Prediction of House Prices

Introduction

The data set housePrice is provided to develop a model to predict the sellng price of a house based on several features.The data set consist of data collected from 5575 sold houses and 12 variables.

1. soldPrice - sold price of house
2. sqftLiving - square footage of living area
3. sqftLand - square footage of land
4. sqftAbove - square footage of area above ground
5. sqftBasement - square footage of basement
6. numBedRooms - number of bed rooms
7. numBathRooms - number of bath rooms
8. numFloors - number of foors
9. builtYear - year of construction
10. grade - construction quality ranked from 1 to 4 where 1 is the lowest grade
11. waterFront- whether the house has a waterfront (1) or not (0)
12. condition - condition of the house (Excellent, Good, Average)

```
## 'data.frame':    5575 obs. of  12 variables:
##  $ soldPrice   : num  221900 180000 604000 510000 229500 ...
##  $ sqftLiving  : int  1180 770 1960 1680 1780 1160 2950 1890 1200 1250 ...
##  $ sqftLand    : int  5650 10000 5000 8080 7470 6000 5000 14040 9850 9774 ...
##  $ sqftAbove   : int  1180 770 1050 1680 1050 860 1980 1890 1200 1250 ...
##  $ sqftBasement: int  0 0 910 0 730 300 970 0 0 0 ...
##  $ numBedRooms : int  3 2 4 3 3 2 4 3 2 3 ...
##  $ numBathRooms: int  1 1 3 2 1 1 3 2 1 1 ...
##  $ numFloors   : int  1 1 1 1 1 1 2 2 1 1 ...
##  $ builtYear   : int  1955 1933 1965 1987 1960 1942 1979 1994 1921 1969 ...
##  $ grade       : int  2 1 2 2 2 2 2 2 2 2 ...
##  $ waterFront  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition   : chr  "average" "average" "excellent" "average" ...
```

Libraries used for this project

```
library(corrplot)
library(car)
library(broom)
library(dplyr)
library(caTools)
library(ggplot2)
library(plotly)
library(MLmetrics)
```

Exploratory Data Analysis

Dataset

```
##   soldPrice sqftLiving sqftLand sqftAbove sqftBasement numBedRooms numBathRooms
## 1    221900       1180     5650      1180            0           3            1
## 2    180000        770    10000       770            0           2            1
## 3    604000       1960     5000      1050          910           4            3
## 4    510000       1680     8080      1680            0           3            2
## 5    229500       1780     7470      1050          730           3            1
## 6    468000       1160     6000       860          300           2            1
##   numFloors builtYear grade waterFront condition
## 1         1      1955     2          0   average
## 2         1      1933     1          0   average
## 3         1      1965     2          0 excellent
## 4         1      1987     2          0   average
## 5         1      1960     2          0   average
## 6         1      1942     2          0      good
```

Checking for any missing values

```
sum(is.na(housePrice))
```

```
## [1] 0
```

There are no missing values

Condition column has to be changed into an ordinal categorical variable

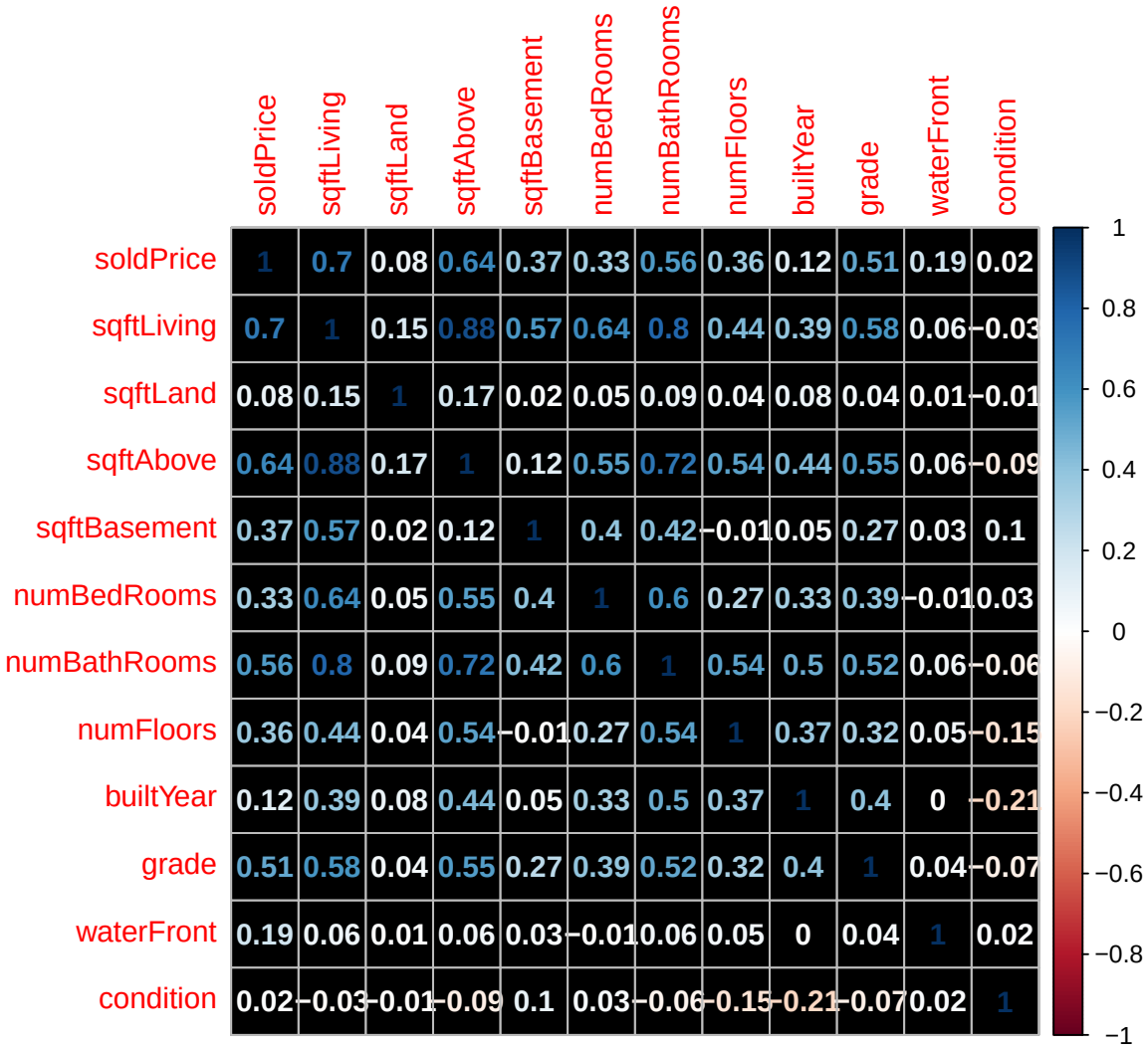
```
housePrice$condition[housePrice$condition=='average']<-1
housePrice$condition[housePrice$condition=='good']<-2
housePrice$condition[housePrice$condition=='excellent']<-3
housePrice$condition = as.integer(housePrice$condition)
```

Summary of the Dataset

```
##    soldPrice          sqftLiving      sqftLand          sqftAbove
##  Min.   :  78000   Min.   : 390   Min.   :    600   Min.   : 390
##  1st Qu.: 253702   1st Qu.:1020   1st Qu.:   5120   1st Qu.: 950
##  Median : 356000   Median :1340   Median :   7425   Median :1160
##  Mean   : 427782   Mean   :1565   Mean   :  13573   Mean   :1352
##  3rd Qu.: 499936   3rd Qu.:1820   3rd Qu.:   9826   3rd Qu.:1510
##  Max.   :5350000   Max.   :8000   Max.   :1651359   Max.   :7850
##   sqftBasement     numBedRooms     numBathRooms     numFloors      builtYear
##  Min.   :   0.0   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1900
##  1st Qu.:   0.0   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1942
##  Median :   0.0   Median :3.000   Median :1.000   Median :1.00   Median :1955
##  Mean   : 212.7   Mean   :2.921   Mean   :1.581   Mean   :1.19   Mean   :1958
##  3rd Qu.: 325.0   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:1.00   3rd Qu.:1975
##  Max.   :2810.0   Max.   :5.000   Max.   :6.000   Max.   :3.00   Max.   :2015
##      grade         waterFront         condition
##  Min.   :1.000   Min.   :0.000000   Min.   :1.000
```

```
##  1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:1.000
##  Median :2.000   Median :0.000000   Median :1.000
##  Mean   :1.758   Mean   :0.004664   Mean   :1.483
##  3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :1.000000   Max.   :3.000
```

checking for any collinearity of explanatory variables

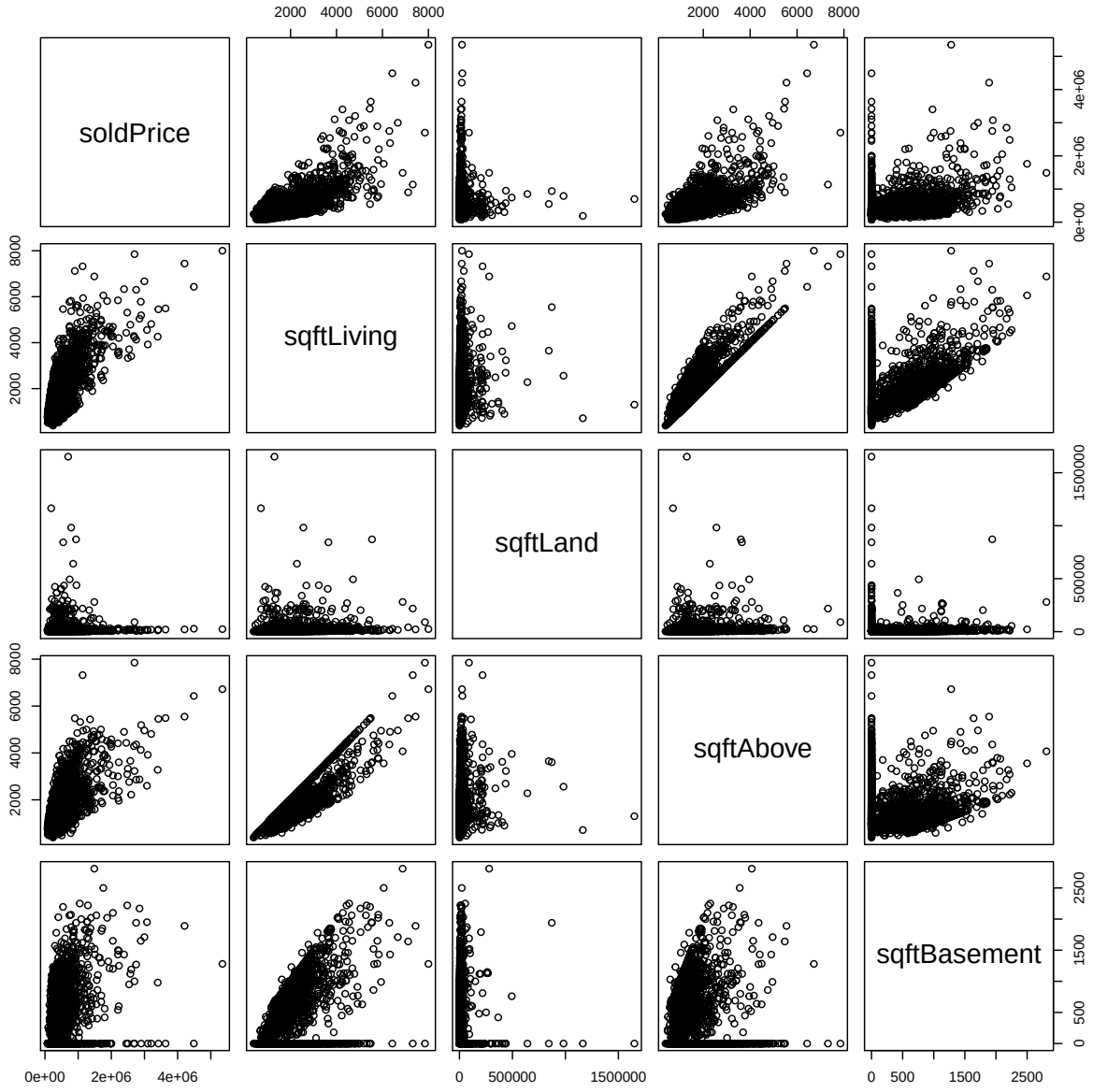


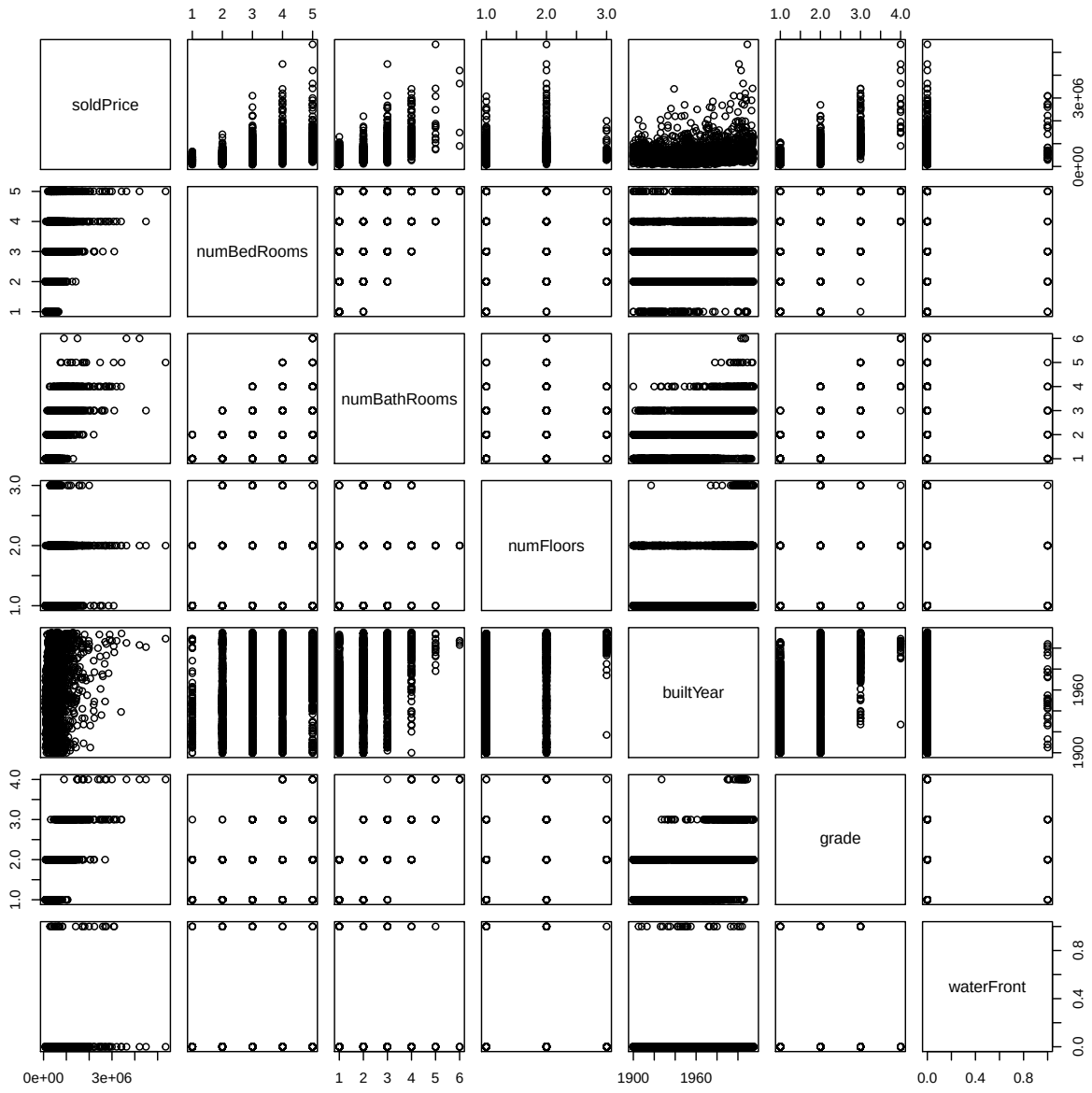

```
##              soldPrice sqftLiving sqftLand sqftAbove sqftBasement numBedRooms
## soldPrice         1.00       0.70     0.08      0.64         0.37        0.33
## sqftLiving        0.70       1.00     0.15      0.88         0.57        0.64
## sqftLand          0.08       0.15     1.00      0.17         0.02        0.05
## sqftAbove         0.64       0.88     0.17      1.00         0.12        0.55
## sqftBasement      0.37       0.57     0.02      0.12         1.00        0.40
```

```
## numBedRooms       0.33       0.64     0.05      0.55         0.40        1.00
## numBathRooms      0.56       0.80     0.09      0.72         0.42        0.60
## numFloors         0.36       0.44     0.04      0.54        -0.01        0.27
## builtYear         0.12       0.39     0.08      0.44         0.05        0.33
## grade             0.51       0.58     0.04      0.55         0.27        0.39
## waterFront        0.19       0.06     0.01      0.06         0.03       -0.01
## condition         0.02      -0.03    -0.01     -0.09         0.10        0.03
##              numBathRooms numFloors builtYear grade waterFront condition
## soldPrice            0.56      0.36      0.12  0.51       0.19      0.02
## sqftLiving           0.80      0.44      0.39  0.58       0.06     -0.03
## sqftLand             0.09      0.04      0.08  0.04       0.01     -0.01
## sqftAbove            0.72      0.54      0.44  0.55       0.06     -0.09
## sqftBasement         0.42     -0.01      0.05  0.27       0.03      0.10
## numBedRooms          0.60      0.27      0.33  0.39      -0.01      0.03
## numBathRooms         1.00      0.54      0.50  0.52       0.06     -0.06
## numFloors            0.54      1.00      0.37  0.32       0.05     -0.15
## builtYear            0.50      0.37      1.00  0.40       0.00     -0.21
## grade                0.52      0.32      0.40  1.00       0.04     -0.07
## waterFront           0.06      0.05      0.00  0.04       1.00      0.02
## condition           -0.06     -0.15     -0.21 -0.07       0.02      1.00
```

There is a very high correlation between sqftLiving and sqftAbove and between sqftLiving and numBathRooms. Hence including sqftAbove and numBathrooms along with sqftLiving in the model is not ideal due to collinearity.

Pairwise comparisons of the variables


soldPrice
sqftLiving
sqftLand
sqftAbove
sqftBasement
2000 4000 6000 8000
0e+00 2e+06 4e+06
0 500000 1500000
0 500 1500 2500



Feature engineering

Removing the predictor sqftAbove from the dataframe

```
housePrice <- subset(housePrice, select = -c(sqftAbove) )
```

Here we decided to drop sqftAbove since it has a high correlation with sqftLiving and we cannot remove sqftLiving because it's has a high correlation with response sold price.

Creating a new interaction term

```
housePrice["interaction"] = (housePrice$sqftLiving*housePrice$numBathRooms)
housePrice <- subset(housePrice, select = -c(sqftLiving,numBathRooms))
```

We created a new interaction term combining sqftLiving and numBathRooms to eliminate collinearity and also maintain the reliability of the model

Checking for collinearity after adjustments

```
##              soldPrice sqftLand sqftBasement numBedRooms numFloors builtYear
## soldPrice         1.00     0.08         0.37        0.33      0.36      0.12
## sqftLand          0.08     1.00         0.02        0.05      0.04      0.08
## sqftBasement      0.37     0.02         1.00        0.40     -0.01      0.05
## numBedRooms       0.33     0.05         0.40        1.00      0.27      0.33
## numFloors         0.36     0.04        -0.01        0.27      1.00      0.37
## builtYear         0.12     0.08         0.05        0.33      0.37      1.00
## grade             0.51     0.04         0.27        0.39      0.32      0.40
## waterFront        0.19     0.01         0.03       -0.01      0.05      0.00
## condition         0.02    -0.01         0.10        0.03     -0.15     -0.21
## interaction       0.72     0.14         0.48        0.58      0.48      0.41
##              grade waterFront condition interaction
## soldPrice     0.51       0.19      0.02        0.72
## sqftLand      0.04       0.01     -0.01        0.14
## sqftBasement  0.27       0.03      0.10        0.48
## numBedRooms   0.39      -0.01      0.03        0.58
## numFloors     0.32       0.05     -0.15        0.48
## builtYear     0.40       0.00     -0.21        0.41
## grade         1.00       0.04     -0.07        0.55
## waterFront    0.04       1.00      0.02        0.07
## condition    -0.07       0.02      1.00       -0.06
## interaction   0.55       0.07     -0.06        1.00
```

VIF values after feature engineering

```
##     sqftLand sqftBasement  numBedRooms    numFloors    builtYear        grade
##     1.028058     1.538362     1.605727     1.488596     1.427247     1.534067
##   waterFront    condition  interaction
##     1.011312     1.073905     2.628963
```

By introducing the interaction term we have reduced the correlation between predictors and further we have the interaction term with the highest correlation with the response soldPrice. And also the vif values suggests that collinearity is very low(1<vif<5)

Model Fitting

Train test split

We split our data set into training and testing parts, training set contain 70% of observations and 10 variables. Following are the variables that we start with

1. soldPrice - sold price of house.(Response Variable)
2. sqftLand - square footage of land
3. sqftBasement - square footage of basement
4. numBedRooms - number of bed rooms
5. numFloors - number of floors
6. builtYear - year of construction
7. grade - construction quality ranked from 1 to 4 where 1 is the lowest grade
8. waterFront- whether the house has a waterfront (1) or not (0)
9. condition - condition of the house (3=Excellent, 2=Good, 1=Average)
10. interaction - interaction variable made using sqftLiving and numBathRooms

Forward selection using AIC value(Feature Selection)

```
## Start:  AIC=98501.69
## soldPrice ~ 1
##
##                Df  Sum of Sq        RSS   AIC
## + interaction   1 1.8524e+14 1.7315e+14 95665
## + grade         1 9.4518e+13 2.6387e+14 97309
## + sqftBasement  1 4.6739e+13 3.1165e+14 97958
## + numFloors     1 4.2844e+13 3.1554e+14 98007
## + numBedRooms   1 3.8402e+13 3.1998e+14 98061
## + builtYear     1 4.5241e+12 3.5386e+14 98454
## + waterFront    1 2.9296e+12 3.5546e+14 98472
## + sqftLand      1 1.9643e+12 3.5642e+14 98482
## + condition     1 2.9551e+11 3.5809e+14 98500
## <none>                       3.5839e+14 98502
##
## Step:  AIC=95665.11
## soldPrice ~ interaction
##
##                Df  Sum of Sq        RSS   AIC
## + builtYear     1 1.4038e+13 1.5911e+14 95337
## + grade         1 6.7776e+12 1.6637e+14 95511
## + numBedRooms   1 4.3851e+12 1.6876e+14 95567
## + waterFront    1 2.2867e+12 1.7086e+14 95615
## + condition     1 1.4623e+12 1.7168e+14 95634
## + numFloors     1 1.0937e+11 1.7304e+14 95665
## <none>                       1.7315e+14 95665
## + sqftBasement  1 7.7155e+10 1.7307e+14 95665
## + sqftLand      1 6.9311e+10 1.7308e+14 95666
```

```
##
## Step:  AIC=95337.2
## soldPrice ~ interaction + builtYear
##
##                Df  Sum of Sq        RSS   AIC
## + grade         1 1.2887e+13 1.4622e+14 95010
## + numBedRooms   1 2.5598e+12 1.5655e+14 95276
## + waterFront    1 1.8192e+12 1.5729e+14 95294
## + numFloors     1 1.5162e+12 1.5759e+14 95302
## + condition     1 1.8533e+11 1.5892e+14 95335
## + sqftBasement  1 1.4964e+11 1.5896e+14 95336
## <none>                       1.5911e+14 95337
## + sqftLand      1 3.2175e+10 1.5908e+14 95338
##
## Step:  AIC=95009.63
## soldPrice ~ interaction + builtYear + grade
##
##                Df  Sum of Sq        RSS   AIC
## + numBedRooms   1 3.7188e+12 1.4250e+14 94911
## + waterFront    1 1.8420e+12 1.4438e+14 94962
## + numFloors     1 1.1829e+12 1.4504e+14 94980
## + sqftBasement  1 3.0116e+11 1.4592e+14 95004
## + condition     1 2.0196e+11 1.4602e+14 95006
## <none>                       1.4622e+14 95010
## + sqftLand      1 1.9609e+09 1.4622e+14 95012
##
## Step:  AIC=94911.11
## soldPrice ~ interaction + builtYear + grade + numBedRooms
##
##                Df  Sum of Sq        RSS   AIC
## + waterFront    1 1.6486e+12 1.4086e+14 94868
## + numFloors     1 9.0045e+11 1.4160e+14 94888
## + condition     1 4.0701e+11 1.4210e+14 94902
## <none>                       1.4250e+14 94911
## + sqftBasement  1 2.4913e+10 1.4248e+14 94912
## + sqftLand      1 6.7805e+09 1.4250e+14 94913
##
## Step:  AIC=94867.7
## soldPrice ~ interaction + builtYear + grade + numBedRooms + waterFront
##
##                Df  Sum of Sq        RSS   AIC
## + numFloors     1 8.1830e+11 1.4004e+14 94847
## + condition     1 3.9672e+11 1.4046e+14 94859
## <none>                       1.4086e+14 94868
## + sqftBasement  1 1.9282e+10 1.4084e+14 94869
## + sqftLand      1 8.3340e+09 1.4085e+14 94869
##
## Step:  AIC=94846.97
## soldPrice ~ interaction + builtYear + grade + numBedRooms + waterFront +
##     numFloors
##
##                Df  Sum of Sq        RSS   AIC
## + condition     1 5.1845e+11 1.3952e+14 94834
## <none>                       1.4004e+14 94847
```

```
## + sqftBasement  1 1.1380e+10 1.4003e+14 94849
## + sqftLand      1 3.6200e+09 1.4003e+14 94849
##
## Step:  AIC=94834.49
## soldPrice ~ interaction + builtYear + grade + numBedRooms + waterFront +
##     numFloors + condition
##
##                Df  Sum of Sq        RSS   AIC
## <none>                       1.3952e+14 94834
## + sqftLand      1 3685749841 1.3951e+14 94836
## + sqftBasement  1 2358105794 1.3952e+14 94836


##
## Call:
## lm(formula = soldPrice ~ interaction + builtYear + grade + numBedRooms +
##     waterFront + numFloors + condition, data = train)
##
## Coefficients:
## (Intercept)  interaction    builtYear        grade  numBedRooms   waterFront
##  5719439.60        64.81     -2905.78    142860.34    -42973.16    322748.23
##   numFloors    condition
##    41518.19     17956.15
```

When comparing models fitted by maximum likelihood to the same data, the smaller the AIC, the better the fit. SO the final model suggested here contains the most relevant predictor variables with respect to the response.And according to that the model contain all variables except sqftBasement and sqftLAnd.

Evaluating and validation of the Final Model

Summary of the final model for the trained data set

```
##
## Call:
## lm(formula = soldPrice ~ interaction + builtYear + grade + numBedRooms +
##     waterFront + numFloors + condition, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1387400   -98988   -11056    74114  2833283
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.719e+06  2.595e+05  22.041  < 2e-16 ***
## interaction  6.481e+01  1.284e+00  50.483  < 2e-16 ***
## builtYear   -2.906e+03  1.349e+02 -21.539  < 2e-16 ***
## grade        1.429e+05  7.296e+03  19.582  < 2e-16 ***
## numBedRooms -4.297e+04  4.361e+03  -9.853  < 2e-16 ***
## waterFront   3.227e+05  4.909e+04   6.575 5.52e-11 ***
## numFloors    4.152e+04  8.106e+03   5.122 3.17e-07 ***
## condition    1.796e+04  4.720e+03   3.804 0.000145 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 189300 on 3894 degrees of freedom
## Multiple R-squared:  0.6107, Adjusted R-squared:   0.61
## F-statistic: 872.7 on 7 and 3894 DF,  p-value: < 2.2e-16
```

This is the summary of the final model for the trained data set

multiple R squared value: 0.6107 suggest that 61.07% of variability of sold prices of the houses are explained by the model.
F-Statistic: 872.7(high f value suggest the strength of the model)

Let us carry out a Partial F-Test to see whether the reduced model is adequate

let the reduced model be the final.model
and full model be:

soldPrice ~ interaction + builtYear + grade + waterFront + numBedRooms + numFloors + condition + sqftBasement +sqftLand

H(Null):Reduced model is adequate
H(Alternative):Reduced model is not adequate

Summary of the full model

```
##
## Call:
## lm(formula = soldPrice ~ interaction + builtYear + grade + waterFront +
##     numBedRooms + numFloors + condition + sqftBasement + sqftLand,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1389450   -98728   -10593    73866  2835559
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.708e+06  2.620e+05  21.790  < 2e-16 ***
## interaction   6.472e+01  1.416e+00  45.707  < 2e-16 ***
## builtYear    -2.900e+03  1.362e+02 -21.289  < 2e-16 ***
## grade         1.427e+05  7.307e+03  19.529  < 2e-16 ***
## waterFront    3.229e+05  4.910e+04   6.577 5.44e-11 ***
## numBedRooms  -4.320e+04  4.439e+03  -9.731  < 2e-16 ***
## numFloors     4.196e+04  8.405e+03   4.992 6.23e-07 ***
## condition     1.786e+04  4.737e+03   3.771 0.000165 ***
## sqftBasement  2.340e+00  9.715e+00   0.241 0.809676
## sqftLand     -2.178e-02  7.064e-02  -0.308 0.757868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 189300 on 3892 degrees of freedom
## Multiple R-squared:  0.6107, Adjusted R-squared:  0.6098
## F-statistic: 678.4 on 9 and 3892 DF,  p-value: < 2.2e-16
```

Anova table

```
## Analysis of Variance Table
##
## Model 1: soldPrice ~ interaction + builtYear + grade + numBedRooms + waterFront +
##     numFloors + condition
## Model 2: soldPrice ~ interaction + builtYear + grade + waterFront + numBedRooms +
##     numFloors + condition + sqftBasement + sqftLand
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1   3894 1.3952e+14
## 2   3892 1.3951e+14  2 5765303489 0.0804 0.9227
```

When we look at the summary of the full model we can clearly see sqftBasement and sqftLand are not significant and also the adjusted R squared of reduced model is slightly greater than that of full model, further the partial F-Test suggest that the null hypothesis cannot be rejected(p=0.9927) and thus the reduced model is adequate.

Correlation matrix for the final model

```
##             soldPrice numBedRooms numFloors builtYear grade waterFront
## soldPrice        1.00        0.33      0.36      0.12  0.51       0.19
## numBedRooms      0.33        1.00      0.27      0.33  0.39      -0.01
## numFloors        0.36        0.27      1.00      0.37  0.32       0.05
## builtYear        0.12        0.33      0.37      1.00  0.40       0.00
## grade            0.51        0.39      0.32      0.40  1.00       0.04
## waterFront       0.19       -0.01      0.05      0.00  0.04       1.00
## condition        0.02        0.03     -0.15     -0.21 -0.07       0.02
## interaction      0.72        0.58      0.48      0.41  0.55       0.07
##             condition interaction
## soldPrice        0.02        0.72
## numBedRooms      0.03        0.58
## numFloors       -0.15        0.48
## builtYear       -0.21        0.41
## grade           -0.07        0.55
## waterFront       0.02        0.07
## condition        1.00       -0.06
## interaction     -0.06        1.00
```

And the correlation matrix suggest that the previously faced issues of multicollinearity is solved and the final model is fit for use.

```
predicted<-predict(final.model,test)
R2_Score(predicted,test$soldPrice)
```

```
## [1] 0.6159655
```

R squared value for the test data set is 0.6159655 that is 61.59% of variability of sold prices of the houses are explained by the model when data are not obtained from the trained data set.

Evaluation of the test set

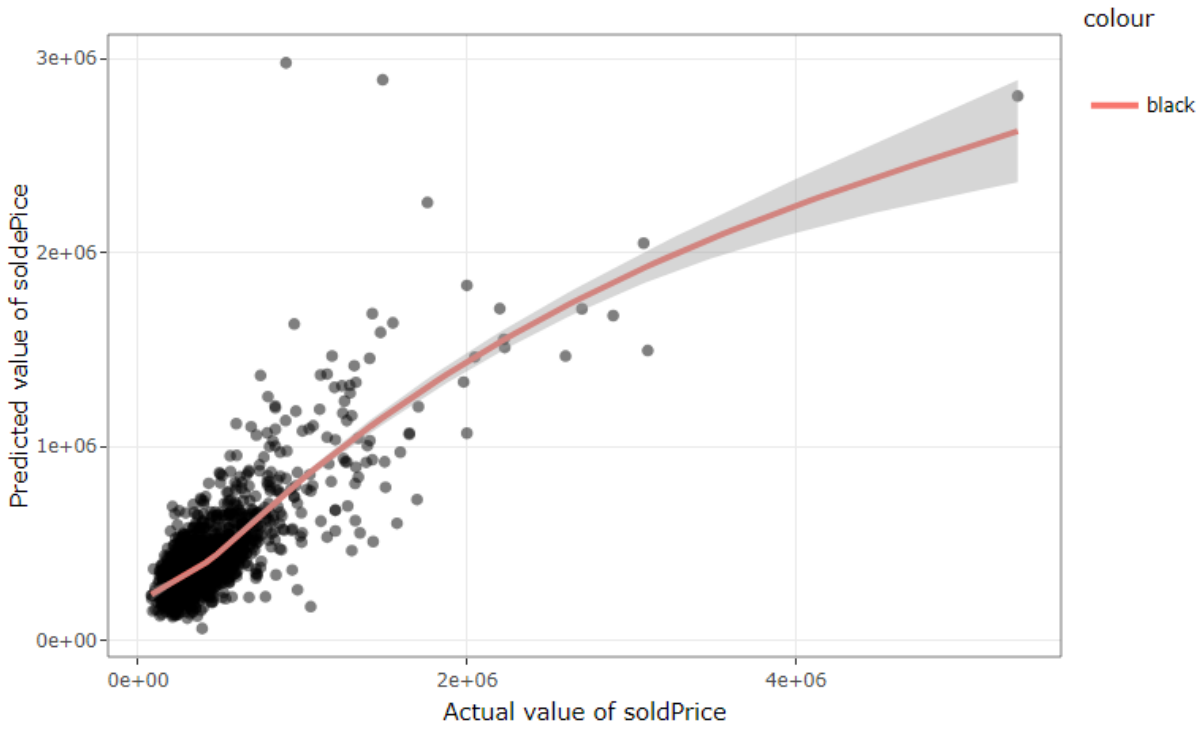


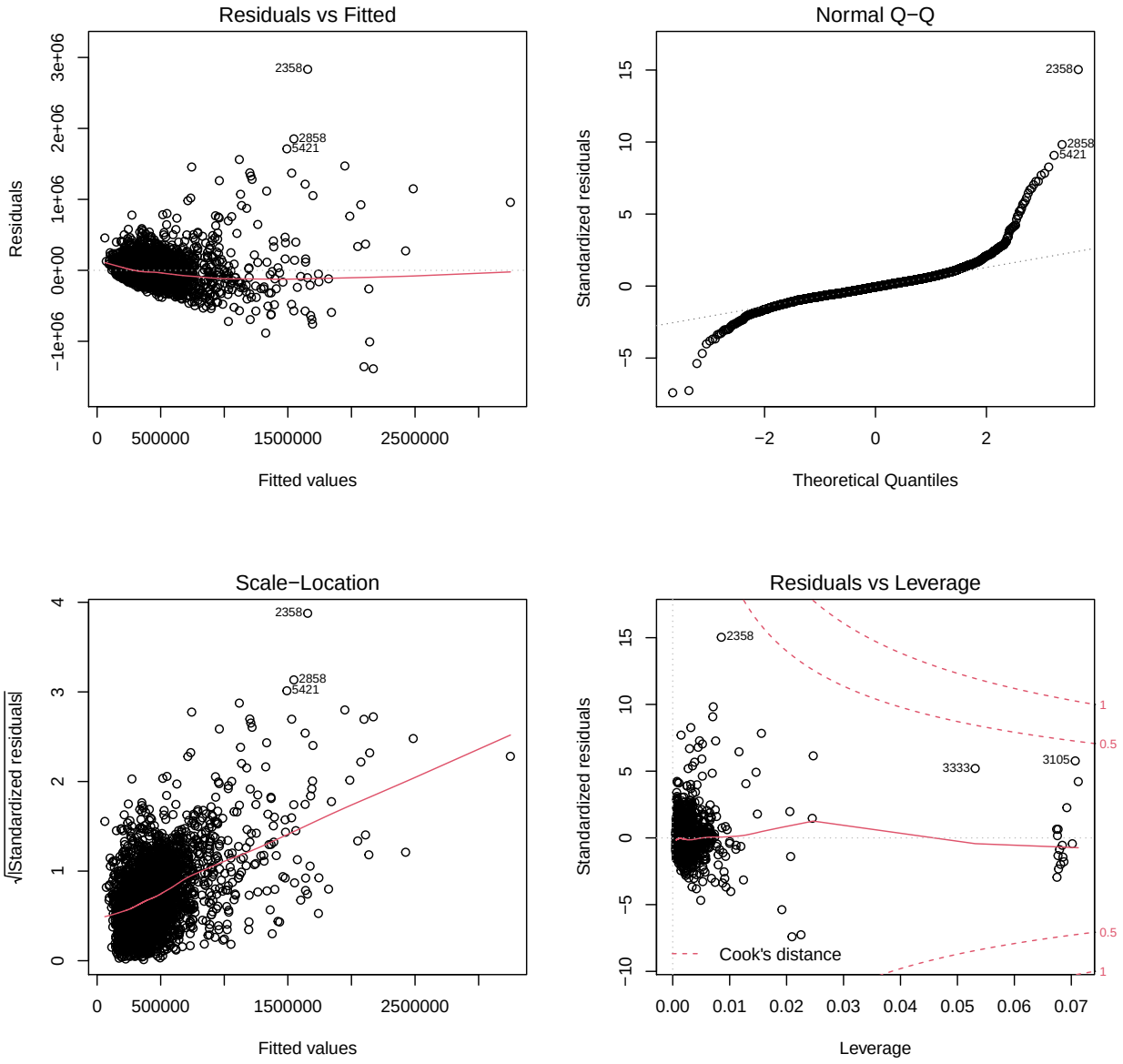
Figure 1: Obseved vs Fitted values(interactive image).

soldPrice	.fitted	.resid	.std.resid	.hat	.sigma	.cooksd
221900	331385.8	-109485.832	-0.5786394	0.0007725	189301.8	0.0000324
604000	599879.8	4120.168	0.0217901	0.0021209	189310.0	0.0000001
229500	355743.6	-126243.641	-0.6671693	0.0006642	189299.2	0.0000370
650000	757294.0	-107294.041	-0.5673007	0.0016358	189302.2	0.0000659
395000	428087.6	-33087.612	-0.1749356	0.0015205	189309.2	0.0000058
189000	492408.0	-303408.000	-1.6040673	0.0014418	189247.4	0.0004644

Gives an idea on the fitted, residuals and some other information when we plot observed data to the final model.

Residual Analysis

Diagnostic plots



1.**Residual vs Fitted plot:** No distinct pattern can be seen and the red line being horizontal indicates that we can assume there is a linear relationship between predictors and response variables

2.**Normal Q-Q plot:**According to the plot, majority of the points fall approximately along the reference line eventhough there is a slight deviation at the end points so we can assume normality.

3.**Scale-Location plot:**The plot shows that the variances of the residual points increases with the value of the fitted outcome suggesting non constant variance in the residuals(heteroscedasticity).

4.**Residual vs Leverage plot:** In this plot there are no points outside of the dashed line(Threshold value) which say that there are no influential points,hence all the points are included in the fitted model.

Discussion and Conclusion

A multiple linear regression model was fitted to predict house prices in relation to 11 predictor variables mentioned above. The final model includes all the necessary predictor variables that significantly impact the house prices.

Feature selection was done based on the AIC values so that the best features are added to the model.

And according to that the model we came up with was(Parsimonious model);

```
##   (Intercept)   interaction     builtYear         grade   numBedRooms
## 5719439.59947      64.81121   -2905.78340  142860.34053  -42973.15673
##    waterFront     numFloors     condition
##  322748.23289   41518.18648   17956.15179
```

soldPrice = 5719439.59947 + (64.81121)interaction - (2905.78340)builtYear+ (142860.3405)grade - (42973.15673)numBedRooms + (322748.23289)waterFront + (41518.1864)numFloors + (17956.15179)condition

The final model suggest that 61.59% of variability of sold prices of the houses are explained by the model and further high F values tells us the strength of the model

There were several key adjustments made such that the final model will give us reliable outputs like, handling the issue of multicollinearity by removing certain predictor variables(sqftAbove) and adding an interaction term so that we reduce VIF values a measure of multicollinearity.

Validation of the fitted model was done by

1.Looking at the correlations of the final model

2.Checking for validation of linear regression assumptions(Residual Analysis)

3.A partial F-test was carried out

4.Used the test data set to see the strength and linear relationship of variables how fitted and observed values lie.

The model predicts the house prices with a R2 score of 61.59% which is an average model this can be improved using feature extraction, rebuilding and training the model. The residual plots like Normal QQ and scale location plots were not satisfactory these can be adjusted through transformation of variables for eg:getting the log values of response variable.

further without using linear models; models like polynomial regression or lasso regression can be used appropriately.