

# Housing Prices Exploration Data Analysis

Pushkar Chougule

Nov 20th 2020

## R Markdown

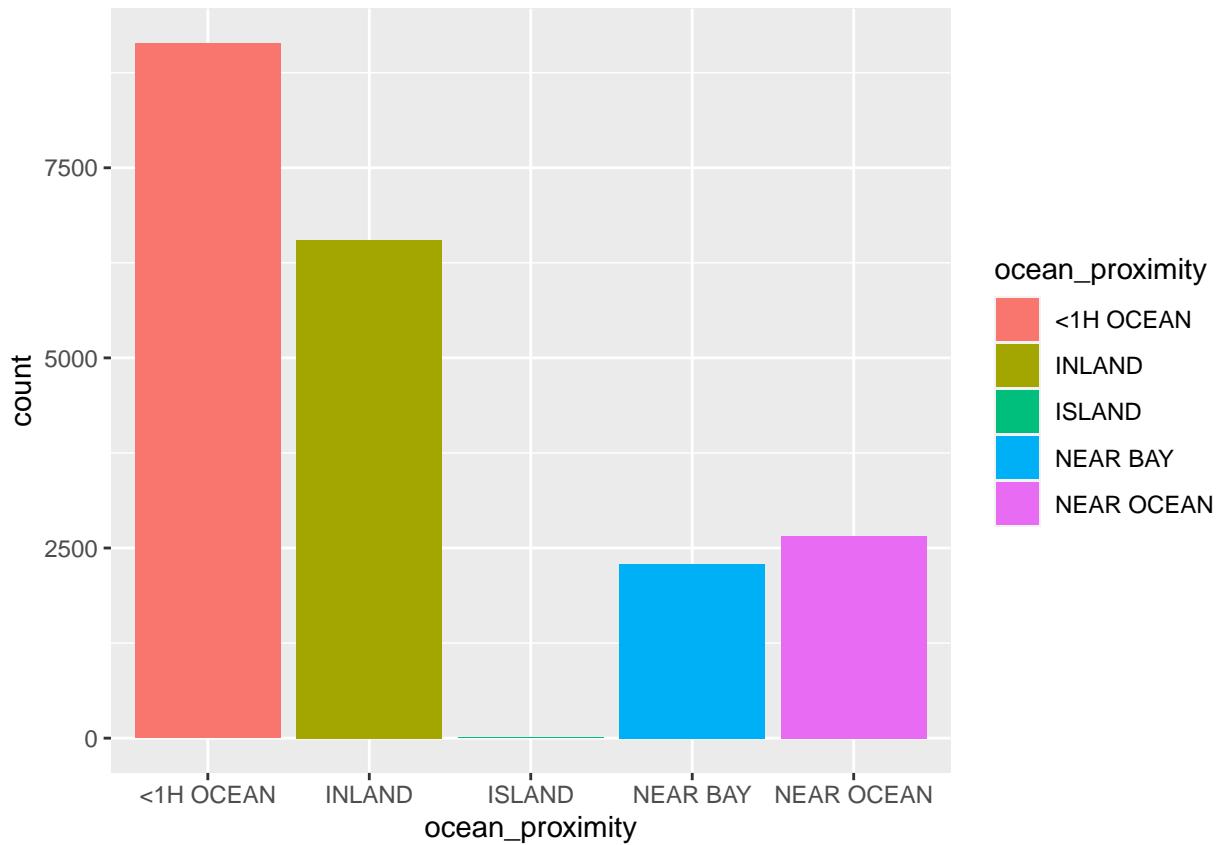
```
library(ggplot2)
library(car)

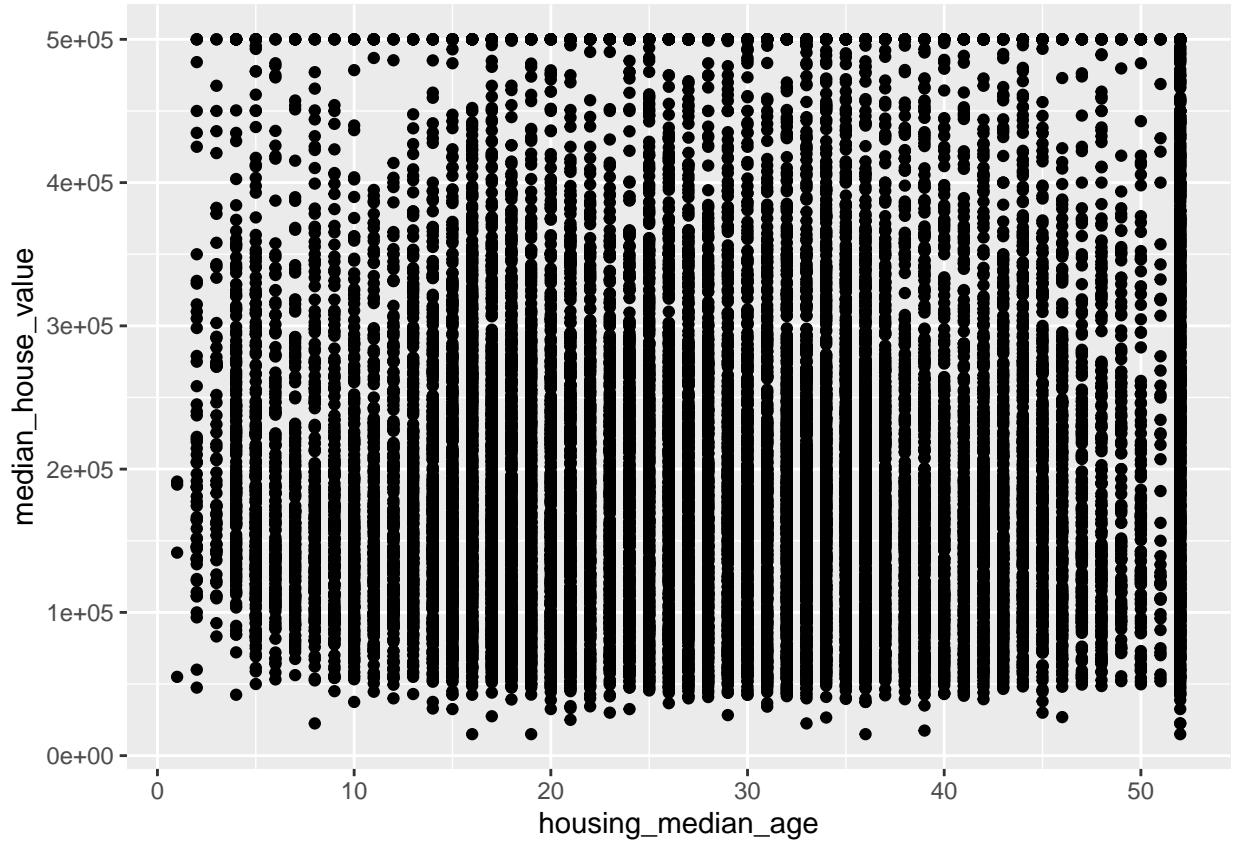
## Loading required package: carData

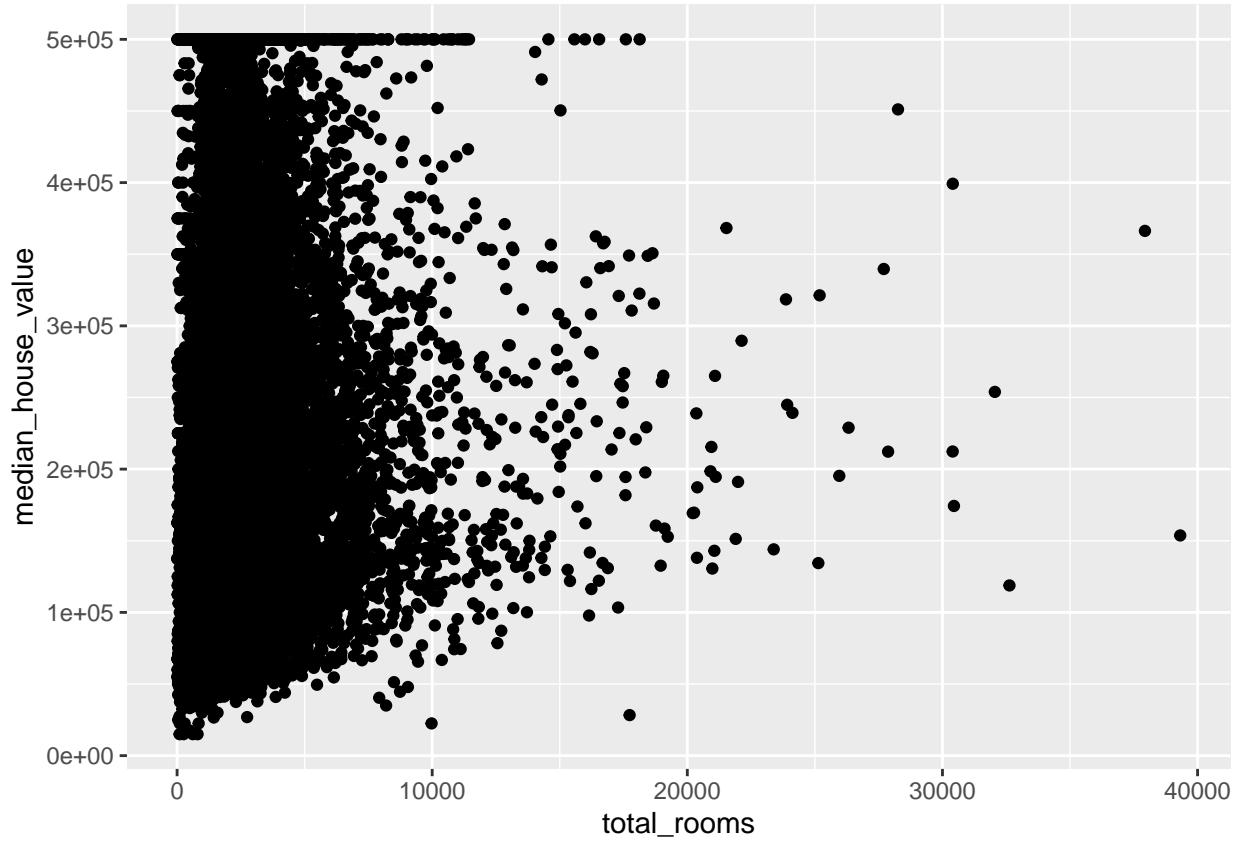
initial reading of California Housing prices csv

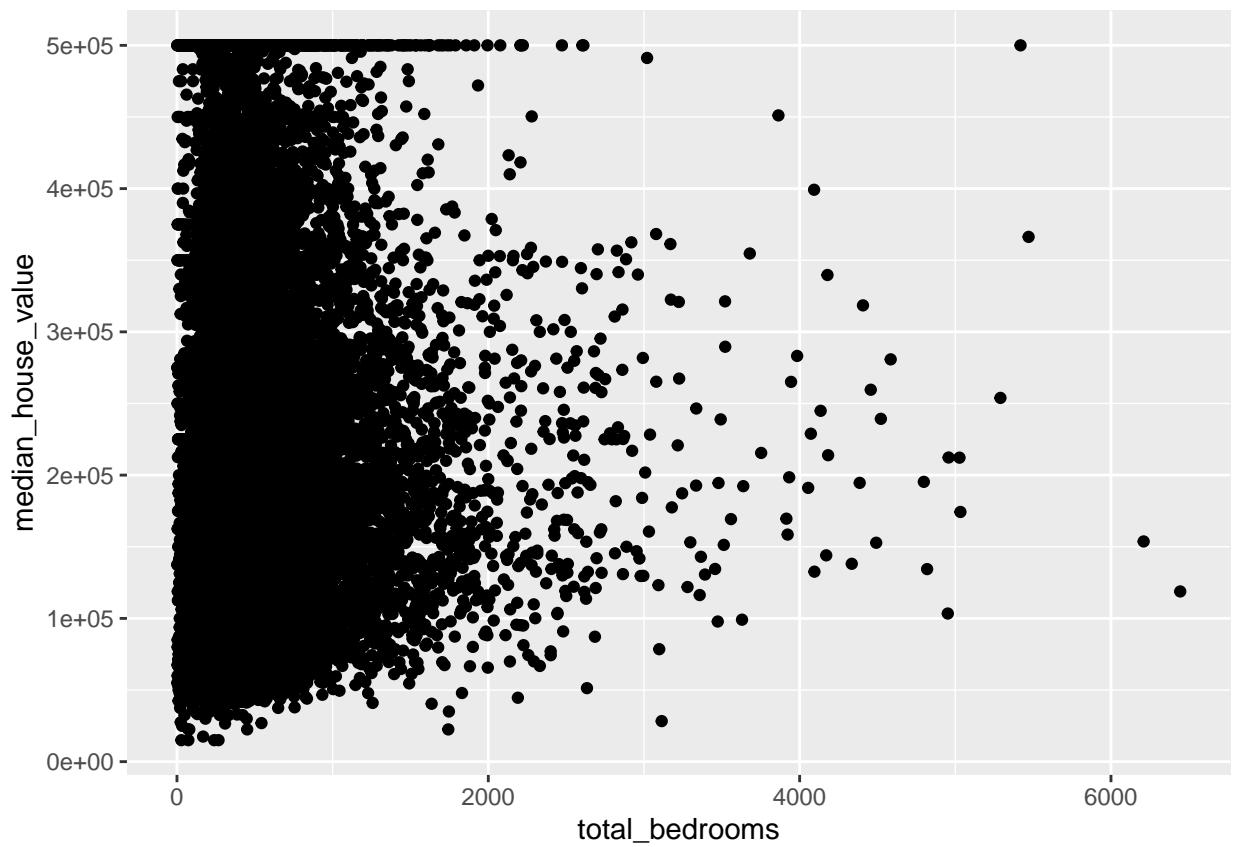
## 'data.frame': 20640 obs. of 10 variables:
## $ longitude      : num -122 -122 -122 -122 -122 ...
## $ latitude       : num 37.9 37.9 37.9 37.9 37.9 ...
## $ housing_median_age: num 41 21 52 52 52 52 52 52 42 52 ...
## $ total_rooms     : num 880 7099 1467 1274 1627 ...
## $ total_bedrooms  : num 129 1106 190 235 280 ...
## $ population      : num 322 2401 496 558 565 ...
## $ households      : num 126 1138 177 219 259 ...
## $ median_income    : num 8.33 8.3 7.26 5.64 3.85 ...
## $ median_house_value: num 452600 358500 352100 341300 342200 ...
## $ ocean_proximity  : Factor w/ 5 levels "<1H OCEAN","INLAND",...: 4 4 4 4 4 4 4 4 4 ...
```

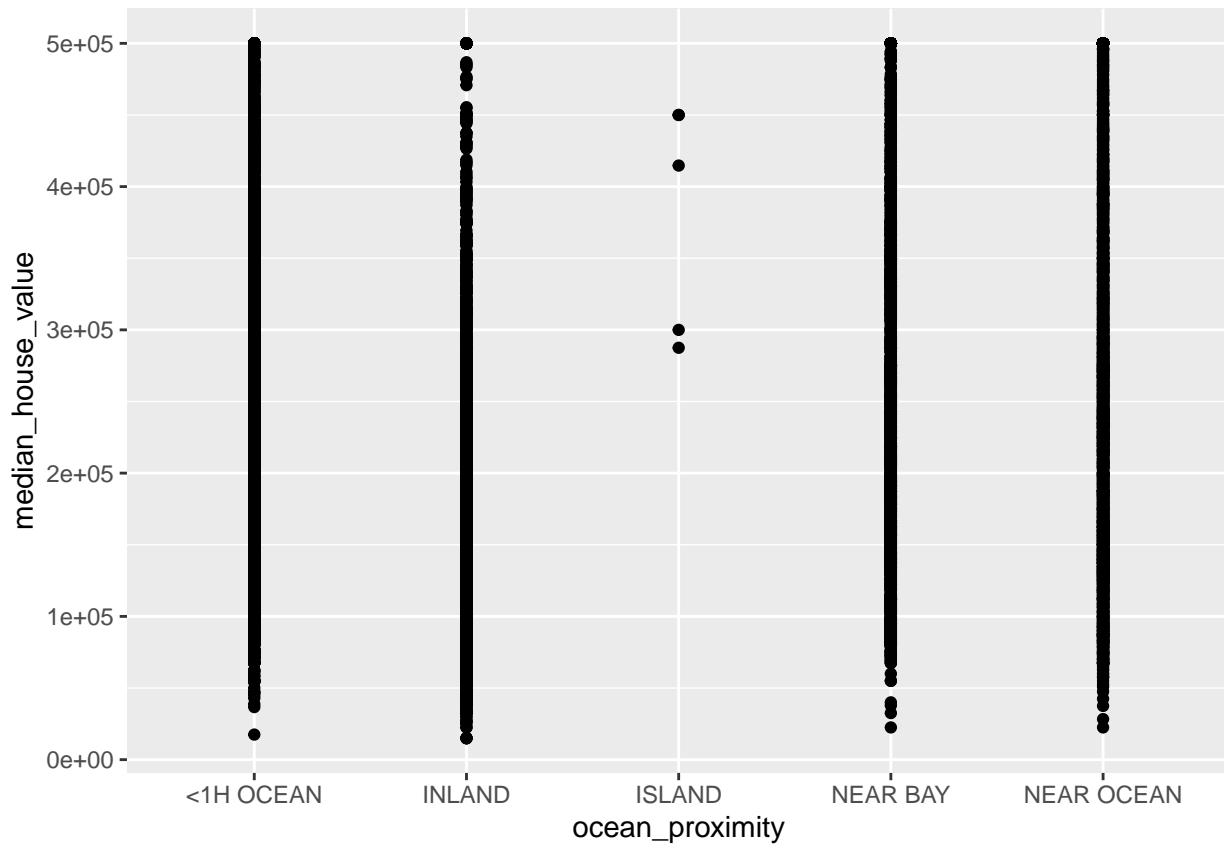
histograms and scatter plots for initial analysis

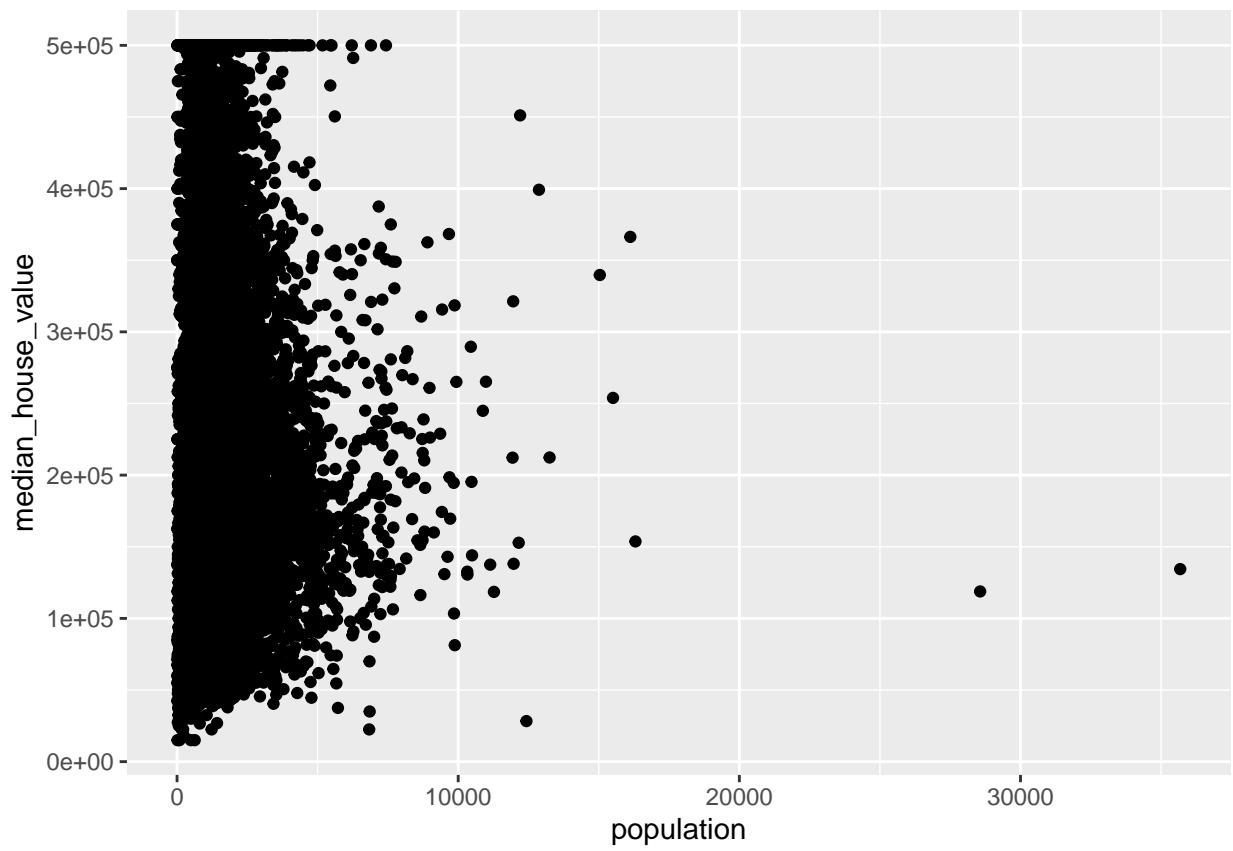


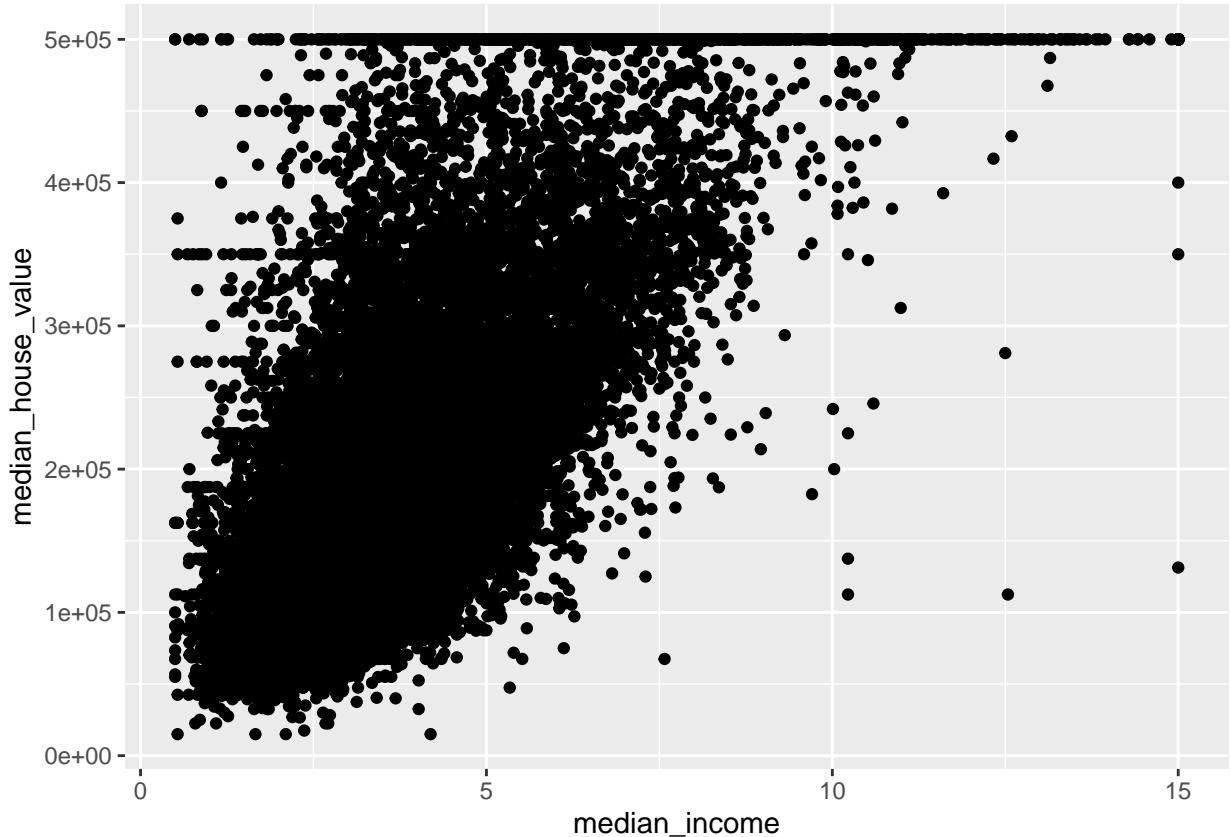












- 1) There are 5 distinct categories in the ocean proximity variables
- 2) Housing median age vs. median house value scatter plot
- 3) scatter plot for total rooms vs. median house value
- 4) scatter plot for total bedrooms vs. median house value
- 5) Median House value price ranges are similar in each of the ocean proximity category, looking at the scatter plot
- 6) scatter plot for population vs median house value
- 7) median income vs median house value

California Housing prices dataset has 10 columns including median house value and population, ocean proximity etc. there was not much significance and data was almost equally distributed. Thus not enough variables available for analysis. Also, when looked at the plots between. The total number of rooms, bedrooms data is not in standard format and would skew other dataset info, if merged with it. Hence decided not to use this dataset.

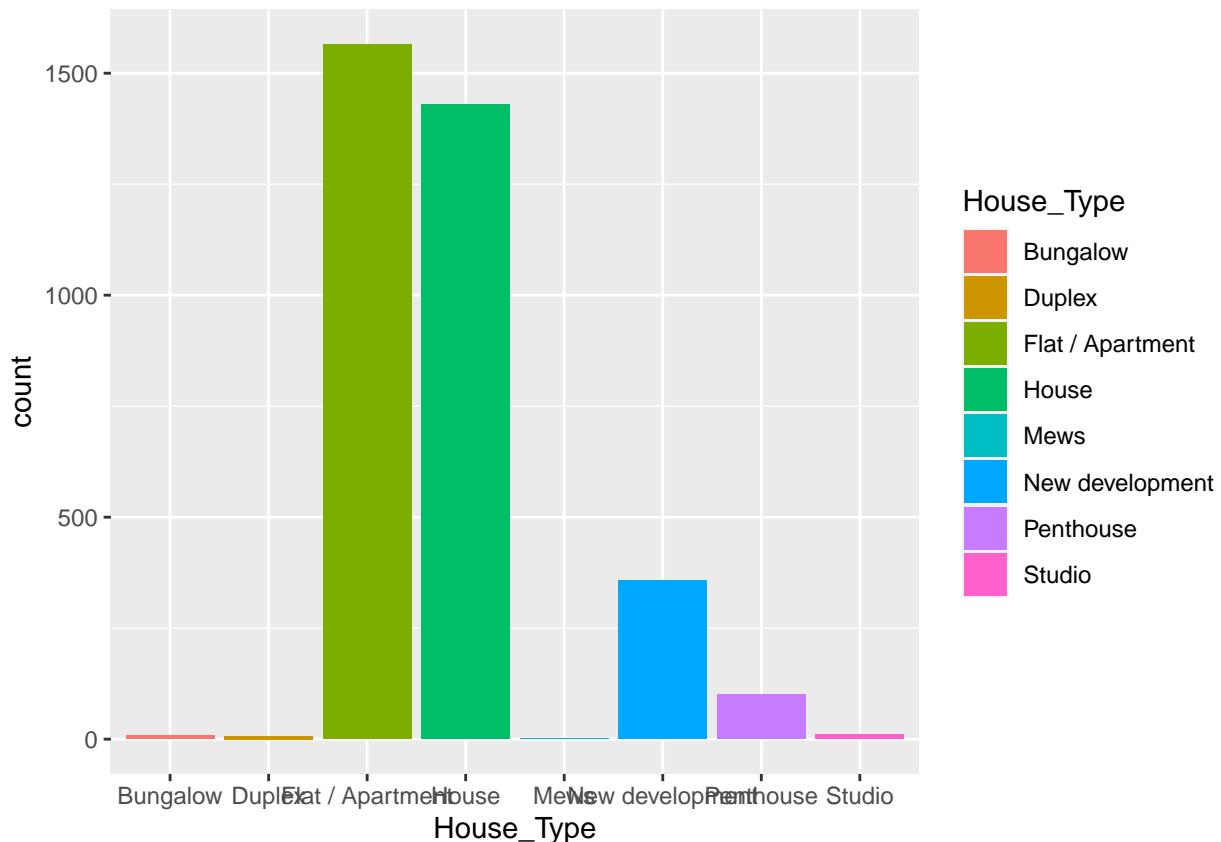
#### initial reading of London Housing prices csv

```

## 'data.frame': 3480 obs. of 6 variables:
## $ Price      : int 1675000 650000 735000 1765000 675000 420000 1475000 650000 2500000 925000 ...
## $ House_Type : Factor w/ 8 levels "Bungalow","Duplex",...: 4 3 3 4 3 3 4 6 4 3 ...
## $ Area_sqft   : int 2716 814 761 1986 700 403 1548 560 1308 646 ...
## $ Num_Bedrooms: int 5 2 2 4 2 1 4 1 3 2 ...
## $ Num_Bathrooms: int 5 2 2 4 2 1 4 1 3 2 ...
## $ Num_Receptions: int 5 2 2 4 2 1 4 1 3 2 ...

```

### London housing data initial analysis histogram



London Housing prices dataset has just about 11 column variables and some of those variables do not seem to have relation for being picked as predictor (Sequence ID, Property name, Blank / Invalid Location values and in some cases consist of partial street address). Also, the sale prices are in Pounds, which may not be relevant. So decided not to use the dataset.

### initial reading of US Housing prices csv file and summary

```

##    MSSubClass     MSZoning     LotFrontage     LotArea     BldgType
##  20      :536     C (all): 10     Min.   : 0.00     Min.   : 1300  1Fam  :1220
##  60      :299     FV       : 65     1st Qu.: 42.00     1st Qu.: 7554  2fmCon: 31
##  50      :144     RH       : 16     Median  : 63.00     Median  : 9478  Duplex: 52
## 120      : 87     RL       :1151     Mean    : 57.62     Mean    :10517   Twnhs : 43
##  30      : 69     RM       :218      3rd Qu.: 79.00     3rd Qu.:11602   TwnhsE: 114
## 160      : 63                         Max.   :313.00     Max.   :215245
## (Other):262

```

```

##   HouseStyle LotConfig      Neighborhood      Condition1
## 1Story :726 Length:1460      Length:1460      Length:1460
## 2Story :445 Class :character Class :character Class :character
## 1.5Fin :154 Mode :character Mode :character Mode :character
## SLvl  : 65
## SFoyer : 37
## 1.5Unf : 14
## (Other): 19
##   Condition2      Foundation      RoofStyle      RoofMatl
## Length:1460      Length:1460      Length:1460      Length:1460
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
## 
## 
## 
##   Exterior1st      Exterior2nd      ExterQual      HeatingQC
## Length:1460      Length:1460      Length:1460      Length:1460
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
## 
## 
## 
##   Electrical      OverallQual      OverallCond      YearBuilt      YearRemodAdd
## Length:1460      5      :397      5      :821      Min.    :1872      Min.    :1950
## Class :character 6      :374      6      :252      1st Qu.:1954      1st Qu.:1967
## Mode :character  7      :319      7      :205      Median   :1973      Median   :1994
##                      8      :168      8      : 72      Mean     :1971      Mean     :1985
##                      4      :116      4      : 57      3rd Qu.:2000      3rd Qu.:2004
##                      9      : 43      3      : 25      Max.    :2010      Max.    :2010
## (Other): 43      (Other): 28
##   MasVnrArea      MasVnrType      WoodDeckSF      OpenPorchSF
## Min.    : 0.0      BrkCmn : 15      Min.    : 0.00      Min.    : 0.00
## 1st Qu.: 0.0      BrkFace:445      1st Qu.: 0.00      1st Qu.: 0.00
## Median  : 0.0      None   :864      Median  : 0.00      Median  :25.00
## Mean    : 103.7     Stone  :128      Mean    : 94.24      Mean    :46.66
## 3rd Qu.: 166.0     Unk    :  8      3rd Qu.:168.00      3rd Qu.: 68.00
## Max.    :1600.0          NA's    :8          Max.    :857.00      Max.    :547.00
##   EnclosedPorch      BsmtFinSF1      TotalBsmtSF      BsmtFinType1
## Min.    : 0.00      Min.    : 0.0      Min.    : 0.0      Length:1460
## 1st Qu.: 0.00      1st Qu.: 0.0      1st Qu.: 795.8      Class :character
## Median  : 0.00      Median : 383.5      Median : 991.5      Mode  :character
## Mean    : 21.95     Mean   :443.6      Mean   :1057.4
## 3rd Qu.: 0.00      3rd Qu.: 712.2      3rd Qu.:1298.2
## Max.    :552.00     Max.   :5644.0      Max.   :6110.0
## 
##   BsmtQual      GrLivArea      FullBath      HalfBath
## Length:1460      Min.    :334      Min.    :0.000      Min.    :0.0000
## Class :character 1st Qu.:1130     1st Qu.:1.000     1st Qu.:0.0000
## Mode  :character  Median :1464      Median :2.000      Median :0.0000
##                      Mean   :1515      Mean   :1.565      Mean   :0.3829
##                      3rd Qu.:1777     3rd Qu.:2.000     3rd Qu.:1.0000

```

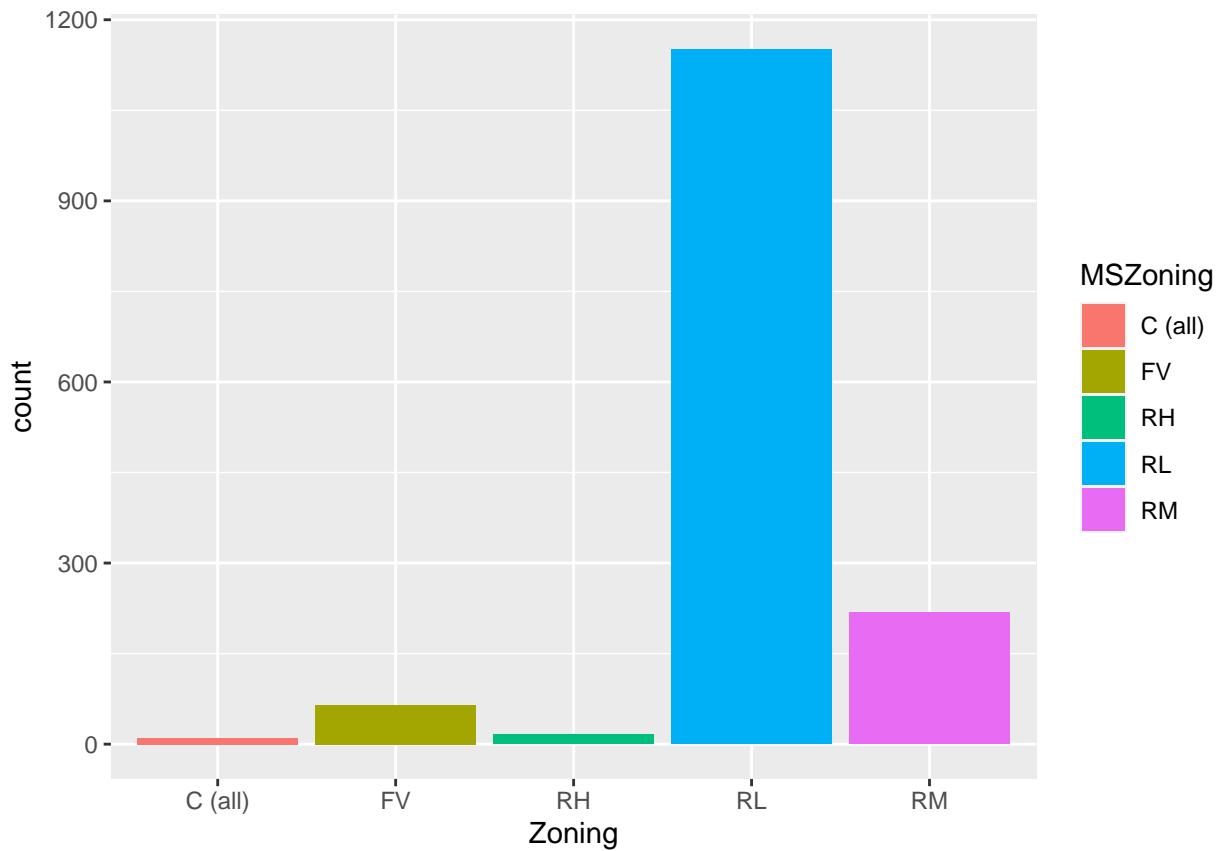
```

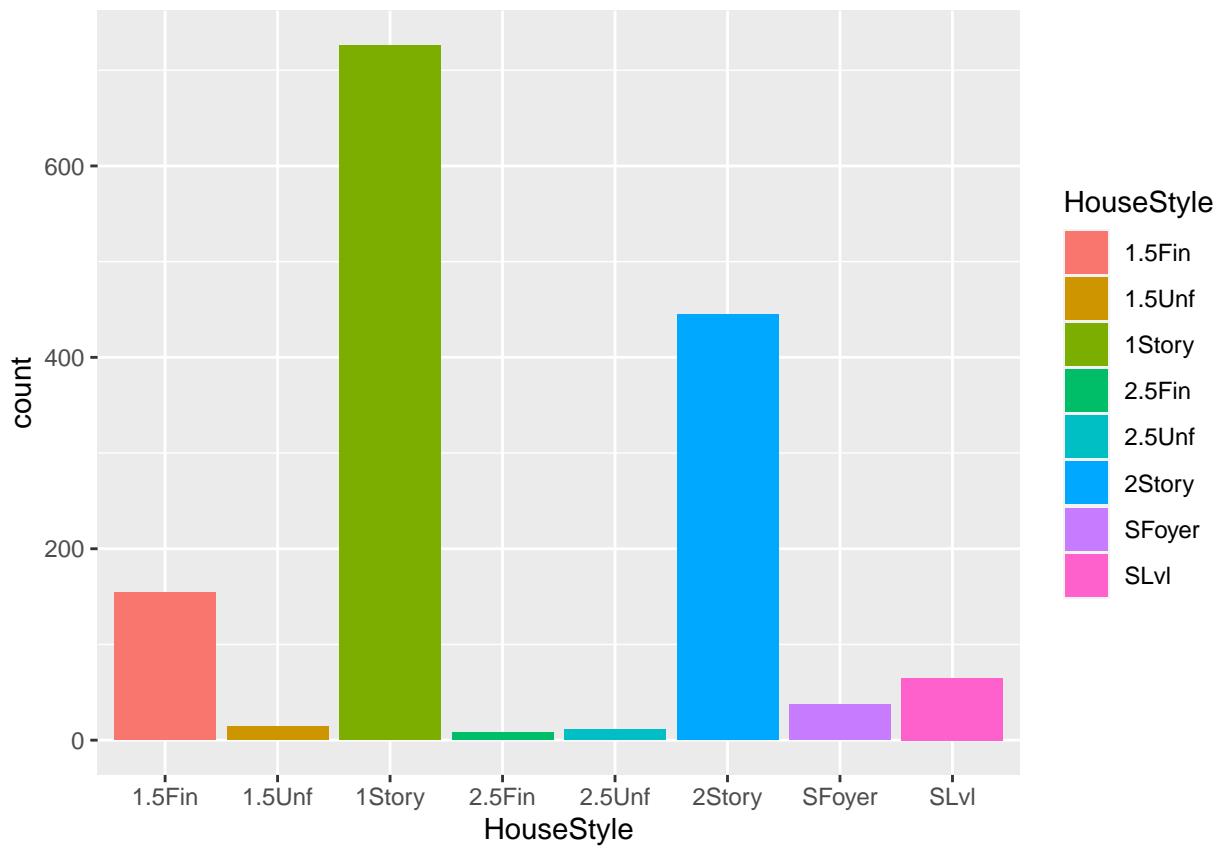
##          Max.    :5642   Max.    :3.000   Max.    :2.0000
##
##      BedroomAbvGr     TotRmsAbvGrd      GarageArea      Fence
##  Min.    :0.000   Min.    : 2.000   Min.    : 0.0   GdPrv:  59
##  1st Qu.:2.000   1st Qu.: 5.000   1st Qu.:334.5   GdWo : 54
##  Median :3.000   Median : 6.000   Median :480.0   MnPrv: 157
##  Mean   :2.866   Mean   : 6.518   Mean   :473.0   MnWw : 11
##  3rd Qu.:3.000   3rd Qu.: 7.000   3rd Qu.:576.0   NA's :1179
##  Max.   :8.000   Max.   :14.000   Max.   :1418.0
##
##      SalePrice
##  Min.   :34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##

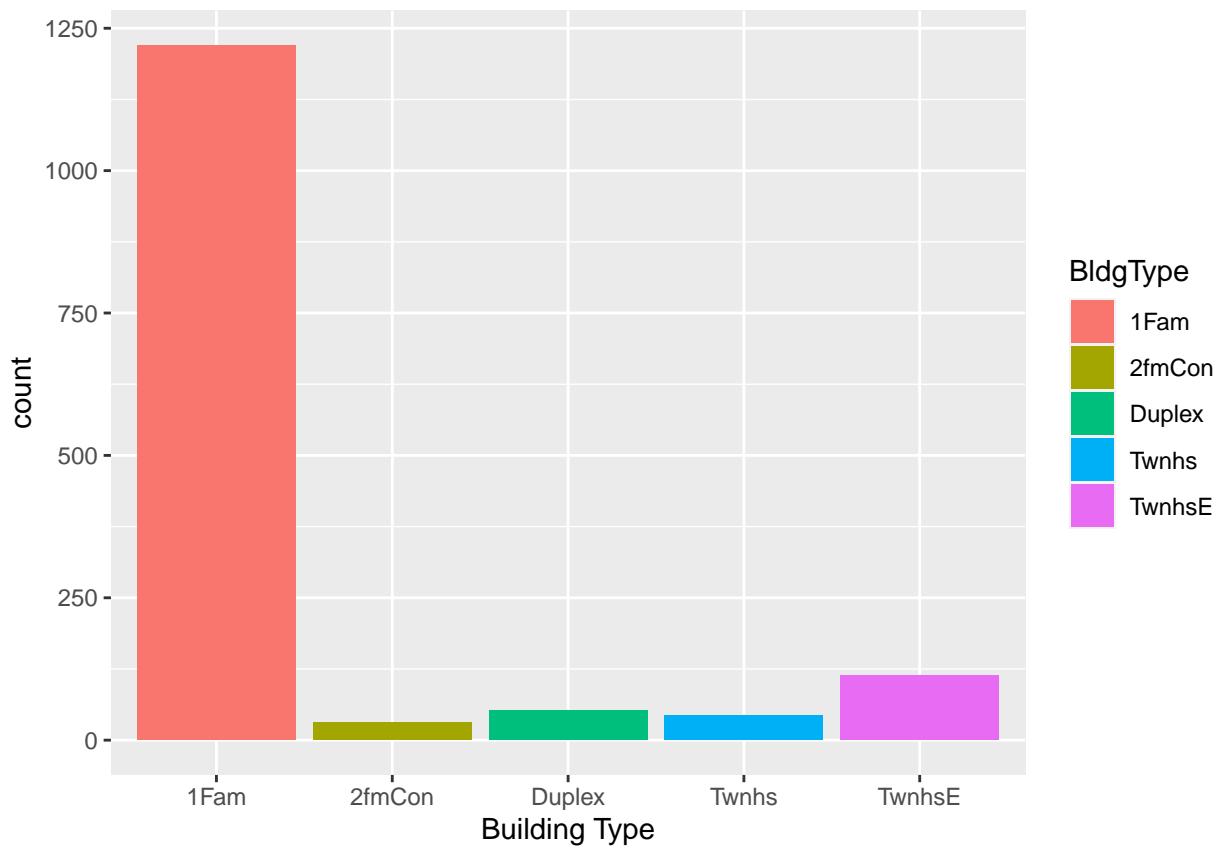
```

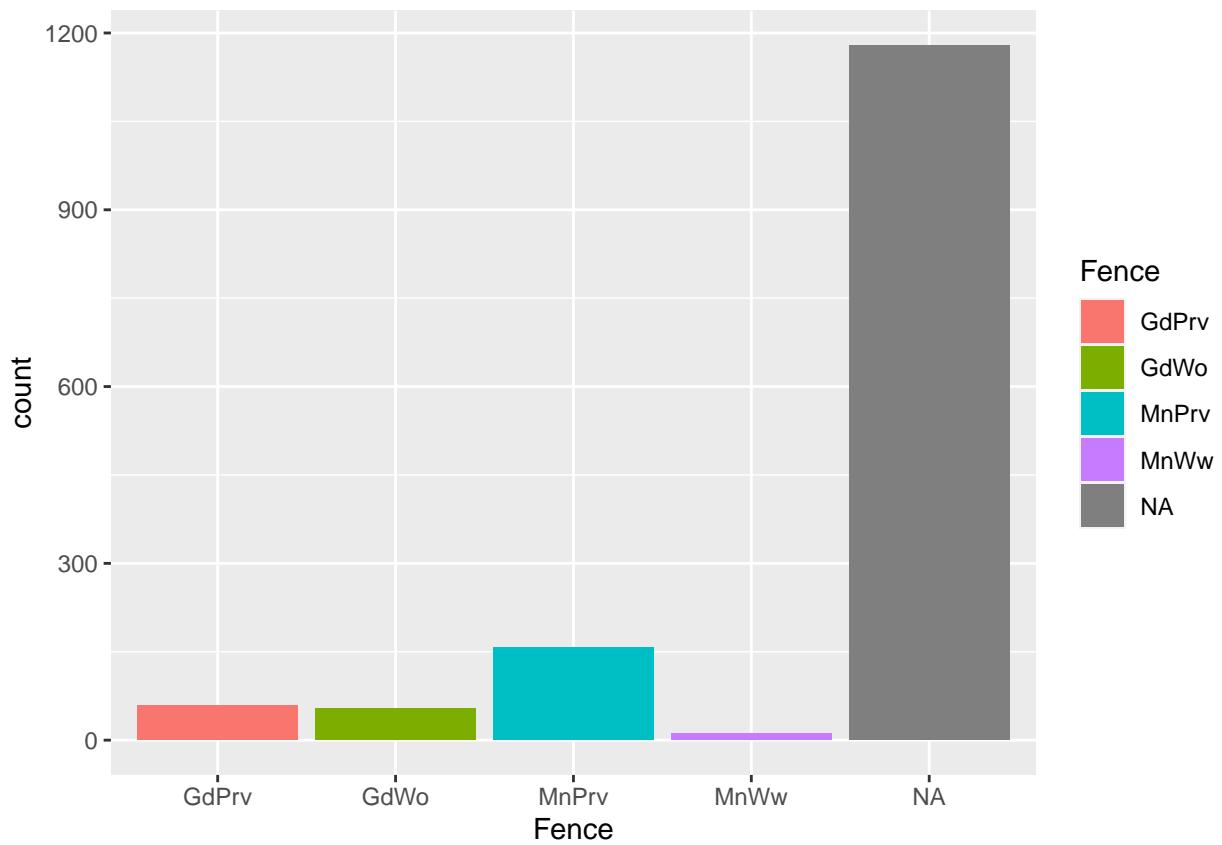
US Housing prices dataset has plenty of variables (81 in all) and hence I will be using this dataset for my analysis purpose. I have manually looked at the data variables and also used some of the plots to understand the data points / variables. I have tried to capture this information below. Noticed that TotalBsmtSF is the addition of BsmtFinSF1, BsmtFinSF2 and BsmtUnfSF. Similarly 1stFlrSF and 2ndFlrSF columns values are combined into and it is present in variable GrLivArea.

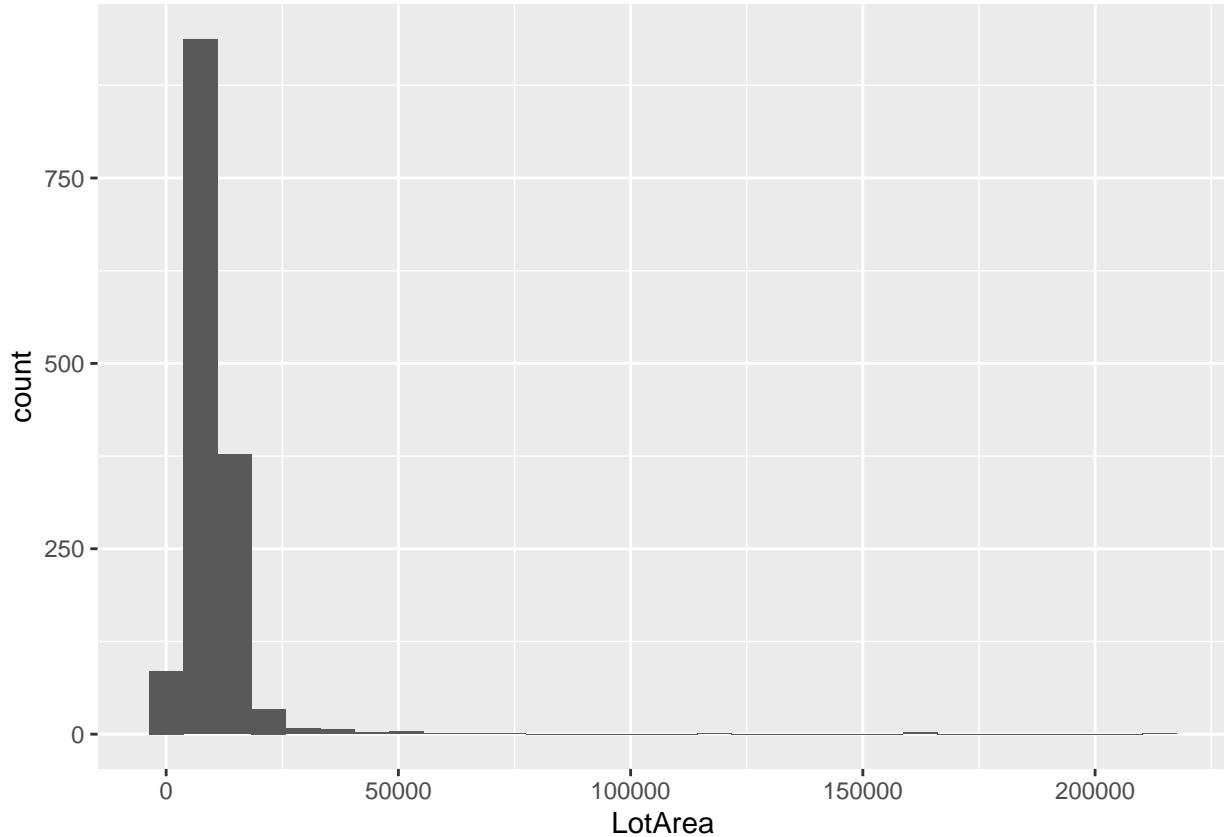
**histogram plots for initial analysis of data to also help with data cleanup steps**

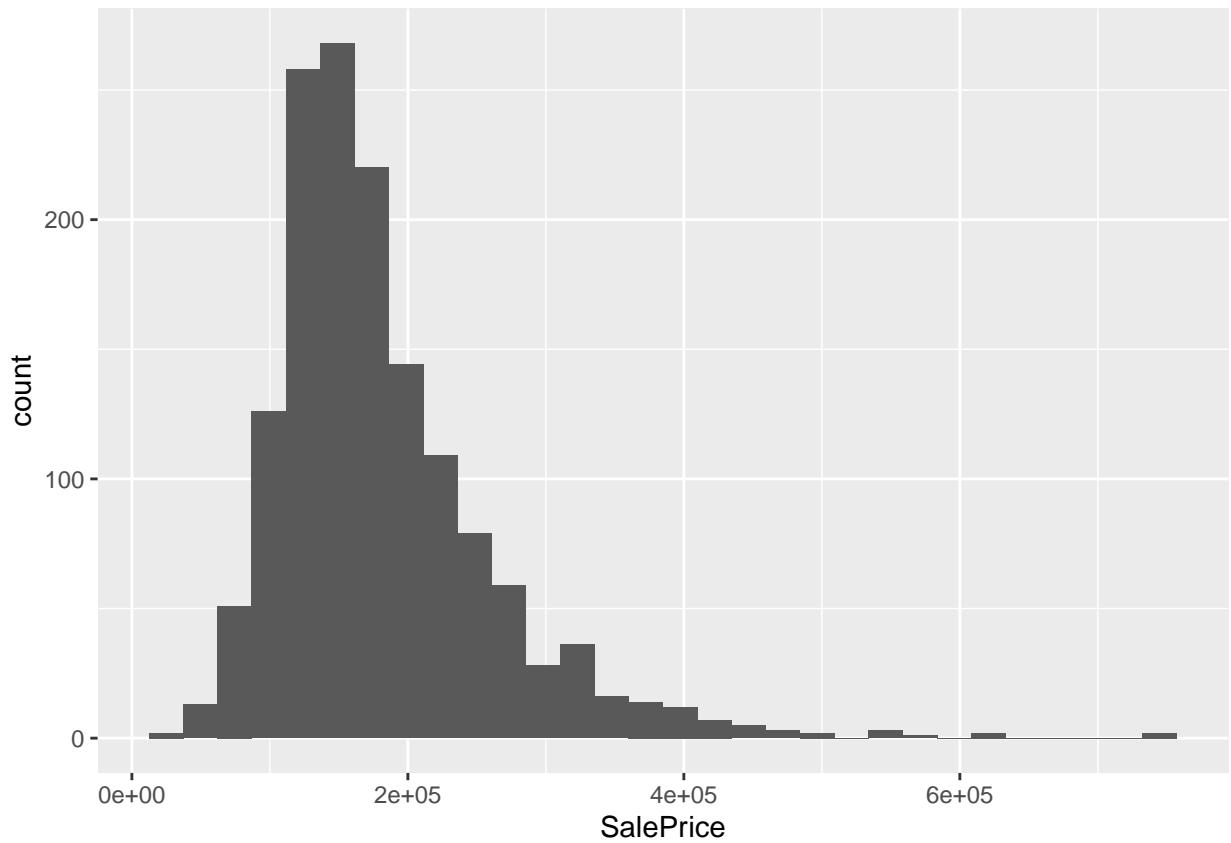




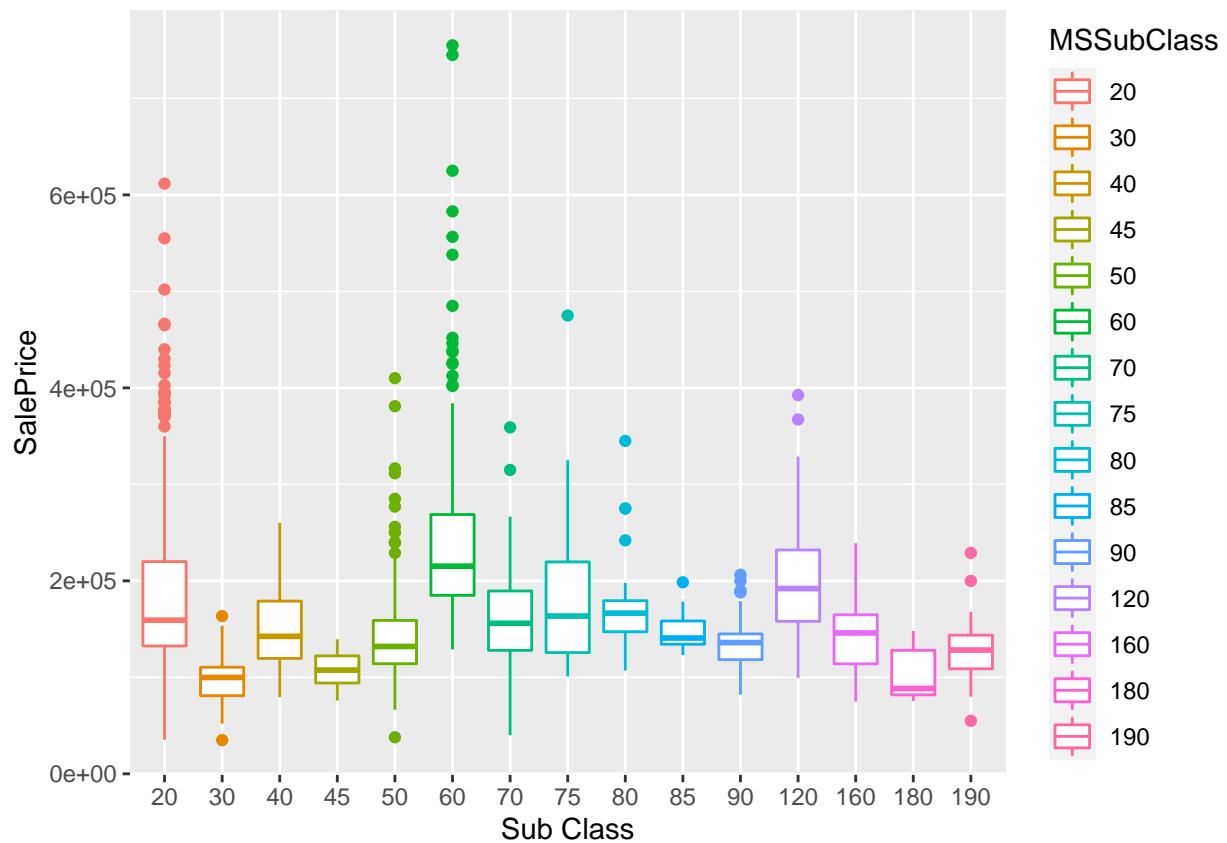


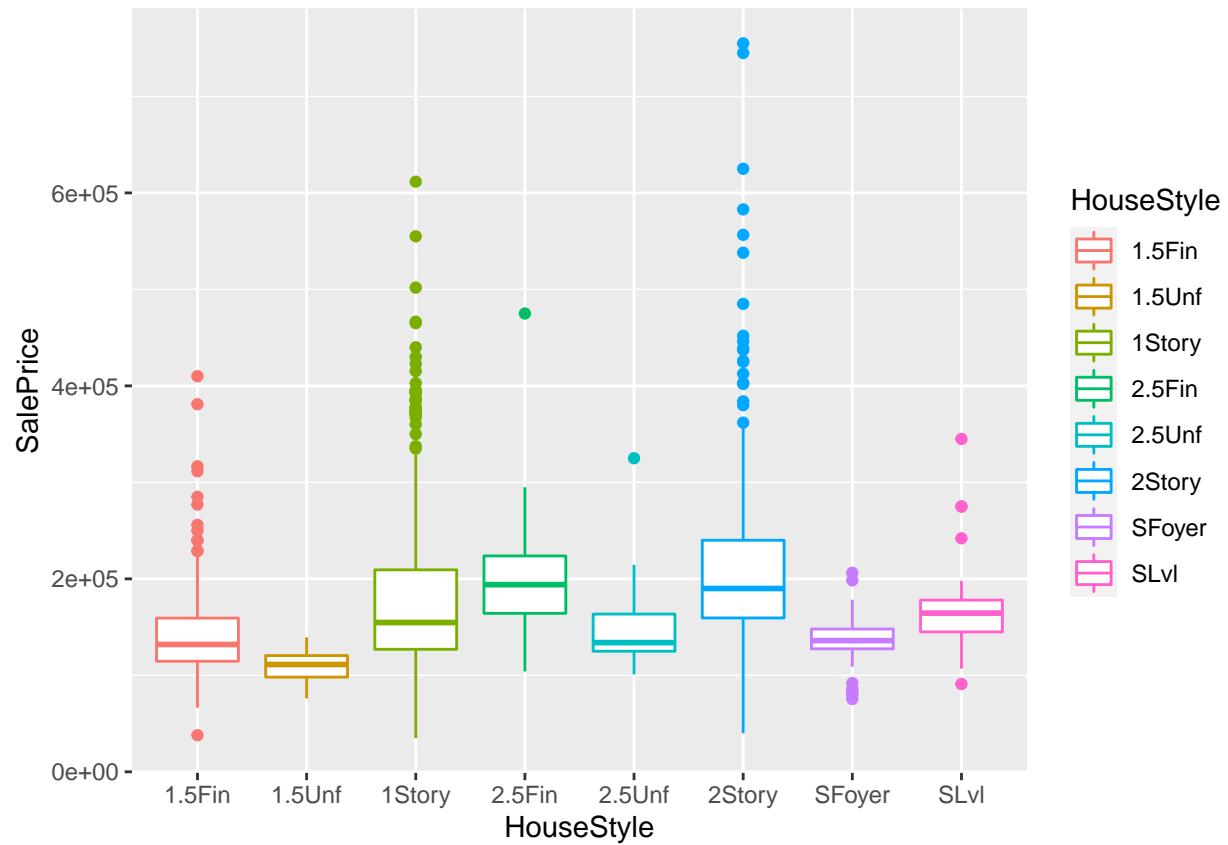


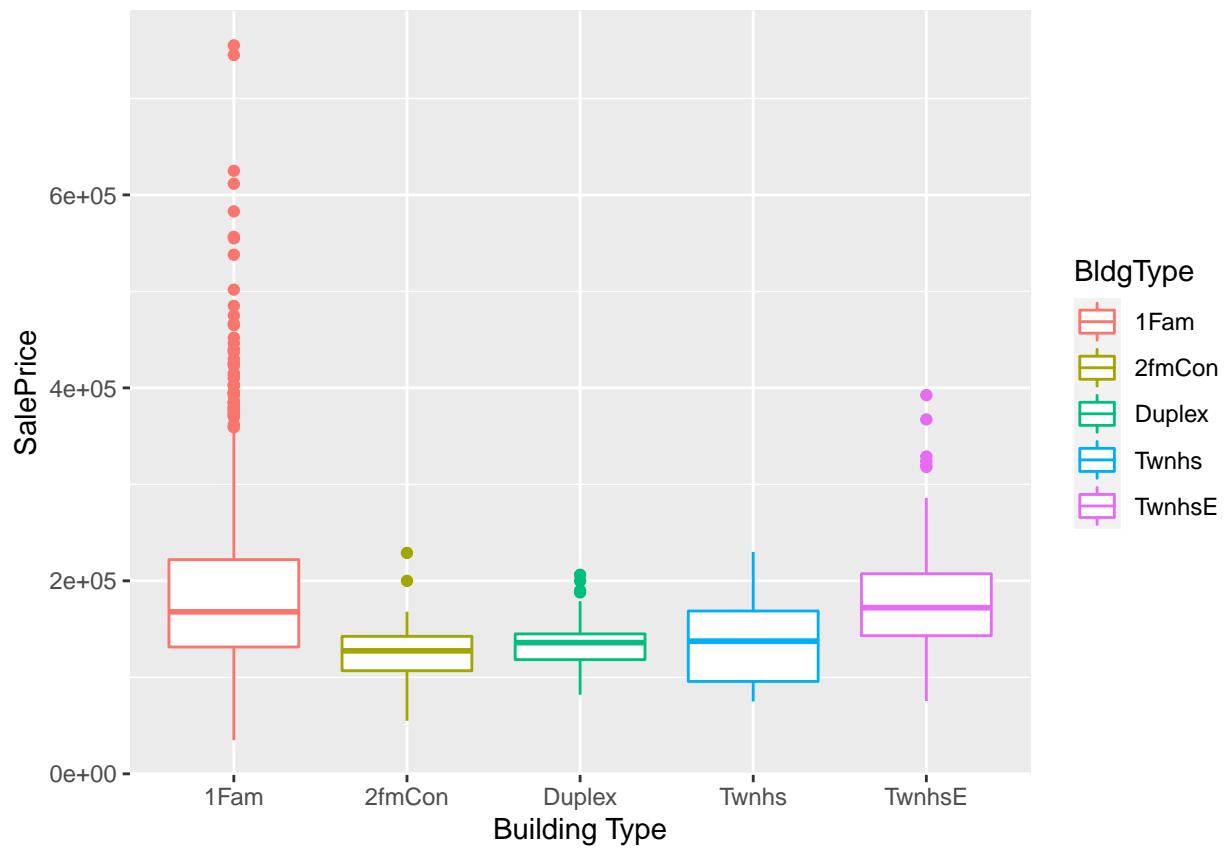


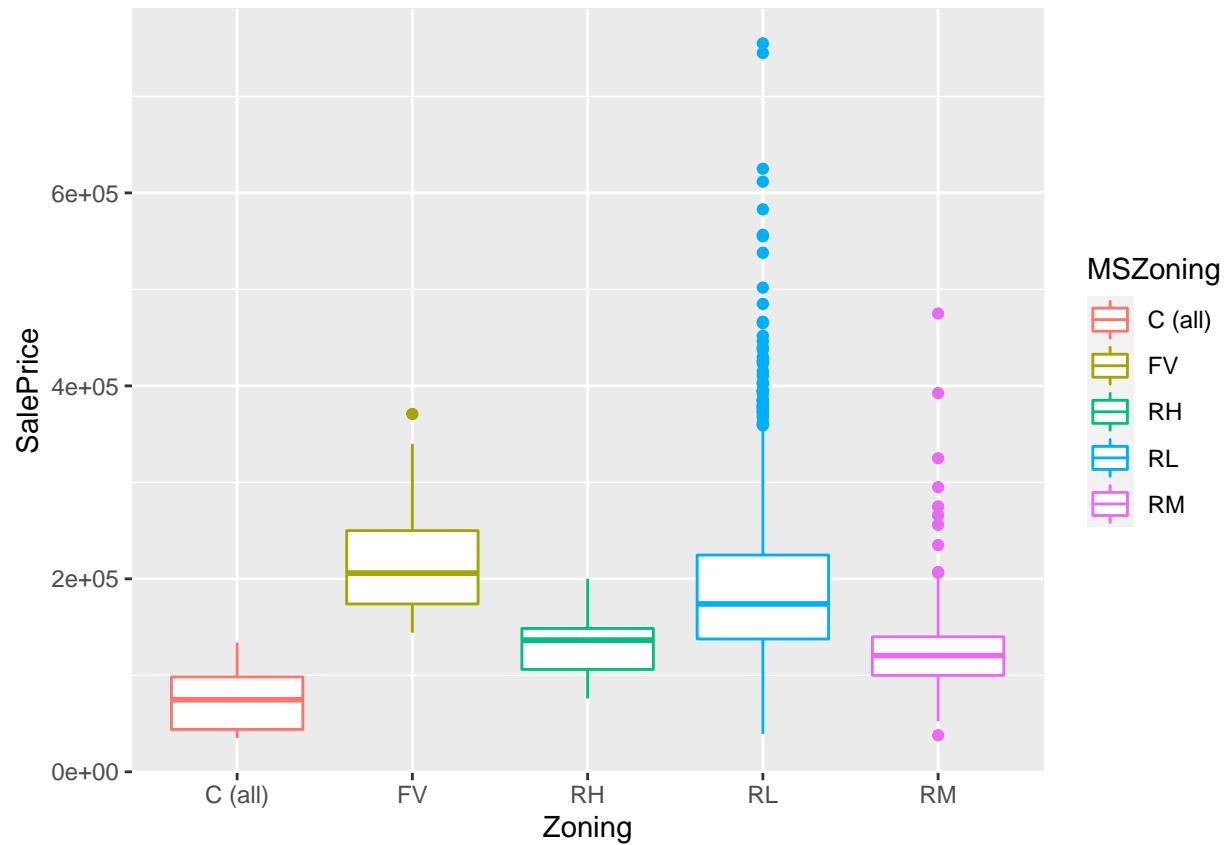


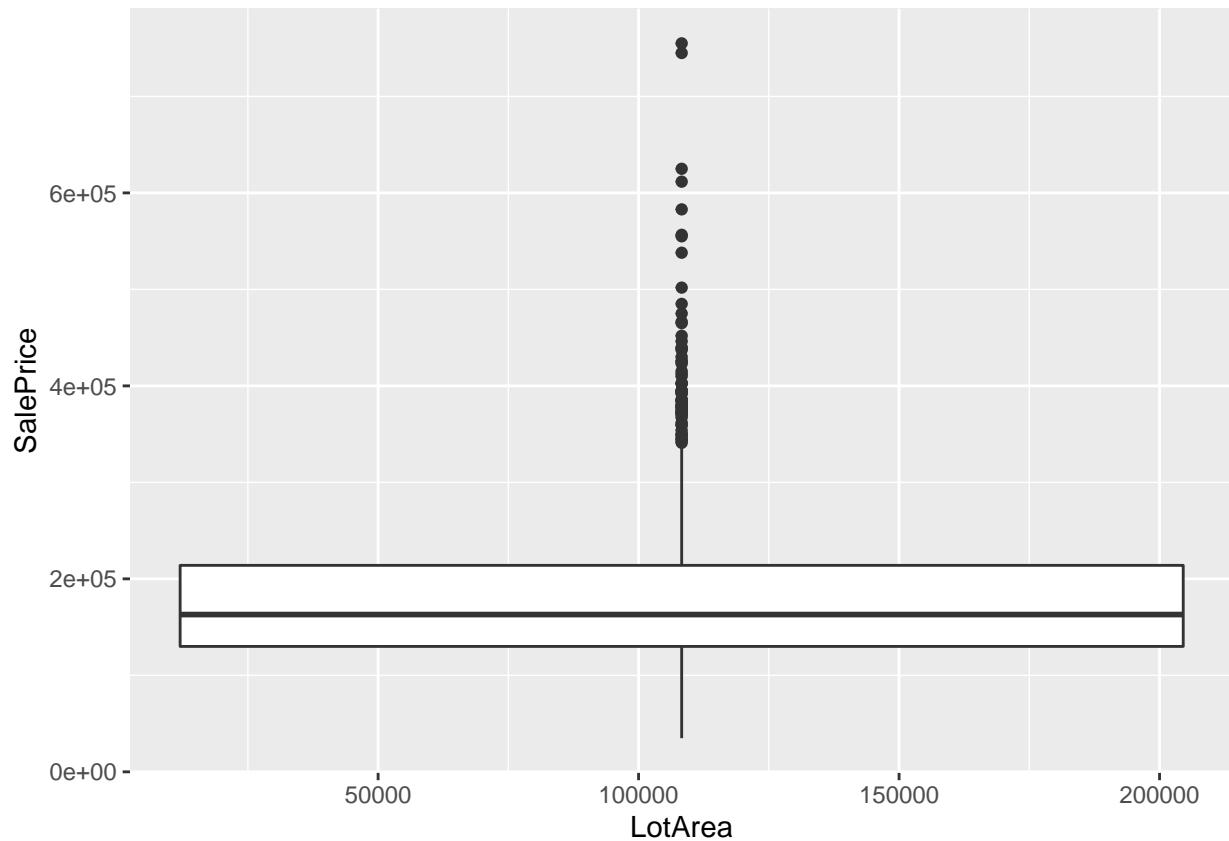
box plots and scatter plots for initial analysis of data to also help with data cleanup steps

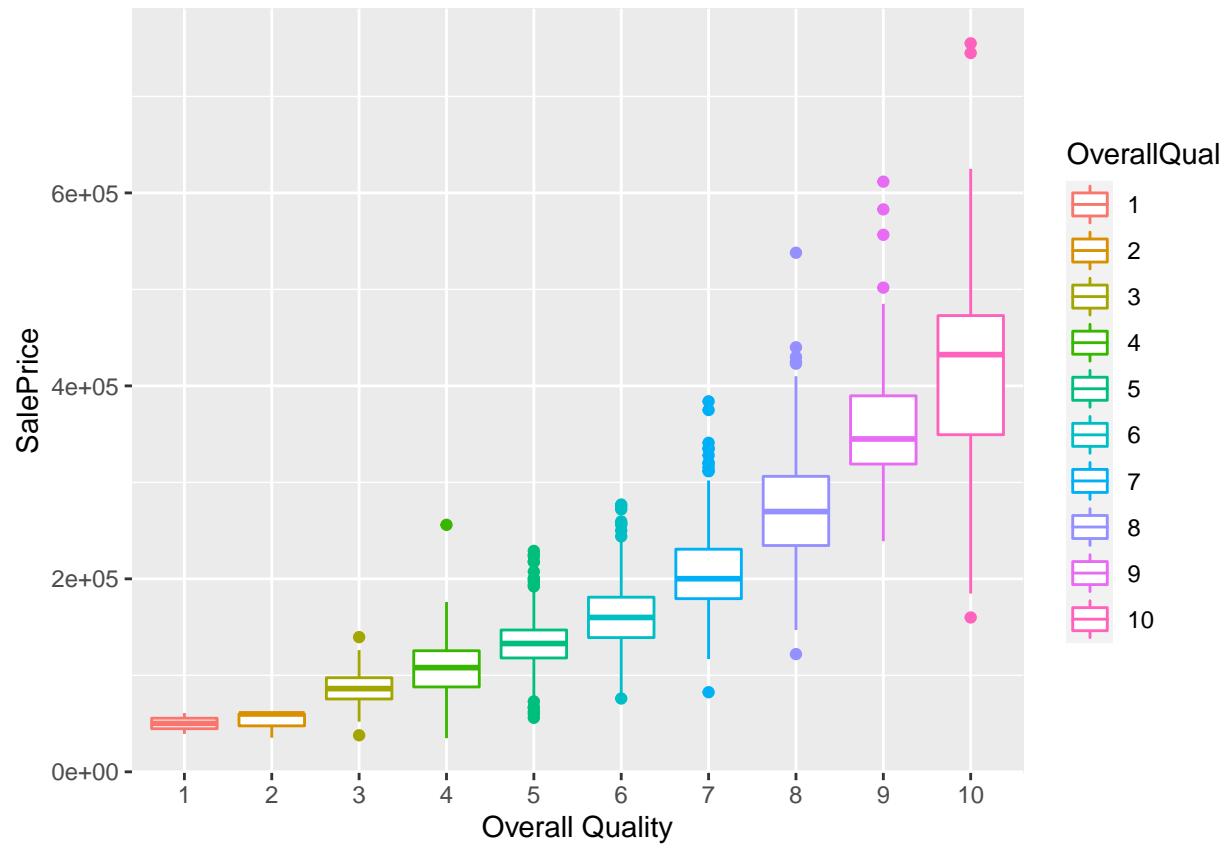


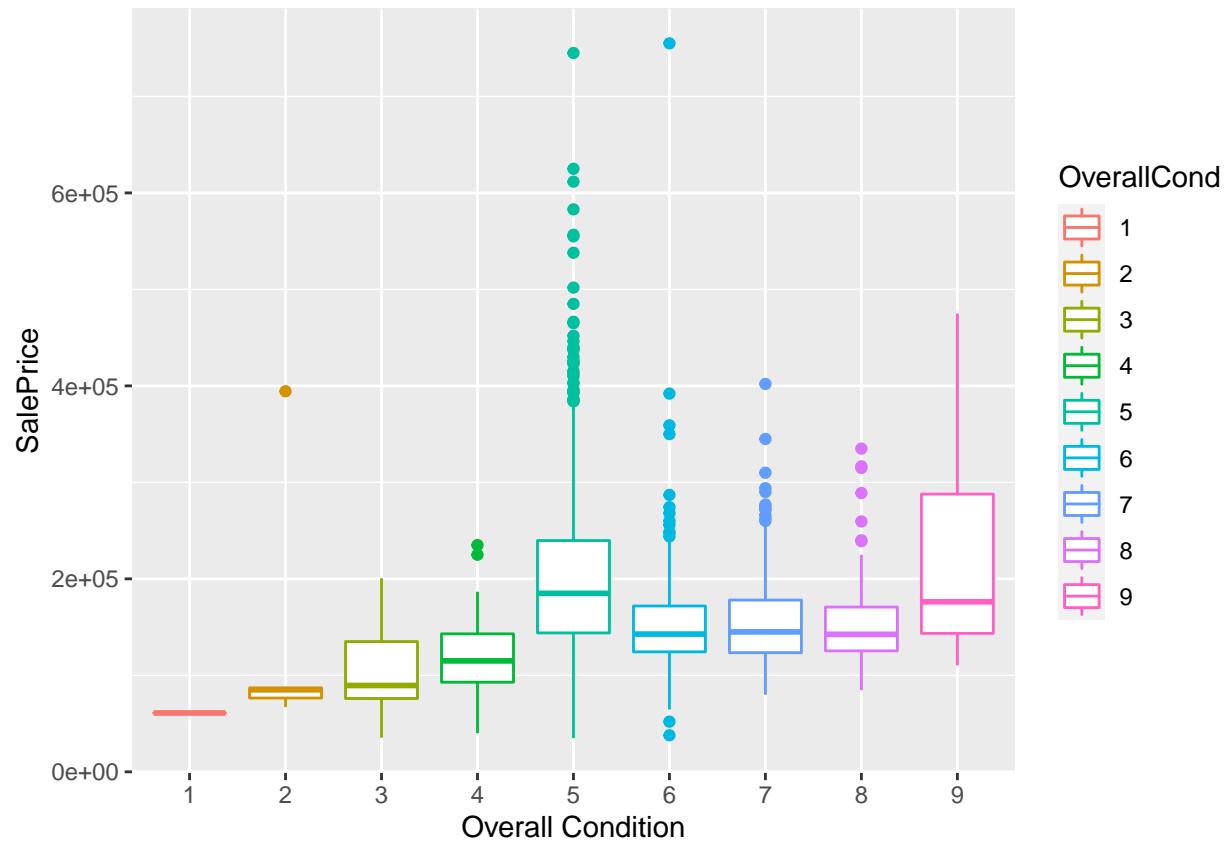


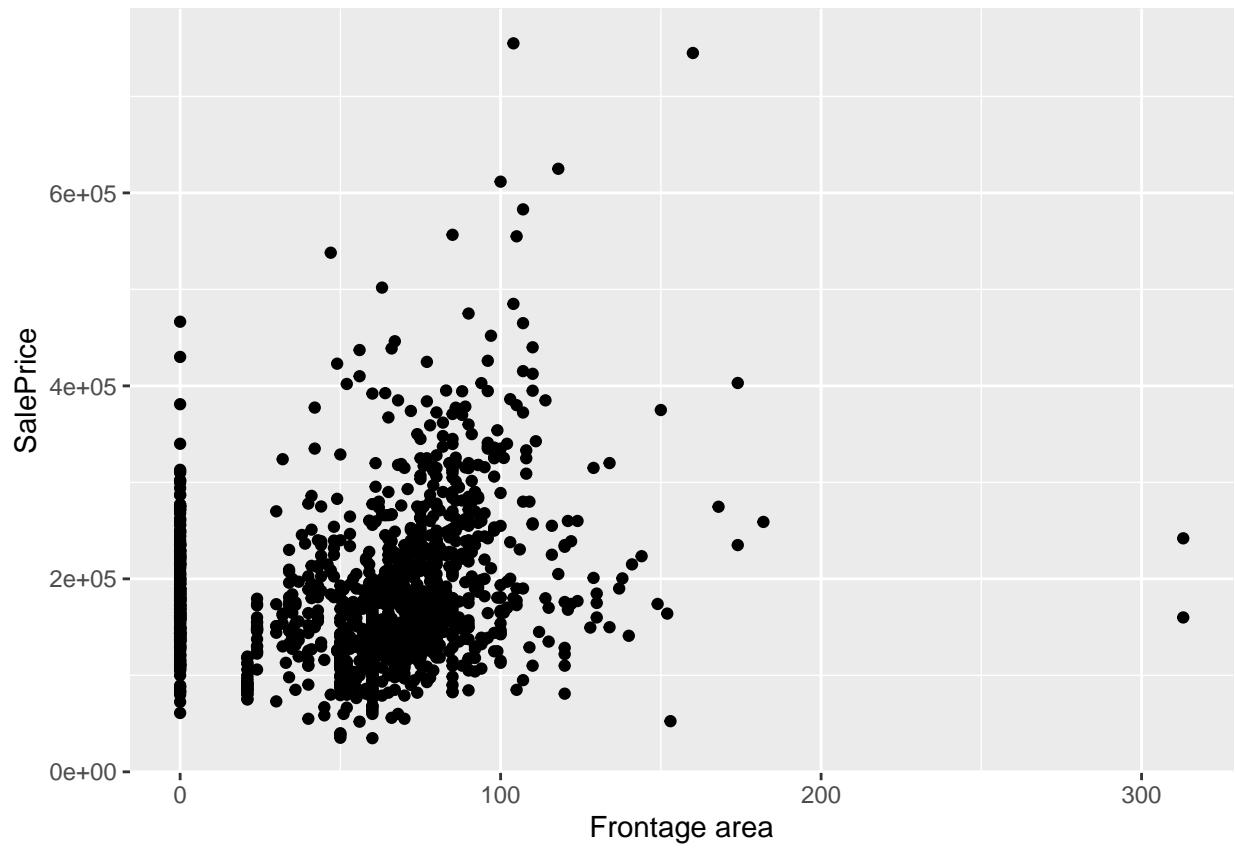


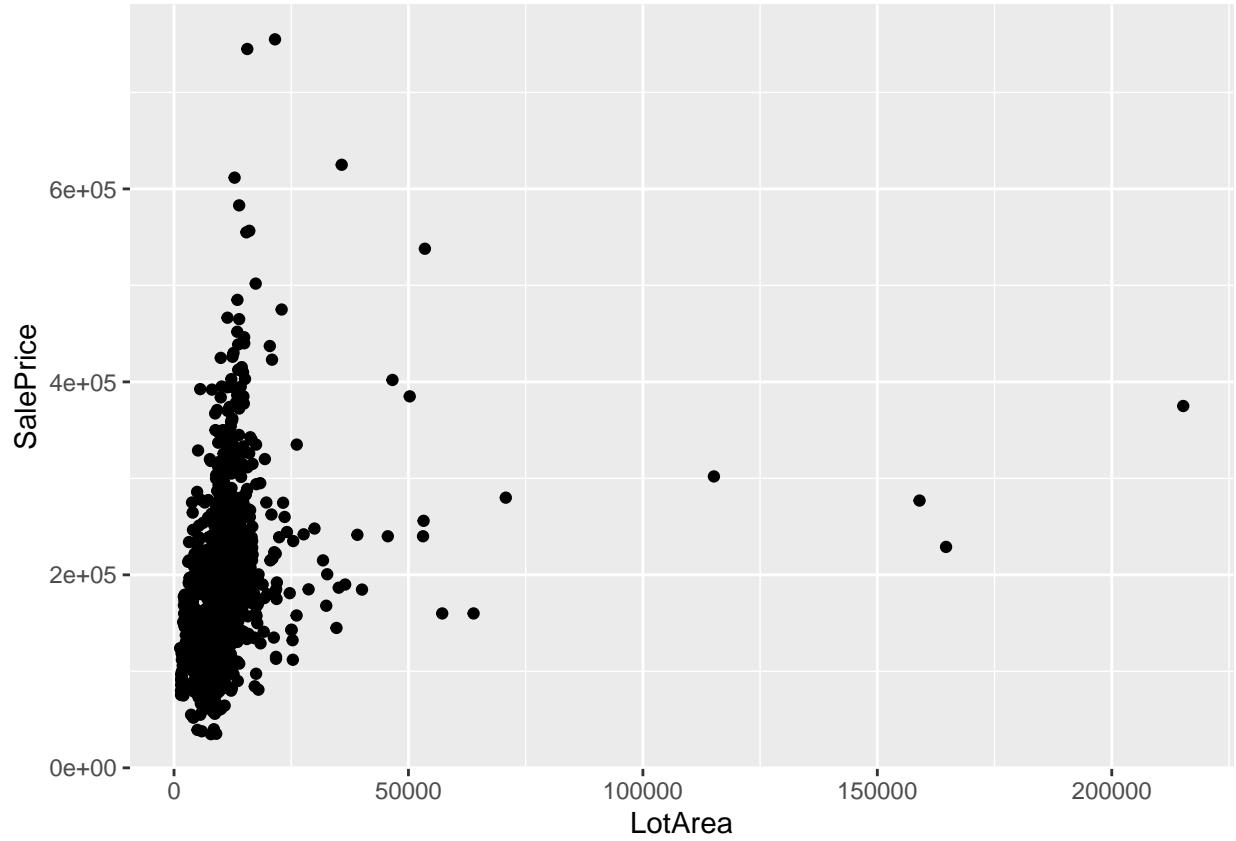


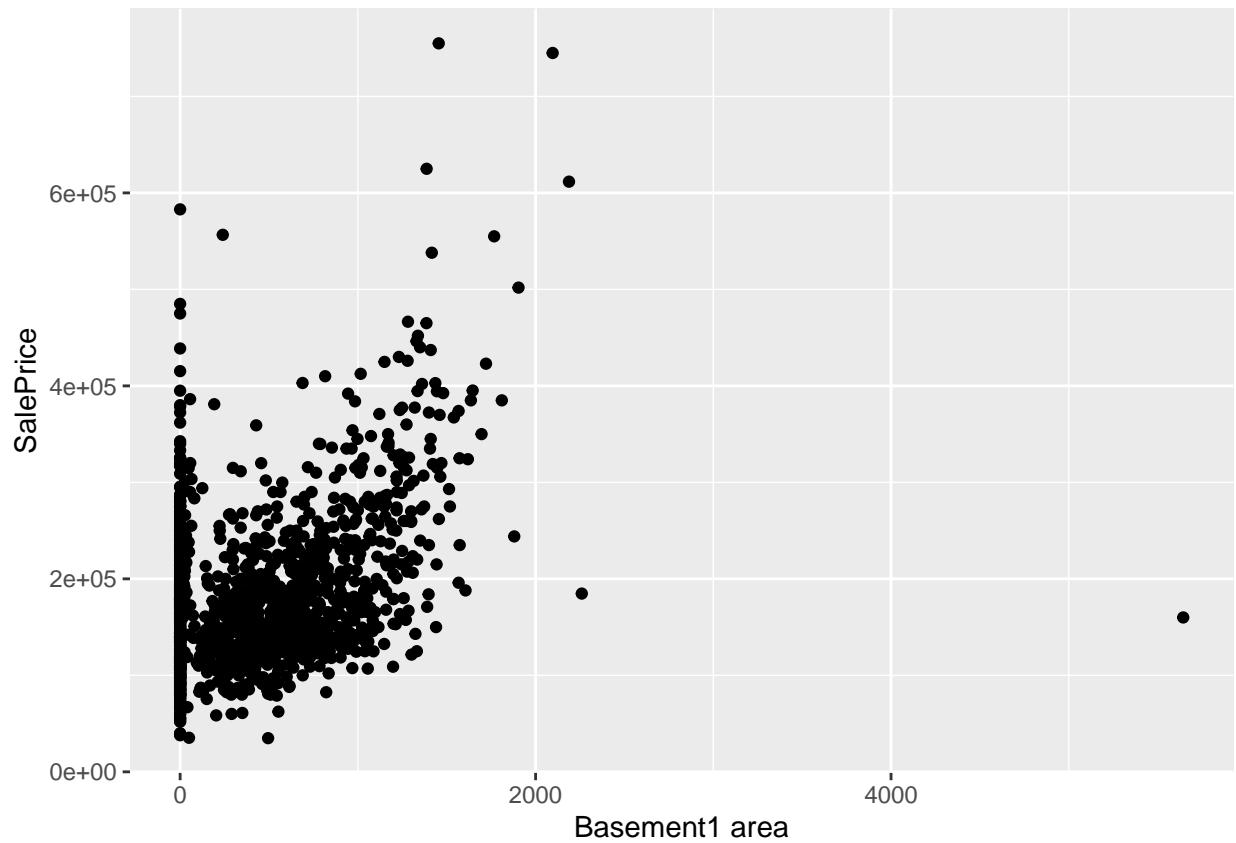


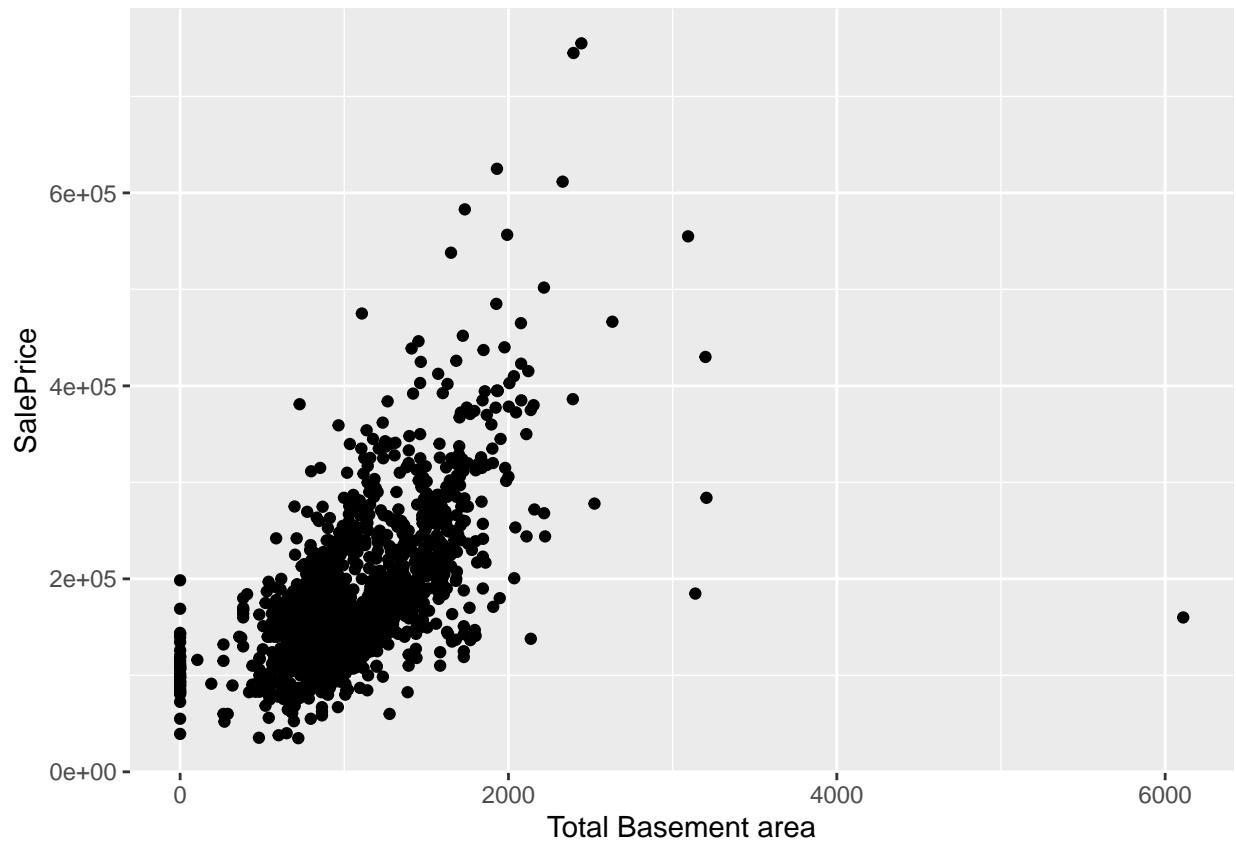


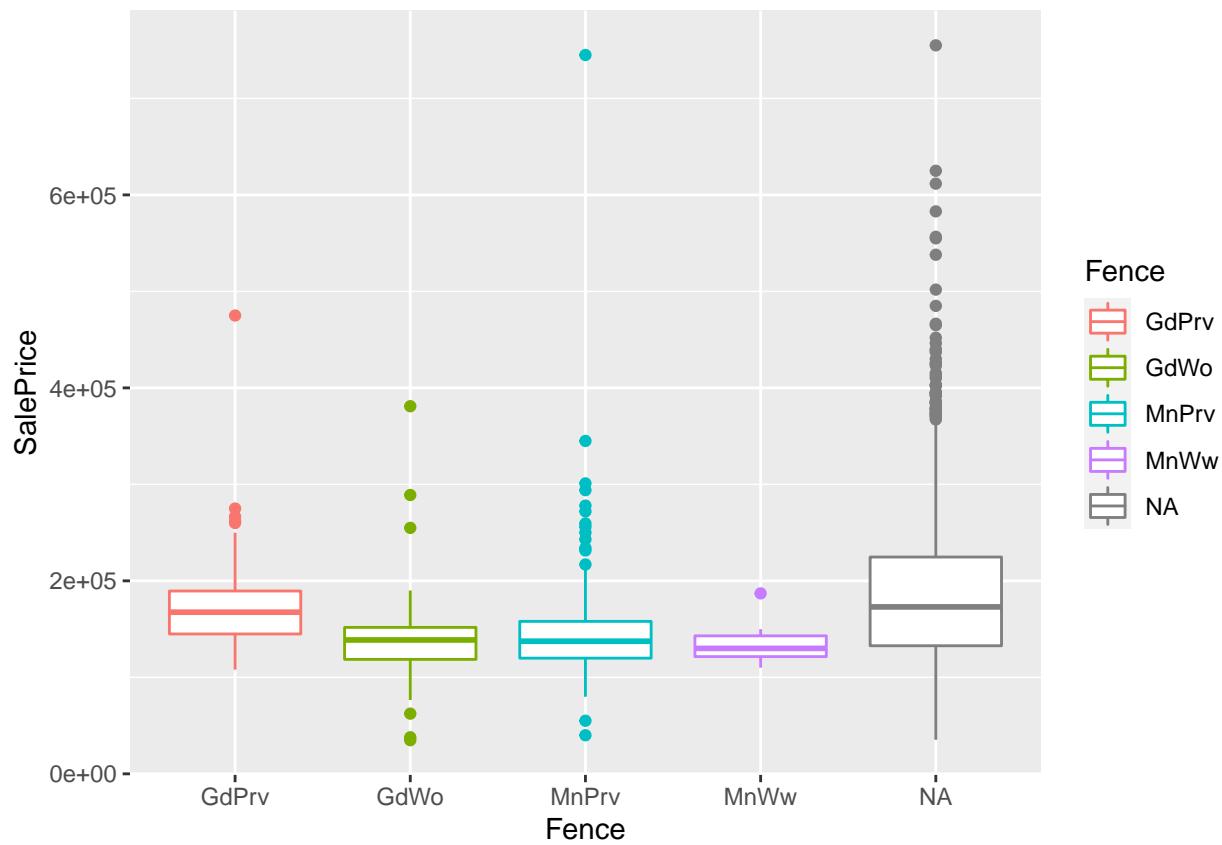


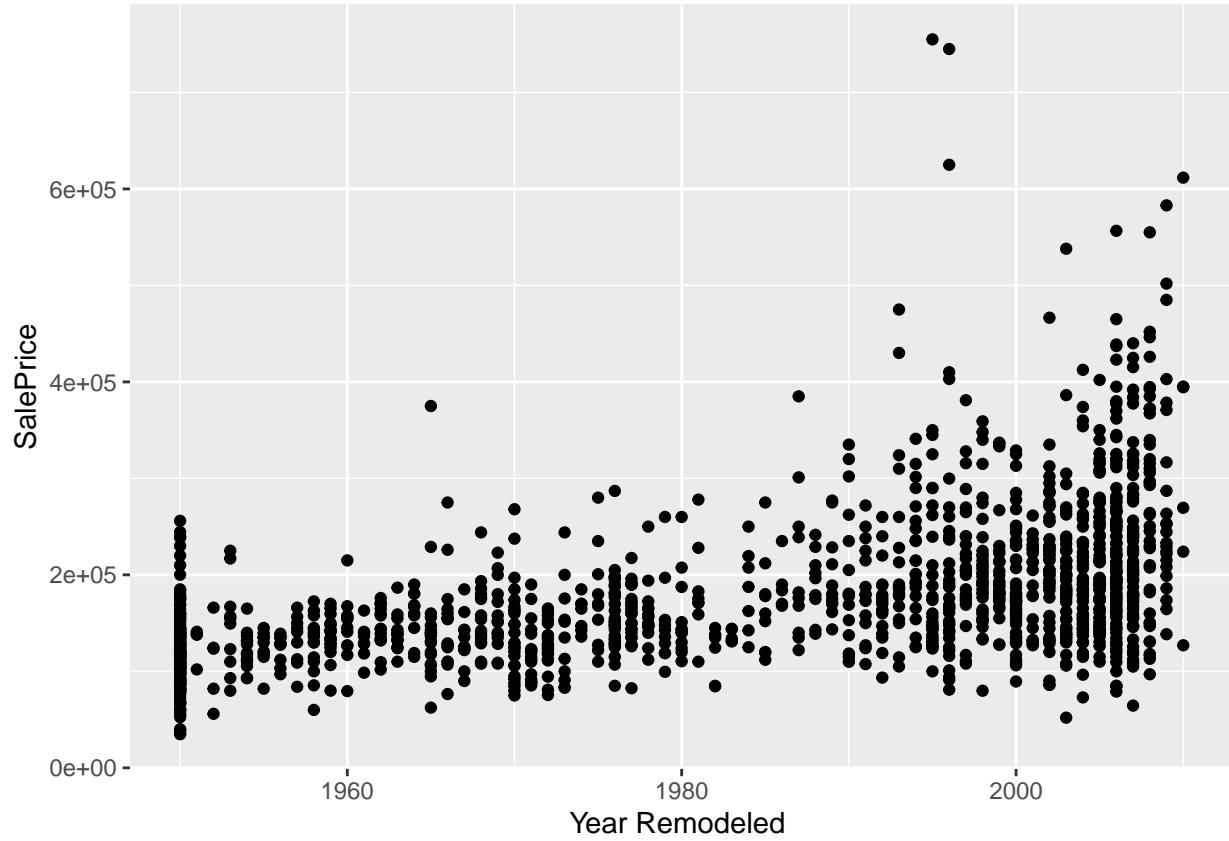


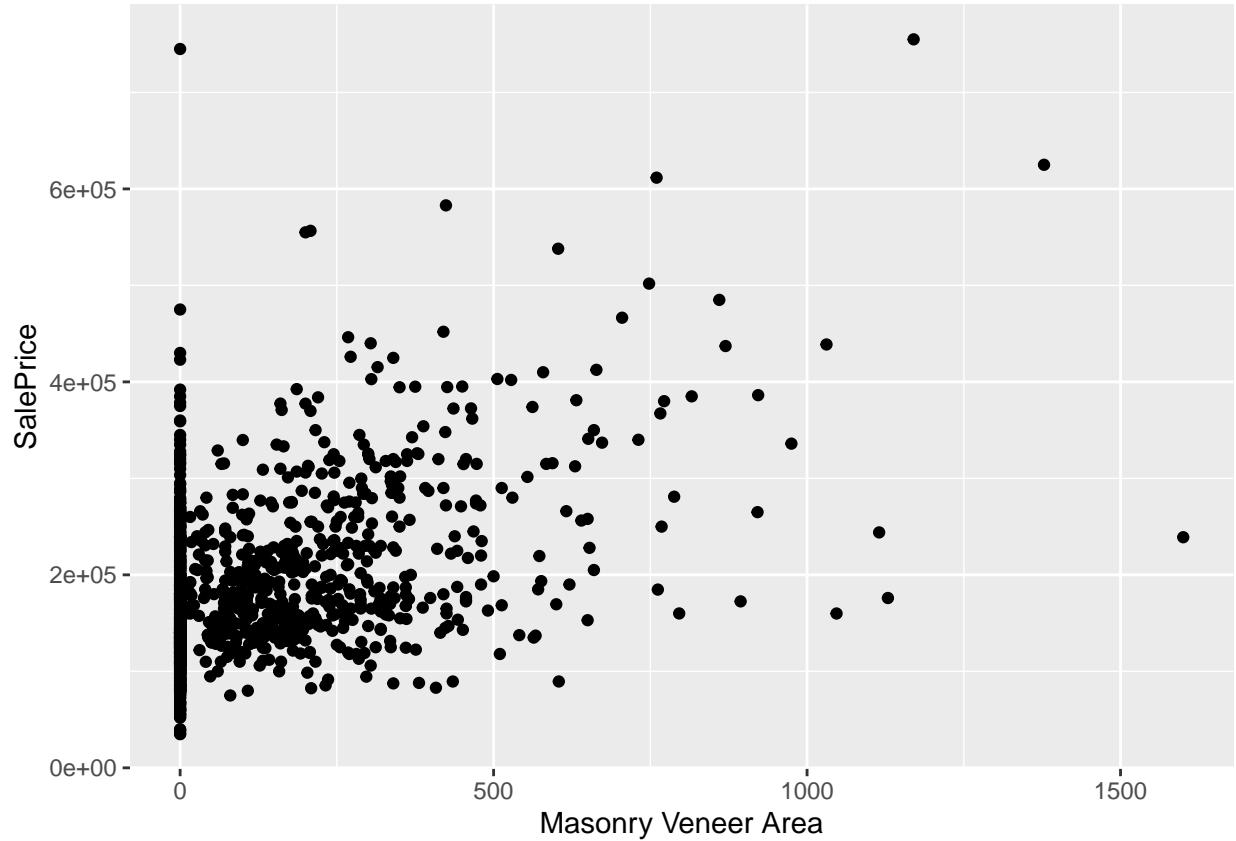


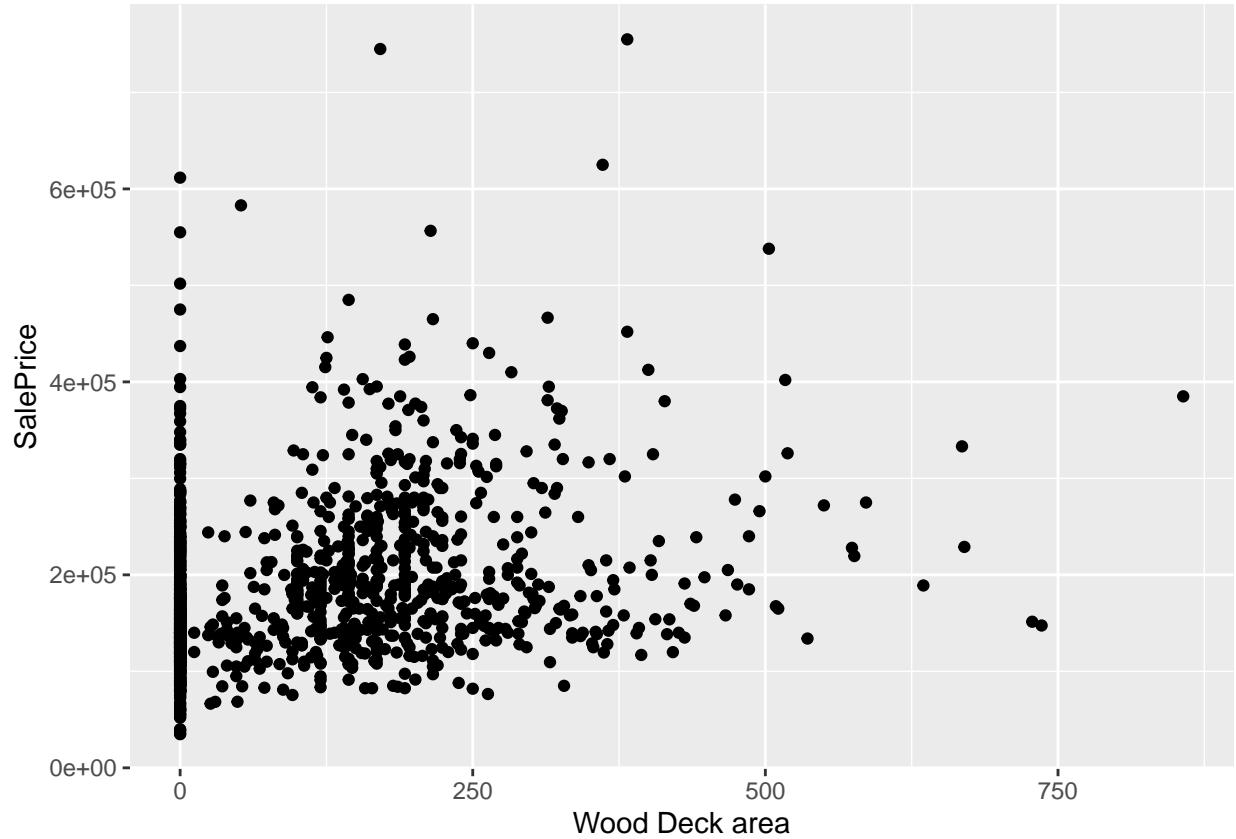


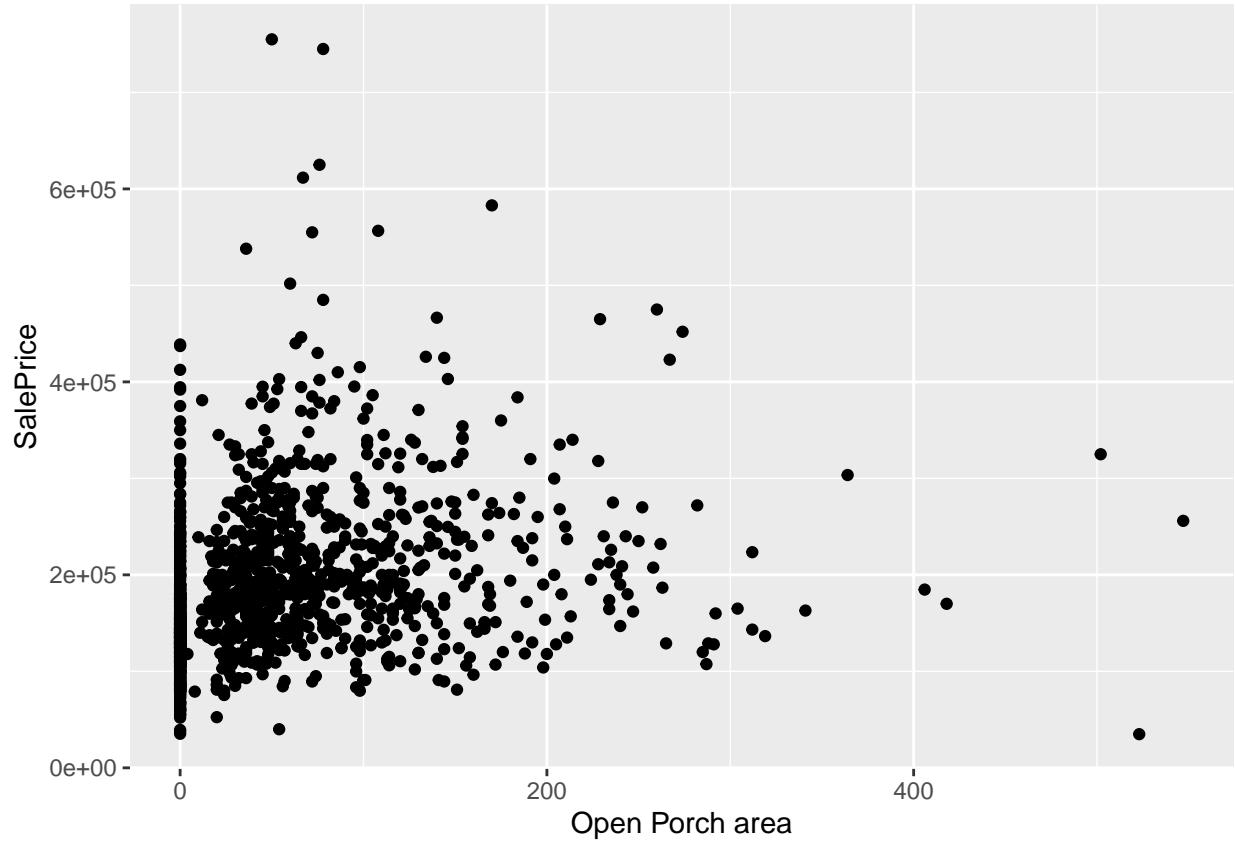


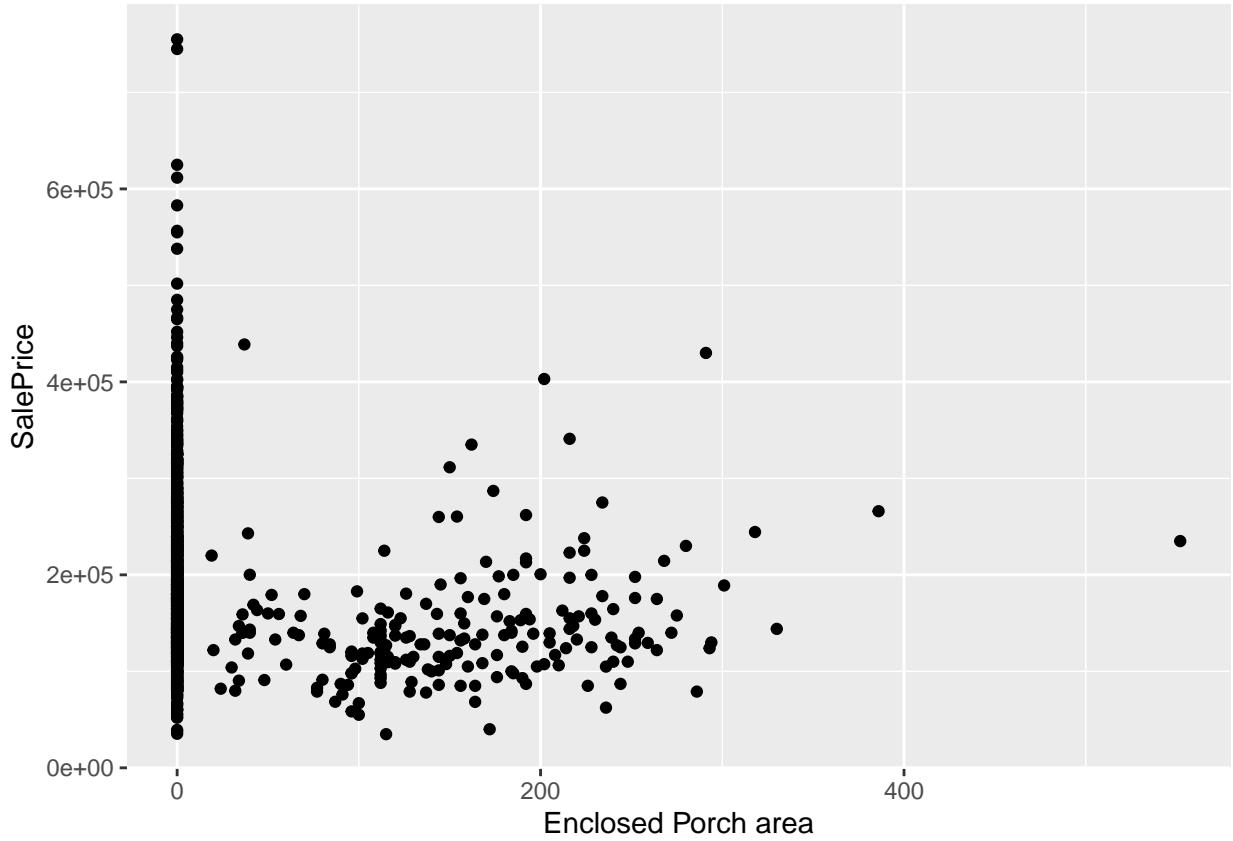












data cleanup / delete outliers and create data subset and summary() of the same subset

the variables selected in the current subset dataframe have been arrived at after multiple iterations of the model and fine tuning to add / drop variables from the subset to achieve better accuracy.

```
## 'data.frame': 1344 obs. of 39 variables:
## $ MSSubClass : Factor w/ 15 levels "20","30","40",...: 6 1 6 7 6 5 1 6 5 15 ...
## $ MSZoning   : Factor w/ 5 levels "C (all)","FV",...: 4 4 4 4 4 4 4 4 4 5 4 ...
## $ BldgType    : Factor w/ 5 levels "1Fam","2fmCon",...: 1 1 1 1 1 1 1 1 1 2 ...
## $ HouseStyle  : Factor w/ 8 levels "1.5Fin","1.5Unf",...: 6 3 6 6 6 1 3 6 1 2 ...
## $ LotConfig   : chr "Inside" "FR2" "Inside" "Corner" ...
## $ Neighborhood: chr "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
## $ Condition1  : chr "Norm" "Feedr" "Norm" "Norm" ...
## $ Condition2  : chr "Norm" "Norm" "Norm" "Norm" ...
## $ Foundation  : chr "PConc" "CBlock" "PConc" "BrkTil" ...
## $ RoofStyle   : chr "Gable" "Gable" "Gable" "Gable" ...
## $ RoofMatl   : chr "CompShg" "CompShg" "CompShg" "CompShg" ...
## $ Exterior1st : chr "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
## $ Exterior2nd : chr "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
## $ ExterQual   : chr "Gd" "TA" "Gd" "TA" ...
## $ HeatingQC   : chr "Ex" "Ex" "Ex" "Gd" ...
## $ Electrical  : chr "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
## $ LotFrontage : num 65 80 68 60 84 85 75 0 51 50 ...
## $ LotArea     : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
## $ OverallQual : Factor w/ 10 levels "1","2","3","4",...: 7 6 7 7 8 5 8 7 7 5 ...
## $ OverallCond : Factor w/ 9 levels "1","2","3","4",...: 5 8 5 5 5 5 5 6 5 6 ...
```

```

## $ YearBuilt      : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
## $ YearRemodAdd  : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
## $ MasVnrArea    : int  196 0 162 0 350 0 186 240 0 0 ...
## $ MasVnrType    : Factor w/ 5 levels "BrkCmn","BrkFace",...: 2 3 2 3 2 3 4 4 3 3 ...
## $ WoodDeckSF    : int  0 298 0 0 192 40 255 235 90 0 ...
## $ OpenPorchSF   : int  61 0 42 35 84 30 57 204 0 4 ...
## $ EnclosedPorch : int  0 0 0 272 0 0 0 228 205 0 ...
## $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
## $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
## $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
## $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
## $ 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 ...
## $ TotBaths     : num  2.5 2 2.5 1 2.5 1.5 2 2.5 2 1 ...
## $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
## $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
## $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
## $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
## $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

##      MSSubClass      MSZoning      BldgType      HouseStyle      LotConfig
## 20      :487      C (all): 9      1Fam :1111      1Story :669      Length:1344
## 60      :260      FV      : 64      2fmCon: 28      2Story :401      Class :character
## 50      :137      RH      : 16      Duplex: 50      1.5Fin :145      Mode  :character
## 120     : 85      RL      :1040      Twnhs : 43      SLvl   : 61
## 30      : 67      RM      : 215      TwnhsE: 112      SFoyer : 37
## 160     : 63                  1.5Unf : 14
## (Other):245                  (Other): 17
## Neighborhood      Condition1      Condition2      Foundation
## Length:1344      Length:1344      Length:1344      Length:1344
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##      RoofStyle      RoofMatl      Exterior1st      Exterior2nd
## Length:1344      Length:1344      Length:1344      Length:1344
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##      ExterQual      HeatingQC      Electrical      LotFrontage
## Length:1344      Length:1344      Length:1344      Min.   : 0.00
## Class :character      Class :character      Class :character      1st Qu.: 42.00
## Mode  :character      Mode  :character      Mode  :character      Median : 61.00
##                               Mean   : 56.19
##                               3rd Qu.: 77.00
##                               Max.   :182.00
##
##      LotArea      OverallQual      OverallCond      YearBuilt      YearRemodAdd
## Min.   : 1300      5       :380      5       :742      Min.   :1872      Min.   :1950

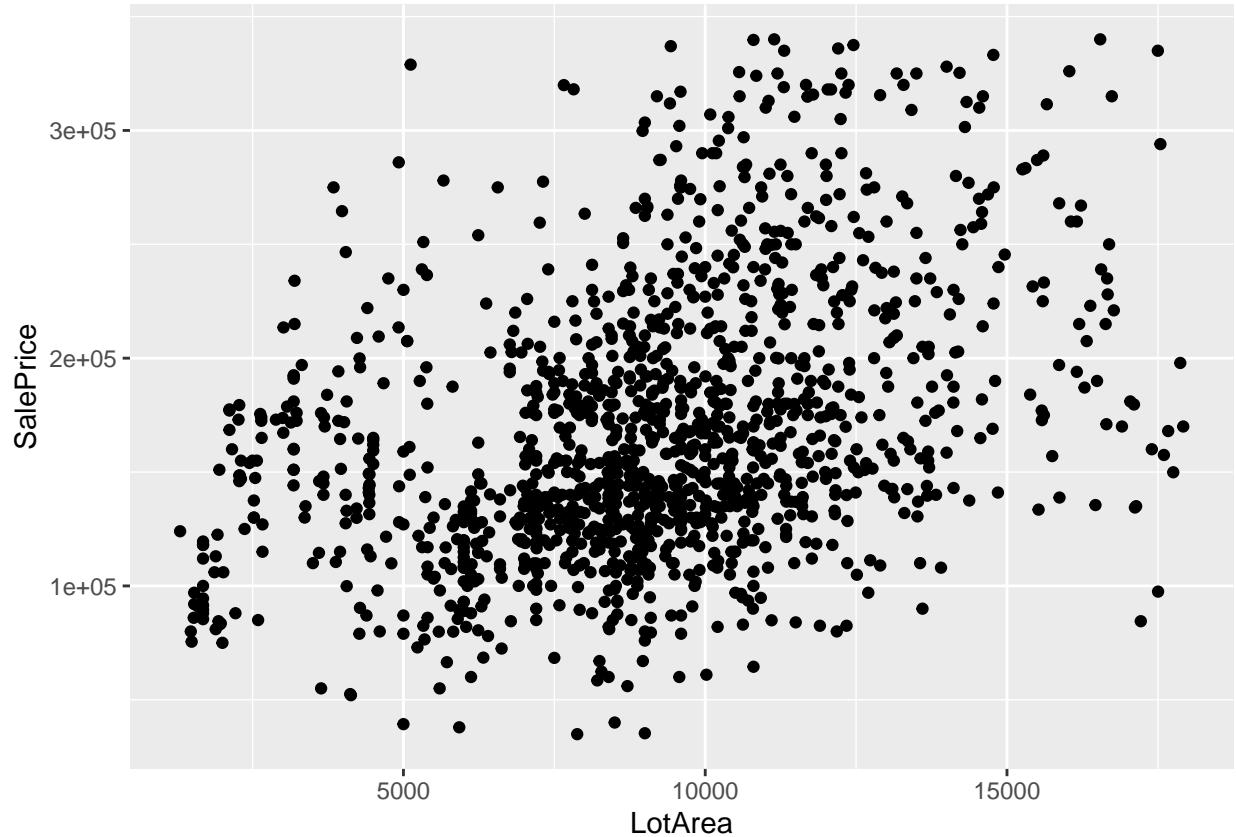
```

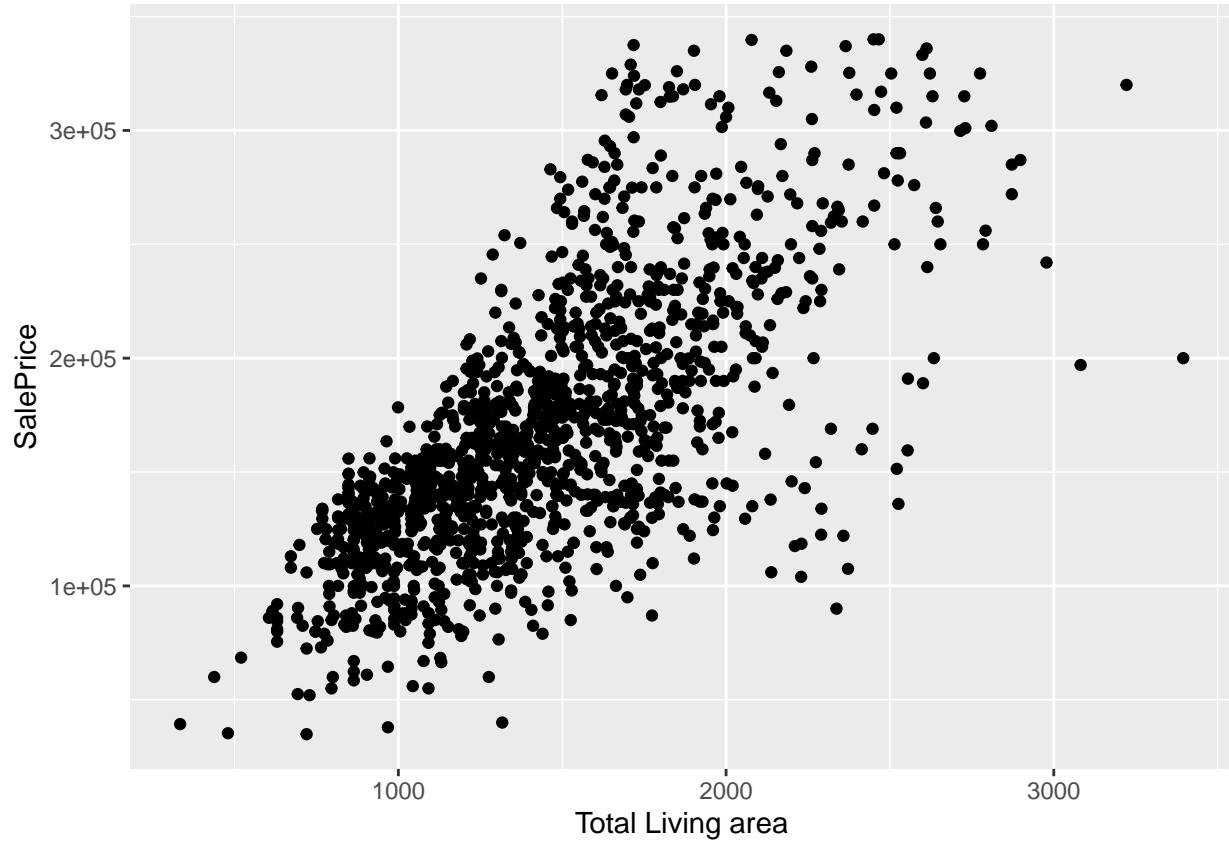
```

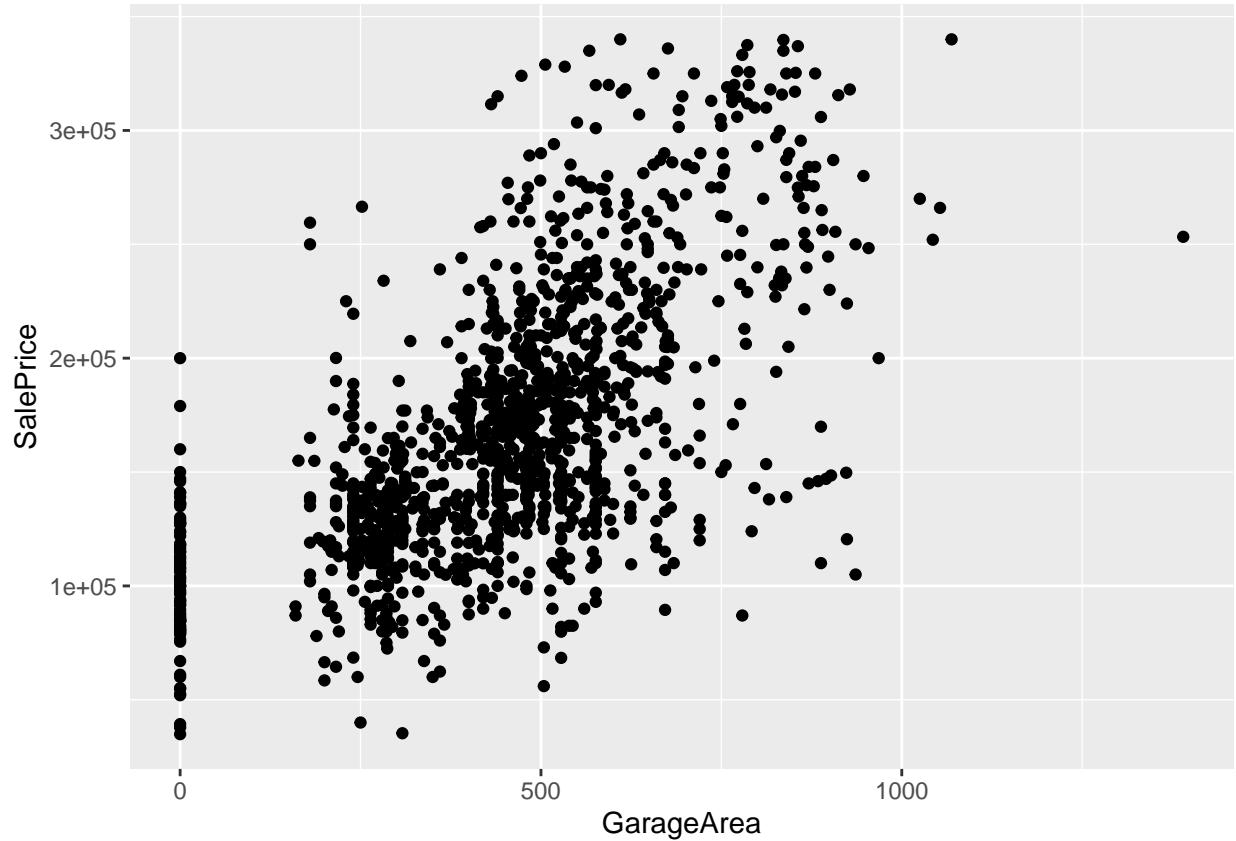
## 1st Qu.: 7310   6      :359   6      :240   1st Qu.:1953   1st Qu.:1966
## Median : 9164   7      :301   7      :194   Median :1972   Median :1992
## Mean   : 9148   8      :145   8      : 70   Mean   :1970   Mean   :1984
## 3rd Qu.:11042   4      :112   4      : 51   3rd Qu.:2000   3rd Qu.:2003
## Max.   :17920   9      : 20   3      : 24   Max.   :2009   Max.   :2010
## (Other): 27   (Other): 23
##      MasVnrArea      MasVnrType      WoodDeckSF      OpenPorchSF
## Min.   : 0.00   BrkCmn : 12   Min.   : 0.0   Min.   : 0.00
## 1st Qu.: 0.00   BrkFace:407   1st Qu.: 0.0   1st Qu.: 0.00
## Median : 0.00   None   :819   Median : 0.0   Median : 22.00
## Mean   : 89.34  Stone  :100   Mean   : 86.9   Mean   : 43.83
## 3rd Qu.:143.75 Unk    : 6    3rd Qu.:160.0 3rd Qu.: 64.00
## Max.   :1600.00
## NA's   :6
## EnclosedPorch      BsmtFinSF1      TotalBsmtSF      BsmtFinType1
## Min.   : 0.00   Min.   : 0.0   Min.   : 0.0   Length:1344
## 1st Qu.: 0.00   1st Qu.: 0.0   1st Qu.: 783.0  Class :character
## Median : 0.00   Median : 366.5  Median : 968.5  Mode  :character
## Mean   : 21.86  Mean   : 407.7  Mean   :1012.8 
## 3rd Qu.: 0.00   3rd Qu.: 679.2 3rd Qu.:1237.8 
## Max.   :386.00  Max.   :1880.0 Max.   :3206.0
##
##      BsmtQual      GrLivArea      FullBath      HalfBath
## Length:1344      Min.   :334   Min.   :0.000   Min.   :0.0000
## Class :character  1st Qu.:1113  1st Qu.:1.000  1st Qu.:0.0000
## Mode  :character  Median :1426   Median :2.000   Median :0.0000
##                  Mean   :1454   Mean   :1.532   Mean   :0.3705
##                  3rd Qu.:1718   3rd Qu.:2.000  3rd Qu.:1.0000
##                  Max.   :3395   Max.   :3.000   Max.   :2.0000
##
##      TotBaths      BedroomAbvGr      TotRmsAbvGrd      GarageArea
## Min.   :0.000   Min.   :0.000   Min.   : 2.000   Min.   : 0.0
## 1st Qu.:1.000   1st Qu.:2.000   1st Qu.: 5.000   1st Qu.: 308.0
## Median :2.000   Median :3.000   Median : 6.000   Median : 464.5
## Mean   :1.717   Mean   :2.854   Mean   : 6.381   Mean   : 454.0
## 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 7.000   3rd Qu.: 572.0
## Max.   :3.500   Max.   :8.000   Max.   :14.000   Max.   :1390.0
##
##      SalePrice
## Min.   : 34900
## 1st Qu.:128000
## Median :157700
## Mean   :168855
## 3rd Qu.:200625
## Max.   :340000
##

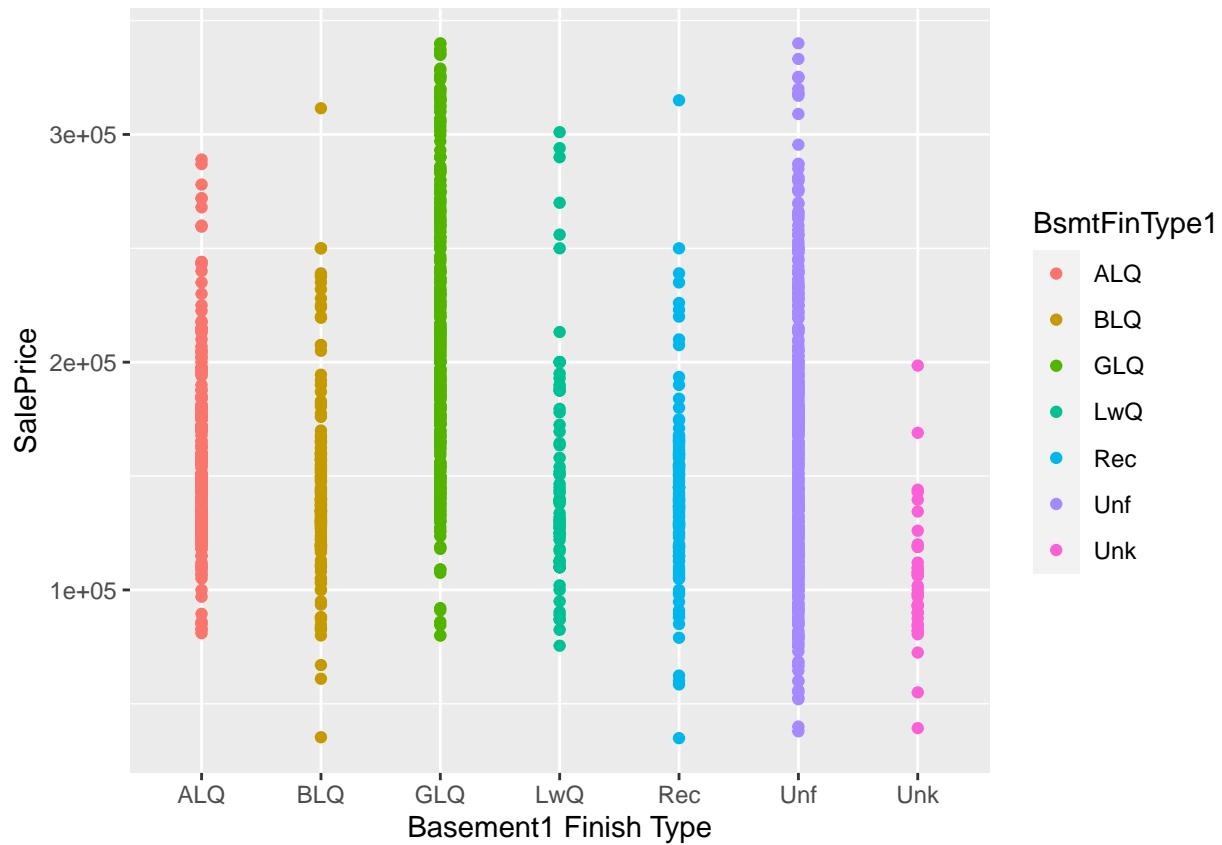
```

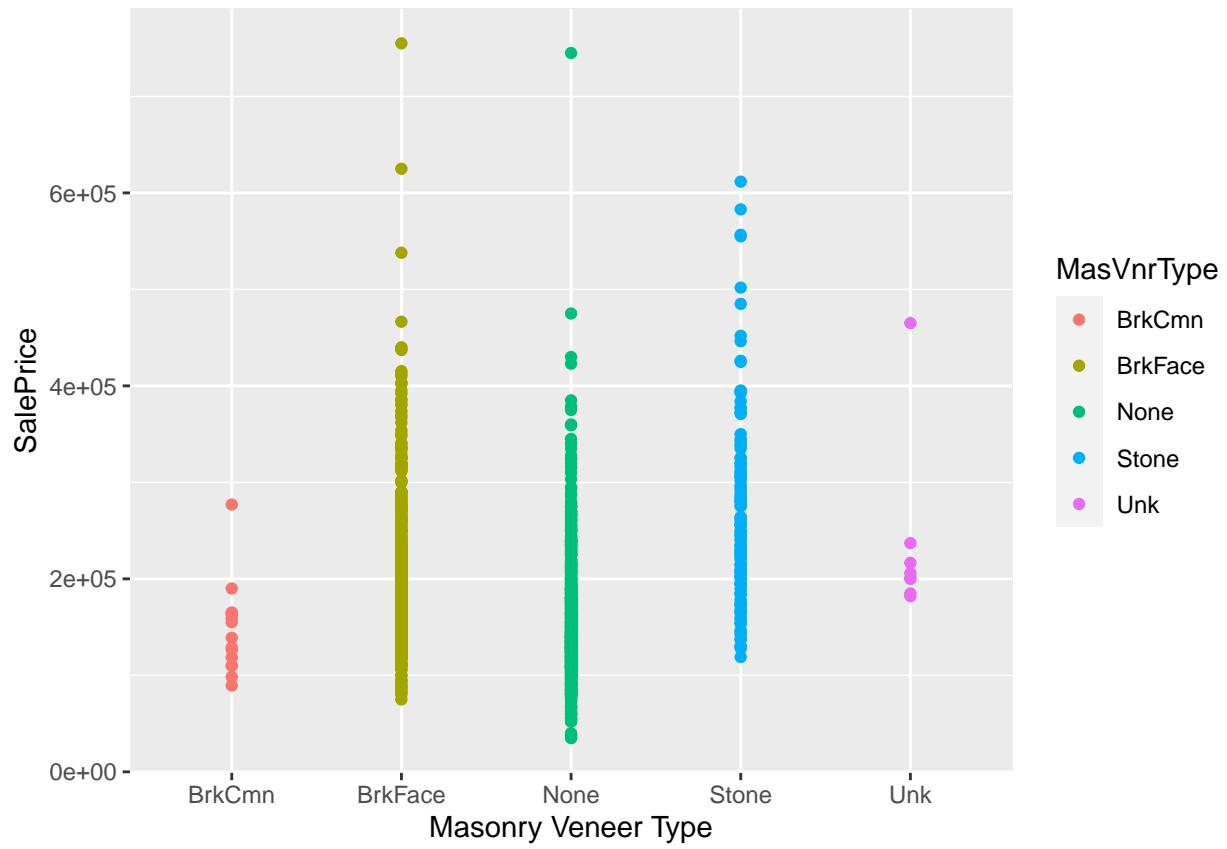
Scatter Plots for relationship between Sale Price and some of the predictor variables

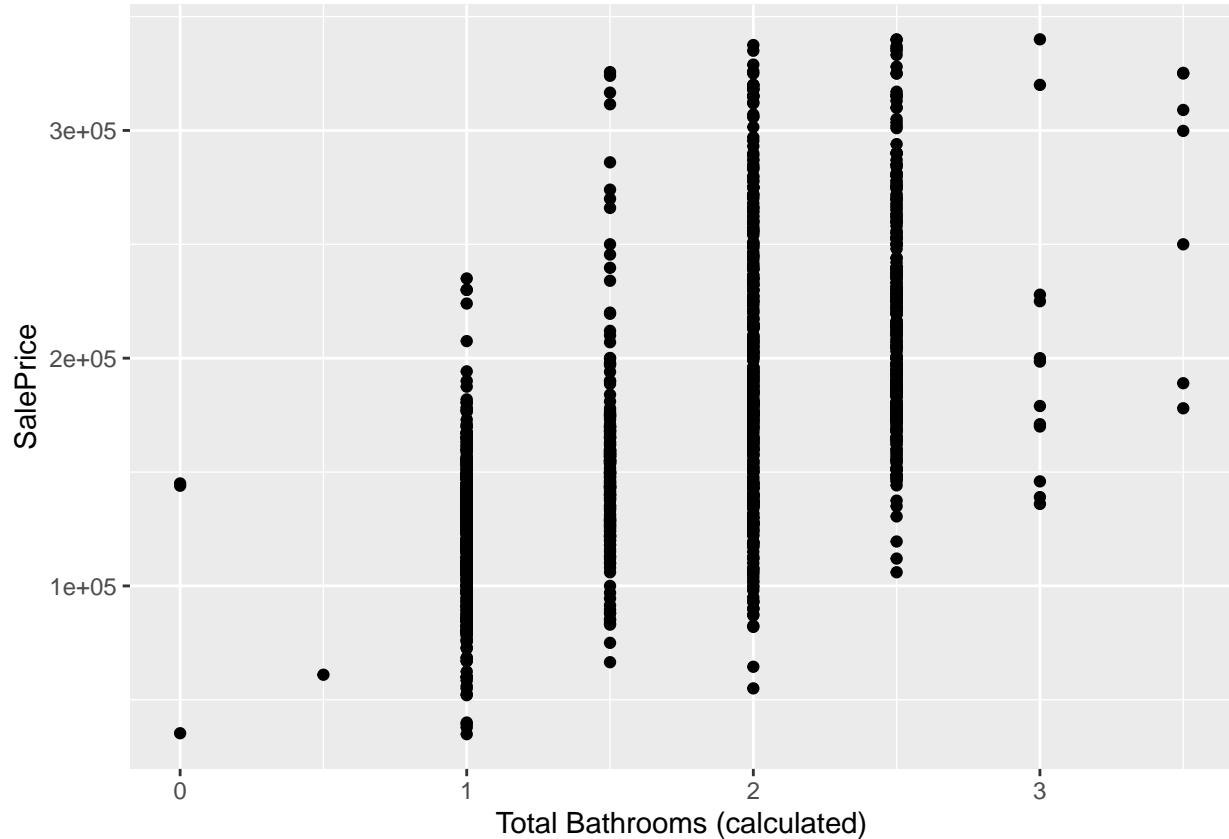


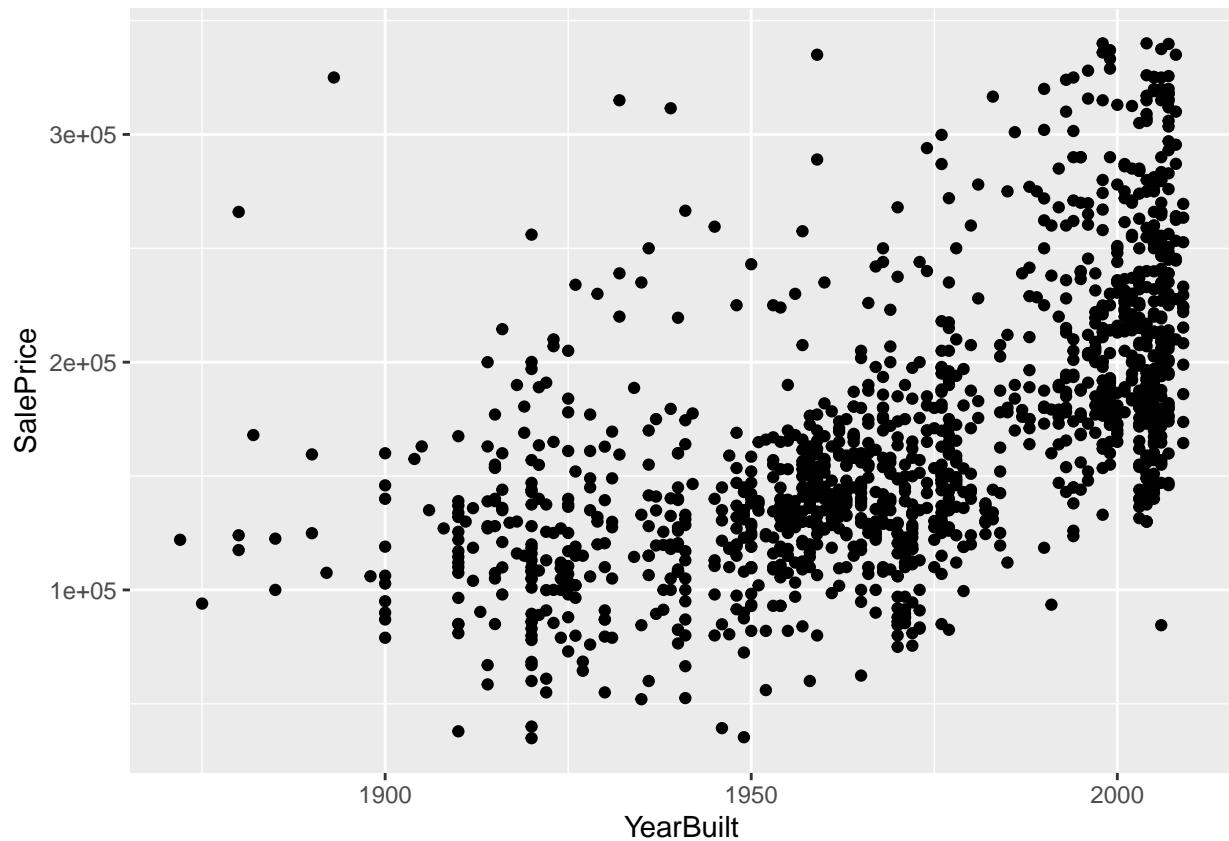


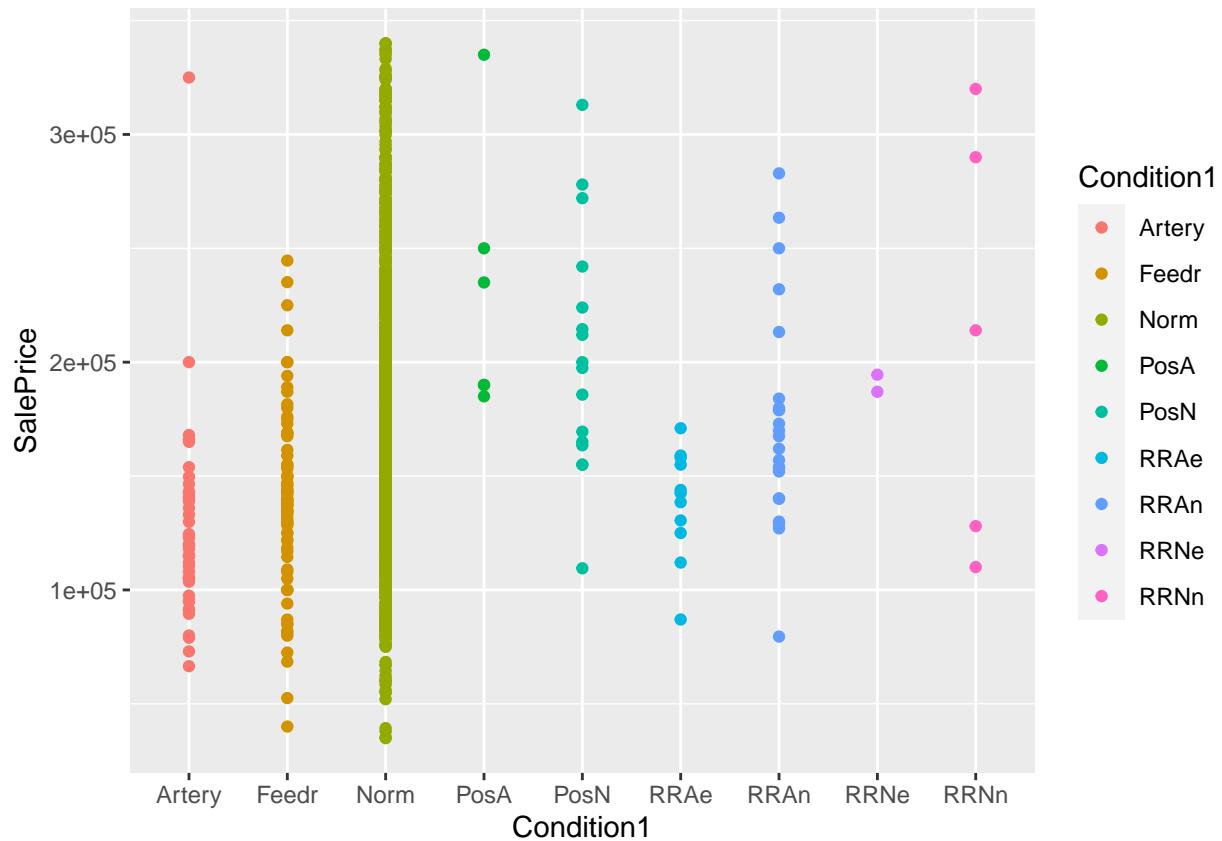


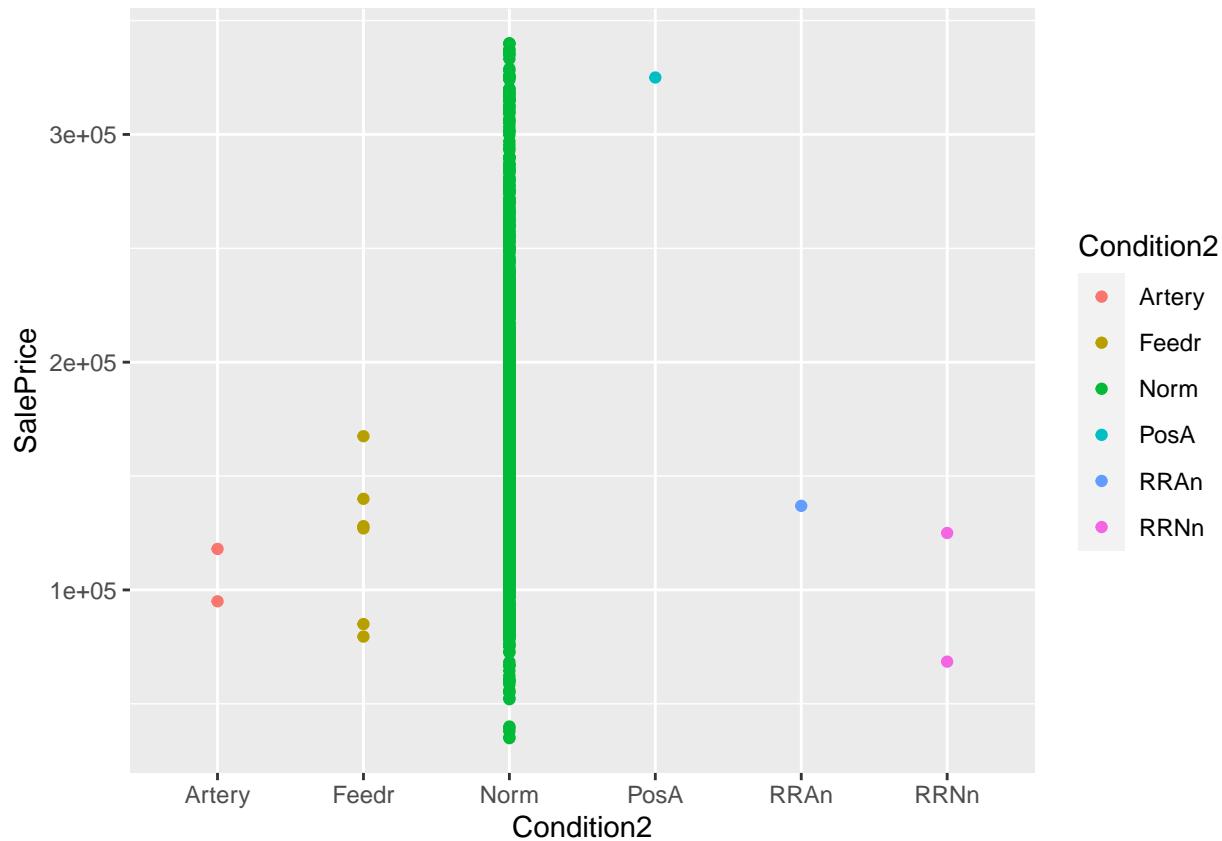


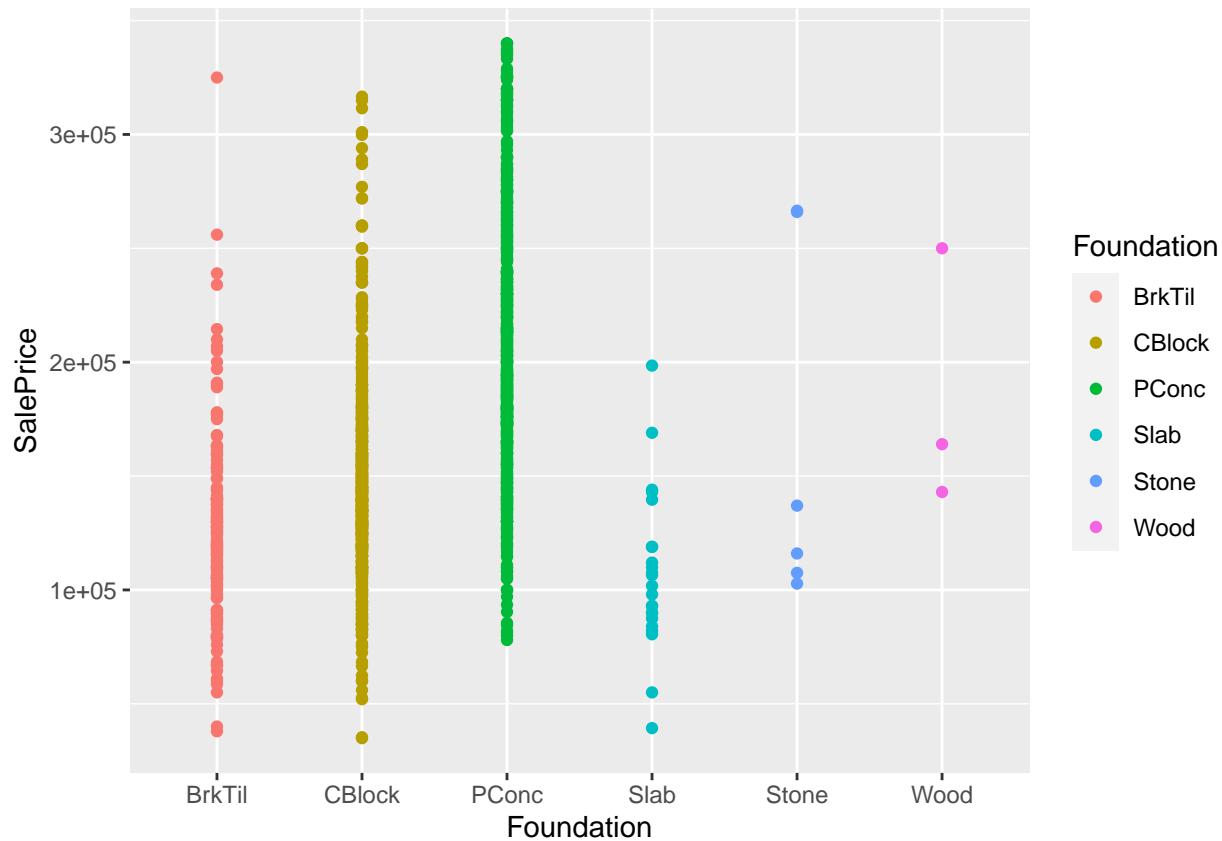












`head()` from the modified subset data frame

```
##   MSSubClass MSZoning BldgType HouseStyle LotConfig Neighborhood Condition1
## 1       60      RL     1Fam    2Story    Inside    CollgCr     Norm
## 2       20      RL     1Fam    1Story     FR2     Veenker     Feedr
## 3       60      RL     1Fam    2Story    Inside    CollgCr     Norm
## 4       70      RL     1Fam    2Story    Corner    Crawfor     Norm
## 5       60      RL     1Fam    2Story     FR2    NoRidge     Norm
## 6       50      RL     1Fam   1.5Fin    Inside    Mitchel     Norm
##   Condition2 Foundation RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual
## 1     Norm      PConc     Gable  CompShg   VinylSd   VinylSd      Gd
## 2     Norm      CBlock     Gable  CompShg   MetalSd   MetalSd      TA
## 3     Norm      PConc     Gable  CompShg   VinylSd   VinylSd      Gd
## 4     Norm      BrkTil     Gable  CompShg   Wd Sdng   Wd Shng      TA
## 5     Norm      PConc     Gable  CompShg   VinylSd   VinylSd      Gd
## 6     Norm       Wood     Gable  CompShg   VinylSd   VinylSd      TA
##   HeatingQC Electrical LotFrontage LotArea OverallQual OverallCond YearBuilt
## 1       Ex        SBrkr        65   8450         7          5     2003
## 2       Ex        SBrkr        80   9600         6          8     1976
## 3       Ex        SBrkr        68  11250         7          5     2001
## 4       Gd        SBrkr        60   9550         7          5     1915
## 5       Ex        SBrkr        84  14260         8          5     2000
## 6       Ex        SBrkr        85  14115         5          5     1993
##   YearRemodAdd MasVnrArea MasVnrType WoodDeckSF OpenPorchSF EnclosedPorch
## 1       2003        196    BrkFace          0         61          0
## 2       1976         0     None        298          0          0
```

```

## 3      2002      162    BrkFace       0      42       0
## 4      1970       0    None        0      35     272
## 5      2000      350    BrkFace      192      84       0
## 6      1995       0    None        40      30       0
##   BsmtFinSF1 TotalBsmtSF BsmtFinType1 BsmtQual GrLivArea FullBath HalfBath
## 1      706      856      GLQ       Gd     1710       2       1
## 2      978     1262      ALQ       Gd     1262       2       0
## 3      486      920      GLQ       Gd     1786       2       1
## 4      216      756      ALQ       TA     1717       1       0
## 5      655     1145      GLQ       Gd     2198       2       1
## 6      732      796      GLQ       Gd     1362       1       1
##   TotBaths BedroomAbvGr TotRmsAbvGrd GarageArea SalePrice
## 1      2.5          3            8      548  208500
## 2      2.0          3            6      460  181500
## 3      2.5          3            6      608  223500
## 4      1.0          3            7      642  140000
## 5      2.5          4            9      836  250000
## 6      1.5          1            5      480  143000

```

**correlation - R2 for various numeric variables with Sale Price**

```

cor(housing_df$SalePrice, housing_df$LotArea)^2

## [1] 0.1528357

cor(housing_df$SalePrice, housing_df$YearBuilt)^2

## [1] 0.3497324

cor(housing_df$SalePrice, housing_df$YearRemodAdd)^2

## [1] 0.3085205

cor(housing_df$SalePrice, housing_df$GrLivArea)^2

## [1] 0.4813659

cor(housing_df$SalePrice, (housing_df$FullBath + (housing_df$HalfBath * 0.5)))^2

## [1] 0.390098

cor(housing_df$SalePrice, housing_df$BedroomAbvGr)^2

## [1] 0.0397595

cor(housing_df$SalePrice, housing_df$TotRmsAbvGrd)^2

## [1] 0.2298562

```

```

cor(housing_df$SalePrice, housing_df$GarageArea)^2

## [1] 0.4066953

cor(housing_df$SalePrice, housing_df$BsmtFinSF1)^2

## [1] 0.09712678

cor(housing_df$SalePrice, housing_df$TotalBsmtSF)^2

## [1] 0.3482836

cor(housing_df$SalePrice, housing_df$WoodDeckSF)^2

## [1] 0.08447678

cor(housing_df$SalePrice, housing_df$OpenPorchSF)^2

## [1] 0.1118559

cor(housing_df$SalePrice, housing_df$EnclosedPorch)^2

## [1] 0.02645305

```

housing price prediction - multiple regression model creation and summary() of the models

```

##
## Call:
## lm(formula = SalePrice ~ Neighborhood + RoofStyle + LotArea +
##     LotFrontage + OverallQual + OverallCond + YearBuilt + OpenPorchSF +
##     GrLivArea + TotRmsAbvGrd + GarageArea + TotalBsmtSF, data = housing_df)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -123224   -9749     443    10078    73042
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.072e+06  8.912e+04 -12.032 < 2e-16 ***
## NeighborhoodBlueste -1.235e+04  1.449e+04  -0.852 0.394427
## NeighborhoodBrDale -1.582e+04  6.970e+03  -2.270 0.023365 *
## NeighborhoodBrkSide  1.272e+03  6.172e+03   0.206 0.836806
## NeighborhoodClearCr 5.621e+03  7.312e+03   0.769 0.442220
## NeighborhoodCollgCr -4.745e+03  5.156e+03  -0.920 0.357529
## NeighborhoodCrawfor  2.259e+04  6.254e+03   3.611 0.000316 ***
## NeighborhoodEdwards -1.115e+04  5.704e+03  -1.955 0.050794 .
## NeighborhoodGilbert -5.734e+03  5.536e+03  -1.036 0.300517
## NeighborhoodIDOTRR  -1.140e+04  6.572e+03  -1.734 0.083185 .

```

```

## NeighborhoodMeadowV -1.856e+04 6.980e+03 -2.659 0.007940 **
## NeighborhoodMitchel -1.229e+04 5.861e+03 -2.097 0.036167 *
## NeighborhoodNAmes -9.416e+03 5.443e+03 -1.730 0.083889 .
## NeighborhoodNoRidge 1.595e+04 6.289e+03 2.537 0.011311 *
## NeighborhoodNPkVill -4.605e+03 8.139e+03 -0.566 0.571647
## NeighborhoodNridgHt 1.215e+04 5.628e+03 2.158 0.031096 *
## NeighborhoodNWAmes -1.484e+04 5.669e+03 -2.618 0.008951 **
## NeighborhoodOldTown -1.371e+04 6.042e+03 -2.270 0.023372 *
## NeighborhoodSawyer -1.154e+04 5.771e+03 -1.999 0.045841 *
## NeighborhoodSawyerW -4.365e+03 5.599e+03 -0.780 0.435785
## NeighborhoodSomerst 4.243e+03 5.291e+03 0.802 0.422704
## NeighborhoodStoneBr 1.046e+04 6.893e+03 1.518 0.129364
## NeighborhoodSWISU -5.618e+03 6.946e+03 -0.809 0.418826
## NeighborhoodTimber -4.514e+02 6.132e+03 -0.074 0.941340
## NeighborhoodVeenker 1.334e+04 7.994e+03 1.668 0.095475 .
## RoofStyleGable -5.562e+03 8.816e+03 -0.631 0.528203
## RoofStyleGambrel 4.837e+02 1.058e+04 0.046 0.963526
## RoofStyleHip -3.948e+03 8.855e+03 -0.446 0.655816
## RoofStyleMansard 4.372e+03 1.153e+04 0.379 0.704539
## LotArea 2.128e+00 2.308e-01 9.221 < 2e-16 ***
## LotFrontage -2.275e+01 1.799e+01 -1.264 0.206324
## OverallQual2 -8.049e+03 2.225e+04 -0.362 0.717556
## OverallQual3 -6.559e+03 2.022e+04 -0.324 0.745743
## OverallQual4 -8.637e+03 1.991e+04 -0.434 0.664549
## OverallQual5 -5.874e+03 1.992e+04 -0.295 0.768082
## OverallQual6 -4.223e+02 1.994e+04 -0.021 0.983105
## OverallQual7 1.377e+04 2.001e+04 0.688 0.491516
## OverallQual8 3.650e+04 2.016e+04 1.811 0.070439 .
## OverallQual9 7.290e+04 2.064e+04 3.532 0.000427 ***
## OverallQual10 9.259e+04 2.310e+04 4.008 6.47e-05 ***
## OverallCond2 -3.949e+03 2.947e+04 -0.134 0.893437
## OverallCond3 4.236e+03 2.732e+04 0.155 0.876825
## OverallCond4 1.496e+04 2.785e+04 0.537 0.591206
## OverallCond5 2.392e+04 2.779e+04 0.861 0.389429
## OverallCond6 3.202e+04 2.779e+04 1.152 0.249393
## OverallCond7 4.267e+04 2.779e+04 1.535 0.124999
## OverallCond8 4.803e+04 2.786e+04 1.724 0.084957 .
## OverallCond9 5.513e+04 2.815e+04 1.958 0.050394 .
## YearBuilt 5.602e+02 4.410e+01 12.701 < 2e-16 ***
## OpenPorchSF 2.742e+01 9.050e+00 3.030 0.002497 **
## GrLivArea 5.162e+01 2.577e+00 20.031 < 2e-16 ***
## TotRmsAbvGrd -2.592e+03 6.528e+02 -3.970 7.57e-05 ***
## GarageArea 2.563e+01 3.548e+00 7.223 8.66e-13 ***
## TotalBsmtSF 2.305e+01 1.813e+00 12.712 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 19080 on 1290 degrees of freedom
## Multiple R-squared: 0.8993, Adjusted R-squared: 0.8952
## F-statistic: 217.4 on 53 and 1290 DF, p-value: < 2.2e-16

```

looking at model statistics iteratively, I refined the model a little bit from the beginning to include more relevant variables with higher impact on the prices and omit the ones having higher p values. Logistic regression is mainly useful for classification rather than predicting numeric

values which can take any number of values, like house prices based on various factors. So, using only multiple linear regression model.

### Finalized model with little better predictive power

```
##
## Call:
## lm(formula = SalePrice ~ Neighborhood + RoofStyle + LotArea +
##     LotFrontage + OverallQual + OverallCond + YearBuilt + OpenPorchSF +
##     GrLivArea + TotRmsAbvGrd + GarageArea + TotalBsmtSF + Exterior1st +
##     ExterQual + YearRemodAdd + MasVnrArea + MasVnrType + WoodDeckSF +
##     BsmtFinType1 + MSSubClass + MSZoning + HouseStyle + LotConfig +
##     Foundation + Condition1 + Condition2, data = housing_df)
##
## Residuals:
##      Min    1Q Median    3Q   Max 
## -108352 -8740    470   8553  68971 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.167e+06 1.202e+05 -9.709 < 2e-16 ***
## NeighborhoodBlueste 1.106e+04 1.381e+04  0.800 0.423615  
## NeighborhoodBrDale  7.856e+03 8.074e+03  0.973 0.330751  
## NeighborhoodBrkSide 1.068e+03 6.743e+03  0.158 0.874225  
## NeighborhoodClearCr -4.503e+03 6.950e+03 -0.648 0.517173  
## NeighborhoodCollgCr -1.226e+04 5.206e+03 -2.355 0.018699 *  
## NeighborhoodCrawfor  1.781e+04 6.234e+03  2.858 0.004341 ** 
## NeighborhoodEdwards -1.502e+04 5.784e+03 -2.598 0.009497 ** 
## NeighborhoodGilbert -1.190e+04 5.631e+03 -2.112 0.034849 *  
## NeighborhoodIDOTRR -5.261e+03 7.628e+03 -0.690 0.490469  
## NeighborhoodMeadowV -1.740e+04 8.783e+03 -1.981 0.047775 *  
## NeighborhoodMitchel -1.360e+04 5.916e+03 -2.299 0.021682 *  
## NeighborhoodNAmes -1.359e+04 5.591e+03 -2.431 0.015213 *  
## NeighborhoodNoRidge  4.181e+03 6.289e+03  0.665 0.506282  
## NeighborhoodNPkVill  1.108e+04 8.017e+03  1.382 0.167082  
## NeighborhoodNridgHt  6.412e+03 5.472e+03  1.172 0.241542  
## NeighborhoodNWAmes -1.715e+04 5.787e+03 -2.964 0.003098 ** 
## NeighborhoodOldTown -1.173e+04 6.864e+03 -1.709 0.087626 .  
## NeighborhoodSawyer -1.125e+04 5.845e+03 -1.924 0.054588 .  
## NeighborhoodSawyerW -6.873e+03 5.615e+03 -1.224 0.221181  
## NeighborhoodSomerset -3.555e+03 6.703e+03 -0.530 0.596023  
## NeighborhoodStoneBr  1.291e+04 6.460e+03  1.998 0.045978 *  
## NeighborhoodSWISU -7.453e+03 6.941e+03 -1.074 0.283126  
## NeighborhoodTimber -5.062e+03 6.013e+03 -0.842 0.400020  
## NeighborhoodVeenker 1.364e+04 7.645e+03  1.784 0.074669 .  
## RoofStyleGable -1.192e+04 8.258e+03 -1.443 0.149287  
## RoofStyleGambrel -3.474e+03 9.911e+03 -0.350 0.726034  
## RoofStyleHip -1.085e+04 8.302e+03 -1.307 0.191299  
## RoofStyleMansard -4.423e+03 1.073e+04 -0.412 0.680311  
## LotArea 1.676e+00 2.395e-01  6.997 4.31e-12 ***
## LotFrontage -6.596e+00 1.747e+01 -0.378 0.705813  
## OverallQual2 1.270e+04 2.103e+04  0.604 0.546011  
## OverallQual3 5.332e+03 1.928e+04  0.277 0.782200  
## OverallQual4 6.855e+03 1.902e+04  0.360 0.718633
```

## OverallQual5	8.354e+03	1.912e+04	0.437	0.662219
## OverallQual6	1.201e+04	1.916e+04	0.627	0.530830
## OverallQual7	2.240e+04	1.922e+04	1.166	0.243958
## OverallQual8	4.016e+04	1.935e+04	2.076	0.038122 *
## OverallQual9	7.250e+04	1.983e+04	3.657	0.000267 ***
## OverallQual10	7.578e+04	2.331e+04	3.251	0.001181 **
## OverallCond2	-1.044e+04	2.669e+04	-0.391	0.695834
## OverallCond3	-5.511e+03	2.486e+04	-0.222	0.824573
## OverallCond4	3.609e+03	2.527e+04	0.143	0.886451
## OverallCond5	1.203e+04	2.527e+04	0.476	0.634153
## OverallCond6	1.818e+04	2.528e+04	0.719	0.472299
## OverallCond7	2.496e+04	2.530e+04	0.986	0.324121
## OverallCond8	2.807e+04	2.538e+04	1.106	0.269006
## OverallCond9	3.380e+04	2.572e+04	1.314	0.189152
## YearBuilt	4.180e+02	5.452e+01	7.667	3.59e-14 ***
## OpenPorchSF	1.760e+01	8.559e+00	2.057	0.039919 *
## GrLivArea	4.937e+01	2.964e+00	16.657	< 2e-16 ***
## TotRmsAbvGrd	-1.695e+03	6.414e+02	-2.643	0.008330 **
## GarageArea	2.333e+01	3.242e+00	7.198	1.07e-12 ***
## TotalBsmtSF	2.163e+01	2.870e+00	7.535	9.53e-14 ***
## Exterior1stAsphShn	-3.045e+03	1.830e+04	-0.166	0.867853
## Exterior1stBrkComm	-3.039e+04	1.397e+04	-2.176	0.029743 *
## Exterior1stBrkFace	1.646e+04	5.017e+03	3.282	0.001060 **
## Exterior1stCBlock	-8.865e+03	1.909e+04	-0.464	0.642470
## Exterior1stCemntBd	8.534e+03	5.512e+03	1.548	0.121801
## Exterior1stHdBoard	-5.218e+03	4.520e+03	-1.154	0.248612
## Exterior1stImStucc	1.025e+02	1.779e+04	0.006	0.995404
## Exterior1stMetalSd	9.382e+02	4.349e+03	0.216	0.829242
## Exterior1stPlywood	-3.588e+03	4.765e+03	-0.753	0.451574
## Exterior1stStone	-2.251e+03	1.315e+04	-0.171	0.864102
## Exterior1stStucco	9.208e+03	5.786e+03	1.591	0.111781
## Exterior1stVinylSd	-1.716e+03	4.459e+03	-0.385	0.700470
## Exterior1stWd Sdng	-5.484e+02	4.377e+03	-0.125	0.900318
## Exterior1stWdShing	-1.475e+02	5.503e+03	-0.027	0.978619
## ExterQualFa	-2.291e+03	8.149e+03	-0.281	0.778630
## ExterQualGd	-3.336e+03	4.805e+03	-0.694	0.487673
## ExterQualTA	-4.133e+03	5.028e+03	-0.822	0.411314
## YearRemodAdd	1.639e+02	3.758e+01	4.360	1.41e-05 ***
## MasVnrArea	1.073e+01	4.587e+00	2.339	0.019515 *
## MasVnrTypeBrkFace	1.065e+04	5.118e+03	2.081	0.037621 *
## MasVnrTypeNone	9.505e+03	5.152e+03	1.845	0.065271 .
## MasVnrTypeStone	1.745e+04	5.480e+03	3.184	0.001491 **
## WoodDeckSF	1.348e+01	4.350e+00	3.099	0.001989 **
## BsmtFinType1BLQ	-1.134e+03	1.947e+03	-0.582	0.560377
## BsmtFinType1GLQ	6.516e+03	1.800e+03	3.619	0.000308 ***
## BsmtFinType1LwQ	-3.599e+03	2.499e+03	-1.441	0.149970
## BsmtFinType1Rec	-2.125e+03	2.099e+03	-1.012	0.311620
## BsmtFinType1Unf	-8.328e+03	1.701e+03	-4.895	1.11e-06 ***
## BsmtFinType1Unk	3.677e+01	6.050e+03	0.006	0.995151
## MSSubClass30	-7.712e+02	3.415e+03	-0.226	0.821342
## MSSubClass40	-1.867e+03	1.044e+04	-0.179	0.858038
## MSSubClass45	-4.801e+03	1.514e+04	-0.317	0.751145
## MSSubClass50	-2.990e+02	6.263e+03	-0.048	0.961934
## MSSubClass60	8.268e+03	5.284e+03	1.565	0.117901

```

## MSSubClass70      4.366e+03  5.870e+03  0.744  0.457188
## MSSubClass75      7.205e+03  1.196e+04  0.602  0.547089
## MSSubClass80     -5.569e+03  8.027e+03 -0.694  0.487961
## MSSubClass85     -5.682e+03  7.168e+03 -0.793  0.428139
## MSSubClass90     -1.694e+04  3.657e+03 -4.633  3.99e-06 ***
## MSSubClass120    -7.718e+03  3.111e+03 -2.481  0.013249 *
## MSSubClass160    -2.080e+04  6.399e+03 -3.251  0.001181 **
## MSSubClass180    -1.270e+04  8.770e+03 -1.448  0.147782
## MSSubClass190    -2.333e+03  5.471e+03 -0.426  0.669850
## MSZoningFV       4.349e+04  8.743e+03  4.974  7.51e-07 ***
## MSZoningRH       3.068e+04  8.593e+03  3.571  0.000370 ***
## MSZoningRL       3.483e+04  7.351e+03  4.738  2.41e-06 ***
## MSZoningRM       2.891e+04  6.890e+03  4.196  2.92e-05 ***
## HouseStyle1.5Unf 1.109e+04  1.520e+04  0.730  0.465690
## HouseStyle1Story -1.245e+03  6.153e+03 -0.202  0.839742
## HouseStyle2.5Fin -3.052e+04  1.184e+04 -2.578  0.010061 *
## HouseStyle2.5Unf -4.657e+03  1.169e+04 -0.399  0.690324
## HouseStyle2Story -3.664e+03  5.582e+03 -0.656  0.511715
## HouseStyleSFoyer  5.860e+03  7.903e+03  0.742  0.458497
## HouseStyleSLvl   4.904e+03  8.947e+03  0.548  0.583666
## LotConfigCulDSac 2.001e+03  2.467e+03  0.811  0.417565
## LotConfigFR2     -6.985e+03  2.959e+03 -2.360  0.018415 *
## LotConfigFR3     -8.706e+03  9.152e+03 -0.951  0.341677
## LotConfigInside   -1.742e+03  1.287e+03 -1.353  0.176208
## FoundationCBlock 1.323e+03  2.261e+03  0.585  0.558571
## FoundationPConc  4.988e+03  2.470e+03  2.020  0.043624 *
## FoundationSlab   5.175e+03  6.607e+03  0.783  0.433616
## FoundationStone  8.711e+03  7.501e+03  1.161  0.245728
## FoundationWood   -2.825e+04  1.027e+04 -2.751  0.006027 **
## Condition1Feedr  3.353e+03  3.603e+03  0.930  0.352313
## Condition1Norm   7.510e+03  3.020e+03  2.487  0.013020 *
## Condition1PosA   1.237e+04  7.868e+03  1.572  0.116122
## Condition1PosN   9.669e+03  5.417e+03  1.785  0.074547 .
## Condition1RRAe   -1.679e+04  6.210e+03 -2.704  0.006941 **
## Condition1RRAn   1.971e+03  5.090e+03  0.387  0.698700
## Condition1RRNe   -1.295e+04  1.266e+04 -1.023  0.306586
## Condition1RRNn   1.849e+04  8.855e+03  2.088  0.037044 *
## Condition2Feedr  1.577e+04  1.725e+04  0.914  0.360901
## Condition2Norm   1.822e+04  1.507e+04  1.209  0.226885
## Condition2PosA   5.505e+04  2.695e+04  2.043  0.041312 *
## Condition2RRAn   -4.469e+02  2.315e+04 -0.019  0.984604
## Condition2RRNn   2.138e+04  1.959e+04  1.092  0.275233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16710 on 1208 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.9275, Adjusted R-squared:  0.9198
## F-statistic: 119.9 on 129 and 1208 DF,  p-value: < 2.2e-16

```

Finally, this derived model consists of about 27 predictor variables and produces little around 92.75% accuracy (Multiple R<sup>2</sup> value). Adjusted R-squared value is around 92%. So, this differences is around 0.0075 a small value - less than 0.8%. Thus this model has slight variation in accuracy but can be a good representation of the large population of real world data with p-value

of 15 zeroes prior to 22 i.e. very very small value and gives us good confidence that this model is a fairly accurate.

## 95 % confidence interval

	2.5 %	97.5 %
##		
## (Intercept)	-1.402535e+06	-9.309811e+05
## NeighborhoodBlueste	-1.604578e+04	3.816150e+04
## NeighborhoodBrDale	-7.984968e+03	2.369748e+04
## NeighborhoodBrkSide	-1.216193e+04	1.429717e+04
## NeighborhoodClearCr	-1.813716e+04	9.131933e+03
## NeighborhoodCollgCr	-2.247117e+04	-2.044505e+03
## NeighborhoodCrawfor	5.583352e+03	3.004279e+04
## NeighborhoodEdwards	-2.637222e+04	-3.677551e+03
## NeighborhoodGilbert	-2.294365e+04	-8.478095e+02
## NeighborhoodIDOTRR	-2.022641e+04	9.703633e+03
## NeighborhoodMeadowV	-3.463597e+04	-1.708544e+02
## NeighborhoodMitchel	-2.520629e+04	-1.993373e+03
## NeighborhoodNAmes	-2.456143e+04	-2.621355e+03
## NeighborhoodNoRidge	-8.157641e+03	1.652019e+04
## NeighborhoodNPkVill	-4.645344e+03	2.681115e+04
## NeighborhoodNridgHt	-4.324095e+03	1.714753e+04
## NeighborhoodNWAmes	-2.850400e+04	-5.797728e+03
## NeighborhoodOldTown	-2.520137e+04	1.733163e+03
## NeighborhoodSawyer	-2.271507e+04	2.216383e+02
## NeighborhoodSawyerW	-1.788886e+04	4.143267e+03
## NeighborhoodSomerst	-1.670596e+04	9.596852e+03
## NeighborhoodStoneBr	2.307892e+02	2.557988e+04
## NeighborhoodSWISU	-2.107106e+04	6.164502e+03
## NeighborhoodTimber	-1.685829e+04	6.734431e+03
## NeighborhoodVeenker	-1.360060e+03	2.863919e+04
## RoofStyleGable	-2.811654e+04	4.285366e+03
## RoofStyleGambrel	-2.291826e+04	1.597092e+04
## RoofStyleHip	-2.714253e+04	5.433244e+03
## RoofStyleMansard	-2.547990e+04	1.663326e+04
## LotArea	1.205986e+00	2.145739e+00
## LotFrontage	-4.086941e+01	2.767751e+01
## OverallQual2	-2.856396e+04	5.396923e+04
## OverallQual3	-3.250164e+04	4.316632e+04
## OverallQual4	-3.046610e+04	4.417662e+04
## OverallQual5	-2.915484e+04	4.586266e+04
## OverallQual6	-2.557269e+04	4.959055e+04
## OverallQual7	-1.530258e+04	6.011114e+04
## OverallQual8	2.203433e+03	7.812247e+04
## OverallQual9	3.359846e+04	1.113953e+05
## OverallQual10	3.005308e+04	1.215148e+05
## OverallCond2	-6.279347e+04	4.192204e+04
## OverallCond3	-5.427586e+04	4.325410e+04
## OverallCond4	-4.597130e+04	5.319013e+04
## OverallCond5	-3.754328e+04	6.159707e+04
## OverallCond6	-3.142518e+04	6.777996e+04
## OverallCond7	-2.468316e+04	7.460244e+04
## OverallCond8	-2.172877e+04	7.786697e+04
## OverallCond9	-1.667127e+04	8.426394e+04

```

## YearBuilt          3.110587e+02  5.249870e+02
## OpenPorchSF       8.120258e-01  3.439635e+01
## GrLivArea         4.355148e+01  5.518074e+01
## TotRmsAbvGrd     -2.953587e+03 -4.367040e+02
## GarageArea        1.697331e+01  2.969310e+01
## TotalBsmtSF      1.599781e+01  2.726110e+01
## Exterior1stAsphShn -3.894715e+04  3.285639e+04
## Exterior1stBrkComm -5.779688e+04 -2.991076e+03
## Exterior1stBrkFace 6.622232e+03  2.630756e+04
## Exterior1stCBlock -4.631784e+04  2.858842e+04
## Exterior1stCemntBd -2.279632e+03  1.934817e+04
## Exterior1stHdBoard -1.408622e+04  3.650797e+03
## Exterior1stImStucc -3.480350e+04  3.500852e+04
## Exterior1stMetalSd -7.594495e+03  9.470889e+03
## Exterior1stPlywood -1.293634e+04  5.760153e+03
## Exterior1stStone   -2.805038e+04  2.354809e+04
## Exterior1stStucco  -2.143913e+03  2.055989e+04
## Exterior1stVinylSd -1.046496e+04  7.033191e+03
## Exterior1stWd_Sdng -9.135300e+03  8.038597e+03
## Exterior1stWdShing -1.094355e+04  1.064853e+04
## ExterQualFa        -1.827897e+04  1.369651e+04
## ExterQualGd        -1.276387e+04  6.091804e+03
## ExterQualTA        -1.399814e+04  5.732722e+03
## YearRemodAdd       9.012503e+01  2.375817e+02
## MasVnrArea         1.727942e+00  1.972469e+01
## MasVnrTypeBrkFace 6.107305e+02  2.069442e+04
## MasVnrTypeNone    -6.019663e+02  1.961224e+04
## MasVnrTypeStone   6.695029e+03  2.819773e+04
## WoodDeckSF         4.944807e+00  2.201348e+01
## BsmtFinType1BLQ   -4.954894e+03  2.686404e+03
## BsmtFinType1GLQ   2.983494e+03  1.004759e+04
## BsmtFinType1LwQ   -8.501300e+03  1.302689e+03
## BsmtFinType1Rec   -6.244133e+03  1.993798e+03
## BsmtFinType1Unf   -1.166538e+04 -4.990189e+03
## BsmtFinType1Unk   -1.183267e+04  1.190622e+04
## MSSubClass30       -7.470418e+03  5.927926e+03
## MSSubClass40       -2.233968e+04  1.860582e+04
## MSSubClass45       -3.449882e+04  2.489597e+04
## MSSubClass50       -1.258696e+04  1.198899e+04
## MSSubClass60       -2.098597e+03  1.863469e+04
## MSSubClass70       -7.150722e+03  1.588186e+04
## MSSubClass75       -1.626493e+04  3.067521e+04
## MSSubClass80       -2.131714e+04  1.017947e+04
## MSSubClass85       -1.974508e+04  8.381524e+03
## MSSubClass90       -2.411921e+04 -9.769781e+03
## MSSubClass120      -1.382161e+04 -1.613869e+03
## MSSubClass160      -3.335851e+04 -8.250124e+03
## MSSubClass180      -2.990743e+04  4.504092e+03
## MSSubClass190      -1.306578e+04  8.399860e+03
## MSZoningFV         2.633466e+04  6.064246e+04
## MSZoningRH         1.382493e+04  4.754114e+04
## MSZoningRL         2.040941e+04  4.925495e+04
## MSZoningRM         1.539175e+04  4.242732e+04
## HouseStyle1.5Unf  -1.872981e+04  4.091457e+04

```

```

## HouseStyle1Story      -1.331614e+04  1.082707e+04
## HouseStyle2.5Fin     -5.373973e+04  -7.290386e+03
## HouseStyle2.5Unf     -2.758564e+04   1.827112e+04
## HouseStyle2Story     -1.461458e+04   7.287303e+03
## HouseStyleSFoyer     -9.644062e+03   2.136459e+04
## HouseStyleSLvl       -1.264825e+04   2.245707e+04
## LotConfigCulDSac    -2.839421e+03   6.840428e+03
## LotConfigFR2        -1.279004e+04  -1.178991e+03
## LotConfigFR3        -2.666193e+04   9.250135e+03
## LotConfigInside      -4.267233e+03   7.833785e+02
## FoundationCBlock    -3.112896e+03   5.758799e+03
## FoundationPConc     1.429359e+02   9.833685e+03
## FoundationSlab      -7.787189e+03   1.813725e+04
## FoundationStone     -6.005010e+03   2.342731e+04
## FoundationWood      -4.838734e+04  -8.103013e+03
## Condition1Feedr     -3.716615e+03   1.042208e+04
## Condition1Norm      1.585439e+03   1.343538e+04
## Condition1PosA      -3.064736e+03   2.780651e+04
## Condition1PosN      -9.596227e+02   2.029698e+04
## Condition1RRAe      -2.897784e+04  -4.610146e+03
## Condition1RRAn     -8.014816e+03   1.195584e+04
## Condition1RRNe      -3.778628e+04   1.188850e+04
## Condition1RRNn      1.112733e+03   3.585899e+04
## Condition2Feedr     -1.807634e+04   4.960833e+04
## Condition2Norm      -1.134419e+04   4.777844e+04
## Condition2PosA      2.173432e+03   1.079338e+05
## Condition2RRAn     -4.587019e+04   4.497647e+04
## Condition2RRNn      -1.705024e+04   5.981909e+04

```

### analysis of variance

```
anova(housing_mod_lm)
```

```

## Analysis of Variance Table
##
## Response: SalePrice
##                    Df  Sum Sq  Mean Sq  F value    Pr(>F)
## Neighborhood      24 2.5828e+12 1.0762e+11 385.4439 < 2.2e-16 ***
## RoofStyle          4 5.4117e+10 1.3529e+10 48.4569 < 2.2e-16 ***
## LotArea             1 3.2302e+11 3.2302e+11 1156.9286 < 2.2e-16 ***
## LotFrontage         1 3.5851e+09 3.5851e+09 12.8406 0.0003527 ***
## OverallQual         9 7.0095e+11 7.7883e+10 278.9506 < 2.2e-16 ***
## OverallCond         8 3.3621e+10 4.2027e+09 15.0524 < 2.2e-16 ***
## YearBuilt            1 7.11414e+10 7.11414e+10 254.8011 < 2.2e-16 ***
## OpenPorchSF          1 2.8734e+10 2.8734e+10 102.9146 < 2.2e-16 ***
## GrLivArea            1 2.9544e+11 2.9544e+11 1058.1680 < 2.2e-16 ***
## TotRmsAbvGrd         1 1.0907e+10 1.0907e+10 39.0656 5.671e-10 ***
## GarageArea            1 2.4246e+10 2.4246e+10 86.8398 < 2.2e-16 ***
## TotalBsmtSF          1 5.8762e+10 5.8762e+10 210.4637 < 2.2e-16 ***
## Exterior1st          14 1.9818e+10 1.4156e+09 5.0701 2.567e-09 ***
## ExterQual              3 1.6559e+09 5.5196e+08 1.9769 0.1156329
## YearRemodAdd          1 5.0122e+09 5.0122e+09 17.9519 2.438e-05 ***
## MasVnrArea             1 3.4172e+09 3.4172e+09 12.2393 0.0004849 ***

```

```

## MasVnrType      3 2.1021e+09 7.0069e+08    2.5096 0.0573383 .
## WoodDeckSF     1 5.1416e+09 5.1416e+09    18.4152 1.918e-05 ***
## BsmtFinType1   6 3.0695e+10 5.1159e+09    18.3233 < 2.2e-16 ***
## MSSubClass     14 3.1769e+10 2.2692e+09    8.1275 < 2.2e-16 ***
## MSZoning       4 8.4356e+09 2.1089e+09    7.5534 5.211e-06 ***
## HouseStyle     7 3.8457e+09 5.4939e+08    1.9677 0.0563050 .
## LotConfig      4 2.1583e+09 5.3957e+08    1.9326 0.1027184
## Foundation     5 4.4389e+09 8.8777e+08    3.1797 0.0073949 **
## Condition1    8 1.0052e+10 1.2565e+09    4.5002 2.110e-05 ***
## Condition2    5 1.5241e+09 3.0483e+08    1.0918 0.3630854
## Residuals     1208 3.3728e+11 2.7920e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Analysis of Variance statistics for the given model help to confirm the significance of selected variables.

#### assumptions of independence durbin watson test

```

##  lag Autocorrelation D-W Statistic p-value
##    1      0.03595457      1.928091    0.246
## Alternative hypothesis: rho != 0

```

From the Durbin-Watson test, we can see that Autocorrelation coefficient is around 0.03 - a small value and D-W statistic value is around 1.92, which is very close to 2. This indicates that the predictor variables we have selected may not be correlations amongst themselves. So, we should be good with these. Only slight concern is p-value is around 0.15 (above 0.05).

#### assumptions of no multicollinearity

```
v <- vif(housing_mod_lm)
```

```
v
```

```

##                               GVIF Df GVIF^(1/(2*Df))
## Neighborhood 6.260709e+04 24      1.258722
## RoofStyle    2.251882e+00 4       1.106798
## LotArea      2.795870e+00 1       1.672085
## LotFrontage  1.462912e+00 1       1.209509
## OverallQual  1.062198e+02 9       1.295887
## OverallCond  1.891167e+01 8       1.201699
## YearBuilt    1.288356e+01 1       3.589367
## OpenPorchSF  1.429299e+00 1       1.195533
## GrLivArea    8.476724e+00 1       2.911481
## TotRmsAbvGrd 4.491318e+00 1       2.119273
## GarageArea   2.037311e+00 1       1.427344
## TotalBsmtSF  5.659159e+00 1       2.378899
## Exterior1st  5.802099e+01 14      1.156073
## ExterQual    1.102684e+01 3       1.491907
## YearRemodAdd 2.884558e+00 1       1.698399
## MasVnrArea   2.506809e+00 1       1.583291
## MasVnrType   4.698417e+00 3       1.294172

```

```

## WoodDeckSF    1.273914e+00   1      1.128678
## BsmtFinType1 1.821869e+01   6      1.273629
## MSSubClass    7.312065e+07  14     1.909231
## MSZoning      4.599684e+01   4      1.613768
## HouseStyle    1.112154e+06   7      2.703142
## LotConfig     1.998494e+00   4      1.090405
## Foundation    2.881236e+01   5      1.399452
## Condition1   4.620963e+00   8      1.100388
## Condition2   5.605796e+00   5      1.188129

```

[1/v](#)

```

##                  GVIF          Df  GVIF^(1/(2*Df))
## Neighborhood 1.597263e-05 0.04166667      0.7944569
## RoofStyle    4.440730e-01 0.25000000      0.9035076
## LotArea      3.576705e-01 1.00000000      0.5980556
## LotFrontage   6.835682e-01 1.00000000      0.8267818
## OverallQual  9.414441e-03 0.11111111      0.7716725
## OverallCond   5.287741e-02 0.12500000      0.8321553
## YearBuilt    7.761832e-02 1.00000000      0.2786006
## OpenPorchSF   6.996435e-01 1.00000000      0.8364470
## GrLivArea    1.179701e-01 1.00000000      0.3434678
## TotRmsAbvGrd 2.226518e-01 1.00000000      0.4718599
## GarageArea   4.908430e-01 1.00000000      0.7006019
## TotalBsmtSF  1.767047e-01 1.00000000      0.4203626
## Exterior1st  1.723514e-02 0.07142857      0.8649974
## ExterQual    9.068780e-02 0.33333333      0.6702829
## YearRemodAdd 3.466736e-01 1.00000000      0.5887899
## MasVnrArea   3.989135e-01 1.00000000      0.6315960
## MasVnrType   2.128376e-01 0.33333333      0.7726950
## WoodDeckSF   7.849822e-01 1.00000000      0.8859922
## BsmtFinType1 5.488869e-02 0.16666667      0.7851577
## MSSubClass   1.367603e-08 0.07142857      0.5237710
## MSZoning     2.174063e-02 0.25000000      0.6196679
## HouseStyle   8.991560e-07 0.14285714      0.3699398
## LotConfig    5.003769e-01 0.25000000      0.9170904
## Foundation   3.470732e-02 0.20000000      0.7145656
## Condition1   2.164051e-01 0.12500000      0.9087705
## Condition2   1.783868e-01 0.20000000      0.8416592

```

```

t <- 0

for(i in 1:24)
{
  t = t + v[i, 3]
}

t/24

```

```

## [1] 1.654531

```

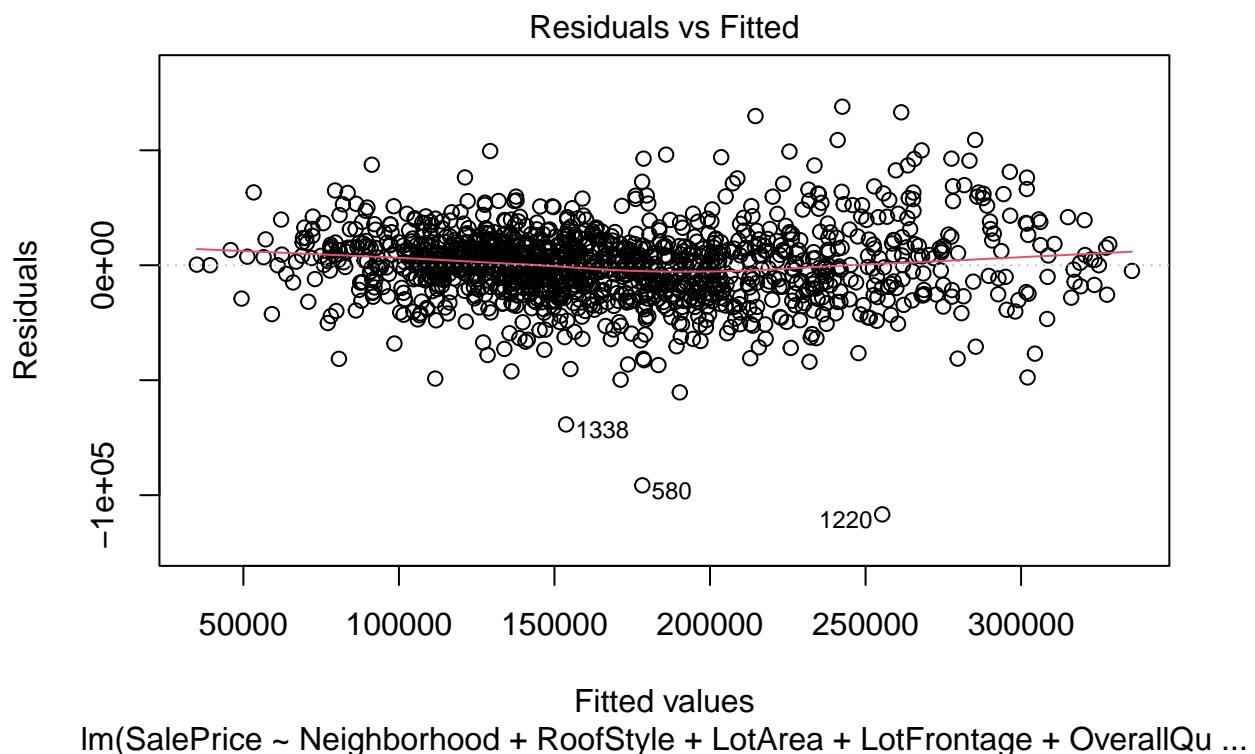
Based on (Field, Miles, and Field 2012) book

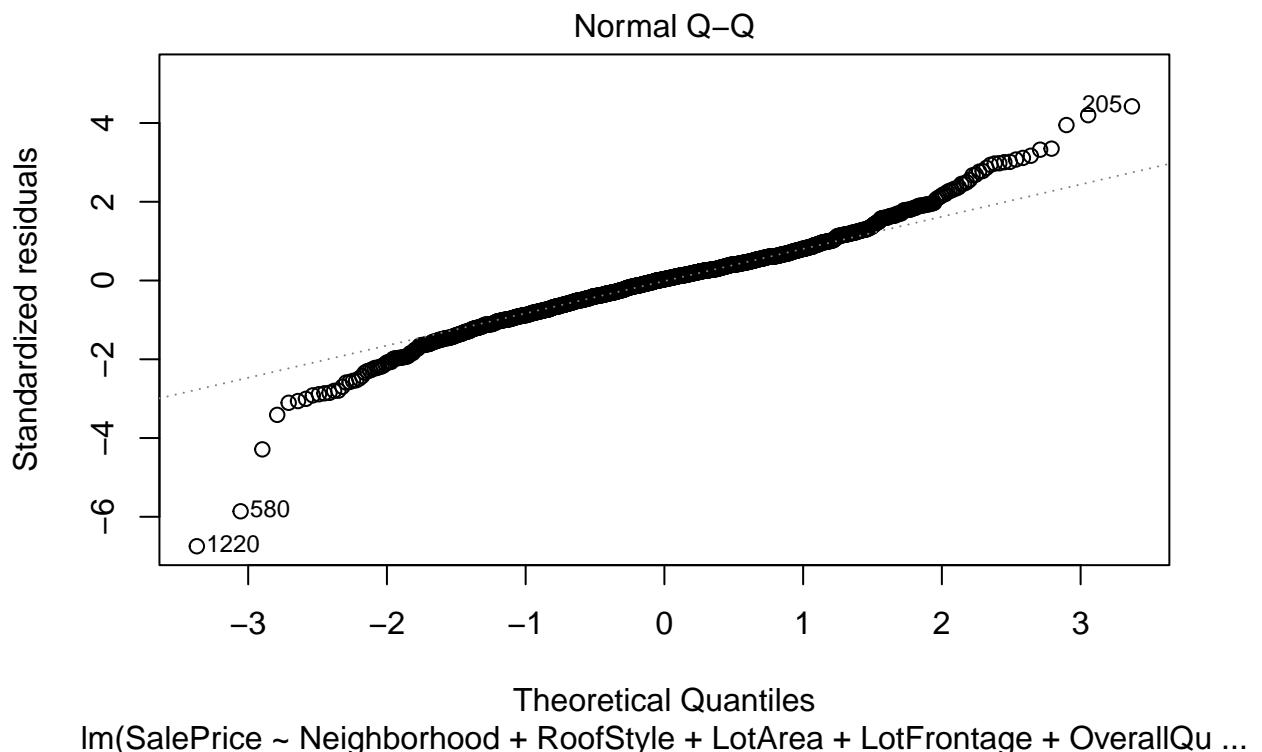
All of the VIF values are well below 10. SO, there is no cause of concerns.

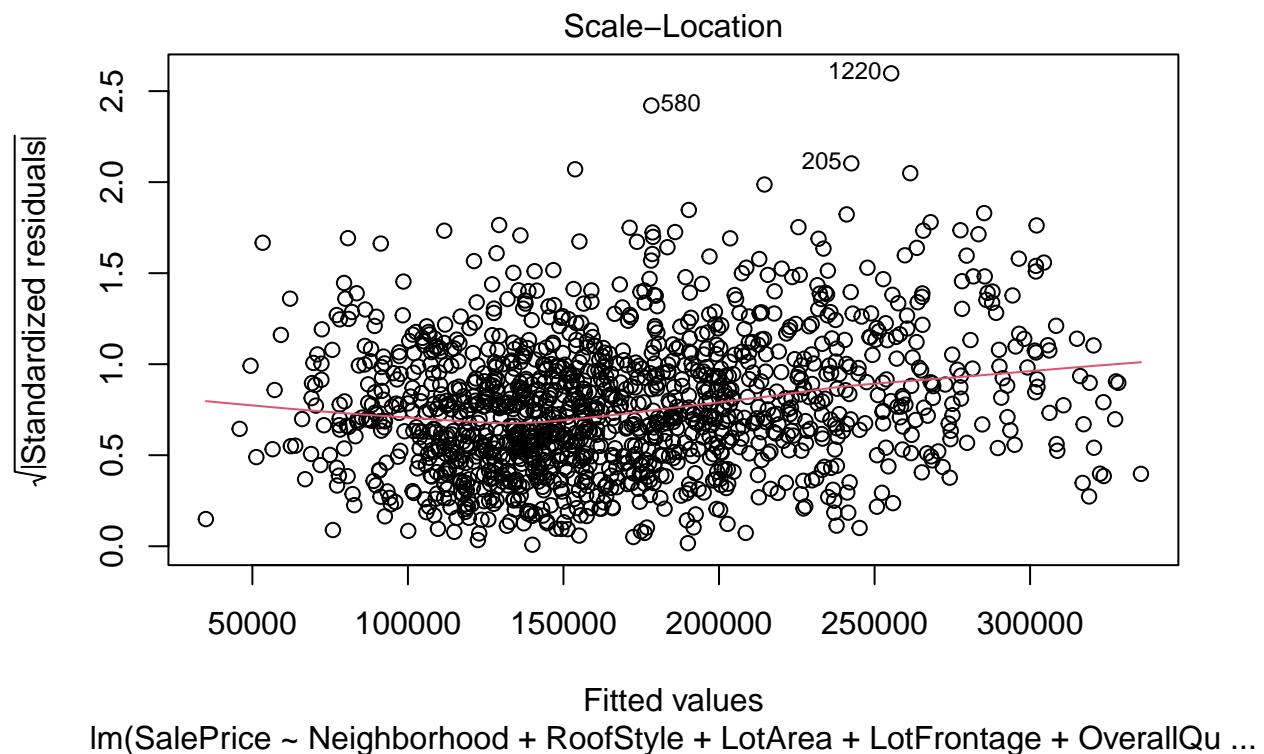
All of the tolerance values are well above 0.1, so there is no cause of concerns.

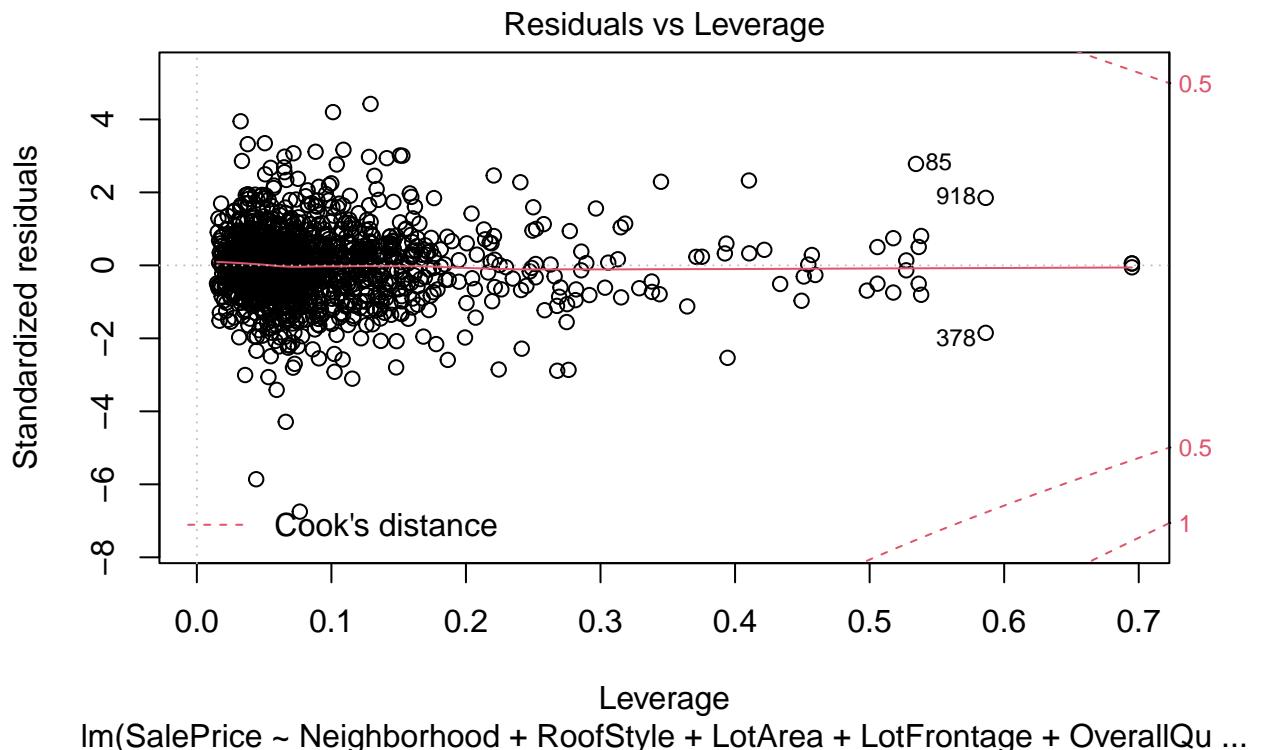
Mean of VIF values is little bit above 1, indicates our model might be slightly biased and may be needs to consider additional predictor variables or more cleaning of data needed or additional / larger size of data is needed. Current data size after filtering became 1344 records.

### plot() and hist() function

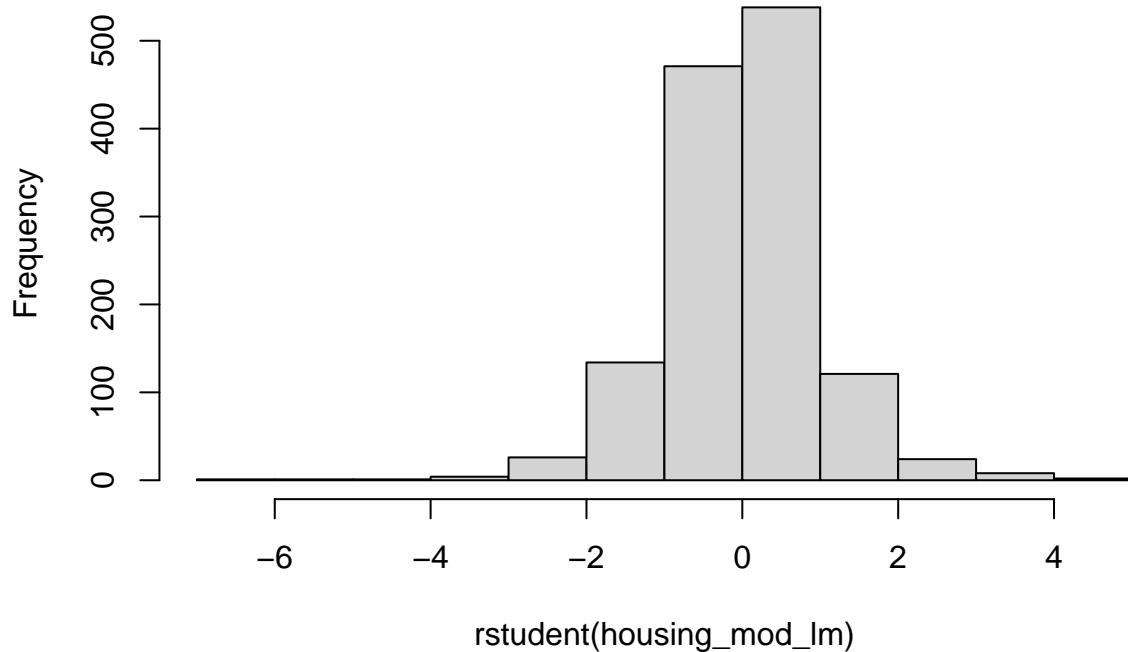




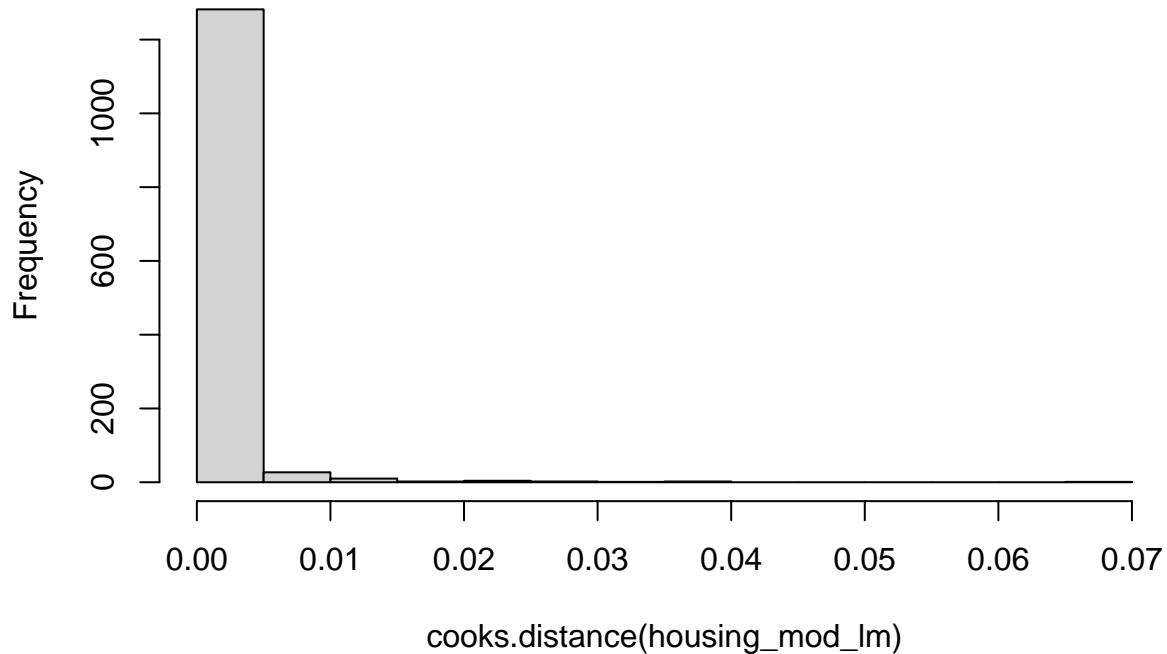




**Histogram of rstudent(housing\_mod\_lm)**



## Histogram of cooks.distance(housing\_mod\_lm)



As we can see, fitted line on Residuals vs Fitted values, is close to 0 residuals line and as we go towards the lower and upper extremes, the line seems to deviate from ideal fitted line. It can be improved with possibly more number of records for model creation and possibly reducing some of outliers on the lower and upper ends. Also in this plot, residuals in our model shows a fairly random pattern, which is indicative of situation in which the assumptions of linearity, randomness and homoscedasticity have been met.

Normal Q-Q plot shows most of the records fall between 2 standard deviations of the mean, mostly along the straight line fit and thus the model is a good representation of the data, for predictions.

Histogram of Studentized residuals also shows a close to normal distribution up to 2 standard deviations around mean, although we can confirm the same cases earlier that we might be having some level of outliers / skewness in the distribution which might be causing slight deviation of the residuals from the straight line.

All of the records have Cook's distance less than 1, hence we should be good about the model.

Looking at the model, it is fairly close representation of the sample and a generalizable model to the larger population. It can be improved further by deleting outliers for model building purpose. We may need to consider additional predictor variables or more cleaning of data is needed. Another way, perhaps could be finding additional / larger size of data, which can smoothen out the normal distribution even further and help improve accuracy of the model. Current data size after filtering became 1344 records. So, possibly a little larger data size could help.

## **Conclusion**

Overall, people interested in buying a house need to consider not only the basic factors like area / size, number of rooms / bathrooms, neighborhood, garage area, year of initial built but also the overall quality of the construction, Foundations well Roof types, when the house is remodeled, if applicable and Exteriors, open areas and masonry work along with Basement finishing. Based on different geographic locations, the proximity to key essentials and amenities like hospitals, schools, shopping malls, commercial buildings etc. are also important.

## **References**

- Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using r*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.