Open Access Library Journal
2022, Volume 9, e8346
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables

Azad Abdulhafedh

University of Missouri, State of Missouri, USA
Email: dr.azad.s.a@gmail.com

How to cite this paper: Abdulhafedh, A. (2022) Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables. *Open Access Library Journal*, **9**: e8346.
https://doi.org/10.4236/oalib.1108346

Received: January 3, 2022
Accepted: January 26, 2022
Published: January 29, 2022

Abstract

This paper uses a multiple linear regression analysis to predict the final price of a house in a big real estate dataset. The data describes the sale of individual properties, various features, and details of each home in Ames, Iowa, USA from 2006 to 2010. The dataset comprises of 80 explanatory variables which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. The goal was to use the training data to predict the sale prices of the houses in the testing data. The most important predictors were determined by random forest and kept in the analysis. The highly correlated predictors were dropped from the dataset. All assumptions of the linear regression were checked, and an optimal final predictive model was achieved by keeping the most influential predictors only. The model accuracy assessments produced very good results with an adjusted R-squared value of 0.9283, a residual standard error (RSE) of 0.094, and a root squared mean error (RSME) of 0.12792. In addition, the prediction error (Mean Squared Error, MSE) of the final model was found to be very small (12%) by applying different cross validation techniques, including the validation set approach, the K-fold approach and the Leave-One-Out-Cross Validation (LOOCV) approach. Results show that multiple linear regression can precisely predict the house prices with big dataset and large number of both categorical and numerical predictors.

Subject Areas

Applied Statistical Mathematics, Civil Engineering

Keywords

Multiple Linear Regression, Ames House Price Prediction, RSE, RSME, MSE, K-Fold, LOOCV

1. Introduction

Multiple Linear Regression is a statistical technique that is designed to explore the relationship between two or more variables. The variable we want to predict is called the dependent variable (or the response variable, or the outcome, or target). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictors, explanatory variables, or regressors). The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or outcome variable. For example, a real estate agent might record for each listing the size of the house, the number of bedrooms, the average income in the respective neighborhood according to data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses, it would be interesting to see whether and how these measures relate to the price for which a house is sold. For instance, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). You may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics [1]-[7]. This technique is also useful in identifying important factors that will impact a dependent variable, and the nature of the relationship between each of the factors and the dependent variable [1] [2] [3] [4]-[11].

1.1. The Mathematical Expression of the Multiple Linear Regression

The formula or mathematical expression for a multiple linear regression is [1] [2] [4] [12]-[18]:

$$y = \beta_0 + \beta_0 X_1 + \cdots + \beta_n X_n + \varepsilon$$

where:

- y = the predicted value of the dependent variable (or response variable),
- β_0 = the y-intercept (value of y when all other parameters are set to 0),
- $\beta_1 X_1$ = the regression coefficient (β_1) of the first independent variable (X_1) (*i.e.*, the effect that increasing the value of the independent variable has on the predicted y value),
- … = do the same for however many independent variables you are testing,
- $\beta_n X_n$ = the regression coefficient of the last independent variable,
- ε = model error (*i.e.*, how much variation there is in our estimate of y).

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t*-statistic of the overall model.
- The associated p-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

1.2. Assumptions of Multiple Linear Regression

Multiple linear regression makes the following assumptions, which should be hold [1] [2] [3] [4] [6] [9] [12] [14] [15]:

1) Homogeneity of variance (homoscedasticity): the size of the error in the prediction doesn't change significantly across the values of the independent variable.

2) Independence of observations: the observations in the dataset are collected using statistically valid methods, and there are no hidden relationships among variables.

3) Collinearity: In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (>~0.5), then only one of them should be used in the regression model.

4) Normality: The data follows a normal distribution.

5) Linearity: the line of best fit through the data points is a straight line, rather than a curve.

1.3. Data

Ames is a city in the state of Iowa in the United States. The Ames housing dataset examines features of houses sold in Ames during the 2006-10 timeframe. Ames data were prepared by Ames Housing Authority, which is a public housing agency that serves the city of Ames, Iowa, US. It helps provide safe rental housing for eligible low-income families, the elderly, and persons with disabilities. The dataset includes 80 assessment parameters which describes every aspect of residential homes in Ames. These variables focus on the quality and quantity of the physical attributes of a property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property. The explanatory variables consist of 23 nominal, 23 ordinal, 14 discrete, and 20 continuous. The 20 continuous variables are related to measurements of area dimensions for each observation, including the sizes of lots, rooms, porches, and garages. The 14 discrete variables are related to the number of bedrooms, bathrooms, kitchens, etc. There are some geographic categorical variables that are related to profiling properties and the neighborhood characteristics. The remaining nominal variables identify characteristics of the property and dwelling type/structure. The ordinal variables are related to the rankings of the quality/condition of rooms and lot characteristics [19]. The Ames Housing dataset was obtained from (http://www.kaggle.com) that contains both training dataset and testing dataset. The two datasets are representation of whole data spilt into 50% - 50% to train and test sets. Test set contains all the predictor variables in train set excluding the response variable "SalePrice". The total number of observations in the Train file is 1460 and in the Test file, is 1459. A total of three data files were used in the analysis, namely (response_id.csv, train_reg_features.csv, and test_reg_features.csv), and one data description file (data_description.txt) with full description of each house feature. The objective of the analysis is to build an optimal linear regression model to predict the final price of each house using the features in the test data. The data description of the response variable and the first 11 predictors are as follows [19]:

- SalePrice: The property's sale price in dollars. This is the response variable or target,
- MSSubClass: The building class,
- MSZoning: The general zoning classification,
- LotFrontage: Linear feet of street connected to property,
- LotArea: Lot size in square feet,
- Street: Type of road access,
- Alley: Type of alley access,
- LotShape: General shape of property,
- LandContour: Flatness of the property,
- Utilities: Type of utilities available,
- Neighborhood: Physical locations within Ames city limits.
- Condition1: Proximity to main road or railroad.

2. Methodology

The analysis was performed using the R software. The following steps were included in the methodology [1] [2] [4] [9] [12] [14] [15] [16] [17]:

- Inspect, and clean the datasets from outliers,
- Combine the training and testing data files,
- Explore the data and produce the summary statistics,
- Check for the completeness of data and missing values,
- Conduct Exploratory Data Analysis (EDA),
- Remove nonimportant features based on EDA,
- Check multicollinearity among different features,
- Remove highly correlated features,
- Remove near zero variance predictors,
- Select a short list of predictor features for prediction of the house price,
- Develop a multiple linear regression model for the prediction,
- Check the assumptions of the model and perform necessary diagnosis,
- Calculate prediction error of the optimal developed model by cross validation approaches,
- Predict the house price using the test data.

These analysis steps are summarized as follows:

Step 1: Inspection of the data was conducted to know how many observations there in the train and test files and how many features are presented in the dataset. It was found that the total number of observations in the Train file is 1460 and in the Test file, is 1459. The total explanatory variables (also called features) are 80 features including the ID of each feature. The datasets were cleaned from undesirable outliers as well.

Step 2: Combining the Train and Test Datasets for conducting the analysis. It was found that there are 36 numerical features and 44 categorical features in the combined dataset.

Step 3: Explore the combined dataset and produce the summary statistics. The response variable and the first 10 features are shown in Table 1.

Step 4: Examine Data Completeness and determine the percent of the Null/or Missing Values. It was found that 11 features have null or missing values, as shown in Table 2.

Step 5: Remove features with more than 40% Missing/or Null Values, so that it does not affect the integrity of the analysis [20] [21] [22]. It was found that the total number of features with more than 40% null values were only 5 features, namely (PoolQC", "MiscFeature", "Alley", "Fence", "FireplaceQu), as shown in Table 2.

Table 1. Data exploration of the response variable and the first 10 features in the dataset.

SalePrice	:	integer 126000 139500 124900 114000 227000 198500...
MSSubClass	:	integer 30 120 30 70 60 85 20 20 20 180 ...
MSZoning	:	Factor w/ 7 levels "A (agr)","C (all)",..: 6 6 2 6 ..
LotFrontage	:	integer NA 42 60 80 70 64 60 53 74 35 ...
LotArea	:	integer 7890 4235 6060 8146 8400 7301 6000 3710 12395
Street	:	Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 ...
Alley	:	Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA ...
Lot Shape	:	Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 4 4 4 ..
Land Contour	:	Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 ..
Utilities	:	Factor w/ 3 levels "AllPub","NoSeWa",..: 1 1 1 1 1 ..
Neighborhood	:	Factor w/ 28 levels "Blmngtn","Blueste",..: 26 8 1 ..

Table 2. The null values of some features in the dataset.

	Null Count	%Null
PoolQC	2909	1.00
MiscFeature	2814	0.96
Alley	2721	0.93
Fence	2348	0.80
FireplaceQu	1420	0.49
LotFrontage	486	0.17
GarageYrBlt	159	0.05
GarageFinish	159	0.05
GarageQual	159	0.05
GarageCond	159	0.05
GarageType	157	0.05

Step 6: Handle the Missing/Null values for Numeric Features with less than 40% for continuous variables by replacing the null values with the mean value of the feature. One of the data imputation techniques for continuous variables is mean imputation in which the missing values are replaced with the mean value of the entire feature column [20] [21] [22].

Step 7: Handle the Missing/Null values for Categorical Features with less than 40% by replacing the null values with the mode of the feature. Another data imputation technique for categorical variables is mode imputation in which the missing values are replaced with the mode value or most frequent value of the entire feature column [20] [21] [22].

Step 8: Conduct Exploratory Data Analysis (EDA) for the response variable and the features as well. The histogram of the response is shown in **Figure 1**.

It can be seen from the histogram of the response variable "SalePrice" that this variable is skewed to the left. Therefore, a Log transformation is necessary to make it normally distributed, which will provide better model performance. However, the normality assumption is not necessary for the explanatory variables (features), because multiple linear regression can handle any type of features, even the skewed features [1] [9] [14] [17]. Therefore, the normality assumption should hold for the response variable only. Applying a log transformation to the response variable resulted in the histogram shown in **Figure 2**.

More Histograms of some explanatory variables are shown in **Figure 3**.

More EDA plots for the features in the dataset were performed, including the following:

Boxplot of the SalePrice vs Neighborhood

Location of the house plays an important role in determining the price of the house. The boxplot in **Figure 4** shows the price distribution in multiples of 1000 in the various neighborhoods of Ames, Iowa. MeadowV is the least expensive locality while StoneBr is the most expensive. Also, the plot indicates that

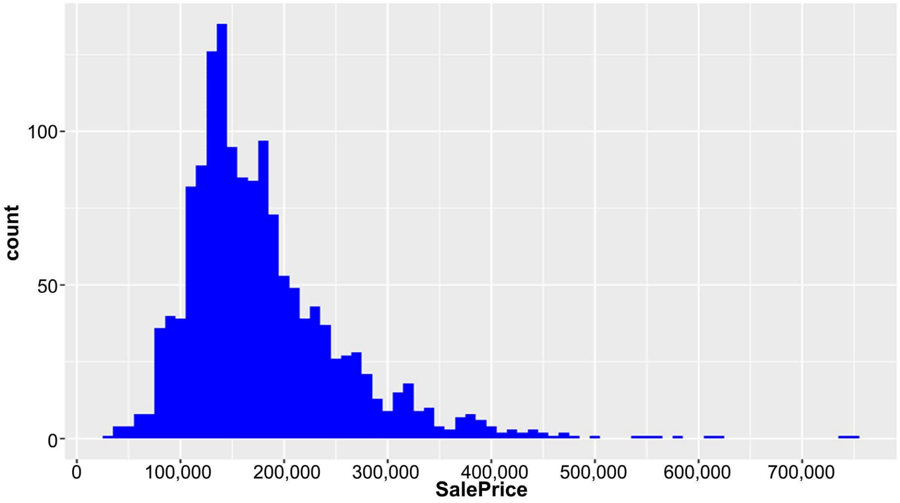

Figure 1. Distribution of the Response Variable "SalePrice".

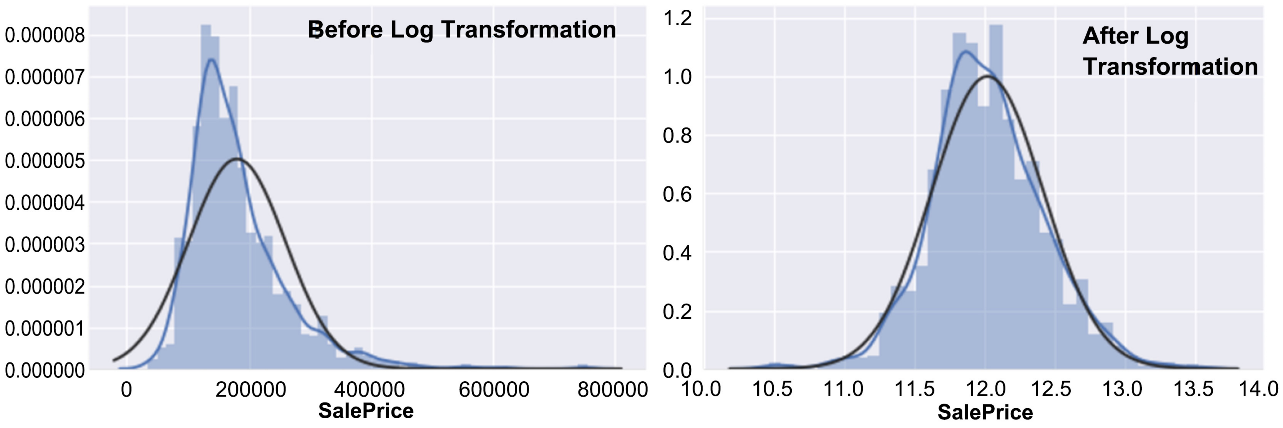

Figure 2. The distribution of the response variable before and after applying a log transformation.

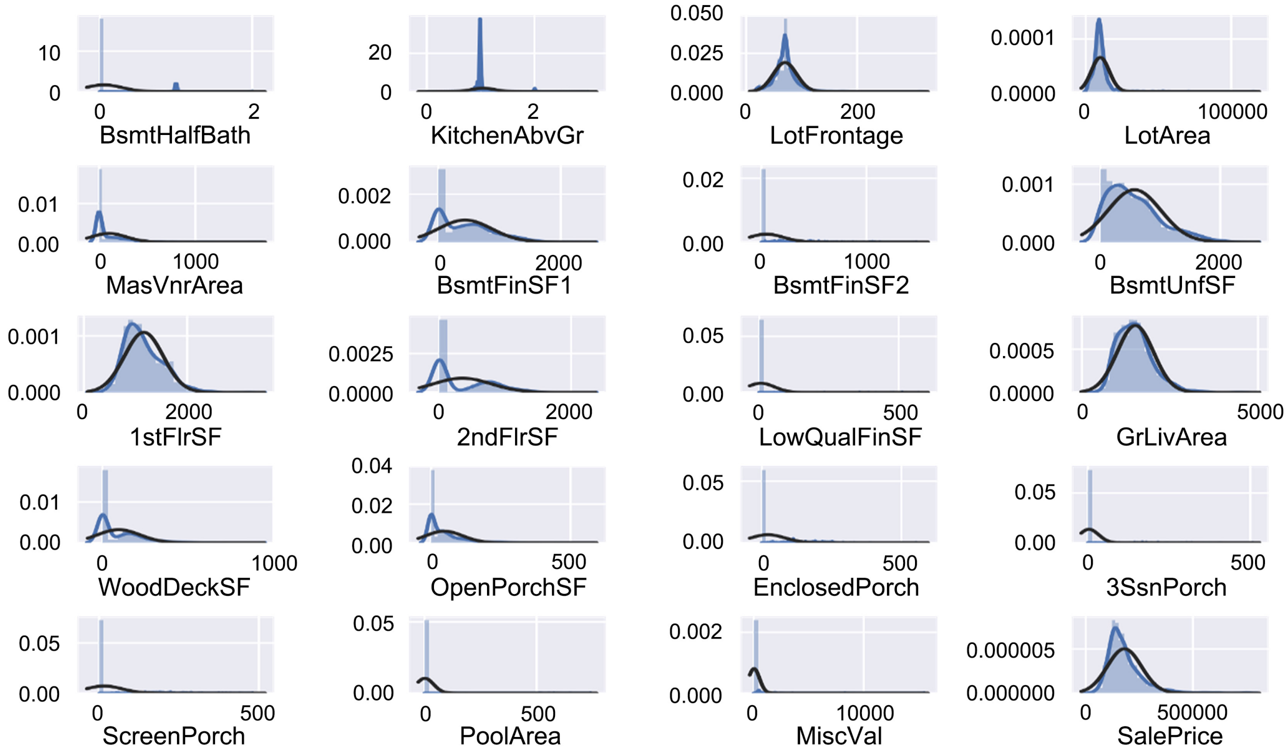

Figure 3. Histograms of some features in the dataset.

the variation in the house price is also quite significant in the StoneBr, NridgHt localities though they are one of the most expensive localities.

Paired Association of SalePrice vs. some features is shown in Figure 5 and Figure 6. It can be seen from Figure 5 that the house price has a negative relationship with the Age of the house. The number of bedrooms above ground and lot configuration does not seem to have a very strong effect on the house price. It can be seen from Figure 6 that the kitchen, basement, and external qualities have medium association with the house price.

Step 9: Remove non-important features from the dataset based on the EDA, and considering the relationship shown above in the boxplots and association

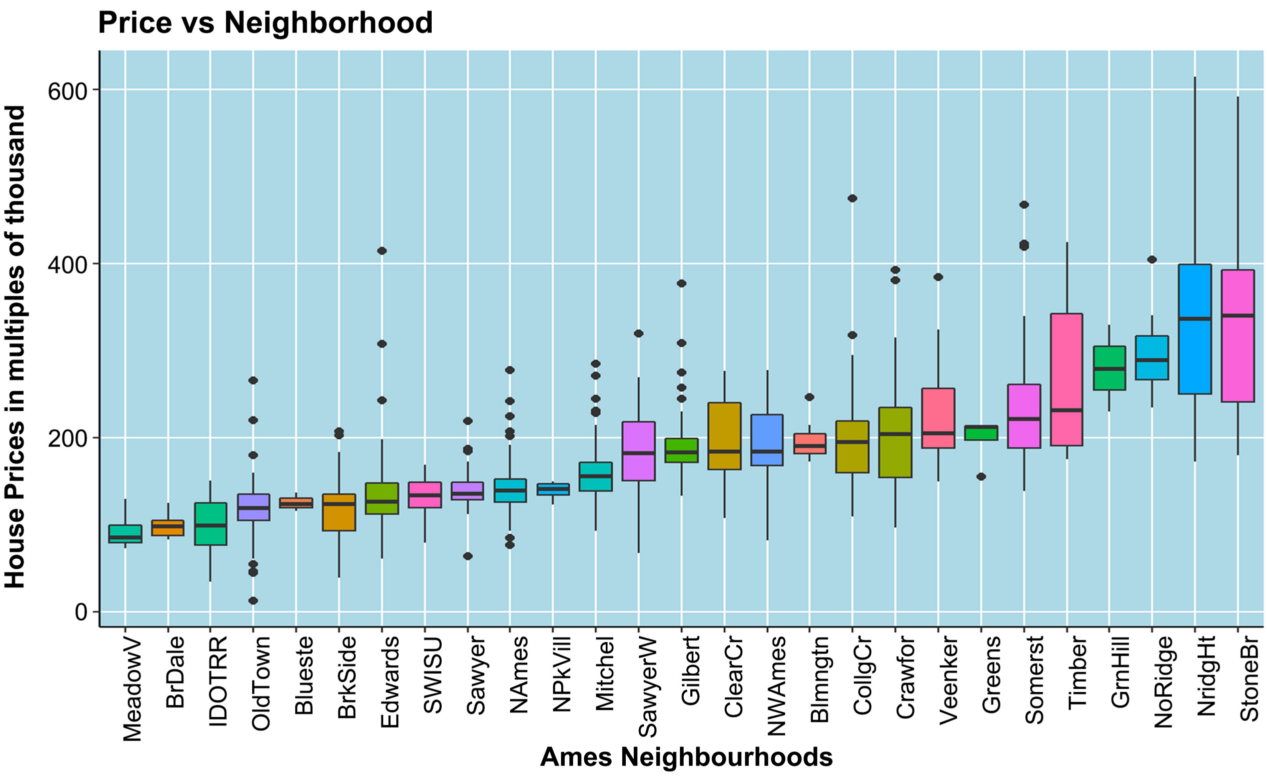

Figure 4. Boxplot of the response "SalePrice vs Neighborhoods".

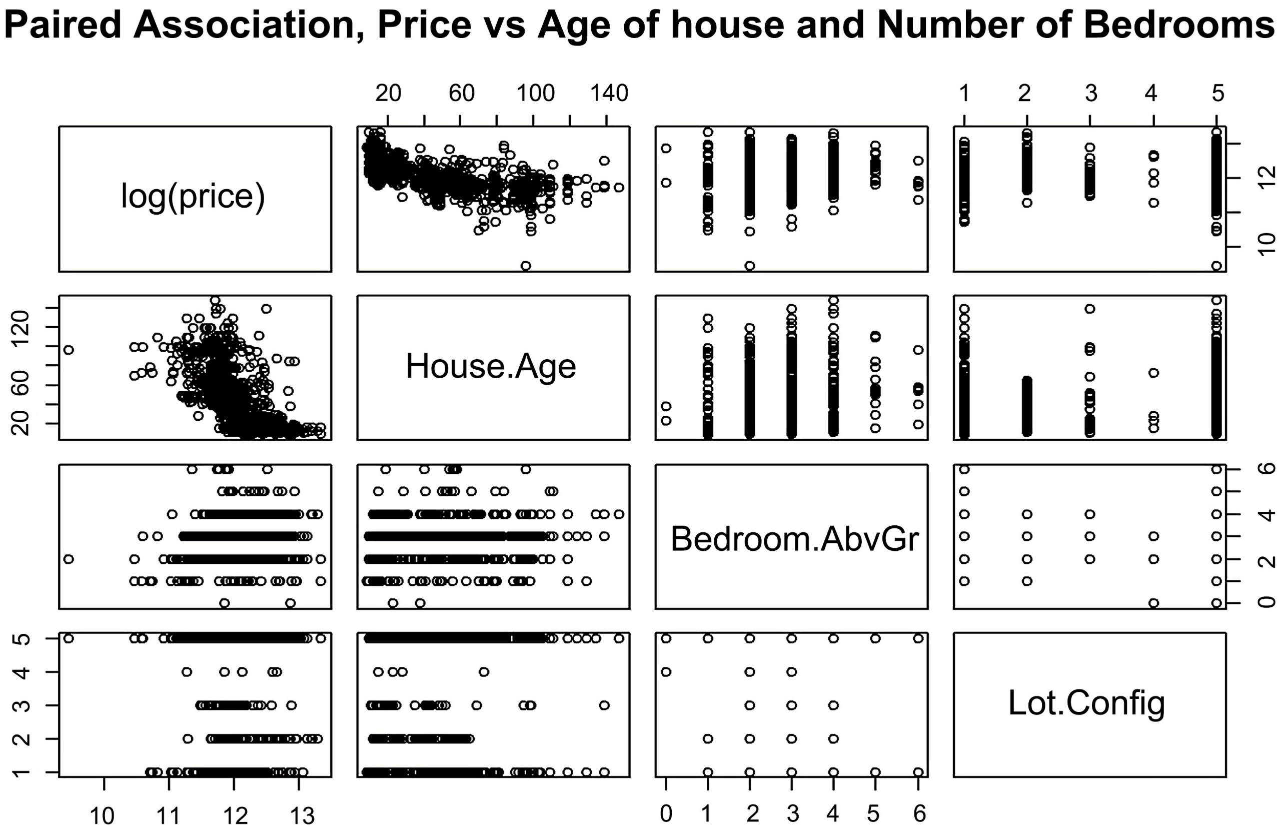

Figure 5. Paired Association of the response vs. house age, bedrooms above ground level, and lot.

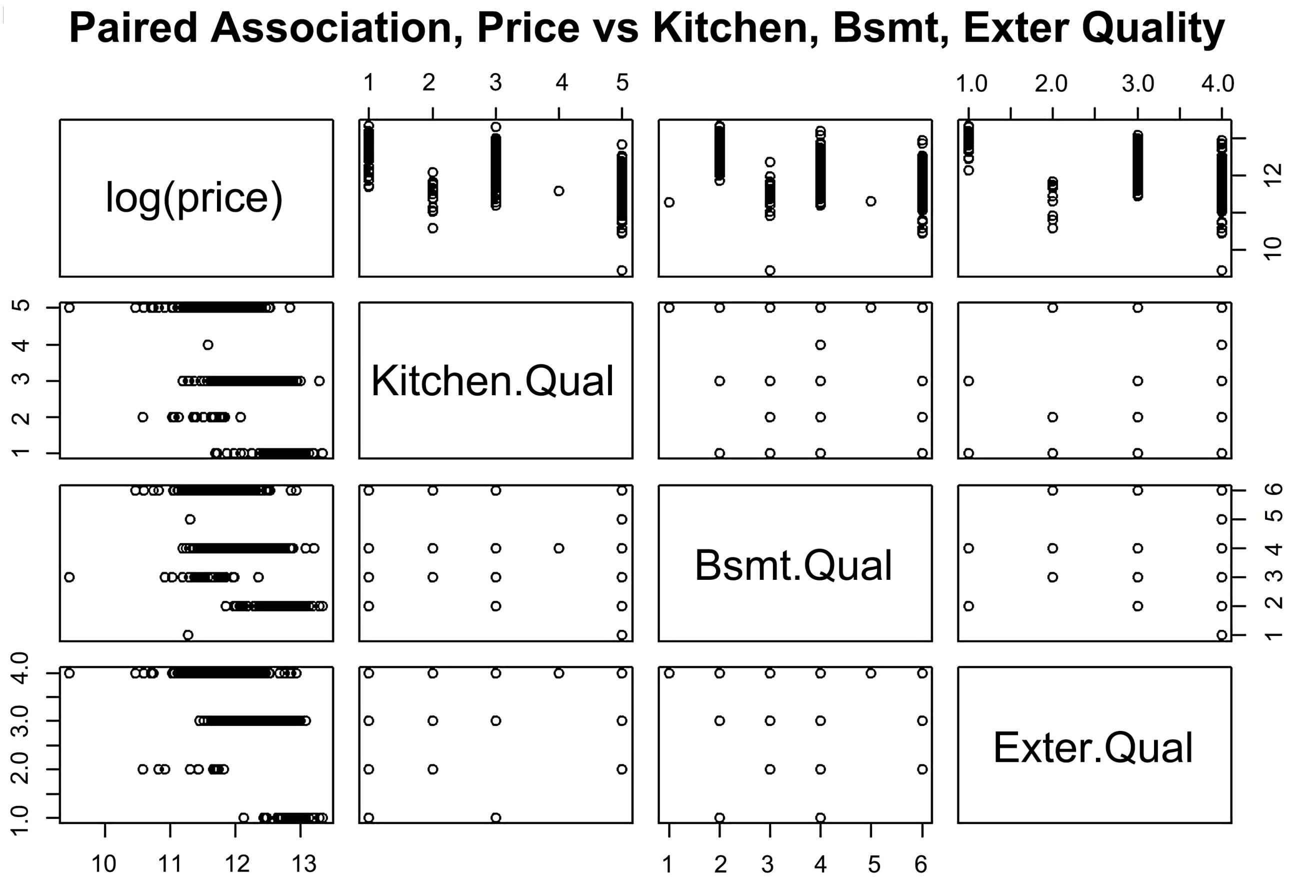

Figure 6. Paired Association of the response vs. Kitchen quality, basement quality, and external quality.

plots [1] [2] [3] [9]. The following 17 features were removal from the dataset, as they have very little effect on the response "SalePrice":

1) BsmtHalfBath,
2) KitchenAbvGr,
3) MaxVnrArea,
4) BsmtFinSF1,
5) BsmtFinSF2,
6) 2ndFlrSF,
7) LowQualFinSF,
8) WoodDeckSF,
9) OpenPorchSF,
10) EnclosedPorch,
11) 3SsnPorch,
12) ScreenPorch,
13) PoolArea,
14) MiscVal,
15) Utilities,
16) Condition2,
17) RootMatl.

Step 10: Check the correlation of Features with the Response "SalePrice". The correlation matrix is shown in Figure 7.

It can be seen from Figure 7 that some numeric variables have low correlation with SalePrice, such as LotFrontage (0.33) and BsmtFullBath (0.28), while other variables show high correlation with SalePrice, such as OverallQual (0.79) and GarageCars (0.64). Figure 8 shows which variables have a high correlation (>0.5) with the response variable SalePrice that should be removed from the dataset.

It can be seen from Figure 8 that there are 10 variables with a correlation of at least 0.5 with the response "SalePrice", which should be removed as follows:

- OverallQual (correlation: 0.79),
- GrLivArea (0.71),
- GarageCars (0.64),
- GarageArea (0.62),
- TotalBsmtSf (0.61),
- X1stFlrSF (0.61),
- FullBath (0.56),
- TotRmsAbvGrd (0.53),
- YearBuilt (0.52),
- YearRemodAdd (0.51).

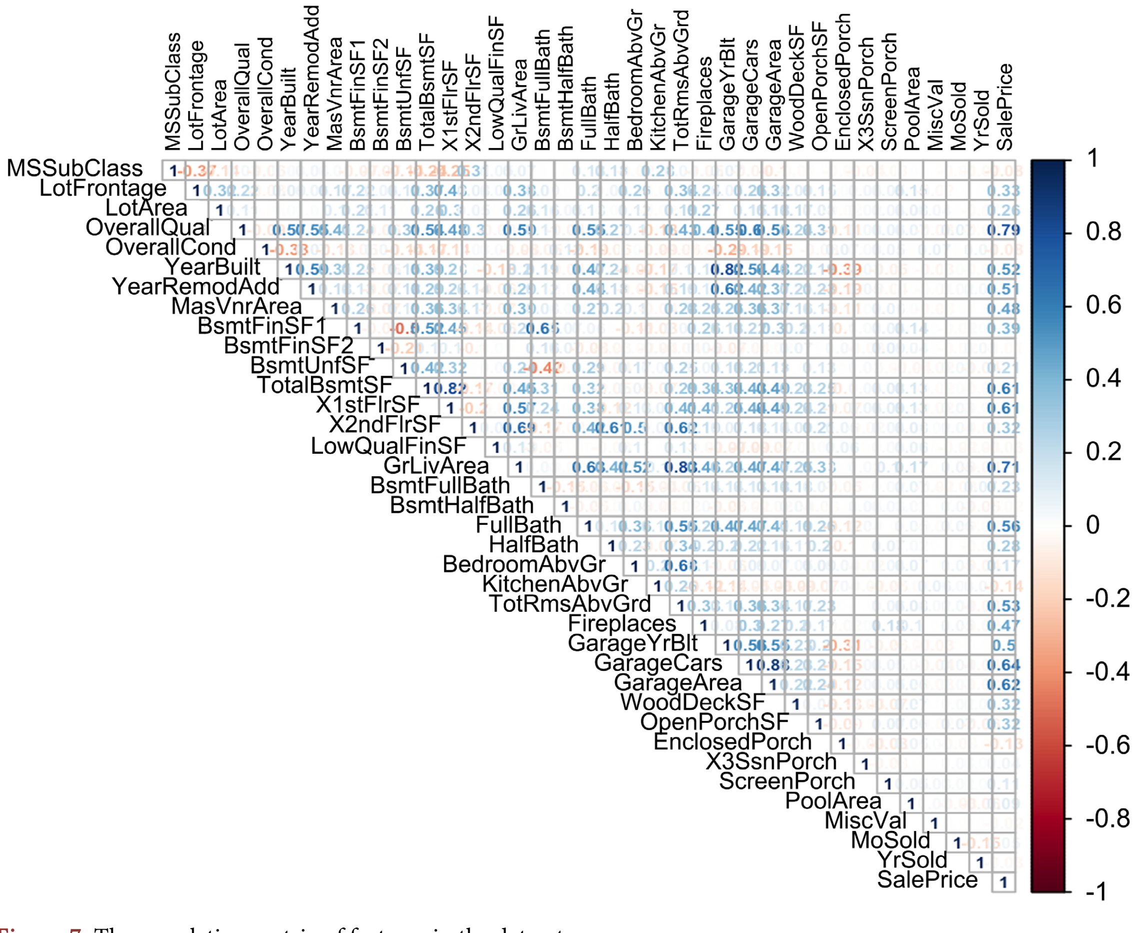

Figure 7. The correlation matrix of features in the dataset.

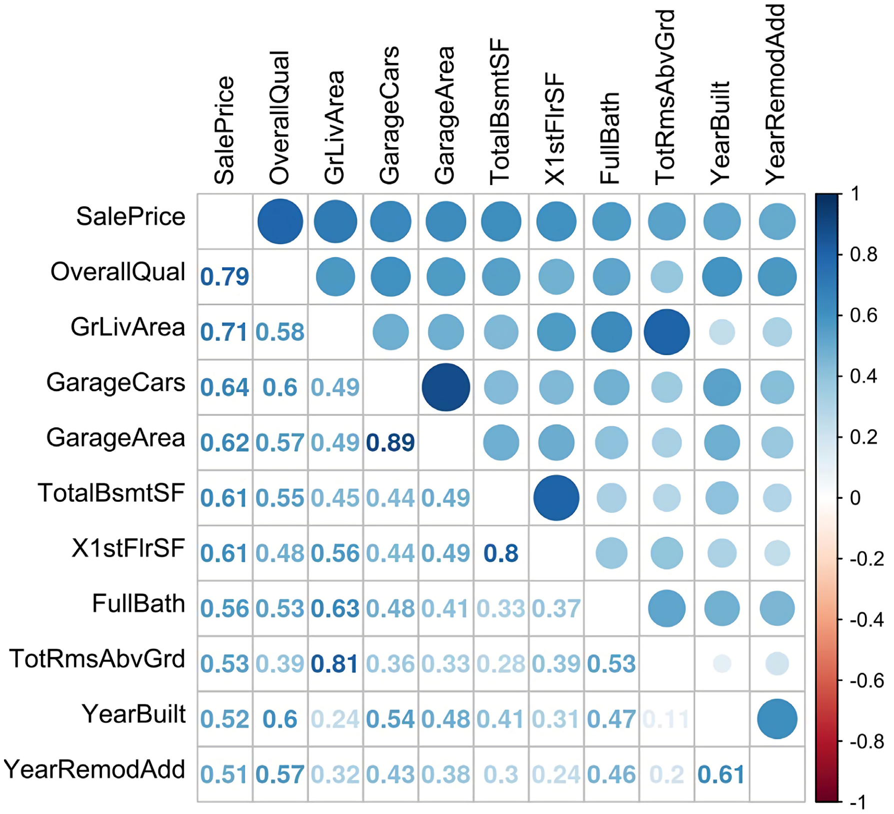

Figure 8. High correlation values of some features with the response "SalePrice".

Figure 8 also shows the correlation between the numeric variables themselves. For example: the correlation between GarageCars and GarageArea is very high (0.89), and both have similar (high) correlations with SalePrice.

Step 11: Find Feature Importance by using the Random Forest (RF) algorithm in R to get a clear idea of the most important features that should be kept in the dataset. **Figure 9** shows the resulted matrix of the 20 most important features using the RF.

Step 12: Remove features with Near Zero Variance as they violate the assumptions of multiple linear regression. Constant and almost constant predictors across samples (called zero and near-zero variance predictors, respectively) happens widely when we usually break a categorical variable with many categories into several dummy variables. Hence, when one of the categories has zero observations, it becomes a dummy variable full of zeros [23] [24] [25] [26]. For example, for 1000 samples, near zero variance has two distinct values and 999 of them are a single value. In the above example, the frequency ratio is 999 and the unique value percentage is 0.0001. Using the function "nearZeroVar" in R, it was found that 8 features have near zero variance, which should be removed from the dataset. These features are:

1) BsmtFinSF1,

2) X1stFlrSF,

3) GarageCars,

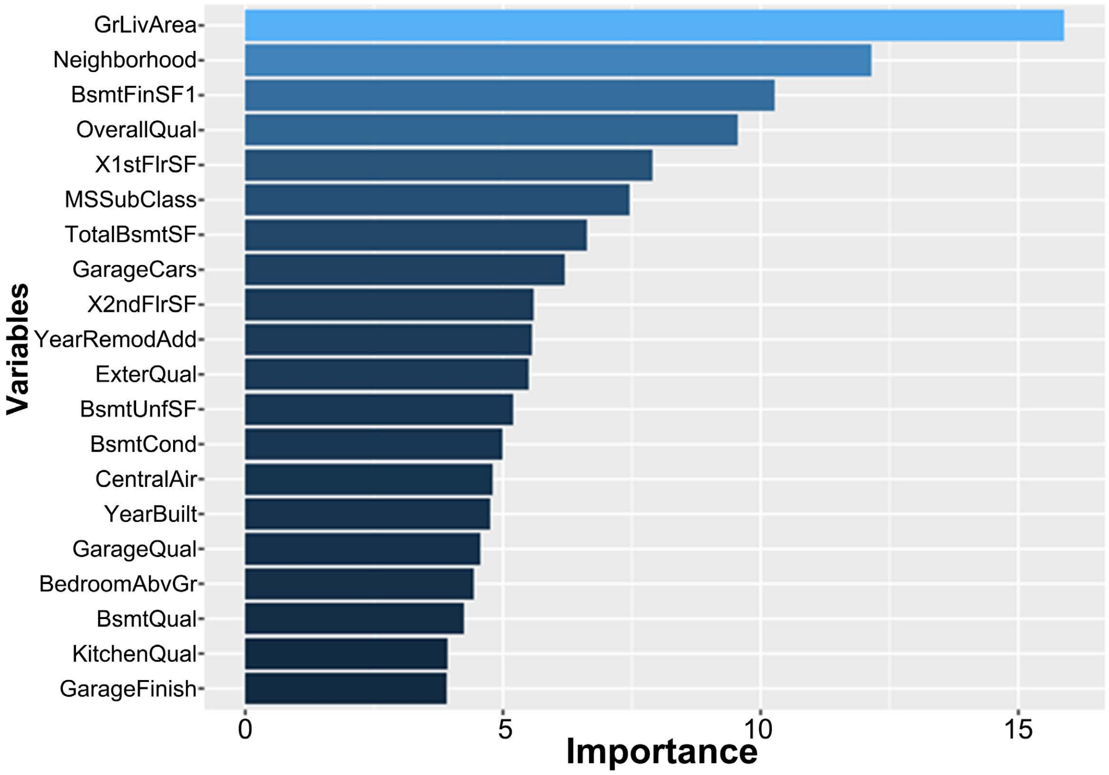

Figure 9. The 20 most important features in the dataset using random forest.

4) X2ndFlrsSF,

5) BsmtUnfSF,

6) CentralAir,

7) BedroomAbvGr,

8) GarageFinish.

Step 13: Split the whole dataset back to Training and Testing files (50% each randomly).

Step 14: Feature selection for the Predictive Model. Feature selection has the following benefits [1] [2] [4] [12] [15]:

- It reduces the variance of the model, and therefore overfitting,
- It reduces the complexity of a model and makes it easier to interpret,
- It improves the accuracy of a model if the right subset is chosen,
- it reduces the computational cost and time of training a model.

Based on the list of important variables above, the following explanatory variables from the training file were selected and an initial multiple linear regression model for predicting the "SalePrice" was created. The following 12 features were selected for the initial regression model:

1) GrLivArea,

2) Neighborhood,

3) OverallQual,

4) MSSubClass,

5) TotalBsmtSF,

6) YearRemodAdd,
7) ExterQual,
8) BsmtCond,
9) FullBath,
10) GarageQual,
11) LotArea,
12) KitchenQual.

3. Results and Discussion

Using the selected 12 features in step 14, an initial Multiple Linear Regression Model was created using the Ames training file. The Ordinary Least Squared (OLS) procedure was used in developing the initial model. The initial model summary is shown in Table 3.

3.1. The Interpretation of the Initial Model

When examining the F-statistic and the associated p-value, at the bottom of model summary, it can be seen that the p-value of the F-statistic is 2.2e−16, which is highly significant [1] [2] [6] [9].

Table 3. The initial multiple regression model summary.

Residuals:				
Min	1Q	Median	3Q	Max
−0.84933	−0.06044	0.00587	0.06465	0.45309
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.601036	0.174386	43.587	<2e−16***
(GrLivArea)	0.140648	0.011665	12.057	<2e−16***
(OverallQual)	0.477700	0.023065	20.711	<2e−16***
(MSSubClass)	−0.161004	0.017717	−9.088	<2e−16***
(TotalBsmtSF)	0.056936	0.006027	9.447	<2e−16***
(YearRemodAdd)	0.054356	0.004498	12.085	<2e−16***
(ExterQual)	−0.120274	0.037620	−3.197	0.001445**
(BsmtCond)	−0.075241	0.021177	−3.553	0.000404***
(FullBath)	−0.239834	0.118946	−2.016	0.064113
(GarageQual)	−0.099868	0.025789	−3.873	0.000117***
(LotArea)	−0.137083	0.040996	−3.344	0.000866***
(KitchenQual)	−0.060986	0.026482	−2.303	0.021549*
(Neighborhood)	−0.128795	0.049195	−2.618	0.009017**

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1; Residual standard error (RSE): 0.113 on 770 degrees of freedom; Multiple R-squared: 0.9123, Adjusted R-squared: 0.9074; F-statistic: 186.2 on 43 and 770 DF, p-value: <2.2e−16; RMSE: 0.12936.

To see which predictor variables are significant, we can examine the coefficients table, which shows the estimate of regression coefficients and the associated t-statistic, and p-values:

For a given predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the response variable, that is whether the beta coefficient of the predictor is significantly different from zero. The regression coefficient can be interpreted as the average effect on the response variable of a one unit increase in predictor, holding all other predictors fixed.

When the p-value is less than 0.05, this tells us that the predictor is statistically significant. In our initial model all predictor variables (except for the predictor FullBath) have p-values less than 0.05, meaning all these predictors are significant.

The predictor FullBath has a p-value of (0.064) which is greater than (0.05), meaning this predictor is insignificant in our model. So, we can improve the initial model by dropping this predictor variable from the initial model and re-run the model again with the remaining predictor variables to see how the new model fit well in the data.

3.2. The Initial Model Accuracy Assessment

We will assess the overall quality of our initial model by examining the R-squared (R^2), the residual standard error (RSE), and the root squared mean error (RSME) [1] [4] [9] [12] [17].

An R^2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable. In our initial model the $R^2 = 0.9123$ which is very good. The adjusted $R^2 = 0.9074$, meaning that "90.74%" of the variance in the measure of the response (SalePrice) can be predicted by the selected predictor variables in the model, which is very good.

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model. In our initial multiple linear regression model, the RSE = 0.113. The RSME of our initial model is (0.12936).

3.3. Checking the Assumptions of Multiple Linear Regression in the Initial Model

Before improving our initial model, we will check the assumptions of multiple linear regression in our developed model. The assumptions of the OLS multiple linear regression include the following [1] [2] [3] [4] [6] [9] [12] [14] [15]:

1) Normal distribution of the response variable. The response variable "SalePrice" was already log transformed in the EDA, so, this assumption holds.

2) Independence of observations in the dataset. Since the dataset was collected correctly by the housing agency. So, this assumption holds.

3) Linearity assumption: linear regression requires that the relationship between the independent and dependent variables to be linear. We will check this assumption by creating plots of some of the predictor variables in the dataset, which were used in the initial mode against the dependent variable SalePrice, as

shown in Figure 10.

It can be seen from the above plots that almost all predictor variables are linearly related to the dependent variable SalePrice. So, this assumption holds.

4) Multicollinearity assumption: linear regression model assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are highly correlated with each other. We have checked the correlation between all predictors. Some predictors were found to be highly correlated with each other and with the response variable SalePrice. We have identified these highly correlated predictors and removed the predictors with more than 50% correlation from the dataset in Step 10. So, this assumption holds.

5) Homoscedasticity assumption: the residuals should be of equal variances across the regression line. We will check this assumption using the Breusch Pagan Test.

The Statistical Breusch Pagan Test:

This test produces a Chi-Square test statistic and a corresponding p-value. If the p-value is below a certain threshold (commonly 0.05) then, there is sufficient evidence to say that heteroscedasticity is present [9] [12] [14] [16] [17]. The results of this test are shown in Table 4.

We can see from the results of the test in Table 4 that the p-value of the test is less than 0.05, therefore we can reject the null hypothesis that the variance of the residuals is constant and conclude that heteroscedasticity is present. In order to rectify for homoscedasticity, we will transform the dataset using Box-Cox approach.

Table 4. The results of Breusch Pagan test.

studentized Breusch-Pagan test
BP = 862.77, df = 770, p-value < 2.2e−16

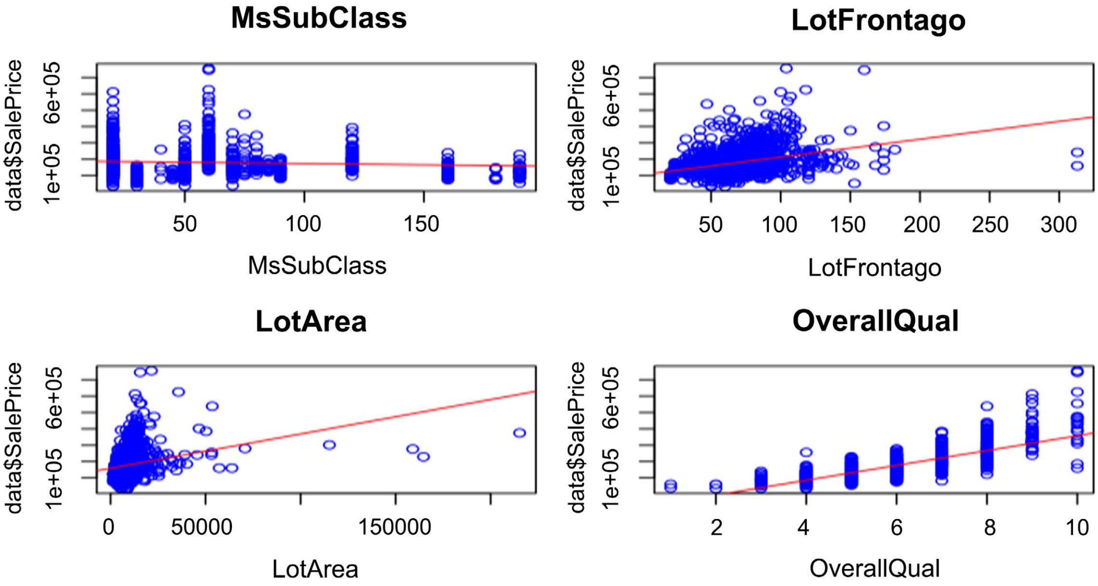

Figure 10. Plots of some features against the response in the dataset.

Box-Cox transformation to Rectify for Homoscedasticity

Box-cox transformation is a mathematical transformation of the variables to make them approximate to a normal distribution [9] [12] [14]. After Box-Cox Transformation, we will rerun the Breusch-Pagan Test again. The results are shown in Table 5.

We can see from the results of rerunning the test in Table 5 that the p-value of the test is now greater than 0.05 after Box-Cox transformation, therefore we can accept the null hypothesis that the variance of the residuals is constant and conclude that the assumption of homoscedasticity holds.

Hence, all the assumptions of multiple linear regression in our initial predictive model hold.

3.4. Developing the Final Predictive Model

We will improve the initial multiple linear regression model using the Ames training file by dropping the predictor variable "FullBath" from the initial model, as it was insignificant, and rerun the initial model. The summary results of the final developed model are shown in Table 6.

Table 5. The results of Breusch Pagan test rerun.

```
studentized Breusch-Pagan test
BP = 969.42, df = 770, p-value 0.91
```

Table 6. The final multiple regression model summary.

Residuals:				
Min	1Q	Median	3Q	Max
−0.74921	−0.05084	0.00629	0.07835	0.39631
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.179216	0.073428	2.441	0.014883*
(GrLivArea)	0.387315	0.121795	3.180	0.001531**
(OverallQual)	0.206114	0.055580	3.708	0.000224***
(MSSubClass)	−0.148740	0.055864	−2.663	0.007918**
(TotalBsmtSF)	0.177196	0.053467	3.314	0.000962***
(YearRemodAdd)	0.138330	0.048847	2.832	0.004748**
(ExterQual)	−0.026555	0.007433	−3.573	0.000375***
(BsmtCond)	−0.096330	0.028374	−3.395	0.000721***
(GarageQual)	−0.011520	0.073236	−0.157	0.000616***
(LotArea)	−0.057779	0.050885	−1.135	0.000866***
(KitchenQual)	−0.037216	0.051101	−0.728	0.011549*
(Neighborhood)	−0.036876	0.053633	−0.688	0.008117**

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1; Residual standard error (RSE): 0.094 on 770 degrees of freedom; Multiple R-squared: 0.9362, Adjusted R-squared: 0.9283; F-statistic: 179.6 on 43 and 770 DF, p-value: <2.2e−16; RMSE: 0.12792.

3.5. The Interpretation of the Final Model

When examining the F-statistic and the associated p-value, at the bottom of model summary, it can be seen that the p-value of the F-statistic is 2.2e−16, which is highly significant.

To see which predictor variables are significant, we can examine the coefficients table, which shows the estimate of regression coefficients and the associated t-statistic, and p-values:

When the p-value is less than 0.05, this tells us that the predictor is statistically significant. In our final model all predictor variables have p-values less than 0.05, meaning all these predictors are significant. So, we can consider this model as our final predictive model.

3.6. The Final Model Accuracy Assessment

We will assess the overall quality of our final model by examining the R-squared (R^2), the residual standard error (RSE), and the root squared mean error (RSME).

An R^2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable. In our final model the $R^2 = 0.9362$ which is very good and greater than the initial $R^2 = 0.9123$. The adjusted $R^2 = 0.9283$, meaning that "92.83%" of the variance in the measure of the response (SalePrice) can be predicted by the selected predictor variables in the model, which is very good and better than the initial adjusted $R^2 = 0.9074$.

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model. In our final multiple linear regression model, the RSE = 0.094, which is smaller than the initial RSE = 0.113 indicating an improvement of the final model accuracy compared to the initial model. The RSME of our final model is (0.12792), which is smaller than the initial RSME = 0.12936, which also indicates an improvement in the final model over the initial model. Moreover, the RSME of the test dataset is lower than the RSME of the train dataset, which implies that there is no overfitting of the data, and this also indicates an improvement of our final model.

3.7. Checking the Prediction Error (MSE) of the Final Model

In addition to the model accuracy assessment that we have performed, we will also calculate the prediction error (Mean Squared Error, MSE) of our final model using the test dataset. We will use three cross validation techniques, namely, the validation set approach, the K-fold approach and the Leave-One-Out-Cross Validation (LOOCV) approach in calculating the prediction error [1] [2] [3] [12] [14] [15]. We know that the total number of observations (rows) in the test dataset is 1459 and total number of features is 80.

3.8. Validation Set Approach

This approach splits the test dataset into two parts randomly; train and test. The calculated MSE was found to be (0.011261), which represent the prediction error of our final regression model based on the validation set approach.

3.9. K-Fold Cross Validation (CV) Approach

The K-fold CV splits the entire test dataset into K parts, of which K − 1 parts are used as training data and a single part is used as test data and the process is repeated so that each part can be treated as a test data. We will use K-10 cross validation in this resampling approach. The prediction error (test error) for our final regression model based on the 10-fold cross validation approach was found to be 0.010685, which is slightly lower test error than the validation set approach.

3.10. Leave-One-Out-Cross Validation (LOOCV) Approach

In LOOCV approach, each observation (row) of the original data is used as the test data and rest of the observations are treated as training data. So, we will use K = 1459 in this approach. The prediction error (test error) for our final regression model based on the LOOCV approach was found to be 0.012428. The mean squared error is very small in the final model (max MSE = 12% using LOOCV), which indicates extremely good results, as shown in Table 7.

3.11. Predicting House Prices Using the Test Dataset with the Final Multiple Linear Regression Model

We will predict the house price using the features in the test dataset with the final predictive model. The first 10 predicted house prices are shown in Table 8.

Table 7. The prediction error of the final model using different approaches.

Approach	Validation Set	K-fold CV	LOOCV
Prediction error (test error), MSE	0.011261	0.010685	0.012428

Table 8. The predicted house prices of the first 10 houses using the final model.

N	Predicted SalePrice in ($)
1	124,016.149
2	159,066.458
3	186,471.543
4	198,486.081
5	187,279.477
6	168,734.369
7	174,938.905
8	164,972.560
9	180,377.353
10	126,628.191

4. Conclusion

This paper presented a real estate dataset describing the sale of individual residential property in Ames, Iowa, USA from 2006 to 2010. Ames data consisted of 80 assessment parameters or explanatory variables which described every aspect of residential homes in Ames. The explanatory variables consisted of 23 nominal, 23 ordinal, 14 discrete, and 20 continuous. The Ames Housing dataset contained both training dataset and testing dataset. The two datasets were representation of whole data spilt into 50% - 50% to train and test sets. The paper used a multiple linear regression analysis to predict the final price of every house in the dataset. The goal was to use the training data to predict the sale prices of the houses in the testing data. The methodology included cleaning the datasets from outliers, combining the training and testing data files, exploring the data and producing the summary statistics, checking for the completeness of data and missing values, conducting Exploratory Data Analysis (EDA), removing non-important features based on EDA, checking multicollinearity among different features, removing highly correlated features, removing near zero variance predictors, selecting a short list of predictor features for prediction of the house price, developing a multiple linear regression model for the prediction, checking the assumptions of the model and performing necessary diagnosis, calculating the prediction error of the optimal developed model by cross validation approaches, and predicting the house price using the test data. An optimal final predictive model was achieved by keeping the most influential predictors only. The model accuracy assessments produced promising results with an adjusted R-squared value of 0.9283, a residual standard error (RSE) of 0.094, and a root squared mean error (RSME) of 0.12792. In addition, the prediction error (Mean Squared Error, MSE) of the final model was found to be very small (12%) by applying different cross validation techniques, including the validation set approach, the K-fold approach and the Leave-One-Out-Cross Validation (LOOCV) approach. The results of this study showed that the multiple linear regression can effectively predict the response variable with big datasets and large number of predictors.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Gareth, J., Witten, D. Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: With Applications in R. Springer, New York.

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2008) The Elements of Statistical Learning. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

[3] Bruce, P. and Andrew, B. (2017) Practical Statistics for Data Scientists. O'Reilly Media, Sebastopol.

[4] Berry, W.D. and Feldman, S. (1985) Multiple Regression in Practice. Sage Universi-

ty Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-050, Sage, Newbury Park.

[5] Abdulhafedh, A. (2017) A Novel Hybrid Method for Measuring the Spatial Autocorrelation of Vehicular Crashes: Combining Moran's Index and Getis-Ord G*i Statistic. *Open Journal of Civil Engineering*, **7**, 208-221. https://doi.org/10.4236/ojce.2017.72013

[6] Cohen, J. and Cohen, P. (1983) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Inc., Hillsdale.

[7] Abdulhafedh, A. (2017) Road Traffic Crash Data: An Over-View on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, **7**, 206-219. https://doi.org/10.4236/jtts.2017.72015

[8] Abdulhafedh, A. (2017) Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, **7**, 190-205. https://doi.org/10.4236/jtts.2017.72014

[9] Pedhazur, E.J. (1997) Multiple Regression in Behavioral Research. 3rd Edition, Harcourt Brace Orlando.

[10] Abdulhafedh, A. (2017) Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview. *Journal of Transportation Technologies*, **7**, 279-303. https://doi.org/10.4236/jtts.2017.73019

[11] Tabachnick, B.G. and Fidell, L.S. (2001). Using Multivariate Statistics. 4th Edition, Allyn and Bacon, Needham Heights.

[12] Montgomery, D.C. and Peck, E.A. (1982) Introduction to Linear Regression Analysis. John Wiley and Sons, Inc., New York.

[13] Abdulhafedh, A. (2016) Crash Frequency Analysis. *Journal of Transportation Technologies*, **6**, 169-180. https://doi.org/10.4236/jtts.2016.64017

[14] Rawlings, J.O. (1988) Applied Regression Analysis: A Research Tool. Wadsworth & Brooks/Cole, Pacific Grove.

[15] Jobson, J.D. (1991) Multiple Linear Regression. In: *Applied Multivariate Data Analysis*, Springer, New York, 219-398. https://doi.org/10.1007/978-1-4612-0955-3

[16] Weisberg, S. (1980) Applied Linear Regression. 2nd Edition, John Wiley and Sons, Inc., New York.

[17] Neter, J., Wasserman, W. and Kutner, M.H. (1983) Applied Linear Regression Models. Richard D. Irwin, Inc., Homewood.

[18] Abdulhafedh, A. (2021) Incorporating K-Means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, **3**, 12-30.

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

[20] Van Buuren, S. (2018) Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9780429492259

[21] Schafer, J.L. and Graham, J.W. (2002) Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-77. https://doi.org/10.1037/1082-989X.7.2.147

[22] Abayomi, K., Gelman, A. and Levy, M. (2008) Diagnostics for Multivariate Imputations. *Journal of the Royal Statistical Society C*, **57**, 273-291. https://doi.org/10.1111/j.1467-9876.2007.00613.x

[23] Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3

[24] Zorn, C. (2005) A Solution to Separation in Binary Response Models. *Political Analysis*, **13**, 157-170. https://doi.org/10.1093/pan/mpi009

[25] Abdulhafedh, A. (2021) Vehicle Crash Frequency Analysis Using Ridge Regression. *International Journal for Science and Advance Research in Technology*, **7**, 254-261.

[26] Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y.S. (2008) A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics*, **2**, 1360-1383. https://doi.org/10.2139/ssrn.1010421