USED CAR PRICES CASE STUDY

Deliverable I: Data Processing, Description, Validation and Profiling

Miquel Parra i Xavier Alaman

March 13, 2022

Contents

1 R libraries imports, useful functions and data loading

In this first section we will load all required packages and libraries, declare additional functions, and load our data.

1.1 Load required packages

```
options(contrasts=c("contr.treatment","contr.treatment"))

requiredPackages <- c("effects","FactoMineR","car",
                      "factoextra","RColorBrewer","ggplot2","dplyr","ggmap",
                      "ggthemes","knitr")

#use this function to check if each package is on the local machine
#if a package is installed, it will be loaded
#if any are not, the missing package(s) will be installed and loaded
package.check <- lapply(requiredPackages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
})

#verify they are loaded
search()
```

1.2 Useful functions

```r
# Mout <- which((df$tax < var_out$mouti)|(df$tax > var_out$mouts))

# Some useful functions
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],
       q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr, souts=s.x[5]+3*iqr ) }

countNA <- function(x) {
  mis_x <- NULL
  for (j in 1:ncol(x)) {mis_x[j] <- sum(is.na(x[,j])) }
  mis_x <- as.data.frame(mis_x)
  rownames(mis_x) <- names(x)
  mis_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j])) }
  list(mis_col=mis_x,mis_ind=mis_i) }

countX <- function(x,X) {
  n_x <- NULL
  for (j in 1:ncol(x)) {n_x[j] <- sum(x[,j]==X) }
  n_x <- as.data.frame(n_x)
  rownames(n_x) <- names(x)
  nx_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {nx_i <- nx_i + as.numeric(x[,j]==X) }
  list(nx_col=n_x,nx_ind=nx_i) }
```

1.3 Sample load

```r
# Clear plots
if(!is.null(dev.list())) dev.off()

# Clean workspace
rm(list=ls())

# Users file path
miquel_fp <- "C:/Users/Miquel/Documents/GitHub/ADEI/"
xavi_fp <- "~/Documents/FIB/ADEI/ADEI/"
filepath <- xavi_fp

# Set working directory
setwd(filepath)

# Load data from file
load(paste0(filepath,"MyOldCars-Raw.RData"))
```

2 Data Description

During this project we will be working with a subset of the pre-treated original dataset "Uk used car dataset". A sample of 5000 cars has been randomly selected from Mercedes, BMW, Volkwagen and Audi manufacturers and stored into a RData file ***MyOldCars-Raw.RData***.

2.1 Original variables description

- **model:** Car model.
- **year:** Car registration year.
- **price:** Car price in £.
- **transmission:** Type of transmission ["Manual", "Automatic", "Semi-Auto"].

- **mileage:** Distance used, accumulated miles.
- **fuelType:** Type of engine fuel [“Petrol”, “Diesel”, “Hybrid”, “Other”].
- **tax:** Applied road tax.
- **mpg:** Miles per gallon.
- **engineSize:** Engine size in liters. The cars with engine size 0 are in fact electric cars, nevertheless Mercedes C class, and other given cars are not electric cars,so data imputation is required.
- **manufacturer:** Car manufacturer [“Audi”, “BMW”, “Mercedes”, “VW”].

```
summary(df)
```

```
##     model               year          price        transmission
##  Length:5000        Min.   :1999   Min.   :   650   Length:5000
##  Class :character   1st Qu.:2016   1st Qu.: 13995   Class :character
##  Mode  :character   Median :2017   Median : 19498   Mode  :character
##                     Mean   :2017   Mean   : 21470
##                     3rd Qu.:2019   3rd Qu.: 26039
##                     Max.   :2020   Max.   :109990
##     mileage         fuelType              tax             mpg
##  Min.   :     4   Length:5000        Min.   :  0.0   Min.   :  8.80
##  1st Qu.:  5999   Class :character   1st Qu.:125.0   1st Qu.: 44.80
##  Median : 16619   Mode  :character   Median :145.0   Median : 53.30
##  Mean   : 23312                      Mean   :125.3   Mean   : 53.89
##  3rd Qu.: 33834                      3rd Qu.:145.0   3rd Qu.: 61.40
##  Max.   :153000                      Max.   :580.0   Max.   :470.80
##    engineSize    manufacturer
##  Min.   :0.000   Length:5000
##  1st Qu.:1.500   Class :character
##  Median :2.000   Mode  :character
##  Mean   :1.917
##  3rd Qu.:2.000
##  Max.   :6.600
```

```
head(df, 3)
```

```
##    model year price transmission mileage fuelType tax  mpg engineSize
## 3     A1 2016 11000       Manual   29946   Petrol  30 55.4        1.4
## 9     A3 2015 10200       Manual   46112   Petrol  20 60.1        1.4
## 26    A4 2017 18500    Automatic   17418   Diesel 145 62.8        2.0
##    manufacturer
## 3          Audi
## 9          Audi
## 26         Audi
```

3 Univariate Descriptive Analysis

In this step of the process original numeric variables corresponding to qualitative concepts have to be converted to factors. New factors grouping original levels will be considered very positively.

Additionally original numeric variables corresponding to real quantitative concepts are kept as numeric but additional factors should also be created as a discretization of each numeric variable.

3.1 model

```
df$model<-factor(paste0(trimws(df$manufacturer),"-",trimws(df$model)))
df$model<-factor(df$model,labels=paste0("f.Model",levels(df$model)))

summary(df$model)
```

```
##            f.ModelAudi-A1            f.ModelAudi-A3            f.ModelAudi-A4
##                       137                       199                       136
##            f.ModelAudi-A5            f.ModelAudi-A6            f.ModelAudi-A7
##                       101                        71                        14
##            f.ModelAudi-A8            f.ModelAudi-Q2            f.ModelAudi-Q3
##                        16                        80                       142
##            f.ModelAudi-Q5            f.ModelAudi-Q7            f.ModelAudi-Q8
##                        93                        32                         8
##            f.ModelAudi-R8           f.ModelAudi-RS3           f.ModelAudi-RS4
##                         2                         5                         2
##           f.ModelAudi-RS5           f.ModelAudi-RS6            f.ModelAudi-S3
##                         3                         3                         1
##            f.ModelAudi-S4           f.ModelAudi-SQ5            f.ModelAudi-TT
##                         4                         5                        31
##       f.ModelBMW-1 Series       f.ModelBMW-2 Series       f.ModelBMW-3 Series
##                       197                       138                       243
##       f.ModelBMW-4 Series       f.ModelBMW-5 Series       f.ModelBMW-6 Series
##                        98                        86                        10
##       f.ModelBMW-7 Series       f.ModelBMW-8 Series             f.ModelBMW-i3
##                        12                         5                         5
##             f.ModelBMW-i8             f.ModelBMW-M2             f.ModelBMW-M3
##                         3                         5                         6
##             f.ModelBMW-M4             f.ModelBMW-M5             f.ModelBMW-M6
##                        11                         4                         1
##             f.ModelBMW-X1             f.ModelBMW-X2             f.ModelBMW-X3
##                        74                        22                        48
##             f.ModelBMW-X4             f.ModelBMW-X5             f.ModelBMW-X6
##                        18                        33                         7
##             f.ModelBMW-X7             f.ModelBMW-Z3             f.ModelBMW-Z4
##                         7                         1                        15
##   f.ModelMercedes-A Class   f.ModelMercedes-B Class   f.ModelMercedes-C Class
##                       270                        63                       367
##  f.ModelMercedes-CL Class f.ModelMercedes-CLA Class f.ModelMercedes-CLC Class
##                        57                        11                         2
## f.ModelMercedes-CLS Class   f.ModelMercedes-E Class  f.ModelMercedes-GL Class
##                        18                       180                        14
## f.ModelMercedes-GLA Class f.ModelMercedes-GLB Class f.ModelMercedes-GLC Class
##                        84                         1                       119
## f.ModelMercedes-GLE Class f.ModelMercedes-GLS Class   f.ModelMercedes-M Class
##                        55                        14                         8
##   f.ModelMercedes-S Class  f.ModelMercedes-SL CLASS       f.ModelMercedes-SLK
##                        16                        38                         5
##   f.ModelMercedes-V Class   f.ModelMercedes-X-CLASS          f.ModelVW-Amarok
##                        18                        11                        16
##          f.ModelVW-Arteon          f.ModelVW-Beetle      f.ModelVW-Caddy Life
##                        27                         7                         2
## f.ModelVW-Caddy Maxi Life      f.ModelVW-California       f.ModelVW-Caravelle
##                         5                         1                         8
##              f.ModelVW-CC            f.ModelVW-Golf         f.ModelVW-Golf SV
##                        11                       488                        25
##           f.ModelVW-Jetta          f.ModelVW-Passat            f.ModelVW-Polo
##                         3                        94                       348
##        f.ModelVW-Scirocco          f.ModelVW-Sharan         f.ModelVW-Shuttle
##                        24                        22                         9
##         f.ModelVW-T-Cross           f.ModelVW-T-Roc          f.ModelVW-Tiguan
##                        25                        63                       164
## f.ModelVW-Tiguan Allspace         f.ModelVW-Touareg          f.ModelVW-Touran
##                        14                        32                        39
##              f.ModelVW-Up
##                        88
```

```
barplot(summary(df$model), main = "Model Barplot", col = "blue", horiz=TRUE)
```