Data Analysis Project

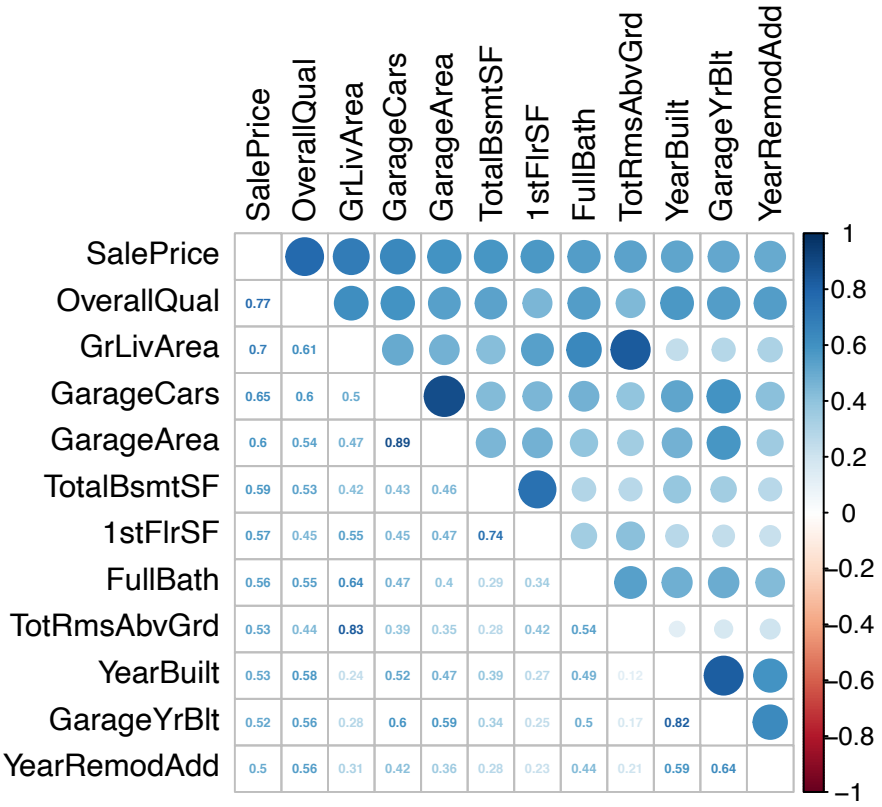
1. Introduction ###### The project consists in predicting the price of a house. We are given two datasets, **train** and **test**, each containing 68 variables (quantitative and qualitative) describing the characteristics of the house. We will try to use to a linear model to solve the problem and hence, in the first step, we will select a few variables using standard techniques such as forward/backward selection and Lasso, but we will also be considering a transformation of the dependent variable SalePrice. To verify the precision of our models, we will analyze the **Diagnostic Plots** and will compare the performances. The hypothesis that we propose is that the logarithm of the dependent variable strongly depends on the numerical ones and some few categorical ones.

#2. Exploratory Data Analysis

First, we print the summary of the training set to remark some general properties.

We notice that all the variables are normalized in the interval [-1, 1], which is what we expected from the dataset since we are starting with the file DataProject.RData.

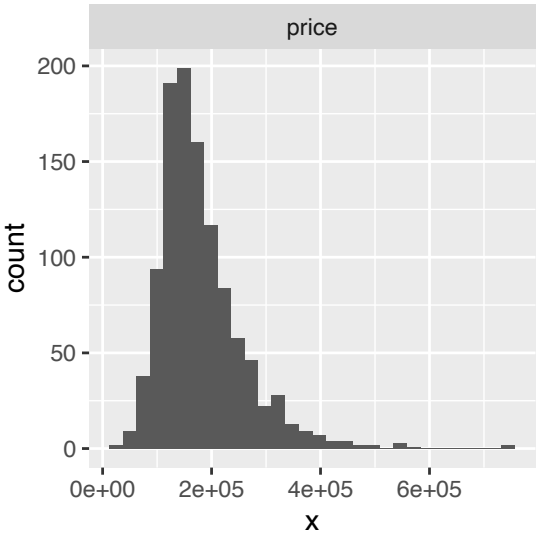
Naturally, we would like to find the most correlated variables with the target variable SalePrice. Since there are some categorical (qualitative) variables, we will ignore them in this first introductory analysis.



We find 29 numerical variables and we notice that some of the most correlated variables are OverallQual, GrLivArea, GarageCars, GarageArea. Naturally, most of these variables will be also found in the next sections when we use forward/bacward selection or Lasso to select the variables for our linear model.

Similarly, we find some very intuitive results such as the correlation of GarageCars and GarageArea, 0.89, since both are just two different ways to measure the same area (in cars and square feet). It is the same case for YearBuilt and GarageYrBuilt, the correlation in this case is 0.82. The explanation is also intutitive, when we build a garage for a house, we usually do it at the same time (the same year) than the house itself.

Finally, we plot the target variable SalePrice and notice that they are skewed. It is what leads to think that a transformation of the variable is necessary to make it more symmetrical and hence appropriate for a linear regression model.

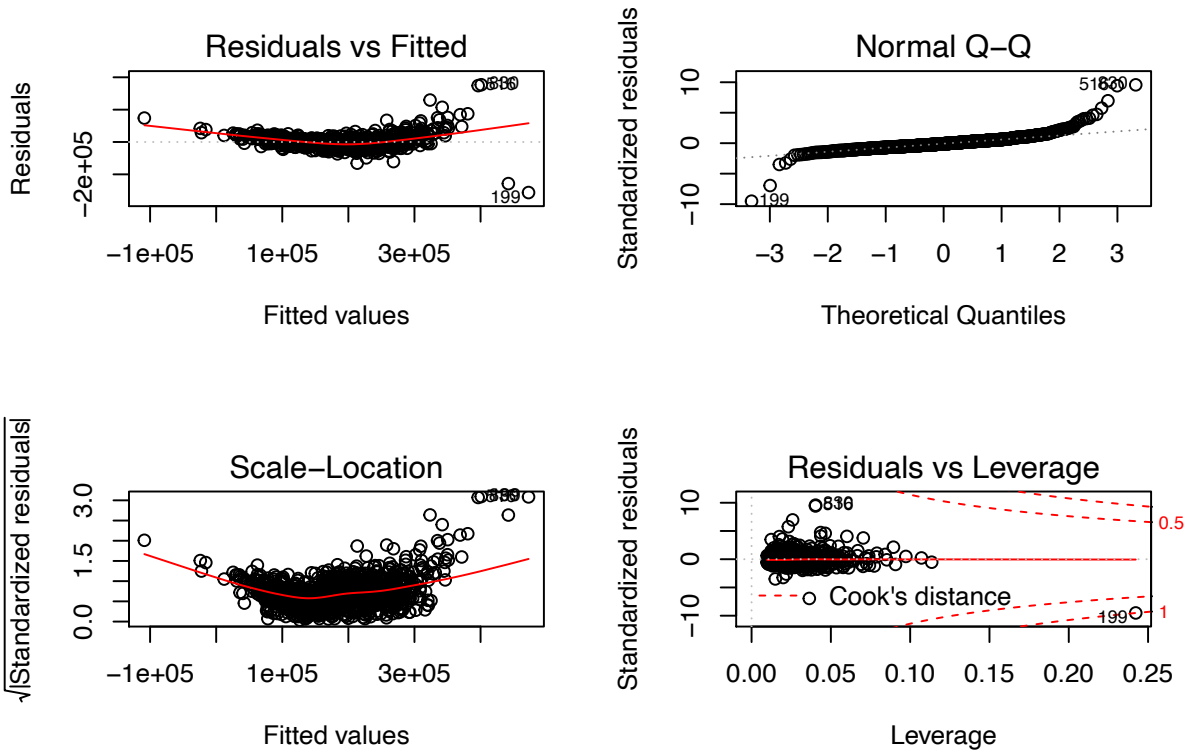


#3. Modeling and Diagnostics

Given that at the beginning, we have both quantitative and qualitative variables, the first model will consist in taking only the quantitative variables which are easier to model (we already use this idea in the previous section to find some correlations between the explicative variables and SalePrice).

The following diagnostic graphs corresponds to the simple model with only numerical variables given by :

fit <- lm(formula = SalePrice ~ ., data = train_numVar)

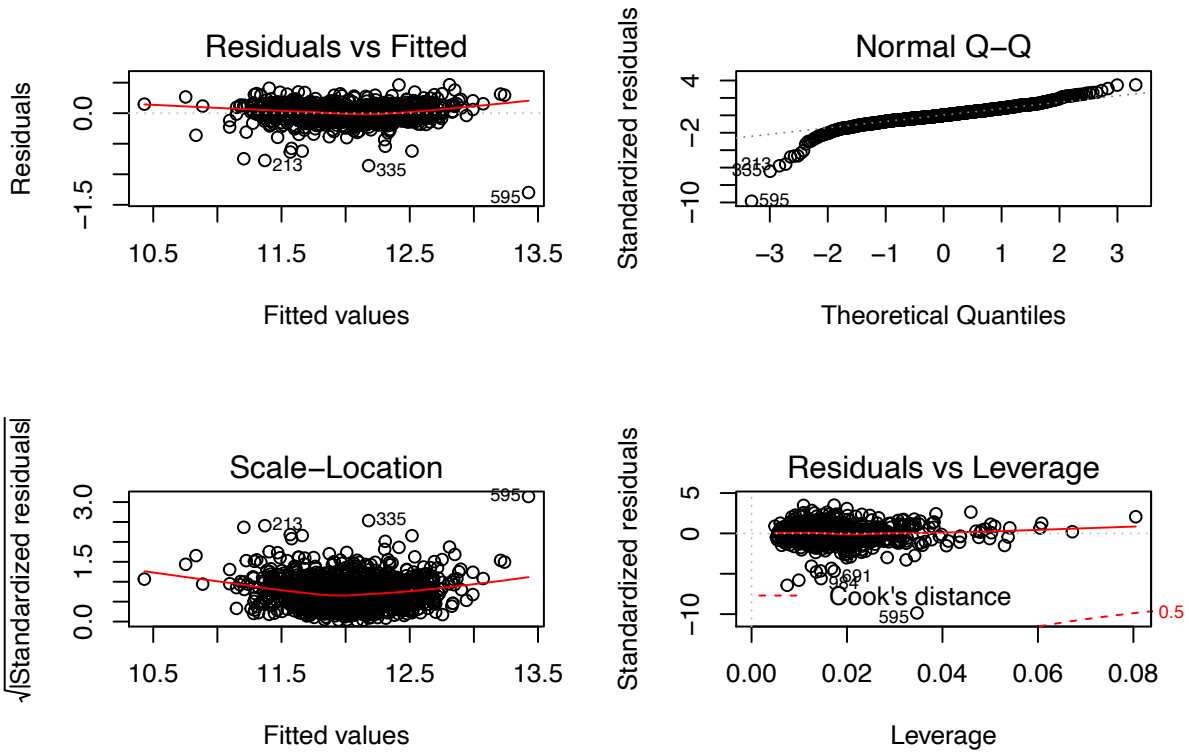


We find the 29 numeric features which were also found before, but the Residuals vs Fitted graph is not accurate at all since the distance to the horizontal line $y = 0$ is very clear (it is of the order of 10^5). This also suggests that we strongly need to use a transformation of SalePrice to reduce the current difference in the model. At the same time, due to the 4th graph Residuals vs Leverage. we notice that 199 is an outlier. That's why we are going to remove it from the set train.

The first non trivial models are constructed using stepAIC (forward/backward selection) and we will consider the familiy of Box-Cox transformations for the target variable SalePrice. Box-Cox family of transformations:

$$Y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \text{if } \lambda \neq 0 \\ \log(y), \text{if } \lambda = 0 \end{cases}$$

The following diagnostic graphs are of the model removed outlier(point 199), applied box-cox transformation to the SalePrice and a forward-backward selectionAIC:



As we can see, the results are quite good. The Residuals vs Fitted graph is precise enough (notice that the scales in Residulas are as small as 0.2). The Q-Q Plot is also a line (not completely at the tails, but the main part is) and then it already suggest that indeed, the residuals are following a gaussian distribution. The Scale-Location graph is probably the less acurate among the 4 graphs obtained, but anyway it's not very far from being a line (notice again the scale of Residuals) Finally, the residuals vs Leverage graph indicates, as expected, that now there are no outliers anymore.

Therefore, we can say that this first model (BoxCox + StepAIC) is accurate enough to measure it's performance with the test dataset and the value of R^2 is about 0.907, which is already pretty good. Unfortunately, the MSE (Mean Square Error) is still very high, it's about 7e8, this is due to the exponential function that must be propagated to the residuals too.

After considering the numeric variables only, we return back to the initial model and will consider again the categorical variables to find some improvement. We follow a similar procedure.

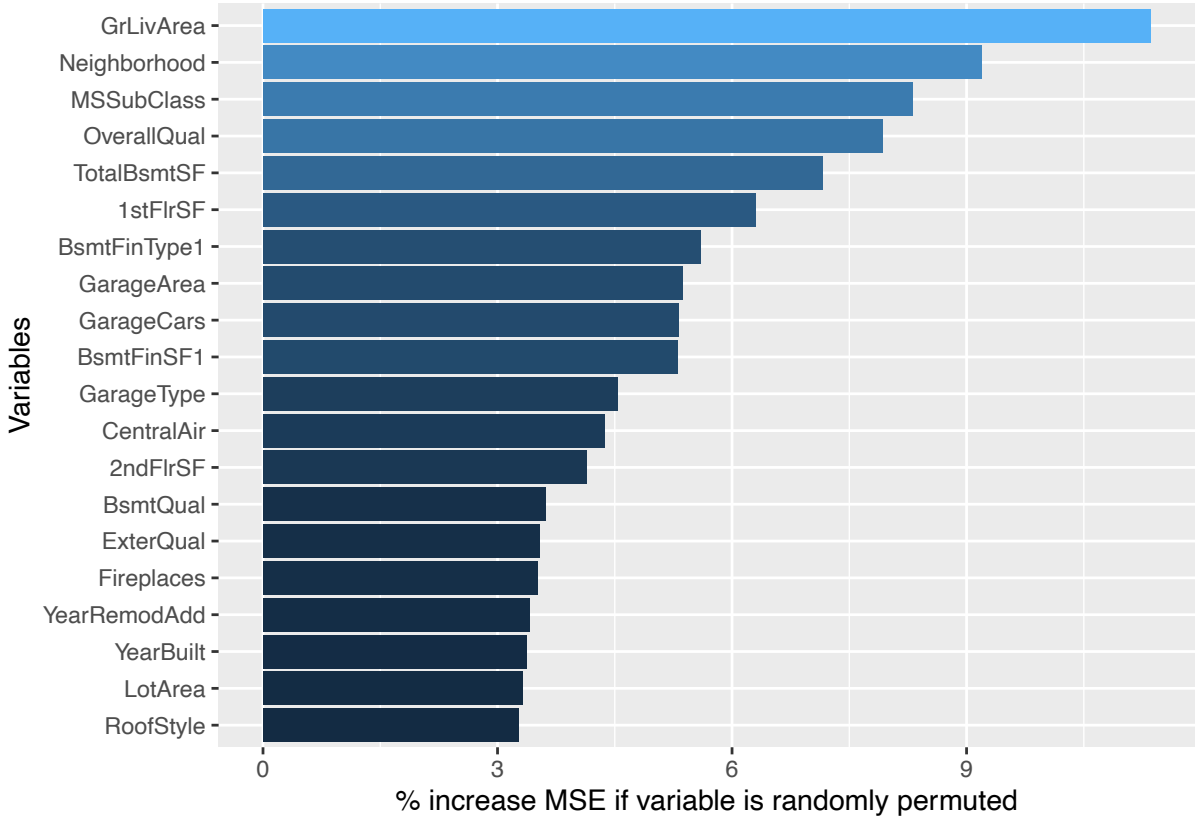
From the above table, some categorical variables are suggested by R: MSSubClass, Neighborhood, ExterQual, KitchenQual, RoofStyle, GarageType.

We also notice that sometimes several categorical variables are correlated, like GarageArea, GarageQual etc.

RandomForest algorithm is a powerful ensemble machine learning algorithm. RandomForest is a ensemble of many independent decision trees. While minimise the loss function, the desicion trees "learn" the best criterion to split the class. For instance, if there are 2 GarageArea, then the value will be larger than 22,000. For each decision tree, there will be several splits for a given observation. And the observation finally belongs to one leaf of the tree with a value. The final prediction value will be determined by all trees in the ensemble.

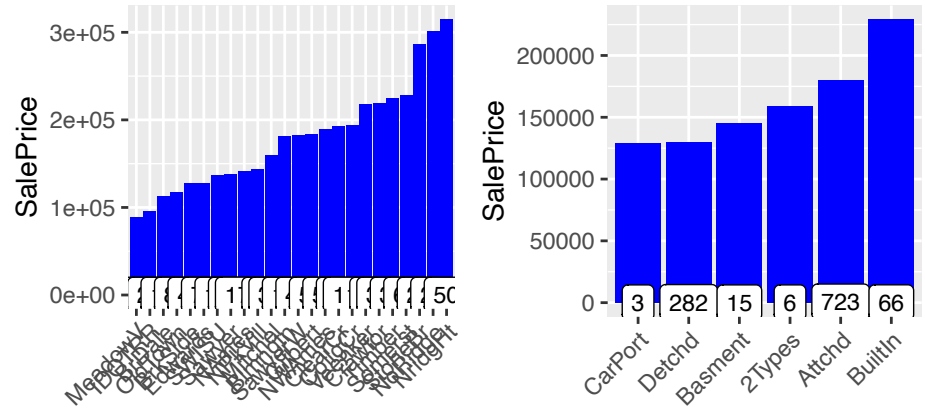
So we could use the RandomForest algorithm to do a first feature selection for us, just to see what kind of influence it will have in introducing the categorical variables. We do this in the whole dataset.

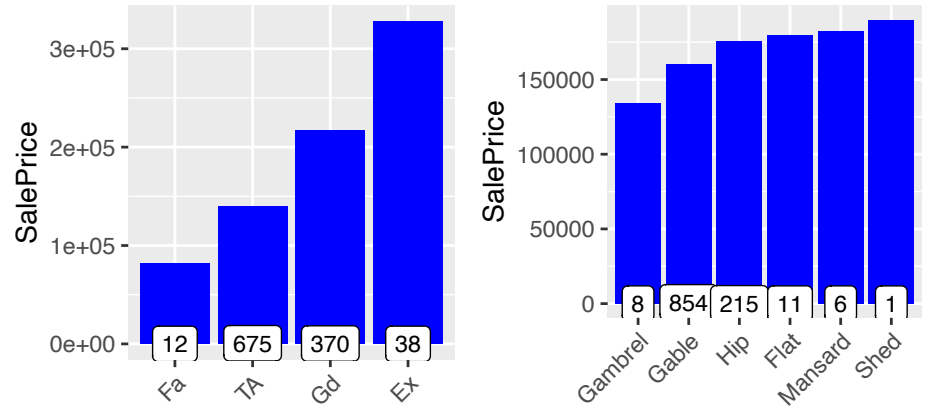
RandomForest suggest the most important 20 variables as follows:



Then, we select the variables both suggested by ANOVA and the random forest regressor, that's to say: Neighborhood, GarageType, ExterQual, RoofStyle.

Take a glimpse at these categorical features:





We do a regression only with the selected categorical variables. R knows transform the string to numerical variables. We find that only with these four categorical variables, we have a $R^2 = 0.70$, which can be interpreted by the existed of linearity: there are indeed useful informations in the categorical variables.

It is logic since the house price depends on many internal and external conditions. The internal conditions like the garage type and roof type (better design cost more); the external conditions like whether it finds itself in a nice resident area. We hope to have a better residual distribution in combaining numerical and categorical variables.

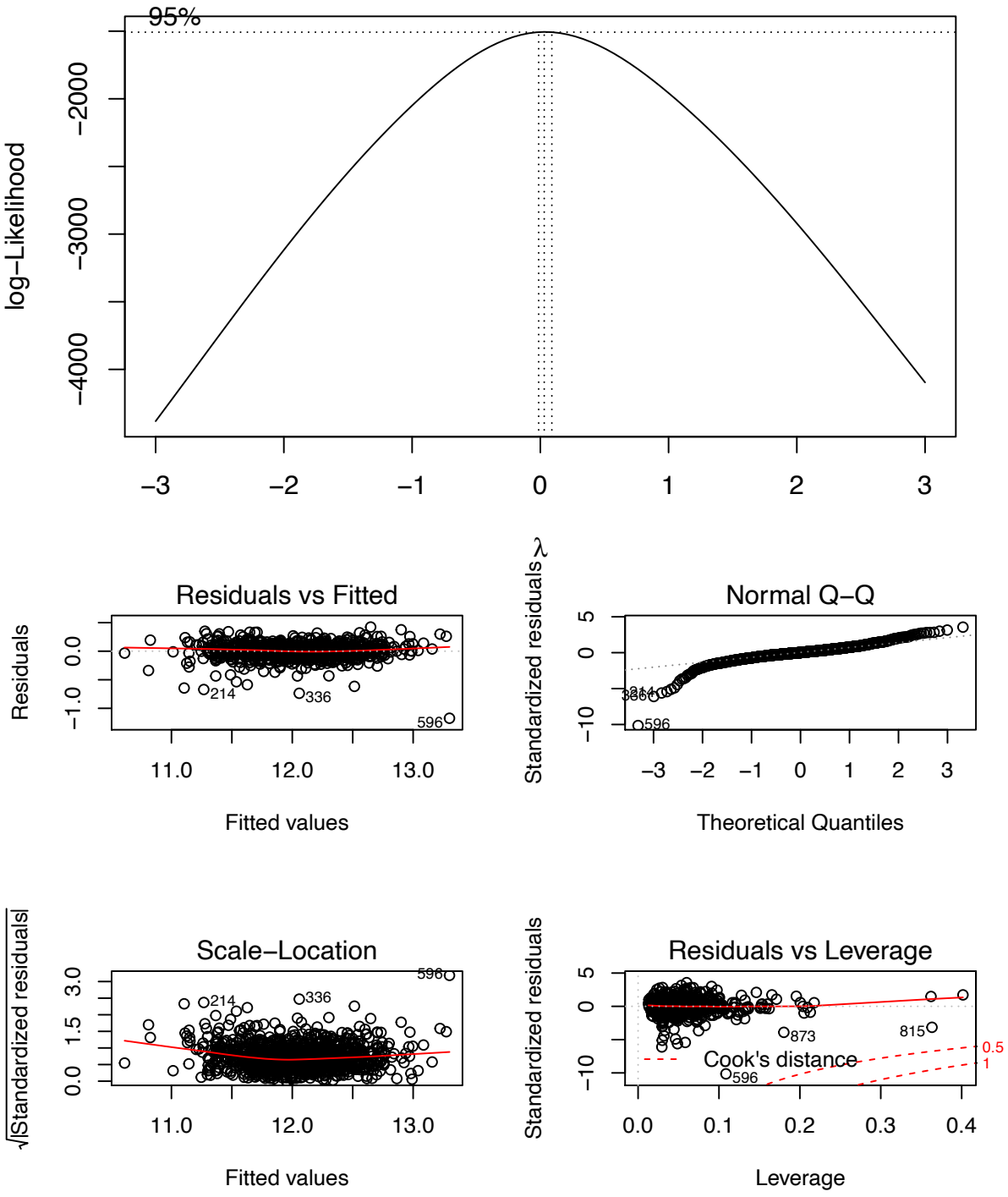
In comparing with the case where on numerical variables are considered, we find that the residual graph is improved. Therefore we get the intuition that when more information is revealed, the residual is more randomly distributed around zero.

We also find that the point 199 seems to be an outlier, because it goes beyond the courbe of Cook's distance equals to 1/2. We will remove this point.

As before, to refine the model, we use a forward selection to select pertinent variables.

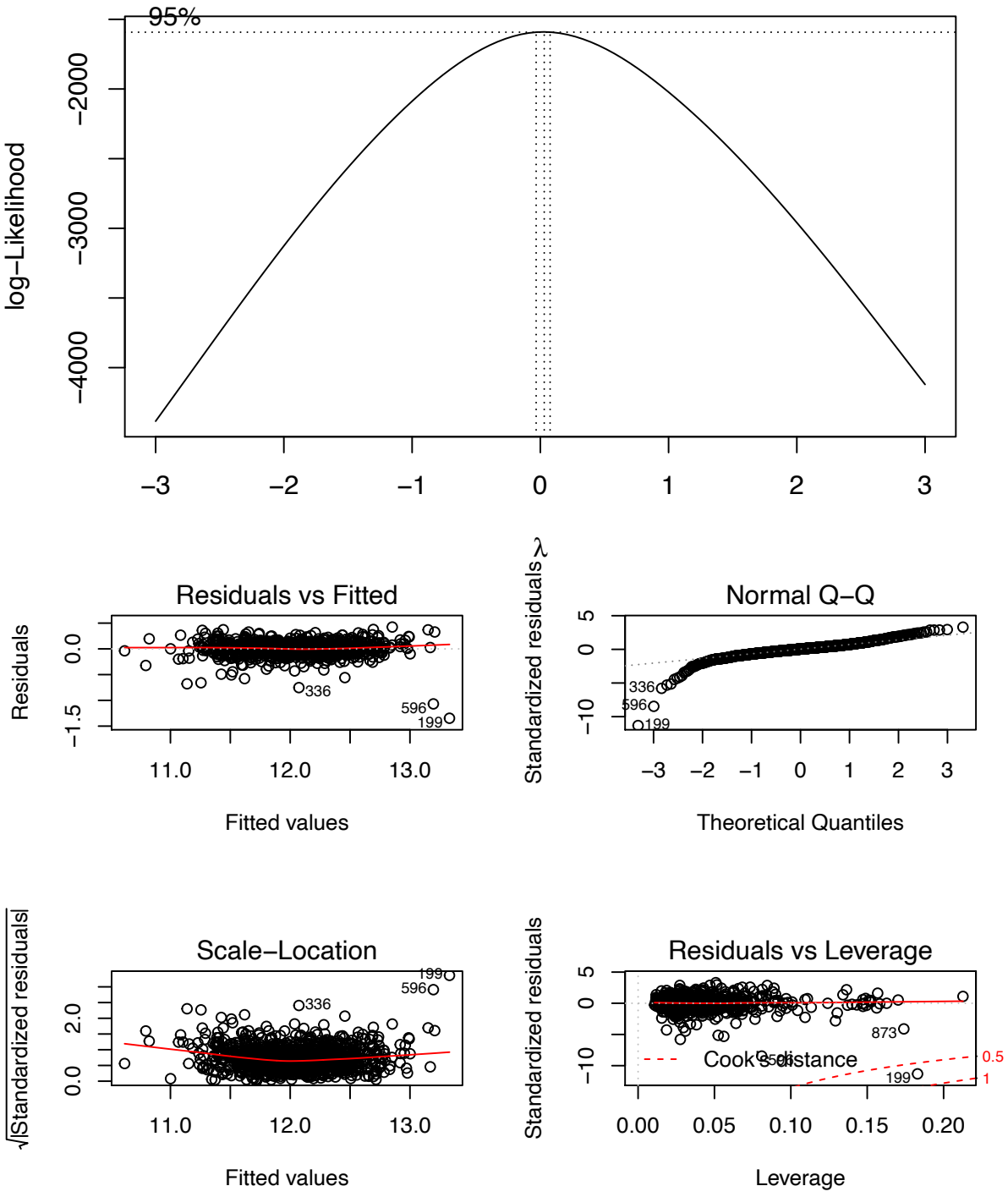
We remark that categorical variables Neighborhood and ExterQual have significant influence and selected by forward and backward selection. Certains values like: NeighborhoodStoneBr, NeighborhoodNoRidge, ExterQualTA are important. GarageType have less but also important influence to SalePrice.

In the new obtained model, we will also do box-cox transformation. We find the best $\lambda = 0.03$. However, for practical reasons, we will take $\lambda = 0$, which is still in the 95% confidence interval.



We will do the same thing, but using the dummy variables for the 4 selected categorical variables. We have in total 29 numerical variables and 38 one-hot dummy variables.

Similarly to precedent case, we apply a forward, backward selection; a box-cox transformation for SalePrice; remove outlier. The diagnostic graph of this model is shown as below:



We remark that with this dummy method, R^2 even increased. Which means too many variables harms the linearity. But we are glad to see that the Residual plot tends to be very good.

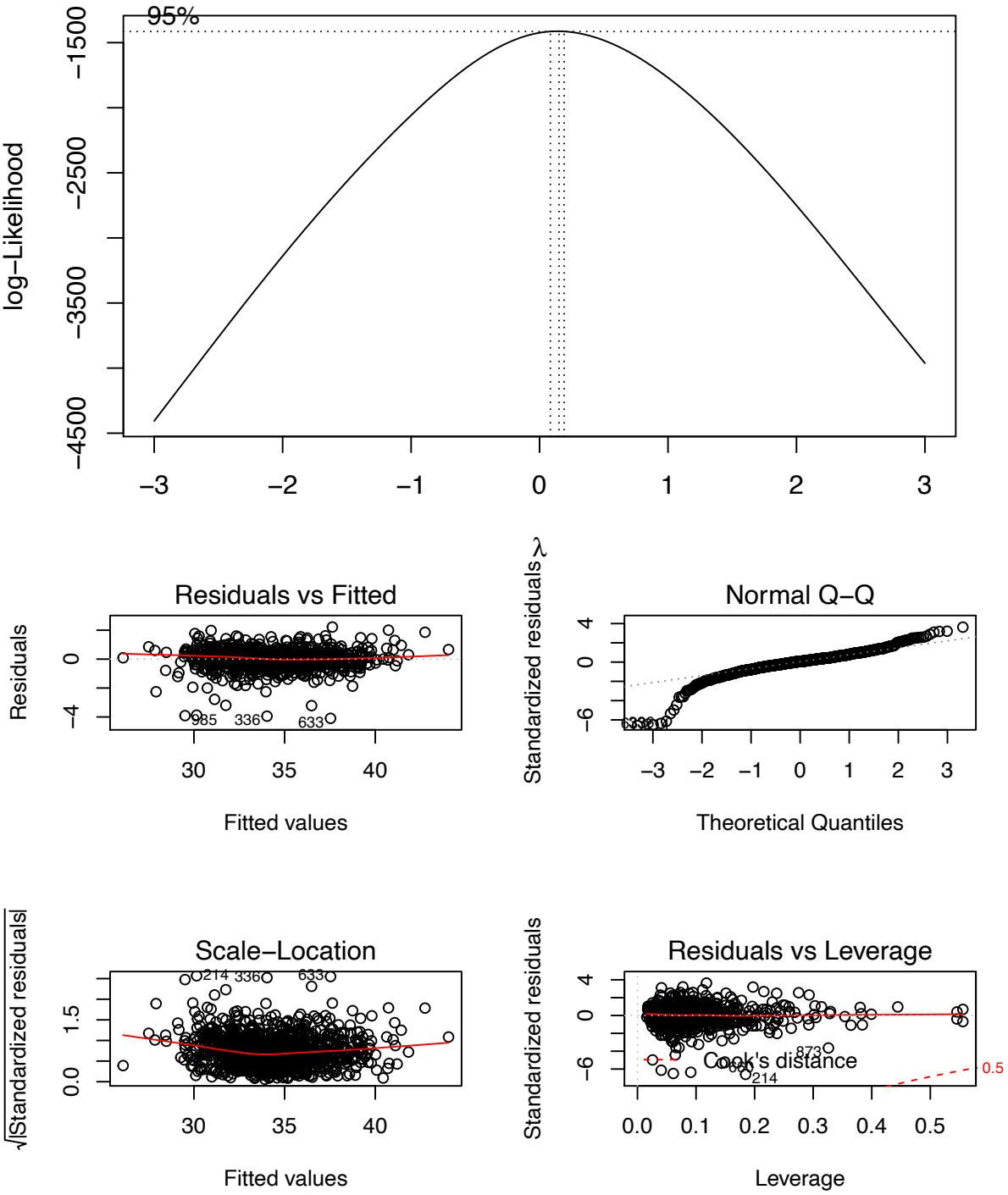
From the above models, we find that increase the numbers of covariables is not a garantee for improving the performance. And therefore using the one-hot variables introduces too many variable and harms the performance. To verify this idea, we will do it over all categorical variables while using a stronger covariable selection method: LASSO. The model.matrix has been used for the one-hot transformation.

This has created 211 one-hot variables.

We will now use lasso regression to select pertinent variables. The fine-tuning yield lambda to be around 1000(ranging from 700 to 1200). It is very large, which implies that a very large penalisation has been used. We will return to this point later. And there are about 100 variables are selected (note that each time the result is not deterministe).

We will now to use the 94 variables for regression.

Similarly to precedent, we remove the outlier and do the box-cox transformation, and we get the diagnostic graph as follows:



We remark that after this effort, the residual distribution looks better, and so as the residual vs Leverage plot.

#4. Final Models ###### The first model studied which took into account the numerical variables only, the Box-Cox transformation for the SalePrice (more particularly, the function $log(SalePrice)$ is taken) and StepAIC (stepforward/backward selection) for the explicative variables provides an $R^2 = 0.907$, a $MSE \sim 7e8$ and F-test = 118.7 when tested with the dataset **test**. As we stated before, these results are already good and the main advantage is that the model is rather simple.

The second one is an improvement of the previous one which consits in adding 4 categorical variables (Neighborhood, GarageType, ExterQual, RoofStyle) selected by using a Random Forest Regressor and also because those were very significative for the ANOVA summary (all had a p-value $<$ 1e-5). The transformation for SalePrice and the StepAIC are kept for this model. In this case, we have $R^2 = 0.919$, $MSE \sim 6e8$ and F-test = 204.4 when tested with the dataset **test**. This is the model with the best performance that we have found in terms of R^2.

The third model is the most complex one. This time, we have taken all categorical variables into consideration, there are 240 in total (211 one-hot variables). After that, we decided to use Lasso method to select the explicative variables instead of stepforward/backward. The results were surprisingly not as good as expected and we ended up with a value of $R^2 = 0.893$, $MSE \sim 7e8$, F-test = 158.9 when tested with the dataset **test**. This results are probably due to an overfitting when selecting the variables using Lasso. We finally, tried to use an stepforward/backward selection over the variables selected by Lasso in the previous model, but unfortunately the results remained the same.

#5. Discussion ###### We have verified th hypothesis stated at the beginning of the report since the $log(SalePrice)$ can be linearly modeled by some explicative variables of the initial set.

Actually, we noticed that increasing blindly the number of explcative variables can lead to worst results even with better methods for the variable selection such as Lasso.

Some clues to improve the current models are to introduce terms of interactions, **variables croisees** ,given that as we remarked at the begining there are several correlated variables and to mix different models for ensemble learning.