ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

BUSINESS
ANALYTICS
Master of Science

Department: Department of Management Science and Technology

M.Sc.: Business Analytics (Full-Time)

Course: Statistics for Business Analytics I

Professor: Ioannis Ntzoufras

Deliverable: Main Assignment (Multiple Regression)

Student Name: Evangelos Lakkas-Pyknis

Registration Number: f2822306

E-mail: eva.lakkaspyknis@aueb.gr

Exploratory Data Analysis (EDA)

Before the presentation of the prediction model, the way that was fitted in my data and its interpretation it is useful to make an analysis and some visualizations of the data in order to gain insight on the values of the most important variables and the associations that they have with each other. This process can be separated into two smaller parts the univariate analysis (analysis which considers only one variable each time) and the bivariate analysis (procedure which considers more than one variables each time).

Univariate Analysis

Let's start analyzing the variable that I want to predict, the price of sales (SalesPrice). In the histogram given below we can see that price deviates from normal distribution and is right-skewed, also the vast majority of the accomplished sales have a price between 0 and 200,000$, while really few houses were sold at a price greater than 400,000$ (25 exactly). The mean price is 180,550$ and the median price is equal to 160,000 (this difference between the mean and the median is another indication that the price doesn't follow the normal distribution).

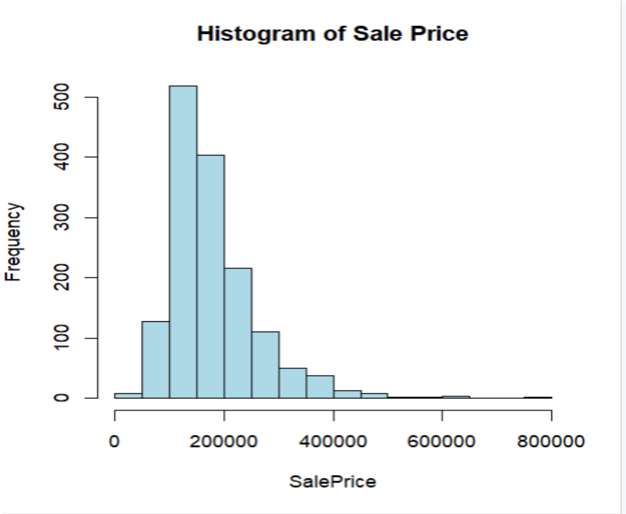


Figure 1

An important characteristic for the price of a house as expected is the year that it has been built, again we notice that this variable isn't normally distributed as the histogram is left-skewed (because there is an increase on the construction of houses almost for every time period). The average house has been constructed in 1972 and the most houses have been built during the periods 1950-1970 and 1990-2000.

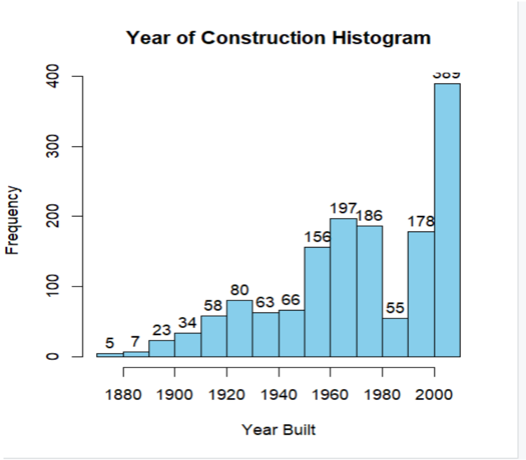


Figure 2

Two more numeric variables that have been proven significant later on my model but are also expected from common logic to affect the price of a house property are Garage Area and Gr.Liv.Area(above the ground living area square feet). Both of these variables don't seem to be normally distributed, garage area is found on a range 0-1500 sq.feet and we can see that most houses have a garage space near 500-600 sq.feet on the same time the average house has a garage of approximately 475 sq.feet. As for the Gr.Liv.Area variable we notice that its mean value is equal to 1493 and differs significantly from the median unlike Garage Area, on top of that most houses have a ground living area between 1000-2000 sq.feet.

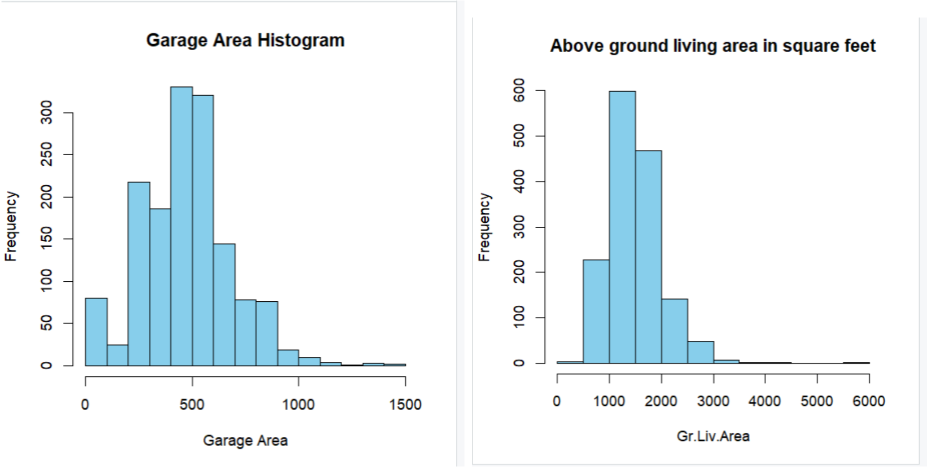


Figure 3: Garage Area-Gr.Liv.Area Histograms

Considering the categorical variables of the dataset the two that seemed to be the most important were the exterior quality of the house (meaning the quality of the material that was used for the outside of the house) and the kitchen quality. In the barplots below we see that the most common level for these variables is the "TA" which stand for typical/average (60% of the sold houses are characterized by this level for external quality and 50% for kitchen's quality), note that these two variables have ordered levels starting from Po/Poor (which doesn't characterizes any of these features in the sample) and ascending to Ex/Excellent (which exists rarely). The vast majority of the houses have kitchen's and exteriors quality at least in an average level and above.

Later on I will also examine the pairwise associations between these two variables and the price.

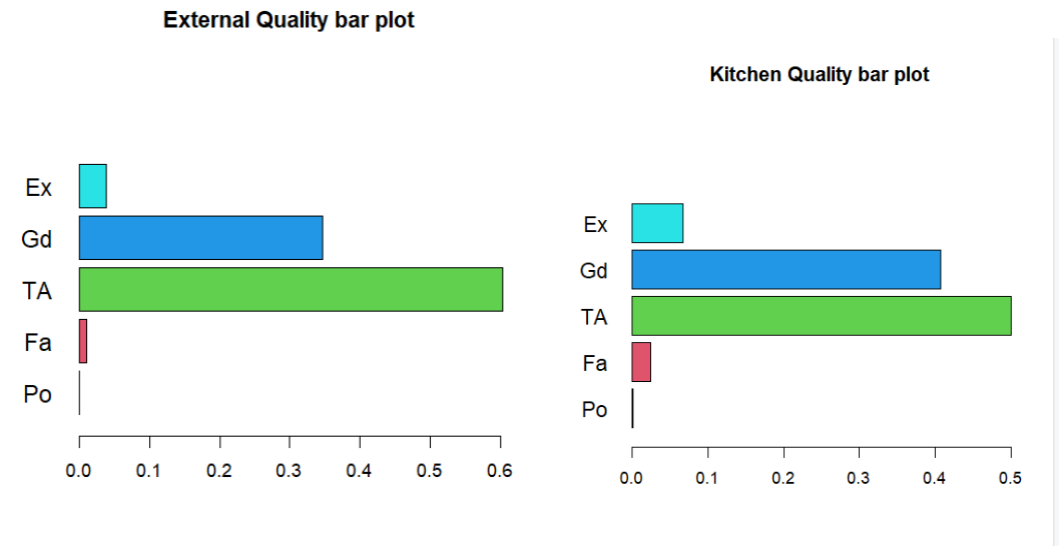


Figure 4: Exterior's and Kitchen's Quality bar plots

Bivariate Analysis

In the table below we can see the significant correlations between price of sales and other numeric variables[1]. We notice that all of them have a positive relationship (meaning that increase in one of these variables can cause an increase in the price). We see that the price has a stronger relationship with the variables that refer to garage features, the construction year of the house and mainly the living area above the ground, so an increase in these variables can have a stronger effect on the response variable.

Lot Area	Year Built	Year Remod.	Mas.Vnr.Area	BsmtFin SF 1	Total Bsmt SF	Garage Area
0.26	0.56	0.52	0.55	0.41	0.62	0.64
Gr.Liv.Area	**Full Bath**	**Total Rooms. Above Groud**	**Fireplaces**	**Garage Year Blt.**	**Garage Cars**	**Wood Deck SF**
0.71	0.55	0.50	0.46	0.54	0.65	0.33

Table 1: Sales Price correlations

As for the associations between price and categorical variables, I examined the variables kitchen quality and exterior quality. Firstly I checked if normality of the residuals of a model that has price as a dependent variable and exterior quality or kitchen quality as a independent variable can be assumed for all the levels of these variables, that hypothesis is rejected for both exterior quality (Lillie test, p-value$<2.2*e^{-16}<0.05$ and Shapiro-Wilk test, p-value$<2.2*e^{-16}<0.05$) and kitchen quality (Lillie test, p-value$<2.2*e^{-16}<0.05$ and SW test, p-value$<2.2*e^{-16}<0.05$). After conducting tests to find if there is significant differences for price between the levels of each factor this hypothesis cannot be rejected for exterior quality (Kruskal-Wallis test, p-value$<2.2*e^{-}$

[1] A table with the correlations between price and all the numeric variables of the dataset can be found in the appendix at Ap.Table 1

$^{16}<0.05$) and kitchen quality (Kruskal-Wallis test, p-value$<2.2*e^{-16}<0.05$)[2]. So given that there is big differences for price between the levels of the quality of these features is an indication that these two factor can play a role in the differentiation of sales price. As for the question which groups have significant differences, in exterior quality we observe that all the groups have significant differences with each other and in the kitchen quality the only combination of levels that doesn't have a significant differences is the Fair-Typical/Average (the results of Tukey HSD test can be found in the appendix in Ap.Table 2 and 3, in these results a p-value smaller than 0.05 indicates a significant differences of Sales Price between the levels of each variable).

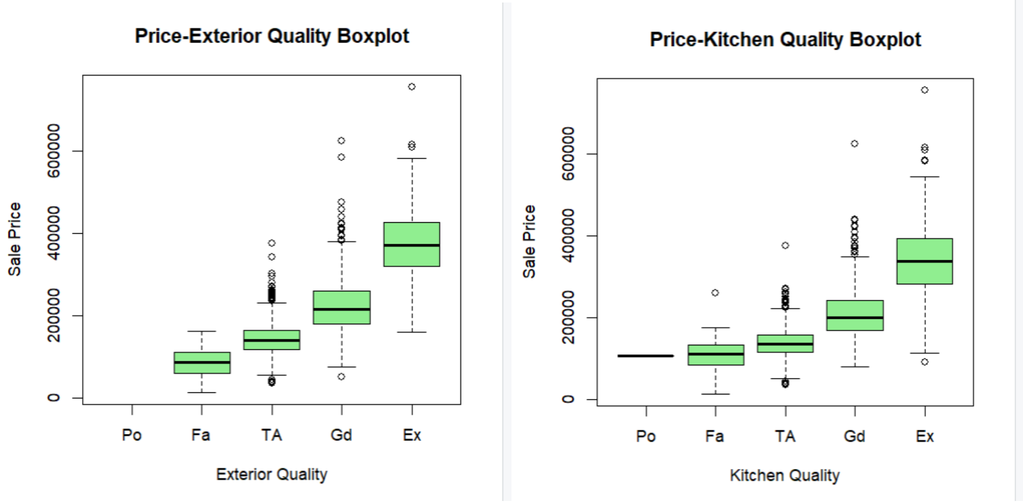


Figure 5:Price-Exterior/Kitchen Quality Boxplots

These differences among the levels of these variables are also appeared in the boxplots above. We can see that the (median) price rises as the quality of the kitchen gets better also there are several price outliers for the levels Fair and Average that indicate that a house can be sold in a higher price than expected for these two levels. For the exterior quality we notice that the (median) price rises as explained before we also see the significant difference between Fair and Average (which doesn't exist in the kitchen quality) and again several outliers (the only level that hasn't an outlier).

Model Selection

The starting point for the fitting of the model in my sample dataset was the construction of the full multiple linear regression model (e.g. using all the available 80 variables that existed in the dataset as predictors of the response variable, Sale Price). However, it is obvious that a model with these many variables is not useful because its interpretation is not easy at all and meaningful and there is the danger that if a prediction is based on this type of model when a different dataset is available with less variables, we will not be able to use the full model. So in order to simplify the model and select the appropriate variables I followed the Least Absolute Shrinkage and Screening Operator (LASSO) method that leads to the subtraction of 29 variables from the full model (here as a parameter of lambda I used the "min" value which may not get rid of

[2] Here a non-parametric test was used because normality of the residuals cannot be assumed

as many as possible variables but I wanted to have a more conservative approach at first steps in order to not lose much of the predictive ability of the model due to a sudden subtraction of predictors). The selected variables from the LASSO can be found in the appendix in Ap.Table 4.

So, the second model that was created had 51 variables which means that a more simplified model has to be found. Also given that adjusted R-squared value of this model is equal to 94% and many variables are statistically insignificant (meaning that they don't add much information in the model) I can "sacrifice" more variables to find a better model). In order to do that I implemented a backward procedure using Akaike Information Criterion to select variables[3] two times and in between I removed the statistically insignificant variables (meaning that in the summary of the model had a p-value>0.05). The first backwards procedure led to a score equal to 30038 and the second one ended with 29763. After that I checked the model that has been created so far, for multicollinearity (that can lead to increased standard errors) using the variance inflation factor (VIF) which gave me the results shown below.

Variable	GVIF	Variable	GVIF
Lot Area	1.6	Bsmt.Fin. SF 1	2.8
Land Contour	2.4	Bsmt.Fin. SF 2	1.2
Neighborhood	1262.0	Total.Bsmt.SF	2.5
Bldg.Type	16.6	Low.Qual.Fin. SF	1.2
Overall Cond.	7.25	Bsmt. Full Bath	4.5
Year Built	8.64	Full Bath	2.0
Year Remod.	3.12	Bedroom Abv.Gr	2.3
Roof Mat.	3.0	Kitchen Abv.Gr	3.8
Exterior 1st	31.5	Kitchen Qual.	8.3
Exter.Quality	9.3	Fireplaces	1.7
Garage Cars	6.5	Sale Type	151.4
Garage Area	6.1	Sale Condition	146.7

Table 2: Multicollinearity table

Based on these results I subtracted some variables that can cause a problem and have a GVIF value larger than 10, that variables were Neighborhood, Garage Cars, Exterior 1st, Sale Type and Bldg.Type. The removing of these variables is most likely not to be an issue even if they were statistically significant because if we need we can get the information we need from the other variable that was collinear with the subtracted (for example in the model created so far sale type was significant at some factor levels but if we subtract it we can still get some information from sale condition that is a similar variable). After these removings the multicollinearity problem was fixed (the table with

[3] Here I preferred to use AIC over BIC (Bayesian Information Criterion) because the first is generally better for predictive models and the second for descriptive models

the new VIF values of the newly constructed model that is referred below is shown in the appendix in Ap.Table 5).

The model that remained after the removing of the variables mentioned above had a more reasonable number of features compared to the previous models but it is still over-parameterized more than twenty variables so it had to be simplified again. Trying the AIC method hadn't an effect this time so I implemented the LASSO[4] procedure once again (its results can be found on the Ap.Table 6) and after that I removed one categorical variable that wasn't statistically significant for the most of each levels (Kitchen Quality) the multicollinearity was once again not a problem. Now we have one model with 10 variables (8 numeric and two categorical), this model fits well to the training dataset (adj. $R^2 = 0.86>0.7$, this means that 86% of the variance of sale price is explained by this model)[5] and has a residual standard error equal to 29720. Now I must check the assumptions of this model.

Assumptions

The assumptions that must be checked is the normality of the residuals, the constant variance of the residuals and the linearity. The residuals can not be assumed to follow a normal distribution for this particular model (Lillie test, p-value$<2.2*10^{-16}$ and SW test, p-value$<2.2*10^{-16}$) this can also been seen in the QQplot below where we see that the residuals are not aligned with the QQline and it deviates from the normal distribution at the tails.

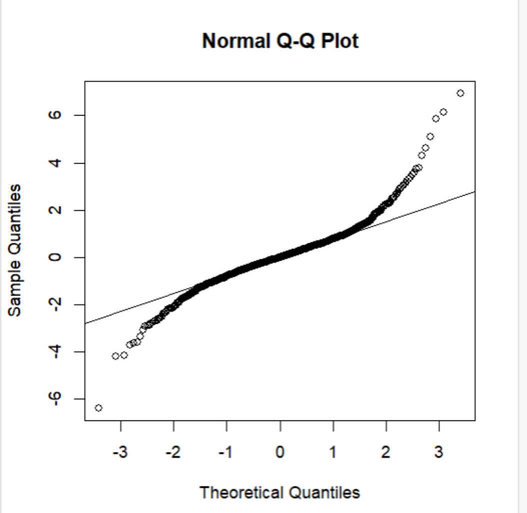


Figure 6:Residuals QQplot

Constant variance assumption is also rejected (Levene's test, p-value$<2.2*10^{-16}$ and non-constant variance test p-value$<2.2*10^{-16}$). The homoscedasticity assumption can also be rejected from the Ap.Figure 1 in the appendix we observe that the many points of the plot are over the dashed line and that they are heavily concentrated on the right. The linearity assumption is also rejected (Tukey test, p-value $<2.2*10^{-16}$) in the

[4] This time I used "1se" as a lambda parameter in order keep as less variables as possible

[5] I consider that 70% is a limit for a sufficient goodness of fit

figure shown below we observe that the blue line is curved and not aligned with the line passes through zero this indicates for a non-linearity problem in my model.

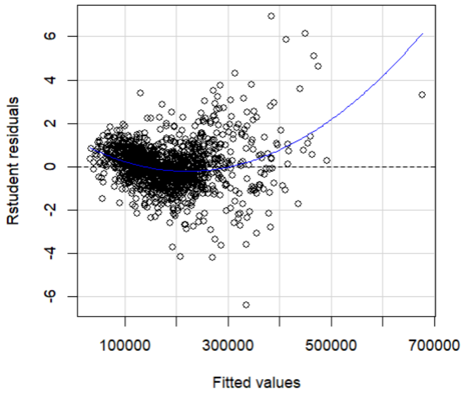


Figure 7

The fact that these assumptions don't hold can lead to problems because if the residuals are not normally distributed and the variance is non-constant the model can be led to calculate non-accurate standard errors and thus decrease its performance. Also linearity is very important for a model to be assumed credible because if there is non-linearity endogeneity is implied and thus no causal interpretation can be drawn from the model. So, I had to make some transformations in the model to try and fix these assumptions. First of all I putted logarithms on the response variable and one predictor (Gr.Liv.Area) to try and fix normality. Unfortunately even though the QQplot got a little better (on Ap.Figure 2 we see that the "tails" on the left are still heavy on the right is now more aligned than before) the normality is still rejected[6] (Lillie test, p-value<$2.2*10^{-16}$). I also added a polynomial term of third degree for garage area that transformation of the model led to the satisfaction of the linearity assumption (Tukey test, p-value=0.84>0.05) this is also shown in the figure 8 below. Notice how the blue line has no slope and is almost completely aligned with zero.

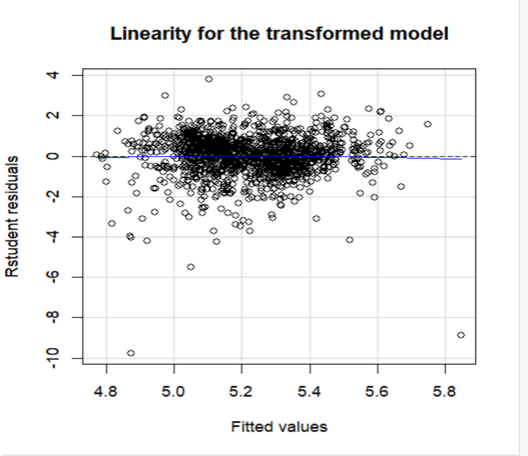


Figure 8: Linearity for the transformed model

[6] Unfortunately, every transformation with logarithms that I tried didn't get me to model with normally distributed residuals.

Constant variance is also rejected (Levene's test, p-value=5.9*10^{-7}<0.05 and ncv-test, p-value=0.0007<0.05). The fact that now there is linearity in the model discharges the model of the endogeneity and causal interpretation can be accessed through the model although certain miscalculations in the standard errors can be made due to the lack of normality and homoscedasticity.

The mathematical formulation of the model is given in the equation below:

$$\begin{aligned}\log(Price) = 0.147 &+ 0.001 * YearBuilt + 0.001 * GarageArea - 0.25 \\ &* GarageArea^2 - 0.25GarageArea^3 + 0.22 * Ext.Qual.L - 0.01 \\ &* Ext.Qual.Q + 0.03 * Ext.Qual.C + 0.57 * \log(Gr.Liv.Area) \\ &+ e\,, where\ e \sim N(0, 0.084^2\ )\end{aligned}$$

Ext. Qual are dummy variables taking the values 0 or 1 and each one indicates a specific level:

Ext.Qual.L: Fair

Ext.Qual.Q: Typical/Average

Ext.Qual.C: Good

If all of those levels are 0 then Ext.Qual. = Excellent

The adj. R-squared is equal to 77% (>0.7) and it implies a good fit in the training dataset

Interpretation of the model

Intercept: If all the others variables are zero then the logarithm of the price is equal to 0.147. The intercept here hasn't an important interpretation because it is not very reasonable for a house to have all the covariates zero.

Garage Area: If two houses have all the other attributes common an one of them has one more sq.feet in the garage then that will add 0.001 to the log(price). The polyonymic degrees are difficult in the interpretation and cannot be explained easily.

Year Built: If two houses have all the other attributes common and one of them is one year newer that will cause 0.001 increase to the log(price).

Log(Gr.Liv.Area): If two houses have all the other attributes common an one of them has one more sq.feet in the ground living area then that will cause an increase of 0.57 to the log(price).

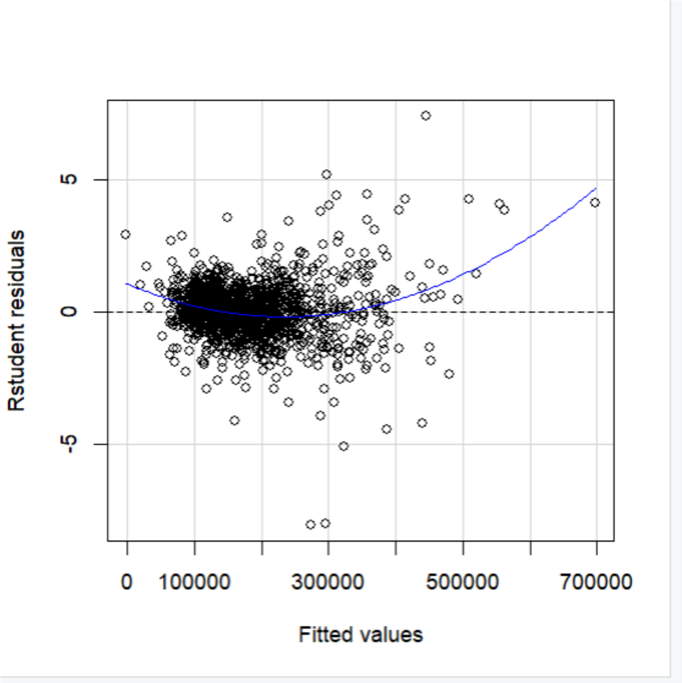
Ext.Qual: When two houses have all the other attributes common a change in the level of exterior quality will lead to an increase equal to the coefficient of the corrensponding level in the log(price)

Comparison with LOO method and predictive ability

As for the predictive ability of the I calculated RMSE for both training and testing dataset, and the fact that these two are almost equal suggests that the model doesn't overfit in the training dataset and has a sufficient predictive ability both datasets have an RMSE=6.8 which is also relatively low (close to 0).

I also compared my model with a model that came from the full model with the method leave one out. This model cannot be characterized better than my model. Because:

1) Doesn't satisfy linearity assumption which leads to problems that I mentioned before



2) Has a higher adjusted R-squared (=0.9) but is overparameterized
3) Also it has a higher RMSE on the training dataset (1996.6)