Trilha 5
Bruna Matos

Problema 1

Utilizando então a base disponibilizada, você deve:

a) Ajustar um modelo de regressão linear sendo price a variável alvo (resposta), como função das demais variáveis citadas acima: horsepower, length, engine.size, city.mpg.

```
Call:
lm(formula = price ~ horsepower + length + engine.size + city.mpg,
    data = autos_imp)

Coefficients:
(Intercept)   horsepower       length  engine.size     city.mpg
 -4719.1036      33.0073      -0.4579     138.0737     -93.1429
```

b) Realizar a análise do modelo ajustado, avaliando o valor do R-quadrado, a significância estatística de cada parâmetro ajustado e a qualidade total do ajuste pela estatística F.

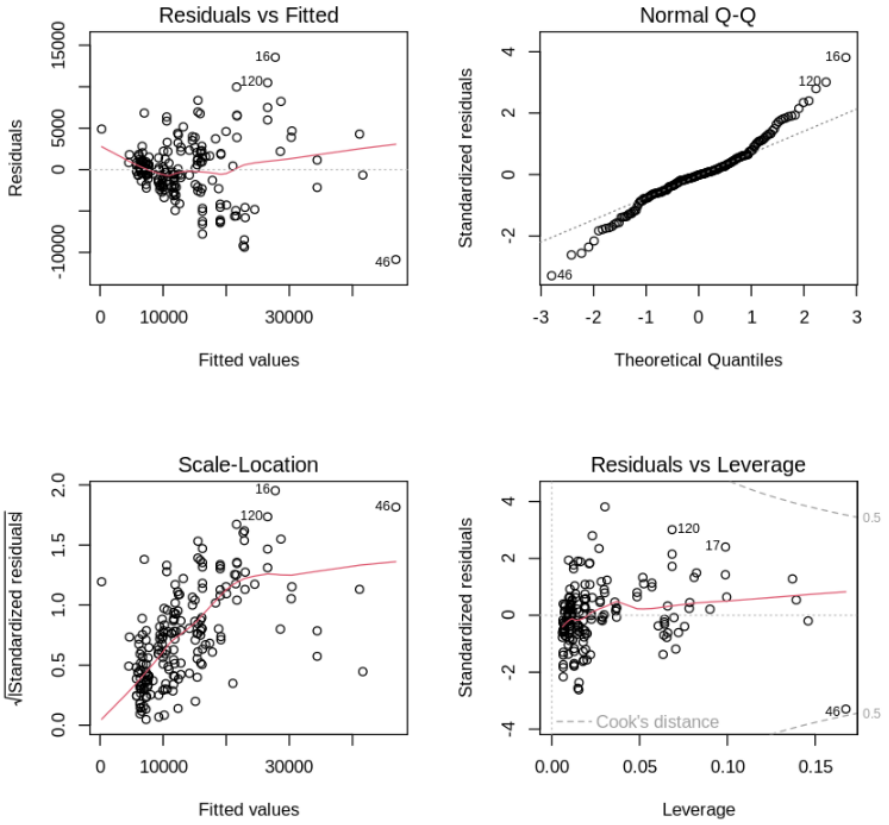
```
Call:
lm(formula = price ~ horsepower + length + engine.size + city.mpg,
    data = autos_imp)

Residuals:
   Min     1Q Median     3Q    Max
-10852  -1860    -78   1617  13547

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4719.1036  3276.8386  -1.440    0.151
horsepower     33.0073    16.2776   2.028    0.044 *
length         -0.4579     0.6081  -0.753    0.452
engine.size   138.0737    11.8043  11.697   <2e-16 ***
city.mpg      -93.1429    74.2520  -1.254    0.211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3609 on 188 degrees of freedom
Multiple R-squared:  0.8051,    Adjusted R-squared:  0.8009
F-statistic: 194.1 on 4 and 188 DF,  p-value: < 2.2e-16
```

A anova: 5 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
horsepower	1	8292710641	8292710641	636.6172631	2.903410e-62
length	1	6640435	6640435	0.5097749	4.761218e-01
engine.size	1	1794410044	1794410044	137.7538010	3.216149e-24
city.mpg	1	20497549	20497549	1.5735619	2.112478e-01
Residuals	188	2448927622	13026211	NA	NA

Podemos observar que o valor de Pr que as variáveis length e city.mpg não estão contribuindo com o meu sistema, ou seja, não possuem nenhuma correlação com o preço. Analisando o R_squared_ajusted percebemos que a equação obtida representa cerca de 80% da amostra e ainda temos um p-value muito abaixo de 5%, o que indica que a princípio a equação poderia ser utilizada para representar o meu sistema.

c) Realizar a verificação de aderência do modelo às premissas estatísticas do método dos mínimos quadrados através dos gráficos diagnósticos, comentando o gráfico dos resíduos x valores ajustados e o gráfico da curva Normal-QQ.



Analisando o gráfico de resíduos podemos perceber que não possuímos uma reta e que os pontos não estão uniformemente distribuídos em torno de curva. Ainda, observando o gráfico Q-Q percebemos que temos uma discrepância significativa no início e no fim do gráfico, ou seja, os pontos não estão em cima da reta e a reta também não possui uma angulação de 45 graus.

Assim, podemos dizer que a função obtida não é boa para utilizarmos para representar os nossos dados.

d) Fazer uma análise dos resultados do ajuste, discorrendo sobre o impacto de cada preditora, significativa do ponto de vista estatístico, no preço do carro.

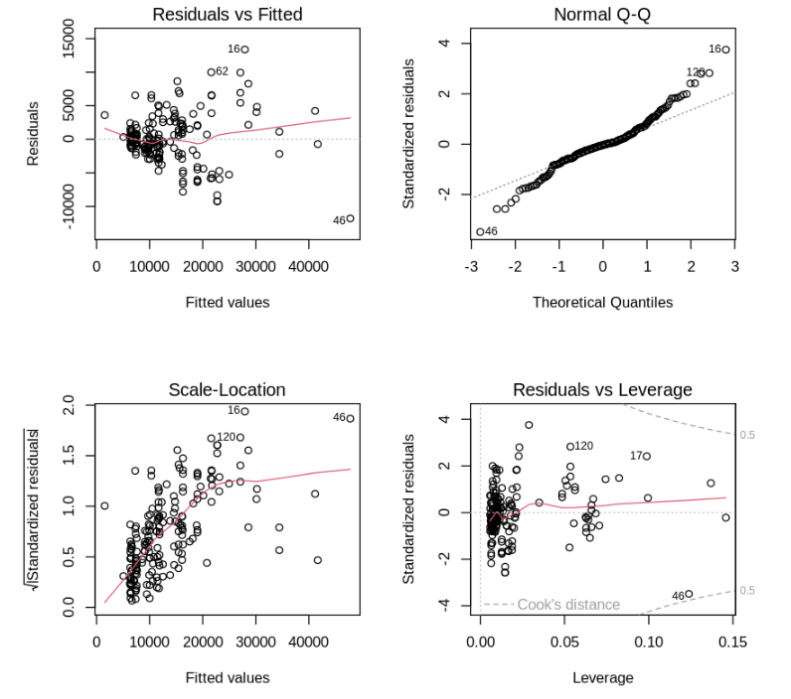
i) removendo city.mpg

```
Call:
lm(formula = price ~ horsepower + length + engine.size, data = autos_imp)

Coefficients:
(Intercept)   horsepower       length  engine.size
 -8552.9568      45.5501      -0.3818     138.4908
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-11797.2  -1875.0   -135.4   1533.7  13373.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8552.9568  1183.6569  -7.226 1.19e-11 ***
horsepower     45.5501    12.8640   3.541 0.000502 ***
length         -0.3818     0.6059  -0.630 0.529361
engine.size   138.4908    11.8175  11.719  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3615 on 189 degrees of freedom
Multiple R-squared:  0.8034,    Adjusted R-squared:  0.8003
F-statistic: 257.5 on 3 and 189 DF,  p-value: < 2.2e-16
```

A anova: 4 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
horsepower	1	8292710641	8292710641	634.6911538	2.540772e-62
length	1	6640435	6640435	0.5082325	4.767842e-01
engine.size	1	1794410044	1794410044	137.3370217	3.405411e-24
Residuals	189	2469425171	13065742	NA	NA



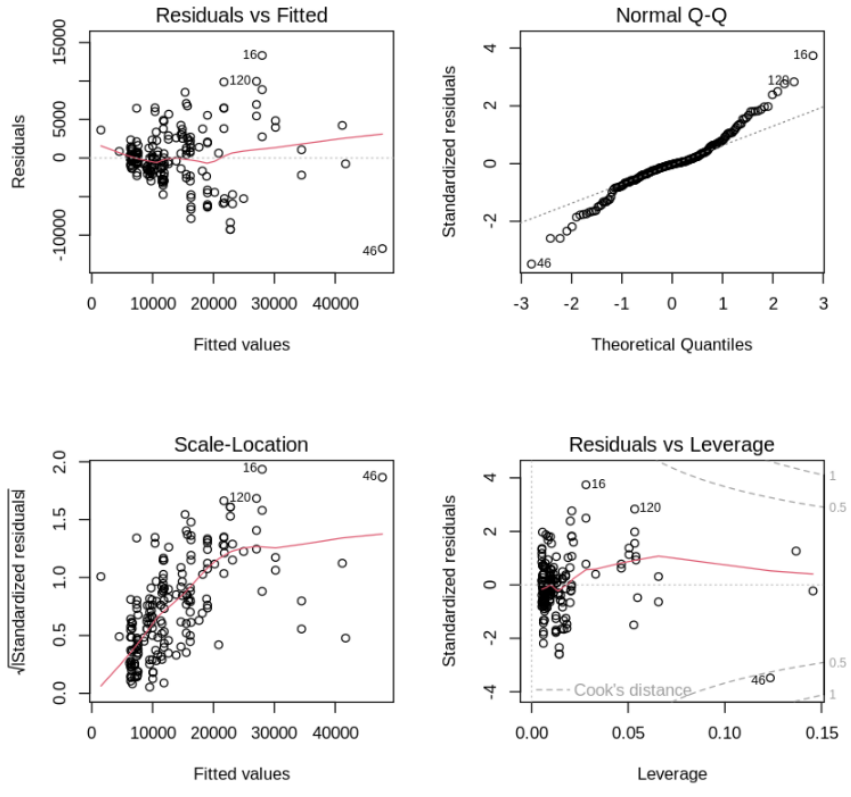
Removendo a variável city.mpg não resultou em nenhum impacto no meu sistema, portanto, ainda é necessário tomar alguma outra ação.

ii) removendo length

```
Call:
lm(formula = price ~ horsepower + engine.size, data = autos_imp)

Coefficients:
(Intercept)   horsepower  engine.size
   -9074.11        45.65       137.64

Residuals:
     Min       1Q   Median       3Q      Max
-11757.2  -1759.4    -81.2   1478.8  13313.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9074.11     845.45 -10.733  < 2e-16 ***
horsepower     45.65      12.84   3.555 0.000478 ***
engine.size   137.64      11.72  11.742  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3609 on 190 degrees of freedom
Multiple R-squared:  0.803,     Adjusted R-squared:  0.801
F-statistic: 387.3 on 2 and 190 DF,  p-value: < 2.2e-16
```

A anova: 3 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
horsepower	1	8292710641	8292710641	636.7116	1.415743e-62
engine.size	1	1795862292	1795862292	137.8857	2.722606e-24
Residuals	190	2474613358	13024281	NA	NA



Removendo length pudemos perceber uma leve melhora no valor de R_ajusted_squared, mas nenhuma outra melhoria foi identificada.

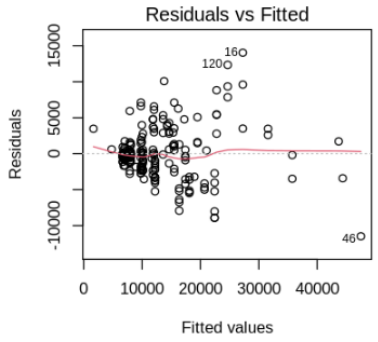
iii) removendo horsepower

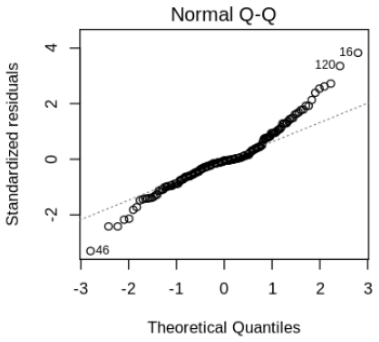
```
Call:
lm(formula = price ~ engine.size, data = autos_imp)

Coefficients:
(Intercept)  engine.size
    -8862.8         172.9
```

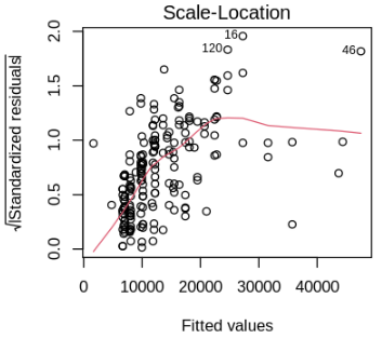
```
Residuals:
   Min     1Q Median     3Q    Max
-11490  -2031   -193   1460  14050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8862.79     868.66   -10.2   <2e-16 ***
engine.size   172.86       6.45    26.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3717 on 191 degrees of freedom
Multiple R-squared:  0.7899,    Adjusted R-squared:  0.7888
F-statistic: 718.2 on 1 and 191 DF,  p-value: < 2.2e-16
```

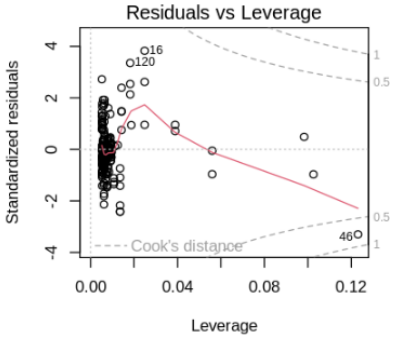
A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
engine.size	1	9924002742	9924002742	718.2087	1.252508e-66
Residuals	191	2639183549	13817715	NA	NA









Usando somente engine.size obtivemos uma piora no nosso modelo, pois analisando o valor de R_squared de agora é 79%. Essa piora também pode ser observada no gráfico de resíduoes.