House Prices: Advanced Regression Techniques

MAP 553-Regression

Aksel Kaastrup Rasmussen

11/30/2019

Contents

1 Introduction.

Write a short introduction describing the research problem. Clearly state the research hypothesis at the end.

2 Exploratory Data Analysis/Inital Modeling.

We are given a dataset, which consists of information of large number of residential homes in Ames, Iowa. More precisely our traning and test dataset consists of 1095 and 365 observations respectively with both 67 features. It is clear both cleaning and basic handling of missing values has already taken place, and we will not explore such techniques further here. There has already been taking care of a great deal of ordinal variables using ordinal encoding and normalizing. We will continue this work.

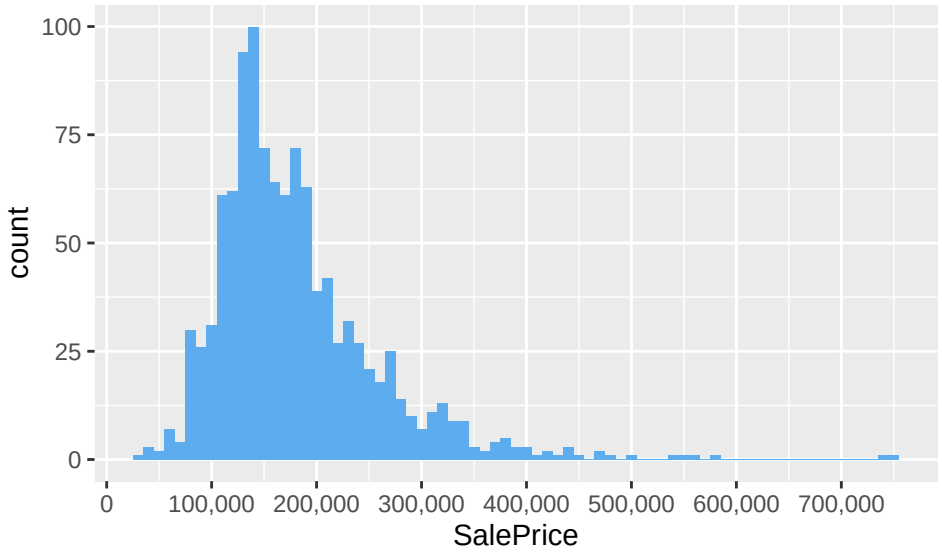
The dataset consists of both numerical, nominal and ordinal variables, and right from the start we will label encode ordinal variables into numbers. This arguably holds for *Street* and *PavedDrive*, since the levels have an order. With this ordinal encoding we ensure the variables keeps the ordered structure. In addition there are lots of categorical *quality* variables, which indeed are ordinal. As an example consider *KitchenQual*. The kitchen quality has an ordered structure ranging from poor to excellent in quality. For this reason we will use ordinal encoding here as well.

The numerical variable *MSSubClass* is clearly a categorical variable, which identifies the type of dwelling involved in the sale. Hence we will also factorize and rename to make this fact clearer.

We have chosen to remove the variable *Utilities*, since this variable is constant for all observations, and finally we have split the data into categorical and numerical variables with dimension 36 and 30 respectively. As a start of the exploratory analysis, let us look into the response variable.

2.1 The response variable

We notice the response variable *SalePrice* is right skewed. This is due to the fact that more people can afford a relatively cheap house compared to an expensive house.



This skewness will justify a log-transformation in the early modelling phase in order to satisfy the assumption that the variables are normally distributed. Let us now take a look into the response variables relationship with the numerical variables. Consider figure 1, where we have compared all numerical variables with correlation coefficient above $\frac{1}{2}$ with *SalePrice*.

It turns out the overall quality of the house and *GrLivArea*, the above ground living area, are the variables of highest correlation with *SalePrice*. We also see there is high intercorrelation between some of this variables of interest. Not surprisingly are features like *GarageCars* and *GarageArea*, and features like *YearBuilt* and *GarageYrBlt* highly correlated. In general we observe that variables concerning the same area (ie. garage, basement and so on) are typically highly correlated. This fact will give reason to a model reduction in the modeling phase by removal of variables that are highly correlated to a more important variable.

2.1.1 Relationship with correlated features

Considering figure 2 there seems to be a growing trend to in SalePrice when increasing the overall quality. Also there does not seem to be any obvious outlier, perhaps apart from the somewhat expensive house with overall quality 4. Considering figure 3 there are no strong trend, but one could argue *SalePrice* is a bit higher for newer houses than old houses.

Considering the relationship between *SalePrice* and *GrLivArea*, the second most correlated variable, we notice in figure 4 that there is a somewhat strong positive trend in SalePrice as function of GrLivArea. One could suggest that house 596 and 199 are outliers, since they are very large properties with a relatively small sale price. Taking a harder look at these houses they have a maximum normalized score of 2.65 in overall quality, which gives further reason to classify them as outliers. We will keep these houses in mind when removing outliers.

2.2 Important predictors

Now that we have explored the response variable and some of its highly correlated predictors we woud like to take a look on the variables of high importance for the prediction. Of course we can do this by building a multiple linear regression model using all the variables and look at the significance of each variable, but right now we would like to get a feel of the data without building a huge model. To this end we will use random forests as inspired by (Bruin 2017), where we will refer to a classic reference for random forests (Breiman

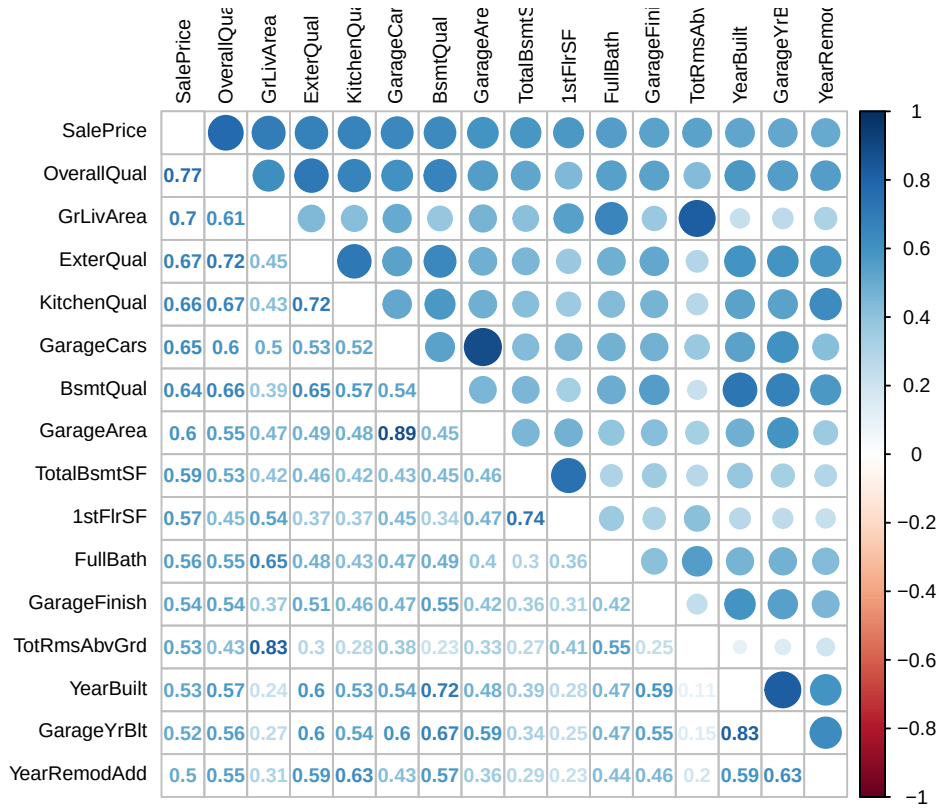


Figure 1: Correlation plot of numerical variables of high correlation with SalePrice

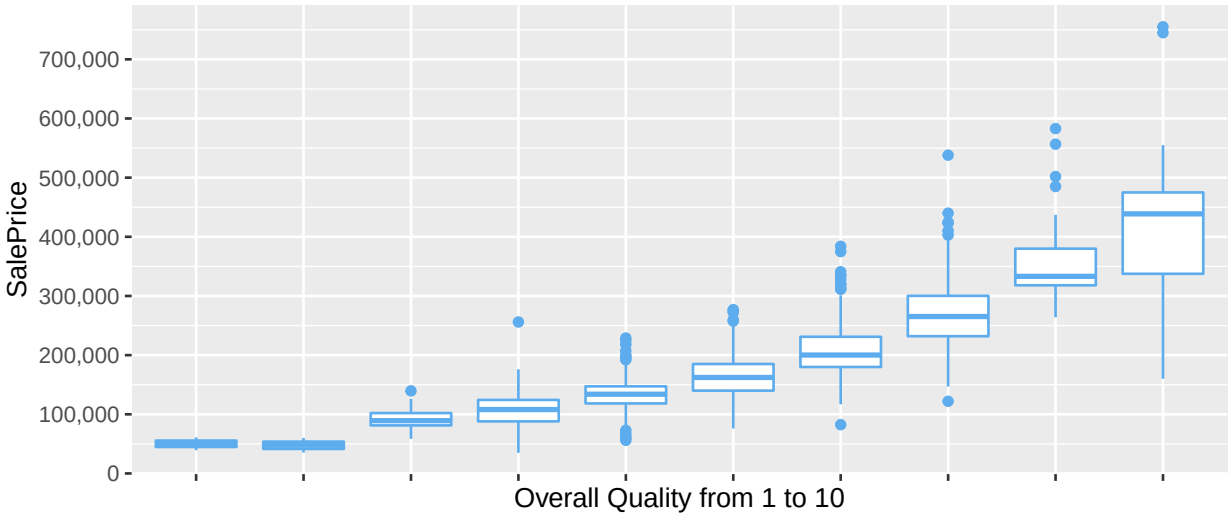


Figure 2: Boxplot of SalePrice divided by their overall quality ranging from 1 to 10.

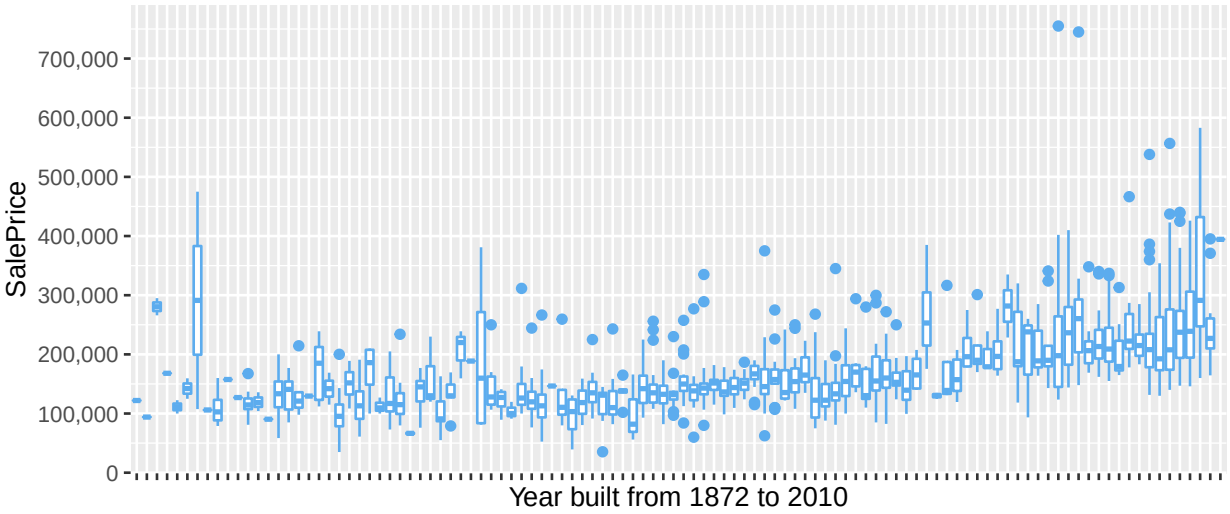


Figure 3: Boxplot of SalePrice divided by the years the houses are built from 1872 to 2010

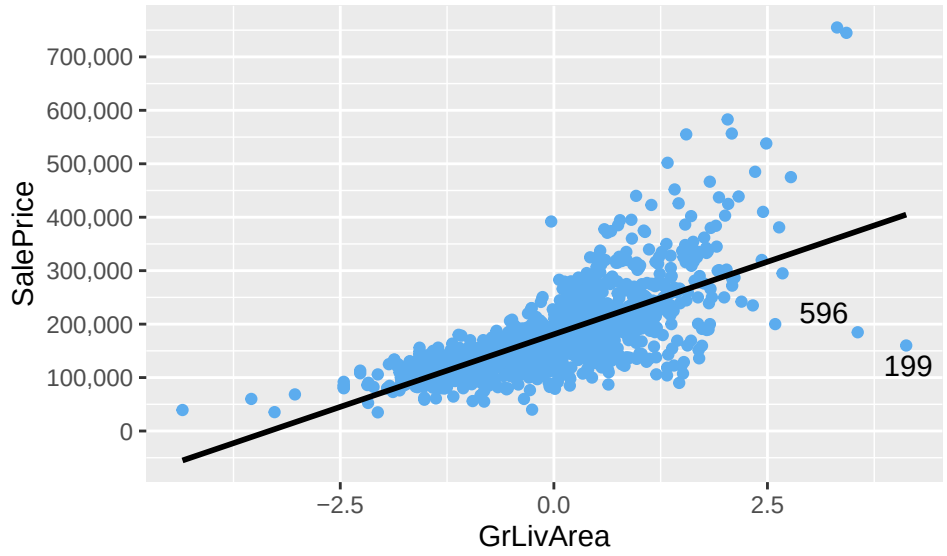


Figure 4: Scatterplot with trend of SalePrice as function of GrLivArea

2001). The technique computes the mean square error before and after permuting the features among the dataset to which we fit a random forest. The technique is implemented in the R-library `randomForest`.

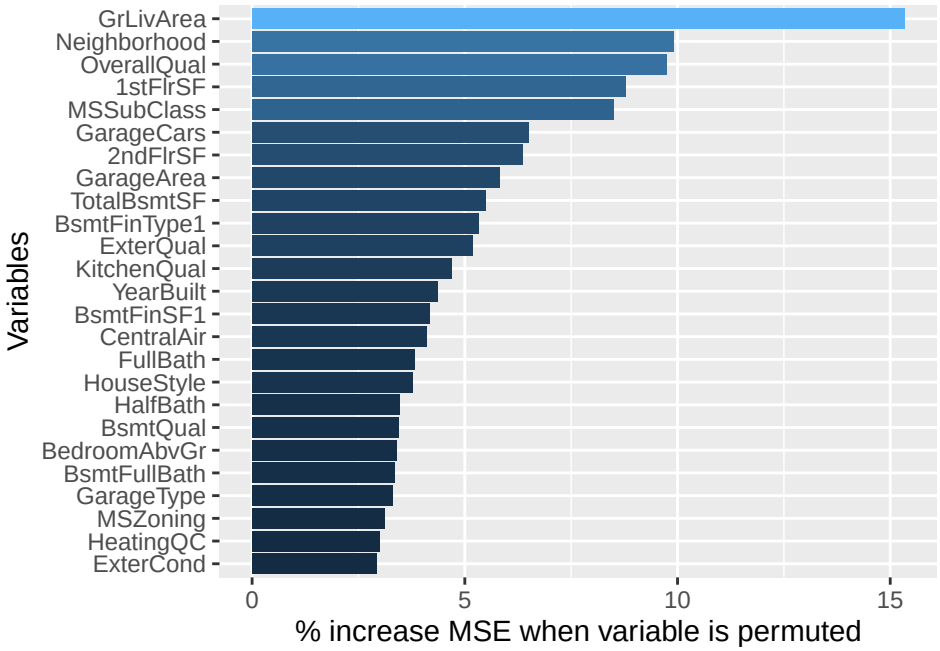


Figure 5: Most important features found by random forest technique

In figure 5 we see the 25 most important variables as ranked by the random forest model. We have already seen some of these variables' relationship with the response variable. Indeed it comes as no surprise that the above ground living area and overall quality are important, since houses scoring high in these variables are expected to be expensive. Let us now take a look on the distribution of the top 10 in figure 6. Some of the variables seem to have a central Gaussian distribution like *GrLivArea* or *1stFlrSF*, while distribution for other variables are skew like *OverallQual*, and yet other variables distribution are harder to determine since they have already undergone some transformation. We will keep this in mind when standardizing and and log-transforming in the modelling phase.

One idea we could explore further is to merge the different neighborhoods in the neighborhood variable if the sale prices are similar enough. In this way we could reduce the dimensionality of this categorical variable.

%%% What about pairwise correlation?

3 Modeling and Diagnostics

Let us start the modeling by a bit of feature engineering in order to prepare our dataset for modeling. This preparation will consists of one-hot encoding of all categorical variables and a standardization of numerical variables.

3.1 Feature engineering

We have encoded the categorical variables, and standardized the numerical variables which combined now amounts to 1460 observations with 199 variables.

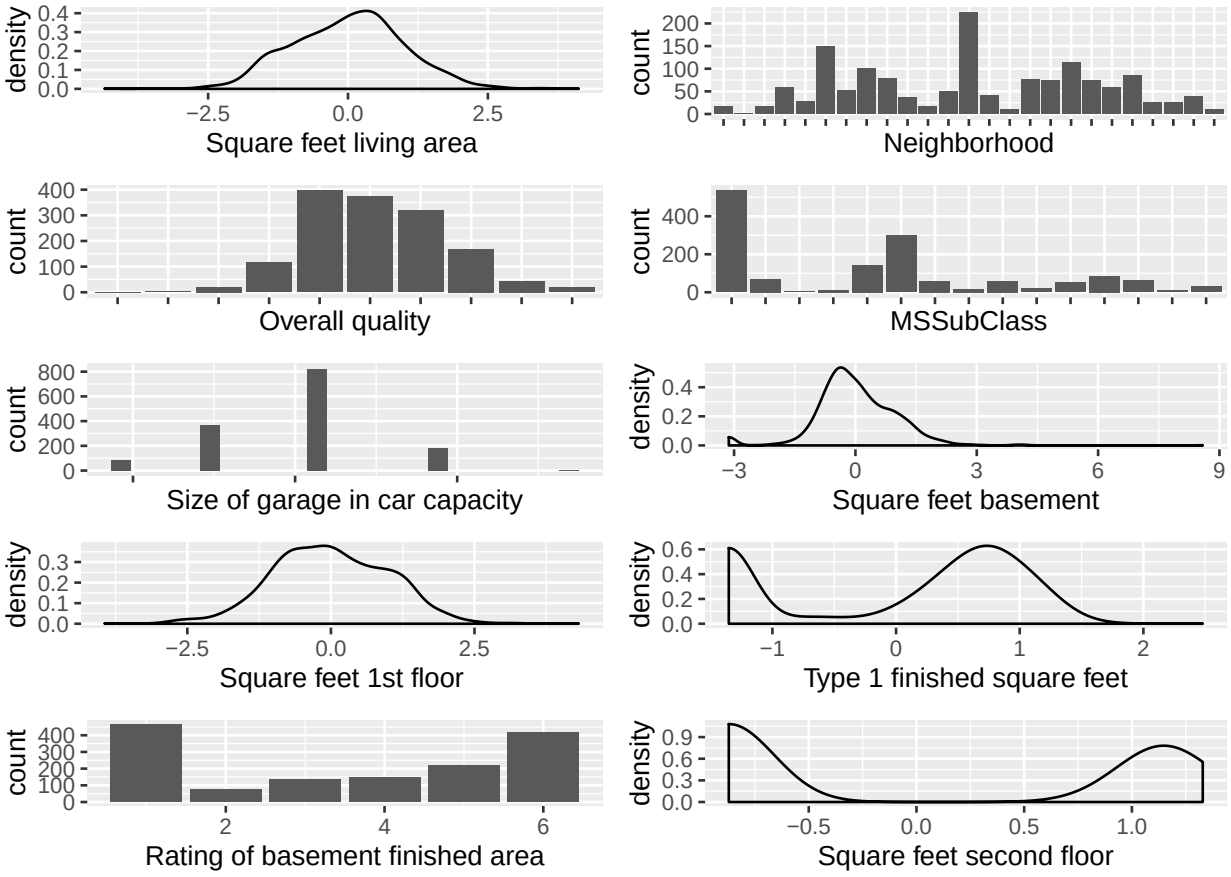


Figure 6: Distributions of important variables

3.2 A multiple linear regression model

We have fitted our training data to a simple multiple linear regression model to the data of $p = 199$ variables and $n = 1095$,

$$Y = X^T \beta^* + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ and $\beta^* \in \mathbb{R}^p$ is the unknown regression vector. We assume we have a sample of i.i.d. copies of $(X, Y) : (X_i, Y_i)_{i=1}^n$. We will explore this assumption further as well as the various assumptions regarding the residuals, see **P1** to **P4** in (Lounici 2019). Using our designmatrix of all our observations $\mathbb{X} = (X_1, \ldots, X_n)^\top = (X_{i,j})_{1<i<n, 1<j<p}$ and using our observed responses $\mathbb{Y} = (Y_1, Y_2, \ldots, Y_n)$, we want to estimate β^* by the following minimization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}\|^2.$$

Notice then that the residuals $\epsilon = \mathbb{Y} - \mathbb{X}\boldsymbol{\beta} \sim N(0, \sigma^2 I_n)$ This is done in `R` using the `lm`-function. The diagnostics for this regression can bee seen in figure 7.

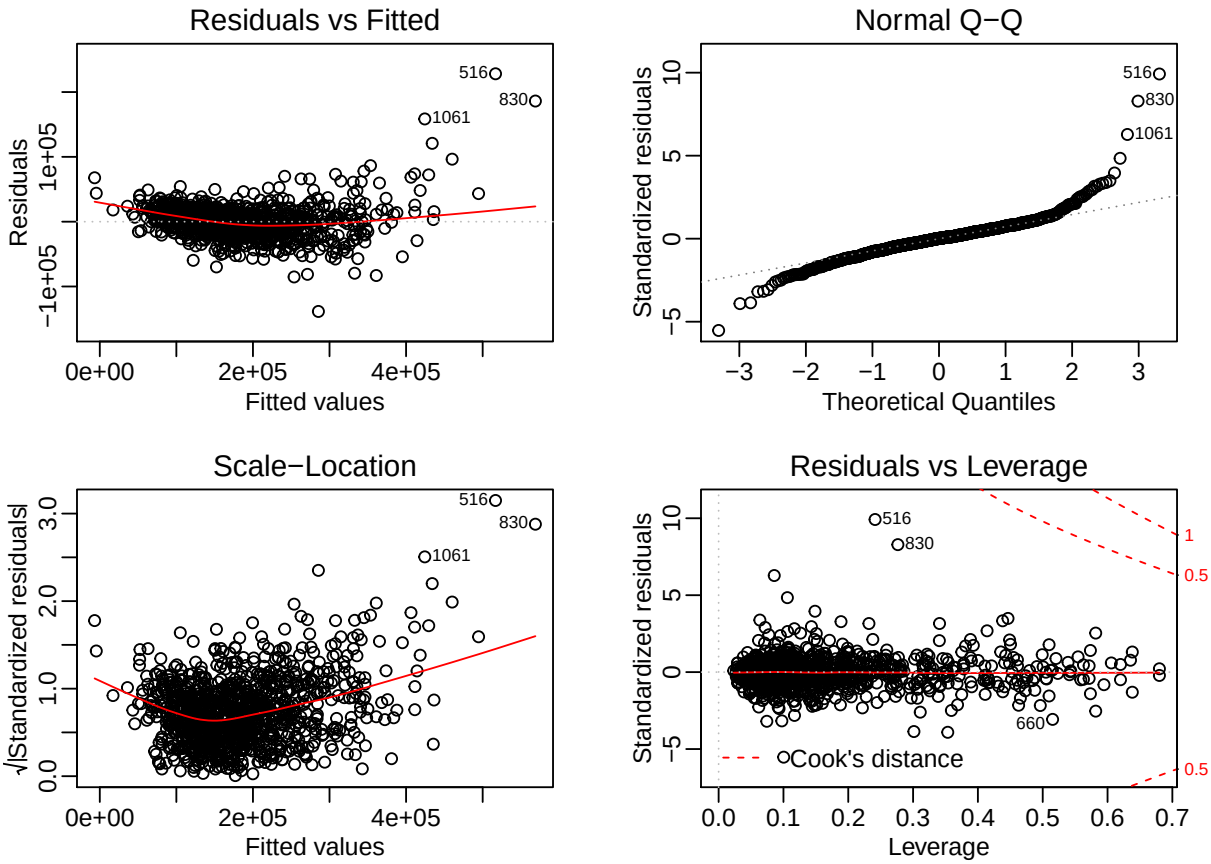


Figure 7: Distributions of important variables

- Outliers
  - Removing levels with few or no observations
- Skewness
- More Feature engineering. Not all information is explained

3.2.1 Diagnostics and feature engineering

- Feature engineering

- Multiple linear regression
-
- Lasso
- XGboost

Start by building any multiple linear regression models you think are appropriate. Create diagnostic plots to determine the appropriateness of your model. Discuss whether the assumptions are met. If not, take any actions you think are justified such as: transformations, removing outliers, removing variables, etc. . . Explain what decisions you make and explain why you made them. Check again your diagnostics for the new models

4 Final Models.

Summarize your final models: report the parameter estimates, standard errors, confidence intervals, and p-values. Interpret the fitted models in the context of the problem. If you have several models, then compare them. Note that there are several ways to compare different regression models: (i) partial F-tests, (ii) residuals and diagnostics and (iii) cross-validation (or other measures of prediction error). Use the test sample to get an estimate of the generalization error of your final model.

5 Discussion.

- Data cleaning should be better - too much information lost

What are your final conclusions? Mention any limitations of your analysis, or possible future directions of research.

Comments from class Goal: Construct a model in order to predict the price of a house given some features. Dataset: 3000 datapoints with 80 features. Training and testing is split equally. Flow: Dataset -> atacleaning -> Exploratory data analysis -> Feature engineering (Construct features) -> Feature selection (p-value, AIC/BIC, Lasso, feature importance (sklearn) with RF) -> Model: (Multiple Linear model, Lasso, boosting, RF, hyperparameter)

Categorical features: 1-out-of-K coding

Interaction based on correlation * Year built x overall quality * quality x remodel * quality x basement size * quality x living area * quality x baths

Construct a benchmark model + its score on the test set and improve it via feature engineering + hyperparameters

References

Breiman, Leo. 2001. “Random Forests.” https://doi.org/10.1023/A:1010933404324.

Bruin, Erik. 2017. “House Prices: Lasso, XGBoost, and a Detailed EDA.” Kaggle. https://www.kaggle.com/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda.

Lounici, Karim. 2019. “Visual Diagnostics and Model Validation.”