USED CAR PRICES CASE STUDY: FROM MULTIVARIANT DATA ANALYSIS TO PREDICTIVE MODELING

Soukaïna Mahboub Mehboub

# 1. Introduction

This comprehensive report encapsulates a rigorous analytical journey through a vast dataset of UK used car data. Beginning with meticulous data validation, including a thorough Univariate Descriptive Analysis for each variable, the narrative unfolds to detail strategic data imputation for both numerical and categorical variables. The exploration deepens with a discerning Feature Selection process, sharpening the focus for both numeric and binary targets. Advanced multivariate techniques, such as Principal Component and Multiple Correspondence Analysis, further refine the dataset, culminating in a robust clustering for population segmentation. The crux of the report is the meticulous Model Building process, adeptly balancing statistical rigor with practical insights for numeric and binary responses, reinforced by stringent model validation techniques.

This report presents an exploratory analysis of the 100,000 UK used car dataset. The dataset includes information from four major car manufacturers: Audi, BMW, Mercedes, and Volkswagen. The data consists of details such as car model, registration year, price, gearbox type, mileage, engine fuel, tax, consumption in miles per gallon, and engine size.

To make the analysis manageable and insightful, a random sample of 5,000 records has been selected from this extensive dataset.

Data from: <https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes>

# 2. Validation of the Data Set

## 2.1. Data Preparation

As our initial step, we’ll start by downloading the essential packages and libraries required for our project. It’s crucial to ensure that these packages are properly installed to avoid any issues later on. Once that’s accomplished, our next task involves creating a subset of our dataset with 5000 specific observations. It’s important to note that during this process, we will maintain the complete set of original variables, ensuring that no data is lost.

We’ll now upload the data and proceed to create our sample by randomly selecting 5000 records.

Sample overview (Dimension of the dataframe (number of rows and columns), the names of variables and brief statistical summary (including measures such as mean, median, quartiles, and counts for each variable)).

str(df) # Variable types

## 'data.frame': 5000 obs. of 10 variables:  
## $ model : chr " 1 Series" " GLE Class" " Caddy Maxi Life" " Golf" ...  
## $ year : int 2017 2018 2019 2019 2016 2019 2018 2017 2018 2019 ...  
## $ price : int 19761 44738 19000 17990 25412 16930 20310 15498 17250 16555 ...  
## $ transmission: chr "Semi-Auto" "Semi-Auto" "Automatic" "Manual" ...  
## $ mileage : int 39681 21276 13191 1201 24346 5317 14863 62140 7629 9451 ...  
## $ fuelType : chr "Petrol" "Diesel" "Diesel" "Diesel" ...  
## $ tax : int 200 150 145 145 160 145 145 145 150 145 ...  
## $ mpg : num 39.8 36.7 44.1 57.7 51.4 49.6 53.3 64.2 56.5 68.9 ...  
## $ engineSize : num 3 3 2 1.6 3 1.6 1.4 2 1.4 2 ...  
## $ manufacturer: chr "BMW" "Mercedes" "VW" "VW" ...

dim(df) # Displays the sample size

## [1] 5000 10

names(df) # Displays the names of the sample variables

## [1] "model" "year" "price" "transmission" "mileage"   
## [6] "fuelType" "tax" "mpg" "engineSize" "manufacturer"

summary(df)

## model year price transmission   
## Length:5000 Min. :1998 Min. : 899 Length:5000   
## Class :character 1st Qu.:2016 1st Qu.: 13994 Class :character   
## Mode :character Median :2017 Median : 19500 Mode :character   
## Mean :2017 Mean : 21573   
## 3rd Qu.:2019 3rd Qu.: 26499   
## Max. :2020 Max. :154998   
## mileage fuelType tax mpg   
## Min. : 1 Length:5000 Min. : 0.0 Min. : 21.10   
## 1st Qu.: 5866 Class :character 1st Qu.:125.0 1st Qu.: 44.10   
## Median : 16698 Mode :character Median :145.0 Median : 52.30   
## Mean : 23309 Mean :125.5 Mean : 53.67   
## 3rd Qu.: 33646 3rd Qu.:145.0 3rd Qu.: 61.40   
## Max. :323000 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 :1.927   
## 3rd Qu.:2.000   
## Max. :6.200

Prior to examining individual variables, we’ll establish counters to track missing values, errors, and outliers within the vectors.

We will also detect all the missing values in the dataframe and store them in two vectors (initial missings for the individuals and for each variable.

mis1<-countNA(df)   
imis<-mis1$mis\_ind   
#mis1$mis\_col   
# Number of missings for the current set of variables   
jmis<-mis1$mis\_col$mis\_x   
iouts<-rep(0,nrow(df))   
# rows - trips   
jouts<-rep(0,ncol(df))   
# columns - variables   
ierrs<-rep(0,nrow(df))   
# rows - trips   
jerrs<-rep(0,ncol(df))   
# columns - variables

Categorical variables should be converted to factors for appropriate analysis to enhance data analysis and enabling effective grouping, summarization, and visualization.

Model (1)

df$model<-factor(paste0(df$manufacturer,"-",df$model))   
levels(df$model)

## [1] "Audi- A1" "Audi- A3" "Audi- A4"   
## [4] "Audi- A5" "Audi- A6" "Audi- A7"   
## [7] "Audi- A8" "Audi- Q2" "Audi- Q3"   
## [10] "Audi- Q5" "Audi- Q7" "Audi- Q8"   
## [13] "Audi- R8" "Audi- RS3" "Audi- RS4"   
## [16] "Audi- RS5" "Audi- RS6" "Audi- S3"   
## [19] "Audi- S4" "Audi- S8" "Audi- SQ5"   
## [22] "Audi- TT" "BMW- 1 Series" "BMW- 2 Series"   
## [25] "BMW- 3 Series" "BMW- 4 Series" "BMW- 5 Series"   
## [28] "BMW- 6 Series" "BMW- 7 Series" "BMW- 8 Series"   
## [31] "BMW- i3" "BMW- M2" "BMW- M3"   
## [34] "BMW- M4" "BMW- M5" "BMW- M6"   
## [37] "BMW- X1" "BMW- X2" "BMW- X3"   
## [40] "BMW- X4" "BMW- X5" "BMW- X6"   
## [43] "BMW- X7" "BMW- Z3" "BMW- Z4"   
## [46] "Mercedes- A Class" "Mercedes- B Class" "Mercedes- C Class"   
## [49] "Mercedes- CL Class" "Mercedes- CLA Class" "Mercedes- CLS Class"  
## [52] "Mercedes- E Class" "Mercedes- G Class" "Mercedes- GL Class"   
## [55] "Mercedes- GLA Class" "Mercedes- GLB Class" "Mercedes- GLC Class"  
## [58] "Mercedes- GLE Class" "Mercedes- GLS Class" "Mercedes- M Class"   
## [61] "Mercedes- S Class" "Mercedes- SL CLASS" "Mercedes- SLK"   
## [64] "Mercedes- V Class" "Mercedes- X-CLASS" "VW- Amarok"   
## [67] "VW- Arteon" "VW- Beetle" "VW- Caddy"   
## [70] "VW- Caddy Maxi" "VW- Caddy Maxi Life" "VW- California"   
## [73] "VW- Caravelle" "VW- CC" "VW- Fox"   
## [76] "VW- Golf" "VW- Golf SV" "VW- Passat"   
## [79] "VW- Polo" "VW- Scirocco" "VW- Sharan"   
## [82] "VW- Shuttle" "VW- T-Cross" "VW- T-Roc"   
## [85] "VW- Tiguan" "VW- Tiguan Allspace" "VW- Touareg"   
## [88] "VW- Touran" "VW- Up"

Transmission (4)

df$transmission <- factor(df$transmission)   
levels( df$transmission )

## [1] "Automatic" "Manual" "Semi-Auto"

df$transmission <- factor( df$transmission, levels = c("Manual","Semi-Auto","Automatic"),labels = paste0("f.Trans-",c("Manual","SemiAuto","Automatic")))

FueltType (6)

df$fuelType <- factor( df$fuelType )

Manufacturer (10)

df$manufacturer <- factor( df$manufacturer )

## 2.2 Exploratory Data Analysis

### 2.2.1 Variable missings, errors & outliers

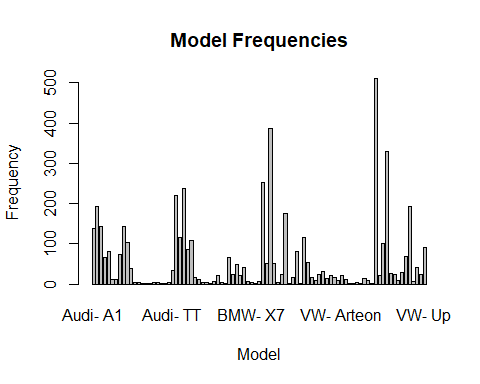
Model (1):

In this variable, the presence of numerous car models makes it challenging to identify missing values through a barplot. To tackle this, we will primarily utilize functions such as table() and is.na() to assess the distribution of cars across each model and employ is.na() for missing value detection.

summary(df$model)

## Audi- A1 Audi- A3 Audi- A4 Audi- A5   
## 137 192 142 65   
## Audi- A6 Audi- A7 Audi- A8 Audi- Q2   
## 81 11 12 73   
## Audi- Q3 Audi- Q5 Audi- Q7 Audi- Q8   
## 143 103 39 5   
## Audi- R8 Audi- RS3 Audi- RS4 Audi- RS5   
## 4 2 1 1   
## Audi- RS6 Audi- S3 Audi- S4 Audi- S8   
## 4 3 1 1   
## Audi- SQ5 Audi- TT BMW- 1 Series BMW- 2 Series   
## 3 34 219 115   
## BMW- 3 Series BMW- 4 Series BMW- 5 Series BMW- 6 Series   
## 237 85 109 17   
## BMW- 7 Series BMW- 8 Series BMW- i3 BMW- M2   
## 12 3 3 2   
## BMW- M3 BMW- M4 BMW- M5 BMW- M6   
## 7 21 3 2   
## BMW- X1 BMW- X2 BMW- X3 BMW- X4   
## 65 24 49 22   
## BMW- X5 BMW- X6 BMW- X7 BMW- Z3   
## 42 7 5 1   
## BMW- Z4 Mercedes- A Class Mercedes- B Class Mercedes- C Class   
## 7 253 51 387   
## Mercedes- CL Class Mercedes- CLA Class Mercedes- CLS Class Mercedes- E Class   
## 51 5 24 175   
## Mercedes- G Class Mercedes- GL Class Mercedes- GLA Class Mercedes- GLB Class   
## 1 15 81 1   
## Mercedes- GLC Class Mercedes- GLE Class Mercedes- GLS Class Mercedes- M Class   
## 115 53 15 8   
## Mercedes- S Class Mercedes- SL CLASS Mercedes- SLK Mercedes- V Class   
## 23 30 14 20   
## Mercedes- X-CLASS VW- Amarok VW- Arteon VW- Beetle   
## 15 10 22 12   
## VW- Caddy VW- Caddy Maxi VW- Caddy Maxi Life VW- California   
## 1 1 4 2   
## VW- Caravelle VW- CC VW- Fox VW- Golf   
## 14 9 1 510   
## VW- Golf SV VW- Passat VW- Polo VW- Scirocco   
## 21 100 328 27   
## VW- Sharan VW- Shuttle VW- T-Cross VW- T-Roc   
## 24 10 28 69   
## VW- Tiguan VW- Tiguan Allspace VW- Touareg VW- Touran   
## 193 6 42 24   
## VW- Up   
## 91

barplot(table(df$model), main = "Model Frequencies", xlab = "Model", ylab = "Frequency")



Detecting any missing values:

#Detecting any missing values as previous barplot cannot show missing values:  
na\_values <- is.na(df$model)  
any(na\_values)

## [1] FALSE

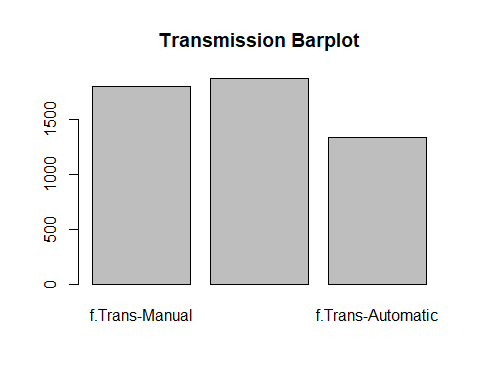
Transmission (2):

Zero missing values, and cars are nearly evenly distributed across three categories. No errors or outliers are present (as these three are the only three possible transmission types in cars).

summary(df$transmission)

## f.Trans-Manual f.Trans-SemiAuto f.Trans-Automatic   
## 1798 1870 1332

barplot(summary(df$transmission),main="Transmission Barplot")



FuelType (6):

As we can see, the summary reveals that there are 15 NA’s in this variable, and very few cars are hybrid

At this stage we will consider missing values as electrical cars if their engine-size are zero (This assumption will help us analyze the “engineSize” variable later).

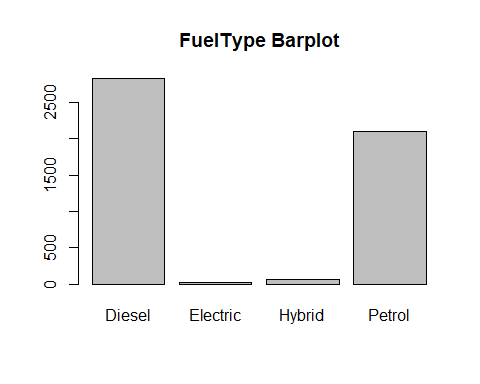
summary(df$fuelType)

## Diesel Hybrid Other Petrol   
## 2825 64 15 2096

#Mark NA's as Electric car  
na\_rows <- which(df$fuelType == 'Other')  
#convert variable back to character (to avoid warnings)  
df$fuelType <- as.character(df$fuelType)  
df$fuelType[na\_rows] <- 'Electric'  
#convert variable back to factor  
df$fuelType <- as.factor(df$fuelType)

FuelType Distribution:

#Barplot  
barplot(summary(df$fuelType),main="FuelType Barplot")



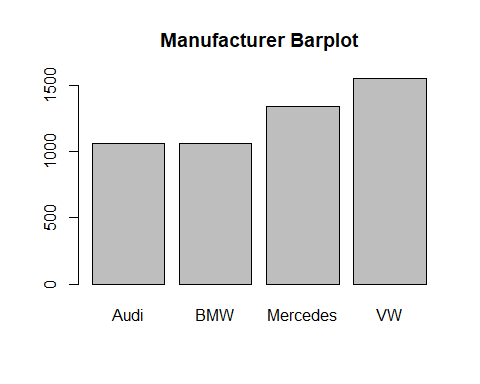
Manufacturer (10):

Every vehicle in our sample is sourced from one of the four manufacturers that contributed to our dataset. We’ve detected no missing values. Since our sample was selected randomly, we have a slightly higher representation of VW and Mercedes cars compared to Audi and BMW. For this variable, no missing, errors, or outliers data has been identified.

summary(df$manufacturer)

## Audi BMW Mercedes VW   
## 1057 1057 1337 1549

barplot(summary(df$manufacturer),main="Manufacturer Barplot")



We will consistently detect missing outliers in all numerical variables using the same method, which involves identifying both low and high outliers. This approach ensures that the R script remains adaptable to changes in datasets or samples without requiring modifications.

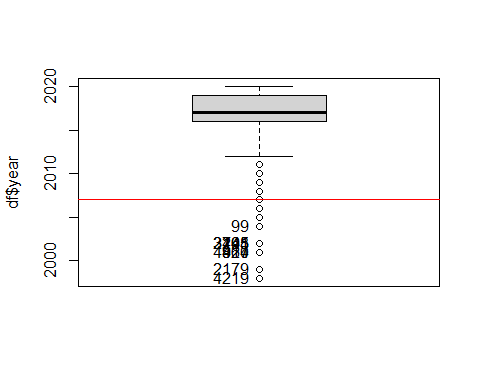
Year (2):

The summary indicates that the ‘year’ values fall within the valid range of 1998 to 2020, demonstrating the absence of errors or inconsistencies. Given that ‘year’ is typically represented as an integer, we’ll ensure any potential decimal values are rounded to maintain data integrity.

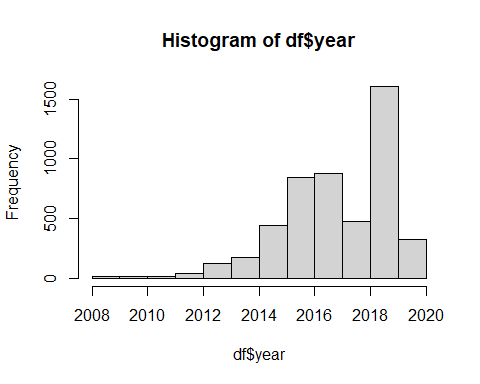
x<-summary(df$year)  
  
# Outlier detection  
Boxplot(df$year)

## [1] 4219 2179 460 814 4927 248 2495 3165 3741 99

var\_out<-calcQ(df$year)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



sel <- which(df$year <= var\_out$souti);  
iouts[sel]<-iouts[sel]+1  
jouts[2]<-jouts[2]+length(sel)  
df[sel, "year"] <- NA  
  
sel <- which(df$year >= var\_out$souts);  
iouts[sel]<-iouts[sel]+1  
jouts[2]<-jouts[2]+length(sel)  
df[sel, "year"] <- NA  
  
hist(df$year) #Distribution of "year"



Price (3):

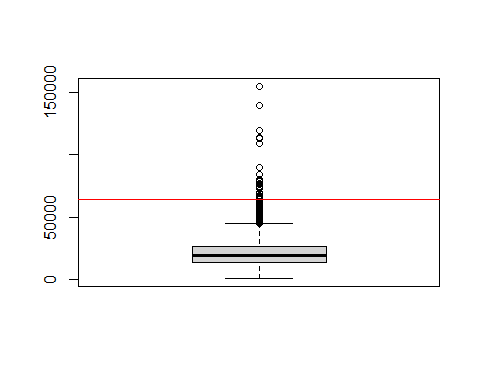
No missing values, no errors identified, and all values fall within a reasonable range, reflecting real car prices in the current market. We’ll focus on excluding only the most extreme outliers.

As “price” is out Target Variable, we won’t do imputations, so we won’t assign NA value to outliers.

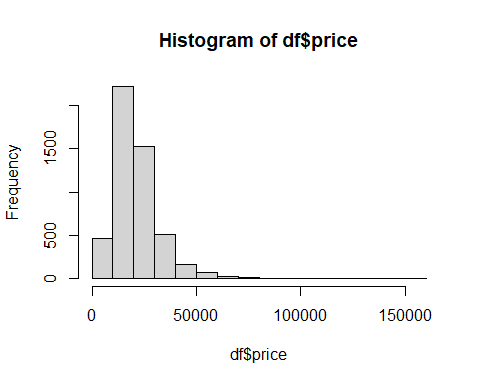
summary(df$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 899 13994 19500 21573 26499 154998

# Outlier detection  
boxplot(df$price)  
var\_out<-calcQ(df$price)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



sel <- which(df$price <= var\_out$souti);  
iouts[sel]<-iouts[sel]+1  
jouts[3]<-jouts[3]+length(sel)  
  
  
sel <- which(df$price >= var\_out$souts);  
iouts[sel]<-iouts[sel]+1  
jouts[3]<-jouts[3]+length(sel)  
  
hist(df$price) #Distribution of "price"



Mileage (5):

No missing values or errors are present, given the logical and positive range of all mileage values. Our focus will be on the exclusion of extreme outliers.

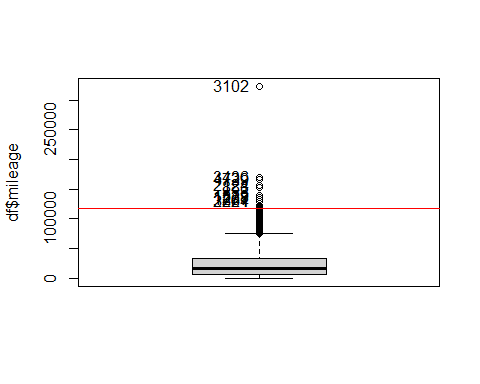
summary(df$mileage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 5866 16698 23309 33646 323000

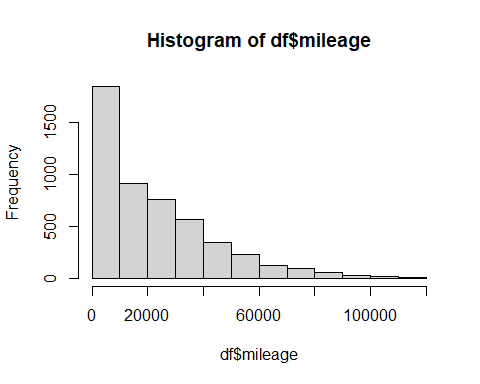
# Outlier detection  
Boxplot(df$mileage)

## [1] 3102 3436 4729 2124 888 1579 1228 1267 2564 3221

var\_out<-calcQ(df$mileage)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



sel <- which(df$mileage >= var\_out$souts);  
iouts[sel]<-iouts[sel]+1  
jouts[5]<-jouts[5]+length(sel)  
df[sel, "mileage"] <- NA  
  
sel <- which(df$mileage <= var\_out$souti);  
iouts[sel]<-iouts[sel]+1  
jouts[5]<-jouts[5]+length(sel)  
df[sel, "mileage"] <- NA  
  
hist(df$mileage) #Distribution of "mileage"



Tax (7):

The summary reveals that there are instances of zero tax values. This is a possibility in specific cases within the UK, considering the dataset’s origin.

The tax values are within expected ranges, so our primary concern is identifying extreme outliers.

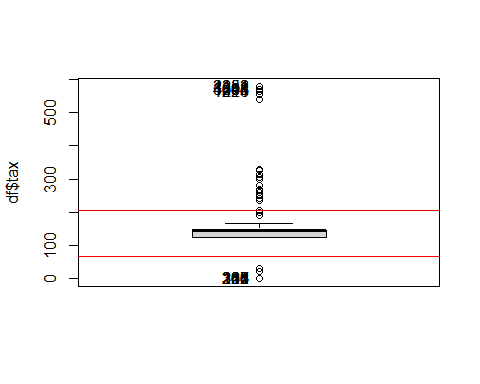
summary(df$tax)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 125.0 145.0 125.5 145.0 580.0

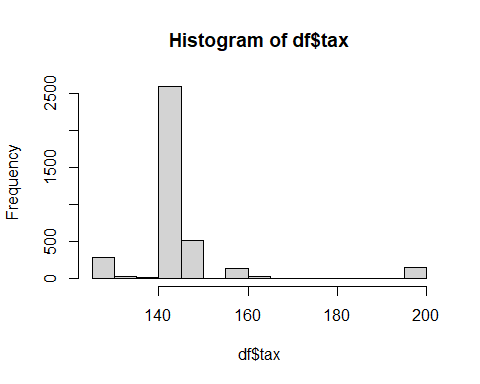
# Outlier detection  
Boxplot(df$tax)

## [1] 101 112 165 206 244 268 316 317 321 381 2131 4252 248 361 3095  
## [16] 4434 4604 4682 1221 1916

var\_out<-calcQ(df$tax)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



sel <- which(df$tax >= var\_out$souts);  
iouts[sel]<-iouts[sel]+1  
jouts[7]<- jouts[7] +length(sel)  
df[sel, "tax"] <- NA  
  
  
sel <- which(df$tax <= var\_out$souti);  
iouts[sel]<-iouts[sel]+1  
jouts[7]<- jouts[7] +length(sel)  
df[sel, "tax"] <- NA  
  
hist(df$tax) #Distribution of "tax"



MPG (8):

As we can observe from the summary, there are no missing values in this variable. However, it’s worth noting that some values are significantly higher than what would be considered normal for miles per gallon (mpg), even though they fall within the possible range. To identify and address these extreme outliers, we will proceed with outlier detection.

Note: We will assume that electric cars, which have an MPG value, are represented as MPGe (Miles Per Gallon Equivalent), in order to prevent any data loss.

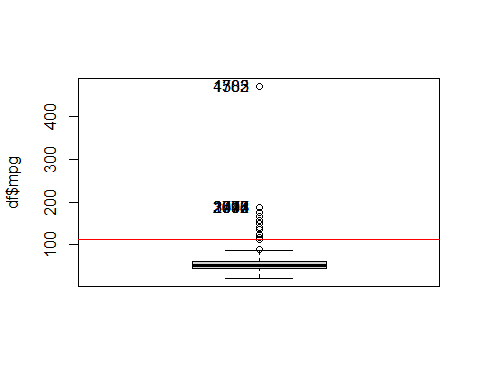
summary(df$mpg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.10 44.10 52.30 53.67 61.40 470.80

# Outlier detection  
Boxplot(df$mpg)

## [1] 383 1785 4502 515 1636 2073 2472 2747 3604 3994

var\_out<-calcQ(df$mpg)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



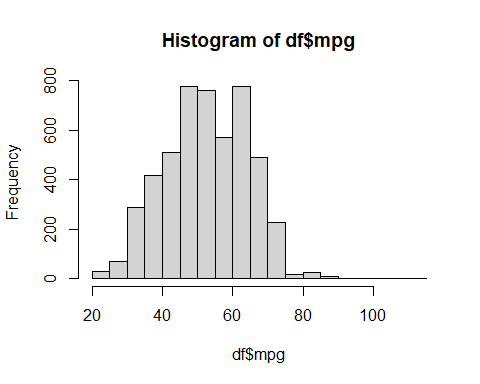
var\_out$souts

## 3rd Qu.   
## 113.3

var\_out$souti

## 1st Qu.   
## -7.8

sel <- which(df$mpg >= var\_out$souts);  
  
iouts[sel]<-iouts[sel]+1  
jouts[8]<- jouts[8] +length(sel)  
df[sel, "mpg"] <- NA  
  
  
sel <- which(df$mpg <= var\_out$souti);  
  
iouts[sel]<-iouts[sel]+1  
jouts[8]<- jouts[8] +length(sel)  
df[sel, "mpg"] <- NA  
  
hist(df$mpg) #Distribution of "mpg"



Engine size (9):

Through summary, we can see that we have no missing values here. However, we spotted some errors. When a car’s engine size is listed as 0, it usually means the car is electric. However, some cars, like the Mercedes C class, might also show 0 as the engine size, but they are not electric; this could be a data issue. It is also an error to find Hybrid, Petrol and Diesel with an engine size 0.

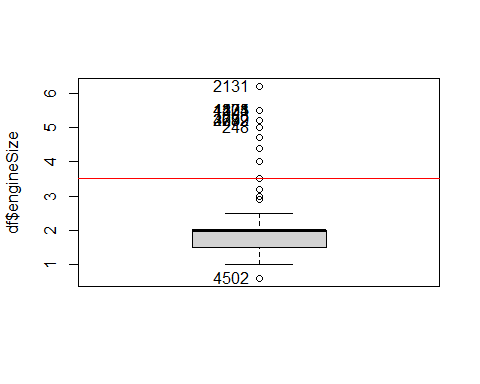
summary(df$engineSize)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.500 2.000 1.927 2.000 6.200

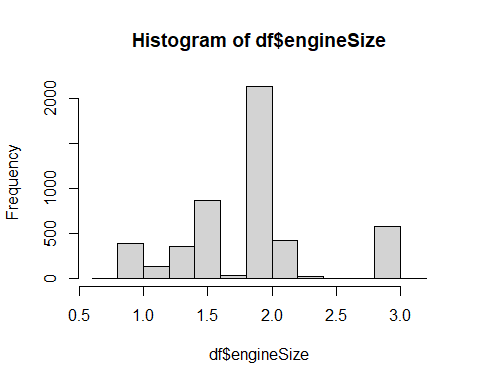
sel <- which(df$engineSize == 0 & (df$model == "Mercedes- C Class" | df$fuelType != "Electric"))   
  
ierrs[sel]<-ierrs[sel]+1  
jerrs[9]<-length(sel)  
df[sel,"engineSize"]<-NA  
  
# Outlier detection  
Boxplot(df$engineSize)

## [1] 4502 2131 1173 1221 4434 4505 799 2272 3032 4682 248

var\_out<-calcQ(df$engineSize)  
abline(h=var\_out$souts,col="red")  
abline(h=var\_out$souti,col="red")



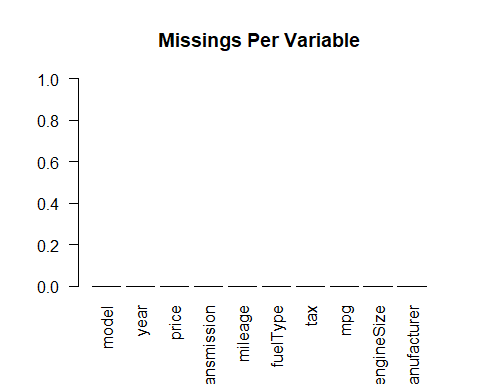
sel <- which(df$engineSize >= var\_out$souts);  
iouts[sel]<-iouts[sel]+1  
jouts[9]<- jouts[9] +length(sel)  
df[sel, "engineSize"] <- NA  
  
  
sel <- which(df$engineSize <= var\_out$souti);  
iouts[sel]<-iouts[sel]+1  
jouts[9]<- jouts[9] +length(sel)  
df[sel, "engineSize"] <- NA  
  
  
hist(df$engineSize) #Distribution of "engineSize"



### 2.2.3 Summary of variable analysis

As we can we see, initially we have no missing values to begin it.

labels <- colnames(df[1:10])  
# Barplot  
barplot(mis1$mis\_col$mis\_x, names.arg = labels, main = "Missings Per Variable", col = "grey", ylim = c(0, max(mis1$mis\_col$mis\_x) + 1), las = 2)

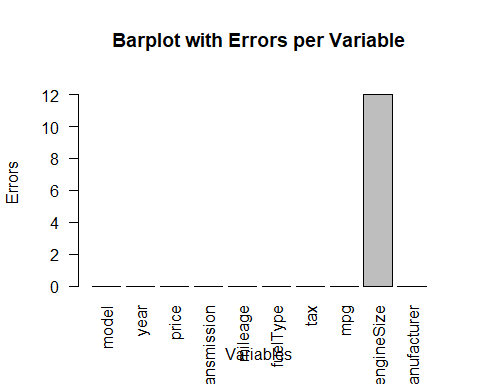


Only 12 errors in engineSize:

jerrs

## [1] 0 0 0 0 0 0 0 0 12 0

# Barplot  
barplot(jerrs[1:10], names.arg = labels,   
 main = "Barplot with Errors per Variable",  
 xlab = "Variables", ylab = "Errors",  
 col = "grey",   
 ylim = c(0, max(jerrs) + 1),  
 las = 2)

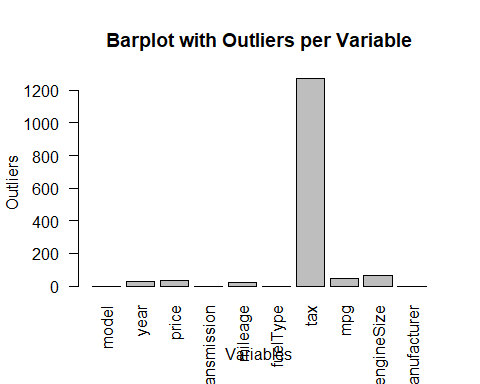


Some scattered outliers in few variables.

jouts

## [1] 0 29 37 0 20 0 1270 49 66 0

# Barplot  
barplot(jouts[1:10], names.arg = labels,   
 main = "Barplot with Outliers per Variable",  
 xlab = "Variables", ylab = "Outliers",  
 col = "grey",   
 ylim = c(0, max(jouts) + 1),  
 las = 2)



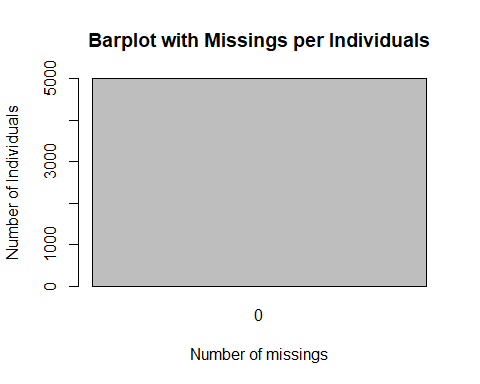
### 2.2.3 Individuals’ missings, errors & outliers

Missings

table(imis)

## imis  
## 0   
## 5000

barplot(table(imis), main = "Barplot with Missings per Individuals",  
 xlab = "Number of missings", ylab = "Number of Individuals",  
 col = "grey",  
 ylim = c(0,5000))

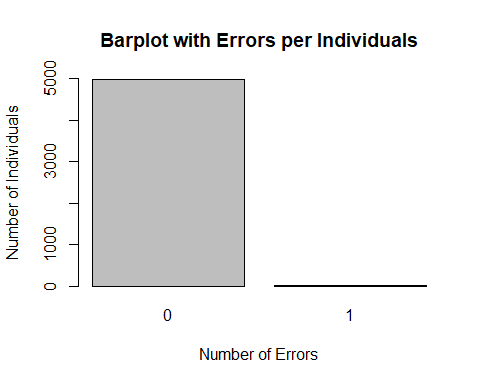


Errors:

table(ierrs)

## ierrs  
## 0 1   
## 4988 12

barplot(table(ierrs), main = "Barplot with Errors per Individuals",  
 xlab = "Number of Errors", ylab = "Number of Individuals",  
 col = "grey",  
 ylim = c(0,5000))

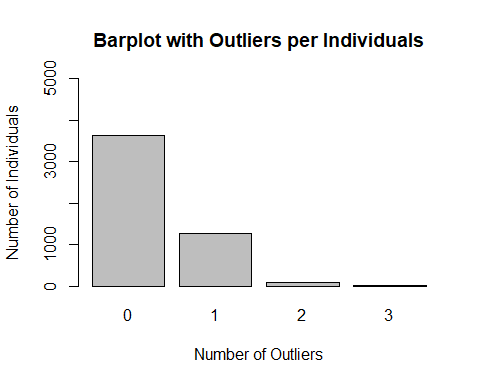


Outliers

table(iouts)

## iouts  
## 0 1 2 3   
## 3640 1258 93 9

barplot(table(iouts), main = "Barplot with Outliers per Individuals",  
 xlab = "Number of Outliers", ylab = "Number of Individuals",  
 col = "grey",  
 ylim = c(0,5000))



### 2.2.4 Summary of inidividuals’ analysis

Summary and totals of missings, errors and outliers:

# TOTAL OF INDIVIDUAL MISSINGS, ERRORS, OUTLIERS:   
total\_missings <- sum(imis); total\_errors <- sum(ierrs); total\_outliers <- sum(iouts);  
total\_missings; total\_errors; total\_outliers;

## [1] 0

## [1] 12

## [1] 1471

## 2.4 Correlation between variables

We observe a strong correlation between ‘year’ and ‘mileage,’ which is intuitively sensible since both increase as years pass and the vehicle is driven. Additionally, the ‘price’ variable shows noteworthy correlations with ‘year’ and ‘engine size’.

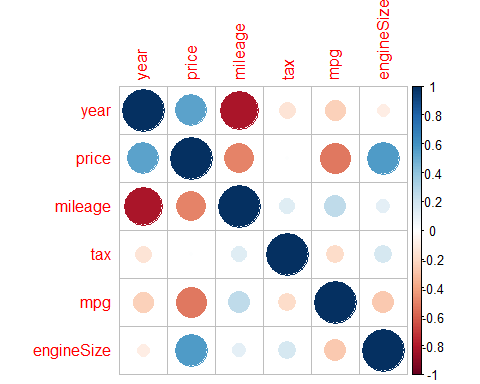
# dataset with numerical variables and individuals without NA values.  
  
df\_temp <- na.omit(df)  
numerical\_df <- df\_temp[, sapply(df\_temp, is.numeric)]  
numerical\_df <- numerical\_df[1:6]  
head(numerical\_df)

## year price mileage tax mpg engineSize  
## 12837 2017 19761 39681 200 39.8 3.0  
## 29357 2018 44738 21276 150 36.7 3.0  
## 47901 2019 19000 13191 145 44.1 2.0  
## 37819 2019 17990 1201 145 57.7 1.6  
## 25588 2016 25412 24346 160 51.4 3.0  
## 22743 2019 16930 5317 145 49.6 1.6

# Coorelation matrix   
correlation\_matrix <- cor(numerical\_df)  
  
# Print the correlation matrix  
library(corrplot)

## corrplot 0.92 loaded

corrplot(correlation\_matrix)



# 3 Imputation & Discretization & Multivariant Outliers Detection

We will refrain from applying imputation to any missing values in the “price” variable. This variable represents the target variable in our study, and altering or filling in missing values in this variable could introduce bias into our data, potentially skewing the results.

Note: in this case of ours we have no missings at all.

## 3.1 Imputation with Numerical Variables

As we can see, missing values are substituted with new values:

library(missMDA)  
quantitative\_vars<-names(df)[c(2,3,5,7:9)]  
  
summary(df[,quantitative\_vars])

## year price mileage tax   
## Min. :2008 Min. : 899 Min. : 1 Min. :125.0   
## 1st Qu.:2016 1st Qu.: 13994 1st Qu.: 5836 1st Qu.:145.0   
## Median :2017 Median : 19500 Median : 16513 Median :145.0   
## Mean :2017 Mean : 21573 Mean : 22834 Mean :146.9   
## 3rd Qu.:2019 3rd Qu.: 26499 3rd Qu.: 33396 3rd Qu.:145.0   
## Max. :2020 Max. :154998 Max. :116000 Max. :200.0   
## NA's :29 NA's :20 NA's :1270   
## mpg engineSize   
## Min. : 21.10 Min. :0.6   
## 1st Qu.: 44.10 1st Qu.:1.5   
## Median : 52.30 Median :2.0   
## Mean : 52.51 Mean :1.9   
## 3rd Qu.: 61.40 3rd Qu.:2.0   
## Max. :113.00 Max. :3.2   
## NA's :49 NA's :78

res.input<-imputePCA(df[,quantitative\_vars],ncp=5)  
  
summary(res.input$completeObs)

## year price mileage tax   
## Min. :2008 Min. : 899 Min. : 1 Min. :125.0   
## 1st Qu.:2016 1st Qu.: 13994 1st Qu.: 5866 1st Qu.:145.0   
## Median :2017 Median : 19500 Median : 16698 Median :145.0   
## Mean :2017 Mean : 21573 Mean : 22977 Mean :146.9   
## 3rd Qu.:2019 3rd Qu.: 26499 3rd Qu.: 33646 3rd Qu.:147.2   
## Max. :2020 Max. :154998 Max. :116000 Max. :200.0   
## mpg engineSize   
## Min. : 21.10 Min. :0.600   
## 1st Qu.: 44.10 1st Qu.:1.500   
## Median : 52.30 Median :2.000   
## Mean : 52.51 Mean :1.923   
## 3rd Qu.: 60.20 3rd Qu.:2.000   
## Max. :113.00 Max. :8.051

df[,"year"] <- res.input$completeObs[,"year"]  
  
df[,"price"] <- res.input$completeObs[,"price"]  
  
df[,"mileage"] <- res.input$completeObs[,"mileage"]  
  
df[,"tax"] <- res.input$completeObs[,"tax"]  
  
df[,"mpg"] <- res.input$completeObs[,"mpg"]  
  
df[,"engineSize"] <- res.input$completeObs[,"engineSize"]

## 3.2 Imputation to factors (Categorical Variables)

categorical\_vars<-names(df)[c(1,4,6,10)]  
summary(df[,categorical\_vars])

## model transmission fuelType   
## VW- Golf : 510 f.Trans-Manual :1798 Diesel :2825   
## Mercedes- C Class: 387 f.Trans-SemiAuto :1870 Electric: 15   
## VW- Polo : 328 f.Trans-Automatic:1332 Hybrid : 64   
## Mercedes- A Class: 253 Petrol :2096   
## BMW- 3 Series : 237   
## BMW- 1 Series : 219   
## (Other) :3066   
## manufacturer   
## Audi :1057   
## BMW :1057   
## Mercedes:1337   
## VW :1549   
##   
##   
##

#nb <- estim\_ncpMCA(df[, categorical\_vars],ncp.max=25) #it stabilizes at ncp = 7  
  
X<-imputeMCA(df[,categorical\_vars],ncp=7)  
summary(res.input$completeObs)

## year price mileage tax   
## Min. :2008 Min. : 899 Min. : 1 Min. :125.0   
## 1st Qu.:2016 1st Qu.: 13994 1st Qu.: 5866 1st Qu.:145.0   
## Median :2017 Median : 19500 Median : 16698 Median :145.0   
## Mean :2017 Mean : 21573 Mean : 22977 Mean :146.9   
## 3rd Qu.:2019 3rd Qu.: 26499 3rd Qu.: 33646 3rd Qu.:147.2   
## Max. :2020 Max. :154998 Max. :116000 Max. :200.0   
## mpg engineSize   
## Min. : 21.10 Min. :0.600   
## 1st Qu.: 44.10 1st Qu.:1.500   
## Median : 52.30 Median :2.000   
## Mean : 52.51 Mean :1.923   
## 3rd Qu.: 60.20 3rd Qu.:2.000   
## Max. :113.00 Max. :8.051

df[,"model"] <- X$completeObs[,"model"]  
df[,"transmission"] <- X$completeObs[,"transmission"]  
df[,"fuelType"] <- X$completeObs[,"fuelType"]  
df[,"manufacturer"] <- X$completeObs[,"manufacturer"]

## 3.3 Discretization

Discretization can be important for profiling as it enhances data interpretability, reduces noise, and making the profiling process more effective and more understandable.

# f.Year :  
table(df$year, useNA="always")

##   
## 2008 2009 2010 2010.34738403483   
## 8 10 14 1   
## 2010.66773380449 2010.78983258164 2010.97377044245 2011   
## 1 1 1 20   
## 2011.1217842681 2011.22075865324 2011.29089233062 2011.35562454868   
## 1 1 1 1   
## 2011.91223446802 2012 2012.54949928848 2012.73973497351   
## 1 38 1 1   
## 2012.79396662919 2012.82148845673 2012.85527406511 2012.96681826786   
## 1 1 1 1   
## 2013 2013.13269141168 2013.22232815254 2013.32869020097   
## 127 1 1 1   
## 2013.49526927193 2013.54228455266 2013.61323177794 2013.82692418114   
## 1 1 1 1   
## 2013.90270400965 2013.93149957525 2014 2014.03047094366   
## 1 1 177 1   
## 2015 2015.18898465221 2016 2016.27386794821   
## 440 1 841 1   
## 2016.33089399059 2016.40530896584 2017 2018   
## 1 1 881 479   
## 2019 2020 <NA>   
## 1607 329 0

quantile(df$year,seq(0,1,0.25))

## 0% 25% 50% 75% 100%   
## 2008 2016 2017 2019 2020

min(df$year)

## [1] 2008

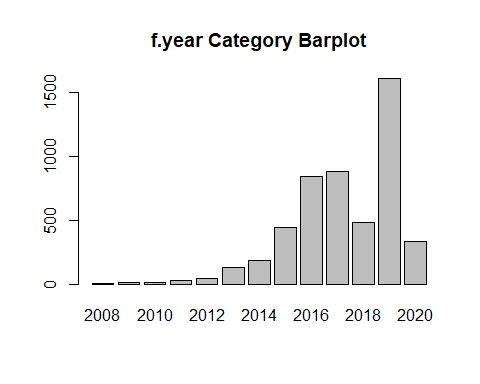
year\_labels <- as.character(seq(2008, 2020))  
year\_breaks <- seq(2007, 2020)  
df$f.year <- cut(df$year, breaks = year\_breaks, labels = year\_labels, include.lowest = TRUE)  
  
summary(df$f.year)

## 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020   
## 8 10 14 24 43 133 186 441 842 884 479 1607 329

table(df$f.year, useNA="always")

##   
## 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 <NA>   
## 8 10 14 24 43 133 186 441 842 884 479 1607 329 0

barplot(summary(df$f.year),main="f.year Category Barplot",col = "Grey")



# f.Price:  
summary(df$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 899 13994 19500 21573 26499 154998

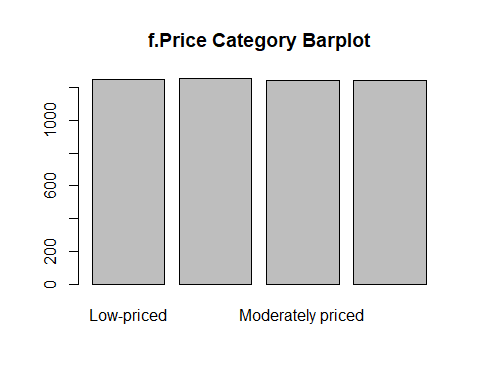
quantile(df$price,seq(0,1,0.25),na.rm=TRUE)

## 0% 25% 50% 75% 100%   
## 899.0 13994.5 19500.0 26499.0 154998.0

df$f.price <- cut(df$price, breaks = c(min(df$price), 13994.5 , 19500 , 26499.0 , max(df$price)), labels = c("Low-priced", "Affordable", "Moderately priced", "Expensive"), include.lowest = TRUE)  
table(df$f.price)

##   
## Low-priced Affordable Moderately priced Expensive   
## 1250 1256 1247 1247

barplot(summary(df$f.price),main="f.Price Category Barplot",col = "Grey")



# f.Mileage: Usage.   
summary(df$mileage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 5866 16698 22977 33646 116000

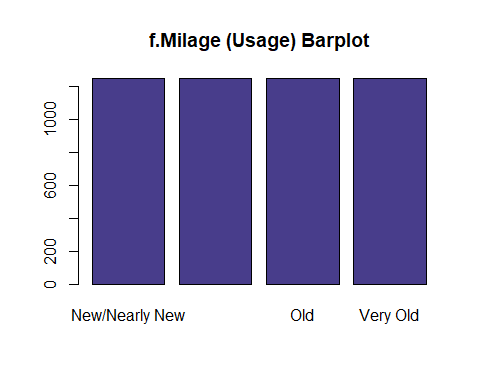
quantile(df$mileage,seq(0,1,0.25),na.rm=TRUE)

## 0% 25% 50% 75% 100%   
## 1.0 5866.5 16697.5 33645.5 116000.0

mileage\_labels <- c("New/Nearly New", "Used", "Old", "Very Old")  
mileage\_intervals <- c(min(df$mileage), 5866.5 , 16697.5, 33645.5, max(df$mileage))  
df$f.miles <- cut(df$mileage, breaks = mileage\_intervals, labels = mileage\_labels, include.lowest = TRUE)  
table(df$f.miles)

##   
## New/Nearly New Used Old Very Old   
## 1250 1250 1250 1250

barplot(summary(df$f.miles),main="f.Milage (Usage) Barplot",col = "DarkSlateBlue")



table(df$f.miles,useNA="always")

##   
## New/Nearly New Used Old Very Old <NA>   
## 1250 1250 1250 1250 0

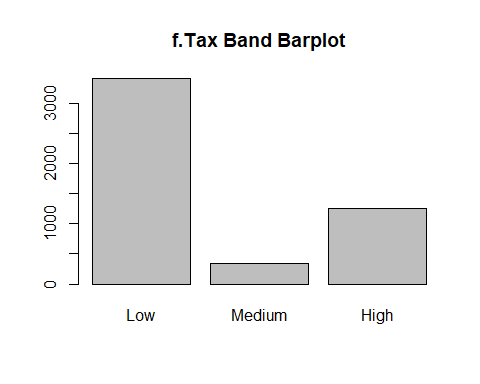
# f.Tax:   
summary(df$tax)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 125.0 145.0 145.0 146.9 147.2 200.0

quantile(df$tax,seq(0,1,0.25),na.rm=TRUE)

## 0% 25% 50% 75% 100%   
## 125.00 145.00 145.00 147.19 200.00

tax\_labels <- c("Low", "Medium", "High")  
tax\_intervals <- c(min(df$tax), 145, 147.19 , max(df$tax))  
df$f.tax <- cut(df$tax, breaks = tax\_intervals, labels = tax\_labels, include.lowest = TRUE)  
barplot(summary(df$f.tax),main="f.Tax Band Barplot",col = "Grey")



# MPG Category: Consumption Category  
summary(df$mpg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.10 44.10 52.30 52.51 60.20 113.00

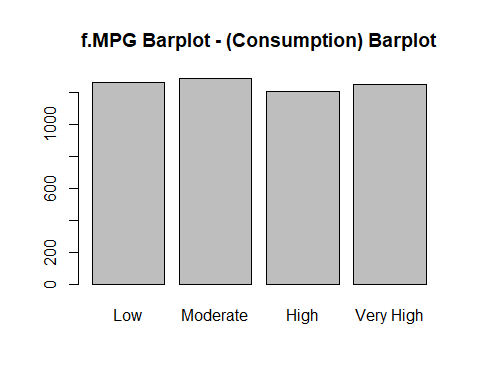
quantile(df$mpg,seq(0,1,0.25),na.rm=TRUE)

## 0% 25% 50% 75% 100%   
## 21.10000 44.10000 52.30000 60.19753 113.00000

mpg\_labels <- c("Low", "Moderate", "High", "Very High")  
mpg\_intervals <- c(min(df$mpg), 44.10, 52.30, 60.20, max(df$mpg))  
df$f.mpg <- cut(df$mpg, breaks = mpg\_intervals, labels = mpg\_labels, include.lowest = TRUE)  
  
table(df$f.mpg)

##   
## Low Moderate High Very High   
## 1260 1286 1204 1250

barplot(summary(df$f.mpg),main="f.MPG Barplot - (Consumption) Barplot",col = "Grey")



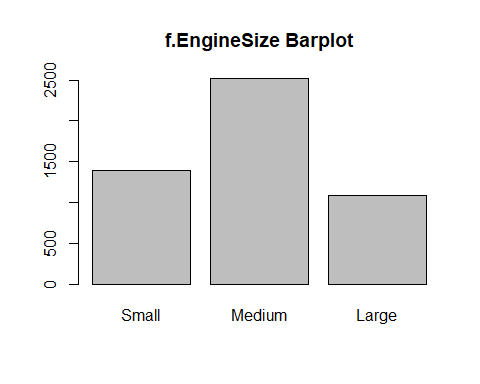
# Engine Size Category: Small, Medium, Large  
summary(df$engineSize)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.600 1.500 2.000 1.923 2.000 8.051

quantile(df$engineSize,seq(0,1,0.25),na.rm=TRUE)

## 0% 25% 50% 75% 100%   
## 0.600000 1.500000 2.000000 2.000000 8.050534

engineSize\_labels <- c("Small", "Medium", "Large")  
engineSize\_intervals <- c(min(df$engineSize), 1.5, 2.0, max(df$engineSize))  
df$f.engineSize <- cut(df$engineSize, breaks = engineSize\_intervals, labels = engineSize\_labels, include.lowest = TRUE)  
barplot(summary(df$f.engineSize),main="f.EngineSize Barplot",col = "Grey")



## 3.4 Multivariant Outliers Detection

We are applying the Mahalanobis method to identify multivariate outliers

doesnt work with tax

library(mvoutlier)

## Loading required package: sgeostat

##   
## Attaching package: 'mvoutlier'

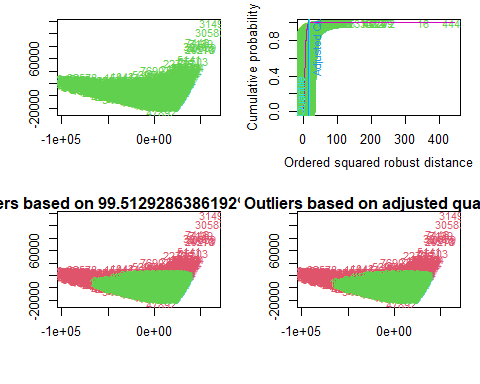
## The following object is masked \_by\_ '.GlobalEnv':  
##   
## X

library(chemometrics)

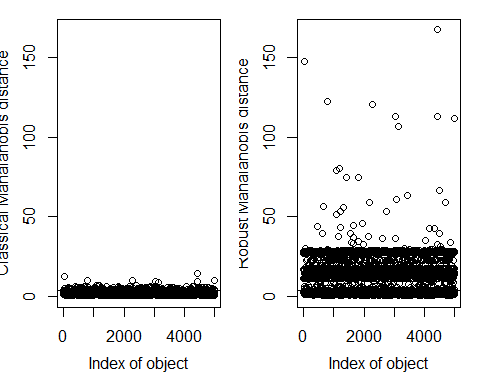
## Loading required package: rpart

vars<-c("year","price", "mileage", "mpg", "engineSize")   
mout<-aq.plot(df[,vars],delta=qchisq(0.99, df= 6), alpha=0.01)

## Projection to the first and second robust principal components.  
## Proportion of total variation (explained variance): 0.9995513



mout<-Moutlier(df[,vars],quantile = 0.99, plot = TRUE)



ll<-which(mout$md > mout$cutoff)

# 4. Profiling

library(FactoMineR)  
summary(df$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 899 13994 19500 21573 26499 154998

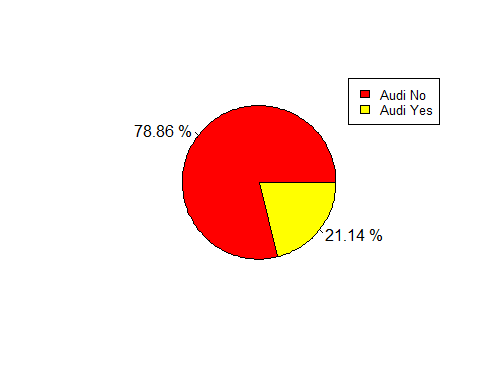
# Binary Target: Audi?  
df$Audi<-ifelse(df$manufacturer == "Audi",1,0)  
df$Audi<-factor(df$Audi,labels=paste("Audi",c("No","Yes")))  
summary(df$Audi)

## Audi No Audi Yes   
## 3943 1057

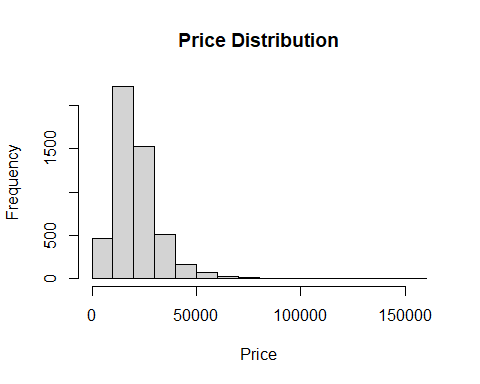
# Pie  
piepercent<-round(100\*(table(df$Audi)/nrow(df)),dig=2); piepercent

##   
## Audi No Audi Yes   
## 78.86 21.14

pie(table(df$Audi),col=heat.colors(2),labels=paste(piepercent,"%"))  
legend("topright", levels(df$Audi), cex = 0.8, fill = heat.colors(2))



# Histogram for Price  
hist(df$price, main = "Price Distribution", xlab = "Price")



With Numeric Target “Price”:

Clearly, each quantitative variable is correlated to “price,” either positively or negatively.

In simple terms, when the year and engine specifications go up, the price tends to rise. On the other hand, an increase in mileage and mpg typically leads to a decrease in price. This straightforward relationship helps us understand how these factors impact pricing.

res.condes<- condes(df, 3)  
  
res.condes$quanti

## correlation p.value  
## engineSize 0.6417973 0  
## year 0.5625867 0  
## mileage -0.5160380 0  
## mpg -0.5809686 0

In this context, it’s evident that the price significantly influences the choice of car category. As the price increases, certain car models become increasingly likely choices compared to others. The same thing happens with the type of transmission.

res.condes$quali

## R2 p.value  
## model 0.520829965 0.000000e+00  
## f.year 0.340415103 0.000000e+00  
## f.price 0.697084268 0.000000e+00  
## f.miles 0.296621674 0.000000e+00  
## f.mpg 0.306345595 0.000000e+00  
## transmission 0.220025494 2.306211e-270  
## f.engineSize 0.179779108 9.040068e-216  
## manufacturer 0.080505068 1.417320e-90  
## f.tax 0.059143057 7.058305e-67  
## fuelType 0.007073811 9.692687e-08  
## Audi 0.003975412 8.131402e-06

There is a lot of information to deduce from this output:

* The price is much likely higher if it’s from 2020 year, and if the MPG is categorized as Low, and the engineSize is Large, if the car is New/Likely New (based on mileage discretization),
* The most expensive cars are: BMW- 8 Series, Audi- R8, VW- California, Audi- Q8, BMW- X6…
* Usually cars that are classed as hybrid tend to be more expensive.
* We can also check the cheapest car models that usually are manual transmission and categorized as affordable.

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

## Estimate p.value  
## model=Mercedes- G Class 124434.538822 3.354380e-32  
## model=Audi- R8 72157.788822 3.749753e-47  
## model=BMW- M5 40365.872156 4.734815e-14  
## model=BMW- X7 39468.138822 1.024919e-21  
## model=Audi- RS4 36436.538822 6.410382e-05  
## model=Audi- Q8 31282.538822 1.930085e-15  
## model=BMW- 8 Series 30420.205489 1.813806e-09  
## model=VW- California 27427.538822 5.807168e-06  
## model=BMW- X6 23078.253108 7.146147e-14  
## model=Audi- Q7 18860.026002 1.979555e-54  
## model=Mercedes- GLS Class 17731.605489 5.676987e-20  
## f.year=2020 17666.424269 3.823079e-70  
## model=Audi- RS5 17331.538822 2.061904e-02  
## model=Audi- RS6 17282.538822 3.723359e-06  
## model=BMW- M2 15184.038822 2.633929e-03  
## f.price=Expensive 14700.291377 0.000000e+00  
## f.year=2019 13933.236973 8.808905e-206  
## model=BMW- 7 Series 10740.205489 1.670965e-09  
## f.mpg=Low 10368.681899 0.000000e+00  
## model=BMW- M4 10144.824537 9.300966e-15  
## model=VW- Caravelle 9841.824537 5.119488e-10  
## model=BMW- X5 9517.372156 1.754686e-26  
## model=Audi- A8 8823.872156 5.335379e-08  
## model=Mercedes- S Class 8702.712735 6.479387e-14  
## f.miles=New/Nearly New 8657.176600 2.879661e-235  
## model=Mercedes- GLE Class 7433.557690 2.216762e-26  
## f.engineSize=Large 7420.722195 5.050978e-168  
## fuelType=Hybrid 5472.170676 2.083920e-05  
## f.tax=Low 4544.683844 1.435068e-48  
## model=Mercedes- SL CLASS 4526.238822 6.060965e-11  
## transmission=f.Trans-SemiAuto 4511.282762 4.575840e-119  
## f.year=2017 4453.911485 2.058227e-14  
## model=BMW- X4 3766.129732 1.297516e-07  
## model=VW- Touareg 3558.729299 6.077266e-13  
## manufacturer=Mercedes 3034.163366 6.291241e-37  
## transmission=f.Trans-Automatic 2735.801771 2.837568e-28  
## f.miles=Used 2579.037400 1.456895e-20  
## model=Mercedes- GLC Class 1765.434475 6.106381e-25  
## f.year=2016 1446.487565 9.289086e-58  
## manufacturer=Audi 1065.774262 8.131402e-06  
## f.price=Moderately priced 1044.051601 1.459049e-04  
## f.tax=High 928.035058 3.502435e-13  
## Audi=Audi Yes 878.066828 8.131402e-06  
## manufacturer=BMW 869.837649 1.284345e-04  
## model=BMW- Z4 636.253108 2.500517e-02  
## model=Audi- Q5 554.635910 5.731692e-18  
## f.year=2015 -9.832546 8.406766e-44  
## fuelType=Diesel -45.421061 5.239279e-04  
## Audi=Audi No -878.066828 8.131402e-06  
## f.engineSize=Medium -1090.556974 1.569369e-02  
## fuelType=Petrol -1352.180708 1.550615e-05  
## model=BMW- X2 -1426.169511 1.084891e-03  
## f.year=2014 -1931.137534 1.200559e-28  
## model=BMW- X3 -2101.746892 2.005844e-05  
## model=Mercedes- X-CLASS -2975.261178 4.019808e-02  
## f.miles=Old -3529.769000 2.359242e-37  
## model=Mercedes- V Class -3718.911178 3.777450e-02  
## f.year=2013 -4246.068612 7.918918e-32  
## model=Mercedes- CLS Class -4327.127844 4.402867e-02  
## f.mpg=High -4409.728455 1.645703e-56  
## f.price=Affordable -4718.097655 3.525089e-66  
## manufacturer=VW -4969.775277 8.411264e-87  
## f.tax=Medium -5472.718902 5.116203e-47  
## f.mpg=Very High -5590.278196 2.796377e-95  
## model=Mercedes- E Class -5887.769749 2.367795e-04  
## f.engineSize=Small -6330.165221 3.123262e-110  
## model=Mercedes- C Class -6660.649808 2.677191e-05  
## f.year=2010 -6672.685153 5.584784e-06  
## f.year=2012 -6676.552263 1.223403e-15  
## f.year=2011 -6749.679201 2.097540e-09  
## transmission=f.Trans-Manual -7247.084533 9.069704e-267  
## f.miles=Very Old -7706.445000 1.411152e-182  
## f.year=2009 -8461.570868 1.459745e-05  
## f.year=2008 -10655.220868 9.740801e-06  
## model=BMW- 3 Series -10798.469616 1.216820e-02  
## f.price=Low-priced -11026.245323 0.000000e+00  
## model=BMW- 2 Series -11574.095960 1.371921e-02  
## model=BMW- X1 -11769.015024 4.744161e-02  
## model=Audi- A3 -13437.482011 3.164646e-08  
## model=VW- Golf -13696.929805 3.852614e-23  
## model=BMW- 1 Series -14278.082182 1.767152e-12  
## model=VW- Passat -14705.281178 3.751870e-07  
## model=Audi- A1 -15944.337090 3.484825e-13  
## model=VW- Golf SV -15963.889749 4.860002e-03  
## model=VW- Scirocco -17626.275992 7.544347e-05  
## model=VW- Polo -19101.698982 5.192534e-64  
## model=Mercedes- SLK -19683.818320 4.244729e-04  
## model=VW- CC -20556.016733 2.256001e-03  
## model=VW- Beetle -22533.127844 3.587711e-05  
## model=VW- Up -22555.966672 6.750578e-31

Profiling binary factor “Audi?” it with all other variables:

res.catdes <- catdes(df,17,proba = 0.05)

We observe a relatively weak correlation between Y.bin-‘Audi’ and the other quantitative variables. However, the presence of very low p-values suggests that there is a connection. It’s important to note that while this connection exists, the limited sample size may prevent us from establishing it.

res.catdes$quanti.var

## Eta2 P-value  
## mpg 0.0092478291 9.489946e-12  
## price 0.0039754125 8.131402e-06  
## tax 0.0019457989 1.809295e-03  
## year 0.0007909104 4.675624e-02

Again, we can deduce plenty of information:

* A robust link emerges between this binary variable and the categories. Notably, Audi cars are distinctly associated with the ‘Medium Size’ engines, ‘Low’ mpg ratings, and the ‘Expensive’ category. Furthermore, they tend to favor manual transmission and ‘Petrol’ as their preferred fuel type.

# 5. Principal Component Analysis

## 5.1 Eigenvalues and dominant axes analysis:

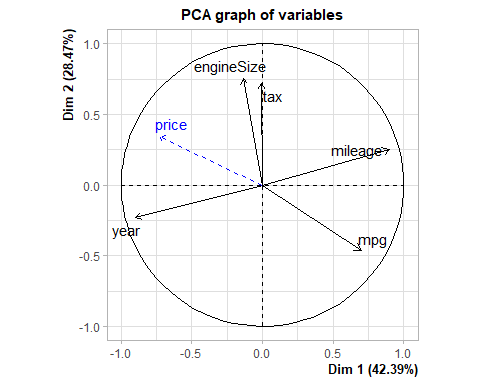
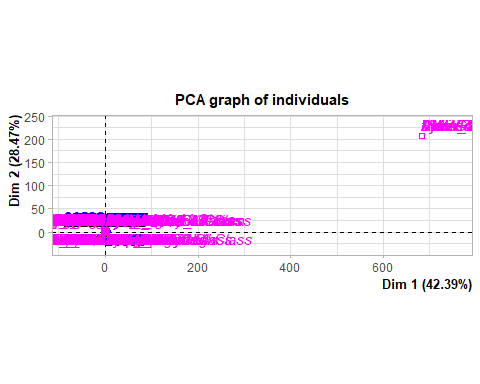
We are asked to perform a PCA taking into account also supplementary that can be quantitative and/or categorical.

We previously applied imputation on our dataframe, and now will apply PCA, passing all categorical/factor variable as qualitative supplementary variables, and pass the target variable “price” is our quantitative supplementary variable. We will also pass the detected multivariant outliers as supplementary individuals to avoid any anomalies.

We deduce from the following graph of variables and the results:

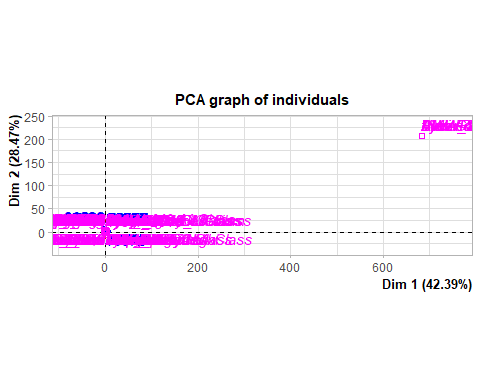
* The two first dimensions explain the 70% of inertia.
* The first component agglutinates 40,8% of variability meanwhile the second component has 30% of variability. We can sense that more than two--thirds of the variability are already inside the first and second component.
* The variables “mileage”, “year” and “price” have a significant impact on the first component, meanwhile “tax” and “engineSize” have an important effect on the second component.
* The variable “mpg” has an insignificant impact on both component compared to the rest.
* “mileage” and “year” are negatively correlated.

library(FactoMineR)  
res.pca<-FactoMineR::PCA(df, quali.sup=c(1,4,6,10:17), quanti.sup= c(3), ind.sup =ll)



* The following graph shows the relationship between individuals with the two axes. Proximity of points in the scatter plot indicates similarity between individuals. Individuals closer together are more similar in terms of the variables used in the analysis.
* In our case, we can spot graphically few individuals that do not follow the common pattern (11053 as an example). But we have many individuals that contribute almost equally to the inertia of each component.
* For example individuals with a model type “Audi-R8” contributes in the variability of the second dimension more than it does in the first component.

plot.PCA(res.pca,choix=c("ind"),invisible=c("ind"))



* As we can see below, 6 components have been created. The choice of retaining the most informative axes in a PCA analysis can be made using various methods, including but not limited to the Kaiser criterion and the Elbow method. These approaches assist in determining the optimal number of principal components to capture and retain, contributing to a more robust interpretation of the underlying patterns in the data.

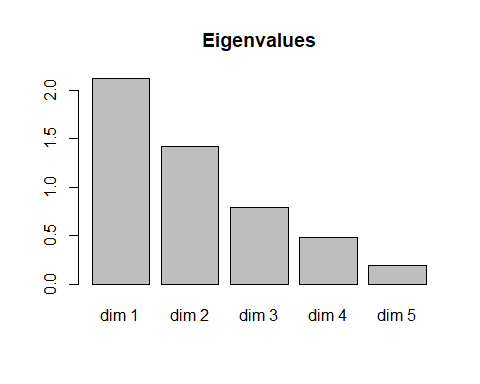
**Kaiser Criteria:**

* The PCA function yields an eigenvector with normalized eigenvalues. Employing the Kaiser criterion, we opt to retain the first two components, given that their eigenvalues surpass the mean of all components’ eignevalues, in our case 0.8.
* The two first componenets meet this criteria and have 70.79% of cumulative percentage of variance. This strategic selection ensures a focused representation of the data’s principal components, enhancing the interpretability of the analysis.

res.pca$eig

## eigenvalue percentage of variance cumulative percentage of variance  
## comp 1 2.1192922 42.385843 42.38584  
## comp 2 1.4237255 28.474510 70.86035  
## comp 3 0.7867375 15.734751 86.59510  
## comp 4 0.4797697 9.595395 96.19050  
## comp 5 0.1904750 3.809501 100.00000

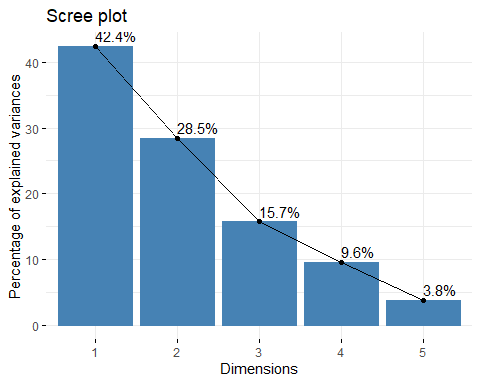
barplot(res.pca$eig[,1],main="Eigenvalues",names.arg=paste("dim",1:nrow(res.pca$eig)))



**Elbow Method:**

* The following graph shows the Eignevalues in a downward curve, from highest to lowest, and by using the Elbow Method we can determine the number of significant axes in this case, we would retain 3 axes.
* The three components englobe 86,4% of the information.

library("factoextra")  
fviz\_eig(res.pca, addlabels = TRUE)



## 5.2 Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables:

**Variables Coordinates:**

* The values in this matrix indicate the strength of the relationship between each variable and the principal components through coordinates values. Notably, the first principal component (Dim.1) exhibits a strong correlation with the ‘year’ and ‘mileage’ variables. This also implies that Dim.1 captures information variability related to the year and mileage of the vehicles. Additionally, the second principal component (Dim.2) shows a notable positive correlation with ‘tax’ and ‘engineSize,’ while being negatively correlated with ‘mpg.’

res.pca$var$cor

## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5  
## year -0.899785514 -0.2315976 0.03533354 0.2047096 0.305931560  
## mileage 0.896593780 0.2504712 -0.08791196 -0.1713041 0.310338810  
## tax -0.006315433 0.7199024 0.67288897 0.1699056 0.007290117  
## mpg 0.697610289 -0.4640649 0.05219770 0.5433127 -0.008388018  
## engineSize -0.138194696 0.7574540 -0.56767632 0.2906235 -0.021148043

**Quality of representation:**

* To measure the quality of representation of each variable on the principal components we use the squared cosines. It provides insights into how well each variable is represented in the reduced-dimensional space created by the principal components.
* In the first dimension, both “year” and “mileage” exhibit strong representation, indicating that they play a substantial role in shaping this principal component. Conversely, “tax” contributes minimally, accounting for only 1% of the representation in this specific dimension. This insight underscores the differential impact of these variables on the overall structure captured by the principal components, emphasizing the significance of “year” and “mileage” in this particular dimension.
* When it comes to the second dimension, “tax,” “mpg,” and “engineSize” showcase some representation, contributing to the structure of the second principal component. However, it’s important to note that “year” and “mileage” have limited influence in this specific dimension.

res.pca$var$cos2[,1:2]

## Dim.1 Dim.2  
## year 0.8096139720 0.05363744  
## mileage 0.8038804063 0.06273580  
## tax 0.0000398847 0.51825950  
## mpg 0.4866601154 0.21535619  
## engineSize 0.0190977739 0.57373659

**Contribution of variables:**

* Analyzing variable contributions provides insights into which variables strongly influence the selected axes. This information aids in interpreting the meaning of each dimension and helps focus on key variables.

res.pca$var$contrib[,1:2]

## Dim.1 Dim.2  
## year 38.202093615 3.767400  
## mileage 37.931552073 4.406454  
## tax 0.001881982 36.401644  
## mpg 22.963333058 15.126243  
## engineSize 0.901139273 40.298258

* In the context of the first principal component, it’s evident that “year” and “mileage” exhibit the highest contributions, aligning with our earlier observation from squared cosines. Conversely, “tax” and “engine size” make minimal contributions, indicating their limited impact on this principal component.
* Regarding the second principal component, “tax” and “engine size” are contributing more significantly compared to other variables. In contrast, “year” and “mileage” show comparatively lower contributions, underscoring their reduced influence on the second principal component. This also matches with previous result we got during the quality representations of variables.

To double-check our earlier findings, we can examine the correlation between each variable and the principal components. This additional step ensures consistency with our previous results:

res.des<-dimdesc(res.pca)  
res.des$Dim.1$quanti

## correlation p.value  
## mileage 0.8965938 0.000000e+00  
## mpg 0.6976103 0.000000e+00  
## engineSize -0.1381947 7.179433e-22  
## price -0.7327479 0.000000e+00  
## year -0.8997855 0.000000e+00

## 5.3 Individuals point of view

Individual analysis in PCA is crucial for understanding how each observation or individual contributes to the overall variation in the dataset and how they are positioned in the reduced-dimensional space defined by the principal components.

**Coordinates analysis:**

* These coordinates express the position of each individual in the reduced-dimensional space created by the principal components. Examining these coordinates helps in visualizing the distribution of individuals in the PCA plot and understanding the relationships and patterns in the data captured by the principal components.
* Examining the top records based on Dim.1 reveals that individuals with IDs 44400, 39750, and 9823 have notably high positive scores on this component. This suggests that these individuals contribute significantly to the variance captured by Dim.1. On the other hand, reviewing the top records based on Dim.2, individuals with IDs 31494, 30583, and 7446 exhibit high positive values, indicating their substantial influence on the variability captured by Dim.2. Conversely, individual 7884 stands out with a high positive value on Dim.2 but in the opposite direction.

head(res.pca$ind$coord[order(-res.pca$ind$coord[, 1]), 1:2])

## Dim.1 Dim.2  
## 40051 4.781053 0.08106568  
## 7951 4.700787 0.04667048  
## 39715 4.448594 0.32814005  
## 9607 4.378025 -0.10136739  
## 39752 4.329937 1.71607910  
## 9633 4.242106 -0.14073144

head(res.pca$ind$coord[order(-res.pca$ind$coord[, 2]), 1:2])

## Dim.1 Dim.2  
## 11766 2.349787 5.511234  
## 21413 2.240263 5.416913  
## 32548 1.821140 5.331052  
## 11797 1.877107 5.293137  
## 4692 1.868136 5.249974  
## 21598 1.304131 5.163942

**Quality of representation:**

* Analyzing the top records based on Dim.1 squared cosines, it is evident that individuals with IDs 45095, 45035, and 15725 have extremely high values close to 1. This indicates a strong and accurate representation of these individuals along Dim.1. Similarly, reviewing the top records based on Dim.2 squared cosines, individuals 17858, 12277, and 10417 exhibit high values close to 1, signifying an excellent representation along Dim.2.

head(res.pca$ind$cos2[order(-res.pca$ind$cos2[, 1]), 1:2])

## Dim.1 Dim.2  
## 19302 0.9955531 4.142808e-03  
## 9826 0.9939067 3.760972e-03  
## 7951 0.9935953 9.793838e-05  
## 9607 0.9925819 5.321162e-04  
## 9584 0.9877533 1.105616e-03  
## 24186 0.9876881 9.525401e-04

head(res.pca$ind$cos2[order(-res.pca$ind$cos2[, 2]), 1:2])

## Dim.1 Dim.2  
## 17858 0.0005470026 0.9357183  
## 12277 0.0091690400 0.9254502  
## 10417 0.0023268795 0.9238301  
## 14597 0.0076065878 0.9121488  
## 49398 0.0092277294 0.8985328  
## 18431 0.0265945077 0.8978619

**Contribution of individuals:**

* Analyzing the top records based on Dim.1 contributions, it is notable that individuals with IDs 44400, 39750, and 9823 have the highest contributions to the variability along Dim.1. In particular, individual 44400 stands out with a substantial contribution of approximately 29.1%, emphasizing its significant role in explaining the variance along Dim.1. Similarly, reviewing the top records based on Dim.2 contributions, individuals 31494, 30583, and 7446 exhibit the highest contributions, suggesting their prominent influence on the variability captured by Dim.2.

head(res.pca$ind$contrib[order(-res.pca$ind$contrib[, 1]), 1:2])

## Dim.1 Dim.2  
## 40051 0.2250344 9.630313e-05  
## 7951 0.2175419 3.191912e-05  
## 39715 0.1948262 1.577919e-03  
## 9607 0.1886940 1.505785e-04  
## 39752 0.1845716 4.315597e-02  
## 9633 0.1771596 2.902342e-04

head(res.pca$ind$contrib[order(-res.pca$ind$contrib[, 2]), 1:2])

## Dim.1 Dim.2  
## 11766 0.05435741 0.4451065  
## 21413 0.04940828 0.4300016  
## 32548 0.03265038 0.4164781  
## 11797 0.03468803 0.4105751  
## 4692 0.03435728 0.4039062  
## 21598 0.01674342 0.3907770

**Analyzing 6 individuals that have a significant contribution to the first component:**

* This results matches the outcome of the variable analysis we did previously, where we concluded that mileage and year are the variables that contributed more in the first component.
* Its observed that cars that mostly contributed to the first component have a really high mileage and and their year is very far so they are very old.

df[which(row.names(df) %in% c(39721 , 18027 , 49012, 9823, 39750, 44400 )), ]

## model year price transmission mileage fuelType tax  
## 18027 BMW- 1 Series 2010.79 2495 f.Trans-Manual 112000.00 Diesel 200.0000  
## 49012 VW- Touran 2008.00 2300 f.Trans-Manual 112304.00 Petrol 162.4420  
## 39721 VW- Golf 2010.00 4995 f.Trans-Manual 111913.00 Diesel 145.0000  
## 9823 Audi- A6 2008.00 2490 f.Trans-Manual 95062.65 Diesel 200.0000  
## 39750 VW- Golf 2011.00 4295 f.Trans-Manual 105000.00 Diesel 147.1385  
## 44400 VW- Polo 2009.00 2995 f.Trans-Manual 89000.00 Diesel 147.4458  
## mpg engineSize manufacturer f.year f.price f.miles f.tax f.mpg  
## 18027 49.6 2.0 BMW 2011 Low-priced Very Old High Moderate  
## 49012 34.5 1.6 VW 2008 Low-priced Very Old High Low  
## 39721 52.3 2.0 VW 2010 Low-priced Very Old Low Moderate  
## 9823 44.1 2.0 Audi 2008 Low-priced Very Old High Low  
## 39750 74.3 1.6 VW 2011 Low-priced Very Old Medium Very High  
## 44400 72.4 1.4 VW 2009 Low-priced Very Old High Very High  
## f.engineSize Audi  
## 18027 Medium Audi No  
## 49012 Medium Audi No  
## 39721 Medium Audi No  
## 9823 Medium Audi Yes  
## 39750 Medium Audi No  
## 44400 Small Audi No

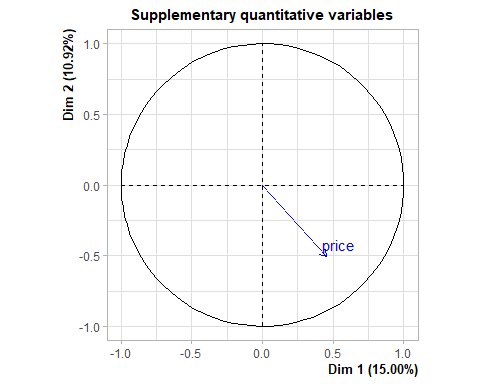
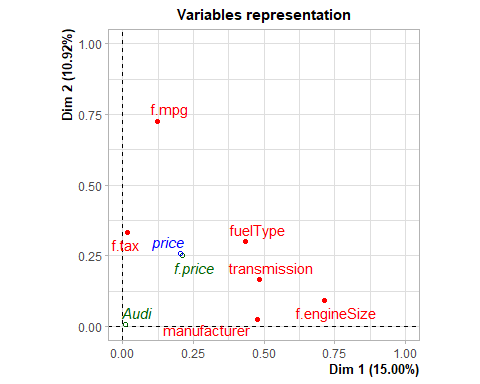
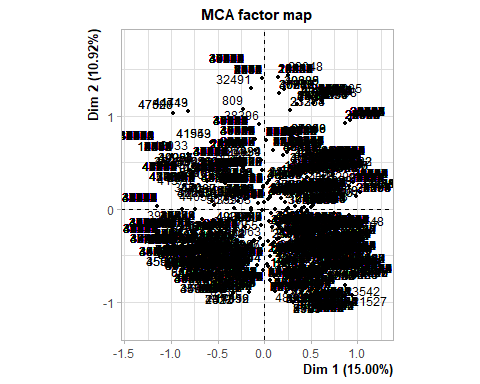
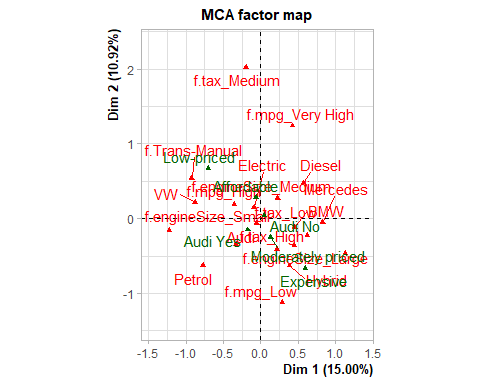
# 6 Multiple Correspondence Analysis

## 6.1 MCA & Eigenvalues & dominant axes analysis

* We’ll utilize a dataframe free of previously identified multivariate outliers to avoid anomalies. The variable “price” will serve as a supplementary quantitative variable, while “f.price” and the binary target “Audi” will function as supplementary qualitative variables. We will also discard f.year and f.miles as they are very related to f.price as we spotted previously.

library(FactoMineR)  
library(factoextra)  
x<-df[,c(3,4,6,10, 12, 14:17)] #Does not include Multivariant Outliers  
res.mca<-MCA(x[-ll,], quanti.sup = c(1), quali.sup = c(5,9))

## Warning: ggrepel: 3 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



* Based on Kaiser Criteria, 7 components should be retained.

length(which(res.mca$eig[,1] > mean(res.mca$eig[,1])))

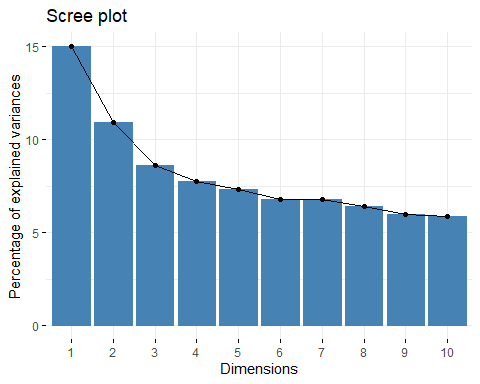
## [1] 7

* In 7 components, it is accumulated 63.13% of variance.

res.mca$eig[1:7,]

## eigenvalue percentage of variance cumulative percentage of variance  
## dim 1 0.3749519 14.998076 14.99808  
## dim 2 0.2728933 10.915730 25.91381  
## dim 3 0.2152289 8.609154 34.52296  
## dim 4 0.1930482 7.721930 42.24489  
## dim 5 0.1822856 7.291424 49.53631  
## dim 6 0.1700351 6.801404 56.33772  
## dim 7 0.1692793 6.771171 63.10889

fviz\_eig(res.mca)



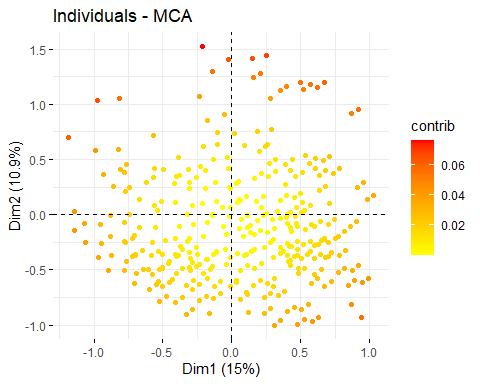
## 6.2 Individuals Point of View

* **Are there any individuals “too contributive”?**
* As we can see, There are some individuals that contribute more in the first component, and others that that do the same in the second component. We can also state the existence of many individuals that contribute equally in both components.

head(res.mca$var$contrib)

## Dim 1 Dim 2 Dim 3 Dim 4 Dim 5  
## f.Trans-Manual 13.617299443 6.337365642 0.095975242 0.16665001 0.4265806  
## f.Trans-SemiAuto 3.435585583 2.940765694 0.099807995 0.85417281 13.7181424  
## f.Trans-Automatic 4.448703962 0.791889618 0.000336414 2.59156179 13.9988314  
## Diesel 8.109997709 7.914145545 0.238732846 0.04850964 0.7663800  
## Electric 0.001127977 0.004252932 0.810558937 15.37618553 12.3930178  
## Hybrid 0.070993242 0.268290041 0.133329382 13.57336616 2.5884583

fviz\_mca\_ind(res.mca, geom=c("point"),col.ind="contrib", gradient.cols =  
c("yellow", "red"))

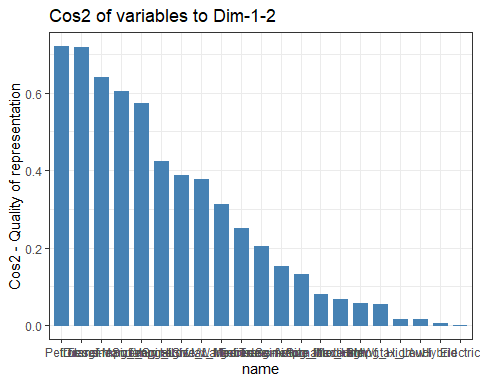


* **Are there any groups?**
* From the the previous graph, we can spot there a few individuals that can form a group as they are scattered equally without any specific pattern.
* As we can see in the following output table, categories tend to contribute equally or contribute very low in different dimensions, so there is not grouping pattern in the plots.

head(res.mca$var$cos2)

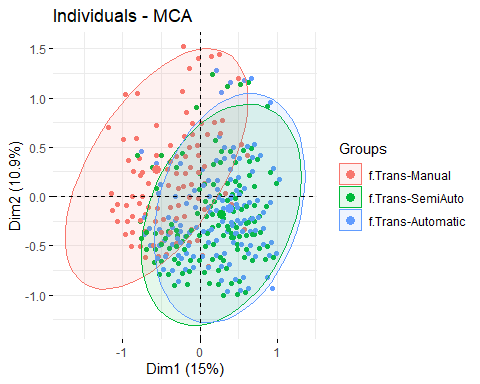
## Dim 1 Dim 2 Dim 3 Dim 4  
## f.Trans-Manual 4.790653e-01 1.622668e-01 1.938152e-03 0.003018557  
## f.Trans-SemiAuto 1.251113e-01 7.794235e-02 2.086345e-03 0.016015181  
## f.Trans-Automatic 1.347466e-01 1.745686e-02 5.849022e-06 0.040414396  
## Diesel 4.200241e-01 2.983148e-01 7.097258e-03 0.001293516  
## Electric 2.545589e-05 6.985441e-05 1.050020e-02 0.178659859  
## Hybrid 1.615001e-03 4.441991e-03 1.741032e-03 0.158976798  
## Dim 5  
## f.Trans-Manual 0.007295947  
## f.Trans-SemiAuto 0.242866696  
## f.Trans-Automatic 0.206135528  
## Diesel 0.019296319  
## Electric 0.135969654  
## Hybrid 0.028626872

fviz\_cos2(res.mca, choice = "var", axes = 1:2)+theme\_bw()

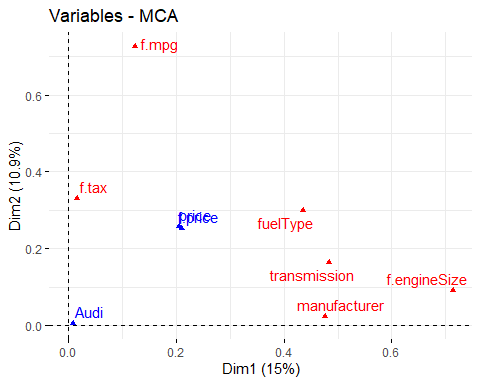


* We can see through the graph that no individual groups are spotted base on each variable, the only one that can show a better grouping is if we depend on transmission types, we can split individuals into two groups Manual and Automatic/Semi-Automatic.
* Please check Annex, to chaeck the other grouping graphs of individuals:

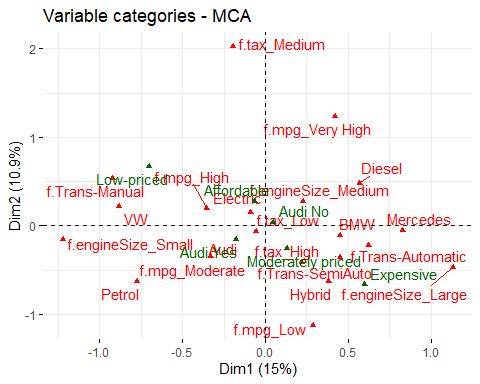
grp<- df[-ll,]$transmission  
fviz\_mca\_ind(res.mca, label="none", habillage = grp, addEllipses = TRUE)



## 6.3 Interpreting map of categories:

* Map of variables: **transmission**, **manufacturer** and **f.engineSize** are better represented in Dim1 , meanwhile **f.tax** and **f.mpg** are better represented in Dim2.
* **f.fueltype** is represented equally and insignificantly compared to other variables.
* fviz\_mca\_var(res.mca, choice="mca.cor", repel=TRUE)
* 
* **f.tax\_Mediumf.mpg\_Very High** categories are better represented in Dim2 is represented
* We can see that **Cheap/Very Chea** **Manual** Transmission and **Small** EngineSize are close to each other and contribute negatively at same way in Dim1.
* **Mercedes**, **Automatic** transmission, **Hybrid** and **Large Engine Size**, **very Expensive** are also gathered in the same area of the Map which makes sense, and contribute positvely in Dim1.

fviz\_mca\_var(res.mca, repel=TRUE)



## 6.4 Interpreting the axes associations to factor map

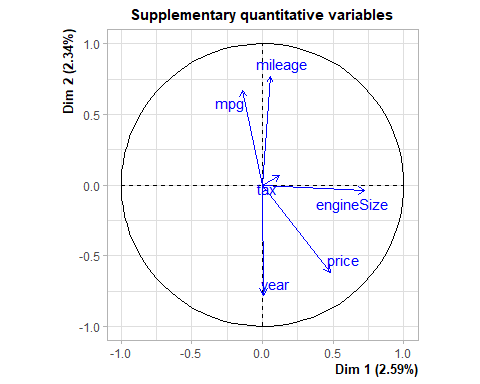
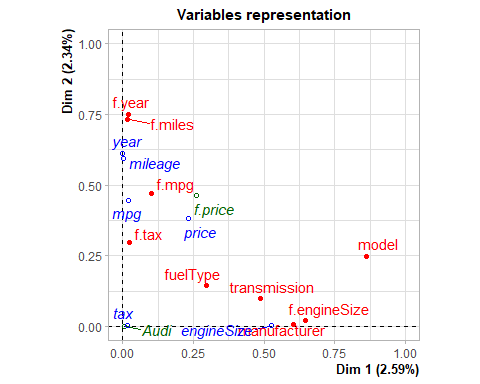
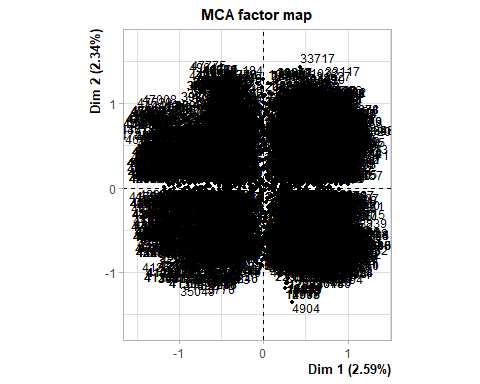
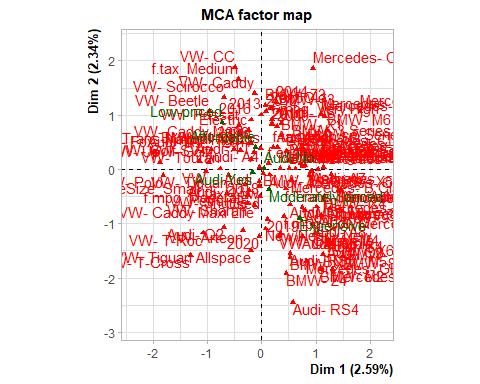
* The following result gives us an insight regarding the variables and categories that are related to the two retained axes:
* Dim1:
  + Variables: **f.engineSize** with R-Squared value of 0.71 and **transmission type** (0.50)
  + Categories: **f.engineSize= f.engineSize\_Large** and **transmission=f.Trans-Manual**

res.desc <- dimdesc(res.mca, axes = c(1,2))  
#res.desc[[1]]

* Dim2:
  + Variables: **f.mpg** with a R-Squared value of 0.73 and **fuelType** (0.35)
  + Categories: **f.mpg=f.mpg\_Very High** and **fuelType=Petrol**

#res.desc[[2]]

## 6.5 MCA with all variables

* We will try MCA taking into account all numerical variables:
* res.mca2 <- MCA(df, quanti.sup=c(2,3,5,7:9),  
  quali.sup=c(12,17))
* 
* Dim1:
  + Variables: **f.engineSize** has a Rsquared value of 0.72 and **transmission** (0.50)
  + Categories: **f.engineSize= f.engineSize\_Large** and **transmission=f.Trans-Manual**

res.desc2 <- dimdesc(res.mca2, axes = c(1,2))  
#res.desc2[[1]]

* Dim2:
  + Variables: **f.mpg** with a R-Squared value of 0.73 and **fuelType** (0.35)
  + Categories: **f.mpg=f.mpg\_Very High** and **fuelType=Petrol**

#res.desc2[[2]]

**Conclusions:**

* Both MCA analysis (either with all numerical variables as supplementary or not) has the same results on both axes. So there was no enhancement in the axis interpretation.

# 7 K-Means Classification:

## 7.1 Optimal Number of Clusters:

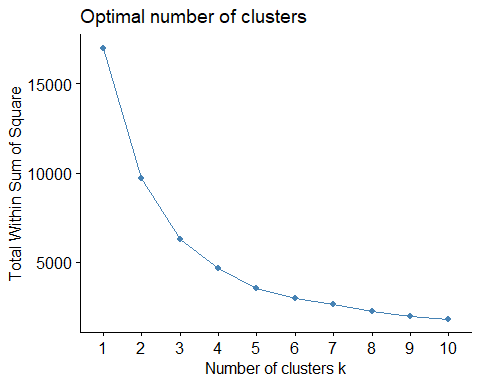
* At this point, after applying the PCA, and retaining the the first and second axes based on Kaiser Criteria we will process with clustering our data by using K-Means:

res.pca<-FactoMineR::PCA(df, quali.sup=c(1,4,6,10:17), quanti.sup= c(3), ind.sup = ll, graph = FALSE)  
ppcc<-res.pca$ind$coord[,1:2] # 2 components principals (based on kaiser criteria)  
dim(ppcc)

## [1] 4793 2

* Using the elbow method we can expect that the optimal number of cluster is 5, as the graph shows that the total of within sum of square starts to slow down.

#Optimal number of clusters  
library("factoextra")  
fviz\_nbclust(ppcc, kmeans, method = "wss")



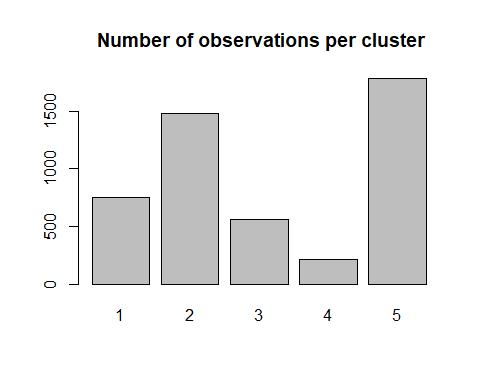
* To understand how the K-means Algorithm was used with 5 clusters, we can see in which iteration it starts to converge, in our case it was in the 4th iteration.

dist<-dist(ppcc) # coordenates are real - Euclidean metric  
kc<-kmeans(dist,5,iter.max=30,trace=TRUE)

## KMNS(\*, k=5): iter= 1, indx=0  
## QTRAN(): istep=4793, icoun=61  
## QTRAN(): istep=9586, icoun=13  
## QTRAN(): istep=14379, icoun=307  
## QTRAN(): istep=19172, icoun=2  
## QTRAN(): istep=23965, icoun=104  
## QTRAN(): istep=28758, icoun=202  
## QTRAN(): istep=33551, icoun=110  
## QTRAN(): istep=38344, icoun=547  
## QTRAN(): istep=43137, icoun=2821  
## KMNS(\*, k=5): iter= 2, indx=25  
## QTRAN(): istep=4793, icoun=2077  
## QTRAN(): istep=9586, icoun=2442  
## KMNS(\*, k=5): iter= 3, indx=4793

* As we set the number of clusters at 5, we can display the number of observations in each cluster as follows:

barplot(table(factor(kc$cluster)),main= " Number of observations per cluster")



## 7.2 Clustering Quality:

* The next chunk shows the quality of clustering. 77.84% is a higher percentage that suggests good and meaningful separation of clusters, indicating that the clustering is explaining a significant portion of the variance in the data.
* 100\*(kc$betweenss/kc$totss)
* ## [1] 77.65374

## 7.3 Clusters Description:

* We will proceed with describing and analyzing each cluster.
* We will assign to each individual in our original datafram with its own K-Means Cluster number. We are not considering Any Multivariant Outliers to avoid anomalies.

save\_df<-df  
df<-df[-ll,]  
df$claKM<-kc$cluster  
df$claKM<-factor(df$claKM)  
  
#res.cat <-catdes(df,18) #!8 is claKM variables, representing each individuals corresponding cluster.

### 7.3.1 **Description of clusters in relation with catgorical variables:**

* As we can see below (Please check [7. Annex:], as the output was very long), we find some categorical variables and factor very related to the cluster variable that we recently created. The low p-values show evidence how some variables have been significant to cluster our data.
* Regarding the first cluster, 78,78% are not Audi cars, 68,09% of the individuals are categorized as “**Low-Tax**”, 50,52% are categorized as “Medium Engine-size” and 56,48% have Diesel as Fuel-Type. Cluster 1 exhibits a distinct profile characterized by a prevalence of newer cars (**2019** and **2020** models) with **low** taxation, **moderate** mpg, and **moderate** pricing. Specific models like **VW-T-Cross** and **VW-T-Roc** show a perfect association with this cluster. Conversely, **very old** category as mpg, **low-price** cars, and certain models like **BMW-X6** and **BMW-M3** are notably absent.
* When it comes to Cluster 2: 83.71% of the cars in this cluster sample, were labelled as **Low-MPG**, 80,34% are labelled as Expensive and 73,27% have **Large Engine-size**. This make sense as cars with large engine size and low mpg are meant to be expensive. Cluster 2 is characterized by several distinct features. It has a strong association with low fuel efficiency (**f.mpg=Low**), expensive pricing (**f.price=Expensive**), and large engine sizes (**f.engineSize=Large**). Specific car models such as VW Touareg, Audi Q7, and BMW M4 are highly prevalent in this cluster. Vehicles in this cluster often have **new** or **nearly new** caracterisitcs. (**f.miles=New/Nearly New**). The presence of semi-automatic transmissions (**transmission=f.Trans-SemiAuto**) is also notable.
* Moreover, the cluster exhibits preferences for specific years, especially 2019 (**f.year=2019**). High tax rates (**f.tax=High**) and certain fuel types like Petrol (**fuelType=Petrol**) are significant features. Luxury brands such as BMW and Mercedes-Benz are well-represented in this cluster. Additionally, there is a prevalence of diesel fuel type (**fuelType=Diesel**) and hybrid fuel type (**fuelType=Hybrid**).
* Overall, Cluster 2 gathers expensive, fuel-inefficient vehicles with large engine sizes, especially those from luxury brands.
* Cluster 3: 70% are **Low Tax**, 27,92% have small Engine-Size Type, and 50,52% have a **medium Engine-Size**, 56,48 are **Diesel** in Fuel-type and 41,89% are **Petrol** cars, an d around 3 quarters fall in the catgegories combined “**Used**”, “**Old**” and “**Very Old**”, this also makes sense with the previous stated categories.
* Cluster 3 is characterized by older vehicles (**f.miles=Old**) and an emphasis on specific years, notably 2017 and 2016 (**f.year=2017**, **f.year=2016**). High fuel efficiency (**f.mpg=High**) and affordability (**f.price=Affordable**) are prominent features. This cluster also includes manual transmissions (**transmission=f.Trans-Manual**) and smaller engine sizes (**f.engineSize=Small**). Noteworthy associations involve manufacturers like VW and specific models such as VW Polo.
* Overall, this cluster represents a distinct group with a focus on older, fuel-efficient, and affordable vehicles.
* Cluster 4 exhibits distinctive features such as very old vehicles (**f.miles=Very Old**) and low-priced options (**f.price=Low-priced**), both strongly associated with the cluster. The statistical significance is evident with low p-values, notably for variables like **f.miles=Very Old** (p=0.0). Specific manufacturers (**manufacturer=Mercedes**, **manufacturer=BMW**) and models (**model=VW- Passat**, **model=BMW- 1 Series**) contribute significantly.
* Cluster 5 is characterized by high tax rates (**f.tax=High**) and a preference for very old vehicles (**f.miles=Very Old**). Large engine sizes (**f.engineSize=Large**) and low fuel efficiency (**f.mpg=Low**) are notable features. The cluster is associated with specific years, especially 2011 and 2009 (**f.year=2011**, **f.year=2009**). Automatic transmissions (**transmission=f.Trans-Automatic**) and moderate fuel efficiency (**f.mpg=Moderate**) are also prevalent. This cluster includes luxury models like Mercedes-M Class and Audi Q5. Notably, the presence of **Audi Yes** indicates a distinction for Audi vehicles. Overall, Cluster 5 represents a group with a focus on high tax rates, older vehicles, and larger engine sizes.

### 7.3.2 **Description of clusters in relation with numerical variables:**

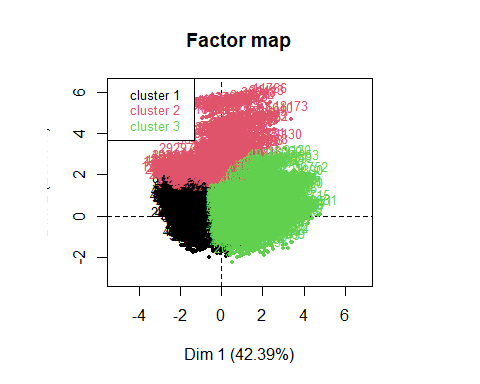
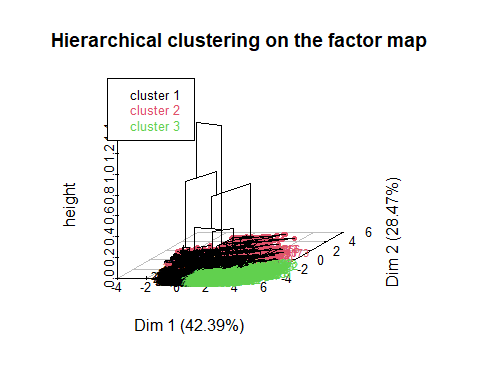
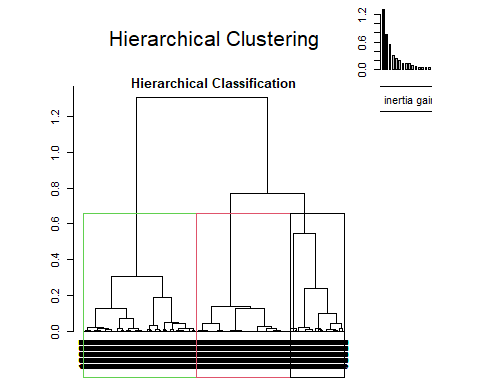
* In Cluster 1, the average price is 16190.02, which is lower than the overall mean of 21,600. These cars have a lower tax rate of 143.09 compared to the overall mean of 146.92. The engine size is smaller, with an average of 1.71, and the fuel efficiency (MPG) is slightly higher at 59.97. The mileage is relatively higher, with an average of 26134 These results match with the cluster description with categorical variables which makes sense. This cluster gathers mainly low tax cars with low prices small engine sizes.
* Cluster 2 is characterized by cars with an average engine size (1.84) and lower-than-average prices (11786). The tax rate is slightly lower than the overall mean at 146.92. Mileage is higher (54501), and the fuel efficiency (MPG) is higher 62.68.
* Cluster 3 includes cars with lowe fuel efficiency (MPG) at 42.59 and high mileage, around 52520. Tax rates are higher than average at 180.35, and the engine size is bigger (2.59). The average price is 16,190, significantly lower than the overall mean.
* Cluster 4 consists of cars with very low mileage (7013.41), slightly low fuel efficiency (MPG) at 48.79, and a relatively small engine size (1.73). Tax rates are low at 145.53, and prices are a bit higher than average.
* Cluster 5 features cars with a an average tax rate, very low mileage (11669.23), and big engine sizes (2.71). Fuel efficiency (MPG) is low at 37.56, and prices are significantly higher than the overall mean.

# 8 Hierarchical Clustering:

## 8.1 Number of Clusters :

* As we observe below, using HCPC, has implicitly detected the optimal number of clusters through the inertia gain barplot, in this case we have three clusters. The quality of this partition is around 41.08%, so we will try to increase it.

res.hcpc <- HCPC(res.pca,nb.clust = -1, order = TRUE)



((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[3])/  
res.hcpc$call$t$within[1])\*100

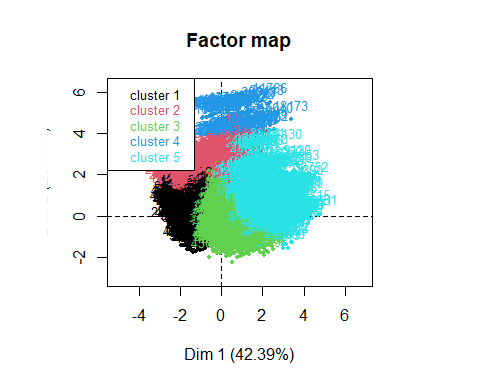
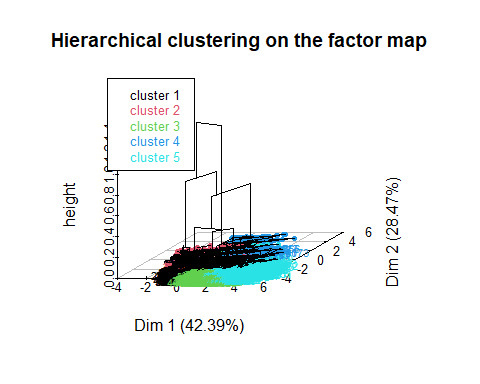
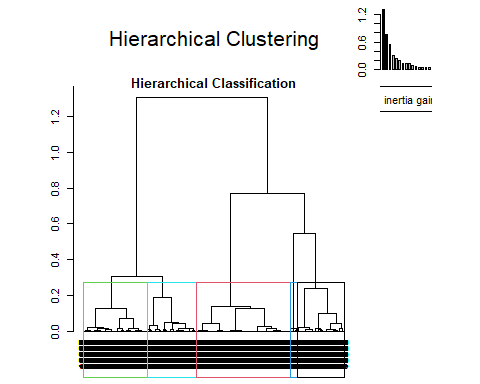
## [1] 41.54846

## 8.2 Clustering Quality:

We will try now with 5 clusters:

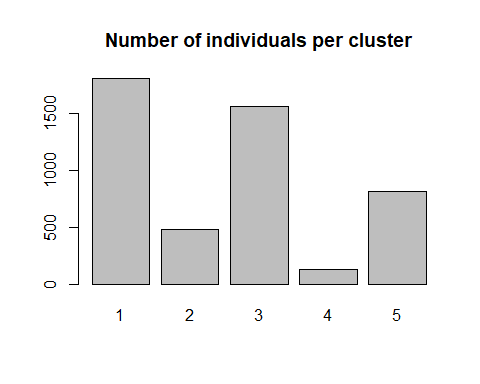
* As we can see the quality of partition has now increased to 55.55%, in case if we want to achieve a quality of 80%, we need 16 clusters at least. In our study, we will keep data in 5 cluster to facilitate the process of study, comparison and analysis.

res.hcpc <- HCPC(res.pca, nb.clust = 5, order = TRUE)



((res.hcpc$call$t$within[1]-res.hcpc$call$t$within[5])/  
res.hcpc$call$t$within[1])\*100

## [1] 58.60638

* We can visualize hos many individuals has each cluster:
* table(res.hcpc$data.clust$clust)
* ##   
  ## 1 2 3 4 5   
  ## 1807 481 1562 131 812
* barplot(table(res.hcpc$data.clust$clust), main = "Number of individuals per cluster")
* 

## 8.3 Clusters Description:

* We will assign to each individual of our dataframe the number of its cluster.

### 8.3.1 **Description of clusters in relation with categorical variables:**

* Description of each cluster in relation with other categorical/factor variables, please check [7. Annex:] for the following output.
  + Notably, Cluster 1 (Cla/Mod) exhibits a strong association with cars from the year 2019, emphasizing a preference for new or nearly new vehicles. Characteristics such as low mileage (**f.miles=New/Nearly New**), moderate fuel efficiency (**f.mpg\_Moderate**), and an inclination towards expensive cars (**f.price=Expensive**) contribute to the distinctive profile of this cluster. Additionally, there is a notable dominance of Volkswagen (**manufacturer=VW**) models, particularly VW-T-Roc and VW-T-Cross. Transmission type (**f.Trans-SemiAuto**) and medium engine size (**f.engineSize\_Medium**) These cars often have **low** tax rates and run on **petrol**.
  + In Cluster 2, the analysis reveals a distinct automotive profile. Cars with **large engine size**, specifically the BMW M4 and Mercedes GLS Class, dominate, showcasing a 100% modulation score. High-end features like **expensive** pricing, **low** fuel efficiency (mpg), and **semi-automatic** transmissions are strongly associated. Notably, the VW Touareg, Audi Q7, BMW M4, and Mercedes GLS Class are the top contributing models. **Diesel** fuel types and **manual transmissions** play a significant role. **BMW** and **Mercedes** emerge as the primary manufacturers, especially in the year **2019** with **New/Nearly New** based on mileage. **High tax** brackets and **expensive** pricing align with this cluster, emphasizing luxury and performance in vehicle characteristics.
  + Cluster 3 reflects a car segment characterized by specific features. **Old** vehicles with **high** mileage dominate. Cars with very **high** fuel efficiency (mpg) are prevalent. The year **2017** and **2016** stand out, indicating a preference for recent models. **Affordable** and **low-priced** options are significant, aligning with a budget-conscious consumer base. **Manual transmissions** and **small** engine sizes are notable features. **Diesel** fuel types are prevalent. The VW Polo emerges as a popular model, epitomizing the practical and efficient characteristics of this cluster.
  + Cluster 4 represents a distinct car segment characterized by specific attributes. **High tax** rates are a defining feature, suggesting a preference for luxury or high-performance vehicles. **Moderate fuel efficiency** (mpg) is observed, indicating a balance between performance and economy. **Very old** cars are significant in this cluster. The Audi Q5 and BMW X5 emerge as prominent models. The year **2015** stands out, reflecting a preference for relatively recent models. **Large engine sizes** are notable. Low fuel efficiency (**Low mpg**) and **high prices** characterize this cluster, aligning with a segment that values performance and luxury over fuel economy and affordability.
  + Cluster 5 is characterized by **very old**, **low-priced** cars with model years **2013**-**2015**, **high** tax rates, and a preference for **diesel** fuel. These cars exhibit very **high** fuel efficiency (mpg) and often feature **manual transmissions**. Prominent models include Audi A6, BMW 5 Series, and Mercedes SLK, reflecting diversity in manufacturers. The cluster encompasses a mix of **diesel** and **petrol** fuel types, various transmission preferences, and a range of budget options. Cars from both **used** and **new/nearly new** categories contribute to this cluster, highlighting a diverse set of preferences.

### 8.3.2 **Description of clusters in relation with numerical variables:**

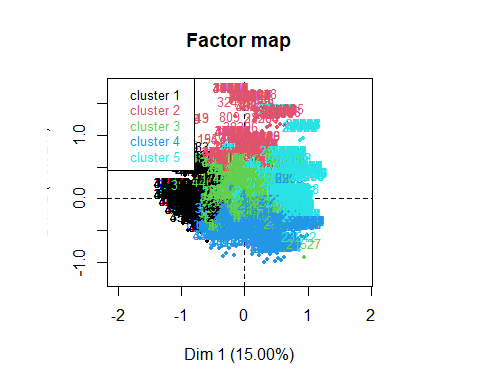
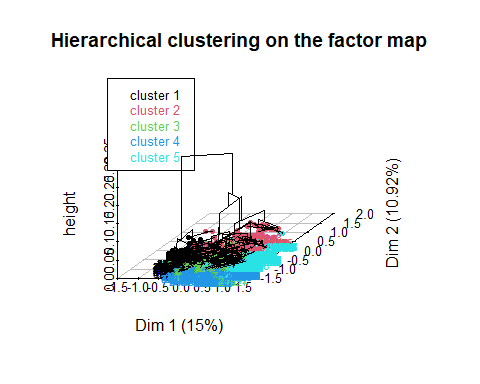
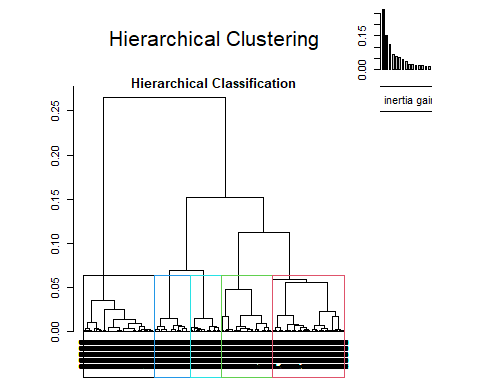
* Now we will proceed to describe the relationship between numerical variables and these hierarchical clusters:
* # res.hcpc$desc.var$quanti Please check annex.
  + This cluster represents vehicles with a relatively recent manufacturing year (2018.91), higher prices (25074.78), lower taxes (145.86), smaller engine sizes (1.75), lower fuel efficiency (46.63 MPG), and lower mileage (7082.22). This cluster represents recent models (mean year 2018.91) with higher prices and smaller engine sizes. Despite lower taxes, these vehicles have lower fuel efficiency and mileage, suggesting a preference for newer cars.
  + Vehicles in this cluster have larger engine sizes (3.08), higher prices (40327.40), slightly higher manufacturing years (2017.93), slightly higher taxes (148.47), lower mileage (14545.83), and lower fuel efficiency (38.94 MPG). Cluster 2 showcases larger, more powerful vehicles with higher prices and slightly newer manufacturing years. Despite slightly higher taxes, these cars exhibit lower mileage and fuel efficiency, appealing to those seeking performance and modern features.
  + this cluster is characterized by vehicles with higher fuel efficiency (61.75 MPG), higher mileage (25448.23), lower manufacturing years (2016.64), lower taxes (143.08), smaller engine sizes (1.70), and lower prices (16126.43). Vehicles in this cluster stand out for their higher fuel efficiency and mileage, lower manufacturing years, and smaller engine sizes. With lower taxes and prices, these cars are likely more economical and environmentally friendly, attracting a conscious consumer base.
  + Vehicles in this cluster have higher taxes (199.93), higher mileage (42536.34), larger engine sizes (2.36), slightly lower fuel efficiency (44.60 MPG), and manufacturing years slightly below the overall mean. Cluster 4 is characterized by higher taxes, elevated mileage, and larger engine sizes. Although slightly below the overall mean in fuel efficiency, these cars may appeal to those valuing power and durability, especially with manufacturing years slightly below average.
  + This cluster includes vehicles with significantly higher mileage (60750.74), moderately higher fuel efficiency (57.71 MPG), slightly smaller engine sizes (1.96), lower prices (11852.06), and manufacturing years notably below the overall mean (2014.28). This cluster represents vehicles with significantly higher mileage, moderately improved fuel efficiency, and slightly smaller engine sizes. With notably lower prices and manufacturing years below the mean, these cars may appeal to budget-conscious consumers seeking reliable, efficient transportation.

# 9 Hierarchical Clustering from MCA

## 9.1 Hierarchical Clustering

* We will make 5 clusters to facilitate the process of further comparision and analysis:

res.hcpcMCA <- HCPC(res.mca,nb.clust = 5, order = TRUE)



## 9.2 Clustering Quality

* This clustering has a total gain of inertia of 47.25%, in case if we wanted to achieve at least 80% we will be needing 22 clusters, which makes the study more complicated.
* ((res.hcpcMCA$call$t$within[1]-res.hcpcMCA$call$t$within[5])/  
  res.hcpcMCA$call$t$within[1])\*100
* ## [1] 48.3903
* ((res.hcpcMCA$call$t$within[1]-res.hcpcMCA$call$t$within[22])/  
  res.hcpcMCA$call$t$within[1])\*100
* ## [1] 80.4319

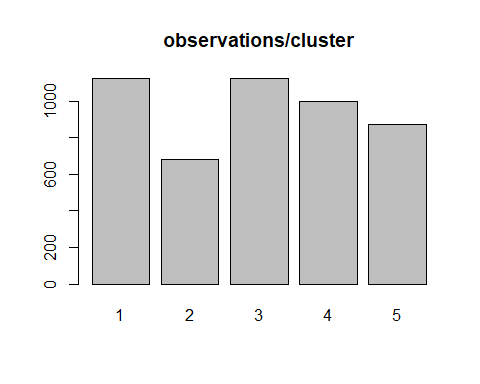
## 9.3 Clustering Description:

* We can see the following barplot describing the individuals distribution through different clusters

table(res.hcpcMCA$data.clust$clust)

##   
## 1 2 3 4 5   
## 1122 680 1122 996 873

barplot(table(res.hcpcMCA$data.clust$clust), main= "observations/cluster")



* The extremely low p-values (approaching zero) suggest a significant association between the categories of the variables, indicating that these variables contribute significantly to the formation of clusters.

res.hcpcMCA$desc.var$test.chi2

## p.value df  
## transmission 0.000000e+00 8  
## fuelType 0.000000e+00 12  
## manufacturer 0.000000e+00 12  
## f.price 0.000000e+00 12  
## f.mpg 0.000000e+00 12  
## f.engineSize 0.000000e+00 8  
## f.tax 3.102628e-154 8  
## Audi 3.064963e-89 4

**Describing each cluster in relation with different categories:**

1. Cluster 1 is characterized by a dominance of smaller engine sizes, with an overwhelming 80.31% of cars falling into the “**Small**” category. The fuel type for this cluster is predominantly **petrol**, constituting 54% of the vehicles. **Manual** transmission is strongly associated, with 47.52% of cars in this cluster featuring this type. The cluster is notably linked to the manufacturer **Volkswagen** 47.47%). Furthermore, high miles per gallon (MPG) is a defining characteristic, as 41.08% of the cars in this cluster fall into the “**High**” category. The pricing spectrum in this cluster is diverse, with a substantial portion (47.43%) classified as “**Very Cheap**.” The tax levels are also noteworthy, with a strong association (25.9%) with the “**Low**” tax band
2. Cluster 2 exhibits a preeminence of **medium-sized engine** cars, representing 60.6% of the vehicles in this category. **Diesel** fuel type is prevalent, constituting 49.1% of the cars. **BMW** stands out as the dominant manufacturer, with 56.41% of vehicles in this cluster associated with the brand. The cluster features a balanced distribution across various miles per gallon (MPG) categories, with a significant presence in both “**Moderate**” (50.2%) and “**Very High**” (42.27%) MPG ranges. **Hybrid** fuel type and **manual transmission** are notably associated with this cluster, each representing 81.25% and 37.54%, respectively. Additionally, there is a considerable concentration of cars with **high tax** (38.76%).
3. In Cluster 3, there is a strong association (69.23%) with cars having a “**Medium**” tax level. **Very high** miles per gallon (MPG) is a defining characteristic, representing 18.97% of the vehicles in this cluster. **Diesel** fuel type is predominant, accounting for 8.43% of the cars. The pricing landscape varies, with a significant presence of cars classified as “**Very Affordable**” (13.01%) and “**Cheap**” (11.73%) and “**Very Cheap**” (6.57%). **Manual transmission** is notable in this cluster, with 7.42% of cars featuring this type. The cluster is also marked by **medium-size engine** cars (6.51%).
4. Cluster 4 is characterized by a dominant association with low miles per gallon (MPG), with 79.69% of cars falling into the “**Low**” category. **Very expensive** cars constitute a substantial portion of this cluster (62.80%). The cluster is strongly associated with **large engine size** (48.06%) and **petrol** fuel type (36.59%). **High** taxes (40.88%) and **automatic transmission** (36.30%) are also notable features. **Audi** emerges as a prominent manufacturer in this cluster
5. Cluster 5 is distinguished by a strong association with **Mercedes** as the manufacturer (60.37%). **Diesel** fuel type is prevalent, constituting 29.88% of the cars in this cluster. **Large engine sizes** (41.90%) are also notable features. The cluster is characterized by a medium association with **low** taxes (22.33%) and **very high** miles per gallon (MPG) (30.25%). Additionally, **automatic** and **semi-automatic** transmissions are significant in this cluster, representing 26.02% and 25.88%, respectively. The pricing spectrum is diverse, with a notable presence of cars in various price ranges.

When it comes to numerical target variable price it a low effect in this clustering creation compared but as p-vale is 0 we can say that this variable has somehow an effect on the this clustering. Note that, we passed this variable as supplementary during MCA.

res.hcpcMCA$desc.var$quanti.var

## Eta2 P-value  
## price 0.3903306 0

We can aslo see how **price** behaves in each cluster:

res.hcpcMCA$desc.var$quanti

## $`1`  
## v.test Mean in category Overall mean sd in category Overall sd  
## price -23.5068 15447.81 21333.57 5560.034 9582.346  
## p.value  
## price 3.474894e-122  
##   
## $`2`  
## v.test Mean in category Overall mean sd in category Overall sd  
## price -21.91066 13874.35 21333.57 3938.32 9582.346  
## p.value  
## price 2.056008e-106  
##   
## $`3`  
## v.test Mean in category Overall mean sd in category Overall sd  
## price 2.398389 21934.1 21333.57 6886.665 9582.346  
## p.value  
## price 0.01646737  
##   
## $`4`  
## v.test Mean in category Overall mean sd in category Overall sd  
## price 35.63749 30965.45 21333.57 10400.53 9582.346  
## p.value  
## price 3.682114e-278  
##   
## $`5`  
## v.test Mean in category Overall mean sd in category Overall sd  
## price 5.502185 22947.5 21333.57 8436.298 9582.346  
## p.value  
## price 3.751137e-08

1. In Cluster 1, the “**price**” variable exhibits a significant negative v-test value of -20.95, suggesting a considerable difference in mean prices compared to the overall dataset. The mean price in this cluster is notably lower at 15,308.06, with a standard deviation of 5,663.19. This substantial deviation is highly significant (p-value = 1.75e-97), emphasizing a distinct pricing pattern in this cluster cheap and affordable.
2. Cluster 2 shows a significant negative v-test value of -10.80 for the “price” variable, indicating a noteworthy difference in mean prices. The mean price in this cluster is $19,079.93, with a standard deviation of $7,331.85. The low p-value (3.53e-27) underscores the significance of the observed price difference in this cluster.
3. For Cluster 3, the “price” variable demonstrates a significant negative v-test value of -10.48, highlighting a substantial difference in mean prices. The mean price in this cluster is $14,011.54, with a standard deviation of $3,161.38. The p-value (1.04e-25) reinforces the significance of the observed pricing distinction. This cluster has cheap and affordable pricing.
4. In Cluster 4, the “price” variable exhibits a substantial positive v-test value of 31.66, indicating a significant difference in mean prices. The mean price in this cluster is $31,238.13, with a standard deviation of $15,382.38. The extremely low p-value (5.27e-220) emphasizes the highly significant pricing difference in this cluster. This cluster has expensive/very expensive pricing.
5. Cluster 5 shows a positive v-test value of 7.77 for the “price” variable, indicating a significant difference in mean prices. The mean price in this cluster is $24,394.22, with a standard deviation of $9,709.73. The p-value (7.93e-15) underscores the significance of the observed price difference in this cluster. This cluster ha affordable pricing.

## 9.4 Paragons & Class-Specific individuals:

We can spot the most contributing individuals and extreme ones as following:

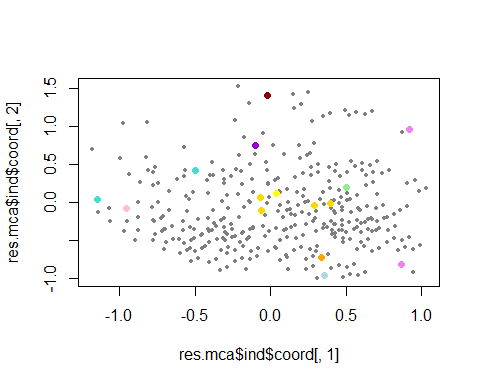
res.hcpcMCA$desc.ind$para

## Cluster: 1  
## 1458 1843 9045 744 8133   
## 0.2987625 0.2987625 0.2987625 0.2987625 0.2987625   
## ------------------------------------------------------------   
## Cluster: 2  
## 49602 48967 49047 40796 36656   
## 0.344365 0.344365 0.344365 0.344365 0.344365   
## ------------------------------------------------------------   
## Cluster: 3  
## 41190 34791 48501 40692 48482   
## 0.186631 0.186631 0.186631 0.186631 0.186631   
## ------------------------------------------------------------   
## Cluster: 4  
## 27151 26274 24132 23750 29095   
## 0.1884203 0.1884203 0.1884203 0.1884203 0.1884203   
## ------------------------------------------------------------   
## Cluster: 5  
## 26018 30618 30293 27490 34299   
## 0.4065199 0.4065199 0.4065199 0.4065199 0.4065199

res.hcpcMCA$desc.ind$dist

## Cluster: 1  
## 36869 44245 42363 46695 47033   
## 1.681039 1.648978 1.648978 1.648978 1.648978   
## ------------------------------------------------------------   
## Cluster: 2  
## 1810 2689 8031 1636 71   
## 1.981722 1.981722 1.981722 1.981722 1.981722   
## ------------------------------------------------------------   
## Cluster: 3  
## 21321 21018 46369 41055 33545   
## 5.014320 5.014320 4.807920 4.648160 4.645408   
## ------------------------------------------------------------   
## Cluster: 4  
## 10150 9145 10649 10417 9739   
## 1.872417 1.872417 1.872417 1.872417 1.872417   
## ------------------------------------------------------------   
## Cluster: 5  
## 23542 31797 33117 34326 33725   
## 1.719936 1.666393 1.666393 1.666393 1.666393

para1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[1]]))  
dist1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[1]]))  
para2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[2]]))  
dist2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[2]]))  
para3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[3]]))  
dist3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[3]]))  
para4<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[4]]))  
dist4<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[4]]))  
para5<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[5]]))  
dist5<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[5]]))  
plot(res.mca$ind$coord[,1],res.mca$ind$coord[,2],col="grey50",cex=0.5,pch=16)  
points(res.mca$ind$coord[para1,1],res.mca$ind$coord[para1,2],col="pink",cex=1,pch=16)  
points(res.mca$ind$coord[dist1,1],res.mca$ind$coord[dist1,2],col="turquoise",cex=1,pch=16)  
points(res.mca$ind$coord[para2,1],res.mca$ind$coord[para2,2],col="darkviolet",cex=1,pch=16)  
points(res.mca$ind$coord[dist2,1],res.mca$ind$coord[dist2,2],col="darkred",cex=1,pch=16)  
points(res.mca$ind$coord[para3,1],res.mca$ind$coord[para3,2],col="yellow",cex=1,pch=16)  
points(res.mca$ind$coord[dist3,1],res.mca$ind$coord[dist3,2],col="gold",cex=1,pch=16)  
points(res.mca$ind$coord[para4,1],res.mca$ind$coord[para4,2],col="orange",cex=1,pch=16)  
points(res.mca$ind$coord[dist4,1],res.mca$ind$coord[dist4,2],col="lightblue",cex=1,pch=16  
)  
points(res.mca$ind$coord[para5,1],res.mca$ind$coord[para5,2],col="lightgreen",cex=1,pch=16)  
points(res.mca$ind$coord[dist5,1],res.mca$ind$coord[dist5,2],col="violet",cex=1,pch=16)



## 9.5 Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on targets

### 9.5.1 General Comparision

* The accuracy of approximately 26.12% suggests a moderate level of alignment between the hierarchical clustering and the actual classes. This indicates that the clusters generated by the hierarchical clustering algorithm capture some of the underlying patterns present in the HC-MCA classes, but there is room for improvement.
* The low accuracy of approximately 18.10% indicates a poor alignment between the hierarchical clustering and these HC-MCA clusters. This raises questions about the effectiveness of the clustering algorithm in capturing the patterns inherent in the **claKM** variable. It’s essential to scrutinize the reasons behind this discrepancy. Potential issues could include the sensitivity of the hierarchical clustering algorithm to certain data patterns, the appropriateness of the clustering parameters chosen, etc.
* If we had a greater concordance, this would mean that they would be more similar.

df$hcpckMCA<-res.hcpcMCA$data.clust$clust  
# With Hierarchical Clustering (PCA)  
t1<-table(df$claH,df$hcpckMCA)  
t2<-table(df$claKM,df$hcpckMCA)  
t1

##   
## 1 2 3 4 5  
## 1 505 48 531 534 189  
## 2 0 0 58 338 85  
## 3 536 329 286 21 390  
## 4 0 0 50 64 17  
## 5 81 303 197 39 192

t2

##   
## 1 2 3 4 5  
## 1 92 325 143 28 168  
## 2 523 283 284 29 363  
## 3 1 0 56 439 62  
## 4 0 15 86 88 24  
## 5 506 57 553 412 256

100\*sum(diag(t1)/sum(t1))

## [1] 21.84436

100\*sum(diag(t2)/sum(t2))

## [1] 16.16941

### 9.5.2 Comparison based on quantitative target: Price

* The results unveil distinctive patterns in the relationship between the “price” variable and the clustering variable across three clustering methods. Hierarchical Clustering based on PCA exhibits the most substantial association, influencing price variation by 51%. In K-Means Clustering, the “price” variable shows a noteworthy 45% impact, indicating a robust relationship. On the other hand, MCA Hierarchical Clustering reveals a comparatively modest influence, with an Eta2 value of 0.28. These numerical insights shed light on the varying degrees of impact that different clustering methodologies exert on the variable of interest, providing a quantitative understanding of their implications.

### 9.5.3 Comparison based on binary target: Audi

* The variable “Audi” consistently shows significant associations within different cluster methods, and it plays a meaningful role in distinguishing clusters and sometimes note.
* The Audi variable exhibits a noteworthy association exclusively in one only cluster during Hierarchical Clustering based on PCA, evident from its considerably higher p-value in comparison to other categorical variables. This elevated p-value implies a diminished linkage and contribution to the formation of these clusters using this method. In contrast, during K-Means clustering, despite a higher p-value of 8.045414e-03, Audi’s impact is relatively modest, particularly in clusters 3 and 4. Surprisingly, this contribution is more substantial than the prior method, despite the lower p-value. Notably, MCA Hierarchical Clustering stands out with the lowest p-value of 1.429662e-67, underscoring the pivotal role played by Audi categories (Yes and No) in shaping clusters 2, 4, and 5, thereby contributing significantly to the underlying structure.

**Conclusion:**

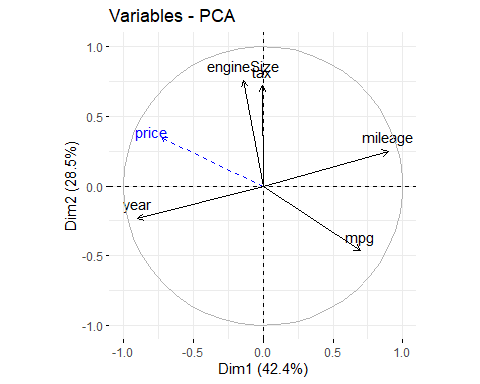
* In conclusion, the optimal clustering method is contingent on our research objectives, emphasizing the nuanced nature of this decision. Rather than asserting superiority of one method over another, the selection should align with the desired data interpretation, research goals, and study conditions. A judicious choice rooted in a comprehensive understanding of these factors ensures a rigorous application of clustering techniques, fostering a more insightful and robust data analysis.

# *10. Prediction model for numeric target “Price”:*

## *10.1 Initial model: price ~ engineSize + mpg*

* *When examining the interplay between price and other variables via Principal Components Analysis, a notable observation surfaces: the most pronounced (negative) correlation is evident between year and mpg (as illustrated in the following graph). Therefore, the most intuitive variable that come to mind for inclusion in our first initial model is mpg.*
* *It’s worth highlighting a robust correlation between Price and the other variables in our dataset. For instance, the correlation coefficients reveal noteworthy associations: 0.64 with Engine Size, 0.56 with Year, -0.52 with Mileage, and -0.58 with MPG. These correlations, coupled with p-values approaching zero, underscore the substantial impact of these variables on our target variable, Price. These insights affirm the significance of Engine Size, Year, Mileage, and MPG as influential factors shaping the main dynamics in our model. The ‘tax’ variable holds minimal significance in this context, which is why it doesn’t appear in the following output.*

# Graph of the variables  
fviz\_pca\_var(res.pca)



res.con <- condes(df,num.var=which(names(df)=="price"))  
res.con$quanti

## correlation p.value  
## year 0.60468636 0.0000000000  
## engineSize 0.58681573 0.0000000000  
## tax 0.05093213 0.0004195395  
## mileage -0.54914570 0.0000000000  
## mpg -0.62361708 0.0000000000

* *Let’s built the first model based on these conclusions:*
* Disclaimer: we won’t include the multivariant outliers that we spotted during the first/second deliverable.

#df<-df[-ll,]  
m0<-lm(price~engineSize+mpg,data=df)  
summary(m0)

##   
## Call:  
## lm(formula = price ~ engineSize + mpg, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26811.0 -3853.8 -415.6 3766.1 25155.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28544.793 677.245 42.15 <2e-16 \*\*\*  
## engineSize 7965.484 187.452 42.49 <2e-16 \*\*\*  
## mpg -422.513 8.836 -47.82 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6386 on 4790 degrees of freedom  
## Multiple R-squared: 0.5562, Adjusted R-squared: 0.556   
## F-statistic: 3002 on 2 and 4790 DF, p-value: < 2.2e-16

* *The model currently exhibits a moderate level of variability, as indicated by the R-squared value of 55.24%. While this suggests that approximately 55.24% of the variability in the dependent variable (price) is explained by the included independent variables (engineSize and mpg), there is room for improvement. Our aim is to enhance the model’s explanatory power, ultimately surpassing an ambitious target of 80% R-squared. Achieving this goal would signify a more robust and accurate representation of the factors influencing the price, thereby enhancing the model’s predictive capabilities.*

## *10.2 Adding more covariates: price ~ mileage + year + engineSize + mpg*

* *To enhance the model further, we will incorporate mileage and year, identified as correlated variables with price through Principal Component Analysis (PCA) deductions.*

m1<-lm(price~mileage+year+engineSize+mpg,data=df)  
summary(m1)

##   
## Call:  
## lm(formula = price ~ mileage + year + engineSize + mpg, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15283.8 -2644.7 -73.8 2386.1 18822.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.881e+06 1.236e+05 -31.39 <2e-16 \*\*\*  
## mileage -9.695e-02 5.896e-03 -16.44 <2e-16 \*\*\*  
## year 1.931e+03 6.121e+01 31.55 <2e-16 \*\*\*  
## engineSize 9.842e+03 1.352e+02 72.79 <2e-16 \*\*\*  
## mpg -1.950e+02 7.040e+00 -27.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4493 on 4788 degrees of freedom  
## Multiple R-squared: 0.7804, Adjusted R-squared: 0.7802   
## F-statistic: 4253 on 4 and 4788 DF, p-value: < 2.2e-16

*The new model, denoted as m1, represents a significant improvement over the initial model (m0). Here are the key findings from the summary of m1:*

* *The residuals exhibit a narrower range compared to m0, with a minimum of -26504 and a maximum of 48651.*
* *The residual standard error has decreased to 5321, indicating a reduction in the variability of the residuals compared to m0.*
* *The R-squared value has significantly increased to 78.33%, and the adjusted R-squared is also high (78.31%). This implies that the new model explains approximately 78.33% of the variability in the dependent variable.*

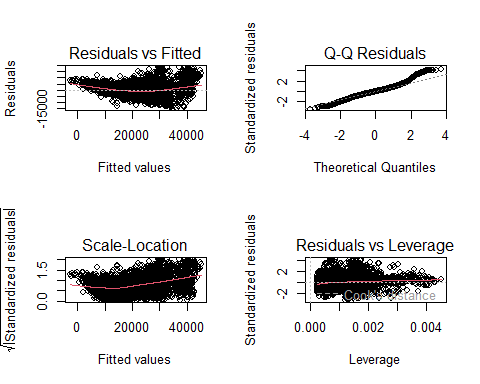
*m1 demonstrates substantial improvement over m0, with a higher R-squared, lower residual standard error, and significant coefficients. This model appears to be a more powerful predictor of price, explaining a substantial portion of the variability.*

vif(m1)

## mileage year engineSize mpg   
## 2.999840 2.944574 1.170204 1.427723

* *The provided values, all falling below 5, suggest that the correlation impact in this regression is relatively modest and not particularly influential. These findings suggest that while some correlation exists among certain predictors, the multicollinearity in the model is generally well-controlled, with VIF values falling within acceptable ranges.*

par(mfrow=c(2,2));  
plot(m1,id.n=0)



* *Even though we achieved a high Multiple R-squared value in this model (78.33% of the variability explained), the residual plots still reveal some imperfections that we need to address and improve upon.*
* *As we observe, none of the previous graphs show some heteroscedasticity as points are not equally scattered.*
  + *“Residuals vs Fitted” graph detects residuals that experience heteroscedasticity.*
  + *“Q-Q Residuals” graph shows us the normality of residuals, as the residuals tend to dodge the normal line this indicates non-normality of these residuals.*
  + *“Scale-Location” graph reassures the heteroscedasticity of residuals as how they are spread.*
  + *“Residuals vs Leverage Fitted” graph detects influential individuals with residuals that might strongly affect the regression model. Some outliers impact significantly impact the normal distribution.*

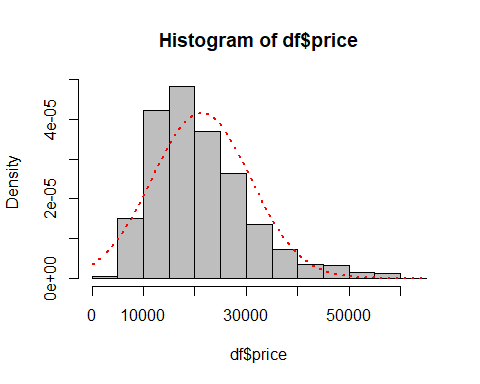
AIC(m0,m1)

## df AIC  
## m0 4 97597.57  
## m1 6 94230.15

***Is our data normal?***

* *Assessing the normality of our target variable would help us improve these graphs or eve our Multiple R-squared value.*

hist(df$price,freq=F,col="grey")  
mm<-mean(df$price);ss<-sd(df$price)  
curve(dnorm(x,mean=mm,sd=ss),col="red",lwd=2,lty=3, add=T)



* *The Shapiro Test performs normality test on the variable “price”. The result is an extremely low p-value (< 2.2e-16). The small p-value indicates strong evidence against the null hypothesis, suggesting that the data does not follow a normal distribution.*

shapiro.test(df$price)

##   
## Shapiro-Wilk normality test  
##   
## data: df$price  
## W = 0.93442, p-value < 2.2e-16

*Other tests:*

* *The following Skewness test result shows a non-null value which indicates non-normality.*

library(e1071)  
skewness(df$price)

## [1] 1.077892

* *The following Kurtosis test result shows a non-null value which indicates non-normality.*

kurtosis(df$price)

## [1] 1.403526

## *10.3 Box-Cox transformation of price:*

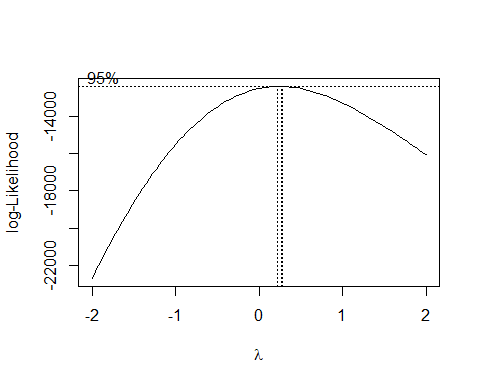
* *In an effort to enhance our linear regression model, we applied a Box-Cox transformation to the target variable. This transformation, guided by the optimal lambda from the boxcox function, seeks to address issues related to variance and normality. By optimizing the target variable’s distribution, we aim to improve the overall performance and reliability of our regression analysis.*
* *Given the considerable deviation of the lambda interval from zero, we’ll refrain from employing a direct logarithmic transformation and, instead, pursue the following strategy:*

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

boxcox\_results<-boxcox(m1,data=df)



# Extract the optimal lambda  
optimal\_lambda <- boxcox\_results$x[which.max(boxcox\_results$y)]  
  
# Target variable transformed  
df$price\_transformed <- (df$price^ optimal\_lambda - 1) / optimal\_lambda

* *Let’s build a second model with price\_transformed:*

m2<-lm(price\_transformed~mileage+year+engineSize+mpg,data=df)  
summary(m2)

##   
## Call:  
## lm(formula = price\_transformed ~ mileage + year + engineSize +   
## mpg, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.988 -1.606 0.072 1.713 8.964   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.701e+03 7.089e+01 -38.10 <2e-16 \*\*\*  
## mileage -6.668e-05 3.381e-06 -19.72 <2e-16 \*\*\*  
## year 1.360e+00 3.509e-02 38.75 <2e-16 \*\*\*  
## engineSize 6.109e+00 7.752e-02 78.80 <2e-16 \*\*\*  
## mpg -1.021e-01 4.036e-03 -25.29 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.576 on 4788 degrees of freedom  
## Multiple R-squared: 0.8145, Adjusted R-squared: 0.8144   
## F-statistic: 5256 on 4 and 4788 DF, p-value: < 2.2e-16

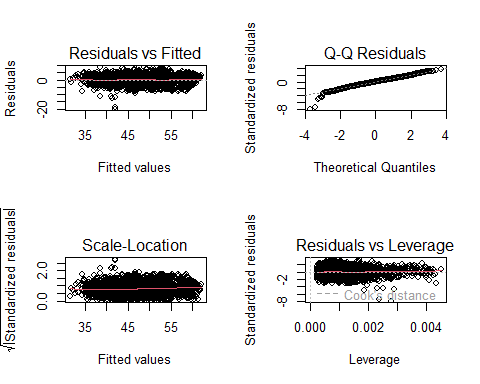
* *The R-squared went from 78% to 83%, it is a small but significant improvement when it comes to only make one only transformation. This indicates a better fit than the first model. This new model explain more variability in the relationship between the predictors and the target variable.*
* *To check, we calculate Variance Inflation Factors for this mode and we can see that the values are constant, no significant change in these values.*

vif(m2)

## mileage year engineSize mpg   
## 2.999840 2.944574 1.170204 1.427723

* *Let’s analyse residual plots:*

par(mfrow=c(2,2));  
plot(m2,id.n=0);



* *As we can see, even though the Multiple R-squared: didn’t increase a lot, we still could sense a lot of improvement through these graphs.*
* *The current plots illustrate a normal distribution of residuals, affirming the selection of this model as a preferred one. Notably, homoscedasticity is now more evident, and also indicating improved normality. However, it’s worth noting that the residuals vs. leverage plot hasn’t shown a better enhancement.*

AIC(m1,m2)

## df AIC  
## m1 6 94230.15  
## m2 6 22680.55

## *10.4 Covariates transformations with BoxTidwell:*

* *Utilizing the boxTidwell function could provide valuable insights into potential transformations that may enhance our model performance across various lambda values.*
* *Note: The variables year and mileage exhibit a higher degree of correlation compared to other variables. Therefore, it is not feasible to incorporate both of them simultaneously during the computation of the boxTidwell function.*

library("carData")  
boxTidwell(price\_transformed~mileage+engineSize+mpg,data=df)

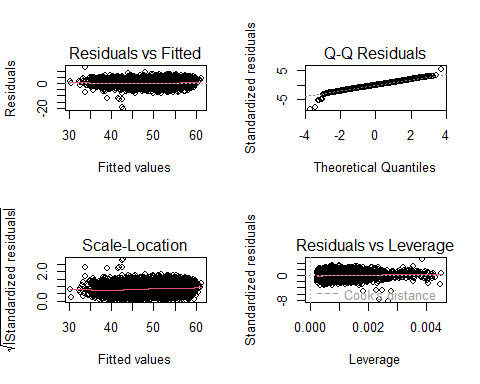
## MLE of lambda Score Statistic (t) Pr(>|t|)   
## mileage 0.698192 13.1840 < 2.2e-16 \*\*\*  
## engineSize -0.060201 -15.7665 < 2.2e-16 \*\*\*  
## mpg -0.874433 9.7996 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 5   
##   
## Score test for null hypothesis that all lambdas = 1:  
## F = 188.45, df = 3 and 4786, Pr(>F) = < 2.2e-16

* As engineSize’s lambda is close to zero than one, we will apply a logarithmic transformation.

library("carData")  
m3<-lm(price\_transformed~mileage+year+log(engineSize)+mpg,data=df)  
summary(m3)

##   
## Call:  
## lm(formula = price\_transformed ~ mileage + year + log(engineSize) +   
## mpg, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.6393 -1.6362 -0.0074 1.6230 13.3076   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.621e+03 6.908e+01 -37.94 <2e-16 \*\*\*  
## mileage -7.045e-05 3.300e-06 -21.35 <2e-16 \*\*\*  
## year 1.323e+00 3.420e-02 38.68 <2e-16 \*\*\*  
## log(engineSize) 1.136e+01 1.379e-01 82.42 <2e-16 \*\*\*  
## mpg -1.137e-01 3.874e-03 -29.34 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.511 on 4788 degrees of freedom  
## Multiple R-squared: 0.8238, Adjusted R-squared: 0.8237   
## F-statistic: 5598 on 4 and 4788 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(m3,id.n=0)



* *While the Multiple R-squared values do not exhibit significant improvement compared to other models, the notable distinction lies in the graphical representation. The plots vividly reveal better homoscedasticity, pinpoint individuals with less leverage, and visible normality.*

***Model Validation:***

* We will employ the Breusch-Pagan test to evaluate homoscedasticity. Since the p-value associated with the test is significantly low, we can reject the null hypothesis of heteroscedasticity. Consequently, this leads us to conclude that the model exhibits **homoscedasticity**, suggesting that the variance of the residuals is consistent across all levels of the explanatory variables.

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(m3)

##   
## studentized Breusch-Pagan test  
##   
## data: m3  
## BP = 83.425, df = 4, p-value < 2.2e-16

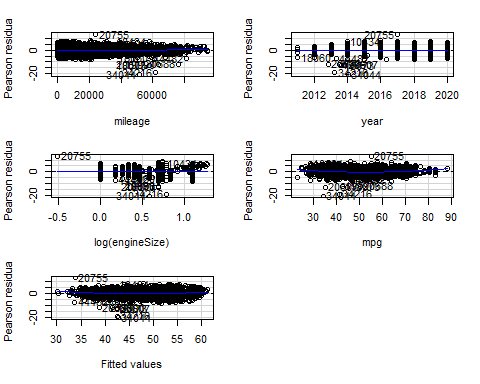
* *The VIF values are within a relatively moderate range, with mileage having a VIF of 2.9, year with 2.79, log-engineSize with 1.16, and mpg with 1.32. While the VIF for mileage and year is slightly above 2, suggesting some correlation,* ***none of the variables exhibit severe multicollinearity*** *(VIF above 5).*

vif(m3)

## mileage year log(engineSize) mpg   
## 3.009602 2.944673 1.139159 1.384596

* *In the below residual plots for model m3 using Cook’s distance, “mileage” and “year” exhibit non-significant test statistics, suggesting well-behaved residuals. However, “engineSize” and “mpg” display highly significant test statistics, indicating deviations from the model assumptions.*
* *The flat line in the scatter plot for mileage and year indicates a steady linear trend, meeting the assumption of consistent variability. But, when looking at engine size and mpg, the pattern is less clear, suggesting possible deviations from these assumptions in that case.*
* *The graphical representations affirm the* ***independence*** *of residuals within this model. The plots indicate that there is no discernible pattern or structure in the residuals.*

residualPlots(m3,id=list(method=cooks.distance(m3),n=10))



## Test stat Pr(>|Test stat|)   
## mileage 2.2942 0.02182 \*   
## year -0.1817 0.85584   
## log(engineSize) 1.0971 0.27265   
## mpg 9.5156 < 2.2e-16 \*\*\*  
## Tukey test 7.4329 1.063e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Let’s proceed to display the box plots of the R-student values, Hat values, and Cook’s distances for the observations in the model.

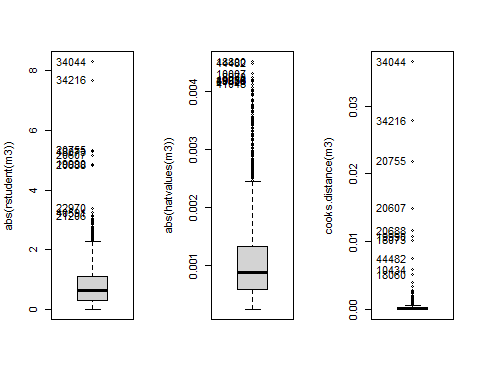
par(mfrow=c(1,3))  
Boxplot(abs(rstudent(m3)),id=list(labels=row.names(df)))

## [1] "34044" "34216" "20755" "18073" "20607" "19990" "20688" "22970" "40591"  
## [10] "21206"

Boxplot(abs(hatvalues(m3)),id=list(labels=row.names(df)))

## [1] "18800" "44482" "10007" "8970" "44258" "19931" "49306" "19676" "39539"  
## [10] "41048"

Boxplot(cooks.distance(m3),id=list(labels=row.names(df)))

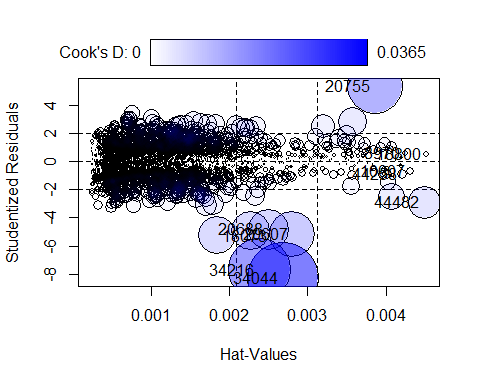


## [1] "34044" "34216" "20755" "20607" "20688" "19990" "18073" "44482" "10434"  
## [10] "18060"

stu <- which(abs(rstudent(m3))>3.0)  
cook <- which(abs(cooks.distance(m3))>0.01)  
hat <- which(abs(hatvalues(m3))>0.004)  
outs<-unique(stu,cook,hat)

* Spotting influential individuals:

x<-influencePlot( m3, id=c(list="noteworthy",n=5))



obs<-rownames(x)  
outs<-unique(outs,obs)  
  
df\_outs<-df[-outs,]

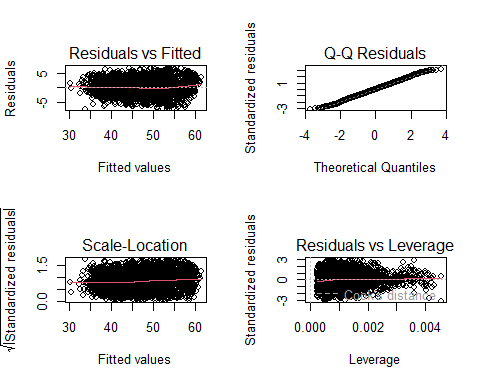
* Building a model without unusual and influential data:

m4<- update(m3, data=df\_outs)  
summary(m4)

##   
## Call:  
## lm(formula = price\_transformed ~ mileage + year + log(engineSize) +   
## mpg, data = df\_outs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.4810 -1.6545 -0.0251 1.5912 7.4319   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.586e+03 6.669e+01 -38.78 <2e-16 \*\*\*  
## mileage -6.822e-05 3.185e-06 -21.42 <2e-16 \*\*\*  
## year 1.306e+00 3.302e-02 39.55 <2e-16 \*\*\*  
## log(engineSize) 1.136e+01 1.332e-01 85.30 <2e-16 \*\*\*  
## mpg -1.183e-01 3.752e-03 -31.54 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.42 on 4771 degrees of freedom  
## Multiple R-squared: 0.8335, Adjusted R-squared: 0.8333   
## F-statistic: 5970 on 4 and 4771 DF, p-value: < 2.2e-16

* The model demonstrates a high degree of explanatory power, about 84% of the variance in the dependent variable. This robust model performance, along with the significant coefficients and the reasonable distribution of residuals, underscores the importance of removing outliers and influential points in data analysis.
* These diagnostic plots indicate that the model **m4** is performing well. The assumptions of linearity, homoscedasticity, and normality of the residuals appear to have been met, and there are no obvious influential outliers. This suggests that the model provides a good fit to the data.

par(mfrow=c(2,2))  
plot(m4,id.n=0)



* Let’s review and compare models m3 and m4 to decide which model is better suited for our analysis and then choose the one to move forward with.

AIC(m3,m4)

## df AIC  
## m3 6 22433.16  
## m4 6 22002.09

* Although model m4 has a lower AIC and might be the better model, we’ll proceed with m3. This way, we keep the outliers in the mix, which will help us spot and address them as we refine our model.

## *10.5 Incorporating Interaction Terms in the Linear Regression Model*

### *10.5.1 Adding qualitative variables as predictors*

* *The following output highlights the categorical variables most correlated with the price. Notably, model emerges as the most influential categorical factor, followed by transmission. It’s essential to note that some factors are derived from previously utilized covariates, so we won’t take them into consideration.*

condes(df,3)$quali

## R2 p.value  
## model 0.501972378 0.000000e+00  
## f.year 0.380402821 0.000000e+00  
## f.price 0.814879203 0.000000e+00  
## f.miles 0.322984657 0.000000e+00  
## f.mpg 0.370479330 0.000000e+00  
## claKM 0.560904586 0.000000e+00  
## claH 0.528503025 0.000000e+00  
## hcpckMCA 0.390330648 0.000000e+00  
## transmission 0.266798626 1.482197e-323  
## f.engineSize 0.191033566 3.115676e-221  
## manufacturer 0.096096343 1.475048e-104  
## f.tax 0.066808573 1.202052e-72  
## fuelType 0.010997182 1.867192e-11  
## Audi 0.003969156 1.271992e-05

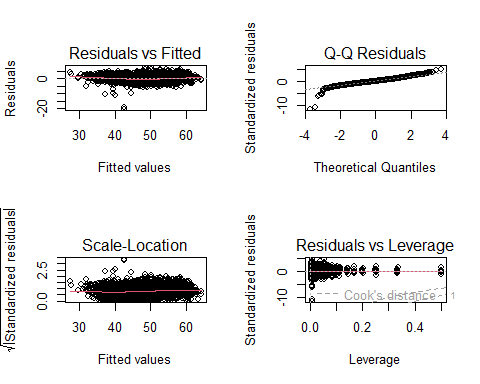
* *Let’s create a new model and analyze the results:*

m5<-lm(price\_transformed~mileage+year+engineSize+mpg + model + transmission,data=df)  
summary(m5)

##   
## Call:  
## lm(formula = price\_transformed ~ mileage + year + engineSize +   
## mpg + model + transmission, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3785 -1.0653 -0.0495 1.0450 8.0854   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.378e+03 5.130e+01 -46.345 < 2e-16 \*\*\*  
## mileage -7.240e-05 2.409e-06 -30.057 < 2e-16 \*\*\*  
## year 1.201e+00 2.540e-02 47.285 < 2e-16 \*\*\*  
## engineSize 2.953e+00 8.647e-02 34.148 < 2e-16 \*\*\*  
## mpg -7.415e-02 3.194e-03 -23.219 < 2e-16 \*\*\*  
## modelAudi- A3 1.257e+00 2.031e-01 6.188 6.63e-10 \*\*\*  
## modelAudi- A4 1.457e+00 2.240e-01 6.501 8.78e-11 \*\*\*  
## modelAudi- A5 2.230e+00 2.792e-01 7.987 1.72e-15 \*\*\*  
## modelAudi- A6 2.835e+00 2.657e-01 10.671 < 2e-16 \*\*\*  
## modelAudi- A7 2.660e+00 5.989e-01 4.442 9.11e-06 \*\*\*  
## modelAudi- A8 4.778e+00 5.557e-01 8.598 < 2e-16 \*\*\*  
## modelAudi- Q2 1.610e+00 2.609e-01 6.170 7.38e-10 \*\*\*  
## modelAudi- Q3 2.614e+00 2.189e-01 11.940 < 2e-16 \*\*\*  
## modelAudi- Q5 4.317e+00 2.480e-01 17.405 < 2e-16 \*\*\*  
## modelAudi- Q7 6.803e+00 3.845e-01 17.692 < 2e-16 \*\*\*  
## modelAudi- Q8 8.785e+00 1.053e+00 8.345 < 2e-16 \*\*\*  
## modelAudi- RS3 5.670e+00 1.277e+00 4.441 9.15e-06 \*\*\*  
## modelAudi- RS5 7.843e+00 1.796e+00 4.367 1.29e-05 \*\*\*  
## modelAudi- RS6 1.044e+01 9.227e-01 11.312 < 2e-16 \*\*\*  
## modelAudi- S3 4.171e+00 1.045e+00 3.991 6.68e-05 \*\*\*  
## modelAudi- S4 4.959e+00 1.797e+00 2.759 0.005813 \*\*   
## modelAudi- S8 6.569e+00 1.796e+00 3.658 0.000257 \*\*\*  
## modelAudi- SQ5 6.012e+00 1.280e+00 4.699 2.69e-06 \*\*\*  
## modelAudi- TT 1.932e+00 3.594e-01 5.376 7.99e-08 \*\*\*  
## modelBMW- 1 Series -5.342e-01 2.011e-01 -2.656 0.007940 \*\*   
## modelBMW- 2 Series -2.736e-01 2.341e-01 -1.169 0.242664   
## modelBMW- 3 Series 6.717e-01 2.071e-01 3.243 0.001190 \*\*   
## modelBMW- 4 Series 9.050e-01 2.630e-01 3.441 0.000585 \*\*\*  
## modelBMW- 5 Series 1.989e+00 2.501e-01 7.951 2.30e-15 \*\*\*  
## modelBMW- 6 Series 1.852e+00 4.757e-01 3.893 0.000100 \*\*\*  
## modelBMW- 7 Series 6.013e+00 5.543e-01 10.848 < 2e-16 \*\*\*  
## modelBMW- 8 Series 9.897e+00 1.796e+00 5.510 3.79e-08 \*\*\*  
## modelBMW- i3 3.165e+00 1.044e+00 3.033 0.002438 \*\*   
## modelBMW- M2 4.945e+00 1.279e+00 3.865 0.000113 \*\*\*  
## modelBMW- M3 5.736e+00 1.051e+00 5.457 5.10e-08 \*\*\*  
## modelBMW- M4 4.958e+00 4.505e-01 11.005 < 2e-16 \*\*\*  
## modelBMW- M6 7.166e+00 1.279e+00 5.601 2.25e-08 \*\*\*  
## modelBMW- X1 1.117e+00 2.774e-01 4.026 5.76e-05 \*\*\*  
## modelBMW- X2 1.643e+00 4.017e-01 4.090 4.39e-05 \*\*\*  
## modelBMW- X3 3.634e+00 3.161e-01 11.497 < 2e-16 \*\*\*  
## modelBMW- X4 3.918e+00 4.262e-01 9.192 < 2e-16 \*\*\*  
## modelBMW- X5 6.463e+00 3.652e-01 17.698 < 2e-16 \*\*\*  
## modelBMW- X6 7.443e+00 8.253e-01 9.019 < 2e-16 \*\*\*  
## modelBMW- Z3 -1.016e+01 1.794e+00 -5.660 1.60e-08 \*\*\*  
## modelBMW- Z4 1.878e+00 6.957e-01 2.699 0.006985 \*\*   
## modelMercedes- A Class 1.134e+00 1.950e-01 5.816 6.44e-09 \*\*\*  
## modelMercedes- B Class 1.815e-01 3.059e-01 0.593 0.553142   
## modelMercedes- C Class 2.213e+00 1.901e-01 11.642 < 2e-16 \*\*\*  
## modelMercedes- CL Class 2.309e+00 2.978e-01 7.754 1.08e-14 \*\*\*  
## modelMercedes- CLA Class 2.744e+00 9.075e-01 3.024 0.002508 \*\*   
## modelMercedes- CLS Class 3.542e+00 4.296e-01 8.245 < 2e-16 \*\*\*  
## modelMercedes- E Class 2.977e+00 2.222e-01 13.398 < 2e-16 \*\*\*  
## modelMercedes- GL Class 3.393e+00 4.921e-01 6.896 6.07e-12 \*\*\*  
## modelMercedes- GLA Class 1.317e+00 2.581e-01 5.105 3.44e-07 \*\*\*  
## modelMercedes- GLB Class 4.189e+00 1.792e+00 2.337 0.019482 \*   
## modelMercedes- GLC Class 4.764e+00 2.400e-01 19.852 < 2e-16 \*\*\*  
## modelMercedes- GLE Class 6.466e+00 3.112e-01 20.778 < 2e-16 \*\*\*  
## modelMercedes- GLS Class 7.520e+00 5.561e-01 13.524 < 2e-16 \*\*\*  
## modelMercedes- M Class 4.708e+00 8.185e-01 5.752 9.38e-09 \*\*\*  
## modelMercedes- S Class 6.791e+00 4.895e-01 13.874 < 2e-16 \*\*\*  
## modelMercedes- SL CLASS 3.399e+00 3.874e-01 8.772 < 2e-16 \*\*\*  
## modelMercedes- SLK 2.460e-01 5.903e-01 0.417 0.676899   
## modelMercedes- V Class 3.452e+00 4.434e-01 7.785 8.50e-15 \*\*\*  
## modelMercedes- X-CLASS 3.543e-01 4.965e-01 0.714 0.475437   
## modelVW- Amarok 1.564e+00 6.268e-01 2.495 0.012636 \*   
## modelVW- Arteon 1.471e+00 4.150e-01 3.544 0.000399 \*\*\*  
## modelVW- Beetle -1.078e+00 6.159e-01 -1.750 0.080127 .   
## modelVW- Caddy -1.235e+00 1.794e+00 -0.688 0.491278   
## modelVW- Caddy Maxi -1.558e+00 1.792e+00 -0.869 0.384726   
## modelVW- Caddy Maxi Life -2.259e+00 9.074e-01 -2.490 0.012818 \*   
## modelVW- Caravelle 7.023e+00 5.439e-01 12.913 < 2e-16 \*\*\*  
## modelVW- CC -1.387e+00 6.174e-01 -2.247 0.024708 \*   
## modelVW- Golf -2.983e-01 1.752e-01 -1.703 0.088686 .   
## modelVW- Golf SV -1.815e+00 4.189e-01 -4.333 1.50e-05 \*\*\*  
## modelVW- Passat -5.927e-01 2.428e-01 -2.441 0.014678 \*   
## modelVW- Polo -2.730e+00 1.835e-01 -14.880 < 2e-16 \*\*\*  
## modelVW- Scirocco -4.999e-01 3.918e-01 -1.276 0.201984   
## modelVW- Sharan 8.298e-01 4.067e-01 2.040 0.041370 \*   
## modelVW- Shuttle 1.126e+00 5.893e-01 1.911 0.056094 .   
## modelVW- T-Cross -2.831e-01 3.734e-01 -0.758 0.448464   
## modelVW- T-Roc 8.671e-01 2.673e-01 3.244 0.001187 \*\*   
## modelVW- Tiguan 1.379e+00 2.068e-01 6.666 2.94e-11 \*\*\*  
## modelVW- Tiguan Allspace 2.113e+00 7.475e-01 2.827 0.004725 \*\*   
## modelVW- Touareg 2.878e+00 3.509e-01 8.201 3.04e-16 \*\*\*  
## modelVW- Touran 1.295e+00 4.032e-01 3.212 0.001326 \*\*   
## modelVW- Up -5.929e+00 2.478e-01 -23.925 < 2e-16 \*\*\*  
## transmissionf.Trans-SemiAuto 1.354e+00 7.642e-02 17.720 < 2e-16 \*\*\*  
## transmissionf.Trans-Automatic 1.130e+00 8.190e-02 13.803 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.784 on 4705 degrees of freedom  
## Multiple R-squared: 0.9126, Adjusted R-squared: 0.911   
## F-statistic: 564.6 on 87 and 4705 DF, p-value: < 2.2e-16

* *Based on the adjusted R-squared and other statistical measures, it appears that adding the categorical variables ‘model’ and ‘transmission’ significantly improved the model’s explanatory power and overall fit to the data. We achieved 91% of variability,*
* *Based on the following plots, the non-horizontal red line in the scale-location graph suggests potential heteroscedasticity, challenging the assumption of constant variance. While the model retains excellent variability, the residuals’ departure from normal distribution in extreme quantiles indicates potential limitations in capturing certain patterns. Additionally, there are influential extreme values with high leverage that could impact the regression and may require removal for model improvement.*

par(mfrow=c(2,2));  
plot(m5,id.n=0);



* *For this model, mileage, year, engineSize, transmission and mpg show low VIF values (below 5), suggesting minimal multicollinearity. The “model” variable exhibits a relatively higher VIF of 5.47, possibly due to the categorical nature of car models.*

vif(m5)

## GVIF Df GVIF^(1/(2\*Df))  
## mileage 3.176135 1 1.782171  
## year 3.216757 1 1.793532  
## engineSize 3.035761 1 1.742344  
## mpg 1.863838 1 1.365225  
## model 5.409976 81 1.010476  
## transmission 1.889205 2 1.172384

* *The ANOVA table for model m5 shows highly significant p-values for all predictor variables, indicating their strong influence on the transformed price. The model is statistically significant overall, and the residuals have a low mean square value, suggesting a well-fitted model.*

anova(m5)

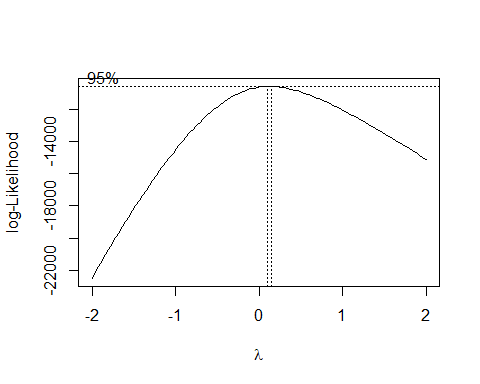
## Analysis of Variance Table  
##   
## Response: price\_transformed  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mileage 1 60133 60133 18893.502 < 2.2e-16 \*\*\*  
## year 1 15016 15016 4717.827 < 2.2e-16 \*\*\*  
## engineSize 1 60152 60152 18899.319 < 2.2e-16 \*\*\*  
## mpg 1 4245 4245 1333.847 < 2.2e-16 \*\*\*  
## model 81 15769 195 61.165 < 2.2e-16 \*\*\*  
## transmission 2 1035 518 162.608 < 2.2e-16 \*\*\*  
## Residuals 4705 14975 3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### *10.5.2 Logarithmic transformation of “price”:*

*Can a target transformation make it better?*

* *Let’s apply box-cox to check for possible transformations:*

boxcox(price~mileage+year+engineSize+mpg +model + transmission,data=df)



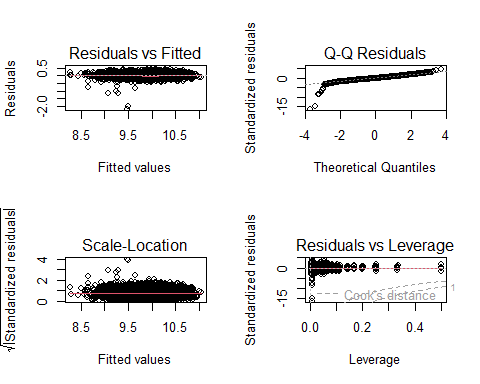
* *Given the proximity of lambda (λ) to zero, it suggests that applying a logarithmic transformation to the target variable would enhance its relationship with the predictor variables.*

m6<-lm(log(price)~mileage+year+engineSize+mpg +model + transmission,data=df)  
summary(m6)

##   
## Call:  
## lm(formula = log(price) ~ mileage + year + engineSize + mpg +   
## model + transmission, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.16369 -0.07463 -0.00202 0.07783 0.58074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.751e+02 3.839e+00 -45.618 < 2e-16 \*\*\*  
## mileage -5.618e-06 1.802e-07 -31.170 < 2e-16 \*\*\*  
## year 9.161e-02 1.901e-03 48.200 < 2e-16 \*\*\*  
## engineSize 2.095e-01 6.470e-03 32.380 < 2e-16 \*\*\*  
## mpg -4.844e-03 2.390e-04 -20.268 < 2e-16 \*\*\*  
## modelAudi- A3 1.059e-01 1.520e-02 6.967 3.69e-12 \*\*\*  
## modelAudi- A4 1.237e-01 1.676e-02 7.380 1.86e-13 \*\*\*  
## modelAudi- A5 1.840e-01 2.089e-02 8.807 < 2e-16 \*\*\*  
## modelAudi- A6 2.273e-01 1.988e-02 11.432 < 2e-16 \*\*\*  
## modelAudi- A7 2.020e-01 4.481e-02 4.507 6.74e-06 \*\*\*  
## modelAudi- A8 3.245e-01 4.158e-02 7.804 7.35e-15 \*\*\*  
## modelAudi- Q2 1.420e-01 1.953e-02 7.274 4.07e-13 \*\*\*  
## modelAudi- Q3 2.149e-01 1.638e-02 13.118 < 2e-16 \*\*\*  
## modelAudi- Q5 3.212e-01 1.856e-02 17.307 < 2e-16 \*\*\*  
## modelAudi- Q7 4.590e-01 2.877e-02 15.953 < 2e-16 \*\*\*  
## modelAudi- Q8 5.497e-01 7.877e-02 6.979 3.39e-12 \*\*\*  
## modelAudi- RS3 4.190e-01 9.553e-02 4.386 1.18e-05 \*\*\*  
## modelAudi- RS5 5.183e-01 1.344e-01 3.857 0.000116 \*\*\*  
## modelAudi- RS6 7.285e-01 6.904e-02 10.551 < 2e-16 \*\*\*  
## modelAudi- S3 3.508e-01 7.820e-02 4.486 7.43e-06 \*\*\*  
## modelAudi- S4 3.239e-01 1.345e-01 2.409 0.016030 \*   
## modelAudi- S8 5.153e-01 1.344e-01 3.834 0.000127 \*\*\*  
## modelAudi- SQ5 4.385e-01 9.574e-02 4.580 4.76e-06 \*\*\*  
## modelAudi- TT 1.684e-01 2.689e-02 6.261 4.18e-10 \*\*\*  
## modelBMW- 1 Series -3.619e-02 1.505e-02 -2.404 0.016235 \*   
## modelBMW- 2 Series -4.415e-03 1.752e-02 -0.252 0.801026   
## modelBMW- 3 Series 6.027e-02 1.550e-02 3.889 0.000102 \*\*\*  
## modelBMW- 4 Series 8.628e-02 1.968e-02 4.384 1.19e-05 \*\*\*  
## modelBMW- 5 Series 1.569e-01 1.871e-02 8.382 < 2e-16 \*\*\*  
## modelBMW- 6 Series 1.601e-01 3.560e-02 4.498 7.01e-06 \*\*\*  
## modelBMW- 7 Series 3.999e-01 4.148e-02 9.641 < 2e-16 \*\*\*  
## modelBMW- 8 Series 6.148e-01 1.344e-01 4.573 4.92e-06 \*\*\*  
## modelBMW- i3 2.566e-01 7.810e-02 3.286 0.001025 \*\*   
## modelBMW- M2 3.138e-01 9.574e-02 3.278 0.001053 \*\*   
## modelBMW- M3 4.298e-01 7.866e-02 5.464 4.90e-08 \*\*\*  
## modelBMW- M4 3.384e-01 3.371e-02 10.039 < 2e-16 \*\*\*  
## modelBMW- M6 5.637e-01 9.573e-02 5.888 4.18e-09 \*\*\*  
## modelBMW- X1 1.004e-01 2.076e-02 4.838 1.36e-06 \*\*\*  
## modelBMW- X2 1.270e-01 3.006e-02 4.224 2.44e-05 \*\*\*  
## modelBMW- X3 2.718e-01 2.365e-02 11.492 < 2e-16 \*\*\*  
## modelBMW- X4 2.798e-01 3.189e-02 8.773 < 2e-16 \*\*\*  
## modelBMW- X5 4.512e-01 2.733e-02 16.512 < 2e-16 \*\*\*  
## modelBMW- X6 4.658e-01 6.175e-02 7.542 5.51e-14 \*\*\*  
## modelBMW- Z3 -9.864e-01 1.343e-01 -7.347 2.37e-13 \*\*\*  
## modelBMW- Z4 1.354e-01 5.206e-02 2.601 0.009332 \*\*   
## modelMercedes- A Class 9.437e-02 1.459e-02 6.469 1.08e-10 \*\*\*  
## modelMercedes- B Class 1.475e-02 2.289e-02 0.644 0.519319   
## modelMercedes- C Class 1.690e-01 1.423e-02 11.879 < 2e-16 \*\*\*  
## modelMercedes- CL Class 1.940e-01 2.229e-02 8.706 < 2e-16 \*\*\*  
## modelMercedes- CLA Class 2.209e-01 6.791e-02 3.253 0.001149 \*\*   
## modelMercedes- CLS Class 2.651e-01 3.214e-02 8.248 < 2e-16 \*\*\*  
## modelMercedes- E Class 2.239e-01 1.663e-02 13.467 < 2e-16 \*\*\*  
## modelMercedes- GL Class 2.803e-01 3.682e-02 7.612 3.25e-14 \*\*\*  
## modelMercedes- GLA Class 1.209e-01 1.931e-02 6.261 4.18e-10 \*\*\*  
## modelMercedes- GLB Class 2.792e-01 1.341e-01 2.082 0.037406 \*   
## modelMercedes- GLC Class 3.485e-01 1.795e-02 19.407 < 2e-16 \*\*\*  
## modelMercedes- GLE Class 4.585e-01 2.329e-02 19.690 < 2e-16 \*\*\*  
## modelMercedes- GLS Class 5.196e-01 4.161e-02 12.487 < 2e-16 \*\*\*  
## modelMercedes- M Class 3.958e-01 6.125e-02 6.462 1.14e-10 \*\*\*  
## modelMercedes- S Class 4.789e-01 3.663e-02 13.076 < 2e-16 \*\*\*  
## modelMercedes- SL CLASS 2.442e-01 2.899e-02 8.425 < 2e-16 \*\*\*  
## modelMercedes- SLK 3.078e-02 4.417e-02 0.697 0.485929   
## modelMercedes- V Class 2.714e-01 3.318e-02 8.181 3.59e-16 \*\*\*  
## modelMercedes- X-CLASS 4.511e-02 3.715e-02 1.214 0.224715   
## modelVW- Amarok 1.426e-01 4.690e-02 3.039 0.002383 \*\*   
## modelVW- Arteon 1.220e-01 3.105e-02 3.928 8.68e-05 \*\*\*  
## modelVW- Beetle -1.084e-01 4.609e-02 -2.351 0.018752 \*   
## modelVW- Caddy -8.632e-02 1.342e-01 -0.643 0.520182   
## modelVW- Caddy Maxi -9.569e-02 1.341e-01 -0.714 0.475494   
## modelVW- Caddy Maxi Life -1.443e-01 6.790e-02 -2.125 0.033616 \*   
## modelVW- Caravelle 4.803e-01 4.070e-02 11.802 < 2e-16 \*\*\*  
## modelVW- CC -1.235e-01 4.620e-02 -2.674 0.007529 \*\*   
## modelVW- Golf -1.013e-02 1.311e-02 -0.773 0.439816   
## modelVW- Golf SV -1.342e-01 3.135e-02 -4.281 1.89e-05 \*\*\*  
## modelVW- Passat -4.750e-02 1.817e-02 -2.614 0.008970 \*\*   
## modelVW- Polo -2.247e-01 1.373e-02 -16.363 < 2e-16 \*\*\*  
## modelVW- Scirocco -1.946e-02 2.931e-02 -0.664 0.506771   
## modelVW- Sharan 8.395e-02 3.043e-02 2.759 0.005823 \*\*   
## modelVW- Shuttle 1.149e-01 4.410e-02 2.605 0.009204 \*\*   
## modelVW- T-Cross -1.521e-03 2.794e-02 -0.054 0.956585   
## modelVW- T-Roc 8.763e-02 2.000e-02 4.381 1.21e-05 \*\*\*  
## modelVW- Tiguan 1.264e-01 1.548e-02 8.168 3.98e-16 \*\*\*  
## modelVW- Tiguan Allspace 1.678e-01 5.593e-02 2.999 0.002720 \*\*   
## modelVW- Touareg 2.072e-01 2.626e-02 7.891 3.69e-15 \*\*\*  
## modelVW- Touran 1.172e-01 3.017e-02 3.883 0.000104 \*\*\*  
## modelVW- Up -5.252e-01 1.855e-02 -28.319 < 2e-16 \*\*\*  
## transmissionf.Trans-SemiAuto 1.136e-01 5.718e-03 19.861 < 2e-16 \*\*\*  
## transmissionf.Trans-Automatic 9.694e-02 6.128e-03 15.818 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1335 on 4705 degrees of freedom  
## Multiple R-squared: 0.9129, Adjusted R-squared: 0.9113   
## F-statistic: 567 on 87 and 4705 DF, p-value: < 2.2e-16

* *Improved normality in the regression is evident after the transformation, yet lower quantiles still exhibit a departure from a normal distribution. The residuals demonstrate a linear distribution, but the presence of influential data points affecting the regression is notable. A thorough analysis, including the removal of influential data, is recommended to refine model performance.*

par(mfrow=c(2,2))  
plot(m6,id.n=0)



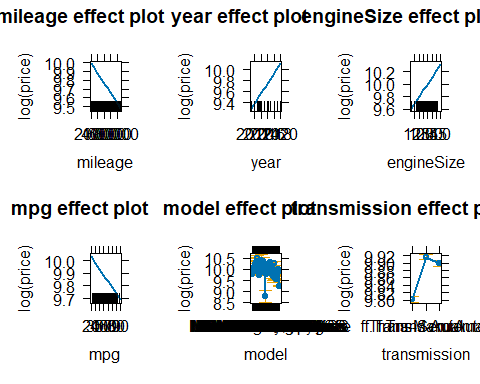
* *The variance inflation factor (VIF) values for the variables in the model (m6) indicate potential issues with multicollinearity. Variables like ‘mileage,’ ‘year,’ ‘engineSize,’ and ‘mpg’ show moderate VIF values, suggesting some correlation with other predictors. However, the ‘model’ variable has a high VIF of 5.47, indicating a substantial level of multicollinearity with other categorical variables.*

vif(m6)

## GVIF Df GVIF^(1/(2\*Df))  
## mileage 3.176135 1 1.782171  
## year 3.216757 1 1.793532  
## engineSize 3.035761 1 1.742344  
## mpg 1.863838 1 1.365225  
## model 5.409976 81 1.010476  
## transmission 1.889205 2 1.172384

* *Given the high VIF for the ‘model’ variable, it is influencing the effects plot in an unexpected way due to multicollinearity.*

plot(allEffects(m6))



* *m6 exhibits a negative and lowest AIC, indicating a potential improvement in model fit compared to others. This suggests that the inclusion of variables or transformations in m6 have enhanced its performance.*

AIC(m0,m1,m2,m3,m4,m5,m6)

## Warning in AIC.default(m0, m1, m2, m3, m4, m5, m6): models are not all fitted  
## to the same number of observations

## df AIC  
## m0 4 97597.573  
## m1 6 94230.152  
## m2 6 22680.546  
## m3 6 22433.157  
## m4 6 22002.090  
## m5 89 19240.208  
## m6 89 -5612.204

* *Each predictor, including mileage, year, engine size, mpg, model, and transmission, exhibits highly significant p-values (< 2.2e-16), indicating their substantial impact on the target variable. The residuals also have a low mean square value, suggesting good model fit. The model’s overall significance is confirmed by a notable F value. The ‘model’ variable and ‘transmission’ both contribute significantly to explaining the variation in log-transformed price. The residuals have a small mean square value, indicating an effective fit.*

anova(m6)

## Analysis of Variance Table  
##   
## Response: log(price)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mileage 1 346.06 346.06 19419.118 < 2.2e-16 \*\*\*  
## year 1 86.79 86.79 4870.126 < 2.2e-16 \*\*\*  
## engineSize 1 323.87 323.87 18174.084 < 2.2e-16 \*\*\*  
## mpg 1 19.33 19.33 1084.867 < 2.2e-16 \*\*\*  
## model 81 95.60 1.18 66.227 < 2.2e-16 \*\*\*  
## transmission 2 7.34 3.67 206.062 < 2.2e-16 \*\*\*  
## Residuals 4705 83.85 0.02   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Can adding other categorical variables improve our model?*

* *The remaining categorical variables have a very low R-squared value in relation to the price, suggesting that they do not significantly contribute to explaining the variability in car prices. Therefore, adding these variables is unlikely to enhance the model’s predictive power regarding price variations.*

### *10.5.3 Interactions*

* *Constructing the following model will reveal which potential interactions play a significant role in explaining the variability. Interactions with lower AIC values indicate greater efficiency in contributing to the model.*
* *In this case, the candidates are: model:transmission and mileage:model.*

mt<-lm(log(price)~(mileage+year+engineSize+mpg + model + transmission)\*(mileage+year+engineSize+mpg + model + transmission),data=df)

mt<-step(mt)

## Start: AIC=-20039.85  
## log(price) ~ (mileage + year + engineSize + mpg + model + transmission) \*   
## (mileage + year + engineSize + mpg + model + transmission)  
##   
## Df Sum of Sq RSS AIC  
## - mileage:model 64 1.1188 61.852 -20080  
## - model:transmission 92 2.2919 63.025 -20046  
## - mileage:engineSize 1 0.0097 60.743 -20041  
## <none> 60.733 -20040  
## - year:engineSize 1 0.0293 60.762 -20040  
## - year:mpg 1 0.0308 60.764 -20039  
## - mileage:transmission 2 0.0752 60.808 -20038  
## - year:transmission 2 0.1007 60.834 -20036  
## - mileage:year 1 0.0951 60.828 -20034  
## - mileage:mpg 1 0.1594 60.893 -20029  
## - engineSize:transmission 2 0.2735 61.007 -20022  
## - mpg:transmission 2 0.3610 61.094 -20015  
## - year:model 61 2.2946 63.028 -19984  
## - engineSize:mpg 1 1.1686 61.902 -19951  
## - engineSize:model 49 4.4776 65.211 -19797  
## - mpg:model 60 5.7399 66.473 -19727  
##   
## Step: AIC=-20080.36  
## log(price) ~ mileage + year + engineSize + mpg + model + transmission +   
## mileage:year + mileage:engineSize + mileage:mpg + mileage:transmission +   
## year:engineSize + year:mpg + year:model + year:transmission +   
## engineSize:mpg + engineSize:model + engineSize:transmission +   
## mpg:model + mpg:transmission + model:transmission  
##   
## Df Sum of Sq RSS AIC  
## - model:transmission 94 2.4249 64.277 -20084  
## - mileage:transmission 2 0.0409 61.893 -20081  
## <none> 61.852 -20080  
## - mileage:engineSize 1 0.0292 61.881 -20080  
## - year:engineSize 1 0.0362 61.888 -20080  
## - mileage:mpg 1 0.0618 61.914 -20078  
## - mileage:year 1 0.0774 61.929 -20076  
## - year:transmission 2 0.1318 61.984 -20074  
## - year:mpg 1 0.1308 61.983 -20072  
## - engineSize:transmission 2 0.2667 62.119 -20064  
## - mpg:transmission 2 0.3535 62.205 -20057  
## - engineSize:mpg 1 1.2458 63.098 -19987  
## - engineSize:model 49 4.6060 66.458 -19834  
## - mpg:model 61 5.7093 67.561 -19779  
## - year:model 63 5.8617 67.714 -19772  
##   
## Step: AIC=-20084.04  
## log(price) ~ mileage + year + engineSize + mpg + model + transmission +   
## mileage:year + mileage:engineSize + mileage:mpg + mileage:transmission +   
## year:engineSize + year:mpg + year:model + year:transmission +   
## engineSize:mpg + engineSize:model + engineSize:transmission +   
## mpg:model + mpg:transmission  
##   
## Df Sum of Sq RSS AIC  
## <none> 64.277 -20084  
## - mileage:transmission 2 0.0564 64.333 -20084  
## - mileage:engineSize 1 0.0340 64.311 -20084  
## - year:engineSize 1 0.0470 64.324 -20083  
## - mileage:mpg 1 0.0597 64.336 -20082  
## - mileage:year 1 0.0678 64.345 -20081  
## - engineSize:transmission 2 0.1432 64.420 -20077  
## - year:transmission 2 0.1461 64.423 -20077  
## - year:mpg 1 0.1418 64.419 -20076  
## - mpg:transmission 2 0.2591 64.536 -20069  
## - engineSize:mpg 1 1.1414 65.418 -20002  
## - mpg:model 67 5.5004 69.777 -19825  
## - engineSize:model 52 5.1937 69.470 -19816  
## - year:model 66 6.8829 71.160 -19729

#### *10.5.3.1 Interaction between two factors*

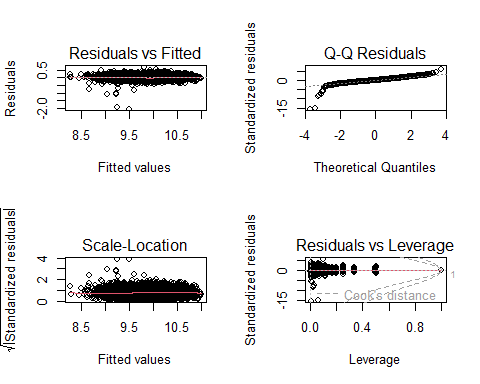
* *Interactions enhance models by capturing nuanced relationships between variables, allowing for non-additive effects. Inclusion of interaction terms, such as between the two most significant factors, model and transmission, enables better representation of complex dependencies, improving predictive accuracy and overall model fit. This allows the model to capture nuanced relationships that may be missed by considering these predictors individually.*

m7<-lm(log(price)~mileage+year+engineSize+mpg + model + transmission + model\*transmission,data=df)  
summary(m7)

##   
## Call:  
## lm(formula = log(price) ~ mileage + year + engineSize + mpg +   
## model + transmission + model \* transmission, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.02919 -0.07484 -0.00145 0.07445 0.70466   
##   
## Coefficients: (60 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) -1.764e+02 3.868e+00  
## mileage -5.546e-06 1.818e-07  
## year 9.226e-02 1.915e-03  
## engineSize 2.116e-01 6.524e-03  
## mpg -4.823e-03 2.418e-04  
## modelAudi- A3 9.784e-02 1.875e-02  
## modelAudi- A4 1.220e-01 2.248e-02  
## modelAudi- A5 2.259e-01 3.698e-02  
## modelAudi- A6 2.587e-01 3.808e-02  
## modelAudi- A7 2.249e-01 8.162e-02  
## modelAudi- A8 3.535e-01 6.928e-02  
## modelAudi- Q2 1.471e-01 2.441e-02  
## modelAudi- Q3 2.140e-01 2.096e-02  
## modelAudi- Q5 3.490e-01 4.899e-02  
## modelAudi- Q7 4.041e-01 6.042e-02  
## modelAudi- Q8 5.492e-01 9.020e-02  
## modelAudi- RS3 4.193e-01 1.048e-01  
## modelAudi- RS5 5.573e-01 1.347e-01  
## modelAudi- RS6 7.786e-01 1.054e-01  
## modelAudi- S3 3.394e-01 1.403e-01  
## modelAudi- S4 3.235e-01 1.407e-01  
## modelAudi- S8 5.169e-01 1.406e-01  
## modelAudi- SQ5 4.384e-01 1.051e-01  
## modelAudi- TT 1.534e-01 3.500e-02  
## modelBMW- 1 Series -3.573e-02 1.844e-02  
## modelBMW- 2 Series -1.221e-03 2.519e-02  
## modelBMW- 3 Series -1.813e-03 2.392e-02  
## modelBMW- 4 Series 1.215e-01 5.585e-02  
## modelBMW- 5 Series 1.481e-01 5.077e-02  
## modelBMW- 6 Series 1.633e-01 6.489e-02  
## modelBMW- 7 Series 3.269e-01 8.164e-02  
## modelBMW- 8 Series 6.528e-01 1.347e-01  
## modelBMW- i3 2.583e-01 8.952e-02  
## modelBMW- M2 3.385e-01 1.407e-01  
## modelBMW- M3 4.098e-01 1.334e-01  
## modelBMW- M4 3.506e-01 7.223e-02  
## modelBMW- M6 5.651e-01 1.050e-01  
## modelBMW- X1 4.137e-02 3.585e-02  
## modelBMW- X2 1.063e-01 1.330e-01  
## modelBMW- X3 2.647e-01 5.455e-02  
## modelBMW- X4 2.919e-01 6.889e-02  
## modelBMW- X5 4.432e-01 5.462e-02  
## modelBMW- X6 5.705e-01 1.407e-01  
## modelBMW- Z3 -1.001e+00 1.332e-01  
## modelBMW- Z4 1.370e-01 1.046e-01  
## modelMercedes- A Class 9.213e-02 2.068e-02  
## modelMercedes- B Class -4.731e-02 5.567e-02  
## modelMercedes- C Class -7.063e-02 3.495e-02  
## modelMercedes- CL Class 2.166e-01 4.059e-02  
## modelMercedes- CLA Class 1.166e-01 9.461e-02  
## modelMercedes- CLS Class 1.861e-01 7.578e-02  
## modelMercedes- E Class 1.879e-01 4.972e-02  
## modelMercedes- GL Class 1.309e-01 7.759e-02  
## modelMercedes- GLA Class 7.710e-02 4.404e-02  
## modelMercedes- GLB Class 3.192e-01 1.344e-01  
## modelMercedes- GLC Class 3.403e-01 5.271e-02  
## modelMercedes- GLE Class 4.327e-01 5.575e-02  
## modelMercedes- GLS Class 4.973e-01 6.697e-02  
## modelMercedes- M Class 4.357e-01 6.418e-02  
## modelMercedes- S Class 5.411e-01 6.342e-02  
## modelMercedes- SL CLASS 1.268e-01 1.330e-01  
## modelMercedes- SLK 8.527e-03 7.159e-02  
## modelMercedes- V Class 1.340e-01 4.436e-02  
## modelMercedes- X-CLASS -1.334e-01 1.331e-01  
## modelVW- Amarok 2.577e-01 7.801e-02  
## modelVW- Arteon 1.340e-01 9.460e-02  
## modelVW- Beetle -1.293e-01 4.875e-02  
## modelVW- Caddy -8.701e-02 1.404e-01  
## modelVW- Caddy Maxi -5.578e-02 1.345e-01  
## modelVW- Caddy Maxi Life -1.505e-01 1.330e-01  
## modelVW- Caravelle 4.950e-01 7.164e-02  
## modelVW- CC -8.760e-02 5.581e-02  
## modelVW- Golf -2.345e-02 1.552e-02  
## modelVW- Golf SV -1.310e-01 4.610e-02  
## modelVW- Passat -9.251e-02 2.348e-02  
## modelVW- Polo -2.409e-01 1.572e-02  
## modelVW- Scirocco -3.288e-04 3.415e-02  
## modelVW- Sharan 1.152e-01 5.199e-02  
## modelVW- Shuttle 1.451e-01 6.091e-02  
## modelVW- T-Cross -8.590e-03 3.413e-02  
## modelVW- T-Roc 8.456e-02 2.356e-02  
## modelVW- Tiguan 1.060e-01 1.923e-02  
## modelVW- Tiguan Allspace 6.541e-02 1.330e-01  
## modelVW- Touareg 2.019e-01 5.826e-02  
## modelVW- Touran 1.300e-01 4.215e-02  
## modelVW- Up -5.401e-01 1.984e-02  
## transmissionf.Trans-SemiAuto 5.914e-02 2.785e-02  
## transmissionf.Trans-Automatic 8.132e-02 4.864e-02  
## modelAudi- A3:transmissionf.Trans-SemiAuto 5.111e-02 3.605e-02  
## modelAudi- A4:transmissionf.Trans-SemiAuto 3.020e-02 3.967e-02  
## modelAudi- A5:transmissionf.Trans-SemiAuto -3.604e-02 5.084e-02  
## modelAudi- A6:transmissionf.Trans-SemiAuto -4.385e-03 4.990e-02  
## modelAudi- A7:transmissionf.Trans-SemiAuto -6.141e-04 1.003e-01  
## modelAudi- A8:transmissionf.Trans-SemiAuto -3.232e-02 9.363e-02  
## modelAudi- Q2:transmissionf.Trans-SemiAuto 1.147e-02 4.484e-02  
## modelAudi- Q3:transmissionf.Trans-SemiAuto 3.599e-02 3.695e-02  
## modelAudi- Q5:transmissionf.Trans-SemiAuto 9.725e-03 5.713e-02  
## modelAudi- Q7:transmissionf.Trans-SemiAuto 1.317e-01 7.134e-02  
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS6:transmissionf.Trans-SemiAuto -6.261e-02 1.423e-01  
## modelAudi- S3:transmissionf.Trans-SemiAuto 5.758e-02 1.703e-01  
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- TT:transmissionf.Trans-SemiAuto 8.899e-02 6.114e-02  
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 2.219e-02 3.457e-02  
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 3.355e-02 4.014e-02  
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 1.165e-01 3.599e-02  
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -6.426e-03 6.416e-02  
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 5.665e-02 5.893e-02  
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 3.129e-02 8.308e-02  
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.471e-01 9.659e-02  
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- M2:transmissionf.Trans-SemiAuto -1.058e-02 1.942e-01  
## modelBMW- M3:transmissionf.Trans-SemiAuto 6.060e-02 1.643e-01  
## modelBMW- M4:transmissionf.Trans-SemiAuto 2.109e-02 8.327e-02  
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- X1:transmissionf.Trans-SemiAuto 1.216e-01 5.060e-02  
## modelBMW- X2:transmissionf.Trans-SemiAuto 9.285e-02 1.397e-01  
## modelBMW- X3:transmissionf.Trans-SemiAuto 5.430e-02 6.551e-02  
## modelBMW- X4:transmissionf.Trans-SemiAuto 2.063e-02 8.021e-02  
## modelBMW- X5:transmissionf.Trans-SemiAuto 6.452e-02 7.223e-02  
## modelBMW- X6:transmissionf.Trans-SemiAuto -9.307e-02 1.569e-01  
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- Z4:transmissionf.Trans-SemiAuto 3.898e-02 1.225e-01  
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 3.894e-02 3.460e-02  
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 1.216e-01 6.667e-02  
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 2.823e-01 4.334e-02  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -8.914e-03 5.334e-02  
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.431e-01 8.579e-02  
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 1.004e-01 5.655e-02  
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 4.106e-01 1.554e-01  
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 6.416e-02 5.376e-02  
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 5.138e-02 5.961e-02  
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 8.198e-02 6.486e-02  
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.037e-01 9.660e-02  
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.271e-01 8.628e-02  
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.464e-01 1.389e-01  
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 9.951e-02 1.003e-01  
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Arteon:transmissionf.Trans-SemiAuto 1.595e-02 1.042e-01  
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 7.852e-02 1.643e-01  
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 1.281e-02 9.274e-02  
## modelVW- CC:transmissionf.Trans-SemiAuto -9.722e-02 9.752e-02  
## modelVW- Golf:transmissionf.Trans-SemiAuto 3.439e-02 3.113e-02  
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 1.870e-02 6.997e-02  
## modelVW- Passat:transmissionf.Trans-SemiAuto 1.225e-01 4.086e-02  
## modelVW- Polo:transmissionf.Trans-SemiAuto 7.778e-02 3.584e-02  
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -2.950e-02 8.703e-02  
## modelVW- Sharan:transmissionf.Trans-SemiAuto -6.994e-03 6.873e-02  
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -3.422e-02 9.291e-02  
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA  
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 3.193e-02 5.215e-02  
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 6.663e-02 3.447e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.825e-01 1.504e-01  
## modelVW- Touareg:transmissionf.Trans-SemiAuto 4.644e-02 6.778e-02  
## modelVW- Touran:transmissionf.Trans-SemiAuto 2.930e-02 6.416e-02  
## modelVW- Up:transmissionf.Trans-SemiAuto 6.938e-02 9.860e-02  
## modelAudi- A3:transmissionf.Trans-Automatic -1.031e-02 5.521e-02  
## modelAudi- A4:transmissionf.Trans-Automatic -4.361e-03 5.537e-02  
## modelAudi- A5:transmissionf.Trans-Automatic -3.770e-02 6.605e-02  
## modelAudi- A6:transmissionf.Trans-Automatic -3.902e-02 6.567e-02  
## modelAudi- A7:transmissionf.Trans-Automatic NA NA  
## modelAudi- A8:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q2:transmissionf.Trans-Automatic -3.175e-02 6.868e-02  
## modelAudi- Q3:transmissionf.Trans-Automatic -2.681e-02 5.836e-02  
## modelAudi- Q5:transmissionf.Trans-Automatic -3.207e-02 7.105e-02  
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA  
## modelAudi- S3:transmissionf.Trans-Automatic NA NA  
## modelAudi- S4:transmissionf.Trans-Automatic NA NA  
## modelAudi- S8:transmissionf.Trans-Automatic NA NA  
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA  
## modelAudi- TT:transmissionf.Trans-Automatic -4.071e-02 8.299e-02  
## modelBMW- 1 Series:transmissionf.Trans-Automatic -1.440e-02 5.511e-02  
## modelBMW- 2 Series:transmissionf.Trans-Automatic -2.077e-02 5.883e-02  
## modelBMW- 3 Series:transmissionf.Trans-Automatic 7.079e-02 5.480e-02  
## modelBMW- 4 Series:transmissionf.Trans-Automatic -3.082e-02 7.609e-02  
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- i3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M2:transmissionf.Trans-Automatic NA NA  
## modelBMW- M3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M4:transmissionf.Trans-Automatic NA NA  
## modelBMW- M6:transmissionf.Trans-Automatic NA NA  
## modelBMW- X1:transmissionf.Trans-Automatic 6.778e-02 6.522e-02  
## modelBMW- X2:transmissionf.Trans-Automatic -2.665e-02 1.476e-01  
## modelBMW- X3:transmissionf.Trans-Automatic NA NA  
## modelBMW- X4:transmissionf.Trans-Automatic NA NA  
## modelBMW- X5:transmissionf.Trans-Automatic NA NA  
## modelBMW- X6:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA  
## modelMercedes- A Class:transmissionf.Trans-Automatic -1.502e-03 5.333e-02  
## modelMercedes- B Class:transmissionf.Trans-Automatic 5.561e-02 7.909e-02  
## modelMercedes- C Class:transmissionf.Trans-Automatic 2.627e-01 5.926e-02  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 5.968e-03 7.355e-02  
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 1.949e-01 1.409e-01  
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.672e-01 9.892e-02  
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 9.891e-02 6.981e-02  
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 1.497e-01 1.477e-01  
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA  
## modelMercedes- V Class:transmissionf.Trans-Automatic 2.746e-01 7.785e-02  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.914e-01 1.452e-01  
## modelVW- Amarok:transmissionf.Trans-Automatic -1.807e-01 1.058e-01  
## modelVW- Arteon:transmissionf.Trans-Automatic 4.602e-03 1.166e-01  
## modelVW- Beetle:transmissionf.Trans-Automatic 8.541e-02 1.485e-01  
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -6.513e-02 1.932e-01  
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA  
## modelVW- CC:transmissionf.Trans-Automatic NA NA  
## modelVW- Golf:transmissionf.Trans-Automatic 4.611e-02 5.168e-02  
## modelVW- Golf SV:transmissionf.Trans-Automatic -3.828e-05 9.316e-02  
## modelVW- Passat:transmissionf.Trans-Automatic 6.712e-02 6.302e-02  
## modelVW- Polo:transmissionf.Trans-Automatic 3.535e-02 5.884e-02  
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.282e-01 8.777e-02  
## modelVW- Sharan:transmissionf.Trans-Automatic -6.210e-02 9.609e-02  
## modelVW- Shuttle:transmissionf.Trans-Automatic -7.879e-02 1.528e-01  
## modelVW- T-Cross:transmissionf.Trans-Automatic 3.631e-04 7.130e-02  
## modelVW- T-Roc:transmissionf.Trans-Automatic -3.057e-02 6.695e-02  
## modelVW- Tiguan:transmissionf.Trans-Automatic 3.274e-02 5.880e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 3.284e-02 1.932e-01  
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA  
## modelVW- Touran:transmissionf.Trans-Automatic -1.642e-01 1.127e-01  
## modelVW- Up:transmissionf.Trans-Automatic 2.202e-01 1.418e-01  
## t value Pr(>|t|)   
## (Intercept) -45.615 < 2e-16 \*\*\*  
## mileage -30.496 < 2e-16 \*\*\*  
## year 48.177 < 2e-16 \*\*\*  
## engineSize 32.429 < 2e-16 \*\*\*  
## mpg -19.948 < 2e-16 \*\*\*  
## modelAudi- A3 5.219 1.88e-07 \*\*\*  
## modelAudi- A4 5.426 6.04e-08 \*\*\*  
## modelAudi- A5 6.109 1.09e-09 \*\*\*  
## modelAudi- A6 6.794 1.23e-11 \*\*\*  
## modelAudi- A7 2.756 0.005872 \*\*   
## modelAudi- A8 5.103 3.48e-07 \*\*\*  
## modelAudi- Q2 6.027 1.80e-09 \*\*\*  
## modelAudi- Q3 10.212 < 2e-16 \*\*\*  
## modelAudi- Q5 7.124 1.21e-12 \*\*\*  
## modelAudi- Q7 6.688 2.53e-11 \*\*\*  
## modelAudi- Q8 6.088 1.24e-09 \*\*\*  
## modelAudi- RS3 4.000 6.45e-05 \*\*\*  
## modelAudi- RS5 4.137 3.57e-05 \*\*\*  
## modelAudi- RS6 7.388 1.76e-13 \*\*\*  
## modelAudi- S3 2.419 0.015600 \*   
## modelAudi- S4 2.300 0.021513 \*   
## modelAudi- S8 3.678 0.000238 \*\*\*  
## modelAudi- SQ5 4.173 3.06e-05 \*\*\*  
## modelAudi- TT 4.382 1.20e-05 \*\*\*  
## modelBMW- 1 Series -1.937 0.052758 .   
## modelBMW- 2 Series -0.048 0.961349   
## modelBMW- 3 Series -0.076 0.939596   
## modelBMW- 4 Series 2.176 0.029570 \*   
## modelBMW- 5 Series 2.917 0.003557 \*\*   
## modelBMW- 6 Series 2.517 0.011865 \*   
## modelBMW- 7 Series 4.004 6.33e-05 \*\*\*  
## modelBMW- 8 Series 4.845 1.31e-06 \*\*\*  
## modelBMW- i3 2.885 0.003932 \*\*   
## modelBMW- M2 2.406 0.016177 \*   
## modelBMW- M3 3.072 0.002142 \*\*   
## modelBMW- M4 4.854 1.25e-06 \*\*\*  
## modelBMW- M6 5.380 7.82e-08 \*\*\*  
## modelBMW- X1 1.154 0.248587   
## modelBMW- X2 0.799 0.424123   
## modelBMW- X3 4.852 1.26e-06 \*\*\*  
## modelBMW- X4 4.237 2.31e-05 \*\*\*  
## modelBMW- X5 8.114 6.22e-16 \*\*\*  
## modelBMW- X6 4.056 5.07e-05 \*\*\*  
## modelBMW- Z3 -7.514 6.85e-14 \*\*\*  
## modelBMW- Z4 1.309 0.190629   
## modelMercedes- A Class 4.454 8.62e-06 \*\*\*  
## modelMercedes- B Class -0.850 0.395471   
## modelMercedes- C Class -2.021 0.043376 \*   
## modelMercedes- CL Class 5.338 9.86e-08 \*\*\*  
## modelMercedes- CLA Class 1.232 0.218062   
## modelMercedes- CLS Class 2.456 0.014086 \*   
## modelMercedes- E Class 3.779 0.000160 \*\*\*  
## modelMercedes- GL Class 1.687 0.091588 .   
## modelMercedes- GLA Class 1.750 0.080100 .   
## modelMercedes- GLB Class 2.374 0.017640 \*   
## modelMercedes- GLC Class 6.457 1.18e-10 \*\*\*  
## modelMercedes- GLE Class 7.761 1.03e-14 \*\*\*  
## modelMercedes- GLS Class 7.426 1.33e-13 \*\*\*  
## modelMercedes- M Class 6.788 1.28e-11 \*\*\*  
## modelMercedes- S Class 8.532 < 2e-16 \*\*\*  
## modelMercedes- SL CLASS 0.953 0.340626   
## modelMercedes- SLK 0.119 0.905193   
## modelMercedes- V Class 3.021 0.002534 \*\*   
## modelMercedes- X-CLASS -1.002 0.316197   
## modelVW- Amarok 3.304 0.000960 \*\*\*  
## modelVW- Arteon 1.417 0.156563   
## modelVW- Beetle -2.652 0.008028 \*\*   
## modelVW- Caddy -0.620 0.535396   
## modelVW- Caddy Maxi -0.415 0.678263   
## modelVW- Caddy Maxi Life -1.132 0.257866   
## modelVW- Caravelle 6.910 5.53e-12 \*\*\*  
## modelVW- CC -1.569 0.116610   
## modelVW- Golf -1.511 0.130896   
## modelVW- Golf SV -2.841 0.004514 \*\*   
## modelVW- Passat -3.940 8.27e-05 \*\*\*  
## modelVW- Polo -15.321 < 2e-16 \*\*\*  
## modelVW- Scirocco -0.010 0.992319   
## modelVW- Sharan 2.216 0.026774 \*   
## modelVW- Shuttle 2.382 0.017246 \*   
## modelVW- T-Cross -0.252 0.801301   
## modelVW- T-Roc 3.589 0.000335 \*\*\*  
## modelVW- Tiguan 5.512 3.75e-08 \*\*\*  
## modelVW- Tiguan Allspace 0.492 0.622902   
## modelVW- Touareg 3.466 0.000533 \*\*\*  
## modelVW- Touran 3.085 0.002045 \*\*   
## modelVW- Up -27.217 < 2e-16 \*\*\*  
## transmissionf.Trans-SemiAuto 2.124 0.033754 \*   
## transmissionf.Trans-Automatic 1.672 0.094628 .   
## modelAudi- A3:transmissionf.Trans-SemiAuto 1.418 0.156332   
## modelAudi- A4:transmissionf.Trans-SemiAuto 0.761 0.446504   
## modelAudi- A5:transmissionf.Trans-SemiAuto -0.709 0.478497   
## modelAudi- A6:transmissionf.Trans-SemiAuto -0.088 0.929982   
## modelAudi- A7:transmissionf.Trans-SemiAuto -0.006 0.995114   
## modelAudi- A8:transmissionf.Trans-SemiAuto -0.345 0.729997   
## modelAudi- Q2:transmissionf.Trans-SemiAuto 0.256 0.798139   
## modelAudi- Q3:transmissionf.Trans-SemiAuto 0.974 0.330067   
## modelAudi- Q5:transmissionf.Trans-SemiAuto 0.170 0.864836   
## modelAudi- Q7:transmissionf.Trans-SemiAuto 1.847 0.064861 .   
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS6:transmissionf.Trans-SemiAuto -0.440 0.660042   
## modelAudi- S3:transmissionf.Trans-SemiAuto 0.338 0.735361   
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- TT:transmissionf.Trans-SemiAuto 1.456 0.145550   
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 0.642 0.520947   
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 0.836 0.403292   
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 3.236 0.001221 \*\*   
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -0.100 0.920223   
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 0.961 0.336479   
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 0.377 0.706477   
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.523 0.127800   
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- M2:transmissionf.Trans-SemiAuto -0.054 0.956550   
## modelBMW- M3:transmissionf.Trans-SemiAuto 0.369 0.712310   
## modelBMW- M4:transmissionf.Trans-SemiAuto 0.253 0.800024   
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- X1:transmissionf.Trans-SemiAuto 2.402 0.016327 \*   
## modelBMW- X2:transmissionf.Trans-SemiAuto 0.665 0.506199   
## modelBMW- X3:transmissionf.Trans-SemiAuto 0.829 0.407192   
## modelBMW- X4:transmissionf.Trans-SemiAuto 0.257 0.797066   
## modelBMW- X5:transmissionf.Trans-SemiAuto 0.893 0.371707   
## modelBMW- X6:transmissionf.Trans-SemiAuto -0.593 0.553094   
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- Z4:transmissionf.Trans-SemiAuto 0.318 0.750360   
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 1.125 0.260504   
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 1.824 0.068205 .   
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 6.514 8.08e-11 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -0.167 0.867300   
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.668 0.095393 .   
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 1.776 0.075846 .   
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 2.643 0.008248 \*\*   
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 1.194 0.232724   
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 0.862 0.388723   
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 1.264 0.206300   
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.073 0.283286   
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.473 0.140800   
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.054 0.291861   
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 0.992 0.321102   
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Arteon:transmissionf.Trans-SemiAuto 0.153 0.878330   
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 0.478 0.632710   
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 0.138 0.890130   
## modelVW- CC:transmissionf.Trans-SemiAuto -0.997 0.318829   
## modelVW- Golf:transmissionf.Trans-SemiAuto 1.105 0.269310   
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 0.267 0.789291   
## modelVW- Passat:transmissionf.Trans-SemiAuto 2.999 0.002723 \*\*   
## modelVW- Polo:transmissionf.Trans-SemiAuto 2.171 0.030014 \*   
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -0.339 0.734665   
## modelVW- Sharan:transmissionf.Trans-SemiAuto -0.102 0.918958   
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -0.368 0.712685   
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA   
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 0.612 0.540388   
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 1.933 0.053281 .   
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.213 0.225032   
## modelVW- Touareg:transmissionf.Trans-SemiAuto 0.685 0.493282   
## modelVW- Touran:transmissionf.Trans-SemiAuto 0.457 0.647885   
## modelVW- Up:transmissionf.Trans-SemiAuto 0.704 0.481709   
## modelAudi- A3:transmissionf.Trans-Automatic -0.187 0.851866   
## modelAudi- A4:transmissionf.Trans-Automatic -0.079 0.937233   
## modelAudi- A5:transmissionf.Trans-Automatic -0.571 0.568191   
## modelAudi- A6:transmissionf.Trans-Automatic -0.594 0.552423   
## modelAudi- A7:transmissionf.Trans-Automatic NA NA   
## modelAudi- A8:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q2:transmissionf.Trans-Automatic -0.462 0.643908   
## modelAudi- Q3:transmissionf.Trans-Automatic -0.459 0.645972   
## modelAudi- Q5:transmissionf.Trans-Automatic -0.451 0.651720   
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA   
## modelAudi- S3:transmissionf.Trans-Automatic NA NA   
## modelAudi- S4:transmissionf.Trans-Automatic NA NA   
## modelAudi- S8:transmissionf.Trans-Automatic NA NA   
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA   
## modelAudi- TT:transmissionf.Trans-Automatic -0.490 0.623810   
## modelBMW- 1 Series:transmissionf.Trans-Automatic -0.261 0.793873   
## modelBMW- 2 Series:transmissionf.Trans-Automatic -0.353 0.724104   
## modelBMW- 3 Series:transmissionf.Trans-Automatic 1.292 0.196480   
## modelBMW- 4 Series:transmissionf.Trans-Automatic -0.405 0.685485   
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- i3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M2:transmissionf.Trans-Automatic NA NA   
## modelBMW- M3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M4:transmissionf.Trans-Automatic NA NA   
## modelBMW- M6:transmissionf.Trans-Automatic NA NA   
## modelBMW- X1:transmissionf.Trans-Automatic 1.039 0.298739   
## modelBMW- X2:transmissionf.Trans-Automatic -0.181 0.856748   
## modelBMW- X3:transmissionf.Trans-Automatic NA NA   
## modelBMW- X4:transmissionf.Trans-Automatic NA NA   
## modelBMW- X5:transmissionf.Trans-Automatic NA NA   
## modelBMW- X6:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA   
## modelMercedes- A Class:transmissionf.Trans-Automatic -0.028 0.977527   
## modelMercedes- B Class:transmissionf.Trans-Automatic 0.703 0.481985   
## modelMercedes- C Class:transmissionf.Trans-Automatic 4.433 9.50e-06 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 0.081 0.935325   
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 1.383 0.166657   
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.690 0.091069 .   
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 1.417 0.156583   
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 1.014 0.310727   
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA   
## modelMercedes- V Class:transmissionf.Trans-Automatic 3.527 0.000424 \*\*\*  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.318 0.187557   
## modelVW- Amarok:transmissionf.Trans-Automatic -1.708 0.087624 .   
## modelVW- Arteon:transmissionf.Trans-Automatic 0.039 0.968523   
## modelVW- Beetle:transmissionf.Trans-Automatic 0.575 0.565238   
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -0.337 0.736050   
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA   
## modelVW- CC:transmissionf.Trans-Automatic NA NA   
## modelVW- Golf:transmissionf.Trans-Automatic 0.892 0.372322   
## modelVW- Golf SV:transmissionf.Trans-Automatic 0.000 0.999672   
## modelVW- Passat:transmissionf.Trans-Automatic 1.065 0.286949   
## modelVW- Polo:transmissionf.Trans-Automatic 0.601 0.548012   
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.461 0.144195   
## modelVW- Sharan:transmissionf.Trans-Automatic -0.646 0.518106   
## modelVW- Shuttle:transmissionf.Trans-Automatic -0.516 0.606089   
## modelVW- T-Cross:transmissionf.Trans-Automatic 0.005 0.995937   
## modelVW- T-Roc:transmissionf.Trans-Automatic -0.457 0.647994   
## modelVW- Tiguan:transmissionf.Trans-Automatic 0.557 0.577746   
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 0.170 0.865014   
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA   
## modelVW- Touran:transmissionf.Trans-Automatic -1.457 0.145204   
## modelVW- Up:transmissionf.Trans-Automatic 1.553 0.120413   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1322 on 4603 degrees of freedom  
## Multiple R-squared: 0.9164, Adjusted R-squared: 0.913   
## F-statistic: 267.1 on 189 and 4603 DF, p-value: < 2.2e-16

* *Despite achieving a 91% explained variability, enhanced homoscedasticity, and improved normality after the interaction addition, lower quantiles in residuals still deviate from normal distribution. While the overall distribution appears linear, influential data points impact the regression, and there’s noticeable leverage.*

par(mfrow=c(2,2))  
plot(m7,id.n=0)



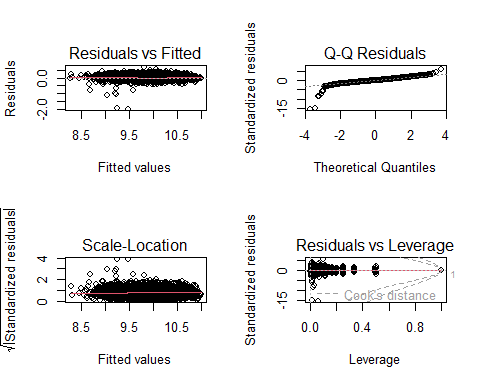
#### *10.5.3.2 Interaction between one factors and one covariate*

* *Pairing a crucial factor with a relevant covariate in model interaction uncovers nuanced relationships, providing deeper insights into their joint impact on the outcome.*
* *This model offered the maximum variability till now.*

library(MASS)  
m8<-lm(log(price) ~ mileage+year+engineSize+mpg+model + transmission + model\*transmission + mileage\*transmission,data=df)  
summary(m8)

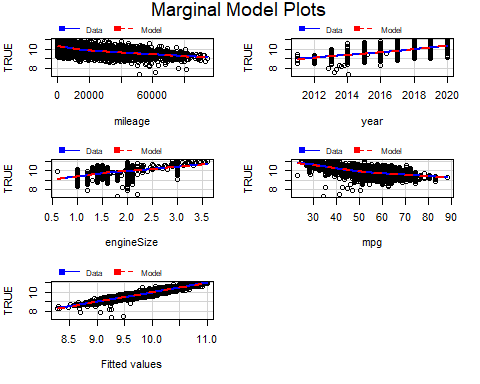
##   
## Call:  
## lm(formula = log(price) ~ mileage + year + engineSize + mpg +   
## model + transmission + model \* transmission + mileage \* transmission,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.99755 -0.07360 -0.00178 0.07602 0.70777   
##   
## Coefficients: (60 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) -1.761e+02 3.873e+00  
## mileage -5.052e-06 2.215e-07  
## year 9.210e-02 1.918e-03  
## engineSize 2.091e-01 6.527e-03  
## mpg -4.907e-03 2.417e-04  
## modelAudi- A3 9.719e-02 1.870e-02  
## modelAudi- A4 1.206e-01 2.243e-02  
## modelAudi- A5 2.219e-01 3.690e-02  
## modelAudi- A6 2.516e-01 3.803e-02  
## modelAudi- A7 2.383e-01 8.145e-02  
## modelAudi- A8 3.433e-01 6.914e-02  
## modelAudi- Q2 1.543e-01 2.444e-02  
## modelAudi- Q3 2.162e-01 2.092e-02  
## modelAudi- Q5 3.405e-01 4.891e-02  
## modelAudi- Q7 4.003e-01 6.027e-02  
## modelAudi- Q8 5.396e-01 9.000e-02  
## modelAudi- RS3 4.210e-01 1.046e-01  
## modelAudi- RS5 5.598e-01 1.344e-01  
## modelAudi- RS6 7.950e-01 1.052e-01  
## modelAudi- S3 3.410e-01 1.399e-01  
## modelAudi- S4 3.122e-01 1.403e-01  
## modelAudi- S8 5.191e-01 1.402e-01  
## modelAudi- SQ5 4.377e-01 1.048e-01  
## modelAudi- TT 1.535e-01 3.491e-02  
## modelBMW- 1 Series -3.712e-02 1.840e-02  
## modelBMW- 2 Series 2.797e-03 2.515e-02  
## modelBMW- 3 Series -7.016e-03 2.391e-02  
## modelBMW- 4 Series 1.290e-01 5.574e-02  
## modelBMW- 5 Series 1.622e-01 5.072e-02  
## modelBMW- 6 Series 1.695e-01 6.474e-02  
## modelBMW- 7 Series 3.475e-01 8.153e-02  
## modelBMW- 8 Series 6.562e-01 1.344e-01  
## modelBMW- i3 2.677e-01 8.931e-02  
## modelBMW- M2 3.223e-01 1.404e-01  
## modelBMW- M3 4.161e-01 1.331e-01  
## modelBMW- M4 3.468e-01 7.205e-02  
## modelBMW- M6 5.718e-01 1.048e-01  
## modelBMW- X1 3.900e-02 3.577e-02  
## modelBMW- X2 1.209e-01 1.328e-01  
## modelBMW- X3 2.720e-01 5.443e-02  
## modelBMW- X4 2.978e-01 6.873e-02  
## modelBMW- X5 4.512e-01 5.451e-02  
## modelBMW- X6 5.591e-01 1.403e-01  
## modelBMW- Z3 -1.017e+00 1.329e-01  
## modelBMW- Z4 1.234e-01 1.044e-01  
## modelMercedes- A Class 9.183e-02 2.063e-02  
## modelMercedes- B Class -5.241e-02 5.555e-02  
## modelMercedes- C Class -7.284e-02 3.487e-02  
## modelMercedes- CL Class 2.142e-01 4.049e-02  
## modelMercedes- CLA Class 1.075e-01 9.441e-02  
## modelMercedes- CLS Class 1.936e-01 7.561e-02  
## modelMercedes- E Class 1.953e-01 4.961e-02  
## modelMercedes- GL Class 1.307e-01 7.739e-02  
## modelMercedes- GLA Class 7.851e-02 4.393e-02  
## modelMercedes- GLB Class 3.224e-01 1.341e-01  
## modelMercedes- GLC Class 3.366e-01 5.258e-02  
## modelMercedes- GLE Class 4.345e-01 5.561e-02  
## modelMercedes- GLS Class 4.940e-01 6.681e-02  
## modelMercedes- M Class 4.280e-01 6.438e-02  
## modelMercedes- S Class 5.424e-01 6.326e-02  
## modelMercedes- SL CLASS 1.264e-01 1.327e-01  
## modelMercedes- SLK 2.136e-02 7.146e-02  
## modelMercedes- V Class 1.348e-01 4.425e-02  
## modelMercedes- X-CLASS -1.264e-01 1.328e-01  
## modelVW- Amarok 2.409e-01 7.793e-02  
## modelVW- Arteon 1.465e-01 9.442e-02  
## modelVW- Beetle -1.369e-01 4.867e-02  
## modelVW- Caddy -5.677e-02 1.401e-01  
## modelVW- Caddy Maxi -5.698e-02 1.341e-01  
## modelVW- Caddy Maxi Life -1.364e-01 1.327e-01  
## modelVW- Caravelle 4.803e-01 7.152e-02  
## modelVW- CC -9.834e-02 5.576e-02  
## modelVW- Golf -2.253e-02 1.549e-02  
## modelVW- Golf SV -1.313e-01 4.598e-02  
## modelVW- Passat -9.789e-02 2.347e-02  
## modelVW- Polo -2.377e-01 1.572e-02  
## modelVW- Scirocco -2.541e-03 3.407e-02  
## modelVW- Sharan 1.156e-01 5.186e-02  
## modelVW- Shuttle 1.451e-01 6.076e-02  
## modelVW- T-Cross 2.231e-03 3.420e-02  
## modelVW- T-Roc 9.375e-02 2.365e-02  
## modelVW- Tiguan 1.087e-01 1.920e-02  
## modelVW- Tiguan Allspace 7.520e-02 1.327e-01  
## modelVW- Touareg 1.971e-01 5.812e-02  
## modelVW- Touran 1.296e-01 4.204e-02  
## modelVW- Up -5.366e-01 1.982e-02  
## transmissionf.Trans-SemiAuto 6.984e-02 2.824e-02  
## transmissionf.Trans-Automatic 1.135e-01 4.897e-02  
## modelAudi- A3:transmissionf.Trans-SemiAuto 5.118e-02 3.598e-02  
## modelAudi- A4:transmissionf.Trans-SemiAuto 3.180e-02 3.959e-02  
## modelAudi- A5:transmissionf.Trans-SemiAuto -3.150e-02 5.075e-02  
## modelAudi- A6:transmissionf.Trans-SemiAuto 9.996e-04 5.000e-02  
## modelAudi- A7:transmissionf.Trans-SemiAuto -1.167e-02 1.001e-01  
## modelAudi- A8:transmissionf.Trans-SemiAuto -1.961e-02 9.343e-02  
## modelAudi- Q2:transmissionf.Trans-SemiAuto 4.684e-03 4.478e-02  
## modelAudi- Q3:transmissionf.Trans-SemiAuto 3.199e-02 3.688e-02  
## modelAudi- Q5:transmissionf.Trans-SemiAuto 1.841e-02 5.706e-02  
## modelAudi- Q7:transmissionf.Trans-SemiAuto 1.366e-01 7.118e-02  
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS6:transmissionf.Trans-SemiAuto -7.727e-02 1.420e-01  
## modelAudi- S3:transmissionf.Trans-SemiAuto 4.695e-02 1.701e-01  
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- TT:transmissionf.Trans-SemiAuto 8.735e-02 6.100e-02  
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 2.415e-02 3.451e-02  
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 3.058e-02 4.004e-02  
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 1.221e-01 3.601e-02  
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -1.238e-02 6.401e-02  
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 4.330e-02 5.887e-02  
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 2.612e-02 8.290e-02  
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.303e-01 9.640e-02  
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- M2:transmissionf.Trans-SemiAuto 9.072e-03 1.938e-01  
## modelBMW- M3:transmissionf.Trans-SemiAuto 5.392e-02 1.639e-01  
## modelBMW- M4:transmissionf.Trans-SemiAuto 2.757e-02 8.307e-02  
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- X1:transmissionf.Trans-SemiAuto 1.232e-01 5.053e-02  
## modelBMW- X2:transmissionf.Trans-SemiAuto 8.072e-02 1.394e-01  
## modelBMW- X3:transmissionf.Trans-SemiAuto 4.783e-02 6.536e-02  
## modelBMW- X4:transmissionf.Trans-SemiAuto 1.639e-02 8.001e-02  
## modelBMW- X5:transmissionf.Trans-SemiAuto 6.048e-02 7.205e-02  
## modelBMW- X6:transmissionf.Trans-SemiAuto -7.898e-02 1.565e-01  
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- Z4:transmissionf.Trans-SemiAuto 5.551e-02 1.222e-01  
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 3.954e-02 3.452e-02  
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 1.279e-01 6.653e-02  
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 2.852e-01 4.326e-02  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -7.431e-03 5.327e-02  
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.364e-01 8.560e-02  
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 9.427e-02 5.644e-02  
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 4.022e-01 1.553e-01  
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 6.292e-02 5.364e-02  
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 5.567e-02 5.946e-02  
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 8.075e-02 6.470e-02  
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.059e-01 9.639e-02  
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.255e-01 8.606e-02  
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.486e-01 1.386e-01  
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 8.349e-02 1.002e-01  
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Arteon:transmissionf.Trans-SemiAuto 4.227e-03 1.040e-01  
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 6.649e-02 1.639e-01  
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 2.757e-02 9.255e-02  
## modelVW- CC:transmissionf.Trans-SemiAuto -9.226e-02 9.760e-02  
## modelVW- Golf:transmissionf.Trans-SemiAuto 3.371e-02 3.105e-02  
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 1.950e-02 6.979e-02  
## modelVW- Passat:transmissionf.Trans-SemiAuto 1.285e-01 4.080e-02  
## modelVW- Polo:transmissionf.Trans-SemiAuto 7.371e-02 3.575e-02  
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -3.076e-02 8.696e-02  
## modelVW- Sharan:transmissionf.Trans-SemiAuto -6.314e-03 6.856e-02  
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -3.420e-02 9.267e-02  
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA  
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 2.428e-02 5.212e-02  
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 6.290e-02 3.440e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.732e-01 1.500e-01  
## modelVW- Touareg:transmissionf.Trans-SemiAuto 5.245e-02 6.762e-02  
## modelVW- Touran:transmissionf.Trans-SemiAuto 2.950e-02 6.400e-02  
## modelVW- Up:transmissionf.Trans-SemiAuto 6.578e-02 9.836e-02  
## modelAudi- A3:transmissionf.Trans-Automatic -6.820e-03 5.507e-02  
## modelAudi- A4:transmissionf.Trans-Automatic -6.384e-03 5.524e-02  
## modelAudi- A5:transmissionf.Trans-Automatic -3.143e-02 6.590e-02  
## modelAudi- A6:transmissionf.Trans-Automatic -2.129e-02 6.561e-02  
## modelAudi- A7:transmissionf.Trans-Automatic NA NA  
## modelAudi- A8:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q2:transmissionf.Trans-Automatic -5.176e-02 6.863e-02  
## modelAudi- Q3:transmissionf.Trans-Automatic -3.410e-02 5.823e-02  
## modelAudi- Q5:transmissionf.Trans-Automatic -3.153e-02 7.090e-02  
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA  
## modelAudi- S3:transmissionf.Trans-Automatic NA NA  
## modelAudi- S4:transmissionf.Trans-Automatic NA NA  
## modelAudi- S8:transmissionf.Trans-Automatic NA NA  
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA  
## modelAudi- TT:transmissionf.Trans-Automatic -4.492e-02 8.279e-02  
## modelBMW- 1 Series:transmissionf.Trans-Automatic -6.710e-03 5.499e-02  
## modelBMW- 2 Series:transmissionf.Trans-Automatic -2.786e-02 5.869e-02  
## modelBMW- 3 Series:transmissionf.Trans-Automatic 8.459e-02 5.473e-02  
## modelBMW- 4 Series:transmissionf.Trans-Automatic -3.705e-02 7.591e-02  
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- i3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M2:transmissionf.Trans-Automatic NA NA  
## modelBMW- M3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M4:transmissionf.Trans-Automatic NA NA  
## modelBMW- M6:transmissionf.Trans-Automatic NA NA  
## modelBMW- X1:transmissionf.Trans-Automatic 7.004e-02 6.506e-02  
## modelBMW- X2:transmissionf.Trans-Automatic -5.732e-02 1.474e-01  
## modelBMW- X3:transmissionf.Trans-Automatic NA NA  
## modelBMW- X4:transmissionf.Trans-Automatic NA NA  
## modelBMW- X5:transmissionf.Trans-Automatic NA NA  
## modelBMW- X6:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA  
## modelMercedes- A Class:transmissionf.Trans-Automatic -2.988e-03 5.319e-02  
## modelMercedes- B Class:transmissionf.Trans-Automatic 5.798e-02 7.890e-02  
## modelMercedes- C Class:transmissionf.Trans-Automatic 2.640e-01 5.911e-02  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 1.391e-02 7.338e-02  
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 2.043e-01 1.406e-01  
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.722e-01 9.867e-02  
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 1.026e-01 6.963e-02  
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 1.476e-01 1.473e-01  
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA  
## modelMercedes- V Class:transmissionf.Trans-Automatic 2.673e-01 7.767e-02  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.716e-01 1.449e-01  
## modelVW- Amarok:transmissionf.Trans-Automatic -1.713e-01 1.056e-01  
## modelVW- Arteon:transmissionf.Trans-Automatic -2.272e-02 1.165e-01  
## modelVW- Beetle:transmissionf.Trans-Automatic 1.104e-01 1.482e-01  
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -8.669e-02 1.928e-01  
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA  
## modelVW- CC:transmissionf.Trans-Automatic NA NA  
## modelVW- Golf:transmissionf.Trans-Automatic 4.146e-02 5.155e-02  
## modelVW- Golf SV:transmissionf.Trans-Automatic -2.449e-03 9.293e-02  
## modelVW- Passat:transmissionf.Trans-Automatic 7.135e-02 6.288e-02  
## modelVW- Polo:transmissionf.Trans-Automatic 2.742e-02 5.872e-02  
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.243e-01 8.756e-02  
## modelVW- Sharan:transmissionf.Trans-Automatic -6.292e-02 9.584e-02  
## modelVW- Shuttle:transmissionf.Trans-Automatic -8.935e-02 1.524e-01  
## modelVW- T-Cross:transmissionf.Trans-Automatic -2.874e-02 7.137e-02  
## modelVW- T-Roc:transmissionf.Trans-Automatic -5.410e-02 6.695e-02  
## modelVW- Tiguan:transmissionf.Trans-Automatic 2.425e-02 5.867e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 8.209e-03 1.928e-01  
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA  
## modelVW- Touran:transmissionf.Trans-Automatic -1.750e-01 1.125e-01  
## modelVW- Up:transmissionf.Trans-Automatic 2.037e-01 1.414e-01  
## mileage:transmissionf.Trans-SemiAuto -2.873e-07 2.707e-07  
## mileage:transmissionf.Trans-Automatic -1.299e-06 2.640e-07  
## t value Pr(>|t|)   
## (Intercept) -45.469 < 2e-16 \*\*\*  
## mileage -22.809 < 2e-16 \*\*\*  
## year 48.023 < 2e-16 \*\*\*  
## engineSize 32.041 < 2e-16 \*\*\*  
## mpg -20.298 < 2e-16 \*\*\*  
## modelAudi- A3 5.197 2.12e-07 \*\*\*  
## modelAudi- A4 5.378 7.91e-08 \*\*\*  
## modelAudi- A5 6.014 1.95e-09 \*\*\*  
## modelAudi- A6 6.616 4.12e-11 \*\*\*  
## modelAudi- A7 2.926 0.003450 \*\*   
## modelAudi- A8 4.965 7.13e-07 \*\*\*  
## modelAudi- Q2 6.314 2.97e-10 \*\*\*  
## modelAudi- Q3 10.338 < 2e-16 \*\*\*  
## modelAudi- Q5 6.960 3.88e-12 \*\*\*  
## modelAudi- Q7 6.642 3.46e-11 \*\*\*  
## modelAudi- Q8 5.995 2.18e-09 \*\*\*  
## modelAudi- RS3 4.025 5.78e-05 \*\*\*  
## modelAudi- RS5 4.166 3.16e-05 \*\*\*  
## modelAudi- RS6 7.559 4.88e-14 \*\*\*  
## modelAudi- S3 2.437 0.014856 \*   
## modelAudi- S4 2.225 0.026137 \*   
## modelAudi- S8 3.702 0.000216 \*\*\*  
## modelAudi- SQ5 4.176 3.02e-05 \*\*\*  
## modelAudi- TT 4.397 1.12e-05 \*\*\*  
## modelBMW- 1 Series -2.017 0.043743 \*   
## modelBMW- 2 Series 0.111 0.911479   
## modelBMW- 3 Series -0.294 0.769151   
## modelBMW- 4 Series 2.315 0.020659 \*   
## modelBMW- 5 Series 3.197 0.001398 \*\*   
## modelBMW- 6 Series 2.619 0.008845 \*\*   
## modelBMW- 7 Series 4.262 2.07e-05 \*\*\*  
## modelBMW- 8 Series 4.882 1.08e-06 \*\*\*  
## modelBMW- i3 2.997 0.002743 \*\*   
## modelBMW- M2 2.296 0.021704 \*   
## modelBMW- M3 3.126 0.001781 \*\*   
## modelBMW- M4 4.813 1.53e-06 \*\*\*  
## modelBMW- M6 5.457 5.10e-08 \*\*\*  
## modelBMW- X1 1.090 0.275581   
## modelBMW- X2 0.911 0.362364   
## modelBMW- X3 4.997 6.04e-07 \*\*\*  
## modelBMW- X4 4.333 1.50e-05 \*\*\*  
## modelBMW- X5 8.278 < 2e-16 \*\*\*  
## modelBMW- X6 3.984 6.88e-05 \*\*\*  
## modelBMW- Z3 -7.653 2.38e-14 \*\*\*  
## modelBMW- Z4 1.182 0.237414   
## modelMercedes- A Class 4.451 8.75e-06 \*\*\*  
## modelMercedes- B Class -0.944 0.345459   
## modelMercedes- C Class -2.089 0.036795 \*   
## modelMercedes- CL Class 5.291 1.27e-07 \*\*\*  
## modelMercedes- CLA Class 1.138 0.255056   
## modelMercedes- CLS Class 2.561 0.010466 \*   
## modelMercedes- E Class 3.936 8.42e-05 \*\*\*  
## modelMercedes- GL Class 1.689 0.091269 .   
## modelMercedes- GLA Class 1.787 0.073994 .   
## modelMercedes- GLB Class 2.404 0.016246 \*   
## modelMercedes- GLC Class 6.402 1.68e-10 \*\*\*  
## modelMercedes- GLE Class 7.813 6.85e-15 \*\*\*  
## modelMercedes- GLS Class 7.394 1.68e-13 \*\*\*  
## modelMercedes- M Class 6.648 3.32e-11 \*\*\*  
## modelMercedes- S Class 8.573 < 2e-16 \*\*\*  
## modelMercedes- SL CLASS 0.953 0.340867   
## modelMercedes- SLK 0.299 0.765072   
## modelMercedes- V Class 3.046 0.002333 \*\*   
## modelMercedes- X-CLASS -0.952 0.341112   
## modelVW- Amarok 3.091 0.002006 \*\*   
## modelVW- Arteon 1.552 0.120752   
## modelVW- Beetle -2.812 0.004941 \*\*   
## modelVW- Caddy -0.405 0.685441   
## modelVW- Caddy Maxi -0.425 0.670992   
## modelVW- Caddy Maxi Life -1.028 0.303956   
## modelVW- Caravelle 6.715 2.11e-11 \*\*\*  
## modelVW- CC -1.764 0.077842 .   
## modelVW- Golf -1.455 0.145723   
## modelVW- Golf SV -2.856 0.004308 \*\*   
## modelVW- Passat -4.170 3.10e-05 \*\*\*  
## modelVW- Polo -15.124 < 2e-16 \*\*\*  
## modelVW- Scirocco -0.075 0.940547   
## modelVW- Sharan 2.230 0.025796 \*   
## modelVW- Shuttle 2.388 0.016962 \*   
## modelVW- T-Cross 0.065 0.947993   
## modelVW- T-Roc 3.963 7.51e-05 \*\*\*  
## modelVW- Tiguan 5.662 1.58e-08 \*\*\*  
## modelVW- Tiguan Allspace 0.567 0.570941   
## modelVW- Touareg 3.391 0.000703 \*\*\*  
## modelVW- Touran 3.083 0.002060 \*\*   
## modelVW- Up -27.074 < 2e-16 \*\*\*  
## transmissionf.Trans-SemiAuto 2.473 0.013435 \*   
## transmissionf.Trans-Automatic 2.318 0.020511 \*   
## modelAudi- A3:transmissionf.Trans-SemiAuto 1.423 0.154933   
## modelAudi- A4:transmissionf.Trans-SemiAuto 0.803 0.421944   
## modelAudi- A5:transmissionf.Trans-SemiAuto -0.621 0.534915   
## modelAudi- A6:transmissionf.Trans-SemiAuto 0.020 0.984053   
## modelAudi- A7:transmissionf.Trans-SemiAuto -0.117 0.907154   
## modelAudi- A8:transmissionf.Trans-SemiAuto -0.210 0.833739   
## modelAudi- Q2:transmissionf.Trans-SemiAuto 0.105 0.916706   
## modelAudi- Q3:transmissionf.Trans-SemiAuto 0.867 0.385775   
## modelAudi- Q5:transmissionf.Trans-SemiAuto 0.323 0.746947   
## modelAudi- Q7:transmissionf.Trans-SemiAuto 1.919 0.054985 .   
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS6:transmissionf.Trans-SemiAuto -0.544 0.586370   
## modelAudi- S3:transmissionf.Trans-SemiAuto 0.276 0.782531   
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- TT:transmissionf.Trans-SemiAuto 1.432 0.152245   
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 0.700 0.484040   
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 0.764 0.445012   
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 3.391 0.000702 \*\*\*  
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -0.193 0.846676   
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 0.736 0.462070   
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 0.315 0.752719   
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.351 0.176625   
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- M2:transmissionf.Trans-SemiAuto 0.047 0.962663   
## modelBMW- M3:transmissionf.Trans-SemiAuto 0.329 0.742240   
## modelBMW- M4:transmissionf.Trans-SemiAuto 0.332 0.739939   
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- X1:transmissionf.Trans-SemiAuto 2.437 0.014848 \*   
## modelBMW- X2:transmissionf.Trans-SemiAuto 0.579 0.562525   
## modelBMW- X3:transmissionf.Trans-SemiAuto 0.732 0.464362   
## modelBMW- X4:transmissionf.Trans-SemiAuto 0.205 0.837727   
## modelBMW- X5:transmissionf.Trans-SemiAuto 0.839 0.401320   
## modelBMW- X6:transmissionf.Trans-SemiAuto -0.505 0.613886   
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- Z4:transmissionf.Trans-SemiAuto 0.454 0.649772   
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 1.145 0.252147   
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 1.922 0.054685 .   
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 6.593 4.79e-11 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -0.140 0.889061   
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.593 0.111167   
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 1.670 0.094917 .   
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 2.590 0.009619 \*\*   
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 1.173 0.240831   
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 0.936 0.349181   
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 1.248 0.212083   
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.099 0.272039   
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.458 0.144944   
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.073 0.283481   
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 0.834 0.404548   
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Arteon:transmissionf.Trans-SemiAuto 0.041 0.967573   
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 0.406 0.685016   
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 0.298 0.765811   
## modelVW- CC:transmissionf.Trans-SemiAuto -0.945 0.344547   
## modelVW- Golf:transmissionf.Trans-SemiAuto 1.086 0.277678   
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 0.279 0.779917   
## modelVW- Passat:transmissionf.Trans-SemiAuto 3.150 0.001642 \*\*   
## modelVW- Polo:transmissionf.Trans-SemiAuto 2.062 0.039298 \*   
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -0.354 0.723541   
## modelVW- Sharan:transmissionf.Trans-SemiAuto -0.092 0.926635   
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -0.369 0.712087   
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA   
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 0.466 0.641296   
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 1.828 0.067559 .   
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.154 0.248470   
## modelVW- Touareg:transmissionf.Trans-SemiAuto 0.776 0.437969   
## modelVW- Touran:transmissionf.Trans-SemiAuto 0.461 0.644848   
## modelVW- Up:transmissionf.Trans-SemiAuto 0.669 0.503666   
## modelAudi- A3:transmissionf.Trans-Automatic -0.124 0.901450   
## modelAudi- A4:transmissionf.Trans-Automatic -0.116 0.907994   
## modelAudi- A5:transmissionf.Trans-Automatic -0.477 0.633382   
## modelAudi- A6:transmissionf.Trans-Automatic -0.324 0.745612   
## modelAudi- A7:transmissionf.Trans-Automatic NA NA   
## modelAudi- A8:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q2:transmissionf.Trans-Automatic -0.754 0.450776   
## modelAudi- Q3:transmissionf.Trans-Automatic -0.586 0.558192   
## modelAudi- Q5:transmissionf.Trans-Automatic -0.445 0.656549   
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA   
## modelAudi- S3:transmissionf.Trans-Automatic NA NA   
## modelAudi- S4:transmissionf.Trans-Automatic NA NA   
## modelAudi- S8:transmissionf.Trans-Automatic NA NA   
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA   
## modelAudi- TT:transmissionf.Trans-Automatic -0.543 0.587402   
## modelBMW- 1 Series:transmissionf.Trans-Automatic -0.122 0.902883   
## modelBMW- 2 Series:transmissionf.Trans-Automatic -0.475 0.635093   
## modelBMW- 3 Series:transmissionf.Trans-Automatic 1.546 0.122277   
## modelBMW- 4 Series:transmissionf.Trans-Automatic -0.488 0.625492   
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- i3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M2:transmissionf.Trans-Automatic NA NA   
## modelBMW- M3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M4:transmissionf.Trans-Automatic NA NA   
## modelBMW- M6:transmissionf.Trans-Automatic NA NA   
## modelBMW- X1:transmissionf.Trans-Automatic 1.077 0.281680   
## modelBMW- X2:transmissionf.Trans-Automatic -0.389 0.697369   
## modelBMW- X3:transmissionf.Trans-Automatic NA NA   
## modelBMW- X4:transmissionf.Trans-Automatic NA NA   
## modelBMW- X5:transmissionf.Trans-Automatic NA NA   
## modelBMW- X6:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA   
## modelMercedes- A Class:transmissionf.Trans-Automatic -0.056 0.955208   
## modelMercedes- B Class:transmissionf.Trans-Automatic 0.735 0.462443   
## modelMercedes- C Class:transmissionf.Trans-Automatic 4.466 8.16e-06 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 0.190 0.849627   
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 1.453 0.146335   
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.745 0.081083 .   
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 1.473 0.140782   
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 1.002 0.316254   
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA   
## modelMercedes- V Class:transmissionf.Trans-Automatic 3.442 0.000583 \*\*\*  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.184 0.236321   
## modelVW- Amarok:transmissionf.Trans-Automatic -1.622 0.104878   
## modelVW- Arteon:transmissionf.Trans-Automatic -0.195 0.845353   
## modelVW- Beetle:transmissionf.Trans-Automatic 0.745 0.456196   
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -0.450 0.652946   
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA   
## modelVW- CC:transmissionf.Trans-Automatic NA NA   
## modelVW- Golf:transmissionf.Trans-Automatic 0.804 0.421277   
## modelVW- Golf SV:transmissionf.Trans-Automatic -0.026 0.978973   
## modelVW- Passat:transmissionf.Trans-Automatic 1.135 0.256589   
## modelVW- Polo:transmissionf.Trans-Automatic 0.467 0.640513   
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.419 0.155833   
## modelVW- Sharan:transmissionf.Trans-Automatic -0.657 0.511532   
## modelVW- Shuttle:transmissionf.Trans-Automatic -0.586 0.557707   
## modelVW- T-Cross:transmissionf.Trans-Automatic -0.403 0.687216   
## modelVW- T-Roc:transmissionf.Trans-Automatic -0.808 0.419067   
## modelVW- Tiguan:transmissionf.Trans-Automatic 0.413 0.679469   
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 0.043 0.966030   
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA   
## modelVW- Touran:transmissionf.Trans-Automatic -1.556 0.119831   
## modelVW- Up:transmissionf.Trans-Automatic 1.440 0.149918   
## mileage:transmissionf.Trans-SemiAuto -1.061 0.288523   
## mileage:transmissionf.Trans-Automatic -4.919 8.99e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1319 on 4601 degrees of freedom  
## Multiple R-squared: 0.9169, Adjusted R-squared: 0.9135   
## F-statistic: 265.8 on 191 and 4601 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(m8,id.n=0)



* *The shape of the marginal models closely mirrors the underlying data, and the fitted values reveal a more robust model. The convergence of the blue and red datasets, along with their alignment with the same underlying function, underscores a compelling consistency between observed and predicted outcomes.*

marginalModelPlots(m8)



* *Which model is better?*
* *Model m8 is preferred over m7 as it has a lower AIC, suggesting a better balance between goodness of fit and model complexity.*

AIC(m7,m8)

## df AIC  
## m7 191 -5606.559  
## m8 193 -5629.022

anova(m7,m8)

## Analysis of Variance Table  
##   
## Model 1: log(price) ~ mileage + year + engineSize + mpg + model + transmission +   
## model \* transmission  
## Model 2: log(price) ~ mileage + year + engineSize + mpg + model + transmission +   
## model \* transmission + mileage \* transmission  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4603 80.447   
## 2 4601 80.004 2 0.44295 12.737 3.046e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## *10.6 Model Validation & Unusual-Influential Data Detection*

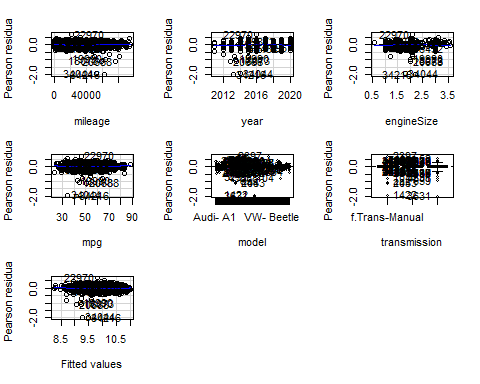
* *We will employ the Breusch-Pagan test to evaluate* ***homoscedasticity****. Since the p-value associated with the test is significantly low, we can reject the null hypothesis of heteroscedasticity. Consequently, this leads us to conclude that the model exhibits homoscedasticity, suggesting that the variance of the residuals is consistent across all levels of the explanatory variables.*

library(lmtest)   
bptest(m8)

##   
## studentized Breusch-Pagan test  
##   
## data: m8  
## BP = 308.15, df = 191, p-value = 1.6e-07

* *The residual plots for the regression model suggest a good fit. Residuals for mileage, year, mpg, car model, and transmission type are randomly spread, indicating these variables are well accounted for in the model.*
* *The lack of clear patterns or systematic trends in these plots suggests that the assumption of* ***independence*** *is met.*

residualPlots(m8,id=list(method=cooks.distance(m8),n=10))



## Test stat Pr(>|Test stat|)   
## mileage 3.8187 0.0001359 \*\*\*  
## year -1.9688 0.0490338 \*   
## engineSize -1.6042 0.1087328   
## mpg 6.3198 2.867e-10 \*\*\*  
## model   
## transmission   
## Tukey test 4.1814 2.897e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Let’s proceed to display the boxplots of the R-student values, Hat values, and Cook’s distances for the observations in the model.

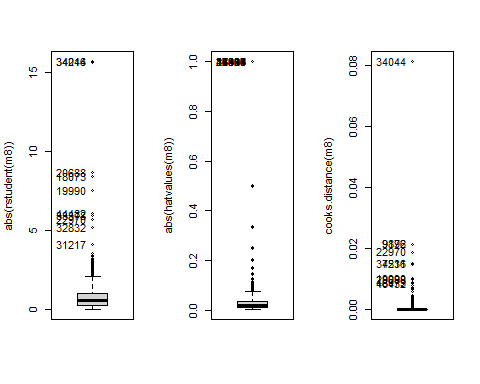
par(mfrow=c(1,3))   
Boxplot(abs(rstudent(m8)),id=list(labels=row.names(df)))

## [1] "34044" "34216" "20688" "18073" "19990" "44482" "33378" "22970" "32832"  
## [10] "31217"

Boxplot(abs(hatvalues(m8)),id=list(labels=row.names(df)))

## [1] "47901" "22486" "16514" "47891" "23307" "31814" "7483" "6466" "15429"  
## [10] "14903"

Boxplot(cooks.distance(m8),id=list(labels=row.names(df)))

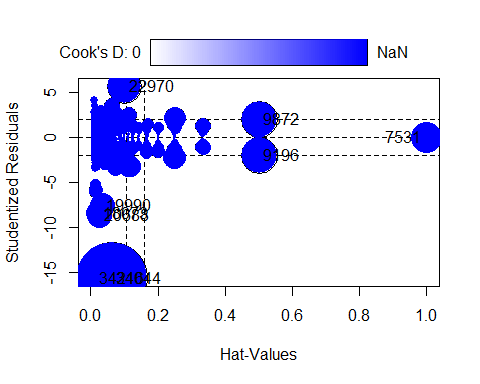


## [1] "34044" "9196" "9872" "22970" "7531" "34216" "19990" "20688" "18073"  
## [10] "46432"

stu <- which(abs(rstudent(m8))>3.0)   
cook <- which(abs(cooks.distance(m8))>0.1)   
hat <- which(abs(hatvalues(m8))>0.2)   
outs<-unique(stu,cook,hat)

* *Spotting influential individuals:*

x<-influencePlot( m8, id=c(list="noteworthy",n=5))



obs<-rownames(x)   
outs<-unique(outs,obs)   
df\_outs<-df[-outs,]

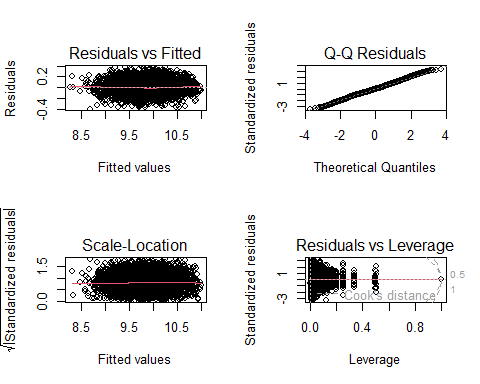
* Building a model without unusual and influential data:

m9<- update(m8, data=df\_outs)   
summary(m9)

##   
## Call:  
## lm(formula = log(price) ~ mileage + year + engineSize + mpg +   
## model + transmission + model \* transmission + mileage \* transmission,   
## data = df\_outs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39674 -0.07435 -0.00172 0.07414 0.38218   
##   
## Coefficients: (60 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) -1.705e+02 3.461e+00  
## mileage -4.803e-06 1.986e-07  
## year 8.936e-02 1.714e-03  
## engineSize 2.025e-01 5.827e-03  
## mpg -5.593e-03 2.177e-04  
## modelAudi- A3 1.018e-01 1.667e-02  
## modelAudi- A4 1.223e-01 1.995e-02  
## modelAudi- A5 2.181e-01 3.283e-02  
## modelAudi- A6 2.491e-01 3.383e-02  
## modelAudi- A7 2.403e-01 7.246e-02  
## modelAudi- A8 3.583e-01 6.151e-02  
## modelAudi- Q2 1.559e-01 2.174e-02  
## modelAudi- Q3 2.137e-01 1.861e-02  
## modelAudi- Q5 3.260e-01 4.352e-02  
## modelAudi- Q7 4.071e-01 5.362e-02  
## modelAudi- Q8 5.472e-01 8.006e-02  
## modelAudi- RS3 4.147e-01 9.303e-02  
## modelAudi- RS5 5.577e-01 1.195e-01  
## modelAudi- RS6 7.798e-01 9.356e-02  
## modelAudi- S3 3.376e-01 1.245e-01  
## modelAudi- S4 3.246e-01 1.248e-01  
## modelAudi- S8 5.051e-01 1.247e-01  
## modelAudi- SQ5 4.362e-01 9.323e-02  
## modelAudi- TT 1.509e-01 3.105e-02  
## modelBMW- 1 Series -3.161e-02 1.640e-02  
## modelBMW- 2 Series 6.636e-03 2.238e-02  
## modelBMW- 3 Series 4.233e-02 2.156e-02  
## modelBMW- 4 Series 1.285e-01 4.958e-02  
## modelBMW- 5 Series 1.631e-01 4.512e-02  
## modelBMW- 6 Series 1.686e-01 5.759e-02  
## modelBMW- 7 Series 3.437e-01 7.253e-02  
## modelBMW- 8 Series 6.626e-01 1.196e-01  
## modelBMW- i3 2.646e-01 7.945e-02  
## modelBMW- M2 3.296e-01 1.249e-01  
## modelBMW- M3 4.139e-01 1.184e-01  
## modelBMW- M4 3.486e-01 6.409e-02  
## modelBMW- M6 5.539e-01 9.321e-02  
## modelBMW- X1 3.720e-02 3.182e-02  
## modelBMW- X2 1.335e-01 1.181e-01  
## modelBMW- X3 2.711e-01 4.842e-02  
## modelBMW- X4 2.979e-01 6.114e-02  
## modelBMW- X5 4.516e-01 4.849e-02  
## modelBMW- X6 5.685e-01 1.248e-01  
## modelBMW- Z3 -1.046e+00 1.182e-01  
## modelBMW- Z4 1.270e-01 9.287e-02  
## modelMercedes- A Class 9.263e-02 1.835e-02  
## modelMercedes- B Class -5.765e-02 4.941e-02  
## modelMercedes- C Class 3.121e-02 3.270e-02  
## modelMercedes- CL Class 2.132e-01 3.602e-02  
## modelMercedes- CLA Class 1.082e-01 8.398e-02  
## modelMercedes- CLS Class 1.956e-01 6.725e-02  
## modelMercedes- E Class 2.069e-01 4.421e-02  
## modelMercedes- GL Class 1.331e-01 6.884e-02  
## modelMercedes- GLA Class 7.830e-02 3.908e-02  
## modelMercedes- GLB Class 3.280e-01 1.193e-01  
## modelMercedes- GLC Class 3.433e-01 4.677e-02  
## modelMercedes- GLE Class 4.343e-01 4.947e-02  
## modelMercedes- GLS Class 4.981e-01 5.943e-02  
## modelMercedes- M Class 4.199e-01 5.727e-02  
## modelMercedes- S Class 5.461e-01 5.628e-02  
## modelMercedes- SL CLASS 1.201e-01 1.180e-01  
## modelMercedes- SLK 1.404e-02 6.357e-02  
## modelMercedes- V Class 5.528e-02 4.129e-02  
## modelMercedes- X-CLASS -1.288e-01 1.181e-01  
## modelVW- Amarok 2.149e-01 6.933e-02  
## modelVW- Arteon 1.607e-01 8.399e-02  
## modelVW- Beetle -1.494e-01 4.330e-02  
## modelVW- Caddy -6.681e-02 1.247e-01  
## modelVW- Caddy Maxi -5.057e-02 1.193e-01  
## modelVW- Caddy Maxi Life -1.236e-01 1.180e-01  
## modelVW- Caravelle 4.823e-01 6.362e-02  
## modelVW- CC -1.056e-01 4.960e-02  
## modelVW- Golf -2.080e-02 1.378e-02  
## modelVW- Golf SV -1.294e-01 4.090e-02  
## modelVW- Passat -8.797e-02 2.103e-02  
## modelVW- Polo -2.344e-01 1.399e-02  
## modelVW- Scirocco -6.777e-03 3.031e-02  
## modelVW- Sharan 1.153e-01 4.613e-02  
## modelVW- Shuttle 1.435e-01 5.405e-02  
## modelVW- T-Cross 7.364e-03 3.042e-02  
## modelVW- T-Roc 9.785e-02 2.104e-02  
## modelVW- Tiguan 1.086e-01 1.708e-02  
## modelVW- Tiguan Allspace 8.117e-02 1.180e-01  
## modelVW- Touareg 2.027e-01 5.171e-02  
## modelVW- Touran 1.282e-01 3.739e-02  
## modelVW- Up -5.352e-01 1.763e-02  
## transmissionf.Trans-SemiAuto 7.424e-02 2.513e-02  
## transmissionf.Trans-Automatic 1.089e-01 4.357e-02  
## modelAudi- A3:transmissionf.Trans-SemiAuto 4.829e-02 3.202e-02  
## modelAudi- A4:transmissionf.Trans-SemiAuto 3.371e-02 3.522e-02  
## modelAudi- A5:transmissionf.Trans-SemiAuto -2.141e-02 4.515e-02  
## modelAudi- A6:transmissionf.Trans-SemiAuto 1.068e-02 4.448e-02  
## modelAudi- A7:transmissionf.Trans-SemiAuto -3.762e-03 8.900e-02  
## modelAudi- A8:transmissionf.Trans-SemiAuto -2.598e-02 8.310e-02  
## modelAudi- Q2:transmissionf.Trans-SemiAuto 2.637e-03 3.984e-02  
## modelAudi- Q3:transmissionf.Trans-SemiAuto 3.229e-02 3.281e-02  
## modelAudi- Q5:transmissionf.Trans-SemiAuto 3.178e-02 5.076e-02  
## modelAudi- Q7:transmissionf.Trans-SemiAuto 1.340e-01 6.331e-02  
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- RS6:transmissionf.Trans-SemiAuto -6.774e-02 1.263e-01  
## modelAudi- S3:transmissionf.Trans-SemiAuto 4.043e-02 1.513e-01  
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA  
## modelAudi- TT:transmissionf.Trans-SemiAuto 8.525e-02 5.426e-02  
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 2.277e-02 3.072e-02  
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 3.222e-02 3.561e-02  
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 7.975e-02 3.222e-02  
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -2.359e-03 5.694e-02  
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 5.192e-02 5.237e-02  
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 3.569e-02 7.374e-02  
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.405e-01 8.575e-02  
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- M2:transmissionf.Trans-SemiAuto 6.067e-04 1.724e-01  
## modelBMW- M3:transmissionf.Trans-SemiAuto 5.229e-02 1.458e-01  
## modelBMW- M4:transmissionf.Trans-SemiAuto 2.580e-02 7.389e-02  
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- X1:transmissionf.Trans-SemiAuto 1.301e-01 4.495e-02  
## modelBMW- X2:transmissionf.Trans-SemiAuto 7.184e-02 1.240e-01  
## modelBMW- X3:transmissionf.Trans-SemiAuto 5.335e-02 5.814e-02  
## modelBMW- X4:transmissionf.Trans-SemiAuto 2.124e-02 7.117e-02  
## modelBMW- X5:transmissionf.Trans-SemiAuto 6.281e-02 6.409e-02  
## modelBMW- X6:transmissionf.Trans-SemiAuto -8.509e-02 1.392e-01  
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA  
## modelBMW- Z4:transmissionf.Trans-SemiAuto 5.245e-02 1.087e-01  
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 3.664e-02 3.077e-02  
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 1.407e-01 5.918e-02  
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 1.883e-01 3.982e-02  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -2.658e-03 4.738e-02  
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.414e-01 7.614e-02  
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 9.368e-02 5.026e-02  
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 3.978e-01 1.381e-01  
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 7.069e-02 4.771e-02  
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 5.345e-02 5.289e-02  
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 8.119e-02 5.755e-02  
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.024e-01 8.574e-02  
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.222e-01 7.655e-02  
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.596e-01 1.233e-01  
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 9.091e-02 8.909e-02  
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA  
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Arteon:transmissionf.Trans-SemiAuto -6.874e-03 9.248e-02  
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 6.223e-02 1.458e-01  
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 1.996e-02 8.233e-02  
## modelVW- CC:transmissionf.Trans-SemiAuto -8.502e-02 8.681e-02  
## modelVW- Golf:transmissionf.Trans-SemiAuto 3.423e-02 2.762e-02  
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 1.975e-02 6.208e-02  
## modelVW- Passat:transmissionf.Trans-SemiAuto 1.112e-01 3.657e-02  
## modelVW- Polo:transmissionf.Trans-SemiAuto 7.084e-02 3.181e-02  
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -1.948e-02 7.735e-02  
## modelVW- Sharan:transmissionf.Trans-SemiAuto -5.846e-03 6.099e-02  
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -3.754e-02 8.243e-02  
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA  
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 1.973e-02 4.636e-02  
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 6.323e-02 3.060e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.650e-01 1.335e-01  
## modelVW- Touareg:transmissionf.Trans-SemiAuto 4.959e-02 6.015e-02  
## modelVW- Touran:transmissionf.Trans-SemiAuto 3.502e-02 5.693e-02  
## modelVW- Up:transmissionf.Trans-SemiAuto 7.054e-02 8.749e-02  
## modelAudi- A3:transmissionf.Trans-Automatic -1.146e-02 4.900e-02  
## modelAudi- A4:transmissionf.Trans-Automatic -1.187e-03 4.914e-02  
## modelAudi- A5:transmissionf.Trans-Automatic -2.791e-02 5.862e-02  
## modelAudi- A6:transmissionf.Trans-Automatic -1.050e-03 5.857e-02  
## modelAudi- A7:transmissionf.Trans-Automatic NA NA  
## modelAudi- A8:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q2:transmissionf.Trans-Automatic -4.493e-02 6.105e-02  
## modelAudi- Q3:transmissionf.Trans-Automatic -3.080e-02 5.179e-02  
## modelAudi- Q5:transmissionf.Trans-Automatic -1.536e-02 6.307e-02  
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA  
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA  
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA  
## modelAudi- S3:transmissionf.Trans-Automatic NA NA  
## modelAudi- S4:transmissionf.Trans-Automatic NA NA  
## modelAudi- S8:transmissionf.Trans-Automatic NA NA  
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA  
## modelAudi- TT:transmissionf.Trans-Automatic -4.232e-02 7.364e-02  
## modelBMW- 1 Series:transmissionf.Trans-Automatic 2.033e-02 4.905e-02  
## modelBMW- 2 Series:transmissionf.Trans-Automatic -2.410e-02 5.221e-02  
## modelBMW- 3 Series:transmissionf.Trans-Automatic 5.166e-02 4.888e-02  
## modelBMW- 4 Series:transmissionf.Trans-Automatic -2.559e-02 6.753e-02  
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA  
## modelBMW- i3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M2:transmissionf.Trans-Automatic NA NA  
## modelBMW- M3:transmissionf.Trans-Automatic NA NA  
## modelBMW- M4:transmissionf.Trans-Automatic NA NA  
## modelBMW- M6:transmissionf.Trans-Automatic NA NA  
## modelBMW- X1:transmissionf.Trans-Automatic 7.707e-02 5.787e-02  
## modelBMW- X2:transmissionf.Trans-Automatic -6.433e-02 1.311e-01  
## modelBMW- X3:transmissionf.Trans-Automatic NA NA  
## modelBMW- X4:transmissionf.Trans-Automatic NA NA  
## modelBMW- X5:transmissionf.Trans-Automatic NA NA  
## modelBMW- X6:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA  
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA  
## modelMercedes- A Class:transmissionf.Trans-Automatic 4.872e-03 4.732e-02  
## modelMercedes- B Class:transmissionf.Trans-Automatic 7.016e-02 7.018e-02  
## modelMercedes- C Class:transmissionf.Trans-Automatic 1.853e-01 5.361e-02  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 1.379e-02 6.527e-02  
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 2.060e-01 1.251e-01  
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.684e-01 8.777e-02  
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 1.037e-01 6.194e-02  
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA  
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 2.152e-01 1.318e-01  
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA  
## modelMercedes- V Class:transmissionf.Trans-Automatic 4.002e-01 7.154e-02  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.785e-01 1.289e-01  
## modelVW- Amarok:transmissionf.Trans-Automatic -1.403e-01 9.394e-02  
## modelVW- Arteon:transmissionf.Trans-Automatic -2.461e-02 1.036e-01  
## modelVW- Beetle:transmissionf.Trans-Automatic 1.210e-01 1.318e-01  
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA  
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -9.281e-02 1.715e-01  
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA  
## modelVW- CC:transmissionf.Trans-Automatic NA NA  
## modelVW- Golf:transmissionf.Trans-Automatic 4.211e-02 4.586e-02  
## modelVW- Golf SV:transmissionf.Trans-Automatic -4.524e-03 8.266e-02  
## modelVW- Passat:transmissionf.Trans-Automatic 1.001e-01 5.666e-02  
## modelVW- Polo:transmissionf.Trans-Automatic 2.321e-02 5.223e-02  
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.202e-01 7.788e-02  
## modelVW- Sharan:transmissionf.Trans-Automatic -5.998e-02 8.525e-02  
## modelVW- Shuttle:transmissionf.Trans-Automatic -8.787e-02 1.356e-01  
## modelVW- T-Cross:transmissionf.Trans-Automatic -2.617e-02 6.349e-02  
## modelVW- T-Roc:transmissionf.Trans-Automatic -5.005e-02 5.955e-02  
## modelVW- Tiguan:transmissionf.Trans-Automatic 2.747e-02 5.219e-02  
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 8.229e-03 1.715e-01  
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA  
## modelVW- Touran:transmissionf.Trans-Automatic -1.599e-01 1.000e-01  
## modelVW- Up:transmissionf.Trans-Automatic 2.061e-01 1.258e-01  
## mileage:transmissionf.Trans-SemiAuto -5.785e-07 2.418e-07  
## mileage:transmissionf.Trans-Automatic -1.196e-06 2.372e-07  
## t value Pr(>|t|)   
## (Intercept) -49.282 < 2e-16 \*\*\*  
## mileage -24.182 < 2e-16 \*\*\*  
## year 52.152 < 2e-16 \*\*\*  
## engineSize 34.759 < 2e-16 \*\*\*  
## mpg -25.690 < 2e-16 \*\*\*  
## modelAudi- A3 6.106 1.11e-09 \*\*\*  
## modelAudi- A4 6.129 9.60e-10 \*\*\*  
## modelAudi- A5 6.643 3.43e-11 \*\*\*  
## modelAudi- A6 7.363 2.13e-13 \*\*\*  
## modelAudi- A7 3.317 0.000918 \*\*\*  
## modelAudi- A8 5.825 6.09e-09 \*\*\*  
## modelAudi- Q2 7.169 8.74e-13 \*\*\*  
## modelAudi- Q3 11.484 < 2e-16 \*\*\*  
## modelAudi- Q5 7.492 8.12e-14 \*\*\*  
## modelAudi- Q7 7.592 3.80e-14 \*\*\*  
## modelAudi- Q8 6.835 9.25e-12 \*\*\*  
## modelAudi- RS3 4.458 8.46e-06 \*\*\*  
## modelAudi- RS5 4.666 3.16e-06 \*\*\*  
## modelAudi- RS6 8.334 < 2e-16 \*\*\*  
## modelAudi- S3 2.713 0.006701 \*\*   
## modelAudi- S4 2.601 0.009331 \*\*   
## modelAudi- S8 4.050 5.21e-05 \*\*\*  
## modelAudi- SQ5 4.679 2.97e-06 \*\*\*  
## modelAudi- TT 4.860 1.21e-06 \*\*\*  
## modelBMW- 1 Series -1.927 0.054059 .   
## modelBMW- 2 Series 0.297 0.766813   
## modelBMW- 3 Series 1.963 0.049668 \*   
## modelBMW- 4 Series 2.592 0.009573 \*\*   
## modelBMW- 5 Series 3.614 0.000304 \*\*\*  
## modelBMW- 6 Series 2.927 0.003437 \*\*   
## modelBMW- 7 Series 4.739 2.21e-06 \*\*\*  
## modelBMW- 8 Series 5.542 3.16e-08 \*\*\*  
## modelBMW- i3 3.331 0.000872 \*\*\*  
## modelBMW- M2 2.640 0.008330 \*\*   
## modelBMW- M3 3.496 0.000478 \*\*\*  
## modelBMW- M4 5.439 5.65e-08 \*\*\*  
## modelBMW- M6 5.943 3.01e-09 \*\*\*  
## modelBMW- X1 1.169 0.242411   
## modelBMW- X2 1.131 0.258287   
## modelBMW- X3 5.600 2.27e-08 \*\*\*  
## modelBMW- X4 4.872 1.14e-06 \*\*\*  
## modelBMW- X5 9.313 < 2e-16 \*\*\*  
## modelBMW- X6 4.554 5.39e-06 \*\*\*  
## modelBMW- Z3 -8.844 < 2e-16 \*\*\*  
## modelBMW- Z4 1.368 0.171509   
## modelMercedes- A Class 5.047 4.66e-07 \*\*\*  
## modelMercedes- B Class -1.167 0.243395   
## modelMercedes- C Class 0.954 0.339994   
## modelMercedes- CL Class 5.918 3.49e-09 \*\*\*  
## modelMercedes- CLA Class 1.288 0.197690   
## modelMercedes- CLS Class 2.908 0.003657 \*\*   
## modelMercedes- E Class 4.681 2.94e-06 \*\*\*  
## modelMercedes- GL Class 1.933 0.053253 .   
## modelMercedes- GLA Class 2.003 0.045184 \*   
## modelMercedes- GLB Class 2.750 0.005991 \*\*   
## modelMercedes- GLC Class 7.339 2.53e-13 \*\*\*  
## modelMercedes- GLE Class 8.781 < 2e-16 \*\*\*  
## modelMercedes- GLS Class 8.381 < 2e-16 \*\*\*  
## modelMercedes- M Class 7.332 2.67e-13 \*\*\*  
## modelMercedes- S Class 9.704 < 2e-16 \*\*\*  
## modelMercedes- SL CLASS 1.018 0.308844   
## modelMercedes- SLK 0.221 0.825245   
## modelMercedes- V Class 1.339 0.180741   
## modelMercedes- X-CLASS -1.090 0.275688   
## modelVW- Amarok 3.099 0.001954 \*\*   
## modelVW- Arteon 1.913 0.055763 .   
## modelVW- Beetle -3.451 0.000564 \*\*\*  
## modelVW- Caddy -0.536 0.592028   
## modelVW- Caddy Maxi -0.424 0.671729   
## modelVW- Caddy Maxi Life -1.047 0.294993   
## modelVW- Caravelle 7.581 4.12e-14 \*\*\*  
## modelVW- CC -2.129 0.033278 \*   
## modelVW- Golf -1.510 0.131188   
## modelVW- Golf SV -3.163 0.001574 \*\*   
## modelVW- Passat -4.184 2.92e-05 \*\*\*  
## modelVW- Polo -16.753 < 2e-16 \*\*\*  
## modelVW- Scirocco -0.224 0.823093   
## modelVW- Sharan 2.499 0.012497 \*   
## modelVW- Shuttle 2.655 0.007963 \*\*   
## modelVW- T-Cross 0.242 0.808740   
## modelVW- T-Roc 4.650 3.42e-06 \*\*\*  
## modelVW- Tiguan 6.356 2.28e-10 \*\*\*  
## modelVW- Tiguan Allspace 0.688 0.491743   
## modelVW- Touareg 3.920 8.98e-05 \*\*\*  
## modelVW- Touran 3.429 0.000612 \*\*\*  
## modelVW- Up -30.356 < 2e-16 \*\*\*  
## transmissionf.Trans-SemiAuto 2.955 0.003145 \*\*   
## transmissionf.Trans-Automatic 2.499 0.012493 \*   
## modelAudi- A3:transmissionf.Trans-SemiAuto 1.508 0.131628   
## modelAudi- A4:transmissionf.Trans-SemiAuto 0.957 0.338521   
## modelAudi- A5:transmissionf.Trans-SemiAuto -0.474 0.635339   
## modelAudi- A6:transmissionf.Trans-SemiAuto 0.240 0.810309   
## modelAudi- A7:transmissionf.Trans-SemiAuto -0.042 0.966288   
## modelAudi- A8:transmissionf.Trans-SemiAuto -0.313 0.754617   
## modelAudi- Q2:transmissionf.Trans-SemiAuto 0.066 0.947222   
## modelAudi- Q3:transmissionf.Trans-SemiAuto 0.984 0.325088   
## modelAudi- Q5:transmissionf.Trans-SemiAuto 0.626 0.531301   
## modelAudi- Q7:transmissionf.Trans-SemiAuto 2.116 0.034371 \*   
## modelAudi- Q8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS3:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- RS6:transmissionf.Trans-SemiAuto -0.536 0.591810   
## modelAudi- S3:transmissionf.Trans-SemiAuto 0.267 0.789307   
## modelAudi- S4:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- S8:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- SQ5:transmissionf.Trans-SemiAuto NA NA   
## modelAudi- TT:transmissionf.Trans-SemiAuto 1.571 0.116226   
## modelBMW- 1 Series:transmissionf.Trans-SemiAuto 0.741 0.458622   
## modelBMW- 2 Series:transmissionf.Trans-SemiAuto 0.905 0.365738   
## modelBMW- 3 Series:transmissionf.Trans-SemiAuto 2.475 0.013346 \*   
## modelBMW- 4 Series:transmissionf.Trans-SemiAuto -0.041 0.966955   
## modelBMW- 5 Series:transmissionf.Trans-SemiAuto 0.991 0.321553   
## modelBMW- 6 Series:transmissionf.Trans-SemiAuto 0.484 0.628459   
## modelBMW- 7 Series:transmissionf.Trans-SemiAuto 1.639 0.101317   
## modelBMW- 8 Series:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- i3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- M2:transmissionf.Trans-SemiAuto 0.004 0.997192   
## modelBMW- M3:transmissionf.Trans-SemiAuto 0.359 0.719897   
## modelBMW- M4:transmissionf.Trans-SemiAuto 0.349 0.726988   
## modelBMW- M6:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- X1:transmissionf.Trans-SemiAuto 2.895 0.003809 \*\*   
## modelBMW- X2:transmissionf.Trans-SemiAuto 0.579 0.562320   
## modelBMW- X3:transmissionf.Trans-SemiAuto 0.918 0.358876   
## modelBMW- X4:transmissionf.Trans-SemiAuto 0.298 0.765376   
## modelBMW- X5:transmissionf.Trans-SemiAuto 0.980 0.327112   
## modelBMW- X6:transmissionf.Trans-SemiAuto -0.611 0.541176   
## modelBMW- Z3:transmissionf.Trans-SemiAuto NA NA   
## modelBMW- Z4:transmissionf.Trans-SemiAuto 0.482 0.629588   
## modelMercedes- A Class:transmissionf.Trans-SemiAuto 1.191 0.233723   
## modelMercedes- B Class:transmissionf.Trans-SemiAuto 2.378 0.017459 \*   
## modelMercedes- C Class:transmissionf.Trans-SemiAuto 4.728 2.33e-06 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-SemiAuto -0.056 0.955278   
## modelMercedes- CLA Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- CLS Class:transmissionf.Trans-SemiAuto 1.857 0.063373 .   
## modelMercedes- E Class:transmissionf.Trans-SemiAuto 1.864 0.062416 .   
## modelMercedes- GL Class:transmissionf.Trans-SemiAuto 2.880 0.003992 \*\*   
## modelMercedes- GLA Class:transmissionf.Trans-SemiAuto 1.481 0.138543   
## modelMercedes- GLB Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-SemiAuto 1.010 0.312317   
## modelMercedes- GLE Class:transmissionf.Trans-SemiAuto 1.411 0.158405   
## modelMercedes- GLS Class:transmissionf.Trans-SemiAuto 1.194 0.232372   
## modelMercedes- M Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- S Class:transmissionf.Trans-SemiAuto -1.596 0.110524   
## modelMercedes- SL CLASS:transmissionf.Trans-SemiAuto 1.295 0.195306   
## modelMercedes- SLK:transmissionf.Trans-SemiAuto 1.020 0.307591   
## modelMercedes- V Class:transmissionf.Trans-SemiAuto NA NA   
## modelMercedes- X-CLASS:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Amarok:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Arteon:transmissionf.Trans-SemiAuto -0.074 0.940753   
## modelVW- Beetle:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-SemiAuto NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-SemiAuto 0.427 0.669524   
## modelVW- Caravelle:transmissionf.Trans-SemiAuto 0.242 0.808476   
## modelVW- CC:transmissionf.Trans-SemiAuto -0.979 0.327471   
## modelVW- Golf:transmissionf.Trans-SemiAuto 1.239 0.215278   
## modelVW- Golf SV:transmissionf.Trans-SemiAuto 0.318 0.750440   
## modelVW- Passat:transmissionf.Trans-SemiAuto 3.041 0.002374 \*\*   
## modelVW- Polo:transmissionf.Trans-SemiAuto 2.227 0.025991 \*   
## modelVW- Scirocco:transmissionf.Trans-SemiAuto -0.252 0.801135   
## modelVW- Sharan:transmissionf.Trans-SemiAuto -0.096 0.923645   
## modelVW- Shuttle:transmissionf.Trans-SemiAuto -0.455 0.648803   
## modelVW- T-Cross:transmissionf.Trans-SemiAuto NA NA   
## modelVW- T-Roc:transmissionf.Trans-SemiAuto 0.426 0.670434   
## modelVW- Tiguan:transmissionf.Trans-SemiAuto 2.066 0.038874 \*   
## modelVW- Tiguan Allspace:transmissionf.Trans-SemiAuto 1.236 0.216426   
## modelVW- Touareg:transmissionf.Trans-SemiAuto 0.825 0.409687   
## modelVW- Touran:transmissionf.Trans-SemiAuto 0.615 0.538438   
## modelVW- Up:transmissionf.Trans-SemiAuto 0.806 0.420122   
## modelAudi- A3:transmissionf.Trans-Automatic -0.234 0.815043   
## modelAudi- A4:transmissionf.Trans-Automatic -0.024 0.980732   
## modelAudi- A5:transmissionf.Trans-Automatic -0.476 0.634043   
## modelAudi- A6:transmissionf.Trans-Automatic -0.018 0.985690   
## modelAudi- A7:transmissionf.Trans-Automatic NA NA   
## modelAudi- A8:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q2:transmissionf.Trans-Automatic -0.736 0.461824   
## modelAudi- Q3:transmissionf.Trans-Automatic -0.595 0.552121   
## modelAudi- Q5:transmissionf.Trans-Automatic -0.244 0.807601   
## modelAudi- Q7:transmissionf.Trans-Automatic NA NA   
## modelAudi- Q8:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS3:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS5:transmissionf.Trans-Automatic NA NA   
## modelAudi- RS6:transmissionf.Trans-Automatic NA NA   
## modelAudi- S3:transmissionf.Trans-Automatic NA NA   
## modelAudi- S4:transmissionf.Trans-Automatic NA NA   
## modelAudi- S8:transmissionf.Trans-Automatic NA NA   
## modelAudi- SQ5:transmissionf.Trans-Automatic NA NA   
## modelAudi- TT:transmissionf.Trans-Automatic -0.575 0.565499   
## modelBMW- 1 Series:transmissionf.Trans-Automatic 0.414 0.678537   
## modelBMW- 2 Series:transmissionf.Trans-Automatic -0.462 0.644455   
## modelBMW- 3 Series:transmissionf.Trans-Automatic 1.057 0.290596   
## modelBMW- 4 Series:transmissionf.Trans-Automatic -0.379 0.704729   
## modelBMW- 5 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 6 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 7 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- 8 Series:transmissionf.Trans-Automatic NA NA   
## modelBMW- i3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M2:transmissionf.Trans-Automatic NA NA   
## modelBMW- M3:transmissionf.Trans-Automatic NA NA   
## modelBMW- M4:transmissionf.Trans-Automatic NA NA   
## modelBMW- M6:transmissionf.Trans-Automatic NA NA   
## modelBMW- X1:transmissionf.Trans-Automatic 1.332 0.182990   
## modelBMW- X2:transmissionf.Trans-Automatic -0.491 0.623653   
## modelBMW- X3:transmissionf.Trans-Automatic NA NA   
## modelBMW- X4:transmissionf.Trans-Automatic NA NA   
## modelBMW- X5:transmissionf.Trans-Automatic NA NA   
## modelBMW- X6:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z3:transmissionf.Trans-Automatic NA NA   
## modelBMW- Z4:transmissionf.Trans-Automatic NA NA   
## modelMercedes- A Class:transmissionf.Trans-Automatic 0.103 0.917998   
## modelMercedes- B Class:transmissionf.Trans-Automatic 1.000 0.317500   
## modelMercedes- C Class:transmissionf.Trans-Automatic 3.456 0.000553 \*\*\*  
## modelMercedes- CL Class:transmissionf.Trans-Automatic 0.211 0.832704   
## modelMercedes- CLA Class:transmissionf.Trans-Automatic 1.647 0.099540 .   
## modelMercedes- CLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- E Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GL Class:transmissionf.Trans-Automatic 1.919 0.055103 .   
## modelMercedes- GLA Class:transmissionf.Trans-Automatic 1.674 0.094229 .   
## modelMercedes- GLB Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLC Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLE Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- GLS Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- M Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- S Class:transmissionf.Trans-Automatic NA NA   
## modelMercedes- SL CLASS:transmissionf.Trans-Automatic 1.633 0.102630   
## modelMercedes- SLK:transmissionf.Trans-Automatic NA NA   
## modelMercedes- V Class:transmissionf.Trans-Automatic 5.594 2.34e-08 \*\*\*  
## modelMercedes- X-CLASS:transmissionf.Trans-Automatic 1.384 0.166298   
## modelVW- Amarok:transmissionf.Trans-Automatic -1.493 0.135419   
## modelVW- Arteon:transmissionf.Trans-Automatic -0.238 0.812224   
## modelVW- Beetle:transmissionf.Trans-Automatic 0.918 0.358602   
## modelVW- Caddy:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi:transmissionf.Trans-Automatic NA NA   
## modelVW- Caddy Maxi Life:transmissionf.Trans-Automatic -0.541 0.588357   
## modelVW- Caravelle:transmissionf.Trans-Automatic NA NA   
## modelVW- CC:transmissionf.Trans-Automatic NA NA   
## modelVW- Golf:transmissionf.Trans-Automatic 0.918 0.358491   
## modelVW- Golf SV:transmissionf.Trans-Automatic -0.055 0.956358   
## modelVW- Passat:transmissionf.Trans-Automatic 1.766 0.077458 .   
## modelVW- Polo:transmissionf.Trans-Automatic 0.444 0.656844   
## modelVW- Scirocco:transmissionf.Trans-Automatic -1.544 0.122738   
## modelVW- Sharan:transmissionf.Trans-Automatic -0.704 0.481754   
## modelVW- Shuttle:transmissionf.Trans-Automatic -0.648 0.516886   
## modelVW- T-Cross:transmissionf.Trans-Automatic -0.412 0.680191   
## modelVW- T-Roc:transmissionf.Trans-Automatic -0.840 0.400702   
## modelVW- Tiguan:transmissionf.Trans-Automatic 0.526 0.598701   
## modelVW- Tiguan Allspace:transmissionf.Trans-Automatic 0.048 0.961725   
## modelVW- Touareg:transmissionf.Trans-Automatic NA NA   
## modelVW- Touran:transmissionf.Trans-Automatic -1.598 0.110091   
## modelVW- Up:transmissionf.Trans-Automatic 1.638 0.101539   
## mileage:transmissionf.Trans-SemiAuto -2.393 0.016770 \*   
## mileage:transmissionf.Trans-Automatic -5.042 4.79e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1173 on 4576 degrees of freedom  
## Multiple R-squared: 0.932, Adjusted R-squared: 0.9292   
## F-statistic: 328.3 on 191 and 4576 DF, p-value: < 2.2e-16

* The residuals of the model are relatively small, ranging from -0.60962 to 0.44434, with a median very close to zero, indicating that the model’s predictions are generally accurate.
* Many coefficients are not defined due to singularities, which often happens when there are categories with very low/null occurrence or when there’s multicollinearity due to the interaction terms.
* The residual standard error is 0.1235 on 4654 degrees of freedom, which is quite low, indicating a good fit. The model explains a substantial amount of variance, with a Multiple R-squared of 93% , suggesting a strong predictive power. The F-statistic is 336.9 with a p-value of less than 2.2e-16, which is indicative of a highly significant model.

par(mfrow=c(2,2))   
plot(m9,id.n=0)



* The diagnostic plots for model m9 indicate a robust linear regression model. The Residuals vs Fitted plot shows a random distribution of points around the horizontal line, suggesting that the model’s assumptions of linearity and homoscedasticity are met. The Q-Q Plot supports the assumption of normally distributed residualss. The Scale-Location plot’s uniform spread indicates stable variance across predictions, reinforcing the model’s homoscedastic nature. Lastly, the Residuals vs Leverage plot reveals no points with high leverage or significant Cook’s distances, pointing to an absence of influential outliers. Collectively, these plots suggest that model m9 is accepted and provides a good fit to the data.

anova(m9)

## Analysis of Variance Table  
##   
## Response: log(price)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mileage 1 327.71 327.71 23819.3133 < 2.2e-16 \*\*\*  
## year 1 84.78 84.78 6162.4824 < 2.2e-16 \*\*\*  
## engineSize 1 322.77 322.77 23460.8047 < 2.2e-16 \*\*\*  
## mpg 1 21.87 21.87 1589.3608 < 2.2e-16 \*\*\*  
## model 81 95.51 1.18 85.7031 < 2.2e-16 \*\*\*  
## transmission 2 6.76 3.38 245.7789 < 2.2e-16 \*\*\*  
## model:transmission 102 3.02 0.03 2.1556 2.630e-10 \*\*\*  
## mileage:transmission 2 0.35 0.18 12.7472 3.015e-06 \*\*\*  
## Residuals 4576 62.96 0.01   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* The ANOVA for model m9 shows that all predictors, including mileage, year, engine size, mpg, model, transmission, and their interactions, are highly significant with p-values less than 0.05.
* Let’s review and compare models m8 and m9 to decide which model is better suited for our analysis and then choose the one to move forward with. As we can see the best one is m9.

AIC(m9,m8)

## df AIC  
## m9 193 -6715.280  
## m8 193 -5629.022

# *11. Prediction model for binary target “Audi”:*

* *In the subsequent section of the assignment, our focus shifts to constructing a predictive model for the binary variable ‘Audi.’ The goal is to develop a model that enables us to estimate the likelihood of a given set of input data being associated with an Audi car or not.*

set.seed(1998)  
  
#df<-intial\_data[-mouts,]  
x <- sample(1:nrow(df),round(0.70\*nrow(df),0))  
  
train <- df[x,]  
test <-df[-x,]

* *Based on the results based on the following analysis using the catdes function, the variables mpg, price, tax, and year exhibit statistically significant relationships with the target variable. Therefore, these variables may be considered as potential predictors for the initial binary classification model.*

res.cat <- catdes(df, num.var = which(names(df)=="Audi"))  
res.cat$quanti.var

## Eta2 P-value  
## mpg 0.009607319 1.042792e-11  
## price\_transformed 0.005001918 9.525018e-07  
## price 0.003969156 1.271992e-05  
## tax 0.001288782 1.293564e-02  
## year 0.001148705 1.895097e-02  
## mileage 0.001016666 2.728291e-02

## *11.1 Initial model*

* *Based on the MCA and previous results, we will be proceeding to choose the suitable variables and built the initial model:*

b1<-glm(Audi~mpg+tax+year,family="binomial",data=train)  
summary(b1)

##   
## Call:  
## glm(formula = Audi ~ mpg + tax + year, family = "binomial", data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 253.009312 52.420440 4.827 1.39e-06 \*\*\*  
## mpg -0.030312 0.004522 -6.703 2.04e-11 \*\*\*  
## tax 0.003156 0.003651 0.864 0.387   
## year -0.125540 0.025863 -4.854 1.21e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3345.7 on 3351 degrees of freedom  
## AIC: 3353.7  
##   
## Number of Fisher Scoring iterations: 4

* *The model (b1) examines how mpg, tax, and year affect the likelihood of the outcome variable “Audi.” The results show that mpg and year significantly influence the odds of the event, with higher mpg and later years associated with certain outcomes. However, the tax variable doesn’t have a significant impact. The model fits the data reasonably well, as indicated by the deviance values, and the AIC is 3504.9, suggesting a decent overall model quality.*
* *As these VIF values are all close to 1, suggesting that there is no severe multicollinearity among the predictor variables in the model.*

vif(b1)

## mpg tax year   
## 1.315976 1.125198 1.281231

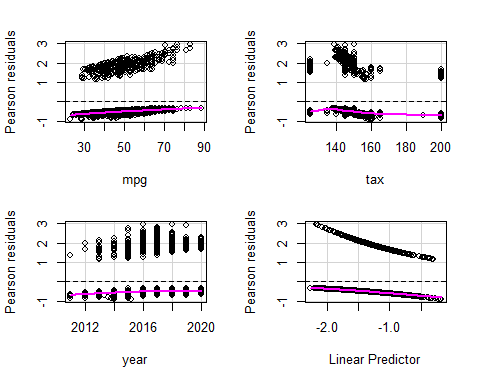
* *Again, mpg and year are valuable predictors in this model, while tax does not appear to play a statistically significant role.*

Anova(b1)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: Audi  
## LR Chisq Df Pr(>Chisq)   
## mpg 45.942 1 1.218e-11 \*\*\*  
## tax 0.741 1 0.3893   
## year 23.186 1 1.471e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* *At this stage, the residual plots show that the variance of the residuals is not constant, which violates the assumption of homoscedasticity.*

residualPlots(b1)



## Test stat Pr(>|Test stat|)   
## mpg 4.7341 0.02957 \*  
## tax 0.0524 0.81890   
## year 1.4214 0.23318   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* *As tax didn’t show much influence, we could try to remove it and built a simpler model:*

b2<-glm(Audi~mpg+year,family="binomial"(link = logit),data=train)  
summary(b2)

##   
## Call:  
## glm(formula = Audi ~ mpg + year, family = binomial(link = logit),   
## data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 264.939814 50.499366 5.246 1.55e-07 \*\*\*  
## mpg -0.031483 0.004316 -7.294 3.01e-13 \*\*\*  
## year -0.131194 0.024989 -5.250 1.52e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3346.4 on 3352 degrees of freedom  
## AIC: 3352.4  
##   
## Number of Fisher Scoring iterations: 4

* *As we can see the AIC value didn’t change much.*

AIC(b1,b2)

## df AIC  
## b1 4 3353.704  
## b2 3 3352.445

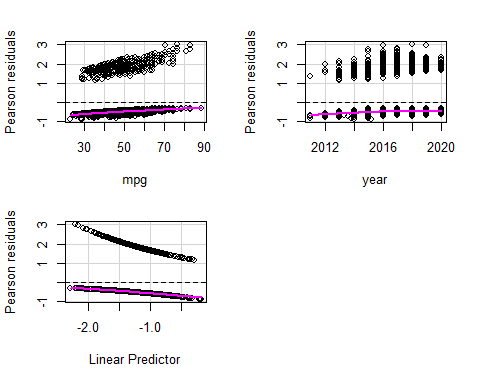
* *While examining the AIC values, it is evident that the inclusion of the variable ‘tax has not substantially altered the model. The marginal change in the AIC suggests that this variable may not play a significant role in shaping the model. It appears that ’tax’ may be perturbing the model without significantly enhancing its overall performance.*

anova(b1,b2)

## Analysis of Deviance Table  
##   
## Model 1: Audi ~ mpg + tax + year  
## Model 2: Audi ~ mpg + year  
## Resid. Df Resid. Dev Df Deviance  
## 1 3351 3345.7   
## 2 3352 3346.4 -1 -0.74106

* *The residual plots further support the decision to exclude the ‘tax’ variable from the model. Without ‘tax,’ the residual graphs exhibit improved characteristics, demonstrating reduced dispersion and less heteroscedastic behavior. This shows better improvement compared to the previous one. The residual plots show a weaker non-linear relationship between the residuals and the fitted values, and the variance of the residuals appears to be more constant.*

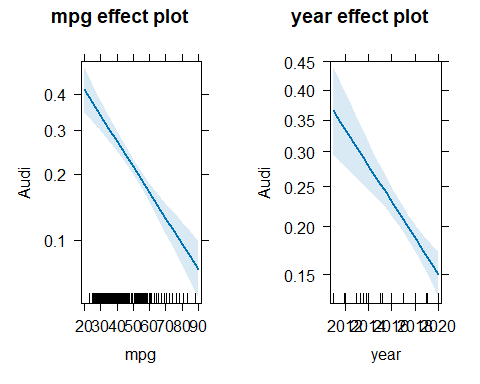
residualPlots(b2)



## Test stat Pr(>|Test stat|)   
## mpg 4.5938 0.03209 \*  
## year 1.6325 0.20136   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

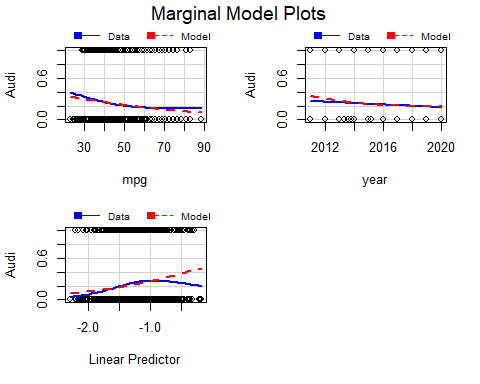
* *How this model works?*
  + *The following plots illustrate the probability of a car being an Audi based on its mpg and year values. Notably, a discernible trend emerges: as mpg increases, the likelihood of the car being an Audi diminishes. Additionally, the passage of time (More recent years) is associated with a decreasing probability of the car being identified as an Audi.*

plot(allEffects(b2))



* *Even though we will chose the second model, some marginal model plots for b2 reveal a noticeable lack of overlap between the observed data and the model predictions, indicating a potential need for model refinement or consideration of additional factors.*

marginalModelPlots(b2)



## *11.2 Adding factors*

* *Adding new factors based on MCA analysis and the following results maybe enhance our model. We won’t take into consideration f.mpg as it is derivated from mpg so the best candidates are: f.engineSize, fuelType and transmission.*

catdes(df,17)$test.chi2

## p.value df  
## model 0.000000e+00 81  
## manufacturer 0.000000e+00 3  
## hcpckMCA 3.064963e-89 4  
## f.mpg 2.417283e-18 3  
## f.engineSize 1.104807e-17 2  
## fuelType 3.413762e-06 3  
## f.price 5.601760e-05 3  
## transmission 6.462237e-05 2  
## claH 3.082361e-02 4  
## f.miles 3.628667e-02 3  
## claKM 4.424360e-02 4

* *Let’s add them and build a new model:*

b3<-glm(Audi~mpg+year+f.engineSize+fuelType+transmission,family="binomial",data=train)  
summary(b3)

##   
## Call:  
## glm(formula = Audi ~ mpg + year + f.engineSize + fuelType + transmission,   
## family = "binomial", data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 350.276167 54.096701 6.475 9.48e-11 \*\*\*  
## mpg -0.058048 0.006075 -9.555 < 2e-16 \*\*\*  
## year -0.172492 0.026733 -6.452 1.10e-10 \*\*\*  
## f.engineSizeMedium 0.008123 0.137672 0.059 0.952948   
## f.engineSizeLarge -1.192807 0.201952 -5.906 3.50e-09 \*\*\*  
## fuelTypeElectric -12.580888 197.937205 -0.064 0.949321   
## fuelTypeHybrid -1.703225 0.735263 -2.316 0.020532 \*   
## fuelTypePetrol -0.464844 0.134285 -3.462 0.000537 \*\*\*  
## transmissionf.Trans-SemiAuto -0.365683 0.116753 -3.132 0.001736 \*\*   
## transmissionf.Trans-Automatic -0.345846 0.125993 -2.745 0.006052 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3226.9 on 3345 degrees of freedom  
## AIC: 3246.9  
##   
## Number of Fisher Scoring iterations: 12

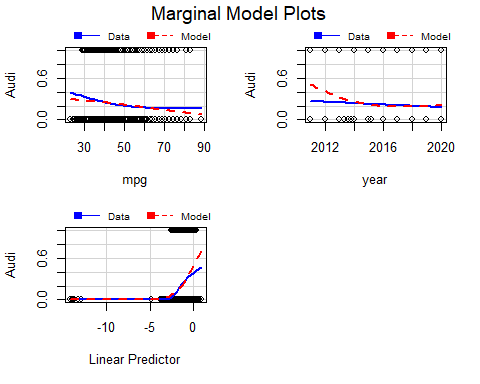
* *The regression model b3 suggests again that cars with lower ‘mpg’ and older ‘year’ are less likely to be Audi. Additionally, factors such as ‘f.engineSizeLarge,’ ‘fuelTypeHybrid,’ ‘transmissionf.Trans-SemiAuto,’ and ‘transmissionf.Trans-Automatic’ influence the likelihood.*
* *The model fits reasonably well, demonstrating a reduction in deviance from the previous model, with an AIC of 3397.3.*
* *No evidence of collinearity is observed among the predictor variables in the model, indicating that they can independently contribute to explaining the variance in the response variable.*

vif(b3)

## GVIF Df GVIF^(1/(2\*Df))  
## mpg 2.198845 1 1.482850  
## year 1.373446 1 1.171941  
## f.engineSize 2.327783 2 1.235195  
## fuelType 2.327265 3 1.151174  
## transmission 1.427650 2 1.093089

* *The marginal model plot of the linear predictor for b3 reveal a noticeable improvement of overlap between the observed data and the model predictions, indicating a potential prigress for model refinement after including new factors.*

marginalModelPlots(b3)



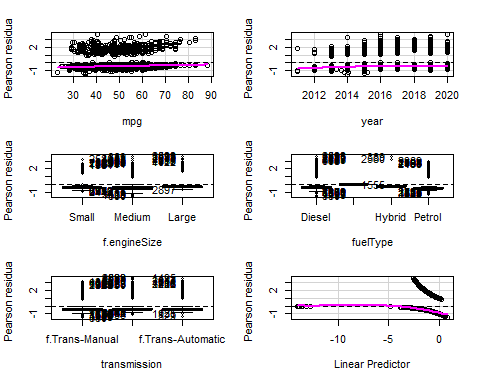
* All variables are significant for the target.

Anova(b3)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: Audi  
## LR Chisq Df Pr(>Chisq)   
## mpg 95.639 1 < 2.2e-16 \*\*\*  
## year 41.222 1 1.359e-10 \*\*\*  
## f.engineSize 81.137 2 < 2.2e-16 \*\*\*  
## fuelType 21.751 3 7.349e-05 \*\*\*  
## transmission 11.340 2 0.003448 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Overall, the residual plots show that the model is not perfectly fitting the data. There are some patterns in the residuals that suggest that the model could be improved.

residualPlots(b3)



## Test stat Pr(>|Test stat|)   
## mpg 13.745 0.0002093 \*\*\*  
## year 12.451 0.0004178 \*\*\*  
## f.engineSize   
## fuelType   
## transmission   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## *11.3 Adding Interactions*

* *We will build a test model with all the possible interactions to see which are the most contribuitng to out model.*
* *Based on AIC values we can see that the best interactions to consider are mpg\*f.engineSize and f.engineSize\*fuelType.*

b\_test<-glm(Audi~(mpg+year+f.engineSize+fuelType+transmission)\*(f.engineSize+fuelType+transmission),family="binomial",data=train)  
summary(b\_test)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + f.engineSize + fuelType +   
## transmission) \* (f.engineSize + fuelType + transmission),   
## family = "binomial", data = train)  
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error z value  
## (Intercept) 5.793e+02 2.620e+03 0.221  
## mpg 4.814e-02 2.663e-02 1.808  
## year -2.996e-01 9.487e-02 -3.158  
## f.engineSizeMedium -4.918e+02 2.620e+03 -0.188  
## f.engineSizeLarge -7.034e+02 5.816e+03 -0.121  
## fuelTypeElectric -2.519e+02 6.937e+07 0.000  
## fuelTypeHybrid 3.555e+05 6.717e+06 0.053  
## fuelTypePetrol -7.138e+01 2.617e+03 -0.027  
## transmissionf.Trans-SemiAuto 5.429e+02 1.466e+02 3.704  
## transmissionf.Trans-Automatic 1.849e+02 1.435e+02 1.289  
## mpg:f.engineSizeMedium -6.392e-02 2.597e-02 -2.461  
## mpg:f.engineSizeLarge -1.254e-01 3.304e-02 -3.795  
## mpg:fuelTypeElectric 4.084e-02 1.654e+03 0.000  
## mpg:fuelTypeHybrid -7.695e+01 1.434e+03 -0.054  
## mpg:fuelTypePetrol -6.096e-02 2.217e-02 -2.750  
## mpg:transmissionf.Trans-SemiAuto -5.553e-02 1.551e-02 -3.581  
## mpg:transmissionf.Trans-Automatic -2.763e-02 1.658e-02 -1.666  
## year:f.engineSizeMedium 2.563e-01 9.279e-02 2.762  
## year:f.engineSizeLarge 3.527e-01 1.287e-01 2.740  
## year:fuelTypeElectric 1.247e-01 3.435e+04 0.000  
## year:fuelTypeHybrid -1.744e+02 3.297e+03 -0.053  
## year:fuelTypePetrol 4.750e-02 7.247e-02 0.655  
## year:transmissionf.Trans-SemiAuto -2.673e-01 7.242e-02 -3.690  
## year:transmissionf.Trans-Automatic -9.088e-02 7.092e-02 -1.282  
## f.engineSizeMedium:fuelTypeElectric -2.102e+01 2.862e+04 -0.001  
## f.engineSizeLarge:fuelTypeElectric NA NA NA  
## f.engineSizeMedium:fuelTypeHybrid -4.168e+01 2.819e+03 -0.015  
## f.engineSizeLarge:fuelTypeHybrid -6.387e+02 1.787e+04 -0.036  
## f.engineSizeMedium:fuelTypePetrol -2.224e+01 2.613e+03 -0.009  
## f.engineSizeLarge:fuelTypePetrol -2.356e+01 2.613e+03 -0.009  
## f.engineSizeMedium:transmissionf.Trans-SemiAuto -1.379e+00 4.026e-01 -3.425  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 1.748e+01 5.190e+03 0.003  
## f.engineSizeMedium:transmissionf.Trans-Automatic -4.545e-01 4.617e-01 -0.984  
## f.engineSizeLarge:transmissionf.Trans-Automatic 1.841e+01 5.190e+03 0.004  
## fuelTypeElectric:transmissionf.Trans-SemiAuto NA NA NA  
## fuelTypeHybrid:transmissionf.Trans-SemiAuto 3.423e+01 5.408e+03 0.006  
## fuelTypePetrol:transmissionf.Trans-SemiAuto -2.233e-01 4.006e-01 -0.557  
## fuelTypeElectric:transmissionf.Trans-Automatic -1.451e+00 7.679e+04 0.000  
## fuelTypeHybrid:transmissionf.Trans-Automatic NA NA NA  
## fuelTypePetrol:transmissionf.Trans-Automatic -2.566e-01 4.311e-01 -0.595  
## Pr(>|z|)   
## (Intercept) 0.825025   
## mpg 0.070582 .   
## year 0.001589 \*\*   
## f.engineSizeMedium 0.851102   
## f.engineSizeLarge 0.903750   
## fuelTypeElectric 0.999997   
## fuelTypeHybrid 0.957798   
## fuelTypePetrol 0.978240   
## transmissionf.Trans-SemiAuto 0.000212 \*\*\*  
## transmissionf.Trans-Automatic 0.197527   
## mpg:f.engineSizeMedium 0.013847 \*   
## mpg:f.engineSizeLarge 0.000148 \*\*\*  
## mpg:fuelTypeElectric 0.999980   
## mpg:fuelTypeHybrid 0.957194   
## mpg:fuelTypePetrol 0.005961 \*\*   
## mpg:transmissionf.Trans-SemiAuto 0.000342 \*\*\*  
## mpg:transmissionf.Trans-Automatic 0.095637 .   
## year:f.engineSizeMedium 0.005745 \*\*   
## year:f.engineSizeLarge 0.006142 \*\*   
## year:fuelTypeElectric 0.999997   
## year:fuelTypeHybrid 0.957806   
## year:fuelTypePetrol 0.512232   
## year:transmissionf.Trans-SemiAuto 0.000224 \*\*\*  
## year:transmissionf.Trans-Automatic 0.200004   
## f.engineSizeMedium:fuelTypeElectric 0.999414   
## f.engineSizeLarge:fuelTypeElectric NA   
## f.engineSizeMedium:fuelTypeHybrid 0.988203   
## f.engineSizeLarge:fuelTypeHybrid 0.971490   
## f.engineSizeMedium:fuelTypePetrol 0.993210   
## f.engineSizeLarge:fuelTypePetrol 0.992806   
## f.engineSizeMedium:transmissionf.Trans-SemiAuto 0.000615 \*\*\*  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 0.997312   
## f.engineSizeMedium:transmissionf.Trans-Automatic 0.324913   
## f.engineSizeLarge:transmissionf.Trans-Automatic 0.997170   
## fuelTypeElectric:transmissionf.Trans-SemiAuto NA   
## fuelTypeHybrid:transmissionf.Trans-SemiAuto 0.994950   
## fuelTypePetrol:transmissionf.Trans-SemiAuto 0.577318   
## fuelTypeElectric:transmissionf.Trans-Automatic 0.999985   
## fuelTypeHybrid:transmissionf.Trans-Automatic NA   
## fuelTypePetrol:transmissionf.Trans-Automatic 0.551688   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3070.8 on 3318 degrees of freedom  
## AIC: 3144.8  
##   
## Number of Fisher Scoring iterations: 20

* *We can see that fuelType does not play a significant role due to to high p-value, even when it interacts with other variables. So we will remove it to shape a new model.*
* *Also, this test model summary indicate coefficients that could not be estimated due to collinearity, something that we will check in further steps.*

### *11.3.1 Interaction between covariates and factors*

* *Let’s build a model without fuelType’s interactions and the strongest possible covariate-factor interactions:*

b4 <- glm(Audi~(mpg+year+ fuelType +f.engineSize+transmission) + (mpg+year)\*(f.engineSize+transmission),family="binomial",data=train)  
summary(b4)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + fuelType + f.engineSize +   
## transmission) + (mpg + year) \* (f.engineSize + transmission),   
## family = "binomial", data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 611.21031 122.69329 4.982 6.31e-07 \*\*\*  
## mpg -0.06212 0.01269 -4.894 9.90e-07 \*\*\*  
## year -0.30174 0.06064 -4.976 6.49e-07 \*\*\*  
## fuelTypeElectric -13.54244 327.69739 -0.041 0.967036   
## fuelTypeHybrid -1.71900 0.74307 -2.313 0.020702 \*   
## fuelTypePetrol -0.49113 0.13593 -3.613 0.000302 \*\*\*  
## f.engineSizeMedium -544.90135 140.05344 -3.891 1.00e-04 \*\*\*  
## f.engineSizeLarge -835.34712 217.85474 -3.834 0.000126 \*\*\*  
## transmissionf.Trans-SemiAuto 494.82748 137.53352 3.598 0.000321 \*\*\*  
## transmissionf.Trans-Automatic 190.23567 138.12785 1.377 0.168437   
## mpg:f.engineSizeMedium 0.05021 0.01331 3.772 0.000162 \*\*\*  
## mpg:f.engineSizeLarge 0.01221 0.01970 0.620 0.535533   
## mpg:transmissionf.Trans-SemiAuto -0.05959 0.01236 -4.820 1.44e-06 \*\*\*  
## mpg:transmissionf.Trans-Automatic -0.03744 0.01279 -2.927 0.003420 \*\*   
## year:f.engineSizeMedium 0.26877 0.06924 3.882 0.000104 \*\*\*  
## year:f.engineSizeLarge 0.41295 0.10775 3.832 0.000127 \*\*\*  
## year:transmissionf.Trans-SemiAuto -0.24388 0.06804 -3.584 0.000338 \*\*\*  
## year:transmissionf.Trans-Automatic -0.09343 0.06834 -1.367 0.171554   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3162.5 on 3337 degrees of freedom  
## AIC: 3198.5  
##   
## Number of Fisher Scoring iterations: 13

* *Let’s check for any possibile collinearity.*
* *As we can see all the interaction have high GVIF values, higher than 5, so it makes it not acceptable and it is an evidence of high multicollineairity.*
* *To solve this we will keep one suitable interaction.*

vif(b4)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

## GVIF Df GVIF^(1/(2\*Df))  
## mpg 1.033768e+01 1 3.215226  
## year 6.782543e+00 1 2.604332  
## fuelType 2.351675e+00 3 1.153178  
## f.engineSize 3.581470e+12 2 1375.673384  
## transmission 2.489790e+12 2 1256.147570  
## mpg:f.engineSize 1.412884e+03 2 6.130934  
## mpg:transmission 9.134198e+02 2 5.497530  
## year:f.engineSize 3.549128e+12 2 1372.557095  
## year:transmission 2.473018e+12 2 1254.026785

b5 <- glm(Audi~(mpg+year+fuelType + f.engineSize+transmission + mpg\*transmission ),family="binomial",data=train)  
summary(b5)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + fuelType + f.engineSize +   
## transmission + mpg \* transmission), family = "binomial",   
## data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 337.880454 54.693640 6.178 6.50e-10 \*\*\*  
## mpg -0.031436 0.008294 -3.790 0.000150 \*\*\*  
## year -0.167068 0.027027 -6.182 6.35e-10 \*\*\*  
## fuelTypeElectric -13.550238 325.553565 -0.042 0.966800   
## fuelTypeHybrid -1.713781 0.736233 -2.328 0.019924 \*   
## fuelTypePetrol -0.498704 0.133819 -3.727 0.000194 \*\*\*  
## f.engineSizeMedium -0.045510 0.137652 -0.331 0.740937   
## f.engineSizeLarge -1.312769 0.204991 -6.404 1.51e-10 \*\*\*  
## transmissionf.Trans-SemiAuto 2.198193 0.566351 3.881 0.000104 \*\*\*  
## transmissionf.Trans-Automatic 1.570882 0.594979 2.640 0.008285 \*\*   
## mpg:transmissionf.Trans-SemiAuto -0.048586 0.010629 -4.571 4.85e-06 \*\*\*  
## mpg:transmissionf.Trans-Automatic -0.034650 0.011073 -3.129 0.001752 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3204.8 on 3343 degrees of freedom  
## AIC: 3228.8  
##   
## Number of Fisher Scoring iterations: 13

vif(b5)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

## GVIF Df GVIF^(1/(2\*Df))  
## mpg 4.373472 1 2.091285  
## year 1.404057 1 1.184929  
## fuelType 2.301266 3 1.149021  
## f.engineSize 2.379795 2 1.242038  
## transmission 712.779613 2 5.167004  
## mpg:transmission 582.442224 2 4.912620

* *Now, all VIF/GVIF values are under 5. So no present collinearity.*

*Let’s compare this model with the model without interactions:*

anova(b3,b5)

## Analysis of Deviance Table  
##   
## Model 1: Audi ~ mpg + year + f.engineSize + fuelType + transmission  
## Model 2: Audi ~ (mpg + year + fuelType + f.engineSize + transmission +   
## mpg \* transmission)  
## Resid. Df Resid. Dev Df Deviance  
## 1 3345 3226.9   
## 2 3343 3204.8 2 22.185

* *The lower AIC value in this model compared to the previous one suggests that the added interaction enhances the model’s overall fit*

### *11.3.2 Interaction between two factors*

* *The addition of this new interaction has once again lowered the AIC values, providing evidence of further improvement in the model.*

b6 <- glm(Audi~(mpg+year+fuelType +f.engineSize+transmission + mpg\*transmission + transmission\*f.engineSize),family="binomial", data=train)  
summary(b6)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + fuelType + f.engineSize +   
## transmission + mpg \* transmission + transmission \* f.engineSize),   
## family = "binomial", data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 330.58281 54.85452 6.027  
## mpg -0.03137 0.00833 -3.766  
## year -0.16352 0.02711 -6.033  
## fuelTypeElectric -14.48537 534.84950 -0.027  
## fuelTypeHybrid -1.71317 0.73893 -2.318  
## fuelTypePetrol -0.46489 0.13476 -3.450  
## f.engineSizeMedium 0.20826 0.16992 1.226  
## f.engineSizeLarge -14.57745 269.48398 -0.054  
## transmissionf.Trans-SemiAuto 2.70782 0.59196 4.574  
## transmissionf.Trans-Automatic 1.33313 0.65276 2.042  
## mpg:transmissionf.Trans-SemiAuto -0.04765 0.01072 -4.447  
## mpg:transmissionf.Trans-Automatic -0.03261 0.01118 -2.917  
## f.engineSizeMedium:transmissionf.Trans-SemiAuto -0.83023 0.24177 -3.434  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 12.88373 269.48406 0.048  
## f.engineSizeMedium:transmissionf.Trans-Automatic 0.03691 0.31731 0.116  
## f.engineSizeLarge:transmissionf.Trans-Automatic 13.58396 269.48414 0.050  
## Pr(>|z|)   
## (Intercept) 1.68e-09 \*\*\*  
## mpg 0.000166 \*\*\*  
## year 1.61e-09 \*\*\*  
## fuelTypeElectric 0.978393   
## fuelTypeHybrid 0.020425 \*   
## fuelTypePetrol 0.000561 \*\*\*  
## f.engineSizeMedium 0.220329   
## f.engineSizeLarge 0.956860   
## transmissionf.Trans-SemiAuto 4.78e-06 \*\*\*  
## transmissionf.Trans-Automatic 0.041122 \*   
## mpg:transmissionf.Trans-SemiAuto 8.72e-06 \*\*\*  
## mpg:transmissionf.Trans-Automatic 0.003529 \*\*   
## f.engineSizeMedium:transmissionf.Trans-SemiAuto 0.000595 \*\*\*  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 0.961869   
## f.engineSizeMedium:transmissionf.Trans-Automatic 0.907405   
## f.engineSizeLarge:transmissionf.Trans-Automatic 0.959798   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3185.7 on 3339 degrees of freedom  
## AIC: 3217.7  
##   
## Number of Fisher Scoring iterations: 14

* *Interaction terms inherently complicate the multicollinearity assessment because they represent the combined effect of two used variables, potentially correlating with their individual main effects.*

vif(b6, type = 'predictor')

## GVIF Df GVIF^(1/(2\*Df))  
## mpg 4.403366e+00 1 2.098420  
## year 1.405621e+00 1 1.185589  
## fuelType 2.326653e+00 3 1.151124  
## f.engineSize 1.241393e+07 2 59.357735  
## transmission 1.076192e+03 2 5.727596  
## mpg:transmission 6.105347e+02 2 4.970815  
## f.engineSize:transmission 1.646401e+08 4 10.643071

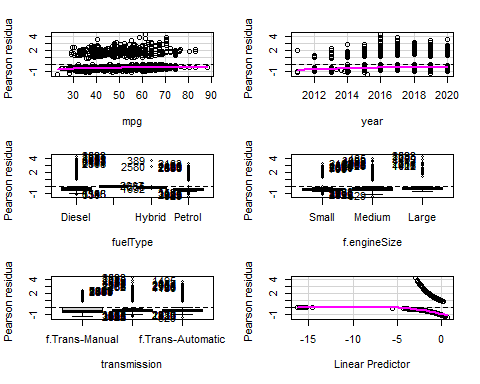
* *Adding the interaction between transmission and f.engineSize to the model significantly improves the fit, as indicated by the decrease in residual deviance and the highly significant p-value (0.0009509).*

anova(b5,b6, test="LR")

## Analysis of Deviance Table  
##   
## Model 1: Audi ~ (mpg + year + fuelType + f.engineSize + transmission +   
## mpg \* transmission)  
## Model 2: Audi ~ (mpg + year + fuelType + f.engineSize + transmission +   
## mpg \* transmission + transmission \* f.engineSize)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 3343 3204.8   
## 2 3339 3185.7 4 19.08 0.0007579 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Residuals look much better than the previous model’s one.

residualPlots(b6)



## Test stat Pr(>|Test stat|)   
## mpg 1.398 0.2370587   
## year 13.901 0.0001927 \*\*\*  
## fuelType   
## f.engineSize   
## transmission   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## *11.4 Diagnosis & Unusual-Influential Data Detection*

* *Till now, we built a model predicting the likelihood of a car being an Audi based on factors like mpg, year, engine size category, fuel type, and transmission, including some interactions between these variables. Most variables do significantly influence the prediction, except for few. Possible due to lack of cars with these characteristics in our train dataset. Still, this model seems to have some predictive power.*

summary(b6)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + fuelType + f.engineSize +   
## transmission + mpg \* transmission + transmission \* f.engineSize),   
## family = "binomial", data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 330.58281 54.85452 6.027  
## mpg -0.03137 0.00833 -3.766  
## year -0.16352 0.02711 -6.033  
## fuelTypeElectric -14.48537 534.84950 -0.027  
## fuelTypeHybrid -1.71317 0.73893 -2.318  
## fuelTypePetrol -0.46489 0.13476 -3.450  
## f.engineSizeMedium 0.20826 0.16992 1.226  
## f.engineSizeLarge -14.57745 269.48398 -0.054  
## transmissionf.Trans-SemiAuto 2.70782 0.59196 4.574  
## transmissionf.Trans-Automatic 1.33313 0.65276 2.042  
## mpg:transmissionf.Trans-SemiAuto -0.04765 0.01072 -4.447  
## mpg:transmissionf.Trans-Automatic -0.03261 0.01118 -2.917  
## f.engineSizeMedium:transmissionf.Trans-SemiAuto -0.83023 0.24177 -3.434  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 12.88373 269.48406 0.048  
## f.engineSizeMedium:transmissionf.Trans-Automatic 0.03691 0.31731 0.116  
## f.engineSizeLarge:transmissionf.Trans-Automatic 13.58396 269.48414 0.050  
## Pr(>|z|)   
## (Intercept) 1.68e-09 \*\*\*  
## mpg 0.000166 \*\*\*  
## year 1.61e-09 \*\*\*  
## fuelTypeElectric 0.978393   
## fuelTypeHybrid 0.020425 \*   
## fuelTypePetrol 0.000561 \*\*\*  
## f.engineSizeMedium 0.220329   
## f.engineSizeLarge 0.956860   
## transmissionf.Trans-SemiAuto 4.78e-06 \*\*\*  
## transmissionf.Trans-Automatic 0.041122 \*   
## mpg:transmissionf.Trans-SemiAuto 8.72e-06 \*\*\*  
## mpg:transmissionf.Trans-Automatic 0.003529 \*\*   
## f.engineSizeMedium:transmissionf.Trans-SemiAuto 0.000595 \*\*\*  
## f.engineSizeLarge:transmissionf.Trans-SemiAuto 0.961869   
## f.engineSizeMedium:transmissionf.Trans-Automatic 0.907405   
## f.engineSizeLarge:transmissionf.Trans-Automatic 0.959798   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3407.0 on 3354 degrees of freedom  
## Residual deviance: 3185.7 on 3339 degrees of freedom  
## AIC: 3217.7  
##   
## Number of Fisher Scoring iterations: 14

-fsdfsdfd

library(lmtest)  
bptest(b6)

##   
## studentized Breusch-Pagan test  
##   
## data: b6  
## BP = 286.44, df = 15, p-value < 2.2e-16

Anova(b6)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: Audi  
## LR Chisq Df Pr(>Chisq)   
## mpg 91.855 1 < 2.2e-16 \*\*\*  
## year 35.944 1 2.031e-09 \*\*\*  
## fuelType 21.282 3 9.197e-05 \*\*\*  
## f.engineSize 87.529 2 < 2.2e-16 \*\*\*  
## transmission 11.340 2 0.0034483 \*\*   
## mpg:transmission 20.858 2 2.957e-05 \*\*\*  
## f.engineSize:transmission 19.080 4 0.0007579 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* *Identifying observations with high leverage, which are those that are unusual or distinct from the rest of the data in terms of the predictor values.*

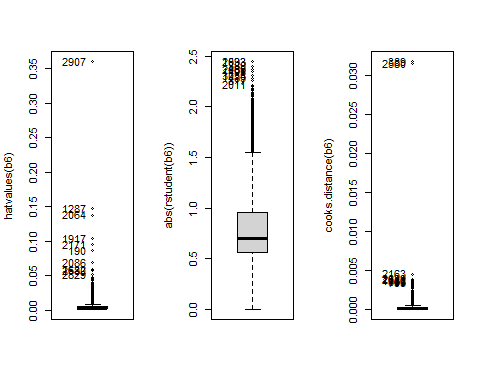
par(mfrow=c(1,3))  
Boxplot(hatvalues(b6),id=c(labels=row.names(train)))

## [1] 2907 1287 2064 1917 2171 190 2086 1632 2580 2829

Boxplot(abs(rstudent(b6)),id=c(labels=row.names(train)))

## [1] 1033 2593 389 485 2089 1495 3038 1483 817 2011

Boxplot(cooks.distance(b6),id=c(labels=row.names(train)))



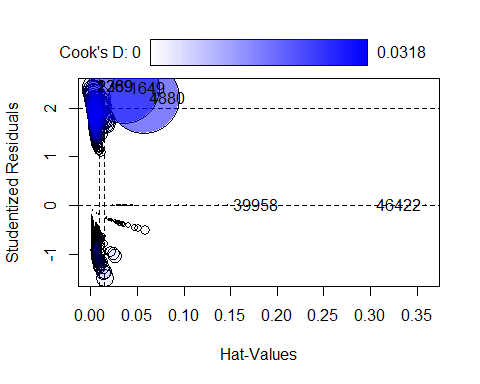
## [1] 389 2580 2163 2629 2811 156 1181 564 403 793

* *These diagnostic plots are crucial for regression analysis as they help in identifying observations that might be unduly influencing the model, either through high leverage, large influence on the model’s predictions, or being outliers in terms of the response variable.*
* *Let’s remove them and shape our model:*

out1 <- which(abs(rstudent(b6))>1.7);  
out2 <- which(abs(cooks.distance(b6))>0.003);  
out3 <- which(abs(hatvalues(b6))>0.03);  
outs<-unique(c(out1,out2,out3))

* *Influence plot can helps us identify outliers, influential data points, or observations that have a large impact on the model’s coefficients.*
* *We can observe how certain points stand out noticeably from the clusters formed, with some exerting significant influence.*

par(mfrow=c(1,1));  
outs2 <- influencePlot(b6, id=c(labels=row.names(train)));



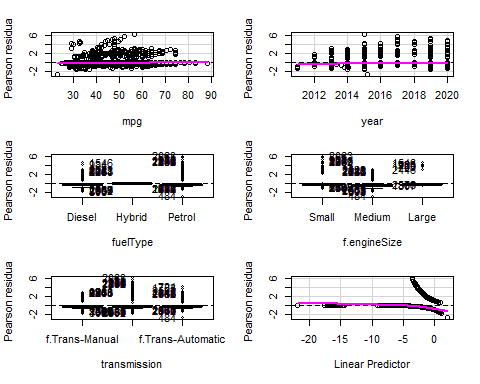
outs2 <- labels(outs2)[[1]];  
outs2 <- as.numeric(outs2);  
outs<-unique(outs,outs2)

b7<-update(b5, data = train[-outs,])  
summary(b7)

##   
## Call:  
## glm(formula = Audi ~ (mpg + year + fuelType + f.engineSize +   
## transmission + mpg \* transmission), family = "binomial",   
## data = train[-outs, ])  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 585.53654 72.73105 8.051 8.23e-16 \*\*\*  
## mpg -0.06637 0.01132 -5.861 4.59e-09 \*\*\*  
## year -0.28936 0.03594 -8.051 8.24e-16 \*\*\*  
## fuelTypeHybrid -12.95171 297.68190 -0.044 0.965296   
## fuelTypePetrol -0.86731 0.17361 -4.996 5.86e-07 \*\*\*  
## f.engineSizeMedium 0.64097 0.20082 3.192 0.001414 \*\*   
## f.engineSizeLarge -3.15124 0.45804 -6.880 5.99e-12 \*\*\*  
## transmissionf.Trans-SemiAuto 4.19256 0.83066 5.047 4.48e-07 \*\*\*  
## transmissionf.Trans-Automatic 3.12655 0.86514 3.614 0.000302 \*\*\*  
## mpg:transmissionf.Trans-SemiAuto -0.10034 0.01700 -5.901 3.61e-09 \*\*\*  
## mpg:transmissionf.Trans-Automatic -0.07527 0.01742 -4.321 1.55e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2159.2 on 2974 degrees of freedom  
## Residual deviance: 1705.2 on 2964 degrees of freedom  
## AIC: 1727.2  
##   
## Number of Fisher Scoring iterations: 14

* *The updated regression model b7 shows improved statistical significance and model fit following the removal of influential observations. This is evident from the reduced AIC and deviances, indicating a better model representation of the underlying data relationship. Significant variables such as ‘mpg’ and ‘year’ demonstrate the model’s enhanced predictiveness after excluding outliers.*
* Overall, the residuals are better clustered around the zero line than the previous model where points were more scattered, so it is a good indication. We can see that some variables still have some outliers, so we will make a test to check if they are significant.

residualPlots(b7)



## Test stat Pr(>|Test stat|)   
## mpg 5.0272 0.02495 \*  
## year 5.5768 0.01820 \*  
## fuelType   
## f.engineSize   
## transmission   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

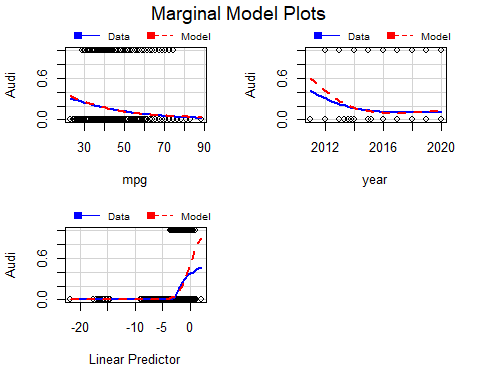
* This suggests that there are no significant outliers in the model based on the Bonferroni-adjusted p-values.

outlierTest(b7)

## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 5977 2.713682 0.006654 NA

* *Some regions in the model predictions align closely with the data, there are discrepancies in others. This implies that while the model may capture the overall trend, there could be room for refinement to improve its accuracy in certain areas.*

marginalModelPlots(b7)



## *11.5 Predictive Power & Quality of Fit*

* *Let’s take a look over the final model that we built.*

Anova(b7)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: Audi  
## LR Chisq Df Pr(>Chisq)   
## mpg 179.308 1 < 2.2e-16 \*\*\*  
## year 64.654 1 8.926e-16 \*\*\*  
## fuelType 28.128 2 7.799e-07 \*\*\*  
## f.engineSize 233.380 2 < 2.2e-16 \*\*\*  
## transmission 29.180 2 4.608e-07 \*\*\*  
## mpg:transmission 42.162 2 6.993e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* *Every variable and interaction term has a highly significant p-value (p < 0.001 for all), which suggests that they all have a statistically significant effect on the response variable. The presence of very low p-values (especially those less than 2.2e-16) suggests that the model has a good fit in terms of the statistical significance of the predictors.*
* *We verify the quality of the fit based on the deviance for b6 (including influential data) and b7 (excluding influential data).*

1-pchisq(b6$deviance, b6$df.residual)

## [1] 0.9711256

1-pchisq(b7$deviance, b7$df.residual)

## [1] 1

* *Given that model* ***b6*** *includes influential data points and model* ***b7*** *excludes them, the p-values can tell us something about the impact of these points on the model fit. The high p-value of 0.784 for model* ***b6*** *suggests that even with the influential data included, the model appears to fit the data well; however, the presence of these points might not be dramatically affecting the overall fit.*
* *For model* ***b7****, the perfect p-value of 1 after excluding influential data may indicate that the model fits the non-influential data exceptionally well, which could be interpreted as the influential data having had a distorting effect on the model.*
* *Below, we can see Pearson’s chi-squared test statistic for model b7 and b6 and then computing the corresponding p-value to assess the goodness of fit of the model. A high p-value suggests that the model has a good fit to the observed data.*

X2\_b7 <- sum((resid(b7, "pearson")^2))  
1-pchisq( X2\_b7, b7$df.res)

## [1] 1

X2\_b6 <- sum((resid(b6, "pearson")^2))  
1-pchisq( X2\_b6, b6$df.res)

## [1] 0.9791007

* *Using the Hosmer-Lemeshow test helps check if our model’s predictions match up with actual data, telling us if the fit is good.*

library(ResourceSelection)

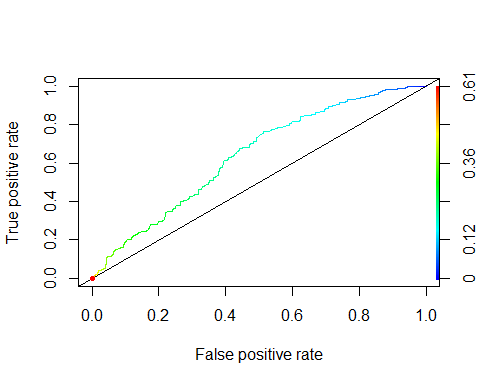
## ResourceSelection 0.3-6 2023-06-27

library(ROCR)  
test$fuelType <- factor(test$fuelType, levels = levels(b6$model$fuelType))  
ll <- which( is.finite(test$fuelType) )  
  
pred\_test <- predict(b6, newdata=test[ll,], type="response")  
ht <- hoslem.test(as.numeric(test$Audi[ll])-1, pred\_test)  
ht

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: as.numeric(test$Audi[ll]) - 1, pred\_test  
## X-squared = 24.213, df = 8, p-value = 0.00211

* This indicates that there is not enough evidence to reject the null hypothesis of the test, which states that the model’s predictions are not significantly different from the actual values — in other words, the model fits well.

pred <- prediction(pred\_test, test$Audi[ll])  
perf <- performance(pred,measure="tpr",x.measure="fpr")  
plot(perf,colorize=TRUE,type="l")  
abline(a=0,b=1)  
# Área bajo la curva  
AUC <- performance(pred,measure="auc")  
AUCaltura <- AUC@y.values  
# Punto de corte óptimo  
cost.perf <- performance(pred, measure ="cost")  
opt.cut <- pred@cutoffs[[1]][which.min(cost.perf@y.values[[1]])]  
#coordenadas del punto de corte óptimo  
x<-perf@x.values[[1]][which.min(cost.perf@y.values[[1]])]  
y<-perf@y.values[[1]][which.min(cost.perf@y.values[[1]])]  
points(x,y, pch=20, col="red")



* *In our plot, the curve is above the line of no discrimination, which suggests that the model has a good ability to distinguish between the positive class and the negative class, suggests a model that outperforms random guessing, as evidenced by the blue line’s ascent above the diagonal line of no discrimination.*
* *While the curve does not hug the upper left corner, which would indicate a perfect model, it still shows a good balance between sensitivity and the ability to avoid false alarms, implying a reasonable level of accuracy.*
* *Overall, the plot indicates a model that is statistically useful.*
* *An AUC score of 0.6459242 indicates that your model has fair discriminative ability to distinguish between the positive and negative classes. While not indicative of a strong predictive model, this level of AUC suggests that the model performs better than random chance.*

# Área bajo la curva  
AUC <- performance(pred,measure="auc")  
AUCaltura <- AUC@y.values  
cat("AUC:", AUCaltura[[1]])

## AUC: 0.638286

## *11.6 Confusion matrix*

audi.est <- ifelse(pred\_test<0.4,0,1)  
tt<-table(audi.est,test$Audi[ll]);tt;

##   
## audi.est Audi No Audi Yes  
## 0 1073 309  
## 1 40 16

* ***What percentage of the model’s predictions were correct?***

100\*sum(diag(tt))/sum(tt)

## [1] 75.73018

* ***How accurate the model’s positive predictions are?***
* *25.64% means that when the model predicts an instance as positive, about 25.64% of these predictions are correct, and the rest are false positives.*

100\*(tt[2,2]/(tt[2,1]+ tt[2,2]))

## [1] 28.57143

* ***Evaluating the binary classification model:***

prob.audi <- b6$fit[ll]  
audi.est <- ifelse(prob.audi<0.5,0,1)  
tt<-table(audi.est,df$Audi[ll]);tt

##   
## audi.est Audi No Audi Yes  
## 0 1139 292  
## 1 6 1

* *The model appears to have a high number of True Negatives and a low number of True Positives, suggesting it is better at identifying ‘No Audi’ than ‘Audi Yes’. This could very well be related to an imbalance in the dataset, where there are far fewer ‘Audi’ cars compared to ‘No Audi’ cars.*
* ***How well the model is at correctly identifying negative instances?***
* *A rate of 79.35% True Negatives suggests that the model is quite effective at correctly identifying instances of the negative class. It indicates a strong ability of the model to recognize situations where the condition it’s trying to predict is absent.*

100\*tt[1,1]/sum(tt)

## [1] 79.20723

* ***How well the model is at correctly identifying positive instances?***
* *The precision of 28.57% indicates that the model’s ability to correctly identify positive instances is limited, and a significant number of its positive predictions are actually false positives.*

100\*(tt[2,2]/(tt[2,1]+ tt[2,2]))

## [1] 14.28571