MAP 535 - Data Analysis Project

Adrien Toulouse & Paul-Antoine Girard

Introduction

Our task is to analyze the dataset named **House Prices: Advanced Regression Techniques**. It contains the sale price of about 1500 residential homes located in Ames, Iowa, along with 79 explanatory variables describing (almost) every aspect of the houses. The dataset has already been preprocessed to deal with missing values, so we will work on a reduced dataset containing 68 variables.

Our aim within this project is to focus on dimensionality reduction by doing a variable selection. Variable selection can be defined as selecting a subset of the most relevant features.
The objectives of feature selection include: building simpler and more comprehensible models, improving performance, and preparing clean, and understandable data. Indeed, with a large number of features, learning models tend to overfit which may cause performance degradation on unseen data.

We can, therefore, address the following question: **What are the most relevant features to explain the sale price of houses of the dataset?**

To answer this question we will first analyze the variables and assess their relevance by looking at their correlation with the regression target: *SalePrice*. We will also build and compare several linear regression models with different numbers of variables and finally conclude on the relevance of the features. Our work will be focused on finding the best linear prediction model using a minimum number of variables. We can, therefore, state our research hypothesis as follows:
Can we construct a performant linear regression model by selecting only the most appropriate variables? How does it compare to larger or other models?

Note: This report condenses all our work. Not all graphs have been included for space issues. Please refer to the file with the full code if necessary.

Exploratory Data Analysis

1. Transformations

The histogram for the response variable *SalePrice* shows that it is skewed. So, a first transformation we do is to take the log of the house prices to reduce the effect of the tail in its density.

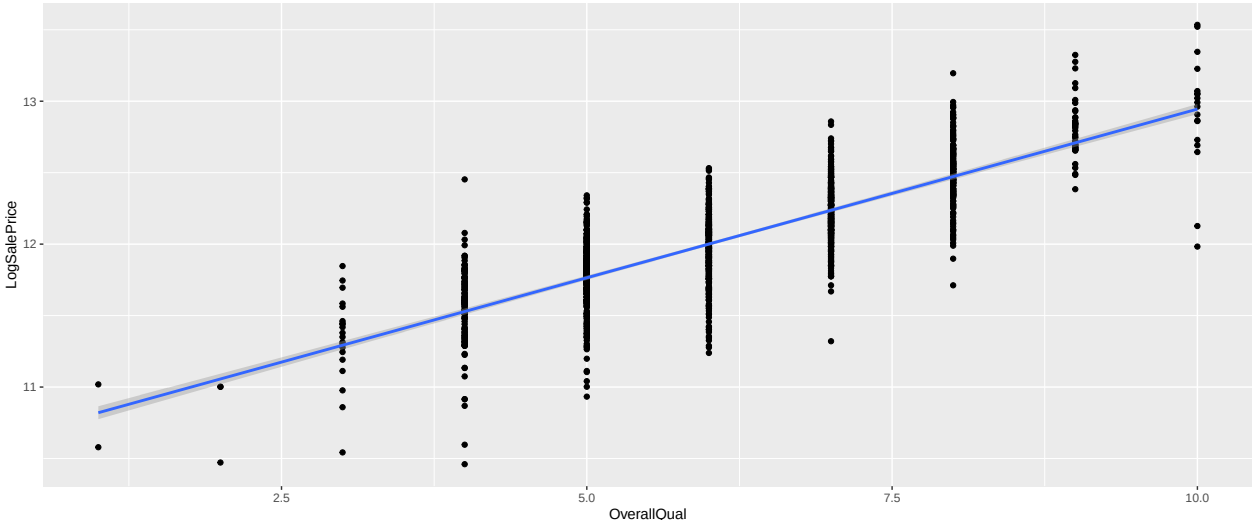
```
trainImputed$LogSalePrice <- log(trainImputed$SalePrice)
trainImputed <- select(trainImputed, -c("SalePrice", "X")) #X is unrelated to our study
trainPP$LogSalePrice <- log(trainPP$SalePrice)
trainPP <- select(trainPP, -c("SalePrice", "X")) #X is unrelated to our study
```

We observe the same particularity for the variables *LotArea*, *TotalBsmtSF* and *GrLivAreado*. So we do similar transformations for these variables.

2. Numeric variables

Looking at the numeric variables, we analyze the correlation between the different variables together as well as their correlation with *LogSalePrice*.
The first step is very important when trying to see which variables are the most important to explain price since there might be multicollinearity problems. Indeed we know that having correlated response variables is not efficient in linear models and detecting strong correlations will allow us to create a reduced model. Some variables like for example *YearBuilt* and *YearRemodAdd* are very strongly correlated. So we will not need to keep both variables.
Secondly, we look at the correlations with our response variables to see which variables explain the house sale

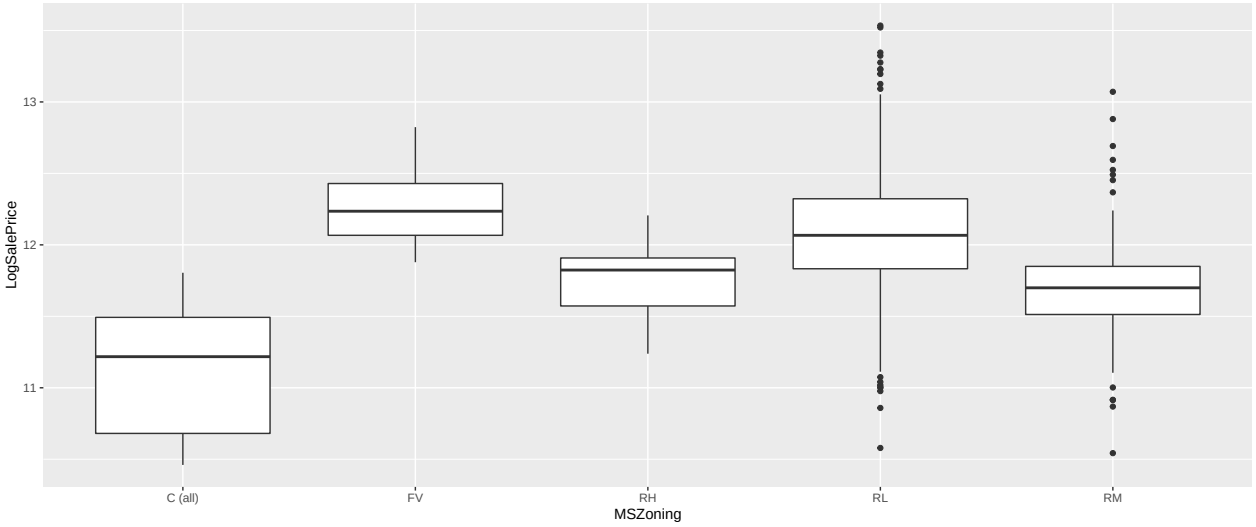
price well.
The regressor *Overallqual* is particularly interesting as it rates the overall material and finish of the house.



The scatter plots indicate a strong positive correlation between the two variables. By plotting other scatter plots, we found that the variables *YearBuilt*, *YearRemodAdd*, *MasvnrArea*, *BsmtFinSF1*, *X1stFlrSF*, *LogGrLiveArea* are also strongly correlated with the price. Including these predictors in our model should therefore be performant.

3. Factor variables

Concerning categorical variables, a variable will be interesting in our model if it has different boxplots for each category when considering *SalePrice*, as this will indicate a clear dependency between the two variables.



We remark that the variable *MSZoning* corresponds to this situation. Indeed, these boxplots are quite different visually from one another which indicates that MSZoning is an important variable to explain the Sale Price.

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## MSZoning       4  40.94  10.234   77.61 <2e-16 ***
## Residuals   1455 191.87   0.132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is confirmed by a one way Anova test as the p-value is less than the significance level 0.05. We can conclude that there are significant differences between the *MSZoning* categories when considering the *LogSalePrice* and this leads us to include *MSZoning* in our future model.

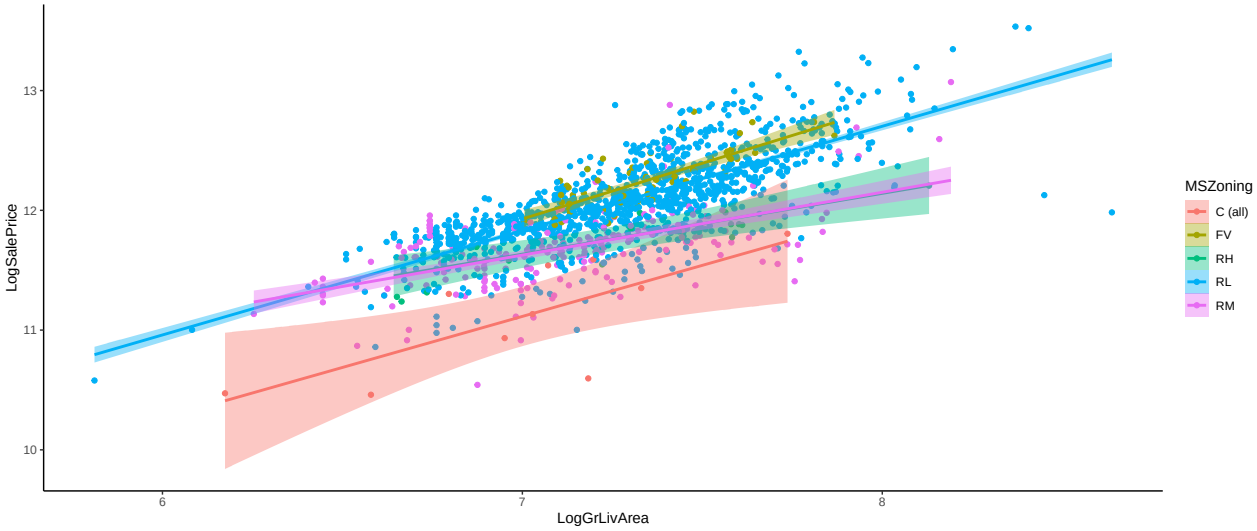
We find similar results when looking at the *GarageQual* variable. However, the boxplots do not completely follow intuition as they indicate that on average houses with garages in excellent quality have a lower sale price that garages in good quality. This reveals that garage quality is not a key variable when trying to explain the sale price of a house.

We also note that some categorical variables like for example *RoofMatl* are not very interesting to explain sale price as they are too heavily unbalanced (almost all the observations take the value Compshg and we have only one observation for some of the other roof material types). This is also the case for *Condition2*.

```
##
## ClyTile CompShg Membran   Metal    Roll Tar&Grv WdShake WdShngl
##       1    1434       1       1       1      11       5       6
```

4. Ancova analysis

We can also plot two quantitative variables together with one factor variable.



This graph helps to make clear that while *GrLiveArea* has a large predictive effect for *LogSalePrice* (the slopes of all the lines are clearly non-zero), there is also an effect of group assignment: for example the houses assigned to the FV MSZoning have a higher Sale Price than the houses assigned to RH.

5. Key findings from EDA

To sum up, our findings from this first part are the following:

1. We used a log transformation on the sale price to reduce the impact of the tail in its distribution.

2. The variables *YearBuilt* and *YearRemodAdd* as well as the variables *LotArea* and *LotFrontage* are highly correlated two by two.

3. The numeric variables *LogTotalBsmtSF*, *LogGrLivArea*, *OverallQual*, *OverallCond*, *LogLotArea* are interesting when explaining Sale Price because they are highly correlated with our response variables.

4. The factor variables *MSZoning*, *CentralAir*, *BsmtQual*, *KitchenQual* are interesting when explaining Sale Price because of the large difference in each category boxplots. Some variables like *Roofmatl* are

highly unbalanced which is not very interesting for our model.

5. We have confirmed our intuitions with statistical tests and have ploted numeric and factor variables together with Ancova plots.

Now that we have more information on our data, let's build multiple linear models. We will start from the full model and then use our findings from this exploratory analysis part as well as other techniques to select variables and build better models.

Modeling and Diagnostics

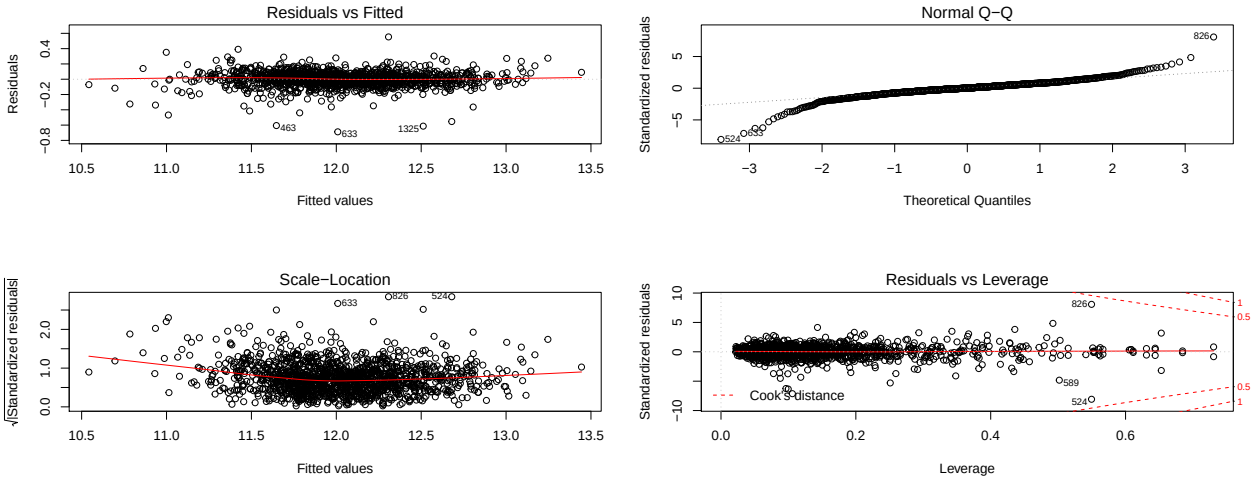
1. Full model

We start by doing a linear regression with all the variables of the dataset.

```
full_model = lm(LogSalePrice ~ ., data = trainPP)
```

As we explained in the introduction, our interest in this study is to select variables that explain the best our model. Some of the variables are irrelevant if we consider the p-values related to Student test. At a significance level of 0.001, this technique suggests us to only keep the following variables: *MSZoning*, *LotArea*, *OverallQual*, *OverallCond*, *YearBuilt*, *YearRemodAdd*, *RoofMatl*, *TotalBsmtSF*, *CentralAir*, *GrLivArea*, *KitchenQual*, *Fireplaces*, and *GarageQual*. Concerning the qualitative variables, we decide to only keep the one that have many categories that are relevant for the model at a significance level of 0.001, and not only one category (that is the case for the followings: *Condition1*, *Condition2*, *Heating*, *Functional*).
Overall, this full model has a R^2 coefficient of 0.94 (can't be improve since we can't add new variables), a R^2_a of 0.93, and a AIC of -2314.3. The F test statistic yields a very low p-value, that shows that the model is meaningful at a level of 0.05.
Let's check if the residuals verify the postulates of the linear model.

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



The residuals seem to have a mean of zero and they are uncorrelated. However, the other assumptions do not look verified. We check it by running a Breush-Pagan, and a Shapiro-Wilk tests. Both of them give a p-value that is lower than 0.05. So, the hypotheses of homoscedastic variance, and gaussian distribution are rejected at a significance level of 5%. Finally, none of the residuals have a cook distance larger than 1. Note that R informs us that it didn't plot a few points that have a leverage of one. These points mean the fitted value corresponds exactly to the observed value. Since we have many regressors in this model, this is explained by the fact that certain combinations of modalities are associated to only 1 observation. Therefore, we need to work on our selection of variables to reduce the number of regressors, and define a reduced model that verifies the assumptions required for its validity.

2. Model using the backward method

We use the different selection methods based on minimizing the AIC to automatically select a reduced number of variables for our model.

The three methods (forward, backward, both) lead to the same model that has an AIC value of -2382.3. Comparing AICs this model is better than the full model, and also selects less variables. The F-test gives a low p-value, so it is meaningful. Finally, if we look at the Student tests done, we obtain the same variables as before and two additional variables: *BsmtQual*, *BsmtFinSF1*.

Now, let's take a look at the residuals to see if the postulates are now valid, and if the number of observations with leverage 1 is reduced.

The residuals still don't verify the assumptions needed for the validity of our model. As before, the mean seem to be zero, but the others assumptions aren't verified. The Breush-Pagan, Durbin-Watson, and Shapiro-Wlik tests give all of them a p-value that is lower than 0.05. In addition, there are still observations that aren't plotted because they have a leverage of one. Therefore, we need to reduce even more the number of variables. To do so, we will use our work from the descriptive statistics.

3. Reduced model based on the Student tests and our EDA work

In this part, we are going to construct a model based on the regressors selected by the p-values of the Student tests realized for each variable in the linear regression model obtained by the forward method, and by our work done in the first part of the study.

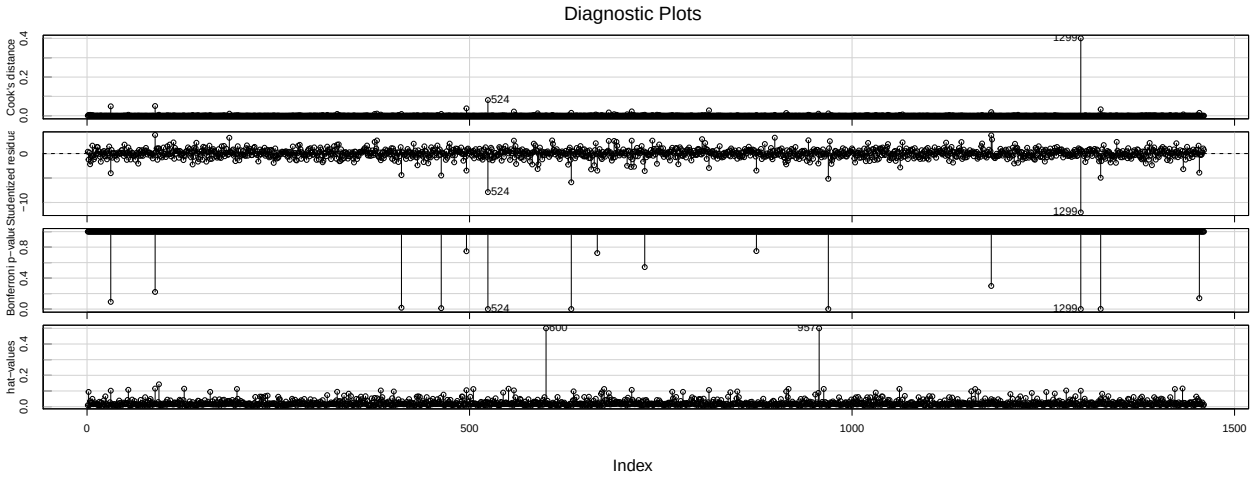
In order to obtain a model that verify the assumptions needed in a linear model, we first wanted to use the transformations made in the first part, but how the variables have been encoded don't let us apply the log transformations (some values are now negative).

Our base model contains all the following variables selected in the part II: *MSZoning*, *LotArea*, *OverallQual*, *OverallCond*, *YearBuilt*, *YearRemodAdd*, *RoofMatl*, *TotalBsmtSF*, *CentralAir*, *GrLivArea*, *KitchenQual*, *Fireplaces*, *GarageQual*, *BsmtQual*, and *BsmtFinSF1*.
From there we decided to remove the variables *YearRemodAdd*, *RoofMatl*, *CentralAir*, *KitchenQual*, *BsmtFinSF1*, *GarageQual* and to add *Neighborhood*, and *GarageCars*.

Indeed, as we saw in our previous analysis, the numeric variables *TotalBsmtSF*, *GrLivArea*, *OverallQual*, *OverallCond*, *LotArea* are interesting when explaining Sale Price because they are highly correlated with our response variables. The factor variables *MSZoning*, *Neighborhood*, *GarageCars* and *Fireplaces* are also interesting when explaining Sale Price because of the large difference in each category boxplots. We removed the factor variable *Roofmatl* which is not very interesting because it is highly unbalanced and the variable *YearRemodAdd* because of its high correlation with *YearBuilt*.

Our reduced model is finally composed of: *MSZoning*, *LotArea*, *OverallQual*, *OverallCond*, *YearBuilt*, *TotalBsmtSF*, *GrLivArea*, *Neighborhood*, *GarageCars*, *Fireplaces*.

Finally, we decided to remove two observations. We observed that the observations 524 and 1299 have a low cook distance compare to others (however not exceeding 1), and their associated studentized residuals are also very low. Therefore, they are regression outliers, and since the p-value of the Bonferroni test are very low, we decided to remove them from the dataset to be sure that they don't influence our predictions.



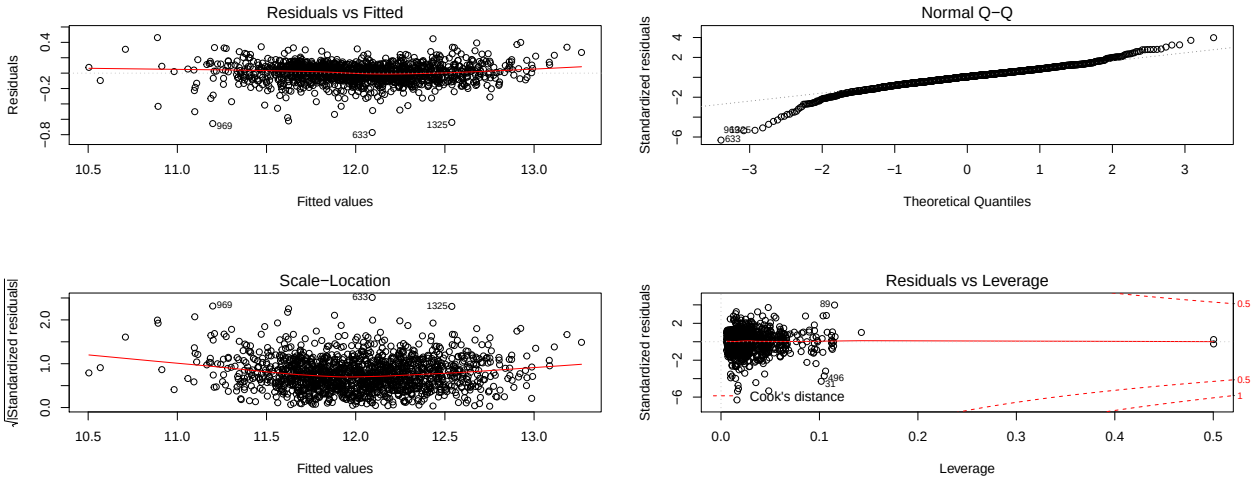
Let's train the model with our selected variables to see how the postulates are verified and if there are outliers, and observations with leverage one.

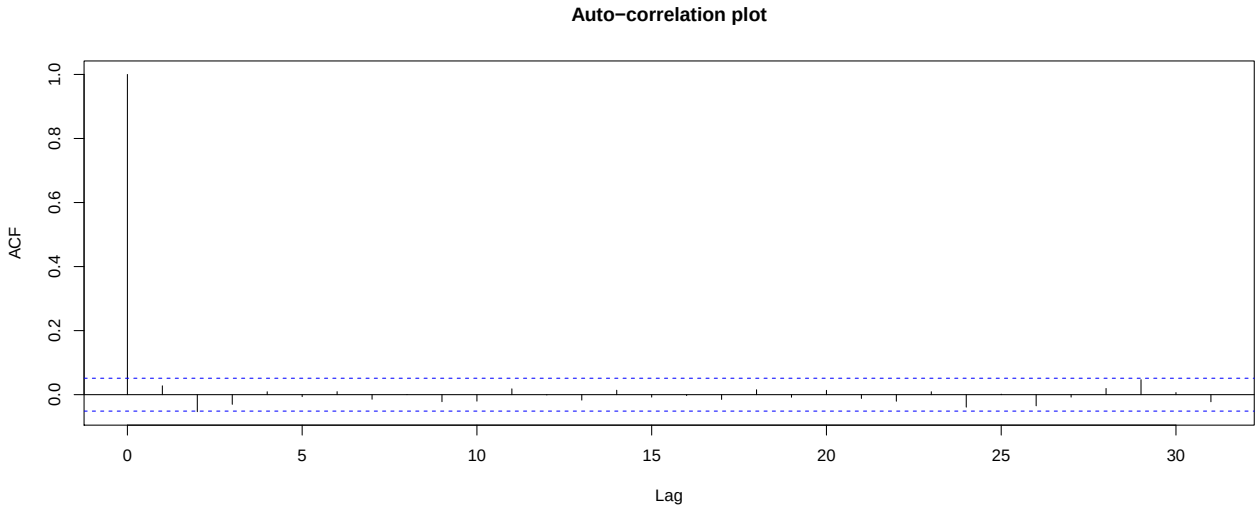
```
##
## Call:
## lm(formula = LogSalePrice ~ MSZoning + LotArea + OverallQual +
##     OverallCond + YearBuilt + TotalBsmtSF + GrLivArea + Neighborhood +
##     GarageCars + Fireplaces, data = trainPP[-c(524, 1299), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77192 -0.06266  0.00666  0.07063  0.46185
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            11.6882955  0.0578889 201.909  < 2e-16 ***
## MSZoningFV              0.4080177  0.0571618   7.138 1.51e-12 ***
## MSZoningRH              0.3314211  0.0575234   5.762 1.02e-08 ***
## MSZoningRL              0.3409601  0.0474916   7.179 1.13e-12 ***
## MSZoningRM              0.3017175  0.0446989   6.750 2.15e-11 ***
## LotArea                 0.0544929  0.0051554  10.570  < 2e-16 ***
## OverallQual             0.0839397  0.0060291  13.922  < 2e-16 ***
## OverallCond             0.0680080  0.0036834  18.463  < 2e-16 ***
## YearBuilt               0.1077733  0.0077241  13.953  < 2e-16 ***
## TotalBsmtSF             0.0632181  0.0044190  14.306  < 2e-16 ***
## GrLivArea               0.1273621  0.0049499  25.730  < 2e-16 ***
## NeighborhoodBlueste    -0.0012487  0.0941799  -0.013  0.98942
## NeighborhoodBrDale     -0.0232943  0.0476750  -0.489  0.62520
## NeighborhoodBrkSide     0.0468492  0.0407393   1.150  0.25035
## NeighborhoodClearCr     0.0354862  0.0423825   0.837  0.40257
## NeighborhoodCollgCr     0.0068256  0.0336632   0.203  0.83935
## NeighborhoodCrawfor     0.1196665  0.0388860   3.077  0.00213 **
## NeighborhoodEdwards    -0.0283327  0.0365608  -0.775  0.43850
## NeighborhoodGilbert    -0.0416566  0.0353728  -1.178  0.23914
## NeighborhoodIDOTRR     -0.0026253  0.0474119  -0.055  0.95585
## NeighborhoodMeadowV    -0.0028118  0.0474692  -0.059  0.95277
## NeighborhoodMitchel    -0.0465751  0.0374687  -1.243  0.21406
## NeighborhoodNAmes      -0.0231440  0.0345644  -0.670  0.50323
## NeighborhoodNoRidge     0.0899789  0.0377732   2.382  0.01735 *
```

```
## NeighborhoodNPkVill -0.0179524  0.0513892  -0.349  0.72688
## NeighborhoodNridgHt  0.1107001  0.0345941   3.200  0.00140 **
## NeighborhoodNWAmes  -0.0840361  0.0357776  -2.349  0.01897 *
## NeighborhoodOldTown -0.0187435  0.0421208  -0.445  0.65639
## NeighborhoodSawyer  -0.0355083  0.0367233  -0.967  0.33375
## NeighborhoodSawyerW -0.0219480  0.0360710  -0.608  0.54298
## NeighborhoodSomerst  0.0088254  0.0418505   0.211  0.83301
## NeighborhoodStoneBr  0.1731618  0.0398586   4.344 1.50e-05 ***
## NeighborhoodSWISU    0.0285194  0.0441927   0.645  0.51881
## NeighborhoodTimber   0.0007673  0.0385848   0.020  0.98414
## NeighborhoodVeenker  0.0540002  0.0497145   1.086  0.27757
## GarageCars           0.0451696  0.0047177   9.574  < 2e-16 ***
## Fireplaces           0.0252898  0.0041008   6.167 9.05e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1233 on 1421 degrees of freedom
## Multiple R-squared:  0.9072, Adjusted R-squared:  0.9048
## F-statistic: 385.8 on 36 and 1421 DF,  p-value: < 2.2e-16
```

```
## [1] -1927.16
```

We can observe that most of the variables still have a p-value for the Student tests that are smaller than 0.01, except for some of the categories of the variable *Neighborhood*. The F-statistic yields that the model is meaningful, but the AIC is larger than before, and the R^2, R_a^2 coefficients are smaller. Therefore, the model selection criterions don't go in favor of this reduced model, but we now have only 10 features out of 68 at the beginning.





Then, by looking at the model assumptions, we observe that the mean of the residuals is still zero, and the residuals are uncorrelated. The homoscedastic variance seems satisfied on the Scale Location plot, but the Breush-Pagan test still returns a p-value lower than the 5% treshold, that suggests us to reject the homoscedastic variance hypothesis. This is the same case concerning the gaussian assumption, but tt could have been better if we could have use the log transformation on some variables as we stated before.

4. Lasso Model

Finally, we do a Lasso regression which does automatic variable selection by putting some of the coefficients to zero. In fact, adding the L^1 penalization induces sparsity in the estimator.

The lasso regression gives us the variables that explain the most its predictions. We can observe that this method uses mainly the following regressors: *Condition2*, *GrLivArea*, *OverallQual*, *Neighborhood*, *Functional*, *CentralAir*, *YearBuilt*, *GarageCars*, *LotArea*, *SaleType*, *OverallCond*, *RoofMatl*, *Condition1*, *BsmtFinSF1*.
We used some of these features in our reduced model, but lasso is less selective than our model and uses a lot more regressors. Therefore, it will be interesting to see how the two different models predict, and to compare their respective RMSE (see file with full code for details).

5. Comparison of different models

The lasso regression model returns the log predictions, and obtains a RMSE of O.127 on the test dataset. By comparison, our reduced model has a RMSE of 0.124 on the same dataset. Therefore, we can state that our variable selection improve the accuracy of our log predictions. If we want to have the real predictions, we need to take the exponential of our predictions.

Conclusion

Our analysis now allows us to provide answers to our initial questions. We have found which variables explained the most the sale price and built a reduced model using a minimum number of variables. We used only 10 variables out of 68 at the beginning, and we reached a correct score, compared to more complex model. Our reduced model is as good as the lasso model when trying to predict Sale Price (similar RMSE), and it has the advantage to use a lot less variables which makes it faster and more easily explainable.
Our work can be improve by adapting the preproccesing to our study. In fact, we think that the way it was preprocessed influenced our work. We couldn't use the log transformations on some variables, and we think it could have helped us to verify the assumptions of a linear regression. Therefore, possible future directions of research can be how to preprocessed the data to have a linear model that verify all the assumptions needed for its validity. Also, it can be on the algorithm used to predict the sale price of the houses. Many different algorithms exists, and some models (for example gradient boosting algorithms) might be more adapted to reach the lowest score of prediction.