House Price Prediction

Syed Faizan

ALY 6015 Module 1 Regression Diagnostics in R by Syed Faizan

Introduction

In this project, I present a detailed exploration of the famous Ames housing dataset, employing exploratory data analysis (EDA) and descriptive statistics to meticulously investigate the data's attributes. After the EDA I proceeded with the imputation of missing values and ensured a robust base for the regression model. A correlation matrix, supplemented by visual aids, facilitated the discernment of variable interrelations. Focused scatter plot analyses illuminated varied associations with 'SalePrice' which was my chosen response variable. My regression model, articulated through a precise equation, enabled the interpretation of each predictor's influence. Diagnostic evaluations and multicollinearity assessments informed iterative enhancements to the model's integrity. Leveraging all subsets regression, I identified the optimal model configuration, which was critically compared to the initial model to validate the effectiveness of the selected modeling approach. The regression model may be employed for sale price prediction and predictor effect analysis. The report takes the form of answers to the 14 questions posed in the assignment rubric, systematically maintaining an analytical workflow that leads to the final adoption of a linear regression model based on continuous variables in the prescribed dataset.

Analysis

I loaded the prescribed Ames housing dataset into R Studio to proceed with the analysis.

The Ames Housing Dataset, derived from the Ames Assessor's Office, details housing sales in Ames, Iowa from 2006 to 2010. It is a comprehensive resource for regression analysis in real estate and was designed with academic purposes in mind. Each observation represents a house sale. It is an extremely popular dataset for data science competitions and in the academic setting for regression analysis.

The Variable breakdown of the Data set is as follows: T

Nominal (23): Categorical without order, e.g., 'Neighborhood'.

Ordinal (23): Categorical with order, e.g., 'Overall Qual'.

Discrete (14): Countable numbers, e.g., 'Full Bath'.

Continuous (20): Measurable range, e.g., 'Lot Area'.

In total, there are 2930 observations with 46 character-like variables, combining nominal and ordinal types.

Right at the outset, since the dataset is large and rich in variables, I sought to focus on the continuous variables since they are most relevant to linear regression models. I created a new 'Age' column based on the year in which the house was built.

I then set about looking for discrepancies in the Dataset such as missing values or inexplicable zeros. As depicted in the visualization below 5 columns represented the major portion of the missing values and were promptly dropped from the dataset as they has upwards of 90% values missing in some cases.

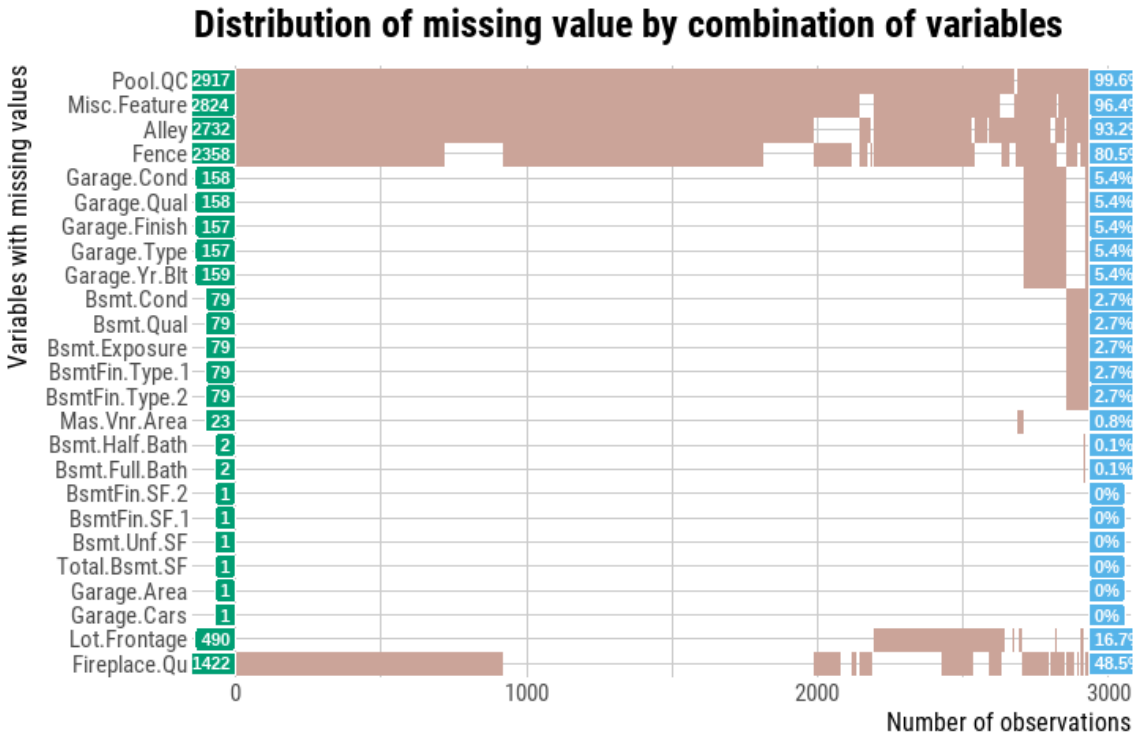


Figure 1 Missing values by variable

As the above visualization renders clear, 5 columns had substantial missing values with two being well over 95% in terms of missing values. I therefore dropped these 5 columns from the dataset.

After dropping the columns mainly contributing to missing values I sought to look at the remaining columns for missing values with special minute attention to continuous variables.

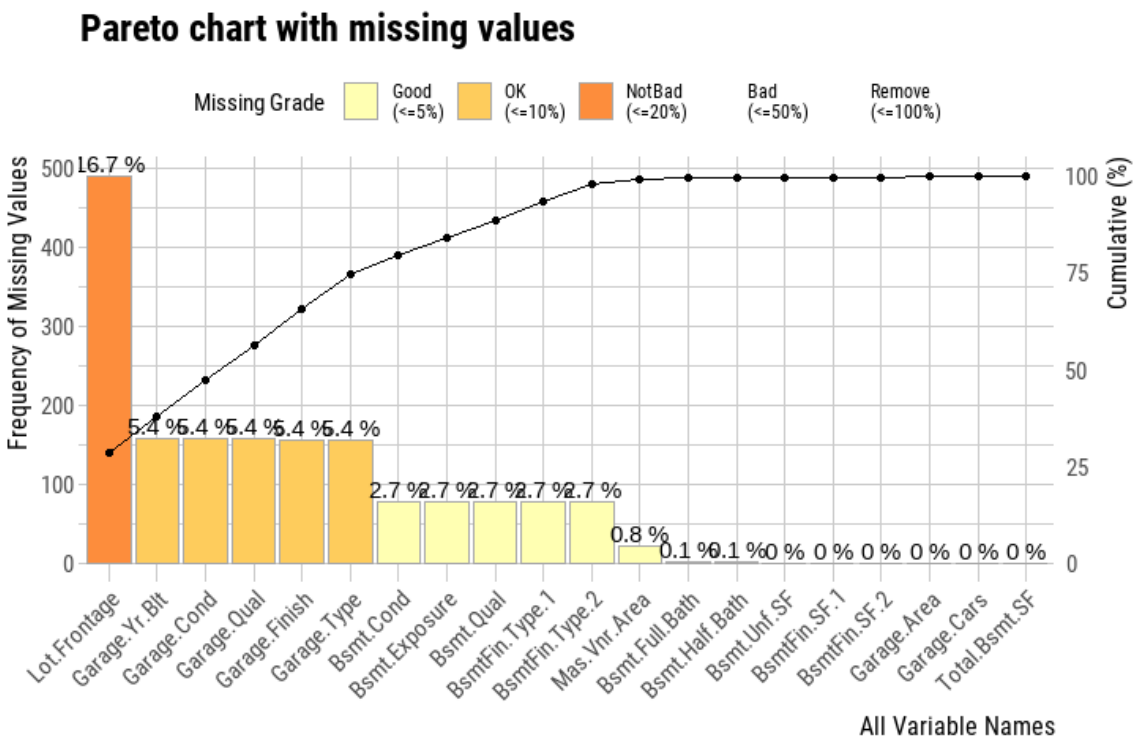


Figure 2 Remaining missing values by variable

As this pareto chart shows, Lot Frontage, Masonry veneer area and Total Basement Area were among the important continuous variables that had substantial missing values. In fullfilment of the Question 3 of the assignment and in order to facilitate further analysis I imputed the missing values. I compared models for imputing missing values for Lot Frontage in particular, as this column had 16.7% missing values.

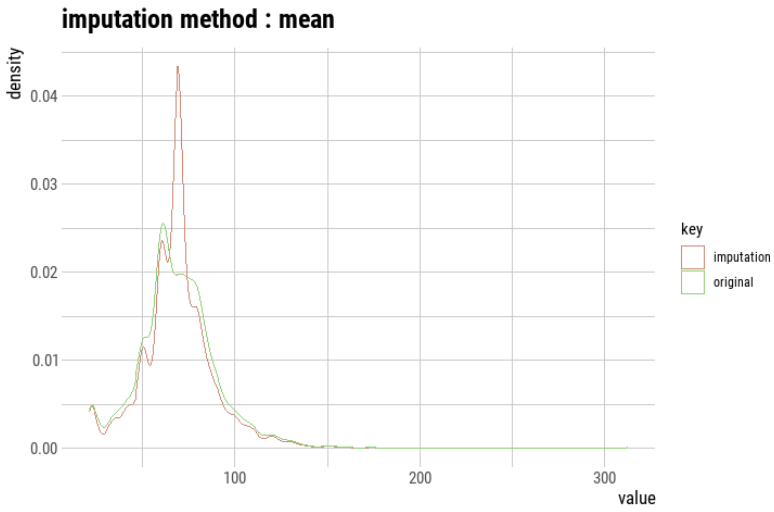


Figure 3 Imputation by mean producing abnormal distortion in the data

I examined the option of imputing means to this column, but as the above plot makes abundantly clear how this distorts the distribution, I settled upon machine learning algorithms(MICE- Multivariate Imputation by Chained Equations and PMM – Predictive Mean Matching) using the 'ranger' and 'mice' packages in R. As seen below the distribution of the Lot Frontage variable was largely preserved.

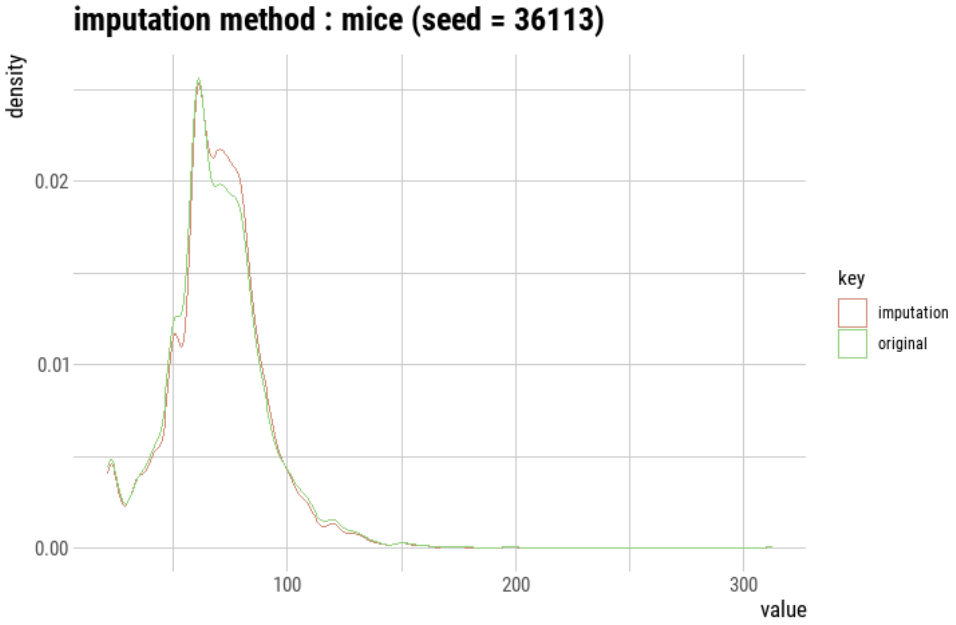


Figure 4 Imputation by Machine Learning algorithms

As the number of missing values in the other two variables of interest were small I employed imputation using means.

As the number of 'integer' variables was 39 and therefore too large to be amenable to visual representation in a descriptive statistics table, I decided to home in on the 7 most important variables with respect to sale price. The choise to focus on these variables ws informed by domain knowledge of housing models. Most, as can be seen below, have to do with the size of some component of the house.

	variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
1	Gr.Liv.Area	334	1126.00	1499.69044	1442.0	1742.75	5642	0	0	74
2	Total.Bsmt.SF	0	793.00	1051.61454	990.0	1301.50	6110	79	0	123
3	Garage.Area	0	320.00	472.81973	480.0	576.00	1488	157	0	42
4	Lot.Frontage	21	60.00	70.13959	70.0	80.00	313	0	0	238
5	Lot.Area	1300	7440.25	10147.92184	9436.5	11555.25	215245	0	0	127
6	SalePrice	12789	129500.00	180796.06007	160000.0	213500.00	755000	0	0	137
7	Age	0	9.00	38.64369	37.0	56.00	138	3	0	9

Figure 5 Descriptive Statistics of important numerical variables

The above table delineates a statistical summary of selected variables from the Ames Housing dataset, providing quintessential measures of central tendency and dispersion. The variables encompass living area above ground (Gr.Liv.Area), total basement area (Total.Bsmt.SF), garage area (Garage.Area), lot frontage (Lot.Frontage), lot area (Lot.Area), sale price (SalePrice), and the age of the property (Age).

For each variable, the table enumerates the minimum (min), first quartile (Q1), mean, median, third quartile (Q3), and maximum (max) values, elucidating the distribution's scope. Notably, Gr.Liv.Area exhibits a mean of approximately 1499.69 square feet, reflecting a higher average than the median, indicating a potential right skew in the data. The Total.Bsmt.SF and Garage.Area have zero values as their minimum, possibly indicating the absence of a basement or garage in certain properties.

The Lot.Frontage and Lot.Area variables show significant variability, as evidenced by their substantial range, with Lot.Frontage spanning from 21 to 313 feet and Lot.Area from 1300 to 215245 square feet. SalePrice displays considerable diversity in property values, ranging from $12,789 to $755,000, with a median sale price markedly less than the mean, further suggesting right-skewed distribution. Outliers will be discussed in a subsequent section.

The Age variable, defined as the number of years since construction, ranges from 0 to 138 years, suggesting the inclusion of both newly constructed and historic homes within the dataset.

This table is also available as an interactive table at https://rpubs.com/SyedFaizan2024/1173195

Visualisation of the numeric variables

Visualisation of continuous variables provides clues often hidden within heaps of values and helps reveal important patterns that pertain to the assumptions of Linear Regression. With

Linearity, Normality and Homoscedasticity in mind, I ventured to visualise the set of chosen numeric variables in box plots, histograms and scatter plots.

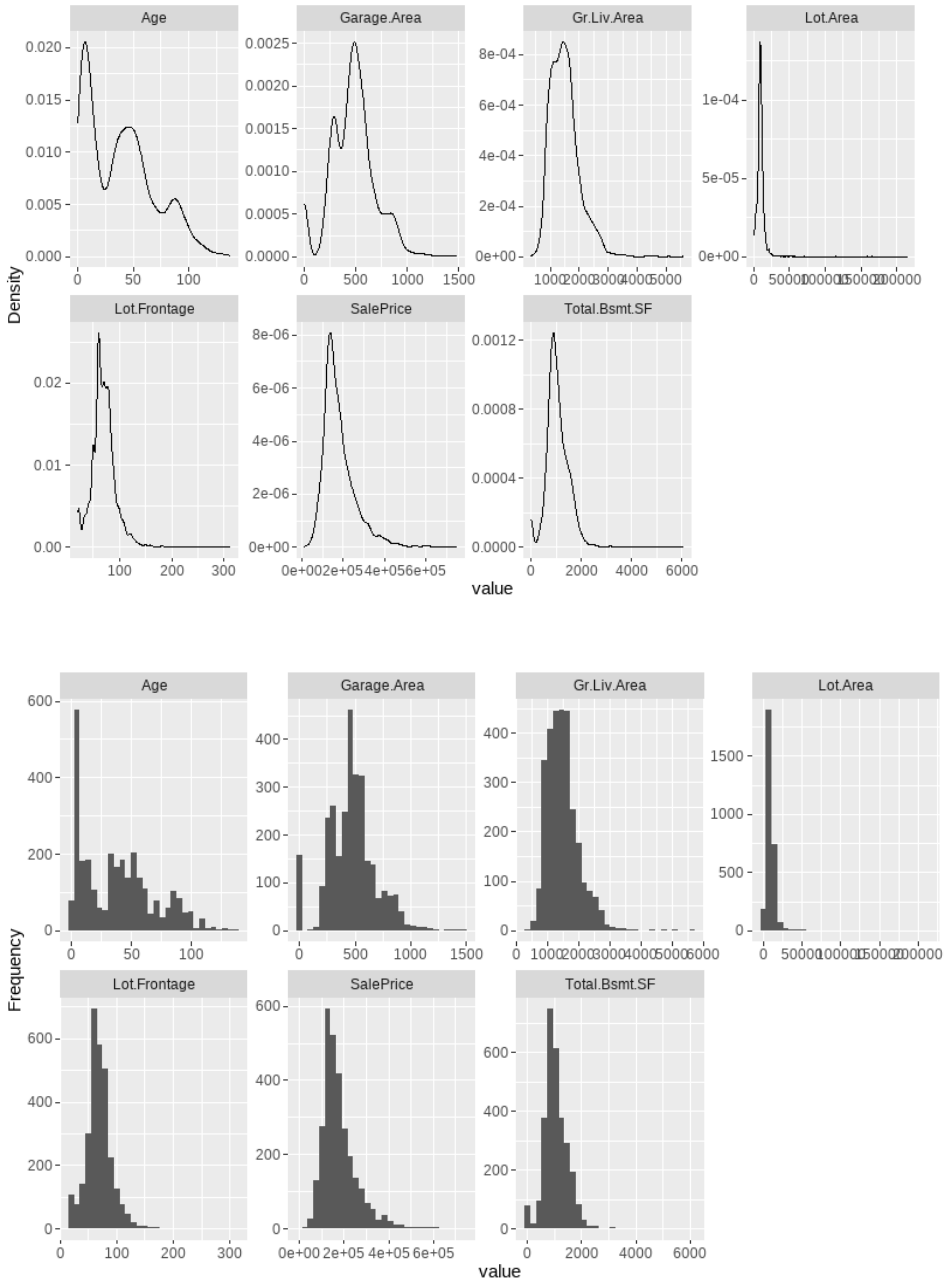


Figure 6 Density plots and Histograms of important numerical variables

Density plots and Histograms

The density plots and histograms above suggested that **SalePrice**, **Lot.Frontage**, **Lot.Area**, and **Total.Bsmt.SF (Total basement area in square feet)** deviate from normality, displaying right-skewed tendencies with a longer tail on the right side. **Age** and **Garage.Area**

exhibit a more complex, multi-modal distribution, implying the presence of distinct subpopulations within the data. Conversely, **Gr.Liv.Area** approaches a unimodal distribution with a slight right skew.

Frequency histograms tend to corroborate these observations, with **SalePrice**, **Total.Bsmt.SF**, **Gr.Liv.Area**, and **Garage.Area** demonstrating asymmetry towards higher values. The histogram for **Lot.Area** underscored its substantial right skewness, with a concentration of values at the lower end and sparse frequency at higher values.

I wanted to further investigate normality of the distributions visually and consider log transformations as a strategy to standardize the variables. I therefore undertook to create more normality visualizations.

Normality Plots

The series of plots below scrutinize the conformity of key continuous variables from the Ames Housing dataset to the normal distribution, an assumption pivotal to linear regression. Histograms and Q-Q plots for **Lot Area**, **Lot Frontage**, **Age**, **Garage Area**, **Gr.Liv.Area**, and **SalePrice** illustrate deviations from normality, with pronounced skewness and kurtosis. Logarithmic and square root transformations attempt to normalize the distributions, with varying success across variables. The Q-Q plots reveal substantial departures from the diagonal line, indicating a mismatch with the theoretical normal distribution. Notably, transformations improve normality for some variables, as evidenced by more symmetrical histograms and more aligned points in the Q-Q plots. This alignment is crucial for meeting linear regression assumptions, which dictate that the residuals—rather than the predictors themselves—should be normally distributed for inference validity.

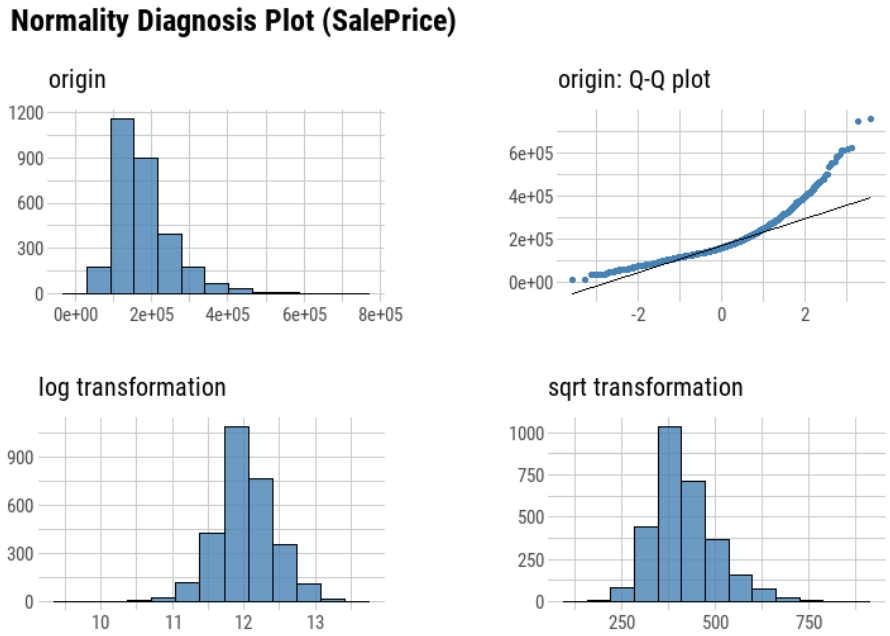


Figure 7 Normality plot of Sale Price

Sale price is the most crucial variable in the dataset as far as our purposes are concerned, given that I chose it as my response variable. It is very important to note in the above plots how the sale price at the higher and the lower end tend to move away from normality. This suggests that the data is skewed rightward due to some very expensive houses. Such a distribution is often noted among econometric variables such as costs, prices, taxes etc. A log transform clearly aids in offsetting some of these departures from the normal distribution. I therefore added a log transformed sale price column in the dataset for potential use later on in the linear regression model.

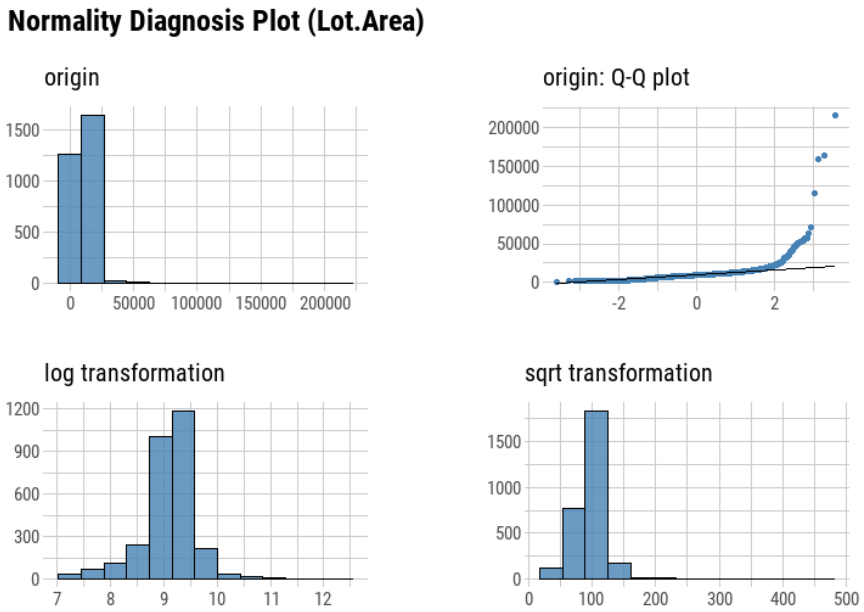


Figure 8 Normality plot of lot area

Lot Area shows the impact of outliers in its distribution.

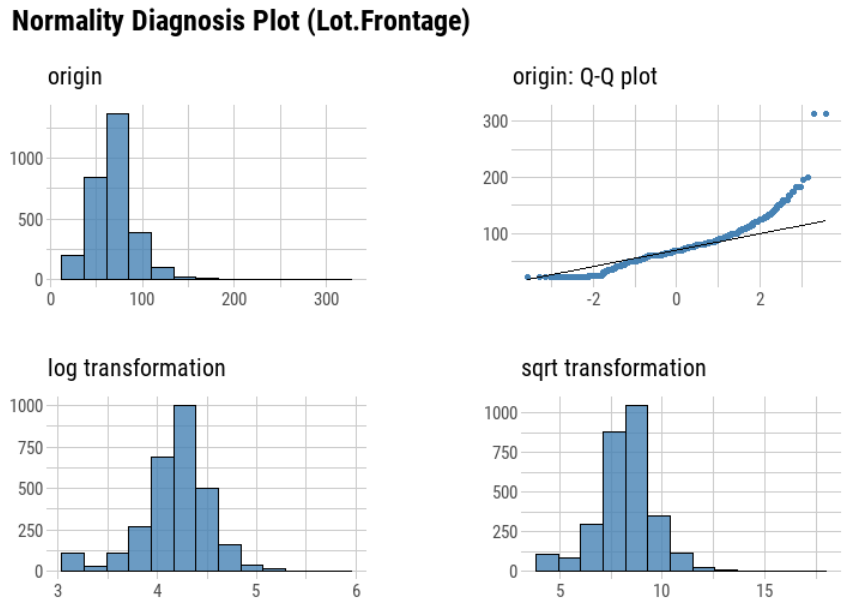


Figure 9 Normality plot of Lot Frontage

The imputed values in Lot Frontage seem to have not altered the fundamental distribution. Square root transform might help standardise this variable.

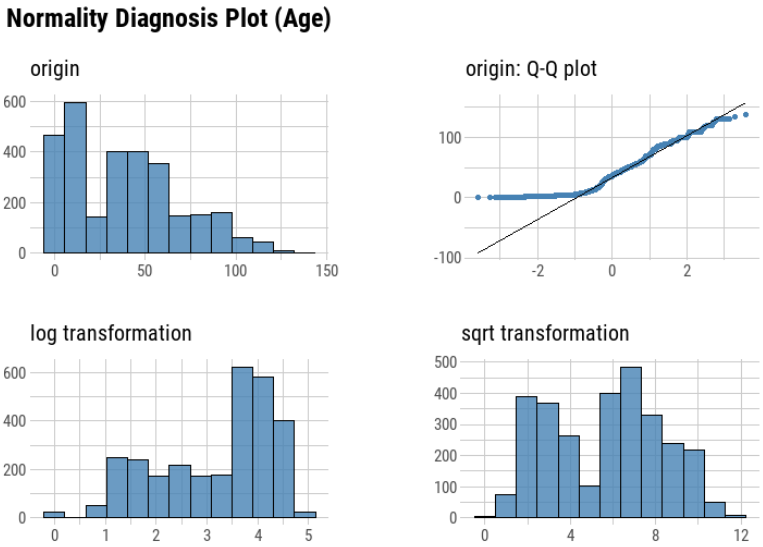


Figure 10 Normality Plot of Age

The zero values in age represent houses that were constructed in the year 2010. It is interesting that age follows a normality line if allowances are made for the zero value.

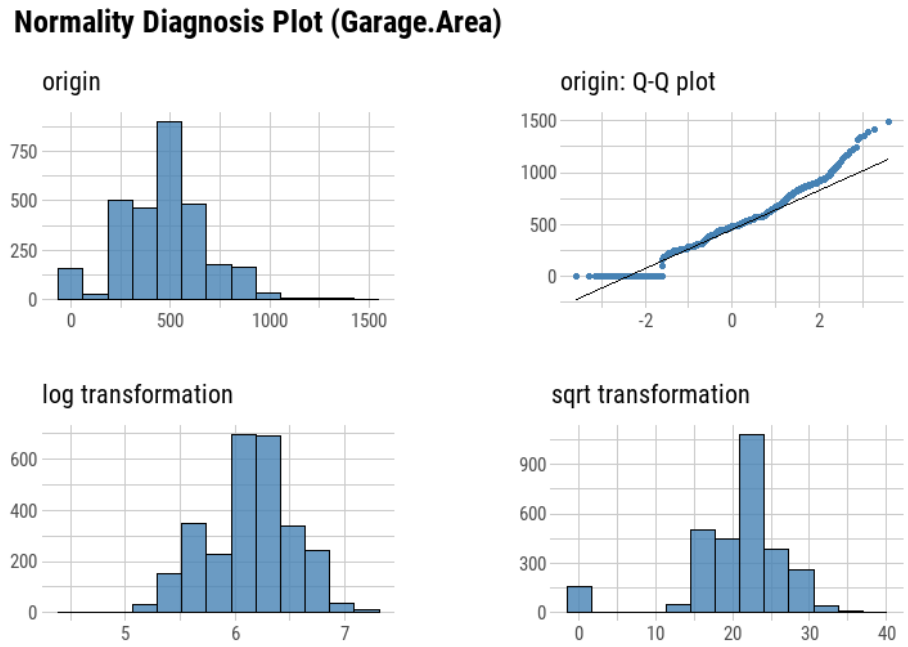


Figure 11 Normality plot of Garage Area

Garage Area shows a deviation from normality at the tails. A log transform aids in standardization.

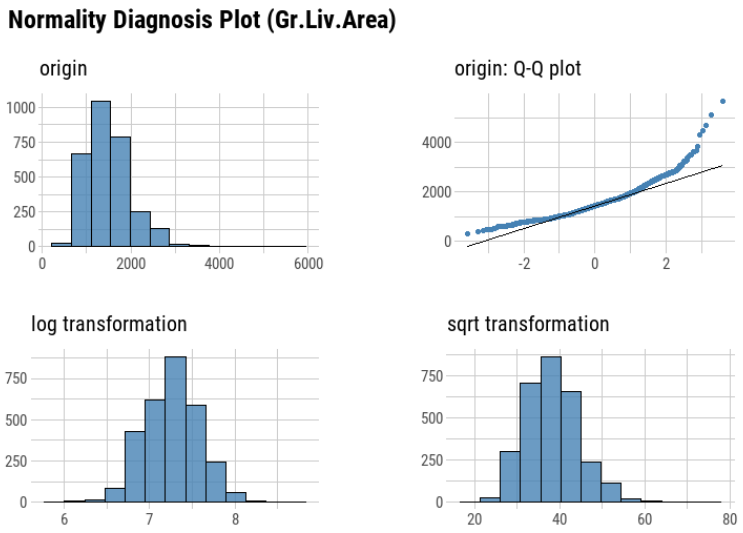


Figure 12 Normality plot of above ground living area

It is not surprising that a square root transformation restores to normality the Above Ground Living Area variable given its near quadratic distribution on the Q-Q plot.

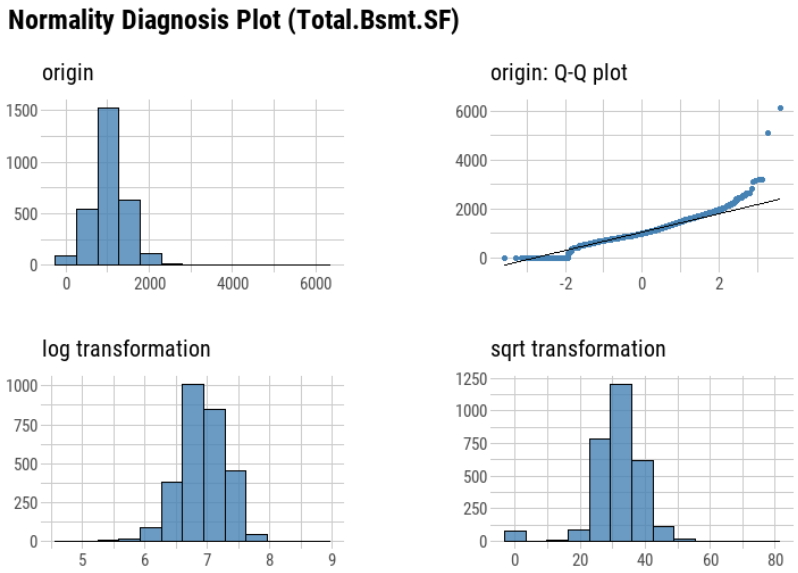


Figure 13 Normality plot of Total Basement Area

The impact of outliers in the data is noticed in the Total Basement Area. Again, zero values represent the absence of a basement in the houses.

Box plots of the numerical continuous variables

Outlier analysis is extremely important prior to linear regression and correlation analysis, since outliers may impact the distribution of data and distort the assumptions of linear regression. We earlier noticed that the important variables had numerous outliers that were computed using z-score analysis.

- **Gr.Liv.Area**: 74 outliers
- **Total.Bsmt.SF**: 123 outliers
- **Garage.Area**: 42 outliers
- **Lot.Frontage**: 238 outliers
- **Lot.Area**: 127 outliers
- **SalePrice**: 137 outliers
- **Age**: 9 outliers

These outliers can significantly alter the sensitivity of regression analysis. So, I carried out a detailed analysis of outliers in the variables using boxplots and envisioned what the distribution of the variable would be without outliers. **Please note that I did not remove these outliers from the dataset and these graphs only represent an automated visualization of potential outlier removal.**

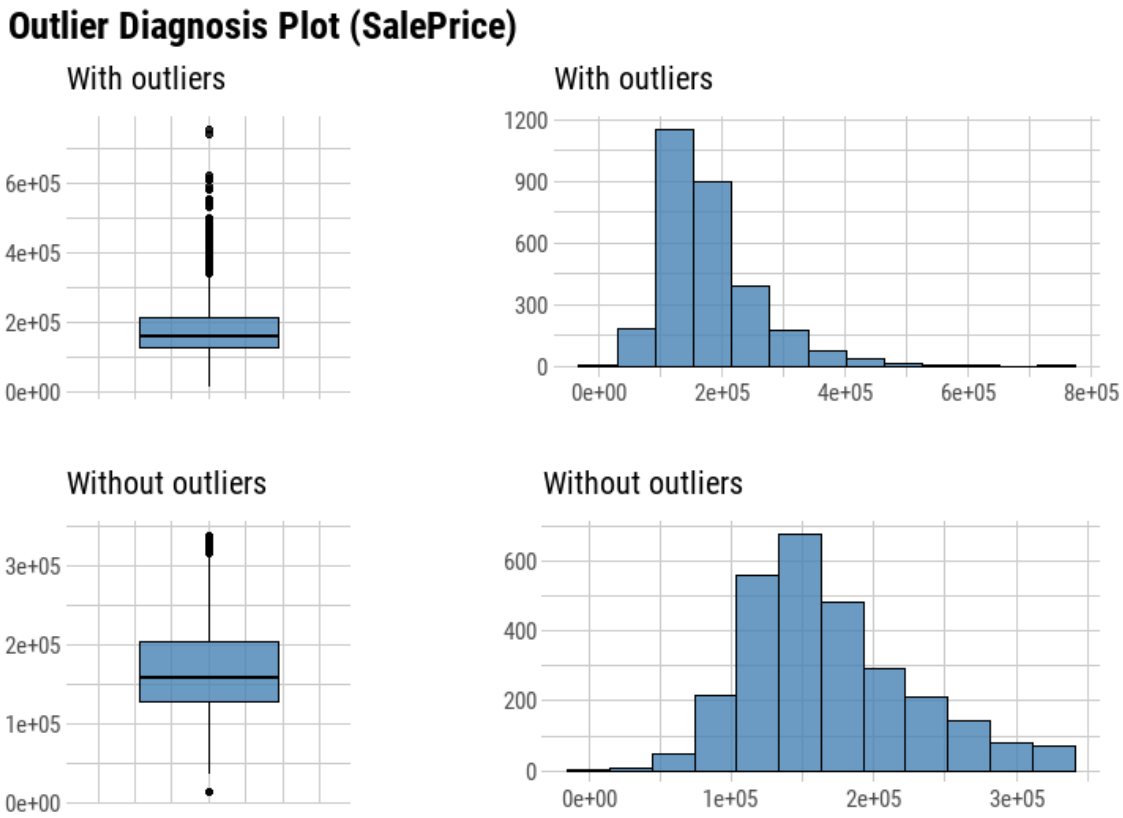


Figure 14 Outlier Diagnosis and Box Plot of Sale Price

The boxplot 'With outliers' shows a significant presence of extreme values above the upper whisker, suggesting a heavy-tailed distribution to the right. These outliers exert substantial influence on the distribution, potentially violating the assumption of homoscedasticity—a key condition for linear regression that requires equal variance across all levels of the predictor variables.

Removing the outliers yields a more symmetric boxplot, indicating a distribution that adheres more closely to normality. The histogram 'Without outliers' exhibits a bell-shaped curve, although still slightly skewed, it suggests an improved alignment with the normality assumption.

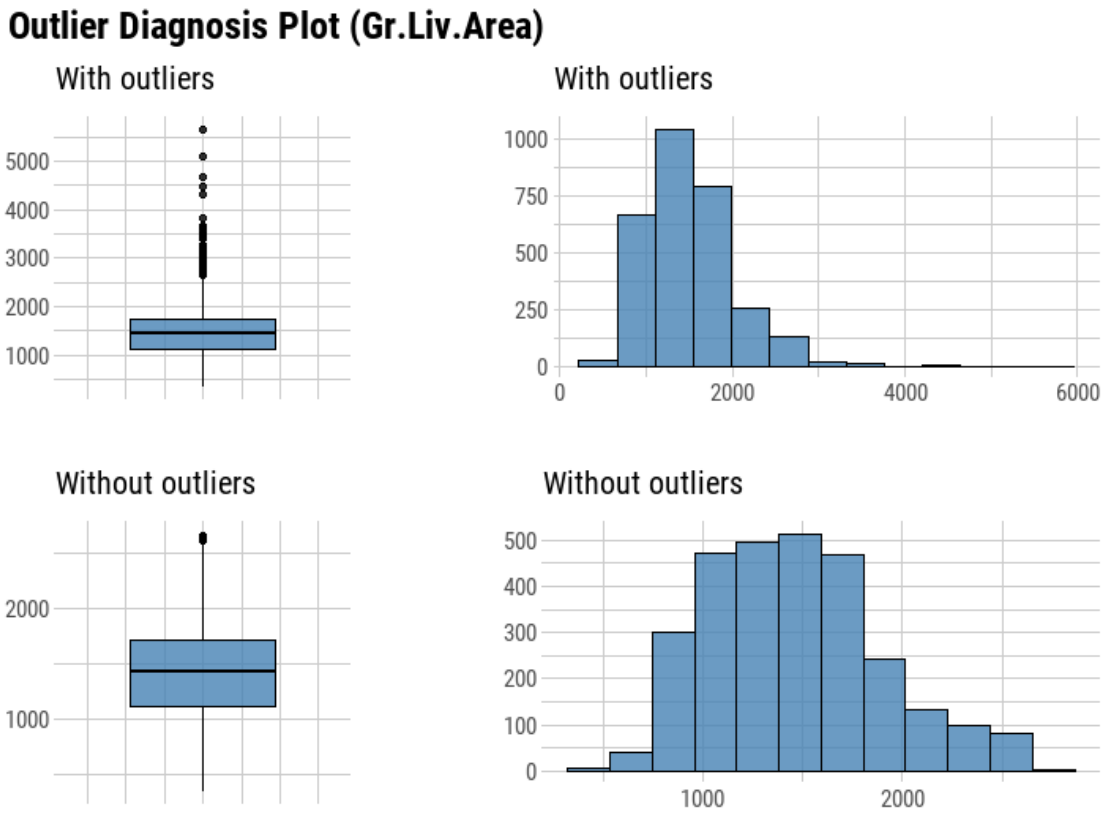
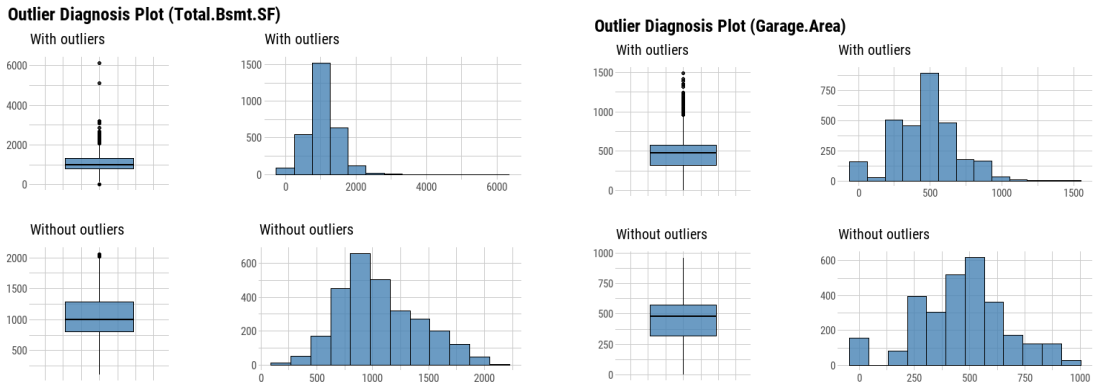


Figure 15 Outlier Diagnosis and Box Plot of Above ground Living Area

The boxplot and histogram of **Gr.Liv.Area** with outliers depict pronounced right skewness, highlighting the presence of extreme values. Excluding outliers results in distributions that more closely approximate normality, a key assumption of linear regression. This adjustment may enhance the potential for homoscedasticity and the accuracy of model predictions.



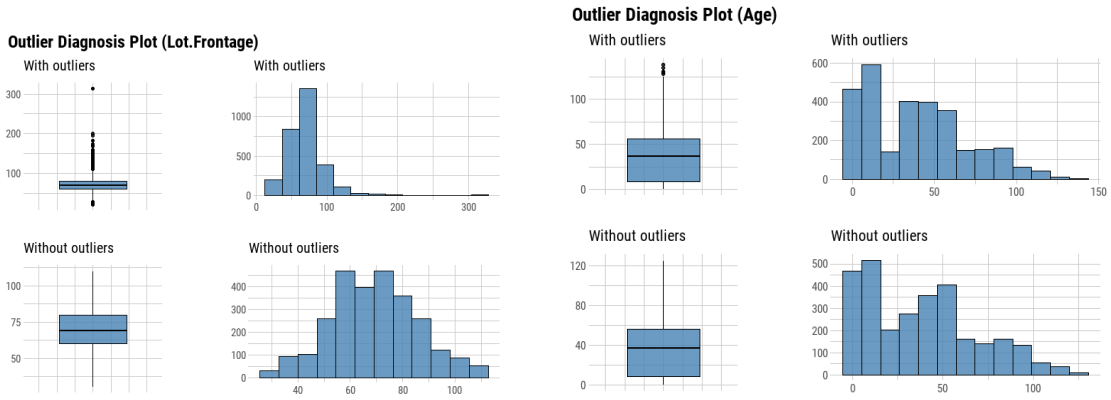


Figure 16 Box Plots and Outlier Analysis of other Variables

To avoid prolixity, I briefly describe the above box plots of the remaining important continuous variables below succinctly.

Total.Bsmt.SF: Initial distribution displays significant right skew; outliers' removal visibly centralizes the median, aiding in linear regression assumptions.

Garage.Area: Outliers stretch the right tail; their exclusion coheres the data, offering a modest symmetry conducive to homoscedasticity in regression analysis.

Lot.Frontage: Heavy right tail with outliers indicates potential heteroscedasticity; post-exclusion, distribution approaches normality, aligning better with regression prerequisites.

Age: With outliers, data hints at left skewness; without outliers, distribution improves but retains slight skew, which could influence regression diagnostics subtly.

Feature Engineering

To feature engineering initiatives have been mentioned already. Namely, adding a log transformed sale price column and secondly adding an age column derived from ‘year’ house was built. The third and arguably the most important feature is the decision to remove outliers. This decision cannot be taken lightly as removing outliers rashly may distort the entire dataset and render any regression analysis moot. It is therefore prudent to err on the side of conservatism as far as outlier removal is concerned. In this spirit, even though I knew that the continuous variables had numerous outliers, I wanted to confirm it visually using scatterplots. The ggpairs() function of the GGally package in R offers a quick scatterplot across the variables of the dataset.

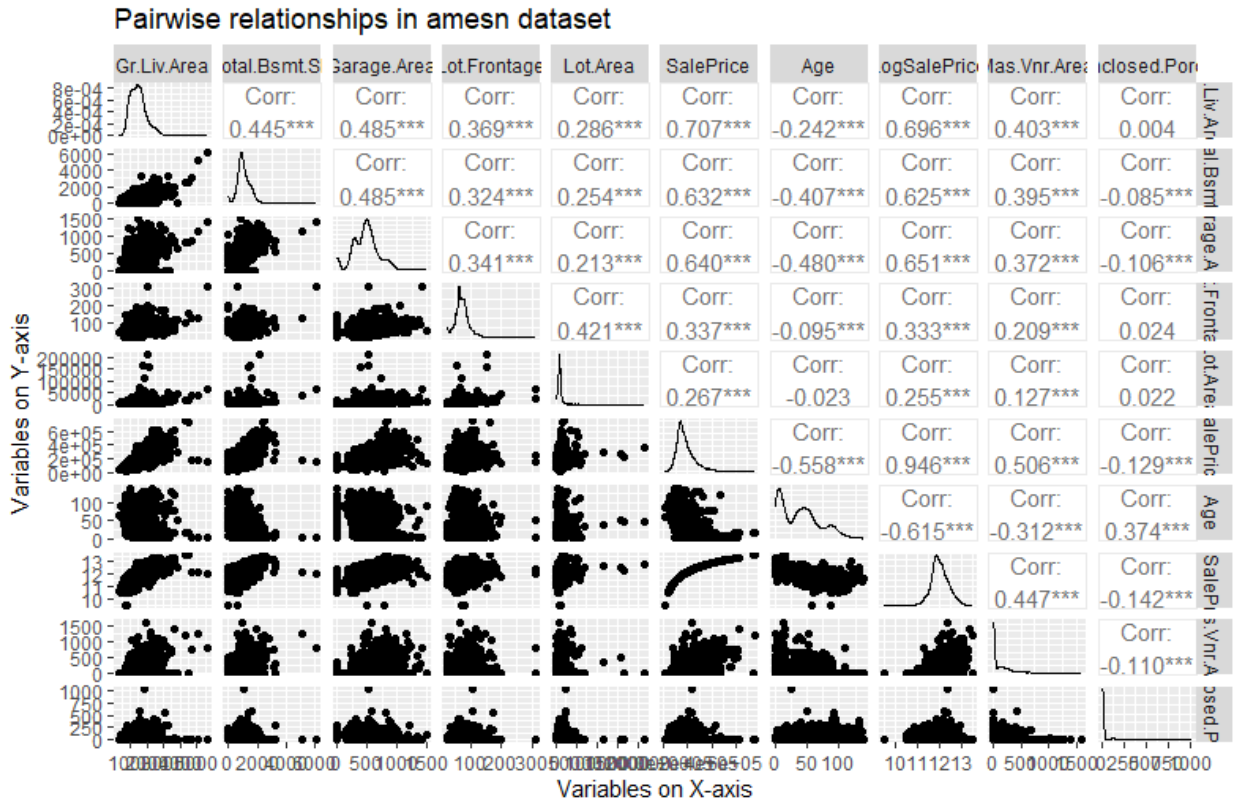


Figure 17 Pair plots of numerical variables. Notice the outliers in the scatter plots.

The scatterplots raised my concern about outliers in the interaction between Gr.Liv.Area (Above Ground Living Area) and Saleprice. I was impelled to take a closer look.

```
> outliers <- ames$Gr.Liv.Area > 4000 & ames$SalePrice < 300000
>   ames[outliers,]
     Order       PID MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Lot.Shape Land.Contour Utilities Lot.Config Land.Slope
1499  1499 908154235          60        RL          313    63887   Pave       IR3          Bnk    AllPub     Corner        Gtl
2181  2181 908154195          20        RL          128    39290   Pave       IR1          Bnk    AllPub     Inside        Gtl
2182  2182 908154205          60        RL          130    40094   Pave       IR1          Bnk    AllPub     Inside        Gtl
     Neighborhood Condition.1 Condition.2 Bldg.Type House.Style Overall.Qual Overall.Cond Year.Built Year.Remod.Add Roof.Style
1499      Edwards       Feedr       Norm      1Fam      2Story           10           5       2008          2008        Hip
2181      Edwards        Norm        Norm      1Fam      1Story           10           5       2008          2009        Hip
2182      Edwards        PosN        PosN      1Fam      2Story           10           5       2007          2008        Hip
     Roof.Matl Exterior.1st Exterior.2nd Mas.Vnr.Type Mas.Vnr.Area Exter.Qual Exter.Cond Foundation Bsmt.Qual Bsmt.Cond
1499   ClyTile       Stucco       Stucco         Stone          796         Ex         TA       PConc         Ex         TA
2181   CompShg     CemntBd      CmentBd         Stone         1224         Ex         TA       PConc         Ex         TA
2182   CompShg     CemntBd      CmentBd         Stone          762         Ex         TA       PConc         Ex         TA
     Bsmt.Exposure BsmtFin.Type.1 BsmtFin.SF.1 BsmtFin.Type.2 BsmtFin.SF.2 Bsmt.Unf.SF Total.Bsmt.SF Heating Heating.QC
1499            Gd            GLQ         5644          Unf             0         466          6110    GasA         Ex
2181            Gd            GLQ         4010          Unf             0        1085          5095    GasA         Ex
2182            Gd            GLQ         2260          Unf             0         878          3138    GasA         Ex
     Central.Air Electrical X1st.Flr.SF X2nd.Flr.SF Low.Qual.Fin.SF Gr.Liv.Area Bsmt.Full.Bath Bsmt.Half.Bath Full.Bath
1499           Y      SBrkr        4692         950               0        5642              2              0         2
2181           Y      SBrkr        5095           0               0        5095              1              1         2
2182           Y      SBrkr        3138        1538               0        4676              1              0         3
     Half.Bath Bedroom.AbvGr Kitchen.AbvGr Kitchen.Qual TotRms.AbvGrd Functional Fireplaces Garage.Type Garage.Yr.Blt
1499         1            3            1           Ex           12        Typ          3      Attchd          2008
2181         1            2            1           Ex           15        Typ          2      Attchd          2008
2182         1            3            1           Ex           11        Typ          1     BuiltIn          2007
     Garage.Finish Garage.Cars Garage.Area Garage.Qual Garage.Cond Paved.Drive Wood.Deck.SF Open.Porch.SF Enclosed.Porch
1499           Fin           2        1418          TA          TA          Y          214           292              0
2181           Fin           3        1154          TA          TA          Y          546           484              0
2182           Fin           3         884          TA          TA          Y          208           406              0
     X3Ssn.Porch Screen.Porch Pool.Area Misc.Val Mo.Sold Yr.Sold Sale.Type Sale.Condition SalePrice Age LogSalePrice
1499           0           0       480        0       1    2008       New       Partial    160000   2     11.98293
2181           0           0         0    17000      10    2007       New       Partial    183850   2     12.12188
2182           0           0         0        0      10    2007       New       Partial    184750   3     12.12676
```

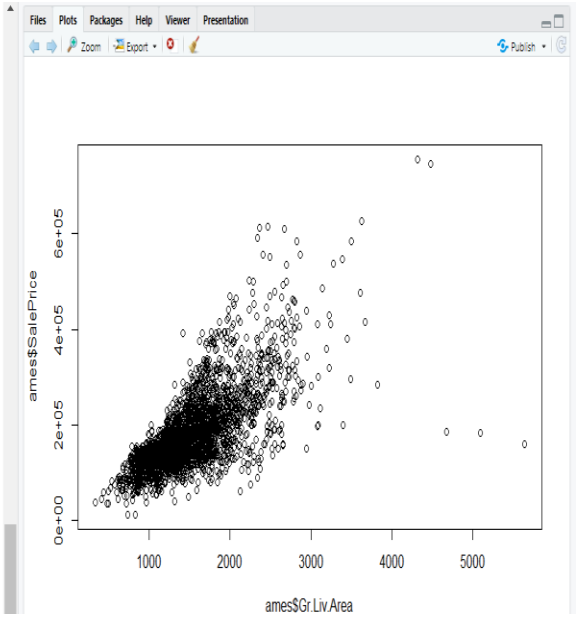


Figure 18 The three most influential outliers

The three most influential outliers in above grade living area were numbered 1499, 2181 and 2182. Interestingly, all of them were from the same neighborhood.

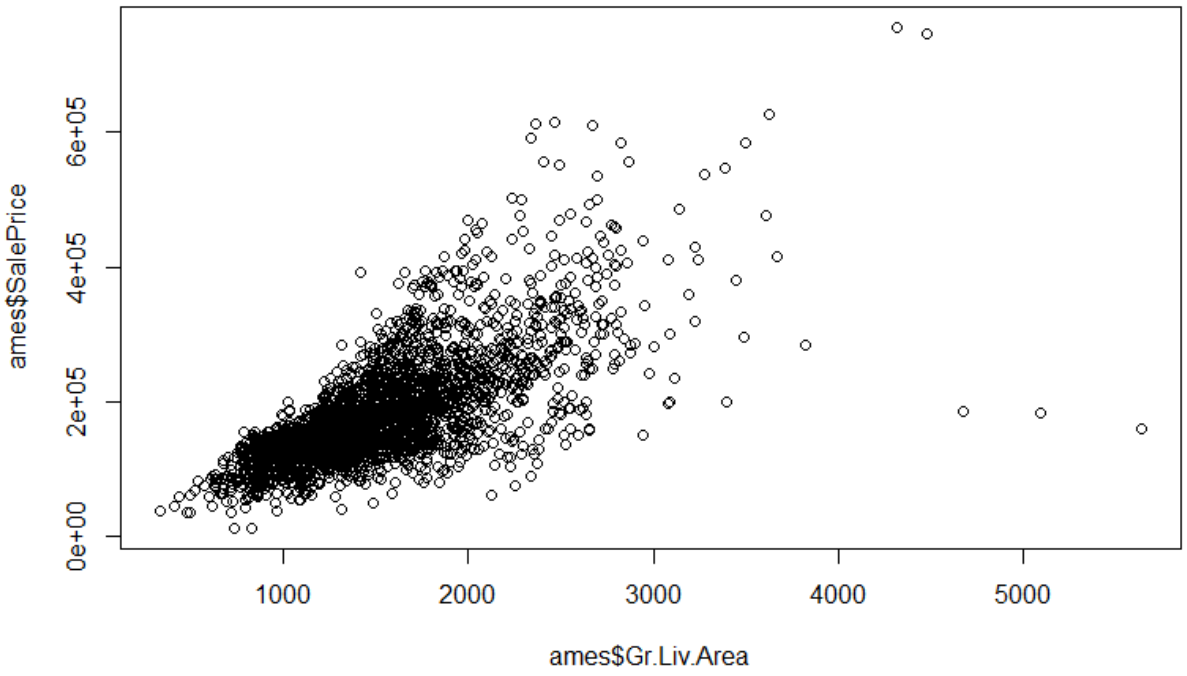


Figure 19 Outliers in the scatter plot between sale price and above ground living area

The scatterplot showed 5 outliers with 3 being particularly influential. I inspected these 5 houses closely to find patterns among them. I looked into whether they were from the same neighborhood. It turned out that 3 of the most influential outlier houses had the same neighborhood, namely 'Edwards', but the mean above ground living area in this neighborhood (1338 square feet) failed to explain the huge sizes of these houses at a relatively usual price.

```
> ames %>% filter( Gr.Liv.Area > 4000 ) %>% arrange( SalePrice )
  Order       PID MS.SubClass MS.Zoning Lot.Frontage Lot.Area Str
1  1499 908154235          60        RL           313    63887   P
2  2181 908154195          20        RL           128    39290   P
3  2182 908154205          60        RL           130    40094   P
4  1761 528320050          60        RL           160    15623   P
5  1768 528351010          60        RL           104    21535   P
  Neighborhood Condition.1 Condition.2 Bldg.Type House.Style Over
1      Edwards       Feedr        Norm      1Fam      2Story
2      Edwards        Norm        Norm      1Fam      1Story
3      Edwards        PosN        PosN      1Fam      2Story
4      NoRidge        Norm        Norm      1Fam      2Story
5      NoRidge        Norm        Norm      1Fam      2Story
  Roof.Matl Exterior.1st Exterior.2nd Mas.Vnr.Type Mas.Vnr.Area E
1   ClyTile       Stucco       Stucco        Stone          796
```

Figure 20 Inspecting the outliers

I then fitted two simple linear regression models on the dataset with and without these 3 influential outliers to observe their pull using the abline() function.

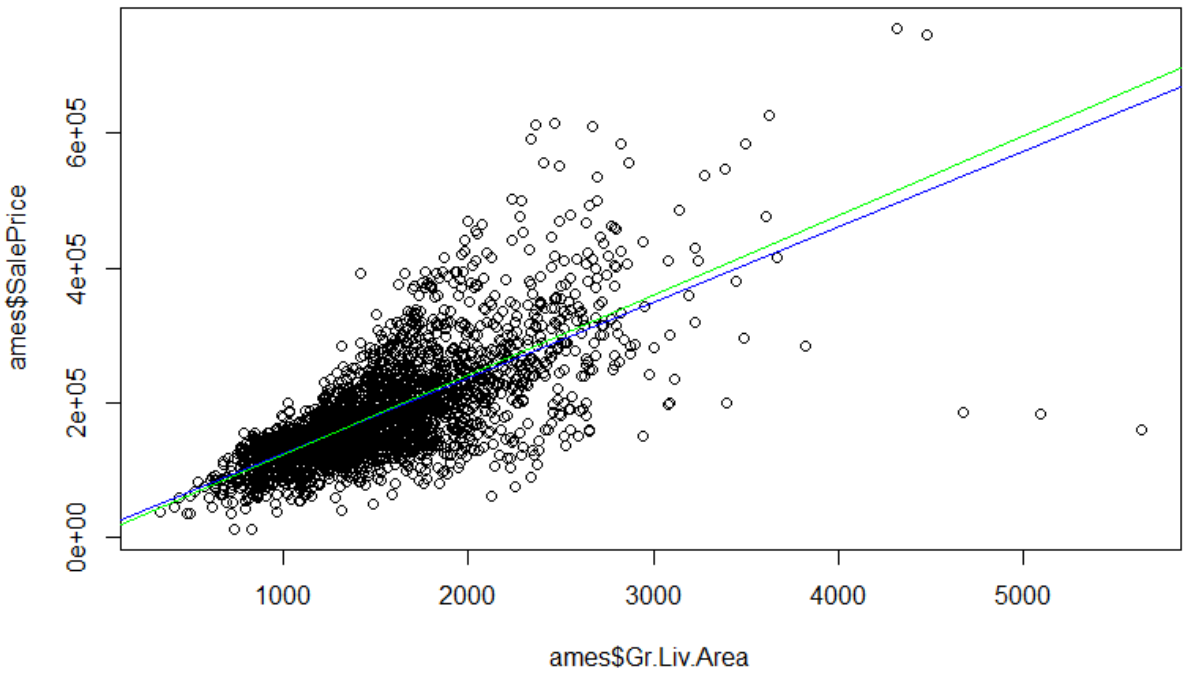


Figure 21 The influence of the outliers observed using two simple linear regression models

As feared, the 3 outliers were exerting a pull owing to their influence on the simple linear model. This pull and the lack of explanation of the large sizes of these outliers by their neighborhood convinced me that these 3 outliers can be safely removed without adversely impacting the integrity of the broader dataset.

Correlation Analysis

	Gr.Liv.Area	Total.Bsmt.SF	Garage.Area	Lot.Frontage	Lot.Area	SalePrice	Age	Mas.Vnr.Area	Enclosed.Porch	LogSalePrice
Gr.Liv.Area	1.000000000	0.40753145	0.4763335	0.34499538	0.25966526	0.7271216	-0.23955378	0.3833424	0.006864981	0.7144387
Total.Bsmt.SF	0.407531452	1.00000000	0.4776089	0.28872229	0.22098101	0.6604202	-0.41456264	0.3742450	-0.085680415	0.6510217
Garage.Area	0.476333535	0.47760891	1.0000000	0.32983602	0.19976807	0.6443566	-0.47908396	0.3634971	-0.105559631	0.6545465
Lot.Frontage	0.344995375	0.28872229	0.3298360	1.00000000	0.40384231	0.3454589	-0.09082505	0.1923660	0.026805193	0.3401986
Lot.Area	0.259665263	0.22098101	0.1997681	0.40384231	1.00000000	0.2705181	-0.01761172	0.1073867	0.023932522	0.2579146
SalePrice	0.727121646	0.66042024	0.6443566	0.34545894	0.27051811	1.0000000	-0.55891815	0.5120099	-0.128818682	0.9463209
Age	-0.239553775	-0.41456264	-0.4790840	-0.09082505	-0.01761172	-0.5589181	1.00000000	-0.3099925	0.374224122	-0.6157857
Mas.Vnr.Area	0.383342355	0.37424501	0.3634971	0.19236600	0.10738674	0.5120099	-0.30999252	1.0000000	-0.110046025	0.4514051
Enclosed.Porch	0.006864981	-0.08568042	-0.1055596	0.02680519	0.02393252	-0.1288187	0.37422412	-0.1100460	1.000000000	-0.1424479
LogSalePrice	0.714438686	0.65102171	0.6545465	0.34019861	0.25791462	0.9463209	-0.61578574	0.4514051	-0.142447877	1.0000000

Figure 22 Correlation Matrix

I created the above correlation matrix for the important continuous variables in the dataset. Although it is not best practice to include more than 10 variables in a correlation matrix, however, for the plot of the correlation matrix (seen below) I decided to include all the numeric variables so as to be able to answer the subsequent questions on making scatterplots for the least, most and variable with correlation closest to 0.5 with respect to sale price.

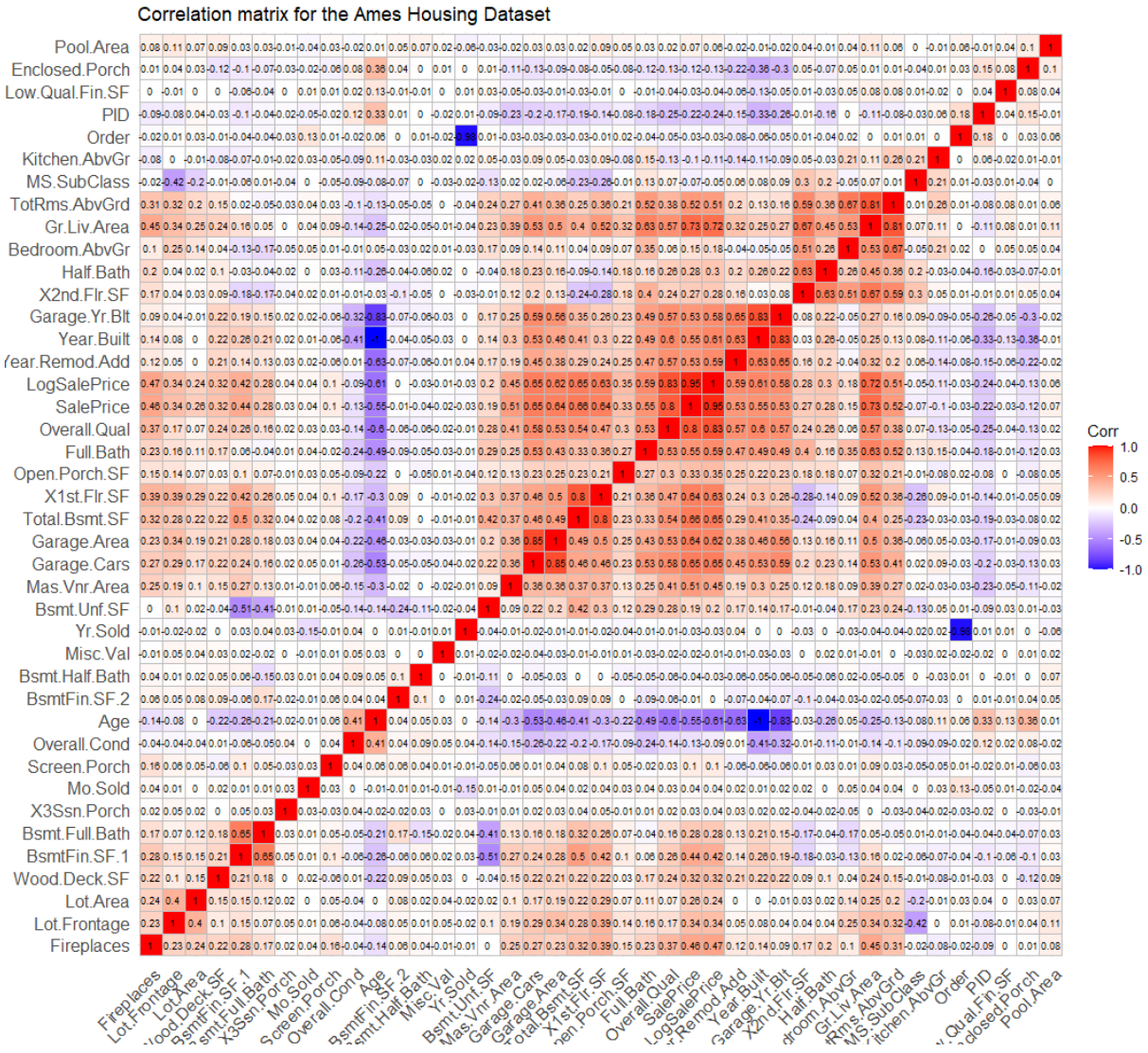


Figure 23 Correlation Matrix plot

Gr.Liv.Area exhibits a strong positive correlations with **SalePrice** (0.727 respectively), indicating good predictive potential for regression models. Conversely, **Age** demonstrates a notable negative correlation with **SalePrice** (-0.559), suggesting that as houses age, their value tends to decrease. Variables such as **Lot.Frontage** and **Enclosed.Porch** show weak correlations with **SalePrice** (0.345 and -0.128 respectively), reflecting a poor association and potentially limited explanatory power within a linear regression framework for predicting sale price.

I drew two conclusions from the correlation analysis-

1. Not to use the log transform of the sale price for regression modelling since its correlation was not particularly better with the important variables that the original sale price. Also, log transformation would be problematic as far as interpreting the coefficients of the regression model is concerned. Exponentiation would complicate interpretation.

2. To use the following continuous variables in my regression model based on a relatively strong correlation with the response variable (sale price) - Gr.Liv.Area (0.727), Total.Bsmt.SF (0.66), X1st.Flr.SF(0.64) , Garage.Area(0.64).

Figure 24Scatterplot for varaible with best correlation

The above scatterplot shows **Gr.Liv.Area**, exhibiting the highest positive correlation (0.727)with SalePrice, with a pattern indicating that as living area increases, so does sale price, albeit with increasing variability. It is also discernable that a large part of the data is scrunched up in the lower sale price range. A clear linear pattern is also visible, although the variation in the area seems to increase with incresing sale price, indicating factors extraneous to above ground living area contributing more as far as the prices of the costlier houses are concerned. Perhaps- it would not be unreasonable to suppose- the neighborhood or access to amenities is playing a role in the higher price ange.

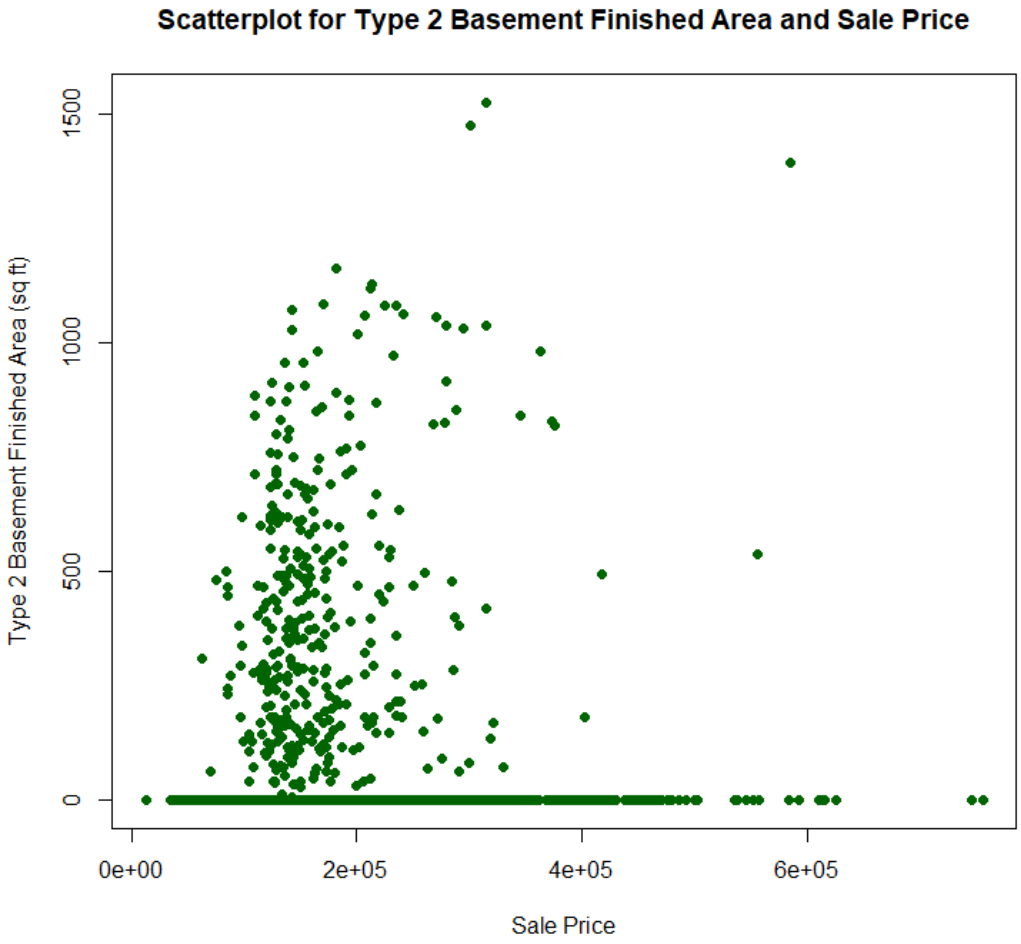


Figure 25 Scatterplot for variable with least correlation

This above scatterplot of **Type 2 Basement Finished Area** displays the lowest correlation with SalePrice (-0.01), with a much weaker and more dispersed relationship, suggesting that this variable has little impact on sale price. The weaker the correlation between variables the more the scatterplot resembles 'noise' or utter 'randomness' in distribution patterns.

Intuitively basement area may be expected to play a role in determining sale price, however the zero values in the dataset indicating that a large number of houses do not even have a basement may be producing this weak correlation. Also, it is interesting to note that according to Fanny Mae and the American National Standards Institute (ANSI) guidelines house evaluation and appraisal ought not to take into account the basement area. Only the above-grade floor space ought to be included in Gross Living Area (GLA).

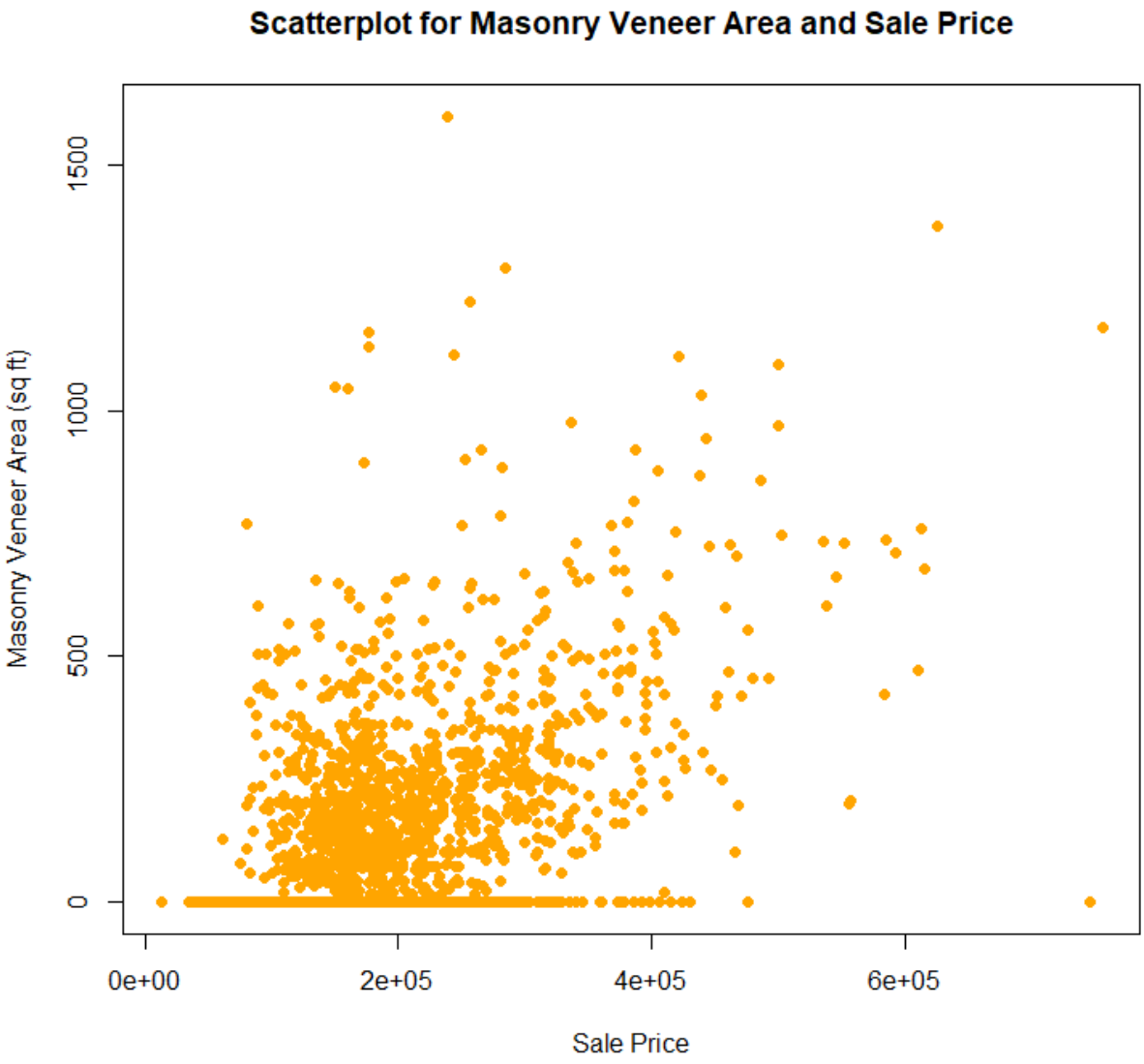


Figure 26 Scatterplot for vaariable with correlation closest to 0.5

This third scatterplot, representing **Masonry Veneer Area**, aligns with a correlation near 0.5 i.e. 0.51 to be exact. This indicates a moderate positive relationship. However, the data points are more scattered around the trend line than in the first plot, reflecting a less consistent effect on SalePrice.

Each of the above plots demonstrates the importance of variable selection in predictive modeling, as the strength and pattern of the relationship can significantly influence model performance.

Question 7

1. **Using at least 3 continuous variables, fit a regression model in R.**

   I used the call below to fit a regression model in R

   model <- lm(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + X1st.Flr.SF + Garage.Area, data = amesclean)

I used the four continuous variables based on a relatively strong correlation with the response variable (sale price) –
Gr.Liv.Area (0.727)- This represents the Above grade Living Area in Square Feet. It has been notified by Fanny Mae and the ANSI (American National Standards Institute) as the most important aspect of the appraisal of the sale price of a house.
Total.Bsmt.SF (0.66) – Total Basement Area in Square Feet. This represents the basement area. It may have value 0 for houses with no basement.
X1st.Flr.SF(0.64) - First Floor square feet. This may have multicollinearity with Gr.Liv.Area. But their correlation was moderate (0.52), and so I included it in the model.
Garage.Area(0.64) - Size of garage in square feet.

```
>   model <- lm(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + X1st.Flr.SF + Garage.Area, data = amesclean)
>   summary(model)

Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + X1st.Flr.SF +
    Garage.Area, data = amesclean)

Residuals:
    Min      1Q  Median      3Q     Max
-212642  -20084     815   20781  247590

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.610e+04  2.784e+03 -16.555   <2e-16 ***
Gr.Liv.Area    7.514e+01  1.932e+00  38.902   <2e-16 ***
Total.Bsmt.SF  6.622e+01  2.984e+00  22.190   <2e-16 ***
X1st.Flr.SF   -2.818e-01  3.546e+00  -0.079    0.937
Garage.Area    9.623e+01  4.358e+00  22.079   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41360 on 2922 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.7322
F-statistic:  2001 on 4 and 2922 DF,  p-value: < 2.2e-16

>   coefficients(model)
  (Intercept)   Gr.Liv.Area Total.Bsmt.SF   X1st.Flr.SF   Garage.Area
-4.609705e+04  7.514213e+01  6.621780e+01 -2.817609e-01  9.622778e+01
```

Figure 27 Summary of the regression model

The above output in the console of the R Studio represents the the raw or unstandardized coefficients from the regression output to show the impact of each independent variable on the dependent variable.

I created the following equation for this model .

SalePrice = -46090 + 75.14*Gr.Liv.Area + 66.28*Total.Bsmt.SF + (-0.8178)*X1st.Flr.SF + 96.23*Garage.Area

With the predictor variables enlarged to include more information about their meaning the equation is –

SalePrice = -46090 + 75.14 x Ground Living Area in Square Feet + 66.28 x Total Basement Area in Square Feet + (-0.8178) x First Floor Square Feet+ 96.23 x Garage Area in Square Feet.

Interpretation of the Coefficients of the model within the context of the Dataset

SalePrice is the dependent variable representing the price of the house in Dollars. The Gr.Liv.Area (above ground living area), Total.Bsmt.SF (total basement area), X1st.Flr.SF (first-floor area), and Garage.Area are the independent variables representing different square footage measurements of the house.

Intercept (Constant Term): The intercept value of -46,090 suggests the baseline value for the sale price. In practical terms, this negative value is not feasible as it would imply a negative sale price. However, in regression analysis, the intercept often does not have a practical interpretation unless all independent variables are zero and that condition itself is meaningful within the context of the dataset. Instead, the intercept adjusts the regression plane for the mean values of the independent variables. This explains the negative value here.

Coefficients:

- Gr.Liv.Area: A coefficient of 75.14 indicates that, holding other factors constant, each additional square foot of above-ground living area is associated with an increase of approximately $75.14 in the sale price of the house.
- Total.Bsmt.SF: A coefficient of 66.28 suggests that for each additional square foot of basement area, there is an increase of roughly $66.28 in the sale price, *ceteris paribus*.
- X1st.Flr.SF: The negative coefficient of -0.8178 for first-floor square footage suggests a marginal decrease in sale price with increasing first-floor area. This could be due to multicollinearity with other floor area variables or may indicate that larger first floors are not as valued in the Ames housing market because, counterintuitively, a greater First Floor Area might be eating into other aspects of the house area the buyers deem more important, such as Garage Area as our model suggests.
- Garage.Area: The positive coefficient of 96.23 suggests that each additional square foot of garage area contributes $96.23 to the sale price, all else equal. This indicates that garage space is highly valued in the Ames housing market.

P-Values and Significance: The p-values associated with the coefficients indicate the probability that the corresponding coefficient is actually zero (no effect), under the assumption that the null hypothesis of no effect is true.

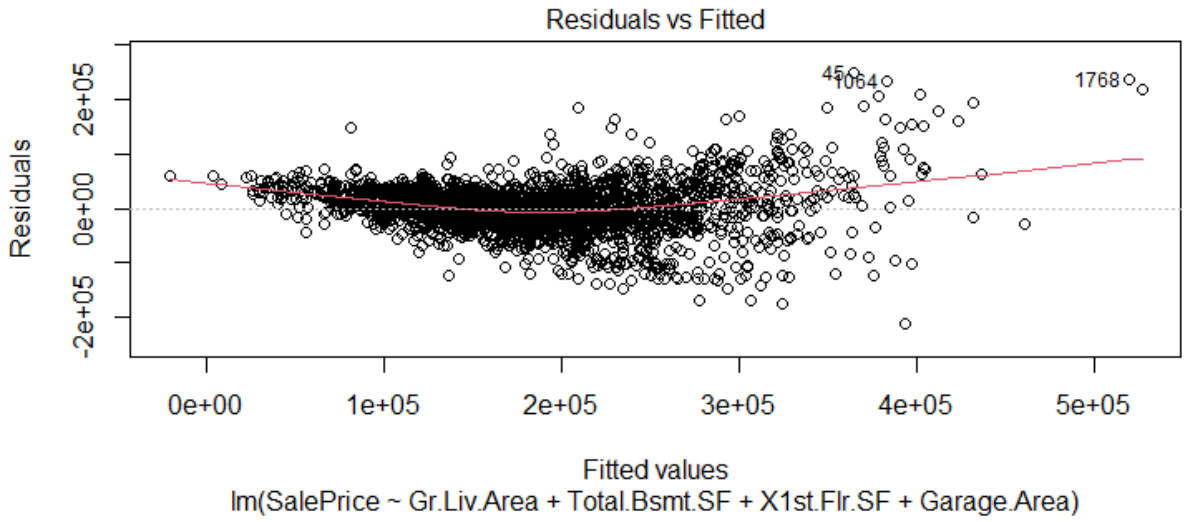
- For Gr.Liv.Area, Total.Bsmt.SF, and Garage.Area, the p-values are extremely low (close to zero), indicating that we can reject the null hypothesis of no effect with a high degree of confidence. These variables are statistically significant predictors of SalePrice.
- The p-value for X1st.Flr.SF is 0.937, which is much higher than the common alpha level of 0.05, suggesting that we cannot reject the null hypothesis for this coefficient. This implies that the first-floor square footage **may not be a statistically significant predictor** of SalePrice in the presence of the other variables.

Adjusted R-Squared: The Adjusted R-squared value is 0.7322, which indicates that approximately 73.22% of the variability in the sale prices can be explained by the model. This is a relatively high value and suggests a good fit of the model to the data. The adjustment in Adjusted R-squared accounts for the number of predictors in the model, thus providing a more accurate measure of model fit than the non-adjusted R-squared when comparing models with a different number of independent variables.

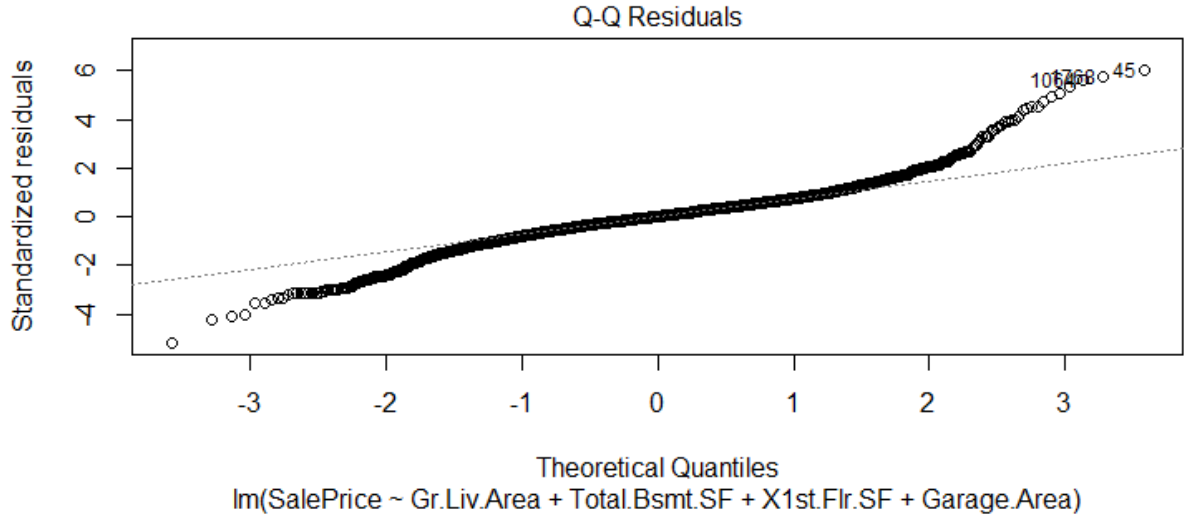
ALY 6015 Module 1 Regression Diagnostics in R by Syed Faizan

Dataset Context: The Ames Housing dataset includes a wide range of houses with diverse features. The regression model quantifies the relationship between the house size (as measured by square footage in different areas) and the sale price.

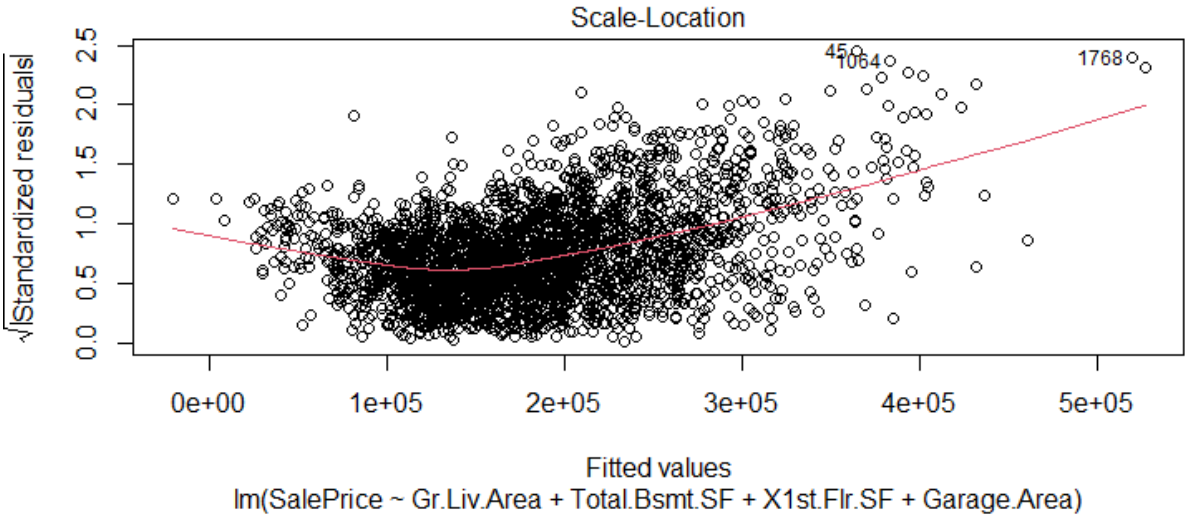
I include below the four graphs of the model and their diagnostic interpretation.



The Residuals vs. Fitted plot is utilized to evaluate the regression model's fit. The ideal pattern is a random scatter of points with a flat red loess line, indicating linearity and homoscedasticity. Here, the curve of the red line suggests non-linearity. A slight funnel shape implies heteroscedasticity, with variance increasing for higher fitted values. Some potential outliers are evident, particularly at higher fitted values. However, the random dispersion of points is consistent with the assumption of independent errors. The plot indicates the model captures SalePrice variability but may benefit from addressing non-linearity, heteroscedasticity, and outliers, potentially through including interaction terms or examining influential data points.

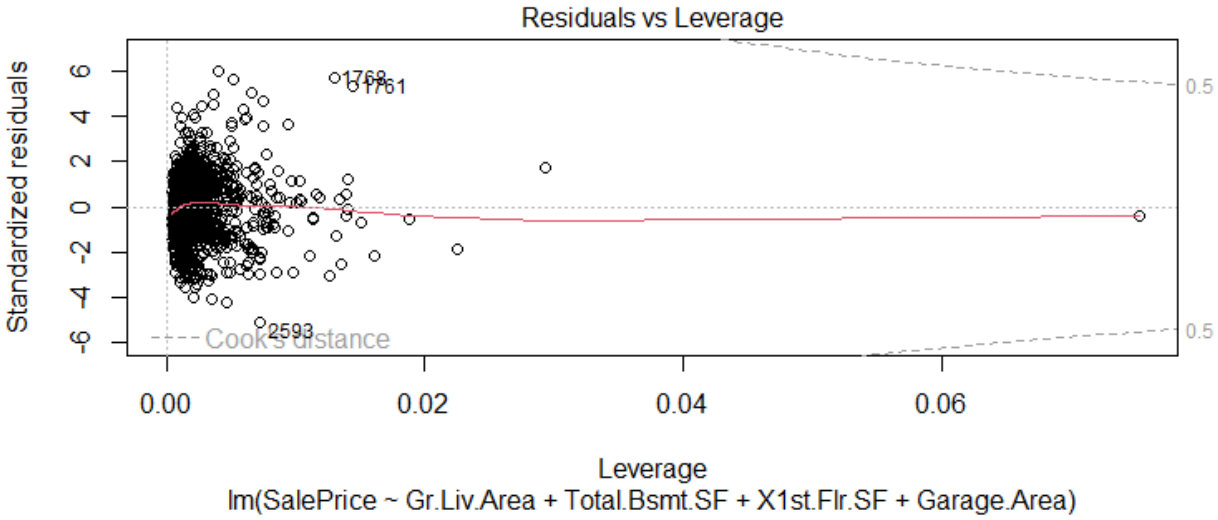


The above Q-Q plot is a graphical tool to assess the normality of residuals in my regression model predicting SalePrice. Residuals aligning with the 45-degree reference line would indicate normal distribution. Here, the deviation from the line, especially at the higher end, suggests that while my model does a good job of predicting Sale price in the middle value range my model is less reliable at the lower and particularly, the higher end of the sale price values. The heavier tails also indicate that extreme SalePrice values are more frequent than a normal distribution would expect. This deviation from normality could affect the reliability of confidence intervals and hypothesis tests associated with the model.



The Scale-Location plot, also known as a Spread-Location plot, is used to assess the homoscedasticity of residuals in the regression model for SalePrice. The increasing spread of residuals at higher fitted values indicates heteroscedasticity, suggesting that the variance of the SalePrice predictions is not constant. The red line's upward trend confirms this pattern, implying

that prediction accuracy varies across the range of SalePrice, with less precision at higher SalePrice values.



The Residuals vs. Leverage plot helps identify influential observations that might have a disproportionate impact on the regression model for SalePrice. Points outside the Cook's distance lines may unduly influence the model's parameters. Here, most data points are within the Cook's distance threshold, suggesting limited influence. However, a few points with higher leverage and significant residuals indicate potential influence on the model's predictions of SalePrice, warranting further investigation as carried out in a subsequent section below.

```
Gr.Liv.Area Total.Bsmt.SF   X1st.Flr.SF   Garage.Area
   1.544403      2.717442      3.070391      1.486286
```

The table above displays the Variance Inflation Factors (VIF) for the predictors in the SalePrice regression model. VIF assesses multicollinearity, with values exceeding 5 or 10 typically indicating a problematic level of collinearity. In this model, all VIF values are below these thresholds, suggesting multicollinearity is not substantially inflating the variance of the estimated coefficients. The model's predictors (Gr.Liv.Area, Total.Bsmt.SF, X1st.Flr.SF, Garage.Area) thus appear to provide independent information in predicting the SalePrice without undue collinearity.

To correct multicollinearity:

1. I would first calculate VIFs as I have done above; values over 5 or 10 suggest issues.
2. Drop correlated predictors or combine them.
3. Apply PCA to reduce dimensions and eliminate correlations.
4. Use Ridge Regression to penalize large coefficients.

5. If possible, increase the sample size to diminish multicollinearity.
6. Center variables by subtracting their means.
7. Try Partial Least Squares Regression for many predictors.
8. Standardize variables to z-scores.
9. Employ regularization like LASSO for variable selection.
10. Use domain expertise to inform variable retention decisions.

One must bear in mind that while multicollinearity impacts coefficient estimates it is less impactful when it comes to predictive accuracy. The purpose of the regression model thus informs our decision as well.

	rstudent	unadjusted p-value	Bonferroni p
45	6.034082	1.8001e-09	5.2689e-06
1768	5.763950	9.0750e-09	2.6562e-05
1064	5.662880	1.6332e-08	4.7805e-05
1761	5.347055	9.6293e-08	2.8185e-04
2593	-5.182192	2.3422e-07	6.8557e-04
433	5.077126	4.0723e-07	1.1920e-03
434	4.972865	6.9784e-07	2.0426e-03
2446	4.715516	2.5241e-06	7.3880e-03
2333	4.542498	5.7845e-06	1.6931e-02
2335	4.506953	6.8350e-06	2.0006e-02

Figure 28 Outliers in the model

The above output from calling the outlierTest() in R on our model presents the studentized residuals with their unadjusted and Bonferroni-adjusted p-values, identifying potential outliers in the regression model for SalePrice. Observations with large absolute studentized residuals and small p-values, particularly after Bonferroni adjustment, suggest they are statistically significant outliers. These points merit closer investigation; they may be candidates for removal or further analysis to ensure they do not unduly influence the model. I therefore proceeded to examine them further.

Whether these observations ought to be removed or left as they are depends upon several factors. General rules-of-thumb to examine and remedy outliers in a regression model are as follows-

1. **Influence**: If the outliers are highly influential to the model, meaning that their removal would significantly change the coefficients, they may be considered for removal.
2. **Substantive Reasons**: If there is a substantive reason or an error (e.g., data entry mistake) behind the outlier, it should be removed or corrected.
3. **Model Fit**: If the outliers are affecting the overall fit and predictive accuracy of the model, you might consider removing them.
4. **Representativeness**: If the outliers represent rare but possible scenarios in the data, they should be kept to maintain the model's representativeness.
5. **Robustness**: Alternatively, instead of removing outliers, you could use a robust regression method that is less sensitive to outliers.

In order to examine the infuence that these outliers were exerting I decided to plot the Hat Value and Cook's Distance on plots.

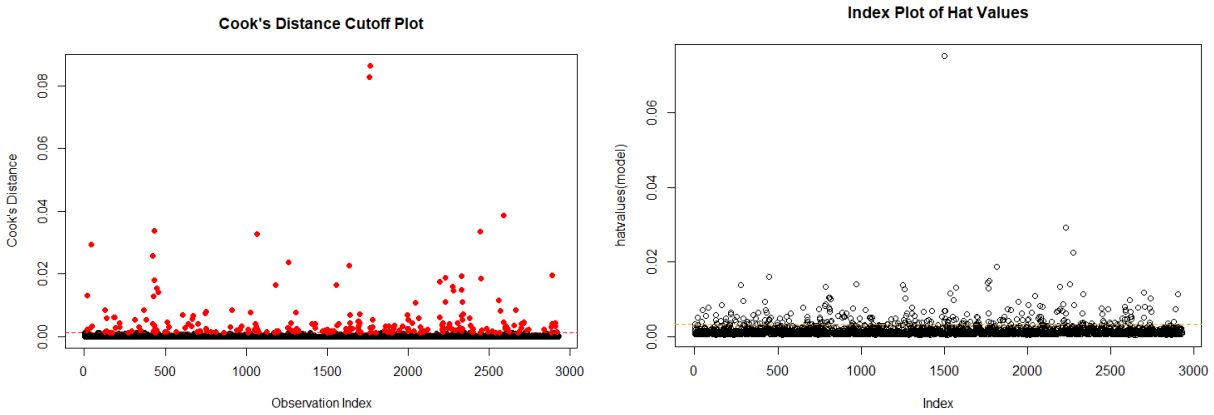


Figure 29 Cook's distance and Hat Values visualized

The Cook's Distance plot identifies influential observations in the model, with points above the red line representing the common threshold indicating potential influence. Only a few observations exceed this cutoff, which may warrant further investigation or exclusion to improve model integrity. The Hat Values plot displays leverage, with the orange line denoting twice the average leverage. Points above this line are considered to have high leverage, possibly exerting undue influence on the model's fit. Together with the outlier test results, these graphs suggest that while most data points do not overly influence the SalePrice model, a few observations with high studentized residuals, leverage, or Cook's distance could be influential and should be examined for validity or potential removal to enhance the model's robustness.

I would like to address two issues with respect to my model.

Firstly, the issue of outliers and secondly, the removal of a predictor variable.

After carefully isolating and visualizing the outliers in our model I concluded that rows with 'Order' numbers 45, 1768, 1064, 1761, 2593, 433, 434, 2446, 2333, 2335 were outliers. The decision to remove outliers ought never to be taken lightly. Therefore, I proceeded to inspect all the variables in these rows to mine all the information in this rich dataset with respect to these particular rows. I found to candidates that could be potentially removed without having a deleterious effect on the overall dataset and analysis. In the sale condition column of the house which was an outlier in row 1761 , I found that the salcondition was listed 'abnormal' this implies a foreclosure sale or similar abnormalitites. Given this condition I decided this outlier may be having undue influence and removed it.Also, the house in row 2593 was a good candidate for removal because its age was very high i.e it was oddly old (85 years) considering the mean age

in the dataset being 38.64 and the median age being 37. These two outliers were removed from the model.

Secondly, I removed the predictor variable 1st Floor Area in square feet from the model owing to its reduced p-value and therefore less independent statistical significance for the model.

```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area,
    data = amesclean2)

Residuals:
    Min      1Q  Median      3Q     Max
-172946  -20090     644   20743  247837

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -45474.956   2667.754  -17.05   <2e-16 ***
Gr.Liv.Area       74.258      1.811   41.01   <2e-16 ***
Total.Bsmt.SF     66.040      2.101   31.44   <2e-16 ***
Garage.Area       97.411      4.309   22.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40980 on 2921 degrees of freedom
Multiple R-squared:  0.733,     Adjusted R-squared:  0.7327
F-statistic:  2673 on 3 and 2921 DF,  p-value: < 2.2e-16
```

Figure 30 Summary of the updated regression model

The above summary of the updated model has the following equation-

SalePrice = -45,474.956 + 74.258 × (Gr.Liv.Area) + 66.040 × (Total.Bsmt.SF) + 97.411 × (Garage.Area)

"Did your changes improve the model, why or why not?"

Comparing the initial and updated regression models:

Initial Model:

- Adjusted R-squared: 0.7322
- p-value for X1st.Flr.SF: 0.937 (not significant)
- Included X1st.Flr.SF in the model

Updated Model:

- Adjusted R-squared: 0.7327
- X1st.Flr.SF removed from the model
- Two outliers removed based on substantive reasons

The changes resulted in a slight improvement in the model. The Adjusted R-squared increased from 0.7322 to 0.7327, which indicates that the model explains the variability of the response data slightly better than before. Removing the X1st.Flr.SF (1st Floor Area in Square Feet)due to its lack of significance has likely reduced noise, focusing the model on more significant predictors. Moreover, the exclusion of two particular outliers based on their unusual characteristics (abnormal sale condition and age) means that the model's estimates are less influenced by these atypical observations.

In summary, by removing non-significant predictors and outliers with justifiable reasons, the updated model's fit has been refined without losing generalizability, making the adjustments overall beneficial for the analysis.

R is a rich statistical programming language that has specific packages to deal with virtually every aspect of Data Analysis. I used two such packages, 'leaps' and 'car' to obtain and visualize the best model after performing a subset regression. I limited the number of predictor variables to three as I did not want to increase the number of parameters beyond the updated models parameters.

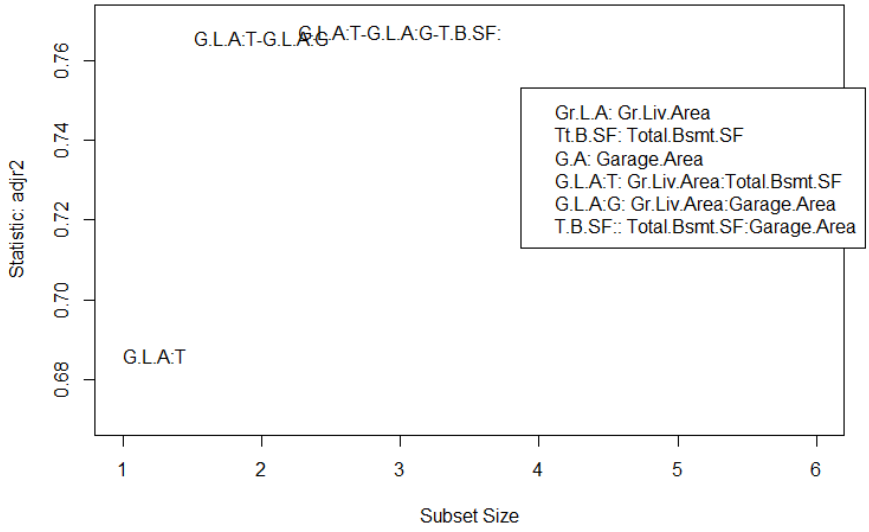


Figure 31 Subset Regression analysis visualized. Upto three predictors were included.

The best model with only one variable was the model with the interaction term between above ground living and total basement Area . This model was able to explain 68.59% of the variation in sale price. This is impressive for a single predictor and essentially gives the strength of house size, both above ground and below ground, in predicting sale price.

Similarly, the model with interaction terms of above ground living area and total basement area and above ground living area and total garage area was the best 2-variable model as it was explaining close to 76.59% of the variation.Incorporating garage size improves the model significantly.

However, the model with highest adjusted R-squared was the model with 3 variables which were the interaction terms of all the three variables in combination. This explained 76.73% of the variation.

```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area:Garage.Area + Total.Bsmt.SF:Garage.Area +
    Gr.Liv.Area:Total.Bsmt.SF, data = amesclean2)

Residuals:
    Min      1Q  Median      3Q     Max
-187938  -17440     743   19393  204188

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.949e+04  1.344e+03   51.69  < 2e-16 ***
Gr.Liv.Area:Garage.Area    6.028e-02  2.372e-03   25.42  < 2e-16 ***
Garage.Area:Total.Bsmt.SF  1.515e-02  3.524e-03    4.30 1.77e-05 ***
Gr.Liv.Area:Total.Bsmt.SF  3.489e-02  1.211e-03   28.82  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38240 on 2921 degrees of freedom
Multiple R-squared:  0.7675,    Adjusted R-squared:  0.7673
F-statistic:  3215 on 3 and 2921 DF,  p-value: < 2.2e-16
```

Figure 32 Refined model after subset regression

Model Interpretation:

- **Intercept**: The model estimates a SalePrice intercept of $69,940, the expected value of Sale when all predictors are zero.
- **Coefficients**:

  Each additional square foot of living area per square foot of garage area is associated with an incremental Sale Price increase of approximately 6.28%.

  Each additional square foot of garage area per square foot of basement area is associated with an incremental Sale Price increase of approximately 15.15%.

  Each additional square foot of living area per square foot of basement area is associated with an incremental Sale Price increase of approximately 34.89%.

Statistical Significance:

- All predictors, including the interaction terms, show statistical significance with p-values < 0.001.

Model Fit:

- The model's Adjusted R-squared is 0.7673, which means it explains 76.73% of the variability in Sale Price, indicating a strong fit.

"State the preferred model in equation form."

SalePrice = 69,940 + 0.0628 × (Above Ground Living Area : Garage Area) + 0.1515 × (Garage Area : Total Basement Area) + 0.3489 × (Above Ground Living Area: Total Basement Area)

Where ":" denotes the following according to the 'lm function documentation' in the offcial R Manual - "A specification of the form first:second indicates the set of terms obtained by taking the interactions of all terms in first with all terms in second." (R Core Team, n.d.)

I used common evaluation metrics. Adjusted R-Squared, AIC, BIC and the F-Statistic. The below table includes these comparisons.

Model	Adjusted_Rsquared	AIC	BIC	Fstatistic	Residual_SE
best_model	0.7673065	70033.64	70063.55	3214.963	38238.52
model_updated	0.7326983	70439.21	70469.11	2672.650	40983.58

Figure 33 Comparing the refined model after subset regression with the model before it using AIC, BIC, F-Statistic and AdjR2.

Upon comparing the two regression models I made the decision to prefer the model from step 13 owing to the following differences in its favor:

- Adjusted R-squared: The model from step 13 has a slightly higher Adjusted R-squared value (0.7673) compared to the model from step 12 (0.7327), indicating a marginally better fit to the data as it accounts for a higher proportion of the variability in the response variable.
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): Both AIC and BIC are lower for the model from step 13 (70033.64 and 70063.55, respectively) than for the model from step 12 (70439.21 and 70469.11, respectively). Lower AIC and BIC values suggest that the model from step 13 is preferred as it potentially offers a better trade-off between goodness-of-fit and complexity.
- F-statistic: The model from step 13 has a higher F-statistic (3214.963), which suggests a more statistically significant fit than the model from step 12 (2672.650).
- Residual Standard Error (RSE): The model from step 13 has a lower RSE (38238.52) in comparison to the model from step 12 (40983.58), indicating that the residuals are, on average, smaller for the model from step 13.

Based on these criteria, the model from step 13 is preferred because it has a better fit to the data without being overly complex, as evidenced by the higher Adjusted R-squared, lower AIC and BIC, higher F-statistic, and lower RSE. These metrics collectively indicate that the model from step 13 is likely to generalize better to new data compared to the model from step 12.

I therefore prefer the model obtained after subsets regression in step 13 to the model as it was in step 12.

Conclusion/Interpretations

This report has delineated an exhaustive analysis of the Ames Housing dataset within the scope of regression diagnostics in R, with a focus on its continuous variables due to their relevance to linear regression models. While the dataset is characterized by a predominance of categorical and discrete variables, the decision to exclude these from summary statistics calculations was based on assignment rubric that explicitly mentioned continuous variables alone are to be included in the regression model designed to predict Sale Price. (Habbous, S. 2024).

The analysis identified that a subset of properties was sold prior to completion, which was a notable discovery within the dataset's breadth. Additionally, it was found that 16% of the Lot Frontage Area data was missing, which required imputation for further analysis.

The Garage Area, Above Ground Living Area, and Total Basement Area and the 1st Floor Area emerged as variables positively correlated with Sale Price. An initial regression model ('model') that included these variables without addressing outliers was capable of explaining 73% of the variation in Sale Price.

Model diagnostics were conducted to ensure the model's integrity. The diagnostic process revealed that while no variables warranted removal due to low Variance Inflation Factors (VIFs less than 3), outliers were identified that violated the assumptions of normality and homoscedasticity and were exerting undue influence on the model's coefficients. Consequently, 2 of these outliers were removed to enhance the model's validity. Also, 1st Floor Area was removed from the model owing to its low contribution statistically.

A subsequent model ('model_updated') was constructed post-outlier removal, which demonstrated a slightly improved capability, explaining 73.27% of the variation in Sale Price. Comparative analysis of models built using subsets of three variables confirmed that the best-performing model was a model based on the interaction terms which utilized all three aforementioned variables. This model ('best_model') had an improved predictive power, capable of explaining 76.6% of the variation In the sale price.

This report's methodical approach to analyzing and improving the regression model has yielded a refined predictive tool for Sale Price within the Ames housing market context.

References

Bluman, A. G. (2023). Elementary Statistics 10th Edition (Instructor's Annotated Edition) (10th ed.). McGraw Hill.

Kabacoff, R., I. (2011). R in Action: Data Analysis and Graphics with R (1st ed.). Manning.

R Core Team. (n.d.). lm - Fitting Linear Models. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

Regression Plots — statsmodels. (n.d.). https://www.statsmodels.org/dev/examples/notebooks/generated/regression_plots.html