BUSINFO 704: PREDICTIVE ANALYTICS

GROUP ASSIGNMENT 2: REPORT

DUE DATE: FRIDAY, 25TH NOVEMBER @ 1:00PM

GROUP NUMBER: 5

GROUP MEMBERS:

SHAKEEL ALI
LAUREN LIANG
JOSHUA MAO
SERENA ZHANG
GEORGE ZHU

Executive Summary:

In terms of predicting house sale prices, we find that our model is apt for predicting sale prices of **under-valued** houses, that is, less than equal to $290,000. The prediction error which comes from predicting the sale price of these houses would, on average, be around $22,000. On the other hand, even though our model has predictive capabilities on **over-valued** houses, that is, more than $290,000, this comes with a higher prediction error of around $58,000 on average. Moreover, we find that in order to maximize sale price, a remodeling company should mainly focus on targeting an overall quality rating of 10 **(OverallQual10)**, maximizing the ground living area **(GrLivArea)**, basement area **(TotalBsmtSF)** and the number of **Fireplaces** as this increases the sale price by roughly $141,100, $45.00 (per square feet), $13.80 (per square feet) and $9815.00 (per fireplace) respectively**.** In addition, we have also found out that the more recently the year built **(YearBuilt)** of a house is, the greater the monetarily influence it has on sale price. For example, given two houses, one which was built in 2021, and the other built in 2022, we would expect to see the house built in 2022 being sold for $330.00 more than the other, holding other determinants constant. Likewise, the more recent the remodeling date **(YearRemodAdd)**, the higher the monetarily influence on sale price, with each subsequent year bringing an additional $345.00 increase in sale price, while holding other determinants constant. Furthermore, we advise our consultee to remodel houses with a zoning classification of "RM", and in the neighborhood of "StoneBr" in comparison to "C(all)" and "MeadowV" respectively, as this has the highest monetarily impact on sale price of roughly $25,700 and $60,600 respectively. Lastly, it is ill-advised for our client to remodel houses which has a dwelling type **(BldgType)** of BldgTypeTwnhs or BldgTypeTwnhsE, as these greatly reduces the selling price by $40,600 and $29,450 in comparison to BldgType2fmCon.

Introduction:

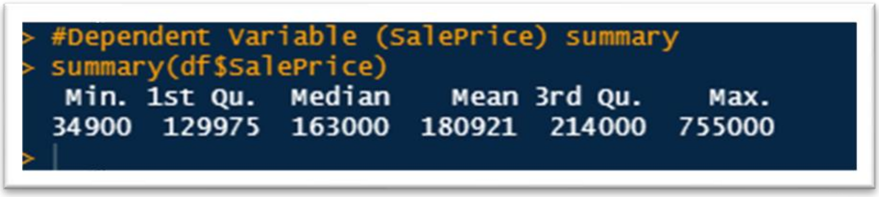
Due to the unpredictable housing market, a house remodeling company sought our help in order to maximize their profitability in the market. Essentially, the consultees' business objective is to understand the critical determinants of house sale price. Hence, the data-mining objectives of this project is to firstly identify statistically significant determinants which influence house sale price, and secondly to build a prediction model which is able to predict house sale price.

Furthermore, it is imperative for any project to set its data-mining success criteria beforehand in order to evaluate at the later stage. The reason for this is to nullify the possibility of any *hindsight biasness* which might cause us to "*bend the truth"* in terms of our interpretation, and conclusion only after we have obtained the results. In this project, a Linear Regression model will be deployed in order to achieve the pre-stated objectives. An independent variable will be considered statistically significant if its associated *probability value (p-value)* is less than equal to 0.05. Moreover, the overall quality of the model will be determined by analyzing its R-Squared, which is a statistical measure that represents the proportion of the variance for sale price, that's explained by the independent variables in our model. This means that the higher the R-Squared value, the better, and hence we would consider a minimum r-squared value of 0.70 as an acceptable benchmark. Lastly, in terms of prediction, we would measure this according to the residuals. The residuals are the difference between the actual sale price and the predicted sale price. If we get a root-mean-square-error of less than $35,000, we would consider our model's prediction ability to be acceptable.

All in all, the purpose of this project is to predict our dependent variable (SalePrice), based on the several independent variables. Moreover, we will also determine if the independent variables are statistically significant or not.

Descriptive Analytics:

The dataset, which comes as a comma-separated values (csv) file contains information which pertains to houses sold. This means that each row within our dataset represents a house sold, while each column represents a specific characteristic of that particular house. Upon inspection, we notice that there are 1460 rows of data within our dataset, and 81 variables (Appendix 1). Of these 81 variables, 38 are of integer (numeric) type whilst the remaining 43 are of character (string) type (Appendix 2). Moving on, the figure below shows the summary statistics of our dependent variable; SalePrice.



```
> #Dependent variable (SalePrice) summary
> summary(df$SalePrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34900  129975  163000  180921  214000  755000 
> 
```

Figure 1.0 showing the summary statistics of SalePrice.

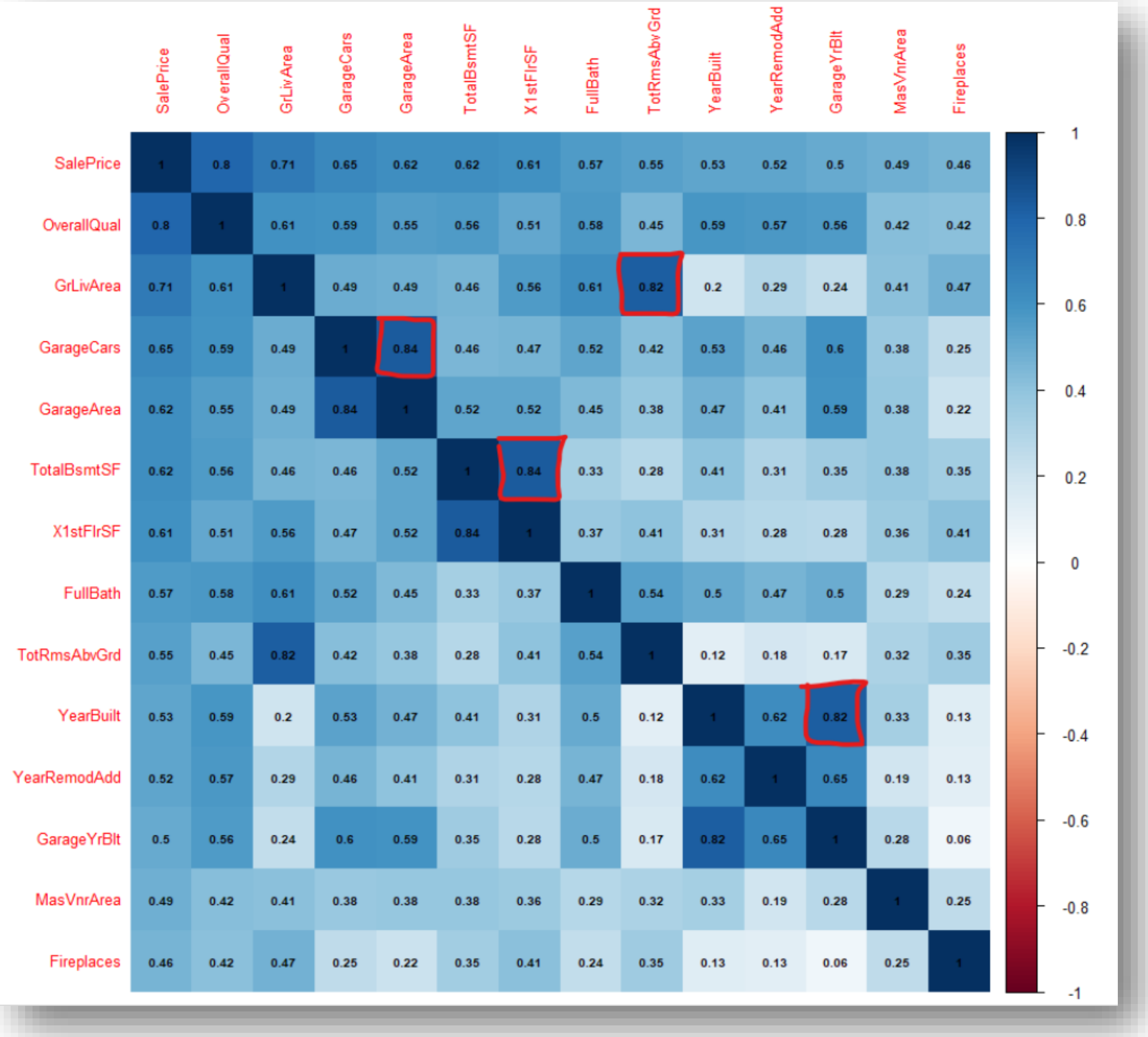
Upon looking at the above figure, we can see that the difference between the minimum sale price value of $34,900 and the median value is $128,100 whilst the difference between the median value and the maximum value of $755,000 is $592,000. This explains why the mean is "stretched" outwards compared to the median. This does indicate that perhaps we are looking at outlier(s) within our data-set in conjunction with a right-skewed distribution, something which will become apparent in data visualization part of this project.

Independent Variables Selection:

In building a Linear Regression model, the selection of important independent variables is necessary. An independent variable is deemed important if it'll have a high influence on the dependent variable, which in our case is SalePrice. We have decided to select independent variable using three methods; **Correlation Matrix**, **Subset Regression**, and **Logic**. The correlation matrix and subset regression solely work on quantitative (numeric) variables, hence the need to use logic to determine significant qualitative (categorical) determinants of SalePrice.

Correlation Matrix

The correlation-matrix helps us to identify correlation through visualization of correlation coefficients between variables. This helps us to identify independent variables which have a high correlation with the dependent variable (SalePrice), as well as identify variables which have a high correlation amongst each other. Those variables which have a high correlation with each other are said to have multi-collinearity, that is, the variables are *not* independent of each other. Variables would be considered significant if its correlation coefficient with SalePrice is above 0.40 or -0.40, while at the same time, variables will be deemed to have multi-collinearity if they have a correlation coefficient above 0.80 or -0.80.



	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	X1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt	YearRemodAdd	GarageYrBlt	MasVnrArea	Fireplaces
SalePrice	1	0.8	0.71	0.65	0.62	0.62	0.61	0.57	0.55	0.53	0.52	0.5	0.49	0.46
OverallQual	0.8	1	0.61	0.59	0.55	0.56	0.51	0.58	0.45	0.59	0.57	0.56	0.42	0.42
GrLivArea	0.71	0.61	1	0.49	0.49	0.46	0.56	0.61	0.82	0.2	0.29	0.24	0.41	0.47
GarageCars	0.65	0.59	0.49	1	0.84	0.46	0.47	0.52	0.42	0.53	0.46	0.6	0.38	0.25
GarageArea	0.62	0.55	0.49	0.84	1	0.52	0.52	0.45	0.38	0.47	0.41	0.59	0.38	0.22
TotalBsmtSF	0.62	0.56	0.46	0.46	0.52	1	0.84	0.33	0.28	0.41	0.31	0.35	0.38	0.35
X1stFlrSF	0.61	0.51	0.56	0.47	0.52	0.84	1	0.37	0.41	0.31	0.28	0.28	0.36	0.41
FullBath	0.57	0.58	0.61	0.52	0.45	0.33	0.37	1	0.54	0.5	0.47	0.5	0.29	0.24
TotRmsAbvGrd	0.55	0.45	0.82	0.42	0.38	0.28	0.41	0.54	1	0.12	0.18	0.17	0.32	0.35
YearBuilt	0.53	0.59	0.2	0.53	0.47	0.41	0.31	0.5	0.12	1	0.62	0.82	0.33	0.13
YearRemodAdd	0.52	0.57	0.29	0.46	0.41	0.31	0.28	0.47	0.18	0.62	1	0.65	0.19	0.13
GarageYrBlt	0.5	0.56	0.24	0.6	0.59	0.35	0.28	0.5	0.17	0.82	0.65	1	0.28	0.06
MasVnrArea	0.49	0.42	0.41	0.38	0.38	0.38	0.36	0.29	0.32	0.33	0.19	0.28	1	0.25
Fireplaces	0.46	0.42	0.47	0.25	0.22	0.35	0.41	0.24	0.35	0.13	0.13	0.06	0.25	1

Figure 2.0 showing the Correlation Matrix.

From the above correlation-matrix, we can see the existence of multi-collinearity between variables as indicated by the red-box. This means, that we only need to select one of these multi-correlated variables, and as such, we'll select the variable which has the highest correlation with SalePrice.

Subset Regression

The subset regression helps us to obtain the best possible adjusted r-squared value by running a simulation based on the variables provided. However, it cannot be assumed that the highest adjusted r-squared value is the best model. This is because there might be multi-collineated variables present which explains a high adjusted r-squared. Hence, it is necessary to consider whether there are multi-correlated variables selected by the subset regression. In the figure below, as outlined by the red boxes, are the independent variables which would give the best adjusted r-squared value *after* considering variables which had multicollinearity amongst each other.

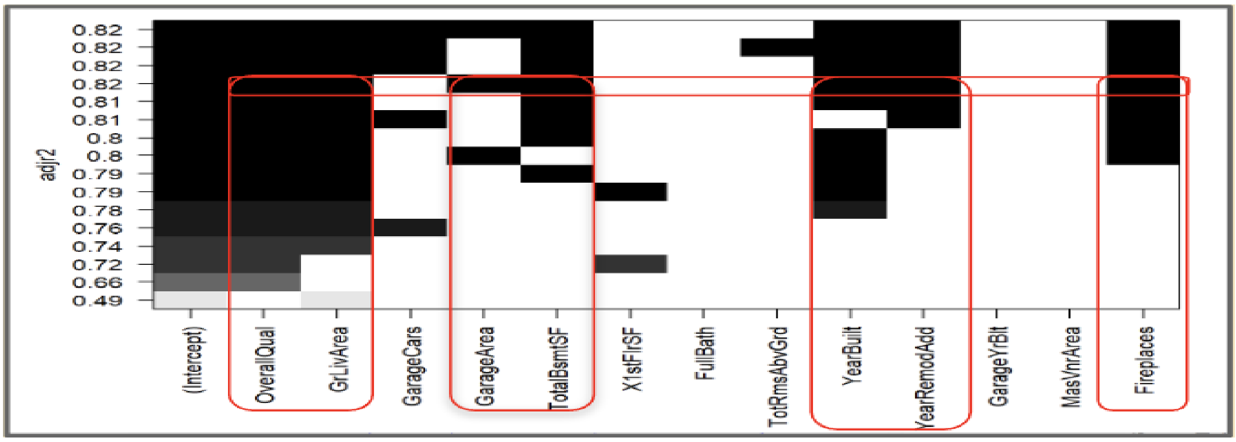


Figure 3.0 showing the Subset Regression Output.

All in all, after the consolidation of the collinearity matrix and the subset regression, we have chosen these **numerical** independent variables to add to our linear regression model:

- OverallQual – Overall material and finish quality of a house.
- GrLivArea – Above grade (ground) living area square feet.
- GarageArea – Size of garage in square feet.
- TotalBsmtSF - Total square feet of basement area.
- YearBuilt – Original construction date.
- YearRemodAdd – Remodel date.
- Fireplaces – Number of fireplaces.

Logic

Since the other two methods works only with numerical variables, we need to logically choose which qualitative variables would improve our model. We selected the following:

- Neighborhood – Physical locations within Ames city limits.
- BldgType – Type of dwelling.
- MSZoning - The general zoning classification.

We know that buyers are willing to pay lucrative prices in order to purchase a house in a neighborhood with close amenities. Moreover, the dwelling type of a house can significantly influence the threat of trespassing, burglary, etc. People are not willing to pay a higher price for building type which is susceptible to these. Lastly, the type of zoning classification determines what an owner can put on or use its land for.

In short, we will include the **seven** numeric variables and **three** qualitative variables in our model. Hence, our alternative hypothesis (H_1) is that these variables **does** influence SalePrice, while our null hypothesis (H_0) is that these variables **does not** influence SalePrice. An independent variable will be deemed statistically significant, if we are able to reject the null hypothesis in favor of our alternative hypothesis by looking at its associated p-value in the regression summary. We will only consider a variable to be statistically significant if its p-value is below the prior-set benchmark of 0.05.

Data Visualization:

Here, we will visualize our dependent and independent variables.

#1a: Visualization of Sale Price (Dependent Variable):

In terms of SalePrice, the red dots in the box-plot below indicates the 56 outliers above $350,000.

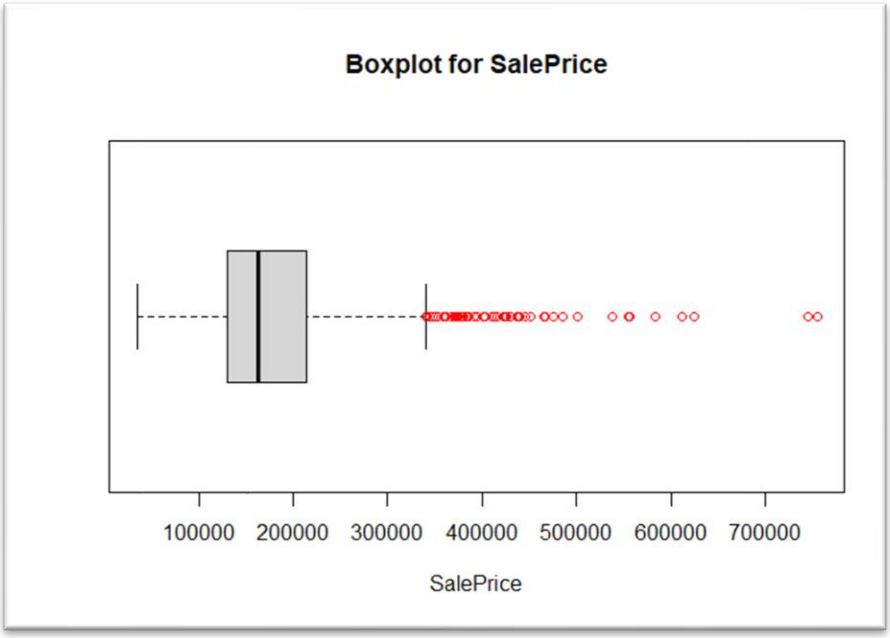


Figure 4.1 Box-plot of Sale Price.

Let's see if we still have outliers if we remove houses which have a sale price of at least $350,000 (**NOTE:** The new "outliers" were still present at the $350,000, $330,000 and $300,000 marks but gets removed at $290,000 as shown below.)

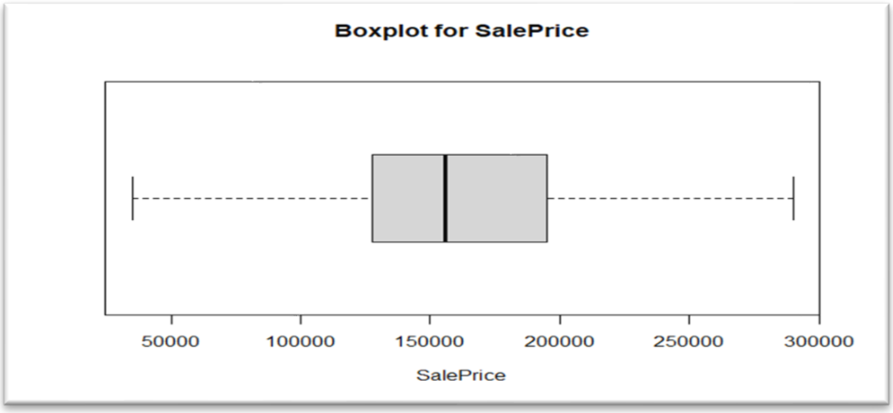


Figure 4.1 Box-plot of Sale Price without outliers (above $290,000)

#1b: Visualization of Sale Price (Dependent Variable):

From the visualization below, we can see that the *outliers* (above $290,000) skew the distribution to the right.

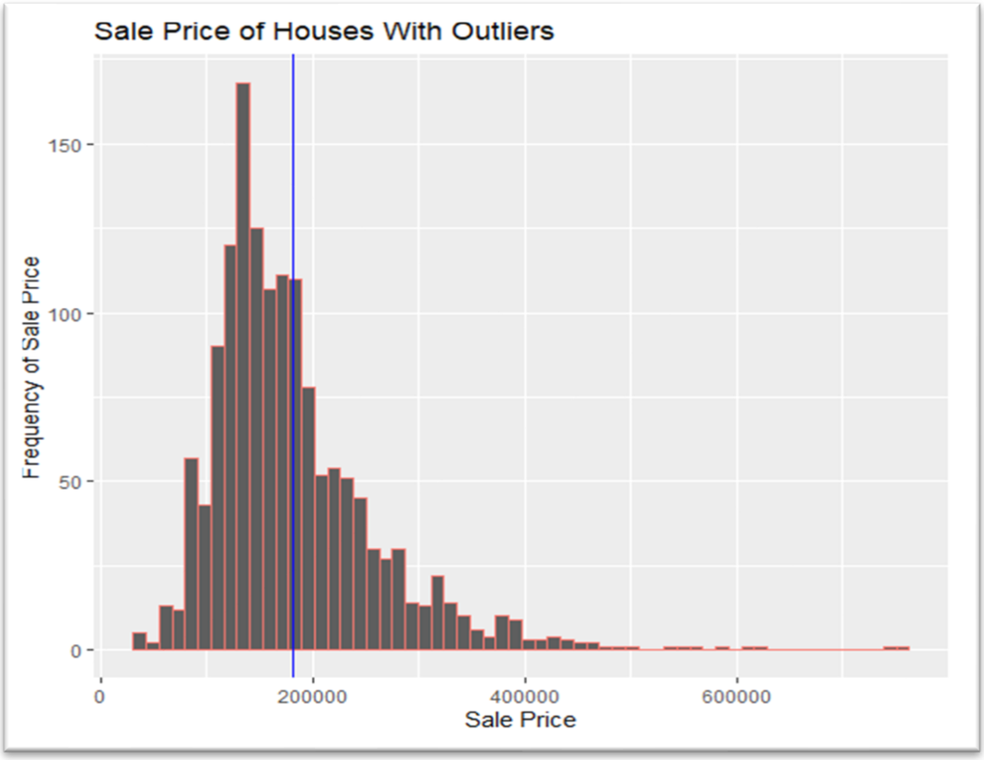


Figure 5.1 Box-plot of Sale Price with outliers (above $290,00)

In the majority of the situations, you would do some sort of transformation on the dependent variable in order to get a more normal distribution. This is **not** one of those situations. The reason for this is because those *outliers* which were detected in the first box-plot are what we deem as *"true outliers"*, i.e., those are just houses which were sold at a relatively higher price than the majority of the houses within our dataset rather than arising due erroneous data. The figure below shows normal distribution without the outliers.

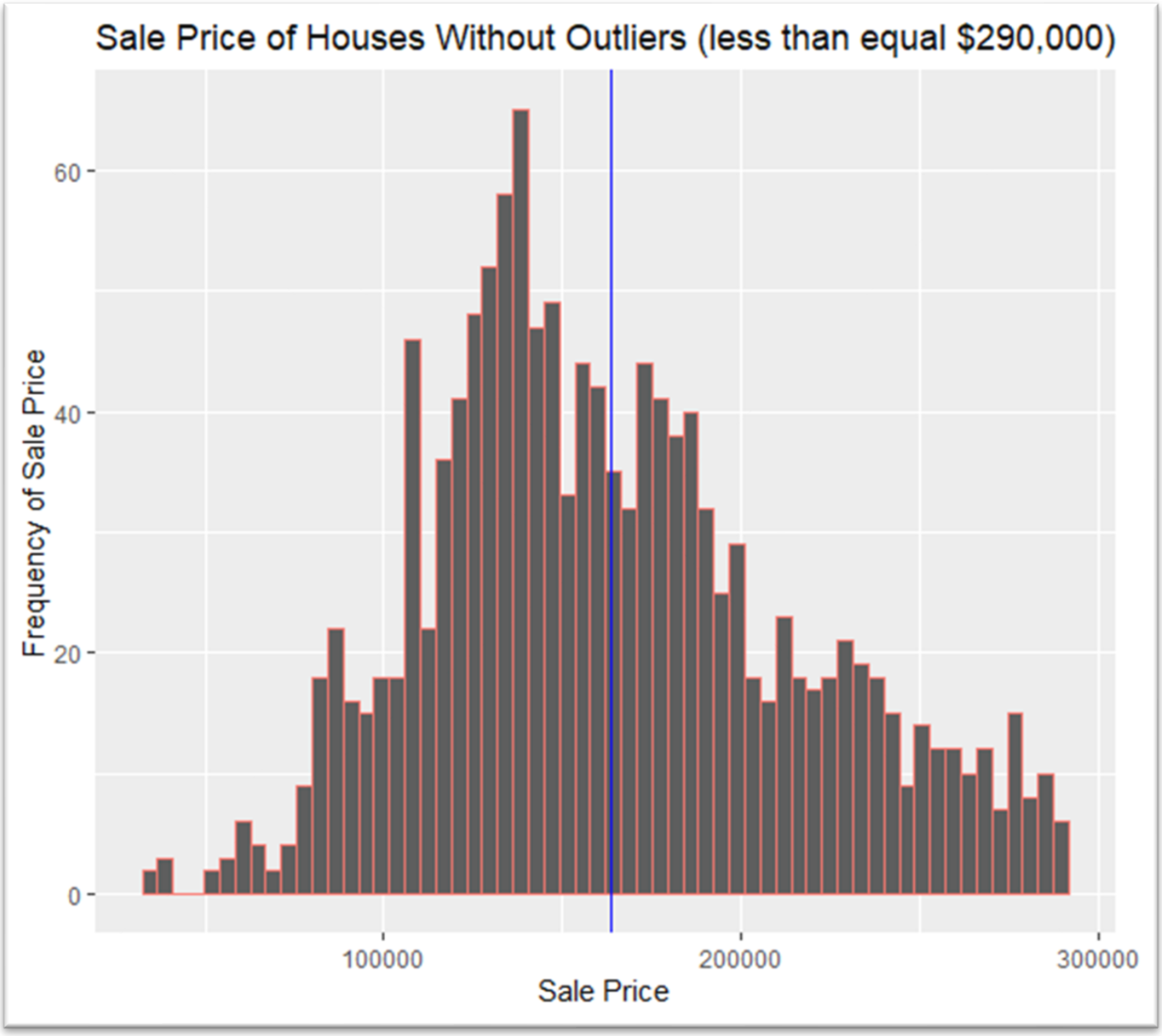


Figure 5.2 Showing normal distribution of Sale Price without outliers (<= $290,00).

However, if we were to do a log transformation, in order to get a normal distribution, you'd notice that even though we'd obtain a normal distribution, we would end up with *outliers* from the left **and** the right side of the box-plot. In comparison, *without* the log transformation, we'd **only** get *outliers* from the right side of the box plot as indicated in *Figure 4.1* above.

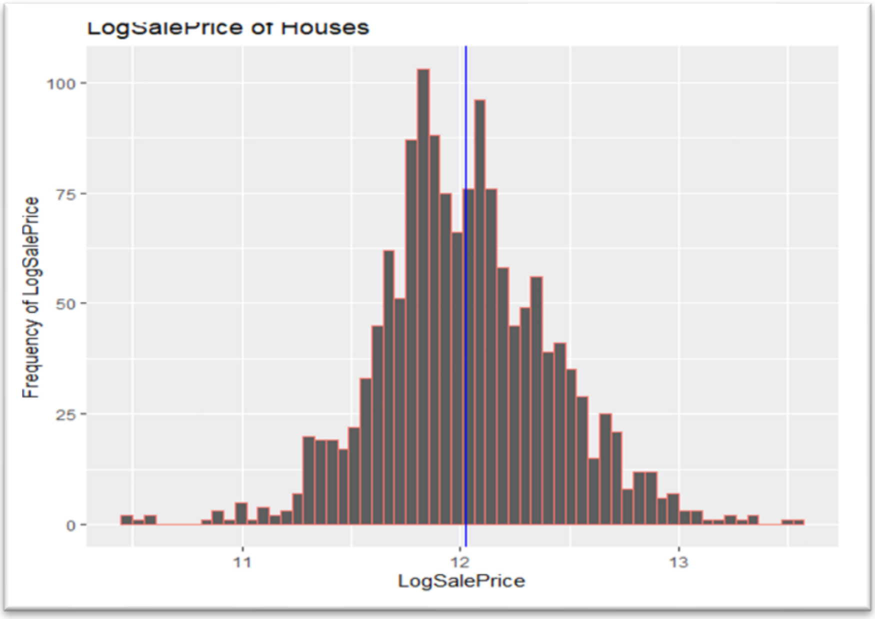


Figure 5.3 Showing normal distribution of LogSalePrice.

However, as stated, we end up with outliers at both sides of the box-plot.

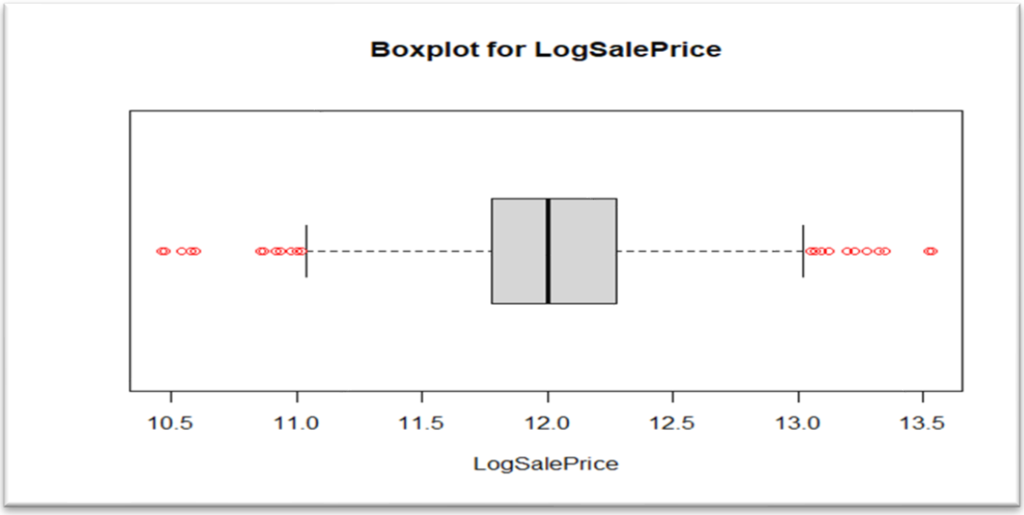


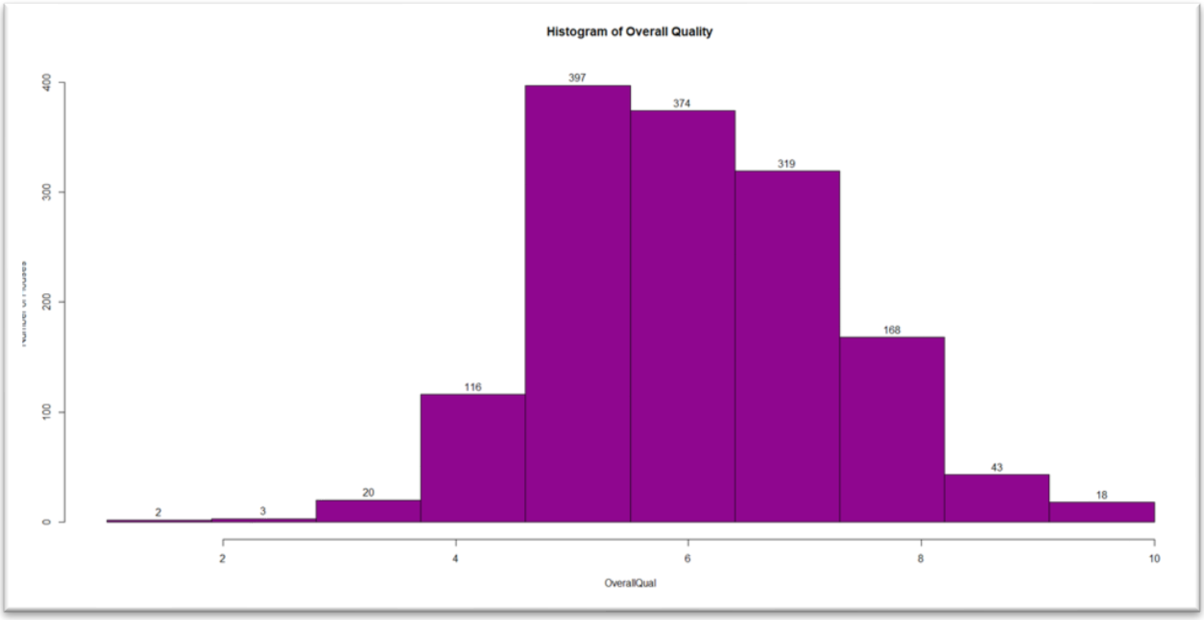
Figure 5.4 Box-plot of LogSalePrice.

Hence, we must first need to understand the implication of the following scenarios on our model:

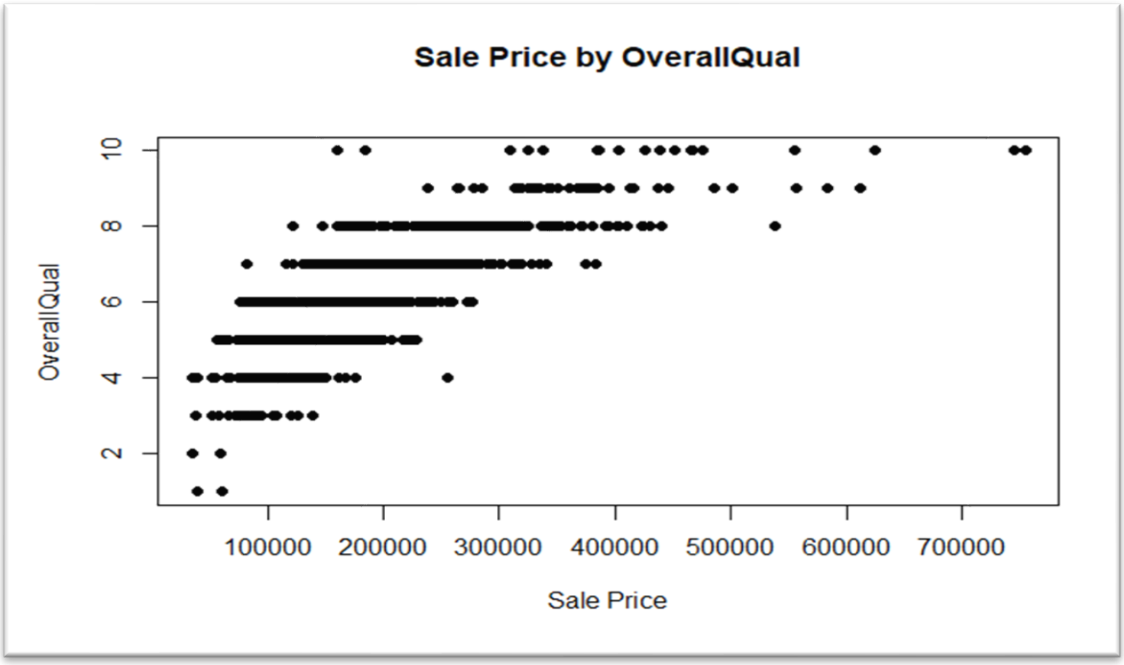
- What happens if we select SalePrice as our dependent variable and remove the outliers of SalePrice (above $290,000)?
    - Our model will be able to predict house sale price far more accurately in terms of having low residuals (errors) for houses with selling price lower than $290,000, however, the trade-off is that the model's prediction accuracy would be greatly reduced if it were to predict houses above $290,000.
- *What happens if we select LogSalePrice as our dependent variable and remove the outliers of LogSalePrice (below 11.0 and above 13.0)?*
    - Same as above, the prediction accuracy would increase for houses with non-outlier *LogSalePrice.* However, the prediction accuracy would greatly decrease for houses with outlier *LogSalePrice.* An important thing to note here is that unlike the outliers of *SalePrice*, which is only in one direction, the outliers for *LogSalePrice is* in both the direction (*Refer Figure 5.4)*!
- *What happens if we keep the outliers in our model?*
    - The prediction accuracy would decrease for **non-outlier** houses as the residuals (errors) would get stretched towards the side of which the outliers lie. However, given the relatively small number of outliers, we do not believe that the residuals would be considerably large.
    - However, the benefit is that, the prediction accuracy for ***outlier*** houses, that is houses with sale price bigger than $290,000 would increase. Although, the prediction accuracy would still be quite large in terms of the residuals, this would nevertheless be lower in comparison to if we built the model without the outliers. Hence, the benefit of including the outliers is that we will be able to get some form of predictive ability on outlier houses, which would make our model, relatively speaking, more robust.

We decided that since we wanted to achieve robustness in terms of our model being able to have some predictive ability on outliers, we will keep the outliers. Moreover, since we solely want these outliers to be on one-side, we will choose SalePrice as the dependent variable rather than LogSalePrice. However, we will still provide visualization and discussion on LogSalePrice's residuals in order to gauge if our decision was correct or not.

#2: Visualization of OverallQual (Independent Variable):

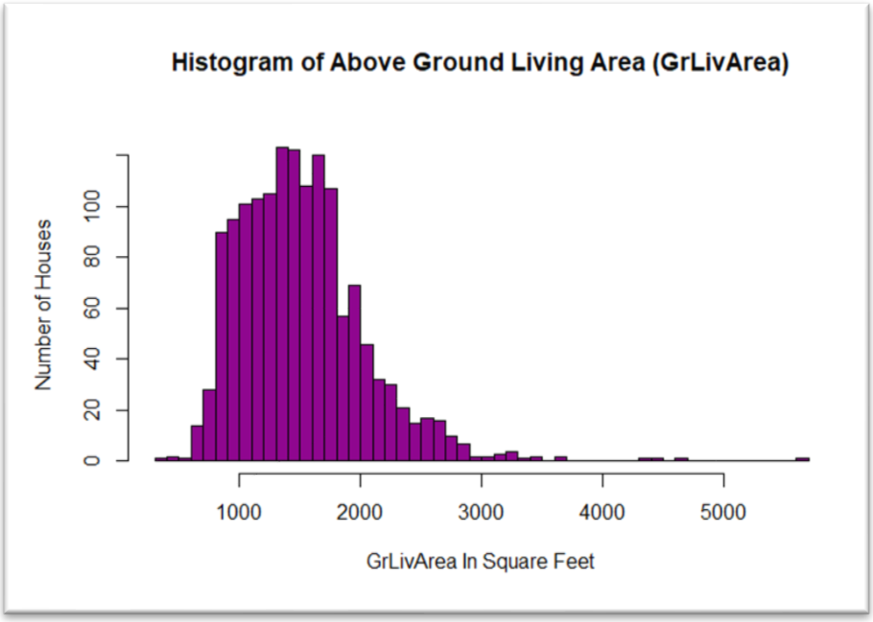


The above histogram shows us that majority of the houses in our dataset have an OverallQual of 5, while the total number of houses which have an OverallQual1 and OverallQual3 are 2 and 3 respectively.

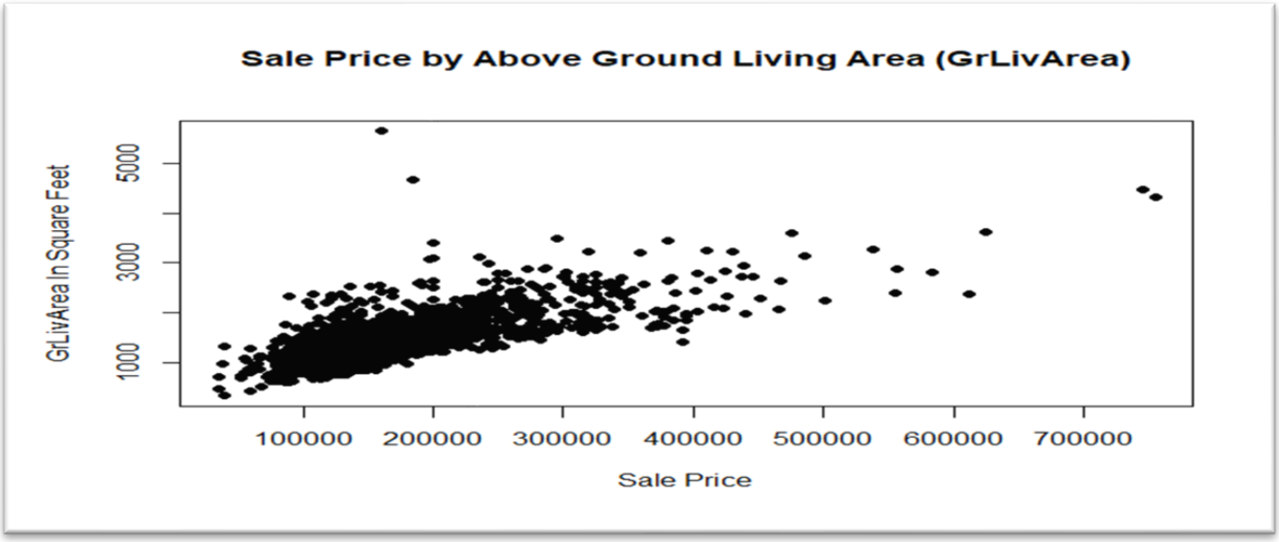


The above plot shows us that the variation in terms of Sale Price does start showing for houses as OverallQual increases, for example, OverallQual10 houses have been sold for as low as $150,000 and as higher as ~$730,000. Surprisingly, we also see that the highest SalePrice of OverallQual1 house is relatively higher than the highest SalePrice of OverallQual2.

#3: Visualization of GrLivArea (Independent Variable):

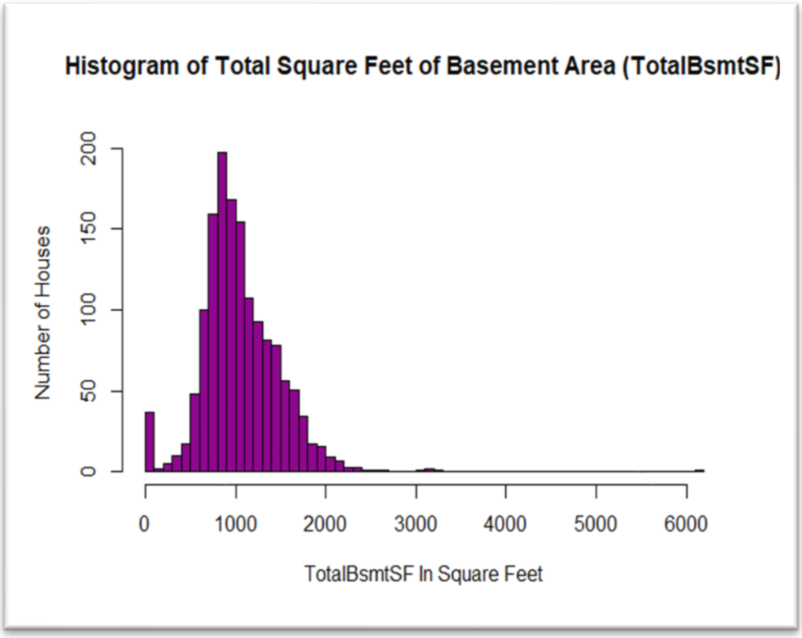


It looks like the majority of the houses' GrLivArea seems to be between 700sqft to 3000sqft. Only a miniscule number of houses have GrLivArea outside of this.

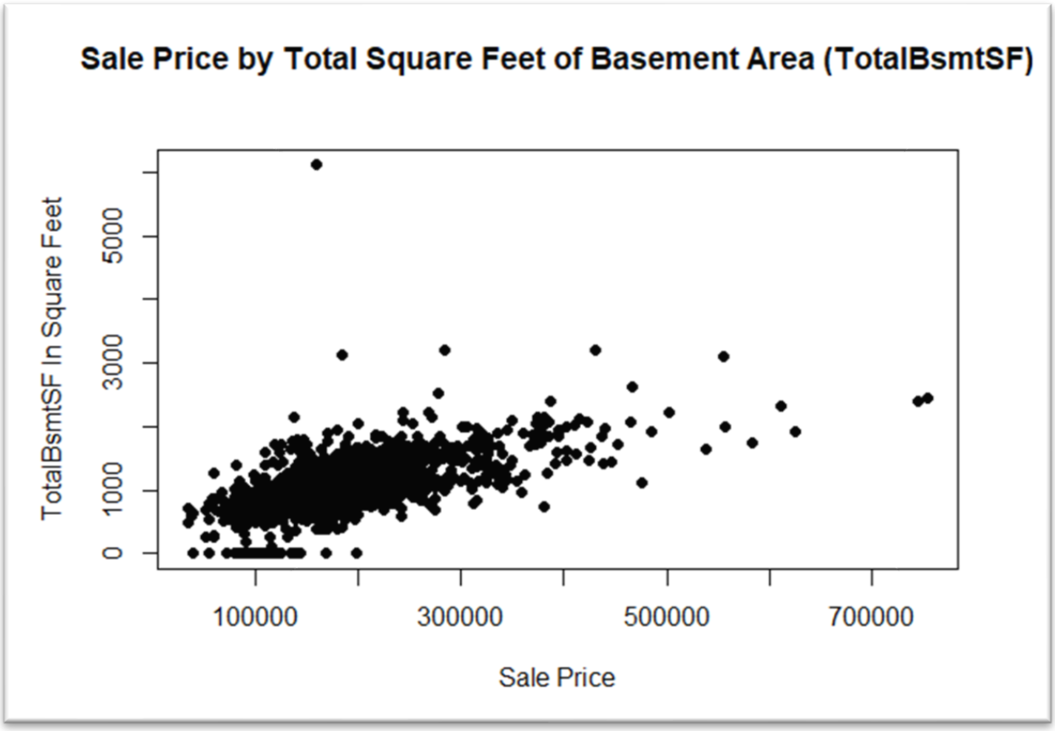


There does seem a linear relationship between GrLivArea and Sale Price because as the GrLivArea increases, so does the SalePrice. Surpvrisingly, the house with the highest GrLivArea was sold at a relatively low price of roughly $180,000.

#4: Visualization of TotalBsmtSF (Independent Variable):

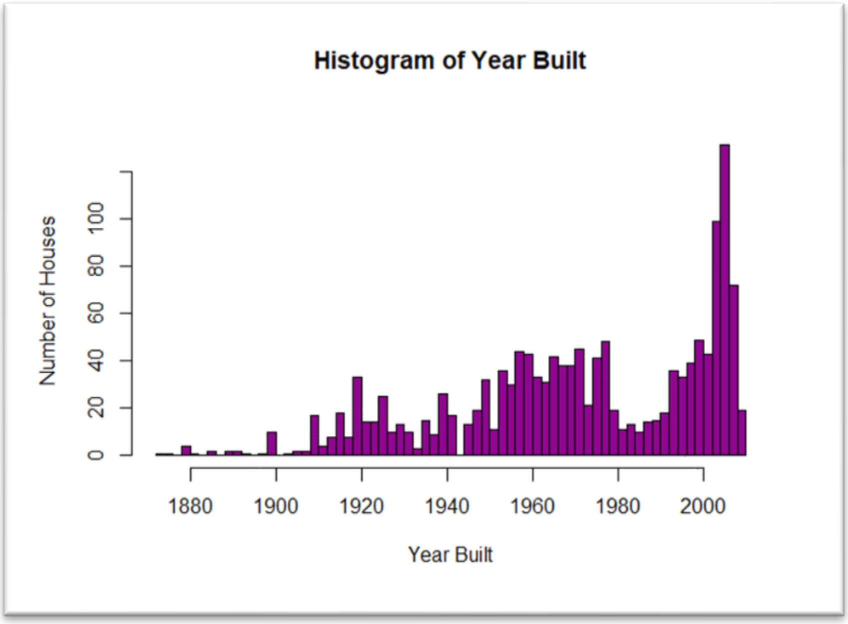


Majority of the houses have a TotalBsmtSF between ~100sqft to 2100sqft. Only a miniscule number of houses have a TotalBsmtSF above 2100sqft.

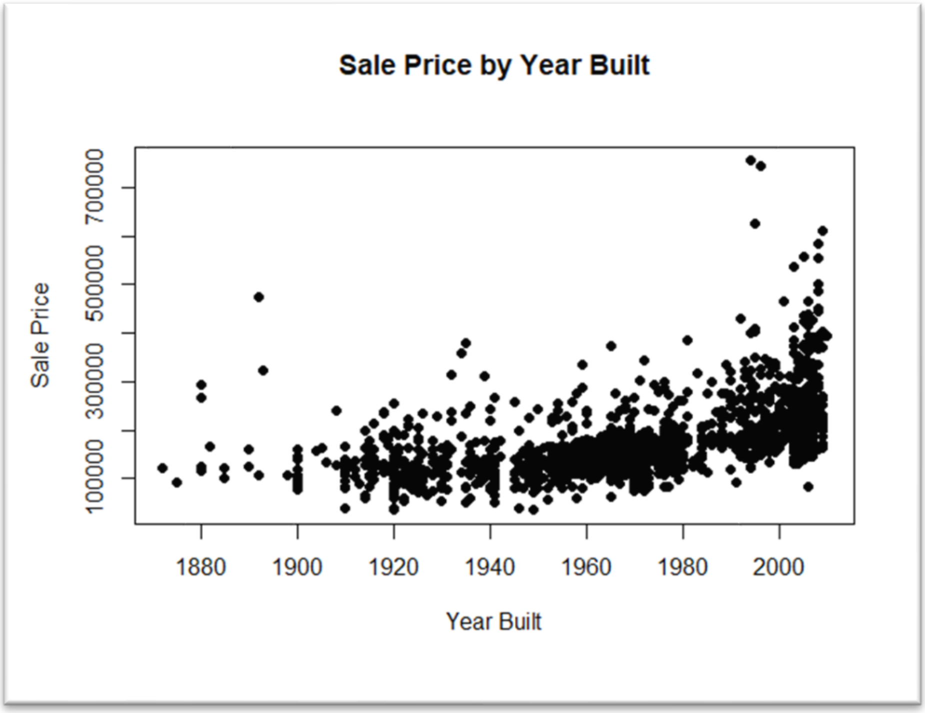


We can see that as the square feet of TotalBsmtSF increases, so does the Sale Price. However, there does seem to be an outlier to this, as the house with the largest TotalBsmtSF was sold for a relatively lower price of roughly $130,000.

#5: Visualization of YearBuilt (Independent Variable):

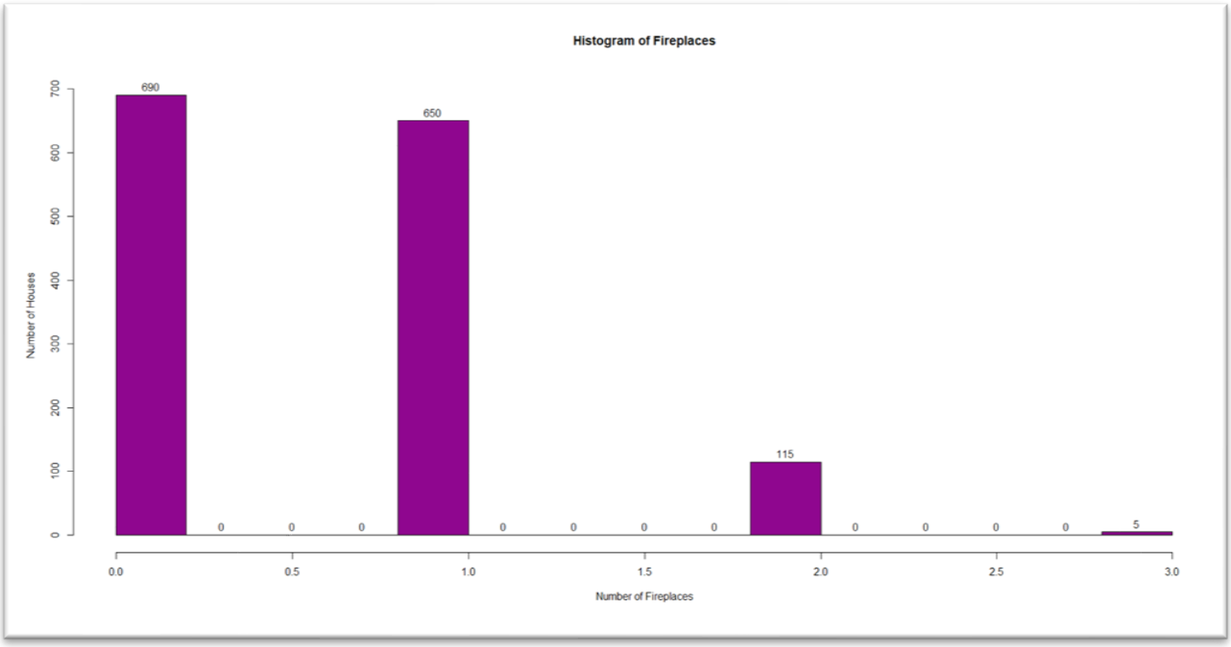


We can see that majority of the houses were built after 2000, with the highest number of houses being built around 2000-2006.

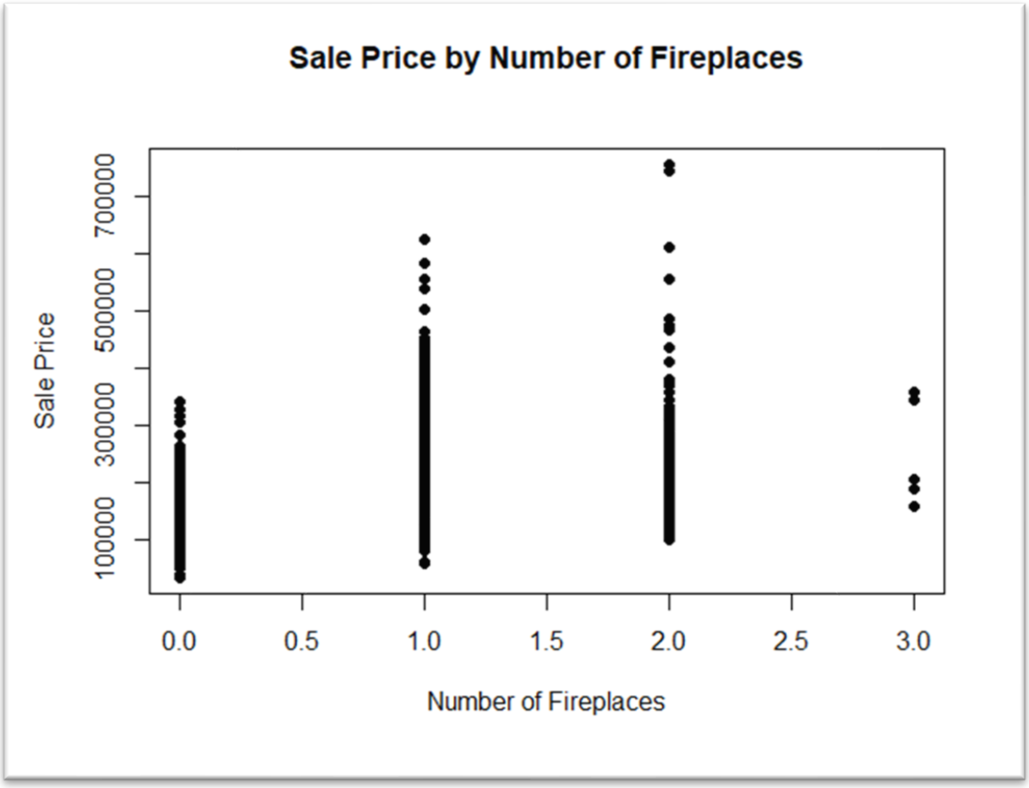


We can see that only after the year 2000, with each subsequent year, the sale price of houses increases. However, the sale prices of houses built in between 1900 to 1990 is stagnant. We also notice that older houses, that is prior to 1900s, does show a tendency to be sold at a higher price.

#6: Visualization of Fireplaces (Independent Variable):

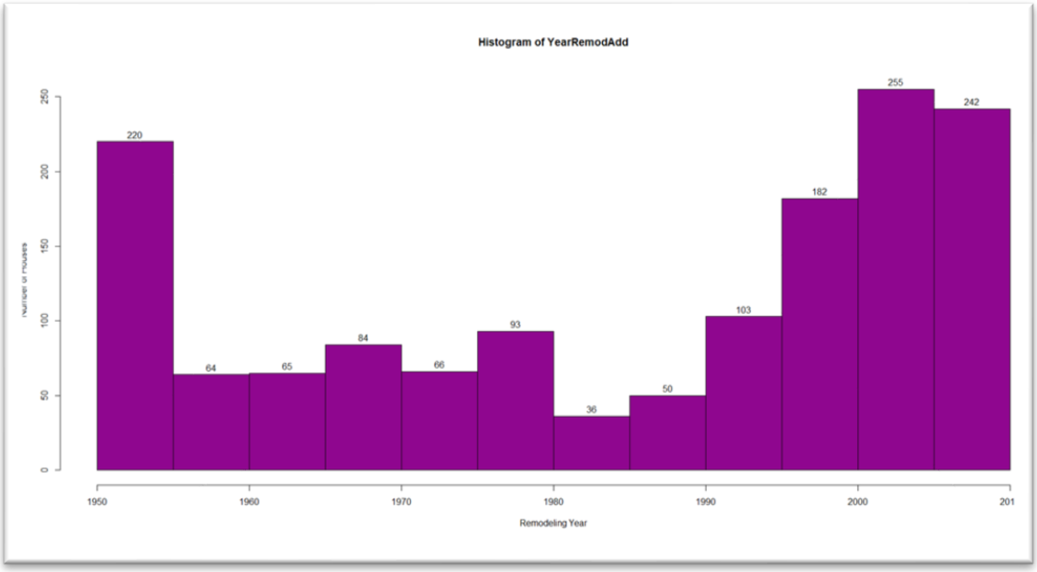


From the above histogram, we can see that majority of the houses within our dataset has either 0 fireplaces or at most only 1. It is very rare to see a house with 3 fireplaces.

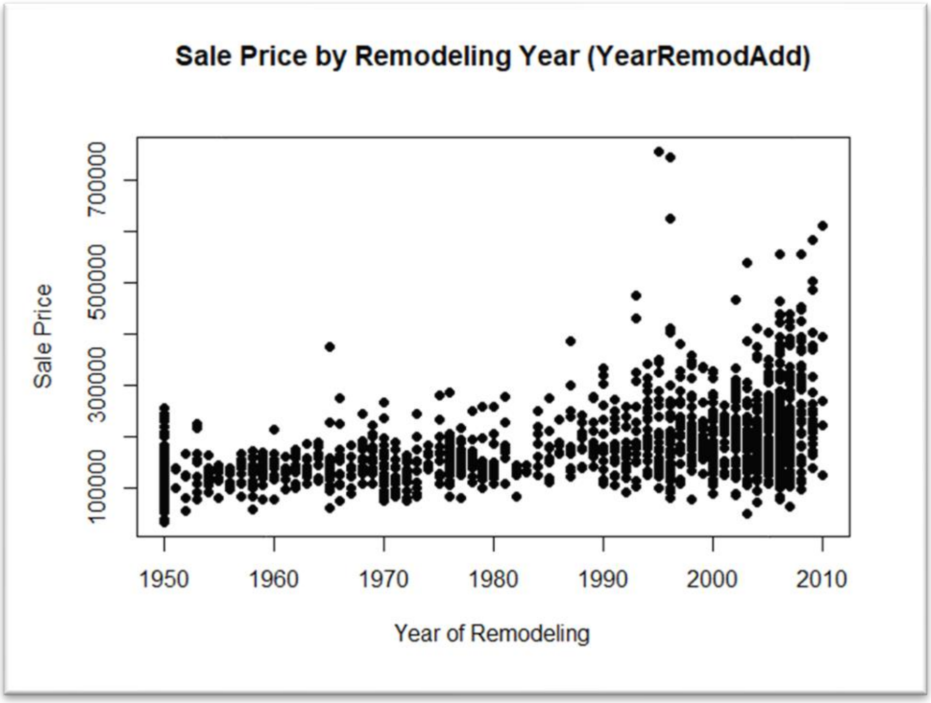


As we can see, that it's not necessary that the more fireplaces a house has, the higher the selling price. Moreover, it seems that the houses which have either 1 or 2 fireplaces, have a greater possibility of being sold for a relatively higher price.

#7: Visualization of YearRemodAdd (Independent Variable):

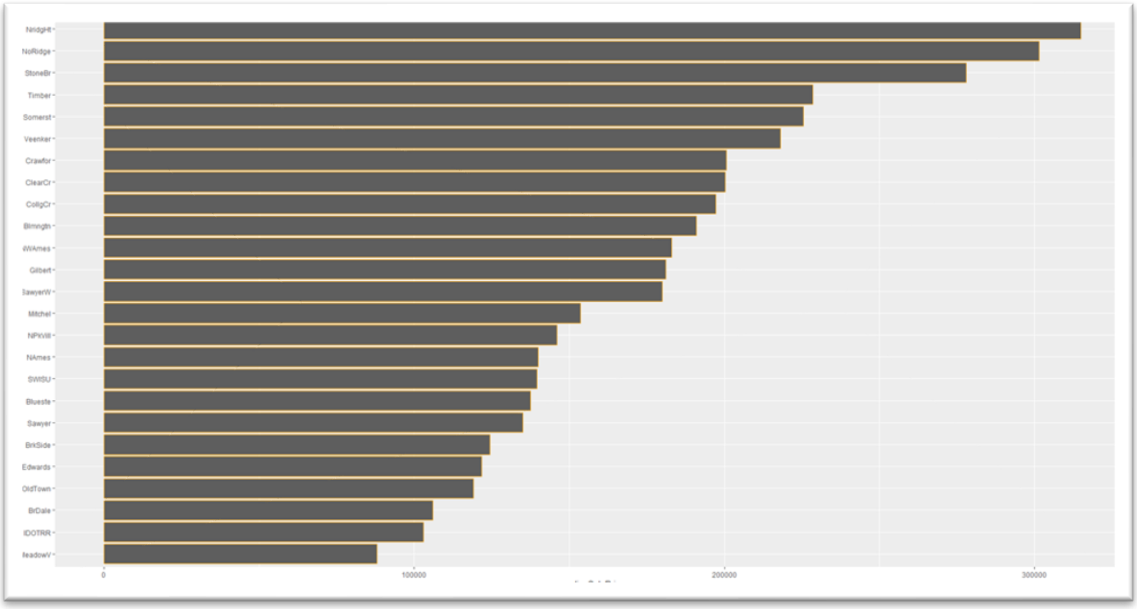


Seems like majority of the houses were remodeled in 1950s, 2000s and the 2010s. The number of houses which were remodeled between these dates were relatively similar.



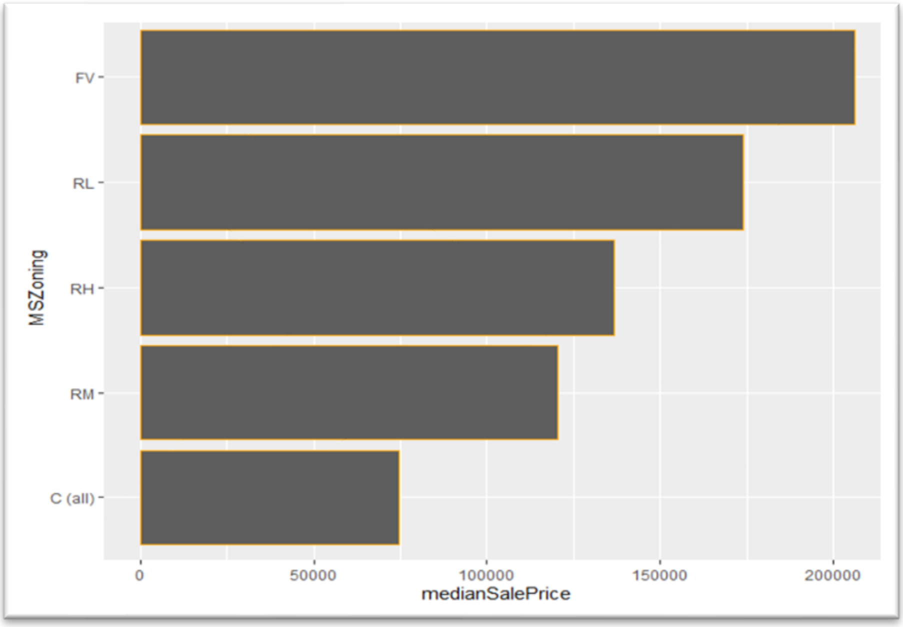
It does seem that if the remodeling was done in the 2000s and 2010s, it does generally result in a higher selling price with a few exceptions between 1990s and 2000s.

#8: Visualization of Neighborhood (Independent Variable):

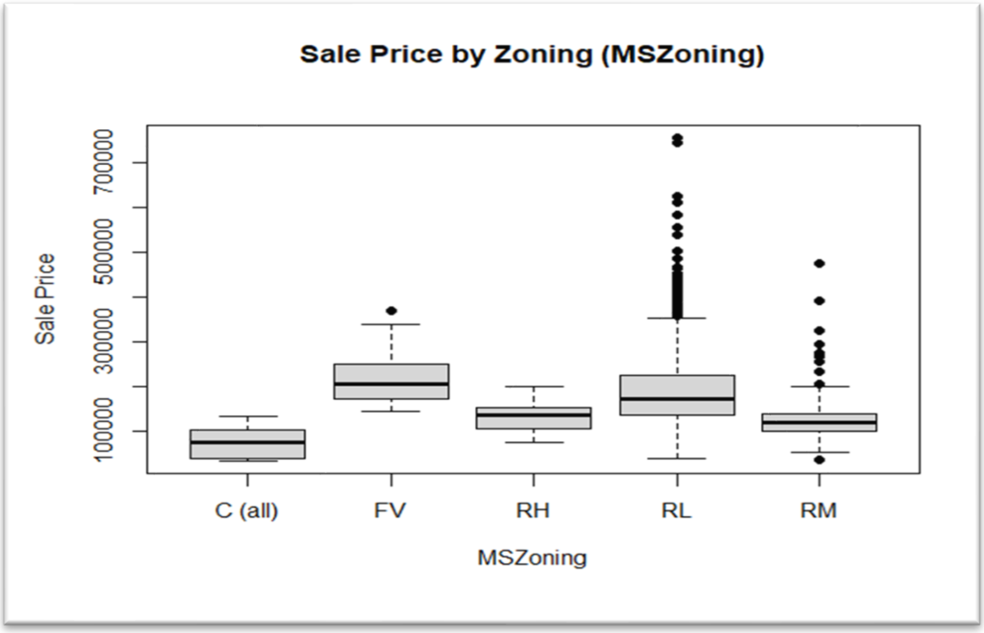


As evidently shown, the median selling price of houses in NridgeHt neighborhood is the highest at around $315,000 while the lowest is MeadowV. It would be interesting to see if this is statistically significant or not.

#9: Visualization of MSZoning (Independent Variable):

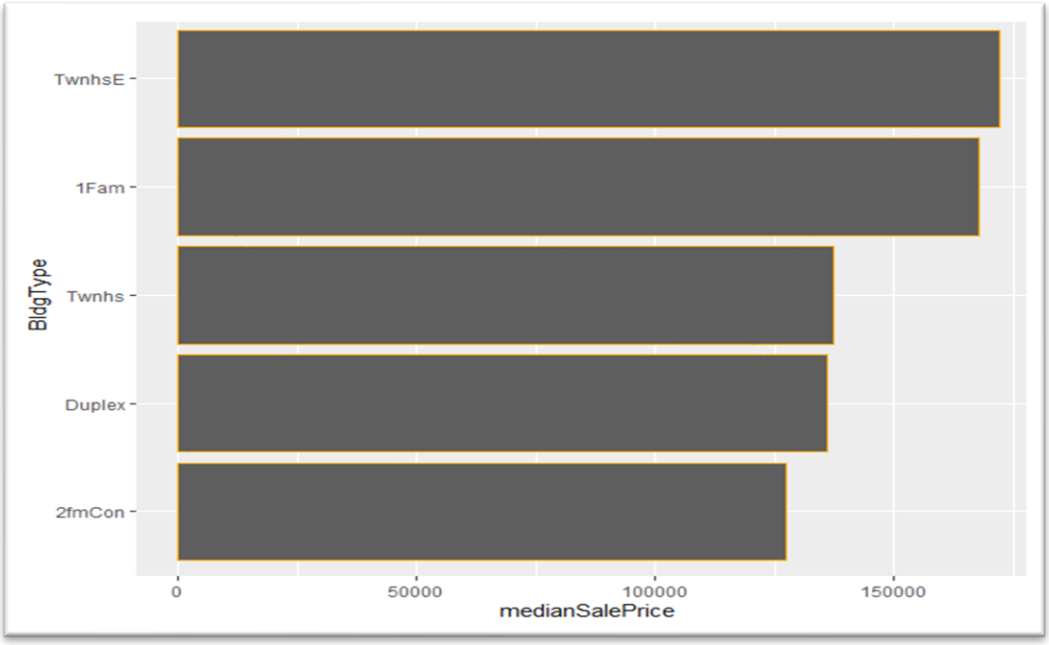


MSZoning of “FV” has the highest median sale price of about $205,000, while “C(all)” has the lowest of about $75,000.

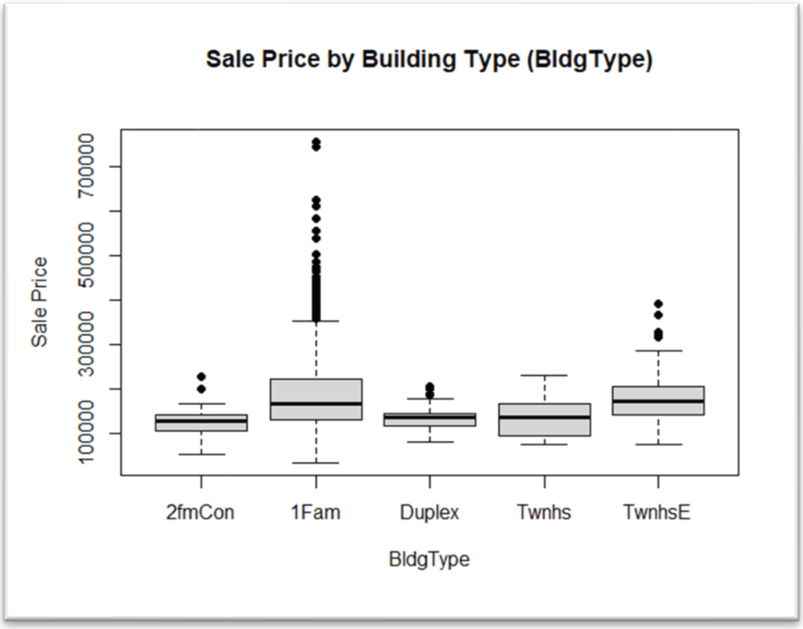


The box-plot shows us that MSZoning "RL" and "RM" have a lot of outliers. Moreover, the upper whisker of "FV" and "RL" are relatively the same, while the lower whisker of "C(all)", "RL" and "RM" are relatively the same as well.

#10: Visualization of BldgType (Independent Variable):



"TwnhsE" has the highest median sale price of about $172,000 followed closely by "1Fam" at around $168,000. While the building type "2fmCon" has the lowest median house sale price.



We can also see that the lower whiskers and median of 2fmCon and 1Fam are relatively the same, and the same can be said for Duplex, Twnhs and TwnhsE.

Modeling:

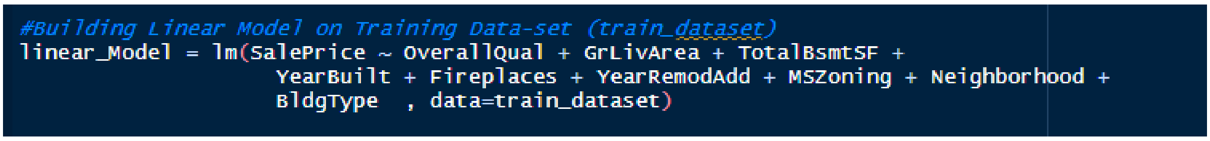
Creating Logical Test(s):

Based on our discussion prior, we want our model to be robust. Hence, we will gauge how our model performs in the following ways:

- *How does the model perform on the training dataset?*
    - The training dataset contains both, outlier and non-outlier houses in terms of SalePrice.
- *How does the model perform on the testing dataset?*
    - The testing dataset contains both; outlier and non-outlier houses which were **not** in the training dataset. Essentially, we want to find out how our model performs on a dataset which it hasn't seen before.
- *How does the model perform on outlier houses (> $290,000)?*
    - This dataset **only** has outlier houses which was in the testing dataset **only**.
- *How does the model perform on non-outlier houses (<= $290,000)?*
    - This dataset **only** has non-outlier houses which was in the testing dataset **only.**

Building The Model:

The formula for our model is as shown in the figure below. Moreover, it is important to note, that for our qualitative variables, we would need to exclude one category from our regression model. This excluded category will serve as a *reference variable*. That is, for example, if our reference category for the independent variable *OverallQual* was *OverallQual1,* we would then compare how the other categories (OverallQual2 to OverallQual10) in our model does in terms of SalePrice in comparison to *OverallQual1 (reference variable).*



```
#Building Linear Model on Training Data-set (train_dataset)
linear_Model = lm(SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF +
                  YearBuilt + Fireplaces + YearRemodAdd + MSZoning + Neighborhood +
                  BldgType  , data=train_dataset)
```

Figure 6.0 Showing the Linear Regression Formula.

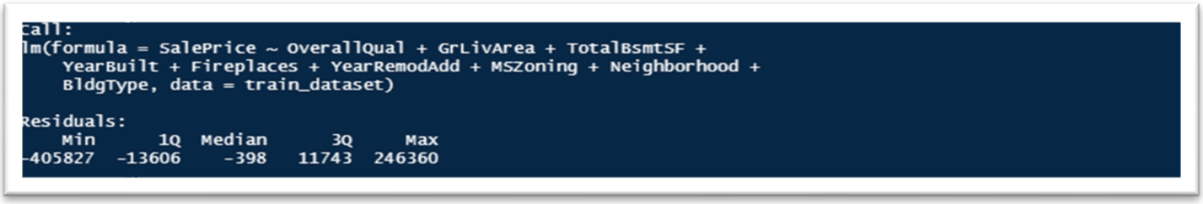
The qualitative variables and its reference categories are given below:

- OverallQual – OverallQual1
- MSZoning – C(all)
- Neighborhood – MeadowV
- BldgType – 2fmCon

Model Interpretation:

Upon running our model, we get the outputs as shown the figures below.

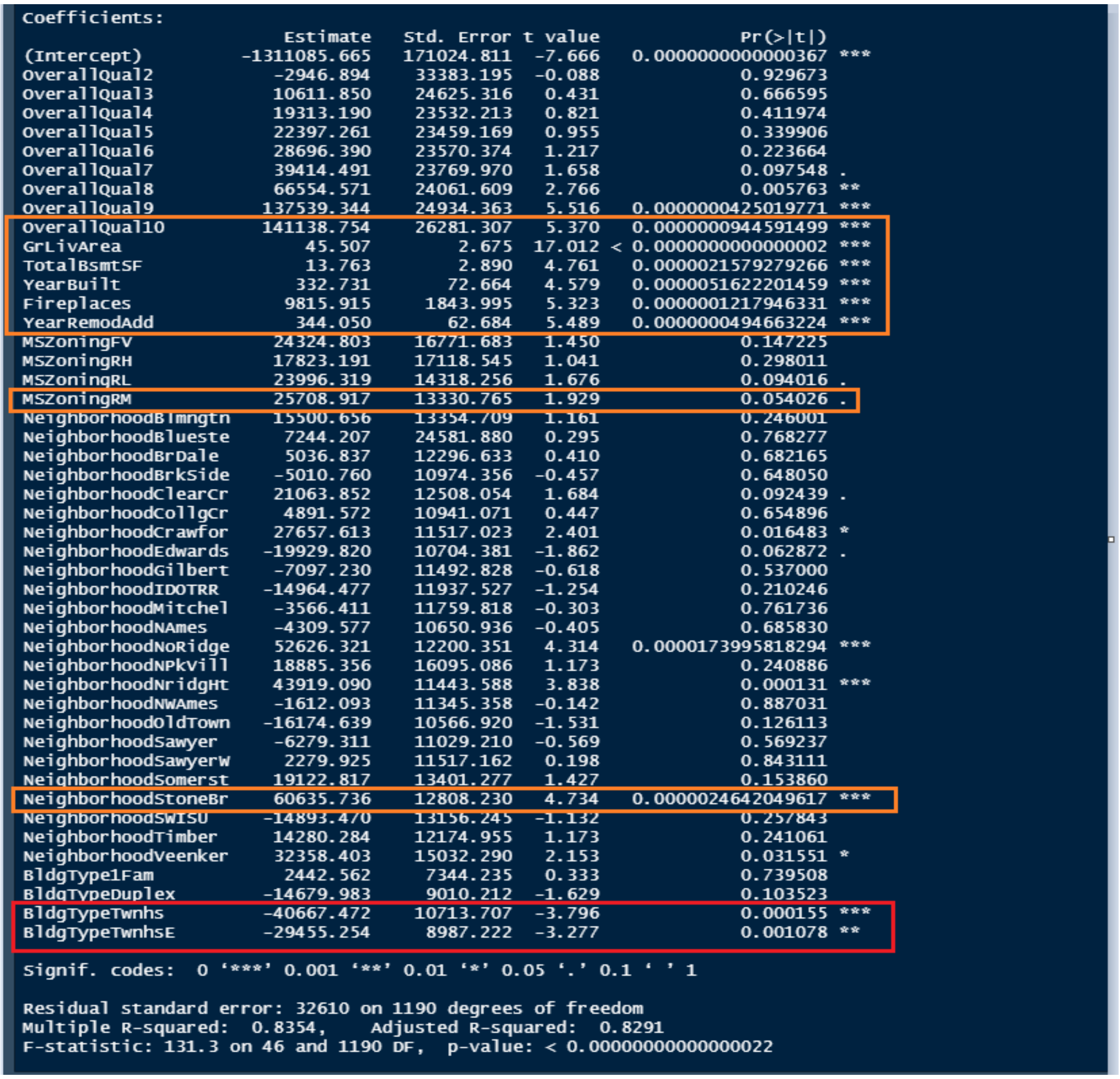
1.0 Summary Statistics:



```
call:
lm(formula = SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF +
    YearBuilt + Fireplaces + YearRemodAdd + MSZoning + Neighborhood +
    BldgType, data = train_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-405827  -13606    -398   11743  246360
```

Figure 6.A Showing Model Summary.



```
Coefficients:
                          Estimate   Std. Error t value             Pr(>|t|)
(Intercept)           -1311085.665   171024.811  -7.666   0.0000000000000367 ***
OverallQual2             -2946.894    33383.195  -0.088             0.929673
OverallQual3             10611.850    24625.316   0.431             0.666595
OverallQual4             19313.190    23532.213   0.821             0.411974
OverallQual5             22397.261    23459.169   0.955             0.339906
OverallQual6             28696.390    23570.374   1.217             0.223664
OverallQual7             39414.491    23769.970   1.658             0.097548 .
OverallQual8             66554.571    24061.609   2.766             0.005763 **
OverallQual9            137539.344    24934.363   5.516   0.0000000425019771 ***
OverallQual10           141138.754    26281.307   5.370   0.0000000944591499 ***
GrLivArea                   45.507        2.675  17.012 < 0.0000000000000002 ***
TotalBsmtSF                 13.763        2.890   4.761   0.0000021579279266 ***
YearBuilt                  332.731       72.664   4.579   0.0000051622201459 ***
Fireplaces                9815.915     1843.995   5.323   0.0000001217946331 ***
YearRemodAdd               344.050       62.684   5.489   0.0000000494663224 ***
MSZoningFV               24324.803    16771.683   1.450             0.147225
MSZoningRH               17823.191    17118.545   1.041             0.298011
MSZoningRL               23996.319    14318.256   1.676             0.094016 .
MSZoningRM               25708.917    13330.765   1.929             0.054026 .
NeighborhoodBlmngtn      15500.656    13354.709   1.161             0.246001
NeighborhoodBlueste       7244.207    24581.880   0.295             0.768277
NeighborhoodBrDale        5036.837    12296.633   0.410             0.682165
NeighborhoodBrkSide      -5010.760    10974.356  -0.457             0.648050
NeighborhoodClearCr      21063.852    12508.054   1.684             0.092439 .
NeighborhoodCollgCr       4891.572    10941.071   0.447             0.654896
NeighborhoodCrawfor      27657.613    11517.023   2.401             0.016483 *
NeighborhoodEdwards     -19929.820    10704.381  -1.862             0.062872 .
NeighborhoodGilbert      -7097.230    11492.828  -0.618             0.537000
NeighborhoodIDOTRR      -14964.477    11937.527  -1.254             0.210246
NeighborhoodMitchel      -3566.411    11759.818  -0.303             0.761736
NeighborhoodNAmes        -4309.577    10650.936  -0.405             0.685830
NeighborhoodNoRidge      52626.321    12200.351   4.314   0.0000173995818294 ***
NeighborhoodNPkVill      18885.356    16095.086   1.173             0.240886
NeighborhoodNridgHt      43919.090    11443.588   3.838             0.000131 ***
NeighborhoodNWAmes       -1612.093    11345.358  -0.142             0.887031
NeighborhoodOldTown     -16174.639    10566.920  -1.531             0.126113
NeighborhoodSawyer       -6279.311    11029.210  -0.569             0.569237
NeighborhoodSawyerW       2279.925    11517.162   0.198             0.843111
NeighborhoodSomerst      19122.817    13401.277   1.427             0.153860
NeighborhoodStoneBr      60635.736    12808.230   4.734   0.0000024642049617 ***
NeighborhoodSWISU       -14893.470    13156.245  -1.132             0.257843
NeighborhoodTimber       14280.284    12174.955   1.173             0.241061
NeighborhoodVeenker      32358.403    15032.290   2.153             0.031551 *
BldgType1Fam              2442.562     7344.235   0.333             0.739508
BldgTypeDuplex          -14679.983     9010.212  -1.629             0.103523
BldgTypeTwnhs           -40667.472    10713.707  -3.796             0.000155 ***
BldgTypeTwnhsE          -29455.254     8987.222  -3.277             0.001078 **

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32610 on 1190 degrees of freedom
Multiple R-squared:  0.8354,	Adjusted R-squared:  0.8291
F-statistic: 131.3 on 46 and 1190 DF,  p-value: < 0.00000000000000022
```

Figure 6.B Showing the Model Summary.

In terms of our model's summary statistics, we do firstly notice that the *Adjusted R-Squared* value is 0.8291 (82.91%). The *Adjusted R-Squared* value informs us about the percentage of variance of our dependent variable, which in our case is *SalePrice,* that is explained by our independent variables.

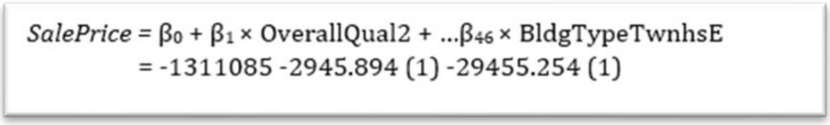
Secondly, the *residuals* show the differences between the actual values and the predicted values in terms of MIN(Minimum), 1Q (1st Quartile), Median, 3Q (3rd Quartile) and MAX(Maximum). The expectation is that the residuals should be symmetrical.

Thirdly, the *Probability Value (p-value)* is used to identify the statistical significance of an independent variable, whereby it tells us the probability of attaining the results as extreme as the ones shown, under the assumption that the null hypothesis is true.

Fourthly, the *Standard Error (Std.Error)* shows the standard error of our coefficients. Given that this is an error, we need this as low as possible.

Fifthly, the *t-value* is derived from dividing the coefficients with its respective standard error. This basically means that the lower the *Std.Error,* the higher the *t-value*, and the higher the *t-value*, the lower the *p-value.*

Lastly, our model's interpretation is as follows:



$$\begin{aligned} \textit{SalePrice} &= \beta_0 + \beta_1 \times \text{OverallQual2} + \ldots\beta_{46} \times \text{BldgTypeTwnhsE} \\ &= -1311085 -2945.894\ (1) -29455.254\ (1) \end{aligned}$$

We can see that the SalePrice is equal to β_0 (Beta 0), which is our constant, plus all the coefficients (β_1 to β_{46}) of the independent variables multiplied by the units of the independent variables itself. It is essential to note here, that for categorical variables, this would be 1, indicating the presence of it, and 0, if one wants to calculate SalePrice in terms of an independent variable's respective reference variable. The interpretations are translatory in accordance to an independent variables' type. For example, for **qualitative variables** such as OverallQual3, we could say that for those houses which has an overall quality of 3 rating, we expect to see an increase of roughly $10,600 in SalePrice compared to those houses which has an overall quality of 1 (OverallQual1 – *reference variable*), while holding other determinants constant. Likewise, for **quantitative variables** such as GrLivArea, we could say that for each unit increase in ground living area, we expect to see an increase of roughly $46.00, holding other variables constant.

Regression Diagnostic Plots:

From the residual plot given in *Figure 7.0* below, we can see that the residuals are considerably large for higher predicted values,v that is, values above $290,000 when compared to lower predicted values. This is something that we expected, given that we had *outliers* (over-valued houses) in our dataset above $290,000. In addition, we can also see that the linearity assumption is met given an almost flat red line.

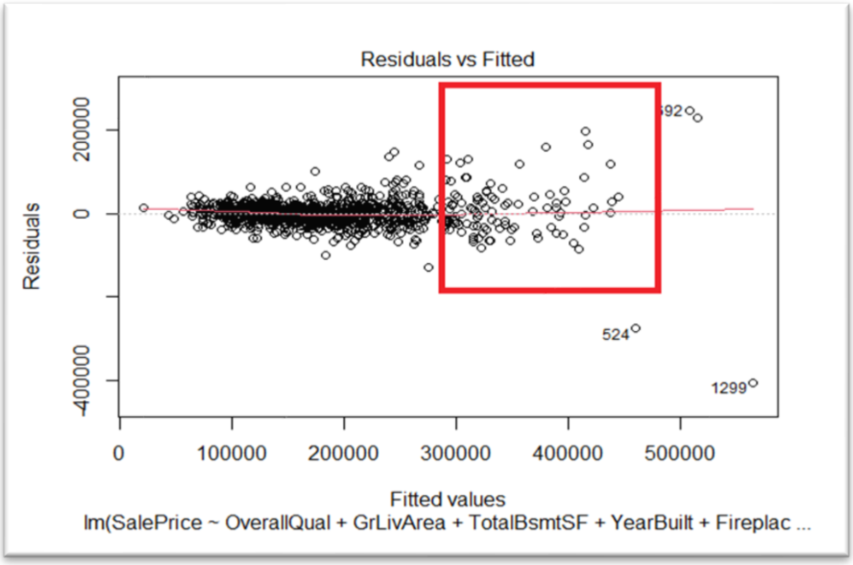


Figure 7.0 Showing high residuals of **SalePrice only** on the high predicted values.

The figure below shows other diagnostic residual plots which helps us to identify non-linearity, non-constant variance as well as other troublesome observations. There is nothing alarming and unexpected shown in these plots.

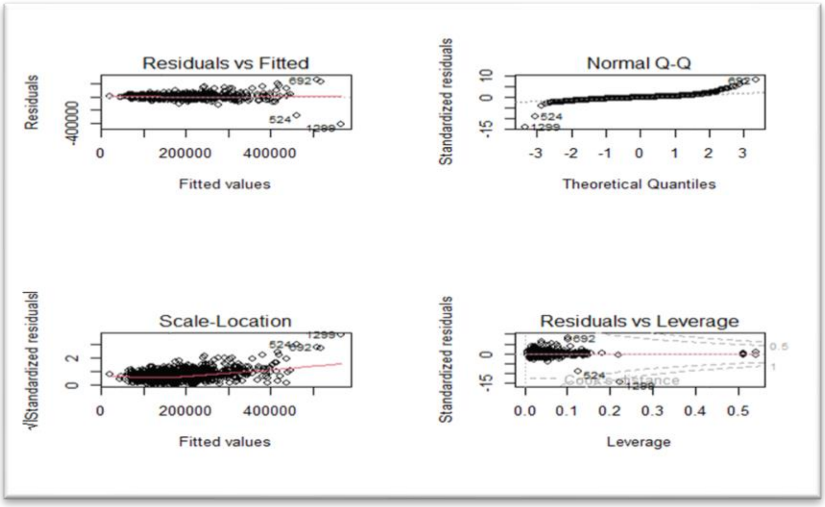


Figure 7.1 Showing other diagnostic plots. Nothing alarming shown.

Furthermore, we also hypothesized that *if* we were to use *LogSalePrice* as our dependent variable, which has *outliers* on both sides of the box-plot, we would see greater residuals on both the lower and the higher predicted values. Hence, due to this reason, we refrained from using *LogSalePrice* as our dependent variable in favor of *SalePrice*, which would only give us higher residuals on higher predicted values. From the residual plot of *LogSalePrice*, we see that our hypothesis holds true. That is, we see greater residuals on both the lower predicted and higher predicted values.

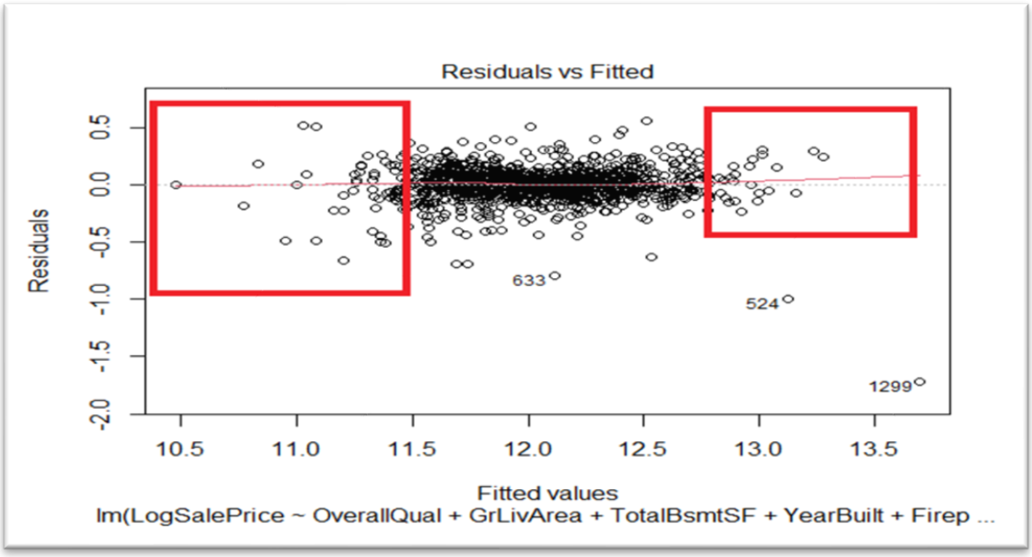
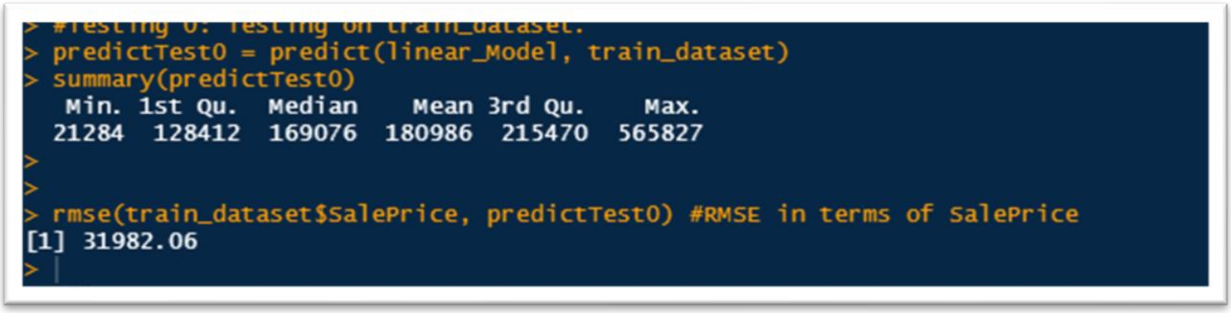


Figure 7.2 Showing high residuals of **LogSalePrice** on **both** low and high predicted values.

Root-Mean-Square-Deviation (RMSE):

The *Root-Mean-Square-Deviation*, also known as *Root-Mean-Square-Error* aggregates each residual by squaring each residual, and then taking the square root of the average residual. We have calculated and tested the RMSE on four datasets in accordance to our discussion prior. These datasets include:

Train_dataset – This is the dataset which our model was built upon. This dataset has a mixture of outlier (> $290,000) and non-outlier (<= $290,000) houses.

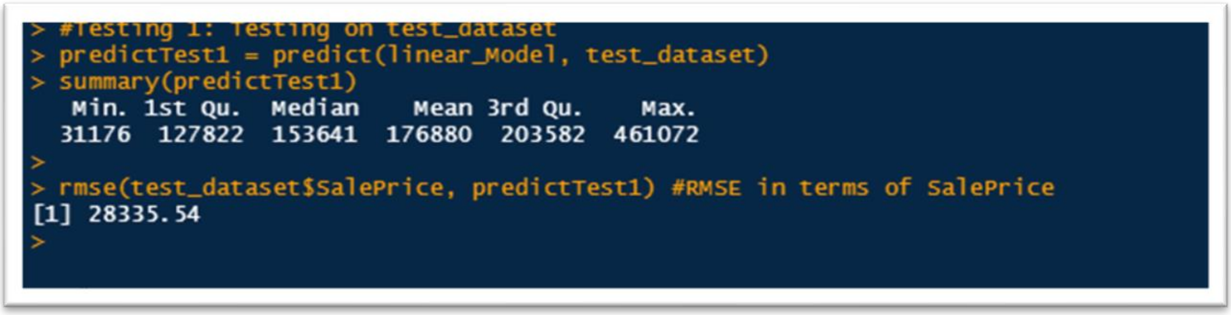


```
> #Testing 0: Testing on train_dataset.
> predictTest0 = predict(linear_Model, train_dataset)
> summary(predictTest0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  21284  128412  169076  180986  215470  565827 
>
>
> rmse(train_dataset$SalePrice, predictTest0) #RMSE in terms of SalePrice
[1] 31982.06
> 
```

Figure 7.0 Showing RMSE on Train_dataset.

We can see that we expect that the predicted values of house sale prices to be $31,982.06 higher or lower than the actual values on average.

Test_dataset – The dataset has a mixture of outlier and non-outlier houses which wasn't included in the training dataset. This dataset is formed by combining the below mentioned datasets.

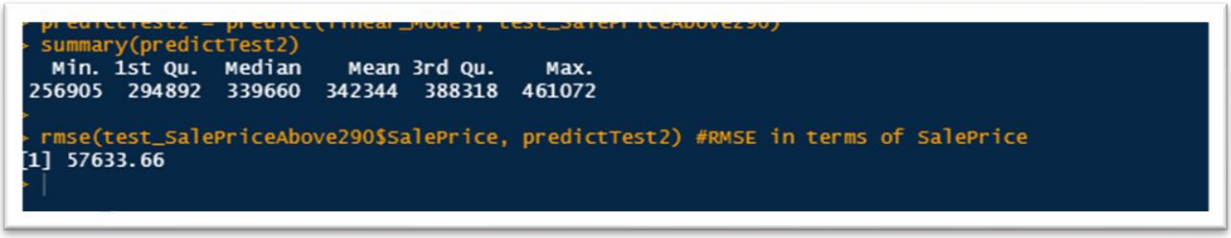


```
> #Testing 1: Testing on test_dataset
> predictTest1 = predict(linear_Model, test_dataset)
> summary(predictTest1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  31176  127822  153641  176880  203582  461072 
>
> rmse(test_dataset$SalePrice, predictTest1) #RMSE in terms of SalePrice
[1] 28335.54
>
```

Figure 7.1 Showing the RMSE on Test_dataset.

We can see that we expect that the predicted values of house sale prices to be $28,335.54 higher or lower than the actual values on average.

Test_SalePriceAbove290 **–** The dataset has outlier houses which can also be found in the test_dataset.

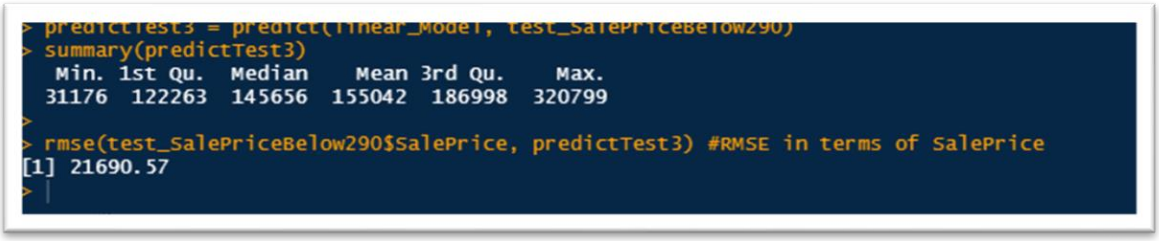


```
> predictTest2 = predict(linear_Model, test_SalePriceAbove290)
> summary(predictTest2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
256905  294892  339660  342344  388318  461072
>
> rmse(test_SalePriceAbove290$SalePrice, predictTest2) #RMSE in terms of SalePrice
[1] 57633.66
> 
```

Figure 7.2 Showing the RMSE on Test_SalePriceAbove290 dataset.

We can see that we expect that the predicted values of house sale prices to be $57,633.66 higher or lower than the actual values on average. This high RMSE was something that was expected and is clearly found to hold true according to regression diagnostic plots which was seen earlier. It is also important to note here that this RMSE would have been even greater had we not included the *outliers*, that is, houses with sale price greater than $290,000 in our training dataset (*train_dataset*). Since we wanted to have a robust model in the sense that it would some predictive capabilities on high-valued (*outlier*) houses, we had to include the *outliers* in our training dataset. However, we also understood and accepted the trade-off of doing so, which was that it would slightly increase the residuals of lower-valued (*non-outlier*) houses as well.

Test_SalePriceBelow290 **–** The dataset has non-outlier houses which can also be found in the test_dataset.



```
> predictTest3 = predict(linear_Model, test_SalePriceBelow290)
> summary(predictTest3)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 31176  122263  145656  155042  186998  320799
>
> rmse(test_SalePriceBelow290$SalePrice, predictTest3) #RMSE in terms of SalePrice
[1] 21690.57
> 
```

Figure 7.3 Showing the RMSE on Test_SalePriceBelow290 dataset.

We can see that we expect that the predicted values of house sale prices to be $21,690.57 higher or lower than the actual values on average. This relatively low RMSE when compared to the RMSE of *Test_SalePriceAbove290* is something which was expected. Moreover, it is also important to note here that we could have reduced this RMSE if we excluded the *outliers* from the training dataset (*train_dataset*).

Discussion and Conclusion:

At the beginning of our report, we stated some benchmarks to evaluate the success of our data-mining project. As per our benchmark, firstly, we see that our *Adjusted R-Squared* value is 0.8291, which exceeds our expectation. Moreover, all our quantitative variables show statistical significance given that its *p-values* are lower than 0.05. However, for our some of our qualitative variables, such as OverQual3, we do not find it statistically significant that a house with overall quality of 3 would increase the sale price of a house by $10611.850 in comparison to OveralQual1. This finding is similar to some of the categories of MSZoning, Neighborhood and BldgType when compared to its respective reference variable given a p-value lower than 0.05.

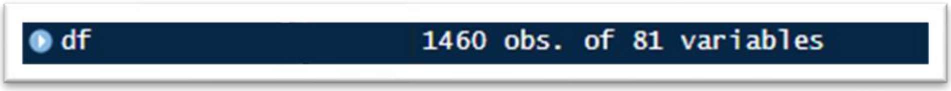
Interestingly, we do notice that some very important factors from a customers' perspective are not in our model such as the number of bedrooms, bathrooms, etc. This is because these variables either had a high multi-collinearity with other variables, such as, number of bedrooms having a high multi-collinearity with ground living area, or had low correlation with sale price. Nevertheless, we do see ground living area (GrLivArea) being statistically significant in influencing sale price. In addition to this, other statistically significant independent variables include TotalBsmtSF, YearBuilt, Fireplaces, and YearRemodAdd. It is no surprise to see that OverallQual10 increases the selling price of a house by roughly $141,100 in comparison to OverallQual1. In hindsight, this tells us to advise our consultees to set a remodeling goal of achieving an overall quality rating of 10 in order to maximize a house's sale price. Furthermore, our client should also seek to maximize the GrLivArea, TotalBsmSF and Fireplaces as possible, as with each unit increase, the sales price increases as well. We also believe that its better to remodel and sell houses which has an MSZoning classification of "RM" and is in the neighborhood of "StoneBr" compared to zoning classification of "C(all)" and "MeadowV" respectively. Moreover, it is ill-advised for the remodeling company to remodel a home with a dwelling of BldgTypeTwnhs or BldgTypeTwnhsE as houses with these two dwelling types decreases the sale price of a house by approximately $40,700 and $29,400 respectively, when compared to BldgType2fmCon. Hence, given a choice, between these three dwellings, it is better to choose BlgType2fmCon.

Moreover, in terms of our models' predictive abilities, we can see that our model performs well on predicting sale prices of houses which have an actual sale price which is lower than equal to $290,000. In contrast, our model performs relatively poorly when on predicting sale prices of house which has an actual sale price above $290,000. The question the arises *how would one know the sale price beforehand, in order to decide whether to use our model or not?* This is where some other relevant variables such as *capital value, market value,* etc. of houses would be useful. Essentially, these attributes help us to gauge whether the sale price would be higher or lower than $290,000.

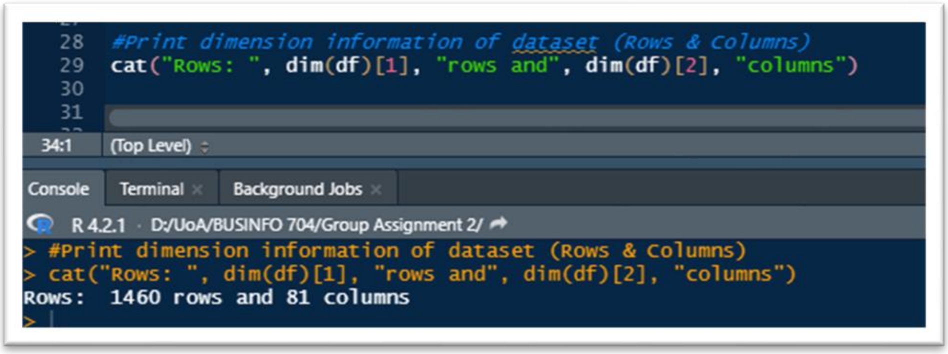
Needless to say, if we acquired these variables, then it would be wise to include these variables in our model, and by implication, would increase the prediction accuracy of our new model.

Appendix:

Appendix 1: Showing the total rows/objects of data (1460) and total variables/columns (81) within our dataset called "df".



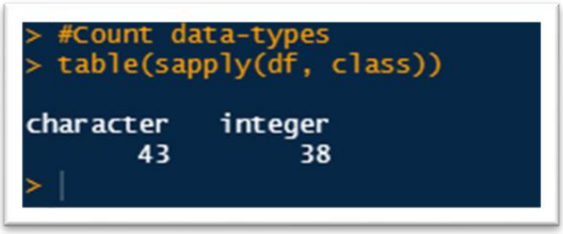
```
df                    1460 obs. of 81 variables
```



```
#Print dimension information of dataset (Rows & Columns)
cat("Rows: ", dim(df)[1], "rows and", dim(df)[2], "columns")
```

```
> #Print dimension information of dataset (Rows & Columns)
> cat("Rows: ", dim(df)[1], "rows and", dim(df)[2], "columns")
Rows:  1460 rows and 81 columns
>
```

Appendix 2: Showing the total character (43) and integer (38) variables.



```
> #Count data-types
> table(sapply(df, class))

character   integer 
       43        38 
>
```