Projeto com Feedback 1

José Aleilson

2023-01-21

Machine Learning em Logística Prevendo o Consumo de Energia de Carros Elétricos

Este Projeto buscou oferecer insight para uma empresa de transporte que deseja migrar sua frota de carros elétricos objetivando a redução de custos.

Com esse objetivo a empresa gostaria de prever o consumo de energia de carros elétricos com base em diversos fatores de utilização dos veículos e características dos veículos.

Para atingir o objetivo dessa pesquisa utilizaremos um conjunto de dados com informações de veículos elétricos adquiridos na Polônia a partir de 2 de novembro de 2020.

Análise Exploratória

Ao verificar se existiam valores *NA* no conjunto de dados, foram detectados o total de 30 valores *NA*. Ao observar isto tomei a decisão de substituir os valores *NA* das variáveis numéricas por suas respectivas médias, e, a única observação com valor *NA* encontrado numa variável categórica decidi por remove-la do dataset.

Variáveis Catégoricas

A tabela a seguir traz informações dos tipos de freios e trações dos carros elétricos.

```
##
##                              2WD (front) 2WD (rear)  4WD
##   disc (front + rear)               40.4       11.5 34.6
##   disc (front) + drum (rear)         3.8        9.6  0.0
```

Analisando essa tabela acima é possível observar que aproximadamente 40% do carros que apresentam tração frontal possuem discos de freio frontal e traseiro, e aproximadamente 35% dos carros que possuem tração nas quatro rodas, tem freios a discos frontais e traseiros. Das 52 observações cerca de 13,04% possuem freio a disco frontal e tambor traseiro.

A seguir será apresentado informações dos tipos de freios dos carros elétricos e sua marca.

```
##
##                 disc (front + rear) disc (front) + drum (rear)
##   Audi                         11.5                        0.0
##   BMW                           5.8                        0.0
##   Citroën                       3.8                        0.0
##   DS                            1.9                        0.0
##   Honda                         3.8                        0.0
##   Hyundai                       5.8                        0.0
##   Jaguar                        1.9                        0.0
##   Kia                           7.7                        0.0
##   Mazda                         1.9                        0.0
##   Mercedes-Benz                 1.9                        0.0
##   Mini                          1.9                        0.0
##   Nissan                        5.8                        0.0
##   Opel                          3.8                        0.0
##   Peugeot                       3.8                        0.0
##   Porsche                       7.7                        0.0
##   Renault                       3.8                        0.0
##   Skoda                         0.0                        1.9
##   Smart                         0.0                        3.8
##   Tesla                        13.5                        0.0
##   Volkswagen                    0.0                        7.7
```

Apenas as marcas Volkswagen, Smart, Skoda utilizam o tipo de freio a disco frontal e tambor traseiro. As demais marcas utilizam freios a disco frontal e traseiro. Com relação as variáveis categóricas tomei a decisão de segui apenas com as variáveis marca, tipo de freio e tração. Acredito que o nome do carro e o seu modelo não sejam fundamentais para determinar o consumo médio de energia dos carros elétricos. E Além disso, o modelo tem informações semelhantes a variável nome.

Engenharia de atributos

O código a seguir buscou atribuir valores numéricos as informações contidas nas variáveis categóricas. Vejamos:

```r
# Atribuindo valores numericos as variáveis categoricas.
dataset$tipo_tracao <- as.numeric(as.factor(dataset$tipo_tracao))
dataset$tipo_freio <- as.numeric(as.factor(dataset$tipo_freio))
dataset$marca <- as.numeric(as.factor(dataset$marca))
```

Com isso os informações antes descritas por um nome de freio ou tração especifica foi substituído por um número, a modo de representar a mesma informação.

Variáveis Numéricas

As variáveis a seguir são o vetores númericos disponibilizados no conjunto de dados.

```
var_num <- c ("preco_min","forca_motor","maximo_torque","capacidade_carga",
              "distancia_max_perco","distancia_eixo","comprimento","largura",
              "altura","peso_vazio_min","peso_aceito","peso_max","assentos",
              "portas","tamanho_pneu","vel_max","cap_carga","acel_s",
              "pot_max_carre", "media_consener"
              )
```

O códico a seguir cria plots de histograma e boxplot de algumas variáveis do conjunto de dados. Vejamos:

```
# plots
col <- c("peso_max", "media_consener", "peso_aceito",
         "cap_carga", "acel_s")
mapply(function(x,col) {
  nf <- layout( matrix(c(1,2), ncol=2) )
  hist(x , breaks=20 , border=F , col=rgb(1,0,0,0.5), main="", xlab=col)
  title(main = "Histograma")
  boxplot(x , col=rgb(0,0,1,0.5) , las=2, xlab=col)
  title(main = "Boxplot")
}, subset_num[col], col)
```

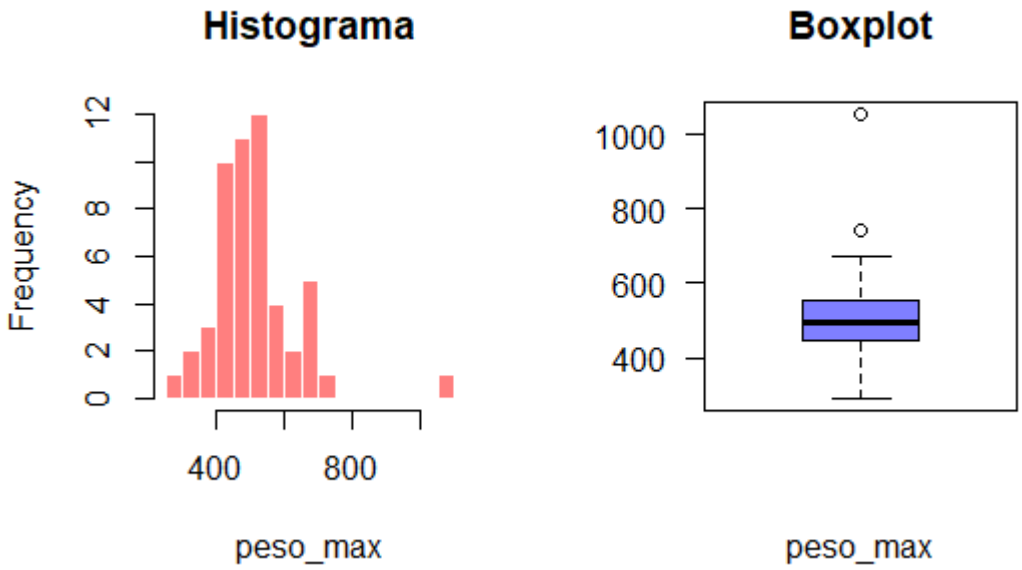
Plot - 1 Peso Maximo

Vejamos algumas estatísticas da variável Peso Máximo.

```
summary(dataset$peso_max)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   290.0   445.0   495.5   513.8   546.2  1056.0
```

Ao observa o summary da variável fica vemos que o menor peso máximo de uma carro elétrico nesse banco de dados é de 290 kg e o maior peso máximo é de 1056 kg. Ainda é possível observar que a mediana foi de 495 kg e a média de 513 kg, isso nos sugere que existem valores extremos nos dados que puxam a média para cima. O plot a seguir apresenta essas informações no formato de histograma e boxplot da variável Peso máximo.



1. Analisando histograma é possível observa que a maior parte das observações estão concentrada em 445 kg e 546 kg. Isto é, o peso máximo dos carros elétricos em sua maioria estão entre esse intervalo.
2. No boxplot foi possível detectar dois valores com potencial outlier, um sendo superior a 1000 kg e outro próximo a 800 kg. Ainda sobre o boxplot e distribuição dos dados parecem ser igualmente distribuídas no primeiro e terceiro quartil.

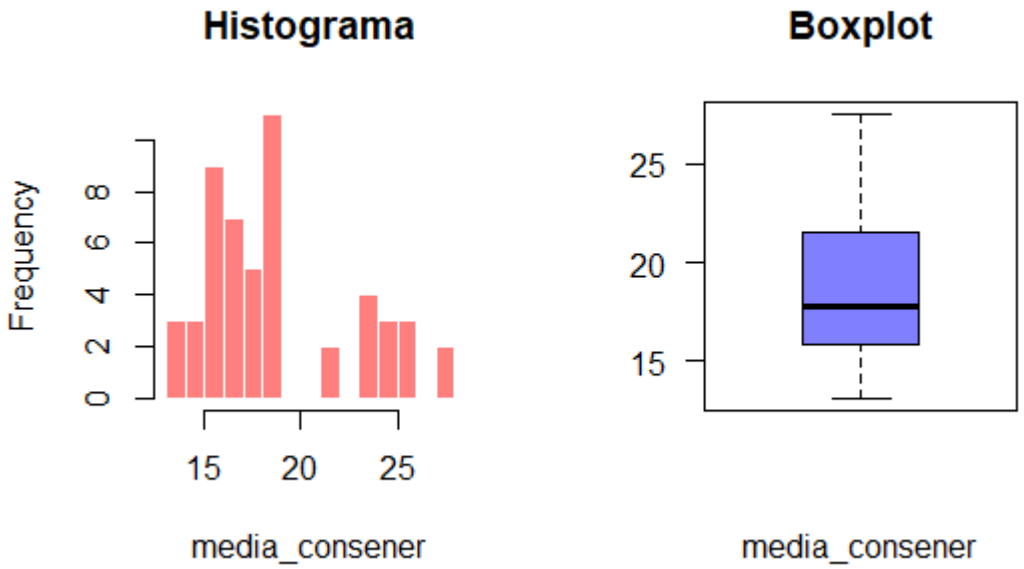
Plot - 2 Média de Consumo de Energia

Vejamos algumas estatísticas da variável média de consumo de energia.

```
summary(dataset$media_consener)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.10   15.85   17.80   18.82   21.36   27.55
```

O valor médio mínimo de consumo energia observado foi de 13.10 kWh/100 km e o valor máximo foi de 27.55 kWh/100 km. Ainda é possível observar que os valores da média e média são muito próximos o que indica que não existe valores extremos nesse vetor. O plot a seguir apresenta informações no formato de histograma e boxplot da variável Média de Consumo de Energia.



1. Ao explorar o histograma de início é observado que a variável media_consener não segue uma destruição normal. Além disso, a maior contagem dos dados está concentrada entre o intervalo de 15 a 21 kWh/100 km. Ou seja, a maioria do carros apresentam um consumo médio de energia de 15 a 21 kWh/100 km.
2. O boxplot dessa variável nos mostra que a maior concentração de dados está no terceiro quartil.

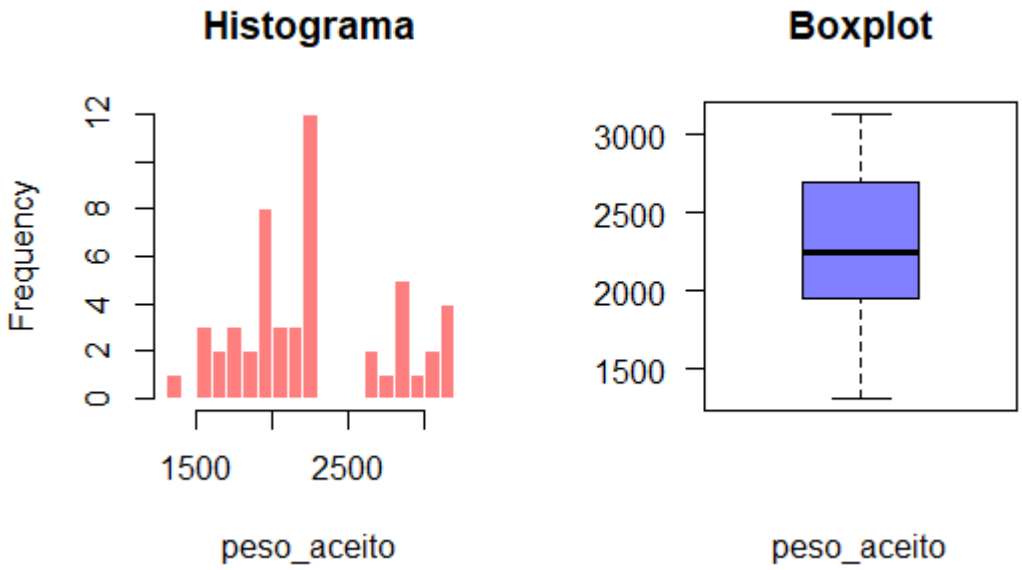
Plot - 3 Peso Aceito

Vejamos algumas estatísticas da Peso aceito.

```
summary(dataset$peso_aceito)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1310    1957    2240    2266    2684    3130
```

Observando a variável peso aceito, vemos que o peso bruto admissível mínimo encontrado nessa dataset foi de 1310 kg e o máximo de 3013 kg. Também é visto que as medidas de tendência central media e mediana são muito próximas, o que nos dá o indicativo de que não existe outliers nessa variável. O plot a seguir apresenta informações no formato de histograma e boxplot da variável Peso Aceito.



1. Observando a histograma da variável peso aceito podemos ver que ela não segue uma distribuição normal.
2. Analisando o boxplot é visto que o terceiro quartil apresentam uma maior concentração de dados em comparação com o primeiro quartil.

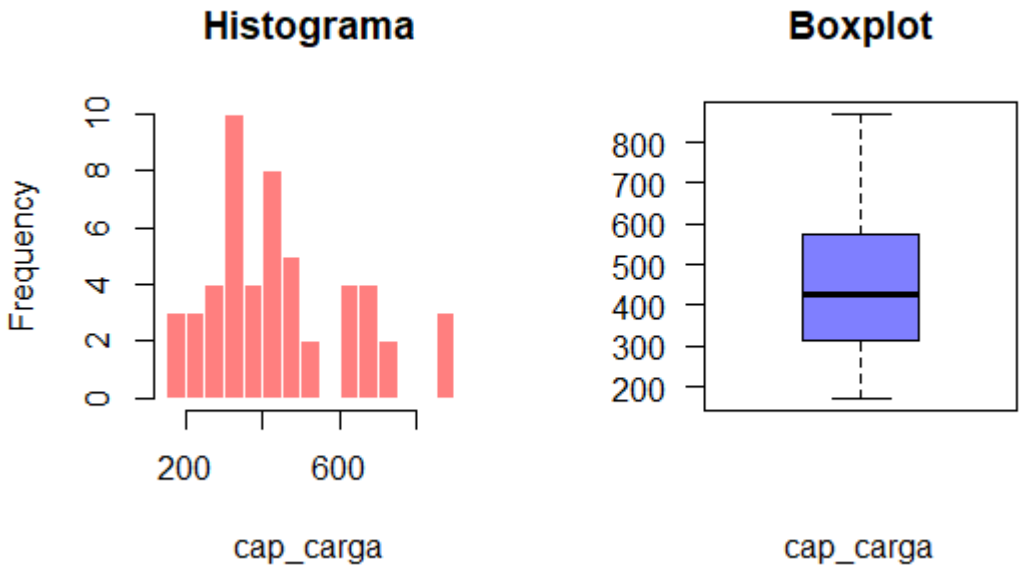
Plot - 4 Capacidade de Carga

Vejamos algumas estatísticas capacidade de carga.

```
summary(dataset$cap_carga)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   171.0   315.0   425.0   445.1   558.0   870.0
```

Ao observar o resultados é visto que é menor valor observado foi de 171 L e o maior valor de 870 L. A média dessa variável foi de 445 L e a mediana 425 L, como o valor de média é superior ao valor de mediana existe a possibilidade é existir valores extremos. O plot a seguir apresenta informações no formato de histograma e boxplot da variável capacidade de carga.



1. Essa variável assim como as demais não segue uma destruição normal. Observando o histograma vemos que a maior contagem dos dados está entre o intervalo de 315 L a 558 L. Ou seja, os carro elétricos em sua maioria para esse conjunto de dados tem uma capacidade de carga medida em litros no intervalo de 315 L a 558 L.
2. Observando o boxplot é possível detectar um pouco de concentração dos dados no terceiro quartil em comparação ao primeiro quartil.

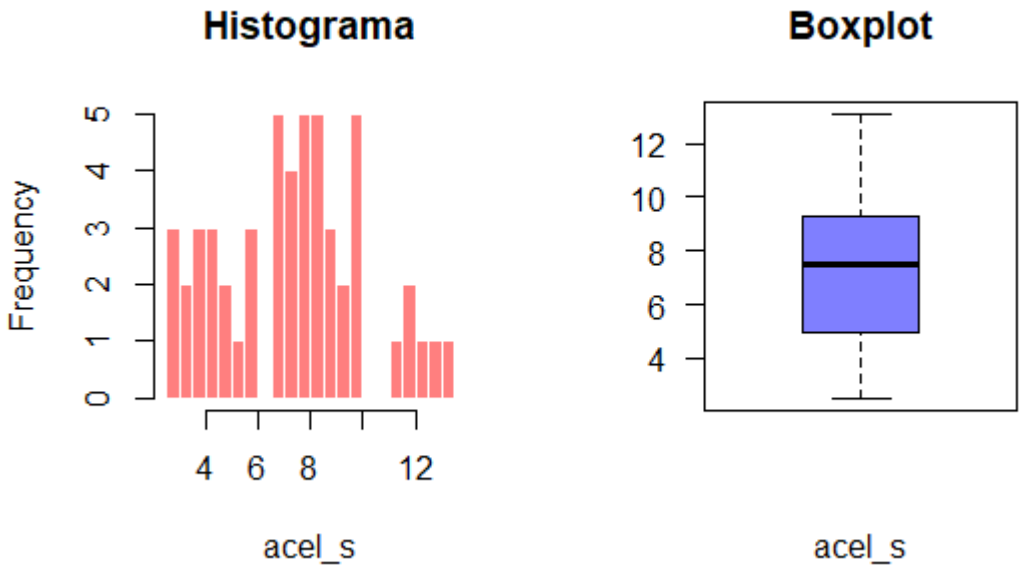
Plot - 5 Aceleração de 0-100 Kph em segundos

Vejamos algumas estatísticas da Aceleração de 0-100 Kph em segundos.

```
summary(dataset$acel_s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.500   5.025   7.480   7.360   9.125  13.100
```

Ao analisar a estatísticas da variável **acel_s**, que é medida em segundos, foi visto que o menor tempo observado para atingir 100kph foi de 2.5 segundos e o maior tempo foi de 13.1 segundos. Os valores da média e mediana são muito próximos o que indica que não existe valores extemos nesse vetor. O plot a seguir apresenta informações no formato de histograma e boxplot da variável capacidade de carga.



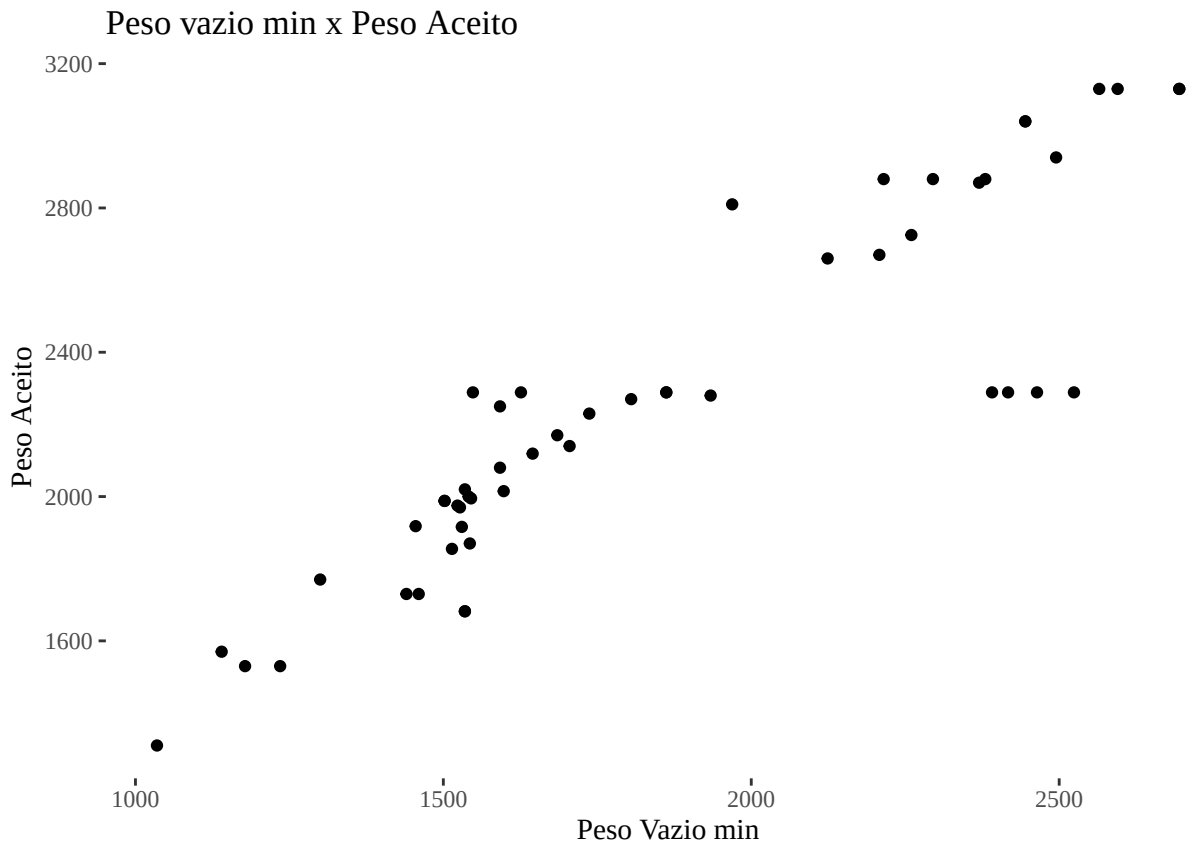
1. De acordo com o histograma a grande maioria dos carros desse conjunto de dados apresentam uma aceleração de 0 a 100 kph medido em segundos de 7 a 10 s. Isto é, o maioria dos carros desse dataset levam para atingir os 100 kph demora cerca de 7 a 10 segundos.
2. Ao observar o boxplot é possível detectar que o primeiro quartil apresenta um maior volume de dados se comparado com o terceiro quartil.

Análise estatística bivariada

A análise estatística bivariada permite fazer uma análise entre duas variáveis de modo a entender o comportamento entre a variáveis estudadas.

plot 1

Vejamos no Plot a seguir o relacionamento entra a variável **peso vazio mínimo** e **peso aceito**.



Como podemos ver no gráfico acima as variáveis **peso vazio mínimo** e **peso aceito** apresentam alta correlação positiva, isto é, à medida que **peso aceito** aumenta o **peso vazio mínimo** também aumenta. Agora veremos o resultado do cálculo de correlação entre essas variáveis para confirmar o que foi visto no gráfico.

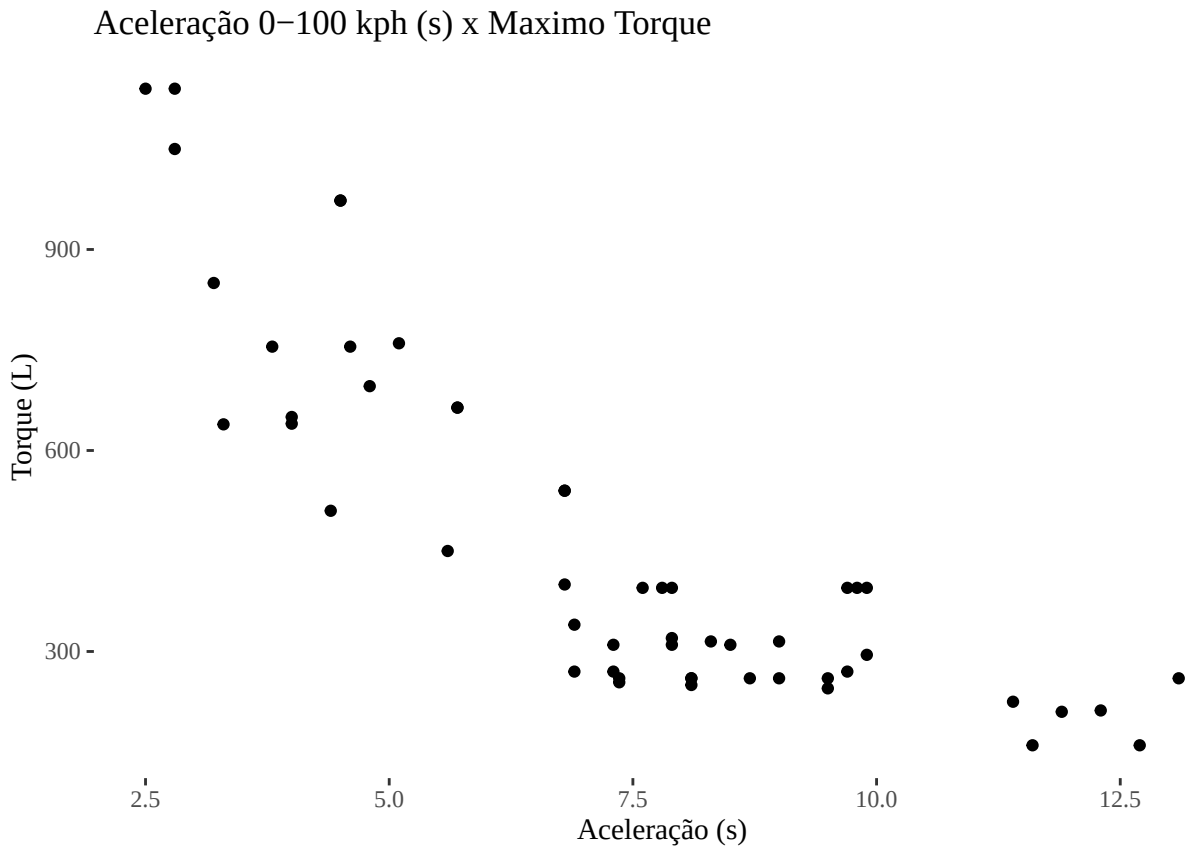
```
cor(subset_num$peso_vazio_min, subset_num$peso_aceito)
```

```
## [1] 0.9035717
```

Sabemos que a correlação é uma medida estatística entre -1 e 1, quanto mais próximo de -1 maior é a correlação negativa entre a variáveis e quanto mais próximo de 1 maior é a correlação positiva entre as variáveis. Nesse exemplo a variáveis **peso vazio mínimo** e **peso aceito** apresentaram uma correlação de aproximadamente 0.90, isto significa que ela apresentam uma forte correlação positiva. Isto posto, essa variáveis independentes podem apresentar problemas de multicolinearidade ao modelo de regressão, por isso, mais afrente tomarei a decisão de escolher a melhor variável para o modelo proposto.

Plot 2

O gráfico a seguir demonstra a relação entre as variáveis **Aceleração de 0-100 kph (s)** e **Máximo Torque**. É importante saber que Torque é uma medida de força, é uma grandeza vetorial associada às forças que produzem rotação em um corpo.



É visto no Plot 2 que a medida que o **Máximo Torque** sobe, o tempo medido em segundos para atingir a aceleração de 0 a 100 kph cai. Logo, as variáveis **Aceleração de 0-100 kph (s)** e **Máximo Torque** possível uma correlação negativa. Veremos a seguir o cálculo de correlação estre essas variáveis.

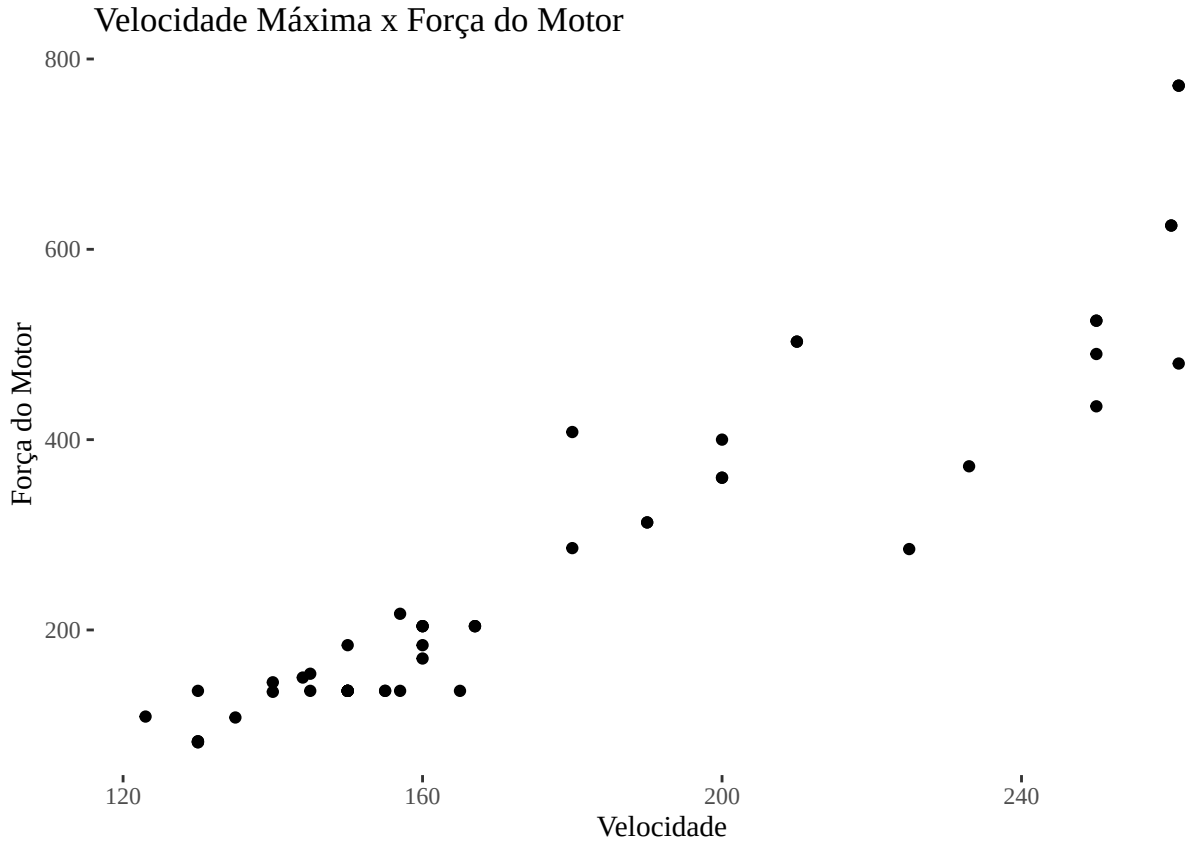
```
cor(subset_num$acel_s, subset_num$maximo_torque)
```

```
## [1] -0.8395833
```

Assim como vimos no plot 2, o cálculo de correlação dessa variáveis nos diz que elas apresentam forte correlação negativa. Esses vetores podem passar informações semelhantes ao modelo e assim ocorrer problemas de multicolinearidade.

Plot 3

O gráfico a seguir traz informações sobre a relação entre as variáveis **velocidade máxima** e **força do motor**.



É observado no gráfico que esses vetores possuem uma relação positiva, isto é, a medida que a **força do motor** aumenta a **velocidade máxima** do carro também aumenta. Vejamos o cálculo de correlação para verificar a força da relação dessas duas variáveis.

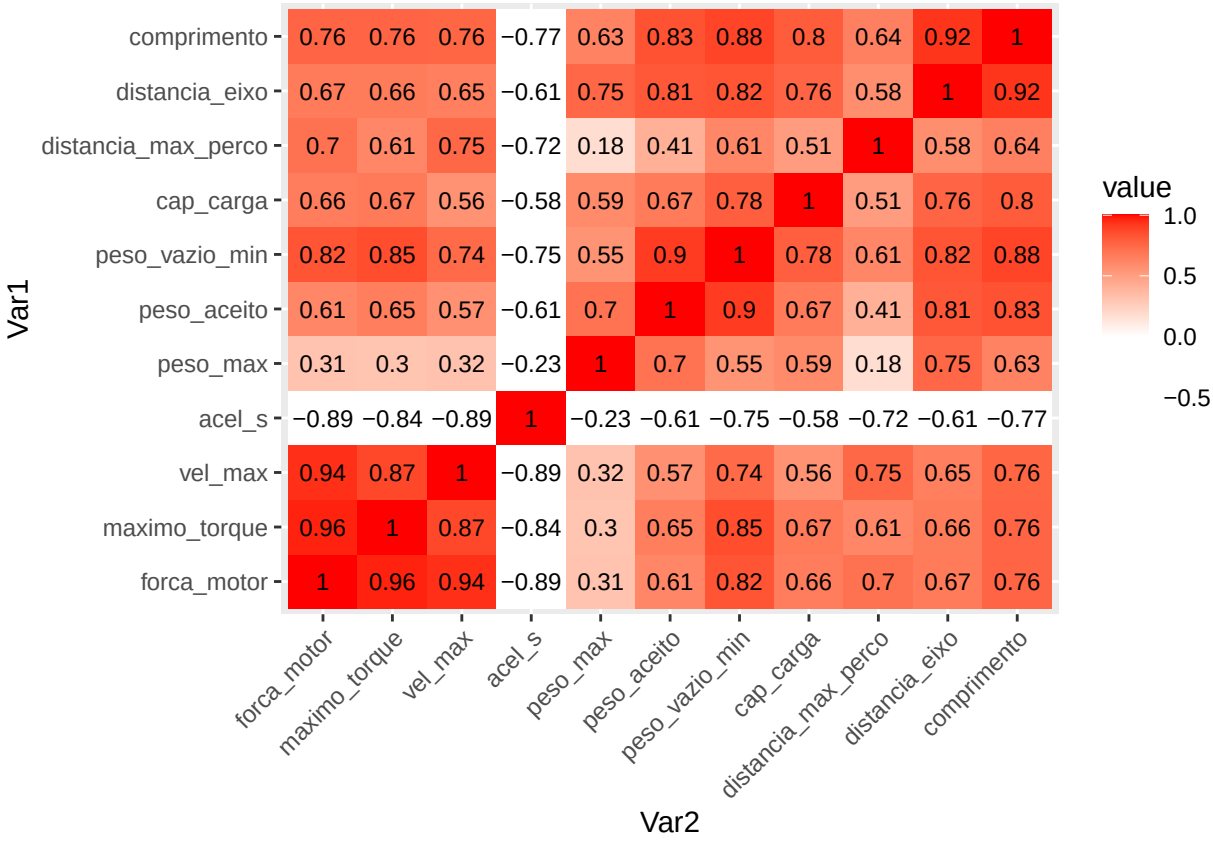
```
cor(subset_num$vel_max, subset_num$forca_motor)
```

```
## [1] 0.9367542
```

O cálculo estatístico obtido entre essas duas variáveis foi de aproximadamente 0.93, isso significa que elas possuem uma forte correlação positiva.

Plot 4

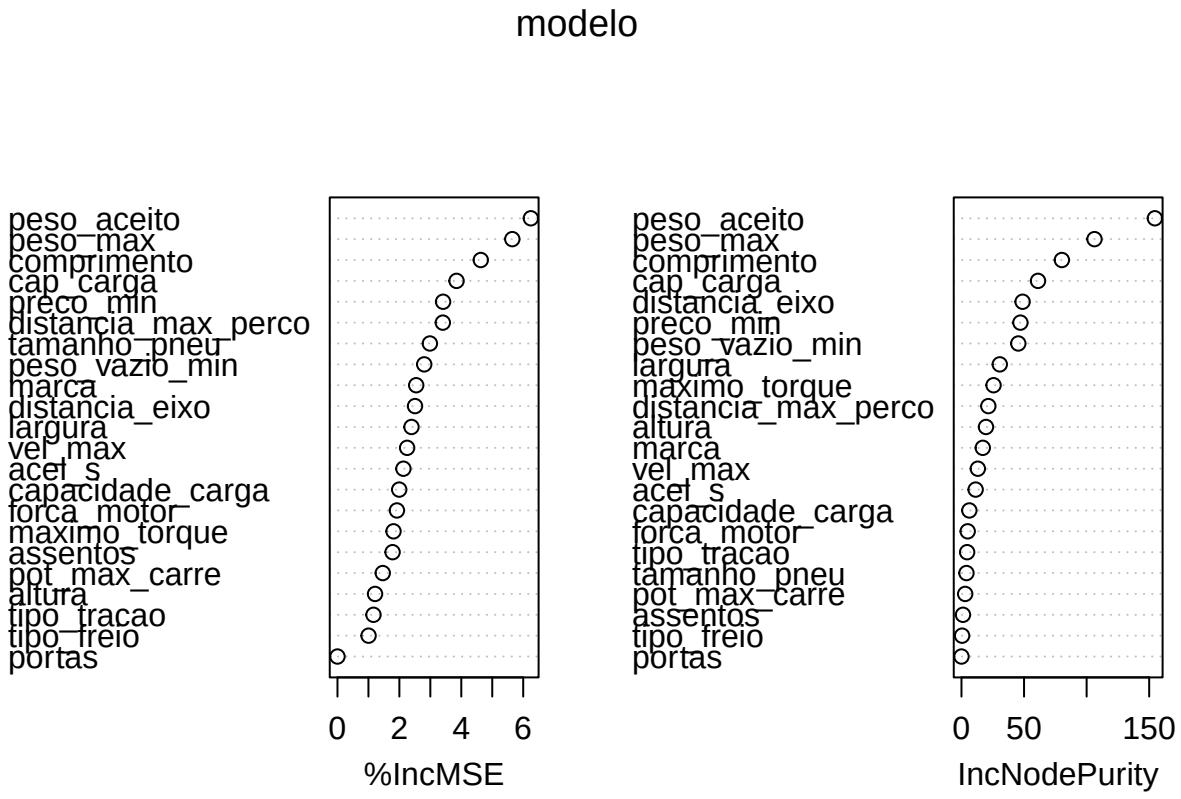
A seguir será apresentado um heatmap de correlação entre as variáveis numéricas disponível nesse banco de dados.



Nesse heatmap as variáveis com um tom em vermelho apresentam correlação positiva e as que apresentam um tom mais próximo ao branco tem correlação negativo. Dessa forma, os vetores **vel_max**, **forca_motor** e **maximo_torque** apresentam uma forte correlação positiva entre si, essas variáveis também apresentam um uma forte correlação negativa com **acel_s**.

Feature Selection

Nessa seção vamos selecionar as variáveis com características mais relevantes e informativas para o modelo proposto mais à frente. E ao mesmo tempo reduzir o número de características a serem utilizadas no modelo. Para isso decidi utilizar o modelo **Random Forest** para calcular a importância das variáveis. Vejamos o gráfico a seguir.



Antes de rodar o modelo Random Forest foi aplicado uma padronização dos dados com a função **scale**. Ao observar o modelo Rf decidi selecionar as 15 variáveis mais importantes de acordo com a métrica MSE do modelo.

```
var_fs <- c ("peso_max","peso_aceito","comprimento","distancia_eixo","cap_carga",
             "vel_max","tamanho_pneu","largura","marca","preco_min","peso_vazio_min",
             "distancia_max_perco","maximo_torque","capacidade_carga","pot_max_carre",
             "media_consener"
              )

# Feature selection
df_fs <- df %>%
  select(all_of(var_fs))
```

Pré processamento

Aqui iremos dividir os dados em treino e teste e em seguir dar início ao processamento do modelo de regressão.

```
# dividindo dados em treino e teste
set.seed(1998)
indice_treinamento <- createDataPartition(df_fs$peso_max, p = 0.7, list = FALSE)
dados_treinamento <- df_fs[indice_treinamento,]
dados_teste <- df_fs[-indice_treinamento,]
```

Processamento do Modelo: Regressão Linear Multipla

Modelo 1

```
# Modelo de Regressão 1
model1 <- lm(media_consener ~ ., data = dados_treinamento)
summary(model1)
```

```
##
## Call:
## lm(formula = media_consener ~ ., data = dados_treinamento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8644 -0.7654  0.1673  0.7210  2.3202
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.54336    0.22755  81.490  < 2e-16 ***
## peso_max             -0.18283    0.65007  -0.281  0.78104
## peso_aceito           2.22508    0.89929   2.474  0.02116 *
## comprimento          -1.93656    1.21829  -1.590  0.12558
## distancia_eixo        2.21470    1.08614   2.039  0.05310 .
## cap_carga            -0.33274    0.93673  -0.355  0.72566
## vel_max               1.53516    0.99238   1.547  0.13552
## tamanho_pneu         -0.68492    0.49433  -1.386  0.17918
## largura              -0.20612    0.26035  -0.792  0.43663
## marca                -0.10778    0.31276  -0.345  0.73352
## preco_min            -0.72607    0.74450  -0.975  0.33958
## peso_vazio_min       -1.11834    1.47647  -0.757  0.45647
## distancia_max_perco -3.11823    0.95729  -3.257  0.00347 **
## maximo_torque         0.08167    1.01125   0.081  0.93633
## capacidade_carga      4.42496    1.60691   2.754  0.01131 *
## pot_max_carre         0.51072    0.90408   0.565  0.57760
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.288 on 23 degrees of freedom
## Multiple R-squared:  0.9386, Adjusted R-squared:  0.8985
## F-statistic: 23.44 on 15 and 23 DF,  p-value: 1.989e-10
```

Observando os resultados obtidos por esse modelo vemos que:

- **Residual standard error** teve um valor baixo de 1.288 o que indica que em termos médio a variabilidade dos erros foram relativamente baixas.
- **Multiple R-squared** obtido nesse modelo foi de 0.9386, isso nos diz que cerca de 93,86% da variação

da variável de estudo é explicada pelas variáveis independentes do modelo. O que é considerado um valor alto, ou seja, o modelo tem um alto valor explicativo.

- **p-value** desse modelo vou muito baixo, 1.989e-10. Dito isso, a hipótese nula que diz que as variáveis independentes são irrelevantes para explicação do modelo é rejeitada. Ou seja, temos evidência de que ao menos um variável seja relevantes na explicação da variável target.

Embora esse modelo de regressão linear múltipla tenha apresentado bons parâmetros em termos explicativos. É importante lembrar que esse modelo proposto utilizou-se de um grande número de variáveis e apenas algumas apresentaram um bom nível de significância. Ainda, foi visto anteriormente que algumas das variáveis independentes desse modelo apresentam alta correlação entre si e isso pode ter afetado o modelo com problemas de multicolinearidade e por conseguinte de overfitting.

Modelo 2

Como foi visto na análise exploratória algumas variáveis independentes apresentaram um alta correlação. Agora, irei excluir alguns desses vetores do modelo com o objetivo de evitar possíveis problemas de multicolinearidade e overfitting.

```
# Modelo de Regressão 2

var2 <- (media_consener ~ peso_max + peso_aceito + capacidade_carga + vel_max +
           distancia_max_perco + pot_max_carre + largura + marca +
           + tamanho_pneu)

#treino
model2 <- lm(var2, data = dados_treinamento)
summary(model2)
```

```
##
## Call:
## lm(formula = var2, data = dados_treinamento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9520 -0.7410  0.1530  0.7662  2.4292
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.64398    0.21705  85.899  < 2e-16 ***
## peso_max              0.58218    0.32064   1.816  0.07977 .
## peso_aceito           1.93651    0.60442   3.204  0.00329 **
## capacidade_carga      2.75738    0.81857   3.369  0.00215 **
## vel_max               1.67152    0.65647   2.546  0.01646 *
## distancia_max_perco -2.51276    0.54330  -4.625 7.19e-05 ***
## pot_max_carre       -0.19299    0.49288  -0.392  0.69825
## largura             -0.17039    0.25404  -0.671  0.50770
## marca               -0.02154    0.28698  -0.075  0.94069
## tamanho_pneu        -1.08405    0.42216  -2.568  0.01565 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.343 on 29 degrees of freedom
## Multiple R-squared:  0.9158, Adjusted R-squared:  0.8897
## F-statistic: 35.06 on 9 and 29 DF,  p-value: 3.203e-13
```

Dos resultados:

- O **Residual standard error** continuou com valor baixo de 1.343 o que indica que em média de variabilidade dos erros foram relativamente baixas.
- O **Multiple R-squared** obtido no modelo 2 foi de 0.9158, ele nos diz que cerca de 91,58% da variação da variável de estudo é explicada pelas variáveis independentes do modelo. Ainda que esse modelo detinha tido uma perda em seu valor explicativo em comparação com o modelo 1, ele ainda tem uma alto valor explicativo.
- **p-value** do modelo 2 continuou muito baixo, 3.203e-13. Assim como o modelo anterior, a hipótese nula foi rejeitada. Ou seja, esse modelo apresenta evidências de que ao menos uma variável seja relevantes na explicação da variável dependente.

A retirada de algumas variáveis desse modelo não teve prejuízo significativo na capacidade de explicação da variável dependente. E além disso, essa redução no número de variáveis independentes reduz os custos de reprodução do modelo. Essa redução aumentou o nível se significância de alguns vetores nesse modelo.

Modelo 3

Com objetivos semelhantes ao modelo anterior tomei a decisão de remover as variáveis marca, largura, peso_max e pot_max_carre. Essas variáveis apresentaram um baixo nível de significância estatística no modelo.

```
# Modelo de Regressão 3

var3 <- (media_consener ~ peso_aceito + vel_max + tamanho_pneu +
           distancia_max_perco + capacidade_carga
)

# Treino
model3 <- lm(var3, data = dados_treinamento)
summary(model3)
```

```
##
## Call:
## lm(formula = var3, data = dados_treinamento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9126 -0.8114  0.1434  0.7299  2.6872
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.6518     0.2153  86.633  < 2e-16 ***
## peso_aceito           2.4997     0.4776   5.234 9.26e-06 ***
## vel_max               1.4715     0.4436   3.317  0.00222 **
## tamanho_pneu         -1.1166     0.3770  -2.962  0.00563 **
## distancia_max_perco  -2.3251     0.4972  -4.676 4.77e-05 ***
## capacidade_carga      2.3118     0.7677   3.011  0.00496 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.334 on 33 degrees of freedom
## Multiple R-squared:  0.9054, Adjusted R-squared:  0.8911
## F-statistic: 63.16 on 5 and 33 DF,  p-value: 6.2e-16
```

Resultados obtidos modelo 3:

- **Residual standard error** do modelo 3 assim como o outros dois modelos apresentou um valor baixo de 1.334, indicando que em termos médio a variabilidade dos erros foram relativamente baixas.
- Já o **Multiple R-squared** obtido no modelo 3 foi de 0.9054, nos indicando que cerca de 90,54% da variação o objeto de estudo pode ser explicado pelas variáveis independentes desse modelo. Isso é considerado um alto valor explicativo.
- A **F-statistic** obtida nesse modelo foi consideravelmente alta. O valor F de 63.16 mostra que a variação explicada pelas variáveis independentes do modelo é maior que a variação que não é explicada pelo modelo.
- O **p-value** continuou com um valor baixo de 3.203e-13. Com isso, a hipótese nula que foi rejeitada. Ou seja, nesse modelo temos evidência suficiente de que ao menos uma variável seja relevantes na explicação da variável de dependente.

Esse modelo de regressão linear também apresentou bons parâmetros em termos explicativos da variável dependente. O modelo 3 apresentou de acordo com a estatística F maior potencial de explicabilidade da variável de estudo. É importante ressaltar que nesse modelo as variáveis independentes tiverem um alto valor de significância no modelo.