Read Data into R:

```
Ames <- read.csv("AmesHousing.csv")
```

Next I do a few preliminary checks to see the specifics of this data set:

Look at variable Names:

```
> names(Ames)
 [1] "MS.Zoning"       "Lot.Frontage"    "Lot.Area"        "Street"          "Alley"           "Lot.Shape"
 [7] "Land.Contour"    "Lot.Config"      "Land.Slope"      "Condition.1"     "Condition.2"     "Bldg.Type"
[13] "House.Style"     "Overall.Qual"    "Overall.Cond"    "Year.Built"      "Year.Remod.Add"  "Roof.Style"
[19] "Roof.Matl"       "Exterior.1st"    "Exterior.2nd"    "Mas.Vnr.Type"    "Mas.Vnr.Area"    "Exter.Qual"
[25] "Exter.Cond"      "Foundation"      "Bsmt.Qual"       "Bsmt.Cond"       "Bsmt.Exposure"   "BsmtFin.Type.1"
[31] "BsmtFin.SF.1"    "BsmtFin.Type.2"  "BsmtFin.SF.2"    "Bsmt.Unf.SF"     "Total.Bsmt.SF"   "Heating"
[37] "Heating.QC"      "Central.Air"     "Electrical"      "X1st.Flr.SF"     "X2nd.Flr.SF"     "Low.Qual.Fin.SF"
[43] "Gr.Liv.Area"     "Bsmt.Full.Bath"  "Bsmt.Half.Bath"  "Full.Bath"       "Half.Bath"       "Bedroom.AbvGr"
[49] "Kitchen.AbvGr"   "Kitchen.Qual"    "TotRms.AbvGrd"   "Functional"      "Fireplaces"      "Fireplace.Qu"
[55] "Garage.Type"     "Garage.Yr.Blt"   "Garage.Finish"   "Garage.Cars"     "Garage.Area"     "Garage.Qual"
[61] "Garage.Cond"     "Paved.Drive"     "Wood.Deck.SF"    "Open.Porch.SF"   "Enclosed.Porch"  "X3Ssn.Porch"
[67] "Screen.Porch"    "Pool.Area"       "Fence"           "Misc.Val"        "Yr.Sold"         "Sale.Type"
[73] "Sale.Condition"  "SalePrice"
```

Looking at the Variable Names. We can get rid of Garage Yr Blt due to multicollinearity. Meaning this variable is redundant since we also have the Yr The House

Look at the structure of the data set: **I normally do this when first looking at a new data set just so I can see how many observations I have and number of variables. I also like to see what variable type each variable is. As we can see from the print out, we have many categorical variables as well as continuous variables.**

```
> # Structure of Data:
> str(Ames)
'data.frame':   2925 obs. of  74 variables:
 $ MS.Zoning      : Factor w/ 5 levels "C (all)","FV",..: 4 3 4 4 4 4 4 4 4 4 ...
 $ Lot.Frontage   : int  141 80 81 93 74 78 41 43 39 60 ...
 $ Lot.Area       : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
 $ Street         : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
 $ Alley          : Factor w/ 3 levels "Grvl","None",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Lot.Shape      : Factor w/ 4 levels "IR1","IR2","IR3",..: 1 4 1 4 1 1 4 1 1 4 ...
 $ Land.Contour   : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 2 4 4 ...
 $ Lot.Config     : Factor w/ 5 levels "Corner","CulDSac",..: 1 5 1 1 5 5 5 5 5 5 ...
 $ Land.Slope     : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
 $ Condition.1    : Factor w/ 7 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 3 3 3 ...
 $ Condition.2    : Factor w/ 3 levels "Neg","Norm","Pos": 2 2 2 2 2 2 2 2 2 2 ...
 $ Bldg.Type      : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 5 5 5 1 ...
 $ House.Style    : Factor w/ 7 levels "1.5Fin","1.5Unf",..: 3 3 3 3 5 5 3 3 3 5 ...
 $ Overall.Qual   : int  6 5 6 7 5 6 8 8 8 7 ...
 $ Overall.Cond   : int  5 6 6 5 5 6 5 5 5 5 ...
 $ Year.Built     : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
 $ Year.Remod.Add : int  1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
 $ Roof.Style     : Factor w/ 5 levels "Flat","Gable",..: 4 2 4 4 2 2 2 2 2 2 ...
 $ Roof.Matl      : Factor w/ 3 levels "CompShg","Other",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Exterior.1st   : Factor w/ 9 levels "AsbShng","BrkFace",..: 2 8 9 2 8 8 3 4 3 8 ...
 $ Exterior.2nd   : Factor w/ 10 levels "AsbShng","BrkFace",..: 6 9 10 2 9 9 3 4 3 9 ...
 $ Mas.Vnr.Type   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 4 3 2 3 3 2 3 3 3 3 ...
 $ Mas.Vnr.Area   : int  112 0 108 0 0 20 0 0 0 0 ...
 $ Exter.Qual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 4 4 4 3 4 4 3 3 3 4 ...
 $ Exter.Cond     : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ Foundation     : Factor w/ 5 levels "BrkTil","CBlock",..: 2 2 2 2 4 4 4 4 4 4 ...
 $ Bsmt.Qual      : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 3 5 3 3 3 5 ...
 $ Bsmt.Cond      : Factor w/ 4 levels "Fa","Gd","None",..: 2 4 4 4 4 4 4 4 4 4 ...
 $ Bsmt.Exposure  : Factor w/ 5 levels "Av","Gd","Mn",..: 2 4 4 4 4 4 3 4 4 4 ...
 $ BsmtFin.Type.1 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 2 6 1 1 3 3 3 1 3 7 ...
 $ BsmtFin.SF.1   : int  639 468 923 1065 791 602 616 263 1180 0 ...
 $ BsmtFin.Type.2 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 7 4 7 7 7 7 7 7 7 7 ...
 $ BsmtFin.SF.2   : int  0 144 0 0 0 0 0 0 0 0 ...
 $ Bsmt.Unf.SF    : int  441 270 406 1045 137 324 722 1017 415 994 ...
 $ Total.Bsmt.SF  : int  1080 882 1329 2110 928 926 1338 1280 1595 994 ...
 $ Heating        : Factor w/ 2 levels "Gas","Other": 1 1 1 1 1 1 1 1 1 1 ...
 $ Heating.QC     : Factor w/ 4 levels "Ex","Fa","Gd",..: 2 4 4 1 3 1 1 1 1 3 ...
 $ Central.Air    : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Electrical     : Factor w/ 4 levels "FuseA","FuseF",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ X1st.Flr.SF    : int  1656 896 1329 2110 928 926 1338 1280 1616 1028 ...
 $ X2nd.Flr.SF    : int  0 0 0 0 701 678 0 0 0 776 ...
 $ Low.Qual.Fin.SF: int  0 0 0 0 0 0 0 0 0 0 ...
```

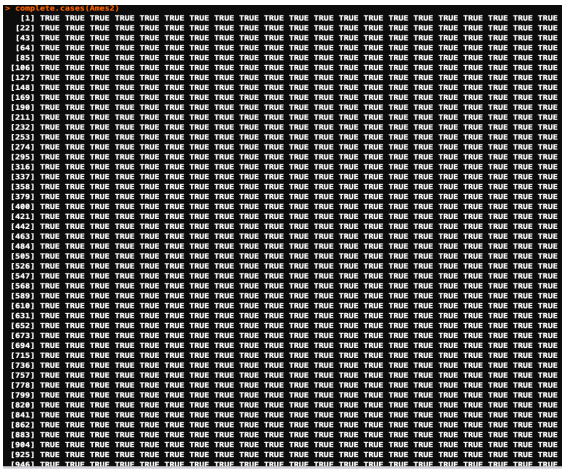
Number of NA Values: **Next, I like to see how many NA's values we have in our dataset. One way I like to do this is by printing a summary of the dataset:**
As we can see from the summary output, there are many variables that have at least one NA value. Some being Exterior.2nd, Mas.Vnr.Area, Bsmt.Qual, etc.

```
> summary(Ames)
   MS.Zoning      Lot.Frontage       Lot.Area      Street      Alley      Lot.Shape
 C (all):  25   Min.   : 21.00   Min.   :  1300   Grvl:  12   Grvl: 120   IR1: 977
 FV     : 139   1st Qu.: 58.00   1st Qu.:  7438   Pave:2913   None:2727   IR2:  76
 RH     :  27   Median : 68.00   Median :  9428               Pave:  78   IR3:  15
 RL     :2270   Mean   : 69.07   Mean   : 10108                           Reg:1857
 RM     : 461   3rd Qu.: 80.00   3rd Qu.: 11520
 NA's   :   3   Max.   :313.00   Max.   :215245
                NA's   :490
 Land.Contour   Lot.Config   Land.Slope Condition.1   Condition.2  Bldg.Type    House.Style
 Bnk: 114     Corner : 510   Gtl:2785   Artery:  92   Neg :  22   1Fam  :2420   1.5Fin: 314
 HLS: 120     CulDSac: 180   Mod: 124   Feedr : 163   Norm:2896   2fmCon:  62   1.5Unf:  19
 Low:  59     FR2    :  85   Sev:  16   Norm  :2519   Pos :   7   Duplex: 109   1Story:1478
 Lvl:2632     FR3    :  14              PosA  :  20               Twnhs : 101   2.5Unf:  24
              Inside :2136              PosN  :  38               TwnhsE: 233   2Story: 879
                                        RRAe  :  34                             SFoyer:  83
                                        RRAn  :  59                             SLvl  : 128
  Overall.Qual     Overall.Cond     Year.Built   Year.Remod.Add   Roof.Style    Roof.Matl
 Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  20   CompShg:2883
 1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1965   Gable  :2319   Other  :  26
 Median : 6.000   Median :5.000   Median :1973   Median :1993   Gambrel:  22   Tar&Grv:  16
 Mean   : 6.094   Mean   :5.565   Mean   :1971   Mean   :1984   Hip    : 548   WdShake:  54
 3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   Other  :  16   WdShngl: 199
 Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010
 Exterior.1st   Exterior.2nd   Mas.Vnr.Type   Mas.Vnr.Area    Exter.Qual Exter.Cond
 VinylSd:1026   VinylSd:1014   BrkCmn :  25   Min.   :   0.0   Ex: 104    Ex:  12
 Wood   : 697   Wood   : 478   BrkFace: 880   1st Qu.:   0.0   Fa:  34    Fa:  66
 MetalSd: 450   MetalSd: 447   None   :1750   Median :   0.0   Gd: 989    Gd: 299
 HdBoard: 442   HdBoard: 406   Stone  : 246   Mean   : 101.1   TA:1798    Po:   2
 CemntBd: 124   Plywood: 274   NA's   :  24   3rd Qu.: 164.0             TA:2546
 BrkFace:  98   (Other): 305                  Max.   :1600.0
 (Other):  88   NA's   :   1                  NA's   :23
```

```
  Foundation   Bsmt.Qual   Bsmt.Cond   Bsmt.Exposure BsmtFin.Type.1  BsmtFin.SF.1
 BrkTil: 310   Ex  : 255   Fa  : 108   Av  : 418     GLQ    :856    Min.   :   0.0
 CBlock:1244   Fa  :  89   Gd  : 125   Gd  : 281     Unf    :850    1st Qu.:   0.0
 Other :  16   Gd  :1219   None:  78   Mn  : 239     ALQ    :429    Median : 370.0
 PConc :1307   None:  78   TA  :2613   No  :1905     Rec    :288    Mean   : 439.3
 Slab  :  48   TA  :1283   NA's:   1   None:  78     BLQ    :269    3rd Qu.: 734.0
               NA's:   1               NA's:   4     (Other):232    Max.   :2288.0
                                                     NA's   :  1    NA's   :1
 BsmtFin.Type.2  BsmtFin.SF.2     Bsmt.Unf.SF     Total.Bsmt.SF   Heating     Heating.QC
 Unf    :2495   Min.   :   0.00   Min.   :   0.0   Min.   :   0   Gas  :2908   Ex:1492
 Rec    : 106   1st Qu.:   0.00   1st Qu.: 219.0   1st Qu.: 793   Other:  17   Fa:  94
 LwQ    :  89   Median :   0.00   Median : 464.5   Median : 990                Gd: 476
 None   :  78   Mean   :  49.81   Mean   : 559.2   Mean   :1048                TA: 863
 BLQ    :  68   3rd Qu.:   0.00   3rd Qu.: 801.2   3rd Qu.:1300
 (Other):  87   Max.   :1526.00   Max.   :2336.0   Max.   :3206
 NA's   :   2   NA's   :1         NA's   :1        NA's   :1
 Central.Air Electrical    X1st.Flr.SF    X2nd.Flr.SF     Low.Qual.Fin.SF     Gr.Liv.Area
 N: 194      FuseA: 187   Min.   : 334   Min.   :   0.0   Min.   :   0.000   Min.   : 334
 Y:2731      FuseF:  50   1st Qu.: 877   1st Qu.:   0.0   1st Qu.:   0.000   1st Qu.:1126
             FuseP:   8   Median :1084   Median :   0.0   Median :   0.000   Median :1442
             SBrkr:2678   Mean   :1157   Mean   : 335.2   Mean   :   4.685   Mean   :1496
             NA's :   2   3rd Qu.:1383   3rd Qu.: 703.0   3rd Qu.:   0.000   3rd Qu.:1742
                          Max.   :3820   Max.   :2065.0   Max.   :1064.000   Max.   :4476

 Bsmt.Full.Bath   Bsmt.Half.Bath     Full.Bath       Half.Bath      Bedroom.AbvGr
 Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
 Median :0.0000   Median :0.0000   Median :2.000   Median :0.0000   Median :3.000
 Mean   :0.4307   Mean   :0.0609   Mean   :1.566   Mean   :0.3791   Mean   :2.855
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :3.0000   Max.   :2.0000   Max.   :4.000   Max.   :2.0000   Max.   :8.000
 NA's   :2        NA's   :2
```

```
 Kitchen.AbvGr   Kitchen.Qual TotRms.AbvGrd    Functional    Fireplaces     Fireplace.Qu
 Min.   :0.000   Ex: 202      Min.   : 2.000   Maj1:  19   Min.   :0.0000   Ex :  43
 1st Qu.:1.000   Fa:  70      1st Qu.: 5.000   Maj2:   9   1st Qu.:0.0000   Fa :  75
 Median :1.000   Gd:1160      Median : 6.000   Min1:  65   Median :1.0000   Gd : 740
 Mean   :1.044   TA:1493      Mean   : 6.438   Min2:  70   Mean   :0.5979   None:1421
 3rd Qu.:1.000                3rd Qu.: 7.000   Mod :  35   3rd Qu.:1.0000   Po :  46
 Max.   :3.000                Max.   :14.000   Typ :2724   Max.   :4.0000   TA : 600
                                               NA's:   3
  Garage.Type   Garage.Yr.Blt  Garage.Finish  Garage.Cars     Garage.Area   Garage.Qual
 2Types :  23   Min.   :1895   Fin : 725     Min.   :0.000   Min.   :   0   Fa  : 127
 Attchd :1728   1st Qu.:1960   None: 157     1st Qu.:1.000   1st Qu.: 320   Gd  :  27
 Basment:  36   Median :1979   RFn : 812     Median :2.000   Median : 480   None: 158
 BuiltIn: 185   Mean   :1978   Unf :1229     Mean   :1.766   Mean   : 472   TA  :2612
 CarPort:  15   3rd Qu.:2002   NA's:   2     3rd Qu.:2.000   3rd Qu.: 576   NA's:   1
 Detchd : 781   Max.   :2207                 Max.   :5.000   Max.   :1488
 None   : 157   NA's   :159                  NA's   :1       NA's   :1
 Garage.Cond Paved.Drive  Wood.Deck.SF     Open.Porch.SF    Enclosed.Porch
 Fa  :  73   N: 214      Min.   :   0.00   Min.   :  0.00   Min.   :   0.00
 Gd  :  18   P:  62      1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00
 None: 158   Y:2649      Median :   0.00   Median : 27.00   Median :   0.00
 Po  :  13               Mean   :  93.58   Mean   : 47.21   Mean   :  23.05
 TA  :2662               3rd Qu.: 168.00   3rd Qu.: 70.00   3rd Qu.:   0.00
 NA's:   1               Max.   :1424.00   Max.   :742.00   Max.   :1012.00

  X3Ssn.Porch      Screen.Porch      Pool.Area         Fence         Misc.Val
 Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   GdPrv: 118   Min.   :    0.00
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   GdWo : 112   1st Qu.:    0.00
 Median :  0.000   Median :  0.00   Median :  0.000   MnPrv: 330   Median :    0.00
 Mean   :  2.597   Mean   : 16.03   Mean   :  2.083   MnWw :  12   Mean   :   44.91
 3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000   None :2353   3rd Qu.:    0.00
 Max.   :508.000   Max.   :576.00   Max.   :800.000                Max.   :15500.00
```

```
    Yr.Sold     Sale.Type  Sale.Condition   SalePrice
 Min.   :2006   COD:  87   Abnorml: 188   Min.   : 34900
 1st Qu.:2007   Con:  48   AdjLand:  12   1st Qu.:129500
 Median :2008   New: 236   Alloca :  24   Median :160000
 Mean   :2008   Oth:   7   Family :  46   Mean   :180916
 3rd Qu.:2009   WD :  13   Normal :2413   3rd Qu.:213500
 Max.   :2010   WD :2534   Partial: 242   Max.   :755000
```



The Above print out is from the code complete.cases(). What this does is it allows me to see if I have removed all of the missing values after running the na.omit() function. If all of the values are set to “TRUE” then that means that the na.omit() function worked and now my data set does not have any missing values.

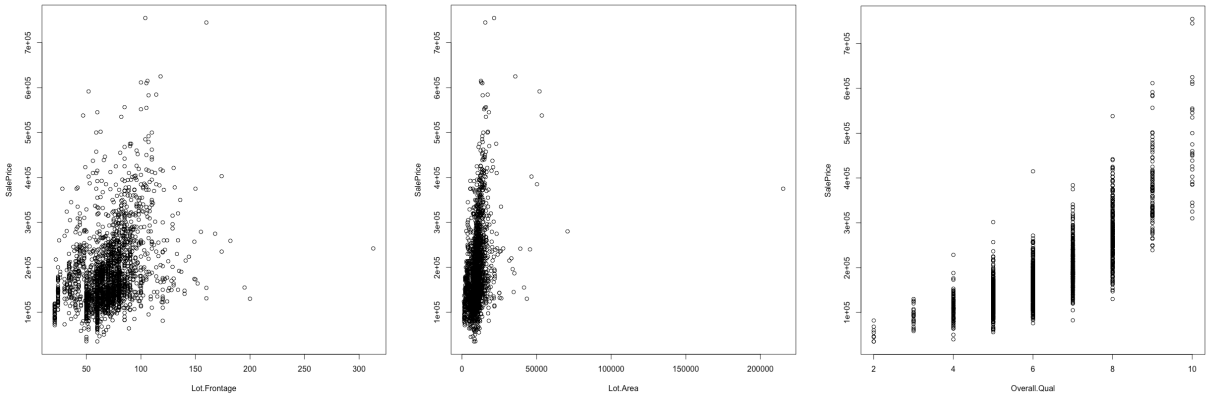
Sale Price vs Independent Variable Plots:

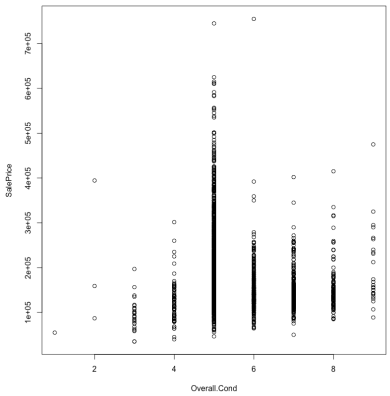
I will now plot SalePrice vs all of the variables in our num.ames data frame and see if there is relationship between these two variables. This will give me a good indication which of these variables would be good to keep in our model.

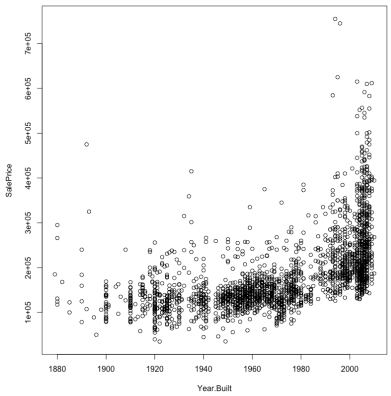
- **Looking at the first model is SalePrice vs Lot.Frontage. Doesn’t look like there is a relationship between these two variables as the variance is not consistent. There is a huge cluster at the bottom left of the plot then the variance starts to spread out. However, there could be a non-linear relationship between these two variables. For now, we will keep it in our model.**
- **Next plot is SalePrice vs Lot.Area. There is no relationship between these variables. Therefore, we will keep this variable out of our model.**
- **Next plot is SalePrice vs Overall. It looks like there could be a linear relationship between these two variables even though the variance isn’t very consistent. Yet, we will keep this variable in our model for now.**
- **Next plot is SalePrice vs Overall.Cond, there could be a relationship between these variables. It looks like the plot is almost linear but then the variance shoots up between 4-6 for Overall.Cond. For now will keep this in our model.**

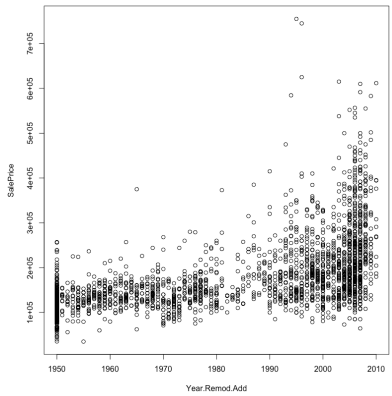
- **Next plot is SalePrice vs Year.Built, we can see that there is some sort of positive relationship happening between SalePrice and Year.Built althought the variance is not very consistent. We will keep this in our model for now and possibly could take the polynomial of this variable to try and make that variance tighter.**
- **Next plot is SalePrice vs Years.Remod.Add, we can see that there is a positive relationship between these two variables so we will keep it in our model.**
- **Next plot is SalePrice vs Mas.Vnr.Area, we can see there there is no relationship between these two variables. Therefore, we will keep this variable out of our model.**
- **Next plot is SalePrice vs BsmtFin.SF.1. There is no relationship between these two variables. Therefore, we will remove it from our model.**
- **Next plot is SalePrice vs BsmtFin.SF.2, there is no relationship between these two variables. Therefore, we will remove it from our model.**
- **Next plot is SalePrice vs Bsmt.Unf.SF, there is no relationship between these two variables. Therefore, we will remove it from our model.**
- **Next plot is SalePrice vs Total.Bsmt.SF, we can see that there is a positive relationship between these two variables. The variance starts to spread out as square footage increases. We could maybe do a polynomial on this variable to help tighten up the variance. We will keep this in our model for now. It's also important to note that we can see an example of "multicollinearity". Since we have Total Square footage of basement, it's redundant to have the previous two variables BsmFin.SF.1 and BsmFin.SF.2.**
- **Next plot is SalePrice vs X1st.Fir.SF, we can see here that there is a positive relationship between these two variables but the variance isn't consistent. We will keep this in our model for now.**
- **Next plot is SalePrice vs X2nd.Flr.SF, There seems to be a lot of outliers for this variable and then a positive relationship between the two. We will keep this in our model for now but will have to address the outliers later.**
- **Next plot is SalePrice vs Low.Qual.Fin.SF, there is no relationship between these two variables.**
- **Next plot is SalePrice vs Gr.Live.Area, there is a positive relationship between these two variables. We will keep in model for now.**
- **Next Plot is SalePrice vs Bsmt.Full.Bath, no relationship here. We will leave this out.**

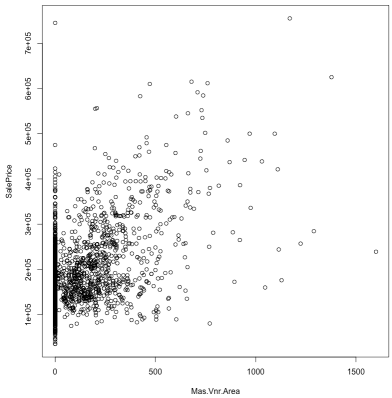
- **Next plot is SalePrice vs Bsmt.Half.Bath, no relationship here, we will leave this out of our model.**
- **Looks like we could have a positive relationship between SalePrice and Full.Bath, but the variance is not consistent. We will keep this in our model for now.**
- **No relationship between SalePrice and Half.Bath**
- **No relationship between SalePrice and Bedroom.AbvGr.**
- **No relationship between SalePrice and Kitchen.AbvGr.**
- **Loooks like there could be a relationship between SalePrice and Total Rooms above ground. We will keep this in our model for now.**
- **No relationship between SalePrice and Fireplaces.**
- **Looks like there could be a positive relationship between SalePrice and Year the garage was built, we will keep this in our model for now.**
- **No relationship between SalePrice and Garage.Cars**
- **Looks like there is a relationship between SalePrice and the Area of the garage. Variance isn't consistent, we will keep this in our model for now.**
- **No relationship between SalePrice and Wood.Deck.SF.**
- **Looks like there could be a negative relationship between SalePrice and Open.Proch.SF, we will keep in model for now.**
- **No relationship between SalePrice and Enclosed.porch**
- **No relationship between SalePrice and X3Ssn.Porch.**
- **No relationship between SalePrice and Screen.Porch.**
- **No relationship between SalePrice and Pool.Area.**

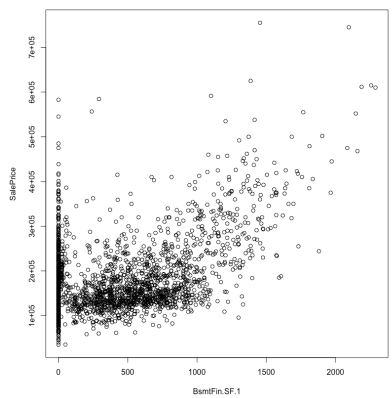


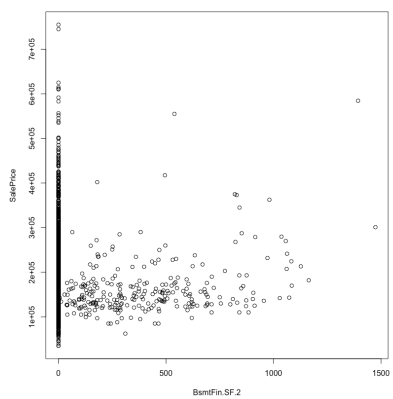

SalePrice
Overall.Cond

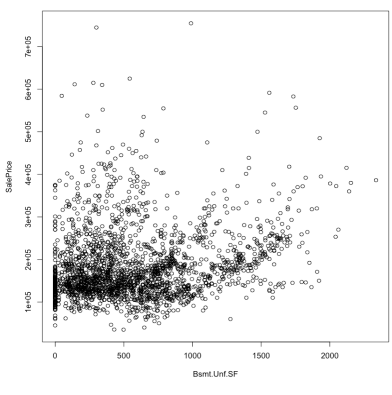

SalePrice
Year.Built

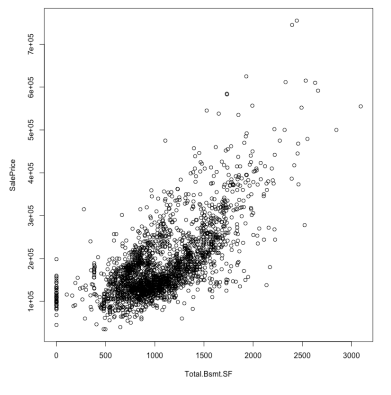

SalePrice
Year.Remod.Add

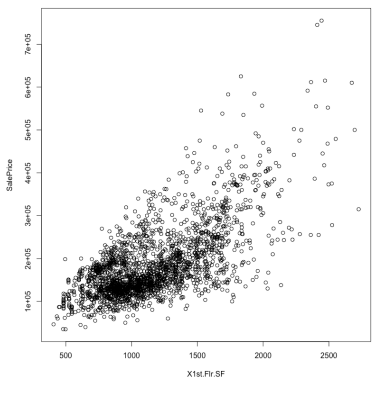

SalePrice
Mas.Vnr.Area

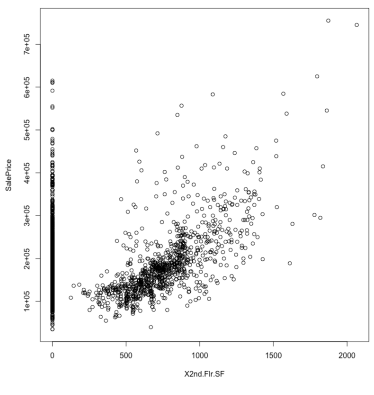

SalePrice
BsmtFin.SF.1

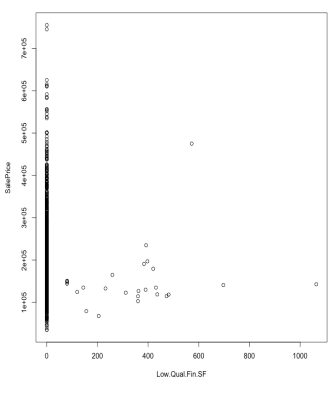

SalePrice
BsmtFin.SF.2

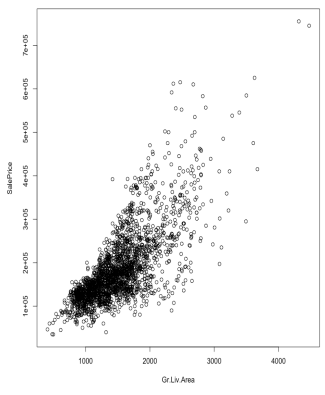

SalePrice
Bsmt.Unf.SF

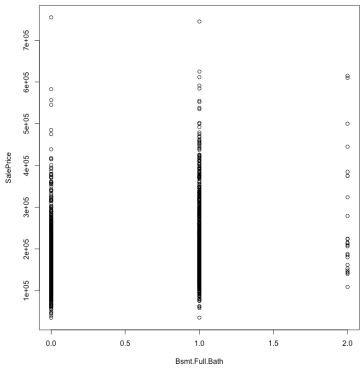

SalePrice
Total.Bsmt.SF

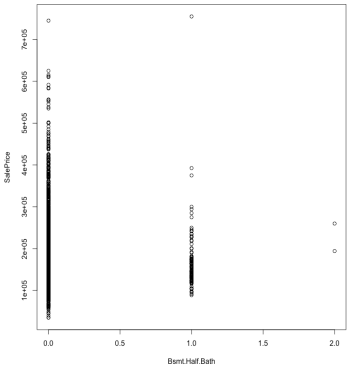

SalePrice
X1st.Flr.SF

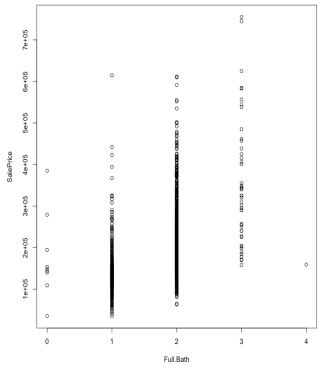

SalePrice
X2nd.Flr.SF

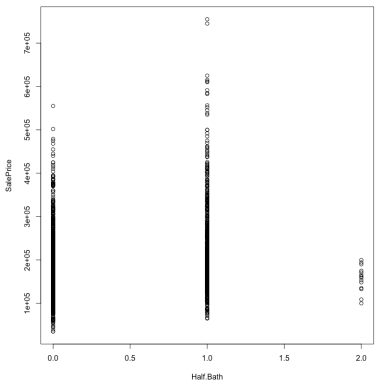

SalePrice
Low.Qual.Fin.SF

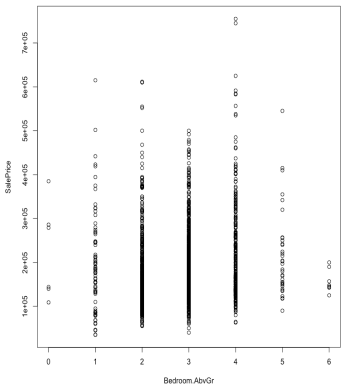

SalePrice
Gr.Liv.Area

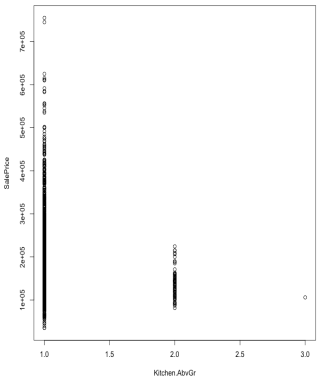

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0.0
0.5
1.0
1.5
2.0
Bsmt.Full.Bath

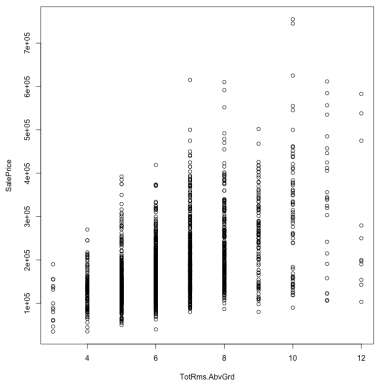

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0.0
0.5
1.0
1.5
2.0
Bsmt.Half.Bath

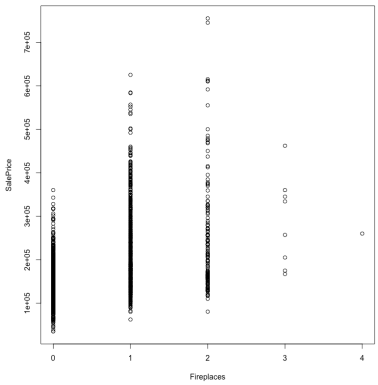

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
1
2
3
4
Full.Bath

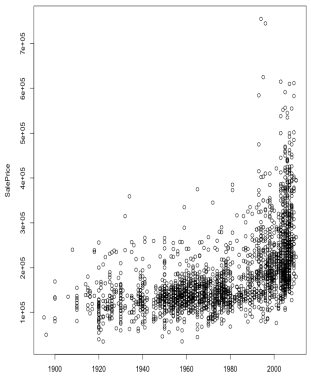

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0.0
0.5
1.0
1.5
2.0
Half.Bath

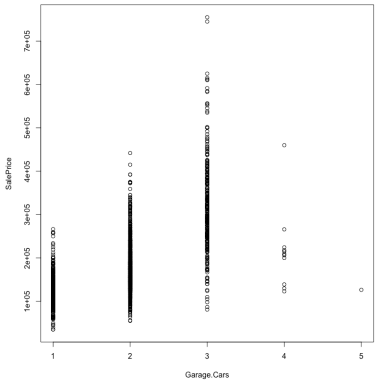

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
1
2
3
4
5
6
Bedroom.AbvGr

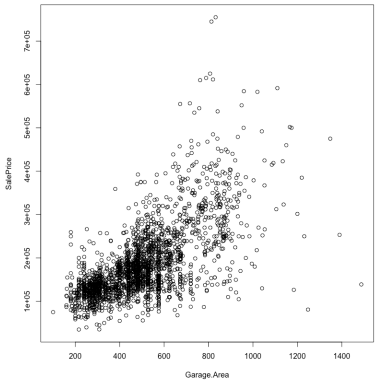

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
1.0
1.5
2.0
2.5
3.0
Kitchen.AbvGr

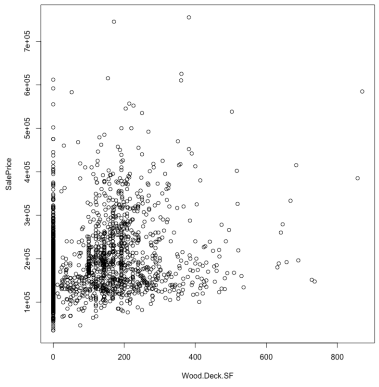

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
4
6
8
10
12
TotRms.AbvGrd

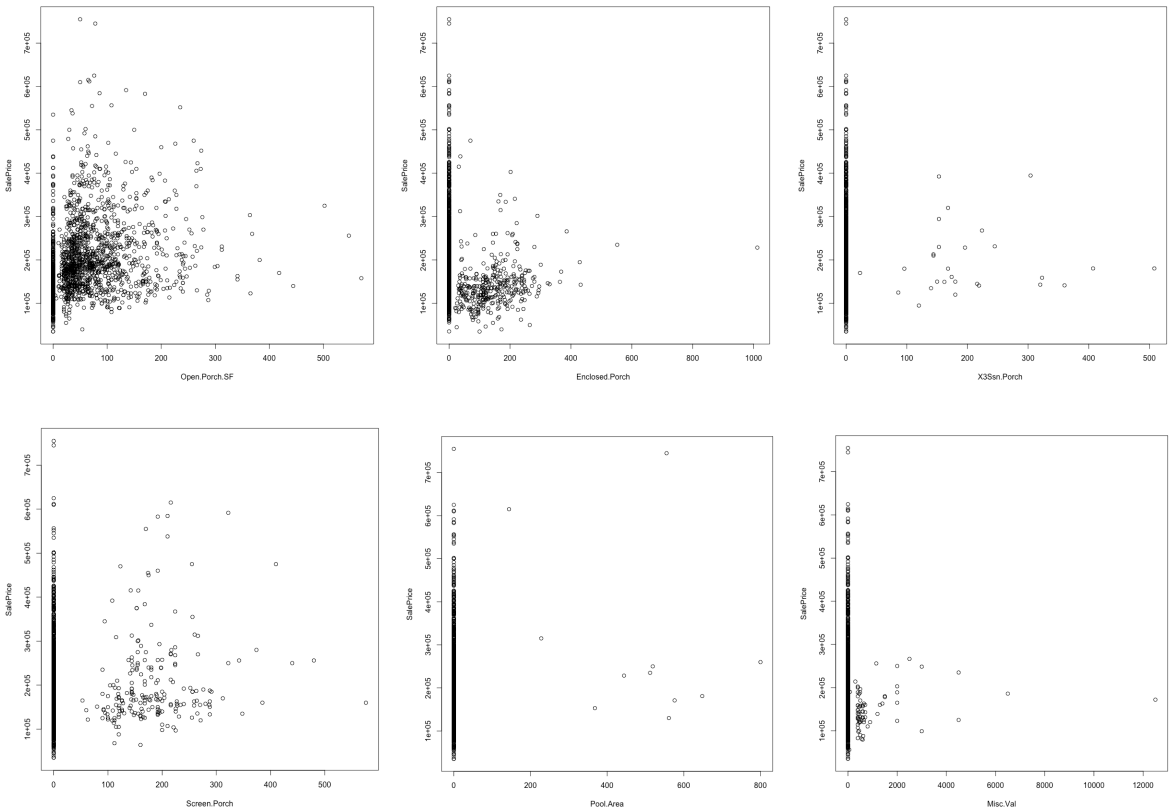

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
1
2
3
4
Fireplaces


SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
1900
1920
1940
1960
1980
2000
Garage.Yr.Blt


SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
1
2
3
4
5
Garage.Cars

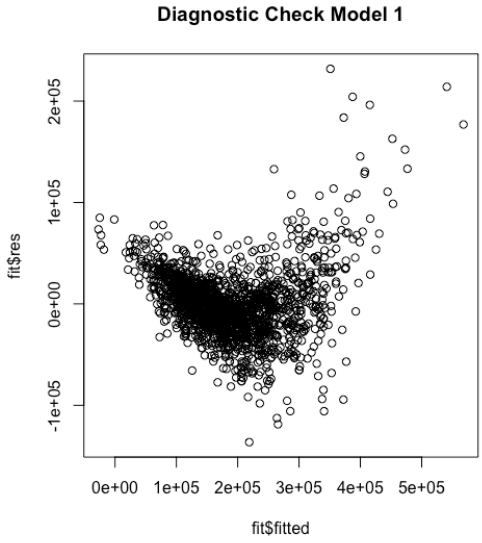

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
200
400
600
800
1000
1200
1400
Garage.Area

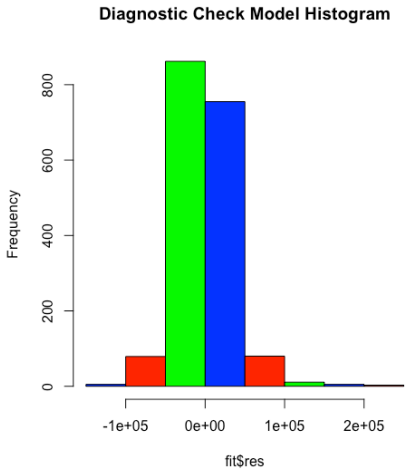

SalePrice
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
200
400
600
800
Wood.Deck.SF

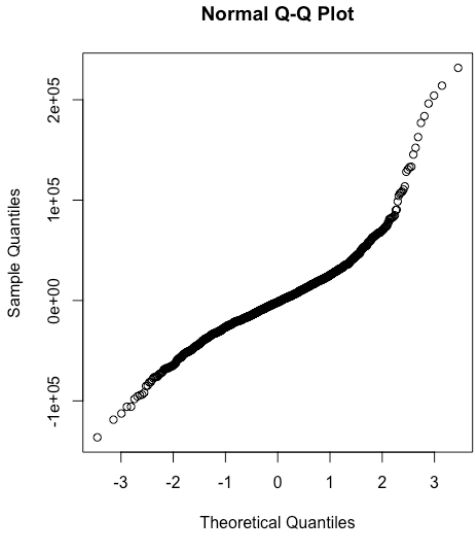


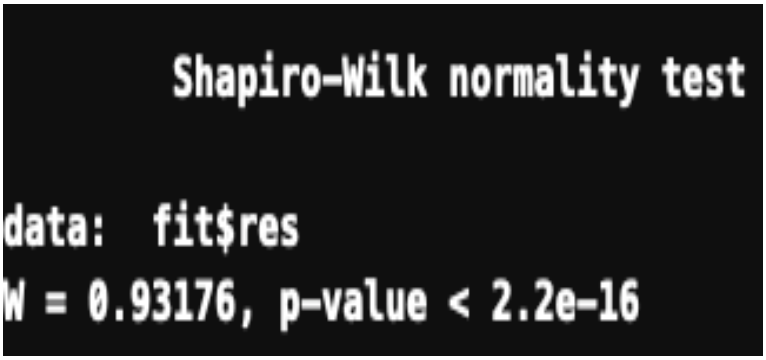
Diagnostic Checks:

I conducted the following checks on my first model after looking at the different plots to see which variables I would want to include. The model is ok but need to do something about TotRms.AbvGrd and Garage.Yr.Blt and Open.Porch.SF as these are not statistically significant.



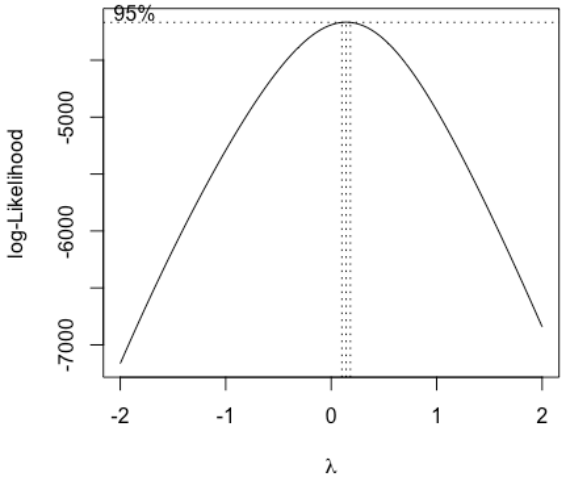






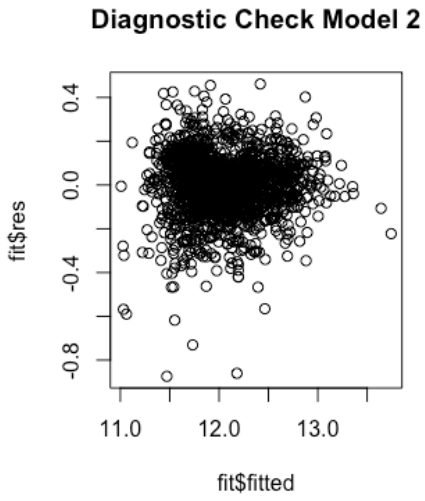
To address both of these Issues. I ran a boxcox to see if I could fix these three variables. After running the boxcox. The the result had me take the log of SalePrice. The summary print out was a stronger model but TotRms.AbvGrd and Garage.Yr.Blt still had a P-Value higher than 0.05. After looking at the relationship between these two variables and SalePrice again, I decided to

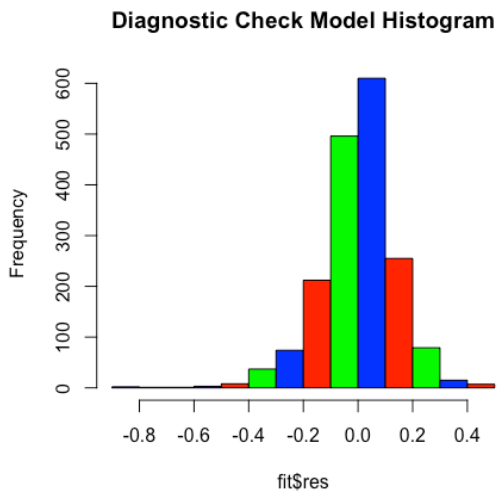
remove these two variables all together. The Diagonostic checks are much better all around and I am happy with this being my model for my numeric values.

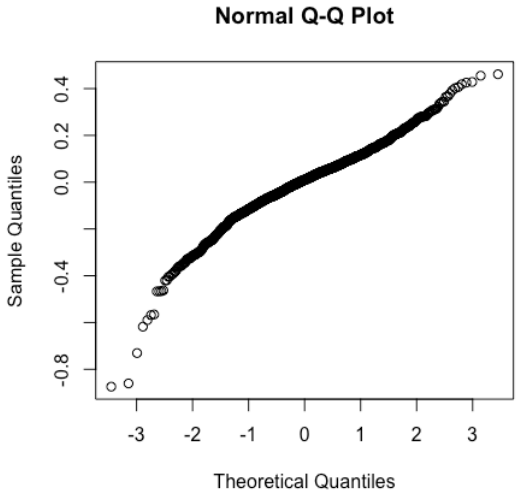


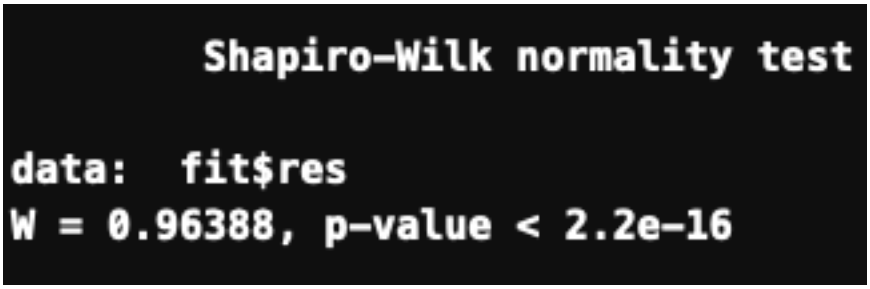
Diagonostic Checks After BoxCox:

```
Call:
lm(formula = num.ames$logSalePrice ~ Overall.Qual + Year.Built +
    Year.Remod.Add + BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF +
    Gr.Liv.Area + Wood.Deck.SF + Open.Porch.SF, data = num.ames,
    subset = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87421 -0.07030  0.00998  0.07985  0.46140

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.423e+00  3.927e-01   6.171 8.35e-10 ***
Overall.Qual   9.480e-02  3.797e-03  24.971  < 2e-16 ***
Year.Built     1.670e-03  1.532e-04  10.901  < 2e-16 ***
Year.Remod.Add 2.550e-03  2.137e-04  11.934  < 2e-16 ***
BsmtFin.SF.1   1.204e-04  8.476e-06  14.208  < 2e-16 ***
Total.Bsmt.SF  1.043e-04  1.446e-05   7.214 7.99e-13 ***
X1st.Flr.SF    6.190e-05  1.566e-05   3.952 8.04e-05 ***
Gr.Liv.Area    2.826e-04  9.192e-06  30.746  < 2e-16 ***
Wood.Deck.SF   1.120e-04  2.808e-05   3.987 6.95e-05 ***
Open.Porch.SF  1.733e-04  5.420e-05   3.198  0.00141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1374 on 1790 degrees of freedom
Multiple R-squared:  0.8848,    Adjusted R-squared:  0.8842
F-statistic:  1527 on 9 and 1790 DF,  p-value: < 2.2e-16
```





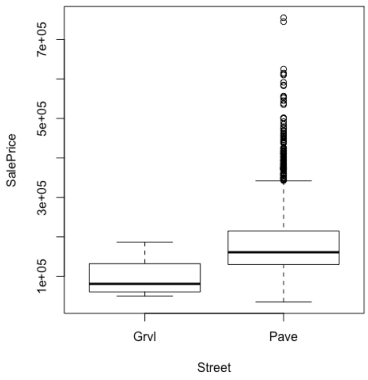


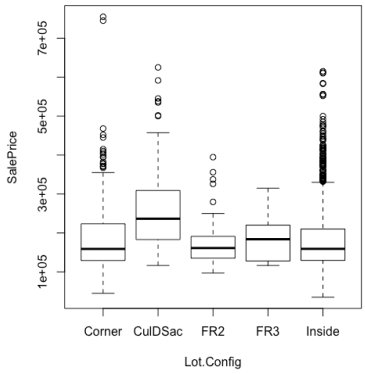


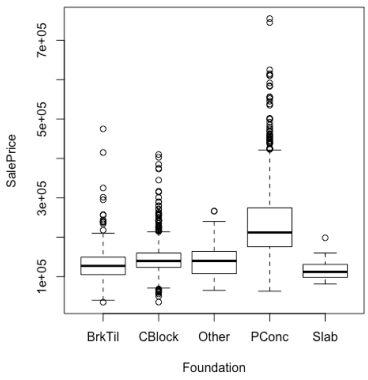
I noticed that there are still outliers in this model. I am ok with that for now though and plan on addressing this at the end once I also have model for my categorical variables.

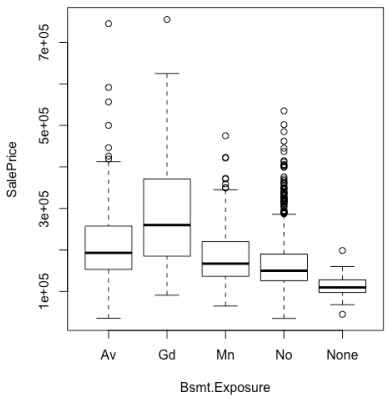
<u>Preliminary Analysis for Categorical Variables:</u>

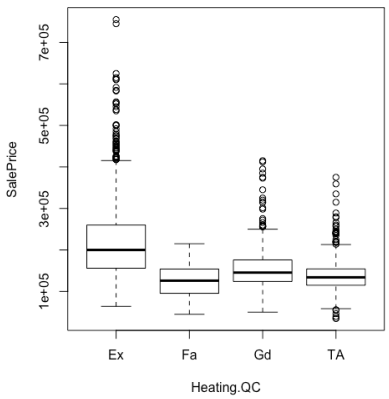
I plotted all of cateogorical variables to see which ones had a relationship with SalePrice. The following plots are the categorical variables I thought had a relationship with SalePrice. There are some variables that was a bit difficult to be able to tell if there was or wasn't a relationship. I went ahead and threw them into my model for now. I will remove them later if needed.

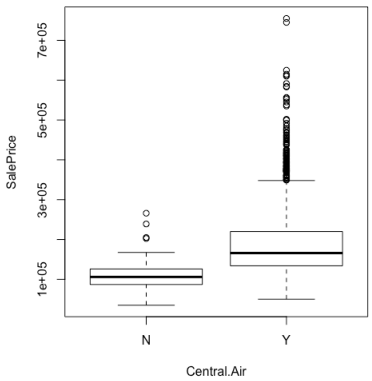

SalePrice
1e+05
3e+05
5e+05
7e+05
Grvl
Pave
Street

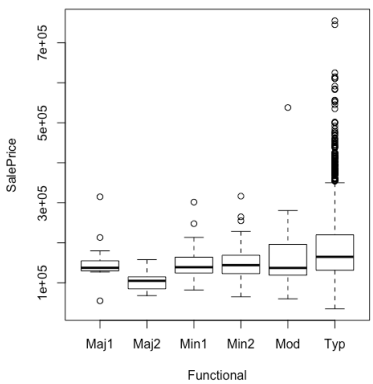

SalePrice
1e+05
3e+05
5e+05
7e+05
Corner
CulDSac
FR2
FR3
Inside
Lot.Config

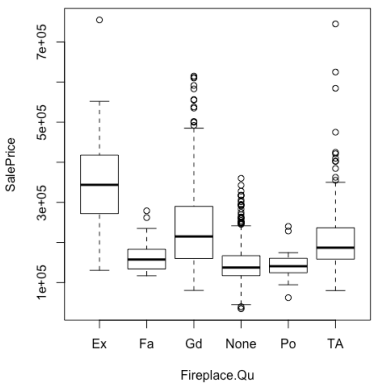

SalePrice
1e+05
3e+05
5e+05
7e+05
BrkTil
CBlock
Other
PConc
Slab
Foundation

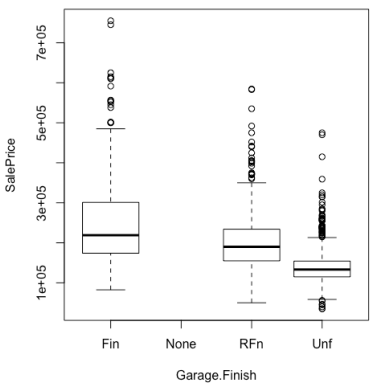

SalePrice
1e+05
3e+05
5e+05
7e+05
Av
Gd
Mn
No
None
Bsmt.Exposure

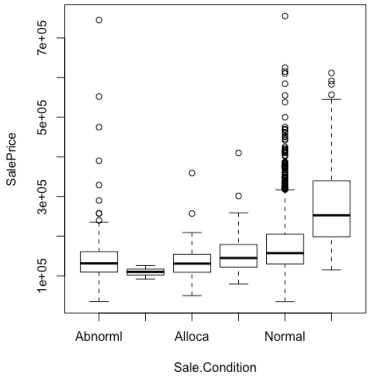

SalePrice
1e+05
3e+05
5e+05
7e+05
Ex
Fa
Gd
TA
Heating.QC


SalePrice
1e+05
3e+05
5e+05
7e+05
N
Y
Central.Air


SalePrice
1e+05
3e+05
5e+05
7e+05
Maj1
Maj2
Min1
Min2
Mod
Typ
Functional


SalePrice
1e+05
3e+05
5e+05
7e+05
Ex
Fa
Gd
None
Po
TA
Fireplace.Qu


SalePrice
1e+05
3e+05
5e+05
7e+05
Fin
None
RFn
Unf
Garage.Finish



Model Comparison:

I will now compare add one categorical variable to the model I created with only the numeric variables. Then use the anova function to compare them both.

```
Analysis of Variance Table

Model 1: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF
Model 2: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1790 33.818
2   1789 33.732  1  0.085418 4.5302 0.03344 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Model 1: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF
Model 2: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1790 33.818
2   1785 33.455  5   0.36261 3.8694 0.00172 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Model 1: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF
Model 2: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config + Foundation
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   1790 33.818
2   1781 33.336  9   0.48108 2.8558 0.002403 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Model 1: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF
Model 2: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config + Foundation +
    Bsmt.Exposure
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1790 33.818
2   1777 32.959 13   0.85834 3.5598 1.522e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

After Running the nova function a few more times. I ended up with this final model.

```
> anova(model1,model2)
Analysis of Variance Table

Model 1: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF
Model 2: num.ames$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config + Foundation +
    Bsmt.Exposure + Heating.QC + Central.Air + Functional
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1790 33.818
2   1768 29.938 22    3.8792 10.413 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

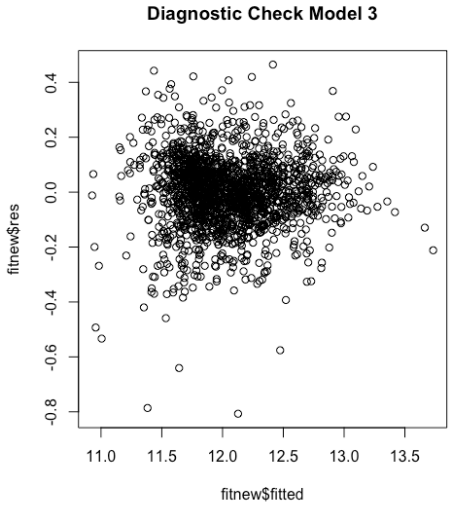
I then conducted a diagnostics checks on my new model which includes both the numeric and non-numeric values: Everything Looks good

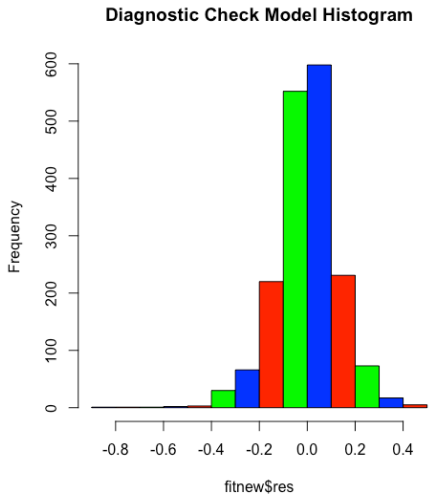
```
Call:
lm(formula = num.ames$logSalePrice ~ Overall.Qual + Year.Built +
    Year.Remod.Add + BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF +
    Gr.Liv.Area + Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config +
    Foundation + Bsmt.Exposure + Heating.QC + Central.Air + Functional,
    data = Ames2, subset = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80369 -0.06770  0.00426  0.07541  0.46330

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.268e+00  4.909e-01   8.694  < 2e-16 ***
Overall.Qual      8.439e-02  3.826e-03  22.055  < 2e-16 ***
Year.Built        1.185e-03  1.982e-04   5.982 2.66e-09 ***
Year.Remod.Add    1.923e-03  2.210e-04   8.704  < 2e-16 ***
BsmtFin.SF.1      1.102e-04  8.361e-06  13.179  < 2e-16 ***
Total.Bsmt.SF     8.064e-05  1.851e-05   4.357 1.40e-05 ***
X1st.Flr.SF       8.403e-05  1.920e-05   4.375 1.28e-05 ***
Gr.Liv.Area       2.855e-04  9.005e-06  31.699  < 2e-16 ***
Wood.Deck.SF      1.045e-04  2.719e-05   3.845 0.000125 ***
Open.Porch.SF     1.889e-04  5.193e-05   3.638 0.000282 ***
StreetPave        1.179e-01  5.438e-02   2.169 0.030231 *
Lot.ConfigCulDSac 4.720e-02  1.734e-02   2.723 0.006536 **
Lot.ConfigFR2    -4.339e-02  2.187e-02  -1.984 0.047390 *
Lot.ConfigFR3    -8.240e-02  4.693e-02  -1.756 0.079277 .
Lot.ConfigInside -1.201e-03  8.381e-03  -0.143 0.886093
FoundationCBlock  5.292e-03  1.302e-02   0.406 0.684478
FoundationOther  -1.805e-02  4.308e-02  -0.419 0.675247
FoundationPConc   1.780e-02  1.570e-02   1.133 0.257276
FoundationSlab    2.328e-02  4.749e-02   0.490 0.624098
Bsmt.ExposureGd   2.937e-02  1.336e-02   2.198 0.028091 *
Bsmt.ExposureMn  -2.027e-02  1.381e-02  -1.468 0.142197
Bsmt.ExposureNo  -2.609e-02  9.752e-03  -2.676 0.007530 **
Bsmt.ExposureNone -4.252e-02  4.399e-02  -0.966 0.333939
Heating.QCFa     -4.888e-02  2.112e-02  -2.315 0.020725 *
Heating.QCGd     -1.891e-02  9.712e-03  -1.948 0.051616 .
Heating.QCTA     -6.149e-02  9.218e-03  -6.670 3.41e-11 ***
Central.AirY      1.146e-01  1.590e-02   7.207 8.43e-13 ***
FunctionalMaj2    4.088e-02  7.637e-02   0.535 0.592543
FunctionalMin1    1.798e-01  4.323e-02   4.158 3.36e-05 ***
FunctionalMin2    1.847e-01  4.287e-02   4.309 1.73e-05 ***
FunctionalMod     1.419e-01  4.808e-02   2.952 0.003204 **
FunctionalTyp     2.270e-01  3.796e-02   5.981 2.67e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1301 on 1768 degrees of freedom
Multiple R-squared:  0.898,     Adjusted R-squared:  0.8962
F-statistic:   502 on 31 and 1768 DF,  p-value: < 2.2e-16
```

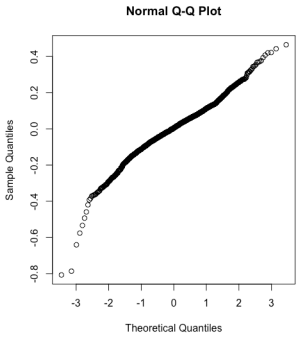
After Looking at the model, It looks like the only variable that needs to be removed is Foundation categorical variable. There are other categorical variables that may have a P-Value over 0.05 but they're still going to be left in our model since they're attached to other variables in their subsidary that "DO" have P-Values less than 0.05.

Final Checks Look Good.

```
lm(formula = num.ames$logSalePrice ~ Overall.Qual + Year.Built +
    Year.Remod.Add + BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF +
    Gr.Liv.Area + Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config +
    Bsmt.Exposure + Heating.QC + Central.Air + Functional, data = Ames2,
    subset = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80698 -0.06840  0.00451  0.07490  0.46453

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.919e+00  4.227e-01   9.271  < 2e-16 ***
Overall.Qual      8.447e-02  3.748e-03  22.539  < 2e-16 ***
Year.Built        1.325e-03  1.553e-04   8.533  < 2e-16 ***
Year.Remod.Add    1.964e-03  2.149e-04   9.138  < 2e-16 ***
BsmtFin.SF.1      1.099e-04  8.294e-06  13.251  < 2e-16 ***
Total.Bsmt.SF     8.206e-05  1.839e-05   4.464 8.57e-06 ***
X1st.Flr.SF       8.236e-05  1.902e-05   4.330 1.57e-05 ***
Gr.Liv.Area       2.869e-04  8.897e-06  32.252  < 2e-16 ***
Wood.Deck.SF      1.019e-04  2.711e-05   3.759 0.000176 ***
Open.Porch.SF     1.904e-04  5.178e-05   3.678 0.000242 ***
StreetPave        1.216e-01  5.392e-02   2.256 0.024218 *
Lot.ConfigCulDSac 4.618e-02  1.731e-02   2.668 0.007705 **
Lot.ConfigFR2    -4.368e-02  2.184e-02  -2.000 0.045681 *
Lot.ConfigFR3    -8.309e-02  4.687e-02  -1.773 0.076451 .
Lot.ConfigInside -1.964e-03  8.357e-03  -0.235 0.814175
Bsmt.ExposureGd   2.924e-02  1.335e-02   2.190 0.028634 *
Bsmt.ExposureMn  -2.007e-02  1.380e-02  -1.455 0.145921
Bsmt.ExposureNo  -2.584e-02  9.744e-03  -2.652 0.008069 **
Bsmt.ExposureNone -3.021e-02  3.091e-02  -0.977 0.328563
Heating.QCFa      -5.043e-02  2.099e-02  -2.403 0.016362 *
Heating.QCGd      -2.033e-02  9.586e-03  -2.121 0.034047 *
Heating.QCTA      -6.408e-02  8.780e-03  -7.298 4.39e-13 ***
Central.AirY       1.129e-01  1.540e-02   7.332 3.44e-13 ***
FunctionalMaj2     3.156e-02  7.563e-02   0.417 0.676484
FunctionalMin1     1.790e-01  4.318e-02   4.144 3.57e-05 ***
FunctionalMin2     1.835e-01  4.281e-02   4.287 1.91e-05 ***
FunctionalMod      1.423e-01  4.796e-02   2.967 0.003049 **
FunctionalTyp      2.269e-01  3.794e-02   5.982 2.66e-09 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1301 on 1772 degrees of freedom
Multiple R-squared:  0.8978,    Adjusted R-squared:  0.8963
F-statistic: 576.8 on 27 and 1772 DF,  p-value: < 2.2e-16
```








```
        Shapiro-Wilk normality test

data:  fitnew$res
W = 0.97235, p-value < 2.2e-16
```

Removing Outliers:

I first had to put the missing data back into my dataframe. By using the na.action=na.exclued function at the end of my linear model.

Afterwards I received the following output once I ran my code:

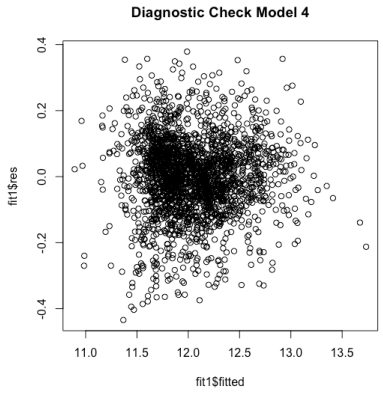
	Misc.Val	Yr.Sold	Sale.Type	Sale.Condition	SalePrice	sres
307	0	2010	WD	Alloca	50138	-3.440812
372	0	2009	WD	Family	82500	-6.349877
561	0	2009	WD	Normal	62383	-4.903205
695	0	2009	WD	Abnorml	66500	-3.998236
790	0	2009	COD	Abnorml	60000	-5.317468
1471	0	2008	WD	Abnorml	185000	-3.072142
1553	0	2008	WD	Normal	40000	-6.248267
1780	0	2007	New	Partial	147000	-4.549078
1947	0	2007	WD	Normal	64500	-3.600840
1996	0	2007	WD	Normal	133900	-3.032386
2598	0	2006	COD	Abnorml	80000	-3.070253
2839	0	2006	Con	Normal	35000	-3.929471
2876	0	2006	WD	Abnorml	35311	-4.251353

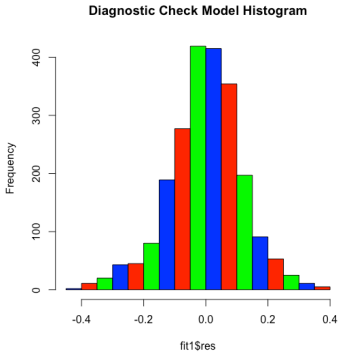
	Misc.Val	Yr.Sold	Sale.Type	Sale.Condition	SalePrice	sres
919	0	2009	WD	Normal	205000	3.245110
929	0	2009	WD	Normal	158000	3.230606
1535	0	2008	WD	Alloca	257076	3.162753
1638	0	2007	New	Partial	392000	3.659438
2005	0	2007	WD	Normal	144000	3.559679
2208	0	2007	WD	Normal	153500	3.005213
2214	0	2007	WD	Normal	315000	3.373858
2854	0	2006	WD	Normal	195000	3.202883

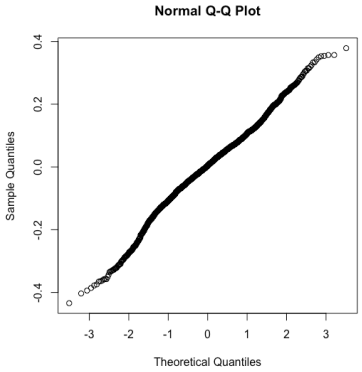
The following 21 Variables I ended up removing. My Model was already in good shape and I figured removing these 21 outliers wouldn't hurt my model since it is such a big dataset.

The following is the new summary output with the outliers now removed:

```
Call:
lm(formula = x$logSalePrice ~ Overall.Qual + Year.Built + Year.Remod.Add +
    BsmtFin.SF.1 + Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area +
    Wood.Deck.SF + Open.Porch.SF + Street + Lot.Config + Bsmt.Exposure +
    Heating.QC + Central.Air + Functional, data = Ames, na.action = na.exclude)

Residuals:
      Min       1Q   Median       3Q      Max
-0.43427 -0.06700  0.00412  0.07138  0.37859

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.615e+00  3.411e-01  10.600  < 2e-16 ***
Overall.Qual      7.955e-02  3.084e-03  25.797  < 2e-16 ***
Year.Built        1.434e-03  1.270e-04  11.291  < 2e-16 ***
Year.Remod.Add    1.985e-03  1.728e-04  11.482  < 2e-16 ***
BsmtFin.SF.1      1.015e-04  6.809e-06  14.913  < 2e-16 ***
Total.Bsmt.SF     1.065e-04  1.520e-05   7.007 3.22e-12 ***
X1st.Flr.SF       6.752e-05  1.573e-05   4.291 1.85e-05 ***
Gr.Liv.Area       2.939e-04  7.301e-06  40.253  < 2e-16 ***
Wood.Deck.SF      8.423e-05  2.216e-05   3.800 0.000148 ***
Open.Porch.SF     1.454e-04  4.231e-05   3.438 0.000597 ***
StreetPave        1.310e-01  4.883e-02   2.683 0.007351 **
Lot.ConfigCulDSac 3.709e-02  1.424e-02   2.605 0.009242 **
Lot.ConfigFR2     -4.076e-02  1.656e-02  -2.461 0.013930 *
Lot.ConfigFR3     -7.558e-02  4.006e-02  -1.887 0.059327 .
Lot.ConfigInside  -2.044e-03  6.799e-03  -0.301 0.763691
Bsmt.ExposureGd    3.599e-02  1.081e-02   3.329 0.000885 ***
Bsmt.ExposureMn   -1.888e-02  1.117e-02  -1.690 0.091085 .
Bsmt.ExposureNo   -2.419e-02  7.781e-03  -3.108 0.001905 **
Bsmt.ExposureNone -2.604e-02  2.489e-02  -1.046 0.295702
Heating.QCFa      -7.449e-02  1.704e-02  -4.372 1.29e-05 ***
Heating.QCGd      -2.148e-02  7.737e-03  -2.777 0.005534 **
Heating.QCTA      -5.807e-02  7.134e-03  -8.140 6.51e-16 ***
Central.AirY       8.985e-02  1.262e-02   7.117 1.48e-12 ***
FunctionalMaj2     4.488e-02  6.430e-02   0.698 0.485235
FunctionalMin1     2.619e-01  3.967e-02   6.602 5.05e-11 ***
FunctionalMin2     2.559e-01  3.978e-02   6.433 1.53e-10 ***
FunctionalMod      2.561e-01  4.316e-02   5.935 3.41e-09 ***
FunctionalTyp      3.065e-01  3.594e-02   8.527  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1181 on 2209 degrees of freedom
Multiple R-squared:  0.9105,    Adjusted R-squared:  0.9094
F-statistic: 832.1 on 27 and 2209 DF,  p-value: < 2.2e-16
```








```
        Shapiro-Wilk normality test

data:  fit1$res
W = 0.98937, p-value = 8.332e-12
```

After removing the outliers from our model, I ran the diagnostic checks one more time on my final model and everything looks great! Fitted vs Residuals has a consistent variance. Our histogram is a normal distribution. Normal Q-Q Plot and Shapiro-Wilk Look good.

AmesHousing_Predict Data:

Now that I have developed my model, I will now predict SalePrice using the AmesHousing_Predict data:

```
> pred = predict(fit1,Predict.Data,interval="prediction")
> pred
       fit      lwr      upr
1 11.40149 11.15050 11.65248
2 11.77428 11.53968 12.00888
```

Thus, I predict that 95% of SalePrice for:
Housing 1, has a Price between 11.2(logofSalePrice) and 11.66(logofSalePrice)
Housing 2, has a Price between 11.5(logofSalePrice) and 12.0(logofSalePrice)