

# Partial Exam ADEI Q2-2022/2023

Josep Franquet and Daniel Villalobos

20/4/2023

## Problem:

Craigslist is the world's largest collection of used vehicles for sale, yet it's very difficult to collect all of them in the same place. The following data set includes every used vehicle entry within the United States on Craigslist. It contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories (details here). Can we determine the relationships between the variables in the data set and which relate best to used car's price? The variables which you should retain from the data set are the following ones:

Feature	Description
manufacturer	Factor: Audi, BMW, Mercedes, or Volkswagen
model	Car model
year	Registration year
price	Price in £
transmission	Type of gearbox
mileage	Distance used
fuelType	Engine fuel
tax	Road tax
mpg	Consumption in miles per gallon
engineSize	Size in litres

1. Using the .Rdata provided, load it to your workspace and you will find a dataframe object (*df*) which only retains the mentioned variables. Validate that all cars manufacturers are Audi, BMW, Mercedes or Volkswagen. Additionally, validate that the number of rows is 49725. (0.5 points)

Càrrega del .Rdata proporcionat al directori de treball:

```
load("Partial_exam_cars.RData")
```

Validació dels manufacturers existents:

```
unique(df$manufacturer)
```

```
## [1] "Audi"      "BMW"       "Mercedes"   "VW"
```

Validació del nombre de files:

```
nrow(df)
## [1] 49725
```

## 2. Make sure you understand the variables in the dataset. (1 point)

- Which variables are numeric and which are categorical (character variables should be treated as expected).
- Make a short description (summarize) of the dataset at hand.
- Is there any *id* variable? (if not, create a new one called *car\_id*)

En primer lloc, mirem quins tipus de variables tenim al nostre *dataframe*:

```
str(df)
```

```
## 'data.frame': 49725 obs. of 10 variables:
## $ model      : chr "A1" "A6" "A1" "A4" ...
## $ year       : int 2017 2016 2016 2017 2019 2016 2016 2016 2015 2016 ...
## $ price      : int 12500 16500 11000 16800 17300 13900 13250 11750 10200 12000 ...
## $ transmission: chr "Manual" "Automatic" "Manual" "Automatic" ...
## $ mileage    : int 15735 36203 29946 25952 1998 32260 76788 75185 46112 22451 ...
## $ fuelType   : chr "Petrol" "Diesel" "Petrol" "Diesel" ...
## $ tax        : int 150 20 30 145 145 30 30 20 20 30 ...
## $ mpg         : num 55.4 64.2 55.4 67.3 49.6 58.9 61.4 70.6 60.1 55.4 ...
## $ engineSize : num 1.4 2 1.4 2 1 1.4 2 2 1.4 1.4 ...
## $ manufacturer: chr "Audi" "Audi" "Audi" "Audi" ...
```

En aquest punt, passem a factor les variables de tipus *chr*:

```
df$model <- factor(df$model)
df$transmission <- factor(df$transmission)
df$manufacturer <- factor(df$manufacturer)
df$fuelType <- factor(df$fuelType)
```

Per tant, ja podem dir que el nostre dataset està format per:

- 6 variables numèriques: year, price, mileage, tax, mpg i engineSize
- 4 variables categòriques: model, transmission, manufacturer i fuelType.

A continuació, fem una descripció de les nostres dades:

```
summary(df)
```

	model	year	price	transmission
##	Golf	4863	Min. :1970	Min. : 650 Automatic:13081
##	C Class	3747	1st Qu.:2016	1st Qu.: 13995 Manual :17757
##	Polo	3287	Median :2017	Median : 19498 Other : 2
##	A Class	2561	Mean :2017	Mean : 21490 Semi-Auto:18885

```

##   3 Series: 2443   3rd Qu.:2019   3rd Qu.: 26090
##   1 Series: 1969   Max.     :2020   Max.     :159999
##  (Other)  :30855
##      mileage          fuelType          tax          mpg
##  Min.    : 1   Diesel   :28163   Min.    : 0.0   Min.    : 0.30
##  1st Qu.: 5891 Electric  :     3   1st Qu.:125.0   1st Qu.: 44.80
##  Median  :16908 Hybrid   : 644   Median  :145.0   Median  : 53.30
##  Mean    :23380 Other    :130   Mean    :124.2   Mean    : 54.06
##  3rd Qu.:33981 Petrol   :20785   3rd Qu.:145.0   3rd Qu.: 61.40
##  Max.    :323000
##      engineSize        manufacturer
##  Min.   :0.000   Audi      :10668
##  1st Qu.:1.500   BMW      :10781
##  Median  :2.000   Mercedes:13119
##  Mean    :1.919   VW       :15157
##  3rd Qu.:2.000
##  Max.    :6.600
##

```

Respecte l'*output* obtingut, s'espera que l'estudiant destaquï alguns aspectes de les dades amb les que ha de treballar. Algunes possibilitats són los següents:

- Destacar les freqüències de les variables categòriques: quines són les més freqüents, hi ha molta diferència entre la freqüència de les diferents categories, quines és la variable categòrica en la qual les categories tenen una freqüència més desigual... entre d'altres.
- Destacar la presència de categories residuals: Other (transmission) o Electric (fuelType)
- Per altra banda, respecte les variables numèriques, l'estudiant hauria d'observar valors màxims i mínims estranys: el màxim de price o el de mpg són molt elevats en comparació amb els altres estadístics.
- En relació amb els valors mínims d'aquestes variables, també observem valors molt petits per a la gran majoria.

(Recordatori: Es tracta d'alguns exemples d'aspectes a destacar)

Per últim, caldria observar i detectar que no tenim cap variable que es pugui utilitzar com a identificadora. Per tant, tal i com s'indica, es crea la variable *car\_id* de la següent manera:

```
df$car_id <- as.character(c(1:nrow(df)))
```

Tal i com es pot veure, les variables identificadores sempre haurien de ser de tipus string.

**3. Make a data quality report. Are there any missing data (apply an adequate function/s to validate it)? Are there any errors (assign them as NA's if it applies)? Which variables are affected? (1 point)**

En primer lloc, cal confirmar que no hi ha missings en el dataframe:

```

if (any(is.na(df))) {
  print("El dataframe té NA's")
} else {
  print("El dataframe no té NA's")
}

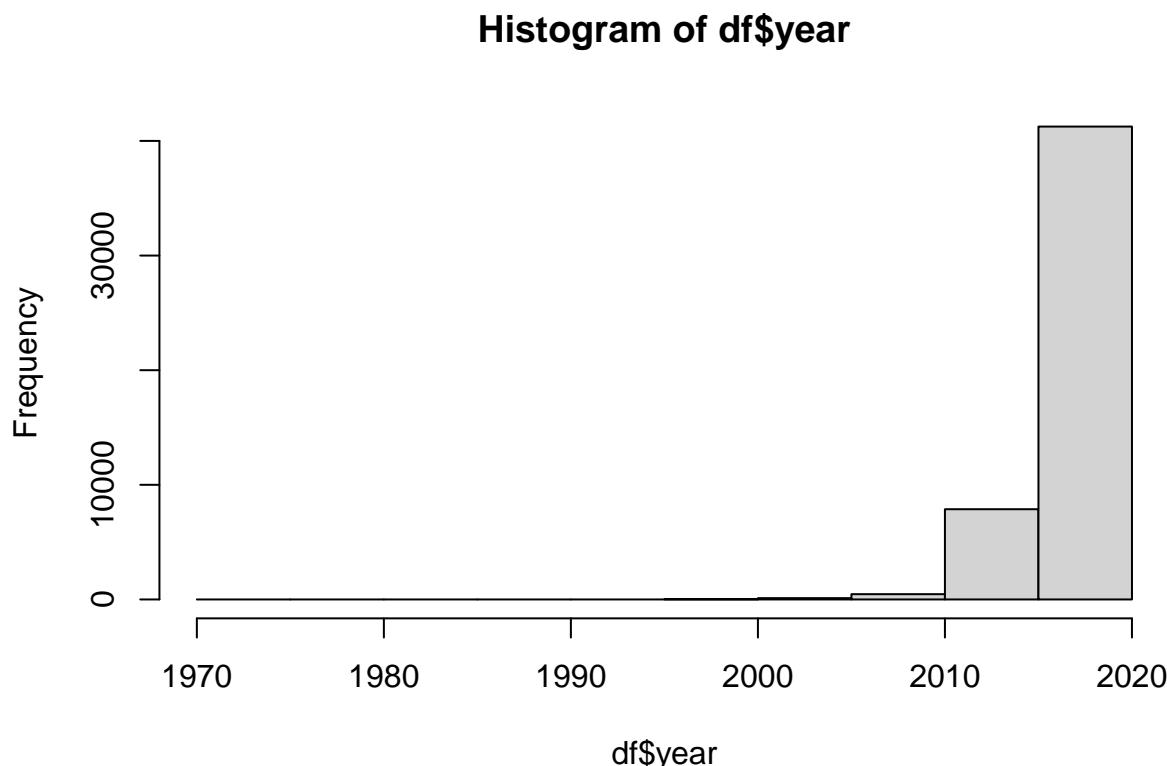
```

```
## [1] "El dataframe no té NA's"
```

En relació a la detecció d'errors, l'estudiant hauria de saber què un error és un valor d'alguna variable que no es pugui donar (o sigui molt poc probable). Per tant, centrant-nos en el nostre dataset, s'hauria de comentar:

- La variable year presenta un valor mínim molt petit. Es vol analitzar la distribució de valors de la mateixa de la següent manera:

```
hist(df$year)
```



Sembla que 1970 és un valor molt petit. Seguidament es mira quins cotxes tenen aquest valor:

```
df [which(df$year == 1970),]
```

```
##          model year price transmission mileage fuelType tax  mpg engineSize
## 33522    M Class 1970 24999     Automatic   14000 Diesel 305 39.2         0
##       manufacturer car_id
## 33522    Mercedes 33522
```

Veient la resta de valors, clarament es considera un error:

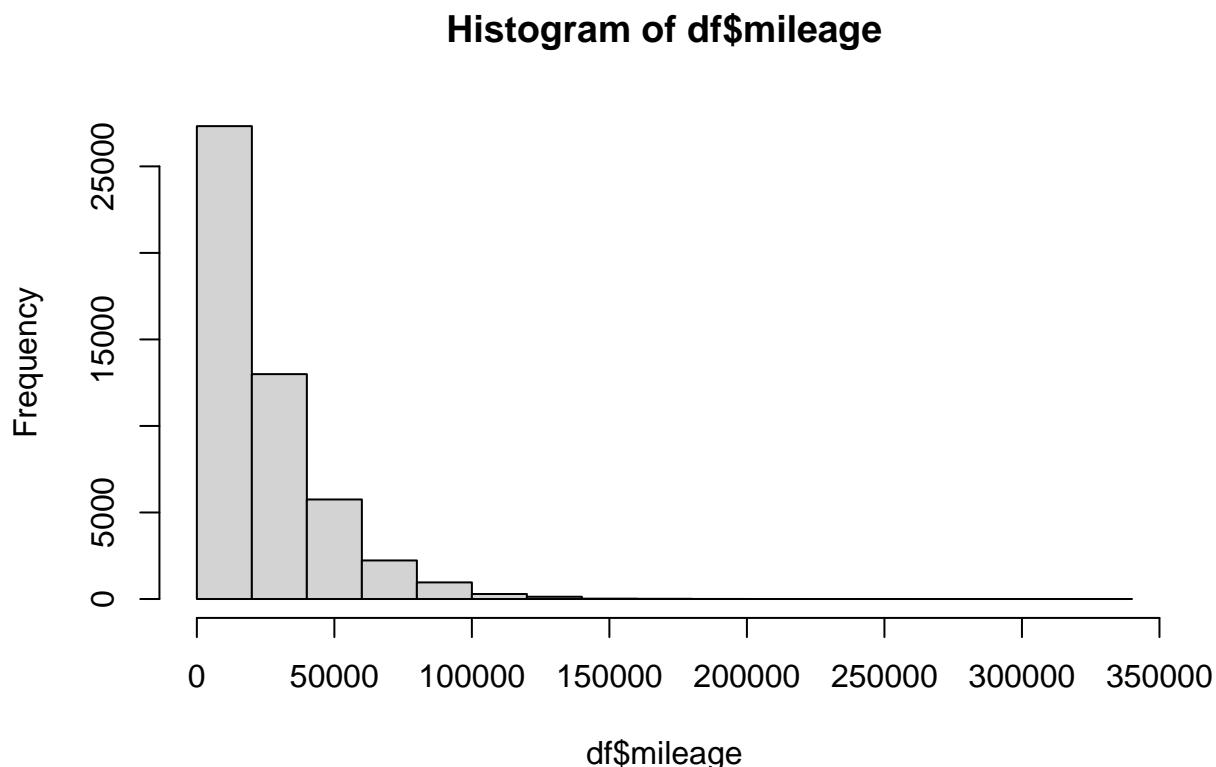
```
df [which(df$year == 1970), "year"] <- NA
```

- En relació a la variable transmission, es tracten com errors els dos valors en els quals la variable pren el valor d'Other ja que es considera que no poden existir altres tipus de transmissions:

```
df[which(df$transmission == "Other"), "transmission"] <- NA
```

- Per a la variable mileage, es podrien analitzar els valors en els quals la mateixa pren un valor molt elevat. Es fa de la següent manera:

```
hist(df$mileage)
```



Veient la distribució de valors, es busquen els cotxes en els quals aquesta variable pren un valor superior a 250000:

```
df[which(df$mileage>250000),]
```

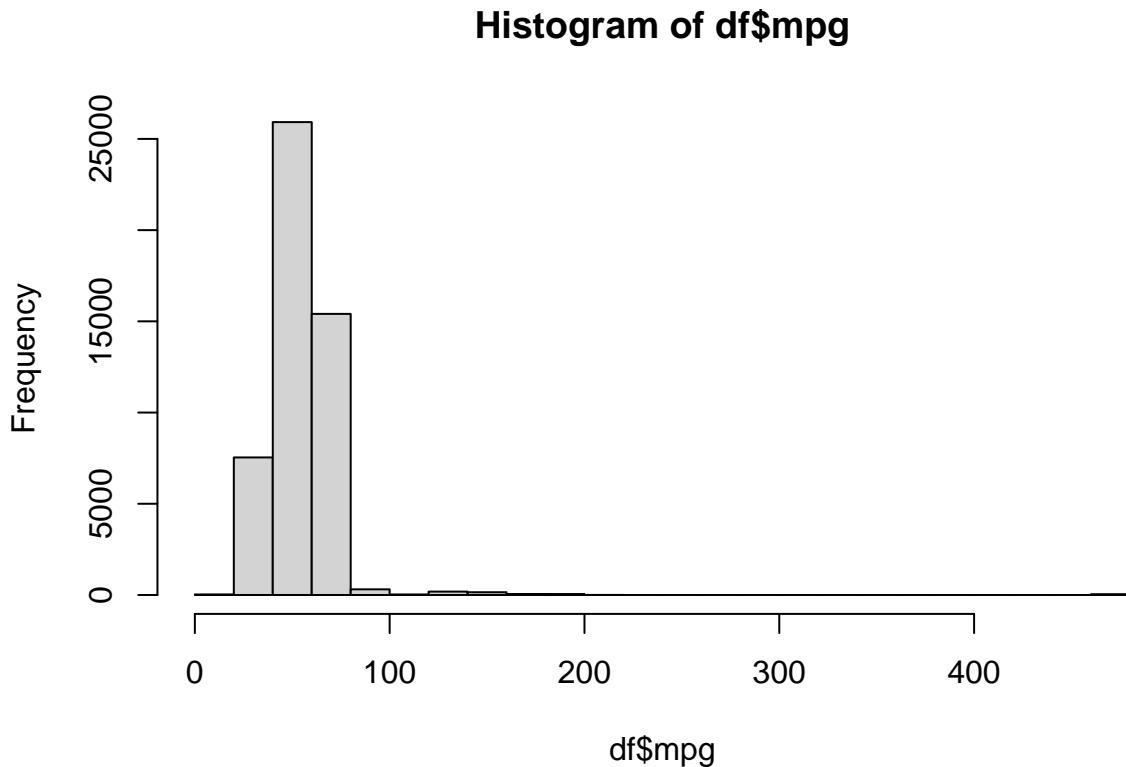
```
##          model year price transmission mileage fuelType tax mpg engineSize
## 9823        A6 2008  2490      Manual  323000 Diesel 200 44.1       2
## 32794    V Class 2010  6949   Automatic  259000 Diesel 540 30.7       3
##           manufacturer car_id
## 9823        Audi     9823
## 32794    Mercedes  32794
```

En aquest cas, es considera que un valor de 323000 és molt elevat i es considera un error:

```
df[which(df$mileage==323000), "mileage"] <- NA
```

- Per últim, el valor màxim de la variable mpg resulta ser extremadament elevat, per això, s'analitza la seva distribució de valors:

```
hist(df$mpg)
```



Confirmant la nostra hipòtesi, es troben aquests cotxes:

```
df [which(df$mpg>300),]
```

```
##      model year price transmission mileage fuelType tax    mpg engineSize
## 10701   i3 2016 17100   Automatic  25269     Other  0 470.8      0.6
## 11290   i3 2017 19998   Automatic  41949    Hybrid 140 470.8      0.0
## 11447   i3 2017 19998   Automatic  41146    Hybrid  0 470.8      0.0
## 11449   i3 2017 21898   Automatic 10839     Hybrid  0 470.8      0.0
## 12535   i3 2017 19980   Automatic 26965     Hybrid 140 470.8      0.0
## 13021   i3 2016 19490   Automatic  8421     Hybrid  0 470.8      0.0
## 13946   i3 2016 16482   Automatic 43695     Hybrid  0 470.8      0.0
## 14297   i3 2015 14285   Automatic 65800     Hybrid  0 470.8      0.0
## 14582   i3 2017 18500   Automatic 36429     Hybrid  0 470.8      0.0
## 14769   i3 2017 19495   Automatic 17338    Hybrid 135 470.8      0.0
## 14778   i3 2015 17481   Automatic 9886     Hybrid  0 470.8      0.0
## 14880   i3 2015 15498   Automatic 33931    Hybrid  0 470.8      0.0
## 15418   i3 2014 15450   Automatic 42479    Hybrid  0 470.8      0.0
## 15535   i3 2014 14495   Automatic 34539    Hybrid  0 470.8      0.0
## 15845   i3 2017 21444   Automatic 22063    Hybrid  0 470.8      0.0
## 16171   i3 2017 18995   Automatic 33021    Hybrid  0 470.8      0.0
## 16289   i3 2017 21490   Automatic 26139    Hybrid 135 470.8      0.0
## 16414   i3 2017 19000   Automatic 23983    Hybrid  0 470.8      0.0
```

```

## 16459 i3 2014 14182 Automatic 37161 Hybrid 0 470.8 0.0
## 16515 i3 2017 21500 Automatic 10900 Hybrid 140 470.8 0.0
## 16739 i3 2017 23751 Automatic 28169 Hybrid 0 470.8 0.0
## 16949 i3 2017 22999 Automatic 3976 Hybrid 135 470.8 0.0
## 17483 i3 2017 21494 Automatic 16867 Hybrid 135 470.8 0.0
## 17543 i3 2016 15990 Automatic 68000 Hybrid 0 470.8 0.0
## 17545 i3 2016 14900 Automatic 59945 Hybrid 0 470.8 0.0
## 17630 i3 2017 20495 Automatic 20082 Hybrid 135 470.8 0.0
## 17716 i3 2017 19948 Automatic 20929 Hybrid 135 470.8 0.0
## 17753 i3 2017 22495 Automatic 21025 Hybrid 0 470.8 0.0
## 17923 i3 2016 19875 Automatic 20013 Hybrid 0 470.8 0.0
## 17968 i3 2017 21495 Automatic 24041 Hybrid 0 470.8 0.0
## 18016 i3 2017 19895 Automatic 29851 Hybrid 0 470.8 0.0
## 18465 i3 2016 19850 Automatic 19995 Hybrid 0 470.8 0.0
## 18857 i3 2015 14940 Automatic 59000 Other 0 470.8 0.6
## 19044 i3 2017 18999 Automatic 20321 Electric 135 470.8 0.0
## 19069 i3 2016 18999 Automatic 9990 Electric 0 470.8 0.0
## 19289 i3 2017 19300 Automatic 32867 Other 0 470.8 0.6
## 19504 i3 2015 17400 Automatic 29465 Electric 0 470.8 1.0
## 19928 i3 2015 12500 Automatic 79830 Hybrid 0 470.8 0.0
## 20593 i3 2016 16500 Automatic 35446 Hybrid 0 470.8 0.0
## 20749 i3 2017 20000 Automatic 19178 Other 0 470.8 0.6
## 20755 i3 2017 19500 Automatic 23956 Other 135 470.8 0.6
## 20994 i3 2016 17000 Automatic 41063 Other 0 470.8 0.6
## 21199 i3 2017 17600 Automatic 50867 Other 135 470.8 0.6

##      manufacturer car_id
## 10701      BMW 10701
## 11290      BMW 11290
## 11447      BMW 11447
## 11449      BMW 11449
## 12535      BMW 12535
## 13021      BMW 13021
## 13946      BMW 13946
## 14297      BMW 14297
## 14582      BMW 14582
## 14769      BMW 14769
## 14778      BMW 14778
## 14880      BMW 14880
## 15418      BMW 15418
## 15535      BMW 15535
## 15845      BMW 15845
## 16171      BMW 16171
## 16289      BMW 16289
## 16414      BMW 16414
## 16459      BMW 16459
## 16515      BMW 16515
## 16739      BMW 16739
## 16949      BMW 16949
## 17483      BMW 17483
## 17543      BMW 17543
## 17545      BMW 17545
## 17630      BMW 17630
## 17716      BMW 17716
## 17753      BMW 17753

```

```

## 17923      BMW  17923
## 17968      BMW  17968
## 18016      BMW  18016
## 18465      BMW  18465
## 18857      BMW  18857
## 19044      BMW  19044
## 19069      BMW  19069
## 19289      BMW  19289
## 19504      BMW  19504
## 19928      BMW  19928
## 20593      BMW  20593
## 20749      BMW  20749
## 20755      BMW  20755
## 20994      BMW  20994
## 21199      BMW  21199

```

En aquest cas, sí que és correcte. Es tracta de cotxes híbrids i/o elèctrics, els quals consumeixen molt poc (amb un gal·ló, poden recorrer moltes milles). Per tant, no es consideren que siguin errors.

Respecte les altres variables, no s'observa cap fenòmen estrany (En aquest punt, es recorda que els criteris per a trobar errors són subjectius, per tant, altres respostes i possibilitats també poden ser correctes).

Per últim, es valida que s'han assignat correctament els NA's i quines són les variables més afectades:

```

n_missing <- colSums(is.na(df))
n_missing

```

```

##       model      year     price transmission      mileage      fuelType
##          0         1        0            2           1             0
##       tax      mpg engineSize manufacturer      car_id
##          0         0         0            0           0

```

**4. Are there any univariate extreme outliers? Check and determine them. Please, after detected, assign them as missing data if corresponds (please be aware). Show the final number of NA's by variable.(1 point)**

A continuació, es defineix la funció presentada en les diferents sessions de laboratori de l'assignatura:

```

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 ) }

```

Havent-la definit, s'utilitza per a detectar outliers univariants de la següent manera:

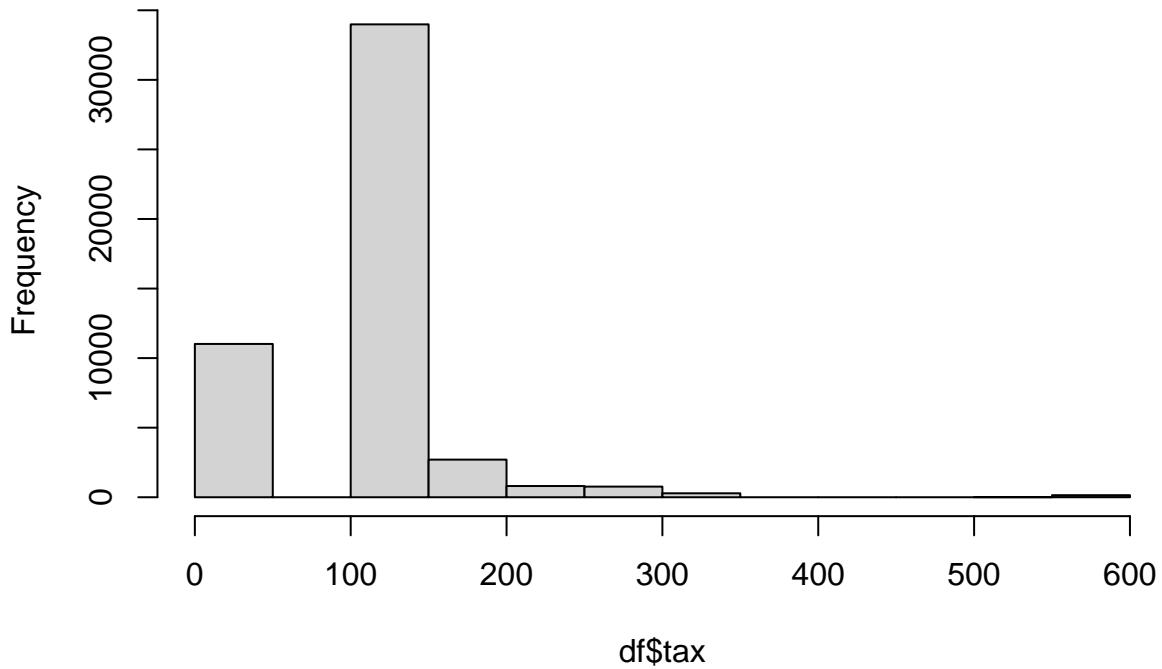
```

#Mileage:
var_out<-calcQ(df$mileage)
llout_mileage<-which((df$mileage<var_out$souti)|(df$mileage>var_out$souts))
df [llout_mileage,"mileage"] <- NA

#Tax:
var_out<-calcQ(df$tax)
llout_tax<-which((df$tax<var_out$souti)|(df$tax>var_out$souts))
# En aquest cas, es replanteja ja que es detecten masses observacions:
hist(df$tax)

```

## Histogram of df\$tax

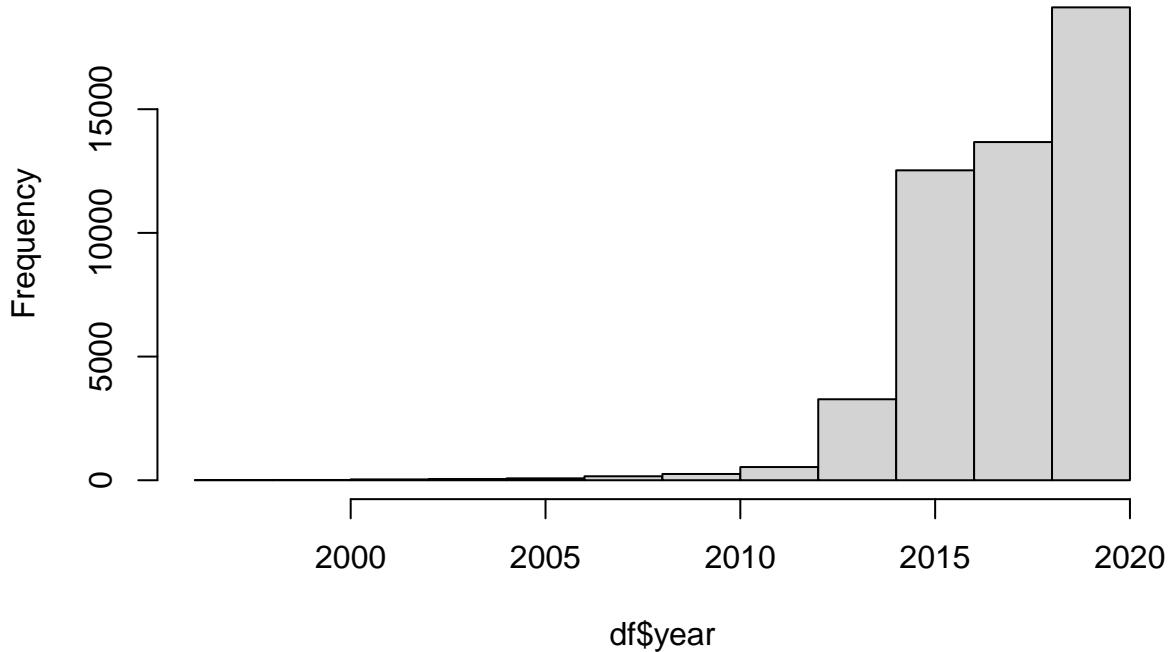


```
llout_tax<-which(df$tax>300)
df [llout_tax,"tax"] <- NA

#mpg:
var_out<-calcQ(df$mpg)
llout_mpg<-which((df$mpg<var_out$souti)|(df$mpg>var_out$souts))
df [llout_mpg,"mpg"] <- NA

#year:
var_out<-calcQ(df$year)
llout_year<-which((df$year<var_out$souti)|(df$year>var_out$souts))
# Degut a que és una variable estranya (any de fabricació),
# abans de considerar outliers, es mira la seva distribució de valors:
hist(df$year)
```

## Histogram of df\$year



```
# No s'observen valors anòmals, per tant, es deixarà tal i com està.

#engineSize:
var_out<-calcQ(df$engineSize)
llout_engineSize<-which((df$engineSize<var_out$souti)|(df$engineSize>var_out$souts))
df[llout_engineSize,"engineSize"] <- NA

#price:
var_out<-calcQ(df$price)
llout_price<-which((df$price<var_out$souti)|(df$price>var_out$souts))
# Degut a que és la variable target, no s'ha de realitzar cap tipus d'imputació.
```

Per últim, es miren el nombre de missings final (després de l'imputació d'errors i outliers univariants):

```
n_missing <- colSums(is.na(df))
n_missing
```

Variable	missings
model	0
year	1
price	0
transmission	2
mileage	192
fuelType	0
car_id	0
manufacturer	0
engineSize	583
mpg	512
tax	440

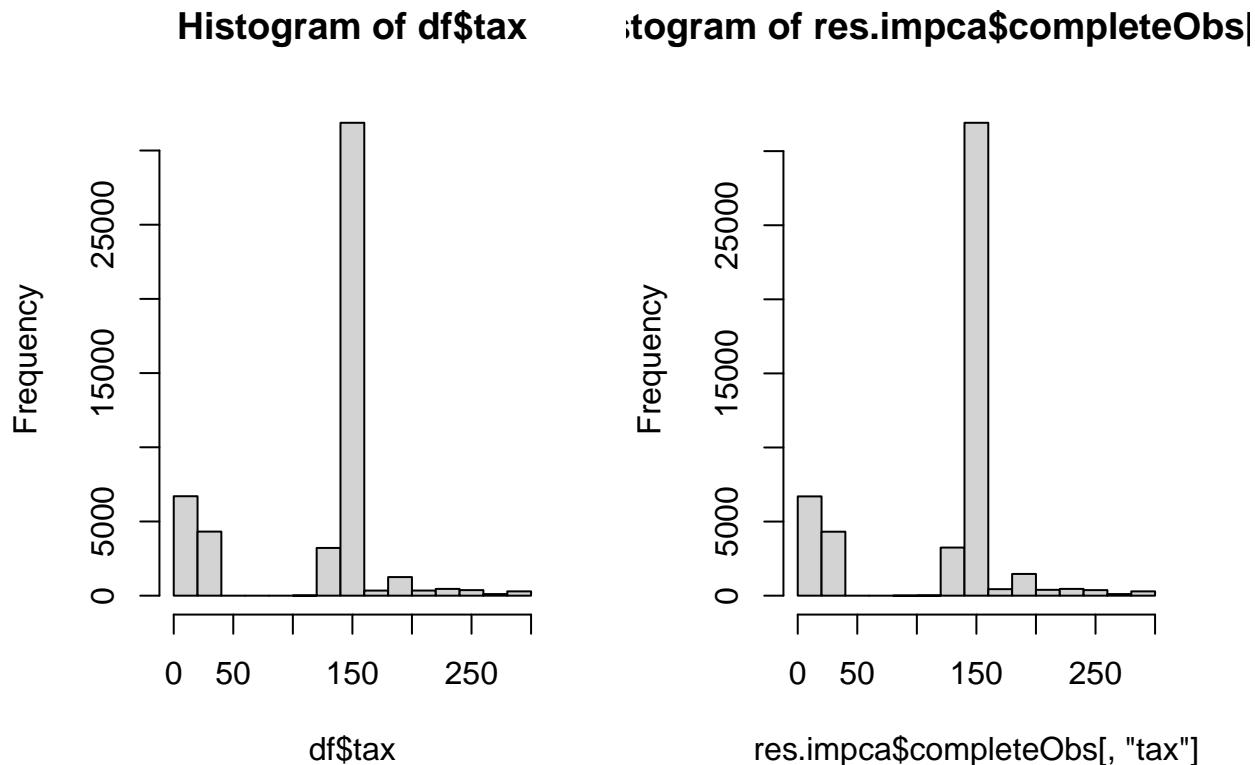
5. Impute missing data with imputePCA or imputeMCA as convenient. Check consistency of the imputation for a numeric variable and a categoric one. (1 point)

Imputació de variables numèriques:

```
library(missMDA)
res.impca<-imputePCA(df[,c("price", "mileage", "mpg", "tax", "year", "engineSize")],ncp=5)
```

Validació per a la variable tax:

```
par(mfrow=c(1,2))
hist(df$tax)
hist(res.impca$completeObs[, "tax"])
```



Assignació al dataframe original:

```
df[,c("price", "mileage", "mpg", "tax", "year", "engineSize")] <-
  res.impca$completeObs[,c("price", "mileage", "mpg", "tax", "year", "engineSize")]
```

Imputació de variables categòriques:

```
res.immca<-imputeMCA(df[,c("model", "fuelType", "manufacturer", "transmission")],ncp=10)
```

Validem que han desaparegut els missings de la variable transmission:

```
summary(df$transmission)
```

	Automatic	Manual	Other	Semi-Auto	NA's
##	13081	17757	0	18885	2

```
summary(res.immca$completeObs[, "transmission"])

## Automatic      Manual   Semi-Auto
##        13083     17757     18885
```

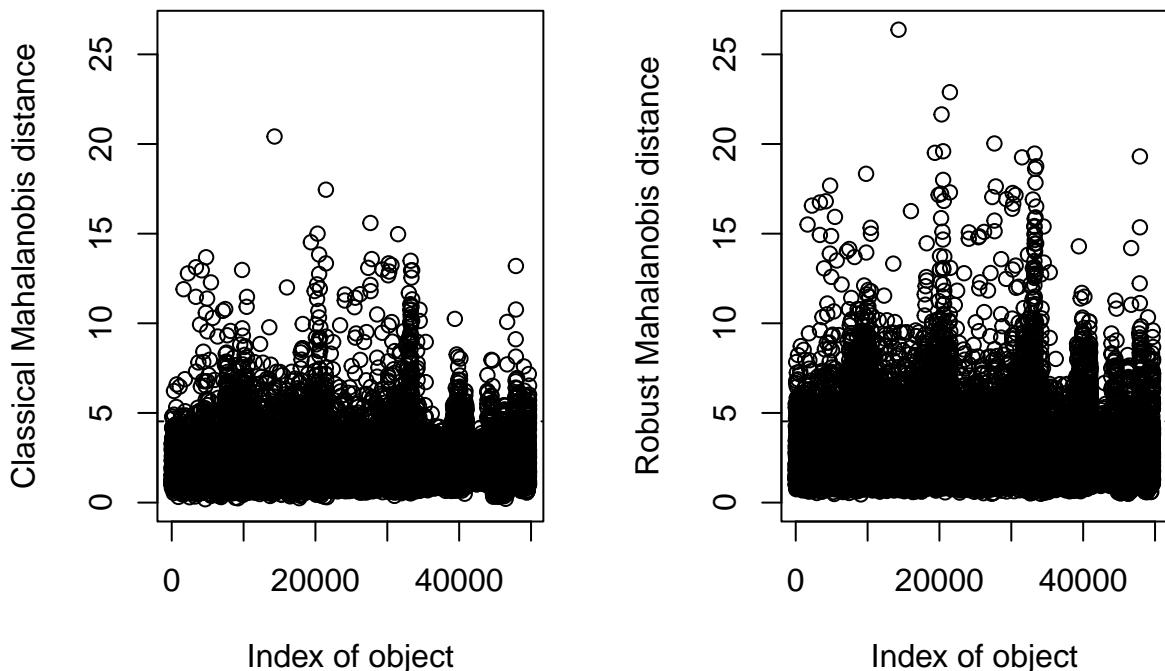
Assignació al dataframe original:

```
df[,c("model", "fuelType", "manufacturer", "transmission")] <-
  res.immca$completeObs[,c("model", "fuelType", "manufacturer", "transmission")]
```

6. Are there multivariate outliers? You will notice that a variable is causing problems to the algorithm, when calculating the mahalanobis distances, please remove it to calculate them correctly. Multivariate outliers, if present, are not going to be treated in this exercise: keep them as supplementary observations for the rest of the exercises and we will use them as it corresponds when performing the PCA. Additionally, provide the number of multivariate outliers detected. Which is the cut-off at 99.9% CI? (1 point)

Càlcul de les distàncies de Mahalanobis:

```
library(chemometrics)
mout<-Moutlier(df[,c("price", "mileage", "mpg", "year", "engineSize")],
                 quantile = 0.999, plot = TRUE)
```



El cutoff amb un nivell de confiança del 99,9% és el següent:

```
mout$cutoff
```

```
## [1] 4.529349
```

Per tant, fent servir les distàncies robustes calculades, trobem les següents observacions, les quals poden ser considerades outliers multivariants:

```
length(which(mout$rd > mout$cutoff))
```

```
## [1] 3868
```

Es tracta aproximadament de la meitat d'observacions, el qual no pot ser, per tant, veient la distribució de valors, es decideix redefinir el límit de la següent manera:

```
length(which(mout$rd > 25))
```

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

Per tant:

```
mo <- which(mout$rd > 25)
df$mout <- 0
df$mout[ mo ]<-1
df$mout <- factor( df$mout, labels=c( "NoMOut","YesMOut"))
table(df$mout)
```

```
##
##   NoMOut YesMOut
##     49724      1
```

7. Indicate by using exploratory data analysis tools which are apparently the most associated variables with the numeric response variable price. Use also FactoMineR profiling tools at 99% significance level. Comment on the results. Hint: You need to exclude car\_id when performing the profiling. (1 point)

Per tant, considerant que tenim un output numèric, caldria utilitzar la funció *condes* per a caracteritzar el meu *target*:

```
library(FactoMineR)
res.condes <- condes(df[,-c(11)], 3, proba = 0.01)
```

En aquest cas, el primer que mirarem seran les diferents correlacions amb les variables quantitatives:

```
res.condes$quanti
```

```
##           correlation p.value
## engineSize    0.5980187    0
## year          0.5536676    0
## tax            0.3803146    0
## mileage       -0.5206251    0
## mpg            -0.5848747    0
```

Veiem que sembla que totes les variables numèriques presenten una correlació estadísticament significativa amb el preu del cotxe. Concretament:

- A major valor d'enginesize, year o tax, el preu serà més elevat (correlació positiva)
- A menor mileage (km recorreguts) i mpg (menys milles recorregudes per gal · ló -> més consum), menor serà el preu.

A continuació, s'analitza l'equivalent a la correlació per a totes les variables categòriques:

```
res.condes$quali
```

```
##          R2      p.value
## model      0.490163449 0.000000e+00
## transmission 0.219390051 0.000000e+00
## manufacturer 0.081807565 0.000000e+00
## fuelType     0.003993016 6.368123e-42
## mout        0.001698703 3.773631e-20
```

Igual que per a les variables numèriques, sembla que totes les variables presenten una correlació amb el preu estadísticament diferent de 0. En aquest punt, es pot mencionar que quan es treballa amb mostres molt grans, aquests contrastos tendeixen a perdre potència pel qual resulta més difícil arribar a no rebutjar la hipòtesi nul · la.

Per últim, es mira quin és l'efecte individual de cadascuna de les categories de les diferents variables categòriques sobre el preu del cotxe:

```
df_categories <- as.data.frame(res.condes$category)
df_categories[order(df_categories$Estimate, decreasing = TRUE),]
```

```
##           Estimate      p.value
## model= G Class    71130.25910 3.487163e-162
## model= R8         69848.27338 6.182604e-293
## mout=YesMOut     50983.93119 3.773631e-20
## model= X7         42038.82273 7.620328e-232
## model= 8 Series   36193.85397 2.462730e-127
## model= Q8         32311.07359 1.117779e-185
## model= M5         29956.05910 1.279543e-69
## model= California 29789.92576 1.746229e-36
## model= i8          29208.64733 7.213974e-40
## model= RS6         28159.93089 2.985296e-84
## model= RS5         23461.26599 1.836459e-47
## model= RS4         22347.67200 4.978073e-47
## model= SQ7         21465.05910 1.400377e-12
## model= GLS Class   19415.59964 5.079432e-89
## model= S Class     17271.51087 3.796789e-198
## model= Q7          16984.37899 0.000000e+00
## model= X6          16034.67230 3.836537e-96
## model= M4          15470.29110 1.387634e-107
## model= M2          15336.39243 3.628431e-19
## model= GLE Class   12379.92244 4.145095e-293
## model= X5          11847.25568 8.136704e-281
## model= Caravelle   11521.72246 5.622684e-59
```

## model= GLB Class	9869.58541	2.010747e-10
## model= 7 Series	9130.37985	8.997215e-47
## model= A8	7177.90655	5.383610e-40
## model= RS3	6246.57425	7.674324e-11
## model= GLC Class	5125.46326	1.070776e-230
## model= Touareg	5078.30152	3.685127e-86
## model= X4	4917.39429	5.233993e-42
## model= V Class	4642.68228	4.347846e-46
## model= M6	4386.05910	6.371451e-03
## transmission=Semi-Auto	4350.55775	0.000000e+00
## fuelType=Hybrid	4147.71191	3.236908e-21
## model= SQ5	3611.87160	3.445829e-04
## model= S4	3444.14243	2.310123e-03
## model= SL CLASS	3413.50140	1.082193e-45
## manufacturer=Mercedes	2906.68612	0.000000e+00
## transmission=Automatic	2726.49793	4.502700e-266
## model= Q5	2641.74781	2.950590e-129
## model= M3	2425.83687	4.230316e-05
## model= X-CLASS	1891.16885	2.032171e-11
## manufacturer=Audi	1104.77424	1.696312e-49
## model= A7	1061.57549	1.932079e-13
## model= Tiguan Allspace	966.35580	3.694845e-10
## manufacturer=BMW	941.49807	1.525231e-39
## model= X2	682.72229	6.674484e-27
## fuelType=Diesel	439.41065	7.904133e-20
## model= Amarok	187.55459	6.341946e-10
## model= X3	-45.63056	1.225877e-40
## fuelType=Petrol	-614.94098	3.646073e-29
## model= Z4	-802.00572	2.353297e-07
## model= CLS Class	-1060.75103	2.706907e-13
## model= Arteon	-1518.92074	8.830900e-12
## model= E Class	-2322.52001	2.589830e-59
## model= 6 Series	-3449.57053	7.235240e-03
## model= C Class	-4108.11865	9.211596e-37
## model= A5	-4226.82639	1.727095e-08
## model= Q3	-4804.67908	2.044195e-07
## manufacturer=VW	-4952.95843	0.000000e+00
## model= T-Roc	-4964.54936	9.096726e-04
## model= A6	-5108.55588	2.756940e-03
## model= 5 Series	-5266.51287	1.931720e-03
## model= Q2	-5286.96524	7.458632e-03
## model= 4 Series	-5305.52281	3.783037e-03
## transmission=Manual	-7077.05568	0.000000e+00
## model= GLA Class	-7376.28447	4.934112e-03
## model= A4	-7548.49051	2.732159e-05
## model= 3 Series	-7923.08294	1.919451e-13
## model= A Class	-7954.34387	1.519725e-14
## model= X1	-7987.37623	1.613225e-05
## model= 2 Series	-8264.56906	4.294627e-10
## model= B Class	-8907.11349	1.081490e-08
## model= Touran	-10250.10852	2.363084e-11
## model= A3	-10395.41887	3.484497e-61
## model= M Class	-10438.00420	9.437812e-04
## model= Golf	-11114.25553	1.056837e-223

```

## model= Passat      -11119.25784  5.471150e-40
## model= Golf SV    -11922.91852  1.026507e-16
## model= 1 Series   -11982.27051  4.979486e-119
## model= A1          -13476.19035  3.124727e-128
## model= Scirocco   -14277.85413  4.034856e-29
## model= Polo        -16462.50829  0.000000e+00
## model= SLK          -17016.62511  4.704647e-21
## model= CC          -17158.29880  1.423783e-21
## model= Beetle      -17922.30235  1.367157e-21
## model= Jetta        -19178.81590  5.276479e-11
## model= Up           -19774.51330  5.889062e-294
## model= Eos          -21187.51233  3.891269e-04
## model= Z3           -21977.51233  1.871276e-04
## model= CLK          -24726.08376  1.126036e-05
## model= Fox          -26070.44090  3.683666e-04
## mout=NoMOut        -50983.93119  3.773631e-20

```

Veient aquests resultats, alguns dels possibles comentaris que pot fer l'estudiant són els següents:

- Tenim alguns models de cotxe que ens fan augmentar molt el preu del cotxe respecte el preu del cotxe mitjà (*baseline*). Entre aquests destaquen el G Class i el R8 (es podria mencionar l'X7).
- Tanmateix, sembla que els cotxes híbrids i els de transmissió automàtica i semiautomàtica tenen un preu més elevat que la resta de transmissions i combustibles.
- Respecte al fabricant, sembla que els audi, mercedes i BMW són més cars que els VW.
- Per altra banda, sembla que els de transmissió manual són unes 7000 lliures més baratos que el cotxe mitjà.
- Per últim, es poden observar una serie de models que són molt més barats que el cotxe mitjà (cotxe *baseline*). Entre aquests, destaquen el FOX, CLK, Z3...

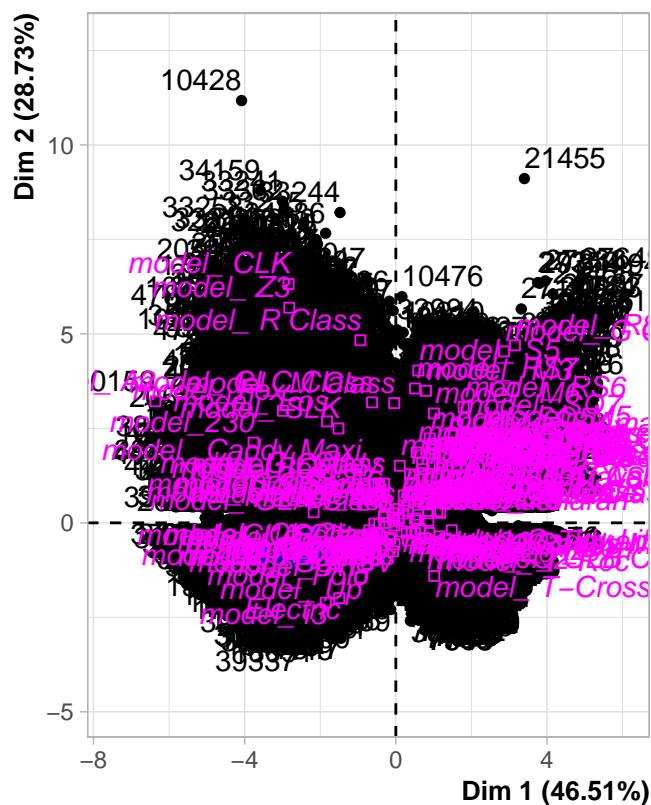
8. A Principal Component Analysis is addressed using as supplementary variables: price and all categorical variables. Multivariate outliers are set as supplementary observations. How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by the components retained Kaiser-based criteria? (0.5 points)

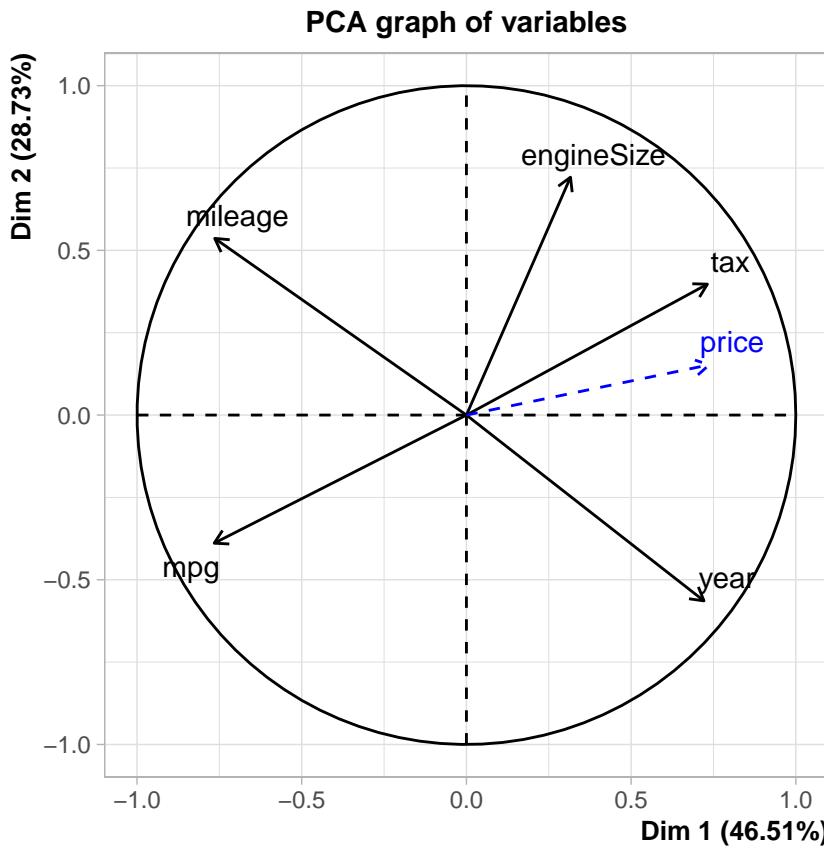
```

res.pca <- PCA(df[,-c(11,12)], quanti.sup = 3, quali.sup = c(1,4,6,10),
                 ind.sup = which(df$mout=="YesMOut"))

```

PCA graph of individuals





En primer lloc, cal mencionar que segons el criteri de *Kaiser*, ens hem de quedar amb **2** components principals (cal quedar-nos amb les components principals que tinguin un valor propi més gran que la mitjana de valors propis -> PCA normalitzat = 1):

```
res.pca$eig
```

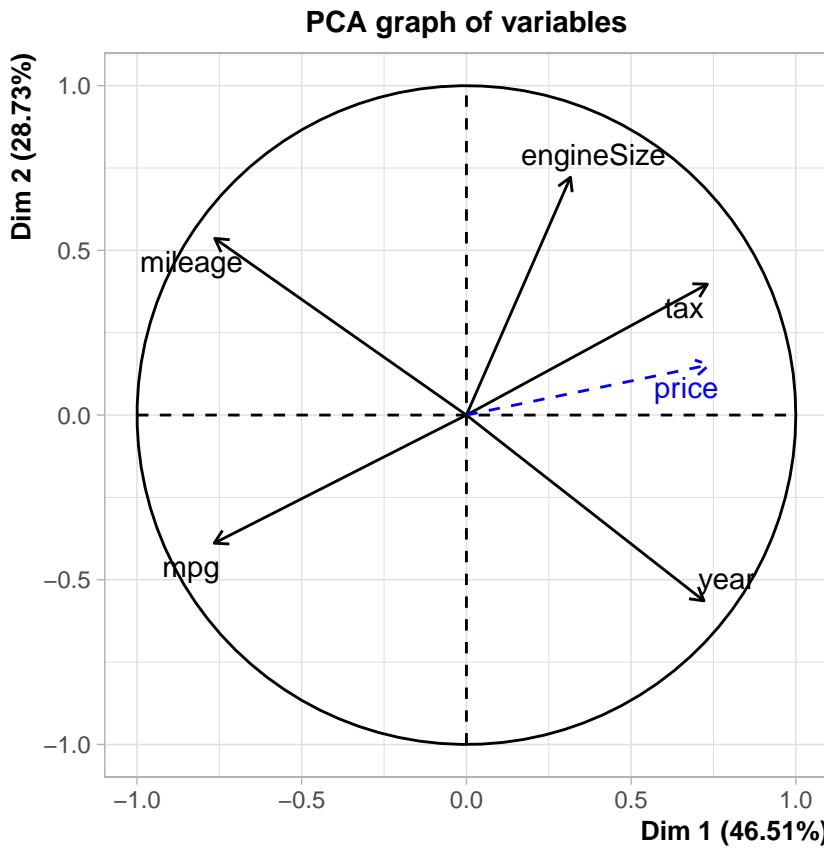
	eigenvalue	percentage of variance	cumulative percentage of variance
## comp 1	2.3254516	46.509033	46.50903
## comp 2	1.4366267	28.732534	75.24157
## comp 3	0.6451947	12.903895	88.14546
## comp 4	0.3768543	7.537085	95.68255
## comp 5	0.2158727	4.317453	100.00000

Tanmateix, cal mencionar que ens quedem amb un **75.37%** de la variabilitat total.

**9. Interpret the first factorial plane. Which variables are well represented in this first plane (interpret contributions and quality of representation)? Which are not? Relate the exposed results with the ones of the previous questions. Make use of plots to support your comments. (1.5 points)**

Respecte el PCA realitzat, el primer gràfic en el qual que ens fixem és el següent:

```
plot.PCA(res.pca, choix = c("var"))
```



Podem veure-hi que la variable target manté una correlació positiva amb la primera component. De la mateixa manera que ho fan les variables tax i year; per tant, com més alts són els valors d'aquestes variables (més nou és el cotxe i més taxes paga), més alt és el preu del cotxe (aquests resultats coincideixen amb els obtinguts amb el *profiling*).

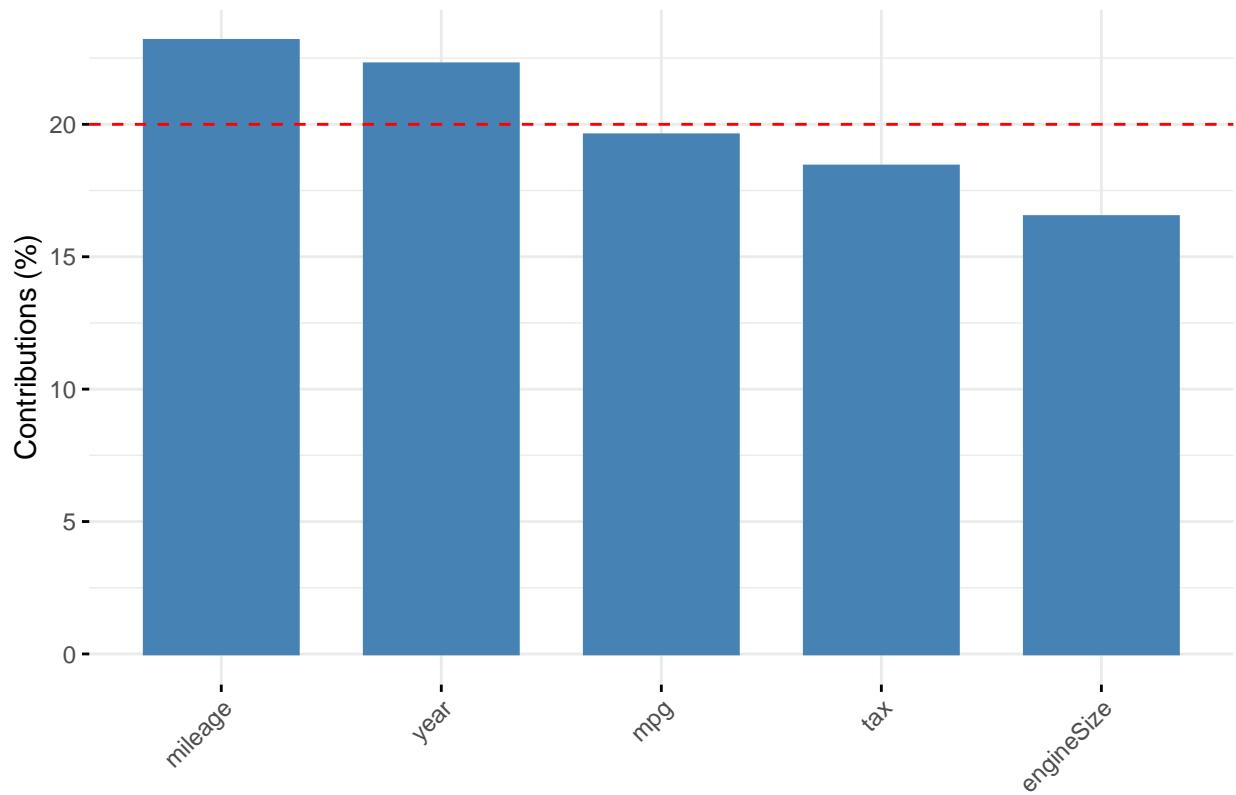
Per altra banda, mileage i mpg tenen una correlació negativa amb aquesta primera component principal. Per tant, es confirma el que hem comentat anteriorment: com més km recorreguts del cotxe i més consum, menor preu del cotxe.

Per últim, cal mencionar que sembla que engineSize és la variable menys correlacionada amb aquesta primera component principal. Es podria dir que, des d'un punt de vista del PCA, sembla ser la variable que afecta menys al preu.

A continuació, si ens centrem en les contribucions i les qualitats de representació de les diferents variables en aquestes dues primeres components principals:

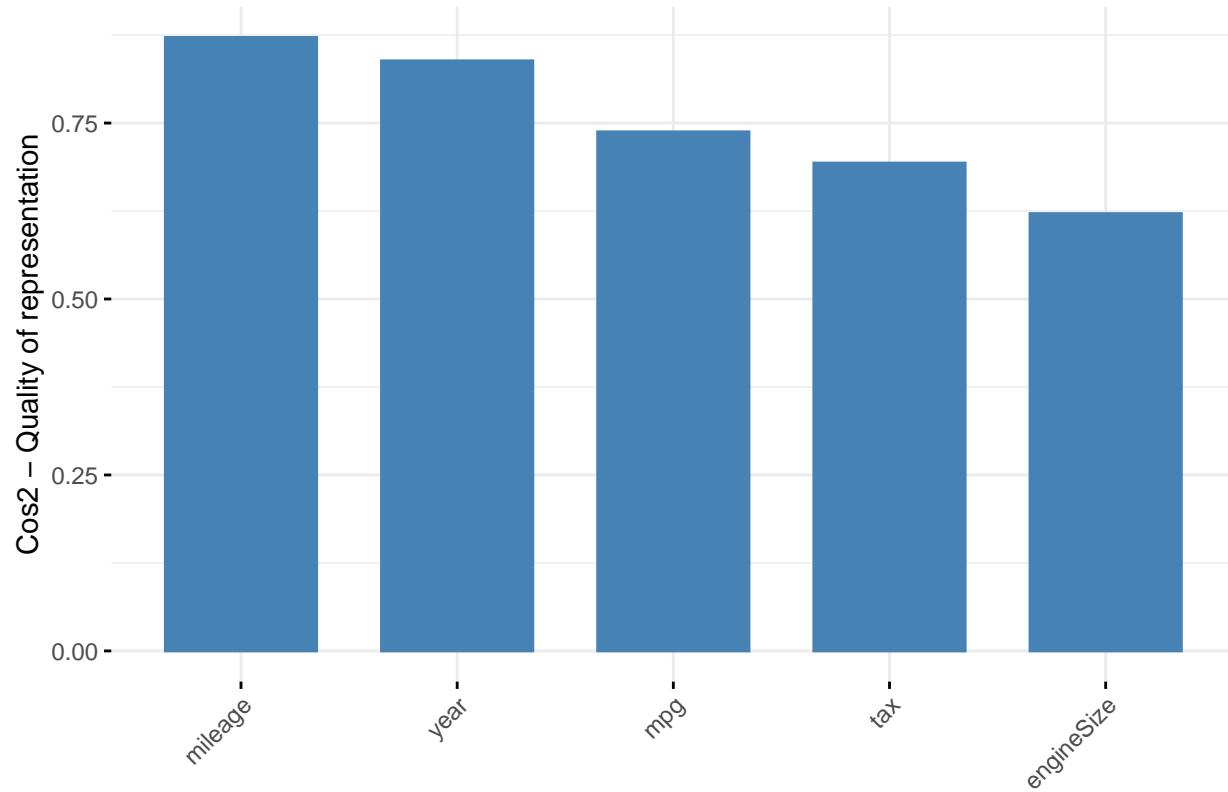
```
library(factoextra)
fviz_contrib(res.pca, choice = "var", axes = 1:2)
```

### Contribution of variables to Dim-1–2



```
fviz_cos2(res.pca, choice = "var", axes = 1:2)
```

## Cos2 of variables to Dim–1–2



Aquests gràfics són *self-explanatory*. Tot i així, s'espera que l'estudiant mencioni que el primer gràfic (o un de semblant) fa referència a la contribució de cadascuna de les variables en la creació de les dues primeres components principals. Tanmateix, la qualitat de representació fa referència la quantitat d'informació de cadascuna de les variables representada en aquestes dues primeres components principals.

L'estudiant també pot detallar mitjançant la següent taula (o una de semblant) com es reparteixen les contribucions i les qualitats de representació entre les dues primeres components principals.

```
res.pca$var$cos2
```

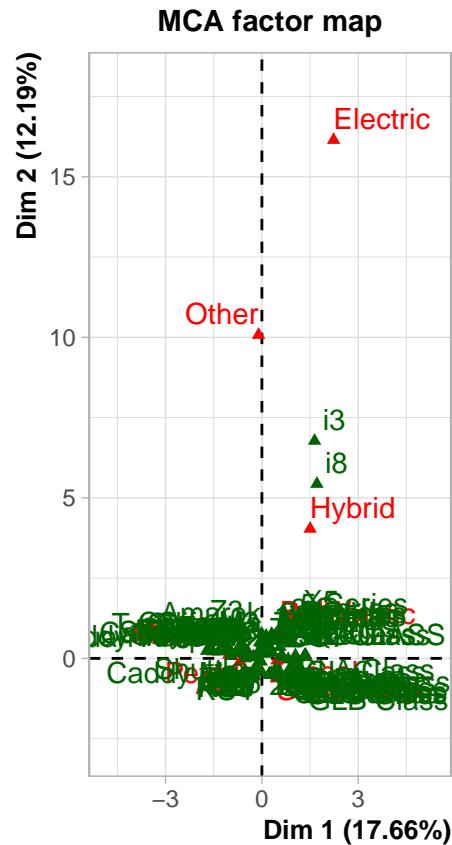
```
##           Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## year      0.52016391 0.3182276 0.05863423 0.006612801 0.096361465
## mileage   0.58396713 0.2876789 0.01386157 0.001627498 0.112864942
## tax        0.53522957 0.1579317 0.14011052 0.166563818 0.000164354
## mpg        0.58624142 0.1512664 0.05787394 0.199592287 0.005025983
## engineSize 0.09984961 0.5215221 0.37471448 0.002457851 0.001455925
```

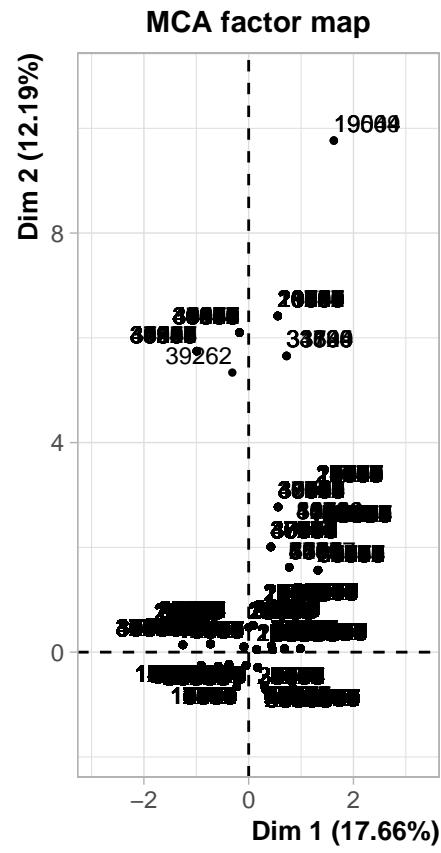
```
res.pca$var$contrib
```

```
##           Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## year      22.368296 22.15103  9.087835  1.7547371 44.63810334
## mileage   25.111988 20.02461  2.148433  0.4318639 52.28310841
## tax        23.016155 10.99323 21.716004 44.1984710  0.07613471
## mpg        25.209788 10.52927  8.969995 52.9627261  2.32821641
## engineSize 4.293773 36.30185 58.077733 0.6522019  0.67443713
```

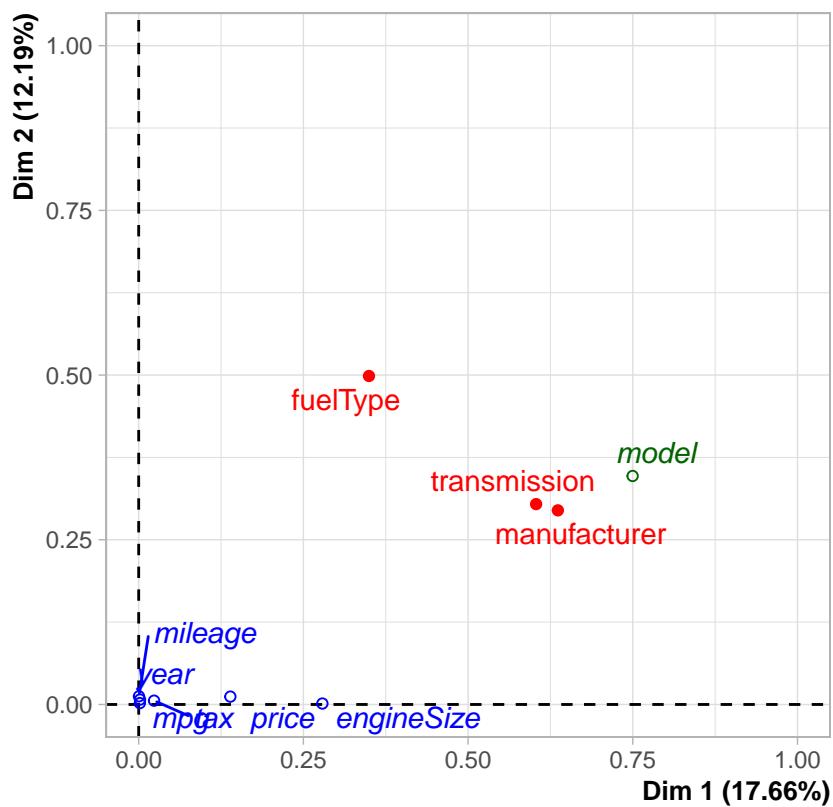
10. A Multiple Correspondence analysis is addressed using as supplementary variables: all numeric variables and model. Following a specific criteria (please explain it), how many axes would you retain? What is the category which has contributed more in the creation of the first factorial axes? After observing the first factorial axes, please provide a sentence which summarizes your findings. (1.5 points)

```
res.mca <- MCA(df[,-c(11,12)], quanti.sup=c(2,3,5,7,8,9), quali.sup = 1)
```

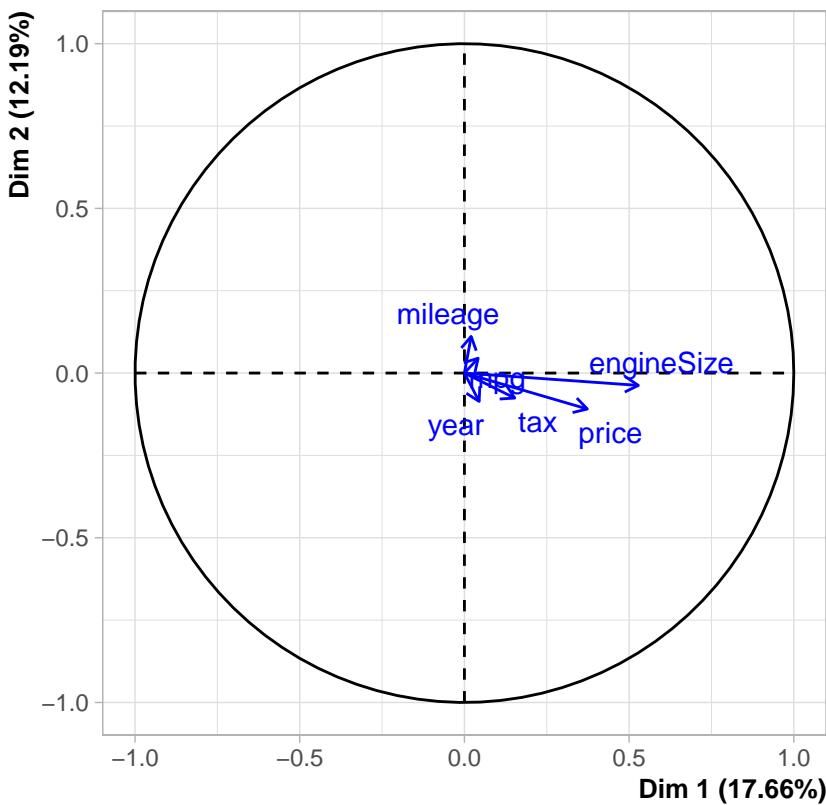




**Variables representation**



### Supplementary quantitative variables



Un cop realitzat l'ACM, es mira el percentatge de variabilitat retingut per cadascuna de les correspondències:

```
res.mca$eig
```

	eigenvalue	percentage of variance	cumulative percentage of variance
## dim 1	0.5297095	17.656983	17.65698
## dim 2	0.3657854	12.192846	29.84983
## dim 3	0.3447567	11.491890	41.34172
## dim 4	0.3387982	11.293275	52.63499
## dim 5	0.3325027	11.083423	63.71842
## dim 6	0.3189813	10.632710	74.35113
## dim 7	0.3115857	10.386189	84.73731
## dim 8	0.2699193	8.997312	93.73463
## dim 9	0.1879612	6.265374	100.00000

```
mean(res.mca$eig[,1])
```

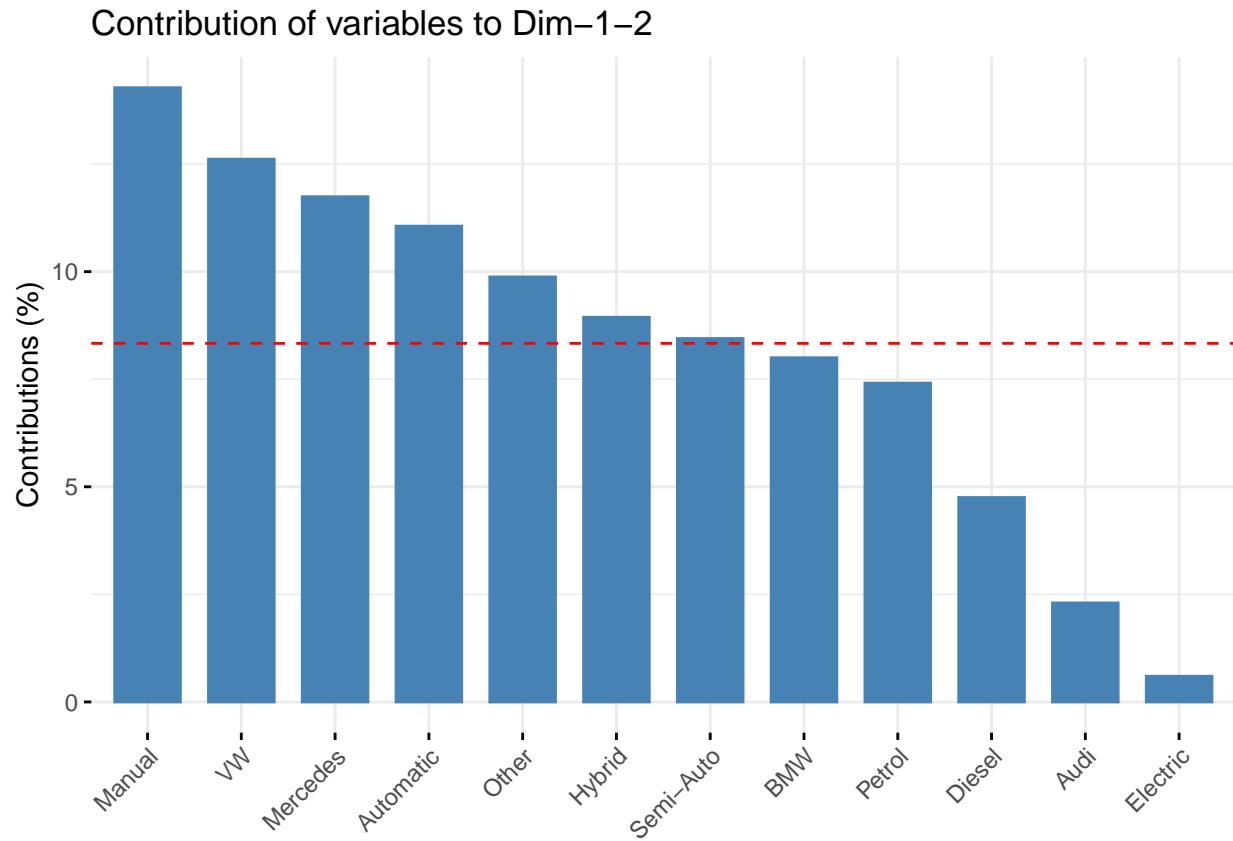
```
## [1] 0.3333333
```

Per a decidir el nombre de correspondències, es poden utilitzar els següents dos criteris:

- Quedar-nos amb el nombre de correspondències que retinguin almenys un 80%: 7 correspondències.
- Quedar-nos amb totes les correspondències que tinguin un valor propi major que la mitjana de valors propis: 4 correspondències.

Per altra banda, respondent a la pregunta de, quina és la categoria que ha contribuït més en la creació de les dues primeres correspondències, es pot veure en el següent gràfic que és la següent:

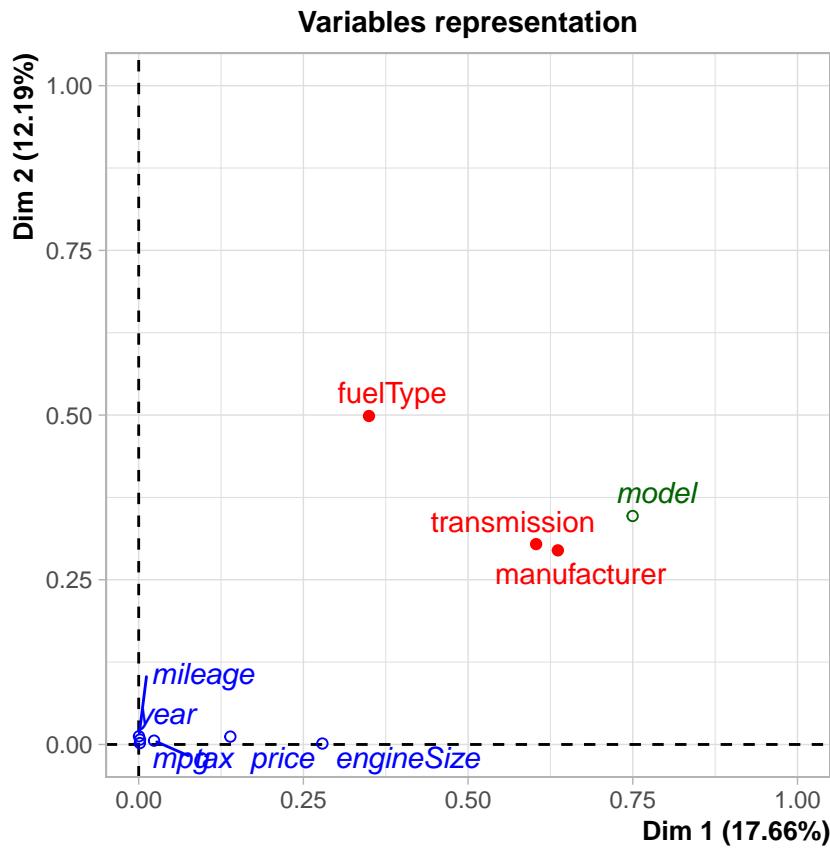
```
fviz_contrib(res.mca, choice = "var", axes = 1:2)
```



En aquest cas, es pot veure que és la categoria Manual de la variable transmission.

Per últim, observant el següent gràfic per a interpretar l'ACM:

```
plot.MCA(res.mca, choix = c("var"))
```



S'espera que l'estudiant digui algunes de les següents idees (o d'altres):

- La primera correspondència està més explicada per les variables transmission i manufacturer mentre que la segona està més definida per la variable tipus de combustible.
- La variable complementària model està més relacionada amb les variables de transmissió i fabricant que no amb el tipus de combustible.
- Observant el gràfic en el qual es visualitzen les categories de les diferents variables categòriques, sembla que els cotxes híbrids i elèctrics són més representatius i causen majors diferències entre els cotxes en la segona correspondència.