Midterm

Alexis Laks

30 octobre 2018

This is my appendix, you will find all my thoughts detailed and illustrated by graphs and summaries. Please check on the code if you doubt some of my results, and contact me if you can't figure it out just by going through the rmd.

Introduction

We are given the task to analyse a dataset (realestate) containing various information on house sales such as the price at which it was sold, the location of the house, various characteristics (pool, garages, etc.) and use this information to create a predictor of sale prices. The idea is to use past data to analyse the variation and links that exist between a set of variables and our outcome variable of interest, the end result being a function that takes in a similar set of characteristics and yields an estimated sale price with a certain level of accuracy.

Exploratory data analysis

First, let's check how are data set is structured :

Data Structure:

The dataset realestate contains 11 variables plus one ID variable to distinguish each of the 522 observations. From the glimpse and head functions we clearly distinguish the information transmitted by each:

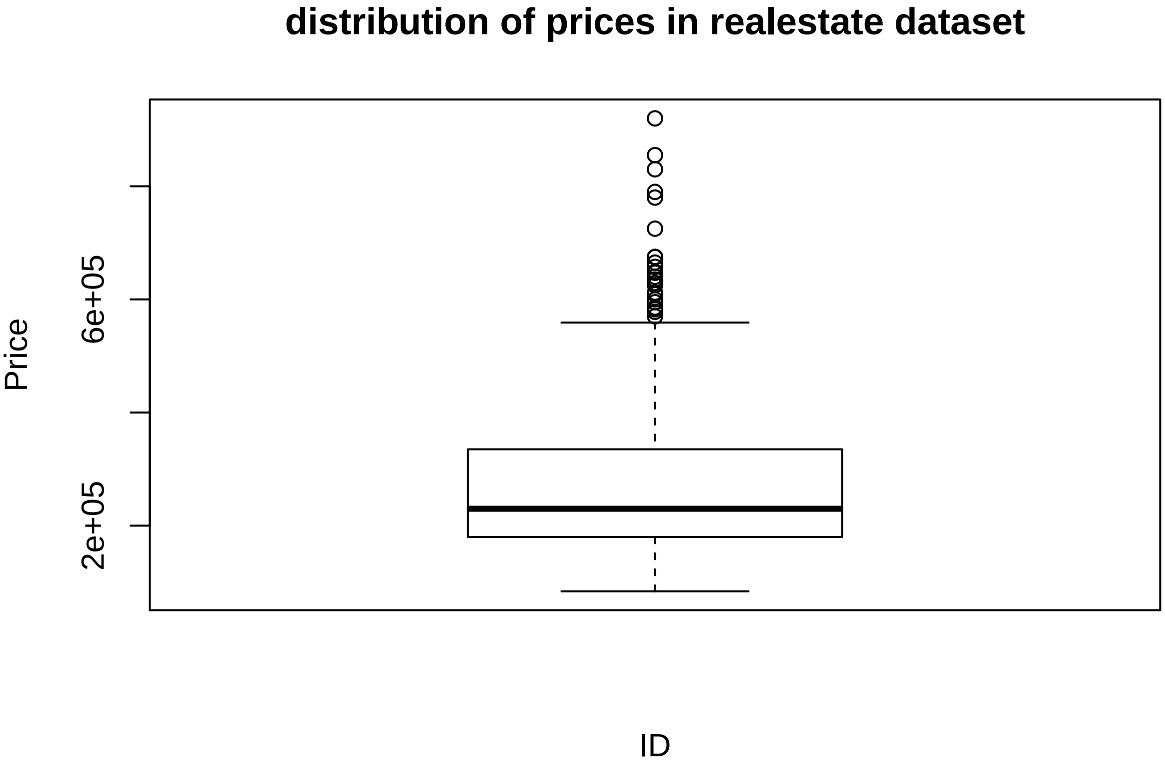
- *ID* : label for characteristics of sale of each house contained in the dataset. We will consider all the following variables for one given house ID.
- *Price* : Price at which house was sold
- *Sqft* : It's surface in square feet
- *Bedroom* : Number of bedrooms in the house
- *Bathroom* : Number of bathrooms in the house
- *Airconditioning* : House is equiped with airconditionning (var = 1) or not (var = 0)
- *Garage* : Number of garages
- *Pool* : Presence of a pool (var = 1) or not (var = 0) <- note: No more than one pool in each house considered since summary shows max of that var is 1.
- *YearBuild* : Year of construction of the house
- *Quality* : Grade going from 1 to 3 evaluating quality schooling nearby (3 is the worst grade)
- *Lot* : Total size of the property.
- *AdjHighway* : takes value 1 if house is near to a highway, 0 otherwise. This variables as well will need to be handled carefully as we don't know the threshold of distance to consider the house close to a highway or not.

Here most qualitative variables have already been converted into binary or quantitave responses (such as the "pool" varaible) which spares us the struggle of doing so.

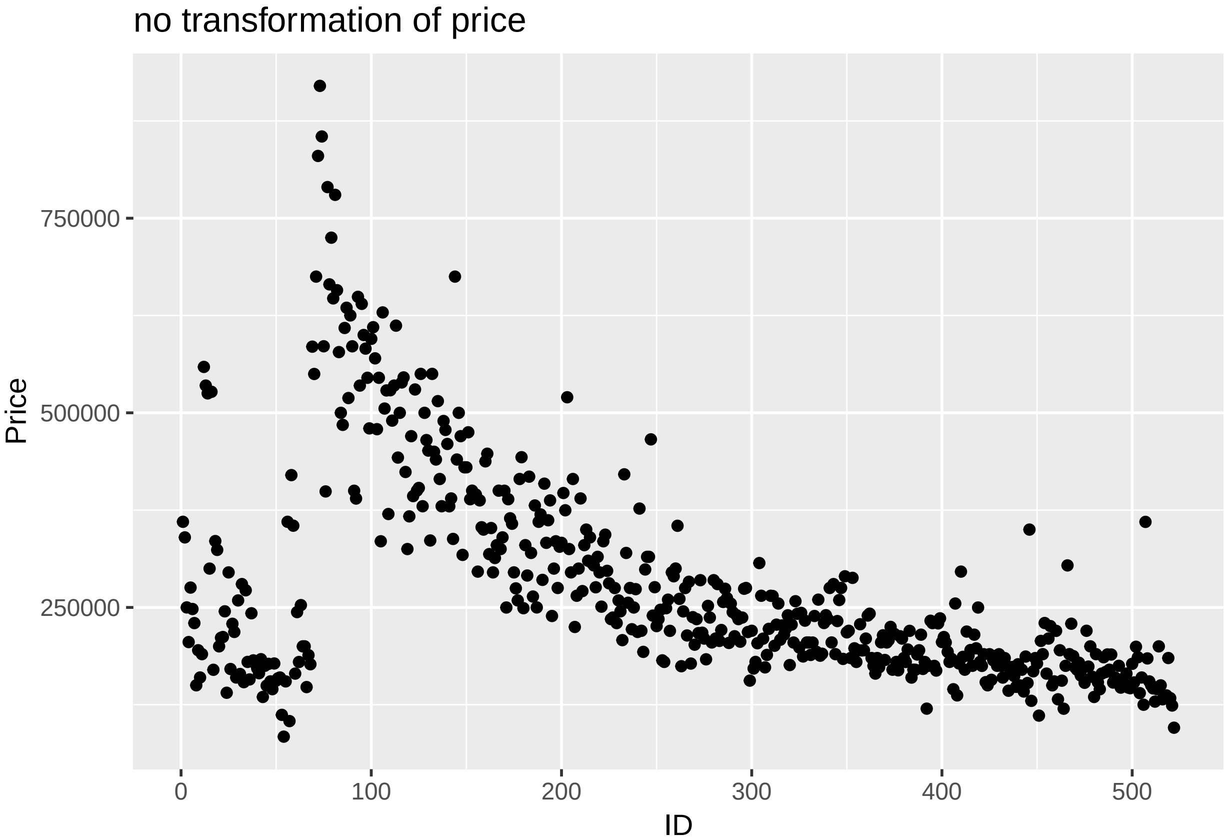
Let's check out the variation within each variable and understand how the variables are structured.

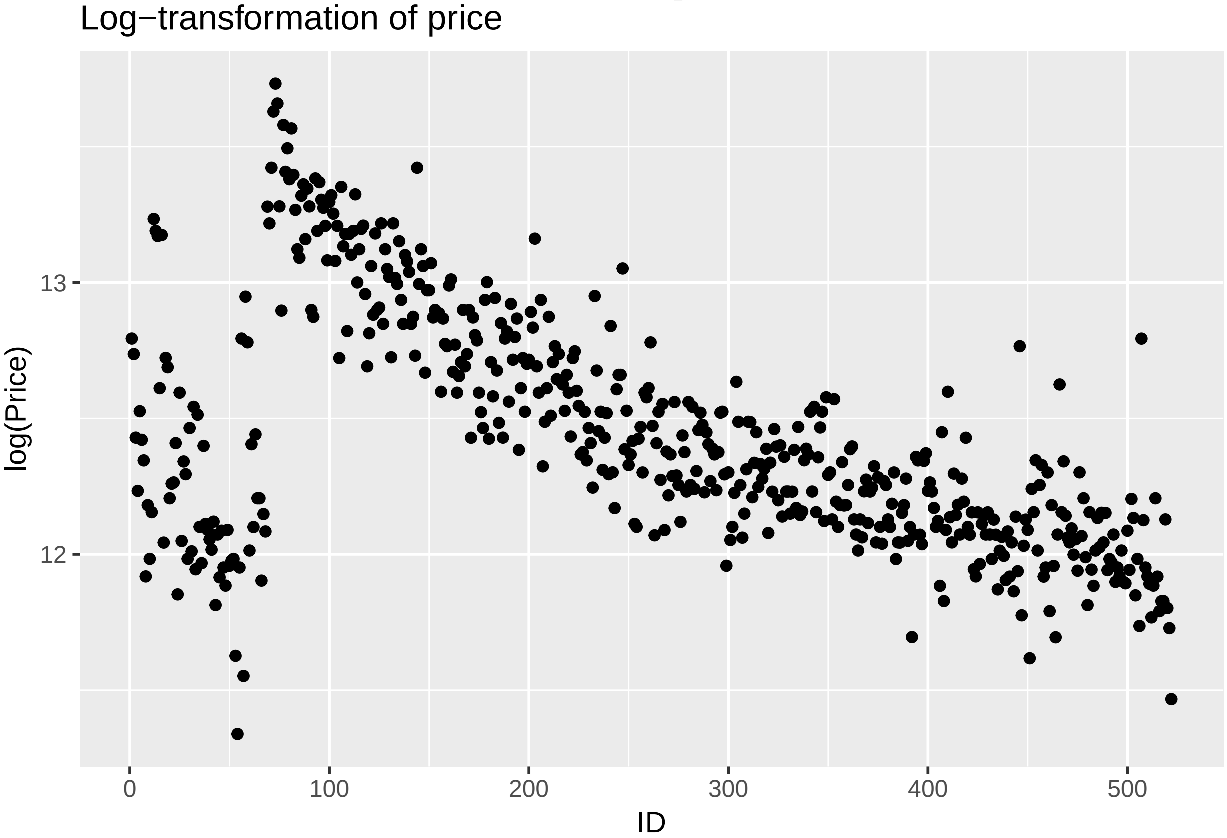
Analysing the variables:

To get a better idea of what house sale market we are in let's look at the distribution of our outcome variable, Price.



Seems there is a positive skewness in house sale prices, so we're dealing with a bulk of houses sold at intermediate prices, and a few very expensive ones. Before we identify them as potential outliers, we can try scaling the price in order to get a better distributions of our data using a log tranformation for example. We'll asses the leverage and influence of these points later on when building our model.

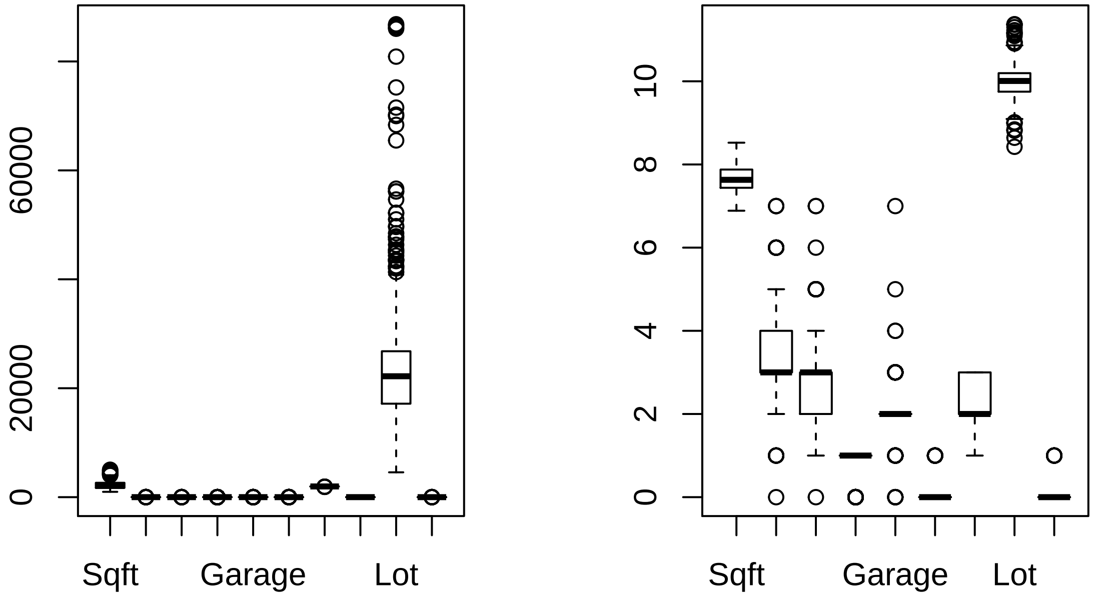




The log transformation seems to have improved the distribution of our data, although transforming the response variable can have big consequences in regards to the interpretation of the model since we aren't looking at the expected value of Y but the expected value of the transformed value of Y.

We'll assess the necessity of this transformation later on, before that let's check the distribution of the data from our predictors. We have different scales for our predictors, so it could be interesting to consider transformations on our predictors to get once again a better distribution of our data between covariates.

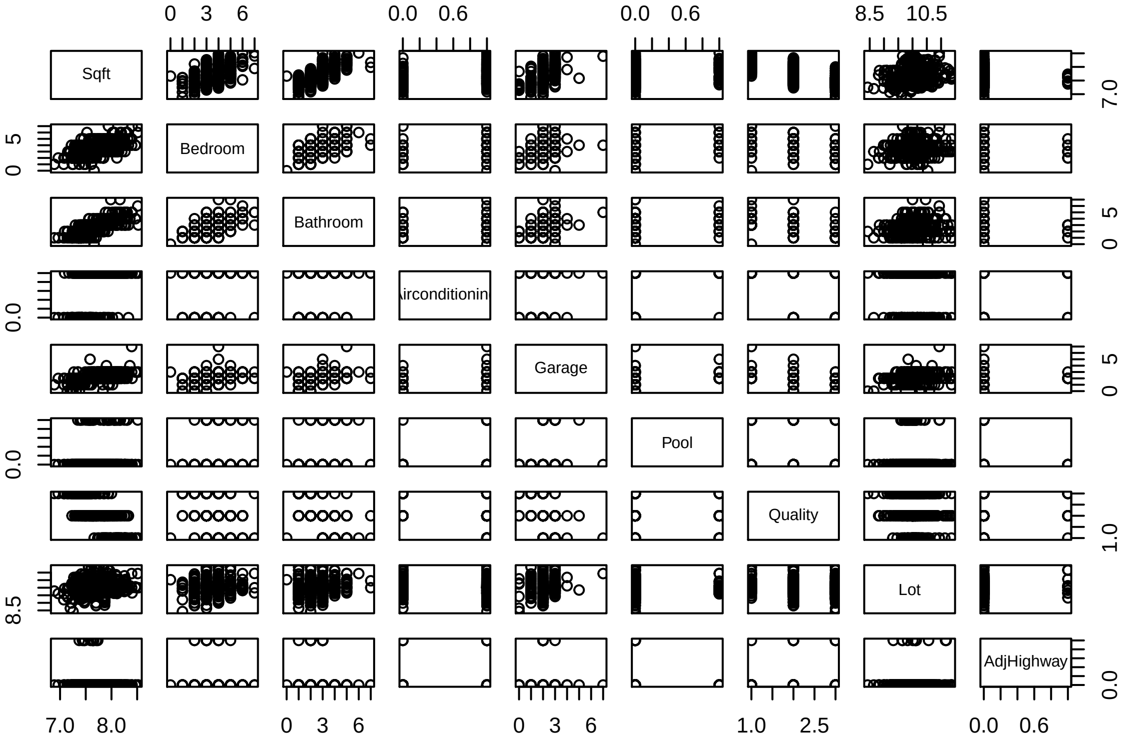
```
par(mfrow=c(1,2))
boxplot(realestate %>% select(-c(ID,Price))) ## Price and ID are not relevant here, we take them out te
boxplot(realestate %>%
  select(-YearBuild, -ID, -Price) %>%
  mutate(Lot = log(Lot), Sqft = log(Sqft))
        )
```



We discard *YearBuild*, *ID* and *Price* to focus on varaibility of only our predictor variables.Here by appling a log transformation to our variables Lot and Sqft seem to improve the distribution of our data which could be easier to work with, we can also check that this transformation has an interest or not.

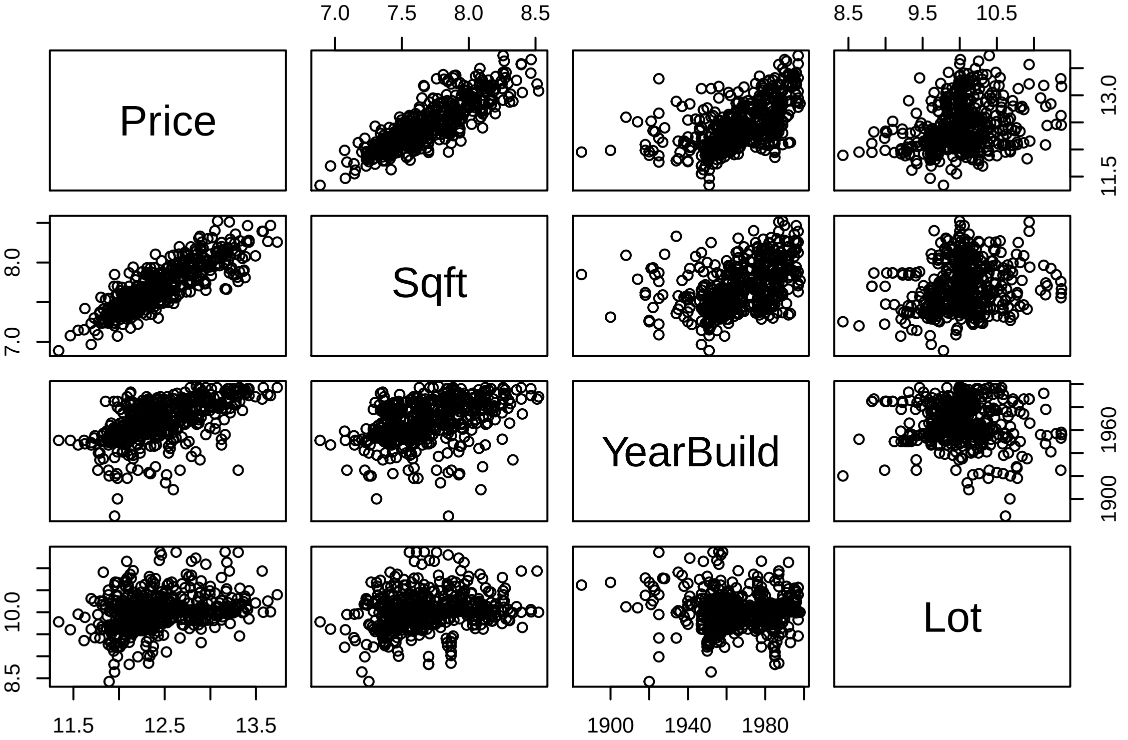
Now that we have an idea of the distribution of our varaibles, let's see how they link to each other, their relations, etc. Also, from now on we'll apply our transformations and consider afterwards if this was a good choice.

Here we note that from now on, we are looking at the LOG of prices, any interpretation of predictions need to be arragned accordingly.



There’s a lot of different variables so it’s hard to read the realtions between variables. To get a more precise look at the relation between Price and the different set of varaibles at hand, I’ll start by looking at the relation with quantitative varaibles then with binary and ordinal variables. We’ll check later for the relation between all variables to verify if there any mulitcolinearity problems, and potentially see if any variable in our data is redundant.

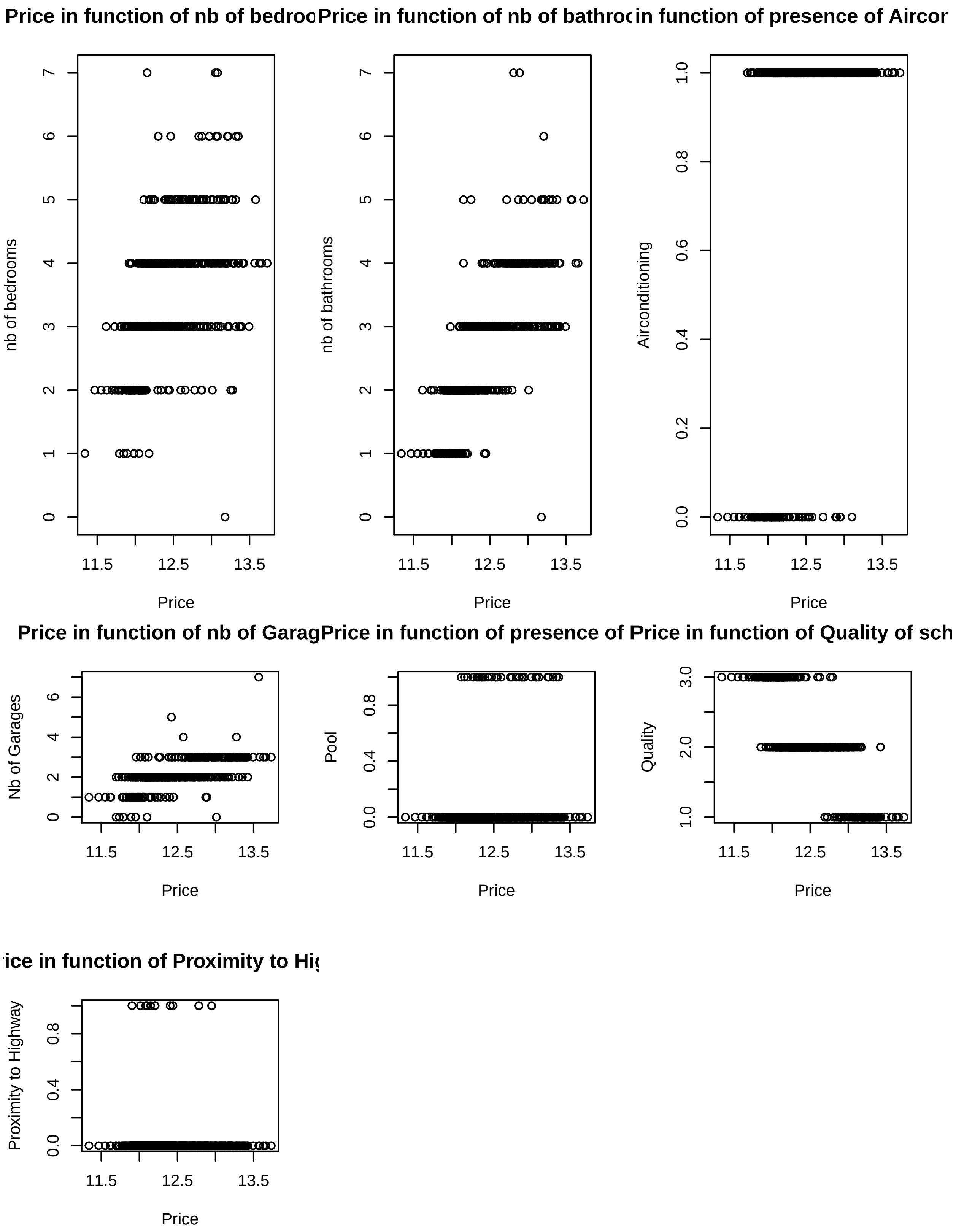
Quantitative Variables



Now relations are much clearer, we can see strong positive relations between Price and each of the other variables. The only relation which isn't clear cut is the "Lot" variable, although we can distinguish a positive relation. Surprisingly there also seems to be a strong relation between sqft and YearBuild, this might meen houses got bigger over the years as families in the studied area grew wealthier. One of them might be a redundant variable as mentionned before, so we'll need to assess later on wether this variable has meaning within our model or not.

Ordinal & Binary variables

Let's look at how the other variables influence the price, each at a time:



First of all, we can notice that there are potential outliers for example in our *Garage* variable, only four houses in our data have more than 4 parking places. Same thing for *AdjHighway*, almost all the houses condsidered are sufficiently far from a Highway to get a 0 in this varaible.

Bedrooms : Surprisingly the number of bedrooms does not show a strong positive relation with house sale price, we can look further into that by studying the correlation coefficient:

```
## [1] 0.4844275
```

The relation is still positive although I expected it to be higher.

Proximity to highway : we can clearly see there are only few houses in our data set which were sufficiently close to a highway to get this variable set to 1, this lack of observations may pose a problem later on when constructing the model.

Pool : no clear difference between houses with and without pools, only that here again the vast majority of our dataset contains houses without pools.

Airconditioning : This seems to be an important factor, our dataset might have been collected in an area where temperatures are high and airconditionning is key in house sales.

Garages : Higher house sale prices are achieved only when increasing the number of garages, althogh the vast majority of the houses considered here have max 3 garages, only 4 houses in our 522 observations go above that threshold, we need to keep this in mind when looking at outliers!

Bathrooms : The relation with price looks similar to that of Bedrooms or Garages although the link is more clear cut. Highest levels for that variables are attained only be a few observations just as with the garage variable. What's surprising is that these don't correspond obviously to the highest prices in our dataset, whereas we could have intuitively thought that the house with 7 garages, 7 bathrooms, 7 bedrooms and a pool would be that extremely high priced mansion down the neighbourhood. That would've made distinguishing outliers easy, too easy.

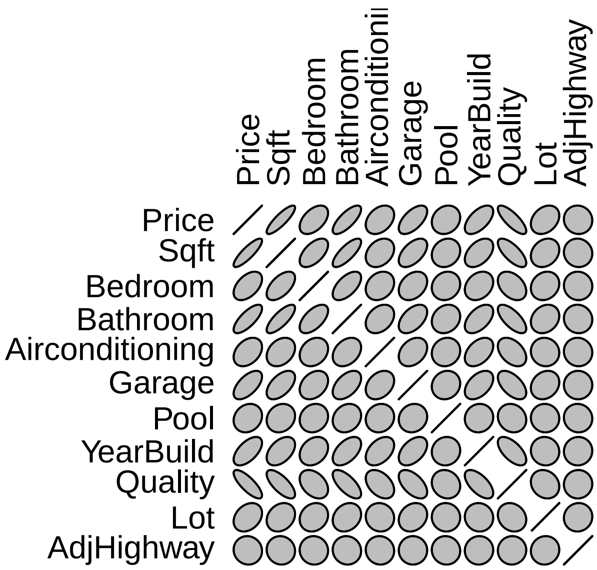
Quality : We can clearly see patterns in function of the value of quality, although I do think we need to modify this variable since it's a grade attributed to school quality, the fact that the grade is 3 doesn't obviously mean that the school is 3 times better. So we'll separate this variable into 3.

Let's check the relations between all the covariates regardless of their type to check for multicolinearity we mentionned beforehand:

Multicolinearity Problems

We saw before that there potentially exists correlation between the prediction variables, which could be a problem of data redundacy as a consequence of overfitting. The best model is the one which has a predictor highly correlated with its explanatory variables but who correlate minamally with each other. If such colinearity exists, by getting rid of redundant variables we can seek to achieve statistical robustness.

Let's first get a better look at the correlation between the varaibles:



We can see that there are 3 main variables which may me source of redundancy: *Quality*, *Bedroom* - *Bathroom* (only apparent correlation between them), and maybe *YearBuild*. To verify if they may truly pose a problem, we can check the variance inflation factors.

```
##
##                    GVIF Df GVIF^(1/(2*Df))
## Sqft            3.39526  1         1.84262
## Bedroom         1.66334  1         1.28970
## Bathroom        3.16695  1         1.77959
## Airconditioning 1.37570  1         1.17290
## Garage          1.66391  1         1.28993
## Pool            1.05044  1         1.02491
## YearBuild       1.91265  1         1.38299
## Quality         3.20212  2         1.33770
## Lot             1.16818  1         1.08082
## AdjHighway      1.01549  1         1.00771
##
##   Mean: 1.46076
```

The mean of the VIF is under 5 and no individual factor is above 10, seems as though there is no multicolinearity problem.

Now that we have an idea of how are covariates and data overall are distributed, we can move on to the construction of our model.

The model

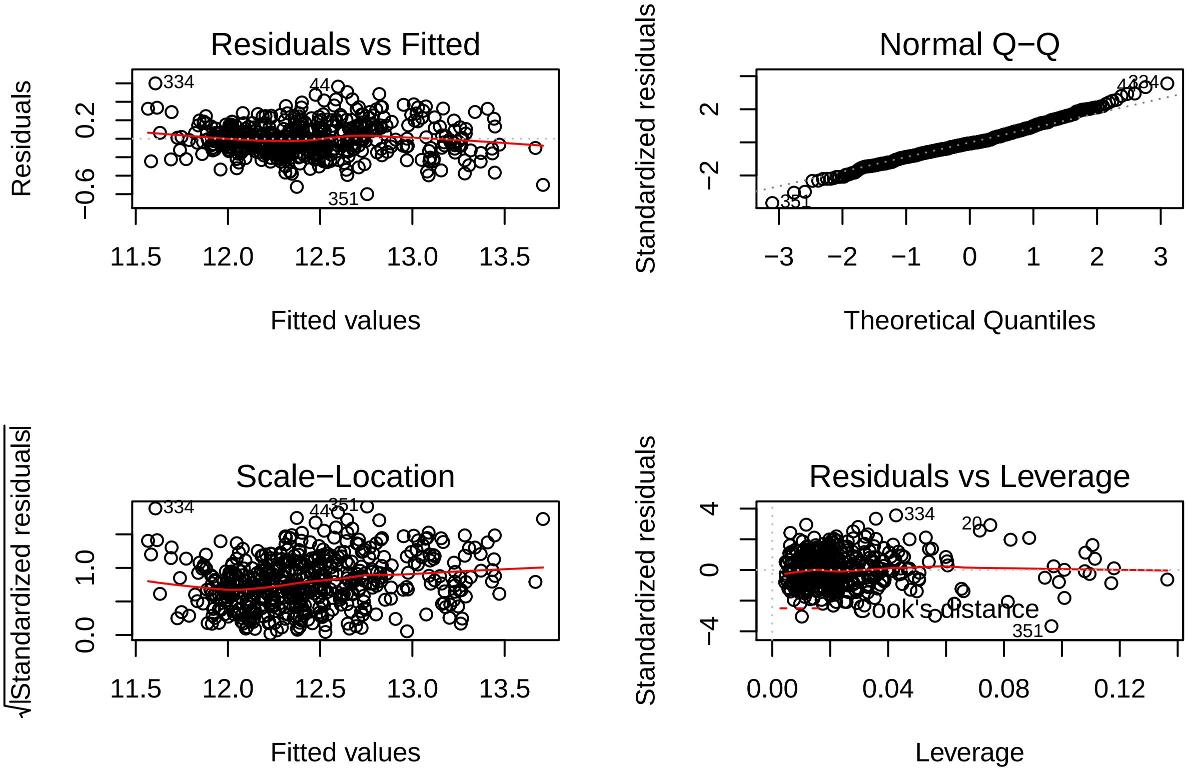
Multivariate regression:

Full Model:

Let's start by fitting a multiple regression using classic least squares method:

```
##
## Call:
## lm(formula = realestate_transf$Price ~ ., data = realestate_transf)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59997 -0.10116 -0.01119  0.09985  0.59950
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.8921624  1.2870815  -1.470   0.1421
## Sqft             0.7023166  0.0474190  14.811  < 2e-16 ***
## Bedroom          0.0001468  0.0095936   0.015   0.9878
## Bathroom         0.0307460  0.0126181   2.437   0.0152 *
## Airconditioning  0.0384870  0.0236164   1.630   0.1038
## Garage           0.0317972  0.0148830   2.136   0.0331 *
## Pool             0.0543066  0.0304898   1.781   0.0755 .
## YearBuild        0.0038832  0.0005916   6.563 1.30e-10 ***
## Quality2        -0.3124255  0.0290197 -10.766  < 2e-16 ***
## Quality3        -0.3720239  0.0409995  -9.074  < 2e-16 ***
## Lot              0.1402016  0.0197268   7.107 4.02e-12 ***
## AdjHighway      -0.0682356  0.0528894  -1.290   0.1976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1722 on 510 degrees of freedom
## Multiple R-squared:  0.8442, Adjusted R-squared:  0.8408
## F-statistic: 251.2 on 11 and 510 DF,  p-value: < 2.2e-16
```



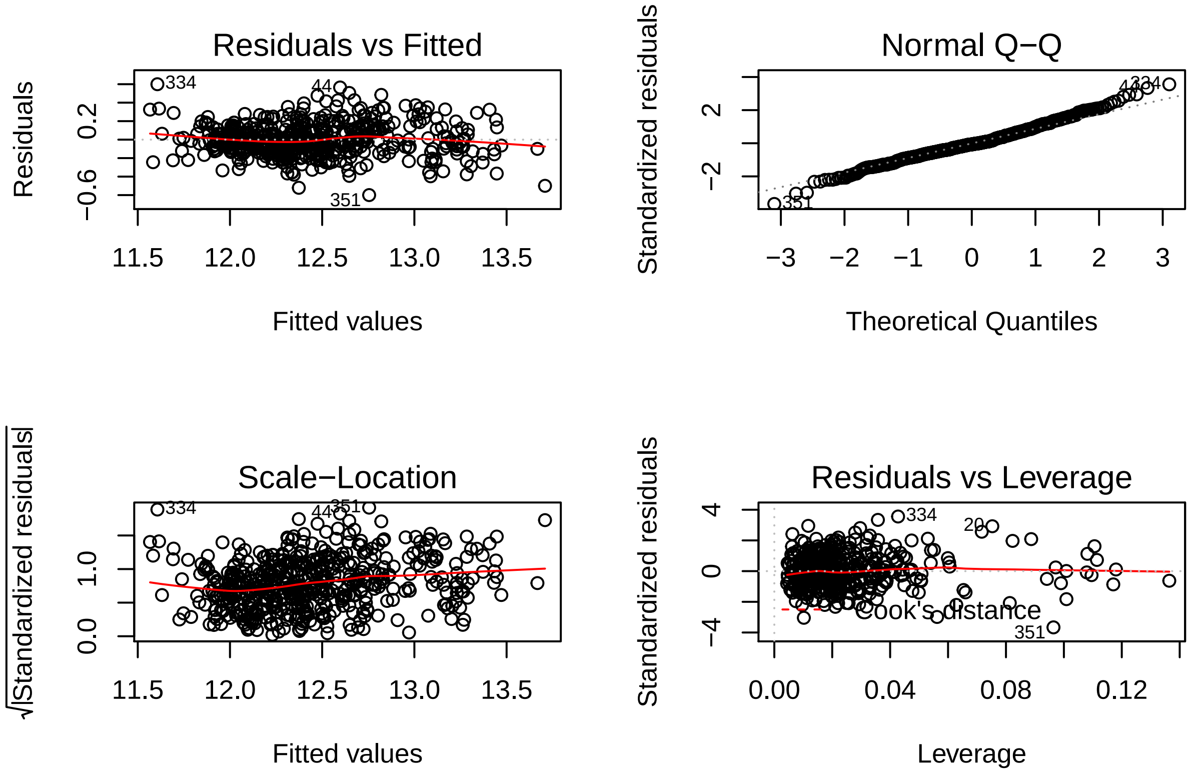
What comes out of our first model is that we have 3 non significant variables: the *Intercept*, *Bedroom*, *Airconditionning* and *AdjHighway*. The F statistic yields a very low p-value so we know that at least one of these individually unsignificant variables have a significant relation with our outcome variable, although we can't directly conclude on which one directly. Concerning the residuals they same to satisfy the hypothesis necessary for our model to be valid:

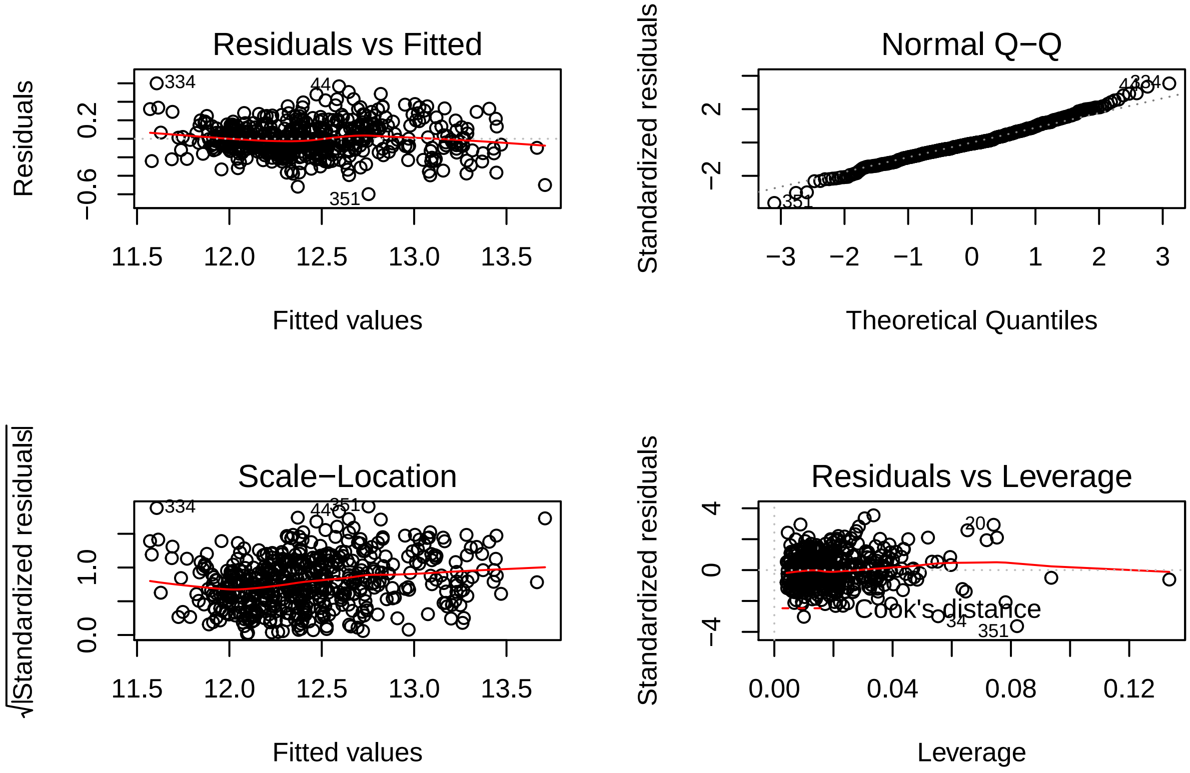
- The residuals are approximately normally distributed from the qqplot.
- There isn't a very recognisible structure within our residuals in the residuals vs. fitted values so there seems to be a non-linear relationship
- The scale location plot shows that the assumption of homoscedasticity (equal variance)
- The residuals vs. leverage plot show that all the data fall within cook's distance so maybe our concerns regarding outliers may not be much of a problem here, which is convenient since taking out outliers from our model must be done with extreme precaution. Although R seems to have detected a few "extreme" points (pt. 1,20,351…)

Reduced Model:

Since we saw in our data before that we did not have much values for houses near a Highway, this could be the origin of its non-significance, although I'm sure it would have been very relevant in our model if we did have more data. Also, it might be that either bedroom or bathroom is a redundant variable although we cheched multicolinearity beforehand. In the end both indicate the capacity of the house so in that sense give info on the same criteria. Thus, I will try reformulating my model while leaving these two variables out and check if my guesses were right.

```
##
## Call:
## lm(formula = realestate_transf$Price ~ . - Bedroom - AdjHighway,
##     data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59843 -0.10139 -0.01079  0.10105  0.59960
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.7825057  1.2825268  -1.390  0.16518
## Sqft             0.7042392  0.0456365  15.431  < 2e-16 ***
## Bathroom         0.0314425  0.0120995   2.599  0.00963 **
## Airconditioning  0.0397340  0.0235336   1.688  0.09194 .
## Garage           0.0315750  0.0148763   2.123  0.03427 *
## Pool             0.0550982  0.0304640   1.809  0.07109 .
## YearBuild        0.0038271  0.0005897   6.490 2.03e-10 ***
## Quality2        -0.3125563  0.0286416 -10.913  < 2e-16 ***
## Quality3        -0.3719335  0.0405091  -9.181  < 2e-16 ***
## Lot              0.1384456  0.0196655   7.040 6.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1722 on 512 degrees of freedom
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.8409
## F-statistic:   307 on 9 and 512 DF,  p-value: < 2.2e-16
```





Now that we've taken out the variables *Bedroom* & *AdjHighway* we get a model with only significant variables according to the T-test, as for the residuals, they verify all the conditions met to justify the hypothesis necessary for our regression method (stated before). They seem to have even improved compared to our previous model taking all variables into account. Although we do see that there is one point which seems to have a large influence. We'll study the decision to make regarding potential outliers later on.

In order to really confirm that we have improved our model with the deleted variables we choosed, we need to run an F-test to confirm our hypothesis:

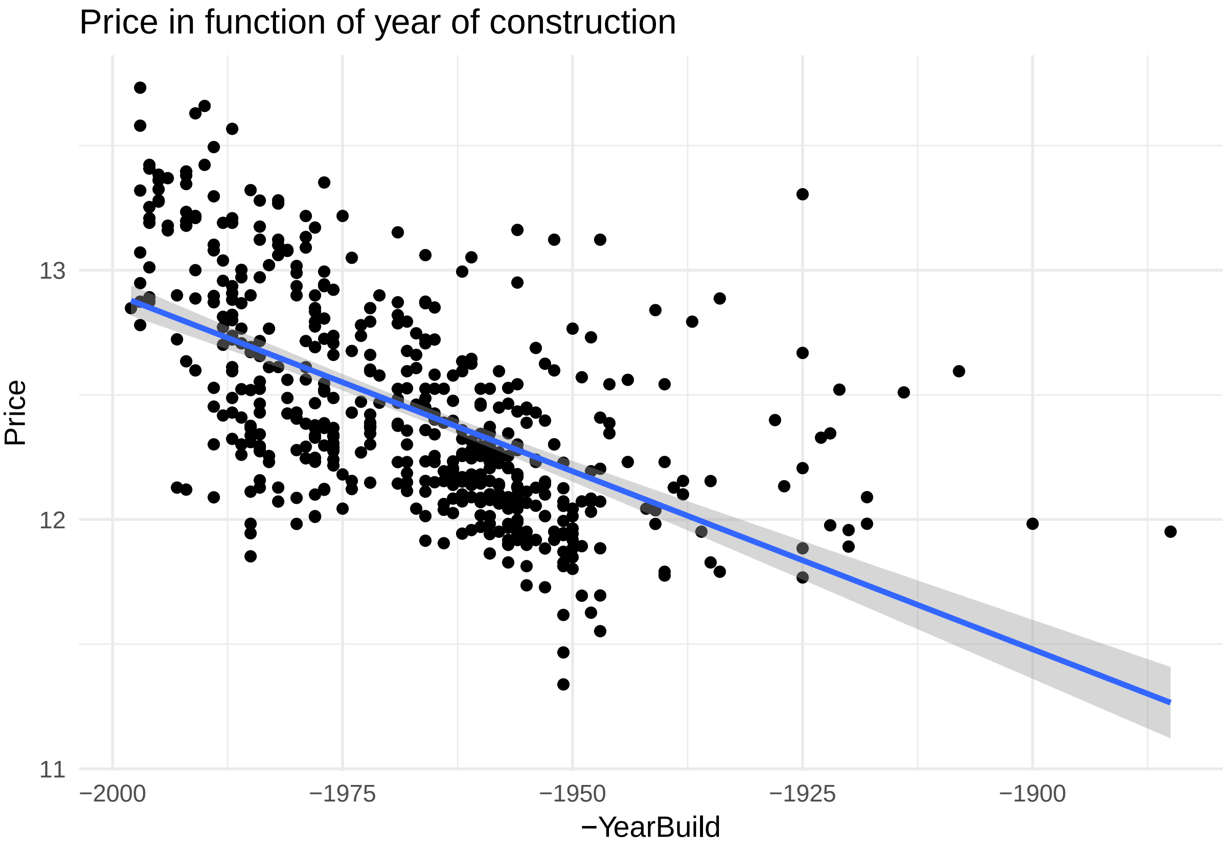
```
## Analysis of Variance Table
##
## Model 1: realestate_transf$Price ~ Sqft + Bedroom + Bathroom + Airconditioning +
##     Garage + Pool + YearBuild + Quality + Lot + AdjHighway
## Model 2: realestate_transf$Price ~ (Sqft + Bedroom + Bathroom + Airconditioning +
##     Garage + Pool + YearBuild + Quality + Lot + AdjHighway) -
##     Bedroom - AdjHighway
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    510 15.128
## 2    512 15.177 -2 -0.049373 0.8323 0.4357
```

The p-value we obtain (greater than any tolerable choice of alpha) indicates that we don't reject H0, so deleting the variables we selected seems to make sense here. Although our model seems significant and the residuals seem to pass the conditions required for a valid model, there may be more at stake than we think. We'll go on testing different hypothesis, those proposed by you and some of my own creation.

Interactions hypothesis:

Hypothesis proposed:

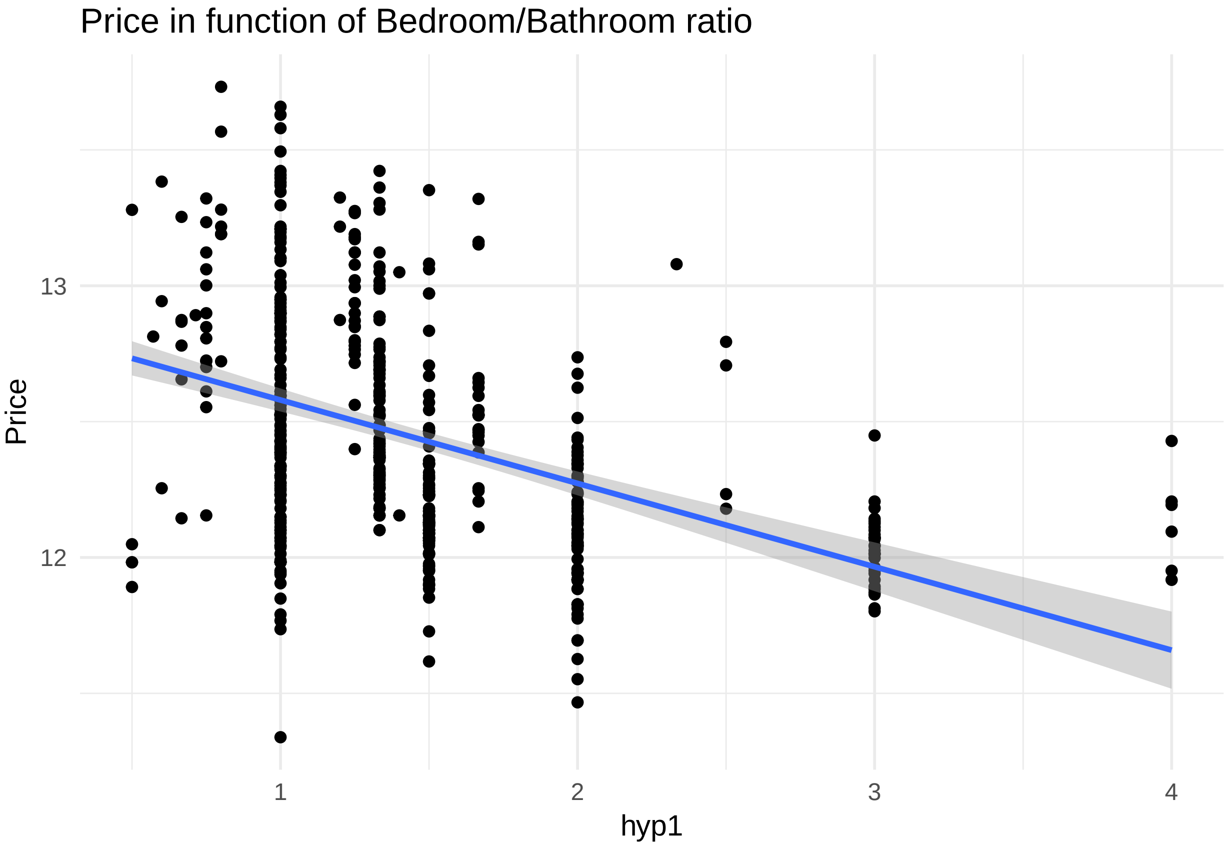
1. **Older houses tend to have lower prices:**



The trend here is indeed downwards, although there are very few data to confirm this hypothesis, we can see the conf interval of the linear fit grows as we go down in years (less data points so obviously). Although if we do look at the prices from 1950 onwards, we do see an increase. Also, in our model the variable YearBuild is very significant, so we can say that prices do tend to be lower for older houses. Although the fact that we lack data for the early part of the XXth century might pose a problem, how can we say we'll be able to predict the price of houses that data from that period if only a dozen data from that time were took into account in our model?

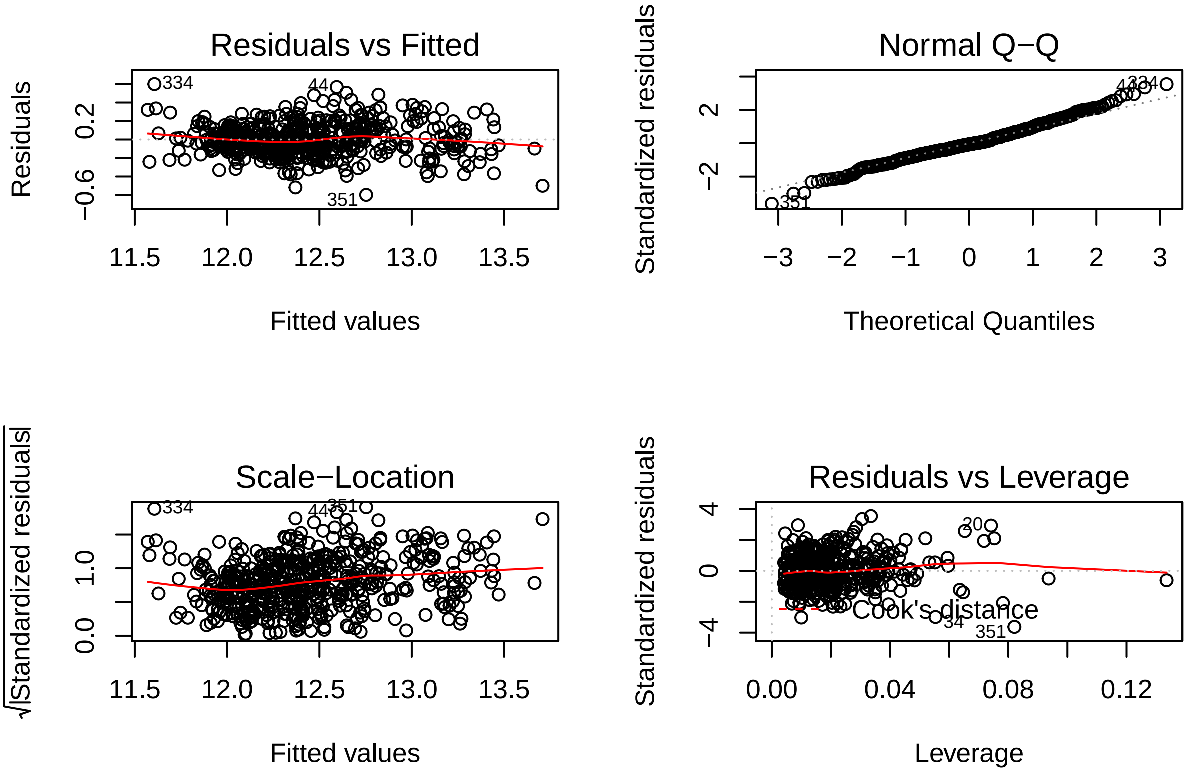
2. House with higher bathroom/bedroom ratio should have higher price:

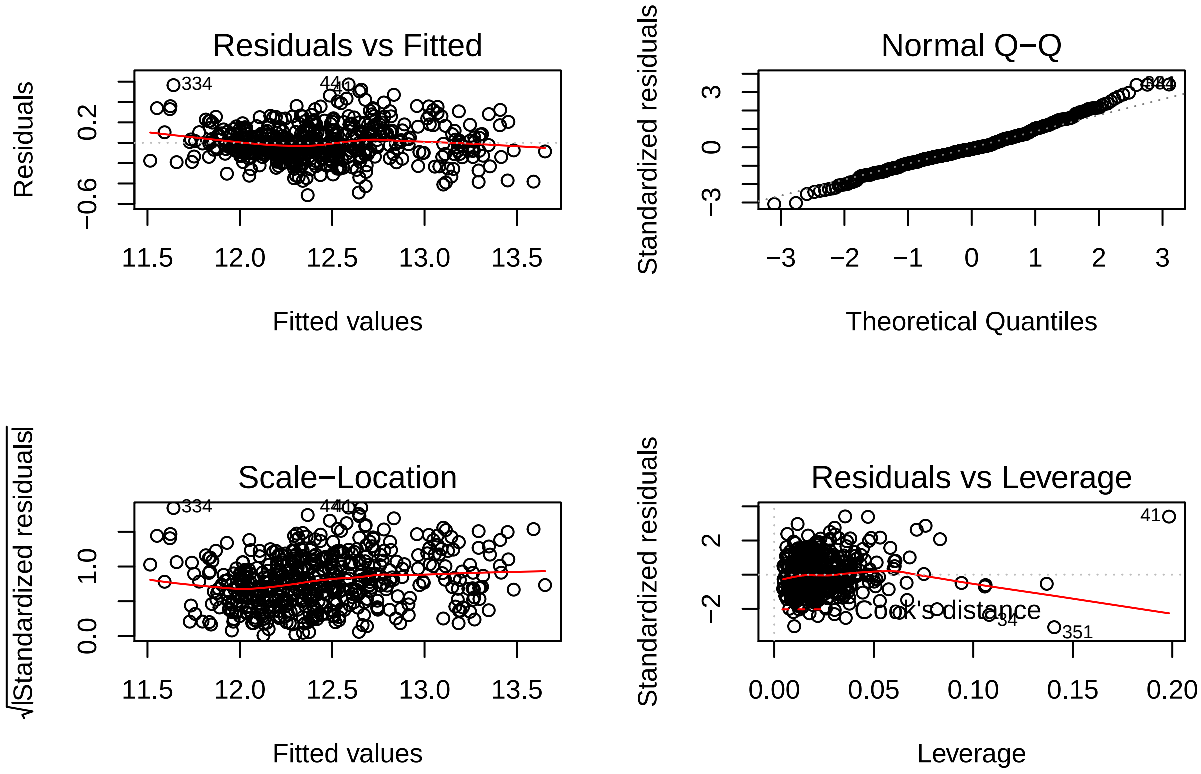
Instead of looking at bathroom/bedroom I would like to test the inverse, in my opinion more bathrooms than bedrooms doesn't present much interest but too many bedrooms for not enough bathrooms can be inconvenient, so we should see a downtrend of prices as that ratio goes up:



As expected, this inconvenience we mentionned does seem to have its effect on price, although we need to see if this variable fits our model:

```
##
## Call:
## lm(formula = Price ~ . + Bedroom/Bathroom - AdjHighway, data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51495 -0.10097 -0.01438  0.09804  0.57261
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.1439351  1.2803562  -1.674 0.094649 .
## Sqft               0.7137272  0.0471505  15.137  < 2e-16 ***
## Bedroom            0.0519284  0.0195378   2.658 0.008111 **
## Bathroom           0.0998453  0.0257216   3.882 0.000117 ***
## Airconditioning    0.0325166  0.0235430   1.381 0.167835
## Garage             0.0273536  0.0148376   1.844 0.065832 .
## Pool               0.0629891  0.0303706   2.074 0.038579 *
## YearBuild          0.0038885  0.0005861   6.635 8.30e-11 ***
## Quality2          -0.3235151  0.0290305 -11.144  < 2e-16 ***
## Quality3          -0.3703423  0.0407016  -9.099  < 2e-16 ***
## Lot                0.1397493  0.0195400   7.152 2.99e-12 ***
## Bedroom:Bathroom -0.0189950  0.0062404  -3.044 0.002456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.171 on 510 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.8431
## F-statistic: 255.6 on 11 and 510 DF,  p-value: < 2.2e-16
```

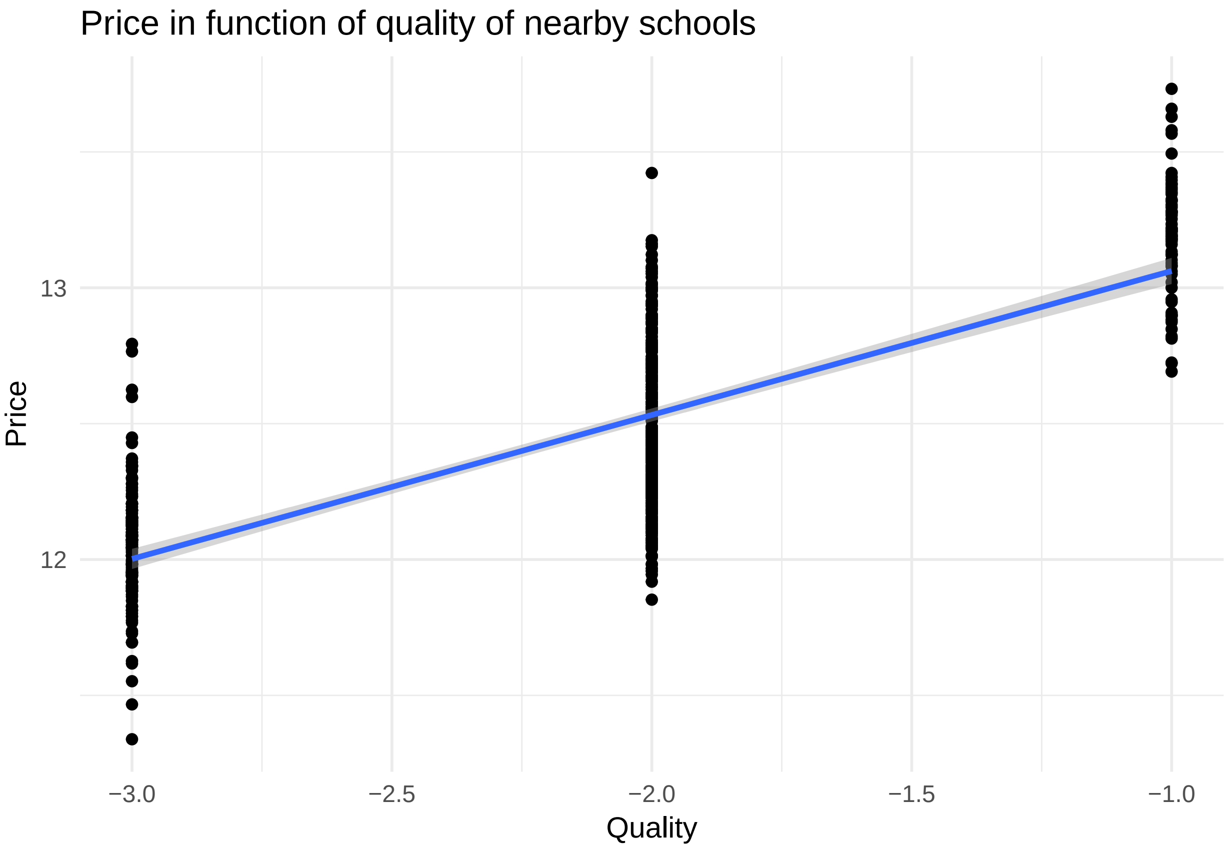




The interaction is very significant in our model, although the variable Airconditioning isn't significant anymore. Nonetheless there is a downgrade in the residuals compared to our reduced model, we'll need to see if this could be integrated in our model or not using Anova.

3. School quality impacts the price positively:

This variable should impact positively the price if we are facing a clientèle which have children were surely the school quality will be a criteria of selection among houses, and in turn will be a price argument for the salesman:

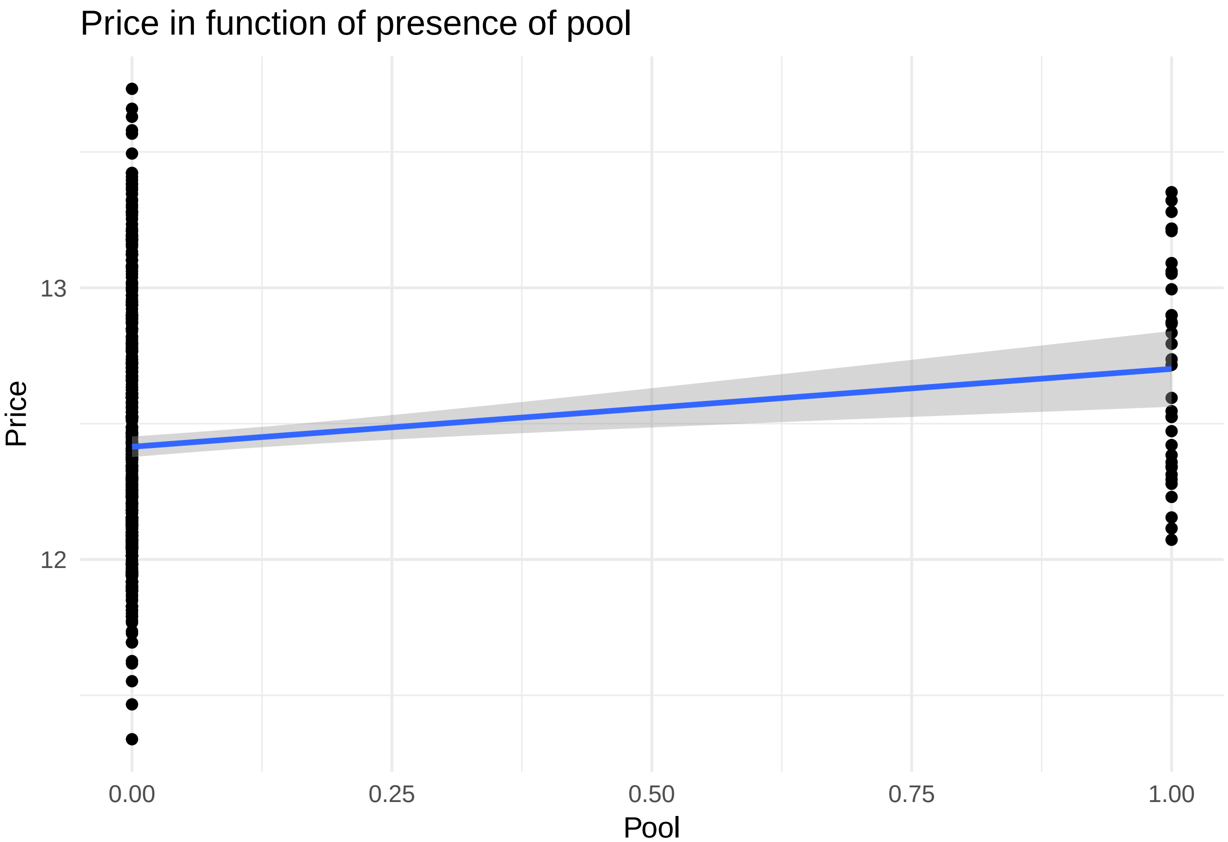


There is indeed a positive relation between the price of the house and the quality of the schools nearby, and the confidence interval of the fit seems pretty precise so the data stick to this relation.

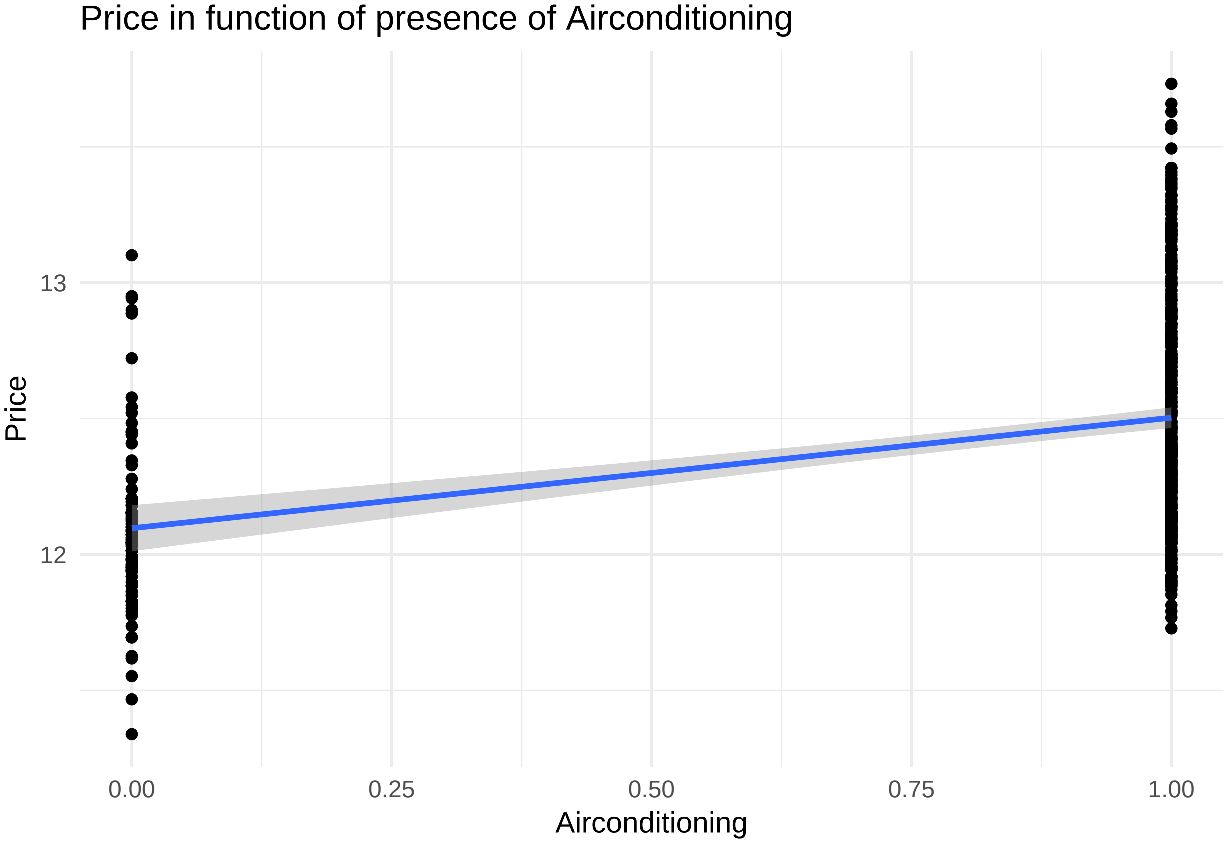
My hypothesis

I want to test two things

1. It seems as though the *Pool* variable is significant in almost all our models whereas *Airconditioning* almost in none, I want to check both that *Pool* have a positive effect on price, *Airconditionning* doesn't really affect but also that *Airconditionning* might have an influence when the house has no pool. I can check that using the following interaction *Airconditioning x Pool*.



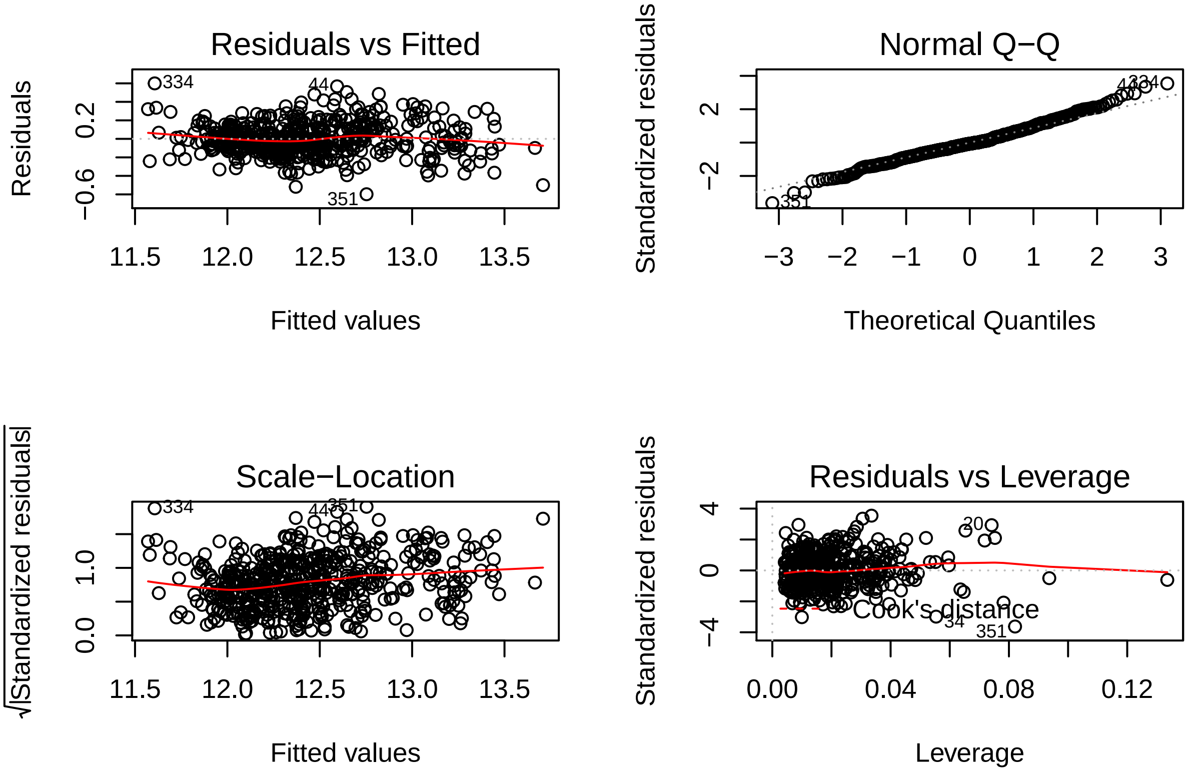
Price in function of presence of pool

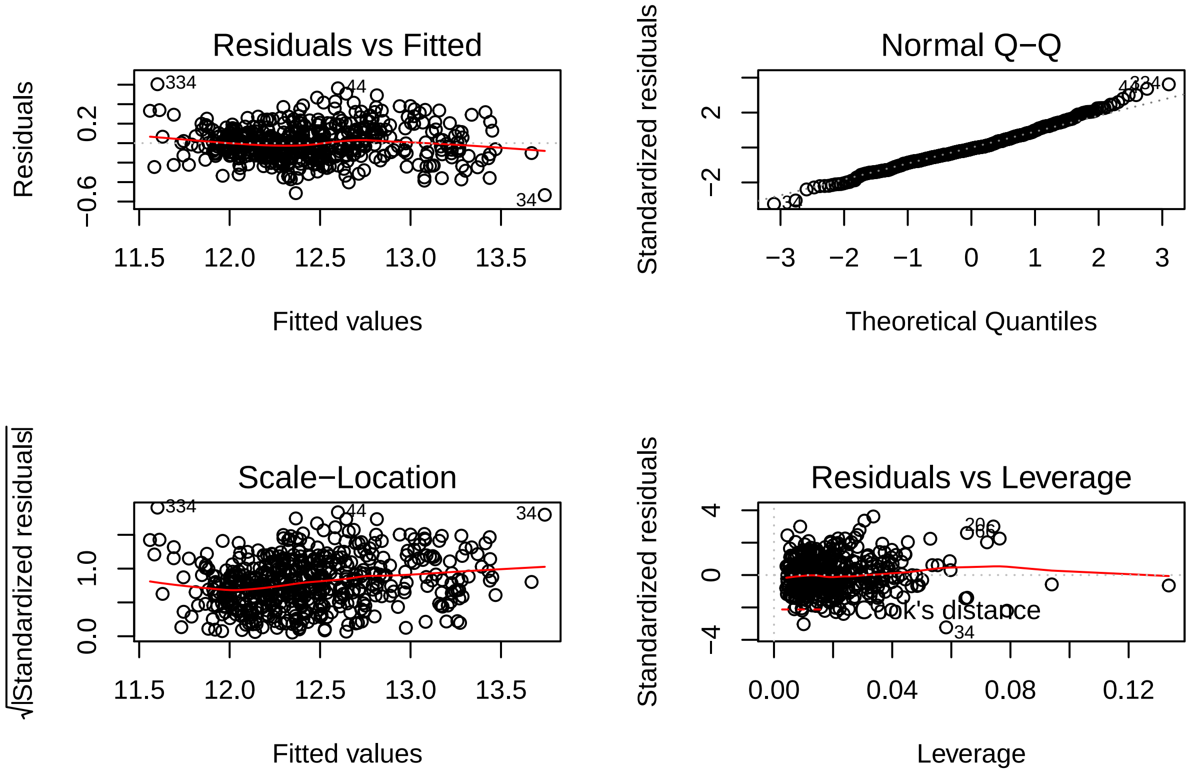


Price in function of presence of Airconditioning

There is a positive relation for both, although there isn't a lot of houses with a pool in our data, this may affect the significance of that variable as well as the interaction in our model. Let's check:

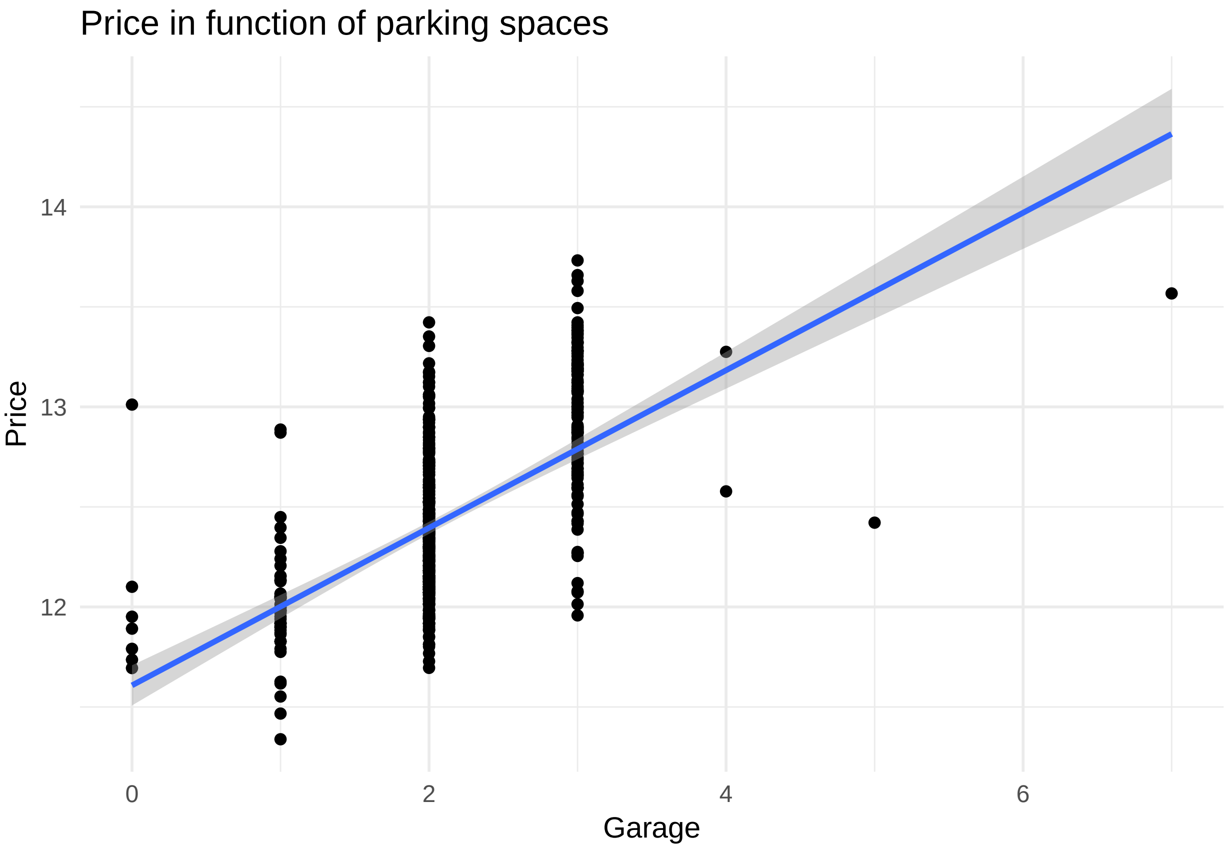
```
##
## Call:
## lm(formula = Price ~ . + Airconditioning * Pool - Bedroom - AdjHighway,
##     data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53343 -0.10365 -0.01318  0.10097  0.60491
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.1465933  1.2710472  -1.689 0.091860 .
## Sqft                  0.7042831  0.0450902  15.619  < 2e-16 ***
## Bathroom              0.0370136  0.0120506   3.072 0.002243 **
## Airconditioning       0.0320037  0.0233469   1.371 0.171043
## Garage                0.0301587  0.0147033   2.051 0.040761 *
## Pool                 -0.5785814  0.1751853  -3.303 0.001025 **
## YearBuild             0.0039680  0.0005839   6.796 3.01e-11 ***
## Quality2             -0.3048734  0.0283760 -10.744  < 2e-16 ***
## Quality3             -0.3525064  0.0403723  -8.731  < 2e-16 ***
## Lot                   0.1455375  0.0195259   7.454 3.91e-13 ***
## Airconditioning:Pool  0.6519363  0.1775523   3.672 0.000266 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1701 on 511 degrees of freedom
## Multiple R-squared:  0.8477, Adjusted R-squared:  0.8447
## F-statistic: 284.4 on 10 and 511 DF,  p-value: < 2.2e-16
```





This interaction is positive, so Airconditioning doesn't affect the same way the price if there's a pool or not as I expected, although the main variable Airconditioning is still not significative. We can try keeping this interaction but not the main variable now that we've seen the main effect is non significant whereas the interaction is.

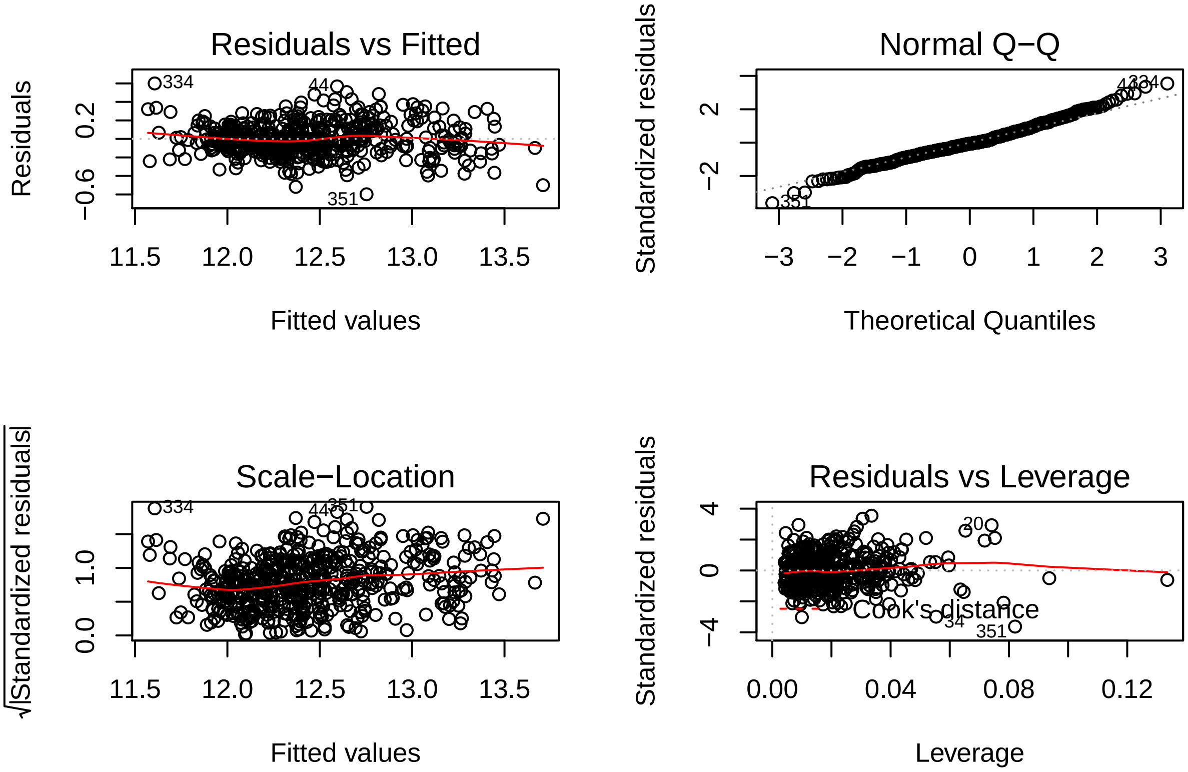
2. It makes sense to have a lot of parking spaces if you have a big house which can welcome many people, but if it's a small house then it's just a waste of space. So in order to see how *Garage* affects prices in function of the capacity of the house, i'll try out the following interaction *Garage x Bedroom*

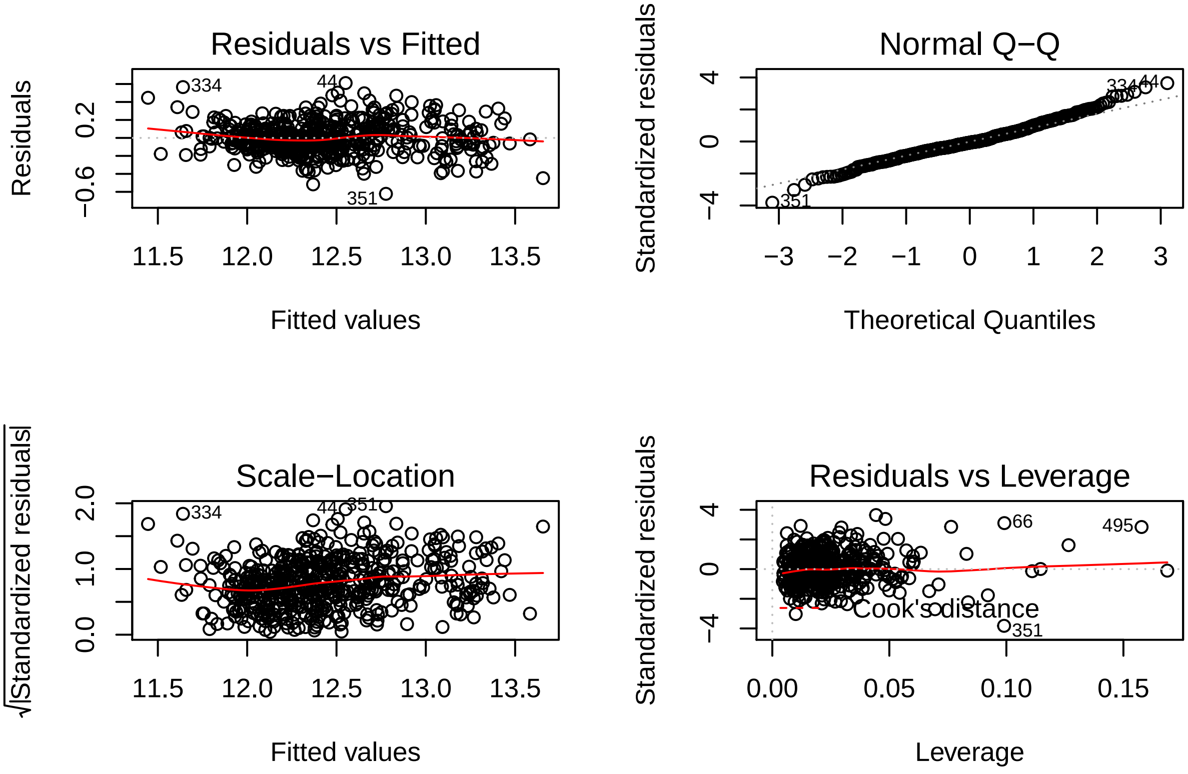


Here again, strong positive realtionship but for parking spaces exceeding 3 we have very few data to confirm our hypothesis. Once again this can affect the fit to the model but it's still worth a try:

```
##
## Call:
## lm(formula = Price ~ . + Garage * Bedroom - AdjHighway, data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62193 -0.10193 -0.01227  0.09796  0.61046
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.8829212  1.2780996  -1.473  0.14131
## Sqft              0.7090754  0.0471822  15.028  < 2e-16 ***
## Bedroom           0.0591958  0.0242220   2.444  0.01487 *
## Bathroom          0.0330238  0.0125546   2.630  0.00879 **
## Airconditioning   0.0346216  0.0235515   1.470  0.14217
## Garage            0.1240204  0.0377881   3.282  0.00110 **
## Pool              0.0520907  0.0303450   1.717  0.08666 .
## YearBuild         0.0037768  0.0005873   6.431 2.91e-10 ***
## Quality2         -0.3229210  0.0291295 -11.086  < 2e-16 ***
## Quality3         -0.3771057  0.0408309  -9.236  < 2e-16 ***
## Lot               0.1366030  0.0195890   6.973 9.63e-12 ***
## Bedroom:Garage   -0.0281711  0.0105948  -2.659  0.00808 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1713 on 510 degrees of freedom
## Multiple R-squared:  0.8458, Adjusted R-squared:  0.8425
## F-statistic: 254.3 on 11 and 510 DF,  p-value: < 2.2e-16
```



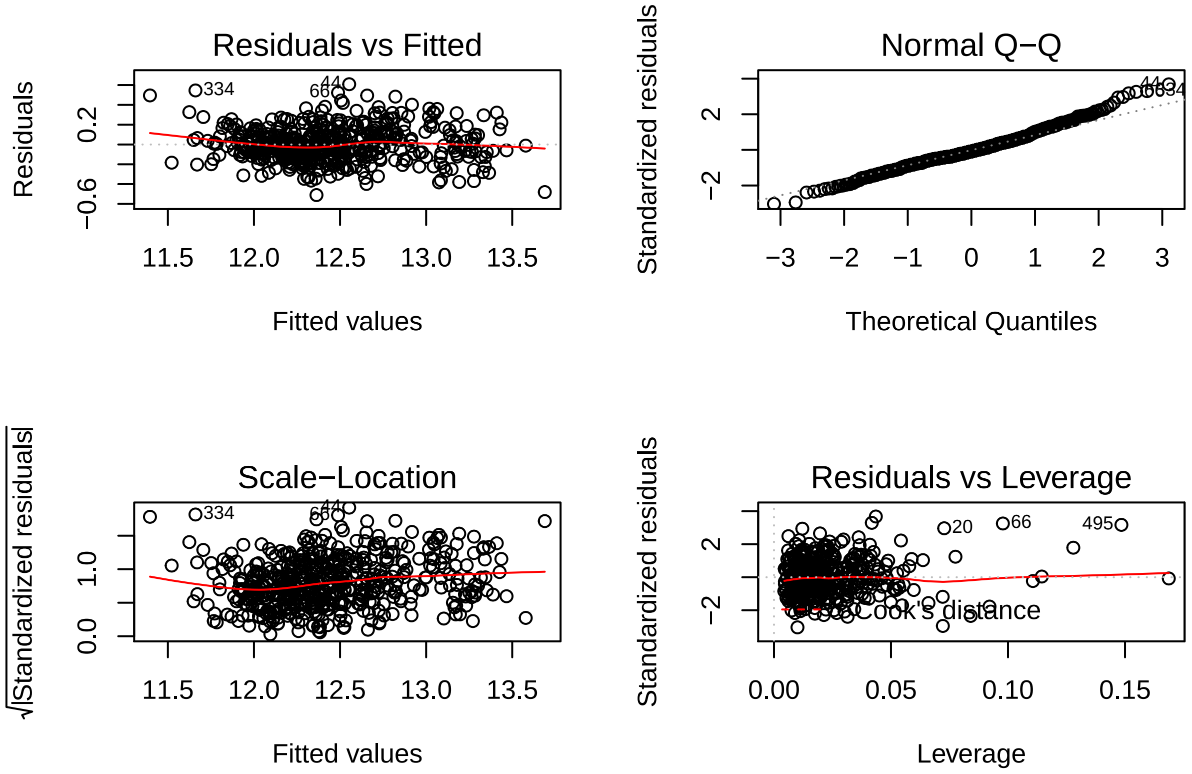


This interaction seems to be significant as well! Although we lost significance for the intercept (not that big of a deal, I'll just let the intercept be the intercept) and Airconditioning is non-significant once again.

MY MODEL

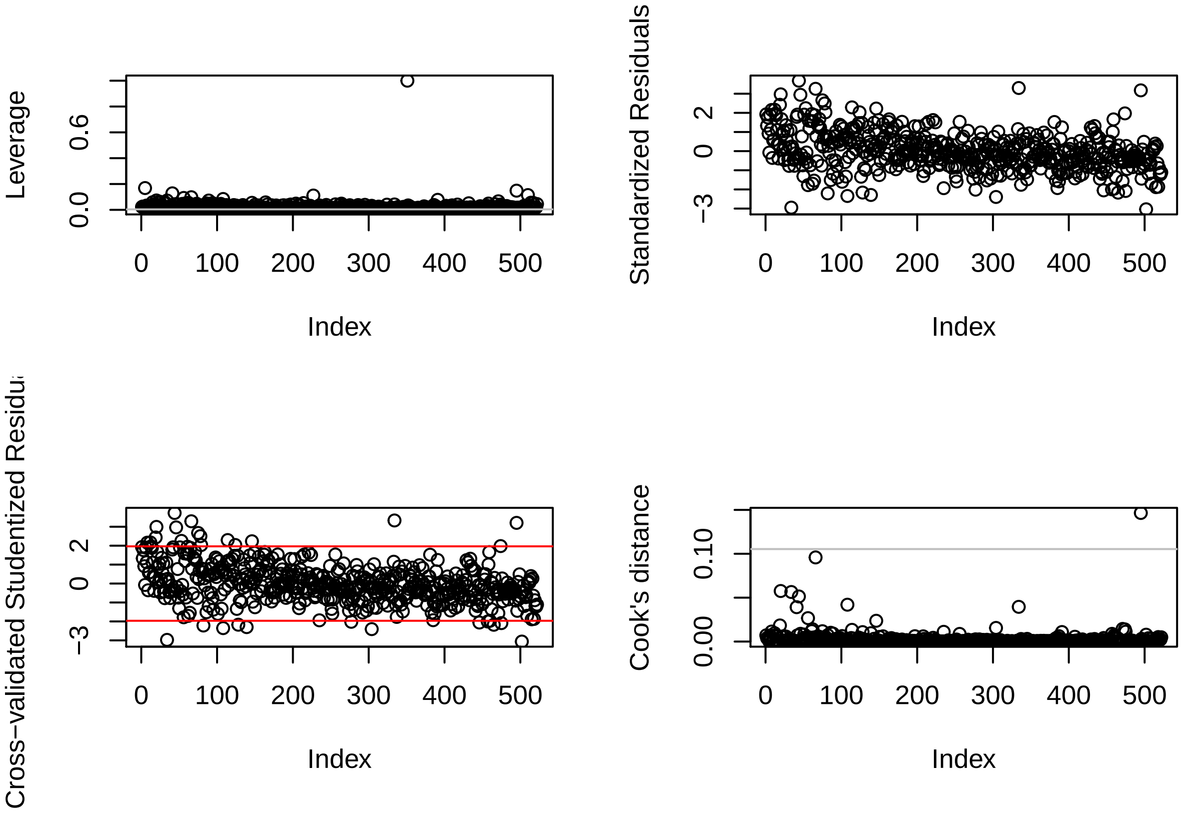
So my thought is that with the interactions we've tested, they are all significant but we're not sure they might all significant together, so we'll integrate them to the model but be very cautious about interpretting them afterwards. Since we distinguished main effects from interaction effects at each hypothesis we tested, we can remove the individual non significant variables related to the interactions and just keep the interaction. This will constitue my final model.

```
##
## Call:
## lm(formula = realestate_transf$Price ~ . + Garage * Bedroom +
##     Airconditioning * Pool - Airconditioning - AdjHighway, data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51086 -0.09747 -0.01534  0.09234  0.60846
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.435332   1.252725  -1.944 0.052442 .
## Sqft                  0.702708   0.046573  15.088  < 2e-16 ***
## Bedroom               0.071304   0.023869   2.987 0.002950 **
## Bathroom              0.037307   0.012438   2.999 0.002836 **
## Garage                0.134710   0.037021   3.639 0.000302 ***
## Pool                 -0.637321   0.175051  -3.641 0.000300 ***
## YearBuild             0.004044   0.000571   7.082 4.73e-12 ***
## Quality2             -0.316982   0.028762 -11.021  < 2e-16 ***
## Quality3             -0.364537   0.040306  -9.044  < 2e-16 ***
## Lot                   0.141214   0.019278   7.325 9.38e-13 ***
## Bedroom:Garage       -0.031355   0.010430  -3.006 0.002775 **
## Airconditioning:Pool  0.710182   0.177138   4.009 7.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.169 on 510 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8466
## F-statistic: 262.5 on 11 and 510 DF,  p-value: < 2.2e-16
```



I decided to keep my own interactions for originality. All the covariates I considered (interactions included) are significant, but I want to go further and confort my choice by going through model selection processes. Concerning the residuals, they seem to confirm gaussian distribution, non linear pattern and homoscedasticity of the residuals. So the assumptions necessary for our model to hold seem to be verified in our model. All good.

Now for outliers, although the residuals vs leverage plot shows that we should have nothing to worry about since all the data fall within cook's distance, there seems to be some points that are redundantly identified as outliers with R. Let's have a closer look.



Seems one point is really out of the lot from the cook's distance plot, could be the point 495 or 334 we saw with our residual plots earlier, let's see what they look like!

```
## # A tibble: 2 x 11
##    Price  Sqft Bedroom Bathroom Airconditioning Garage  Pool YearBuild
##    <dbl> <dbl>   <int>    <int>           <int>  <int> <int>     <int>
## 1 2.00e5 1370.       4        1               0      1     0      1925
## 2 1.46e5 1412        1        2               1      0     0      1920
## # ... with 3 more variables: Quality <fct>, Lot <dbl>, AdjHighway <int>
```

The points identified as potential outliers are the only few data we have for very old houses, I don't think it's a good idea to take them as from the graphs they only get our attention but don't seem to be much of a problem. To be sure we can run a robust linear regression and see if that really improves anythbing. We'll confirm our choice of model by looking at selection methods. Note: I won't run this robust regression here since there is a conflict between the package MASS and dplyr, it breaks my whole project. Although I checked for it and saw that there was not much difference in models.

```
# library(MASS)
# Model_final_robust <- rlm(realestate_transf$Price ~. +Garage*Bedroom +Airconditioning*Pool -Aircondit
# summary(Model_final_robust)
# par(mfrow=c(2,2))
# plot(Model_final_robust)
```

In the end, we already confirmed that our reduced model was better than the model taking all covariates, so I'll compare my final model with my reduced model to choose upon both. I'll also run the global criterion which tests all sub-models of an original model for the Cp and AIC criterion.

```
##
## Call:
## lm(formula = Price ~ . - Airconditioning - AdjHighway, data = realestate_test)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51086 -0.09747 -0.01534  0.09234  0.60846
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.435332   1.252725  -1.944 0.052442 .
## Sqft         0.702708   0.046573  15.088  < 2e-16 ***
## Bedroom      0.071304   0.023869   2.987 0.002950 **
## Bathroom     0.037307   0.012438   2.999 0.002836 **
## Garage       0.134710   0.037021   3.639 0.000302 ***
## Pool        -0.637321   0.175051  -3.641 0.000300 ***
## YearBuild    0.004044   0.000571   7.082 4.73e-12 ***
## Quality2    -0.316982   0.028762 -11.021  < 2e-16 ***
## Quality3    -0.364537   0.040306  -9.044  < 2e-16 ***
## Lot          0.141214   0.019278   7.325 9.38e-13 ***
## GB          -0.031355   0.010430  -3.006 0.002775 **
## AP           0.710182   0.177138   4.009 7.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.169 on 510 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8466
## F-statistic: 262.5 on 11 and 510 DF,  p-value: < 2.2e-16
```

```
##
## -----------------------------------------------------
##   GLOBAL VARIABLE SELECTION PROCEDURE
##
##   ( Data =  realestate_test )
##
##   A = Sqft
##   B = Bedroom
##   C = Bathroom
##   D = Airconditioning
##   E = Garage
##   F = Pool
##   G = YearBuild
##   H = Quality
##   I = Lot
##   J = AdjHighway
##   K = GB
##   L = AP
##
##   Models        |  Cp            |  AIC           |
##   -----------------------------------------------
##   ABCEFGHIKL    |     11.33  ( 2) |  - 362.58  ( 2) |
##   ABEFGHIJKL    |     17.68  ( 9) |  - 356.13  ( 9) |
##   ACDEFGHIJL    |     17.02  ( 8) |  - 356.80  ( 8) |
##   ABCDEFGHIKL   |     12.16  ( 4) |  - 361.77  ( 4) |
##   ABCEFGHIJKL   |     11.03  ( 1) |  - 362.93  ( 1) |
##   ABCDEFGHIJKL  |     12.00  ( 3) |  - 361.99  ( 3) |
##   ACEFGHIL      |     16.72  ( 5) |  - 357.08  ( 5) |
##   ACDEFGHIL     |     16.81  ( 7) |  - 356.99  ( 7) |
##   ACEFGHIJL     |     16.78  ( 6) |  - 357.02  ( 6) |
##   ACEFGHIKL     |     18.28  (10) |  - 355.52  (10) |
## 
## -------------------------------------------------------
```

```
## 
## -------------------------------------------------------
##   GLOBAL VARIABLE SELECTION PROCEDURE
## 
##   ( Data =  realestate_test )
## 
##   A = Sqft
##   B = Bedroom
##   C = Bathroom
##   D = Airconditioning
##   E = Garage
##   F = Pool
##   G = YearBuild
##   H = Quality
##   I = Lot
##   J = AdjHighway
##   K = GB
##   L = AP
## 
##   Models        |  Cp            |  AIC           |
##   ---------------------------------------------------
##   ABCEFGHIKL    |     11.33  ( 2) |  - 362.58  ( 2) |
##   ABEFGHIJKL    |     17.68  ( 9) |  - 356.13  ( 9) |
##   ACDEFGHIJL    |     17.02  ( 8) |  - 356.80  ( 8) |
##   ABCDEFGHIKL   |     12.16  ( 4) |  - 361.77  ( 4) |
##   ABCEFGHIJKL   |     11.03  ( 1) |  - 362.93  ( 1) |
##   ABCDEFGHIJKL  |     12.00  ( 3) |  - 361.99  ( 3) |
##   ACEFGHIL      |     16.72  ( 5) |  - 357.08  ( 5) |
##   ACDEFGHIL     |     16.81  ( 7) |  - 356.99  ( 7) |
##   ACEFGHIJL     |     16.78  ( 6) |  - 357.02  ( 6) |
##   ACEFGHIKL     |     18.28  (10) |  - 355.52  (10) |
## 
## -------------------------------------------------------
```

The following function should work on models built on dataframes that have not been altered by data tidying with dplyr. Unfortunately most of my data manipulation was done with dplyr so this must be the source of the problem. Excellent function though, I'll use it in the future.

The Cp and AIC Both indicate that my model is close to to the best model according to the Cp and AIC criteria (2nd model out of 10 best, they recommend plugging back AdjHighway) So seems we did a good job! With my final model, we should be able to predict House sale price with a certain level of confidence.