MAP 535 - Data Analysis Project - Report

Adrien Toulouse & Paul-Antoine Girard

Introduction

Our task is to analyse the 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 aim within this project is to focus on dimensionality reduction by doing 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, understandable data. Indeed, with a large number of features, learning models tend to overfit which may cause performance degradation on unseen data. Moreover, data of high dimensionality can significantly increase the memory storage requirements and computational costs for data analytics.

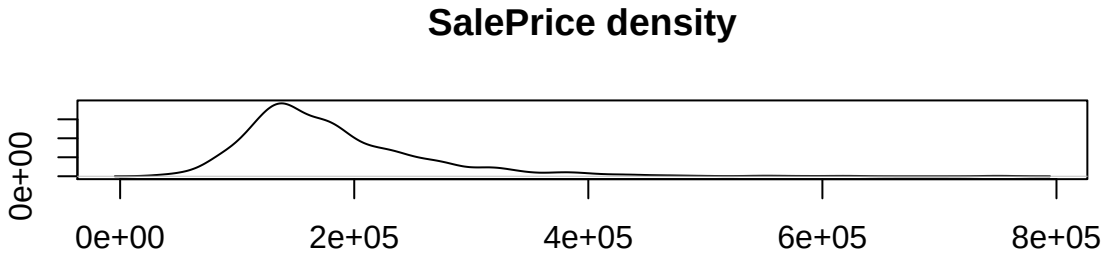
We can therefore adress the following question: **What are the most relevant features to explain the sale price of houses in our dataset?**

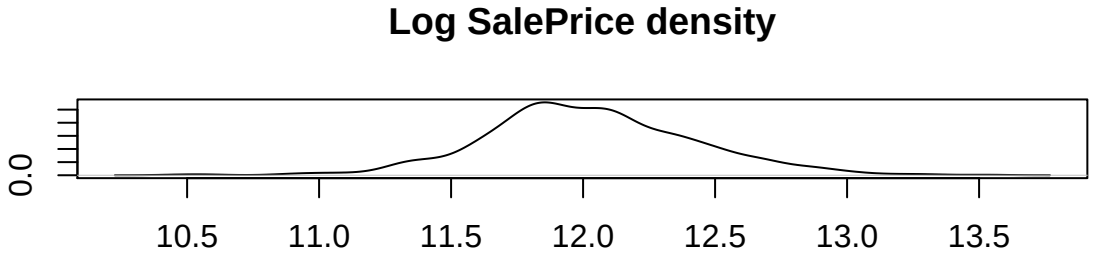
To answer our question we will first analyse the variables and assess their relevance by looking at the correlation with the regression target: *SalePrice*. We will also build and compare several linear regression models with different number of variables and finally conclude on the relevance of the 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?**

Exploratory Data Analysis

1. Transformations

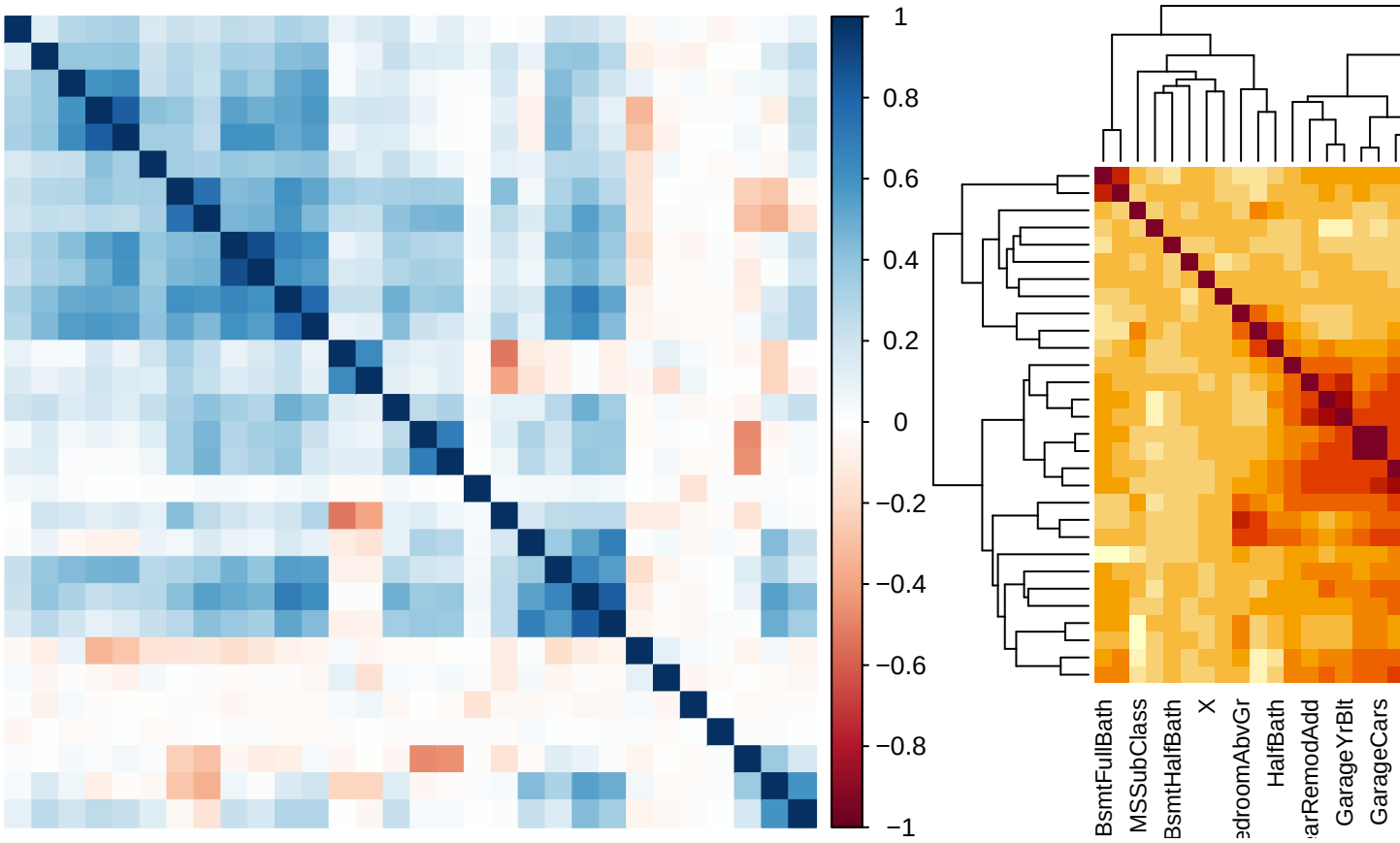
We start by looking at the data to see how it is structured.
We find that the *SalePrice* has skewness in its distribution. So, a first transformation we do is to take the log of the house prices to reduce the effect of the tail in the density of our response variable. We do the same transformation for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



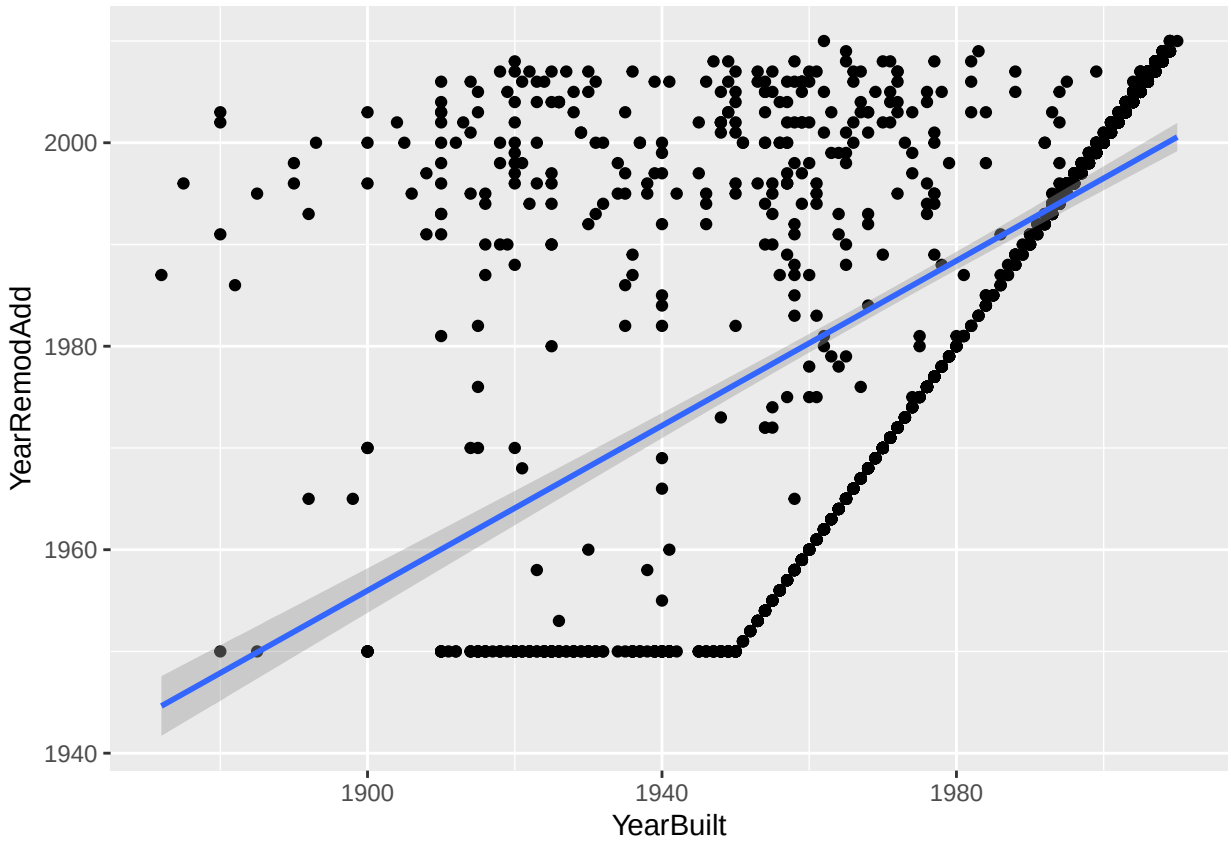


2. Numeric variables

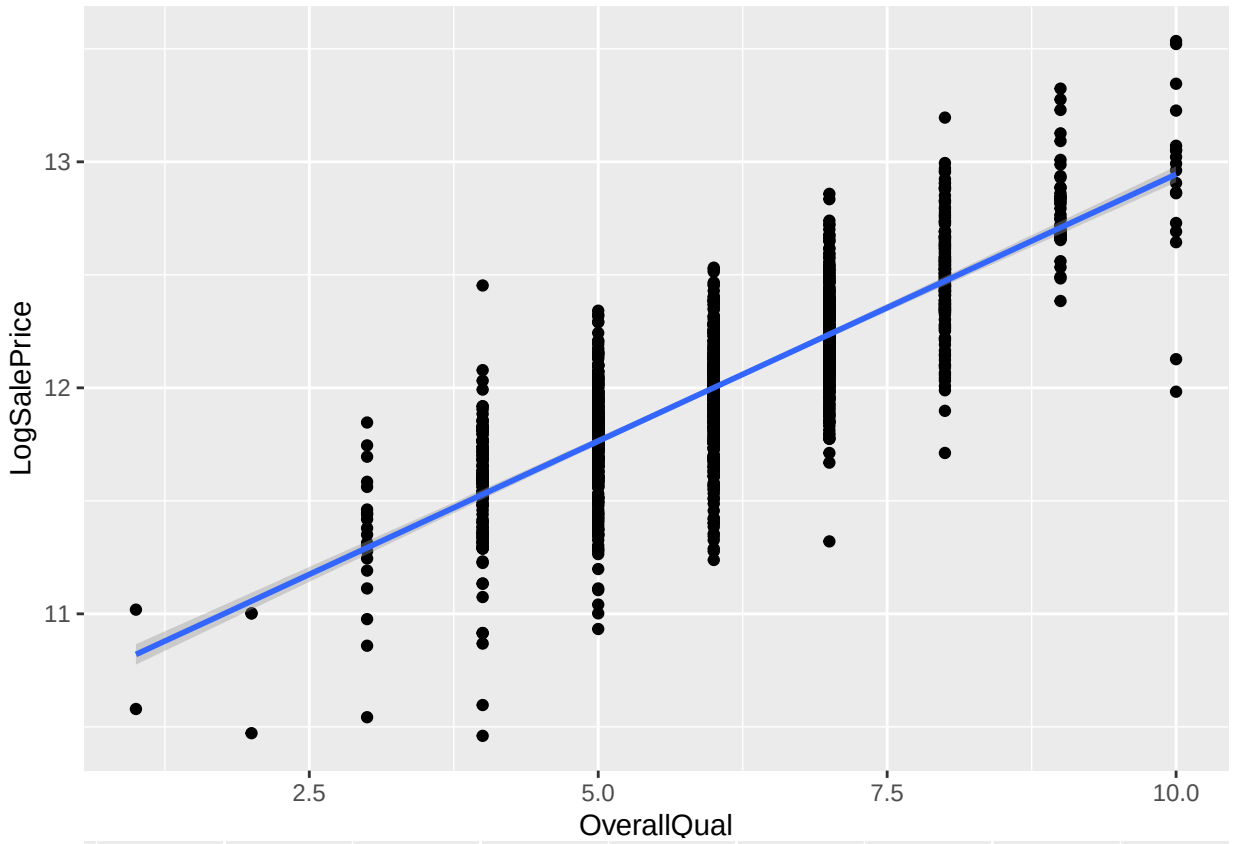
Looking at the numeric variables, we check correlation between the different variables together and then the correlation of SalePrice with the variables. The first step is not useful when you try to predict a variable. It is however very important when you try to see which variables are the most important to explain price.
To select regressors, it is important to analyze the correlation between numerical variables, since there might be multicollinearity problems, and we can create 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.

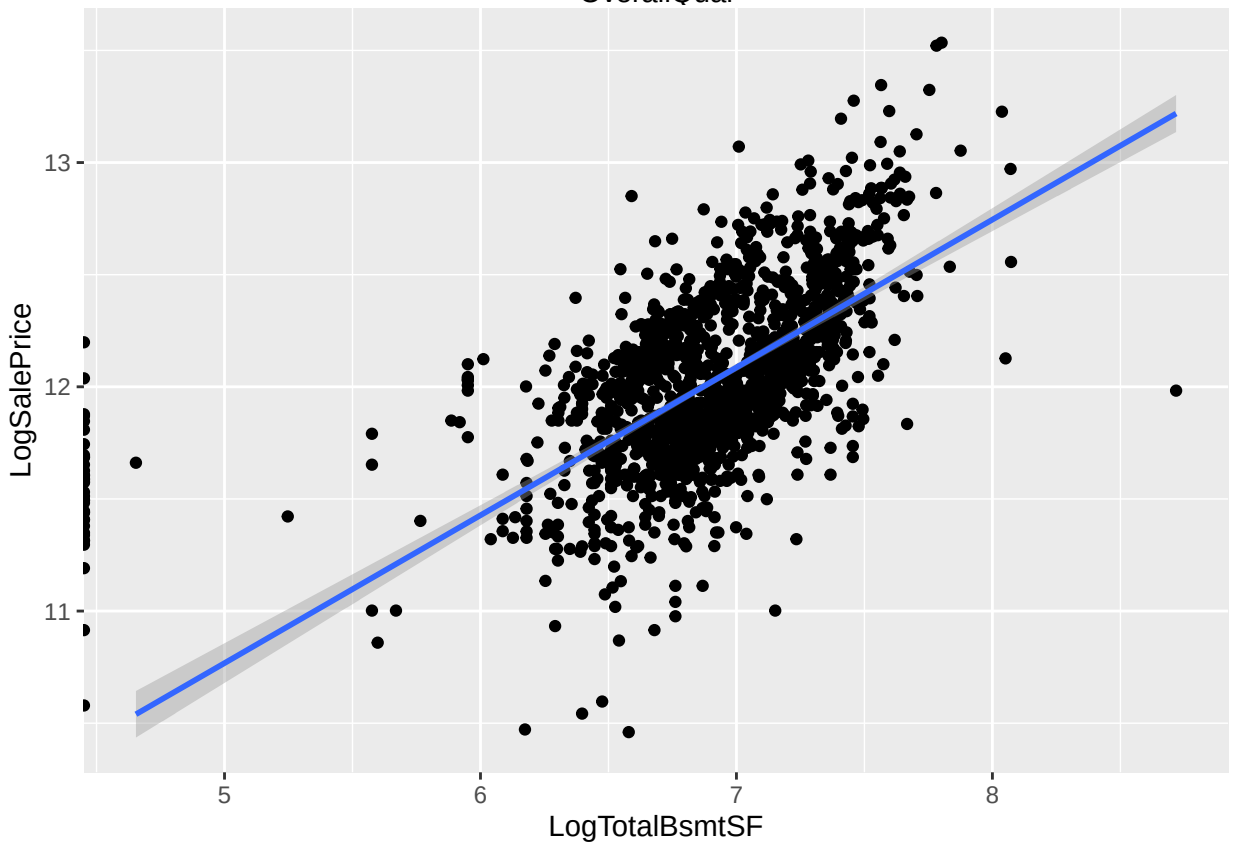


Some variables like for example YearBuilt and YearRemodAdd are very strongly correlated and we will not need to keep both variables.



Let's now look at the correlations with our response variables to see which variables explain the house sale price well. For example, Overall quality and LogTotalBsmtSF impact Sale Price as the scatter plots show what looks like a linear relationship. By plotting other scatter plots, we found that the variables YearBuilt, LogLotArea, OverallCond are also strongly correlated with the price. Including these predictors in our model should be a good idea.



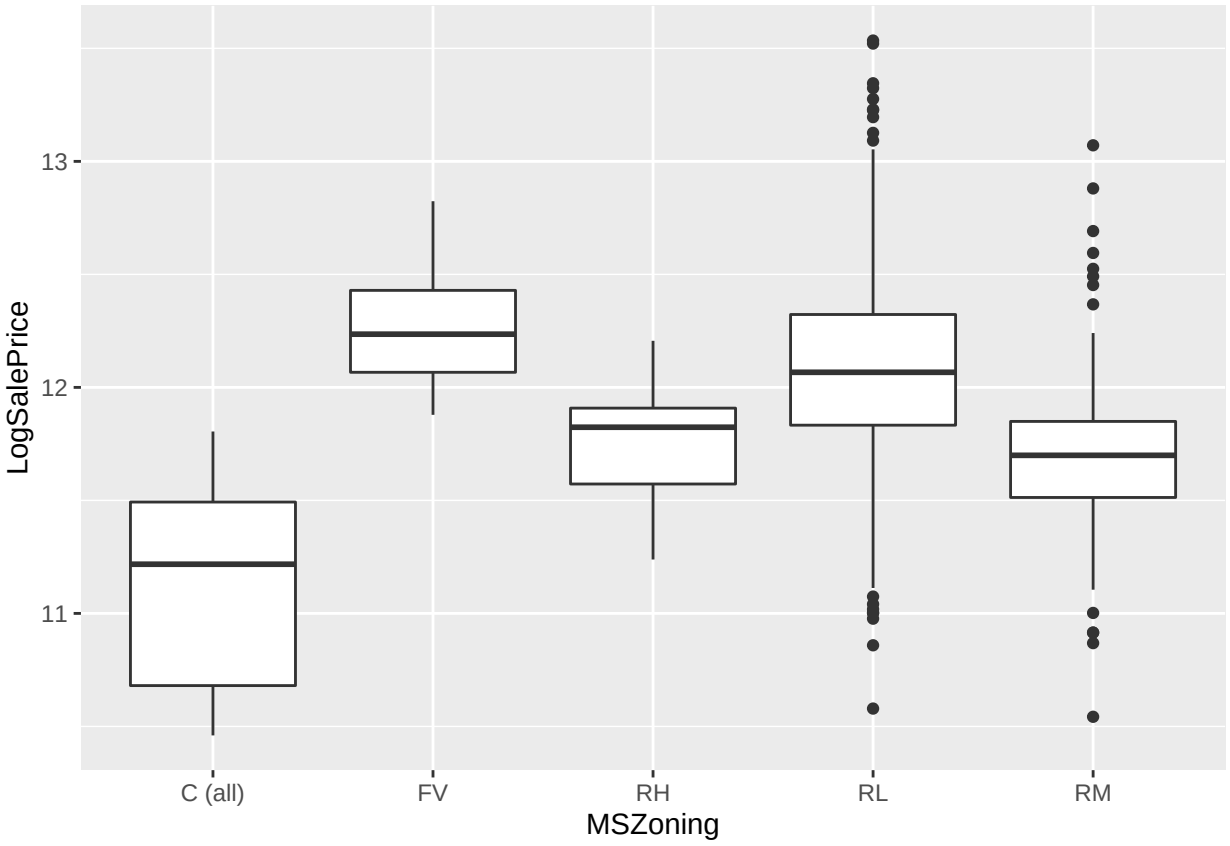


The regressor *Overallqual* is particularly interesting as it rates the overall material and finish of the house. It

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.

3. Factor variables

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 remark that the variable *MSZoning* corresponds to this situation. Indeed, these boxplots are quite different visually from one another which indicates clearly 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
```

The one way Anova test confirmes what we thought, as the p-value is less than the significance level of 0.05. We can conclude that there are significant differences between the MSZoning categories when considering the Sale Price and make us say that we should include MSZoning in our model.

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.

By plotting other boxplots for the different variables, we found that the variables CentralAir, RoofMatl and Lot Config could also be good to explain the Sale Price.