Ames Housing Analysis

Heesu Choi, Sickandar Akhthar, & Kylie Green

Contents:

1 Introduction

This project's goal is to build out a model that clients can use when they estimate house sale prices. This is a project originally hosted on the Kaggle website, as part of a regression analysis competition. There are 2 parts to this project. First, we will build a multiple linear regression model, and then, build a predictive model from using source data collected on houses sold in Ames, IA, between 2006 and 2010. As part of this project, we also generated an output of predicted results, on the test data from our best predictive model. We then uploaded these predicted results on the Kaggle website to generate our team's Kaggle score (Listed below). Based on the result of these two analysis, we will present our case for which variables, and what kind, are best to accurately predict a home SalePrice in Ames, IA, but also caution about what variables are least effective for the model.

Team Kaggle score: 0.13870

2 Data Description

We were initially given 80 variables to work with, with 1460 entries of data. For our first problem we chose to pick SalePrice as the variable we are trying to predict.

The data set for this project originally came from the Assessor's Office in Ames, IA. The initial file had 113 variables describing 3970 sales interactions/transactions that occurred there between 2006 and 2010. For this project, the variables from the Kaggle data set were reduced to 80 variables, that were directly related to the property sales. It is described as having 23 nominal, 23 ordinal, 14 discrete and 20 continuous. The kaggle dataset has 2919 entries, split into training dataset (1460) and testing dataset (1459). More information could be found out in this file. https://jse.amstat.org/v19n3/decock/DataDocumentation.txt

The data is sourced from Ames, IA Assessor office. It was collected from the Ames, IA's Assessor's Office by Dr. Dean De Cock of Truman State University. More information on this dataset, and the original collector's intent can be found here (https://jse.amstat.org/v19n3/decock.pdf).

Specific Variables used for our analysis for question 1/part 1 are:

- SalePrice (This is the dependent variable, continuous)
- GrLivArea (This is a continuous variable)
- Neighborhood. Categorical variable

The dataset does have some issues, in that there are some NA/Empty values in the data set that we needed to account for. More information on how we handled these NA/Empty values will be covered in the data cleaning section.

3 Data Cleaning

Prior to building any models, a crucial step is making sure we have clean data. For both Analysis I and Analysis II we used a logarithmic transformation on both GrLivArea and SalePrice.

In Analysis 1, we filtered the dataset for 3 neighborhoods (Brookside, North Ames, and Edwards). Once this was complete, the data for Analysis I was ready.

For our second analysis, we reviewed the variables/columns to check which columns had NULL values in the training dataset. The following columns had missing values and were removed from our training dataset. (LotFrontage, Alley, FireplaceQu, PoolQC, Fence, MiscFeature, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Electrical, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond). We also removed the Utilities column because the value remained constant throughout the datasets.

We created R code that dummy coded all of the predictors. After reviewing the dataset, we found several numeric columns that we dummy encoded to be categorized. These columns were MSSubClass, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MoSold, YrSold. We created code that grouped these variables into categorical data and created new dummy variables in our dataset.

For our testing data, we chose to replace missing values with measures of center from the respective columns from the training dataset.

We processed our data for Analysis II manually using Excel and created 3 files. We applied these changes to the train and test data separately. These new files (included in the supplemental zip folder) are cleaned_train_data_q2.csv, cleaned_test_data_q2.csv, and combined_dataset_q2.csv. The combined dataset includes both the testing and training data in one file for processing.

4 Analysis I

4.1 Problem Statement

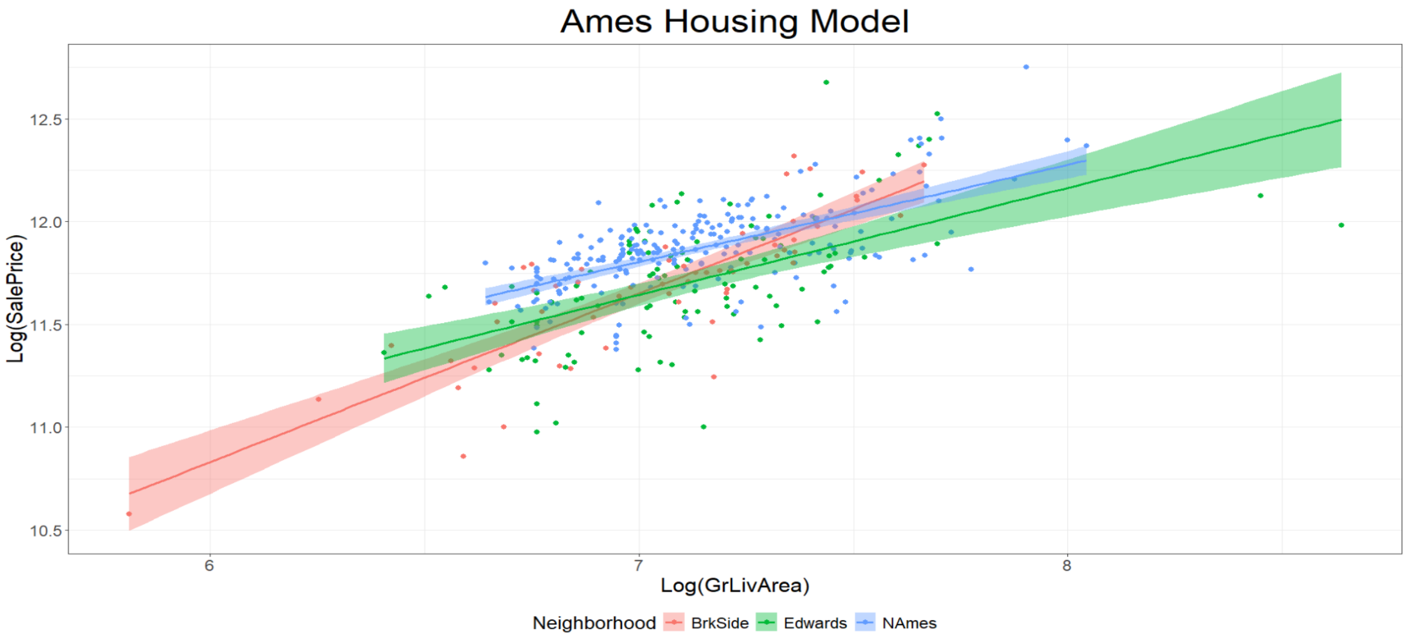
We would like to estimate the sales price of homes in the 3 neighborhoods sold by Century 21 in Ames, Iowa (North Ames, Edwards, and Brookside) based on the square footage of the home and the neighborhood in which the home is located in. For this part, we will build a multiple linear regression model, with an estimated regression equation and confidence intervals.

4.2 Model

4.2.1 Equation

$$\log(\hat{S}alePrice) = 5.91 + 0.82 * \log(GrLivArea) + 2.09 * Neighborhood{:}\, Edwards + 2.58 * Neighborhood{:}\, NorthAmes - 0.30 * \log(GrLivArea) * Neighborhood{:}\, Edwards - 0.35 * \log(GrLivArea) * Neighborhood{:}\, NorthAmes$$

4.2.2 Scatterplot

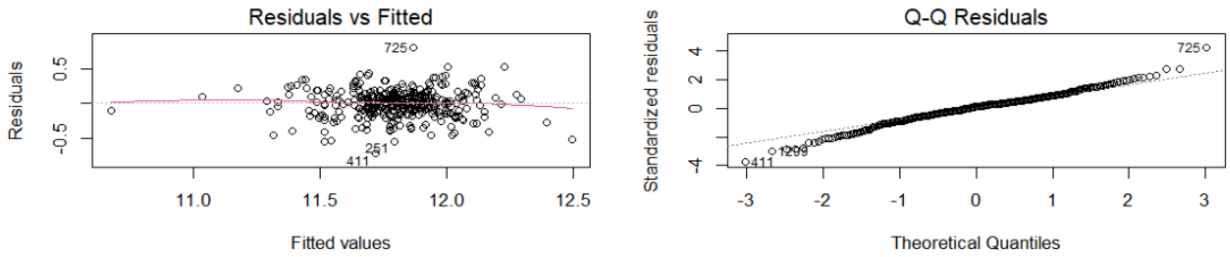


The scatterplot is a good representation of the data points in this analysis and highlights the difference in how square footage relates to sale price in each neighborhood.

4.3 Assumptions and Diagnostics

The log transformation of both the living area square footage (GrLivArea) and the sales price (SalePrice) of the homes best meets the assumptions required for multiple linear regression.

4.3.1 Diagnostic Plots

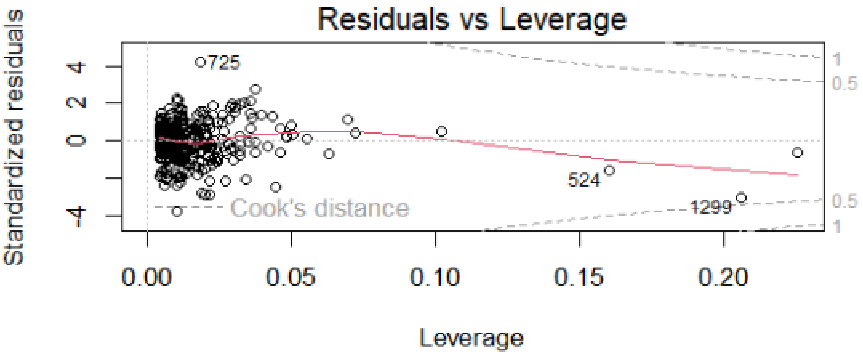


Judging by the Residuals plot, there is not strong evidence against equal standard deviation and linear trend. The Q-Q plot suggests that there is not strong evidence against normality. We are assuming that house sales are independent of each other. Therefore, the assumptions of multiple linear regression are met.

4.4 Influential Points

Although the plots above may raise concern of influential points altering the model, we ultimately decided not to remove any data points.

4.4.1 Cook's D and Leverage Plot



We observed the plot above and added leverage, studentized residuals, and Cook's D values to the observations in our data set. There are no data points with a Cook's D greater than one, so we decided to leave all the observations in our filtered data for the model.

4.5 Competing Models

4.5.1 Interaction Terms

In real estate, the sale price of a home could be related to the neighborhood the home is located. For that reason, it is valid to add interaction terms to a possible model. We completed an extra sum of squares test to determine if adding interaction terms is necessary.

$$Full\ model: \log\left(\hat{S}alePrice\right) = \beta_0 + \beta_1 * \log(GrLivArea) + \beta_2 * Neighborhood{:}\ Edwards + \beta_3 * Neighborhood{:}\ NorthAmes + \beta_4 * \log(GrLivArea) * Neighborhood{:}\ Edwards + \beta_5 * \log\ (GrLivArea) * Neighborhood{:}\ NorthAmes$$

$$Reduced\ model: \log\left(\hat{S}alePrice\right) = \beta_0 + \beta_1 * \log(GrLivArea) + \beta_2 * Neighborhood{:}\ Edwards + \beta_3 * Neighborhood{:}\ NorthAmes$$

The results suggest that there is enough evidence to conclude that at least one of β_4 and β_5 are not equal to zero (p=0.0002, df=2, F-statistic=8.649). This indicates strong evidence that there is an interaction and a model including the interaction term would be more accurate than without.

4.6 Statistical Conclusion and Interpretation

4.6.1 Estimates

Brookside: $\log\ (\hat{S}alePrice) = 5.91 + 0.82 * \log\ (GrLivArea)$

Edwards: $\log\ (\hat{S}alePrice) = 8.00 + 0.52 * \log\ (GrLivArea)$

North Ames: $\log\ (\hat{S}alePrice) = 8.49 + 0.47 * \log\ (GrLivArea)$

4.6.2 95% Confidence Intervals For Model Coefficients

Intercept (4.92, 6.91)
logGrLivArea (0.68, 0.96)
NeighborhoodEdwards (0.82, 3.36)
NeighborhoodNAmes (1.40, 3.76)
logGrLivArea:NeighborhoodEdwards (-0.48,-0.12)
logGrLivArea:NeighborhoodNAmes (-0.51, -0.18)

4.6.3 Conclusion

It is estimated that the doubling of square footage of a home results in a ($2^{0.68}$ = 1.60) 60% increase in the median sales price of a home. A confidence interval for this multiplicative effect is (1.60, 1.95) (p< 2e-16). If a home is in Edwards, the estimated median sale price is ($e^{2.09}$=8.11) is $8,000 more than the median sale price of a home in Brookside a CI for this increase is ($2,270, $28,789) (p=.0013). If a home is in North Ames, it is estimated that the median sale price is ($e^{2.58}$=13.20) is $13,200 more than the median sale price of a home in Brookside (p=2.17e-05). More importantly, for a house in Edwards, it is estimated that doubling the square footage results in ($2^{-.30}$=0.81) 19% decrease than the median sales price of a house in Brookside. A 95% CI for this multiplicative effect is (0.72, 0.92) (p=0.0011). For a house in North Ames, it is estimated that doubling the square footage results in a ($2^{-.35}$=0.78) 22% decrease than the median sales price of a house in Brookside. A 95% CI for this multiplicative effect is (0.70, 0.88) (p=5.35e-05). It is important to note that zero is not included in any of these confidence intervals which reveals that none of the coefficients are equal to zero and each of the variables are associated with sales price. We can make casual conclusions on these results but the results can only be generalized to the houses in these three neighborhoods.

4.7 Client Message

To summarize, our model shows that sales price is associated with square footage and is different for Brookside, North Ames, and Edwards. For further understanding, these relationships can be explored by examining the calculating the estimated sales price for a house of the same size in each of the three different neighborhoods.

An estimated sale price of a 1,000 square foot home in each of the neighborhoods:

Brookside- $105,874

Edwards- $108,012

North Ames- $125,492

For visual aid of these relationships, see the scatterplot in Section 4.2.2. The main takeaways from this analysis would be that an increase in square footage of a home increases the value. The amount of increase is dependent on which neighborhood the home is in.

5 Analysis II

5.1 Problem Statement

We would like to build a predictive multiple regression model that allows us to make predictions on sale prices of homes in Ames, Iowa, for all neighborhoods. For this section, we will be limited to methods that have been covered in the lecture notes for STAT 6301.

5.2 Model Selection

The model we chose was an Elastic Net regression model with a log transformation in square footage with adjusted variables from the original training set based on what variables the method selected.

5.3 Data Cleaning and Candidate Variables Considered

An exhaustive explanation for the data cleaning performed for this analysis is covered in Section 3 – Data Cleaning. A full list of variables that were considered as part of the ElasticNet prediction model is included in Appendix, Section 6.4. A full list of variables removed in the model can be viewed in the Appendix Section 6.5.

5.4 Competing Models

When choosing our models, our goal was to try to minimize the residual sum of squares. After referencing our lecture notes, we decided to build our model based on Penalized Regression Method, since it offered a more robust alternative to the traditional framework. Our competing models that we considered were those built using Lasso, Ridge, and Elastic Net. They used the exact same data set with cleaning steps listed in the Data Cleaning section.

Our RMSE scores for the models are listed below:

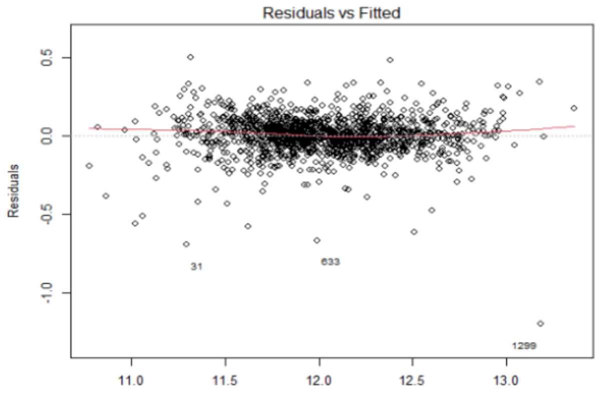
Model Name	RMSE Value
Lasso	.1237262
EN	.1213677
Ridge	.1215444

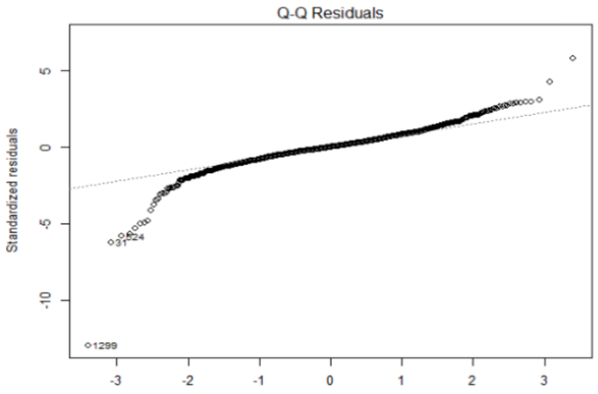
Elastic net gave us the lowest RMSE so we chose this to be our final model.

5.5 Addressing Assumptions

Once we had our model built, we made sure that our model fits the assumptions necessary for multiple linear regression. Similar to Analysis 1, we decided to draw residual plots and check for our assumptions of our model using the training data after removing columns that the ElasticNet model excluded.

5.5.1 Diagnostic Plots



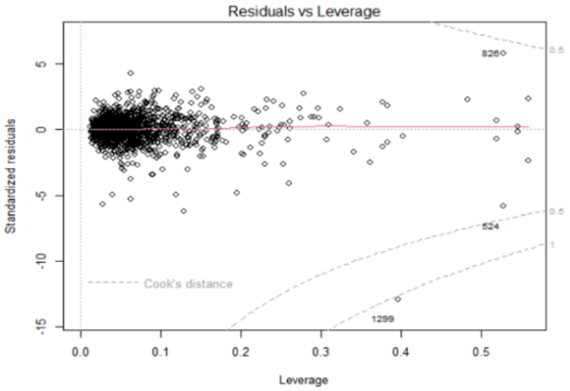


There is some concern of the normality assumption, however, with a large sample size, CLT allows us to assume normality is okay. Judging by the Residuals plot, there is not strong evidence against equal standard deviation and linear trend. We are assuming that house sales are independent of each other. Therefore, the assumptions of multiple linear regression are met for our model.

5.6 Influential Points

Although the plots above may raise concern of influential points altering the model, we ultimately decided not to remove any data points.

5.6.1 Cook's D and Leverage Plot



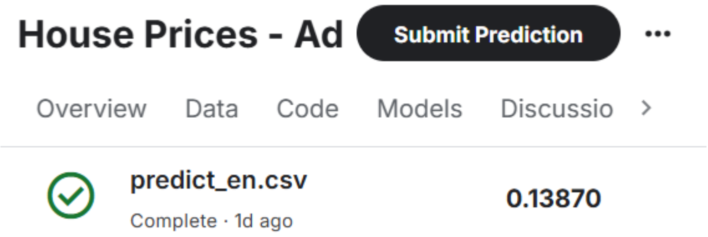
We observed the plot above and noticed the point 1299 is a potential influential point. We ran a model with and without this data point and were more accurate with the point included so we proceeded with caution.

5.7 Conclusion

Our model used a penalized regression model called elastic net that removed variables that are not associated to the sales price of homes. Based on our model, we noticed external aesthetics were less significant to the sales price of the home. Details such as roof material and exterior material were one of the heaviest removed variables. Another result is that the model included every variable that had square footage values. Allowing us to conclude that measures of area are associated with the sale price of a home. Some of these measures include total square footage of the home, lot size, deck area, basement area and more. Another finding was that sales price changes based on the neighborhood a home is in. An additional note is that the year sold was a variable not used in our model indicating that overall market of home sales likely remained constant throughout 2006-2010. Using our model, it is possible to predict the sale price of a house in Ames, Iowa with close accuracy. It is important to note that this data was collected from house sales in 2006-2010 so this model cannot be used to accurately predict sales of homes in other years or regions.

6 Appendix

6.1 Screencap of our Kaggle Score



predict_en.csv

Complete · 1d ago

0.13870

6.2 Analysis 1 Code

library(tidyverse)
library(car)
library(MASS)
library(extrafont)
library(ggplot2)
	Set-up

question1data <- read.csv("C:/Users/19725/Desktop/SMU MASDA/F24/Exp1/Final Project/train.csv")
filtereddata <-
question1data[question1data$Neighborhood %in% c('NAmes', 'Edwards', 'BrkSide'),]
finalq1 <- filtereddata[c('Neighborhood', 'GrLivArea', 'SalePrice')]
	Importing data

finalq1$logGrLivArea <- log(finalq1$GrLivArea)
finalq1$logSalePrice <- log(finalq1$SalePrice)
salesprice.lm_bothlog <- lm(logSalePrice ~ logGrLivArea + Neighborhood + logGrLivArea * Neighborhood, data = finalq1)
par(mfrow=c(2,2))
plot(salesprice.lm_bothlog)
	Fit model of log-log transformation and made diagnostic plots

#Fit a model with all possible interactions
salesprice.lm_bothlog <- lm(logSalePrice ~ logGrLivArea + Neighborhood + logGrLivArea * Neighborhood, data = finalq1)
#Reduced model
salesprice.lm2_bothlog <- lm(logSalePrice ~ logGrLivArea + Neighborhood, data = finalq1)
#Extra sum of squares test
#Must list the reduced model first
anova(salesprice.lm2_bothlog, salesprice.lm_bothlog)
	Extra sum of squares test on interaction terms

summary(salesprice.lm_bothlog)	Print model summary

ggplot(finalq1, aes(x=logGrLivArea, y=logSalePrice, shape=Neighborhood, color=Neighborhood)) +
  geom_point(size=2, shape = 16) +
  geom_smooth(method=lm, aes(fill=Neighborhood)) + # You can remove this line if you don't want regression lines
  ggtitle("Ames Housing Model") +
  theme_bw() +
  xlab("Log(GrLivArea)") +
  ylab("Log(SalePrice)") +
  theme(
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, size = 30), # Increase plot title size
    axis.title = element_text(size = 18), # Increase axis title size
    axis.text = element_text(size = 14), # Increase axis tick label size
	Model scatterplot

legend.title = element_text(size = 16), # Increase legend title size
legend.text = element_text(size = 14) # Increase legend text size
) +
labs(color = 'Neighborhood', shape = 'Neighborhood', fill = 'Neighborhood')	

```
{r}
confint(salesprice.lm_bothlog)
```
	Confidence interval

6.3 Analysis 2 Code

library(tidyverse)
library(dplyr)
library(glmnet)
library(mice)
library(caret)
library(leaps)
library(bestglm)
library(MASS)
	Loading our required libraries

comb <- read.csv("combined_dataset_q2.csv")
test <- read.csv("cleaned_test_data_q2.csv")
train <- read.csv("cleaned_train_data_q2.csv")
	Loading Data into R

test.index <- test$Id
train.index <- train$Id
	Creating index vectors to have Ids for testing and training data, for reference in the combined data.

predictiondata <- comb[, !(names(comb) %in% c("SalePrice","GrLivArea"))]
predictiondata$yearbuiltsub1950 <- ifelse(predictiondata$YearBuilt < 1950, 1, 0)
predictiondata$yearbuiltgreater75 <- ifelse(predictiondata$YearBuilt > 1975, 1, 0)
predictiondata$yearremodsub75 <- ifelse(predictiondata$YearRemodAdd < 1975, 1, 0)
predictiondata$yearremodgreater95 <- ifelse(predictiondata$YearRemodAdd > 1995, 1, 0)
predictiondata$monthsoldjantojune <- ifelse(predictiondata$MoSold > 7, 1, 0)
predictiondata$yearsold06 <- ifelse(predictiondata$YrSold == 2006, 1, 0)
predictiondata$yearsold07 <- ifelse(predictiondata$YrSold == 2007, 1, 0)
predictiondata$yearsold08 <- ifelse(predictiondata$YrSold == 2008, 1, 0)
predictiondata$yearsold09 <- ifelse(predictiondata$YrSold == 2009, 1, 0)
predictiondata$overallqualsub5 <- ifelse(predictiondata$OverallQual < 5, 1, 0)
predictiondata$overallqualabv7 <- ifelse(predictiondata$OverallQual > 7, 1, 0)
predictiondata$overallcondsub5 <- ifelse(predictiondata$OverallCond < 5, 1, 0)
predictiondata <- predictiondata %>% mutate(MSSubclassgroups = case_when(
MSSubClass==20 ~ "1-STORY 1946 & NEWER ALL STYLES",
	Here, we are adding some numerical versions of the factor variables, since glmnet required a matrix input. Any variables we decided to use, after accounting for ones with NA/Empty values were kept.

MSSubClass == 30 ~ "1-STORY 1945 & OLDER", MSSubClass == 40 ~ "1-STORY W/FINISHED ATTIC ALL AGES", MSSubClass == 45 ~ "1-1/2 STORY - UNFINISHED ALL AGES", MSSubClass == 50 ~ "1-1/2 STORY FINISHED ALL AGES", MSSubClass == 60 ~ "2-STORY 1946 & NEWER", MSSubClass == 70 ~ "2-STORY 1945 & OLDER", MSSubClass == 75 ~ "2-1/2 STORY ALL AGES", MSSubClass == 80 ~ "SPLIT OR MULTI-LEVEL", MSSubClass == 85 ~ "SPLIT FOYER", MSSubClass == 90 ~ "DUPLEX - ALL STYLES AND AGES", MSSubClass == 120 ~ "1-STORY PUD (Planned Unit Development) - 1946 & NEWER", MSSubClass == 150 ~ "1-1/2 STORY PUD - ALL AGES", MSSubClass == 160 ~ "2-STORY PUD - 1946 & NEWER", MSSubClass == 180 ~ "PUD - MULTILEVEL - INCL SPLIT LEV/FOYER", MSSubClass == 190 ~ "2 FAMILY CONVERSION - ALL STYLES AND AGES"))	

predictiondata_dummycoded1 <- predictiondata[, !(names(predictiondata) %in% c('OverallQual','OverallCond','YearBuilt','YearRemodAdd','MoSold','YrSold','MSSubClass'))]
alldata_dummycoded <- model.matrix(~., data=predictiondata_dummycoded1)[,-1]
dfalldata_dummycoded <- as.data.frame(alldata_dummycoded)
	These are variables that were removed, after numerical version of them were created.

train_subset_2 <- subset(dfalldata_dummycoded, Id %in% train.index)
test_subset_2 <- subset(dfalldata_dummycoded, Id %in% test.index)
	Splitting our combined dataset back to their respective testing and train versions.

set.seed(123)
predictiondata.glmnet2 <- cv.glmnet(as.matrix(train_subset_2[,3:209]), train_subset_2[,1], alpha=1)
##Optimal tuning parameter
best.lambda2 <- predictiondata.glmnet2$lambda.min
##Check parameter estimates for the optimal model
coef(predictiondata.glmnet2, s=best.lambda2)
	Creating prediction training data using LASSO method.

set.seed(123)
predictionridge.glmnet <- cv.glmnet(as.matrix(train_subset_2[,3:209]), train_subset_2[,1], alpha=0)
attributes(predictionridge.glmnet)
##Optimal tuning parameter
best.lambda <- predictionridge.glmnet$lambda.min
##Check parameter estimates for the optimal model
coef(predictionridge.glmnet, s=best.lambda)
	Creating prediction training data using Ridge method.

tcontrol <- trainControl(method="repeatedcv", number=10, repeats=5)
set.seed(123)
predictiondataenet.glmnet3 <- train(as.matrix(train_subset_2[,3:209]), train_subset_2[,1], trControl=tcontrol, method="glmnet", tuneLength=10)
attributes(predictiondataenet.glmnet3)
predictiondataenet.glmnet3$results
predictiondataenet.glmnet3$bestTune
predictiondataenet.glmnet4 <- predictiondataenet.glmnet3$finalModel
	Creating prediction training data using Elastic Net method.

coef(predictiondataenet.glmnet4, s=predictiondataenet.glmnet3$bestTune$lambda)	

ridge.pred<- predict(predictionridge.glmnet, as.matrix(train_subset_2[,3:209]), s=best.lambda)
ridge.rmse <- sqrt(mean((ridge.pred - train_subset_2[,1])^2))
lasso.pred <- predict(predictiondata.glmnet2, as.matrix(train_subset_2[,3:209]), s=best.lambda2)
lasso.rmse <- sqrt(mean((lasso.pred - train_subset_2[,1])^2))
en.pred <- predict(predictiondataenet.glmnet3,as.matrix(train_subset_2[,3:209]), s=predictiondataenet.glmnet3$bestTune$lambda)
en.rmse <- sqrt(mean((en.pred - train_subset_2[,1])^2))
data.frame(Method = c("Ridge", "Lasso", "EN"), RMSE = c(ridge.rmse, lasso.rmse, en.rmse))
	Running code to get RMSE scores for all 3 methods.

salesprice_predict.en <- predict(predictiondataenet.glmnet3, test_subset_2)
salesprice_predict.ridge <- predict(predictionridge.glmnet, newx=as.matrix(test_subset_2[,3:209]),s=best.lambda)
salesprice_predict.lasso <- predict(predictiondata.glmnet2, newx=as.matrix(test_subset_2[,3:209]), s=best.lambda2)
	Generating the predicted values for the test data.

salesprice_predict.en <- exp(salesprice_predict.en)
salesprice_predict.ridge <- exp(salesprice_predict.ridge)
salesprice_predict.lasso <- exp(salesprice_predict.lasso)
	Change from log(SalePrice) to actual SalePrice. ("Unlog" the data).

write.csv(salesprice_predict.en,"predict_elasticnet.csv")
write.csv(salesprice_predict.lasso,"predict_lasso.csv")
write.csv(salesprice_predict.ridge,"predict_ridge.csv")
	Exporting these to CSV.
	Please note column names have to be manually added to CSV.

6.4 Columns Included for Analysis 2 from training dataset

BldgTypeTwnhs, BsmtFinSF1, BsmtFinSF2, BsmtFullBath, BsmtHalfBath, CentralAirY, Condition1Feedr, Condition1Norm, Condition1PosA, Condition1RRAe, Condition2PosA, Condition2PosN, ExterCondFa, ExterCondPo, Exterior1stBrkComm, Exterior1stBrkFace, Exterior1stWd Sdng, Exterior2ndCmentBd, Exterior2ndImStucc, Exterior2ndStucco, Exterior2ndWd Shng, ExterQualFa, ExterQualTA, Fireplaces, FoundationPConc, FoundationSlab, FoundationStone, FoundationWood, FullBath, FunctionalMaj2, FunctionalSev, FunctionalTyp, GarageArea, GarageCars, HalfBath, HeatingGasW, HeatingGrav, HeatingOthW, HeatingQCFa, HeatingQCGd, HeatingQCTA, HouseStyle1.5Unf, HouseStyleSFoyer, HouseStyleSLvl, KitchenAbvGr, KitchenQualFa, KitchenQualGd, KitchenQualTA, LandContourHLS, log.GrvLivArea., LotArea, LotConfigCulDSac, LotConfigInside, LotShapeIR2, LotShapeIR3, LotShapeReg, LowQualFinSF, monthsoldjantojune, MSSubclassgroups1-1/2 STORY FINISHED ALL AGES, MSSubclassgroups1-STORY 1945 & OLDER, MSSubclassgroups1-STORY 1946 & NEWER ALL STYLES, MSSubclassgroups2-STORY PUD - 1946 & NEWER, MSSubclassgroupsSPLIT OR MULTI-LEVEL, MSZoningFV, MSZoningRL, NeighborhoodClearCr, NeighborhoodCollgCr, NeighborhoodCrawfor, NeighborhoodEdwards, NeighborhoodIDOTRR, NeighborhoodMeadowV, NeighborhoodMitchel, NeighborhoodNoRidge, NeighborhoodNridgHt, NeighborhoodOldTown, NeighborhoodSawyer, NeighborhoodSomerst, NeighborhoodStoneBr, NeighborhoodSWISU, NeighborhoodVeenker, OpenPorchSF, overallcondsub5, overallqualabv7, overallqualsub5, PavedDriveY, RoofMatlMembran, RoofMatlWdShngl, RoofStyleGable, SaleConditionNormal, SaleTypeCon, SaleTypeCWD, SaleTypeNew, SaleTypeOth, ScreenPorch, StreetPave, TotalBsmtSF, WoodDeckSF, X3SsnPorch, yearbuiltsub1950, yearremodgreater95, yearremodsub75, yearsold09

6.5 Removed Columns for Analysis 2 from training dataset

MSZoningRH, MSZoningRM, LandContourLow, LandContourLvl, LotConfigFR2, LotConfigFR3, LandSlopeMod, LandSlopeSev, NeighborhoodBlueste, NeighborhoodBrDale, NeighborhoodBrkSide, NeighborhoodGilbert, NeighborhoodNAmes, NeighborhoodNPkVill, NeighborhoodNWAmes, NeighborhoodSawyerW, NeighborhoodTimber, Condition1PosN, Condition1RRAn, Condition1RRNe, Condition1RRNn, Condition2Feedr, Condition2Norm, Condition2RRAe, Condition2RRAn. Condition2RRNn, BldgType2fmCon, BldgTypeDuplex, BldgTypeTwnhsE, HouseStyle1Story, HouseStyle2.5Fin, HouseStyle2.5Unf, HouseStyle2Story, RoofStyleGambrel, RoofStyleHip, RoofStyleMansard, RoofStyleShed, RoofMatlCompShg, RoofMatlMetal, RoofMatlRoll, RoofMatlTar&Grv, RoofMatlWdShake, Exterior1stAsphShn, Exterior1stCBlock, Exterior1stCemntBd, Exterior1stHdBoard, Exterior1stImStucc, Exterior1stMetalSd, Exterior1stPlywood, Exterior1stStone, Exterior1stStucco, Exterior1stVinylSd, Exterior1stWdShing, Exterior2ndAsphShn, Exterior2ndBrk Cmn, Exterior2ndBrkFace, Exterior2ndCBlock, Exterior2ndHdBoard, Exterior2ndMetalSd, Exterior2ndOther, Exterior2ndPlywood, Exterior2ndStone, Exterior2ndVinylSd, Exterior2ndWd Sdng, ExterQualGd, ExterCondGd, ExterCondTA, FoundationCBlock, BsmtUnfSF, HeatingGasA, HeatingWall, HeatingQCPo, X1stFlrSF, X2ndFlrSF, BedroomAbvGr, TotRmsAbvGrd, FunctionalMin1, FunctionalMin2, FunctionalMod, PavedDriveP, EnclosedPorch, PoolArea, MiscVal, SaleTypeConLD, SaleTypeConLI, SaleTypeConLw, SaleTypeWD, SaleConditionAdjLand, SaleConditionAlloca, SaleConditionFamily, SaleConditionPartial, yearbuiltgreater75, yearsold06, yearsold07, yearsold08, MSSubclassgroups1-1/2 STORY, PUD - ALL AGES, MSSubclassgroups1-STORY PUD (Planned Unit Development) - 1946 & NEWER, MSSubclassgroups1-STORY W/FINISHED ATTIC ALL AGES, MSSubclassgroups2, FAMILY CONVERSION - ALL STYLES AND AGES, MSSubclassgroups2-1/2 STORY ALL AGES, MSSubclassgroups2-STORY 1945 & OLDER, MSSubclassgroups2-STORY 1946 & NEWER, MSSubclassgroupsDUPLEX - ALL STYLES AND AGES, MSSubclassgroupsPUD - MULTILEVEL - INCL SPLIT LEV/FOYER, MSSubclassgroupsSPLIT FOYER