Trabalho pratico Analise de dados

Sumário

Proposta:

A essência desse trabalho é selecionar, analisar e interpretar um conjunto de dados utilizando conceitos teóricos e práticos lecionados na classe de analise de dados (2023 - 2024) do Instituto politécnico de Bragança (IPB).

Conjunto de dados:

O conjunto de dados escolhido foi "combustiveis-regioes" o qual compara preços dos principais combustíveis automobilístico e de uso domestico entre as regiões do Brasil. Retirado do site "Kaggle".

Metodologia:

A manipulação dos dados presentes neste trabalho como: gráficos, estatísticas e textos interpretativos. São da autoria de: Dennis de Sousa Farias e Pedro Eunísio Vieira de Souza. E seguem os parâmetros ensinados em sala aplicados ao softwere "R".

INSTITUTO POLITÉCNICO DE BRAGANÇA

Docente:

Maria Prudência Gonçalves Martins

Alunos:

Dennis de Sousa Farias

Pedro Eunísio Vieira de Souza

Código geral:

Estatísticas Descritivas

```
> summary(dados_final$gasolina_comum_preco_revenda_avg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.490   2.506   2.792   3.244   3.877   7.540
> summary(dados_final$etanol_hidratado_preco_revenda_avg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8229  1.6493  2.0590  2.3458  2.8860  6.0100
> summary(dados_final$oleo_diesel_preco_revenda_avg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7926  1.8813  2.1660  2.6310  3.2212  7.6800
> summary(dados_final$gas_cozinha_glp_preco_revenda_avg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.69   32.85   40.35   49.16   62.97  123.44
```

Nas estatísticas descritivas temos a representação matemática dos conceitos de diferentes valores:

- **Mínimo**
- **Máximo**
- **Média**
- **Quartis**

Código geral:

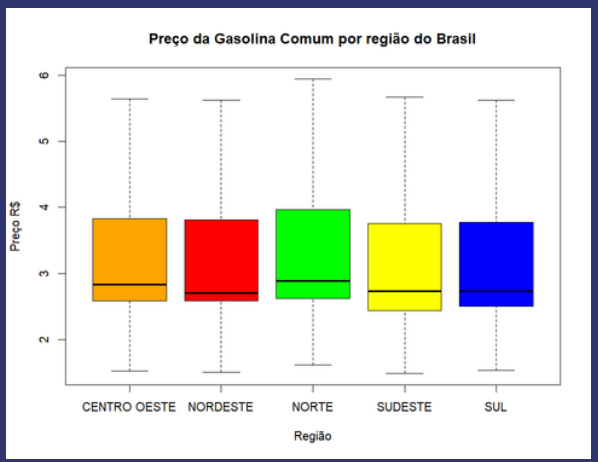
BoxPlot

```
## Boxplot ##

# Cores
cores_regioes = c('orange', 'red', 'green', 'yellow', 'blue')

boxplot(dados_final$gasolina_comum_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço da Gasolina Comum por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)

boxplot(dados_final$etanol_hidratado_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Etanol Hidratado por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)

boxplot(dados_final$oleo_diesel_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Óleo Diesel por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)

boxplot(dados_final$gas_cozinha_glp_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Gás de Cozinha por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)
```

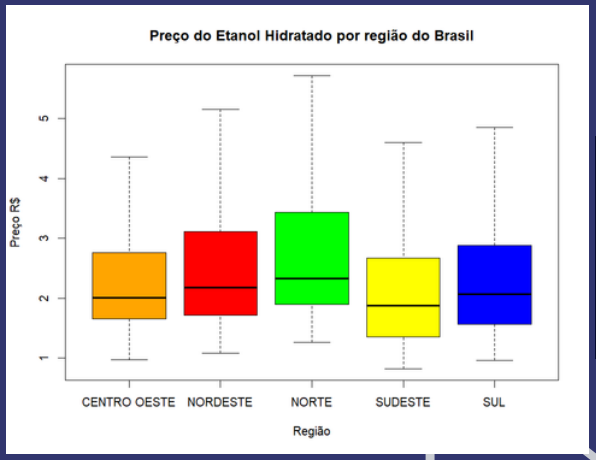
Estes são os códigos utilizados para gerar os gráficos “BoxPlot” (Caixa de Bigodes) abaixo.

O objetivos dos gráficos gerados neste capitulo é demonstrar o preço dos diferentes combustíveis domésticos entre regiões exibindo os valores já apresentados nas Estatísticas descritivas.

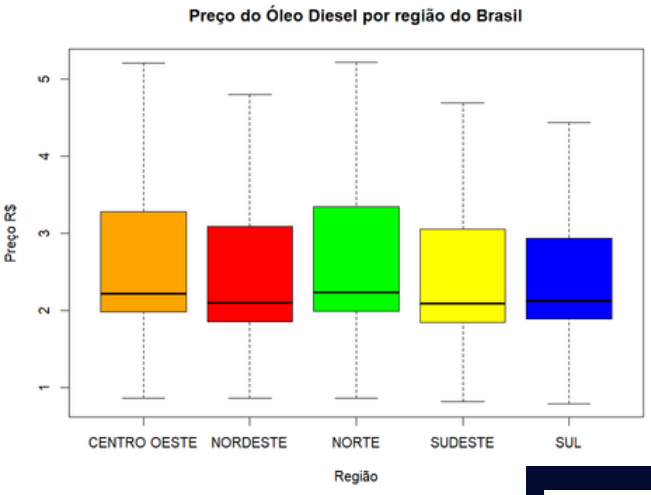
Perceba o padrão de cores atribuído as diferentes regiões do Brasil.



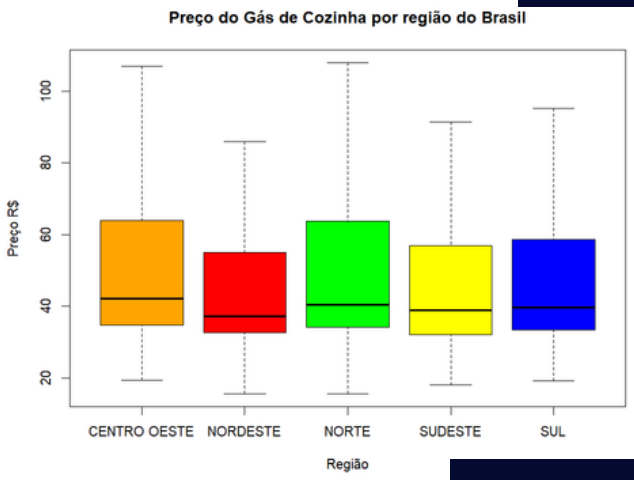

```
boxplot(dados_final$gasolina_comum_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço da Gasolina Comum por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)
```




```
boxplot(dados_final$etanol_hidratado_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Etanol Hidratado por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)
```




```
boxplot(dados_final$oleo_diesel_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Óleo Diesel por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)
```




```
boxplot(dados_final$gas_cozinha_glp_preco_revenda_avg ~ dados_final$regiao,
        outline = F,
        main = 'Preço do Gás de Cozinha por região do Brasil',
        ylab = 'Preço R$',
        xlab = 'Região',
        col = cores_regioes)
```

Interpreta-se que a região norte possui um valor geral dos combustíveis domésticos mais elevado enquanto a região sul em contrapartida possui um valor geral mais baixo. Além de possuir uma mediana geral com valores “similares”.

Código geral:

Histogramas

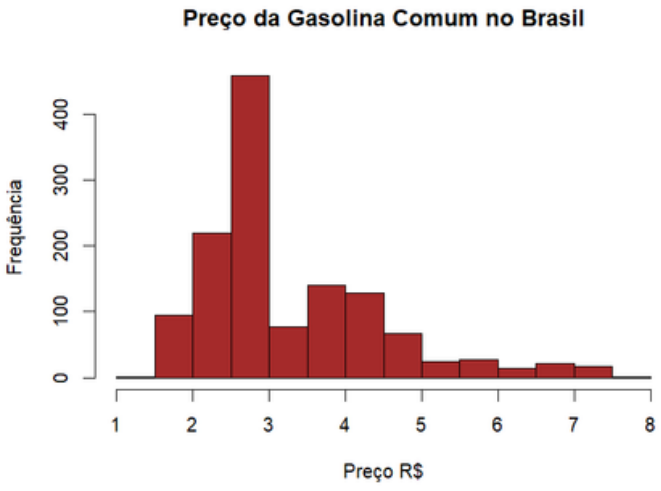
```
## Histogramas ##

hist(dados_final$gasolina_comum_preco_revenda_avg,
     main = 'Preço da Gasolina Comum no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$',
     col = 'brown')

hist(dados_final$etanol_hidratado_preco_revenda_avg,
     main = 'Preço do Etanol Hidratado no Brasil',
     ylab = 'Frequência',
       xlab = 'Preço R$',
     col = 'darkolivegreen3')

hist(dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Preço do Óleo Diesel no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$',
     col = 'darkblue')

hist(dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Preço do Gás de Cozinha no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$',
     col = 'cyan')
```

Estes são os códigos utilizados para gerar os gráficos “Histogramas”.

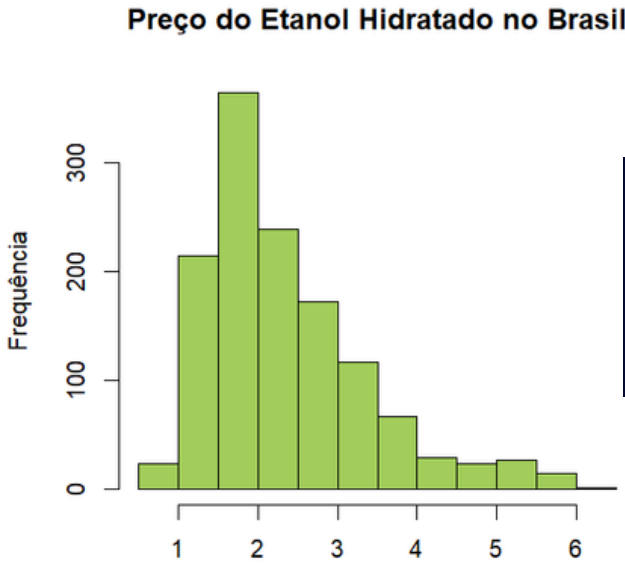
O objetivos dos gráficos gerados neste capitulo é demonstrar a Frequência dos preços dos diferentes combustíveis domésticos.

Observe o padrão de cores atribuído aos diferentes combustíveis.

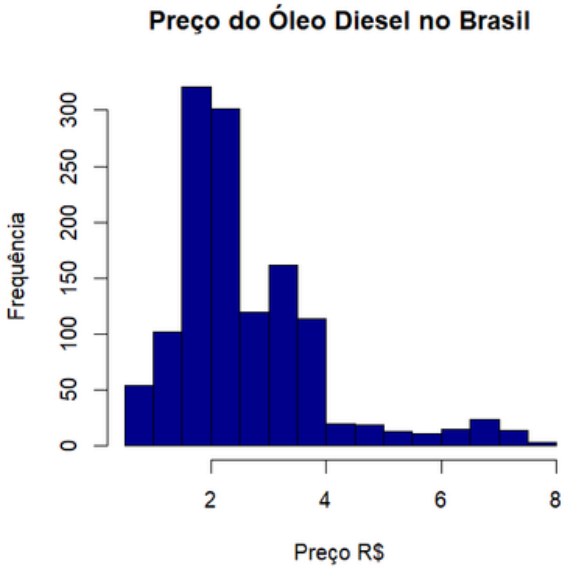
A terceira coluna representa o valor mais repetido dentro da serie histórica.



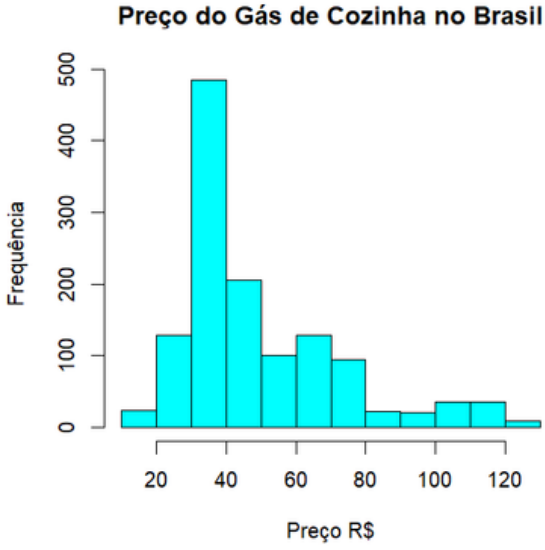

```
hist(dados_final$gasolina_comum_preco_revenda_avg,
     main = 'Preço da Gasolina Comum no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$',
     col = 'brown')
```




```
hist(dados_final$etanol_hidratado_preco_revenda_avg,
     main = 'Preço do Etanol Hidratado no Brasil',
     ylab = 'Frequência',
       xlab = 'Preço R$',
     col = 'darkolivegreen3')
```




```
hist(dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Preço do Óleo Diesel no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$?',
     col = 'darkblue')
```




```
hist(dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Preço do Gás de Cozinha no Brasil',
     ylab = 'Frequência',
     xlab = 'Preço R$',
     col = 'cyan')
```

Código geral:

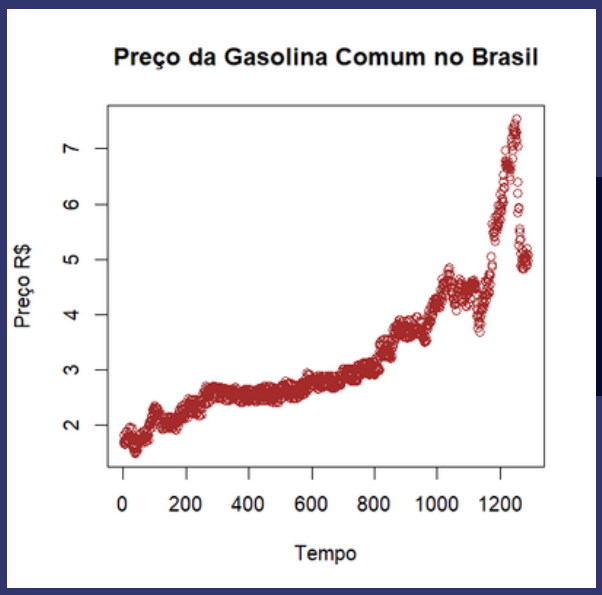
Plot Dispersão

```
## Plot ##
# Gráficos para mostrar a variação do preço ao longo do tempo

plot(dados_final$gasolina_comum_preco_revenda_avg,
     main = 'Preço da Gasolina Comum no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'brown')

plot(dados_final$etanol_hidratado_preco_revenda_avg,
     main = 'Preço do Etanol Hidratado no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'darkolivegreen3')

plot(dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Preço do Óleo Diesel no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'darkblue')

plot(dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Preço do Gás de Cozinha no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'cyan')
```

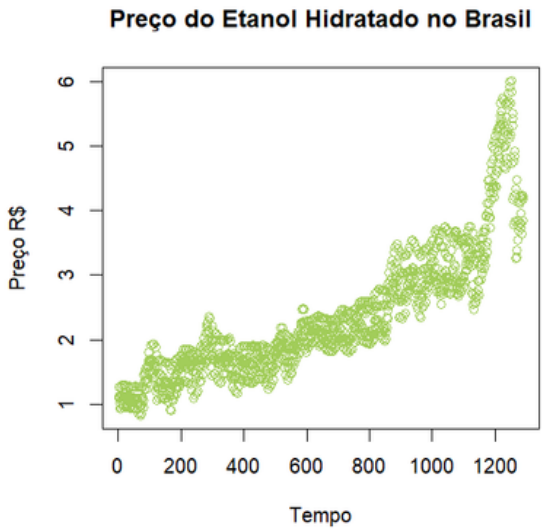
Estes são os códigos utilizados para gerar os gráficos “Plot” de Dispersão.

O objetivos dos gráficos gerados neste capitulo é demonstrar a variação dos preços dos diferentes combustíveis domésticos ao longo do tempo.

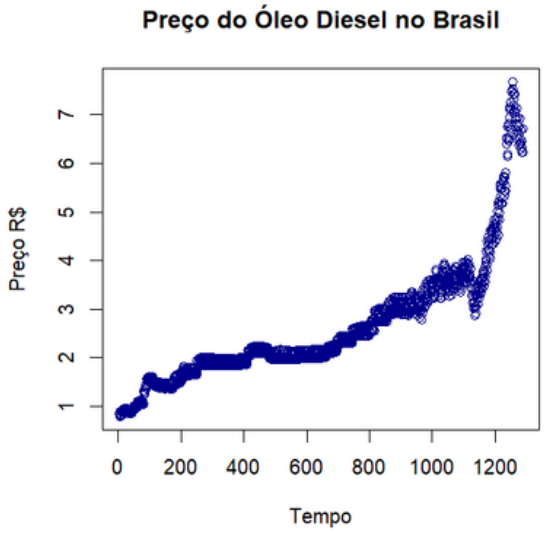
Mantém o padrão de cores atribuído aos diferentes combustíveis.



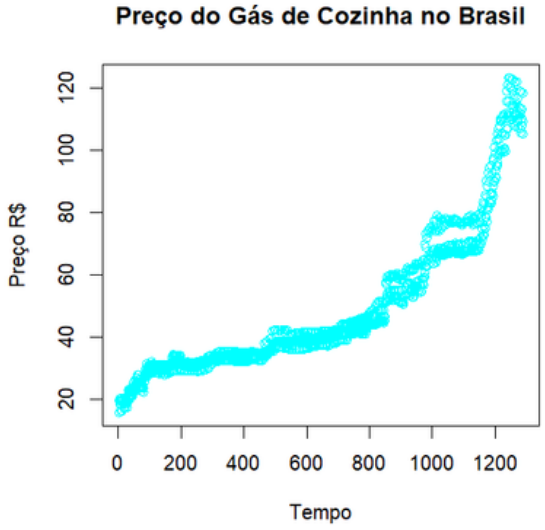

```
plot(dados_final$gasolina_comum_preco_revenda_avg,
     main = 'Preço da Gasolina Comum no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'brown')
```




```
plot(dados_final$etanol_hidratado_preco_revenda_avg,
     main = 'Preço do Etanol Hidratado no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'darkolivegreen3')
```




```
plot(dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Preço do Óleo Diesel no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'darkblue')
```

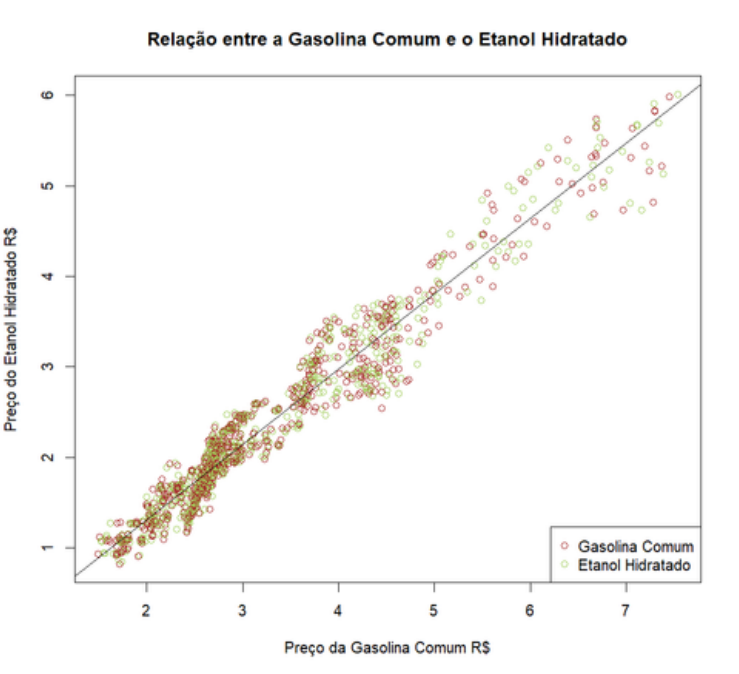



```
plot(dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Preço do Gás de Cozinha no Brasil',
     xlab = 'Tempo',
     ylab = 'Preço R$',
     col = 'cyan')
```

O preço de todos os itens são “igualmente” correlacionados com o tempo tendo em vista que eles crescem da ”mesma forma” ao longo do tempo.

Código Por Relação:

Plot dispersão

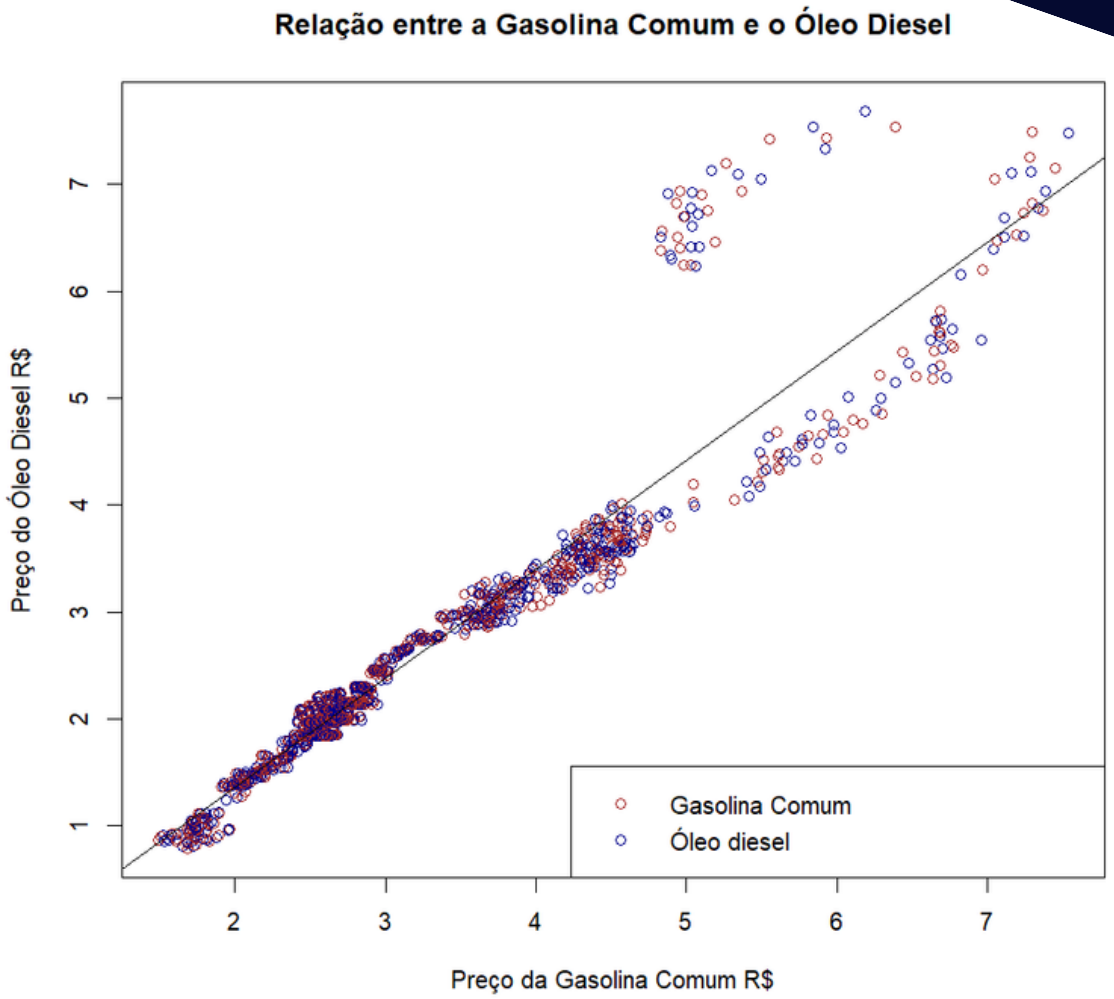


Estes são os códigos utilizados para gerar os gráficos “Plot” de Dispersão relacionando duas variáveis.

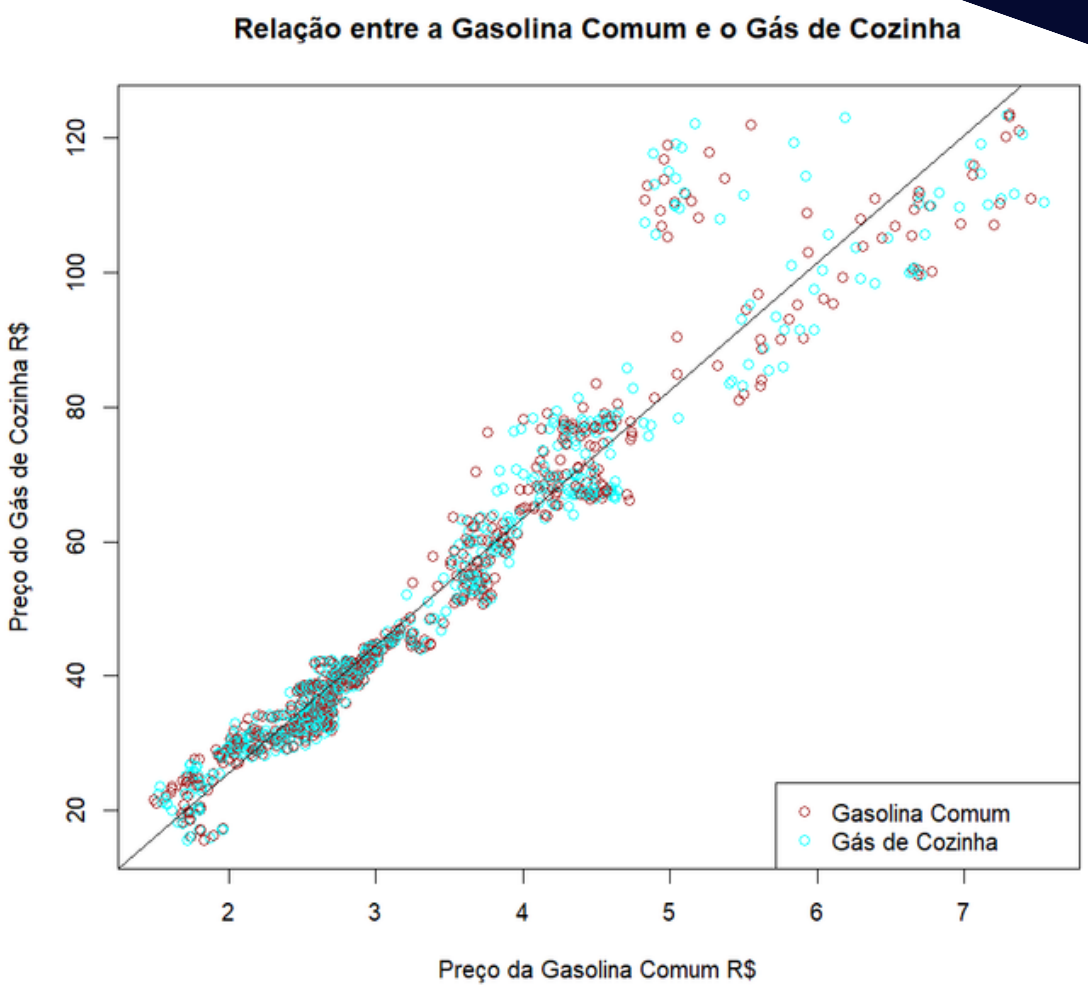
O objetivos dos gráficos gerados neste capitulo é demonstrar a Relação dos preços de dois diferentes combustíveis domésticos ao longo do tempo.

Mantém o padrão de cores atribuído aos diferentes combustíveis.

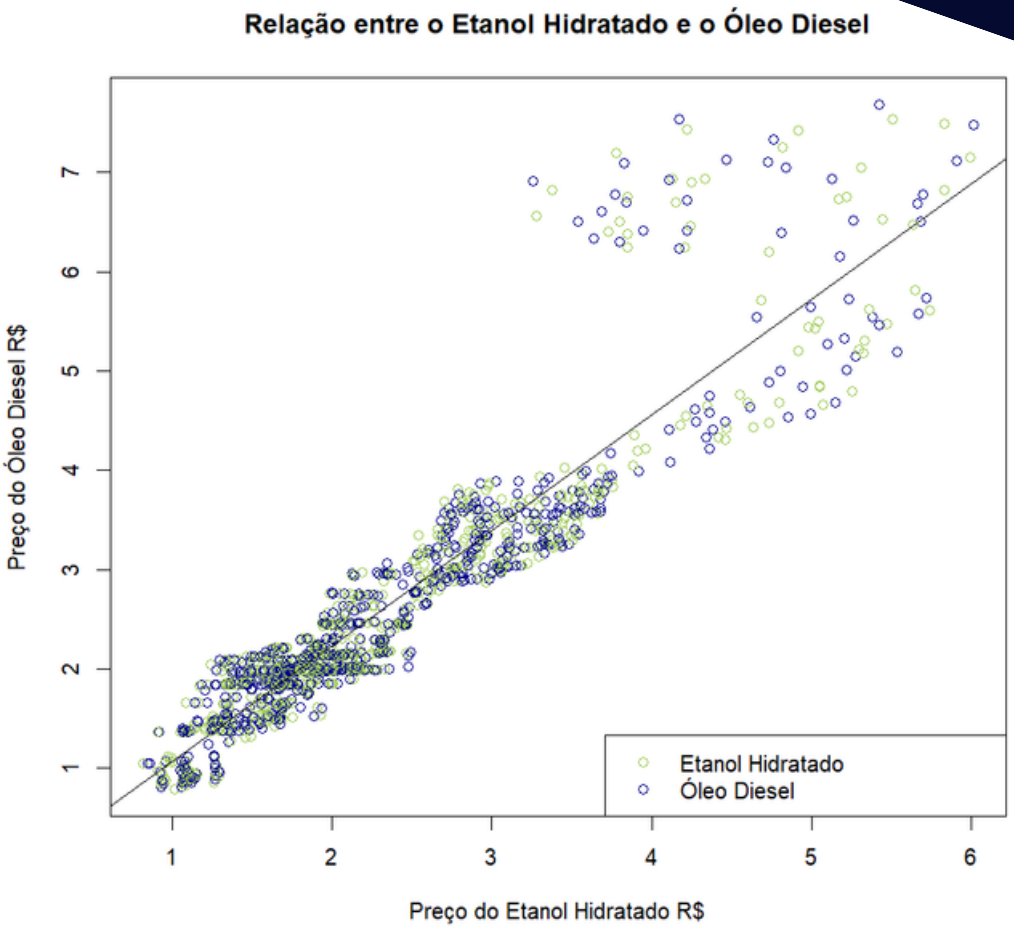
```
plot(dados_final$gasolina_comum_preco_revenda_avg, dados_final$etanol_hidratado_preco_revenda_avg,
     main = 'Relação entre a Gasolina Comum e o Etanol Hidratado',
     col = cores_variaveis[c(1,2)],
     xlab = 'Preço da Gasolina Comum R$',
     ylab = 'Preço do Etanol Hidratado R$'
)
legend(
  'bottomright',
  legend = c('Gasolina Comum', 'Etanol Hidratado'),
  pch = 1,
  col = cores_variaveis[c(1,2)]
)
abline(reg = lm(dados_final$etanol_hidratado_preco_revenda_avg ~ dados_final$gasolina_comum_preco_revenda_avg,
               lty = 'dashed'))
```



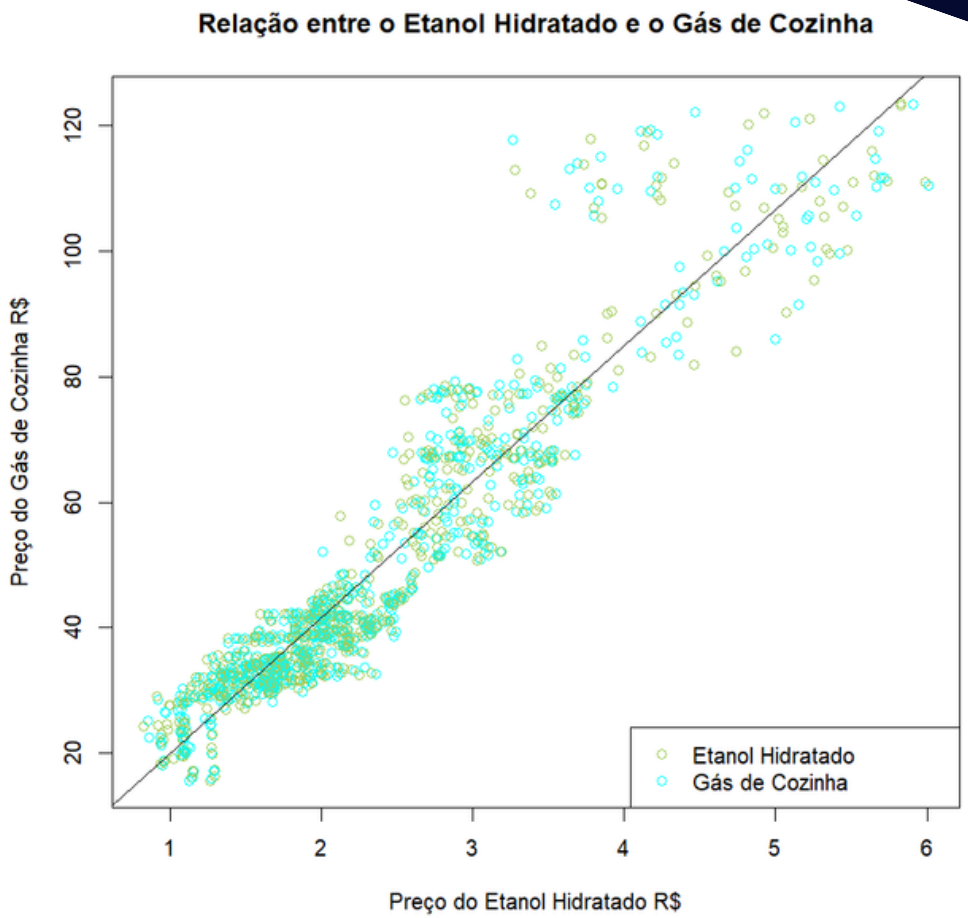

```r
# Relação entre gasolina comum e óleo diesel

plot(dados_final$gasolina_comum_preco_revenda_avg, dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Relação entre a Gasolina Comum e o Óleo Diesel',
     col = cores_variaveis[c(1,3)],
     xlab = 'Preço da Gasolina Comum R$',
     ylab = 'Preço do Óleo Diesel R$')
legend(
  'bottomright',
  legend = c('Gasolina Comum', 'Óleo diesel'),
  pch = 1,
  col = cores_variaveis[c(1,3)]
)
abline(reg = lm(dados_final$oleo_diesel_preco_revenda_avg ~ dados_final$gasolina_comum_preco_revenda_avg,
                lty = 'dashed'))
```



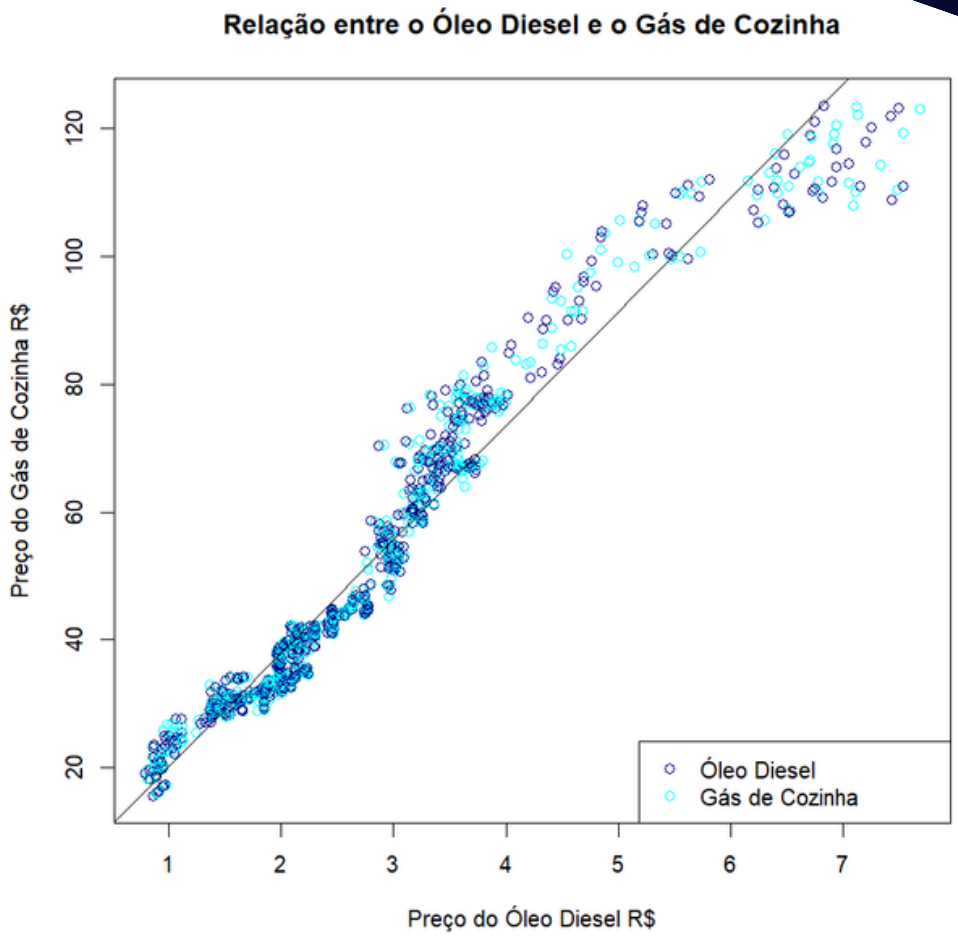

```r
# Relação entre gasolina comum e gás de cozinha

plot(dados_final$gasolina_comum_preco_revenda_avg, dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Relação entre a Gasolina Comum e o Gás de Cozinha',
     col = cores_variaveis[c(1,4)],
     xlab = 'Preço da Gasolina Comum R$',
     ylab = 'Preço do Gás de Cozinha R$')
legend(
  'bottomright',
  legend = c('Gasolina Comum', 'Gás de Cozinha'),
  pch = 1,
  col = cores_variaveis[c(1,4)]
)
abline(reg = lm(dados_final$gas_cozinha_glp_preco_revenda_avg ~ dados_final$gasolina_comum_preco_revenda_avg,
                lty = 'dashed'))
```




```r
# Relação entre etanol hidratado e óleo diesel

plot(dados_final$etanol_hidratado_preco_revenda_avg, dados_final$oleo_diesel_preco_revenda_avg,
     main = 'Relação entre o Etanol Hidratado e o Óleo Diesel',
     col = cores_variaveis[c(2,3)],
     xlab = 'Preço do Etanol Hidratado R$',
     ylab = 'Preço do Óleo Diesel R$')
legend(
  'bottomright',
  legend = c('Etanol Hidratado', 'Óleo Diesel'),
  pch = 1,
  col = cores_variaveis[c(2,3)]
)
abline(reg = lm(dados_final$oleo_diesel_preco_revenda_avg ~ dados_final$etanol_hidratado_preco_revenda_avg,
                lty = 'dashed'))
```




```r
# Relação entre etanol hidratado e gás de cozinha

plot(dados_final$etanol_hidratado_preco_revenda_avg, dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Relação entre o Etanol Hidratado e o Gás de Cozinha',
     col = cores_variaveis[c(2,4)],
     xlab = 'Preço do Etanol Hidratado R$',
     ylab = 'Preço do Gás de Cozinha R$')
legend(
  'bottomright',
  legend = c('Etanol Hidratado', 'Gás de Cozinha'),
  pch = 1,
  col = cores_variaveis[c(2,4)]
)
abline(reg = lm(dados_final$gas_cozinha_glp_preco_revenda_avg ~ dados_final$etanol_hidratado_preco_revenda_avg,
                lty = 'dashed'))
```




```
# Relação entre óleo diesel e gás de cozinha

plot(dados_final$oleo_diesel_preco_revenda_avg, dados_final$gas_cozinha_glp_preco_revenda_avg,
     main = 'Relação entre o Óleo Diesel e o Gás de Cozinha',
     col = cores_variaveis[c(3,4)],
     xlab = 'Preço do Óleo Diesel R$',
     ylab = 'Preço do Gás de Cozinha R$')
legend(
  'bottomright',
  legend = c('Óleo Diesel', 'Gás de Cozinha'),
  pch = 1,
  col = cores_variaveis[c(3,4)]
)
abline(reg = lm(dados_final$gas_cozinha_glp_preco_revenda_avg ~ dados_final$oleo_diesel_preco_revenda_avg,
                lty = 'dashed'))
```

Podemos inferir que os preços são correlacionados entre si já que os valores variam de forma semelhante.

Analise de variância

```
## Análise de variância ##
## Teste para ver se existe diferença no preço da gasolina por região
# H0 - Não existe diferença significativa entre o preço da gasolina por região
# H1 - Existe diferença significativa entre o preço da gasolina por região

an = aov(gasolina_comum_preco_revenda_avg ~ regiao, data=dados_final)
summary(an)

# p-valor = 0.49, não rejeitar H0, não há evidência estatística para afirmar que existe diferença
significativa entre o preço da gasolina por região
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
regiao        4    4.8   1.190   0.847  0.496
Residuals  1285 1806.5   1.406
```

```
## Teste para ver se existe diferença no preço do etanol por região
# H0 - Não existe diferença significativa entre o preço do etanol por região
# H1 - Existe diferença significativa entre o preço do etanol por região

an = aov(etanol_hidratado_preco_revenda_avg ~ regiao, data=dados_final)
summary(an)

# p-valor = 0.00 < 0.05, rejeitar H0, há evidência estatística para afirmar que existe diferença
significativa no preço do etanol dependendo da região
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
regiao        4   52.9   13.22   13.36 1.15e-10 ***
Residuals  1285 1271.7    0.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
# Como houve diferença significativa por região no preço do etanol, vamos ver quais regiões tiveram
mais diferença significativa
# Teste de tukey compara região por região

TukeyHSD(an)

# Diferença mais significativa de preço do etanol entre as regiões sudeste e norte
```

```
> TukeyHSD(an)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = etanol_hidratado_preco_revenda_avg ~ regiao, data = dados_final)

$regiao
                           diff          lwr         upr     p adj
NORDESTE-CENTRO OESTE  0.2343205 -0.004936729  0.47357781 0.0582192
NORTE-CENTRO OESTE     0.4159376  0.176680326  0.65519487 0.0000224
SUDESTE-CENTRO OESTE  -0.1688310 -0.408088279  0.07042626 0.3031128
SUL-CENTRO OESTE       0.0305031 -0.208754171  0.26976037 0.9968455
NORTE-NORDESTE         0.1816171 -0.057640217  0.42087433 0.2322010
SUDESTE-NORDESTE      -0.4031516 -0.642408822 -0.16389428 0.0000450
SUL-NORDESTE          -0.2038174 -0.443074713  0.03543983 0.1369691
SUDESTE-NORTE         -0.5847686 -0.824025876 -0.34551133 0.0000000
SUL-NORTE             -0.3854345 -0.624691767 -0.14617722 0.0001140
SUL-SUDESTE            0.1993341 -0.039923163  0.43859138 0.1533792
```

A diferença entre a região sudeste e norte é a maior representando que o sudeste possui o menor preço enquanto o norte o maior.

Regressão linear

```
## Regressão linear ##
# Modelo para prever o preço do gás de cozinha baseado no preço da gasolina comum

cor(dados_final$gas_cozinha_glp_preco_revenda_avg, dados_final$gasolina_comum_preco_revenda_avg)
# Correlação = 0.96, forte e positiva

modelo = lm(gas_cozinha_glp_preco_revenda_avg ~ gasolina_comum_preco_revenda_avg, data = dados_final)
modelo

# Coeficiente de determinação (% que a variável dependente é explicada pela variável explanatória ou independente)

summary(modelo)$r.squared
# R2 = 0.93, ou seja, 93% do preço do gás de cozinha consegue ser explicado pelo preço da gasolina comum

# O resultado é o preço do gás de cozinha quando a gasolina está valendo 1,2,3,4,5,6
predict(modelo, data.frame(gasolina_comum_preco_revenda_avg = c(1,2,3,4,5,6)))
```

```
> cor(dados_final$gas_cozinha_glp_preco_revenda_avg, dados_final$gasolina_comum_preco_revenda_avg)
[1] 0.9647457
> modelo = lm(gas_cozinha_glp_preco_revenda_avg ~ gasolina_comum_preco_revenda_avg, data = dados_final)
> modelo

Call:
lm(formula = gas_cozinha_glp_preco_revenda_avg ~ gasolina_comum_preco_revenda_avg,
    data = dados_final)

Coefficients:
                     (Intercept)  gasolina_comum_preco_revenda_avg
                          -12.33                             18.95

> summary(modelo)$r.squared
[1] 0.9307342
> predict(modelo, data.frame(gasolina_comum_preco_revenda_avg = c(1,2,3,4,5,6)))
        1         2         3         4         5         6
 6.626421 25.578021 44.529620 63.481219 82.432818 101.384418
```

Os dados possuem correlação de forma que: O preço do gás de cozinha varia de forma linear positiva acompanhando a gasolina.