Report

Alexis Laks

16/11/2018

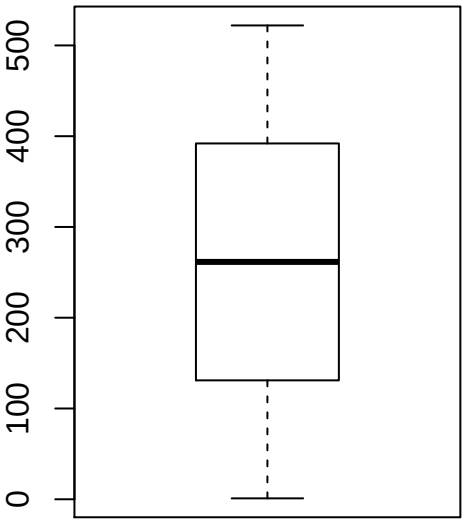
This is my report, you will find here the bulk of my research. I invite you to check the appendix if you want to get further details on my interpretations/conclusions.

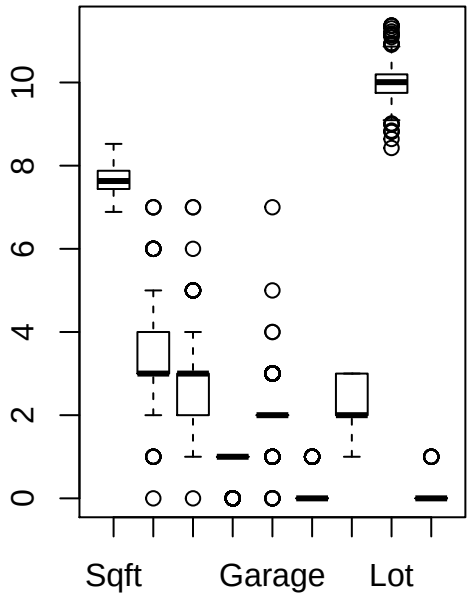
Introduction

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

I started by looking at the data and how it's structured. When looking at only **Price**, I saw that there was skewness in the distribution of the data which led me to considering a log transformation of the prices. Doing this transformation changes all the interpretation of my models, which I will take into account. I then checked for the distribution of our potential covariates. I saw that the scales of the data were very different as well (values around 5000 for **Sqft**, whereas max of Bedroom was around 7 or 8). I decided then to apply log transformations on the two covariates who had very different scales, namely **Lot** and **Sqft**. I also decided to transform the quality variable into factors to consider them not as levels but as unitary contributors to the model.





Now that I've gone through transformation of my covariates, I decided to check for any multicolinearity problems. To do that I plotted the matrix of correlation of matrices to detect any potential redundant variable. Some variables such as Quality, Bedroom and Bathroom showed strong correlation so to be sure I

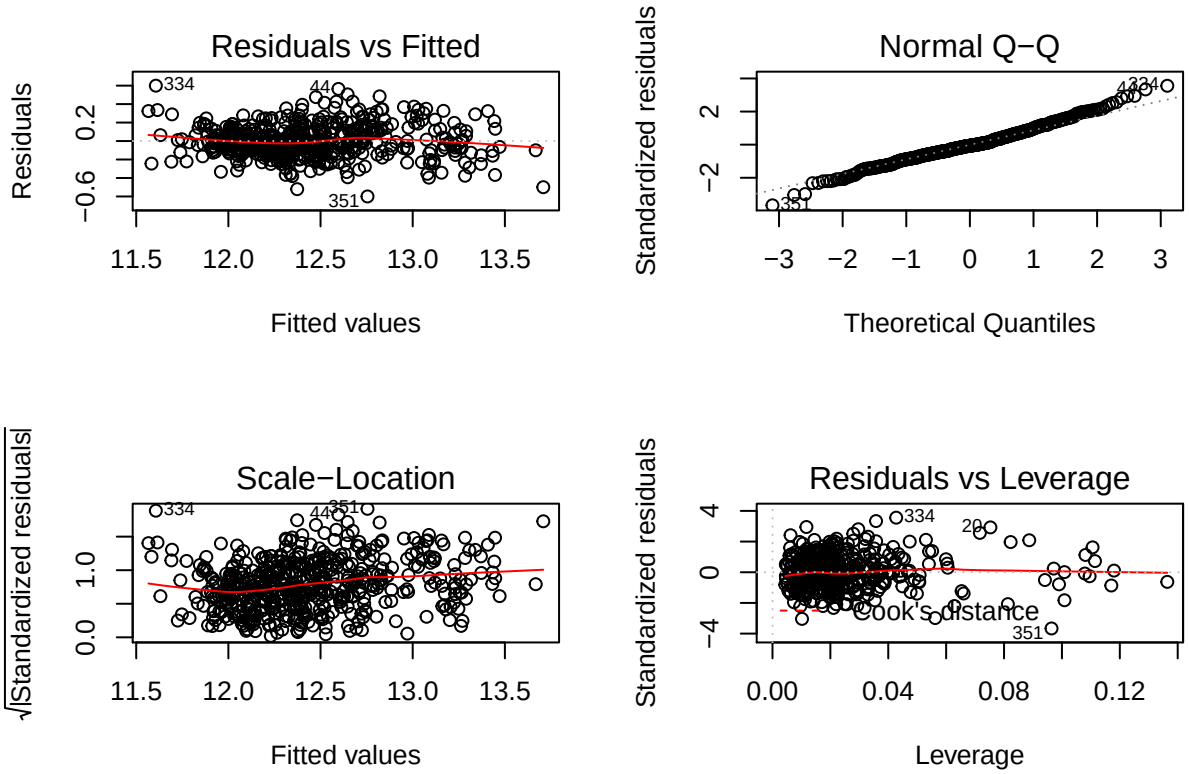
could keep them, I checked the varaince inflation factors. This test showed that all the factors were below 10 and the mean of factors was below 5 so nothing to worry about.

The Model

Full Model

I started out with a regression on all the covariates available to us, from which I worked down to a more optimal solution of there is one.

```
##
## 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 my first model is there are 4 non significant variables: the *Intercept*, *Bedroom*, *Air-conditionning* 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 seem 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 I saw in our data before that I did not have much values for houses near a Highway, I though that 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. I also thought that 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.

After taking out the variables *Bedroom* & *AdjHighway* I 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, I discuss this potential problem later on.

In order to really confirm that I have improved my model with the deleted variables I choose, I ran an F-test to confort me in my choice.

```
## 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 I obtain (greater than any tolerable choice of alpha) indicates that I don't reject H0, so deleting the variables I selected seems to make sense here. Although my model seems significant and the residuals seem to pass the conditions required for a valid model, there may be more at stake in **Price** prediction. So I went 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. And as expected, this inconvenience we mentionned does seem to have its effect on price, and the interaction is very significant in our model, although the variable Airconditioning lost its significance. 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 later on.

3. School quality impacts the price positively:

This variable should impact positively the price if we are facing a clientèle which have children where surely the school quality will be a criteria of selection among houses, and in turn will be a price argument for the salesman. When fitting a linear relationship between both we see 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*.

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.

After checking the fit, 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, there is a strong positive realtionship but for parking spaces exceeding 3 we have very few data to confirm our hypothesis. This can affect the fit to the model but it's still worth a try.

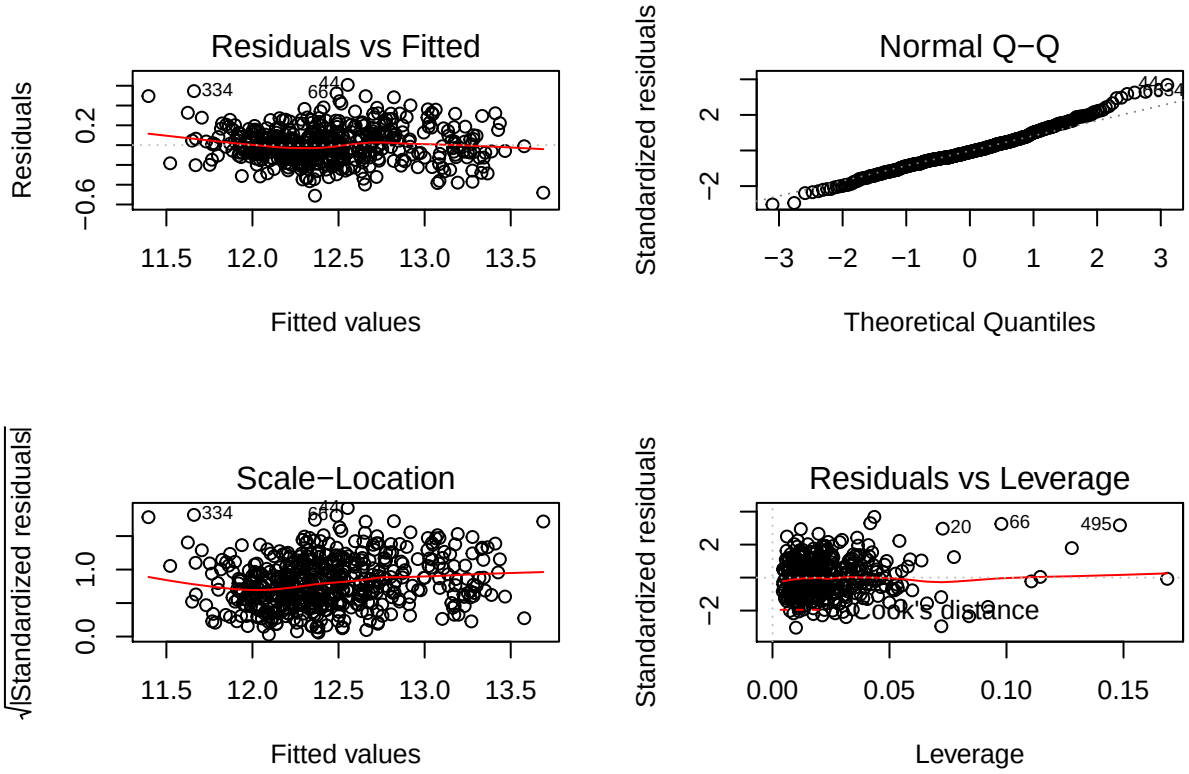
This interaction seems to be significant as well! Although we lost significance for the intercept (not that big of a deal, I'll just let the intercept be the intercept) and Airconditioning is non-significant once again.

MY MODEL

So my thought is that the interactions I've tested are all significant but I'm not sure they might all be 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.

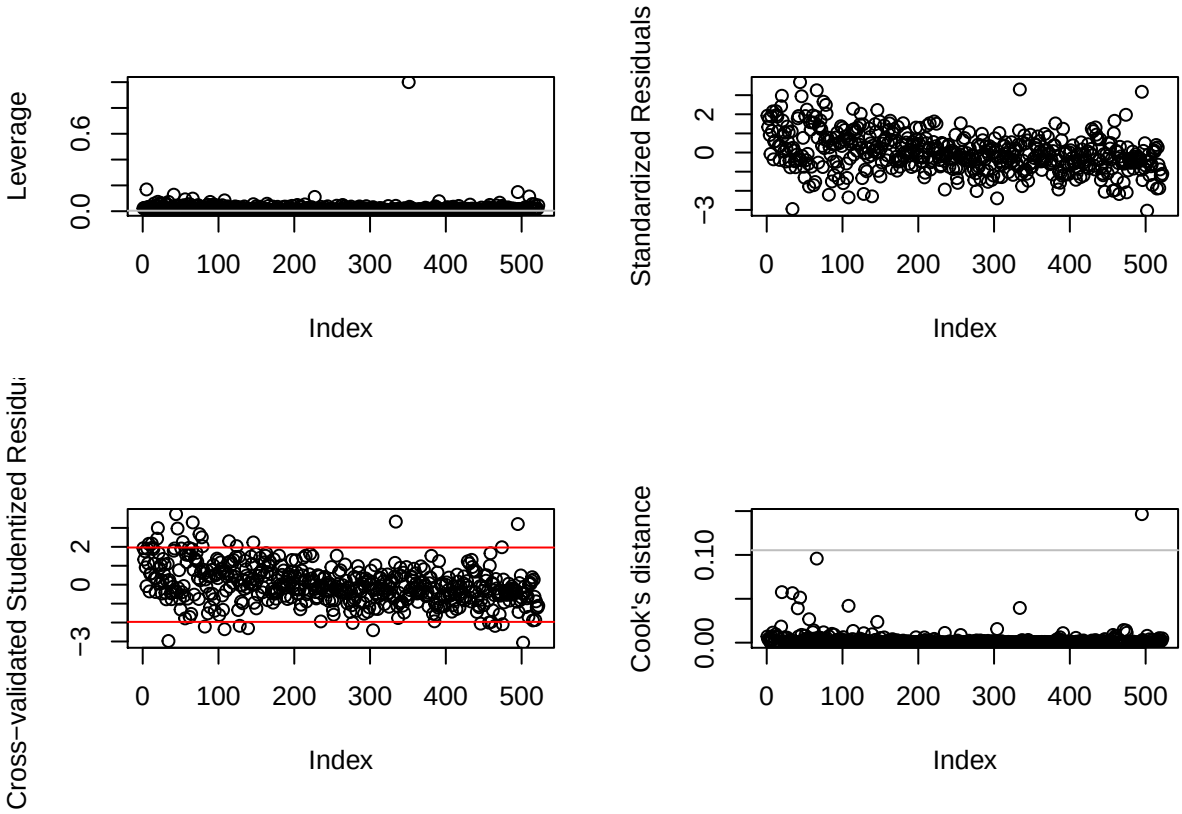
```
##
## Call:
## lm(formula = realestate_transf$Price ~ . + Garage * Bedroom +
##     Airconditioning * Pool - Airconditioning - AdjHighway, data = realestate_transf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51086 -0.09747 -0.01534  0.09234  0.60846
##
## Coefficients:
##                      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.

Before I move on to model selection I want to check 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. I'll confirm my choice of model by looking at selection method criterions.

```
##
## ------------------------------------------------------
##   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) |
##
## ---------------------------------------------------

## [1] -285.7499 -296.5644 -305.2265 -305.2265
```

The GlobalCrit function that tests all combinations of variables in a regression indicates 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! We could choose the first one in the list but the deviation in Cp and AIC Between my model (2nd) and the 1st model in the list isn't that big. I prefer staying with less covariates but which are all significant. We see the same thing with thte BIC, the difference isn't big so I'd just stick with my model. Ready to be a realestate agent?