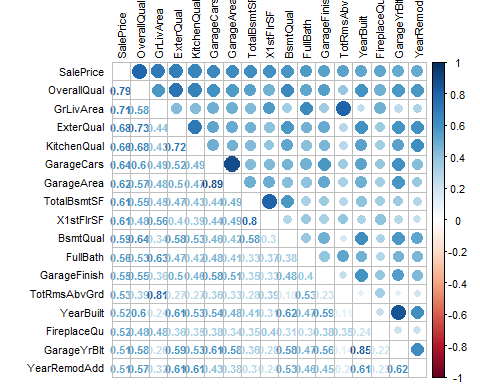
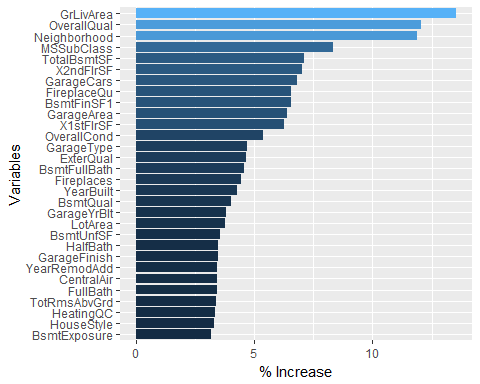
EDA & Stats Work

##Level Set High Correlations  
df.numeric.Vars <- df.combined[, numeric.Vars]  
correlation.numeric.Vars <- cor(df.numeric.Vars, use = "pairwise.complete.obs")  
correlation.sorted <- as.matrix(sort(correlation.numeric.Vars[, 'SalePrice'], decreasing = TRUE))  
  
correlation.high <- names(which(apply(correlation.sorted, 1, function(x) abs(x) > 0.5)))  
correlation.numeric.Vars <- correlation.numeric.Vars[correlation.high, correlation.high]  
  
corrplot.mixed(correlation.numeric.Vars, tl.col = "black", tl.pos = "lt", tl.cex = 0.7, cl.cex = 0.7, number.cex = 0.7)



##Quick RandomForest to Identify Most Important Variables(Numeric & Categorical)  
set.seed(2018)  
RandomForest.draft <- randomForest(x = df.combined[1:1460, -79], y = df.combined$SalePrice[1:1460], ntree = 100, importance = TRUE)  
RandomForest.important <- importance(RandomForest.draft)  
df.important <- data.frame(Variables = row.names(RandomForest.important), MSE = RandomForest.important[, 1])  
df.important <- df.important[order(df.important$MSE, decreasing = TRUE),]  
  
##Visualize Quick RandomForest (df.important)  
ggplot(df.important[1:30,], aes(x = reorder(Variables, MSE), y = MSE, fill = MSE)) +  
 geom\_bar(stat = 'identity') +  
 labs(x = 'Variables', y = '% Increase') +  
 coord\_flip() +  
 theme(legend.position = "none")

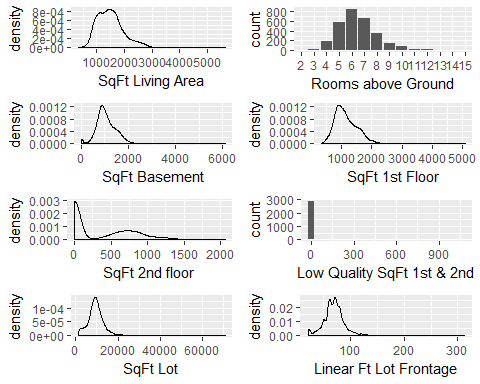


##EDA of GrLivArea & Surface Related Variables  
s1 <- ggplot(data = df.combined, aes(x = GrLivArea)) +  
 geom\_density() + labs(x = 'SqFt Living Area')  
s2 <- ggplot(data = df.combined, aes(x = as.factor(TotRmsAbvGrd))) +  
 geom\_histogram(stat = 'count') + labs(x = 'Rooms above Ground')

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

s3 <- ggplot(data = df.combined, aes(x = X1stFlrSF)) +  
 geom\_density() + labs(x = 'SqFt 1st Floor')  
s4 <- ggplot(data = df.combined, aes(x = X2ndFlrSF)) +  
 geom\_density() + labs(x='SqFt 2nd floor')  
s5 <- ggplot(data = df.combined, aes(x = TotalBsmtSF)) +  
 geom\_density() + labs(x = 'SqFt Basement')  
s6 <- ggplot(data = df.combined[df.combined$LotArea < 100000,], aes(x=LotArea)) +  
 geom\_density() + labs(x = 'SqFt Lot')  
s7 <- ggplot(data = df.combined, aes(x=LotFrontage)) +  
 geom\_density() + labs(x = 'Linear Ft Lot Frontage')  
s8 <- ggplot(data = df.combined, aes(x = LowQualFinSF)) +  
 geom\_histogram() + labs(x = 'Low Quality SqFt 1st & 2nd')  
  
d.layout <- matrix(c(1, 2, 5, 3, 4, 8, 6, 7), 4, 2, byrow=TRUE)  
multiplot(s1, s2, s3, s4, s5, s6, s7, s8, layout = d.layout)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



##Investigating Correlation between GrLivArea to X1stFlrSF, X2ndFlrSF & LowQualFinSF  
cor(df.combined$GrLivArea, (df.combined$X1stFlrSF + df.combined$X2ndFlrSF + df.combined$LowQualFinSF))

## [1] 1

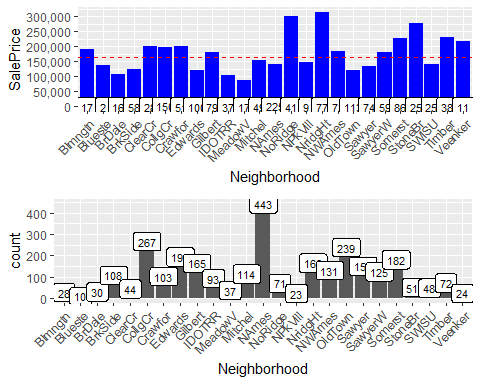
head(df.combined[df.combined$LowQualFinSF > 0, c('GrLivArea', 'X1stFlrSF', 'X2ndFlrSF', 'LowQualFinSF')])

## GrLivArea X1stFlrSF X2ndFlrSF LowQualFinSF  
## 52 1176 816 0 360  
## 89 1526 1013 0 513  
## 126 754 520 0 234  
## 171 1382 854 0 528  
## 186 3608 1518 1518 572  
## 188 1656 808 704 144

##EDA of Neighborhood Variable  
n1 <-ggplot(df.combined[!is.na(df.combined$SalePrice), ], aes(x = Neighborhood, y = SalePrice)) +  
 geom\_bar(stat = 'summary', fun.y = "median", fill = 'blue') +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 scale\_y\_continuous(breaks = seq(0, 800000, by = 50000), labels = scales::comma) +  
 geom\_label(stat = "count", aes(label = ..count.., y = ..count..), size = 3) +  
 geom\_hline(yintercept = 163000, linetype = "dashed", color = "red")  
  
n2 <- ggplot(data = df.combined, aes(x = Neighborhood)) +  
 geom\_histogram(stat = 'count') +  
 geom\_label(stat = "count", aes(label = ..count.., y = ..count..), size = 3) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

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

grid.arrange(n1, n2)



##EDA of OverallQuall & Quality Variables  
q1 <- ggplot(data = df.combined, aes(x = as.factor(OverallQual))) +  
 geom\_histogram(stat = 'count')

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

q2 <- ggplot(data = df.combined, aes(x=as.factor(ExterQual))) +  
 geom\_histogram(stat='count')

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

q3 <- ggplot(data = df.combined, aes(x=as.factor(BsmtQual))) +  
 geom\_histogram(stat='count')

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

q4 <- ggplot(data = df.combined, aes(x=as.factor(KitchenQual))) +  
 geom\_histogram(stat='count')

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

q5 <- ggplot(data = df.combined, aes(x=as.factor(GarageQual))) +  
 geom\_histogram(stat='count')

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

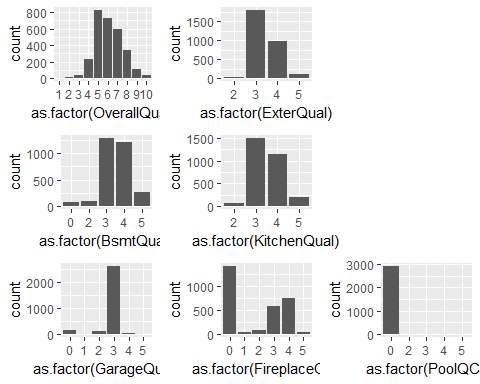
q6 <- ggplot(data = df.combined, aes(x=as.factor(FireplaceQu))) +  
 geom\_histogram(stat='count')

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

q7 <- ggplot(data = df.combined, aes(x=as.factor(PoolQC))) +  
 geom\_histogram(stat='count')

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

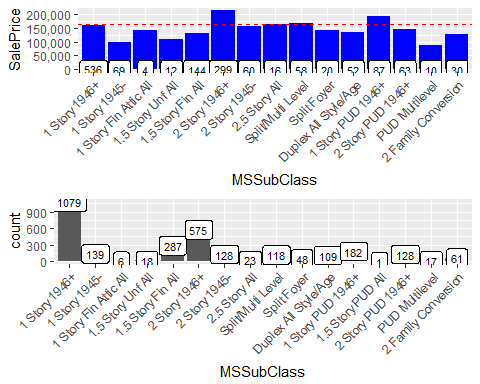
layout <- matrix(c(1,2,8,3,4,8,5,6,7),3,3,byrow=TRUE)  
multiplot(q1, q2, q3, q4, q5, q6, q7, layout=layout)



##EDA of MSSubCLass  
ms1 <- ggplot(df.combined[!is.na(df.combined$SalePrice),], aes(x = MSSubClass, y = SalePrice)) +  
 geom\_bar(stat = 'summary', fun.y = "median", fill = 'blue') +  
 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")   
  
ms2 <- ggplot(data = df.combined, aes(x = MSSubClass)) +  
 geom\_histogram(stat = 'count') +  
 geom\_label(stat = "count", aes(label = ..count.., y = ..count..), size = 3) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

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

grid.arrange(ms1, ms2)



##EDA of Garage Variables  
g1 <- ggplot(data = df.combined[df.combined$GarageCars !=0,], aes(x = GarageYrBlt)) +   
 geom\_histogram() +   
 xlim(1890, 2015)  
g2 <- ggplot(data = df.combined, aes(x = as.factor(GarageCars))) +  
 geom\_histogram(stat = 'count')

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

g3 <- ggplot(data = df.combined, aes(x = GarageArea)) +  
 geom\_density()  
g4 <- ggplot(data = df.combined, aes(x = as.factor(GarageCond))) +  
 geom\_histogram(stat = 'count')

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

g5 <- ggplot(data = df.combined, aes(x = GarageType)) +  
 geom\_histogram(stat = 'count')

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

g6 <- ggplot(data = df.combined, aes(x = as.factor(GarageQual))) +  
 geom\_histogram(stat = 'count')

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

g7 <- ggplot(data = df.combined, aes(x = as.factor(GarageFinish))) +  
 geom\_histogram(stat = 'count')

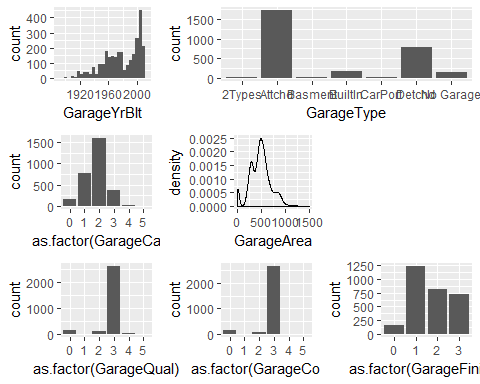
## Warning: Ignoring unknown parameters: binwidth, bins, pad

layout <- matrix(c(1, 5, 5, 2, 3, 8, 6, 4, 7), 3 , 3, byrow = TRUE)  
multiplot(g1, g2, g3, g4, g5, g6, g7, layout = layout)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat\_bin).

## Warning: Removed 1 rows containing missing values (geom\_bar).



##EDA of Basement Variables  
b1 <- ggplot(data = df.combined, aes(BsmtFinSF1)) +   
 geom\_histogram() +  
 labs(x = 'Type 1 Finished SqFt')  
b2 <- ggplot(data = df.combined, aes(x = BsmtFinSF2)) +  
 geom\_histogram() +  
 labs(x = 'Type 2 Finished SqFt')  
b3 <- ggplot(data = df.combined, aes(x = BsmtUnfSF)) +   
 geom\_histogram() +   
 labs(x = 'Unfinished SqFt')  
b4 <- ggplot(data = df.combined, aes(x = as.factor(BsmtFinType1))) +  
 geom\_histogram(stat = 'count') +  
 labs(x = 'Type 1 Finished Area Rating')

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

b5 <- ggplot(data = df.combined, aes(x = as.factor(BsmtFinType2))) +  
 geom\_histogram(stat = 'count') +  
 labs(x = 'Type 2 Finished Area Rating')

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

b6 <- ggplot(data = df.combined, aes(x = as.factor(BsmtQual))) +  
 geom\_histogram(stat = 'count') +  
 labs(x = 'Height of Bsmt')

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

b7 <- ggplot(data = df.combined, aes(x = as.factor(BsmtCond))) +  
 geom\_histogram(stat = 'count') +  
 labs(x = 'Bsmt Condition Rating')

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

b8 <- ggplot(data = df.combined, aes(x = as.factor(BsmtExposure))) +  
 geom\_histogram(stat = 'count') +  
 labs(x = 'Walkout/Garden Level Walls')

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

layout <- matrix(c(1, 2, 3, 4, 5, 9, 6, 7, 8), 3, 3, byrow = TRUE)  
multiplot(b1, b2, b3, b4, b5, b6, b7, b8, layout = layout)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

