# USED CAR PRICES: CASE STUDY

Anàlisi de Dades i Explotació de la Informació

Joan Domingo Navarro

FIB - UPC Q1 21/22

# ÍNDEX

# 1. Data Processing, Description, Validation and Profiling

# 1.1. Data description

- Llistat de cotxes usats
- 5000 mostres aleatòries
- 10 variables inicials
  - 4 numériques
  - 6 categòriques
- Target numèric: Price
- Target categòric: Audi Si/No

# 1.1. Data description

```
> summary(df)
   model               year          price        transmission          mileage      
 Length:5000        Min.   :1999   Min.   :  1975   Length:5000        Min.   :     1  
 Class :character   1st Qu.:2016   1st Qu.: 13998   Class :character   1st Qu.:  5785  
 Mode  :character   Median :2017   Median : 19498   Mode  :character   Median : 16741  
                    Mean   :2017   Mean   : 21552                      Mean   : 23457  
                    3rd Qu.:2019   3rd Qu.: 26350                      3rd Qu.: 34125  
                    Max.   :2020   Max.   :139559                      Max.   :240494  
  fuelType              tax              mpg            engineSize    manufacturer      
 Length:5000        Min.   :  0.0   Min.   : 11.00   Min.   :0.000   Length:5000       
 Class :character   1st Qu.:125.0   1st Qu.: 44.80   1st Qu.:1.500   Class :character  
 Mode  :character   Median :145.0   Median : 53.30   Median :2.000   Mode  :character  
                    Mean   :125.5   Mean   : 53.81   Mean   :1.927                     
                    3rd Qu.:145.0   3rd Qu.: 61.40   3rd Qu.:2.000                     
                    Max.   :580.0   Max.   :470.80   Max.   :6.600                     
```

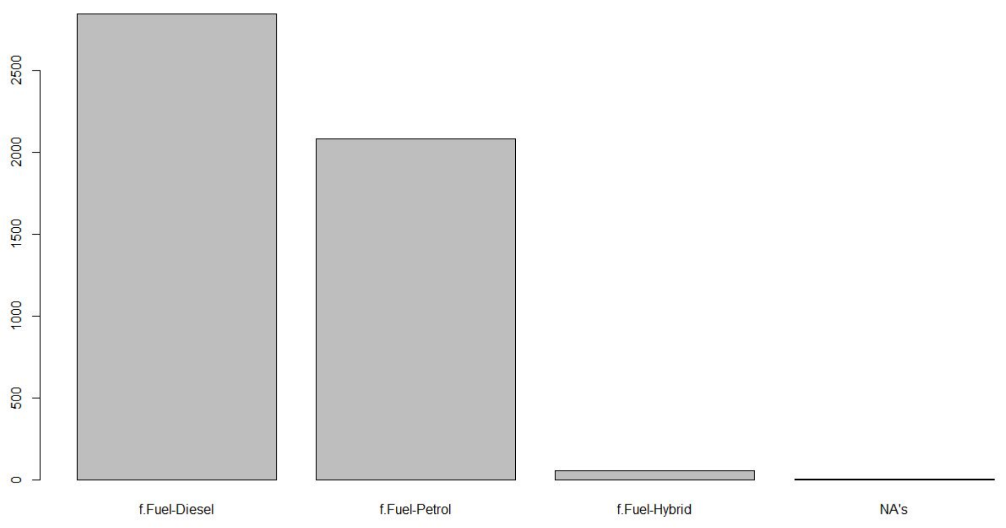
# 1.2. Variables qualitatives

- Convertim la variable year a categòrica
- Factoritzem totes les variables qualitatives

```
> summary(df[c("model", "year", "transmission", "fuelType", "manufacturer")])
              model              year               transmission
 VW- Golf         : 493    2019    :1620   f.Trans-Manual    :1780
 Mercedes- C Class: 373    2017    : 925   f.Trans-SemiAuto  :1979
 VW- Polo         : 337    2016    : 825   f.Trans-Automatic:1241
 Mercedes- A Class: 275    2018    : 461
 BMW- 3 Series    : 249    2015    : 395
 BMW- 1 Series    : 201    2020    : 325
 (Other)          :3072    (Other): 449
          fuelType        manufacturer
 f.Fuel-Diesel:2848    Audi    :1077
 f.Fuel-Petrol:2086    BMW     :1094
 f.Fuel-Hybrid:   58   Mercedes:1316
 NA's         :    8   VW      :1513
```

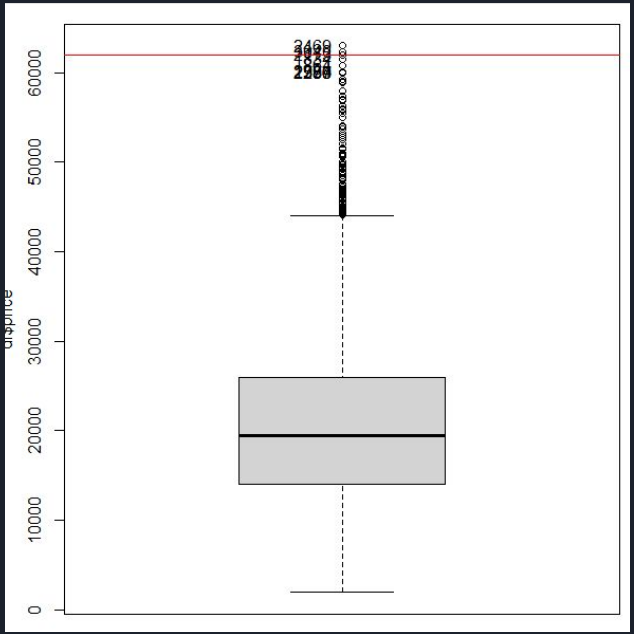
# 1.2. Variables qualitatives

- Convertim la variable year a categòrica
- Factoritzem totes les variables qualitatives
- Identifiquem errors i missing values



# 1.3. Variables quantitatives

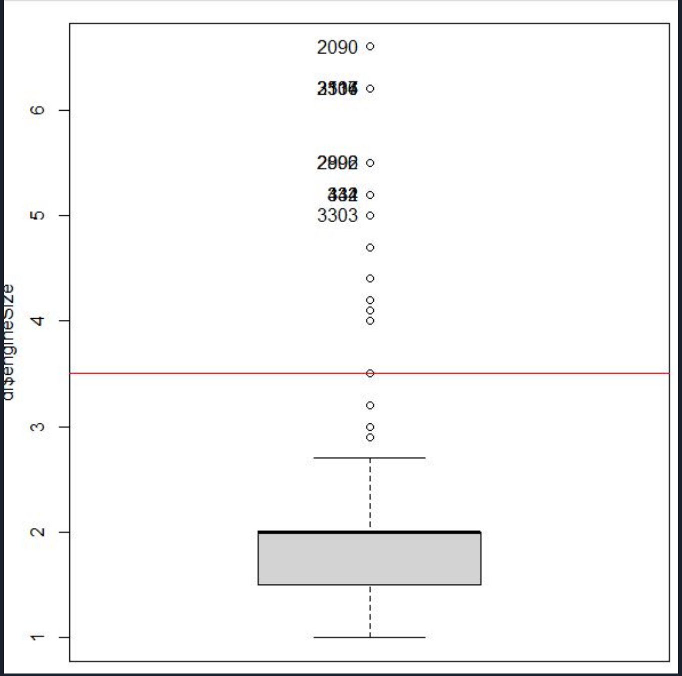
- Identifiquem errors i missing values
- Detecter outliers



# 1.3. Variables quantitatives

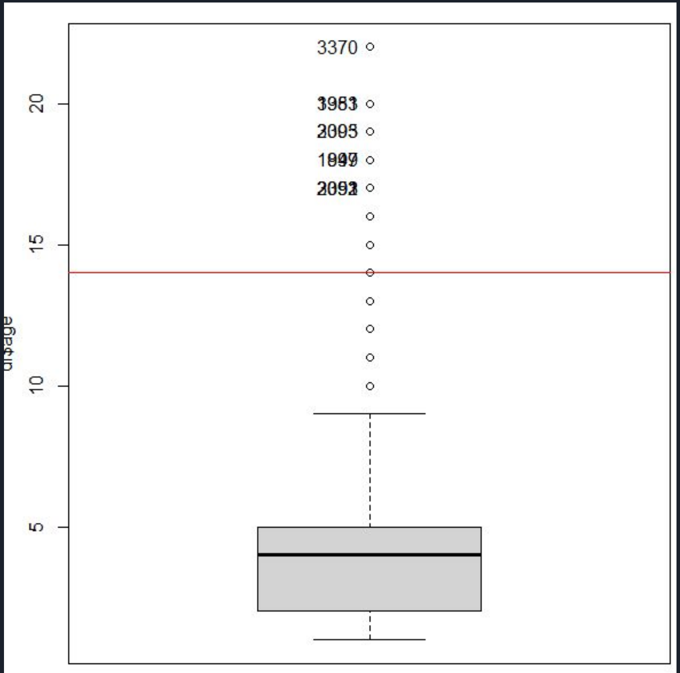
- Identifiquem errors i missing values
- Detecter outliers
- Convertim engineSize a factor

```
> summary(df$engineSize)
 Small Medium    Big   NA's 
  1400   2526    995     79 
```



# 1.3. Variables quantitatives

- Identifiquem errors i missing values
- Detecter outliers
- Convertim engineSize a factor
- Creem la variable age a partir de year



# 1.4. Data quality report

Missings:

```
[1] "tax  :  1275"
[1] "engineSize  :  79"
[1] "mpg  :  42"
[1] "price  :  40"
[1] "age  :  23"
[1] "mileage  :  20"
[1] "fuelType  :  8"
[1] "model  :  0"
[1] "year  :  0"
[1] "transmission  :  0"
[1] "manufacturer  :  0"
```

Errors:

```
[1] "engineSize  :  12"
[1] "model  :  0"
[1] "year  :  0"
[1] "price  :  0"
[1] "transmission  :  0"
[1] "mileage  :  0"
[1] "fuelType  :  0"
[1] "tax  :  0"
[1] "mpg  :  0"
[1] "manufacturer  :  0"
[1] "age  :  0"
```

Outliers:

```
[1] "tax  :  1275"
[1] "engineSize  :  67"
[1] "mpg  :  42"
[1] "price  :  40"
[1] "age  :  23"
[1] "mileage  :  20"
[1] "model  :  0"
[1] "year  :  0"
[1] "transmission  :  0"
[1] "fuelType  :  0"
[1] "manufacturer  :  0"
```

```
> missing_before <- total_m - total_o - total_e
> missing_before #els vists a fuelType
[1] 8
```

També detectem multivariate outliers i el emmagatzemem a una nova variable binaria

# 1.5. Imputació de missing

- Borrem les observacions amb NA’s de la variable target PRICE
- Imputem els missing de la resta de variables

```
> summary(df[,vars_dis])
                  model           year            transmission
 VW- Golf          : 493   2019    :1595   f.Trans-Manual   :1780
 Mercedes- C Class: 373   2017    : 925   f.Trans-SemiAuto :1951
 VW- Polo          : 337   2016    : 824   f.Trans-Automatic:1229
 Mercedes- A Class: 268   2018    : 457
 BMW- 3 Series     : 248   2015    : 395
 BMW- 1 Series     : 201   2020    : 315
 (Other)           :3040   (Other): 449
        fuelType      engineSize     manufacturer
 f.Fuel-Diesel:2830   Small :1400   Audi     :1067
 f.Fuel-Petrol:2066   Medium:2526   BMW      :1082
 f.Fuel-Hybrid:  56   Big   : 978   Mercedes:1298
 NA's         :   8   NA's  :  56   VW       :1513


> res.immca<-imputeMCA(df[,vars_dis],ncp=10)
> summary(res.immca$completeObs)
                  model           year            transmission
 VW- Golf          : 493   2019    :1595   f.Trans-Manual   :1780
 Mercedes- C Class: 373   2017    : 925   f.Trans-SemiAuto :1951
 VW- Polo          : 337   2016    : 824   f.Trans-Automatic:1229
 Mercedes- A Class: 268   2018    : 457
 BMW- 3 Series     : 248   2015    : 395
 BMW- 1 Series     : 201   2020    : 315
 (Other)           :3040   (Other): 449
        fuelType      engineSize     manufacturer
 f.Fuel-Diesel:2836   Small :1409   Audi     :1067
 f.Fuel-Petrol:2068   Medium:2553   BMW      :1082
 f.Fuel-Hybrid:  56   Big   : 998   Mercedes:1298
                                    VW       :1513
```

```
> summary(df[,vars_con])
     price          mileage             tax              mpg
 Min.   : 1975   Min.   :     1   Min.   :125.0   Min.   : 19.00
 1st Qu.:13995   1st Qu.:  5905   1st Qu.:145.0   1st Qu.: 44.80
 Median :19490   Median : 16869   Median :145.0   Median : 53.30
 Mean   :21080   Mean   : 23138   Mean   :147.1   Mean   : 52.95
 3rd Qu.:25997   3rd Qu.: 34026   3rd Qu.:145.0   3rd Qu.: 61.40
 Max.   :62980   Max.   :119000   Max.   :205.0   Max.   :100.90
                 NA's   :20       NA's   :1274    NA's   :40
      age
 Min.   : 1.000
 1st Qu.: 2.000
 Median : 4.000
 Mean   : 3.749
 3rd Qu.: 5.000
 Max.   :14.000
 NA's   :23
> res.impca<-imputePCA(df[,vars_con],ncp=4)
> summary(res.impca$completeObs)
     price          mileage             tax              mpg
 Min.   : 1975   Min.   :     1   Min.   :125.0   Min.   : 19.00
 1st Qu.:13995   1st Qu.:  5936   1st Qu.:145.0   1st Qu.: 45.60
 Median :19490   Median : 16994   Median :145.0   Median : 53.30
 Mean   :21080   Mean   : 23275   Mean   :147.3   Mean   : 52.95
 3rd Qu.:25997   3rd Qu.: 34228   3rd Qu.:148.2   3rd Qu.: 61.40
 Max.   :62980   Max.   :119000   Max.   :205.0   Max.   :100.90
      age
 Min.   : 1.000
 1st Qu.: 2.000
 Median : 4.000
 Mean   : 3.769
 3rd Qu.: 5.000
 Max.   :14.000
```

# 1.6. Discretization

- Discretitzem les variables numeriques per convertir-les a factors

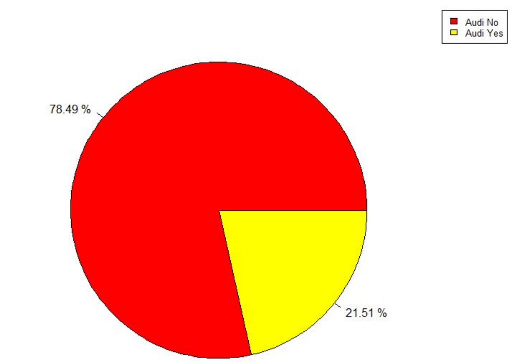
```
         f.price            f.miles                 f.tax                 f.mpg               f.age
f.price-[0,15] :1468   f.miles-[0,6]    :1287   f.tax-[0,125]  : 279   f.mpg-[0,45]   :1239   f.age-[0,2]     :1910
f.price-(15,20]:1228   f.miles-(6,18]   :1261   f.tax-(125,145]:2943   f.mpg-(45,54]  :1430   f.age-(2,4.1]   :1383
f.price-(20,26]:1035   f.miles-(18,36]  :1264   f.tax-(145,570]:1738   f.mpg-(54,62]  :1163   f.age-(4.1,5.1]: 825
f.price-(26,90]:1229   f.miles-(36,195]:1148                           f.mpg-(62,101]:1128   f.age-(5.1,15] : 842
```

# 1.7. Profiling

## Variable PRICE

```
> condes(df[,c(vars_res,vars_con,vars_dis)],3)
$quanti
        correlation       p.value
age       0.8016308   0.000000e+00
mpg       0.3304369 1.130881e-126
tax       0.1405900   2.564861e-23
price.1  -0.5597160   0.000000e+00
price    -0.5597160   0.000000e+00

$quali
                     R2        p.value
year        0.654968100 0.000000e+00
model       0.096690680 3.220007e-59
fuelType    0.048648606 2.080017e-54
transmission 0.048508064 2.999643e-54
engineSize  0.019672549 4.106100e-22
manufacturer 0.005974259 1.595680e-06
```

## Variable AUDI



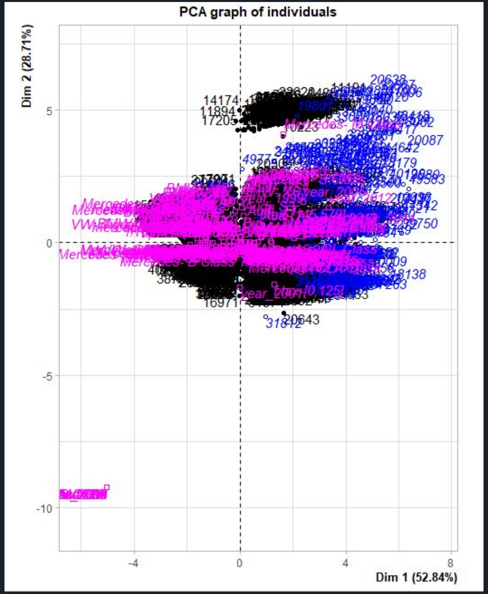

```
> catdes(df,18)

Link between the cluster variable and the categorical variables (chi-square test)
=================================================================================
                  p.value df
model        0.000000e+00 87
manufacturer 0.000000e+00  3
engineSize   2.963844e-20  2
f.mpg        5.001165e-17  3
f.miles      2.467586e-05  3
fuelType     9.706971e-05  2
f.price      5.015511e-04  3
transmission 2.730973e-03  2
aux          3.887650e-02  3
f.age        3.887650e-02  3
```

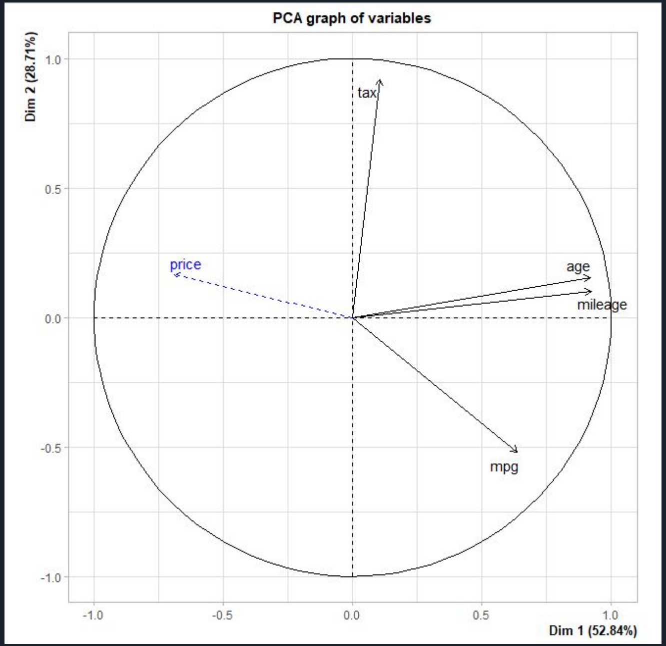
# 2. PCA & Clustering

# 2.1. PCA

Creació del PCA:

```r
ll <- which( df$mout == "YesMOut")
res.pca<-PCA(df[,c(vars_res, vars_dis, vars_con)],quali.sup=c(2:13),quanti.sup= c(1), ind.sup = ll )
```



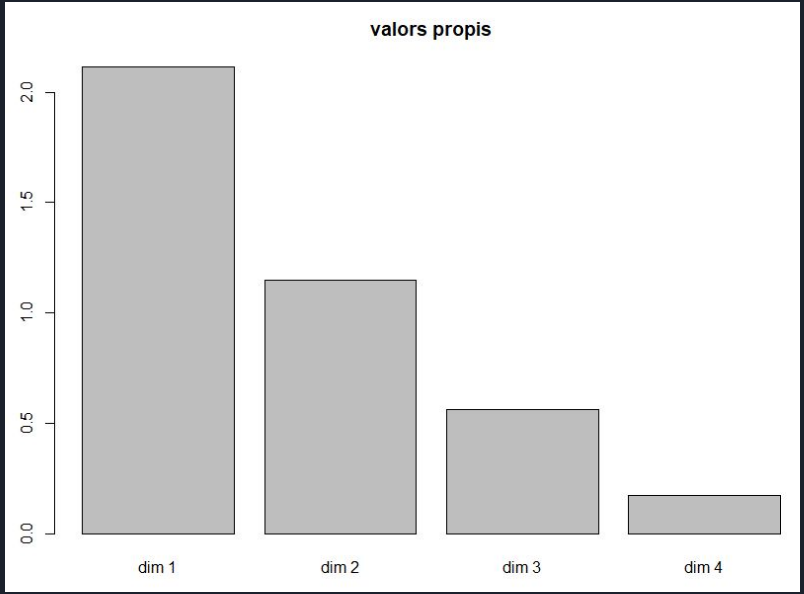
# 2.2. Valors propis i Regla de Kaiser

Mirant la taula de valors propis i seguint la regla de Kaiser, es tindran en compte les dues primeres component principals. Segons la Elbow rule, les dues o tres primeres. Representen el 81.55 percent de les dades.

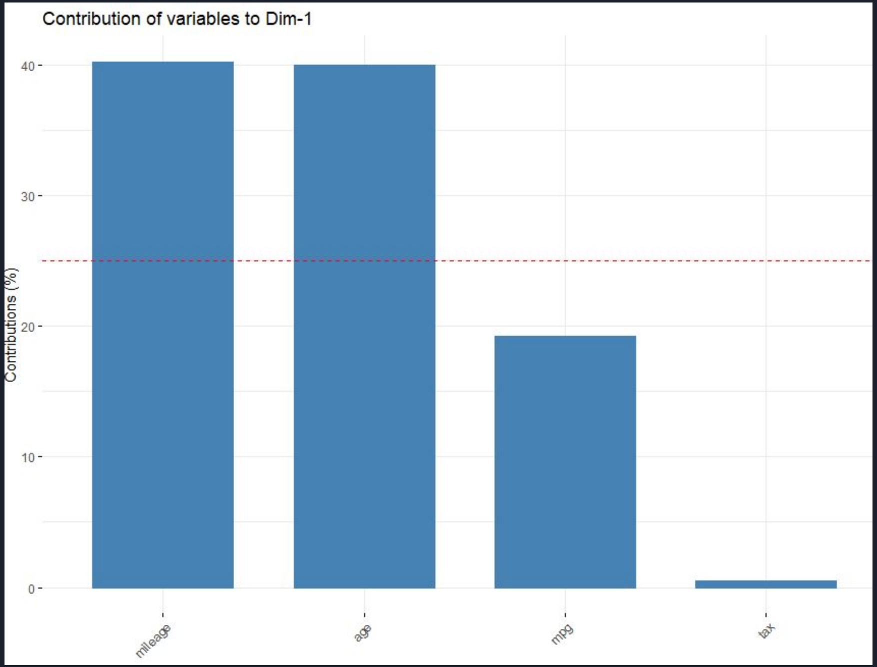
| | eigenvalue | percentage of variance | cumulative percentage of variance |
|---|---|---|---|
| comp 1 | 2.11 | 52.84 | 52.84 |
| comp 2 | 1.15 | 28.71 | 81.55 |
| comp 3 | 0.56 | 14.09 | 95.63 |
| comp 4 | 0.17 | 4.37 | 100.00 |

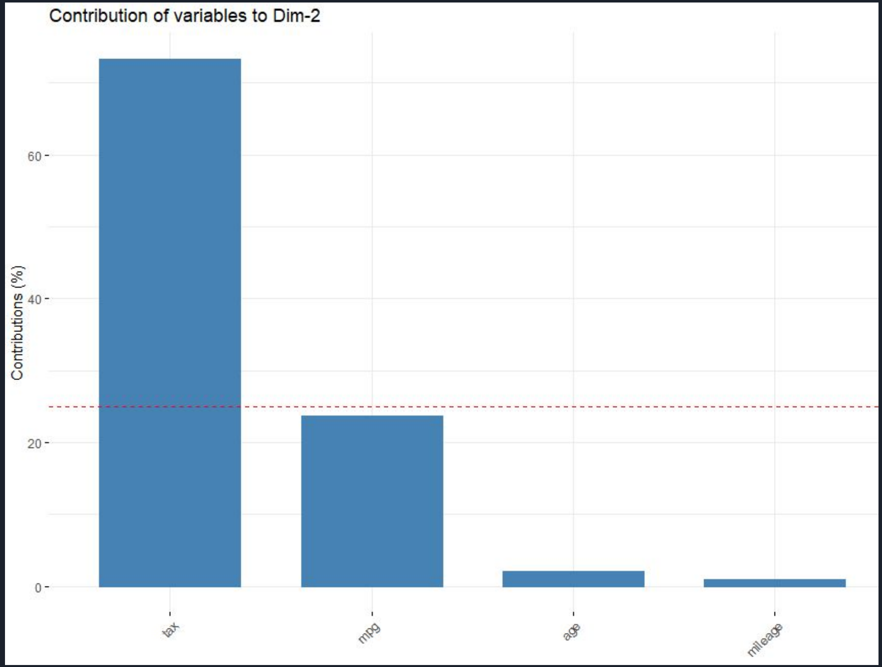
# 2.2. Valors propis i Regla de Kaiser

Mirant la taula de valors propis i seguint la regla de Kaiser, es tindran en compte les dues primeres component principals. Segons la Elbow rule, les dues o tres primeres. Representen el 81.55 percent de les dades.



# 2.2. Valors propis i Regla de Kaiser



# 2.3. K-Means Clustering

- Amb 14 clusters obtenim un 92% d’informació representada

```
> kc<-kmeans(dist(ppcc),14, iter.max = 30, trace=F)
> kc$betweenss/kc$totss
[1] 0.9226075
```

# 2.4. Hierarchical Clustering

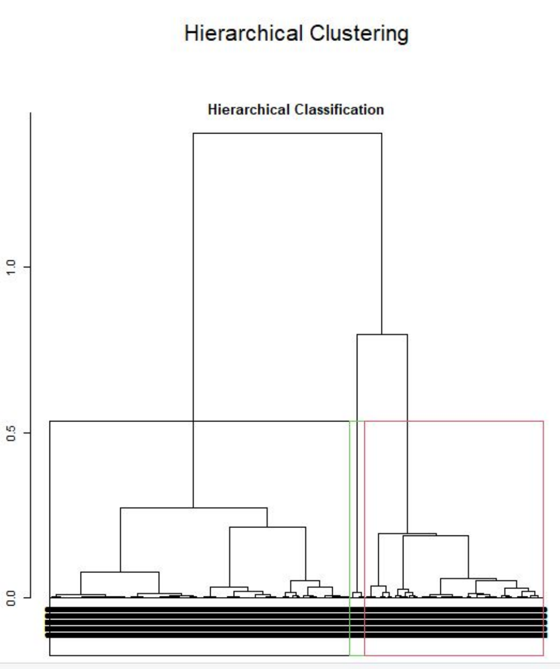
```
> res.hcpc$desc.var$test.chi2
                   p.value  df
year          0.000000e+00  28
f.price       0.000000e+00   6
f.miles       0.000000e+00   6
f.mpg         0.000000e+00   6
f.age         0.000000e+00   6
f.tax        3.935152e-260   4
model        2.609099e-143 164
transmission  4.515742e-74   4
fuelType      3.247317e-37   4
engineSize    6.878244e-23   4
manufacturer  3.070179e-08   6
Audi          6.886327e-06   2
```

```
> res.hcpc$desc.var$quanti.var
             Eta2 P-value
price   0.3677924       0
mileage 0.6019362       0
tax     0.7099223       0
mpg     0.3419401       0
age     0.6479452       0
```

```
> res.hcpc$desc.var$category
$`1`
                       Cla/Mod    Mod/Cla     Global      p.value    v.test
f.age=f.age-[0,2]   99.9476166 76.68810289 40.2997678 0.000000e+00       Inf
f.miles=f.miles-[0,6] 99.7668998 51.60771704 27.1690944 0.000000e+00     Inf
year=year_2019      99.9372647 64.02733119 33.6499894 0.000000e+00       Inf
```

```
$`2`
                          Cla/Mod     Mod/Cla     Global       p.value    v.test
f.price=f.price-[0,15]    86.94638695 53.10868533 27.1690944 1.274980e-301 37.121274
f.miles=f.miles-(36,195]  93.27641409 41.48077836 19.7804518 5.120846e-275 35.433976
f.miles=f.miles-(18,36]   79.80922099 47.65068818 26.5568925 6.430720e-197 29.937340
f.mpg=f.mpg-(62,101]      82.80802292 41.14855244 22.1025966 9.268256e-184 28.908751
year=year_2016            88.93229167 32.41575700 16.2127929 3.160859e-174 28.140175
```

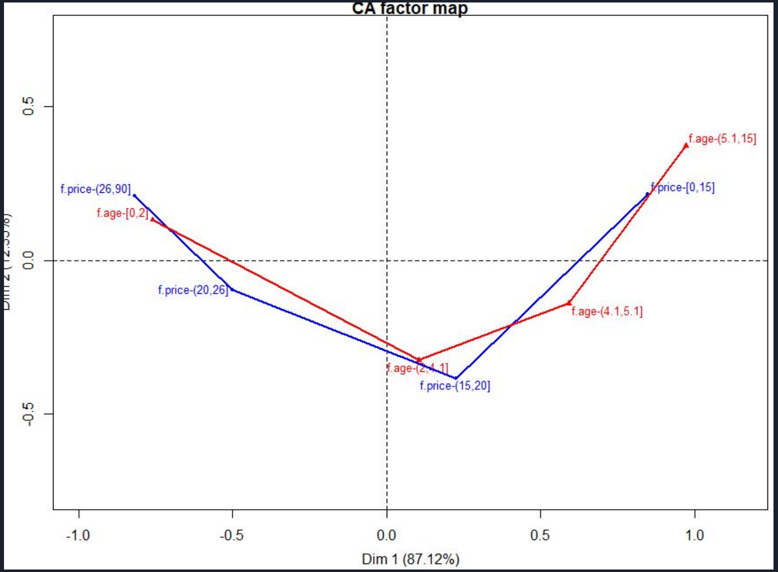
```
$`3`
                          Cla/Mod    Mod/Cla     Global      p.value    v.test
f.tax=f.tax-(145,570]    8.947700 100.000000 33.5022166 4.599373e-70 17.694791
f.age=f.age-(5.1,15]     9.339080  45.774648 14.6928436 2.152164e-19  9.005229
year=year_2015          11.080332  28.169014  7.6208571 8.032479e-14  7.469787
model=Audi- Q5          26.027397  13.380282  1.5410597 1.477102e-13  7.389204
year=year_2016           7.421875  40.140845 16.2127929 4.460407e-12  6.921768
```



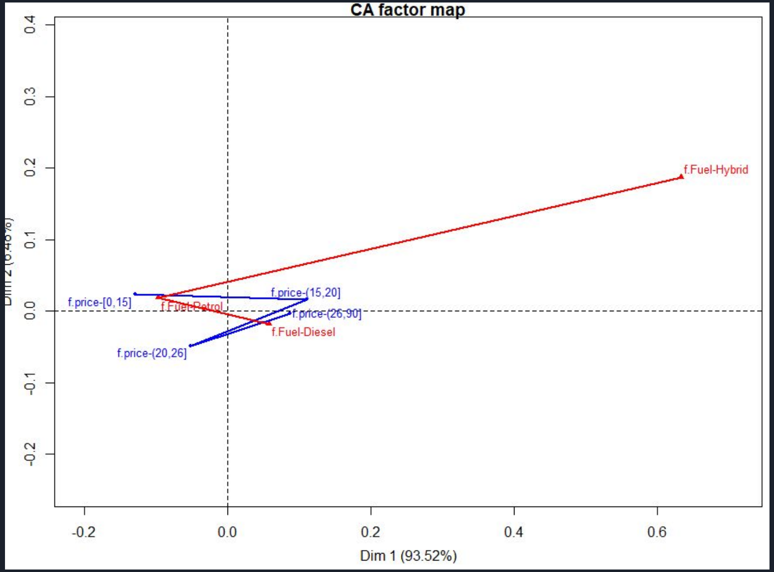
# 3. CA, MCA & Clustering

# 3.1. Correspondence analysis

Fem CA sobre f.price - f.age i sobre f.price - fuelType




X-squared = 2528.1, df = 9, p-value < 2.2e-16

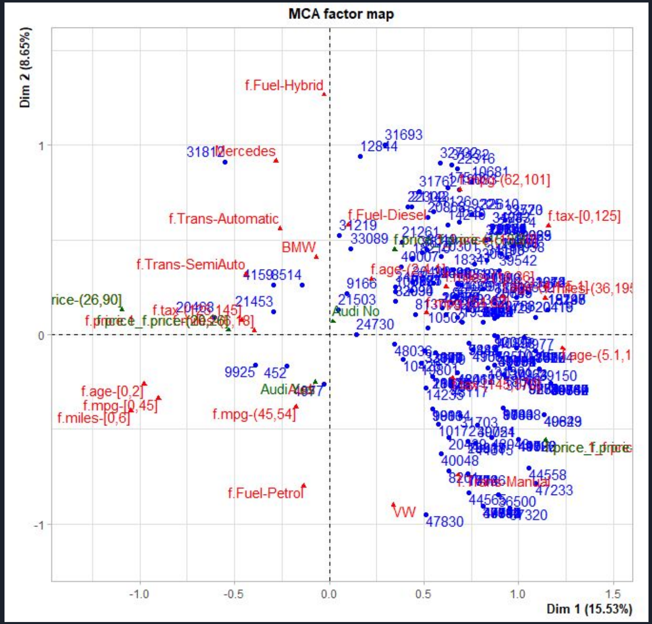



X-squared = 54.864, df = 6, p-value = 4.937e-10

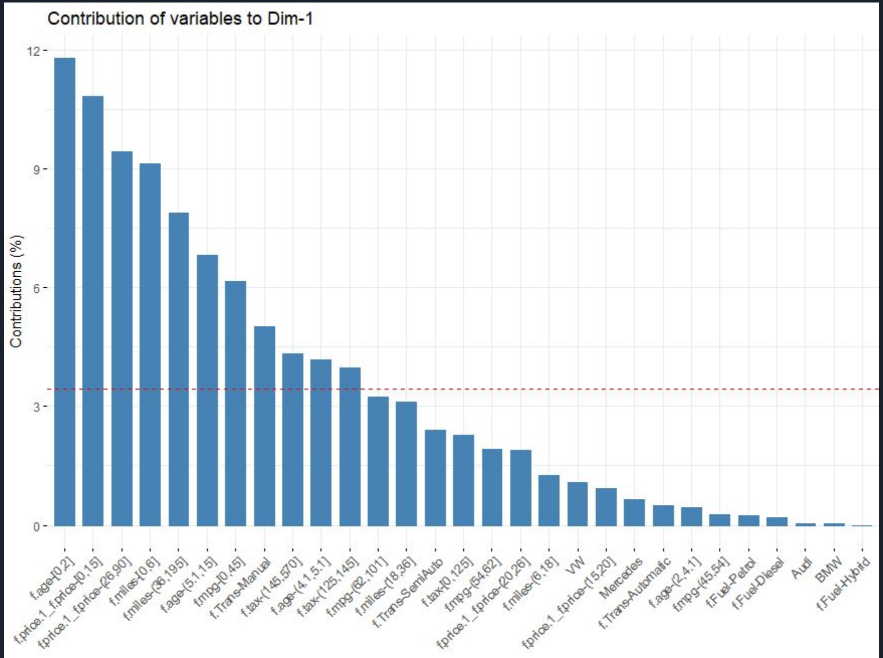
# 3.2. Multiple Correspondence Analysis

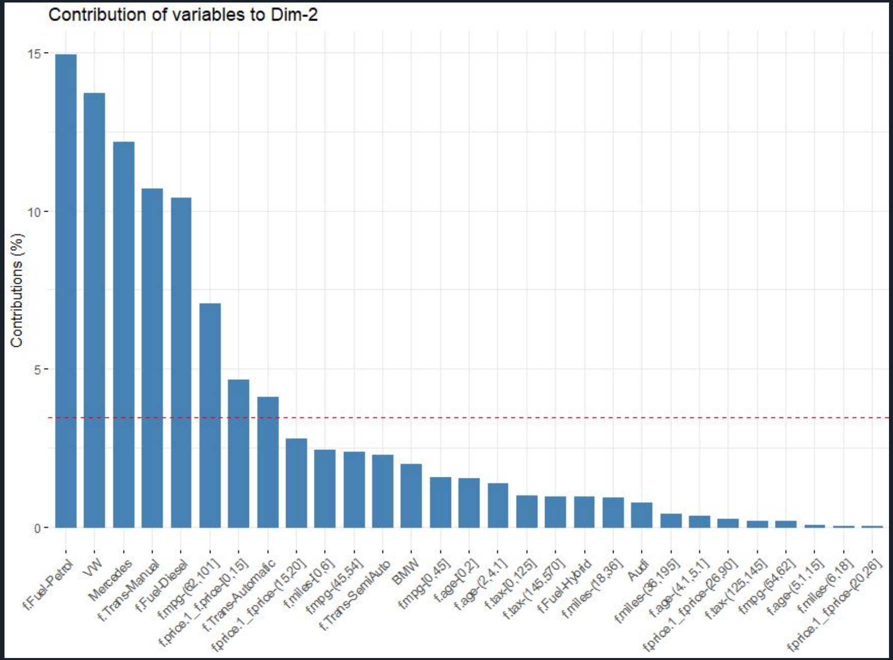
Agafem les dimensions que tenen el valor propi més gran que la mitjana d'aquests

```
> which(res.mca$eig[,1] > mean(res.mca$eig[,1]))
dim 1 dim 2 dim 3 dim 4 dim 5 dim 6 dim 7 dim 8 dim 9
    1     2     3     4     5     6     7     8     9
```



# 3.2. Multiple Correspondence Analysis

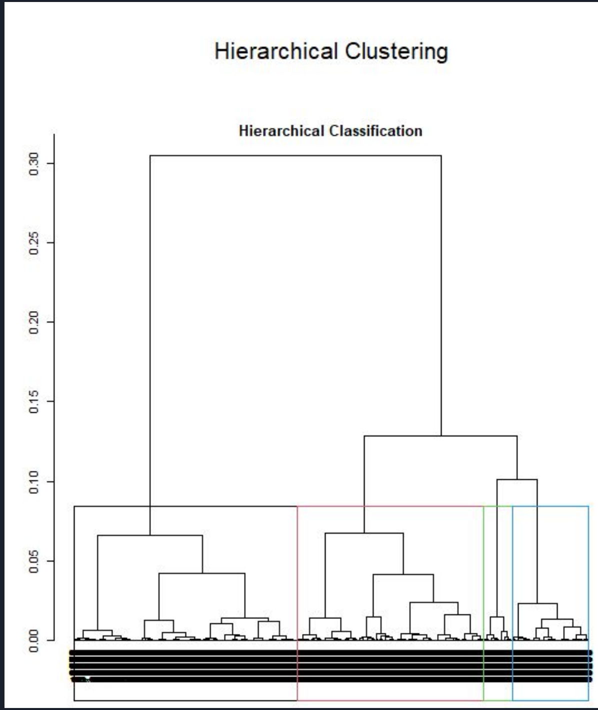


# 3.3. Hierarchical Clustering (MCA)

```
> res.hcmc$desc.var$test.chi2
                  p.value df
f.price      0.000000e+00  9
f.price.1    0.000000e+00  9
f.miles      0.000000e+00  9
f.tax        0.000000e+00  6
f.mpg        0.000000e+00  9
f.age        0.000000e+00  9
transmission 1.943284e-96  6
fuelType     2.180553e-26  6
manufacturer 7.551773e-21  9
```

```
> res.hcmc$desc.var$quanti.var
             Eta2       P-value
mileage 0.7024493  0.000000e+00
mpg     0.2793801  0.000000e+00
age     0.7498242  0.000000e+00
price   0.4453126  0.000000e+00
tax     0.2649183 1.589976e-315
```




```
> res.hcmc$desc.var$category
$`1`
                                     Cla/Mod    Mod/Cla    Global      p.value    v.test
f.age=f.age-[0,2]                 96.0712415 91.8837675 40.308277 0.000000e+00       Inf
f.miles=f.miles-[0,6]             98.1351981 63.2765531 27.174831 0.000000e+00       Inf
f.price.1=f.price.1_f.price-(26,90] 88.9071487 54.2084168 25.696791 0.000000e+00     Inf
f.price=f.price_f.price-(26,90]   88.9071487 54.2084168 25.696791 0.000000e+00       Inf
```

```
$`2`
                                      Cla/Mod   Mod/Cla    Global       p.value    v.test
f.age=f.age-(2,4.1]                 80.924431 68.170581 28.779561  0.000000e+00       Inf
f.miles=f.miles-(18,36]             78.775835 61.248455 26.562500  0.000000e+00       Inf
f.price.1=f.price.1_f.price-(15,20] 60.033306 44.561187 25.358953 3.143172e-102 21.467369
```

```
$`3`
                           Cla/Mod    Mod/Cla    Global      p.value    v.test
f.tax=f.tax-[0,125]    99.61832061 93.2142857  5.532095 0.000000e+00       Inf
f.mpg=f.mpg-(54,62]    16.87388988 67.8571429 23.775338 7.031478e-59 16.179543
f.miles=f.miles-(36,195] 16.66666667 55.7142857 19.763514 2.760629e-43 13.794236
f.age=f.age-(4.1,5.1]  15.60468140 42.8571429 16.237331 7.203238e-28 10.942688
```

```
$`4`
                                     Cla/Mod    Mod/Cla    Global       p.value    v.test
f.age=f.age-(5.1,15]              77.5539568 64.0142518 14.674831  0.000000e+00       Inf
f.miles=f.miles-(36,195]          70.8333333 78.7410926 19.763514  0.000000e+00       Inf
f.price.1=f.price.1_f.price-[0,15] 50.5058366 77.0783848 27.132601 9.731685e-255 34.091879
```

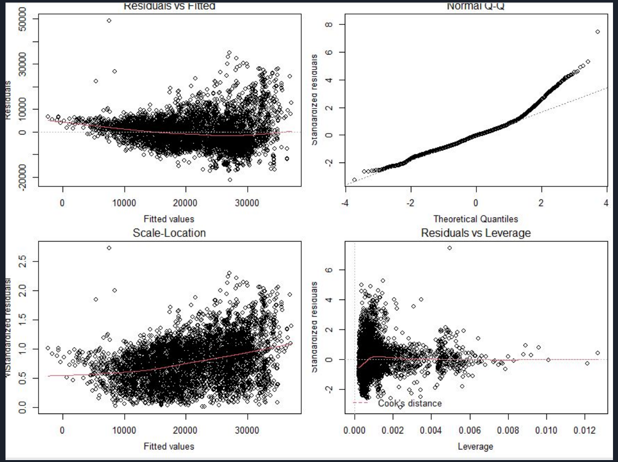
# 4. Modelling of the numeric target

# 4.1. Initial model

Model inicial :

```
lm(formula = price ~ tax + mpg + age + mileage, data = df)
```

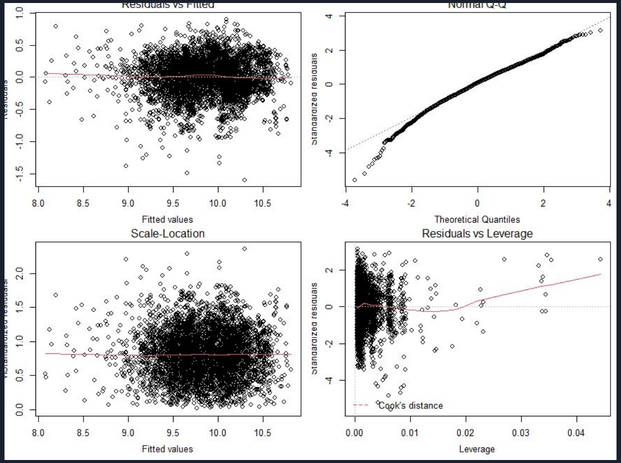
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.872e+04  1.450e+03  33.595  < 2e-16 ***
tax          1.795e+00  8.711e+00   0.206    0.837    
mpg         -3.654e+02  9.185e+00 -39.777  < 2e-16 ***
age         -2.051e+03  7.923e+01 -25.882  < 2e-16 ***
mileage     -3.560e-02  7.345e-03  -4.847 1.29e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6633 on 4955 degrees of freedom
Multiple R-squared:  0.5427,	Adjusted R-squared:  0.5423 
F-statistic:  1470 on 4 and 4955 DF,  p-value: < 2.2e-16
```




Amb transformacions:

```
lm(formula = log(price) ~ poly(tax, 2) + poly(mpg, 2) + poly(age, 
    2) + sqrt(mileage), data = df)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.976e+00  1.506e-02 662.320  < 2e-16 ***
poly(tax, 2)1  1.453e+00  3.095e-01   4.694 2.75e-06 ***
poly(tax, 2)2  2.238e+00  2.970e-01   7.537 5.71e-14 ***
poly(mpg, 2)1 -1.195e+01  3.401e-01 -35.126  < 2e-16 ***
poly(mpg, 2)2  4.603e+00  2.915e-01  15.793  < 2e-16 ***
poly(age, 2)1 -1.685e+01  5.398e-01 -31.213  < 2e-16 ***
poly(age, 2)2 -3.457e+00  3.275e-01 -10.555  < 2e-16 ***
sqrt(mileage) -9.382e-04  1.080e-04  -8.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2869 on 4952 degrees of freedom
Multiple R-squared:  0.6311,	Adjusted R-squared:  0.6306 
F-statistic:  1210 on 7 and 4952 DF,  p-value: < 2.2e-16
```

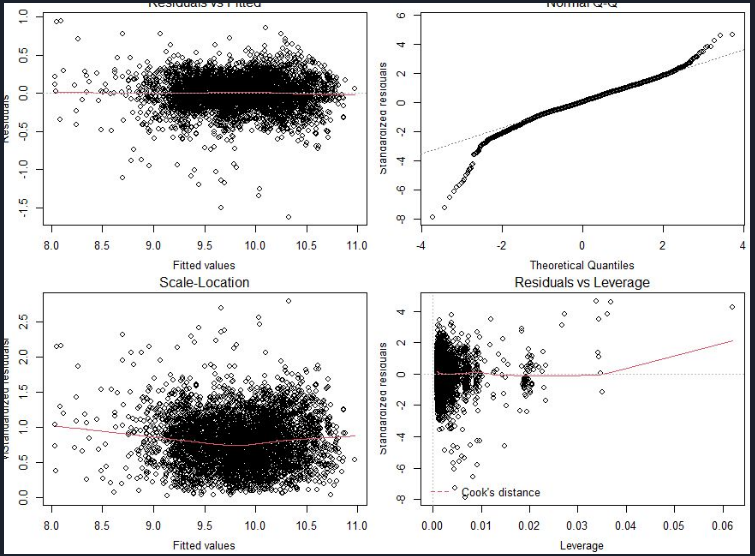


# 4.2. Adding factors

Afegint els factors fuelType i transmission

```
lm(formula = log(price) ~ poly(tax, 2) + poly(mpg, 2) + poly(age,
    2) + sqrt(mileage) + fuelType + transmission, data = df)
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.003e+01  1.286e-02 779.465  < 2e-16 ***
poly(tax, 2)1                2.377e-01  2.249e-01   1.057    0.290    
poly(tax, 2)2                1.054e+00  2.157e-01   4.885 1.07e-06 ***
poly(mpg, 2)1               -1.525e+01  2.879e-01 -52.978  < 2e-16 ***
poly(mpg, 2)2                3.901e+00  2.114e-01  18.456  < 2e-16 ***
poly(age, 2)1               -1.226e+01  3.973e-01 -30.861  < 2e-16 ***
poly(age, 2)2               -5.202e+00  2.399e-01 -21.684  < 2e-16 ***
sqrt(mileage)               -1.536e-03  7.883e-05 -19.480  < 2e-16 ***
fuelTypef.Fuel-Petrol       -3.040e-01  7.235e-03 -42.013  < 2e-16 ***
fuelTypef.Fuel-Hybrid        5.451e-02  2.816e-02   1.936    0.053 .  
transmissionf.Trans-SemiAuto  2.492e-01  7.485e-03  33.289  < 2e-16 ***
transmissionf.Trans-Automatic 2.284e-01  8.465e-03  26.985  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2077 on 4948 degrees of freedom
Multiple R-squared:  0.8069,    Adjusted R-squared:  0.8065 
F-statistic:  1880 on 11 and 4948 DF,  p-value: < 2.2e-16
```



# 4.3. Final model

Model final:

```
lm(formula = log(price) ~ poly(tax, 2) + poly(mpg, 2) + poly(age,
    2) + mileage + mpg * fuelType + fuelType:transmission, data = df)
```

```
Coefficients: (2 not defined because of singularities)
                                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                   9.952e+00  1.012e-02 983.211  < 2e-16 ***
poly(tax, 2)1                                 1.766e-01  2.212e-01   0.798  0.42474    
poly(tax, 2)2                                 8.869e-01  2.133e-01   4.158 3.27e-05 ***
poly(mpg, 2)1                                -1.398e+01  3.471e-01 -40.279  < 2e-16 ***
poly(mpg, 2)2                                 2.474e+00  2.574e-01   9.612  < 2e-16 ***
poly(age, 2)1                                -1.221e+01  3.611e-01 -33.813  < 2e-16 ***
poly(age, 2)2                                -4.208e+00  2.245e-01 -18.742  < 2e-16 ***
mileage                                      -5.215e-06  2.322e-07 -22.458  < 2e-16 ***
mpg                                                  NA         NA      NA       NA    
fuelTypef.Fuel-Petrol                        -9.348e-03  4.366e-02  -0.214  0.83048    
fuelTypef.Fuel-Hybrid                        -4.921e-01  1.554e-01  -3.167  0.00155 ** 
mpg:fuelTypef.Fuel-Petrol                    -6.199e-03  7.898e-04  -7.849 5.09e-15 ***
mpg:fuelTypef.Fuel-Hybrid                     1.317e-02  2.471e-03   5.332 1.02e-07 ***
fuelTypef.Fuel-Diesel:transmissionf.Trans-SemiAuto   2.283e-01  9.842e-03  23.192  < 2e-16 ***
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto   2.341e-01  1.127e-02  20.774  < 2e-16 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto   4.281e-02  5.824e-02   0.735  0.46240    
fuelTypef.Fuel-Diesel:transmissionf.Trans-Automatic  2.144e-01  1.058e-02  20.259  < 2e-16 ***
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic  2.162e-01  1.402e-02  15.415  < 2e-16 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic         NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2041 on 4943 degrees of freedom
Multiple R-squared:  0.8137,	Adjusted R-squared:  0.8131 
F-statistic:  1349 on 16 and 4943 DF,  p-value: < 2.2e-16
```

```
> AIC(m1,m2,m3,m7)
   df        AIC
m1  6 101377.556
m2  9   1701.126
m3 13  -1502.138
m7 18  -1668.921
```

# 5. Modelling of the categorical target

# 5.1. Initial model

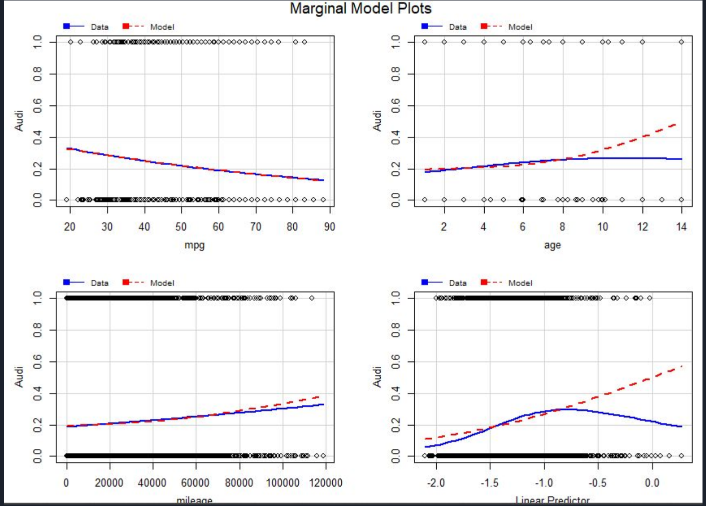
Separem les dades entre train i test

```
llwork <- sample(1:nrow(df),round(0.70*nrow(df),0))

dfall<-df
df_train <- dfall[llwork,]
df_test <-dfall[-llwork,]
```

Model inicial :

```
glm(formula = Audi ~ mpg + age + mileage, family = "binomial",
    data = df_train)
```

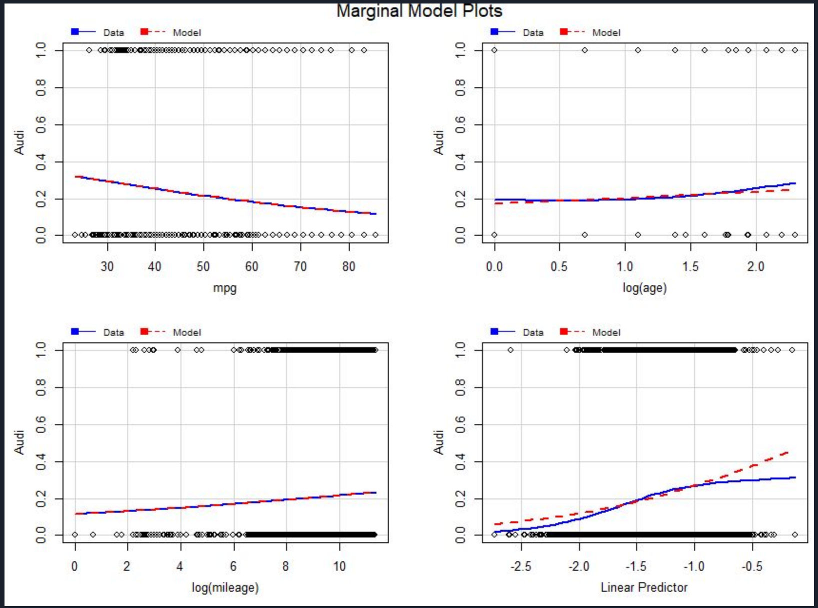
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2971  -0.7074  -0.6435  -0.5558   2.0638

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.390e-01  2.081e-01  -1.629  0.10327
mpg         -2.419e-02  3.894e-03  -6.214 5.17e-10 ***
age          2.192e-02  3.339e-02   0.657  0.51149
mileage      8.576e-06  3.173e-06   2.703  0.00688 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3584.4  on 3471  degrees of freedom
Residual deviance: 3533.4  on 3468  degrees of freedom
AIC: 3541.4
```



# 5.2. Afegim transformacions

```
glm(formula = Audi ~ mpg + log(age) + log(mileage), family = "binomial",
    data = df_train[!df_train$mout == "YesMOut", ])
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -0.794426   0.337283  -2.355   0.0185 *  
mpg           -0.031437   0.004342  -7.240  4.5e-13 ***
log(age)       0.312067   0.123128   2.534   0.0113 *  
log(mileage)   0.078792   0.040983   1.923   0.0545 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3391.3  on 3320  degrees of freedom
Residual deviance: 3329.4  on 3317  degrees of freedom
AIC: 3337.4
```



# 5.3. Afegim factors

Separem les dades entre train i test

Afegim els factors fuelType, transmission i engineSize

```
glm(formula = Audi ~ mpg + log(age) + log(mileage) + fuelType +
    transmission + engineSize, family = "binomial", data = df_train[!df_train$mout ==
    "YesMOut", ])
```

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.253947   0.477462   0.532 0.594817
mpg                         -0.052067   0.006123  -8.504  < 2e-16 ***
log(age)                     0.454958   0.129510   3.513 0.000443 ***
log(mileage)                 0.093161   0.041194   2.262 0.023727 *
fuelTypef.Fuel-Petrol       -0.242165   0.136986  -1.768 0.077092 .
fuelTypef.Fuel-Hybrid       -1.330076   0.743787  -1.788 0.073736 .
transmissionf.Trans-SemiAuto -0.347086  0.112973  -3.072 0.002124 **
transmissionf.Trans-Automatic -0.381977 0.130388  -2.930 0.003395 **
engineSizeMedium             0.395873   0.137874   2.871 0.004088 **
engineSizeBig               -0.988871   0.209104  -4.729 2.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3391.3  on 3320  degrees of freedom
Residual deviance: 3191.9  on 3311  degrees of freedom
AIC: 3211.9
```

```
> AIC(mb1,mb2,mb4)
    df      AIC
mb1  4 3541.359
mb2  4 3337.430
mb4 10 3211.877
```

# 5.4. Model final

```
glm(formula = Audi ~ mpg + log(age) + log(mileage) + fuelType +
    transmission + engineSize + (fuelType * engineSize) + mpg *
    transmission, family = "binomial", data = df_train[!df_train$mout ==
    "YesMout", ])
```

```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -17.54035  339.91145  -0.052 0.958845
mpg                                  -0.01613    0.00834  -1.934 0.053128 .
log(age)                              0.45599    0.13034   3.499 0.000468 ***
log(mileage)                          0.09245    0.04144   2.231 0.025678 *
fuelTypef.Fuel-Petrol                15.66679  339.91092   0.046 0.963238
fuelTypef.Fuel-Hybrid                16.58678  339.91204   0.049 0.961081
transmissionf.Trans-SemiAuto          2.49355    0.57721   4.320 1.56e-05 ***
transmissionf.Trans-Automatic         3.09734    0.64954   4.769 1.86e-06 ***
engineSizeMedium                     16.20166  339.91091   0.048 0.961984
engineSizeBig                        14.78600  339.91095   0.043 0.965303
fuelTypef.Fuel-Petrol:engineSizeMedium -16.13317 339.91093 -0.047 0.962144
fuelTypef.Fuel-Hybrid:engineSizeMedium -33.05999 977.75896 -0.034 0.973027
fuelTypef.Fuel-Petrol:engineSizeBig  -16.40771  339.91106  -0.048 0.961501
fuelTypef.Fuel-Hybrid:engineSizeBig  -30.52772 1495.55045  -0.020 0.983714
mpg:transmissionf.Trans-SemiAuto     -0.05176    0.01065  -4.858 1.19e-06 ***
mpg:transmissionf.Trans-Automatic    -0.06538    0.01245  -5.253 1.49e-07 ***
---
```

```
> AIC(mb4,mb5,mb2)
    df      AIC
mb4 10 3211.877
mb5 16 3145.865
mb2  4 3337.430
> Anova(mb5, test="LR")
Analysis of Deviance Table (Type II tests)

Response: Audi
                    LR Chisq Df Pr(>Chisq)
mpg                   67.850  1  < 2.2e-16 ***
log(age)              12.220  1  0.0004727 ***
log(mileage)           5.311  1  0.0211903 *
fuelType               7.888  2  0.0193696 *
transmission           9.547  2  0.0084494 **
engineSize           113.229  2  < 2.2e-16 ***
fuelType:engineSize   53.492  4  6.722e-11 ***
mpg:transmission      36.151  2  1.412e-08 ***
---
```

# 5.5. Predictions (ENG)

```
> tt<-table(audi.est,df_train$Audi);
> 100*sum(diag(tt))/sum(tt)
[1] 78.42742
```

```
> pred_train <- predict(mb5, newdata=df_train, type="response")
> hoslem.test(as.numeric(df_train$Audi)-1, pred_train)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  as.numeric(df_train$Audi) - 1, pred_train
X-squared = 33.521, df = 8, p-value = 4.961e-05
```

```
> pred_test <- predict(mb5, newdata=df_test, type="response")
> hoslem.test(as.numeric(df_test$Audi)-1, pred_test)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  as.numeric(df_test$Audi) - 1, pred_test
X-squared = 41.849, df = 8, p-value = 1.445e-06
```

```
> audi.est <- ifelse(pred_test<0.5,0,1)
> tt<-table(audi.est,df_test$Audi);tt

audi.est Audi No Audi Yes
       0    1144      326
       1      12        6
> 100*sum(diag(tt))/sum(tt)
[1] 77.28495
```

# Preguntes?

Moltes gràcies per escoltar-me!