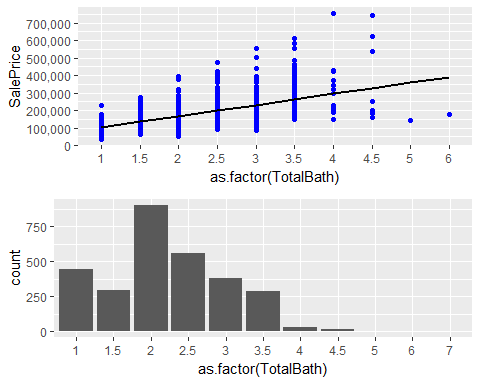
Feature Engineering

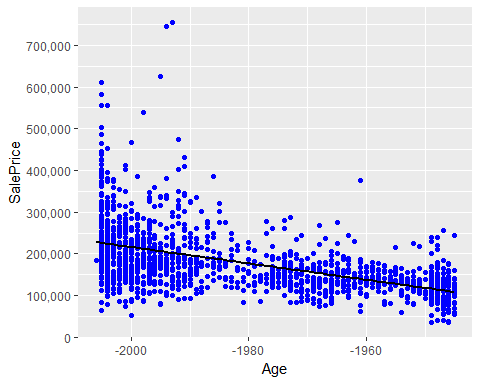
##Feature Engineering  
  
##Total Bathrooms  
df.combined$TotalBath <- df.combined$FullBath + (df.combined$HalfBath\*0.5) + df.combined$BsmtFullBath + (df.combined$BsmtHalfBath\*0.5)  
  
tb1 <- ggplot(data = df.combined[!is.na(df.combined$SalePrice), ], aes(x = as.factor(TotalBath), y = SalePrice)) +  
 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 = df.combined, aes(x = as.factor(TotalBath))) +  
 geom\_histogram(stat = 'count')

## Warning: Ignoring unknown parameters: binwidth, bins, pad

grid.arrange(tb1, tb2)



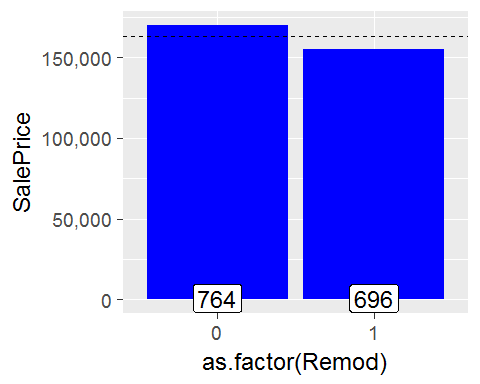
##Age, Remodel & New Variables  
df.combined$Remod <- ifelse(df.combined$YearBuilt == df.combined$YearRemodAdd, 0, 1) ## 0 = NA Remodel, 1 = Remodel  
df.combined$Age <- as.numeric(df.combined$YrSold) - df.combined$YearRemodAdd  
  
ggplot(data = df.combined[!is.na(df.combined$SalePrice), ], aes(x = 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)



cor(df.combined$SalePrice[!is.na(df.combined$SalePrice)], df.combined$Age[!is.na(df.combined$SalePrice)])

## [1] -0.5090787

ggplot(df.combined[!is.na(df.combined$SalePrice),], aes(x = as.factor(Remod), y = SalePrice)) +  
 geom\_bar(stat = 'summary', fun.y = "median", fill = 'blue') +  
 geom\_label(stat = "count", aes(label = ..count.., y = ..count..), size = 6) +  
 scale\_y\_continuous(breaks = seq(0, 800000, by = 50000), labels = comma) +  
 theme\_grey(base\_size = 18) +  
 geom\_hline(yintercept = 163000, linetype = "dashed")

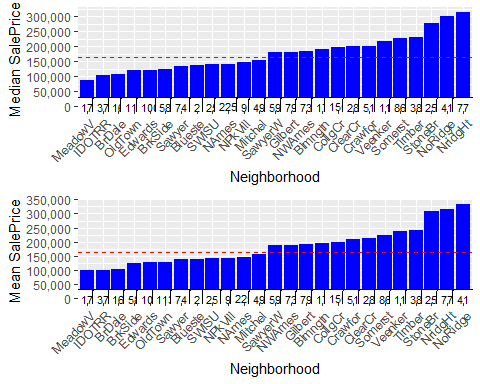


df.combined$New <- ifelse(df.combined$YrSold == df.combined$YearBuilt, 1, 0) ##1 = New, 0 = Not New  
table(df.combined$New)

##   
## 0 1   
## 2803 116

##Factorize YrSold  
df.combined$YrSold <- as.factor(df.combined$YrSold)

##Binning Neighborhoods  
nb1 <- ggplot(df.combined[!is.na(df.combined$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")  
  
nb2 <- ggplot(df.combined[!is.na(df.combined$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")  
  
grid.arrange(nb1, nb2)



##Categorizing Neighborhood Wealth (Poor - 0, Middle - 1, Wealthy - 2)  
df.combined$NeighborhoodWealth[df.combined$Neighborhood %in% c('StoneBr', 'NridgHt', 'NoRidge')] <- 2  
df.combined$NeighborhoodWealth[!df.combined$Neighborhood %in% c('MeadowV', 'IDOTRR', 'BrDale', 'StoneBr', 'NridgHt', 'NoRidge')] <- 1  
df.combined$NeighborhoodWealth[df.combined$Neighborhood %in% c('MeadowV', 'IDOTRR', 'BrDale')] <- 0  
table(df.combined$NeighborhoodWealth)

##   
## 0 1 2   
## 160 2471 288

sum(table(df.combined$NeighborhoodWealth))

## [1] 2919

## Total SqFt  
df.combined$TotalSqFt <- df.combined$GrLivArea + df.combined$TotalBsmtSF  
  
cor(df.combined$SalePrice, df.combined$TotalSqFt, use = "pairwise.complete.obs")

## [1] 0.7789588

##Porch Variables  
df.combined$TotalPorchSqFt <- df.combined$OpenPorchSF + df.combined$EnclosedPorch + df.combined$X3SsnPorch + df.combined$ScreenPorch  
  
cor(df.combined$SalePrice, df.combined$TotalPorchSqFt, use = "pairwise.complete.obs")

## [1] 0.1957389