Multiple Linear Regression on Ames Housing Data

Stat 632

Zhaoshan Duan

May 12, 2021

Contents

1 Abstract **2**

2 Problem and Motivation **3**
2.1 Background of the Dataset 3
2.2 Motivation 3

3 Data Description **4**

4 Question of Interest **4**

5 Exploratory Data Analysis **5**
5.1 Missingness 5
5.2 Response Variable `Sale_Price` 9
5.3 Numeric Predictors 11

6 Regression Analysis, Results and Interpretation **13**
6.1 Feature Engineering 13
6.2 Variable Selection 13
6.3 Diagnostics 15
6.4 Results 20
6.5 Interpretation 23

7 Conclusion **25**

8 Appendices **25**
8.1 R scripts 25

Reference **26**

1 Abstract

This project uses the a dataset that describes residential property sales in Ames, Iowa between 2006 and 2021. The dataset contains 2970 observations and 81 variables describing nearly every aspect of a residential property. We are interested in investigating what fixed characteristics influence the sale price of a property in Ames, Iowa using multiple linear regression. Through various variable selection techniques, we reduced the amount of predictors in our model to 10. Although the model explains 81.6% of the variability in sale price, it is more likely that we are overfitting the model.

We concluded that in Ames, Iowa, a house is more valuable with bigger living area and better overall condition. Having a basement would decrease the home value according to our model. The fixed characteristics that influences the sale price of a house the most is the overall quality of the materials and finish, and the size of the living area.

We also find that multiple linear regression may not be the best regression model for this dataset since the normality assumption could not be improved. This finding agrees with the conventional regression model used for real state pricing - hedonic regression method, a variation of Lasso regression.

2 Problem and Motivation

2.1 Background of the Dataset

This project uses the Ames Housing Dataset by Sean De Cock (2011), a contemporary alternative to the well known Boston Housing dataset. The dataset describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The original dataset contains 2930 observations and 82 explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous). Most of the explanatory variables are information that a home buyer typically sought out when purchasing properties.

The continuous numeric variables of the dataset describe the various area dimensions for each property and the discrete numeric variables quantify the number of typical items within the house such as amount of bedrooms. The nominal categorical variables identify types of dwellings, garages, materials and environmental conditions while ordinal categorical variables rate items within the property. (Dean De Cock 2011)

2.2 Motivation

For many people, homeownership is a both a dream and an achievement. It is a serious purchasing decision that requires meticulous research and careful Pro & Con analysis. Accurate prediction of housing prices provides great help in buyer's decision-making process and informs them what characteristics of the properties generally affect the prices. The same can be said for realtors, sellers and developers. Housing prices also reflect the health of the economy, which can be insightful for police makers.

While there are many external factors influencing the housing price of a given property such as crime rate in the region, proximity to public schools, hospitals and busy areas, fixed characteristics of the house are often the first things people look at. Therefore, we think it will be interesting to examine the sale price of the property using a multiple linear regression model with explanatory variables that contain information about many common aspects of the house such as number of bathrooms, size of the basement and so on. It would also be fascinating to observe the quantitative relationship between these characteristics and the sale price, and see which ones have the most influence as well as its implication to buyers' purchasing behavior.

3 Data Description

We use the `AmesHousing` package on CRAN to access the data. The package provides two version of the data: `ames_raw` and processed `ames`. Our preprocessing and analysis are done on the processed version as it removes unique identifiers such as `Order` and `PID` , arranges all factors unordered, and engineers features with large missing values. This results in a dataset with 2930 observations and 81 explanatory variables. Prior to analysis, we removed 168 observations since they are of non-residential types as indicated in the table below.

MS_Zoning	n
Floating_Village_Residential	139
Residential_High_Density	27
Residential_Low_Density	2273
Residential_Medium_Density	462
A_agr	2
C_all	25
I_all	2

The response variable is `Sale_Price` and our predictors and the description of the predictors in the final model are in listed in the Table 1.

Table 1: Predictors in the Final Model

Variable Name	Description
`Total_Area`	Total area (including basement)
`Total_Bsmt_SF`	Total square feet of basement area
`Garage_Area`	Size of garage in square feet
`Overall_Qual_Good`	Good overall material and finish of the house
`Overall_Qual_Poor`	Poor overall material and finish of the house
`TotRms_AbvGrd_15`	The number of total rooms above grade (does not include bathrooms) that is 15
`Bsmt_Qual_Typical`	Typical quality and height of the basement
`Bsmt_Qual_Good`	Good quality and height of the basement
`Bsmt_Qual_No_Basement`	Property that does not have basement
`Year_Built`	The year the property was built

4 Question of Interest

Our primary question of interest is to identify what fixed characteristics of a property has the largest effect on the sale price in Ames, Iowa. Secondarily, we are interested in knowing what types of a home would have lower market values in Ames, Iowa.

5 Exploratory Data Analysis

In this section, we investigate access the missingness of the dataset, create new variables that could help us answer our research question, examine summary statistics on the response variable, correlation between the numeric predictors. We also record features and observations that are removed from our analysis in this section.

5.1 Missingness

We first investigate the missing values of the dataset using `DataExplorer` package. From Figure 1, it can be observed that all the missing values have been imputed to 0 since this project is using the processed version of the data.

This is somewhat misleading since some of the numerical variables have mostly 0 values as Figure 2 indicates. Hence, We remove the following numeric variables: `BsmtFin_SF_2`, `Enclosed_Porch`, `Low_Qual_Fin_SF`, `Misc_Val`, `Pool_Area`, `Screen_Porch`, `Three_season_porch`, `Bsmt_Half_Bath`, `Kitchen_AbvGr`, `Open_Porch_SF`

Additionally, some of the categorical variables have majority of the data concentrated in one level as Figure 3 indicated. Therefore we remove the following categorical variables: `Utilities`, `Street`, `Alley`, `Roof_Matl`, `Land_Contour`, `Land_Slope`, `Condition_1`, `Condition_2`, `Heating`, `Functional`, `Pool_QC`, `Bsmt_Cond`, `Misc_Feature`, `Central_Air`, `Electrical`, `Garage_Qual`, `Garage_Cond`, `Paved_Drive`. We also remove `Longtitude` and `Latitude` as they are computationally expensive (each with 2700+ levels) and irrelevant to our research questions.

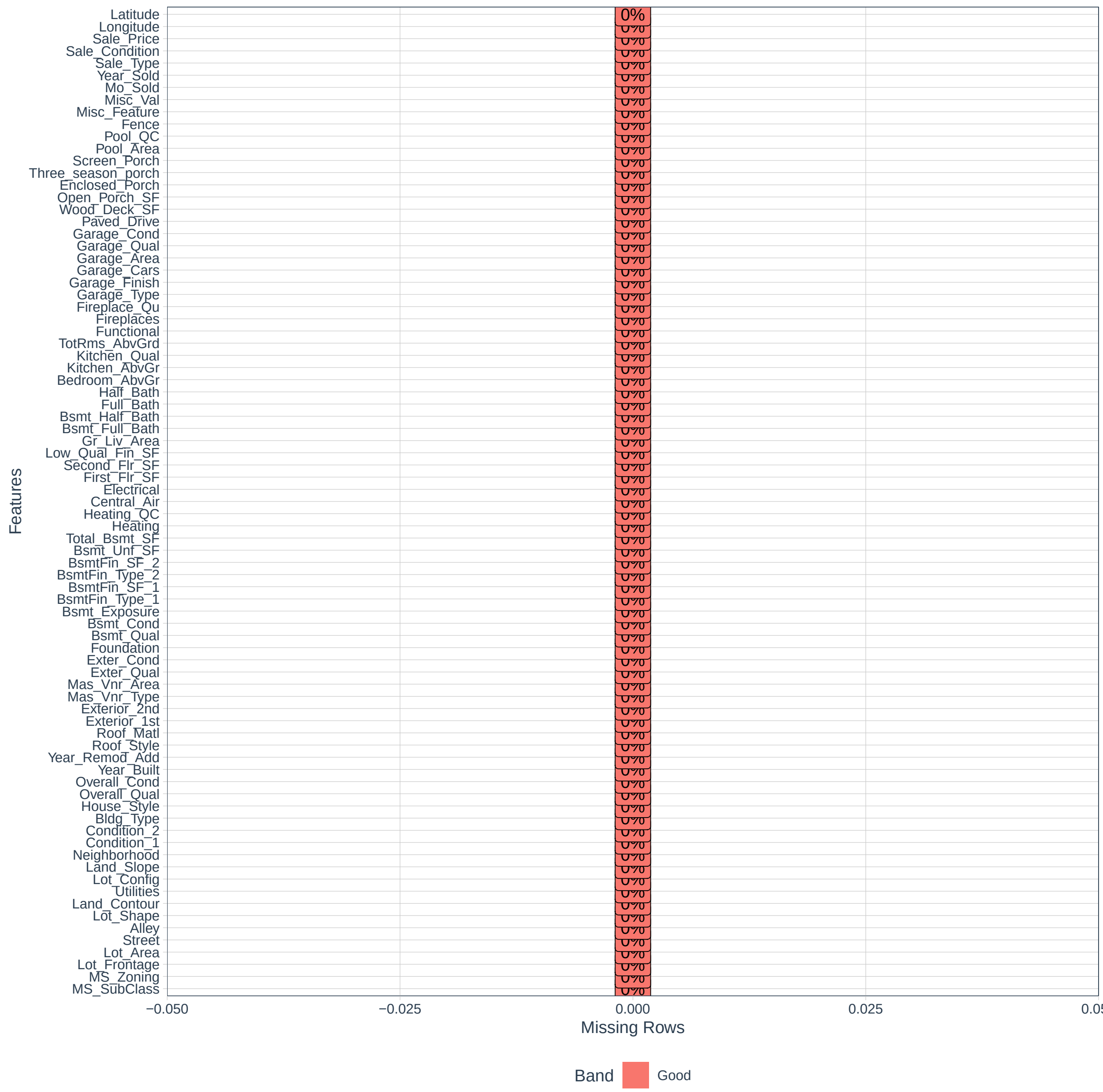


Figure 1: Missing Value in the Data Set

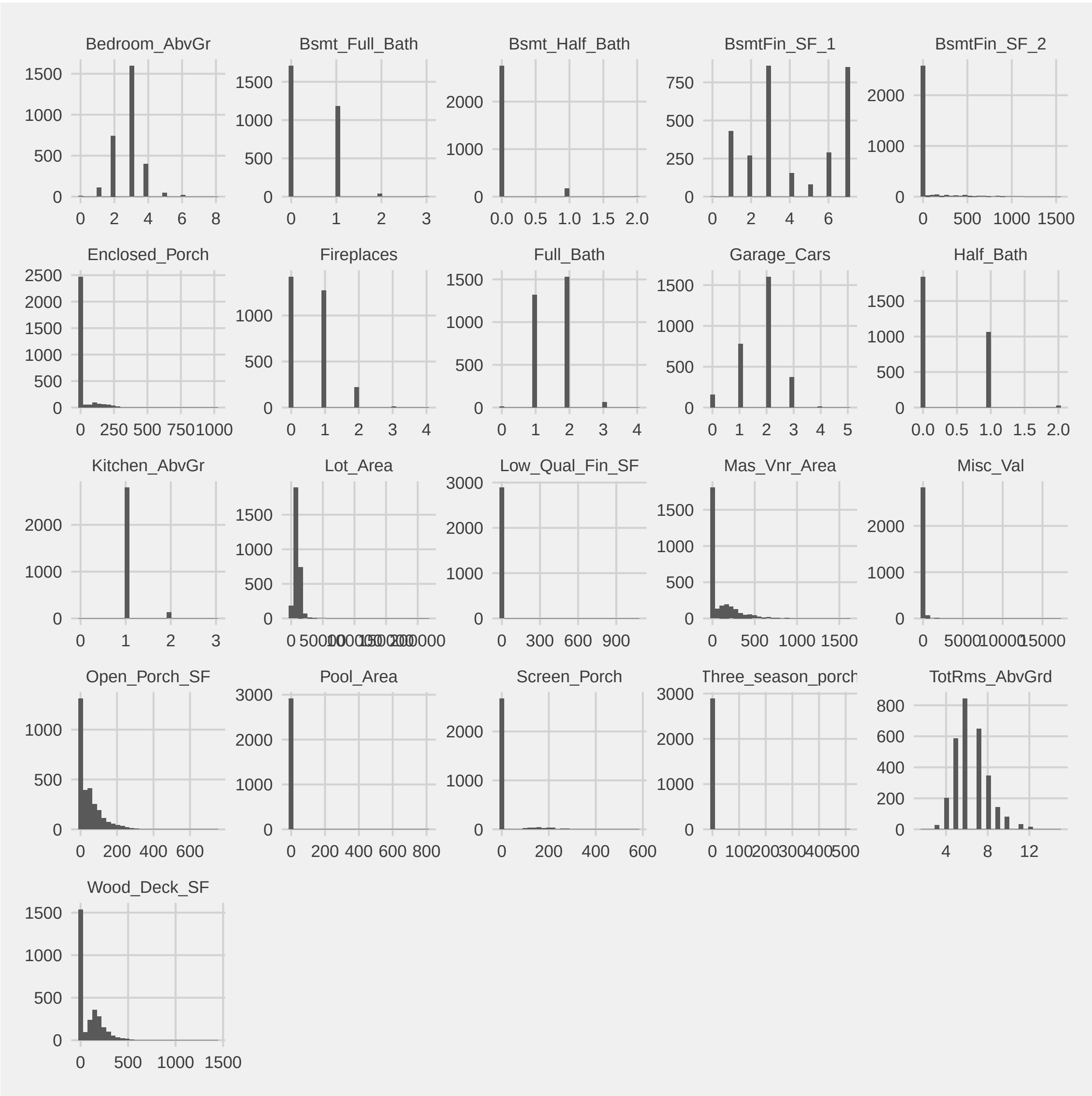


Figure 2: Numerical Variables Overview

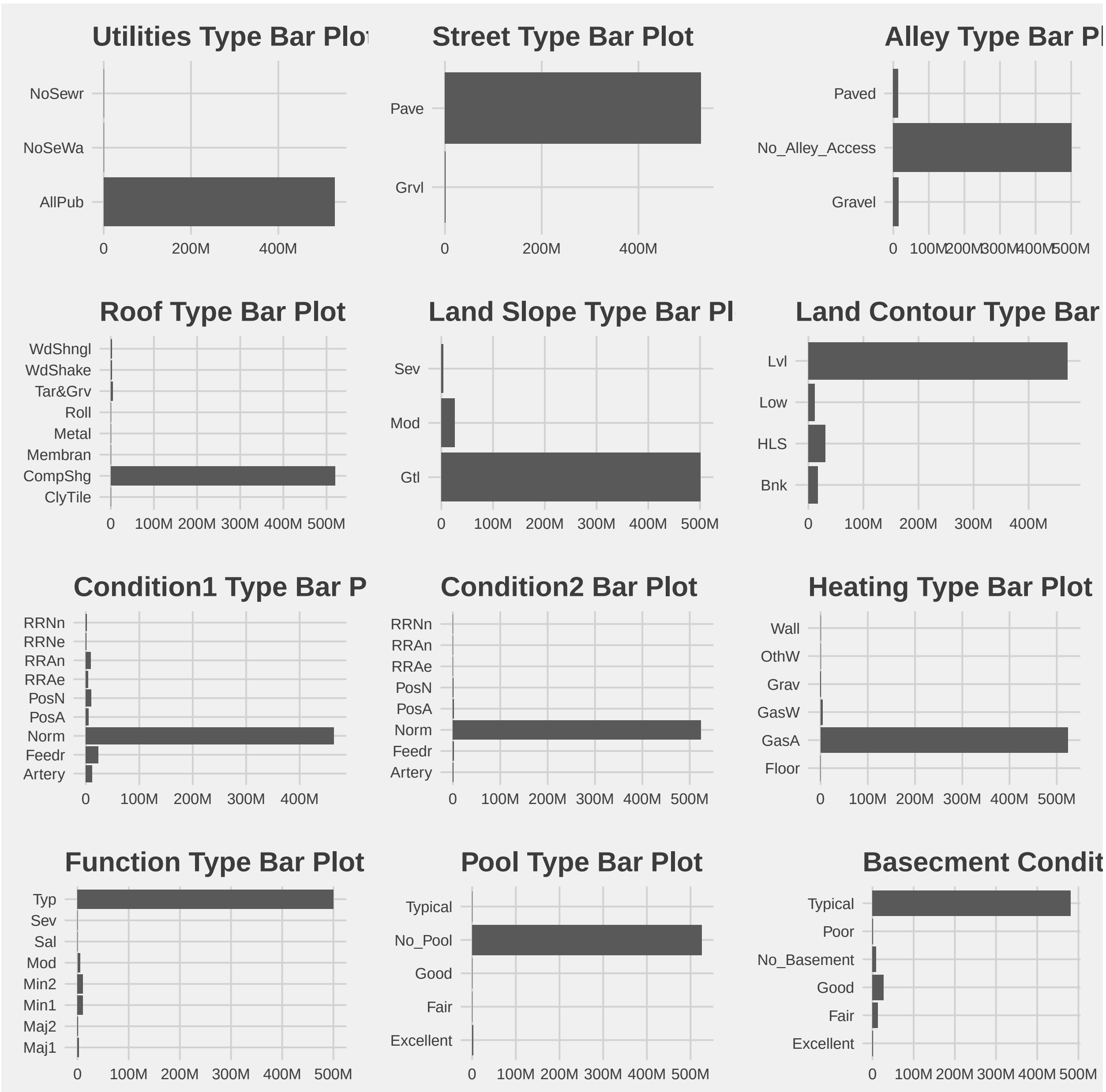


Figure 3: Removable Categorical Variables

5.2 Response Variable `Sale_Price`

We plot `Sale_Price` against `Gv_Liv_Area` [Above grade (ground) living area in square feet] since, base on intuition, size of the property may be positively associated with its value. We also plot the distribution of the response in a histogram and a box plot, check the normality of the response variable in normal Q-Q plot.

From Figure 4, we can observe a clear right skewness of the response which suggest some types of transformation should be considered to improve its normality. This is confirmed in Table 2 since the response has an observed mean of $179,957.7 and an observed median of $159,000. Overall, the response variable is heavily right skewed with some potentially influential data points.

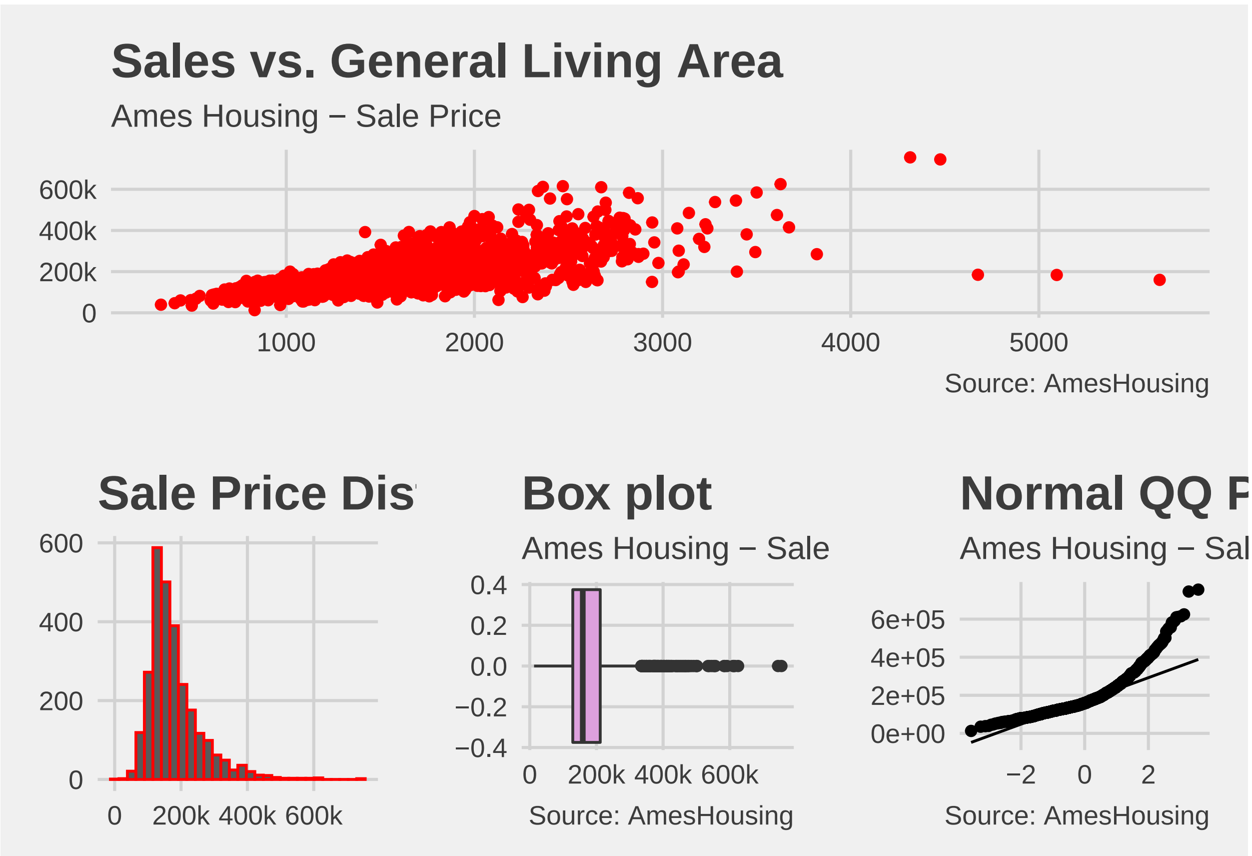


Figure 4: Summary Statistics of Sale_Price

Table 2: Sale_Price Summary Statistics

Statistics	Values
Mean	$179,957.7
Median	$159,000
Standard Error	80,219

We test out log transformation on the response. It is evident from the Figure 5 that the transformation has greatly improved the normality of the response. Therefore, we should consider applying log transformation on the response when we fit the model.

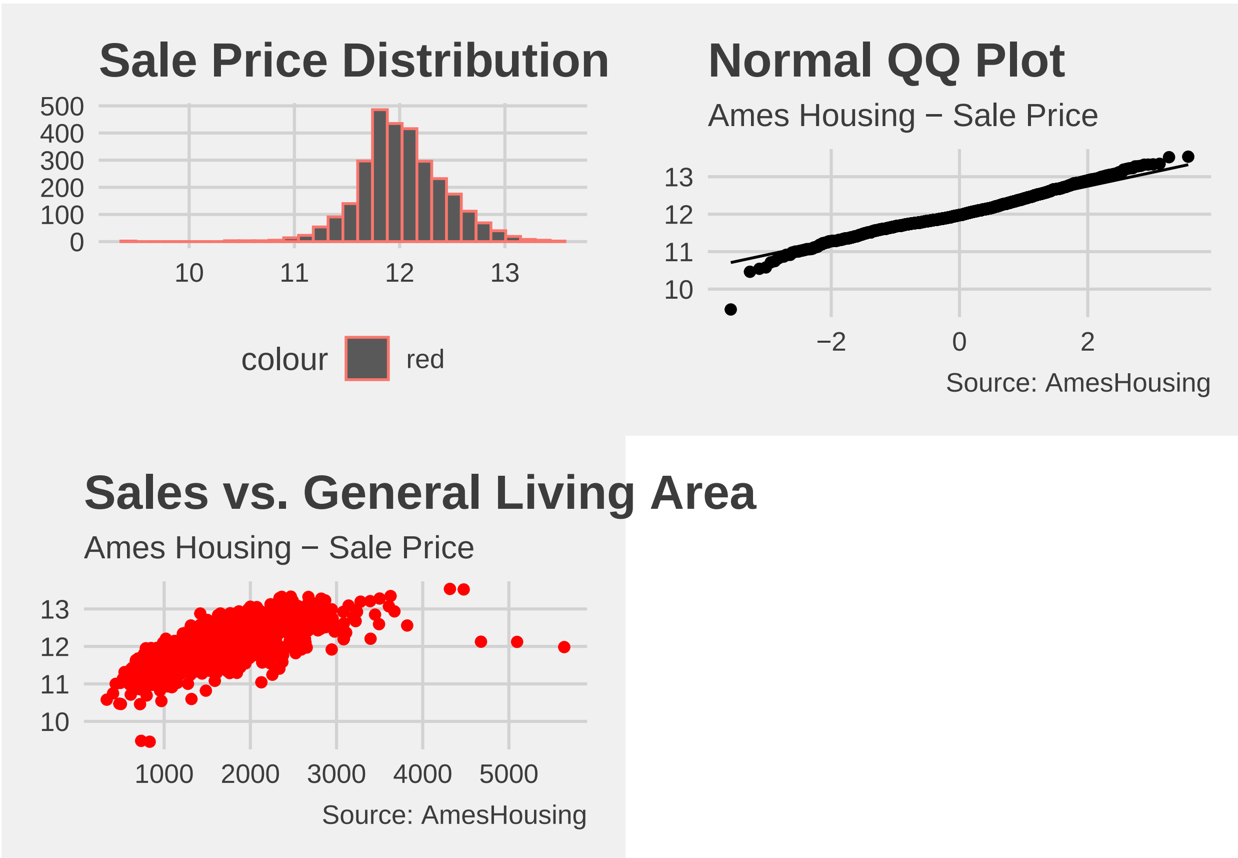


Figure 5: Log Transformation on Sale_Price

Moreover, from the scatter plot in Figure 4, we can observe some exceedingly large data points. The author suggested to remove points that have general living area larger than 4000 sqft. We need to further investigate these points.

5.3 Numeric Predictors

We investigate the correlation between the predictors and the response in a correlation matrix (Figure 6). Some correlation between the predictors are intuitive such as `Gv_Liv_Area` is linear combination of `First_Flr_SF` and `Second_Flr_SF` as well as `Bedroom_AbvGr`.

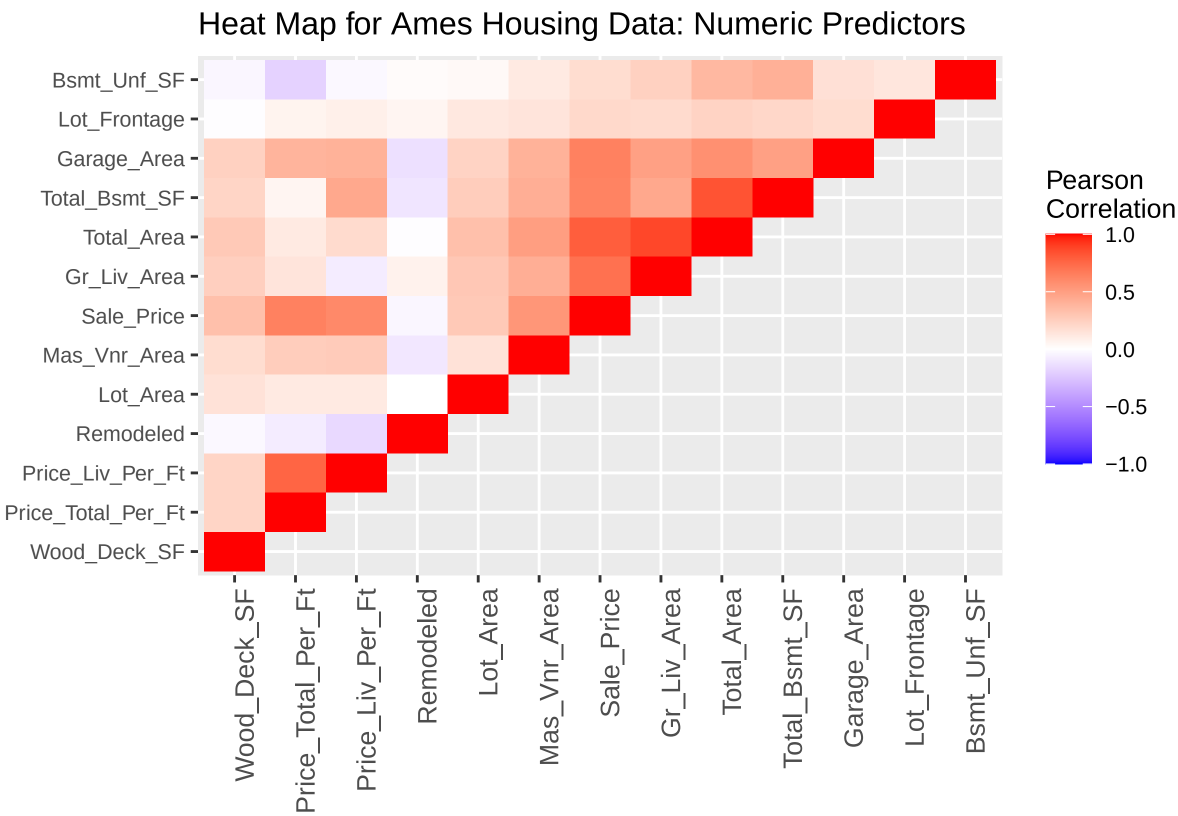


Figure 6: Correlation Matrix

We further investigate muticollinearity with Variance Inflation Factor table below and found that, as expected, `Gr_Liv_Area`, `Firest_Flr_SF` and `Second_Flr_SF` all have VIF scores much larger than 5. We decided to remove `Firest_Flr_SF` and `Second_Flr_SF` and keep `Gr_Liv_Area` since it is contextually important. `Garage_Area` and `Garage_Cars` have VIF scores slightly above 5 as well. However, we decided to keep both variables and convert `Garage_Cars` to a factor.

```
##    Lot_Frontage       Lot_Area     Year_Built Year_Remod_Add   Mas_Vnr_Area
##            1.12           1.20           2.70           1.83           1.37
##    BsmtFin_SF_1    Bsmt_Unf_SF  Total_Bsmt_SF   First_Flr_SF  Second_Flr_SF
##            1.77           2.76           4.50          77.87          91.38
##     Gr_Liv_Area Bsmt_Full_Bath      Full_Bath      Half_Bath  Bedroom_AbvGr
##          125.55           1.90           2.61           2.12           2.17
##   TotRms_AbvGrd     Fireplaces    Garage_Cars    Garage_Area   Wood_Deck_SF
```

```
##          4.00          1.47          5.69          5.43          1.17
##       Mo_Sold     Year_Sold
##          1.03          1.04
```

6 Regression Analysis, Results and Interpretation

6.1 Feature Engineering

We create 4 new variables that could help us answer our research questions: `Total_Area` that combines general living area and basement area; `Price_Total_Per_Ft` calculates the sale price of the total property area per square feet; `Price_Liv_Per_Ft` calculates sale price of the living area per square feet; `Remodeled` that determines whether the house was remodeled.

6.2 Variable Selection

Before fitting the model, we apply forward, backward and stepwsie variable selection algorithms to reduce the amount of variables we will use in the model. Since the computation takes considerable amount of time, we run the algorithms in a separate script, `model.R` and save the result in as an RDS object. From the RDS, we are left with 28 variables. We categorized them and list their types, descriptions in Table 3.

We further reduce the amount of variables and remove those that do not address our research questions. We remove the following variables from our analysis: `Mas_Vnr_Area`, `MS_Zoning`, `Garage_Cars`, `Fireplace_Qu`, `Neighborhood`, `Exterior_2nd`.

Table 3: Variables Keep After Selection Algorithms

Variable Name	Variable Type	Description	Category
`Lot_Frontage`	Numeric	Linear feet of street connected to property	
`Overall_Qual`	Factor	Rates the overall material and finish of the house	
`Overall_Cond`	Factor	Rates the overall condition of the house	
`TotRms_AbvGrd`	Factor	Total rooms above grade (does not include bathrooms)	
`Bedroom_AbvGr`	Factor	Bedrooms above grade (does NOT include basement bedrooms)	
`Total_Area`	Numeric	Total area (including basement)	
`Price_Total_Per_Ft`	Numeric	Sale price of the total property area per square feet	
`Price_Liv_Per_Ft`	Numeric	sale price of the living area per square feet	
`Neighborhood`	Factor	Physical locations within Ames city limits	
`MS_SubClass`	Factor	Identifies the type of dwelling involved in the sale.	
`Lot_Shape`	Factor	General shape of property	
`MS_Zoning`	Factor	Identifies the general zoning classification of the sale.	
`Year_Built`	Date	Original construction date	
`Mas_Vnr_Area`	Numeric	Masonry veneer area in square feet	
`Remodeled`	Factor	Whether the house is modeled or not	Remodel

Variable Name	Variable Type	Description	Category
Year_Remod_Add	Date	Remodel date (same as construction date if no remodeling or additions)	Remodel
Half_Bath	Factor	Half baths above grade	Bathroom
Full_Bath	Factor	Full bathrooms above grade	Bathroom
Exterior_2nd	Factor	Exterior 2: Exterior covering on house (if more than one material)	Exterior
Exter_Cond	Factor	Evaluates the present condition of the material on the exterior	Exterior
Total_Bsmt_SF	Numeric	Total square feet of basement area	Basement
Bsmt_Qual	Factor	Evaluates the height of the basement	Basement
Bsmt_Exposure	Factor	Walkout or garden level walls	Basement
Bsmt_Full_Bath	Factor	Basement full bathrooms	Basement
Garage_Area	Numeric	Size of garage in square feet	Garage
Garage_Cars	Factor	Size of garage in car capacity	Garage
Fireplaces	Factor	Number of fireplaces	Fireplace
Fireplace_Qu	Factor	Fireplace quality	Fireplace

6.3 Diagnostics

Our preliminary multiple linear regression model has 22 variables. Before checking the diagnostics, we apply log transformation on `Sale_Price` as suggested from our **Exploratory Data Analysis**, and investigate the exceedingly large points we observed.

6.3.1 Influence Points & Outliers

We identify observations with absolute standardized residuals greater than 2 and observations with hat-values greater than 3. Through checking their intersections, it can be concluded that the following observations are outliers and thus can be removed: 18, 106, 162, 274, 721, 723, 894, 1090, 1405, 2047, 2089, 2197, 2708 .

Evidently, the author recommended "removing any houses with more than 4000 square feet from the dataset." (Dean De Cock 2011) Based on this recommendation, we investigated the indices of the observations with general living area greater than 4000: 1405, 1660, 1667, 2046, 2047 . Four out of five of these observations are outliers.

6.3.2 Normality and Equal Variance Check

Since the assumption of independency has been satisfied according to the author's documentation, we check the normality and equal variance assumptions of our model after applying log transformation on the response and removing outliers. It can be observe from Figure 7 that the equal variance assumption is satisfied as most of the points are somewhat scattered around 0. However, based on Figure 8, the assumption of normality is not satisfied.

6.3.3 Transformation

Since we already apply log transformation on the response, we examine possible transformations on the numerical predictors. Based on the result below, we can apply square root transformation on `Total_Area`, and choose to apply no transformation on the rest of the predictors since their lambda values are close 1. After fitting the model again, we can observe Figure 9 and 10 and see that normality improved slightly but still heavy tailed. It is plausible that multiple linear regression is not appropriate to this dataset. However, we continue our analysis with this model for now.

```
## bcPower Transformations to Multinormality
##                Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Total_Area        0.4188        0.42       0.3500       0.4877
## Total_Bsmt_SF     0.8678        0.87       0.8370       0.8987
## Garage_Area       0.8635        0.86       0.8315       0.8955
## Lot_Frontage      0.8152        0.82       0.7819       0.8485
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df       pval
## LR test, lambda = (0 0 0 0) 14673.83  4 < 2.22e-16
```

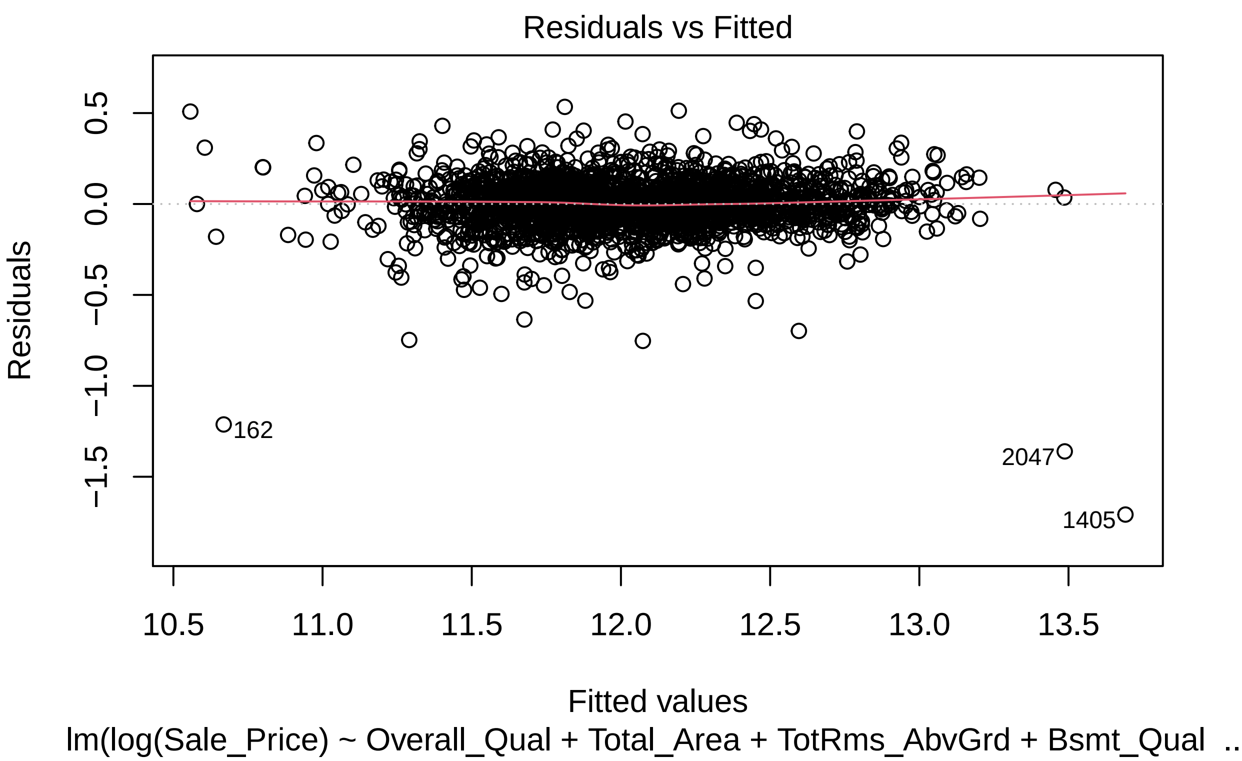


Figure 7: Equal Variance

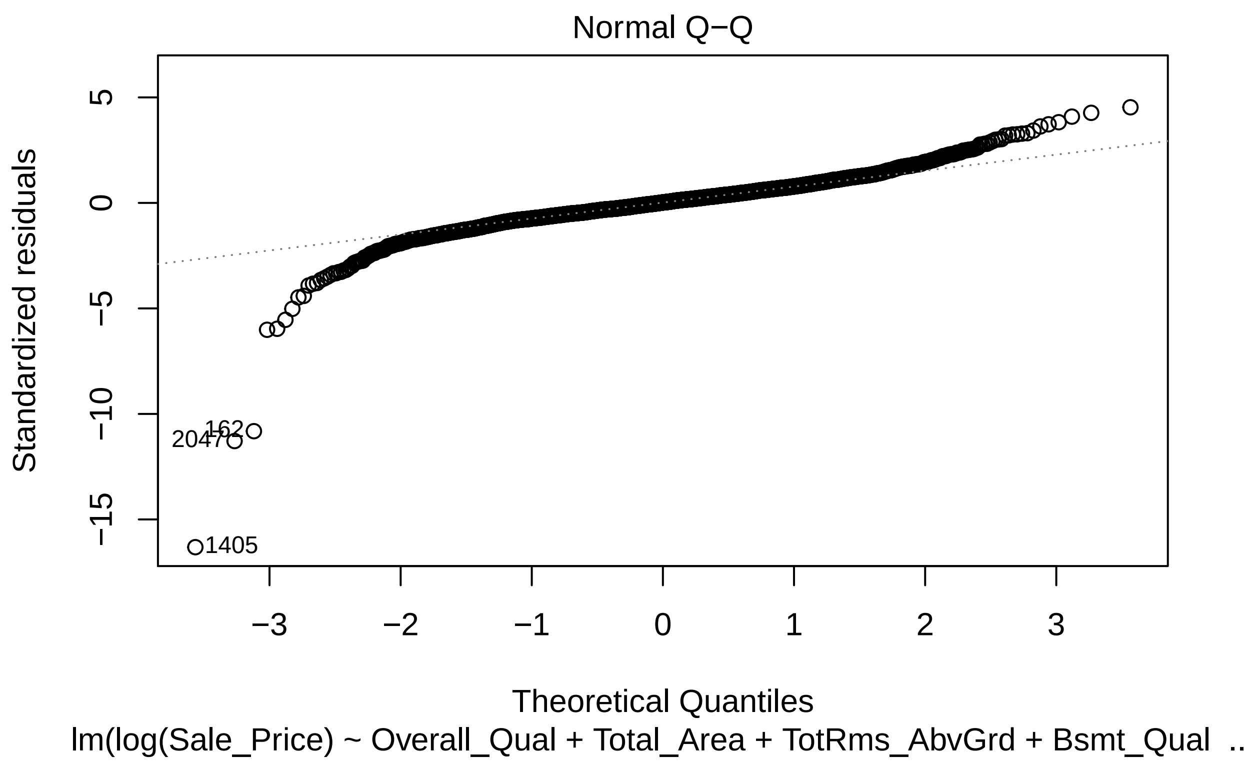


Figure 8: Normality

```
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1 1) 443.8562  4 < 2.22e-16
```

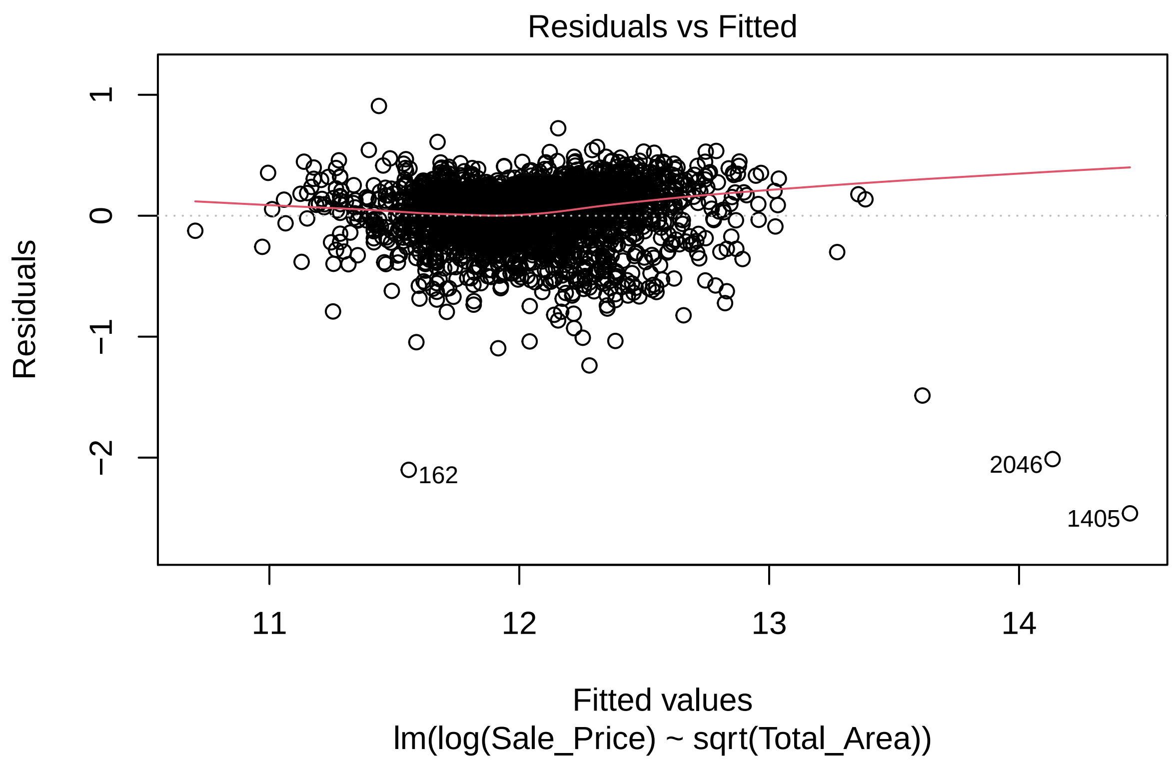


Figure 9: Equal Variacne Check on Transformed Model

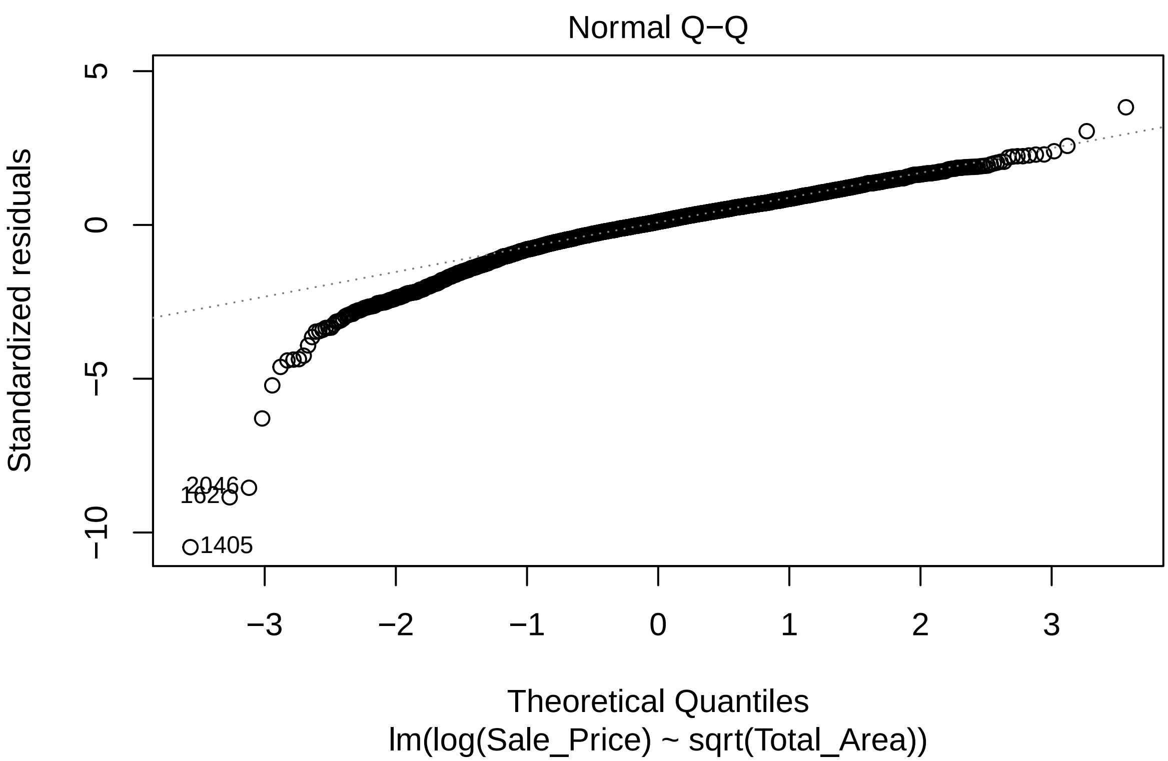


Figure 10: Normality Check on Transformed Model

6.4 Results

For the final model, we create dummy variables for all the categorical variables and keep the significant ones in the model. This resulted in 10 predictors in our final model. We check the assumptions of multiple linear regression again. It can be observed from Figure 11 that the model somewhat satisfies the equal variance assumption. However, normality improved somewhat but still do not satisfy the assumption as indicated in Figure 12.

```
##
## Call:
## lm(formula = log(Sale_Price) ~ Total_Bsmt_SF + Garage_Area +
##     Overall_Qual_Good + Overall_Qual_Poor + TotRms_AbvGrd_15 +
##     Bsmt_Qual_Typical + Bsmt_Qual_Good + Bsmt_Qual_No_Basement +
##     Year_Built + sqrt(Total_Area), data = ames_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89611 -0.08481  0.00740  0.09609  0.66426
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.214e+00  3.143e-01  10.226  < 2e-16 ***
## Total_Bsmt_SF         -1.273e-04  1.458e-05  -8.728  < 2e-16 ***
## Garage_Area            3.179e-04  1.979e-05  16.063  < 2e-16 ***
## Overall_Qual_Good      2.581e-02  8.988e-03   2.872  0.00411 **
## Overall_Qual_Poor     -3.030e-01  5.743e-02  -5.276 1.42e-07 ***
## TotRms_AbvGrd_15      -1.676e+00  1.734e-01  -9.666  < 2e-16 ***
## Bsmt_Qual_Typical     -1.319e-01  1.212e-02 -10.878  < 2e-16 ***
## Bsmt_Qual_Good        -1.007e-01  1.163e-02  -8.663  < 2e-16 ***
## Bsmt_Qual_No_Basement -1.014e-01  2.518e-02  -4.029 5.76e-05 ***
## Year_Built             3.598e-03  1.589e-04  22.641  < 2e-16 ***
## sqrt(Total_Area)       3.606e-02  8.115e-04  44.438  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1695 on 2738 degrees of freedom
## Multiple R-squared:  0.8173, Adjusted R-squared:  0.8166
## F-statistic:  1225 on 10 and 2738 DF,  p-value: < 2.2e-16
```

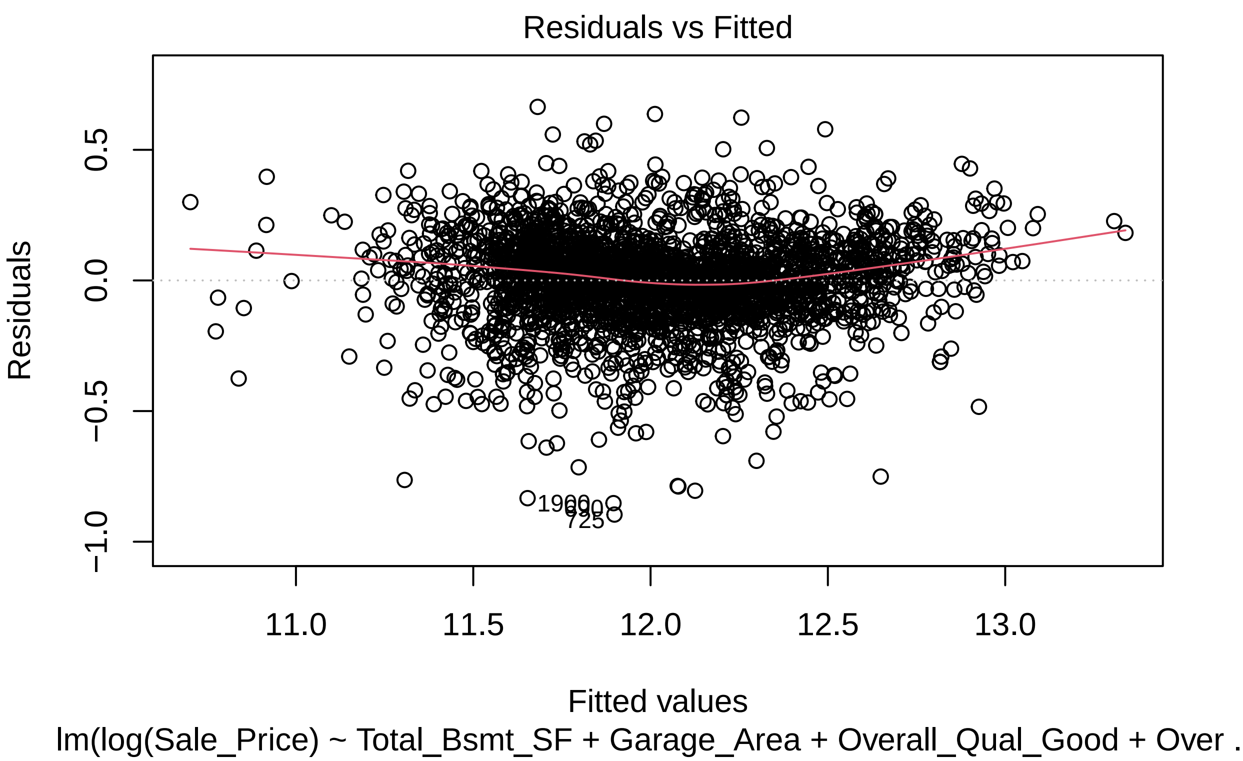


Figure 11: final model diagnostic - Equal Variance

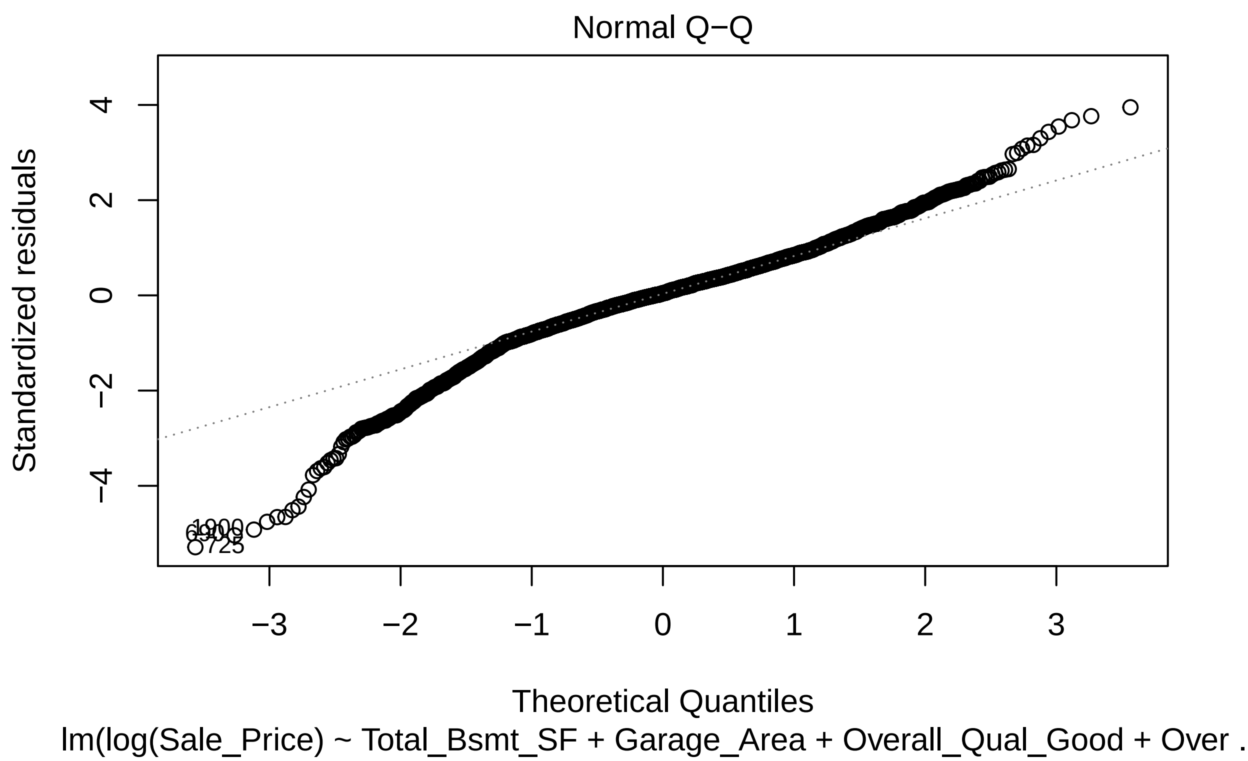


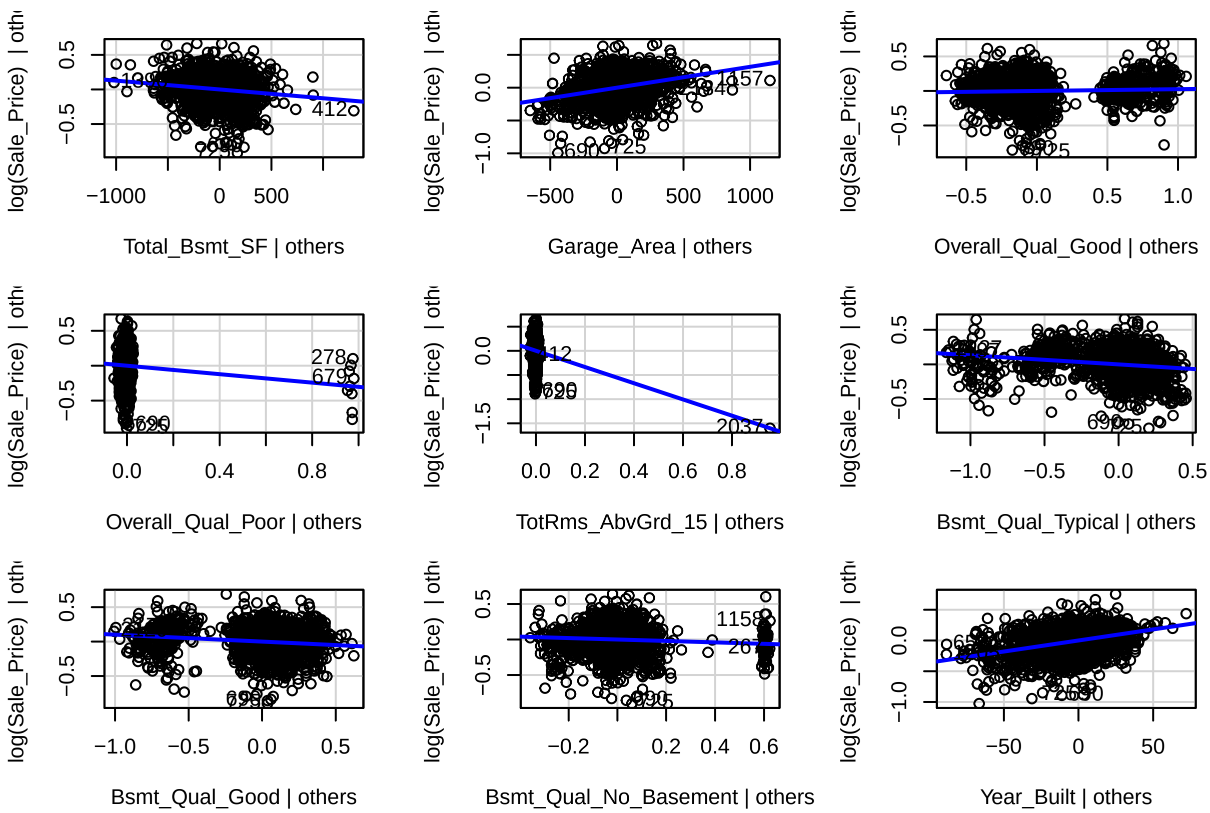
Figure 12: final model diagnostic - Normality

6.5 Interpretation

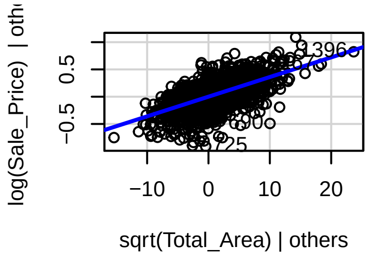
Our final fitted model is:

Log(Sale_Price) = 3.214 - 0.0001 Total_Bsmt_SF + 0.0003 Garage_Area + 0.0258 Overall_Qual_Good - 0.303 Overall_Qual_Poor - 1.676 TotRms_AbvGrd_15 - 0.132 Bsmt_Qual_Typical- 0.101 Bsmt_Qual_Good - 0.101 Bsmt_Qual_No_Basement + 0.004 Year_Built + 0.036 sqrt(Total_Area)

From looking at the Added Variables Plots, we can conclude that `Garage_Area`, `Year_Built` and `sqrt(Total_Area)` are all positively associated with the response after controlling the effect of others, while `Overall_Qual_Poor`, `TotRms_AbvGrd_15`, `Total_Bsmt_SF` are negatively associated with the response after controlling the effects of other predictors. Other predictors that we find to be significant do not seem to be able to explain variability in the response individually. This could be due to overfitting.



Added−Variable Plots



Now looking at the summary table of the fitted model, we have the following interpretation:

- The predictors in this model collectively explain 81.7% of variability in the response, `Sale_Price`.
- A 1 unit increase in total basement area in square feet, with the other predictors held fixed, is associated with an decrease in Sale Price by nearly 1 dollar.
- A 1 unit increase in garage area in square feet, with the other predictors held fixed, is associated with an decrease in Sale Price by nearly 1 dollar.
- Sale price of a property is positively associated with the good quality of overall material and finish of the house ,total area and the year it was built.
- Sale price of a property is negatively associated with poor quality of overall material and finish of the house, number of the rooms if it's 15 and more, and basement quality.
- Since `Year_Built` is a significant predictor, time series analysis should be done based on this finding. We do not preform a time series analysis as it is out of the scope of this project.
- Overall, a house is more valuable if it has less than 15 rooms, no basement or smaller basement, good overall finish, bigger garage and total area.

7 Conclusion

We can conclude that in Ames, Iowa, the real states market values a house with bigger area and living area than basement. The fixed characteristics that influences the sale price of a house the most is the overall quality of the materials and finish, and the size of the living area. However, multiple linear regression may not be the best regression method for this dataset even though the model explains 81.6% of the variability in Sale Price. A more general regression technique is preferred since our model doesn't satisfy all the assumptions of a multiple linear regression model.

8 Appendices

8.1 R scripts

The report is rendered with the following scripts arranged in order of dependency:

- `script.R` : the main script containing all of the data objects and model objects. Need to be run first. The script makes calls to three other scripts: `packages.R` , `functions.R` and `dummy.R`. The project RMD file automatically calls this script.
- `figure_generator.R`: generate all of the figures used in the report and presentation.
- `dummy.R` : creates dummy variables for all of the categorical variables. Called by `script.R`
- `model.R`: runs forward, backward and stepwise AIC and BIC. Do not need to be called.
- `packages.R` : contains all of the packages used in this project. Called by `packages.R`.
- `functions.R` : contains helper functions. Called by `script.R` .

Reference

Dean De Cock. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." http://jse.amstat.org/v19n3/decock.pdf.