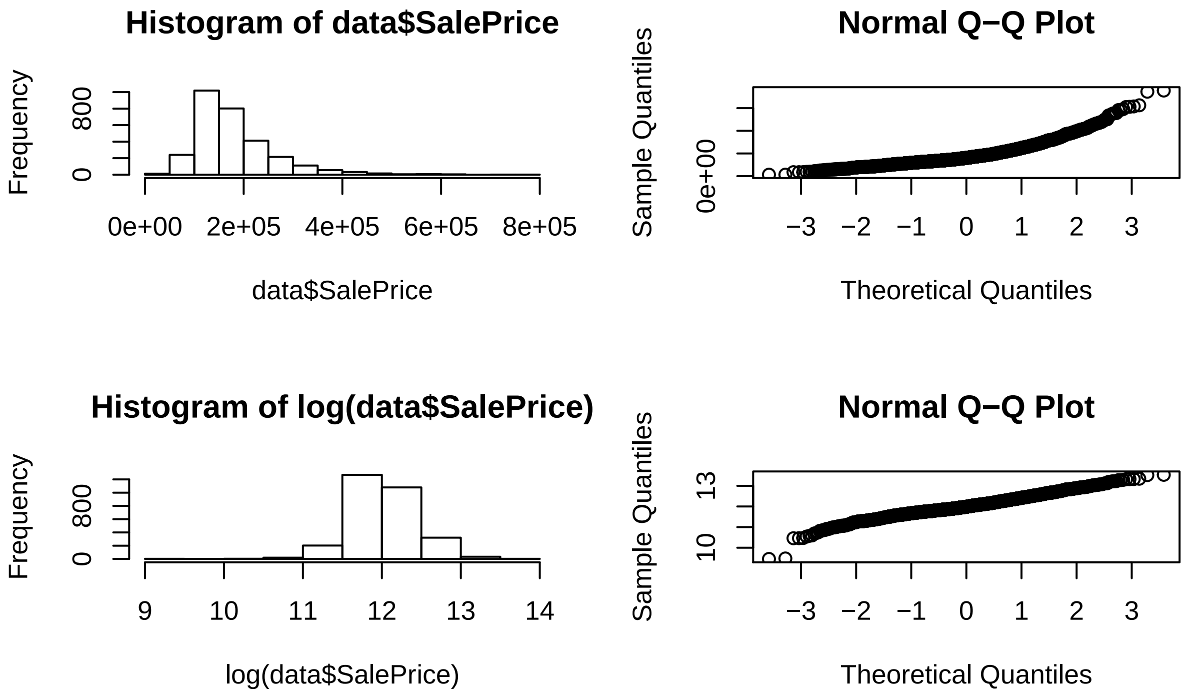
Ames Housing EDA and Cleaning

Joel Bracken

The goal of this analysis is to predict the sale price of a piece of property in Ames, Iowa. First, we examine the distribution of both Sale Price and the log of Sale Price.

```
setwd("~/PracProj/Ames Housing Data/Ames-Housing-Data") #set working directory
data <- read.csv("AmesHousing.csv",row.names = 1)
par(mfrow=c(2,2))
hist(data$SalePrice)
qqnorm(data$SalePrice)
hist(log(data$SalePrice))
qqnorm(log(data$SalePrice))
```



Each potential predictor was analyzed by itself (this was excluded due to length). Variables in the 'remove' group were removed for various reasons including a high percentage of NA or low variability between levels. For example, the utilities variable contained 2927 lots labeled 'AllPub' , 1 labeled 'NoseWa', and 2 labeled 'NoSewr'. Variables in the 'convert' group were used to make new variables and then removed afterwards. For example, the second floor square footage variable was changed into a yes/no variable to indicate if the house had a second floor or not.

```
remove <- c('PID','MS.Zoning', 'MS.SubClass', "Street", "Alley", "Land.Contour","Utilities" , "Land.Slop
            "Condition.1", "Condition.2", "Bldg.type", "Roof.Style", "Roof.Matl", "Bsmt.Cond",
            "BsmtFin.Type.2", "BsmtFin.SF.1", "BsmtFin.SF.2", "Heating", "CentralAir", "Electrical",
            "Low.Qual.Fin.SF", "X1st.Flr.SF", "Kitchen.AbvGr", "Garage.Qual",
            "Garage.Cond", "Pool.QC", "Misc.Feature", "Pool.Area","Fireplace.Qu", "Neighborhood",
            "Exterior.1st", "Exterior.2nd", 'Lot.Frontage')
```

```r
convert <- c("Mas.Vnr.Type", "Bsmt.Full.Bath", "Bsmt.Half.Bath","Mas.Vnr.Area", "Wood.Deck.SF",
             "Full.Bath", "Half.Bath", "Open.Porch.SF", "X2nd.Flr.SF", "Fireplaces",
             "Enclosed.Porch", "X3Ssn.Porch", "Screen.Porch", "Fence", "Mo.Sold", 'Garage.Type',
             'Garage.Yr.Blt', 'Garage.Finish')

# Function to change zeros to No and others to Yes
num.to.bin.cat <- function(list){
  no <- which(list %in% 0)
  yes <- which(!(list %in% 0))
  list[no] <- "No"
  list[yes] <- "Yes"
  return(as.factor(list))
}
# Change continuous variables to categrical
data$Mas.Vnr <- num.to.bin.cat(data$Mas.Vnr.Area)
data$Two.Floors <- num.to.bin.cat(data$X2nd.Flr.SF)
data$Fireplace <- num.to.bin.cat(data$Fireplaces)
data$Deck <- num.to.bin.cat(data$Wood.Deck.SF)

data$Porch <- data$Open.Porch.SF + data$Enclosed.Porch + data$X3Ssn.Porch + data$Screen.Porch
data$Porch <- num.to.bin.cat(data$Porch)

# Convert categorical variables down to two level
data$Fence <- as.vector(data$Fence)
data$Fence[which(!is.na(data$Fence))] <- "Yes"
data$Fence[which(is.na(data$Fence))] <- "No"
data$Fence <- as.factor(data$Fence)

# Represent bathrooms by single number
data$Bath <- (data$Bsmt.Full.Bath + data$Bsmt.Half.Bath/2)+ (data$Full.Bath + data$Half.Bath/2)

# Garage or not
data$Garage[which(!is.na(data$Garage.Yr.Blt))] <- "Yes"
data$Garage[which(is.na(data$Garage.Yr.Blt))] <- "No"

# month to season
data$Mo.Sold <- data$Mo.Sold +1
data$Mo.Sold[which(data$Mo.Sold %in% 13)] = 1
data$Season <-  ceiling( data$Mo.Sold/3)
seasons <- c("Winter", "Spring", "Summer", "Fall")
seas_int <- c(1,2,3,4)

winter <- c(12,1,2)
spring <- c(3,4,5)
summer <- c(6.7,8)
fall <- c(9,10,11)

for (i in 1:length(seas_int)) {
 row <- which( data$Season %in%  seas_int[i])
 data$Season[row] = seasons[i]
}

#Remove variables
```

```
data.clean <- data[,-c( which(names(data) %in% remove),which(names(data) %in% convert))]
```

The top 5 largest houses were removed from the data set since they were high outliers. Eighty-one observations were removed due to missing data.

```
tail(sort(data.clean$Gr.Liv.Area), n = 50) # exclude houses over 4,000 sq ft as suggested in handout
```

```
##  [1] 2787 2787 2787 2790 2792 2794 2795 2798 2799 2810 2814 2822 2826 2828
## [15] 2840 2855 2868 2872 2872 2898 2944 2945 2956 2978 3005 3078 3082 3086
## [29] 3086 3112 3140 3194 3222 3228 3238 3279 3390 3395 3447 3493 3500 3608
## [43] 3627 3672 3820 4316 4476 4676 5095 5642
```

```
data.clean <- data.clean[ which(data.clean$Gr.Liv.Area < 4000), ]    #remove houses over 4,000 sq ft

na_count <-sapply(data.clean, function(y) sum(length(which(is.na(y)))))
na_count
```

```
##       Lot.Area      Lot.Shape     Lot.Config      Bldg.Type    House.Style
##              0              0              0              0              0
##   Overall.Qual   Overall.Cond     Year.Built Year.Remod.Add     Exter.Qual
##              0              0              0              0              0
##     Exter.Cond     Foundation      Bsmt.Qual  Bsmt.Exposure BsmtFin.Type.1
##              0              0             79             79             79
##    Bsmt.Unf.SF  Total.Bsmt.SF     Heating.QC    Central.Air    Gr.Liv.Area
##              1              1              0              0              0
##   Bedroom.AbvGr   Kitchen.Qual  TotRms.AbvGrd     Functional    Garage.Cars
##              0              0              0              0              1
##    Garage.Area    Paved.Drive       Misc.Val        Yr.Sold      Sale.Type
##              1              0              0              0              0
## Sale.Condition      SalePrice        Mas.Vnr     Two.Floors      Fireplace
##              0              0              0              0              0
##           Deck          Porch           Bath         Garage         Season
##              0              0              2              0              0
```

```
data.final = na.omit(data.clean)
```

Create 80% train and 20% test set and store in new CSV files.

```
set.seed(555)
trainindex <- sample.int(nrow(data.final),size = nrow(data.final)*.80, replace= FALSE) # 80/20 split

Amestrain <- data.final[trainindex,]
Amestest <- data.final[-trainindex,]

#write.csv(Amestrain, file="Amestrain.csv", quote=F, row.names=F)
#write.csv(Amestest, file="Amestest.csv", quote=F, row.names=F)
```