Assignment #2:
Oscar Hernandez

Introduction:

The purpose of this report is to provide results from the fitting of regression models that use the Ames Housing data. The housing data set contains 2930 observations spread across 83 variables which are either continuous, discrete, ordinal or nominal. The variables are representative of housing characteristics such as bathroom square footage, size of garage, and exterior quality. The report will summarize all aspects of the model building process. Specifically, the report will cover how the sample population data was determined, results from exploratory data analysis and finally a discussion of the (multiple) linear regression models that were built. The final goal is to determine which model is most appropriate to predict the value of a "typical" home.

Data:

As mentioned, it is necessary to define a "typical" home. The "typical" home will represent our population of interest and will ultimately become our sample population data. Therefore, we will create "drop conditions" from the original housing data. The "drop conditions" that have been chosen to represent our "typical" home will be placed in a waterfall which is displayed in Figure 1.

Drop Condition	Observations
01: Not Residential	168
02: Not SFR	440
03: Lot Area Greater than 20,000 Square Feet	81
04: Poor House Condition	11
05: Built Pre-1950	533
06: Poor Exterior Condition	152
07: Does not have all utilities	1
08: Home Non-Functional	5
09: Non-Normal Sale	267
10: Sale Price Greater than $400,000	26
99: Eligible Sample	1246

Figure 1: Drop Condition Waterfall

The "drop conditions" listed above represent items that are not representative of a "typical home" in Ames, Iowa. For example, the decision was made to exclude homes priced over $400,000 which resulted in 26 non-eligible observations. Furthermore, it makes sense to remove observations where the property sale took place in a non-residential zone (168 observations) and the property was not a single-family home which is indicated by the "SFR" tag (440 observations). The waterfall created provides us with a sample of 1246 eligible observations from the original 2930 in the housing data.

Prior to moving further along in the model building process, the decision was made to remove observations in which a missing value was present for any variable. The "na.omit" function was applied to our dataframe which reduced the 1246 observations to 973.

Exploratory Data Analysis

Before fitting the regression model, it was necessary to conduct a exploratory data analysis on two promising predictor variables – "LotArea" and "GrLivArea." First, boxplots were created for each of the predictor variables to check for any outliers, skewness and overall spread of the distribution. Figure 2 displays the boxplots that were created.

From Figure 2, we can see that there are several outliers for both variables which are heavily concentrated on the top half of the plots. The presence of these outliers can be problematic since they may affect the results of our model(s). Furthermore, it would appear that both variables are positively-skewed.

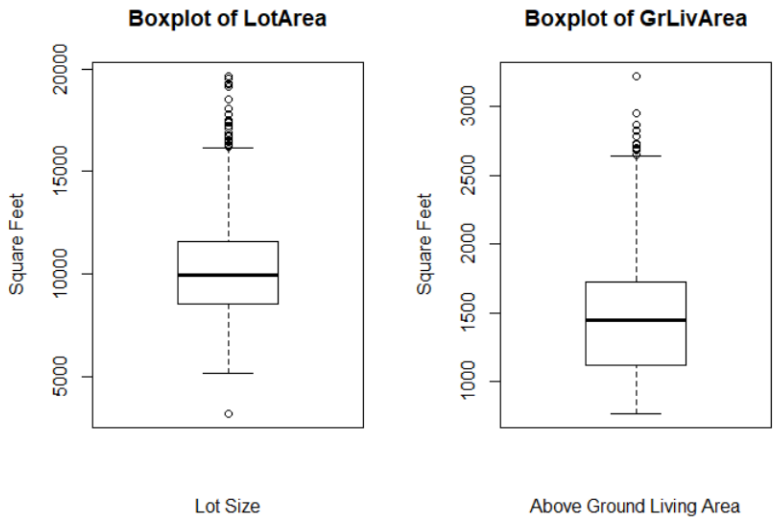

Figure 2: Side by Side Boxplot of LotArea and GrLivArea

Next, it is necessary to create a scatterplot between the predictor variables and the response variable SalePrice. The scatterplots will provide further evidence of a linear relationship between the two variables. Figure 3 displays the scatterplot of SalePrice and LotArea.

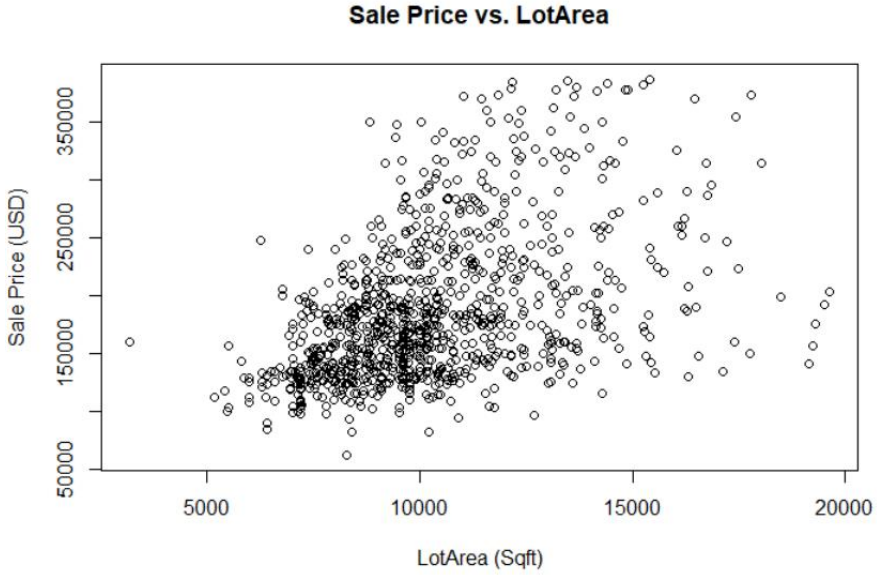

Figure 3: Scatterplot of SalePrice vs. LotArea

From Figure 3, we can see limited evidence of a linear relationship between SalePrice and LotArea. Specifically, the variation in the data becomes greater as LotArea increases past 10,000 square feet. It may be necessary to transform SalePrice to see if that offers better results. Overall, our initial exploration of these two variables leads us to think that LotArea isn't a strong predictor of SalePrice on its own.

Figure 4 is a scatterplot of SalePrice and GrLivArea. From the plot, we can see clear evidence of a linear relationship between the two variables. Specifically, it would appear that there is a strong correlation for observations between 500 and 1700 square feet and $100,000 and $200,000 sale price. Overall, there is some evidence here that leads us to believe that GrLivArea will be an appropriate predictor variable for SalePrice.

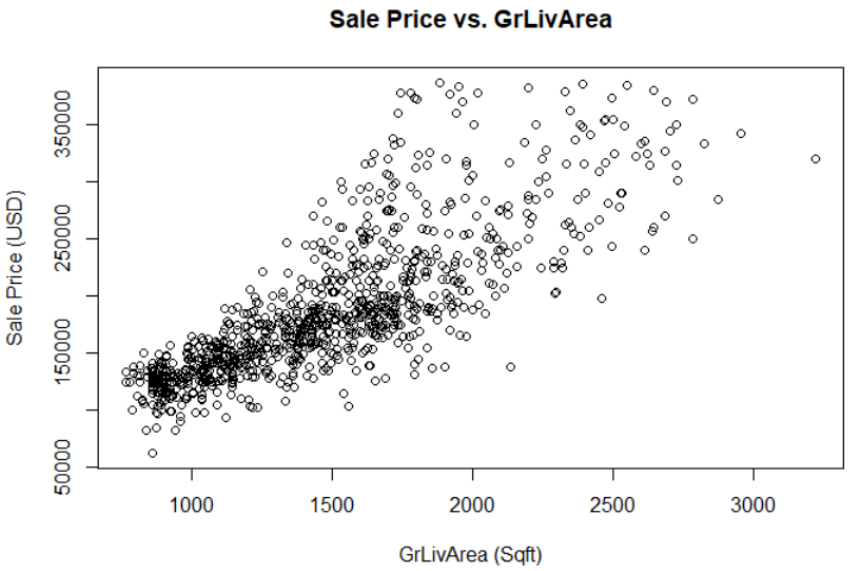

Figure 4: Scatterplot of SalePrice vs. GrLivArea

Simple Linear Regression Models

Now that an exploratory data analysis of our predictor variables is complete, we can fit our simple linear regression models. Using the "lm()" function in R, we will create "LotArea_Model," which is our first simple linear regression model. This model uses SalePrice as the response variable and LotArea as the predictor variable. Our second linear regression will be called "GrLivArea_Model" which keeps SalePrice as the response variable and uses GrLivArea as the predictor variable.

Figure 5 displays a side by side comparison of the key outputs from each model. From the table, we can see that each coefficient estimate was statistically significant as indicated by their corresponding p-values. An interesting part of the output is the significant difference in the R^2/Adj. R^2 for each model.

	SalePrice			SalePrice		
	B	*CI*	*p*	*B*	*CI*	*p*
(Intercept)	68690.06	53301.21 – 84078.92	<.001	19504.58	11270.23 – 27738.93	<.001
LotArea	11.50	10.05 – 12.96	<.001			
GrLivArea				112.92	107.60 – 118.23	<.001
Observations	973			973		
R^2 / adj. R^2	.199 / .198			.642 / .641		

Figure 5: Table of Model Outputs

LotArea_Model (LotArea Predictor Variable)

The next step in the model building process is to assess the model fit. The assessment will review normality and homoscedasticity using a Q-Q Plot and scatterplot with regards to the residuals from LotArea_Model. Figure 6 shows that the residuals from LotArea_Model are not necessarily normally distributed given that the data points are not tightly fit on the line.

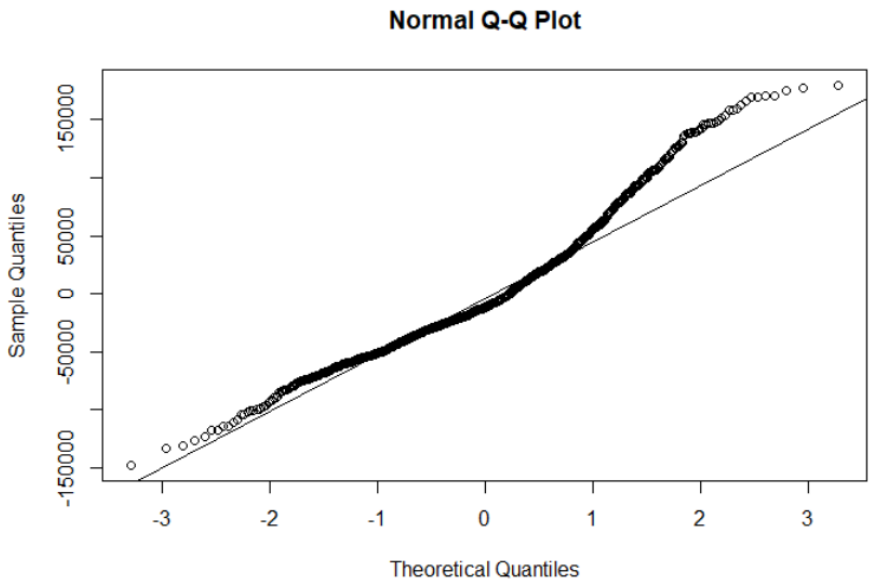

Figure 6: Q-Q Plot LotArea_Model Residuals

Next, we must view the relationship between LotArea_Model residuals and the LotArea variable. Figure 7 displays the scatterplot that is necessary to determine the level of homoscedasticity present. The plot provides limited evidence to support the goodness-of-fit for LotArea_Model.

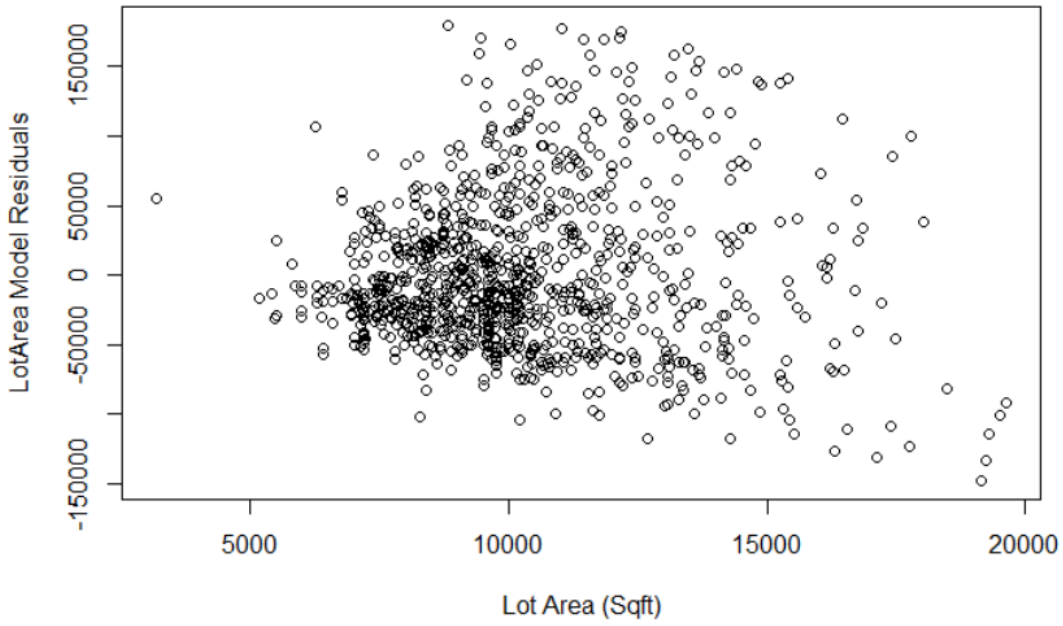

Figure 7: Scatterplot of LotArea_Model residuals and LotArea

GrLivArea_Model (GrLivArea Predictor Variable)

Figure 8 and 9 were created to explore the same assumptions for the GrLivArea_Model. From Figure 8, we can see the data points further along the line deviate in a more extreme fashion than in the LotArea_Model. As such, using Figure 8, it can be stated there is evidence to overturn the assumption of normality for the GrLivArea_Model.

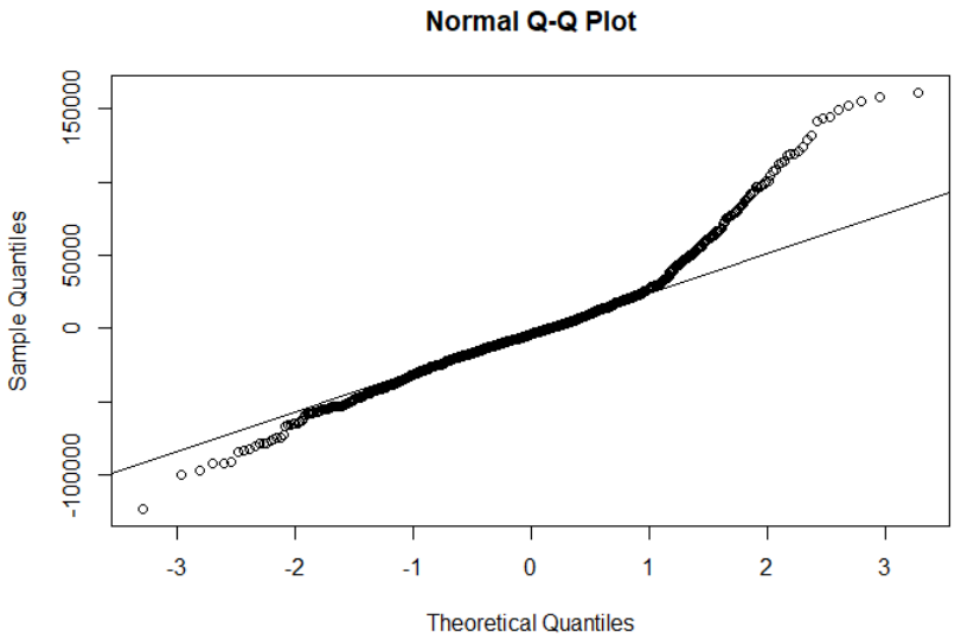

Figure 8: Q-Q Plot GrLivArea_Model Residuals

Figure 9 offers better evidence of the homoscedasticity assumption being valid for the GrLivArea_Model. With regards to goodness-of-fit, the GrLivArea_Model also shows mixed results.

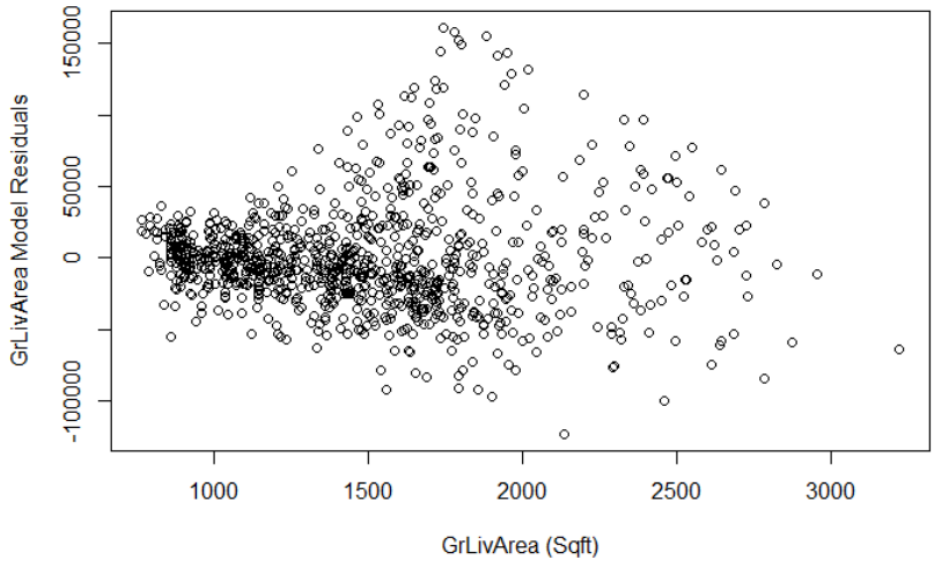

Figure 9: Scatterplot of GrLivArea_Model residuals and GrLivArea

Multiple Linear Regression Model

As part of the model building process, it makes sense to build a multiple linear regression model that includes the effect of both LotArea and GrLivArea on the response variable SalesPrice (Note: LotArea and GrLivArea have a correlation coefficient of 0.37). The same programming function was used to create our multiple linear regression model – "MLR_model." Figure 10 displays the relevant output for this new model. It would appear that the intercept coefficient is not as statistically significant as prior models based on its p-value but that's not overly concerning. Furthermore, we see that R^2/Adj. R^2 slightly improved by using both predictor variables.

	SalePrice		
	B	*CI*	*p*
(Intercept)	-12431.42	-23258.38 – -1604.46	.024
LotArea	4.40	3.38 – 5.41	<.001
GrLivArea	103.95	98.42 – 109.48	<.001
Observations		973	
R^2 / adj. R^2		.667 / .666	

Figure 10: Table of Model Outputs

To assess the goodness-of-fit of the MLR_model, the first step that will be taken is using a Q-Q Plot with its residuals. From Figure 11, we can see that the normality assumption appears to be violated for the residuals as indicated by the data points not tightly fitting on the line.

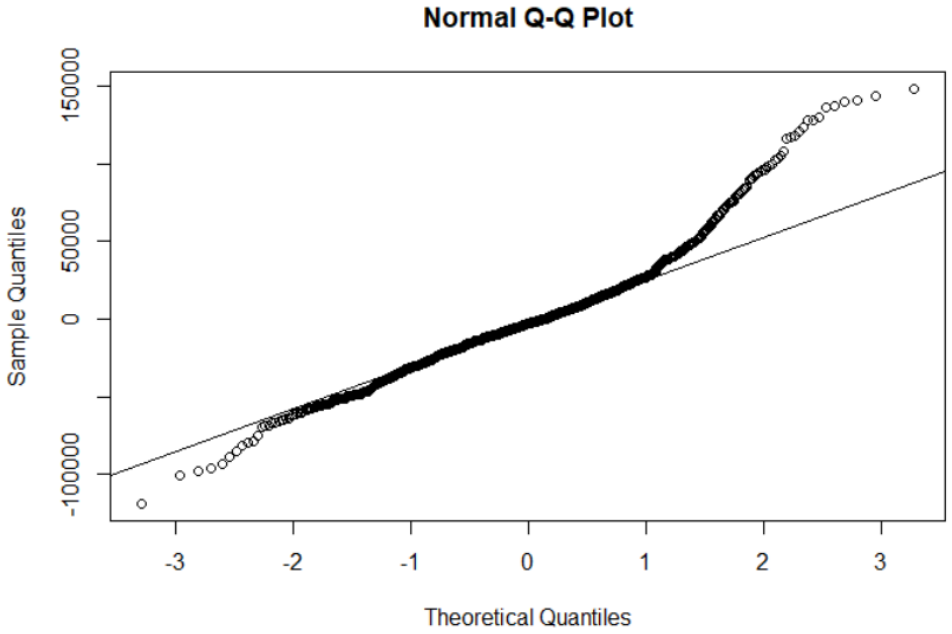

Figure 11: Q-Q Plot MLR_Model Residuals

Secondly, Figure 12 is a scatterplot of MLR_Model residuals versus MLR_Model fitted values. In Figure 12, we are looking to see if we can predict non-zero values for residuals based on the fitted values. Based on the variation in the data points, it would be difficult to predict those non-zero values. There is evidence to believe that the predictor variables in the MLR_Model can be good predictors of SalePrice.

Overall, the MLR_Model does fit better than the simple linear regression models from earlier in the report. More predictor variables in a model don't always mean that the model will fit better. However, in this case, the main criteria used when comparing the model fit was the results from Figure 12. This display showed us that the fitted values offer very limited ability to predict a residual value, which is a desirable quality for a regression model.

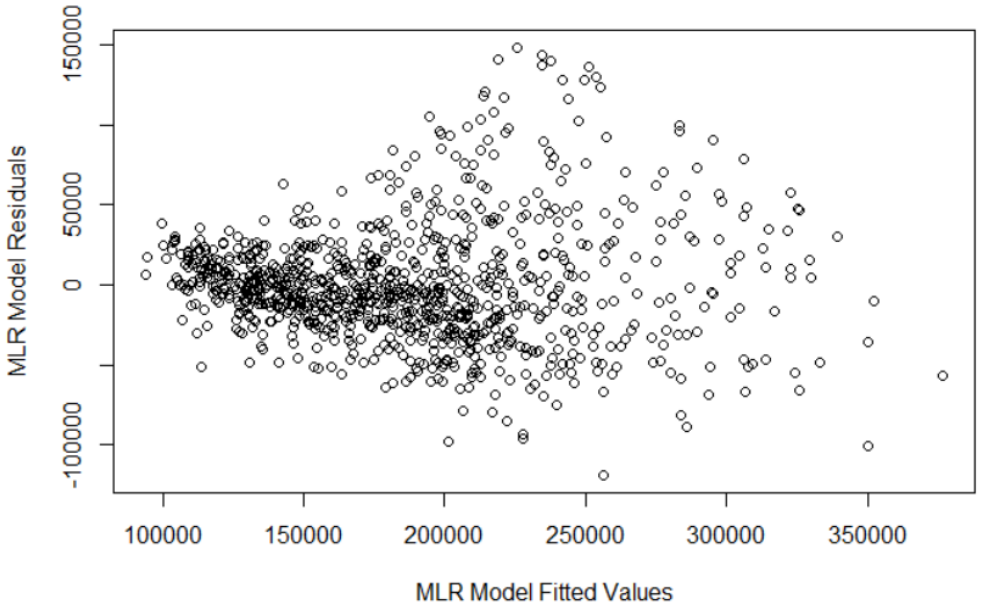

Figure 12: Scatterplot of MLR_Model Residuals and Fitted Values

Log SalePrice Response Models

As noted earlier, at times it may be necessary to transform variables to see how the transformation improves the model with regards to addressing our statistical problem – predicting the price of a "typical" home. Therefore, a log transformation of SalePrice was applied and all three of the previously created models were refitted with log(SalePrice).

With regards to the transformed models, the one major change that occurred with evaluating the normality assumption was with the LotArea_Model. In this case, the log(SalePrice) transformation resulted in the residuals from the refitted LotArea_Model being more closely fitted to the line in the plot. Figure 13 displays these changes.

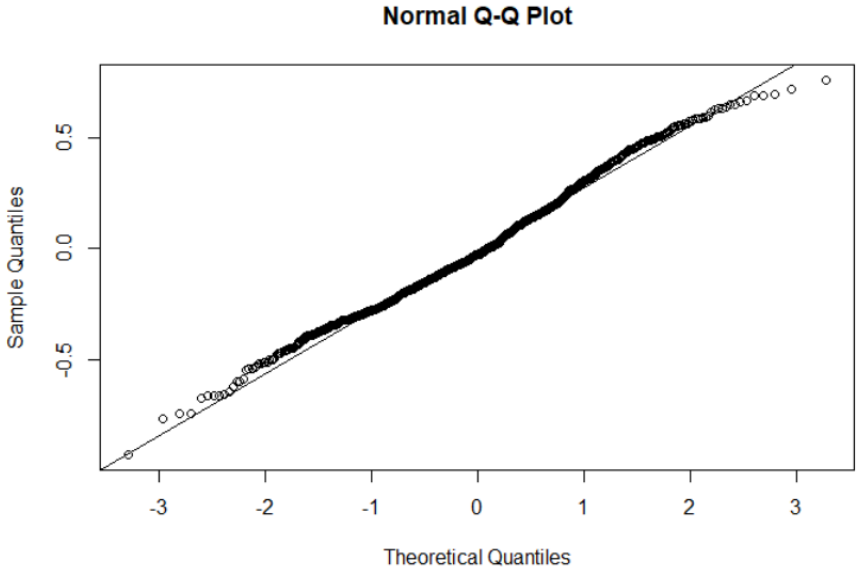

Figure 13: Q-Q Plot Refitted LotArea_Model Residuals

Another important change that transpired after the transformation was with the refitted MLR_Model residuals and fitted values plot. Figure 14 shows the changes that occurred as a result of the log(SalePrice) transformation. In this plot, the fitted values show a decreased correlation with the residuals. Furthermore, the increased variation among the data points make it so that a fitted value is less useful in predicting a residual, which is a noted desired characteristic.

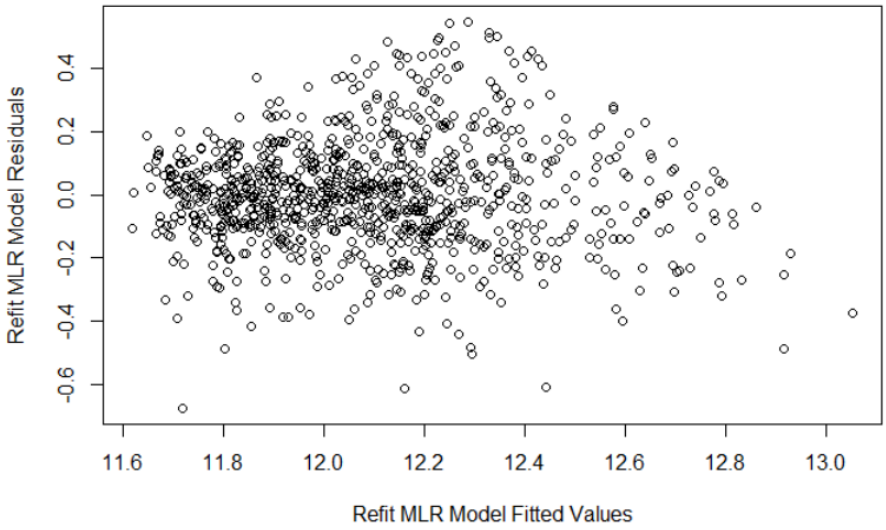

Figure 14: Scatterplot of Refitted MLR_Model Residuals and Fitted Values

Conclusion:

Overall, between LotArea and GrLivArea, it would appear that GrLivArea is a better predictor of SalePrice. The evidence outlined in this report gives reason as to why that is the case. Simply thinking about what the variables represent, it makes sense to think that the square footage of above ground living area (i.e. space where people can actually live in) would be better at predicting the value of a "typical" home as opposed to knowing the square footage of the lot where the home resides. Furthermore, in terms of goodness-of-fit, doing a log transformation of SalePrice didn't change much. Finally, in terms of answering our statistical problem, using both predictor variables together in a multiple linear regression model is the best approach.

Code:

```
#Read in our Ames Housing Data
ames_df <- read.csv('ames_housing_data.csv', header = TRUE, stringsAsFactors = FALSE)

ames_df$dropCondition <- ifelse(!ames_df$Zoning %in% c('RH','RL','RP','RM'), '01: Not Residential',
 ifelse(ames_df$BldgType!='1Fam','02: Not SFR',
 ifelse(ames_df$LotArea>20000,'03: Lot Area Greater than 20,000 Square Feet',
 ifelse(ames_df$OverallCond<3,'04: Poor House Condition',
 ifelse(ames_df$YearBuilt<1950,'05: Built Pre-1950',
 ifelse(!ames_df$ExterCond %in% c('Ex', 'GD', 'TA', 'Fa'),'06: Poor Exterior Condition',
 ifelse(ames_df$Utilities!='AllPub','07: Does not have all utilities',
 ifelse(!ames_df$Functional %in% c('Typ', 'Min1', 'Min2', 'Mod'),'08: Home Non Functional',
 ifelse(ames_df$SaleCondition!='Normal','09: Non-Normal Sale',
 ifelse(ames_df$SalePrice>400000, '10: Sale Price Greater than $400,000',
 '99: Eligible Sample'))))))))))


#Create a table summarizing Drop Conditions
table(ames_df$dropCondition)

#Save table as 'waterfall'
waterfall <- table(ames_df$dropCondition)

#Display the table as a column matrix
as.matrix(waterfall, 11,1)

# Eliminate all observations that are not part of the eligible sample population
eligible_population <- subset(ames_df,dropCondition=='99: Eligible Sample')

# Check that all remaining observations are eligible
table(eligible_population$dropCondition)

# Create a new DataFrame that only includes variables we like
keep.vars <- c('SID','PID','LotFrontage','LotArea','LotConfig','Neighborhood',
          'HouseStyle','OverallQual','OverallCond','YearBuilt','YearRemodel','Exterior1',
          'BsmtFinSF1','BsmtFinSF2','CentralAir','GrLivArea','BsmtFullBath','BsmtHalfBath',
          'FullBath','HalfBath','BedroomAbvGr','TotRmsAbvGrd','Fireplaces','GarageCars',
          'GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','ThreeSsnPorch',
          'ScreenPorch','PoolArea','MoSold','YrSold','SaleCondition','SalePrice')

#Subset eligible_population DataFrame with keep.vars vector
skinny_ames_df <- eligible_population[, keep.vars]

#Delete observations with missing values
skinny_ames_df <- na.omit(skinny_ames_df)


#Complete EDA on promising predictor variables in skinny_ames_df
#First, use boxplots to inspect distribution of two most promising variables
par(mfrow=c(1,2))
boxplot(skinny_ames_df$LotArea, main = 'Boxplot of LotArea', xlab= 'Lot Size', ylab='Square Feet')
```

```
boxplot(skinny_ames_df$GrLivArea, main = 'Boxplot of GrLivArea', xlab='Above Ground Living Area',
ylab = 'Square Feet')
par(mfrow=c(1,1))

#Next create scatterplot of SalePrice versus each promising predictor variable
plot(skinny_ames_df$LotArea, skinny_ames_df$SalePrice, ylab='Sale Price (USD)', main='Sale Price vs.
LotArea', xlab='LotArea (Sqft)')
plot(skinny_ames_df$GrLivArea, skinny_ames_df$SalePrice, ylab='Sale Price (USD)', main='Sale Price vs.
GrLivArea', xlab='GrLivArea (Sqft)')

#Fit a linear regression model using LotArea variable
LotArea_model <- lm(SalePrice ~ LotArea, data=skinny_ames_df)

#Fit a linear regression model using GrLivArea
GrLivArea_model <- lm(SalePrice ~ GrLivArea, data=skinny_ames_df)

#Display summary of LotArea_model
sjt.lm(LotArea_model)

#Display summary of GrLivArea_model
sjt.lm(GrLivArea_model)

#Display summary of both models
sjt.lm(LotArea_model, GrLivArea_model)

#Now, we need to produce some diagnostic plots to assess the goodness-of-fit of each model
#We will check the assumptions of normality and homoscedasticity with:
# 1. QQ Plot
# 2. Scatterplot of residuals versus predictor variable

#GOF plots for LotArea_model
#Plot for normaility
qqnorm(LotArea_model$residuals)
qqline(LotArea_model$residuals)
#Plot for homoscedasticity
plot(skinny_ames_df$LotArea,LotArea_model$residuals, xlab='Lot Area (Sqft)', ylab='LotArea Model
Residuals')

#GOF plots for GrLivArea_model
#Plot for normaility
qqnorm(GrLivArea_model$residuals)
qqline(GrLivArea_model$residuals)


#Plot for homoscedasticity
plot(skinny_ames_df$GrLivArea,GrLivArea_model$residuals, xlab='GrLivArea (Sqft)', ylab='GrLivArea
Model Residuals')


#Now we will built a multiple linear regrssion model using the same promising predictor variables
#Let's first look at the correlation between our predictor variables
cor(skinny_ames_df$LotArea, skinny_ames_df$GrLivArea)
```

```
#Fit the multiple linear regression model with both predictor variables
MLR_model <- lm(SalePrice ~ LotArea+GrLivArea, data=skinny_ames_df)

#GOF plots for MLR_model
#Plot for normaility
qqnorm(MLR_model$residuals)
qqline(MLR_model$residuals)
#Plot for homoscedasticity
plot(skinny_ames_df$LotArea,MLR_model$residuals, xlab='Lot Area (Sqft)', ylab='MLR Model Residuals')
plot(skinny_ames_df$GrLivArea,MLR_model$residuals, xlab='GrLivArea (Sqft)', ylab='MLR Model
Residuals')
plot(MLR_model$fitted.values,MLR_model$residuals, xlab='MLR Model Fitted Values', ylab='MLR Model
Residuals')

#Refit models for the transformed response variable log(SalePrice)
Refit_LotArea_model <- lm(log(SalePrice) ~ LotArea, data=skinny_ames_df)
Refit_GrLivArea_model <- lm(log(SalePrice) ~ GrLivArea, data=skinny_ames_df)
Refit_MLR_model <- lm(log(SalePrice) ~ LotArea+GrLivArea, data=skinny_ames_df)

#GOF plots for refitted models using log(SalePrice)
qqnorm(Refit_LotArea_model$residuals)
qqline(Refit_LotArea_model$residuals)
qqnorm(Refit_GrLivArea_model$residuals)
qqline(Refit_GrLivArea_model$residuals)
qqnorm(Refit_MLR_model$residuals)
qqline(Refit_MLR_model$residuals)
#Plot for homoscedasticity
plot(skinny_ames_df$LotArea,Refit_LotArea_model$residuals, xlab='Lot Area (Sqft)', ylab='Refit LotArea
Model Residuals')
plot(skinny_ames_df$GrLivArea,Refit_GrLivArea_model$residuals, xlab='GrLivArea (Sqft)', ylab='Refit
GrLivArea Model Residuals')
plot(skinny_ames_df$LotArea,Refit_MLR_model$residuals, xlab='Lot Area (Sqft)', ylab='Refit MLR Model
Residuals')
plot(skinny_ames_df$GrLivArea,Refit_MLR_model$residuals, xlab='GrLivArea (Sqft)', ylab='Refit MLR
Model Residuals')
plot(Refit_MLR_model$fitted.values,Refit_MLR_model$residuals, xlab='Refit MLR Model Fitted Values',
ylab='Refit MLR Model Residuals')
```