

Assignment 2: Regression Model Building

Sri Seshadri

7/1/2017

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 models were limited only two predictor variables. Total basement area and above grade living area (sometimes referred as living area in this report) were found to be good candidates for predicting sale price of typical homes. The model was formulated as Sale Price = 84.11(Total Basement Area) + 97.6(Above Grade Living Area) - 50244.02. The model explains 70% of the variation in Sale Price. But is limited in application due to prediction errors ranging up to +/- \$ 86,000 (+/- 2 standard deviation of residuals).

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	TotalBsmtSF	KitchenQual
LotConfig	GrLivArea	TotRmsAbvGrd
Neighborhood	BsmtFullBath	GarageArea
BldgType	BsmtHalfBath	MoSold
HouseStyle	FullBath	YrSold

OverallCond	HalfBath	SaleCondition
YearRemodel	BedroomAbvGr	SalePrice

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.

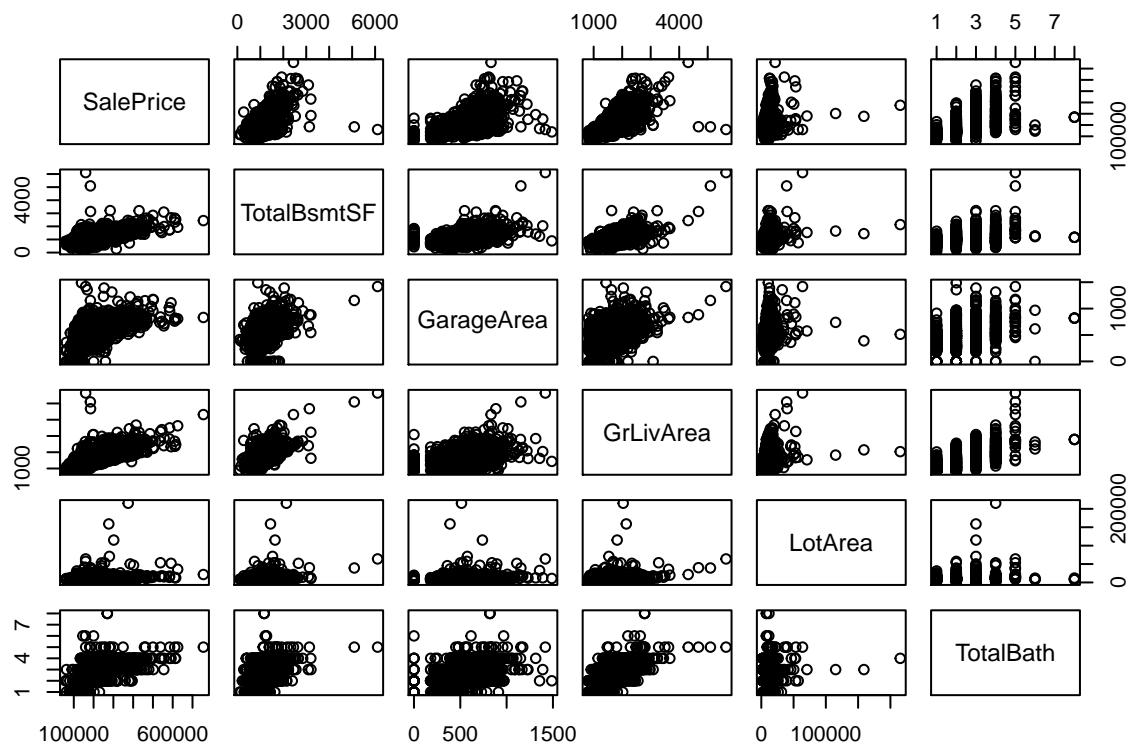


Figure 1: Multicollinearity check amongst potential predictors

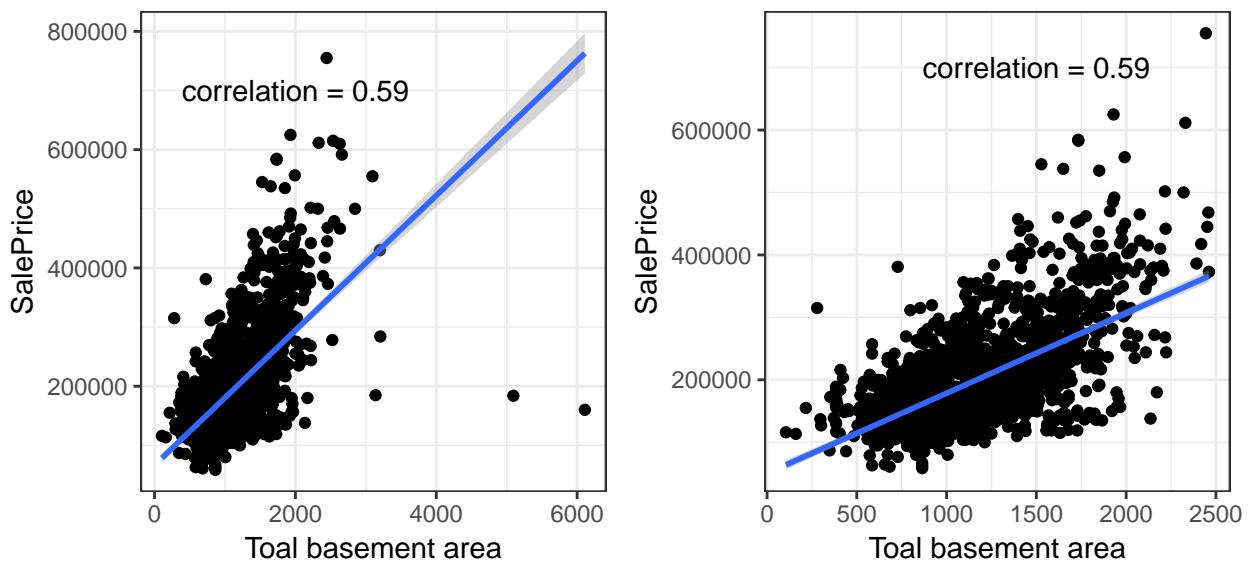


Figure 2: Sale price vs Basement area

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.

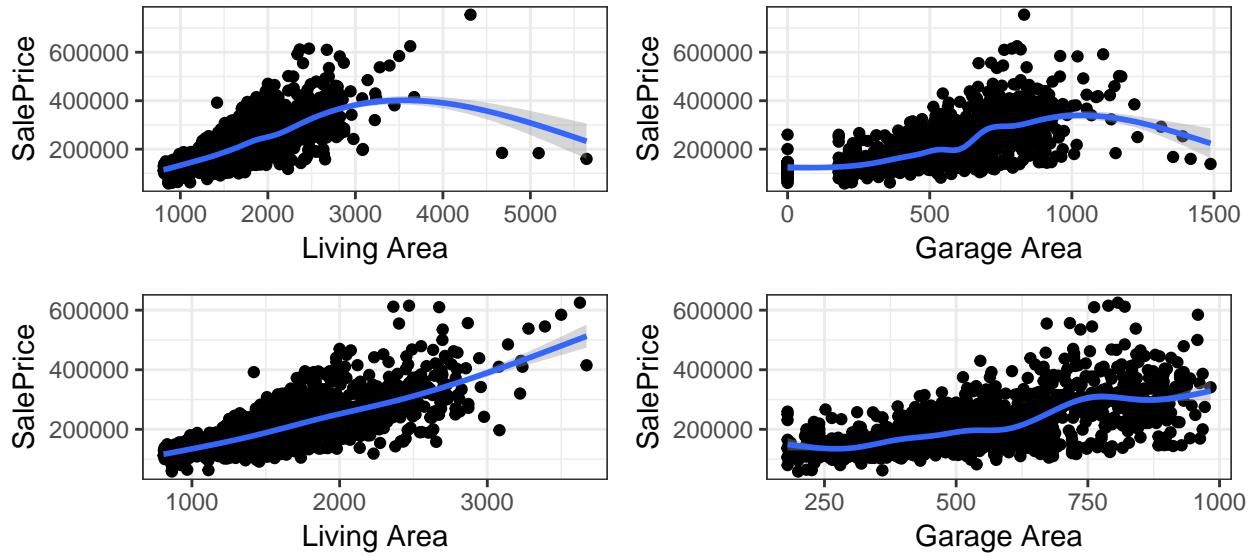


Figure 3: Sales Price vs Garage Area and LivingArea

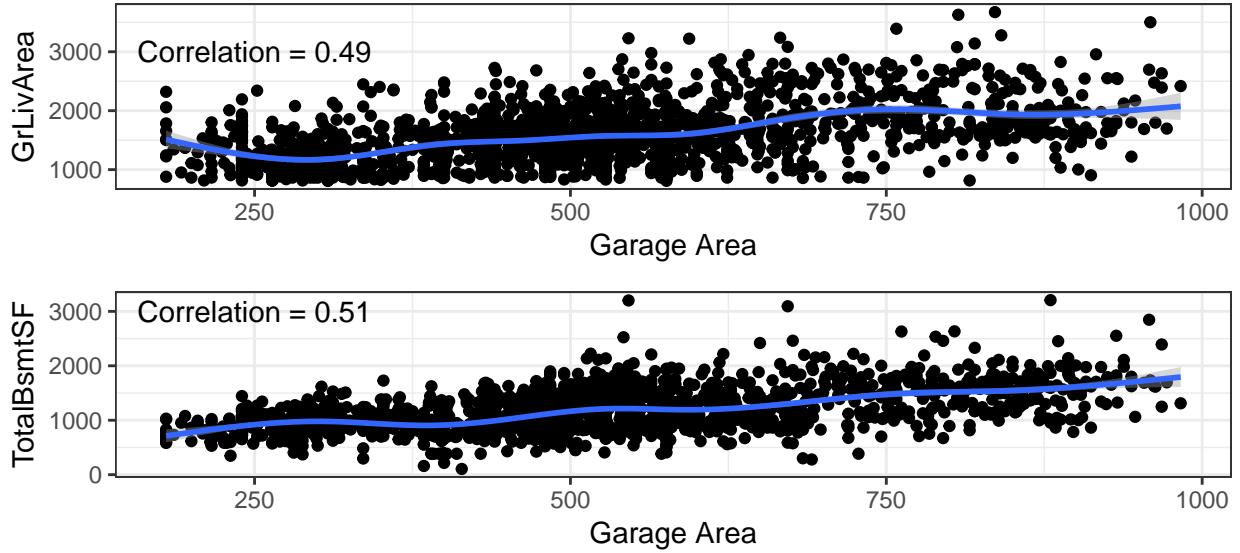


Figure 4: GarageArea vs Living and Basement area

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 with Living area above grade as predictor is seen below. We fail to reject the null hypothesis of intercept = 0. It would be appropriate to fit a no intercept model. From a goodness of fit perspective, 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 a bullhorn 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
```

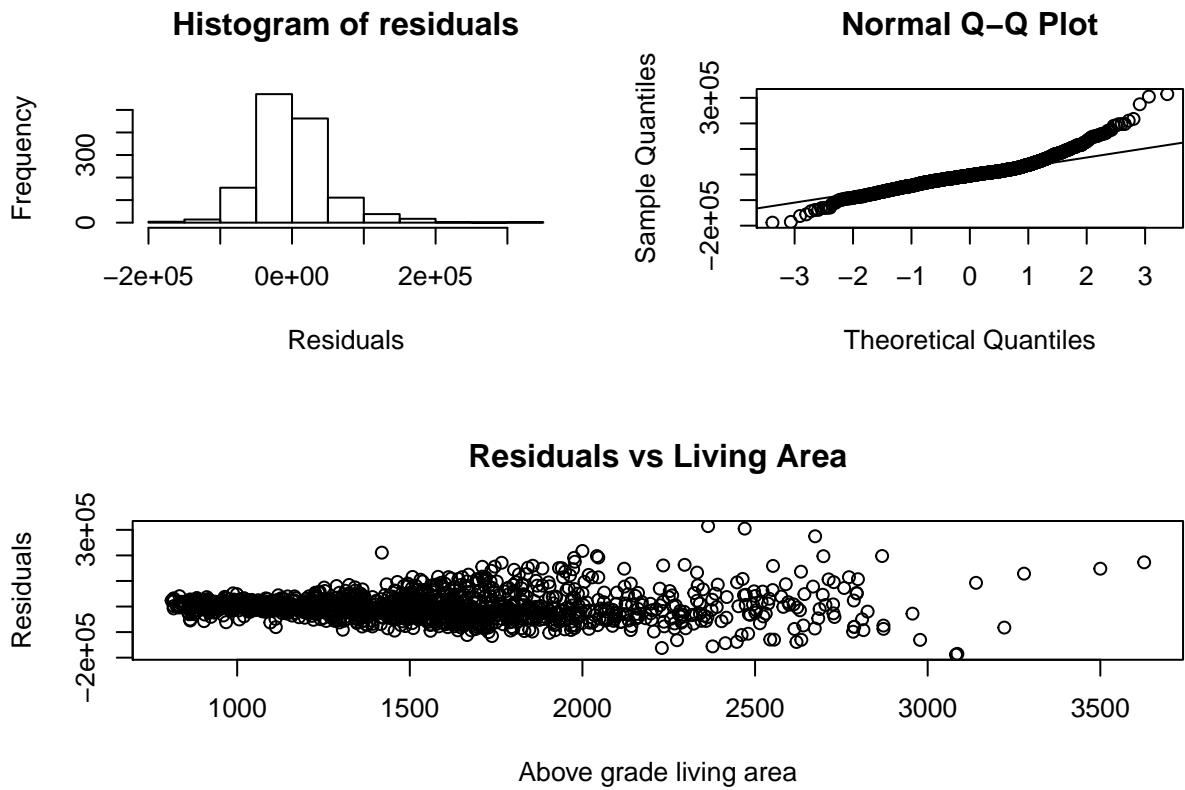


Figure 5: Model diagnostics Sale Price ~ GrLivArea

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. The no intercept model seem to have a better Adjusted R-Squared value of 0.94; however, there is no practical implication of this fit. When the residual plots are compared with the previous model, the kurtosis of the residuals has increased, but the bullhorn pattern still remains.

```
##
## 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
```

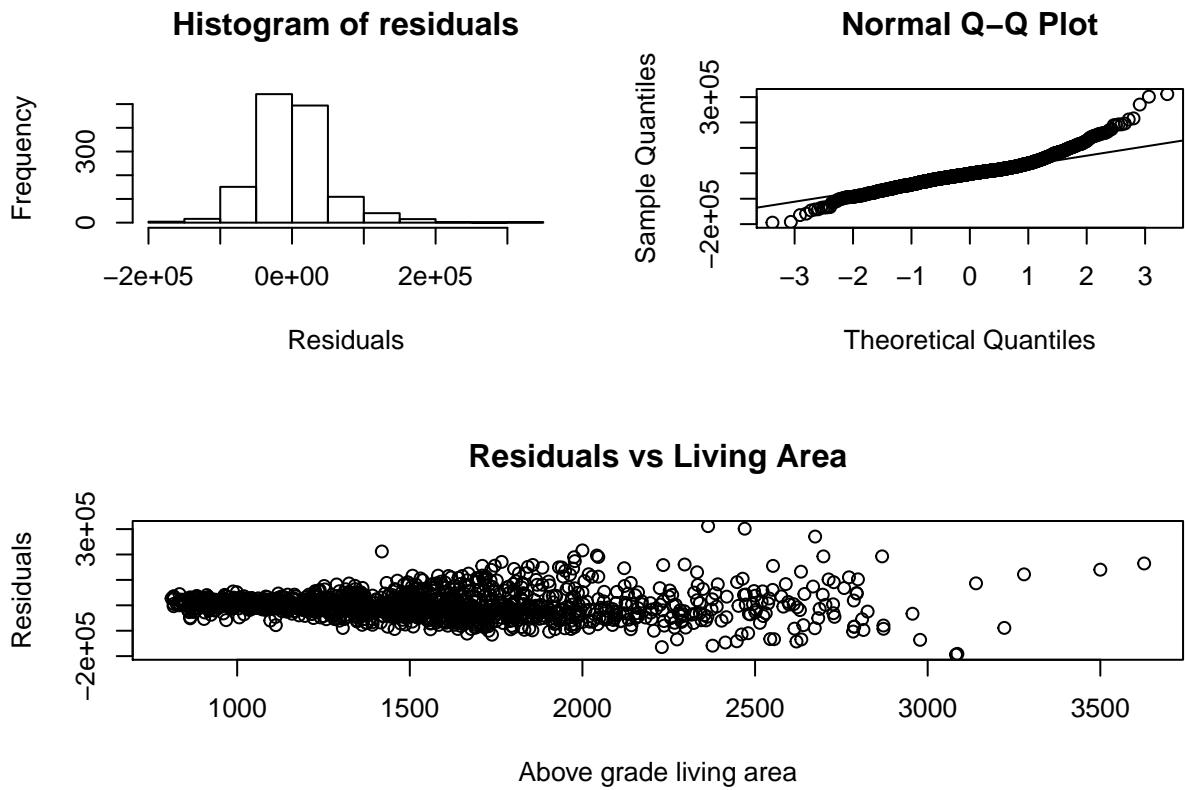


Figure 6: No intercept model

4.2 Simple linear regression with Total Basement area as predictor

The results of a simple linear model with Total basement area as predictor is shown below. The sum of squares error is much greater than the one with living area as predictor. Making basement area a poor predictor. 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
```

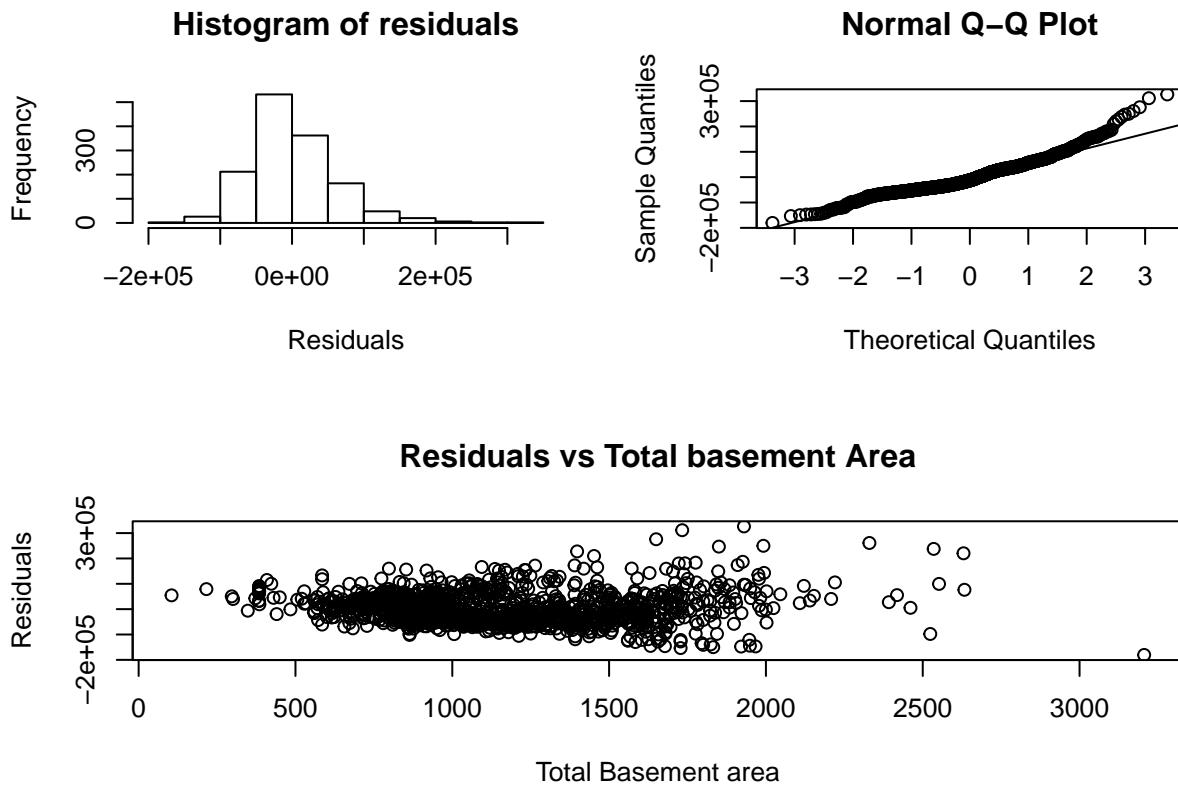


Figure 7: Simple Linear Regression - Total Basement Area as predictor

5. Multiple linear regression model

The below results is of multiple linear model fit on the training data with both the above grade area and the total basement area as predictors. The model has an adjusted R-squared of 70% and the residuals seem to fairly follow the normal distribution with mean zero. The residuals are random. 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

```

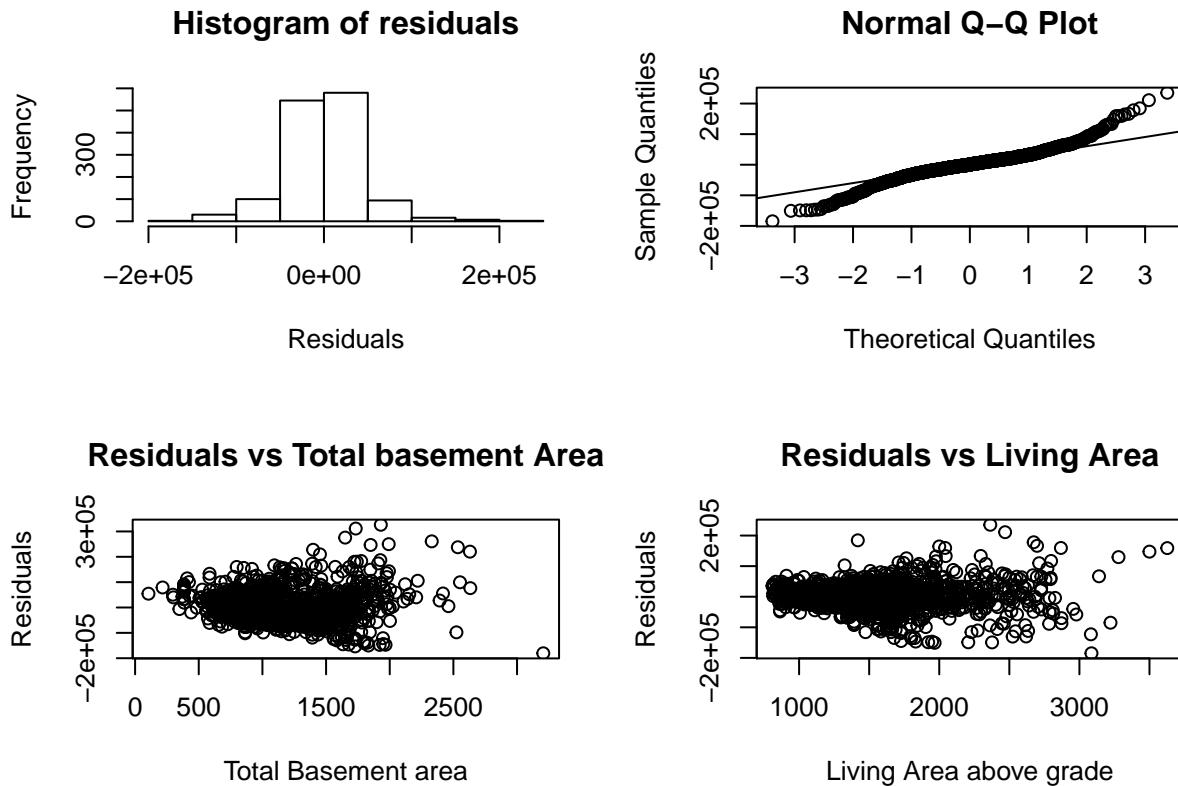


Figure 8: Multiple Linear regression

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.

The approach used to compare the models is F test for nested models. 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.

5.2 Nested model comparison.

The ANOVA table below compares the Reduced Model(RM) $\text{SalePrice} \sim \text{GrLivArea}$ to the Full Model(FM) $\text{SalePrice} \sim \text{TotalBsmtSF} + \text{GrLivArea}$. 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

```

6. Regression models for transformed response - log(SalePrice)

Logarithmic transformations of variables help with handling distributions that have heavy tails by normalizing them. In this section we will explore if modelling the logarithmic transformation of SalePrice makes the regression models better.

6.1 Comparison of Regressions of log(SalePrice) by Living Area Vs SalePrice by Living Area

It is seen from figure 9 and 10 that for the simple linear regression with the Living area as the predictor, the logarithmic transformation has not yielded a practical difference. The residuals look more normal and the variance inflation with increase in living area has been addressed by the transformation, however the model has yielded very little practical significance.

```

## 
## Call:
## lm(formula = LogSalePrice ~ GrLivArea, data = trainFiltered)
## 
## Residuals:
##       Min     1Q     Median      3Q     Max 
## -0.36035 -0.05068 -0.00282  0.05039  0.35892 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.885e+00 9.188e-03 531.67  <2e-16 ***
## GrLivArea   2.462e-04 5.702e-06   43.19  <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.09937 on 1374 degrees of freedom
## Multiple R-squared:  0.5758, Adjusted R-squared:  0.5755 
## F-statistic: 1865 on 1 and 1374 DF,  p-value: < 2.2e-16

```

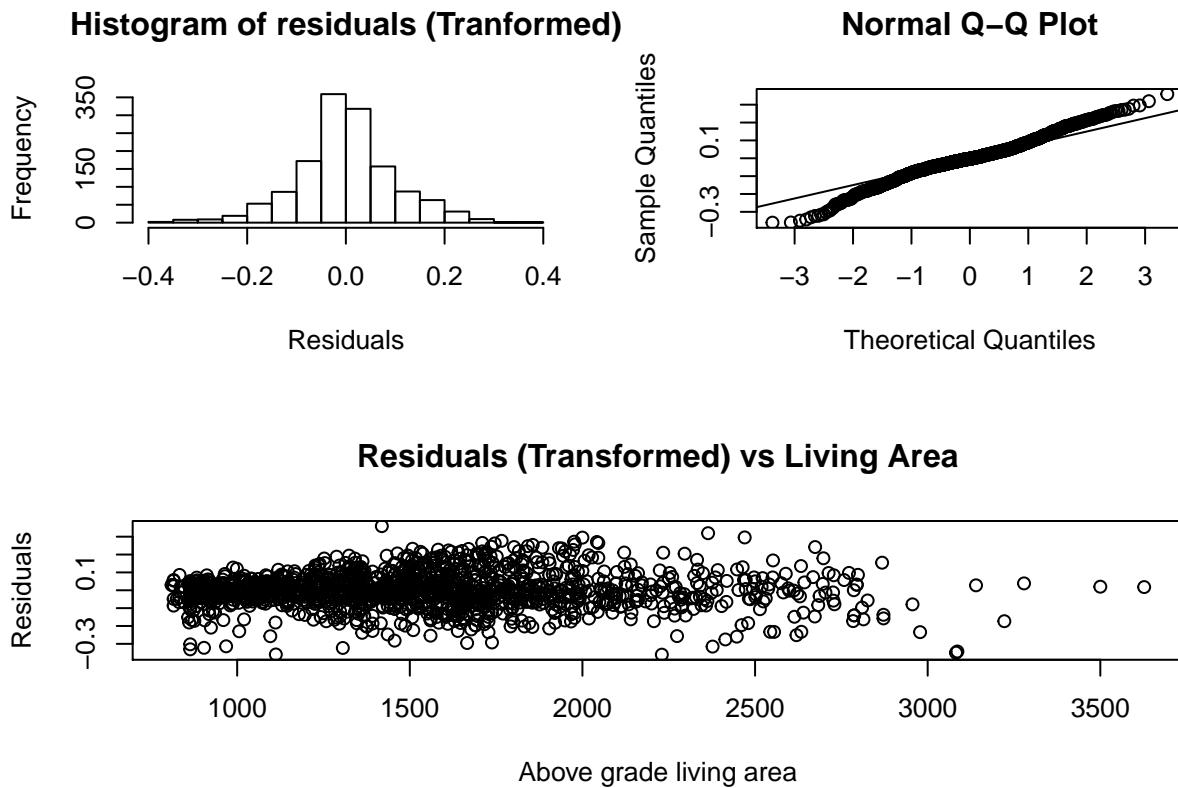


Figure 9: Simple linear regression of Log(SalePrice) by above grade living area

6.2 Comparison of Regressions of log(SalePrice) by Total basement area Vs SalePrice by Total basement area

The log(SalePrice) when regressed by total basement area, yields a better model relative to the regression of non-transformed sale price. The Adjusted R squared are approximately the same for both the models and the residuals of the model with transformation meets the normality and homoscedastic assumption as much as the non-transformed model. The adjusted R squared of 70% is good, however as mentioned above, this model needs improvement.

```
##
## Call:
## lm(formula = LogSalePrice ~ TotalBsmtSF, data = trainFiltered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42845 -0.08858 -0.00951  0.08891  0.35564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.981e+00  1.049e-02 475.07  <2e-16 ***
## TotalBsmtSF 2.480e-04  8.727e-06 28.42  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1211 on 1374 degrees of freedom
## Multiple R-squared:  0.3702, Adjusted R-squared:  0.3698
```

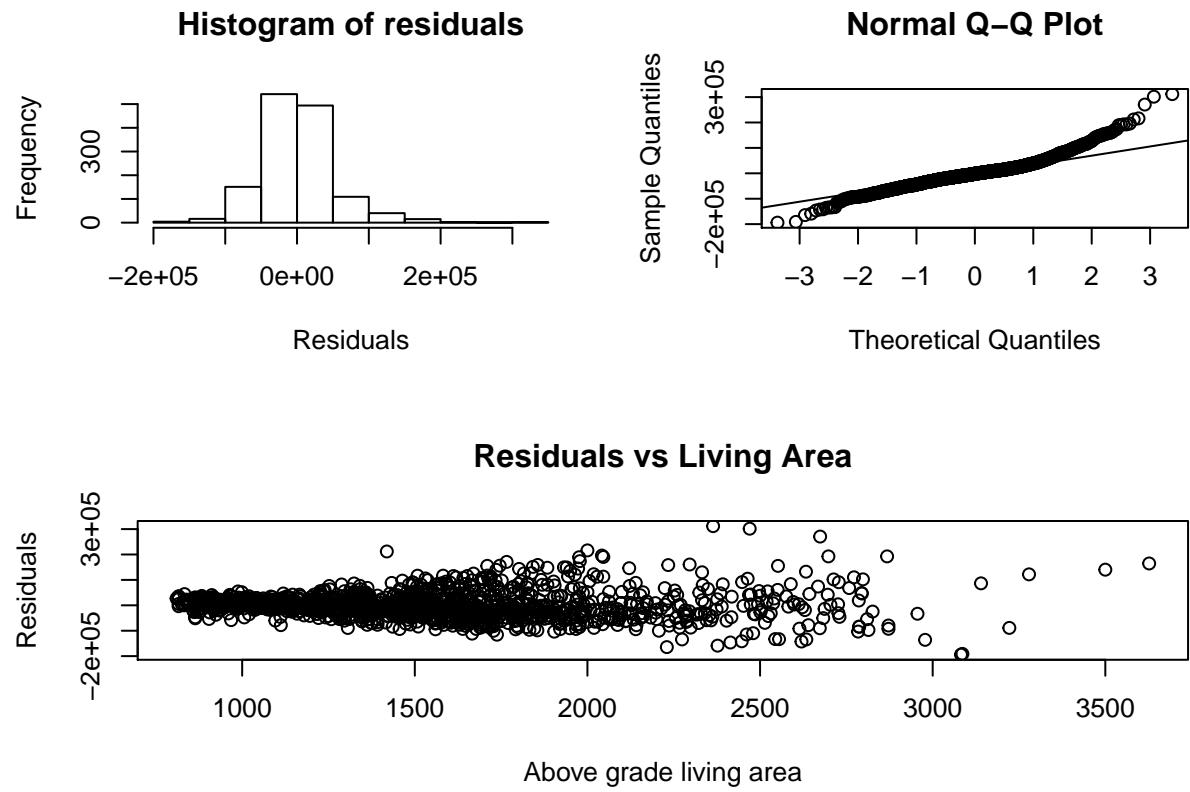


Figure 10: Linear regression of SalePrice by above grade living area

```
## F-statistic: 807.8 on 1 and 1374 DF,  p-value: < 2.2e-16
```

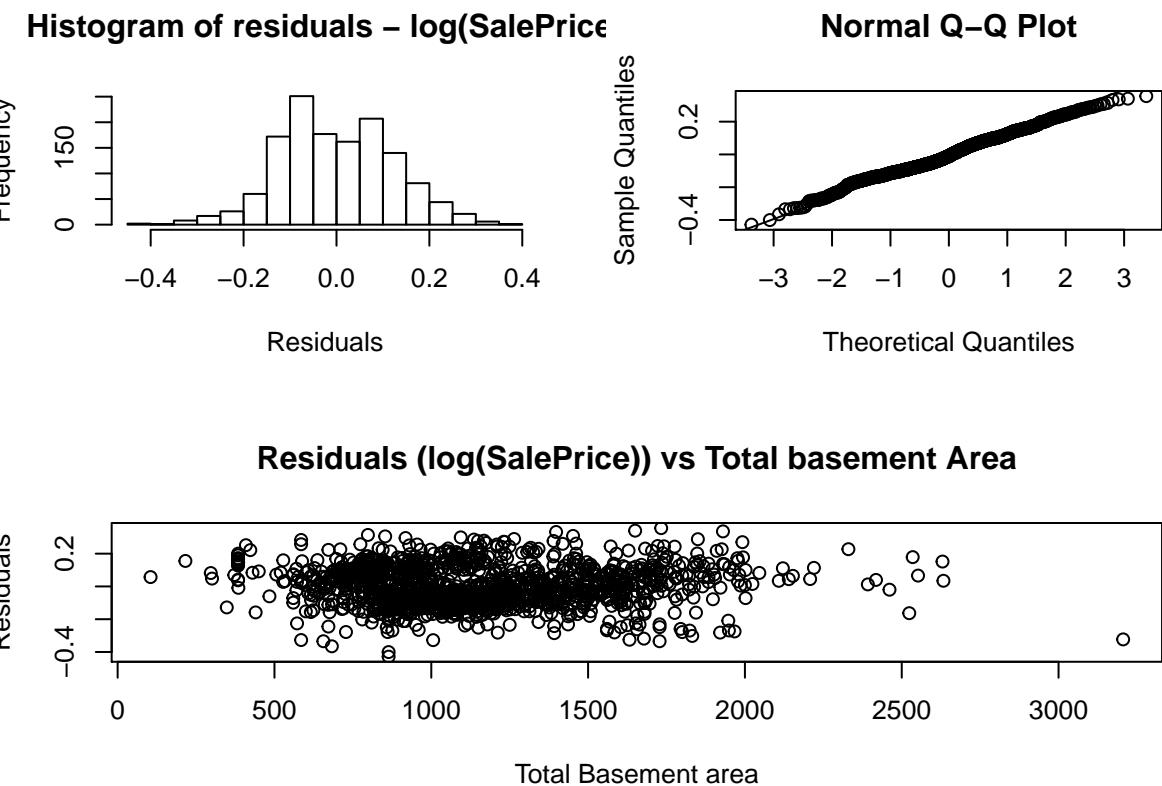
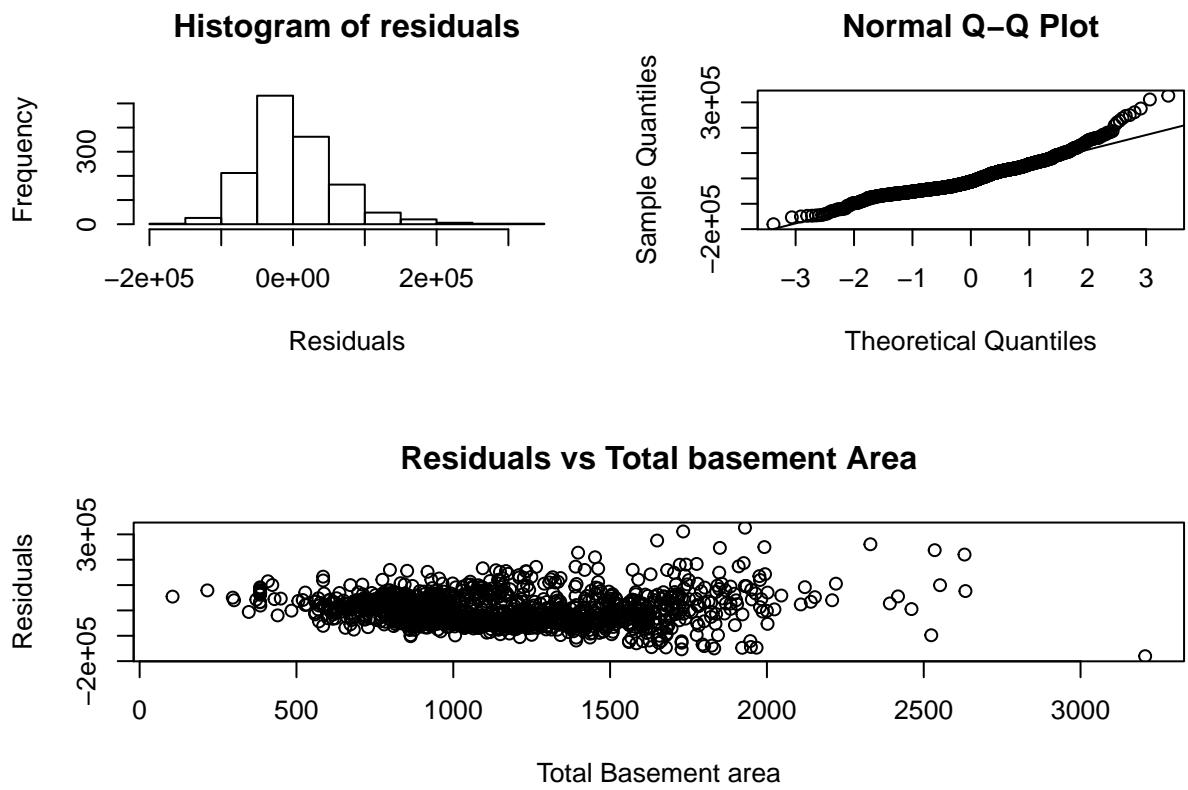


Figure 11: SLR log(SalePrice) by Total basement area



6.3 Multiple linear regression of Sale Price and log(Sale Price) by Basement area and above grade living area

It is seen from the ANOVA table and figures 12 and 13 that the logarithmic transformation has not provided us with a significant advantage from a practical perspective. The normality and homoscedasticity assumption was satisfied reasonable in both models. Therefore no transformation is recommended for this model.

```
##  
## Call:  
## lm(formula = LogSalePrice ~ TotalBsmtSF + GrLivArea, data = trainFiltered)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.34218 -0.03987  0.00861  0.05312  0.31066  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.780e+00 8.936e-03 534.88 <2e-16 ***  
## TotalBsmtSF 1.539e-04 6.514e-06 23.62 <2e-16 ***  
## GrLivArea   2.004e-04 5.186e-06 38.65 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.08382 on 1373 degrees of freedom  
## Multiple R-squared:  0.6984, Adjusted R-squared:  0.698  
## F-statistic:  1590 on 2 and 1373 DF, p-value: < 2.2e-16
```

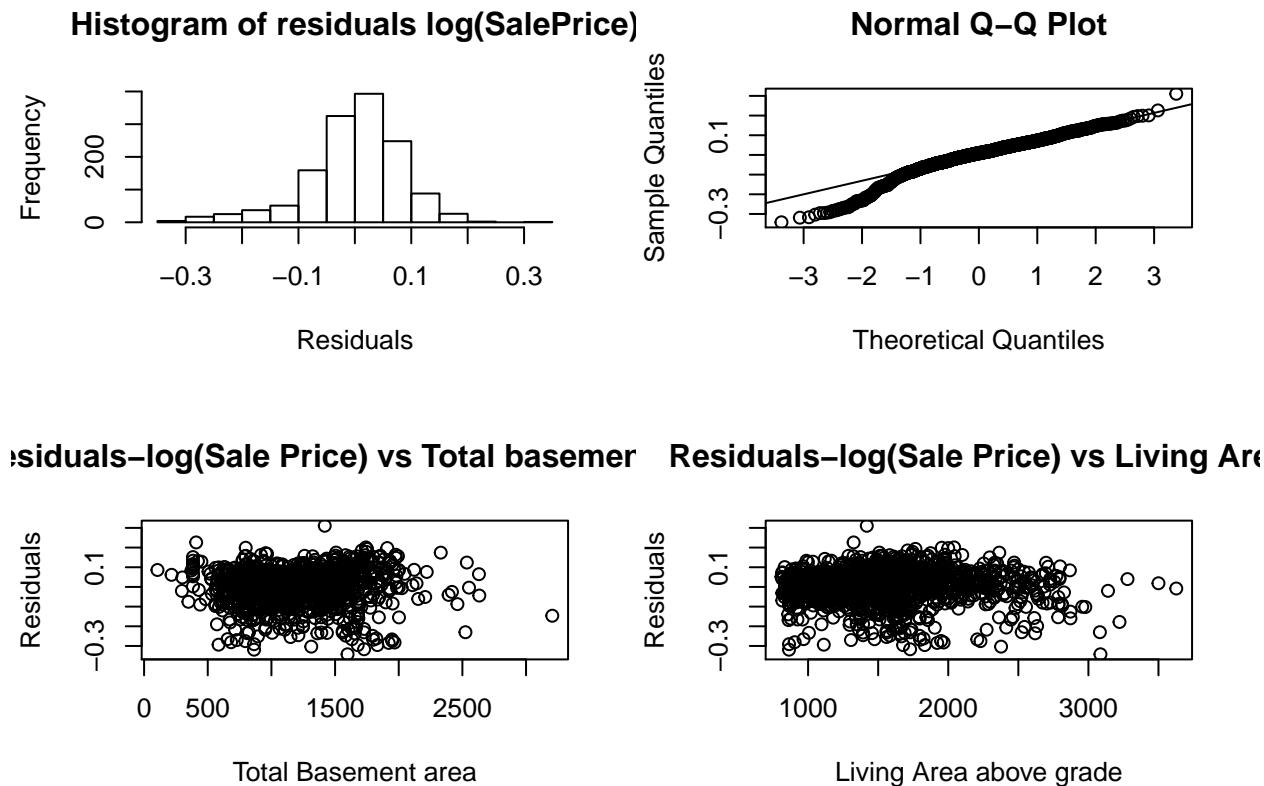


Figure 12: Multiple Linear regression of log(Sale Price)

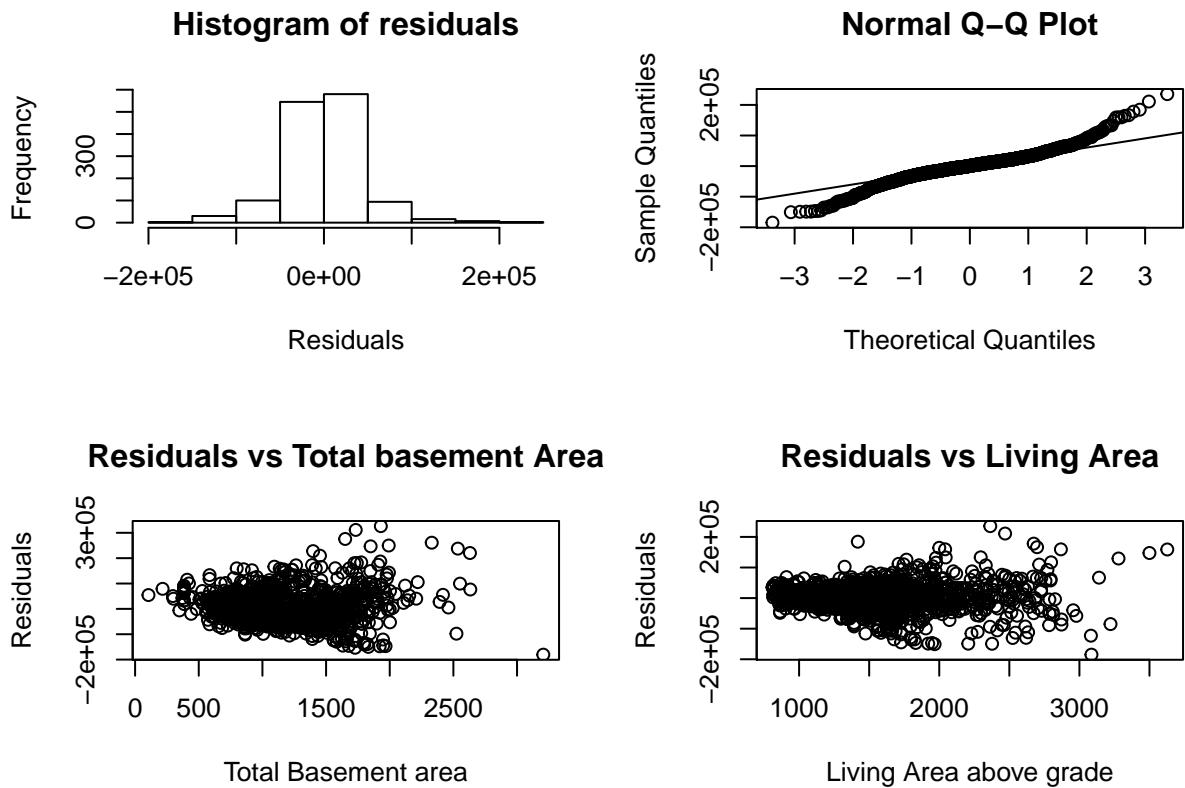


Figure 13: Multiple Linear regression - SalePrice

7. Conclusion

A multiple linear regression model with Total basement area and above grade living area explains 70% of the variation in Sales Price of typical homes in Ames, Iowa. It is formulated as Sale Price = 84.11(Total Basement Area) + 97.6(Above Grade Living Area) - 50244.02. The model meet it's statistical assumptions. However, from the perspective of practical use of this model, the prediction could be off by \$86,000; which is 49% of median Sale Price of the sample of typical houses. The model needs improvement.

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 4: 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
TotalBath	1	2.00	3.0	3.00	8	3e+00	9e-01	1992	0

Table 5: 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 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")

# 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)
# 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 basment 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, GarageArea, LivingAreaRestricted, GarageAreaRestricted,
             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)
# 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)
# 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)
# 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, ResdfBsmt$.resid, main = "Residuals vs Total basement Area",
     xlab = "Total Basement area", ylab = "Residuals")
plot(MLR_Model$GrLivArea, MLR_Model$.resid, main = "Residuals vs Living Area",
     xlab = "Living Area above grade", ylab = "Residuals")
anova(SLR_LivingArea_NoIntercept, MLR)
anova(SLR_BsmtArea, MLR)
SLR_LivingArea_Log <- lm(data = trainFiltered, LogSalePrice ~
                           GrLivArea)
print(summary(SLR_LivingArea_Log), caption = "ANOVA Simple Linear Regression of LogSalePrice by Above ground floor area")

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

```

```

  xlab = "Above grade living area", ylab = "Residuals")
SLR_BsmtArea_Log <- lm(data = trainFiltered, LogSalePrice ~ TotalBsmtSF)
print(summary(SLR_BsmtArea_Log), caption = "ANOVA Simple Linear Regression Log(SalePrice) vs Total basm"

# Model diagnostics Tidy store of model results
ResdfBsmt_Log <- broom::augment(SLR_BsmtArea_Log)
# Plot of residuals
layout(matrix(c(1, 2, 3, 3), 2, 2, byrow = TRUE))
hist(ResdfBsmt_Log$.resid, main = "Histogram of residuals - log(SalePrice)",
     xlab = "Residuals")
qqnorm(ResdfBsmt_Log$.resid, title = "Normal Q-Q plot of residuals log(Sale Price)")
qqline(ResdfBsmt_Log$.resid)
plot(ResdfBsmt_Log$TotalBsmtSF, ResdfBsmt_Log$.resid, main = "Residuals (log(SalePrice)) vs Total basem
  xlab = "Total Basement area", ylab = "Residuals")
# Model diagnostics Tidy store of model results
ResdfBsmt <- broom::augment(SLR_BsmtArea)
# Plot of residuals
layout(matrix(c(1, 2, 3, 3), 2, 2, byrow = TRUE))
hist(ResdfBsmt$.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_log <- lm(data = trainFiltered, LogSalePrice ~ TotalBsmtSF +
  GrLivArea)
print(summary(MLR_log), caption = "ANOVA Multiple Linear Regression Total basment area")

# Model diagnostics Tidy store of model results
MLR_Model_log <- broom::augment(MLR_log)
# Plot of residuals
layout(matrix(c(1, 2, 3, 4), 2, 2, byrow = TRUE))
hist(MLR_Model_log$.resid, main = "Histogram of residuals log(SalePrice)",
     xlab = "Residuals")
qqnorm(MLR_Model_log$.resid, title = "Normal Q-Q plot of residuals log(Sale Price)")
qqline(MLR_Model_log$.resid)
plot(MLR_Model_log$TotalBsmtSF, MLR_Model_log$.resid, main = "Residuals-log(Sale Price) vs Total baseme
  xlab = "Total Basement area", ylab = "Residuals")
plot(MLR_Model_log$GrLivArea, MLR_Model_log$.resid, main = "Residuals-log(Sale Price) vs Living Area",
      xlab = "Living Area above grade", ylab = "Residuals")

# 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, ResdfBsmt$.resid, main = "Residuals vs Total basement Area",
      xlab = "Total Basement area", ylab = "Residuals")
plot(MLR_Model$GrLivArea, MLR_Model$.resid, main = "Residuals vs Living Area",
      xlab = "Living Area above grade", ylab = "Residuals")
library(mosaic)
sanitycheck <- do.call(rbind, dfapply(SampleFrame, favstats,
  select = is.numeric))
sanitycheck$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"))

```