Advanced Statistics and Programming - Individual Assignment 1

Thanh Dat Nguyen - 532618tn

15 September 2022

Data Selection and Preparation

The dataset contains a total of 81 variables, including the dependent variable of interest - sales price of the house. For this analysis, I make a selection of 4 commonly used independent variables in house pricing models based on previous research (Sirmans, Macpherson, & Zietz, 2005) - Lot Area, Year of Construction, Garage Size and Type of Dwelling. Summary tables are shown below:

Table 1: Summary Table - Quantitative Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
LotArea	1,460	10,516.830	9,981.265	1,300	7,553.5	11,601.5	215,245
YearBuilt	1,460	1,971.268	30.203	1,872	1,954	2,000	2,010
GarageArea	1,460	472.980	213.805	0	334.5	576	1,418

Table 2: Summary Table - Categorical Variables

1Fam	2fmCon	Duplex	Twnhs	TwnhsE
1220	31	52	43	114

There are 1460 observations in total. To provide a brief description of the chosen variables: LotArea gives the total lot size in square feet, YearBuilt is the year of the original construction of the building, GarageArea gives the size of the garage in square feet (all the values are integer values, LotArea and Garage Area are ratio variables, while YearBuilt is interval) and the Type of Dwelling denotes the type of building (nominal variables). There are five types in total - Single-family Detached houses (1Fam), Two-Family Conversion houses (2fmCon), Duplexes (Duplx), Townhouse End Units (TwnhsE) and Townhouse Inside Units (Twnhs).

We can observe multiple things from the summary tables. First off, the standard deviation for the variable LotArea is almost equal to its mean - this suggests a large variation in the lot sizes and a possible presence of extreme values. Comparing the variable's 75th percentile and its maximum supports this notion. Similar comments can be made about the GarageArea variable. Furthermore, we can see that the observations are mostly made up of Single Family Detached houses, which account for more than 80% of houses in the sample. To further our insights on the data, three plots are provided below:

Figure 1: Lot Area against House Price - by Dwelling Type

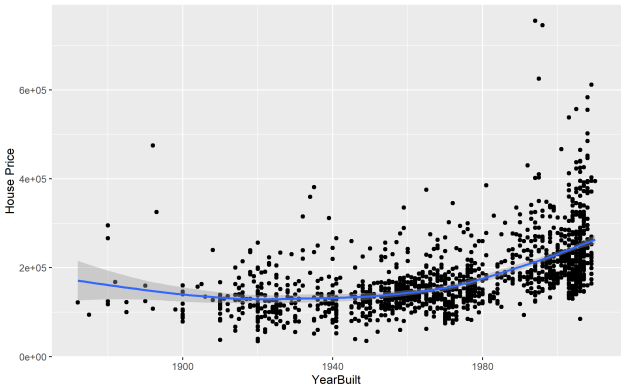


Figure 2: Year Built against House Price

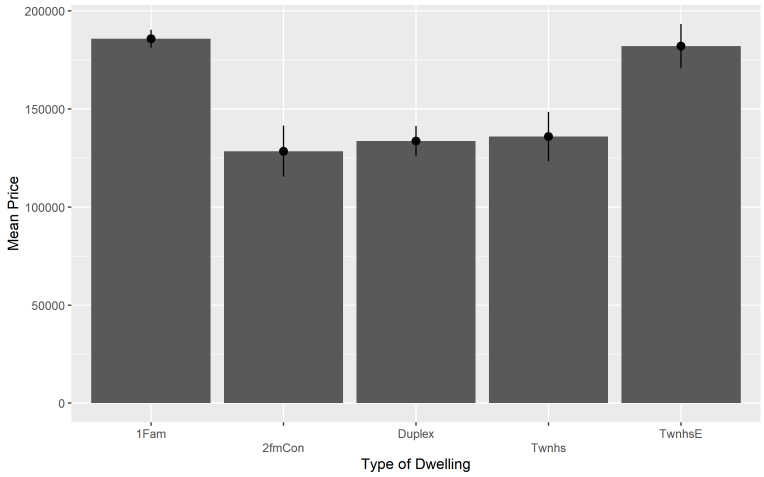
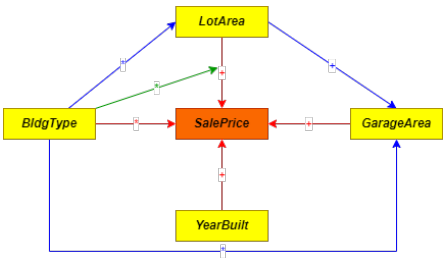


Figure 3: Average House Price Across Building Types

Theoretical Model and OLS Assumptions

Based on previous research and insights gathered from the above plots, I present the causal diagram, which is the basis for the complete model.



The following three hypotheses are made based on the causal diagram and the three plots above:

H1: Year of construction has a positive influence on the house price.
H2: Lot size has a positive influence on the house price.
H3: The positive influence of lot size on the house price is expected to be higher for Townhouse Inside Units than for the other house types.

The model can be represented by the below equation:

$$\begin{aligned} SalePrice = \beta_0 + \beta_{11} LotArea + \beta_{12} LotAreadType2 + \beta_{13} LotAreadType3 \\ + \beta_{14} LotAreadType4 + \beta_{15} LotAreadType5 + \beta_2 YearBuilt \\ + \beta_3 YearBuilt^2 + \beta_4 GarageArea + \beta_{52} dType2 + \beta_{53} dType3 \\ + \beta_{54} dType4 + \beta_{55} dType5 + \epsilon \end{aligned} \tag{1}$$

Note: dType1 = 1Fam, dType2 = 2FmCon, dType3 = Duplx, dType4 = TwnhsE, dType5 = Twnhs

As I am be using the ordinary least squares linear regression model, there are a number of assumptions my model has to abide by for its results to be reliable. There are six key assumptions:

A1: Linearity - The relationship between the dependent variable Y (in this case SalePrice) and independent variables X is linear. As could be seen in the formulation of the model, this may be violated by for the variables YearBuilt and LotArea. The scatter plot of YearBuilt against SalePrice suggests a more exponential relationship and the LotArea scatter plot with a fitted line above shows that the observations are over-predicted for higher lot sizes.

A2: Full rank - there are no exact linear relations among the independent variables. This may be violated by variables LotArea and GarageArea as it could be assumed that GarageArea will be larger as LotArea grows.

A3: Exogeneity of independent variables - the error term is independent of any variable X, and therefore, the expected value of the error term is zero conditional on the the variable X_i. This may be violated as the number of observations decrease as we get to higher values of GarageArea and LotArea and as there are a number of outliers present. This assumption may also be violated by omitted variables.

A4: Homoskedasticity - the conditional variance of the disturbance is constant. This may be violated for the independent variable LotArea, as from the scatter plot, we can see the disturbance is higher for the lower lot sizes and higher for the higher lot sizes.

A5: The data may be a mixture of constants and random variables. For the analysis, I assume this holds.

A6: Normal distribution of disturbance - the error term is normally distributed. This may again be violated due to the observations of extreme values for the quantitative variables.

OLS Regression and Model Fit

Now that the model has been defined, the analysis can be performed. Both the model with and without the interaction and non-linearity effects is run. The summary of the results is presented in Table 3.

Let us first examine the model without any additional terms. As can be seen the model has moderate explanatory power, with 49.9% of variation in house prices being explained by the chosen independent variables (the corresponding $F = 206.836$, $p < 0.01$). Furthermore, all of the chosen variables are significant ($p < 0.01$), with the exception of the Two-Family Conversion house type ($\hat{\beta}_{52} = -6573.452, p = 0.532$). This is in line with Figure 1, as the fitted line for Two-Family Conversion houses is relatively flat and there is a relatively low number of observations for this building type. The lack of significance can also be explained by looking at the standard error of the coefficient ($s_{52} = 10497.96$), as the error is higher than the coefficient itself, which is indicative of large variance in values.

Evaluating the model in terms of the stated hypotheses, the null hypothesis for H1 can be rejected as YearBuilt significantly affects the sales price of a house ($\hat{\beta}_2 = 876.510, p < 0.01$). As for the interpretation of the coefficient, while holding all other variables constant, one additional year of construction (in more understandable terms, a house newer by one year) is expected to increase the price of a house by 876.51 dollars. A similar interpretation of coefficients can be applied to the other quantitative variables - LotArea ($\hat{\beta}_{11} = 1.293, p < 0.01$) and GarageArea ($\hat{\beta}_4 = 157.200, p < 0.01$). Thus, the null hypothesis for H2 can also be rejected, as LotArea is a significant variable, whose one square foot increase will lead to an additional 1.293 dollars on the house price.

To make conclusions about H3, we need to look at the model, which includes the interaction and non-linearity terms. As can be seen, the model's explanatory power improved, with now 55% of variation in house prices being explained by the newly added terms ($F = 147,483, p < 0.01$). One thing we immediately notice is that the constant is now extremely high and positive ($\hat{\beta}_0 = 67987512, p < 0.01$), which is caused by the inclusion of the significant non-linear term $YearBuilt^2$ ($\hat{\beta}_3 = 17.09, p < 0.01$), which causes the original variable YearBuilt to have a significant negative coefficient ($\hat{\beta}_2 = -66103.84, p < 0.01$). Inspecting the table further also confirms that including the non-linear term was indeed the correct choice, as its inclusion increased the base model's explanatory power from 49.9% to 54.1% ($F = 214.055, p < 0.01$) (column 3). Looking at the complete model on the right and studying the interaction terms, it can be concluded that the H0 should not be rejected for H3, as for Townhouse Inside Units, the interaction effect is negative and insignificant ($\hat{\beta}_{55} = -5.498, p = 0.511$). From the table results, it is actually the Townhouse End Units, for which the positive influence on the lot size-house price relationships is the greatest ($\hat{\beta}_{54} = 8.33, p < 0.01$). This coefficient can be interpreted as following: for Townhouse End Units, an increase of square foot in the lot size yields a 10.003 dollar (1.673 + 8.33) increase in house price. The 1.673 value comes from the LotArea variable in the full model, which is the coefficient for Single-Family Detached houses (our base case in the model).

Finally, we determine which of the chosen independent variables have larger effect sizes. As the coefficients for all the variables are not dimensionless and have an associated unit of measurement (e.g. dollar per square foot for LotArea and Garage Area, dollar per year for YearBuilt), it is difficult to compare each variable's effects in the overall model. Therefore, standardized effects are used (column 5). As can be seen the highest contributing variables are $YearBuilt^2$ ($beta = 25.497$) and GarageArea ($beta = 0.361$), followed by BldgTypeTwnhsE ($|beta| = 0.211$) and LotArea ($beta = 0.210$).

Table 3: Building the Full Regression Model

	Dependent variable:				
	SalePrice				
	(Base model)	(Model - non-linearity)	(Model - interaction)	(Full model)	(Standardied coefficients)
LotArea	1.293***	1.466***	1.520***	1.673***	0.210
	(0.155)	(0.149)	(0.169)	(0.163)	(0.163)
BldgType2fmCon	−6,573.452	−19,181.960*	15,253.990	1,245.230	0.002
	(10,497.960)	(10,110.370)	(11,984.420)	(11,544.740)	(11,544.740)
BldgTypeDuplex	−39,348.950***	−27,492.400***	−29,100.570	−19,645.810	−0.046
	(7,990.521)	(7,719.160)	(25,493.450)	(24,433.830)	(24,433.830)
BldgTypeTwnhs	−33,069.040***	−32,673.180***	−37,277.600	−17,909.110	−0.038
	(8,997.397)	(8,614.465)	(22,668.710)	(21,779.870)	(21,779.870)
BldgTypeTwnhsE	−17,399.650***	−23,963.390***	−58,503.850***	−62,363.560***	−0.211
	(5,830.950)	(5,611.679)	(12,361.440)	(11,845.680)	(11,845.680)
YearBuilt	876.510***	−66,869.620***	880.097***	−66,103.840***	−25.132
	(59.652)	(5,874.976)	(59.149)	(5,850.106)	(5,850.106)
I(YearBuilt^2)		17.284***		17.090***	25.497***
		(1.499)		(1.492)	(1.492)
GarageArea	157.200***	135.126***	155.334***	134.159***	0.361
	(8.151)	(8.036)	(8.107)	(7.983)	(7.983)
LotArea:BldgType2fmCon			−1.455***	−1.344***	−0.079
			(0.396)	(0.380)	(0.380)
LotArea:BldgTypeDuplex			−1.029	−0.794	−0.019
			(2.491)	(2.386)	(2.386)
LotArea:BldgTypeTwnhs			2.475	−5.498	−0.030
			(8.687)	(8.350)	(8.350)
LotArea:BldgTypeTwnhsE			8.909***	8.330***	0.150
			(2.276)	(2.181)	(2.181)
Constant	−1,630,991.000***	64,743,759.000***	−1,639,705.000***	63,987,512.000***	
	(115,613.900)	(5,756,842.000)	(114,662.800)	(5,732,399.000)	(5,732,399.000)
Observations	1,460	1,460	1,460	1,460	1,460
R^2	0.499	0.541	0.509	0.550	0.550
Adjusted R^2	0.497	0.539	0.506	0.546	0.546
Residual Std. Error	56,349.840 (df = 1452)	53,951.140 (df = 1451)	55,853.940 (df = 1448)	53,501.830 (df = 1447)	53,501.830 (df = 1447)
F Statistic	206.836*** (df = 7; 1452)	214.055*** (df = 8; 1451)	136.688*** (df = 11; 1448)	147.483*** (df = 12; 1447)	147.483*** (df = 12; 1447)
Note:					*p<0.1; **p<0.05; ***p<0.01

Diagnostic Checking

Now that the model has been built and interpreted, it is time to check whether it is in line with the six standard assumptions of the linear regression model. In the model, the non-linear behaviour of the YearBuilt has been already accounted for so checks can be performed for the other two quantitative variables. Scatter plots of X against Y have been performed with inconclusive results so residual analysis can also be performed to see if any patterns in the data surface. No patterns showed themselves through this analysis.

Next, A2, the full rank assumption will be checked. Here, we are checking for multicollinearity between the chosen variables with the VIF value:

Looking at the very last column, we can see that most values are quite low (hovering betwen the ranges

Table 4: VIF Comparison

	GVIF	Df	GVIF^(1/(2*Df))
LotArea	1.347	1	1.161
BldgType	519.707	4	2.185
YearBuilt	15,912.660	1	126.145
I(YearBuilt^2)	15,948.900	1	126.289
GarageArea	1.485	1	1.218
LotArea:BldgType	562.133	4	2.207

of 1 and 2), with the only large VIFs being present for the YearBuilt variable and its transformation, which was expected. This test, therefore, does not suggest multicollinearity between the variables.

To perform a check on A3, exogeneity, we can make use of its definition and see whether the expected (mean) value of errors (conditional on variables X) equals zero. To do this, the following table in R was generated:

Table 5: Exogeneity Check

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
resid_lot	1,460	−0.0002	1.006	−3.953	−0.628	0.408	7.352
resid_gar	1,460	0.001	1.005	−4.550	−0.532	0.394	8.084
resid_year	1,460	0.001	1.005	−2.133	−0.606	0.335	8.198
resid_type	1,460	0.001	1.002	−1.933	−0.650	0.411	7.419
resid_fullmod	1,460	0.001	1.008	−5.679	−0.575	0.337	9.270

As can be seen, the mean values are around 0, which suggests the exogeneity assumption holds.

Next, a check can be performed for homoskedasticity. Here, a plot of residuals against the fitted values of the model can be generated with R to see if there are any systematic patterns (which would suggest heteroskedasticity). The graph shows no obvious pattern. The Breusch-Pagan test can further be performed, which proves to be significant ($p < 0.01$) and suggests there is indeed heteroskedasticity in the model. To remedy this problem, White's robust standard errors are implemented.

Finally, there is the normal distribution assumption. This we can test with the Shapiro-Wilk test, which yields a significant result ($p < 0.01$), suggesting the residuals are not normally distributed. Although multiple (log, square root, inverse) transformations on the dependent variable and independent variables were performed, none yielded a normal distribution as the Shapiro-Wilk tests kept being significant. Judging based on the histograms and QQ plots, however, it was decided to work with the log transformation of SalePrice as the dependent variable in the final model, which is presented in Table 6 below including the robust standard errors (the right column is the original model).

Table 6: Final Regression Model

	Dependent variable:	
	lnprice	SalePrice
	(Final Model)	(Original Model)
LotArea	0.00001***	1.673***
	(0.00000)	(0.163)
BldgType2fmCon	−0.003	1,245.230
	(0.062)	(11,544.740)
BldgTypeDuplex	−0.154**	−19,645.810
	(0.072)	(24,433.830)
BldgTypeTwnhs	−0.279***	−17,909.110
	(0.051)	(21,779.870)
BldgTypeTwnhsE	−0.323***	−62,363.560***
	(0.045)	(11,845.680)
YearBuilt	−0.285***	−66,103.840***
	(0.028)	(5,850.106)
I(YearBuilt^2)	0.0001***	17.090***
	(0.00001)	(1.492)
GarageArea	0.001***	134.159***
	(0.0001)	(7.983)
LotArea:BldgType2fmCon	−0.00001***	−1.344***
	(0.00000)	(0.380)
LotArea:BldgTypeDuplex	0.00000	−0.794
	(0.00001)	(2.386)
LotArea:BldgTypeTwnhs	0.00003**	−5.498
	(0.00001)	(8.350)
LotArea:BldgTypeTwnhsE	0.00005***	8.330***
	(0.00001)	(2.181)
Constant	285.656***	63,987,512.000***
	(27.826)	(5,732,399.000)
Observations	1,460	1,460
R^2	0.607	0.550
Adjusted R^2	0.604	0.546
Residual Std. Error (df = 1447)	0.251	53,501.830
F Statistic (df = 12; 1447)	186.456***	147.483***
Note:		*p<0.1; **p<0.05; ***p<0.01

There are more significant variables now - specifically BldgTypeDuplex ($p = 0.032$), BldgTypeTwnhs ($p < 0.01$) and the interaction term betwen LotArea and the Townhouse Inside Units ($p = 0.017$). All previously significant variables are also significant in the new model. The model's explanatory power has increased to 60.7% ($F = 186.456, p < 0.01$). As the dependent variable is now log transformed, the interpretation of coefficients will differ - a one unit increase is corresponding to the coefficient*100 percent increase in SalePrice. For example, when holding all variables constant, having a Duplex will decrease the Sale Price by 15.4%.

Subset Analyses

In the final section, I perform a number of subset analyses. For the sake of clarity and ease of interpretation, I use SalePrice instead of lnprice as the dependent variable in these subset analyses and the linear form of the variable YearBuilt. The categorical variables for Building Type are left out for some analyses because the small number of observations for certain categories were split even further through the subsetting (i.e. after certain data set transformations, some building types would have less than 10 observations, which is

not enough to generate reliable linear regression coefficients). Past research (Zietz, Zietz, & Sirmans, 2008) has shown that the significance of variables can vary wildly across different price ranges. I divide the entire data set on its median price and perform the analysis on the two resulting data sets - see Table 7.

Table 7: Subset Analysis - On Sale Price

	Dependent variable:	
	SalePrice	
	(Low Priced Houses)	(High Priced Houses)
LotArea	0.794***	0.972***
	(0.187)	(0.182)
YearBuilt	288.754***	230.544**
	(32.127)	(105.482)
GarageArea	36.185***	221.628***
	(4.475)	(14.010)
Constant	−458,849.400***	−362,659.800*
	(62,526.300)	(207,734.000)
Observations	732	728
R^2	0.246	0.324
Adjusted R^2	0.243	0.321
Residual Std. Error	21,762.210 (df = 728)	62,924.620 (df = 724)
F Statistic	79.354*** (df = 3; 728)	115.594*** (df = 3; 724)
Note:		*p<0.1; **p<0.05; ***p<0.01

As can be seen, all the variables remain significant, but the coefficients are mostly higher for the higher priced houses, which somewhat contradicts Zietz et al.'s (2008) research. The exception to this would be YearBuilt, as it shows that for higher priced houses, the effect of an additional year of construction on house price is lower (so for higher priced houses, "newness" has less of a price premium). Otherwise, the coefficients are quite comparable (suggesting that this particular subset analysis may be redundant), with only the variable GarageArea having a much greater positive effect on house prices for high priced houses ($\hat{\beta} = 221.628, p < 0.01$).

Another worthwhile question to explore would be whether the larger houses (determined by the LotArea) skew the results (as we could see from the scatter plot on SalePrice, the larger the LotArea, the greater the spread of house prices). This time, I use the 3rd quartile to split the data set to still sufficiently separate exceptionally large houses, while keeping the number of observations relevant. See Table 8 for results.

Here, we can immediately see that there are large differences in coefficients between the two subsets. For small houses, the coefficient for LotArea is multiple times higher than that for Large Houses ($\hat{\beta} = 6.242, p < 0.01$). On the other hand, large houses place a much larger premium on every additional square foot of GarageArea ($\hat{\beta} = 217.696, p < 0.01$). Comparing the new models to our original base model, it is safe to say that large houses do indeed skew the regression results.

Table 8: Subset Analysis - On House Size

	Dependent variable:		
	SalePrice		
	(Small Houses)	(Large Houses)	(Base Model)
LotArea	6.242***	0.723***	1.293***
	(0.743)	(0.246)	(0.155)
BldgType2fmCon	13,452.520	−88,559.610***	−6,573.452
	(8,378.094)	(31,790.980)	(10,497.960)
BldgTypeDuplex	−31,200.170***	−58,165.930**	−39,348.950***
	(6,101.232)	(29,094.630)	(7,990.521)
BldgTypeTwnhs	3,973.902		−33,069.040***
	(7,916.255)		(8,997.397)
BldgTypeTwnhsE	14,917.340***	10,811.020	−17,399.650***
	(5,388.255)	(57,620.110)	(5,830.950)
YearBuilt	815.453***	949.352***	876.510***
	(49.748)	(167.844)	(59.652)
GarageArea	101.707***	217.696***	157.200***
	(7.153)	(21.801)	(8.151)
Constant	−1,538,877.000***	−1,777,714.000***	−1,630,991.000***
	(95,451.450)	(327,826.800)	(115,613.900)
Observations	1,095	365	1,460
R^2	0.565	0.390	0.499
Adjusted R^2	0.562	0.380	0.497
Residual Std. Error	39,473.420 (df = 1087)	81,048.430 (df = 358)	56,349.840 (df = 1452)
F Statistic	201.440*** (df = 7; 1087)	38.118*** (df = 6; 358)	206.836*** (df = 7; 1452)
Note:			*p<0.1; **p<0.05; ***p<0.01

Finally, we can also study how the effect of the variables varies across the different building types. From previous sections, we know the effect of lot size on house price differs across different building types but what about the other variables in our model? As such, I split the data set based on the different building types and present the results in Table 9.

For one, we confirm our previous findings, where the coefficients for LotArea greatly differ across building types, with Townhouses benefiting from an increase in the variable the most in terms of house price. The other variables also differ greatly across the different building types, although the Townhouse End Units benefit the most in sale price from both increases in the year of construction ($\hat{\beta} = 1285.799, p < 0.01$) and garage area ($\hat{\beta} = 211.144, p < 0.01$).

Table 9: Subset Analysis - On Type of Dwelling

	Dependent variable:				
	SalePrice				
	(1Fam)	(2fmCon)	(Duplex)	(Twnhs)	(TwnhsE)
LotArea	1.458***	0.715***	1.375	9.091**	9.330***
	(0.177)	(0.206)	(1.141)	(3.765)	(1.720)
YearBuilt	830.920***	−143.013	573.270**	990.493***	1,285.799***
	(64.072)	(232.401)	(236.501)	(285.241)	(350.859)
GarageArea	173.607***	20.469	22.745*	102.048***	211.144***
	(9.222)	(18.562)	(13.522)	(25.575)	(38.369)
Constant	−1,550,955.000***	385,891.900	−1,017,252.000**	−1,889,523.000***	−2,529,268.000***
	(124,001.800)	(446,655.700)	(463,718.800)	(558,322.000)	(692,025.100)
Observations	1,220	31	52	43	114
R^2	0.507	0.373	0.208	0.848	0.559
Adjusted R^2	0.506	0.304	0.159	0.836	0.547
Residual Std. Error	58,095.670 (df = 1216)	29,588.800 (df = 27)	25,526.920 (df = 48)	16,608.680 (df = 39)	40,813.360 (df = 110)
F Statistic	417.032*** (df = 3; 1216)	5.361*** (df = 3; 27)	4.211** (df = 3; 48)	72.370*** (df = 3; 39)	46.447*** (df = 3; 110)
Note:					*p<0.1; **p<0.05; ***p<0.01

In conclusion, I would recommend the stakeholders to use the final model presented in Table 6 with the transformed dependent variable and subset any analyses on lot sizes and types of buildings, as the variable relationships differ greatly in these sub-groups.

References

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of real estate literature*, *13*(1), 1–44.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics*, *37*(4), 317–333.

Appendix - R code

```
####Session set-up####

setwd("C:/Users/ThanhDat/Desktop/School/BAM/Block 1/ASP/Assignments/Assignment 1/Programs")
library(tidyverse)
library(stargazer)
library(lm.beta)
library(car)
library(lmtest)

data<-read.csv(file = "train.csv", header = TRUE)




####Collect and prepare data####
str(data)
#dependent variable: SalePrice
#selection - continuous: LotArea, YearBuilt, GarageArea
#selection - categorical: BldgType

#Convert categoricals to factor variables
data$Neighborhood <- as.factor(data$Neighborhood)
data$BldgType <- as.factor(data$BldgType)

##Summary tables##
categorical <- table(data$BldgType)
stargazer(data[c("LotArea", "YearBuilt", "GarageArea", "BldgType")], type = "latex")


##Graphs##
#Scatterplot for Garage Area with fitted line
scatter <- ggplot(data, aes(GarageArea, SalePrice))
scatter + geom_point() + geom_smooth(method = "lm", colour = "red") +
  labs (x = "Garage Size (sqft)", y = "House Price")
ggsave ("GarageSize.png", width=8, height=5)

#Scattterplot for Lot Area - zoomed in on y up to 500000 and x up to 40000
scatter_lot <- ggplot(data, aes(LotArea, SalePrice, colour = BldgType))
scatter_lot + geom_point() + geom_smooth(method = "lm", aes(fill = BldgType), alpha = 0.1) +
  labs (x = "Lot Size", y = "House Price", colour = "BldgType") +
  coord_cartesian(xlim = c(0,40000), ylim = c(0,500000))
ggsave ("LotArea.png", width=8, height=5)

#Bar charts for Building Types
bar <- ggplot(data, aes(BldgType, SalePrice))
bar + stat_summary(fun = mean, geom = "bar")+
  stat_summary(fun.data = mean_cl_normal, geom = "pointrange") +
  labs(x = "Type of Dwelling", y = "Mean Price") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
ggsave ("BldgType-diffs.png", width=8, height=5)
```

```
#Scatterplot with line for YearBuilt
scatter_year <- ggplot(data, aes(YearBuilt, SalePrice))
scatter_year + geom_point() + geom_smooth() + labs (x = "YearBuilt", y = "House Price")
ggsave ("YearBuilt.png", width=8, height=5)
```

```
####OLS regression and model fit####
model_without <- lm(SalePrice~LotArea + BldgType + YearBuilt + GarageArea, data = data)
model_withnon <- lm(SalePrice~LotArea + BldgType + YearBuilt + I(YearBuilt^2) +
                      GarageArea, data = data)
model_withint <- lm(SalePrice~LotArea*BldgType + YearBuilt + GarageArea, data = data)
model_withboth <- lm(SalePrice~LotArea*BldgType + YearBuilt + I(YearBuilt^2) +
                       GarageArea, data = data)

#stargazer with p values visible for each variable
stargazer(model_without, model_withboth, type = "text", report=('vc*p'))
summary(model_without)

##Determining size effects - standardized coefficient model##
betafullmod <- lm.beta(model_withboth)
betamodnon <- lm.beta(model_withnon)
summary(betafullmod)
stargazer(model_withnon, betamodnon,
          coef = list(model_withnon$coefficients, betamodnon$standardized.coefficients),
          type = "text")
stargazer(model_withboth, betafullmod,
          coef = list(model_withboth$coefficients, betafullmod$standardized.coefficients),
          type = "text")
betamodwithout <- lm.beta(model_without)

#comparing standardized coefficients of the basic model
summary(betamodwithout)
stargazer(model_without, betamodwithout,
          coef = list(model_without$coefficients, betamodwithout$standardized.coefficients),
          type = "text")

#stargazer table used in the assignment
stargazer(model_without,model_withnon, model_withint, model_withboth, betafullmod,
          coef = list(model_without$coefficients,
                      model_withnon$coefficients,
                      model_withint$coefficients,
                      model_withboth$coefficients,
                      betafullmod$standardized.coefficients))
```

```
####Diagnostic checks####
##A1- residuals against X - GarageArea - linearity##
```

```
resid_fullmod <- rstudent(model_withboth)
scatter_resid <- ggplot(data, aes(GarageArea, resid_fullmod))
scatter_resid + geom_point(colour="blue4") +
  ylab("Studentized deleted residuals, sdresid") +
  xlab("Explanatory variable, GarageArea") +
  geom_hline(yintercept = 0, colour="grey") +
  geom_hline(yintercept = qt(0.995,df=model_withboth$df.residual),
             colour="grey", linetype=2) +
  geom_hline(yintercept = qt(0.005,df=model_withboth$df.residual),
             colour="grey", linetype=2) +
  theme(axis.text.x = element_text(size=rel(1.25)),
        axis.text.y = element_text(size=rel(1.25)))
ggsave("residualsagainstgarage.png")

##residuals against X - LotArea##
scatter_resid <- ggplot(data, aes(LotArea, resid_fullmod))
scatter_resid + geom_point(colour="blue4") +
  ylab("Studentized deleted residuals, sdresid") +
  xlab("Explanatory variable, LotArea") +
  geom_hline(yintercept = 0, colour="grey") +
  geom_hline(yintercept = qt(0.995,df=model_withboth$df.residual),
             colour="grey", linetype=2) +
  geom_hline(yintercept = qt(0.005,df=model_withboth$df.residual),
             colour="grey", linetype=2) +
  theme(axis.text.x = element_text(size=rel(1.25)),
        axis.text.y = element_text(size=rel(1.25))) +
  coord_cartesian(xlim = c(0,50000))
ggsave("residualsagainstlot.png")

##A2 - checking full-rank##
library(car)
vif <- vif(model_withboth)
stargazer(vif)

##A3 - checking exogeneity##
mod_lot <- lm(SalePrice ~ LotArea, data = data)
resid_lot <- rstudent(mod_lot)

mod_gar <- lm(SalePrice ~ GarageArea, data = data)
resid_gar <- rstudent(mod_gar)

mod_year <- lm(SalePrice ~ YearBuilt, data = data)
resid_year <- rstudent(mod_year)

mod_type <- lm(SalePrice ~ BldgType, data = data)
resid_type <- rstudent(mod_type)

resdf <- data.frame(resid_lot, resid_gar, resid_year, resid_type, resid_fullmod)

stargazer(resdf)

##A4 - Heteroskedasticity##
plot(fitted(model_withboth), resid_fullmod, main = "Residuals vs. Fitted")
library(lmtest)
```

```r
bptest(model_withboth)

#Remedy: robust standard errors
library(sandwich)
seBasic <- sqrt(diag(vcov(model_withboth)))
seWhite <- sqrt(diag(vcovHC(model_withboth, type = "HC0")))
seClust <- sqrt(diag(vcovHC(model_withboth, cluster = "BldgType")))
stargazer(model_withboth, model_withboth, model_withboth, se = list(seBasic, seWhite, seClust))

##A5 - Normality distribution tests##

shapiro.test(resid_fullmod)
#the residuals are not normally distributed

#SalePrice
shapiro.test(data$SalePrice)
data$lnprice <- log(data$SalePrice)
histogram <- ggplot(data, aes(lnprice))
histogram + geom_histogram()
qplot(sample = data$lnprice, stat = "qq")
qqnorm(data$lnprice)
qqline(data$lnprice)

shapiro.test(data$lnprice)
shapiro.test(res_lnsale)
shapiro.test(sqrt(data$SalePrice))
shapiro.test(1/(data$SalePrice))
#none of the transformations help - the model keeps the Sale Price variable as it is

#LotArea
shapiro.test(data$LotArea)
shapiro.test(log(data$LotArea))
histogram <- ggplot(data, aes(LotArea))
histogram + geom_histogram()
qqnorm(log(data$LotArea))
qqline(log(data$LotArea))

shapiro.test(sqrt(data$LotArea))
shapiro.test(1/(data$LotArea))

#GarageArea
shapiro.test(data$GarageArea)
histogram <- ggplot(data, aes(GarageArea))
histogram + geom_histogram()
shapiro.test(log(data$GarageArea + 1))
shapiro.test(sqrt(data$GarageArea))
shapiro.test(1/(data$GarageArea + 1))

#YearBuilt
shapiro.test(data$YearBuilt)
histogram <- ggplot(data, aes(log(YearBuilt)))
histogram + geom_histogram()
shapiro.test(log(data$YearBuilt))
shapiro.test(sqrt(data$YearBuilt))
```

```r
shapiro.test(1/(data$YearBuilt))

#none of the transformations of the independent variables yielded a normal distribution
#looking at the histogram and the qq plot however, they suggest that a
#log transformation of the dependent variable sale price would be suitable
#as it gets it very close to a normal distribution

##New model based on diagnostic checks##
lnsale_mod <- lm(lnprice ~ LotArea*BldgType + YearBuilt +
                   I(YearBuilt^2) + GarageArea, data = data)
res_lnsale <- rstudent(lnsale_mod)

seBasic <- sqrt(diag(vcov(lnsale_mod)))
seWhite <- sqrt(diag(vcovHC(lnsale_mod, type = "HC0")))
seClust <- sqrt(diag(vcovHC(lnsale_mod, cluster = "BldgType")))

stargazer(lnsale_mod, model_withboth, se = list(seWhite))
stargazer(lnsale_mod, model_withboth, se = list(seWhite), type = "text",report=('vc*p'))
```

```r
####Subset Analyses####
##Subsetting on house prices##
#Checking the data
stargazer(data, type = "text")
hist(data$SalePrice)
median(data$SalePrice)

#Subsetting based on median price
highpset <- subset(data, SalePrice > 163000)
lowpset <- subset(data, SalePrice <= 163000)

#Model creation
highp_model <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = highpset)
lowp_model <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = lowpset)
stargazer(lowp_model,highp_model)
summary(highp_model)

##Subsetting on lot area##
stargazer(data["LotArea"], type = "text")

#use 3rd quartile/75 percentile = 11601.5 to subset
bigset <- subset(data, LotArea > 11601.5)
smallset <- subset(data, LotArea <= 11601.5)

#Model Creation
big_model <- lm(SalePrice ~ LotArea + BldgType + YearBuilt + GarageArea, data = bigset)
small_model <- lm(SalePrice ~ LotArea + BldgType + YearBuilt + GarageArea, data = smallset)
stargazer(small_model,big_model, model_without)

##Subsetting on BldgType##
```

```r
table(data$BldgType)
ds1fam <- subset(data, BldgType == "1Fam")
ds2fmcon <- subset(data, BldgType == "2fmCon")
dsDuplex <- subset(data, BldgType == "Duplex")
dsTwnhs <- subset(data, BldgType == "Twnhs")
dsTwnhsE <- subset(data, BldgType == "TwnhsE")

#Model Creation
model_1fam <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = ds1fam)
model_2fmcon <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = ds2fmcon)
model_duplex <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = dsDuplex)
model_twnhs <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = dsTwnhs)
model_twnhse <- lm(SalePrice ~ LotArea + YearBuilt + GarageArea, data = dsTwnhsE)
stargazer(model_1fam, model_2fmcon, model_duplex, model_twnhs, model_twnhse)
```