How to predict house prices?

Assignment 2 for Data Analysis 2

Abduvosid Malikov

Introduction

This is the Assignment 2 for `Data Analysis 2` and `Coding` course. The aim of this project is to predict house price based on its area and other variables. House price prediction can help the sellers of a house to determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

Data collection

Data set was obtained from Kaggle. It contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, Iowa (IA) from 2006 to 2010. This data was collected by Ames Assessor's Office, one of the cities of Iowa. This data is representative only for this state. Since this is administrative data that was collected by government office, we assume that there was no (or few) mistakes in entering data. Also, we have almost all necessary variables that matters for the house sale price (such as area, quality, rooms, etc).

Data descriptives

Variables capture the house price and other conditions such as area of a house and garage, number of rooms, availability of fireplace, pool and others. I want to use these variables to predict house price given the certain house condition (area, rooms, fireplaces). The data has 2930 observations. It consists of 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers, Order and Parcel ID). These variables describe the sales price of a house and its condition, such as area, quality, rooms, etc. As an outcome variable, we have chosen **SalePrice** (in US dollars) variable. The main parameter of interest (explanatory variable) is **Above Ground Area** (in square feet) of a house. Full summary statistics and distributions of explanatory variables are given in Appendix.

The histogram below shows the house sale prices. We can see that on average, most of the houses are priced from 100,000 to 200,000 dollars. It's skewed to the right, meaning that only few houses were priced extremely high.

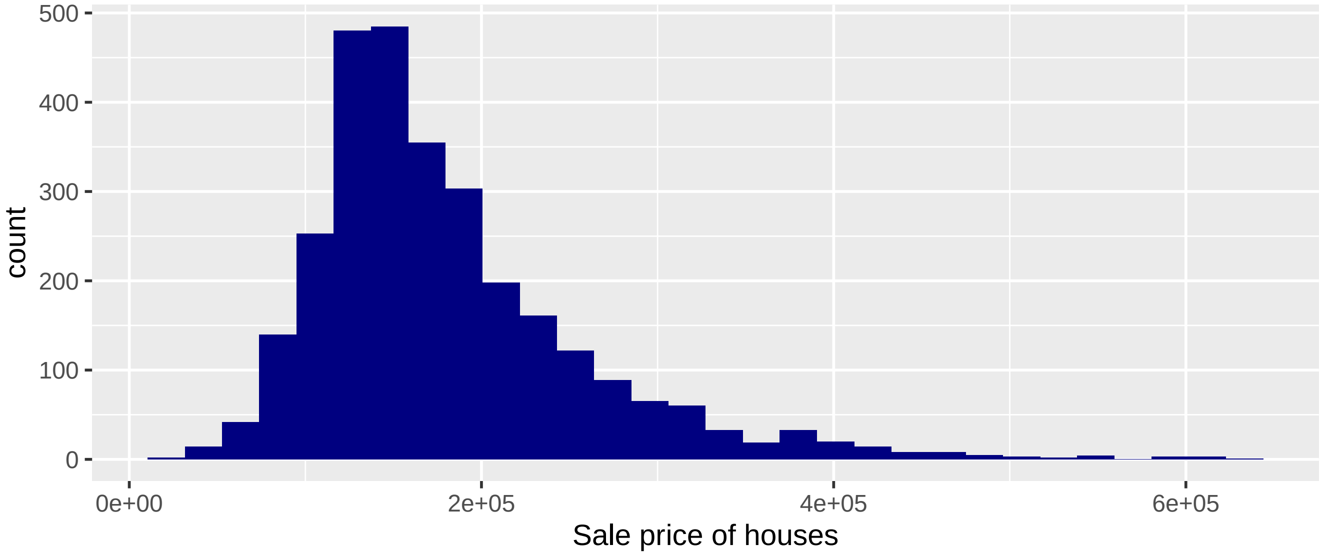


Figure 1: Histogram of house sale prices

Data Cleaning

Outliers

When outcome variable SALE PRICE and parameter of interest GR LIV AREA (above ground living area square feet) were plotted, there were 5 observations with extreme values. When inspected, it was clear that two of these outliers have more than 5000 square feet area (which is extremely high) but nevertheless priced relatively appropriately. Also, three of them had Partial Sales that likely don't represent actual market values. Therefore, I removed all houses with more than 4000 square feet from the data set to avoid these five unusual observations (Appendix, Figure 4.

Therefore, re-iterated research question is the following: **Does a house with higher above ground living area (considering it's less than 4000 square feet) have higher price from 2006 - 2010 in Ames (Iowa) ?**

Data Type for variables

An appropriate format was given to categorical (numeric and ordinal) variables (see Appendix).

Missing values

Missing values in qualitative variables As it was described in original data description, in some categorical variables, NA didn't mean missing value. Therefore, corresponding name was assigned to the observations with such values (such as "No garage" or "No garage").

I decided to drop columns "Alley", "Misc.Feature", "Fence", "Pool.Qu" and "Fireplace.Qu" because more than 90% (48% of Fireplace.Qu) of their values were missing. Observations with NAs in all the garage related columns, were completed with the value "No garage" in case of nominal attribute. I have not to used Street variable (Type of road access to property) as there are only 12 houses with Gravel and 2913 of them are Paved

Pattern of association

With the help of scatter-plots and LOESS method, we want to check the pattern of association between y and each potential x variables.

First, we see the plot of Sale Price and Above Ground Area since it's our main parameter of interest. The plot below shows that there is a strong linear relationship between these two variables.

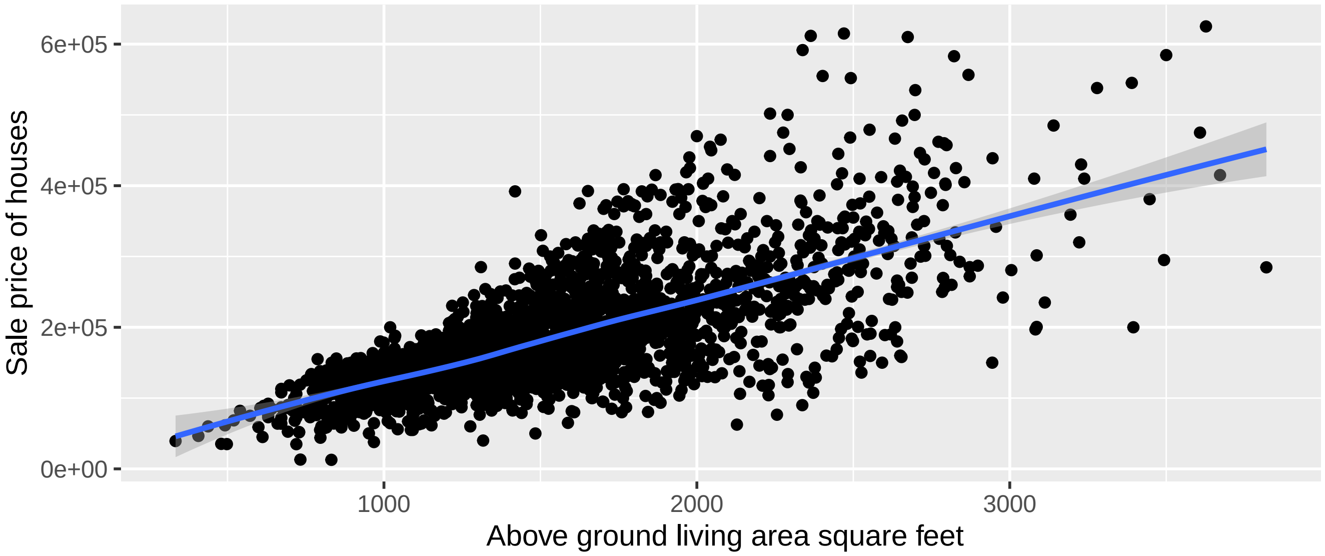


Figure 2: Plot of Above Ground Area of houses

The scatterplots that shows the pattern of association between *y variable* (Sale Price) and other explanatory variables are given in Appendix.

After we saw which kind of relationship the variables have, now we have an idea on how build a regression model and what variables to include into our model.

Table 1: Correlation table

	SalePrice	TotalArea	Lot.Area	HasFireplace	Garage.Area	Total.Bsmt.SF	Gr.Liv.Area	Bsmt.Qual	MS.Zoning	Overall.Qual	BsmtFin.Type.1	Year.Built	Year.Remod.Add	Garage.Yr.Blt	Garage.Cars	TotRms.AbvGrd
SalePrice	1.00	0.83	0.26	0.47	0.64	0.66	0.73	0.62	-0.15	0.80	0.32	0.56	0.54	0.53	0.65	0.52
TotalArea	0.83	1.00	0.28	0.46	0.59	0.81	0.86	0.54	-0.08	0.66	0.22	0.38	0.37	0.37	0.59	0.65
Lot.Area	0.26	0.28	1.00	0.17	0.19	0.21	0.25	0.04	0.00	0.07	0.02	0.00	0.00	-0.02	0.16	0.20
HasFireplace	0.47	0.46	0.17	1.00	0.26	0.30	0.45	0.28	-0.01	0.42	0.15	0.22	0.19	0.17	0.31	0.32
Garage.Area	0.64	0.59	0.19	0.26	1.00	0.49	0.50	0.41	-0.16	0.53	0.20	0.46	0.38	0.55	0.85	0.36
Total.Bsmt.SF	0.66	0.81	0.21	0.30	0.49	1.00	0.39	0.58	-0.07	0.54	0.34	0.41	0.29	0.35	0.45	0.24
Gr.Liv.Area	0.73	0.86	0.25	0.45	0.50	0.39	1.00	0.34	-0.06	0.56	0.05	0.25	0.32	0.27	0.52	0.81
Bsmt.Qual	0.62	0.54	0.04	0.28	0.41	0.58	0.34	1.00	-0.16	0.64	0.40	0.63	0.53	0.59	0.47	0.18
MS.Zoning	-0.15	-0.08	0.00	-0.01	-0.16	-0.07	-0.06	-0.16	1.00	-0.17	-0.08	-0.31	-0.20	-0.26	-0.14	0.02
Overall.Qual	0.80	0.66	0.07	0.42	0.53	0.54	0.56	0.64	-0.17	1.00	0.26	0.60	0.57	0.57	0.58	0.38
BsmtFin.Type.1	0.32	0.22	0.02	0.15	0.20	0.34	0.05	0.40	-0.08	0.26	1.00	0.36	0.24	0.29	0.18	-0.06
Year.Built	0.56	0.38	0.00	0.22	0.46	0.41	0.25	0.63	-0.31	0.60	0.36	1.00	0.63	0.83	0.53	0.13
Year.Remod.Add	0.54	0.37	0.00	0.19	0.38	0.29	0.32	0.53	-0.20	0.57	0.24	0.63	1.00	0.65	0.45	0.20
Garage.Yr.Blt	0.53	0.37	-0.02	0.17	0.55	0.35	0.27	0.59	-0.26	0.57	0.29	0.83	0.65	1.00	0.59	0.16
Garage.Cars	0.65	0.59	0.16	0.31	0.85	0.45	0.52	0.47	-0.14	0.58	0.18	0.53	0.45	0.59	1.00	0.41
TotRms.AbvGrd	0.52	0.65	0.20	0.32	0.36	0.24	0.81	0.18	0.02	0.38	-0.06	0.13	0.20	0.16	0.41	1.00

Table 2: Variables with high correlation

Var1	Var2	corr_val
TotalArea	SalePrice	0.83
Overall.Qual	SalePrice	0.80
SalePrice	TotalArea	0.83
Total.Bsmt.SF	TotalArea	0.81
Gr.Liv.Area	TotalArea	0.86
Garage.Cars	Garage.Area	0.85
TotalArea	Total.Bsmt.SF	0.81
TotalArea	Gr.Liv.Area	0.86
TotRms.AbvGrd	Gr.Liv.Area	0.81
SalePrice	Overall.Qual	0.80
Garage.Yr.Blt	Year.Built	0.83
Year.Built	Garage.Yr.Blt	0.83
Garage.Area	Garage.Cars	0.85
Gr.Liv.Area	TotRms.AbvGrd	0.81

Comparing explanatory variables

We selected x variables that has shown strong association with y variables. Total Area, Above Ground Area, Garage Area are among them. Now we explore how these x-s are related to each other.

Table 1 shows the selected x variables and shows correlation between them.

From the correlation table, we retained the variables that has correlation less than 0.8 and removed others. The reason for this is that highly correlated variables may be confounders that leads us to multicollinearity issue (high Standard Error). Table 2 shows highly correlated variables.

The following variables showed correlation less than 0.8, therefore I decided to retain them for further analysis and include in regression model:

SalePrice: Lot.Area, HasFireplace, Garage.Area, Total.Bsmt.SF, Gr.Liv.Area, Bsmt.Qual, MS.Zoning, BsmtFin.Type.1, Year.Built, Year.Remod.Add, Garage.Yr.Blt, Garage.Cars, TotRms.AbvGrd

I also decided to include HasFireplace variable (as a dummy variable, 1 if house has one or more fireplaces, 0 otherwise). I want to interact HasFireplace variable with Gr.Liv.Area variable (Above ground area). The reason for controlling for such interaction is that building a fireplace in the house requires additional expenditure from the owners. Therefore, owners may tend to increase house price because of the fireplace.

Model choice

I decided to build several regression models in order to see how they perform and pick up the best one later.

First regression model is:

SalePrice = Beta0 + Beta1 * **Above Ground Area**

R^2 = 51

Second model is:

SalePrice = Beta0 + Beta1 * **Basement Area** + Beta2 * **Above Ground Area**

R^2 = 0.68

Third model is the interaction of HasFireplace dummy variable with Above Ground Area:

Regression 3.1

SalePrice = Beta0 + Beta1 * **Above Ground Area** + Beta2 * **HasFireplace** + Beta3 * **Above Ground Area** * **HasFireplace** R2 = 57

Fourth model is:

Regression 3.2

SalePrice = Beta0 + Beta1 * **Total Basement Area** + Beta2 * **Above Ground Area** + Beta3 * **HasFireplace** + Beta3 * **Above Ground Area** * **HasFireplace**

R2 = 0.71

Earlier, we have seen strong linear relationship between these variables. Even though the model fits the data pretty well, it has low R^2.

Residual Analysis

For these 5 countries in table 2(see Appendix), the model overestimated life expectancy, as the actual value is smaller than the predicted value; in another word, these countries have short life span than average. The explanation could be extreme weather(temperature too hot or too cold) or worsened social safety conditions.

For these 5 countries in table 3(see Appendix), the model underestimated life expectancy, as the actual value is larger than the predicted value; in another word, these countries have longer life span than average. The explanation could be healthy(Mediterranean and Eastern Asian diet) or complete social security administration.

Also, we check the y and y_hat plot(see Appendix) to examine the model fit. We can see that most scatters fall aside the line, indicating a good fit of the model.

Appendix

Variables

Variables describe house area, garage area, availability of fireplace, pool and other conditions. To see a full list of explanation of variables, go to Data Documentation

Data descriptives

Data Cleaning

clean_data.R script contains all the steps for data cleaning. Exact number of missing values in certain variables are also given as a comment

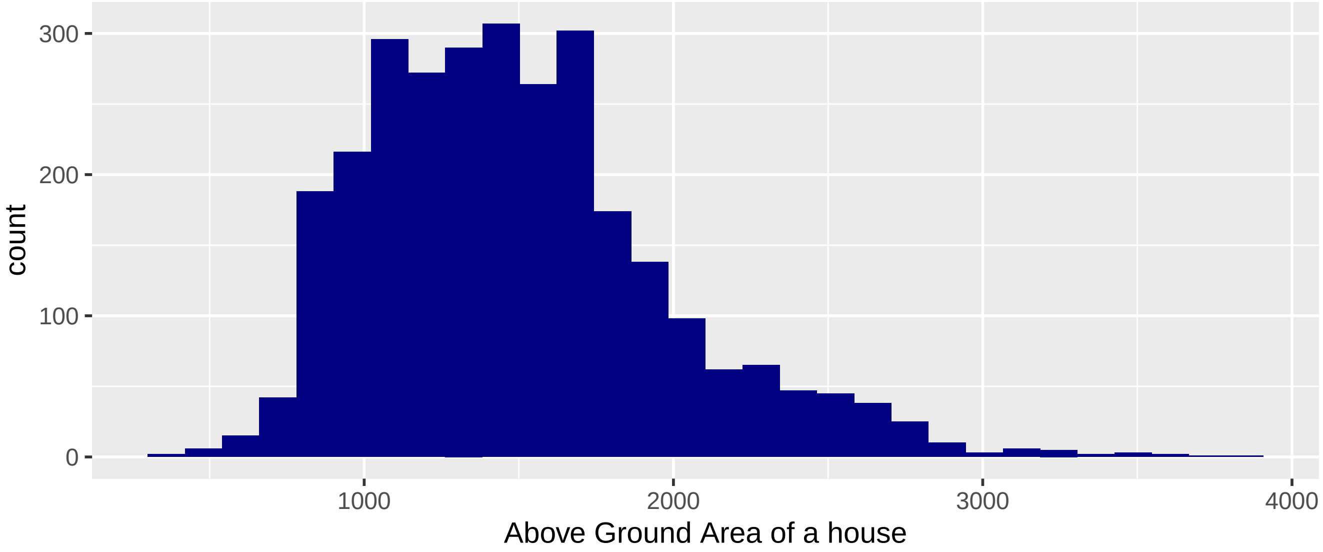


Figure 3: Histogram of Above Ground Area of a house

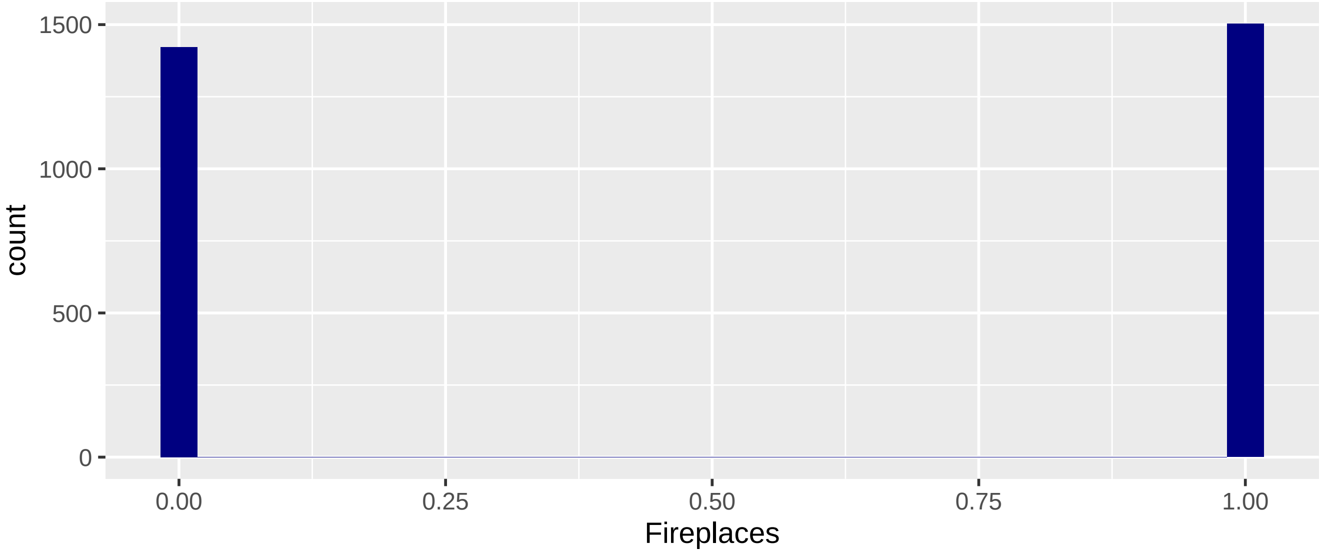


Figure 4: Histogram of houses with and without fireplace

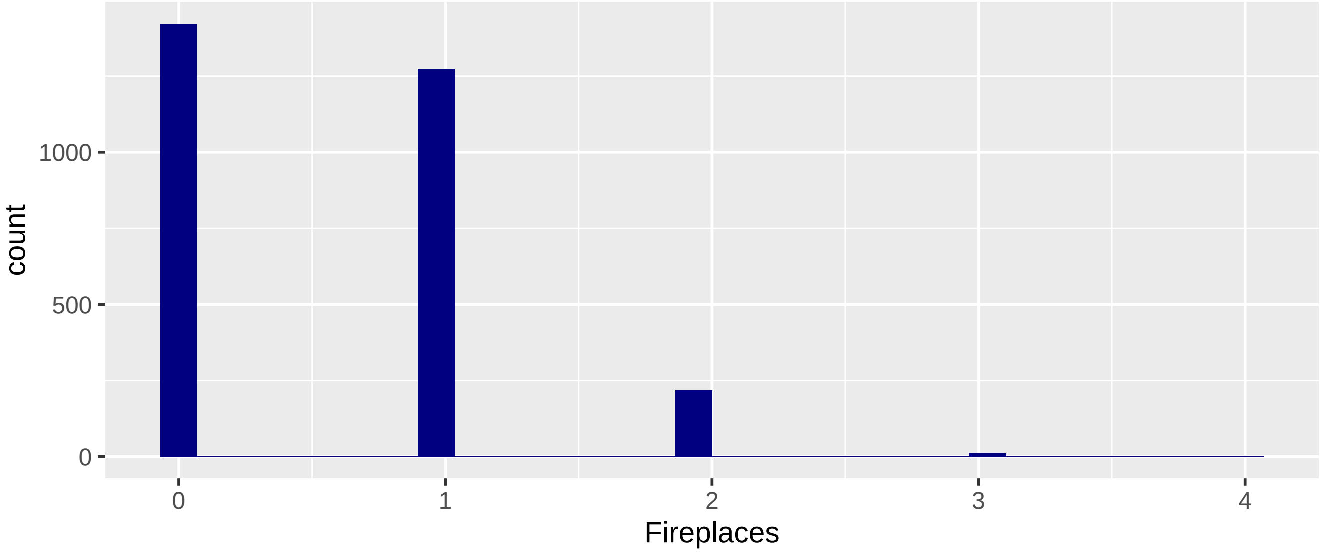


Figure 5: Histogram of houses with 0, 1, 2, 3 and 4 fireplaces

	(1) First Model	(2) Second Model	(3) Third Model	(4) Fourth Model
(Intercept)	6773.48	-36647.54 ***	57113.38 ***	14070.74 **
	(3917.34)	(3942.30)	(4141.03)	(4280.13)
Gr.Liv.Area	116.23 ***	87.63 ***	66.19 ***	43.91 ***
	(3.00)	(2.41)	(3.57)	(3.40)
Total.Bsmt.SF		82.30 ***		78.33 ***
		(2.98)		(2.84)
HasFireplace			-46162.58 ***	-57566.96 ***
			(7731.68)	(6713.80)
Gr.Liv.Area:HasFireplace			54.92 ***	55.15 ***
			(5.58)	(4.88)
nobs	2925	2924	2925	2924
r.squared	0.52	0.68	0.57	0.72
adj.r.squared	0.52	0.68	0.57	0.72
statistic	1497.66	1218.66	860.63	1009.73
p.value	0.00	0.00	0.00	0.00
df.residual	2923.00	2921.00	2921.00	2919.00
nobs.1	2925.00	2924.00	2925.00	2924.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Data Type for variables

An appropriate format was given to categorical (numeric and ordinal) variables. In order to make it easier for further use, categorical variables were converted from character (or numeric) vector into nominal and ordinal variables (using factor and ordered factor) in clean_data.R script

Pattern of association

Plot shows bunch of dots near 0. Probably the Sale Price and Lot Area variables are not correlated.

Ln(Sale Price) - Ln(Lot Area). Probably linear spline, with knots at 8 and 10

The more regular shape house has, the higher the price. But some Irregular houses has same price as Regular houses.

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

Houses with 1 or more fireplaces have higher price

Almost positive linear, but some outliers at the end causing the line to curve. Seems Garage Area variable is important

There is a linear relationship between Sale Price and Total square feet of basement area

Table 3: Summary Statistics of Variables

n	Min	1st IQR	Median	3rd IQR	Max	Mean	Std.	Skew	Name
2925	12789	129500.00	160000	213500.00	625000	180411.57	78554.86	1.59	Sale Price
2925	334	1126.00	1441	1740.00	3820	1493.98	486.27	0.88	Above Ground Living Area
2924	NA	792.75	NA	1299.25	NA	NA	NA	NA	Basement Area
2925	0	0.00	1	1.00	1	0.51	0.50	-0.06	Fireplaces
2925	1300	7438.00	9428	11515.00	215245	10103.58	7782.00	13.19	Lot Area
2924	NA	320.00	NA	576.00	NA	NA	NA	NA	Garage Area

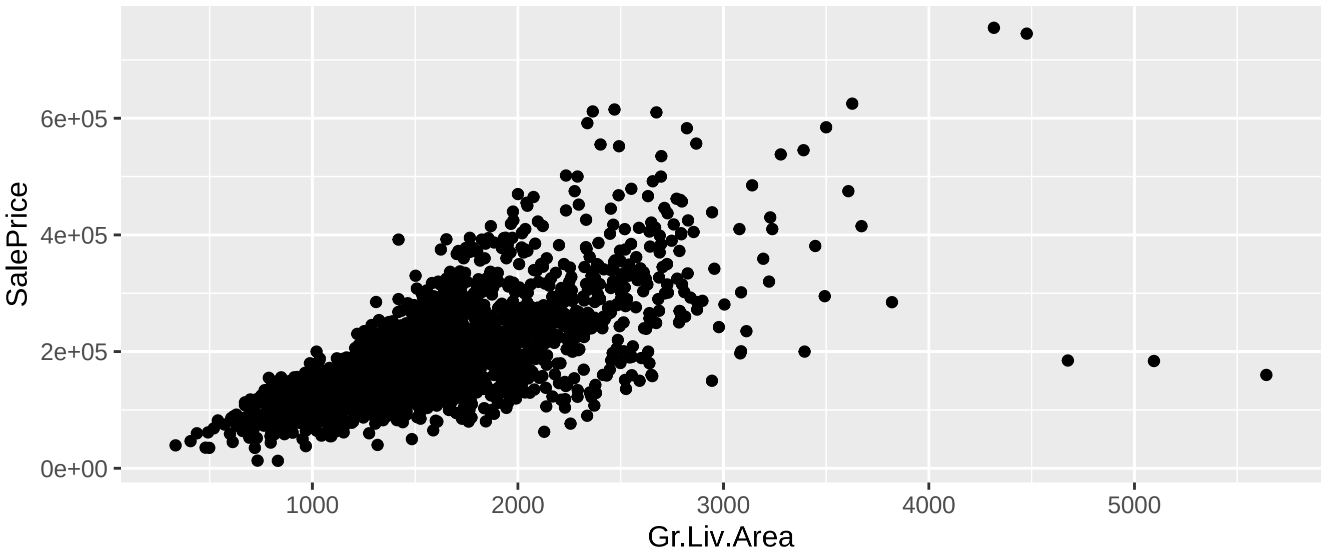


Figure 6: Plot of houses more than 4000 square feet Above Ground Area

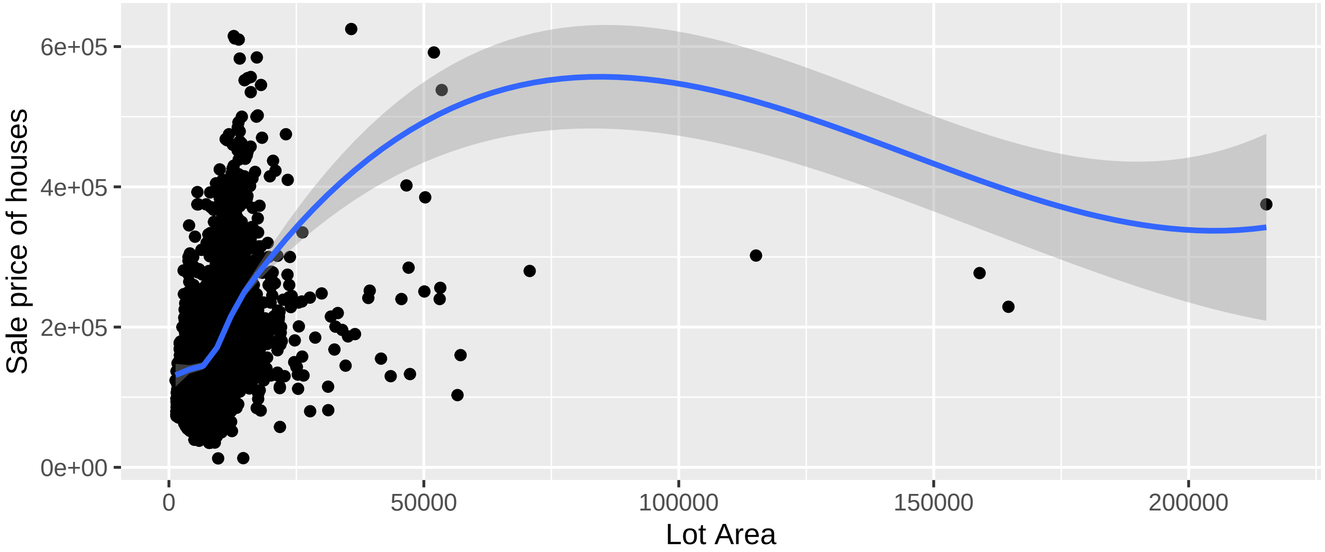


Figure 7: Plot of Sale Price and Lot Area

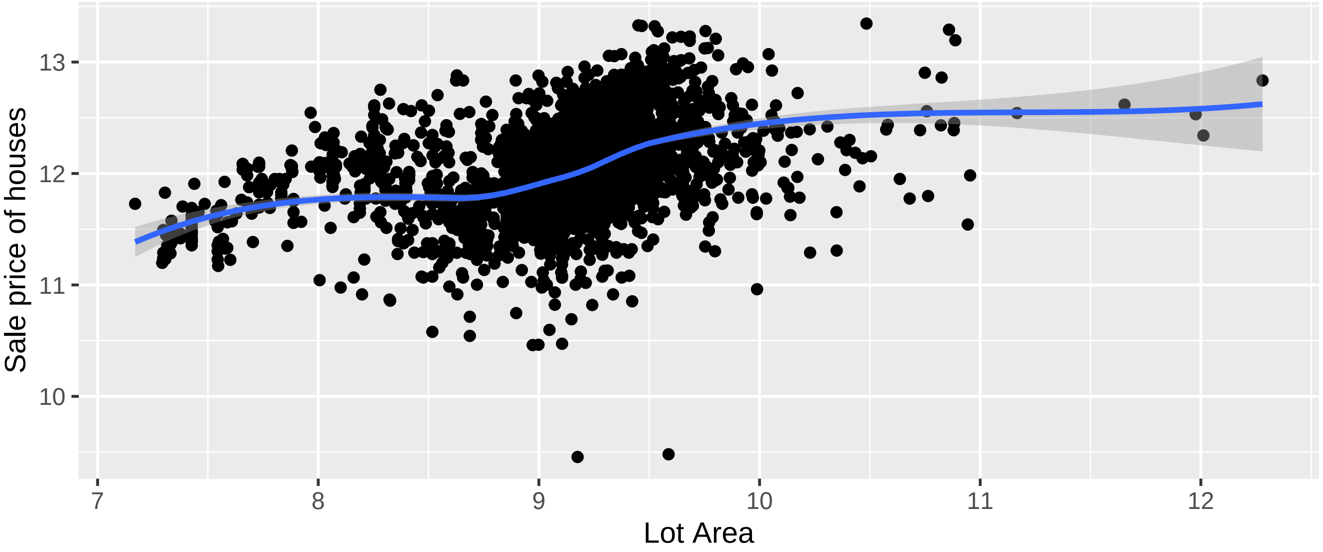


Figure 8: Plot of Ln(Sale Price) - Ln(Lot Area).

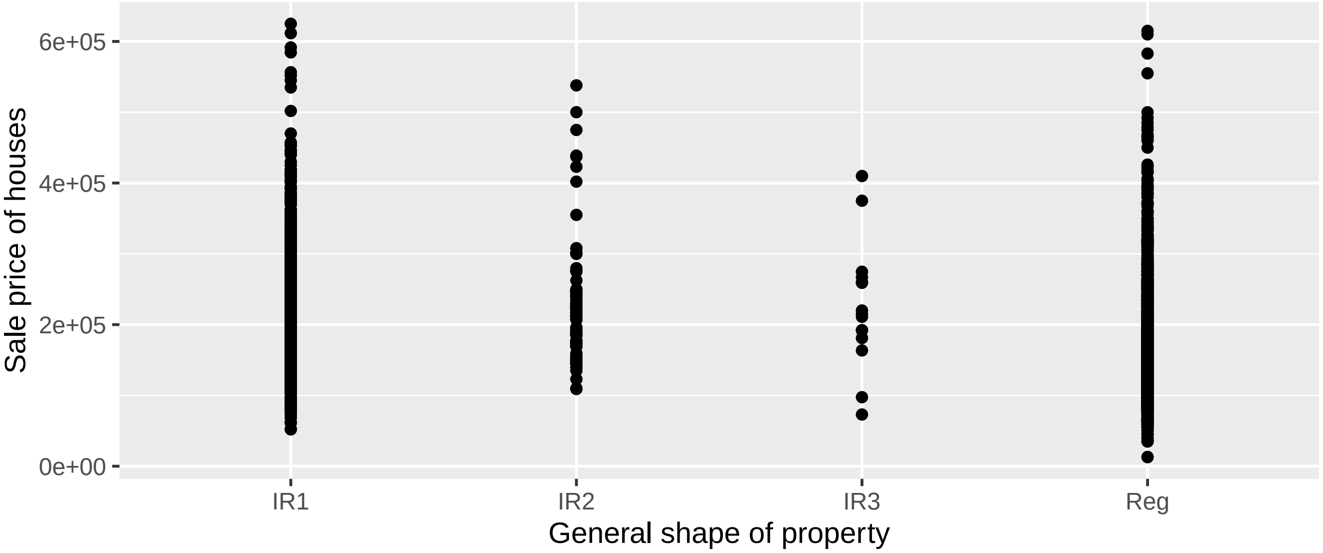


Figure 9: Plot of Sale Price - Lot Shape

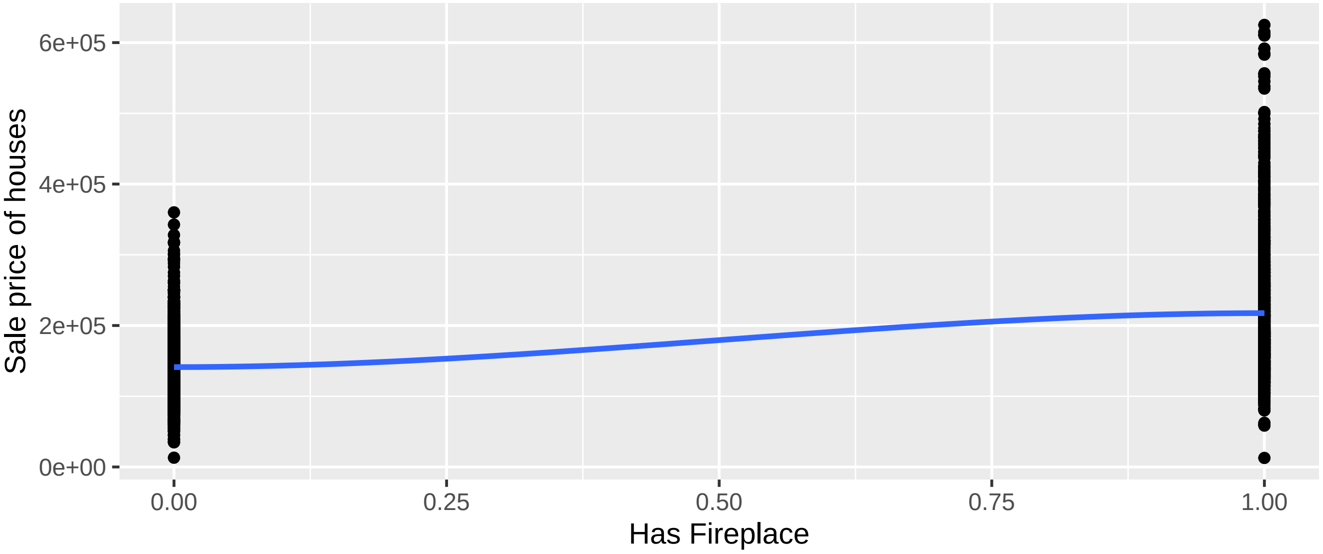


Figure 10: Plot of Sale Price - Fireplace

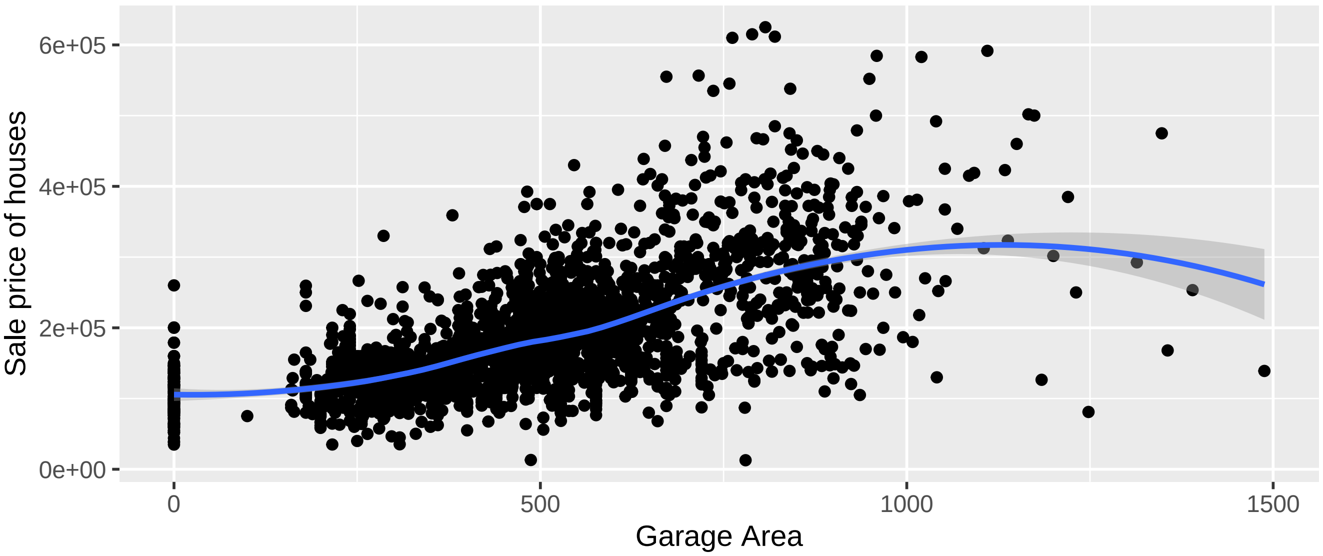


Figure 11: Plot of Sale Price - Garage Area

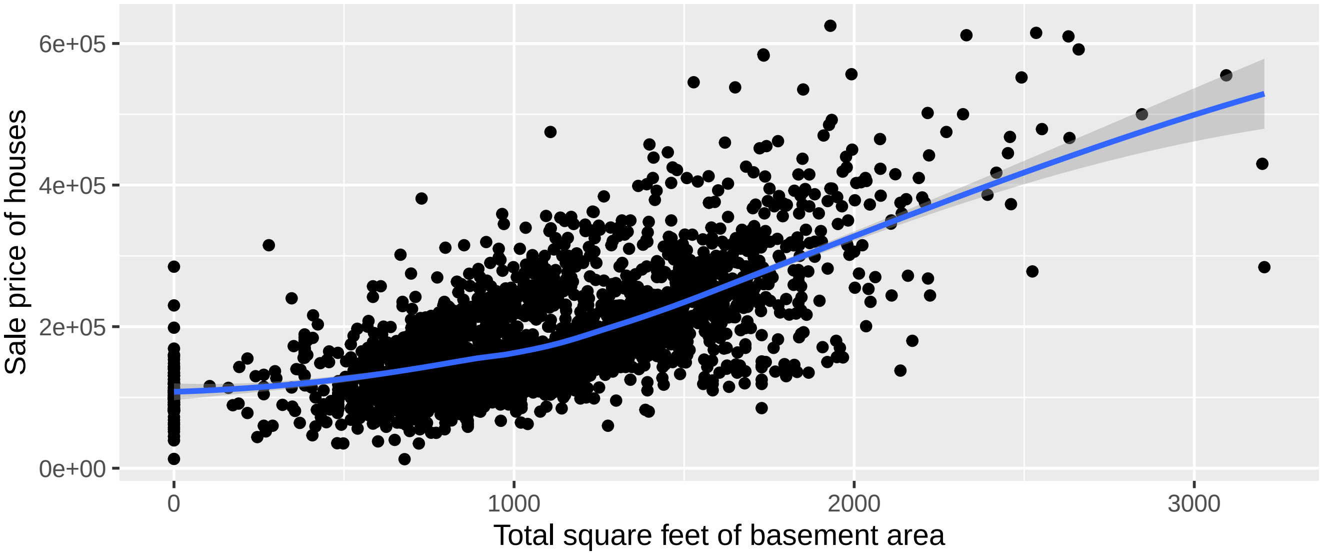
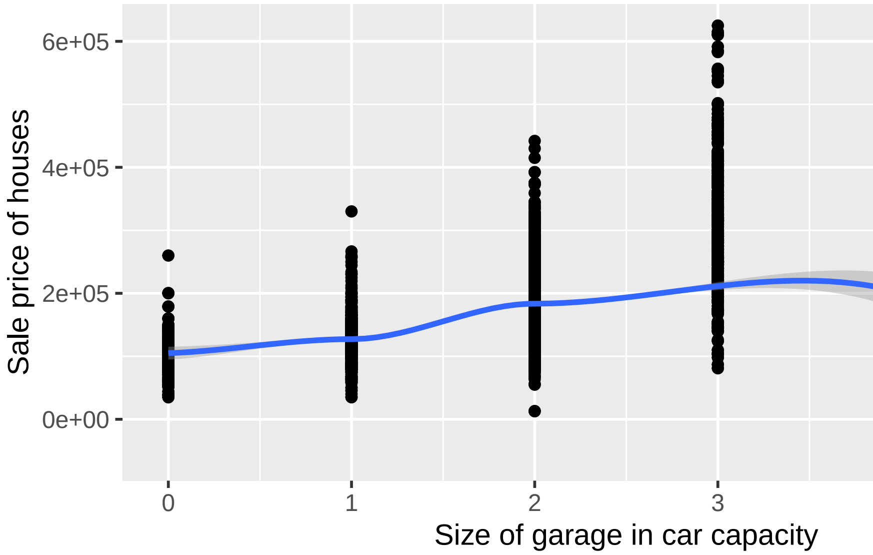


Figure 12: Plot of Sale Price - Basement area



On average, houses with more garage cars have higher price

Near Flat/Level and Hillside houses have higher prices.

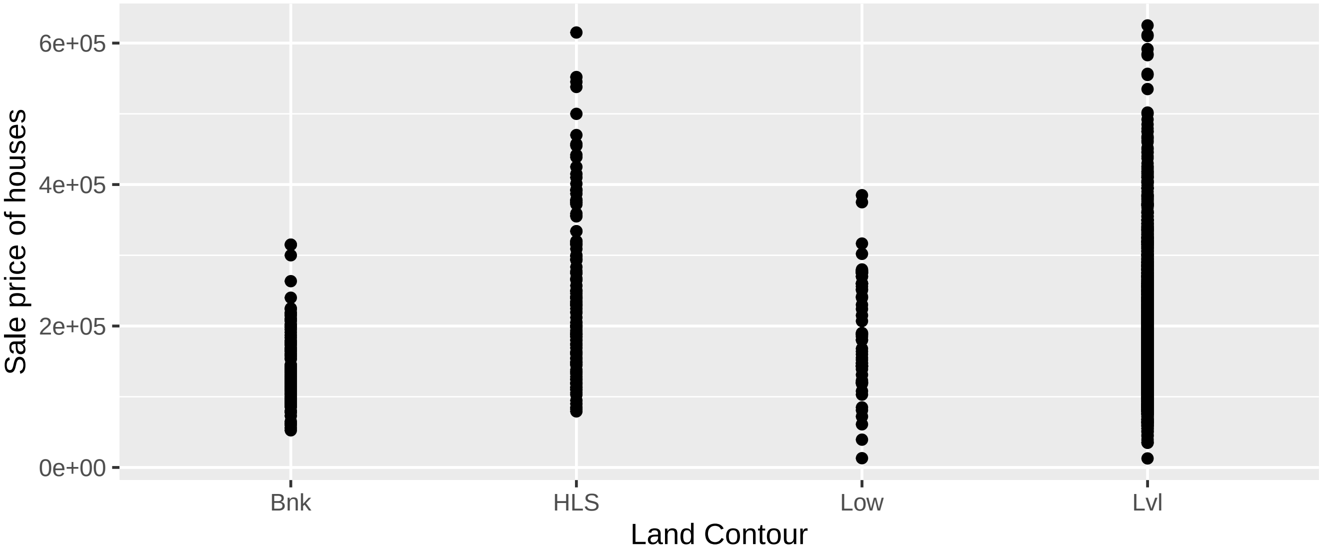


Figure 13: Plot of Sale Price - Land.Contour

There are 2922 houses with All Public *Utilities*. Since we don't have much observations with different utilities, it was decided not to check the pattern of association and not to include this variable in the model.

Cul-de-sac and Inside lot has higher prices

Houses with *Gentle slope* and *Moderate Slope* have higher prices

Houses from Northridge, Northridge Heights, and Stone Brook neighborhoods have relatively high prices

Houses with Normal and "Adjacent to postive off-site feature" conditions have relatively high prices

Most of the houses are Single-family Detached. Single-family Detached and Townhouse End Unit houses have relatively high prices

Most of the houses are One story and Two story. One story and Two story houses have relatively high prices

Higher rating - higher sale price. But some houses has quality 6 but has same price as quality 2

Most of the houses have Average (5) condition. Average conditioned houses have relatively high prices

The newer the house (the later is the year built) - the higher the price. But some houses that are built before 1900 were priced realtively high. There is an upward trend

I also checked the pattern of association between Sale Price and Total Area of a house. TotalArea = Total Basement Area + Above Ground Area

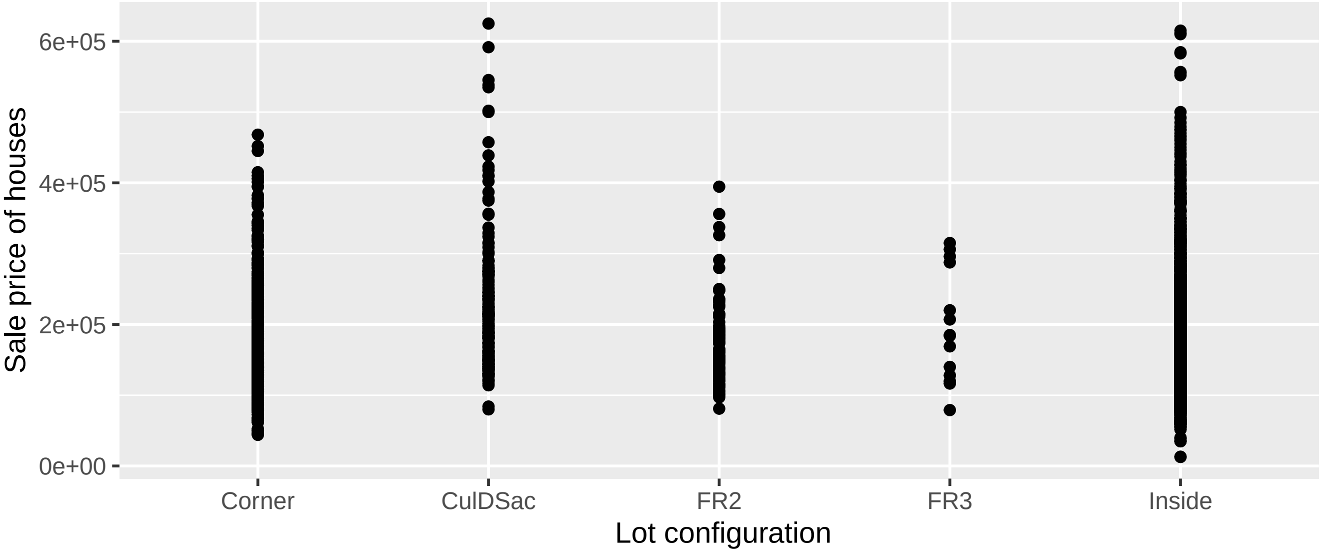


Figure 14: Plot of Sale Price - Land Configuration

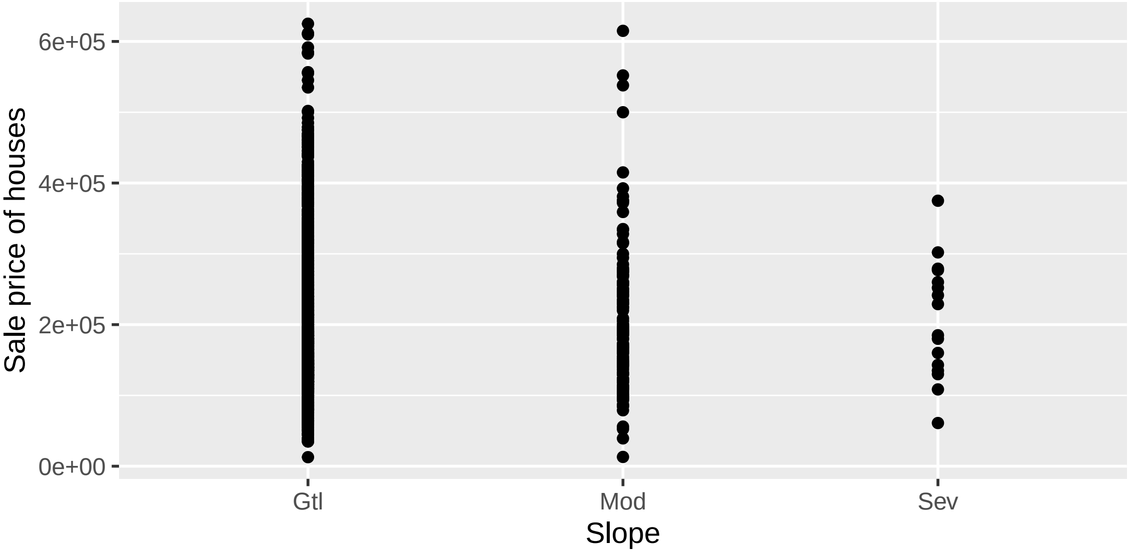


Figure 15: Plot of Sale Price - Land Slope

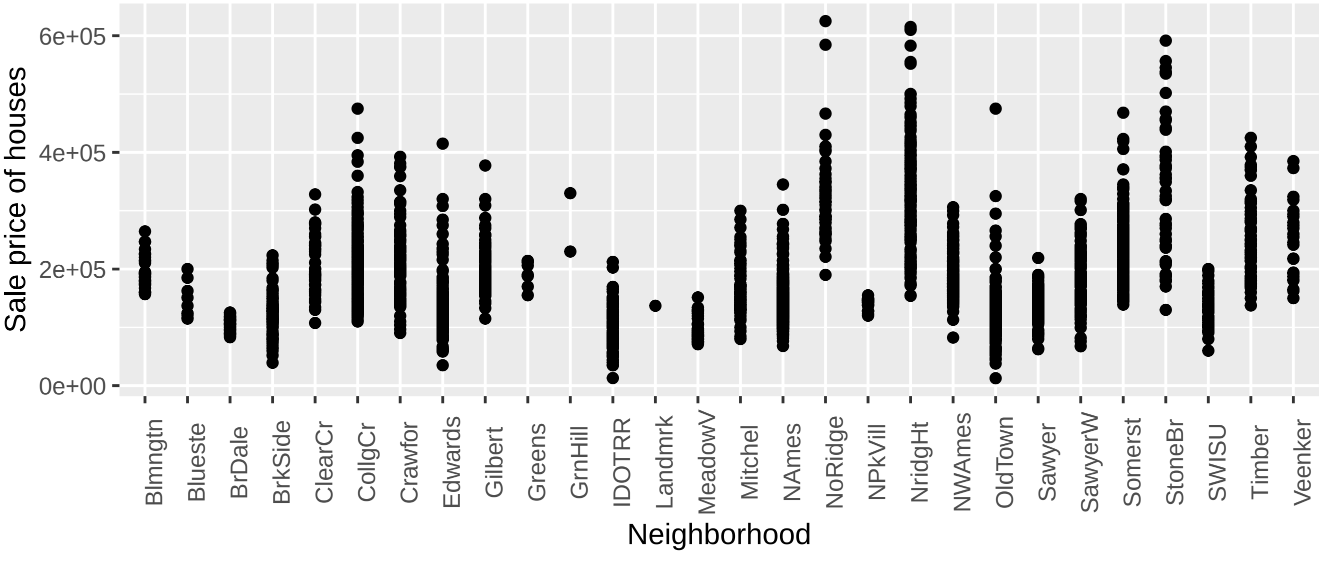


Figure 16: Plot of Sale Price - Neighborhood

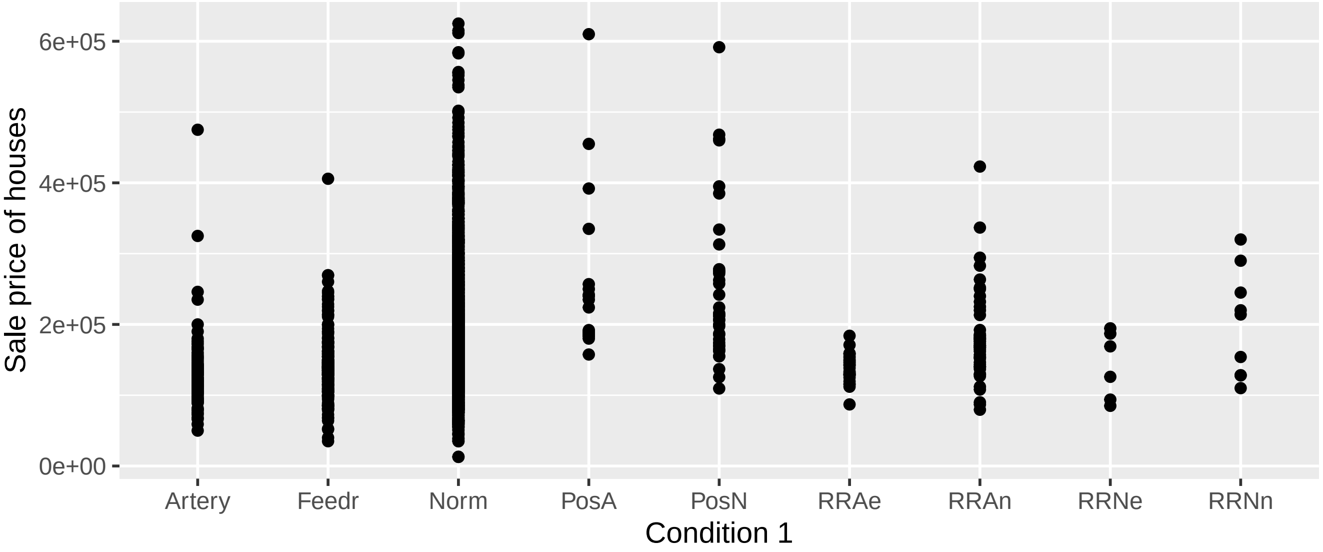


Figure 17: Plot of Sale Price - House Condition

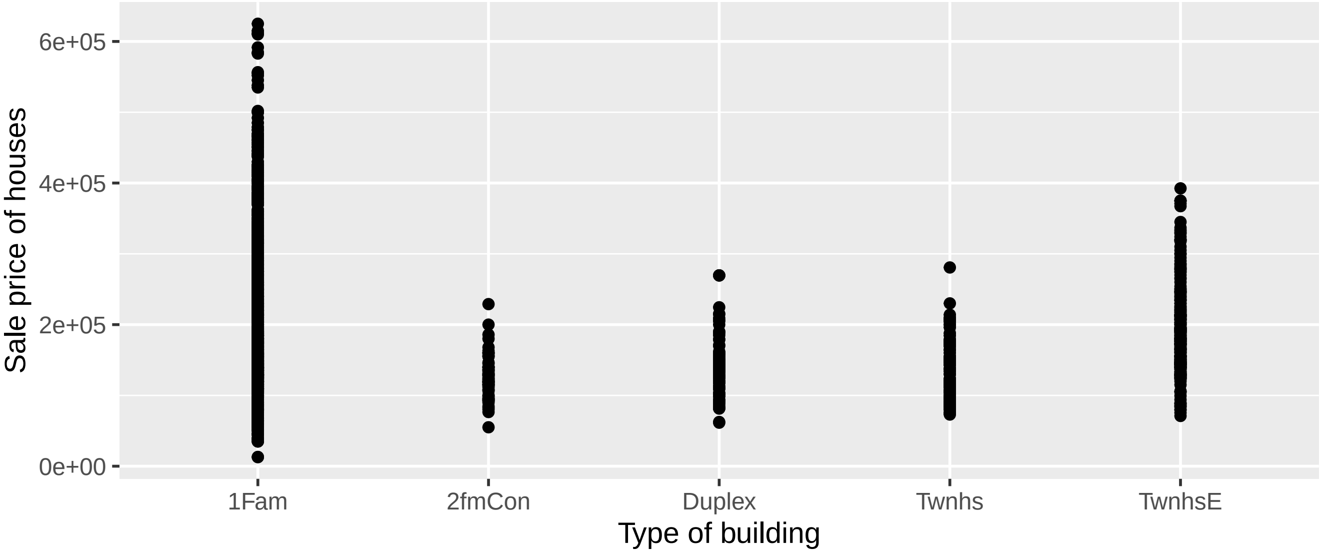


Figure 18: Plot of Sale Price - Type of building

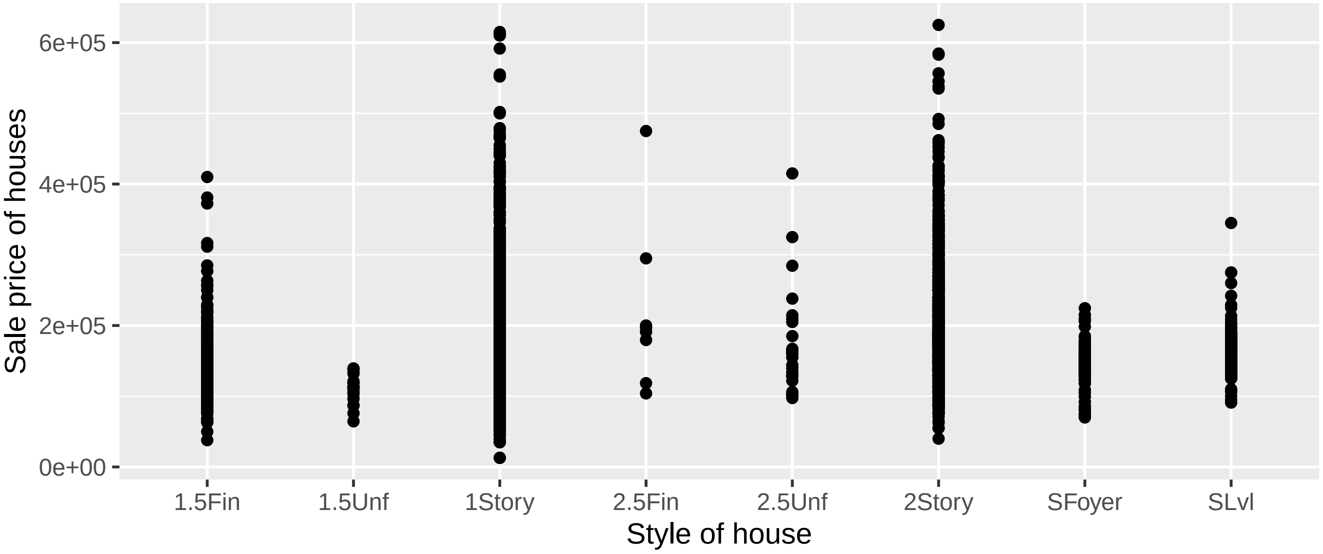


Figure 19: Plot of Sale Price - Style of house

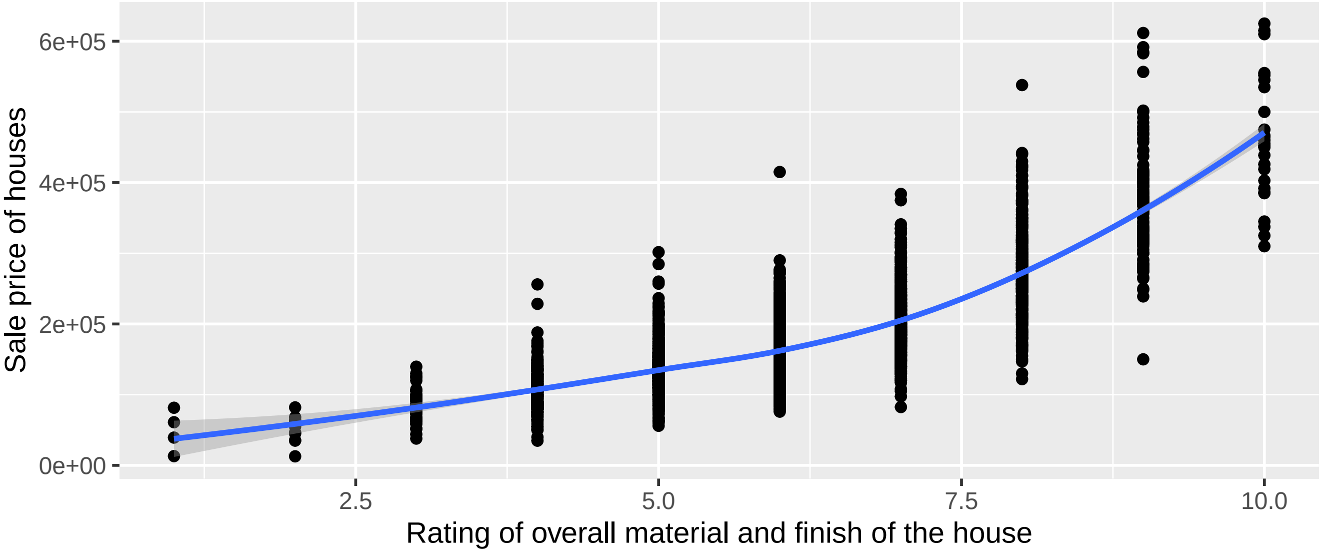


Figure 20: Plot of Sale Price - House rating

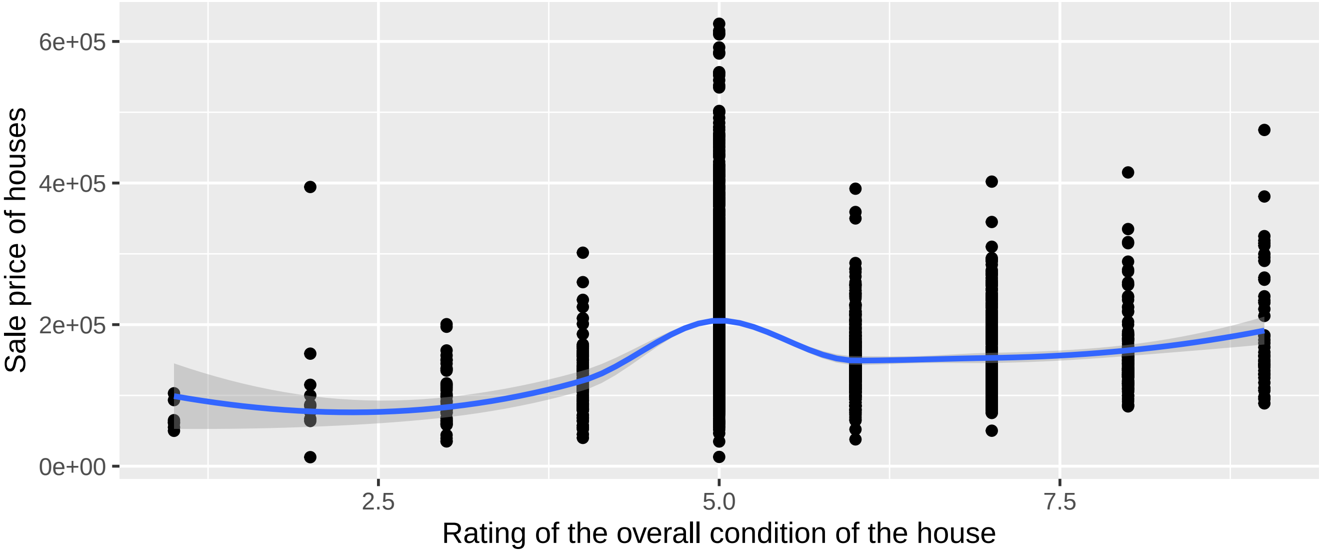


Figure 21: Plot of Sale Price - House condition

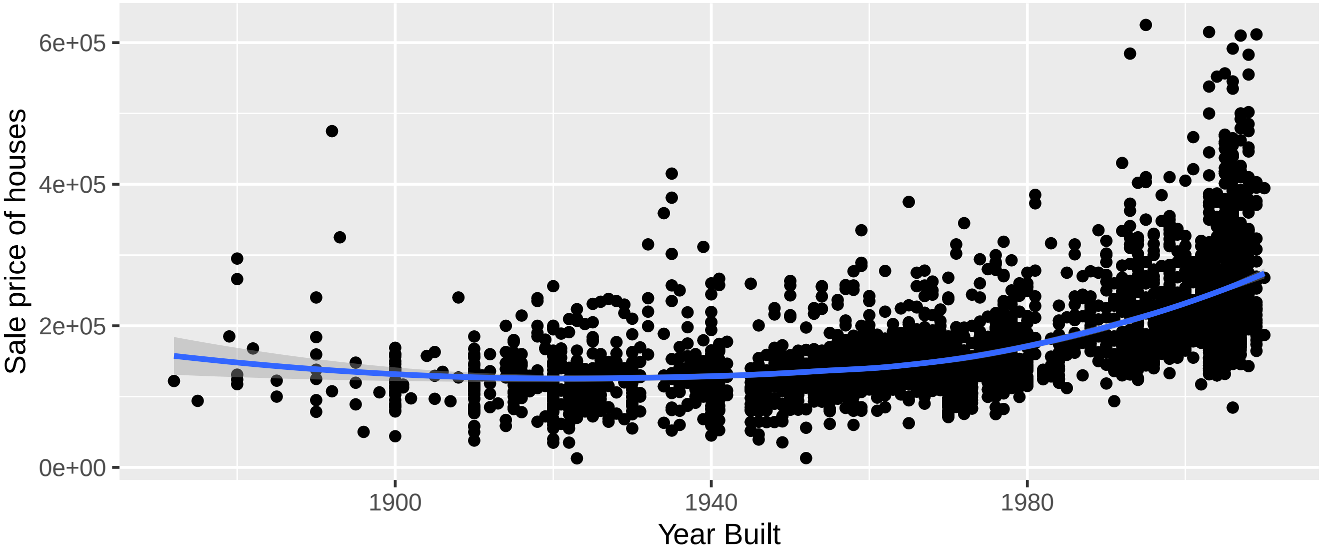


Figure 22: Plot of Sale Price - Year Built

Total Bsmt SF - Total square feet of basement area. (basement - below the ground floor)

Gr Liv Area - Above ground living area square feet

The total square footage model indicates some possible curvature (convex) which could be better interpreted with quadratic variables.

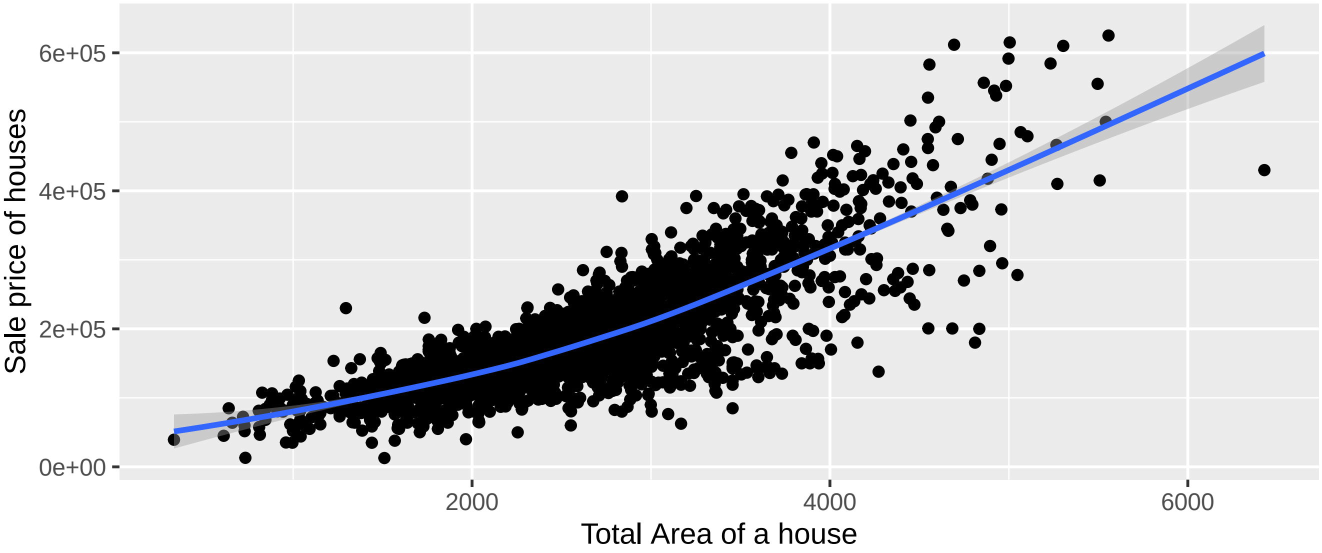


Figure 23: Plot of Sale Price - Total Area

Total basement area and Above ground area with Sale Condition

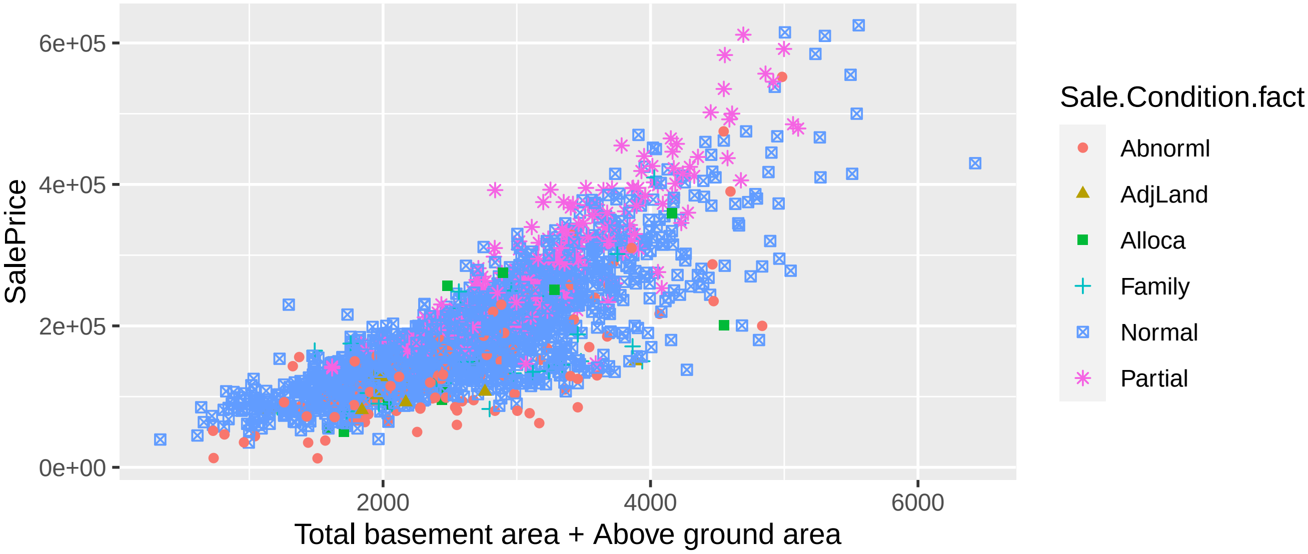


Figure 24: Plot of Sale Price - Total Area with Sale Condition

Model choice

Regression 1.1:

SalePrice = Beta0 + Beta1 * Lot Area + Beta2 * HasFireplace

R2 = 0.27

Regression 1.2: SalePrice = Beta0 + Beta1 * Total Area

R2 = 68

Even though this model fits the data very well, we dont take this model as TotalArea is higly correlated to SalePrice - multicollinearity issue.

Regression 1.3: SalePrice = Beta0 + Beta1 * Garage Area

R2 = 42

Regression 1.4: SalePrice = Beta0 + Beta1 * Total square feet of basement area

R2 = 43

Regression 1.5 SalePrice = Beta0 + Beta1 * Basement quality

R2 = 37

Regression 1.6 SalePrice = Beta0 + Beta1 * MS.Zoning

R2 = 0.01

Regression 1.7 SalePrice = Beta0 + Beta1 * BsmtFin.Type.1

R2 = 0.11

Regression 1.8 SalePrice = Beta0 + Beta1 * Year.Built

R2 = 0.31

Regression 1.9 SalePrice = Beta0 + Beta1 * Year.Remod.Add

R2 = 0.29

Regression 2.0 SalePrice = Beta0 + Beta1 * Garage.Yr.Blt

R2 = 0.28

Regresssion 2.1 SalePrice = Beta0 + Beta1 * Garage.Cars

R2 = 0.42

Regression 2.2 SalePrice = Beta0 + Beta1 * TotRms.AbvGrd

R2 = 0.24

Since SalePrice - TotalArea plot had a curvature, I tried to make quadratic model as well.

Regression 2.4 SalePrice = Beta0 + Beta1 * TotalArea + Beta2 * TotalAreaˆ2

R2 = 0.68 - same as Regression 2.3

log - log: ln(Price) - ln(Lot Area) piecewise linear spline

Regression 2.5 log(SalePrice) = Alpha1 + Beta1 * log(Lot.Area)[if log(Lot.Area) < 8] + (Alpha2 + Beta2 * log(Lot.Area)) * [if 8 <= log(Lot.Area) <= 10]

R2 = 0.14

Model: Weighted linear regression, using rooms as weights.

Regression 2

SalePrice = Above Ground Area (weights = Total Rooms Above Ground)

R2 = 0.49

Regression 3 SalePrice = Beta0 + Beta1 * Above Ground Area + Beta2 * HasFireplace

R2 = 55

Regression 4 SalePrice = Beta0 + Lot.Area + Beta1 * Total Basement Area + Beta2 * Above Ground Area + Beta3 * Garage Cars + Beta4 * HasFireplace

R2 = 74

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	124597.47 ***	-36147.77 ***	-2713214.78 ***	61328.87 ***	18686.95 **
	(4428.37)	(3967.50)	(89612.76)	(2976.85)	(6940.71)
Lot.Area	1.91 ***				
	(0.50)				
HasFireplace	71009.76 ***				
	(2594.67)				
TotalArea		85.24 ***			
		(1.74)			
Year.Built			1467.88 ***		
			(45.64)		
Garage.Cars				67472.84 ***	
				(1939.55)	
TotRms.AbvGrd					25135.20 ***
					(1158.60)
nobs	2925	2924	2925	2924	2925
r.squared	0.27	0.68	0.32	0.43	0.25
adj.r.squared	0.27	0.68	0.32	0.43	0.25
statistic	487.43	2404.64	1034.49	1210.20	470.65
p.value	0.00	0.00	0.00	0.00	0.00
df.residual	2922.00	2922.00	2923.00	2922.00	2923.00
nobs.1	2925.00	2924.00	2925.00	2924.00	2925.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00	HC2.00

*** p < 0.001; ** p < 0.01; * p < 0.05.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	-36147.77 ***	10.80 ***	5536.80	12350.04 **	-41975.51 ***
	(3967.50)	(0.51)	(5124.24)	(3925.43)	(3931.52)
TotalArea	85.24 ***				
	(1.74)				
lspline(log(Lot.Area), c(8, 10))1		0.11			
		(0.07)			
lspline(log(Lot.Area), c(8, 10))2		0.34 ***			
		(0.02)			
lspline(log(Lot.Area), c(8, 10))3		0.07			
		(0.07)			
Gr.Liv.Area			116.00 ***	101.43 ***	64.90 ***
			(3.73)	(3.21)	(2.57)
HasFireplace				32173.07 ***	15526.05 ***
				(2063.99)	(1630.98)
Lot.Area					0.27 *
					(0.12)
Total.Bsmt.SF					64.33 ***
					(2.85)
Garage.Cars					26830.53 ***
					(1278.32)
nobs	2924	2925	2925	2925	2923
r.squared	0.68	0.14	0.50	0.55	0.74
adj.r.squared	0.68	0.14	0.50	0.55	0.74
statistic	2404.64	193.46	966.20	1028.33	788.89
p.value	0.00	0.00	0.00	0.00	0.00
df.residual	2922.00	2921.00	2923.00	2922.00	2917.00
nobs.1	2924.00	2925.00	2925.00	2925.00	2923.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.