Data Cleaning

Karine Almeida

2023-07-19

Business Understanding

A fictitious car buying and selling company is having difficulties reselling used cars in its catalogue. With the aim of pricing your catalog more competi- tively and thus recovering the poor performance in this sector, I will analyze the data to answer business questions and create a predictive model that prices the customer's cars so that they are as close to market values. **In this notebook you will have access to a descriptive analysis of the data, insights and answers to some business questions.**

Data cleaning

Installing packages

```
pacotes <- c('tidyverse','knitr','kableExtra', 'ggplot2', "paletteer",
             "scales","DT", "kableExtra", 'gridExtra', 'xlsx')

options(rgl.debug = TRUE)

if(sum(as.numeric(!pacotes %in% installed.packages())) != 0){
  instalador <- pacotes[!pacotes %in% installed.packages()]
  for(i in 1:length(instalador)) {
    install.packages(instalador, dependencies = T)
    break()}
  sapply(pacotes, require, character = T)
} else {
  sapply(pacotes, require, character = T)
}
```

```
##  tidyverse      knitr kableExtra    ggplot2  paletteer     scales         DT
##       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE
## kableExtra  gridExtra       xlsx
##       TRUE       TRUE      FALSE
```

Viewing the database

```
cars_train <- read_delim("../data/cars_train.txt", show_col_types = FALSE)
glimpse(cars_train)
```

```
## Rows: 29,584
## Columns: 29
## $ id                       <dbl> 3.007162e+38, 2.796398e+38, 5.641446e+37, 5.68~
## $ num_fotos                <dbl> 8, 8, 16, 14, 8, 13, 14, 15, 8, 15, 8, 8, 16, ~
## $ marca                    <chr> "NISSAN", "JEEP", "KIA", "VOLKSWAGEN", "SSANGY~
## $ modelo                   <chr> "KICKS", "COMPASS", "SORENTO", "AMAROK", "KORA~
## $ versao                   <chr> "1.6 16V FLEXSTART SL 4P XTRONIC", "2.0 16V FL~
## $ ano_de_fabricacao        <dbl> 2017, 2017, 2018, 2013, 2013, 2017, 2019, 2016~
## $ ano_modelo               <dbl> 2017, 2017, 2019, 2015, 2015, 2018, 2019, 2017~
## $ hodometro                <dbl> 67772, 62979, 44070, 85357, 71491, 85314, 2783~
## $ cambio                   <chr> "CVT", "Automática", "Automática", "Automática~
## $ num_portas               <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ tipo                     <chr> "Sedã", "Sedã", "Sedã", "Picape", "Utilitário ~
## $ blindado                 <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "~
## $ cor                      <chr> "Branco", "Branco", "Preto", "Branco", "Preto"~
## $ tipo_vendedor            <chr> "PF", "PF", "PJ", "PJ", "PF", "PJ", "PJ", "PJ"~
## $ cidade_vendedor          <chr> "Rio de Janeiro", "Belo Horizonte", "Santos", ~
## $ estado_vendedor          <chr> "São Paulo (SP)", "Minas Gerais (MG)", "São Pa~
## $ anunciante               <chr> "Pessoa Física", "Pessoa Física", "Loja", "Loj~
## $ entrega_delivery         <lgl> FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, F~
## $ troca                    <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, ~
## $ elegivel_revisao         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ dono_aceita_troca        <chr> NA, "Aceita troca", "Aceita troca", "Aceita tr~
## $ veiculo_único_dono       <chr> NA, NA, NA, NA, NA, NA, NA, NA, "Único dono", ~
## $ revisoes_concessionaria <chr> "Todas as revisões feitas pela concessionária"~
## $ ipva_pago                <chr> "IPVA pago", "IPVA pago", NA, "IPVA pago", NA,~
## $ veiculo_licenciado       <chr> "Licenciado", NA, NA, "Licenciado", NA, NA, NA~
## $ garantia_de_fábrica      <chr> NA, NA, NA, NA, "Garantia de fábrica", NA, NA,~
## $ revisoes_dentro_agenda   <chr> NA, NA, NA, NA, "Todas as revisões feitas pela~
## $ veiculo_alienado         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ preco                    <dbl> 74732.59, 81965.33, 162824.81, 123681.36, 8241~
```

Identifying and removing missing data

```
sapply(cars_train, function(x) sum(is.na(x)))
```

```
##                       id                num_fotos                    marca
##                        0                      177                        0
##                   modelo                   versao        ano_de_fabricacao
##                        0                        0                        0
##              ano_modelo                hodometro                   cambio
##                        0                        0                        0
##              num_portas                     tipo                 blindado
##                        0                        0                        0
##                      cor             tipo_vendedor          cidade_vendedor
##                        0                        0                        0
##         estado_vendedor                anunciante         entrega_delivery
##                        0                        0                        0
##                   troca          elegivel_revisao        dono_aceita_troca
##                        0                        0                     7662
##      veiculo_único_dono revisoes_concessionaria                ipva_pago
##                   19161                    20412                     9925
```

```
##      veiculo_licenciado     garantia_de_fábrica  revisoes_dentro_agenda
##                   13678                   25219                   23674
##        veiculo_alienado                   preco
##                   29584                       0
```

veiculo_único_dono: I will consider that the missing values represent "mais de um dono" (more than one owner), so I will replace it with that.

```
cars_train$veiculo_único_dono <-
  cars_train$veiculo_único_dono %>% replace_na("mais de um dono")
```

veiculo_licenciado: Null values were considered as "não licenciado"

```
cars_train$veiculo_licenciado <-
  cars_train$veiculo_licenciado %>% replace_na("não licenciado")
```

dono_aceita_troca: missing values replaced by "não aceita troca"

```
cars_train$dono_aceita_troca <-
  cars_train$dono_aceita_troca %>% replace_na("não aceita troca")
```

ipva_pago: missing values replaced by "IPVA não pago"

```
cars_train$ipva_pago <-
  cars_train$ipva_pago %>% replace_na("ipva não pago")
```

```
cars_train$num_fotos <-
  cars_train$num_fotos %>% replace_na(0)
```

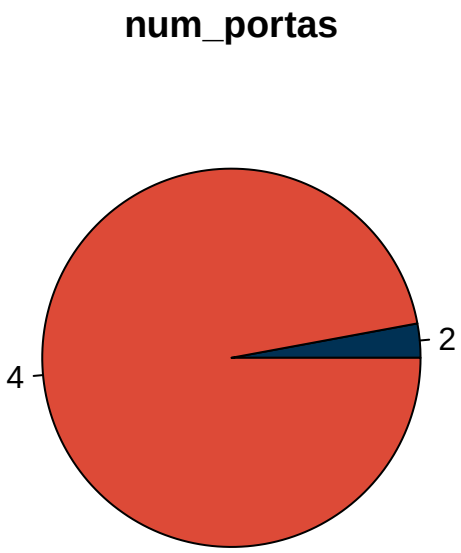
The missing items identified in the other variables, in addition to being very numerous, are redundant because even if corrected they would provide information without variation. These variables will be deleted.

Removing variables with redundant values, excess missing values and many categories

```
cars_train <- subset(cars_train,
                     select = -c(id,veiculo_alienado,revisoes_concessionaria,
                     garantia_de_fábrica,revisoes_dentro_agenda,
                     elegivel_revisao))
```

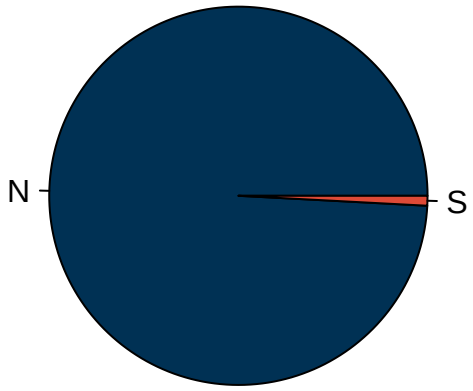
Check whether the levels of the categorical variables are balanced

```
#The brands FERRARI, IVECO, JAC, BRM and EFFA have very few observations, which therefore does not have

#marca
cars_train <- subset(cars_train, marca != "FERRARI")
cars_train <- subset(cars_train, marca != "IVECO")
cars_train <- subset(cars_train, marca != "JAC")
cars_train <- subset(cars_train, marca != "BRM")
cars_train <- subset(cars_train, marca != "EFFA")

#ano_de_fabricação
cars_train <- subset(cars_train, ano_de_fabricacao != "1985")
cars_train <- subset(cars_train, ano_de_fabricacao != "1988")
cars_train <- subset(cars_train, ano_de_fabricacao != "1990")

#ano_modelo
cars_train <- subset(cars_train, ano_modelo != "1997")
cars_train <- subset(cars_train, ano_modelo != "2006")
cars_train <- subset(cars_train, ano_modelo != "2008")
cars_train <- subset(cars_train, ano_modelo != "2010")

#num_portas
cars_train <- subset(cars_train, num_portas != 3)

#cor
cars_train <- subset(cars_train, cor != "Dourado")
cars_train <- subset(cars_train, cor != "Verde")
cars_train <- subset(cars_train, cor != "Vermelho")

#estado_vendedor
cars_train <- subset(cars_train, estado_vendedor != "Roraima (RR)")
cars_train <- subset(cars_train, estado_vendedor != "Maranhão (MA)")
cars_train <- subset(cars_train, estado_vendedor != "Rondônia (RO)")
cars_train <- subset(cars_train, estado_vendedor != "Piauí (PI)")

#anunciante
cars_train$anunciante[cars_train$anunciante == "Acessórios e serviços para autos"] <- "Concessionária"
cars_train$anunciante[cars_train$anunciante == "Loja"] <- "Concessionária"


#num_portas, blindado, entrega_delivery, troca, dono_aceita_troca: These are variables that present a s
pie(table(cars_train$num_portas), main = "num_portas", col= c("#003154","#dd4a37"))
```

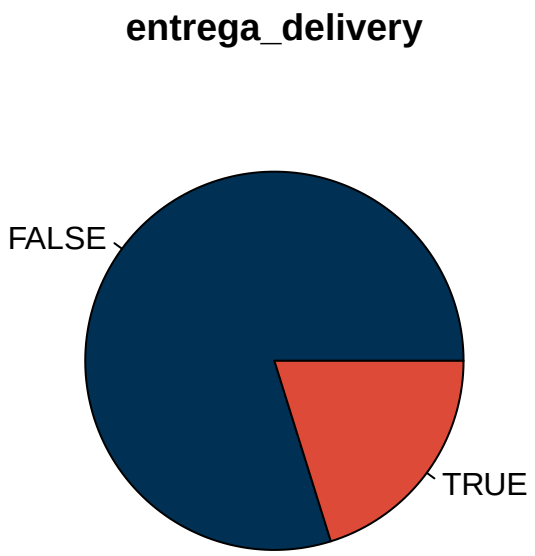



```
pie(table(cars_train$blindado), main = "blindado", col= c("#003154","#dd4a37"))
```

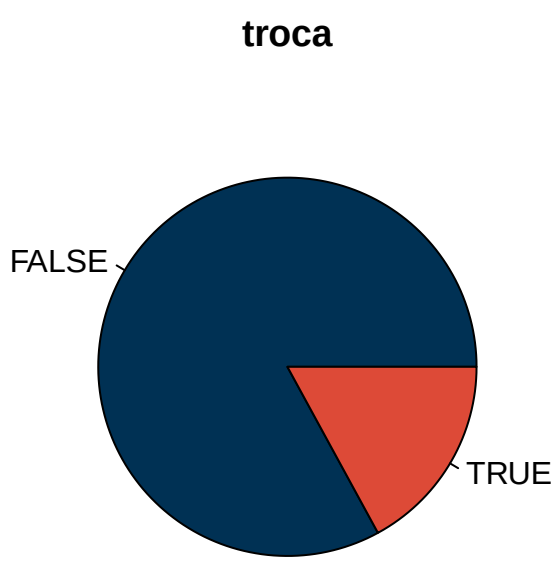
blindado



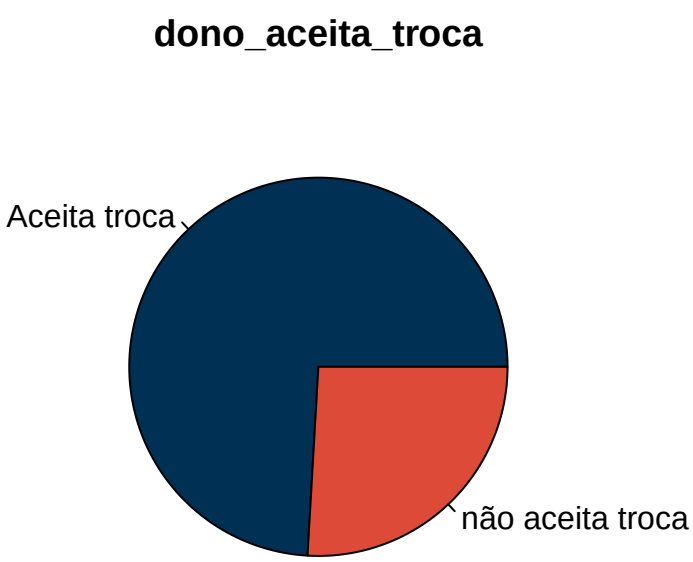

```
pie(table(cars_train$entrega_delivery), main = "entrega_delivery", col= c("#003154","#dd4a37"))
```




```
pie(table(cars_train$troca), main = "troca", col= c("#003154","#dd4a37"))
```

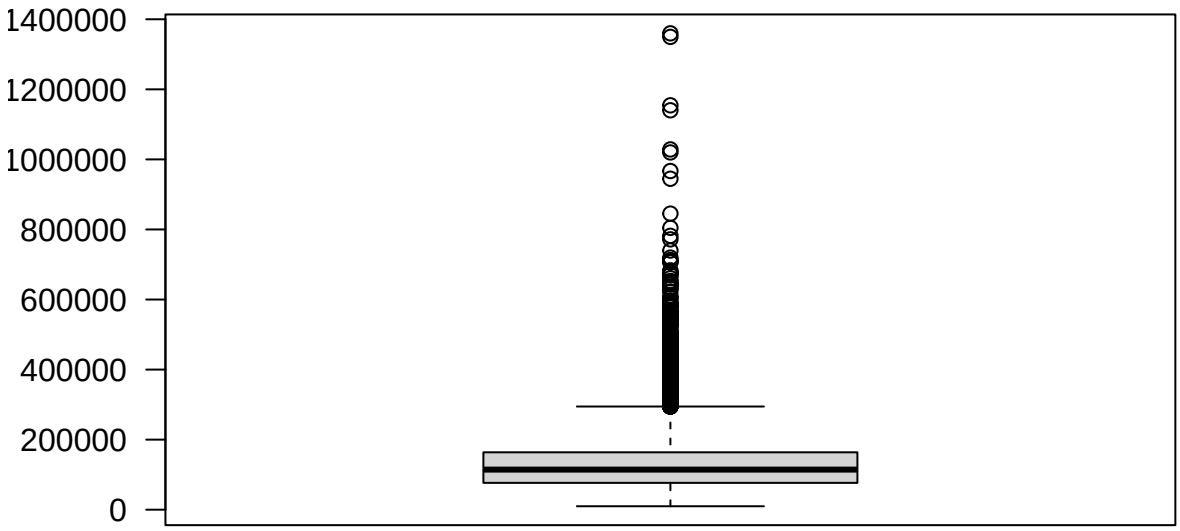



```
pie(table(cars_train$dono_aceita_troca), main = "dono_aceita_troca", col= c("#003154","#dd4a37"))
```

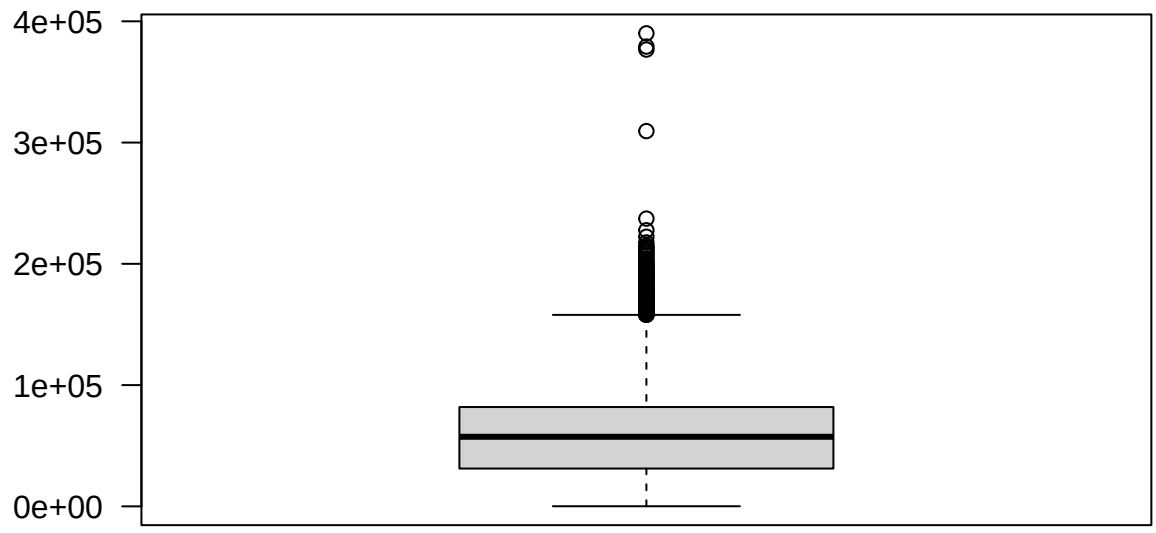


Identifying outliers

```
boxplot(cars_train$preco, las=2, xlab="preco")
```




```
boxplot(cars_train$hodometro,las=2, xlab="hodometro")
```



hodometro

Quartile function

```
quartil <- function(column){

  q1 <- quantile(column, 0.25, na.rm = TRUE) #1º quartil
  q3 <- quantile(column, 0.75, na.rm = TRUE) #3º quartil
  iq <- q3 - q1 #interquartil
  lim_sup <- q3 + 1.5*iq #limite superior
  return(lim_sup)
}
```

Calculating outliers across the top quartile

```
max_preco <- quartil(cars_train$preco)
max_hodo <- quartil(cars_train$hodometro)
print(paste("Preço:",max_preco, "Hodometro:", max_hodo))
```

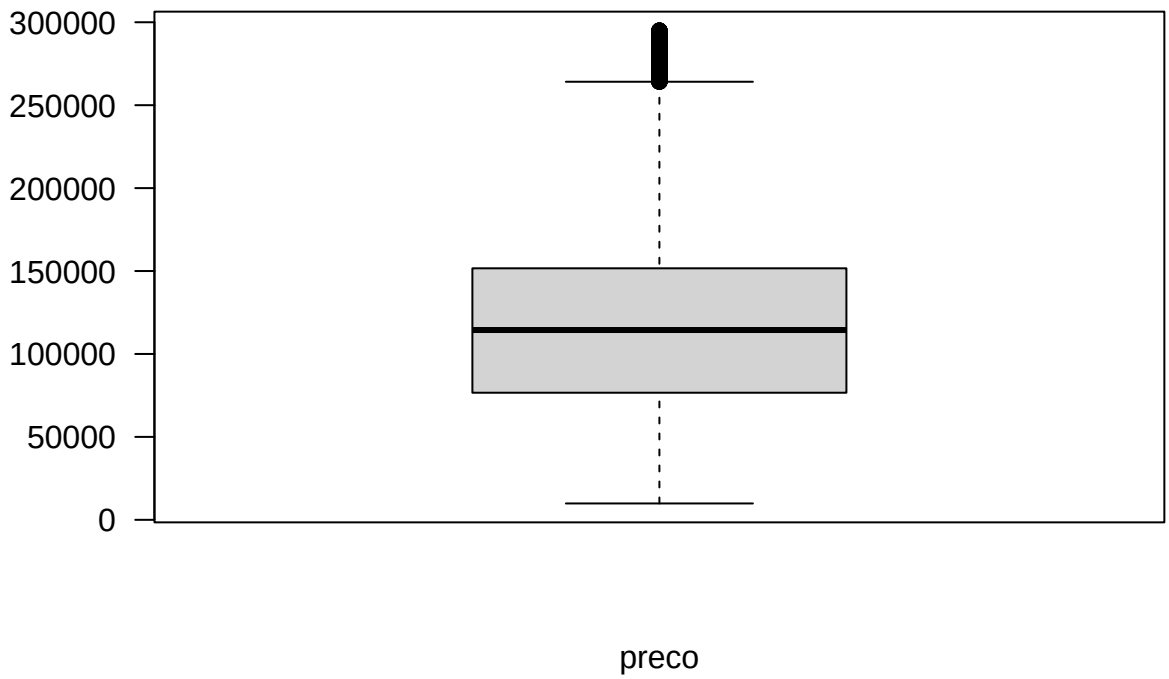
```
## [1] "Preço: 294610.784413666 Hodometro: 157969.5"
```

Now we will discard the lines where price and odometer are above the upper limit

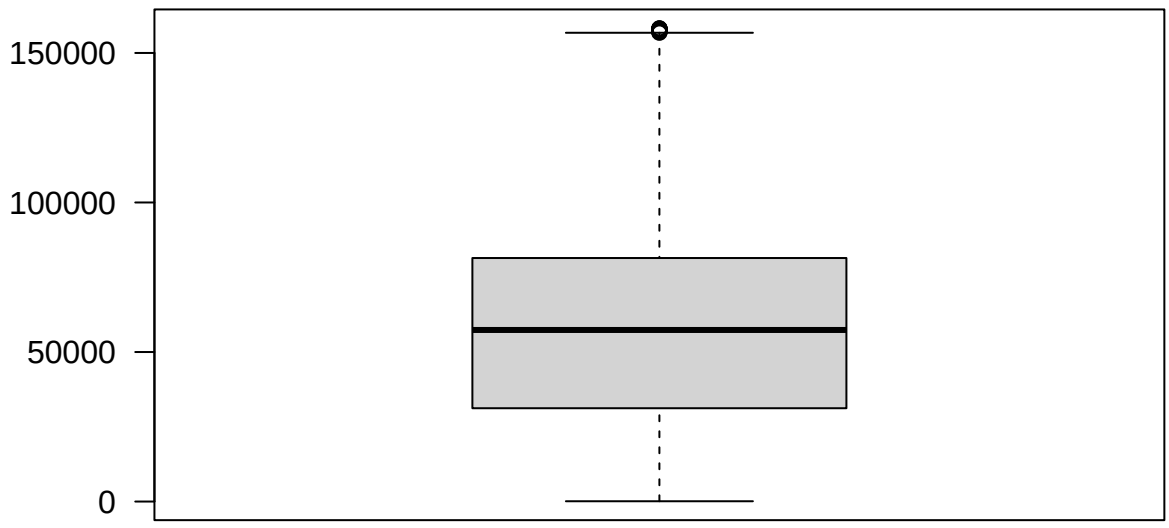
```
for (i in seq_along(cars_train$preco)){
  if (cars_train$preco[i] > 295085.69){
    cars_train$preco[i] <- mean(cars_train$preco)
  }
}

for (i in seq_along(cars_train$hodometro)){
  if (cars_train$hodometro[i] > 158264){
    cars_train$hodometro[i] <- mean(cars_train$hodometro)
  }
}
```

Visualize the distribution of variables, now without the outliers

```
boxplot(cars_train$preco, las=2, xlab="preco")
```




```
boxplot(cars_train$hodometro,las=2, xlab="hodometro")
```



Transforming variables into factors

```
cars_train[,c("marca", "cor", "cambio", "tipo", "ano_de_fabricacao",
              "ano_modelo", "dono_aceita_troca", "veiculo_único_dono",
              "ipva_pago", "veiculo_licenciado","blindado",
              "tipo_vendedor","estado_vendedor",
              "anunciante")]=
  lapply(cars_train[,c("marca", "cor", "cambio", "tipo", "ano_de_fabricacao",
              "ano_modelo", "dono_aceita_troca", "veiculo_único_dono",
              "ipva_pago", "veiculo_licenciado","blindado",
              "tipo_vendedor","estado_vendedor",
              "anunciante")], as.factor)

summary(cars_train)
```

```
##    num_fotos           marca          modelo            versao
##  Min.   : 0.00   VOLKSWAGEN: 4584   Length:29513       Length:29513
##  1st Qu.: 8.00   CHEVROLET : 3011   Class :character   Class :character
##  Median : 8.00   TOYOTA    : 2178   Mode  :character   Mode  :character
##  Mean   :10.26   HYUNDAI   : 2043
##  3rd Qu.:14.00   JEEP      : 2000
##  Max.   :21.00   FIAT      : 1903
##                  (Other)   :13794
```

```
##  ano_de_fabricacao   ano_modelo     hodometro
##  2020   :4715      2021   :5056   Min.   :   100
##  2017   :4366      2017   :4512   1st Qu.: 31197
##  2019   :3873      2018   :4218   Median : 57373
##  2018   :3815      2019   :3580   Mean   : 57731
##  2021   :2611      2020   :3536   3rd Qu.: 81433
##  2013   :2443      2015   :2384   Max.   :158228
##  (Other):7690      (Other):6227
##                    cambio        num_portas                       tipo
##  Automática           :22514   Min.   :2.000   Cupê                :   26
##  Automática Sequencial:   24   1st Qu.:4.000   Hatchback           : 4910
##  Automatizada         :  139   Median :4.000   Minivan             :    7
##  Automatizada DCT     :   53   Mean   :3.942   Perua/SW            :   26
##  CVT                  : 1791   3rd Qu.:4.000   Picape              : 4817
##  Manual               : 4951   Max.   :4.000   Sedã                :16406
##  Semi-automática      :   41                   Utilitário esportivo: 3321
##  blindado      cor        tipo_vendedor cidade_vendedor
##  N:29266   Branco:20914   PF:17900      Length:29513
##  S:  247   Cinza : 1632   PJ:11613      Class :character
##            Prata : 1725                 Mode  :character
##            Preto : 5242
##
##
##
##                estado_vendedor           anunciante    entrega_delivery
##  São Paulo (SP)        :16344   Concessionária:11540   Mode :logical
##  Rio de Janeiro (RJ)   : 2541   Pessoa Física :17973   FALSE:23558
##  Paraná (PR)           : 2525                          TRUE :5955
##  Santa Catarina (SC)   : 2299
##  Minas Gerais (MG)     : 1773
##  Rio Grande do Sul (RS): 1644
##  (Other)               : 2387
##    troca                dono_aceita_troca       veiculo_único_dono
##  Mode :logical   Aceita troca    :21864   mais de um dono:19121
##  FALSE:24480     não aceita troca: 7649   Único dono     :10392
##  TRUE :5033
##
##
##
##
##          ipva_pago          veiculo_licenciado     preco
##  ipva não pago: 9899   Licenciado    :15864    Min.   :  9870
##  IPVA pago    :19614   não licenciado:13649    1st Qu.: 76631
##                                                Median :114435
##                                                Mean   :121209
##                                                3rd Qu.:151654
##                                                Max.   :295002
##
```

Saving clean dataset

```r
write.csv(cars_train, "../data/cars_train_clean1.csv", row.names = FALSE)
```