Multivariate Analysis

House price prediction

Student:
Patrick Schneider

Date:26.06.2018

Table of Content

1.Introduction

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. This competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing almost every aspect of residential homes in Ames, Iowa, this competition challenges to predict the final price of each home.

General research conclusion about house pricing:

- **Location** - location is key for high valuations, therefore having a safe, well facilitated and well-positioned house within a good neighborhood, is a large contributing factor.
- **Size** - The more space, rooms, and land that the house contains, the higher the valuation.
- **Features** - the latest utilities and extras (such as a garage) are highly desirable.

This insight built the starting point for this project and built the special focus on this project.

2.Objectives

The objective of this project was to find the best prediction algorithm for house prices.
To get to this point, the milestones were set as follow:

1. Data analysis
    a. General exploration of the predictors and target variable
    b. Principal component analysis
2. Data preprocessing & Feature engineering
    a. Analysis and handling of missing values
    b. Analysis and handling of outliers
    c. Analysis and handling of Multicollinear
    d. Analysis and handling of near zero variance predictors
    e. Hot encoding of categorical variables
3. Prediction
    a. Evaluation of different prediction algorithm
    b. Tuning of the algorithm
    c. Cross-validation

4. Conclusion

3.Data dictionary

Data fields

This is a brief overview of the data description file. The full description of the levels can be found in appendix 1:

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet

- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

4.Exploration of data

a) Univariate analysis

The following analysis will be split into numerical and categorical variables.

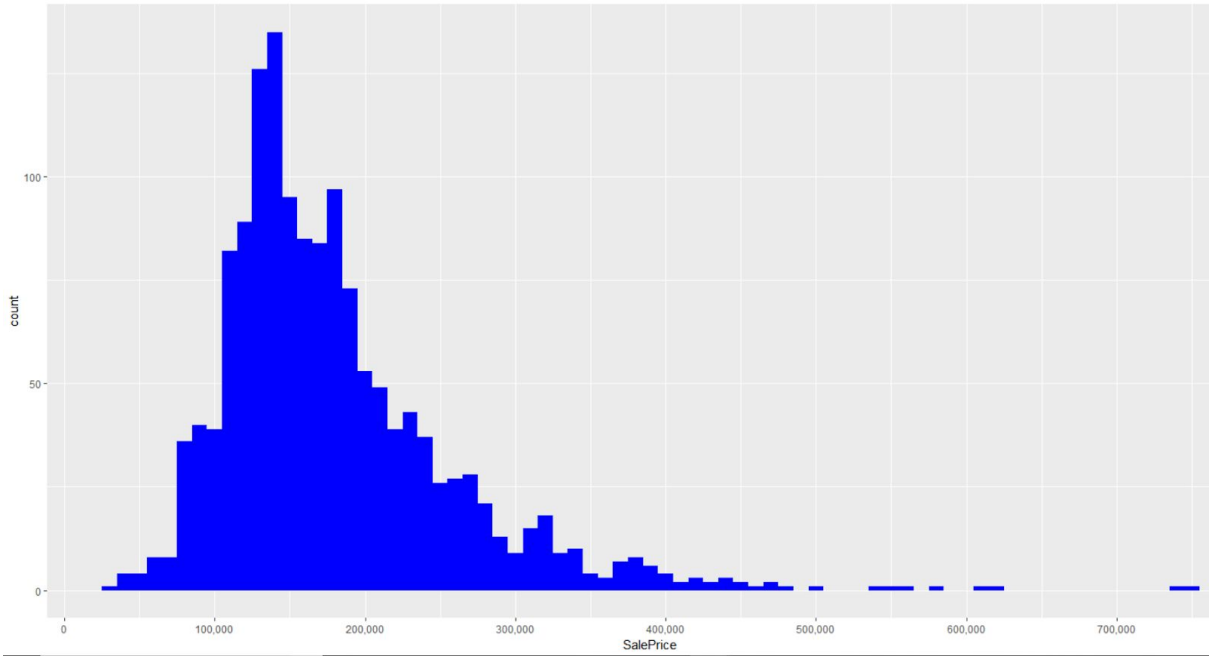
Numerical

```
> numericVarNames
 [1] "MSSubClass"    "LotFrontage"   "LotArea"       "OverallQual"   "OverallCond"   "YearBuilt"     "YearRemodAdd"  "MasVnrArea"
 [9] "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"   "X1stFlrSF"     "X2ndFlrSF"     "LowQualFinSF"  "GrLivArea"
[17] "BsmtFullBath"  "BsmtHalfBath"  "FullBath"      "HalfBath"      "BedroomAbvGr"  "KitchenAbvGr"  "TotRmsAbvGrd"  "Fireplaces"
[25] "GarageYrBlt"   "GarageCars"    "GarageArea"    "WoodDeckSF"    "OpenPorchSF"   "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"
[33] "PoolArea"      "MiscVal"       "MoSold"        "YrSold"        "SalePrice"
```

Illustration of all numerical variables.

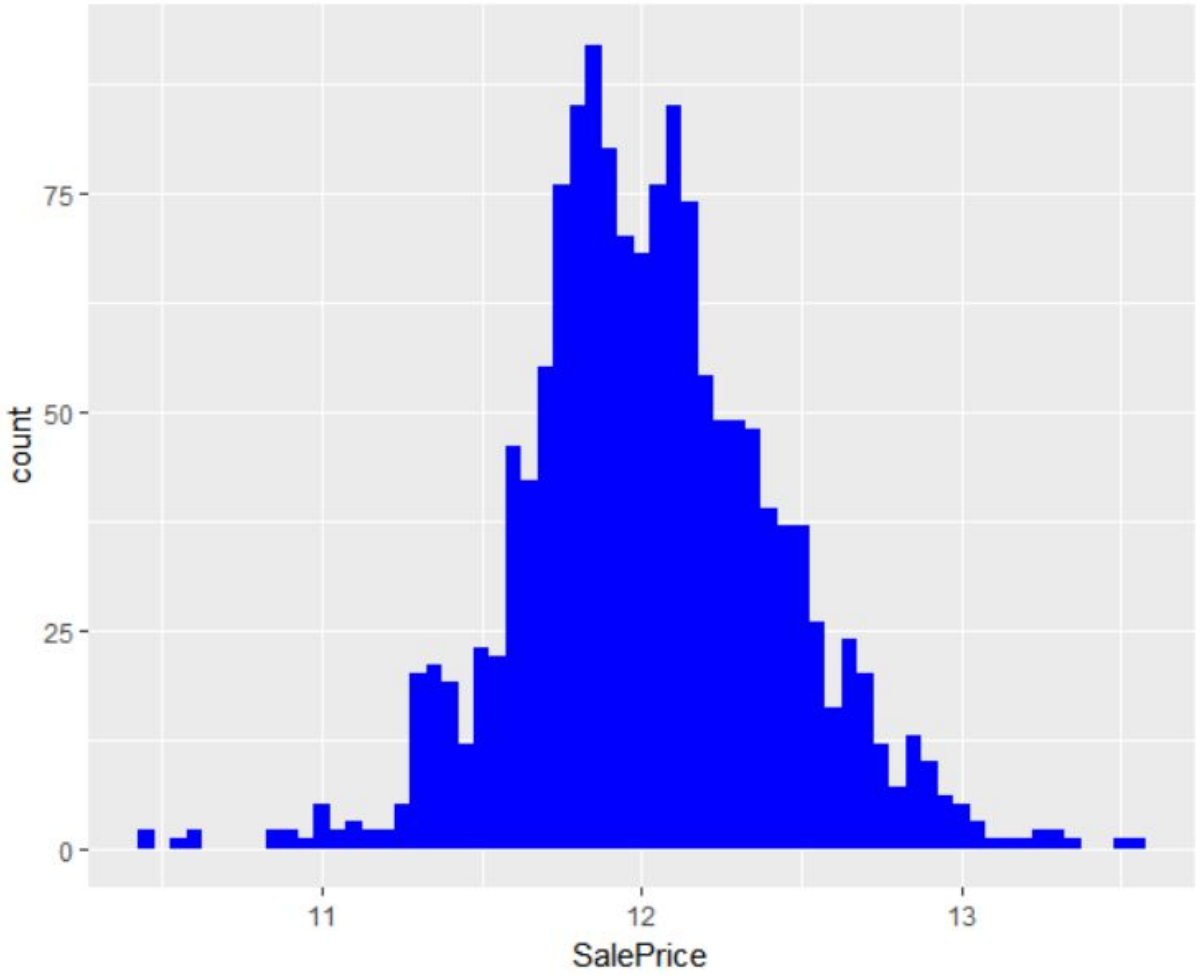
Response variable - SalePrice

In the first analysis contained insights into the sales price distribution, where a right-skewed distribution was detected. This plot can be explained by the fact, that not many people can offer an expensive house, where they offer demand leads to this distribution.




```
summary(alldata$SalePrice)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
34900  129975  163000  180921  214000  755000    1459
```

We see that the distribution is skewed towards cheaper homes, with a relatively long tail at high prices. Applying a logarithmic transformation makes the distribution symmetric.



Besides making the distribution more symmetric, working with the log of the sale price ensures that the relative error for cheaper and more expensive homes are treated on an equal footing. The Log SalePrice will be our true target variable.

In appendix 2, the other variables can be found with their distribution (also it can be found in the script analysis-processing.r by executing the code in chronological order).

Categorical

```
> catVars
 [1] "MSZoning"      "Street"        "Alley"         "LotShape"      "LandContour"   "LotConfig"
 [7] "LandSlope"     "Neighborhood"  "Condition1"    "Condition2"    "BldgType"      "HouseStyle"
[13] "RoofStyle"     "RoofMatl"      "Exterior1st"   "Exterior2nd"   "MasVnrType"    "ExterQual"
[19] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"      "BsmtExposure"  "BsmtFinType1"
[25] "BsmtFinType2"  "Heating"       "HeatingQC"     "CentralAir"    "Electrical"    "KitchenQual"
[31] "Functional"    "FireplaceQu"   "GarageType"    "GarageFinish"  "GarageQual"    "GarageCond"
[37] "PavedDrive"    "Fence"         "MiscFeature"   "SaleType"      "SaleCondition"
```

For categorical variables, bar charts and frequency counts are the way of illustration. The overview of the categorical distributions can be found in Appendix 2, or also in the script analysis-processing.r by executing the code in chronological order.

b) Bi variate analysis

Having looked at some variables individually, the relationships and correlation can be analyzed. The most interesting insight will be the relationship between the target variable (sale price) and the prediction features.

numerical (37 variables)

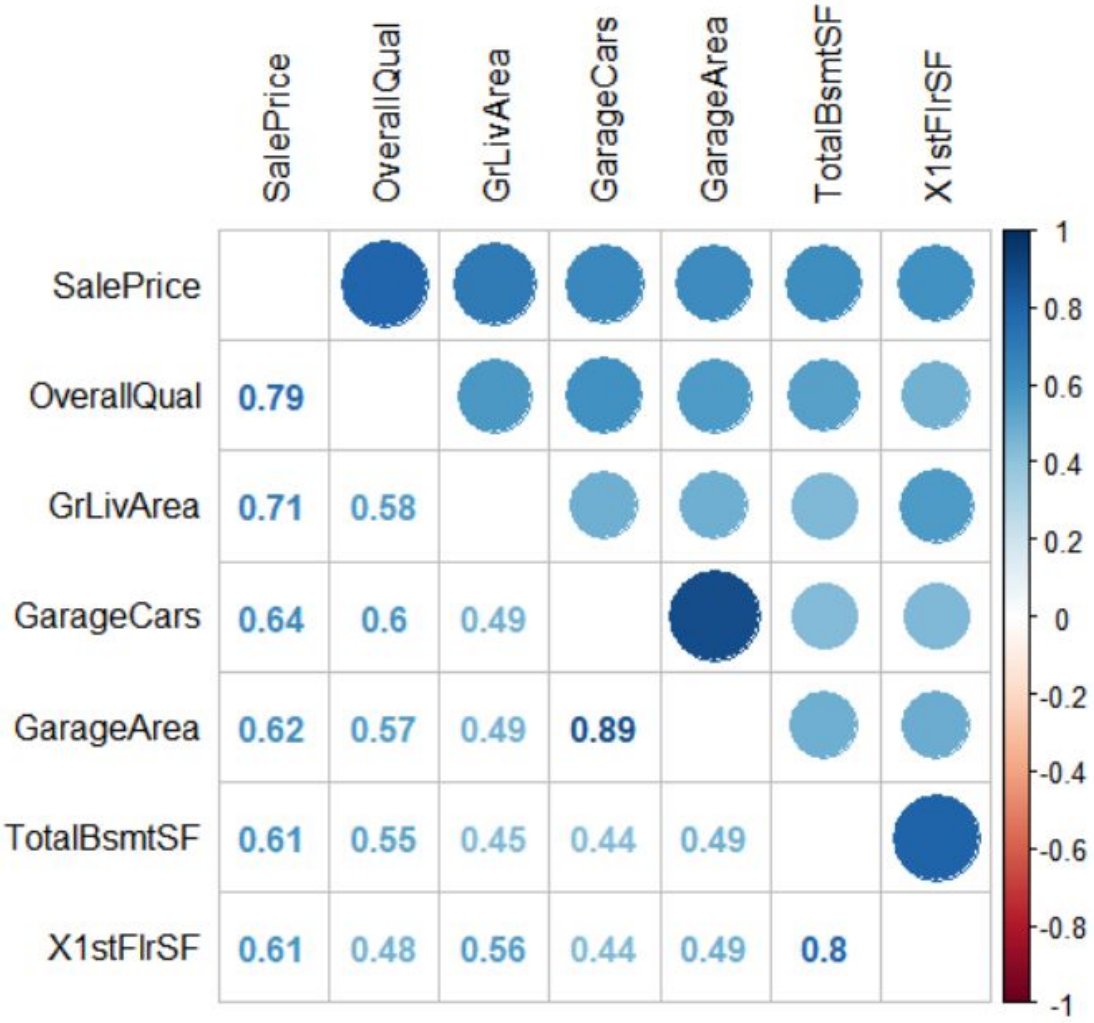
To determine which features have the strongest relationship with SalePrice we can compute the sample correlation coefficient between the 2 variables,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])$$

is the sample covariance and sx, sy are the sample standard deviations. The correlation coefficient measures how linearly the 2 variables are related. A coefficient of 0 means that the 2 variables show no linear relationship, a coefficient between (0,1) shows that they have positive relationship and a coefficient between (-1,0) means they have a negative relationship. Here we are interested in variables that show strong relationship with SalePrice so we only retain the feature with a coefficient > .6, that can be seen in the following illustration:



6 variables were found that are highly correlated with the target variable.

The ranking of the features to the target variables:

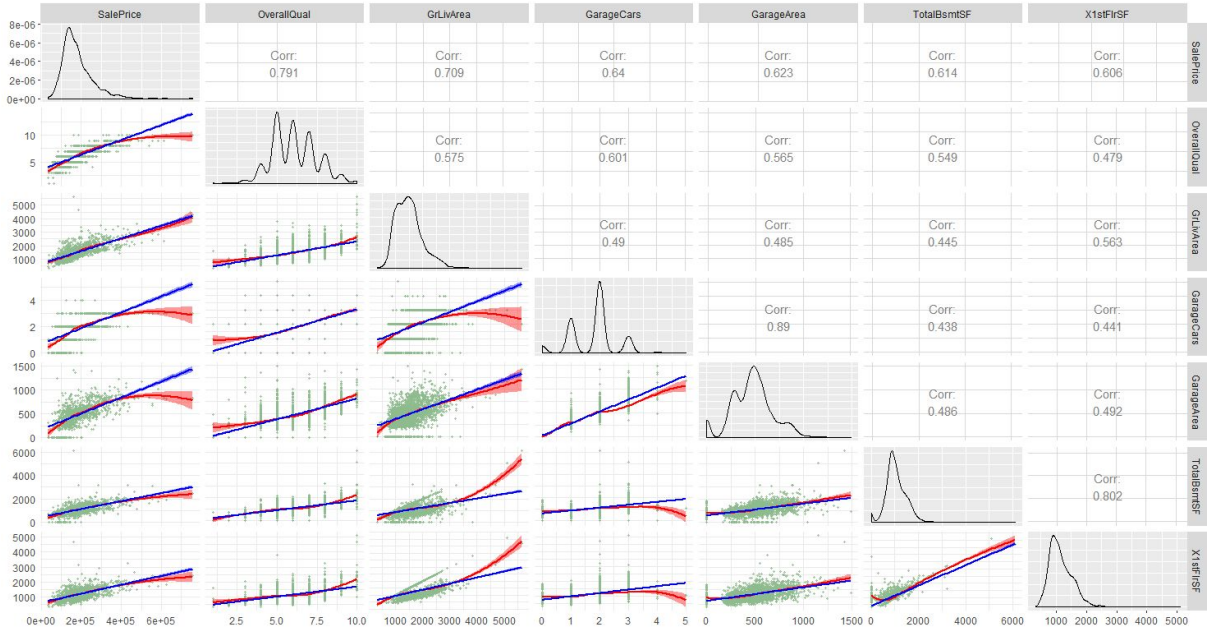
- 1. OveralQuality: Is the overall quality of the house. This has the highest correlation to the sales price.
- 2. GrLivArea: Here it was needed to research the definition of the meaning. The definition of the living Area is the area that is not underground and has access to the sunlight.
- 3. GarageCars: Size of garage in car capacity
- 4. TotalBsmtSF: Total square feet of basement area
- 5. TotalBsmtSF: Total square feet of basement area
- 6. x1stFlrSF: First Floor square feet

Another important step is to study the relationships among features.

Multicollinearity between each other:

- GarageCars and GarageArea: We can see that the correlation between the 2 is extremely high, 0.88, which makes sense as the area of the garage is a constraint on how many cars can fit.
- The TotalBsmtSF and 1stFlrSF also make intuitive sense as anything corresponding to area/size of the house will have an effect on the price. We can also see that these 2 features have a strong linear relationship with one another, which makes sense as the size of the basement can certainly depend on the size of the houses 1st floor.

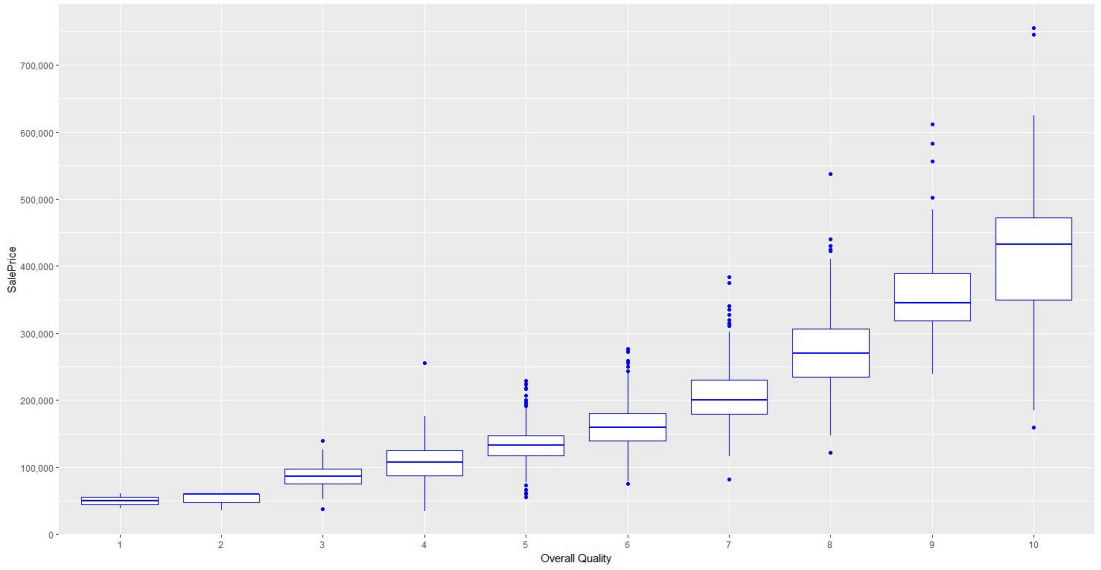
The following scatterplot shows the correlation with a linear and polynomial regression to give an overview of the distribution of the features to each other:

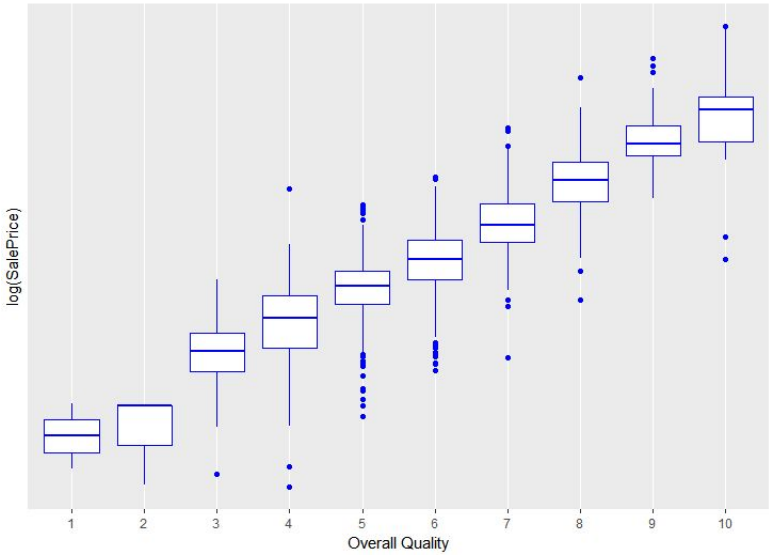


The blue lines in the scatter plots represent a simple linear regression fit while the red lines represent a local polynomial fit. OverallQual and GrLivArea and TotalBsmtSF follow a linear model, but have some outliers we need to check later on. For instance, there are multiple houses with an overall quality of 10, but have suspisciously low prices. We can see similar behavior in GrLivArea and TotalBsmtSF. GarageCars and GarageArea both follow more of a quadratic function. It seems that having a 4 car garage does not result in a higher house price and same with an extremely large area. The remaining feature 1stFlrSF follows a linear model.

Analysis of 1. OverallQuality:

In the next step we observe the overall quality distribution on SalePrice in a boxplot. The plot can be seen below:

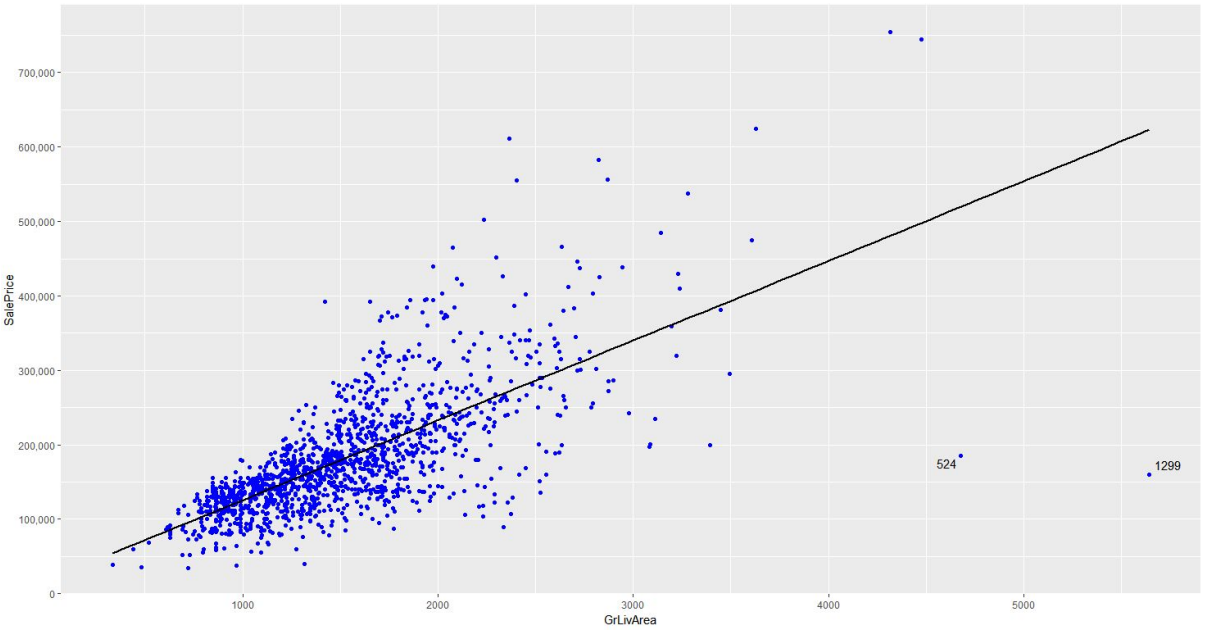




It can be observed that the price of the house is close aligned with quality level in an upwards curve. Single outliers are detected, while it can be said that there are no extreme cases. The quality category 4 and 8 should still be checked for outliers.

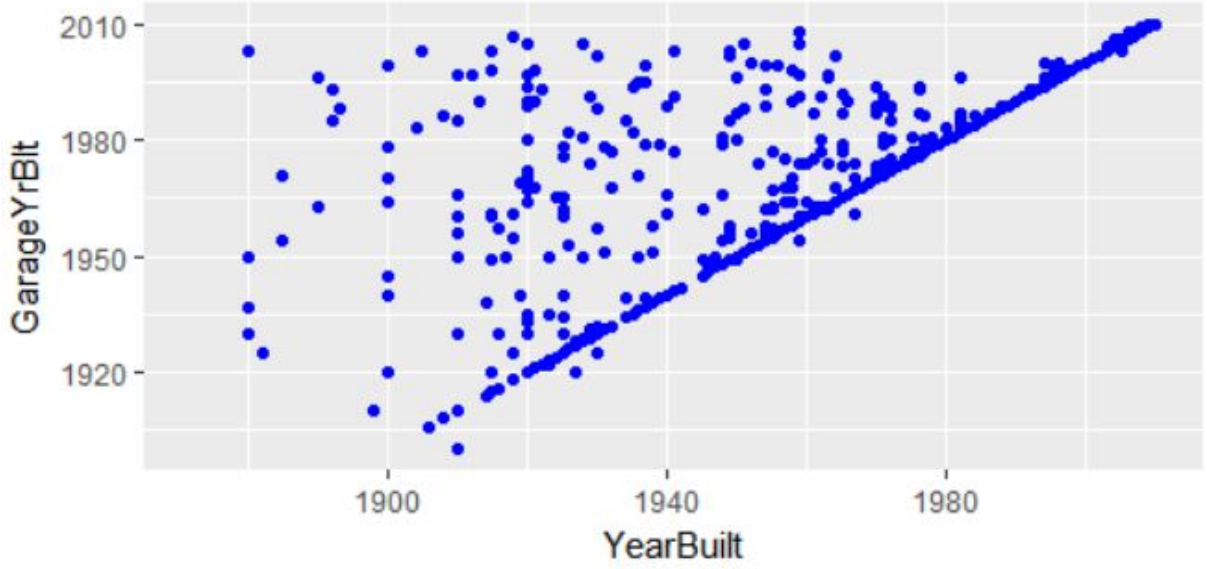
Analysis of 2. GrLivingArea:

Houses that offer a bigger living area with sunlight, are generally expensive.
In the following regression plot we can see the distribution of the points to a regression line:



While there is clearly a trend of sale price increasing with area, we see that there are two points that don't seem to fit in with the rest. Towards the lower right part of the plot, there are two very large houses (bigger than 4500 sqft) with unusually low sale prices. Left untreated, those points can have a huge impact on the accuracy of the later model. In the simplest case, if we have a good reason to believe that the outliers represent false values or mistakes in the data, we can simply remove them.

Analysis 2: Feature correlation YearBuilt and GarageYrBlt

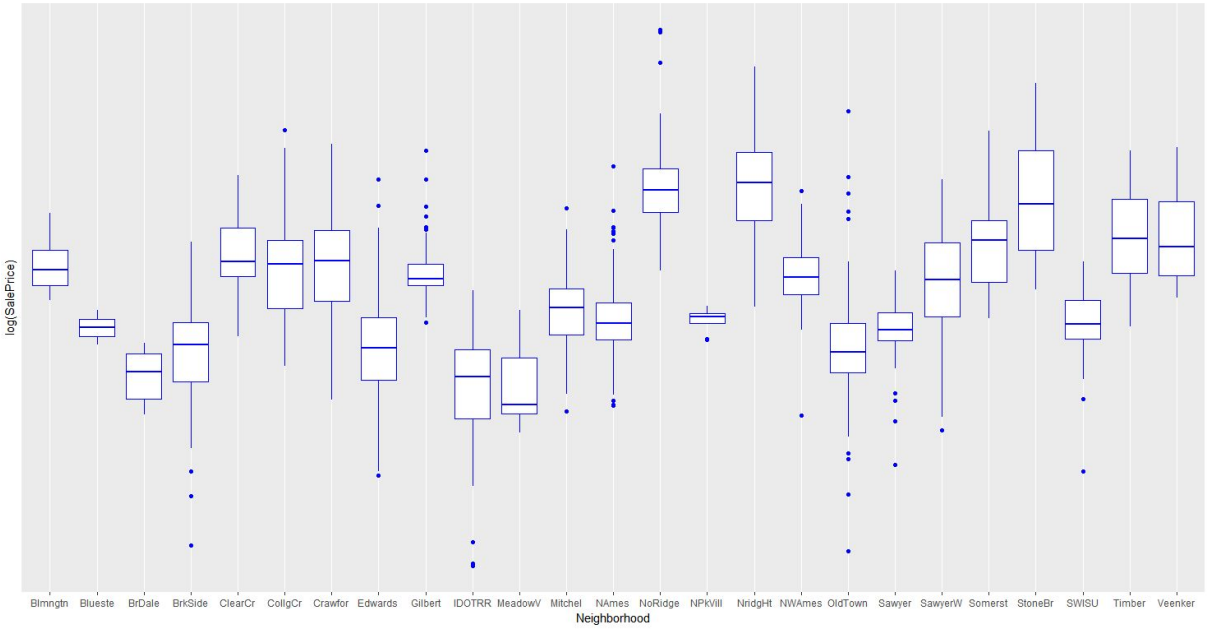


As we might expect, the figure tells us that the majority of garages were built at the same time as the houses they belong to (diagonal line). A significant number were also added later (points above the line). I considered creating a new feature that tells us whether or not a garage was originally constructed with the house or how many years later on the garage was added.

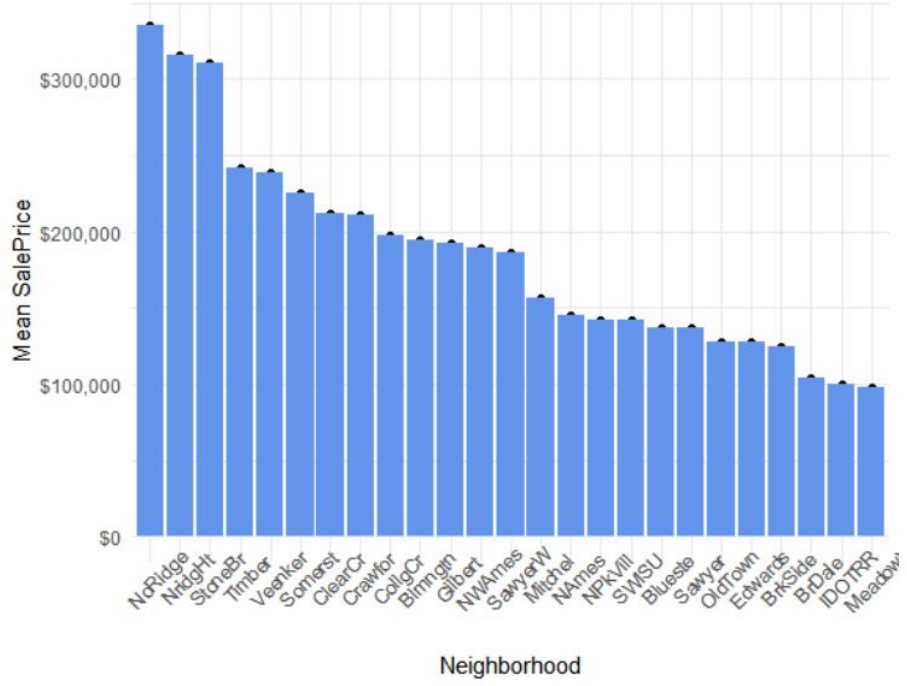
We have garages that were built 20 years earlier than their houses (the points below the diagonal line).

Categorical

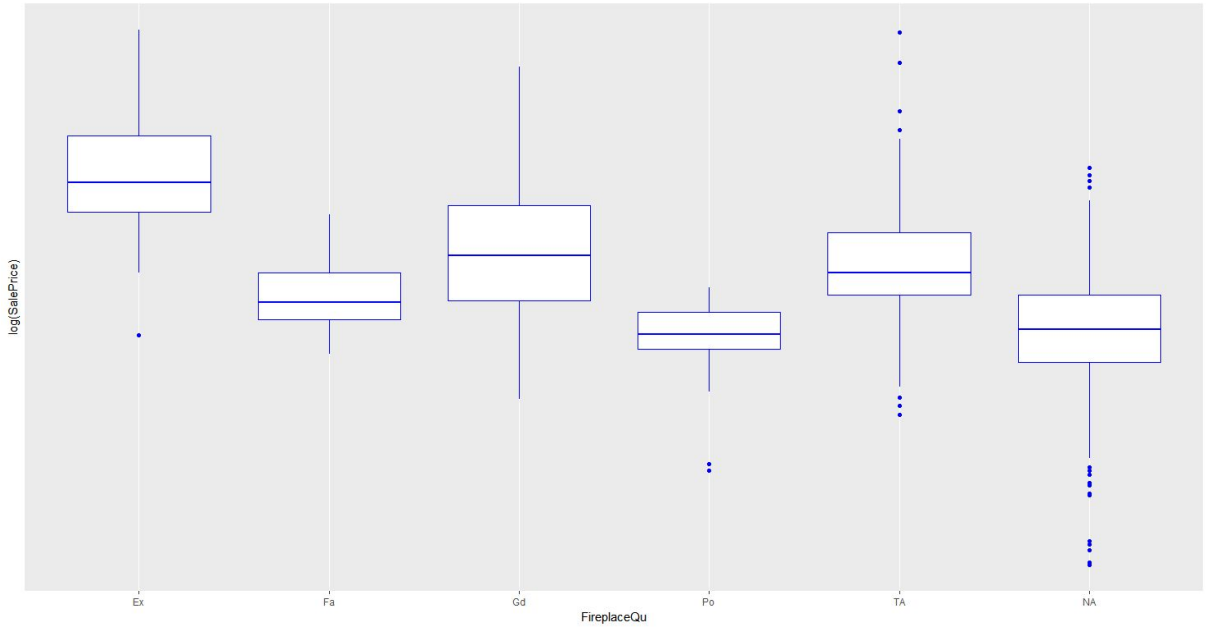
In this section I concentrated on the categorical values that have a high correlation, as well as follow the pre assumption of the general research done about house pricing. In the following we analyse the value by the neighborhood a house is located:



As expect, there is considerable variation in price between neighbourhoods.

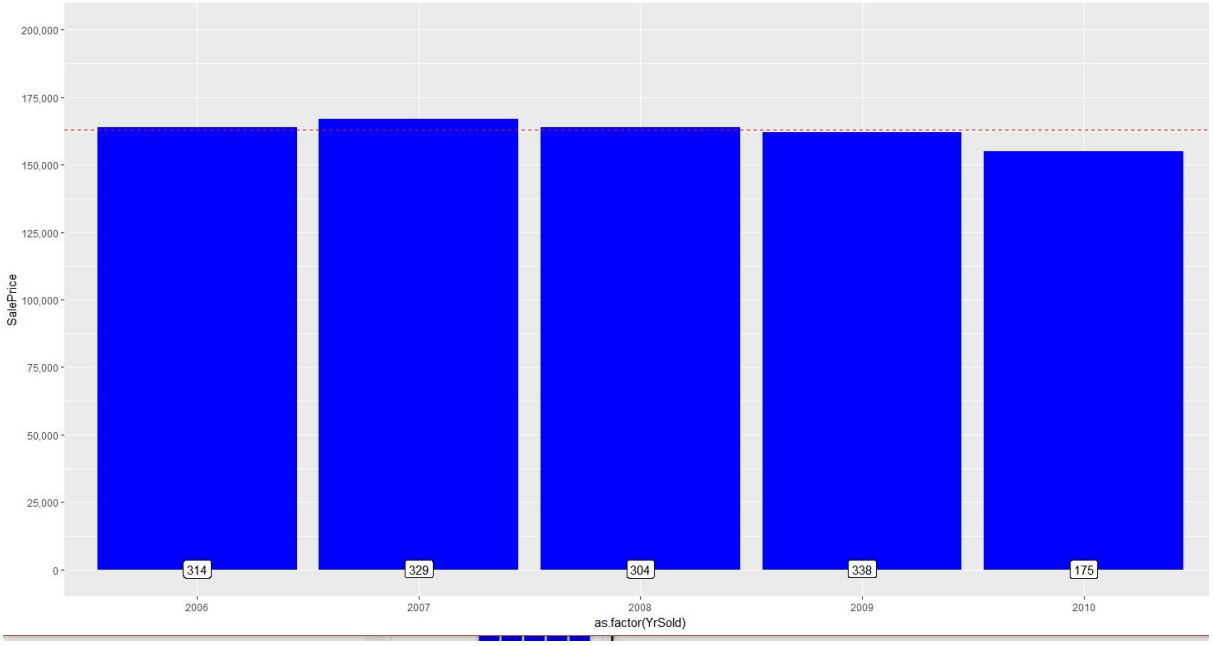


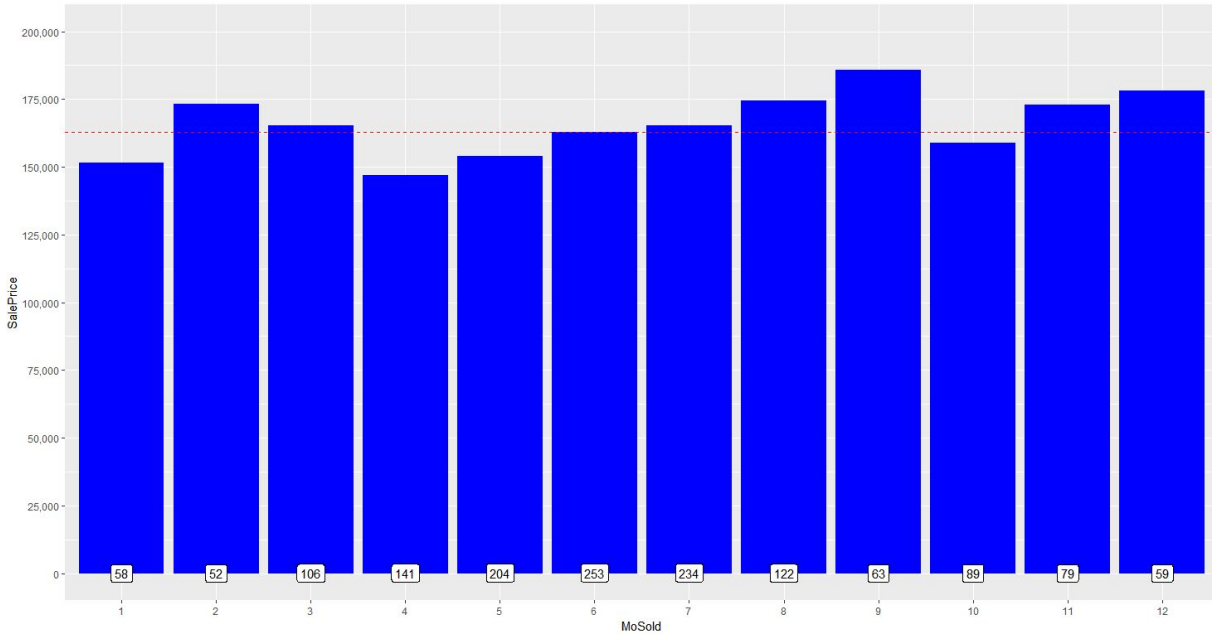
Having a fireplace and the quality of the fireplace was analysed in a second step:



Here we can see that the quality has indeed an impact on the sales price.

In the third step I took a look on the year and month sold of the house.





We can clearly see that the year and especially the month has a huge factor on the target variable. This should be taken into consideration for later feature engineering.

Further researches were done, but left out in this report to not extend the size to much. The r script contains the illustration of the other values.

5.Preprocessing

a) Missing Data analysis

Since the total number of entries in the train and test set is 2919, we can see that for some features nearly all entries are missing, while for others it is just one or two. How we proceed to treat these missing values depends very much on the reasons the data is missing, the problem and the type of model we want to use. The decision fell on imputing most values.

Completeness of the data

The dataset contains 35 variables with missing values. The ranked order of variables by missing data can be seen in the following overview:

PoolQC	MiscFeature	Alley	Fence	
2909	2814	2721	2348	
FireplaceQu	LotFrontage	GarageYrBlt	GarageFinish	GarageQual
1420	486	159	159	159
GarageCond	GarageType	BsmtCond	BsmtExposure	BsmtQual
159	157	82	82	81
BsmtFinType2	BsmtFinType1	MasVnrType	MasVnrArea	MSZoning
80	79	24	23	4
Utilities	BsmtFullBath	BsmtHalfBath	Functional	Exterior1st
2	2	2	2	1
Exterior2nd	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF
1	1	1	1	1
Electrical	KitchenQual	GarageCars	GarageArea	SaleType
1	1	1	1	1

Missing for a reason

For example, missing values for garage, pool or basement-related features simply imply that the house does not have a garage, pool or basement respectively. In this case, I decided to capture these missing values with a proper information level. For categorical features, we can replace missing values in such cases with a new value called ‘None’. For most numerical features it makes sense to replace the missing values with zero. It follows the overview of the structured features and the action and reasoning about our action:

Variable	**Action & Reasoning**
PoolQC (2909 NA)	N/A entries most likely represent the non existence of a pool. That's why we impute for every N/A value a “None” level in PoolQC. After the imputation of the PoolQC, we checked if there is a existing PoolArea with “None” under PoolQC. For three existing cases we imputed the poolQC based on their overalQuality scaled on the PoolQC levels.
MiscFeature (2814 NA)	Here the frequencies are very low for the different levels, which are very specific like “having a tennis court”(1 entry). Here we just set the NA values to “None”.
Alley (2721 NA)	Alley has the level of Grvl and Pave. The 2721 entries most likely represent that there is no alley access. That’s why I decided to impute “none” as the value for alley
Fence(2348 NA)	NA values most likely represent the non existence of a fence, that's why I imputed “none” for those values. Side note: The values do not seem to be ordinal (no fence is best).

FirePlace(1420 NA)	FireplaceQuality N/A matches the houses with noFireplace. ->imputation of “None”
LotFrontage(486 NA)	For the linear feet of street connected to property a proper imputation can be the median per neighborhood.
Garage variables(159 NA) c('GarageYrBlt','GarageArea', 'GarageCars', 'GarageQual', 'GarageFinish', 'GarageCond', 'GarageType')	The NA values of GarageYrBlt were imputed by the value of the year the house was built. Later we will take a closer look on those values in the feature engineering process.
Basement (Bsmt) variables(5 vars each ~80NA) c('BsmtCond','BsmtExposure', 'BsmtQual', 'BsmtFinType2', 'BsmtFinType1')	The 80 missing values occurred in all variables that were related to basement specifications. This leads to the conclusion that for those cases the basement was not existing. -> imputation of “none”.

Missing at random

The reasons for missing values are not clear for the following features and having no further information, I assume that they are missing at random. In this case, there are three main options: delete, impute or leave.
The following table shows the reasoning for every variable imputation:

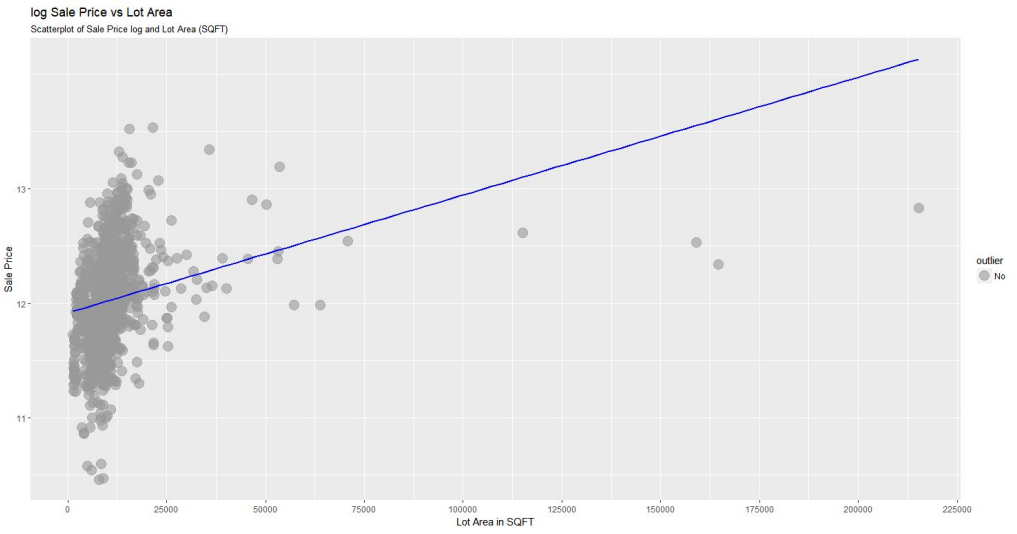
Variable	**Action & Reasoning**
Masonry variables(24 NA) c('MasVnrType','MasVnrArea')	If a house did not have a masonry area, it also had noMasVnrType. Here we imputed the value “0” for the area in square meter. In one case I found that one individual value was missing the MasVnrType. Here I imputed the mode.
MSZoning (4 NA)	MSZoning is a categorical variable about the zoning of the house. Here I imputed the value via random forest.
Kitchen variables (1 NA)	Set the most common value for the ordinal quality level

Utilities(2 NA)	For utilities I noticed that the whole data set (train and test) have 2916 entries for “allPub” and 1 for “NoSeWa”. This makes this column quite useless for the prediction and that's why I decided to not deal with the NA values and take the variable early on out. Further removal of other variables will follow later.
Home Functional (1 NA)	Home functionality is a ordinal variable, where I decided to impute the value via random forest.
Exterior variables(2 NA)	Categorical variables where I decided to impute the mode of the neighborhood.
Electrical system (1 NA)	Categorical variable where I decided to impute the value via random forest.
Sale Type and Condition (1 NA)	Categorical variable where I decided to impute the value via random forest.

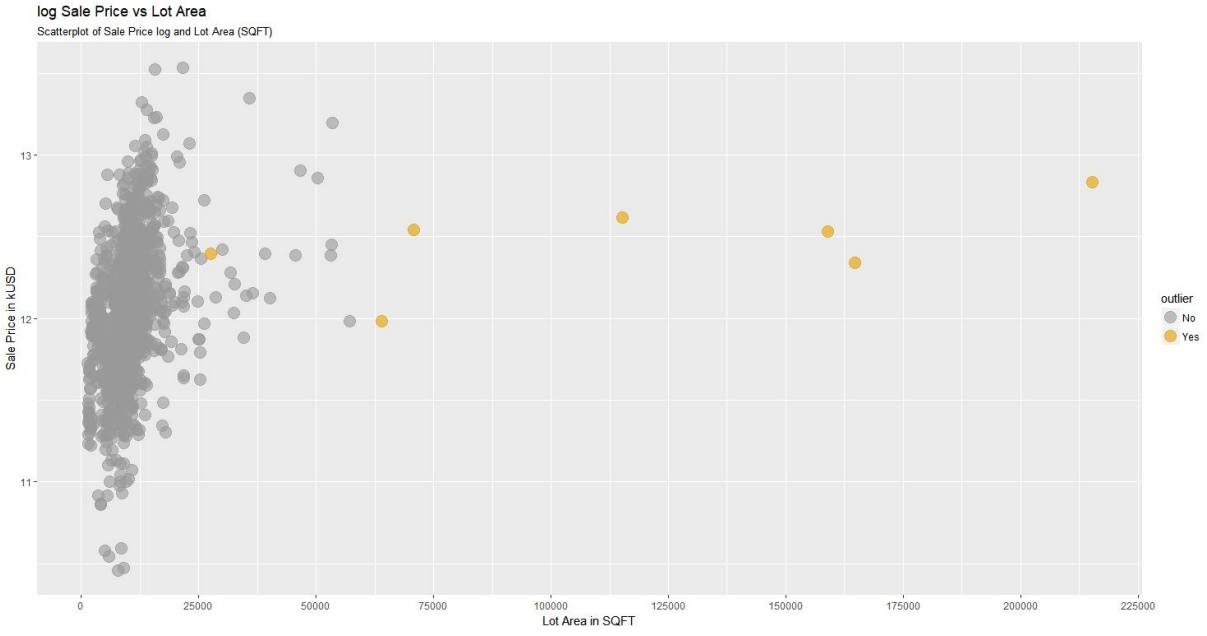
Furthermore, did I convert character integers into ordinal integers if a ordinality can be observed. For categories without ordinality, I decided to convert them into numerical factors.

b) Outliers detection

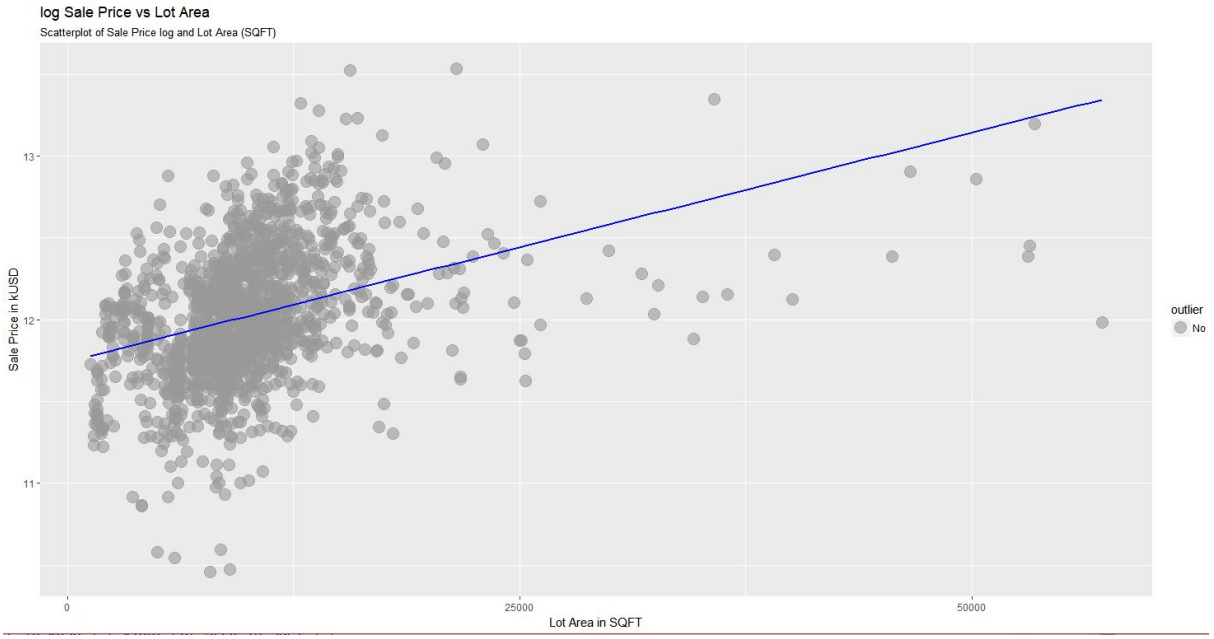
For the training data we can see 4 houses whose GrLivArea is greater than 4000 yet there is one in the testing set. Outliers are usually telling us something more about what is going on with the data. These houses in the training set are very large and ultimately do not add much value and are causing heavy skewness in both the SalePrice and GrLivArea and particularly the 2 values that have above a 4000 GrLivArea, but low SalePrice are putting a limits on the correlation between the 2 variables. Here I use the Mahalanobis distance to measure the similarity/dissimilarity for the detection of the outliers. The Mahalanobis distance is a measure of the distance between a point P and a distribution D. When plotting the data and adding the linear model regression line, it shows how strongly a few outliers distort the model.



With a threshold of 30 on the distance, 7 outliers were detected. On the following plot can be seen, that the biggest outliers are detected.



The analysis gives, by removing those outliers, the following resulting plot, were the linear function got optimized.



c) One - Hot encoding of categorical values

A way to ensured that all predictors are converted into numeric columns is by applying 'one-hot encoding' on the categorical variables. This means that all (non ordinal) factor values are getting a seperate columns with 1s and 0s (1 basically means Yes/Present). Dimension was increased from 78 features to 110.

Afterwards I removed levels with few or no observations in train or test set.

d) Normality

For linear regression models, we need to check for normality in any of the dependant variables. I used a Kolmogorov-Smirnov test and compute the skewness/kurtosis in each column to verify normality.

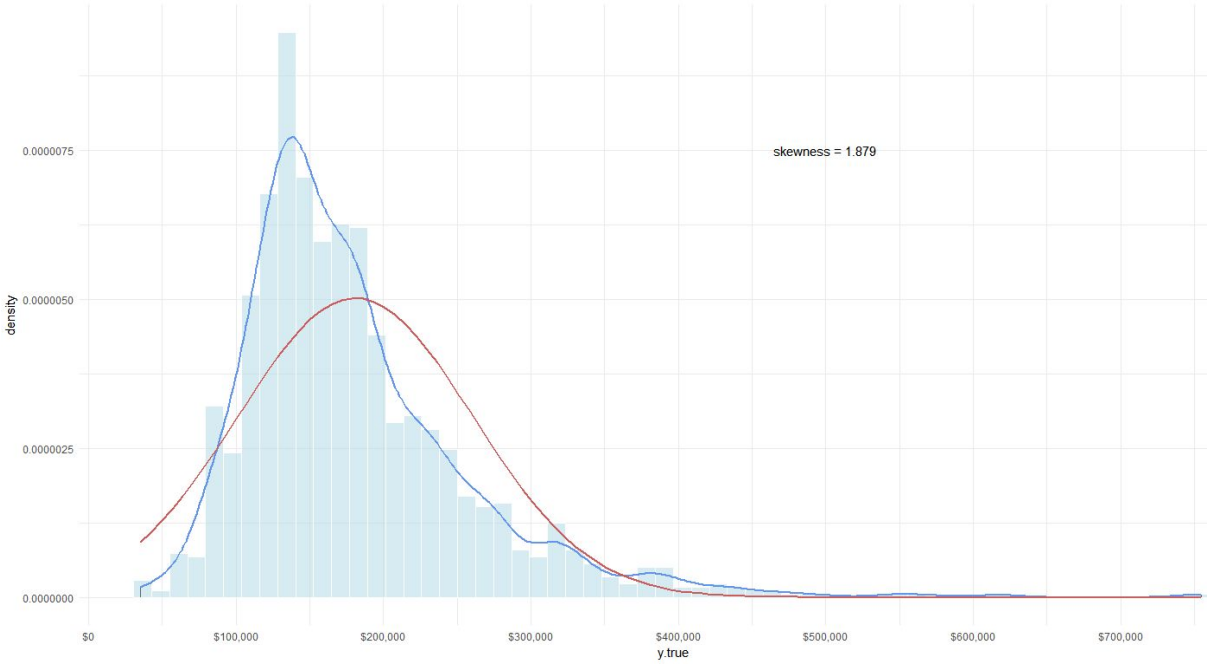
Kolmogorov-Smirnov test: Compares the sample distribution to a normal and returns a p-value determining if the 2 distributions are similar. Skewness is a measure of symmetry where distributions with 0 skew follow a normal shape. Kurtosis measures the taildness of the distribution. For skewnesses outside the range of -0.8 to 0.8 and kurtosises outside the of -3.0 to 3.0 do not satisfy the assumption of normality. For any features that are not normally distributed I

made a non-linear transformation with $f(x)=\log(x+1)$ when a column has 0 and $f(x)=\log(x)$ when there are no 0's in the column. I did this because $\log(0)=-\infty$.
I scaled all of the numeric data by standardizing the data, as we know our data has potential outliers we don't want to bound each column. To standardize the observations at individual columns, we compute

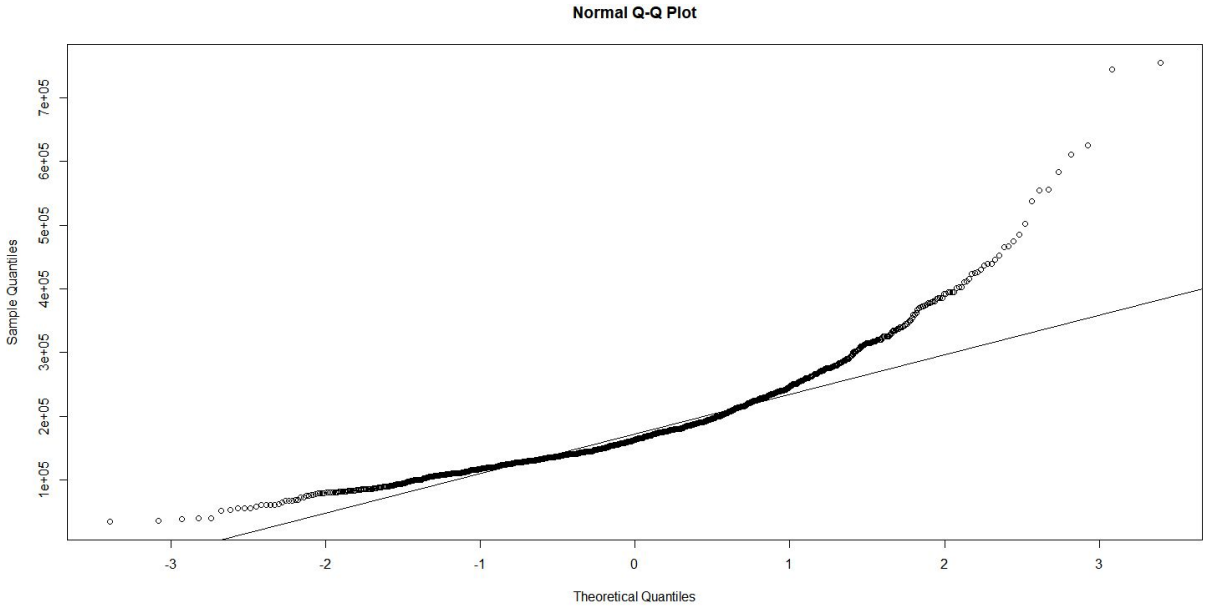
$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j}$$

where $\bar{x_j}$ is the sample mean at column *j* and s_j is the sample deviation at column *j*.

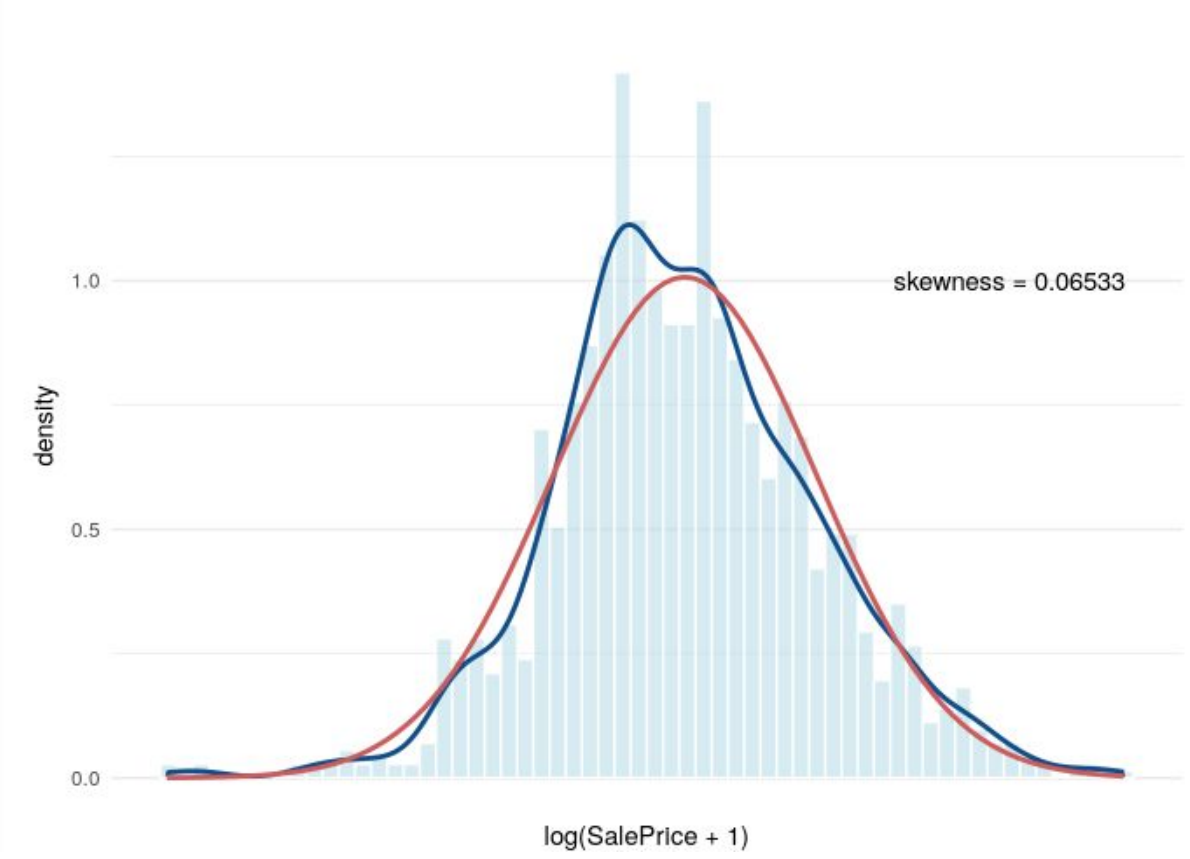
The sales price looks the follow:

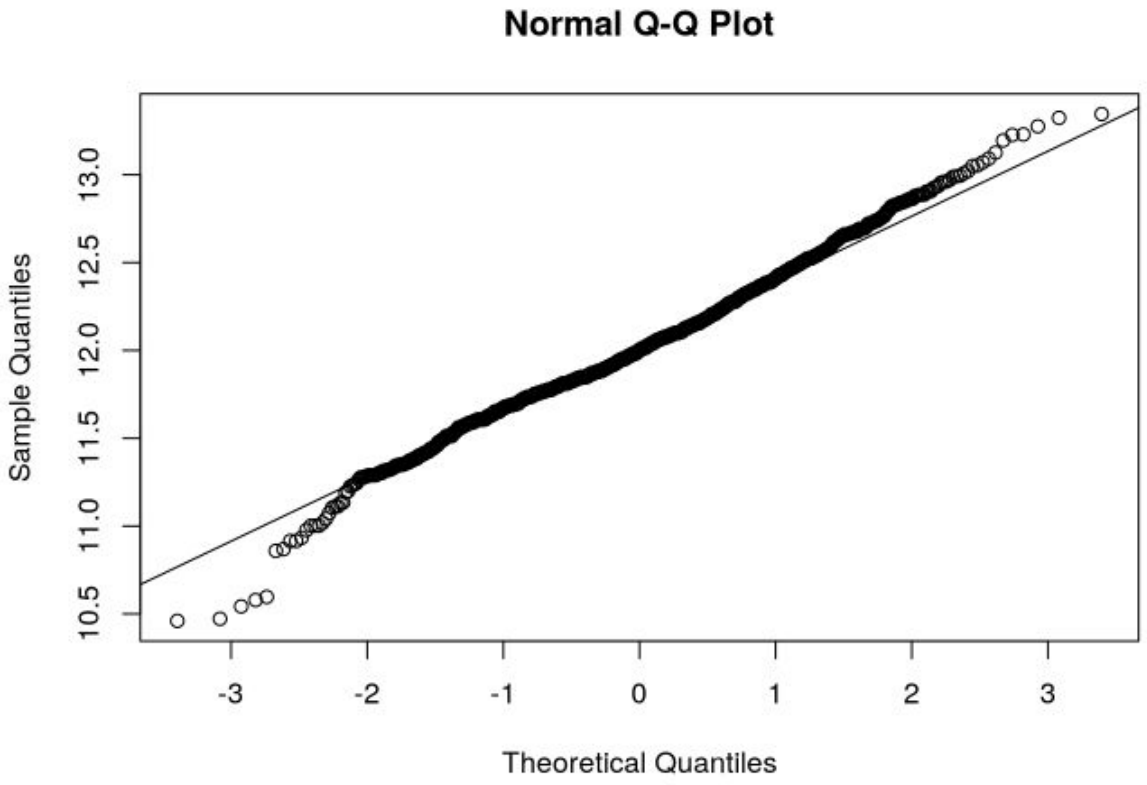


By checking the quantiles, we receive following plot:



We can see from the histogram and the quantile-quantile plot that the distribution of sale prices is right-skewed and does not follow a normal distribution. The early announced log transformation gave us following result:

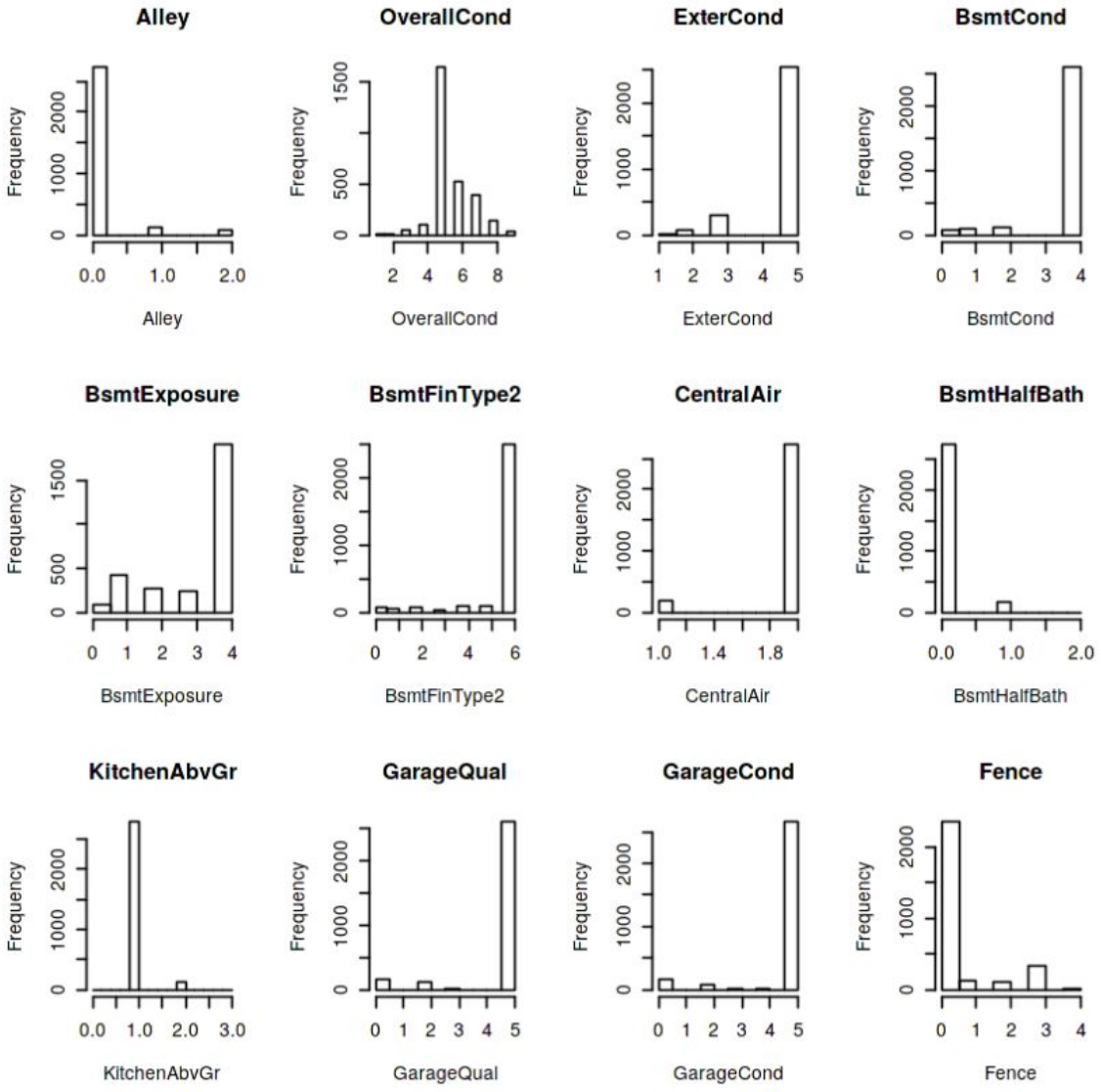




I reduced the skewness of 1.879 to 0.065.

e) Near Zero Variance

Now I have manipulated many variables and added many features in the dataset and their were some variables that won't give the data any value when modeling it. Some of these features have become zero-variance predictors, such that a few samples may have an insignificant influence on the model. These near-zero-variance may cause overfitting or will prevent the model from generalizing over the data. I checked the frequency of the most common value over the second most frequent values, which would be closer to 1 for well-behaved predictors and very large for highly-unbalanced features. It also checks the number of unique values divided by the number of samples which will approach zero (when the level of detail in the feature increases). I removed all of the near-zero-variance variables from our prediction set. In the following is an overview of the some of the features:

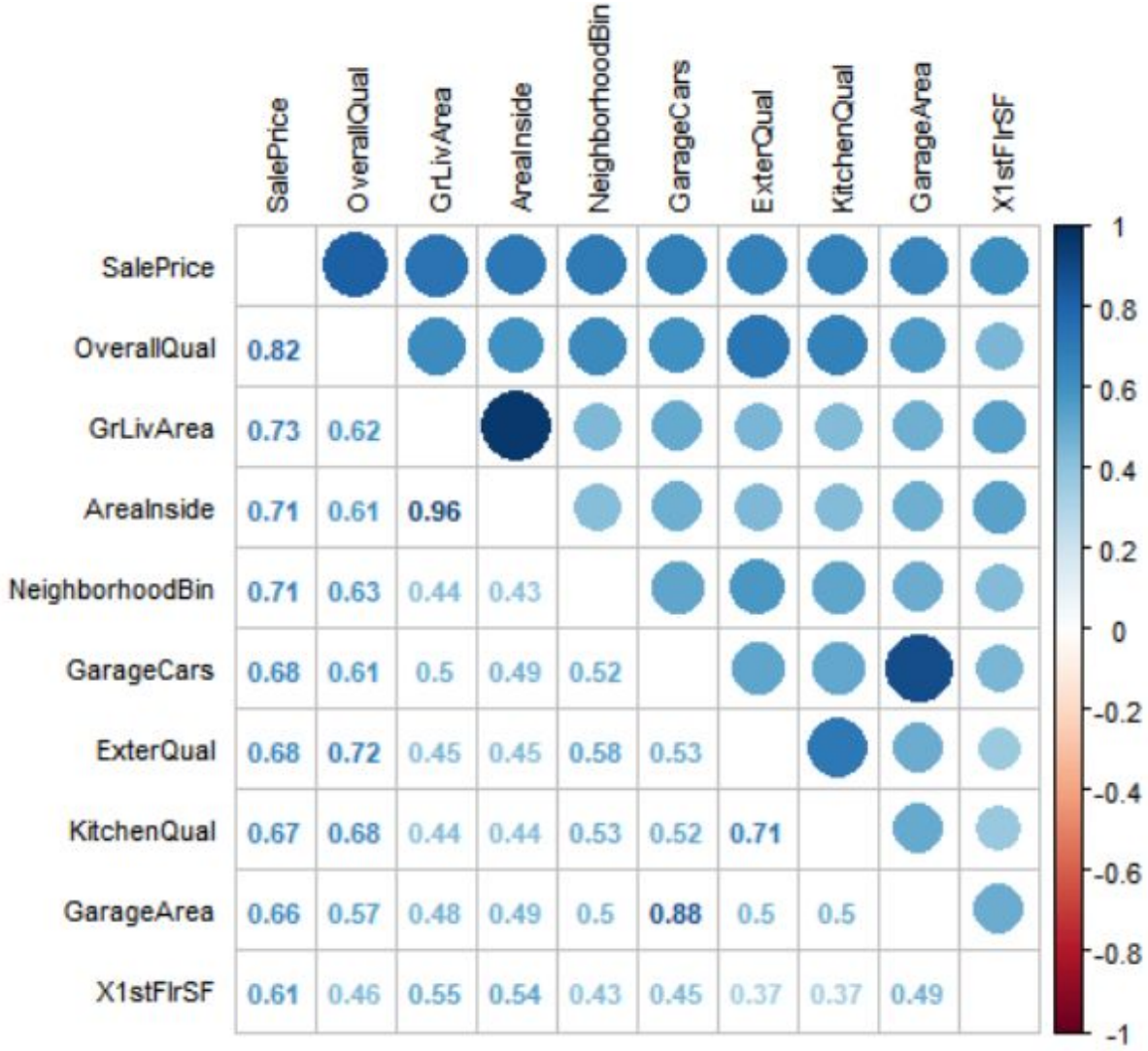


f) Reevaluation & Multicollinearity

If several predictor variables are highly linear correlated, the problem of multicollinearity occurs. Multicollinearity can lead to an overfitted model and imprecise parameter estimations.
I explored the relationship between the potential predictor variables and SalePrice again after the preprocessing. Here I will just include these variables in the model which show a clear correlation with the target variable. Further did I transform some variable to make the relationship linear.

I first dealt with the numerical variables by using the correlation coefficients by Pearson and Spearman. The Pearson-coefficient shows linear correlation while the Spearman-coefficient

shows monoton correlation. So if the Spearman-correlation is higher than the Pearson-correlation the variable has probably a nonlinear relationship with SalePrice and needs to be transformed. If the Spearman-correlation is low there can still be a non-monotonic relation. We will inspect all variable which have no linear correlation with SalePrice using scatterplots. The following scatterplot illustrated a correlation of >0.6 to the target variable:



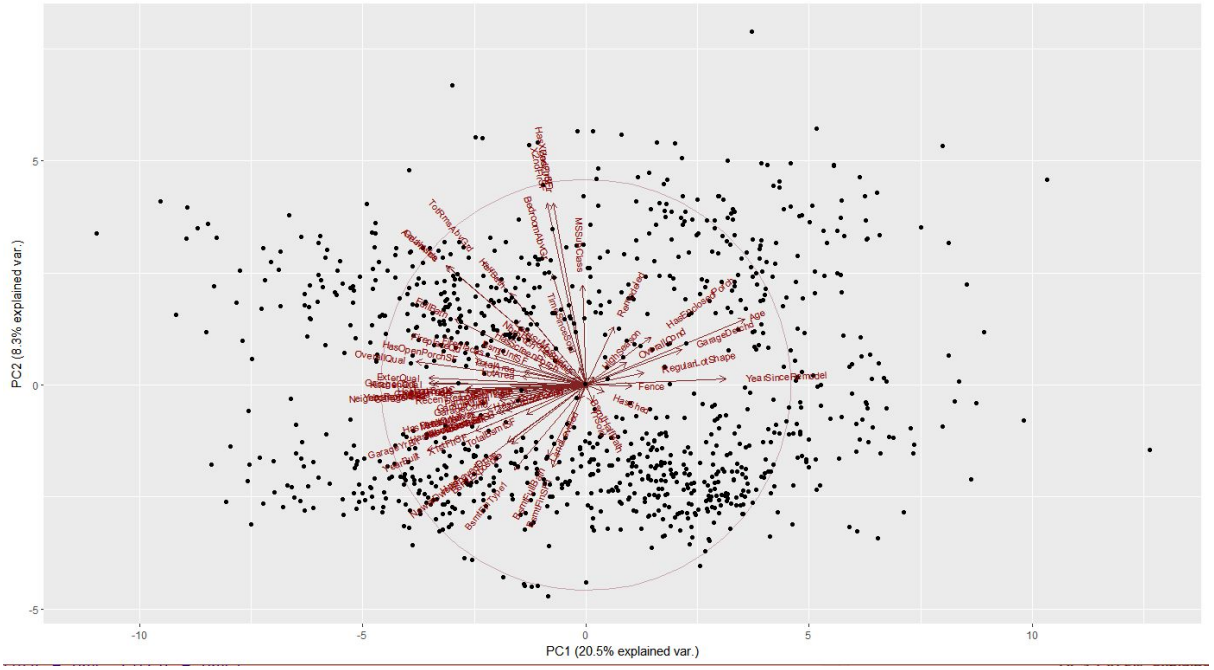
In an future step, the multicollinearity needs to be dealt with. Whether deleting one of the extremely high correlated predictors, or to engineer new features out of those.

6. Principal Component Analysis

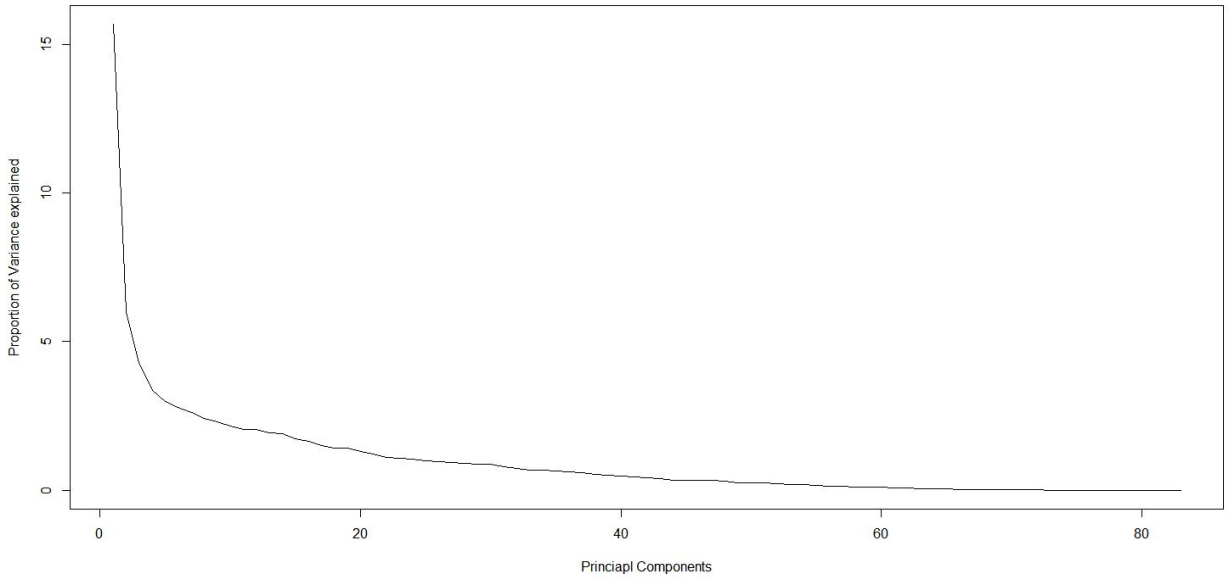
To get highly variance explanatory features, I used principal component analysis. R provides two PCA algorithms. One implements Singular Value Decomposition (SVD) and the other one implements Eigen Decomposition. For numerical accuracy, R recommends SVD, so we use this algorithm in this script.

One method is to retain only those Principal Components that have their eigenvalues higher than 1. An eigenvalue > 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. We retain all the PCs up till the cumulative variance explained is 85%.

After retaining the number of principal components, I analyzed the variables that have correlations higher than -0.50 or 0.50 with the corresponding individual Principal Components. The variables that have high correlations with the first Principal Components will be highly informative. The variables having high correlation with second Principal Component Analysis will be less informative than the variables we gain from first PC but more important than the variables we will get from third PC and so on.



The 1 PC explains 20.5%, 2 PC explains 8.3% of variance in the data and so on. Due to our big amount of features, it is quite messy to analyze the resulting plot with our eyes only. Later on follows a categorization of the variables to the principal components.



With our initial 83 variables, I found ~40 principal components explaining 85% of the variance.

PCA has helped to reduce 83 explanatory variables to 40 without compromising on to much variance.

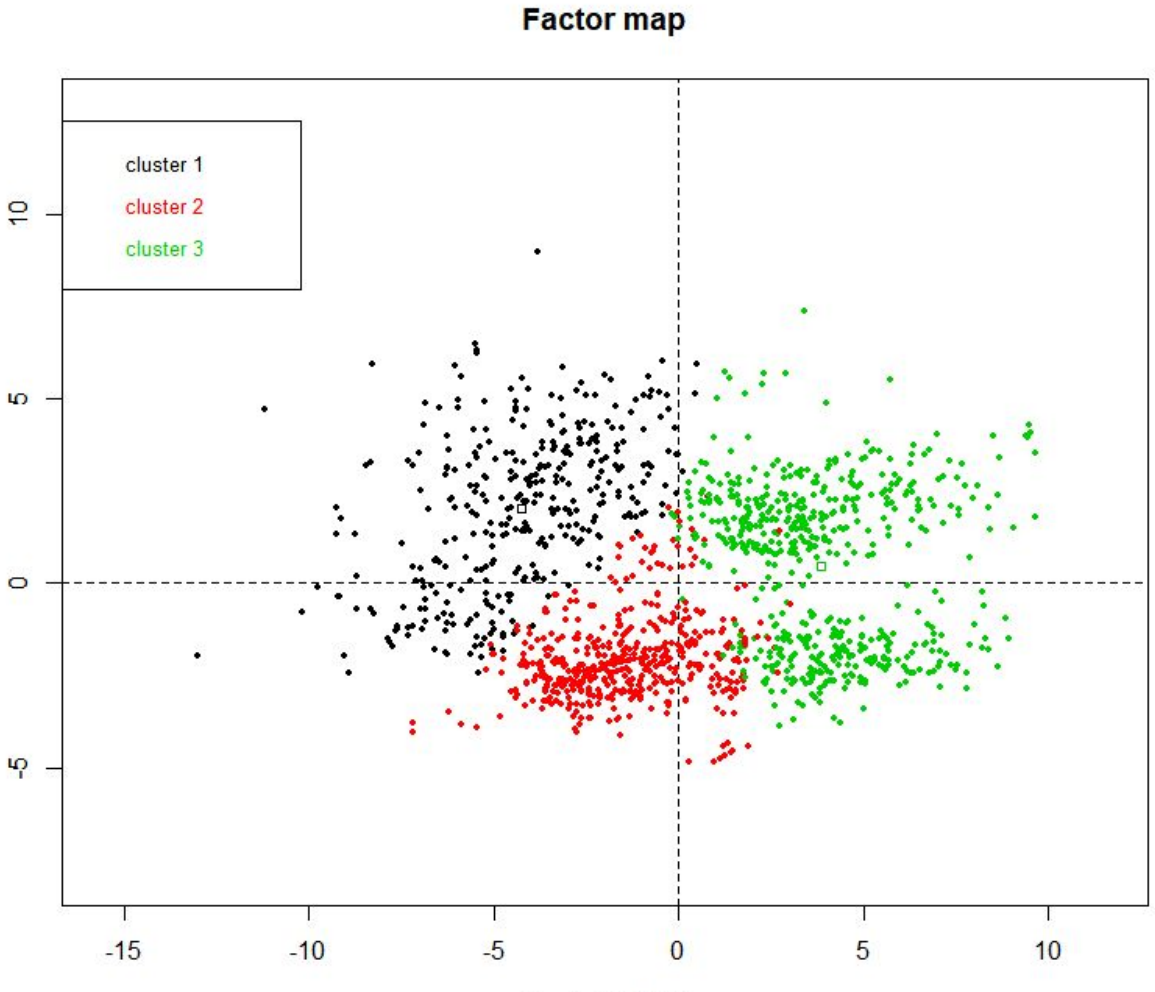
The following rank table shows us the significant features for the principal components between 1 and 9.

	coord.Dim.1	coord.Dim.2	coord.Dim.3	coord.Dim.4	coord.Dim.5	coord.Dim.6	coord.Dim.7	coord.Dim.8	coord.Dim.9
1	OverallQual	X2ndFlrSF	LotFrontage	WoodDeckSF	GarageQual	OverallCond	WoodDeckSF	YrSold	BsmtUnfSF
2	YearBuilt	GrLivArea	LotArea	HasWoodDeck	GarageCond		HasWoodDeck	TimeSinceSold	TotalBsmtSF
3	YearRemodAdd	BedroomAbvGr	X1stFlrSF	HasWoodDeckSF			HasWoodDeckSF		
4	X1stFlrSF	TotRmsAbvGrd	TotalArea						
5	GrLivArea	Has2ndFlr							
6	FullBath	HasX2ndFlrSF							
7	TotRmsAbvGrd	AreaInside							
8	GarageYrBlt								
9	GarageCars								
10	GarageArea								

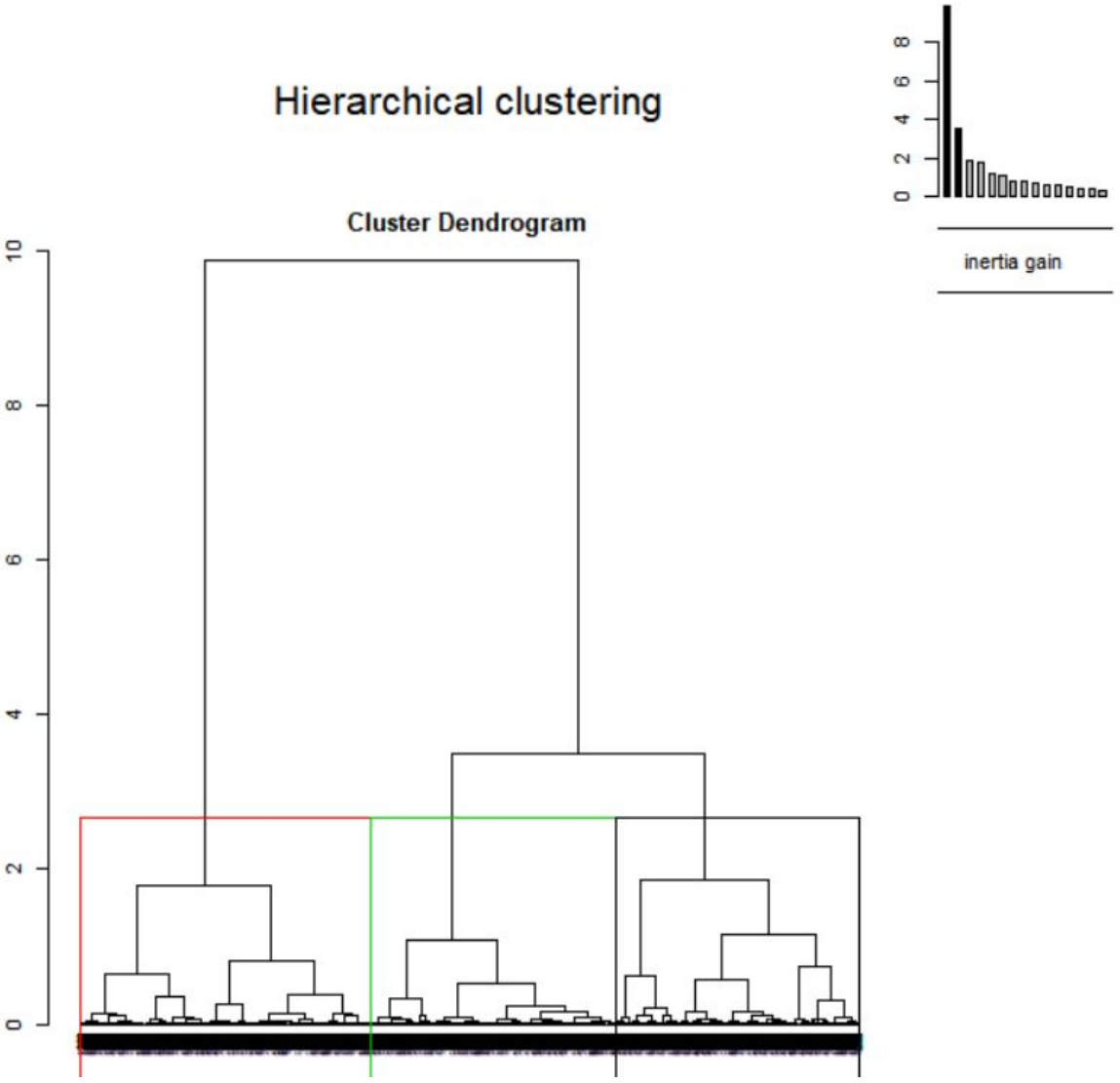
7. Hierarchical clustering

The principal components were used for the clustering.

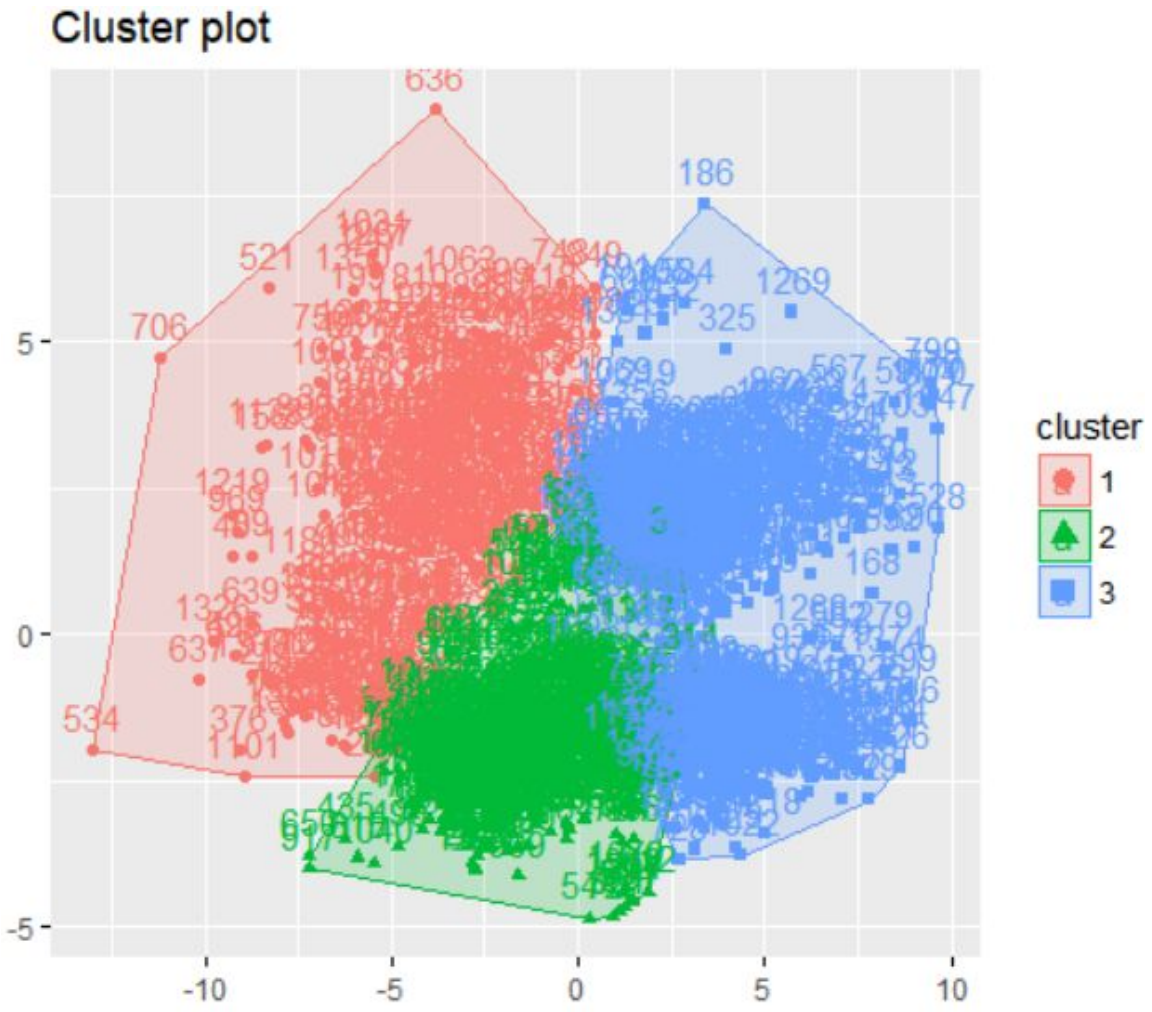
The following result was retrieved with an automatic cluster evaluation.



We can see that a cluster distinction was found on the first 2 principal components.



The dendrogram shows us that a clusterfication with c=4 can be done too. The Inertia gain diagram shows us that an acceptable gain can be found between 3 and 4 cluster. I decided to take 3 cluster into consideration, because the gain was still quite high.



This leads us to the cluster distinction between:

- low value housing objects- mean (157470€)

```
$`1`
                   v.test Mean in category Overall mean sd in category Overall sd       p.value
Age              27.575886       1.16895437 -0.001485600      0.7702345  1.0029121 2.166097e-167
GarageDetchd     19.244031       0.81901132  0.003306367      1.0925534  1.0015676  1.583982e-82
HasEnclosedPorch 16.038993       0.74423573  0.040621338      1.3590439  1.0365752  6.825210e-58
YearSinceRemodel 15.905514       0.70832981  0.027697576      1.0130087  1.0111324  5.802514e-57
HasX2ndFlrSF     15.300605       0.63940591 -0.007319500      0.8805490  0.9987448  7.574974e-53
Has2ndFlr        15.300605       0.63940591 -0.007319500      0.8805490  0.9987448  7.574974e-53
Remodeled        14.731596       0.59988352 -0.022348356      0.8500556  0.9980348  4.040493e-49
```

Low value houses are mainly characterised by their age and the attached garage. The age is on average at 70 years.

- medium value housing objects - mean (170673 €)

```
$`2`
                v.test Mean in category Overall mean sd in category Overall sd       p.value
ExterQual    28.446374       0.94540989  0.001415412      0.8541702  1.0102536 5.403230e-178
YearBuilt    28.106375       0.92741923  0.001485600      0.3688624  1.0029121 8.187432e-174
OverallQual  27.112938       0.90000043 -0.007270123      0.7430411  1.0187042 6.930489e-162
GarageYrBlt  26.694551       0.89056696 -0.009387972      0.4717359  1.0263276 5.444852e-157
YearRemodAdd 25.530542       0.81902312 -0.028796855      0.3440509  1.0109542 9.031869e-144
KitchenQual  24.780191       0.81072331 -0.001288290      0.8065298  0.9975747 1.466070e-135
FullBath     24.502416       0.81312647  0.005309985      0.6161670  1.0036717 1.392116e-132
```

Mainly for the cluster assignment is responsible the ExterQual, YearBuilt and OveralQual. In the following are the summaries to each:

```
   YearBuilt      OverallQual
 Min.   :1892   Min.   : 2.000
 1st Qu.:1957   1st Qu.: 5.000
 Median :1969   Median : 6.000
 Mean   :1970   Mean   : 5.826
 3rd Qu.:1993   3rd Qu.: 7.000
 Max.   :2009   Max.   :10.000
```

- high value housing objects (203424 €)

```
$`3`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
YearSinceRemodel 11.272307       0.44875541  0.027697576      0.8175067  1.0111324 1.797971e-29
BsmtFinSF1       10.071872       0.36919333 -0.001602670      0.7527508  0.9965610 7.356430e-24
NewerDwelling     8.633285       0.31408894 -0.005228189      0.8563350  1.0012100 5.961455e-18
BsmtHalfBath      6.498316       0.26243940  0.015691607      1.3769456  1.0278541 8.122395e-11
HasPavedDrive     6.398852       0.20573332 -0.044467068      0.6245514  1.0584369 1.565499e-10
Fence             6.332370       0.24387308  0.008563029      1.1474900  1.0058964 2.414234e-10
LotArea           6.138417       0.19165991 -0.031500566      0.7773460  0.9841016 8.334794e-10
LotFrontage       5.695259       0.18431215 -0.029326384      0.7983232  1.0154187 1.231847e-08
```

High value objects are mainly characterized by the year since remodeled, which was on an mean of 1991(median 1999). Good old buildings are more likely to be remodeled(simple assumption, like a palast).

8. Predictive Analysis

The predictive analysis was done with the following regression models:

- XGBoost
- Elastic-Net
- Lasso
- Ridge
- Random forest

XGBoost fits shallow regression trees to the data and then additional trees to the residuals, we'll repeat this process for 30000 rounds so that the model has learned from the data as much as possible without overfitting. XGBoost is a Gradient Boosted Method (GBM), which is an ensemble learning method that uses a very large number of decision trees, which are typically weak learners and combines them into one final prediction. For gradient boosted trees, the new trees added to the model are the weak learnings where

$$F_0 = 0, \quad F_t(x) = F_{t-1}(x) + h(x)$$

such that F(x) is the entire model after time t−1 and h(x) is the new weak learner (new tree) that will be added to our model. In particular for XGBoost we will minimize the following objective function at time *t* such that:

$$Obj(F_t) = L(F_{t-1} + F_t) + \Omega(F_t)$$

such that L(Ft) is the loss function and Ω(Ft) is the regularization function. In the XGBoost package the regularization function is computed with

$$\Omega(F_t) = \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2$$

where w_j is the score of the jth leaf, λ is the leaf weight penalty, γ is the tree size penalty and *T* represents the number of leaves in the tree.

I did choose which parameters minimize the loss in the model by searching through a grid of parameters. Following parameters were evaluated:

- η - Shrinkage term
- γ - Tree size penalty
- λ - L2 leaf node weight penalty
- *Max Depth* - The maximum depth of each tree
- *Subsample* - proportion of data for bagging (randomly sampling with replacement)
- *Min Child Weight* - The minimum weight that a node can have

One limitation to using GBM's and XGBoost is its inability to extrapolate and because of this I can use linear models to better predict any sale prices outside the range of prices given in our training set.

Ridge Lasso

We know there exists multicollinearity in our dependent variables as area features are a determinant for a house having a certain amount of rooms, we know that the garage variables have heavy dependency and the list goes on. Due to this a simple linear regression model will not be of much help to predict accurate sale prices, which is why we can make use of both ridge and lasso regression. The Ridge penalty is known to shrink the coefficients of correlated predictors towards each other with the use of the ℓ2-norm, while the lasso tends to pick one of them and discard the others by construct of the ℓ1-norm. Adding the ℓ1 and ℓ2 penalties give a good constraint on the coefficients of our model and solves our problem of having collinear variables. Another useful property of these regression models is their ability to *extrapolate*, i.e.

they can predict house prices that where outside of the range of prices which were given in the training set. Along with Ridge and Lasso we can use elastic-net regularization which can be derived such that,

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{n=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$$

where l(y,n) is the negative log-likelihood contribution at observation *i* and λ is a shrinkage parameter. When λ=0, no shrinkage will be performed and and if λ increases, the coefficients are shrunken at a larger rate. Elastic-net has the property that the equation is controlled by α, which leverages the gap between ridge and lasso. When α=0 we have

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{n=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2]$$

which is our equation for ridge and when α=1 we have,

$$min_{\beta_0,\beta}\frac{1}{N}\sum_{n=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[\alpha||\beta||_1]$$

which is our equation for lasso.

To better estimate the test error of a predictive model we'll use K-fold cross-validation. For K-fold cross-validation, the observations in the data are split into K partitions, then the model is trained on the K – 1 partitions, where the test error is predicted on the left out partition k. We repeat this process for k = 1,2…K times and we will average the results. To score how well our model predicts output we will take the root mean squared error (RMSE). The RMSE can be computed the following:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

To avoid overfit, I used a function called early_stopping_rounds to find at which point the model begins to overfit, which will change based on the parameters used.

I used nrounds = 10000 and eta = 0.01. A very good way of choosing optimal parameters is with the use of expand grid search.
Feature importance is computed by averaging the gain of each feature for all split and all trees in the model. We can take a look at the 10 most important features used in the model.

Having the opportunity to see which features are most important allows us to give an emphasis on making use of specific features to improve our model even more. We can also look deeper in the model to see which features are adding little to no value for predictions.

Cv.glmnet function which will run cross validation for us and pick the λ the produces the smallest error rate. We can call on the same glmnet function and use a ridge regression when α=0, lasso when α=1 and elastic-net when $\alpha \in (0,1)$.

Initial evaluation:

Method	**test-rmse**
XgBoost	6540
RandomForest	10321.12
lasso	21133.27
ridge	20179.57
net	21184.43

We see a RMSE for elastic-net, lasso and ridge are all very similar yet not as low as the RMSE we see from XGBoost, *however*, ridge, lasso, and elastic-nets ability to extrapolate sale prices

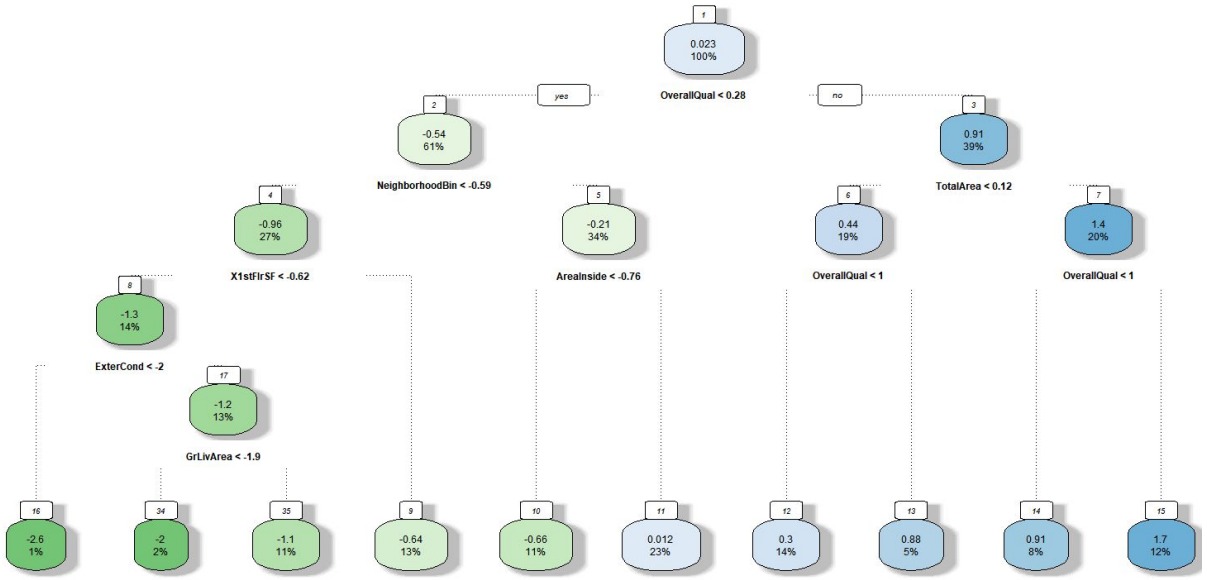
beyond the range of the training set will be crucial for predicting more accurately towards unseen data.

Winner - full evaluation:

Method	train-rmsle	test-rmsle	cross-validation-rmsle
xgBoost	0.055603	0.191394	0.202694
0.5*xgboost+ 0.25*regression+ 0.25*randomforest			0.195643

I further followed a simple regression, which I gave up for the xgboost method due to early evaluations and better results of the boosted method.

The following tree was retrieved based on the normalized data:



We can see that the overall quality followed by the total area and neighborhood have the biggest decision factor on the price prediction. The initial assumption about the quality, neighborhood and area was right.

9. Conclusion

The analysis and pre-processing took most of the time. I approximate about 90% of the total time of 100 hours. The first evaluation approach landed in the first 12% of the leaderboard, which was a motivation to further research on improvements. The learned methods and insight in multivariate analysis helped to find different approaches to the general Kaggle community, like the cluster analysis or the missing data evaluation. The final result was a simple form of stacking the prediction results of XgBoost with the results of the linear regression. The reasoning for this was to give the prediction space for extrapolation. This stacking step can be in a future step improved and a prediction computed on the stacked result for better predictions.

In a future outlook, I would make optimizations in the part of the One-hot encoding of categorical variables. I did not check exhaustively every single categorical value for an ordinal order - only for some of them. The same for outlier detection. I took a focus on the most obvious features that contained outliers. A look into the other features could result in finding further outliers.

In the field of feature selection and engineering I want in a future step to make more out of the principal component analysis, deeper reasoning on the features itself in the real world, as well as using the cluster analysis results as additional features.

Furthermore, could an evaluation of different prediction algorithms be helpful to find better fitting prediction methods.

The final idea was to also use the predictions on the specific cluster. I have the assumption that cluster-specific predictions could archive an improved result.

Appendix 1: Data_description

MSSubClass: Identifies the type of dwelling involved in the sale.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl	Gravel
Pave	Paved

Alley: Type of alley access to property

Grvl	Gravel

Pave	Paved
NA	No alley access

LotShape: General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

LandContour: Flatness of the property

LvlNear	Flat/Level
Bnk	Banked - Quick and significant rise from street grade to building
HLS	Hillside - Significant slope from side to side
Low	Depression

Utilities: Type of utilities available

AllPub	All public Utilities (E,G,W,& S)
NoSewr	Electricity, Gas, and Water (Septic Tank)
NoSeWa	Electricity and Gas Only
ELO	Electricity only

LotConfig: Lot configuration

Inside	Inside lot
Corner	Corner lot
CulDSac	Cul-de-sac
FR2	Frontage on 2 sides of property
FR3	Frontage on 3 sides of property

LandSlope: Slope of property

GtlGentle	slope
Mod	Moderate Slope
Sev	Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek

Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Condition1: Proximity to various conditions

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam	Single-family Detached
2FmCon	Two-family Conversion; originally built as one-family dwelling
Duplx	Duplex
TwnhsE	Townhouse End Unit
TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

OverallQual: Rates the overall material and finish of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

OverallCond: Rates the overall condition of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

RoofMatl: Roof material

ClyTile	Clay or Tile
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
Tar&Grv	Gravel & Tar
WdShake	Wood Shakes
WdShngl	Wood Shingles

Exterior1st: Exterior covering on house

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common

BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

Foundation: Type of foundation

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete

Slab	Slab
Stone	Stone
Wood	Wood

BsmtQual: Evaluates the height of the basement

Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches
NA	No Basement

BsmtCond: Evaluates the general condition of the basement

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd	Good Exposure
Av	Average Exposure (split levels or foyers typically score average or above)
Mn	Mimimum Exposure
No	No Exposure
NA	No Basement

BsmtFinType1: Rating of basement finished area

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ	Good Living Quarters
ALQ	Average Living Quarters

BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace

HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

CentralAir: Central air conditioning

N	No
Y	Yes

Electrical: Electrical system

SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2
Sev	Severely Damaged
Sal	Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level
TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa	Fair - Prefabricated Fireplace in basement
Po	Poor - Ben Franklin Stove
NA	No Fireplace

GarageType: Garage location

2Types	More than one type of garage
Attchd	Attached to home
Basment	Basement Garage
BuiltIn	Built-In (Garage part of house - typically has room above garage)
CarPort	Car Port
Detchd	Detached from home
NA	No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin	Finished
RFn	Rough Finished
Unf	Unfinished
NA	No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage

GarageCond: Garage condition

Ex	Excellent
Gd	Good
TA	ypical/Average
Fa	Fair
Po	Poor
NA	No Garage

PavedDrive: Paved driveway

Y	Paved
P	Partial Pavement
N	Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No Pool

Fence: Fence quality

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev	Elevator
Gar2	2nd Garage (if not described in garage section)
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

SaleCondition: Condition of sale

Normal	Normal Sale
Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family	Sale between family members
Partial	Home was not completed when last assessed (associated with New Homes)

Appendix 2: Numerical feature distribution

-taken out for printed version-

Appendix 3: Categorical feature distribution

-taken out for printed version-