A Regression Model to Understand the Effects on Housing Prices

Jagdish Chhabria, Diego Correa, Stephen Haslett, Orli Khaimova, and Richard Zheng

M.S. in Data Science, CUNY School of Professional Studies

DATA 621: Business Analytics and Data Mining

Professor Nasrin Khansari

May 23, 2021

Abstract

The data set that was used describes the sales of residential properties from 2006 to 2010 in Ames, Iowa. There are 1460 observations with 79 predictor variables describing aspects of residential homes and the response variable being the sales price. The problem was to see whether or not the sales price can be predicted using a combination of these attributes. After exploring the data set to see how some attributes may be correlated, four multiple linear regression models were built using different methods. Afterwards, they were assessed on their ability to meet the assumptions of a linear model and compared using the adjusted R-squared in order to choose the best fitting multiple linear regression model. The paper concludes by using the final model to predict housing sales prices in a test data set.

Keywords: real estate, house prices, multiple regression, assessed value, home buyers, linear models

A Regression Model to Understand the Effects on Housing Prices

When hunting in the real estate market, many home buyers assume that the assessed value of the home is based on just a few factors such as location, the number of rooms, property, and so on. In reality, the sales price of the house is based on many factors that may not be independent of each other and fluctuate based on other factors. Furthermore, housing prices also affect consumption, are incorporated in the gross domestic product, and affect the economy as a whole.

It is also interesting to note how the current pandemic is affecting the housing market. According to the Congressional Research Service, contrary to most beliefs, “many housing market indicators have thus far remained strong” (Weinstock, 2021). Prices and sales have risen in the real estate market but the supply of homes on the market has decreased.

The goal of this study is to be able to predict the sales price of each house given the different variables and to find which variables may greatly influence the sales price. In order to explore this topic, we used the Ames Housing dataset, as produced by Dean De Cock, and later modified by Kaggle, in which it was separated by a training and test data set. The variables in the data set describe the quantity and quality of most of the physical attributes of a home.

Literature Review

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock is a professor of regression and statistics courses. He has previously used the Boston Housing Data Set for his own project and assignments. However, he was looking for a more updated version and finally came across a data set that was use by the Ames City Assessor’s Office. He removed variables that required further knowledge or previous calculations that were specific to the city’s modeling system and selected only the residential properties.

He offers advice throughout his publication such as removing outliers that may indicate partial sales or uncommon large houses. He also suggests testing each model manually on at least one observation to ensure that the model gives reasonable predictions. Additionally, he advises to keep track of complex variables that require transformations and interactions.

He also compares his students’ models at the end of each semester and highlights any special transformations. He evaluates each model by comparing the actual house prices to the predicated assessed value. He evaluates it based on the following four criteria: bias, maximum deviation, mean absolute deviation, and mean square error. It highlights where the model overestimated or underestimated the model, where it made its worst prediction, the average error, and the mean square error was used to “compare its calculate to the methodology used to obtain the coefficient estimates from the original data set” (De Cock, 2011).

In order to simplify the data, he suggests eliminating sales that are abnormal, such as foreclosure or newly built homes as they may skew the predictions. In order to remedy growing variation as the square footage of each house increases, he suggests to square root the sales price. He has also found the coefficients for continuous variables are affected when incorporating the neighborhood variable as it produces more realistic and expected coefficients. When dealing with ordinal variables, he suggests refactoring the levels in order to see how the succeeding level affects the model. When there are too many levels, he mentions it can be remedied by combining them. He has also found inconsistencies in the discrete variables that quantify the real estate property which can be fixed by “treating the variables as covariates which results in equal increases per item… or by once again collapsing the number of categories” (De Cock, 2011). He has created models that were simple with only three variables with an adjusted R^2 of 0.80 and models that were complex with 36 variables that produced an adjusted R^2 of 0.92.

Hedonic Housing Prices and the Demand for Clean Air

David Harrison, Jr. and Daniel L. Rubinfeld set out to measure how willing people are to pay for clean air using housing data from the Boston area in 1970. Problems with air pollution tend to increase with the concentration of air pollution and the household income. They mentioned that their results are “relatively sensitive to the specification of the hedonic housing price equation, but insensitive to the specification of the air quality demand equation” (Harrison & Rubinfeld, 1978).

They have assumed that individuals in the real estate market will pay more for a house in an area with better air quality. Their methodology was to first “estimate a hedonic housing value equation with air pollution as one house attribute” and then compute how much each household is willing to pay for a “marginal change” in the air pollution that was calculated from the previous equation (Harrison & Rubinfeld, 1978). Third, they approximated a “marginal willingness-to-pay function” which is similar to the demand curve. Lastly, they used the function and the estimated air pollution concentrations from before and after to compute the dollar amount.

Their model took into consideration the air pollution level and typical house characteristics that denote the quality and quantity, as well as the neighborhood characteristics. Most people do not view a house as a single item, but rather, as a combination of many attributes. The hedonic housing value equation “translates a vector of housing attributes at each location into a price which influences the decisions of both supplies and demanders of housing attributes” (Harrison & Rubinfeld, 1978). They also mention that it is not a linear relationship between housing attributes and its assessed price. They found that taking the logarithm of the sale price produced a better fit. Their model was used to account for problems that other researchers has in which they disregarded that improved air pollution concentration is dependent on household income and other housing attributes. In the end, they conclude that “marginal air pollution damages…. increase with the level of air pollution and increase with the level of household income” (Harrison & Rubinfeld, 1978).

Understanding Recent Trends in House Prices and Home Ownership

Robert J. Shiller investigates some of the factors that affect housing booms. As well as physical factors influencing housing prices, Shiller argues that there are also psychological factors at play such as society's insistence that housing is an important investment.

Cracking the Ames Housing Dataset with Linear Regression

Alvin T. Tan Investigates the Ames Housing dataset from the point of view of the hedonic pricing regression technique. According to the Organisation For Economic Co-Operation and Development, the hedonic method is "A regression technique in which observed prices of different qualities or models of the same generic good or service are expressed as a function of the characteristics of the goods or services in question. It is based on the hypothesis that products can be treated as bundles of characteristics and that prices can be attached to the characteristics" (2005). Basically, items can be broken into their constituent parts and used to predict the target value based on how much influence those parts bear.

Methodology

In our investigation to predict house prices, we are given eighty-one categorical and numerical independent variables to use in a multiple linear regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_1 X_n + \hat{\epsilon}_i$$

We begin our exploratory analysis by taking a look into all the variables in our data set. We take a glimpse into how the data looks like within the variables in the data set. From there, we identify the categorical and numerical variables and the number of missing values by variable. Next in our data exploration, we visualize the variables and its distribution using boxplots. Lastly, we illustrate the correlation of our independent variables to our dependent variable, house prices.

The preparation for the data set only includes the imputation of our missing values using classification regression tree. The data set used is otherwise clean -- variables' data types are properly stored.

In creating multiple linear models, we begin with our first model that includes all eighty independent variables to predict the house prices. The next model created is the model that only includes the numerical values, ignoring the categorical independent variables. The third model takes model one and performs feature engineering using backward stepwise elimination. The fourth and last model performs transformations on the training set to eliminate outliers and condenses the number of features. The four assumptions used in our model are:

1) Residuals of the model are nearly normal.
2) Variability of the residuals is nearly constant.
3) Residuals are independent.
4) Each variable is linearly related to the outcome.

We evaluate these assumptions through plotting the residuals in a quantile-quantile (q-q) plot, distribution plot, and by taking the absolute values against the fitted values to determine that the variability and distribution are normal.

In choosing the best fitting multiple linear regression model, we prioritize the model's adjusted R-squared values and the variability of the residuals.

Experimentation and Results

The Ames Housing training data set consists of 81 variables describing the characteristics of 1,460 homes in Ames, Iowa sold between 2006 and 2010. The dataset is available for download via the Kaggle website. The Ames Housing dataset is feature rich, and contains many of the features that home buyers consider when buying a house such as overall condition, location, number of rooms, etc. Table 1 is a summary of the variables contained within the dataset. Further descriptions of the variables can be found in the AMES dataset description.

The first step in our data analysis is to get a feel for the data by generating a glimpse of the dataset. As we can see from Table 2, the 81 variables contained within the dataset are a mixture of integer, and factor variables.

The dataset glimpse results above also reveal something else to us - many of the columns contain missing values which could be problematic when it comes to generating our models. This deserves further investigation, so we will now hone in on these columns to get an idea of the quantity of missing values contained within each column. Table 3 represents a count of missing values per column in descending order.

The first order of business when it comes to data preparation is to deal with the missing values in the data. Looking at Table 3, it appears that there are quite a few columns containing NA values (19 columns in all). However, according to the AMES dataset description, some variables contain genuine NA values that have meaning within the context of the data. For example, an "NA" value in the Alley column represents "No alley access", an "NA" value within the "BsmtQual" column represents "No Basement", and so on. Therefore, to prevent these from being interpreted as true empty NA values, we imputed them to have more meaningful values (i.e. NoAlleyAccess, NoBasement), and ran the empty values check again. Table 4 shows the effects after the NA values were replaced.

We explored the data further by looking at the correlations. Table 5 shows the correlations with the numeric predictors and Table 6 shows the correlations with numeric predictors against the sales price.The amount of cars in the garage is correlated with the garage area. As expected, the year the house was built is highly correlated with the year the garage was built. Similarly, the basement is correlated to the first floor's square footage. The more rooms a house has, the greater the living area is. There is also a correlation between the finished square footage of the basement and the amount of full baths it has.

The overall quality of the material and finish of the house seems to have the greatest effect on the sales price. The second variable to have a great effect is the total living square footage of the house, followed by the amount of cars in the garage and its square footage. Other variables that are correlated with the sales price are the total square footage of the basement and first floors, the amount of full bathrooms, total rooms above grade, the year the house was built in, and the year it was remodeled. It should be noted that if the year the house was remodeled is equal to the year the house was built in, it means that the house has not been remodeled since building.

Table 7 shows different graphs that were explored. There seems to be an overall positive relationship between the total living square footage and sales prices. As shown by the graph, there is an outlier due to a house being much larger than all the other houses in the city.

The sales price for each house tends to vary more as the overall quality increases, which rate the overall material and finish of the house. The overall condition of the house does not affect the sales price as one would expect because houses with only a condition of 5 out of 10 tend to sell for more whereas homes with an overall condition of 6-9 show greater variation in price compared to homes with an overall condition of 1-4. Single family homes also tend to vary in price compared to the other building types. The sales prices increase as the amount of cars in the garage increase from 0 to 3, but then decrease when there are 4 cars. Homes that have amenities or special features such as central air conditioning and fireplaces are more likely to sell for more.

From the imputation of all of the genuine NA values in the dataset and re-counting, we can see that the top offending variables are no longer listed as having missing values - great news. However, we are still left with 38 variables that contain missing values, so our next order of business is to deal with these variables.

There are several imputation options available to us at this point. We can do nothing, which will hinder the quality of our models, remove observations that contain missing values, which may affect the accuracy of the results but its best to avoid the option, use Multivariate Imputation by Chained Equation (MICE), k-nearest neighbors, or impute using mean/median values.

Taking a look at the Glimpse report we generated at the beginning of our study in Table 2, our dataset consists of both numerical and categorical variables. For this reason, the MICE imputation method would appear to be our best option as it deals with both numerical and categorical variables.

After running the MICE algorithm on our dataset and re-running the empty values check, we were left with zero missing values as reflected in Table 8.

Now that we have a dataset that is free from empty values, we can move on to building our models.

Model 1 was built using all the variables in the training data set. The coefficients are not reasonable as you cannot have a negative sales price with no attributes. There were some NA values produced in the model and the adjusted R-squared is 0.9063 with a small p-value and an F-statistic of 62.07on 231 and 1228 degrees of freedom.

Model 2 was built by filtering out non numeric values and keeping only the numeric variables. Some NA values were still generated and the adjusted R-squared worsened as it became 0.8086 with a small p-value and an F-statistic of 182.3 on 34 and 1425 degrees of freedom.

Model 3 utilizes model 1, but uses backward stepwise elimination, by filtering out insignificant variables. It should be noted that this model is the most costly as it takes the longest to run. The adjusted R-squared improved as it became 0.9076 with a small p-value and an F-statistic of 118.5 on 122 and 1337 degrees of freedom.

Model 4 utilized a mix of forward selection and recommendations from De Cock’s findings. He suggested including the neighborhood in the model as sales prices vary from one model to another. Also, less than a handful of outliers were removed in which the living square footage exceeded 4,000 as it can be seen that it skews our data in Table 7. The bathrooms were combined to form a new variable for *TotalBath*, giving half baths a weight of 0.5. The porch square footage was also combined into one variable. The age of the houses were included as it can be found by subtracting the year it was built from the year it was sold. The sale condition was refactored into normal and other, and some homes were traded, foreclosed, partially not completed, or sold between family members. Another variable to depict new homes was created since new homes tend to sell for more compared to other homes. Most importantly, the sales price is transformed logarithmically, to remedy the increasing variation as the square footage of each house increases. The adjusted R-squared improved with a small p-value and an F-statistic of 346.1 on 45 and 1410 degrees of freedom.

Comparing these four models, we selected model 4 due to it fitting the data well as its adjusted R-squared is the highest at 0.9143, which is slightly more than model 3 with an adjusted R-squared of 0.9077, which was more computationally expensive. We think model 4 would be a better fit for a production environment where such a model might be used, to predict housing prices on demand, as it is a great fit and computes reasonably fast. Table 9 shows the first 5 predicted sales prices for the testing data set.

Discussion and Conclusions

Applying transformations to the variables improved model performance. Square footage of the living area and neighborhood were most impactful when determining housing prices as they explain a great amount of variation alone. Some limitations included that this dataset only contained houses in Ames, Iowa and a different city may require a different model. Many variables were heavily unbalanced towards one level.

References

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, *19*(3). https://doi.org/10.1080/10691898.2011.11889627

De Cock, D. (n.d.). *Data Documentation* . Journal of Statistics Education (JSE) Home Page. http://jse.amstat.org/v19n3/decock/DataDocumentation.txt. This source clarifies the variables in the data set further.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Kaggle. (2012). House Prices - Advanced Regression Techniques, Version 1. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

OECD Statistics. (2005, July 8). *Hedonic Method*. OECD Glossary of Statistical Terms. https://stats.oecd.org/glossary/detail.asp?ID=1225.

Shiller, R. J. (2007). Understanding Recent Trends in House Prices and Homeownership. In *Proceedings - Economic Policy Symposium - Jackson Hole* (pp. 89–123), Federal Reserve Bank of Kansas City. https://www.kansascityfed.org/documents/3224/pdf-Shiller_0415.pdf

Tan, A. T. (2021, May 6). *Cracking the Ames Housing Dataset with Linear Regression*. Medium. https://towardsdatascience.com/wrangling-through-dataland-modeling-house-prices-in-ames-iowa-75b9b4086c96.

Weinstock, L. R. (2021, May 3). *Introduction to U.S. Economy: Housing Market*. Congressional Research Service. https://fas.org/sgp/crs/misc/IF11327.pdf.

Appendices

Predicting Property Prices

Group 2

5/16/2021

Contents

Model Building 12

Group 2 members: *Diego Correa, Jagdish Chhabria, Orli Khaimova, Richard Zheng, Stephen Haslett.*

Table 1: Data Set Variables

- PID: Parcel identification number - can be used with city web site for parcel review.
- MS SubClass: Identifies the type of dwelling involved in the sale.
- MS Zoning: Identifies the general zoning classification of the sale.
- Lot Frontage: Linear feet of street connected to property
- Lot Area: Lot size in square feet
- Street: Type of road access to property
- Alley: Type of alley access to property
- Lot Shape: General shape of property
- Land Contour: Flatness of the property
- Utilities: Type of utilities available
- Lot Config: Lot configuration
- Land Slope: Slope of property
- Neighborhood: Physical locations within Ames city limits (map available)
- Condition 1: Proximity to various conditions
- Condition 2: Proximity to various conditions (if more than one is present)
- Bldg Type: Type of dwelling
- House Style: Style of dwelling
- Overall Qual: Rates the overall material and finish of the house
- Overall Cond: Rates the overall condition of the house
- Year Built: Original construction date
- Year Remod/Add: Remodel date (same as construction date if no remodeling or additions)
- Roof Style: Type of roof
- Roof Matl: Roof material
- Exterior 1: Exterior covering on house
- Exterior 2: Exterior covering on house (if more than one material)
- Mas Vnr Type: Masonry veneer type
- Mas Vnr Area: Masonry veneer area in square feet
- Exter Qual: Evaluates the quality of the material on the exterior
- Exter Cond: Evaluates the present condition of the material on the exterior
- Foundation: Type of foundation
- Bsmt Qual: Evaluates the height of the basement

- Bsmt Cond: Evaluates the general condition of the basement
- Bsmt Exposure: Refers to walkout or garden level walls
- BsmtFin Type 1: Rating of basement finished area
- BsmtFin SF 1: Type 1 finished square feet
- BsmtFinType 2: Rating of basement finished area (if multiple types)
- BsmtFin SF 2: Type 2 finished square feet
- Bsmt Unf SF: Unfinished square feet of basement area
- Total Bsmt SF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- Central Air: Central air conditioning
- Electrical: Electrical system
- 1st Flr SF: First Floor square feet
- 2nd Flr SF: Second floor square feet
- Low Qual Fin SF: Low quality finished square feet (all floors)
- Gr Liv Area: Above grade (ground) living area square feet
- Bsmt Full Bath: Basement full bathrooms
- Bsmt Half Bath: Basement half bathrooms
- Full Bath: Full bathrooms above grade
- Half Bath: Half baths above grade
- Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
- Kitchen: Kitchens above grade
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality (Assume typical unless deductions are warranted)
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- Garage Type: Garage location
- Garage Yr Blt: Year garage was built
- Garage Finish: Interior finish of the garage
- Garage Cars: Size of garage in car capacity
- Garage Area: Size of garage in square feet
- Garage Qual: Garage quality
- Garage Cond: Garage condition
- Paved Drive: Paved driveway
- Wood Deck SF: Wood deck area in square feet
- Open Porch SF: Open porch area in square feet
- Enclosed Porch: Enclosed porch area in square feet
- 3-Ssn Porch: Three season porch area in square feet
- Screen Porch: Screen porch area in square feet
- Pool Area: Pool area in square feet
- Pool QC: Pool quality
- Fence: Fence quality
- Misc Feature: Miscellaneous feature not covered in other categories
- Misc Val: $Value of miscellaneous feature
- Mo Sold: Month Sold
- Yr Sold: Year Sold
- Sale Type: Type of sale
- Sale Condition: Condition of sale

Table 2: Glimpse of the Data Set The first step in our data analysis is to get a feel for the data by generating a glimpse of the dataset. As we can see from the below results, the 81 variables contained within the dataset are a mixture of integer, and factor variables.

```
## Rows: 1,460
## Columns: 81
## $ Id            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ MSSubClass    <int> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, ...
## $ MSZoning      <fct> RL, RL, RL, RL, RL, RL, RL, RL, RM, RL, RL, RL, RL, R...
## $ LotFrontage   <int> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, NA, 9...
## $ LotArea       <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, ...
## $ Street        <fct> Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave,...
## $ Alley         <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ LotShape      <fct> Reg, Reg, IR1, IR1, IR1, IR1, Reg, IR1, Reg, Reg, Reg...
## $ LandContour   <fct> Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl...
## $ Utilities     <fct> AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllPu...
## $ LotConfig     <fct> Inside, FR2, Inside, Corner, FR2, Inside, Inside, Cor...
## $ LandSlope     <fct> Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl...
## $ Neighborhood  <fct> CollgCr, Veenker, CollgCr, Crawfor, NoRidge, Mitchel,...
## $ Condition1    <fct> Norm, Feedr, Norm, Norm, Norm, Norm, Norm, PosN, Arte...
## $ Condition2    <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm,...
## $ BldgType      <fct> 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam,...
## $ HouseStyle    <fct> 2Story, 1Story, 2Story, 2Story, 2Story, 1.5Fin, 1Stor...
## $ OverallQual   <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4,...
## $ OverallCond   <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5,...
## $ YearBuilt     <int> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931,...
## $ YearRemodAdd  <int> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950,...
## $ RoofStyle     <fct> Gable, Gable, Gable, Gable, Gable, Gable, Gable, Gabl...
## $ RoofMatl      <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg,...
## $ Exterior1st   <fct> VinylSd, MetalSd, VinylSd, Wd Sdng, VinylSd, VinylSd,...
## $ Exterior2nd   <fct> VinylSd, MetalSd, VinylSd, Wd Shng, VinylSd, VinylSd,...
## $ MasVnrType    <fct> BrkFace, None, BrkFace, None, BrkFace, None, Stone, S...
## $ MasVnrArea    <int> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 30...
## $ ExterQual     <fct> Gd, TA, Gd, TA, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, G...
## $ ExterCond     <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ Foundation    <fct> PConc, CBlock, PConc, BrkTil, PConc, Wood, PConc, CBl...
## $ BsmtQual      <fct> Gd, Gd, Gd, TA, Gd, Gd, Ex, Gd, TA, TA, TA, Ex, TA, G...
## $ BsmtCond      <fct> TA, TA, TA, Gd, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ BsmtExposure  <fct> No, Gd, Mn, No, Av, No, Av, Mn, No, No, No, No, No, A...
## $ BsmtFinType1  <fct> GLQ, ALQ, GLQ, ALQ, GLQ, GLQ, GLQ, ALQ, Unf, GLQ, Rec...
## $ BsmtFinSF1    <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906,...
## $ BsmtFinType2  <fct> Unf, Unf, Unf, Unf, Unf, Unf, Unf, BLQ, Unf, Unf, Unf...
## $ BsmtFinSF2    <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ BsmtUnfSF     <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140, 134,...
## $ TotalBsmtSF   <int> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952, 991,...
## $ Heating       <fct> GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA,...
## $ HeatingQC     <fct> Ex, Ex, Ex, Gd, Ex, Ex, Ex, Ex, Gd, Ex, Ex, Ex, TA, E...
## $ CentralAir    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ Electrical    <fct> SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrk...
## $ X1stFlrSF     <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 107...
## $ X2ndFlrSF     <int> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142,...
## $ LowQualFinSF  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ GrLivArea     <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774,...
## $ BsmtFullBath  <int> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...
## $ BsmtHalfBath  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ FullBath      <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2,...
## $ HalfBath      <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,...
## $ BedroomAbvGr  <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2,...
## $ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ KitchenQual   <fct> Gd, TA, Gd, Gd, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, G...
## $ TotRmsAbvGrd  <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6...
## $ Functional    <fct> Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Min1, Typ, Ty...
## $ Fireplaces    <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0,...
## $ FireplaceQu   <fct> NA, TA, TA, Gd, TA, NA, Gd, TA, TA, TA, NA, Gd, NA, G...
## $ GarageType    <fct> Attchd, Attchd, Attchd, Detchd, Attchd, Attchd, Attch...
## $ GarageYrBlt   <int> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, 1931,...
## $ GarageFinish  <fct> RFn, RFn, RFn, Unf, RFn, Unf, RFn, RFn, Unf, RFn, Unf...
## $ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2,...
## $ GarageArea    <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384...
## $ GarageQual    <fct> TA, TA, TA, TA, TA, TA, TA, TA, Fa, Gd, TA, TA, TA, T...
## $ GarageCond    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ PavedDrive    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ WoodDeckSF    <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, ...
## $ OpenPorchSF   <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 2...
## $ EnclosedPorch <int> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, 176, ...
## $ X3SsnPorch    <int> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ScreenPorch   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0, 0, ...
## $ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ PoolQC        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Fence         <fct> NA, NA, NA, NA, NA, MnPrv, NA, NA, NA, NA, NA, NA, NA...
## $ MiscFeature   <fct> NA, NA, NA, NA, NA, Shed, NA, Shed, NA, NA, NA, NA, N...
## $ MiscVal       <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 7...
## $ MoSold        <int> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3,...
## $ YrSold        <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008,...
## $ SaleType      <fct> WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, New, WD, ...
## $ SaleCondition <fct> Normal, Normal, Normal, Abnorml, Normal, Normal, Norm...
## $ SalePrice     <int> 208500, 181500, 223500, 140000, 250000, 143000, 30700...
```

Table 3: Missing Values

Table 4: Replacing NA values with meaningful values

	x
PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	690
LotFrontage	259
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
BsmtExposure	38
BsmtFinType2	38
BsmtQual	37
BsmtCond	37
BsmtFinType1	37
MasVnrType	8
MasVnrArea	8
Electrical	1
Id	0
MSSubClass	0
MSZoning	0
LotArea	0
Street	0
LotShape	0
LandContour	0
Utilities	0
LotConfig	0
LandSlope	0
Neighborhood	0
Condition1	0
Condition2	0
BldgType	0
HouseStyle	0
OverallQual	0
OverallCond	0
YearBuilt	0
YearRemodAdd	0
RoofStyle	0
RoofMatl	0
Exterior1st	0
Exterior2nd	0
ExterQual	0
ExterCond	0
Foundation	0
BsmtFinSF1	0
BsmtFinSF2	0
BsmtUnfSF	0
TotalBsmtSF	0
Heating	0
HeatingQC	0
CentralAir	0
X1stFlrSF	0
X2ndFlrSF	0
LowQualFinSF	0
GrLivArea	0
BsmtFullBath	0
BsmtHalfBath	0

	x
LotFrontage	259
GarageYrBlt	81
MasVnrType	8
MasVnrArea	8
Electrical	1
Id	0
MSSubClass	0
MSZoning	0
LotArea	0
Street	0
Alley	0
LotShape	0
LandContour	0
Utilities	0
LotConfig	0
LandSlope	0
Neighborhood	0
Condition1	0
Condition2	0
BldgType	0
HouseStyle	0
OverallQual	0
OverallCond	0
YearBuilt	0
YearRemodAdd	0
RoofStyle	0
RoofMatl	0
Exterior1st	0
Exterior2nd	0
ExterQual	0
ExterCond	0
Foundation	0
BsmtQual	0
BsmtCond	0
BsmtExposure	0
BsmtFinType1	0
BsmtFinSF1	0
BsmtFinType2	0
BsmtFinSF2	0
BsmtUnfSF	0
TotalBsmtSF	0
Heating	0
HeatingQC	0
CentralAir	0
X1stFlrSF	0
X2ndFlrSF	0
LowQualFinSF	0
GrLivArea	0
BsmtFullBath	0
BsmtHalfBath	0
FullBath	0
HalfBath	0
BedroomAbvGr	0
KitchenAbvGr	0
KitchenQual	0
TotRmsAbvGrd	0
Functional	0
Fireplaces	0

Table 1: Correlations of numeric predictors

row	column	cor	p
GarageCars	GarageArea	0.8824754	0
YearBuilt	GarageYrBlt	0.8256675	0
GrLivArea	TotRmsAbvGrd	0.8254894	0
TotalBsmtSF	X1stFlrSF	0.8195300	0
OverallQual	SalePrice	0.7909816	0
GrLivArea	SalePrice	0.7086245	0
X2ndFlrSF	GrLivArea	0.6875011	0
BedroomAbvGr	TotRmsAbvGrd	0.6766199	0
BsmtFinSF1	BsmtFullBath	0.6492118	0
YearRemodAdd	GarageYrBlt	0.6422768	0

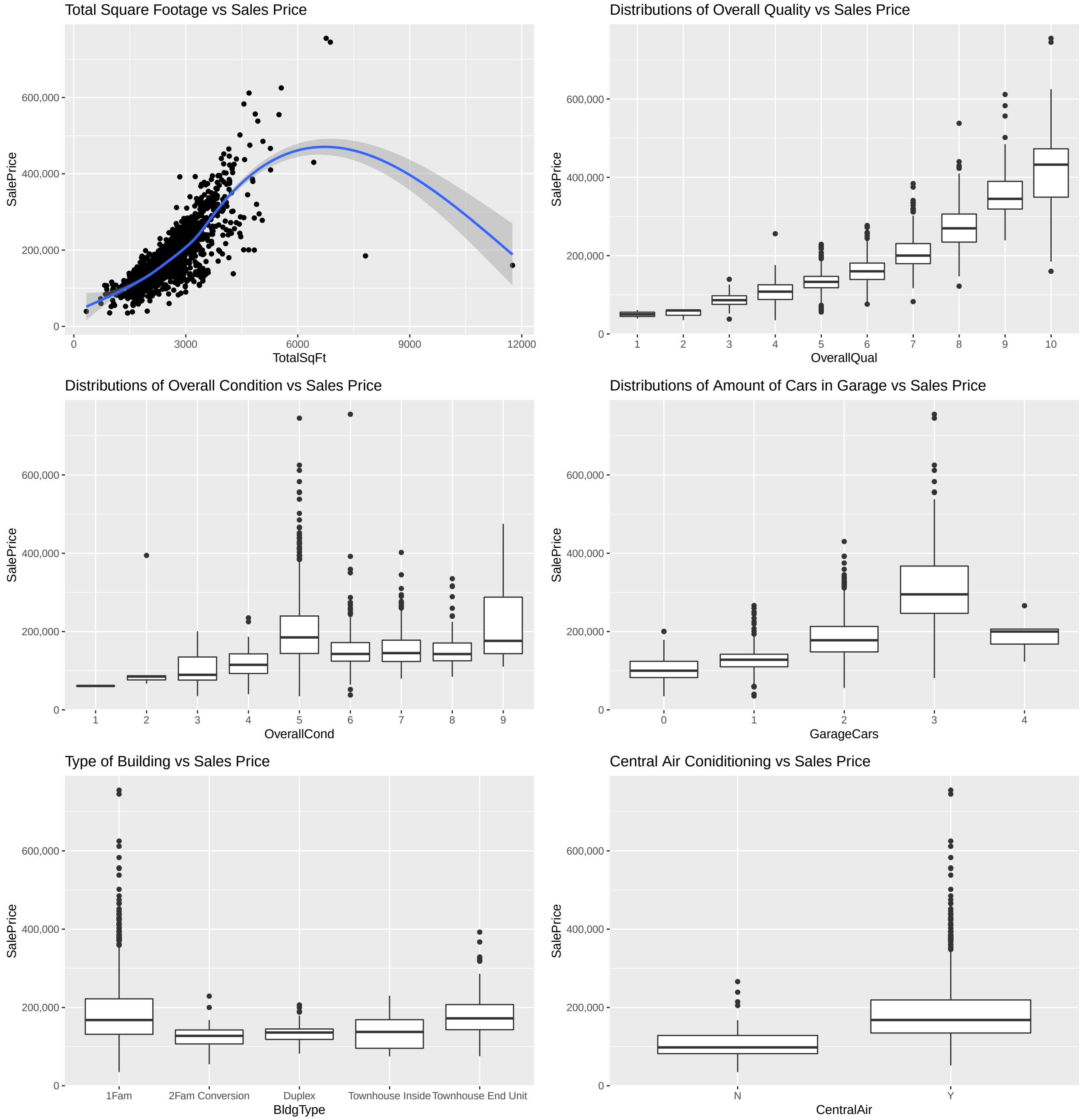
Table 5: Correlations of Numeric Predictors

Table 2: Correlations of numeric predictors against the Sales Price

row	column	cor	p
OverallQual	SalePrice	0.7909816	0
GrLivArea	SalePrice	0.7086245	0
GarageCars	SalePrice	0.6404092	0
GarageArea	SalePrice	0.6234314	0
TotalBsmtSF	SalePrice	0.6135806	0
X1stFlrSF	SalePrice	0.6058522	0
FullBath	SalePrice	0.5606638	0
TotRmsAbvGrd	SalePrice	0.5337232	0
YearBuilt	SalePrice	0.5228973	0
YearRemodAdd	SalePrice	0.5071010	0

Table 6: Correlations of Numeric Predictors Against the Sales Price

Table 7: Various Plots .



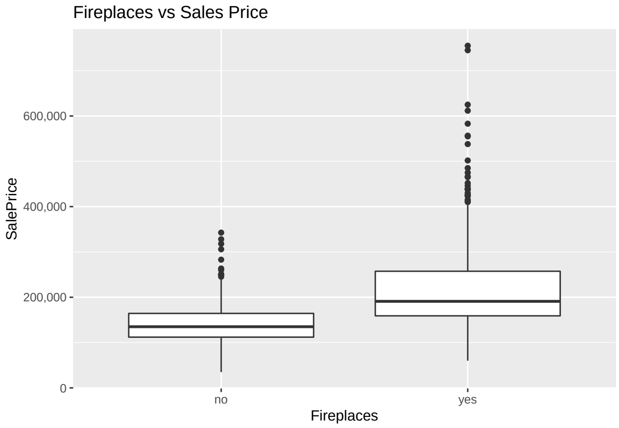

Fireplaces vs Sales Price
SalePrice
600,000
400,000
200,000
0
no
yes
Fireplaces

Table 8: Missing Values Post-MICE Imputation

```
##
##  iter imp variable
##   1   1  LotFrontage  MasVnrType  MasVnrArea  Electrical  GarageYrBlt
##   2   1  LotFrontage  MasVnrType  MasVnrArea  Electrical  GarageYrBlt
##   3   1  LotFrontage  MasVnrType  MasVnrArea  Electrical  GarageYrBlt
##   4   1  LotFrontage  MasVnrType  MasVnrArea  Electrical  GarageYrBlt
##   5   1  LotFrontage  MasVnrType  MasVnrArea  Electrical  GarageYrBlt
```

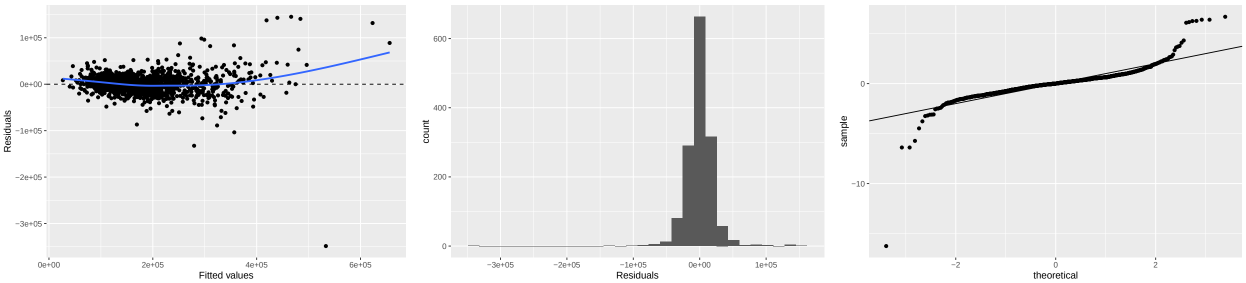
	x
MSSubClass	0
MSZoning	0
LotFrontage	0
LotArea	0
Street	0
Alley	0
LotShape	0
LandContour	0
Utilities	0
LotConfig	0
LandSlope	0
Neighborhood	0
BldgType	0
HouseStyle	0
OverallQual	0
OverallCond	0
YearBuilt	0
YearRemodAdd	0
RoofMatl	0
Exterior1st	0
Exterior2nd	0
MasVnrType	0
MasVnrArea	0
ExterQual	0
ExterCond	0
Foundation	0
BsmtQual	0
BsmtCond	0
BsmtExposure	0
BsmtFinType1	0
BsmtFinSF1	0
BsmtFinType2	0
BsmtFinSF2	0
BsmtUnfSF	0
TotalBsmtSF	0
Heating	0
HeatingQC	0
CentralAir	0
Electrical	0
X1stFlrSF	0
X2ndFlrSF	0
LowQualFinSF	0
GrLivArea	0
BsmtFullBath	0
BsmtHalfBath	0
FullBath	0
HalfBath	0
BedroomAbvGr	0
KitchenAbvGr	0
KitchenQual	0
TotRmsAbvGrd	0
Functional	0
Fireplaces	0
FireplaceQu	0
GarageType	0
GarageYrBlt	0
GarageFinish	0
GarageC	0

Model Building

Now that we have a dataset that is free from empty values, we can move on to building our models.

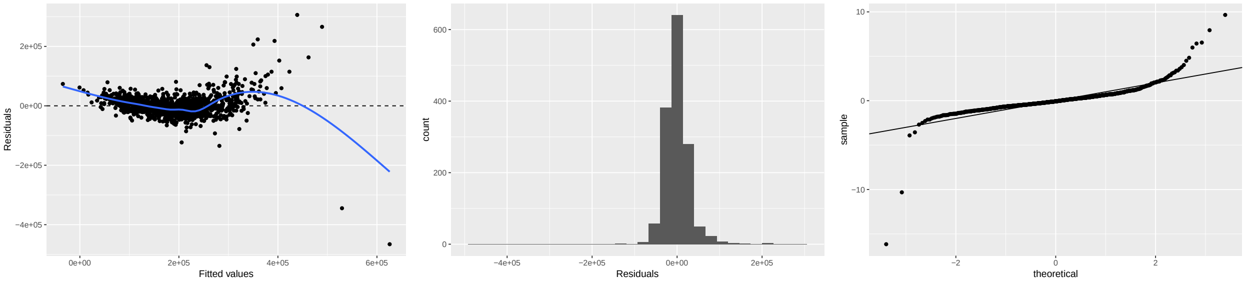
Model One

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348585   -9806       0    9812  145215
##
## Coefficients: (8 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -4.240e+05  1.127e+06  -0.376 0.706920
## MSSubClass             -8.925e+01  8.806e+01  -1.014 0.310964
## MSZoningFV              3.625e+04  1.274e+04   2.844 0.004523 **
## MSZoningRH              2.577e+04  1.264e+04   2.039 0.041687 *
## MSZoningRL              2.672e+04  1.078e+04   2.480 0.013272 *
## MSZoningRM              2.322e+04  1.011e+04   2.296 0.021848 *
## LotFrontage             6.985e+00  4.403e+01   0.159 0.873961
## LotArea                 5.254e-01  1.135e-01   4.628 4.08e-06 ***
## StreetPave              2.827e+04  1.283e+04   2.204 0.027681 *
## AlleyPave               1.101e+03  6.417e+03   0.172 0.863798
## AlleyNoAlleyAccess     -5.309e+02  4.504e+03  -0.118 0.906186
## LotShapeIR2             6.016e+03  4.508e+03   1.335 0.182285
## LotShapeIR3             9.383e+03  9.379e+03   1.000 0.317315
## LotShapeReg             1.741e+03  1.693e+03   1.028 0.304088
## LandContourHLS          1.234e+04  5.428e+03   2.273 0.023227 *
## LandContourLow         -2.606e+03  6.754e+03  -0.386 0.699669
## LandContourLvl          1.004e+04  3.933e+03   2.554 0.010780 *
## UtilitiesNoSeWa        -3.986e+04  2.819e+04  -1.414 0.157705
## LotConfigCulDSac        7.716e+03  3.680e+03   2.097 0.036215 *
## LotConfigFR2           -8.611e+03  4.270e+03  -2.017 0.043946 *
## LotConfigFR3           -1.864e+04  1.293e+04  -1.442 0.149461
## LotConfigInside        -1.392e+03  1.910e+03  -0.729 0.466309
## LandSlopeMod            8.003e+03  4.235e+03   1.890 0.059050 .
## LandSlopeSev           -2.296e+04  1.161e+04  -1.978 0.048113 *
## NeighborhoodBlueste     6.749e+03  2.062e+04   0.327 0.743476
## NeighborhoodBrDale      3.114e+03  1.168e+04   0.267 0.789863
## NeighborhoodBrkSide    -1.865e+03  1.001e+04  -0.186 0.852308
## NeighborhoodClearCr    -9.424e+03  9.752e+03  -0.966 0.334046
## NeighborhoodCollgCr    -4.636e+03  7.762e+03  -0.597 0.550481
## NeighborhoodCrawfor     1.757e+04  9.109e+03   1.929 0.053902 .
## NeighborhoodEdwards    -2.176e+04  8.552e+03  -2.545 0.011048 *
## NeighborhoodGilbert    -7.915e+03  8.190e+03  -0.966 0.334026
## NeighborhoodIDOTRR     -8.251e+03  1.142e+04  -0.723 0.469964
## NeighborhoodMeadowV     4.947e+03  1.193e+04   0.415 0.678494
## NeighborhoodMitchel    -1.633e+04  8.740e+03  -1.869 0.061874 .
## NeighborhoodNAmes      -1.572e+04  8.381e+03  -1.876 0.060866 .
## NeighborhoodNoRidge     3.752e+04  8.968e+03   4.184 3.06e-05 ***
## NeighborhoodNPkVill     1.139e+04  1.503e+04   0.757 0.448936    
## NeighborhoodNridgHt     2.543e+04  8.010e+03   3.174 0.001539 ** 
## NeighborhoodNWAmes     -1.459e+04  8.478e+03  -1.721 0.085571 .  
## NeighborhoodOldTown    -1.308e+04  1.030e+04  -1.271 0.204010    
## NeighborhoodSawyer     -1.223e+04  8.691e+03  -1.407 0.159750    
## NeighborhoodSawyerW    -2.289e+03  8.316e+03  -0.275 0.783197    
## NeighborhoodSomerst     3.891e+03  9.527e+03   0.408 0.683026    
## NeighborhoodStoneBr     4.897e+04  8.855e+03   5.530 3.91e-08 ***
## NeighborhoodSWISU      -4.544e+03  1.038e+04  -0.438 0.661583    
## NeighborhoodTimber     -4.416e+03  8.672e+03  -0.509 0.610726    
## NeighborhoodVeenker     9.607e+02  1.123e+04   0.086 0.931832    
## BldgType2fmCon          3.947e+03  1.320e+04   0.299 0.764940    
## BldgTypeDuplex         -6.951e+03  7.849e+03  -0.886 0.376055    
## BldgTypeTwnhs          -1.716e+04  1.067e+04  -1.608 0.108044    
## BldgTypeTwnhsE         -1.200e+04  9.605e+03  -1.249 0.211887    
## HouseStyle1.5Unf        1.112e+04  8.372e+03   1.329 0.184157    
## HouseStyle1Story        6.664e+03  4.657e+03   1.431 0.152667    
## HouseStyle2.5Fin       -1.088e+04  1.315e+04  -0.828 0.408027    
## HouseStyle2.5Unf       -1.221e+03  9.446e+03  -0.129 0.897148    
## HouseStyle2Story       -4.531e+03  3.672e+03  -1.234 0.217466    
## HouseStyleSFoyer        5.985e+03  6.636e+03   0.902 0.367244    
## HouseStyleSLvl          7.510e+03  5.890e+03   1.275 0.202541    
## OverallQual             6.415e+03  1.077e+03   5.958 3.33e-09 ***
## OverallCond             5.661e+03  9.323e+02   6.072 1.68e-09 ***
## YearBuilt               2.750e+02  8.323e+01   3.303 0.000983 ***
## YearRemodAdd            9.550e+01  5.948e+01   1.605 0.108660    
## RoofMatlCompShg         4.919e+05  5.578e+04   8.818  < 2e-16 ***
## RoofMatlMembran         5.517e+05  6.395e+04   8.629  < 2e-16 ***
## RoofMatlMetal           5.215e+05  6.338e+04   8.228 4.79e-16 ***
## RoofMatlRoll            4.729e+05  6.207e+04   7.617 5.14e-14 ***
## RoofMatlTar&Grv         4.812e+05  5.791e+04   8.310 2.51e-16 ***
## RoofMatlWdShake         5.045e+05  5.754e+04   8.767  < 2e-16 ***
## RoofMatlWdShngl         5.510e+05  5.690e+04   9.683  < 2e-16 ***
## Exterior1stAsphShn     -2.915e+04  3.456e+04  -0.844 0.399034    
## Exterior1stBrkComm     -2.907e+03  2.966e+04  -0.098 0.921949    
## Exterior1stBrkFace      3.189e+03  1.351e+04   0.236 0.813411    
## Exterior1stCBlock      -2.588e+04  2.886e+04  -0.897 0.370004    
## Exterior1stCemntBd     -1.048e+04  2.026e+04  -0.517 0.604928    
## Exterior1stHdBoard     -1.797e+04  1.362e+04  -1.320 0.187242    
## Exterior1stImStucc     -2.894e+04  3.006e+04  -0.963 0.335909    
## Exterior1stMetalSd     -1.217e+04  1.536e+04  -0.792 0.428223    
## Exterior1stPlywood     -1.627e+04  1.344e+04  -1.210 0.226534    
## Exterior1stStone       -1.214e+04  2.533e+04  -0.479 0.631765    
## Exterior1stStucco      -6.954e+03  1.499e+04  -0.464 0.642792    
## Exterior1stVinylSd     -1.885e+04  1.410e+04  -1.337 0.181596    
## Exterior1stWd Sdng     -1.671e+04  1.306e+04  -1.280 0.200857    
## Exterior1stWdShing     -1.441e+04  1.412e+04  -1.020 0.307763    
## Exterior2ndAsphShn      2.428e+04  2.250e+04   1.079 0.280752    
## Exterior2ndBrk Cmn      6.845e+03  2.143e+04   0.319 0.749451    
## Exterior2ndBrkFace      6.097e+03  1.402e+04   0.435 0.663690    
## Exterior2ndCBlock              NA         NA      NA       NA    
## Exterior2ndCmentBd     -1.429e+03  1.994e+04  -0.072 0.942901    
## Exterior2ndHdBoard      9.805e+03  1.312e+04   0.747 0.455004    
## Exterior2ndImStucc      1.815e+04  1.513e+04   1.200 0.230514    
## Exterior2ndMetalSd      8.710e+03  1.503e+04   0.579 0.562478    
## Exterior2ndOther       -1.537e+04  2.904e+04  -0.529 0.596664    
## Exterior2ndPlywood      8.143e+03  1.275e+04   0.639 0.523149    
## Exterior2ndStone       -6.949e+03  1.823e+04  -0.381 0.703168    
## Exterior2ndStucco       4.382e+03  1.447e+04   0.303 0.762011    
## Exterior2ndVinylSd      1.503e+04  1.359e+04   1.106 0.268953    
## Exterior2ndWd Sdng      1.251e+04  1.267e+04   0.987 0.323856    
## Exterior2ndWd Shng      4.737e+03  1.322e+04   0.358 0.720185    
## MasVnrTypeBrkFace       3.988e+03  7.326e+03   0.544 0.586247    
## MasVnrTypeNone          5.100e+03  7.393e+03   0.690 0.490449    
## MasVnrTypeStone         1.014e+04  7.759e+03   1.307 0.191476    
## MasVnrArea              1.216e+01  6.131e+00   1.984 0.047495 *  
## ExterQualFa            -8.729e+03  1.171e+04  -0.745 0.456239    
## ExterQualGd            -1.716e+04  5.044e+03  -3.402 0.000691 ***
## ExterQualTA            -1.695e+04  5.579e+03  -3.038 0.002431 ** 
## ExterCondFa            -2.131e+04  1.672e+04  -1.274 0.202812    
## ExterCondGd            -2.446e+04  1.544e+04  -1.584 0.113494    
## ExterCondPo            -1.708e+04  3.234e+04  -0.528 0.597495    
## ExterCondTA            -2.073e+04  1.540e+04  -1.346 0.178411    
## FoundationCBlock        3.237e+03  3.393e+03   0.954 0.340252    
## FoundationPConc         3.739e+03  3.650e+03   1.025 0.305778    
## FoundationSlab         -4.481e+03  1.076e+04  -0.416 0.677174    
## FoundationStone         1.512e+04  1.188e+04   1.273 0.203312    
## FoundationWood         -2.062e+04  1.586e+04  -1.300 0.193848    
## BsmtQualFa             -1.452e+04  6.781e+03  -2.141 0.032447 *  
## BsmtQualGd             -1.990e+04  3.553e+03  -5.601 2.63e-08 ***
## BsmtQualTA             -1.631e+04  4.429e+03  -3.683 0.000241 ***
## BsmtQualNoBasement      3.125e+04  3.930e+04   0.795 0.426653    
## BsmtCondGd              2.010e+03  5.628e+03   0.357 0.721084    
## BsmtCondPo              5.075e+04  3.164e+04   1.604 0.108968    
## BsmtCondTA              3.369e+03  4.518e+03   0.746 0.455950    
## BsmtCondNoBasement             NA         NA      NA       NA    
## BsmtExposureGd          1.606e+04  3.206e+03   5.009 6.28e-07 ***
## BsmtExposureMn         -2.339e+03  3.215e+03  -0.728 0.466977    
## BsmtExposureNo         -4.246e+03  2.329e+03  -1.823 0.068540 .  
## BsmtExposureNoBasement -1.075e+04  2.473e+04  -0.434 0.664014    
## BsmtFinType1BLQ         2.080e+03  2.987e+03   0.697 0.486233    
## BsmtFinType1GLQ         4.592e+03  2.685e+03   1.710 0.087485 .  
## BsmtFinType1LwQ        -4.369e+03  3.987e+03  -1.096 0.273395    
## BsmtFinType1Rec        -5.587e+02  3.205e+03  -0.174 0.861642    
## BsmtFinType1Unf         2.205e+02  3.102e+03   0.071 0.943346    
## BsmtFinType1NoBasement         NA         NA      NA       NA    
## BsmtFinSF1              3.422e+01  5.652e+00   6.056 1.86e-09 ***
## BsmtFinType2BLQ        -1.019e+04  8.038e+03  -1.268 0.204937    
## BsmtFinType2GLQ        -1.044e+03  9.903e+03  -0.105 0.916047    
## BsmtFinType2LwQ        -1.231e+04  7.834e+03  -1.571 0.116498    
## BsmtFinType2Rec        -8.620e+03  7.481e+03  -1.152 0.249429    
## BsmtFinType2Unf        -5.625e+03  8.055e+03  -0.698 0.485077    
## BsmtFinType2NoBasement -2.481e+04  2.681e+04  -0.925 0.354955    
## BsmtFinSF2              3.135e+01  9.591e+00   3.269 0.001111 ** 
## BsmtUnfSF               2.134e+01  5.190e+00   4.113 4.17e-05 ***
## TotalBsmtSF                    NA         NA      NA       NA    
## HeatingGasA             1.989e+04  2.717e+04   0.732 0.464302    
## HeatingGasW             1.924e+04  2.805e+04   0.686 0.492842    
## HeatingGrav             1.388e+04  2.988e+04   0.464 0.642425    
## HeatingOthW            -4.709e+02  3.350e+04  -0.014 0.988787    
## HeatingWall             3.528e+04  3.158e+04   1.117 0.264119    
## HeatingQCFa            -7.270e+02  4.916e+03  -0.148 0.882462    
## HeatingQCGd            -3.891e+03  2.197e+03  -1.771 0.076781 .  
## HeatingQCPo            -3.117e+04  2.708e+04  -1.151 0.249865    
## HeatingQCTA            -2.554e+03  2.200e+03  -1.161 0.245912    
## CentralAirY            -2.359e+02  4.075e+03  -0.058 0.953842    
## ElectricalFuseF         9.683e+02  6.154e+03   0.157 0.875001    
## ElectricalFuseP        -1.078e+04  1.979e+04  -0.545 0.586187    
## ElectricalMix          -2.306e+04  4.755e+04  -0.485 0.627767    
## ElectricalSBrkr        -1.255e+03  3.151e+03  -0.398 0.690534    
## X1stFlrSF               3.727e+01  5.969e+00   6.245 5.85e-10 ***
## X2ndFlrSF               5.656e+01  6.053e+00   9.344  < 2e-16 ***
## LowQualFinSF           -7.456e+00  2.004e+01  -0.372 0.709882    
## GrLivArea                      NA         NA      NA       NA    
## BsmtFullBath            2.292e+03  2.114e+03   1.084 0.278610    
## BsmtHalfBath           -1.955e+02  3.209e+03  -0.061 0.951420    
## FullBath                4.630e+03  2.336e+03   1.982 0.047677 *  
## HalfBath                2.619e+03  2.210e+03   1.185 0.236132    
## BedroomAbvGr           -3.880e+03  1.449e+03  -2.679 0.007494 ** 
## KitchenAbvGr           -1.635e+04  5.992e+03  -2.729 0.006449 ** 
## KitchenQualFa          -2.287e+04  6.592e+03  -3.469 0.000540 ***
## KitchenQualGd          -2.408e+04  3.726e+03  -6.463 1.48e-10 ***
## KitchenQualTA          -2.438e+04  4.201e+03  -5.803 8.27e-09 ***
## TotRmsAbvGrd            3.179e+03  1.012e+03   3.143 0.001715 ** 
## FunctionalMaj2         -1.145e+03  1.540e+04  -0.074 0.940726    
## FunctionalMin1          5.170e+03  9.179e+03   0.563 0.573380    
## FunctionalMin2          7.031e+03  9.221e+03   0.762 0.445918    
## FunctionalMod          -2.091e+03  1.123e+04  -0.186 0.852403    
## FunctionalSev          -2.798e+04  3.130e+04  -0.894 0.371615    
## FunctionalTyp           1.635e+04  7.951e+03   2.056 0.039964 *  
## Fireplaces              7.382e+03  2.724e+03   2.710 0.006823 ** 
## FireplaceQuFa          -3.754e+03  7.372e+03  -0.509 0.610742    
## FireplaceQuGd          -2.592e+03  5.687e+03  -0.456 0.648655    
## FireplaceQuPo           6.659e+03  8.470e+03   0.786 0.431881    
## FireplaceQuTA          -9.439e+02  5.913e+03  -0.160 0.873200    
## FireplaceQuNoFireplace  4.892e+03  6.652e+03   0.735 0.462224    
## GarageTypeAttchd        1.945e+04  1.179e+04   1.650 0.099139 .  
## GarageTypeBasment       2.512e+04  1.359e+04   1.847 0.064921 .  
## GarageTypeBuiltIn       1.769e+04  1.230e+04   1.439 0.150392    
## GarageTypeCarPort       2.319e+04  1.540e+04   1.506 0.132415    
## GarageTypeDetchd        2.088e+04  1.180e+04   1.770 0.077002 .  
## GarageTypeNoGarage      1.750e+04  2.226e+04   0.786 0.432090    
## GarageYrBlt            -2.587e+01  6.236e+01  -0.415 0.678326    
## GarageFinishRFn        -3.219e+03  2.093e+03  -1.538 0.124193    
## GarageFinishUnf        -1.960e+03  2.602e+03  -0.753 0.451414    
## GarageFinishNoGarage           NA         NA      NA       NA    
## GarageCars              6.096e+03  2.418e+03   2.522 0.011811 *  
## GarageArea              1.086e+01  8.368e+00   1.298 0.194436    
## GarageQualFa           -1.024e+05  3.204e+04  -3.197 0.001422 ** 
## GarageQualGd           -9.448e+04  3.281e+04  -2.880 0.004048 ** 
## GarageQualPo           -1.133e+05  4.061e+04  -2.790 0.005350 ** 
## GarageQualTA           -9.661e+04  3.176e+04  -3.042 0.002397 ** 
## GarageQualNoGarage             NA         NA      NA       NA
## GarageCondFa            8.686e+04  3.708e+04   2.342 0.019317 *
## GarageCondGd            8.318e+04  3.847e+04   2.162 0.030805 *
## GarageCondPo            8.602e+04  3.951e+04   2.177 0.029650 *
## GarageCondTA            8.812e+04  3.674e+04   2.398 0.016611 *
## GarageCondNoGarage             NA         NA      NA       NA
## PavedDriveP            -2.079e+03  5.931e+03  -0.350 0.726034
## PavedDriveY             1.049e+03  3.661e+03   0.286 0.774625
## WoodDeckSF              1.661e+01  6.255e+00   2.655 0.008034 **
## OpenPorchSF            -5.774e+00  1.217e+01  -0.474 0.635375
## EnclosedPorch           6.928e-01  1.332e+01   0.052 0.958532
## X3SsnPorch              3.720e+01  2.356e+01   1.579 0.114670
## ScreenPorch             4.379e+01  1.333e+01   3.286 0.001045 **
## PoolArea                7.612e+02  2.425e+02   3.139 0.001736 **
## PoolQCFa               -1.737e+05  4.365e+04  -3.979 7.32e-05 ***
## PoolQCGd               -1.442e+05  3.949e+04  -3.651 0.000272 ***
## PoolQCNoPool            2.897e+05  1.312e+05   2.209 0.027377 *
## FenceGdWo               6.732e+03  5.247e+03   1.283 0.199691
## FenceMnPrv              7.682e+03  4.271e+03   1.798 0.072348 .
## FenceMnWw               2.472e+03  8.777e+03   0.282 0.778288
## FenceNoFence            7.724e+03  3.887e+03   1.987 0.047145 *
## MiscFeatureOthr         4.996e+04  5.153e+04   0.970 0.332455
## MiscFeatureShed         3.890e+04  5.027e+04   0.774 0.439136
## MiscFeatureTenC         6.989e+04  6.568e+04   1.064 0.287446
## MiscFeatureNone         3.788e+04  5.277e+04   0.718 0.472956
## MiscVal                 2.249e+00  4.074e+00   0.552 0.581027
## MoSold                 -3.939e+02  2.609e+02  -1.510 0.131371
## YrSold                 -5.608e+02  5.502e+02  -1.019 0.308283
## SaleTypeCon             2.530e+04  1.885e+04   1.342 0.179854
## SaleTypeConLD           1.570e+04  1.029e+04   1.526 0.127290
## SaleTypeConLI           4.410e+03  1.226e+04   0.360 0.719223
## SaleTypeConLw          -1.304e+03  1.295e+04  -0.101 0.919817
## SaleTypeCWD             1.094e+04  1.378e+04   0.794 0.427525
## SaleTypeNew             2.690e+04  1.647e+04   1.633 0.102645
## SaleTypeOth             1.215e+04  1.550e+04   0.784 0.433241
## SaleTypeWD             -1.408e+03  4.471e+03  -0.315 0.752938
## SaleConditionAdjLand    1.447e+04  1.535e+04   0.943 0.346037
## SaleConditionAlloca     4.491e+03  9.422e+03   0.477 0.633692
## SaleConditionFamily     1.324e+03  6.520e+03   0.203 0.839116
## SaleConditionNormal     8.397e+03  3.092e+03   2.716 0.006698 **
## SaleConditionPartial   -8.254e+03  1.587e+04  -0.520 0.603089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24310 on 1228 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9063
## F-statistic: 62.11 on 231 and 1228 DF,  p-value: < 2.2e-16
```


Residuals
1e+05
0e+00
−1e+05
−2e+05
−3e+05
0e+00
2e+05
4e+05
6e+05
Fitted values
count
600
400
200
0
−3e+05
−2e+05
−1e+05
0e+00
1e+05
Residuals
sample
0
−10
−2
0
2
theoretical

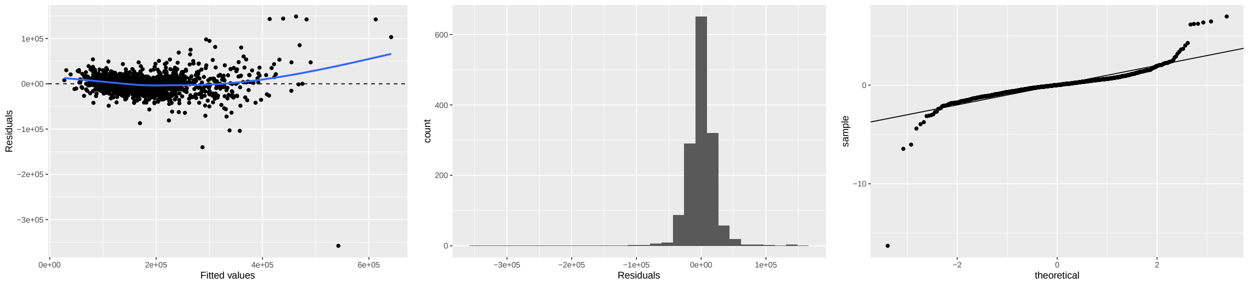
Model Two Model two filters out non numeric values.

```
##
## Call:
## lm(formula = SalePrice ~ ., data = numeric_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -465729  -16851   -2230   13535  305959
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.677e+05  1.413e+06   0.402 0.687804
## MSSubClass    -1.853e+02  2.775e+01  -6.676 3.50e-11 ***
## LotFrontage   -8.896e+01  4.711e+01  -1.889 0.059158 .
## LotArea        4.313e-01  1.018e-01   4.238 2.40e-05 ***
## OverallQual    1.738e+04  1.187e+03  14.643  < 2e-16 ***
## OverallCond    4.433e+03  1.025e+03   4.325 1.63e-05 ***
## YearBuilt      3.577e+02  7.273e+01   4.918 9.74e-07 ***
## YearRemodAdd   1.775e+02  6.745e+01   2.632 0.008578 **
## MasVnrArea     3.217e+01  5.945e+00   5.411 7.33e-08 ***
## BsmtFinSF1     1.877e+01  4.668e+00   4.021 6.10e-05 ***
## BsmtFinSF2     7.714e+00  7.055e+00   1.093 0.274373
## BsmtUnfSF      9.033e+00  4.195e+00   2.153 0.031463 *
## TotalBsmtSF           NA         NA      NA       NA
## X1stFlrSF      4.880e+01  5.809e+00   8.402  < 2e-16 ***
## X2ndFlrSF      4.865e+01  4.974e+00   9.781  < 2e-16 ***
## LowQualFinSF   3.321e+01  1.980e+01   1.677 0.093724 .
## GrLivArea             NA         NA      NA       NA
## BsmtFullBath   9.280e+03  2.611e+03   3.554 0.000391 ***
## BsmtHalfBath   1.563e+03  4.088e+03   0.382 0.702257
## FullBath       4.209e+03  2.818e+03   1.493 0.135544
## HalfBath      -1.788e+03  2.661e+03  -0.672 0.501664
## BedroomAbvGr  -1.019e+04  1.700e+03  -5.996 2.55e-09 ***
## KitchenAbvGr  -1.216e+04  5.207e+03  -2.336 0.019639 *
## TotRmsAbvGrd   5.176e+03  1.236e+03   4.189 2.98e-05 ***
## Fireplaces     3.402e+03  1.774e+03   1.917 0.055378 .
## GarageYrBlt   -6.544e+01  7.645e+01  -0.856 0.392153
## GarageCars     1.039e+04  2.855e+03   3.639 0.000283 ***
## GarageArea     2.866e+00  1.015e+01   0.282 0.777767
## WoodDeckSF     2.548e+01  7.994e+00   3.187 0.001467 **
## OpenPorchSF   -1.366e+00  1.514e+01  -0.090 0.928135
## EnclosedPorch  1.251e+01  1.687e+01   0.742 0.458265
## X3SsnPorch     2.094e+01  3.137e+01   0.668 0.504516
## ScreenPorch    5.492e+01  1.718e+01   3.197 0.001417 **
## PoolArea      -3.010e+01  2.369e+01  -1.271 0.204061
## MiscVal       -8.460e-01  1.856e+00  -0.456 0.648521
## MoSold        -6.981e+01  3.447e+02  -0.203 0.839516
## YrSold        -7.724e+02  7.020e+02  -1.100 0.271368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34730 on 1425 degrees of freedom
## Multiple R-squared:  0.8133, Adjusted R-squared:  0.8089
## F-statistic: 182.6 on 34 and 1425 DF,  p-value: < 2.2e-16
```



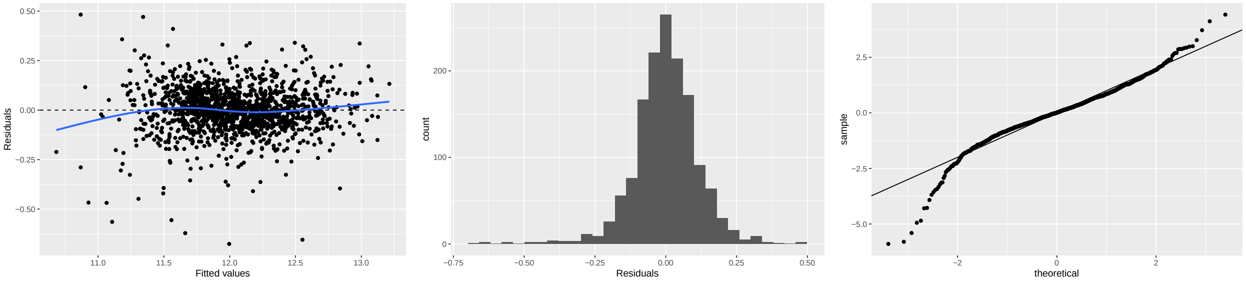
Model Three Model three utilizes model one, but uses backward stepwise elimination.

```
## 
## Call:
## lm(formula = SalePrice ~ MSSubClass + MSZoning + LotArea + Street + 
##     LandContour + LotConfig + LandSlope + Neighborhood + BldgType + 
##     OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofMatl + 
##     Exterior1st + MasVnrArea + ExterQual + BsmtQual + BsmtExposure + 
##     BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF + X1stFlrSF + X2ndFlrSF + 
##     BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + 
##     TotRmsAbvGrd + Functional + Fireplaces + GarageCars + GarageQual + 
##     GarageCond + WoodDeckSF + ScreenPorch + PoolArea + PoolQC + 
##     SaleCondition, data = train_set)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -357858  -10185     128   10137  148581 
## 
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -1.603e+06  1.731e+05  -9.263  < 2e-16 ***
## MSSubClass             -1.043e+02  4.831e+01  -2.159 0.031061 *  
## MSZoningFV              3.371e+04  1.175e+04   2.870 0.004172 ** 
## MSZoningRH              2.212e+04  1.181e+04   1.873 0.061342 .  
## MSZoningRL              2.488e+04  9.971e+03   2.495 0.012717 *  
## MSZoningRM              2.015e+04  9.307e+03   2.165 0.030536 *  
## LotArea                 5.346e-01  9.722e-02   5.499 4.58e-08 ***
## StreetPave              2.946e+04  1.174e+04   2.509 0.012236 *  
## LandContourHLS          1.338e+04  5.112e+03   2.617 0.008960 ** 
## LandContourLow         -1.990e+03  6.245e+03  -0.319 0.750022    
## LandContourLvl          1.035e+04  3.653e+03   2.833 0.004683 ** 
## LotConfigCulDSac        6.480e+03  3.140e+03   2.063 0.039263 *  
## LotConfigFR2           -7.664e+03  4.031e+03  -1.901 0.057498 .  
## LotConfigFR3           -1.672e+04  1.263e+04  -1.323 0.185910    
## LotConfigInside        -1.494e+03  1.748e+03  -0.854 0.393022    
## LandSlopeMod            7.260e+03  3.914e+03   1.855 0.063837 .  
## LandSlopeSev           -1.897e+04  1.021e+04  -1.859 0.063231 .  
## NeighborhoodBlueste     7.820e+02  1.910e+04   0.041 0.967355    
## NeighborhoodBrDale      5.699e+03  1.088e+04   0.524 0.600377    
## NeighborhoodBrkSide     4.550e+02  9.114e+03   0.050 0.960194    
## NeighborhoodClearCr    -7.123e+03  9.158e+03  -0.778 0.436826    
## NeighborhoodCollgCr    -2.714e+03  7.252e+03  -0.374 0.708274    
## NeighborhoodCrawfor     1.753e+04  8.541e+03   2.052 0.040379 *  
## NeighborhoodEdwards    -1.710e+04  7.999e+03  -2.137 0.032774 *  
## NeighborhoodGilbert    -8.026e+03  7.646e+03  -1.050 0.294040    
## NeighborhoodIDOTRR     -2.810e+03  1.048e+04  -0.268 0.788699    
## NeighborhoodMeadowV     5.024e+03  1.099e+04   0.457 0.647656    
## NeighborhoodMitchel    -1.637e+04  8.176e+03  -2.003 0.045395 *  
## NeighborhoodNAmes      -1.319e+04  7.803e+03  -1.691 0.091096 .  
## NeighborhoodNoRidge     4.065e+04  8.411e+03   4.833 1.50e-06 ***
## NeighborhoodNPkVill     7.627e+03  1.100e+04   0.693 0.488135    
## NeighborhoodNridgHt     2.498e+04  7.312e+03   3.416 0.000654 ***
## NeighborhoodNWAmes     -1.437e+04  7.968e+03  -1.803 0.071554 .  
## NeighborhoodOldTown    -9.136e+03  9.452e+03  -0.967 0.333936    
## NeighborhoodSawyer     -1.040e+04  8.152e+03  -1.275 0.202372    
## NeighborhoodSawyerW    -1.806e+03  7.813e+03  -0.231 0.817201    
## NeighborhoodSomerst     7.353e+03  8.708e+03   0.844 0.398601    
## NeighborhoodStoneBr     4.627e+04  8.250e+03   5.608 2.48e-08 ***
## NeighborhoodSWISU      -1.497e+03  9.561e+03  -0.157 0.875630    
## NeighborhoodTimber     -8.252e+03  8.210e+03  -1.005 0.315001    
## NeighborhoodVeenker     2.717e+03  1.043e+04   0.261 0.794464    
## BldgType2fmCon          6.712e+03  8.703e+03   0.771 0.440729    
## BldgTypeDuplex         -4.937e+03  6.516e+03  -0.758 0.448726    
## BldgTypeTwnhs          -1.622e+04  7.307e+03  -2.220 0.026618 *  
## BldgTypeTwnhsE         -1.056e+04  5.885e+03  -1.794 0.072981 .  
## OverallQual             6.165e+03  9.821e+02   6.277 4.65e-10 ***
## OverallCond             5.540e+03  7.980e+02   6.943 5.97e-12 ***
## YearBuilt               3.493e+02  6.299e+01   5.546 3.52e-08 ***
## YearRemodAdd            1.096e+02  5.330e+01   2.056 0.040005 *  
## RoofMatlCompShg         5.070e+05  4.574e+04  11.084  < 2e-16 ***
## RoofMatlMembran         5.654e+05  5.438e+04  10.396  < 2e-16 ***
## RoofMatlMetal           5.376e+05  5.378e+04   9.996  < 2e-16 ***
## RoofMatlRoll            5.002e+05  5.264e+04   9.502  < 2e-16 ***
## RoofMatlTar&Grv         4.943e+05  4.713e+04  10.489  < 2e-16 ***
## RoofMatlWdShake         5.212e+05  4.760e+04  10.950  < 2e-16 ***
## RoofMatlWdShngl         5.645e+05  4.657e+04  12.121  < 2e-16 ***
## Exterior1stAsphShn     -3.688e+03  2.568e+04  -0.144 0.885845    
## Exterior1stBrkComm      4.359e+03  1.977e+04   0.221 0.825500    
## Exterior1stBrkFace      1.515e+04  7.050e+03   2.150 0.031756 *  
## Exterior1stCBlock      -2.555e+04  2.672e+04  -0.956 0.339150    
## Exterior1stCemntBd     -7.663e+03  7.402e+03  -1.035 0.300712    
## Exterior1stHdBoard     -5.535e+03  6.393e+03  -0.866 0.386768    
## Exterior1stImStucc     -2.106e+04  2.553e+04  -0.825 0.409578    
## Exterior1stMetalSd     -1.187e+03  6.235e+03  -0.190 0.849085    
## Exterior1stPlywood     -6.099e+03  6.756e+03  -0.903 0.366881    
## Exterior1stStone       -1.011e+04  1.999e+04  -0.506 0.613042    
## Exterior1stStucco       1.730e+02  7.925e+03   0.022 0.982582    
## Exterior1stVinylSd     -1.515e+03  6.287e+03  -0.241 0.809591    
## Exterior1stWd Sdng     -2.824e+03  6.210e+03  -0.455 0.649424    
## Exterior1stWdShing     -6.679e+03  7.755e+03  -0.861 0.389215    
## MasVnrArea              1.130e+01  4.709e+00   2.401 0.016502 *  
## ExterQualFa            -7.873e+03  1.011e+04  -0.779 0.436280    
## ExterQualGd            -2.013e+04  4.812e+03  -4.183 3.06e-05 ***
## ExterQualTA            -2.124e+04  5.311e+03  -3.998 6.73e-05 ***
## BsmtQualFa             -1.669e+04  6.185e+03  -2.699 0.007043 ** 
## BsmtQualGd             -2.285e+04  3.338e+03  -6.845 1.17e-11 ***
## BsmtQualTA             -2.045e+04  4.059e+03  -5.038 5.35e-07 ***
## BsmtQualNoBasement      5.149e+03  2.544e+04   0.202 0.839629    
## BsmtExposureGd          1.743e+04  3.050e+03   5.713 1.36e-08 ***
## BsmtExposureMn         -2.318e+03  3.044e+03  -0.761 0.446516    
## BsmtExposureNo         -5.681e+03  2.163e+03  -2.627 0.008708 ** 
## BsmtExposureNoBasement -1.629e+04  2.433e+04  -0.670 0.503257    
## BsmtFinSF1              3.337e+01  4.714e+00   7.080 2.32e-12 ***
## BsmtFinSF2              2.384e+01  5.914e+00   4.031 5.86e-05 ***
## BsmtUnfSF               1.916e+01  4.466e+00   4.291 1.91e-05 ***
## X1stFlrSF               4.161e+01  5.136e+00   8.101 1.22e-15 ***
## X2ndFlrSF               4.942e+01  3.614e+00  13.673  < 2e-16 ***
## BsmtFullBath            3.383e+03  1.831e+03   1.847 0.064938 .  
## FullBath                3.399e+03  2.009e+03   1.692 0.090852 .  
## BedroomAbvGr           -4.664e+03  1.329e+03  -3.510 0.000463 ***
## KitchenAbvGr           -1.713e+04  5.478e+03  -3.127 0.001807 ** 
## KitchenQualFa          -2.115e+04  6.042e+03  -3.501 0.000480 ***
## KitchenQualGd          -2.351e+04  3.537e+03  -6.647 4.36e-11 ***
## KitchenQualTA          -2.374e+04  3.983e+03  -5.962 3.18e-09 ***
## TotRmsAbvGrd            3.157e+03  9.390e+02   3.362 0.000796 ***
## FunctionalMaj2         -3.915e+03  1.362e+04  -0.288 0.773772    
## FunctionalMin1          3.216e+03  8.382e+03   0.384 0.701278    
## FunctionalMin2          6.635e+03  8.226e+03   0.807 0.420017    
## FunctionalMod           1.696e+03  9.784e+03   0.173 0.862404    
## FunctionalSev          -3.400e+04  2.737e+04  -1.243 0.214250    
## FunctionalTyp           1.566e+04  7.169e+03   2.185 0.029073 *  
## Fireplaces              3.353e+03  1.332e+03   2.518 0.011927 *  
## GarageCars              8.191e+03  1.556e+03   5.264 1.64e-07 ***
## GarageQualFa           -9.938e+04  2.836e+04  -3.505 0.000472 ***
## GarageQualGd           -8.828e+04  2.905e+04  -3.039 0.002420 ** 
## GarageQualPo           -9.469e+04  3.456e+04  -2.740 0.006225 ** 
## GarageQualTA           -9.657e+04  2.814e+04  -3.432 0.000617 ***
## GarageQualNoGarage     -8.000e+03  1.776e+04  -0.450 0.652525    
## GarageCondFa            7.665e+04  3.345e+04   2.291 0.022093 *  
## GarageCondGd            6.860e+04  3.438e+04   1.996 0.046191 *  
## GarageCondPo            7.444e+04  3.569e+04   2.086 0.037167 *  
## GarageCondTA            8.076e+04  3.309e+04   2.440 0.014808 *  
## GarageCondNoGarage             NA         NA      NA       NA    
## WoodDeckSF              1.144e+01  5.882e+00   1.945 0.051945 .  
## ScreenPorch             3.952e+01  1.260e+01   3.137 0.001746 ** 
## PoolArea                5.907e+02  1.751e+02   3.374 0.000762 ***
## PoolQCFa               -1.535e+05  2.715e+04  -5.653 1.92e-08 ***
## PoolQCGd               -1.290e+05  3.283e+04  -3.929 8.97e-05 ***
## PoolQCNoPool            1.953e+05  9.521e+04   2.051 0.040488 *  
## SaleConditionAdjLand    1.459e+04  1.341e+04   1.088 0.276818    
## SaleConditionAlloca     2.384e+03  8.605e+03   0.277 0.781791    
## SaleConditionFamily     7.813e+02  6.159e+03   0.127 0.899079    
## SaleConditionNormal     7.439e+03  2.743e+03   2.712 0.006775 ** 
## SaleConditionPartial    1.765e+04  3.852e+03   4.581 5.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 24140 on 1337 degrees of freedom
## Multiple R-squared:  0.9154, Adjusted R-squared:  0.9077 
## F-statistic: 118.6 on 122 and 1337 DF,  p-value: < 2.2e-16
```



Model Four Model four utilizes forward selection for the most part with recommendations from the maker of the data set where he suggested to include the neighborhood in the model. Bathrooms were combined together as `TotalBath` and the age of the home is denoted by `Age`. `SaleCondition` was regrouped as normal and other. `NewHome` is whether or not the house sold is new.

```
##
## Call:
## lm(formula = log(SalePrice) ~ GrLivArea + TotalBsmtSF + OverallQual +
##     Neighborhood + NewHome + Age + CentralAir + Fireplaces +
##     GarageArea + TotalBath + PorchSqFt + PoolArea + SaleCondition +
##     MSZoning + BldgType + OverallCond, data = transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67570 -0.05984  0.00113  0.06766  0.48222
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.047e+01  6.565e-02 159.447  < 2e-16 ***
## GrLivArea            2.285e-04  1.018e-05  22.439  < 2e-16 ***
## TotalBsmtSF          1.445e-04  1.007e-05  14.355  < 2e-16 ***
## OverallQual          5.924e-02  4.140e-03  14.308  < 2e-16 ***
## NeighborhoodBlueste -6.831e-02  8.949e-02  -0.763 0.445364
## NeighborhoodBrDale  -7.007e-02  4.820e-02  -1.454 0.146299
## NeighborhoodBrkSide  2.420e-02  4.112e-02   0.589 0.556266
## NeighborhoodClearCr  9.036e-02  3.988e-02   2.266 0.023615 *
## NeighborhoodCollgCr  1.179e-02  3.333e-02   0.354 0.723639
## NeighborhoodCrawfor  1.463e-01  3.792e-02   3.858 0.000119 ***
## NeighborhoodEdwards -2.647e-02  3.602e-02  -0.735 0.462526
## NeighborhoodGilbert  6.295e-03  3.505e-02   0.180 0.857488
## NeighborhoodIDOTRR  -1.373e-02  4.752e-02  -0.289 0.772690
## NeighborhoodMeadowV -1.264e-01  4.591e-02  -2.754 0.005967 **
## NeighborhoodMitchel -1.610e-02  3.669e-02  -0.439 0.660749
## NeighborhoodNAmes   -4.839e-03  3.440e-02  -0.141 0.888164
## NeighborhoodNoRidge  5.908e-02  3.751e-02   1.575 0.115519
## NeighborhoodNPkVill -3.175e-02  4.915e-02  -0.646 0.518314
## NeighborhoodNridgHt  1.073e-01  3.319e-02   3.233 0.001252 **
## NeighborhoodNWAmes  -5.634e-02  3.545e-02  -1.590 0.112167
## NeighborhoodOldTown -3.500e-02  4.283e-02  -0.817 0.413936
## NeighborhoodSawyer  -1.770e-02  3.637e-02  -0.487 0.626532
## NeighborhoodSawyerW -8.178e-03  3.530e-02  -0.232 0.816835
## NeighborhoodSomerst  4.295e-02  4.067e-02   1.056 0.291163
## NeighborhoodStoneBr  1.493e-01  3.746e-02   3.984 7.12e-05 ***
## NeighborhoodSWISU   -1.877e-02  4.333e-02  -0.433 0.664904
## NeighborhoodTimber   4.682e-02  3.710e-02   1.262 0.207141
## NeighborhoodVeenker  9.483e-02  4.670e-02   2.031 0.042475 *
## NewHomeother        -1.175e-01  1.664e-02  -7.065 2.52e-12 ***
## Age                 -2.696e-03  2.522e-04 -10.692  < 2e-16 ***
## CentralAirY          7.178e-02  1.494e-02   4.806 1.70e-06 ***
## Fireplaces           3.437e-02  6.004e-03   5.725 1.26e-08 ***
## GarageArea           1.942e-04  1.985e-05   9.785  < 2e-16 ***
## TotalBath            5.454e-02  5.959e-03   9.153  < 2e-16 ***
## PorchSqFt            1.271e-04  2.257e-05   5.631 2.16e-08 ***
## PoolArea             1.265e-04  8.758e-05   1.444 0.148977
## SaleConditionother  -6.331e-02  1.085e-02  -5.836 6.63e-09 ***
## MSZoningFV           3.253e-01  5.413e-02   6.010 2.36e-09 ***
## MSZoningRH           3.097e-01  5.445e-02   5.689 1.55e-08 ***
## MSZoningRL           3.164e-01  4.538e-02   6.972 4.78e-12 ***
## MSZoningRM           2.784e-01  4.249e-02   6.553 7.90e-11 ***
## BldgType2fmCon      -5.730e-03  2.253e-02  -0.254 0.799276
## BldgTypeDuplex      -6.799e-02  1.819e-02  -3.738 0.000193 ***
## BldgTypeTwnhs       -1.107e-01  2.467e-02  -4.485 7.87e-06 ***
## BldgTypeTwnhsE      -4.382e-02  1.624e-02  -2.698 0.007053 **
## OverallCond          5.052e-02  3.313e-03  15.247  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1159 on 1410 degrees of freedom
## Multiple R-squared:  0.917,  Adjusted R-squared:  0.9143
## F-statistic: 346.1 on 45 and 1410 DF,  p-value: < 2.2e-16
```



Model Selection

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

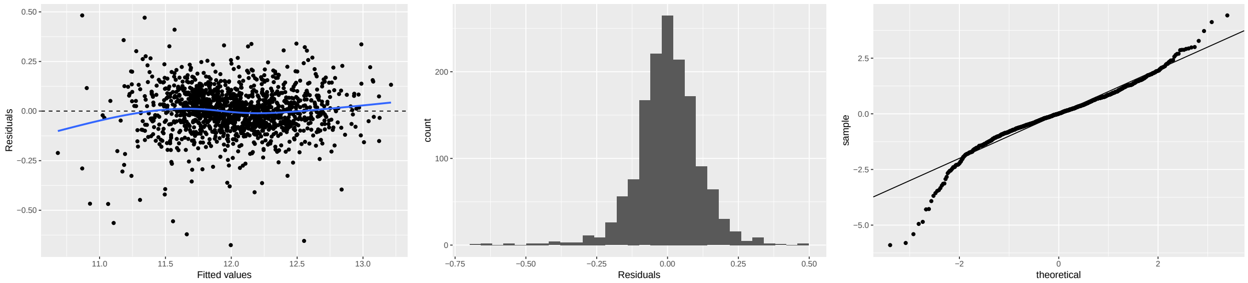


Table 9: Model Evaluation

```
##
##  iter imp variable
##   1   1  LotFrontage  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  BsmtFullBath  Bsmt
##   2   1  LotFrontage  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  BsmtFullBath  Bsmt
##   3   1  LotFrontage  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  BsmtFullBath  Bsmt
##   4   1  LotFrontage  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  BsmtFullBath  Bsmt
##   5   1  LotFrontage  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  BsmtFullBath  Bsmt
```

```
## Warning: Number of logged events: 81
```

```
##   MSSubClass MSZoning LotFrontage LotArea Street         Alley LotShape
## 1         20       RH          80   11622   Pave NoAlleyAccess      Reg
## 2         20       RL          81   14267   Pave NoAlleyAccess      IR1
## 3         60       RL          74   13830   Pave NoAlleyAccess      IR1
## 4         60       RL          78    9978   Pave NoAlleyAccess      IR1
## 5        120       RL          43    5005   Pave NoAlleyAccess      IR1
## 6         60       RL          75   10000   Pave NoAlleyAccess      IR1
##   LandContour Utilities LotConfig LandSlope Neighborhood BldgType HouseStyle
## 1         Lvl    AllPub    Inside       Gtl        NAmes     1Fam     1Story
## 2         Lvl    AllPub    Corner       Gtl        NAmes     1Fam     1Story
## 3         Lvl    AllPub    Inside       Gtl      Gilbert     1Fam     2Story
## 4         Lvl    AllPub    Inside       Gtl      Gilbert     1Fam     2Story
## 5         HLS    AllPub    Inside       Gtl      StoneBr   TwnhsE     1Story
## 6         Lvl    AllPub    Corner       Gtl      Gilbert     1Fam     2Story
##   OverallQual OverallCond YearBuilt YearRemodAdd RoofMatl Exterior1st
## 1           5           6      1961         1961  CompShg     VinylSd
## 2           6           6      1958         1958  CompShg     Wd Sdng
## 3           5           5      1997         1998  CompShg     VinylSd
## 4           6           6      1998         1998  CompShg     VinylSd
## 5           8           5      1992         1992  CompShg     HdBoard
## 6           6           5      1993         1994  CompShg     HdBoard
##   Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual
## 1     VinylSd       None          0        TA        TA     CBlock       TA
## 2     Wd Sdng    BrkFace        108        TA        TA     CBlock       TA
## 3     VinylSd       None          0        TA        TA      PConc       Gd
## 4     VinylSd    BrkFace         20        TA        TA      PConc       TA
## 5     HdBoard       None          0        Gd        TA      PConc       Gd
## 6     HdBoard       None          0        TA        TA      PConc       Gd
##   BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2
## 1       TA           No          Rec        468          LwQ        144
## 2       TA           No          ALQ        923          Unf          0
## 3       TA           No          GLQ        791          Unf          0
## 4       TA           No          GLQ        602          Unf          0
## 5       TA           No          ALQ        263          Unf          0
## 6       TA           No          Unf          0          Unf          0
##   BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical X1stFlrSF
## 1       270         882    GasA        TA          Y      SBrkr       896
## 2       406        1329    GasA        TA          Y      SBrkr      1329
## 3       137         928    GasA        Gd          Y      SBrkr       928
## 4       324         926    GasA        Ex          Y      SBrkr       926
## 5      1017        1280    GasA        Ex          Y      SBrkr      1280
## 6       763         763    GasA        Gd          Y      SBrkr       763
##   X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
## 1         0            0       896            0            0        1        0
## 2         0            0      1329            0            0        1        1
## 3       701            0      1629            0            0        2        1
## 4       678            0      1604            0            0        2        1
## 5         0            0      1280            0            0        2        0
## 6       892            0      1655            0            0        2        1
##   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces
## 1            2            1          TA            5        Typ          0
## 2            3            1          Gd            6        Typ          0
## 3            3            1          TA            6        Typ          1
## 4            3            1          Gd            7        Typ          1
## 5            2            1          Gd            5        Typ          0
## 6            3            1          TA            7        Typ          1
##   FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea
## 1 NoFireplace     Attchd        1961          Unf          1        730
## 2 NoFireplace     Attchd        1958          Unf          1        312
## 3          TA     Attchd        1997          Fin          2        482
## 4          Gd     Attchd        1998          Fin          2        470
## 5 NoFireplace     Attchd        1992          RFn          2        506
## 6          TA     Attchd        1993          Fin          2        440
##   GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch
## 1         TA         TA          Y        140           0             0
## 2         TA         TA          Y        393          36             0
## 3         TA         TA          Y        212          34             0
## 4         TA         TA          Y        360          36             0
## 5         TA         TA          Y          0          82             0
## 6         TA         TA          Y        157          84             0
##   X3SsnPorch ScreenPorch PoolArea PoolQC   Fence MiscFeature MiscVal MoSold
## 1          0         120        0 NoPool   MnPrv        None       0      6
## 2          0           0        0 NoPool NoFence        Gar2   12500      6
## 3          0           0        0 NoPool   MnPrv        None       0      3
## 4          0           0        0 NoPool NoFence        None       0      6
## 5          0         144        0 NoPool NoFence        None       0      1
## 6          0           0        0 NoPool NoFence        None       0      4
##   YrSold SaleType SaleCondition TotalBath Age PorchSqFt NewHome SalePrice
## 1   2010       WD        normal       1.0  49       260   other  127504.9
## 2   2010       WD        normal       1.5  52       429   other  154023.9
## 3   2010       WD        normal       2.5  13       246   other  172933.0
## 4   2010       WD        normal       2.5  12       396   other  195604.1
## 5   2010       WD        normal       2.0  18       226   other  206002.9
## 6   2010       WD        normal       2.5  17       241   other  176736.1
```

```
glimpse(prices)
```

```
sapply(prices, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>%
  kable_styling(latex_options = "hold_position")
```

```
#’ na_replace - NA Replace.
#’
#’ Given the Ames Housing dataset, converts genuine NA values that have
#’ meaning within the context of the data to more meaningful values, and
#’ returns the altered dataset to illiminate the mistaken interpretation
#’ of the term "NA" as a genuine missing value.
#’
#’ @param dataframe The Ames dataset as a dataframe.
#’
#’ @return The Ames dataset with genuine NA values imputed to human friendly values.
na_replace <- function(dataframe) {
  dataframe %>%
    mutate(Alley = fct_explicit_na(Alley, na_level = ’NoAlleyAccess’),
           BsmtQual = fct_explicit_na(BsmtQual, na_level = ’NoBasement’),
           BsmtCond = fct_explicit_na(BsmtCond, na_level = ’NoBasement’),
           BsmtExposure = fct_explicit_na(BsmtExposure, na_level = ’NoBasement’),
           BsmtFinType1 = fct_explicit_na(BsmtFinType1, na_level = ’NoBasement’),
           BsmtFinType2 = fct_explicit_na(BsmtFinType2, na_level = ’NoBasement’),
           FireplaceQu = fct_explicit_na(FireplaceQu, na_level = ’NoFireplace’),
           GarageType = fct_explicit_na(GarageType, na_level = ’NoGarage’),
           GarageFinish = fct_explicit_na(GarageFinish, na_level = ’NoGarage’),
           GarageQual = fct_explicit_na(GarageQual, na_level = ’NoGarage’),
           GarageCond = fct_explicit_na(GarageCond, na_level = ’NoGarage’),
           PoolQC = fct_explicit_na(PoolQC, na_level = ’NoPool’),
           Fence = fct_explicit_na(Fence, na_level = ’NoFence’),
           MiscFeature = fct_explicit_na(MiscFeature, na_level = ’None’)
    )
}



prices <- na_replace(prices)
# Check for empty values once again to see what affect this has on the data.
sapply(prices, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>%
  kable_styling(latex_options = "hold_position")
```

```
corr_data<- select_if(prices,is.numeric) %>%
  select(-Id) %>%
  as.matrix(.) %>%
  rcorr(.)
corr_p <- round(corr_data$P,4)
```

```r
# this takes the values and correlations and makes it into a 2 column dataframe
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
}


# sorted the pairs of correlations by their p value to show variables with the biggest
# relationships, with the p showing the significance value
flattenCorrMatrix(corr_data$r, corr_data$P) %>%
  arrange(desc(abs(cor))) %>%
  head(10) %>%
  kable(caption = ’Correlations of numeric predictors’) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

flattenCorrMatrix(corr_data$r, corr_data$P) %>%
  arrange(desc(abs(cor))) %>%
  filter(column == ’SalePrice’) %>%
  head(10) %>%
  kable(caption = ’Correlations of numeric predictors against the Sales Price’) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

```r


prices %>%
  mutate(TotalSqFt = GrLivArea + TotalBsmtSF) %>%
  ggplot(., aes(x = TotalSqFt, y = SalePrice)) +
  geom_point() +
  geom_smooth() +
  ggtitle("Total Square Footage vs Sales Price") +
  scale_y_continuous(labels = scales::label_comma())



prices %>%
  mutate(OverallQual= as.factor(OverallQual)) %>%
  ggplot(., aes(x = OverallQual, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Distributions of Overall Quality vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())

prices %>%
  mutate(OverallCond= as.factor(OverallCond)) %>%
  ggplot(., aes(x = OverallCond, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Distributions of Overall Condition vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())
```

```
prices %>%
  mutate(GarageCars = as.factor(GarageCars)) %>%
  ggplot(., aes(x = GarageCars, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Distributions of Amount of Cars in Garage vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())

prices %>%
  mutate(BldgType = recode(BldgType, ‘2fmCon‘ = "2Fam Conversion",
                           ‘Twnhs‘ = "Townhouse Inside",
                           ‘TwnhsE‘ = "Townhouse End Unit")) %>%
  ggplot(., aes(x = BldgType, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Type of Building vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())

prices %>%
  ggplot(., aes(x = CentralAir, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Central Air Coniditioning vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())

prices %>%
  mutate(Fireplaces = ifelse(Fireplaces == 0, "no", "yes")) %>%
  ggplot(., aes(x = Fireplaces, y = SalePrice)) +
  geom_boxplot() +
  labs(title = ’Fireplaces vs Sales Price’) +
  scale_y_continuous(labels = scales::label_comma())
```

```
drop = c(drop, "Id",’Condition1’, ’Condition2’, ’RoofStyle’)
dropped = prices[,!(names(prices) %in% drop)]
```

```
#’ mice_imputation- Mice Imputation.
#’
#’ Given the Ames Housing dataset, runs the MICE algorithm on the dataset
#’ to impute both numerical and categorical missing values.
#’
#’ @param dataframe The Ames dataset as a dataframe.
#’
#’ @return The Ames dataset with missing values imputed to complete values.
#’
mice_imputation <- function(dataframe) {
  imputation <- mice(dataframe, m = 1, method = ’cart’)
  imputed <- mice::complete(imputation)
}
imputed <- mice_imputation(dropped)
# Check for empty values once again to see what affect MICE had on our data.
sapply(imputed, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>% kable_styling()
```

```
train_set = imputed
model1 = lm(SalePrice~., train_set)
summary(model1)
```

```
ggplot(data = model1, aes(x = .fitted, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = ’dashed’) +
  geom_smooth(se = FALSE) + xlab(’Fitted values’) + ylab(’Residuals’)
ggplot(data = model1, aes(x = .resid)) + geom_histogram() + xlab(’Residuals’)
ggplot(data = model1) + stat_qq(aes(sample = .stdresid)) + geom_abline()
```

```
# Using features with only numeric values.
numeric_df <- train_set %>% dplyr::select(where(is.numeric))
model2 <- lm(SalePrice~., numeric_df)
summary(model2)
```

```
# Looking at residuals.
ggplot(data = model2, aes(x = .fitted, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = "dashed") +
  geom_smooth(se = FALSE) + xlab("Fitted values") + ylab("Residuals")
ggplot(data = model2, aes(x = .resid)) + geom_histogram() + xlab("Residuals")
ggplot(data = model2) + stat_qq(aes(sample = .stdresid)) + geom_abline()
```

```
model3 <- step(model1, direction = ’backward’, trace = 0)
summary(model3)
```

```
ggplot(data = model3, aes(x = .fitted, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = ’dashed’) +
  geom_smooth(se = FALSE) + xlab(’Fitted values’) + ylab(’Residuals’)
ggplot(data = model3, aes(x = .resid)) + geom_histogram() + xlab(’Residuals’)
ggplot(data = model3) + stat_qq(aes(sample = .stdresid)) + geom_abline()
```

```
transform = train_set %>%
  filter(GrLivArea < 4000) %>%
  mutate(TotalBath = BsmtFullBath + 0.5 * BsmtHalfBath + FullBath + 0.5 * HalfBath,
         Age = YrSold - YearBuilt,
         SaleCondition = ifelse(SaleCondition == "Normal", "normal", "other"),
         PorchSqFt = ScreenPorch + X3SsnPorch + EnclosedPorch + OpenPorchSF + WoodDeckSF,
         NewHome = ifelse(SaleType == ’New’, ’new’, ’other’))
model4 = lm(log(SalePrice) ~ GrLivArea + TotalBsmtSF + OverallQual + Neighborhood + NewHome +
     Age + CentralAir + Fireplaces + GarageArea + TotalBath  + PorchSqFt + PoolArea +
     SaleCondition + MSZoning + BldgType + OverallCond , data = transform)
summary(model4)
```

```
ggplot(data = model4, aes(x = .fitted, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = ’dashed’) +
  geom_smooth(se = FALSE) + xlab(’Fitted values’) + ylab(’Residuals’)
ggplot(data = model4, aes(x = .resid)) + geom_histogram() + xlab(’Residuals’)
ggplot(data = model4) + stat_qq(aes(sample = .stdresid)) + geom_abline()
```

```
# the model is suitable for linaer regression as residuals meet the following criteria
# residuals are clustered around 0
ggplot(data = model4, aes(x = .fitted, y = .resid)) +
  geom_point() + geom_hline(yintercept = 0, linetype = ’dashed’) +
  geom_smooth(se = FALSE) + xlab(’Fitted values’) + ylab(’Residuals’)
# residuals are normally distributed
ggplot(data = model4, aes(x = .resid)) + geom_histogram() + xlab(’Residuals’)
# qq plot also shows that residuals are almost normal
ggplot(data = model4) + stat_qq(aes(sample = .stdresid)) + geom_abline()
```

```
test = prices.test[,!(names(prices.test) %in% drop)]
test.impute = test%>%
  na_replace()%>%
  mice_imputation%>%
  filter(GrLivArea < 4000) %>%
  mutate(TotalBath = BsmtFullBath + 0.5 * BsmtHalfBath + FullBath + 0.5 * HalfBath,
         Age = YrSold - YearBuilt,
         SaleCondition = ifelse(SaleCondition == "Normal", "normal", "other"),
         PorchSqFt = ScreenPorch + X3SsnPorch + EnclosedPorch + OpenPorchSF + WoodDeckSF,
         NewHome = ifelse(SaleType == ’New’, ’new’, ’other’))
test.impute$SalePrice = exp(predict(model4, test.impute))
head(test.impute)
```

R statistical programming code