Ames Iowa Housing Prices (Analysis Two)

Paul Adams

August 3, 2019

library(pacman)

## Warning: package 'pacman' was built under R version 3.6.1

p\_load(tidyr,dplyr,purrr,stringr,ggplot2, MASS,DAAG, sjPlot, Hmisc)

### Handle NAs

#See structure for replacing NAs  
str(df)

## 'data.frame': 1460 obs. of 81 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ MSSubClass : int 60 20 60 70 60 50 20 60 50 190 ...  
## $ MSZoning : chr "RL" "RL" "RL" "RL" ...  
## $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  
## $ Street : chr "Pave" "Pave" "Pave" "Pave" ...  
## $ Alley : chr NA NA NA NA ...  
## $ LotShape : chr "Reg" "Reg" "IR1" "IR1" ...  
## $ LandContour : chr "Lvl" "Lvl" "Lvl" "Lvl" ...  
## $ Utilities : chr "AllPub" "AllPub" "AllPub" "AllPub" ...  
## $ LotConfig : chr "Inside" "FR2" "Inside" "Corner" ...  
## $ LandSlope : chr "Gtl" "Gtl" "Gtl" "Gtl" ...  
## $ Neighborhood : chr "CollgCr" "Veenker" "CollgCr" "Crawfor" ...  
## $ Condition1 : chr "Norm" "Feedr" "Norm" "Norm" ...  
## $ Condition2 : chr "Norm" "Norm" "Norm" "Norm" ...  
## $ BldgType : chr "1Fam" "1Fam" "1Fam" "1Fam" ...  
## $ HouseStyle : chr "2Story" "1Story" "2Story" "2Story" ...  
## $ OverallQual : int 7 6 7 7 8 5 8 7 7 5 ...  
## $ OverallCond : int 5 8 5 5 5 5 5 6 5 6 ...  
## $ YearBuilt : int 2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...  
## $ YearRemodAdd : int 2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...  
## $ RoofStyle : chr "Gable" "Gable" "Gable" "Gable" ...  
## $ RoofMatl : chr "CompShg" "CompShg" "CompShg" "CompShg" ...  
## $ Exterior1st : chr "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...  
## $ Exterior2nd : chr "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...  
## $ MasVnrType : chr "BrkFace" "None" "BrkFace" "None" ...  
## $ MasVnrArea : int 196 0 162 0 350 0 186 240 0 0 ...  
## $ ExterQual : chr "Gd" "TA" "Gd" "TA" ...  
## $ ExterCond : chr "TA" "TA" "TA" "TA" ...  
## $ Foundation : chr "PConc" "CBlock" "PConc" "BrkTil" ...  
## $ BsmtQual : chr "Gd" "Gd" "Gd" "TA" ...  
## $ BsmtCond : chr "TA" "TA" "TA" "Gd" ...  
## $ BsmtExposure : chr "No" "Gd" "Mn" "No" ...  
## $ BsmtFinType1 : chr "GLQ" "ALQ" "GLQ" "ALQ" ...  
## $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...  
## $ BsmtFinType2 : chr "Unf" "Unf" "Unf" "Unf" ...  
## $ BsmtFinSF2 : int 0 0 0 0 0 0 0 32 0 0 ...  
## $ BsmtUnfSF : int 150 284 434 540 490 64 317 216 952 140 ...  
## $ TotalBsmtSF : int 856 1262 920 756 1145 796 1686 1107 952 991 ...  
## $ Heating : chr "GasA" "GasA" "GasA" "GasA" ...  
## $ HeatingQC : chr "Ex" "Ex" "Ex" "Gd" ...  
## $ CentralAir : chr "Y" "Y" "Y" "Y" ...  
## $ Electrical : chr "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...  
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...  
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...  
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...  
## $ BsmtFullBath : int 1 0 1 1 1 1 1 1 0 1 ...  
## $ BsmtHalfBath : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...  
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...  
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...  
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...  
## $ KitchenQual : chr "Gd" "TA" "Gd" "Gd" ...  
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  
## $ Functional : chr "Typ" "Typ" "Typ" "Typ" ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ FireplaceQu : chr NA "TA" "TA" "Gd" ...  
## $ GarageType : chr "Attchd" "Attchd" "Attchd" "Detchd" ...  
## $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...  
## $ GarageFinish : chr "RFn" "RFn" "RFn" "Unf" ...  
## $ GarageCars : int 2 2 2 3 3 2 2 2 2 1 ...  
## $ GarageArea : int 548 460 608 642 836 480 636 484 468 205 ...  
## $ GarageQual : chr "TA" "TA" "TA" "TA" ...  
## $ GarageCond : chr "TA" "TA" "TA" "TA" ...  
## $ PavedDrive : chr "Y" "Y" "Y" "Y" ...  
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...  
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...  
## $ EnclosedPorch: int 0 0 0 272 0 0 0 228 205 0 ...  
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...  
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolQC : chr NA NA NA NA ...  
## $ Fence : chr NA NA NA NA ...  
## $ MiscFeature : chr NA NA NA NA ...  
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...  
## $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...  
## $ YrSold : int 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...  
## $ SaleType : chr "WD" "WD" "WD" "WD" ...  
## $ SaleCondition: chr "Normal" "Normal" "Normal" "Abnorml" ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

# UDF for flattening and ordering the correlation matrix  
flattenCorrMatrix <- function(cormat, pmat) {  
 ut <- upper.tri(cormat)  
 data.frame(  
 row = rownames(cormat)[row(cormat)[ut]],  
 column = rownames(cormat)[col(cormat)[ut]],  
 cor =(cormat)[ut],  
 p = pmat[ut]  
 )  
}  
  
# Snapshot of all NAs  
na\_count <- sapply(df, function(cnt) sum(length(which(is.na(cnt)))))  
na\_count

## Id MSSubClass MSZoning LotFrontage LotArea   
## 0 0 0 259 0   
## Street Alley LotShape LandContour Utilities   
## 0 1369 0 0 0   
## LotConfig LandSlope Neighborhood Condition1 Condition2   
## 0 0 0 0 0   
## BldgType HouseStyle OverallQual OverallCond YearBuilt   
## 0 0 0 0 0   
## YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
## 0 0 0 0 0   
## MasVnrType MasVnrArea ExterQual ExterCond Foundation   
## 8 8 0 0 0   
## BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1   
## 37 37 38 37 0   
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating   
## 38 0 0 0 0   
## HeatingQC CentralAir Electrical X1stFlrSF X2ndFlrSF   
## 0 0 1 0 0   
## LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath   
## 0 0 0 0 0   
## HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd   
## 0 0 0 0 0   
## Functional Fireplaces FireplaceQu GarageType GarageYrBlt   
## 0 0 690 81 81   
## GarageFinish GarageCars GarageArea GarageQual GarageCond   
## 81 0 0 81 81   
## PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch   
## 0 0 0 0 0   
## ScreenPorch PoolArea PoolQC Fence MiscFeature   
## 0 0 1453 1179 1406   
## MiscVal MoSold YrSold SaleType SaleCondition   
## 0 0 0 0 0   
## SalePrice   
## 0

########################################################################################################  
#Start Area for handling NAs############################################################################  
########################################################################################################  
  
##############################################################################  
#Start Checking Electrical NAs################################################  
##############################################################################  
  
# Which neighborhood is the NA for electrical in?  
noElectric <- df[which(is.na(df$Electrical)),c("SalePrice","Neighborhood")]  
noElectric

## SalePrice Neighborhood  
## 1380 167500 Timber

# Who services electricity in this neighborhood?  
timberElectric <- arrange(df[which(df$Neighborhood=="Timber"),c("SalePrice","Electrical")], SalePrice)  
head(timberElectric)

## SalePrice Electrical  
## 1 137500 FuseA  
## 2 160000 SBrkr  
## 3 167500 <NA>  
## 4 170000 SBrkr  
## 5 175000 SBrkr  
## 6 175000 SBrkr

tail(timberElectric)

## SalePrice Electrical  
## 33 315000 SBrkr  
## 34 315500 SBrkr  
## 35 335000 SBrkr  
## 36 369900 SBrkr  
## 37 375000 SBrkr  
## 38 378500 SBrkr

# Timber looks to mostly use SBrkr so this should not be an issue to assume for the NA value for Electrical - it also fits the SalePrice range of SBrkr  
  
##############################################################################  
#End Checking Electrical NAs##################################################  
##############################################################################  
  
df$LotFrontage[is.na(df$LotFrontage)] <- 0  
df$Alley[is.na(df$Alley)] <- "None"  
df$MasVnrType[is.na(df$MasVnrType)] <- "None"  
df$MasVnrArea[is.na(df$MasVnrArea)] <- 0  
df$BsmtQual[is.na(df$BsmtQual)] <- 0  
df$BsmtCond[is.na(df$BsmtCond)] <- 0  
df$BsmtExposure[is.na(df$BsmtExposure)] <- 0  
df$BsmtFinType1[is.na(df$BsmtFinType1)] <- 0  
df$BsmtFinType2[is.na(df$BsmtFinType2)] <- 0  
df$Electrical[is.na(df$Electrical)] <- "SBrkr"  
df$FireplaceQu[is.na(df$FireplaceQu)] <- "None"  
df$GarageType[is.na(df$GarageType)] <- "None"  
df$GarageYrBlt[is.na(df$GarageYrBlt)] <- mean(df$GarageYrBlt)  
df$GarageFinish[is.na(df$GarageFinish)] <- "None"  
df$GarageQual[is.na(df$GarageQual)] <- "None"  
df$GarageCond[is.na(df$GarageCond)] <- "None"  
df$PoolQC[is.na(df$PoolQC)] <- "None"  
df$Fence[is.na(df$Fence)] <- "None"  
df$MiscFeature[is.na(df$MiscFeature)] <- "None"  
  
########################################################################################################  
#End Area for handling NAs##############################################################################  
########################################################################################################  
  
# create dataframe for numeric data  
dfTest.numeric <- dplyr::select\_if(df, is.numeric) %>% data.frame()  
  
# create dataframe for non-numeric data  
dfTest.nonnumeric <- dplyr::select\_if(df, is.factor) %>% data.frame()

### Correlation matrix for quantitative data

#See what variables are correlated with eachother, p-values  
correlation.matrix <- rcorr(as.matrix(dfTest.numeric))  
corDF <- data.frame(flattenCorrMatrix(correlation.matrix$r, correlation.matrix$P))

### Testing Quantitative Data Model

#################################################################################  
###########Start testing quantitative data model variable selection##############  
#################################################################################  
#Order the correlation matrix to show the highest correlated  
data.frame(corDF[order(-corDF$cor),])  
quantDataModel <- corDF[which(corDF$cor >= 0.5),]  
  
fitQuant <- lm(log(SalePrice) ~ OverallQual + YearBuilt + YearRemodAdd + TotalBsmtSF + GrLivArea + X1stFlrSF + FullBath + TotRmsAbvGrd + GarageCars + GarageArea + Neighborhood, data = df)  
summary(fitQuant)  
  
# Model using all variables  
fitFull <- lm(log(SalePrice) ~ Id + MSSubClass + MSZoning + LotFrontage + LotArea + Street + Alley + LotShape + LandContour + Utilities + LotConfig + LandSlope + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl + Exterior1st + Exterior2nd + MasVnrType + MasVnrArea + ExterQual + ExterCond + Foundation + BsmtQual + BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtFinType2 + BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + Heating + HeatingQC + CentralAir + Electrical + X1stFlrSF + X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional + Fireplaces + FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars + GarageArea + GarageQual + GarageCond + PavedDrive + WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + PoolQC + Fence + MiscFeature + MiscVal + MoSold + YrSold + SaleType + SaleCondition, data = df)  
  
# Forward Selection  
stepForward.Quant.model <- stepAIC(fitQuant, direction = "forward", trace = F)  
summary(stepForward.Quant.model)  
stepForward.model$anova  
  
# Backward Selection  
stepBackward.Quant.model <- stepAIC(fitQuant, direction = "backward", trace = F)  
summary(stepBackward.Quant.model)  
stepBackward.model$anova  
  
# Stepwise Selection  
stepwise.Quant.model <- stepAIC(fitQuant, direction = "both", trace = F)  
summary(stepwise.Quant.model)  
stepBoth.model$anova  
# Forward and Stepwise selection both indicate to remove X1stFlrSF and FullBath  
  
#################################################################################  
###############End testing quantitative data model variable selection############  
#################################################################################

### Testing Qualitative Data Model

#################################################################################  
#############Start testing qualitative data model variable selection#############  
#################################################################################  
fitQual <- lm(log(SalePrice) ~ MSZoning + Street + Alley + LotShape + LandContour + Utilities + LotConfig + LandSlope + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + RoofStyle + RoofMatl + Exterior1st + Exterior2nd + MasVnrType + ExterQual + ExterCond + Foundation + BsmtQual + BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinType2 + Heating + HeatingQC + CentralAir + Electrical + KitchenQual + FireplaceQu + GarageType + GarageFinish + GarageQual + GarageCond + PavedDrive + PoolQC + Fence+ MiscFeature + SaleType + SaleCondition, data=df)  
summary(fitQual)  
  
fitQual\_Filtered <- lm(log(SalePrice) ~ Neighborhood + LotConfig + Condition1 + BldgType + HouseStyle + RoofMatl + Heating + KitchenQual + GarageQual + GarageCond + PoolQC, data=df)  
  
fitQual\_double.Filtered <- lm(log(SalePrice) ~ Neighborhood + LotConfig + BldgType + HouseStyle + RoofMatl + KitchenQual + PoolQC, data=df) ##LotConfig mostly only matters when it's a Culdesac  
  
# Forward Selection  
stepForward.Qual.model <- stepAIC(fitQual\_double.Filtered, direction = "forward", trace = F)  
summary(stepForward.Qual.model)  
stepForward.Qual.model$anova  
  
# Backward Selection  
stepBackward.Qual.model <- stepAIC(fitQual\_double.Filtered, direction = "backward", trace = F)  
summary(stepBackward.Qual.model)  
stepBackward.Qual.model$anova  
  
# Stepwise Selection  
stepwise.Qual.model <- stepAIC(fitQual\_double.Filtered, direction = "both", trace = F)  
summary(stepBoth.model)  
stepBoth.model$anova  
  
#################################################################################  
##############End testing qualitative data model variable selection##############  
#################################################################################

### Start testing model with qual and quant data

#################################################################################  
##############Start testing quant+qual data model variable selection#############  
#################################################################################  
fit\_model.Filtered <- lm(log(SalePrice) ~ OverallQual + YearBuilt + YearRemodAdd + TotalBsmtSF + GrLivArea + TotRmsAbvGrd + GarageCars + GarageArea + Neighborhood + LotConfig + BldgType + HouseStyle + RoofMatl + KitchenQual + PoolQC, data = df)  
  
# Forward Selection  
stepForward.model <- stepAIC(fit\_model.Filtered, direction = "forward", trace = F)  
summary(stepForward.model)  
stepForward.model$anova  
  
# Backward Selection  
stepBackward.model <- stepAIC(fit\_model.Filtered, direction = "backward", trace = F)  
summary(stepBackward.model)  
stepBackward.model$anova  
  
# Stepwise Selection  
stepwise.model <- stepAIC(fit\_model.Filtered, direction = "both", trace = F)  
summary(stepwise.model)  
stepwise.model$anova  
  
#Both backward and stepwise selection indicate to remove HouseStyle, TotRmsAbvGrd, X1stFlrSF, and FullBath  
  
#################################################################################  
################End testing quant+qual data model variable selection#############  
#################################################################################