Final Project Web Mining

May 4, 2020

IS 688 Web Mining - Final Project Submission

by **Group 3**
New Jersey Institute Of Technology - **Spring 2020**

- **Members of Group 3 :**

  Haisam Ezmat

  Pratik Parija

  Sai Akhilesh Chunduri

  Munazzam Mirza

Prediction of House Price

Dataset we are using is available on Kraggle and here's a brief version of what data file looks like:

• SalePrice — the property's sale price in dollars. (This is the target variable that we trying to predict) • MSSubClass — the building class • MSZoning — the general zoning classification • LotFrontage — linear feet of street connected to property • LotArea — Lot size in square feet • Street — Type of road access • Alley — Type of alley access • LotShape — General shape of property • LandContour — Flatness of the property • Utilities — Type of utilities available

There are 1460 observations with 79 explanatory variables describing almost every aspect of residential homes in Ames, Iowa.

Among explanatory variables, there are 37 integer variables, such as Id, MSSubClass, LotFrontage. There are 43 factor variables such as MSZoning, Street, LotShape. Our goal is to predict sale price of each house. For each Id in the test set, we will predict the value of Sale Price variable.

We will divide our project in 3 sections.

Section 1: Exploratory Data Analysis

Section 2: Feature Engineering

Section 3: Model Building - Training and Testing

Section 1 - Exploratory Data Analysis

[40]:

```r
library(ggplot2)
library('ggplot2')
library('ggthemes')
library('scales')
library('dplyr')
library('mice')
library('randomForest')
library('data.table')
library('gridExtra')
library('corrplot')
library('GGally')
library('e1071')
path='C:/Users/Munazzam/Downloads/train.csv'
data=data.frame(read.csv(path))
train <-read.csv('C:/Users/Munazzam/Downloads/train.csv', stringsAsFactors = F)
summary(data)
```

```
       Id           MSSubClass        MSZoning     LotFrontage    
 Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00  
 1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00  
 Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00  
 Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05  
 3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00  
 Max.   :1460.0   Max.   :190.0                  Max.   :313.00  
                                                 NA's   :259     
    LotArea          Street      Alley       LotShape  LandContour  Utilities   
 Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459  
 1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1  
 Median :  9478               NA's:1369   IR3: 10   Low:  36                 
 Mean   : 10517                           Reg:925   Lvl:1311                 
 3rd Qu.: 11602                                                              
 Max.   :215245                                                              

   LotConfig    LandSlope   Neighborhood   Condition1     Condition2  
 Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445  
 CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6  
 FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2  
 FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2  
 Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2  
                           Gilbert: 79   RRAe   :  11   PosA   :   1  
                           (Other):707   (Other):  15   (Other):   2  
  BldgType      HouseStyle   OverallQual      OverallCond      YearBuilt   
 1Fam  :1220   1Story :726   Min.   : 1.000   Min.   :1.000   Min.   :1872  
 2fmCon:  31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954  
 Duplex:  52   1.5Fin :154   Median : 6.000   Median :5.000   Median :1973  
```

```
Twnhs :  43   SLvl   : 65   Mean   : 6.099   Mean   :5.575   Mean   :1971
TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
              1.5Unf : 14   Max.   :10.000   Max.   :9.000   Max.   :2010
              (Other): 19
 YearRemodAdd    RoofStyle       RoofMatl     Exterior1st   Exterior2nd
Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515   VinylSd:504
1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222   MetalSd:214
Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220   HdBoard:207
Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206   Wd Sdng:197
3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108   Plywood:142
Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61   CmentBd: 60
                              (Other):   2   (Other):128   (Other):136
  MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual
BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146   Ex  :121
BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634   Fa  : 35
None   :864   Median :   0.0   Gd:488    Gd: 146   PConc :647   Gd  :618
Stone  :128   Mean   : 103.7   TA:906    Po:   1   Slab  : 24   TA  :649
NA's   :  8   3rd Qu.: 166.0             TA:1282   Stone :  6   NA's: 37
              Max.   :1600.0                       Wood  :  3
              NA's   :8
BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2
Fa  :  45   Av  :221     ALQ :220     Min.   :   0.0   ALQ :  19
Gd  :  65   Gd  :134     BLQ :148     1st Qu.:   0.0   BLQ :  33
Po  :   2   Mn  :114     GLQ :418     Median : 383.5   GLQ :  14
TA  :1311   No  :953     LwQ : 74     Mean   : 443.6   LwQ :  46
NA's:  37   NA's: 38     Rec :133     3rd Qu.: 712.2   Rec :  54
                         Unf :430     Max.   :5644.0   Unf :1256
                         NA's: 37                      NA's:  38
  BsmtFinSF2        BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC
Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741
1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49
Median :   0.00   Median : 477.5   Median : 991.5   GasW :  18   Gd:241
Mean   :  46.55   Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
Max.   :1474.00   Max.   :2336.0   Max.   :6110.0   Wall :   4

CentralAir Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
N:  95     FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
Y:1365     FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
           FuseP:   3   Median :1087   Median :   0   Median :  0.000
           Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
           SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
           NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000

  GrLivArea     BsmtFullBath     BsmtHalfBath       FullBath
Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
Median :1464   Median :0.0000   Median :0.00000   Median :2.000
```

```
Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000

   HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd
Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100      Min.   : 2.000
1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39      1st Qu.: 5.000
Median :0.0000   Median :3.000   Median :1.000   Gd:586      Median : 6.000
Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735      Mean   : 6.518
3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000               3rd Qu.: 7.000
Max.   :2.0000   Max.   :8.000   Max.   :3.000               Max.   :14.000

Functional    Fireplaces    FireplaceQu   GarageType   GarageYrBlt
Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6   Min.   :1900
Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870   1st Qu.:1961
Min1:  31   Median :1.000   Gd  :380    Basment: 19   Median :1980
Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88   Mean   :1979
Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9   3rd Qu.:2002
Sev :   1   Max.   :3.000   NA's:690    Detchd :387   Max.   :2010
Typ :1360                               NA's   : 81   NA's   :81
GarageFinish   GarageCars      GarageArea     GarageQual  GarageCond
Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3   Ex  :   2
RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48   Fa  :  35
Unf :605     Median :2.000   Median : 480.0   Gd  :  14   Gd  :   9
NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3   Po  :   7
             3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311   TA  :1326
             Max.   :4.000   Max.   :1418.0   NA's:  81   NA's:  81

PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch     X3SsnPorch
N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
Y:1340     Median :  0.00   Median : 25.00   Median :  0.00   Median :  0.00
           Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41
           3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00
           Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00

 ScreenPorch        PoolArea        PoolQC       Fence      MiscFeature
Min.   :  0.00   Min.   :  0.000   Ex  :   2   GdPrv:  59   Gar2:   2
1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2   GdWo :  54   Othr:   2
Median :  0.00   Median :  0.000   Gd  :   3   MnPrv: 157   Shed:  49
Mean   : 15.06   Mean   :  2.759   NA's:1453   MnWw :  11   TenC:   1
3rd Qu.:  0.00   3rd Qu.:  0.000               NA's :1179   NA's:1406
Max.   :480.00   Max.   :738.000

   MiscVal            MoSold          YrSold        SaleType
Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :1267
1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   New    : 122
Median :    0.00   Median : 6.000   Median :2008   COD    :  43
```

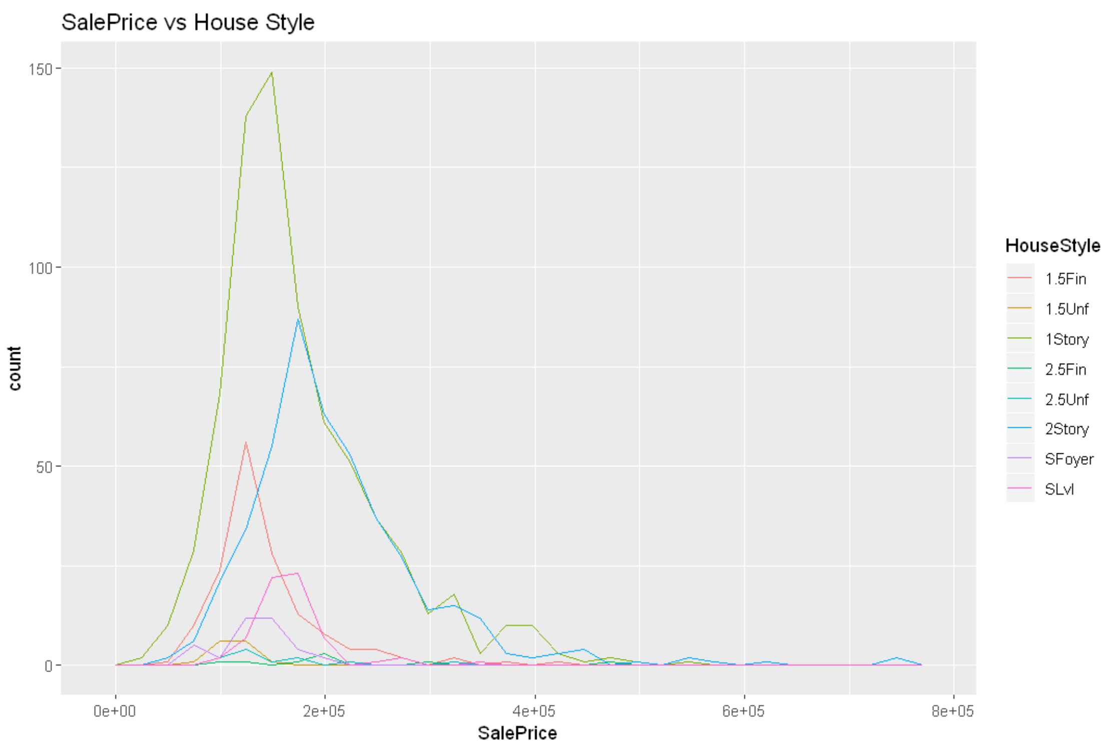
```
Mean   :   43.49   Mean   : 6.322   Mean   :2008   ConLD  :   9
3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   5
Max.   :15500.00   Max.   :12.000   Max.   :2010   ConLw  :   5
                                                   (Other):   9
SaleCondition    SalePrice
Abnorml: 101   Min.   : 34900
AdjLand:   4   1st Qu.:129975
Alloca :  12   Median :163000
Family :  20   Mean   :180921
Normal :1198   3rd Qu.:214000
Partial: 125   Max.   :755000
```

1.1 - Plotting SalePrice vs House Style

[21]:

```
ggplot(data, aes(SalePrice, color=HouseStyle)) + geom_freqpoly() +
ggtitle("SalePrice vs House Style")
```

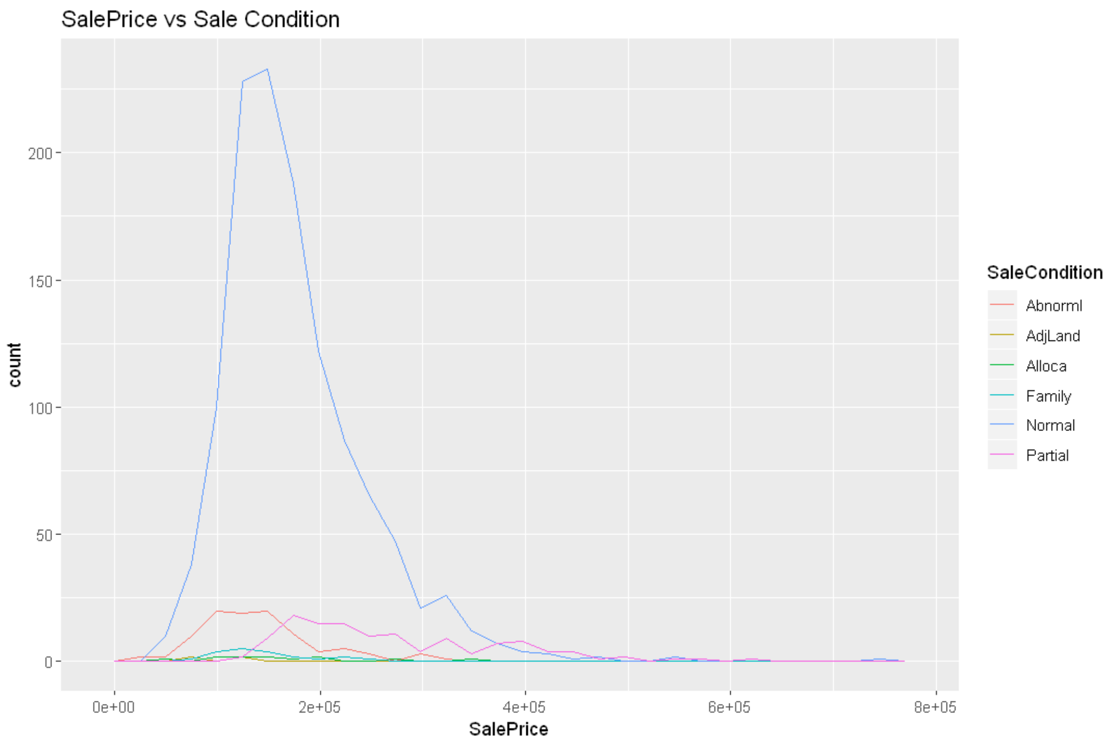
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.2 - Plotting SalePrice vs Sale Condition

[22]:
```
ggplot(data, aes(SalePrice, color=SaleCondition)) + geom_freqpoly() +␣
↪geom_freqpoly() + ggtitle("SalePrice vs Sale Condition")
```

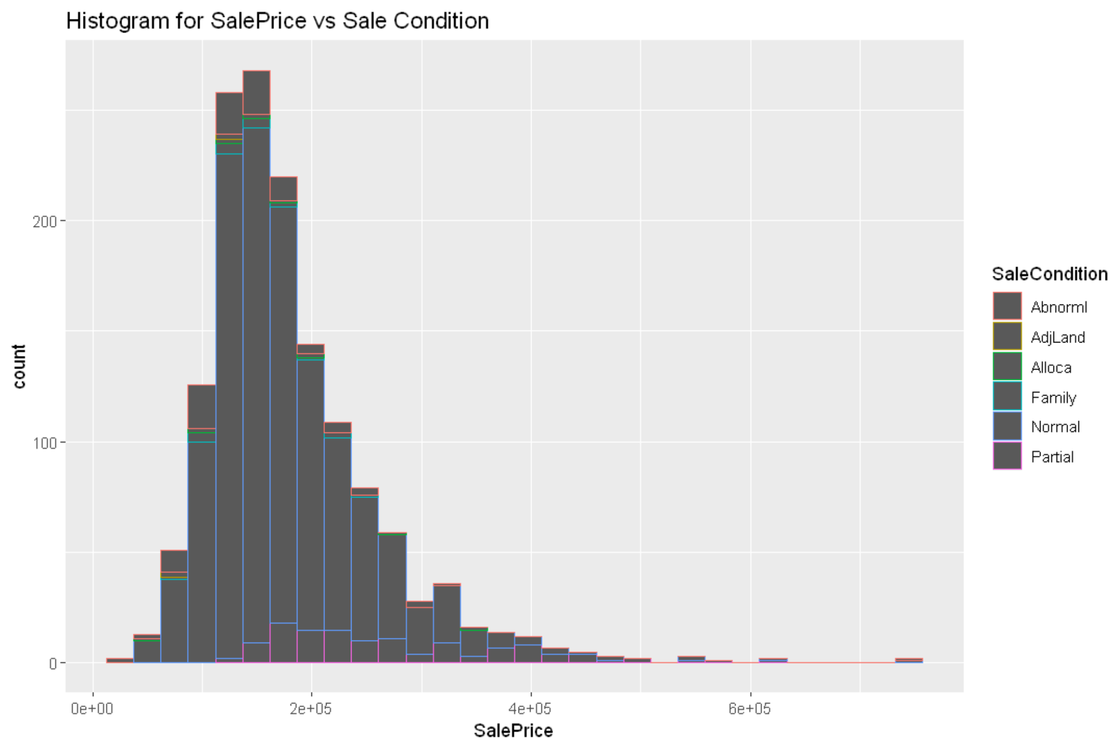
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.3 - Plotting Histogram for SalePrice vs Sale Condition

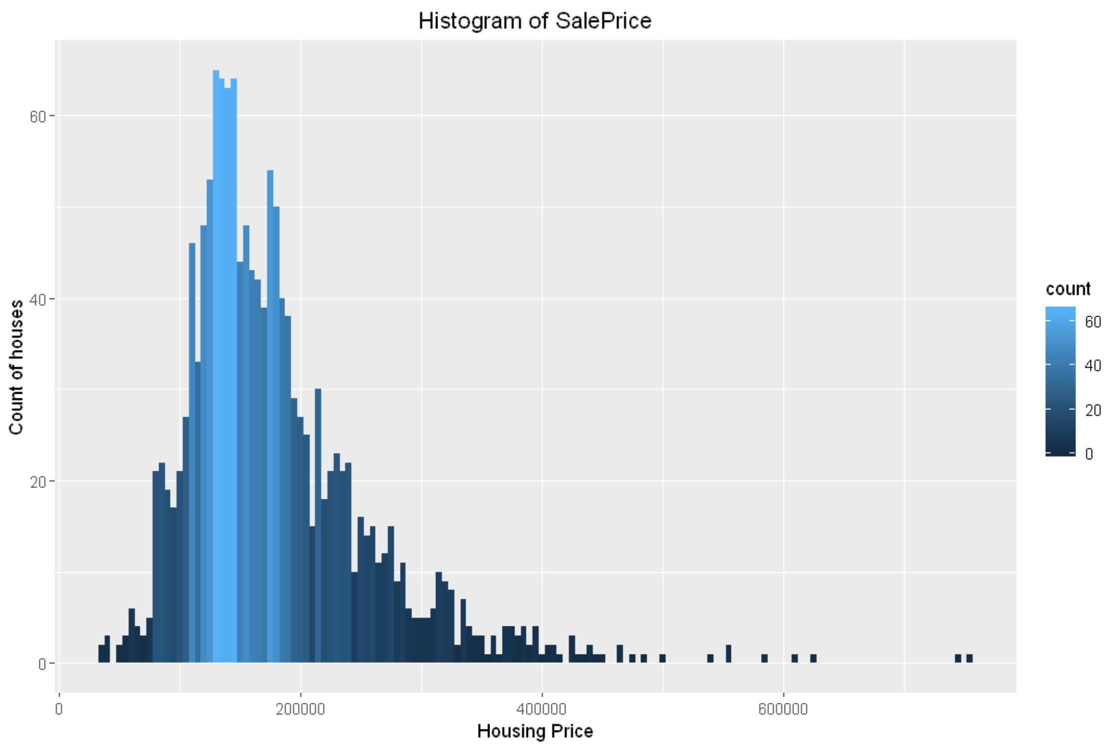
[23]:
```
ggplot(data, aes(SalePrice, color=SaleCondition)) +
    geom_histogram() + ggtitle("Histogram for SalePrice vs Sale Condition")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.4 - Plotting Higtogram to figure out the distribution of SalePrice

[24]:
```
options(scipen=10000)
ggplot(data, aes(x = SalePrice, fill = ..count..)) +
  geom_histogram(binwidth = 5000) +
  ggtitle("Histogram of SalePrice") +
  ylab("Count of houses") +
  xlab("Housing Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

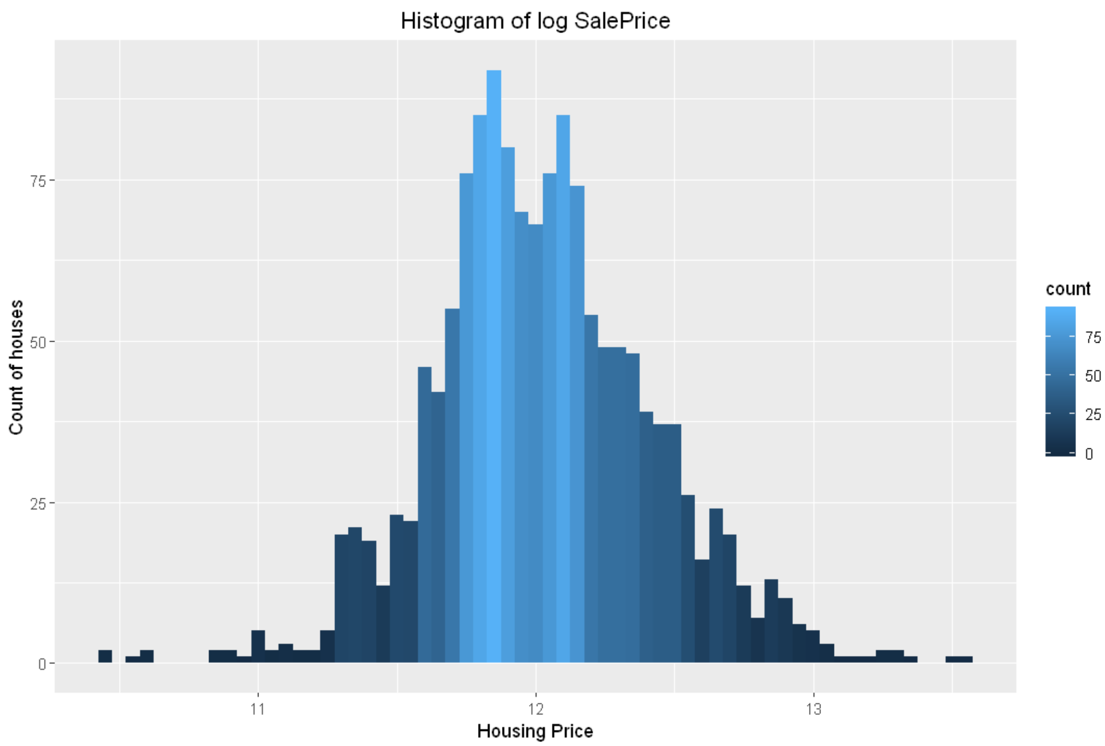


Histogram above is skewed to right. Lets do a normal distrubution to fix it.

[25]:
```
#Taking log of SalePrice

data$lSalePrice <- log(data$SalePrice)
```

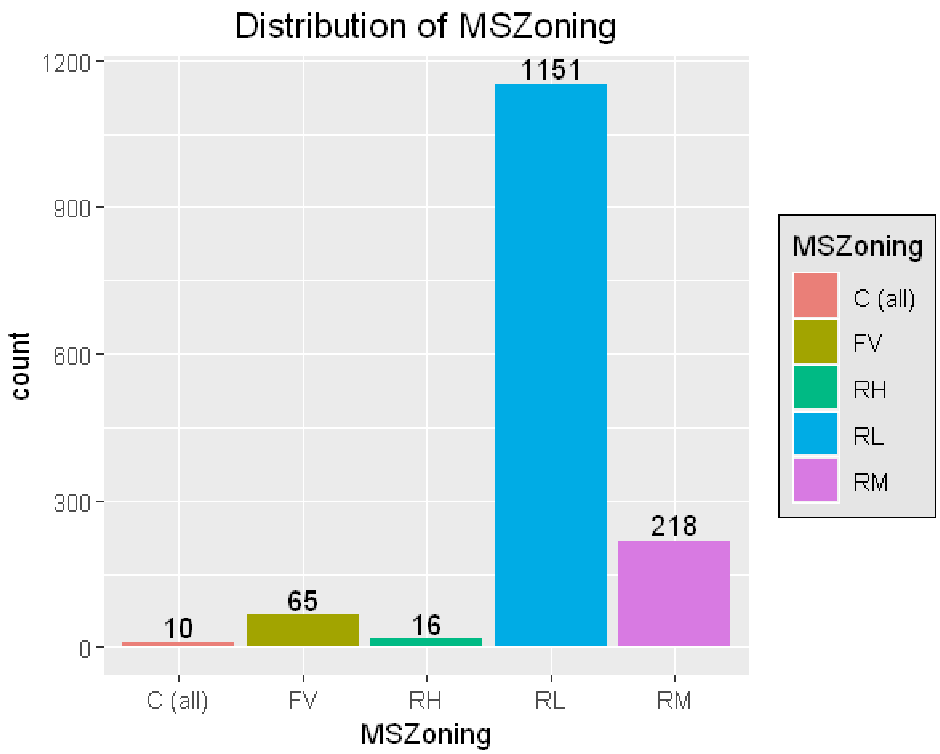
1.5 - Plotting Higtogram of log SalePrice

[26]:
```
ggplot(data, aes(x = lSalePrice, fill = ..count..)) +
  geom_histogram(binwidth = 0.05) +
  ggtitle("Histogram of log SalePrice") +
  ylab("Count of houses") +
  xlab("Housing Price") +
  theme(plot.title = element_text(hjust = 0.5))
```



1.6 - Bar Chart Counting houses by MSZoning

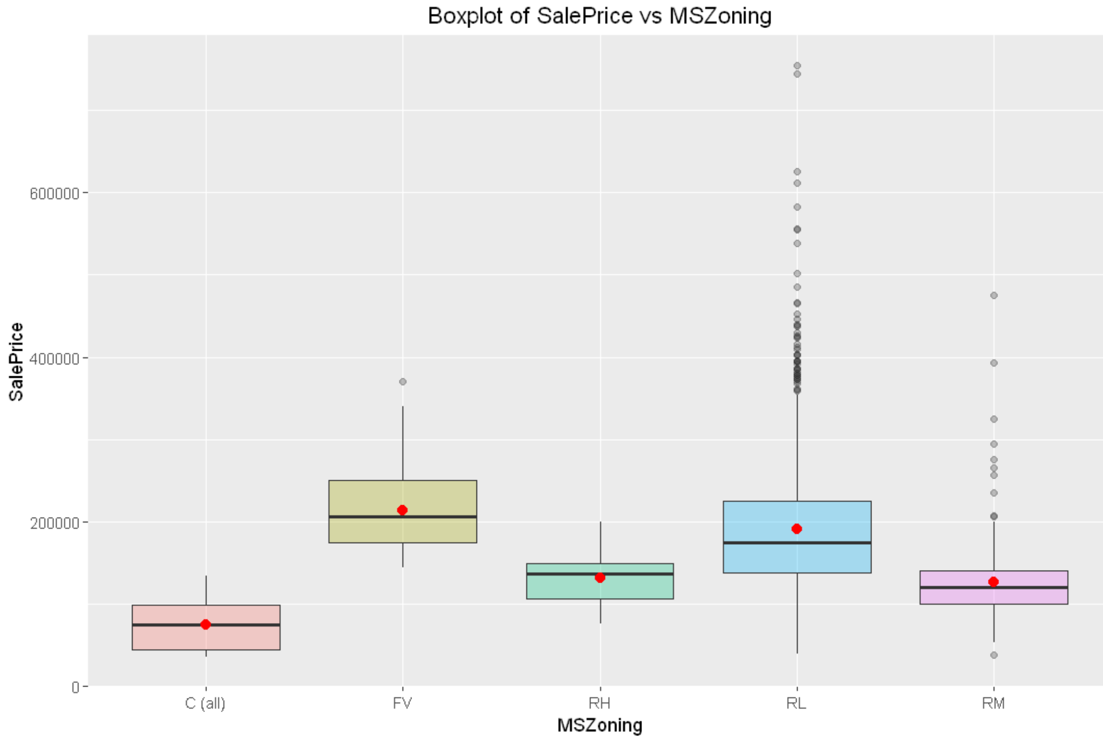
[27]:
```
options(repr.plot.width=5, repr.plot.height=4)
ggplot(data, aes(x = MSZoning, fill = MSZoning )) +
geom_bar()+
scale_fill_hue(c = 80)+
ggtitle("Distribution of MSZoning")+
theme(plot.title = element_text(hjust = 0.5),legend.position="right", legend.
 ↪background = element_rect(fill="grey90",
                                                                              ␣
 ↪                                size=0.5, linetype="solid",
                                                                              ␣
 ↪                                colour ="black"))+
geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



It can be deduced from the graph above that most of houses in this dataset are built in the area of Residential Low Density(1151 houses), and follows by Residential Medium Density(218 houses). Few houes are built in Commercial, Floating Village and Residential High Density.

1.7 - Boxplot Distrubution of Price in each MSZoning

[28]:
```
# Change plot size to 9 x 6
options(repr.plot.width=9, repr.plot.height=6)

#boxplot of SalePrice by MSZoning
#add average value of SalePrice as red point

ggplot(data, aes(x=MSZoning, y=SalePrice, fill=MSZoning)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red",␣
↪fill="red")+
  theme(legend.position="none")+
  ggtitle("Boxplot of SalePrice vs MSZoning")+
  theme(plot.title = element_text(hjust = 0.5))
```



The graph above shows the distribution of SalePrice by MSZoning. The sales in "Floating Village Residential" area have the highest average sale price, and then followed by "Residential Low Density". While "Commercial" sales have the lowest average sale price

Lets visualize SalePrice by different cateogries of BldfType.

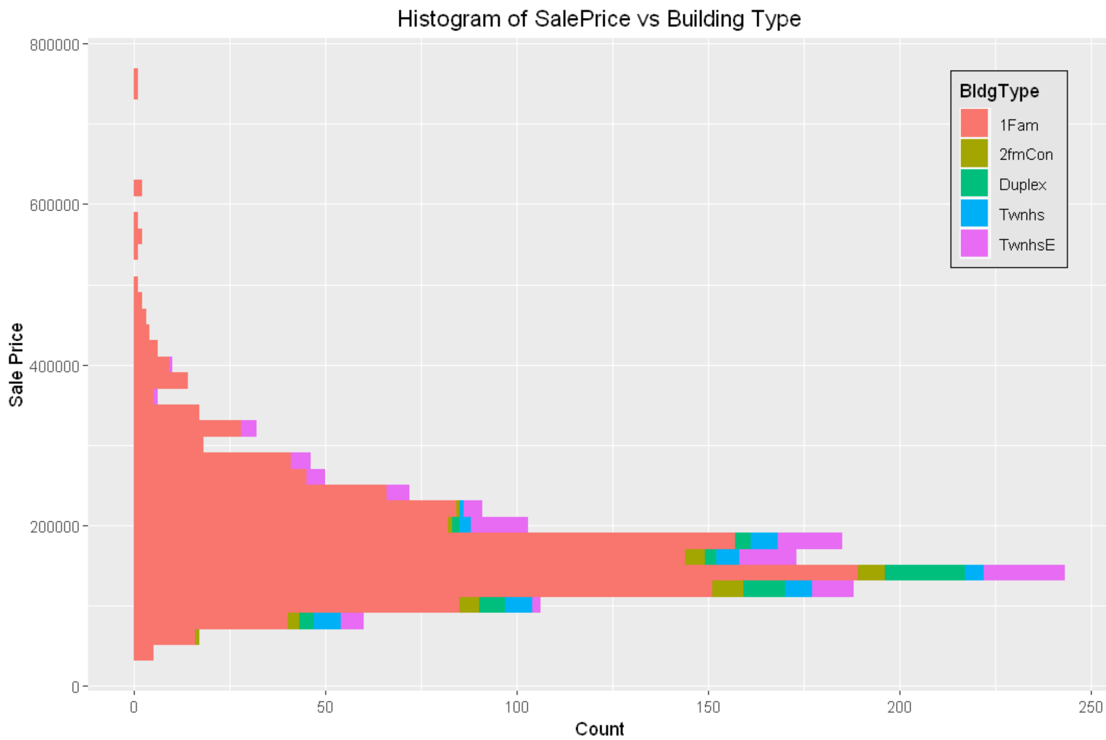
BldgType: Type of dwelling

1Fam : Single-family Detached
2FmCon : Two-family Conversion; originally built as one-family dwelling Duplx : Duplex TwnhsE : Townhouse End Unit TwnhsI : Townhouse Inside Unit

1.8 - Plotting Historgram of Sale Price vs BldgType

[29]:
```
ggplot(data, aes(SalePrice)) +
 geom_histogram(aes(fill = BldgType), position = position_stack(reverse = TRUE),␣
 ↪binwidth = 20000) +
 coord_flip() + ggtitle("Histogram of SalePrice vs Building Type") +
 ylab("Count") +
 xlab("Sale Price") +
 theme(plot.title = element_text(hjust = 0.5),legend.position=c(0.9,0.8), legend.
 ↪background = element_rect(fill="grey90",
                                                                                ␣
 ↪                           size=0.5, linetype="solid",
```

```
↪                           colour ="black"))
```

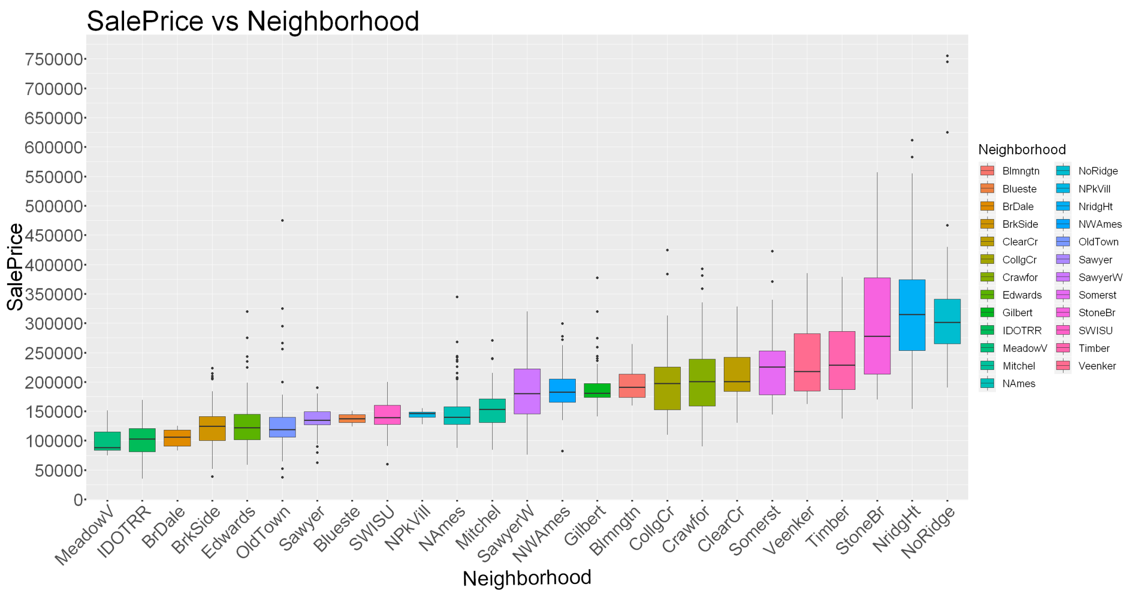


As we can see from the graph above: 1. Single-family Detached price range from 50,000 to 300,000. 2. Two-family Conversion, Duplex, Townhouse End Unit and Townhouse Inside Unit has price ranging from 75000 to 210000.

1.9 - Plotting SalePrice vs Neighborhood

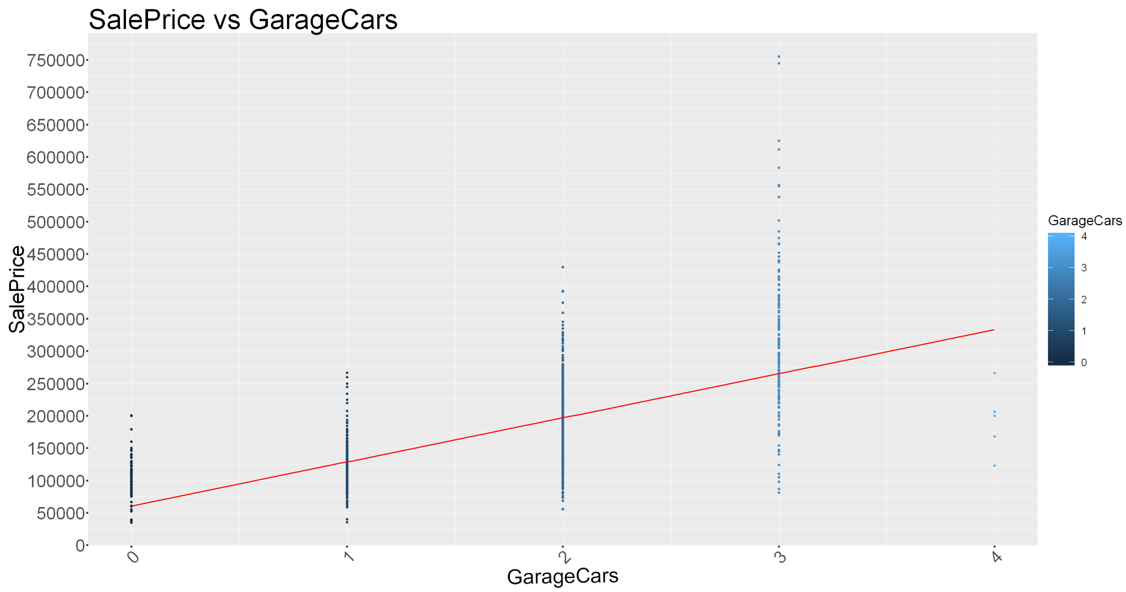
[30]:
```
options(repr.plot.width = 25, repr.plot.height = 13) # Defyning plot size
ggplot(aes(x = reorder(Neighborhood,SalePrice), y = SalePrice,fill =
↪Neighborhood),,data = data) +
  geom_boxplot() + labs(x='Neighborhood', y='SalePrice') +
  ggtitle('SalePrice vs Neighborhood')+
  scale_y_continuous(breaks= seq(0, 800000, by=50000))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1)
        ,axis.title = element_text(size = rel(3), angle = 1)
       ,plot.title = element_text(size = rel(4))
        ,axis.text =  element_text(size = rel(2.5))
        ,axis.ticks = element_line(size = 1.5)
       ,legend.key.size = unit(1, "cm")
       ,legend.title = element_text(size=22)
```

```
       ,legend.text = element_text(size=16))
```



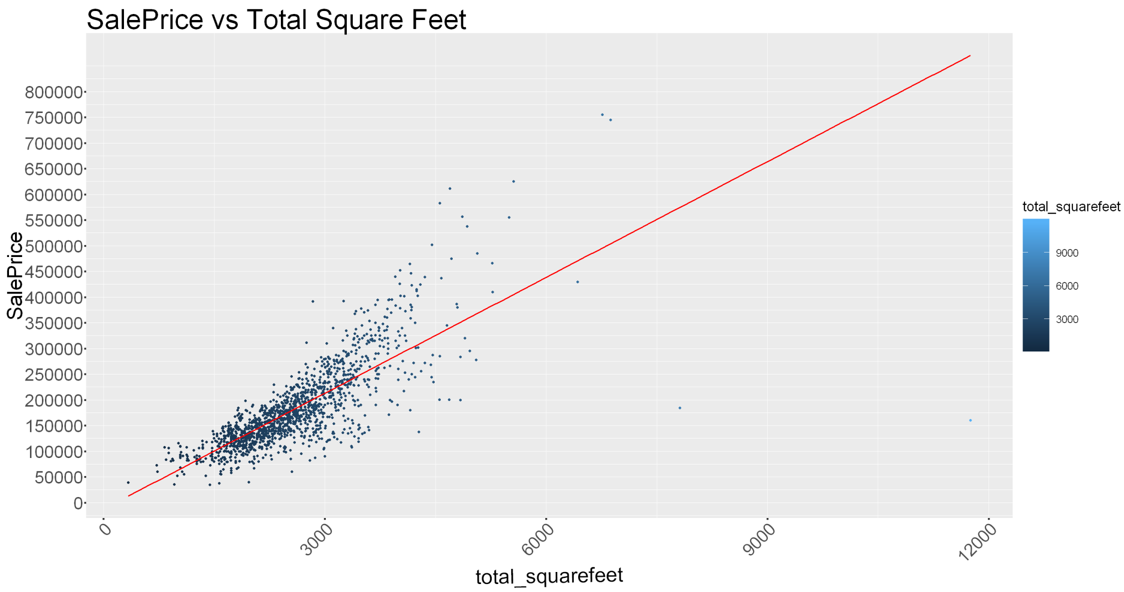
1.10 - Plotting SalePrice vs GarageCars

[31]:
```
options(repr.plot.width = 25, repr.plot.height = 13) # Defyning plot size
ggplot(aes(x = GarageCars, y = SalePrice,color = GarageCars),,data = data) +
  geom_point() +
  geom_smooth(method = "lm",col ='red', se = FALSE)+
  ggtitle('SalePrice vs GarageCars')+
  scale_y_continuous(breaks= seq(0, 800000, by=50000))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1)
        ,axis.title = element_text(size = rel(3), angle = 1)
       ,plot.title = element_text(size = rel(4))
        ,axis.text =  element_text(size = rel(2.5))
        ,axis.ticks = element_line(size = 1.5)
       ,legend.key.size = unit(1.5, "cm")
       ,legend.title = element_text(size=22)
       ,legend.text = element_text(size=16))
```



1.11 - Plotting SalePrice vs Total Square Feet

[32]:
```
total_squarefeet <- data$GrLivArea + data$TotalBsmtSF

options(repr.plot.width = 25, repr.plot.height = 13) # Defyning plot size
ggplot(aes(x = total_squarefeet, y = SalePrice,color = total_squarefeet),,data =
 ↪data) +
  geom_point() +
  geom_smooth(method = "lm",col ='red', se = FALSE)+
  ggtitle('SalePrice vs Total Square Feet')+
  scale_y_continuous(breaks= seq(0, 800000, by=50000))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1)
        ,axis.title = element_text(size = rel(3), angle = 1)
       ,plot.title = element_text(size = rel(4))
        ,axis.text =  element_text(size = rel(2.5))
        ,axis.ticks = element_line(size = 1.5)
       ,legend.key.size = unit(1.5, "cm")
       ,legend.title = element_text(size=22)
       ,legend.text = element_text(size=16))
```



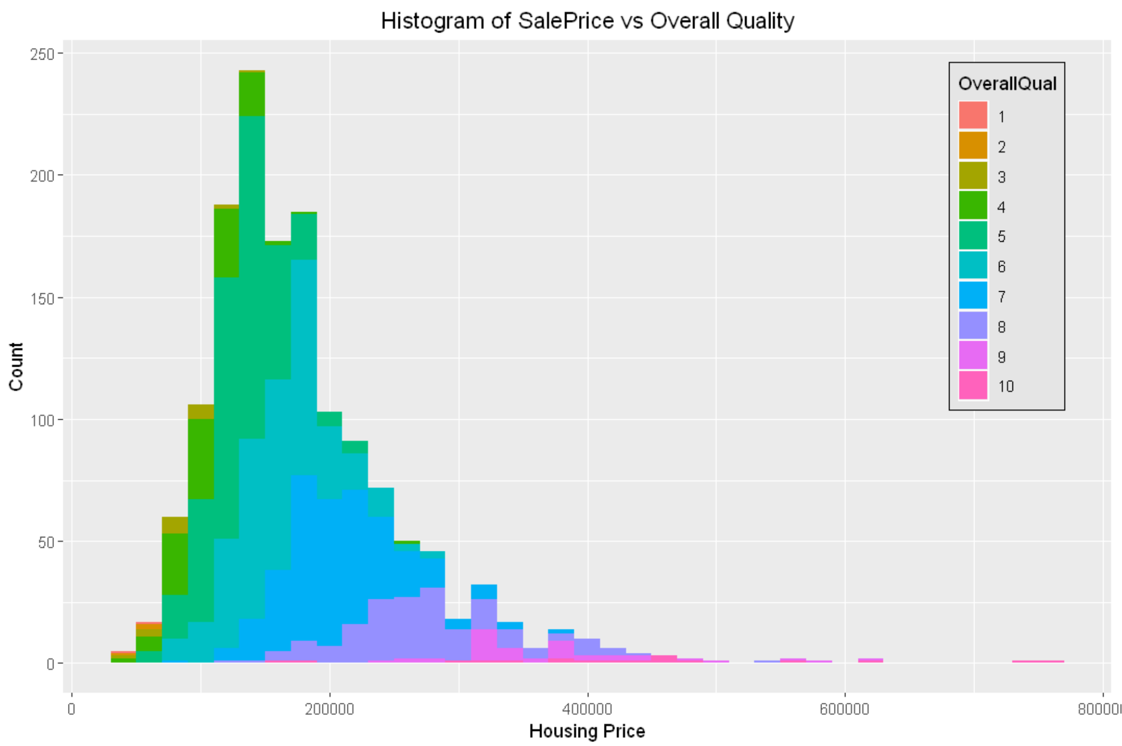
1.12 - Plotting Histogram of SalePrice vs Overall Quality

Lets visualize Sale Price by OverallQual.

OverallQual: Rates the overall material and finish of the house

10 Very Excellent 9 Excellent 8 Very Good 7 Good 6 Above Average 5 Average 4 Below Average 3 Fair 2 Poor 1 Very Poor

[50]:
```
ggplot(data, aes(x = SalePrice,fill = as.factor(OverallQual))) +
  geom_histogram(position = "stack", binwidth = 20000) +
  ggtitle("Histogram of SalePrice vs Overall Quality") +
  ylab("Count") +
  xlab("Housing Price") +
  scale_fill_discrete(name="OverallQual")+
  theme(plot.title = element_text(hjust = 0.5), legend.position=c(0.9,0.7),
    legend.background = element_rect(fill="grey90",size=0.5,␣
↪linetype="solid",colour ="black"))
```



As we see in graph above most houses are with OverallQuall of 4,5,6 and 7 which is equivalent to “Below Average”, “Average”, “Above Average” and “Good”. Sale Price increases as Overall Quality increases. For each rate level of overall quality, the distribution of house price is almost symmetric.

1.13 - Bar Plots

Lets create some Bar Plots for more insights into the data.

MSZoning bar plot indicates that majority of the houses are located in low density residential areas and medium density residential area.

The type of road access to the property tends to be paved and the houses do not have alleys.

Landcontour bar plot shows that the houses are built on flat properties.

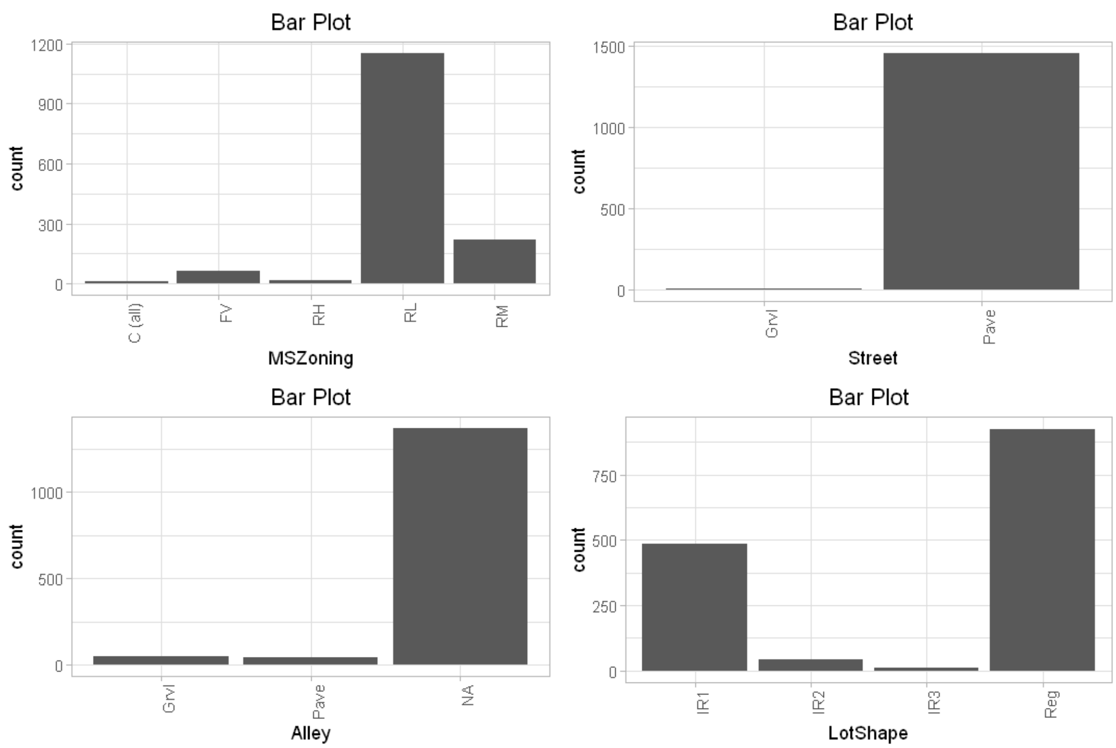
Utilities bar plot shows that almost all homes have all public utilities (E,G,W & S).

LandSlope bar plot shows that most of the properties have a gentle slope.

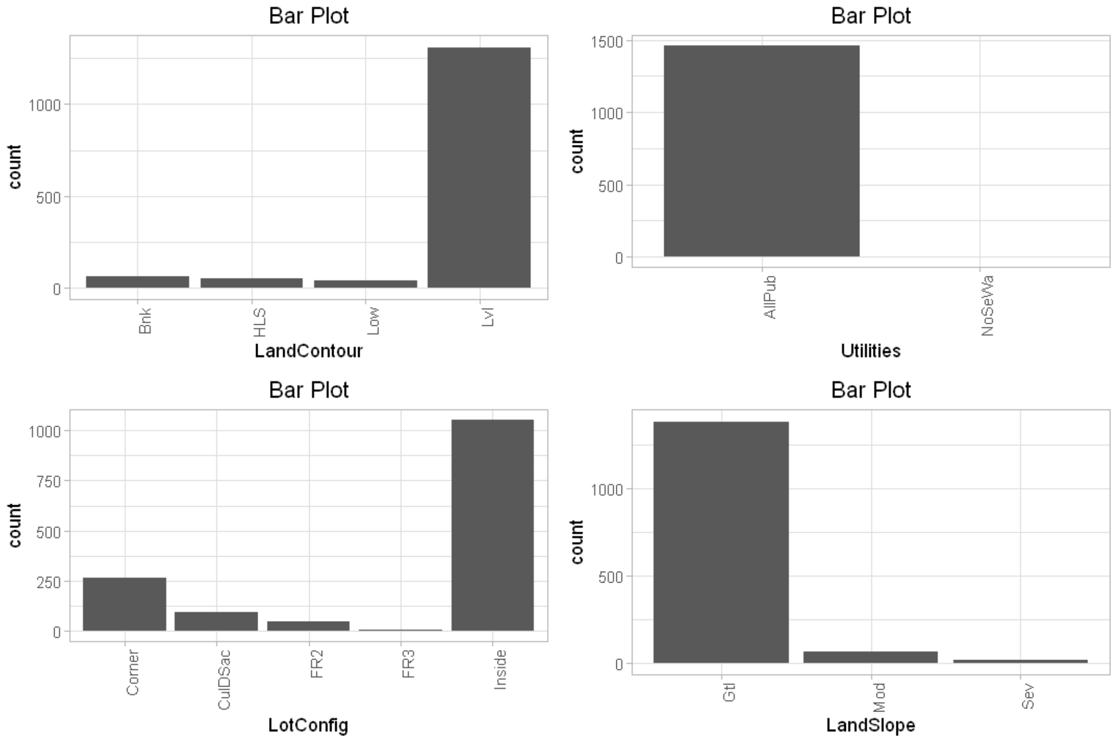
[51]:
```
cat_var <- names(train)[which(sapply(train, is.character))]
cat_car <- c(cat_var, 'BedroomAbvGr', 'HalfBath', '␣
↪KitchenAbvGr','BsmtFullBath', 'BsmtHalfBath', 'MSSubClass')
numeric_var <- names(train)[which(sapply(train, is.numeric))]
```

```
## Creating one training dataset with categorical variable and one with numeric␣
↪variable. We will use this for data visualization.

train1_cat<-train[cat_var]
train1_num<-train[numeric_var]

## Bar plot/Density plot function

## Bar plot function

plotHist <- function(data_in, i)
{
  data <- data.frame(x=data_in[[i]])
  p <- ggplot(data=data, aes(x=factor(x))) + stat_count() +␣
↪xlab(colnames(data_in)[i]) + theme_light() +
    ggtitle("Bar Plot") +
    theme(plot.title = element_text(hjust = 0.5), legend.position=c(0.9,0.7),
    legend.background = element_rect(fill="grey90",size=0.5,␣
↪linetype="solid",colour ="black"),
          axis.text.x = element_text(angle = 90, hjust =1))
  return (p)
}

## Density plot function

plotDen <- function(data_in, i){
  data <- data.frame(x=data_in[[i]], SalePrice = data_in$SalePrice)
  p <- ggplot(data= data) + geom_line(aes(x = x), stat = 'density', size =␣
↪1,alpha = 1.0) +
    xlab(paste0((colnames(data_in)[i]), '\n', 'Skewness:␣
↪',round(skewness(data_in[[i]], na.rm = TRUE), 2))) + ggtitle("Density Plot") +
    theme(plot.title = element_text(hjust = 0.5), legend.position=c(0.9,0.7),
    legend.background = element_rect(fill="grey90",size=0.5,␣
↪linetype="solid",colour ="black"))
  return(p)

}

## Function to call both Bar plot and Density plot function

doPlots <- function(data_in, fun, ii, ncol=3)
{
  pp <- list()
  for (i in ii) {
    p <- fun(data_in=data_in, i=i)
    pp <- c(pp, list(p))
```

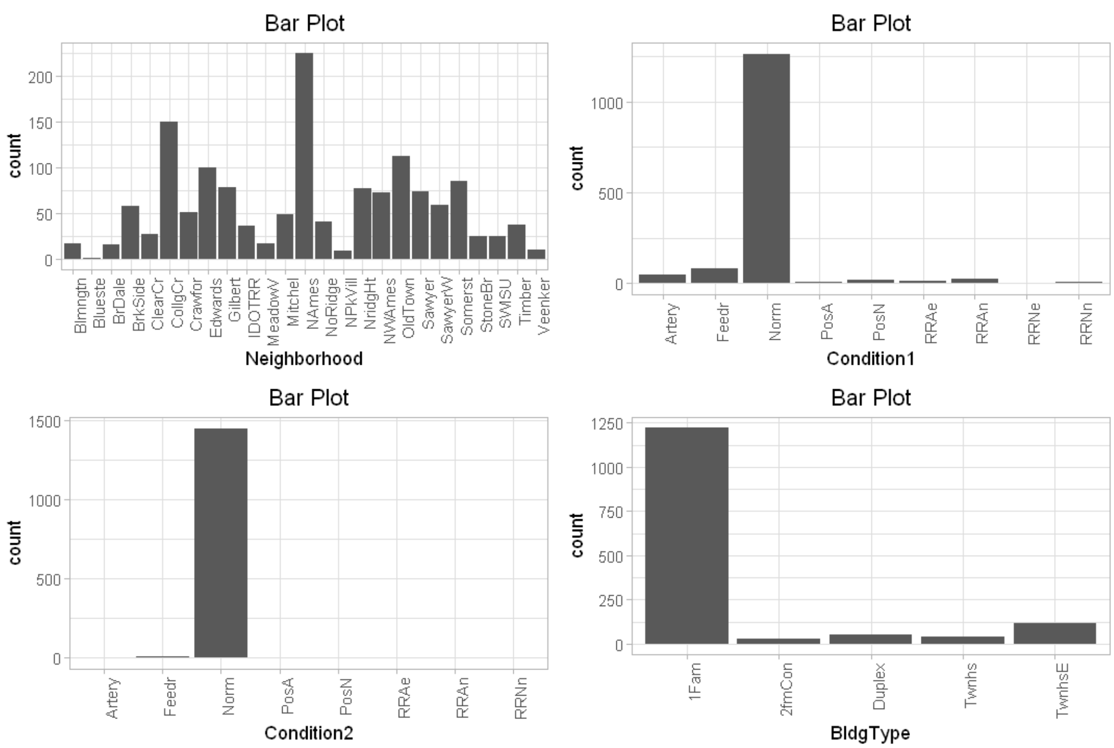
```
  }
  do.call("grid.arrange", c(pp, ncol=ncol))
}


## Barplots for the categorical features

doPlots(train1_cat, fun = plotHist, ii = 1:4, ncol = 2)
```



[28]:
```
doPlots(train1_cat, fun = plotHist, ii  = 5:8, ncol = 2)
```

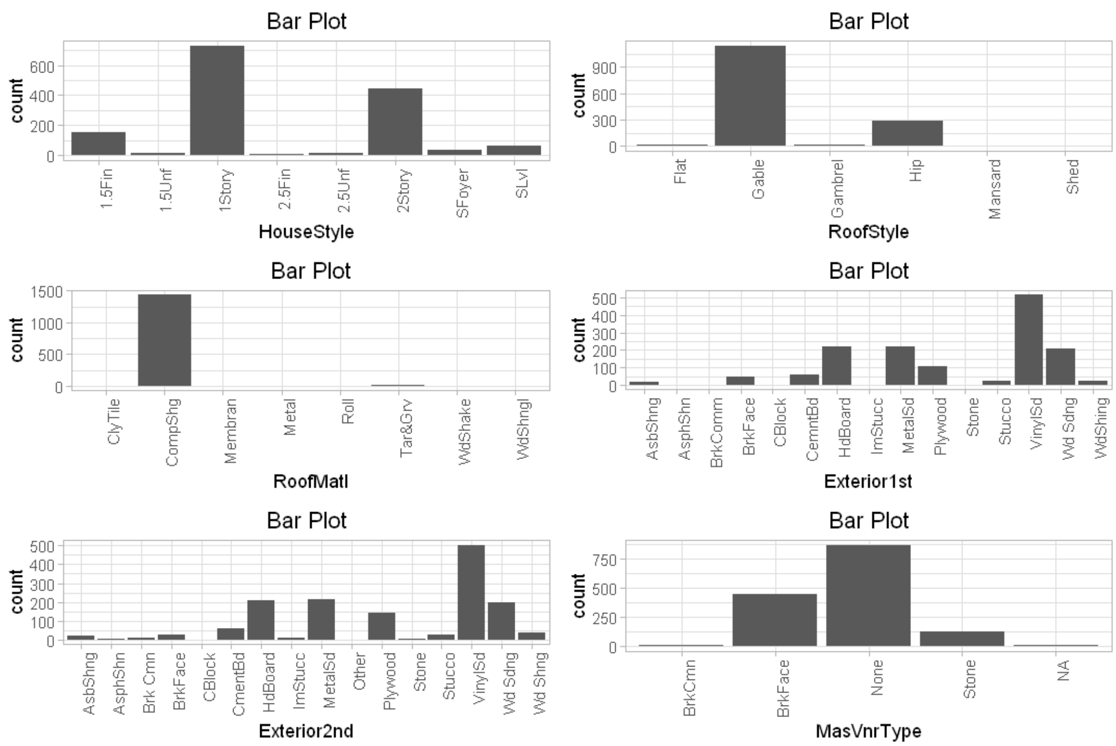



```
[29]: doPlots(train1_cat, fun = plotHist, ii = 9:12, ncol = 2)
```

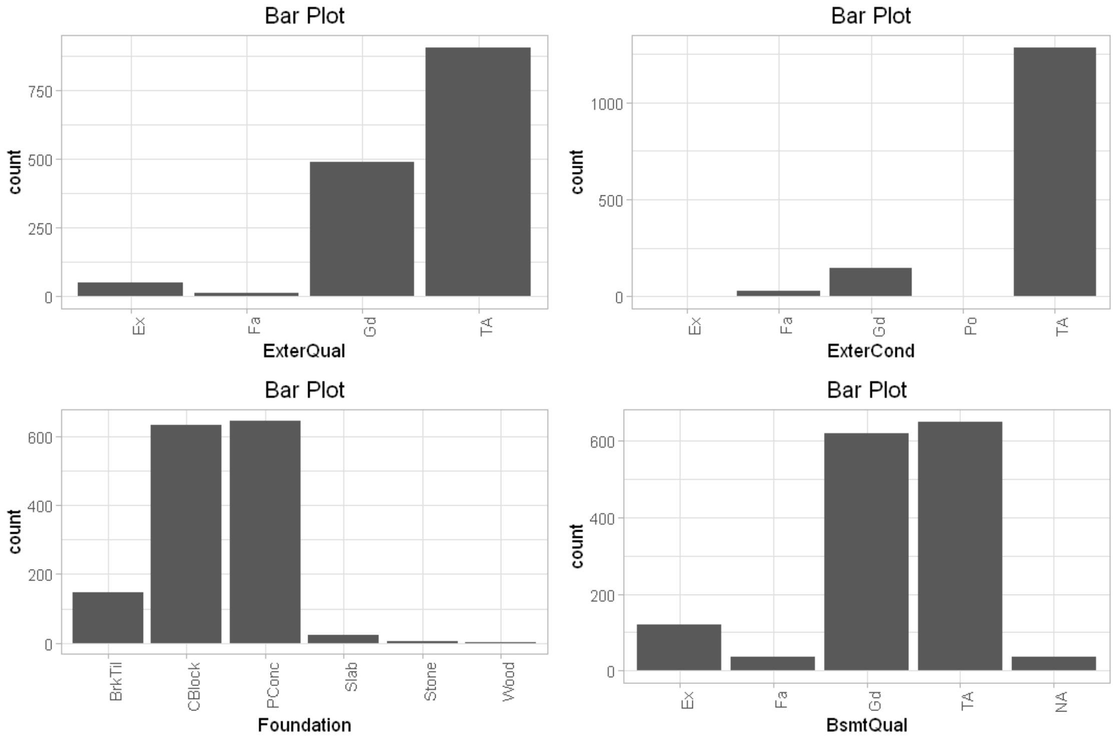


It can be deduced from the graphs above that there a few houses that have severe landslope. The houses with moderate landslope are present in more neighborhoods.

[30]:
```
doPlots(train1_cat, fun = plotHist, ii = 13:18, ncol = 2)
```

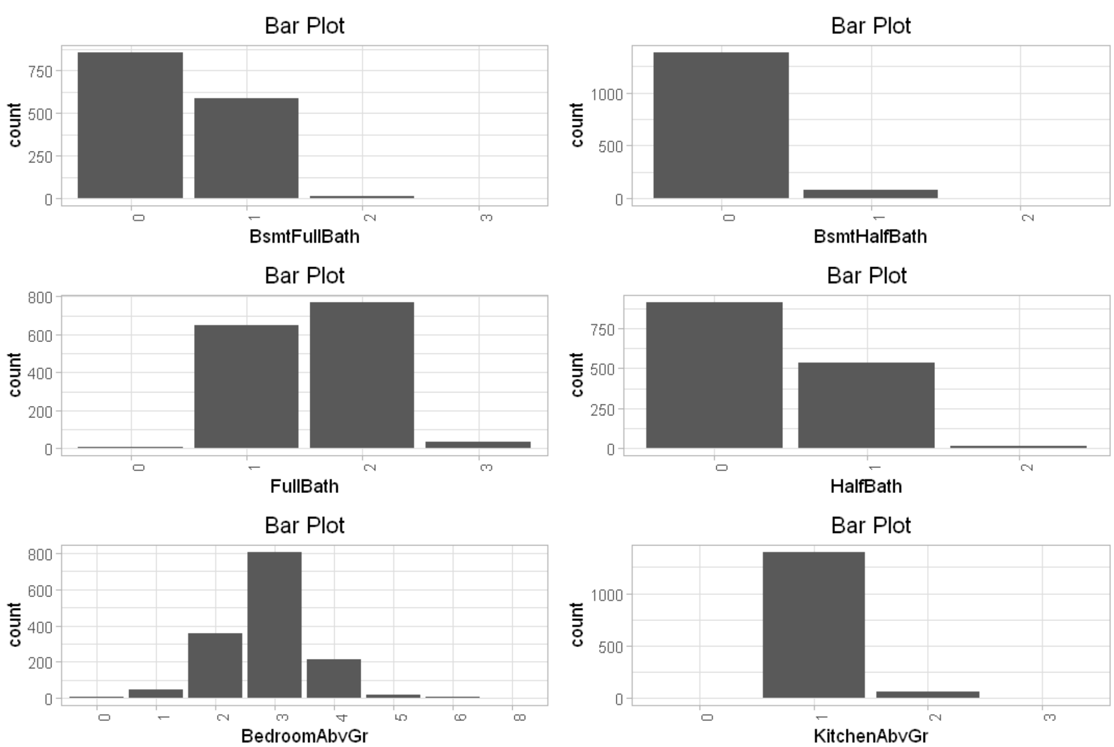


[31]:
```
doPlots(train1_cat, fun = plotHist, ii = 19:22, ncol = 2)
```



[32]:

```
#Histogram for numeric variable
doPlots(train1_num, fun = plotHist, ii = 18:23, ncol = 2)
```



The histograms above show that majority of the houses have 2 full baths, 0 half baths, and have an average of 3 bedrooms.

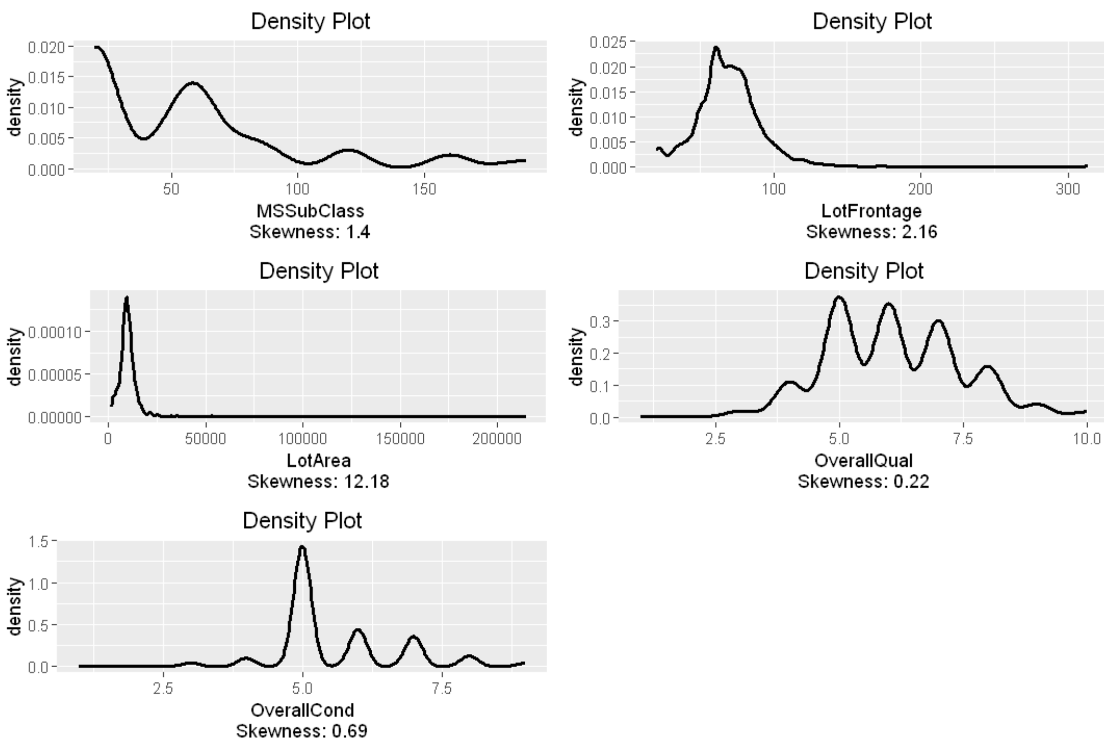
1.14 - Density Plots

Lets create some density plots for numeric variables.

The denisty plot below for YearBuilt shows that the data set contains a mix of new and old houses. It shows a downturn in the number of houses in recent years, possibily due to the housing crisis.

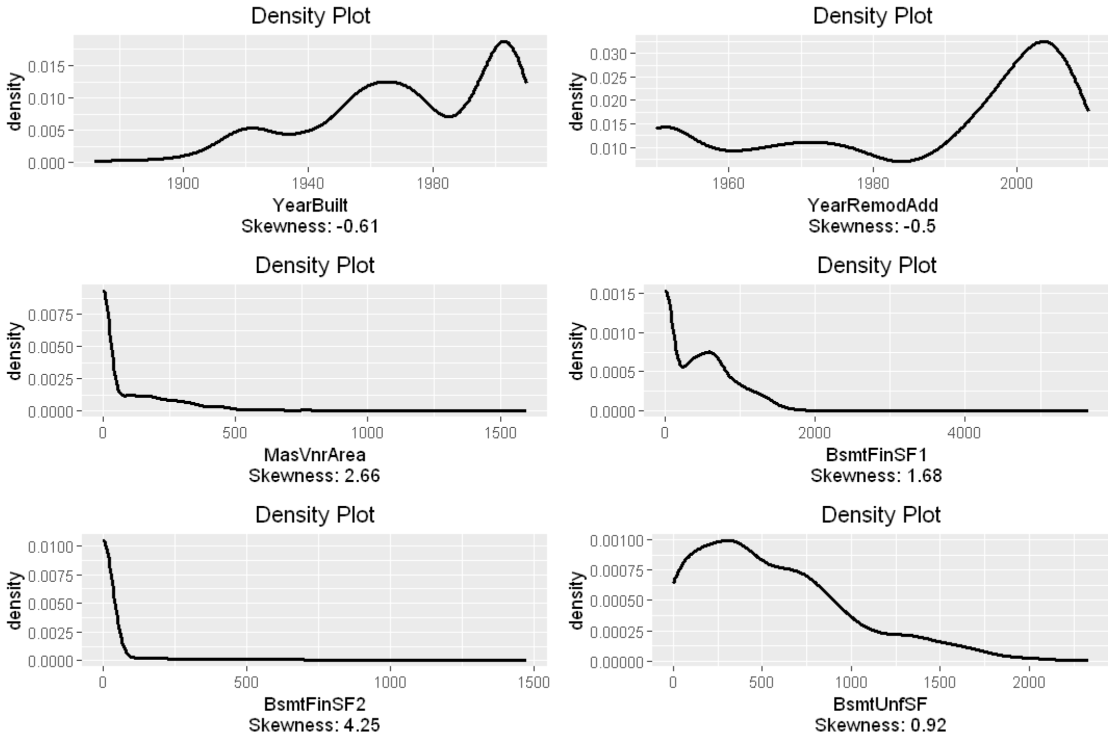
[35]:
```
doPlots(train1_num, fun = plotDen, ii = 2:6, ncol = 2)
```

```
Warning message:
"Removed 259 rows containing non-finite values (stat_density)."
```

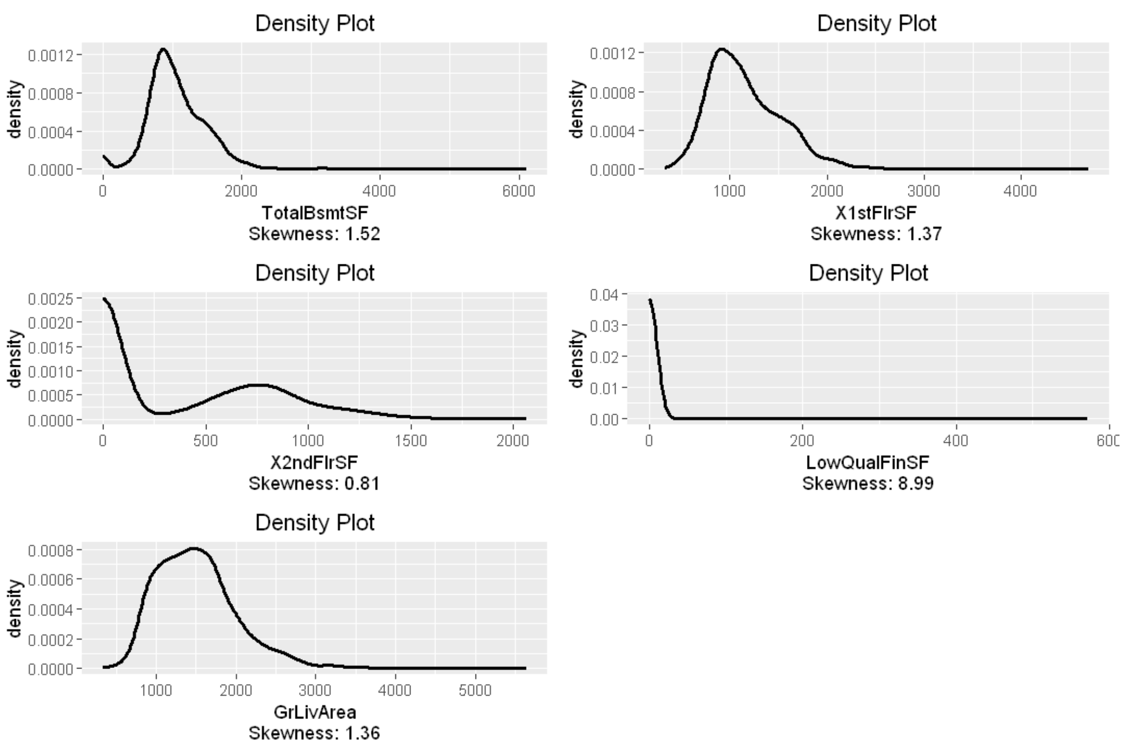


[18]:
```
doPlots(train1_num, fun = plotDen, ii = 7:12, ncol = 2)
```

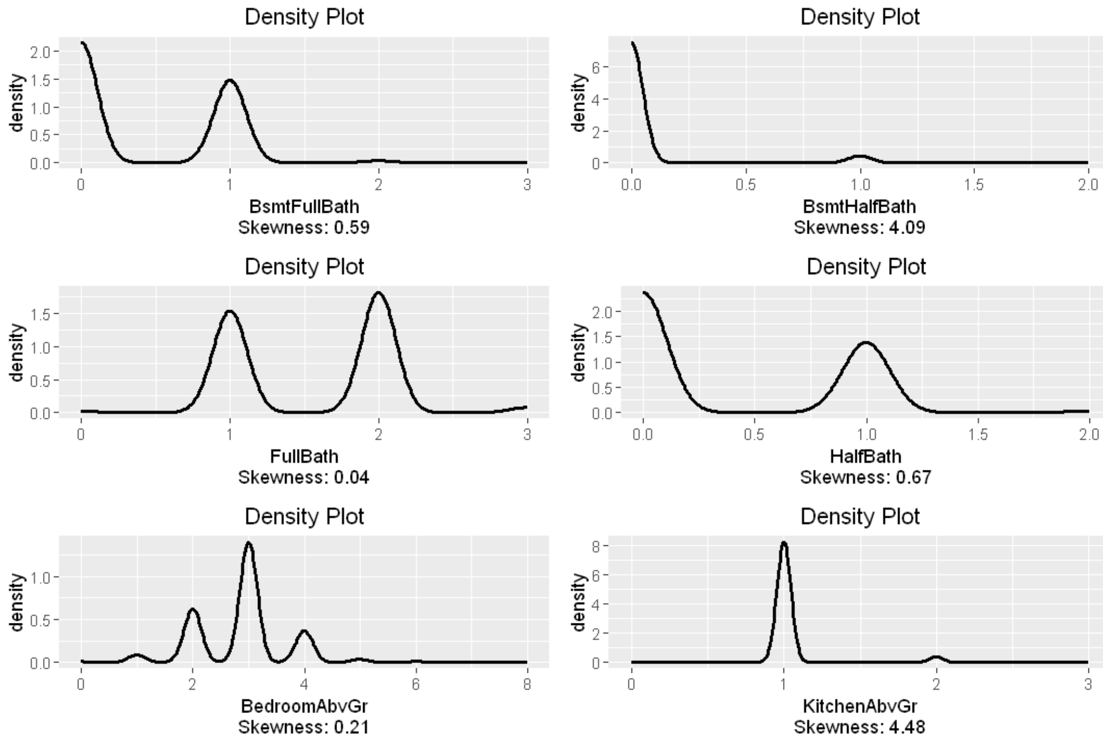
```
Warning message:
"Removed 8 rows containing non-finite values (stat_density)."
```



[19]:
```
doPlots(train1_num, fun = plotDen, ii = 13:17, ncol = 2)
```



[21]:
```
doPlots(train1_num, fun = plotDen, ii = 18:23, ncol = 2)
```



Section 2 - Feature Engineering

We will be doing Feature Engineering of the following 3 categories and will analyze it against Sale Price.

1. Number of Bathrooms
2. House Age
3. Neighbourhood

Later we will create a corelation heatmap.

We will create a feature where we will select the following variables: SalePrice','OverallQual','OverallCond','YearBuilt','ExterCond2','TotalBsmtSF','HeatingQC2'.

Some of these variables needs to be converted to numeric first. We will evaluate quality of the house with ordered levels, such as "Ex", "Fa","Gd", "TA", and "Po", and we will match to numbers: "1","2","3","4", and "5".

[8]:
```
all <- rbind(train, data)
```

[10]:
```
numericVars <- which(sapply(all, is.numeric)) #index vector numeric variables
numericVarNames <- names(numericVars) #saving names vector for use later on
cat('There are', length(numericVars), 'numeric variables')
```

```
There are 38 numeric variables
```

[11]:
```
all_numVar <- all[, numericVars]
cor_numVar <- cor(all_numVar, use="pairwise.complete.obs") #correlations of all␣
↪numeric variables
```

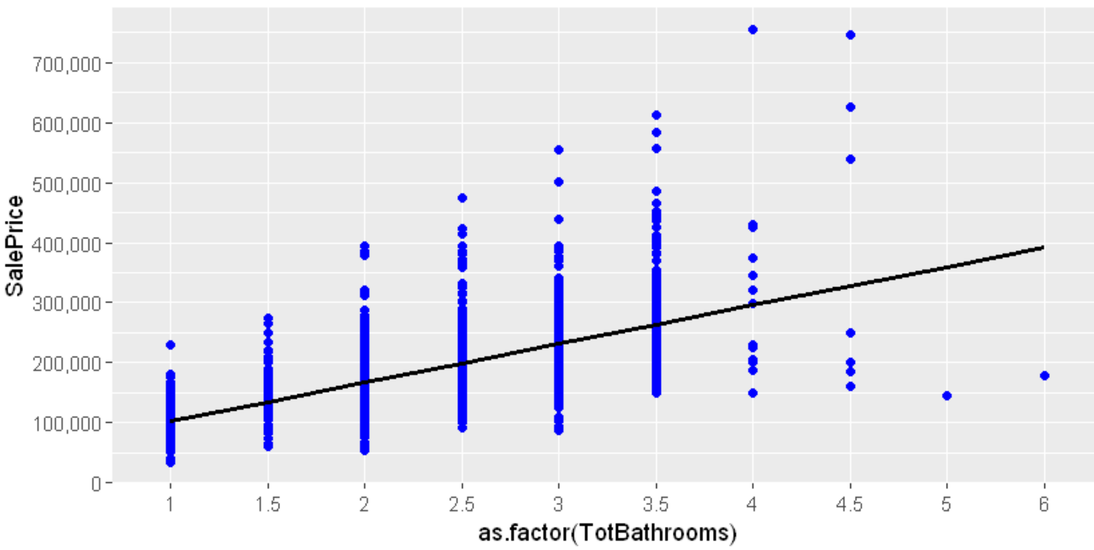
2.1 - Number of Bathrooms

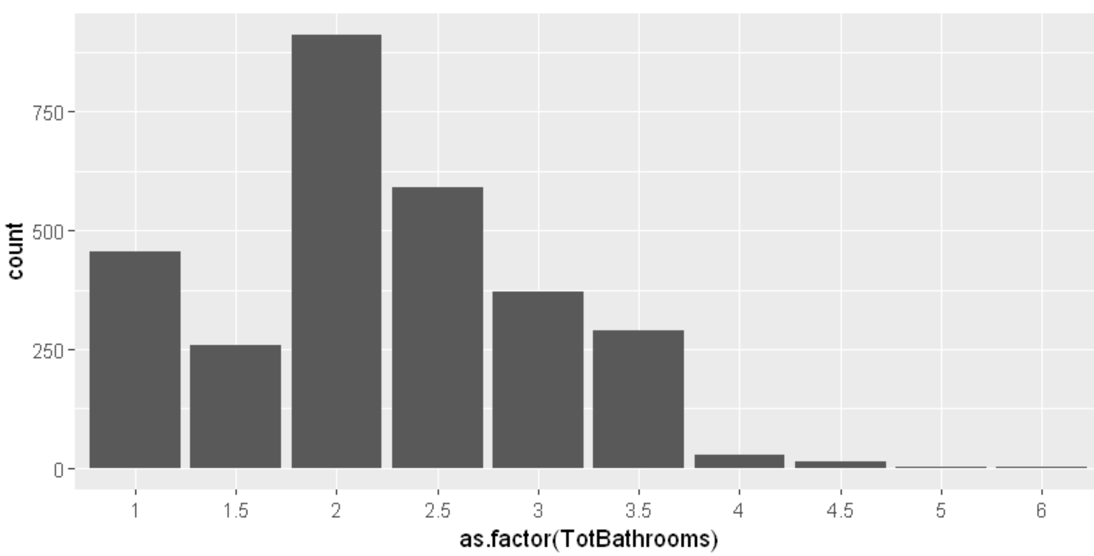
There are 4 bathroom variables. Individually, these variables are not very important. However, assume if I add them up into one predictor, this predictor is likely to become a strong one.

[14]:
```
all$TotBathrooms <- all$FullBath + (all$HalfBath*0.5) + all$BsmtFullBath +␣
↪(all$BsmtHalfBath*0.5)
```

[15]:
```
tb1 <- ggplot(data=all[!is.na(all$SalePrice),], aes(x=as.factor(TotBathrooms),␣
↪y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE,␣
↪color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
tb2 <- ggplot(data=all, aes(x=as.factor(TotBathrooms))) +
        geom_histogram(stat='count')
grid.arrange(tb1, tb2)
```

```
Warning message:
"Ignoring unknown parameters: binwidth, bins, pad"
```





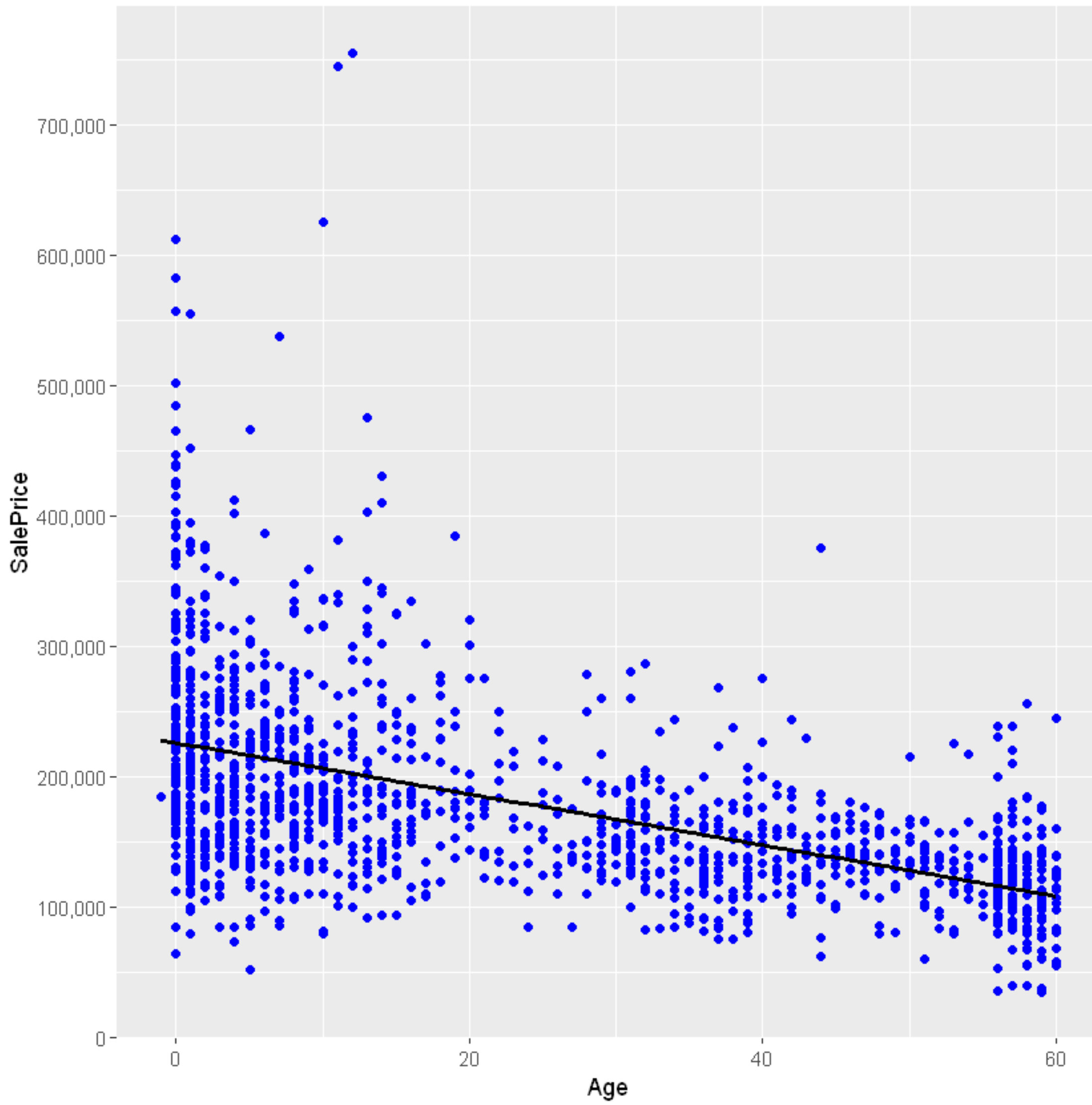
As you can see in the first graph, there now seems to be a clear correlation. The frequency distribution of Bathrooms in all data is shown in the second graph.

2.2 - House Age

[16]:
```
all$Remod <- ifelse(all$YearBuilt==all$YearRemodAdd, 0, 1) #0=No Remodeling,␣
↪1=Remodeling
all$Age <- as.numeric(all$YrSold)-all$YearRemodAdd
```

[17]:
```
ggplot(data=all[!is.na(all$SalePrice),], aes(x=Age, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE,␣
↪color="black", aes(group=1)) +
```

```
    scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```

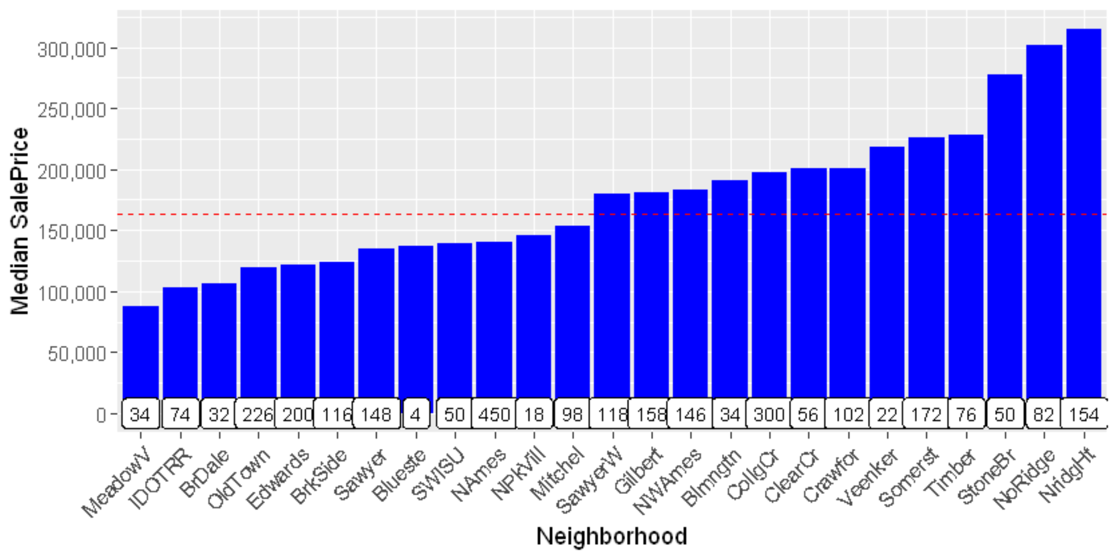


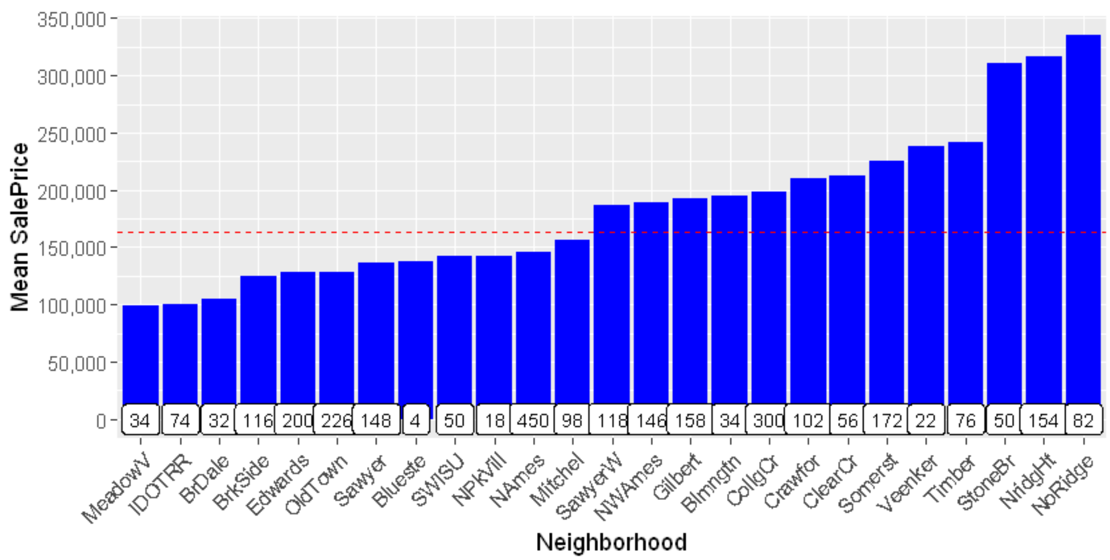
As expected, the graph shows a negative correlation with Age (old house are worth less).

2.3 - Neighbourhood

[18]:
```
nb1 <- ggplot(all[!is.na(all$SalePrice),], aes(x=reorder(Neighborhood,
 ↪SalePrice, FUN=median), y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
 ↪labs(x='Neighborhood', y='Median SalePrice') +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
```

```
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePrice
nb2 <- ggplot(all[!is.na(all$SalePrice),], aes(x=reorder(Neighborhood, SalePrice, FUN=mean), y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "mean", fill='blue') + labs(x='Neighborhood', y="Mean SalePrice") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePrice
grid.arrange(nb1, nb2)
```





As we can see from the graphs above that 3 neighborhoods are relatively cheap.

2.4 - Correlation Heatmap

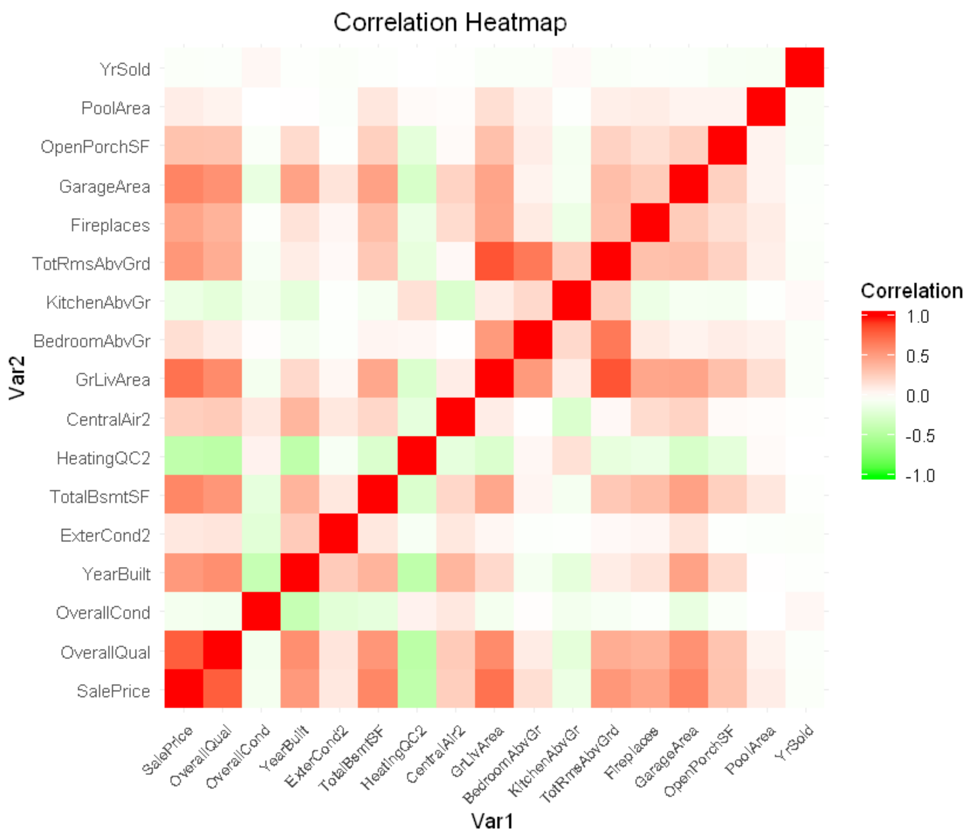
[25]:
```
# convert factor to numeric

data$ExterCond2 <- as.numeric(factor(data$ExterCond,
                                  levels = c("Ex", "Fa","Gd", "TA","Po"),
                                  labels = c(5,2,4,3,1) ,ordered = TRUE))
data$HeatingQC2 <- as.numeric(factor(data$HeatingQC,
                                  levels = c("Ex", "Fa","Gd", "TA","Po"),
                                  labels = c(5,2,4,3,1) ,ordered = TRUE))
data$CentralAir2 <- as.numeric(factor(data$CentralAir,
                                  levels = c("N", "Y"),
                                  labels = c(0,1) ,ordered = TRUE))
```

[26]:
```
#select variables that be used for model buidling and heat map

model_var <- c('SalePrice',
                'OverallQual','OverallCond','YearBuilt','ExterCond2',
                'TotalBsmtSF','HeatingQC2',
                'CentralAir2','GrLivArea','BedroomAbvGr','KitchenAbvGr',
                'TotRmsAbvGrd','Fireplaces',
                'GarageArea','OpenPorchSF','PoolArea',
                 'YrSold')
heat <- data[,model_var]
```

[27]:
```
#Ploting Correlation Heatmap for SalePrice

options(repr.plot.width=8, repr.plot.height=6)
library(ggplot2)
library(reshape2)
qplot(x=Var1, y=Var2, data=melt(cor(heat, use="p")), fill=value, geom="tile") +
   scale_fill_gradient2(low = "green", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Correlation") +
   theme_minimal()+
   theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 8, hjust = 1))+
   coord_fixed()+
   ggtitle("Correlation Heatmap") +
   theme(plot.title = element_text(hjust = 0.4))
```

```
Attaching package: 'reshape2'
```

```
The following objects are masked from 'package:data.table':

    dcast, melt
```



In this graph above, Red indicates perfect positive correlation and Green indicates perfect negative correlation.

As we can see, there are several variables should be paid attention to: GarageArea, Fireplaces, TotRmsAbvGrd, GrLivArea, HeatingQC, TotalBsmtSF and YearBuild.

Section 3 - Model Building - Training and Testing

3.1 - Linear Regression Model

We are selecting the following 16 variables to fit into this model. Variables include:

SalePrice, OverallQual, OverallCond, YearBuilt, ExterQual2, ExterCond2, TotalBsmtSF, HeatingQC2, CentralAir2, GrLivArea, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageArea, OpenPorchSF, PoolArea,YrSold

In Linear Regresion Model, the relationships between Dependent and Indepedent Variables is expressed by equation with coefficients. The aim of this model is to minimize the sum of the squared residuals.

Steps: 1- We will select variables and tranfer SalePrice into log term. 2- We will divide dataset into two parts. Training and Validation. 3- Run regression. 4- Check for accuracy.

[28]:
```
#prediction of lm
#build model dataset for linear regression
model_lin <- data[, model_var]
model_lin$lSalePrice <- log(model_lin$SalePrice)
```

[29]:
```
#partition data

set.seed(10000)
data.index <- sample(c(1:dim(model_lin)[1]), dim(model_lin)[1]*0.8)
model_lin_data = model_lin[data.index,]
model_lin_valid <- model_lin[-data.index,]
```

[30]:
```
linreg <- lm(lSalePrice~.-SalePrice, data = model_lin_data)
summary(linreg)
```

```
Call:
lm(formula = lSalePrice ~ . - SalePrice, data = model_lin_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98613 -0.07164  0.00209  0.08015  0.55020

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.750e+01  7.114e+00   2.460  0.01402 *
OverallQual   8.057e-02  5.757e-03  13.996  < 2e-16 ***
OverallCond   5.664e-02  4.893e-03  11.576  < 2e-16 ***
YearBuilt     3.177e-03  2.422e-04  13.120  < 2e-16 ***
ExterCond2    2.627e-02  1.171e-02   2.244  0.02503 *
TotalBsmtSF   1.115e-04  1.344e-05   8.301 2.86e-16 ***
HeatingQC2   -1.828e-02  4.076e-03  -4.486 7.99e-06 ***
CentralAir2   6.343e-02  2.300e-02   2.757  0.00592 **
GrLivArea     2.026e-04  1.946e-05  10.414  < 2e-16 ***
BedroomAbvGr -4.556e-03  8.486e-03  -0.537  0.59143
KitchenAbvGr -6.642e-02  2.534e-02  -2.621  0.00887 **
TotRmsAbvGrd  1.726e-02  6.232e-03   2.770  0.00570 **
Fireplaces    6.900e-02  8.546e-03   8.074 1.70e-15 ***
GarageArea    2.384e-04  2.956e-05   8.064 1.83e-15 ***
OpenPorchSF   1.953e-05  7.922e-05   0.247  0.80529
PoolArea     -7.814e-04  1.405e-04  -5.561 3.34e-08 ***
YrSold       -6.645e-03  3.539e-03  -1.878  0.06069 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1585 on 1151 degrees of freedom
Multiple R-squared:  0.8467,Adjusted R-squared:  0.8446
F-statistic: 397.3 on 16 and 1151 DF,  p-value: < 2.2e-16
```

[31]:
```
install.packages("forecast")
```

```
Installing package into 'C:/Users/Munazzam/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
```

```
package 'forecast' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Munazzam\AppData\Local\Temp\RtmpwXZmsS\downloaded_packages
```

[32]:
```
library(forecast)

#use predict() to make prediction on a new set

pred1 <- predict(linreg,model_lin_valid,type = "response")
residuals <- model_lin_valid$lSalePrice - pred1
linreg_pred <- data.frame("Predicted" = pred1, "Actual" =
 ↪model_lin_valid$lSalePrice, "Residual" = residuals)
accuracy(pred1, model_lin_valid$lSalePrice)
```

```
Warning message:
"package 'forecast' was built under R version 3.6.3"Registered S3 method
overwritten by 'xts':
  method     from
  as.zoo.xts zoo
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

	ME	RMSE	MAE	MPE	MAPE
Test set	0.007261273	0.1538444	0.1075528	0.04271564	0.9029266

ME: Mean Error

RMSE: Root Mean Squared Error

MAE: Mean Absolute Error

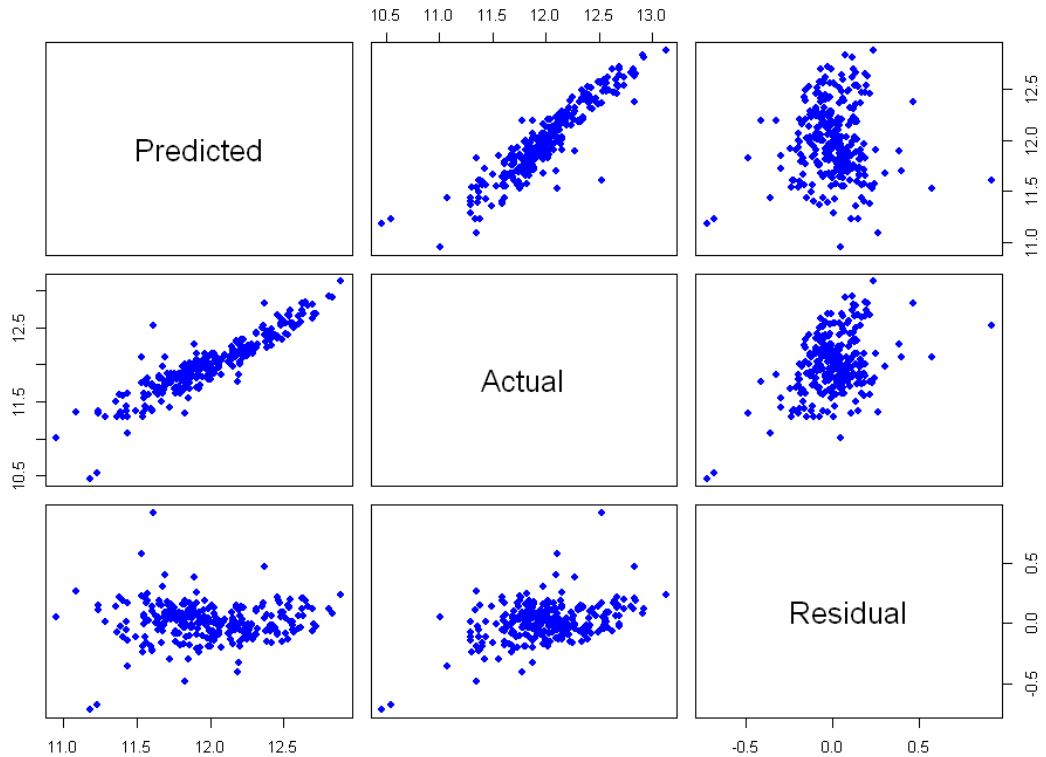
MPE: Mean Percentage Error

MAPE: Mean Absolute Percentage Error

As we can see from the results above, RMSE value is very small.

RMSE values < 0.1 is very satisfactory. RMSE value 0.5 reflects the poor ability of the model to accurately predict the data.

[38]:
```r
#Scatter Plot
plot(linreg_pred, pch = 16, col = "blue")
```



Scatter plots are one of the richest form of data visualization. You can tell pretty much everything from it. Ideally, all your points should be close to a regressed diagonal line.

As we can see from the plot above all the actual data lies between 11 and 13. and so are the predictions.

3.2 - Random Forest

In Random Forest, idea is to:

1- Draw multiple random samples with replacement from the data. 2- Using random subset of predictors at each stage, fit a classification (regression) tree to each sample and create a forest. 3- Combine predictions/classifications from each tree to get improved predictions.

[42]:
```r
library(randomForest)
RF <- randomForest(lSalePrice ~.-SalePrice, data = model_lin_data,
```

```
                          importance =TRUE,ntree=500,nodesize=7, na.action=na.roughfix)
```

[43]:
```
# variable importance plot from Random Forest

options(repr.plot.width=9, repr.plot.height=6)
varImpPlot(RF, type=1)
```

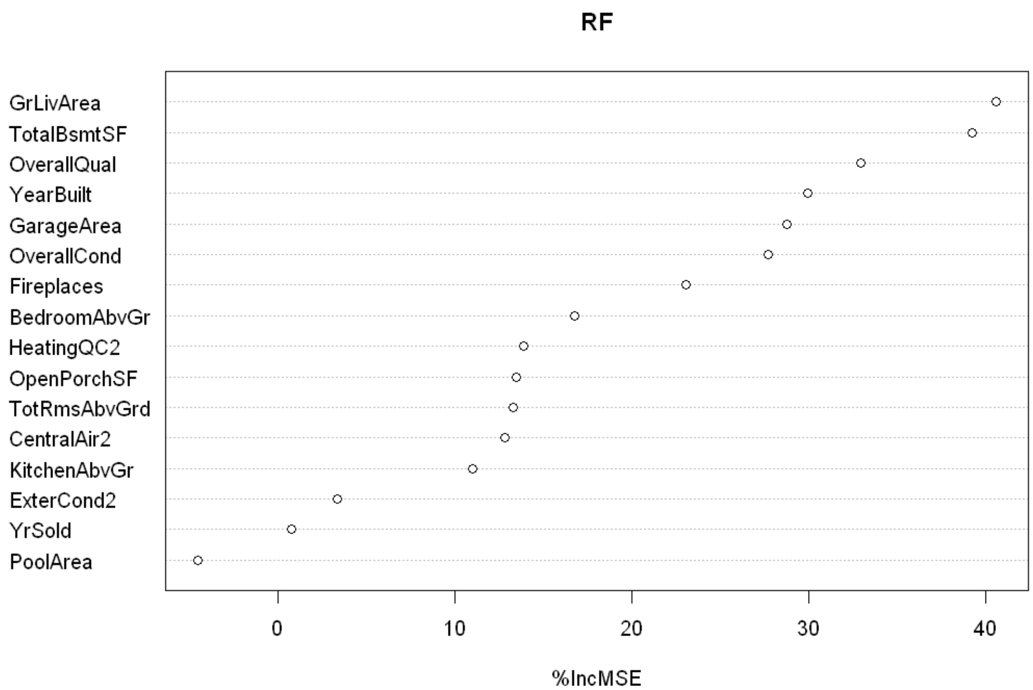


Figure above shows the variable importance plots generated from the random forest model for SalePrice. We see GrLivArea and TotalBsmtSF has the highest score.

[44]:
```
#prediction

rf.pred <- predict(RF, newdata=model_lin_valid )
accuracy(rf.pred, model_lin_valid$lSalePrice)
```

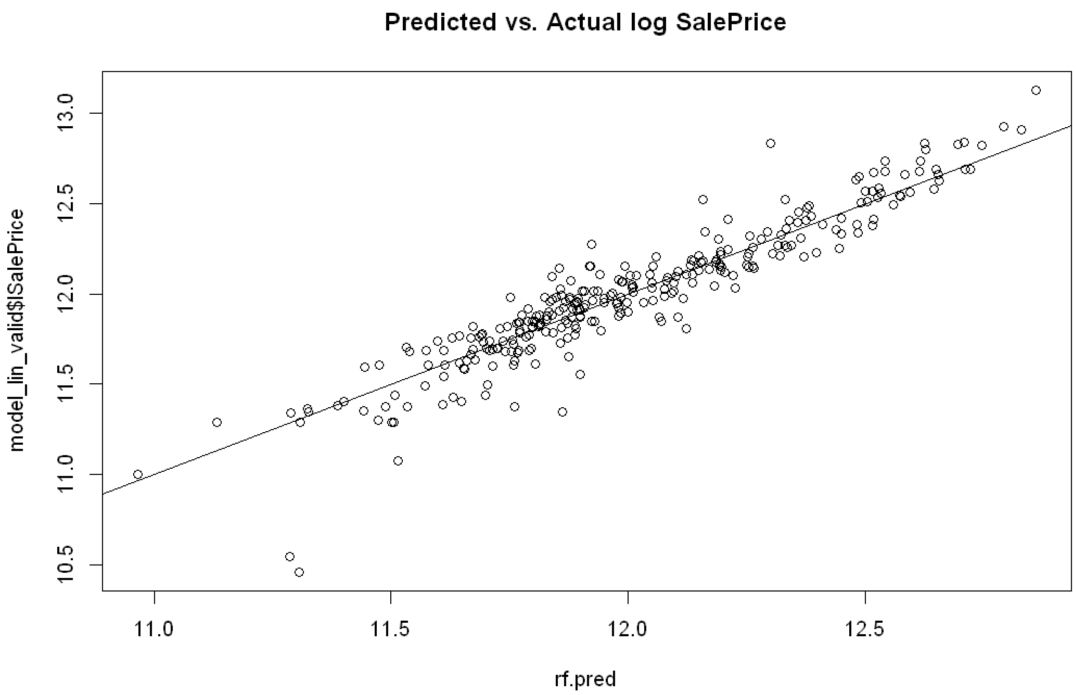
	ME	RMSE	MAE	MPE	MAPE
Test set	-0.0004281985	0.1384497	0.09486207	-0.02296528	0.7980985

As we can see from the results above, RMSE value is very small.

RMSE values < 0.1 is very satisfactory. RMSE value 0.5 reflects the poor ability of the model to accurately predict the data.

Graph below shows predicted vs actual Sale Price.

[45]:
```
plot(rf.pred, model_lin_valid$lSalePrice, main = "Predicted vs. Actual log␣
↪SalePrice")
abline(0,1)
```



Thank You
