Code Pt. 1

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(corrplot)

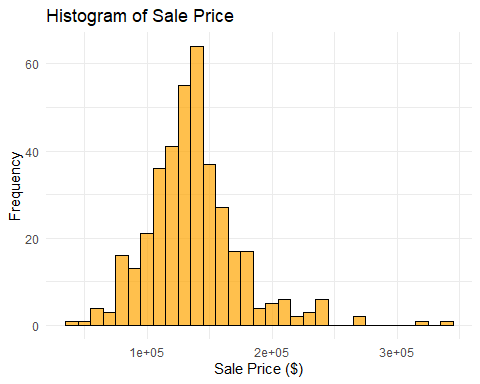
## Warning: package 'corrplot' was built under R version 4.4.2

## corrplot 0.95 loaded

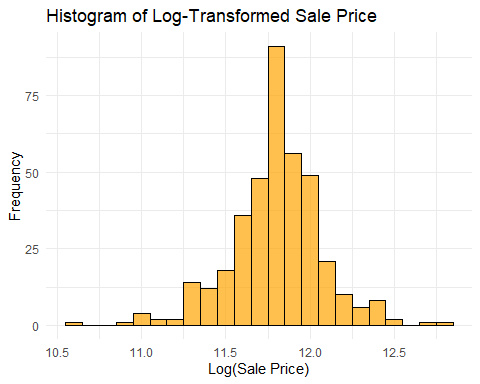
library(lattice)  
  
#load dataset  
trainData <- read.csv("/Users/katie/Documents/MSDS/Statistical Science/Unit 13/Project/train.csv")  
  
#clean up data  
train <- trainData %>%  
 filter(!is.na(SalePrice) & !is.na(GrLivArea)  
 & Neighborhood %in% c("NAmes", "Edwards", "BrkSide"))  
head(train)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 1 1299 60 RL 313 63887 Pave <NA> IR3  
## 2 524 60 RL 130 40094 Pave <NA> IR1  
## 3 198 75 RL 174 25419 Pave <NA> Reg  
## 4 325 80 RL 96 11275 Pave <NA> Reg  
## 5 643 80 RL 75 13860 Pave <NA> Reg  
## 6 1351 90 RL 91 11643 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2  
## 1 Bnk AllPub Corner Gtl Edwards Feedr Norm  
## 2 Bnk AllPub Inside Gtl Edwards PosN PosN  
## 3 Lvl AllPub Corner Gtl NAmes Artery Norm  
## 4 Lvl AllPub Corner Gtl NAmes PosN Norm  
## 5 Lvl AllPub Inside Gtl NAmes Norm Norm  
## 6 Lvl AllPub Inside Gtl NAmes Artery Norm  
## BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 1Fam 2Story 10 5 2008 2008 Hip  
## 2 1Fam 2Story 10 5 2007 2008 Hip  
## 3 1Fam 2Story 8 4 1918 1990 Gable  
## 4 1Fam SLvl 7 7 1967 2007 Mansard  
## 5 1Fam SLvl 8 7 1972 1995 Gable  
## 6 Duplex 2Story 5 5 1969 1969 Gable  
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond  
## 1 ClyTile Stucco Stucco Stone 796 Ex TA  
## 2 CompShg CemntBd CmentBd Stone 762 Ex TA  
## 3 CompShg Stucco Stucco None 0 Gd Gd  
## 4 WdShake Wd Sdng Wd Sdng BrkFace 300 Gd Gd  
## 5 CompShg Plywood Wd Sdng None 0 Gd TA  
## 6 CompShg MetalSd MetalSd BrkFace 368 TA TA  
## Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1  
## 1 PConc Ex TA Gd GLQ 5644  
## 2 PConc Ex TA Gd GLQ 2260  
## 3 PConc TA TA No GLQ 1036  
## 4 CBlock Gd TA No Unf 0  
## 5 CBlock Gd TA Gd GLQ 1410  
## 6 CBlock TA TA No LwQ 500  
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir  
## 1 Unf 0 466 6110 GasA Ex Y  
## 2 Unf 0 878 3138 GasA Ex Y  
## 3 LwQ 184 140 1360 GasA Gd Y  
## 4 Unf 0 710 710 GasA Ex Y  
## 5 Unf 0 542 1952 GasA Gd Y  
## 6 Unf 0 748 1248 GasA TA Y  
## Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath  
## 1 SBrkr 4692 950 0 5642 2  
## 2 SBrkr 3138 1538 0 4676 1  
## 3 SBrkr 1360 1360 392 3112 1  
## 4 SBrkr 1898 1080 0 2978 0  
## 5 SBrkr 2000 704 0 2704 1  
## 6 SBrkr 1338 1296 0 2634 1  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual  
## 1 0 2 1 3 1 Ex  
## 2 0 3 1 3 1 Ex  
## 3 1 2 0 4 1 Gd  
## 4 0 2 1 5 1 Gd  
## 5 0 2 1 4 1 Ex  
## 6 1 2 2 6 2 TA  
## TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt  
## 1 12 Typ 3 Gd Attchd 2008  
## 2 11 Typ 1 Gd BuiltIn 2007  
## 3 8 Typ 1 Ex Detchd 1918  
## 4 11 Typ 1 Gd BuiltIn 1961  
## 5 9 Typ 3 TA Attchd 1972  
## 6 12 Typ 0 <NA> Detchd 1969  
## GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive  
## 1 Fin 2 1418 TA TA Y  
## 2 Fin 3 884 TA TA Y  
## 3 Unf 2 795 TA TA Y  
## 4 Fin 2 564 TA TA Y  
## 5 Fin 2 538 TA TA Y  
## 6 Unf 4 968 TA TA Y  
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 1 214 292 0 0 0 480 Gd  
## 2 208 406 0 0 0 0 <NA>  
## 3 0 16 552 0 0 512 Ex  
## 4 240 0 0 0 0 0 <NA>  
## 5 269 111 0 0 0 0 <NA>  
## 6 0 0 0 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 <NA> <NA> 0 1 2008 New Partial 160000  
## 2 <NA> <NA> 0 10 2007 New Partial 184750  
## 3 GdPrv <NA> 0 3 2006 WD Abnorml 235000  
## 4 <NA> <NA> 0 6 2010 WD Normal 242000  
## 5 MnPrv <NA> 0 7 2009 WD Normal 345000  
## 6 <NA> <NA> 0 8 2009 WD Normal 200000

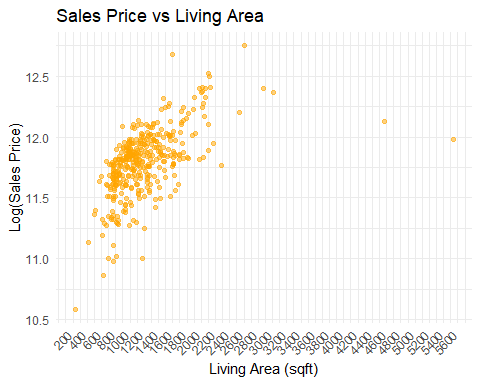
# Observe normality  
ggplot(train, aes(x = SalePrice)) +  
 geom\_histogram(binwidth = 10000, fill = "orange", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Sale Price",   
 x = "Sale Price ($)",   
 y = "Frequency") +  
 theme\_minimal()



# check for normality bc OG looks skewed  
train$LogSalePrice <- log(train$SalePrice)  
  
ggplot(train, aes(x = LogSalePrice)) +  
 geom\_histogram(binwidth = 0.1, fill = "orange", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Log-Transformed Sale Price",   
 x = "Log(Sale Price)",   
 y = "Frequency") +  
 theme\_minimal()



# plot square footage of the living area of the house vs Sales Price  
ggplot(train, aes(x = GrLivArea, y = LogSalePrice)) +  
 geom\_point(alpha = 0.5, col = "orange") +  
 labs(title = "Sales Price vs Living Area", x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 scale\_x\_continuous(breaks = seq(0, max(train$GrLivArea), by = 200)) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#remove outliers  
trainNoOutliers <- train %>%  
 filter(!Id %in% c(1299, 524)) #remove the 2 data points on RHS of plot  
head(trainNoOutliers)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 1 198 75 RL 174 25419 Pave <NA> Reg  
## 2 325 80 RL 96 11275 Pave <NA> Reg  
## 3 643 80 RL 75 13860 Pave <NA> Reg  
## 4 1351 90 RL 91 11643 Pave <NA> Reg  
## 5 667 60 RL NA 18450 Pave <NA> IR1  
## 6 911 90 RL 80 11600 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2  
## 1 Lvl AllPub Corner Gtl NAmes Artery Norm  
## 2 Lvl AllPub Corner Gtl NAmes PosN Norm  
## 3 Lvl AllPub Inside Gtl NAmes Norm Norm  
## 4 Lvl AllPub Inside Gtl NAmes Artery Norm  
## 5 Lvl AllPub Inside Gtl NAmes Norm Norm  
## 6 Lvl AllPub Corner Gtl NAmes Feedr Norm  
## BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 1Fam 2Story 8 4 1918 1990 Gable  
## 2 1Fam SLvl 7 7 1967 2007 Mansard  
## 3 1Fam SLvl 8 7 1972 1995 Gable  
## 4 Duplex 2Story 5 5 1969 1969 Gable  
## 5 1Fam 2Story 6 5 1965 1979 Flat  
## 6 Duplex 2Story 5 5 1960 1960 Gable  
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond  
## 1 CompShg Stucco Stucco None 0 Gd Gd  
## 2 WdShake Wd Sdng Wd Sdng BrkFace 300 Gd Gd  
## 3 CompShg Plywood Wd Sdng None 0 Gd TA  
## 4 CompShg MetalSd MetalSd BrkFace 368 TA TA  
## 5 Tar&Grv Plywood Plywood BrkCmn 113 TA Gd  
## 6 CompShg MetalSd MetalSd BrkFace 361 TA TA  
## Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1  
## 1 PConc TA TA No GLQ 1036  
## 2 CBlock Gd TA No Unf 0  
## 3 CBlock Gd TA Gd GLQ 1410  
## 4 CBlock TA TA No LwQ 500  
## 5 CBlock Gd TA No LwQ 187  
## 6 CBlock TA TA No Rec 443  
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir  
## 1 LwQ 184 140 1360 GasA Gd Y  
## 2 Unf 0 710 710 GasA Ex Y  
## 3 Unf 0 542 1952 GasA Gd Y  
## 4 Unf 0 748 1248 GasA TA Y  
## 5 Rec 723 111 1021 GasA TA Y  
## 6 Unf 0 662 1105 GasA TA Y  
## Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath  
## 1 SBrkr 1360 1360 392 3112 1  
## 2 SBrkr 1898 1080 0 2978 0  
## 3 SBrkr 2000 704 0 2704 1  
## 4 SBrkr 1338 1296 0 2634 1  
## 5 SBrkr 1465 915 0 2380 0  
## 6 FuseA 1105 1169 0 2274 0  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual  
## 1 1 2 0 4 1 Gd  
## 2 0 2 1 5 1 Gd  
## 3 0 2 1 4 1 Ex  
## 4 1 2 2 6 2 TA  
## 5 0 2 1 3 1 TA  
## 6 0 2 0 5 2 TA  
## TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt  
## 1 8 Typ 1 Ex Detchd 1918  
## 2 11 Typ 1 Gd BuiltIn 1961  
## 3 9 Typ 3 TA Attchd 1972  
## 4 12 Typ 0 <NA> Detchd 1969  
## 5 7 Sev 1 Po CarPort 1965  
## 6 12 Typ 0 <NA> Detchd 1960  
## GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive  
## 1 Unf 2 795 TA TA Y  
## 2 Fin 2 564 TA TA Y  
## 3 Fin 2 538 TA TA Y  
## 4 Unf 4 968 TA TA Y  
## 5 Unf 2 596 TA TA Y  
## 6 Unf 2 480 TA TA Y  
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 1 0 16 552 0 0 512 Ex  
## 2 240 0 0 0 0 0 <NA>  
## 3 269 111 0 0 0 0 <NA>  
## 4 0 0 0 0 0 0 <NA>  
## 5 0 265 0 0 0 0 <NA>  
## 6 0 0 0 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 GdPrv <NA> 0 3 2006 WD Abnorml 235000  
## 2 <NA> <NA> 0 6 2010 WD Normal 242000  
## 3 MnPrv <NA> 0 7 2009 WD Normal 345000  
## 4 <NA> <NA> 0 8 2009 WD Normal 200000  
## 5 <NA> <NA> 0 8 2007 WD Abnorml 129000  
## 6 <NA> <NA> 0 1 2010 WD Normal 154300  
## LogSalePrice  
## 1 12.36734  
## 2 12.39669  
## 3 12.75130  
## 4 12.20607  
## 5 11.76757  
## 6 11.94665

#Build the regression model considering GrLivArea and Neighborhood  
model <- lm(LogSalePrice ~ GrLivArea, data = trainNoOutliers)  
summary(model)

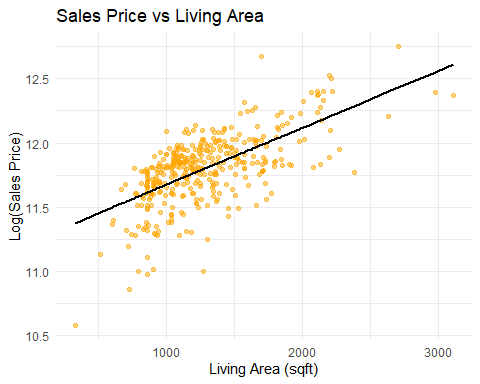
##   
## Call:  
## lm(formula = LogSalePrice ~ GrLivArea, data = trainNoOutliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.79801 -0.12784 0.04049 0.14489 0.69375   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.123e+01 3.642e-02 308.32 <2e-16 \*\*\*  
## GrLivArea 4.438e-04 2.715e-05 16.34 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2097 on 379 degrees of freedom  
## Multiple R-squared: 0.4134, Adjusted R-squared: 0.4119   
## F-statistic: 267.2 on 1 and 379 DF, p-value: < 2.2e-16

confint(model)

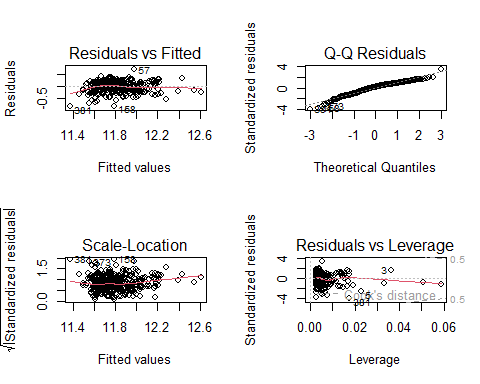
## 2.5 % 97.5 %  
## (Intercept) 1.115716e+01 1.130037e+01  
## GrLivArea 3.904036e-04 4.971779e-04

#plot the predicted SalePrice vs. the actual SalePrice  
trainNoOutliers$LogpredictedSalePrice <- predict(model, newdata = trainNoOutliers)  
trainNoOutliers$predictedSalePrice <- exp(trainNoOutliers$LogpredictedSalePrice)  
  
ggplot(trainNoOutliers, aes(x = GrLivArea, y = LogSalePrice)) +  
 geom\_point(alpha = 0.5, color = "orange") +  
 geom\_line(aes(y = LogpredictedSalePrice), color = "black", size = 1) +  
 labs(title = "Sales Price vs Living Area",   
 x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid  
plot(model)



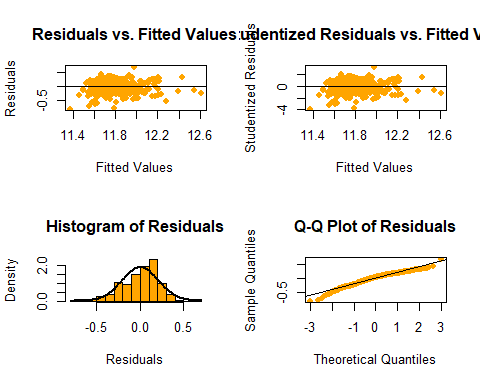
# Find CV PRESS  
cvpress <- sum((trainNoOutliers$SalePrice - trainNoOutliers$predictedSalePrice)^2)  
print(paste("CV PRESS:", cvpress))

## [1] "CV PRESS: 308771210722.523"

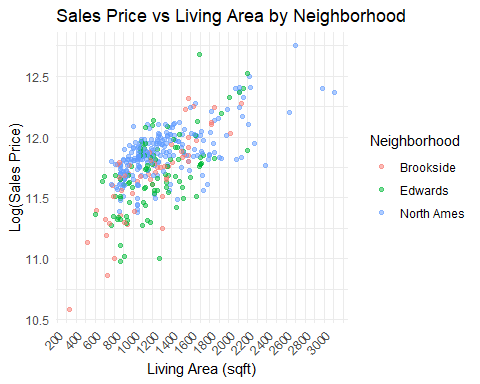
# Calculate AIC   
aic <- AIC(model)  
print(paste("AIC:", aic))

## [1] "AIC: -104.938344932316"

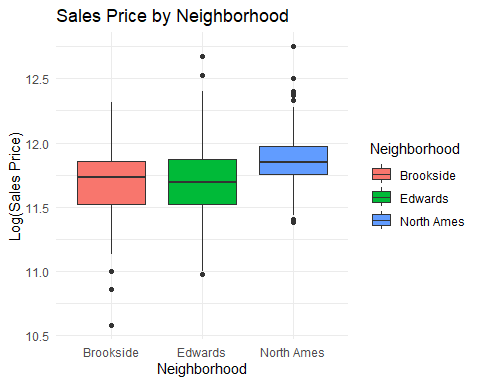
#Check assumptions  
  
par(mfrow = c(2, 2))  
  
# Residuals vs Fitted values  
plot(fitted(model), residuals(model),  
 xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs. Fitted Values", pch = 19, col="orange")  
abline(h = 0, col = "black")  
  
 # Studentized Residuals  
plot(fitted(model), rstudent(model),  
 xlab = "Fitted Values", ylab = "Studentized Residuals", main = "Studentized Residuals vs. Fitted Values",   
 pch = 19, col="orange")  
abline(h = 0, col = "black")  
   
# Histogram of residuals  
hist(residuals(model),  
 xlab = "Residuals", main = "Histogram of Residuals", col = "orange", border = "black", breaks = 20, probability=TRUE)  
curve(dnorm(x, mean = mean(residuals(model)), sd = sd(residuals(model))), col = "black", lwd = 2, add = TRUE)  
  
# Q-Q plot of residuals  
qqnorm(residuals(model), main = "Q-Q Plot of Residuals", col="orange", pch=19)   
qqline(residuals(model), col = "black")



#View plot separated by neighborhood w colors  
ggplot(trainNoOutliers, aes(x = GrLivArea, y = LogSalePrice, color = Neighborhood)) +  
 geom\_point(alpha = 0.5) +  
 labs(title = "Sales Price vs Living Area by Neighborhood",   
 x = "Living Area (sqft)",   
 y = "Log(Sales Price)") +  
 scale\_color\_discrete(labels = c("NAmes" = "North Ames", "Edwards" = "Edwards", "BrkSide" = "Brookside")) +  
 scale\_x\_continuous(breaks = seq(0, max(trainNoOutliers$GrLivArea), by = 200)) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Check for normality/variance with boxplot  
ggplot(trainNoOutliers, aes(x = Neighborhood, y = LogSalePrice, fill=Neighborhood)) +  
 geom\_boxplot() +  
 labs(title = "Sales Price by Neighborhood", x = "Neighborhood", y = "Log(Sales Price)") +   
 scale\_x\_discrete(labels = c("NAmes" = "North Ames", "Edwards" = "Edwards", "BrkSide" = "Brookside")) +  
 scale\_fill\_discrete(labels = c("NAmes" = "North Ames", "Edwards" = "Edwards", "BrkSide" = "Brookside")) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 theme\_minimal()



# See correlations, box plots, density plot, histogram, etc. all at once  
library(GGally)

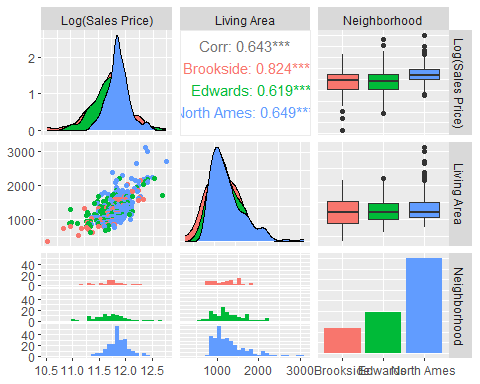
## Warning: package 'GGally' was built under R version 4.4.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(dplyr)  
matrixPlot<-trainNoOutliers %>%  
 select(LogSalePrice, GrLivArea, Neighborhood) %>%  
 rename(  
 `Log(Sales Price)` = LogSalePrice, # Change "SalePrice" to "Sales Price"  
 `Living Area` = GrLivArea, # Change "GrLivArea" to "Living Area"  
 ) %>%  
 mutate(  
 Neighborhood = recode(Neighborhood,   
 "NAmes" = "North Ames",   
 "BrkSide" = "Brookside")  
 )  
ggpairs(matrixPlot, mapping = aes(color = Neighborhood))

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

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



NAmes <- trainNoOutliers %>%  
 filter(Neighborhood %in% c("NAmes"))  
head(NAmes)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 1 198 75 RL 174 25419 Pave <NA> Reg  
## 2 325 80 RL 96 11275 Pave <NA> Reg  
## 3 643 80 RL 75 13860 Pave <NA> Reg  
## 4 1351 90 RL 91 11643 Pave <NA> Reg  
## 5 667 60 RL NA 18450 Pave <NA> IR1  
## 6 911 90 RL 80 11600 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2  
## 1 Lvl AllPub Corner Gtl NAmes Artery Norm  
## 2 Lvl AllPub Corner Gtl NAmes PosN Norm  
## 3 Lvl AllPub Inside Gtl NAmes Norm Norm  
## 4 Lvl AllPub Inside Gtl NAmes Artery Norm  
## 5 Lvl AllPub Inside Gtl NAmes Norm Norm  
## 6 Lvl AllPub Corner Gtl NAmes Feedr Norm  
## BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 1Fam 2Story 8 4 1918 1990 Gable  
## 2 1Fam SLvl 7 7 1967 2007 Mansard  
## 3 1Fam SLvl 8 7 1972 1995 Gable  
## 4 Duplex 2Story 5 5 1969 1969 Gable  
## 5 1Fam 2Story 6 5 1965 1979 Flat  
## 6 Duplex 2Story 5 5 1960 1960 Gable  
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond  
## 1 CompShg Stucco Stucco None 0 Gd Gd  
## 2 WdShake Wd Sdng Wd Sdng BrkFace 300 Gd Gd  
## 3 CompShg Plywood Wd Sdng None 0 Gd TA  
## 4 CompShg MetalSd MetalSd BrkFace 368 TA TA  
## 5 Tar&Grv Plywood Plywood BrkCmn 113 TA Gd  
## 6 CompShg MetalSd MetalSd BrkFace 361 TA TA  
## Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1  
## 1 PConc TA TA No GLQ 1036  
## 2 CBlock Gd TA No Unf 0  
## 3 CBlock Gd TA Gd GLQ 1410  
## 4 CBlock TA TA No LwQ 500  
## 5 CBlock Gd TA No LwQ 187  
## 6 CBlock TA TA No Rec 443  
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir  
## 1 LwQ 184 140 1360 GasA Gd Y  
## 2 Unf 0 710 710 GasA Ex Y  
## 3 Unf 0 542 1952 GasA Gd Y  
## 4 Unf 0 748 1248 GasA TA Y  
## 5 Rec 723 111 1021 GasA TA Y  
## 6 Unf 0 662 1105 GasA TA Y  
## Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath  
## 1 SBrkr 1360 1360 392 3112 1  
## 2 SBrkr 1898 1080 0 2978 0  
## 3 SBrkr 2000 704 0 2704 1  
## 4 SBrkr 1338 1296 0 2634 1  
## 5 SBrkr 1465 915 0 2380 0  
## 6 FuseA 1105 1169 0 2274 0  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual  
## 1 1 2 0 4 1 Gd  
## 2 0 2 1 5 1 Gd  
## 3 0 2 1 4 1 Ex  
## 4 1 2 2 6 2 TA  
## 5 0 2 1 3 1 TA  
## 6 0 2 0 5 2 TA  
## TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt  
## 1 8 Typ 1 Ex Detchd 1918  
## 2 11 Typ 1 Gd BuiltIn 1961  
## 3 9 Typ 3 TA Attchd 1972  
## 4 12 Typ 0 <NA> Detchd 1969  
## 5 7 Sev 1 Po CarPort 1965  
## 6 12 Typ 0 <NA> Detchd 1960  
## GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive  
## 1 Unf 2 795 TA TA Y  
## 2 Fin 2 564 TA TA Y  
## 3 Fin 2 538 TA TA Y  
## 4 Unf 4 968 TA TA Y  
## 5 Unf 2 596 TA TA Y  
## 6 Unf 2 480 TA TA Y  
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 1 0 16 552 0 0 512 Ex  
## 2 240 0 0 0 0 0 <NA>  
## 3 269 111 0 0 0 0 <NA>  
## 4 0 0 0 0 0 0 <NA>  
## 5 0 265 0 0 0 0 <NA>  
## 6 0 0 0 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 GdPrv <NA> 0 3 2006 WD Abnorml 235000  
## 2 <NA> <NA> 0 6 2010 WD Normal 242000  
## 3 MnPrv <NA> 0 7 2009 WD Normal 345000  
## 4 <NA> <NA> 0 8 2009 WD Normal 200000  
## 5 <NA> <NA> 0 8 2007 WD Abnorml 129000  
## 6 <NA> <NA> 0 1 2010 WD Normal 154300  
## LogSalePrice LogpredictedSalePrice predictedSalePrice  
## 1 12.36734 12.60984 299491.9  
## 2 12.39669 12.55037 282201.0  
## 3 12.75130 12.42878 249890.0  
## 4 12.20607 12.39771 242246.4  
## 5 11.76757 12.28499 216422.5  
## 6 11.94665 12.23795 206477.4

#Build the NAmes regression model considering GrLivArea and Neighborhood  
model\_NAmes <- lm(LogSalePrice ~ GrLivArea, data = NAmes)  
summary(model\_NAmes)

##   
## Call:  
## lm(formula = LogSalePrice ~ GrLivArea, data = NAmes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.44719 -0.08610 0.03249 0.09751 0.43153   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.144e+01 3.495e-02 327.42 <2e-16 \*\*\*  
## GrLivArea 3.241e-04 2.544e-05 12.74 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1575 on 223 degrees of freedom  
## Multiple R-squared: 0.4212, Adjusted R-squared: 0.4186   
## F-statistic: 162.3 on 1 and 223 DF, p-value: < 2.2e-16

confint(model\_NAmes)

## 2.5 % 97.5 %  
## (Intercept) 1.137447e+01 1.151222e+01  
## GrLivArea 2.739874e-04 3.742617e-04

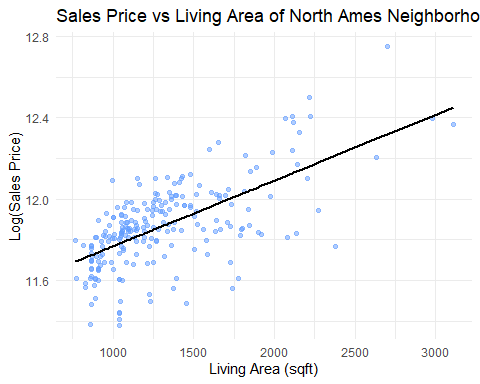
#plot the predicted SalePrice vs. the actual SalePrice  
NAmes$LogpredictedPrice <- predict(model\_NAmes, newdata = NAmes)  
NAmes$predictedPrice <- exp(NAmes$LogpredictedPrice)  
  
# Find CV PRESS  
cvpress\_N <- sum((NAmes$SalePrice - NAmes$predictedPrice)^2)  
print(paste("CV PRESS:", cvpress\_N))

## [1] "CV PRESS: 132149776632.633"

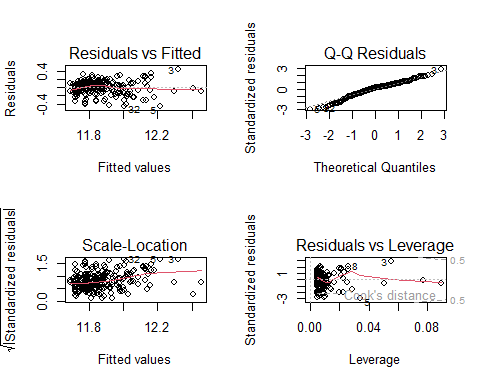
# Calculate AIC   
aic\_N <- AIC(model\_NAmes)  
print(paste("AIC:", aic\_N))

## [1] "AIC: -189.37473848245"

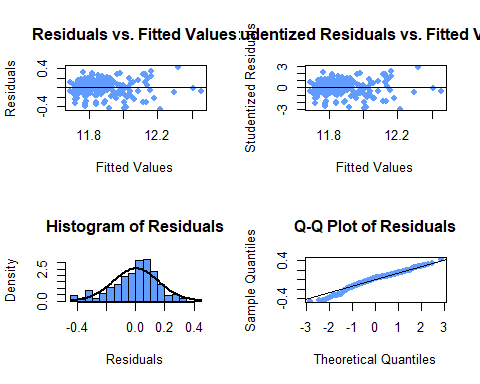
# Plot NAmes GrLiveArea vs SalePrice  
ggplot(NAmes, aes(x = GrLivArea, y = LogSalePrice)) +  
 geom\_point(alpha = 0.5, color = "#619CFF") +  
 geom\_line(aes(y = LogpredictedPrice), color = "black", size = 1) +  
 labs(title = "Sales Price vs Living Area of North Ames Neighborhood",   
 x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 theme\_minimal()



#Check Assumptions  
par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid  
plot(model\_NAmes)



# Check Assumptions  
par(mfrow = c(2, 2))  
  
# Residuals vs Fitted values  
plot(fitted(model\_NAmes), residuals(model\_NAmes),  
 xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs. Fitted Values", pch = 19, col="#619CFF")  
abline(h = 0, col = "black")  
  
 # Studentized Residuals  
plot(fitted(model\_NAmes), rstudent(model\_NAmes),  
 xlab = "Fitted Values", ylab = "Studentized Residuals", main = "Studentized Residuals vs. Fitted Values",   
 pch = 19, col="#619CFF")  
abline(h = 0, col = "black")  
   
# Histogram of residuals  
hist(residuals(model\_NAmes),  
 xlab = "Residuals", main = "Histogram of Residuals", col = "#619CFF", border = "black", breaks = 20, probability=TRUE)  
curve(dnorm(x, mean = mean(residuals(model\_NAmes)), sd = sd(residuals(model\_NAmes))), col = "black", lwd = 2, add = TRUE)  
  
# Q-Q plot of residuals  
qqnorm(residuals(model\_NAmes), main = "Q-Q Plot of Residuals", col="#619CFF", pch=19)   
qqline(residuals(model\_NAmes), col = "black")



Edwards <- trainNoOutliers %>%  
 filter(Neighborhood %in% c("Edwards"))  
head(Edwards)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 1 1424 80 RL NA 19690 Pave <NA> IR1  
## 2 922 90 RL 67 8777 Pave <NA> Reg  
## 3 176 20 RL 84 12615 Pave <NA> Reg  
## 4 1169 70 RL 120 13728 Pave <NA> Reg  
## 5 608 20 RL 78 7800 Pave <NA> Reg  
## 6 363 85 RL 64 7301 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2  
## 1 Lvl AllPub CulDSac Gtl Edwards Norm Norm  
## 2 Lvl AllPub Inside Gtl Edwards Feedr Norm  
## 3 Lvl AllPub Corner Gtl Edwards Norm Norm  
## 4 Lvl AllPub Corner Gtl Edwards Norm Norm  
## 5 Bnk AllPub Inside Mod Edwards Norm Norm  
## 6 Lvl AllPub Corner Gtl Edwards Norm Norm  
## BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 1Fam SLvl 6 7 1966 1966 Flat  
## 2 Duplex 1.5Fin 5 7 1900 2003 Gable  
## 3 1Fam 1Story 6 7 1950 2001 Gable  
## 4 1Fam 2Story 6 7 1935 1986 Hip  
## 5 1Fam 2Story 5 8 1948 2002 Gable  
## 6 1Fam SFoyer 7 5 2003 2003 Gable  
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond  
## 1 Tar&Grv Plywood Plywood None 0 Gd Gd  
## 2 CompShg MetalSd MetalSd None 0 TA TA  
## 3 CompShg WdShing Wd Shng None 0 TA TA  
## 4 CompShg Stucco Stucco None 0 TA TA  
## 5 CompShg MetalSd MetalSd None 0 TA Gd  
## 6 CompShg HdBoard HdBoard BrkFace 500 Gd TA  
## Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1  
## 1 CBlock Gd TA Av Unf 0  
## 2 CBlock TA TA No ALQ 1084  
## 3 CBlock TA Gd Av ALQ 477  
## 4 CBlock TA TA No Rec 626  
## 5 CBlock TA Gd No GLQ 603  
## 6 Slab <NA> <NA> <NA> <NA> 0  
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir  
## 1 Unf 0 697 697 GasA TA Y  
## 2 Unf 0 188 1272 GasA Gd Y  
## 3 Unf 0 725 1202 GasA TA Y  
## 4 Unf 0 501 1127 GasA Ex Y  
## 5 Unf 0 293 896 GasA Ex Y  
## 6 <NA> 0 0 0 GasA Ex Y  
## Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath  
## 1 SBrkr 1575 626 0 2201 0  
## 2 SBrkr 1272 928 0 2200 2  
## 3 SBrkr 2158 0 0 2158 1  
## 4 SBrkr 1236 872 0 2108 0  
## 5 SBrkr 1112 896 0 2008 1  
## 6 SBrkr 495 1427 0 1922 0  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual  
## 1 0 2 0 4 1 Gd  
## 2 0 2 2 4 2 TA  
## 3 0 2 0 4 1 Gd  
## 4 0 2 0 4 1 Gd  
## 5 0 3 0 3 1 Ex  
## 6 0 3 0 4 1 Gd  
## TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt  
## 1 8 Typ 1 Gd Attchd 1966  
## 2 9 Typ 0 <NA> <NA> NA  
## 3 7 Typ 1 Gd Attchd 1950  
## 4 7 Typ 2 TA Basment 1935  
## 5 8 Typ 0 <NA> Attchd 1948  
## 6 7 Typ 1 Ex BuiltIn 2003  
## GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive  
## 1 Unf 2 432 Gd Gd Y  
## 2 <NA> 0 0 <NA> <NA> N  
## 3 Unf 2 576 TA TA Y  
## 4 Unf 2 540 TA TA Y  
## 5 Unf 1 230 TA TA Y  
## 6 RFn 2 672 TA TA Y  
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 1 586 236 0 0 0 738 Gd  
## 2 0 70 0 0 0 0 <NA>  
## 3 0 29 39 0 0 0 <NA>  
## 4 0 0 0 0 90 0 <NA>  
## 5 103 0 0 0 0 0 <NA>  
## 6 0 0 177 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 GdPrv <NA> 0 8 2006 WD Alloca 274970  
## 2 GdPrv <NA> 0 9 2008 WD Normal 145900  
## 3 MnPrv <NA> 0 6 2007 WD Normal 243000  
## 4 <NA> <NA> 0 7 2008 WD Normal 235000  
## 5 <NA> <NA> 0 8 2006 WD Normal 225000  
## 6 <NA> <NA> 0 7 2009 ConLD Normal 198500  
## LogSalePrice LogpredictedSalePrice predictedSalePrice  
## 1 12.52442 12.20555 199895.4  
## 2 11.89068 12.20511 199806.7  
## 3 12.40082 12.18647 196116.9  
## 4 12.36734 12.16428 191813.1  
## 5 12.32386 12.11990 183486.7  
## 6 12.19854 12.08173 176615.7

#Build the Edwards regression model considering GrLivArea and Neighborhood  
model\_Edwards <- lm(LogSalePrice ~ GrLivArea, data = Edwards)  
summary(model\_Edwards)

##   
## Call:  
## lm(formula = LogSalePrice ~ GrLivArea, data = Edwards)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.71071 -0.16825 0.00629 0.18214 0.73596   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.103e+01 9.133e-02 120.720 < 2e-16 \*\*\*  
## GrLivArea 5.387e-04 6.968e-05 7.731 1.05e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2441 on 96 degrees of freedom  
## Multiple R-squared: 0.3837, Adjusted R-squared: 0.3773   
## F-statistic: 59.76 on 1 and 96 DF, p-value: 1.051e-11

confint(model\_Edwards)

## 2.5 % 97.5 %  
## (Intercept) 1.084419e+01 1.120677e+01  
## GrLivArea 4.003457e-04 6.769635e-04

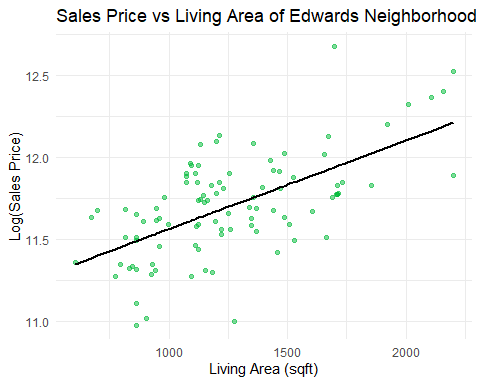
#plot the predicted SalePrice vs. the actual SalePrice  
Edwards$LogpredictedPrice <- predict(model\_Edwards, newdata = Edwards)  
Edwards$predictedPrice <- exp(Edwards$LogpredictedPrice)  
  
# Find CV PRESS  
cvpress\_E <- sum((Edwards$SalePrice - Edwards$predictedPrice)^2)  
print(paste("CV PRESS:", cvpress\_E))

## [1] "CV PRESS: 108435928010.778"

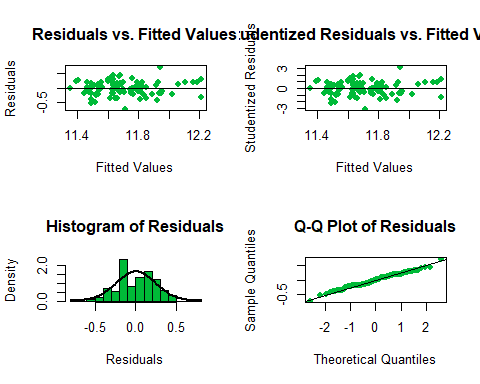
# Calculate AIC   
aic\_E <- AIC(model\_Edwards)  
print(paste("AIC:", aic\_E))

## [1] "AIC: 5.67226565590208"

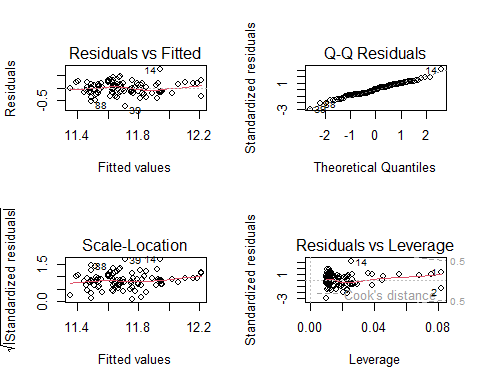
# Plot Edwards GrLiveArea vs SalePrice  
ggplot(Edwards, aes(x = GrLivArea, y = LogSalePrice)) +  
 geom\_point(alpha = 0.5, color = "#00BA38") +  
 geom\_line(aes(y = LogpredictedPrice), color = "black", size = 1) +  
 labs(title = "Sales Price vs Living Area of Edwards Neighborhood",   
 x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 theme\_minimal()



# Check Assumptions  
par(mfrow = c(2, 2))  
  
# Residuals vs Fitted values  
plot(fitted(model\_Edwards), residuals(model\_Edwards),  
 xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs. Fitted Values", pch = 19, col="#00BA38")  
abline(h = 0, col = "black")  
  
 # Studentized Residuals  
plot(fitted(model\_Edwards), rstudent(model\_Edwards),  
 xlab = "Fitted Values", ylab = "Studentized Residuals", main = "Studentized Residuals vs. Fitted Values",   
 pch = 19, col="#00BA38")  
abline(h = 0, col = "black")  
   
# Histogram of residuals  
hist(residuals(model\_Edwards),  
 xlab = "Residuals", main = "Histogram of Residuals", col = "#00BA38", border = "black", breaks = 20, probability=TRUE)  
curve(dnorm(x, mean = mean(residuals(model\_Edwards)), sd = sd(residuals(model\_Edwards))), col = "black", lwd = 2, add = TRUE)  
  
# Q-Q plot of residuals  
qqnorm(residuals(model\_Edwards), main = "Q-Q Plot of Residuals", col="#00BA38", pch=19)   
qqline(residuals(model\_Edwards), col = "black")



# Check Assumptions  
par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid  
plot(model\_Edwards)



BrkSide <- trainNoOutliers %>%  
 filter(Neighborhood %in% c("BrkSide"))  
head(BrkSide)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 1 329 75 RL NA 11888 Pave Pave IR1  
## 2 975 70 RL 60 11414 Pave <NA> IR1  
## 3 1415 50 RL 64 13053 Pave Pave Reg  
## 4 1264 70 RL 60 13515 Pave Pave Reg  
## 5 498 50 RL 60 9120 Pave Pave Reg  
## 6 523 50 RM 50 5000 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2  
## 1 Bnk AllPub Inside Gtl BrkSide PosN Norm  
## 2 Lvl AllPub Corner Gtl BrkSide RRAn Feedr  
## 3 Bnk AllPub Inside Gtl BrkSide Norm Norm  
## 4 Lvl AllPub Inside Gtl BrkSide Norm Norm  
## 5 Lvl AllPub Inside Gtl BrkSide Norm Norm  
## 6 Lvl AllPub Corner Gtl BrkSide Feedr Norm  
## BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle  
## 1 1Fam 2.5Unf 6 6 1916 1994 Gable  
## 2 1Fam 2Story 7 8 1910 1993 Gable  
## 3 1Fam 1.5Fin 6 7 1923 2000 Gambrel  
## 4 1Fam 2Story 6 6 1919 1950 Gambrel  
## 5 1Fam 1.5Fin 7 6 1925 1950 Gable  
## 6 1Fam 1.5Fin 6 7 1947 1950 Gable  
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond  
## 1 CompShg Wd Sdng Wd Shng None 0 TA TA  
## 2 CompShg HdBoard HdBoard None 0 TA Gd  
## 3 CompShg Wd Sdng Wd Sdng None 0 TA TA  
## 4 CompShg Wd Sdng Wd Sdng None 0 TA TA  
## 5 CompShg Wd Sdng Wd Sdng None 0 TA Gd  
## 6 CompShg CemntBd CmentBd None 0 TA Gd  
## Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1  
## 1 BrkTil TA TA No Unf 0  
## 2 BrkTil Gd TA No Unf 0  
## 3 BrkTil TA TA No Unf 0  
## 4 PConc TA TA No Unf 0  
## 5 PConc TA TA No Rec 329  
## 6 CBlock TA TA No ALQ 399  
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir  
## 1 Unf 0 844 844 GasA Gd N  
## 2 Unf 0 728 728 GasA TA N  
## 3 Unf 0 833 833 GasA Gd Y  
## 4 Unf 0 764 764 GasA Ex Y  
## 5 Unf 0 697 1026 GasA Ex Y  
## 6 Unf 0 605 1004 GasA Ex Y  
## Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath  
## 1 FuseA 1445 689 0 2134 0  
## 2 SBrkr 1136 883 0 2019 0  
## 3 SBrkr 1053 795 0 1848 0  
## 4 FuseA 1060 764 0 1824 0  
## 5 SBrkr 1133 687 0 1820 1  
## 6 SBrkr 1004 660 0 1664 0  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual  
## 1 0 2 0 5 1 Gd  
## 2 0 1 0 3 1 Gd  
## 3 0 1 1 4 1 Gd  
## 4 0 1 0 3 1 TA  
## 5 0 2 0 4 1 TA  
## 6 0 2 0 3 1 TA  
## TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt  
## 1 10 Typ 0 <NA> Detchd 1930  
## 2 8 Typ 0 <NA> Detchd 1997  
## 3 8 Typ 1 Gd Detchd 1922  
## 4 8 Typ 1 Gd Detchd 1940  
## 5 8 Typ 0 <NA> Detchd 1925  
## 6 7 Typ 2 Gd Detchd 1950  
## GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive  
## 1 Unf 2 441 TA TA Y  
## 2 Unf 2 532 TA TA Y  
## 3 Unf 2 370 TA TA N  
## 4 Unf 2 520 TA TA N  
## 5 Unf 1 240 TA TA N  
## 6 Unf 2 420 TA TA Y  
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 1 0 60 268 0 0 0 <NA>  
## 2 509 135 0 0 0 0 <NA>  
## 3 0 0 0 0 220 0 <NA>  
## 4 0 0 126 0 0 0 <NA>  
## 5 0 100 0 0 0 0 <NA>  
## 6 0 24 36 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice  
## 1 <NA> <NA> 0 7 2009 WD Normal 214500  
## 2 GdPrv <NA> 0 10 2009 WD Normal 167500  
## 3 <NA> <NA> 0 6 2008 WD Normal 207000  
## 4 GdPrv <NA> 0 7 2007 WD Normal 180500  
## 5 GdPrv <NA> 0 6 2008 WD Normal 184000  
## 6 <NA> <NA> 0 10 2006 WD Normal 159000  
## LogSalePrice LogpredictedSalePrice predictedSalePrice  
## 1 12.27607 12.17582 194039.2  
## 2 12.02874 12.12478 184384.7  
## 3 12.24047 12.04889 170909.8  
## 4 12.10349 12.03824 169099.1  
## 5 12.12269 12.03647 168799.2  
## 6 11.97666 11.96723 157508.3

#Build the Brookside regression model considering GrLivArea and Neighborhood  
model\_BrkSide <- lm(LogSalePrice ~ GrLivArea, data = BrkSide)  
summary(model\_BrkSide)

##   
## Call:  
## lm(formula = LogSalePrice ~ GrLivArea, data = BrkSide)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.51067 -0.08746 -0.01194 0.12308 0.36852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.079e+01 8.561e-02 126.05 < 2e-16 \*\*\*  
## GrLivArea 7.382e-04 6.780e-05 10.89 1.92e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1979 on 56 degrees of freedom  
## Multiple R-squared: 0.6792, Adjusted R-squared: 0.6734   
## F-statistic: 118.5 on 1 and 56 DF, p-value: 1.917e-15

confint(model\_BrkSide)

## 2.5 % 97.5 %  
## (Intercept) 1.062009e+01 1.096310e+01  
## GrLivArea 6.023957e-04 8.740498e-04

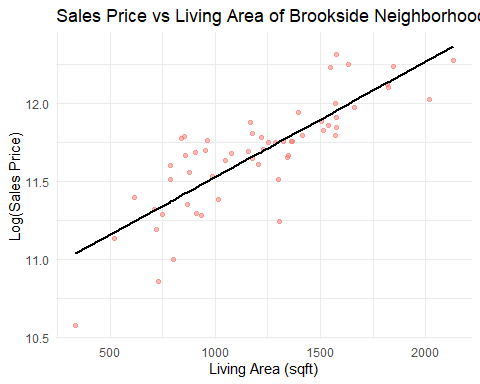
# Obtain the predicted SalePrice  
BrkSide$LogpredictedPrice <- predict(model\_BrkSide, newdata = BrkSide)  
BrkSide$predictedPrice <- exp(BrkSide$LogpredictedPrice)  
  
# Find CV PRESS  
cvpress\_B <- sum((BrkSide$SalePrice - BrkSide$predictedPrice)^2)  
print(paste("CV PRESS:", cvpress\_B))

## [1] "CV PRESS: 29704361945.6668"

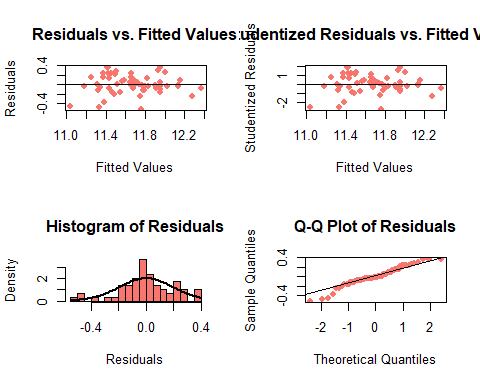
# Calculate AIC   
aic\_B <- AIC(model\_BrkSide)  
print(paste("AIC:", aic\_B))

## [1] "AIC: -19.3516339791371"

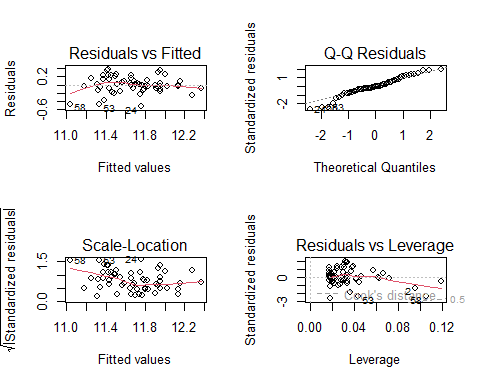
# Plot Brookside GrLivArea vs LogSale Price  
ggplot(BrkSide, aes(x = GrLivArea, y = LogSalePrice)) +  
 geom\_point(alpha = 0.5, color = "#F8766D") +  
 geom\_line(aes(y = LogpredictedPrice), color = "black", size = 1) +  
 labs(title = "Sales Price vs Living Area of Brookside Neighborhood",   
 x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 theme\_minimal()



# Check Assumptions  
  
par(mfrow = c(2, 2))  
  
# Residuals vs Fitted values  
plot(fitted(model\_BrkSide), residuals(model\_BrkSide),  
 xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs. Fitted Values", pch = 19, col="#F8766D")  
abline(h = 0, col = "black")  
  
 # Studentized Residuals  
plot(fitted(model\_BrkSide), rstudent(model\_BrkSide),  
 xlab = "Fitted Values", ylab = "Studentized Residuals", main = "Studentized Residuals vs. Fitted Values",   
 pch = 19, col="#F8766D")  
abline(h = 0, col = "black")  
   
# Histogram of residuals  
hist(residuals(model\_BrkSide),  
 xlab = "Residuals", main = "Histogram of Residuals", col = "#F8766D", border = "black", breaks = 20, probability=TRUE)  
curve(dnorm(x, mean = mean(residuals(model\_BrkSide)), sd = sd(residuals(model\_BrkSide))), col = "black", lwd = 2, add = TRUE)  
  
# Q-Q plot of residuals  
qqnorm(residuals(model\_BrkSide), main = "Q-Q Plot of Residuals", col="#F8766D", pch=19)   
qqline(residuals(model\_BrkSide), col = "black")



# Check assumptions  
par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid  
plot(model\_BrkSide)



# Fit linear regression models separately for each neighborhood  
lmmodel\_NAmes <- lm(SalePrice ~ GrLivArea, data = filter(trainNoOutliers, Neighborhood == "NAmes"))  
lmmodel\_Edwards <- lm(SalePrice ~ GrLivArea, data = filter(trainNoOutliers, Neighborhood == "Edwards"))  
lmmodel\_Brkside <- lm(SalePrice ~ GrLivArea, data = filter(trainNoOutliers, Neighborhood == "BrkSide"))  
  
# Create a base plot  
ggplot(trainNoOutliers, aes(x = GrLivArea, y = LogSalePrice, color = Neighborhood, shape=Neighborhood)) +  
 geom\_point(alpha = 0.5) + # Scatter plot  
 geom\_smooth(data = filter(trainNoOutliers, Neighborhood == "NAmes"), method = "lm", se = FALSE, color = "blue") +  
 geom\_smooth(data = filter(trainNoOutliers, Neighborhood == "Edwards"), method = "lm", se = FALSE, color = "green3") +   
 geom\_smooth(data = filter(trainNoOutliers, Neighborhood == "BrkSide"), method = "lm", se = FALSE, color = "red") +   
 labs(title = "Sales Price vs Living Area by Neighborhood",  
 x = "Living Area (sqft)", y = "Log(Sales Price)") +  
 scale\_color\_discrete(labels = c("NAmes" = "North Ames", "Edwards" = "Edwards", "BrkSide" = "Brookside")) +  
 scale\_shape\_discrete(labels = c("NAmes" = "North Ames", "Edwards" = "Edwards", "BrkSide" = "Brookside")) +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'  
## `geom\_smooth()` using formula = 'y ~ x'  
## `geom\_smooth()` using formula = 'y ~ x'

