Main Assignment in Statistics for Business Analytics I

A real estate analysis of property sales in Ames, Iowa

Anastasios Theodorou

Master of Science in Business Analytics Student

(AM: p2822007 – Dataset: No 53)

Date: January 5th, 2021

Contents

1. Introduction

This analysis focuses on property sales and specifically it refers to 1500 property sales that had occurred in Ames, Iowa between the years 2006 and 2010. The main project includes 2930 property transactions, but it was shrinked to 1500 observations and other 500 which consist our test-sample, details will be discussed later. The problem which was the initial motivation to write this report was that it has been observed many variations as far as the pricing of the houses is concerned and for real estate there wasn't any formula to include all the parameters that form a housing price. So, the aim of this report is to learn more about the field of real estate and perform a drill through analysis in order to find a model that will predict the houses' pricing.

So, we collected data from the Ames City Assessor's Office that contain all the useful information that a buyer wants to have and generally what attributes a typical property has. These data include many measurements and calculated variables in order to have a total view of a property. To be more specific, these variables include general information concerning the location of the property (the area's density, its neighbourhood, access of the road etc), the type of the house (style, shape and others), outside and inside features of it (type of roof, of exterior covering of the house, of fence, pool and type of heating, of air condition, number of bathrooms, kitchens, fireplaces, utilities, etc respectively), the age of the house and information about the transaction (price, year, month of sale, sale type and condition).

It worth to be mentioned that the first purpose of these data collection was the tax assessment of real estate field, but it ended up helping the formation of model prediction in housing price. Also, these data were initially used in a Regression Project of Truman State University, created by Dean De Cock (2011); see References.

This report starts with descriptive and exploratory data analysis, where will be discussed few things about the clearing of the initial data and what was found from a first insight. It will be mentioned some general information on what characteristics a typical property has and in addition to that some graphical visualizations will be presented. After that, will be made some basic comparisons between the variables of the data and then it will follow the regression analysis. This report will focus on the model that was selected and in brief the steps which were followed. Then will be

discussed the predictive analysis of the final model, the assumptions that are fulfilled and will interpret the model in order to understand the typical profile of a property. Finally, some comparisons will be done between the final model and other linear models and the selection of the model will be justified, based upon the indicators and the metrics we have. Will be given some open issues that were found and will be proposed a few suggestions for further research.

2. Results

2.1 Descriptive Analysis and Exploratory Data Analysis

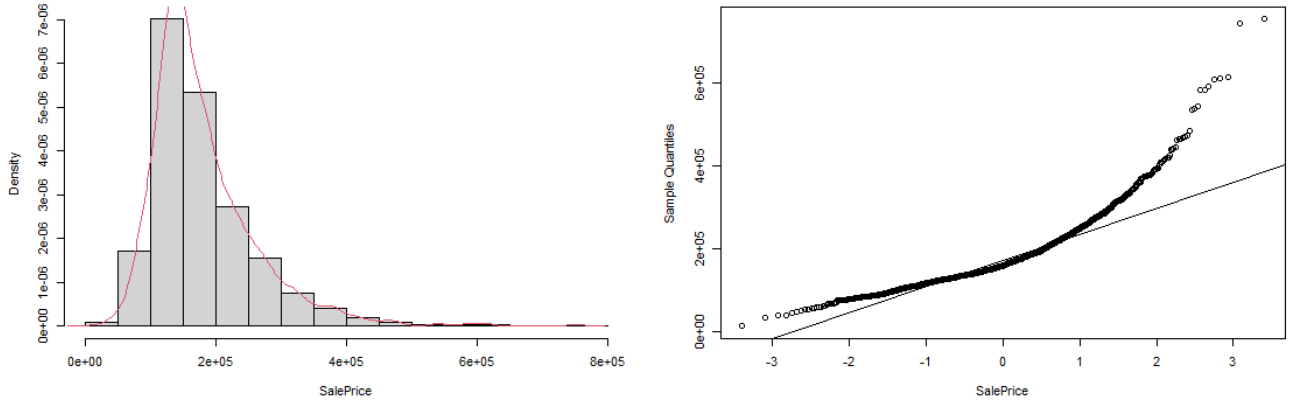
The dataset, that is investigated, wanted some cleaning before it was used for further analysis. The original structure of data frame, which contained all the attributes and calculated measures of the properties, consisted of “character” and “integers”, so they had to be converted as “factors” and “numeric” respectively. Further to this, there were some variables, such as “MS.SubClass”, “Overall.Cond” and “Overall.Qual” which were numeric vectors, but had visible levels. That is why, they have been converted as “factors” too.

The next thing which had to be done was the conversion of missing values. These NA values in numeric variables were altered to “0” and in the factors have been created a new level, which is called “No”. The above transformations were made and in the test sample of 500 observations.

After that, it was necessary to have an insight in the given data and perform a descriptive analysis. For that exact reason, histograms were created for numeric vectors and boxplots for factors (see Appendix for some examples). From this visualization and from the description function that package “psych” in R offers, it is worth to discuss the following:

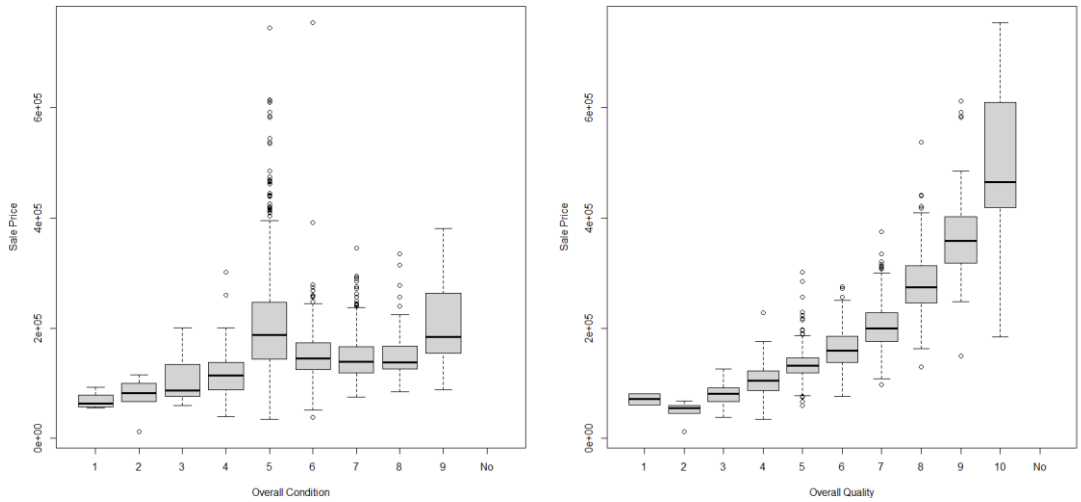
- Lot.Frontage: on average (avg) 69.5 feet (ft) of street relate to a property
- Lot.Area: close to symmetric variable, nearly all the houses have, as lot size, between 10,000 and 25,000 square (sq) feet
- Year.built: the majority of the houses have been built in and after of 2000

- Year.remodel: most of the houses have been renovated in 1950 and after 2000
- Mas.Vnr.Area: many outliers and as we can assume from the data most of the properties won't have a masonry veneer area
- The total sq ft of basement is close to be, as variable, normally distributed, which means that an average house will have approximately 1051 sq ft of basement area. Most of the properties have good exposure to walkout or garden level walls, but they are in a poor condition with severe cracking or wetness.
- An average house will have nearly 1165 sq ft as first floor and most of the properties either they haven't a second floor or they have more than 500 sq ft.
- Rarely a house has one or two bathrooms in the basement. Above the ground most of the houses own two full bathrooms and no half bathrooms.
- A typical house above the ground has nearly three bedrooms, one kitchen and generally it has in total up to 6 rooms. Also, most of the houses do not own a fireplace, but if they do, it will be at most one per house.
- Garage: most of them were built in 1977 or after 2000. An avg garage has capacity of two cars and its area is close to 500 sq ft. It is built in (part of the house) and in a good condition and quality.
- The majority of the houses does not own an open porch and if they do it is very small (nearly 25 sq ft). As far as the enclosed porch is concerned, this variable has many outliers, which mean that most of the houses do not own one or they have a very big (range: 1,012 sq ft).
- Most of the properties do not have a screen porch, fence, alley access, neither a pool, but nearly all the houses have central air conditioning and gas hot water or stream heat as heating.
- Most of them have been sold before their construction have been completed or when they had just finished and the most transactions took place in June with small difference of those in May and July. They have been sold at 2007 with 2008 and 2009 to follow. The avg price is 181,285$ (12.789$ the cheapest and 755,000$ the most expensive). The figure 2.1 shows some of the above details and from that, we can conclude it is not normal, but positive skewed having many outliers as well.



2.1 Histogram and QQ Plot of Sale Price Variable

- Generally, the majority of the houses have severe slope, but they have very excellent overall quality and excellent overall condition. In boxplots we can also observe that the medians of every level in the latter feature (Overall Condition) follow the normal distribution and in the first (Overall Quality) follow the exponential distribution, as the below 2.2 figure shows.



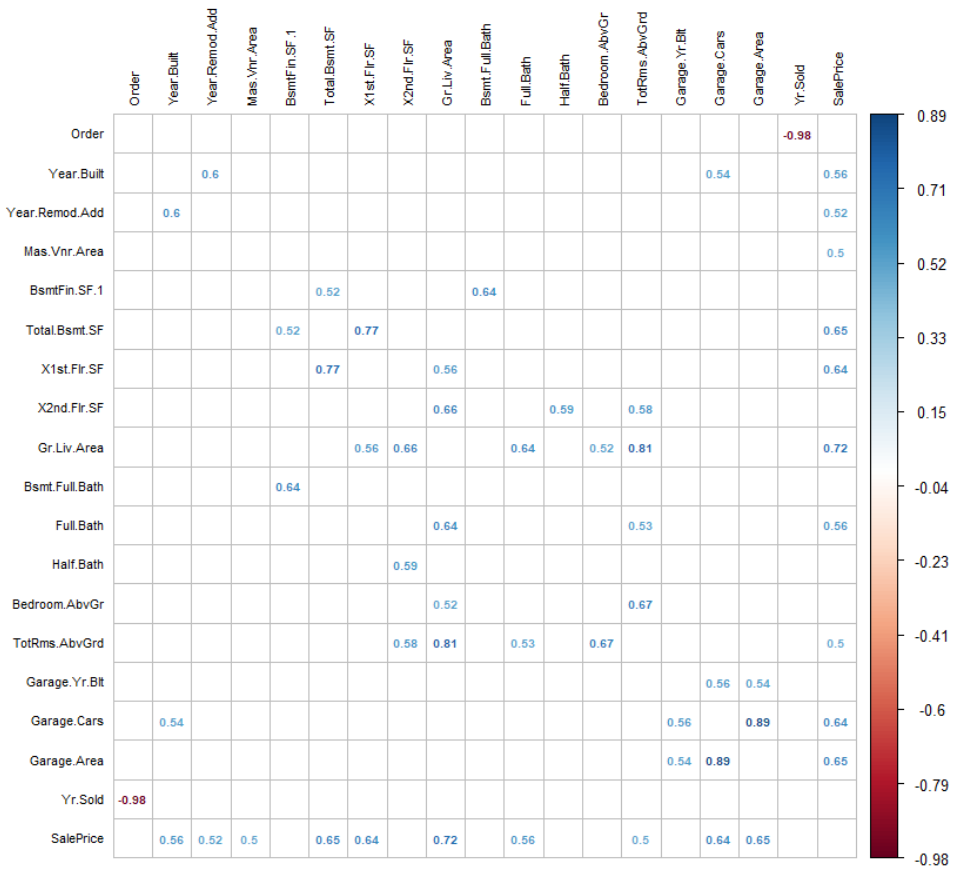
2.2 Boxplot of Overall Condition and Overall Quality

2.2 Pairwise Comparisons

In order to check the relationships between the variables and especially correlations between the Sale Price and all the others it was created the following 2.3 table. This table includes all these associations which make sense with the help of Pearson's Correlation Coefficient test (cor). From this table it can be assumed, that the only

variables that are fairly related to Sale Price is “Gr.Liv.Area” (the living area above the ground measured in sq ft) (cor = 0.72). This means that every change in the second covariate, will affect the first and the estimations will not be so accurate, increasing simultaneously the standard error. Other variables that the “Sale Price” is medium correlated with, is “Total Basement Sq Ft” (cor = 0.65), the sq ft in the 1st floor (“X1st.Flr.SF”) and the “Garage Cars” (cor = 0.64 both of them).

Furthermore, from 2.3 table it can be assumed that there are also more other variables that are highly correlated. For example, the “Order” is nearly perfect negative associated with “Year Sold” (cor = -0.98) which means that the first attribute maybe was created based on the second feature, the year that a property had been sold. Also, other variables that have strong relationship, by descending order, is “Garage Area” and “Garage Cars” (cor = 0.89), “Total Rooms Above Ground” and “Gr. Liv. Area” (cor = 0.81) and “Total Bsmt SF” and “X1st.Flr.SF” (cor = 0.77).

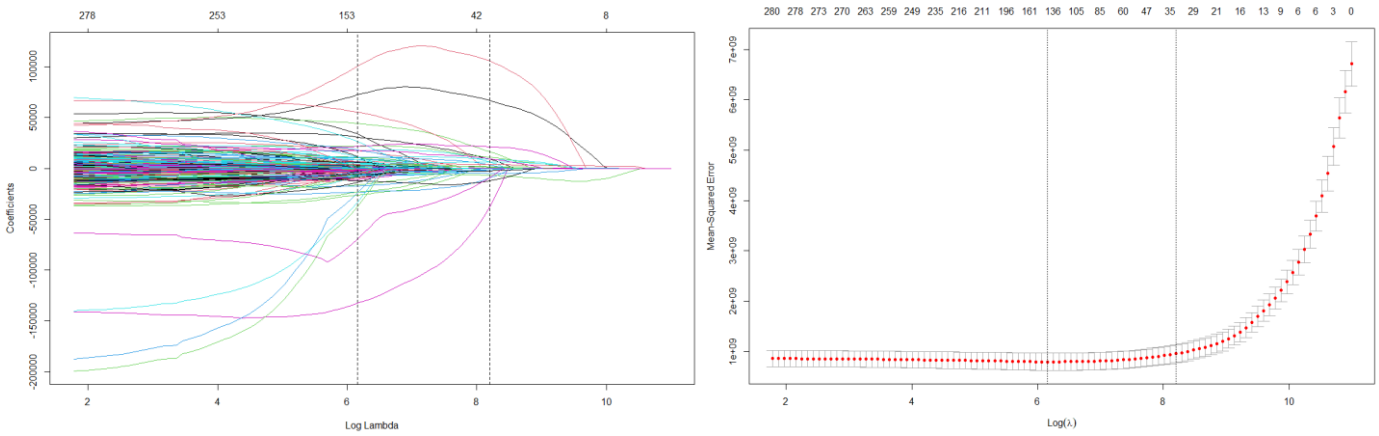


2.3 Pearson’s Correlation Coefficient Plot

2.3 Predictive and Descriptive Models

In order to predict the price of the houses it is necessary to create a model which contains the basic characteristics that form the final price. For that reason, many assumptions and transformations had to be made.

First of all, it was created a matrix [function in R: *model.matrix()*], which contained all the numeric and factors of the data. For the levels of the factors this matrix created dummies data (of 0 and 1) in order to be easier their processing. Next, it was used LASSO (using "*glmnet*" package in R) and stepwise procedure of Akaike Information Criterion (AIC) in order to drop less significant variables, as the following 2.4 figure shows. So, it was formed an initial model.



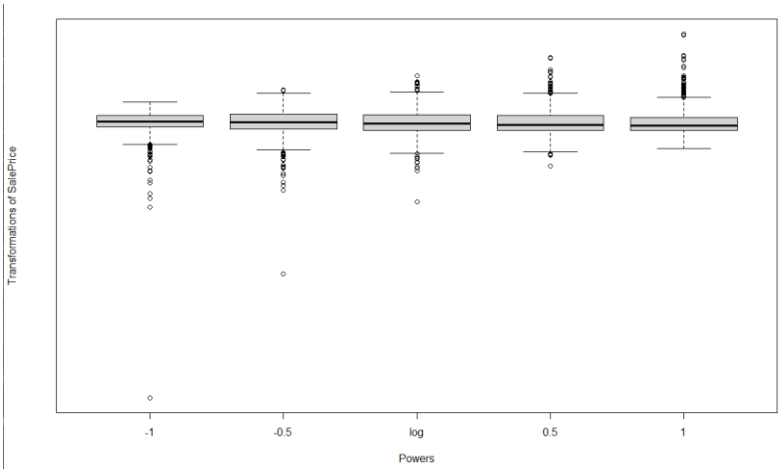
2.4 LASSO Plots: In the left is plot with coefficients and logarithmic (log) lambda (λ) and in the right is a plot with mean sq. error (MSE) and log(λ)

After that, it had to be checked for the residuals' assumptions (normality, homoscedasticity, constant variance, independence and linearity). Because all the above assumptions were violated, was made a lot of transformations to end up, with bootstrapping, to this final model:

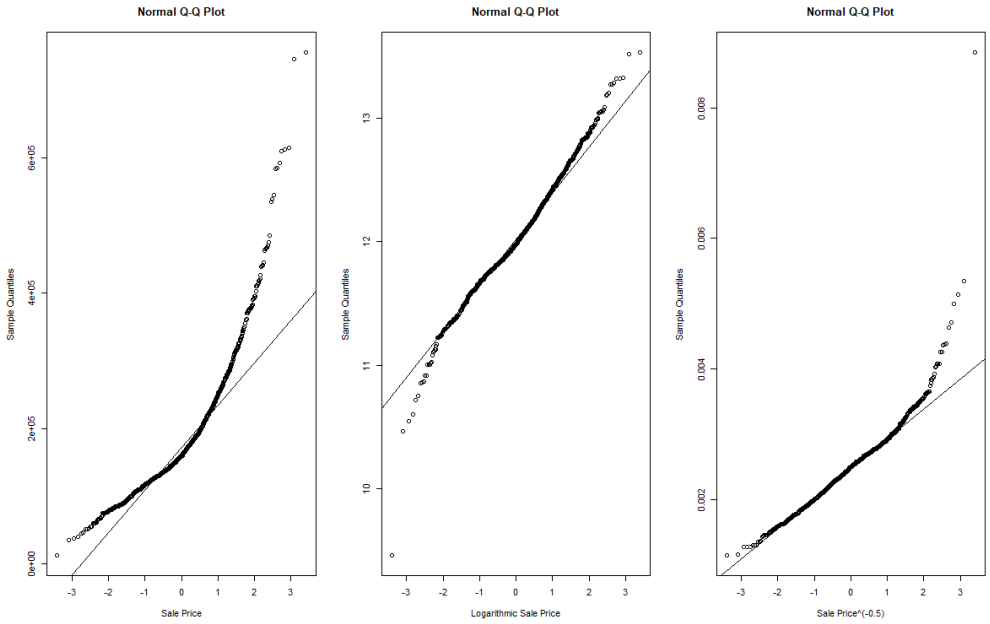
$$\log(\text{SalePrice}) = 12{,}03 + 0.15*\text{Garage.Cars} - 0.004*(\text{Garage.Cars})^4 - 0.13*\text{MS.ZoningRM} + 0.19*\text{Overall.Qual8} + 0.32*\text{Overall.Qual9} - 0.26*\text{Overall.Cond3} + 9.53*10^{-4}*\text{Year.Built} + 4.31*10^{-3}*\text{Year.Remod.Add} + 2.7*10^{-4}*\text{Total.Bsmt.SF} + \varepsilon$$

with $\varepsilon \sim N(0, 0.21^2)$

For this linear model, the “Sale Price” was altered to logarithmic, in order to be normalized, as the 2.5 and 2.6 figures point us and again was used LASSO and AIC to drop further meaningless variables. Also, they had to be removed another few features that influenced a lot the residuals’ assumptions and it was added a fourth grade polynomial [I(Garage.Cars^4)] to ameliorate these violated assumptions.



2.5 Boxplots for Transformations to Symmetry



2.6 Comparisons of various “SalePrice” QQ Plots [simple, log and BoxCox Trasformation (λ = -0,5)]

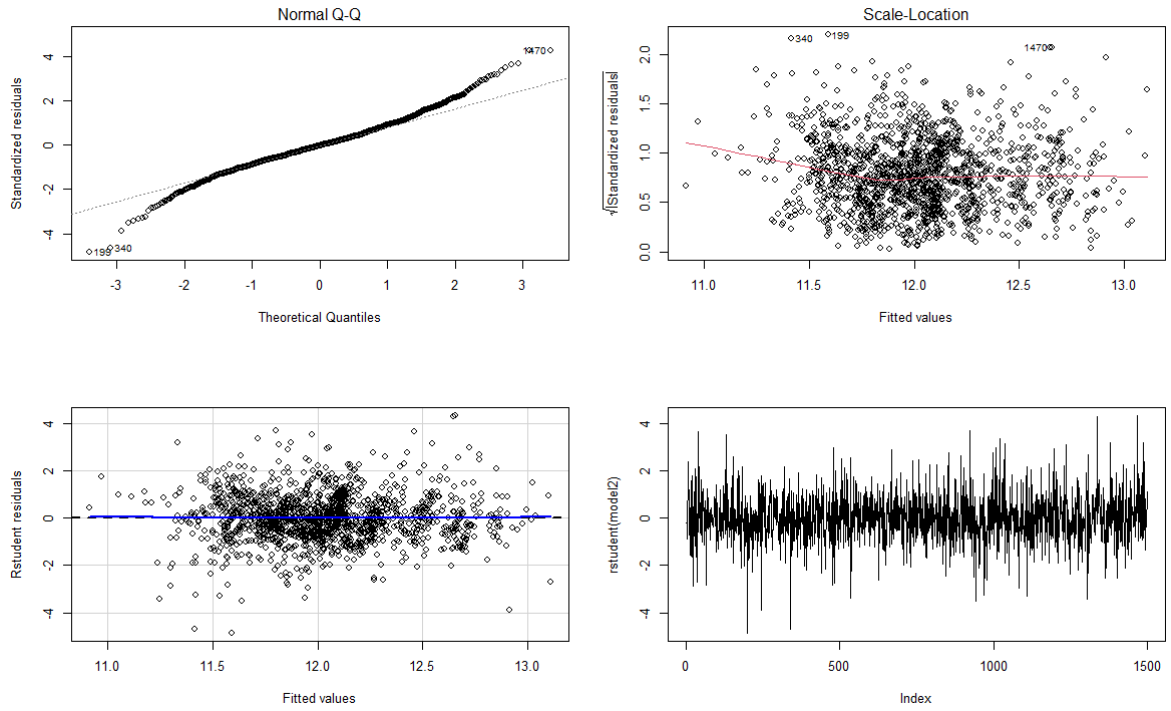
Finally, the variables were centered to zero, in order to make the constant variable significant and one outlier had to be removed, because it affected a lot the accuracy of our model. This outlier was the number Order No 182 with PID 902207130. It was a

property extremely old (1923 year of built) that was renovated in 1970 and was located on the first floor. Its overall condition and quality were very poor and the fact that this transaction, which took place in 2010, were done under abnormal conditions (trade, foreclosure or short sale) and in the price of only 12,789$, made it an influential point.

This final model fits very good to the data (Multiple R^2 = 74.8% > 70%) and in comparison, with other models fulfills the majority of the residuals' assumptions (see table 2.7 for more details and the figure 2.8 for visualization of the above).

Residuals' Assumptions	**Tests**	**p-value**	**Significance Level (α)**	**Null Hypothesis**
Normality of Unstandardized	Kolmogorov Smirnov	5.06×10^{-8}	0.05	Reject
	Shapiro Wilk	4.02×10^{-13}	0.05	Reject
Normality of Standardized	Kolmogorov Smirnov	3.86×10^{-8}	0.05	Reject
	Shapiro Wilk	3.34×10^{-13}	0.05	Reject
Normality of Ext. Studentized	Kolmogorov Smirnov	3.06×10^{-8}	0.05	Reject
	Shapiro Wilk	2.22×10^{-13}	0.05	Reject
Homoscedasticity	Non Constant Variance Score	0.07	0.05	Do not Reject
	Levene's	0.23	0.05	Do not Reject
	Fligner-Killeen	0.64	0.05	Do not Reject
Independence	Runs	0.18	0.05	Do not Reject
Non-Linearity	Tukey	0.63	0.05	Do not Reject
Multi-Collinearity	Variance Inflation Factors (VIF)	All the variables are uncorrelated because the VIF of each of them is smaller than 10.		

2.7 Residuals' Assumptions



2.8 Graphs of Normality, Homogeneity, Non-Linearity and Independence of the Final Model

This model is also really in prediction both with train sample and test sample (for details see 2.9 and 2.10 tables). As far as the first sample is concerned, this model has very good fitness (R^2 > 0.7), which means that can predict the prices with accuracy. Specifically, this measure means that nearly 74% of the variability can be explained by the model. Moreover, both the other two indicators are very low (RMSE = 0.21 & MAE = 0.15 both are lower than 0.5) which means that this model is really good predictor of the response (SalePrice). These measurements show the average absolute/squared differences between predictors (model's coefficients) and actual observations. The low values indicate that the firsts do not differ a lot from the seconds and especially MAE shows that the prediction data are only 0.15 away from their actual value.

Finally, after conducting various alterations in the test sample, the out-of sample predictive ability of our model is not very good. This conclusion is drawn from the fact that the measurements' values are too big, as 2.10 table indicates. In other words, the predictors are not very similar with actual values of the population. Although, this realization is very discouraging, the final model generally is better than all the other models that have been tested and that is why it is the most accurate of all (see paragraph 3 for details).

Cross Validation	Root Mean Squared Error (RMSE)	R Squared (R^2)	Mean Absolute Error (MAE)
Leave-One-Out	0.21	73.7%	0.15
10-fold	0.21	73.9%	0.15

2.9 Prediction Measures with Train Sample

Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)
$1.78* 10^5$	$3.74*10^{10}$	$1.93*10^5$	$7.71*10^3$

2.10 Prediction Measures with Test Sample

To sum up, taking into consideration all the above analysis, the interpretation of the final model is the below:

- If we add an additional car in the garage (Garage.Cars) there will be 16.2% increase of the expected price.
- When from 2 cars we add 6 (2^3 = 8 in total), the approximate difference between the initial and the final sale price will be nearly 20,4% decrease (Garage.Cars^4).
- If we choose a resident with medium density (MS.ZoningRM) there will be 12,2% decrease of the expected price.
- If we choose a property with an overall very good quality (Overall.Qual8) there will be 20.9% increase of the expected price, but if we select an excellent overall quality (Overall.Qual9) the increase will be 37.7%.
- The youngest the property is, the more expensive it becomes and to be more exact it will have, this one year, an impact of 0.095% increase in the expected price (Year.Built).
- The same applies to the year of property's remodelling. In other words, an extra year of remodelling (Year.Remod.Add) has an impact of 0.43% increase in the expected price. If there is not a remodelling in the house this variable is equal to the previous one (Year.Built).
- One more sq ft of the basement of the house (Total.Bsmt.SF) will have 0.027% increase in the expected price.
- Finally, if all the features and generally all the variables of a house are equal to 0, then the expected price will be 167,711.41$. This is also interpreted as the fixed costs that a typical property has.

The typical profile of a property is a combination of what was described in paragraph 2.1 and if we centre all the variables to the to mean (with function *scale()* in R). If we do that, the expected Sale Price of a typical house when all the independent covariates are equal to the sample mean is nearly $0.34* 10^{-10}$$.

3. Conclusions and Discussion

In this paragraph we will discuss the problems that was encountered through the regression analysis and it will be justified the choice of our final model. The model which was selected for price prediction is the best from others for multiple reasons. First of all, it covers all the residuals' assumptions, minus the normality. This means that:

1. The fulfilment of independence has a huge effect on the performance of the model and means that our data are not correlated. That is why, this makes more robust the model and with this, we can be sure that each covariate is not affected from the changes of others. This fact is confirmed also by the lack of multi-collinearity. This term if exists, the estimates will be unstable and the standard error will get higher than its normal value.
2. Homoscedasticity shows that estimators of coefficients are biased, which means that the expected value of the coefficients is really close to its true value and generally the value of the population (of all the properties which are been sold). Also, with homogeneity we can be sure that the error variance and the standard errors are estimated correctly, and this enhances the accuracy and the performance of our model.
3. The non-linearity shows that our model is very good and appropriate for prediction. In addition to homoscedasticity. and non-linearity provides the information that error variance is constant, which means that our model is specified well and the finally predicted price is very near to its true value.

It has to be clarified, that none of the models which were tested were near to these assumptions.

Furthermore, as 3.1 and 3.2 tables depict, our final model has not as low valued error indicators as the other models, which were tested. To begin with, these two tables

show the R^2 and some error measurements, both with training sample and with test sample.

- The first model is the null including all the numeric and factors, after excluded some covariates (with LASSO and AIC),
- the second model is the initial logarithmic model, with same variables as the null,
- the third is one good model that was tested as alternative to our final and
- the last model is what was selected in the end.

The third model that was close to the final has the below form:

$$\log(\text{SalePrice}) = 12.02 - 0.09*\text{MS.ZoningRM} + 0.12*\text{Overall.Qual8} + 0.22*\text{Overall.Qual9} - 0.29*\text{Overall.Cond3} + 0.0022*\text{Year.Built} + 0.0035*\text{Year.Remod.Add} + 0.19*10^{-3}*\text{Total.Bsmt.SF} + 0.32*10^{-3}*\text{Gr.Liv.Area} + 0.07*\text{Fireplaces} + 0.063*\text{Garage.Cars} - 1.66*10^{-11}*(\text{Gr. Liv.Area}^3) - 0.02*\text{Garage.Cars}^2 + 3.63*10^{-5}*\text{Year.Remod.Add}^2 + \varepsilon \text{ with } \varepsilon \sim N(0, 0.15^2)$$

Model	Cross Validation	Root Mean Squared Error (RMSE)	R Squared (R^2)	Mean Absolute Error (MAE)
Null Model	Leave-One-Out	27,857.64	88.5%	17,167.79
	10-fold	27,156.18	88.9%	17,328.05
Initial Log Model	Leave-One-Out	0.15	86.7%	0.10
	10-fold	0.15	86.9%	0.10
Random Log Model	Leave-One-Out	0.16	85%	0.11
	10-fold	0.16	85.5%	0.11
Final Model	Leave-One-Out	0.21	73.7%	0.15
	10-fold	0.21	73.9%	0.15

3.1 Comparing Prediction Measures with Train Sample

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)
Null Model	$1.88*10^4$	$2.15*10^9$	$4.63*10^4$	0.11
Initial Log Model	$1.78*10^5$	$3.74*10^{10}$	$1.93*10^5$	$1.46*10^4$
Random Log Model	$1.77*10^5$	$3.73*10^{10}$	$1.93*10^5$	$1.06*10^3$
Final Model	$1.78*10^5$	$3.74*10^{10}$	$1.93*10^5$	$7.71*10^3$

3.2 Comparing Prediction Measures with Test Sample

As it was mentioned at paragraph 2.3 this model as far as the out of sample prediction is concerned, is not very good, in comparison with others. Table 3.1 shows that the initial log model or the one that presented above, have lower error indicators than the final model, which mean that maybe they will predict better the final price. Despite this, the fact that they do not comply with the residuals' assumptions mean that these measurements possibly are wrong, and their true value will be much higher than our final model. This is also confirmed and in the 3.2 table, where our final model 's measures have the biggest values than the other three formulas. According to the value of the RMSE, the predictions for Sale Price made by the linear regression model are probably off by about 193,337.7$. This is approximately more than the double of the Sale Price's standard deviation (sd) in the training data (sd = 82,073.28$). To sum up, if we take into consideration all the outcomes, that we have been confronted, the model, which was final selected, is the best.

For this model selection, there had to be overcome many problems. To begin with, there were a lot of influential points (in total 90) that had to be checked. Because of their big number I ended up excluding only the most extreme, which influenced much the residuals' assumptions. Next, after trying to test a lot of models and a lot of different ways of presenting the linear model, none of the residuals' assumptions could be fulfilled. Also, there was a problem with big values in error indicators as it was shown before. That is why, I ended up with the model, which is moderately good, but fulfils the majority of the regression analysis conditions.

Based on those that were described above there a lot of things that they have not been covered, because they were not included in the purpose of this report. This linear model can be further investigated, and it would be a good idea to focus more on the test data and in the out of sample predictive ability of the model. Despite the fact that in the field of real estate the sale price prediction is very difficult, more efforts should be done in order to generalize the results that have been found. This field has many parameters that constantly change and on the most occasions the criteria are objective. That is why, is really difficult and needs a lot of struggle to fulfil all the regression analysis conditions, but efforts must be made to ameliorate the current situation. There are many open issues that need to be checked thoroughly, such as the outliers, both the leverages and influential points, and a deeper descriptive analysis need to be made. Finally, is of paramount importance and it must be done further research on this

issue, on predicting housing sale price, and especially concerning different locations. With this will be created a more general view and the regression analysis will become more accurate.

References

BOOKS

- Faraway, J. J., (2002). *Practical Regression and Anova using R.* 3rd Edition
- Fox J., Weisberg S., (2011). *An R companion to Applied Regression. 2nd Edition. University of Minnesota. Minnesota*

SCIENTIFIC ARTICLES

- Carmines E., Stimson J., Zeller R. (1978). *Interpreting Polynomial Regression*. 11.

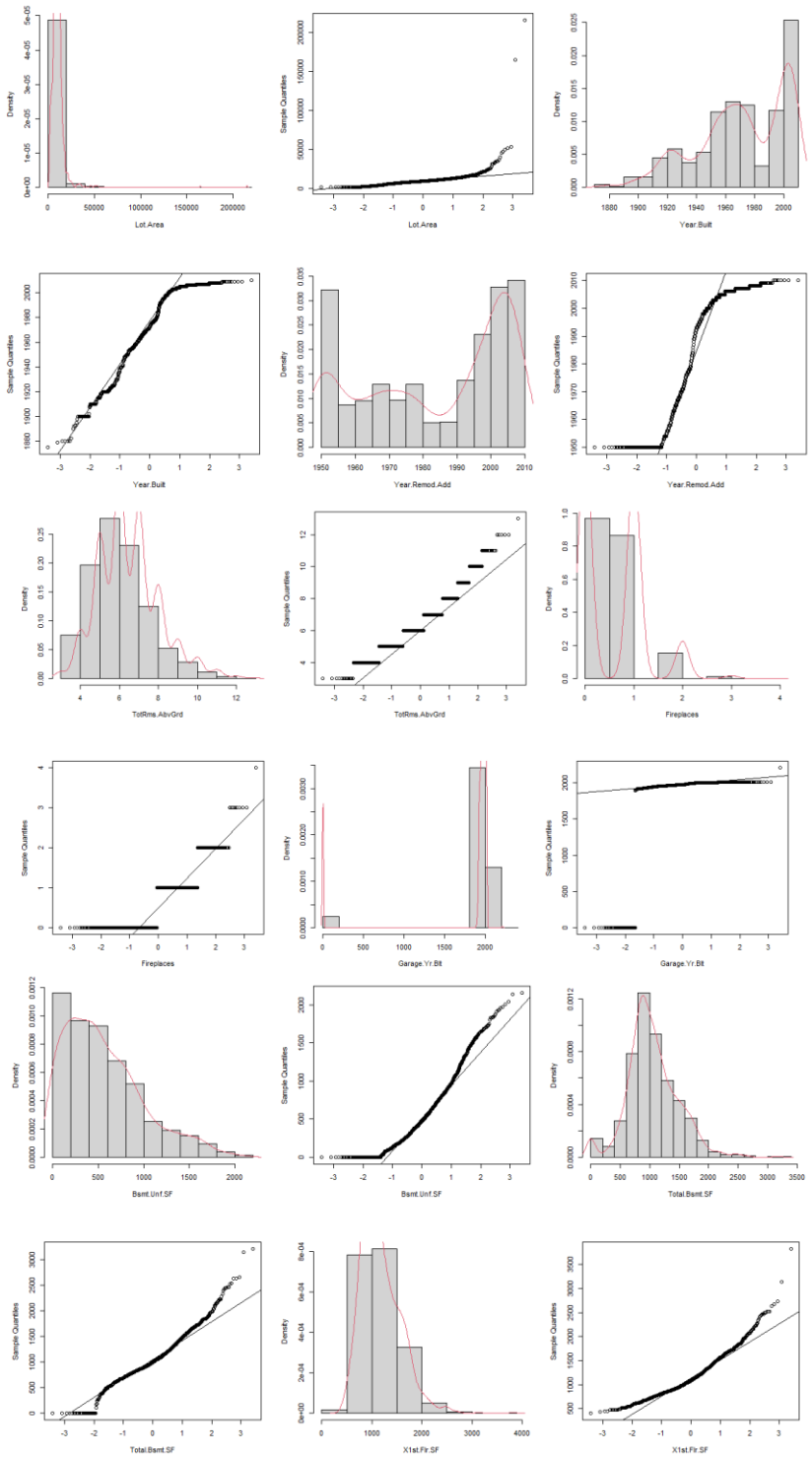
ONLINE SOURCES

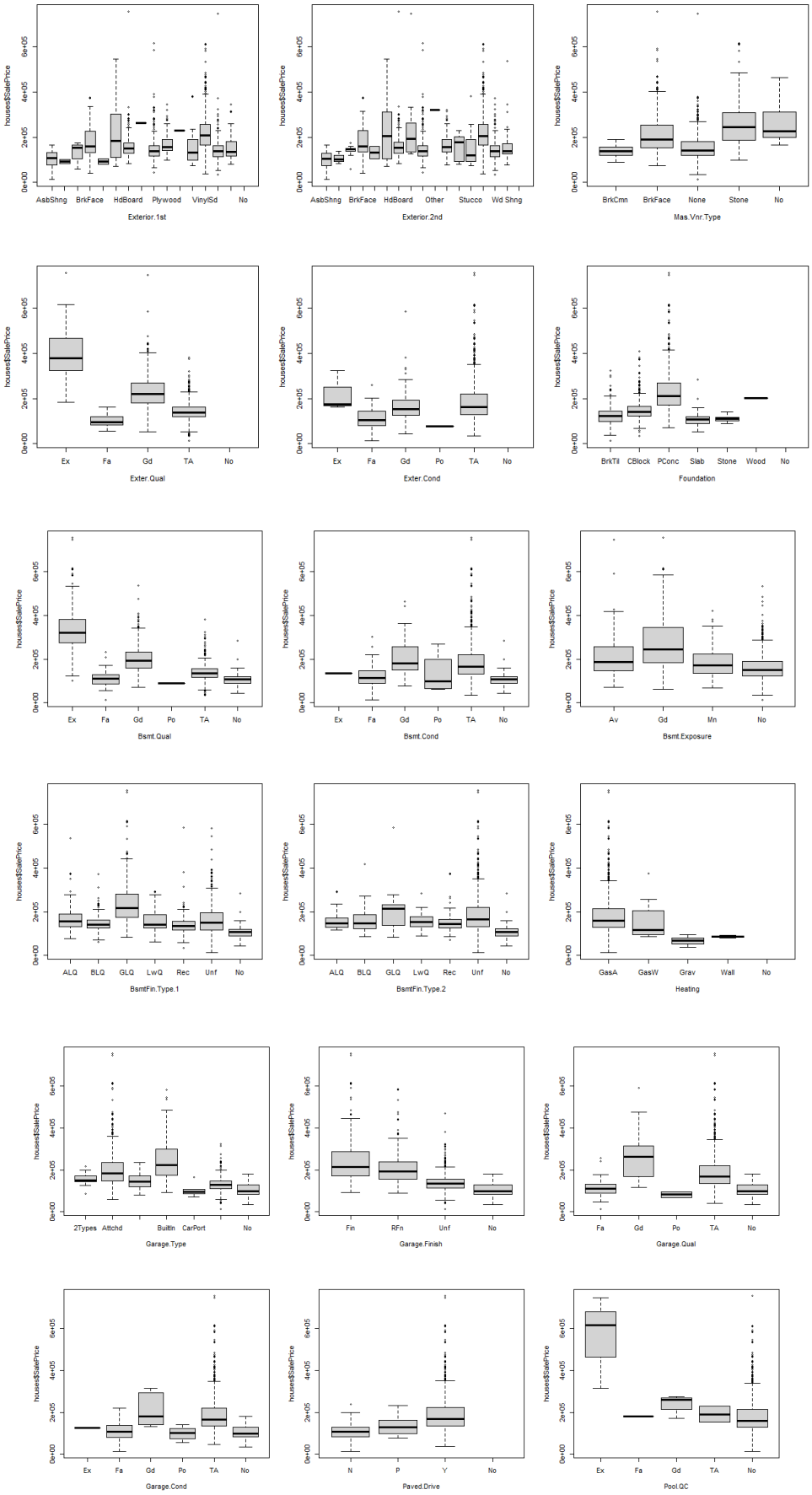
- Cock D.D, (2011). Ames Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. URL: www.amstat.org/publications/jse/v19n3/decock.pdf. Access Date: 14th December 2020
- Hastie T., Qian J., (2014). *Glmnet Vignette*. Sranford University. URL: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html. Access Date: 16th December 2020.

Appendix

In this part of the report will be presented some figures that have been drawn for the purpose of the above regression analysis.

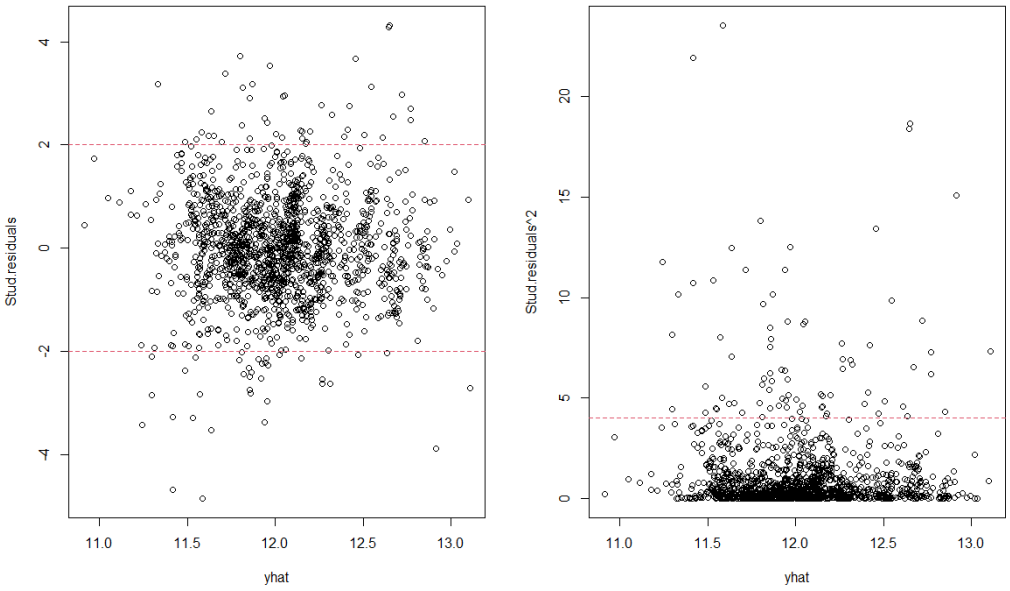
- Descriptive Analysis: here are few plots both for numeric variables and for factors, that descriptive analysis gave us.



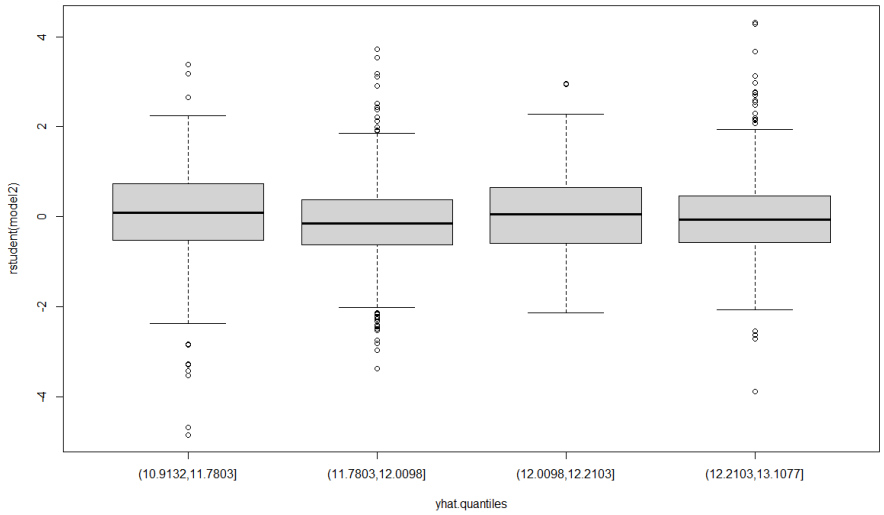


1. Histograms for Numeric Variables and Boxplots for Factors

- Regression Analysis



2. Plot of Studentized Residuals against Fitted Values in Final Model



3. Plot of Final Model's Residuals Quantiles