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       TRUE
```

Viewing the database

```
cars_train <- read_delim("../datasets/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)
```

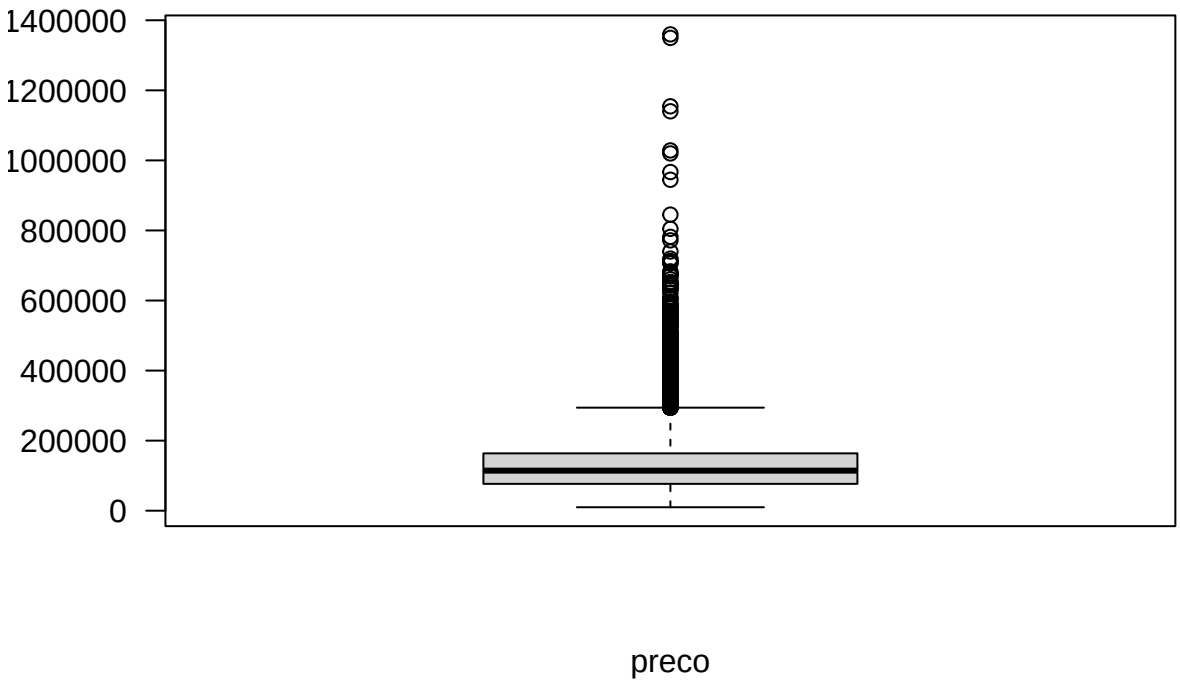
The missing items identified in the other variables, in addition to being very numerous, are redundant because even if corrected they would provide information without variation. These variables will be deleted.

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

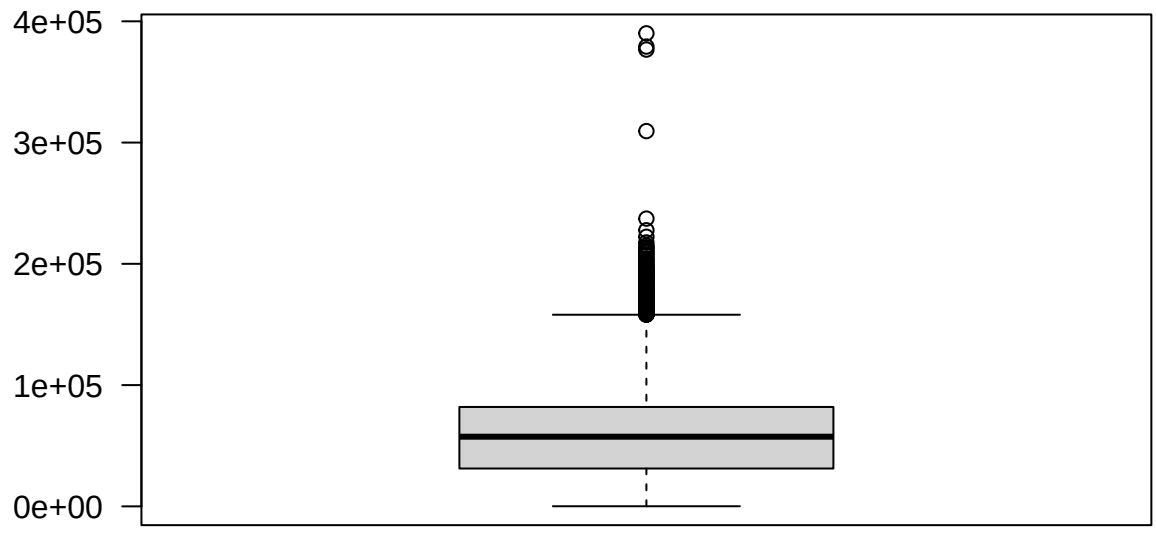
```
cars_train <- subset(cars_train,
                     select = -c(id,veiculo_alienado,revisoes_concessionaria,
                     garantia_de_fábrica,revisoes_dentro_agenda,
                     elegivel_revisao))
```

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))
```

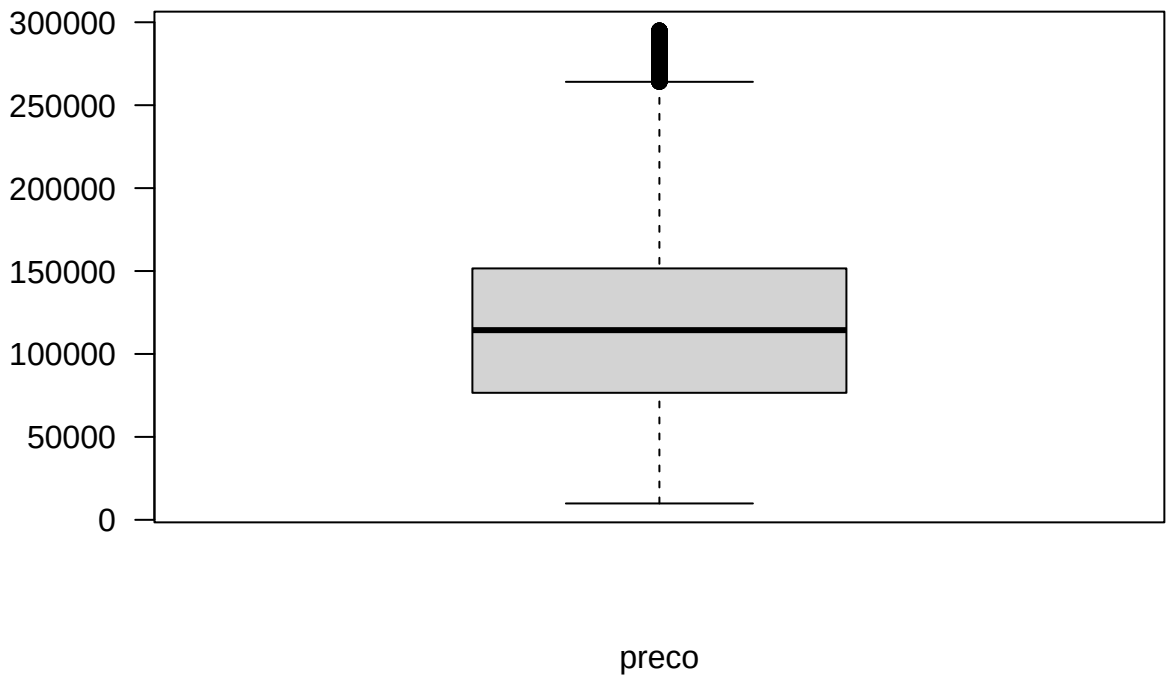
```
## [1] "Preço: 294341.390870122 Hodometro: 158062.75"
```

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

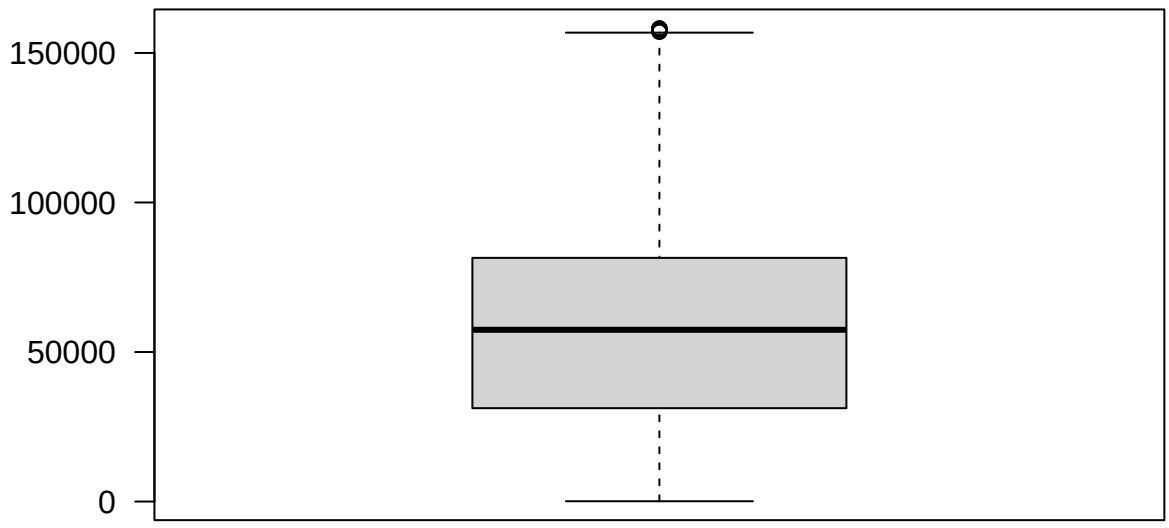
```
for (i in seq_along(cars_train$preco)){
  if (cars_train$preco[i] > 295085.69){
    cars_train$preco[i] <- mean(cars_train$preco)
  }
}

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: 4594   Length:29584       Length:29584
##  1st Qu.: 8.00   CHEVROLET : 3020   Class :character   Class :character
##  Median : 8.00   TOYOTA    : 2180   Mode  :character   Mode  :character
##  Mean   :10.26   HYUNDAI   : 2043
##  3rd Qu.:14.00   JEEP      : 2000
##  Max.   :21.00   FIAT      : 1918
##                  (Other)   :13829
```

```
##  ano_de_fabricacao   ano_modelo     hodometro
##  2020   :4729      2021   :5071   Min.   :   100
##  2017   :4369      2017   :4519   1st Qu.: 31214
##  2019   :3880      2018   :4221   Median : 57434
##  2018   :3820      2019   :3587   Mean   : 57776
##  2021   :2614      2020   :3541   3rd Qu.: 81484
##  2013   :2443      2015   :2386   Max.   :158228
##  (Other):7729      (Other):6259
##                    cambio        num_portas                       tipo
##  Automática           :22545   Min.   :2.000   Cupê                :   26
##  Automática Sequencial:   25   1st Qu.:4.000   Hatchback           : 4924
##  Automatizada         :  139   Median :4.000   Minivan             :    7
##  Automatizada DCT     :   53   Mean   :3.941   Perua/SW            :   27
##  CVT                  : 1792   3rd Qu.:4.000   Picape              : 4849
##  Manual               : 4989   Max.   :4.000   Sedã                :16429
##  Semi-automática      :   41                   Utilitário esportivo: 3322
##  blindado        cor        tipo_vendedor cidade_vendedor
##  N:29336   Branco  :20949   PF:17926      Length:29584
##  S:  248   Cinza   : 1634   PJ:11658      Class :character
##            Dourado :    2                 Mode  :character
##            Prata   : 1741
##            Preto   : 5256
##            Verde   :    1
##            Vermelho:    1
##                estado_vendedor                              anunciante
##  São Paulo (SP)        :16378   Acessórios e serviços para autos:    4
##  Rio de Janeiro (RJ)   : 2548   Concessionária                   : 1702
##  Paraná (PR)           : 2526   Loja                             : 9879
##  Santa Catarina (SC)   : 2302   Pessoa Física                    :17999
##  Minas Gerais (MG)     : 1775
##  Rio Grande do Sul (RS): 1646
##  (Other)               : 2409
##  entrega_delivery   troca               dono_aceita_troca
##  Mode :logical    Mode :logical   Aceita troca    :21922
##  FALSE:23601      FALSE:24523     não aceita troca: 7662
##  TRUE :5983       TRUE :5061
##
##
##
##
##       veiculo_único_dono         ipva_pago          veiculo_licenciado
##  mais de um dono:19161    ipva não pago: 9925   Licenciado    :15906
##  Único dono     :10423    IPVA pago    :19659   não licenciado:13678
##
##
##
##
##
##      preco
##  Min.   :  9870
##  1st Qu.: 76572
##  Median :114356
##  Mean   :121138
##  3rd Qu.:151584
```

```
##  Max.   :295002
##
```

Saving clean dataset

```r
write.csv(cars_train, "cars_train_clean.csv", row.names = FALSE)
```