MATH 531T-B : Assignment 4

Christopher Bradbury, Louis Bensard, Anthony Keys

July 24, 2018

Analysis of the Housing Dataset

Cleaning Data

First of all, the dataset contiains many missing values, so we decide to delete any predictor that has more than 10% of missing values, and then we delete any row that contains at least one missing value.

```
nobs = dim(data)[1]
nvar = dim(data)[2]

#cleaning data
count_na = rep(0,nvar)

for(i in 1:nvar) count_na[i] = sum(is.na(data[,i])==TRUE)

#this vector tell me the proportion of NA for each predictor
count_na = count_na/nobs

#We decide to drop the predictors that contain more than 10% of missing values
var_to_drop = which(count_na>0.10)

data1 = data[,-var_to_drop]

#we drop the rows for which at least one of the remaining predictor have a missing value.
data2 = na.omit(data1)
```

The resulting dataset *data*2 is now clean ans without missing value. We are aware this is a very aggressive way of dealing with missing values. With more time, we could have use an EM approach instead.

Analysis of the Categorical part

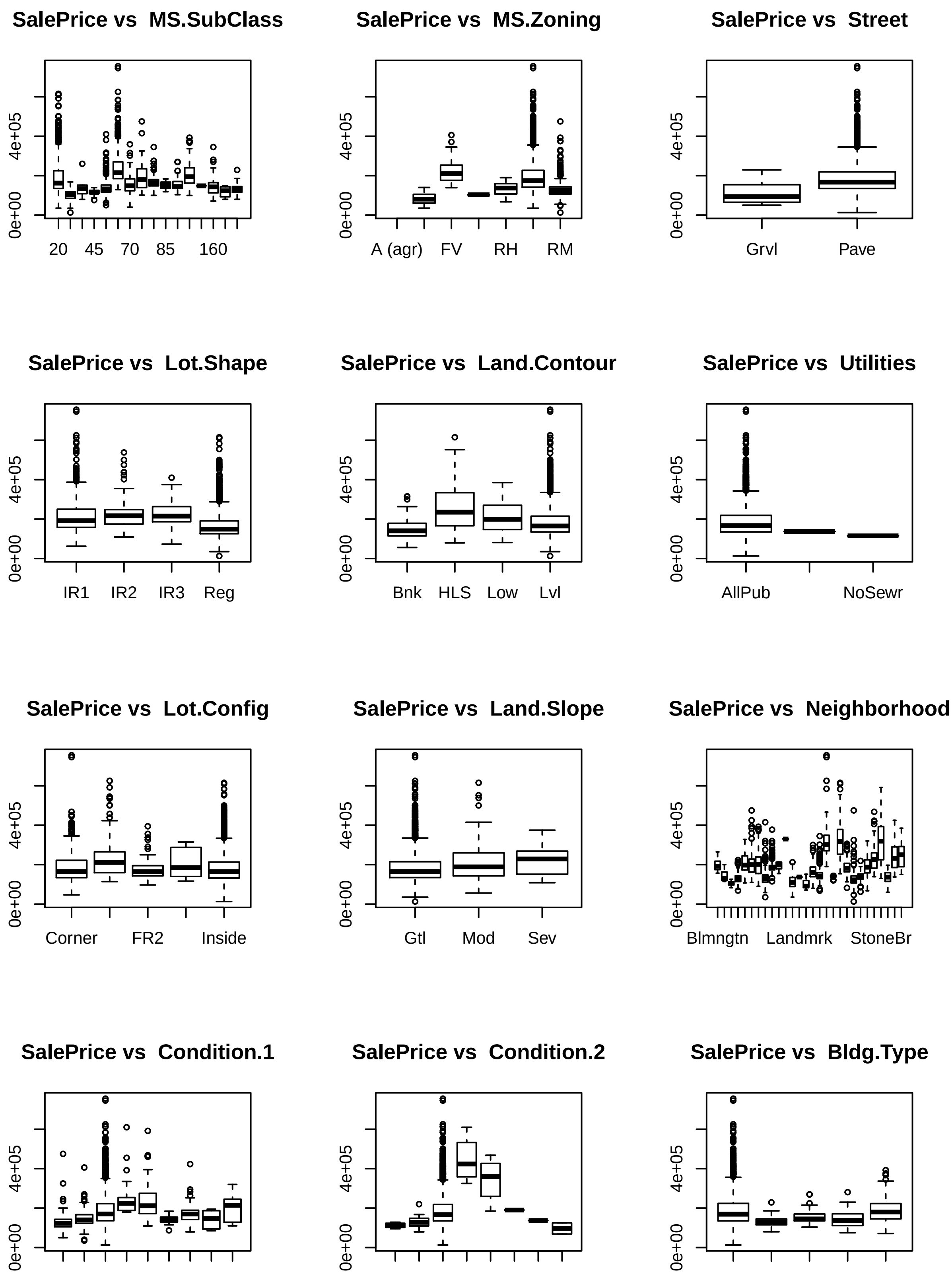
The dataset is a mix of categorical and continuous variables. We decided to split the analysis into 2 parts: An analysis of the cegorical variables of the dataset and another analysis of the continous variables of the dataset. We have 53 categorical variables and 22 continous variables left in the dataset.

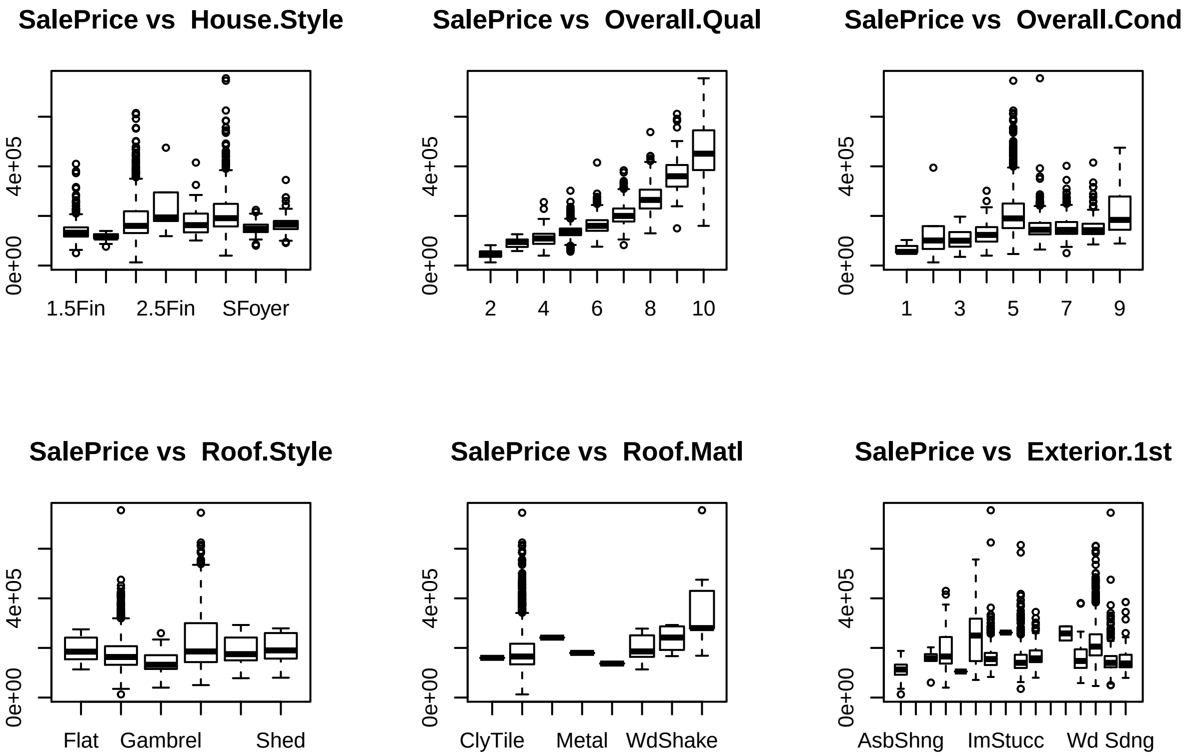
```
#splitting the data into continous and categorical

cont_var = colnames(data2) %in%  c('Lot.Frontage','Pool.Area','Garage.Area','Total.Bsmt.SF',
                                 'Lot.Area','Misc.Val', 'PoolArea',
                                 'Screen.Porch', 'X3Ssn.Porch', 'Enclosed.Porch',
                                 'Open.Porch.SF', 'Wood.Deck.SF', 'Garage.Area',
                                 'Gr.Liv.Area', 'Low.Qual.Fin.SF', 'X2nd.Flr.SF',
                                 'X1st.Flr.SF', 'Total.Bsmt.SF', 'Bsmt.Unf.SF',
                                 'BsmtFin.SF.2', 'BsmtFin.SF.1', 'Mas.Vnr.Area',
                                 'Lot.Area', 'Year.Built', 'Year.Remod.Add',
                                 'Garage.Yr.Blt', 'Yr.Sold')
```

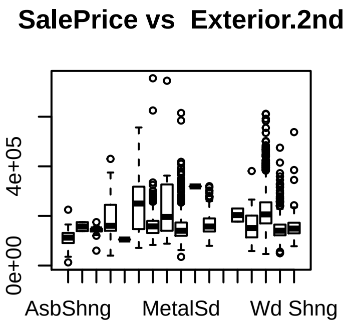
```
#categorical predictors
x_cat = data2[,(!cont_var)]
x_cat = x_cat[,-c(1,2)] #getting rid of order and PID that are irrelevant

#continuous predictors
x_cont = data2[, cont_var]
```

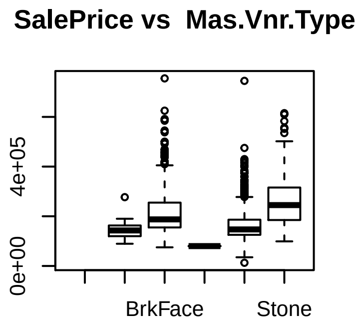
EDA Categorical part

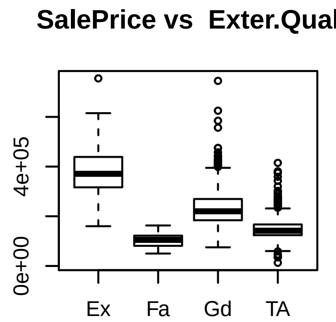
First, we are goin to plot a boxplot of each categorical predictor vs SalePrice. A rough overview of those plots can give us a lead on which predictors to focus on.

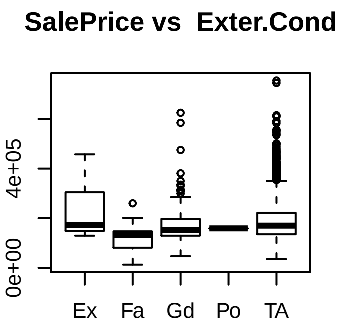

SalePrice vs MS.SubClass
0e+00
4e+05
20 45 70 85 160
SalePrice vs MS.Zoning
0e+00
4e+05
A (agr) FV RH RM
SalePrice vs Street
0e+00
4e+05
Grvl Pave
SalePrice vs Lot.Shape
0e+00
4e+05
IR1 IR2 IR3 Reg
SalePrice vs Land.Contour
0e+00
4e+05
Bnk HLS Low Lvl
SalePrice vs Utilities
0e+00
4e+05
AllPub NoSewr
SalePrice vs Lot.Config
0e+00
4e+05
Corner FR2 Inside
SalePrice vs Land.Slope
0e+00
4e+05
Gtl Mod Sev
SalePrice vs Neighborhood
0e+00
4e+05
Blmngtn Landmrk StoneBr
SalePrice vs Condition.1
0e+00
4e+05
Artery PosA RRAn
SalePrice vs Condition.2
0e+00
4e+05
Artery PosA RRAn
SalePrice vs Bldg.Type
0e+00
4e+05
1Fam Duplex TwnhsE

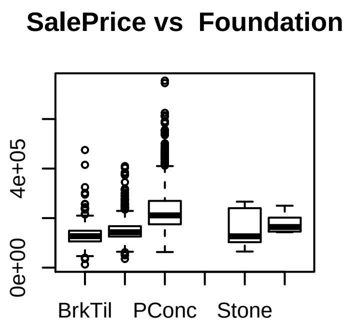

SalePrice vs House.Style
0e+00
4e+05
1.5Fin
2.5Fin
SFoyer
SalePrice vs Overall.Qual
0e+00
4e+05
2
4
6
8
10
SalePrice vs Overall.Cond
0e+00
4e+05
1
3
5
7
9
SalePrice vs Roof.Style
0e+00
4e+05
Flat
Gambrel
Shed
SalePrice vs Roof.Matl
0e+00
4e+05
ClyTile
Metal
WdShake
SalePrice vs Exterior.1st
0e+00
4e+05
AsbShng
ImStucc
Wd Sdng

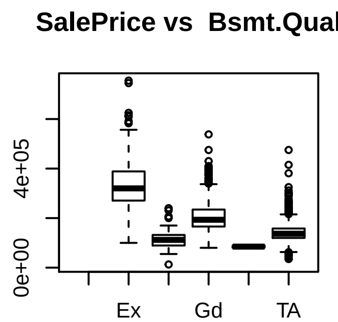

SalePrice vs Exterior.2nd
0e+00
4e+05
AsbShng
MetalSd
Wd Shng

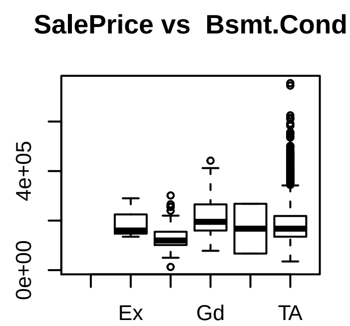

SalePrice vs Mas.Vnr.Type
0e+00
4e+05
BrkFace
Stone

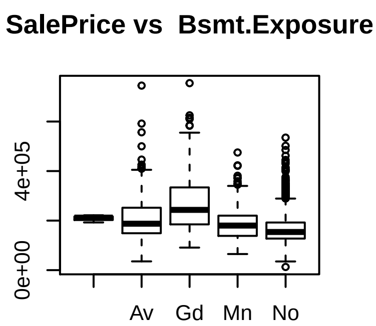

SalePrice vs Exter.Qual
0e+00
4e+05
Ex
Fa
Gd
TA

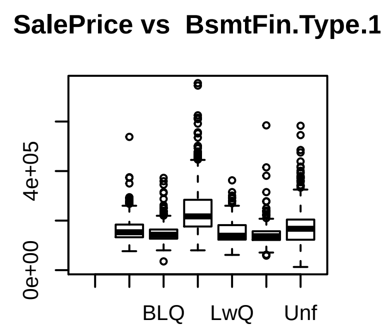

SalePrice vs Exter.Cond
0e+00
4e+05
Ex
Fa
Gd
Po
TA

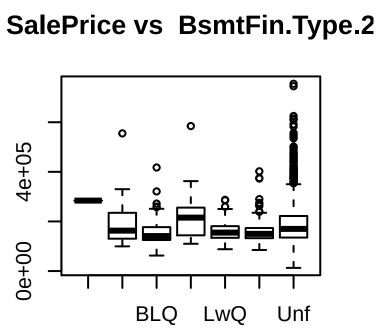

SalePrice vs Foundation
0e+00
4e+05
BrkTil
PConc
Stone

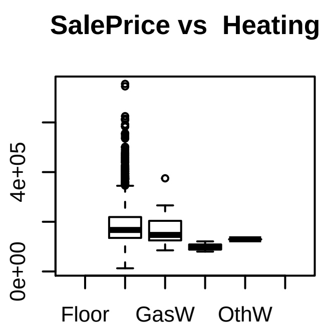

SalePrice vs Bsmt.Qual
0e+00
4e+05
Ex
Gd
TA

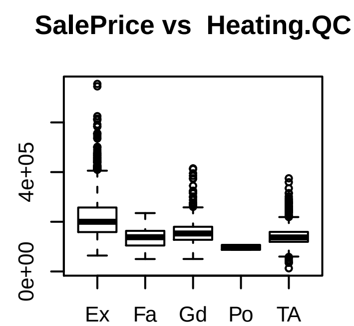

SalePrice vs Bsmt.Cond
0e+00
4e+05
Ex
Gd
TA

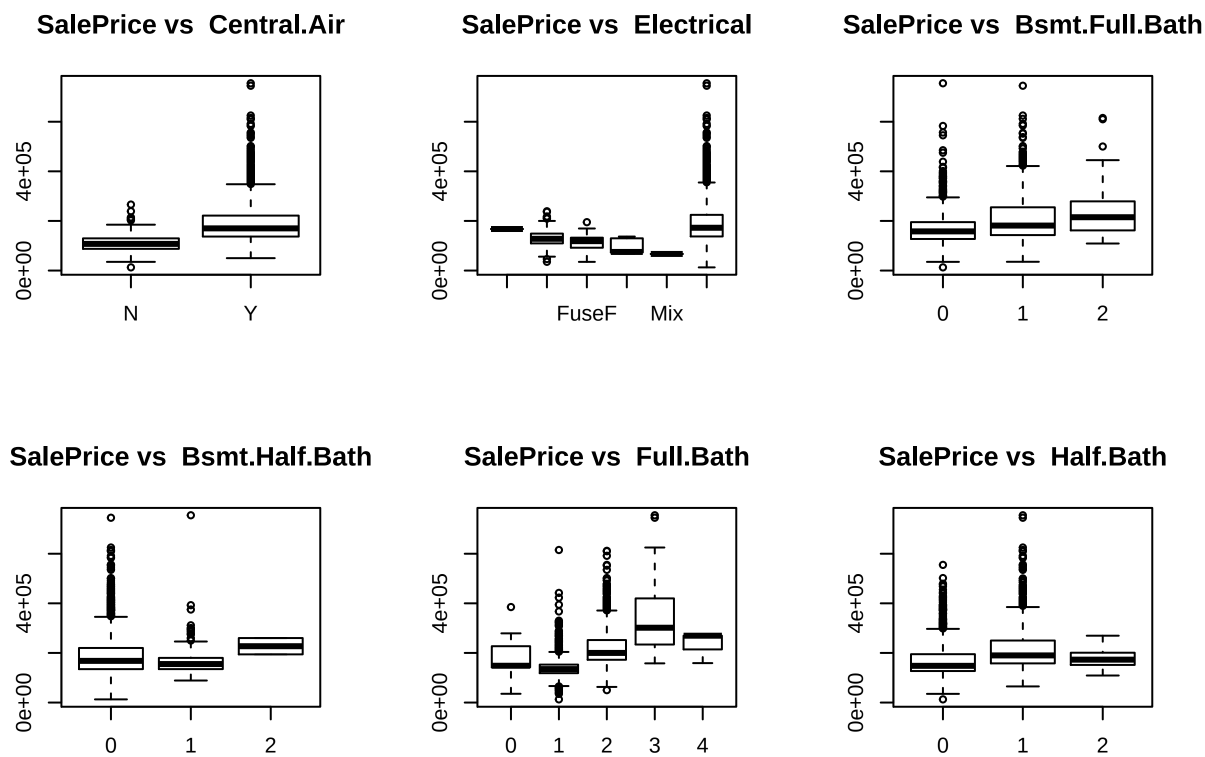

SalePrice vs Bsmt.Exposure
0e+00
4e+05
Av
Gd
Mn
No

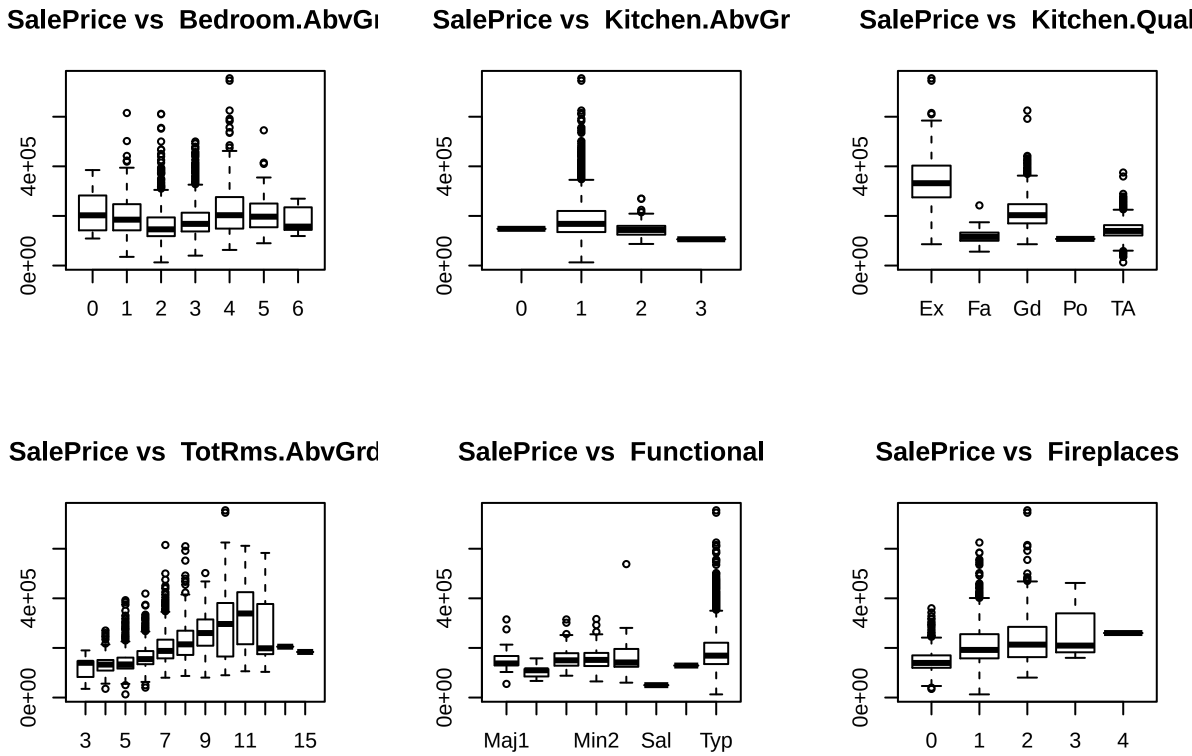

SalePrice vs BsmtFin.Type.1
0e+00
4e+05
BLQ
LwQ
Unf

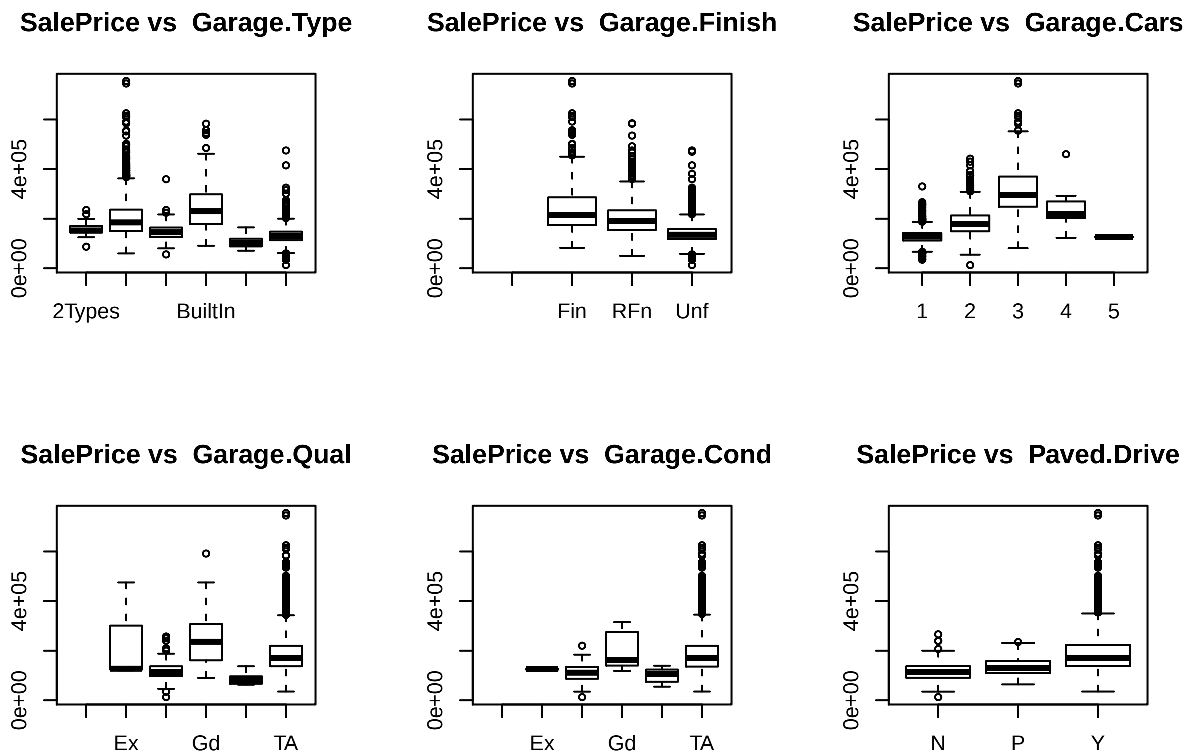

SalePrice vs BsmtFin.Type.2
0e+00
4e+05
BLQ
LwQ
Unf

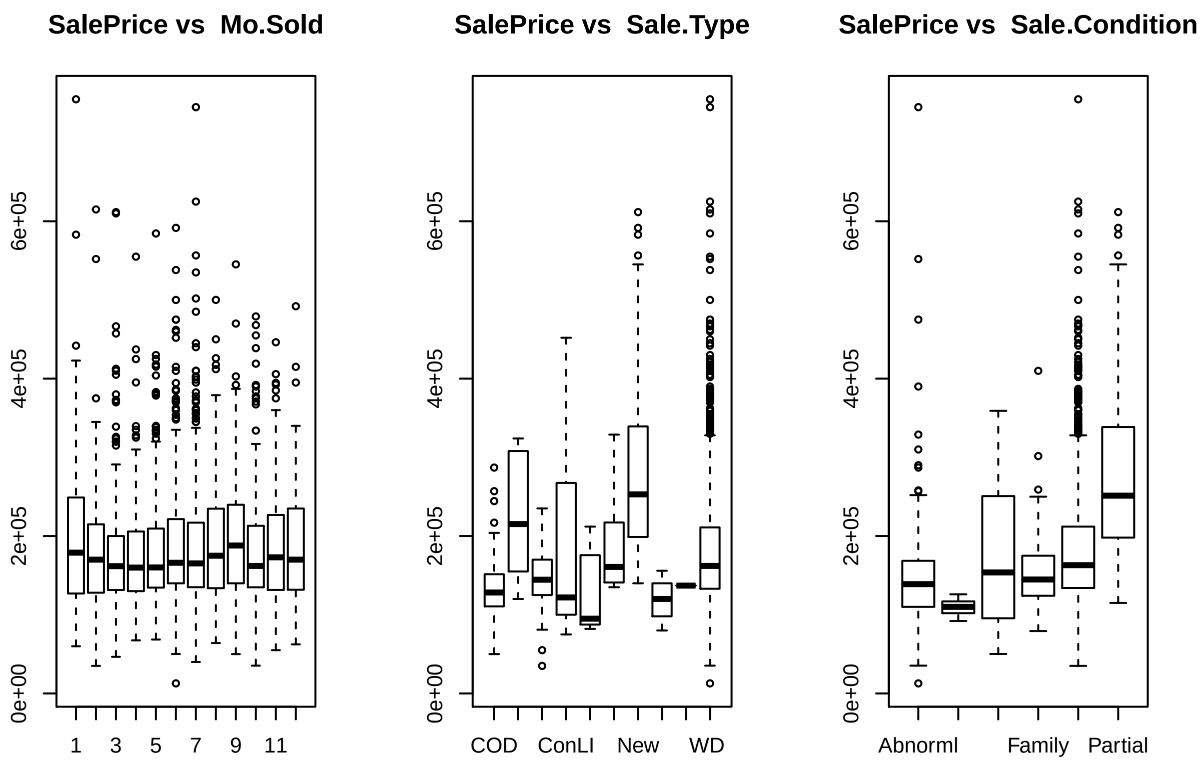

SalePrice vs Heating
0e+00
4e+05
Floor
GasW
OthW


SalePrice vs Heating.QC
0e+00
4e+05
Ex
Fa
Gd
Po
TA


SalePrice vs Central.Air
0e+00
4e+05
N
Y
SalePrice vs Electrical
0e+00
4e+05
FuseF
Mix
SalePrice vs Bsmt.Full.Bath
0e+00
4e+05
0
1
2
SalePrice vs Bsmt.Half.Bath
0e+00
4e+05
0
1
2
SalePrice vs Full.Bath
0e+00
4e+05
0
1
2
3
4
SalePrice vs Half.Bath
0e+00
4e+05
0
1
2


SalePrice vs Bedroom.AbvGr
SalePrice vs Kitchen.AbvGr
SalePrice vs Kitchen.Qual
SalePrice vs TotRms.AbvGrd
SalePrice vs Functional
SalePrice vs Fireplaces
0e+00
4e+05
0 1 2 3 4 5 6
0 1 2 3
Ex Fa Gd Po TA
3 5 7 9 11 15
Maj1 Min2 Sal Typ
0 1 2 3 4


SalePrice vs Garage.Type
0e+00
4e+05
2Types
BuiltIn
SalePrice vs Garage.Finish
0e+00
4e+05
Fin
RFn
Unf
SalePrice vs Garage.Cars
0e+00
4e+05
1
2
3
4
5
SalePrice vs Garage.Qual
0e+00
4e+05
Ex
Gd
TA
SalePrice vs Garage.Cond
0e+00
4e+05
Ex
Gd
TA
SalePrice vs Paved.Drive
0e+00
4e+05
N
P
Y



A box that stands out of the group in one of the above boxplots might indicate that the variable in question has a significant impacton the Sale Price. Here is the exhaustive list of what we think are the 23 possibly significant variables:

Bsmt.Exposure, Bsmt.Qual, Bsmt.Fin.Type1, Ms.Subclass, Ms.Zoning, Land.Contour, Neighborhood, Condition1, Condition2, House.Style, Overall.Qual, Overall.Cond, Exterior1st, Exterior2nd, Exterior.Qual, Foundation, Central.Air, Kitchen.Qual, Kitchen.AbvGr, Garage.Cars, Sale.Type, Sale.Cond and RoofMatl.

Now let's perform a simple ANOVA on all categorical predictors and see which predictors are considered statistically significant:

```
model_cat = lm(SalePrice~., data=x_cat)
anova(model_cat)
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##                   Df    Sum Sq   Mean Sq F value   Pr(>F)
## MS.SubClass        1 9.19e+10 9.19e+10  116.52 < 2e-16 ***
## MS.Zoning          5 1.66e+12 3.33e+11  422.10 < 2e-16 ***
## Street             1 3.20e+09 3.20e+09    4.06 0.04395 *
## Lot.Shape          3 8.62e+11 2.87e+11  364.71 < 2e-16 ***
## Land.Contour       3 5.33e+11 1.78e+11  225.43 < 2e-16 ***
## Utilities          2 1.04e+10 5.21e+09    6.61 0.00137 **
## Lot.Config         4 4.59e+10 1.15e+10   14.56 9.2e-12 ***
## Land.Slope         2 1.15e+10 5.76e+09    7.31 0.00068 ***
## Neighborhood      27 6.82e+12 2.52e+11  320.26 < 2e-16 ***
## Condition.1        8 1.50e+11 1.87e+10   23.76 < 2e-16 ***
```

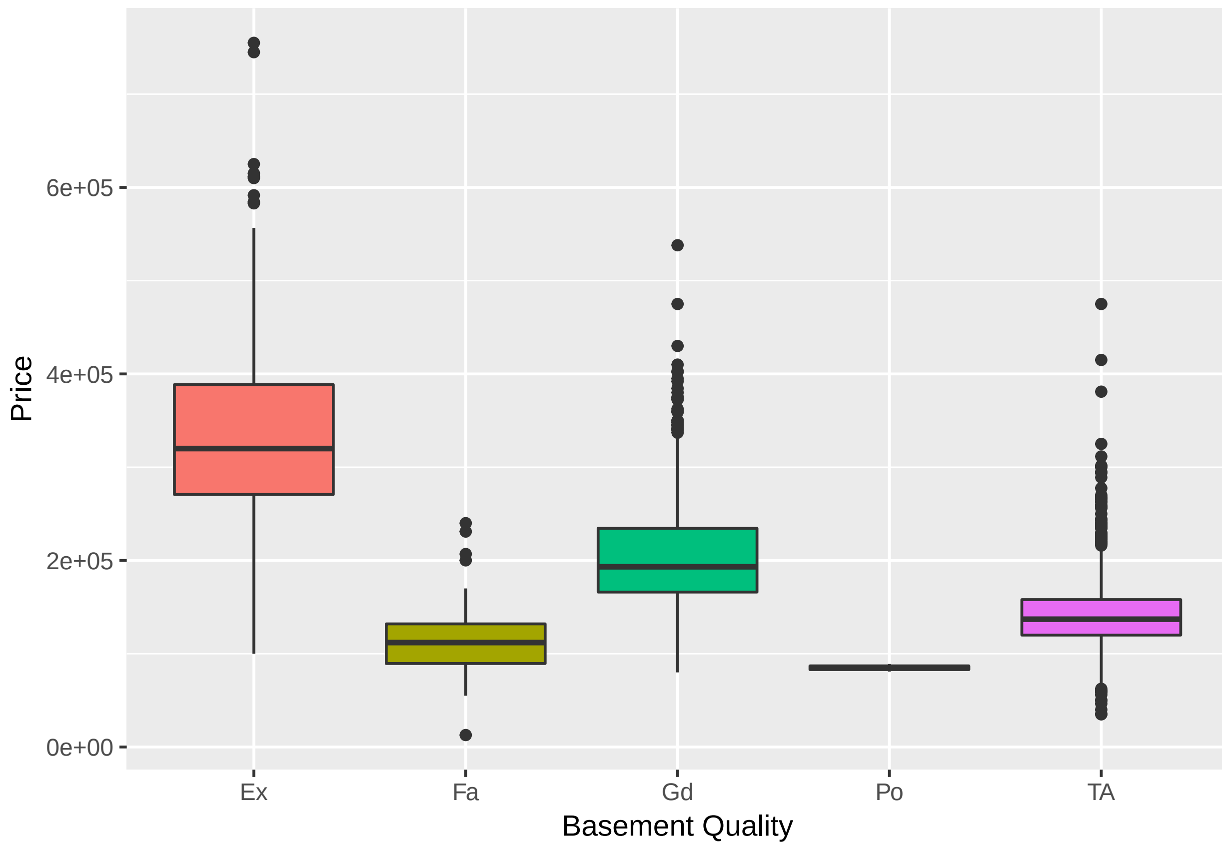
```
## Condition.2       7 1.05e+11 1.51e+10   19.09 < 2e-16 ***
## Bldg.Type         4 4.10e+11 1.03e+11  130.14 < 2e-16 ***
## House.Style       7 1.96e+11 2.80e+10   35.51 < 2e-16 ***
## Overall.Qual      1 2.02e+12 2.02e+12 2560.53 < 2e-16 ***
## Overall.Cond      1 8.93e+09 8.93e+09   11.33 0.00078 ***
## Roof.Style        5 9.41e+10 1.88e+10   23.88 < 2e-16 ***
## Roof.Matl         7 9.13e+10 1.30e+10   16.55 < 2e-16 ***
## Exterior.1st     13 1.29e+11 9.94e+09   12.60 < 2e-16 ***
## Exterior.2nd     14 3.36e+10 2.40e+09    3.04 0.00011 ***
## Mas.Vnr.Type      4 8.03e+10 2.01e+10   25.47 < 2e-16 ***
## Exter.Qual        3 2.47e+11 8.23e+10  104.45 < 2e-16 ***
## Exter.Cond        4 7.95e+09 1.99e+09    2.52 0.03935 *
## Foundation        4 3.43e+10 8.58e+09   10.89 9.4e-09 ***
## Bsmt.Qual         4 2.34e+11 5.85e+10   74.24 < 2e-16 ***
## Bsmt.Cond         4 3.11e+09 7.78e+08    0.99 0.41348
## Bsmt.Exposure     4 1.54e+11 3.86e+10   48.92 < 2e-16 ***
## BsmtFin.Type.1    5 7.69e+10 1.54e+10   19.51 < 2e-16 ***
## BsmtFin.Type.2    6 2.00e+09 3.33e+08    0.42 0.86462
## Heating           3 4.63e+09 1.54e+09    1.96 0.11829
## Heating.QC        4 2.12e+10 5.30e+09    6.73 2.2e-05 ***
## Central.Air       1 1.20e+10 1.20e+10   15.28 9.5e-05 ***
## Electrical        5 3.79e+09 7.58e+08    0.96 0.44037
## Bsmt.Full.Bath    1 4.86e+10 4.86e+10   61.65 6.1e-15 ***
## Bsmt.Half.Bath    1 8.08e+06 8.08e+06    0.01 0.91936
## Full.Bath         1 2.10e+11 2.10e+11  265.87 < 2e-16 ***
## Half.Bath         1 8.99e+10 8.99e+10  114.09 < 2e-16 ***
## Bedroom.AbvGr     1 2.60e+10 2.60e+10   32.99 1.0e-08 ***
## Kitchen.AbvGr     1 8.85e+08 8.85e+08    1.12 0.28949
## Kitchen.Qual      4 9.44e+10 2.36e+10   29.93 < 2e-16 ***
## TotRms.AbvGrd     1 1.45e+11 1.45e+11  183.56 < 2e-16 ***
## Functional        7 3.64e+09 5.21e+08    0.66 0.70572
## Fireplaces        1 7.52e+10 7.52e+10   95.38 < 2e-16 ***
## Garage.Type       5 1.54e+09 3.08e+08    0.39 0.85542
## Garage.Finish     2 4.74e+09 2.37e+09    3.01 0.04955 *
## Garage.Cars       1 7.85e+10 7.85e+10   99.56 < 2e-16 ***
## Garage.Qual       4 1.71e+10 4.27e+09    5.42 0.00024 ***
## Garage.Cond       4 7.67e+09 1.92e+09    2.43 0.04543 *
## Paved.Drive       2 9.26e+08 4.63e+08    0.59 0.55578
## Mo.Sold           1 2.03e+09 2.03e+09    2.57 0.10873
## Sale.Type         9 2.08e+10 2.31e+09    2.93 0.00183 **
## Sale.Condition    5 6.10e+09 1.22e+09    1.55 0.17187
## Residuals      2464 1.94e+12 7.88e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems that most of them have at least one significant component, the lack of time does not allow us to dive deeper in the EDA, thus we will jump to the analysis of a selected few variables.

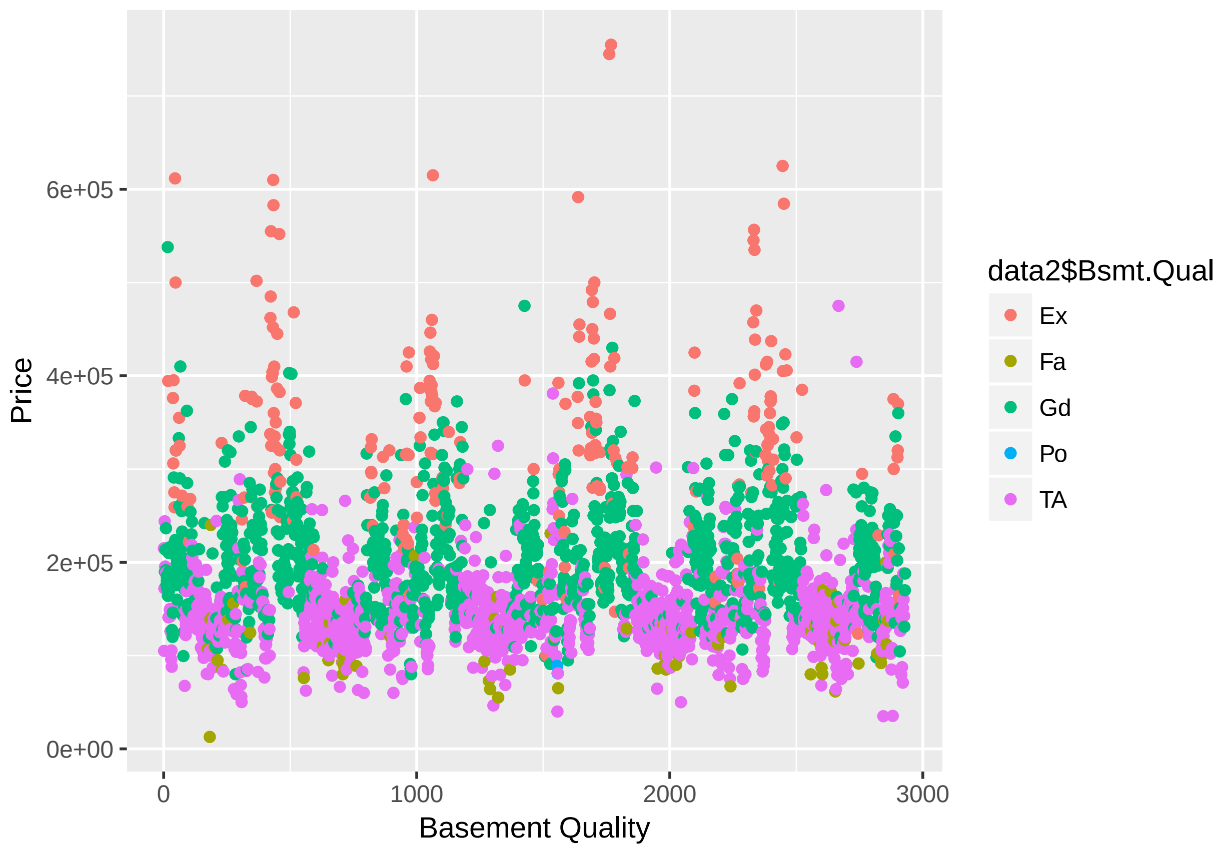
Analysis of the Categorical part

We decided to use the Tukey method on some of the possibly significant variables found above: Bsmt.Qual, Neighborhood, Condition.2, Roof.Matl, Overall.Qual, Bsmt.Exposure, Kitchen.AbvGr, Exterior1st, Sale.Type, Land.Contour. Note that all those variables were found significant from ANOVA and our boxplot analysis. This is why we will use the Tukey method in most of the following analysis.

Bsmt.Qual vs SalePrice

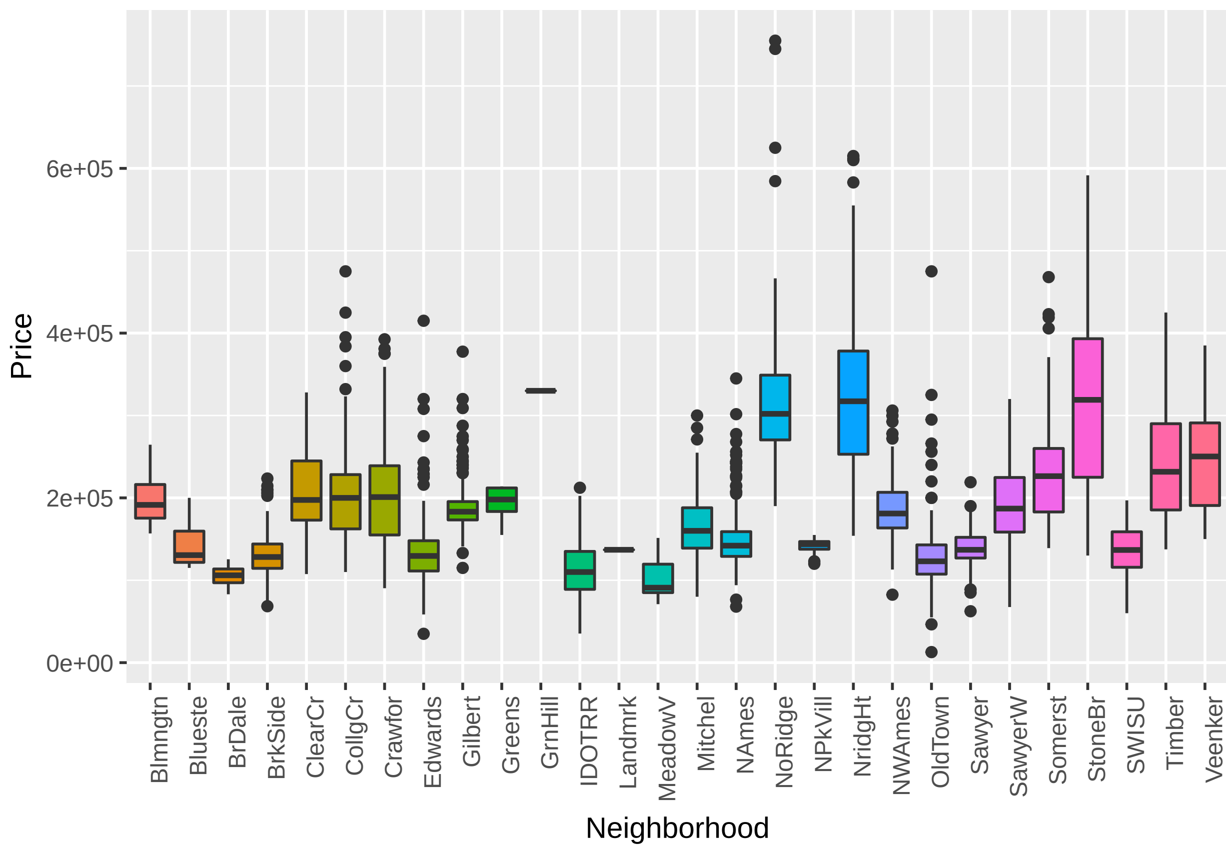



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $Bsmt.Qual
##           diff     lwr     upr  p adj
## Fa-Ex -219641 -239886 -199397 0.0000
## Gd-Ex -129913 -140559 -119267 0.0000
## Po-Ex -249009 -357658 -140360 0.0000
## TA-Ex -191392 -202025 -180759 0.0000
## Gd-Fa   89728   71383  108073 0.0000
## Po-Fa  -29368 -139040   80303 0.9493
## TA-Fa   28249    9912   46587 0.0003
## Po-Gd -119096 -227407  -10785 0.0227
## TA-Gd  -61479  -67785  -55172 0.0000
## TA-Po   57617  -50692  165927 0.5939
```



The boxplot show a distinct difference between Excellent and the rest of the levels of the predictors, with even Good being above the rest as well. Using the Tukey method of comparing the levels we see that Excellent indeed has a significantly different mean than the rest of the levels.

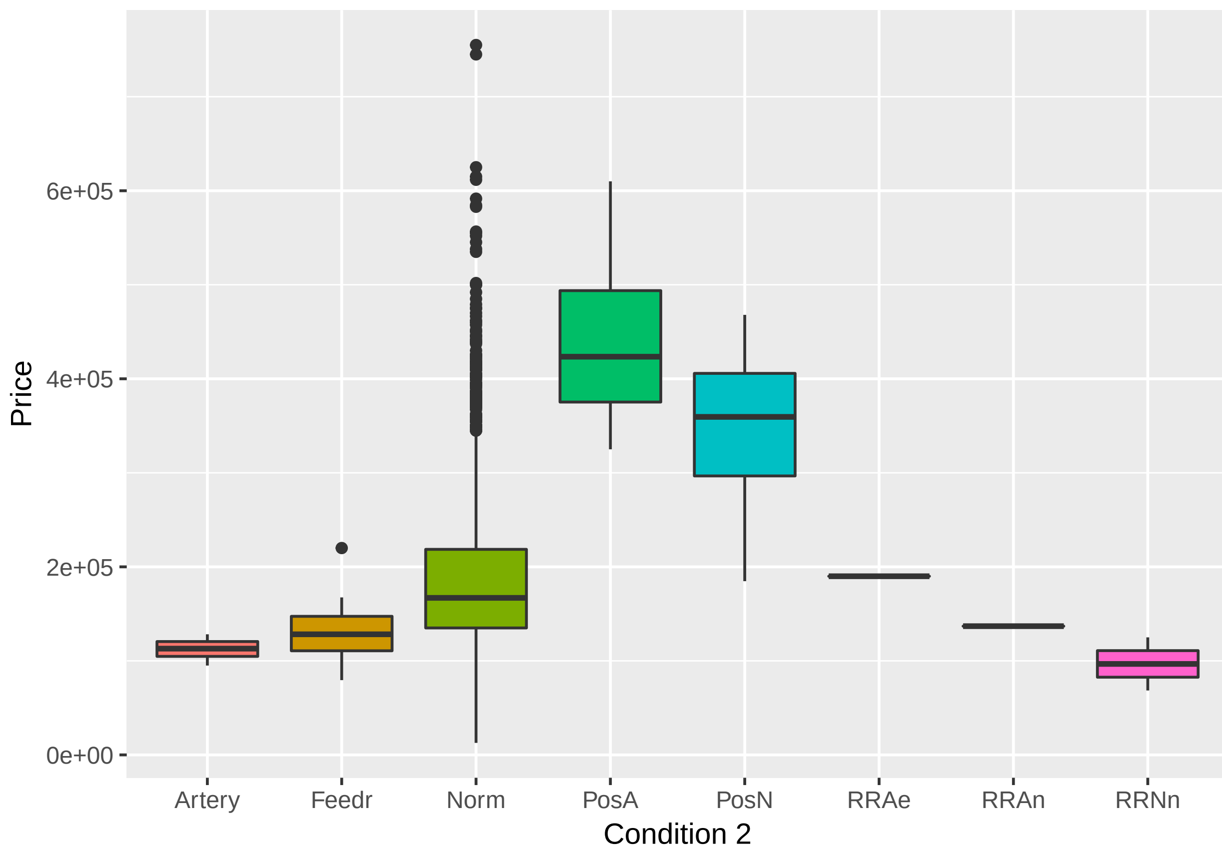
Neighborhood vs SalePrice




```
##                 Df   Sum Sq  Mean Sq F value Pr(>F)
## Neighborhood   27 9.45e+12 3.50e+11     124 <2e-16 ***
## Residuals    2655 7.48e+12 2.82e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As to be expected of a city with varying levels of wealth, some of the areas of the city seem to have higher average price for the sales of the houses. There are seeming large differences with the rest of the data for Green Hill, North Park Villa, North Ridge, and Stone Brook. There are also various other locations that possibly have slightly higher mean than the majority. Running an Anova shows that there is an significant evidence to support that there is infact adifference amongst the prices based of the location.

<u>Condition.2 vs SalePrice</u>

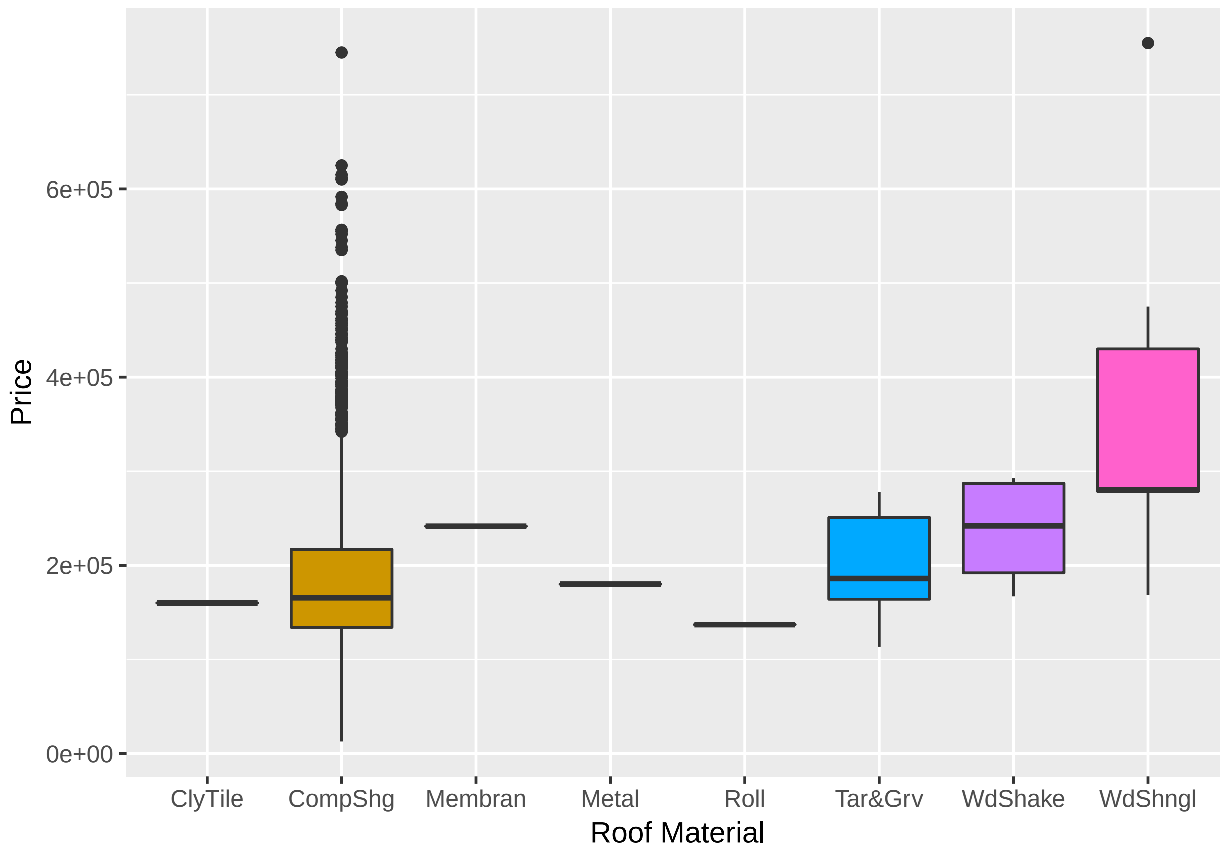



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $Condition.2
##                  diff     lwr     upr  p adj
## Feedr-Artery    20429 -117071  157930 0.9998
## Norm-Artery     73961  -45208  193129 0.5631
## PosA-Artery    333187  164785  501590 0.0000
## PosN-Artery    230625   62222  399028 0.0009
## RRAe-Artery     77687 -188581  343956 0.9874
## RRAn-Artery     24592 -241676  290861 1.0000
## RRNn-Artery    -15563 -221813  190688 1.0000
## Norm-Feedr      53532  -15374  122437 0.2635
## PosA-Feedr     312758  175258  450259 0.0000
## PosN-Feedr     210196   72695  347696 0.0001
## RRAe-Feedr      57258 -190624  305141 0.9970
## RRAn-Feedr       4163 -243719  252046 1.0000
## RRNn-Feedr     -35992 -217888  145904 0.9989
## PosA-Norm      259227  140058  378395 0.0000
## PosN-Norm      156664   37496  275833 0.0018
## RRAe-Norm        3727 -234476  241929 1.0000
## RRAn-Norm      -49368 -287571  188834 0.9985
## RRNn-Norm      -89523 -257990   78943 0.7433
## PosN-PosA     -102563 -270965   65840 0.5875
```

```
## RRAe-PosA    -255500 -521768   10768 0.0708
## RRAn-PosA    -308595 -574863  -42327 0.0105
## RRNn-PosA    -348750 -555001 -142499 0.0000
## RRAe-PosN    -152937 -419206  113331 0.6593
## RRAn-PosN    -206032 -472301   60236 0.2685
## RRNn-PosN    -246187 -452438  -39937 0.0072
## RRAn-RRAe     -53095 -389901  283711 0.9997
## RRNn-RRAe     -93250 -384932  198432 0.9786
## RRNn-RRAn     -40155 -331837  251527 0.9999
```

By just looking at the box plot we can already see the level indicating near or adjacent to positive offsite feature are much higher than the rest of the levels. Using Tukey we see that both of these values have significant evidence to show that their mean sales price for home with these two levels are higher than the rest.

Roof.Matl vs SalePrice

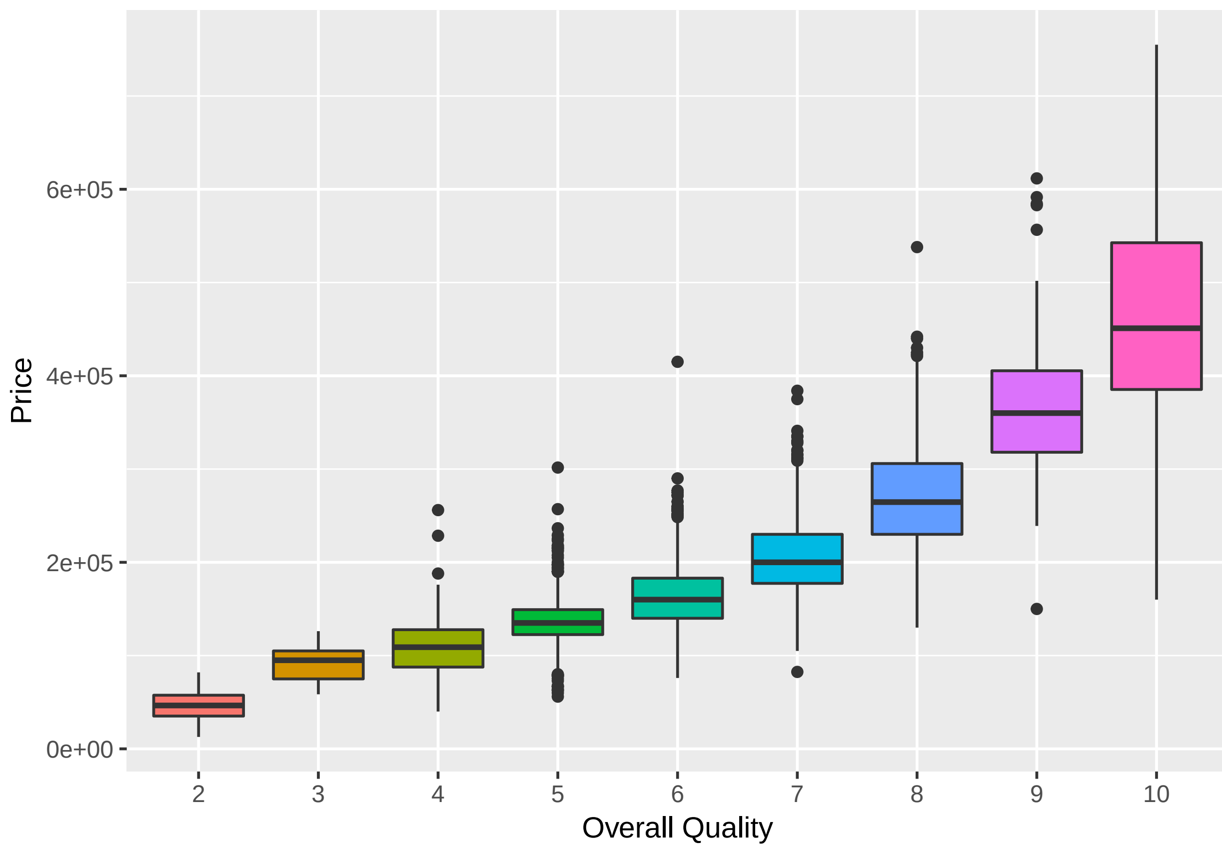



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $Roof.Matl
##                       diff     lwr    upr  p adj
## CompShg-ClyTile   25723 -213635 265080 1.0000
## Membran-ClyTile   81500 -256939 419939 0.9961
## Metal-ClyTile     20000 -318439 358439 1.0000
## Roll-ClyTile     -23000 -361439 315439 1.0000
```

```
## Tar&Grv-ClyTile   37955 -208295 284205 0.9998
## WdShake-ClyTile   78444 -173813 330702 0.9818
## WdShngl-ClyTile  214357  -41478 470193 0.1785
## Membran-CompShg   55777 -183580 295135 0.9968
## Metal-CompShg     -5723 -245080 233635 1.0000
## Roll-CompShg     -48723 -288080 190635 0.9987
## Tar&Grv-CompShg   12232  -45996  70460 0.9984
## WdShake-CompShg   52722  -27184 132628 0.4810
## WdShngl-CompShg  188635   98064 279206 0.0000
## Metal-Membran    -61500 -399939 276939 0.9994
## Roll-Membran    -104500 -442939 233939 0.9825
## Tar&Grv-Membran  -43545 -289795 202705 0.9995
## WdShake-Membran   -3056 -255313 249202 1.0000
## WdShngl-Membran  132857 -122978 388693 0.7653
## Roll-Metal       -43000 -381439 295439 0.9999
## Tar&Grv-Metal     17955 -228295 264205 1.0000
## WdShake-Metal     58444 -193813 310702 0.9969
## WdShngl-Metal    194357  -61478 450193 0.2914
## Tar&Grv-Roll      60955 -185295 307205 0.9954
## WdShake-Roll     101444 -150813 353702 0.9262
## WdShngl-Roll     237357  -18478 493193 0.0919
## WdShake-Tar&Grv   40490  -58162 139141 0.9183
## WdShngl-Tar&Grv  176402   68930 283875 0.0000
## WdShngl-WdShake  135913   15311 256515 0.0148
```

As we see from the box plot the levels of the predictor are much closer this time. Using the anova function in R we see that their is indeed a significant difference between the factor levels. Utilizing Tukey we see that specifically wind shingles does in fact have a small enough p-value showing that it has a different mean from all but wind shakes

Overall.Qual vs SalePrice

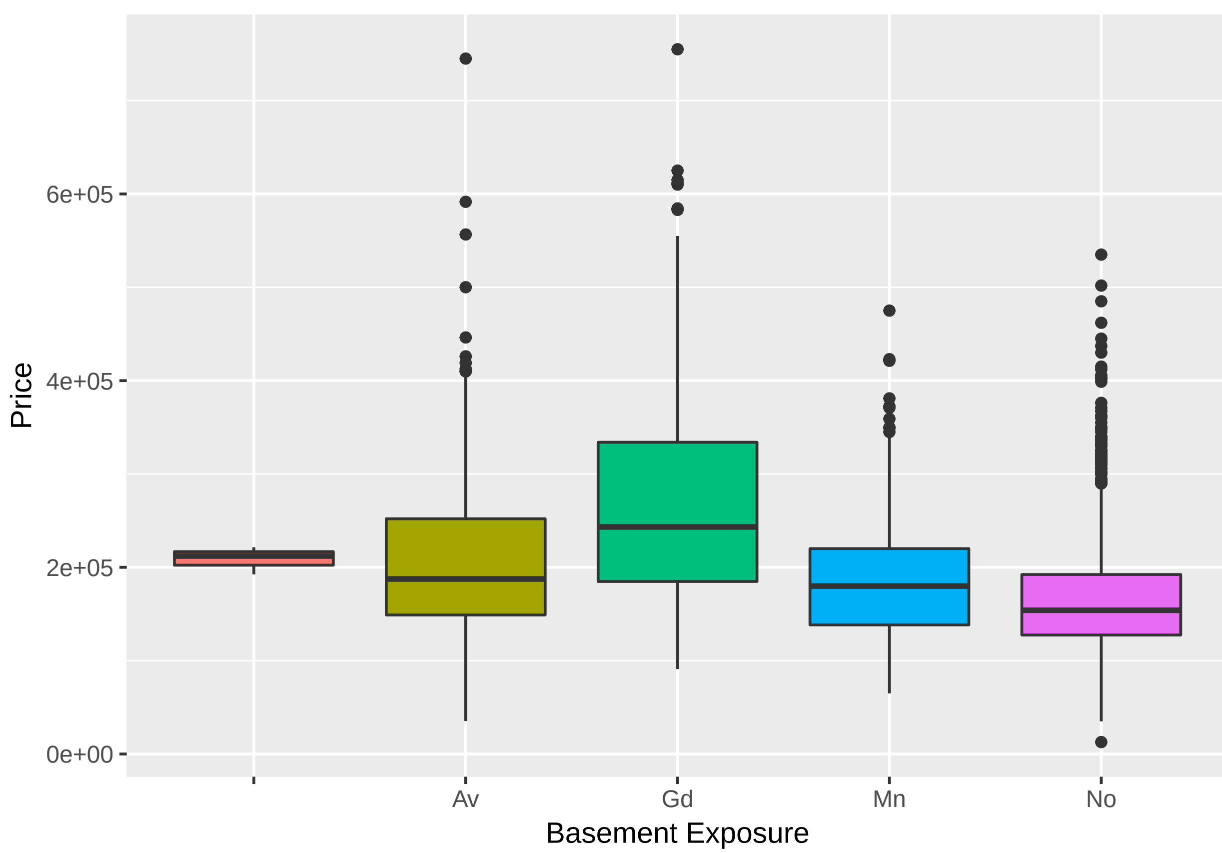



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $`as.ordered(Overall.Qual)`
##         diff    lwr    upr  p adj
## 3-2    44618 -17087 106322 0.3769
## 4-2    64382  11340 117424 0.0053
## 5-2    89925  37745 142105 0.0000
## 6-2   116625  64436 168815 0.0000
## 7-2   159079 106837 211320 0.0000
## 8-2   224576 172114 277038 0.0000
## 9-2   321680 268076 375283 0.0000
## 10-2 403067 345394 460741 0.0000
## 4-3    19764 -15265  54793 0.7142
## 5-3    45307  11597  79017 0.0010
## 6-3    72008  38283 105732 0.0000
## 7-3   114461  80656 148266 0.0000
## 8-3   179958 145813 214103 0.0000
## 9-3   277062 241188 312936 0.0000
## 10-3 358450 316739 400160 0.0000
## 5-4    25543  13610  37476 0.0000
## 6-4    52244  40271  64216 0.0000
## 7-4    94697  82498 106895 0.0000
## 8-4   160194 147083 173305 0.0000
```

```
## 9-4   257298 240181 274414 0.0000
## 10-4 338685 311376 365995 0.0000
## 6-5    26701  19443  33958 0.0000
## 7-5    69154  61529  76778 0.0000
## 8-5   134651 125639 143663 0.0000
## 9-5   231755 217532 245978 0.0000
## 10-5 313142 287547 338738 0.0000
## 7-6    42453  34766  50140 0.0000
## 8-6   107950  98885 117016 0.0000
## 9-6   205054 190798 219311 0.0000
## 10-6 286442 260827 312056 0.0000
## 8-7    65497  56136  74859 0.0000
## 9-7   162601 148154 177048 0.0000
## 10-7 243989 218268 269710 0.0000
## 9-8    97104  81879 112329 0.0000
## 10-8 178491 152326 204657 0.0000
## 10-9  81388  53003 109773 0.0000
```

As the boxplot shows, as well as what intuition should tell us, as the quality goes up so does the price of the house. In order to confirm this Tukey method was used to show mathematically where there are differences between levels. Tukey shows that increasing the Overall Quality of even just 1 unit will almost always have a significant impact on the Sale Price.

Bsmt.Exposure vs SalePrice

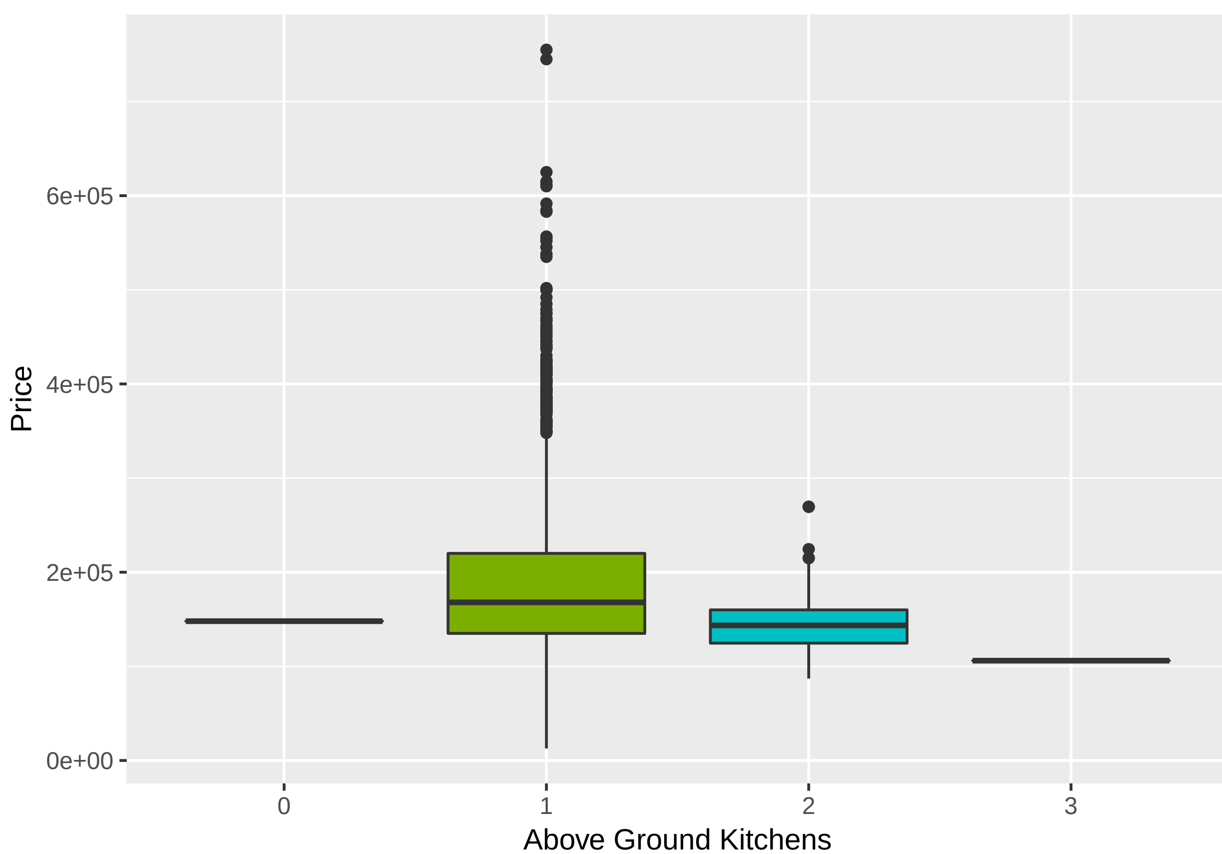



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = lm)
##
## $`as.ordered(Bsmt.Exposure)`
##            diff     lwr    upr  p adj
## Av-        2529 -112227 117285 1.0000
## Gd-       62097  -52853 177046 0.5792
## Mn-      -18893 -133958  96172 0.9917
## No-      -41244 -155666  73177 0.8626
## Gd-Av     59568   44016  75120 0.0000
## Mn-Av    -21422  -37808  -5036 0.0034
## No-Av    -43773  -54765 -32781 0.0000
## Mn-Gd    -80990  -98677 -63302 0.0000
## No-Gd   -103341 -116193 -90490 0.0000
## No-Mn    -22351  -36200  -8502 0.0001
```

The boxplot itself shows no drastic in the difference of one of the levels or the others. Based on the idea that the good is always preferable it would stand to reason good exposure should be the highest. As it turns out utilizing Tukey, this is actually the case with the Good exposure having a significantly higher mean.

<u>Kitchen.AbvGr vs SalePrice</u>

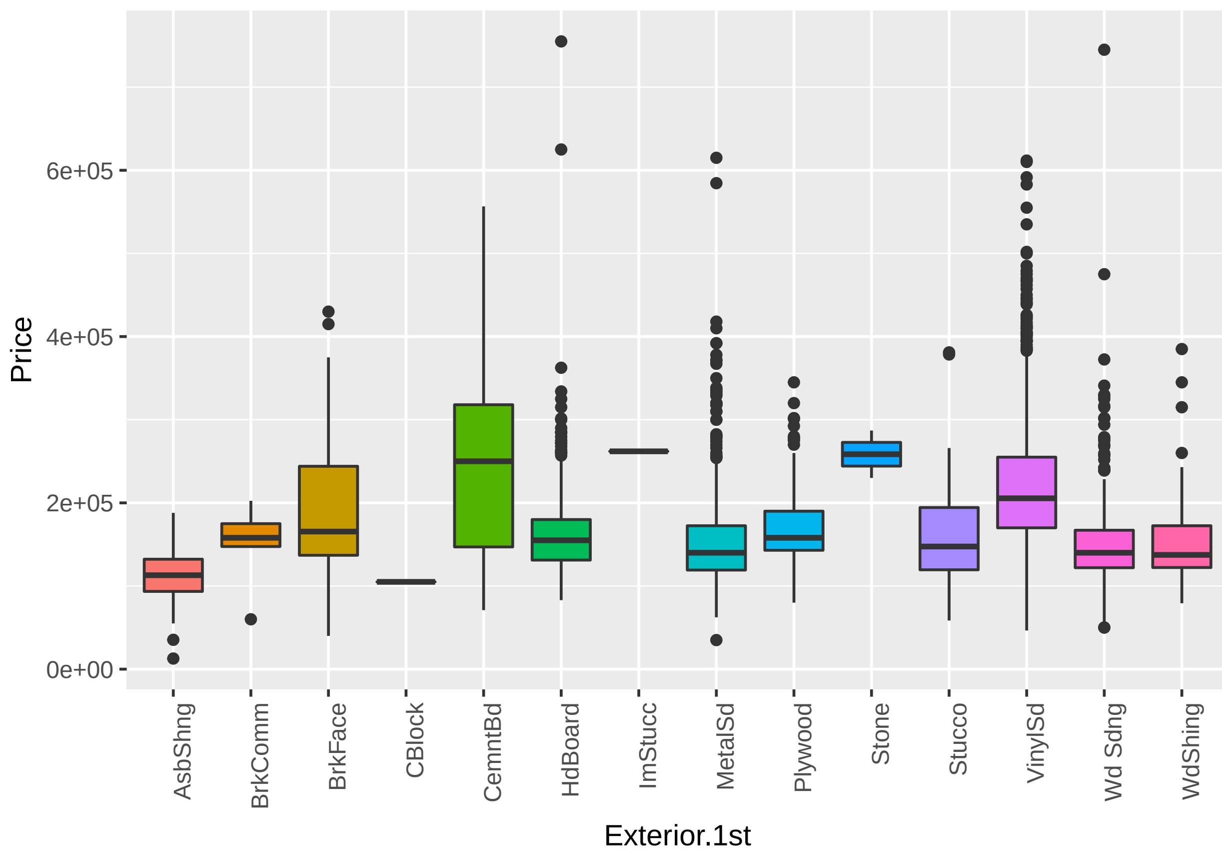



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $`as.factor(Kitchen.AbvGr)`
```

```
##       diff     lwr    upr  p adj
## 1-0  39549 -164185 243282 0.9593
## 2-0   1714 -203352 206780 1.0000
## 3-0 -42000 -330067 246067 0.9821
## 2-1 -37835  -61847 -13822 0.0003
## 3-1 -81549 -285282 122185 0.7325
## 3-2 -43714 -248780 161352 0.9471
```

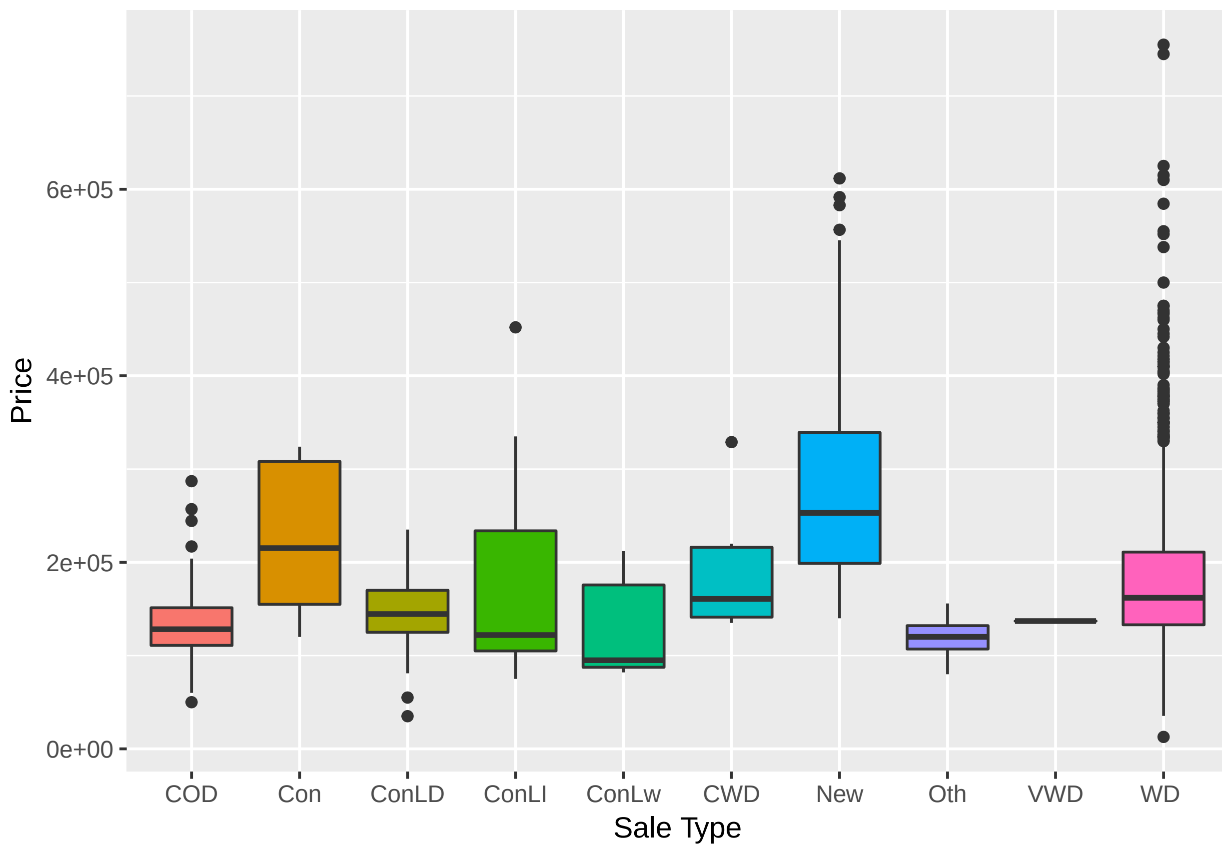
As the boxplot shows we see that aside from level: 1 having a slightly higher mean we can draw no conclusions without further analysis. The Anova of Price vs Kitchen we see that their is significant evidence of a difference and thus we use the Tukey method. The Tukey method returns that their is significant difference in the mean of levels 1 and 2, with 1 having a higher mean. As a note most houses have a singular kitchen and could possibly account for the large number of values observed for 1 or rather lack observations for the others.

Exterior.1st vs Price vs SalePrice



This predictor has a large number of levels that increase the complexity of analyzing both visually and with Tukey. We know for a fact that Anova shows significant evidence that their is a difference. However, simplifying this variable from sixteen levels to around 3-5 would greatly reduce this issue.

Sale.Type vs SalePrice

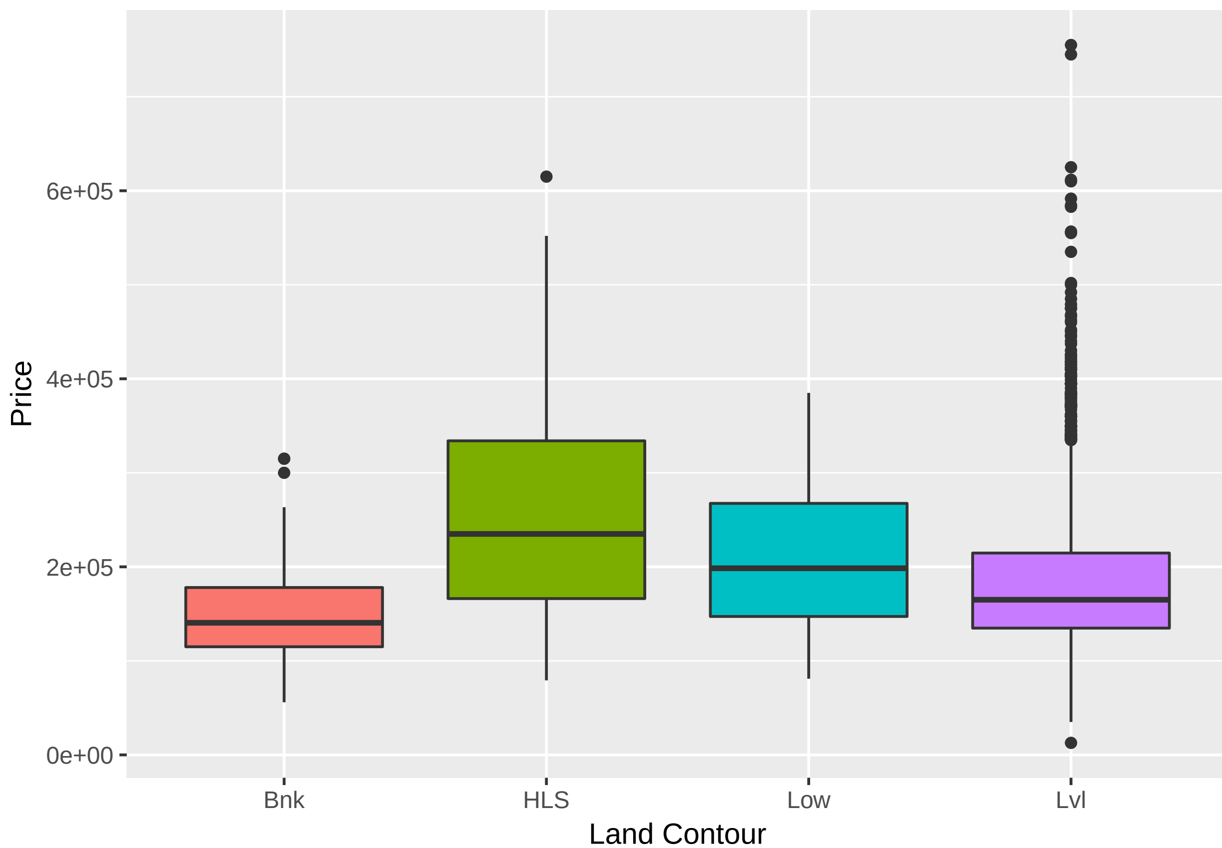



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $`as.factor(Sale.Type)`
##                    diff     lwr    upr  p adj
## Con-COD         90796  -16986 198578 0.1876
## ConLD-COD        8534  -53735  70803 1.0000
## ConLI-COD       54594  -32040 141227 0.6024
## ConLw-COD       -2864  -94975  89246 1.0000
## CWD-COD         50587  -21670 122845 0.4453
## New-COD        144972  115015 174928 0.0000
## Oth-COD        -14640 -134465 105186 1.0000
## VWD-COD          3350 -232180 238880 1.0000
## WD -COD         46380   20375  72385 0.0000
## ConLD-Con      -82262 -201380  36856 0.4662
## ConLI-Con      -36202 -169683  97278 0.9976
## ConLw-Con      -93660 -230759  43438 0.4827
## CWD-Con        -40209 -164839  84422 0.9910
## New-Con         54176  -51697 160049 0.8385
## Oth-Con       -105436 -262502  51631 0.5094
## VWD-Con        -87446 -343934 169042 0.9867
## WD -Con        -44416 -149240  60408 0.9439
## ConLI-ConLD     46060  -54327 146447 0.9102
## ConLw-ConLD    -11398 -116548  93752 1.0000
```

```
## CWD-ConLD      42053  -46226 130333 0.8892
## New-ConLD     136438   77535 195341 0.0000
## Oth-ConLD     -23174 -153290 106943 0.9999
## VWD-ConLD      -5184 -246113 235745 1.0000
## WD -ConLD      37846  -19149  94841 0.5258
## ConLw-ConLI   -57458 -178637  63721 0.8920
## CWD-ConLI      -4006 -110876 102864 1.0000
## New-ConLI      90378    6132 174624 0.0241
## Oth-ConLI     -69233 -212615  74148 0.8807
## VWD-ConLI     -51244 -299587 197100 0.9997
## WD -ConLI      -8214  -91138  74710 1.0000
## CWD-ConLw      53452  -57904 164808 0.8845
## New-ConLw     147836   57967 237705 0.0000
## Oth-ConLw     -11775 -158531 134980 1.0000
## VWD-ConLw       6214 -244093 256521 1.0000
## WD -ConLw      49244  -39386 137874 0.7608
## New-CWD        94384   25007 163762 0.0007
## Oth-CWD       -65227 -200408  69954 0.8812
## VWD-CWD       -47238 -290939 196464 0.9998
## WD -CWD        -4208  -71973  63558 1.0000
## Oth-New      -159611 -277722 -41501 0.0008
## VWD-New      -141622 -376284  93041 0.6616
## WD -New       -98592 -114973 -82210 0.0000
## VWD-Oth        17990 -243787 279767 1.0000
## WD -Oth        61020  -56151 178191 0.8237
## WD -VWD        43030 -191161 277221 0.9999
```

As we can see from the plot New and Contract 15% Down payment regular terms seem to be the only two levels away from the rest of the levels. According to the Anova test there is a difference between the levels and a pairwise test is needed. Using Tukey we can see that in fact both of these levels have a higher mean than the rest of levels. This could be due to multitude of reasons but does show that a lot of houses where bought new or on reagular terms.

<u>Land Countour vs SalePrice</u>



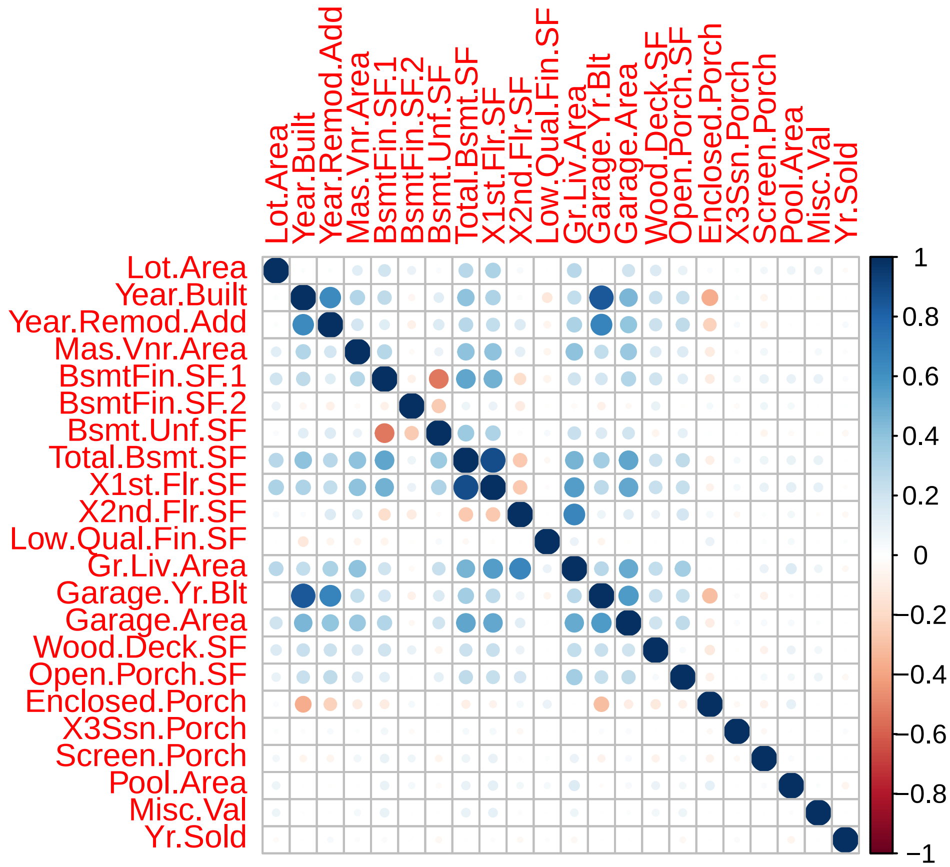

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm)
##
## $`as.factor(Land.Contour)`
##           diff    lwr    upr  p adj
## HLS-Bnk 108261  80963 135560 0.0000
## Low-Bnk  58491  24775  92207 0.0001
## Lvl-Bnk  33539  13165  53914 0.0001
## Low-HLS -49771 -82707 -16835 0.0006
## Lvl-HLS -74722 -93778 -55666 0.0000
## Lvl-Low -24952 -52423   2520 0.0905
```

Looking at the boxplot we see that their is very little difference in the sales price based off levels that dicernable with the eye. The Anova shows that their is indeed a difference amongst the levels. With Tukey we confirm that in fact HLS (Hill Side) is significantly higher in mean than the rest of the levels. From this we can see that homes residing on a hill side can fetch a higher sales price on average than other Homes.

Analysis of the continuous part of the data

EDA of the continous part of the data

Let's look at the correlation matrix:

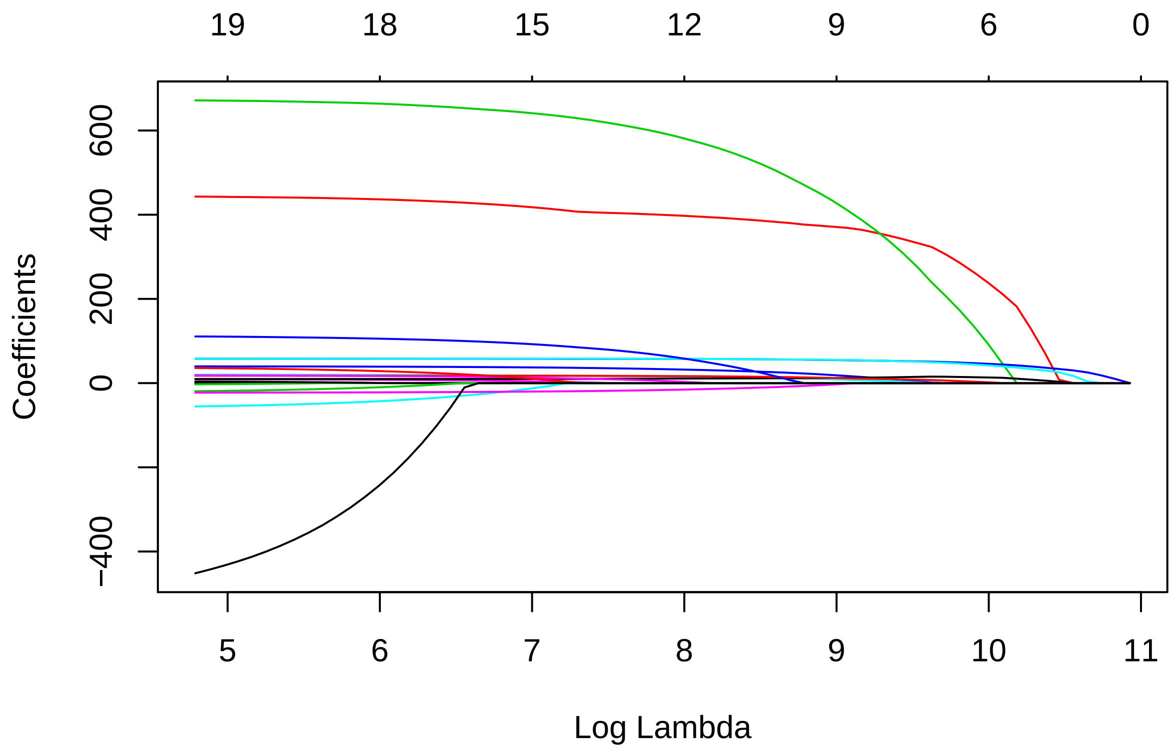


Only the correlation between Garage.Yr.Blt and Year.Built is a problem. Having those 2 variables corrolated is not a surprise. I'm going to drop the var Garage.Yr.Blt as I judge that Garare.Yr.Blt is just a consequence of Year.Built.

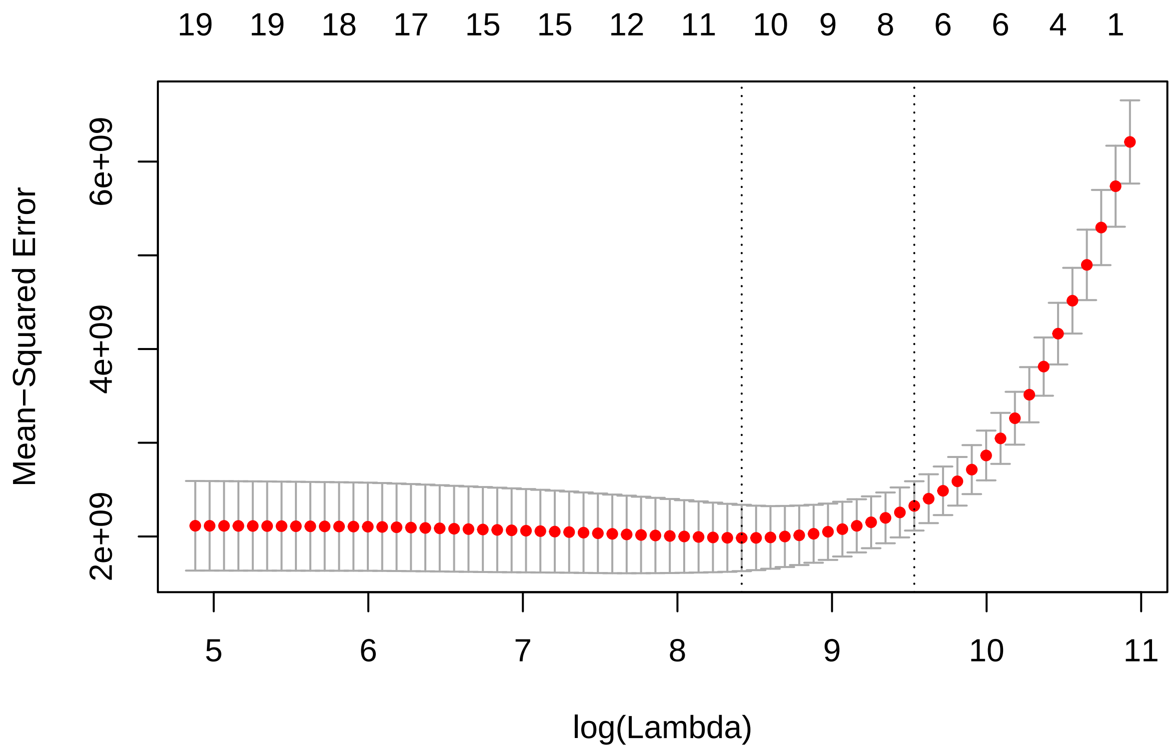
Analysis of the continous part of the data

I am going to run a LASSO regression on ly continous variables to determine which of them have a significant impact on the Sale Price.

```
x=model.matrix(data2$SalePrice~.,data=x_cont)[,-1]
y=data2$SalePrice

set.seed(11)
train=sample(1:nrow(x), round(nrow(x)/2))
test=(-train)
x_train = x[train,]; y_train = y[train]
x_test = x[test,]; y_test=y[test]


lasso.mod=glmnet(x_train, y_train, alpha=1)
plot(lasso.mod, xvar="lambda")
```




```
#this shows that some coeff will be 0, depending on the choice of the
#tuning parameter lambda

#let's use CV to find the optimal lamnda that minimizes test MSE
cv.out=cv.glmnet(x_train, y_train, alpha=1)
plot(cv.out)
```




```
bestlam=cv.out$lambda.min

#now that we found our optiamal lambda, let's use lasso to shrink the number of predictors
out=glmnet(x,y,alpha=1)
lasso.coef=predict(out,type="coefficients",s=bestlam)

#these are the variable with shrunk coefficients different than 0.
sig_var = data.frame(lasso.coef[which(lasso.coef!=0),])


#here is the sorted list of significant predictors
sig_var
```

```
##                 lasso.coef.which.lasso.coef....0....
## (Intercept)                                -1.800e+06
## Year.Built                                  3.717e+02
## Year.Remod.Add                              5.476e+02
## Mas.Vnr.Area                                3.527e+01
## BsmtFin.SF.1                                1.309e+01
## Total.Bsmt.SF                               2.661e+01
## X1st.Flr.SF                                 5.802e+00
## Gr.Liv.Area                                 6.190e+01
## Garage.Area                                 5.097e+01
## Wood.Deck.SF                                4.284e+00
## Screen.Porch                                1.468e+00
## Misc.Val                                   -1.972e+00
```

Thus, 11 out of the original 22 continous variables have an significant impact on the Sale Price. They are listed right above.

Conclusion

To conclude this assignment we will first discuss the issues with the methods used. First and foremost, We first tried to use LASSO but the number of categorical predictors barely dropped from 50 to 40 and the lack of interpretability of the resulting coefficients made us choose another route. Thus, for the categorical part of the data, we used visual analysis on the boxplots and an ANOVA to determine which variables might have a significant impact on Sale Price. Then we applied the Tukey method to those possibly significant variables to have a more in-depth view of which class or those predictors actually impacted the Sale Price. We see that certain neighborhoods affect the price in a significant way putting the cost of a house above houses in other areas. This is not an uncommon occurrence, nearly all cities have area that are considered "nicer" and thus it would make sense in this data set. High quality and excellent conditions of various aspects of a home can also affect the price at which it is sold for. We see that not only the overall quality of the home affecting the price but various other aspect such as the kitchen or basement. There is also the number of amenities that a home offers which can shift the price upward or downward when there is a lack of said amenities. This can include number of bathrooms, garage types, or even number of kitchens. It should be noted that the lack of data points for certain levels of the predictors have left the results for said predictors incomplete, specifically kitchen number. The style of home and lot style can also help increase the price of the home. It is not hard to see this as people do have preferences for certain styles of homes or how they are situated on the lot they are built. These various categorical variables are excellent indicators of an increase in mean price.

As for the continous variables of the data, LASSO was able to be applied and resulted in reducing the number of continous predictor by 50%! As it stands in the above list from LASSO, all of these variables are aspects of a home that make sense to affect the price. Using this model in tandem with the significant categorical levels.

Using our conclusions for the categorical part and continous part of the data, we could somewhat predict the Sale Price for the value of the significant predictors of a given property.