Para a análise do conjunto de dados proposto foi utilizado o software R, por meio das seguintes bibliotecas:

```
library(readr)
library(tidyverse)

library(visdat)

library(naniar)

library(rpart)
library(caret)

library(fastDummies)

require(lmtest)
```

Carregando os dados disponibilizados no Kaggle.

```
original = read_csv("C:/Users/Juliana/Desktop/Icarros/vehicles.csv")
```

Para facilitar a visualização dos dados e a própria análise, algumas colunas que não apresentavam importância foram retiradas.

```
dados = subset(original,select=-
c(X1,id,url,region_url,VIN,image_url,lat,long,posting_date,description))
attach(dados)
```

A partir da primeira visualização dos dados é possível notar a presença de muitos valores faltantes, estes valores podem ser substituídos ou excluídos do banco de dados, a análise deve continuar para tomar essa decisão.

```
sapply(dados, function(x) (sum(is.na(x))*100/length(dados$region)))
```

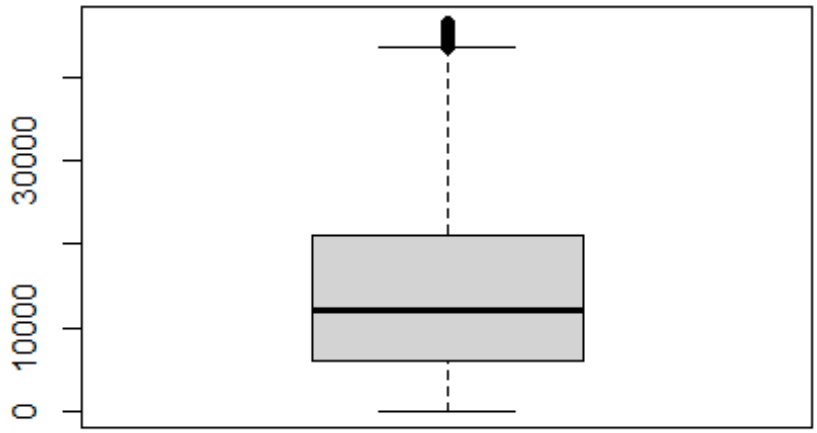
```
##       region        price         year manufacturer        model
condition
##    0.0000000    0.0000000    0.2291511    3.9763167    1.0575868
42.1070550
##    cylinders         fuel     odometer title_status transmission
drive
##   37.3494423    0.7064400   12.0692778    0.5624022    0.5329399
29.2850705
##         size         type  paint_color        state
##   70.1307034   24.6038414   30.7374518    0.0000000
```

Para as linhas que apresentam mais de 45% de observações faltantes, é interessante excluirmos, pois estimar mais da metade das variáveis de uma observação pode trazer complicações para o modelo.

```
dados_sem_NA_45 = data.frame(0)
vetor = c()
for (i in 1:length(dados$region)) {
  if ((sum(is.na(dados[i,]))/length(dados))>=0.45){
    vetor=append(vetor,i)
    dados_sem_NA_45=dados[-c(vetor),]
  }
}
```

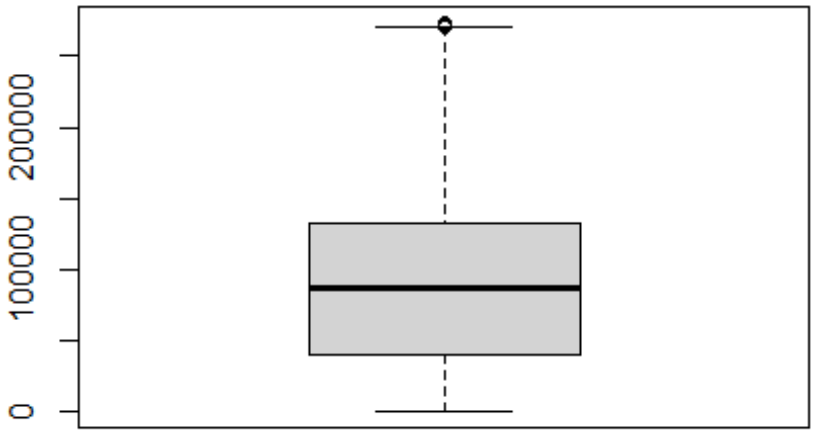
Verificando o boxplot para a variável “preço”, nota-se a presença de alguns outliers, que após uma analise individual aparentavam ser erros de digitação, ou opções para não deixar a parcela em branco (Ex: 123456789), portanto foram retiradas algumas observações que não estavam corretas, como as que apresentavam preços iguais a zero.

```
dados_sem_NA_45_2 = dados_sem_NA_45
q1 = quantile(dados_sem_NA_45$price,0.25)
q3 = quantile(dados_sem_NA_45$price,0.75)
iq = q3-q1
lim_inf = q1-1.5*iq
lim_sup = q3+1.5*iq
out = (dados_sem_NA_45_2$price>lim_sup)|(dados_sem_NA_45_2$price<=0)
dados_sem_NA_45_2$price[out] = NA
boxplot(dados_sem_NA_45_2$price)
```



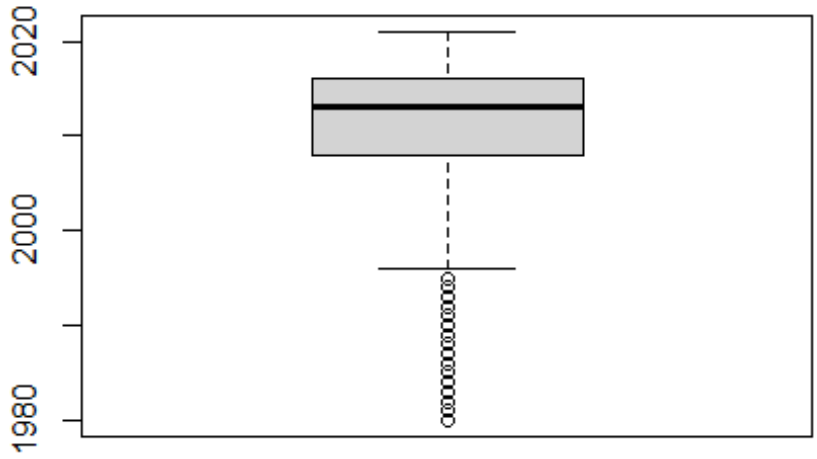
Verificando o boxplot para a variável “odômetro”, nota-se a presença de alguns outliers que após uma analise individual aparentavam estar incorretos, portanto foram retirados.

```
q1o = quantile(dados_sem_NA_45$odometer,0.25,na.rm = T)
q3o = quantile(dados_sem_NA_45$odometer,0.75,na.rm = T)
iqo = q3o-q1o
lim_info = q1o-1.5*iqo
lim_supo = q3o+1.5*iqo
outo =
(dados_sem_NA_45_2$odometer>lim_supo)|(dados_sem_NA_45_2$odometer<lim_inf
o)
dados_sem_NA_45_2$odometer[outo] = NA
boxplot(dados_sem_NA_45_2$odometer)
```



Assim como para as variáveis “preço” e “odômetro”, a coluna “ano” também apresentava dados que estavam incorretos e atrapalhava a análise, estes também foram retirados.

```
q1a = quantile(dados_sem_NA_45$year,0.25,na.rm = T)
q3a = quantile(dados_sem_NA_45$year,0.75,na.rm = T)
iqa = q3a-q1a
lim_infa = q1a-1.5*iqa
outa = (dados_sem_NA_45_2$year<1980)
dados_sem_NA_45_2$year[outa] = NA
boxplot(dados_sem_NA_45_2$year)
```



Após o estudo individual das três variáveis anteriores, a base de dados foi tratada, retirando todos os valores que não eram condizentes com o restante.

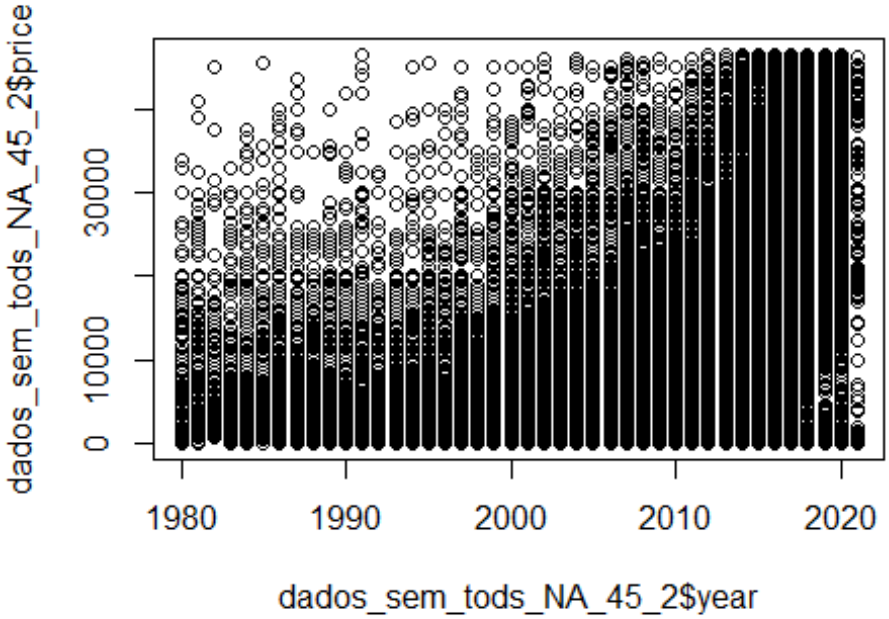
```
dados_sem_tods_NA_45_2 =
dados_sem_NA_45_2[!is.na(dados_sem_NA_45_2$price),]
dados_sem_tods_NA_45_2 =
dados_sem_tods_NA_45_2[!is.na(dados_sem_tods_NA_45_2$year),]
sapply(dados_sem_tods_NA_45_2, function(x) (sum(is.na(x))))
```

```
##       region        price         year manufacturer        model
condition
##            0            0            0        11297         2871
156922
##    cylinders         fuel     odometer title_status transmission
drive
##       143311         2031        45769         1343         1432
114231
##         size         type  paint_color        state
##       279818        96618       118683            0
```

Para escolher as variáveis que são significantes para a criação do modelo, é necessário realizar uma analise da correlação individual com a variável reposta.

Preço x Ano

```
plot(dados_sem_tods_NA_45_2$year,dados_sem_tods_NA_45_2$price)
```



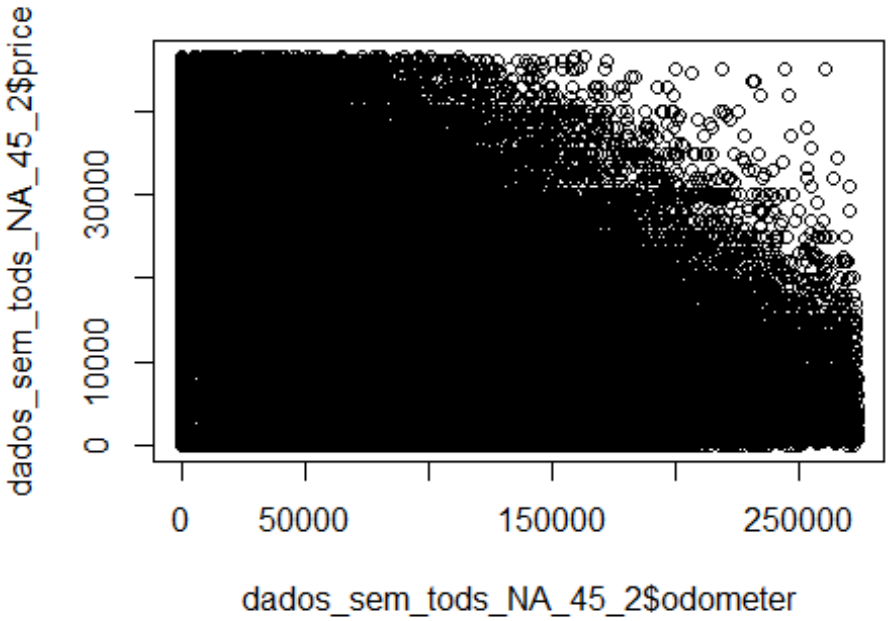
Nota-se que quanto maior o ano, maior é o preço de cada carro, indicando que há uma relação entre as variáveis, o que pode ser comprovado pelo coeficiente de correlação calculado a seguir.

```
cor(dados_sem_tods_NA_45_2$year,dados_sem_tods_NA_45_2$price)
```

```
## [1] 0.5406958
```

Preço x odômetro

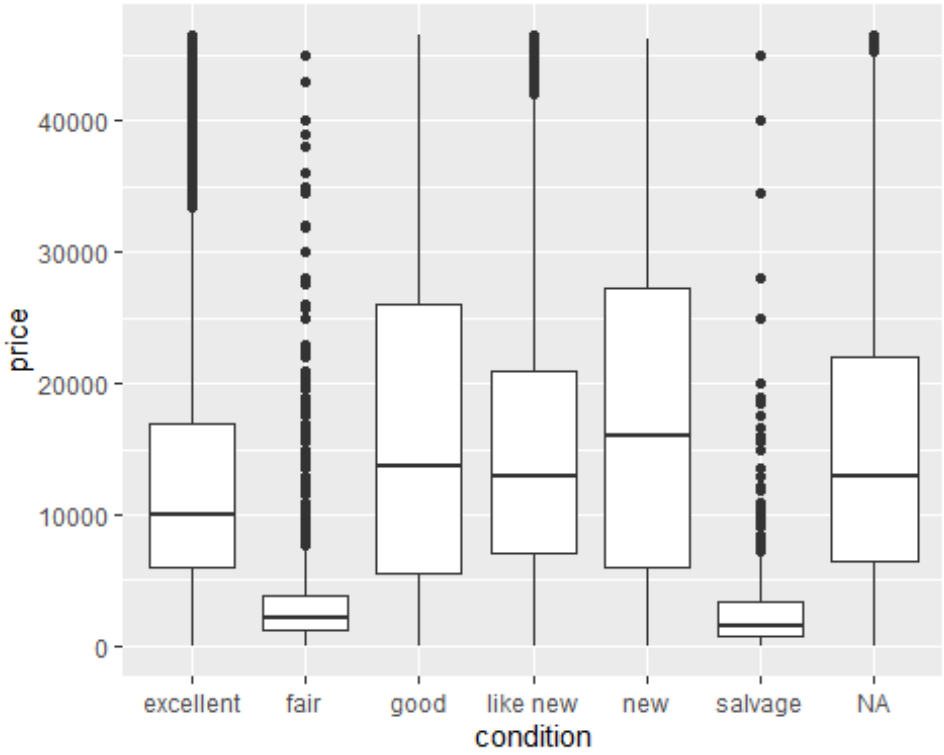
```
plot(dados_sem_tods_NA_45_2$odometer,dados_sem_tods_NA_45_2$price)
```



É possível notar pelo gráfico que quanto maior a distância percorrida pelo carro, menor é seu preço, indicando que há uma relação entre as variáveis “preço” e “odômetro”.

Condição x preço

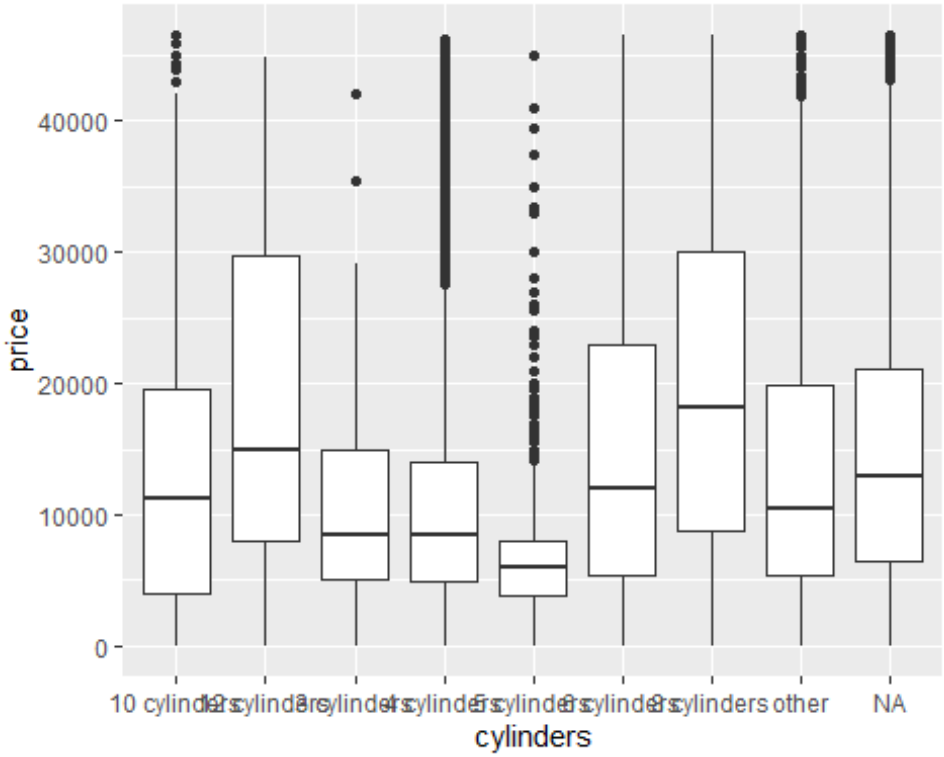
```
ggplot(dados_sem_tods_NA_45_2,aes(x=condition,y=price))+
  geom_boxplot()
```



Nota-se pela figura que os carros que apresentam melhores condições, também apresentam maiores preços.

Cilindros x preço

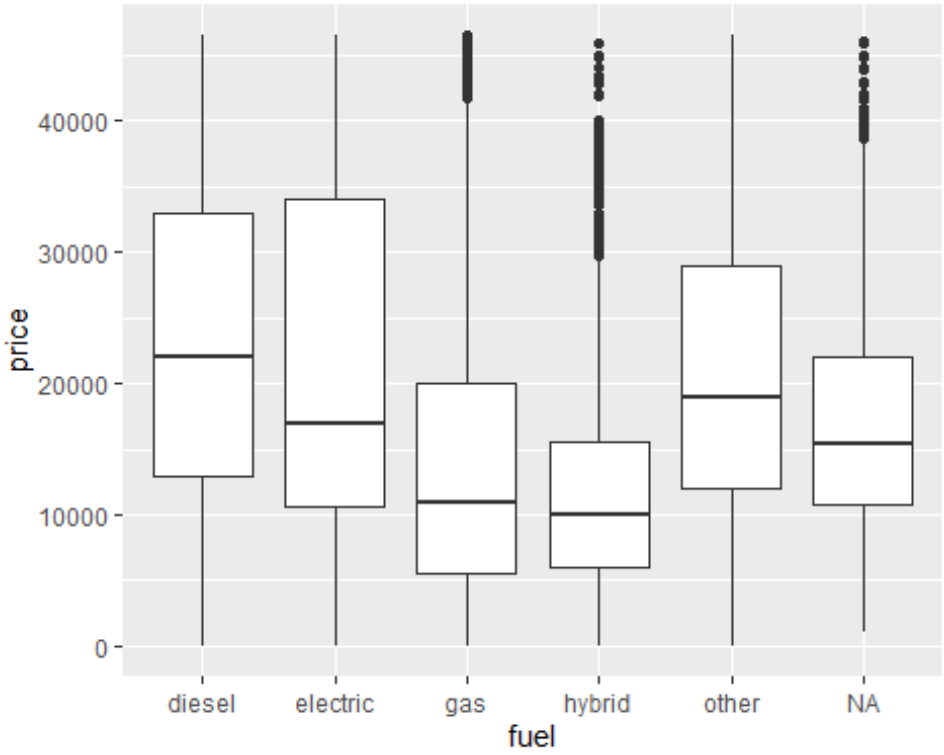
```
ggplot(dados_sem_tods_NA_45_2,aes(x=cylinders,y=price))+
  geom_boxplot()
```



Nota-se pela figura que o número de cilindros do carro não apresenta relação com o preço, pois as médias estão próximas.

Combustível x preço

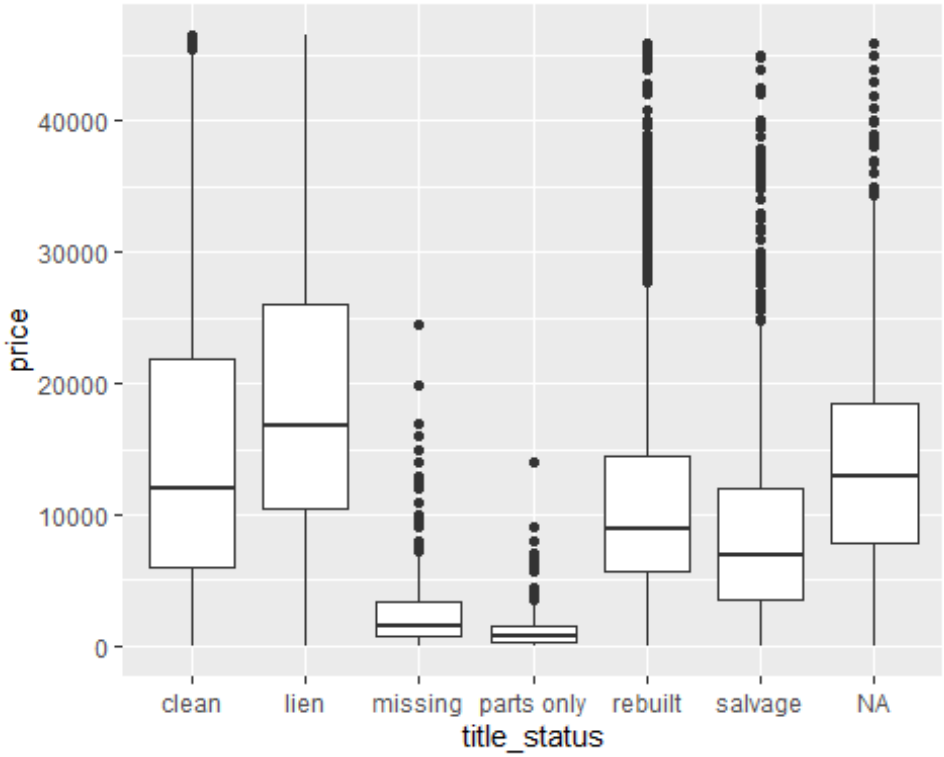
```
ggplot(dados_sem_tods_NA_45_2,aes(x=fuel,y=price))+
  geom_boxplot()
```



Nota-se pela figura que os carros que são abastecidos com diesel apresentam maiores preços, porém não há diferença significativa entre os outros métodos.

Status x preço

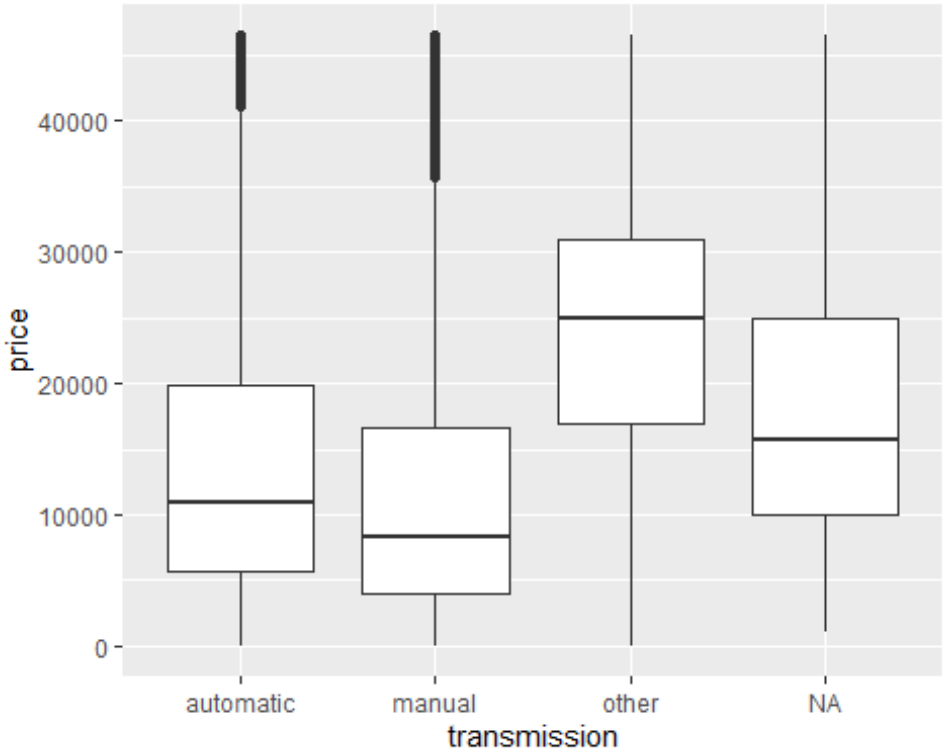
```
ggplot(dados_sem_tods_NA_45_2,aes(x=title_status,y=price))+
  geom_boxplot()
```



Nota-se pela figura que os carros que apresentam melhores condições, também apresentam maiores preços, indicando que há relação entre as condições do carro e seu preço.

Transmissão x preço

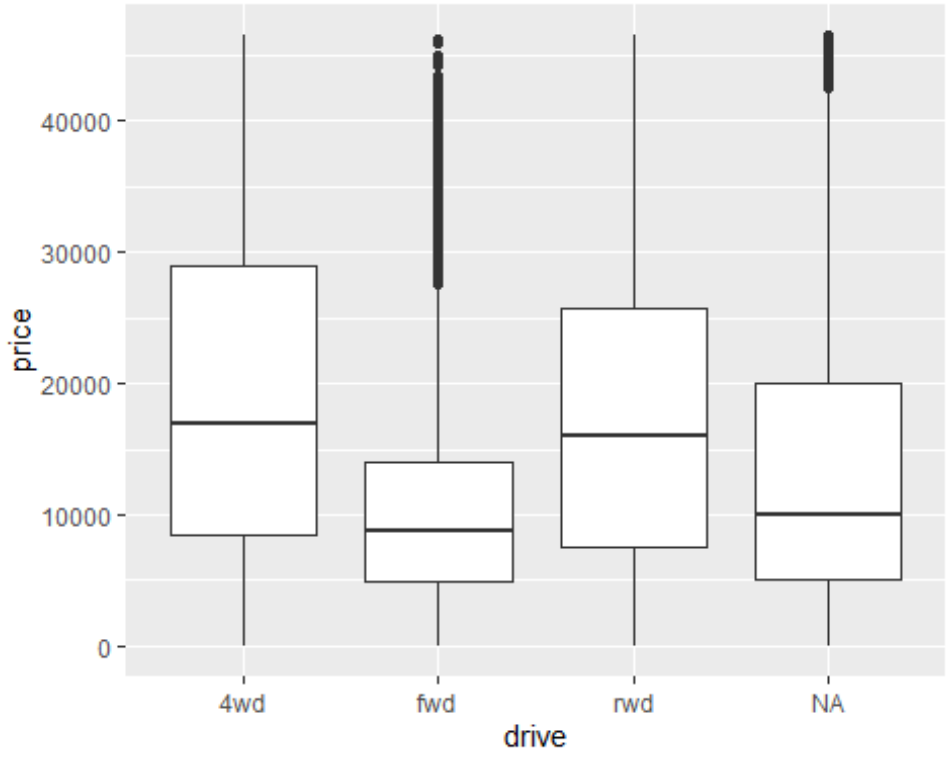
```
ggplot(dados_sem_tods_NA_45_2,aes(x=transmission,y=price))+
  geom_boxplot()
```



Nota-se pela figura que os carros com câmbio manual e automático não diferem significativamente.

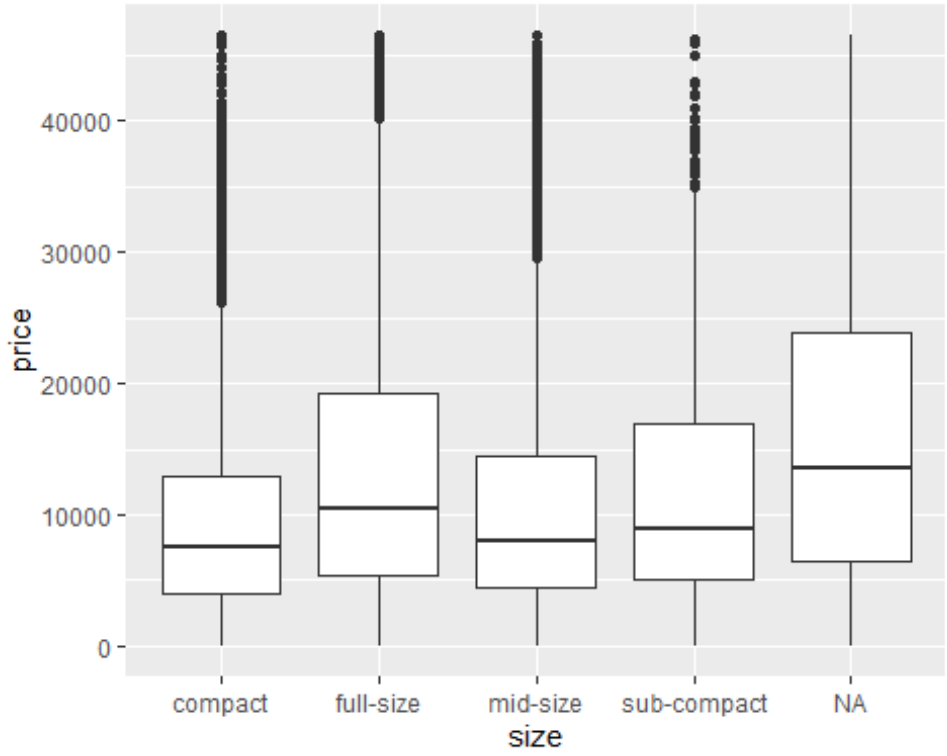
Direção x preço

```
ggplot(dados_sem_tods_NA_45_2,aes(x=drive,y=price))+
  geom_boxplot()
```



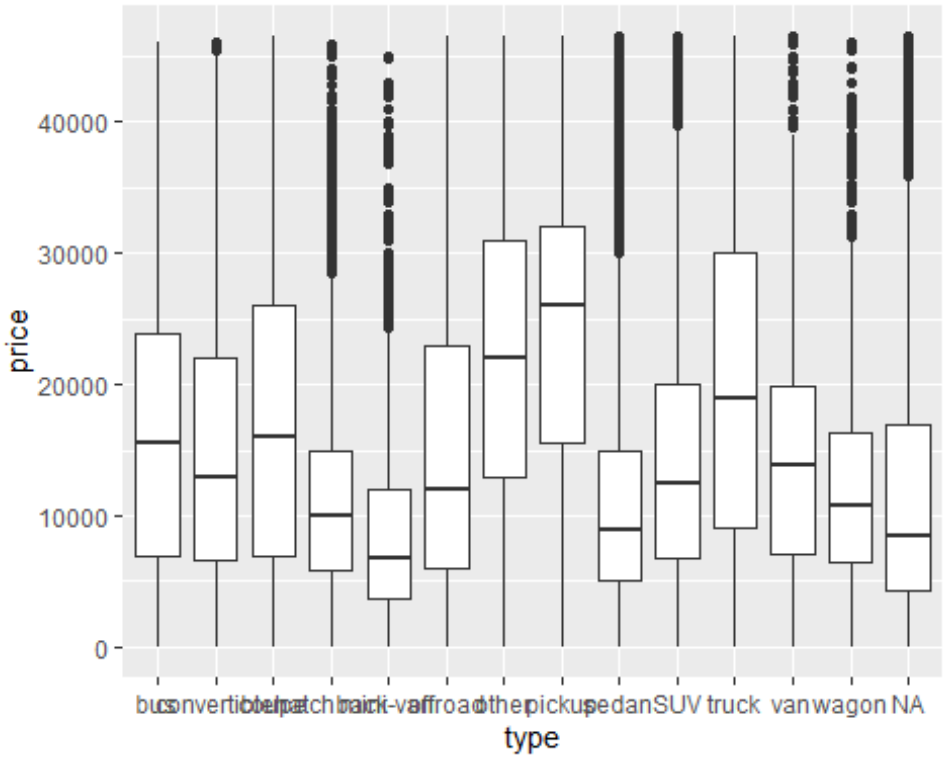
Tamanho x preço

```
ggplot(dados_sem_tods_NA_45_2,aes(x=size,y=price))+
  geom_boxplot()
```



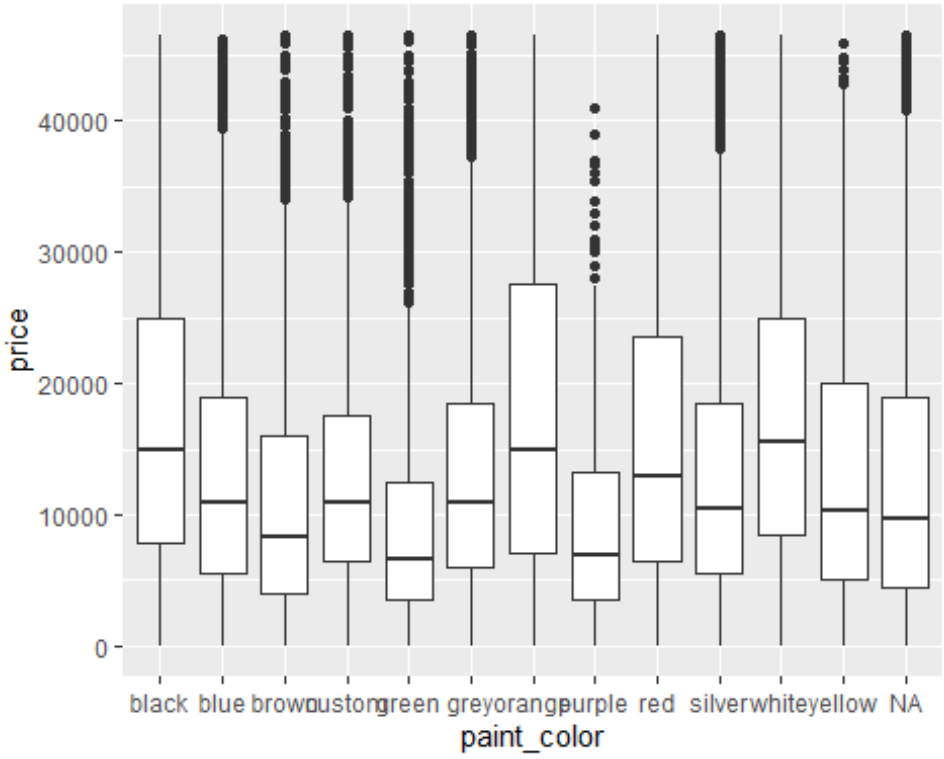
Tipo x preço

```
ggplot(dados_sem_tods_NA_45_2,aes(x=type,y=price))+
  geom_boxplot()
```



Cor x preço

```
ggplot(dados_sem_tods_NA_45_2,aes(x=paint_color,y=price))+
  geom_boxplot()
```



Notam-se pelas ultimas figuras que a cor dos carros, o tipo, o tamanho e a direção não apresentam relação com o preço.

Após a análise gráfica e análise da porcentagem de dados faltantes, optamos pela retirada de algumas variáveis da base de dados, como o tamanho, a cor, tipo, direção, modelo, fabricação, estado, região e transmissão. `dados_sem_tods_NA_45_2=subset(dados_sem_tods_NA_45_2,select=-c(size,paint_color,type,drive,model,manufacturer,state,region,transmission))`

Para solucionar o problema de valores faltantes na variável “odômetro”, que apresentou ser importante para o modelo, os valores “NA” foram substituídos pela mediana.

```
mean(dados_sem_tods_NA_45_2$odometer,na.rm = T)
## [1] 94830.19
median(dados_sem_tods_NA_45_2$odometer,na.rm = T)
```

```
## [1] 91500

sd(dados_sem_tods_NA_45_2$odometer,na.rm = T)

## [1] 59404.27

dados_sem_tods_NA_45_2$odometer[is.na(dados_sem_tods_NA_45_2$odometer)] =
median(dados_sem_tods_NA_45_2$odometer,na.rm=TRUE)
```

Como as colunas “combustível” e “title_status”, são importantes para o modelo e apresentam poucos valores nulos, podemos retirar essas parcelas, a fim de melhorar a precisão dos resultados.

```
dados_sem_tods_NA_45_2 =
dados_sem_tods_NA_45_2[!is.na(dados_sem_tods_NA_45_2$fuel),]
dados_sem_tods_NA_45_2 =
dados_sem_tods_NA_45_2[!is.na(dados_sem_tods_NA_45_2$title_status),]
```

Será aplicado o modelo de regressão linear múltipla, para isso, devemos substituir as variáveis não numéricas por variáveis dummies.

```
var_dummy =
dummy_cols(dados_sem_tods_NA_45_2,select_columns=c("fuel","title_status")
,remove_first_dummy = TRUE)
dados_modelo = subset(var_dummy,select=-
c(condition,cylinders,fuel,title_status))
```

Agora que a base de dados já esta limpa e tratada, podemos separar em conjunto de teste e treino para a criação do modelo, uma amostra aleatória de 70% foi para treino e 30% para teste.

```
set.seed(181254247)
linhas =
sample(1:length(dados_modelo$price),length(dados_modelo$price)*0.7)

#70% treino
treino = dados_modelo[linhas,]
#30% teste
teste = dados_modelo[-linhas,]
```

Aplicando o modelo.

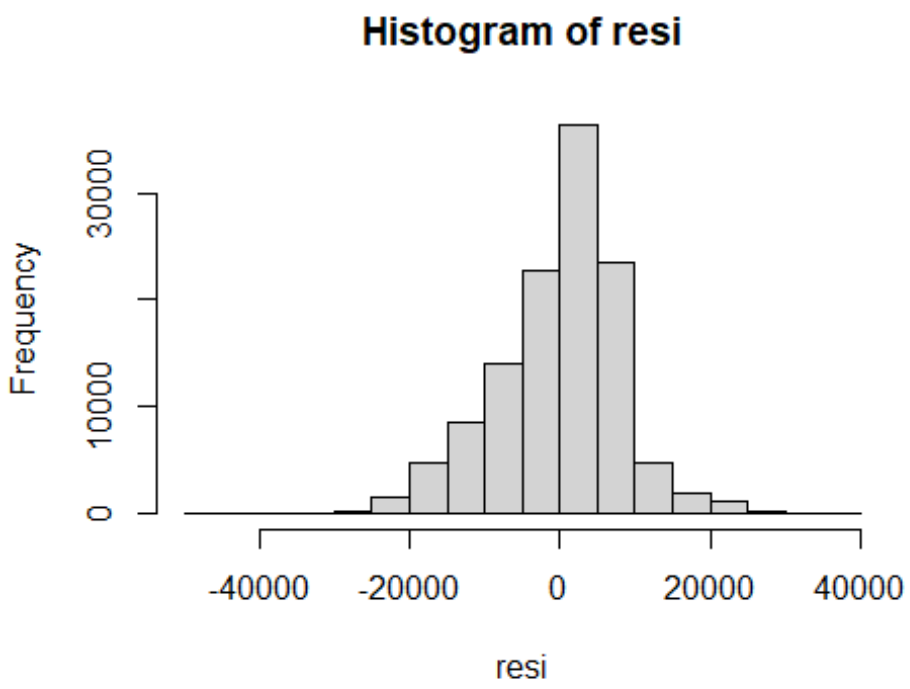
```
fm = lm(price~.,data = treino)
summary(fm)

##
## Call:
## lm(formula = price ~ ., data = treino)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -37915  -5234  -1218   4739  49615
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                -1.376e+06  6.002e+03 -229.287  < 2e-16 ***
## year                        7.001e+02  2.977e+00  235.172  < 2e-16 ***
## odometer                   -6.253e-02  3.271e-04 -191.175  < 2e-16 ***
## fuel_electric              -1.131e+04  2.288e+02  -49.447  < 2e-16 ***
## fuel_gas                   -1.198e+04  6.905e+01 -173.524  < 2e-16 ***
## fuel_hybrid                -1.494e+04  1.513e+02  -98.728  < 2e-16 ***
## fuel_other                 -8.898e+03  1.120e+02  -79.481  < 2e-16 ***
## title_status_lien           1.497e+03  2.251e+02    6.650 2.93e-11 ***
## title_status_missing       -2.741e+03  5.037e+02   -5.443 5.26e-08 ***
## `title_status_parts only` -7.662e+03  8.076e+02   -9.488  < 2e-16 ***
## title_status_rebuilt       -4.982e+03  1.094e+02  -45.557  < 2e-16 ***
## title_status_salvage       -4.671e+03  1.561e+02  -29.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8271 on 278742 degrees of freedom
## Multiple R-squared:  0.4287, Adjusted R-squared:  0.4286
## F-statistic: 1.901e+04 on 11 and 278742 DF,  p-value: < 2.2e-16
```

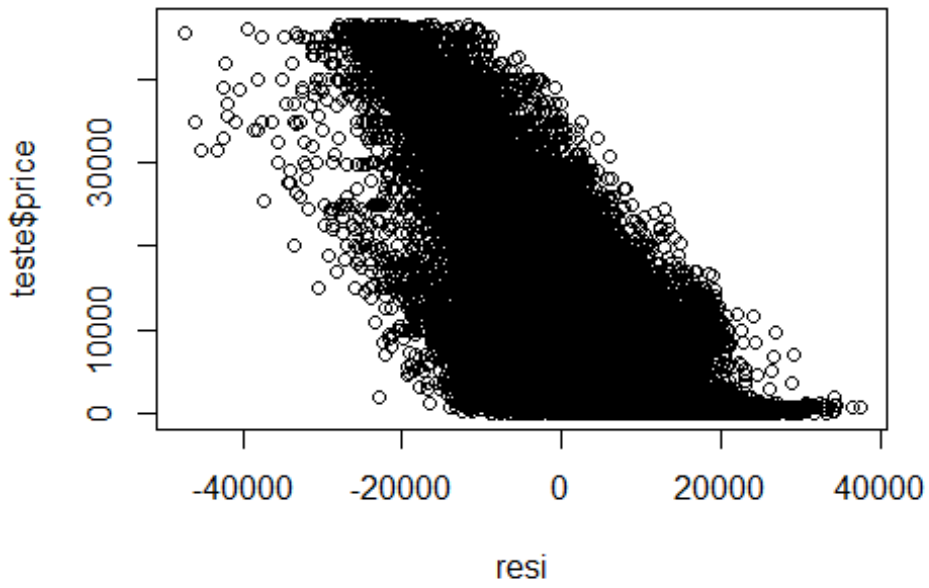
Notamos pelos valores do p-valor que foram fornecidos pela tabela anterior, que todas as variáveis são significantes para o modelo.

Os resíduos apresentam aproximadamente distribuição normal, como é possível visualizar pelo histograma a seguir.

```
teste$previsao = predict(fm,teste)
resi = teste$previsao-teste$price
hist(resi)
```




```r
plot(resi,teste$price)
```




```r
dwtest(lm(price~.,data=treino))
```

```
##
##  Durbin-Watson test
##
## data:  lm(price ~ ., data = treino)
## DW = 1.9952, p-value = 0.102
## alternative hypothesis: true autocorrelation is greater than 0
```

Pelo teste de Durbin-Watson, é possível verificar se os resíduos são autocorrelacionados, pelos resultados obtidos, podemos concluir que a correlação é igual à zero, portanto não há relação entre os valores dos resíduos.

R-quadrado

A porcentagem de variação na resposta que é explicada pelo modelo é calculada pelo r-quadrado, logo para o modelo em questão, temos que 57% da porcentagem de variação dos preços dos carros é explicado pelo modelo.

```
R2 =
((1.9068e+13)/((9.7566e+12)+(2.1474e+12)+(1.9505e+08)+(1.3190e+12)+(4.440
6e+11)+(4.2699e+11)+(3.7155e+09)+(1.8077e+09)+(5.9422e+09)+(1.3921e+11)+(
6.1256e+10)+(1.9068e+13)))

R2

[1] 0.5713399
```