Property Sales

Sourav Dutta

4/6/2021

Load Required Packages

```
library(ggplot2)
library(corrplot)
#library(plyr)
#library(dplyr)
#library(caret)
#library(car)
#library(Rmisc)
#library(leaps)
#library(MASS)
#library(psych)
```

Load the data file in R Envoronment

```
getwd()
```

```
## [1] "/Users/souravdutta/Downloads"
```

```
property <- read.csv("/Users/souravdutta/Downloads/property-sales.csv")
head(property,5)
```

```
##   MSZoning LotArea BldgType HouseStyle OverallQual OverallCond YearBuilt
## 1       RL    8450     1Fam     2Story           7           5      2003
## 2       RL    9600     1Fam     1Story           6           8      1976
## 3       RL   11250     1Fam     2Story           7           5      2001
## 4       RL    9550     1Fam     2Story           7           5      1915
## 5       RL   14260     1Fam     2Story           8           5      2000
##   CentralAir GrLivArea FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
## 1          Y      1710        2        1            3            1          Gd
## 2          Y      1262        2        0            3            1          TA
## 3          Y      1786        2        1            3            1          Gd
## 4          Y      1717        1        0            3            1          Gd
## 5          Y      2198        2        1            4            1          Gd
##   Fireplace GarageArea SaleCondition SalePrice
## 1         N        548        Normal    208500
## 2         Y        460        Normal    181500
## 3         Y        608        Normal    223500
## 4         Y        642       Abnorml    140000
## 5         Y        836        Normal    250000
```

Question 1: Explore the dataset. #### Column Names

```
colnames(property)
```

```
##  [1] "MSZoning"      "LotArea"       "BldgType"      "HouseStyle"   
##  [5] "OverallQual"   "OverallCond"   "YearBuilt"     "CentralAir"   
##  [9] "GrLivArea"     "FullBath"      "HalfBath"      "BedroomAbvGr" 
## [13] "KitchenAbvGr"  "KitchenQual"   "Fireplace"     "GarageArea"   
## [17] "SaleCondition" "SalePrice"
```

As per the document, we have 18 columns in our Dataset.

Structure of DataSet

```
str(property)
```

```
## 'data.frame':    1460 obs. of  18 variables:
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ Fireplace    : chr  "N" "Y" "Y" "Y" ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129
900 118000 ...
```

```
dim(property)
```

```
## [1] 1460   18
```

Our Dataset has **1460** rows and *18 columns

Missing Data

```
colSums(is.na(property))
```

```
##      MSZoning       LotArea      BldgType    HouseStyle   OverallQual 
##             0             0             0             0             0 
##   OverallCond     YearBuilt    CentralAir     GrLivArea      FullBath 
##             0             0             0             0             0 
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual     Fireplace 
##             0             0             0             0             0 
##    GarageArea SaleCondition     SalePrice 
##             0             0             0
```

The dataset is clean and does not have any missing values.

Summary Statistics

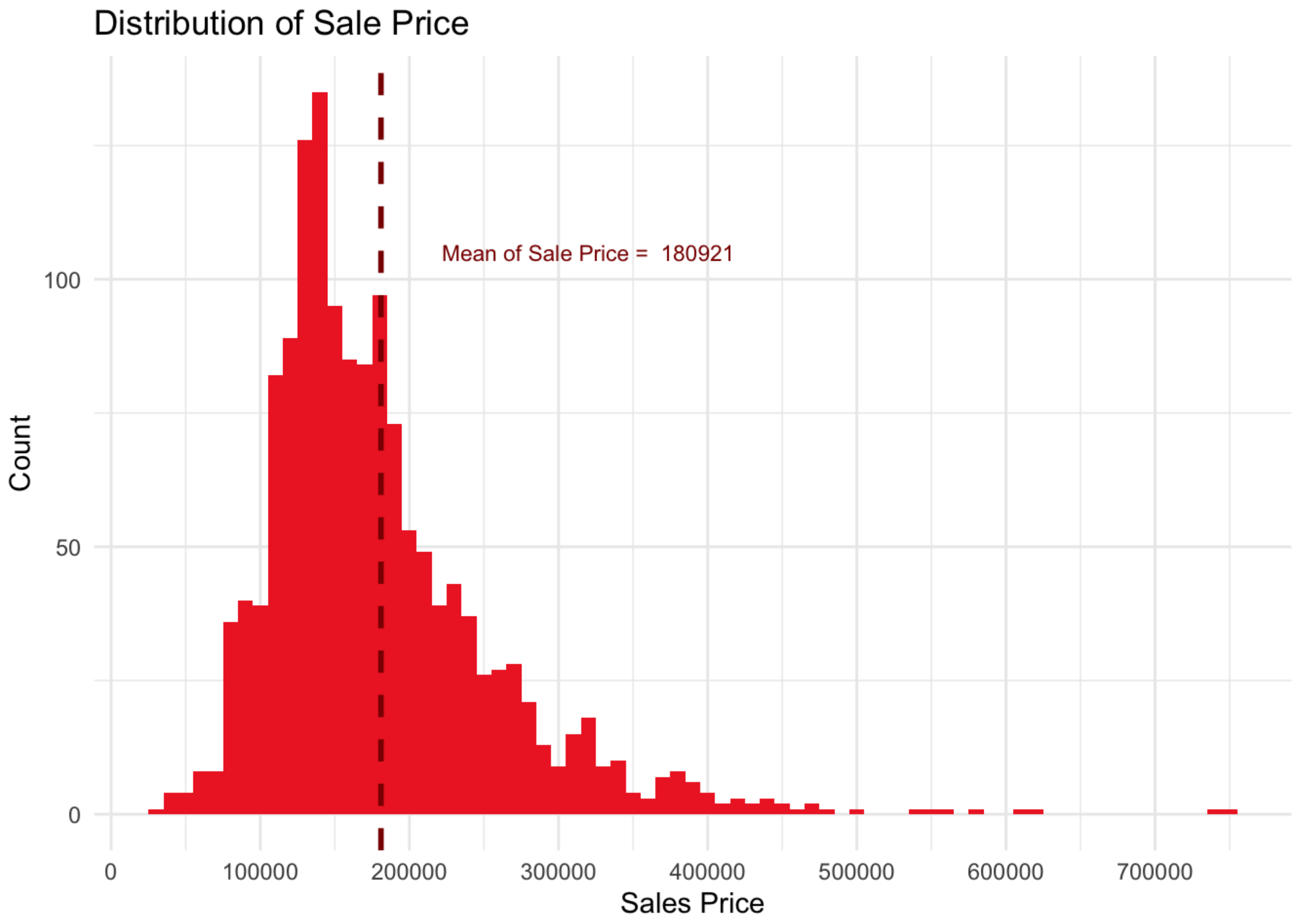
```
summary(property)
```

```
##    MSZoning            LotArea          BldgType          HouseStyle       
##  Length:1460        Min.   :  1300   Length:1460        Length:1460       
##  Class :character   1st Qu.:  7554   Class :character   Class :character  
##  Mode  :character   Median :  9478   Mode  :character   Mode  :character  
##                     Mean   : 10517                                        
##                     3rd Qu.: 11602                                        
##                     Max.   :215245                                        
##   OverallQual      OverallCond      YearBuilt     CentralAir       
##  Min.   : 1.000   Min.   :1.000   Min.   :1872   Length:1460       
##  1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   Class :character  
##  Median : 6.000   Median :5.000   Median :1973   Mode  :character  
##  Mean   : 6.099   Mean   :5.575   Mean   :1971                     
##  3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000                     
##  Max.   :10.000   Max.   :9.000   Max.   :2010                     
##    GrLivArea       FullBath        HalfBath        BedroomAbvGr  
##  Min.   : 334   Min.   :0.000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:1130   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000  
##  Median :1464   Median :2.000   Median :0.0000   Median :3.000  
##  Mean   :1515   Mean   :1.565   Mean   :0.3829   Mean   :2.866  
##  3rd Qu.:1777   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000  
##  Max.   :5642   Max.   :3.000   Max.   :2.0000   Max.   :8.000  
##   KitchenAbvGr   KitchenQual         Fireplace           GarageArea    
##  Min.   :0.000   Length:1460        Length:1460        Min.   :   0.0  
##  1st Qu.:1.000   Class :character   Class :character   1st Qu.: 334.5  
##  Median :1.000   Mode  :character   Mode  :character   Median : 480.0  
##  Mean   :1.047                                         Mean   : 473.0  
##  3rd Qu.:1.000                                         3rd Qu.: 576.0  
##  Max.   :3.000                                         Max.   :1418.0  
##  SaleCondition        SalePrice     
##  Length:1460        Min.   : 34900  
##  Class :character   1st Qu.:129975  
##  Mode  :character   Median :163000  
##                     Mean   :180921  
##                     3rd Qu.:214000  
##                     Max.   :755000  
```

Exploring Some of the most important variables

Sale Price

```
ggplot(data = property, aes(SalePrice)) +
  geom_histogram(fill = "firebrick2", binwidth = 10000) +
  scale_x_continuous(breaks = seq(0, 800000, by = 100000)) +
  geom_vline(aes(xintercept = mean(SalePrice)), color = "darkred", linetype = "dashe
d", size = 1) +
  annotate("text",
          x = 320000,
          y = 105,
          label = paste("Mean of Sale Price = ", round(mean(property$SalePrice))),
          col = "darkred",
          size = 3) +
  labs(title = "Distribution of Sale Price",
      x = "Sales Price",
      y = "Count") + theme_minimal()
```



As can be seen from the graph, the Sale Price variable is highly skewed which means that only few people can afford very expensive houses, so the majority of houses costs under 300000.

```
summary(property$SalePrice)
```

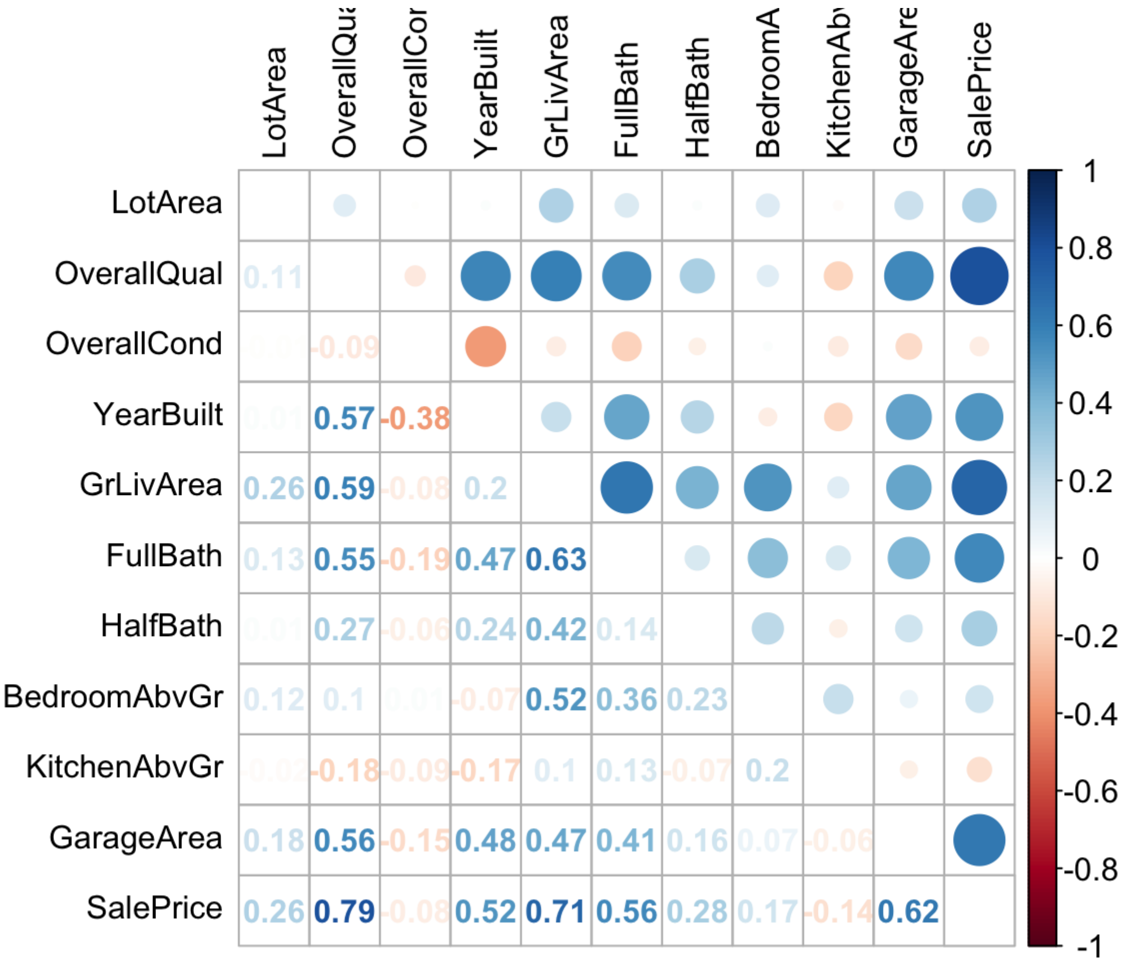
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   34900  129975  163000  180921  214000  755000
```

### Correlations among the variables

```
numericvars <- which(sapply(property, is.numeric))
numericVarNames <- names(numericvars)
cat('There are', length(numericvars), 'numeric Variables')
```

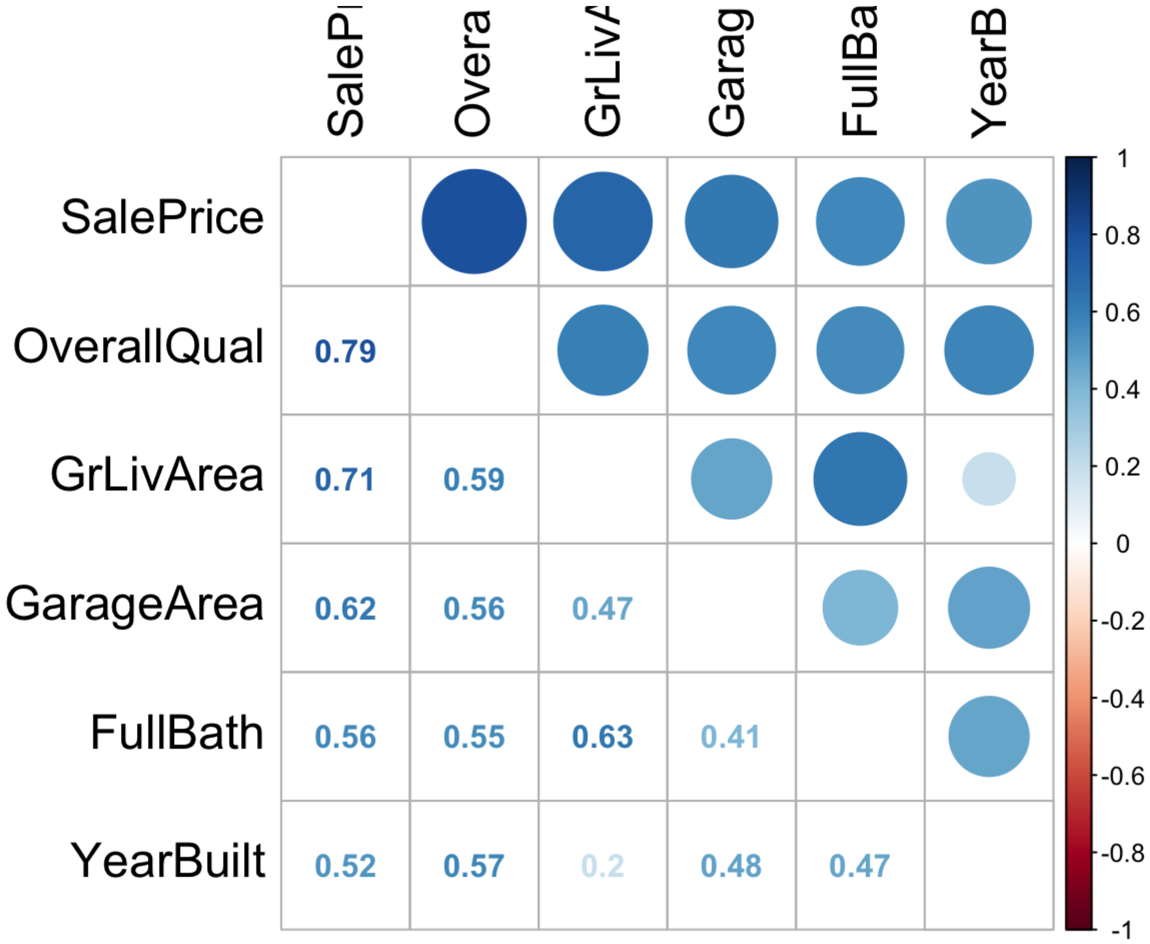
```
## There are 11 numeric Variables
```

```
all_nvar <- property[, numericvars]
cor_nvar <- cor(all_nvar, use = "pairwise.complete.obs")
# Sort on decreasing correlations with SalePrice
corrplot.mixed(cor_nvar, tl.col = "black", tl.pos = "lt",
               tl.cex = 1, cl.cex = 1)
```



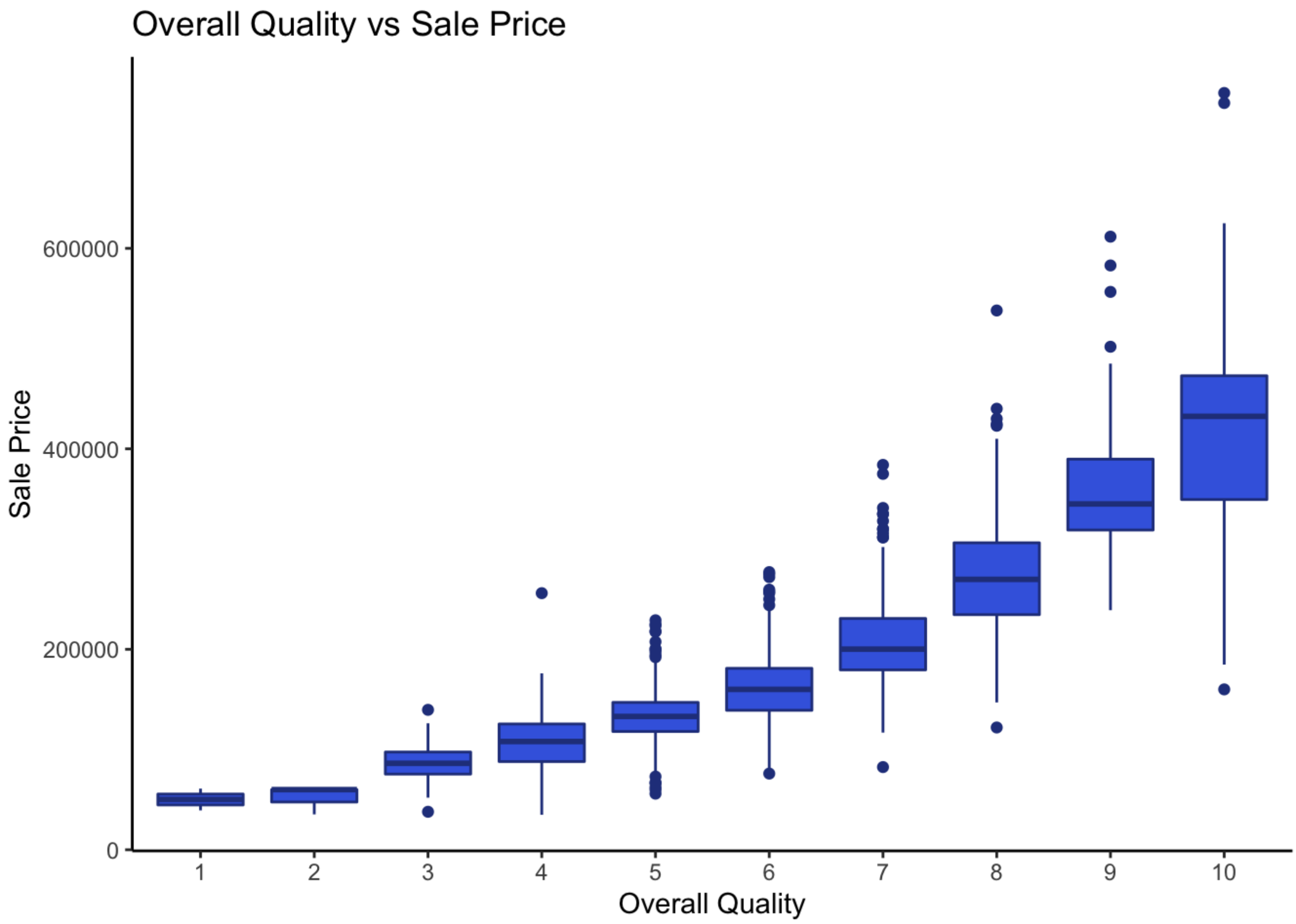
### Correlations with Sale Price

```
# Sort on decreasing correlations with SalePrice
col_sor <- as.matrix(sort(cor_nvar[,'SalePrice'], decreasing = TRUE))
# Select only high correlations with SalePrice
high_cor <- names(which(apply(col_sor, 1, function(x) abs(x) > 0.5)))
cor_nvar <- cor_nvar[high_cor, high_cor]
corrplot.mixed(cor_nvar, tl.col = "black", tl.pos = "lt",
               tl.cex = 1.5, cl.cex = 0.8)
```

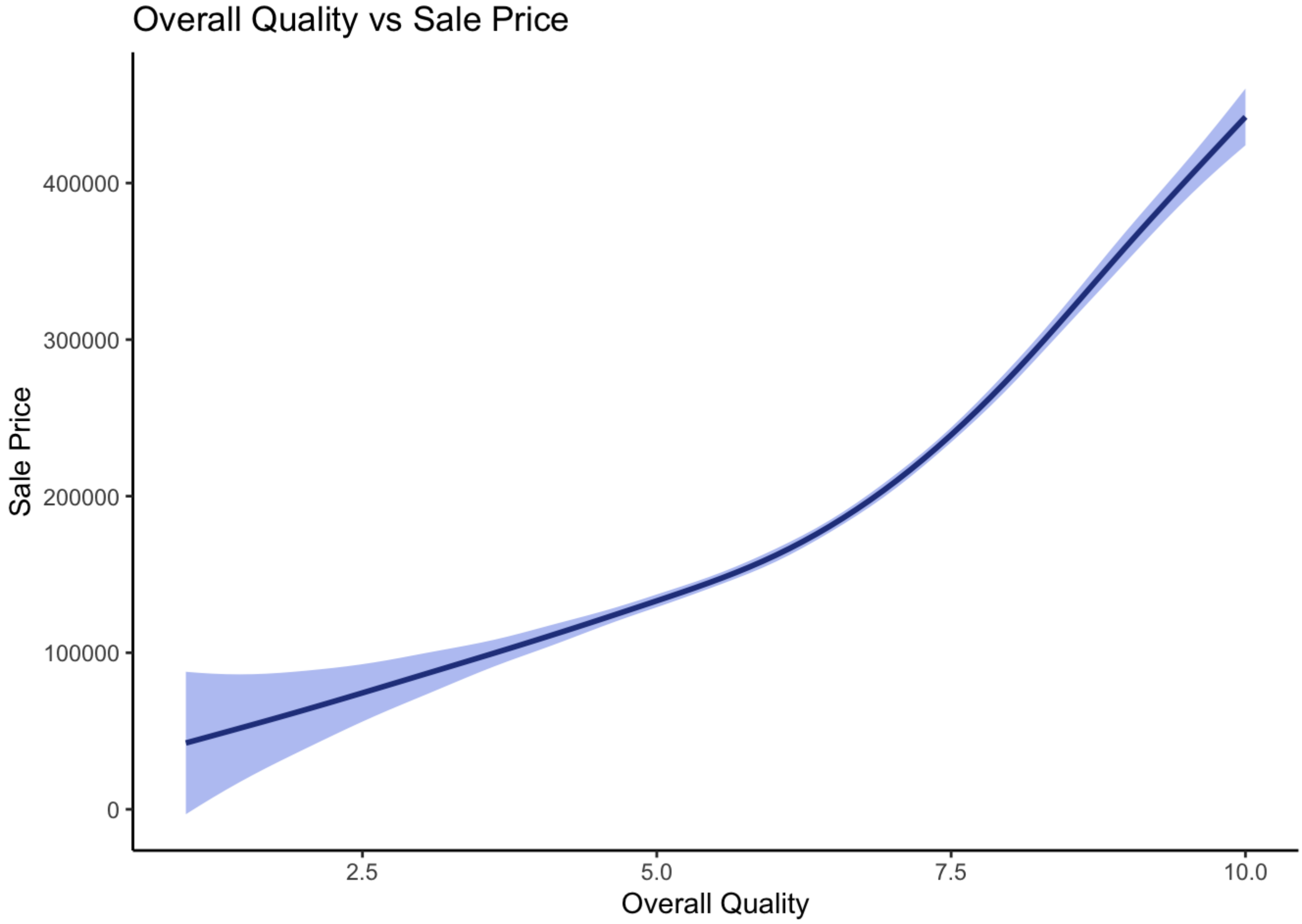


#### Overall Quality

```
ggplot(data = property, aes(x = factor(OverallQual), y = SalePrice)) +
  geom_boxplot(fill = "royalblue", col = "royalblue4") +
  labs(title = "Overall Quality vs Sale Price",
       x = "Overall Quality",
       y = "Sale Price") +
  theme_classic()
```



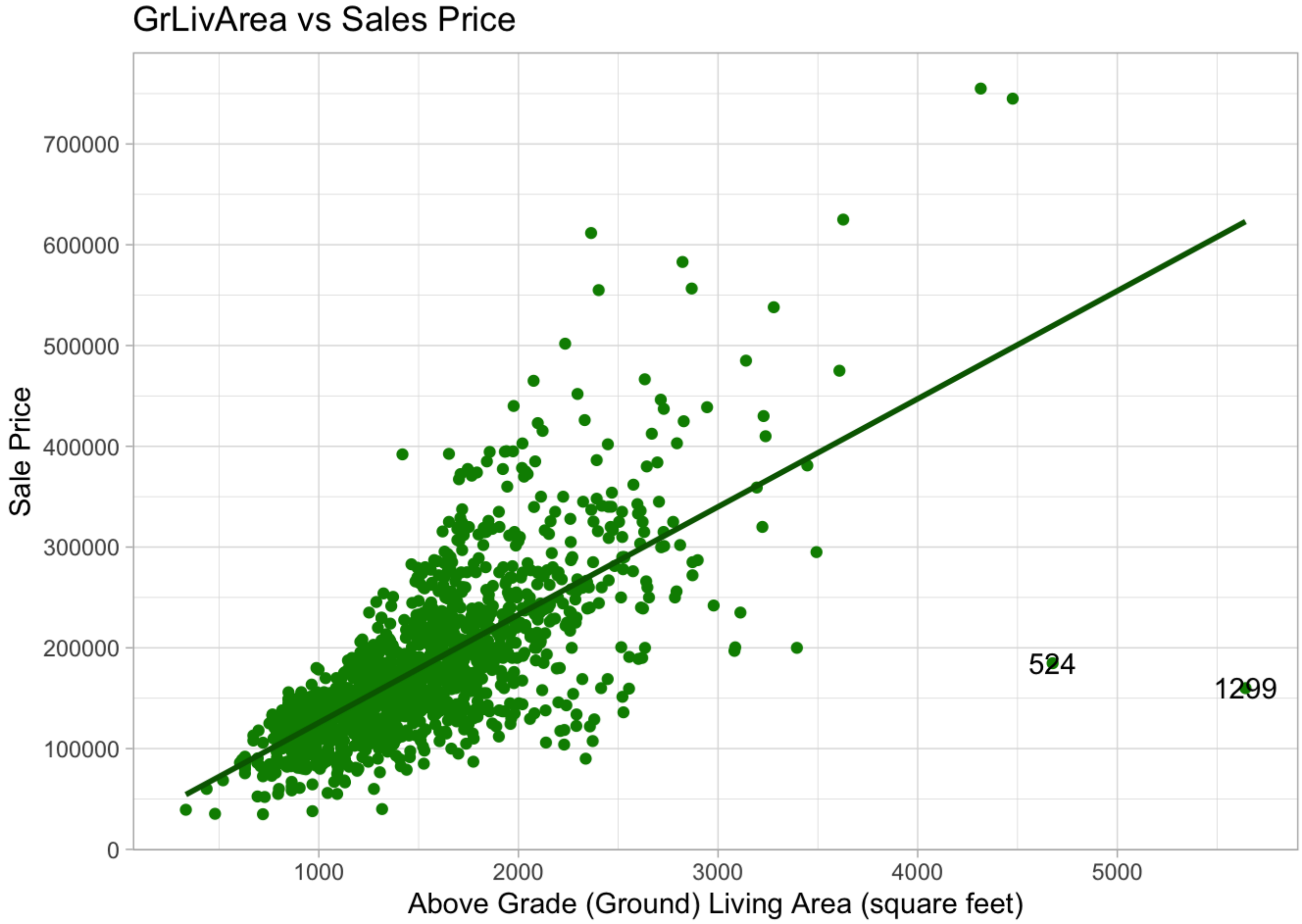
```
ggplot(data = property, aes(x = OverallQual, y = SalePrice)) +
  geom_smooth(fill = "royalblue", col = "royalblue4") +
  labs(title = "Overall Quality vs Sale Price",
       x = "Overall Quality",
       y = "Sale Price") +
  theme_classic()
```



There is an overall increasing trend, with increasing house quality sales price also goes up.

Above Grade (Ground) Living Area (square feet) - GrLivArea

```
ggplot(data = property, aes(x = GrLivArea, y = SalePrice)) +
  geom_point(col = "green4") +
  geom_smooth(method = "lm", se = FALSE, color = "darkgreen", aes(group = 1)) +
  scale_y_continuous(breaks = seq(0, 800000, by = 100000)) +
  geom_text(aes(label = ifelse(property$GrLivArea[!is.na(property$SalePrice)] > 4500,
rownames(property), ''))) +
  labs(title = "GrLivArea vs Sales Price",
       x = "Above Grade (Ground) Living Area (square feet)",
       y = "Sale Price") + theme_light()
```



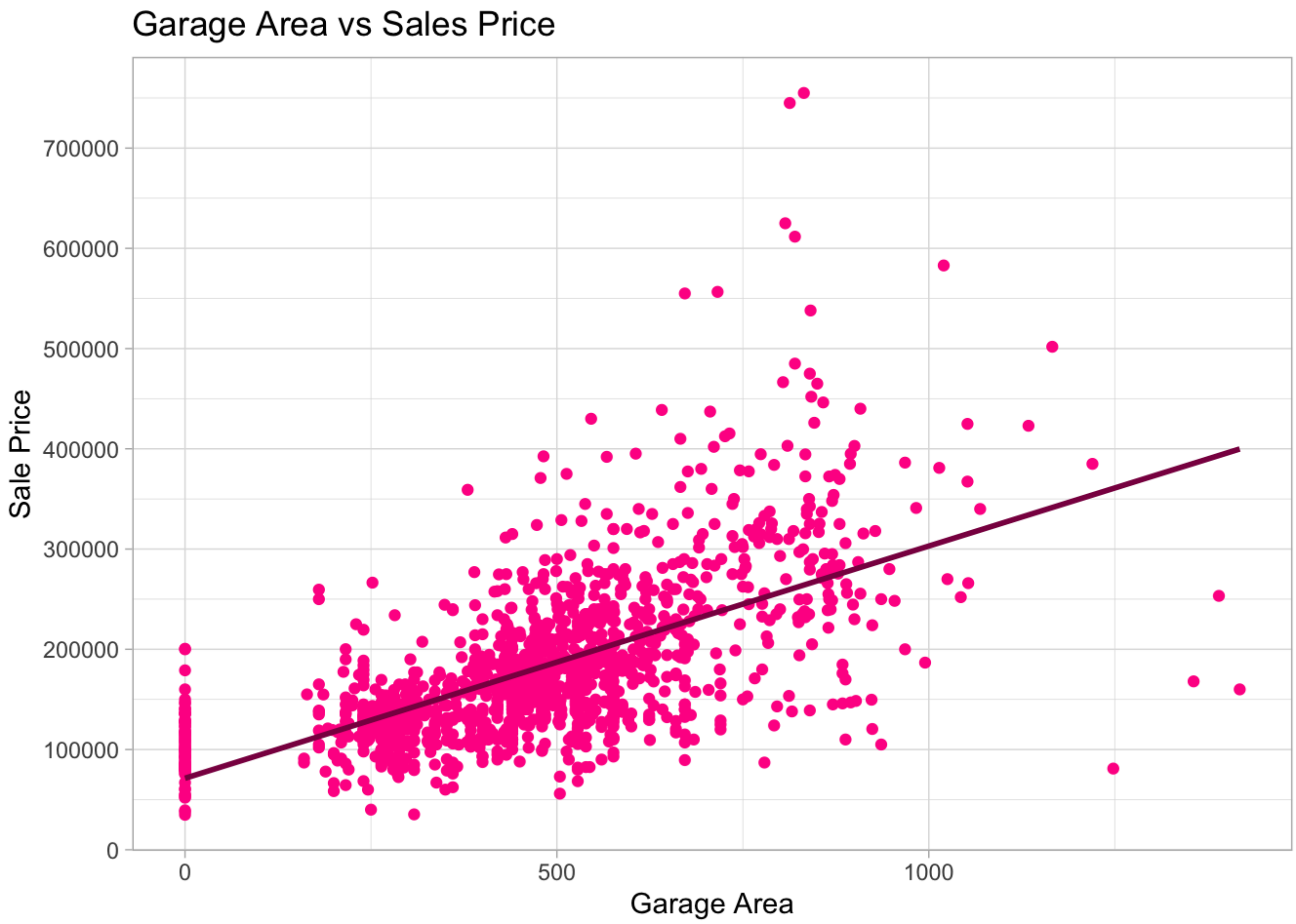
The two houses with really big living area and very low sales price seems like and outlier. Also the Overall Quality can be biased because of its low price. As we have seen that Overall Quality has the highest correlation with Sale Price, bias in Overall Quality might negatively impact the final model.

```
property[c(524, 1299), c('SalePrice', 'GrLivArea', 'OverallQual')]
```

```
##      SalePrice GrLivArea OverallQual
## 524     184750      4676          10
## 1299    160000      5642          10
```

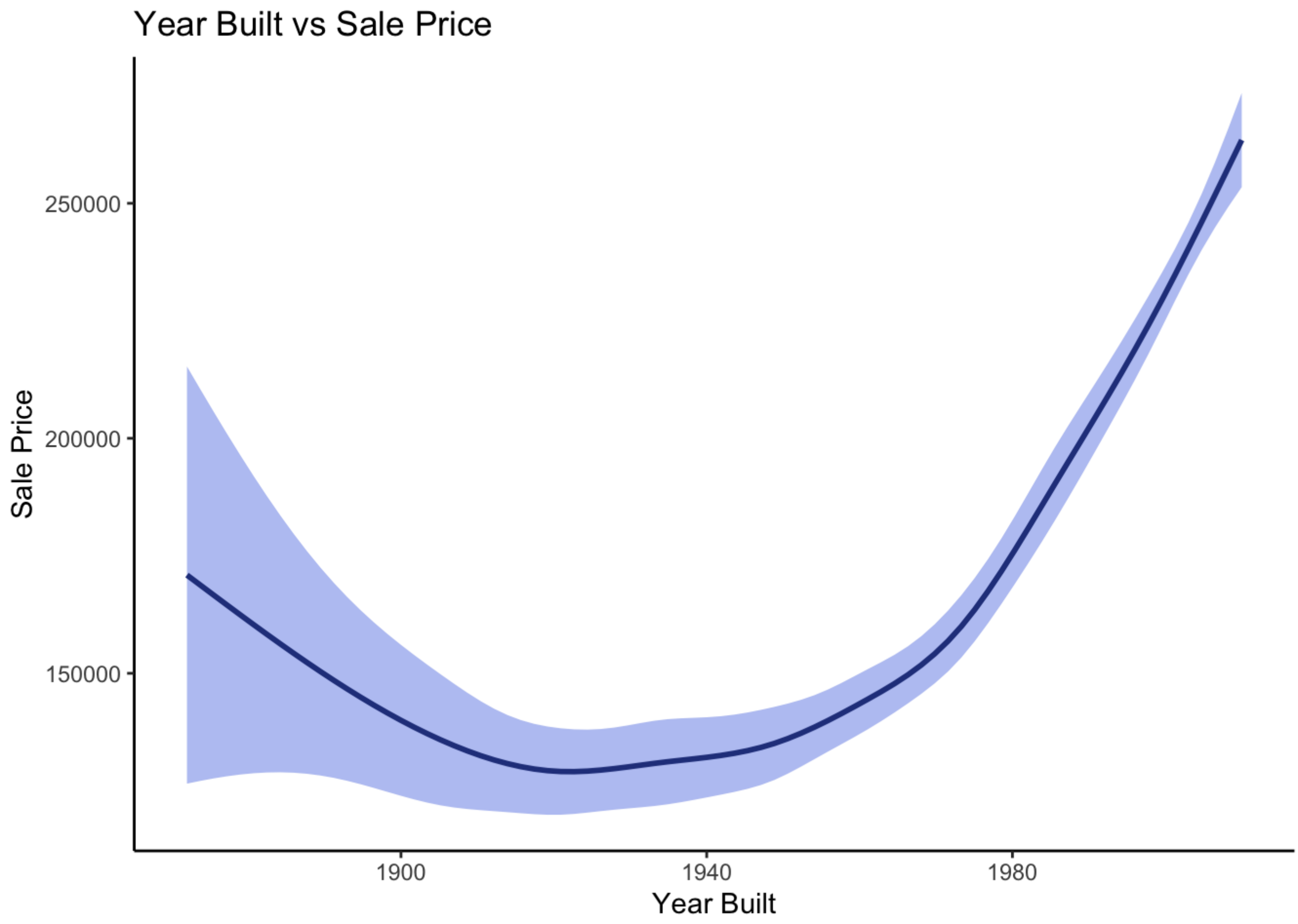
Garage Area vs Sale Price

```
ggplot(data = property, aes(x = GarageArea, y = SalePrice)) +
  geom_point(col = "deeppink") +
  geom_smooth(method = "lm", se = FALSE, color = "deeppink4", aes(group = 1)) +
  scale_y_continuous(breaks = seq(0, 800000, by = 100000)) +
  labs(title = "Garage Area vs Sales Price",
       x = "Garage Area",
       y = "Sale Price") + theme_light()
```



#### Year Built

```
ggplot(data = property, aes(x = YearBuilt, y = SalePrice)) +
  geom_smooth(fill = "royalblue", col = "royalblue4") +
  labs(title = "Year Built vs Sale Price",
       x = "Year Built",
       y = "Sale Price") +
  theme_classic()
```



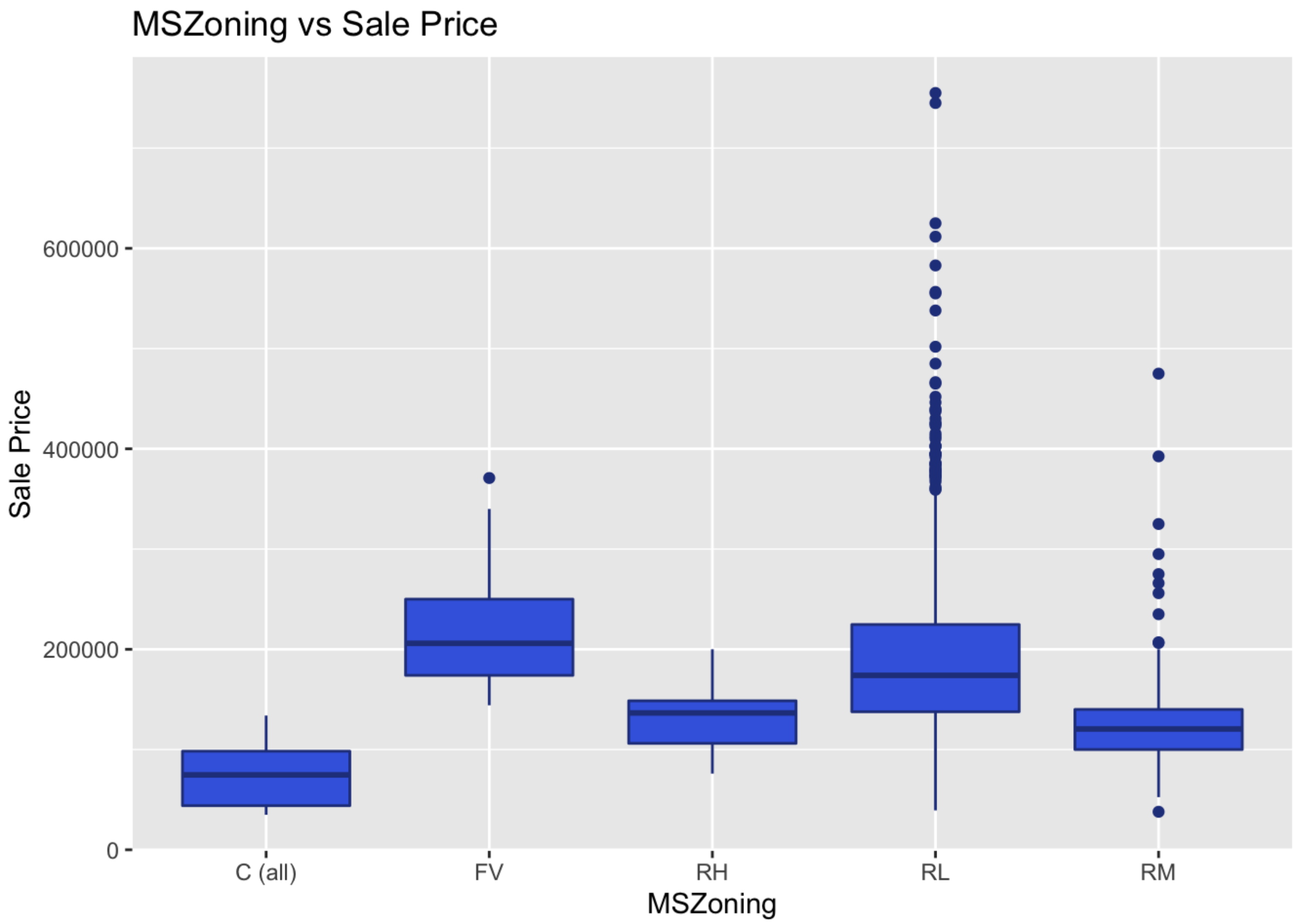
### Categorical Variable with Sale Price

```
cat_vars <- which(sapply(property, is.character))
catVarNames <- names(cat_vars)
cat('There are ', length(cat_vars), 'categorical Variables')
```

```
## There are  7 categorical Variables
```

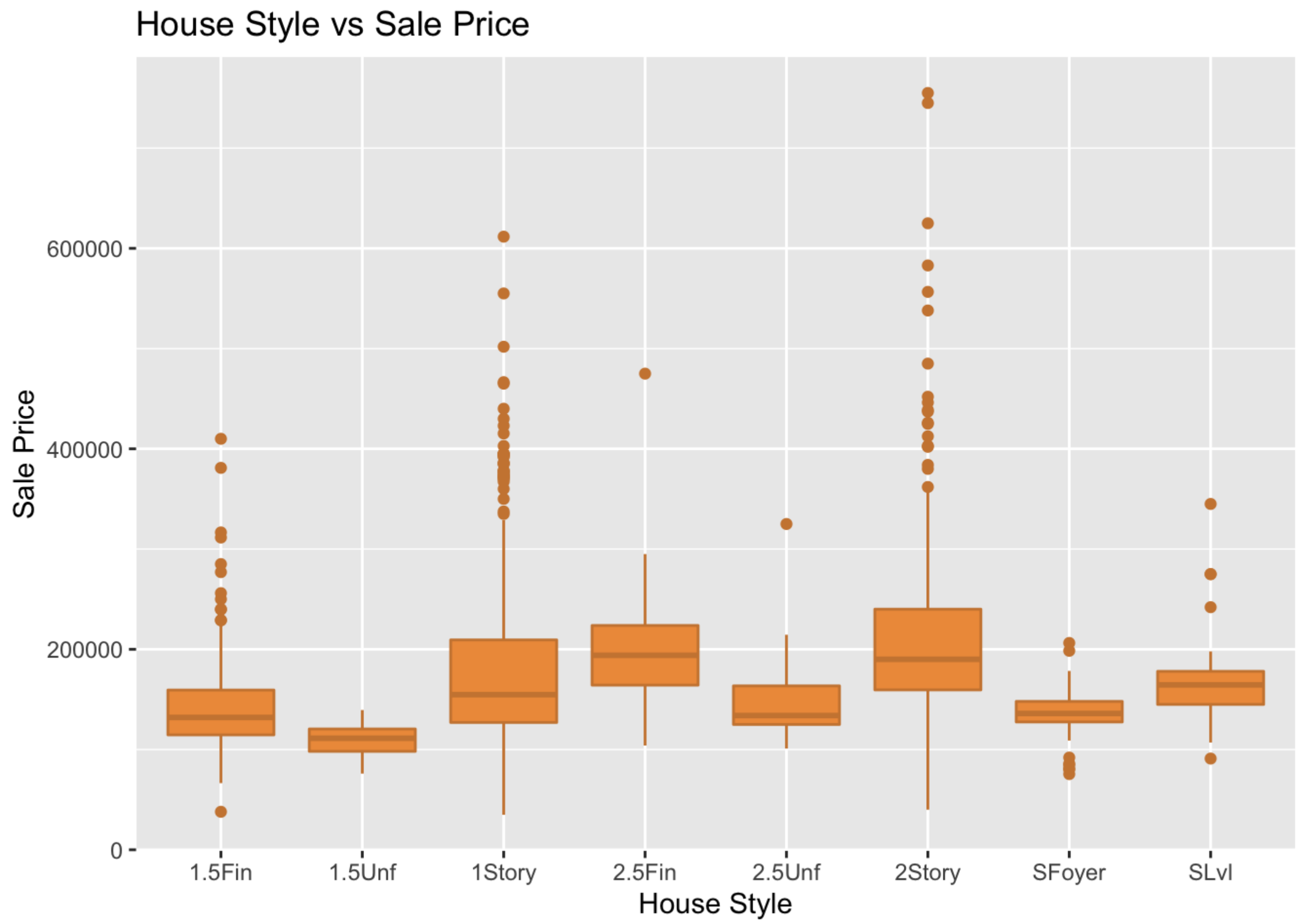
#### Important Categorical Variables

```
ggplot(data = property, aes(x = factor(MSZoning), y = SalePrice)) +
  geom_boxplot(fill = "royalblue", col = "royalblue4") +
  labs(title = "MSZoning vs Sale Price",
       x = "MSZoning",
       y = "Sale Price") + theme_gray()
```



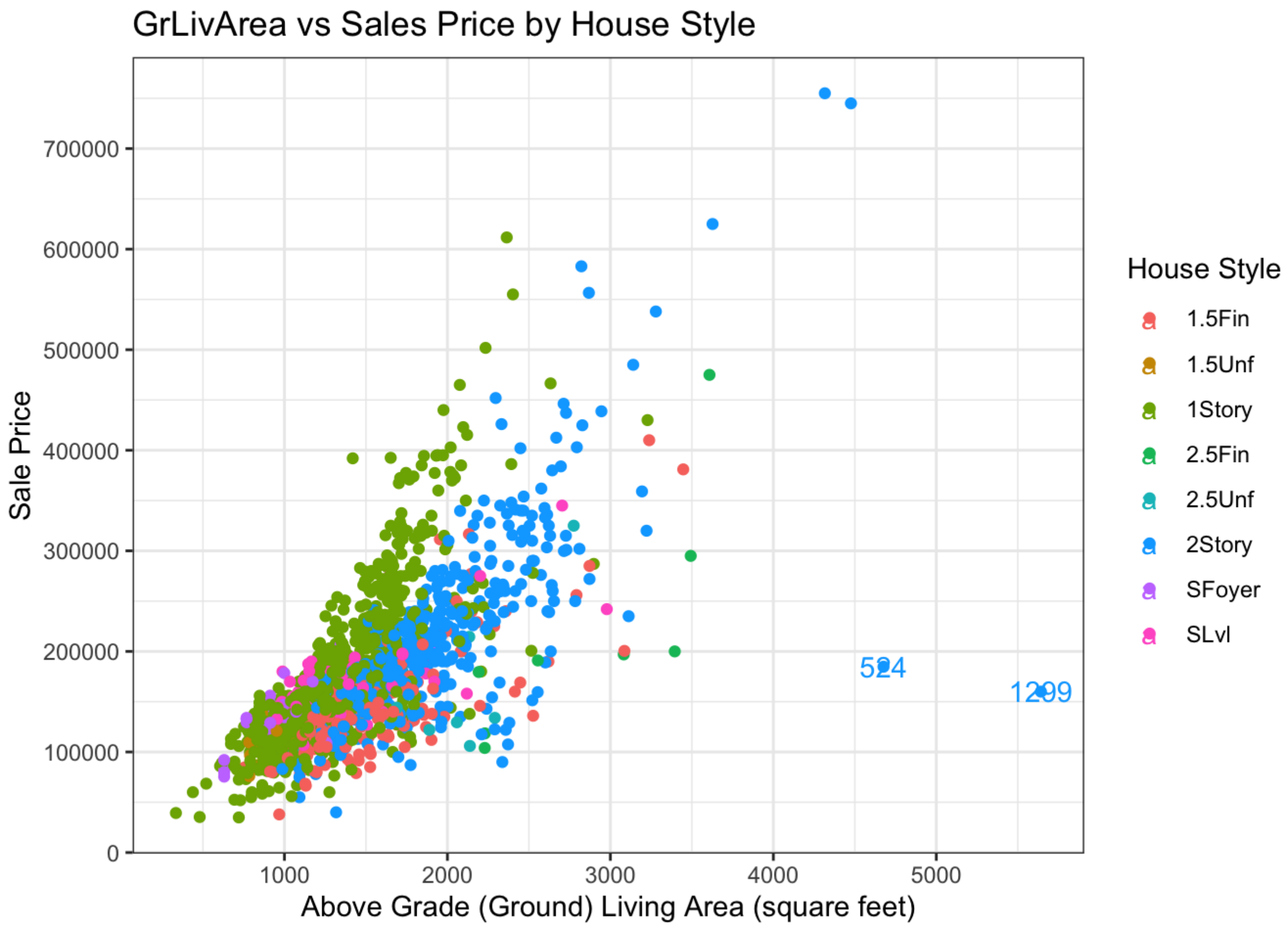
The MSZonig does not necessarily indicates the relationship between SalePrice and MSZoning of the house

```
ggplot(data = property, aes(x = factor(HouseStyle), y = SalePrice)) +
  geom_boxplot(fill = "tan2", col = "tan3") +
  labs(title = "House Style vs Sale Price",
       x = "House Style",
       y = "Sale Price") + theme_gray()
```



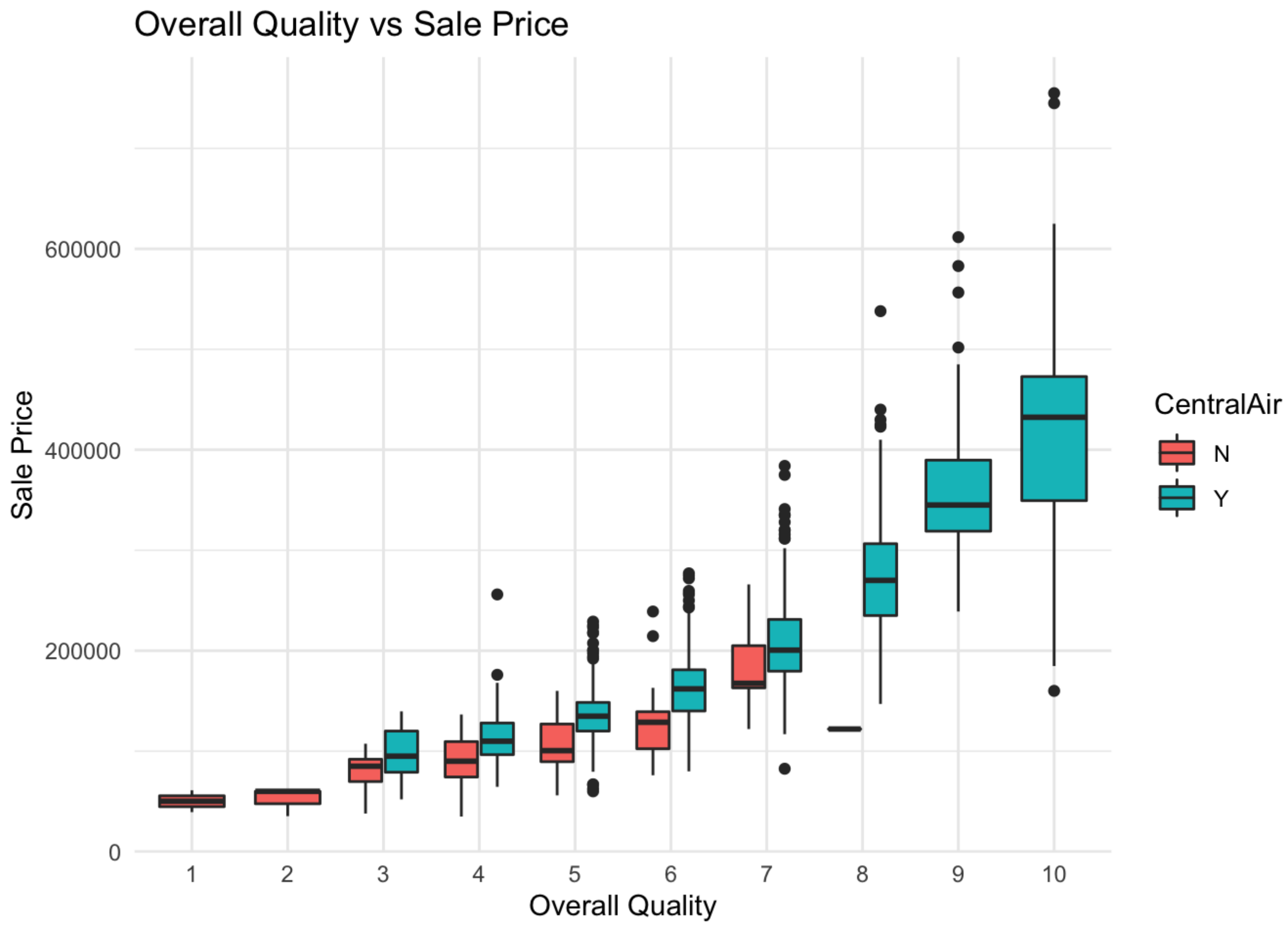
#### Bi-Variate Relationships

```
ggplot(data = property, aes(x = GrLivArea, y = SalePrice, col = factor(HouseStyle)))
+
  geom_point() +
  scale_y_continuous(breaks = seq(0, 800000, by = 100000)) +
  geom_text(aes(label = ifelse(property$GrLivArea[!is.na(property$SalePrice)] > 4500,
rownames(property), ''))) +
  labs(title = "GrLivArea vs Sales Price by House Style",
       x = "Above Grade (Ground) Living Area (square feet)",
       y = "Sale Price",
       col = 'House Style') + theme_bw()
```



The above graph does not show any significant relationship of Sales price by House style but it does give us an insight that 2 story houses, as expected, are more expensive and have more living area. it also gives us the clarity about the outliers here which are also 2 story house with exceptionally large living area but very low sale price.

```
ggplot(data = property, aes(x = factor(OverallQual), y = SalePrice, fill = CentralAi
r)) +
  geom_boxplot() +
  labs(title = "Overall Quality vs Sale Price",
       x = "Overall Quality",
       y = "Sale Price") +
  theme_minimal()
```



The Above graph clearly states that in any quality house, if the house has Central Air conditioning, it is more expensive and as expected as the quality goes up, all the houses has Central Air conditioning.

Question 2: Develop a regression model to predict SalePrice from one or more of the other variables. ### Linear Regression Model We start with the variables having correlation more than 0.5 #### Fit a Linear regression Model with Outliers

```
reg_sales <- lm(SalePrice ~ OverallQual + GrLivArea + GarageArea + FullBath + YearBui
lt, data = property)
summary(reg_sales)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageArea +
##     FullBath + YearBuilt, data = property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -424848  -21012   -2355   17693  295926
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -911065.986   91184.681  -9.991 < 0.0000000000000002 ***
## OverallQual   23242.964    1160.471  20.029 < 0.0000000000000002 ***
## GrLivArea        59.886       3.043  19.683 < 0.0000000000000002 ***
## GarageArea       56.652       6.255   9.058 < 0.0000000000000002 ***
## FullBath      -7174.235    2716.719  -2.641              0.00836 **
## YearBuilt       428.100      48.026   8.914 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39550 on 1454 degrees of freedom
## Multiple R-squared:  0.753,  Adjusted R-squared:  0.7522
## F-statistic: 886.6 on 5 and 1454 DF,  p-value: < 0.00000000000000022
```

We get the Adjusted R square value as 0.7522 which means these variables explain 75.22% variabiltiy in Sale Price.

We check the same model after removing outliers. ##### Removing Outliers

```
property_mod <- property[-c(524, 1299),]
```

#### Linear regression Model without Outliers

```
reg_sales <- lm(SalePrice ~ OverallQual + GrLivArea + GarageArea + FullBath + YearBui
lt, data = property_mod)
summary(reg_sales)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageArea +
##     FullBath + YearBuilt, data = property_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132164  -21488   -2273   17990  272272
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) -1023085.504    85308.638 -11.993 < 0.0000000000000002 ***
## OverallQual    22096.678     1084.397  20.377 < 0.0000000000000002 ***
## GrLivArea         72.123        2.952  24.430 < 0.0000000000000002 ***
## GarageArea        57.790        5.834   9.905 < 0.0000000000000002 ***
## FullBath      -13006.482     2564.799  -5.071           0.000000446 ***
## YearBuilt        483.702       44.912  10.770 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36850 on 1452 degrees of freedom
## Multiple R-squared:  0.7858, Adjusted R-squared:  0.7851
## F-statistic:  1065 on 5 and 1452 DF,  p-value: < 0.00000000000000022
```

Clearly omitting the outliers improve the model as we get the Adjusted R square value as 0.7851 which means these 5 variables explains 78.51% variabiltiy in Sale Price.

```
reg_sales <- lm(SalePrice ~ OverallQual + GrLivArea + GarageArea + YearBuilt, data =
property_mod)
summary(reg_sales)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageArea +
##     YearBuilt, data = property_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132927  -22005   -2052   18922  278168
##
## Coefficients:
##                Estimate  Std. Error t value            Pr(>|t|)
## (Intercept) -862947.392   79920.026 -10.798 <0.0000000000000002 ***
## OverallQual   21928.658    1093.071  20.062 <0.0000000000000002 ***
## GrLivArea        64.089       2.512  25.511 <0.0000000000000002 ***
## GarageArea       59.351       5.875  10.101 <0.0000000000000002 ***
## YearBuilt       398.448      42.000   9.487 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37170 on 1453 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7814
## F-statistic:  1303 on 4 and 1453 DF,  p-value: < 0.00000000000000022
```

Here we see that omitting FullBath has not really affected the model as such with 78.14% of variability still explained by the remaining 4 variables. This can be explained by the significant correlation between the variables FullBath and GrLivArea.

Let us check with nonlinear models. From the graphs earlier, it seemed that OverallQual and YearBuilt had some nonlinear relationship with SalePrice.

```
reg_sales <- lm(SalePrice ~ poly(OverallQual,2) + GrLivArea + GarageArea + poly(YearB
uilt,3), data = property_mod)
summary(reg_sales)
```

```
##
## Call:
## lm(formula = SalePrice ~ poly(OverallQual, 2) + GrLivArea + GarageArea +
##     poly(YearBuilt, 3), data = property_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137085  -19644     298   16910  241887
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              63601.60    3828.58  16.612 < 0.0000000000000002 ***
## poly(OverallQual, 2)1 1288039.10   55718.33  23.117 < 0.0000000000000002 ***
## poly(OverallQual, 2)2  642622.36   34279.84  18.746 < 0.0000000000000002 ***
## GrLivArea                   62.49       2.27  27.532 < 0.0000000000000002 ***
## GarageArea                  48.59       5.27   9.220 < 0.0000000000000002 ***
## poly(YearBuilt, 3)1    446593.71   44824.55   9.963 < 0.0000000000000002 ***
## poly(YearBuilt, 3)2   -174978.52   37750.89  -4.635           0.00000389 ***
## poly(YearBuilt, 3)3     73285.26   33998.46   2.156               0.0313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33100 on 1450 degrees of freedom
## Multiple R-squared:  0.8274, Adjusted R-squared:  0.8266
## F-statistic: 993.2 on 7 and 1450 DF,  p-value: < 0.00000000000000022
```

Clearly, the polynomial terms improves the model fit with the Adjusted R square being 0.8266, which means that these variables explain 82.66% of variability in Sale Price.

In a graph earlier we saw that Central Air has some effect on Sale Price when checked with Overall Quality. Also House Style might have some influence on Ground Living Area. Let us check if including the interaction improves the model or not.

```
reg_sales <- lm(SalePrice ~ poly(OverallQual,2) * CentralAir +  GrLivArea * HouseStyl
e + GarageArea + poly(YearBuilt, 3), data = property_mod)
summary(reg_sales)
```

```
## 
## Call:
## lm(formula = SalePrice ~ poly(OverallQual, 2) * CentralAir + 
##     GrLivArea * HouseStyle + GarageArea + poly(YearBuilt, 3), 
##     data = property_mod)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -142152  -16042      74   14840  229919 
## 
## Coefficients:
##                                     Estimate Std. Error t value
## (Intercept)                        32836.642  11607.229   2.829
## poly(OverallQual, 2)1             764126.362 278120.713   2.747
## poly(OverallQual, 2)2             189537.268 130121.130   1.457
## CentralAirY                        23243.915   6673.264   3.483
## GrLivArea                             66.450      5.693  11.672
## HouseStyle1.5Unf                   24283.798  70598.449   0.344
## HouseStyle1Story                    8394.141  10135.587   0.828
## HouseStyle2.5Fin                  -91517.838  56633.955  -1.616
## HouseStyle2.5Unf                   62769.910  44171.702   1.421
## HouseStyle2Story                  -53189.036  11142.163  -4.774
## HouseStyleSFoyer                   39939.879  20870.028   1.914
## HouseStyleSLvl                     39088.050  17049.489   2.293
## GarageArea                            31.760      5.075   6.258
## poly(YearBuilt, 3)1               387103.000  47879.513   8.085
## poly(YearBuilt, 3)2                -9445.161  37563.401  -0.251
## poly(YearBuilt, 3)3               127229.287  34420.489   3.696
## poly(OverallQual, 2)1:CentralAirY 374534.500 280778.609   1.334
## poly(OverallQual, 2)2:CentralAirY 450731.983 135800.357   3.319
## GrLivArea:HouseStyle1.5Unf            -8.689     77.687  -0.112
## GrLivArea:HouseStyle1Story             7.944      6.492   1.224
## GrLivArea:HouseStyle2.5Fin            22.959     20.154   1.139
## GrLivArea:HouseStyle2.5Unf           -41.533     22.855  -1.817
## GrLivArea:HouseStyle2Story            29.410      6.463   4.551
## GrLivArea:HouseStyleSFoyer           -21.314     19.212  -1.109
## GrLivArea:HouseStyleSLvl             -17.463     11.412  -1.530
##                                               Pr(>|t|)    
## (Intercept)                                   0.004735 ** 
## poly(OverallQual, 2)1                         0.006081 ** 
## poly(OverallQual, 2)2                         0.145440    
## CentralAirY                                   0.000510 ***
## GrLivArea                     < 0.0000000000000002 ***
## HouseStyle1.5Unf                              0.730919    
## HouseStyle1Story                              0.407704    
## HouseStyle2.5Fin                              0.106324    
## HouseStyle2.5Unf                              0.155522    
## HouseStyle2Story                   0.00000199403649372 ***
## HouseStyleSFoyer                              0.055852 .  
## HouseStyleSLvl                                0.022014 *  
## GarageArea                         0.00000000051368756 ***
## poly(YearBuilt, 3)1                0.00000000000000131 ***
## poly(YearBuilt, 3)2                           0.801505    
## poly(YearBuilt, 3)3                           0.000227 ***
## poly(OverallQual, 2)1:CentralAirY             0.182444    
## poly(OverallQual, 2)2:CentralAirY             0.000926 ***
## GrLivArea:HouseStyle1.5Unf                    0.910965    
## GrLivArea:HouseStyle1Story                    0.221254
## GrLivArea:HouseStyle2.5Fin                    0.254821
## GrLivArea:HouseStyle2.5Unf                    0.069395 .
## GrLivArea:HouseStyle2Story         0.00000579735854334 ***
## GrLivArea:HouseStyleSFoyer                    0.267438
## GrLivArea:HouseStyleSLvl                      0.126185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30710 on 1433 degrees of freedom
## Multiple R-squared:  0.8532, Adjusted R-squared:  0.8508
## F-statistic: 347.1 on 24 and 1433 DF,  p-value: < 0.00000000000000022
```

Clearly, the including the interactions improve the model as we see that the Adjusted R Square increases to 0.8508 meaning 85.08% of variability in Sale Price can be explained by these variables.

What if, if we use the complete dataset to see if there are any other variable which can improve the Adjusted R square value.

```
reg_sales_mod <- lm(SalePrice ~ ., data = property_mod)
summary(reg_sales_mod)
```

```
## 
## Call:
## lm(formula = SalePrice ~ ., data = property_mod)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -148265  -15597     -30   12939  205992 
## 
## Coefficients:
##                             Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)           -1029191.64698   101356.14666 -10.154 < 0.0000000000000002
## MSZoningFV               13851.43579    10606.49522   1.306             0.191785
## MSZoningRH               14452.56922    12209.58037   1.184             0.236727
## MSZoningRL               12731.97443     9872.64392   1.290             0.197392
## MSZoningRM                8646.07731     9879.16424   0.875             0.381622
## LotArea                      0.66796        0.08441   7.913  0.00000000000000501
## BldgType2fmCon            6098.18965     6445.18180   0.946             0.344226
## BldgTypeDuplex           -4152.48777     6748.80541  -0.615             0.538460
## BldgTypeTwnhs           -14451.98032     5077.24352  -2.846             0.004485
## BldgTypeTwnhsE          -12846.44445     3394.22645  -3.785             0.000160
## HouseStyle1.5Unf         21717.11004     8442.16508   2.572             0.010199
## HouseStyle1Story         20210.37252     3167.55420   6.380  0.00000000023834006
## HouseStyle2.5Fin        -39608.00206    11165.34787  -3.547             0.000402
## HouseStyle2.5Unf        -11414.19279     9624.02803  -1.186             0.235817
## HouseStyle2Story         -2444.32752     3215.29158  -0.760             0.447249
## HouseStyleSFoyer         19950.03952     6074.25332   3.284             0.001047
## HouseStyleSLvl            9143.63165     4681.74656   1.953             0.051011
## OverallQual              13726.51694     1016.37380  13.505 < 0.0000000000000002
## OverallCond               6180.08003      836.25914   7.390  0.00000000000024918
## YearBuilt                  510.86460       51.86537   9.850 < 0.0000000000000002
## CentralAirY              -2734.12197     3879.09769  -0.705             0.481029
## GrLivArea                   94.72356        3.34547  28.314 < 0.0000000000000002
## FullBath                 -4305.42435     2349.41752  -1.833             0.067079
## HalfBath                  -986.54103     2296.52329  -0.430             0.667566
## BedroomAbvGr            -10308.20025     1347.31302  -7.651  0.00000000000003659
## KitchenAbvGr            -22860.52767     6035.22663  -3.788             0.000158
## KitchenQualFa           -45657.35387     6595.84464  -6.922  0.00000000000671164
## KitchenQualGd           -48652.00918     3547.77322 -13.713 < 0.0000000000000002
## KitchenQualTA           -50279.08737     4085.35588 -12.307 < 0.0000000000000002
## FireplaceY                 872.27724     1895.82195   0.460             0.645511
## GarageArea                  26.21781        4.95475   5.291  0.00000014037680481
## SaleConditionAdjLand     24837.55674    15531.31929   1.599             0.110000
## SaleConditionAlloca       7531.21212     9423.14262   0.799             0.424293
## SaleConditionFamily       2028.06337     7277.73491   0.279             0.780541
## SaleConditionNormal       9735.50056     3150.01831   3.091             0.002036
## SaleConditionPartial     28445.12308     4320.04363   6.584  0.00000000006406545
## 
## (Intercept)          ***
## MSZoningFV              
## MSZoningRH              
## MSZoningRL              
## MSZoningRM              
## LotArea              ***
## BldgType2fmCon          
## BldgTypeDuplex          
## BldgTypeTwnhs        ** 
## BldgTypeTwnhsE       ***
## HouseStyle1.5Unf     *  
## HouseStyle1Story     ***
## HouseStyle2.5Fin     ***
## HouseStyle2.5Unf        
## HouseStyle2Story        
## HouseStyleSFoyer     ** 
## HouseStyleSLvl       .  
## OverallQual          ***
## OverallCond          ***
## YearBuilt            ***
## CentralAirY             
## GrLivArea            ***
## FullBath             .  
## HalfBath                
## BedroomAbvGr         ***
## KitchenAbvGr         ***
## KitchenQualFa        ***
## KitchenQualGd        ***
## KitchenQualTA        ***
## FireplaceY              
## GarageArea           ***
## SaleConditionAdjLand    
## SaleConditionAlloca     
## SaleConditionFamily     
## SaleConditionNormal  ** 
## SaleConditionPartial ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 29270 on 1422 degrees of freedom
## Multiple R-squared:  0.8677, Adjusted R-squared:  0.8644 
## F-statistic: 266.4 on 35 and 1422 DF,  p-value: < 0.00000000000000022
```

This Model gives us the Adjusted R Square value of .8644, which means this models improves the accuracy by almost 2%. However, this model includes variables which are not significant. When we make a model with only the significant variables:

```
reg_sales_lin <- lm(SalePrice ~ LotArea + BldgType + HouseStyle + OverallQual + Overa
llCond + YearBuilt + GrLivArea + BedroomAbvGr + KitchenAbvGr + KitchenQual + GarageAr
ea + SaleCondition, data = property_mod)
summary(reg_sales_lin)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + BldgType + HouseStyle + OverallQual +
##     OverallCond + YearBuilt + GrLivArea + BedroomAbvGr + KitchenAbvGr +
##     KitchenQual + GarageArea + SaleCondition, data = property_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148048  -15473      33   12651  205123
##
## Coefficients:
##                          Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)          -998949.65647   87211.74904 -11.454 < 0.0000000000000002
## LotArea                    0.67707       0.08377   8.082  0.0000000000000134
## BldgType2fmCon          6679.66193    6367.22461   1.049             0.294323
## BldgTypeDuplex         -3766.09615    6706.80832  -0.562             0.574522
## BldgTypeTwnhs         -16971.78205    4879.70150  -3.478             0.000520
## BldgTypeTwnhsE        -14313.25263    3245.58932  -4.410  0.0000111073975985l
## HouseStyle1.5Unf       22644.29255    8384.04065   2.701             0.006997
## HouseStyle1Story       21572.54825    3019.04720   7.145  0.0000000000142427
## HouseStyle2.5Fin      -38954.02558   11108.38468  -3.507             0.000468
## HouseStyle2.5Unf      -12507.78482    9424.81009  -1.327             0.184683
## HouseStyle2Story       -2171.99316    3071.39747  -0.707             0.479577
## HouseStyleSFoyer       21473.64667    5953.06259   3.607             0.000320
## HouseStyleSLvl         10598.99568    4600.91245   2.304             0.021384
## OverallQual            13684.94335     997.52586  13.719 < 0.0000000000000002
## OverallCond             6111.88559     802.78667   7.613  0.0000000000004827
## YearBuilt                499.50642      44.24405  11.290 < 0.0000000000000002
## GrLivArea                 93.25144       2.93293  31.795 < 0.0000000000000002
## BedroomAbvGr          -10549.87513    1313.00354  -8.035  0.0000000000000194
## KitchenAbvGr          -24526.89666    5950.16420  -4.122  0.0000397084417768З
## KitchenQualFa         -45442.01804    6549.42297  -6.938  0.0000000000599485
## KitchenQualGd         -49152.35468    3514.46941 -13.986 < 0.0000000000000002
## KitchenQualTA         -50356.07157    4066.73376 -12.382 < 0.0000000000000002
## GarageArea                25.51733       4.90999   5.197  0.0000002318108805б
## SaleConditionAdjLand   25087.49248   15435.50117   1.625             0.104317
## SaleConditionAlloca     6146.56421    9369.66765   0.656             0.511925
## SaleConditionFamily     1788.87460    7231.65146   0.247             0.804659
## SaleConditionNormal     9904.12477    3094.43500   3.201             0.001401
## SaleConditionPartial   28169.81966    4260.65930   6.612  0.0000000000535346 5
##
## (Intercept)          ***
## LotArea              ***
## BldgType2fmCon
## BldgTypeDuplex
## BldgTypeTwnhs        ***
## BldgTypeTwnhsE       ***
## HouseStyle1.5Unf     **
## HouseStyle1Story     ***
## HouseStyle2.5Fin     ***
## HouseStyle2.5Unf
## HouseStyle2Story
## HouseStyleSFoyer     ***
## HouseStyleSLvl       *
## OverallQual          ***
## OverallCond          ***
## YearBuilt            ***
## GrLivArea            ***
## BedroomAbvGr         ***
## KitchenAbvGr         ***
## KitchenQualFa        ***
## KitchenQualGd        ***
## KitchenQualTA        ***
## GarageArea           ***
## SaleConditionAdjLand    
## SaleConditionAlloca     
## SaleConditionFamily     
## SaleConditionNormal  ** 
## SaleConditionPartial ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 29260 on 1430 degrees of freedom
## Multiple R-squared:  0.867,  Adjusted R-squared:  0.8645 
## F-statistic: 345.2 on 27 and 1430 DF,  p-value: < 0.00000000000000022
```

The Adjusted R square value remains the same while we could eliminate the variables which are not significant. Therefore, this is the best linear model. However, as we have seen earlier, few variables show quadriatic relationship with Sale Price (like Overall Quality and Year Built). Also from the graphs earlier and correlation matrix, we can assume that some of the variables will have some interaction with other variables. For example, Overall Quality of the house might have interaction with that of Central Air, or Ground Living Area might depend on the House Style. For that we need to check with the nonlinear model including the interactions.

```
reg_sales_non <- lm(SalePrice ~ LotArea + BldgType + poly(OverallQual,3) : CentralAir
+ OverallCond + GrLivArea * HouseStyle + BedroomAbvGr + KitchenAbvGr + KitchenQual +
 poly(YearBuilt,3) + GarageArea + SaleCondition, data = property_mod)
summary(reg_sales_non)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + BldgType + poly(OverallQual,
##     3):CentralAir + OverallCond + GrLivArea * HouseStyle + BedroomAbvGr +
##     KitchenAbvGr + KitchenQual + poly(YearBuilt, 3) + GarageArea +
##     SaleCondition, data = property_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146839  -13409      45   12242  197631
##
## Coefficients:
##                                       Estimate   Std. Error t value
## (Intercept)                        51626.28953  11738.24785   4.398
## LotArea                                0.74715      0.07569   9.871
## BldgType2fmCon                      3983.49109   5778.77521   0.689
## BldgTypeDuplex                     -7772.18419   6044.13119  -1.286
## BldgTypeTwnhs                     -10943.25128   4507.10366  -2.428
## BldgTypeTwnhsE                     -7597.15483   2937.82577  -2.586
## OverallCond                         7916.09427    774.13466  10.226
## GrLivArea                             73.44479      5.09326  14.420
## HouseStyle1.5Unf                   -8928.05915  60539.67907  -0.147
## HouseStyle1Story                    -683.57815   8702.18666  -0.079
## HouseStyle2.5Fin                  -76255.24611  48415.68288  -1.575
## HouseStyle2.5Unf                   69667.89677  37983.03836   1.834
## HouseStyle2Story                  -44819.58855   9745.24826  -4.599
## HouseStyleSFoyer                   12800.93290  18154.09702   0.705
## HouseStyleSLvl                     35096.45087  14606.67088   2.403
## BedroomAbvGr                       -7427.60681   1204.05689  -6.169
## KitchenAbvGr                      -20936.48115   5401.47940  -3.876
## KitchenQualFa                     -28224.62573   6090.62747  -4.634
## KitchenQualGd                     -20421.39700   3530.56405  -5.784
## KitchenQualTA                     -23405.20311   3980.91816  -5.879
## poly(YearBuilt, 3)1               647878.37665  49748.70940  13.023
## poly(YearBuilt, 3)2                12574.96735  36355.69313   0.346
## poly(YearBuilt, 3)3                84745.78144  31353.34032   2.703
## GarageArea                            20.39091      4.44780   4.584
## SaleConditionAdjLand               23128.43173  13855.84893   1.669
## SaleConditionAlloca                10463.30835   8481.75188   1.234
## SaleConditionFamily                 5037.21628   6472.78750   0.778
## SaleConditionNormal                10627.44945   2770.65058   3.836
## SaleConditionPartial               26655.40274   3990.56826   6.680
## poly(OverallQual, 3)1:CentralAirN 802429.12132 219914.31351   3.649
## poly(OverallQual, 3)2:CentralAirN 442705.98355 228338.94229   1.939
## poly(OverallQual, 3)3:CentralAirN 123377.88965 122808.38274   1.005
## poly(OverallQual, 3)1:CentralAirY 835188.26375  52099.40589  16.031
## poly(OverallQual, 3)2:CentralAirY 431128.45374  43093.13542  10.005
## poly(OverallQual, 3)3:CentralAirY 284777.40440  45326.60290   6.283
## GrLivArea:HouseStyle1.5Unf            24.29902     66.60654   0.365
## GrLivArea:HouseStyle1Story            13.06525      5.57120   2.345
## GrLivArea:HouseStyle2.5Fin            19.32640     17.22693   1.122
## GrLivArea:HouseStyle2.5Unf           -41.26657     19.77327  -2.087
## GrLivArea:HouseStyle2Story            26.47270      5.62140   4.709
## GrLivArea:HouseStyleSFoyer             1.40216     16.75406   0.084
## GrLivArea:HouseStyleSLvl             -16.67403      9.80221  -1.701
##                                             Pr(>|t|)
## (Intercept)                           0.0000117356110 ***
## LotArea                            < 0.0000000000000002 ***
## BldgType2fmCon                                 0.490728    
## BldgTypeDuplex                                 0.198686    
## BldgTypeTwnhs                                  0.015306 *  
## BldgTypeTwnhsE                                 0.009809 ** 
## OverallCond                        < 0.0000000000000002 ***
## GrLivArea                          < 0.0000000000000002 ***
## HouseStyle1.5Unf                               0.882778    
## HouseStyle1Story                               0.937400    
## HouseStyle2.5Fin                               0.115477    
## HouseStyle2.5Unf                               0.066836 .  
## HouseStyle2Story                        0.0000046214572 ***
## HouseStyleSFoyer                               0.480848    
## HouseStyleSLvl                                 0.016399 *  
## BedroomAbvGr                            0.0000000008962 ***
## KitchenAbvGr                                   0.000111 ***
## KitchenQualFa                           0.0000039145800 ***
## KitchenQualGd                           0.0000000089548 ***
## KitchenQualTA                           0.0000000051299 ***
## poly(YearBuilt, 3)1                < 0.0000000000000002 ***
## poly(YearBuilt, 3)2                            0.729479    
## poly(YearBuilt, 3)3                            0.006955 ** 
## GarageArea                              0.0000049519729 ***
## SaleConditionAdjLand                           0.095295 .  
## SaleConditionAlloca                            0.217547    
## SaleConditionFamily                            0.436573    
## SaleConditionNormal                            0.000131 ***
## SaleConditionPartial                    0.0000000000343 ***
## poly(OverallQual, 3)1:CentralAirN              0.000273 ***
## poly(OverallQual, 3)2:CentralAirN              0.052723 .  
## poly(OverallQual, 3)3:CentralAirN              0.315243    
## poly(OverallQual, 3)1:CentralAirY < 0.0000000000000002 ***
## poly(OverallQual, 3)2:CentralAirY < 0.0000000000000002 ***
## poly(OverallQual, 3)3:CentralAirY      0.0000000004416 ***
## GrLivArea:HouseStyle1.5Unf                     0.715304    
## GrLivArea:HouseStyle1Story                     0.019157 *  
## GrLivArea:HouseStyle2.5Fin                     0.262107    
## GrLivArea:HouseStyle2.5Unf                     0.037068 *  
## GrLivArea:HouseStyle2Story              0.0000027299065 ***
## GrLivArea:HouseStyleSFoyer                     0.933314    
## GrLivArea:HouseStyleSLvl                       0.089153 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 26120 on 1416 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8921 
## F-statistic: 294.7 on 41 and 1416 DF,  p-value: < 0.00000000000000022
```

The model fit increases by almost 2.8% when we introduce non-linearity as well as interaction. Also all the variables are significant. Hence, this model seems to be the right fit.

We had earlier removed the outliers. Now we check whether removing them the outliers has created any negative impact on the final model or not.

```
reg_sales_non <- lm(SalePrice ~ LotArea + BldgType + poly(OverallQual,3) : CentralAir
+ OverallCond + GrLivArea * HouseStyle + BedroomAbvGr + KitchenAbvGr + KitchenQual +
 poly(YearBuilt,3) + GarageArea + SaleCondition, data = property)
summary(reg_sales_non)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + BldgType + poly(OverallQual, 
##     3):CentralAir + OverallCond + GrLivArea * HouseStyle + BedroomAbvGr + 
##     KitchenAbvGr + KitchenQual + poly(YearBuilt, 3) + GarageArea + 
##     SaleCondition, data = property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -507804  -14129    -816   11982  219686 
##
## Coefficients:
##                                       Estimate   Std. Error t value
## (Intercept)                         61204.1414   14731.8792   4.155
## LotArea                                 0.5427       0.0946   5.737
## BldgType2fmCon                       4269.5662    7257.5382   0.588
## BldgTypeDuplex                      -8442.3127    7590.0993  -1.112
## BldgTypeTwnhs                      -24331.4958    5628.7127  -4.323
## BldgTypeTwnhsE                     -12021.3962    3683.7956  -3.263
## OverallCond                          7423.5013     971.9116   7.638
## GrLivArea                              67.0490       6.3897  10.493
## HouseStyle1.5Unf                    -5252.1125   76026.2007  -0.069
## HouseStyle1Story                     -619.9233   10927.8378  -0.057
## HouseStyle2.5Fin                   -79972.8522   60797.7285  -1.315
## HouseStyle2.5Unf                    52234.5116   47690.3500   1.095
## HouseStyle2Story                     9923.8372   11998.8529   0.827
## HouseStyleSFoyer                    25873.1473   22789.8474   1.135
## HouseStyleSLvl                      38764.4325   18341.6104   2.113
## BedroomAbvGr                        -3125.8192    1500.1870  -2.084
## KitchenAbvGr                       -20078.0667    6782.8625  -2.960
## KitchenQualFa                      -35335.8173    7641.8836  -4.624
## KitchenQualGd                      -24860.6445    4429.3021  -5.613
## KitchenQualTA                      -30642.0125    4988.9981  -6.142
## poly(YearBuilt, 3)1                591138.0169   62536.2692   9.453
## poly(YearBuilt, 3)2                -30730.7468   45664.2624  -0.673
## poly(YearBuilt, 3)3                 22294.0517   39288.0885   0.567
## GarageArea                             23.7840       5.5770   4.265
## SaleConditionAdjLand                19268.1189   17398.6068   1.107
## SaleConditionAlloca                 16526.3108   10647.6655   1.552
## SaleConditionFamily                  -303.7499    8124.8666  -0.037
## SaleConditionNormal                  9562.1871    3478.9467   2.749
## SaleConditionPartial                20129.0627    5002.3933   4.024
## poly(OverallQual, 3)1:CentralAirN 1010509.4795  280034.6395   3.609
## poly(OverallQual, 3)2:CentralAirN  539463.2665  292023.5683   1.847
## poly(OverallQual, 3)3:CentralAirN  125809.8492  155695.6995   0.808
## poly(OverallQual, 3)1:CentralAirY  930797.0388   65770.8815  14.152
## poly(OverallQual, 3)2:CentralAirY  487933.4915   54066.0007   9.025
## poly(OverallQual, 3)3:CentralAirY   62559.5150   56967.4592   1.098
## GrLivArea:HouseStyle1.5Unf             20.1367      83.6443   0.241
## GrLivArea:HouseStyle1Story             12.1599       6.9959   1.738
## GrLivArea:HouseStyle2.5Fin             20.4572      21.6328   0.946
## GrLivArea:HouseStyle2.5Unf            -32.1625      24.8266  -1.295
## GrLivArea:HouseStyle2Story             -5.2244       6.9204  -0.755
## GrLivArea:HouseStyleSFoyer            -11.7422      21.0311  -0.558
## GrLivArea:HouseStyleSLvl              -21.8798      12.3070  -1.778
##                                                 Pr(>|t|)    
## (Intercept)                         0.0000345421146625 ***
## LotArea                             0.0000000117740991 ***
## BldgType2fmCon                                0.556429    
## BldgTypeDuplex                                0.266206    
## BldgTypeTwnhs                       0.0000164851327383 ***
## BldgTypeTwnhsE                                0.001127 ** 
## OverallCond                         0.0000000000000404 ***
## GrLivArea                         < 0.0000000000000002 ***
## HouseStyle1.5Unf                              0.944933    
## HouseStyle1Story                              0.954769    
## HouseStyle2.5Fin                              0.188591    
## HouseStyle2.5Unf                              0.273578    
## HouseStyle2Story                              0.408339    
## HouseStyleSFoyer                              0.256445    
## HouseStyleSLvl                                0.034735 *  
## BedroomAbvGr                                  0.037373 *  
## KitchenAbvGr                                  0.003126 ** 
## KitchenQualFa                       0.0000041075216660 ***
## KitchenQualGd                       0.0000000239165710 ***
## KitchenQualTA                       0.0000000010568851 ***
## poly(YearBuilt, 3)1               < 0.0000000000000002 ***
## poly(YearBuilt, 3)2                           0.501075    
## poly(YearBuilt, 3)3                           0.570498    
## GarageArea                          0.0000213493831900 ***
## SaleConditionAdjLand                          0.268286    
## SaleConditionAlloca                           0.120860    
## SaleConditionFamily                           0.970183    
## SaleConditionNormal                           0.006061 ** 
## SaleConditionPartial                0.0000602668662298 ***
## poly(OverallQual, 3)1:CentralAirN             0.000319 ***
## poly(OverallQual, 3)2:CentralAirN             0.064908 .  
## poly(OverallQual, 3)3:CentralAirN             0.419198    
## poly(OverallQual, 3)1:CentralAirY < 0.0000000000000002 ***
## poly(OverallQual, 3)2:CentralAirY < 0.0000000000000002 ***
## poly(OverallQual, 3)3:CentralAirY             0.272320    
## GrLivArea:HouseStyle1.5Unf                    0.809790    
## GrLivArea:HouseStyle1Story                    0.082403 .  
## GrLivArea:HouseStyle2.5Fin                    0.344485    
## GrLivArea:HouseStyle2.5Unf                    0.195363    
## GrLivArea:HouseStyle2Story                    0.450417    
## GrLivArea:HouseStyleSFoyer                    0.576709    
## GrLivArea:HouseStyleSLvl                      0.075645 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 32790 on 1418 degrees of freedom
## Multiple R-squared:  0.8344, Adjusted R-squared:  0.8296 
## F-statistic: 174.2 on 41 and 1418 DF,  p-value: < 0.00000000000000022
```

We see that the Adjusted R Square value drastically decreases when we include the outliers. Also, the interaction variable of Overall Quality with that of Full Bath becomes less significant. Thus we can conclude that removing the outliers was a good call as they would have made a statistically very significant variable less significant.

Now we check the individual model performances with the help of PRESS statistic #### Non-linear model without Outliers

```
# now calculate cross-validated residuals
n <- nrow(property_mod)
cv_res1 = vector(length=n)
for(i in 1:n){
  fiti = lm(SalePrice ~ LotArea + BldgType + poly(OverallQual,3) : CentralAir + Overa
llCond + GrLivArea * HouseStyle + BedroomAbvGr + KitchenAbvGr + KitchenQual + poly(Ye
arBuilt,3) + GarageArea + SaleCondition, data = property_mod[-i,])
  predi = predict(fiti, newdata=property_mod[i,])
  cv_res1[i] = property_mod$SalePrice[i] - predi
}

# PRESS is sum of squared cross-validated residuals
PRESS1 = sum(cv_res1^2)
PRESS1
```

```
## [1] 1047549643073
```

Non-linear model with Outliers

```
# now calculate cross-validated residuals
n <- nrow(property)
cv_res1 = vector(length=n)
for(i in 1:n){
  fiti = lm(SalePrice ~ LotArea + BldgType + poly(OverallQual,3) : CentralAir + Overa
llCond + GrLivArea * HouseStyle + BedroomAbvGr + KitchenAbvGr + KitchenQual + poly(Ye
arBuilt,3) + GarageArea + SaleCondition, data = property[-i,])
  predi = predict(fiti, newdata=property[i,])
  cv_res1[i] = property$SalePrice[i] - predi
}

# PRESS is sum of squared cross-validated residuals
PRESS2 = sum(cv_res1^2)
PRESS2
```

```
## [1] 1765807640717
```

Linear model without Outliers

```
# now calculate cross-validated residuals
n <- nrow(property_mod)
cv_res1 = vector(length=n)
for(i in 1:n){
  fiti = lm(SalePrice ~ LotArea + BldgType + HouseStyle + OverallQual + OverallCond +
YearBuilt + GrLivArea + BedroomAbvGr + KitchenAbvGr +                       KitchenQu
al + GarageArea + SaleCondition, data=property_mod[-i,])
  predi = predict(fiti, newdata=property_mod[i,])
  cv_res1[i] = property_mod$SalePrice[i] - predi
}

# PRESS is sum of squared cross-validated residuals
PRESS3 = sum(cv_res1^2)
PRESS3
```

```
## [1] 1290540563702
```

Clearly, the PRESS statistics also shows that the non-linear model without outliers has the best performance as it has the least PRESS value. Hence it is the best model.

Question 3: Develop a classification model to predict whether a property has a fireplace or not. The variable Fireplace has already been set as a factor variable earlier in the analysis.

As Fireplace is a binary variable taking only the values of 0 and 1, a logit model is used to model the probability of having a fireplace in a property (ranging from 0 to 1).

The model is p(x) = P(Fireplace=1|X=x) where X is a vector of all explanatory variables used in the model and x corresponds to the value of explanatory variables for a given property.

The model equation gives p(x) = [e^(β0+ β’X)] / [1+ e^(β0+ β’X)], where X is a vector of all explanatory variables and β is a vector of corresponding coefficients to the explanatory variables. The coefficients of the model are estimated by using the method of maximum likelihood.

To specify a logit model with useful predictors, we first include all of the variables and then remove one variable with the highest p-value each time from the model until all remaining variables are highly significant (i.e., at 0.1% level as indicated by *** in the R output).

The R script showing the process of finding the right regression model with all useful predictors is as follows:

```
##Setting all categorical variables as factors

property$MSZoning <- as.factor(property$MSZoning)
property$BldgType <- as.factor(property$BldgType)
property$HouseStyle <- as.factor(property$HouseStyle)
property$CentralAir <- as.factor(property$CentralAir)
property$KitchenQual <- as.factor(property$KitchenQual)
property$Fireplace <- as.factor(property$Fireplace)
property$SaleCondition <- as.factor(property$SaleCondition)
```

Finding the model

```
# all variables included in the model

logreg1 <- glm(Fireplace ~ MSZoning+LotArea+BldgType+HouseStyle+OverallQual+OverallCo
nd+YearBuilt+CentralAir+GrLivArea+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+Kitchen
Qual+GarageArea+SaleCondition+SalePrice,family=binomial, data=property)

summary(logreg1)
```

```
##
## Call:
## glm(formula = Fireplace ~ MSZoning + LotArea + BldgType + HouseStyle +
##     OverallQual + OverallCond + YearBuilt + CentralAir + GrLivArea +
##     FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual +
##     GarageArea + SaleCondition + SalePrice, family = binomial,
##     data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.06383  -0.75641   0.08211   0.73601   3.08919
##
## Coefficients:
##                          Estimate    Std. Error z value     Pr(>|z|)
## (Intercept)           25.745319699   9.689926752   2.657     0.007886 **
## MSZoningFV            -0.303602347   1.320637037  -0.230     0.818177
## MSZoningRH            -1.175303070   1.733431859  -0.678     0.497758
## MSZoningRL             1.119989404   1.280169809   0.875     0.381642
## MSZoningRM             0.519019530   1.276738700   0.407     0.684361
## LotArea                0.000038189   0.000021643   1.765     0.077645 .
## BldgType2fmCon         0.308332031   0.632697182   0.487     0.626025
## BldgTypeDuplex        -2.188969037   0.884491058  -2.475     0.013330 *
## BldgTypeTwnhs          0.315395124   0.453495069   0.695     0.486757
## BldgTypeTwnhsE         1.183331614   0.312968468   3.781     0.000156 ***
## HouseStyle1.5Unf       1.532414143   0.666861156   2.298     0.021565 *
## HouseStyle1Story       0.419167780   0.283036646   1.481     0.138616
## HouseStyle2.5Fin      -1.595013913   1.197227971  -1.332     0.182776
## HouseStyle2.5Unf       1.568043456   0.996151097   1.574     0.115464
## HouseStyle2Story      -0.219150176   0.282621355  -0.775     0.438092
## HouseStyleSFoyer       0.612393422   0.550217899   1.113     0.265708
## HouseStyleSLvl         1.038779076   0.395796976   2.625     0.008677 **
## OverallQual            0.239131791   0.100607852   2.377     0.017460 *
## OverallCond           -0.162215687   0.078346167  -2.070     0.038406 *
## YearBuilt             -0.017225939   0.004946547  -3.482     0.000497 ***
## CentralAirY            1.780959978   0.474967690   3.750     0.000177 ***
## GrLivArea              0.002290587   0.000420700   5.445 0.0000000519 ***
## FullBath              -0.094143802   0.209902476  -0.449     0.653784
## HalfBath               0.484603899   0.200199740   2.421     0.015495 *
## BedroomAbvGr          -0.363817016   0.131146254  -2.774     0.005535 **
## KitchenAbvGr          -0.878258523   0.655554062  -1.340     0.180337
## KitchenQualFa          0.516036177   0.686215794   0.752     0.452049
## KitchenQualGd          0.128846464   0.468708373   0.275     0.783395
## KitchenQualTA          0.610784889   0.493949417   1.237     0.216260
## GarageArea            -0.000827447   0.000481057  -1.720     0.085422 .
## SaleConditionAdjLand -11.923700329 382.847998853  -0.031     0.975154
## SaleConditionAlloca    0.978527513   1.246111513   0.785     0.432298
## SaleConditionFamily    1.080132670   0.609817937   1.771     0.076521 .
## SaleConditionNormal    0.434531146   0.301704427   1.440     0.149795
## SaleConditionPartial   0.235565279   0.409878565   0.575     0.565481
## SalePrice              0.000015661   0.000003588   4.365 0.0000127083 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1336.3  on 1424  degrees of freedom
## AIC: 1408.3
##
## Number of Fisher Scoring iterations: 13
```

```
# remove MSZoning

logreg1a <- glm(Fireplace ~ LotArea+BldgType+HouseStyle+OverallQual+OverallCond+YearB
uilt+CentralAir+GrLivArea+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+KitchenQual+Gar
ageArea+SaleCondition+SalePrice,family=binomial, data=property)

summary(logreg1a)
```

```
## 
## Call:
## glm(formula = Fireplace ~ LotArea + BldgType + HouseStyle + OverallQual + 
##     OverallCond + YearBuilt + CentralAir + GrLivArea + FullBath + 
##     HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + GarageArea + 
##     SaleCondition + SalePrice, family = binomial, data = property)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.95310  -0.79585   0.09147   0.76773   2.87345  
## 
## Coefficients:
##                           Estimate    Std. Error z value     Pr(>|z|)    
## (Intercept)           21.095114381   9.315488288   2.265      0.02354 *  
## LotArea                0.000058183   0.000022164   2.625      0.00866 ** 
## BldgType2fmCon         0.185682816   0.633510133   0.293      0.76944    
## BldgTypeDuplex        -2.074382195   0.873652834  -2.374      0.01758 *  
## BldgTypeTwnhs          0.030873455   0.433632916   0.071      0.94324    
## BldgTypeTwnhsE         0.883543777   0.293277735   3.013      0.00259 ** 
## HouseStyle1.5Unf       1.590124715   0.666182411   2.387      0.01699 *  
## HouseStyle1Story       0.624961035   0.276945538   2.257      0.02403 *  
## HouseStyle2.5Fin      -2.125790060   1.176098878  -1.807      0.07069 .  
## HouseStyle2.5Unf       1.430879720   0.970360652   1.475      0.14032    
## HouseStyle2Story      -0.274562105   0.278439738  -0.986      0.32410    
## HouseStyleSFoyer       0.790653773   0.545752427   1.449      0.14741    
## HouseStyleSLvl         1.254809380   0.393417471   3.190      0.00143 ** 
## OverallQual            0.227867280   0.098714424   2.308      0.02098 *  
## OverallCond           -0.148115796   0.077472512  -1.912      0.05590 .  
## YearBuilt             -0.014485966   0.004786360  -3.027      0.00247 ** 
## CentralAirY            1.853222085   0.468212458   3.958 0.0000755548 ***
## GrLivArea              0.002459495   0.000415017   5.926 0.0000000031 ***
## FullBath              -0.076484097   0.207963947  -0.368      0.71304    
## HalfBath               0.521301181   0.197392526   2.641      0.00827 ** 
## BedroomAbvGr          -0.344859957   0.129449479  -2.664      0.00772 ** 
## KitchenAbvGr          -1.072900318   0.662934665  -1.618      0.10557    
## KitchenQualFa          0.367145994   0.681539299   0.539      0.59009    
## KitchenQualGd         -0.003392794   0.464938925  -0.007      0.99418    
## KitchenQualTA          0.517101109   0.491039639   1.053      0.29231    
## GarageArea            -0.001010869   0.000472141  -2.141      0.03227 *  
## SaleConditionAdjLand -11.657437568 381.879097862  -0.031      0.97565    
## SaleConditionAlloca    0.948458533   1.224895320   0.774      0.43874    
## SaleConditionFamily    1.153338972   0.607659754   1.898      0.05770 .  
## SaleConditionNormal    0.564361899   0.292670249   1.928      0.05382 .  
## SaleConditionPartial   0.148352892   0.396559142   0.374      0.70833    
## SalePrice              0.000014465   0.000003454   4.188 0.0000281079 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1363.3  on 1428  degrees of freedom
## AIC: 1427.3
## 
## Number of Fisher Scoring iterations: 13
```

```
# remove KitchenQual

logreg1b <- glm(Fireplace ~ LotArea+BldgType+HouseStyle+OverallQual+OverallCond+YearB
uilt+CentralAir+GrLivArea+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+GarageArea+Sale
Condition+SalePrice,family=binomial, data=property)

summary(logreg1b)
```

```
## 
## Call:
## glm(formula = Fireplace ~ LotArea + BldgType + HouseStyle + OverallQual + 
##     OverallCond + YearBuilt + CentralAir + GrLivArea + FullBath + 
##     HalfBath + BedroomAbvGr + KitchenAbvGr + GarageArea + SaleCondition + 
##     SalePrice, family = binomial, data = property)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.87152  -0.79613   0.09755   0.75877   2.71316  
## 
## Coefficients:
##                           Estimate    Std. Error z value      Pr(>|z|)    
## (Intercept)           25.908270213   9.012258561   2.875      0.004043 ** 
## LotArea                0.000064942   0.000022108   2.938      0.003309 ** 
## BldgType2fmCon         0.194528785   0.632488087   0.308      0.758416    
## BldgTypeDuplex        -1.975164924   0.873313947  -2.262      0.023717 *  
## BldgTypeTwnhs          0.060790777   0.434115524   0.140      0.888633    
## BldgTypeTwnhsE         0.895205473   0.293744810   3.048      0.002307 ** 
## HouseStyle1.5Unf       1.495648275   0.664295969   2.251      0.024355 *  
## HouseStyle1Story       0.621350479   0.275377394   2.256      0.024048 *  
## HouseStyle2.5Fin      -2.105009765   1.144415277  -1.839      0.065860 .  
## HouseStyle2.5Unf       1.502567736   0.962176013   1.562      0.118374    
## HouseStyle2Story      -0.275907346   0.278251227  -0.992      0.321404    
## HouseStyleSFoyer       0.798002005   0.544445764   1.466      0.142726    
## HouseStyleSLvl         1.320882864   0.392110399   3.369      0.000755 ***
## OverallQual            0.191338180   0.097377387   1.965      0.049424 *  
## OverallCond           -0.182722362   0.075402996  -2.423      0.015381 *  
## YearBuilt             -0.016544897   0.004639387  -3.566      0.000362 ***
## CentralAirY            1.935986357   0.463752646   4.175 0.00002984972 ***
## GrLivArea              0.002373525   0.000409341   5.798 0.00000000669 ***
## FullBath              -0.127523630   0.205889008  -0.619      0.535666    
## HalfBath               0.538713768   0.196053195   2.748      0.006000 ** 
## BedroomAbvGr          -0.303416934   0.127676481  -2.376      0.017480 *  
## KitchenAbvGr          -1.019702360   0.662414655  -1.539      0.123714    
## GarageArea            -0.001038332   0.000468690  -2.215      0.026733 *  
## SaleConditionAdjLand -11.723978213 377.874708755  -0.031      0.975249    
## SaleConditionAlloca    1.012393476   1.229116924   0.824      0.410124    
## SaleConditionFamily    1.179980396   0.609069893   1.937      0.052703 .  
## SaleConditionNormal    0.608701999   0.290057917   2.099      0.035856 *  
## SaleConditionPartial   0.133875985   0.394259326   0.340      0.734185    
## SalePrice              0.000013182   0.000003288   4.010 0.00006081739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1371.4  on 1431  degrees of freedom
## AIC: 1429.4
## 
## Number of Fisher Scoring iterations: 13
```

```
# remove SaleCondition

logreg1c <- glm(Fireplace ~ LotArea+BldgType+HouseStyle+OverallQual+OverallCond+YearB
uilt+CentralAir+GrLivArea+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+GarageArea+Sale
Price,family=binomial, data=property)

summary(logreg1c)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + BldgType + HouseStyle + OverallQual +
##     OverallCond + YearBuilt + CentralAir + GrLivArea + FullBath +
##     HalfBath + BedroomAbvGr + KitchenAbvGr + GarageArea + SalePrice,
##     family = binomial, data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.99603  -0.79600   0.09084   0.76702   2.73041
##
## Coefficients:
##                     Estimate   Std. Error z value      Pr(>|z|)
## (Intercept)      27.121107556  8.894417310   3.049      0.002294 **
## LotArea           0.000066295  0.000022124   2.997      0.002731 **
## BldgType2fmCon    0.174267351  0.618815090   0.282      0.778239
## BldgTypeDuplex   -1.864978840  0.783115539  -2.381      0.017243 *
## BldgTypeTwnhs     0.118050152  0.430420313   0.274      0.783879
## BldgTypeTwnhsE    0.919261244  0.292436696   3.143      0.001670 **
## HouseStyle1.5Unf  1.577437019  0.663545579   2.377      0.017441 *
## HouseStyle1Story  0.648291192  0.272878617   2.376      0.017513 *
## HouseStyle2.5Fin -2.370709808  1.128440378  -2.101      0.035652 *
## HouseStyle2.5Unf  1.561241146  0.954371983   1.636      0.101864
## HouseStyle2Story -0.265665374  0.276096512  -0.962      0.335939
## HouseStyleSFoyer  0.869505509  0.533109717   1.631      0.102889
## HouseStyleSLvl    1.330160879  0.387242595   3.435      0.000593 ***
## OverallQual       0.177202117  0.096701482   1.832      0.066882 .
## OverallCond      -0.159500756  0.074357256  -2.145      0.031948 *
## YearBuilt        -0.016941910  0.004580678  -3.699      0.000217 ***
## CentralAirY       1.930828554  0.456878211   4.226 0.00002377404 ***
## GrLivArea         0.002414140  0.000406327   5.941 0.00000000283 ***
## FullBath         -0.132456231  0.206104089  -0.643      0.520440
## HalfBath          0.561216028  0.195750630   2.867      0.004144 **
## BedroomAbvGr     -0.292807863  0.125788662  -2.328      0.019924 *
## KitchenAbvGr     -1.042538176  0.613080774  -1.700      0.089039 .
## GarageArea       -0.001036309  0.000467059  -2.219      0.026500 *
## SalePrice         0.000012999  0.000003269   3.976 0.00007003091 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1381.3  on 1436  degrees of freedom
## AIC: 1429.3
##
## Number of Fisher Scoring iterations: 6
```

```
# remove FullBath

logreg1d <- glm(Fireplace ~ LotArea+BldgType+HouseStyle+OverallQual+OverallCond+YearB
uilt+CentralAir+GrLivArea+HalfBath+BedroomAbvGr+KitchenAbvGr+GarageArea+SalePrice,fam
ily=binomial, data=property)

summary(logreg1d)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + BldgType + HouseStyle + OverallQual +
##     OverallCond + YearBuilt + CentralAir + GrLivArea + HalfBath +
##     BedroomAbvGr + KitchenAbvGr + GarageArea + SalePrice, family = binomial,
##     data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.01052  -0.79589   0.09395   0.76601   2.71337
##
## Coefficients:
##                      Estimate   Std. Error z value     Pr(>|z|)
## (Intercept)      29.282185708  8.228161711   3.559     0.000373 ***
## LotArea           0.000066588  0.000022129   3.009     0.002621 **
## BldgType2fmCon    0.181982179  0.616535191   0.295     0.767865
## BldgTypeDuplex   -1.874631286  0.778692484  -2.407     0.016066 *
## BldgTypeTwnhs     0.102283073  0.429990967   0.238     0.811980
## BldgTypeTwnhsE    0.909570391  0.292022667   3.115     0.001841 **
## HouseStyle1.5Unf  1.586483582  0.663380341   2.392     0.016779 *
## HouseStyle1Story  0.671332706  0.270655799   2.480     0.013124 *
## HouseStyle2.5Fin -2.343861888  1.128901259  -2.076     0.037872 *
## HouseStyle2.5Unf  1.569041218  0.956267321   1.641     0.100839
## HouseStyle2Story -0.277361154  0.275297234  -1.007     0.313696
## HouseStyleSFoyer  0.905267287  0.531784776   1.702     0.088696 .
## HouseStyleSLvl    1.359455560  0.385749551   3.524     0.000425 ***
## OverallQual       0.172661576  0.096451059   1.790     0.073430 .
## OverallCond      -0.159056124  0.074396245  -2.138     0.032520 *
## YearBuilt        -0.018061418  0.004234562  -4.265 0.0000199689 ***
## CentralAirY       1.944435580  0.455754883   4.266 0.0000198647 ***
## GrLivArea         0.002341842  0.000389931   6.006 0.0000000019 ***
## HalfBath          0.608554146  0.181673689   3.350     0.000809 ***
## BedroomAbvGr     -0.300413328  0.125242709  -2.399     0.016456 *
## KitchenAbvGr     -1.077184011  0.606148451  -1.777     0.075552 .
## GarageArea       -0.001038770  0.000466429  -2.227     0.025943 *
## SalePrice         0.000012914  0.000003256   3.966 0.0000732129 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1381.7  on 1437  degrees of freedom
## AIC: 1427.7
##
## Number of Fisher Scoring iterations: 6
```

```
# remove BldgType

logreg1e <- glm(Fireplace ~ LotArea+HouseStyle+OverallQual+OverallCond+YearBuilt+Cent
ralAir+GrLivArea+HalfBath+BedroomAbvGr+KitchenAbvGr+GarageArea+SalePrice,family=binom
ial, data=property)

summary(logreg1e)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + HouseStyle + OverallQual +
##     OverallCond + YearBuilt + CentralAir + GrLivArea + HalfBath +
##     BedroomAbvGr + KitchenAbvGr + GarageArea + SalePrice, family = binomial,
##     data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.97693  -0.78378   0.09393   0.78468   2.59773
##
## Coefficients:
##                     Estimate  Std. Error z value       Pr(>|z|)
## (Intercept)      28.26573142  7.94131255   3.559       0.000372 ***
## LotArea           0.00004524  0.00001892   2.392       0.016776 *
## HouseStyle1.5Unf  1.44964816  0.66185569   2.190       0.028504 *
## HouseStyle1Story  0.64945809  0.26843882   2.419       0.015547 *
## HouseStyle2.5Fin -1.86630263  1.09069827  -1.711       0.087061 .
## HouseStyle2.5Unf  2.04567277  1.00218361   2.041       0.041229 *
## HouseStyle2Story -0.20736120  0.27278486  -0.760       0.447157
## HouseStyleSFoyer  0.49940039  0.51416141   0.971       0.331403
## HouseStyleSLvl    1.31238622  0.38235735   3.432       0.000598 ***
## OverallQual       0.19581039  0.09514717   2.058       0.039593 *
## OverallCond      -0.16645939  0.07383283  -2.255       0.024162 *
## YearBuilt        -0.01681141  0.00408456  -4.116 0.000038575874 ***
## CentralAirY       1.82051875  0.44180579   4.121 0.000037783579 ***
## GrLivArea         0.00240158  0.00038518   6.235 0.000000000452 ***
## HalfBath          0.52222429  0.17567717   2.973       0.002953 **
## BedroomAbvGr     -0.44594184  0.11638321  -3.832       0.000127 ***
## KitchenAbvGr     -1.82140974  0.41697508  -4.368 0.000012530337 ***
## GarageArea       -0.00099926  0.00045606  -2.191       0.028447 *
## SalePrice         0.00001251  0.00000319   3.921 0.000088157871 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1400.8  on 1441  degrees of freedom
## AIC: 1438.8
##
## Number of Fisher Scoring iterations: 6
```

```
# remove HouseStyle

logreg1f <- glm(Fireplace ~ LotArea+OverallQual+OverallCond+YearBuilt+CentralAir+GrLi
vArea+HalfBath+BedroomAbvGr+KitchenAbvGr+GarageArea+SalePrice,family=binomial, data=p
roperty)

summary(logreg1f)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + CentralAir + GrLivArea + HalfBath + BedroomAbvGr +
##     KitchenAbvGr + GarageArea + SalePrice, family = binomial,
##     data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.00588  -0.81070   0.09798   0.84888   2.46757
##
## Coefficients:
##                  Estimate   Std. Error z value    Pr(>|z|)
## (Intercept)  28.758158919  7.432588539   3.869    0.000109 ***
## LotArea       0.000062635  0.000018661   3.357    0.000789 ***
## OverallQual   0.189961388  0.092151628   2.061    0.039265 *
## OverallCond  -0.190186813  0.071050220  -2.677    0.007433 **
## YearBuilt    -0.016529424  0.003799425  -4.351 0.000013582 ***
## CentralAirY   1.558243557  0.381018771   4.090 0.000043198 ***
## GrLivArea     0.001610985  0.000323737   4.976 0.000000648 ***
## HalfBath      0.206520005  0.145370903   1.421    0.155421
## BedroomAbvGr -0.463406813  0.113405773  -4.086 0.000043836 ***
## KitchenAbvGr -1.483410686  0.391731945  -3.787    0.000153 ***
## GarageArea   -0.000837843  0.000445019  -1.883    0.059740 .
## SalePrice     0.000015244  0.000003075   4.958 0.000000713 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1432.8  on 1448  degrees of freedom
## AIC: 1456.8
##
## Number of Fisher Scoring iterations: 6
```

```
# remove HalfBath

logreg1g <- glm(Fireplace ~ LotArea+OverallQual+OverallCond+YearBuilt+CentralAir+GrLi
vArea+BedroomAbvGr+KitchenAbvGr+GarageArea+SalePrice,family=binomial, data=property)

summary(logreg1g)
```

```
## 
## Call:
## glm(formula = Fireplace ~ LotArea + OverallQual + OverallCond + 
##     YearBuilt + CentralAir + GrLivArea + BedroomAbvGr + KitchenAbvGr + 
##     GarageArea + SalePrice, family = binomial, data = property)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.99134  -0.80054   0.09704   0.85381   2.49374  
## 
## Coefficients:
##                  Estimate   Std. Error z value      Pr(>|z|)    
## (Intercept)  26.062591003  7.153950019   3.643      0.000269 ***
## LotArea       0.000061599  0.000018653   3.302      0.000959 ***
## OverallQual   0.183220571  0.092113918   1.989      0.046694 *  
## OverallCond  -0.179922065  0.070732726  -2.544      0.010969 *  
## YearBuilt    -0.015168393  0.003658993  -4.146 0.00003390581 ***
## CentralAirY   1.571080299  0.380868260   4.125 0.00003707402 ***
## GrLivArea     0.001771692  0.000304663   5.815 0.00000000605 ***
## BedroomAbvGr -0.461329293  0.113547292  -4.063 0.00004847045 ***
## KitchenAbvGr -1.530842710  0.389309764  -3.932 0.00008417302 ***
## GarageArea   -0.000878916  0.000443862  -1.980      0.047686 *  
## SalePrice     0.000014654  0.000003045   4.812 0.00000149119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1434.8  on 1449  degrees of freedom
## AIC: 1456.8
## 
## Number of Fisher Scoring iterations: 6
```

```
# remove GarageArea

logreg1h <- glm(Fireplace ~ LotArea+OverallQual+OverallCond+YearBuilt+CentralAir+GrLivArea+BedroomAbvGr+KitchenAbvGr+SalePrice,family=binomial, data=property)

summary(logreg1h)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + CentralAir + GrLivArea + BedroomAbvGr + KitchenAbvGr +
##     SalePrice, family = binomial, data = property)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.99380  -0.80212   0.09985   0.85048   2.47565
##
## Coefficients:
##                 Estimate  Std. Error z value      Pr(>|z|)
## (Intercept)  27.43249369  7.09269248   3.868      0.000110 ***
## LotArea       0.00005599  0.00001825   3.067      0.002160 **
## OverallQual   0.18454237  0.09199745   2.006      0.044861 *
## OverallCond  -0.17223328  0.07053012  -2.442      0.014607 *
## YearBuilt    -0.01590637  0.00362626  -4.386 0.00001152231 ***
## CentralAirY   1.54537739  0.38211404   4.044 0.00005248345 ***
## GrLivArea     0.00176482  0.00030453   5.795 0.00000000682 ***
## BedroomAbvGr -0.43797769  0.11265969  -3.888      0.000101 ***
## KitchenAbvGr -1.60535019  0.38932345  -4.123 0.00003732627 ***
## SalePrice     0.00001303  0.00000291   4.479 0.00000750045 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1438.7  on 1450  degrees of freedom
## AIC: 1458.7
##
## Number of Fisher Scoring iterations: 6
```

```
# remove OverallQual

logreg1i <- glm(Fireplace ~ LotArea+OverallCond+YearBuilt+CentralAir+GrLivArea+Bedroo
mAbvGr+KitchenAbvGr+SalePrice,family=binomial, data=property)

summary(logreg1i)
```

```
## 
## Call:
## glm(formula = Fireplace ~ LotArea + OverallCond + YearBuilt + 
##     CentralAir + GrLivArea + BedroomAbvGr + KitchenAbvGr + SalePrice, 
##     family = binomial, data = property)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.98294  -0.81848   0.09416   0.84841   2.47054  
## 
## Coefficients:
##                  Estimate   Std. Error z value       Pr(>|z|)    
## (Intercept)  25.536096479  7.014252604   3.641       0.000272 ***
## LotArea       0.000046659  0.000017503   2.666       0.007682 ** 
## OverallCond  -0.165757719  0.070207397  -2.361       0.018227 *  
## YearBuilt    -0.014558413  0.003554456  -4.096 0.000042067610 ***
## CentralAirY   1.507250966  0.378708968   3.980 0.000068923479 ***
## GrLivArea     0.001854603  0.000301798   6.145 0.000000000799 ***
## BedroomAbvGr -0.454575664  0.111957962  -4.060 0.000049023293 ***
## KitchenAbvGr -1.684786369  0.388502422  -4.337 0.000014469215 ***
## SalePrice     0.000015634  0.000002633   5.938 0.000000002890 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1442.8  on 1451  degrees of freedom
## AIC: 1460.8
## 
## Number of Fisher Scoring iterations: 6
```

```
# remove OverallCond

logreg1j <- glm(Fireplace ~ LotArea+YearBuilt+CentralAir+GrLivArea+BedroomAbvGr+KitchenAbvGr+SalePrice,family=binomial, data=property)

summary(logreg1j)
```

```
##
## Call:
## glm(formula = Fireplace ~ LotArea + YearBuilt + CentralAir +
##     GrLivArea + BedroomAbvGr + KitchenAbvGr + SalePrice, family = binomial,
##     data = property)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9548  -0.8264   0.1014   0.8591   2.4755
##
## Coefficients:
##                 Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)  16.50511517  5.84049089   2.826         0.004714 **
## LotArea       0.00005123  0.00001744   2.937         0.003315 **
## YearBuilt    -0.01033522  0.00304856  -3.390         0.000698 ***
## CentralAirY   1.34305518  0.37418419   3.589         0.000332 ***
## GrLivArea     0.00203117  0.00029305   6.931 0.00000000000417 ***
## BedroomAbvGr -0.47810371  0.11234095  -4.256 0.00002082772605 ***
## KitchenAbvGr -1.66407538  0.38909680  -4.277 0.00001896292396 ***
## SalePrice     0.00001374  0.00000246   5.586 0.00000002329950 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1448.3  on 1452  degrees of freedom
## AIC: 1464.3
##
## Number of Fisher Scoring iterations: 6
```

```
# remove LotArea

logreg_final <- glm(Fireplace ~ YearBuilt+CentralAir+GrLivArea+BedroomAbvGr+KitchenAb
vGr+SalePrice,family=binomial, data=property)

summary(logreg_final)
```

```
## 
## Call:
## glm(formula = Fireplace ~ YearBuilt + CentralAir + GrLivArea + 
##     BedroomAbvGr + KitchenAbvGr + SalePrice, family = binomial, 
##     data = property)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.9372  -0.8292   0.1194   0.8672   2.4575  
## 
## Coefficients:
##                  Estimate   Std. Error z value         Pr(>|z|)    
## (Intercept)  20.267234899  5.703730901   3.553         0.000380 ***
## YearBuilt    -0.012241052  0.002982431  -4.104 0.00004053880374 ***
## CentralAirY   1.384082109  0.377326029   3.668         0.000244 ***
## GrLivArea     0.001982795  0.000290198   6.833 0.00000000000834 ***
## BedroomAbvGr -0.420277650  0.109500514  -3.838         0.000124 ***
## KitchenAbvGr -1.647231254  0.385425421  -4.274 0.00001921692248 ***
## SalePrice     0.000015717  0.000002383   6.595 0.00000000004249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1458.3  on 1453  degrees of freedom
## AIC: 1472.3
## 
## Number of Fisher Scoring iterations: 5
```

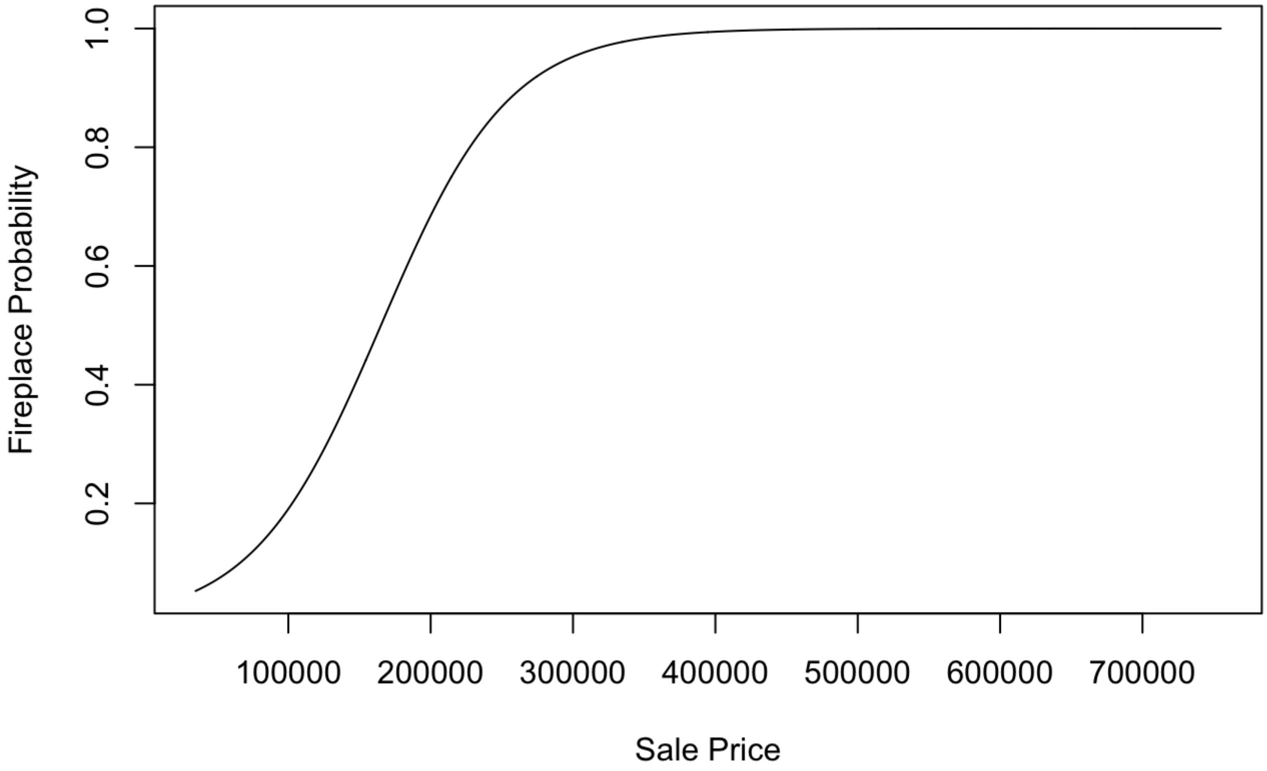
By excluding the variable with the highest p-value at each step, we obtain the final model (logreg_final) with all explanatory variables that are highly significant. Predictors included in this model are YearBuilt (original construction date), CentralAirY (central air conditioning), GrLivArea (above grade/ground living area square feet), BedroomAbvGr (bedrooms above grade), KitchenAbvGr (kitchens above grade) and SalePrice.

The following plots explore the relationship between Fireplace and SalePrice and between Fireplace and GrLivArea respectively under the specified logit model.

```
# plot SalePrice
logreg2 <- glm(Fireplace ~ SalePrice, family=binomial, data=property)

summary(logreg2)
```

```
##
## Call:
## glm(formula = Fireplace ~ SalePrice, family = binomial, data = property)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8155  -0.8933   0.1308   0.9729   2.2031
##
## Coefficients:
##                 Estimate   Std. Error z value            Pr(>|z|)
## (Intercept) -3.667076910  0.231675331  -15.83 <0.0000000000000002 ***
## SalePrice    0.000022214  0.000001378   16.12 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1564.9  on 1458  degrees of freedom
## AIC: 1568.9
##
## Number of Fisher Scoring iterations: 5
```

```
SalePricevals_logit <- seq(from=min(property$SalePrice),to=max(property$SalePrice),length=1200)

Fireplacevals_logit <- predict(logreg2, newdata=data.frame(SalePrice=SalePricevals_logit),type="response")

plot(x=SalePricevals_logit,y=Fireplacevals_logit,type="l",xlab="Sale Price",ylab="Fireplace Probability")
```

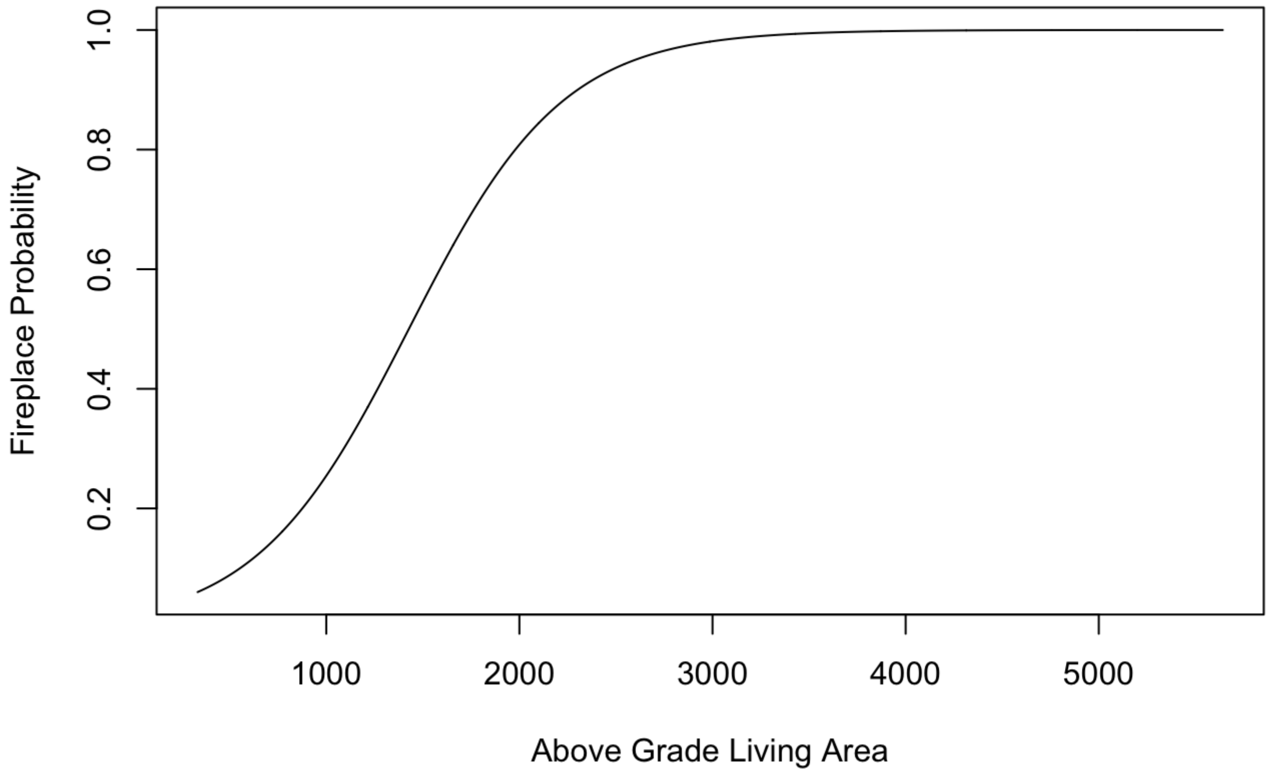


The plot above illustrates that for a property with a sale price above $300,000, there is a very high probability that a fireplace would come with the property.

```
# plot GrLivArea
logreg3 <- glm(Fireplace ~ GrLivArea, family=binomial, data=property)

summary(logreg3)
```

```
## 
## Call:
## glm(formula = Fireplace ~ GrLivArea, family = binomial, data = property)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.1473  -0.9317   0.3029   0.9630   1.9955  
## 
## Coefficients:
##               Estimate Std. Error z value            Pr(>|z|)    
## (Intercept) -3.5887642  0.2379747  -15.08 <0.0000000000000002 ***
## GrLivArea    0.0025138  0.0001606   15.65 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2019.6  on 1459  degrees of freedom
## Residual deviance: 1660.7  on 1458  degrees of freedom
## AIC: 1664.7
## 
## Number of Fisher Scoring iterations: 4
```

```
GrLivAreavals_logit <- seq(from=min(property$GrLivArea), to=max(property$GrLivArea),
length=1200)

Fireplacevals_logit_1 <- predict(logreg3, newdata=data.frame(GrLivArea=GrLivAreavals_
logit), type="response")

plot(x=GrLivAreavals_logit,y=Fireplacevals_logit_1, type="l", xlab="Above Grade Livin
g Area", ylab="Fireplace Probability")
```



Meanwhile, the plot with GrLivArea indicates that for a property with above grade living area greater than 3,000 square feet, there is a very high probability that the property has a fireplace.

Assessing the performance of logreg_final model

```
# create a test sample
n <- nrow(property)
testindex <- sample(1:n, size=n/3)
# test dataset
test <- property[testindex,]
nrow(test)
```

```
## [1] 486
```

```
# training dataset
train <- property[-testindex,]
nrow(train)
```

```
## [1] 974
```

```
# fit the logreg_final model to training data
logreg <- glm(Fireplace ~ YearBuilt+CentralAir+GrLivArea+BedroomAbvGr+KitchenAbvGr+Sa
lePrice,family=binomial,data=train)
# calculate predicted probabilities for the test data
testprob <- predict(logreg, newdata=test,type="response")
length(testprob)
```

```
## [1] 486
```

```
# compare the prediction from the classifier using the test data predictors with the
 actual responses of the test data
testpred <- rep("No",nrow(test))
testpred[testprob>0.5] <- "Yes"
table(testpred)
```

```
## testpred
##  No Yes
## 216 270
```

```
table(test$Fireplace)
```

```
##
##   N   Y
## 237 249
```

```
# Confusion Matrix
confmatrix <- table(test$Fireplace, testpred)
# True Positive Rate
TPR <- confmatrix[2,2]/(confmatrix[2,2]+confmatrix[2,1])
#False Positive Rate
FPR <- confmatrix[1,2]/(confmatrix[1,1]+confmatrix[1,2])
# Misclassification Rate
MR <- (confmatrix[1,2]+confmatrix[2,1])/nrow(test)
```

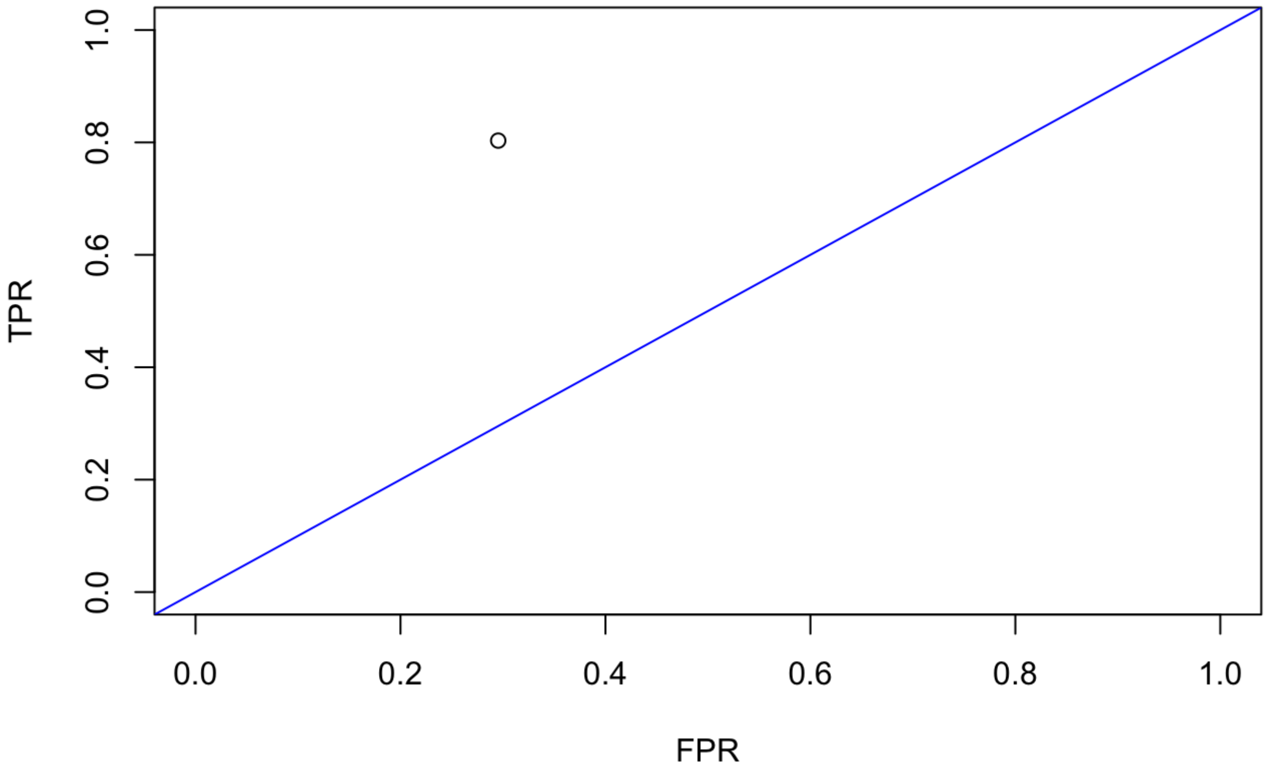
In order to assess the performance of the model logreg_final, we test the model by splitting the dataset into training and test dataset. A training dataset is created by randomly taking 2/3 of the whole dataset while the rest 1/3 of the dataset is used to test the model created using the training data.

To assess the performance of the model, a confusion matrix is created to compare the vector testpred with the default column in the test dataset.

The result for the confusion matrix is presented as below.

TPR = TP / (TP + FN) = 184 / 262 = 0.7022901 FPR = FP / (TN + FP) = 50 / 224 = 0.2232143 MR = (FP + FN) / n = (50 + 78) / 486 = 0.2633745

```
# ROC plot
plot(FPR, TPR, xlim=c(0,1), ylim=c(0,1) )
abline(0,1, col="blue")
```



When the result is presented in a ROC plot as shown below, we can see that the classifier is in the top left of the graph and above the 45-degree line, indicating that this is a relatively good classifier.

In addition, the mis-classification rate is 0.2633745, which is also moderately low.