IE 590: Final Report

Advanced Regression Techniques: House Pricing

Authors: Brian Baller, Hrishikesh Ganesan, Kaushik Manchella

Contents

Abstract

With a housing assessment data set from Ames, Iowa, this project assumed the objective of studying the factors that influence housing prices and make accurate predictions on the same quantity. The project objective caters to the primary stakeholder who is a member of the team that has sold many homes and is currently in the process of selling another one. Other parties such as real estate agencies may benefit from the outcomes of this project as well.

In addressing the stated objective, the team had to adopt a data science implementation structure that emphasized iterative improvement in prediction and insight generation. The implementation involved a comprehensive exploratory data analysis that enabled feature engineering and model building for each feature set.

Exploratory data analysis was crucial in not only identifying influential predictors but prepared the team for model implementation by revealing aspects such as imperfections in the data, and potential outliers.

Additionally, this paper intends to describe how different predictive modeling algorithms fared and discusses why so.

Finally, inferences from models will be drawn to add to the collective domain knowledge on what factors influence housing prices.

Background

The data set is from the Ames City Assessor's Office, which performs yearly tax assessments of residential properties in the Ames area. Although Ames, Iowa is of no particular interest to the authors, its similarity to the Lafayette area is noteworthy: both are mid-size midwestern towns which house large public universities (Iowa State and Purdue). The 80 predictor variables in the data set are quantitative and qualitative details about the property itself (e.g. square footage, number of bathrooms, type of foundation) as well as qualitative assessments of property condition made by on-site tax assessors. Two other key variables include month and year sold. Our team will set out to predict the sale prices of a test data set with a model constructed from the full Kaggle dataset.

Our hypothesis is that home prices are largely dictated by three key variables: location, home size and date of sale. The conventional idiom of "location, location, location" details the importance of location in housing sale prices. Safe neighborhoods, good schools, and access to amenities are all desirable characteristics of a good neighborhood. We feel home size will also be a determinant of sales price because there is a direct correlation between home size and the cost of materials/labor needed to construct a house. Lastly, date of sale will likely be important, as in general home prices increase over time. However, the time period covered by this data includes the "Great Recession" in 2007, so this general rule may not be applicable. Location information is captured in the Neighborhood field and consists of 25 categorical names of neighborhoods.

Methodology

The objective of this project entails not only maximizing predictive accuracy, but also explaining influential factors on housing prices. Keeping the explanatory focus in the modeling allows training a model that is scalable. ie. applicable to other cities such as West Lafayette.

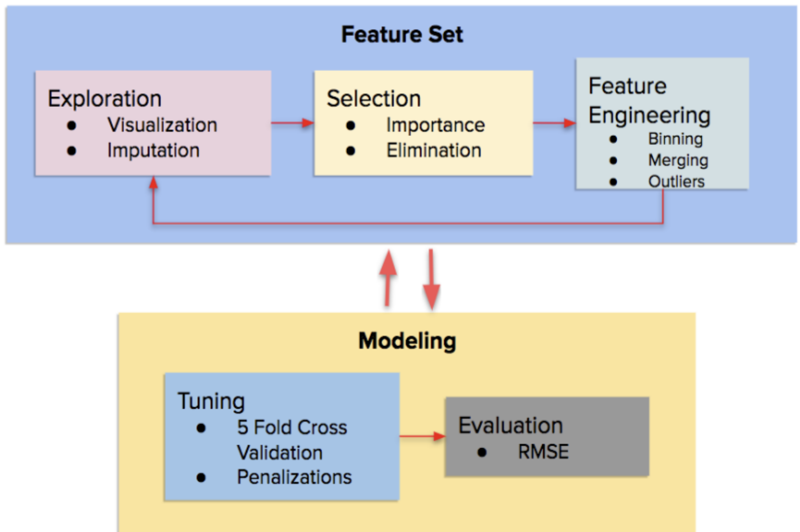


Figure 1: Methodology

Figure 1 shows the structure that the team adopted in addressing the stated objectives. Driven by the end goal of building a model which addressed the predictive as well as explanatory requirement, the feature set methodology was chosen to allow easy tracking of the influence of variables. In the above process, different feature sets were generated and modeled on individually. This iterative process allowed for continuous improvement in predictive and explanatory power.

Exploratory Data Analysis

Exploratory Data Analysis was a crucial exercise that revealed certain characteristics of the data set. The key outcomes of the EDA included NA imputations, identification of important predictors, elimination of unhelpful predictors, potential outliers and insight onto feature engineering and modeling.

While the team did a thorough visualization of every variable in the data set, this section of the report intends to walk through the highlights of the EDA and describe how that allowed the team to manipulate the data set and prepare for modeling. A full detailed EDA walk through can be found in APPENDIX.

Target Variable

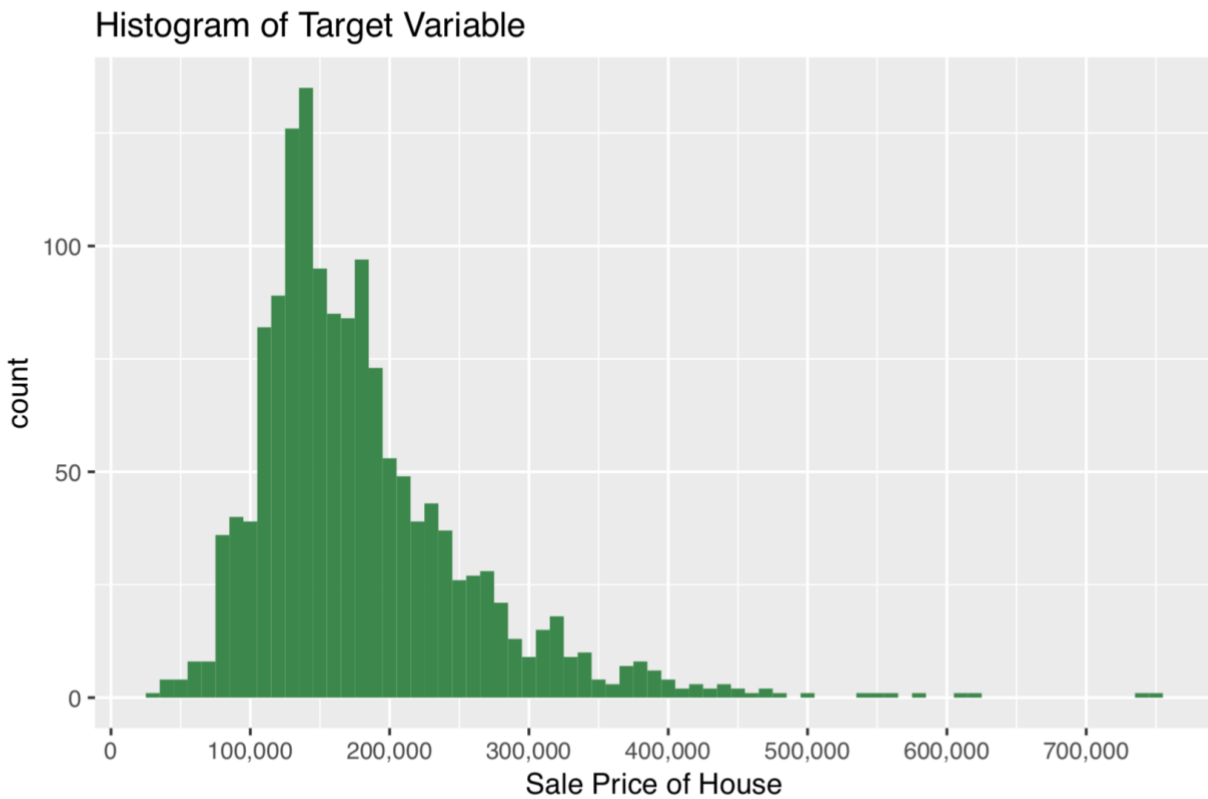


Figure 2: Sale Price Histogram

Figure 2 provides a qualitative judgement of sales price for houses listed in the Ames data set. Key points to not from here is that, there is no normality in the data. The sale price distribution is a little skewed with majority of houses costing in the lower price quartile within the entire range.
This is in accordance with the qqplot shown on page 2 of the EDA Appendix. As a result, a log transformation on sales price was ultimately used in the model implementation which is explained further in this paper.

Important Predictors

A study of the most influential predictors was commenced with a correlation plot of the numeric variables which showed the strongest correlation to Sale Price.

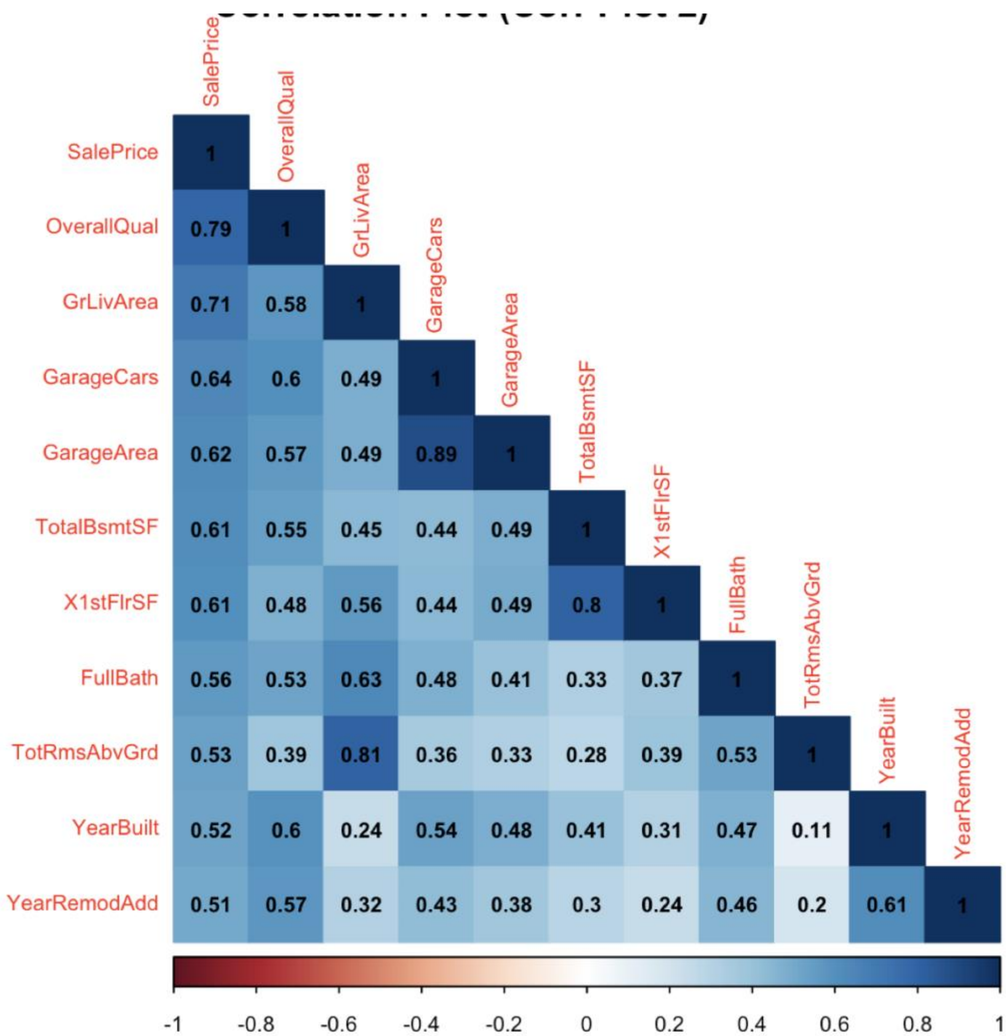


Figure 3: Correlation Matrix

From the correlation matrix in Figure 3, important variables such as Overall Quality (OverallQual), Above Ground Livable Area (GrLivArea) are explored.

Overall Quality

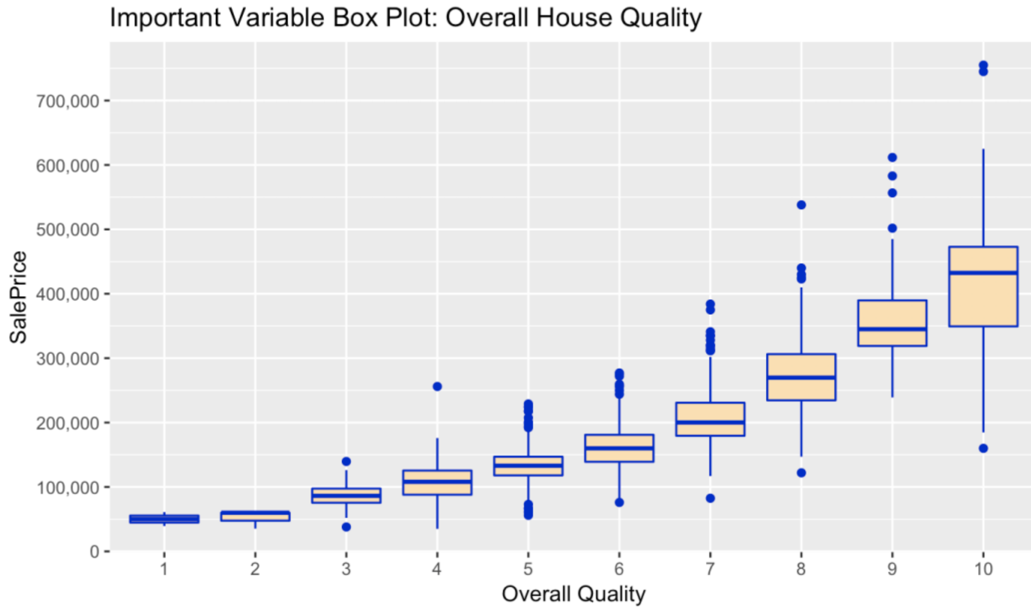


Figure 4: Overall Quality Box Plot

Quality rating is given on a scale of 10 as seen above, and as expected, the most expected houses generally tend to be of the highest Overall Quality. It is worth noting however that there is a variance and outliers for every quality box. eg. The $250k house which is of quality = 4 which seems odd. This potential outlier was removed later on to generate Feature Set 4.

Above Ground Liveable Area

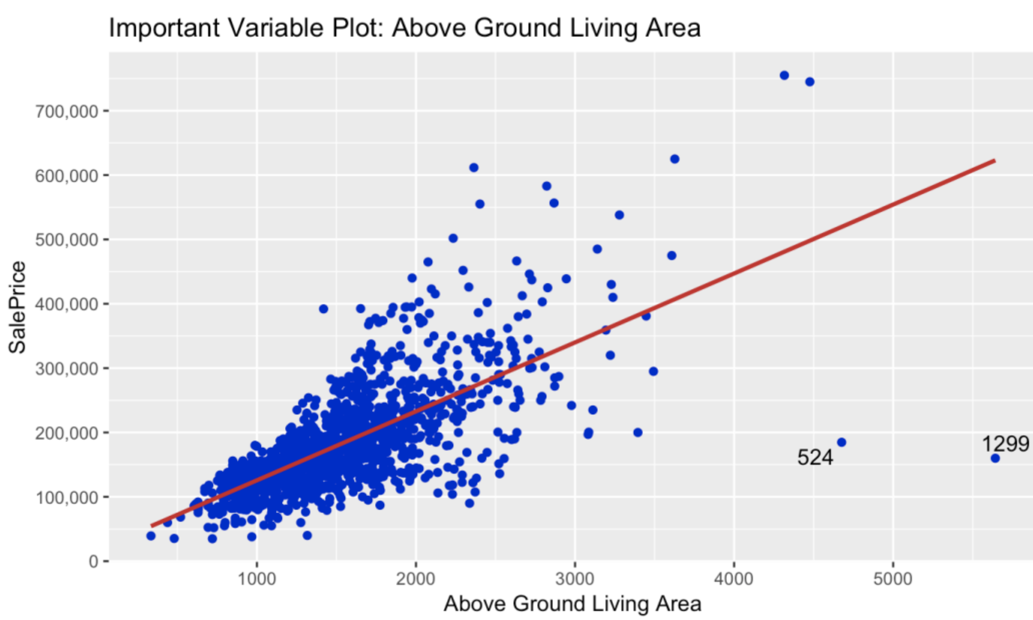


Figure 5: Linear Plot of Above Ground Livable Area

Given this predictor is a continuous variable, a simple linear relationship test was performed to generate figure 5. The two houses with really big living areas and low Sale Prices appeared to be outliers (houses 524 and 1299, see labels in graph), and were removed later during the modeling process to observe their influence in the predictive accuracy (RMSE delta).

Feature Engineering

In summary, the intent of feature engineering was to evaluate the existing variables in our data set and manipulate them to facilitate the training of a model with better performance. Figure 6 below describes the different feature sets used for modeling.

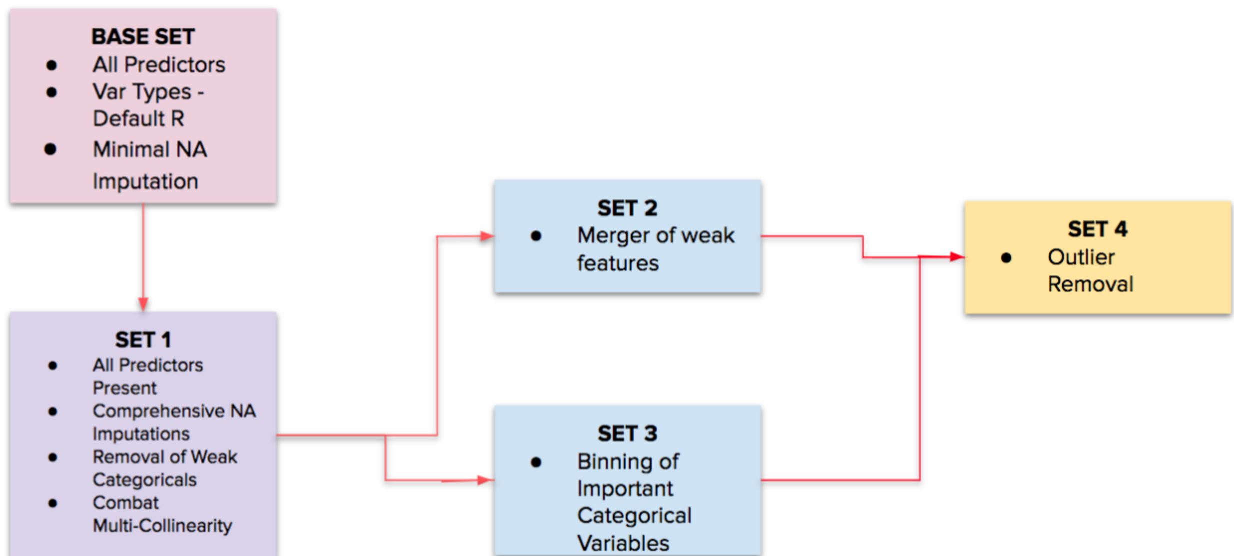


Figure 6: Feature Engineering Methodology

1) Merger of Weak Features
2) Binning of Important Categorical Variable – Neighborhood
3) Outlier Removal

A detailed description and assessment of the methodology behind each of the above feature engineering method may be found in Appendix II (Feature Engineering).

Model Evaluation

The approach to modeling for this project was to run applicable models that we have used in class and evaluate their predictive and explanatory performance. Analysis was performed using a training / test set split of 70/30, yielding 1020 observations in the training set and 438 in the test set (with the two outliers deleted). The training set was used to fit the model and tune any hyperparameters. The hyperparameters were picked with K-fold cross-validation within the training set only. The test set was solely used to evaluate the final model for each model type (Figure 7).

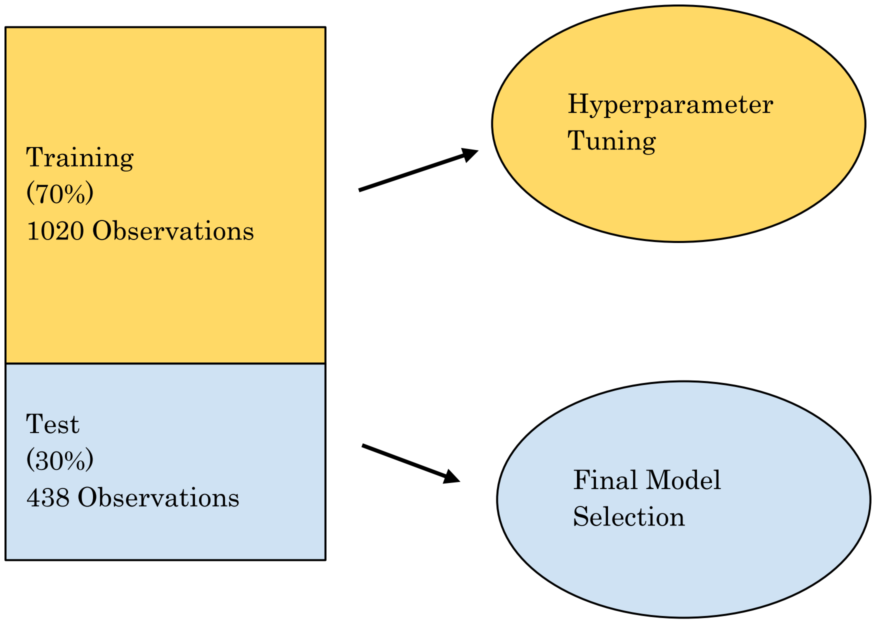


Figure 7: Test/Train Split

All models were run on all the sets described above. The best set was Set #4 (outliers removed) and the best model was LASSO. Predictions with the null model yielded a root mean square error (RMSE) of ~$87K against an ~23K RMSE with the LASSO model. Final results are in Table 1:

Table 1

Model Performance		
Model	Hyperparameters	RMSE
Null Model	NA	87070
Linear (LASSO)	lambda	22622
Decision Tree	Cp	44772
Random Forest	ntree, mtry	27797
BART	ntree, k, q, **ν**	24429
Gradient Boost	n.trees, interaction.depth, shrinkage	24801
MARS	number of predictors	23930
SVR	epsilon, cost	38173

The RMSE listed here is the test set RMSE. These RMSE numbers will vary based on the seed chosen to divide the set into training and test set (see section on robustness) and the seed used in cross-validation. Improvements from set to set can be viewed in Table 2. Significant improvement was made going from the Base Set to Set #1, but no appreciable improvement was made going to Sets #2 and #3.

Table 2

	BASE SET	SET 1	SET 2	SET 3	SET 4
Model	RMSE	RMSE	RMSE	RMSE	RMSE
Null	87070	87070	87070	87070	87070
Linear (Lasso)	29160	28346	29436	28346	22622
Decision Tree	41290	41616	46701	41287	44772
Random Forest	28580	28526	29584	30998	27797
BART	25640	28146	25918	25176	24429
Gradient Boost	26710	24412	24652	25559	24801
MARS	29820	29863	29920	29482	23930
SVR	43130	28100	27824	27637	38173

The LASSO model performs best, but only when the two significant outliers are removed. Both outlying points have high leverage and thus "pull" the linear fit toward the outlying points, altering the predictions for the test set. Accordingly, the non-linear models perform better with the outliers included. A plot of LASSO-model predicted price versus observed price is in Figure 8. Overall, the model predictions are close to the observed values, especially in the lower band of home values. Note this plot will change slightly with different seed values for the training and test sets.

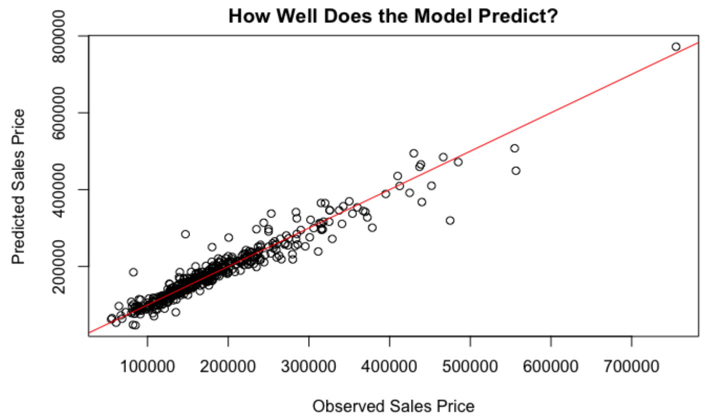


Figure 8: LASSO Model Performance

Inference

To ascertain the relative importance of the predictor variables, our top two models, LASSO and MARS, were used. For LASSO, the predictors were standardized and coefficients were calculated. Although glmnet() will standardize the predictors upon calling the function, it outputs the coefficients in "unstandardized form". Using the absolute value of the standardized coefficients, we can ascertain variable importance. Our second place model, MARS, has a variable importance function, evimp(). Both are given in the below table. *Total Square Footage*, *Overall Quality* and *Overall Condition* feature highly in both models. As before, the importance of these variables will change based on the seed chosen to divide the set into training and test sets (see Table XX).

Table 3

Variable Importance		
Rank	LASSO	MARS
1	Total Square Footage	Overall Quality
2	Overall Quality	Total Square Footage
3	Overall Condition	Overall Condition

Much to our surprise, neighborhood nor date of sale factored heavily into our predictions. Although neighborhood does effect the sale price, the three richest neighborhoods contain less than 5% of the total homes sold from 2006 to 2010 (Figure 9). It's possible that these neighborhoods are small, but also possible that these houses change hands less frequently than their cheaper counterparts, and thus are underrepresented in the sample data.

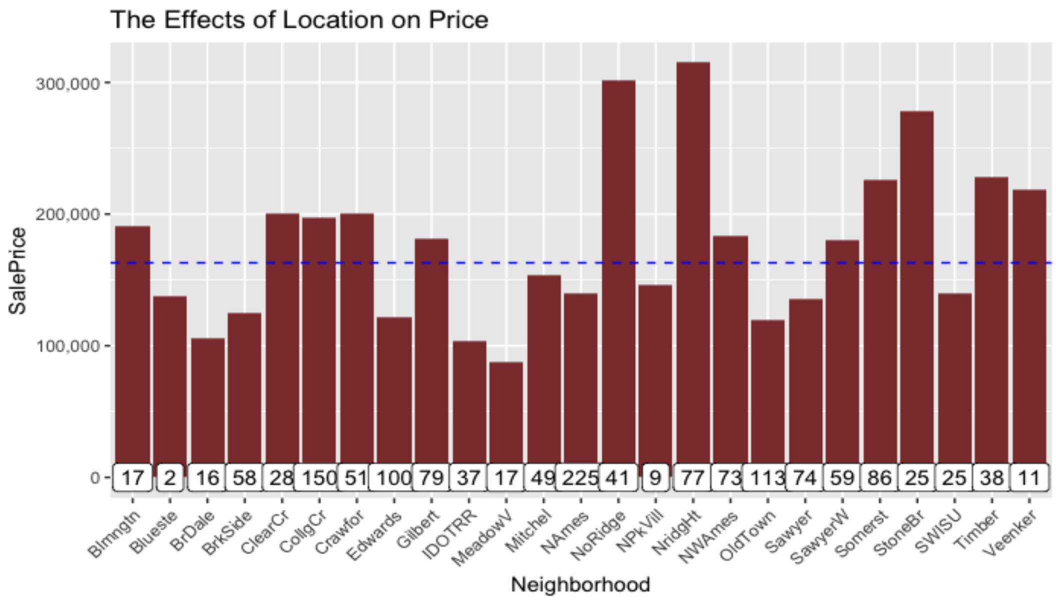


Figure 9: Effect of Neighborhood on Sale Price

Despite the housing crisis, Ames, Iowa wasn't greatly affected during the housing crisis. Although there was a slight downtrend in prices over the period, this down trend was limited and overshadowed by the natural variability of seasonal home prices (see Figure 10). Note that the x-axis of the below chart starts at Month 0 (Jan 2006) and ends at Month 54 (July 2010).

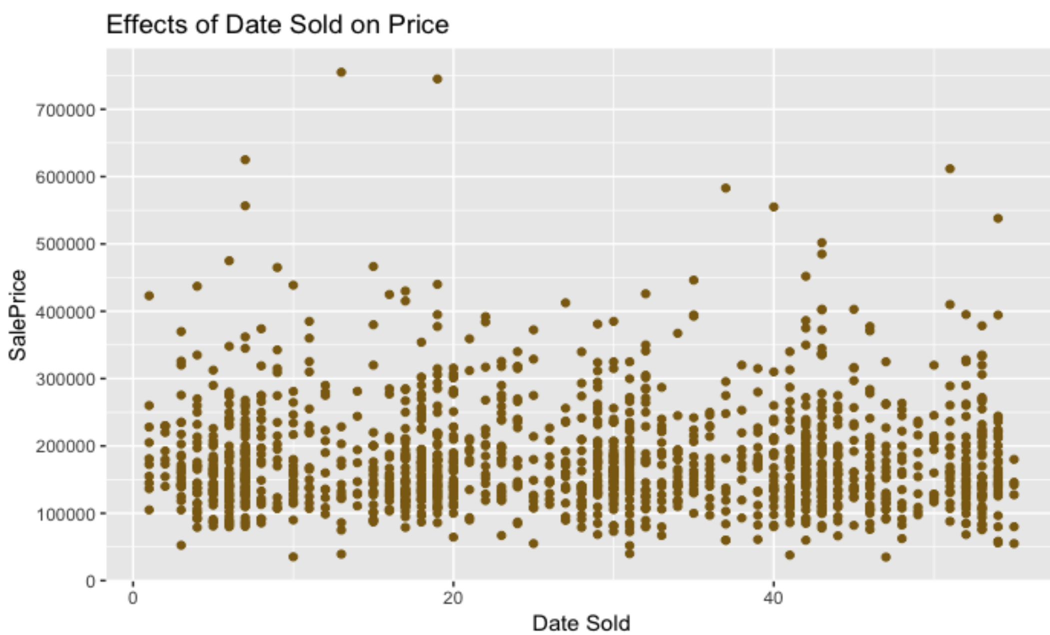


Figure 10: Effect of Date Sold on Sale Price

Model Robustness

Randomness plays a role in model performance. The training and test sets are randomly selected as are the k-folds in our cross-validation algorithms. If the training / test set split is favorable to a linear model, the RMSE for LASSO will be lower than a split which is unfavorable (for instance, all outliers in the training set). To ascertain the robustness of our winning models, we varied the seeds used to select the training / test sets from 1 to 100. Test RMSE of the LASSO and MARS models are depicted in Figure 11. The LASSO is centered at $21,400 (lower than our above results, which used seed = 41) with a standard deviation of $1480. Although LASSO outperforms MARS on average, it's possible that MARS will outperform LASSO with particular training / test set splits. A t-test of two means was performed on the results and the difference in performance is statistically significant.

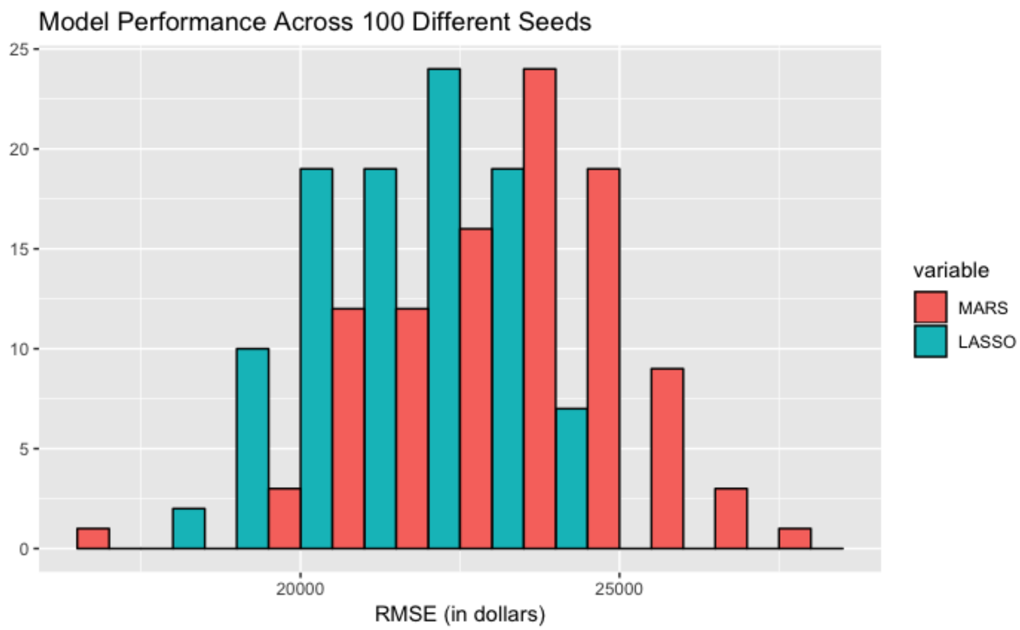


Figure 11: Model Performance Across 100 Seeds

Variable importance within the model will also change as the training/test set seed varies. For seeds 1-20, *Total Square Footage* was most important 10 times, while *Overall Quality* claimed the other 10 spots (see Table 4). This variability is heightened due to multicollinearity between the predictors.

Table 4

Variable Importance over 20 Seed Values			
	#1	#2	#3
Total Square Footage	10	10	0
Overall Quality	10	9	1
Overall Condition	0	0	6

Conclusion

The business of predicting home prices is difficult business, but limiting the scope of the project to Ames, Iowa yields satisfactory results. The best model average error is approximately $22,000 in a market with a median home price of $163,000. The most important predictors of home price were *Total Square Footage*, *Overall Quality* and *Overall Condition.* Both *Overall Quality* and *Overall Condition* are subjective assessments from the local tax assessor. In the absence these assessments, predictive power would be significantly lower than what was achieved. With these predictors however, the results of the analysis will be difficult to extend to other localities with dissimilar assessment practices.

Exploratory Data Analysis

Group 4

4/29/2019

APPENDIX I - Imputations, Detailed EDA, Label Encoding

After visualizing the extent NA values in the main section of this project draft, this Appendix goes into further depth by exploring each variable and imputing NAs appropriately.

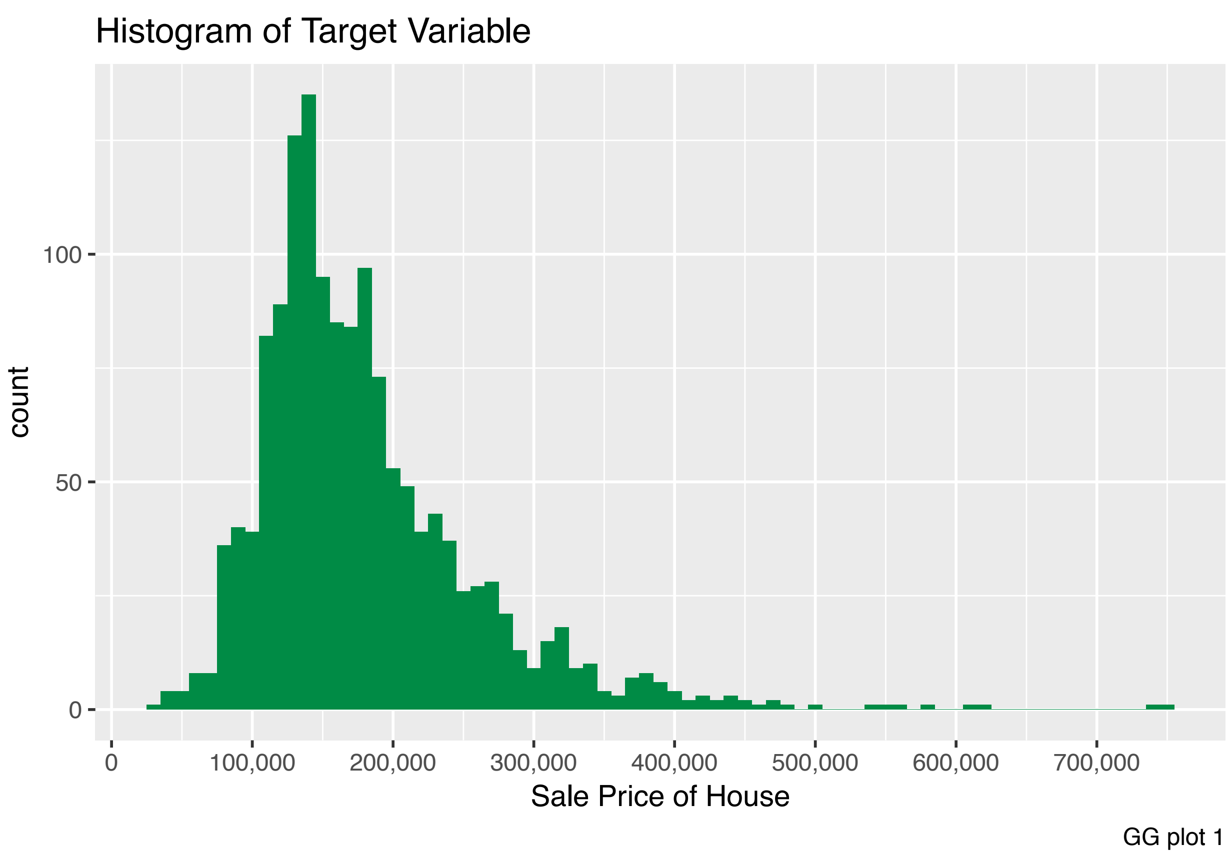
Upon imputation, a quick EDA is performed.

Each variable is studied individually and categorized into one of the following groups:

```
1. Numeric
2. Integer
3. Ordinal
4. Factor (Categoric Object)
```

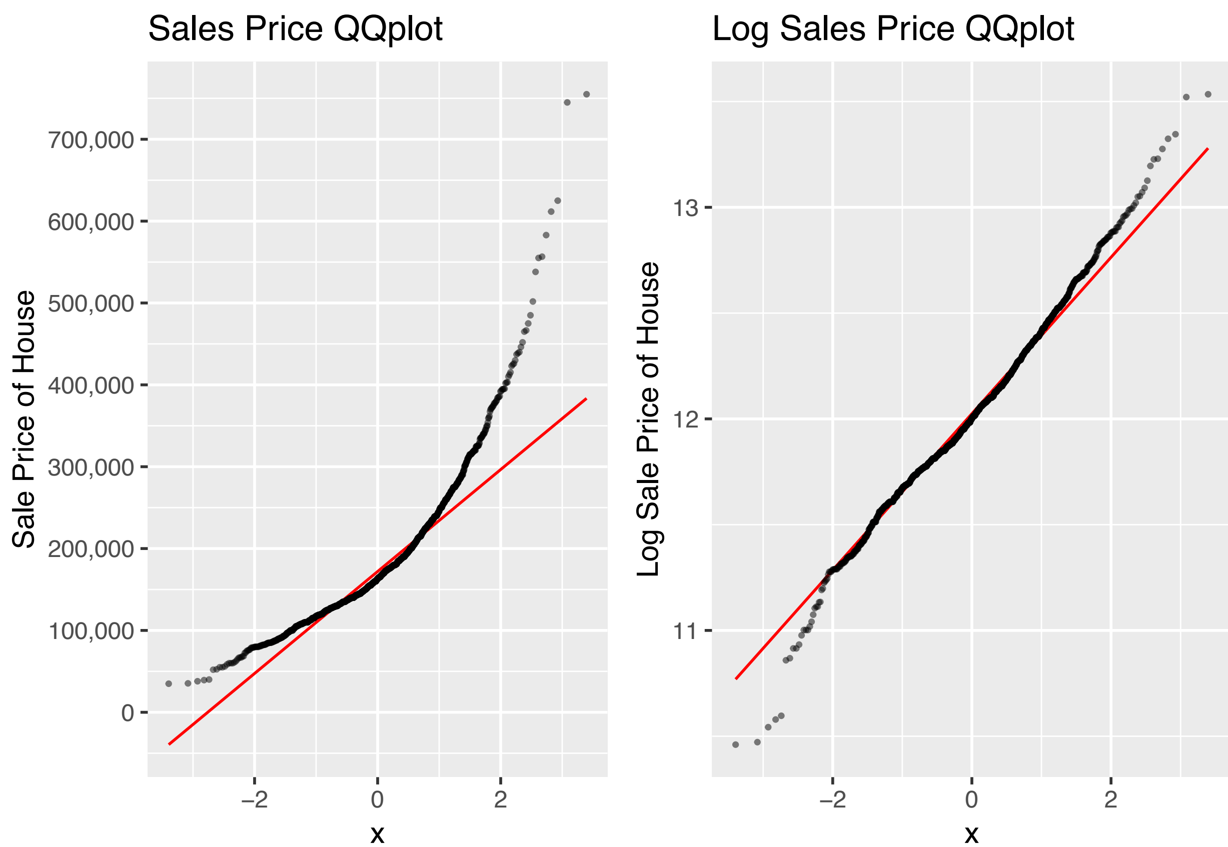
Numeric and Integer variables did not go any for of encoding. Ordinal variables underwent a Label Encoding. Categoric Objects were kept as the text code and transformed to factors within the data set.

Observing the Target Variable



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

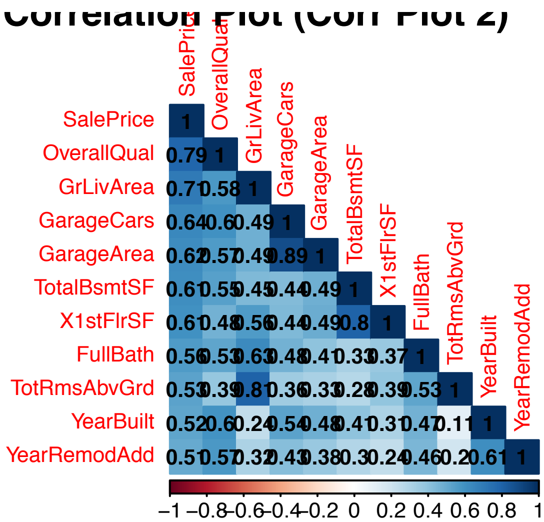
The above histogram and summary gives us a qualitative judgement of Sales Price for the houses listed in the Ames data set. Checking the qq plot of the target variable:



The variables in the dataset can be grouped by the type of feature they are representing. Eg. 'Pool Area' and 'Pool Quality' can both be grouped as pool variables. We shall do our imputation, visualization and label encoding group by group.

Moving on to the most imporatnt numeric predictors. To get a feel for the dataset, we decided to first see which numeric variables have a high correlation with the SalePrice. In total, there are 10 numeric variables with a correlation of at least 0.5 with SalePrice.

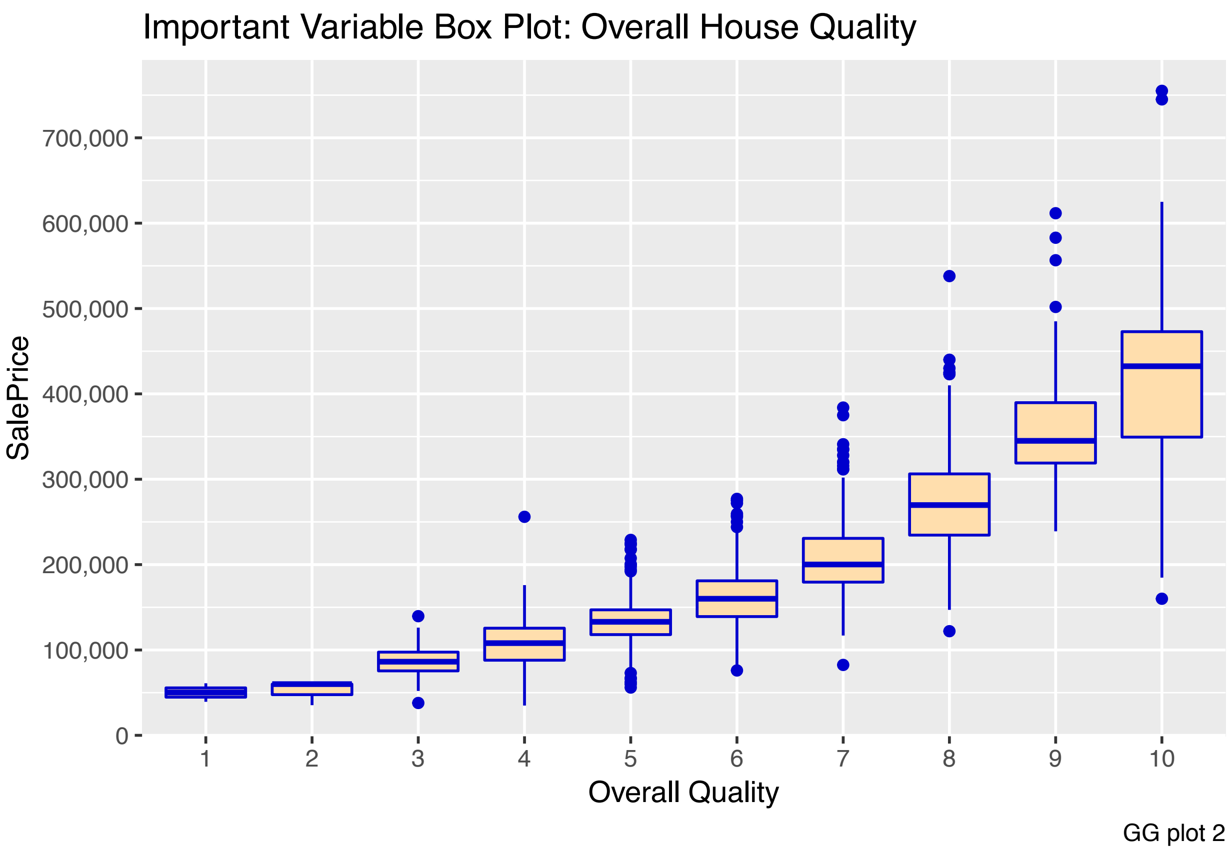
Correlations



Note that all those correlations are positive. We have a total of 11 variables with a correlation of 0.5 or greater with Sales Price.

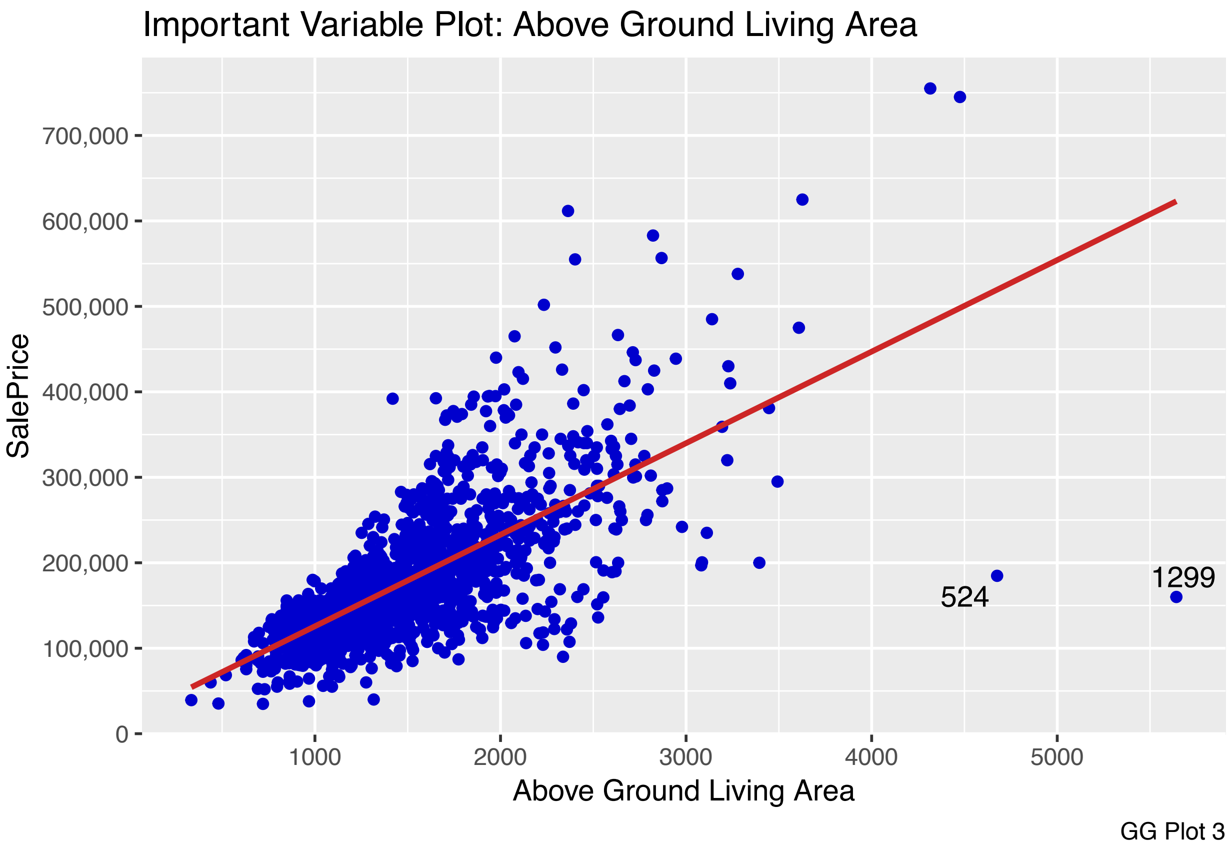
Given that the highest correlated variables are Overall Quality ('OverallQual') and Total Above Ground Living Area ('GrLivArea'), it is a good idea to get a closer look at these.

Visualizating Attribute 'Overall Quality'



Quality rating is given on a scale of 10 as seen above, and as expected, the most expected houses generally tend to be of the highest Overall Quality. It is worth noting however that there is a variance and outliers for every quality box. eg. The $250k house which is of quality = 4 which seems odd. This could be a potential outlier to be studied later on.

Above Ground Living Area Visualization



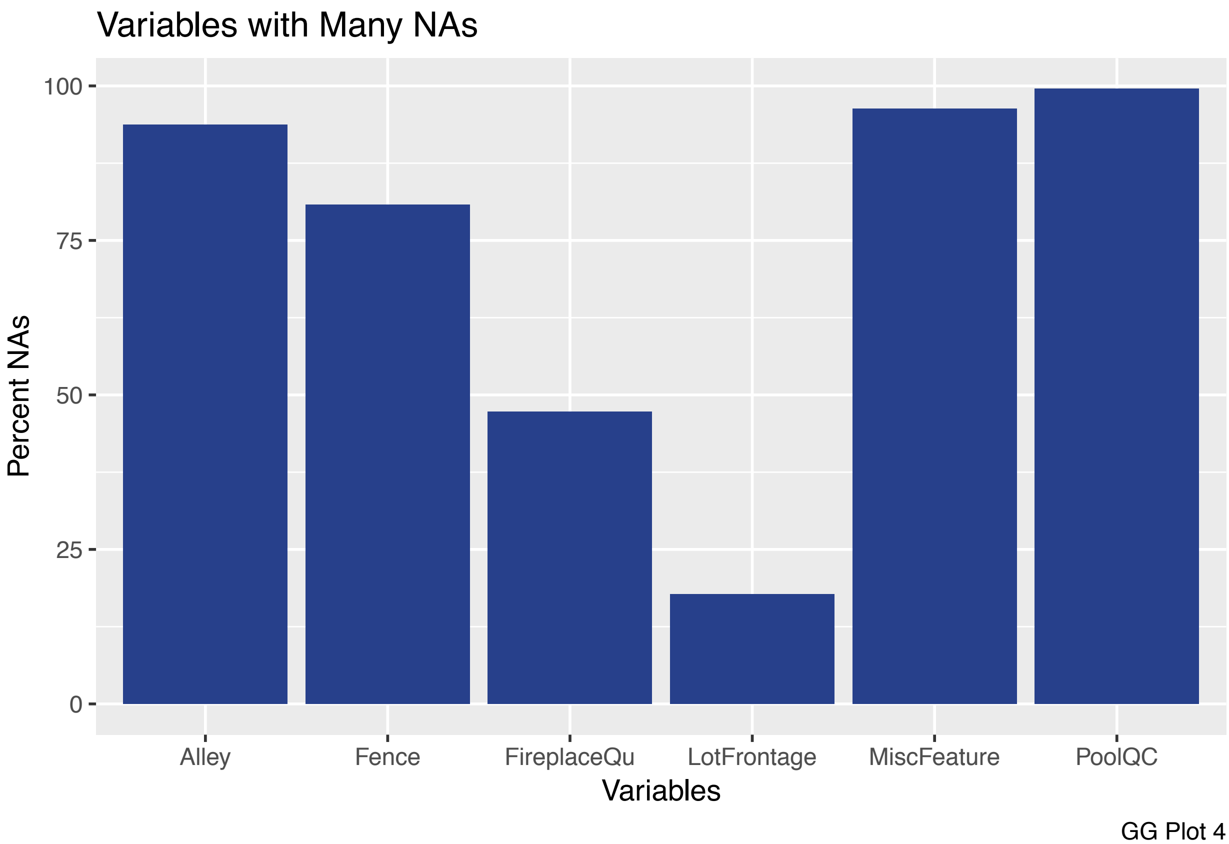
Given this predictor is a continuous variable, a simple linear relationship test was appropriate. The geom_text_repel tool was labeled to highlight the index numbers of potential outliers. Especially the two houses with really big living areas and low SalePrices seem outliers (houses 524 and 1299, see labels in graph).

EDAs for the other highly correlated numerical variables will be attached in the Appendix for further review if needed.

Given that the categorical variables in the set contain a lot of NA's, we will proceed to some pre-processing before visually observing the categorical variables.

Preprocessing Stage 1 - Handling NAs, Label Encoding, Factorizing Variables

Visualizing NA's



The barplot above visualizes the variables with the largest percent of missing values. While the summary below shows the full NA story.

```
##       PoolQC  MiscFeature        Alley        Fence    SalePrice
##         2909         2814         2721         2348         1459
##  FireplaceQu  LotFrontage  GarageYrBlt GarageFinish   GarageQual
##         1420          486          159          159          159
##   GarageCond   GarageType     BsmtCond BsmtExposure     BsmtQual
##          159          157           82           82           81
## BsmtFinType2 BsmtFinType1   MasVnrType   MasVnrArea     MSZoning
##           80           79           24           23            4
##    Utilities BsmtFullBath BsmtHalfBath   Functional  Exterior1st
##            2            2            2            2            1
##  Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            1            1            1            1            1
##   Electrical  KitchenQual   GarageCars   GarageArea     SaleType
##            1            1            1            1            1
```

```
## There are 35  total columns with missing values
```

After doing the high level EDA, we dove deeper into each variables to handle NAs and apply the necessary encoding going forward.

In summary, imputations for NAs were handled differently for each variable depending on the nature of the variable and the quantity of NAs. The three main scenarios were as follows: Categorical Variable with very few NAs - Imputed by Mode of Categorical Object Categorical Variable with many NAs - These NAs were legitimate NAs. eg. Houses without pools had NAs for attribute 'Pool Quality' Numeric/Integer variable with very few 10 NAs - Imputed by Median of Numeric Variable Numeric/Integer varibable with many NAs - Legitimate NA's. eg. 'Pool Area'. These were made to be Zero.

The Detailed Visualization section of this appendix provides an indepth walk through as to how imputation was done for every single in the dataset. Please review at your convenience to get a better understanding.

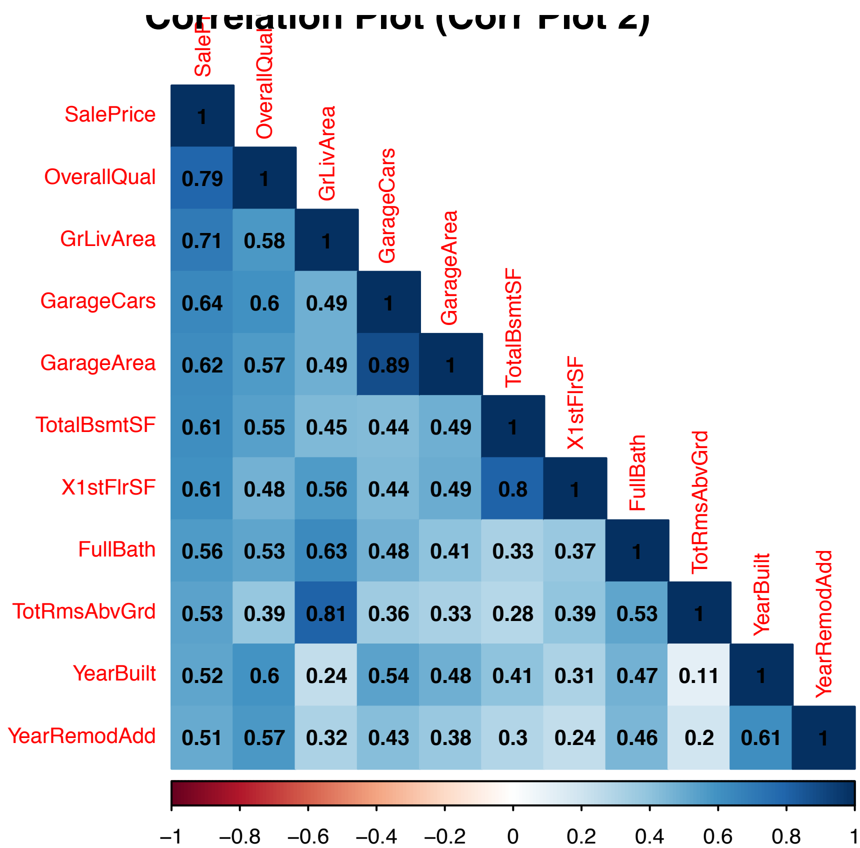
Each variable is studied individually and categorized into one of the following groups:

```
1. Numeric
2. Integer
3. Ordinal
4. Factor (Categoric Object)
```

Numeric and Integer variables did not undergo any form of encoding. Ordinal variables underwent a Label Encoding. Categoric Objects were kept as the text code and transformed to factors within the data set.

Post Processing EDA

Correlations after imputations and variable encoding



Upon generating the new correlation plot (after preprocessing), it is worth noting that there are **six** more variables with a greater correlation than 0.5.

Detailed Visualization

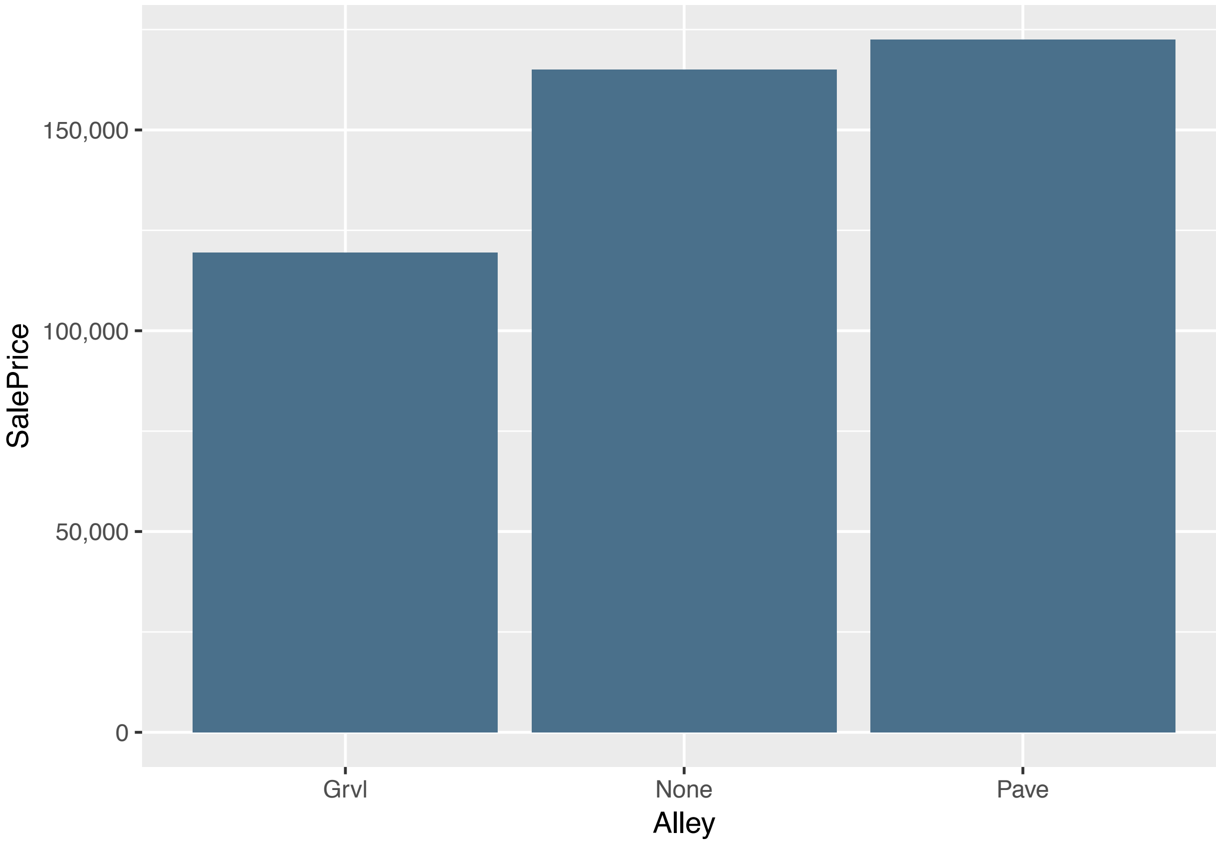
Note: Before each plot, a variable dictionary is provided for ease of interpretation. Please refer to the dictionary before looking at the plots

Alley

Within Alley, there are 2721 NAs. Values:

```
Grvl Gravel
Pave Paved
NA   No fulley access
```

Value Type: Factors (Not Ordinal)



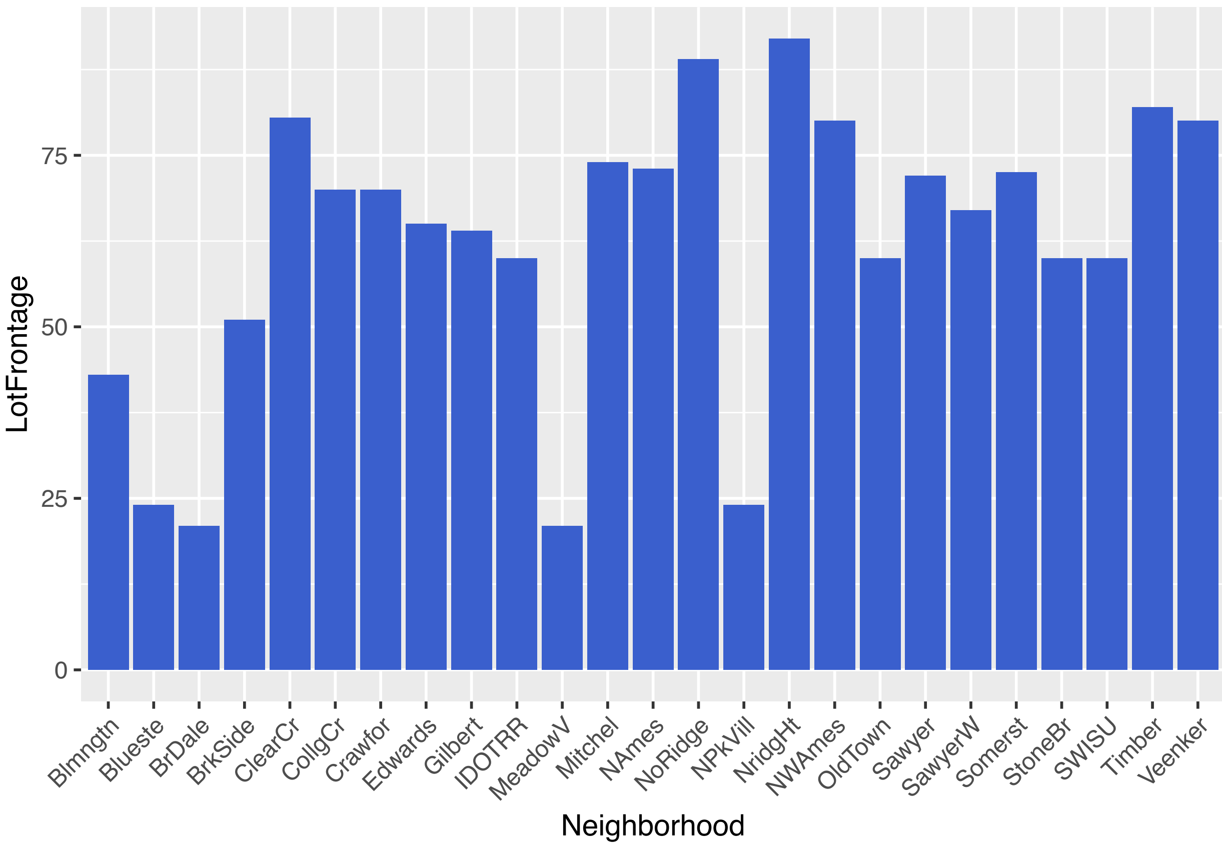

```
## < table of extent 0 >
```

Lot variables

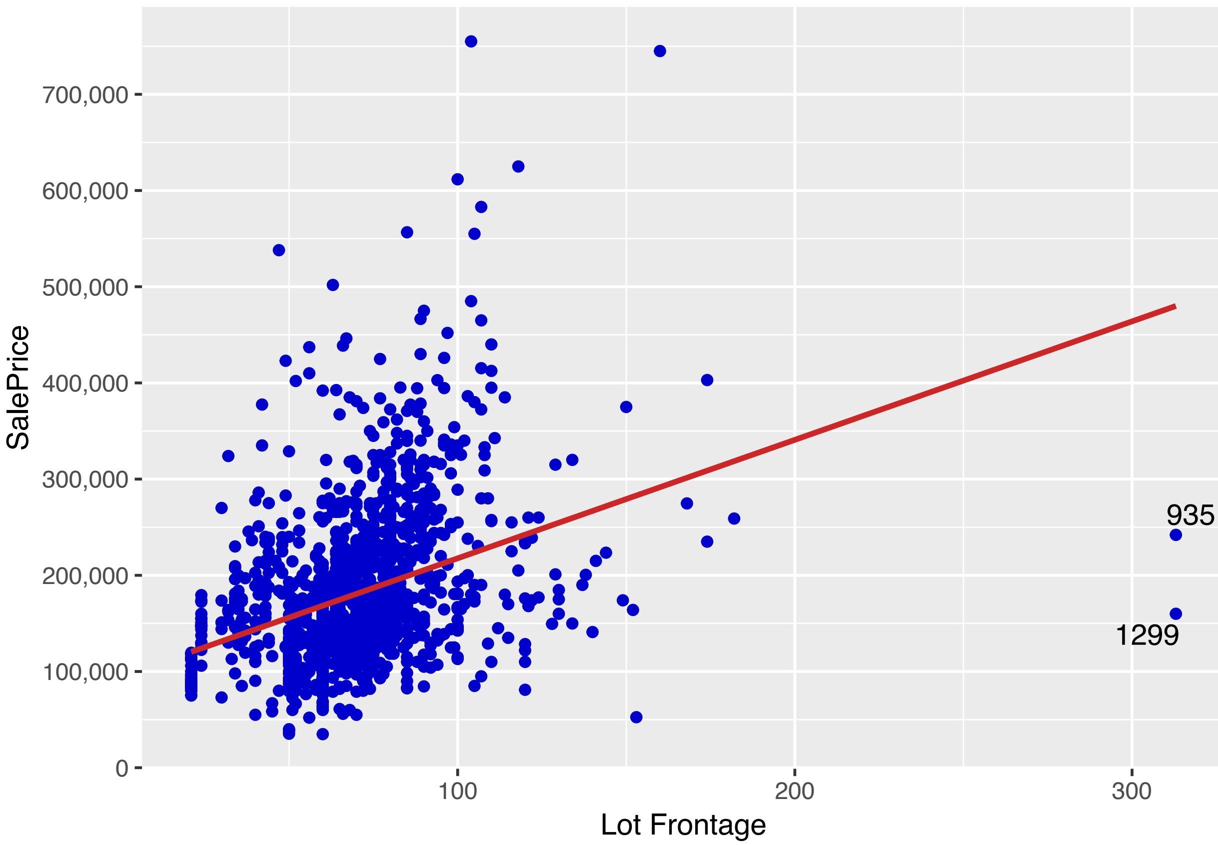
LotFrontage: Linear feet of street connected to property

Value Type: Numeric

486 NAs in this. Seems like these NAs root from lot frontage values actually not being recorded. To impute these, we can take the median of the lot frontage for each neighborhood.



A quick check to see how it varies with Sales Price:



Row numbers '935' and '1299' are potential outliers. Can be studied later to improve accuracy.

LotShape: General shape of property

```
Reg  Regular
IR1  Slightly irregular
IR2  Moderately Irregular
IR3  Irregular
```

** Value Type: Ordinal **

```
##
##    0    1    2    3
##   16   76  968 1859
```

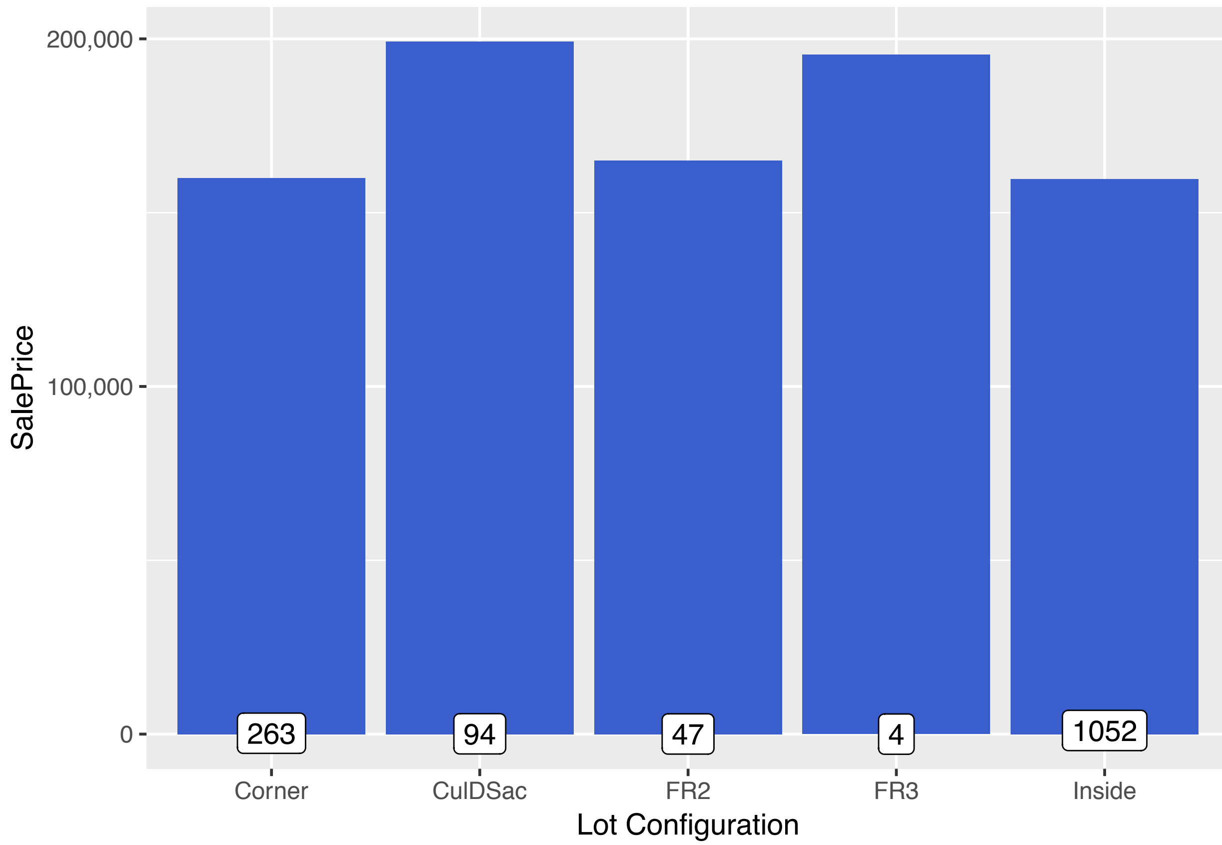
```
## [1] 2919
```

LotConfig: Lot configuration

No NAs.

```
Inside     Inside lot
Corner       Corner lot
CulDSac    Cul-de-sac
FR2        Frontage on 2 sides of property
FR3        Frontage on 3 sides of property
```

Value Type: Factor




```
##
##  Corner CulDSac     FR2     FR3  Inside
##     511     176      85      14    2133
```

```
## [1] 2919
```

Pool variables

The PoolQC is the variable with most NAs. The description is as follows:

PoolQC: Pool quality

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
NA   No Pool
```

Value Type: Ordinal The imputation for this would involve representing NAs as "None" - aka No Pool

Next w shall apply label encoding. Given many of the variables in this data set are assigned ratings as per the following format, it is worth storing the variable:

```
##
##    0    2    4    5
## 2909    2    4    4
```

Pool Area

Value Type: Numeric

```
##      PoolArea PoolQC OverallQual
## 2421      368      0           4
## 2504      444      0           6
## 2600      561      0           3
```

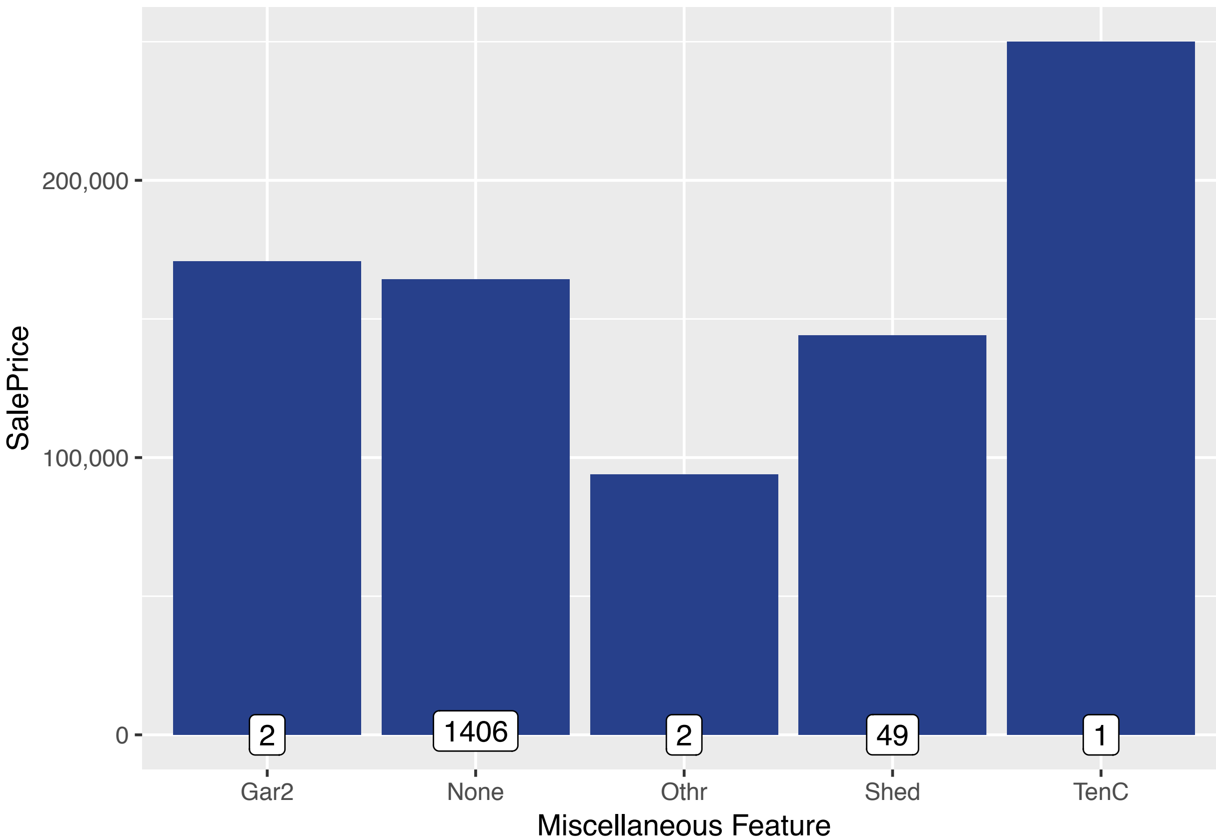
Miscellaneous Feature

Within Miscellaneous Feature, there are 2814 NAs. Values:

```
Elev   Elevator
Gar2   2nd Garage (if not described in garage section)
Othr   Other
Shed   Shed (over 100 SF)
TenC   Tennis Court
NA       None
```

Value Type: Factors (Not Ordinal)

```
##
## Gar2 None Othr Shed TenC
##    5 2814    4   95    1
```



Interesting observation is that the one house with a tennis court is the most expensive..

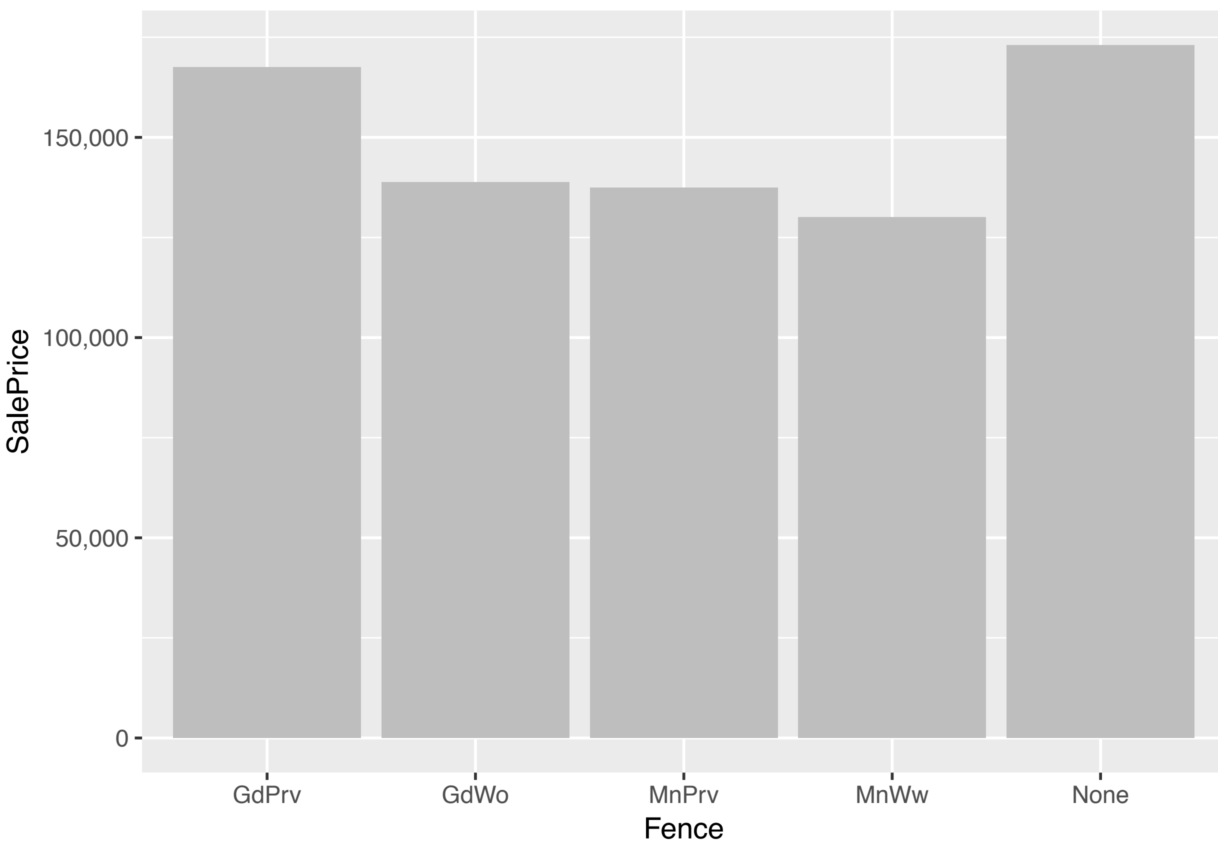
Fence

2348 NAs

```
GdPrv    Good Privacy
MnPrv    Minimum Privacy
GdWo   Good Wood
MnWw   Minimum Wood/Wire
NA       No Fence
```

Value Type: Ordinal

```
##
## GdPrv  GdWo MnPrv  MnWw  None
##   118   112   329    12  2348
```



This variable does not show much variation wrt Sales Price. Leads us to believe that it probably is not very important. Feature importance study to be conducted separately however.

Fireplace variables

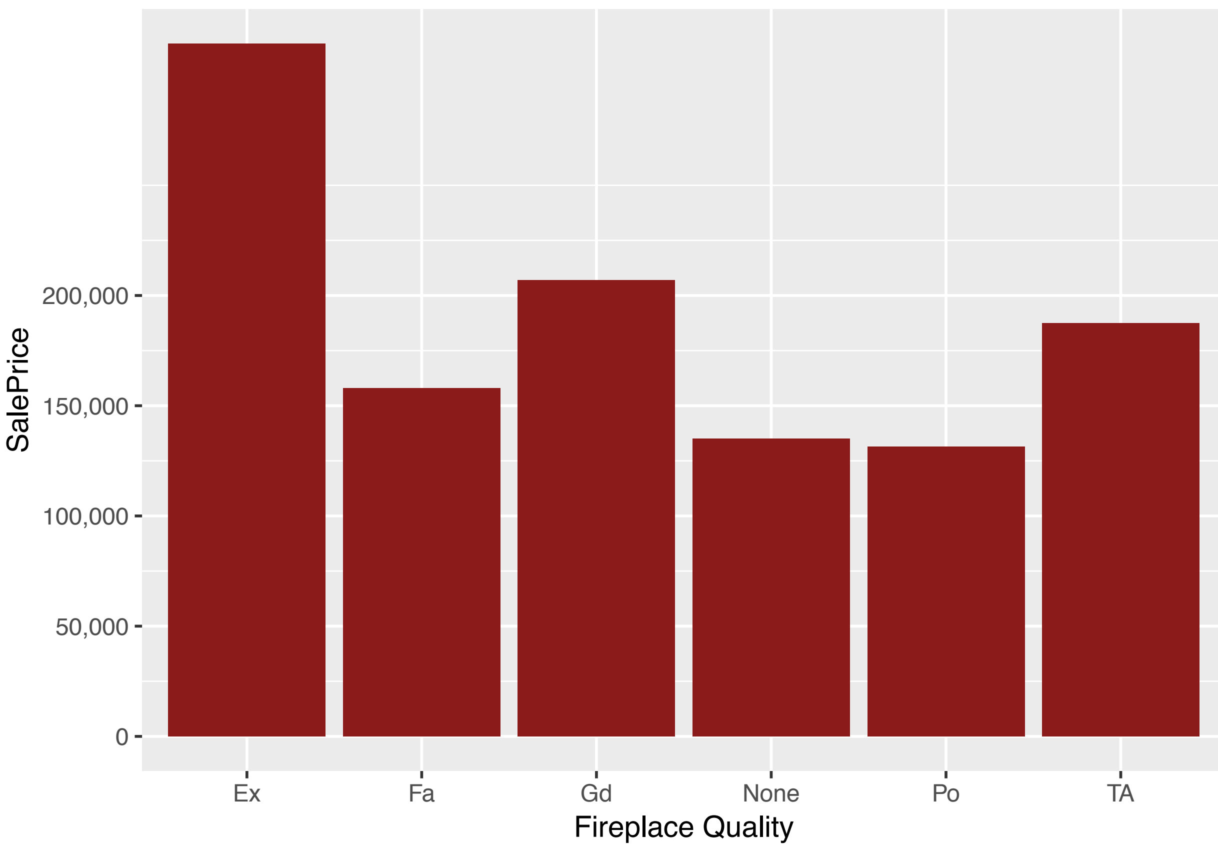
1420 NAs

Fireplace quality

The number of NAs in FireplaceQu matches the number of houses with 0 fireplaces. This means that I can safely replace the NAs in FireplaceQu with ‘no fireplace’. The values are ordinal, and I can use the Qualities vector that I have already created for the Pool Quality. Values:

```
Ex      Excellent - Exceptional Masonry Fireplace
Gd      Good - Masonry Fireplace in main level
TA      Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa      Fair - Prefabricated Fireplace in basement
Po      Poor - Ben Franklin Stove
NA      No Fireplace
```

Value Type: Ordinal



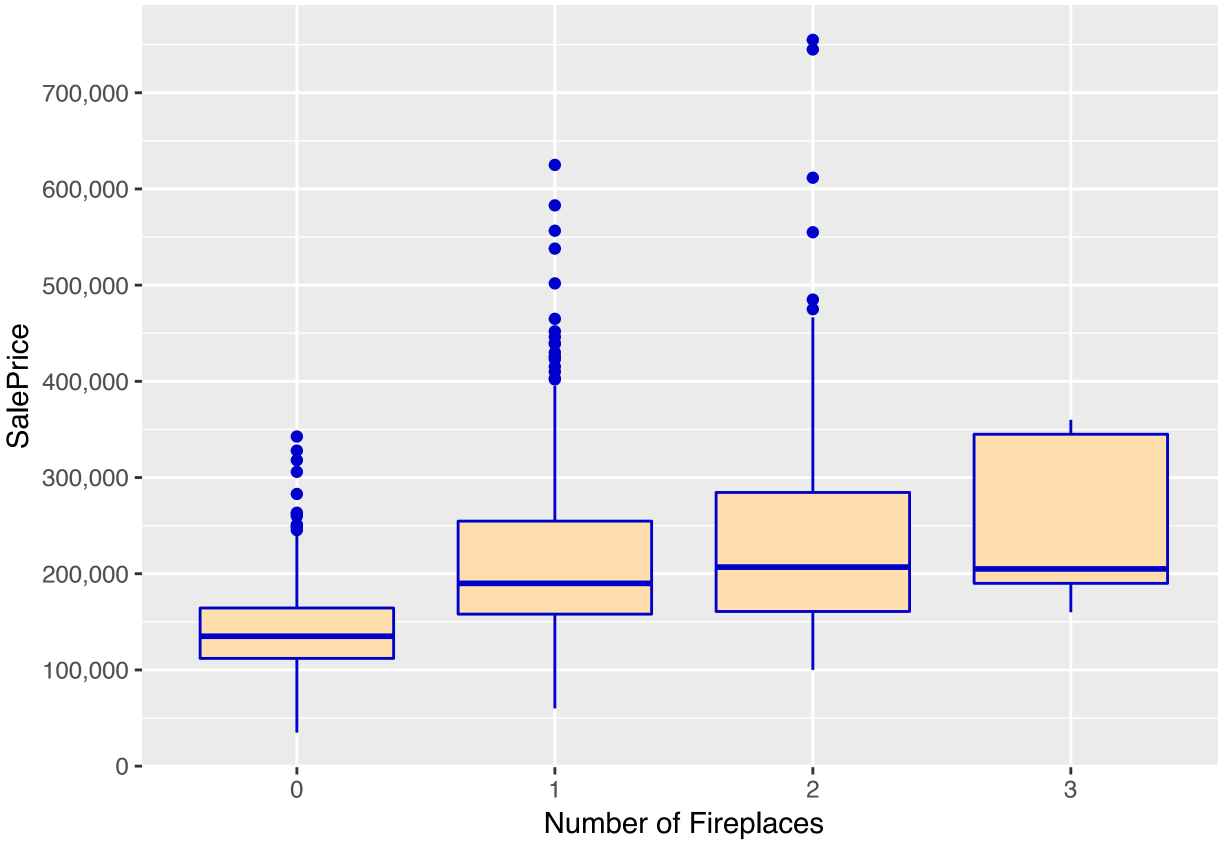

```
##
##    0    1    2    3    4    5
## 1420   46   74  592  744   43
```

It may be interpreted that the Sales Price increases with Fireplace quality as the Fireplace quality is indicative of overall quality. (We saw earlier that the Overall quality was the most important factor in determining the Sales Price)

Number of fireplaces

Value Type: Integer - No Missing Values

```
##
##    0    1    2    3    4
## 1420 1268  219   11    1
```




```
## [1] 2919
```

While there is some correlation with increased price with increasing number of fireplaces, it is not very strong. The variance and outliers also suggest that this isn't a very important factor.

Garage variables

7 total variables

```
GarageCars  - 1 NA
GarageArea  - 1 NA
GarageType  - 157 NAs
GarageYrBlt - 159 NAs
GarageCond  - 159 NAs
GarageQual  - 159 NAs
GarageFinish- 159 NAs
```

GarageYrBlt: Year garage was built Replacing 159 missing values with the values in YearBuilt. Some of the missing data implies that the Year Garage Built was not recorded and we can infer by the Year the house was built.

```
## [1] 157
```

	GarageCars	GarageArea	GarageType	GarageCond	GarageQual	GarageFinish
2127	1	360	Detchd	NA	NA	NA
2577	NA	NA	Detchd	NA	NA	NA

Imputing Modes for Garage Condition, Garage Quality and Garage Finish.

	GarageYrBlt	GarageCars	GarageArea	GarageType	GarageCond	GarageQual	GarageFinish
2127	1910	1	360	Detchd	TA	TA	Unf

GarageCars and GarageArea: Size of garage in car capacity and Size of garage in square

The remaining 4 character variables related to garage full have the same set of 158 NAs, which correspond to 'No Garage'.

GarageType: Garage location

```
2Types     More than one type of garage
Attchd     Attached to home
Basment  Basement Garage
BuiltIn  Built-In (Garage part of house - typicfully has room above garage)
CarPort  Car Port
Detchd     Detached from home
NA         No Garage
```

Value Type: Factor

```
##
##    2Types    Attchd   Basment   BuiltIn   CarPort    Detchd No Garage
##        23      1723        36       186        15       778       158
```

GarageFinish: Interior finish of the garage

```
Fin  Finished
RFn  Rough Finished
Unf  Unfinished
NA   No Garage
```

Value Type: Ordinal

```
##
##    0    1    2    3
##  158 1231  811  719
```

GarageQual: Garage quality

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

Value Type: Ordinal

```
##
##    0    1    2    3    4    5
##  158    5  124 2605   24    3
```

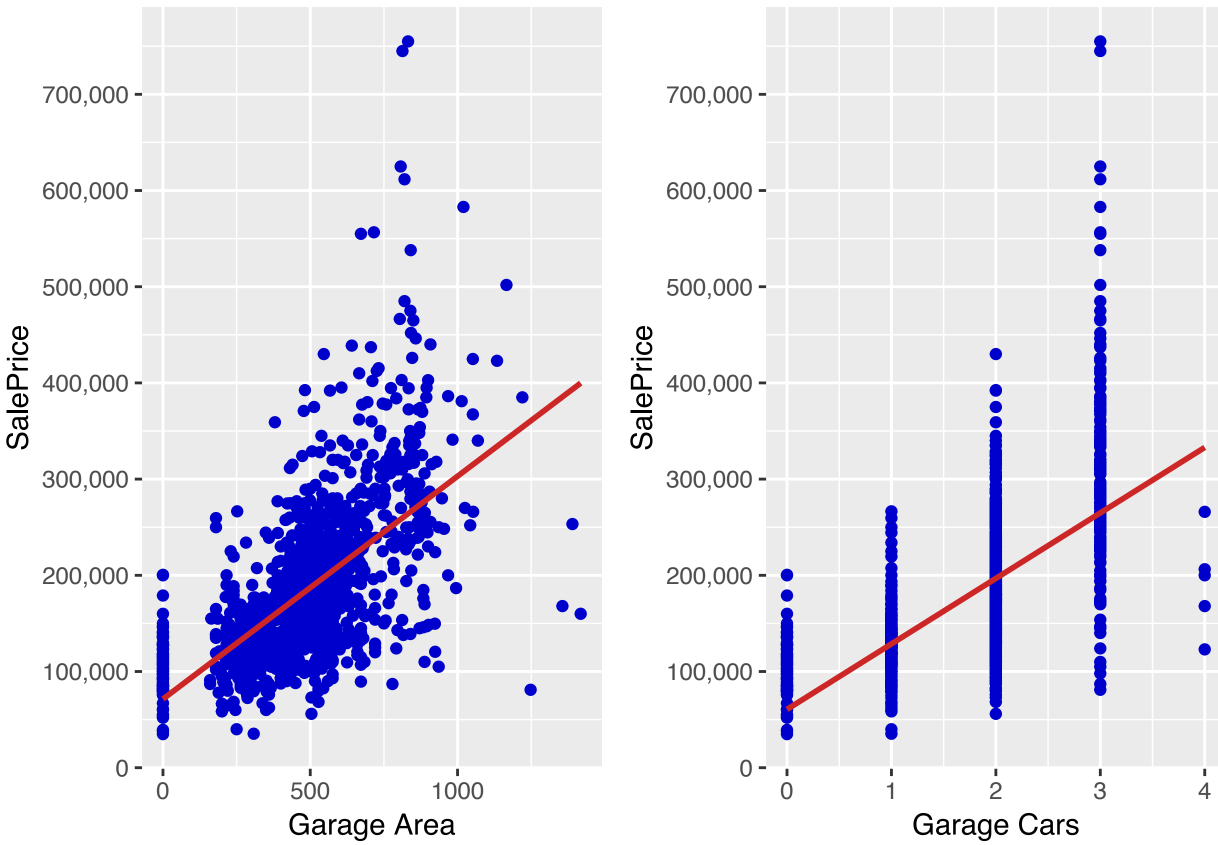
GarageCond: Garage condition

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

Value Type: Ordinal

```
##
##    0    1    2    3    4    5
##  158   14   74 2655   15    3
```

Plots for Garage Area and Garage Cars. In our initial correlation study, these variables showed quite a bit of importance. Lets see this visually:



The relationships with Sales Price between the two are very similar. The Correlation plot also showed a significant multicollinearity. We may consider eliminating one of these variables for certain models.

Basement Variables

11 Variables

```
## [1] 79
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2
## 333        Gd       TA           No          GLQ         <NA>
## 949        Gd       TA         <NA>          Unf          Unf
## 1488       Gd       TA         <NA>          Unf          Unf
## 2041       Gd     <NA>           Mn          GLQ          Rec
## 2186       TA     <NA>           No          BLQ          Unf
## 2218     <NA>       Fa           No          Unf          Unf
## 2219     <NA>       TA           No          Unf          Unf
## 2349       Gd       TA         <NA>          Unf          Unf
## 2525       TA     <NA>           Av          ALQ          Unf
```

There are 79 houses without a basement, because the basement variables of the other houses with missing values are full 80% complete (missing 1 out of 5 values). Imputging the modes to fix those 9 houses.

```r
#Imputing modes.
full$BsmtFinType2[333] <- names(sort(-table(full$BsmtFinType2)))[1]
full$BsmtExposure[c(949, 1488, 2349)] <- names(sort(-table(full$BsmtExposure)))[1]
full$BsmtCond[c(2041, 2186, 2525)] <- names(sort(-table(full$BsmtCond)))[1]
full$BsmtQual[c(2218, 2219)] <- names(sort(-table(full$BsmtQual)))[1]
```

Now that the 5 variables considered agree upon 79 houses with 'no basement', I am going to factorize/hot encode them below.

BsmtQual: Evaluates the height of the basement

```
   Ex   Excellent (100+ inches)
   Gd   Good (90-99 inches)
   TA   Typical (80-89 inches)
   Fa   Fair (70-79 inches)
   Po   Poor (<70 inches
   NA   No Basement
```

Value Type: Ordinal

```
##
##    0    2    3    4    5
##   79   88 1285 1209  258
```

BsmtCond: Evaluates the general condition of the basement

```
   Ex   Excellent
   Gd   Good
   TA   Typical - slight dampness fullowed
   Fa   Fair - dampness or some cracking or settling
   Po   Poor - Severe cracking, settling, or wetness
   NA   No Basement
```

Value Type: Ordinal

```
##
##    0    1    2    3    4
##   79    5  104 2609  122
```

BsmtExposure: Refers to walkout or garden level wfulls

```
Gd   Good Exposure
Av   Average Exposure (split levels or foyers typicfully score average or above)
Mn   Mimimum Exposure
No   No Exposure
NA   No Basement
```

Value Type: Ordinal

```
##
##    0    1    2    3    4
##   79 1907  239  418  276
```

BsmtFinType1: Rating of basement finished area

```
GLQ  Good Living Quarters
ALQ  Average Living Quarters
BLQ  Below Average Living Quarters
Rec  Average Rec Room
LwQ  Low Quality
Unf  Unfinshed
NA     No Basement
```

Value Type: Ordinal

```
##
##   0   1   2   3   4   5   6
##  79 851 154 288 269 429 849
```

BsmtFinType2: Rating of basement finished area (if multiple types)

```
GLQ  Good Living Quarters
ALQ  Average Living Quarters
BLQ  Below Average Living Quarters
Rec  Average Rec Room
LwQ  Low Quality
Unf  Unfinshed
NA     No Basement
```

Value Type: Ordinal

```
##
##    0    1    2    3    4    5    6
##   79 2494   87  105   68   52   34
```

Imputing all the remaining Basement variables with 'None' since they have two NAs

```
#display remaining NAs. Using BsmtQual as a reference for the 79 houses without basement agreed upon ea
full[(is.na(full$BsmtFullBath)|is.na(full$BsmtHalfBath)|is.na(full$BsmtFinSF1)|is.na(full$BsmtFinSF2)|i
```

```
##      BsmtQual BsmtFullBath BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
## 2121        0           NA           NA         NA         NA        NA
## 2189        0           NA           NA          0          0         0
##      TotalBsmtSF
## 2121          NA
## 2189           0
```

BsmtFullBath: Basement full bathrooms Value Type: Integer

```
##
##    0    1    2    3
```

```
## 1707 1172   38    2
```

BsmtHalfBath: Basement half bathrooms Value Type: Integer

```
##
##    0    1    2
## 2744  171    4
```

BsmtFinSF1: Type 1 finished square feet Value Type: Integer

BsmtFinSF2: Type 2 finished square feet Value Type: Integer

BsmtUnfSF: Unfinished square feet of basement area Value Type: Integer

TotalBsmtSF: Total square feet of basement area Value Type: Integer

An integer variable.

Checking the relationship between total basement square footage and Sale Price:

We have one potential outlier which may be considered for removal later.

Masonry variables

```
Masonry veneer type - 24 NAs.
Masonry veneer area  - 23 NAs.
```

```
## [1] 23
```

```
##      MasVnrType MasVnrArea
## 2611       <NA>        198
```

This particular row has an area but no type. Imputing the type by the mode.

```
##      MasVnrType MasVnrArea
## 2611    BrkFace        198
```

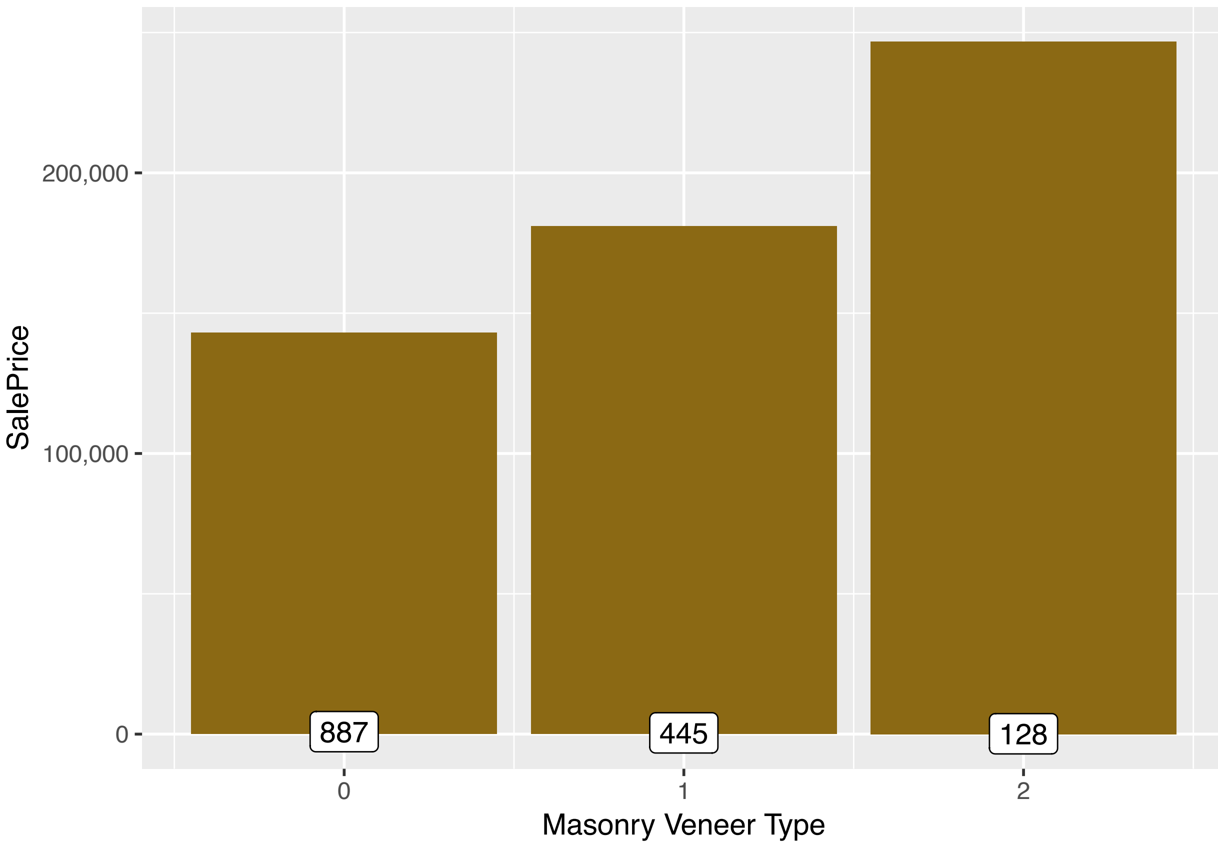
Masonry veneer type

```
   BrkCmn       Common Brick
   BrkFace    Face Brick
   CBlock       Cinder Block
   None       None
   Stone        Stone
```

Value Type: Ordinal Assiuming Ordinality by inferring quality of veneer type. ie. Stone is most expensive and Common Brick is cheapest. Also assuming that there is no difference in quality between Common Brick and None. Ordinalkty will be built accordingly.

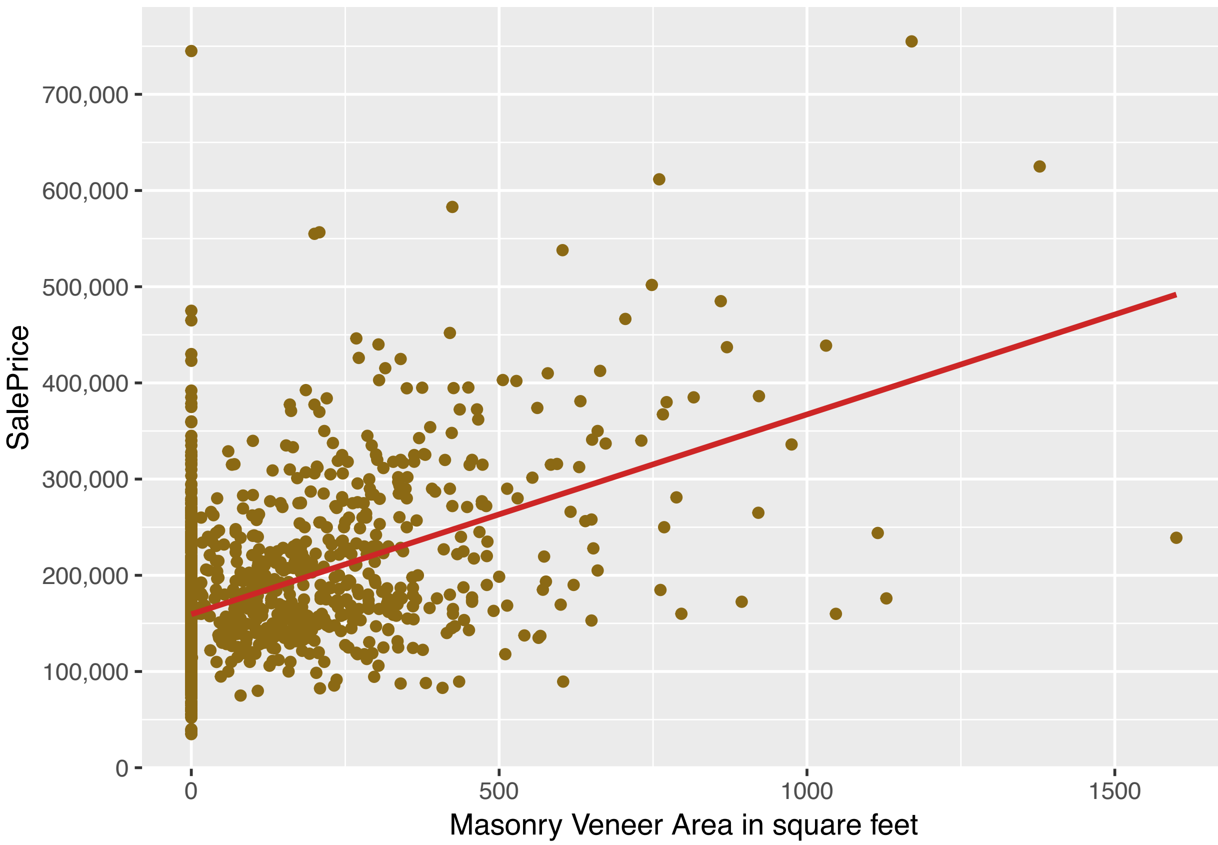
```
##
##    0    1    2
## 1790  880  249
```

Plot to prove ordinality



MasVnrArea: Masonry veneer area in square feet

Value Type: Integer Imputing NAs as integer '0'



700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
SalePrice
0
500
1000
1500
Masonry Veneer Area in square feet

MS Zoning

MSZoning identifies the general zoning classification of the sale

4 NAs

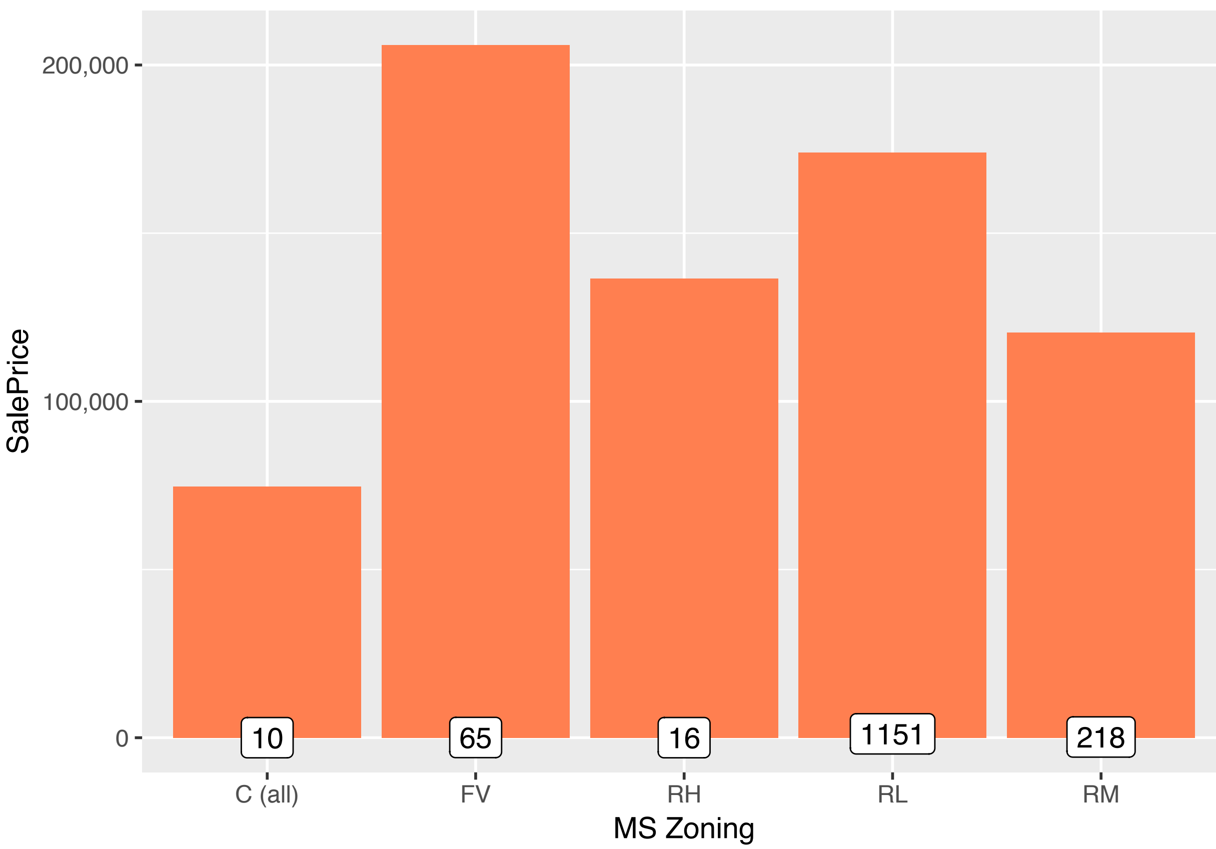
```
A    Agriculture
C    Commercial
FV   Floating Village Residential
I    Industrial
RH   Residential High Density
RL   Residential Low Density
RP   Residential Low Density Park
RM   Residential Medium Density
```

Value Type: Factor

Imputing NA's with overall mode since there are only 4 NAs.

```
##
## C (all)      FV      RH      RL      RM
##      25     139      26    2269     460
```

```
## [1] 2919
```



Kitchen variables

Kitchen quality and number of Kitchens above grade

1 NA

Kitchen quality

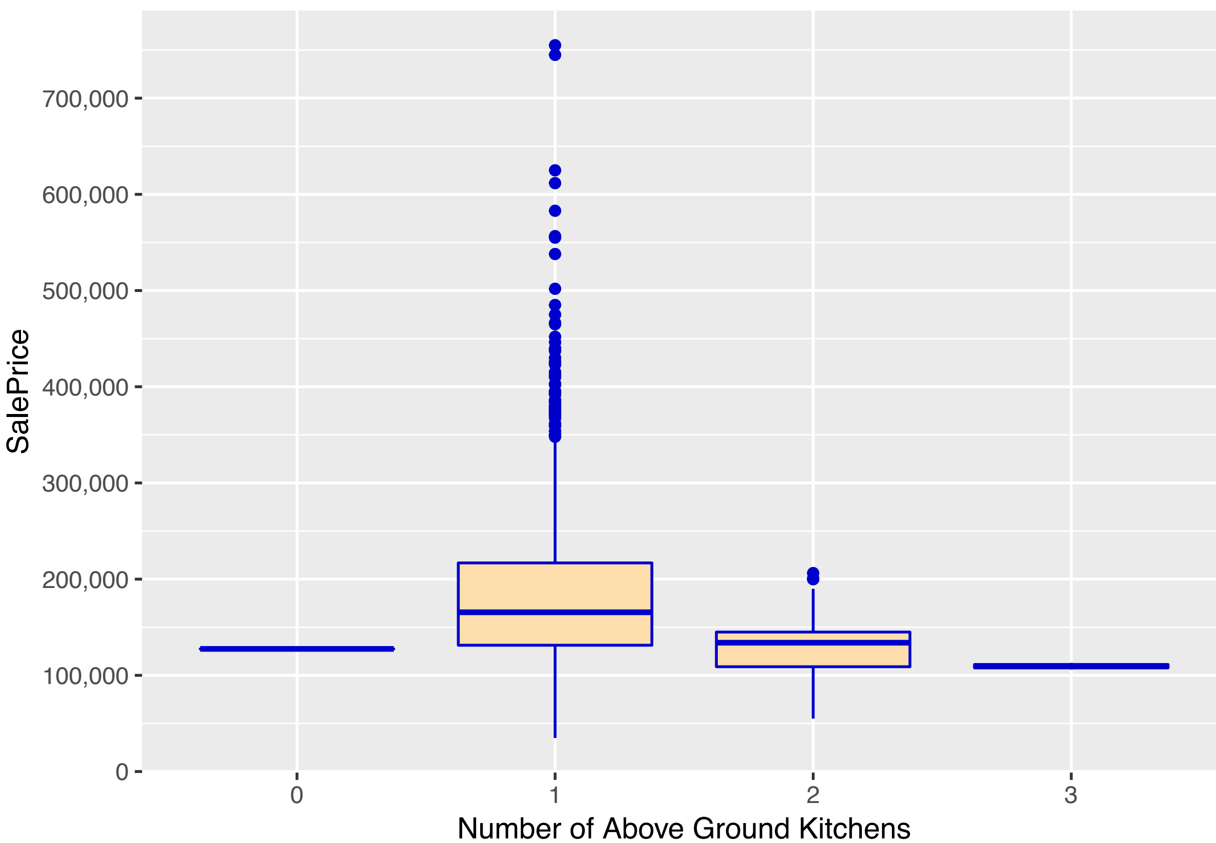
```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
```

Value Type: Ordinal

```
##
##    2    3    4    5
##   70 1493 1151  205
```

```
## [1] 2919
```

Number of Kitchens above grade No NAs.




```
##
##    0    1    2    3
##    3 2785  129    2
```

```
## [1] 2919
```

This doesn't prove to be a useful variable

Utilities

Utilities: Type of utilities available

2 NAs

```
fullPub  full public Utilities (E,G,W,& S)
NoSewr     Electricity, Gas, and Water (Septic Tank)
NoSeWa     Electricity and Gas Only
ELO      Electricity only
```

Value Type: Ordinal

```
##
## AllPub NoSeWa
##   2916      1
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
945	20	RL	82	14375	Pave	None	2	Lvl	NoSeWa
1916	30	RL	109	21780	Grvl	None	3	Lvl	NA
1946	20	RL	64	31220	Pave	None	2	Bnk	NA

The above table shows that only one house in the entire dataset has no full public utilities. This means that the variable will be useless for prediction (no variance at all). It has been removed as a result.

Home functionality

1 NA - Impute with mode **Functional: Home functionality**

```
Typ    Typical Functionality
Min1   Minor Deductions 1
Min2   Minor Deductions 2
Mod    Moderate Deductions
Maj1   Major Deductions 1
Maj2   Major Deductions 2
Sev    Severely Damaged
Sal    Salvage only
```

Value Type: Ordinal

```
##
##    1    2    3    4    5    6    7
##    2    9   19   35   70   65 2719
```

```
## [1] 2919
```

Exterior variables

4 exterior variables.

Exterior1st: Exterior covering on house

1 NA - Impute by mode

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock     Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other      Other
Plywood  Plywood
PreCast  PreCast
Stone      Stone
Stucco     Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

Value Type: Factor

```
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd
##      44       2       6      87       2     126     442       1     450
## Plywood   Stone  Stucco VinylSd Wd Sdng WdShing
##     221       2      43    1026     411      56
```

```
## [1] 2919
```

Exterior2nd: Exterior covering on house (if more than one material)

1 NA - Impute by mode

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock     Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other      Other
Plywood  Plywood
PreCast  PreCast
Stone      Stone
Stucco     Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

Value Type: Factor

```
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd
##      38       4      22      47       3     126     406      15     447
##   Other Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng
##       1     270       6      47    1015     391      81
```

```
## [1] 2919
```

ExterQual: Evaluates the quality of the material on the exterior

No NA

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

Value Type: Ordinal

```
##
##    2    3    4    5
##   35 1798  979  107
```

```
## [1] 2919
```

ExterCond: Evaluates the present condition of the material on the exterior

No NAs.

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

Value Type: Ordinal

```
##
##    1    2    3    4    5
##    3   67 2538  299   12
```

```
## [1] 2919
```

Electrical system

Electrical: Electrical system

1 NA - Impute by mode

```
SBrkr    Standard Circuit Breakers & Romex
FuseA    Fuse Box over 60 AMP and full Romex wiring (Average)
FuseF    60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP    60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix    Mixed
```

Value Type: Factor

```
##
## FuseA FuseF FuseP   Mix SBrkr
##   188    50     8     1  2672
```

```
## [1] 2919
```

Sale Type and Condition

SaleType: Type of sale

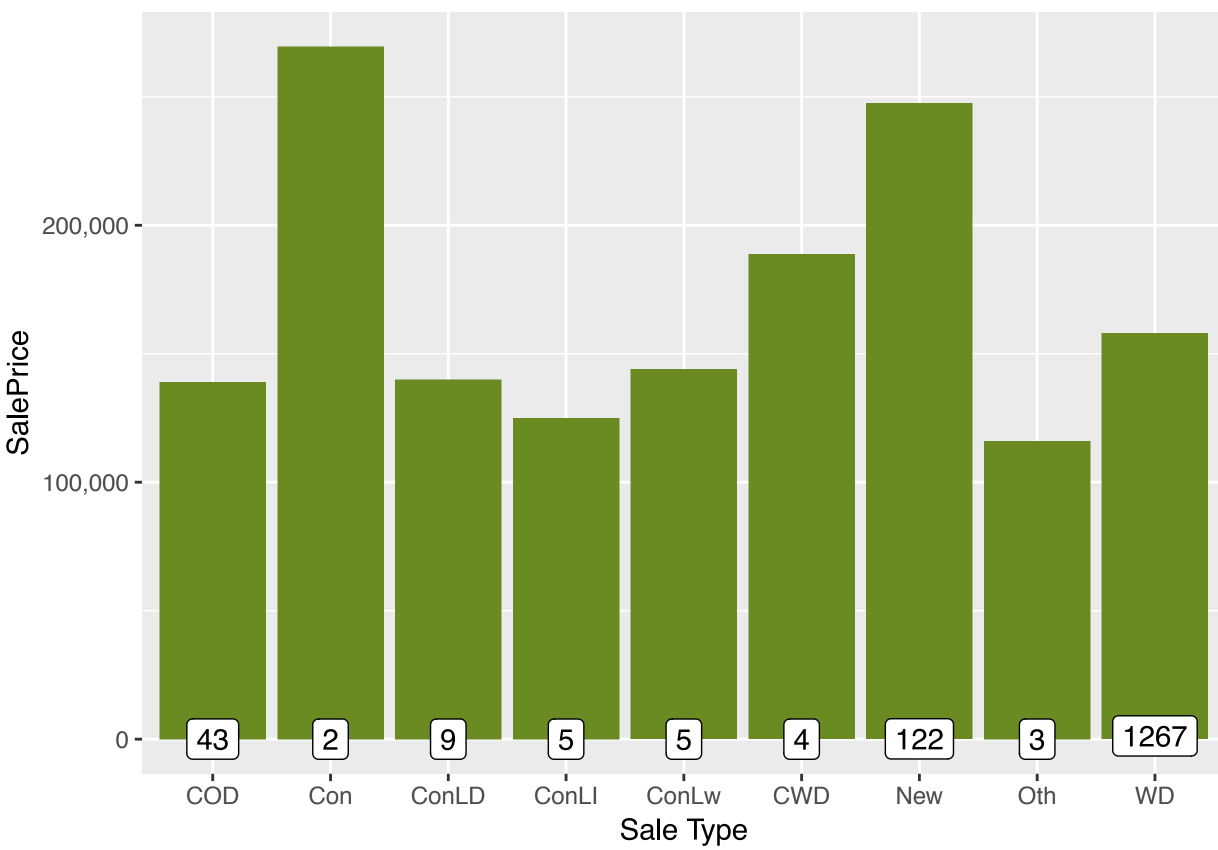
1 NA

```
WD      Warranty Deed - Conventional
CWD     Warranty Deed - Cash
VWD     Warranty Deed - VA Loan
New     Home just constructed and sold
COD     Court Officer Deed/Estate
Con     Contract 15% Down payment regular terms
ConLw     Contract Low Down payment and low interest
ConLI     Contract Low Interest
ConLD     Contract Low Down
Oth     Other
```

Value Type: Factor

```
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD
##    87     5    26     9     8    12   239     7  2526
```

```
## [1] 2919
```



SaleCondition: Condition of sale

No NAs

```
Normal     Normal Sale
```

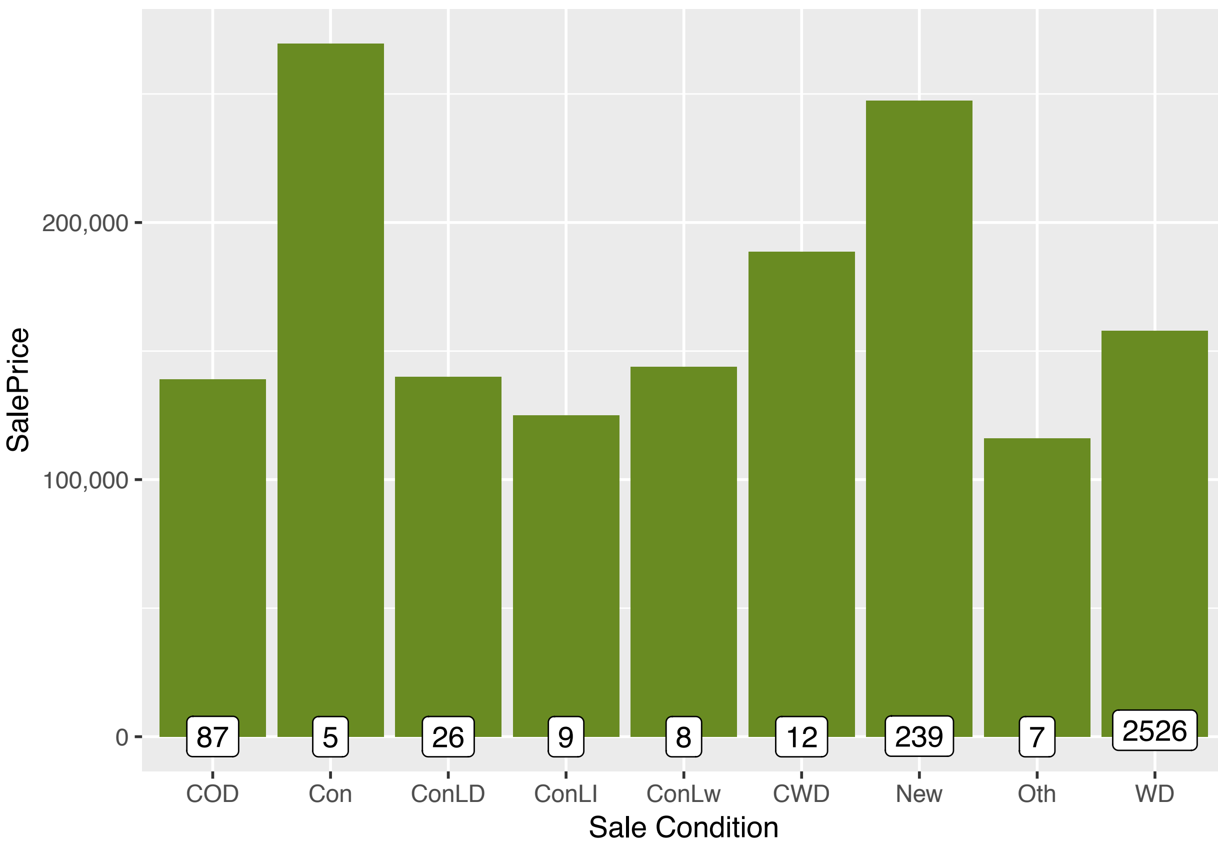
```
Abnorml  Abnormal Sale -  trade, foreclosure, short sale
AdjLand  Adjoining Land Purchase
fulloca  fullocation - two linked properties with separate deeds, typicfully condo with a garage uni
Family     Sale between family members
Partial  Home was not completed when last assessed (associated with New Homes)
```

Value Type: Factor

```
##
## Abnorml AdjLand  Alloca  Family  Normal Partial
##     190      12      24      46    2402     245
```

```
## [1] 2919
```

```
## Warning: Removed 1459 rows containing non-finite values (stat_summary).
```



This covers all the variables which contain NAs. Imputation has been complete. We shall do a label encoding for all the remaining character variables.

```
##  [1] "Street"       "LandContour"  "LandSlope"    "Neighborhood"
##  [5] "Condition1"   "Condition2"   "BldgType"     "HouseStyle"
##  [9] "RoofStyle"    "RoofMatl"     "Foundation"   "Heating"
## [13] "HeatingQC"    "CentralAir"   "PavedDrive"
```

```
## There are 15 remaining columns with character values
```

Foundation

Foundation: Type of foundation

```
BrkTil        Brick & Tile
CBlock        Cinder Block
PConc           Poured Contrete
Slab            Slab
Stone           Stone
Wood            Wood
```

Value Type: Factor

```
##
## BrkTil CBlock  PConc   Slab  Stone   Wood
##    311   1235   1308     49     11      5
```

```
## [1] 2919
```

Heating and Air Conditioning

Heating: Type of heating

```
Floor      Floor Furnace
GasA     Gas forced warm air furnace
GasW     Gas hot water or steam heat
Grav     Gravity furnace
OthW     Hot water or steam heat other than gas
Wfull      With full furnace
```

Value Type: Factor

```
##
## Floor  GasA  GasW  Grav  OthW  Wall
##     1  2874    27     9     2     6
```

```
## [1] 2919
```

HeatingQC: Heating quality and condition

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

Value Type: Ordinal

```
##
##    1    2    3    4    5
##    3   92  857  474 1493
```

```
## [1] 2919
```

CentralAir: Central air conditioning

```
N    No
Y    Yes
```

Value Type: Ordinal

```
##
##    0    1
##  196 2723
```

```
## [1] 2919
```

Roof

RoofStyle: Type of roof

```
Flat     Flat
Gable    Gable
Gambrel  Gabrel (Barn)
Hip      Hip
Mansard  Mansard
Shed     Shed
```

Value Type: Factor

```
##
##     Flat    Gable Gambrel      Hip Mansard     Shed
##       20     2310      22      551      11        5
```

```
## [1] 2919
```

RoofMatl: Roof material

```
ClyTile  Clay or Tile
CompShg  Standard (Composite) Shingle
Membran  Membrane
Metal      Metal
Roll     Roll
Tar&Grv  Gravel & Tar
WdShake  Wood Shakes
WdShngl  Wood Shingles
```

Value Type: Factor

```
##
## ClyTile CompShg Membran   Metal    Roll Tar&Grv WdShake WdShngl
##       1    2876       1       1       1      23       9       7
```

```
## [1] 2919
```

Land

LandContour: Flatness of the property

```
Lvl  Near Flat/Level
Bnk  Banked - Quick and significant rise from street grade to building
HLS  Hillside - Significant slope from side to side
Low  Depression
```

Value Type: Factor

```
##
##  Bnk  HLS  Low  Lvl
##  117  120   60 2622
```

```
## [1] 2919
```

LandSlope: Slope of property

```
Gtl  Gentle slope
Mod  Moderate Slope
Sev  Severe Slope
```

Value Type: Ordinal

```
##
##    0    1    2
##   16  125 2778
```

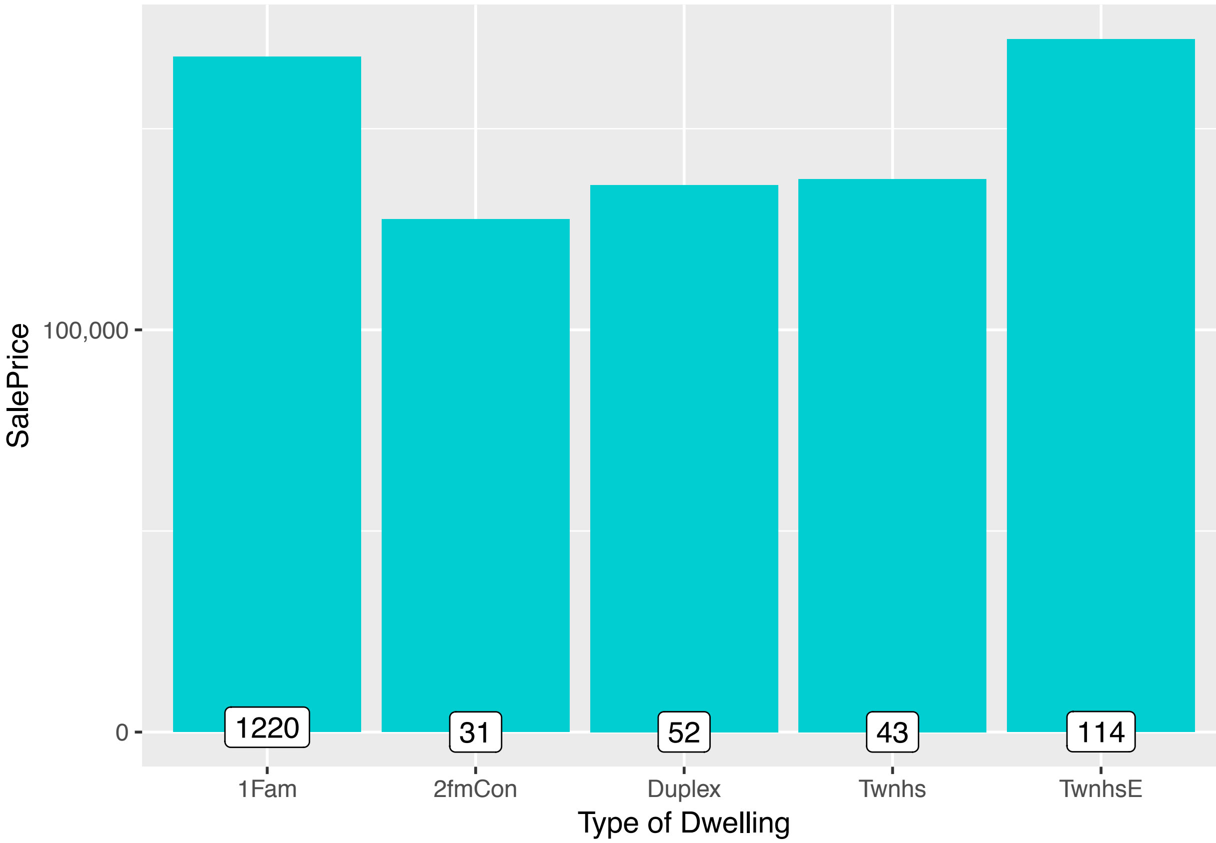
```
## [1] 2919
```

Dwelling

BldgType: Type of dwelling

```
1Fam Single-family Detached
2FmCon   Two-family Conversion; originfully built as one-family dwelling
Duplx    Duplex
TwnhsE   Townhouse End Unit
TwnhsI   Townhouse Inside Unit
```

Value Type: Factor




```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   2425     62    109     96    227
```

```
## [1] 2919
```

HouseStyle: Style of dwelling

```
1Story   One story
1.5Fin   One and one-half story: 2nd level finished
1.5Unf   One and one-half story: 2nd level unfinished
2Story   Two story
2.5Fin   Two and one-half story: 2nd level finished
2.5Unf   Two and one-half story: 2nd level unfinished
SFoyer   Split Foyer
SLvl   Split Level
```

Value Type: Factor

```
##
## 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer   SLvl
##    314     19   1471      8     24    872     83    128
```

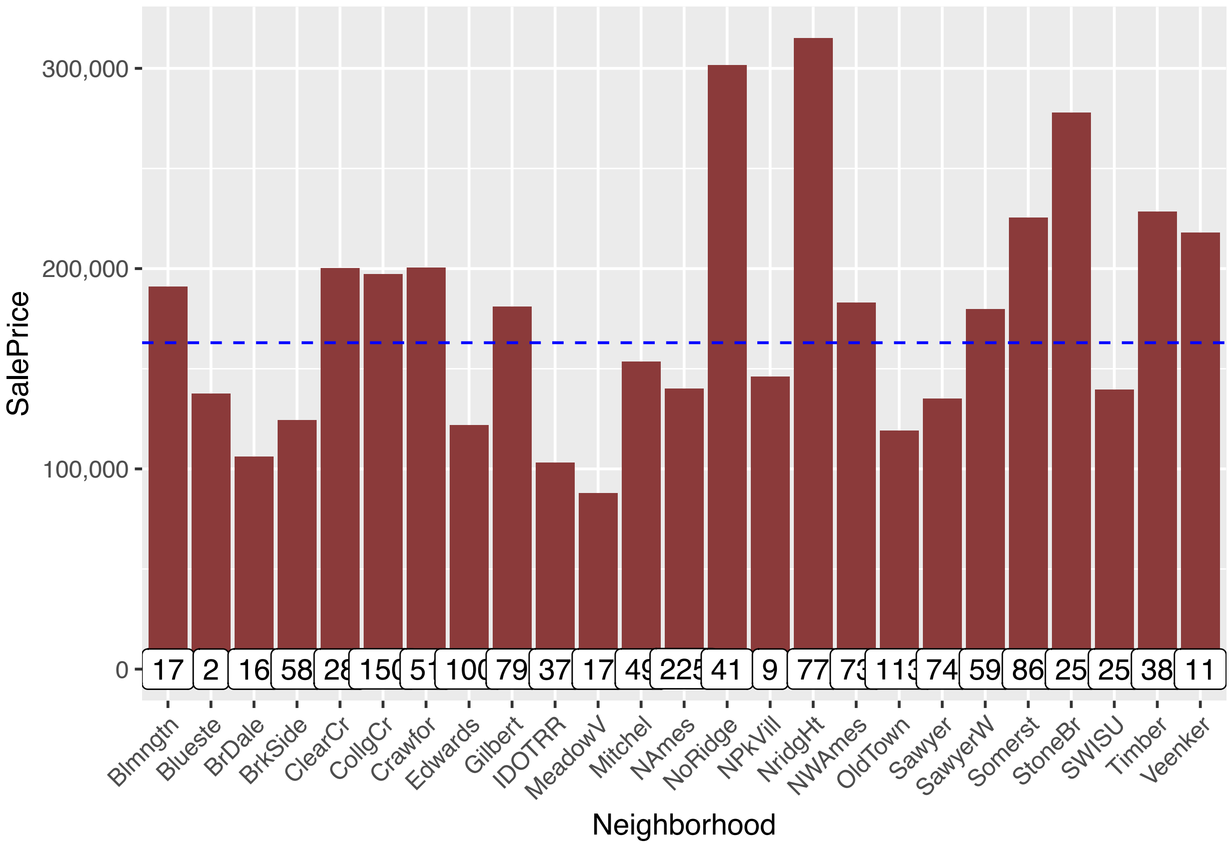
```
## [1] 2919
```

Neighborhood and Conditions

Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Value Type: Factor




```
##
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert
##      28      10      30     108      44     267     103     194     165
##  IDOTRR MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown
##      93      37     114     443      71      23     166     131     239
##  Sawyer SawyerW Somerst StoneBr   SWISU  Timber Veenker
##     151     125     182      51      48      72      24
```

```
## [1] 2919
```

Condition1: Proximity to various conditions

```
Artery     Adjacent to arterial street
Feedr      Adjacent to feeder street
Norm     Normal
RRNn     Within 200' of North-South Railroad
RRAn     Adjacent to North-South Railroad
PosN     Near positive off-site feature--park, greenbelt, etc.
PosA     Adjacent to postive off-site feature
RRNe     Within 200' of East-West Railroad
RRAe     Adjacent to East-West Railroad
```

Value Type: Factor

```
##
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNe   RRNn
##     92    164   2511     20     39     28     50      6      9
```

```
## [1] 2919
```

Condition2: Proximity to various conditions (if more than one is present)

```
Artery     Adjacent to arterial street
Feedr      Adjacent to feeder street
Norm     Normal
RRNn     Within 200' of North-South Railroad
RRAn     Adjacent to North-South Railroad
PosN     Near positive off-site feature--park, greenbelt, etc.
PosA     Adjacent to postive off-site feature
RRNe     Within 200' of East-West Railroad
RRAe     Adjacent to East-West Railroad
```

Value Type: Factor

```
##
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNn
##      5     13   2889      4      4      1      1      2
```

```
## [1] 2919
```

Pavement of Street & Driveway

Street: Type of road access to property

```
Grvl Gravel
Pave Paved
```

Value Type: Ordinal

```
##
##    0    1
##   12 2907
```

```
## [1] 2919
```

PavedDrive: Paved driveway

```
Y    Paved
P    Partial Pavement
N    Dirt/Gravel
```

Value Type: Ordinal

```
##
##    0    1    2
##  216   62 2641
```

```
## [1] 2919
```

MSSubClass

MSSubClass: Identifies the type of dwelling involved in the sale.
Provided as numbers in the data set but are actually categorical variables. Need to be changed

```
    20  1-STORY 1946 & NEWER fuLL STYLES
    30  1-STORY 1945 & OLDER
    40  1-STORY W/FINISHED ATTIC fuLL AGES
    45  1-1/2 STORY - UNFINISHED fuLL AGES
    50  1-1/2 STORY FINISHED fuLL AGES
    60  2-STORY 1946 & NEWER
    70  2-STORY 1945 & OLDER
    75  2-1/2 STORY fuLL AGES
    80  SPLIT OR MULTI-LEVEL
    85  SPLIT FOYER
    90  DUPLEX - fuLL STYLES AND AGES
   120  1-STORY PUD (Planned Unit Development) - 1946 & NEWER
   150  1-1/2 STORY PUD - fuLL AGES
   160  2-STORY PUD - 1946 & NEWER
   180  PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
   190  2 FAMILY CONVERSION - fuLL STYLES AND AGES
```

Value Type: Factor

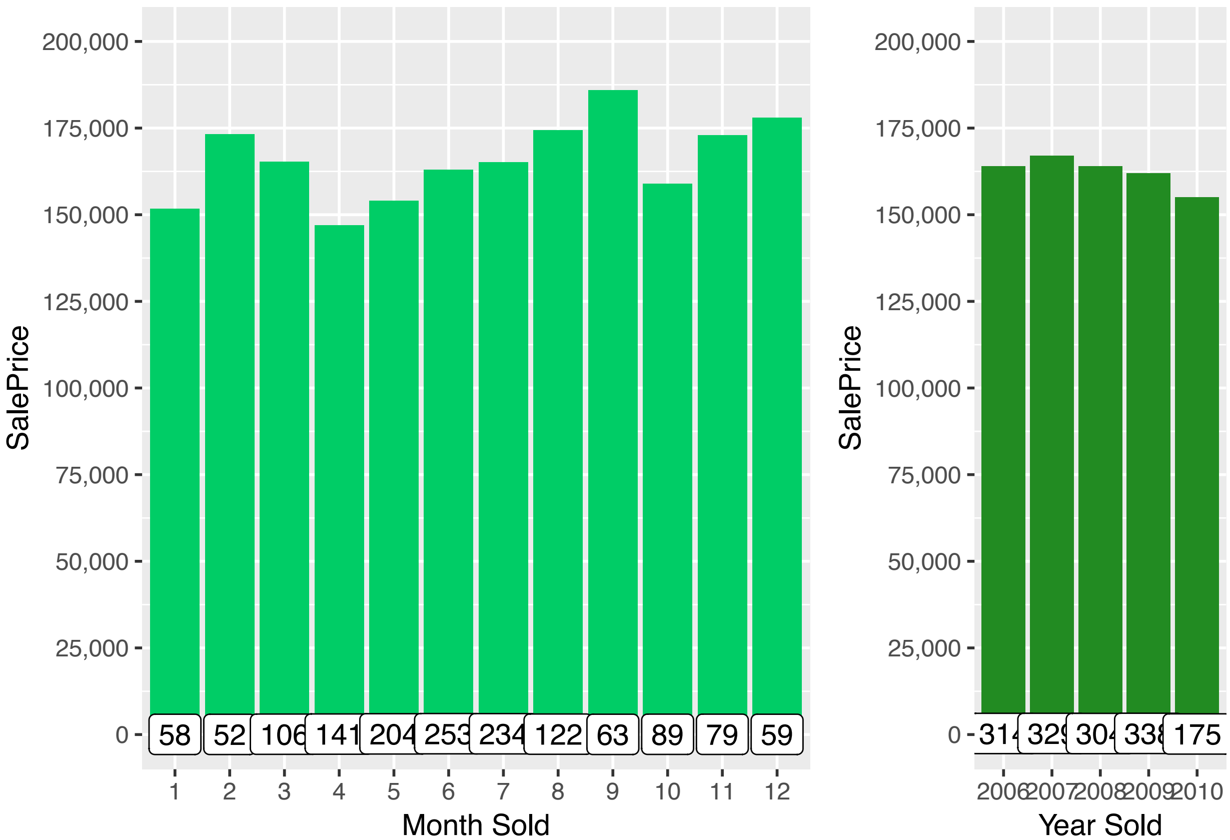
```
##  int [1:2919] 60 20 60 70 60 50 20 60 50 190 ...
```

```
##  Factor w/ 16 levels "1 story 1946+",..: 6 1 6 7 6 5 1 6 5 16 ...
```

Year and Month Sold

Changing Year and Month from Numerical to Factors

```
##  int [1:2919] 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  int [1:2919] 2 5 9 2 12 10 8 11 4 1 ...
```



On the lookout for the housing crisis of 2007. There seems to be a slight drop, but not as dramatic as expected.

The months Graph also shows that the summer season is the best sale price.