

# Linear Regression Analysis: Regression Case Study

*Neerja Doshi, Sri Santhosh Hari, Ker-Yu Ong, Nicha Ruchirawat*

## Part 0: Exploratory Data Analysis

```
rawDF <- read.csv("/Users/booranium/usf/601_regression/project/housing.txt", stringsAsFactors = T)
#rawDF <- read.csv("/Users/santhoshhari/Documents/Coursework/LinearRegression/IowaHousing/Data/housing.
```

### Structure of Data:

The Iowa housing dataset contains 1460 rows and 81 variables, a glimpse of which is as follows:

```
str(rawDF)

## '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     : Factor w/ 5 levels "C (all)","FV",...: 4 4 4 4 4 4 4 5 4 ...
##  $ 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       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",...: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",...: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",...: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",...: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",...: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",...: 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1fam","2fmCon",...: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",...: 6 3 6 6 6 1 3 6 1 2 ...
##  $ 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    : Factor w/ 6 levels "Flat","Gable",...: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",...: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",...: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",...: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",...: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",...: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",...: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",...: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",...: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
```

```

## $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",...: 6 6 6 6 6 6 6 2 6 6 ...
## $ 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      : Factor w/ 6 levels "Floor","GasA",...: 2 2 2 2 2 2 2 2 2 ...
## $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",...: 1 1 1 3 1 1 1 1 3 1 ...
## $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 ...
## $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",...: 5 5 5 5 5 5 5 5 2 5 ...
## $ 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  : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 4 3 3 3 4 3 4 4 4 ...
## $ TotRmsAbvGrd : int   8 6 6 7 9 5 7 7 8 5 ...
## $ Functional   : Factor w/ 7 levels "Maj1","Maj2",...: 7 7 7 7 7 7 7 3 7 ...
## $ Fireplaces   : int   0 1 1 1 1 0 1 2 2 2 ...
## $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",...: NA 5 5 3 5 NA 3 5 5 5 ...
## $ GarageType   : Factor w/ 6 levels "2Types","Attchd",...: 2 2 2 6 2 2 2 2 6 2 ...
## $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
## $ GarageFinish : Factor w/ 3 levels "Fin","Rfn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
## $ 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   : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 2 3 ...
## $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 5 5 ...
## $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
## $ 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       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
## $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",...: NA NA NA NA NA 3 NA NA NA NA ...
## $ MiscFeature   : Factor w/ 4 levels "Gar2","Othr",...: NA NA NA NA NA 3 NA 3 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     : Factor w/ 9 levels "COD","Con","ConLD",...: 9 9 9 9 9 9 9 9 9 9 ...
## $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",...: 5 5 5 1 5 5 5 5 1 5 ...
## $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

```

At first glance, we see that most of the variables are categorical - both numeric and character types - and only a handful are continuous. The response variable for our analysis is `SalePrice`, and the remaining 79 variables (excluding the record ID column) are considered potential predictor variables. Checking the data dictionary, we found the following distribution for the predictor variables:

- 49 categorical
- 19 are continuous, e.g. area, price
- 11 are discrete, e.g. count, year

There are 0 duplicate rows in the dataset.

## Handling NA Values

Below, we compute that number and percentage of NAs per variable in the dataset having at least 1 NA.

```
NA_columns <- colnames(rawDF)[unique(which(is.na(rawDF), arr.ind = T)[,2])]

NA_count <- rawDF %>%
  select(NA_columns) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  gather(key = "Variable", value = "num_na", everything()) %>%
  arrange(desc(num_na))

NA_count %<>% mutate(perc_na = paste(round(num_na/nrow(rawDF),4)*100,"%"))
colnames(NA_count) <- c("**Variable**", "**Number of NA**", "**Percentage of NA**")
row.names(NA_count) <- NULL
knitr::kable(NA_count, caption = "\\label{tab:NACount} Variable NA Count and Percentage",
format.args = list(big.mark = ','))
```

Table 1: Variable NA Count and Percentage

Variable	Number of NA	Percentage of NA
PoolQC	1,453	99.52 %
MiscFeature	1,406	96.3 %
Alley	1,369	93.77 %
Fence	1,179	80.75 %
FireplaceQu	690	47.26 %
LotFrontage	259	17.74 %
GarageType	81	5.55 %
GarageYrBlt	81	5.55 %
GarageFinish	81	5.55 %
GarageQual	81	5.55 %
GarageCond	81	5.55 %
BsmtExposure	38	2.6 %
BsmtFinType2	38	2.6 %
BsmtQual	37	2.53 %
BsmtCond	37	2.53 %
BsmtFinType1	37	2.53 %
MasVnrType	8	0.55 %
MasVnrArea	8	0.55 %
Electrical	1	0.07 %

The data dictionary tells us that for most of the fields in Table 1, NA is actually meaningful, indicating non-applicability or a lack of the feature rather than missing data. After checking the data dictionary for the meaning of each field, we imputed - for every categorical variable for which NA was meaningful - NAs with 0s.

```
# Create a copy of rawDF to be our working data frame
housingDF <- rawDF

# Update NAs with 0s for applicable fields
levels(housingDF$PoolQC) <- c("0", levels(housingDF$PoolQC))
housingDF$PoolQC[is.na(housingDF$PoolQC)] <- "0"
```

```

levels(housingDF$MiscFeature) <- c("0", levels(housingDF$MiscFeature))
housingDF$MiscFeature[is.na(housingDF$MiscFeature)] <- "0"
levels(housingDF$Alley) <- c("0", levels(housingDF$Alley))
housingDF$Alley[is.na(housingDF$Alley)] <- "0"
levels(housingDF$Fence) <- c("0", levels(housingDF$Fence))
housingDF$Fence[is.na(housingDF$Fence)] <- "0"
levels(housingDF$FireplaceQu) <- c("0", levels(housingDF$FireplaceQu))
housingDF$FireplaceQu[is.na(housingDF$FireplaceQu)] <- "0"
levels(housingDF$GarageType) <- c("0", levels(housingDF$GarageType))
housingDF$GarageType[is.na(housingDF$GarageType)] <- "0"
levels(housingDF$GarageFinish) <- c("0", levels(housingDF$GarageFinish))
housingDF$GarageFinish[is.na(housingDF$GarageFinish)] <- "0"
levels(housingDF$GarageQual) <- c("0", levels(housingDF$GarageQual))
housingDF$GarageQual[is.na(housingDF$GarageQual)] <- "0"
levels(housingDF$GarageCond) <- c("0", levels(housingDF$GarageCond))
housingDF$GarageCond[is.na(housingDF$GarageCond)] <- "0"
levels(housingDF$BsmtExposure) <- c("0", levels(housingDF$BsmtExposure))
housingDF$BsmtExposure[is.na(housingDF$BsmtExposure)] <- "0"
levels(housingDF$BsmtFinType2) <- c("0", levels(housingDF$BsmtFinType2))
housingDF$BsmtFinType2[is.na(housingDF$BsmtFinType2)] <- "0"
levels(housingDF$BsmtQual) <- c("0", levels(housingDF$BsmtQual))
housingDF$BsmtQual[is.na(housingDF$BsmtQual)] <- "0"
levels(housingDF$BsmtCond) <- c("0", levels(housingDF$BsmtCond))
housingDF$BsmtCond[is.na(housingDF$BsmtCond)] <- "0"
levels(housingDF$BsmtFinType1) <- c("0", levels(housingDF$BsmtFinType1))
housingDF$BsmtFinType1[is.na(housingDF$BsmtFinType1)] <- "0"

```

We then re-check the count and percentage of NAs per variable left in the dataset.

```

NA_columns <- colnames(housingDF)[unique(which(is.na(housingDF), arr.ind = T)[,2])]
NA_count <- housingDF %>%
  select(NA_columns) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  gather(key = "Variable", value = "num_na", everything()) %>%
  arrange(desc(num_na))

NA_count %<>% mutate(perc_na = paste(round(num_na/nrow(housingDF),4)*100,"%"))
colnames(NA_count) <- c("**Variable**", "**Number of NA**", "**Percentage of NA**")
row.names(NA_count) <- NULL
knitr::kable(NA_count, caption = "\\label{tab:NACount1} Variable NA Count and Percentage(after replacing
format.args = list(big.mark = ','))

```

Table 2: Variable NA Count and Percentage(after replacing NAs with 0s, where appropriate)

Variable	Number of NA	Percentage of NA
LotFrontage	259	17.74 %
GarageYrBlt	81	5.55 %
MasVnrType	8	0.55 %
MasVnrArea	8	0.55 %
Electrical	1	0.07 %

Table 2 shows the list of remaining variables where NA indicates missing data. We impute NAs in these

variables with

- mean of the data, for continuous variables (LotFrontage)
- median of the data, for discrete variables (GarageYrBlt)
- mode of the data, for categorical variables (MasVnrType, Electrical)

```
# Function to get mode of data
getmode <- function(v) {
  univq <- unique(v)
  univq[which.max(tabulate(match(v, univq)))]
}

# Impute NAs
housingDF$LotFrontage[is.na(housingDF$LotFrontage)] <- mean(housingDF$LotFrontage, na.rm = T)
housingDF$GarageYrBlt[is.na(housingDF$GarageYrBlt)] <- median(housingDF$GarageYrBlt, na.rm=T)
housingDF$MasVnrType[is.na(housingDF$MasVnrType)] <- getmode(housingDF$MasVnrType)
housingDF$MasVnrArea[is.na(housingDF$MasVnrArea)] <- 0
housingDF$Electrical[is.na(housingDF$Electrical)] <- getmode(housingDF$Electrical)
```

Since Masonry veneer area (**MasVnrArea**) is directly related to **MasVnrType**, we impute for area based on the mode of **MasVnrType**, which is None. Our cleaned dataset is named **housingDF**.

## Data Visualization

With our clean dataset, we perform exploratory data visualization of the distribution of key measures such as volume and sale price of houses by what we hypothesize to be key predictor variables.

To begin with, we check the distribution of sale prices using a histogram and box-plot.

```
#hist(housingDF$SalePrice, main = "Histogram of Sale Price")
#boxplot(housingDF$SalePrice, main = "Boxplot of Sale Price")
summary(housingDF$SalePrice)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 34900 129975 163000 180921 214000 755000
```

Intuition suggests the neighborhood is a key determining factor in a house's sale price, hence below, we plot the distribution of sale price by neighborhood.

```
housingDF %>%
  select(Neighborhood, SalePrice) %>%
  ggplot(aes(factor(Neighborhood), SalePrice)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  xlab('Neighborhoods')
```

From Figure 1, we can observe that Brookside and Meadow Vista have the lowest median house price while Northridge and Northridge Height have the highest median house price as well as several outliers.

We then distribution of houses by a number of key features we hypothesize to be important in determining housing price: the property's zoning class (**MSZoning**), type of road access to the property (**Street**), type of alley access to the property (**Alley**), and type of utilities available (**Utilities**)

```
plotHist <- function(data_in, i) {
  data <- data.frame(x=data_in[[i]])
  p <- ggplot(data=data, aes(x=factor(x))) +
    stat_count() +
    xlab(colnames(data_in)[i]) +
```

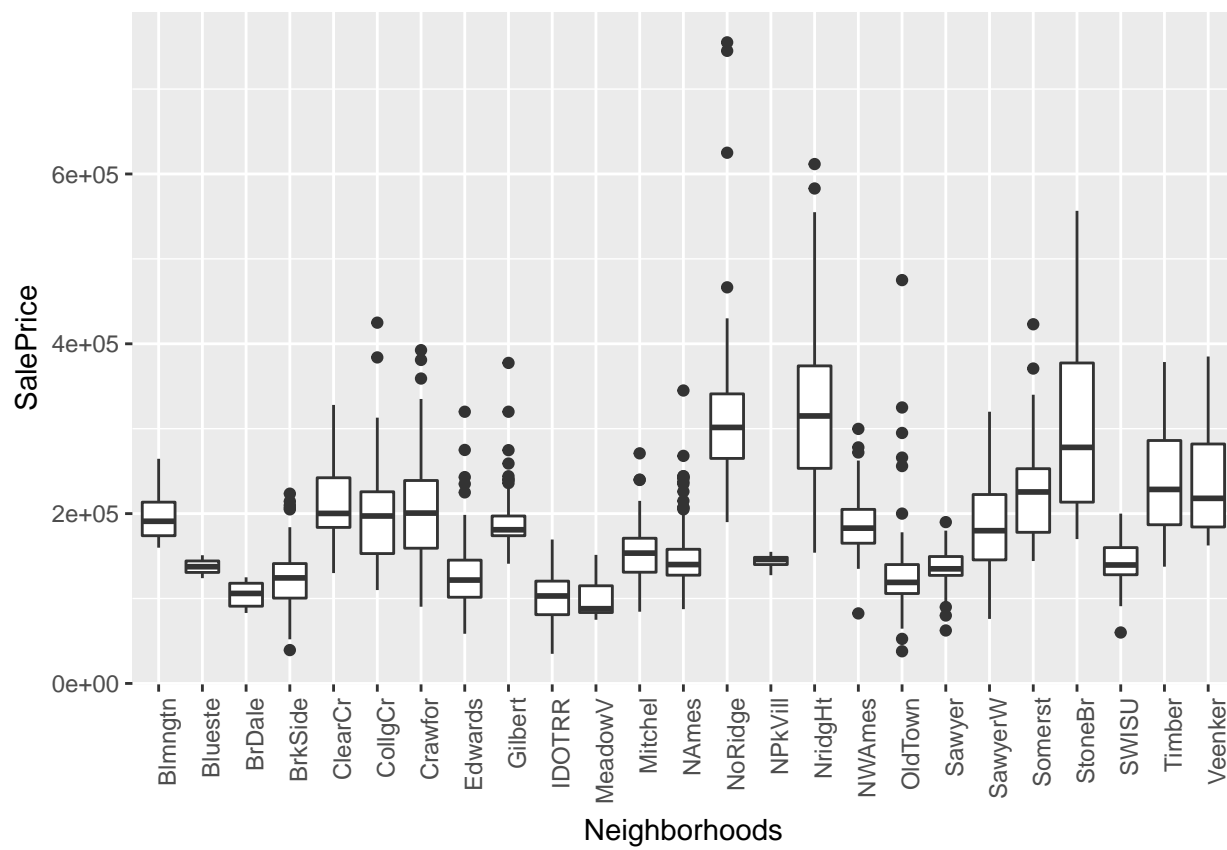


Figure 1: SalePrice distribution per neighborhood

```

    theme_light() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
  return (p)
}

doPlots <- function(data_in, fun, ii, ncol=3) {
  pp <- list()
  for (i in ii) {
    p <- fun(data_in=data_in, i=i)
    pp <- c(pp, list(p))
  }
  do.call("grid.arrange", c(pp, ncol=ncol))
}

plotDen <- function(data_in, i){
  data <- data.frame(x=data_in[[i]], SalePrice = data_in$SalePrice)
  p <- ggplot(data= data) +
    geom_line(aes(x = x), stat = 'density', size = 1,alpha = 1.0) +
    xlab(paste0((colnames(data_in)[i]), '\n', 'Skewness: ',
                round(skewness(data_in[[i]], na.rm = TRUE), 2))) +
    theme_light()
  return(p)
}

plotCorr <- function(data_in, i){
  data <- data.frame(x = data_in[[i]], SalePrice = data_in$SalePrice)
  p <- ggplot(data, aes(x = x, y = SalePrice)) +
    geom_point(na.rm = TRUE) +
    geom_smooth(method = lm ) +
    xlab(paste0(colnames(data_in)[i], '\n', 'R-Squared: ',
                round(cor(data_in[[i]], data$SalePrice, use = 'complete.obs'), 2))) +
    theme_light()
  return(suppressWarnings(p))
}

doPlots(housingDF, fun = plotHist, ii = c(3,6,7,10), ncol = 2)

```

Figure 2 suggests that most of the houses are located in Medium/Low Density residential areas. We can also observe that most of the houses have paved road access, do not have alleys and have all public utilities(E,G,W,&S). From Figure ??{fig:hist2}, we can notice that most of the properties are regular or slightly irregular in shape, built on level surfaces with gentle slope.

We also plot the distribution of houses against a number of features related to the physical geography of the property:

```
doPlots(housingDF, fun = plotHist, ii = c(8,9,11,12), ncol = 2)
```

We see that ...

```

housingDF %>%
  select(LandSlope, Neighborhood) %>%
  arrange(Neighborhood) %>%
  group_by(Neighborhood, LandSlope) %>%
  summarize(Count = n()) %>%
  ggplot(aes(Neighborhood, Count)) +
  geom_bar(aes(fill = LandSlope), position = 'dodge', stat = 'identity')+

```

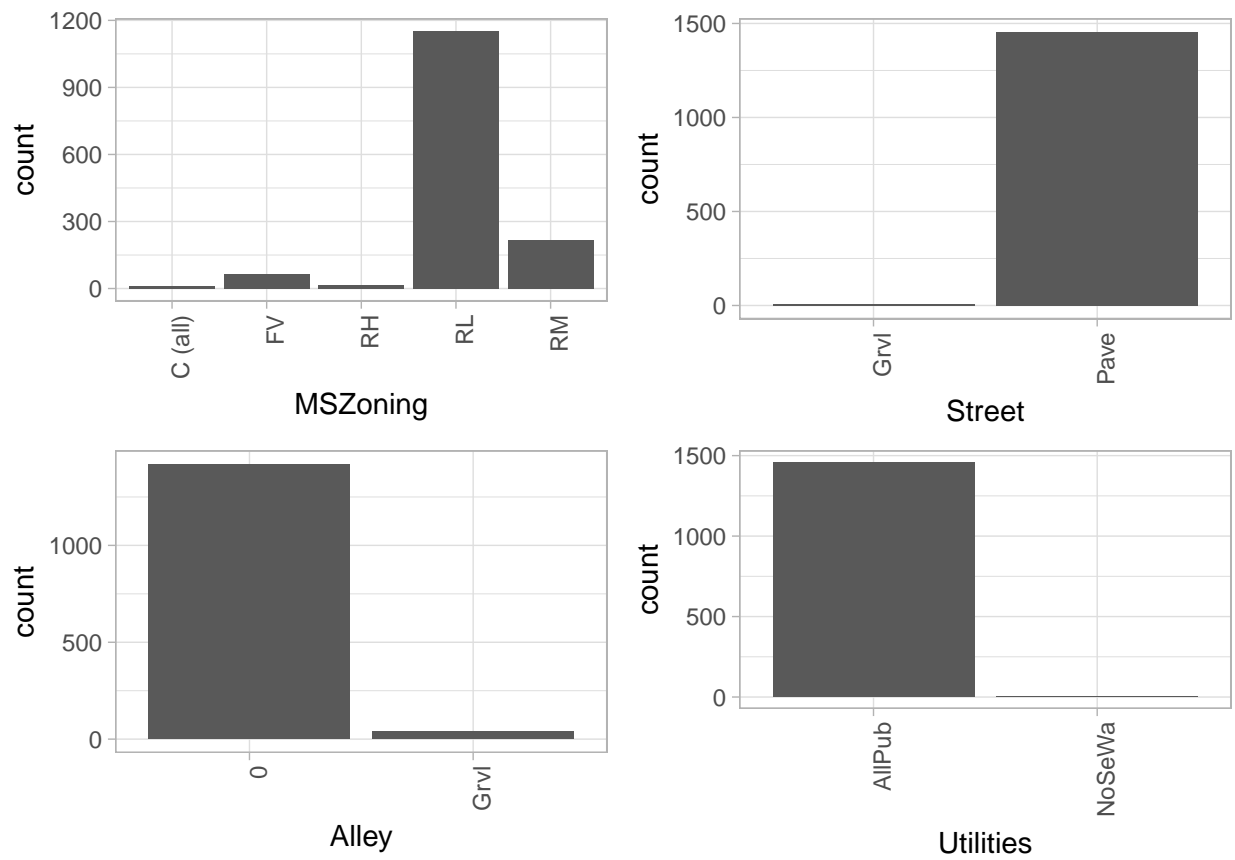


Figure 2: Locality, access, utility features distribution



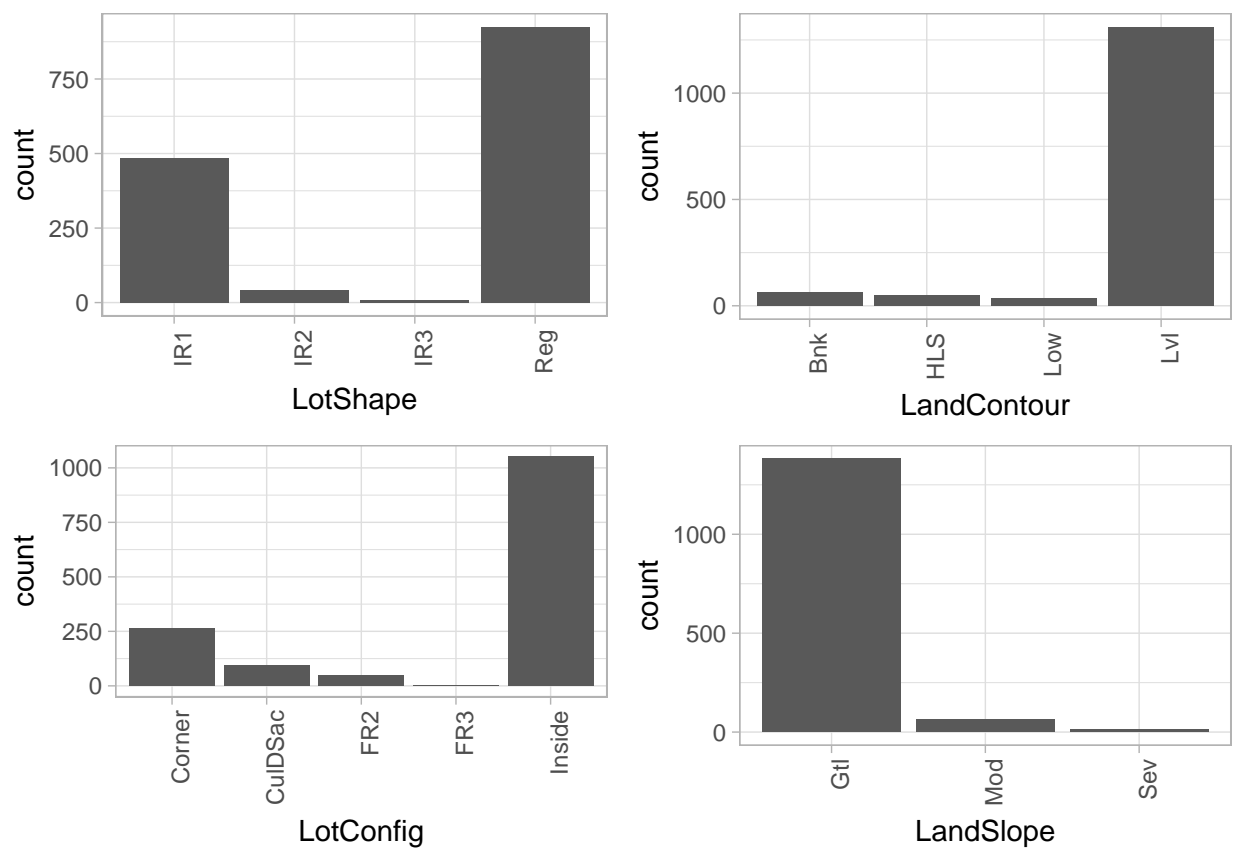


Figure 3: Lot/Land feature distribution

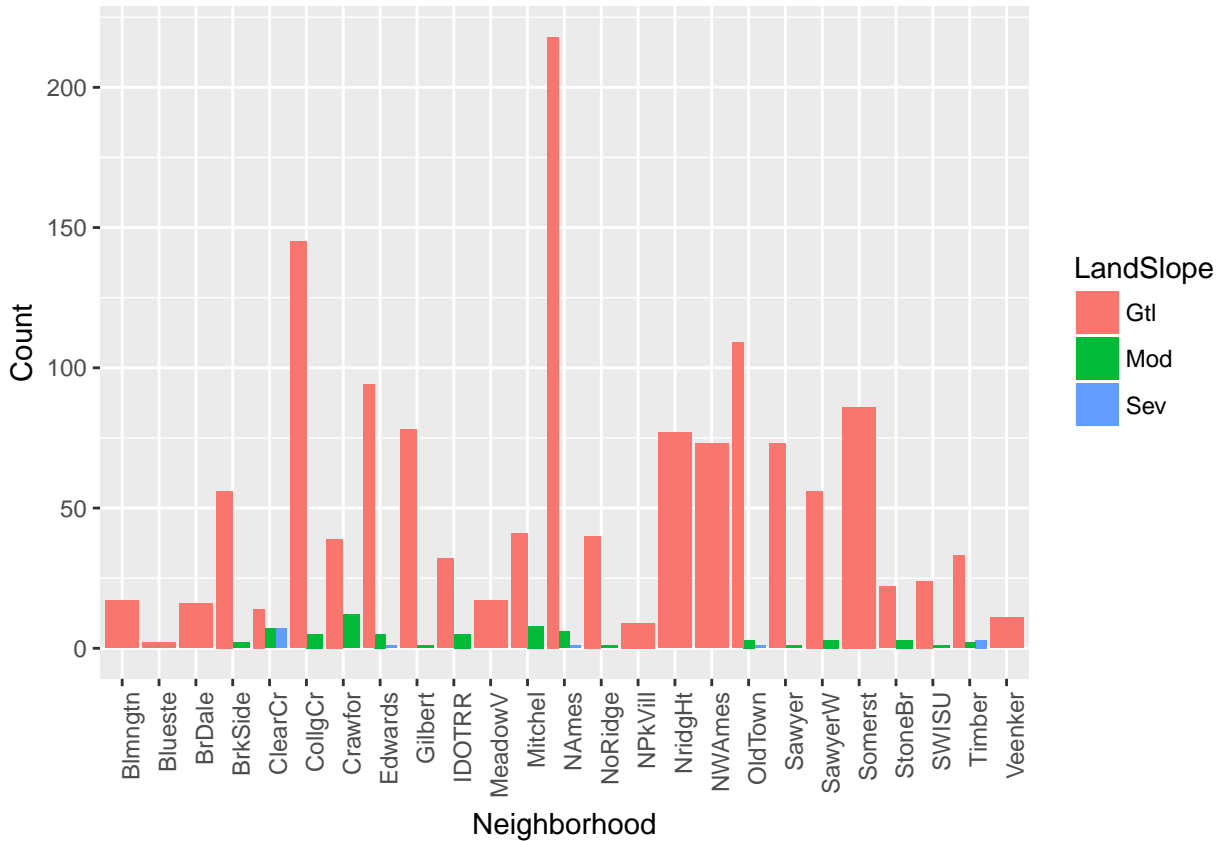


Figure 4: Neighborhood level slope distribution

```
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

From Figure ??{fig:hist3}, we can see that houses with severe slope are located only in Clear Creek and Timberland while more than 10 neighborhoods have properties with moderate slope.

## Testing for Influential Points

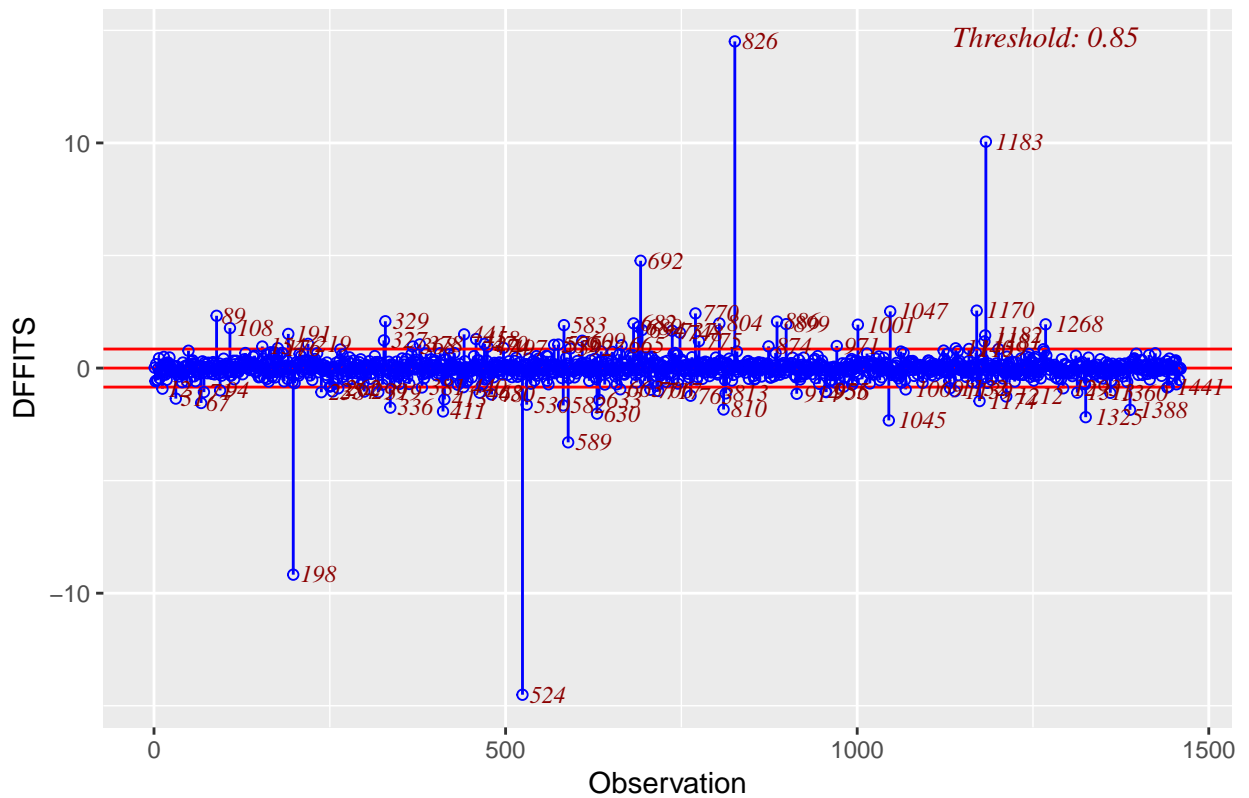
Having dealt with the NAs in our dataset, we use the `model.matrix()` function from the `glmnet` package to convert each categorical variable into an appropriate set of binary indicators: for a categorical variable that takes  $k$  levels, `model.matrix()` produces  $k-1$  binary indicators. We then reappend our response vector `SalePrice` to the resulting wide design matrix `designDF` to create `workingDF`, which includes both the converted predictors and response variables.

```
designDF <- model.matrix(SalePrice ~ ., data = housingDF)[,-1]
designDF <- as.data.frame(designDF)
workingDF <- cbind(designDF, SalePrice = housingDF$SalePrice)
```

In looking for influential points, we leverage the `OLSRR` package to test observations for influence according to both the DFFITS and studentized residuals criterions. We do this by first fitting a saturated model on `workingDF` and then calling `ols_dffits_plot()` and `ols_srsd_plot()` to plot the DFFITS and studentized residuals plots respectively.

```
ols_model <- lm(SalePrice ~ ., data = workingDF)
ols_dffits_plot(ols_model)
```

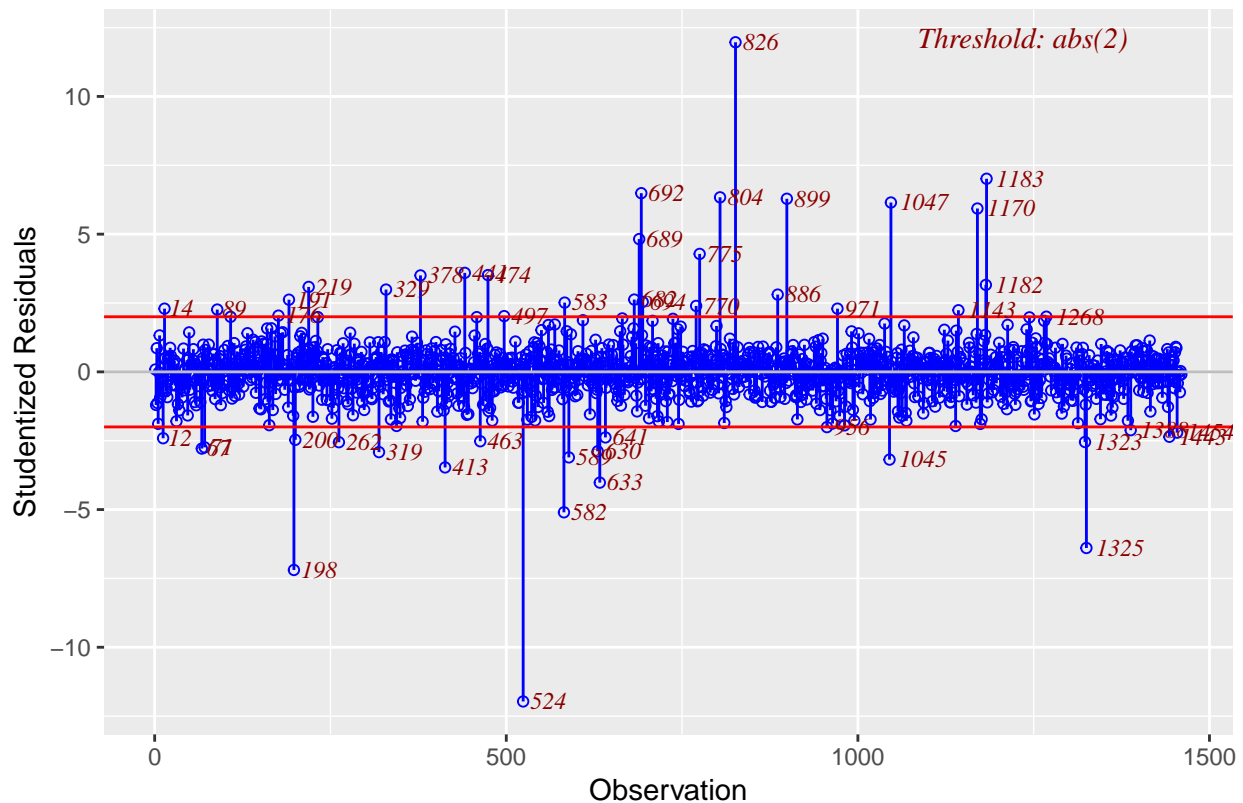
## Influence Diagnostics for SalePrice



The DFFITS plot shows 5 influential observations - 198, 418, 524, 826 and 1183 - with observations 524 and 826 being the most outstanding. This is corroborated by the studentized residuals plot, as shown below:

```
ols_srsd_chart(ols_model)
```

## Standardized Residuals Chart



As a remedial measure, we investigate data points 524 and 826 per the original dataset:

```
filter(workingDF, Id == 524 | Id == 826)
```

##	Id	MSSubClass	MSZoningFV	MSZoningRH	MSZoningRL	MSZoningRM	LotFrontage
## 1	524	60	0	0	1	0	130
## 2	826	20	0	0	1	0	114
##	LotArea	StreetPave	AlleyGrvl	AlleyPave	LotShapeIR2	LotShapeIR3	
## 1	40094	1	0	0	0	0	
## 2	14803	1	0	0	0	0	
##	LotShapeReg	LandContourHLS	LandContourLow	LandContourLvl	UtilitiesNoSeWa		
## 1	0	0	0	0	0		0
## 2	1	0	0	0	1		0
##	LotConfigCulDSac	LotConfigFR2	LotConfigFR3	LotConfigInside	LandSlopeMod		
## 1	0	0	0	1	0		
## 2	0	0	0	1	0		
##	LandSlopeSev	NeighborhoodBlueste	NeighborhoodBrDale	NeighborhoodBrkSide			
## 1	0	0	0	0			0
## 2	0	0	0	0			0
##	NeighborhoodClearCr	NeighborhoodCollgCr	NeighborhoodCrawfor				
## 1	0	0	0				
## 2	0	0	0				
##	NeighborhoodEdwards	NeighborhoodGilbert	NeighborhoodIDOTRR				
## 1	1	0	0				
## 2	0	0	0				
##	NeighborhoodMeadowV	NeighborhoodMitchel	NeighborhoodNames				
## 1	0	0	0				
## 2	0	0	0				

##	NeighborhoodNoRidge	NeighborhoodNPkVill	NeighborhoodNridgHt		
## 1	0	0	0		
## 2	0	0	1		
##	NeighborhoodNWAmes	NeighborhoodOldTown	NeighborhoodSawyer		
## 1	0	0	0		
## 2	0	0	0		
##	NeighborhoodSawyerW	NeighborhoodSomerst	NeighborhoodStoneBr		
## 1	0	0	0		
## 2	0	0	0		
##	NeighborhoodSWISU	NeighborhoodTimber	NeighborhoodVeenker	Condition1Feedr	
## 1	0	0	0	0	
## 2	0	0	0	0	
##	Condition1Norm	Condition1PosA	Condition1PosN	Condition1RR Ae	
## 1	0	0	1	0	
## 2	0	0	1	0	
##	Condition1RRAn	Condition1RRNe	Condition1RRNn	Condition2Feedr	
## 1	0	0	0	0	
## 2	0	0	0	0	
##	Condition2Norm	Condition2PosA	Condition2PosN	Condition2RR Ae	
## 1	0	0	1	0	
## 2	0	0	1	0	
##	Condition2RRAn	Condition2RRNn	BldgType2fmCon	BldgTypeDuplex	
## 1	0	0	0	0	
## 2	0	0	0	0	
##	BldgTypeTwnhs	BldgTypeTwnhsE	HouseStyle1.5Unf	HouseStyle1Story	
## 1	0	0	0	0	
## 2	0	0	0	1	
##	HouseStyle2.5Fin	HouseStyle2.5Unf	HouseStyle2Story	HouseStyleSFoyer	
## 1	0	0	1	0	
## 2	0	0	0	0	
##	HouseStyleSLvl	OverallQual	OverallCond	YearBuilt	YearRemodAdd
## 1	0	10	5	2007	2008
## 2	0	10	5	2007	2008
##	RoofStyleGable	RoofStyleGambrel	RoofStyleHip	RoofStyleMansard	
## 1	0	0	1	0	
## 2	0	0	1	0	
##	RoofStyleShed	RoofMatlCompShg	RoofMatlMembran	RoofMatlMetal	RoofMatlRoll
## 1	0	1	0	0	0
## 2	0	1	0	0	0
##	RoofMatlTar&Grv	RoofMatlWdShake	RoofMatlWdShngl	Exterior1stAsphShn	
## 1	0	0	0	0	
## 2	0	0	0	0	
##	Exterior1stBrkComm	Exterior1stBrkFace	Exterior1stCBlock		
## 1	0	0	0		
## 2	0	0	0		
##	Exterior1stCemntBd	Exterior1stHdBoard	Exterior1stImStucc		
## 1	1	0	0		
## 2	1	0	0		
##	Exterior1stMetalSd	Exterior1stPlywood	Exterior1stStone	Exterior1stStucco	
## 1	0	0	0	0	
## 2	0	0	0	0	
##	Exterior1stVinylSd	Exterior1stWd Sdng	Exterior1stWdShing		
## 1	0	0	0		
## 2	0	0	0		

##	Exterior2ndAsphShn	Exterior2ndBrk	Cmn	Exterior2ndBrkFace		
## 1	0		0		0	
## 2	0		0		0	
##	Exterior2ndCBlock	Exterior2ndCmentBd	Exterior2ndHdBoard			
## 1	0		1		0	
## 2	0		1		0	
##	Exterior2ndImStucc	Exterior2ndMetalSd	Exterior2ndOther			
## 1	0		0		0	
## 2	0		0		0	
##	Exterior2ndPlywood	Exterior2ndStone	Exterior2ndStucco	Exterior2ndVinylSd		
## 1	0		0		0	0
## 2	0		0		0	0
##	Exterior2ndWd	Sdng	Exterior2ndWd	Shng	MasVnrTypeBrkFace	MasVnrTypeNone
## 1	0		0		0	0
## 2	0		0		1	0
##	MasVnrTypeStone	MasVnrArea	ExterQualFa	ExterQualGd	ExterQualTA	
## 1	1	762	0	0	0	
## 2	0	816	0	0	0	
##	ExterCondFa	ExterCondGd	ExterCondPo	ExterCondTA	FoundationCBlock	
## 1	0	0	0	1	0	
## 2	0	0	0	1	0	
##	FoundationPConc	FoundationSlab	FoundationStone	FoundationWood	BsmtQualEx	
## 1	1	0	0	0	0	0
## 2	1	0	0	0	0	0
##	BsmtQualFa	BsmtQualGd	BsmtQualTA	BsmtCondFa	BsmtCondGd	BsmtCondPo
## 1	0	0	0	0	0	1
## 2	0	0	0	0	0	1
##	BsmtCondTA	BsmtExposureAv	BsmtExposureGd	BsmtExposureMn	BsmtExposureNo	
## 1	0	1	0	0	0	
## 2	0	0	0	0	0	
##	BsmtFinType1ALQ	BsmtFinType1BLQ	BsmtFinType1GLQ	BsmtFinType1LwQ		
## 1	0	1	0	0		
## 2	0	1	0	0		
##	BsmtFinType1Rec	BsmtFinType1Unf	BsmtFinSF1	BsmtFinType2ALQ		
## 1	0	0	2260	0		
## 2	0	0	1636	0		
##	BsmtFinType2BLQ	BsmtFinType2GLQ	BsmtFinType2LwQ	BsmtFinType2Rec		
## 1	0	0	0	1		
## 2	0	0	0	1		
##	BsmtFinType2Unf	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	HeatingGasA	HeatingGasW
## 1	0	0	878	3138	1	0
## 2	0	0	442	2078	1	0
##	HeatingGrav	HeatingOthW	HeatingWall	HeatingQCFa	HeatingQCGd	HeatingQCPo
## 1	0	0	0	0	0	0
## 2	0	0	0	0	0	0
##	HeatingQCTA	CentralAirY	ElectricalFuseF	ElectricalFuseP	ElectricalMix	
## 1	0	1	0	0	0	
## 2	0	1	0	0	0	
##	ElectricalSBrkr	X1stFlrSF	X2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath
## 1	1	3138	1538	0	4676	1
## 2	1	2084	0	0	2084	1
##	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	KitchenQualFa
## 1	0	3	1	3	1	0
## 2	0	2	0	2	1	0

```

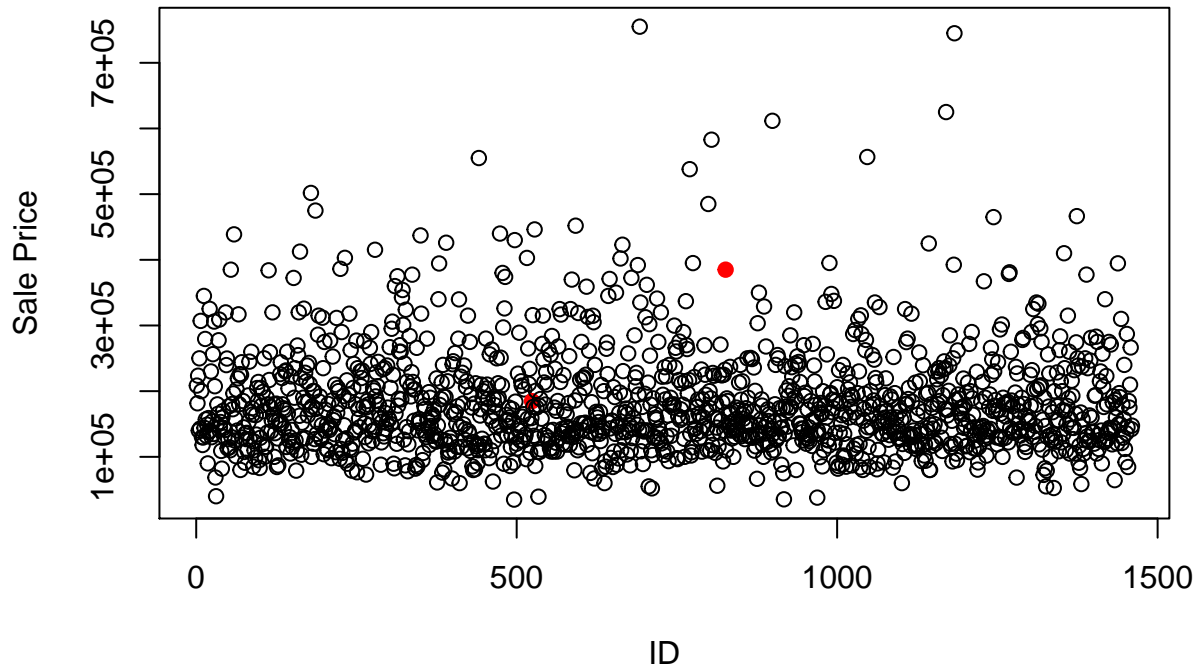
## KitchenQualGd KitchenQualTA TotRmsAbvGrd FunctionalMaj2 FunctionalMin1
## 1 0 0 11 0 0
## 2 0 0 7 0 0
## FunctionalMin2 FunctionalMod FunctionalSev FunctionalTyp Fireplaces
## 1 0 0 0 1 1
## 2 0 0 0 1 1
## FireplaceQuEx FireplaceQuFa FireplaceQuGd FireplaceQuPo FireplaceQuTA
## 1 0 1 0 0 0
## 2 0 1 0 0 0
## GarageType2Types GarageTypeAttchd GarageTypeBasment GarageTypeBuiltIn
## 1 0 0 1 0
## 2 1 0 0 0
## GarageTypeCarPort GarageTypeDetchd GarageYrBlt GarageFinishFin
## 1 0 0 2007 0
## 2 0 0 2007 0
## GarageFinishRfn GarageFinishUnf GarageCars GarageArea GarageQualEx
## 1 0 0 3 884 0
## 2 0 0 3 1220 0
## GarageQualFa GarageQualGd GarageQualPo GarageQualTA GarageCondEx
## 1 0 0 1 0 0
## 2 0 0 1 0 0
## GarageCondFa GarageCondGd GarageCondPo GarageCondTA PavedDriveP
## 1 0 0 1 0 0
## 2 0 0 1 0 0
## PavedDriveY WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch
## 1 1 208 406 0 0 0
## 2 1 188 45 0 0 0
## PoolArea PoolQCEx PoolQCFa PoolQCGd FenceGdPrv FenceGdWo FenceMnPrv
## 1 0 0 0 0 0 0
## 2 0 0 0 0 0 0
## FenceMnWw MiscFeatureGar2 MiscFeatureOthr MiscFeatureShed
## 1 0 0 0 0
## 2 0 0 0 0
## MiscFeatureTenC MiscVal MoSold YrSold SaleTypeCon SaleTypeConLD
## 1 0 0 10 2007 0 0
## 2 0 0 6 2008 0 0
## SaleTypeConLI SaleTypeConLw SaleTypeCWD SaleTypeNew SaleTypeOth
## 1 0 0 0 1 0
## 2 0 0 0 1 0
## SaleTypeWD SaleConditionAdjLand SaleConditionAlloca SaleConditionFamily
## 1 0 0 0 0
## 2 0 0 0 0
## SaleConditionNormal SaleConditionPartial SalePrice
## 1 0 1 184750
## 2 0 1 385000

```

Specifically, let's look at the values they take on a number of continuous variables as compared to other observations.

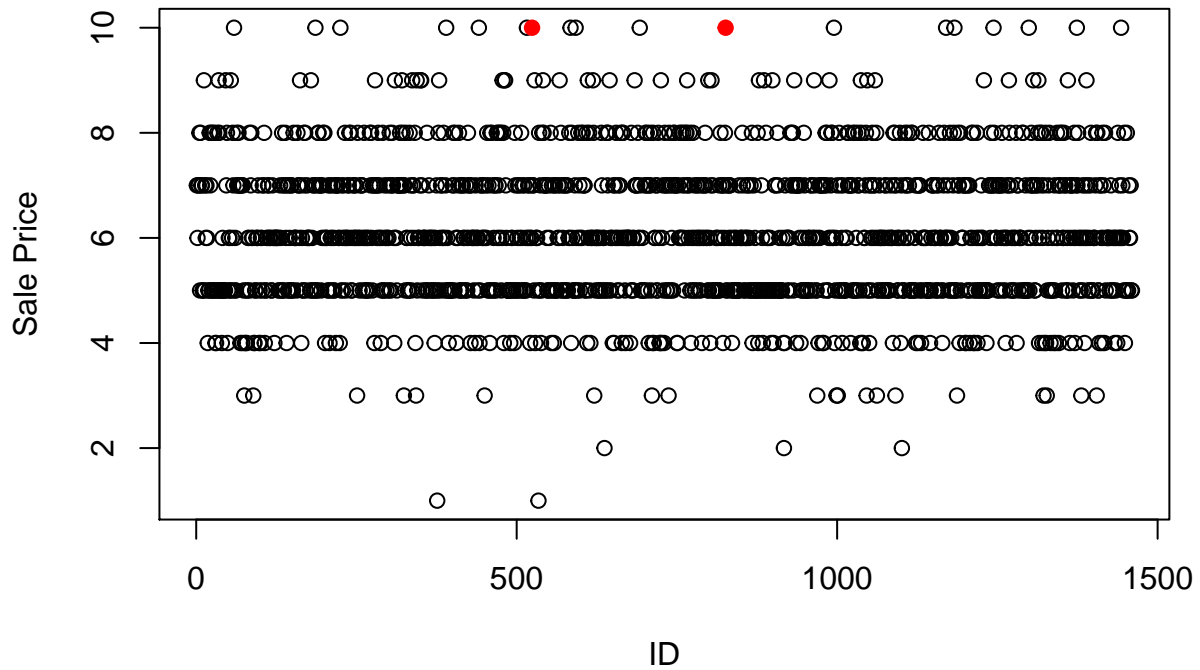
SalePrice:

```
plot(workingDF$Id, workingDF$SalePrice, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red",
```



OverallQual: they are in the bucket of 'Excellent' overall material and finish of the house.

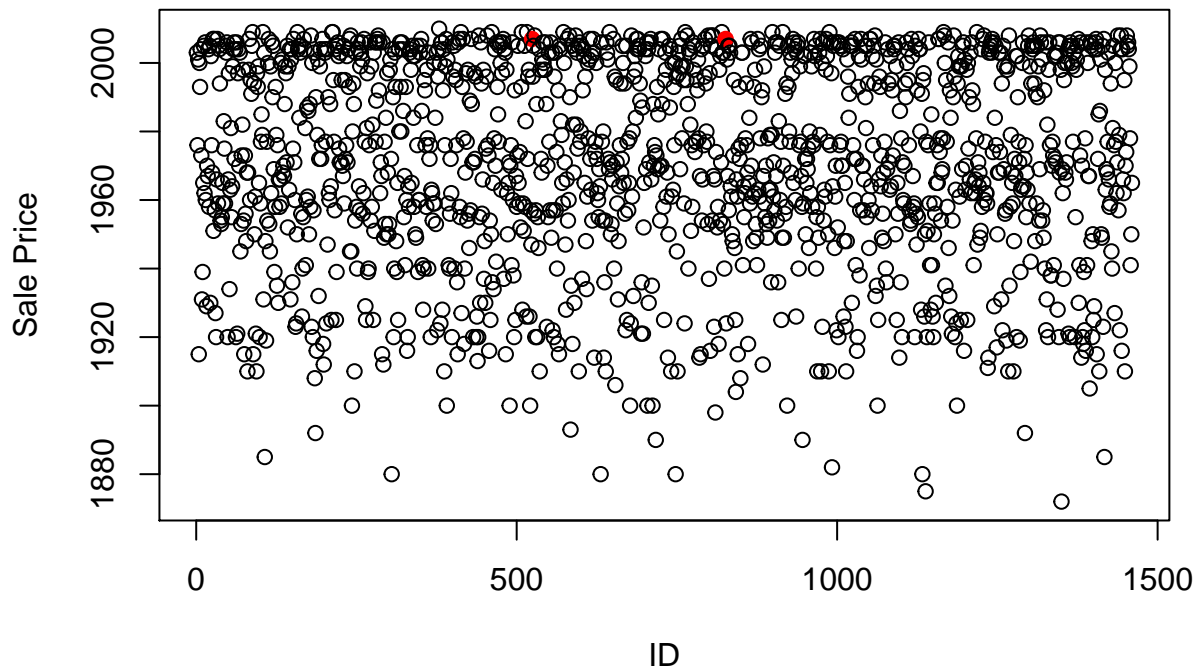
```
plot(workingDF$Id, workingDF$OverallQual, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red",
```



YearBuilt: they appear to be recently built

```
plot(workingDF$Id, workingDF$YearBuilt, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red",
```



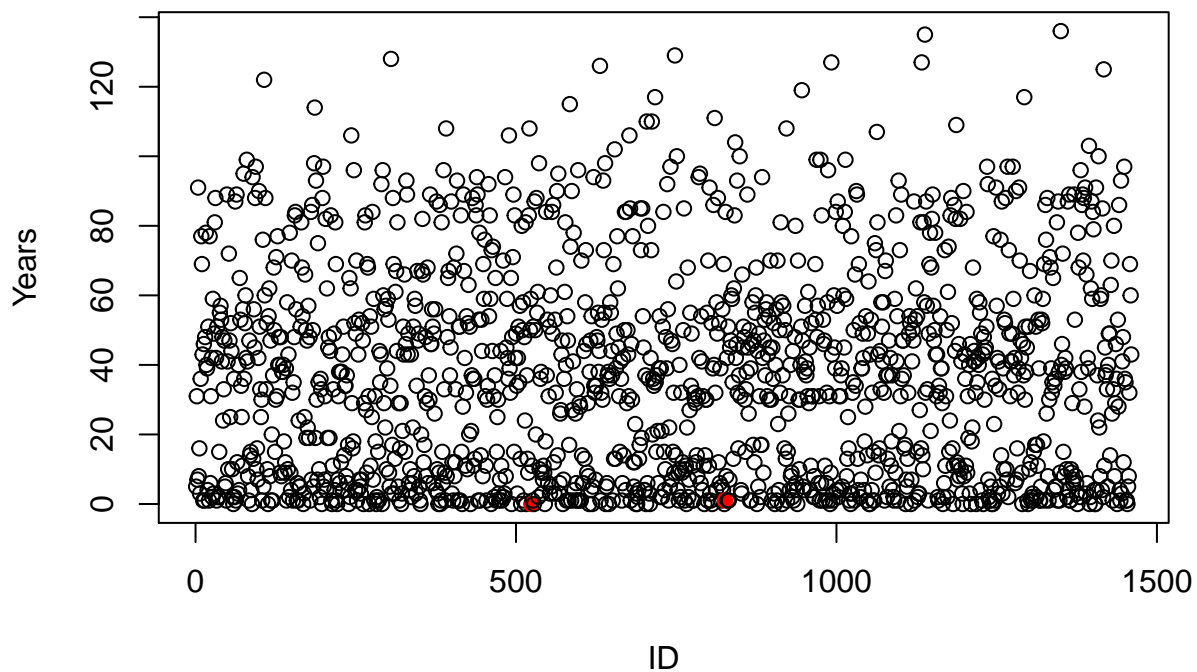


In fact: the difference between the year built and year sold for these two houses is small compared to the same difference for other houses:

Plot of year built vs. year sold - the two observations appear to be both built and sold recently

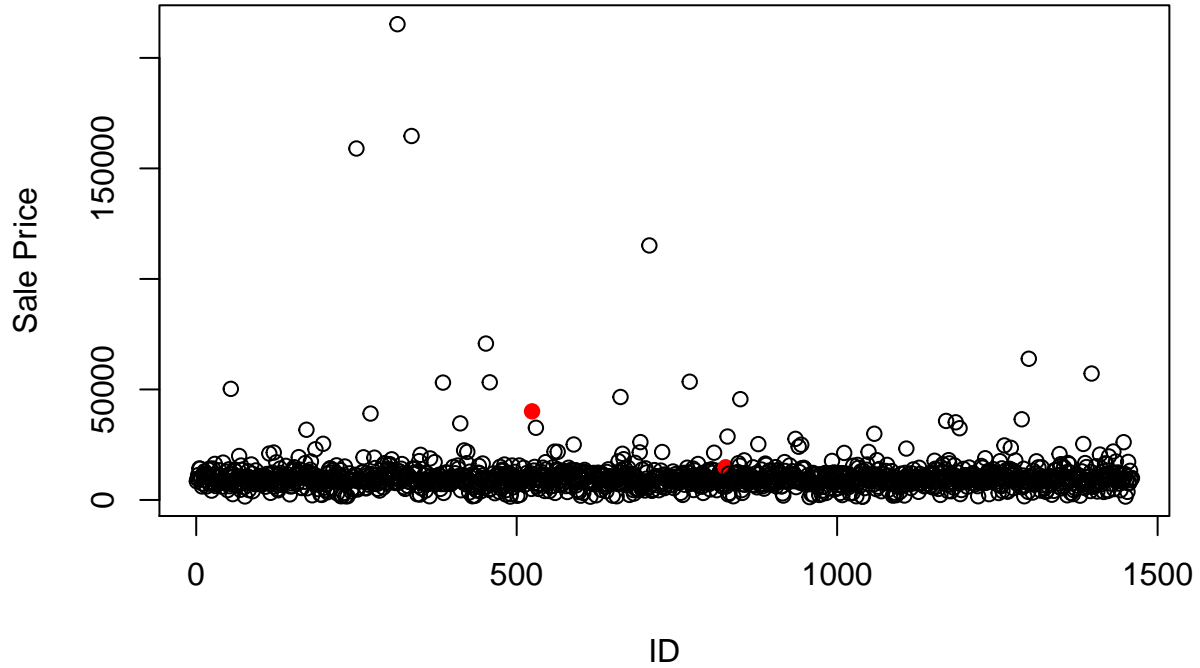
```
plot(workingDF$Id, workingDF$YrSold-workingDF$YearBuilt, col = ifelse(workingDF$Id == 524 | workingDF$Id == 824, "red", "black"))
```

## Plot of Years Between House Built vs. Sold



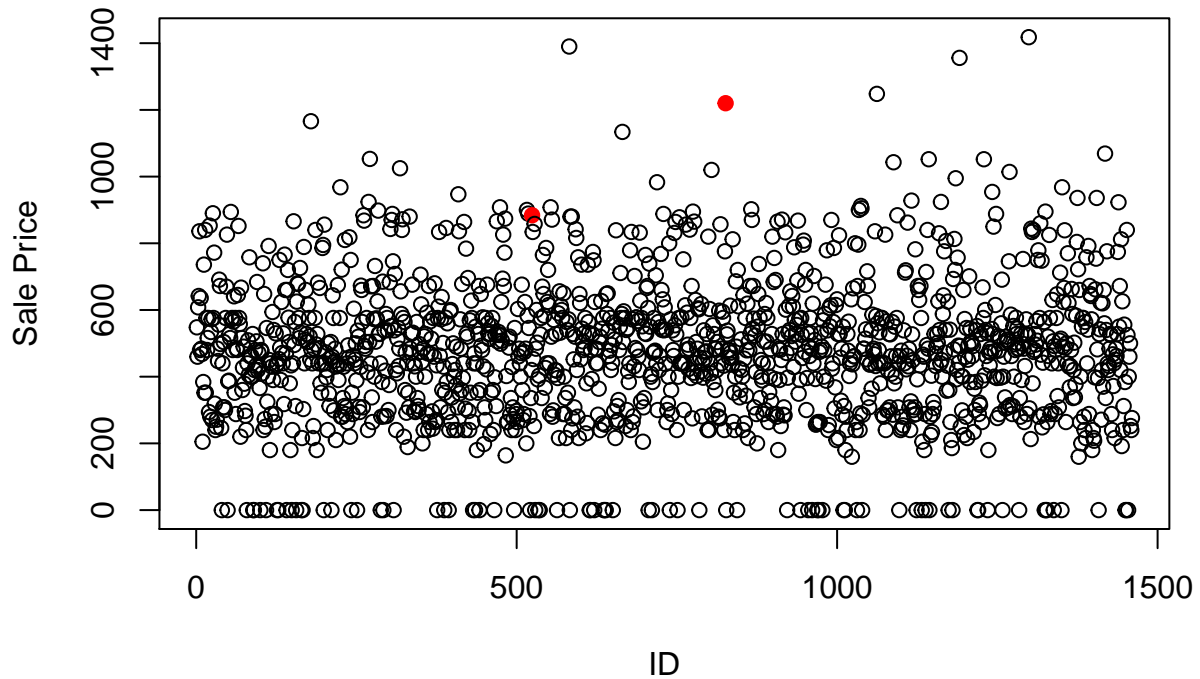
LotArea: Relatively large

```
plot(workingDF$Id, workingDF$LotArea, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red", "black"))
```

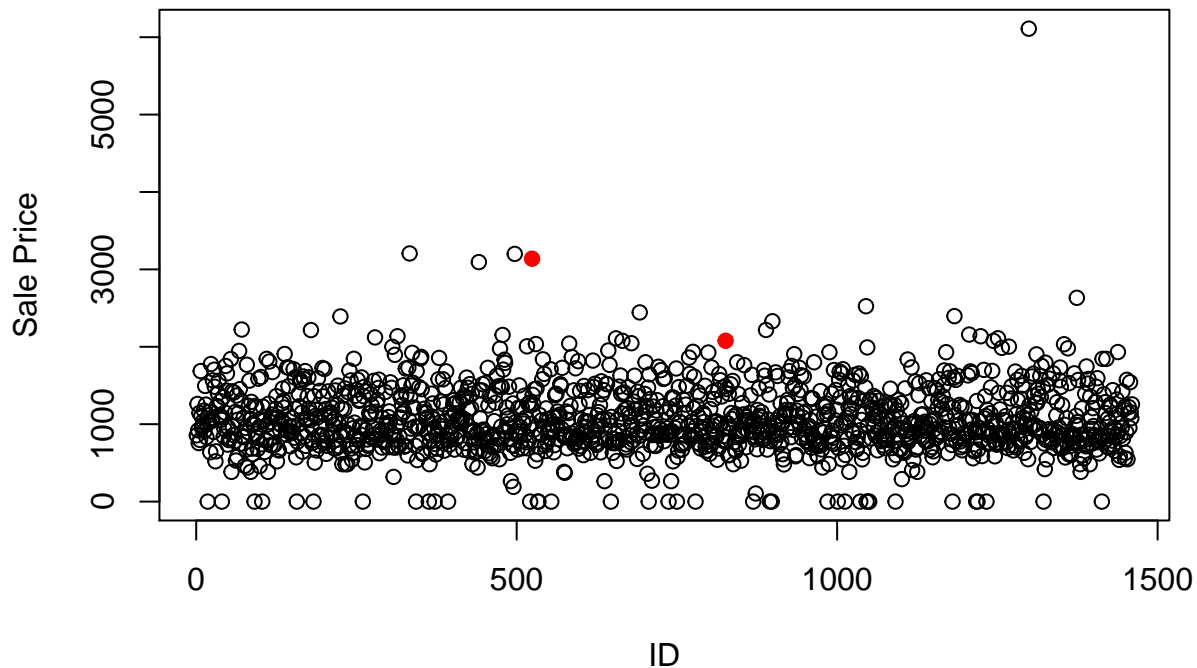


GarageArea: Relatively large

```
plot(workingDF$Id, workingDF$GarageArea, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red", "black"))
```



```
plot(workingDF$Id, workingDF$TotalBsmtSF, col = ifelse(workingDF$Id == 524 | workingDF$Id == 826, "red", "black"))
```



We can now plot the distribution of values for the variables in our `workingDF` and spotcheck for outliers:

```
cols <- colnames(workingDF)

for (c in cols){
  print(c)
  data <- workingDF[[c]]
  plot(data)
}
```

## Part I: Explanatory Modelling

For the purposes of variable selection, we refer to the saturated OLS model created above and perform stepwise model selection according to both AIC and BIC criteria.

```
model_aic <- step(ols_model, direction = 'backward', trace = F)
model_bic <- step(ols_model, k = log(nrow(workingDF)), direction = 'backward', trace = F)

## save the models
save(model_aic, file = "/Users/booranium/usf/601_regression/project/IowaHousing/AIC_model_v2.rda")
save(model_bic, file = "/Users/booranium/usf/601_regression/project/IowaHousing/BIC_model_v2.rda")
```

Per the AIC criterion, the following are the predictor variables significant at the  $\alpha = 0.05$  level.

```
# load the saved model to speed up processing
# load("/Users/booranium/usf/601_regression/project/IowaHousing/AIC_model_v2.rda")
# model loaded as model_aic

# find coefficients significant at the alpha = 0.05 level
bool_aic <- summary(model_aic)$coeff[-1,4] < 0.05
sig_var_aic <- names(bool_aic)[bool_aic == TRUE]
sig_var_aic
```

## [1]	"MSSubClass"	"MSZoningFV"	"MSZoningRH"
## [4]	"MSZoningRL"	"MSZoningRM"	"LotArea"
## [7]	"StreetPave"	"LandContourHLS"	"LandContourLvl"
## [10]	"LotConfigCulDSac"	"LandSlopeMod"	"LandSlopeSev"
## [13]	"NeighborhoodCollgCr"	"NeighborhoodCrawfor"	"NeighborhoodEdwards"
## [16]	"NeighborhoodGilbert"	"NeighborhoodMitchel"	"NeighborhoodNames"
## [19]	"NeighborhoodNoRidge"	"NeighborhoodNridgHt"	"NeighborhoodNWames"
## [22]	"NeighborhoodSawyer"	"NeighborhoodStoneBr"	"Condition1Norm"
## [25]	"Condition1PosN"	"Condition1RRAn"	"Condition2PosA"
## [28]	"Condition2PosN"	"Condition2RRAE"	"BldgTypeTwnhs"
## [31]	"BldgTypeTwnhsE"	"OverallQual"	"OverallCond"
## [34]	"YearBuilt"	"YearRemodAdd"	"RoofStyleShed"
## [37]	"RoofMatlCompShg"	"RoofMatlMembran"	"RoofMatlMetal"
## [40]	"RoofMatlRoll"	"`RoofMatlTar&Grv`"	"RoofMatlWdShake"
## [43]	"RoofMatlWdShngl"	"Exterior1stBrkFace"	"Exterior1stPlywood"
## [46]	"`Exterior1stWd Sdng`"	"MasVnrTypeNone"	"MasVnrTypeStone"
## [49]	"MasVnrArea"	"ExterQualGd"	"ExterQualTA"
## [52]	"FoundationWood"	"BsmtQualEx"	"BsmtQualFa"
## [55]	"BsmtQualGd"	"BsmtExposureAv"	"BsmtExposureMn"
## [58]	"BsmtFinType1BLQ"	"BsmtFinSF1"	"BsmtFinSF2"
## [61]	"BsmtUnfSF"	"HeatingQCGd"	"X1stFlrSF"
## [64]	"X2ndFlrSF"	"BedroomAbvGr"	"KitchenAbvGr"
## [67]	"KitchenQualFa"	"KitchenQualGd"	"KitchenQualTA"
## [70]	"TotRmsAbvGrd"	"FunctionalTyp"	"GarageCars"
## [73]	"GarageArea"	"GarageQualEx"	"GarageQualFa"
## [76]	"GarageQualGd"	"GarageQualPo"	"GarageCondEx"
## [79]	"GarageCondFa"	"GarageCondGd"	"GarageCondPo"
## [82]	"WoodDeckSF"	"ScreenPorch"	"PoolArea"
## [85]	"PoolQCEX"	"PoolQCFa"	"SaleTypeNew"
## [88]	"SaleConditionNormal"		

Per the BIC criterion, the following are the predictor variables significant at the  $\alpha = 0.05$  level.

```
# load the saved model to speed up processing
# load("/Users/booranium/usf/601_regression/project/IowaHousing/BIC_model.rda")
# model loaded as model_bic <- l

# find coefficients significant at the alpha = 0.05 level
bool_bic <- summary(model_aic)$coeff[-1,4] < 0.05
sig_var_bic <- names(bool_bic)[bool_bic == TRUE]
sig_var_bic
```

## [1]	"MSSubClass"	"MSZoningFV"	"MSZoningRH"
## [4]	"MSZoningRL"	"MSZoningRM"	"LotArea"
## [7]	"StreetPave"	"LandContourHLS"	"LandContourLvl"
## [10]	"LotConfigCulDSac"	"LandSlopeMod"	"LandSlopeSev"
## [13]	"NeighborhoodCollgCr"	"NeighborhoodCrawfor"	"NeighborhoodEdwards"
## [16]	"NeighborhoodGilbert"	"NeighborhoodMitchel"	"NeighborhoodNames"
## [19]	"NeighborhoodNoRidge"	"NeighborhoodNridgHt"	"NeighborhoodNWames"
## [22]	"NeighborhoodSawyer"	"NeighborhoodStoneBr"	"Condition1Norm"
## [25]	"Condition1PosN"	"Condition1RRAn"	"Condition2PosA"
## [28]	"Condition2PosN"	"Condition2RRAE"	"BldgTypeTwnhs"
## [31]	"BldgTypeTwnhsE"	"OverallQual"	"OverallCond"
## [34]	"YearBuilt"	"YearRemodAdd"	"RoofStyleShed"
## [37]	"RoofMatlCompShg"	"RoofMatlMembran"	"RoofMatlMetal"

```
## [40] "RoofMatlRoll"      "`RoofMatlTar&Grv`"  "RoofMatlWdShake"
## [43] "RoofMatlWdShngl"   "Exterior1stBrkFace" "Exterior1stPlywood"
## [46] "`Exterior1stWd Sdng`" "MasVnrTypeNone"     "MasVnrTypeStone"
## [49] "MasVnrArea"        "ExterQualGd"        "ExterQualTA"
## [52] "FoundationWood"    "BsmtQualEx"         "BsmtQualFa"
## [55] "BsmtQualGd"        "BsmtExposureAv"     "BsmtExposureMn"
## [58] "BsmtFinType1BLQ"   "BsmtFinSF1"         "BsmtFinSF2"
## [61] "BsmtUnfSF"         "HeatingQCGd"        "X1stFlrSF"
## [64] "X2ndFlrSF"         "BedroomAbvGr"       "KitchenAbvGr"
## [67] "KitchenQualFa"     "KitchenQualGd"      "KitchenQualTA"
## [70] "TotRmsAbvGrd"     "FunctionalTyp"       "GarageCars"
## [73] "GarageArea"        "GarageQualEx"       "GarageQualFa"
## [76] "GarageQualGd"      "GarageQualPo"       "GarageCondEx"
## [79] "GarageCondFa"     "GarageCondGd"       "GarageCondPo"
## [82] "WoodDeckSF"       "ScreenPorch"        "PoolArea"
## [85] "PoolQCEx"         "PoolQCFa"          "SaleTypeNew"
## [88] "SaleConditionNormal"
```

Furthermore, we see that both criterions select the same set of variables.

```
setdiff(sig_var_aic, sig_var_bic)
```

```
## character(0)
```

We can now perform OLS regression with our subset of significant variables, and the model summary is as follows:

```
model_sig_formula <- as.formula(paste("SalePrice ~ ", paste(sig_var_bic, collapse = "+")))
model_sig <- lm(formula = model_sig_formula, data = workingDF)
summary(model_sig)
```

```
##
## Call:
## lm(formula = model_sig_formula, data = workingDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -181937  -10032    171    10116   181937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.586e+06  1.236e+05  -12.826  < 2e-16 ***
## MSSubClass     -7.254e+01   2.848e+01   -2.547  0.010962 *
## MSZoningFV      3.850e+04   8.650e+03   4.451  9.23e-06 ***
## MSZoningRH      2.344e+04   9.584e+03   2.445  0.014590 *
## MSZoningRL      2.685e+04   7.975e+03   3.367  0.000782 ***
## MSZoningRM      1.996e+04   7.912e+03   2.523  0.011763 *
## LotArea        6.119e-01   8.683e-02   7.047  2.88e-12 ***
## StreetPave      3.467e+04   1.015e+04   3.414  0.000658 ***
## LandContourHLS   1.121e+04   4.197e+03   2.671  0.007643 **
## LandContourLvl   9.652e+03   2.896e+03   3.333  0.000881 ***
## LotConfigCulDSac 7.494e+03   2.585e+03   2.899  0.003805 **
## LandSlopeMod     6.720e+03   3.502e+03   1.919  0.055225 .
## LandSlopeSev    -3.303e+04   9.602e+03  -3.439  0.000600 ***
## NeighborhoodCollgCr -3.116e+03   2.663e+03  -1.170  0.242260
## NeighborhoodCrawfor 1.896e+04   3.776e+03   5.020  5.83e-07 ***
```

## NeighborhoodEdwards	-1.201e+04	2.797e+03	-4.293	1.88e-05	***
## NeighborhoodGilbert	-2.892e+03	3.390e+03	-0.853	0.393654	
## NeighborhoodMitchel	-1.743e+04	3.767e+03	-4.626	4.07e-06	***
## NeighborhoodNames	-1.124e+04	2.449e+03	-4.590	4.84e-06	***
## NeighborhoodNoRidge	3.327e+04	4.502e+03	7.390	2.54e-13	***
## NeighborhoodNridgHt	2.099e+04	3.677e+03	5.710	1.38e-08	***
## NeighborhoodNWAmes	-1.222e+04	3.327e+03	-3.672	0.000249	***
## NeighborhoodSawyer	-6.754e+03	3.282e+03	-2.058	0.039782	*
## NeighborhoodStoneBr	3.858e+04	5.162e+03	7.475	1.37e-13	***
## Condition1Norm	1.124e+04	2.069e+03	5.431	6.63e-08	***
## Condition1PosN	9.451e+03	6.025e+03	1.569	0.116947	
## Condition1RRAn	9.848e+03	5.006e+03	1.967	0.049369	*
## Condition2PosA	4.547e+04	2.430e+04	1.871	0.061494	.
## Condition2PosN	-2.261e+05	1.782e+04	-12.689	< 2e-16	***
## Condition2RRAE	-8.003e+04	3.556e+04	-2.251	0.024574	*
## BldgTypeTwnhs	-1.451e+04	5.074e+03	-2.859	0.004320	**
## BldgTypeTwnhsE	-9.863e+03	3.967e+03	-2.486	0.013025	*
## OverallQual	6.601e+03	8.730e+02	7.561	7.26e-14	***
## OverallCond	5.804e+03	7.176e+02	8.087	1.33e-15	***
## YearBuilt	3.700e+02	4.973e+01	7.440	1.76e-13	***
## YearRemodAdd	9.632e+01	4.787e+01	2.012	0.044393	*
## RoofStyleShed	5.918e+04	2.702e+04	2.191	0.028641	*
## RoofMatlCompShg	6.285e+05	3.295e+04	19.071	< 2e-16	***
## RoofMatlMembran	6.969e+05	4.163e+04	16.741	< 2e-16	***
## RoofMatlMetal	6.676e+05	4.152e+04	16.078	< 2e-16	***
## RoofMatlRoll	6.275e+05	4.029e+04	15.576	< 2e-16	***
## `RoofMatlTar&Grv`	6.163e+05	3.323e+04	18.546	< 2e-16	***
## RoofMatlWdShake	6.239e+05	3.488e+04	17.886	< 2e-16	***
## RoofMatlWdShngl	6.773e+05	3.391e+04	19.971	< 2e-16	***
## Exterior1stBrkFace	1.782e+04	3.608e+03	4.940	8.79e-07	***
## Exterior1stPlywood	-3.469e+03	2.609e+03	-1.329	0.183914	
## `Exterior1stWd Sdng`	1.486e+02	1.961e+03	0.076	0.939630	
## MasVnrTypeNone	4.696e+03	1.983e+03	2.368	0.018025	*
## MasVnrTypeStone	5.814e+03	2.576e+03	2.257	0.024192	*
## MasVnrArea	2.332e+01	5.314e+00	4.389	1.23e-05	***
## ExterQualGd	-2.102e+04	3.857e+03	-5.450	5.98e-08	***
## ExterQualTA	-2.141e+04	4.039e+03	-5.301	1.34e-07	***
## FoundationWood	-2.376e+04	1.325e+04	-1.793	0.073242	.
## BsmtQualEx	-1.082e+04	5.108e+03	-2.118	0.034346	*
## BsmtQualFa	-1.947e+04	2.723e+03	-7.147	1.43e-12	***
## BsmtQualGd	-1.581e+04	3.141e+03	-5.032	5.50e-07	***
## BsmtExposureAv	1.699e+04	2.633e+03	6.454	1.51e-10	***
## BsmtExposureMn	-4.197e+03	1.594e+03	-2.633	0.008556	**
## BsmtFinType1BLQ	3.708e+03	1.891e+03	1.961	0.050028	.
## BsmtFinSF1	3.749e+01	3.328e+00	11.265	< 2e-16	***
## BsmtFinSF2	2.598e+01	4.792e+00	5.421	6.98e-08	***
## BsmtUnfSF	2.130e+01	3.236e+00	6.582	6.58e-11	***
## HeatingQCGd	-3.434e+03	1.688e+03	-2.034	0.042133	*
## X1stFlrSF	4.992e+01	4.184e+00	11.930	< 2e-16	***
## X2ndFlrSF	5.601e+01	3.060e+00	18.305	< 2e-16	***
## BedroomAbvGr	-5.026e+03	1.175e+03	-4.276	2.03e-05	***
## KitchenAbvGr	-2.015e+04	3.667e+03	-5.494	4.67e-08	***
## KitchenQualFa	-1.907e+04	5.306e+03	-3.594	0.000338	***
## KitchenQualGd	-2.426e+04	3.179e+03	-7.632	4.30e-14	***

```
## KitchenQualTA      -2.380e+04  3.601e+03  -6.610  5.50e-11 ***
## TotRmsAbvGrd       2.429e+03  8.440e+02   2.878  0.004065 **
## FunctionalTyp      1.437e+04  2.660e+03   5.401  7.79e-08 ***
## GarageCars         5.378e+03  2.065e+03   2.604  0.009304 **
## GarageArea         1.260e+01  6.786e+00   1.856  0.063641 .
## GarageQualEx      -1.086e+05  2.613e+04  -4.158  3.41e-05 ***
## GarageQualFa      -9.902e+04  2.686e+04  -3.686  0.000237 ***
## GarageQualGd      -1.099e+05  3.174e+04  -3.464  0.000548 ***
## GarageQualPo      -1.049e+05  2.590e+04  -4.049  5.43e-05 ***
## GarageCondEx       9.179e+04  2.662e+04   3.448  0.000581 ***
## GarageCondFa       8.454e+04  2.754e+04   3.069  0.002187 **
## GarageCondGd       9.478e+04  2.900e+04   3.269  0.001108 **
## GarageCondPo       9.580e+04  2.624e+04   3.650  0.000272 ***
## WoodDeckSF         1.341e+01  5.359e+00   2.502  0.012479 *
## ScreenPorch        3.520e+01  1.141e+01   3.086  0.002073 **
## PoolArea           2.350e+02  3.109e+01   7.557  7.49e-14 ***
## PoolQCEX          -1.370e+05  2.466e+04  -5.556  3.32e-08 ***
## PoolQCFA          -8.998e+04  2.655e+04  -3.389  0.000722 ***
## SaleTypeNew        2.231e+04  3.317e+03   6.726  2.55e-11 ***
## SaleConditionNormal 7.492e+03  2.152e+03   3.481  0.000516 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22410 on 1371 degrees of freedom
## Multiple R-squared:  0.9252, Adjusted R-squared:  0.9204
## F-statistic: 192.7 on 88 and 1371 DF,  p-value: < 2.2e-16
```

The summary suggests a high-performing explanatory model: our multiple R-squared value of 0.9118 can be interpreted as indicating that 91.19% of the variability in SalePrice around its mean is explained by the model, i.e. by the predictor variables that have been included. In order to test this conclusion, we perform some goodness-of-fits tests as follows.

Residual plot:

```
res = resid(model_sig)
options(scipen=999)
plot(workingDF$SalePrice, res, main = "Plot of Residuals vs. Sale Price", xlab = "Sale Price", ylab = "Residuals")
abline(h=c(0,0), col = 'red')
abline(v=350000, col = 'blue')
```

The residual plot shows that for the most part, the residuals are evenly distributed across the  $y = 0$  line. However, we see that as Sale Price increases, the residuals start to deviate homoscedasticity. More specifically, we see this deviation happen at approximately Sale Price = \$350K, which our earlier summary of SalePrice showed to be between the variable's 3rd quartile and maximum. This suggests that for the last quartile of high-selling houses, the fitted regression model may not be as appropriate as it is for the remainder of the house population.

One option is segment the dataset into two: houses with SalePrice  $\leq$  \$350K, and houses with SalePrice  $>$  \$350K.

## Part II: Predictive Modelling