Adriana Eva Fernandes da Silva

```
set.seed(1)
library(ggpubr)
library(conflicted)
library(data.table)
library(tm)
library(glmnet)
library(tidyverse)
library(dplyr)
dados <- read.csv("~/mineração de dados/atividade-1-precificacao-veiculos.csv")
#transformando em dummies:
df_encoded <- model.matrix(~ . - 1, data = dados)

# Exibindo o resultado

dtm.matrix <-df_encoded %>% as.matrix()
dtm.matrix <- dtm.matrix[, -c(1, ncol(dtm.matrix))]
dim(dtm.matrix)
```

```
## [1] 205 190
```

Podemos notar que, filtrando os termos da matriz de dados, teremos um total de 190 termos.

Item A:

Vamos dividir os dados em 60% para treinamento e 40% para teste:

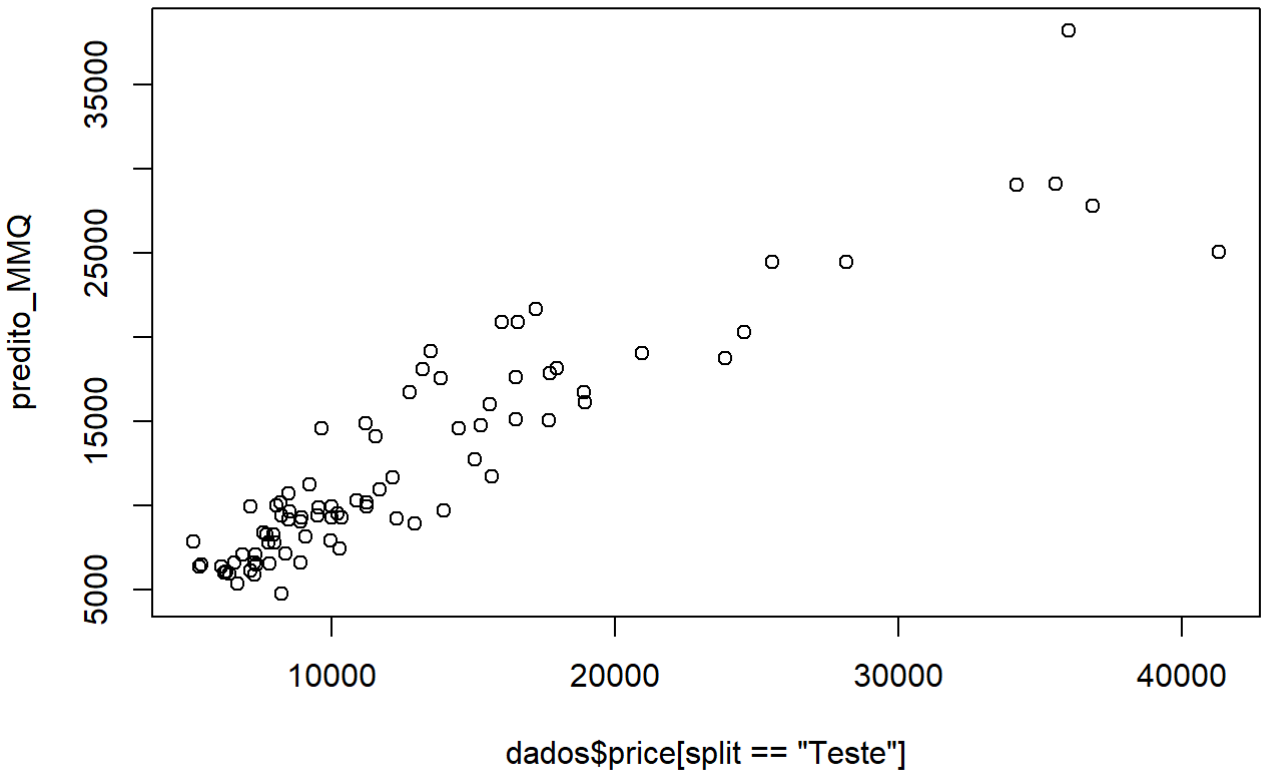
```
set.seed(1)
split<- sample(c("Treinamento","Teste"),prob=c(0.6,0.4),size = nrow(dados),replace = TRUE)
```

Item B:

Mínimos quadrados

A função glmnet é empregada para realizar o ajuste do modelo do lasso, que é uma técnica estatística utilizada para a seleção de variáveis. Por outro lado, o método de mínimos quadrados representa um caso específico do lasso, em que o parâmetro lambda é igual a zero. A seguir temos o ajuste para o modelo de mínimos quadrados:

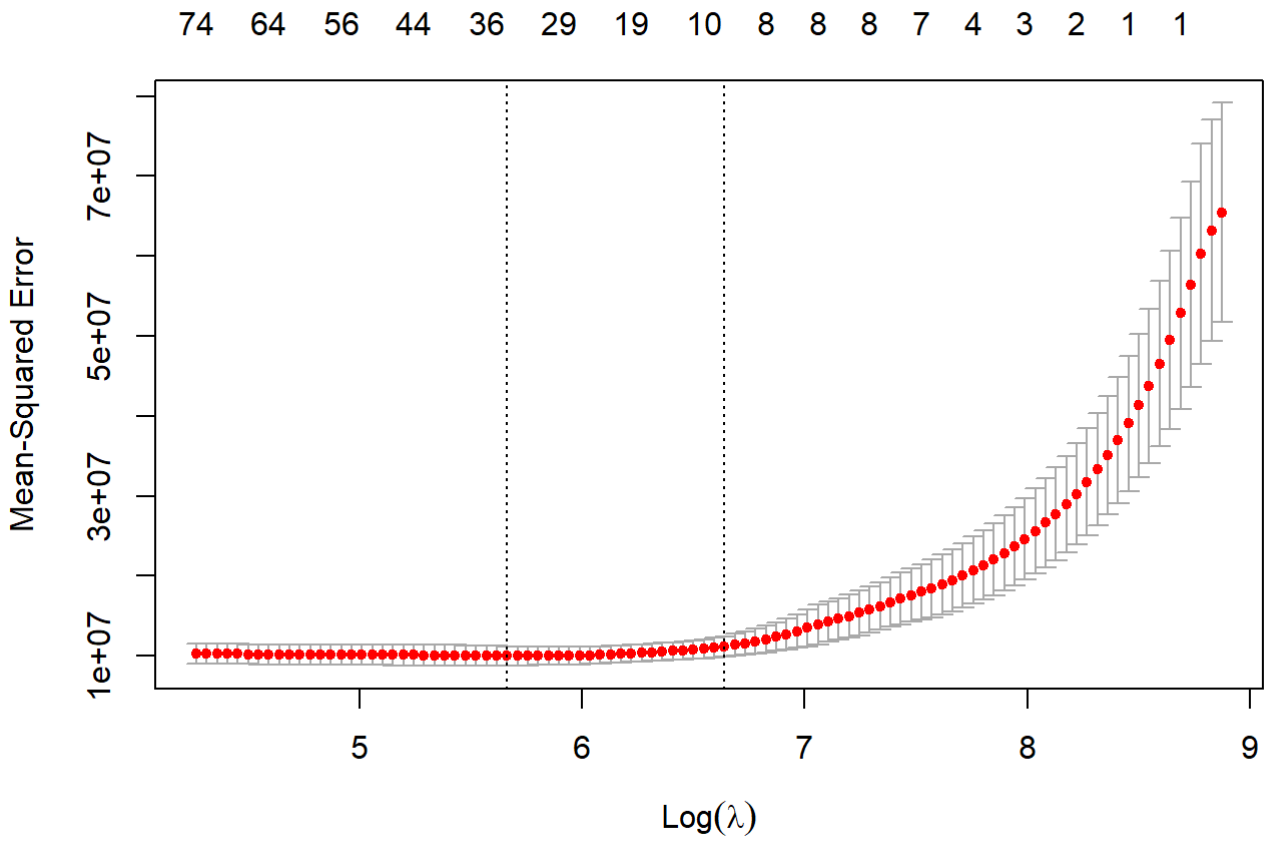
```
riscos<- list()

vc_lasso<- cv.glmnet(dtm.matrix[split=="Treinamento",],dados$price[split=="Treinamento"],alph
a=1)

predito_MMQ<- predict(vc_lasso,s=0, newx= dtm.matrix[split=="Teste",])

plot(dados$price[split == "Teste"],predito_MMQ)
```



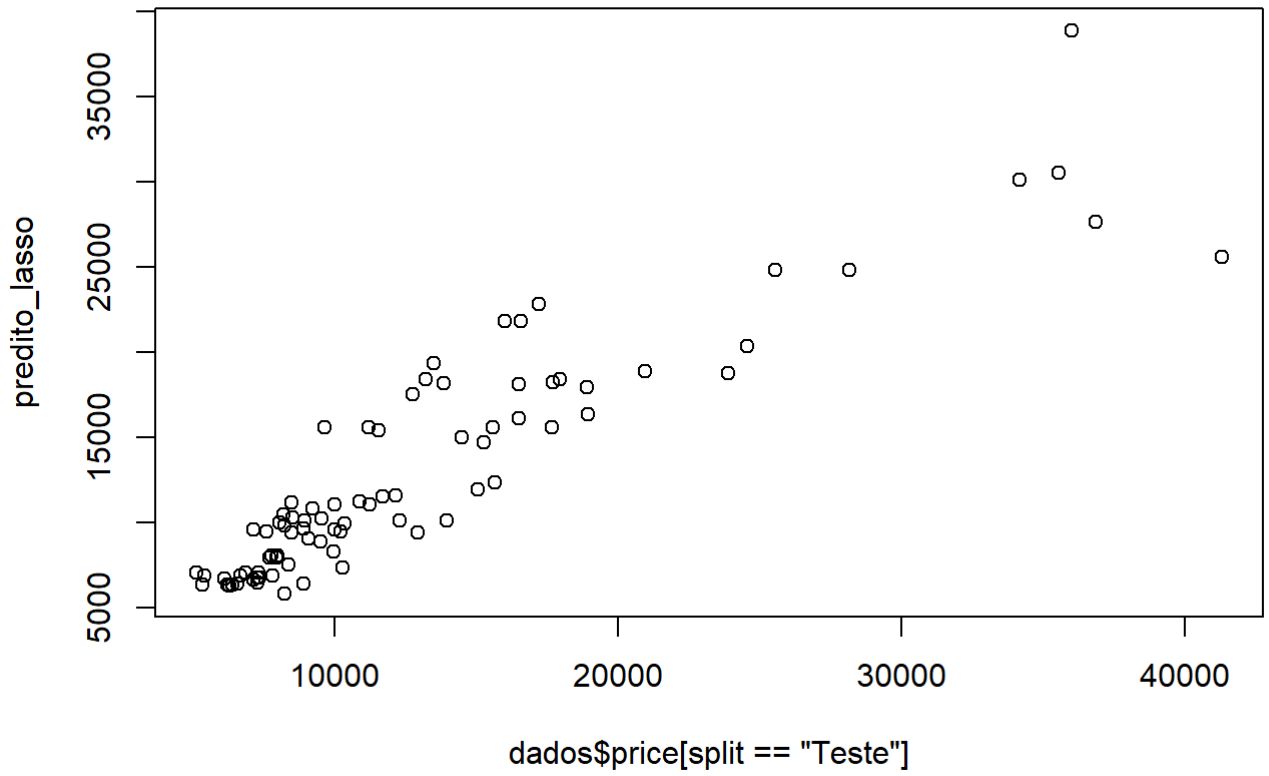
Ao analisarmos o gráfico de dispersão entre os valores reais do preço e os valores preditos pelo modelo de mínimos quadrados, observamos que os pontos no gráfico estão mais próximos de uma linha diagonal. Essa proximidade indica que o modelo possui uma boa capacidade de previsão, pois os valores preditos estão em concordância com os valores reais. Isso sugere que o modelo de mínimos quadrados está sendo eficaz na estimativa do preço com base nas variáveis consideradas.

Ajustando o Lasso:

```
plot(vc_lasso)
```




```
predito_lasso<- predict(vc_lasso,s=vc_lasso$lambda.min, newx= dtm.matrix[split=="Teste",])

plot(dados$price[split == "Teste"],predito_lasso)
```




```
vc_lasso$lambda.min
```

```
## [1] 288.0768
```

De forma similar ao método de mínimos quadrados os pontos estão muito próximos de uma linha diagonal. Essa semelhança nos gráficos de dispersão nos traz informação de que o Lasso está se comportando de maneira similar ao método de mínimos quadrados no que diz respeito à relação entre os valores preditos e os valores reais.

Melhor valor de lambda foi de 288.0768

O gráfico apresenta como o Erro Quadrático Médio varia em função do lambda escolhido. Esse lambda vai selecionar em torno de 34 variáveis. Esse gráfico nos mostra que quanto maior o lambda menos variáveis ele seleciona e quanto menor o lambda, mais variáveis serão selecionadas e mais próximo fica do método de mínimos quadrados.

Item C:

Estimativa dos coeficientes para o método Mínimos Quadrados:

```
table(coef(vc_lasso,s=0)[,1]!=0)
```

```
##
## FALSE  TRUE
##   116    75
```

```
coef_estimates<- coef(vc_lasso,s=0)
coefs<- coef_estimates %>% as.matrix %>% as_tibble
names(coefs)="Estimativa"
coefs %>% mutate(variavel=rownames(coef_estimates)) %>% arrange(desc(abs(Estimativa)))
```

```
## # A tibble: 191 × 2
##    Estimativa variavel
##         <dbl> <chr>
##  1    -24830. (Intercept)
##  2      9438. namebuick regal sport coupe (turbo)
##  3      9145. enginelocationrear
##  4      5665. namebuick skylark
##  5     -5580. nametoyota corona liftback
##  6      5562. namebuick skyhawk
##  7      5482. namebmw x4
##  8     -4901. namenissan dayz
##  9      4755. namebuick century special
## 10      4206. nameporsche macan
## # ℹ 181 more rows
```

Estimativa dos coeficientes para o método Lasso:

```
table(coef(vc_lasso,s=vc_lasso$lambda.min)[,1]!=0)
```

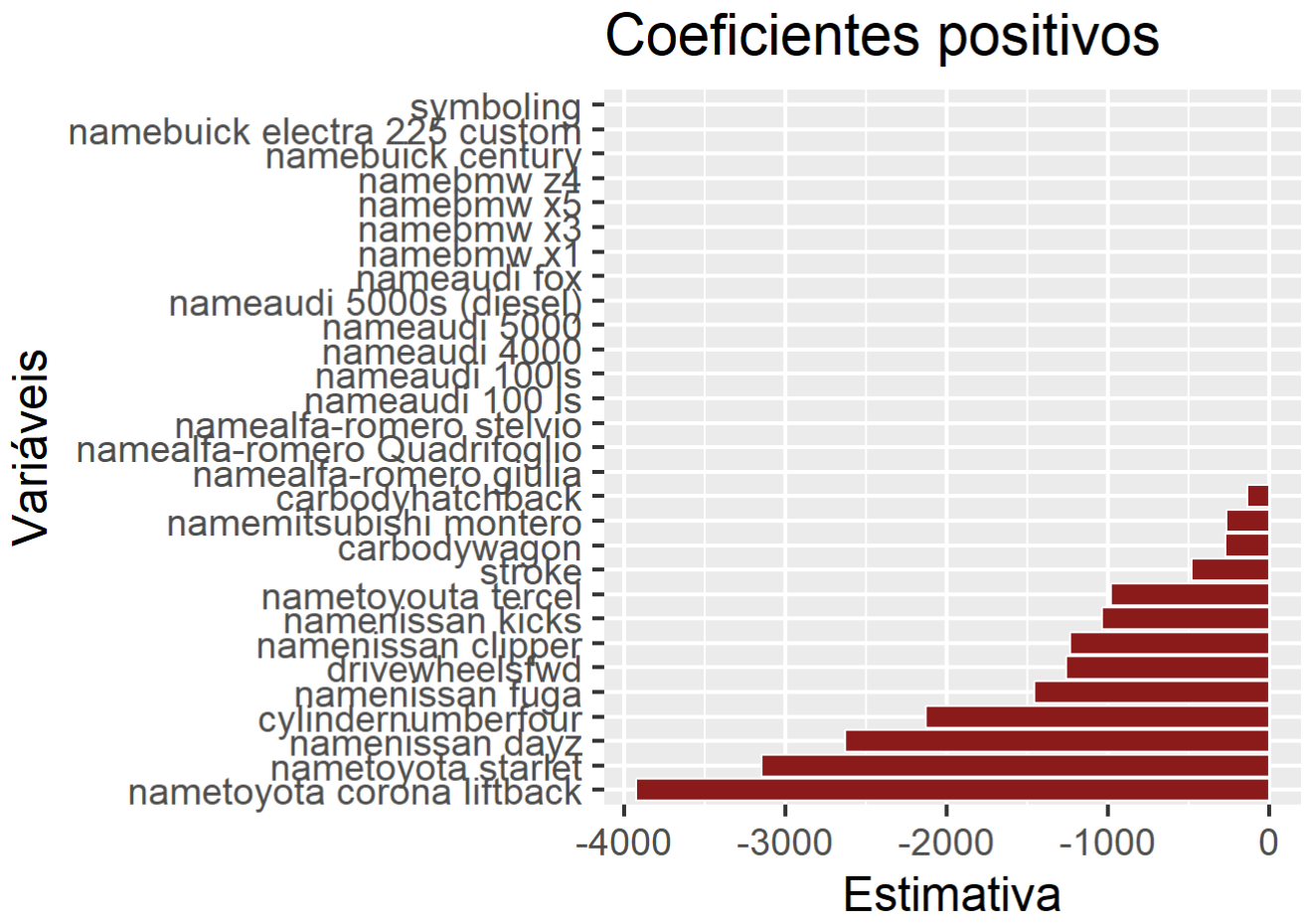
```
##
## FALSE  TRUE
##   156    35
```

```
coef_estimates<- coef(vc_lasso,s=vc_lasso$lambda.min)
coefs<- coef_estimates %>% as.matrix %>% as_tibble
names(coefs)="Estimativa"
coefs %>% mutate(variavel=rownames(coef_estimates)) %>% arrange(desc(abs(Estimativa)))
```

```
## # A tibble: 191 × 2
##    Estimativa variavel
##         <dbl> <chr>
##  1    -41261. (Intercept)
##  2      9291. enginelocationrear
##  3      5659. namebuick regal sport coupe (turbo)
##  4     -3927. nametoyota corona liftback
##  5     -3147. nametoyota starlet
##  6      2723. namebmw x4
##  7      2656. namebuick skyhawk
##  8     -2628. namenissan dayz
##  9      2605. namebuick skylark
## 10      2574. namebmw 320i
## # ℹ 181 more rows
```

Gráfico dos coeficientes para o método Lasso:

```
theme_set(theme_gray(base_size = 18))
coefs_estimates<- coef(vc_lasso,s=vc_lasso$lambda.min)
coefs_estimates = data.frame(Palavra=rownames(coefs_estimates),Coeficientes= coef_estimates[,
1])
coef_pos = coefs_estimates %>% arrange(desc(Coeficientes))
coef_neg = coefs_estimates %>% arrange(Coeficientes)

graf_pos = ggplot(data=coef_pos[2:30,],aes(x=reorder(Palavra,Coeficientes),y=Coeficientes))+
geom_bar(stat = "identity",col="white", fill="dodgerblue")+ coord_flip()+xlab("")+ ylab("Esti
mativa lasso")+ ggtitle("Coeficientes positivos")

graf_neg = ggplot(data=coef_neg[2:30,],aes(x=reorder(Palavra,Coeficientes),y=Coeficientes))+
geom_bar(stat = "identity",col="white", fill="firebrick4")+ coord_flip()+xlab("Variáveis")+ y
lab("Estimativa")+ ggtitle("Coeficientes positivos")

graf_neg
```




```
graf_pos
```

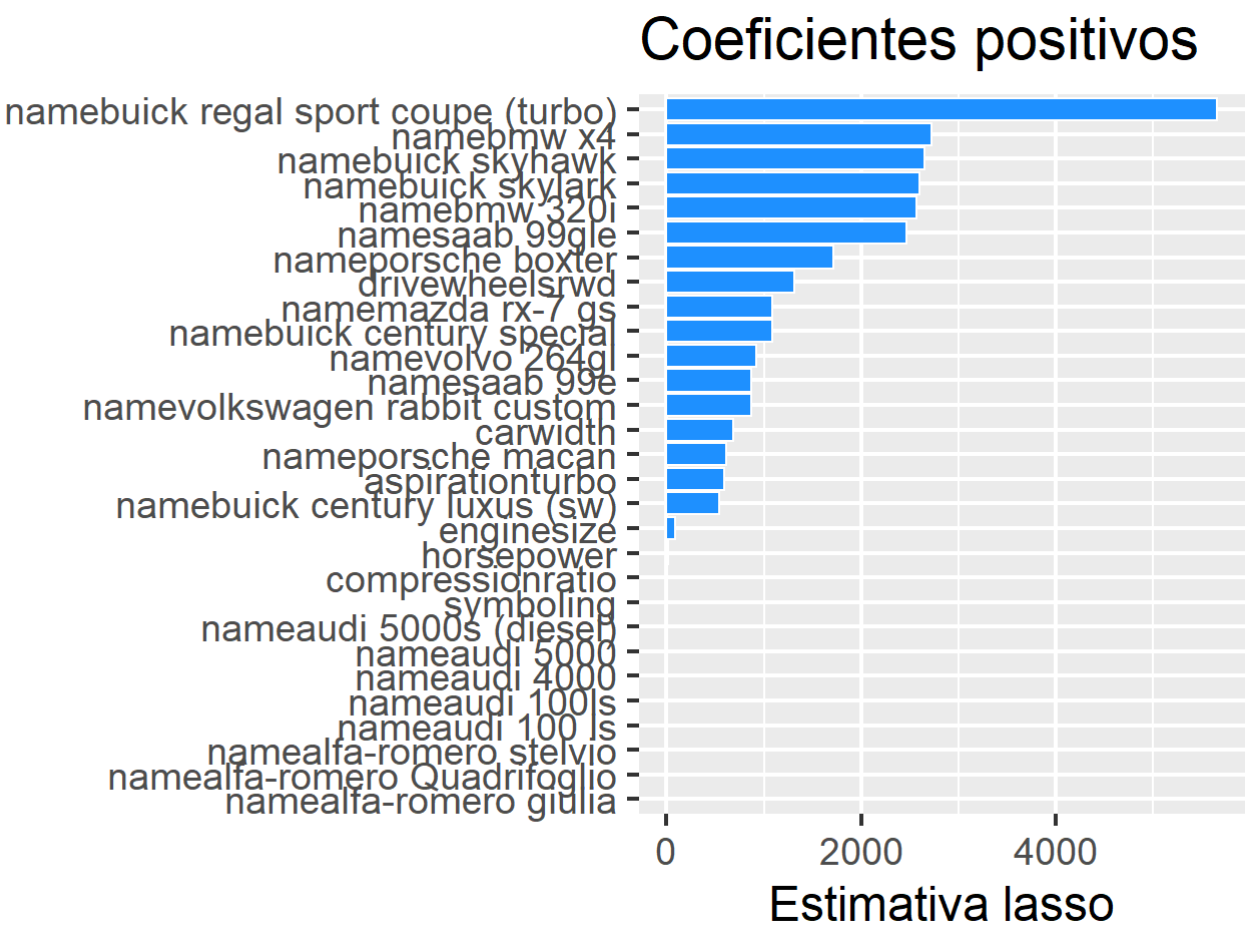
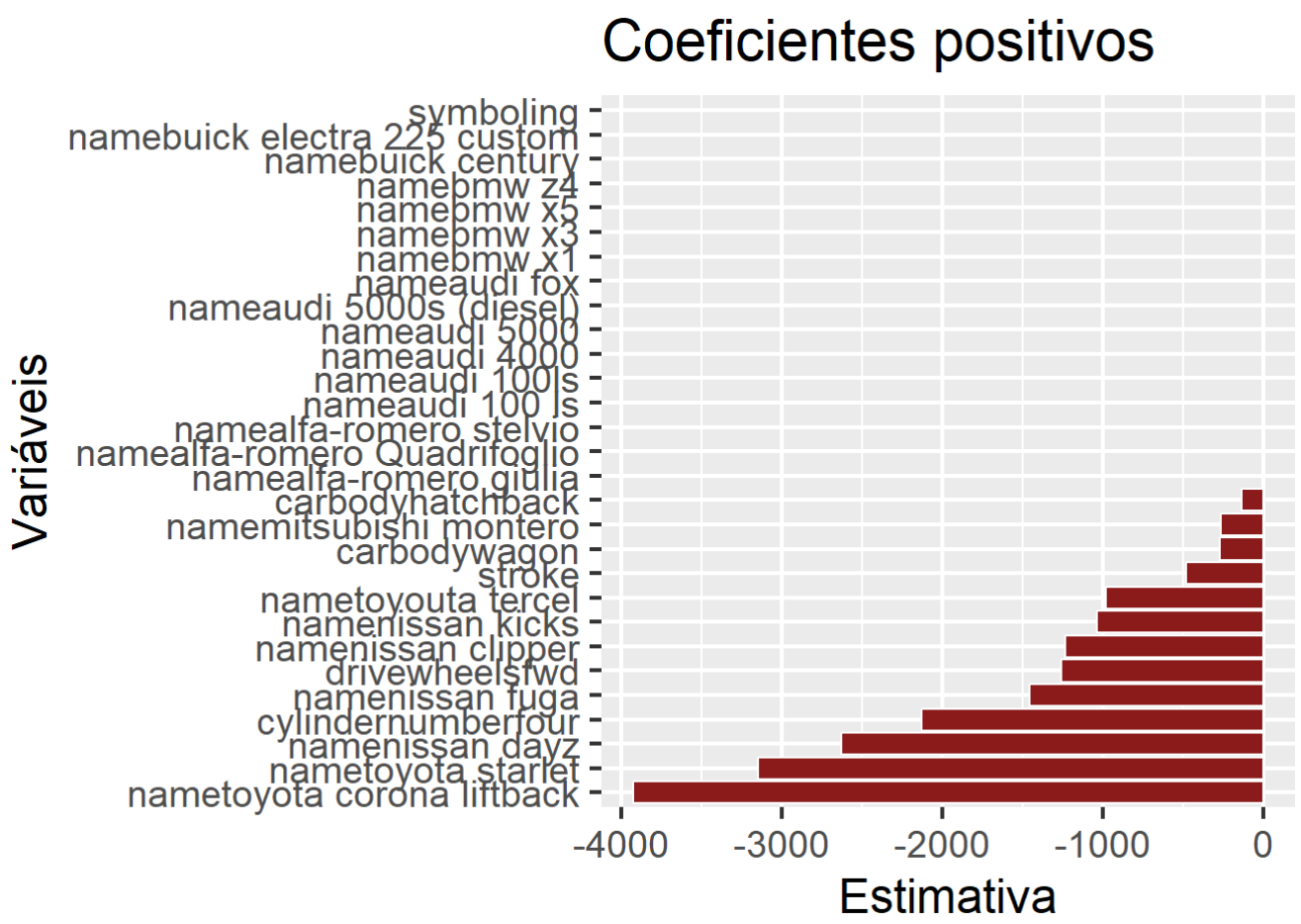
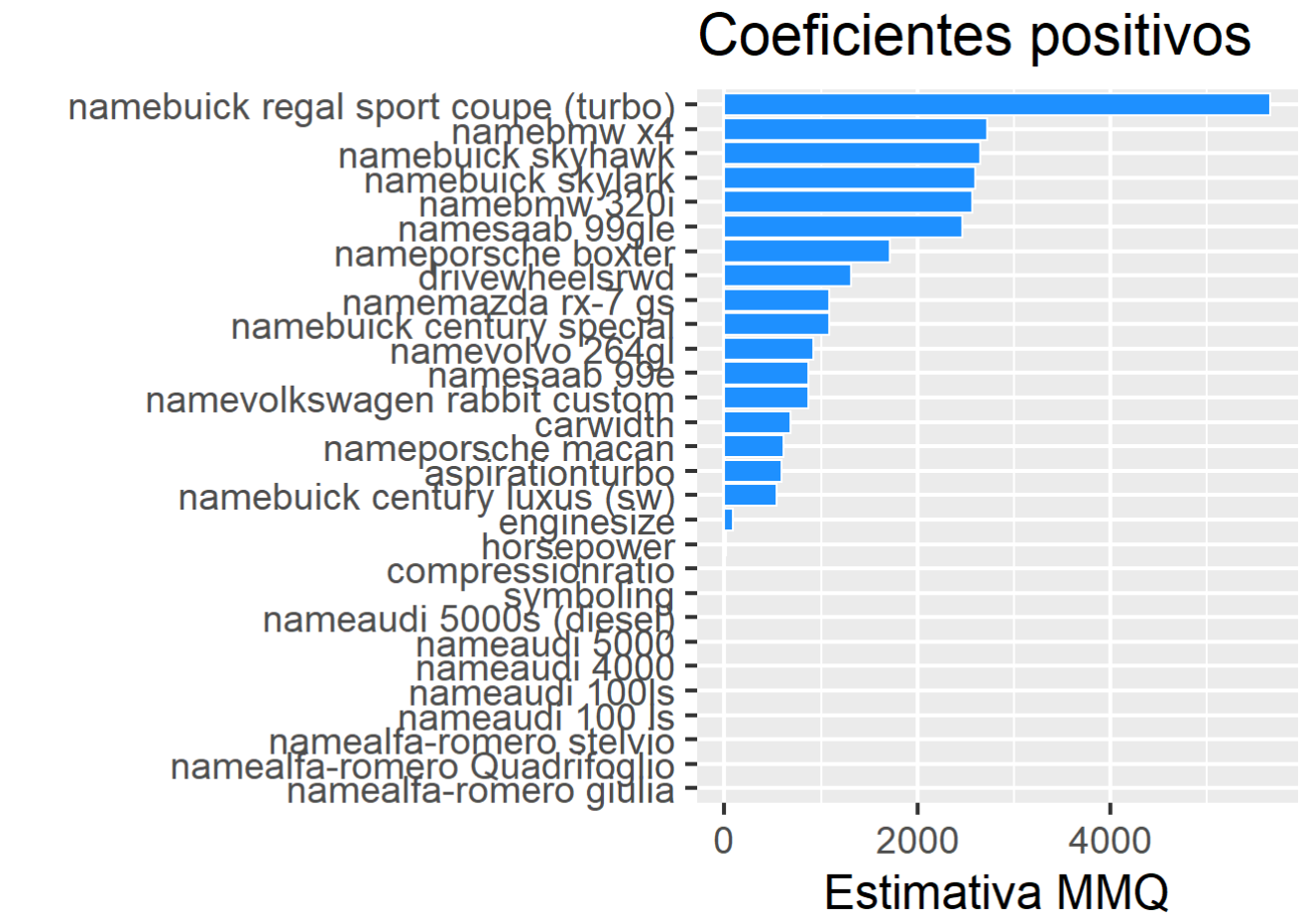


Gráfico dos coeficientes para o método Mínimos Quadrados:

```
theme_set(theme_gray(base_size = 18))
coefs_estimates<- coef(vc_lasso,s=0)
coefs_estimates = data.frame(Palavra=rownames(coefs_estimates),Coeficientes= coef_estimates[,
1])
coef_pos = coefs_estimates %>% arrange(desc(Coeficientes))
coef_neg = coefs_estimates %>% arrange(Coeficientes)

graf_pos = ggplot(data=coef_pos[2:30,],aes(x=reorder(Palavra,Coeficientes),y=Coeficientes))+
geom_bar(stat = "identity",col="white", fill="dodgerblue")+ coord_flip()+xlab("")+ ylab("Esti
mativa MMQ")+ ggtitle("Coeficientes positivos")

graf_neg = ggplot(data=coef_neg[2:30,],aes(x=reorder(Palavra,Coeficientes),y=Coeficientes))+
geom_bar(stat = "identity",col="white", fill="firebrick4")+ coord_flip()+xlab("Variáveis")+ y
lab("Estimativa")+ ggtitle("Coeficientes positivos")

graf_neg
```


Coeficientes positivos
symboling
namebuick electra 225 custom
namebuick century
namebmw z4
namebmw x5
namebmw x3
namebmw x1
nameaudi fox
nameaudi 5000s (diesel)
nameaudi 5000
nameaudi 4000
nameaudi 100ls
nameaudi 100 ls
namealfa-romero stelvio
namealfa-romero Quadrifoglio
namealfa-romero giulia
carbodyhatchback
namemitsubishi montero
carbodywagon
stroke
nametoyouta tercel
namenissan kicks
namenissan clipper
drivewheelsfwd
namenissan fuga
cylindernumberfour
namenissan dayz
nametoyota starlet
nametoyota corona liftback
Variáveis
-4000
-3000
-2000
-1000
0
Estimativa

```
graf_pos
```


Coeficientes positivos
namebuick regal sport coupe (turbo)
namebmw x4
namebuick skyhawk
namebuick skylark
namebmw 320i
namesaab 99gle
nameporsche boxter
drivewheelsrwd
namemazda rx-7 gs
namebuick century special
namevolvo 264gl
namesaab 99e
namevolkswagen rabbit custom
carwidth
nameporsche macan
aspirationturbo
namebuick century luxus (sw)
enginesize
horsepower
compressionratio
symboling
nameaudi 5000s (diesel)
nameaudi 5000
nameaudi 4000
nameaudi 100ls
nameaudi 100 ls
namealfa-romero stelvio
namealfa-romero Quadrifoglio
namealfa-romero giulia
0
2000
4000
Estimativa MMQ

As figuras das estimativas dos coeficientes mostram quais variáveis apresentam maior importância para a previsão e quais tem menor importância. Consideramos os valores positivos e negativos dos coeficientes estimados ordenados. Com base nessas estimativas, podemos concluir que não há diferença entre eles pois o lambda minimo é relativamente proximo de zero.

Item D:

Risco estimado para mínimos quadrados:

```
riscos$minimos.quadrados<- (predito_MMQ-dados$price[split=="Teste"])^2 %>% mean()
riscos$minimos.quadrados
```

```
## [1] 10238423
```

Risco estimado para o lasso:

```
riscos$lasso<-(predito_lasso-dados$price[split=="Teste"])^2 %>% mean()
riscos$lasso
```

```
## [1] 10248768
```

Conclusão para os riscos:

O Método de mínimos quadrados deu um erro preditivo menor no conjunto de teste, ou seja, o risco estimado está sendo menor para o Método de mínimos quadrados do que para o método Lasso. Uma explicação para o risco estar sendo menor para o método de mínimos quadrados é que o valor do parâmetro lambda do Lasso está muito próximo de zero. Quando lambda se aproxima de zero no Lasso, a penalização aplicada aos coeficientes se torna muito pequena, resultando em um modelo que se aproxima dos mínimos quadrados. Assim, é esperado que o método de mínimos quadrados apresente um desempenho semelhante ou melhor quando lambda está próximo de zero.

Intervalo de confiança para o risco estimado do método Lasso:

```
n_teste <- sum(split == "Teste")
residuos <- predito_lasso - dados$price[split == "Teste"]
desvio_padrao <- sqrt(sum(residuos^2) / (n_teste))

z_critico <- qnorm(0.975)
erro_padrao <- desvio_padrao / sqrt(n_teste)
intervalo_confianca <- c(riscos$lasso - z_critico * erro_padrao, riscos$lasso + z_critico * e
rro_padrao)
intervalo_confianca
```

```
## [1] 10248080 10249457
```

Intervalo de confiança para o risco estimado do método

Mínimos quadrados:

```
n_teste <- sum(split == "Teste")
residuos <- predito_MMQ - dados$price[split == "Teste"]
desvio_padrao <- sqrt(sum(residuos^2) / (n_teste))

z_critico <- qnorm(0.975)
erro_padrao <- desvio_padrao / sqrt(n_teste)
intervalo_confianca_mmq <- c(riscos$minimos.quadrados - z_critico * erro_padrao, riscos$minim
os.quadrados + z_critico * erro_padrao)
intervalo_confianca_mmq
```

```
## [1] 10237735 10239112
```

Com base nos intervalos de confiança apresentados para as estimativas dos riscos podemos observar que eles se sobrepõem. Sendo assim, não há evidência suficiente para concluir que existe uma diferença significante entre as estimativas de risco dos dois modelos. Portanto, pode-se dizer que os dois modelos têm um desempenho semelhante.

Item E:

Lasso:

```
conflicts_prefer(glmnetUtils::cv.glmnet)
library(glmnet)

library(glmnetUtils)

# Carregue seu banco de dados
dados <- read.csv("~/mineração de dados/atividade-1-precificacao-veiculos.csv")
ma<- as.matrix(dados)



# Armazene as covariáveis em uma matriz
X <- as.matrix(dados[, -ncol(dados)])  # Exclua a primeira coluna se for a variável resposta
X<- X[,-1]
dim(X)
```

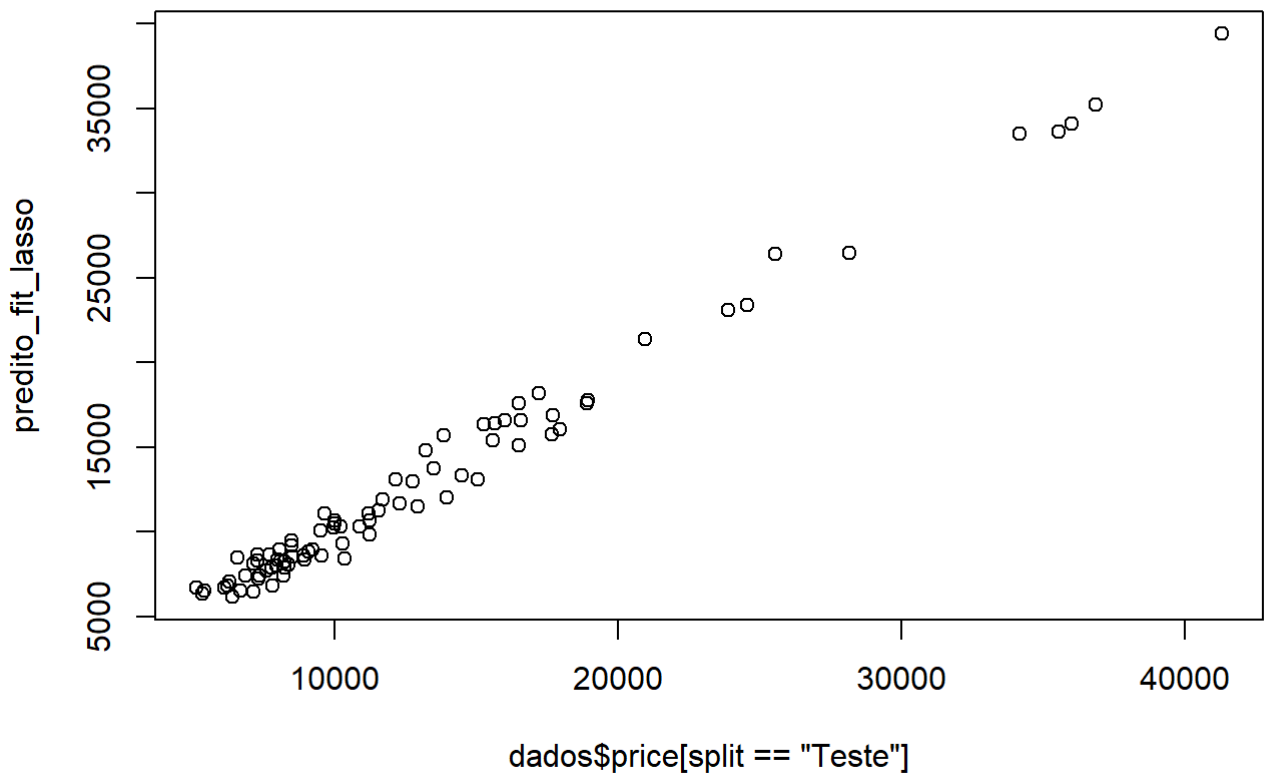
```
## [1] 205  24
```

```
# Armazene a variável resposta
y <- dados[, ncol(dados)]  # Substitua "1" pelo índice correto da coluna da variável resposta

# Adicione interações duas a duas entre as covariáveis
X_interactions <- model.matrix(~ .^2, data = data.frame(X))


# Ajuste o modelo com validação cruzada usando cv.glmnet do pacote glmnetUtils
fit <- cv.glmnet(X_interactions, y, alpha = 1)



# Exiba o valor de lambda ótimo selecionado pela validação cruzada
fit$lambda.min
```

```
## [1] 134.5761
```

Intervalo de confiança para as estimativas do risco para o método Lasso:

```
predito_fit_lasso<- predict(fit,s=fit$lambda.min, newx= X_interactions[split=="Teste",])

plot(dados$price[split == "Teste"],predito_fit_lasso)
```




```
fit$lambda.min
```

```
## [1] 134.5761
```

```
riscos$lasso_fit<-(predito_fit_lasso-dados$price[split=="Teste"])^2 %>% mean()
riscos$lasso_fit
```

```
## [1] 1055160
```

```
n_teste <- sum(split == "Teste")
residuos <- predito_fit_lasso - dados$price[split == "Teste"]
desvio_padrao <- sqrt(sum(residuos^2) / (n_teste))

z_critico <- qnorm(0.975)
erro_padrao <- desvio_padrao / sqrt(n_teste)
intervalo_confianca_lasso <- c(riscos$lasso_fit - z_critico * erro_padrao, riscos$lasso_fit +
z_critico * erro_padrao)
intervalo_confianca_lasso
```

```
## [1] 1054939 1055381
```

Mínimos quadrados:

```
library(glmnet)
library(glmnetUtils)

# Carregue seu banco de dados
dados <- read.csv("~/mineração de dados/atividade-1-precificacao-veiculos.csv")
ma<- as.matrix(dados)


# Armazene as covariáveis em uma matriz
X <- as.matrix(dados[, -ncol(dados)])  # Exclua a primeira coluna se for a variável resposta
X<- X[,-1]
dim(X)
```

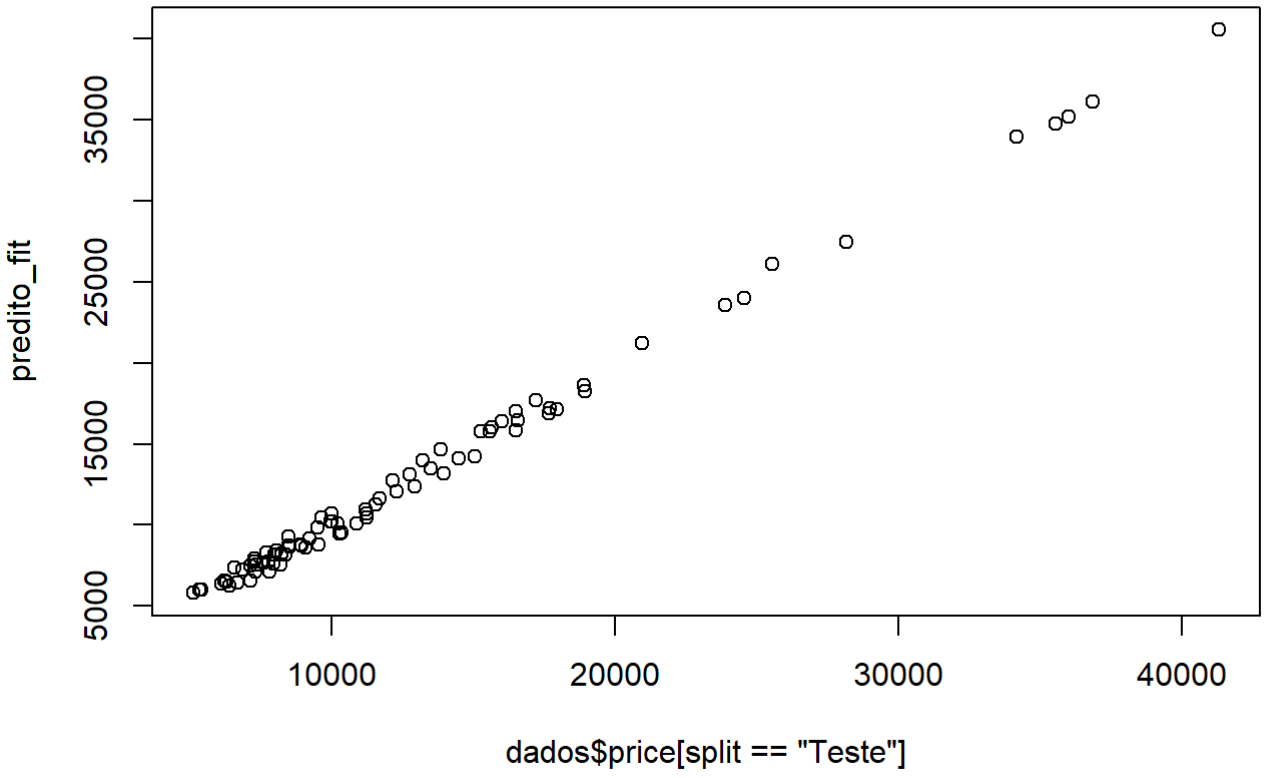
```
## [1] 205  24
```

```
# Armazene a variável resposta
y <- dados[, ncol(dados)]  # Substitua "1" pelo índice correto da coluna da variável resposta

# Adicione interações duas a duas entre as covariáveis
X_interactions <- model.matrix(~ .^2, data = data.frame(X))


# Ajuste o modelo com validação cruzada usando cv.glmnet do pacote glmnetUtils
fit <- cv.glmnet(X_interactions, y, alpha = 1)
```

Intervalo de confiança para as estimativas do risco para o

método MMQ:

```
predito_fit<- predict(fit,s=0, newx= X_interactions[split=="Teste",])

plot(dados$price[split == "Teste"],predito_fit)
```




```
fit$lambda.min
```

```
## [1] 134.5761
```

```
riscos$mmq_fit<-(predito_fit-dados$price[split=="Teste"])^2 %>% mean()
riscos$mmq_fit
```

```
## [1] 263703.3
```

```
n_teste <- sum(split == "Teste")
residuos <- predito_fit - dados$price[split == "Teste"]
desvio_padrao <- sqrt(sum(residuos^2) / (n_teste))

z_critico <- qnorm(0.975)
erro_padrao <- desvio_padrao / sqrt(n_teste)
intervalo_confianca_mmq <- c(riscos$mmq_fit - z_critico * erro_padrao, riscos$mmq_fit + z_cri
tico * erro_padrao)
intervalo_confianca_mmq
```

```
## [1] 263592.9 263813.8
```

Conclusão:

Para que os intervalos se sobreponham, é necessário que haja interseção entre eles. Isso ocorre quando o valor máximo de um intervalo é maior ou igual ao valor mínimo do outro intervalo, e vice-versa. No caso em questão, o valor máximo do primeiro intervalo (56462.11) é menor do que o valor mínimo do segundo intervalo (263592.9), o que significa que não há sobreposição entre os intervalos. Portanto, podemos concluir que esses intervalos não se sobrepõem. Assim, com as interações o Lasso foi melhor. Introduzir interações no processo de avaliação dos intervalos trouxe benefícios significativos, reduzindo o risco associado ao modelo e proporcionando uma estimativa mais precisa. Comparado ao modelo sem interações, a inclusão das interações levou a um intervalo de confiança menor para a estimativa.