A study into the predictive ability of property descriptors on sale price in Ames, Iowa

Executive Summary

This study analyses the sale price of houses, conducted under the request of the city government. It primarily aims to provide insight into the value of a given property based on a series of descriptors. Identifying key characteristics associated within high value houses, a predictive model is created to estimate the value of a given property, without actually putting it on the market.

Numerous methods are considered as part of this study; however, a ridge regression is proposed due to its predictive accuracy, reasonable interpretability, and awareness and counteraction of overfitting. A few of the important predictors identified for the government of Ames to consider are as expected such as the size and quality of the house, as well as a good heating system for the cold climate. Others are less obvious such as the predictive importance of the basement attributes.

Some limitations in the model are considered. These primarily exist when extrapolating the model for use outside the range of the training data, either by date or geography.

Data processing and exploratory data analysis

Missing data handling

Some fields within the observed dataset return *blank* or *NaN* entries. The absence of a data point may hold information in itself, as missing data may identify the given property does not possess the relevant feature specific to a particular variable (see Appendix: Table 1 for full details of missing data) Sauro (2015) suggests imputation should be used to recover some information in partially erroneous data collection. Judgemental analysis of *LotFrontage* suggests that missing values in this variable occur when a given property has no land between the street and itself, where the frontage is equal to 0. As a result, this analysis has replaced blank entries with ‘0’.

Similarly, multiple categorical variables have missing entries where the given variable does not apply to that property. For example, A house with no alley being asked what material their alley is made out of. In instances such as this, a new category, ‘*NA’*, is created and replaces missing entries.

However, not all missing data can be attributed to systematic errors in collection; instead, isolated blanks may be caused by inputter error, employee mistakes, technical issue, or misplaced information. Allison (2001) suggests that data missing completely at random should be omitted, as there is no systematic reason to explain its absence, e.g. It does not appear to have any dependency on the other covariates. Some entries in the variables *MasVnrArea* & *MasVnrType* are blank, despite having a possible option for the user to select if this variable does not apply. It is possible that the integrity of these data points are compromised, and the user has not been able to successfully complete the entire survey. For this reason, 4 data-points with missing values in *MasVnrArea* & *MasVnrType* were omitted from further analysis.

The errors in the data itself and the challenges in regressing samples where the subjects self-select their answer give rise issues of reliability, potentially risking the quality of subsequent analysis. Future data collection should consider safeguarding against data inputting errors, by incorporating a checking procedure. Explanation of the reasons for missing data would aid future analysts, as an understanding of why some data is non-existent would help them to confirm an approach to tackle the issue.

Similarly, effective study design should always allow for subjects to identify when a question does not apply to them, so that these adjustments are not required.

Transformation of Response

Judgemental analysis theorizes the need for a transformation in the response variable, as the majority of sales are likely to be above a minimum value, but a few properties may be sold for a significantly greater amount.

Mean	175,192.47	Max	615,000
Median	158,225	Min	35,000
Skewness	1.60	Kurtosis	4.23

Table 1. Statistical information on *SalePrice*

Table1 shows a positive value for skewness, indicating a reasonably strong right skew in the response. This is supported by the sample mean being greater than the sample median, as well as visual analysis of the sampling distribution of *SalePrice* in Figure1. Alongside this, Kurtosis is ~1.3 greater than that of a normal distribution, suggesting the existence of thinner tails. Together, these measures suggest the existence of outlying datapoints at the upper end of *SalePrice.* To account for this, this analysis has elected to perform a logarithmic transformation of the response. Although this has a beneficial effect on the distribution, correcting for skew and kurtosis, this study is aware that improving model accuracy in this way will reduce the interpretability.

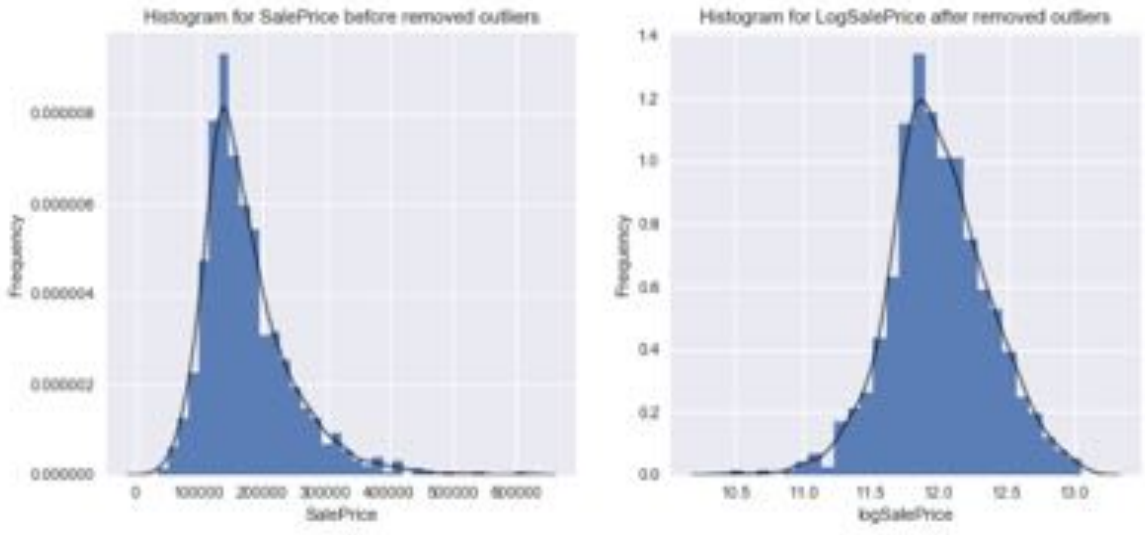


Figure1. Histogram for *SalePrice* and *LogSalePrice* after removing outliers

Despite this transformation, some observations remain more than 3 standard deviations from the mean, identifying them as outliers (Howell, 1998) and they have subsequently been removed from further analysis. In practise, this removes 3 out of the 804 original observations from the upper end of the training set.

Exploratory Analysis of Predictors

Analysis of numerical and categorical predictors

Exploratory analysis was used to gain a greater understanding of predictor variables. Scatter plots, boxplots and histograms were used to understand the sampling distribution of each x variable, as well as indicate whether a relationship with the dependant existed. Some box plots also show that few extreme values exist. Full details of the graphs can be found in the appendix.

Majority of numerical variables have reasonably strong linear relationships with the response, for example, *TotalBsmtSF* and *GrLivArea* (Appendix: Figures 1-7). This indicates that further study for model building will concentrate on linear methods, and these variables are considered to be included. However, non-linear relationships also exist between Sale Price and some variables such as *BsmtFinSF1* and *OpenporchSF* (Appendix: Figures 8-14). In that case, a nonlinear model should also be taken into consideration.

It has been observed that some variables have insufficient non-zero values (Appendix: Table 2). For example, Figure 2 shows that *PoolArea* only has 3 non-zero values out of all observations, which means that only 3 houses in the sample have a pool. These non-zero values cannot fairly represent the variable's influence on sale price, as insufficient variation in the predictor is captured to make accurate predictions. Including variables of this kind in model can generate bias, as well as affect the predictive performance. The insufficient variance captured by the sample size issue also exists in some classes of categorical variables. For example, there is only 1 house is in the *NoSewr* class of utility type, while all the others are in the AllPub class. To account for this, those variables will not be used to build the model.

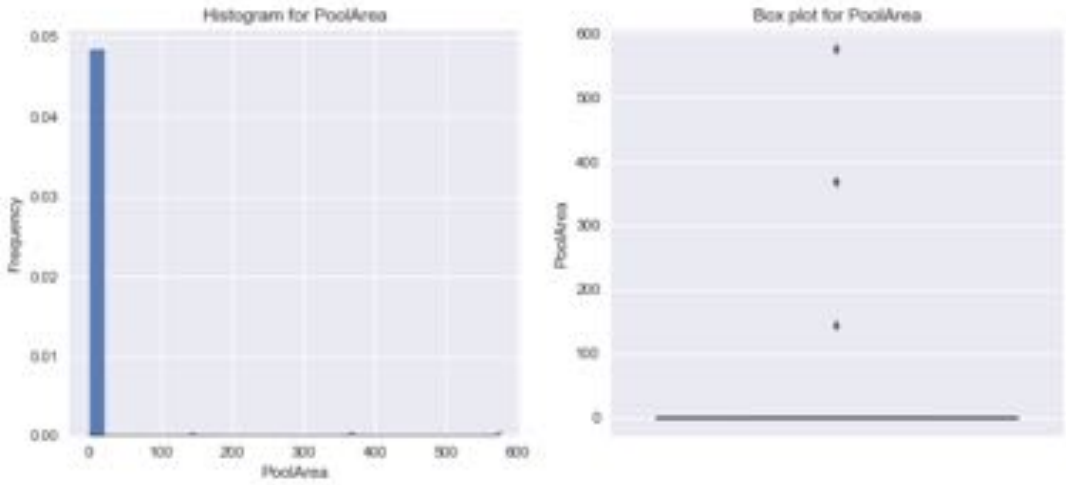


Figure 2. Histogram and boxplot for *PoolArea*

Boxplots for categorical variables indicates that there are linear trends between some ordinal variables and sale price (Appendix: Figures 15-25). Most of this kind of variables are quality terms, such as *ExterQual*, *BsmtQual* and *HeatingQC*. These boxplots illustrate that a house with better quality elements were sold for a higher price.

Correlation Analysis

A correlation matrix was used to help understand which variables might have the greatest predictive ability in regression, and to investigate the existence of collinearity between predictors. The 25 variables with the greatest absolute correlation with *LogSalePrice* were identified and illustrated in a heatmap correlation matrix, shown in Figure 3.

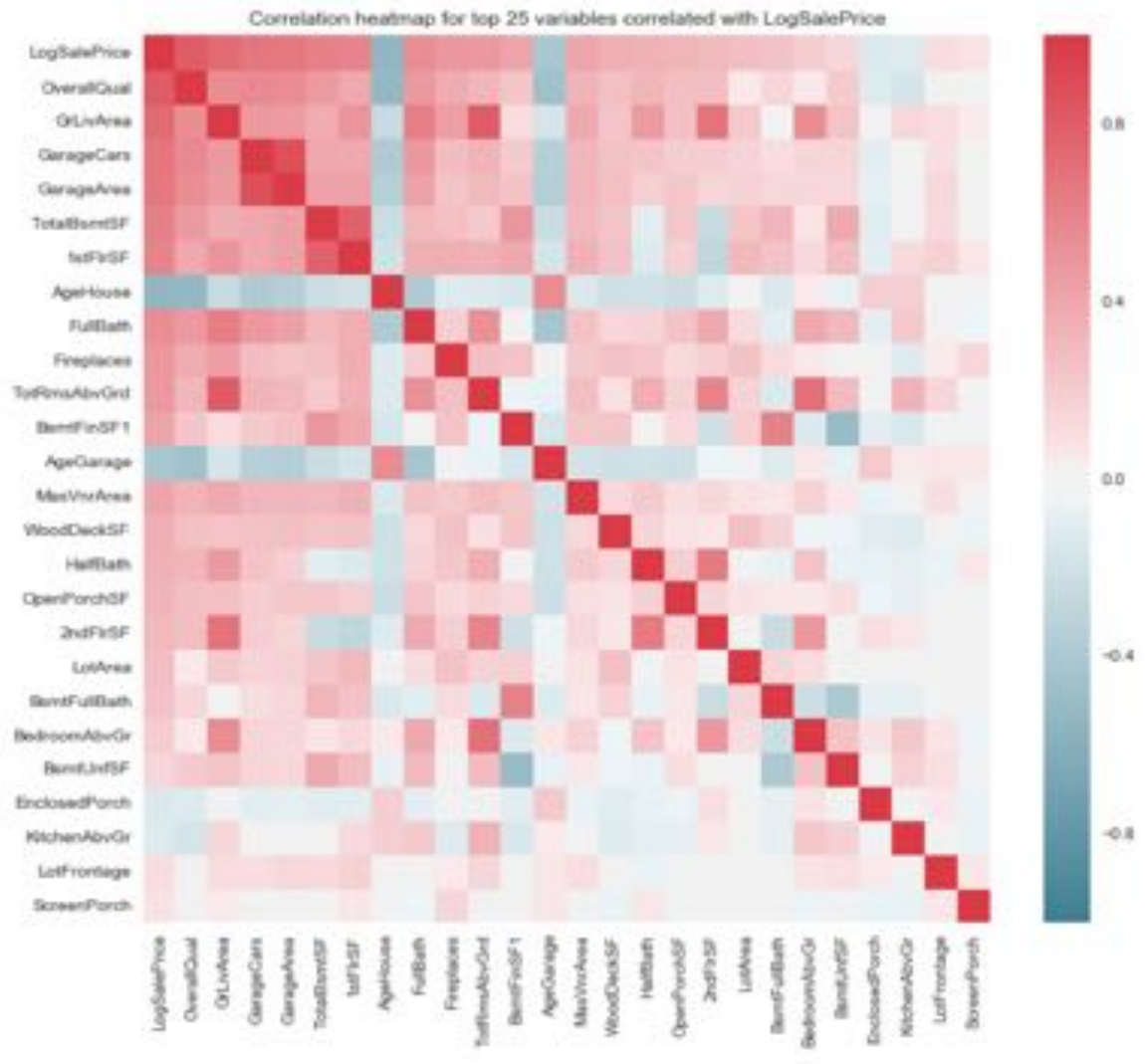


Figure 3. Correlation heatmap for top 25 variables correlated with *LogSalePrice*

As shown, the strongest correlation with the *LogSalePrice* is the overall quality of the property, the size of the property, and the size of the garage. Information on the size of the garage is captured by 2 variables, *GarageCars* and *GarageArea*. Judgemental analysis suggests that these predictors are likely to account for the same variation in the response; this is supported by the reasonably high collinearity shown in Figure 3. A similar interaction can be seen in *GrLivingArea* and *TotRmsAbvGrd*, which both indicate the size the of a property.

Negatively correlated predictors, shown in blue are those for which one would expect property value to fall as the explanatory variable rises, such as the age of the property.

Feature Engineering

Nominal Variables

The training data collected contains many different data types, some which require transformation to be appropriate for regression analysis. This study has opted to split the majority of both ordinal and nominal categorical variables, each with *k* categories, into *k-1* binary variables, commonly referred to as dummy variables.

Ordinal Variables

Ordinal variables, such as *OverallQual,* are based on a progressive scale (Figure 4), where the gaps of each interval cannot be confirmed. This study has opted to assume equally spaced intervals for each possible score. Although this may impact predictive accuracy, the variance will be reduced as significantly fewer coefficients will be estimated in a less complex model; moreover, the interpretability of a given model will be improved.

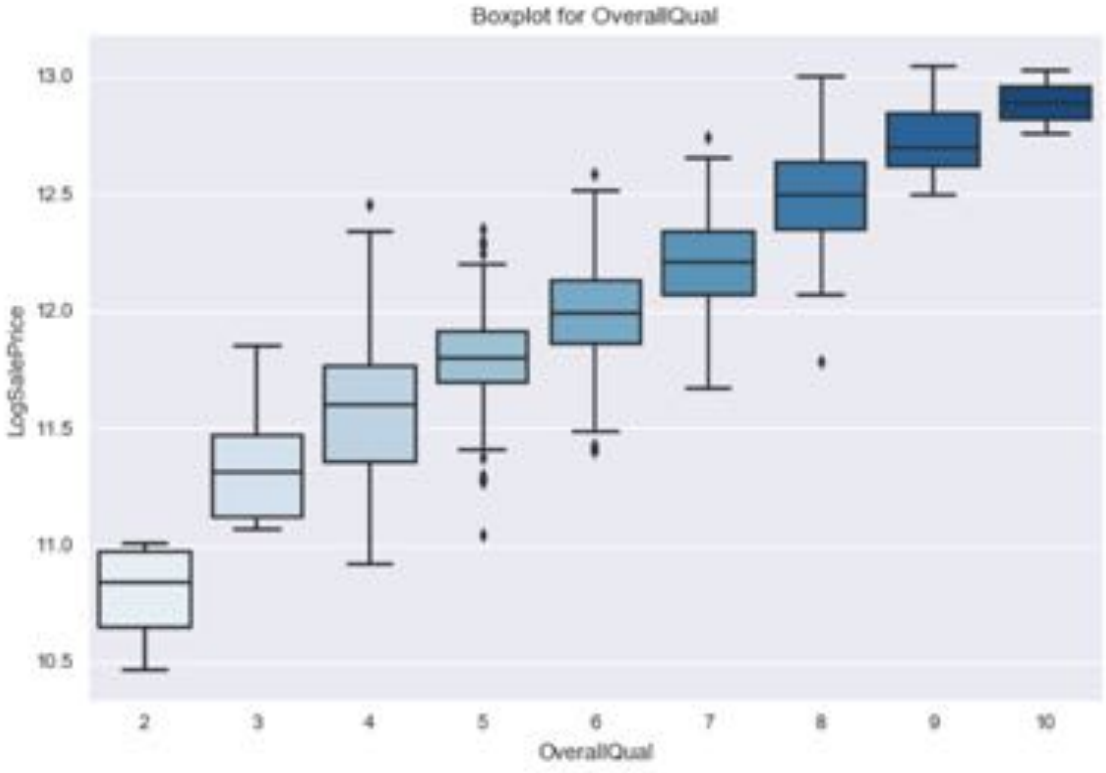


Figure 4. Box plot for *OverallQual*

Transformations of Dates

The sample of property sales was collected over 4 years, causing possible issues creating a model using any timestamp as a predictor. Judgemental analysis suggests that the age of a given property is a more likely to be an effective predictor, rather than the year of construction. It can be argued that property may be subject to value depreciation over time, which is a function of age, rather than year of construction. At a given time *t*, a property's age and year of construction will hold the same information; however, over time age increases whilst year of construction remains static. Not only does this allow any predictive model to be more dynamic, and robust to changes in time; it also increases interpretability.

As a result, 2 new predictors are formed to measure the age of the garage and house, based on either its original construction or remodelling where applicable. A value of 0 is given to *AgeGarage* if there is no garage at the house.

AgeGarage = YrSold - GarageYrBlt
AgeHouse = YrSold – max{YearBuilt, YearRemod/Add}

Clustering of Neighborhoods

Location is often considered a key attribute in influencing housing prices, and as such we would expect there to be a large amount of variation in *Sale Price* based on where the properties are located. This can be seen below.

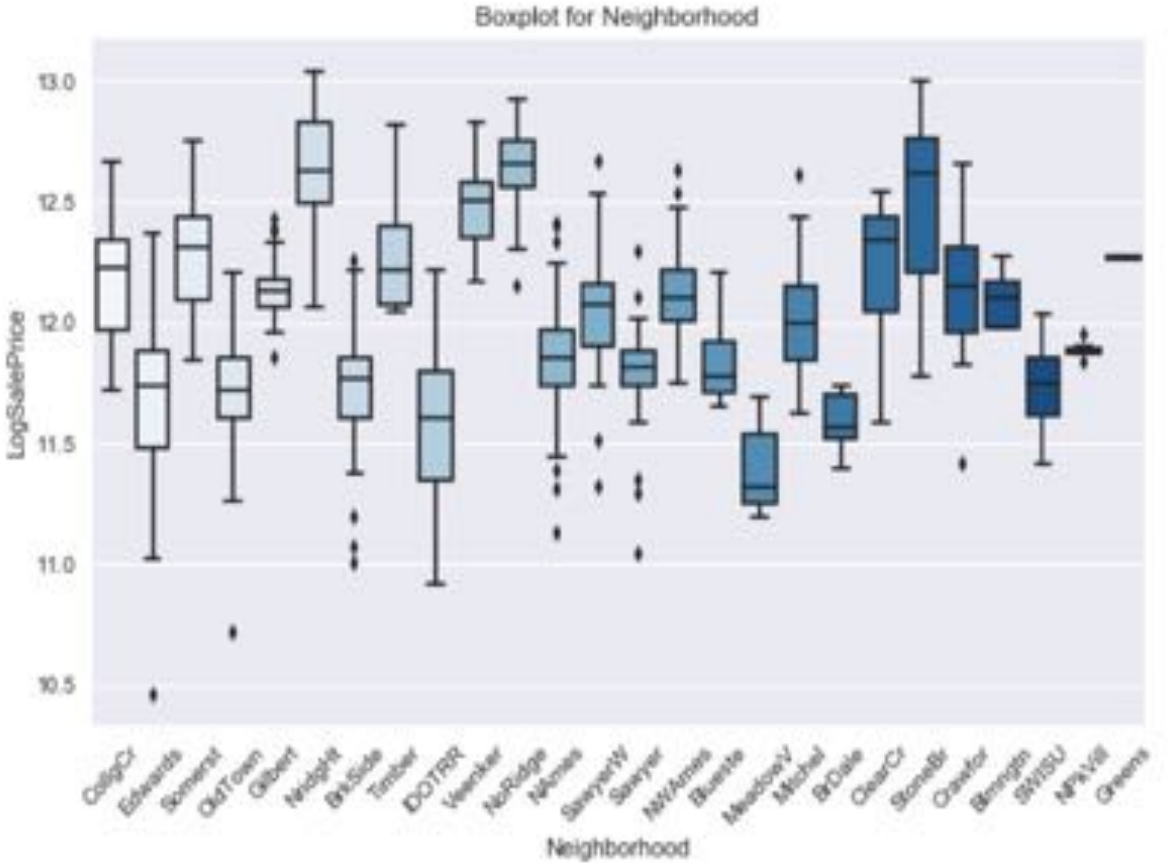


Figure 5. Boxplot for *Neighborhood* before clustering

However, the *Neighborhood* variable contains too many individual categories, many of which have a rather small number of samples or are extremely similar to other categories. Certain neighborhoods were grouped based on their housing prices to handle this issue and reduce the number of groups. This reduces model complexity and is beneficial to predictive performance since there is a smaller number of coefficients to estimate within a single categorical variable, with minimal loss of information.

Neighborhoods were arranged by mean of *Sale Price* and compared using a Welch's t-test, assuming unequal sample variances, at a 10% significance level. The mean for each class and p-value for each t-test, are shown in Table 3 and 4 of the appendix. If all the p-values between any two or more neighborhoods were significant, they were grouped together. Welch's t-test assumes the samples are drawn from a normal distribution, which can be an issue especially with neighborhoods with only a few observations. However, in the interest of clustering these smaller neighborhoods with similar ones, this

method was carried out to reduce the category size to 9, shown in Figure 6. The neighborhood groupings are given in Table 5 of the appendix.

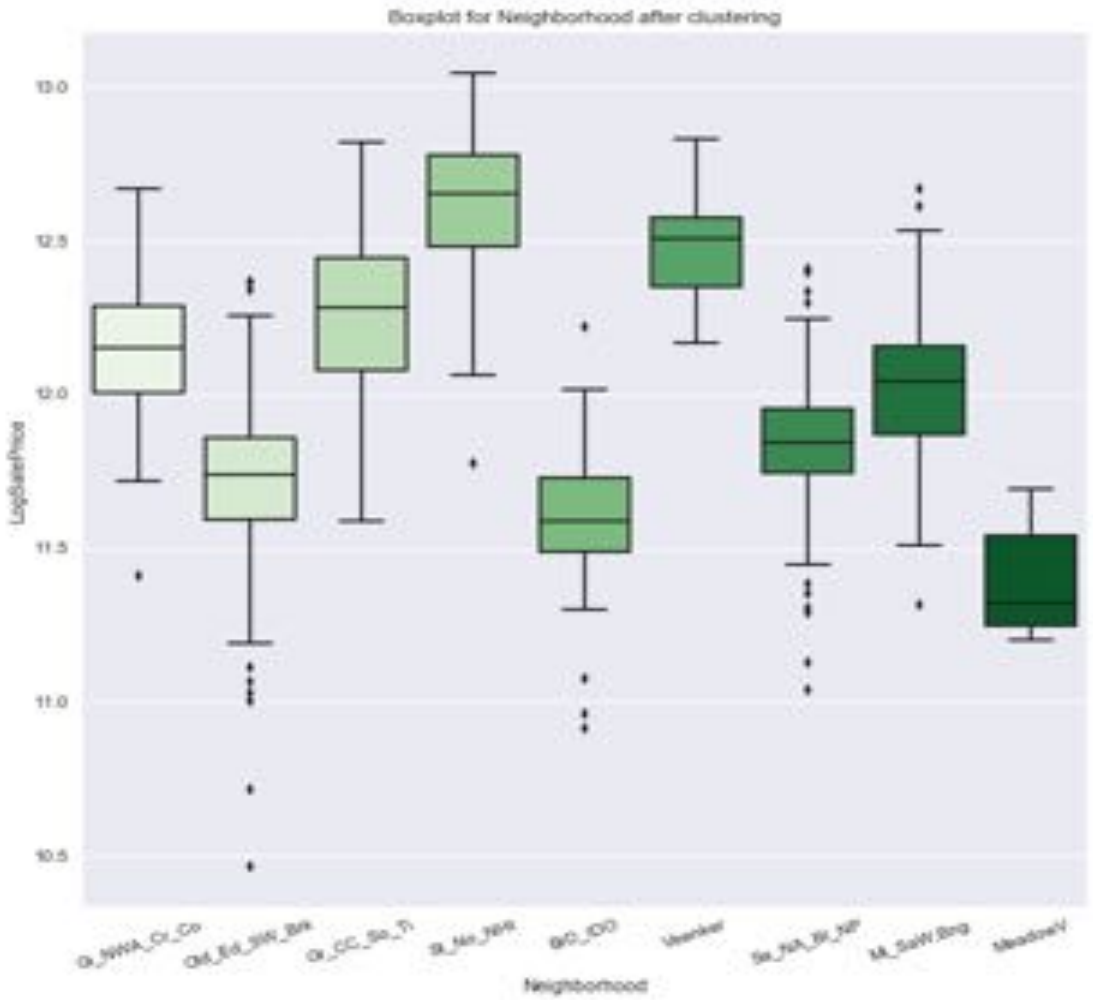


Figure 6. Boxplot for *Neighborhood* after clustering

Interactions

Another point of consideration was the interaction between certain predictors. In this context, an interaction arises when the expected effect on *Sale Price* due to a change in one of the predictors is dependent on the value of another predictor. Judgemental analysis would suggest an effect like this to be seen in the quality and size of an area within a house. For example, an increase in the size of the garage would be expected to lead to an increase in price, but this price increase is dependent on the quality of the garage. Though the actual size increase in the garages of two houses are identical, the one with higher quality i.e. better materials, finish, or simply perceived as higher quality would expect to see a larger price increase.

Interaction terms were only considered among predictors that were highly related to the response since introducing too many interactions, even if they are plausible, increases model complexity and will lead to overfitting. The following 4 interactions were considered:

*BsmtSF_Qual = TotalBsmtSF * BsmtQual*
*GarageArea_Qual = GarageArea * GarageQual*

*GrLiv_Rooms = GrLivArea * TotalRmsAbvGrd*
*GarageArea_Cars = GarageArea * GarageCars*

Methodology

Analysis considered a range of predictive modelling methods in order to balance predictive accuracy and interpretability. Although non-parametric models have considerable advantages in fewer assumptions made about the distribution of data, they exhibited a lower level of predictive accuracy than alternatives in this instance, as well as being considerably limited in terms of interpretability. As a result, the two methods considered in greater detail are both variations of parametric multivariate regression.

Similar methods of varying levels of complexity are compared using cross validation. To reflect the generalisation/test error, the loss function selected for use during supervised learning is the mean absolute error. This loss function accounts for negative and positive error cancelling each other out, but does not emphasise greater errors by squaring them.

Forward Selection with Ridge Regression

Exploratory data analysis identified many predictor variables demonstrating a linear relationship with the response, *LogSalePrice*. However, limited computer processing power is unable to accommodate all variables observed by the dataset. Based on correlation, a reduced subset of 70 predictors (Appendix: Table 6) was selected from the complete list for further consideration.

Although all considered predictors in this subset have a correlation with the response above 20%, a correlation matrix identifies issues of multicollinearity between some predictors. Collinear variables may account for the same variation in the response, and including them in a given model will have the adverse effect of increasing variation in coefficient estimates, in turn making predictions less accurate. Although OLS assumptions only prohibit predictors with perfect collinearity (=1), this study has opted not to consider some variables due to high collinearity before variable selection methods. Additionally, we removed some variables for which there were only a small number of observations (example shown below), as well as binary variables that were already represented by the inclusion of a different binary variable within that predictor i.e removing CentralAir_N because CentralAir_Y was already included. This leaves us with a subset of 51 predictors (Appendix: Table 7)

Variation in predictions are greatly increased when an insufficient variation in predictors is captured by the training sample. This may occur as a result of splitting categorical variables into the k-1 binary variables; for example, in *PoolQC* shown in Table 2.

PoolQC	**NA**	**Fair**	**Typical**	**Good**	**Excellent**
# of observations	795	0	1	1	0

Table 2. The number of observations for each class in *PoolQC* variable

Regularisation is a variation of linear multiple regression that is aware of ordinary least squares tendency to overfit a model to the training data. Conceptually, a penalty term is introduced to the existing loss function which penalizes additional and potentially unnecessary complexity. The complexity function and its weight is dependant on model chosen, as well as a hyperparameter (λ) selected before training begins. In this instance, it is likely that OLS will select too many factors to predict house price under the false pretence that it improves accuracy; regularisation methods counter this by shrinking the

coefficients of predictors. The hyperparameter is determined using 5-fold cross validation over a grid of potential shrinkage coefficients.

Ridge regression is a regularisation method where the penalty term is based on the ℓ_2 norm. To ensure the intercept B0 is not penalised, as well as ensuring all explanatory variables are measurable on the same scale, we centre and standardise predictors by subtracting the values in the predictors by their respective means, and dividing by the standard deviations. For numerical predictors, we divide by two standard deviations to put them on roughly the same scale as the binary predictors.

Stepwise forward selection is used to identify the most effective predictors to include in the model from the previously identified subset of 51. Like the smaller subset of variables, a stepwise method of selection is preferred over the more reliable best subsets methodology due to its significantly smaller computer processing requirement.

Ridge regression will shrink together the coefficients of correlated predictors, and is therefore best suited to situations where many variables contribute evenly to the predictive power of the model. In this instance, the large set of predictors all explain some variation in *LogSalePrice,* making Ridge methodology particularly appropriate. Moreover, this method will further address some of the partially collinear predictors seen in this dataset.

In order to reduce the bias of estimating coefficients based on a single training and validation set, supervised learning is conducted using cross validation. This study has opted not to conduct leave one out cross validation (LOOCV) despite it's 'unbiased' prediction, due to the significant computing power it requires. Instead, 10-fold cross validation is preferred as a result of it' comparatively lower variance and ease of execution.

Of the 51 variables originally considered, this method has identified 21 deemed to be significant in predicting *LogSalePrice*, as well as calculated their respective, regularised coefficients which were shrunk to 91.7% of the standard OLS coefficients. This relatively low shrinkage is expected since forward selection was first applied to the predictors. The greatest of the coefficients are *OverallQual* and *GrLivArea,* which are measure of the overall quality of a given property, and the overall size of a given property. Both of these predictors are applicable to every property, and therefore are more likely to capture the majority of variation in the population. Moreover, it fits with preliminary analysis that measures that describe entire property, rather than smaller aspects, may account for more variation in the sale price. It is of interest that one of these is a subjective measure. *OverallQual* is an opinion based metric, and although at current appears to be a significant predictor, may not if the assessor's criteria changes, or consumer preference trends change over time.

As all predictors are standardised, they have no scale; thus, the absolute value of coefficients are also a measure of relative influence. The standardization along with the log transformation of *Sale Price* changes the interpretation of the coefficients. For example, among houses in which all the predictors besides *OverallQual* are equal, we would expect a house with overall quality *one standard deviation* higher to have about a 15.2% higher *Sale Price* under this model. Note that this does not mean a predictor is necessarily more *important* than one with a lower coefficient; rather a change in the standard deviation corresponds to a higher percentage change. The coefficients and predictors are given in Table 3.

Predictor	Coefficient	Predictor	Coefficient
OverallQual	0.152	BsmtExposure	0.035
GrLivArea	0.215	PavedDrive_Y	0.015
BsmtSF_Qual	0.085	KitchenQual	0.043
BsmtFinSF1	0.077	FullBath	-0.011
AgeHouse	-0.064	Neighborhood_Gr_CC_So_Ti	0.024
GarageCars	0.054	Neighborhood_Gi_NWA_Cr_Co	0.023
MSZoning_RM	-0.028	Neighborhood_St_No_NHt	0.022
CentralAir_Y	0.023	MSSubClass_30	-0.016
LotArea	0.044	AgeGarage	0.026
HeatingQC	0.045	GarageArea	0.047
FireplaceQu	0.046		

Table 3. The coefficient for predictors in ridge regression (after forward selection)

There are a few interesting things to note about the above output. There are some predictors related to general characteristics of a house that we are expecting to see with stronger associations such *OverallQual, GrLivArea, LotArea, GarageArea*. Furthermore, we note that *FireplaceQu* and *HeatingQC* are included in the model. Ames, Iowa tends to get very cold in the winter, so having a good source of warmth is important in determining the value of a house. There also seems to be conflicting information in *AgeGarage* and *AgeHouse.* We might expect there to be a negative coefficient for both, as an older house tends to have a lower price, but the effect of *AgeGarage* could be reversed due to other predictors in the model. This is the same case with *FullBath.*

This model selected some binary variables such as *MSZoning_RM*, *MSSubClass_30* and a few neighborhood groups. While some of these binaries help identify external characteristics unrelated to the actual house that are responsible for variations in the house price as well as helpful towards predicting home values, the interpretations are tricky because these dummy variables are only *some* of the categories in the original variable. It is not usual to only include some dummies of a categorical variable and exclude others because it changes the interpretation of the reference group to be all the excluded dummy variables combined together. However, with that said, we could tentatively say these particular dummy variables have predictive power over others i.e. whether or not a home is one story and built before 1945 (*MSSubClass_30)* is important to consider when predicting home value.

Another variable selected is *BsmtSF_Qual*, an interaction term between the basement size and quality. The main effects (TotalBsmtSF and *BsmtQual*) however, are not included in the model. This goes against the hierarchical principle of interaction terms. However, our model forgos these rules in the interest of simplicity and emphasis on more prediction rather than interpretability. While there are more complex methods to deal with these issues, we do not consider them here.

Elastic Net

The elastic net is a similar, but crucially different variation of the regularisation method seen in ridge regression. An elastic net approach continues to apply a penalty term to the loss function of the model, the result of which shrinks the coefficients of predictor variables to reduce complexity, and thus variance of predictions. The complexity term, however, is a weighted average of the ℓ_2 regularisation used in ridge, and the ℓ_1 regularisation used in the lasso method.

The advantage of this is that ℓ_1 regularisation has the capability to set coefficients to '0', as it is based on the sum of absolute coefficients. This gives it the capability to conduct its own variable selection. The advantage of this over traditional variable selection methods, is that this is a continuous procedure which generally leads to lower variance. However, ℓ_1 regularisation alone is less able to perform smooth reductions, and so by combining ℓ_2 regularisation, variable selection and smooth shrinkage of collinear predictors can be achieved.

The process in determining the initial set of predictors is the same as with the ridge model, except we do not put them through forward selection first since one of the features of elastic net is variable selection. We standardize the predictors as before. Fitting this model to our initial set of 51 predictors results in 19 of the predictors being set to 0, and thus 32 predictors used in the model. Of these 32, the coefficients have been shrunk a moderate amount to 78.6% of the OLS coefficients. Though the number of predictors and thus model complexity is higher than with the ridge model, there is a greater amount of shrinkage to account for the increase in variance and potential overfitting. As with the ridge model, the greatest coefficients of the selected predictors came from *OverallQual* and *GrLivArea.* Much of the concepts and interpretability when using the ridge model pertain to this model as well. The chosen predictors and their coefficients are below. Refer to Table 8 in the appendix for the full list of predictors including those set to 0.

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
OverallQual	0.158	BsmtFinSF1	0.068	Electrical_SBrkr	0.002
BsmtSF_Qual	0.053	LotArea	0.031	Neighborhood_Old_Ed_SW_Brk	-0.008
GrLivArea	0.221	CentralAir_Y	0.022	Neighborhood_Gr_CC_So_Ti	0.017
GarageCars	0.046	PavedDrive_Y	0.014	Neighborhood_Sa_NA_Bl_NP	-0.004
GarageArea	0.037	Exterior2nd_VinylSd	0.003	BsmtFinType1_NA	-0.004
TotalBsmtSF	0.023	GarageQual	0.004	BsmtFinType2_NA	-0.001
KitchenQual	0.038	BsmtExposure	0.023	Neighborhood_Gi_NWA_Cr_Co	0.012
AgeHouse	-0.045	MSZoning_RM	-0.022	Neighborhood_BrD_IDO	-0.006
FireplaceQu	0.036	GarageFinish_Fin	0.004	GarageType_Attchd	0.007
OpenPorchSF	0.004	LotShape_Reg	-0.009	Neighborhood_St_No_NHt	0.017
HeatingQC	0.032	MSSubClass_30	-0.010		

Table 4. The coefficient for predictors in elastic net regression

The elastic net method selects many of the same predictors as the forward selection with ridge model, which is a good indicator to the predictive importance of those variables. Because of the larger set of predictors and higher shrinkage, many of the same variables in this model have smaller coefficients than the ridge model and some are very close to zero. This model considers some variables that were not included in the previous model such as *OpenPorchSF*, *GarageFinish_Fin*, *LotShape_Reg*, and *Electrical_SBkr.* Some of these coefficients are shrunk very close to 0, but they are still deemed important to consider in home value. Something like an open porch or a nice finished garage would be appealing to consumers and can be sold at a higher price, but it is also important to consider aspects of a home consumers may not even think about such as the lot shape or electrical system. This model also considers more of the neighborhood predictors which allows for better interpretation of the effects of location. We can see that among similar houses, those in neighborhoods like Greens, Clear Creek, Somerset and Timberland would be worth more on average than those in Oldtown, Edwards, Sawyer West and Brookside.

Interestingly, elastic net chose to keep *BsmtFinType1_NA* and *BsmtFinType2_NA* even though they are effectively saying the same thing which is whether or not there is a basement in the home. Furthermore,

this information is already captured in *TotalBsmtSF* in 0 values. While it is clear that homes with basements are on average worth more, and that basement attributes are important when predicting prices, he inclusion of all these together could have an impact on predictive power. Similar to the ridge model, we have issues where an interaction term is included without the full main effects, and only some dummy variables of a categorical variable are included while others are excluded. This latter issue is a product of variable selection as well as us using a reduced subset of initial variables. As before, more complex methods are needed to account for this, so we do not consider this given the methodology and favor predictability.

It's worthy to mention that neither model selected the predictor *ExterQual* which assess the external quality of the house. Though this is again a subjective predictor, one might expect this to be important given the relatively high correlation (.64) with the response and the fact that the exterior look and quality of a house are typically important when it comes to price. The fact it is excluded does not mean it isn't important, but rather it is likely well captured in other variables, such as *OverallQual*. Overall, both models perform similarly in terms of predictive power, but forward ridge includes fewer variables and is less complex, while elastic net has smaller coefficients but considers more aspects of a house that could be important to predicting the home value.

Assumptions

The assumptions are checked through residual analysis and correlation matrix. There are six assumptions for multiple linear regression models, the first being linearity of the true model. In linear regression, if $X = x_1, \ldots, x_p = \mathbf{x}$, then $Y = B_0 + B_1*x_1 + \ldots B_p*x_p + \varepsilon$, for some population parameter B_0, B_1, …, B_p and a random error. Assumption 2, 3, 4: Conditional mean of ε given X is zero: $E(\varepsilon|X) = 0$. Variance of ε given X is constant: $Var(\varepsilon|X) = \sigma^2$. 3). All the error pairs ε_i and ε_j (i is not equal to j) are independent. These assumptions are checked by the Residuals vs Fitted Value plot and Absolute Residuals vs Fitted Values plots. From Figure 7 and 8, it can be observed that the residuals are scattered randomly around zero, so it can be assumed that the conditional mean of error is zero. There is no pattern in residual plots, which means that the residuals are not changed with fitted values. Thus, the constant error variance assumption, and independence of errors assumption hold. Linearity is difficult to check given the number of predictors in the model, but the randomness of the residual plots would support this assumption as well.

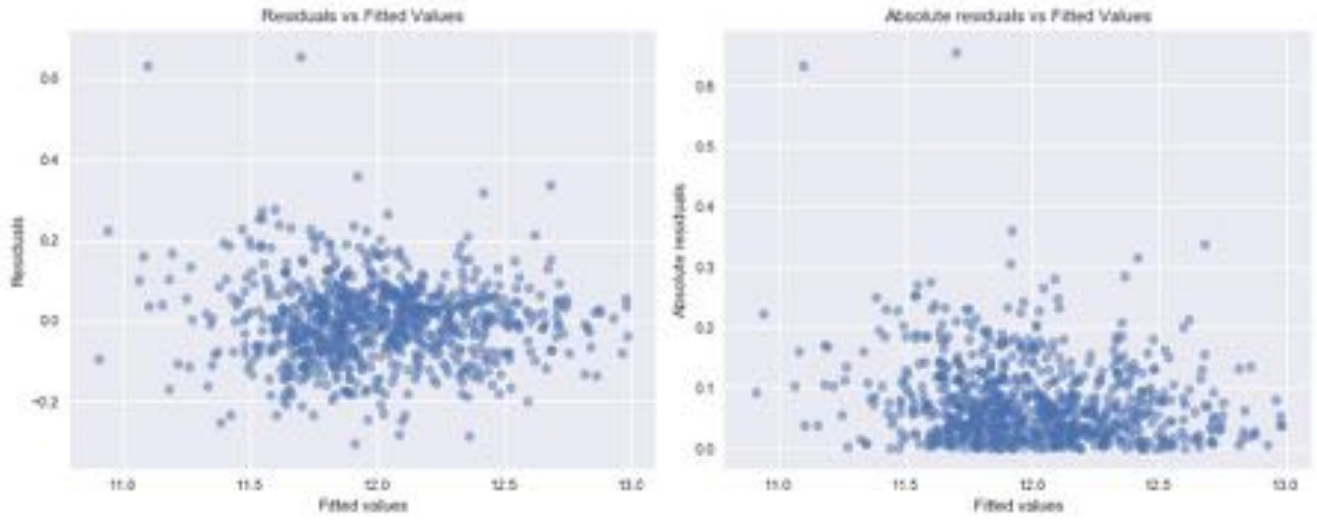


Figure 7. Residual plots for ridge (with forward selection)

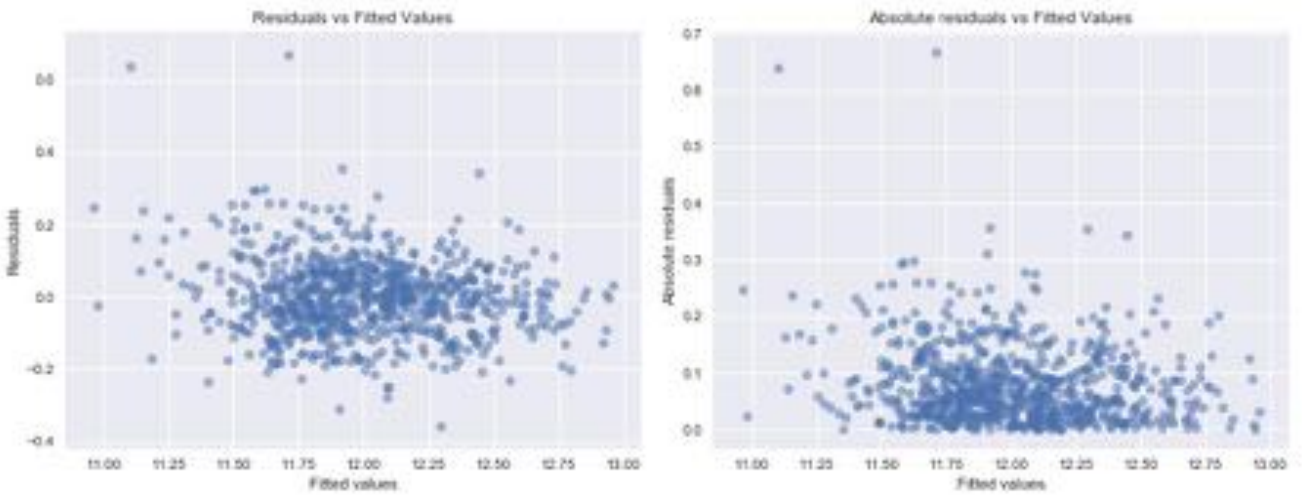


Figure 8. Residual plots for elastic net

Assumption 5 holds in that the distribution of $X_1, \ldots, X_p$ is arbitrary, as well as assumption 6 stating there is no perfect multicollinearity. The correlation between variables use in the model are checked before model building. To satisfy this assumption, variables which have severe collinearity with others are deducted. The predictors used to build models have no perfect collinearity, except for some interaction terms. The correlation heatmap for assumption checking is in Figures 26 and 27 of the appendix.

Validation set results and comparison with other approaches.

Six predictive models were trained and considered as part of this analysis. The validation scores of each can be seen in Table 5. It is important to note the cross-validation scores were calculated based on *logSalePrice,* but are still comparable.

The validation score indicates that ridge regression combined with forward selection can be the model that predicts best. Compared to others, models used shrinkage methods performed better, while KNN performs worst. Although ridge regression is favoured by this study, alternatives are briefly discussed below.

Model	**KNN**	**OLS**	**Ridge**	**Lasso**	**Elastic Net**	**PCR**
Validation Score	52483.60	14621.89	13225.03	13277.61	13280.37	20918.68
CV Score	.128	.089	.081	.083	.082	.081

Table 5. The validation scores and CV scores of each model

K-Nearest Neighbours

The most notably different model compared to alternatives was the K-nearest neighbours model. A non-parametric approach commonly used to estimate a regression, without making the same assumptions regarding the distribution of the underlying data. The forward selection is combined with KNN model, so that it will minimize the cv score, and return the model with a subset of most promising predictors and k value.

KNN was unable to perform as accurately as a ridge regression in this instance. It is likely due to the number of predictors required to estimate house price. As additional predictors are added, it become exponentially more difficult to find observations in the training data which are 'local' to a given query point, causing predictions to be generated on decreasingly similar observations.

Non-parametric models such as this would not be ideal in this situation regardless of predictive accuracy. They require significant computer processing power as well as memory, to store an entire training set of data and recalculate distances from each observation to every query point. It was for these reasons the KNN predictive model was not taken forward.

Ordinary Least Squares

An ordinary least squares (OLS) approach to the multiple linear regression provides an unbiased, consistent, and best linear unbiased estimator of the regression parameters when all assumptions of MLR are satisfied. By checking the correlation matrix, there is no perfect collinearity (Appendix: Figure 28), but some correlation exists between interaction terms. The principle of marginality suggests that a model including interaction terms should also include all the lower-order relatives of that term. Therefore, those interaction terms are kept in OLS model.

The form of an OLS predictive model is a closed form expression, and therefore allows the computation of predictions to be fast and convenient. However, training an OLS model on training data, and setting the model complexity with the aim of minimising the expected loss leads to overfitting. OLS was trained on a smaller subset of only 25 predictors (Appendix: Table 9) since there is no variable selection or regularization term. This subset was chosen by first identifying the top 30, then removing variables with high multicollinearity. Even on the reduced subset, there are concerns of overfitting. This is due to additional predictors that explain little variation in the response being added unnecessarily, such as *MasVnrArea.*

Principal Component Regression

An alternative approach to the OLS and ridge regression may seek to reduce the number of estimated parameters by dimension reduction; in turn, reducing complexity aims to decrease the variance of prediction errors. The principal components regression ("PCR") is a method generating new variables based on linear combinations of existing variables, without giving direct consideration to their effect on the response. As a result of this, there is an implicit assumption that directions of highest variance in the original predictors are the ones most associated with the response.

PCR is most effective when a small number of components account for a large part of the variation in the predictor data. In this instance, this is primarily *OverallQual* and *GrLivingArea.* However, adding additional components has the positive effect of decreasing bias at the cost of increasing variance.

The principal components regression methodology is not capable of performing variable selection alone, as such, a stepwise forward selection approach has been taken to estimate the most effective predictors to use in the model. As previously discussed, forward selection is preferred over a best subsets algorithm due to the significantly reduced computing requirement, despite in marginally less reliable selection of variables.

PCR and ridge regression are very similar in that ridge smoothly and progressively shrinks the coefficients of all principal components. PCR on the other hand does not effect coefficients with the largest variance, but discards those with smaller variance. It's because of this smooth shrinkage that we

preferred to use ridge even though the CV scores were the same. Furthermore, PCR performed significantly worse than ridge in the validation set. This could be due to PCR discarding predictors with low variance that are actually important to predicting home prices.

Lasso

Similar to the ridge regression model, the lasso approach applies a penalty term for complexity when training the model, which reduces overfitting. As discussed with the elastic net, a lasso regression is able to shrink regression parameters to '0' using the ℓ_1 regularization, performing variable selection as such. The lasso method applied to our initial set of 51 predictors selects 30 to be used in the model and shrinks the coefficients to about 58% of OLS. Though the validation score for the lasso model is slightly lower and the shrinkage is higher than the elastic net method, we opt for elastic net since it also incorporates the ℓ_2 regularization and thus also adjusts for collinearity between predictors. Although not selected as the favoured model, the lasso approach selects a similar set of predictors as forward selection. These are primarily measures of overall house size and quality, reinforcing the importance of these predictors.

Limitations

Despite a sufficiently large sample size, this analysis has been conducted based on highly restricted training set. Observations within the sample are house sales that span 5 years, 2006 - 2010. At the time of analysis, observations of 7 - 11 years old.

Outdated training data may raise concerns with validity in application of predictive models to future house sales. It is highly conceivable that consumer preferences may have changed over time, and given characteristics of property that were previously associated with higher value sales may not increase price going forward. Similarly, new preferences may have come to light since the original observations were taken which are more significant now than in the past. Examples of this may be energy efficiency, which is becoming increasingly important both economically and socio-environmentally. The city government may be less able to accurately predict house prices using a model trained on outdated information.

Moreover, a limited dataset ranging only 5 years may not span a great enough period to observe any cyclical trends or variation based on time. Economic boom and bust cycles can span decades; such variation is not captured by this data set. Housing bubbles, or increased demand during times of prosperity may inflate house prices far above their value during times of recession, limiting the effectiveness of extrapolating the proposed predictive model outside dates covered. It is noted that the 2006 - 2010 range covers the global credit crisis in 2007, before which housing demand was high due to the ease of acquiring mortgages. The crisis was primarily triggered by defaulting subprime loans, which over subsequent years drastically changed the methods for which financial lenders assessed eligibility for mortgages. As such, this study argues that a post 2010 data set would provide consistent and far more representative view of today's housing market.

The training data containing 804 observations is strictly limited to Ames, Iowa, USA; and the validity of any predictive model created based on this to make predictions in other locations should be considered. The city government is not advised to extrapolate this model to other locations, as the conditions of the area affecting house price may not be similar. For example, weather, or local culture. However, for ease of data collection, it is not possible to sample multiple every towns; it is therefore suggested that predictive models be used in comparatively similar locations, most notably in terms of urbanisation, and

demographic. In practise, this might apply to american towns of similar size, rather than large cities in other countries, such as Singapore, or rural settlements in remote locations, such as Alaska. Similarly, this dataset appears to apply specifically to domestic dwellings, and so it would not be advised to apply such a model to commercial properties or land.

Although the training set captures many characteristics of a given property, it would be challenging to fully understand all areas that are considered by prospective buyers. As a result, it is highly likely that some significant variables are not recorded by the sample, as so cannot be considered as a predictor. Key examples might be: proximity to high performing schools (not captured by neighbourhood), transport links, or energy efficiency. Qualitative analysis before sampling quantitatively may provide additional insight to the areas of greatest importance to prospective buyers.

The reliability of the training data must also be considered when assessing the limitations of any models. The sample collected has many technical measurements, such as total sqft. Such measurements are potentially hard to accurately take, and systematic errors is data collection will have adverse effects on trained models.

As mentioned before, the models used and our methodology do not treat dummy variables and interaction terms in the usual manner. Because of selecting a subset of predictors based on correlation with the response as well as using forward selection, each variable in a sense is treated independently, even the dummy variables and interaction terms. This is a limitation in that is reduces the general interpretability of the model, however, we allow this for our purposes of prediction and reducing the number of variables used in the model.

Business Implications

The use of predictive models to provide insight to the expected outcome of an unknown can be of great value to businesses as well as the city government. However, it is highly unlikely that point predictions, i.e. a specific figure, are going to be correct. Confidence intervals of predictions, or even probability density of predictions can provide additional information toward the uncertainty of a given expected value. This is because there is lower variation of a given prediction when the query point is closer to the centre of the training set. These more advanced uses of predictive models may provide additional benefit to users.

It may be within the public interest to allow city governments to have a better understanding of property value, should it lead to fairer taxes. However, many individuals may not wish to share the information required in order to generate predictions.

A predictive model for estimating sale price of a property is likely to have a few alternative commercial uses:

1. The first will help to identify where a given property is undervalued (or overvalued) by the market, recognising the opportunity to buy for the less than the true value, wait for the market to adjust, and then sell back at a profit. Economists however, may argue that market forces of supply and demand are greater at accurately valuing property at that given time. Suggesting that any model that disagrees is likely to contain errors or inaccuracies.
2. The second may be to help to financial lenders to accurately assess the value of the property, so that it may be used as collateral against a loan. Models may be applicable, however managerial analysis for local experts may provide additional insight. Moreover, mortgages are commonly taken out to purchase the property they are collateralised against. In these cases, the Naive-1 estimate of house price (i.e. what it just sold for) is likely to be accurate.

3. To understand if the value of a property can be increased by making changes that cost less than the expected increase. Such use may not be recommended, models were trained based on observation data, rather than by actively changing variables in A/B testing.

Final Remarks

The most appropriate model to solve the business case in question is the ridge regression. Its power to shrink coefficients to reduce overfitting is key, as well as limiting the influence of collinear predictors. The details of the model suggest that the key influencing factors of house price centre around the size and quality. Higher quality, larger properties are worth more, as fit with expectation. The model suggests that the presence of specific characteristics only present in certain properties are less associated, such as a masonry veneer, or pool.

This analysis, however, has many limitations. Sampling methods used to collect training data allow for specificity appropriate to the city government, however this reduces the validity and applicability for other users. The city government should be aware of extrapolating this model outside the given geographical and date range seen in the sample, as trends may be isolated in time or location.

References

Allison, P. D., 2001. *Missing data - Quantiative applications in the social sciences.* Thousand Oaks, CA: Sage.

Howell, D. C. (1998). *Statistical methods in human sciences.* New York: Wadsworth

Sauro, J., 2015. *Measuring U - 7 Ways to Handle Missing Data.* [Online] Available at: http://www.measuringu.com/blog/handle-missing-data.php [Accessed 21 September 2017].

Appendix

Tables

Table 1. The number of missing value for each kind of variables

The number of missing values for each kind of variables	
Continuous Variables	
LotFrontage	157
MasVnrArea	4
Discrete Variables	
GarageYrBlt	36
Ordinal Variables	
BsmtQual	21
BsmtCond	21
BsmtExposure	22
BsmtFinType1	21
BsmtFinType2	21
FireplaceQu	386
GarageFinish	36
GarageQual	36
GarageCond	36
PoolQC	801
Fence	644
Nominal Variables	
Alley	749
MasVnrType	4
GarageType	36
MiscFeature	781

Table 2. Number of non-zero values for each continuous variable (outliers removed)

Number of non-zero values for each continuous variable	
LotFrontage	640
LotArea	797
MasVnrArea	305
BsmtFinSF1	567
BsmtFinSF2	107
BsmtUnfSF	730
TotalBsmtSF	776
1stFlrSF	797
2ndFlrSF	341
LowQualFinSF	8
GrLivArea	797
GarageArea	762
WoodDeckSF	374
OpenPorchSF	415
EnclosedPorch	133
3SsnPorch	12
ScreenPorch	70
PoolArea	2
MiscVal	22
SalePrice	797
AgeHouse	789

AgeGarage	760

Table 3. The mean of each class in *Neighborhood* variable

Class in *Neighborhood*	Mean of *SalePrice*
MeadowV	89790.00
BrDale	109272.73
IDOTRR	111095.00
OldTown	123706.82
Edwards	126203.64
SWISU	126291.67
BrkSide	126909.82
Sawyer	135542.88
NAmes	143742.12
Blueste	144000.00
NPkVill	145664.29
Mitchel	169368.18
SawyerW	178059.68
Blmngtn	181077.00
Gilbert	187879.89
NWAmes	189233.72
Crawfor	193894.57
CollgCr	200402.86
Greens	212750.00
ClearCr	214072.22
Somerst	220924.39
Timber	223256.67
Veenker	268154.55
StoneBr	285132.90
NoRidge	313386.36
NridgHt	313850.00

Table 4. The p-value for two sample t-test (classes in *Neighborhood* variable)

Tested classes	p-value
MeadowV vs BrDale	0.007
BrDale vs IDOTRR	0.872
OldTown vs Edwards	0.693
OldTown vs SWISU	0.712
OldTown vs BrkSide	0.674
Edwards vs SWISU	0.991
Edwards vs BrkSide	0.936
SWISU vs BrkSide	0.946

Sawyer vs NAmes	0.090
Sawyer vs Blueste	0.693
Sawyer vs NPkVill	0.030
NAmes vs Blueste	0.990
NAmes vs NPkVill	0.642
Blueste vs NPkVill	0.937
Mitchel vs SawyerW	0.418
Mitchel vs Blmngtn	0.353
SawyerW vs Blmgtn	0.826
Gilbert vs NWAmes	0.849
Gilbert vs Crawfor	0.564
Gilbert vs CollgCr	0.076
NWAmes vs Crawfor	0.679
NWAmes vs CollgCr	0.176
Crawfor vs CollgCr	0.561
Greens vs ClearCr	0.913
Greens vs Somerst	0.233
Greens vs Timber	0.515
ClearCr vs Somerst	0.619
ClearCr vs Timber	0.644
Somerst vs Timber	0.893
Veenker vs StoneBr	0.634
StoneBr vs NoRidge	0.417
StoneBr vs NridgHt	0.420
NoRidge vs NridgHt	0.981

Table 5. The groups of *Neighborhood* after clustering

Neighborhood	**Group**
MeadowV	MeadowV
BrDale	BrD_IDO
IDOTRR	
OldTown	Old_Ed_SW_Brk
Edwards	
SWISU	
BrkSide	
Sawyer	Sa_NA_Bl_NP
NAmes	
Blueste	
NPkVill	
Mitchel	Mi_SaW_Bng
SawyerW	
Blmngtn	
Gilbert	Gi_NWA_Cr_Co
NWAmes	
Crawfor	
CollgCr	
Greens	Gr_CC_So_Ti

ClearCr	
Somerst	
Timber	
Veenker	Veenker
StoneBr	
NoRidge	St_No_NHt
NridgHt	

Table 6. Reduced subset of top 70 predictors absolute value correlated with *LogSalePrice*

Predictor	Corr. With LogSalePrice	Predictor	Corr. With LogSalePrice
OverallQual	0.811	BsmtExposure	0.369
BsmtSF_Qual	0.728	PavedDrive_N	0.369
GrLivArea	0.727	MSZoning_RM	0.368
GarageCars	0.671	WoodDeckSF	0.367
GarageArea_Cars	0.669	GarageFinish_Fin	0.363
GarageArea_Qual	0.669	LotShape_Reg	0.36
GarageArea	0.662	GarageCond	0.358
ExterQual	0.636	MasVnrType_None	0.357
BsmtQual	0.616	HalfBath	0.347
TotalBsmtSF	0.616	MSSubClass_30	0.337
KitchenQual	0.616	OpenPorchSF	0.336
1stFlrSF	0.596	Electrical_SBrkr	0.326
AgeHouse	0.546	LotShape_IR1	0.322
FullBath	0.543	Foundation_CBlock	0.321
FireplaceQu	0.533	2ndFlrSF	0.307
Foundation_PConc	0.53	GarageType_NA	0.304
Fireplaces	0.492	GarageFinish_NA	0.304
TotRmsAbvGrd	0.482	BsmtCond	0.294
Neighborhood_St_No_NHt	0.47	MSZoning_RL	0.282
HeatingQC	0.465	GarageFinish_RFn	0.279
GarageType_Attchd	0.444	LotArea	0.276
GarageFinish_Unf	0.429	MasVnrType_BrkFace	0.272
BsmtFinType1_GLQ	0.421	Foundation_BrkTil	0.268
GarageType_Detchd	0.416	HouseStyle_2Story	0.26
BsmtFinSF1	0.414	Neighborhood_Gr_CC_So_Ti	0.256
AgeGarage	0.411	BsmtFullBath	0.254
Neighborhood_Old_Ed_SW_Brk	0.41	Electrical_FuseA	0.246
MSSubClass_60	0.41	Neighborhood_Sa_NA_Bl_NP	0.236
MasVnrArea	0.407	BsmtFinType1_NA	0.231
CentralAir_Y	0.39	BsmtFinType2_NA	0.231
CentralAir_N	0.39	MasVnrType_Stone	0.229
PavedDrive_Y	0.379	Neighborhood_Gi_NWA_Cr_Co	0.229
Exterior1st_VinylSd	0.377	Neighborhood_BrD_IDO	0.216
Exterior2nd_VinylSd	0.376	BedroomAbvGr	0.2
GarageQual	0.372	Fence_NA	0.197

Table 7. Subset of 51 predictors from original 70

OverallQual	Exterior1st_VinylSd
BsmtSF_Qual	Exterior2nd_VinylSd
GrLivArea	GarageQual
GarageCars	BsmtExposure
GarageArea_Cars	MSZoning_RM
GarageArea	WoodDeckSF
ExterQual	GarageFinish_Fin
BsmtQual	LotShape_Reg
TotalBsmtSF	MasVnrType_None
KitchenQual	HalfBath
AgeHouse	MSSubClass_30
FullBath	OpenPorchSF
FireplaceQu	Electrical_SBrkr
Foundation_PConc	GarageFinish_RFn
Neighborhood_St_No_NHt	LotArea
HeatingQC	Foundation_BrkTil
GarageType_Attchd	Neighborhood_Gr_CC_So_Ti
GarageFinish_Unf	Neighborhood_Sa_NA_Bl_NP
BsmtFinType1_GLQ	BsmtFinType1_NA
BsmtFinSF1	BsmtFinType2_NA
AgeGarage	MasVnrType_Stone
Neighborhood_Old_Ed_SW_Brk	Neighborhood_Gi_NWA_Cr_Co
MSSubClass_60	Neighborhood_BrD_IDO
MasVnrArea	BedroomAbvGr
CentralAir_Y	Fence_NA
PavedDrive_Y	

Table 8. Full list of Elastic Net Coefficients

Predictor	**Coefficient**	**Predictor**	**Coefficient**
OverallQual	0.158	Exterior1st_VinylSd	0.000
BsmtSF_Qual	0.053	Exterior2nd_VinylSd	0.003
GrLivArea	0.221	GarageQual	0.004
GarageCars	0.046	BsmtExposure	0.023
GarageArea_Cars	0.000	MSZoning_RM	-0.022
GarageArea	0.037	WoodDeckSF	0.000
ExterQual	0.000	GarageFinish_Fin	0.004
BsmtQual	0.000	LotShape_Reg	-0.009
TotalBsmtSF	0.023	MasVnrType_None	0.000
KitchenQual	0.038	HalfBath	0.000
AgeHouse	-0.045	MSSubClass_30	-0.010
FullBath	0.000	OpenPorchSF	0.004
FireplaceQu	0.036	Electrical_SBrkr	0.002
Foundation_PConc	0.000	GarageFinish_RFn	0.000
Neighborhood_St_No_NHt	0.017	LotArea	0.031
HeatingQC	0.032	Foundation_BrkTil	0.000
GarageType_Attchd	0.007	Neighborhood_Gr_CC_So_Ti	0.017
GarageFinish_Unf	0.000	Neighborhood_Sa_NA_Bl_NP	-0.004
BsmtFinType1_GLQ	0.000	BsmtFinType1_NA	-0.004

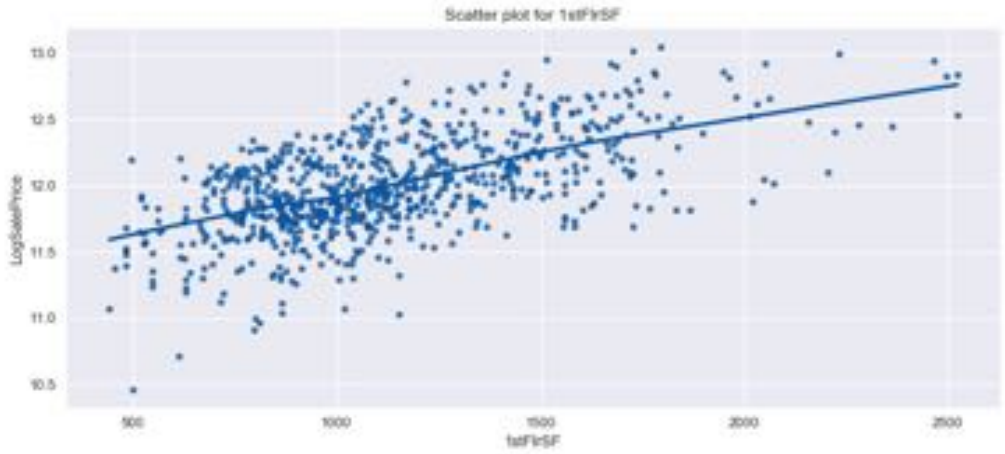
BsmtFinSF1	0.068	BsmtFinType2_NA	-0.001
AgeGarage	0.000	MasVnrType_Stone	0.000
Neighborhood_Old_Ed_SW_Brk	-0.008	Neighborhood_Gi_NWA_Cr_Co	0.012
MSSubClass_60	0.000	Neighborhood_BrD_IDO	-0.006
MasVnrArea	0.000	BedroomAbvGr	0.000
CentralAir_Y	0.022	Fence_NA	0.000
PavedDrive_Y	0.014		

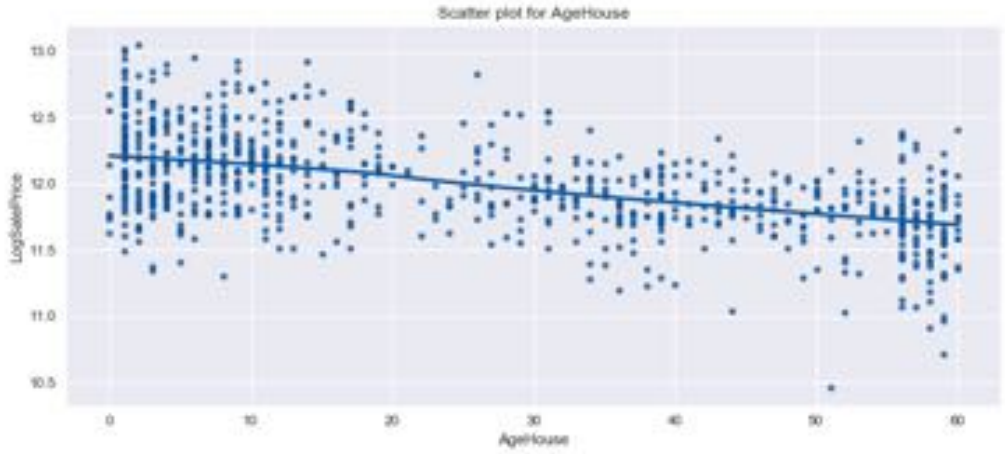
Table 9. Reduced subset of 25 predictors used for OLS

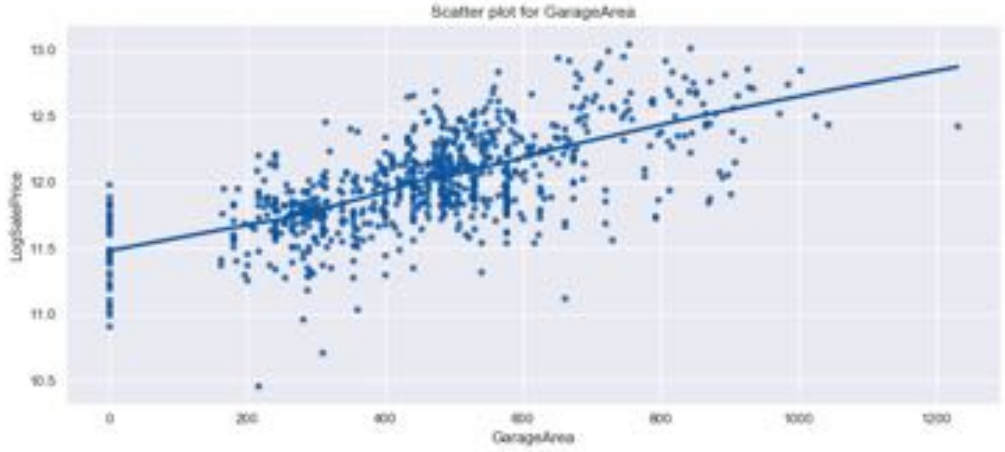
OverallQual	Foundation_PConc
BsmtSF_Qual	Neighborhood_St_No_NHt
GrLivArea	HeatingQC
GarageCars	GarageType_Attchd
GarageArea_Cars	GarageFinish_Unf
GarageArea	BsmtFinType1_GLQ
ExterQual	BsmtFinSF1
BsmtQual	AgeGarage
TotalBsmtSF	Neighborhood_Old_Ed_SW_Brk
KitchenQual	MSSubClass_60
AgeHouse	MasVnrArea
FullBath	CentralAir_Y
FireplaceQu	

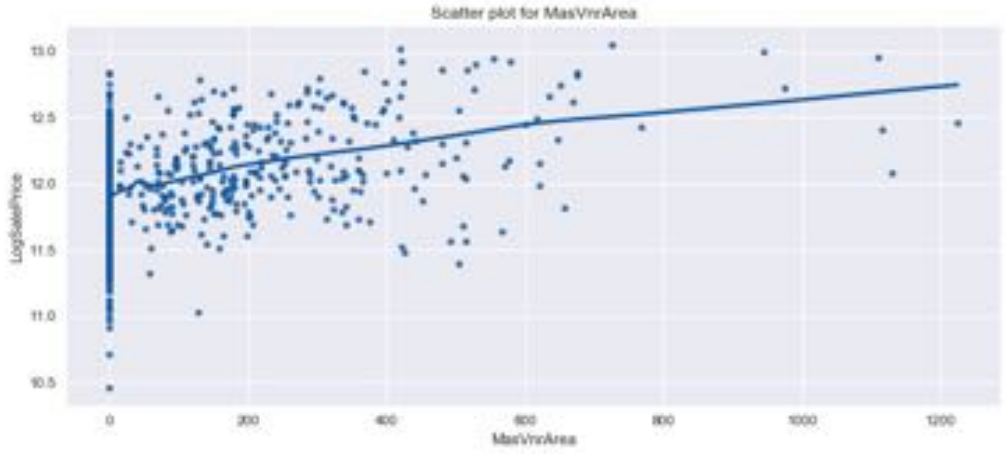
Figures

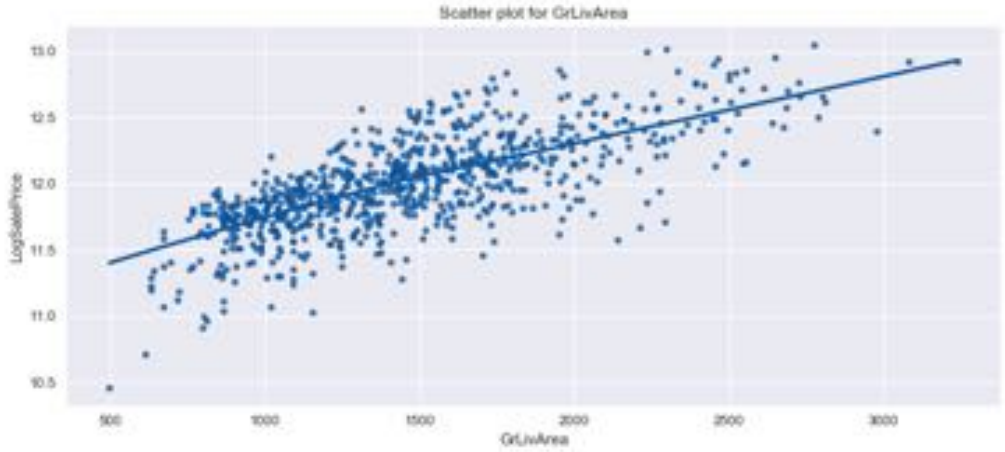
Figure 1-7. Scatter plots for numerical variables (linear relationship)

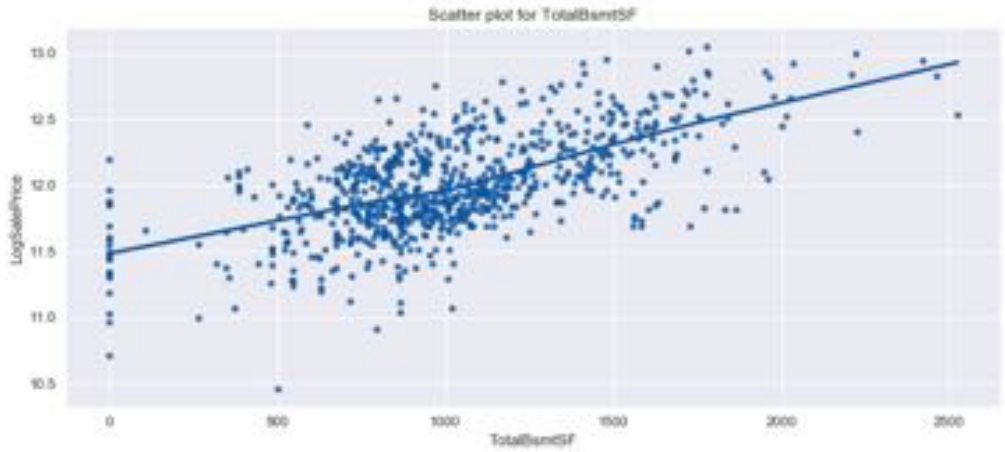



Scatter plot for AgeHouse
LogSalePrice
AgeHouse


Scatter plot for GarageArea
LogSalePrice
GarageArea


Scatter plot for MasVnrArea
LogSalePrice
MasVnrArea


Scatter plot for GrLivArea
LogSalePrice
GrLivArea


Scatter plot for TotalBsmtSF
LogSalePrice
TotalBsmtSF

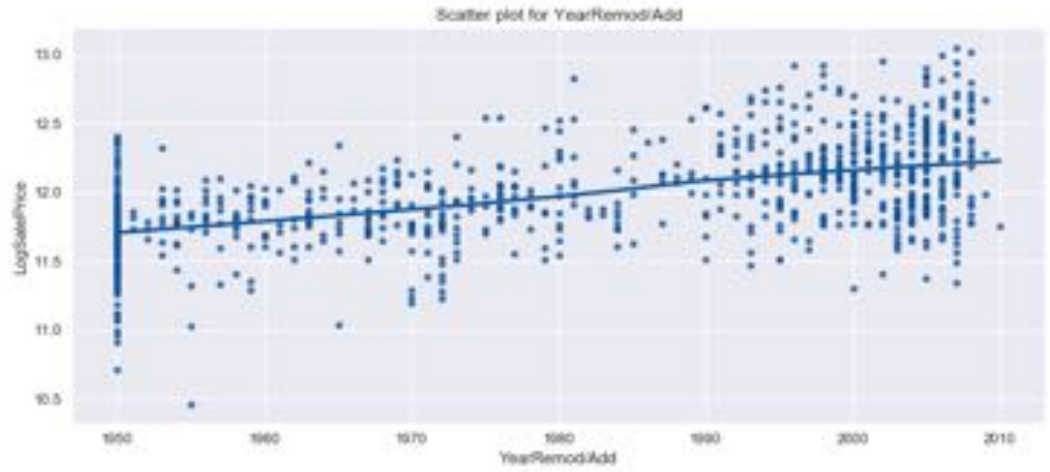
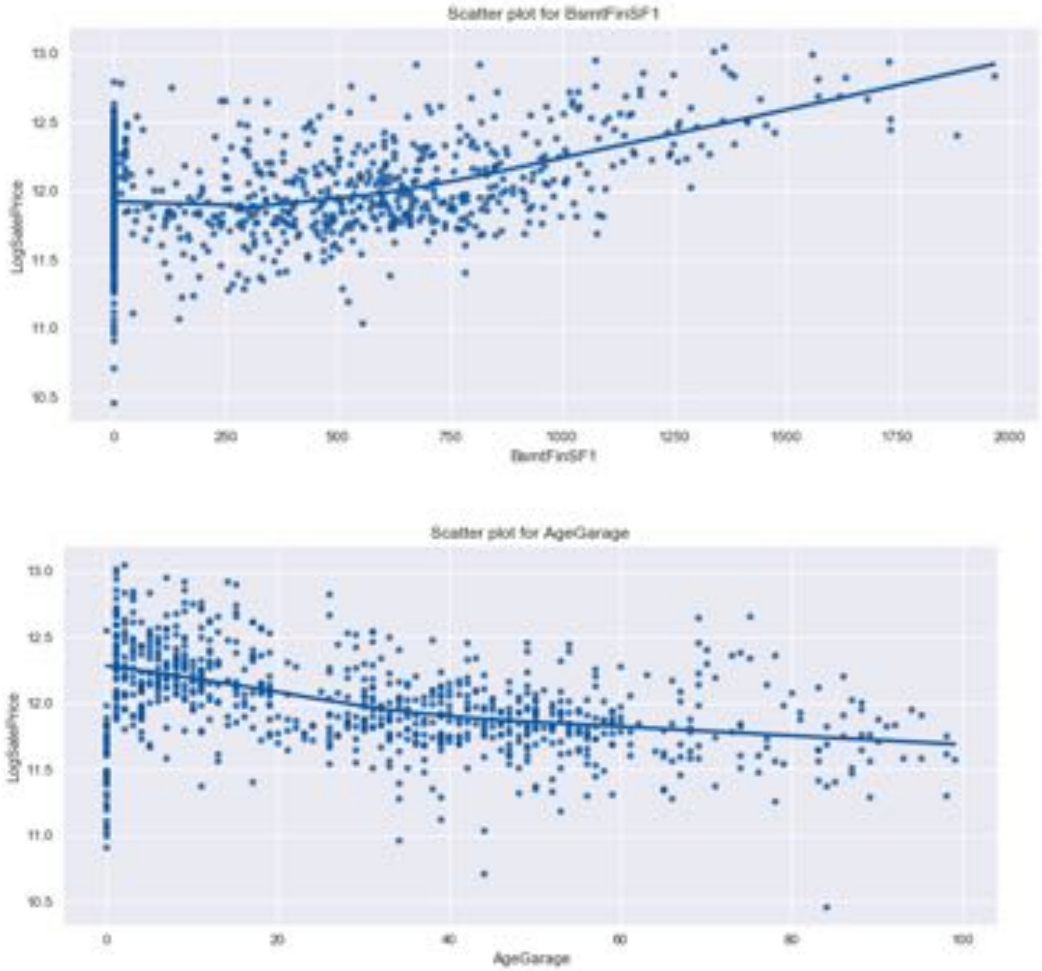
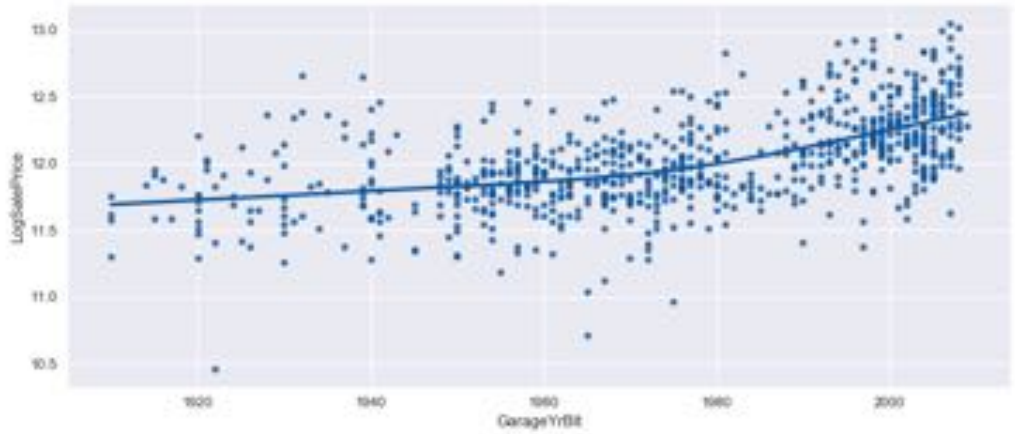
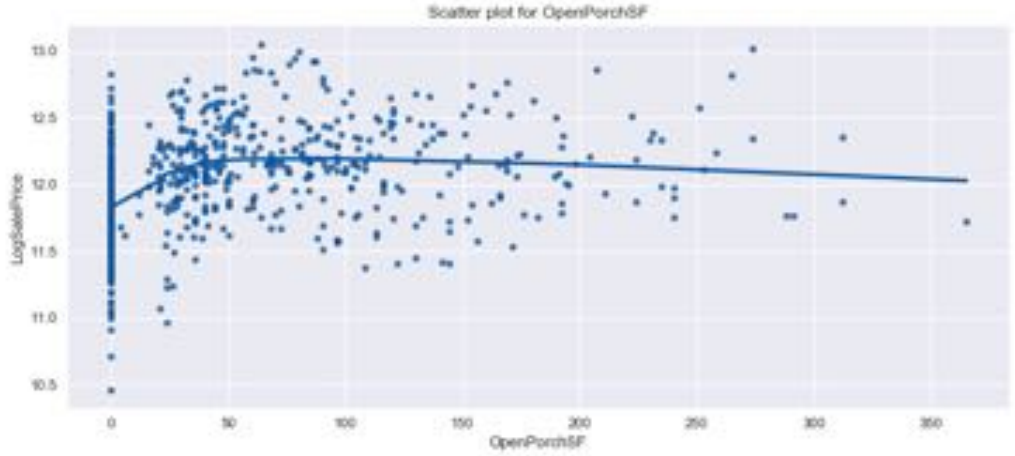

Scatter plot for YearRemodAdd
LogSalePrice
YearRemodAdd

Figure 8-14. Scatter plots for numerical variables (non-linear relationship)




LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
1920
1940
1960
1980
2000
GarageYrBlt


Scatter plot for LotFrontage
LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
0
20
40
60
80
100
120
140
160
LotFrontage


Scatter plot for OpenPorchSF
LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
0
50
100
150
200
250
300
350
OpenPorchSF

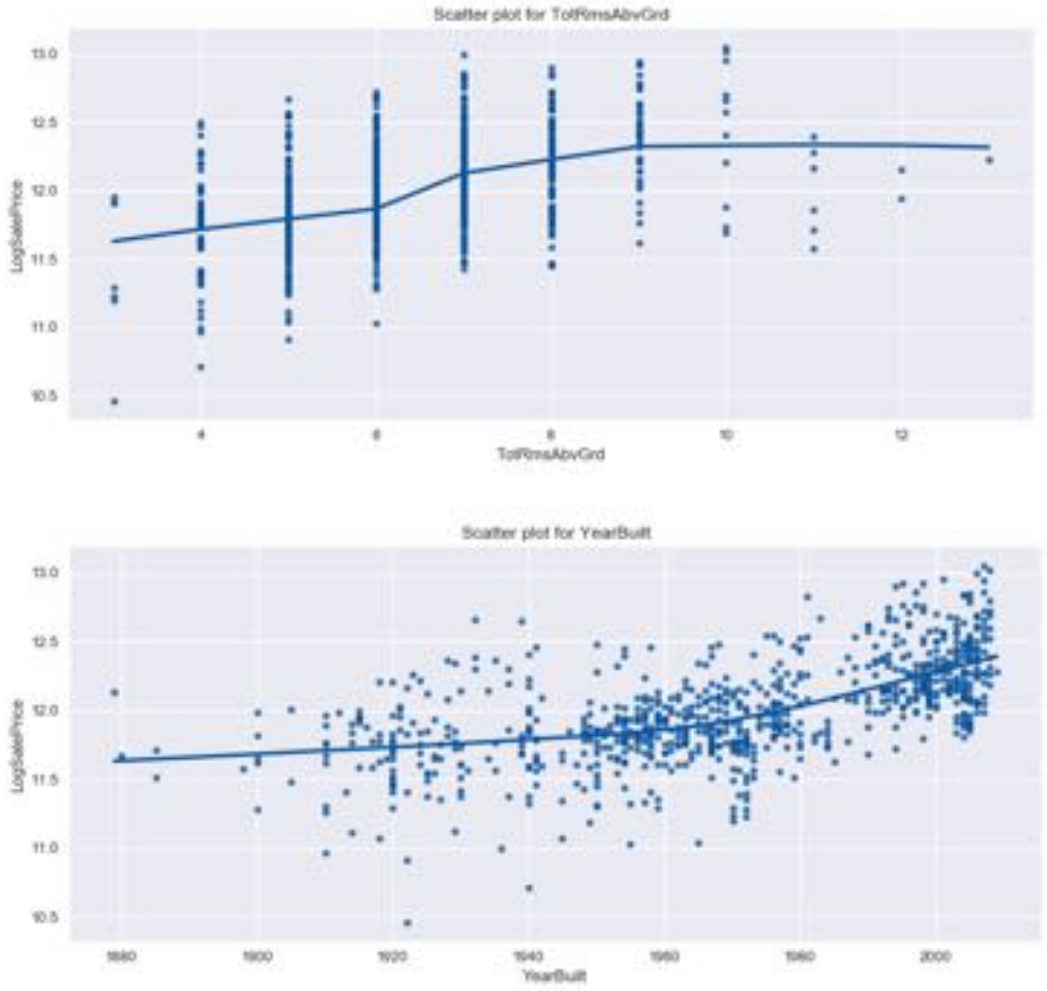
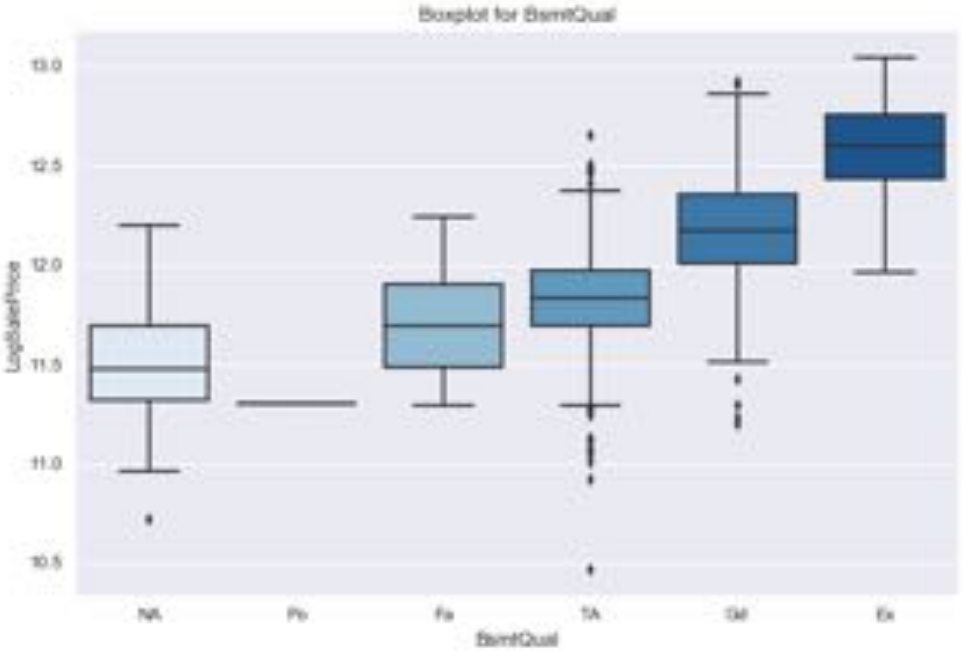
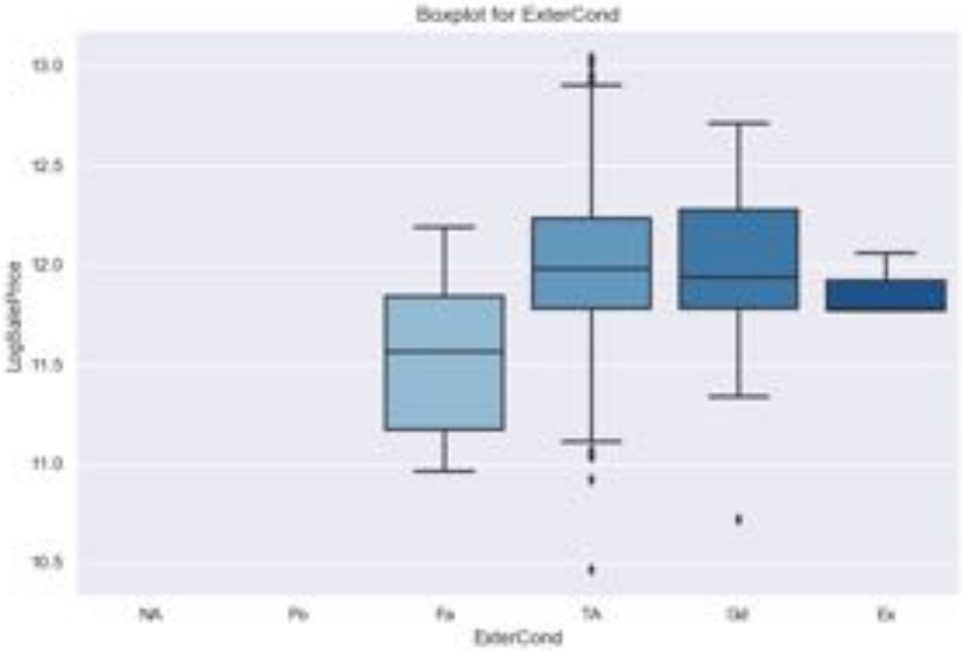
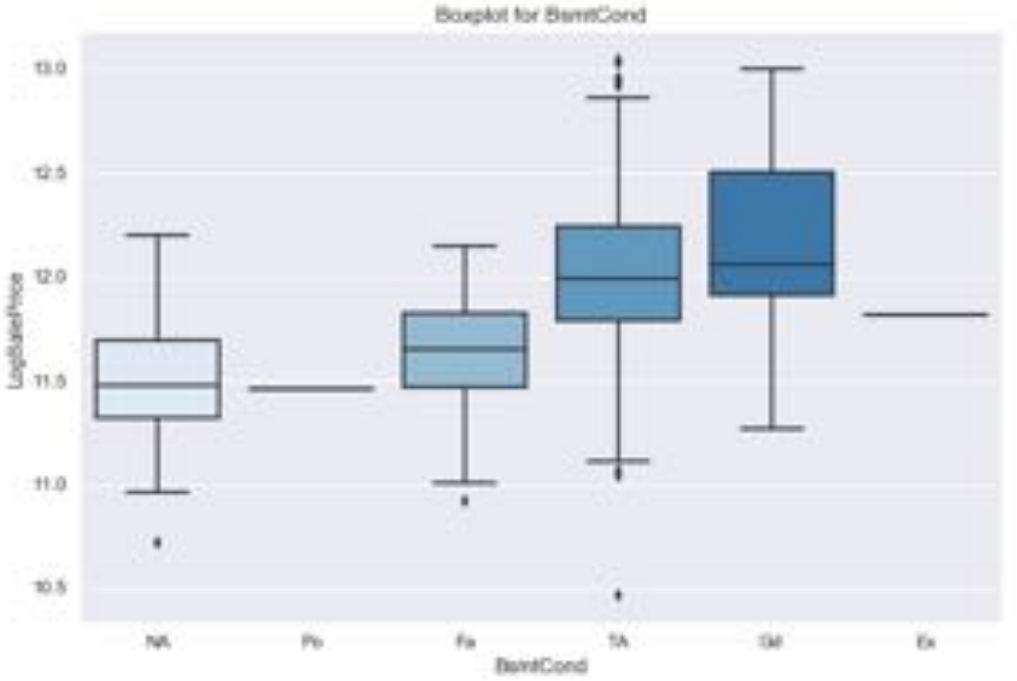
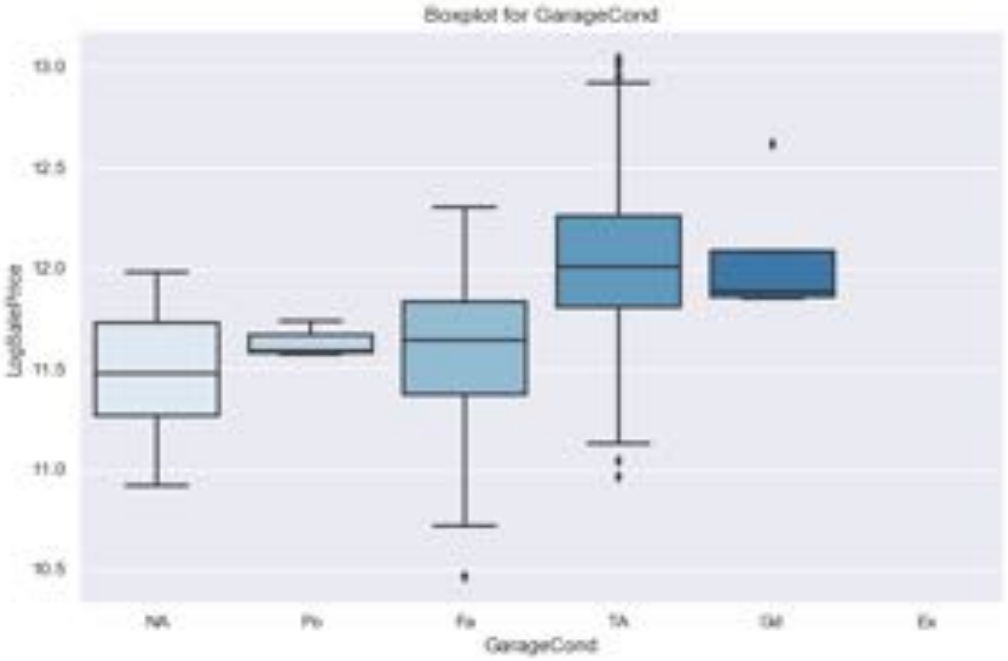


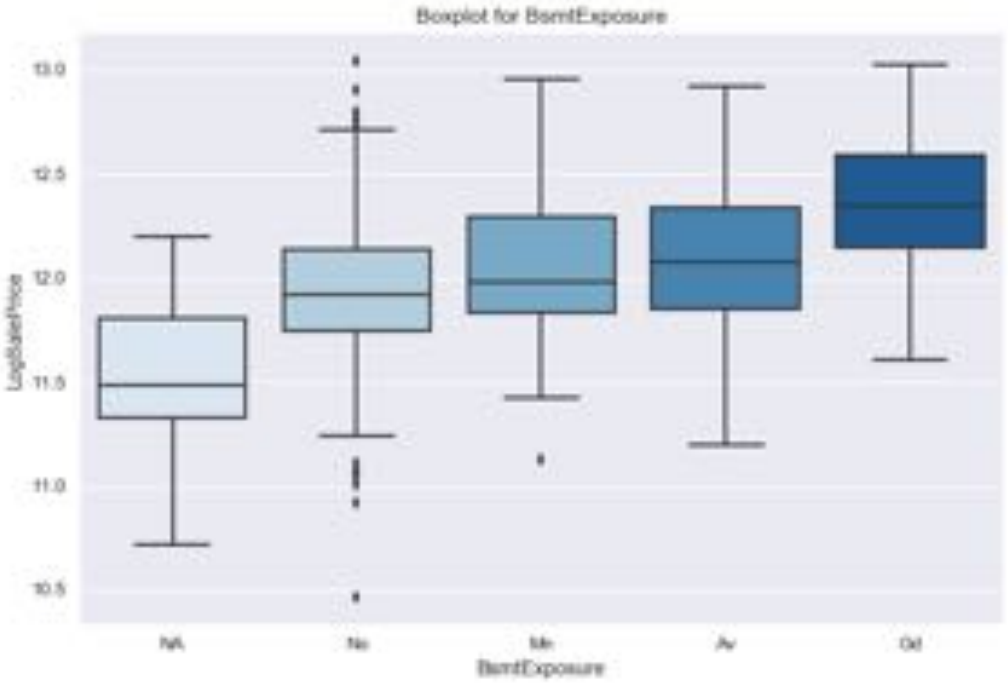
Figure15-25. Boxplots for categorical variables which indicate a linear trend

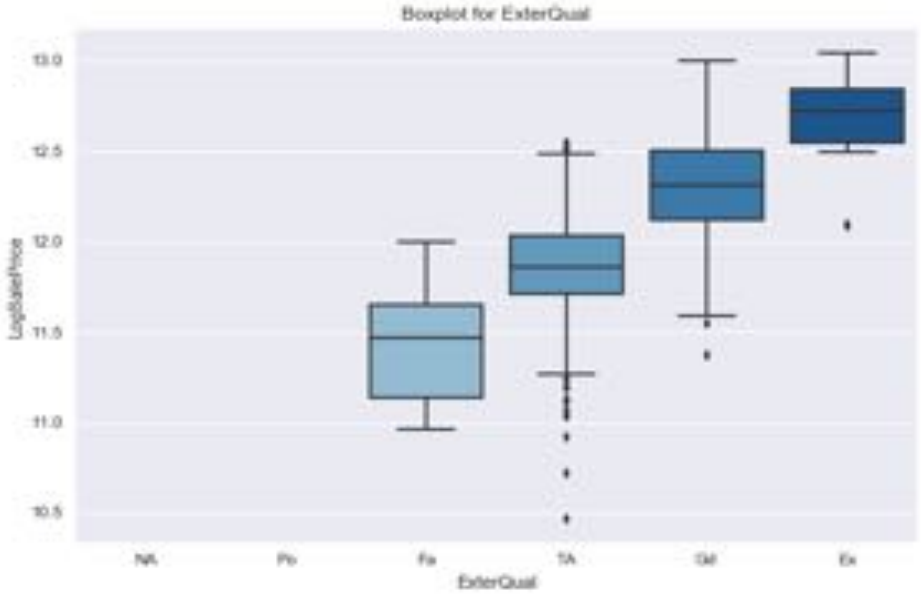


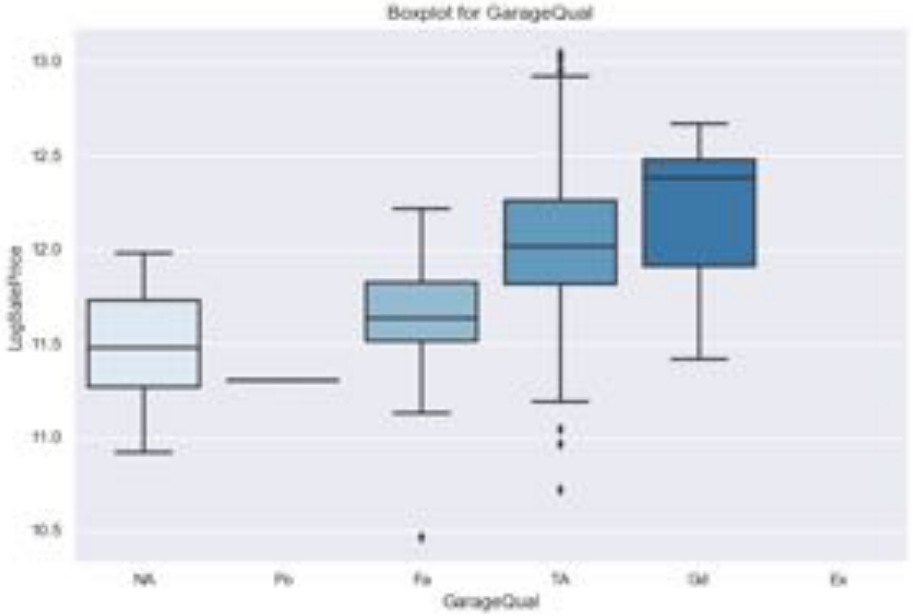

Boxplot for ExterCond
LogSalesPrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
ExterCond

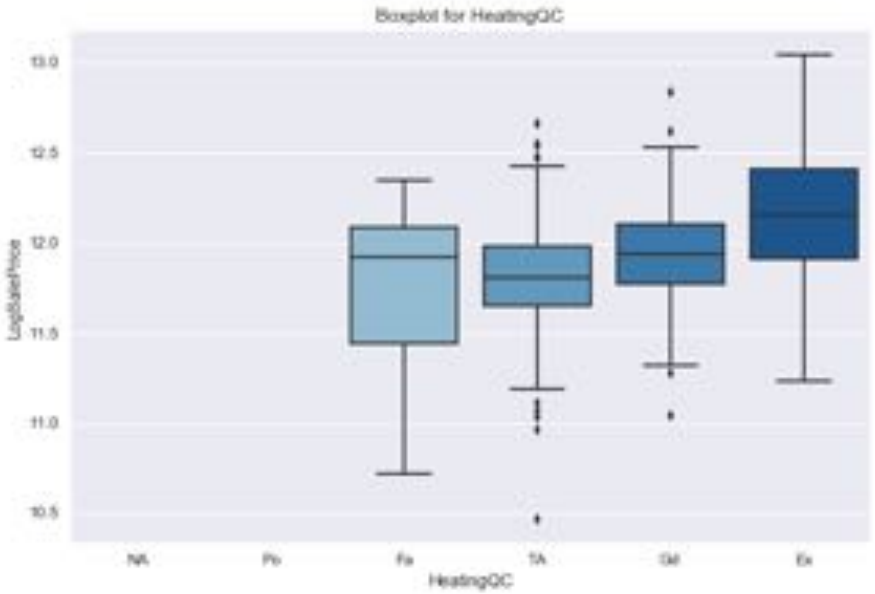

Boxplot for BsmtCond
LogSalesPrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
BsmtCond

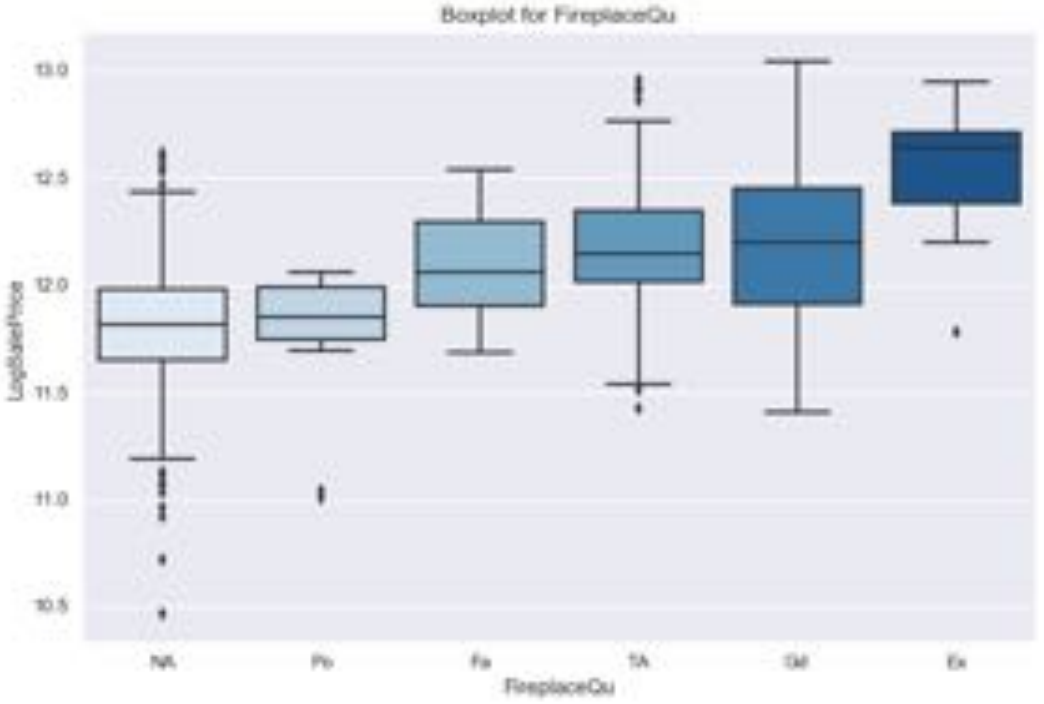

Boxplot for GarageCond
13.0
12.5
12.0
11.5
11.0
10.5
LogSalePrice
NA
Po
Fa
TA
Gd
Ex
GarageCond

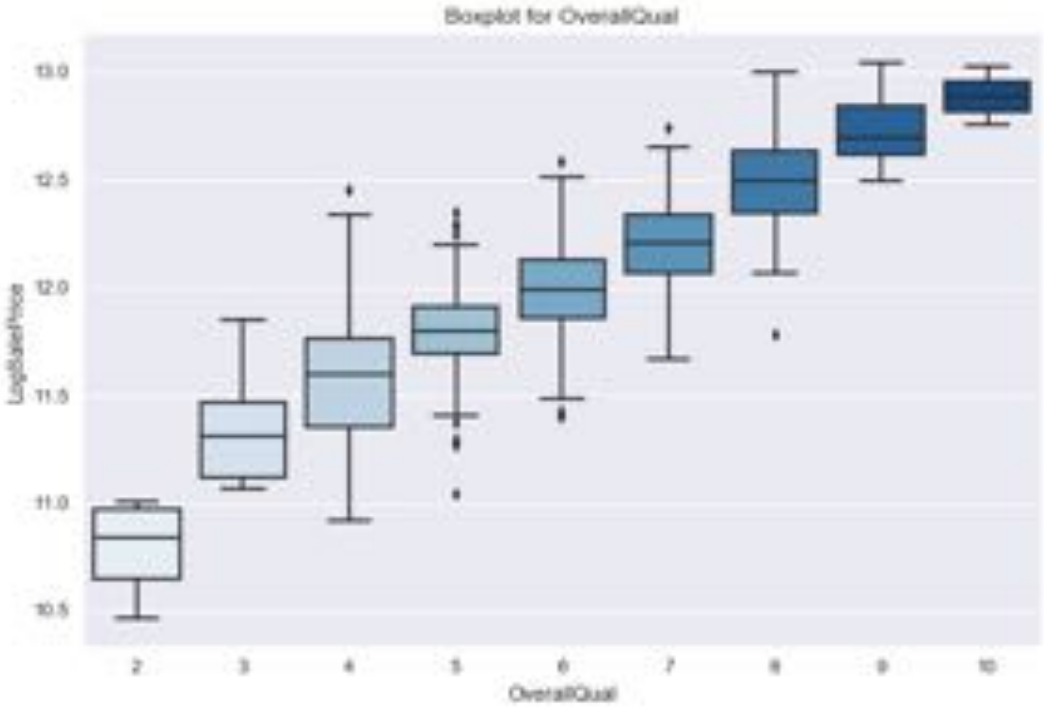

Boxplot for BsmtExposure
13.0
12.5
12.0
11.5
11.0
10.5
LogSalePrice
NA
No
Mn
Av
Gd
BsmtExposure


Boxplot for ExterQual
LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
ExterQual


Boxplot for GarageQual
LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
GarageQual


Boxplot for HeatingQC
LogSalePrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
HeatingQC


Boxplot for FireplaceQu
LogSalesPrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
FireplaceQu


Boxplot for OverallQual
LogSalesPrice
13.0
12.5
12.0
11.5
11.0
10.5
2
3
4
5
6
7
8
9
10
OverallQual

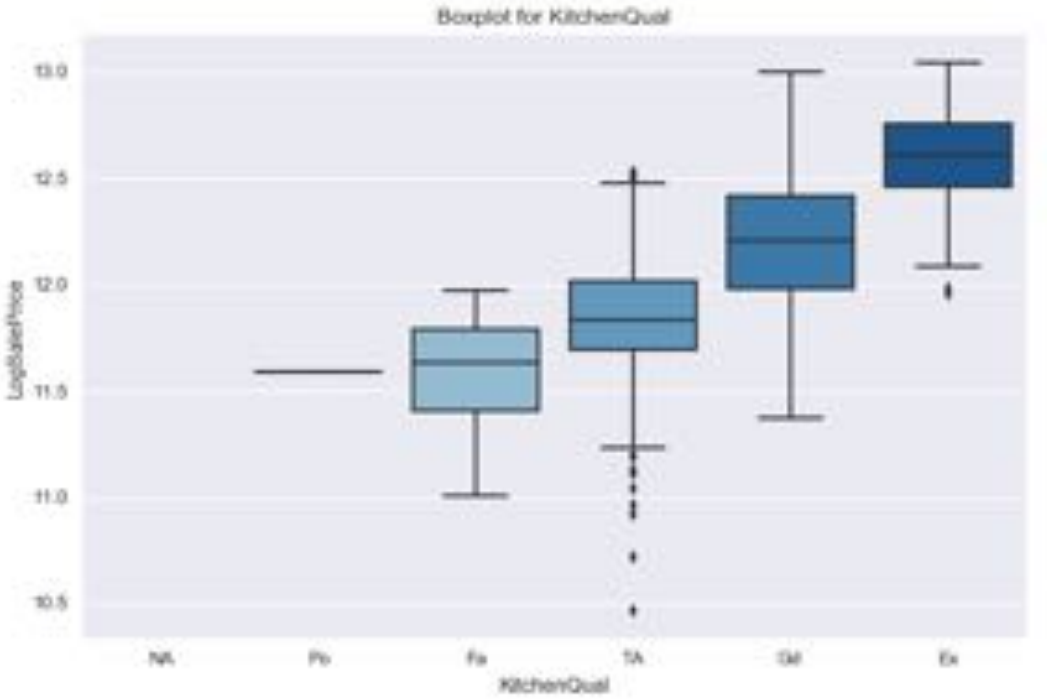
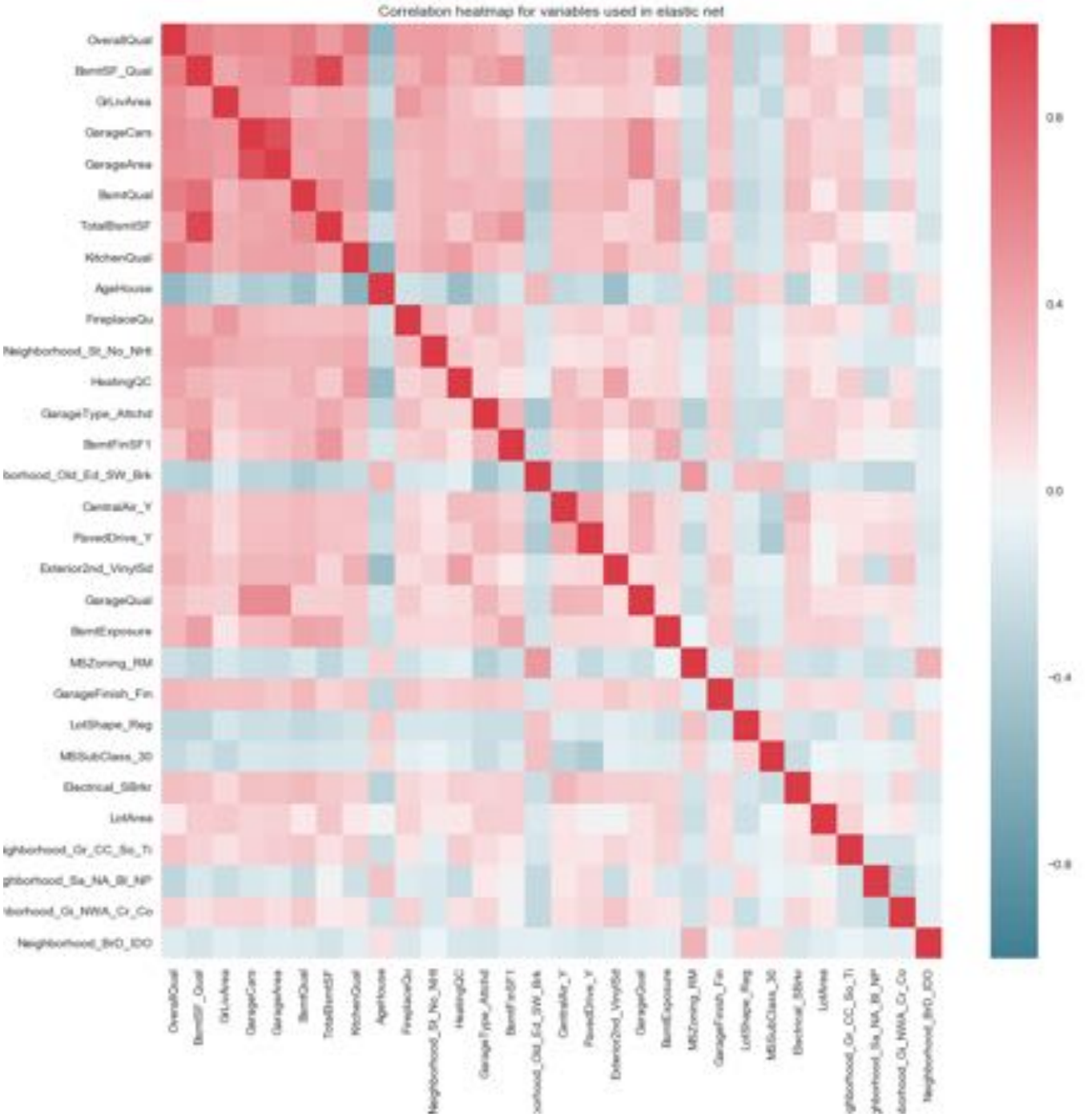

Boxplot for KitchenQual
LogSalesPrice
13.0
12.5
12.0
11.5
11.0
10.5
NA
Po
Fa
TA
Gd
Ex
KitchenQual

Figure 26 and 27. Correlation heatmap for assumption checking (Enet and Ridge)



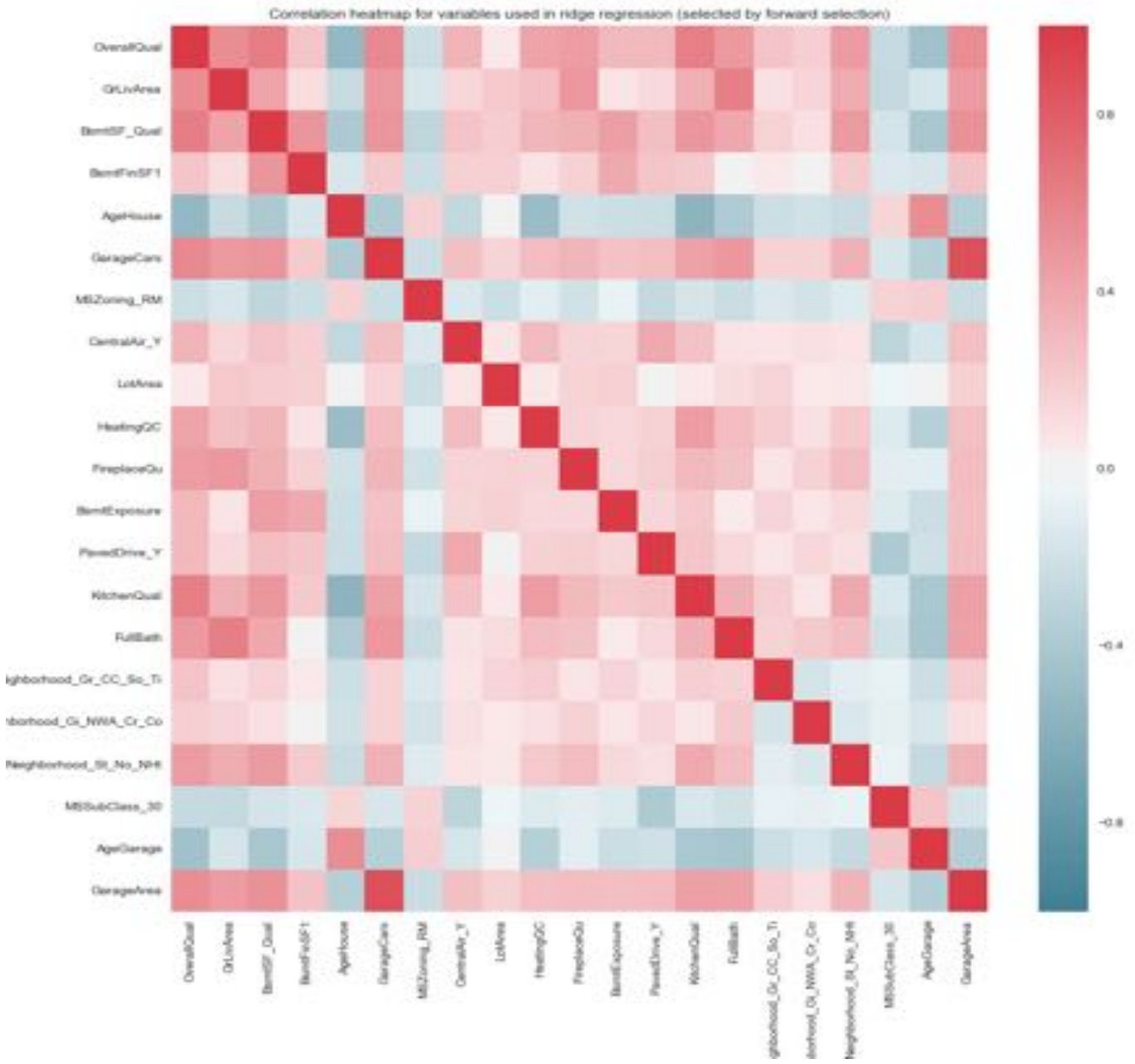

Correlation heatmap for variables used in ridge regression (selected by forward selection)
OverallQual
GrLivArea
BsmtSF_Qual
BsmtFinSF1
AgeHouse
GarageCars
MSZoning_RM
CentralAir_Y
LotArea
HeatingQC
FireplaceQu
BsmtExposure
PavedDrive_Y
KitchenQual
FullBath
AgeGarage
GarageArea
0.8
0.4
0.0
−0.4
−0.8

Figure 28. Correlation heatmap for assumption checking (OLS)