Determining Factors Influencing House Sale Prices

Kyle Gilde, Jaan Bernberg, Kai Lukowiak,
Michael Muller, Ilya Kats
DATA 621, Master of Science in Data Science,
City University of New York

Abstract

TBD: Since it summarizes the work, it will be written at the end. 250 words or less summarizing the problem, methodology, and major outcomes.

Key Words

house prices, regression, linear models, assessed value

Introduction

This project stems out of the Business Analytics and Data Mining class in the Master of Science in Data Science program at CUNY. This paper is the result of the final class group project in applying regression methods to real-world data. Our team chose housing data because it promissed to be an interesting and useful subject. In addition, this research is based on a well studied data set which makes it an excellent educational resource allowing our team to study various approaches.

The data set was prepared by Dean De Cock in an effort to create a real-world data set to test and practice regression methods (De Cock 2011). It describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. Ames, Iowa was founded in 1864 as a station stop. It has the population of about 60,000 people and covers about 24.27 sq mi. It was ranked ninth on the *Best Places to Live* list (CNNMoney 2010).

The data came directly from the Assessor's Office in the form of a data dump from their records system and it included information for calculation of assessed values in the city's assessment process. The data is recent and it covers the period of housing bubble collapse that led to the subprime mortgage crisis. 2008 saw one of the largest housing price drops in history.

Each of over 2,900 total observations in the data represent attributes of a residental property sold. For properties that exchanged ownership multiple times during the collection period (2006 through 2010), only the last sale is included in the data since it represents the most current value of the property. The attributes that make up the sale price of a house can seem daunting given a myriad of factors that can impact its value. There are about 80 variables included in the data set. Most variables describe physical attributes of the property. There is a variety of variable types - discrete, continous, categorical (both nominal and ordinal).

The data was originally published in the Journal of Statistics Education (Volume 19, Number 3). Data set was downloaded from Kaggle.com which gave us the ability to compare our results with results of other teams working with this data set (Kaggle 2016).

Literature Review

Building regression models to predict house prices is not a new undertaking. Quite the opposite, a lot of research went into this area. There is a clear financial benefit to buyers, sellers and other parties in knowing which attributes influence final sale price. There is also a lot of data readily available with some cleanup work. Data is kept by local governments to be used in the assessment process for property taxes. There is a lot of data captured by realors when a property is listed on the market. Additionally, in large part thanks to information revolution, data is easily accessible via many aggregators such as MLS.

There are many attributes that factor into a house price. For example, environmental attributes can impact the price substantially. A garden facing water, a pleasant view whether it overlooks water or open space, attractive landscaping all increase house prices (Luttik 2000). Neighborhood attributes such as schools and public services also play a factor.

Our data set deals mostly with physical characteristics of the house itself. Even here there is a lot of room for variation. For example, one study counted half-bathrooms as 0.1 out of belief that buyers do not value them as much as full bathrooms (Pardoe 2008).

. . .

- 1 page
- Discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are.
- Explain how your investigation is similar or different to the state-of-theart.

Methodology

Data Description

The data set includes 2,919 observation and 79 indepedent variables. Out of those 36 are numeric, such as lot area or pool area in square feet, and 43 are categorical, such as garage type (attached to home, built-in, carport, etc.) or utilities (gas, sewer, both, etc.).

Data Imputation

Original data set included no complete observations (*see table 1*). However, many `NA` values found in the data carry useable information. For example, `NA` in the `PoolQC` variable (pool quality) implies that the property has no pool. Often this logic carried across multiple variables - for example, `NA` in `GarageQual` (garage quality), `GarageCond` (garage condition) and `GarageType` variables all imply that the property has no garage. This type of missing values was replaced with a new category - *No Pool*, *No Garage* or similar. This work was accomplished using the `forcats` `R` package.

After this substitution the number of complete observations went up significantly to 2,861 or about 98% of all observations. There remained only 58 observations with true missing values (about 2% of the total observations). These observations contained 180 missing values in 32 variables. None of the variables contained a large number of missing values. The top one was `MasVnrType` with 24 observations containing `NA` (0.8% of all observations). None of the variables were close to the 5% missing threshold that would suggest that we should drop them from analysis.

Consider the pattern to the missing values. In addition to the quantity of missingness being important, why and how the values are missing can give us insight into whether we have a biased sample. There are three types of missing data (Faraway 2014): 1) Missing Completely at Random (MCAR), 2) Missing at Random (MAR), and 3) Missing Not at Random (MNAR). MCAR is when the probability of missingness is the same for all cases. This is the ideal type of missingness because we could delete these cases without incurring bias. MAR occurs when the probability of a value being missing depends upon a known mechanism. In this scenario, we could delete these observations and compensate by weighting by group membership. Finally, MNAR occurs when the values are missing because of an unknown variable. This is the type of missingness that is most likely to bias our sample. Faraway asserts that ascertaining the exact nature of the missingness is not possible and must be inferred. Figure 1 displays the combinations of missing values in the predictor variables. We may not have MCAR because we can see that the missingness is not more dispersed across all variables and cases. Only 32 of the 79 predictors have a missing value, and we notice that the missingness occurs most often in some of the masonry, basement and garage variables. There is no indication that values are missing not at random and given the small number of missing values, we believe the bias, if any, will be limited.

There are four ways to deal with missing values (Prabhakaran 2017):

- **Deleting the cases:** This is not a preferred method because one could introduce bias or the model could lose power from being based upon fewer cases.
- **Deleting the variables:** If the missingness is concentrated in a relatively small number of variables, then deleting the variables may be a good option. The downside to this approach is that we lose the opportunity to include the observed values in the model.
- **Imputation via mean, median and mode:** An expedient way to retain all of the cases and variables is to insert the mean or median for continuous variables or the mode for categorical or discrete variables. This approach may suffice for a small number of values, but has the potential to introduce bias in the form of decreasing the variance.
- **Prediction:** This more advanced approach involves using the other variables to predict the missing values.

For our data set we used multiple imputation by chained equations (MICE). The technique involves imputing multiple iterations of values in order to account for statistical uncertainty with standard errors (Azur 2012). Since it uses chained equations, MICE has the ability to impute both numerical and categorical variables. The ideal scenario to use MICE is when less than 5% of the values are missing and when values are missing at random. We used the `mice` `R` package with the `cart` (classification and regression trees) method. CART is one of the five `mice` methods that can impute both numerical and categorical variables. Figure 2 shows the density plots of the observed and imputed values. The imputed distributions have more variance and extremes than the observed distributions. If we were to run, multiple imputations, hopefully we would begin to see more convergence between the imputed and observed values.

Additional Data Preparation

All categorical variables were inspected and their order (or order of levels in R) was changed to match the most likely low-to-high order. These variables for the most part do not rely on the order of categories, so this step was not critical to modeling; however, it makes modeling output more readable and easier to interpret.

As is the case with most data sets, we found several values that were clearly typos and input errors. For instance, one observation had the year when garage was built listed as 2207. There were 6 negative values in age related variables (see data transformations below). Those were set to 0.

Data Transformation

Prior to modeling, we have extensively analyzed available variables and took a few approaches to variable transformations. They were ment to both simplify existing variables and add new variables that may be helpful in modeling.

Generally, it is more common to think about the age of the house than the year it was built. Each age related variable was stored in the data set in two related variables - year built and year sold. Rather than trying to work with original variables we have converted them to a single *age* variable. For house age the value was $YrSold - YearBuilt$. Similarly the age of garage and remodeling was added to the data set. Original variables were dropped from analysis.

Because we are not dealing with a time series data set, we have converted `YrSold` and `MoSold` variables from numeric to nominal. It is important to catch seasonality, but does not make sense to regress on these variables as continous variables.

Using the side-by-side box plots in Figure 3, we examined the categorical variables with more than two values to see if the variable can be simplified by combining the values into two groups. Our criteria for this simplification is if the variables' inner quartile ranges of the response variable distinctly and logically bifurcate. For example, in FireplaceQu (fireplace quality), HeatingQC (heating quality) and PoolQC (pool quality), we can notice that only the inner quartiles are bifurcated into two groups that do not overlap: the highest *Excellent* value and all other lesser quality conditions. Additional values that are distinct from other values in the same variables are the *Wood Shingle* value in the roof material variable (RoofMatl), the above average values in the garage quality variable (GarageQual), the gas-related values in the heating variable (Heating), and the *Partial* value in the sale condition variable (SaleCondition). Consequently, we transformed these into dummy variables with appropriate names. This allowed us to preserve some degrees of freedom that would otherwise be subtracted if each and every one of the original values were turned into dummy variables.

We examined whether our modeling would benefit from transforming any of the predictor variables. To do so, we have automated creation of several different versions of the predictor variables using R. We took natural logarithms, square roots and squares of the numerical variables, and then we calculated every possible pairwise interaction between these transformations, the original numerical variables and categorical variables. We then calculated all pairwise correlations between the interactions and the response variable SalePrice. The top correlations can be seen in table 4, which is sorted descendingly by R-squared. We observed that there are several correlation values higher than the highest correlation between the original predictor and the response, which is OverallQual at 0.79 (*see table 3*). Most promising transformations involved taking the square of OverallQual and multiplying it by the log-transformed or square-root-transformed one of the area variables. We added top five interactions to our training data set.

We have created several potential training sets to give use flexibility in training the model. The **first** of the three training data sets we created includes only the original variables with the missing values imputed. In model building and selection this set is referred to as the *original* data set. The **second** training data set includes seven "simplified" dummy variables instead of original variables. It also includes five highly-correlated interactions. This set is referred to as the *transformed* data set. The **third** training data set includes the same predictor variables as in the second set with a transformed response variable. While creating all interactions, we noticed that the correlation values appeared to increase vis-a-vis the square root of the response variable. Consequently, since the response variable contains only positive values, we created a simple BIC step model and used it to calculate the Box-Cox λ value and transform the response variable. According to Box-Cox, a λ value of approximately 0.184 should help the final model meet the normality assumption. This set is referred as the *Box-Cox* data set.

Modeling

Since we are dealing with trying to predict a continous variable, house sale price, we relied on building and optimizing general linear model.

After fitting three baseline (all k-parameters) models to all three training data sets, ANOVA demonstrated statistical significance between the original data set and the transformed data set. While all multiple R^2 values were within some neglible deviation of each other, adding a Box-Cox transformation of the response variable improved the R^2 beyond the model based on the original data set.

We took the strongest model, and applied stepwise regression. Since we started with the baseline model containing all variables we applied backward elimination in order to settle on a model with the lowest Akaike information criterion (AIC) value.

For non-transformed response variable, we expetimented with applying log-transformation as it tends to bring sales data closer to normal distribution.

We ended up with six representative models:

1. **Model 1** is based on the fully transformed data set with Box-Cox transformed response variable. It includes all available predictor variables including any interactions created in data preparation. This

model explains nearly 94% of variability of the response variable. A good starting point, but we can remove some insignificant variables for a more parsimonious model and lower chances of overfitting.

2. **Model 2** is based on Model 1 modified with stepwise regression (backward elimination). It is an improvement with lower number of parameters (156 comparing to 237). The multiple R^2 value is similar. Comparing two models using ANOVA indicates that they are not significantly different.
3. **Model 3** selects only statistically highly significant variables from the previous model (p-value is nearly 0). R^2 drops and F-statistic rises, so even though the model is simplier with only 58 parameters, it may not be an improvement. Comparing this model with the first one using ANOVA, shows that there is significant difference between the two.
4. **Model 4** expands on the previous model by using statistically significant variables, but with less strict criteria (p-value < 0.01). Number of parameters is increased, but R^2 is also increased. Similarly, per ANOVA, this model is significantly different from models 1 and 3.
5. **Model 5** takes variables identified in the previous model, but it is trained on the original data set without interactions. It uses only log-transformation of `LotArea` predictor variable and `SalePrice` response variable. This model represents the best results based on R^2 for any model we have tried using the original data set.
6. **Model 6** is based on Model 4, but it is trained on the transformed data set that includes interactions, but not the Box-Cox transformation of the response variable. Similarly to model 5, this model uses log-transformed `LotArea` and `SalePrice`.

For all models the F-statistic's p-value shows a drastic improvement over an intercept-only model, so we can infer that these models are statistically significant.

Experimentation and Results

- 4-5 pages
- Key figures and tables may be included here
- Additional figures and tables should be added to appendices
- Discuss data prepatation details not mentioned under Methodology
- Discuss model building and selection
- Discuss model validation
- Discuss results of statistical analysis
- Describe final model (coefficients, interpretation)
- Discuss upload of results to Kaggle

Discussion

- 1-2 pages
- Discuss limitations
- Discuss areas for future work
- Discuss detailed findings
- May be combined with Conclusion section below

Based on one town's data

Conclusion

- 1 paragraph
- Quick summary of findings

Appendix A. Figures

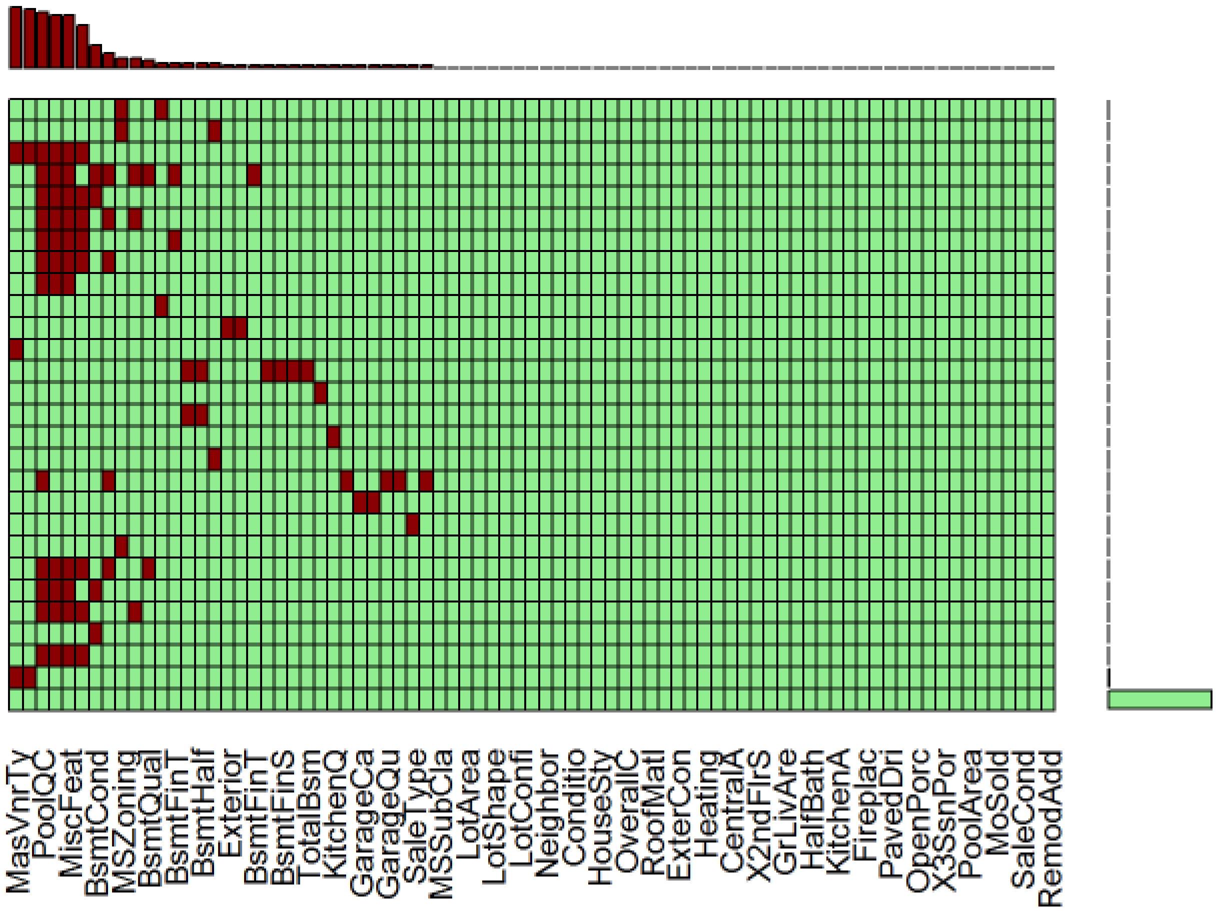

Figure 1. Missing values.

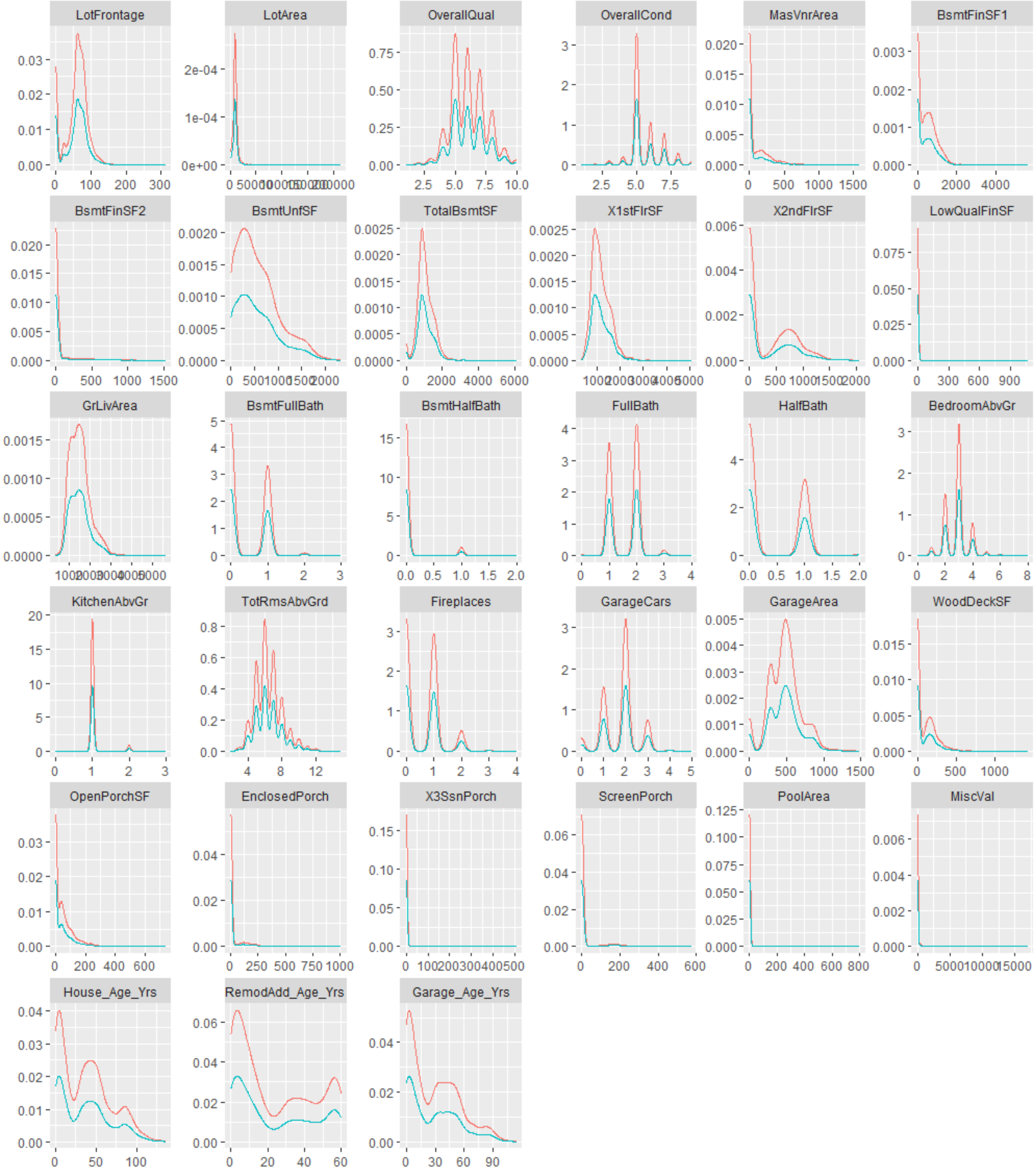

Figure 2. Density plots of observed (blue) and imputed (red) values.

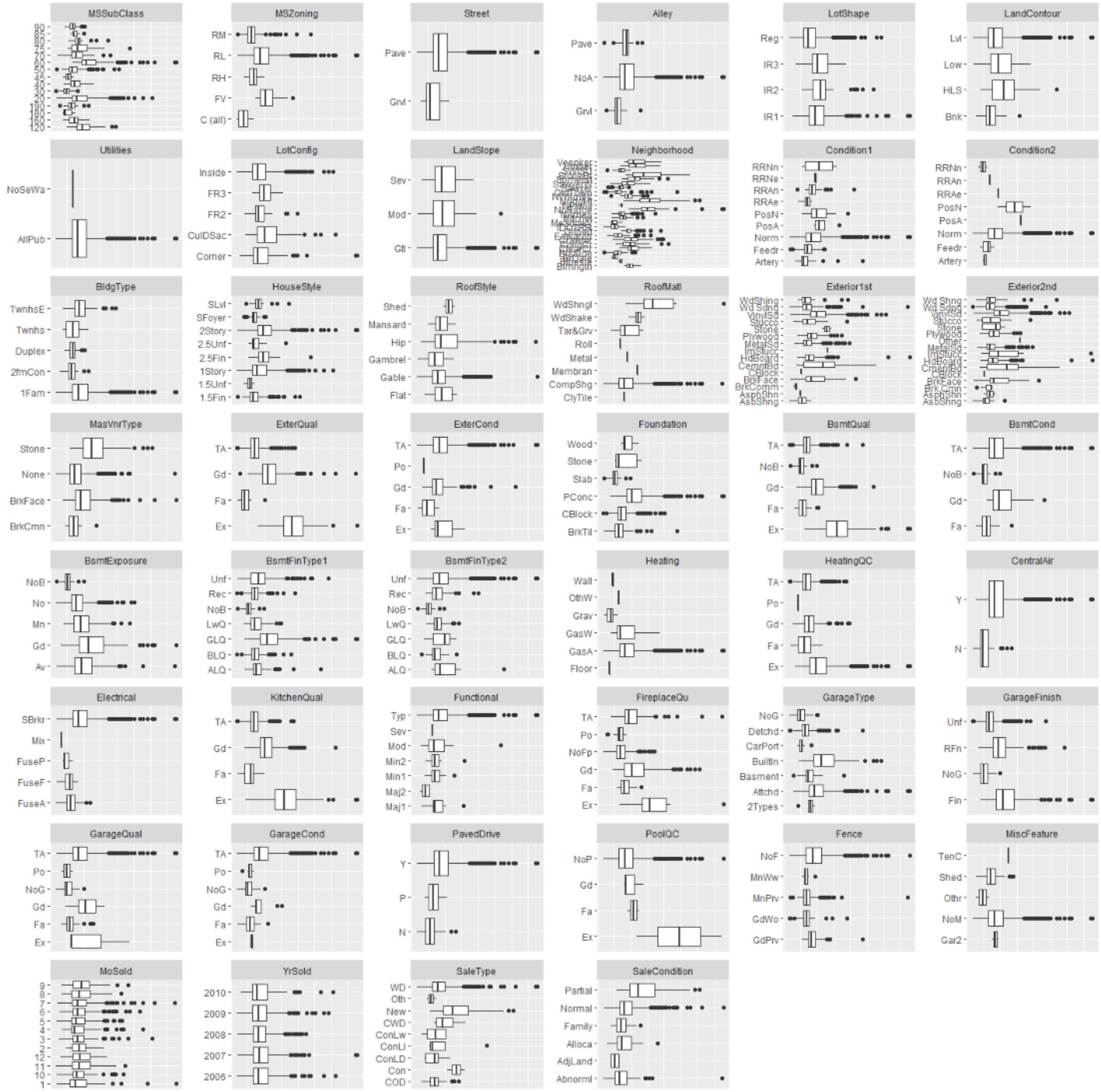

Figure 3. Box plots of categorical variables against the response variable.

Appendix B. Tables

Table 1: Number of NA values in original data.

Variable	No of NAs	Percent of Total Obs
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageYrBlt	159	5.45
GarageFinish	159	5.45
GarageQual	159	5.45
GarageCond	159	5.45
Garage_Age_Yrs	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
Functional	2	0.07
Exterior1st	1	0.03
Exterior2nd	1	0.03
BsmtFinSF1	1	0.03
BsmtFinSF2	1	0.03
BsmtUnfSF	1	0.03
TotalBsmtSF	1	0.03
Electrical	1	0.03
KitchenQual	1	0.03
GarageCars	1	0.03
GarageArea	1	0.03
SaleType	1	0.03

Table 2: Descriptive statistics for numerical variables.

Variable	Count	Mean	SD	Median	Min	Max	Kurtosis
LotFrontage	2919	57.77	33.48	63	0	313	2.169
LotArea	2919	10168	7887	9453	1300	215245	264.3
OverallQual	2919	6.089	1.41	6	1	10	0.06295
OverallCond	2919	5.565	1.113	5	1	9	1.472
YearBuilt	2919	1971	30.29	1973	1872	2010	-0.5142
YearRemodAdd	2919	1984	20.89	1993	1950	2010	-1.347
MasVnrArea	2896	102.2	179.3	0	0	1600	9.228
BsmtFinSF1	2918	441.4	455.6	368.5	0	5644	6.884
BsmtFinSF2	2918	49.58	169.2	0	0	1526	18.79
BsmtUnfSF	2918	560.8	439.5	467	0	2336	0.3985
TotalBsmtSF	2918	1052	440.8	989.5	0	6110	9.125
X1stFlrSF	2919	1160	392.4	1082	334	5095	6.936
X2ndFlrSF	2919	336.5	428.7	0	0	2065	-0.4254
LowQualFinSF	2919	4.694	46.4	0	0	1064	174.5
GrLivArea	2919	1501	506.1	1444	334	5642	4.108
BsmtFullBath	2917	0.4299	0.5247	0	0	3	-0.738
BsmtHalfBath	2917	0.06136	0.2457	0	0	2	14.81
FullBath	2919	1.568	0.553	2	0	4	-0.5409
HalfBath	2919	0.3803	0.5029	0	0	2	-1.035
BedroomAbvGr	2919	2.86	0.8227	3	0	8	1.933
KitchenAbvGr	2919	1.045	0.2145	1	0	3	19.73
TotRmsAbvGrd	2919	6.452	1.569	6	2	15	1.162
Fireplaces	2919	0.5971	0.6461	1	0	4	0.07213
GarageYrBlt	2918	2412	1816	1984	1895	9999	13.51
GarageCars	2918	1.767	0.7616	2	0	5	0.2335
GarageArea	2918	472.9	215.4	480	0	1488	0.9334
WoodDeckSF	2919	93.71	126.5	0	0	1424	6.721
OpenPorchSF	2919	47.49	67.58	26	0	742	10.91
EnclosedPorch	2919	23.1	64.24	0	0	1012	28.31
X3SsnPorch	2919	2.602	25.19	0	0	508	149
ScreenPorch	2919	16.06	56.18	0	0	576	17.73
PoolArea	2919	2.252	35.66	0	0	800	297.9
MiscVal	2919	50.83	567.4	0	0	17000	562.7
MoSold	2919	6.213	2.715	6	1	12	-0.4574
YrSold	2919	2008	1.315	2008	2006	2010	-1.156
House_Age_Yrs	2919	36.48	30.34	35	-1	136	-0.5058
RemodAdd_Age_Yrs	2919	23.53	20.89	15	-2	60	-1.339
Garage_Age_Yrs	2918	28.07	25.8	25	-200	114	1.614

Table 3: Predictor variables most correlated with original response variable.

Predictor Variable	Response Variable	Correlation	R^2
OverallQual	SalePrice	0.79	0.63
GrLivArea	SalePrice	0.71	0.50
GarageCars	SalePrice	0.64	0.41
GarageArea	SalePrice	0.62	0.39
TotalBsmtSF	SalePrice	0.61	0.38
X1stFlrSF	SalePrice	0.61	0.37
FullBath	SalePrice	0.56	0.31
TotRmsAbvGrd	SalePrice	0.53	0.28
House_Age_Yrs	SalePrice	-0.52	0.27
RemodAdd_Age_Yrs	SalePrice	-0.51	0.26

Table 4: Predictor transformations most correlated with transformed response variable.

Response Variable	Predictor Transformation	Correlation	R^2
SalePrice_sqrt	LotArea_log:OverallQual	0.856	0.732
SalePrice_sqrt	GrLivArea_log:OverallQual	0.852	0.727
SalePrice_sqrt	OverallQual_2:GarageCars	0.851	0.724
SalePrice_sqrt	OverallQual_sqrt:GarageCars	0.851	0.724
SalePrice_sqrt	OverallQual_2:TotRmsAbvGrd_log	0.851	0.724
SalePrice_sqrt	OverallQual_sqrt:TotRmsAbvGrd_log	0.851	0.724
SalePrice_sqrt	X1stFlrSF_log:OverallQual	0.851	0.724
SalePrice_sqrt	OverallQual_2:LotArea_log	0.850	0.723
SalePrice_sqrt	OverallQual_sqrt:LotArea_log	0.850	0.723
SalePrice_sqrt	OverallQual_2:GrLivArea_log	0.847	0.717
SalePrice_sqrt	OverallQual_sqrt:GrLivArea_log	0.847	0.717
SalePrice_sqrt	OverallQual_2:X1stFlrSF_log	0.844	0.713
SalePrice_sqrt	OverallQual_sqrt:X1stFlrSF_log	0.844	0.713
SalePrice_sqrt	TotRmsAbvGrd_log:OverallQual	0.841	0.707
SalePrice_sqrt	OverallQual_log:GrLivArea_log	0.837	0.700
SalePrice_sqrt	OverallQual_2:TotRmsAbvGrd	0.835	0.698
SalePrice_sqrt	OverallQual_sqrt:TotRmsAbvGrd	0.835	0.698
SalePrice_sqrt	OverallQual_log:X1stFlrSF_log	0.834	0.695
SalePrice_sqrt	OverallQual_2	0.828	0.685
SalePrice_sqrt	OverallQual_sqrt	0.828	0.685
SalePrice_sqrt	OverallQual:GrLivArea	0.827	0.685
SalePrice_sqrt	UtilitiesAllPub:OverallQual_2	0.827	0.684
SalePrice_sqrt	UtilitiesAllPub:OverallQual_sqrt	0.827	0.684
SalePrice_sqrt	OverallQual_2:OverallQual_log	0.827	0.684
SalePrice_sqrt	OverallQual_sqrt:OverallQual_log	0.827	0.684
SalePrice_sqrt	LotArea_log:OverallQual_log	0.827	0.683
SalePrice_sqrt	OverallQual:GarageCars	0.826	0.682
SalePrice_sqrt	OverallQual_2:GrLivArea	0.826	0.682
SalePrice_sqrt	OverallQual_sqrt:GrLivArea	0.826	0.682
SalePrice_sqrt	StreetPave:OverallQual_2	0.825	0.680
SalePrice_sqrt	StreetPave:OverallQual_sqrt	0.825	0.680
SalePrice_sqrt	OverallQual_2:GarageArea	0.823	0.678
SalePrice_sqrt	OverallQual_sqrt:GarageArea	0.823	0.678

Response Variable	Predictor Transformation	Correlation	R^2
SalePrice_sqrt	OverallQual_log:OverallQual	0.823	0.677
SalePrice_sqrt	OverallQual_2:OverallQual	0.822	0.676
SalePrice_sqrt	OverallQual_sqrt:OverallQual	0.822	0.676
SalePrice_sqrt	OverallQual_2:TotalBsmtSF_log	0.822	0.675
SalePrice_sqrt	OverallQual_sqrt:TotalBsmtSF_log	0.822	0.675
SalePrice_sqrt	OverallQual_2:FullBath	0.821	0.674
SalePrice_sqrt	OverallQual_sqrt:FullBath	0.821	0.674
SalePrice_sqrt	CentralAirY:OverallQual_2	0.817	0.667
SalePrice_sqrt	CentralAirY:OverallQual_sqrt	0.817	0.667
SalePrice_sqrt	OverallQual	0.816	0.666
SalePrice_sqrt	UtilitiesAllPub:OverallQual	0.813	0.660
SalePrice_sqrt	Condition2Norm:OverallQual_2	0.812	0.659
SalePrice_sqrt	Condition2Norm:OverallQual_sqrt	0.812	0.659
SalePrice_sqrt	OverallQual_2:OverallCond_log	0.810	0.656
SalePrice_sqrt	OverallQual_sqrt:OverallCond_log	0.810	0.656
SalePrice_sqrt	OverallQual_2:GarageCars_2	0.809	0.655
SalePrice_sqrt	OverallQual_2:GarageCars_sqrt	0.809	0.655

Appendix C. R Code

```r
# TBD
```

References

Azur, Stuart, M. 2012. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" March. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/.

Chang, W., J. Cheng, JJ. Allaire, Y. Xie, and J. McPherson. 2015. "Shiny: Web Application Framework for R. R Package Version 0.12.1." Computer Program. http://CRAN.R-project.org/package=shiny.

CNNMoney. 2010. "Best Places to Live." http://money.cnn.com/magazines/moneymag/bplive/2010/snapshots/PL1901855.html.

De Cock, D. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education, Vol. 19, No. 3.* https://ww2.amstat.org/publications/jse/v19n3/decock.pdf.

Faraway, J. 2014. *Linear Models with R.* 2nd Edition. New York, NY: Chapman; Hall/CRC.

Kaggle. 2016. "House Prices: Advanced Regression Techniques," August. https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Luttik, J. 2000. "The Value of Trees, Water and Open Space as Reflected by House Prices in the Netherlands." *Landscape and Urban Planning, Vol. 48, Issues 3-4*, May. https://ww2.amstat.org/publications/jse/v16n2/datasets.pardoe.html.

Pardoe, I. 2008. "Modeling Home Prices Using Realtor Data." *Journal of Statistics Education, Vol. 16, No. 2.* https://ww2.amstat.org/publications/jse/v16n2/datasets.pardoe.html.

Prabhakaran, S. 2017. "Missing Value Treatment," April. https://datascienceplus.com/missing-value-treatment/.

R Core Team. 2015. "R: A Language and Environment for Statistical Computing." Journal Article. http://www.R-project.org.

Wickham, H. 2009. *Ggplot2: Elegant Graphics for Data Analysis (Use R!).* New York, NY. http://ggplot2.org.