# Assignment 1

Enric Reverter & Gerard Pons

21/11/2021

## Contents

## Data Preparation and Exploration

In this assignment, a dataset about car prices has been used and explored, to which statistical inference and modeling techniques have been applied. The aim of the project is to gain insights on the price of the cars using the exploratory variables and also create models to try to explain or predict this response variable.

***The code for this first section is not displayed within the same document to avoid having too many pages. It is considered that it does not contribute enough as the visualizations are already sufficient. If necessary, it can be checked in the uploaded .Rmd file.***

The first and most crucial step has been the data preparation and exploration phase, in which the structural consistency of the data is checked, missing values and outliers are treated, and the variables are explored so as to make the proper decisions in future steps.

Regarding general data cleaning, the following points have been assessed:

1. The four homogeneous datasets are merged into a single one.
2. Structural errors are fixed: categorical values are mutated into factors and extra blanks are removed in *Model*.
3. Duplicated observations are removed.
4. A sample of 5000 observations is selected.
5. *Year* variable is transformed into *Age*.
6. Wrongly classified cars are labeled as electric according to their engine size. Out of the 14 cars with *engineSize* 0, only one of them is electric, and the correct *engineSize* is imputed to the database according to the online car specifications.
7. Missing values are explored. 0 NA's are found.

After that, outlier detection has been addressed, which is first done with a univariate exploration and then with a multivariate one. For that, another attribute is created in the dataset, *Outliers*, in order to keep track of the number of univariate outliers each individual has. Henceforth, observations which only have one univariate outlier are imputed with NA while the ones with more are kept so the multivariate outliers do not become biased. That is, multivariate outliers might be part of a different group, which can be treated and modelled separately. All the individuals with more than one univariate outlier have been detected with the Mahalanobis distance and thus, have been removed from the main analysis. The details can be depicted below.

<u>Response Variable: Price</u>

The extreme outliers for the response variable have been treated differently from the explanatory variables, as they can highly influence our models and analyses. Hence, the 46 outlying observations (0,9%) have been completely removed. The boxplot can be depicted in Figure 1.

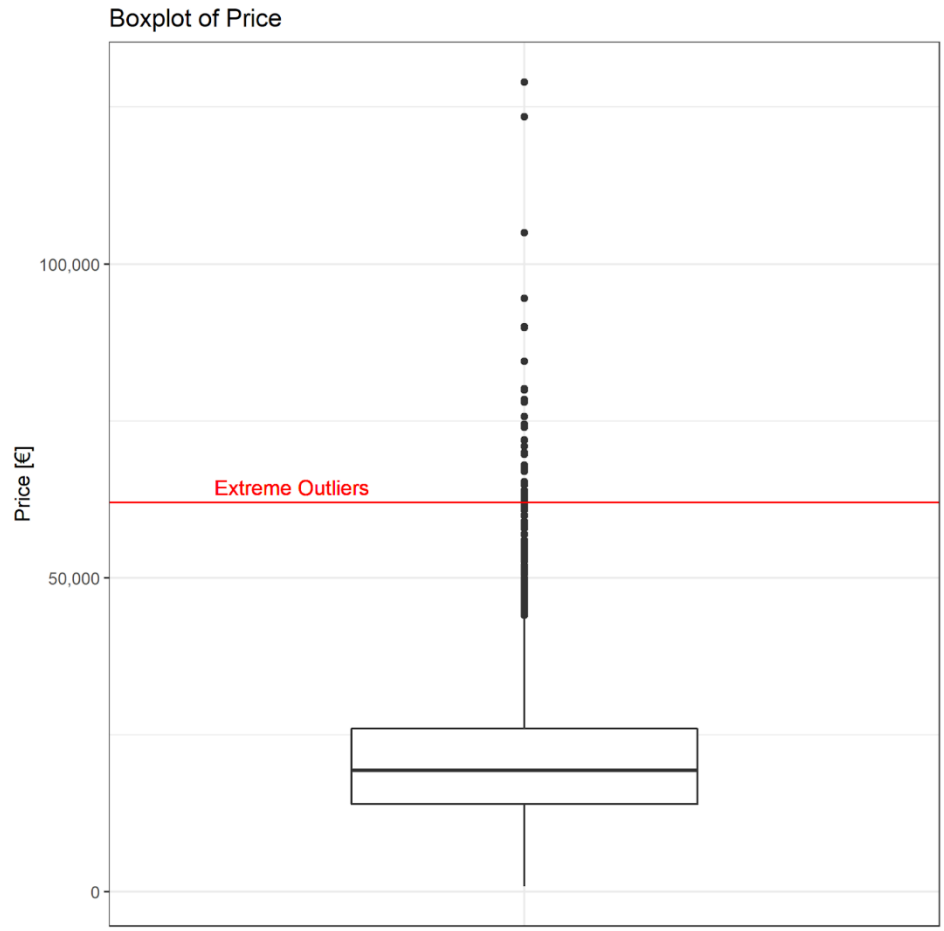



*Figure 1. Boxplot of Price. Extreme outliers are those above the displayed red line.*

Quantitative explanatory variables

It has been observed that all variables have extreme outliers and that tax is the only one which has outliers in both directions. The number of extreme outliers for each category is summarized in the following table and can be depicted in Figure 2:

| | **Mileage** | **Tax** | **MPG** | **Age** |
|---|---|---|---|---|
| **#Outliers** | 15 | 160* | 62 | 16 |

*These are only the extreme outliers in the right tail.

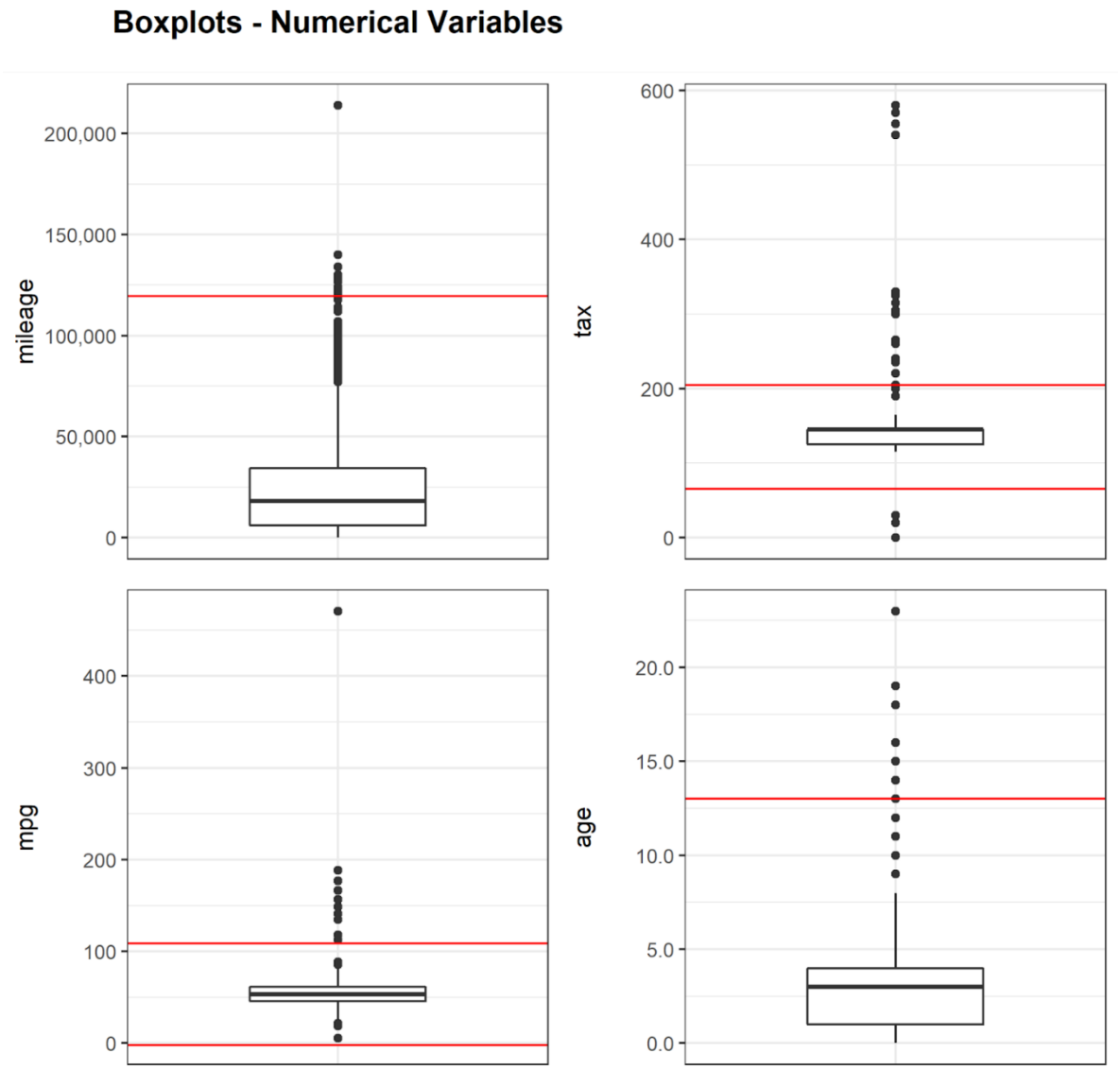



*Figure 2. Boxplots of the explanatory numerical variables. Extreme outliers are those above the displayed red line.*

It has been detected that 4.72% of the observations have at least a value considered an outlier, and that the vast majority of them (93,2%) contain only one outlier. As stated before, these values have automatically been imputed with NA. In the case where an individual has more than one univariate outlier, it has not been assigned NA's yet, since it might be detected by the multivariate outlier analysis. Thus, in order to not influence the Mahalanobis distances, it will be imputed only if it is not seen as a multivariate outlier.

In the case of the variable *tax*, only the extreme outliers from the right tail are accounted for, since it has been seen the bottom range contains more than 20% of the total observations. This can be depicted in Figure 3. Thus, it has been decided to keep these values in the dataset, aware of the consequences it entails. Hence, this variable has to be treated with some precautions.

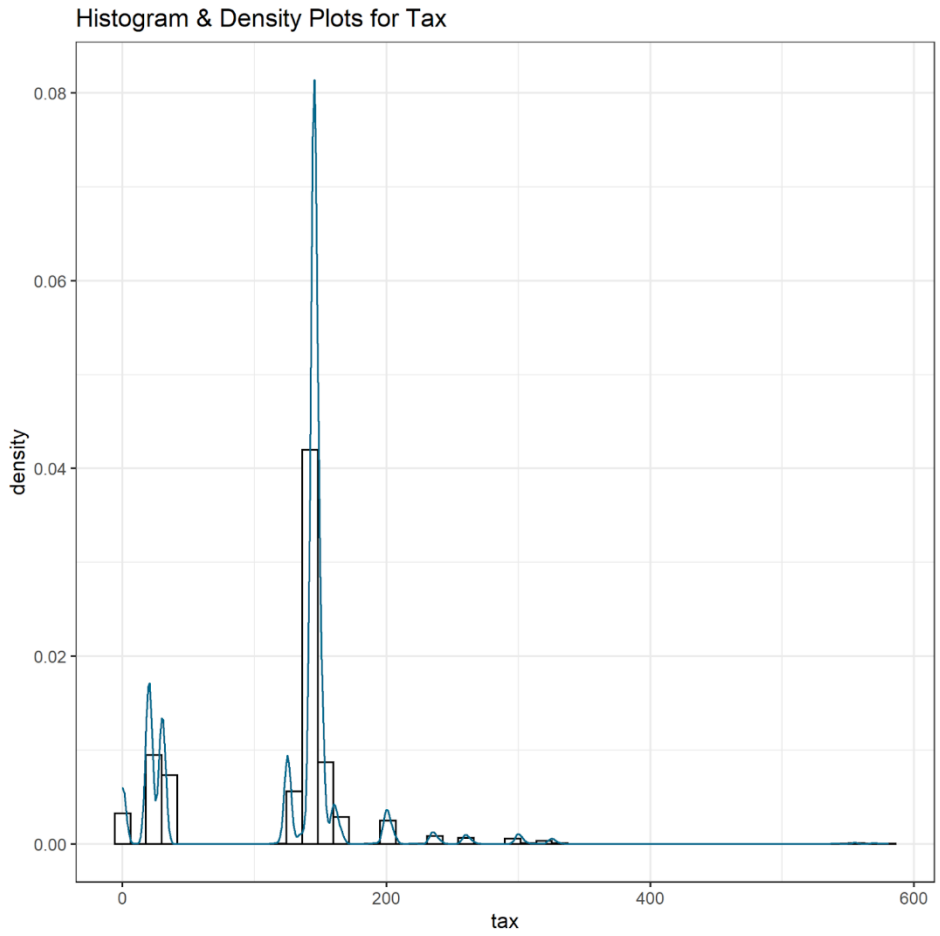



*Figure 3. Histogram and density plot of the variable tax. Notice how there is a clear gap between values.*

Qualitative explanatory variables

Regarding qualitative variables, both *transmissionType* and *manufacturer* have not displayed any interesting remarks, having 3 and 4 distinct categories respectively and no significant count differences. The counts are depicted in Figure 4. The remaining three variables, however, have been assessed differently.

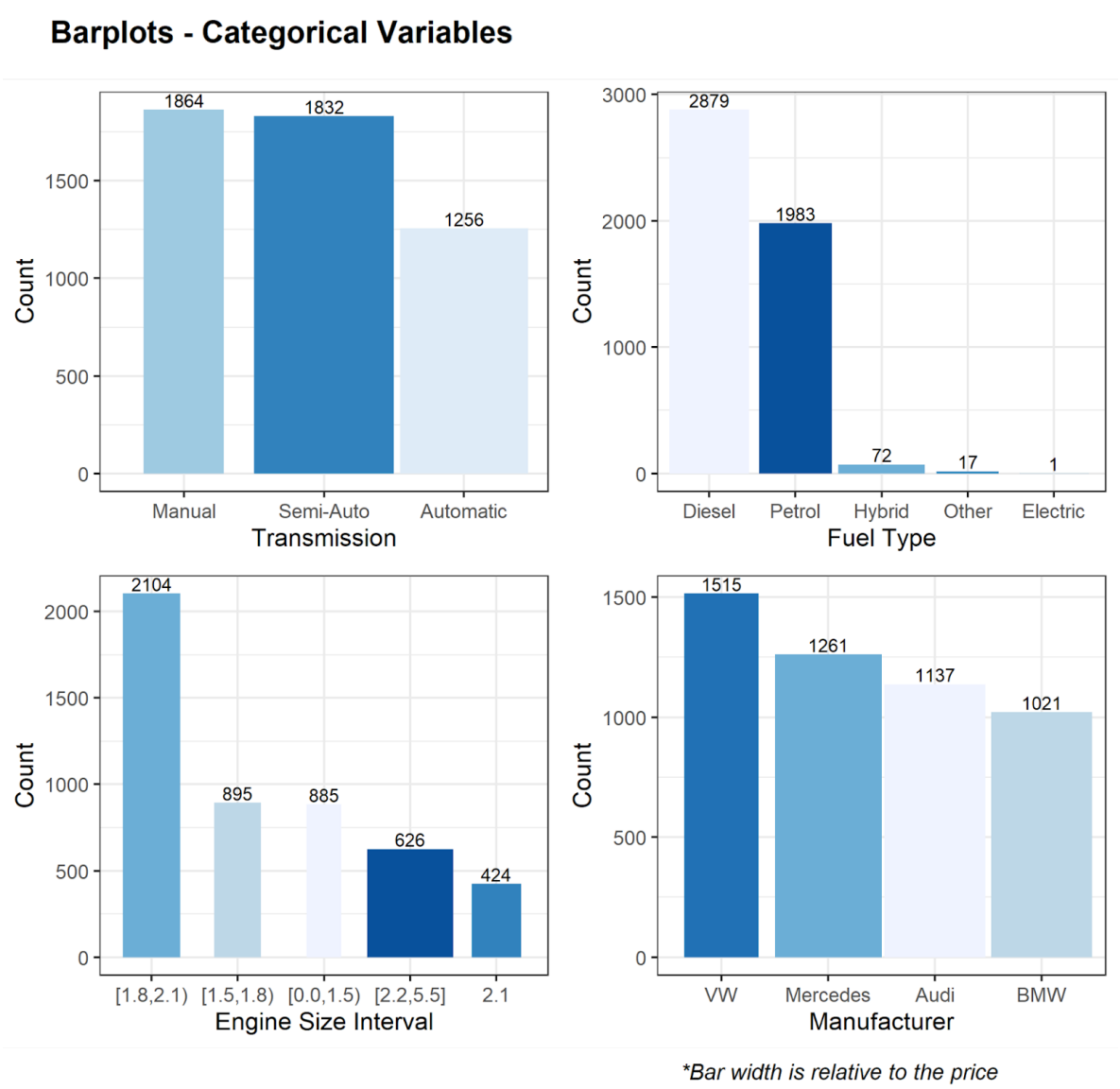



*Figure 4. Bar plots of the explanatory categorical variables. Notice how engine size has already been collapsed into less levels.*

Model

This quantitative variable has 91 different categories. Some of them are very representative, such as the model *Golf*, which accounts for nearly 10% of the dataset, while a great number of other models are only represented once or twice. The most sensible thing to do, as it is not feasible to work with such a high number of categories, is to leave models aside when considering analyses.

Engine Size

As in the variable *Model*, it can be observed that there are a lot of distinct categories that are not uniformly distributed, which makes it not feasible to work with them as a factor. As in this case the variable is ordered categorical (size), it makes sense to create a new 'discrete' variable for engine size, *engineSizeInt*, where size intervals are grouped together according to frequency. The plot in Figure 4 has the groups already collapsed.

Fuel Type

It can be seen that there is only one electric vehicle among the 5000 observations. Also, the category '*Other*' has been addressed because, as of winter 2021, the four manufacturers considered in the database do not produce cars with other engines (i.e., Hydrogen or Biodegradable fuel). After studying the options, manual imputation has been discarded due to the number of ambiguities (i.e., one model can have different types of fuel), so the values are marked as NA and addressed later in the imputation process.

Imputation

After finding all the outliers, the imputation of the missing values is the step which follows. First of all, quantitative variables are managed: PCA imputation is applied; then, it is checked whether the imputed values are incoherent (i.e., negative age) or not, and tested if the distribution of the values is preserved, which it is. After that, categorical variables are dealt with: MCA imputation is applied; similarly, it is checked if the proportions of the variables remain the same after the imputation process.

Multivariate Outliers

Multivariate outliers are detected using the Mahalanobis distance, both classic and robust. Values whose distances are greater than the cut-off threshold are removed from the dataset. As stated before, it is checked if the observations that have two or more univariate outliers are also detected in the multivariate outliers, which turns out to be the case. It is worth mentioning that a small group of very old cars with very high values of mileage are detected within the multivariate outliers, which is not surprising. In a more complete study, this group would be analyzed separately. However, it has been decided to completely remove them from the dataset since it falls out of the scope of this assignment. Multivariate outliers are depicted in Figure 5.

**Henceforth, the dataset is clean and ready to be analyzed in the question that follows.**

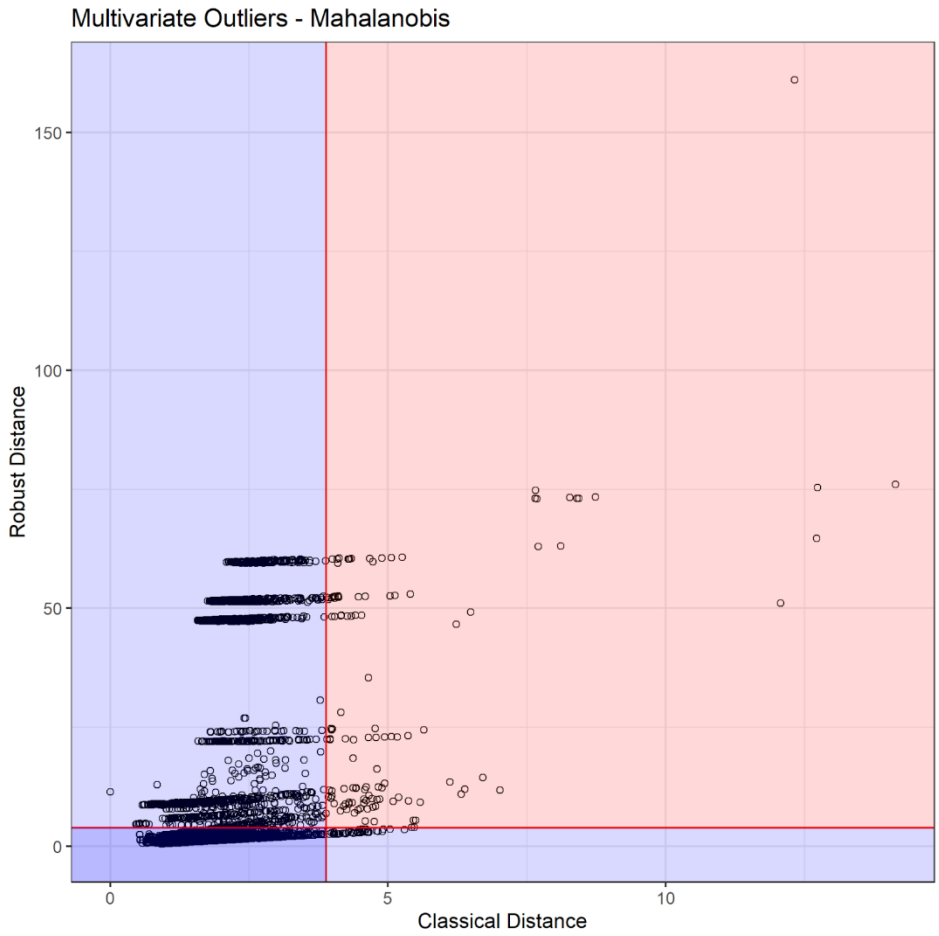



*Figure 5. Multivariate outliers using Mahalanobis distances. The individuals within the red rectangle are the ones deleted from the main analysis.*

Read the cleaned data

```
df = read.csv(file.path(data_path, "cars.csv"))
```

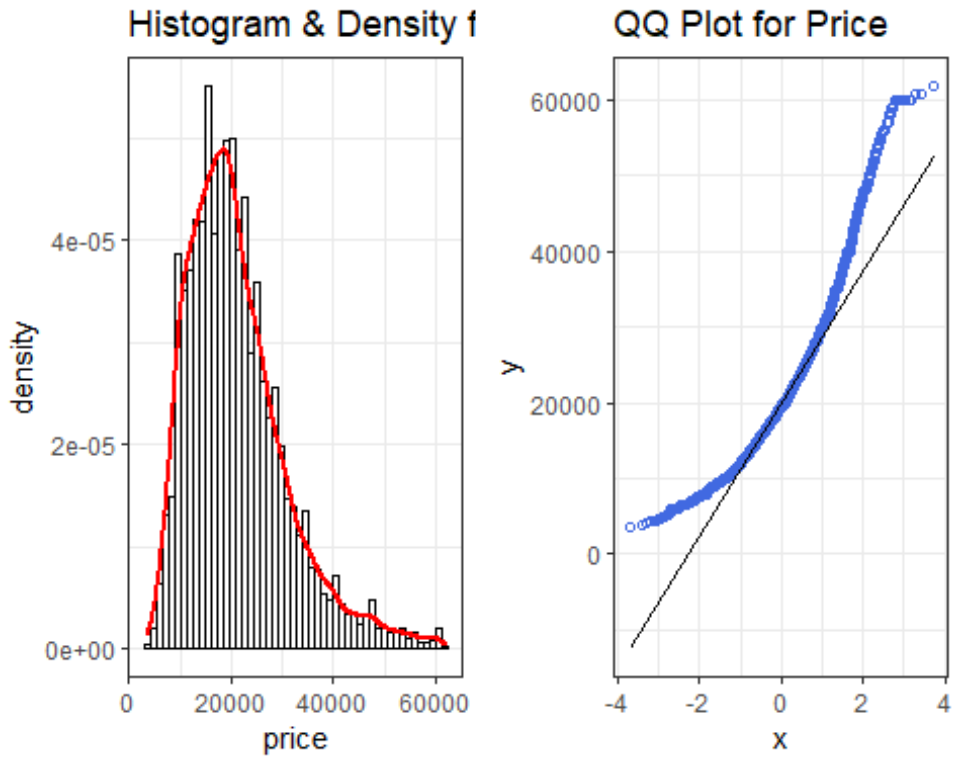
## Question 1

**Q: Determine if the response variable (price) has an acceptably normal distribution. Address test to discard serial correlation.**

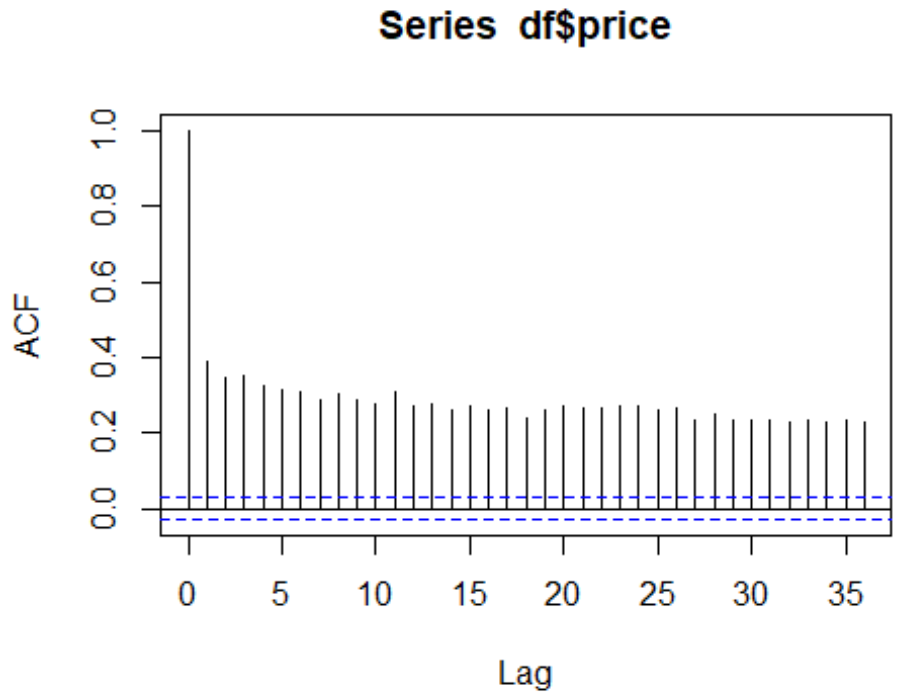
First, the normality of the response variable *price* has been assessed, both graphically and numerically. A Shapiro-Wilk test yields a p-value of $2*10^{-16}$, which does not make it possible to accept the null hypothesis and thus, it is concluded that the variable does not follow a normal distribution. It can also be seen in the plots that its distribution is far from being normal.

Also, the autocorrelation of the response variable has been addressed. A Durbin-Watson test yields a p-value significantly below the threshold (0.05), so the null hypothesis cannot be accepted. In other words, serial correlation is found in the response variable of the dataset.

```
p1 = ggplot(data = df, aes(x=price)) +
  geom_histogram(aes(y=..density..), color = 1, fill = "white", bins = 50
, ) +
  geom_density(color = "red", lwd = 1) +
  labs(title = "Histogram & Density for Price")

p2 = ggplot(data = df, aes(sample=price)) +
  geom_qq(shape = 1, color = "royalblue") + geom_qq_line() +
  labs(title = "QQ Plot for Price")

plot_grid(plotlist = list(p1,p2), align = "h", ncol = 2)
```




```
shapiro.test(df$price)

##
##  Shapiro-Wilk normality test
##
## data:  df$price
## W = 0.9278, p-value < 2.2e-16

acf(df$price)
```




```
dwtest(price ~ 1, data = df)

##
##  Durbin-Watson test
##
## data:  price ~ 1
## DW = 1.2211, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

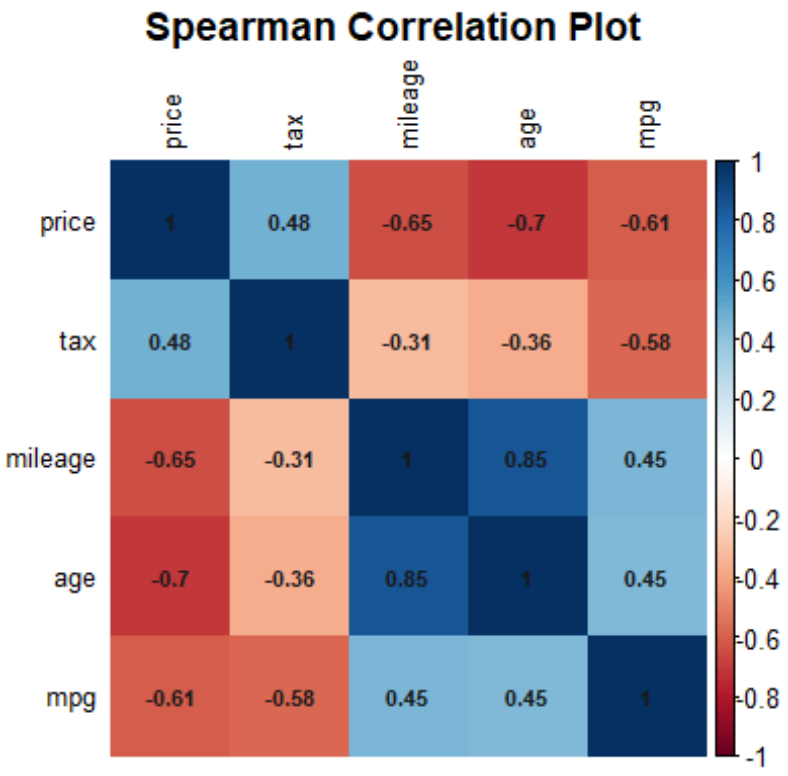
## Question 2

**Q: Indicate by exploration of the data which are apparently the variables most associated with the response variable (use only the indicated variables).**

To address the correlation with the *price* variable, quantitative and qualitative variables are treated differently. Regarding quantitative variables, the association with the response variable has been inspected by means of a correlation test. As seen in the previous step, the price of cars is not normally distributed and thus, the non-parametric method, Spearman, is performed. Therefore, it can be seen that three of the variables: *mpg, mileage*, and *age* have a significant negative correlation. This indicates that old, very used cars, and cars that do not consume a lot of fuel have lower prices. On the other hand, it can be observed that the variable *tax* is positively correlated with price, indicating that cars which have to pay high taxes (usually high-end cars) are more expensive.

Regarding qualitative variables, it can be observed that there is a significant association between them (*transmission*, *fuelType*, and *enginezise_int*) and the price. A Kruskal test has been performed to assess this, where in all cases the obtained p-values are much lower than the standard threshold of 0.05.

```
correlation = cor(df[, sapply(df, is.numeric)],method="spearman")
corrplot(correlation, order="AOE", method="color", addCoef.col = "gray10"
, tl.col = "black", tl.cex = .8, number.digits = 2, number.cex = 0.7, tit
le = "Spearman Correlation Plot", mar=c(0,0,1.5,0))
```



```r
cor.test(df$price, df$tax,method="spearman")
```

```
## Warning in cor.test.default(df$price, df$tax, method = "spearman"): Ca
nnot
## compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  df$price and df$tax
## S = 9827184477, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4757379
```

```r
cor.test(df$price, df$mpg,method="spearman")
```

```
## Warning in cor.test.default(df$price, df$mpg, method = "spearman"): Ca
nnot
## compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  df$price and df$mpg
## S = 3.0157e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.6088215
```

```r
cor.test(df$price, df$age,method="spearman")
```

```
## Warning in cor.test.default(df$price, df$age, method = "spearman"): Ca
nnot
## compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  df$price and df$age
## S = 3.1925e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.7031316
```

```r
cor.test(df$price, df$mileage,method="spearman")
```

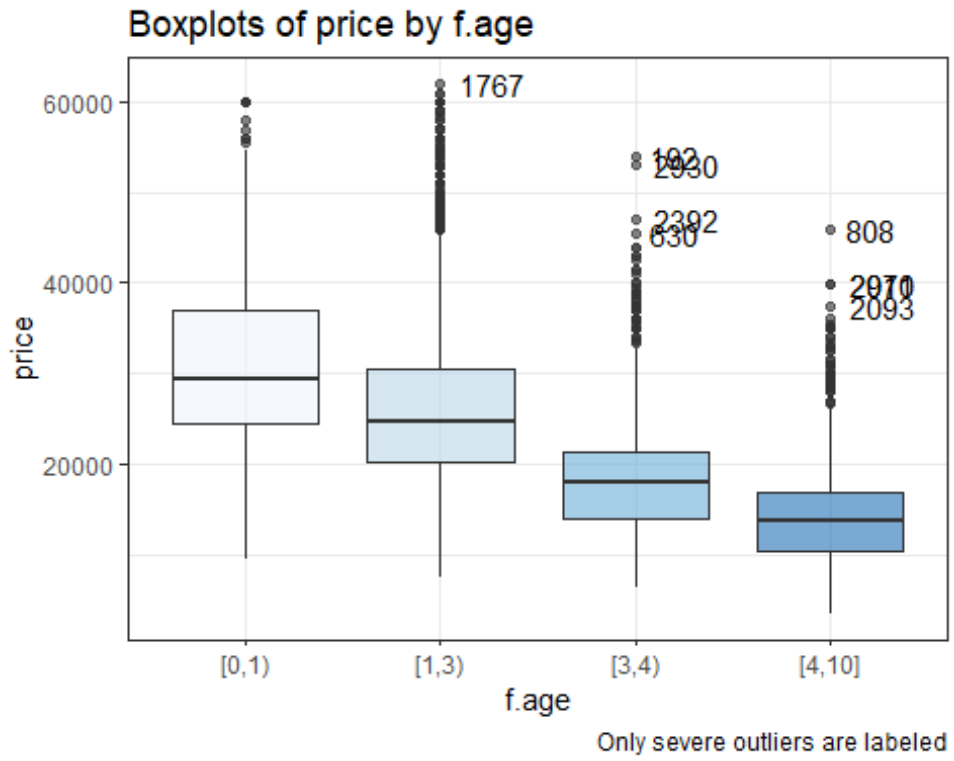
```
## Warning in cor.test.default(df$price, df$mileage, method = "spearman")
: Cannot
## compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  df$price and df$mileage
## S = 3.0909e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.6489515

kruskal.test(df$price,df$manufacturer)

##
##  Kruskal-Wallis rank sum test
##
## data:  df$price and df$manufacturer
## Kruskal-Wallis chi-squared = 521.69, df = 3, p-value < 2.2e-16

kruskal.test(df$price, df$transmission)

##
##  Kruskal-Wallis rank sum test
##
## data:  df$price and df$transmission
## Kruskal-Wallis chi-squared = 1448.7, df = 2, p-value < 2.2e-16

kruskal.test(df$price, df$fuelType)

##
##  Kruskal-Wallis rank sum test
##
## data:  df$price and df$fuelType
## Kruskal-Wallis chi-squared = 51.559, df = 2, p-value = 6.37e-12

kruskal.test(df$price, df$engineSize_int)

##
##  Kruskal-Wallis rank sum test
##
## data:  df$price and df$engineSize_int
## Kruskal-Wallis chi-squared = 1490, df = 4, p-value < 2.2e-16
```

## Question 3

**Q: Define a polytomic factor f.age for the covariate car age according to its quartiles and argue if the average price depends on the level of age. Statistically justify the answer.**

As previously seen, the categorical variable *age* is, in fact, negatively correlated with the response variable. Hence, in order to further explore this relationship, the *age* variable has been discretized to its four quantiles, creating a new factor. It can be observed, both graphically and numerically, the downward trend of the price as age increases. A Kruskal test to corroborate this is also performed. As expected, the p-value is significantly below the threshold, so the null hypothesis cannot be accepted. Moreover, the difference between categories has been studied with a pairwise one-sided Wilcox test. The results are expected, and it can be confidently said that the mean price of the cars from greater quantiles is higher than the one from lower quantiles of age.

```
df$f.age = discretize(df$age,method = "frequency",breaks=4)
table(df$f.age)

##
##  [0,1)  [1,3)  [3,4) [4,10]
##    291   1984    910   1642

round(prop.table(table(df$f.age))*100)

##
##  [0,1)  [1,3)  [3,4) [4,10]
##      6     41     19     34

tapply(df$price, df$f.age, mean)

##    [0,1)    [1,3)    [3,4)   [4,10]
## 31044.32 26479.32 18687.19 14304.35

is_outlier = function(x) {
  return(x < quantile(x, 0.25) - 3 * IQR(x) | x > quantile(x, 0.75) + 3 *
IQR(x))
}

df %>%
  group_by(f.age) %>%
  tibble::rownames_to_column(var = "id") %>%
  mutate(outlier = ifelse(is_outlier(price), id, NA)) %>%
  ggplot(., aes(x = f.age, y = price, fill = f.age)) +
    geom_boxplot(alpha = 0.6, outlier.alpha = NULL) +
    scale_fill_brewer(palette = "Blues") +
    geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3) +
    labs(title = "Boxplots of price by f.age",
         caption = "Only severe outliers are labeled") +
    theme(legend.position = "none")
```




```
kruskal.test(df$price~df$f.age)

##
##  Kruskal-Wallis rank sum test
##
## data:  df$price by df$f.age
## Kruskal-Wallis chi-squared = 2185.8, df = 3, p-value < 2.2e-16

pairwise.wilcox.test(df$price,df$f.age,alternative='less')

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity cor
rection
##
## data:  df$price and df$f.age
##
##         [0,1)  [1,3)  [3,4)
## [1,3)   <2e-16 -      -
## [3,4)   <2e-16 <2e-16 -
## [4,10] <2e-16 <2e-16 <2e-16
##
## P value adjustment method: holm
```

## Question 4

**Q: Calculate and interpret the anova model that explains car price according to the age factor and the fuel type.**

It can clearly be seen that both factors have an influence on the explained variance. However, as explained by the F values, the effect of the *age factor* is more significant than the *fuelType* one. The interpretation of the coefficients of the resulting model is that the older a car is, the lower its predicted price. Also, regarding fuel type, a car being Hybrid is associated with a higher price than a Diesel one, and a Petrol car is associated with a lower price. Also, it should be noted that this model explains only 40% of the variability ($R^2$=0.40).

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

m0 = lm(df$price~1)
m1 = lm(df$price~df$f.age)
m2 = lm(df$price~df$fuelType)
m3 = lm(df$price~df$f.age+df$fuelType)
m4 = lm(df$price~df$f.age*df$fuelType)

anova(m0,m3) #General effect of the proposed model
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ 1
## Model 2: df$price ~ df$f.age + df$fuelType
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   4826 4.5173e+11
## 2   4821 2.6943e+11  5 1.823e+11 652.41 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: df$price
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## df$f.age       3 1.6731e+11 5.5768e+10  997.90 < 2.2e-16 ***
## df$fuelType    2 1.4998e+10 7.4992e+09  134.19 < 2.2e-16 ***
## Residuals   4821 2.6943e+11 5.5886e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = df$price ~ df$f.age + df$fuelType)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19715  -4870  -1102   3024  36468
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32972.4      455.8  72.339   <2e-16 ***
## df$f.age[1,3)      -4880.5      470.3 -10.376   <2e-16 ***
## df$f.age[3,4)     -12825.6      504.5 -25.425   <2e-16 ***
## df$f.age[4,10]    -17501.6      478.2 -36.600   <2e-16 ***
## df$fuelTypeHybrid    336.2      939.6   0.358    0.721
## df$fuelTypePetrol  -3615.9      222.3 -16.268   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7476 on 4821 degrees of freedom
## Multiple R-squared:  0.4036, Adjusted R-squared:  0.4029
## F-statistic: 652.4 on 5 and 4821 DF,  p-value: < 2.2e-16
```

## Question 5

**Q: Do you think that the variability of the price depends on both factors? Does the relation between price and age factor depend on fuel type?**

With the anova test over the base model, the gross effect is calculated for each factor, and it can be seen that both of them explain some variability of the response variable. Additionally, the net effects of both variables are calculated, where the results show that the models are statistically different. Hence, it is a good decision to work with the addition of both. However, as found in the previous exercise, the contribution of *f.age* is significantly larger than the one of *fuelType*.
In addition, when the term *f.age*f.fuelType* is introduced, it appears to be statistically significant, which means the relation between price and age is affected by the fuel type. The adjusted R-squared barely increases with this extension.
Also, the interaction plot can be used to graphically assess this question, where it can be seen that although the price behaviour for all *fuelTypes* is to decrease the older a car is, the price is also influenced by them. The lines of hybrid and diesel cars are close to being parallel while the line of diesel cars deviate from those. More concretely, petrol cars are always the cheaper ones; then, although new diesel cars are more expensive than Hybrid ones, this tendency is reversed as the cars age, so old hybrid cars are more expensive than old diesel cars.

```
anova(m0,m1) #Gross effect factor age
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ 1
## Model 2: df$price ~ df$f.age
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4826 4.5173e+11
## 2   4823 2.8442e+11  3 1.6731e+11 945.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m0,m2) #Gross effect fuelType
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ 1
## Model 2: df$price ~ df$fuelType
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4826 4.5173e+11
## 2   4824 4.4672e+11  2 5006654790 27.032 2.115e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m3,m2) #Net fuelType effect
```

```
## Analysis of Variance Table
##
```

```
## Model 1: df$price ~ df$f.age + df$fuelType
## Model 2: df$price ~ df$fuelType
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4821 2.6943e+11
## 2   4824 4.4672e+11 -3 -1.773e+11 1057.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

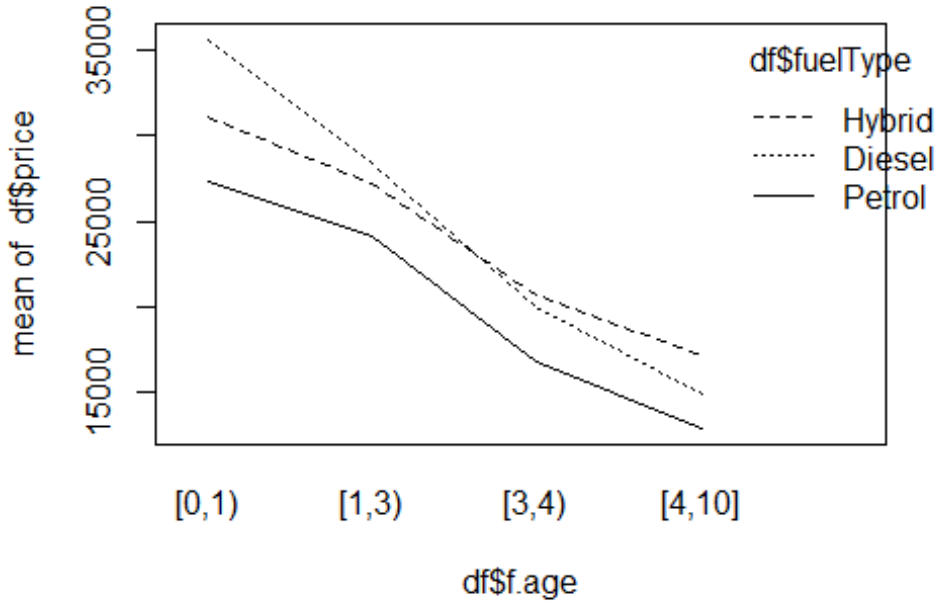
```
anova(m3,m1) #Net factor age effect
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ df$f.age + df$fuelType
## Model 2: df$price ~ df$f.age
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1   4821 2.6943e+11
## 2   4823 2.8442e+11 -2 -1.4998e+10 134.19 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m3,m4) #Check interactions
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ df$f.age + df$fuelType
## Model 2: df$price ~ df$f.age * df$fuelType
##   Res.Df        RSS Df  Sum of Sq    F   Pr(>F)
## 1   4821 2.6943e+11
## 2   4815 2.6665e+11  6 2771224546 8.34 5.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,1))
interaction.plot(df$f.age,df$fuelType,df$price)
```

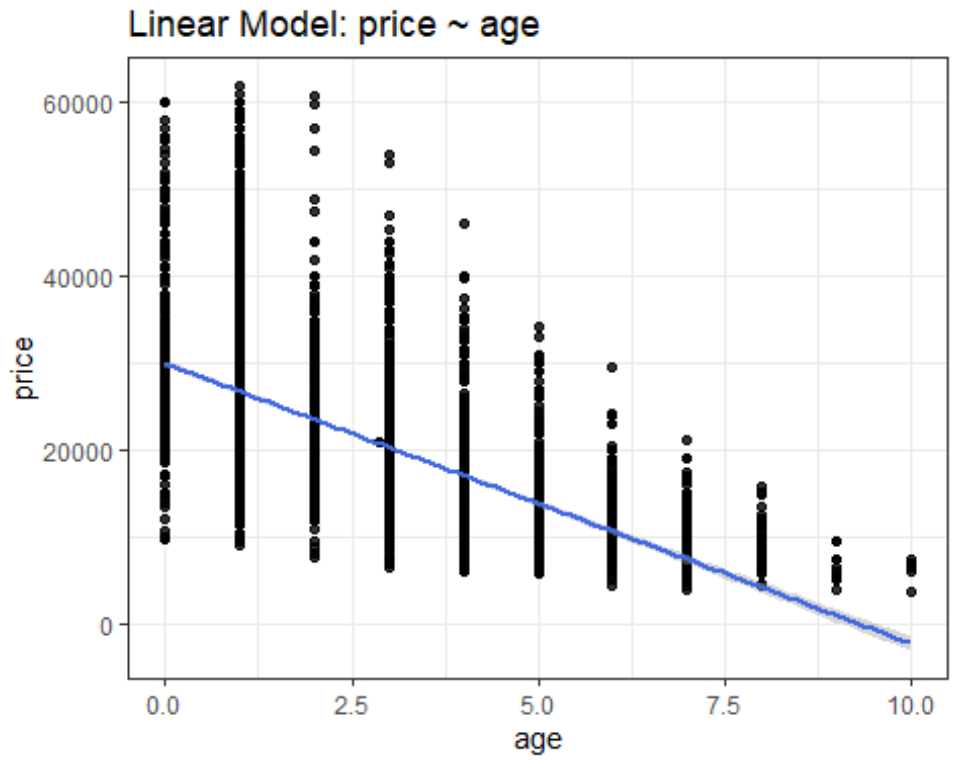

mean of df$price
35000
25000
15000
df$fuelType
Hybrid
Diesel
Petrol
[0,1)
[1,3)
[3,4)
[4,10]
df$f.age

## Question 6 & 7

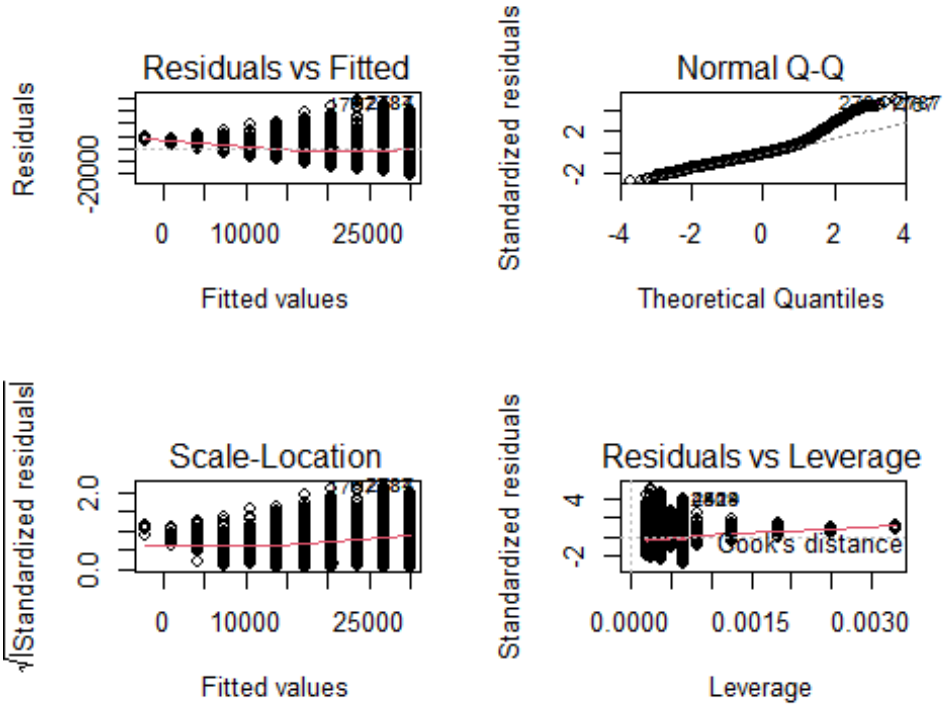
**Q: Calculate the linear regression model that explains the price from the age: interpret the regression line and assess its quality. What is the percentage of the price variability that is explained by the age of the car?**

The resulting linear regression line has a very high intercept (meaning that new cars, i.e age=*0* have a high price) and a negative slope, which coincides with the previously found negative correlation of these two variables. However, as age has a very high granularity (years as integers) a very high-quality line tracing the different points cannot be seen. Moreover, as can be observed in the graphic below, for *age=10*, which is a value inside the established predicting range, the predicted price is negative, which is obviously impossible.
The variability explained by this variable alone is $R^2$=0.39, not an extremely high value and comparable with the one in the previous model with the two factors.
Regarding the study of the residuals, it can clearly be seen that there is a high deviation from a normal distribution and thus, it suggests that a transformation might be needed. A quick exploration with *boxcox* reveals a very narrow interval that does not contain the value 1, and that is near 0. Hence, a proposed transformation would be to apply the logarithm to the response variable, but this effect will be explored in future questions.

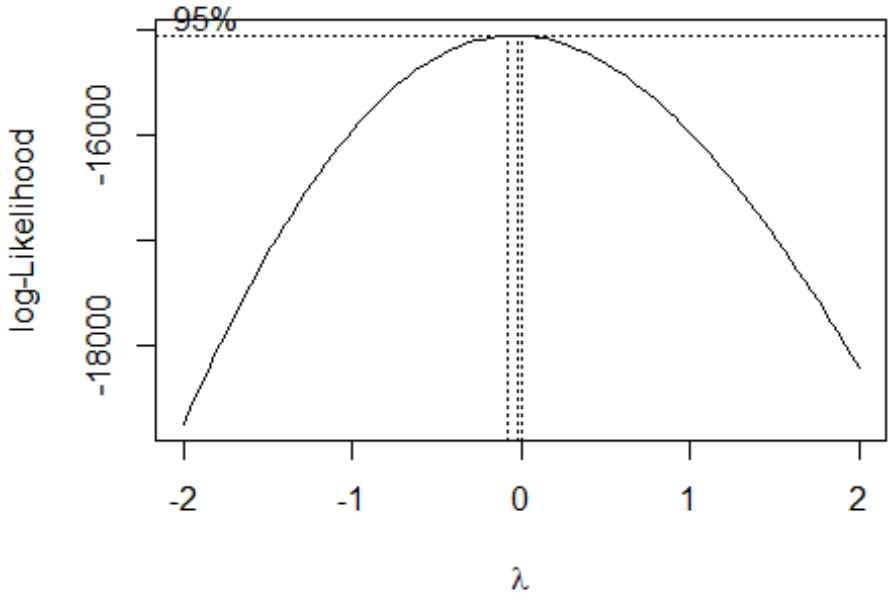
```
m2 = lm(df$price~df$age)

ggplot(df, aes(x = age, y = price)) +
  geom_point(alpha = 0.8) +
  stat_smooth(method = "lm", col = "royalblue") +
  labs(title = "Linear Model: price ~ age")

## `geom_smooth()` using formula 'y ~ x'
```




```
summary(m2)

##
## Call:
## lm(formula = df$price ~ df$age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20332  -4821   -895   3002  37261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29972.89     192.33  155.84   <2e-16 ***
## df$age      -3217.65      57.83  -55.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7552 on 4825 degrees of freedom
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3907
## F-statistic:  3096 on 1 and 4825 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m2)
```




```
par(mfrow=c(1,1))

boxcox(df$price ~ df$age)
```



## Question 8

**Q: Do you think it is necessary to introduce a quadratic term in the equation that relates the price to its age?**

From the anova test comparing both models, it can be seen that they are statistically different and that there is a decrease of the residuals when the quadratic term is used. However, it must be noted that the improvement with the second model is not very significant, going from $R^2$=0.39 to $R^2$=0.4. Moreover, the same problems were encountered as before with regard to the residual behavior. In fact, when applying BoxTidwell method it can be observed that a quadratic transformation is not recommended. Hence, this is not maintained through the next questions.

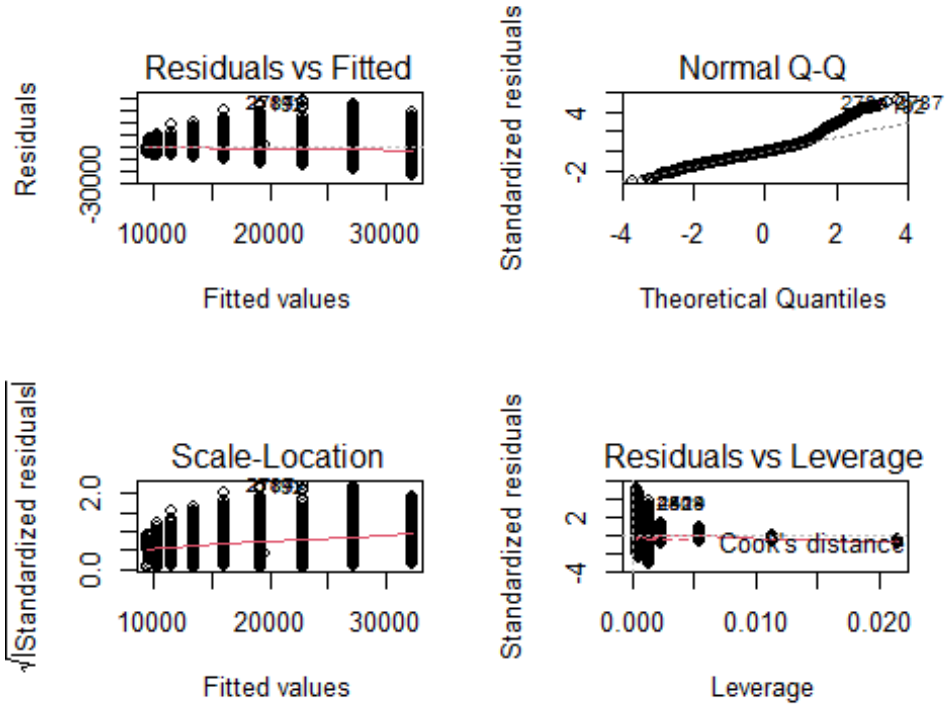
```
m22 = lm(df$price~df$age+I(df$age^2))
summary(m22)
```

```
##
## Call:
## lm(formula = df$price ~ df$age + I(df$age^2))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22531  -4560   -940   2809  37891
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32171.88     262.00  122.79   <2e-16 ***
## df$age      -5235.54     175.55  -29.82   <2e-16 ***
## I(df$age^2)   301.75      24.83   12.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7440 on 4824 degrees of freedom
## Multiple R-squared:  0.409,  Adjusted R-squared:  0.4087
## F-statistic:  1669 on 2 and 4824 DF,  p-value: < 2.2e-16
```

```
anova(m2,m22)
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ df$age
## Model 2: df$price ~ df$age + I(df$age^2)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4825 2.7517e+11
## 2   4824 2.6699e+11  1 8173829948 147.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(m22)
```



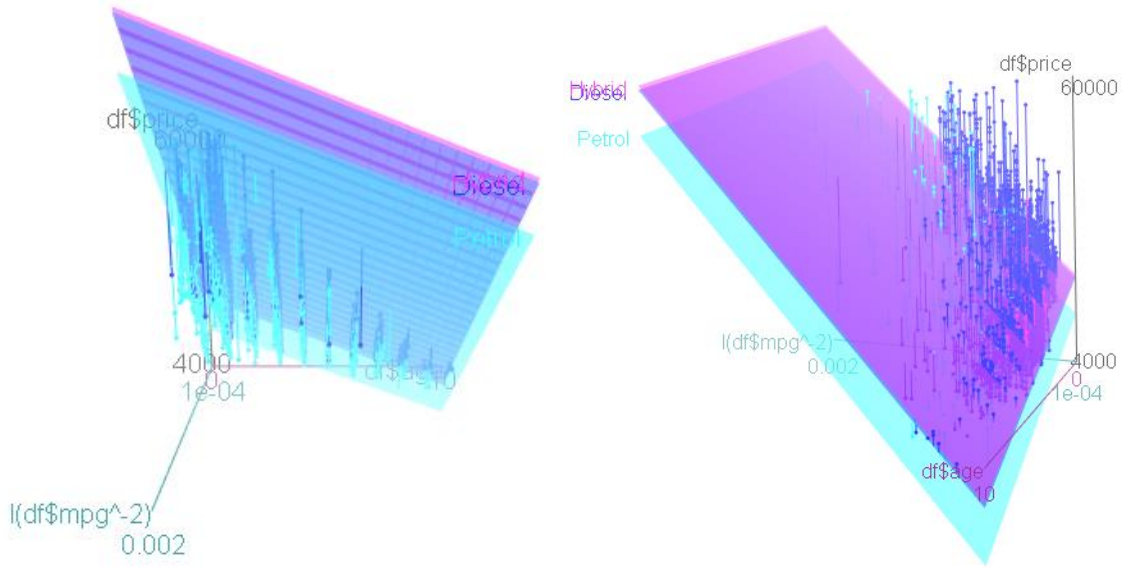

```
par(mfrow=c(1,1))

boxTidwell(df$price~I(df$age+0.1))

##  MLE of lambda Score Statistic (z)  Pr(>|z|)
##        0.64376               11.16 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  5
```

## Question 9

**Q: Are there any additional explanatory numeric variables needed to the car price? Study collinearity effects.**

Using all of the numeric variables and performing an Anova test, it has been observed that adding tax after having all the other variables already in the model is not relevant, as the variance explained by adding *mpg* already covers the amount explained by *tax*. Moreover, the collinearity test showed values slightly over the threshold (a value of 3 has been stated as the threshold) between age and mileage. Therefore, since it has been observed that age explains more variance than mileage, the latter has been removed from the model, which has resulted in the following linear model: *price ~ age + mpg*, having $R^2$=0.55. Then, additional transformations have been studied by means of the boxTidwell method. It can be observed that a transformation in *mpg* of $mpg^{-2}$ is suggested, which clearly increases the $R^2$ by a 5%. Again, collinearity has been addressed after the transformations, where no issues are detected.

The relation between the three variables can be graphically depicted below using the `scatter3d` function. It can be observed how the price is greater for cars with lower values of *age* and higher values of $1/mpg^2$, which is equivalent to lower values of *mpg*. Also, prices are higher for hybrids and diesels than for petrol cars.




```
m3 = lm(df$price~df$age+df$mileage+df$tax+df$mpg)
summary(m3)

##
## Call:
## lm(formula = df$price ~ df$age + df$mileage + df$tax + df$mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -17454  -4588   -213   3291  33804
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.735e+04  7.663e+02  61.787  < 2e-16 ***
## df$age      -2.080e+03  8.561e+01 -24.301  < 2e-16 ***
## df$mileage  -2.582e-02  8.121e-03  -3.179  0.00149 **
## df$tax       2.233e+00  2.215e+00   1.008  0.31344
## df$mpg      -3.773e+02  1.087e+01 -34.714  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6440 on 4822 degrees of freedom
## Multiple R-squared:  0.5572, Adjusted R-squared:  0.5569
## F-statistic:  1517 on 4 and 4822 DF,  p-value: < 2.2e-16
```

```
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: df$price
##                Sum Sq   Df   F value    Pr(>F)
## df$age     2.4493e+10    1  590.5220 < 2.2e-16 ***
## df$mileage 4.1928e+08    1   10.1085  0.001485 **
## df$tax     4.2156e+07    1    1.0164  0.313436
## df$mpg     4.9984e+10    1 1205.0864 < 2.2e-16 ***
## Residuals  2.0000e+11 4822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(m3)
```

```
##     df$age df$mileage     df$tax     df$mpg
##   3.013448   2.921973   1.733609   1.689732
```

```
m3 = lm(df$price~df$age+df$mpg)
summary(m3)
```

```
##
## Call:
## lm(formula = df$price ~ df$age + df$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17507  -4551   -254   3338  33834
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48178.053    459.698  104.80   <2e-16 ***
## df$age      -2301.698     53.884  -42.72   <2e-16 ***
## df$mpg       -387.300      9.135  -42.40   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6446 on 4824 degrees of freedom
## Multiple R-squared:  0.5562, Adjusted R-squared:  0.556
## F-statistic:  3023 on 2 and 4824 DF,  p-value: < 2.2e-16
```

```
vif(m3)
```

```
##   df$age   df$mpg
## 1.191526 1.191526
```

```
boxTidwell(df$price~I(df$age+0.1)+df$mpg)
```

```
##                  MLE of lambda Score Statistic (z) Pr(>|z|)
## I(df$age + 0.1)         0.9073              1.7095  0.08735 .
## df$mpg                 -2.2150             24.8495  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  5
```

```
m33 = lm(df$price~ df$age + I(df$mpg^-2))
summary(m33)
```

```
##
## Call:
## lm(formula = df$price ~ df$age + I(df$mpg^-2))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25468  -4058   -320   3027  35370
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.756e+04  2.865e+02   61.31   <2e-16 ***
## df$age       -2.250e+03  5.013e+01  -44.89   <2e-16 ***
## I(df$mpg^-2)  2.411e+07  4.686e+05   51.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6069 on 4824 degrees of freedom
## Multiple R-squared:  0.6067, Adjusted R-squared:  0.6066
## F-statistic:  3721 on 2 and 4824 DF,  p-value: < 2.2e-16
```

```
vif(m33)
```

```
##       df$age I(df$mpg^-2)
##     1.163608     1.163608
```

```
BIC(m3, m33)
```

```
##     df      BIC
## m3   4 98407.23
## m33  4 97824.13

# scatter3d(df$price ~ df$age + I(df$mpg^-2), fov=0)
```

## Question 10

**Q: After controlling by numerical variables, indicate whether the additive effect of the available factors on the price are statistically significant.**

Using the BIC method (k=log(n)) on the *step* function over a model with all the explanatory variables, both quantitative and qualitative, it can be observed that the addition of the factors is significant and thus, they are kept. *f.age* is not considered due to the high collinearity with age. It is worth mentioning this model raises the $R^2$ value slightly over 80%.
The added variable effect plot shows how none of the variables need any significant transformations, since all of the plots contain straight lines. Also, *age* and $mpg^{-2}$ are the variables which contribute more to the model, as their slopes are the more salient. Regarding factors, some levels within the same factor seem to contribute more than the others, which is expected.

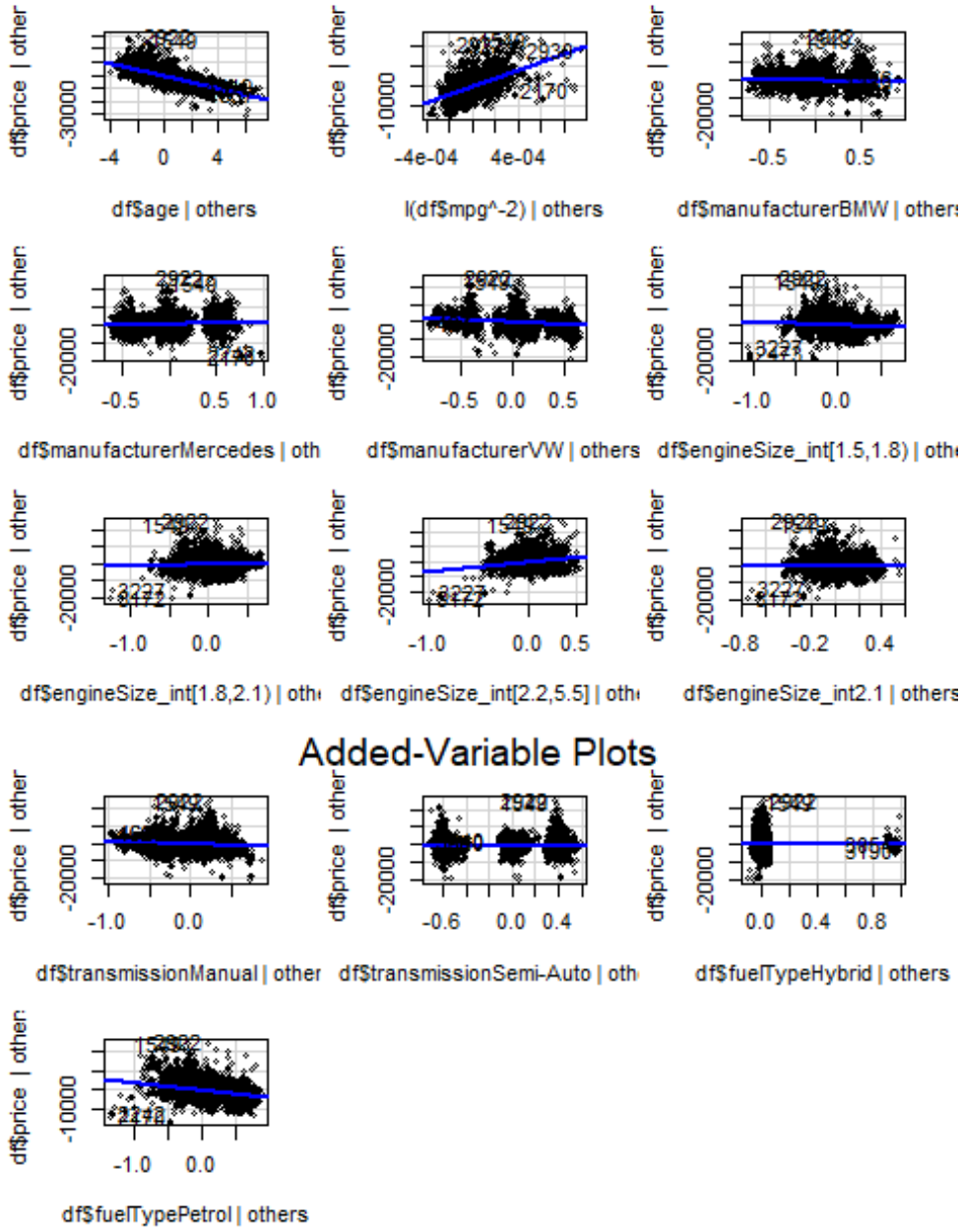
```
m12= lm(df$price~df$age+I(df$mpg^-2)+df$manufacturer+df$engineSize_int+df
$transmission+df$fuelType)
full_model = step(m12,k=log(nrow(df)))

## Start:  AIC=80353.38
## df$price ~ df$age + I(df$mpg^-2) + df$manufacturer + df$engineSize_int
+
##     df$transmission + df$fuelType
##
##                       Df  Sum of Sq        RSS   AIC
## <none>                                7.9899e+10 80353
## - df$transmission      2 2.4176e+09 8.2317e+10 80480
## - df$fuelType          2 7.6496e+09 8.7549e+10 80778
## - df$manufacturer      3 9.4113e+09 8.9310e+10 80865
## - df$engineSize_int    4 1.5543e+10 9.5442e+10 81177
## - I(df$mpg^-2)         1 3.0923e+10 1.1082e+11 81924
## - df$age               1 7.5322e+10 1.5522e+11 83550

summary(full_model)

##
## Call:
## lm(formula = df$price ~ df$age + I(df$mpg^-2) + df$manufacturer +
##     df$engineSize_int + df$transmission + df$fuelType)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19879  -2262   -197   2065  24812
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.105e+04  3.241e+02  64.927  < 2e-16 ***
## df$age                      -2.413e+03  3.582e+01 -67.359  < 2e-16 ***
## I(df$mpg^-2)                 2.020e+07  4.681e+05  43.160  < 2e-16 ***
```

```
## df$manufacturerBMW          -8.524e+02  1.879e+02  -4.535 5.90e-06 ***
## df$manufacturerMercedes      1.294e+03  1.967e+02   6.581 5.18e-11 ***
## df$manufacturerVW           -3.042e+03  1.693e+02 -17.972  < 2e-16 ***
## df$engineSize_int[1.5,1.8) -1.351e+03  2.306e+02  -5.857 5.01e-09 ***
## df$engineSize_int[1.8,2.1)  1.407e+03  2.499e+02   5.630 1.90e-08 ***
## df$engineSize_int[2.2,5.5]  6.998e+03  3.609e+02  19.389  < 2e-16 ***
## df$engineSize_int2.1        1.964e+01  3.514e+02   0.056 0.955435
## df$transmissionManual      -1.383e+03  1.790e+02  -7.727 1.33e-14 ***
## df$transmissionSemi-Auto    5.946e+02  1.533e+02   3.880 0.000106 ***
## df$fuelTypeHybrid           1.094e+03  5.211e+02   2.100 0.035756 *
## df$fuelTypePetrol          -4.006e+03  1.922e+02 -20.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4074 on 4813 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8226
## F-statistic:  1723 on 13 and 4813 DF,  p-value: < 2.2e-16
```

```
anova(m3,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: df$price ~ df$age + df$mpg
## Model 2: df$price ~ df$age + I(df$mpg^-2) + df$manufacturer + df$engin
eSize_int +
##     df$transmission + df$fuelType
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4824 2.0046e+11
## 2   4813 7.9899e+10 11 1.2057e+11 660.25 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
avPlots(m12)
```


Added-Variable Plots
df$price | others
df$age | others
I(df$mpg^-2) | others
df$manufacturerBMW | others
df$manufacturerMercedes | oth
df$manufacturerVW | others
df$engineSize_int[1.5,1.8) | oth
df$engineSize_int[1.8,2.1) | oth
df$engineSize_int[2.2,5.5] | oth
df$engineSize_int2.1 | others
df$transmissionManual | other
df$transmissionSemi-Auto | oth
df$fuelTypeHybrid | others
df$fuelTypePetrol | others

## Question 11

**Q: Select the best model available so far. Interpret the equations that relate the explanatory variables to the answer (rate).**

As observed in the last exercise, where the best model in terms of the BIC criterion is obtained, the model which includes almost all the explanatory variables is capable of explaining 80% of the price variability. None of the previous ones has the $R^2$ that high and thus, is selected as the best one: *price ~ age + $mpg^{-2}$ + manufacturer + engineSize_int + transmission + fuelType*.

Regarding the residual plots, some issues can be observed:

- A curve pattern in both Residual-vs-Fitted and Scale-Location plots can be detected, which suggests some problems with the linearity and homoscedasticity assumption.
- Big deviations from the normal behavior of residuals.
- There is a group of observations with high leverage that could be influential.

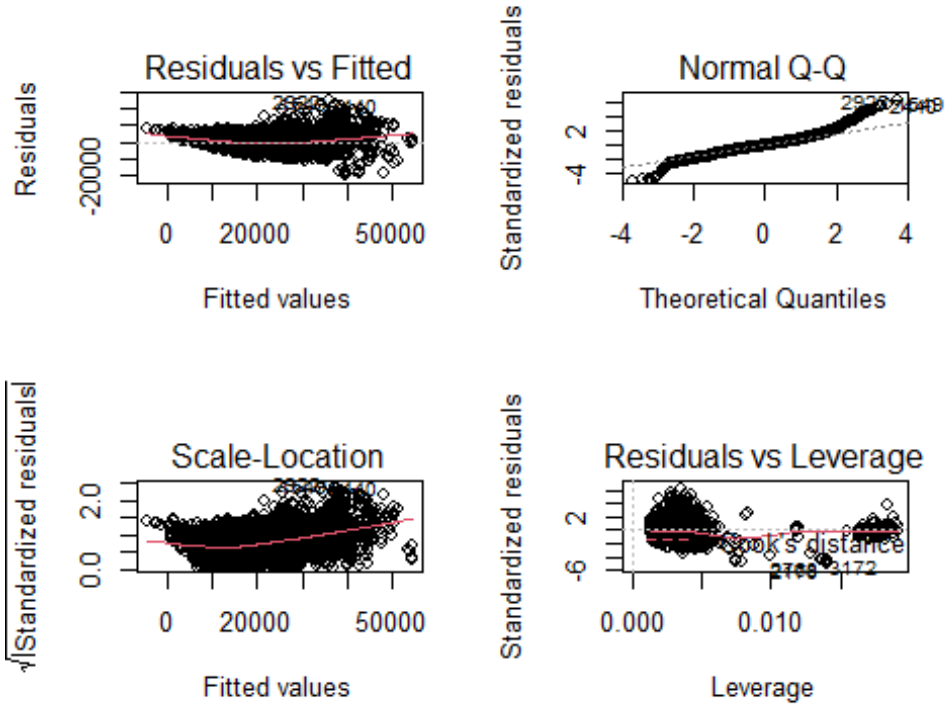
Regarding the computed model, it can be seen that, as before, the coefficient for *age* is negative, as it is negatively correlated with price. For the same reason, *$mpg^{-2}$*'s coefficient is positive. Taking these numerical values into account, the reference levels of the factors are Automatic (*transmission*), Diesel (*fuelType*), small Engine Size (*(0, 1.5)*), and Audi (manufacturer) cars. The most expensive cars seem to be those with a semi-automatic transmission, manufactured by Mercedes, with a big engine, and fueled by a hybrid system. On the contrary, the less expensive cars seem to be those with a manual transmission, manufactured by BMW, with an engine between 1.5 and 1.8 $dm^3$, and fueled by petrol. One would expect VW to be the manufacturer associated with the cheapest cars, but for some reason this is not the case. It might be due to the fact that most cars manufactured by VW already have the above-mentioned characteristics that decrease the price.

```
summary(full_model)
```

```
##
## Call:
## lm(formula = df$price ~ df$age + I(df$mpg^-2) + df$manufacturer +
##     df$engineSize_int + df$transmission + df$fuelType)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19879  -2262   -197   2065  24812
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.105e+04  3.241e+02  64.927  < 2e-16 ***
## df$age                     -2.413e+03  3.582e+01 -67.359  < 2e-16 ***
## I(df$mpg^-2)                2.020e+07  4.681e+05  43.160  < 2e-16 ***
## df$manufacturerBMW         -8.524e+02  1.879e+02  -4.535 5.90e-06 ***
```

```
## df$manufacturerMercedes     1.294e+03  1.967e+02   6.581 5.18e-11 ***
## df$manufacturerVW          -3.042e+03  1.693e+02 -17.972  < 2e-16 ***
## df$engineSize_int[1.5,1.8) -1.351e+03  2.306e+02  -5.857 5.01e-09 ***
## df$engineSize_int[1.8,2.1)  1.407e+03  2.499e+02   5.630 1.90e-08 ***
## df$engineSize_int[2.2,5.5]  6.998e+03  3.609e+02  19.389  < 2e-16 ***
## df$engineSize_int2.1        1.964e+01  3.514e+02   0.056 0.955435
## df$transmissionManual      -1.383e+03  1.790e+02  -7.727 1.33e-14 ***
## df$transmissionSemi-Auto    5.946e+02  1.533e+02   3.880 0.000106 ***
## df$fuelTypeHybrid           1.094e+03  5.211e+02   2.100 0.035756 *
## df$fuelTypePetrol          -4.006e+03  1.922e+02 -20.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4074 on 4813 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8226
## F-statistic:  1723 on 13 and 4813 DF,  p-value: < 2.2e-16
```

```
# pairwise.wilcox.test(df$age, df$manufacturer, alternative = "less")

# pairwise.wilcox.test(df$mileage, df$manufacturer, alternative = "less")

# pairwise.wilcox.test(df$mileage, df$manufacturer, alternative = "less")

# table(df$manufacturer, df$transmission)

# table(df$manufacturer, df$engineSize_int)

# table(df$manufacturer, df$fuelType)

par(mfrow=c(2,2))
plot(full_model)
```




```
par(mfrow=c(1,1))
```

## Question 12

**Q: Study the model that relates the logarithm of the price to the numerical variables.**

As previously suggested by the boxcox method, a model with a natural logarithm transformation on the response variable is assessed. The model is found to be statistically different and more suitable for modelling the price than the one without the transformed response variable. Moreover, with the Anova test it can be observed that contrary to the previous model, keeping the variable tax is statistically significant. Additionally, as seen before, there is collinearity between age and mileage. Hence, the latter is also removed by the same criteria as before. Regarding the transformations, it can be observed that the same power transformation to the *mpg* variable is suggested, and applying it results in a better model. In this new model no transformation is suggested for the variable *tax*, while for *age*, if the suggested transformation is applied the resulting model is worse.

Regarding the residuals, improvements in their overall behavior can be clearly observed: both the Residual-vs -Fitted and Scale-Location now display straight lines, where it can be seen that linearity and homoscedasticity assumptions are met. Moreover, the Normal Q-Q plot displays almost all the points over the straight line, although the left tail seems to slightly deviate from the normal behavior. Influence of observations is also different, as almost all of them are concentrated in the same zone. Some observations still have a high leverage.

```
mlog = lm(log(df$price)~df$age+df$mileage+df$tax+df$mpg)
summary(mlog)
```

```
##
## Call:
## lm(formula = log(df$price) ~ df$age + df$mileage + df$tax + df$mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04346 -0.17717  0.01573  0.19340  0.88771
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.076e+01  3.317e-02 324.501  < 2e-16 ***
## df$age      -1.069e-01  3.705e-03 -28.860  < 2e-16 ***
## df$mileage  -1.795e-06  3.515e-07  -5.107  3.4e-07 ***
## df$tax       1.035e-03  9.587e-05  10.797  < 2e-16 ***
## df$mpg      -1.293e-02  4.704e-04 -27.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2788 on 4822 degrees of freedom
```

```
## Multiple R-squared:  0.6182, Adjusted R-squared:  0.6179
## F-statistic:  1952 on 4 and 4822 DF,  p-value: < 2.2e-16

Anova(mlog)

## Anova Table (Type II tests)
##
## Response: log(df$price)
##             Sum Sq   Df F value    Pr(>F)
## df$age      64.72    1 832.871 < 2.2e-16 ***
## df$mileage   2.03    1  26.078 3.404e-07 ***
## df$tax       9.06    1 116.581 < 2.2e-16 ***
## df$mpg      58.75    1 756.040 < 2.2e-16 ***
## Residuals  374.68 4822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

vif(mlog)

##     df$age df$mileage     df$tax     df$mpg
##   3.013448   2.921973   1.733609   1.689732

mlog = lm(log(df$price)~df$age+df$tax+df$mpg)
summary(mlog)

##
## Call:
## lm(formula = log(df$price) ~ df$age + df$tax + df$mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03594 -0.18054  0.01472  0.19342  0.89307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.078e+01  3.312e-02  325.41   <2e-16 ***
## df$age      -1.213e-01  2.425e-03  -50.00   <2e-16 ***
## df$tax       1.035e-03  9.611e-05   10.77   <2e-16 ***
## df$mpg      -1.323e-02  4.680e-04  -28.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2795 on 4823 degrees of freedom
## Multiple R-squared:  0.6162, Adjusted R-squared:  0.6159
## F-statistic:  2581 on 3 and 4823 DF,  p-value: < 2.2e-16

boxTidwell(log(df$price)~I(df$age+0.1)+I(df$tax+0.1)+df$mpg)

##                  MLE of lambda Score Statistic (z)  Pr(>|z|)
## I(df$age + 0.1)         1.4385             -7.4870 7.047e-14 ***
## I(df$tax + 0.1)         0.8851             -3.3263   0.00088 ***
## df$mpg                 -2.0811             19.1904 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  8
```

```
mlog1 = lm(log(df$price)~df$age+df$tax+I(df$mpg^-2))
```

```
anova(mlog,mlog1)
```

```
## Analysis of Variance Table
##
## Model 1: log(df$price) ~ df$age + df$tax + df$mpg
## Model 2: log(df$price) ~ df$age + df$tax + I(df$mpg^-2)
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   4823 376.71
## 2   4823 352.30  0    24.406
```

```
BIC(mlog, mlog1)
```

```
##       df      BIC
## mlog   5 1429.569
## mlog1  5 1106.250
```

```
boxTidwell(log(df$price)~I(df$age+0.1)+I(df$tax+0.1)+I(df$mpg^-2))
```

```
##                 MLE of lambda Score Statistic (z)  Pr(>|z|)
## I(df$age + 0.1)       1.43844             -7.1730 7.336e-13 ***
## I(df$tax + 0.1)       0.88541             -0.1321    0.8949
## I(df$mpg^-2)          1.04042              0.6505    0.5154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  6
```

```
mlog2 = lm(log(df$price)~I(df$age^2)+df$tax+I(df$mpg^-2))
```

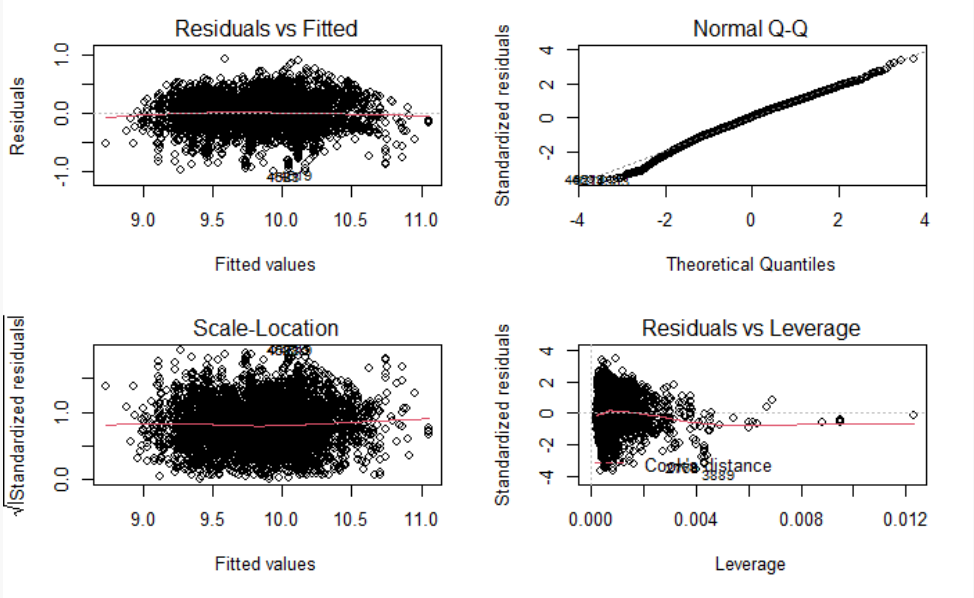
```
anova(mlog1,mlog2)
```

```
## Analysis of Variance Table
##
## Model 1: log(df$price) ~ df$age + df$tax + I(df$mpg^-2)
## Model 2: log(df$price) ~ I(df$age^2) + df$tax + I(df$mpg^-2)
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   4823 352.30
## 2   4823 353.24  0  -0.93653
```

```
BIC(mlog1, mlog2)
```

```
##       df      BIC
## mlog1  5 1106.250
## mlog2  5 1119.065
```

```
par(mfrow=c(2,2))
plot(mlog1)
```




```
par(mfrow=c(1,1))
```

## Question 13

**Q: Once explanatory numerical variables are included in the model, are there any main effects from factors needed?**

Using the step function as before, it can be seen that all the additional factors are worth keeping. Moreover, collinearity effects are not observed. With the addition of factors $R^2$ increases to 86%. The added variable effect plot is depicted in the following question.

```
mlog = lm(log(df$price)~df$age+df$tax+I(df$mpg^-2)+df$engineSize_int+df$m
anufacturer+df$transmission+df$fuelType)
logmodel = step(mlog, k = log(nrow(df)))
```

```
## Start:  AIC=-17154.07
## log(df$price) ~ df$age + df$tax + I(df$mpg^-2) + df$engineSize_int +
##     df$manufacturer + df$transmission + df$fuelType
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             134.54 -17154
## - df$tax               1     1.835 136.37 -17097
## - df$fuelType          2     6.772 141.31 -16934
## - df$transmission      2     9.799 144.34 -16832
## - df$manufacturer      3    23.705 158.24 -16396
## - df$engineSize_int    4    24.682 159.22 -16375
## - I(df$mpg^-2)         1    27.742 162.28 -16258
## - df$age               1   199.196 333.73 -12777
```

```
summary(logmodel)
```

```
##
## Call:
## lm(formula = log(df$price) ~ df$age + df$tax + I(df$mpg^-2) +
##     df$engineSize_int + df$manufacturer + df$transmission + df$fuelTyp
e)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63731 -0.10508 -0.00059  0.10748  0.56951
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   9.898e+00  1.438e-02 688.153  < 2e-16 ***
## df$age                       -1.313e-01  1.555e-03 -84.407  < 2e-16 ***
## df$tax                        4.440e-04  5.481e-05   8.101 6.85e-16 ***
## I(df$mpg^-2)                  6.135e+02  1.948e+01  31.500  < 2e-16 ***
## df$engineSize_int[1.5,1.8)    4.748e-02  9.481e-03   5.008 5.71e-07 ***
## df$engineSize_int[1.8,2.1)    1.986e-01  1.036e-02  19.170  < 2e-16 ***
## df$engineSize_int[2.2,5.5]    3.908e-01  1.496e-02  26.125  < 2e-16 ***
## df$engineSize_int2.1          1.736e-01  1.445e-02  12.015  < 2e-16 ***
## df$manufacturerBMW           -8.296e-02  7.715e-03 -10.753  < 2e-16 ***
```

```
## df$manufacturerMercedes     2.224e-02  8.073e-03   2.755 0.005892 **
## df$manufacturerVW          -1.714e-01  6.946e-03 -24.674  < 2e-16 ***
## df$transmissionManual      -1.033e-01  7.376e-03 -14.010  < 2e-16 ***
## df$transmissionSemi-Auto    2.073e-02  6.291e-03   3.295 0.000991 ***
## df$fuelTypeHybrid           1.355e-01  2.151e-02   6.298 3.28e-10 ***
## df$fuelTypePetrol          -1.050e-01  7.920e-03 -13.261  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1672 on 4812 degrees of freedom
## Multiple R-squared:  0.8629, Adjusted R-squared:  0.8625
## F-statistic:  2164 on 14 and 4812 DF,  p-value: < 2.2e-16
```

```
vif(logmodel)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## df$age            1.475029  1        1.214508
## df$tax            1.575143  1        1.255047
## I(df$mpg^-2)      2.647582  1        1.627139
## df$engineSize_int 5.803429  4        1.245835
## df$manufacturer   2.129562  3        1.134266
## df$transmission   1.638246  2        1.131344
## df$fuelType       2.688830  2        1.280533
```

## Question 14

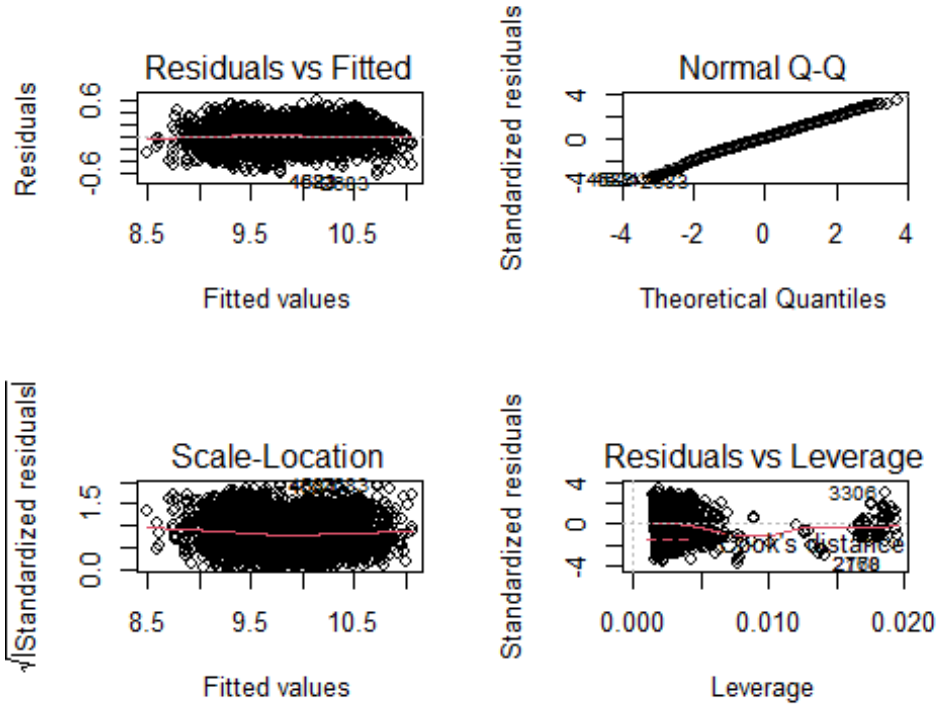
**Q: Graphically assess the best model obtained so far.**

It cannot be observed a pattern in the residuals nor in the standardized residuals as it was observed in the previous best model, hence the linear and homoscedastic assumptions hold. However, the model has normality problems in the left tail and a small group of observations that can be influential.

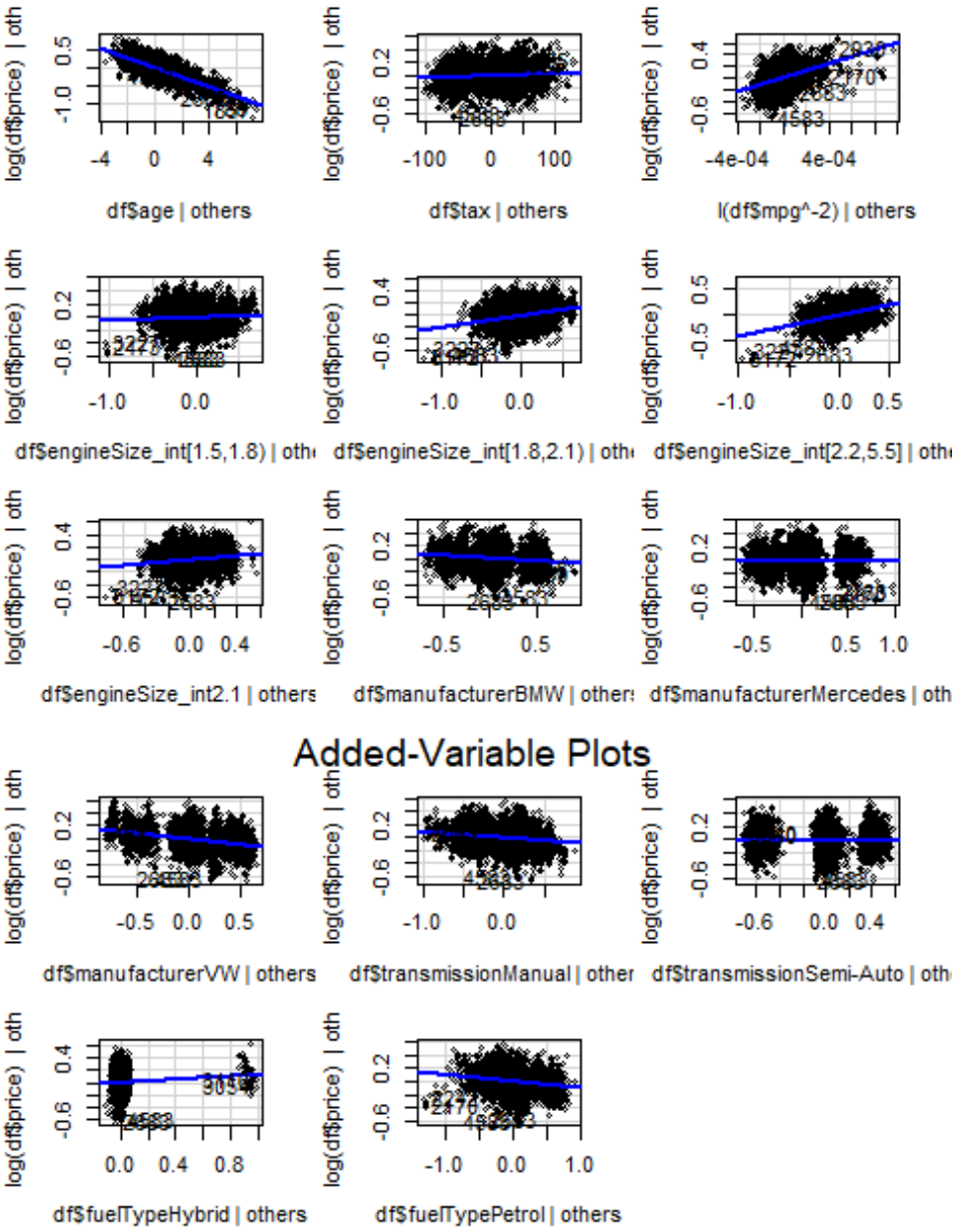
The added variable effect plot shows how none of the variables need any significant transformations, since all of the plots contain straight lines. Also, *age* and $mpg^{-2}$ are the variables which contribute more to the model, as their slopes are the more salient. *tax* seems to not contribute much to the model. In fact, according to the previous BIC step it is the one that contributes the least. Regarding factors, *engineSize_int* seems to be the one that contributes the most. Also, some levels within the same factor seem to provide more information than the others, which is expected.

Marginal plots from the model display how well it fits the points in comparison to a smoother. In this case, the model does not deviate from the smoothed line at all and thus, it can be said that the model properly reproduces the data.

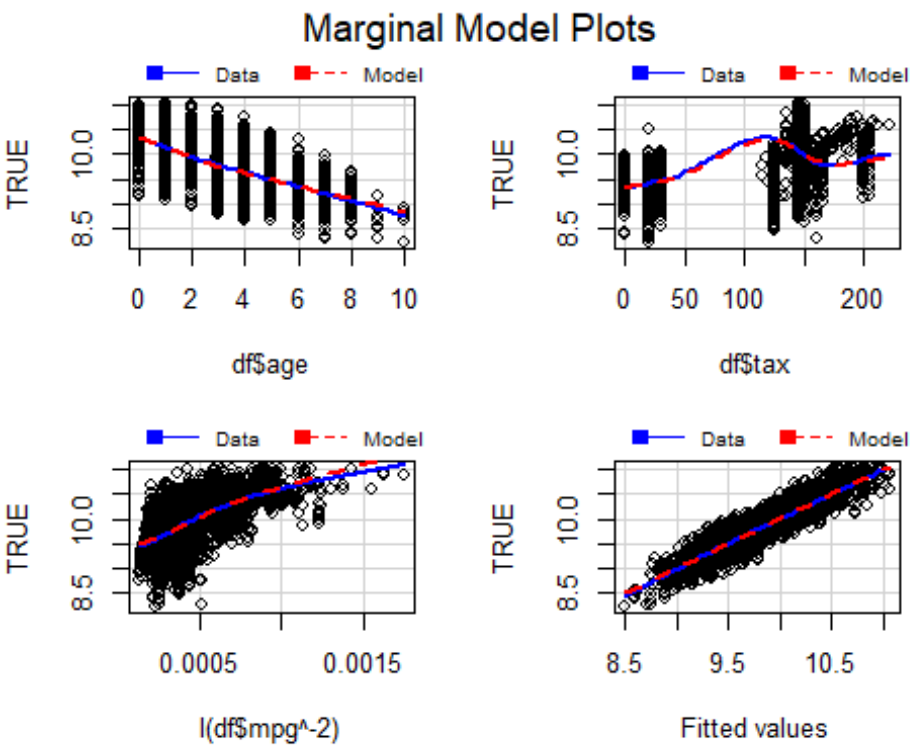
```
par(mfrow=c(2,2))
plot(logmodel)
```




```
par(mfrow=c(1,1))

avPlots(logmodel)
```




```
marginalModelPlots(logmodel, terms = ~df$age+df$tax+I(df$mpg^-2))
```


Marginal Model Plots
Data
Model
TRUE
8.5
10.0
0
2
4
6
8
10
df$age
0
50
100
200
df$tax
0.0005
0.0015
I(df$mpg^-2)
8.5
9.5
10.5
Fitted values

## Question 15

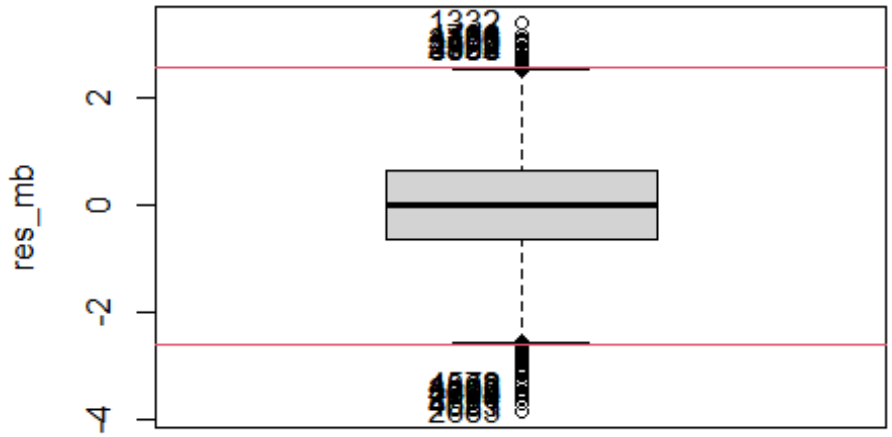
**Q: Assess the presence of outliers in the studentized residuals at a 99% confidence level. Indicate what those observations are.**

It can be observed that the inferential method studied and the boxplot descriptive method are equally restrictive, giving a total of 69 outliers. First of all, it can be seen that the model *'Up'*, which is manufactured by VW, is overrepresented in the outliers, as it accounts for over 30% of them, while in the original dataset it is barely 1%. A general description of the rest of outliers would be manual cars, which are fueled by petrol, and have a lower price. That is, there are no observed differences in engine size, age, tax, and mpg.

```
mb = logmodel # mb is best model
n = dim(df)[1]
p = mb$rank
res_mb = rstudent(mb)
cut_off = qt(0.995,n-p-1)

Boxplot(res_mb)
```

```
##  [1] 2683 4583 4621 2788 4594 4605 4619 3182 4802 4572 1332 4736  192
3306 4475
## [16] 2392 2922 3887 4754  808
```

```
abline(h=cut_off,col=2)
abline(h=-cut_off,col=2)
```




```
nrow(df[which(abs(res_mb)>cut_off),])

## [1] 69

aux = df[which(abs(res_mb)>cut_off),]
summary(aux)

##      price          mileage           tax              mpg
##  Min.   : 3995   Min.   :   10   Min.   :  0.0   Min.   :28.30
##  1st Qu.: 7699   1st Qu.: 4215   1st Qu.:145.0   1st Qu.:46.30
##  Median :10791   Median :14215   Median :145.0   Median :54.30
##  Mean   :19037   Mean   :23028   Mean   :123.3   Mean   :52.43
##  3rd Qu.:22342   3rd Qu.:38244   3rd Qu.:145.0   3rd Qu.:62.80
##  Max.   :60849   Max.   :83587   Max.   :204.4   Max.   :83.10
##       age           model           transmission         fuelType
##  Min.   :0.000   Length:77          Length:77          Length:77
##  1st Qu.:1.000   Class :character   Class :character   Class :characte
r
##  Median :3.000   Mode  :character   Mode  :character   Mode  :characte
r
##  Mean   :2.792
##  3rd Qu.:4.000
##  Max.   :9.000
##  manufacturer       model_collapsed    engineSize_int        f.age
##  Length:77          Length:77          Length:77          [0,1) : 8
##  Class :character   Class :character   Class :character   [1,3) :26
##  Mode  :character   Mode  :character   Mode  :character   [3,4) :18
```

```
##                                                                 [4,10]:25
##
##

# table(aux$model)

prop.table(table(aux$transmission)); prop.table(table(df$transmission))

##
## Automatic    Manual Semi-Auto
## 0.1688312 0.6363636 0.1948052

##
## Automatic    Manual Semi-Auto
## 0.2475658 0.3755956 0.3768386

prop.table(table(aux$fuelType)); prop.table(table(df$fuelType))

##
##     Diesel     Hybrid     Petrol
## 0.33766234 0.01298701 0.64935065

##
##     Diesel     Hybrid     Petrol
## 0.58400663 0.01346592 0.40252745

prop.table(table(aux$engineSize_int)); prop.table(table(df$engineSize_int
))

##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.54545455 0.06493506 0.18181818 0.15584416 0.05194805

##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.18127201 0.18168635 0.42614460 0.12388647 0.08701057

prop.table(table(aux$manufacturer)); prop.table(table(df$manufacturer))

##
##       Audi        BMW   Mercedes         VW
## 0.14285714 0.09090909 0.19480519 0.57142857

##
##      Audi       BMW  Mercedes        VW
## 0.2276776 0.2034390 0.2581313 0.3107520

mean(aux$price); mean(df$price)

## [1] 19036.57

## [1] 21143.97

mean(aux$age); mean(df$age)
```

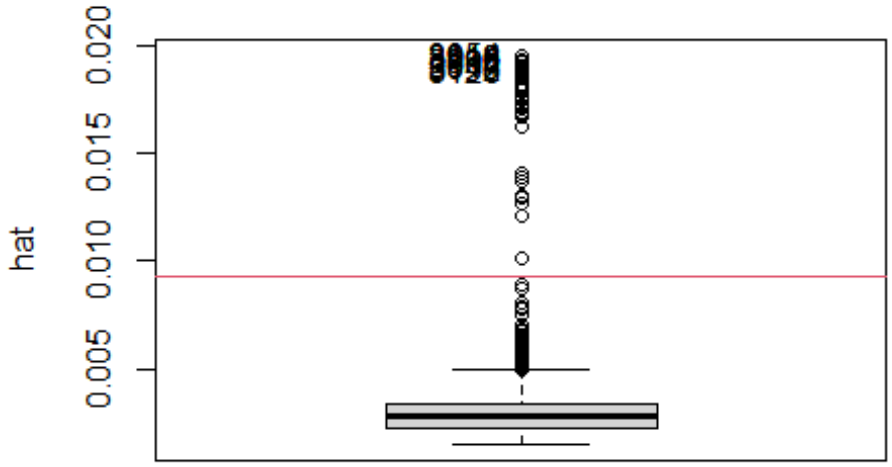
```
## [1] 2.792208

## [1] 2.743903

mean(aux$tax); mean(df$tax)

## [1] 123.2564

## [1] 117.7828

mean(aux$mpg); mean(df$mpg)

## [1] 52.43348

## [1] 53.49462
```

## Question 16

**Q: Study the presence of a priori influential data observations, indicating their number according to the criteria studied in class.**

The a priori influential observations have been calculated by inspecting the leverage through the use of the hat values, considering the dataset to be large enough so that the threshold is 3p/n. In this case, 77 values with a significantly high leverage are obtained. The detected outliers can be clearly seen to be hybrid cars, hence newer, not manual, more expensive and with small engine size.

```
hat = hatvalues(mb)
hat_cut = 3*p/n #if we assume the dataset is small, then 2p/n

Boxplot(hat)

##  [1] 3054 3912 3915 3116 3190 3890 3914 3043 3192 3126

abline(h=hat_cut,col=2)
```




```
sum(hat>hat_cut)

## [1] 77

aux = df[which(hat>hat_cut),]

table(aux$model)
```

```
##
##  2 Series  3 Series  5 Series   A Class        A3   C Class   E Class
GLE Class
##         3        13         7         6         3        19         1
1
##      Golf        i3    Passat        Q7   S Class        X3        X4
X5
##        12         2         4         1         1         3         1
1
```

```
prop.table(table(aux$transmission)); prop.table(table(df$transmission))
```

```
##
##  Automatic     Manual  Semi-Auto
## 0.50000000 0.05128205 0.44871795
```

```
##
## Automatic    Manual Semi-Auto
## 0.2475658 0.3755956 0.3768386
```

```
prop.table(table(aux$fuelType)); prop.table(table(df$fuelType))
```

```
##
##     Diesel     Hybrid     Petrol
## 0.01282051 0.83333333 0.15384615
```

```
##
##     Diesel     Hybrid     Petrol
## 0.58400663 0.01346592 0.40252745
```

```
prop.table(table(aux$engineSize_int)); prop.table(table(df$engineSize_int
))
```

```
##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.30769231 0.07692308 0.43589744 0.08974359 0.08974359
```

```
##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.18127201 0.18168635 0.42614460 0.12388647 0.08701057
```

```
prop.table(table(aux$manufacturer)); prop.table(table(df$manufacturer))
```

```
##
##       Audi        BMW   Mercedes         VW
## 0.05128205 0.38461538 0.35897436 0.20512821
```

```
##
##      Audi       BMW  Mercedes        VW
## 0.2276776 0.2034390 0.2581313 0.3107520
```

```
mean(aux$price); mean(df$price)
```

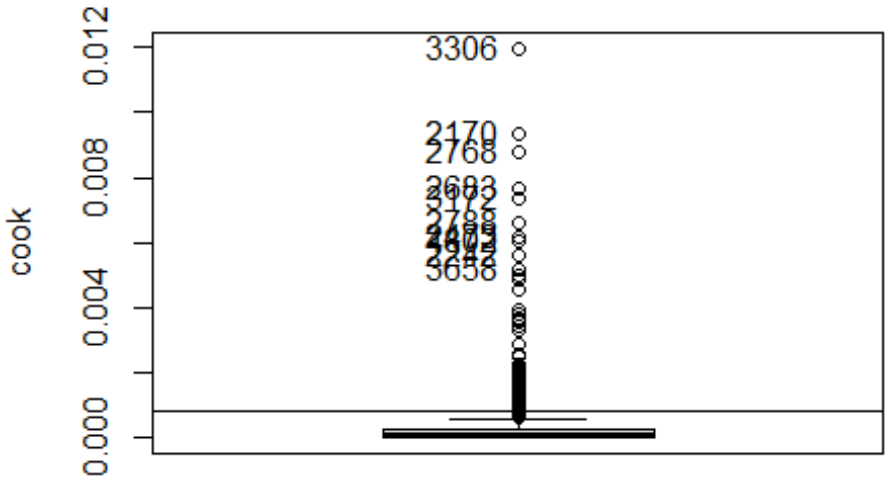
```
## [1] 24410.14
## [1] 21143.97
mean(aux$age); mean(df$age)
## [1] 2.589744
## [1] 2.743903
mean(aux$tax); mean(df$tax)
## [1] 82.56954
## [1] 117.7828
mean(aux$mpg); mean(df$mpg)
## [1] 53.87379
## [1] 53.49462
```

## Question 17

**Q: Study the presence of a posteriori influential values, indicating the criteria studied in class and the actual atypical observations.**

The a posteriori influential data was studied by calculating the Cook's distance of the values, which takes into account leverage and discrepancy, selecting the values whose Cook's distance is greater than 4/(n-p) to be examined. In our case 247 values above the threshold were obtained, which is a number greater than the one found for student residuals and a priori influential data. However, it must be noted that if the sample is assumed to be large enough and the Chatterjee and Hadi cut-off is discarded, there are no observations with a Cook's distance greater than 0.5.

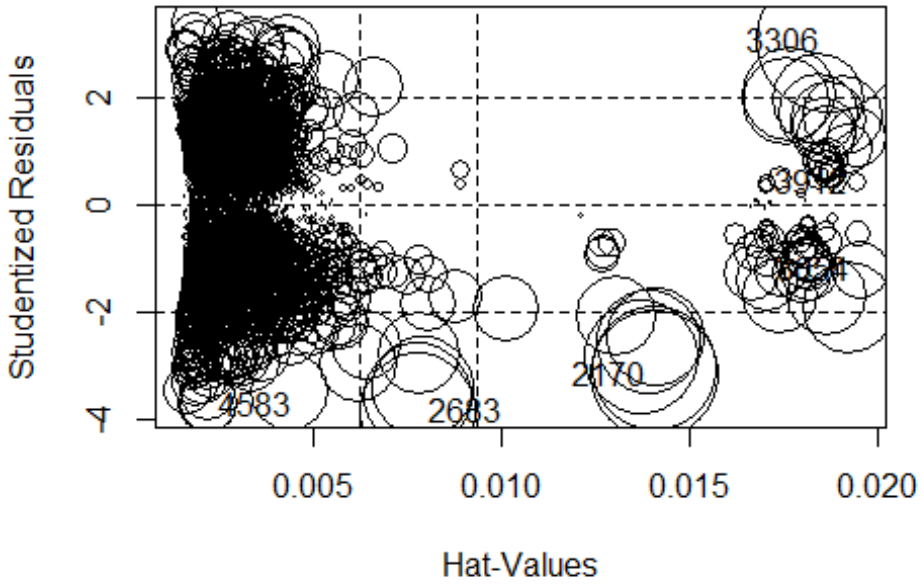
```
cook = cooks.distance(mb)
cook_cut = 4/(n-p)

Boxplot(cook)

##  [1] 3306 2170 2768 2683 3172 2788 2473 4802 2242 3658

abline(h=cook_cut)
```




```
nrow(df[which(cook>cook_cut),])

## [1] 247
```

```
aux = df[which(cook>cook_cut),]

table(aux$model)

##
##        00  1 Series  2 Series  3 Series  4 Series  7 Series  8 Series
A Class
##         1         9         4         5         2         4         2
12
##        A3        A4        A5        A6        A8    Amarok   B Class
Beetle
##        23         4         2         2         2         5         4
1
##   C Class Caravelle        CC  CL Class   E Class GLC Class GLE Class
Golf
##        24         4         1         5         6         3         5
15
##        i3     Jetta   M Class        M4    Passat      Polo        Q3
Q5
##         2         1         1         1         7         6         4
2
##        Q7   S Class  Scirocco       SQ7     T-Roc    Tiguan    Touran
Up
##         4         9         1         1         3         3         6
29
##   V Class   X-CLASS        X1        X2        X3        X4        X5
X6
##         4         2         2         1         2         2         5
3
##        X7
##         1

prop.table(table(aux$transmission)); prop.table(table(df$transmission))

##
## Automatic    Manual Semi-Auto
## 0.3765182 0.3400810 0.2834008

##
## Automatic    Manual Semi-Auto
## 0.2475658 0.3755956 0.3768386

prop.table(table(aux$fuelType)); prop.table(table(df$fuelType))

##
##    Diesel    Hybrid    Petrol
## 0.4493927 0.1174089 0.4331984

##
##     Diesel     Hybrid     Petrol
## 0.58400663 0.01346592 0.40252745
```

```
prop.table(table(aux$engineSize_int)); prop.table(table(df$engineSize_int
))

##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.32793522 0.08097166 0.19433198 0.27935223 0.11740891

##
##  [0.0,1.5)  [1.5,1.8)  [1.8,2.1)  [2.2,5.5]        2.1
## 0.18127201 0.18168635 0.42614460 0.12388647 0.08701057

prop.table(table(aux$manufacturer)); prop.table(table(df$manufacturer))

##
##      Audi       BMW  Mercedes        VW
## 0.1781377 0.1821862 0.3076923 0.3319838

##
##      Audi       BMW  Mercedes        VW
## 0.2276776 0.2034390 0.2581313 0.3107520

mean(aux$price); mean(df$price)

## [1] 22436.2

## [1] 21143.97

mean(aux$age); mean(df$age)

## [1] 3.186235

## [1] 2.743903

mean(aux$tax); mean(df$tax)

## [1] 122.0308

## [1] 117.7828

mean(aux$mpg); mean(df$mpg)

## [1] 50.98647

## [1] 53.49462

df[which(cook>0.5),]

##  [1] price           mileage         tax             mpg
##  [5] age             model           transmission    fuelType
##  [9] manufacturer    model_collapsed engineSize_int  f.age
## <0 rows> (or 0-length row.names)

influencePlot(mb)
```




```
##          StudRes         Hat        CookD
## 2170 -3.1392777 0.014087192 0.0093703476
## 2683 -3.8317824 0.007770501 0.0076438853
## 3054 -1.1907084 0.019496065 0.0018792280
## 3306  3.0718612 0.018693575 0.0119629673
## 3912  0.4321114 0.019444591 0.0002468882
## 4583 -3.6994050 0.002202656 0.0020087846
```

## Question 18

**Q: Given a 5-year old car, the rest of numerical variables on the mean and factors on the reference level, what would be the expected price with a 95% confidence interval?**

The predicted price has been of 13844.91£, and the confidence interval is CI = (10017.85£, 19133.99£). It can be seen that the prediction is not very narrow and that it is nearly at the middle of the average price for 5-year old cars, which is 13919.33£. Hence although the prediction output is coherent, the interval range is quite big to be satisfied with it.

```
mpred = lm(log(price)~age+mpg+engineSize_int+manufacturer+transmission+fu
elType,data=df)
pred_df<-data.frame(age=5,tax=mean(df$tax),mpg=mean(df$mpg),engineSize_in
t='[0.0,1.5)',manufacturer='Audi',
                    transmission='Automatic',fuelType='Diesel')
exp(predict(mpred,pred_df,conf=0.95,interval=c("pred"),se.fit=T)$fit)

##        fit      lwr      upr
## 1 13844.91 10017.85 19133.99
```

## Question 19

**Q: Summarize what you have learned by working with this interesting real dataset.**

The original dataset was far from being adequate for modeling: outliers were found in the vast majority of the variables, including the response one, and some initial transformations had to be done to the dataset in order to increase its usability. The modelling has been explored in an evolutive manner, assessing variables of each kind and exploring their transformations and interactions starting from a basic model and building up to a more complex one. With this, it has been seen how even though there has been a previous data processing step, the initial models had some undesired behavior on the residuals, which has been gradually corrected as transformations and variables were added to the model. Concretely, the potential of Box Cox and Box Tidwell methods when suggesting transformations has been noticed. Moreover, the convenience and usability of performing analysis of variance methods to assess the differences between models and the usage of the step function has also been greatly noted. Additionally, assessing the model in terms of the influence of the observations, both a priori and posteriori, is found to be really interesting.

Overall, this project has served to note that, even though linear modeling could sound a non-complex problem, a great number of statistical methods are to be taken into account. Sound mathematical background has been developed throughout the project, showing us the real potential and capabilities of this modelling techniques.