

Assignment 3: Data Analysis and Regression

Sri Seshadri

7/8/2017

1. Introduction

This report discusses linear regression models and the model generation process, for estimating or predicting the sales price of “typical” homes in Ames, Iowa. The following were deemed good predictors : 1. Above grade living area - Continuous variable 2. Total basement area - Continuous variable 3. Total Bathrooms - Continuous variable 4. Lot Area - Continuous variable 5. Neighborhood Tiers - Discrete. (Neighborhoods grouped into tiers, refer section 6.1) Several models were fit and their goodness of fit were assessed and their performances were compared. The model $\text{Log}(Sale\ Price) = 4.979e+00 + 1e-04(\text{TotalBsmtSF}) + 1.236e-04(\text{GrLivArea}) + 1.251e-06(\text{LotArea}) + 2.809e-02(\text{TotalBath}) - 2.160e-01(\text{Tier1}) - 1.585e-01(\text{Tier2}) - 9.391e-02(\text{Tier3}) - 2.740e-02(\text{Tier4})$ was found to be better performing with Mean Absolute Error (MAE) of \$21,524.

2. Sample definition

It is assumed that typical home buyers are those that move from apartments to single family or town homes. Also apartments are less likely to be sold to individuals as they remain holdings of owners for rental income. Single family and town homes belong to “Residential Low density” (RL) zoning classification in the city of Ames. Data belonging to only to the RL zone is considered for analysis and model development. Also, it is assumed that typical homes have paved streets for access and above grade living area greater than 800 square feet. Sales data belonging to homes that were sold in abnormal conditions such as trade in, foreclosure or short sale are not included in the analysis. Also, sales between family members, sale of adjoining lot, linked properties are omitted from the data. Homes with no basements are excluded from the analysis at this time. Table 1 shows the waterfall of the data not included in the data and the eligible samples.

Table 1: Drop waterfall

DropCondition	counts
01: Not LowDensityZone	657
02: Not Normal/Partial Sale	189
03: Street Not Paved	3
04: Less than 800 SqFt	41
05: No Basement	48
99: Eligible Sample	1992

2.1 Variables of interest for modelling

The following variables in the data were deemed to be of interest for model building. The choice of parameters was based upon initial Exploratory Data Analysis (EDA) and subject matter expertise. See appendix A.1 for data quality checks.

Table 2: Variables of interest

LotArea	GrLivArea	GarageArea
LotConfig	BsmtFullBath	MoSold
Neighborhood	BsmtHalfBath	YrSold
BldgType	FullBath	SaleCondition
HouseStyle	HalfBath	SalePrice
OverallCond	BedroomAbvGr	FirstFlrSF
YearRemodel	KitchenQual	SecondFlrSF
TotalBsmtSF	TotRmsAbvGrd	LotArea

2.2 Training and validation samples.

From the eligible samples, 70% of the data is randomly sampled to be used as the dataset for model development. This dataset would be referred to as training dataset. The remaining 30% is used as the validation set to evaluate the model performance of predicting sale price on data that is outside the training set. Table 3 shows the split of the total eligible samples.

Table 3: Training and Validation sampling

Data	Samples
Training set	1394
Validation set	597

3. Exploratory Data Analysis (EDA)

For exploratory analysis, entire sample frame is used, so any anomalies that were missed surfaces in this process. While EDA by itself does not become an data cleaning step, but certainly allows the opportunity to identify issues in the data. In this section we will focus only on the continuous variable. Exploratory analysis on categorical variables can be found in https://github.com/srivathsesh/RegressionAnalysis/blob/master/Assignment1_Seshadri.pdf.

3.1 EDA of continuous variables

It is hypothesized that bigger the house, more likely the occupancy and sale value. Higher occupancy means likely higher lot area, living area, basement area, garage space for parking, total number of rooms and baths. Also it is hypothesized that newer homes are likely to be more valued than the older homes. Before exploring each of the potential predictors, it will be useful to see if there is multicollinearity amongst the predictors. Figure 1 shows, potential relationship between Above grade living area, Sale price and Total basement area.

Figure 2 shows the relationship between total basement area and the sale price. There are few outliers in the basement area. To explore the relationship better, data corresponding to basement area above 2500 square feet is removed. It is seen that total basement area is very promising predictor. It may be used along with other predictors to model sales price.

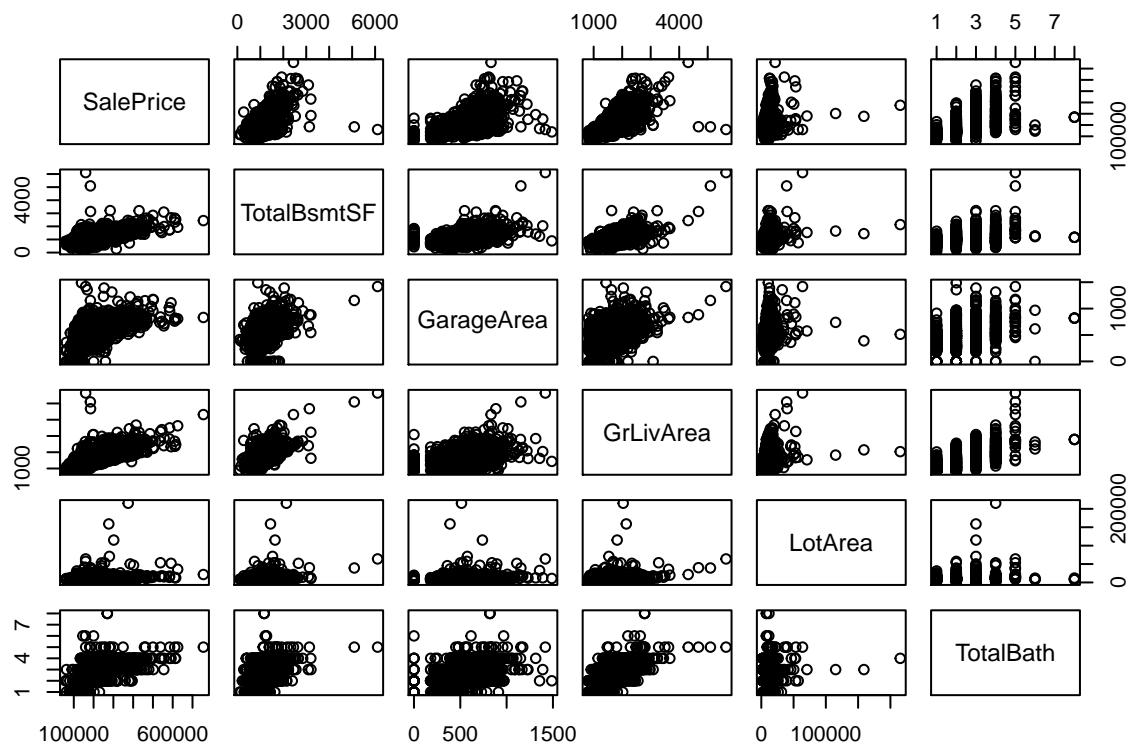


Figure 1: Multicollinearity check amongst potential predictors

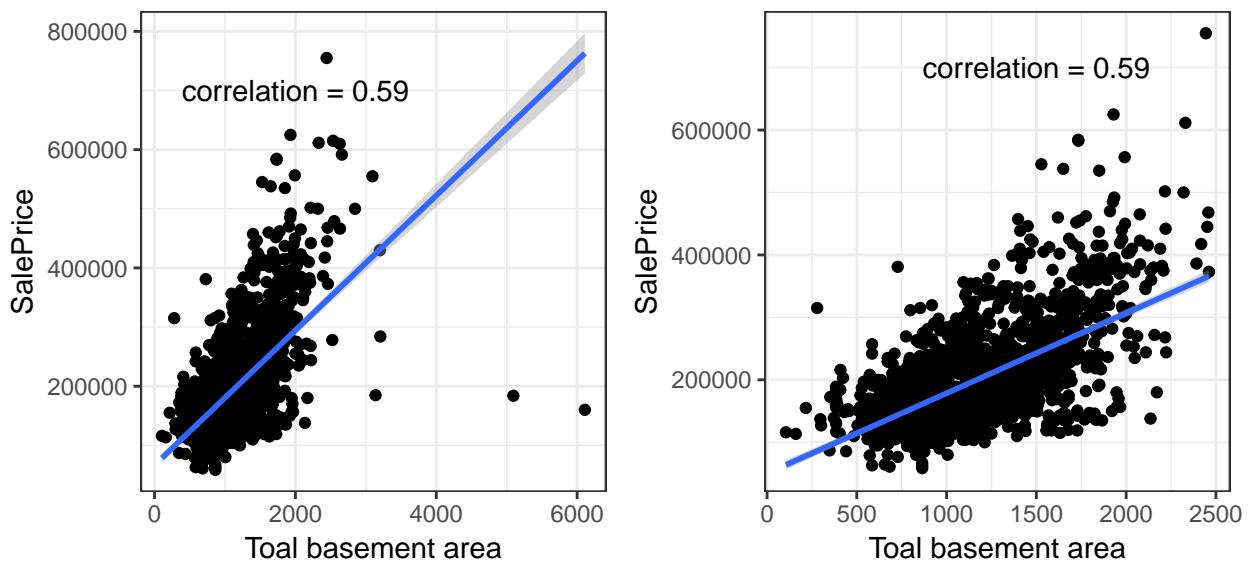


Figure 2: Sale price vs Basement area

Figure 3 shows the relationship between Living area and Total basement area with Sale price. There seem to be a linear relationship between living area and sale price when living area is less than 4000 square feet. Similarly with Garage area when greater than zero and less than 1000 square feet. Figure 4 explores the relationship between Garage Area, Living area and Total basement area. There is correlation amongst the three variables.

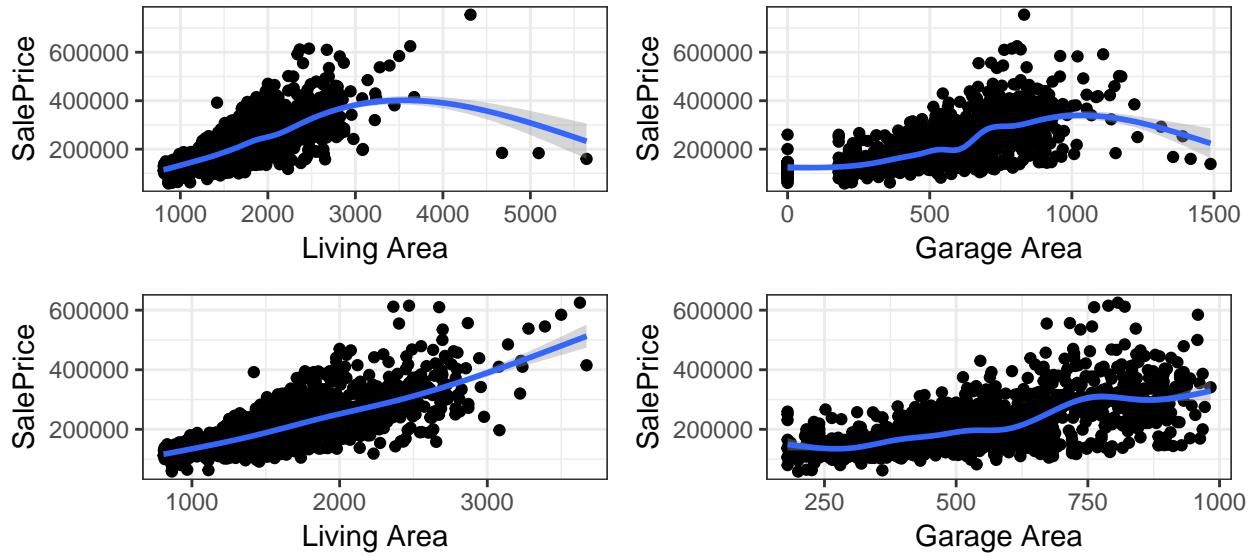


Figure 3: Sales Price vs Garage Area and LivingArea

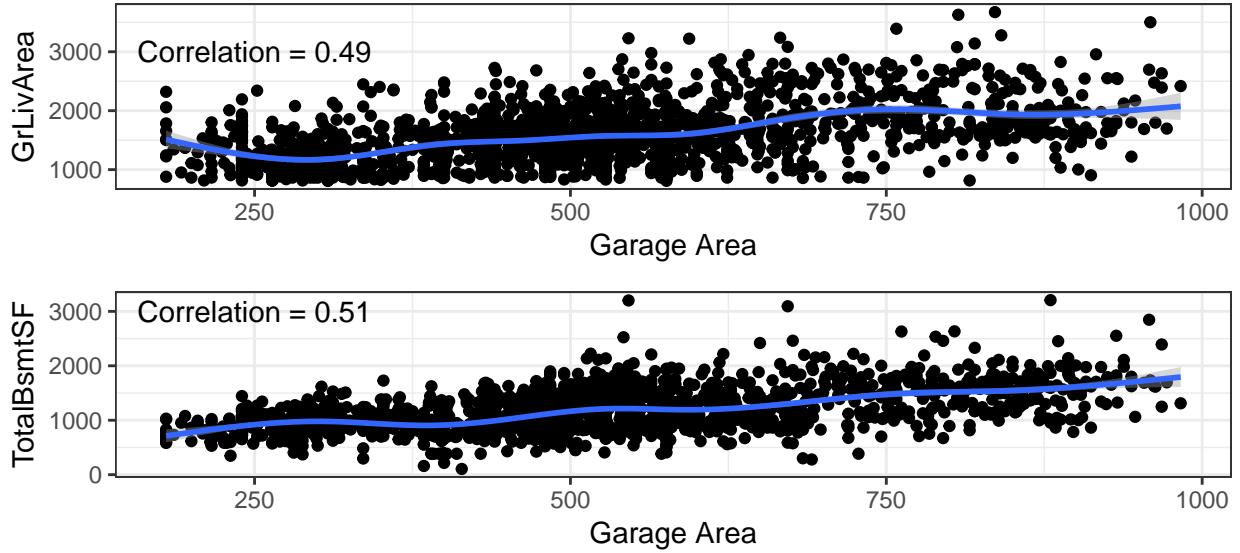


Figure 4: GarageArea vs Living and Basement area

4. Simple Linear Regression Models

From the Exploratory Data analysis, it is seen that the above grade living area and total basement area are good candidates for predictor variables for regression modeling. The regression models would be built on the training data set. Since linear relationships between Sale price and basement area and living are are more pronounced when above grade living area is less than 4000 square feet and garage area less than 1000 square feet; we will remove data corresponding to above grade living area greater 4000 square feet and total garage area greater 10000 square feet. While it is best to update the drop conditions in section 2, at this time, the additional drop conditions would be applied to the training set. The training set data with new drop conditions applied is called “trainFiltered” (The R output reference this as the data)”

4.1 Simple linear regression model with “above grade living area” as predictor

The model fit results for a single linear regression model; **SalePrice = 7200.4 + 122.65 * AboveGradeLivingArea** is seen below. We fail to reject the null hypothesis of intercept = 0; at 95% significance level with t-statistic being 1.5. The overall F test for regression, shows the regression effect to be statistically significant with F statistic of 1712 (p value = 0.000). It would be appropriate to fit a no intercept model. From a goodness of fit perspective (refer Figure 5), the residuals are of mean 0 and distributed fairly normally based on “thick pen test” on Q-Q plot. However, the residulas are not homoscedastic, the residuals vs predictor have an open funnel shape, i.e. the varaiton in the residuals increases as the living area increases. The model is not suitable.

```
##  
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = trainFiltered)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -188199  -27705   -3701   19890  314519  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 7200.401   4776.743   1.507   0.132  
## GrLivArea    122.647      2.964  41.375 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 51660 on 1374 degrees of freedom  
## Multiple R-squared:  0.5547, Adjusted R-squared:  0.5544  
## F-statistic:  1712 on 1 and 1374 DF,  p-value: < 2.2e-16
```

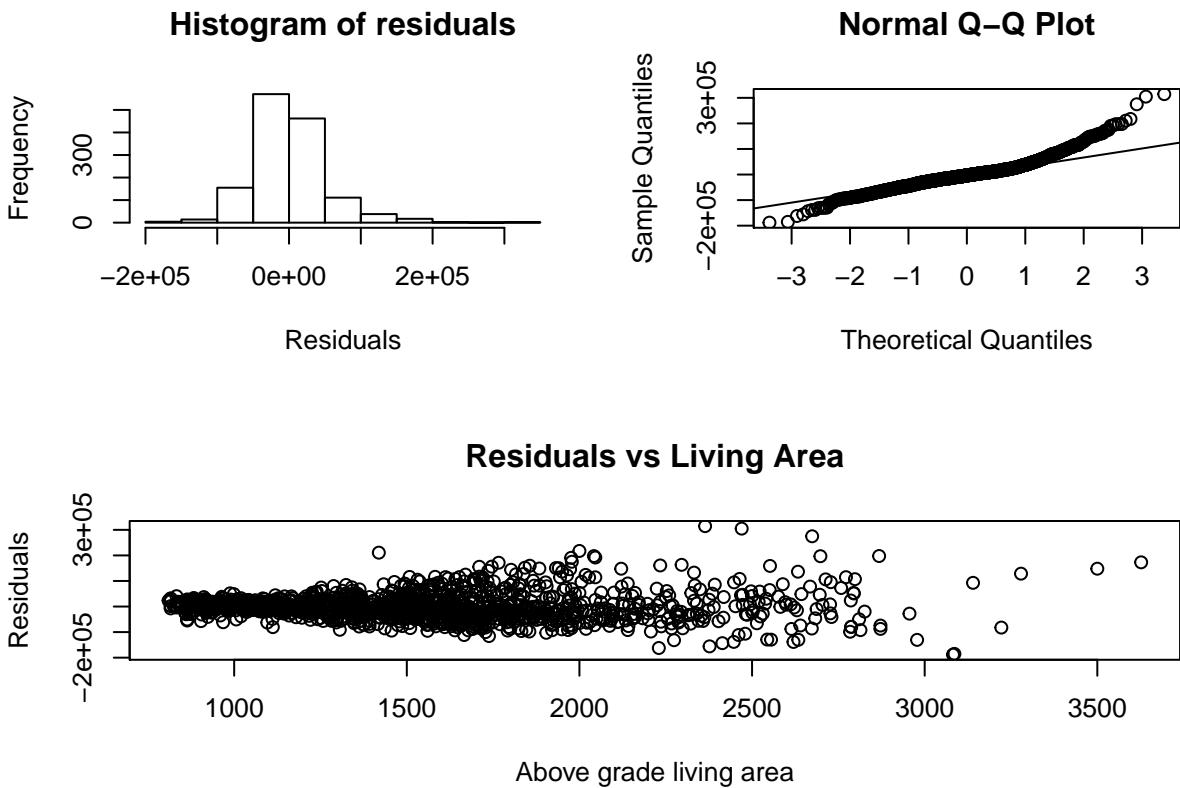


Figure 5: Model diagnostics Sale Price \sim GrLivArea

4.1.1 Non - intercept model with Above grade living area as predictor

In the previous section , it was shown that the intercept was not significant, therefore a no-intercept model is fit. Below are the results of **SalePrice = 126.9 * GrLivArea** . The coefficient for slope (GrLivArea) is statistically significant with t-value of 146.8. The overall regression effect is statistically significant F statistic of 21550 (p value = 0.000) and has a better Adjusted R-Squared value of 0.94. When the residual plots(Figure 6) are compared with the previous model (Figure 5), the kurtosis of the residuals has increased, but the funnel pattern still remains. There is no practical implication of this fit.

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea + 0, data = trainFiltered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -194172 -27499 -1858  21393 311615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## GrLivArea 126.9214     0.8647   146.8    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51690 on 1375 degrees of freedom
## Multiple R-squared:  0.94, Adjusted R-squared:  0.94
## F-statistic: 2.155e+04 on 1 and 1375 DF, p-value: < 2.2e-16
```

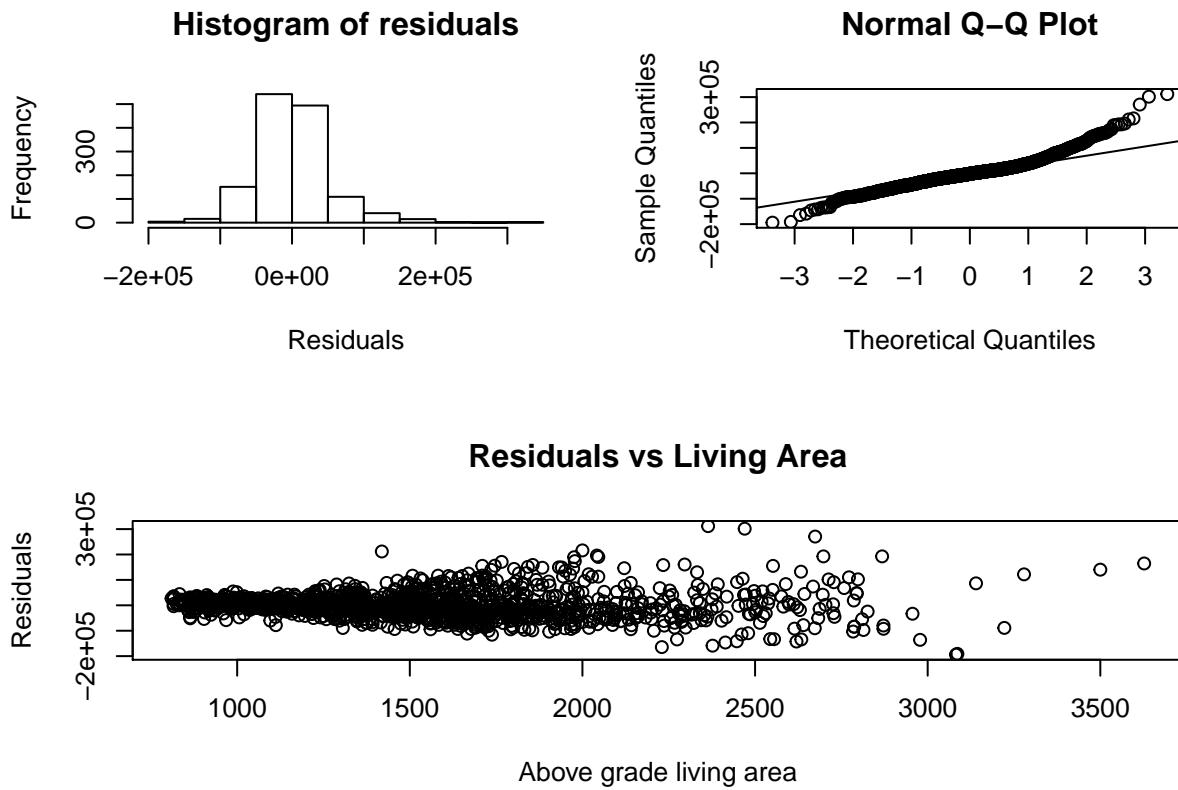


Figure 6: No intercept model

4.2 Simple linear regression with Total Basement area as predictor

The results of a simple linear model with Total basement area as predictor (**SalePrice = 47859.31 + 129.96 * TotalBsmtSF**) is shown below. The t-tests for statistical significance of regression coefficients prove the coefficients to be significant with t statistics for intercept and slope at 9.17 and 29.93 respectively. The overall F test for significance of regression effect shows the regression to be statistically significant with F statistic at 896.2 (p value = 0.0000). The sum of squares error (60230) is much greater than the one with living area as predictor. Making basement area a poor predictor by itself. While the model diagnostics pass the normality tests, but is not homoscedastic as the variance increases at higher basement area. Moreover, the residuals are within +/- 120,000 dollars (2 sigma level or 95 out of 100 times). Which means the estimation of home price could be off by upto 120,000 dollars 95 out of 100 of instances of estimation. Therefore the model cannot be used in practice.

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF, data = trainFiltered)
##
## Residuals:
##    Min     1Q   Median     3Q    Max
## -180515 -42291 -12265  36108 326316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47859.313   5216.177   9.175 <2e-16 ***
## TotalBsmtSF   129.961      4.341  29.936 <2e-16 ***
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60230 on 1374 degrees of freedom
## Multiple R-squared:  0.3948, Adjusted R-squared:  0.3943
## F-statistic: 896.2 on 1 and 1374 DF,  p-value: < 2.2e-16

```

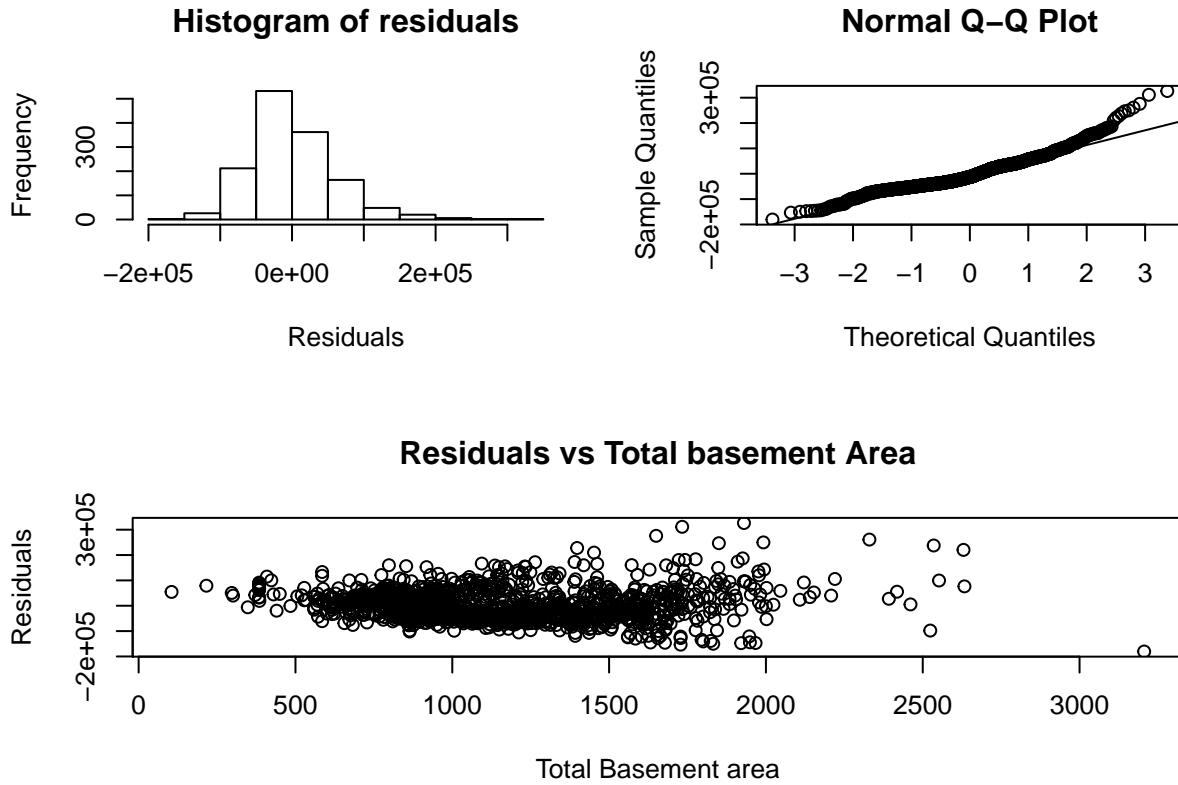


Figure 7: Simple Linear Regression - Total Basement Area as predictor

5. Multiple linear regression model

The above simple linear regression models by themselves were not sufficient in modeling the sale price of typical homes. In this section we will explore multiple linear regression models with total basement area and above grade living area as predictors.

Below are the goodness of fit results for the model $\text{SalePrice} = -502244.02 + 84.11 * \text{TotalBsmtSF} + 97.608 * \text{GrLivArea}$ on the training set. The t-tests show that partial regression coefficients are statistically significant with P values = 0.000. Also the overall F test for regression effect, show the model to be statistically significant in explaining the variation in sale price with F statistic at 1579 (P value = 0.000). The model has an adjusted R-squared of 70% and the residuals seem to fairly follow the normal distribution with mean zero. However the tails of the distributions of the residuals are thicker. With this model, we are likely (95% of the time) to be estimating sale price with an error of =/ \$86000 (44% of mean Sale price or 48% of median sale price). From a practical use perspective, this model needs improvement.

```

##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea, data = trainFiltered)
##
## Residuals:

```

```

##      Min      1Q   Median      3Q     Max
## -184885 -20085     730    20554  235175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50244.022   4545.332 -11.05 <2e-16 ***
## TotalBsmtSF    84.112     3.313   25.39 <2e-16 ***
## GrLivArea      97.608     2.638   37.01 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42640 on 1373 degrees of freedom
## Multiple R-squared:  0.697, Adjusted R-squared:  0.6965
## F-statistic: 1579 on 2 and 1373 DF, p-value: < 2.2e-16

```

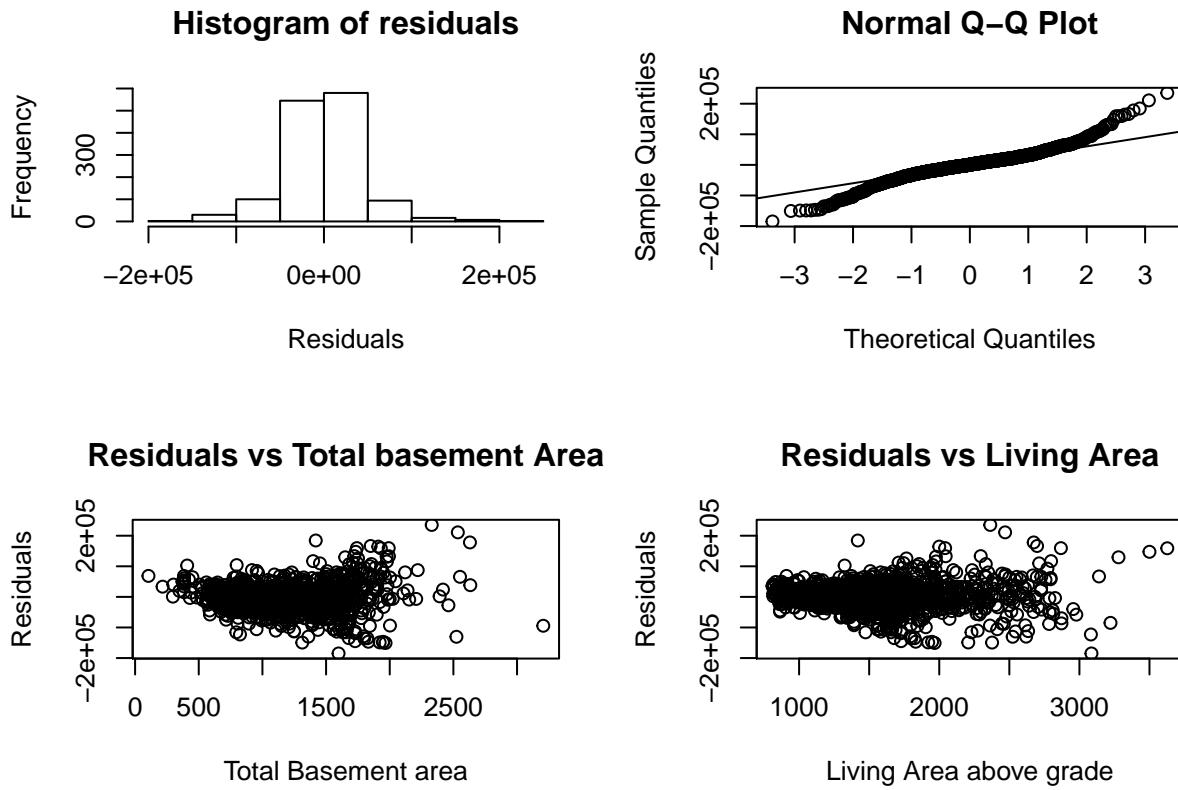


Figure 8: Multiple Linear regression

5.1 Model comparisons

In section 4, it was seen that the no-intercept model had a better fit compared to the one with intercept when above grade living area was used as the predictor. It was seen that the sum of square errors of the model with living area as predictor was less than that of the model with total basement area as the predictor. The adjusted R-Squared value of the no-intercept model with living area as the predictor was 94%. In this section the multiple linear regression model is compared with the linear regression models.

In the table 4 below, the simple and multiple linear models are compared. *For all the models in the table below, the residuals exhibit a funnel shaped pattern when plotted against the predictors and they fairly meet normality by thick pen test on QQ plot. However the residuals have*

heavier tails, contributed by the variation at higher levels of the predictors. Note, that the total variance of the response variable is the same for all the models.

Table 4: Model comparison

Model	MSres	F Stat	P Value	Adj R^2	Fig#_Reference
SalePrice = 126.9 * GrLivArea	51686.54	21545.58	0	0.94	6
SalePrice = 47859.31 + 129.96 * TotalBsmtSF	60233.68	896.15	0	0.39	7
-502244.02 + 84.11 * TotalBsmtSF + 97.608 * GrLivArea	42635.05	1579.03	0	0.7	8

Since the residuals are exhibiting funnel shaped patterns, the models above are not fit for application. *The models either needs transformation of response variable (Sale Price) or additional predictors.* See appendix A.2 for Nested model approach for comparing the simple linear and multiple linear regression models. In the next section, we will explore residuals of the Multiple linear regression with potential predictors that are not in the model.

6. Neighborhood Accuracy

If residuals of the model show structure or pattern when plotted against omitted variable in the model, then the omitted variable could be a potential predictor that may be added to the model. Figure 9 shows that there is a structure between residuals and neighborhood. This affirms the conclusion during EDA of sales price by neighborhood. (refer page 2 of https://github.com/srivathsesh/RegressionAnalysis/blob/master/Assignment1_Seshadri.pdf). The Sawyer and Bloomington Neighborhoods are better predicted. In Figure 9, the neighborhoods below the red lines are over predicted and the neighborhoods above the blue lines are under predicted. Note that the standardized residuals are plotted for easier grouping of neighborhoods.

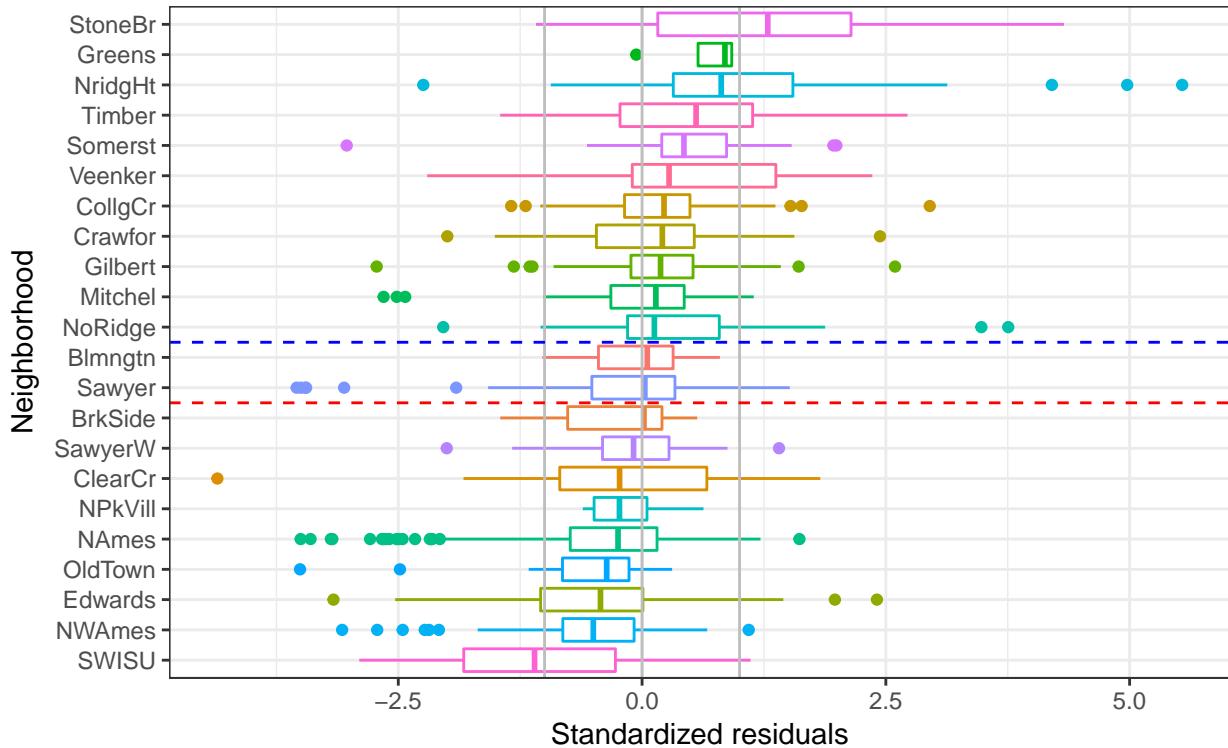


Figure 9: Standardized residuals Vs Neighborhood

6.1 Effect of Neighborhood's mean price per square foot on Mean Absolute Error (MAE)

Figure 10, shows the effect of mean price per square foot on Mean Absolute Error (MAE). It can be seen that the MAE is inflated at the lower end of the spectrum of price per square foot; giving it a non-linear pattern. Therefore it would be useful to include the effect of price per square foot into the model. We also see that the price per square foot is related to the neighborhood. One way to include the effect of price per square foot in the model is to categorize the neighborhoods by price per square foot and use the categories as predictor variables.

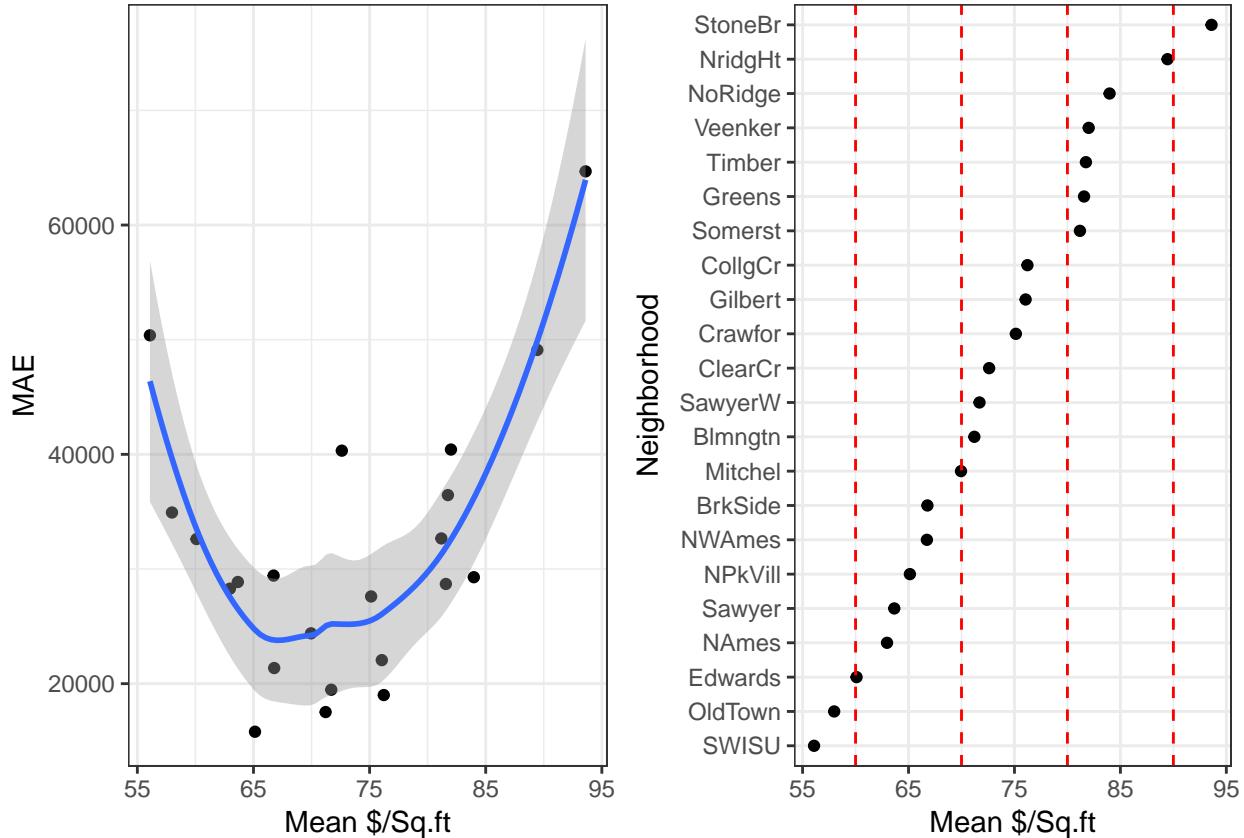


Figure 10: Effect on MAE by Neighborhood (\$/Square foot)

The neighborhoods are grouped into 5 tiers. ***Tier 5 is base category.***

Table 5: Neighborhood classification

Tier	Price.per.sq.ft
1	≤ 60
2	$> 60 \text{ and } \leq 70$
3	$> 70 \text{ and } \leq 80$
4	$> 80 \text{ and } \leq 90$
5	> 90

6.2 Neighborhood categories inclusion as predictors.

The tiers 1 through 4 were included as additional predictors into the multiple linear regression model discussed in section 5. The fit for the model $Sale\ Price = 85820 + 57.4\ TotalBsmtSF + 78.28\ GrLivArea - 1.2e+04\ Tier1 - 9.8e-04\ Tier2 - 9.8e04\ Tier3 - 2.97e04\ Tier4$ is shown below in figure 11. The regression coefficients are statistically significant and so is the overall regression effect with P-value of 0.000.

```
##  
## Call:  
## lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea + Tier1 + Tier2 +  
##       Tier3 + Tier4, data = trainFiltered)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -152766   -16886     -322    16756   236600  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 8.582e+04 8.328e+03 10.306 < 2e-16 ***  
## TotalBsmtSF 5.748e+01 3.016e+00 19.060 < 2e-16 ***  
## GrLivArea   7.828e+01 2.367e+00 33.075 < 2e-16 ***  
## Tier1      -1.229e+05 8.513e+03 -14.434 < 2e-16 ***  
## Tier2      -9.800e+04 6.432e+03 -15.236 < 2e-16 ***  
## Tier3      -7.304e+04 6.398e+03 -11.416 < 2e-16 ***  
## Tier4      -2.974e+04 6.398e+03 -4.648 3.67e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 35430 on 1369 degrees of freedom  
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7905  
## F-statistic: 865.5 on 6 and 1369 DF,  p-value: < 2.2e-16
```

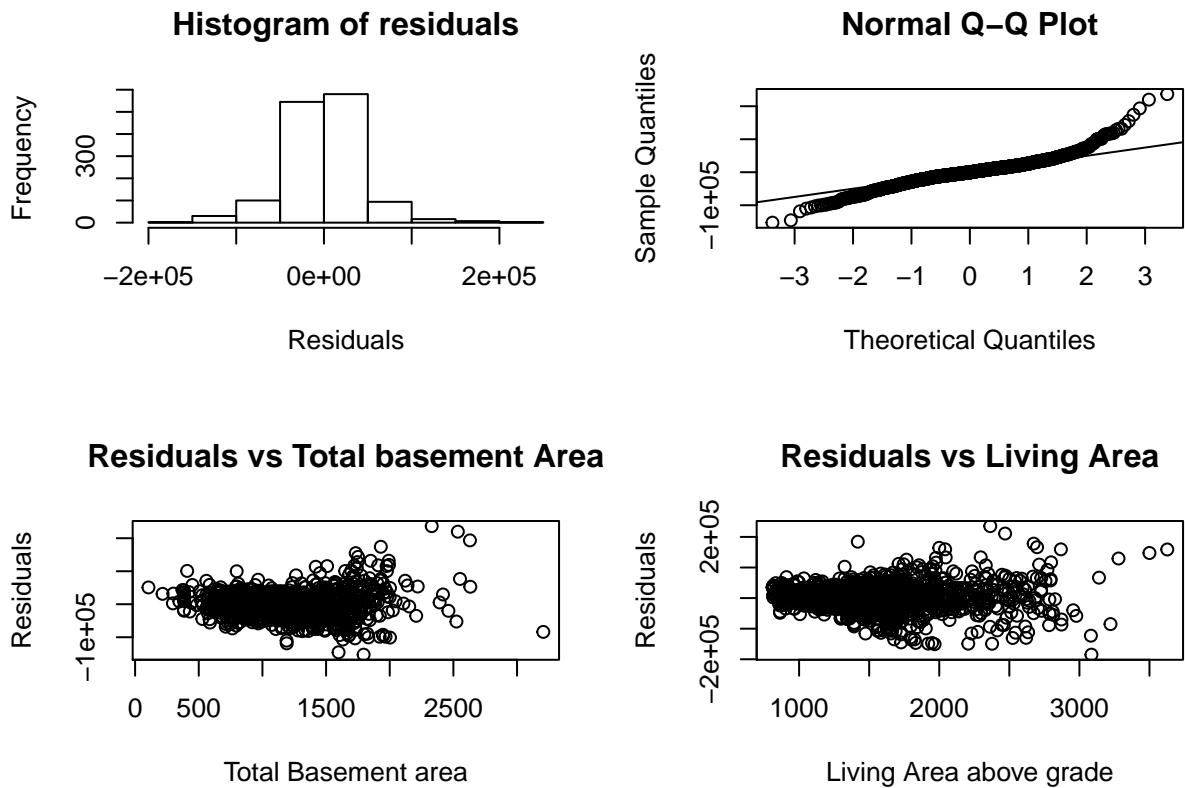


Figure 11: Multiple linear regressions with Neighborhood tiers as predictors

It can be seen from the table below comparing the model in section 5 to the current model, that $\text{Sale Price} = 85820 + 57.4 * \text{TotalBsmtSF} + 78.28 * \text{GrLivArea} - 1.2e+04 * \text{Tier1} - 9.8e-04 * \text{Tier2} - 9.8e04 * \text{Tier3} - 2.97e04 * \text{Tier4}$ is a better model with much lower MAE (24650 vs 29905) and adjusted R squared at 79%. The funnel shape in the residuals is still seen. It would be useful to employ transformations to the response variable. Which we will explore in the next section.

```
##
## -----
##          Model           MAE      MSres   Adj R^2    Fig#_Reference
## -----
## -502244.02 + 84.11 *      29905.22 42635.05     0.7         8
## TotalBsmtSF + 97.608 *
## GrLivArea
##
## 85820 + 57.4 * TotalBsmtSF + 24650.47 35427.61     0.79        11
## 78.28*GrLivArea -1.2e+04*Tier1
## -9.8e-04*Tier2 - 9.8e04*Tier3
## - 2.97e04*Tier4
## -----
## Table: Model comparison
```

7. Model comparison of Sale price and Log(Sale Price)

From the exploratory analysis in section 3.1 and exploratory analysis on categorical variables (refer section 4 of https://github.com/srivathsesh/RegressionAnalysis/blob/master/Assignment1_Seshadri.pdf), the following predictors could be useful for modelling sale price:

1. Above grade living area - Continuous variable
2. Total basement area - Continuous variable
3. Total Bathrooms - Continuous variable
4. Lot Area - Continuous variable
5. Neighborhood Tiers - Discrete. (Neighborhoods grouped into tiers, refer section 6.1)

The below shows the results of multiple linear regression fit for $\text{SalePrice} = 7.17e+04 + 5.681e+01(\text{TotalBsmtSF}) + 6.3e+01(\text{GrLivArea}) + 8.22e-01(\text{LotArea}) + 1.11e+04(\text{TotalBath}) - 1.14e+05(\text{Tier1}) - 9.57e+04(\text{Tier2}) - 7.44e+04(\text{Tier3}) - 2.9e+04(\text{Tier4})$ using the above predictors. All the partial regression coefficients are statistically significant at 95% significance level (P value <<< 0.05). The overall regression effect is also statistically significant with F statistic = 702 and adjusted R squared at 80%.

The residuals continue to show systemic funnel pattern when plotted against basement area, total living area and total rooms (Figure 12). The residuals are not strictly normal and is left skewed. It will be useful to fit the model against a transformed Sales price; i.e.log(Sale Price)

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea + LotArea +
##     TotalBath + Tier1 + Tier2 + Tier3 + Tier4, data = trainFiltered)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -140571 -15895     301   16429  233105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.178e+04 8.261e+03  8.689 < 2e-16 ***
## TotalBsmtSF 5.681e+01 2.953e+00 19.241 < 2e-16 ***
## GrLivArea    6.304e+01 2.840e+00 22.199 < 2e-16 ***
## LotArea      8.212e-01 1.462e-01  5.618 2.34e-08 ***
## TotalBath    1.115e+04 1.424e+03  7.832 9.54e-15 ***
## Tier1       -1.143e+05 8.332e+03 -13.718 < 2e-16 ***
## Tier2       -9.573e+04 6.272e+03 -15.263 < 2e-16 ***
## Tier3       -7.449e+04 6.217e+03 -11.981 < 2e-16 ***
## Tier4       -2.985e+04 6.202e+03 -4.813 1.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34330 on 1367 degrees of freedom
## Multiple R-squared:  0.8044, Adjusted R-squared:  0.8033
## F-statistic: 702.7 on 8 and 1367 DF,  p-value: < 2.2e-16
```

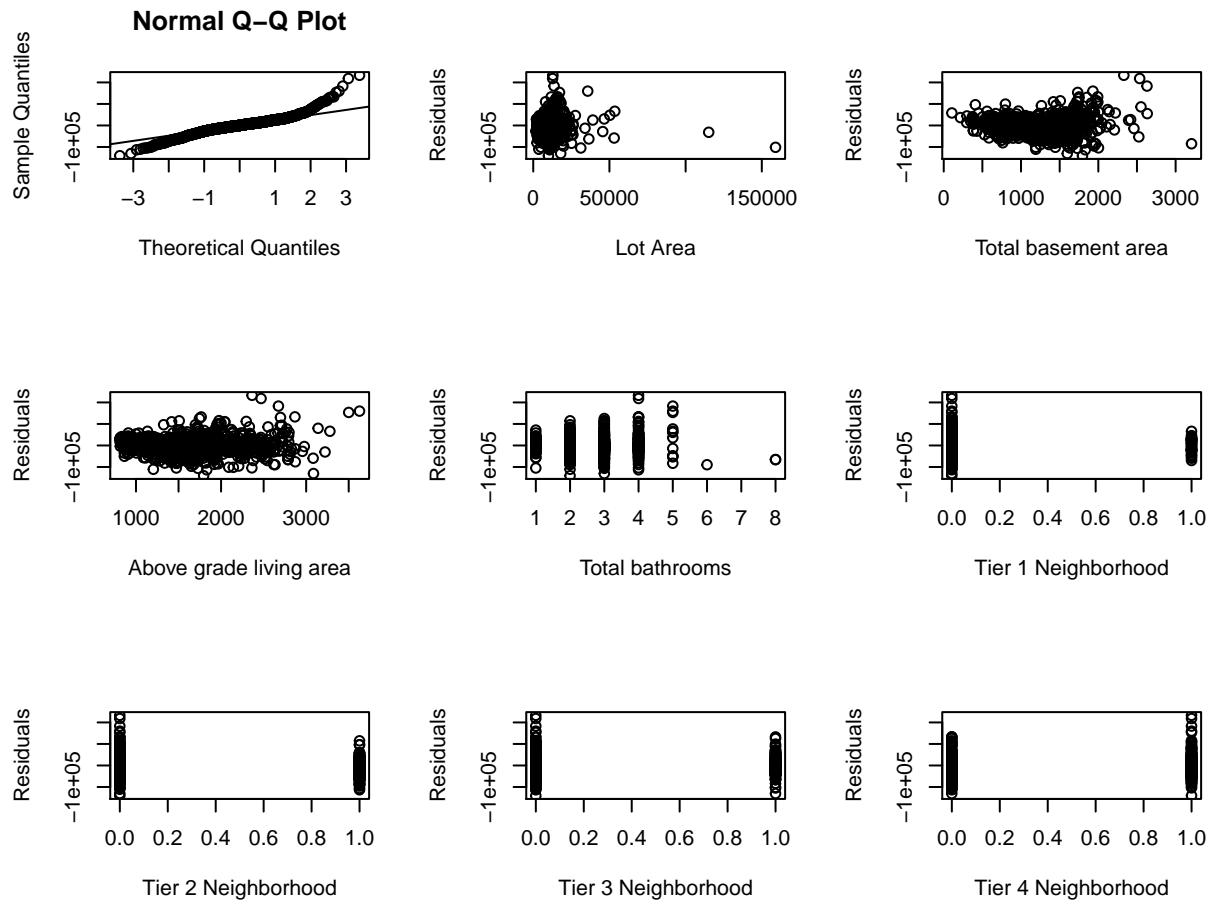


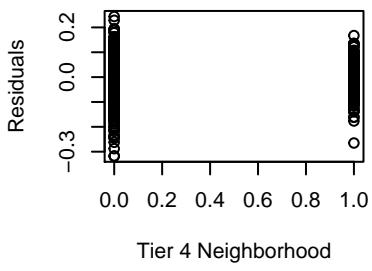
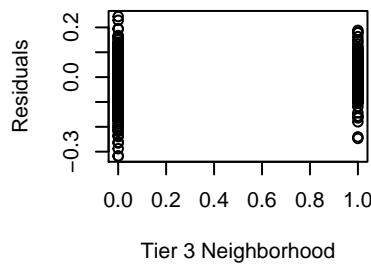
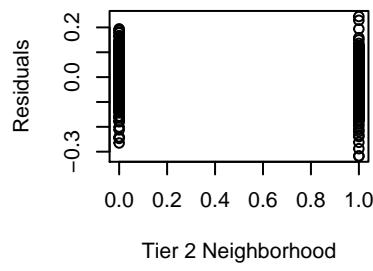
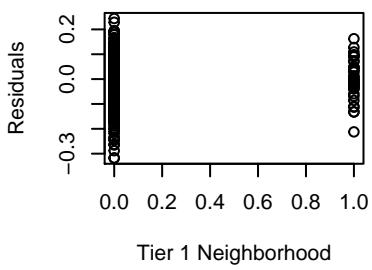
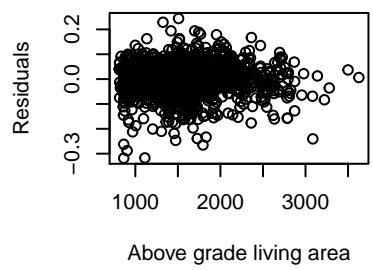
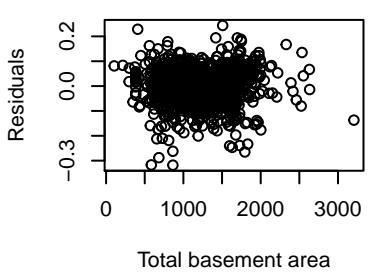
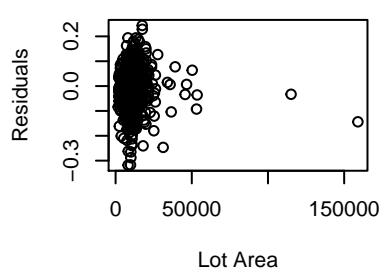
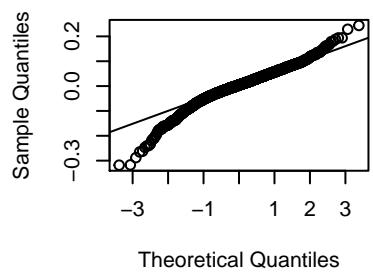
Figure 12: Multiple linear regression of Saleprice with categorical and continuous variables

7.1. Transformed Sale price fit.

A multiple linear regression model of logarithmically transformed Sale price is fit. Below are the results for the model $\text{Log}(Sale\ Price) = 4.979e+00 + 1e-04(\text{TotalBsmtSF}) + 1.236e-04(\text{GrLivArea}) + 1.251e-06(\text{LotArea}) + 2.809e-02(\text{TotalBath}) - 2.160e-01(\text{Tier1}) - 1.585e-01(\text{Tier2}) - 9.391e-02(\text{Tier3}) - 2.740e-02(\text{Tier4})$. The partial regression coefficients are statistically significant at 95% significance, with P values <<< 0.05. The overall regression is significant with F statistic at 796. The Adjusted-R-Squared value is 82%. The residuals are more normal than the non-transformed sale price and the residuals do not have as much as a pattern like it did in the model without transformation of sale price. The residuals are less heteroscedastic. The models are compared in the next section.

```
##
## Call:
## lm(formula = LogSalePrice ~ TotalBsmtSF + GrLivArea + LotArea +
##     TotalBath + Tier1 + Tier2 + Tier3 + Tier4, data = trainFiltered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31821 -0.03063  0.00403  0.03959  0.24379
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.979e+00 1.548e-02 321.708 < 2e-16 ***
## TotalBsmtSF 1.060e-04 5.532e-06 19.164 < 2e-16 ***
## GrLivArea    1.236e-04 5.321e-06 23.221 < 2e-16 ***
## LotArea      1.251e-06 2.739e-07  4.567 5.40e-06 ***
## TotalBath    2.809e-02 2.668e-03 10.530 < 2e-16 ***
## Tier1        -2.160e-01 1.561e-02 -13.835 < 2e-16 ***
## Tier2        -1.585e-01 1.175e-02 -13.488 < 2e-16 ***
## Tier3        -9.391e-02 1.165e-02 -8.063 1.62e-15 ***
## Tier4        -2.740e-02 1.162e-02 -2.358  0.0185 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06432 on 1367 degrees of freedom
## Multiple R-squared:  0.8232, Adjusted R-squared:  0.8222
## F-statistic: 795.6 on 8 and 1367 DF,  p-value: < 2.2e-16
```

Normal Q-Q Plot



7.1.1 Transformed model interpretation

The regression model $\text{Log}(Sale\ Price) = 4.979e+00 + 1e-04(\text{TotalBsmtSF}) + 1.236e-04(\text{GrLivArea}) + 1.251e-06(\text{LotArea}) + 2.809e-02(\text{TotalBath}) - 2.160e-01(\text{Tier1}) - 1.585e-01(\text{Tier2}) - 9.391e-02(\text{Tier3}) - 2.740e-02(\text{Tier4})$ is a model for predicting the *logarithmic* value of sales price. The predicted value is in logarithmic scale of dollars. The exponent or antilog of the fitted values need to be calculated for the actual prediction or fitted value.

7.2 Model comparison of Multiple linear regression models of Sale Price and Log(Sale price)

The below table shows the multiple linear models fitted in the above sections 7 and 7.1. It is seen that the log transformation of sale price yields a better model with Mean Absolute Error (MAE) of \$21,524 and 82% adjusted R squared. The residuals are relatively more random in the log transformed model ($\text{Log}(Sale\ Price) = 4.979e+00 + 1e-04(\text{TotalBsmtSF}) + 1.236e-04(\text{GrLivArea}) + 1.251e-06(\text{LotArea}) + 2.809e-02(\text{TotalBath}) - 2.160e-01(\text{Tier1}) - 1.585e-01(\text{Tier2}) - 9.391e-02(\text{Tier3}) - 2.740e-02(\text{Tier4})$).

```
## 
## -----
##          Model      MAE   Adj.R^2   F.Stat   P.Value   Fig# reference
## -----
##     SalePrice = 7.17e+04 +    23829  0.8033    702.7      0        12
##     5.681e+01*TotalBsmtSF +
##     6.3e+01*GrLivArea +
##     8.22e-01*LotArea +
##     1.11e+04*TotalBath -
##     1.14e+05*Tier1 -
##     9.57e+04*Tier2 -
##     7.44e+04*Tier3 - 2.9e+04*
##     Tier4
## 
## Log(Sale Price) = 4.979e+00 + 21524  0.8222    795.6      0        13
##     1e-04*TotalBsmtSF +
##     1.236e-04*GrLivArea +
##     1.251e-06*LotArea +
##     2.809e-02*TotalBath -
##     2.160e-01*Tier1 -
##     1.585e-01*Tier2 -
##     9.391e-02*Tier3 -
##     2.740e-02*Tier4
## 
## 
```

Table: Comparison of MLR models of Sale Price and Log(Sale Price)

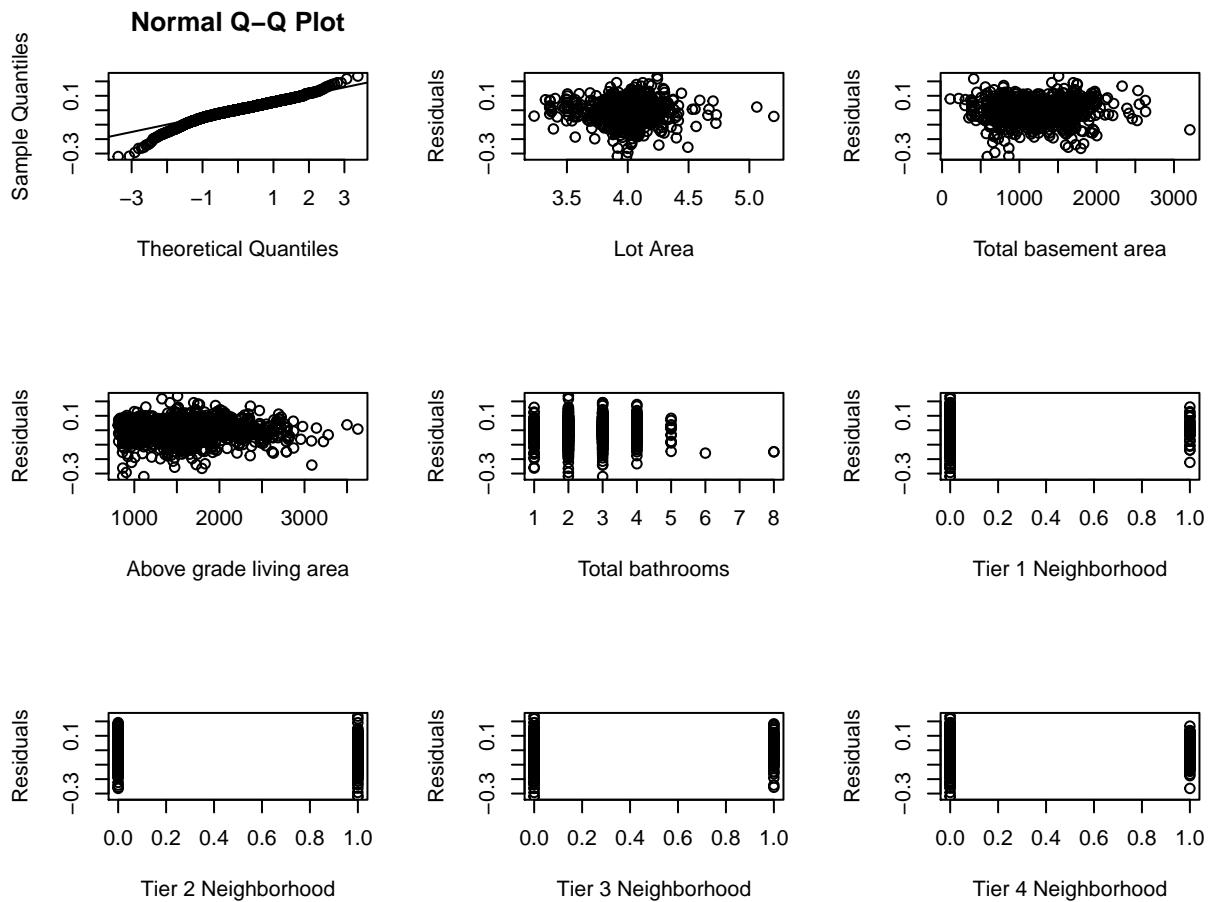
The transformation of response variable is useful when there is explicit patterns or non-randomness in the residuals. Transformation reduces heteroscedasticity. Which is evident in the model performance results below. However, we do see outliers in the predictor “Lot Area”. It’ll be useful to log transform lot area and fit the model again.

7.3. Transformation of Lot Area (Predictor)

To address the skewness in Lot Area, the lot area was logarithmically transformed and a multiple linear regression model was fit. Below are the results of the model $\text{Log}(Sale\ Price) = 4.716e+00 + \text{TotalBsmtSF}(1.037e-04) + \text{GrLivArea}(1.181e-04) + \log(\text{LotArea})(7.309e-02) + \text{TotalBath}(2.847e-02) - \text{Tier1}(2.180e-01) - \text{Tier2}(1.654e-01) - \text{Tier3}(9.951e-02) - \text{Tier4}(3.200e-02)$. It is seen that the partial regression coefficients are statistically significant at 95% significance and so is regression effect with F statistic at 812. The residuals are skewed left, but is fairly normal and they show a funnel shape behavior with Lot area.

The model performance is compared with other models from sections 7, 7.1 and 7.2 below.

```
##  
## Call:  
## lm(formula = LogSalePrice ~ TotalBsmtSF + GrLivArea + logLotArea +  
##       TotalBath + Tier1 + Tier2 + Tier3 + Tier4, data = trainFiltered)  
##  
## Residuals:  
##      Min        1Q     Median        3Q       Max  
## -0.32051 -0.03098  0.00388  0.03851  0.23627  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.716e+00 4.214e-02 111.919 < 2e-16 ***  
## TotalBsmtSF 1.037e-04 5.501e-06 18.854 < 2e-16 ***  
## GrLivArea   1.181e-04 5.391e-06 21.908 < 2e-16 ***  
## logLotArea  7.309e-02 1.088e-02  6.715 2.75e-11 ***  
## TotalBath    2.847e-02 2.646e-03 10.758 < 2e-16 ***  
## Tier1      -2.180e-01 1.548e-02 -14.081 < 2e-16 ***  
## Tier2      -1.654e-01 1.173e-02 -14.105 < 2e-16 ***  
## Tier3      -9.951e-02 1.160e-02 -8.576 < 2e-16 ***  
## Tier4      -3.200e-02 1.155e-02 -2.771  0.00567 **  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.06376 on 1367 degrees of freedom  
## Multiple R-squared:  0.8262, Adjusted R-squared:  0.8252  
## F-statistic: 812.5 on 8 and 1367 DF,  p-value: < 2.2e-16
```



```

## -----
##          Model      MAE  Adj.R^2   F.Stat  P.Value Fig# Reference
## -----
## SalePrice = 7.17e+04 +    23829  0.8033    702.7     0       12
## 5.681e+01*TotalBsmtSF +
## 6.3e+01*GrLivArea +
## 8.22e-01*LotArea +
## 1.11e+04*TotalBath -
## 1.14e+05*Tier1 -
## 9.57e+04*Tier2 -
## 7.44e+04*Tier3 - 2.9e+04*
##                  Tier4
##
## Log(Sale Price) = 4.979e+00 + 21524  0.8222    795.6     0       13
## 1e-04*TotalBsmtSF +
## 1.236e-04*GrLivArea +
## 1.251e-06*LotArea +
## 2.809e-02*TotalBath -
## 2.160e-01*Tier1 -
## 1.585e-01*Tier2 -
## 9.391e-02*Tier3 -
## 2.740e-02*Tier4
##
## Log(Sale Price) = 4.716e+00 + 21254  0.8252    812.5     0       14
## TotalBsmtSF*1.037e-04 +
## GrLivArea*1.181e-04 +
## log(LotArea)*7.309e-02 +
## TotalBath*2.847e-02 -
## Tier1*2.180e-01 -
## Tier2*1.654e-01 -
## Tier3*9.951e-02 -
## Tier4*3.200e-02
## -----
## 
```

Table: Comparison of MLR models of Sale Price and Log(Sale Price)

It is seen that the transformation of lot area does not provide us with significant advantage and performs more like the non-transformed predictor model.

8. Conclusion

Several Multiple linear regression models were fit for predicting sales price of typical homes in Ames Iowa. The models' goodness of fit was evaluated by rigorously checking the assumptions of the model by evaluating the residuals. Transformations of sale price were made to arrive at a better model. $\text{Log}(\text{Sale Price}) = 4.979e+00 + 1e-04(\text{TotalBsmtSF}) + 1.236e-04(\text{GrLivArea}) + 1.251e-06(\text{LotArea}) + 2.809e-02(\text{TotalBath}) - 2.160e-01(\text{Tier1}) - 1.585e-01(\text{Tier2}) - 9.391e-02(\text{Tier3}) - 2.740e-02(\text{Tier4})$ was deemed a better model with Mean Absolute Error (MAE) of \$21,524. The model has a adjusted R squared value of 82%.

APPENDIX

A.1 Data quality check

Tables below shows the summary statistics of the numeric variables and it is noted that statistics are within reasonable bounds and appear to be in the units of measure as described in the data dictionary with only 2 rows missing. Also shown are the number of levels or categories in the nominal variables and the number of missing data (0 missing). The data is deemed usable.

Table 6: Data sanity check for numeric variables

	min	Q1	median	Q3	max	mean	sd	n	missing
SID	1	675.75	1467.5	2171.25	2930	1e+03	8e+02	1992	0
LotArea	1700	8400.00	10030.5	12227.25	215245	1e+04	8e+03	1992	0
OverallCond	1	5.00	5.0	6.00	9	6e+00	1e+00	1992	0
YearRemodel	1950	1969.75	1995.0	2004.00	2010	2e+03	2e+01	1992	0
TotalBsmtSF	105	864.00	1076.0	1393.25	6110	1e+03	4e+02	1992	0
GrLivArea	808	1196.00	1499.5	1800.25	5642	2e+03	5e+02	1992	0
BsmtFullBath	0	0.00	0.0	1.00	2	5e-01	5e-01	1992	0
BsmtHalfBath	0	0.00	0.0	0.00	2	6e-02	3e-01	1992	0
FullBath	0	1.00	2.0	2.00	4	2e+00	5e-01	1992	0
HalfBath	0	0.00	0.0	1.00	2	4e-01	5e-01	1992	0
BedroomAbvGr	0	3.00	3.0	3.00	6	3e+00	8e-01	1992	0
TotRmsAbvGrd	3	6.00	6.0	7.00	15	7e+00	1e+00	1992	0
GarageArea	0	390.00	490.0	605.25	1488	5e+02	2e+02	1992	0
MoSold	1	4.00	6.0	8.00	12	6e+00	3e+00	1992	0
YrSold	2006	2007.00	2008.0	2009.00	2010	2e+03	1e+00	1992	0
SalePrice	58500	143000.00	177547.0	230000.00	755000	2e+05	8e+04	1992	0
FirstFlrSF	453	936.00	1152.5	1470.50	5095	1e+03	4e+02	1992	0
SecondFlrSF	0	0.00	0.0	728.00	1872	3e+02	4e+02	1992	0
TotalBath	1	2.00	3.0	3.00	8	3e+00	9e-01	1992	0
TotalSQFT	1015	2150.00	2590.5	3146.00	11752	3e+03	8e+02	1992	0
SQFTNoBsmt	808	1193.00	1496.0	1800.00	5642	2e+03	5e+02	1992	0

Table 7: Data sanity check for nominal variables

	# Unique	n	missing
LotConfig	5	1992	0
Neighborhood	22	1992	0
BldgType	5	1992	0
HouseStyle	8	1992	0
KitchenQual	5	1992	0
SaleCondition	2	1992	0

A.2 Nested Model

In section 5.1 we compared the simple linear and multiple linear regression models. Here we use a Nested models approach to test if the additional predictor is statistically significant to add value to the model. It is hypothesized that the regression coefficient of the additional parameter in the model is zero. Thereby, hypothesizing that the additional parameter has no significant contribution towards explaining the variation in the response variable.

The ANOVA table below compares the Reduced Model(RM) SalePrice ~ GrLivArea to the Full Model(FM) SalePrice ~ TotalBsmtSF + GrLivArea. It is seen that the regression coefficient of total basement is statistically significant at 95% significance level. Hence the FM is warranted.

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ GrLivArea + 0
## Model 2: SalePrice ~ TotalBsmtSF + GrLivArea
##   Res.Df      RSS Df  Sum of Sq    F    Pr(>F)
## 1    1375 3.6733e+12
## 2    1373 2.4958e+12  2 1.1775e+12 323.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Further, RM of SalePrice ~ TotalBsmtSF is tested against the FM of SalePrice ~ TotalBsmtSF + GrLivArea. In the below ANOVA table it is seen that the regression coefficient of the additional parameter in the FM is statistically significant at 95% significance level. Also it is seen in Figure 8 that the residuals are more random in the multiple linear regression model than that for simple linear regression models (Figure 6 and Figure 7). Hence the FM is deemed better than the RM.

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ TotalBsmtSF
## Model 2: SalePrice ~ TotalBsmtSF + GrLivArea
##   Res.Df      RSS Df  Sum of Sq    F    Pr(>F)
## 1    1374 4.9850e+12
## 2    1373 2.4958e+12  1 2.4892e+12 1369.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.3 R code

```

knitr::opts_chunk$set(echo = TRUE, tidy.opts = list(width.cutoff = 60),
tidy = TRUE)
ames <- readr::read_delim(file = "ames_housing_data.csv", delim = ",")
# change from scientific notations, to restore to default
# options(scipen = 0)
options(scipen = 999)
library(magrittr)
LivingAreaCutoff <- 800
# Adding drop conditions variable insert dummy variable to
# code SaleCondition being either equal to 'Normal' or
# 'Partial'
ames$Sale_NrmPar <- ifelse(ames$SaleCondition == "Normal" | ames$SaleCondition ==
    "Partial", 1, 0)
ames$DropCondition <- ifelse(ames$Zoning != "RL", "01: Not LowDensityZone",
    ifelse(ames$Sale_NrmPar == 0, "02: Not Normal/Partial Sale",
        ifelse(ames$Street != "Pave", "03: Street Not Paved",
            ifelse(ames$GrLivArea < LivingAreaCutoff, "04: Less than 800 SqFt",
                ifelse(ames$TotalBsmtSF < 1, "05: No Basement",
                    "99: Eligible Sample"))))

# Waterfall
waterfall <- ames %>% dplyr::group_by(DropCondition) %>% dplyr::summarise(counts = n())

# Print waterfall table
knitr::kable(waterfall, align = c("l", "r"), caption = "Drop waterfall")
# Define training portion of the data
trainPercent <- round(0.7, 1)
# Columns of interest
colsofinterest <- c("SID", "LotArea", "LotConfig", "Neighborhood",
    "BldgType", "HouseStyle", "OverallCond", "YearRemodel", "TotalBsmtSF",
    "GrLivArea", "BsmtFullBath", "BsmtHalfBath", "FullBath",
    "HalfBath", "BedroomAbvGr", "KitchenQual", "TotRmsAbvGrd",
    "GarageArea", "MoSold", "YrSold", "SaleCondition", "SalePrice",
    "FirstFlrSF", "SecondFlrSF")

# Cleanly show the columns of interest in pdf. Making the
# colsofinterest as matrix for easy printing.
colsmatrix <- matrix(colsofinterest[2:length(colsofinterest)],
    ncol = 3)
# printing on pdf
knitr::kable(colsmatrix, caption = "Variables of interest")
# Get sample frame.
SampleFrame <- ames %>% dplyr::filter(DropCondition == "99: Eligible Sample") %>%
    dplyr::select_(.dots = colsofinterest)
SampleFrame <- SampleFrame %>% dplyr::mutate(TotalBath = BsmtFullBath +
    BsmtHalfBath + FullBath + HalfBath) %>% dplyr::mutate(TotalSQFT = TotalBsmtSF +
    FirstFlrSF + SecondFlrSF) %>% dplyr::mutate(SQFTNoBsmt = FirstFlrSF +
    SecondFlrSF)
# training set
train <- dplyr::sample_n(SampleFrame, size = trainPercent * nrow(SampleFrame),
    replace = F, set.seed(2000))

```

```

train <- train %>% dplyr::arrange(SID)
# Validation set
Validation <- dplyr::sample_n(SampleFrame, size = (1 - trainPercent) *
  nrow(SampleFrame), replace = F, set.seed(2000))
Validation <- Validation %>% dplyr::arrange(SID)
# Check row counts
df <- cbind(Data = c("Training set", "Validation set"), Samples = c(nrow(train),
  nrow(Validation)))
knitr::kable(df, align = c("l", "r"), caption = "Training and Validation sampling")
# Multicollinearity exploration
test <- SampleFrame[, c("SalePrice", "TotalBsmtSF", "GarageArea",
  "GrLivArea", "LotArea", "TotalBath")]
pairs(test)
# Plot of basement area vs sale price
library(ggplot2)
library(gridExtra)
BasementArea <- ggplot(data = SampleFrame, mapping = aes(x = TotalBsmtSF,
  y = SalePrice)) + geom_point() + geom_smooth(method = "lm",
  se = T) + xlab("Total basement area") + theme_bw()
# annotation addition
BasementArea <- BasementArea + annotate("text", x = 2000, y = 7e+05,
  label = paste0("correlation = ", round(cor(SampleFrame$TotalBsmtSF,
    SampleFrame$SalePrice), 2)))
# Restrict basement to < 2500
RestrictedBsmtSF <- SampleFrame %>% dplyr::filter(TotalBsmtSF >
  0 & TotalBsmtSF < 2500)
# Basement < 2500 vs SalePrice
RestrictedBasement <- ggplot(data = RestrictedBsmtSF, mapping = aes(x = TotalBsmtSF,
  y = SalePrice)) + geom_point() + geom_smooth(method = "lm",
  se = T) + xlab("Total basement area") + theme_bw()
# Annotations
RestrictedBasement <- RestrictedBasement + annotate("text", x = 1500,
  y = 7e+05, label = paste0("correlation = ", round(cor(SampleFrame$TotalBsmtSF,
    SampleFrame$SalePrice), 2)))
# Print plot
grid.arrange(BasementArea, RestrictedBasement, ncol = 2)
LivingArea <- ggplot(SampleFrame) + geom_point(mapping = aes(x = GrLivArea,
  y = SalePrice)) + xlab("Living Area") + geom_smooth(mapping = aes(x = GrLivArea,
  y = SalePrice), se = T) + theme_bw()

GarageArea <- ggplot(SampleFrame) + geom_point(mapping = aes(x = GarageArea,
  y = SalePrice)) + xlab("Garage Area") + geom_smooth(mapping = aes(x = GarageArea,
  y = SalePrice), se = T) + theme_bw()

RestrictedLivingGarageArea <- SampleFrame %>% dplyr::filter(GrLivArea <
  4000 & GarageArea > 0 & GarageArea < 1000)

LivingAreaRestricted <- ggplot(RestrictedLivingGarageArea) +
  geom_point(mapping = aes(x = GrLivArea, y = SalePrice)) +
  xlab("Living Area") + geom_smooth(mapping = aes(x = GrLivArea,
  y = SalePrice), se = T) + theme_bw()

GarageAreaRestricted <- ggplot(RestrictedLivingGarageArea) + geom_point(mapping = aes(x = GarageArea,

```

```

y = SalePrice)) + xlab("Garage Area") + geom_smooth(mapping = aes(x = GarageArea,
y = SalePrice), se = T) + theme_bw()

grid.arrange(LivingArea, GargeArea, LivingAreaRestricted, GargeAreaRestricted,
ncol = 2)

RestGarageVsLivArea <- ggplot(RestrictedLivingGarageArea) + geom_point(mapping = aes(x = GarageArea,
y = GrLivArea)) + xlab("Garage Area") + geom_smooth(mapping = aes(x = GarageArea,
y = GrLivArea), se = T) + theme_bw()

RestGarageVsLivArea <- RestGarageVsLivArea + annotate("text",
x = 250, y = 3000, label = paste0("Correlation = ", round(cor(RestrictedLivingGarageArea$GarageArea,
RestrictedLivingGarageArea$GrLivArea), 2)))

RestGarageVsBsmt <- ggplot(RestrictedLivingGarageArea) + geom_point(mapping = aes(x = GarageArea,
y = TotalBsmtSF)) + xlab("Garage Area") + geom_smooth(mapping = aes(x = GarageArea,
y = TotalBsmtSF), se = T) + theme_bw()

RestGarageVsBsmt <- RestGarageVsBsmt + annotate("text", x = 250,
y = 3000, label = paste0("Correlation = ", round(cor(RestrictedLivingGarageArea$GarageArea,
RestrictedLivingGarageArea$TotalBsmtSF), 2)))

grid.arrange(RestGarageVsLivArea, RestGarageVsBsmt, nrow = 2)

options(scipen = 0)
# Filtering Living Area < 4000 and Garage area < 1000 and
# creating additional variable of LogSalePrice in preparation
# of the next section.
trainFiltered <- train %>% dplyr::filter(GrLivArea < 4000 & GarageArea <
1000) %>% dplyr::mutate(LogSalePrice = log10(SalePrice))

SLR_LivingArea <- lm(data = trainFiltered, SalePrice ~ GrLivArea)
print(summary(SLR_LivingArea), caption = "ANOVA Simple Linear Regression Above grade living area")

# Model diagnostics Tidy store of model results
RedDf <- broom::augment(SLR_LivingArea)
# Plot of residuals
layout(matrix(c(1, 2, 3, 3), 2, 2, byrow = TRUE))
hist(RedDf$.resid, main = "Histogram of residuals", xlab = "Residuals")
qqnorm(RedDf$.resid, title = "Normal Q-Q plot of residuals (Sale Price)")
qqline(RedDf$.resid)

plot(RedDf$GrLivArea, RedDf$.resid, main = "Residuals vs Living Area",
xlab = "Above grade living area", ylab = "Residuals")

# No intercept model
SLR_LivingArea_NoIntercept <- lm(data = trainFiltered, SalePrice ~
GrLivArea + 0)
# print model
summary(SLR_LivingArea_NoIntercept)

```

```

# Model diagnostics
ResdfNI <- broom::augment(SLR_LivingArea_NoIntercept)
FtestsSLR_GrLivArea <- broom::glance(SLR_LivingArea_NoIntercept)
# Plot of residuals
layout(matrix(c(1, 2, 3, 3), 2, 2, byrow = TRUE))
hist(ResdfNI$.resid, main = "Histogram of residuals", xlab = "Residuals")
qqnorm(ResdfNI$.resid, title = "Normal Q-Q plot of residuals (Sale Price)")
qqline(ResdfNI$.resid)
plot(ResdfNI$GrLivArea, ResdfNI$.resid, main = "Residuals vs Living Area",
     xlab = "Above grade living area", ylab = "Residuals")

SLR_BsmtArea <- lm(data = trainFiltered, SalePrice ~ TotalBsmtSF)
print(summary(SLR_BsmtArea), caption = "ANOVA Simple Linear Regression Total basement area")

# Model diagnostics Tidy store of model results
ResdfBsmt <- broom::augment(SLR_BsmtArea)
FTestsSLR_BsmtArea <- broom::glance(SLR_BsmtArea)
# Plot of residuals
layout(matrix(c(1, 2, 3, 3), 2, 2, byrow = TRUE))
hist(ResdfBsmt$.resid, main = "Histogram of residuals", xlab = "Residuals")
qqnorm(ResdfBsmt$.resid, title = "Normal Q-Q plot of residuals (Sale Price)")
qqline(ResdfBsmt$.resid)
plot(ResdfBsmt$TotalBsmtSF, ResdfBsmt$.resid, main = "Residuals vs Total basement Area",
     xlab = "Total Basement area", ylab = "Residuals")

MLR <- lm(data = trainFiltered, SalePrice ~ TotalBsmtSF + GrLivArea)
print(summary(MLR), caption = "ANOVA Multiple Linear Regression Total basement area")

# Model diagnostics Tidy store of model results
MLR_Model <- broom::augment(MLR)
FtestsMLR <- broom::glance(MLR)
# Plot of residuals
layout(matrix(c(1, 2, 3, 4), 2, 2, byrow = TRUE))
hist(MLR_Model$.resid, main = "Histogram of residuals", xlab = "Residuals")
qqnorm(MLR_Model$.resid, title = "Normal Q-Q plot of residuals (Sale Price)")
qqline(MLR_Model$.resid)
plot(MLR_Model$TotalBsmtSF, MLR_Model$.resid, main = "Residuals vs Total basement Area",
     xlab = "Total Basement area", ylab = "Residuals")
plot(MLR_Model$GrLivArea, MLR_Model$.resid, main = "Residuals vs Living Area",
     xlab = "Living Area above grade", ylab = "Residuals")
## Create table for comparing models
Models <- c("SalePrice = 126.9 * GrLivArea", "SalePrice = 47859.31 + 129.96 * TotalBsmtSF",
           "-502244.02 + 84.11 * TotalBsmtSF + 97.608 * GrLivArea")
MSres <- c(round(FtestsSLR_GrLivArea$sigma, 2), round(FTestsSLR_BsmtArea$sigma,
            2), round(FtestsMLR$sigma, 2))
F_Stat <- c(round(FtestsSLR_GrLivArea$statistic, 2), round(FTestsSLR_BsmtArea$statistic,
              2), round(FtestsMLR$statistic, 2))
P_Value <- c(round(FtestsSLR_GrLivArea$p.value, 2), round(FTestsSLR_BsmtArea$p.value,
               2), round(FtestsMLR$p.value, 2))
Adj_R_Squared <- c(round(FtestsSLR_GrLivArea$adj.r.squared, 2),
                    round(FTestsSLR_BsmtArea$adj.r.squared, 2), round(FtestsMLR$adj.r.squared,
                     2))
Normality <- c("Satisfied thick pen test; heavy tails", "Satisfied thick pen test; heavy tails",

```

```

    "Satisfied thick pen test; heavy tails")
ResidualPattern <- c("Funnel shaped", "Funnel shaped", "Funnel shaped")
FigureRef <- c(6, 7, 8)
# Comparisonondf <-
# cbind(Models,MSres,F_Stat,P_Value,Adj_R_Squared,Normality,ResidualPattern)
# colnames(Comparisonondf) <- c('Model','MSres','F Stat','P
# Value', 'Adj R^2', 'Residuals Normality', 'Residuals
# Pattern')
Comparisonondf <- cbind(Models, MSres, F_Stat, P_Value, Adj_R_Squared,
FigureRef)
colnames(Comparisonondf) <- c("Model", "MSres", "F Stat", "P Value",
"Adj R^2", "Fig#_Reference")
# print table
knitr::kable(Comparisonondf, caption = "Model comparison")
# pander::pander(Comparisonondf)
library(tidyverse)
library(forcats)
# prepare dataframe with residuals
MLR_Model$Neighborhood <- trainFiltered$Neighborhood
MLR_Model$TotRmsAbvGrd <- trainFiltered$TotRmsAbvGrd
MLR_Model$KitchenQual <- trainFiltered$KitchenQual
MLR_Model$FirstFlrSF <- trainFiltered$FirstFlrSF
MLR_Model$SecondFlrSF <- trainFiltered$SecondFlrSF
MLR_Model$TotalSQFT <- trainFiltered$TotalSQFT
MLR_Model$SQFTNoBsmt <- trainFiltered$SQFTNoBsmt
# Boxplot of residuals by neighborhood - DO NOT KNOW HOW TO
# CREATE LEGEND FOR DASHED LINES
p <- ggplot(MLR_Model, mapping = aes(x = fct_reorder(Neighborhood,
.std.resid), y = .std.resid, color = Neighborhood, guides(color = F)))
residboxplot <- p + geom_boxplot() + coord_flip() + theme_bw() +
scale_colour_discrete(guide = FALSE)
residboxplot + geom_hline(yintercept = c(-1, 0, 1), linetype = "solid",
color = "grey") + geom_vline(xintercept = c(9.5, 11.5), linetype = "dashed",
color = c("red", "blue")) + ylab("Standardized residuals") +
xlab("Neighborhood")

# MAE by neighborhood
MAE_Neighborhood <- MLR_Model %>% group_by(Neighborhood) %>%
summarise(MAE = mean(abs(.resid)), TotalSalePrice = sum(SalePrice),
TotalLivArea = sum(GrLivArea), Tot.sqft = sum(TotalSQFT)) %>%
mutate(MeanPricePerSqft = TotalSalePrice/TotalLivArea) %>%
mutate(MeanPricePerSqft2 = TotalSalePrice/Tot.sqft)

p1 <- ggplot(data = MAE_Neighborhood, mapping = aes(x = MeanPricePerSqft2,
y = MAE)) + geom_point() + theme_bw() + xlab("Mean $/Sq.ft") +
geom_smooth()
p2 <- ggplot(data = MAE_Neighborhood, mapping = aes(x = fct_reorder(Neighborhood,
MeanPricePerSqft2), y = MeanPricePerSqft2)) + geom_point() +
coord_flip() + theme_bw() + geom_hline(yintercept = c(60,
70, 80, 90), linetype = "dashed", color = "red") + xlab("Neighborhood") +
ylab("Mean $/Sq.ft")
grid.arrange(p1, p2, ncol = 2)

```

```

# Grouping Neighborhoods
Less60Neigh <- MAE_Neighborhood$Neighborhood[which(MAE_Neighborhood$MeanPricePerSqft2 <=
60)]
Bet60_70 <- MAE_Neighborhood$Neighborhood[which(MAE_Neighborhood$MeanPricePerSqft2 <=
70 & MAE_Neighborhood$MeanPricePerSqft2 > 60)]
Bet70_80 <- MAE_Neighborhood$Neighborhood[which(MAE_Neighborhood$MeanPricePerSqft2 <=
80 & MAE_Neighborhood$MeanPricePerSqft2 > 70)]
Bet80_90 <- MAE_Neighborhood$Neighborhood[which(MAE_Neighborhood$MeanPricePerSqft2 <=
90 & MAE_Neighborhood$MeanPricePerSqft2 > 80)]
Greater_90 <- MAE_Neighborhood$Neighborhood[which(MAE_Neighborhood$MeanPricePerSqft2 >
90)]

trainFiltered$Tier1 <- ifelse(trainFiltered$Neighborhood %in%
  Less60Neigh, 1, 0)
trainFiltered$Tier2 <- ifelse(trainFiltered$Neighborhood %in%
  Bet60_70, 1, 0)
trainFiltered$Tier3 <- ifelse(trainFiltered$Neighborhood %in%
  Bet70_80, 1, 0)
trainFiltered$Tier4 <- ifelse(trainFiltered$Neighborhood %in%
  Bet80_90, 1, 0)
trainFiltered$Tier5 <- ifelse(trainFiltered$Neighborhood %in%
  Greater_90, 1, 0)

tierdf <- data.frame(Tier = c(1, 2, 3, 4, 5), `Price per sq.ft` = c("<= 60",
  "> 60 and <= 70", "> 70 and <= 80", "> 80 and <= 90", "> 90"))
MLR_Neighborhoods <- lm(data = trainFiltered, SalePrice ~ TotalBsmtSF +
  GrLivArea + Tier1 + Tier2 + Tier3 + Tier4)
summary(MLR_Neighborhoods)
MLR_Neighborhoods_data <- broom::augment(MLR_Neighborhoods)
FtestsMLR_neighborhoods <- broom::glance(MLR_Neighborhoods)
# Plot of residuals
layout(matrix(c(1, 2, 3, 4), 2, 2, byrow = TRUE))
hist(MLR_Model$.resid, main = "Histogram of residuals", xlab = "Residuals")
qqnorm(MLR_Neighborhoods_data$.resid, title = "Normal Q-Q plot of residuals (Sale Price)")
qqline(MLR_Neighborhoods_data$.resid)
plot(MLR_Neighborhoods_data$TotalBsmtSF, MLR_Neighborhoods_data$.resid,
  main = "Residuals vs Total basement Area", xlab = "Total Basement area",
  ylab = "Residuals")
plot(MLR_Model$GrLivArea, MLR_Model$.resid, main = "Residuals vs Living Area",
  xlab = "Living Area above grade", ylab = "Residuals")
# Model comparison
Models_Neighborhood <- c("-502244.02 + 84.11 * TotalBsmtSF + 97.608 * GrLivArea",
  "85820 + 57.4 * TotalBsmtSF + 78.28*GrLivArea -1.2e+04*Tier1 -9.8e-04*Tier2 - 9.8e04*Tier3 - 2.97e04*Tier4")
MAE_comparison <- c(round(mean(abs(MLR_Model$.resid)), 2), round(mean(abs(MLR_Neighborhoods_data$.resid)),
  2))
MSres_Neighborhood <- c(round(FtestsMLR$sigma, 2), round(FtestsMLR_neighborhoods$sigma,
  2))
Adj_R_Squared_Neighborhood <- c(round(FtestsMLR$adj.r.squared,
  2), round(FtestsMLR_neighborhoods$adj.r.squared, 2))
MAEFigRef <- c(8, 11)
Comparisondf_Neighborhood <- cbind(Models_Neighborhood, MAE_comparison,
  MSres_Neighborhood, Adj_R_Squared_Neighborhood, MAEFigRef)
colnames(Comparisondf_Neighborhood) <- c("Model", "MAE", "MSres",

```

```

    "Adj R^2", "Fig#_Reference")
# print table
pander::pandoc.table(Comparisondf_Neighborhood, caption = "Model comparison")
# trainFiltered$KitchenQual <-
# as.factor(trainFiltered$KitchenQual)
# trainFiltered$NeighRecode <-
# as.integer(as.factor(trainFiltered$Neighborhood))
# trainFiltered$RecodeKQ <-
# as.integer(trainFiltered$KitchenQual)
# table(trainFiltered$KitchenQual, trainFiltered$RecodeKQ)
# table(trainFiltered$Neighborhood, trainFiltered$KitchenQual)
# table(trainFiltered$NeighRecode, trainFiltered$KitchenQual)
trainFiltered$logLotArea <- log10(trainFiltered$LotArea)
ComplexModel <- lm(data = trainFiltered, SalePrice ~ TotalBsmtSF +
  GrLivArea + LotArea + TotalBath + Tier1 + Tier2 + Tier3 +
  Tier4)
summary(ComplexModel)
ComplexModel_Data <- broom::augment(ComplexModel)
FtestsComplexModel <- broom::glance(ComplexModel)
MAE_Complex <- mean(abs(ComplexModel_Data$.resid))
layout(matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), 3, 3, byrow = TRUE))
qqnorm(ComplexModel_Data$.resid)
qqline(ComplexModel_Data$.resid)
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$LotArea,
  xlab = "Lot Area", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$TotalBsmtSF,
  xlab = "Total basement area", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$GrLivArea,
  xlab = "Above grade living area", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$TotalBath,
  xlab = "Total bathrooms", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$Tier1,
  xlab = "Tier 1 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$Tier2,
  xlab = "Tier 2 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$Tier3,
  xlab = "Tier 3 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_Data$.resid, x = ComplexModel_Data$Tier4,
  xlab = "Tier 4 Neighborhood", ylab = "Residuals")

ComplexModel_log <- lm(data = trainFiltered, LogSalePrice ~ TotalBsmtSF +
  GrLivArea + LotArea + TotalBath + Tier1 + Tier2 + Tier3 +
  Tier4)
summary(ComplexModel_log)
ComplexModel_log_data <- broom::augment(ComplexModel_log)
FtestsComplexModel_log <- broom::glance(ComplexModel_log)
ComplexModel_log_data$AntiLogFitted <- 10^(ComplexModel_log_data$.fitted)
ComplexModel_log_data$SalePrice <- trainFiltered$SalePrice
MAE_Complex_log <- mean(abs(ComplexModel_log_data$SalePrice -
  ComplexModel_log_data$AntiLogFitted))
layout(matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), 3, 3, byrow = TRUE))
qqnorm(ComplexModel_log_data$.resid)

```

```

qqline(ComplexModel_log_data$.resid)
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$LotArea,
     xlab = "Lot Area", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$TotalBsmtSF,
     xlab = "Total basement area", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$GrLivArea,
     xlab = "Above grade living area", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$TotalBath,
     xlab = "Total bathrooms", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$Tier1,
     xlab = "Tier 1 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$Tier2,
     xlab = "Tier 2 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$Tier3,
     xlab = "Tier 3 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_log_data$.resid, x = ComplexModel_log_data$Tier4,
     xlab = "Tier 4 Neighborhood", ylab = "Residuals")
ComplexModels <- c("SalePrice = 7.17e+04 + 5.681e+01*TotalBsmtSF + 6.3e+01*GrLivArea + 8.22e-01*LotArea
                   "Log(Sale Price) = 4.979e+00 + 1e-04*TotalBsmtSF + 1.236e-04*GrLivArea + 1.251e-06*LotArea + 2.809e
ComplexMAEs <- c(MAE_Complex, MAE_Complex_log)
ComplexRSquared <- c(FtestsComplexModel$adj.r.squared, FtestsComplexModel_log$adj.r.squared)
ComplexFstat <- c(FtestsComplexModel$statistic, FtestsComplexModel_log$statistic)
ComplexPvalues <- c(FtestsComplexModel$p.value, FtestsComplexModel_log$p.value)
FigRef <- c(12, 13)
CompariosnComplex <- data.frame(ComplexModels, ComplexMAEs, ComplexRSquared,
                                  ComplexFstat, ComplexPvalues, FigRef)
colnames(CompariosnComplex) <- c("Model", "MAE", "Adj.R^2", "F.Stat",
                                  "P.Value", "Fig# reference")
pander::pandoc.table(CompariosnComplex, caption = "Comparison of MLR models of Sale Price and Log(Sale Price)

ComplexModel_log_lot <- lm(data = trainFiltered, LogSalePrice ~
                           TotalBsmtSF + GrLivArea + logLotArea + TotalBath + Tier1 +
                           Tier2 + Tier3 + Tier4)
summary(ComplexModel_log_lot)
ComplexModel_log_lot_data <- broom::augment(ComplexModel_log_lot)
FtestsComplexModel_log_lot <- broom::glance(ComplexModel_log_lot)
ComplexModel_log_lot_data$SalePrice <- trainFiltered$SalePrice
ComplexModel_log_lot_data$AntiLogFitted <- 10^(ComplexModel_log_lot_data$.fitted)
MAE_Complex_log_lot <- mean(abs(ComplexModel_log_lot_data$SalePrice -
                                   ComplexModel_log_lot_data$AntiLogFitted))
layout(matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), 3, 3, byrow = TRUE))
qqnorm(ComplexModel_log_lot_data$.resid)
qqline(ComplexModel_log_lot_data$.resid)
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$logLotArea,
      xlab = "Lot Area", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$TotalBsmtSF,
      xlab = "Total basement area", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$GrLivArea,
      xlab = "Above grade living area", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$TotalBath,
      xlab = "Total bathrooms", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$Tier1,
      xlab = "Tier 1 Neighborhood", ylab = "Residuals")

```

```

plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$Tier2,
      xlab = "Tier 2 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$Tier3,
      xlab = "Tier 3 Neighborhood", ylab = "Residuals")
plot(y = ComplexModel_log_lot_data$.resid, x = ComplexModel_log_lot_data$Tier4,
      xlab = "Tier 4 Neighborhood", ylab = "Residuals")
ComplexModels <- c("SalePrice = 7.17e+04 + 5.681e+01*TotalBsmtSF + 6.3e+01*GrLivArea + 8.22e-01*LotArea",
                    "Log(Sale Price) = 4.979e+00 + 1e-04*TotalBsmtSF + 1.236e-04*GrLivArea + 1.251e-06*LotArea + 2.809e-02",
                    "Log(Sale Price) = 4.716e+00 + TotalBsmtSF*1.037e-04 + GrLivArea*1.181e-04 + log(LotArea)*7.309e-02")
ComplexMAEs <- c(MAE_Complex, MAE_Complex_log, MAE_Complex_log_lot)
ComplexRSquared <- c(FtestsComplexModel$adj.r.squared, FtestsComplexModel_log$adj.r.squared,
                      FtestsComplexModel_log_lot$adj.r.squared)
ComplexFstat <- c(FtestsComplexModel$statistic, FtestsComplexModel_log$statistic,
                   FtestsComplexModel_log_lot$statistic)
ComplexPvalues <- c(FtestsComplexModel$p.value, FtestsComplexModel_log$p.value,
                     FtestsComplexModel_log_lot$p.value)
ComplexFigRef <- c(12, 13, 14)
CompariosnComplex <- data.frame(ComplexModels, ComplexMAEs, ComplexRSquared,
                                  ComplexFstat, ComplexPvalues, ComplexFigRef)
colnames(CompariosnComplex) <- c("Model", "MAE", "Adj.R^2", "F.Stat",
                                 "P.Value", "Fig# Reference")
pander::pandoc.table(CompariosnComplex, caption = "Comparison of MLR models of Sale Price and Log(Sale Price)")
library(mosaic)
sanitycheck <- do.call(rbind, dfapply(SampleFrame, favstats,
                                         select = is.numeric))
sanitycheck$mean <- as.numeric(format(sanitycheck$mean, digits = 1,
                                         nsmall = 2))
sanitycheck$sd <- as.numeric(format(sanitycheck$sd, digits = 1,
                                      nsmall = 2))
knitr::kable(sanitycheck, caption = "Data sanity check for numeric variables")
sanitycheckcharacter <- select(SampleFrame, colnames(SampleFrame[1,
                                                               sapply(SampleFrame, class) == "character"]))
library(purrr)
UniqueVals <- sanitycheckcharacter %>% map(unique)
# s <-
# data.frame(names(tst), sapply(tst, function(x){paste(x, collapse
# = ',')})), row.names = NULL)
Counts <- data.frame(sapply(UniqueVals, length), do.call(rbind,
               dfapply(sanitycheckcharacter, length, select = is.character)),
                     do.call(rbind, dfapply(sanitycheckcharacter, n_missing, select = is.character)),
                     row.names = names(UniqueVals))
colnames(Counts) <- c("# Unique", "n", "missing")
knitr::kable(Counts, caption = "Data sanity check for nominal variables",
            align = c("l", "r", "r", "r"))
anova(SLR_LivingArea_NoIntercept, MLR)
anova(SLR_BsmtArea, MLR)

```