In [1]:

```
# pacotes necessarios para executar funções e plotar gráficos
library(ggplot2)
library(dplyr)
library(RVAideMemoire)
library(rstatix)
library(tidyverse)
library(car)
```

```
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


*** Package RVAideMemoire v 0.9-80 ***


Attaching package: 'rstatix'


The following object is masked from 'package:stats':

    filter


-- Attaching packages --------------------------------------------------------------------------------- tidyverse
 1.3.1 --

v tibble  3.1.4     v purrr   0.3.4
v tidyr   1.1.3     v stringr 1.4.0
v readr   2.0.1     v forcats 0.5.1

-- Conflicts ------------------------------------------------------------------------------------ tidyverse_confl
icts() --
x rstatix::filter() masks dplyr::filter(), stats::filter()
x dplyr::lag()      masks stats::lag()

Carregando pacotes exigidos: carData


Attaching package: 'car'


The following object is masked from 'package:purrr':

    some


The following object is masked from 'package:dplyr':

    recode
```

In [2]:

```
# setando configs padrao para o display dos gráficos
options(repr.plot.width = 12, repr.plot.height = 9)
```

Primeiro, vamos importar o dataset reduzido e fazer algumas alterações para melhorar a visualização dos dados

In [3]:

```
# importando o dataset reduzido apenas com as variáveis que iremos usar
dados <- read.csv('hepatiteC-reduzido.csv')
head(dados)
```

A data.frame: 6 × 4

	X29.38	X39.48	X49.58	X59.68
	<dbl>	<dbl>	<dbl>	<dbl>
1	3.23	5.24	4.82	4.46
2	4.80	6.26	4.98	4.86
3	5.20	6.26	5.61	3.09
4	4.74	4.66	5.68	3.69
5	4.32	4.64	4.49	6.89
6	6.05	5.55	5.31	4.33

In [4]:

```
colnames(dados) <- c('De 29 a 38 anos', 'De 39 a 48 anos', 'De 49 a 58 anos', 'De 59 a 68 anos')
head(dados)
```

A data.frame: 6 × 4

	De 29 a 38 anos	De 39 a 48 anos	De 49 a 58 anos	De 59 a 68 anos
	<dbl>	<dbl>	<dbl>	<dbl>
1	3.23	5.24	4.82	4.46
2	4.80	6.26	4.98	4.86
3	5.20	6.26	5.61	3.09
4	4.74	4.66	5.68	3.69
5	4.32	4.64	4.49	6.89
6	6.05	5.55	5.31	4.33

In [5]:

```
nrow(dados)
```

197

Agora, vamos extrair algumas informações importantes das nossas variáveis para ter certeza que elas seguem uma distribuição semelhante à uma curva normal

In [6]:

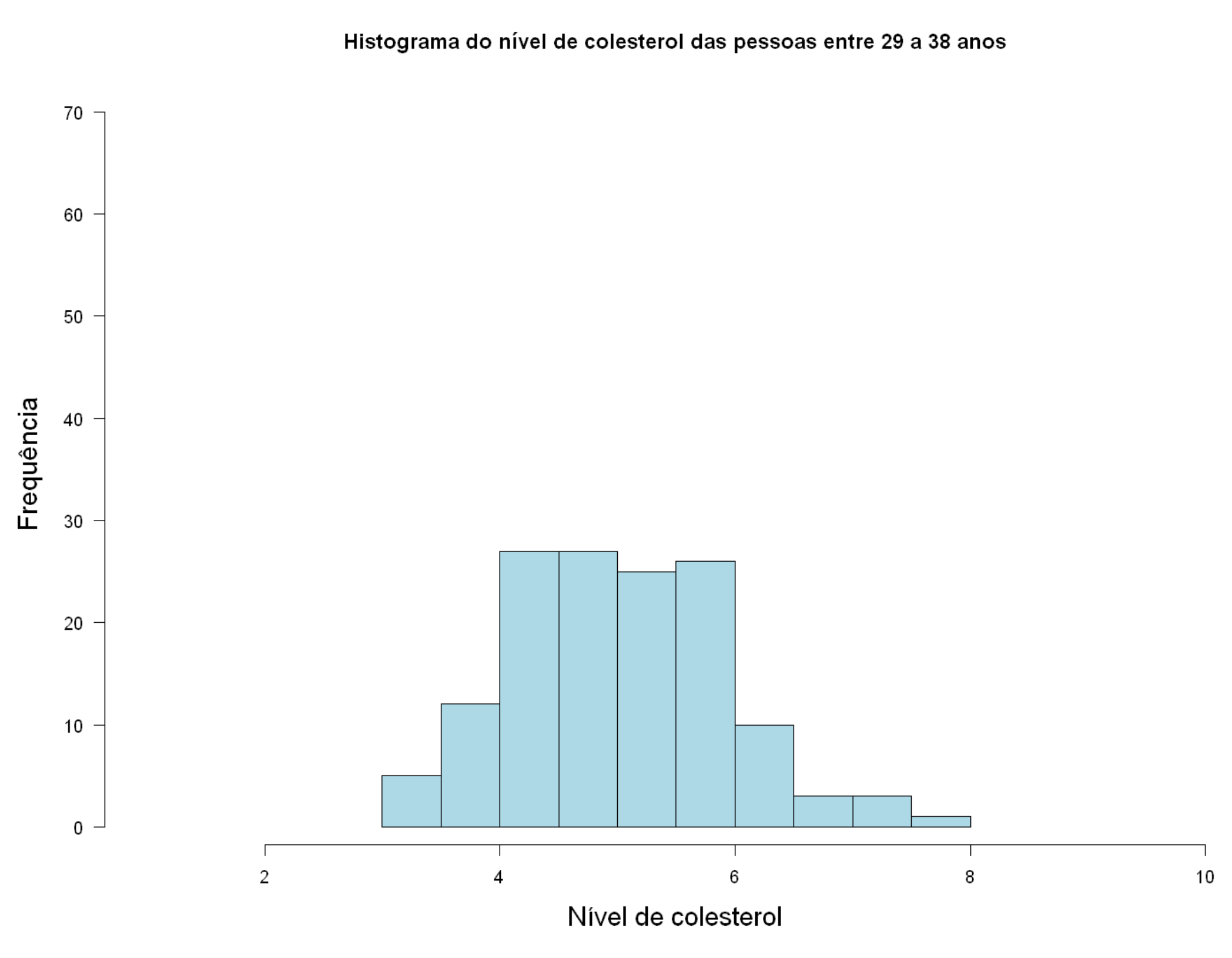
```
summary(dados)
```

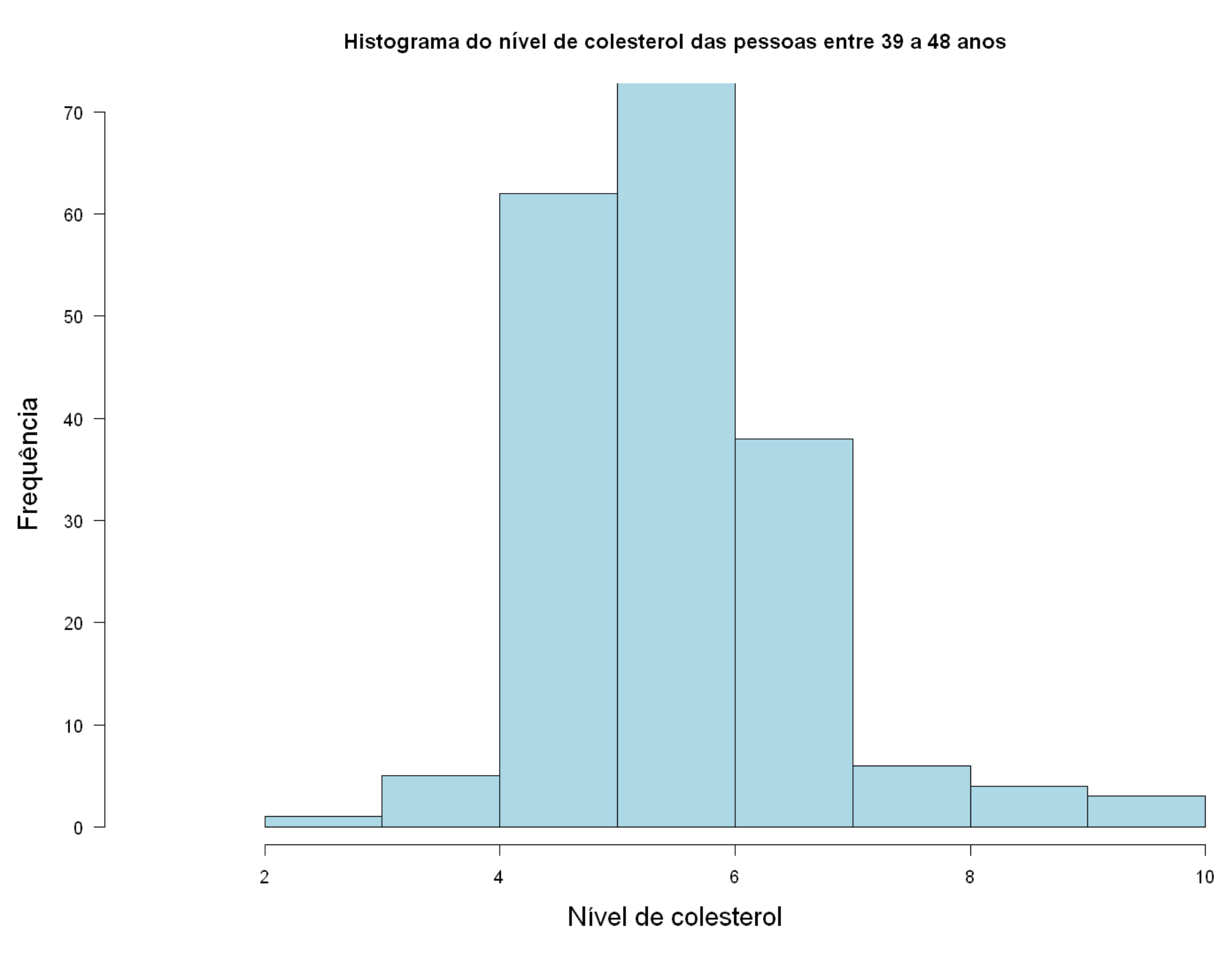
```
 De 29 a 38 anos De 39 a 48 anos De 49 a 58 anos De 59 a 68 anos
 Min.   :3.200   Min.   :2.860   Min.   :1.430   Min.   :3.020
 1st Qu.:4.345   1st Qu.:4.702   1st Qu.:4.860   1st Qu.:4.612
 Median :4.970   Median :5.300   Median :5.680   Median :5.475
 Mean   :5.006   Mean   :5.459   Mean   :5.598   Mean   :5.505
 3rd Qu.:5.620   3rd Qu.:6.022   3rd Qu.:6.390   3rd Qu.:6.393
 Max.   :7.510   Max.   :9.670   Max.   :8.800   Max.   :7.700
 NA's   :58      NA's   :5       NA's   :18      NA's   :117
```

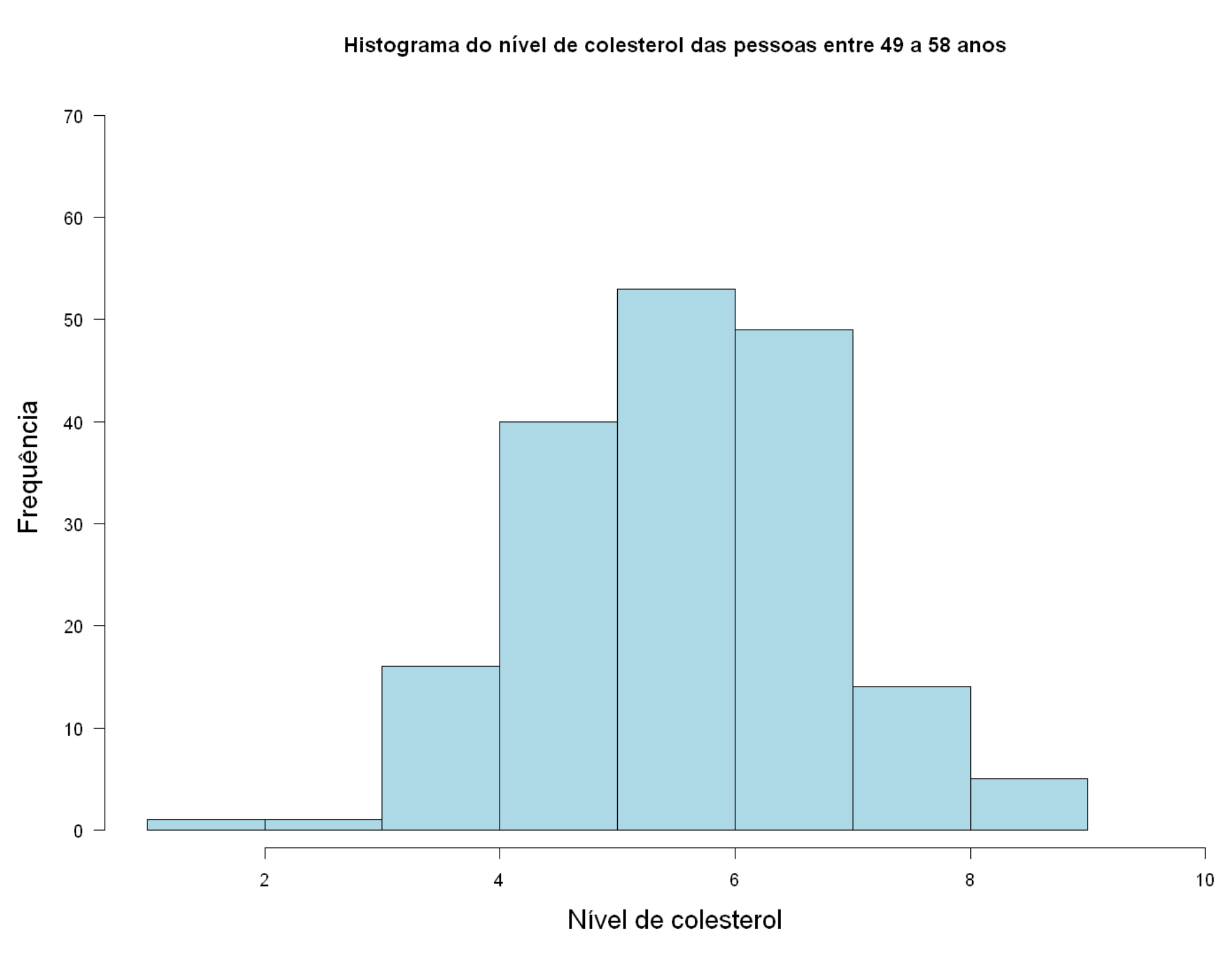
In [7]:

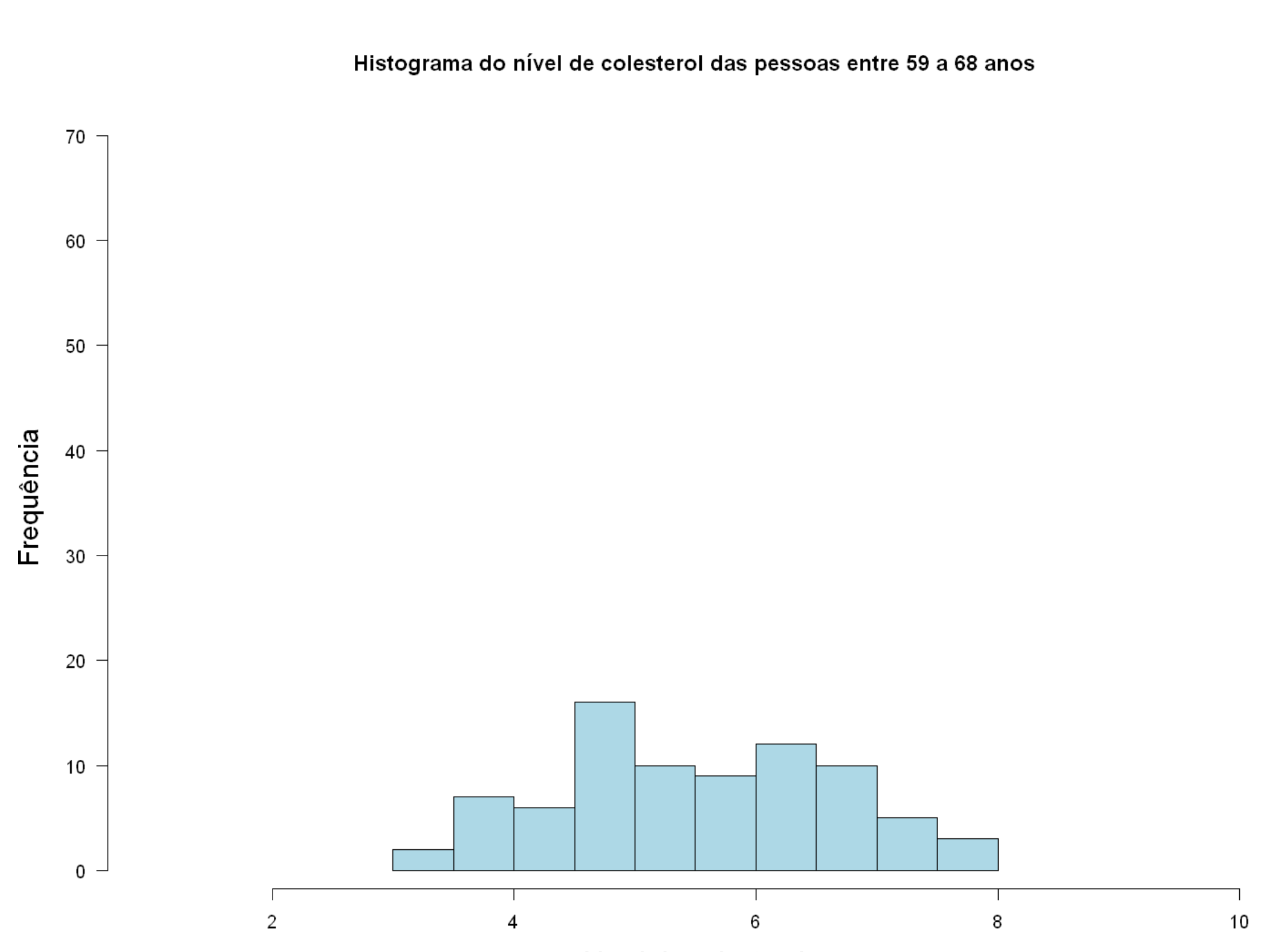
```
hist(
    x = dados$`De 29 a 38 anos`,
    breaks = 'Sturges',
    col = 'lightblue',
    main = 'Histograma do nível de colesterol das pessoas entre 29 a 38 anos',
    xlab = 'Nível de colesterol',
    ylab = 'Frequência',
    cex.lab=1.5,
    xlim=c(1, 10),
    ylim=c(1, 70),
    las = 1
)

hist(
    x = dados$`De 39 a 48 anos`,
    breaks = 'Sturges',
    col = 'lightblue',
    main = 'Histograma do nível de colesterol das pessoas entre 39 a 48 anos',
    xlab = 'Nível de colesterol',
    ylab = 'Frequência',
    cex.lab=1.5,
    xlim=c(1, 10),
    ylim=c(1, 70),
    las = 1
)
hist(
    x = dados$`De 49 a 58 anos`,
    breaks = 'Sturges',
    col = 'lightblue',
    main = 'Histograma do nível de colesterol das pessoas entre 49 a 58 anos',
    xlab = 'Nível de colesterol',
    ylab = 'Frequência',
    cex.lab=1.5,
    xlim=c(1, 10),
    ylim=c(1, 70),
    las = 1
)
hist(
    x = dados$`De 59 a 68 anos`,
    breaks = 'Sturges',
    col = 'lightblue',
    main = 'Histograma do nível de colesterol das pessoas entre 59 a 68 anos',
    xlab = 'Nível de colesterol',
    ylab = 'Frequência',
    cex.lab=1.5,
    xlim=c(1, 10),
    ylim=c(1, 70),
    las = 1
)
```









Com os histogramas acima, podemos ter uma noção intuitiva de que as variáveis seguem o formato de sino. Ou seja, seguem uma distribuição normal

---

No entanto, ter apenas uma noção intuitiva dessa propriedade não é o ideal. Para termos certeza e verificar a **normalidade por grupo**, devemos aplicar o teste de Shapiro-Wilk.

Antes disso, precisamos tranformar/inverter os dados do nosso dataset reduzido, isto é, ao invés de termos colunas para cada grupo e observações denotando o nível de colesterol de cada um, precisamos transformá-las em apenas duas:

1. **GrupoIdade**: Variável de agrupamento
2. **NivelColesterol**: Variável dependente

Dessa forma, cada observação do dataset passará a ser o registro de grupo de idade e nível de colesterol de um único indivíduo

In [8]:

```
dados_modificado <-
dados %>%
    pivot_longer(
        cols = everything(),
        names_to = "GrupoIdade",
        values_to = "NivelColesterol"
    )
```

In [9]:

```
head(dados_modificado)
```

A tibble: 6 × 2

GrupoIdade	NivelColesterol
<chr>	<dbl>
De 29 a 38 anos	3.23
De 39 a 48 anos	5.24
De 49 a 58 anos	4.82
De 59 a 68 anos	4.46
De 29 a 38 anos	4.80
De 39 a 48 anos	6.26

In [10]:

```
byf.shapiro(NivelColesterol ~ GrupoIdade, dados_modificado)
```

```
        Shapiro-Wilk normality tests

data:  NivelColesterol by GrupoIdade

                     W   p-value    
De 29 a 38 anos 0.9888    0.3284    
De 39 a 48 anos 0.9339 1.149e-07 ***
De 49 a 58 anos 0.9939    0.6700    
De 59 a 68 anos 0.9797    0.2353    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com esse teste, já temos uma informação importante:

> Ao contrário do que estávamos achando, agora sabemos com uma certeza muito grande que o grupo de idade de 39 a 48 anos **não** segue uma curva normal, já que o *p-value* < 0.05.

Agora, vamos submeter o nosso dataset ao teste de levene para testarmos a homogeneidade de variâncias dos grupos de idade.

In [11]:

```
leveneTest(NivelColesterol ~ GrupoIdade, dados_modificado, center=mean)
leveneTest(NivelColesterol ~ GrupoIdade, dados_modificado, center=median)
```

```
Warning message in leveneTest.default(y = y, group = group, ...):
"group coerced to factor."
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	3	4.276275	0.005329383
	586	NA	NA

```
Warning message in leveneTest.default(y = y, group = group, ...):
"group coerced to factor."
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	3	4.273967	0.005346238
	586	NA	NA

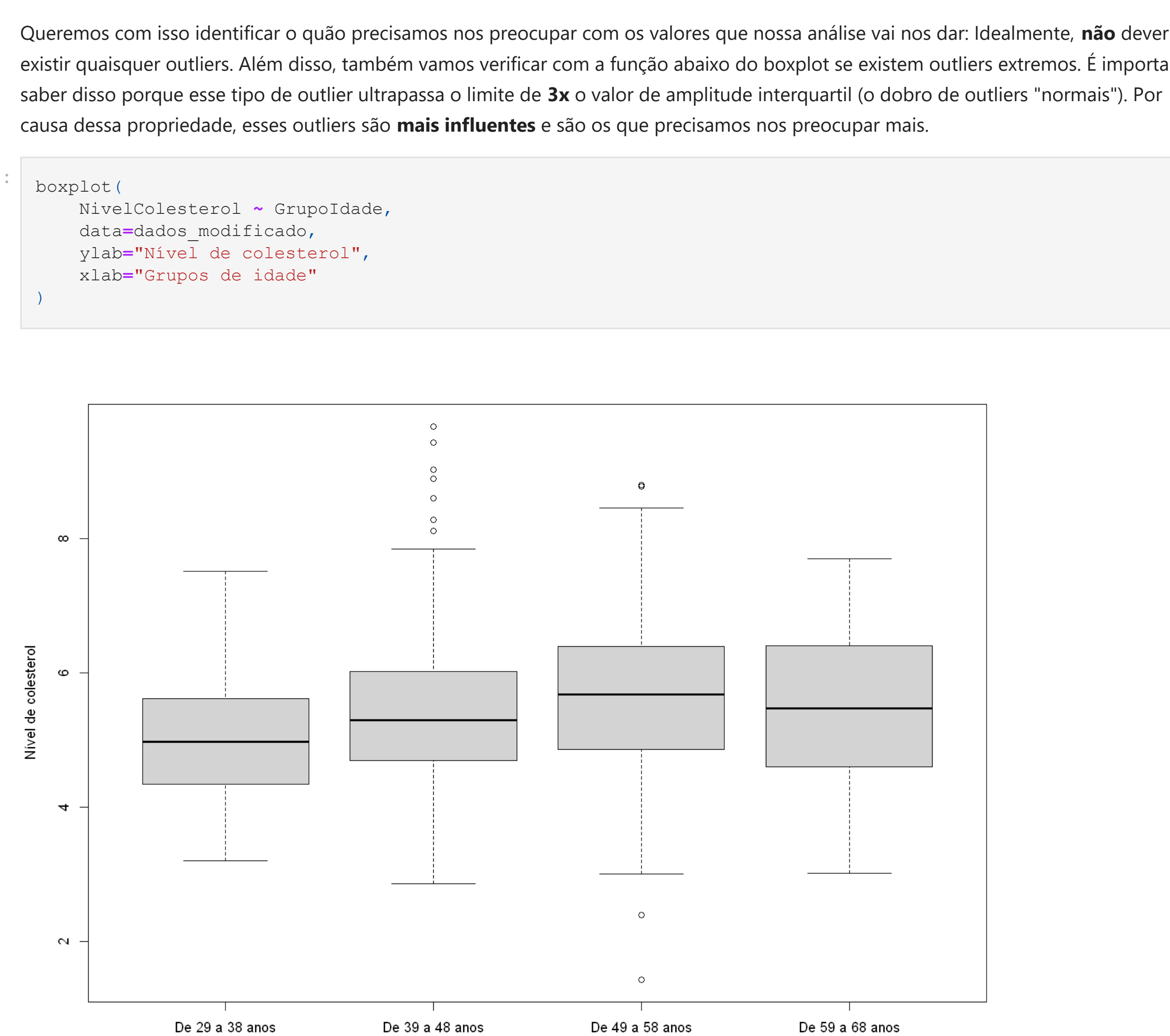
Com isso, também sabemos que as variâncias dos grupos são héterogêneas, já que o *p-value* < 0.05, levando em consideração o centro sendo tanto a mediana quanto a média.

Agora, vamos identificar os outliers das nossas amostras

Queremos com isso identificar o quão precisamos nos preocupar com os valores que nossa análise vai nos dar: Idealmente, **não** deveriam existir quaisquer outliers. Além disso, também vamos verificar com a função abaixo do boxplot se existem outliers extremos. É importante saber disso porque esse tipo de outlier ultrapassa o limite de **3x** o valor de amplitude interquartil (o dobro de outliers "normais"). Por causa dessa propriedade, esses outliers são **mais influentes** e são os que precisamos nos preocupar mais.

In [12]:

```
boxplot(
    NivelColesterol ~ GrupoIdade,
    data=dados_modificado,
    ylab="Nível de colesterol",
    xlab="Grupos de idade"
)
```



In [13]:

```
dados_modificado %>%
    group_by(GrupoIdade) %>%
    identify_outliers(NivelColesterol)
```

A tibble: 11 × 4

GrupoIdade	NivelColesterol	is.outlier	is.extreme
<chr>	<dbl>	<lgl>	<lgl>
De 39 a 48 anos	8.89	TRUE	FALSE
De 39 a 48 anos	8.11	TRUE	FALSE
De 39 a 48 anos	8.60	TRUE	FALSE
De 39 a 48 anos	9.43	TRUE	FALSE
De 39 a 48 anos	8.28	TRUE	FALSE
De 39 a 48 anos	9.03	TRUE	FALSE
De 39 a 48 anos	9.67	TRUE	FALSE
De 49 a 58 anos	8.80	TRUE	FALSE
De 49 a 58 anos	8.78	TRUE	FALSE
De 49 a 58 anos	2.40	TRUE	FALSE
De 49 a 58 anos	1.43	TRUE	FALSE

Com isso, sabemos que temos um total de 11 outliers do total de 590 indivíduos da amostra, **nenhum** deles sendo extremos. Logo, podemos calcular a anova.

In [14]:

```
anova <- aov(NivelColesterol ~ GrupoIdade, dados_modificado)
summary(anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
GrupoIdade    3   30.2   10.07   8.387 1.82e-05 ***
Residuals   586  703.3    1.20                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
198 observations deleted due to missingness
```

Por fim, podemos verificar que o *p-value* da nossa anova é muito inferior ao valor 0.05, o que quer dizer que a média de nível de colesterol dos grupos de idade são estatisticamente diferentes. Isto também quer dizer que nós rejeitamos a hipótese nula H0 e concluímos que a idade influencia sim no nível de colesterol do indivíduo.

Para corroborar com a nossa conclusão, ainda temos a aplicação do teste não paramétrico de Kruskal-Wallis, que pelo fato de não levar em consideração todas as premissas que a ANOVA tem, ele acaba sendo um teste alternativo para casos em que as nossas amostras **não** seguem uma distribuição normal.

No nosso caso, como observamos nos resultados do teste de **Shapiro-Wilk**, o grupo de indivíduos que possuem 39 a 48 anos **não** segue a curva gaussiana. Dessa forma, é importante que testemos através deste método, pois ele pode corroborar ou rejeitar a nossa conclusão baseada no resultado da ANOVA.

Uma observação válida é que o teste de Kruskal-Wallis feito pelo R é uma **versão simplificada**: Ao invés do teste comparar a distribuição dos grupos, ele apenas compara os valores de **mediana** dos grupos, colocando todos os componentes numa espécie de **ranking**.

In [15]:

```
kruskal.test(NivelColesterol ~ GrupoIdade, dados_modificado)
```

```
        Kruskal-Wallis rank sum test

data:  NivelColesterol by GrupoIdade
Kruskal-Wallis chi-squared = 27.191, df = 3, p-value = 5.369e-06
```

Como a hipótese nula *h0* do teste de kruskal-wallis é que a mediana dos grupos é estatisticamente igual, percebemos que devemos rejeitar tal hipótese tendo em vista que o nosso *p-value* resultante foi muito inferior ao valor 0,05 com 3 níveis de confiança. Isso quer dizer que mesmo aplicando um teste diferente, o resultado obtido só fez corroborar ainda mais a nossa conclusão de que **a idade influencia sim no nível de colesterol dos indivíduos do nosso dataset**.