# Regression - Final project

*Victoire de Termont and Sarah Jallot*

*12/3/2019*

## I. Introduction

Our objective is to predict the prices of residential homes in Ames, Iowa accurately by exploiting an existing dataset to its best. Our key metric will be the RMSE on test data.
The training data consists of a set of 66 categorical and quantitative variables extensively describing 1,095 residential homes in Ames, Iowa, and their corresponding price in dollars. We want to preprocess and explore our data smartly, select the best performing features, and fit the most adapted regression model after validating the key assumptions needed to implement it.

We will first explore our data to get intuition of an efficient model. Understanding and preprocessing the data implies that we understand what we are given, check that there are no missing values, and that R correctly categorises each regressor.

The data we are handling is heterogeneous, although it is presented in a structured manner. We are dealing both with categorical and quantitative variables, which we will have to preprocess separately.

Displaying the summary of our dataframe, there are 67 columns including the price, meaning a full model would include 66 features and an intercept, and 1,095 observations.
The list of features is extensive: it is probable that not all regressors are useful to predict house price.

The quantitative variables differ in scale and range: prices start from ~35,000 dollars, and can attain 755,000, whereas bedrooms above ground range from 1 to 8 only for instance.
This means that scaling our data could optimise model performance.

We first treat missing values before launching into the analysis.

```
## There are 0 missing values in our dataframe.
```

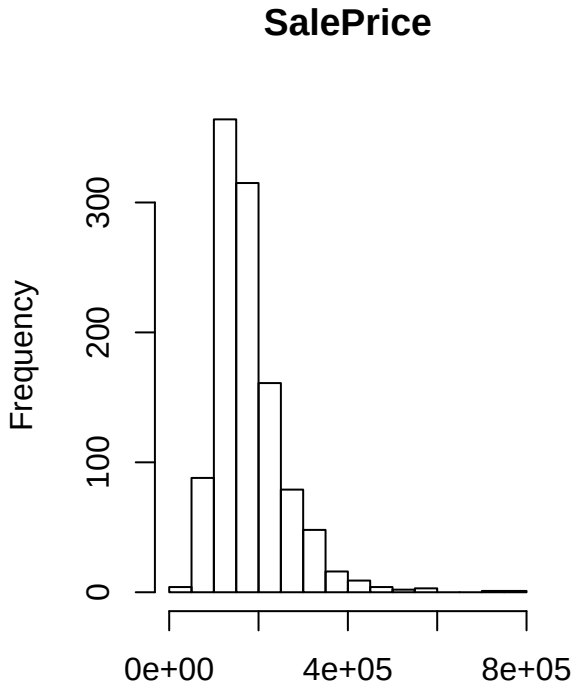
First intuition: looking at this dataset, we can see that we have a lot of regressors and that some might be more relevant than others to explain the SalePrice. For instance, the variable MSZoning is expected to have much more impact on the price that the variable Heating, as the heating system is something that can be changed, whereas the location is permanent.
Thus, we will perform different types of analysis to keep only the most relevant variables in our model.

Additionally, R appears to classify some features as integers when they could be considered as factors: mainly factors related to areas and quality ratings. After debating whether we would consider Years as categories, we decided against given that we might come across a previously unencountered year in the test data which would nullify the interest of using year as a factor.

All quantitative features are integers and price, the output, is the only numeric.
Factor variables are automatically numerically encoded by R, which assigns them to ascending levels using their alphabetical order. Note that this might bias the data by assigning a higher value to one level versus another without any ground to do so.

## II. Exploratory data analysis / initial modeling

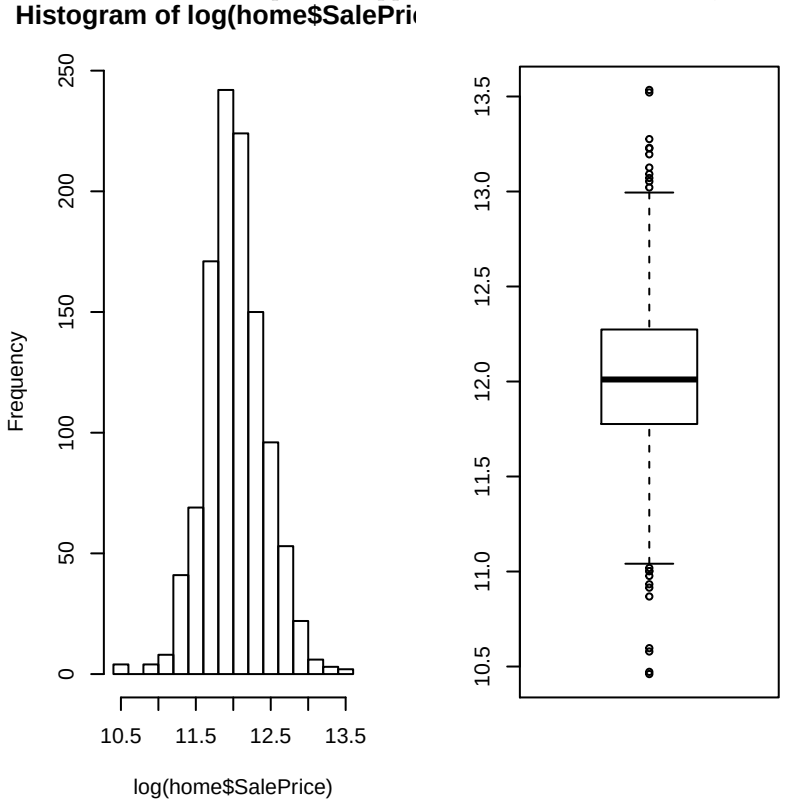
1. Numerical analysis
   We will extract trends and relationships among the regressors. Comparing them is also essential, for instance using pairplots.




We first set out to observe the output, SalePrice.

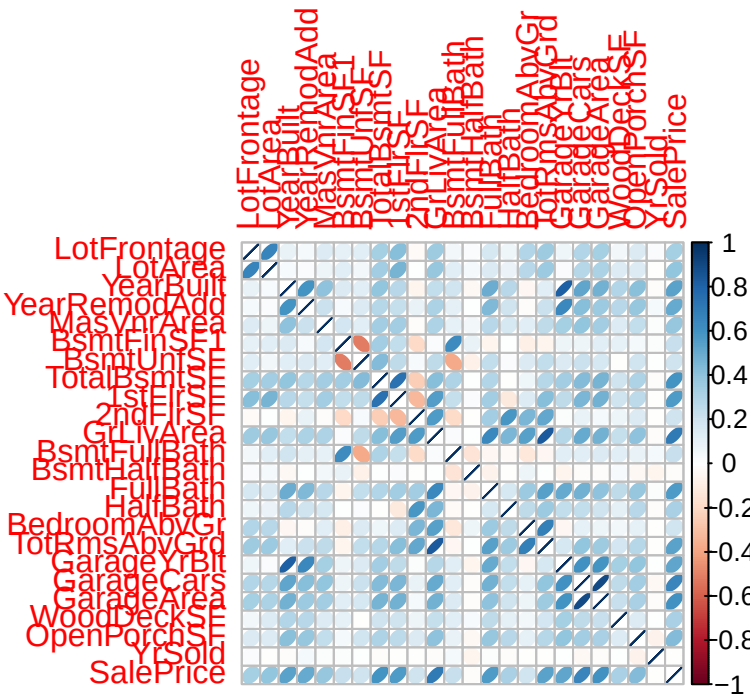
```
## The mean of Saleprice is 181196.1 whereas the median is higher at 164500 indicating that extreme valu
```

SalePrice is volatile with many houses to the left hand side, but with an important number of outliers to the right with extreme values which skew the data. To smoothen the output and approach a normal distribution, we

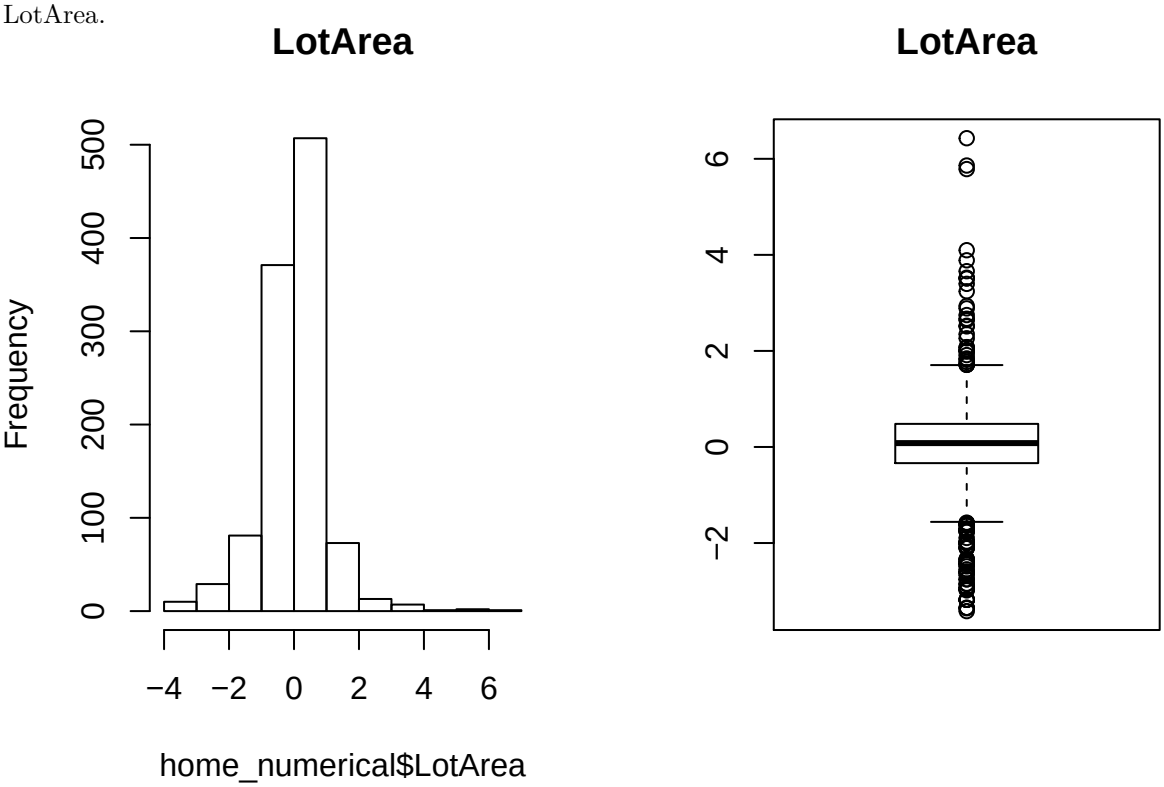



will consider the log when fitting our model.
log(SalePrice) is pretty close to a normal distribution, except for extreme values.

```
## corrplot 0.84 loaded
```




Our intuition is that three feature types mainly drive SalePrices: location, area/surface and quality. Let's start by analysing the numerical variable LotArea.




We believe that areas, particularly LotArea, will be important when predicting SalePrice. Most areas are located within [-2,2] in our pre-processed dataset, with many outliers towards the left- or the right-hand-side.

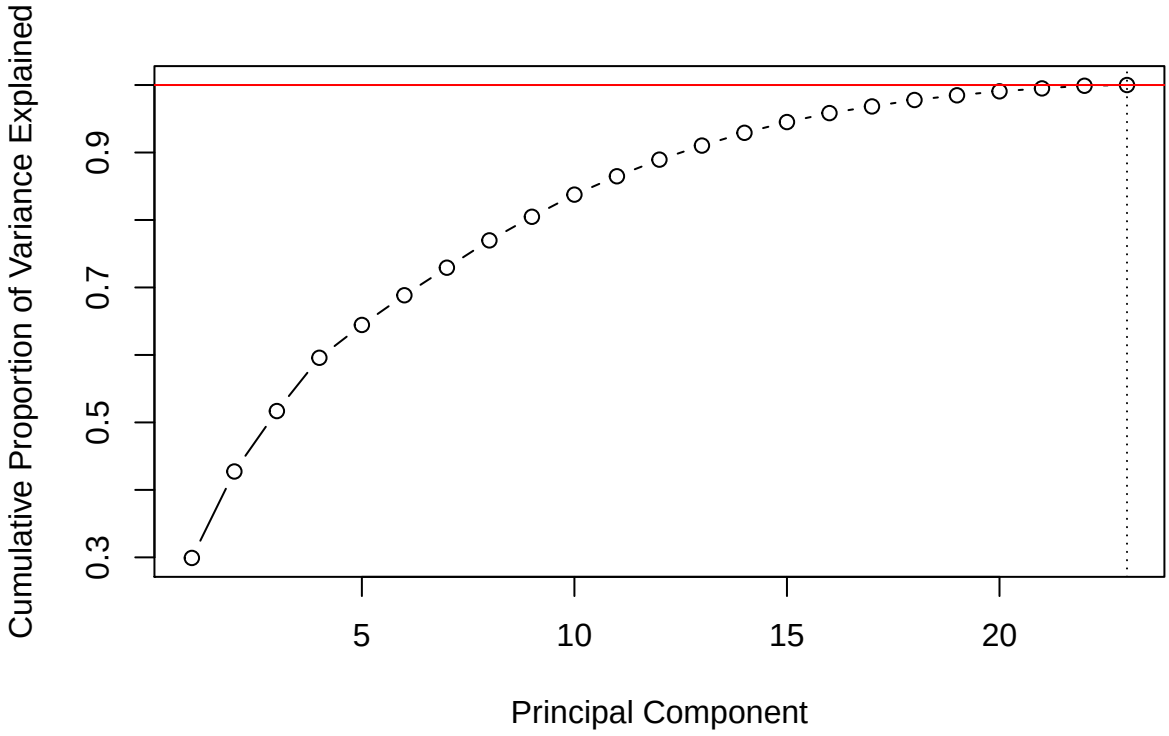
Looking at inter-feature correlations first, we observe that OverallQual is correlated to may other features on the correlation matrix. This indicates that there is a link between it and other variables. This is encouraging

as it means that we would be able to remove some columns without losing too much information.

The numerical features that are the most correlated with SalePrice are OverallQual, TotalBsmtSE, X1stElrSF, GrLivArea, GarageCars and GarageArea. On the contrary, MoSold, YrSold and BsmtHalfBath are poorly correlated to SalePrice. As the correlation matrix does not take into account interactions between our features in predicting sales price, we will not infer from this which variables we will keep in our final model.

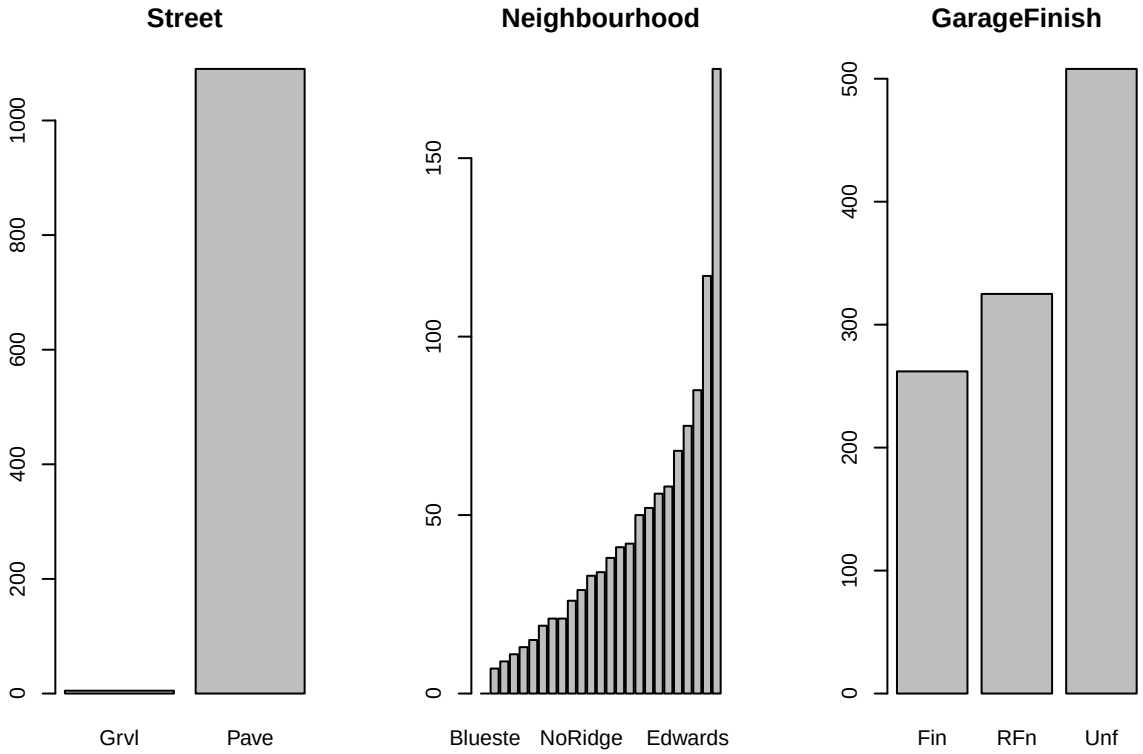
## PCA

Given that we have 67 columns, of which 29 are quantitative, let's explore the possibilities we have to reduce our dataframe.
Because PCA works best with (scaled) numerical data, we will perform it on our scaled numerical feature columns.




23 of our 28 numerical features account for 99% of the variance.
For 20 features, we still get an acceptable explained variance of ~96% > 95%.

3. Factor analysis
   For our first contact with the factors, let us get a closer look at level repartition to get an idea of fragmentation within the factors.



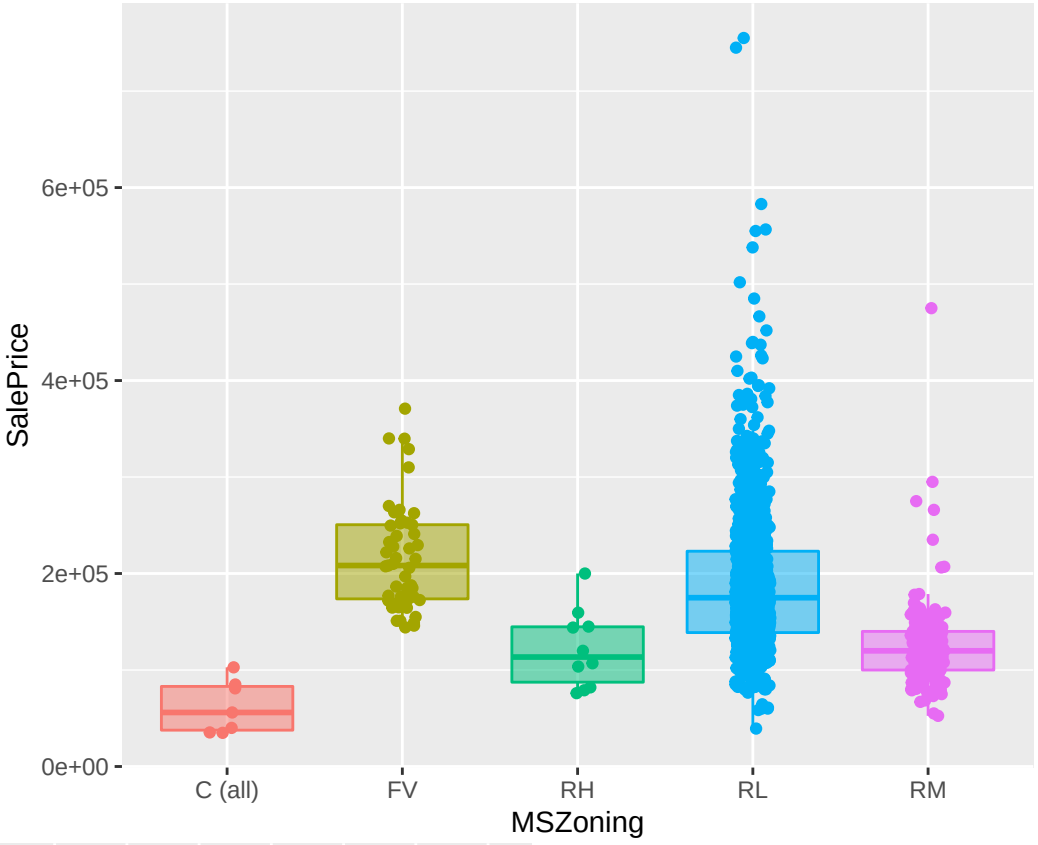

Observing the factor columns, we see three types of repartitions.

i) Clear surrepresentation of one level versus the others. In Street and Utilities, which are binary, this is conspicuous. From this we infer that these factors won't be very useful in predicting house price in general: nearly all houses will be in the same category along these factors (and for those who aren't, data is too sparse to generalise well).
Note that this is also the case for RoofMatl, Heating, BsmtFinType2, Electrical, GarageCond, GarageQual.

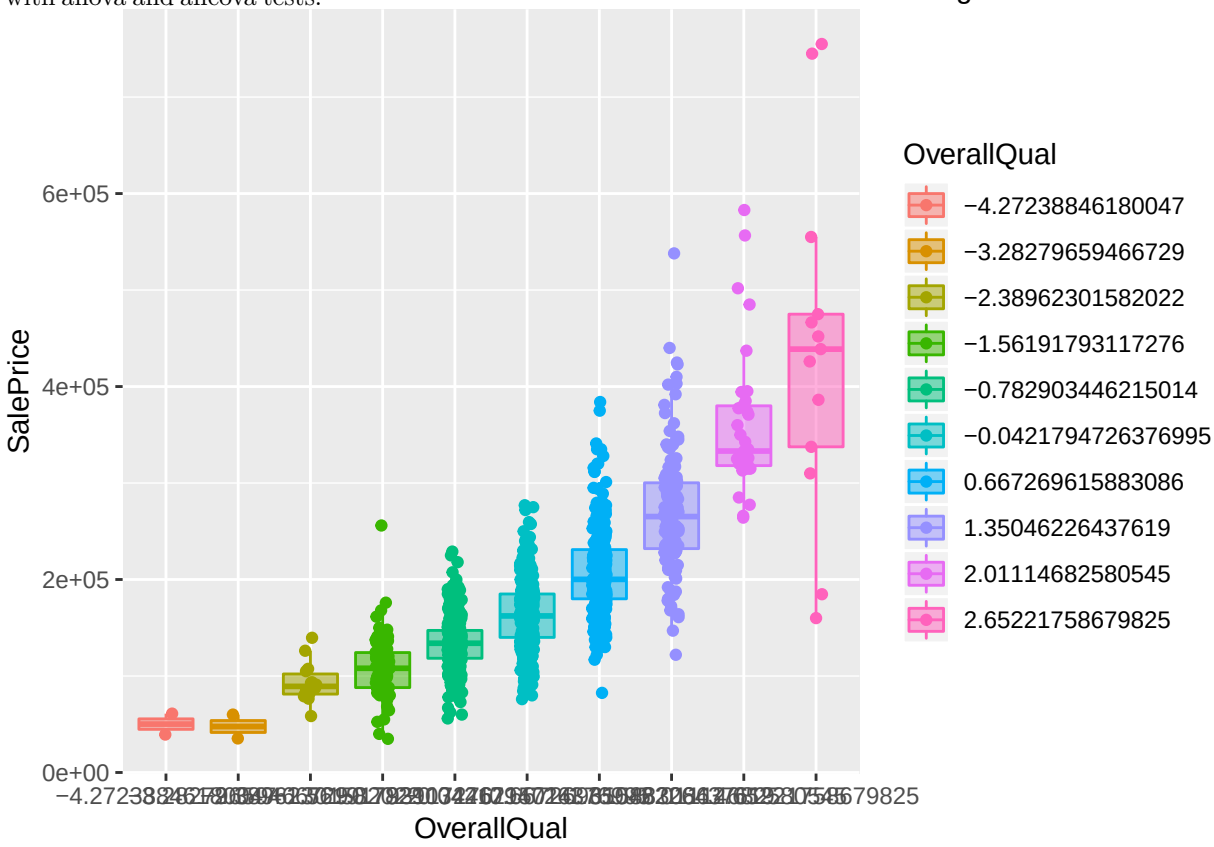
ii) High cardinality in the number of levels: this is especially the case for neighbourhood, which has 25 levels. We will regroup some of these levels together to improve our predictions. Note that this is also the case for Exterior1st, Exterior2nd, BsmtExposure.

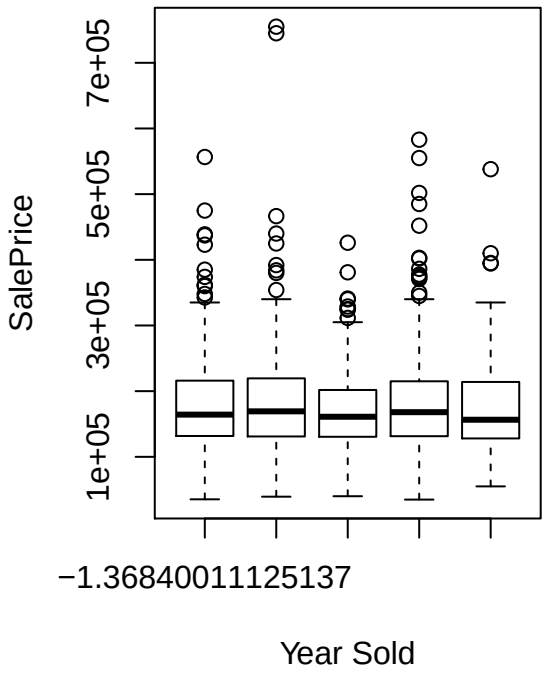
iii) Classic factor repartition with reasonable representation of each modality, as is the case for Housestyle for instance. Note that this is also the case for HeatingQC, GarageFinish, BsmtFinType1.

Intuitively, we said that both location and overall quality will impact SalePrice significantly. Let us check this




with anova and ancova tests.






It appears that houses from the RM and RH areas are less expensive than the ones from FV and RL areas.

```
## Analysis of Variance Table
##
## Response: SalePrice
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## MSZoning     5 3.6722e+13 7.3445e+12  1330.1 < 2.2e-16 ***
## Residuals 1090 6.0185e+12 5.5216e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## OverallQual   10 4.0471e+13 4.0471e+12  1934.1 < 2.2e-16 ***
## Residuals   1085 2.2703e+12 2.0925e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova tests seem to show that both area and overall quality have a strong effect on SalePrice. However, here we have not accounted for the interaction between MSzoning and overall quality: to validate our conclusions, we must show that it is not significant with an ANCOVA test.

Since the last p-value is big, we will consider that SalePrice will depend in a similar manner on OverallQual and MSZoning.

### III. Modeling and diagnostics

Given the volatility of Saleprice, we will work on its logarithm to smooth its distribution and improve estimated residuals homoscedasticity. After manually removing poorly informative factors and other highly correlated numerical features, we start by fitting a full linear model to predict log(SalePrice) and check whether the postulates are validated.

XXX and then add a Lasso penalization as we have a lot of regressors. Using those two models,the variable selection methods will enable us to select the appropriate regressors.

Then, we will compare our models using criteria seen in class. We will also test the gaussian assumption of our model.

Looking at our dataframe, we realised that some factor columns had a lot of differents modalities, with some that appeared very few times. Thus, for more efficiency of our algorithm, we decided to group them. We performed this for Neighbourhoood, RoofStyle, Condition1 and OverallQual.

Then, using previous analyses and deeper data exploration, we found at that some columns did not help to predict our model, as they could add redundancies or errors. Thus, we decided to remove them:
- We removed some features that took almost a single value for all oberservations: Street, Utilities, RoofMatl, Condition2, Heating, Electrical, Functional, GarageCond.
- For some features, we were not sure wether or not they had an influence, so we tested the model with and without them and removed them if they were not improving our score : HeatingQC, SaleCondition, BsmtFinType2, RoofStyle, ExterQual.
- We also removed some features based on Anova tests: Exterior1st, Exterior2nd, GarageFinish.
- We discovered that OverallCond was extremely correlated to many other features and repeated OverallQual, so after an anova test we decided to remove it.
- Finally, we found out that OpenPorshSF was lowering our ajusted R2 and was not correlated to the SalePrice (cf. correlation matrix), so running the model without this feature was better.

Running a model with all the variables excluding the ones we just removed, we obtain an adjusted R2 of 0.91.

According to the T test above, the variables having the most impact to explain SalePrice are the ones describing: - the location of the home (e.g. MSZoning, Neighborhood) - the area of the home (LotArea) - overall quality and condition ( OverallQual, OverallCond) - # Comment/reword the specificities of some aspects of the home such as the roof (e.g. RoofStyleShed, RoofMat) and security functions (e.g. Fireplaces). - the construction period (e.g. YearBuilt, YearRemodAdd)

Many variables could be removed from our model while marginally affecting its efficiency to explain SalePrice. To improve model efficiency, we will perform a selection of variables based on forward, backward and both methods.

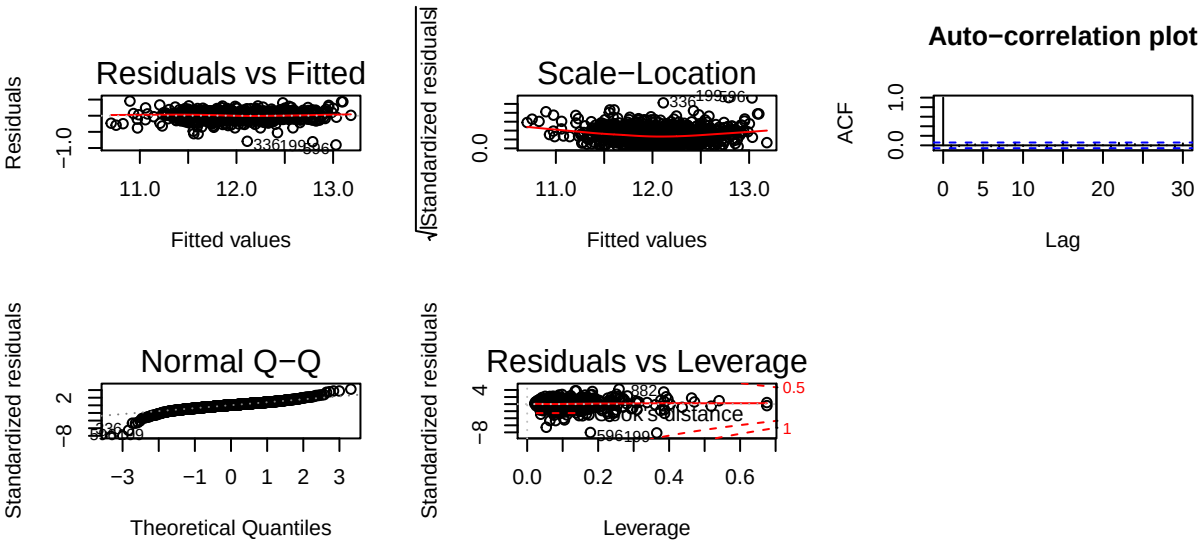
To choose among the three methods, we retrieve the AIC of each model and choose the one with the smallest AIC.

```
## [1]   145.000 -4424.838
```

```
## [1]   106.000 -4469.426
```

```
## [1]   106.000 -4469.426
```

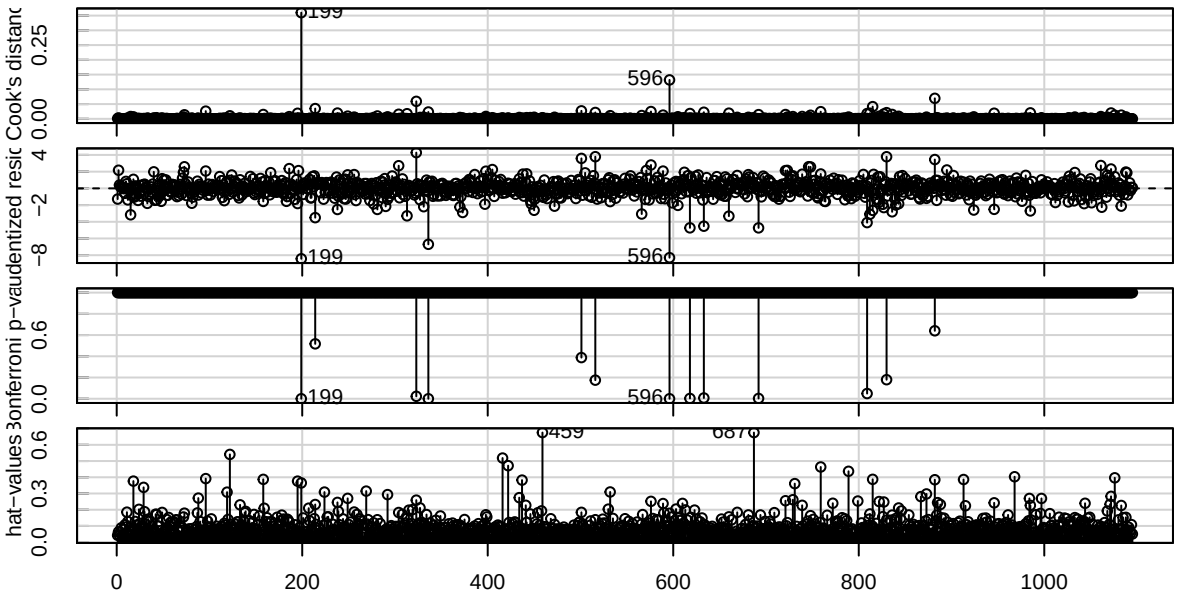
We choose the model extracted by the backward or the both method as they have the same AIC, smaller than the one of the forward method. We choose arbitrarily the backward one. For this specific model, let's verify that the postulates are verified.




Looking at the graphs, the postulates are verified for this model.

Now, we want to verify that we don't have outliers in our model.

```
## Loading required package: carData
```



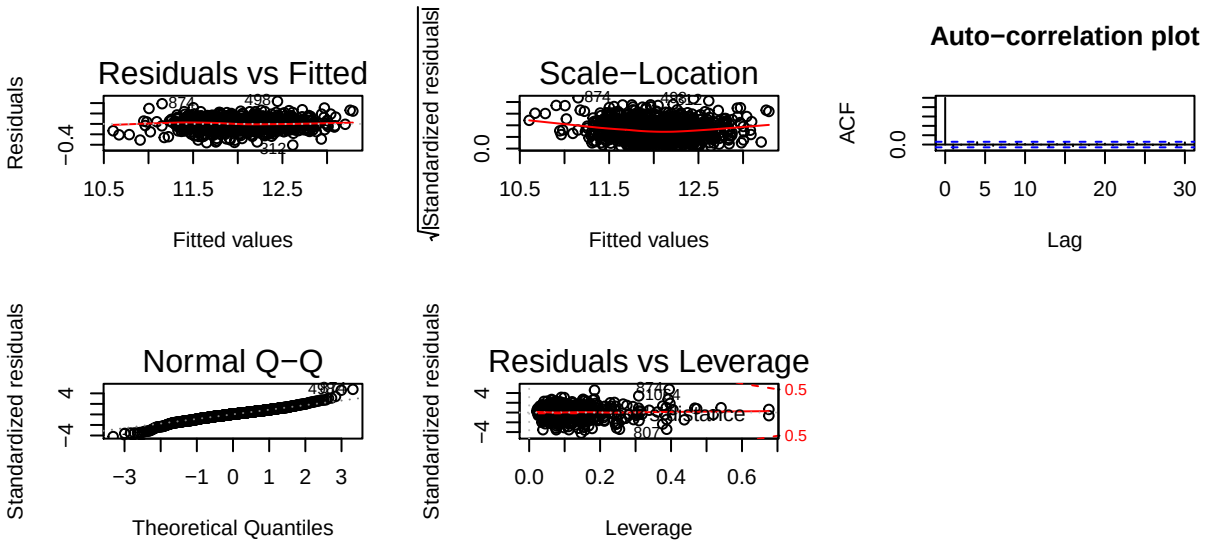

Cook's distance plot: according to the Cook's criteria, we don't observe any leverage point or regression outlier.
Studentized residuals plot: according to the plot, there are a lot of outliers (many points below -2), which is confusing. The two main outliers according to this criteria oare 524 and 1299. Bonferroni's plot : we notice that 13 points have a p-value inferior to 0.5, so they are outliers according to this criteria.
Hat plot: two points (327 and 783) seem to be leverage points according to this criteria.

Based on those results, we decided to run an outlier test for more precision:

```
##      rstudent unadjusted p-value Bonferroni p
## 199 -8.429606         1.2154e-16   1.3309e-13
```

```
## 596 -8.299318         3.4108e-16   3.7348e-13
## 336 -6.699113         3.5154e-11   3.8493e-08
## 692 -4.731168         2.5573e-06   2.8002e-03
## 618 -4.726779         2.6119e-06   2.8601e-03
## 633 -4.528832         6.6545e-06   7.2867e-03
## 323  4.274605         2.1002e-05   2.2997e-02
## 809 -4.099323         4.4839e-05   4.9098e-02
```

Using this test, 7 possible outliers are given according to the studentized criterion and Bonferroni p-value. We can now test our model without them. Note that those tests are based on confidence intervals and then for this reason, given the number of observations, we could have up to 70 outliers.




According to the graphs, P1 and P3 are validated without hesitation. For P2, we notice an slight elliptic behaviour in the middle, but it seems slight enough to validate the postulate. For P4, we have tail observations that do not fit the normal distribution. Based on the high number of observations in our dataframe and the few number of points not aligned with the normal distribution quantiles, we validate the postulate.

**IV. Final models**

Parameters of our model:

Finally, our final model without outliers is the one we built reaffecting some features, regrouping some categories for factors and retreving repretitive or non significant variables. Then, we used a backward selection method to reduce the number of features and keep only the necessary ones. Our model had Our final model had 7 outliers that we decided to remove after outlier tests.

It can be written: lm(formula = log(SalePrice) ~ MSSubClass + MSZoning + LotFrontage + LotArea + LandContour + LotConfig + LandSlope + Neighborhood + Condition1 + OverallQual + YearBuilt + YearRemodAdd + MasVnrType + MasVnrArea + ExterCond + Foundation + BsmtCond + BsmtExposure + BsmtFinSF1 + BsmtUnfSF + CentralAir + GrLivArea + BsmtFullBath + FullBath + HalfBath + KitchenQual + Fireplaces + GarageType + GarageCars + GarageQual + WoodDeckSF + MoSold + YrSold, data = home_no_outliers)

Looking at the coefficients, the most impactful features are MSZoning, OverallQual and Fireplaces by far, which is not suprising. Indeed, we stated in the beginning that the location was key. Also, the quality is very important because if conditions are not satisfied, then the flat would require some work, which cost would be deducted from the final price. Finally, having a fireplace is a good indicator of the standard of a home, as they tend to be built in expensive flats.

Testing our regression model on the test set, we fail to predict around 13% of the sale price, which corresponds

to a RMSE of ~24k, for homes with an average price avec ~180k. This seems acceptable for the price of a home.
We obtain a p-value that is below 2.2e-16, which is much less than 0.05. We reject the null hypothesis that the intercept only is better to explain our SalePrice than the model we have built.

Confidence interval:

**V. Discussion**

Final conclusions on our model: In the end, we improved a lot the model that we have compared to the one we had when running a linear regression on the original dataframe, as our postulates are now better validated with twice as less features.

This project enabled us to apply all the methods we saw in class to analyze a given dataframe and find an acceptable manner to transform it into a more efficient dataframe. Working in pair was very positive, because it allowed us to get more ideas, and to share our understanding of the functions. We could be more efficient next time now that we have gained insights into the R manner to solve a regression problem.
To go further, we could have used a lasso regression because we still have around 30 features. We tried to do so but had struggles plotting our postulates afterwards, and therefore could not conclude. However, we will learn how to do it as it could have been very useful.