Regression - Final project

Victoire de Termont and Sarah Jallot

12/3/2019

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 MSE on test data.
The training data consists of 66 categorical and quantitative variables extensively describing 1,095 residential homes in Ames, Iowa, and the houses' corresponding price in dollars. Discriminating between relevant and non relevant regressors to foster sparsity in our model is paramount for an efficient and robust model.

To achieve this, we first investigate numerical and factor variables sequentially. Applying the log to SalePrice likens it to a Gaussian distribution, with a few outliers. A PCA analysis on numerical variables showed us that 23/28 numerical variables accounted for 99% of the sample variance. Factorwise, we distinguish three types of regressors: factors with an overrepresented level, some with high cardinality, and others with standard level repartition.

We then perform elementary factor pruning and intra-factor modality regrouping before using ANOVA to remove factors manually. We do not touch numerical variables but instead rely on our model to make a selection.

We choose to fit two linear models using stepwise selection procedure with the AIC criterion to foster sparsity. In the first model, the only modification we impose on the output is applying the logarithm, before removing outliers after testing them. In the second model, we apply a winsorisation method to log(SalePrice) as we noticed that most outliers located to the LHS of price made our model less robust. In both cases, our best model according to AIC is the backward selection model. Model 1 predicts the test data much more accurately than model 2 (24 000$ vs 33 000$), but model 2 validates the regression assumptions better. XXXX what do we predict better?? We choose to retain model 1 as we want our model to generalise well, and model 1 is satisfactory enough on the postulates. Our final model has an MSE of XX.

I. Data exploration We first explore the pre-processed 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 consider separately.

We observe in the data summary that a full model would include 66 features and an intercept, and 1,095 observations.
The list of features is extensive: not all regressors will be useful to predict house price. For instance, we expect the variable MSZoning 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.
Quantitative variables differ in scale and range: prices start from ~35,000 dollars, and can attain 755,000. Before pre-processing, surfaces took higher values than bedrooms above ground which ranged from 1 to 8. Scaling the data allows to harmonise it, so we keep the scaling.

R appears to cast some factors as integers: mainly areas and ratings. We decide not to consider years as factors as we want our predictions to generalise to other, unencountered years. We recast OverallQual, OverallCond, MoSold, and MSSubClass as factors.

All quantitative features are integers and price, the output, is the only numeric. We know that R automatically encodes factors, and we choose to keep the by default levels, which are alphabetically ordered.

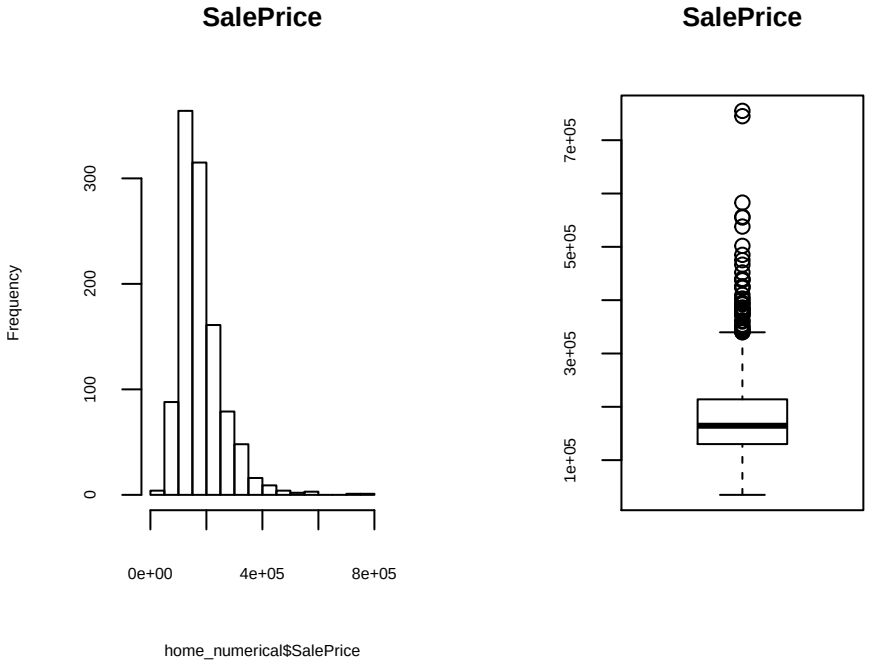
We first treat missing values before launching into the analysis.

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

II. Exploratory data analysis / initial modeling

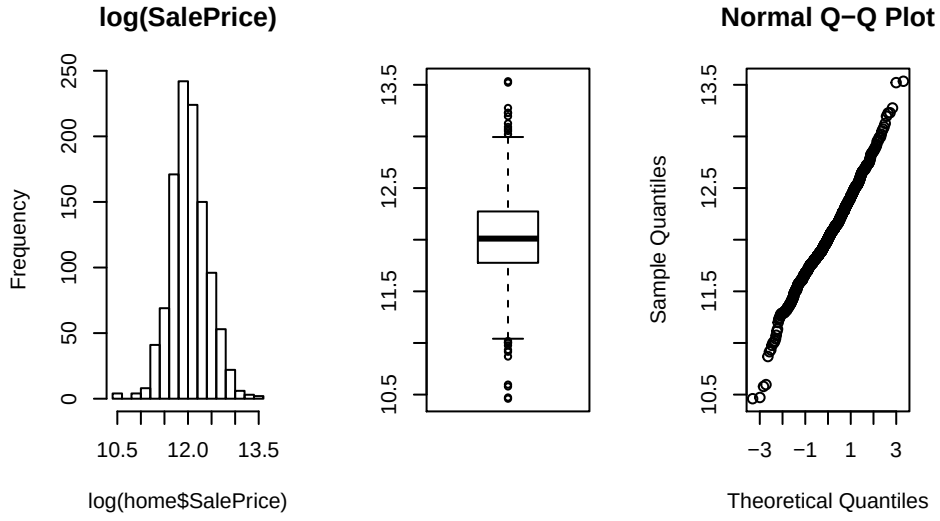
1. Numerical analysis.
   We describe the output data before analysing the correlation matrix. A PCA analysis we performed showed us that 23/28 numerical variables accounted for 99% of the variance. We left it out for concision.

We first set out to observe the output, SalePrice.

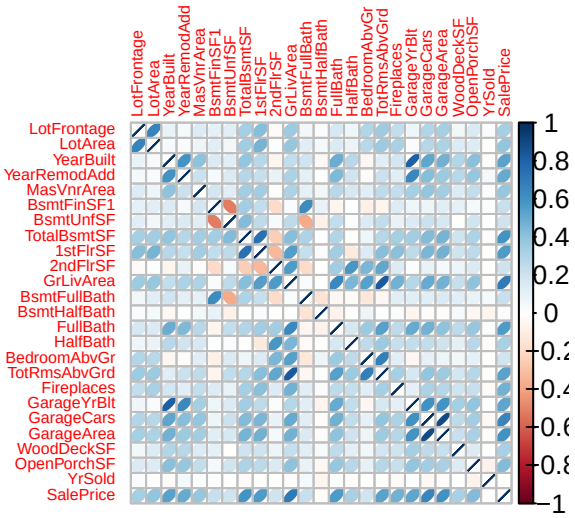



```
## Data skewness is 1.91628
```

SalePrice is highly skewed to the right : we confirm this by noting that Saleprice mean is ~181,196 whereas the median is much lower at ~164,500. SalePrice is volatile with many houses to the left hand side, but with a number of outliers to the right with extreme values. To smoothen the output and approach a normal distribution, we will consider the log when fitting our model. If this isn't enough, we could go a step further by either trimming or modifying outlier values to improve our generalisation error.

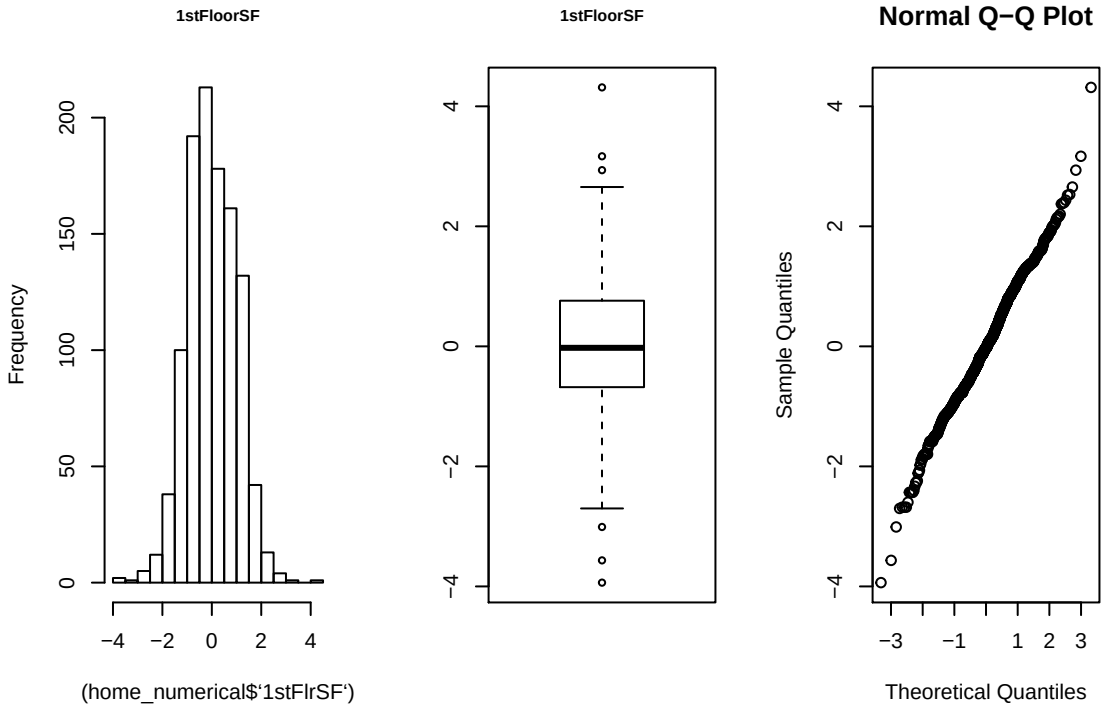


log(SalePrice) is pretty close to a normal distribution, except for extreme values.



The numerical features that are the most correlated with SalePrice: GrLivArea, GarageArea and GarageCars, 1st & 2ndFlrSF, YearBuilt GarageYrBuilt & YearRemodAdd. Areas and surfaces are all related to square feet, which we know is a key driver in house sales. The construction or modernisation works are an indicator of overall quality of the housing and the investments that went into it, so it makes sense for them to be correlated to SalePrice. On the contrary, YrSold and BsmtHalfBath are poorly correlated to SalePrice. YrSold is correlated to none of the other features, so it is an irrelevant regressor: the decision to sell a house doesn't have much to do with what drives its price or the price one can sell it at. The correlation matrix does not take into account feature interactions, so we will leave numerical feature selection to our stepwise model procedures.

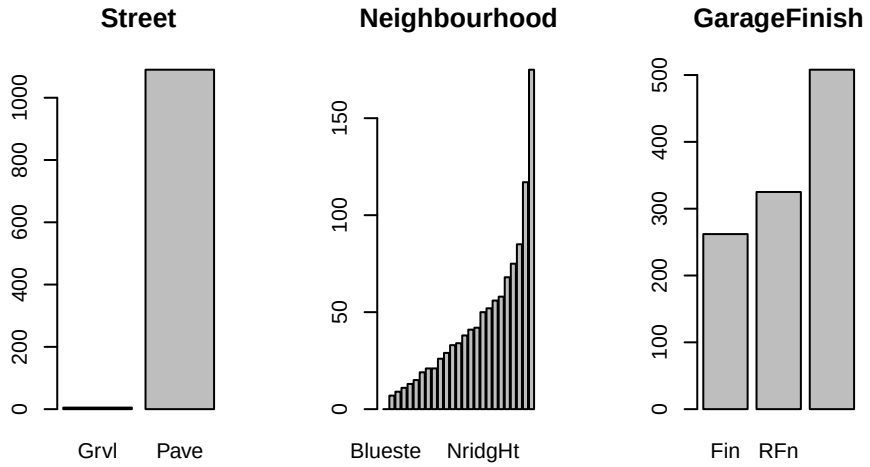
Our intuition is that three feature types mainly drive SalePrices: area/surface, location and quality. Let's describe 1stFlrSF as it is the closest feature to square feet with 2ndFloorSF.



From these three graphs, we can assume a Gaussian distribution on 1stFloorSF. ## 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.

3. Factor analysis
   We first investigate level fragmentation within the factors. We discover three types of factors, of which examples are given below.

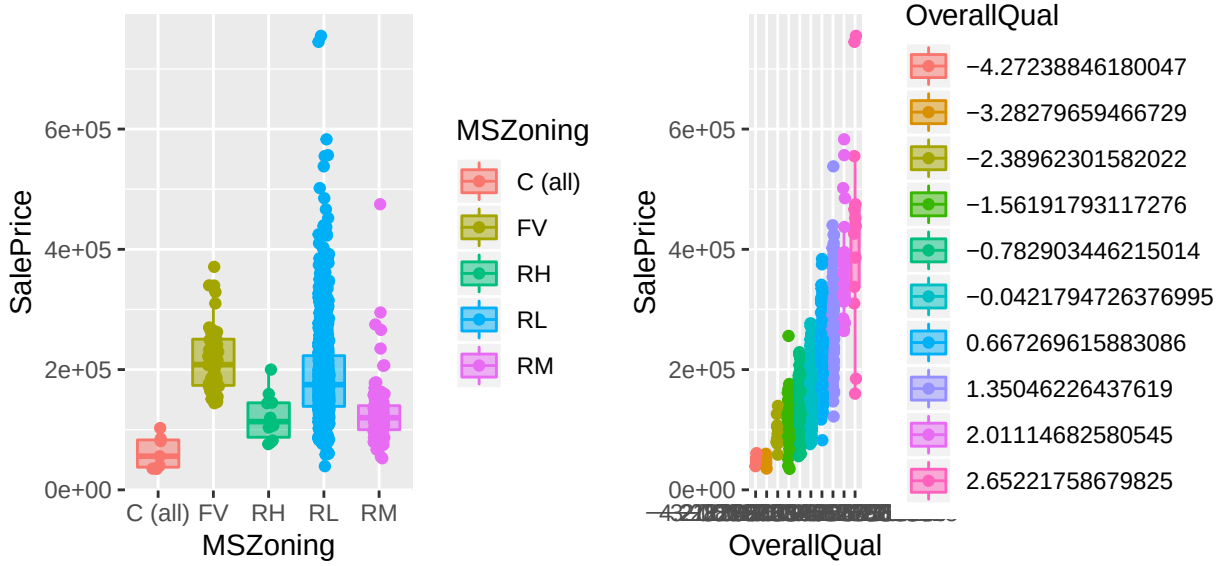


   i) Clear surrepresentation of one level versus the others. In Street and Utilities, which are binary, this is conspicuous. From this we infer that these factors won't be very useful in predicting house price in general: nearly all houses will be in the same category along these factors (and for those 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 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
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## OverallCond    9 3.6792e+13 4.0880e+12  746.28 < 2.2e-16 ***
## Residuals   1086 5.9489e+12 5.4778e+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.

```
## Analysis of Variance Table
##
## Response: SalePrice
##                          Df     Sum Sq    Mean Sq F value    Pr(>F)
## MSZoning                  4 7.7127e+11 1.9282e+11 101.058 < 2.2e-16 ***
## OverallQual               9 3.9224e+12 4.3582e+11 228.416 < 2.2e-16 ***
## MSZoning:OverallQual     16 6.4171e+10 4.0107e+09   2.102  0.006682 **
## Residuals              1065 2.0320e+12 1.9080e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Analysis of Variance Table
##
## Response: SalePrice
##                            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## OverallQual                 9 4.5195e+12 5.0216e+11 251.2801 < 2.2e-16 ***
## OverallCond                 8 2.1085e+10 2.6357e+09   1.3189    0.2299
## OverallQual:OverallCond    32 1.6090e+11 5.0280e+09   2.5160 8.352e-06 ***
## Residuals                1045 2.0884e+12 1.9984e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

III. Modeling and diagnostics

Given the number of regressors, we choose a model favouring feature sparsity. We choose to use a stepwise selection procedure with an AIC criterion.

We will predict SalePrice's logarithm to improve our target smoothness and estimated residuals homoscedasticity. In initial models we ran, we noticed that our 8 outliers were located towards the extreme bottom values of SalePrice. They prevented us from validating our postulates, mainly homoscedasticity and gaussianity. Instead of removing them, we use the winsor method to reaffect extreme values and normalise our data. Doing this greatly improves model robustness (the postulates) while diminishing our MSE on test data.

We noted earlier that many categorical variables could be considered uninformative or redundant. - We remove factors with massively overrepresented categories: Street, Utilities, RoofMatl, Condition2, Heating, Electrical, Functional, GarageCond. - Based on Anova tests, we remove other factors to improve model robustness: OverallCond, Exterior1st, Exterior2nd. OverallCond in particular was too specific and weakened our model by creating observations with leverage one. - For some features, we were not sure wether or not they had an influence, so we test the model with and without them and remove them if they do not improve our score : HeatingQC, SaleCondition, BsmtFinType2, RoofStyle, ExterQual.

```
## 5   8   21   13   37   40   51   59   22   23   17   38   66   34   20   26
```

Running a model with all the variables excluding the ones we just removed, we obtain an adjusted R2 of 0.90.

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)

To improve model efficiency, we will perform variable step selection.

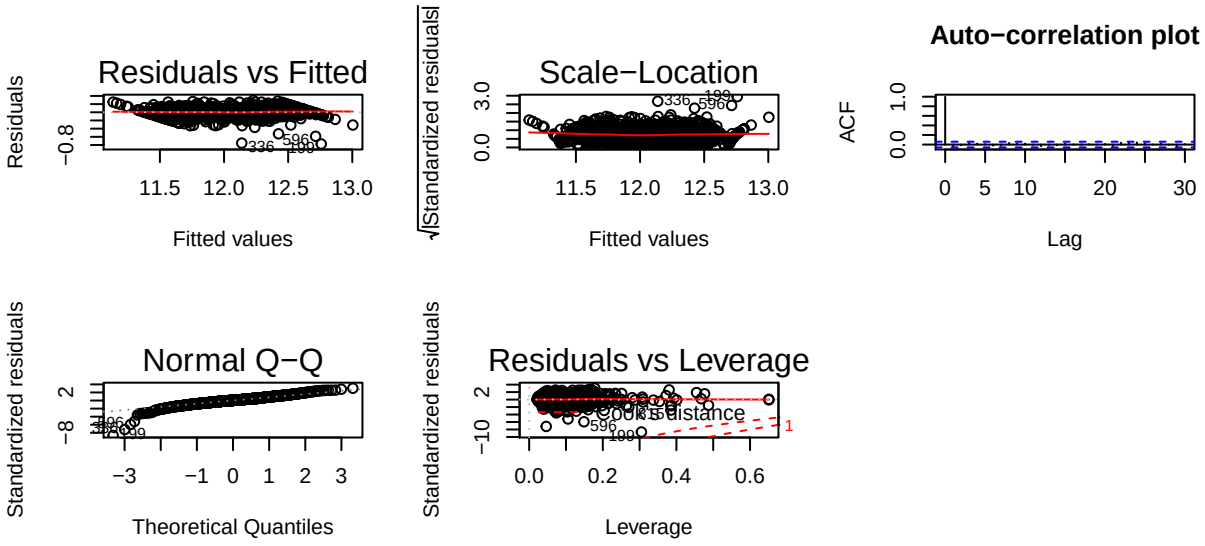
To choose among the three methods, we retrieve the AIC of each model and choose the one with the smallest AIC.

```
## [1]   146.000 -4732.393
```

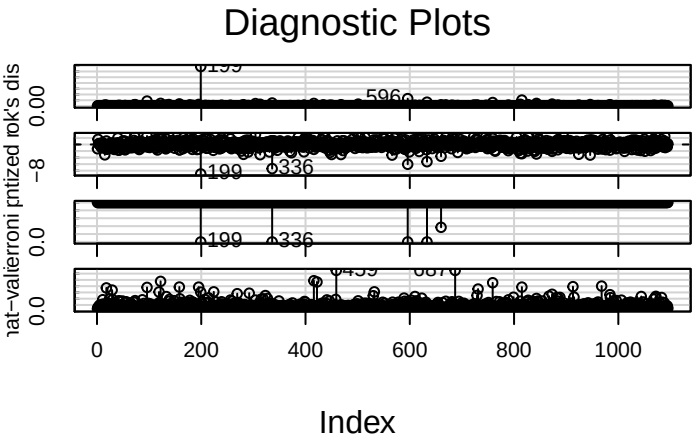
```
## [1]    96.000 -4796.873
```

```
## [1]    96.000 -4796.873
```

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 arbitrarily choose the backward model. 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.



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 -9.003915         1.0801e-18   1.1827e-15
## 336 -7.375215         3.4445e-13   3.7717e-10
## 596 -6.037520         2.2047e-09   2.4141e-06
## 633 -5.258263         1.7789e-07   1.9479e-04
```

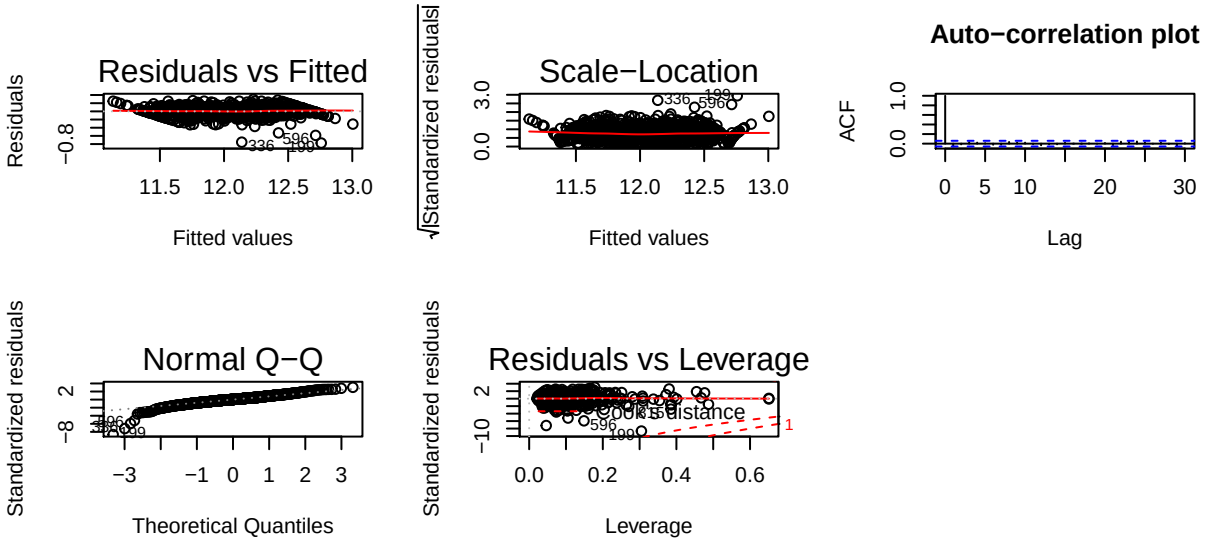
Using this test, 4 possible outliers are given according to the studentized criterion and Bonferroni p-value. We can now test our model without them. Note that those tests are based on confidence intervals and then for this reason, given the number of observations, we could have up to 70 outliers.

```
## # A tibble: 4 x 51
##   MSSubClass MSZoning LotFrontage LotArea LotShape LandContour LotConfig
##   <fct>      <fct>          <dbl>   <dbl> <fct>    <fct>       <fct>
## 1 60         RL             6.19    3.89  IR3      Bnk         Corner
## 2 20         RL             0.701   0.523 Reg      Lvl         Inside
## 3 60         RL             2.19    2.94  IR1      Bnk         Inside
## 4 20         RL             0.308   0.182 Reg      Lvl         Inside
## # ... with 44 more variables: LandSlope <fct>, Neighborhood <fct>,
## #   Condition1 <fct>, BldgType <fct>, HouseStyle <fct>, OverallQual <fct>,
## #   YearBuilt <dbl>, YearRemodAdd <dbl>, MasVnrType <fct>,
## #   MasVnrArea <dbl>, ExterCond <fct>, Foundation <fct>, BsmtQual <fct>,
## #   BsmtCond <fct>, BsmtExposure <fct>, BsmtFinType1 <fct>,
## #   BsmtFinSF1 <dbl>, BsmtUnfSF <dbl>, TotalBsmtSF <dbl>,
## #   CentralAir <fct>, `1stFlrSF` <dbl>, `2ndFlrSF` <dbl>, GrLivArea <dbl>,
## #   BsmtFullBath <dbl>, BsmtHalfBath <dbl>, FullBath <dbl>,
## #   HalfBath <dbl>, BedroomAbvGr <dbl>, KitchenQual <fct>,
## #   TotRmsAbvGrd <dbl>, Fireplaces <dbl>, GarageType <fct>,
## #   GarageYrBlt <dbl>, GarageFinish <fct>, GarageCars <dbl>,
## #   GarageArea <dbl>, GarageQual <fct>, PavedDrive <fct>,
## #   WoodDeckSF <dbl>, OpenPorchSF <dbl>, MoSold <fct>, YrSold <dbl>,
```

```
## #   SaleType <fct>, SalePrice <dbl>
##
## Call:
## lm(formula = log(SalePrice) ~ MSSubClass + MSZoning + LotFrontage +
##     LotArea + LotConfig + Neighborhood + Condition1 + OverallQual +
##     YearBuilt + YearRemodAdd + MasVnrType + MasVnrArea + BsmtCond +
##     BsmtExposure + BsmtFinSF1 + BsmtUnfSF + TotalBsmtSF + CentralAir +
##     `1stFlrSF` + GrLivArea + BsmtFullBath + FullBath + HalfBath +
##     KitchenQual + Fireplaces + GarageType + GarageYrBlt + GarageCars +
##     GarageQual + WoodDeckSF + MoSold + YrSold, data = home_streamlined)
##
## Coefficients:
##                     (Intercept)                     MSSubClass30
##                       11.917175                        -0.080852
##                    MSSubClass40                     MSSubClass45
##                       -0.081549                        -0.034448
##                    MSSubClass50                     MSSubClass60
##                       -0.075474                        -0.063908
##                    MSSubClass70                     MSSubClass75
##                       -0.069136                        -0.062733
##                    MSSubClass80                     MSSubClass85
##                       -0.015248                        -0.011692
##                    MSSubClass90                    MSSubClass120
##                       -0.101923                        -0.023670
##                   MSSubClass160                    MSSubClass180
##                       -0.204410                        -0.074275
##                   MSSubClass190                       MSZoningFV
##                       -0.111513                         0.148316
##                      MSZoningRH                       MSZoningRL
##                         0.121321                         0.114918
##                      MSZoningRM                      LotFrontage
##                         0.056130                        -0.011001
##                         LotArea                   LotConfigCulDSac
##                         0.034098                         0.016712
##                     LotConfigFR2                     LotConfigFR3
##                       -0.035256                        -0.056607
##                  LotConfigInside                NeighborhoodCollgCr
##                       -0.015288                        -0.092616
##              NeighborhoodCrawfor                NeighborhoodEdwards
##                         0.064801                        -0.159345
##              NeighborhoodGilbert                  NeighborhoodNAmes
##                       -0.110303                        -0.107447
##              NeighborhoodNridgHt                 NeighborhoodNWAmes
##                       -0.018157                        -0.112807
##              NeighborhoodOldTown                 NeighborhoodSawyer
##                       -0.087358                        -0.105861
##              NeighborhoodSawyerW                NeighborhoodSomerst
##                       -0.107660                        -0.037495
##      NeighborhoodN_Under20Sales         NeighborhoodN_Under50Sales
##                       -0.085849                        -0.077407
##                  Condition1Norm                   Condition1C_Other
##                         0.030198                        -0.011111
##  OverallQual-0.782903446215014  OverallQual-0.0421794726376995
```

```
##                       0.006930                       0.057213
##   OverallQual0.667269615883086    OverallQual1.35046226437619
##                       0.133803                       0.204031
##    OverallQual2.01114682580545    OverallQual2.65221758679825
##                       0.262383                      -0.012081
##           OverallQualVery_Low                      YearBuilt
##                      -0.030592                       0.032600
##                   YearRemodAdd              MasVnrTypeBrkFace
##                       0.032114                       0.096602
##                MasVnrTypeNone                MasVnrTypeStone
##                       0.023277                       0.101564
##                    MasVnrArea                    BsmtCondGd
##                      -0.032167                       0.096559
##                    BsmtCondPo                    BsmtCondTA
##                      -0.008007                       0.064245
##                BsmtExposureGd                BsmtExposureMn
##                       0.048867                       0.004259
##                BsmtExposureNo                    BsmtFinSF1
##                      -0.015894                       0.039054
##                     BsmtUnfSF                   TotalBsmtSF
##                       0.013027                       0.012576
##                   CentralAirY                    `1stFlrSF`
##                       0.040556                      -0.026535
##                     GrLivArea                  BsmtFullBath
##                       0.121743                       0.015709
##                      FullBath                      HalfBath
##                       0.018749                       0.016740
##                  KitchenQualFa                  KitchenQualGd
##                      -0.065183                      -0.053213
##                  KitchenQualTA                    Fireplaces
##                      -0.084106                       0.016637
##               GarageTypeAttchd              GarageTypeBasment
##                       0.181055                       0.180410
##              GarageTypeBuiltIn              GarageTypeCarPort
##                       0.179749                       0.086710
##               GarageTypeDetchd                    GarageYrBlt
##                       0.171497                      -0.014742
##                     GarageCars                   GarageQualFa
##                       0.046131                      -0.339776
##                   GarageQualGd                   GarageQualPo
##                      -0.196304                      -0.316996
##                   GarageQualTA                     WoodDeckSF
##                      -0.264658                       0.009581
##        MoSold-1.69365012137611        MoSold-1.25178437070736
##                       0.020208                       0.038224
##       MoSold-0.837792285242915       MoSold-0.444737988378689
##                       0.046209                       0.066631
##      MoSold-0.0682823411762393        MoSold0.294523839408755
##                       0.053717                       0.049159
##        MoSold0.645803668048188        MoSold0.987151736743589
##                       0.053533                       0.032916
##         MoSold1.31980535250846         MoSold1.64474974582165
##                       0.017692                       0.034410
##         MoSold1.96278610828061                         YrSold
```

```
##                0.024539                -0.005854
```



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 + LotConfig + Neighborhood + Condition1 + OverallQual + YearBuilt + YearRemodAdd + MasVnrType + MasVnrArea + BsmtCond + BsmtExposure + BsmtFinSF1 + BsmtUnfSF + TotalBsmtSF + CentralAir + `1stFlrSF` + GrLivArea + BsmtFullBath + FullBath + HalfBath + KitchenQual + Fireplaces + GarageType + GarageYrBlt + 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.

```
## [1] 33908.07
```

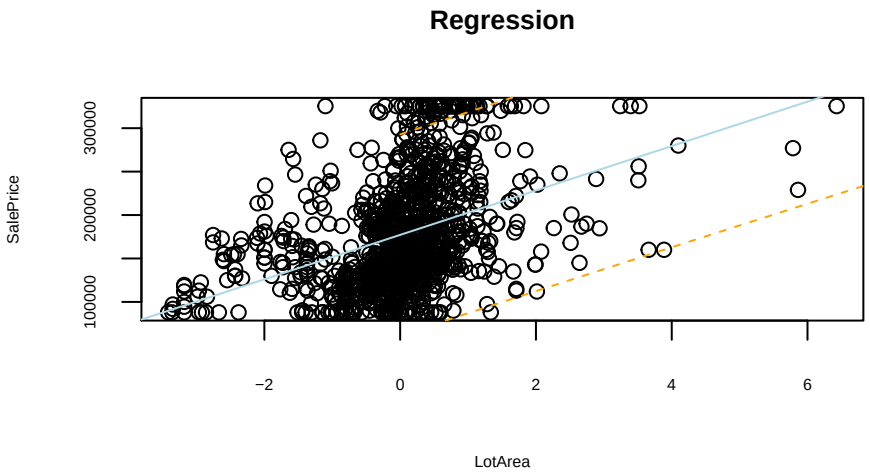
```
## [1] 0.3150128
```

```
## [1] 33908.07
```

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: as our regression model is mostly impacted by factors, doing confidence intervals does

not help us visualize the proximity among SalePrice and the most important features. However, doing a 95% confidence interval with LotArea, we notice that this variable is not good at emplaning SalePrice by itself, as the price can double for a same value of LotArea.



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.

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. Key learnings:
Improvements for future projects: