EXPLORATORY DATA ANALYSIS

Project Name: Prediction of Sales Prices of Houses

GitHub link: https://github.com/tnutalapati/prediction-of-sales-prices-of-houses

Team:

Tejaswini Nutalapati: https://github.com/tnutalapati

Aditi Bhargava : https://github.com/aditibhargava14

Loading libraries required and reading the data into R

```
library(corrplot)
library(knitr)
library(ggplot2)
library(gridExtra)
library(scales)
library(Rmisc)
library(ggrepel)
library(rlang)
```

Import the datasets into R

Import dataset->from text->select file from folder->open

Data size and structure

The train dataset consists of character and integer variables. Most of the character variables are actually (ordinal) factors, but we need to read them into R as character strings as most of them require cleaning and/or feature engineering first. In total, there are 81 columns/variables, of which the last one is the response variable (Sale Price).

```
View(train)
str(train)
View(test)
str(test)
```

```
> #data loading
> View(train)
> str(train)
Classes 'tbl_df', 'tbl' and 'data.frame':       1460 obs. of  80 variables:
 $ MSSubClass   : num  60 20 60 70 60 50 20 60 50 190 ...
 $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
 $ LotFrontage  : chr  "65" "80" "68" "60" ...
 $ LotArea      : num  8450 9600 11250 9550 14260 ...
 $ 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  : num  7 6 7 7 8 5 8 7 7 5 ...
 $ OverallCond  : num  5 8 5 5 5 5 5 6 5 6 ...
 $ YearBuilt    : num  2003 1976 2001 1915 2000 ...
 $ YearRemodAdd : num  2003 1976 2002 1970 2000 ...
 $ 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   : chr  "196" "0" "162" "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   : num  706 978 486 216 655 ...
 $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
```

```
#Getting rid of the IDs but keeping the test IDs in a vector.
test_labels <- test$Id
test$Id <- NULL
train$Id <- NULL

test$SalePrice <- NA
all <- rbind(train, test)
dim(all)
```

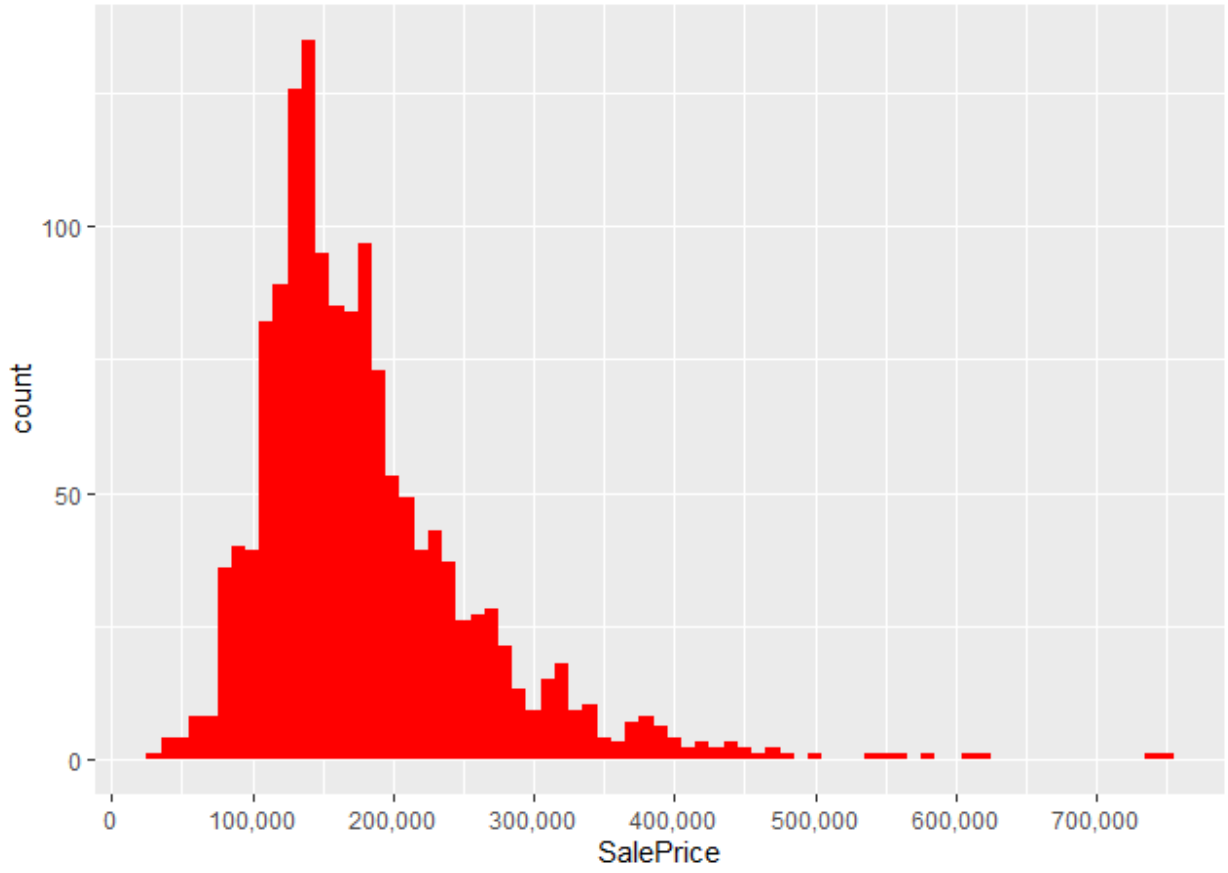
```
[1] 2919   80
>
```

Exploring some of the most important variables

The response variable; SalePrice

As you can see, the sale prices are right skewed. This was expected as few people can afford very expensive houses.

```
#exploring
ggplot(data=all[!is.na(all$SalePrice),], aes(x=SalePrice))
               + geom_histogram(fill="blue", binwidth = 10000)
               + scale_x_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
summary(all$SalePrice)
```

```
> summary(all$SalePrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  34900  129975  163000  180921  214000  755000    1459
>
```

The most important numeric predictors

The character variables need some work before we can use them. To get a feel for the dataset, first we see which numeric variables have a high correlation with the SalePrice.

Correlations with SalePrice

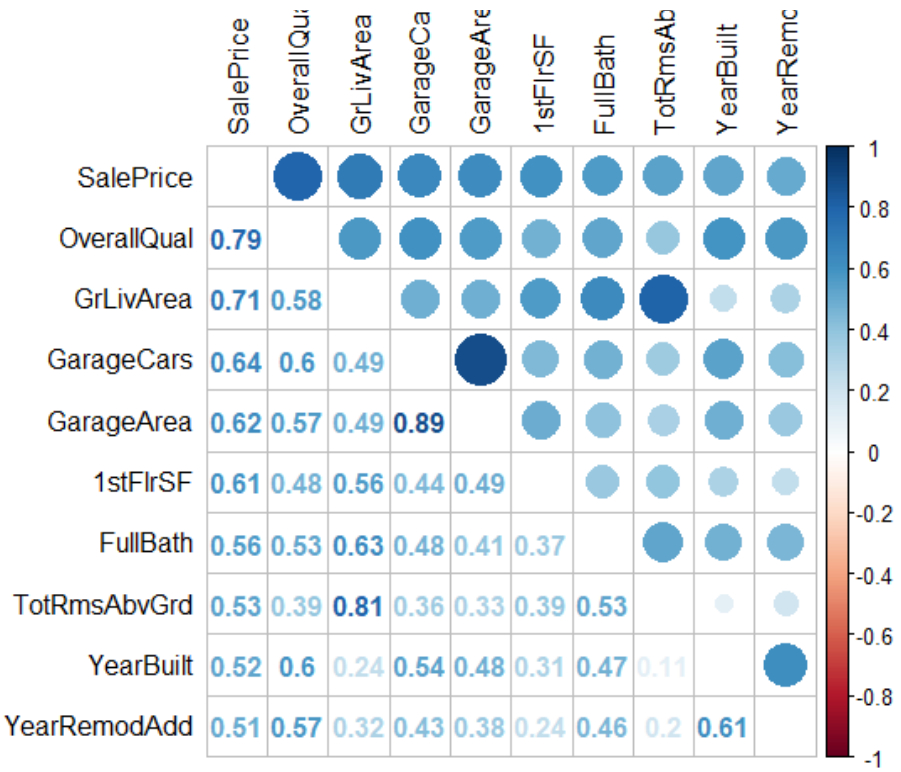
Altogether, there are 10 numeric variables with a correlation of at least 0.5 with SalePrice. All those correlations are positive.

```r
#correlation
numericVars <- which(sapply(all, is.numeric)) #index vector numeric variables
numericVarNames <- names(numericVars) #saving names vector for use later on
cat('There are', length(numericVars), 'numeric variables')
```

```
There are 28 numeric variables
> |
```

```r
all_numVar <- all[, numericVars]
cor_numVar <- cor(all_numVar, use="pairwise.complete.obs") #correlations of all numeric variables
#sort on decreasing correlations with SalePrice
cor_sorted <- as.matrix(sort(cor_numVar[,'SalePrice'], decreasing = TRUE))
#select only high corelations
CorHigh <- names(which(apply(cor_sorted, 1, function(x) abs(x)>0.5)))
cor_numVar <- cor_numVar[CorHigh, CorHigh]

corrplot.mixed(cor_numVar, tl.col="black", tl.pos = "lt")
```



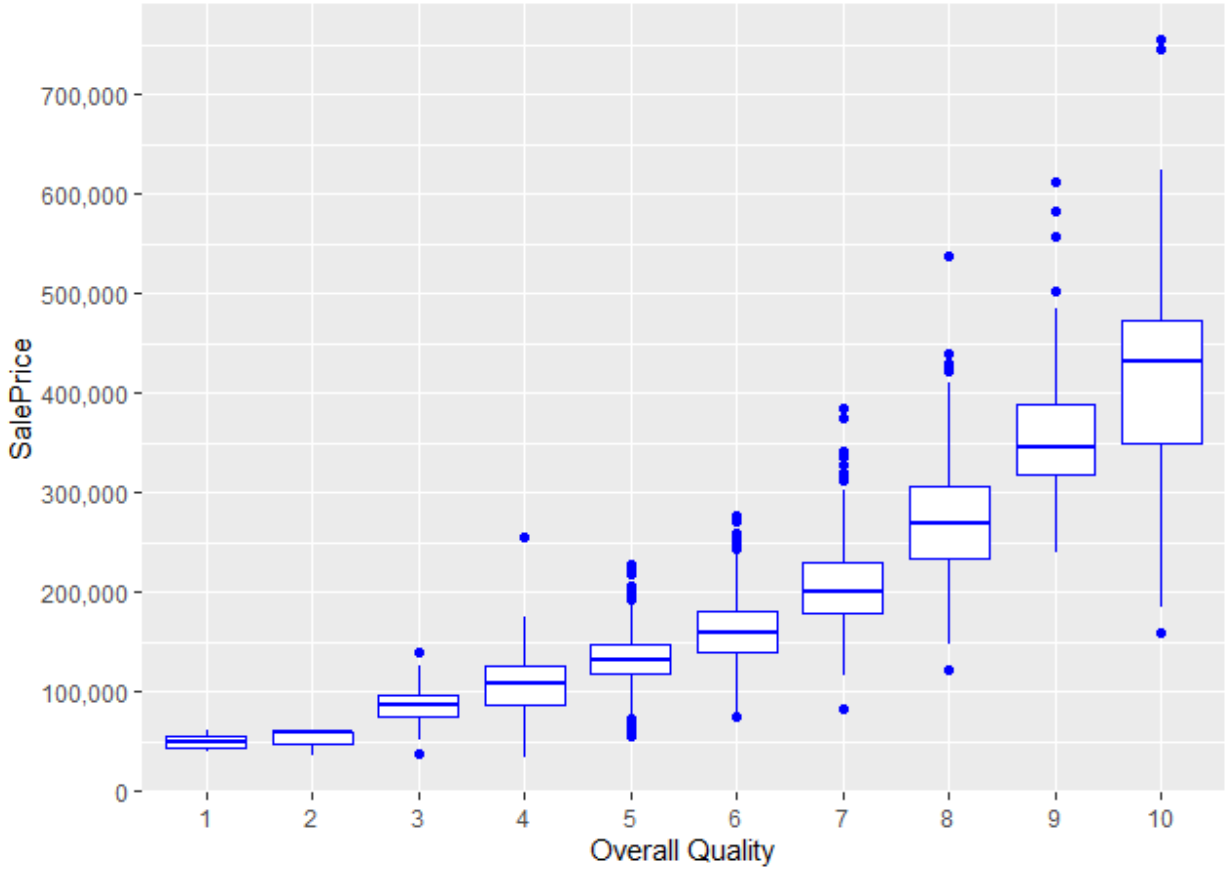
It also becomes clear the multicollinearity is an issue. For example: the correlation between GarageCars and GarageArea is very high (0.89), and both have similar (high) correlations with SalePrice.
The other 6 six variables with a correlation higher than 0.5 with SalePrice are:
TotalBsmtSF: Total square feet of basement area -1stFlrSF

First Floor square feet -FullBath
Full bathrooms above grade -TotRmsAbvGrd
Total rooms above grade (does not include bathrooms) -YearBuilt
Original construction date -YearRemodAdd
Remodel date (same as construction date if no remodeling or additions)

Overall Quality

Overall Quality has the highest correlation with SalePrice among the numeric variables (0.79). It rates the overall material and finish of the house on a scale from 1 (very poor) to 10 (very excellent).

```
#overall quality
ggplot(data=all[!is.na(all$SalePrice),], aes(x=factor(OverallQual), y=SalePrice))
        +geom_boxplot(col='blue') + labs(x='Overall Quality')
        +scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```

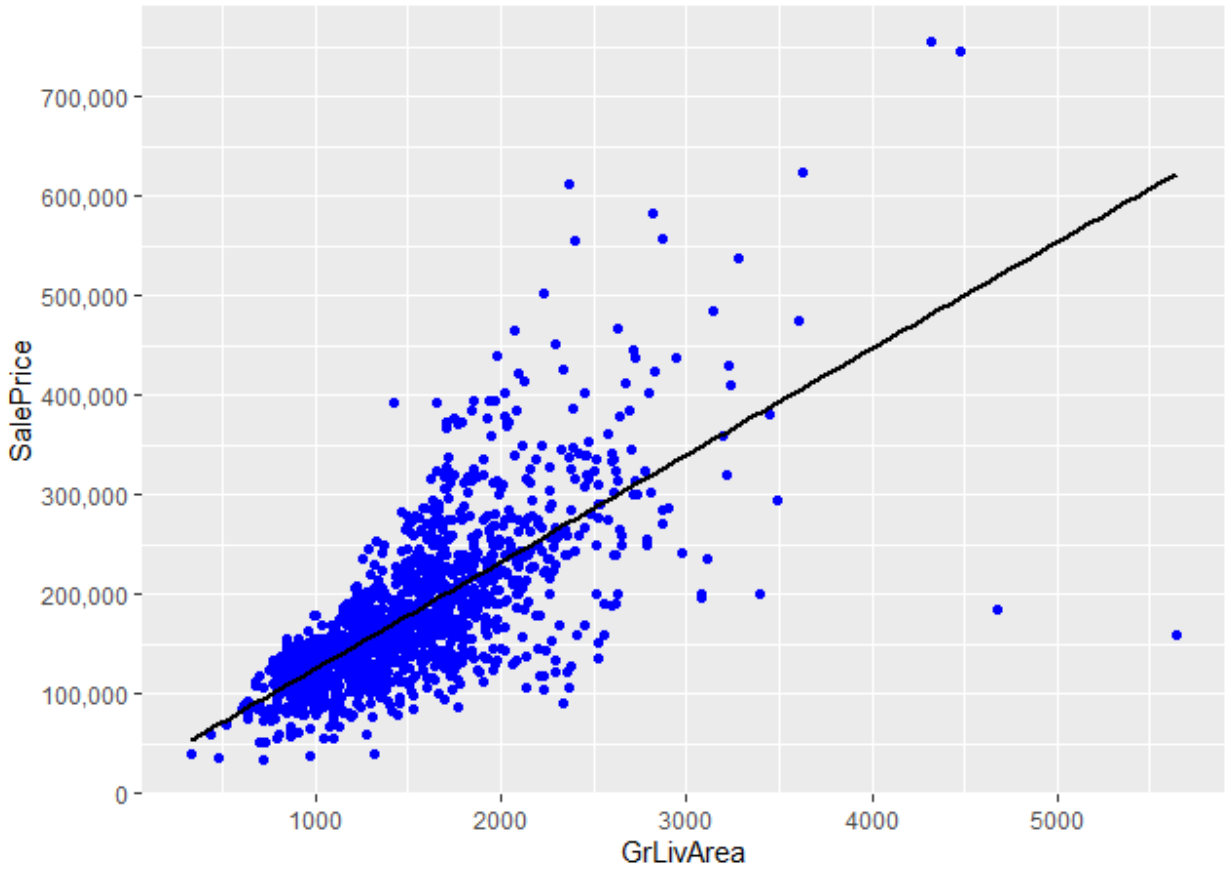


The positive correlation is certainly there indeed and seems to be a slightly upward curve. Regarding outliers, there is no extreme values. If there is a candidate to take out as an outlier later on, it seems to be the expensive house with grade 4.

Above Grade (Ground) Living Area (square feet)

The numeric variable with the second highest correlation with SalesPrice is the Above Grade Living Area. This make a lot of sense; big houses are generally more expensive.

```
#above ground
ggplot(data=all[!is.na(all$SalePrice),], aes(x=GrLivArea, y=SalePrice))
                        + geom_point(col='blue')
                        + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1))
                        + scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
                        + geom_text_repel(aes(label = ifelse(all$GrLivArea[!is.na(all$SalePrice)]>4500, rownames(all), '')))
```



Especially the two houses with really big living areas and low SalePrices seem outliers (houses 524 and 1299, see labels in graph). No need to take them out yet, as taking outliers can be dangerous. For instance, a low score on the Overall Quality could explain a low price.

However, as you can see below, these two houses actually also score maximum points on Overall Quality. Therefore, I will keep houses 1299 and 524 in mind as prime candidates to take out as outliers.

```
all[c(524, 1299), c('SalePrice', 'GrLivArea', 'OverallQual')]
```

```
##      SalePrice GrLivArea OverallQual
## 524     184750      4676          10
## 1299    160000      5642          10
```