

# 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 19/23 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  $\log(\text{SalePrice})$  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 description** 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 than 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, MSSubClass and Fireplaces 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.

Recasting some of our features as factors improves our predictions but decreases model robustness, so we will have to select the factors we keep carefully.

We check there are no missing values in the pre-processed data before launching into the analysis.

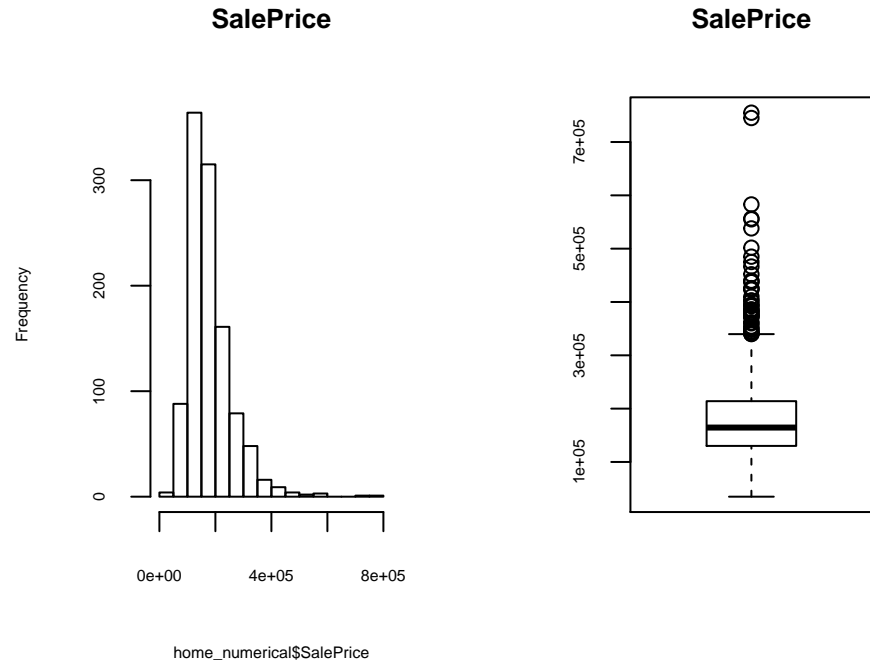
```
## There are 0 missing values in our dataframe.
```

## II. Data exploration

### 1. Numerical analysis.

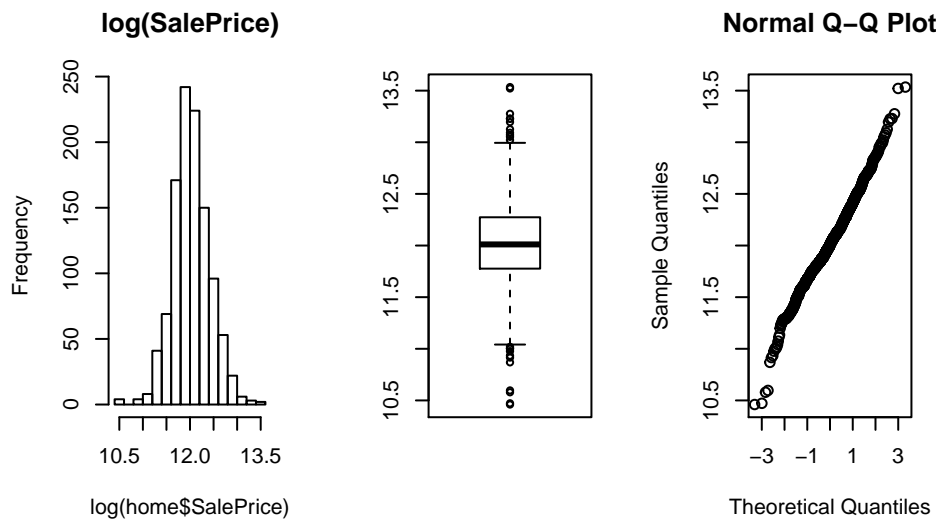
We describe the output data before analysing the correlation matrix. A PCA analysis we performed showed us that 23/28 numerical variables accounted for 99% of the variance. We left it out for concision.

We first set out to observe the output, SalePrice.

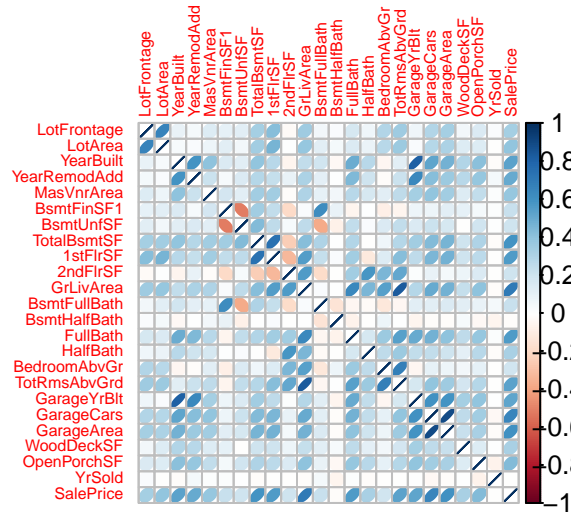


```
## Data skewness is 1.91628
```

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.

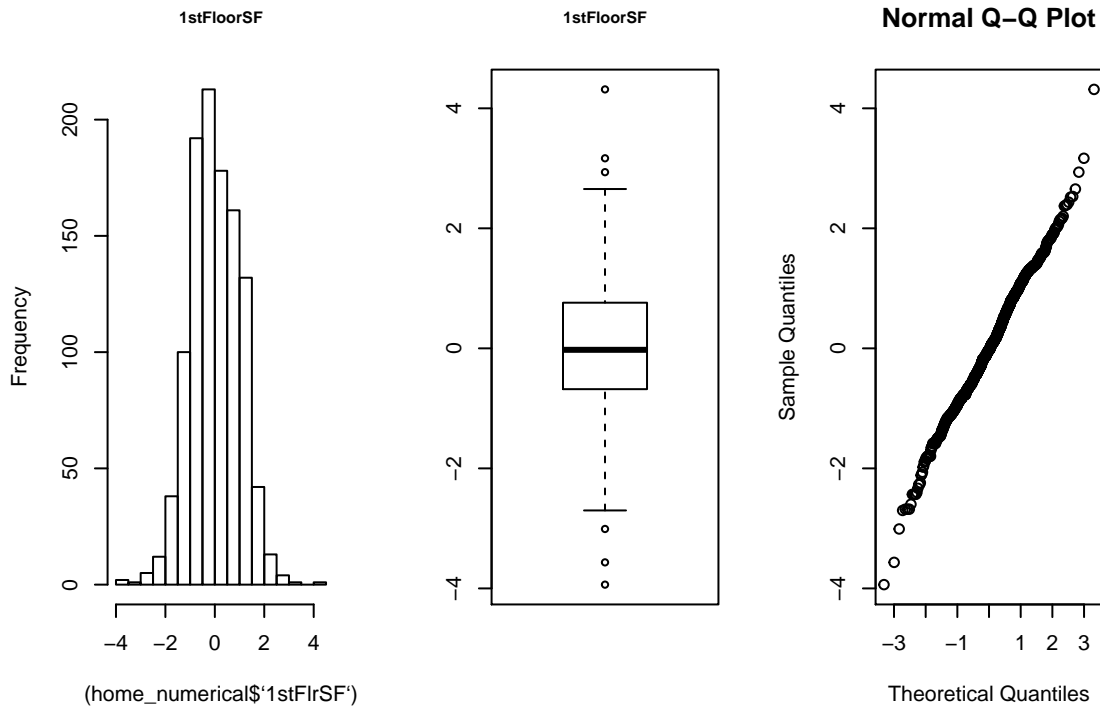


log(SalePrice) is pretty close to a normal distribution, except for extreme values.



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 BsmHalfBath 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.

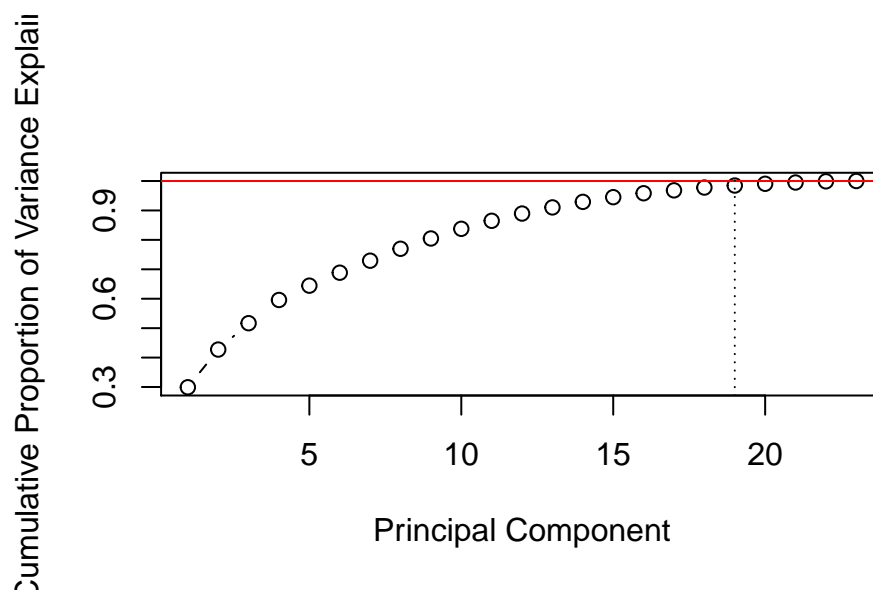
Our intuition is that three feature types mainly drive SalePrices: area/surface, location and quality. Let's describe 1stFlrSF as it is the closest feature to square feet with 2ndFloorSF.



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

PCA Because PCA works best with (scaled) numerical data, we will perform it on the numerical features of

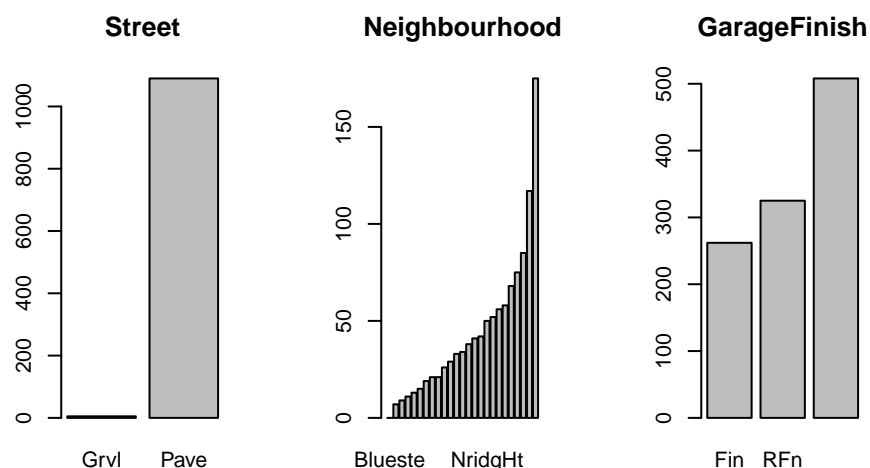
our preprocessed data.



19/23 numerical variables account for ~99% of the variance.

### 3. Factor analysis

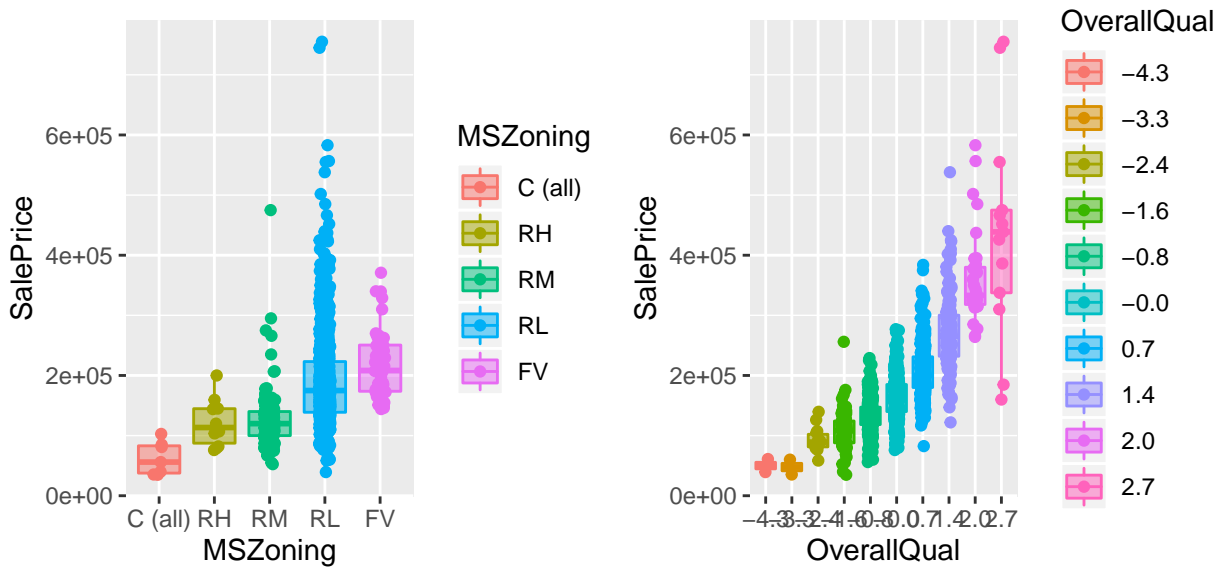
We first investigate level fragmentation within the factors. We discover three types of factors, as per examples below.



- i) Clear underrepresentation of some levels. In Street and Utilities, which are binary, this is conspicuous. From this we infer that these factors will not be useful in predicting house price in general: nearly all houses will be in the same category along these factors (and for those who are not, 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
```

Both area and overall quality have a strong effect on SalePrice. Let's perform an ancova to evaluate interaction importance.

```
## 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
```

SalePrice highly depends on MSZoning, OverallQual and their interaction. So we will keep these factors.

### III. Modelling and diagnostics

Given the number of regressors, we choose models favouring feature sparsity. We implement stepwise selection with an AIC criterion. We predict SalePrice's logarithm to improve our target smoothness and estimated residuals homoscedasticity. In our first model, we notice that our 8 outliers are located towards the extreme bottom values of SalePrice. They prevent us from validating our postulates, mainly homoscedasticity and gaussianity. Removing them gives a more robust model, but we thought of testing a second approach. We fit a second model where instead of removing the outliers after model fit, we use the winsor method to reffect extreme values and normalise our data before model fit. Doing this greatly improves model robustness (the postulates) but also increases our MSE on test data. We choose to keep model1 for our fit as the aim is to predict data accurately.

We noted earlier that many categorical variables could be considered uninformative or redundant. - We removed factors with massively overrepresented categories: Street, Utilities, RoofMatl, Condition2, Heating, Electrical, Functional, GarageCond.

We also made some regroupments within factors to diminish cardinality, mostly for neighbourhood. - Based on Anova tests, we removed 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.

```
## 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
```

- 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.

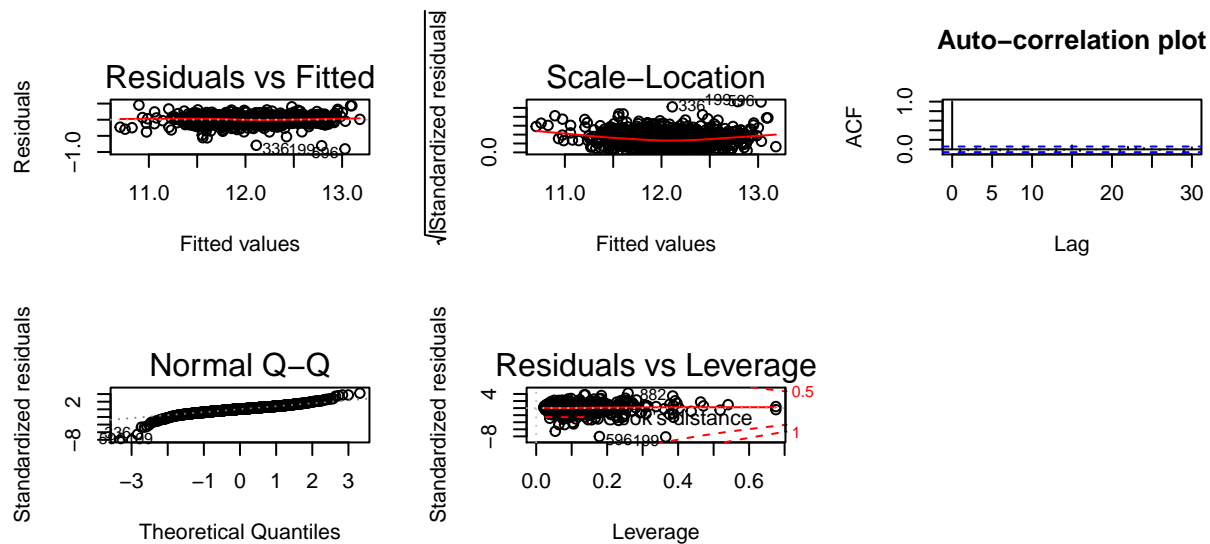
After working on our factors, we start by fitting a full linear model to predict  $\log(\text{SalePrice})$ .

Running a model with all the variables excluding the ones we just removed, we obtain an adjusted R-squared of 0.91. Many variables could be removed from our model while marginally affecting its efficiency to explain SalePrice. To improve model efficiency, we 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.

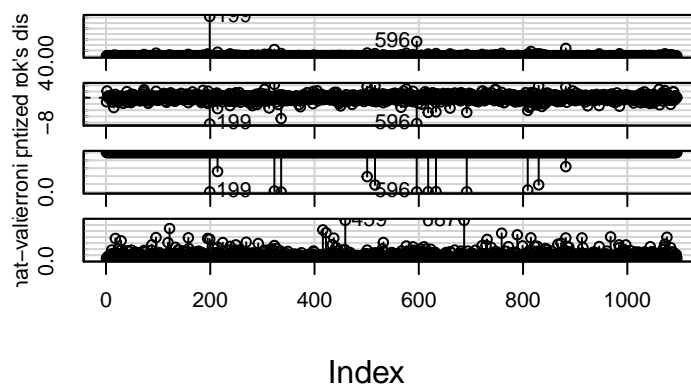
```
## [1] 148.000 -4420.497
## [1] 106.000 -4469.426
## [1] 106.000 -4469.426
```

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 arbitrarily choose the backward model. Let's ensure 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.

### Diagnostic Plots



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 are 524 and 1299. Studentised residuals : big values Bonferroni's plot : we notice that 6 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:

##	student	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, 8 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, meaning it

is normal that a proportion of our observations should be outliers.

```
##
## Call:
## 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_streamlined)
##
## Coefficients:
##               (Intercept)               MSSubClass30
##               11.350506                -0.077322
##               MSSubClass40               MSSubClass45
##               -0.090111                 0.025474
##               MSSubClass50               MSSubClass60
##               -0.036479                -0.046432
##               MSSubClass70               MSSubClass75
##               -0.051225                -0.019121
##               MSSubClass80               MSSubClass85
##               -0.012152                -0.003077
##               MSSubClass90               MSSubClass120
##               -0.099026                -0.027979
##               MSSubClass160              MSSubClass180
##               -0.173679                -0.064564
##               MSSubClass190              MSZoningRH
##               -0.087888                 0.556874
##               MSZoningRM                 MSZoningRL
##               0.483250                 0.534114
##               MSZoningFV                 LotFrontage
##               0.574965                -0.012574
##               LotArea                 LandContourHLS
##               0.041623                 0.087359
##               LandContourLow             LandContourLvl
##               0.049665                 0.061569
##               LotConfigCulDSac           LotConfigFR2
##               0.027451                -0.045210
##               LotConfigFR3              LotConfigInside
##               -0.062062                -0.009755
##               LandSlopeMod              LandSlopeSev
##               -0.018481                -0.092259
##               NeighborhoodCollgCr        NeighborhoodCrawfor
##               -0.067561                 0.086173
##               NeighborhoodEdwards        NeighborhoodGilbert
##               -0.155320                -0.090103
##               NeighborhoodNames          NeighborhoodNridgHt
##               -0.081238                -0.002297
##               NeighborhoodNWames        NeighborhoodOldTown
##               -0.103781                -0.074285
##               NeighborhoodSawyer        NeighborhoodSawyerW
##               -0.091245                -0.093502
##               NeighborhoodSomerst      NeighborhoodN_Under20Sales
```



##	-0.027608	-0.060547
##	NeighborhoodN_Under50Sales	Condition1Norm
##	-0.044703	0.036216
##	Condition1C_Other	OverallQual-0.8
##	-0.015303	0.015744
##	OverallQual-0.0	OverallQual0.7
##	0.065721	0.133070
##	OverallQual1.4	OverallQual2.0
##	0.209206	0.324974
##	OverallQual2.7	OverallQualVery_Low
##	0.224853	-0.045904
##	YearBuilt	YearRemodAdd
##	0.026406	0.035886
##	MasVnrTypeBrkFace	MasVnrTypeNone
##	0.085118	-0.005724
##	MasVnrTypeStone	MasVnrArea
##	0.089817	-0.045255
##	ExterCondEC_Other	FoundationCBlock
##	-0.021848	0.014234
##	FoundationPConc	FoundationSlab
##	0.003180	-0.063356
##	FoundationStone	FoundationWood
##	0.151518	-0.099073
##	BsmtCondGd	BsmtCondPo
##	0.104931	-0.163461
##	BsmtCondTA	BsmtExposureGd
##	0.075263	0.068427
##	BsmtExposureMn	BsmtExposureNo
##	0.011306	-0.013430
##	BsmtFinSF1	BsmtUnfSF
##	0.043217	0.014431
##	CentralAirY	GrLivArea
##	0.078982	0.129656
##	BsmtFullBath	FullBath
##	0.015273	0.019761
##	HalfBath	KitchenQualFa
##	0.018521	-0.105635
##	KitchenQualGd	KitchenQualTA
##	-0.099793	-0.117536
##	Fireplaces0.783736098878231	Fireplaces1.63788722614665
##	0.024607	0.048001
##	Fireplaces2.16737616223006	GarageTypeAttchd
##	-0.203292	0.197488
##	GarageTypeBasment	GarageTypeBuiltIn
##	0.218296	0.212803
##	GarageTypeCarPort	GarageTypeDetchd
##	0.075524	0.205749
##	GarageCars	GarageQualFa
##	0.045096	-0.321068
##	GarageQualGd	GarageQualPo
##	-0.189143	-0.344484
##	GarageQualTA	WoodDeckSF
##	-0.248701	0.009288
##	MoSold-1.69365012137611	MoSold-1.25178437070736

```

##              0.023123              0.024008
##      MoSold-0.837792285242915      MoSold-0.444737988378689
##              0.045720              0.060891
##      MoSold-0.0682823411762393      MoSold0.294523839408755
##              0.051950              0.048830
##      MoSold0.645803668048188      MoSold0.987151736743589
##              0.043002              0.022899
##      MoSold1.31980535250846      MoSold1.64474974582165
##              0.014952              0.020988
##      MoSold1.96278610828061      YrSold
##              0.005324              -0.005742

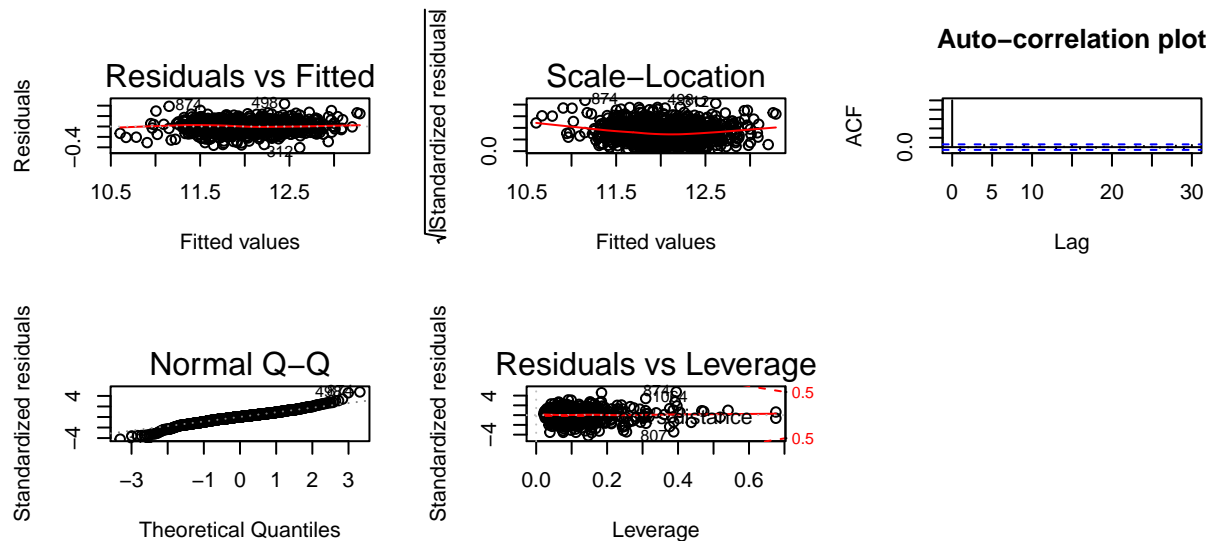
##
## Call:
## lm(formula = log(SalePrice) ~ MSSubClass + MSZoning + LotFrontage +
##      LotArea + LandContour + LotConfig + LandSlope + Neighborhood +
##      Condition1 + OverallQual + YearBuilt + YearRemodAdd + MasVnrType +
##      MasVnrArea + ExterCond + Foundation + BsmtCond + BsmtExposure +
##      BsmtFinSF1 + BsmtUnfSF + CentralAir + GrLivArea + BsmtFullBath +
##      FullBath + HalfBath + KitchenQual + Fireplaces + GarageType +
##      GarageCars + GarageQual + WoodDeckSF + MoSold + YrSold, data = home_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40661 -0.05719  0.00455  0.06063  0.43414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.3034828   0.1138074   99.321 < 2e-16 ***
## MSSubClass30      -0.0601748   0.0224591    -2.679 0.007501 **
## MSSubClass40      -0.0902199   0.0635652    -1.419 0.156121
## MSSubClass45        0.0252195   0.0401676     0.628 0.530244
## MSSubClass50      -0.0435977   0.0178529    -2.442 0.014780 *
## MSSubClass60      -0.0386042   0.0148492    -2.600 0.009469 **
## MSSubClass70      -0.0454124   0.0255024    -1.781 0.075269 .
## MSSubClass75      -0.0308055   0.0401118    -0.768 0.442676
## MSSubClass80      -0.0218480   0.0187507    -1.165 0.244229
## MSSubClass85      -0.0115652   0.0300692    -0.385 0.700602
## MSSubClass90      -0.1095573   0.0228313    -4.799 1.85e-06 ***
## MSSubClass120       0.0071513   0.0196633     0.364 0.716170
## MSSubClass160      -0.1255077   0.0289736    -4.332 1.63e-05 ***
## MSSubClass180      -0.0187411   0.0496866    -0.377 0.706117
## MSSubClass190      -0.0977195   0.0288848    -3.383 0.000745 ***
## MSZoningRH         0.6051180   0.0599447   10.095 < 2e-16 ***
## MSZoningRM         0.5325383   0.0496644   10.723 < 2e-16 ***
## MSZoningRL         0.5852759   0.0484257   12.086 < 2e-16 ***
## MSZoningFV         0.5995738   0.0584237   10.263 < 2e-16 ***
## LotFrontage       -0.0014254   0.0061411    -0.232 0.816508
## LotArea            0.0555982   0.0064528     8.616 < 2e-16 ***
## LandContourHLS      0.0260397   0.0264579     0.984 0.325264
## LandContourLow     -0.0030692   0.0302994    -0.101 0.919337
## LandContourLvl      0.0006548   0.0181557     0.036 0.971238
## LotConfigCulDSac    0.0273224   0.0156145     1.750 0.080464 .
## LotConfigFR2       -0.0523807   0.0215435    -2.431 0.015219 *

```

## LotConfigFR3	-0.0515285	0.0554965	-0.929	0.353376	
## LotConfigInside	-0.0023012	0.0090665	-0.254	0.799697	
## LandSlopeMod	-0.0314202	0.0194478	-1.616	0.106499	
## LandSlopeSev	-0.1129887	0.0403212	-2.802	0.005175	**
## NeighborhoodCollgCr	-0.0981782	0.0249593	-3.934	8.96e-05	***
## NeighborhoodCrawfor	0.0650044	0.0277313	2.344	0.019273	*
## NeighborhoodEdwards	-0.1372407	0.0236648	-5.799	8.97e-09	***
## NeighborhoodGilbert	-0.1254898	0.0278178	-4.511	7.23e-06	***
## NeighborhoodNames	-0.0970275	0.0228298	-4.250	2.34e-05	***
## NeighborhoodNridgHt	-0.0725456	0.0298031	-2.434	0.015104	*
## NeighborhoodNWAms	-0.1106621	0.0272557	-4.060	5.30e-05	***
## NeighborhoodOldTown	-0.0726649	0.0232579	-3.124	0.001835	**
## NeighborhoodSawyer	-0.0978080	0.0260576	-3.754	0.000185	***
## NeighborhoodSawyerW	-0.1241208	0.0281828	-4.404	1.18e-05	***
## NeighborhoodSomerst	-0.0383775	0.0359541	-1.067	0.286053	
## NeighborhoodN_Under20Sales	-0.0695249	0.0286976	-2.423	0.015587	*
## NeighborhoodN_Under50Sales	-0.0754999	0.0221106	-3.415	0.000665	***
## Condition1Norm	0.0320815	0.0151795	2.113	0.034812	*
## Condition1C_Other	-0.0124832	0.0188825	-0.661	0.508705	
## OverallQual-0.8	0.0146045	0.0140812	1.037	0.299916	
## OverallQual-0.0	0.0616287	0.0159161	3.872	0.000115	***
## OverallQual0.7	0.1359265	0.0183084	7.424	2.46e-13	***
## OverallQual1.4	0.2116265	0.0219651	9.635	< 2e-16	***
## OverallQual2.0	0.3267251	0.0316171	10.334	< 2e-16	***
## OverallQual2.7	0.3959621	0.0440920	8.980	< 2e-16	***
## OverallQualVery_Low	-0.0760761	0.0311256	-2.444	0.014694	*
## YearBuilt	0.0422051	0.0105702	3.993	7.02e-05	***
## YearRemodAdd	0.0354792	0.0052050	6.816	1.63e-11	***
## MasVnrTypeBrkFace	0.1100902	0.0315622	3.488	0.000508	***
## MasVnrTypeNone	0.0554322	0.0497103	1.115	0.265078	
## MasVnrTypeStone	0.1284775	0.0339556	3.784	0.000164	***
## MasVnrArea	-0.0243996	0.0192837	-1.265	0.206065	
## ExterCondEC_Other	-0.0120585	0.0108424	-1.112	0.266342	
## FoundationCBBlock	0.0070629	0.0159742	0.442	0.658483	
## FoundationPConc	0.0017332	0.0177280	0.098	0.922140	
## FoundationSlab	-0.0535781	0.0344657	-1.555	0.120380	
## FoundationStone	0.1899690	0.0486767	3.903	0.000102	***
## FoundationWood	-0.1188683	0.0669172	-1.776	0.075985	.
## BsmtCondGd	0.0854387	0.0263179	3.246	0.001208	**
## BsmtCondPo	-0.1477945	0.0930524	-1.588	0.112542	
## BsmtCondTA	0.0657172	0.0211560	3.106	0.001949	**
## BsmtExposureGd	0.0638864	0.0151680	4.212	2.77e-05	***
## BsmtExposureMn	0.0014972	0.0152230	0.098	0.921673	
## BsmtExposureNo	-0.0198406	0.0110677	-1.793	0.073335	.
## BsmtFinSF1	0.0467153	0.0055907	8.356	< 2e-16	***
## BsmtUnfSF	0.0148813	0.0054967	2.707	0.006901	**
## CentralAirY	0.0855719	0.0166204	5.149	3.17e-07	***
## GrLivArea	0.1301474	0.0075770	17.177	< 2e-16	***
## BsmtFullBath	0.0131721	0.0044599	2.953	0.003218	**
## FullBath	0.0204215	0.0058947	3.464	0.000554	***
## HalfBath	0.0157234	0.0053082	2.962	0.003129	**
## KitchenQualFa	-0.1149781	0.0288919	-3.980	7.41e-05	***
## KitchenQualGd	-0.0922609	0.0173534	-5.317	1.31e-07	***
## KitchenQualTA	-0.1022705	0.0193915	-5.274	1.64e-07	***

```
## Fireplaces0.783736098878231 0.0261696 0.0086219 3.035 0.002467 **
## Fireplaces1.63788722614665 0.0310891 0.0150600 2.064 0.039247 *
## Fireplaces2.16737616223006 0.0176830 0.0554439 0.319 0.749844
## GarageTypeAttchd 0.1716695 0.0474233 3.620 0.000310 ***
## GarageTypeBasment 0.1660180 0.0565207 2.937 0.003388 **
## GarageTypeBuiltIn 0.1983511 0.0498179 3.982 7.35e-05 ***
## GarageTypeCarPort 0.0636236 0.0811098 0.784 0.432987
## GarageTypeDetchd 0.1965417 0.0467125 4.207 2.82e-05 ***
## GarageCars 0.0360690 0.0053853 6.698 3.56e-11 ***
## GarageQualFa -0.2400155 0.0671412 -3.575 0.000368 ***
## GarageQualGd -0.1278881 0.0724200 -1.766 0.077720 .
## GarageQualPo -0.2892958 0.0976897 -2.961 0.003136 **
## GarageQualTA -0.1918312 0.0645793 -2.970 0.003046 **
## WoodDeckSF 0.0081391 0.0037684 2.160 0.031026 *
## MoSold-1.69365012137611 0.0134931 0.0232442 0.580 0.561714
## MoSold-1.25178437070736 0.0078354 0.0204973 0.382 0.702348
## MoSold-0.837792285242915 0.0313157 0.0193420 1.619 0.105758
## MoSold-0.444737988378689 0.0492762 0.0185203 2.661 0.007926 **
## MoSold-0.0682823411762393 0.0381092 0.0179652 2.121 0.034149 *
## MoSold0.294523839408755 0.0327119 0.0182158 1.796 0.072834 .
## MoSold0.645803668048188 0.0343832 0.0196728 1.748 0.080821 .
## MoSold0.987151736743589 0.0048039 0.0227853 0.211 0.833063
## MoSold1.31980535250846 0.0080089 0.0208154 0.385 0.700501
## MoSold1.64474974582165 -0.0034528 0.0219638 -0.157 0.875116
## MoSold1.96278610828061 0.0069220 0.0226878 0.305 0.760355
## YrSold -0.0051390 0.0034269 -1.500 0.134036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1047 on 981 degrees of freedom
## Multiple R-squared:  0.9361, Adjusted R-squared:  0.9293
## F-statistic: 136.9 on 105 and 981 DF, p-value: < 2.2e-16
```

We now have a 93% of RMSE on train data.



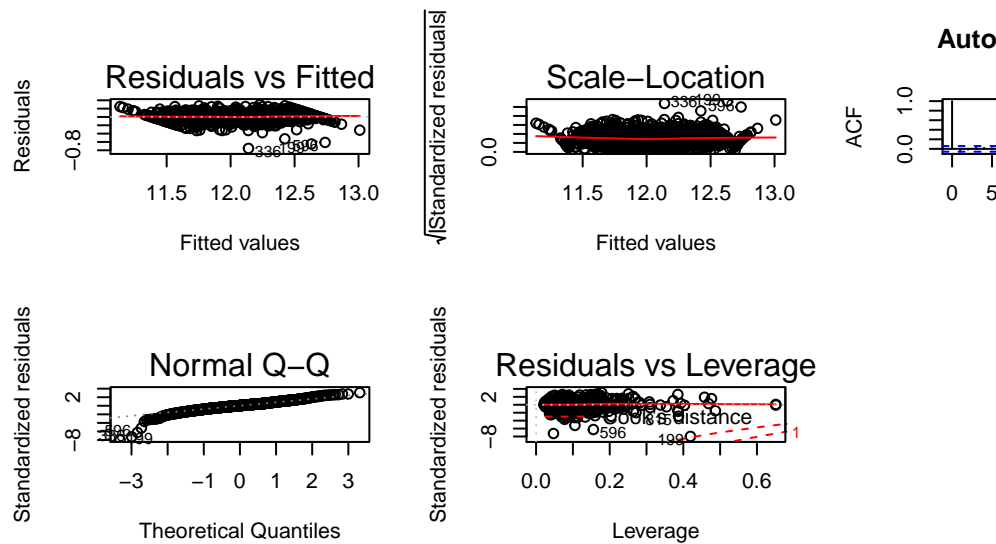
According to the graphs, P1 and P3 are validated without hesitation. For P2, we notice a slight elliptic behaviour in the middle, but it seems limited enough to validate the postulate. For P4, tail observations are further from the normal distribution. Based on the high number of observations in our dataframe and the

few number of points not aligned with the normal distribution quantiles, we validate the postulate.

```
##
## 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)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75466 -0.05378  0.00444  0.05621  0.29604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.907095   0.110388  107.865 < 2e-16 ***
## MSSubClass30    -0.079796   0.022115   -3.608 0.000324 ***
## MSSubClass40    -0.082519   0.064745   -1.275 0.202776
## MSSubClass45    -0.032554   0.040293   -0.808 0.419319
## MSSubClass50    -0.077831   0.021036   -3.700 0.000228 ***
## MSSubClass60    -0.063311   0.021814   -2.902 0.003786 **
## MSSubClass70    -0.072169   0.030533   -2.364 0.018287 *
## MSSubClass75    -0.068995   0.044367   -1.555 0.120240
## MSSubClass80    -0.009682   0.019482   -0.497 0.619311
## MSSubClass85    -0.009299   0.030620   -0.304 0.761426
## MSSubClass90    -0.102402   0.023017   -4.449 9.60e-06 ***
## MSSubClass120   -0.022108   0.019584   -1.129 0.259225
## MSSubClass160   -0.198678   0.032790   -6.059 1.94e-09 ***
## MSSubClass180   -0.065662   0.050092   -1.311 0.190218
## MSSubClass190   -0.114341   0.031321   -3.651 0.000275 ***
## MSZoningRH      0.122984   0.057707    2.131 0.033319 *
## MSZoningRM      0.058880   0.047315    1.244 0.213638
## MSZoningRL      0.116153   0.045889    2.531 0.011520 *
## MSZoningFV      0.150233   0.056377    2.665 0.007828 **
## LotFrontage    -0.009619   0.006120   -1.572 0.116309
## LotArea         0.034293   0.006174    5.555 3.57e-08 ***
## LotConfigCulDSac 0.016087   0.015860    1.014 0.310678
## LotConfigFR2    -0.035921   0.021916   -1.639 0.101524
## LotConfigFR3    -0.056306   0.056631   -0.994 0.320341
## LotConfigInside -0.014286   0.009127   -1.565 0.117850
## NeighborhoodCollgCr -0.092098   0.025070   -3.674 0.000252 ***
## NeighborhoodCrawfor 0.062969   0.027635    2.279 0.022901 *
## NeighborhoodEdwards -0.157959   0.023795   -6.638 5.20e-11 ***
## NeighborhoodGilbert -0.112590   0.027985   -4.023 6.17e-05 ***
## NeighborhoodNames -0.106131   0.022637   -4.688 3.14e-06 ***
## NeighborhoodNridgHt -0.019355   0.029710   -0.651 0.514900
## NeighborhoodNWAmes -0.116084   0.027006   -4.298 1.89e-05 ***
## NeighborhoodOldTown -0.088466   0.023364   -3.786 0.000162 ***
## NeighborhoodSawyer -0.108266   0.025974   -4.168 3.34e-05 ***
## NeighborhoodSawyerW -0.108970   0.028339   -3.845 0.000128 ***
## NeighborhoodSomerst -0.038943   0.035624   -1.093 0.274585
```

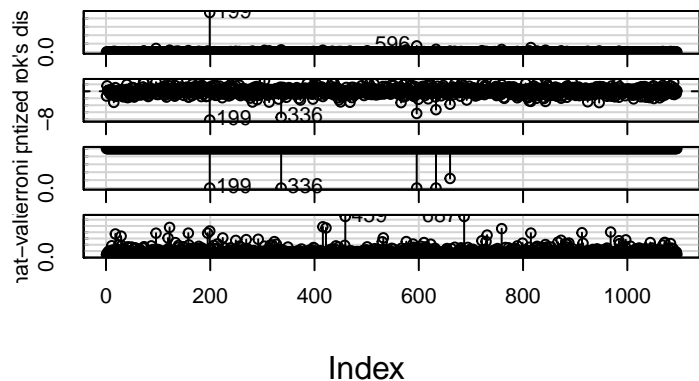
## NeighborhoodN_Under20Sales	-0.089241	0.028724	-3.107	0.001945	**
## NeighborhoodN_Under50Sales	-0.078632	0.022108	-3.557	0.000393	***
## Condition1Norm	0.027270	0.015238	1.790	0.073816	.
## Condition1C_Other	-0.014836	0.018976	-0.782	0.434494	
## OverallQual-0.8	0.006239	0.014136	0.441	0.659077	
## OverallQual-0.0	0.055248	0.015861	3.483	0.000517	***
## OverallQual0.7	0.133554	0.018335	7.284	6.56e-13	***
## OverallQual1.4	0.202552	0.022075	9.176	< 2e-16	***
## OverallQual2.0	0.261865	0.031934	8.200	7.34e-16	***
## OverallQual2.7	-0.010124	0.043282	-0.234	0.815106	
## OverallQualVery_Low	-0.028461	0.030794	-0.924	0.355589	
## YearBuilt	0.032158	0.010646	3.021	0.002588	**
## YearRemodAdd	0.033529	0.005315	6.308	4.24e-10	***
## MasVnrTypeBrkFace	0.095720	0.031848	3.006	0.002717	**
## MasVnrTypeNone	0.019031	0.050202	0.379	0.704702	
## MasVnrTypeStone	0.099786	0.034289	2.910	0.003693	**
## MasVnrArea	-0.034838	0.019503	-1.786	0.074352	.
## BsmtCondGd	0.096690	0.026102	3.704	0.000224	***
## BsmtCondPo	-0.011382	0.090578	-0.126	0.900030	
## BsmtCondTA	0.065912	0.020879	3.157	0.001643	**
## BsmtExposureGd	0.052361	0.015098	3.468	0.000547	***
## BsmtExposureMn	0.005158	0.015424	0.334	0.738125	
## BsmtExposureNo	-0.014428	0.011165	-1.292	0.196568	
## BsmtFinSF1	0.038184	0.006389	5.977	3.17e-09	***
## BsmtUnfSF	0.011101	0.006879	1.614	0.106922	
## TotalBsmtSF	0.016730	0.008417	1.988	0.047115	*
## CentralAirY	0.038337	0.016798	2.282	0.022686	*
## `1stFlrSF`	-0.028174	0.010898	-2.585	0.009869	**
## GrLivArea	0.123848	0.012318	10.054	< 2e-16	***
## BsmtFullBath	0.015207	0.004608	3.300	0.001000	**
## FullBath	0.018377	0.005944	3.091	0.002048	**
## HalfBath	0.016597	0.005457	3.041	0.002418	**
## KitchenQualFa	-0.067797	0.029109	-2.329	0.020054	*
## KitchenQualGd	-0.057284	0.017618	-3.251	0.001187	**
## KitchenQualTA	-0.085970	0.019569	-4.393	1.24e-05	***
## Fireplaces0.783736098878231	0.035445	0.008787	4.034	5.91e-05	***
## Fireplaces1.63788722614665	0.043178	0.015124	2.855	0.004394	**
## Fireplaces2.16737616223006	-0.126517	0.051775	-2.444	0.014715	*
## GarageTypeAttchd	0.177988	0.047830	3.721	0.000209	***
## GarageTypeBasement	0.175423	0.056795	3.089	0.002066	**
## GarageTypeBuiltIn	0.172125	0.050495	3.409	0.000679	***
## GarageTypeCarPort	0.080799	0.082053	0.985	0.325003	
## GarageTypeDetchd	0.169592	0.047159	3.596	0.000339	***
## GarageYrBlt	-0.014931	0.007239	-2.063	0.039408	*
## GarageCars	0.044770	0.005581	8.022	2.90e-15	***
## GarageQualFa	-0.337715	0.069643	-4.849	1.44e-06	***
## GarageQualGd	-0.195419	0.074114	-2.637	0.008501	**
## GarageQualPo	-0.313724	0.098968	-3.170	0.001571	**
## GarageQualTA	-0.261645	0.066371	-3.942	8.64e-05	***
## WoodDeckSF	0.009807	0.003830	2.561	0.010598	*
## MoSold-1.69365012137611	0.018258	0.023353	0.782	0.434513	
## MoSold-1.25178437070736	0.035541	0.020773	1.711	0.087402	.
## MoSold-0.837792285242915	0.043255	0.019461	2.223	0.026464	*
## MoSold-0.444737988378689	0.064477	0.018673	3.453	0.000578	***

```
## MoSold-0.0682823411762393    0.051714    0.018054    2.864 0.004266 **
## MoSold0.294523839408755      0.047998    0.018352    2.615 0.009049 **
## MoSold0.645803668048188      0.050469    0.019830    2.545 0.011076 *
## MoSold0.987151736743589      0.030069    0.022938    1.311 0.190198
## MoSold1.31980535250846       0.018422    0.020818    0.885 0.376422
## MoSold1.64474974582165       0.038203    0.022092    1.729 0.084073 .
## MoSold1.96278610828061       0.022346    0.022803    0.980 0.327341
## YrSold                        -0.005205    0.003470   -1.500 0.133967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1067 on 997 degrees of freedom
## Multiple R-squared:  0.9158, Adjusted R-squared:  0.9076
## F-statistic: 111.8 on 97 and 997 DF,  p-value: < 2.2e-16
```



We have an RMSE of 91% on train data.

### Diagnostic Plots



```
##          rstudent unadjusted p-value Bonferroni p
## 199 -8.240963      5.3435e-16    5.8511e-13
## 336 -7.441034      2.1554e-13    2.3601e-10
## 596 -6.361993      3.0313e-10    3.3193e-07
## 633 -5.278166      1.6015e-07    1.7536e-04
```

### IV. Final models

Model parameters:

Finally, our final model without outliers is the one we built reffecting some features, regrouping some categories for factors and retrieving repetitive 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 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 surprising. 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.

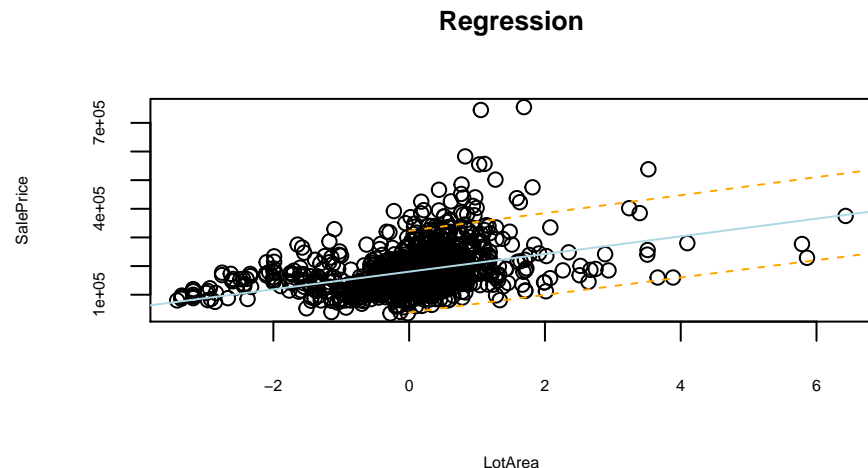
```
## [1] 29260.96
```

```
## [1] 32157.83
```

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 explaining 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.