Model Prep & Normalization

##Data Prep for Modeling  
  
##Remove Highly Correlated Variables  
rmVars <- c('YearRemodAdd', 'GarageYrBlt', 'GarageArea', 'GarageCond', 'TotalBsmtSF', 'TotRmsAbvGrd')  
  
df.combined <- df.combined[, !(names(df.combined) %in% rmVars)]  
  
##Remove Outliers  
df.combined <- df.combined[-c(524, 1299),]

##Prepping Predictor Variables  
  
numeric.VarNames <- numeric.VarNames[!(numeric.VarNames %in% c('MSSubClass', 'MoSold', 'YrSold', 'SalePrice', 'OverallQual', 'OverallCond'))]  
numeric.VarNames <- append(numeric.VarNames, c('Age', 'TotalPorchSqFt', 'TotalBath', 'TotalSqFt'))  
  
df.numeric <- df.combined[, names(df.combined) %in% numeric.VarNames]  
str(df.numeric)

## 'data.frame': 2913 obs. of 30 variables:  
## $ LotFrontage : int 65 80 68 60 84 85 75 80 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  
## $ YearBuilt : int 2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...  
## $ MasVnrArea : num 196 0 162 0 350 0 186 240 0 0 ...  
## $ BsmtFinSF1 : num 706 978 486 216 655 ...  
## $ BsmtFinSF2 : num 0 0 0 0 0 0 0 32 0 0 ...  
## $ BsmtUnfSF : num 150 284 434 540 490 64 317 216 952 140 ...  
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...  
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...  
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...  
## $ BsmtFullBath : num 1 0 1 1 1 1 1 1 0 1 ...  
## $ BsmtHalfBath : num 0 1 0 0 0 0 0 0 0 0 ...  
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...  
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...  
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...  
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ GarageCars : num 2 2 2 3 3 2 2 2 2 1 ...  
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...  
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...  
## $ EnclosedPorch : int 0 0 0 272 0 0 0 228 205 0 ...  
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...  
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...  
## $ TotalBath : num 3.5 2.5 3.5 2 3.5 2.5 3 3.5 2 2 ...  
## $ Age : num -2000 -1974 -1999 -1969 -1997 ...  
## $ TotalSqFt : num 2566 2524 2706 2473 3343 ...  
## $ TotalPorchSqFt: int 61 0 42 307 84 350 57 432 205 4 ...

df.categoric <- df.combined[, !(names(df.combined) %in% numeric.VarNames)]  
df.categoric <- df.categoric[, names(df.categoric) != 'SalePrice']  
str(df.categoric)

## 'data.frame': 2913 obs. of 49 variables:  
## $ MSSubClass : Factor w/ 16 levels "1 Story 1946+",..: 6 1 6 7 6 5 1 6 5 16 ...  
## $ MSZoning : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...  
## $ Street : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Alley : Factor w/ 3 levels "Grvl","None",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ LotShape : int 3 3 2 2 2 2 3 2 3 3 ...  
## $ LandContour : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ LotConfig : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...  
## $ LandSlope : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...  
## $ Condition1 : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...  
## $ Condition2 : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...  
## $ BldgType : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...  
## $ HouseStyle : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...  
## $ OverallQual : int 7 6 7 7 8 5 8 7 7 5 ...  
## $ OverallCond : int 5 8 5 5 5 5 5 6 5 6 ...  
## $ RoofStyle : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ RoofMatl : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Exterior1st : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...  
## $ Exterior2nd : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...  
## $ MasVnrType : int 1 0 1 0 1 0 2 2 0 0 ...  
## $ ExterQual : int 4 3 4 3 4 3 4 3 3 3 ...  
## $ ExterCond : int 3 3 3 3 3 3 3 3 3 3 ...  
## $ Foundation : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...  
## $ BsmtQual : int 4 4 4 3 4 4 5 4 3 3 ...  
## $ BsmtCond : int 3 3 3 4 3 3 3 3 3 3 ...  
## $ BsmtExposure : int 1 4 2 1 3 1 3 2 1 1 ...  
## $ BsmtFinType1 : int 6 5 6 5 6 6 6 5 1 6 ...  
## $ BsmtFinType2 : int 1 1 1 1 1 1 1 4 1 1 ...  
## $ Heating : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ HeatingQC : int 5 5 5 4 5 5 5 5 4 5 ...  
## $ CentralAir : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electrical : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...  
## $ KitchenQual : int 4 3 4 4 4 3 4 3 3 3 ...  
## $ Functional : int 7 7 7 7 7 7 7 7 6 7 ...  
## $ FireplaceQu : int 0 3 3 4 3 0 4 3 3 3 ...  
## $ GarageType : Factor w/ 7 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...  
## $ GarageFinish : int 2 2 2 1 2 1 2 2 1 2 ...  
## $ GarageQual : int 3 3 3 3 3 3 3 3 2 4 ...  
## $ PavedDrive : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ PoolQC : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Fence : Factor w/ 5 levels "GdPrv","GdWo",..: 5 5 5 5 5 3 5 5 5 5 ...  
## $ MiscFeature : Factor w/ 5 levels "Gar2","None",..: 2 2 2 2 2 4 2 4 2 2 ...  
## $ MoSold : Factor w/ 12 levels "1","2","3","4",..: 2 5 9 2 12 10 8 11 4 1 ...  
## $ YrSold : Factor w/ 5 levels "2006","2007",..: 3 2 3 1 3 4 2 4 3 3 ...  
## $ SaleType : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...  
## $ SaleCondition : Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...  
## $ Remod : num 0 0 1 1 0 1 1 0 1 1 ...  
## $ New : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NeighborhoodWealth: num 1 1 1 1 2 1 1 1 1 1 ...

cat(length(df.numeric), 'numeric variables &', length(df.categoric), 'categoric variables')

## 30 numeric variables & 49 categoric variables

##Skewness of Numeric Variables  
for(i in 1:ncol(df.numeric)){  
 if(abs(skew(df.numeric[,i])) > 0.8){  
 df.numeric[,i] <- log(df.numeric[,i] + 1)  
 }  
}  
  
##Normalization of Numeric Variables  
Predictor.Vars <- preProcess(df.numeric, method = c("center", "scale"))  
print(Predictor.Vars)

## Created from 2913 samples and 30 variables  
##   
## Pre-processing:  
## - centered (30)  
## - ignored (0)  
## - scaled (30)

df.normal <- predict(Predictor.Vars, df.numeric)  
dim(df.normal)

## [1] 2913 30

##Encoding of Categoric Varibles  
df.dummy <- as.data.frame(model.matrix(~.-1, df.categoric))  
dim(df.dummy)

## [1] 2913 201

##Removing Levels with Few/None Values  
  
##Absent Values in Test Set  
Values.Absent.Test <- which(colSums(df.dummy[1459:2917, ]) == 0)  
colnames(df.dummy[Values.Absent.Test])

## character(0)

##Removing Predictor Values  
df.dummy <- df.dummy[, -Values.Absent.Test]  
  
##Absent Values in Train Set  
Values.Absent.Train <- which(colSums(df.dummy[1:1458, ]) == 0)  
colnames(df.dummy[Values.Absent.Train])

## character(0)

##Removing Predictor Values  
df.dummy <- df.dummy[, -Values.Absent.Train]  
  
##Few Values (<10) in Train Set  
Values.Few.Train <- which(colSums(df.dummy[1:1458, ]) <10)  
colnames(df.dummy[Values.Few.Train])

## character(0)

##Removing Predictor Values  
df.dummy <- df.dummy[, -Values.Few.Train]  
dim(df.dummy)

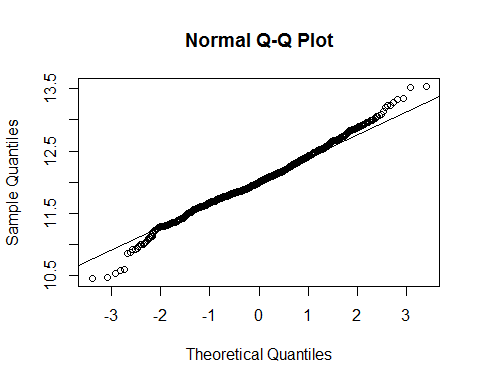
## [1] 2913 0

##Combine Predictors into Data Frame   
df.combined.predictors <- cbind(df.normal, df.dummy)

##Verifying Skewness of Response Variable (SalePrice)  
skew(df.combined$SalePrice)

## [1] 0.1248346

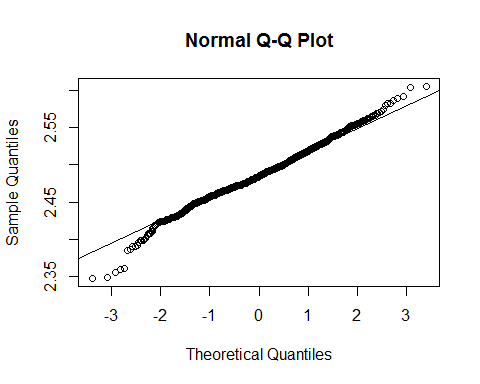
qqnorm(df.combined$SalePrice)  
qqline(df.combined$SalePrice)



##Normally Distribute SalePrice  
df.combined$SalePrice <- log(df.combined$SalePrice)  
  
##Verifying Skewness of Normalized Response Variable (SalePrice)  
skew(df.combined$SalePrice)

## [1] -0.01683804

qqnorm(df.combined$SalePrice)  
qqline(df.combined$SalePrice)



##Saving Train & Test Sets  
train.set <- df.combined.predictors[!is.na(df.combined$SalePrice), ]  
str(train.set)

## 'data.frame': 1454 obs. of 30 variables:  
## $ LotFrontage : num -0.0491 0.5756 0.0865 -0.2894 0.7227 ...  
## $ LotArea : num -0.101 0.15 0.462 0.139 0.928 ...  
## $ YearBuilt : num 1.048 0.157 0.982 -1.857 0.949 ...  
## $ MasVnrArea : num 1.227 -0.792 1.154 -0.792 1.447 ...  
## $ BsmtFinSF1 : num 0.785 0.894 0.66 0.389 0.76 ...  
## $ BsmtFinSF2 : num -0.363 -0.363 -0.363 -0.363 -0.363 ...  
## $ BsmtUnfSF : num -0.3241 0.0154 0.2414 0.358 0.3062 ...  
## $ X1stFlrSF : num -0.778 0.43 -0.553 -0.418 0.127 ...  
## $ X2ndFlrSF : num 1.199 -0.862 1.203 1.162 1.263 ...  
## $ LowQualFinSF : num -0.117 -0.117 -0.117 -0.117 -0.117 ...  
## $ GrLivArea : num 0.572 -0.368 0.706 0.584 1.349 ...  
## $ BsmtFullBath : num 1.09 -0.818 1.09 1.09 1.09 ...  
## $ BsmtHalfBath : num -0.252 3.884 -0.252 -0.252 -0.252 ...  
## $ FullBath : num 0.784 0.784 0.784 -1.026 0.784 ...  
## $ HalfBath : num 1.235 -0.755 1.235 -0.755 1.235 ...  
## $ BedroomAbvGr : num 0.171 0.171 0.171 0.171 1.385 ...  
## $ KitchenAbvGr : num -0.201 -0.201 -0.201 -0.201 -0.201 ...  
## $ Fireplaces : num -0.923 0.627 0.627 0.627 0.627 ...  
## $ GarageCars : num 0.309 0.309 0.309 1.622 1.622 ...  
## $ WoodDeckSF : num -0.944 1.256 -0.944 -0.944 1.087 ...  
## $ OpenPorchSF : num 0.835 -1.082 0.665 0.583 0.982 ...  
## $ EnclosedPorch : num -0.428 -0.428 -0.428 2.735 -0.428 ...  
## $ X3SsnPorch : num -0.113 -0.113 -0.113 -0.113 -0.113 ...  
## $ ScreenPorch : num -0.309 -0.309 -0.309 -0.309 -0.309 ...  
## $ PoolArea : num -0.0641 -0.0641 -0.0641 -0.0641 -0.0641 ...  
## $ MiscVal : num -0.189 -0.189 -0.189 -0.189 -0.189 ...  
## $ TotalBath : num 1.592 0.352 1.592 -0.268 1.592 ...  
## $ Age : num -0.888 0.356 -0.841 0.595 -0.745 ...  
## $ TotalSqFt : num 0.175 0.1219 0.3458 0.0563 1.0258 ...  
## $ TotalPorchSqFt: num 0.434 -1.453 0.267 1.167 0.579 ...

View(train.set)  
test.set <- df.combined.predictors[is.na(df.combined$SalePrice), ]  
str(test.set)

## 'data.frame': 1459 obs. of 30 variables:  
## $ LotFrontage : num 0.576 0.613 0.341 0.499 -1.286 ...  
## $ LotArea : num 0.526 0.929 0.868 0.226 -1.132 ...  
## $ YearBuilt : num -0.338 -0.437 0.85 0.883 0.685 ...  
## $ MasVnrArea : num -0.792 1 -0.792 0.371 -0.792 ...  
## $ BsmtFinSF1 : num 0.647 0.875 0.823 0.731 0.454 ...  
## $ BsmtFinSF2 : num 2.284 -0.363 -0.363 -0.363 -0.363 ...  
## $ BsmtUnfSF : num -0.0115 0.2059 -0.3722 0.0856 0.6959 ...  
## $ X1stFlrSF : num -0.636 0.591 -0.526 -0.533 0.474 ...  
## $ X2ndFlrSF : num -0.862 -0.862 1.139 1.129 -0.862 ...  
## $ LowQualFinSF : num -0.117 -0.117 -0.117 -0.117 -0.117 ...  
## $ GrLivArea : num -1.428 -0.208 0.422 0.374 -0.324 ...  
## $ BsmtFullBath : num -0.818 -0.818 -0.818 -0.818 -0.818 ...  
## $ BsmtHalfBath : num -0.252 -0.252 -0.252 -0.252 -0.252 ...  
## $ FullBath : num -1.026 -1.026 0.784 0.784 0.784 ...  
## $ HalfBath : num -0.755 1.235 1.235 1.235 -0.755 ...  
## $ BedroomAbvGr : num -1.044 0.171 0.171 0.171 -1.044 ...  
## $ KitchenAbvGr : num -0.201 -0.201 -0.201 -0.201 -0.201 ...  
## $ Fireplaces : num -0.923 -0.923 0.627 0.627 -0.923 ...  
## $ GarageCars : num -1.003 -1.003 0.309 0.309 0.309 ...  
## $ WoodDeckSF : num 0.966 1.362 1.125 1.328 -0.944 ...  
## $ OpenPorchSF : num -1.082 0.595 0.57 0.595 0.971 ...  
## $ EnclosedPorch : num -0.428 -0.428 -0.428 -0.428 -0.428 ...  
## $ X3SsnPorch : num -0.113 -0.113 -0.113 -0.113 -0.113 ...  
## $ ScreenPorch : num 2.973 -0.309 -0.309 -0.309 3.097 ...  
## $ PoolArea : num -0.0641 -0.0641 -0.0641 -0.0641 -0.0641 ...  
## $ MiscVal : num -0.189 7.409 -0.189 -0.189 -0.189 ...  
## $ TotalBath : num -1.507 -0.887 0.352 0.352 -0.268 ...  
## $ Age : num 1.217 1.361 -0.553 -0.553 -0.266 ...  
## $ TotalSqFt : num -1.005 0.288 0.164 0.13 0.167 ...  
## $ TotalPorchSqFt: num 0.74 0.198 0.173 0.198 1.028 ...

View(test.set)