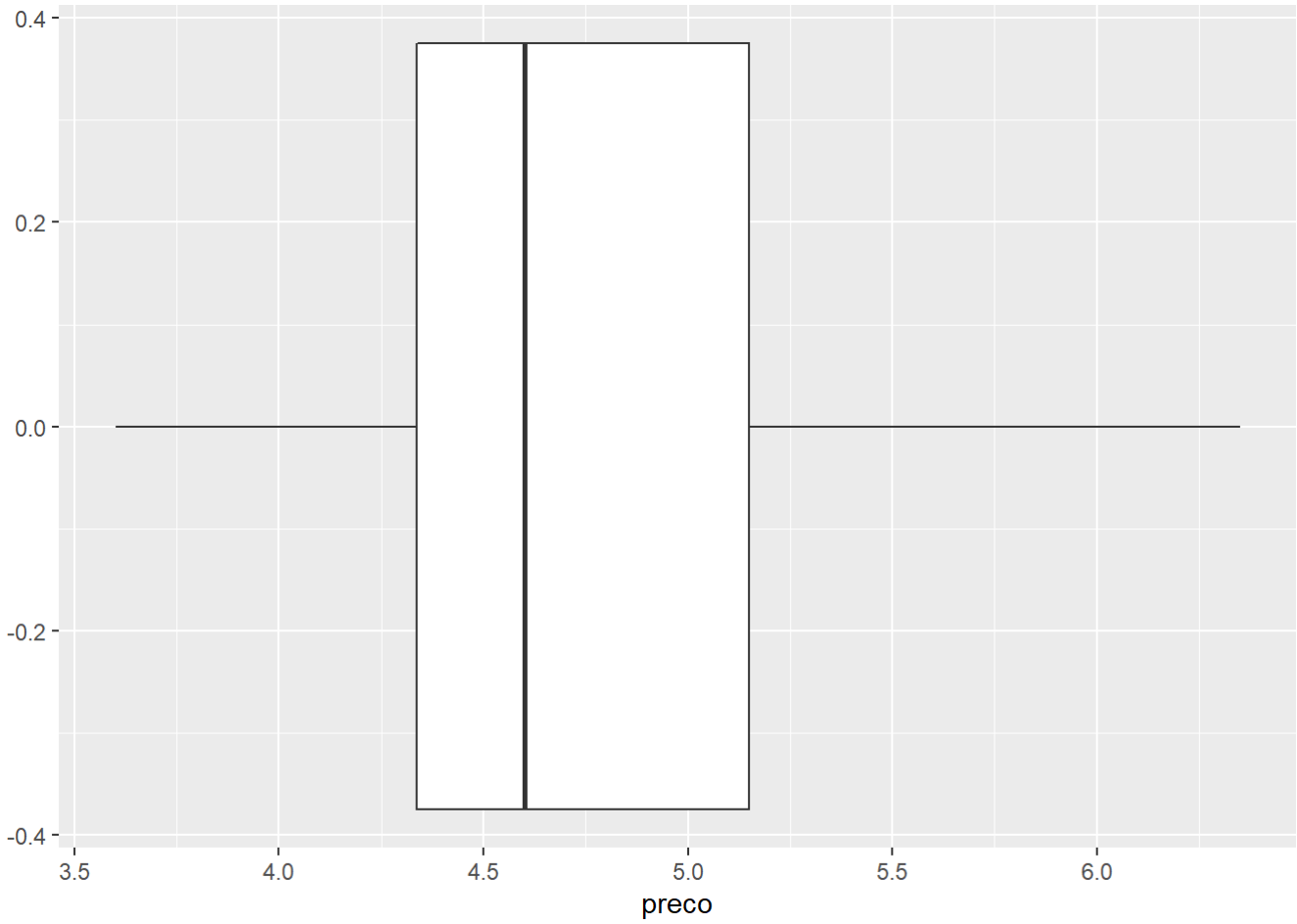
```
# Tirando os quartils Q1 e Q3 para análise de outliers
preco_Q1 = quantile(df$preco  , 0.25)
preco_Q3 = quantile(df$preco  , 0.75)
preco_IQR = preco_Q3 - preco_Q1
```

Gráficos excursões

```
ggplot(df, mapping = aes(x = `preco`)) +
  geom_histogram(bins = 7)
```




```
ggplot(df, mapping = aes(x = `preco`)) +
  geom_boxplot()
```



Pelos gráficos e valores apurados, a variável “Preço” possui:

Média: 4.741 Mediana: 4.600

Como a média é maior que a mediana, e também pelo histograma os dados possuem assimetria a direita
Também pelo gráfico de boxplot não verificamos outliers.

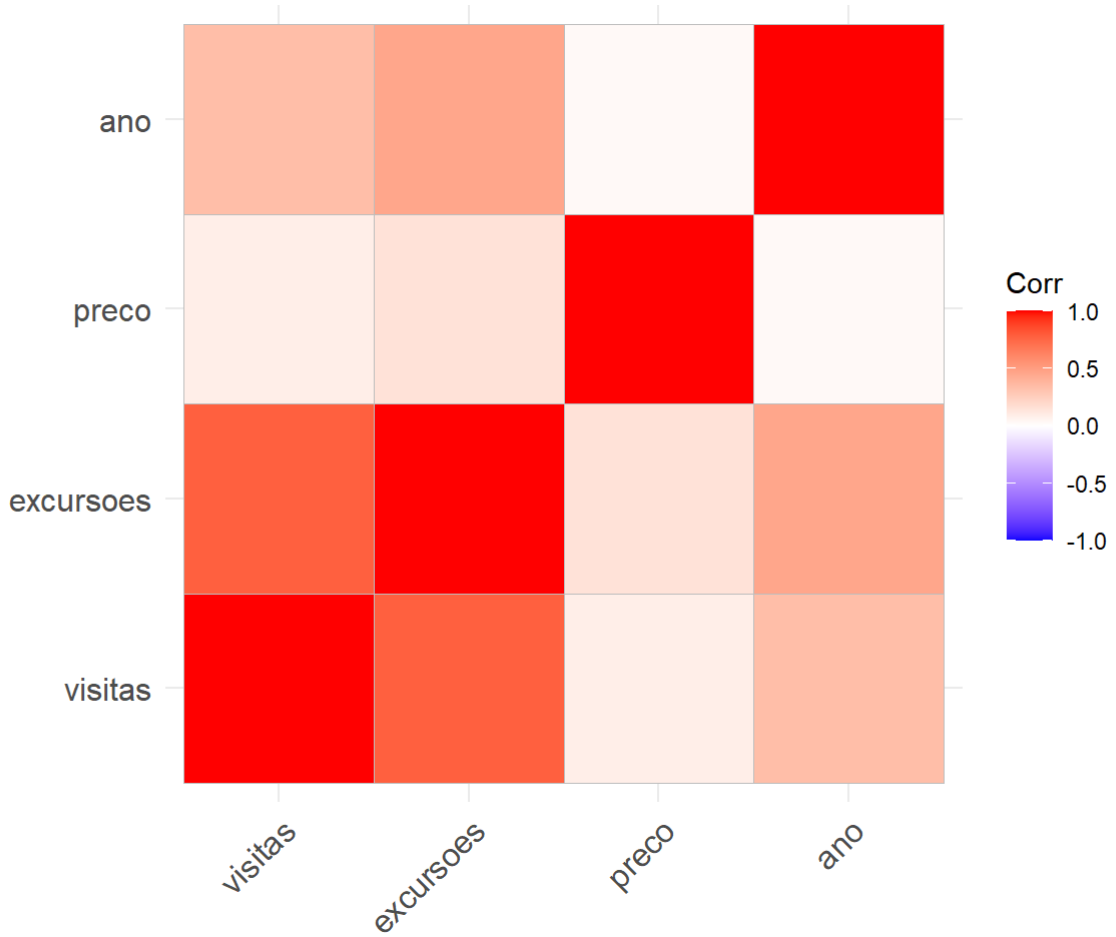
- Ano (Qualitativa)

```
ano_tabela <- table(df$ano);ano_tabela
```

```
##
## 2014 2015 2016 2017 2018 2019 2020
##    4    4    4    4    4    4    4
```

Realizando análise de correlações das variáveis quantitativas

```
df_numericos <- select_if(df, is.numeric)
correl <-cor(df_numericos)
ggcorrplot(correl)
```



Pelo gráfico e valores de correlações, temos que as variáveis Excursões e Visitas possuem uma correlação de média para quase fortemente correlacionada. Sendo que nossa variável Visitas é a target. Entre as variáveis independentes elas não possuem uma alta correlação sendo para nossa modelo mante-las.

```
cor(df_numericos)
```

```
##              visitas excursoes      preco        ano
## visitas   1.00000000 0.7822459 0.08829592 0.34171769
## excursoes 0.78224587 1.0000000 0.14890031 0.46199067
## preco     0.08829592 0.1489003 1.00000000 0.03380983
## ano       0.34171769 0.4619907 0.03380983 1.00000000
```

Fazendo transformações nas nossas variáveis.

Transformando ano para integer e criando as dummies da coluna de “trimestre” e descartando nossa variável qualitativa de “data”

```
df$ano = as.numeric(df$ano)
df$data = NULL
```

Criando dummies com a variável “trimestre”

```
df$trimestre_q1 <- ifelse(df$trimestre == 1, 1,0)
df$trimestre_q2 <- ifelse(df$trimestre == 2, 1,0)
df$trimestre_q3 <- ifelse(df$trimestre == 3, 1,0)
df$trimestre_q4 <- ifelse(df$trimestre == 4, 1,0)
df$trimestre <-  NULL

# Retirando um dos trimestres
df$trimestre_q1 <- NULL
```

Verificando o DataFrame final com as transformações

```
view(df)
str(df)
```

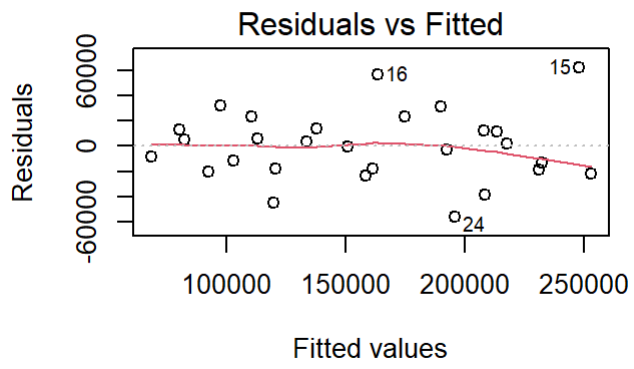
```
## tibble [28 x 7] (S3: tbl_df/tbl/data.frame)
##  $ visitas     : num [1:28] 86947 134868 143617 102210 93407 ...
##  $ excursoes   : num [1:28] 115 135 155 157 110 ...
##  $ preco       : num [1:28] 4.6 5.1 5.3 4.6 4.5 5.6 6.1 6.35 3.6 3.7 ...
##  $ ano         : num [1:28] 2014 2014 2014 2014 2015 ...
##  $ trimestre_q2: num [1:28] 0 1 0 0 0 1 0 0 0 1 ...
##  $ trimestre_q3: num [1:28] 0 0 1 0 0 0 1 0 0 0 ...
##  $ trimestre_q4: num [1:28] 0 0 0 1 0 0 0 1 0 0 ...
```

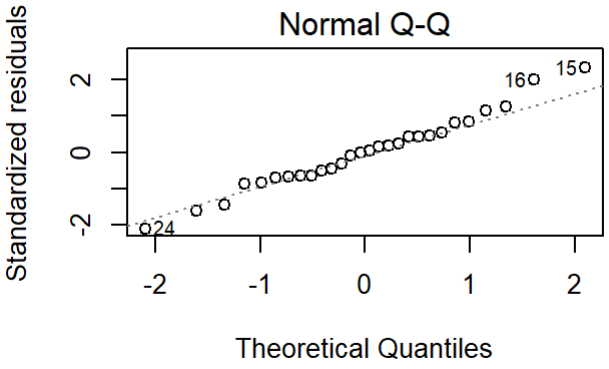
Criando o modelo de regressão linear

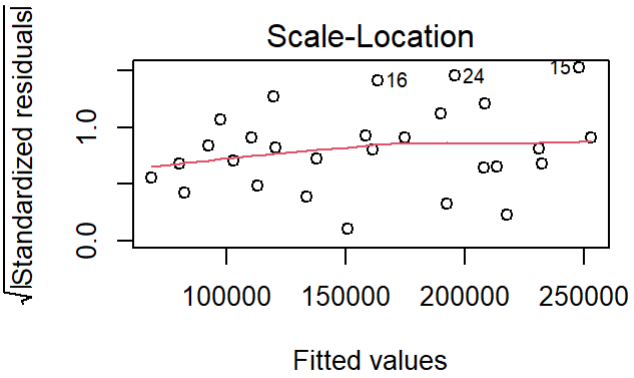
1º MODELO

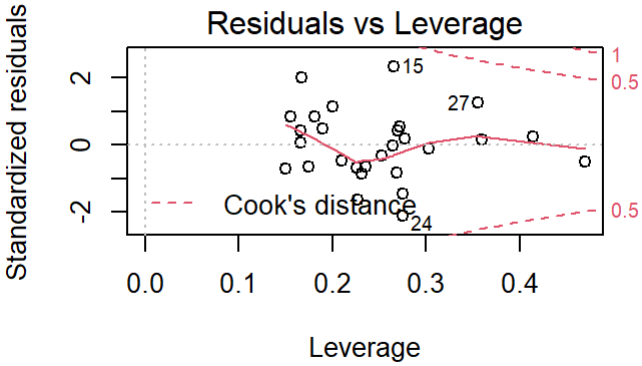
```
modelo_1 <- lm(visitas ~ ., data = df)
```

```
par(mfrow=c(2,2))
plot(modelo_1)
```









Testando a normalidade dos resíduos.

Ho: distribuição dos dados = normal -> p > 0.05 H1: distribuição dos dados <> normal -> p < 0.05

```
shapiro.test(modelo_1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_1$residuals
## W = 0.97961, p-value = 0.841
```

Escolhendo variáveis atráves do stepAIC - backward

```
mod.simples  <- lm(visitas ~ 1, data = df)
stepAIC(modelo_1, scope = list(upper = modelo_1,
                               lower = mod.simples, direction = "backward"))
```

```
## Start:  AIC=585.37
## visitas ~ excursoes + preco + ano + trimestre_q2 + trimestre_q3 +
##     trimestre_q4
##
##                Df  Sum of Sq        RSS    AIC
## - trimestre_q4  1 8.1020e+05 2.0387e+10 583.37
## - ano           1 1.7760e+08 2.0564e+10 583.61
## - preco         1 1.2715e+09 2.1658e+10 585.06
## <none>                       2.0387e+10 585.37
## - trimestre_q3  1 3.4455e+09 2.3832e+10 587.74
## - trimestre_q2  1 8.4179e+09 2.8804e+10 593.04
## - excursoes     1 1.3148e+10 3.3534e+10 597.30
##
## Step:  AIC=583.37
## visitas ~ excursoes + preco + ano + trimestre_q2 + trimestre_q3
##
##                Df  Sum of Sq        RSS    AIC
## - ano           1 2.3474e+08 2.0622e+10 581.69
## - preco         1 1.3931e+09 2.1780e+10 583.22
## <none>                       2.0387e+10 583.37
## + trimestre_q4  1 8.1020e+05 2.0387e+10 585.37
## - trimestre_q3  1 8.7078e+09 2.9095e+10 591.33
## - trimestre_q2  1 1.7520e+10 3.7907e+10 598.73
## - excursoes     1 2.4965e+10 4.5352e+10 603.75
##
## Step:  AIC=581.69
## visitas ~ excursoes + preco + trimestre_q2 + trimestre_q3
##
##                Df  Sum of Sq        RSS    AIC
## - preco         1 1.3775e+09 2.2000e+10 581.50
## <none>                       2.0622e+10 581.69
## + ano           1 2.3474e+08 2.0387e+10 583.37
## + trimestre_q4  1 5.7952e+07 2.0564e+10 583.61
## - trimestre_q3  1 8.5193e+09 2.9141e+10 589.37
## - trimestre_q2  1 1.7289e+10 3.7912e+10 596.74
## - excursoes     1 3.6995e+10 5.7617e+10 608.46
##
## Step:  AIC=581.5
## visitas ~ excursoes + trimestre_q2 + trimestre_q3
##
##                Df  Sum of Sq        RSS    AIC
## <none>                       2.2000e+10 581.50
## + preco         1 1.3775e+09 2.0622e+10 581.69
## + trimestre_q4  1 2.6674e+08 2.1733e+10 583.16
## + ano           1 2.1913e+08 2.1780e+10 583.22
## - trimestre_q3  1 7.4368e+09 2.9436e+10 587.65
## - trimestre_q2  1 1.6195e+10 3.8195e+10 594.95
## - excursoes     1 3.6422e+10 5.8422e+10 606.85
```

```
##
## Call:
## lm(formula = visitas ~ excursoes + trimestre_q2 + trimestre_q3,
##     data = df)
##
## Coefficients:
##  (Intercept)     excursoes  trimestre_q2  trimestre_q3
##     -24298.4         959.6       59770.2       43877.7
```

Criando o modelo com as variáveis selecionadas pelo metodo stepAIC usando método backward

```
modelo_2  <- lm(formula = visitas ~ excursoes + trimestre_q2 + trimestre_q3,
                  data = df)
```

Comparando os modelos

Modelo 1

```
summary(modelo_1)
```

```
##
## Call:
## lm(formula = visitas ~ ., data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56104 -18416    629  13360  62251
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3262115.5  7670942.9  -0.425  0.67498
## excursoes          927.8      252.1   3.680  0.00139 **
## preco            -10611.3     9272.1  -1.144  0.26532
## ano                1631.8     3815.2   0.428  0.67321
## trimestre_q2     62810.9    21330.3   2.945  0.00774 **
## trimestre_q3     49495.8    26272.6   1.884  0.07350 .
## trimestre_q4      -683.6    23663.7  -0.029  0.97723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31160 on 21 degrees of freedom
## Multiple R-squared:  0.8021, Adjusted R-squared:  0.7456
## F-statistic: 14.19 on 6 and 21 DF,  p-value: 1.978e-06
```

Modelo 2 - stepAIC backward

```
summary(modelo_2)
```

```
##
## Call:
## lm(formula = visitas ~ excursoes + trimestre_q2 + trimestre_q3,
##     data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61578 -18976  -2401  18026  65120
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -24298.4    24194.8  -1.004 0.325262
## excursoes       959.6      152.2   6.304 1.62e-06 ***
## trimestre_q2  59770.2    14219.7   4.203 0.000315 ***
## trimestre_q3  43877.7    15404.6   2.848 0.008872 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30280 on 24 degrees of freedom
## Multiple R-squared:  0.7864, Adjusted R-squared:  0.7598
## F-statistic: 29.46 on 3 and 24 DF,  p-value: 3.248e-08
```

Conclusões:

Em nosso modelo 1 utilizando todas as variáveis do nosso dataframe:

excursoes + preco + ano + trimestre_q2 + trimestre_q3 + trimestre_q4

chegamos a uma acurácia de 74,56

Já em nosso modelo 2 utilizando menos variáveis conseguimos chegar a uma acurácia maior, e utilizando as seguintes variáveis:

excursoes + trimestre_q2 + trimestre_q3

nossa acurácia deste modelo foi de 75,98

uma diferença de 1,42 para melhor, mas com um mínimo de variáveis.

Chegando ao nosso modelo final escolhido:

```
  y (visitas) =  -24298.4 + (excursoes) * 959.6 + (trimestre_q2) * 59770.2 + (trimestre_q3) *
43877.7
```

Predizendo um registro de nossa base de dados

```
linha_selecionada = df[1,]
linha_selecionada
```

visitas <dbl>	excursoes <dbl>	preco <dbl>	ano <dbl>	trimestre_q2 <dbl>	trimestre_q3 <dbl>	trimestre_q4 <dbl>
86947	115	4.6	2014	0	0	0

1 row

```
y <- -24298.4 + linha_selecionada$excursoes * 959.6 + linha_selecionada$trimestre_q2 * 59770.
2 + linha_selecionada$trimestre_q3 * 43877.7
y
```

```
## [1] 86055.6
```