Assignment 1

Kevin Gardner

1/30/2022

# ————————————————

# Following is the link to my GitHub account:

# <https://github.com/Kgardner22/64060_-kgardner>

# ————————————————

# 1. Download a dataset from the web.

# My data source is Kaggle. Following are the details:

# House Prices – Advanced Regression Techniques

# Predict sales prices and practice feature engineering, RFs, and gradient boosting

# <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

# 2. Import the dataset into R

House\_Prices\_train <- read.csv("C:/R/MyData/House\_Prices\_train.csv", header=TRUE)

# 3. Print out descriptive statistics for a selection of quantitative and categorical variables.

# The summary command will show a variety of descriptive statistics for each variable in the data set including the minimum, 1st quartile, median, mean, 3rd quartile, maximum values and if any NAs are present  
  
 summary(House\_Prices\_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   
##

# Following are individual descriptive statistics for quantitative variables:

mean(House\_Prices\_train$SalePrice) # Mean Sale Price

## [1] 180921.2

median(House\_Prices\_train$SalePrice) # Median Sale Price

## [1] 163000

sd(House\_Prices\_train$SalePrice) # Standard Deviation of Sale Price

## [1] 79442.5

min(House\_Prices\_train$SalePrice) # Minimum Sale Price

## [1] 34900

max(House\_Prices\_train$SalePrice) # Maximum Sale Price

## [1] 755000

# Following are descriptive statistics for categorical variables:

table(House\_Prices\_train$Street) # shows the frequency of homes located on gravel streets vs paved streets

##   
## Grvl Pave   
## 6 1454

table(House\_Prices\_train$CentralAir) # shows the frequency of homes with and without central air

##   
## N Y   
## 95 1365

table(House\_Prices\_train$CentralAir, House\_Prices\_train$Electrical) # cross classification of homes with and without central air (Y/N) and the type of electrical for the home (fuse box, electrical)

##   
## FuseA FuseF FuseP Mix SBrkr  
## N 22 18 3 0 52  
## Y 72 9 0 1 1282

# To show the percentage of the frequency for each value in a specific categorical variable (such as SaleCondition)

table1 <- table(House\_Prices\_train$SaleCondition)  
 prop.table(table1)

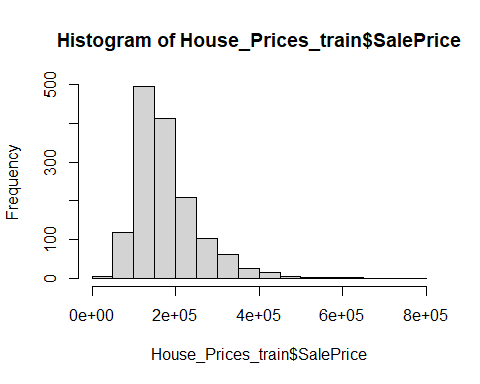
##   
## Abnorml AdjLand Alloca Family Normal Partial   
## 0.069178082 0.002739726 0.008219178 0.013698630 0.820547945 0.085616438

# 4. Transform at least one variable. It doesn’t matter what the transformation is.

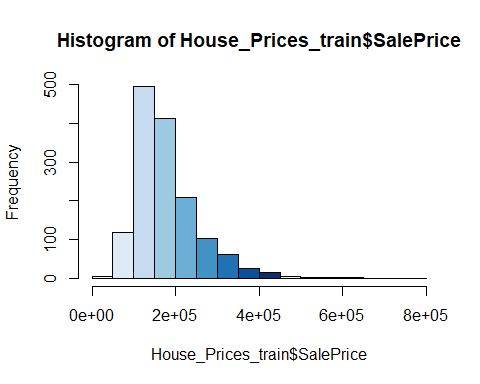
House\_Prices\_train$SalePrice\_Transformed <- (House\_Prices\_train$SalePrice - mean(House\_Prices\_train$SalePrice))/sd(House\_Prices\_train$SalePrice)  
  
   
# Create a new variable for total square feet (TotalLivingSF) which is the square footage of the 1st and 2nd floor combined:  
   
 House\_Prices\_train$TotalLivingSF <- (House\_Prices\_train$X1stFlrSF + House\_Prices\_train$X2ndFlrSF)

# 5. Plot at least one quantitative variable, and one scatterplot.

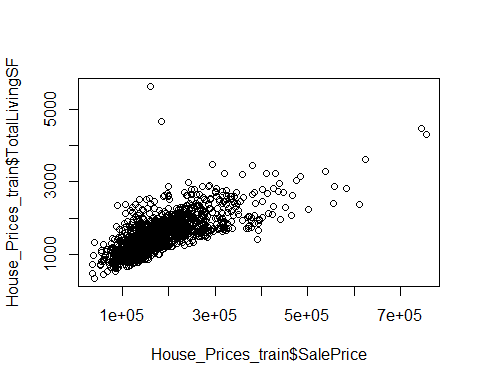
# Show histogram of SalePrice  
   
 hist(House\_Prices\_train$SalePrice)



# We could also add color to the Histogram to improve the visualization  
  
 hist(House\_Prices\_train$SalePrice, col = blues9)



# Show scatterplot of sales price (SalePrice) to total square footage (TotalLivingSF):  
   
 plot(House\_Prices\_train$SalePrice, House\_Prices\_train$TotalLivingSF)



# We can add color to the scatterplot as well to improve the visualization  
  
 plot(House\_Prices\_train$SalePrice, House\_Prices\_train$TotalLivingSF, col = blues9)

