Ames Housing Data: Analysis Question 1

Burton-Cordova & Thibeaux

2023-04-04

Load Data

#Load Test Data  
test = read.csv('https://github.com/athibeaux/MSDS-DDS/raw/main/Project/test.csv', header = TRUE, fill = TRUE)  
  
#Load Train Data  
train = read.csv('https://github.com/athibeaux/MSDS-DDS/raw/main/Project/train.csv', header = TRUE, fill = NA)  
  
#Summary of Train Data  
summary(train)

## Id MSSubClass MSZoning LotFrontage   
## Min. : 1.0 Min. : 20.0 Length:1460 Min. : 21.00   
## 1st Qu.: 365.8 1st Qu.: 20.0 Class :character 1st Qu.: 59.00   
## Median : 730.5 Median : 50.0 Mode :character Median : 69.00   
## Mean : 730.5 Mean : 56.9 Mean : 70.05   
## 3rd Qu.:1095.2 3rd Qu.: 70.0 3rd Qu.: 80.00   
## Max. :1460.0 Max. :190.0 Max. :313.00   
## NA's :259   
## LotArea Street Alley LotShape   
## Min. : 1300 Length:1460 Length:1460 Length:1460   
## 1st Qu.: 7554 Class :character Class :character Class :character   
## Median : 9478 Mode :character Mode :character Mode :character   
## Mean : 10517   
## 3rd Qu.: 11602   
## Max. :215245   
##   
## LandContour Utilities LotConfig LandSlope   
## Length:1460 Length:1460 Length:1460 Length:1460   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Neighborhood Condition1 Condition2 BldgType   
## Length:1460 Length:1460 Length:1460 Length:1460   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## HouseStyle OverallQual OverallCond YearBuilt   
## Length:1460 Min. : 1.000 Min. :1.000 Min. :1872   
## Class :character 1st Qu.: 5.000 1st Qu.:5.000 1st Qu.:1954   
## Mode :character Median : 6.000 Median :5.000 Median :1973   
## Mean : 6.099 Mean :5.575 Mean :1971   
## 3rd Qu.: 7.000 3rd Qu.:6.000 3rd Qu.:2000   
## Max. :10.000 Max. :9.000 Max. :2010   
##   
## YearRemodAdd RoofStyle RoofMatl Exterior1st   
## Min. :1950 Length:1460 Length:1460 Length:1460   
## 1st Qu.:1967 Class :character Class :character Class :character   
## Median :1994 Mode :character Mode :character Mode :character   
## Mean :1985   
## 3rd Qu.:2004   
## Max. :2010   
##   
## Exterior2nd MasVnrType MasVnrArea ExterQual   
## Length:1460 Length:1460 Min. : 0.0 Length:1460   
## Class :character Class :character 1st Qu.: 0.0 Class :character   
## Mode :character Mode :character Median : 0.0 Mode :character   
## Mean : 103.7   
## 3rd Qu.: 166.0   
## Max. :1600.0   
## NA's :8   
## ExterCond Foundation BsmtQual BsmtCond   
## Length:1460 Length:1460 Length:1460 Length:1460   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2   
## Length:1460 Length:1460 Min. : 0.0 Length:1460   
## Class :character Class :character 1st Qu.: 0.0 Class :character   
## Mode :character Mode :character Median : 383.5 Mode :character   
## Mean : 443.6   
## 3rd Qu.: 712.2   
## Max. :5644.0   
##   
## BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating   
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Length:1460   
## 1st Qu.: 0.00 1st Qu.: 223.0 1st Qu.: 795.8 Class :character   
## Median : 0.00 Median : 477.5 Median : 991.5 Mode :character   
## Mean : 46.55 Mean : 567.2 Mean :1057.4   
## 3rd Qu.: 0.00 3rd Qu.: 808.0 3rd Qu.:1298.2   
## Max. :1474.00 Max. :2336.0 Max. :6110.0   
##   
## HeatingQC CentralAir Electrical X1stFlrSF   
## Length:1460 Length:1460 Length:1460 Min. : 334   
## Class :character Class :character Class :character 1st Qu.: 882   
## Mode :character Mode :character Mode :character Median :1087   
## Mean :1163   
## 3rd Qu.:1391   
## Max. :4692   
##   
## X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath   
## Min. : 0 Min. : 0.000 Min. : 334 Min. :0.0000   
## 1st Qu.: 0 1st Qu.: 0.000 1st Qu.:1130 1st Qu.:0.0000   
## Median : 0 Median : 0.000 Median :1464 Median :0.0000   
## Mean : 347 Mean : 5.845 Mean :1515 Mean :0.4253   
## 3rd Qu.: 728 3rd Qu.: 0.000 3rd Qu.:1777 3rd Qu.:1.0000   
## Max. :2065 Max. :572.000 Max. :5642 Max. :3.0000   
##   
## BsmtHalfBath FullBath HalfBath BedroomAbvGr   
## Min. :0.00000 Min. :0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000   
## Median :0.00000 Median :2.000 Median :0.0000 Median :3.000   
## Mean :0.05753 Mean :1.565 Mean :0.3829 Mean :2.866   
## 3rd Qu.:0.00000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :2.00000 Max. :3.000 Max. :2.0000 Max. :8.000   
##   
## KitchenAbvGr KitchenQual TotRmsAbvGrd Functional   
## Min. :0.000 Length:1460 Min. : 2.000 Length:1460   
## 1st Qu.:1.000 Class :character 1st Qu.: 5.000 Class :character   
## Median :1.000 Mode :character Median : 6.000 Mode :character   
## Mean :1.047 Mean : 6.518   
## 3rd Qu.:1.000 3rd Qu.: 7.000   
## Max. :3.000 Max. :14.000   
##   
## Fireplaces FireplaceQu GarageType GarageYrBlt   
## Min. :0.000 Length:1460 Length:1460 Min. :1900   
## 1st Qu.:0.000 Class :character Class :character 1st Qu.:1961   
## Median :1.000 Mode :character Mode :character Median :1980   
## Mean :0.613 Mean :1979   
## 3rd Qu.:1.000 3rd Qu.:2002   
## Max. :3.000 Max. :2010   
## NA's :81   
## GarageFinish GarageCars GarageArea GarageQual   
## Length:1460 Min. :0.000 Min. : 0.0 Length:1460   
## Class :character 1st Qu.:1.000 1st Qu.: 334.5 Class :character   
## Mode :character Median :2.000 Median : 480.0 Mode :character   
## Mean :1.767 Mean : 473.0   
## 3rd Qu.:2.000 3rd Qu.: 576.0   
## Max. :4.000 Max. :1418.0   
##   
## GarageCond PavedDrive WoodDeckSF OpenPorchSF   
## Length:1460 Length:1460 Min. : 0.00 Min. : 0.00   
## Class :character Class :character 1st Qu.: 0.00 1st Qu.: 0.00   
## Mode :character Mode :character Median : 0.00 Median : 25.00   
## Mean : 94.24 Mean : 46.66   
## 3rd Qu.:168.00 3rd Qu.: 68.00   
## Max. :857.00 Max. :547.00   
##   
## EnclosedPorch X3SsnPorch ScreenPorch PoolArea   
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.00 Median : 0.00 Median : 0.00 Median : 0.000   
## Mean : 21.95 Mean : 3.41 Mean : 15.06 Mean : 2.759   
## 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :552.00 Max. :508.00 Max. :480.00 Max. :738.000   
##   
## PoolQC Fence MiscFeature MiscVal   
## Length:1460 Length:1460 Length:1460 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 0.00   
## Mode :character Mode :character Mode :character Median : 0.00   
## Mean : 43.49   
## 3rd Qu.: 0.00   
## Max. :15500.00   
##   
## MoSold YrSold SaleType SaleCondition   
## Min. : 1.000 Min. :2006 Length:1460 Length:1460   
## 1st Qu.: 5.000 1st Qu.:2007 Class :character Class :character   
## Median : 6.000 Median :2008 Mode :character Mode :character   
## Mean : 6.322 Mean :2008   
## 3rd Qu.: 8.000 3rd Qu.:2009   
## Max. :12.000 Max. :2010   
##   
## SalePrice   
## Min. : 34900   
## 1st Qu.:129975   
## Median :163000   
## Mean :180921   
## 3rd Qu.:214000   
## Max. :755000   
##

#Select Relevant Columns and Neighborhoods for Analysis Question 1  
C21 = train %>% select(GrLivArea,Neighborhood,SalePrice) %>% filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood == "BrkSide")  
  
C21$Neighborhood <- as.factor(C21$Neighborhood)  
summary(C21)

## GrLivArea Neighborhood SalePrice   
## Min. : 334 BrkSide: 58 Min. : 39300   
## 1st Qu.:1003 Edwards:100 1st Qu.:116000   
## Median :1200 NAmes :225 Median :135500   
## Mean :1302 Mean :138063   
## 3rd Qu.:1496 3rd Qu.:155000   
## Max. :5642 Max. :345000

#Check and Remove NA's  
sum(is.na(C21$GrLIvArea))

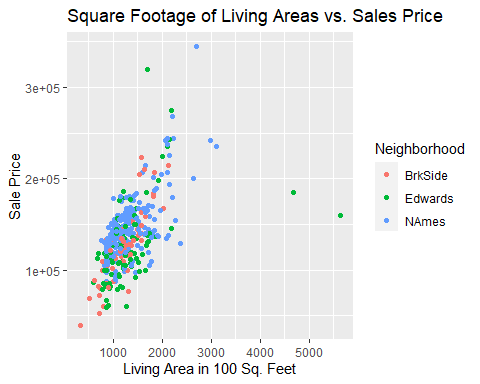
## [1] 0

Checking Assumptions Residual Plots Influential point analysis (Cook’s D and Leverage) Make sure to address each assumption.

# Addressing Assumptions (Linearity, Variance, Normality, and Independence)

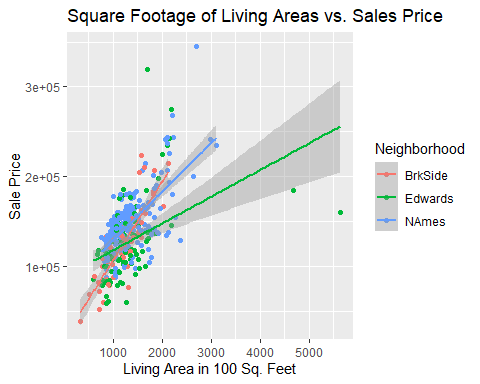
## Linearity

# Without Lines  
C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) + geom\_point() +  
 xlab("Living Area in 100 Sq. Feet") + ylab("Sale Price") +  
 ggtitle("Square Footage of Living Areas vs. Sales Price")



# With Lines for each Neighborhood  
C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) + geom\_point() +  
 geom\_smooth(method = "lm") +   
 xlab("Living Area in 100 Sq. Feet") + ylab("Sale Price") +  
 ggtitle("Square Footage of Living Areas vs. Sales Price")

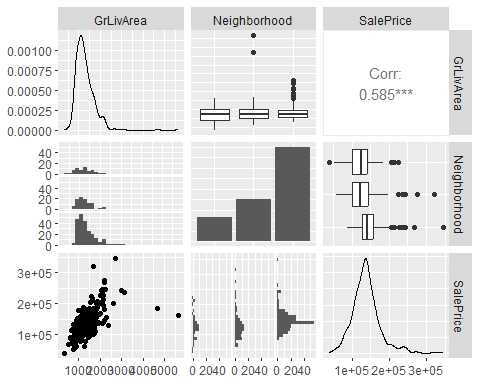
## `geom\_smooth()` using formula = 'y ~ x'



## Normality

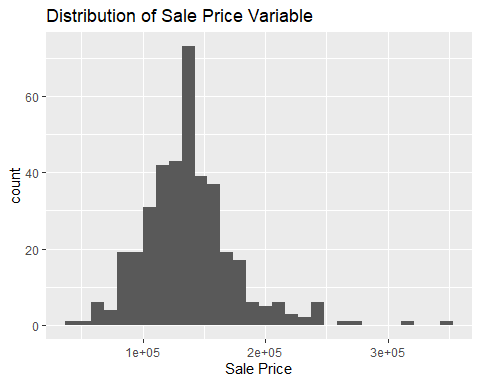
ggpairs(C21)

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



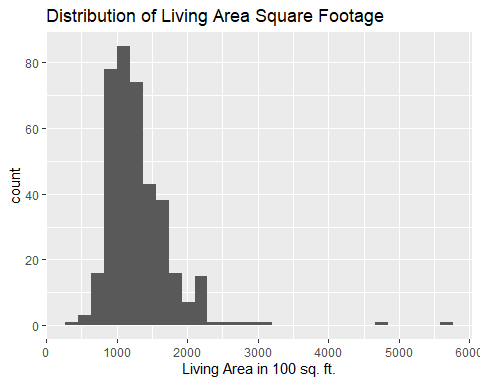
# Histogram for Sale Price  
C21 %>% ggplot() + geom\_histogram(aes(x = SalePrice)) +   
 ggtitle("Distribution of Sale Price Variable") + xlab("Sale Price")

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

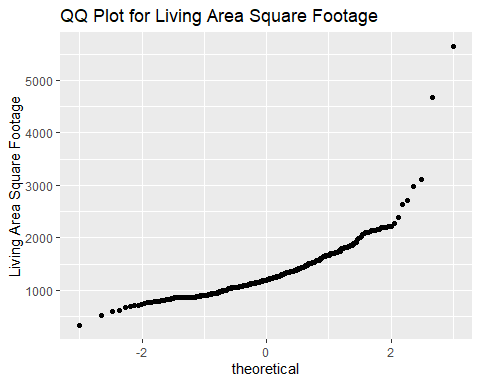


# Histogram for Living Area Square Footage  
C21 %>% ggplot() + geom\_histogram(aes(x = GrLivArea)) +  
 ggtitle("Distribution of Living Area Square Footage") +  
 xlab("Living Area in 100 sq. ft.")

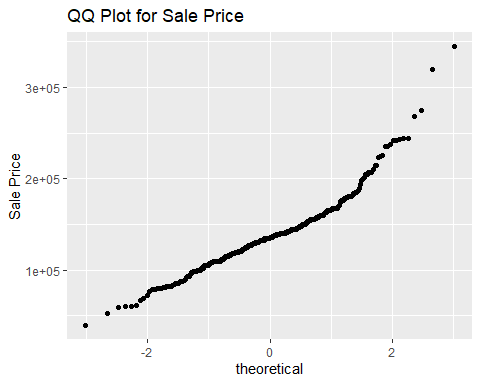
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# QQ Plot for Living Area Square Footage  
C21 %>% ggplot() + geom\_qq(aes(sample = GrLivArea)) +   
 ggtitle("QQ Plot for Living Area Square Footage") + ylab("Living Area Square Footage")

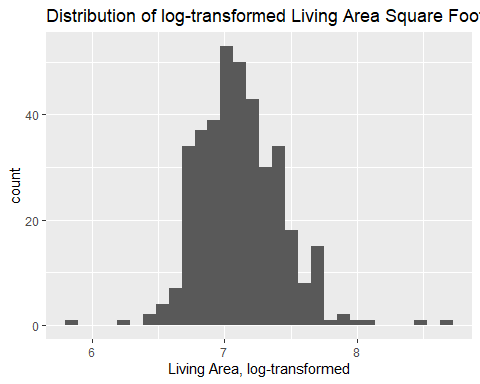


# QQ Plot for Sale Price  
C21 %>% ggplot() + geom\_qq(aes(sample = SalePrice)) +   
 ggtitle("QQ Plot for Sale Price") + ylab("Sale Price")

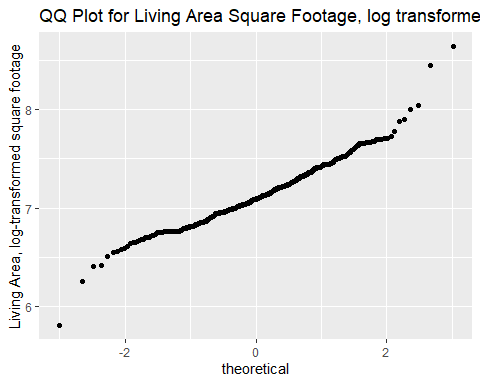


# Log Transformation on GrLivArea  
C21$logLivArea = log(C21$GrLivArea)  
  
# Histogram for Log-Transformed Living Area Square Footage  
C21 %>% ggplot() + geom\_histogram(aes(x = logLivArea)) +  
 xlab("Living Area, log-transformed") +  
 ggtitle("Distribution of log-transformed Living Area Square Footage")

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

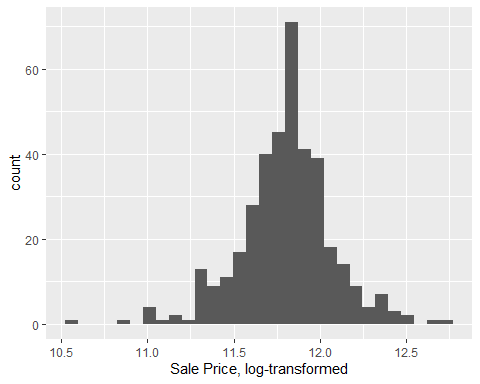


# QQ Plot for Log-Transformed Living Area Square Footage  
C21 %>% ggplot() + geom\_qq(aes(sample = logLivArea)) +   
 ggtitle("QQ Plot for Living Area Square Footage, log transformed") +   
 ylab("Living Area, log-transformed square footage")



# Log Transformation on Sale Price  
C21$logprice = log(C21$SalePrice)  
  
# Histogram for Log-Transformed Sale Price  
C21 %>% ggplot() + geom\_histogram(aes(x = logprice)) +  
 xlab("Sale Price, log-transformed")

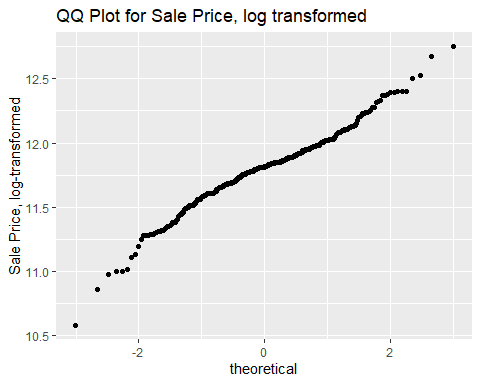
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggtitle("Distribution of log-transformed Sale Price")

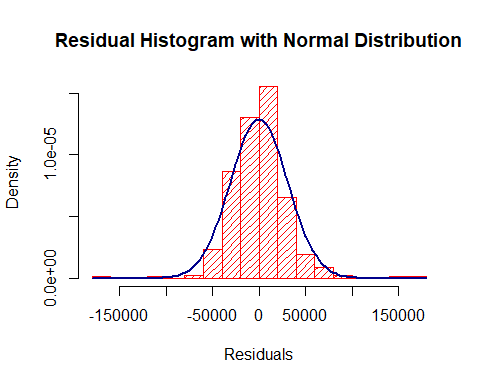
## $title  
## [1] "Distribution of log-transformed Sale Price"  
##   
## attr(,"class")  
## [1] "labels"

# QQ Plot for Log-Transformed Living Area Square Footage  
C21 %>% ggplot() + geom\_qq(aes(sample = logprice)) +   
 ggtitle("QQ Plot for Sale Price, log transformed") +   
 ylab("Sale Price, log-transformed")

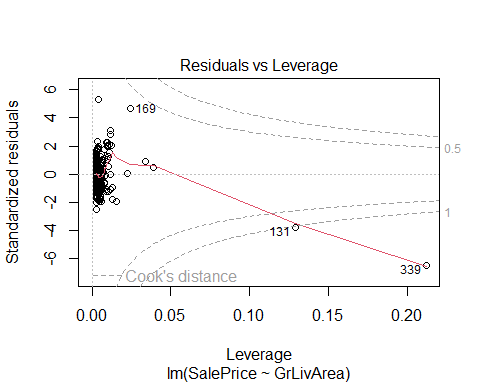
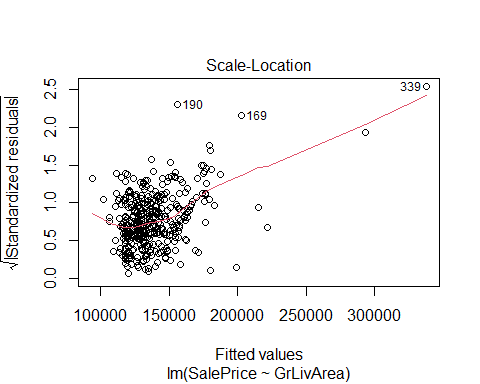
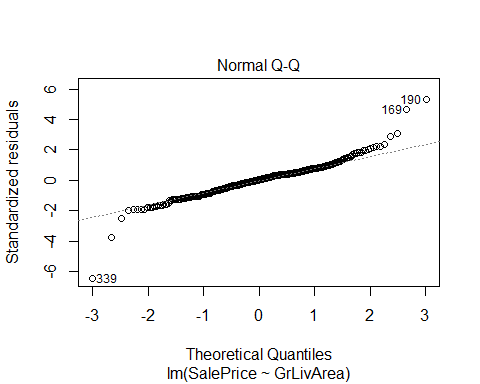
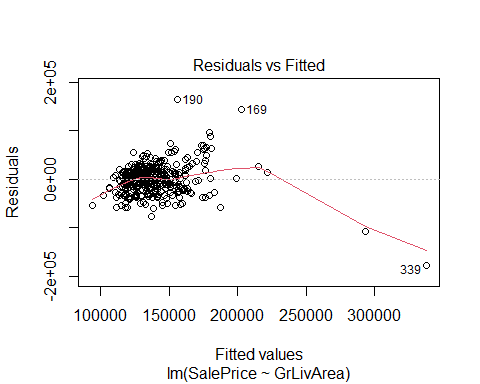


## Equal Variance

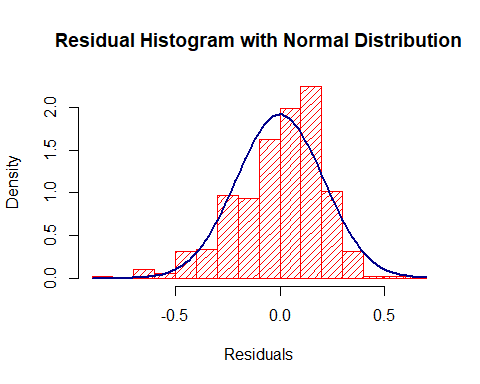
# Non Transformed Data  
rd <- lm(SalePrice ~ GrLivArea, data = C21)  
g = rd$residuals  
m<-mean(g)  
std<-sqrt(var(g))  
hist(g, density=20, breaks=20, prob=TRUE, col="red",  
 xlab="Residuals",   
 main="Residual Histogram with Normal Distribution")  
curve(dnorm(x, mean=m, sd=std),   
 col="darkblue", lwd=2, add=TRUE, yaxt="n")



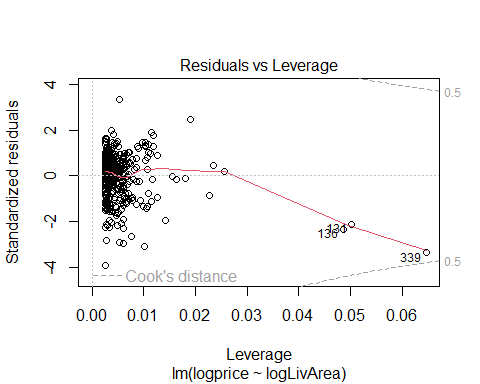
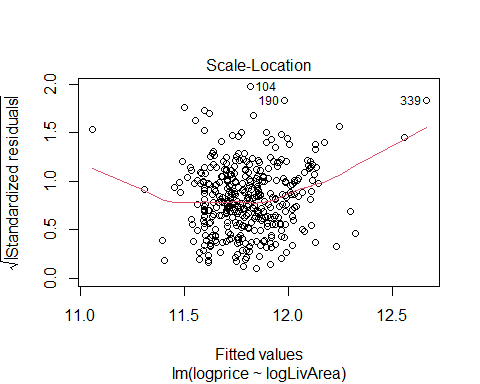
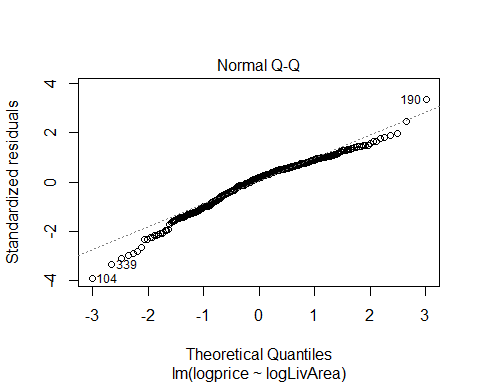
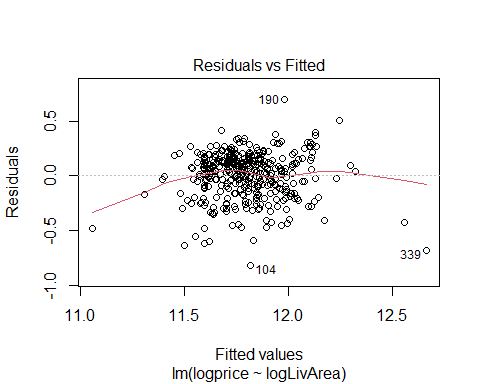
# Display residual plots for non-transformed data  
plot(rd)



# Notice cluster effect of residuals!  
  
# Log-Log Model   
  
log\_rd <- lm(logprice ~ logLivArea, data = C21)  
log\_g = log\_rd$residuals  
log\_m<-mean(log\_g)  
log\_std<-sqrt(var(log\_g))  
hist(log\_g, density=20, breaks=20, prob=TRUE, col="red",  
 xlab="Residuals",   
 main="Residual Histogram with Normal Distribution")  
curve(dnorm(x, mean=log\_m, sd=log\_std),   
 col="darkblue", lwd=2, add=TRUE, yaxt="n")



# Display residual plots for log-transformed data  
plot(log\_rd)

 We have a few high leverage residuals. Let’s take a closer look.

# Examine the largest values in the explanatory variable  
SortBySqFt <- C21[order(C21$GrLivArea, decreasing = TRUE),]  
head(SortBySqFt)

## GrLivArea Neighborhood SalePrice logLivArea logprice  
## 339 5642 Edwards 160000 8.637994 11.98293  
## 131 4676 Edwards 184750 8.450198 12.12676  
## 53 3112 NAmes 235000 8.043021 12.36734  
## 78 2978 NAmes 242000 7.999007 12.39669  
## 169 2704 NAmes 345000 7.902487 12.75130  
## 351 2634 NAmes 200000 7.876259 12.20607

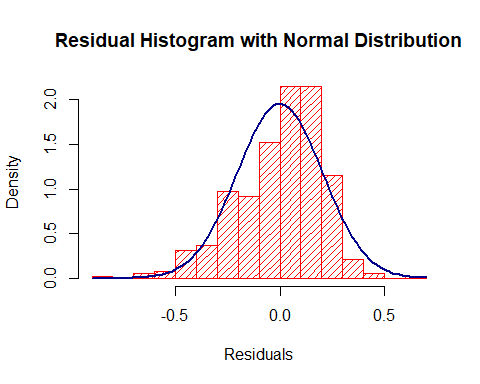
# Wow! Two houses with >4000 square feet were sold for under $200k?? Either that's a mistake (like someone left out a zero) or I need the number of that real estate agent IMMEDIATELY so he can get me that kind of price on a mansion! If these two outliers are an valid observation, we may be looking at a short sale or foreclosure. Since these kinds of sales are not what Century 21 wants to measure, we feel confident in removing them from our dataset.  
  
# Save complete dataset as a separate object in case we need it later  
C21\_full <- C21  
  
# Find rows with outlier data  
which(C21[ , 1] > 4000)

## [1] 131 339

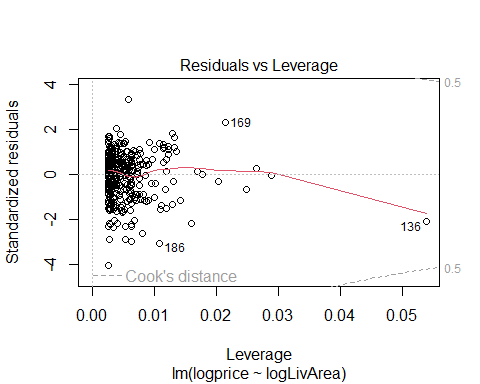
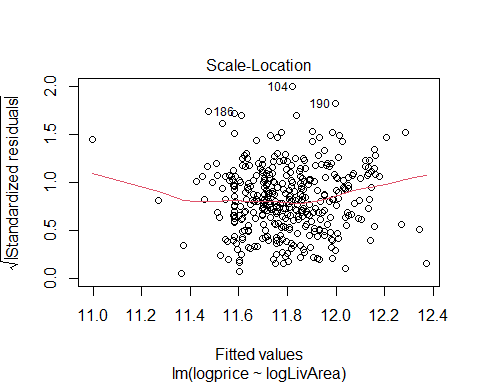
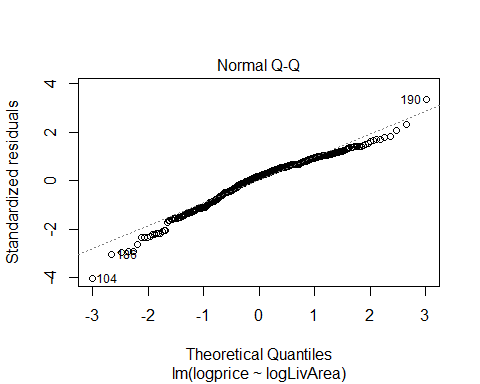
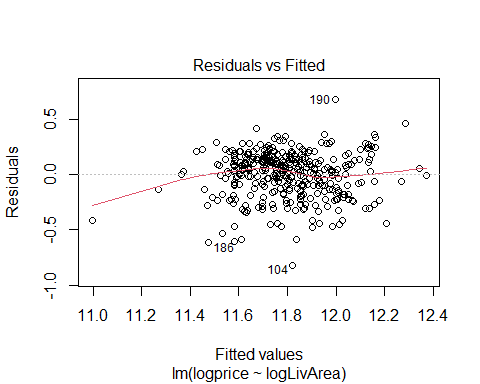
# Those are the same datapoints highlighted in our Residuals vs Leverage plot, so let's remove them.  
C21 <- C21[-c(131,339),]

Now that we’ve removed the invalid data points, let’s look again at the residuals.

# Log-Log Model   
  
log\_rd <- lm(logprice ~ logLivArea, data = C21)  
log\_g = log\_rd$residuals  
log\_m<-mean(log\_g)  
log\_std<-sqrt(var(log\_g))  
hist(log\_g, density=20, breaks=20, prob=TRUE, col="red",  
 xlab="Residuals",   
 main="Residual Histogram with Normal Distribution")  
curve(dnorm(x, mean=log\_m, sd=log\_std),   
 col="darkblue", lwd=2, add=TRUE, yaxt="n")



# Display new residual plots  
plot(log\_rd)

 Now let’s take another look at the fits

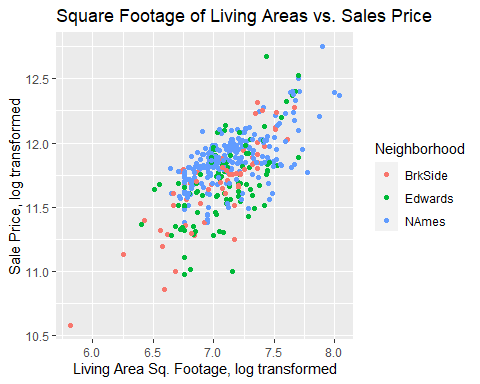
# Parameter Estimates  
# Overall  
fit = lm(logprice ~ logLivArea, data = C21)  
summary(fit)

##   
## Call:  
## lm(formula = logprice ~ logLivArea, data = C21)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.81988 -0.12331 0.03752 0.13699 0.67784   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.41038 0.25290 29.30 <2e-16 \*\*\*  
## logLivArea 0.61688 0.03553 17.36 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2044 on 379 degrees of freedom  
## Multiple R-squared: 0.443, Adjusted R-squared: 0.4415   
## F-statistic: 301.4 on 1 and 379 DF, p-value: < 2.2e-16

confint(fit)

## 2.5 % 97.5 %  
## (Intercept) 6.9131180 7.9076462  
## logLivArea 0.5470181 0.6867395

# Without Lines  
C21 %>% ggplot(aes(logLivArea, logprice, color = Neighborhood)) +  
 geom\_point() +  
 ylab("Sale Price, log transformed") +   
 xlab("Living Area Sq. Footage, log transformed") +   
 ggtitle("Square Footage of Living Areas vs. Sales Price")



# By Neighborhood  
fit\_hoods = C21 %>% group\_by(Neighborhood) %>% do(model = lm(logprice ~ logLivArea, data = .))  
  
# Brookside  
summary(fit\_hoods[[2]][[1]])

##   
## Call:  
## lm(formula = logprice ~ logLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5486 -0.0973 -0.0205 0.1025 0.3695   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.91292 0.49539 11.94 <2e-16 \*\*\*  
## logLivArea 0.81965 0.07032 11.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1888 on 56 degrees of freedom  
## Multiple R-squared: 0.7081, Adjusted R-squared: 0.7029   
## F-statistic: 135.9 on 1 and 56 DF, p-value: < 2.2e-16

confint(fit\_hoods[[2]][[1]])

## 2.5 % 97.5 %  
## (Intercept) 4.9205338 6.9053077  
## logLivArea 0.6787747 0.9605214

# Edwards  
summary(fit\_hoods[[2]][[2]])

##   
## Call:  
## lm(formula = logprice ~ logLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73636 -0.16251 0.01752 0.17602 0.74523   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.92309 0.64232 10.778 < 2e-16 \*\*\*  
## logLivArea 0.67334 0.09037 7.451 4.05e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2475 on 96 degrees of freedom  
## Multiple R-squared: 0.3664, Adjusted R-squared: 0.3598   
## F-statistic: 55.52 on 1 and 96 DF, p-value: 4.049e-11

confint(fit\_hoods[[2]][[2]])

## 2.5 % 97.5 %  
## (Intercept) 5.6481051 8.1980807  
## logLivArea 0.4939554 0.8527198

# North Ames  
summary(fit\_hoods[[2]][[3]])

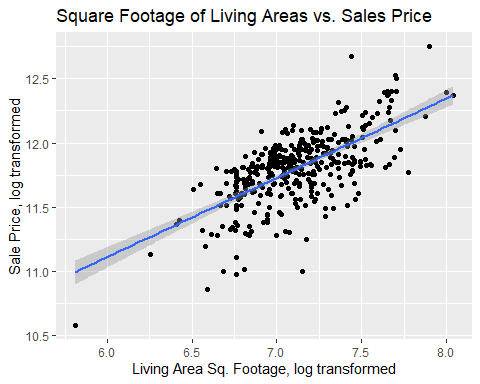
##   
## Call:  
## lm(formula = logprice ~ logLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46095 -0.07958 0.02882 0.09510 0.52051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.49273 0.26603 31.92 <2e-16 \*\*\*  
## logLivArea 0.47302 0.03725 12.70 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1577 on 223 degrees of freedom  
## Multiple R-squared: 0.4196, Adjusted R-squared: 0.417   
## F-statistic: 161.2 on 1 and 223 DF, p-value: < 2.2e-16

confint(fit\_hoods[[2]][[3]])

## 2.5 % 97.5 %  
## (Intercept) 7.9684756 9.0169796  
## logLivArea 0.3996113 0.5464359

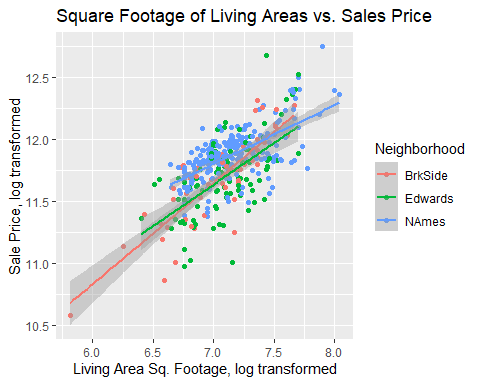
# With Regression Line  
C21 %>% ggplot(aes(logLivArea, logprice)) +   
 geom\_point() + geom\_smooth(method = "lm") +  
 ylab("Sale Price, log transformed") +   
 xlab("Living Area Sq. Footage, log transformed") +   
 ggtitle("Square Footage of Living Areas vs. Sales Price")

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



# With Lines for each Neighborhood  
C21 %>% ggplot(aes(logLivArea, logprice, color = Neighborhood)) +   
 geom\_point() + geom\_smooth(method = "lm") +  
 ylab("Sale Price, log transformed") +   
 xlab("Living Area Sq. Footage, log transformed") +   
 ggtitle("Square Footage of Living Areas vs. Sales Price")

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

 Compare competing models

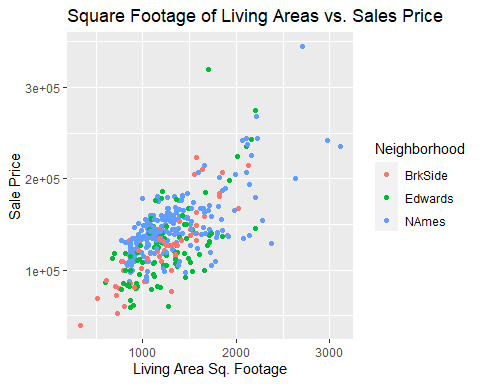
# Fit on non-transformed data  
# Parameter Estimates  
# Overall  
fit = lm(SalePrice ~ GrLivArea, data = C21)  
summary(fit)

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = C21)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80420 -19186 2318 17146 154997   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54415.162 4888.348 11.13 <2e-16 \*\*\*  
## GrLivArea 65.128 3.644 17.87 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28150 on 379 degrees of freedom  
## Multiple R-squared: 0.4573, Adjusted R-squared: 0.4559   
## F-statistic: 319.3 on 1 and 379 DF, p-value: < 2.2e-16

confint(fit)

## 2.5 % 97.5 %  
## (Intercept) 44803.48173 64026.84149  
## GrLivArea 57.96214 72.29407

# Without Lines  
C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +  
 geom\_point() +  
 ylab("Sale Price") +   
 xlab("Living Area Sq. Footage") +   
 ggtitle("Square Footage of Living Areas vs. Sales Price")



# By Neighborhood  
fit\_hoods = C21 %>% group\_by(Neighborhood) %>% do(model = lm(SalePrice ~ GrLivArea, data = .))  
  
# Brookside  
summary(fit\_hoods[[2]][[1]])

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -57306 -12367 -4445 11780 66160   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19971.51 9684.72 2.062 0.0438 \*   
## GrLivArea 87.16 7.67 11.364 3.63e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22390 on 56 degrees of freedom  
## Multiple R-squared: 0.6975, Adjusted R-squared: 0.6921   
## F-statistic: 129.1 on 1 and 56 DF, p-value: 3.635e-16

confint(fit\_hoods[[2]][[1]])

## 2.5 % 97.5 %  
## (Intercept) 570.69729 39372.3303  
## GrLivArea 71.79729 102.5278

# Edwards  
summary(fit\_hoods[[2]][[2]])

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -68374 -23330 -2316 20416 159563   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31428.555 12650.543 2.484 0.0147 \*   
## GrLivArea 75.976 9.651 7.872 5.29e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33810 on 96 degrees of freedom  
## Multiple R-squared: 0.3923, Adjusted R-squared: 0.386   
## F-statistic: 61.97 on 1 and 96 DF, p-value: 5.291e-12

confint(fit\_hoods[[2]][[2]])

## 2.5 % 97.5 %  
## (Intercept) 6317.42478 56539.68441  
## GrLivArea 56.81882 95.13405

# North Ames  
summary(fit\_hoods[[2]][[3]])

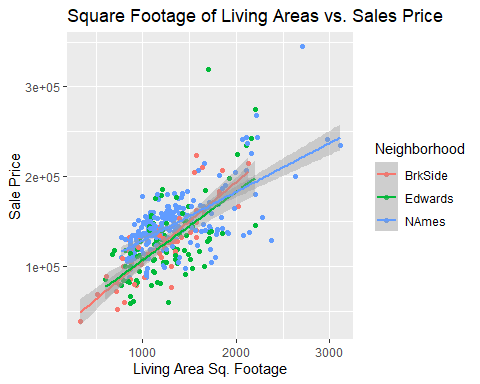
##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -74948 -12263 2796 12598 123454   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 74676.402 5401.761 13.82 <2e-16 \*\*\*  
## GrLivArea 54.316 3.932 13.81 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24340 on 223 degrees of freedom  
## Multiple R-squared: 0.4611, Adjusted R-squared: 0.4587   
## F-statistic: 190.8 on 1 and 223 DF, p-value: < 2.2e-16

confint(fit\_hoods[[2]][[3]])

## 2.5 % 97.5 %  
## (Intercept) 64031.37338 85321.42969  
## GrLivArea 46.56687 62.06485

# With Lines for each Neighborhood  
C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +   
 geom\_point() + geom\_smooth(method = "lm") +  
 ylab("Sale Price") +   
 xlab("Living Area Sq. Footage") +   
 ggtitle("Square Footage of Living Areas vs. Sales Price")

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

 Hold Neighborhood Constant

# Transformed Data  
fit = lm(logprice ~ logLivArea + Neighborhood, data = C21)  
summary(fit)

##   
## Call:  
## lm(formula = logprice ~ logLivArea + Neighborhood, data = C21)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73253 -0.10572 0.02277 0.12232 0.77125   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.48892 0.23954 31.263 < 2e-16 \*\*\*  
## logLivArea 0.59565 0.03386 17.594 < 2e-16 \*\*\*  
## NeighborhoodEdwards -0.01405 0.03211 -0.438 0.662   
## NeighborhoodNAmes 0.12881 0.02867 4.492 9.39e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1934 on 377 degrees of freedom  
## Multiple R-squared: 0.5041, Adjusted R-squared: 0.5002   
## F-statistic: 127.8 on 3 and 377 DF, p-value: < 2.2e-16

confint(fit)

## 2.5 % 97.5 %  
## (Intercept) 7.01791044 7.9599297  
## logLivArea 0.52907953 0.6622165  
## NeighborhoodEdwards -0.07719457 0.0490859  
## NeighborhoodNAmes 0.07242562 0.1851868

# Raw Data  
fit = lm(SalePrice ~ GrLivArea + Neighborhood, data = C21)  
summary(fit)

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = C21)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -84825 -17044 978 13747 164980   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48379.397 5565.139 8.693 < 2e-16 \*\*\*  
## GrLivArea 63.550 3.544 17.933 < 2e-16 \*\*\*  
## NeighborhoodEdwards -1267.014 4517.747 -0.280 0.779284   
## NeighborhoodNAmes 14197.824 4029.477 3.523 0.000478 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27240 on 377 degrees of freedom  
## Multiple R-squared: 0.4945, Adjusted R-squared: 0.4905   
## F-statistic: 123 on 3 and 377 DF, p-value: < 2.2e-16

confint(fit)

## 2.5 % 97.5 %  
## (Intercept) 37436.79570 59321.99922  
## GrLivArea 56.58165 70.51772  
## NeighborhoodEdwards -10150.15274 7616.12542  
## NeighborhoodNAmes 6274.75739 22120.88993

Internal CV as a measure for competing models

# LOOCV method from library(caret), transformed data  
  
train(logprice ~ logLivArea, method = "lm", data = C21, trControl = trainControl(method = "LOOCV"))

## Linear Regression   
##   
## 381 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 380, 380, 380, 380, 380, 380, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.2051113 0.436098 0.162391  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

# LOOCV method from library(caret), raw data  
  
train(SalePrice ~ GrLivArea, method = "lm", data = C21, trControl = trainControl(method = "LOOCV"))

## Linear Regression   
##   
## 381 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 380, 380, 380, 380, 380, 380, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 28303.77 0.4485779 21806.77  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

# LOOCV method from library(caret), transformed data, holding Neighborhood constant  
  
train(logprice ~ logLivArea + Neighborhood, method = "lm", data = C21, trControl = trainControl(method = "LOOCV"))

## Linear Regression   
##   
## 381 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 380, 380, 380, 380, 380, 380, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.1948059 0.4913816 0.1481098  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

# LOOCV method from library(caret), raw data, holding Neighborhood constant  
  
train(SalePrice ~ GrLivArea + Neighborhood, method = "lm", data = C21, trControl = trainControl(method = "LOOCV"))

## Linear Regression   
##   
## 381 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 380, 380, 380, 380, 380, 380, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 27481.93 0.4801935 20199  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE