MAP 535 - Data Analysis Project

Adrien Toulouse & Paul-Antoine Girard

Introduction

Our task is to analyse a dataset, named **House Prices: Advanced Regression Techniques**. It contains our response variable, 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 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 10 most appropriate variables? How does it compare to larger models?**

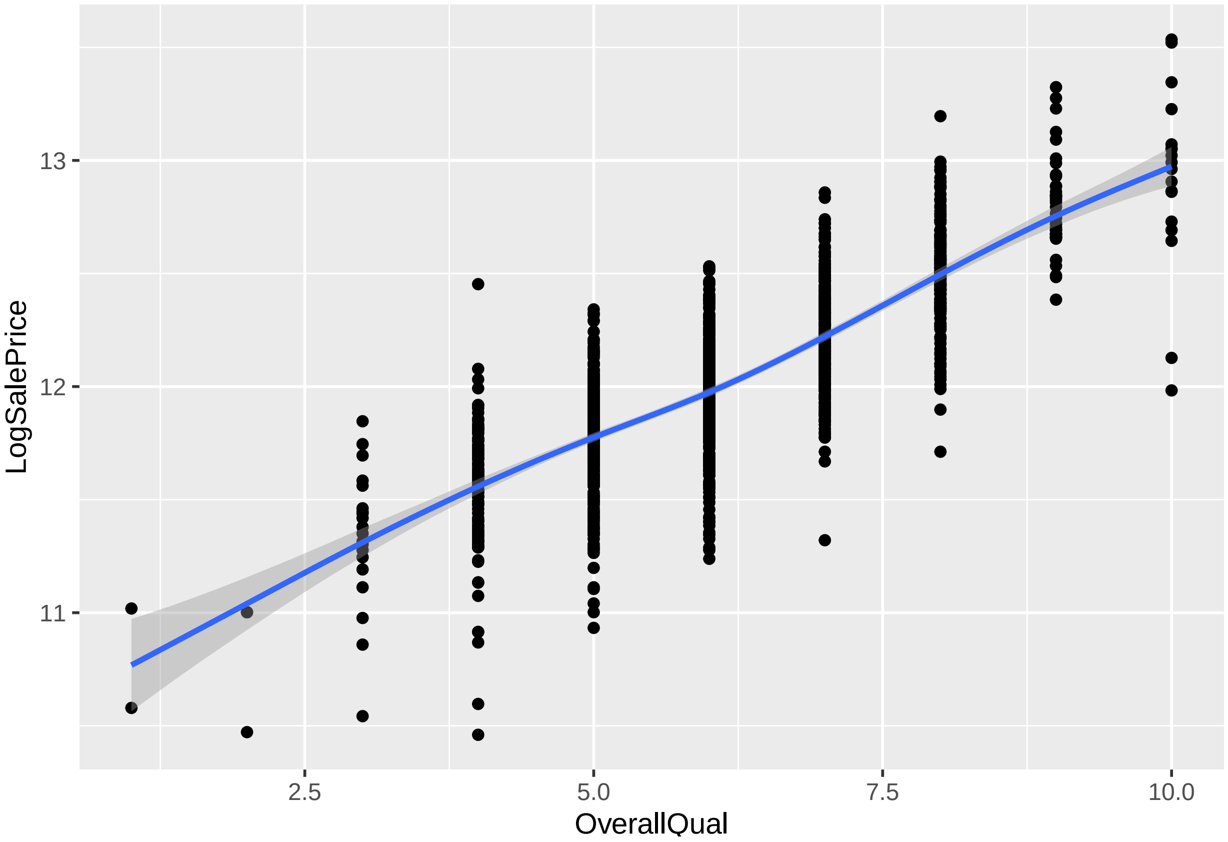
To test our hypothesis, we will start by describing the data and apply descriptive statistics to better apprehend it and preprocess the variables if necessary. We will then try to select the most important variables and after checking if the linear model assumptions are verified, we will build multiple linear regression models and compare them.

Exploratory Data Analysis

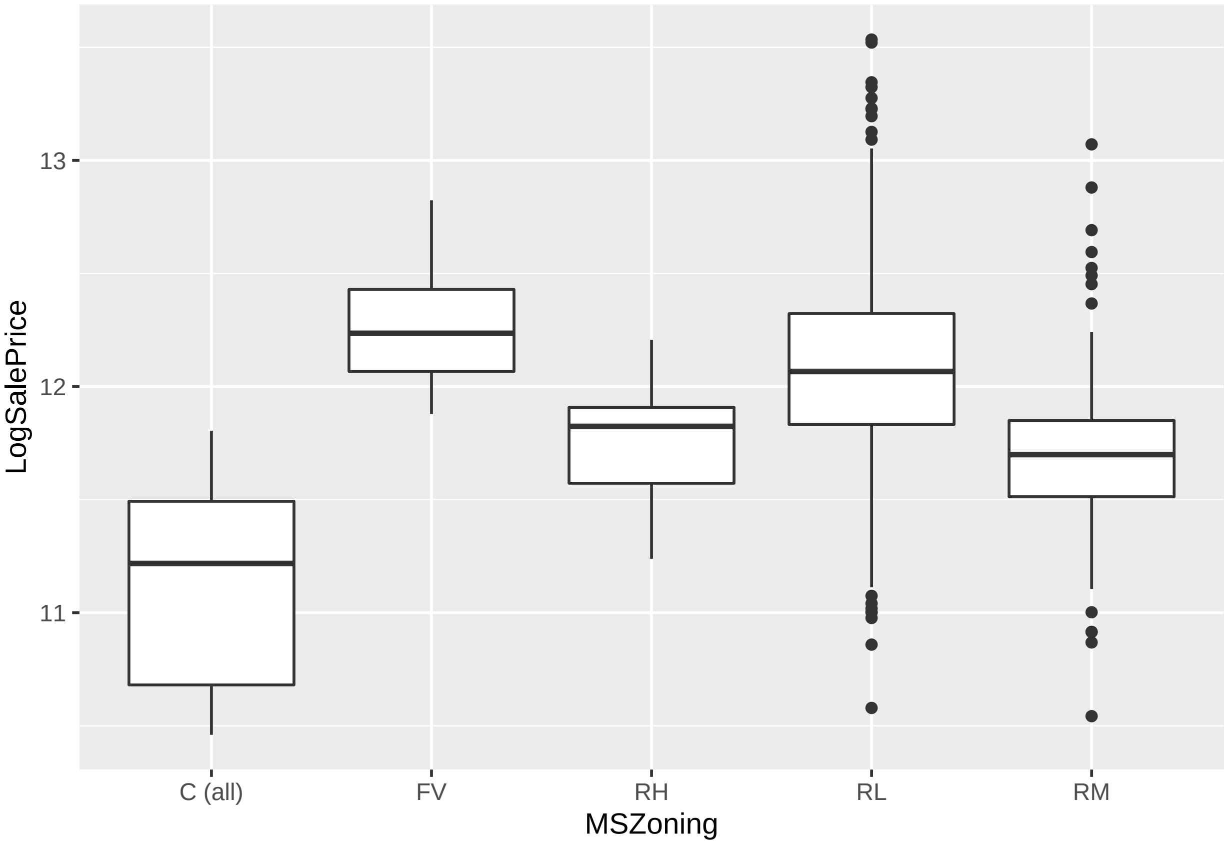
We start by looking at the data to see how it is structured. We find that the *SalePrice* has skewness in its distribution. So, a first transformation we do is to take the log of the house prices to reduce the effect of the tail in the density of our response variable.

Then, we plot scatter plots for quantitative variables and boxplots for categorical variables. These plots help us to determine if some of them are highly related to the *SalePrice*, and might be added to our linear model. We determine that *Overallqual* seems to be strongly correlated with the *SalePrice*. This regressor rates the overall material and finish of the house. It is a kind of summary of other variables, so it can help us to reduce the size of our model as it explains well the price. Concerning categorical variables, a variable may need to be considered in our model if it has different boxplots for each category when considering the *SalePrice*, as this will indicate a clear dependency between the two variables. We then remark that the variable *MSZoning* corresponds to this situation and strength to be important to deal with in our model. **By going this way, we also determine that the following variables might be important to add to ou reduced linear model… However, some of them has outliers (need to deal with) or might be interesting to take the log of them or transform qualitative into levels…**

```r
par(mfrow=c(1,2))
ggplot(trainImputed, aes(OverallQual, LogSalePrice)) + geom_point() + geom_smooth()
```



```r
(ggplot(data = trainImputed) +
  geom_boxplot(aes(y=LogSalePrice, x = MSZoning)))
```



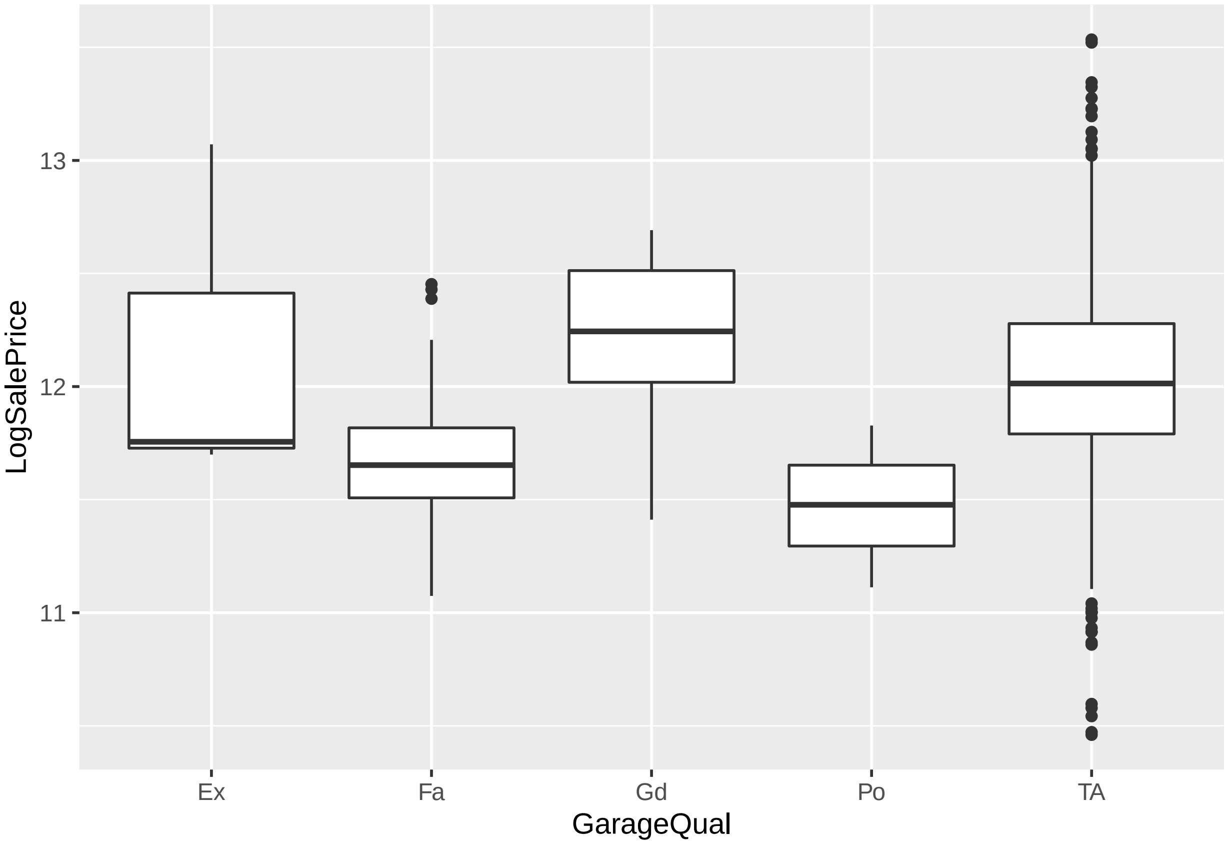
This first analysis visually indicates clearly that MSZoning is an important variable to explain the Sale Price.

```
res.aov <- aov(LogSalePrice ~ MSZoning, data = trainImputed)
summary(res.aov)
```

```
##                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 the 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 caregories when considering the Sale Price and make us say that we should include MSZoning in our model.

```
ggplot(data = trainImputed) +
  geom_boxplot(aes(y=LogSalePrice, x = GarageQual))
```



```r
res.aov <- aov(LogSalePrice ~ GarageQual, data = trainImputed)
summary(res.aov)
```

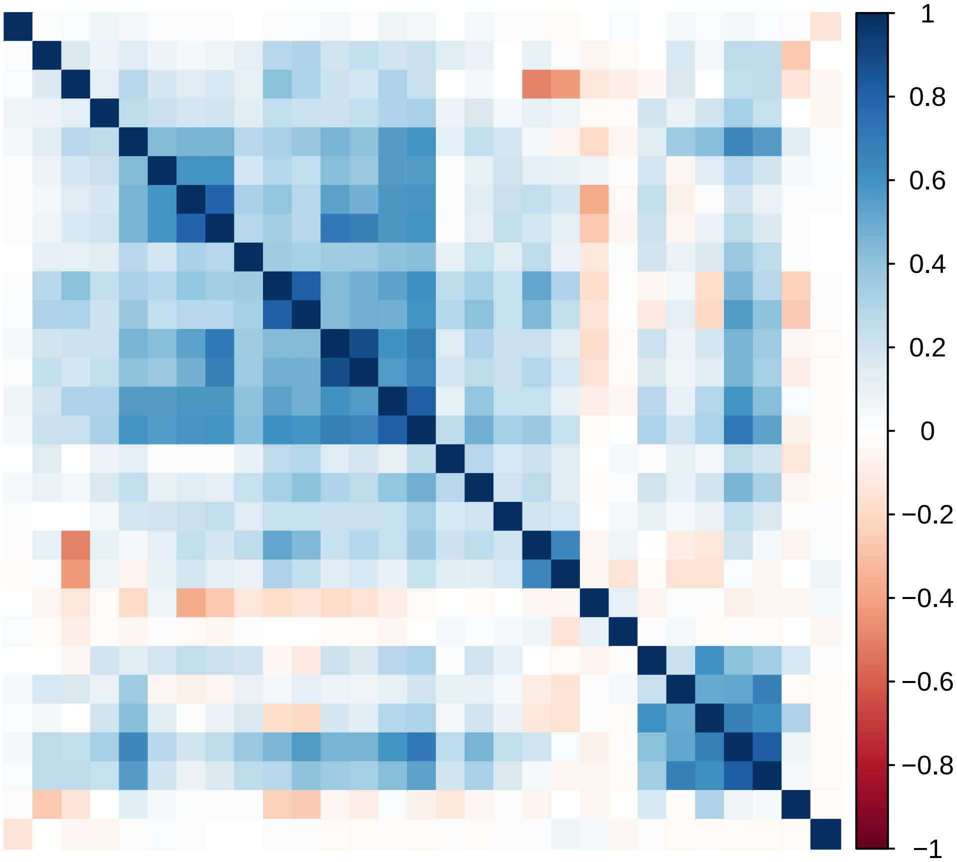
```
##               Df Sum Sq Mean Sq F value   Pr(>F)    
## GarageQual     4   7.54  1.8848   12.17 9.75e-10 ***
## Residuals   1455 225.26  0.1548                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

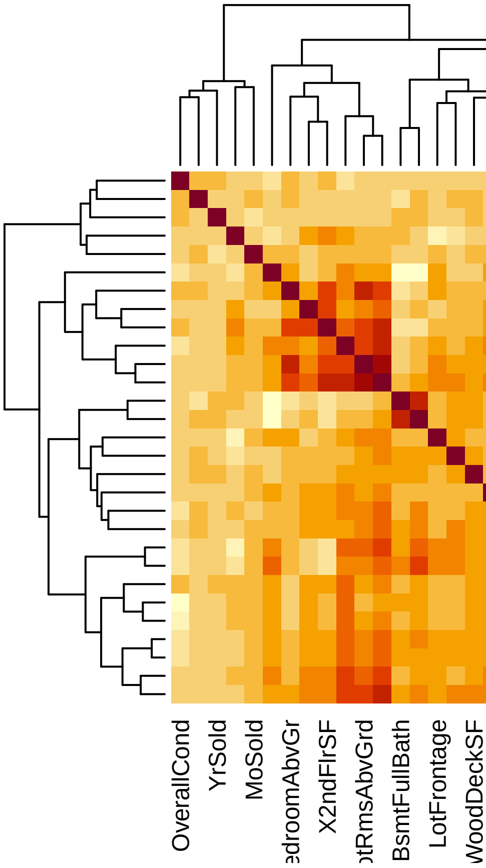
We find similar results when looking at the GarageQual variable. However, the boxplots do not completely follow intuition. Indeed the boxplots indicate that on average houses with garages in excellent quality have a lower sale price that garages in good quality. This seems to indicate that garage quality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

To select regressors, it is also important to analyze the correlation between numerical variables. There might be multicollinearity problems, but this will help us to reduce the number of our regressors by creating clusters of variables based on correlations. Indeed we know that having correlated response variables is not efficient in linear models and we also want to create a reduced model.

```r
var.numeric <- colnames(trainImputed)[sapply(trainImputed, is.numeric)]

trainImputed %>%
  select(var.numeric) %>%
  cor() %>%
  corrplot(method = 'color', order = "hclust", tl.pos = 'n') %>%
  heatmap (symm=T)
```





Now that we have more information on our data, let's try to build multiple linear models. We will use what we found in this exploratory analysis part to select variables and build these different models.

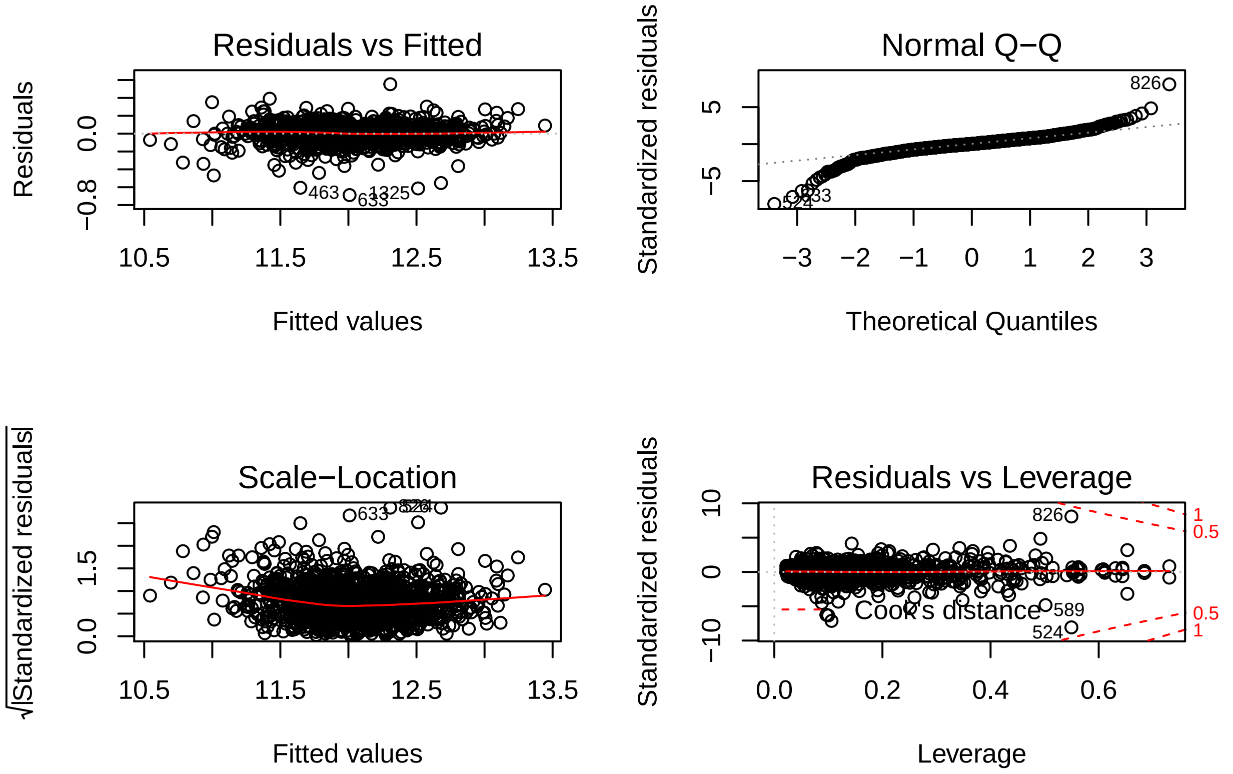
Modeling and Diagnostics

In this part, we are going to build different linear regression models and analyse their differences to select the one that we think fits the best our research hypothesis.

Full model

First, we start by doing a linear regression with all the variables of the dataset.

The results of this model show us that some of the variables are irrelevant if we consider the p-value of the Student test statistic realized for each variable. 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, GarageQual. Concerning the qualitative variables, we decided to only keep the one that have many categories that are relevant for the model, and not only one (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_a^2 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 relevant at a level of 0.1. **may need to look at other criterion**



We need to verify if the postulates verify the hypothesis necessary for the validity of our regression model. The residuals seem to have a mean of zero, and a variance near 1 (even if some residuals aren't equally spread along the line). Also, they are gaussian, except for some extreme values. In fact, before -2 and after 2, the errors don't seem to follow the line generated by the quantiles of a gaussian distribution. Finally, none of the residuals have a cook distance larger than 0.5, so we can assume that there isn't any leverage point or outlier.

Model using the forward method

We now use a method based on minimizing the AIC to automatically select a reduced number of variables for our model. At each step, the model adds the variable which reduces the most the criterion, and stops when adding a variable doesn't improve the accuracy of the model.

This model still selects a lot of variables, and still has a AIC value of -2314.3. So this suggests that the forward method deletes only the variables which are really irrelevant for the model, and the AIC criterion for selection model didn't change. Also, by looking at the p-value of the Student test statistic realized for each variable, we remark that if we take only the regressors that have a p-value inferior to 0.01, we select the same variables as before.
Let's try to reduce again the number of regressors based on the p-values obtained for the different regressors, and what we discovered in the previous exploratory data analysis..

Model reduce based on the Student tests

Model based on our analysis in the first part

Final model and prediction

Conclusion