Final Project, Data 605, Spring 2018

Mehdi Khan

May 17, 2018

load libraries

```
suppressMessages(suppressWarnings(library(ggplot2)))
suppressMessages(suppressWarnings(library(gridExtra)))
suppressMessages(suppressWarnings(library(scales)))
suppressMessages(suppressWarnings(library(corrplot)))
suppressMessages(suppressWarnings(library(RColorBrewer)))
suppressMessages(suppressWarnings(library(Matrix)))
suppressMessages(suppressWarnings(library(MASS)))
```

Data:

The data was downloaded from https://www.kaggle.com/c/house-prices-advanced-regression-techniques,

```
DF <- read.csv("train.csv", sep = ",", stringsAsFactors = FALSE)
head(DF)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1  1         60       RL          65    8450   Pave  <NA>      Reg
## 2  2         20       RL          80    9600   Pave  <NA>      Reg
## 3  3         60       RL          68   11250   Pave  <NA>      IR1
## 4  4         70       RL          60    9550   Pave  <NA>      IR1
## 5  5         60       RL          84   14260   Pave  <NA>      IR1
## 6  6         50       RL          85   14115   Pave  <NA>      IR1
##   LandContour Utilities LotConfig LandSlope Neighborhood Condition1
## 1         Lvl    AllPub    Inside       Gtl      CollgCr       Norm
## 2         Lvl    AllPub       FR2       Gtl      Veenker      Feedr
## 3         Lvl    AllPub    Inside       Gtl      CollgCr       Norm
## 4         Lvl    AllPub    Corner       Gtl      Crawfor       Norm
## 5         Lvl    AllPub       FR2       Gtl      NoRidge       Norm
## 6         Lvl    AllPub    Inside       Gtl      Mitchel       Norm
##   Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt
## 1       Norm     1Fam     2Story           7           5      2003
## 2       Norm     1Fam     1Story           6           8      1976
## 3       Norm     1Fam     2Story           7           5      2001
## 4       Norm     1Fam     2Story           7           5      1915
## 5       Norm     1Fam     2Story           8           5      2000
## 6       Norm     1Fam     1.5Fin           5           5      1993
##   YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
## 1         2003     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 2         1976     Gable  CompShg     MetalSd     MetalSd       None
## 3         2002     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 4         1970     Gable  CompShg     Wd Sdng     Wd Shng       None
## 5         2000     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 6         1995     Gable  CompShg     VinylSd     VinylSd       None
##   MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
## 1        196        Gd        TA      PConc       Gd       TA           No
```

```
## 2          0        TA        TA     CBlock       Gd       TA           Gd
## 3        162        Gd        TA      PConc       Gd       TA           Mn
## 4          0        TA        TA     BrkTil       TA       Gd           No
## 5        350        Gd        TA      PConc       Gd       TA           Av
## 6          0        TA        TA       Wood       Gd       TA           No
##   BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 1          GLQ        706          Unf          0       150         856
## 2          ALQ        978          Unf          0       284        1262
## 3          GLQ        486          Unf          0       434         920
## 4          ALQ        216          Unf          0       540         756
## 5          GLQ        655          Unf          0       490        1145
## 6          GLQ        732          Unf          0        64         796
##   Heating HeatingQC CentralAir Electrical X1stFlrSF X2ndFlrSF LowQualFinSF
## 1    GasA        Ex          Y      SBrkr       856       854            0
## 2    GasA        Ex          Y      SBrkr      1262         0            0
## 3    GasA        Ex          Y      SBrkr       920       866            0
## 4    GasA        Gd          Y      SBrkr       961       756            0
## 5    GasA        Ex          Y      SBrkr      1145      1053            0
## 6    GasA        Ex          Y      SBrkr       796       566            0
##   GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
## 1      1710            1            0        2        1            3
## 2      1262            0            1        2        0            3
## 3      1786            1            0        2        1            3
## 4      1717            1            0        1        0            3
## 5      2198            1            0        2        1            4
## 6      1362            1            0        1        1            1
##   KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu
## 1            1          Gd            8        Typ          0        &lt;NA&gt;
## 2            1          TA            6        Typ          1          TA
## 3            1          Gd            6        Typ          1          TA
## 4            1          Gd            7        Typ          1          Gd
## 5            1          Gd            9        Typ          1          TA
## 6            1          TA            5        Typ          0        &lt;NA&gt;
##   GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual
## 1     Attchd        2003          RFn          2        548         TA
## 2     Attchd        1976          RFn          2        460         TA
## 3     Attchd        2001          RFn          2        608         TA
## 4     Detchd        1998          Unf          3        642         TA
## 5     Attchd        2000          RFn          3        836         TA
## 6     Attchd        1993          Unf          2        480         TA
##   GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch
## 1         TA          Y          0          61             0          0
## 2         TA          Y        298           0             0          0
## 3         TA          Y          0          42             0          0
## 4         TA          Y          0          35           272          0
## 5         TA          Y        192          84             0          0
## 6         TA          Y         40          30             0        320
##   ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold
## 1           0        0   &lt;NA&gt;  &lt;NA&gt;        &lt;NA&gt;       0      2   2008
## 2           0        0   &lt;NA&gt;  &lt;NA&gt;        &lt;NA&gt;       0      5   2007
## 3           0        0   &lt;NA&gt;  &lt;NA&gt;        &lt;NA&gt;       0      9   2008
## 4           0        0   &lt;NA&gt;  &lt;NA&gt;        &lt;NA&gt;       0      2   2006
## 5           0        0   &lt;NA&gt;  &lt;NA&gt;        &lt;NA&gt;       0     12   2008
## 6           0        0   &lt;NA&gt; MnPrv        Shed     700     10   2009
```

```
##   SaleType SaleCondition SalePrice
## 1       WD        Normal    208500
## 2       WD        Normal    181500
## 3       WD        Normal    223500
## 4       WD       Abnorml    140000
## 5       WD        Normal    250000
## 6       WD        Normal    143000
```

Pick one of the quantitative independent variables from the training data set (train.csv) , and define that variable as X. Make sure this variable is skewed to the right! Pick the dependent variable and define it as Y.

The variable 'GrLivArea' was picked as the independent variable and defined as X and 'SalePrice' was picked as dependent variable and defined as Y

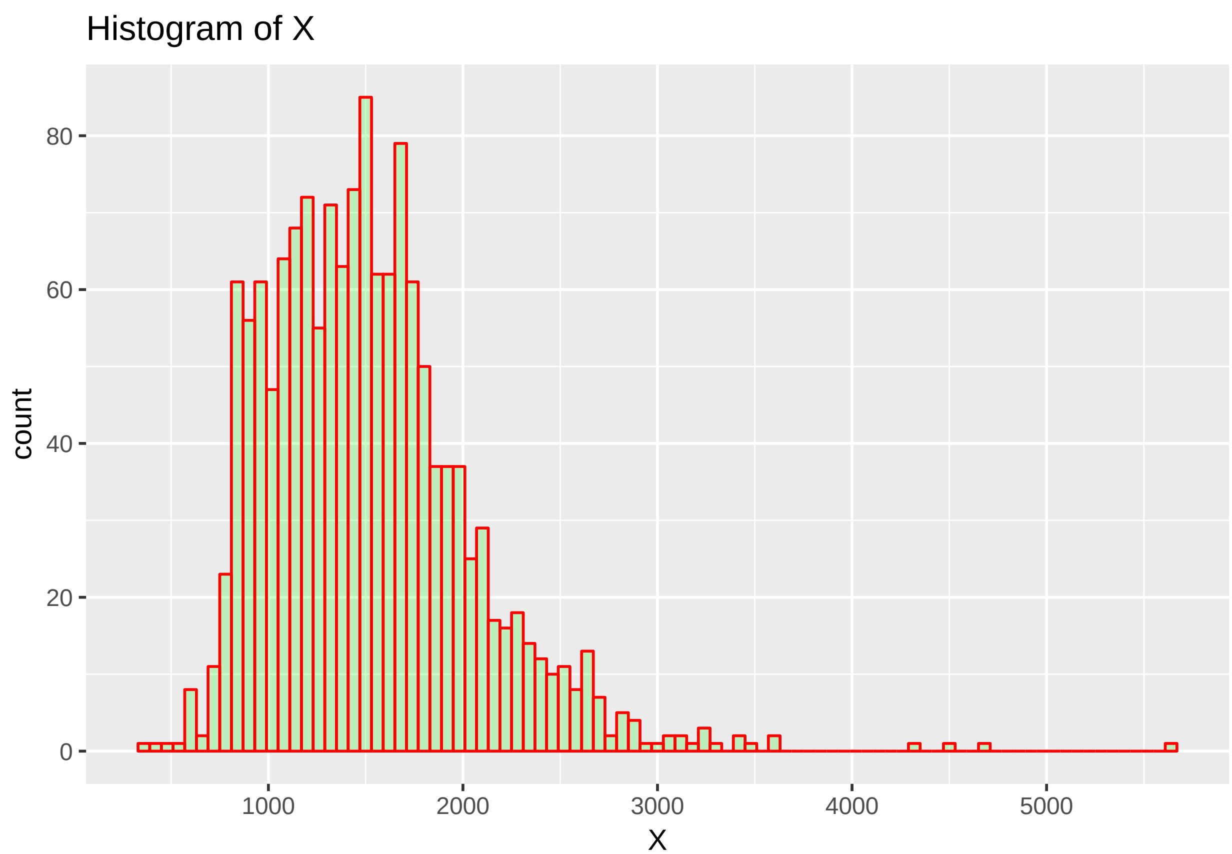
```r
X <- DF["GrLivArea"]
X <- X[!is.na(X)]

Y <- DF["SalePrice"]
Y <- Y[!is.na(Y)]

# creating a dataframe with X and Y

XYdf <- data.frame(cbind(X, Y))

head(XYdf)
```

```
##      X      Y
## 1 1710 208500
## 2 1262 181500
## 3 1786 223500
## 4 1717 140000
## 5 2198 250000
## 6 1362 143000
```

Check if X variable is right skewed

A histogram of X variable was created to see if the data was skewed to the right.

```r
ggplot(XYdf, aes(XYdf$X)) + geom_histogram(col = "red", fill = "green",
    alpha = 0.2, binwidth = 60) + labs(title = "Histogram of X") +
    labs(x = "X")
```



From the histogram it can be seen that the X variable is right skewed.

Probability:

Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the 1st quartile of the X variable, and the small letter "y" is estimated as the 1st quartile of the Y variable. Interpret the meaning of all probabilities. In addition, make a table of counts.

```
# a. P(X>x | Y>y) b. P(X>x, Y>y) c. P(X<x, | Y>y)
```

get the statistics of the variables:

```
summary(XYdf)
```

```
##        X               Y
##  Min.   : 334   Min.   : 34900
##  1st Qu.:1130   1st Qu.:129975
##  Median :1464   Median :163000
##  Mean   :1515   Mean   :180921
##  3rd Qu.:1777   3rd Qu.:214000
##  Max.   :5642   Max.   :755000
```

The 1st quartile of the X variable = 1130 The 1st quartile of the Y variable = 129975 So, x = 1130 and y = 129975

```
x <- 1130
y <- 129975
```

we know P(A|B) = P(A and B)/P(B), by substituting X>x and Y>y for A and B, we get

P(X>x|Y>y) = P(X>x and Y>y)/P(Y>y)

```
Prob_A1_and_B1 <- nrow(subset(XYdf, X > x & Y > y))/nrow(XYdf)
Prob_A1 <- nrow(subset(XYdf, X > x))/nrow(XYdf)
Prob_B1 <- nrow(subset(XYdf, Y > y))/nrow(XYdf)
Prob_C1 <- nrow(subset(XYdf, X < x))/nrow(XYdf)
Prob_C1_and_B1 <- nrow(subset(XYdf, X < x & Y > y))/nrow(XYdf)
```

probability: a

$P(X > x \quad | \quad Y > y)$

```
# a. P(X>x | Y>y)
prob_A1_given_B1 <- Prob_A1_and_B1/Prob_B1
print(prob_A1_given_B1)
```

```
## [1] 0.8712329
```

So P(X>x | Y>y) = .87 or 87%, which means that there is 87% probablity of X>x or Gross living area (GrLivArea) will be bigger than than it 1st quartile value of 1130 given that the Sale price (SalePrice) is bigger than its 1st quartile value of 129975.

probability: b

$P(X > x, \quad Y > y):$

```
# b. P(X>x, Y>y)
print(Prob_A1_and_B1)
```

```
## [1] 0.6534247
```

So P(X>x, Y>y) is 65.34%, which means that there is 65.34% probablity of having X>x or Gross living area (GrLivArea) is bigger than than it's 1st quartile value of 1130 while having the Sale price (SalePrice) bigger than its 1st quartile value of 129975.

probability: c

$P(X < x \quad | \quad Y > y)$

```
### c. P(X<x|Y>y)

prob_C1_given_B1 <- Prob_C1_and_B1/Prob_B1
print(prob_C1_given_B1)
```

```
## [1] 0.1287671
```

The result for c is .1287671 or 12.88%, which means that there is 12.88% probablity of X less than x or Gross living area (GrLivArea) will be smaller than than it 1st quartile value of 1130 given that the Sale price (SalePrice) is bigger than its 1st quartile value of 129975.

Table of counts

```
A1 <- c(sum(X <= x & Y <= y), sum(X > x & Y <= y))
B1 <- c(sum(X <= x & Y > y), sum(X > x & Y > y))
ct_matrix <- matrix(c(A1, B1), nrow = 2)
ct_matrix <- rbind(ct_matrix, apply(ct_matrix, 2, sum))
ct_matrix <- cbind(ct_matrix, apply(ct_matrix, 1, sum))

xy <- c("<=1st quartile", ">1st quartile", "Total")
countDF <- data.frame(xy, ct_matrix)
colnames(countDF) <- c("x/y", "<=1st quartile", ">1st quartile", "Total")
print(countDF)
```

```
##              x/y <=1st quartile >1st quartile Total
## 1 <=1st quartile            225           141   366
## 2  >1st quartile            140           954  1094
## 3          Total            365          1095  1460
```

Does P(AB)=P(A)P(B)?

Let A be the new variable counting those observations above the 1st quartile for X, and let B be the new variable counting those observations above the 1st quartile for Y

```
A <- countDF[2, 4]
B <- countDF[3, 3]
A_B <- countDF[2, 3]
tot <- countDF[3, 4]

Prob_A <- A/tot
Prob_B <- B/tot
prob_A_B <- A_B/tot

print(prob_A_B)
```

```
## [1] 0.6534247
```

So P(AB) = 0.6534247

```
Prob_A_Prob_B <- Prob_A * Prob_B
print(Prob_A_Prob_B)
```

```
## [1] 0.5619863
```

So P(A)P(B) = 0.5625

So, here P(AB) is NOT equal to P(A)P(B). Therefore, variable A and B are not independent and obviously splitting the training data did not make them independent.

Chi Square test

create a matrix from the above observations

```
chiMatrix <- matrix(c(A1, B1), nrow = 2)
chisq.test(chiMatrix)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
```

```
## data:  chiMatrix
## X-squared = 344, df = 1, p-value < 2.2e-16
```

Since the p-value is significantly smaller we can reject the null hypothesis, which agree with the above mathmatical test that the variables are dependent.

Descriptive and Inferential Statistics:

Descriptive statistics:

Subset of data from the train dataset with only numeric columns

```
numcolumns <- unlist(lapply(DF, is.numeric))

numTrain <- DF[, numcolumns]
```

Descriptive statistics of all the numeric columns of train dataset:

```
summary(numTrain)
```

```
##        Id           MSSubClass     LotFrontage       LotArea
##  Min.   :   1.0   Min.   : 20.0   Min.   : 21.00   Min.   :  1300
##  1st Qu.: 365.8   1st Qu.: 20.0   1st Qu.: 59.00   1st Qu.:  7554
##  Median : 730.5   Median : 50.0   Median : 69.00   Median :  9478
##  Mean   : 730.5   Mean   : 56.9   Mean   : 70.05   Mean   : 10517
##  3rd Qu.:1095.2   3rd Qu.: 70.0   3rd Qu.: 80.00   3rd Qu.: 11602
##  Max.   :1460.0   Max.   :190.0   Max.   :313.00   Max.   :215245
##                                   NA's   :259
##   OverallQual      OverallCond      YearBuilt     YearRemodAdd
##  Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950
##  1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967
##  Median : 6.000   Median :5.000   Median :1973   Median :1994
##  Mean   : 6.099   Mean   :5.575   Mean   :1971   Mean   :1985
##  3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004
##  Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010
##
##    MasVnrArea       BsmtFinSF1       BsmtFinSF2        BsmtUnfSF
##  Min.   :   0.0   Min.   :   0.0   Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 223.0
##  Median :   0.0   Median : 383.5   Median :   0.00   Median : 477.5
##  Mean   : 103.7   Mean   : 443.6   Mean   :  46.55   Mean   : 567.2
##  3rd Qu.: 166.0   3rd Qu.: 712.2   3rd Qu.:   0.00   3rd Qu.: 808.0
##  Max.   :1600.0   Max.   :5644.0   Max.   :1474.00   Max.   :2336.0
##  NA's   :8
##   TotalBsmtSF       X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  Min.   :   0.0   Min.   : 334   Min.   :   0   Min.   :  0.000
##  1st Qu.: 795.8   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  Median : 991.5   Median :1087   Median :   0   Median :  0.000
##  Mean   :1057.4   Mean   :1163   Mean   : 347   Mean   :  5.845
##  3rd Qu.:1298.2   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  Max.   :6110.0   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
```

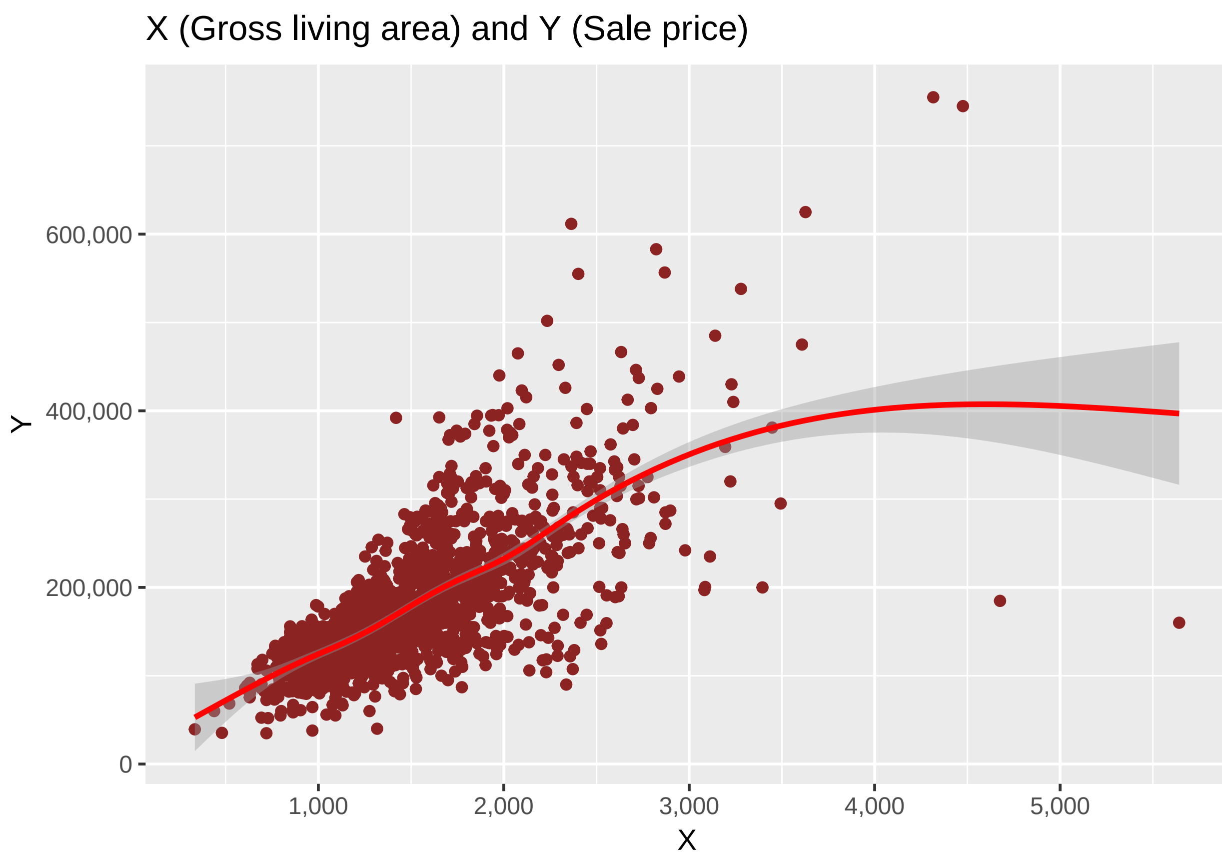
```
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr    TotRmsAbvGrd
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   : 2.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 5.000
##  Median :0.0000   Median :3.000   Median :1.000   Median : 6.000
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   Mean   : 6.518
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.: 7.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000   Max.   :14.000
##
##    Fireplaces     GarageYrBlt     GarageCars      GarageArea
##  Min.   :0.000   Min.   :1900   Min.   :0.000   Min.   :   0.0
##  1st Qu.:0.000   1st Qu.:1961   1st Qu.:1.000   1st Qu.: 334.5
##  Median :1.000   Median :1980   Median :2.000   Median : 480.0
##  Mean   :0.613   Mean   :1979   Mean   :1.767   Mean   : 473.0
##  3rd Qu.:1.000   3rd Qu.:2002   3rd Qu.:2.000   3rd Qu.: 576.0
##  Max.   :3.000   Max.   :2010   Max.   :4.000   Max.   :1418.0
##                  NA's   :81
##    WoodDeckSF      OpenPorchSF     EnclosedPorch     X3SsnPorch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  0.00   Median : 25.00   Median :  0.00   Median :  0.00
##  Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41
##  3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00
##  Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00
##
##   ScreenPorch        PoolArea          MiscVal             MoSold
##  Min.   :  0.00   Min.   :  0.000   Min.   :    0.00   Min.   : 1.000
##  1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:    0.00   1st Qu.: 5.000
##  Median :  0.00   Median :  0.000   Median :    0.00   Median : 6.000
##  Mean   : 15.06   Mean   :  2.759   Mean   :   43.49   Mean   : 6.322
##  3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:    0.00   3rd Qu.: 8.000
##  Max.   :480.00   Max.   :738.000   Max.   :15500.00   Max.   :12.000
##
##      YrSold       SalePrice
##  Min.   :2006   Min.   : 34900
##  1st Qu.:2007   1st Qu.:129975
##  Median :2008   Median :163000
##  Mean   :2008   Mean   :180921
##  3rd Qu.:2009   3rd Qu.:214000
##  Max.   :2010   Max.   :755000
##
```

3 Visualization of data

Scatterplot of X and Y.

```
ggplot(XYdf, aes(X, Y)) + geom_point(color = "brown4") + geom_smooth(method = "auto",
    col = "red") + ggtitle("X (Gross living area) and Y (Sale price)") +
    xlab("X") + ylab("Y") + scale_x_continuous(labels = comma) + scale_y_continuous(labels = comma)
```

```
## `geom_smooth()` using method = 'gam'
```



The above scatterplot shows a positive linear relationship between X and Y but there are some outliers that forces the relationship line almost horizonatl.

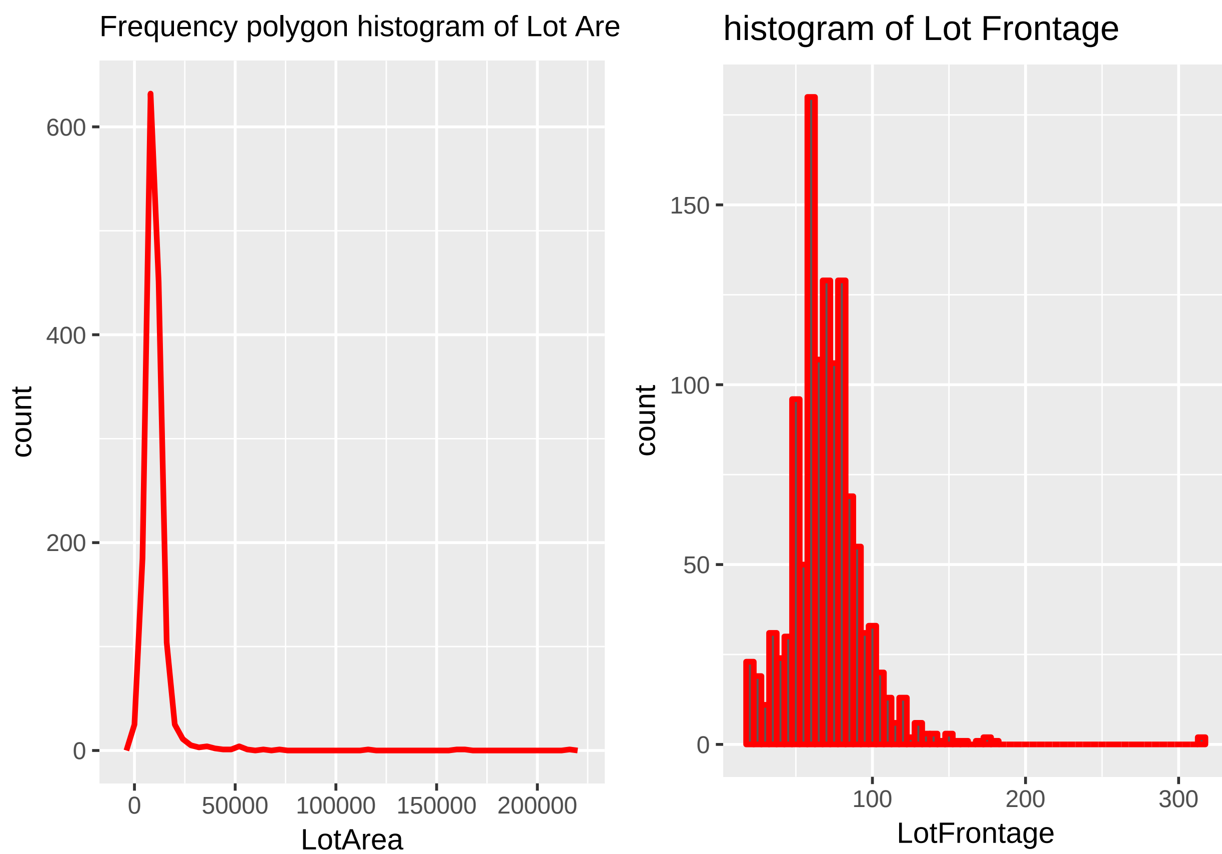
```
ggplot(XYdf[X < 4500, ], aes(X, Y)) + geom_point(color = "brown4") +
    geom_smooth(method = "auto", col = "red") + ggtitle("X (Gross living area) and Y (Sale price)") +
    xlab("X") + ylab("Y") + scale_x_continuous(labels = comma) + scale_y_continuous(labels = comma)
```

```
## `geom_smooth()` using method = 'gam'
```

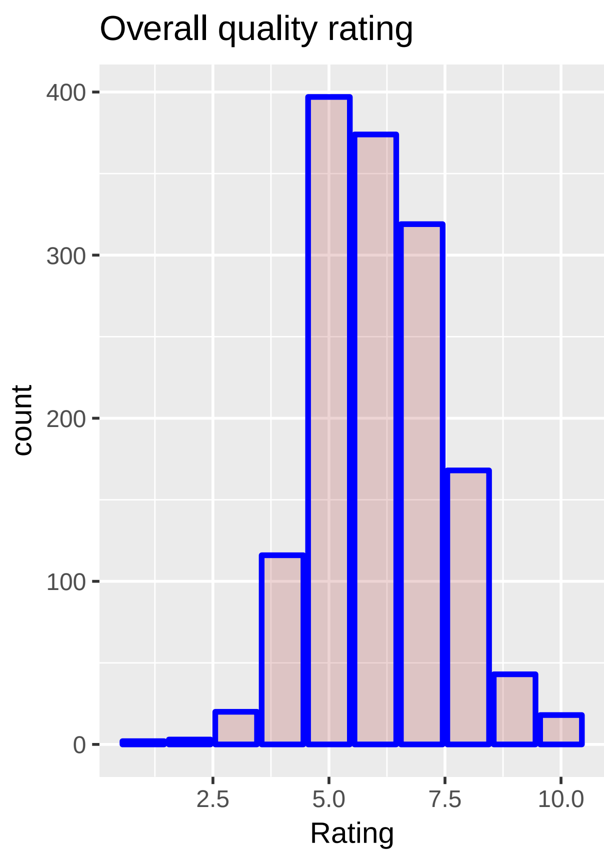
Once the outliers are removed, it does show a strong positive relationship between X and Y.

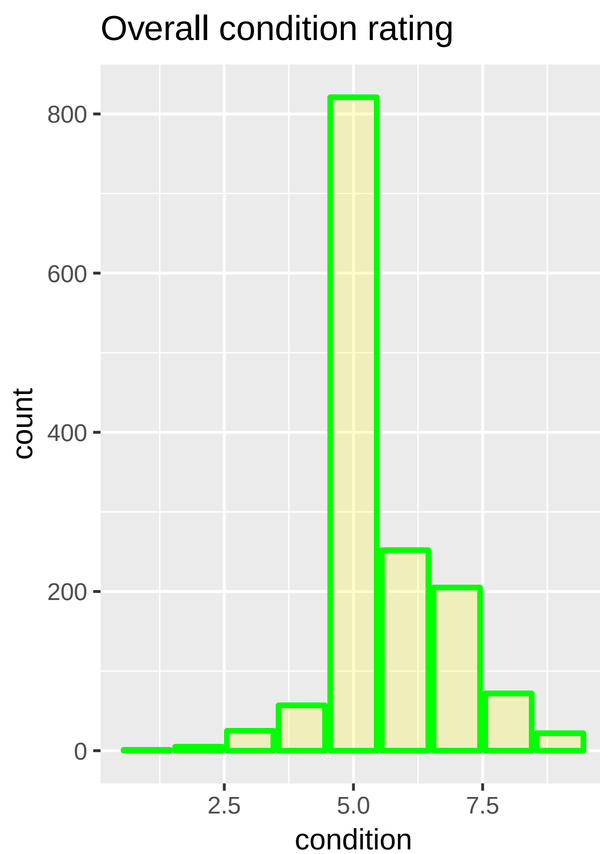
Below are some Plots to visually describe some variables of the dataset:

```
p1 = ggplot(numTrain, aes(LotArea, color = )) + geom_freqpoly(col = "red",
    binwidth = 4000, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "Frequency polygon histogram of Lot Area") + labs(x = "LotArea") +
    theme(plot.title = element_text(size = 11))

p2 = ggplot(numTrain, aes(numTrain$LotFrontage, color = )) + geom_histogram(col = "red",
    binwidth = 5, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "histogram of Lot Frontage") + labs(x = "LotFrontage")

grid.arrange(p1, p2, nrow = 1)
```

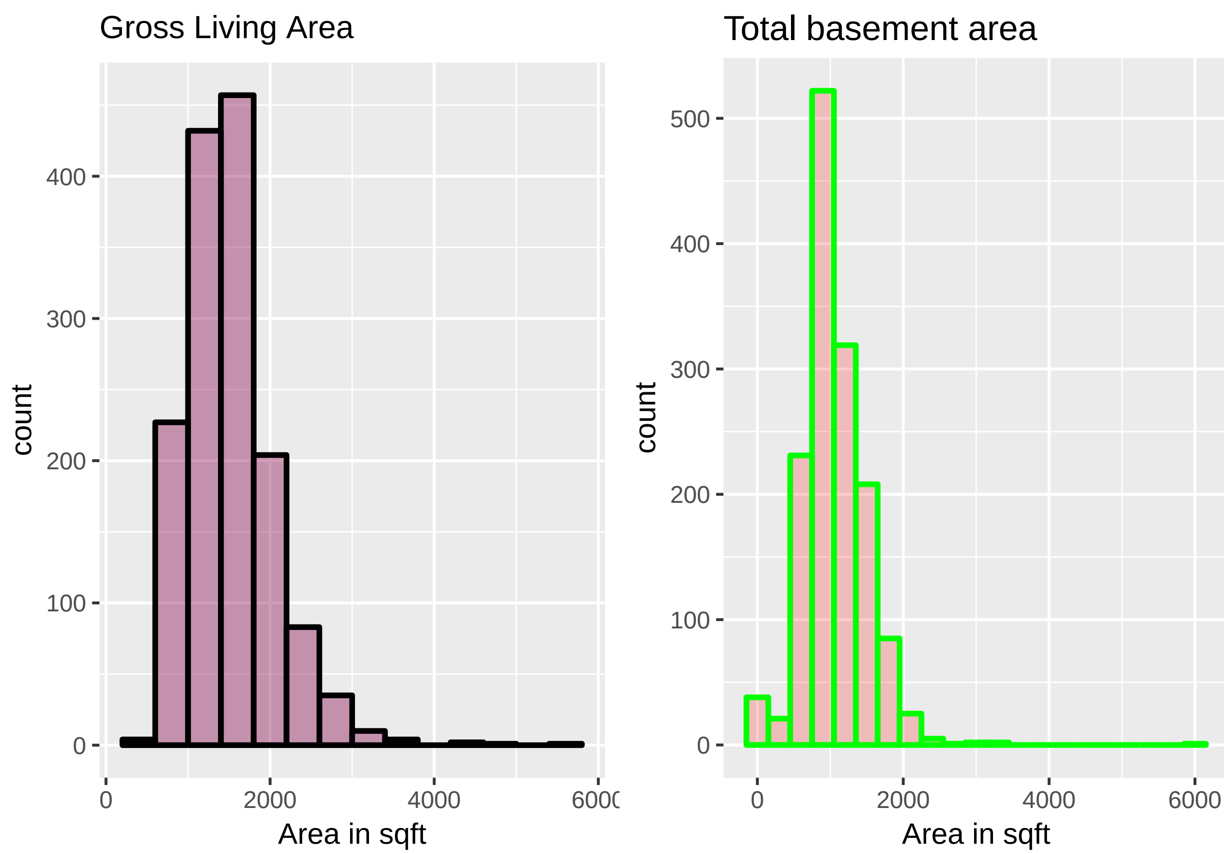



```r
p3 = ggplot(numTrain, aes(numTrain$OverallQual)) + geom_bar(col = "blue",
    fill = "brown", alpha = 0.2, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "Overall quality rating") + labs(x = "Rating")

p4 = ggplot(numTrain, aes(numTrain$OverallCond)) + geom_bar(col = "green",
    fill = "yellow", alpha = 0.2, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "Overall condition rating") + labs(x = "condition")

grid.arrange(p3, p4, nrow = 1)
```

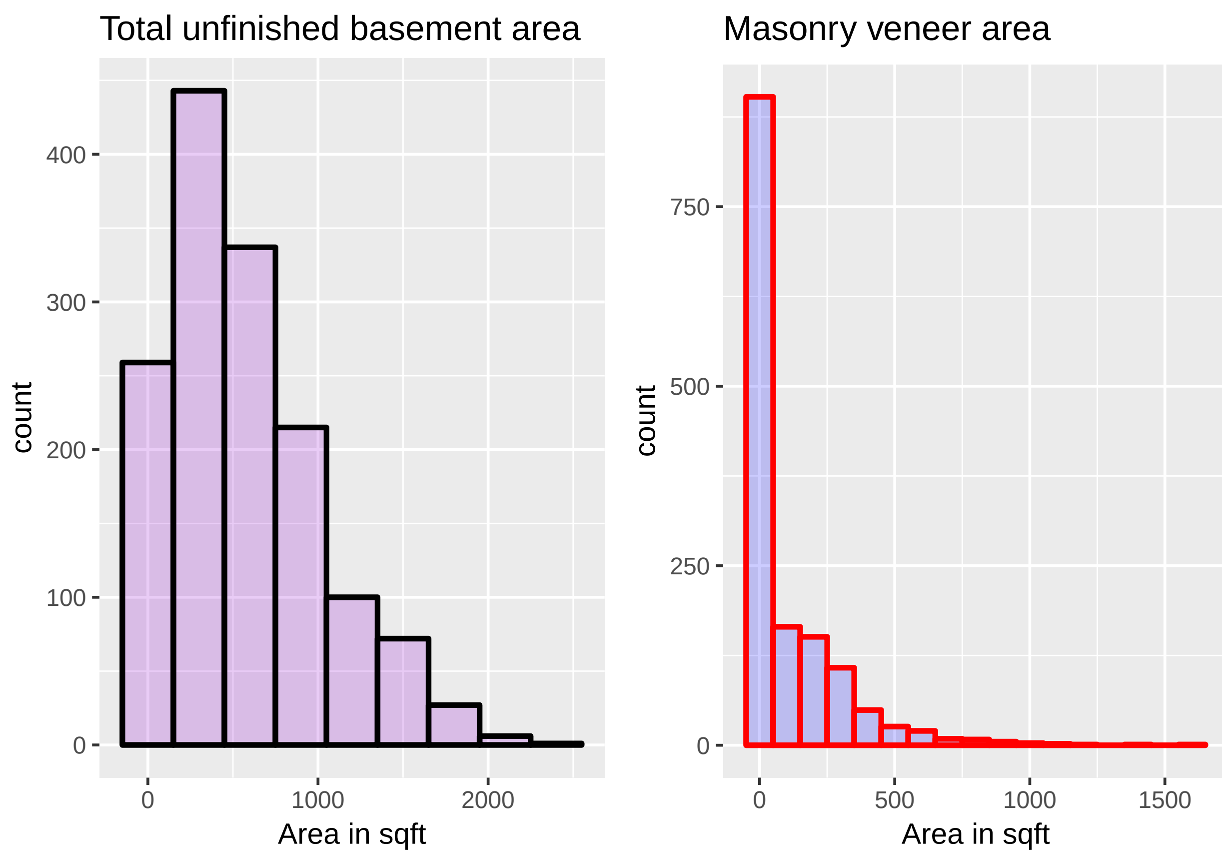




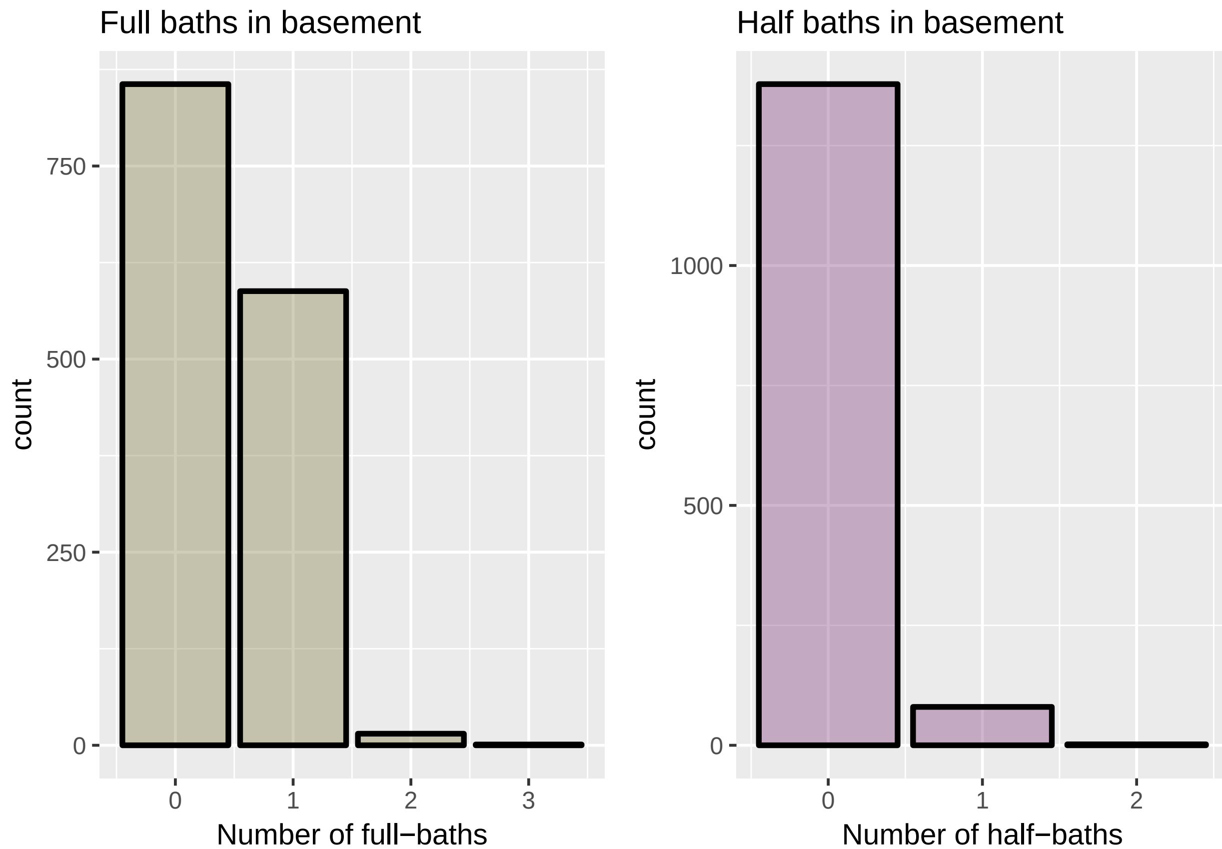

```
p5 = ggplot(numTrain, aes(numTrain$GrLivArea)) + geom_histogram(col = "black",
    binwidth = 400, fill = "deeppink4", alpha = 0.4, lwd = 1, na.rm = TRUE,
    position = "identity") + labs(title = "Gross Living Area") + labs(x = "Area in sqft") +
    theme(plot.title = element_text(size = 12))

p6 = ggplot(numTrain, aes(numTrain$TotalBsmtSF)) + geom_histogram(col = "green",
    binwidth = 300, fill = "red", alpha = 0.2, lwd = 1, na.rm = TRUE,
    position = "identity") + labs(title = "Total basement area") +
    labs(x = "Area in sqft")

grid.arrange(p5, p6, nrow = 1)
```



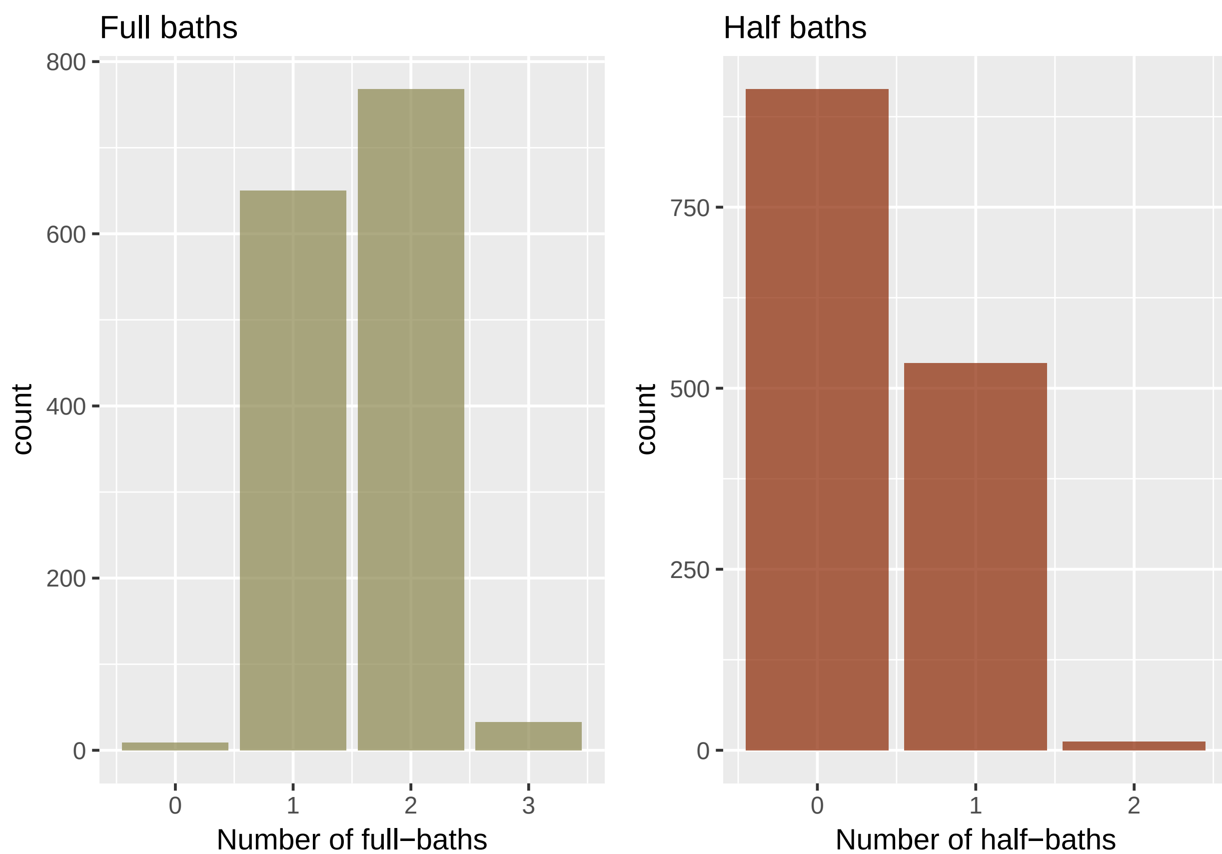

```
p7 = ggplot(numTrain, aes(numTrain$BsmtUnfSF)) + geom_histogram(col = "black",
    binwidth = 300, fill = "darkviolet", alpha = 0.2, lwd = 1, na.rm = TRUE,
    position = "identity") + labs(title = "Total unfinished basement area") +
    labs(x = "Area in sqft")

p8 = ggplot(numTrain, aes(numTrain$MasVnrArea)) + geom_histogram(col = "red",
    binwidth = 100, fill = "blue", alpha = 0.2, lwd = 1, na.rm = TRUE,
    position = "identity") + labs(title = "Masonry veneer area") +
    labs(x = "Area in sqft")

grid.arrange(p7, p8, nrow = 1)
```



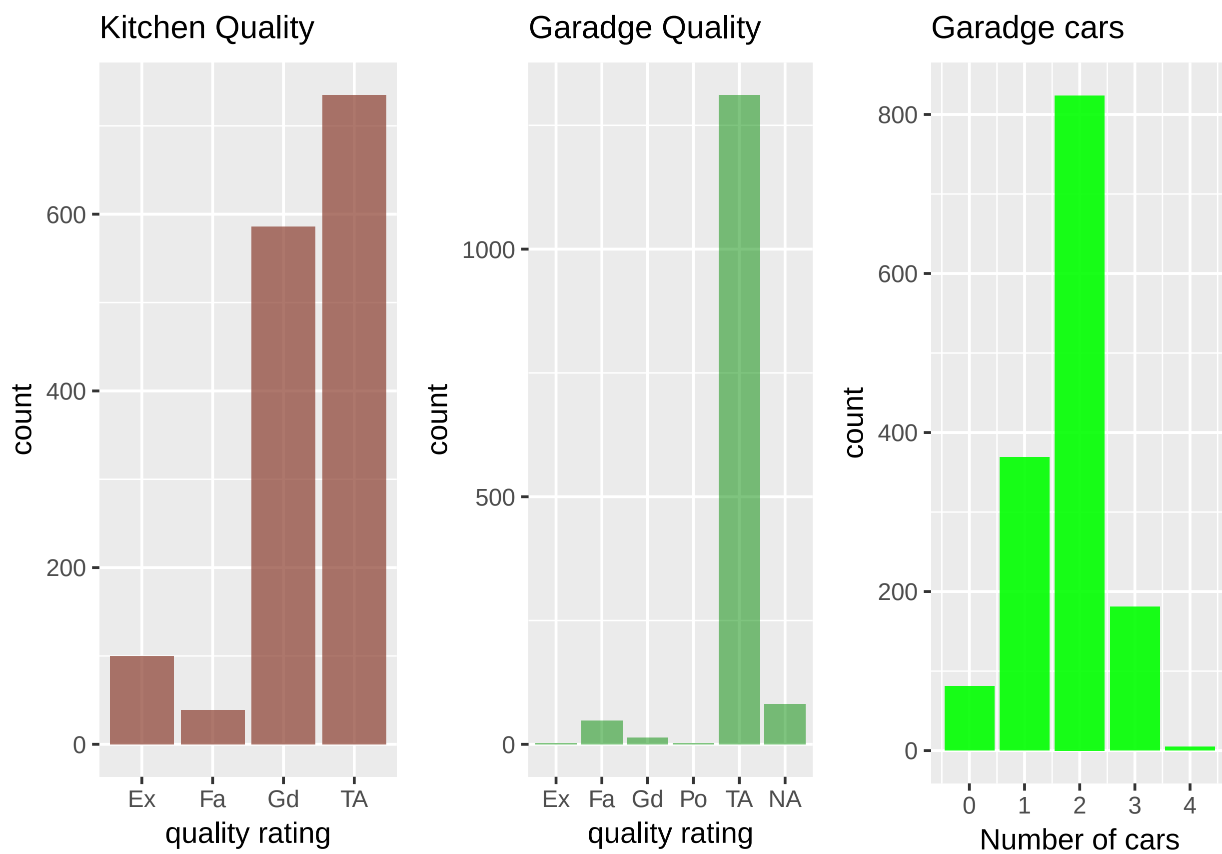

```r
p9 = ggplot(numTrain, aes(numTrain$BsmtFullBath)) + geom_bar(col = "black",
    fill = "khaki4", alpha = 0.4, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "Full baths in basement") + labs(x = "Number of full-baths") +
    theme(plot.title = element_text(size = 12))

p10 = ggplot(numTrain, aes(numTrain$BsmtHalfBath)) + geom_bar(col = "black",
    fill = "orchid4", alpha = 0.4, lwd = 1, na.rm = TRUE, position = "identity") +
    labs(title = "Half baths in basement") + labs(x = "Number of half-baths") +
    theme(plot.title = element_text(size = 12))

grid.arrange(p9, p10, nrow = 1)
```



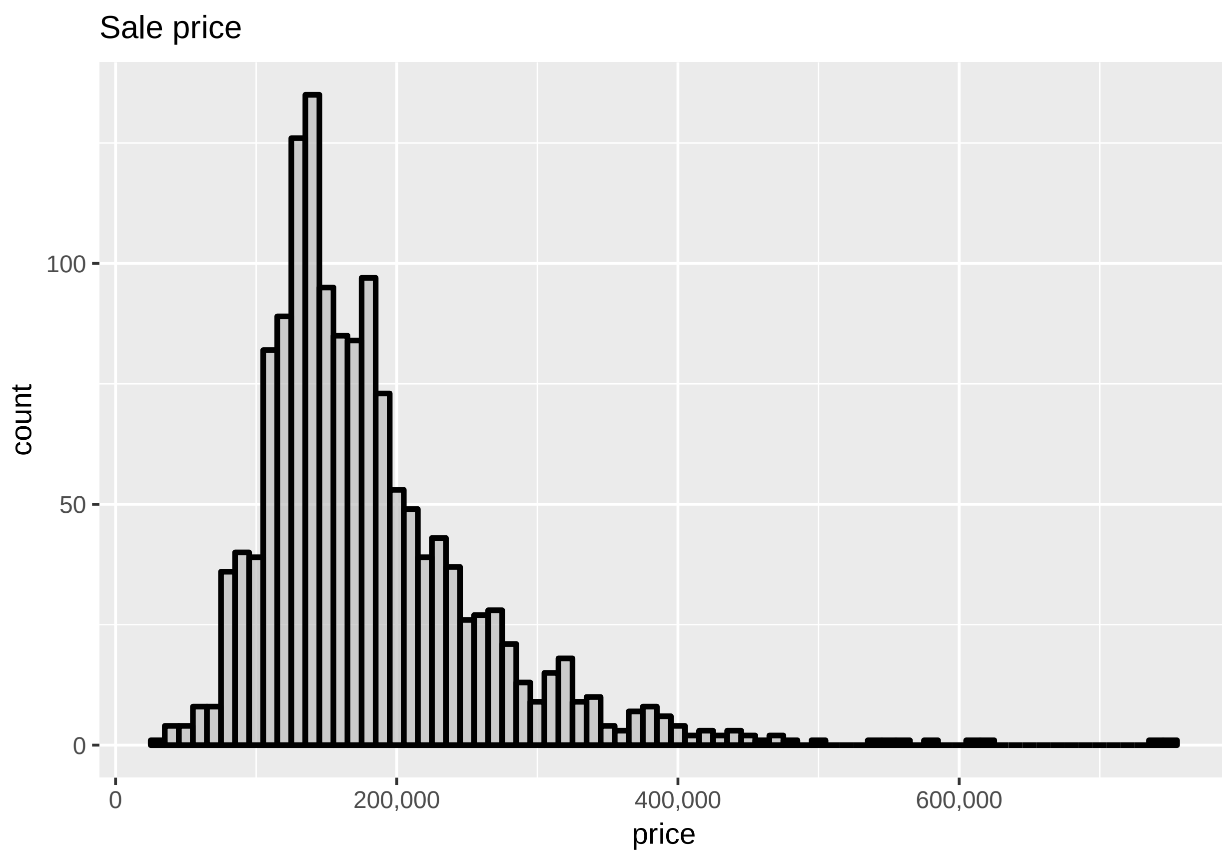

```
p11 = ggplot(numTrain, aes(numTrain$FullBath)) + geom_bar(fill = "khaki4",
    alpha = 0.7, lwd = 1, na.rm = TRUE, position = "identity") + labs(title = "Full baths") +
    labs(x = "Number of full-baths") + theme(plot.title = element_text(size = 12))

p12 = ggplot(numTrain, aes(numTrain$HalfBath)) + geom_bar(fill = "orangered4",
    alpha = 0.7, lwd = 1, na.rm = TRUE, position = "identity") + labs(title = "Half baths ") +
    labs(x = "Number of half-baths") + theme(plot.title = element_text(size = 12))

grid.arrange(p11, p12, nrow = 1)
```




```
p13 = ggplot(DF, aes(DF$KitchenQual)) + geom_bar(fill = "coral4",
    alpha = 0.7, lwd = 1, na.rm = TRUE, position = "identity") + labs(title = "Kitchen Quality") +
    labs(x = "quality rating") + theme(plot.title = element_text(size = 12))

p14 = ggplot(DF, aes(DF$GarageQual)) + geom_bar(fill = "green4", alpha = 0.5,
    lwd = 1, na.rm = TRUE, position = "identity") + labs(title = "Garadge Quality") +
    labs(x = "quality rating") + theme(plot.title = element_text(size = 12))

p15 = ggplot(DF, aes(DF$GarageCars)) + geom_bar(fill = "green", alpha = 0.9,
    lwd = 1, na.rm = TRUE, position = "identity") + labs(title = "Garadge cars") +
    labs(x = "Number of cars") + theme(plot.title = element_text(size = 12))

grid.arrange(p13, p14, p15, nrow = 1)
```




```
ggplot(DF, aes(DF$SalePrice)) + geom_histogram(col = "black", fill = "grey",
    alpha = 0.7, lwd = 1, na.rm = TRUE, position = "identity", binwidth = 10000) +
    labs(title = "Sale price") + labs(x = "price") + theme(plot.title = element_text(size = 12)) +
    scale_x_continuous(labels = comma)
```




```
p16 <- ggplot(DF, aes(x = DF$TotalBsmtSF, y = DF$SalePrice)) + geom_point(color = "blue") +
    ggtitle("Basement size vs Sale price") + xlab("basement sqft") +
    ylab("Sale price") + geom_smooth(method = "auto", col = "red") +
    scale_y_continuous(labels = comma)

p17 <- ggplot(DF, aes(x = DF$OverallCond, y = DF$SalePrice)) + geom_point(color = "brown4") +
    ggtitle("Overall condition vs Sale price") + xlab("Quality rating") +
    ylab("Sale price") + scale_x_continuous(labels = comma) + scale_y_continuous(labels = comma)

grid.arrange(p16, p17, nrow = 1)
```

```
## `geom_smooth()` using method = 'gam'
```

The above two plots are interesting. The figure on the left shows the size of basement and the sale price have a positive corelation until the basement size reaches around little more than 3000 sqft, then the price decreases. This probably is caused by one outlier with a very big basement. The second plot on the right depicts that the price reaches highest around the mid point of quality ratings, which correctly suggests that the house quality is one of many factors for a sale price to go high or low.

```
p18 <- ggplot(DF, aes(x = DF$LotArea, y = DF$SalePrice)) + geom_point(color = "blue") +
    ggtitle("Living area vs Sale price") + xlab("Living area") + ylab("Sale price") +
    geom_smooth(method = "auto", col = "red") + scale_y_continuous(labels = comma)

p19 <- ggplot(DF, aes(x = DF$KitchenQual, y = DF$SalePrice)) + geom_point(color = "brown4") +
    ggtitle("kitchen condition vs Sale price") + xlab("kitchen quality rating") +
    ylab("Sale price") + scale_y_continuous(labels = comma)

grid.arrange(p18, p19, nrow = 1)
```

```
## `geom_smooth()` using method = 'gam'
```

The plot 'Lot area vs Sale price' shows a positive corelation between the variables, although the slope of the corelation line abruptly changes reaffirming some outliers. The second plot on the right shows that the really expensive houses have excellent kitchens but mid priced to low priced houses have kitchens of all quality ratings.

Derive a correlation matrix for any THREE quantitative variables in the dataset

Three selected variables are: SalePrice,TotalBsmtSF,GrLivArea

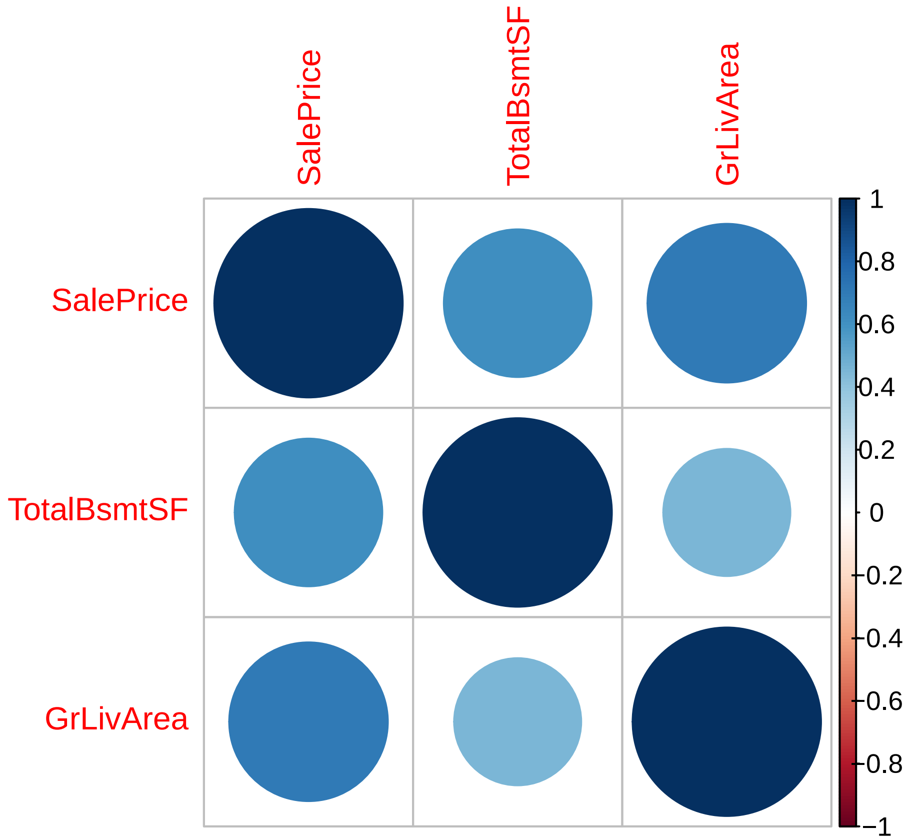
```
corDF <- DF[c("SalePrice", "TotalBsmtSF", "GrLivArea")]
corMatrix <- cor(corDF, use = "complete.obs")
print(corMatrix)
```

```
##             SalePrice TotalBsmtSF GrLivArea
## SalePrice   1.0000000   0.6135806 0.7086245
## TotalBsmtSF 0.6135806   1.0000000 0.4548682
## GrLivArea   0.7086245   0.4548682 1.0000000
```

The above Co-relation matrix suggests that there are strong to moderate corelation exists between these three variables. 'Saleprice' has strong corelations with 'TotalBsmtSF' and 'GrLivArea' with corelation coefficients of .61 and .708 respectively while 'TotalBsmtSF' and 'GrLivArea' have moderate corelation between them with coefficient of .45

Co-relation matrix visualization:

```r
corrplot(corMatrix, method = "circle")
```



Co-relation test bwteen each pair:

Test between 'TotalBsmtSF' and 'SalePrice'

```r
cor.test(DF$TotalBsmtSF, DF$SalePrice, method = "pearson", conf.level = 0.92)
```

```
##
##  Pearson's product-moment correlation
##
## data:  DF$TotalBsmtSF and DF$SalePrice
## t = 29.671, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 92 percent confidence interval:
##  0.5841762 0.6413763
## sample estimates:
##       cor
## 0.6135806
```

Test between 'GrLivArea' and 'SalePrice'

```r
cor.test(DF$GrLivArea, DF$SalePrice, method = "pearson", conf.level = 0.92)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  DF$GrLivArea and DF$SalePrice
## t = 38.348, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 92 percent confidence interval:
##  0.6850407 0.7307245
## sample estimates:
##       cor
## 0.7086245
```

Test between 'GrLivArea' and 'TotalBsmtSF'

```r
cor.test(DF$GrLivArea, DF$TotalBsmtSF, method = "pearson", conf.level = 0.92)
```

```
##
##  Pearson's product-moment correlation
##
## data:  DF$GrLivArea and DF$TotalBsmtSF
## t = 19.503, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 92 percent confidence interval:
##  0.4177447 0.4904754
## sample estimates:
##       cor
## 0.4548682
```

Corelation tests were done above for all three pairs of variables using pearson method, which estimate the association between paired samples and compute a test of the value being zero. Since all three p-values are less than the significance level alpha = 0.08, We can conclude that each pair of those variables are significantly correlated with correlation coefficients showing above.

Would you be worried about familywise error?

Yes, becuse there are many variables in this dataset that might have impact on the corelation of the the pairs of selected variables that are being tested here. Unless all other variables are not considered there is a scope for familywise error which might cause rejecting of true Null hypothesis.

Linear Algebra and Correlation:

Correlation matrix

```r
print(corMatrix)
```

```
##             SalePrice TotalBsmtSF GrLivArea
## SalePrice   1.0000000   0.6135806 0.7086245
## TotalBsmtSF 0.6135806   1.0000000 0.4548682
## GrLivArea   0.7086245   0.4548682 1.0000000
```

precision matrix:

```r
preci_matrix <- solve(corMatrix)
print(preci_matrix)
```

```
##               SalePrice TotalBsmtSF   GrLivArea
## SalePrice    2.5582310 -0.93946422 -1.38549273
## TotalBsmtSF -0.9394642  1.60588442 -0.06473842
## GrLivArea   -1.3854927 -0.06473842  2.01124151
```

Multiplication of correlation matrix by the precision matrix:

```r
round((corMatrix %*% preci_matrix), 2)
```

```
##             SalePrice TotalBsmtSF GrLivArea
## SalePrice           1           0         0
## TotalBsmtSF         0           1         0
## GrLivArea           0           0         1
```

Multiplication of precision matrix by the correlation matrix:

```r
round((preci_matrix %*% corMatrix), 2)
```

```
##             SalePrice TotalBsmtSF GrLivArea
## SalePrice           1           0         0
## TotalBsmtSF         0           1         0
## GrLivArea           0           0         1
```

Both of the above multiplications produce indentity matrix

LU decomposition of corelation matrix:

```r
lud_cor <- lu(corMatrix)
elu_cor <- expand(lud_cor)

cor_L <- elu_cor$L
cor_U <- elu_cor$U
```

lower triangular matrix for corelation matrix:

```r
print(cor_L)
```

```
## 3 x 3 Matrix of class "dtrMatrix" (unitriangular)
##      [,1]       [,2]       [,3]      
## [1,] 1.00000000          .          .
## [2,] 0.61358055 1.00000000          .
## [3,] 0.70862448 0.03218829 1.00000000
```

upper triangular matrix for corelation matrix:

```r
print(cor_U)
```

```
## 3 x 3 Matrix of class "dtrMatrix"
##      [,1]      [,2]      [,3]     
## [1,] 1.0000000 0.6135806 0.7086245
## [2,]         . 0.6235189 0.0200700
## [3,]         .         . 0.4972053
```

LU decomposition of precision matrix:

```
lud_precision <- lu(preci_matrix)
elu_precision <- expand(lud_precision)

precision_L <- elu_precision$L
precision_U <- elu_precision$U
```

lower triangular matrix for precision matrix:

```
print(precision_L)
```

```
## 3 x 3 Matrix of class "dtrMatrix" (unitriangular)
##      [,1]       [,2]       [,3]
## [1,]  1.0000000          .          .
## [2,] -0.3672320  1.0000000          .
## [3,] -0.5415823 -0.4548682  1.0000000
```

upper triangular matrix for precision matrix:

```
print(precision_U)
```

```
## 3 x 3 Matrix of class "dtrMatrix"
##      [,1]       [,2]       [,3]
## [1,]  2.5582310 -0.9394642 -1.3854927
## [2,]          .  1.2608831 -0.5735356
## [3,]          .          .  1.0000000
```

Since A = LU, the abover lower and upper triangular matrices should return the射 origimal matrices after multiplications:

```
cor_L %*% cor_U
```

```
## 3 x 3 Matrix of class "dgeMatrix"
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.6135806 0.7086245
## [2,] 0.6135806 1.0000000 0.4548682
## [3,] 0.7086245 0.4548682 1.0000000
```

```
precision_L %*% precision_U
```

```
## 3 x 3 Matrix of class "dgeMatrix"
##            [,1]        [,2]        [,3]
## [1,]  2.5582310 -0.93946422 -1.38549273
## [2,] -0.9394642  1.60588442 -0.06473842
## [3,] -1.3854927 -0.06473842  2.01124151
```

As expected multiplications of L and U matrices returned their corresponding original matrices.

Calculus-Based Probability & Statistics

Check if shifting is necessary of the X variable that was selected earlier:

```r
min(XYdf$X)
```

```
## [1] 334
```

Since minimum value (334) is above zero, no shifting is necessary.

run fitdistr to fit an exponential probability density function, Find the optimal value of 'lambda' for this distribution

```r
fit_expo <- fitdistr(X, densfun = "exponential")
options(scipen = 999)
print(fit_expo$estimate)
```

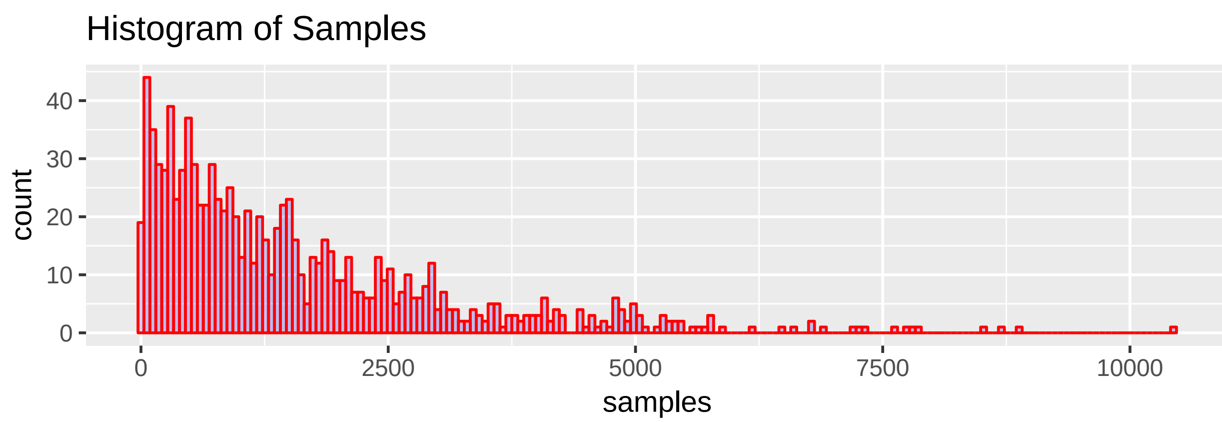
```
##        rate
## 0.000659864
```

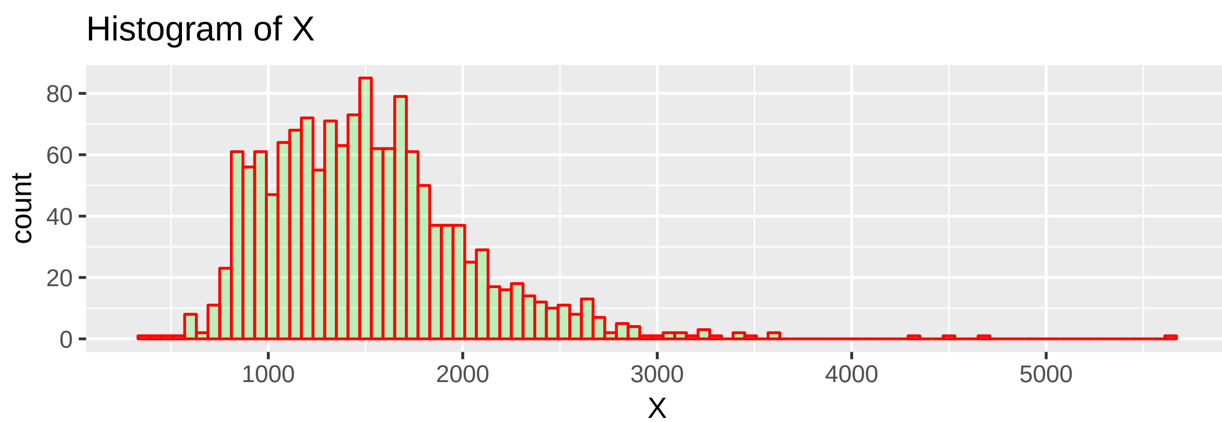
take 1000 samples from this exponential distribution:

```r
samples <- rexp(1000, fit_expo$estimate)
```

Histogram of the samples (simulated data) and the original(observed data), X :

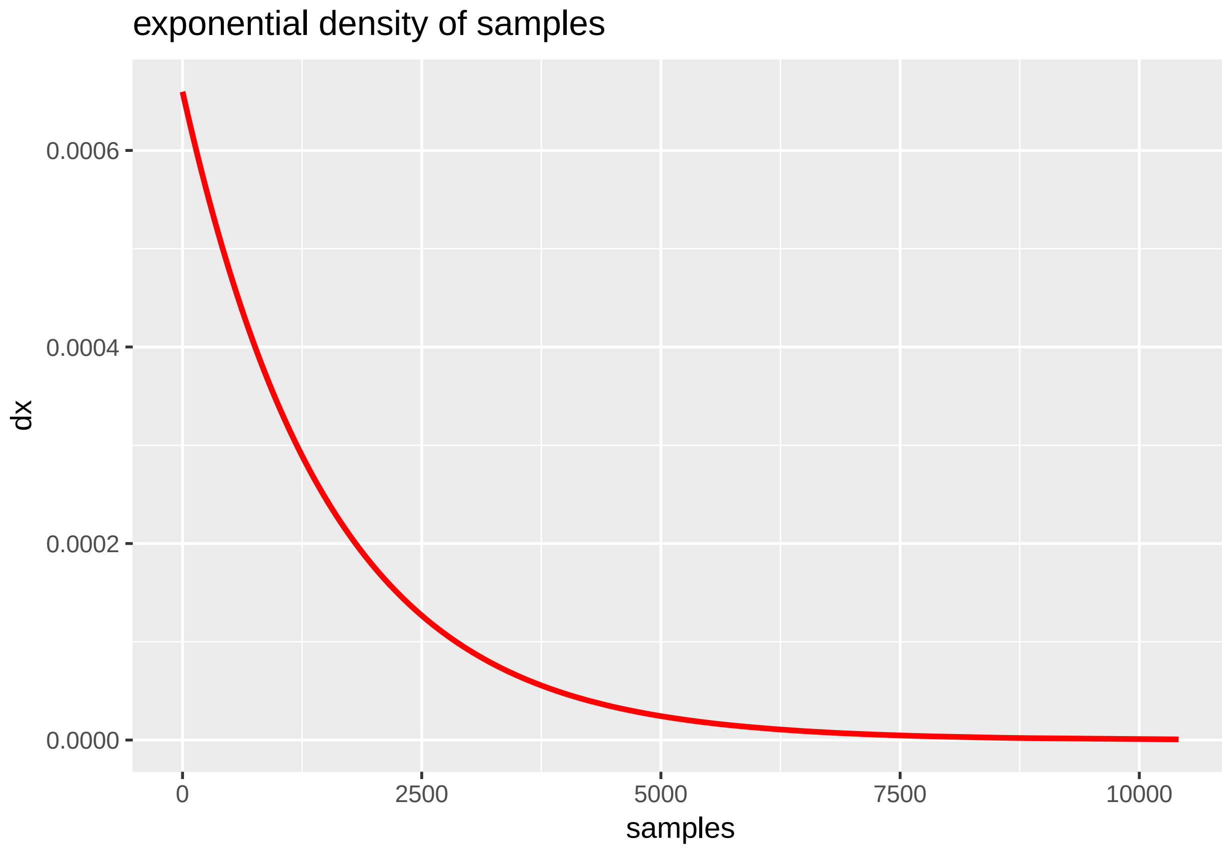
```r
sampldata <- data.frame(samples)


p_samples <- ggplot(sampldata, aes(samples)) + geom_histogram(col = "red",
    fill = "blue", alpha = 0.2, binwidth = 60) + labs(title = "Histogram of Samples") +
    labs(x = "samples")

p_original <- ggplot(XYdf, aes(XYdf$X)) + geom_histogram(col = "red",
    fill = "green", alpha = 0.2, binwidth = 60) + labs(title = "Histogram of X") +
    labs(x = "X")
grid.arrange(p_samples, p_original)
```



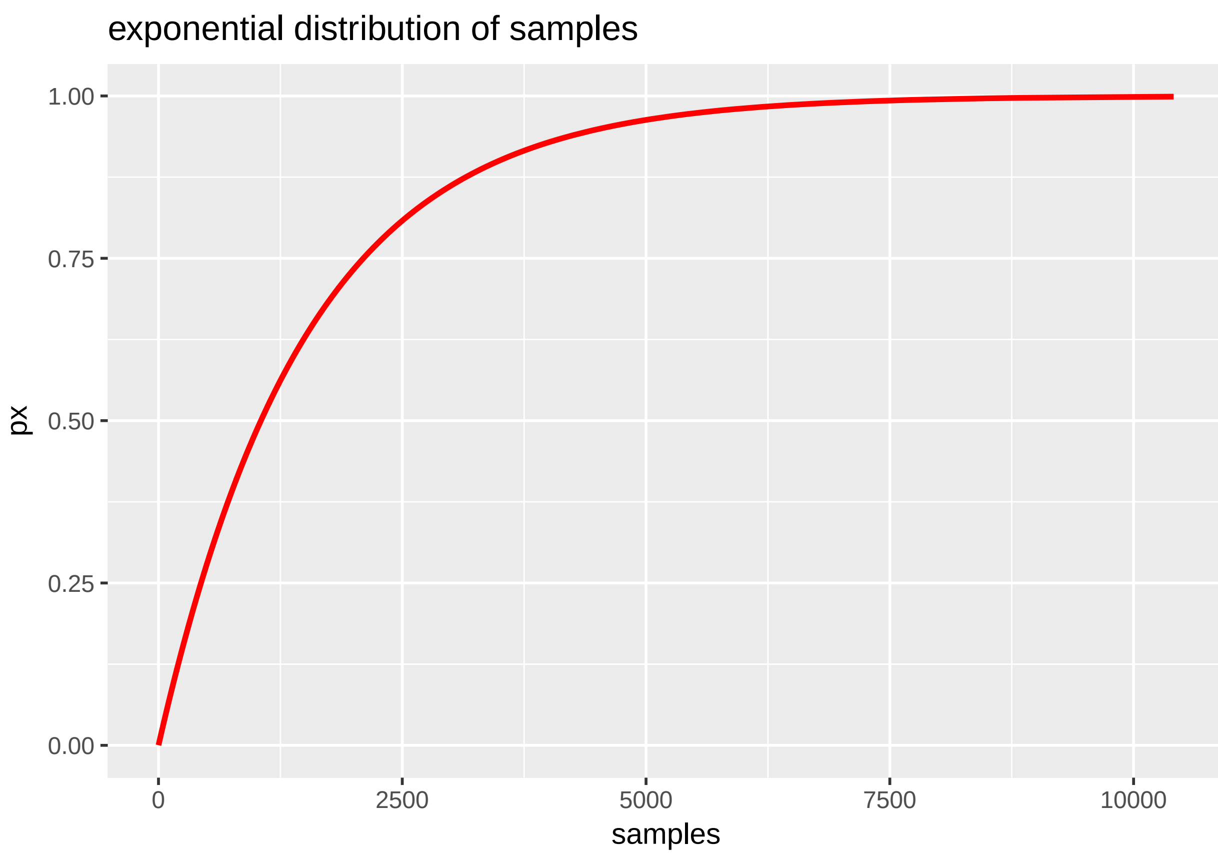


Both of the histograms show similar right skewed pattern but the samples (simulated data) have the highest frequency near zero it is also more skewed than the observed data.

```
dat <- data.frame(samples, dx = dexp(samples, rate = fit_expo$estimate))
ggplot(dat, aes(x = samples, y = dx)) + geom_line(lwd = 1, col = "red") +
    ggtitle("exponential density of samples")
```




```
dat <- data.frame(samples, px = pexp(samples, rate = fit_expo$estimate))
ggplot(dat, aes(x = samples, y = px)) + geom_line(lwd = 1, col = "red") +
    ggtitle("exponential distribution of samples")
```



find the 5th and 95th percentiles of the observed data (X)

```
quantile(XYdf$X, probs = c(0.05, 0.95))
```

```
##     5%    95%
##  848.0 2466.1
```

find the 5th and 95th percentiles of the samples (simulated data)

```
# 5th percentile
qexp(0.05, fit_expo$estimate)
```

```
## [1] 77.73313
```

```
# 95th percentile
qexp(0.95, fit_expo$estimate)
```

```
## [1] 4539.924
```

The 5th and 95th percentiles of the observed data (X) is 848.0 and 2466.1 respectively. The 5th and 95th percentiles of the samples (simulated data) is 77.73313 and 4539.924 respectively.

These differences in percentiles explain why the histograms of these two dataset looked different.

generate a 95% confidence interval from the empirical data, assuming normality:

```r
X_mean <- mean(XYdf$X)
X_std <- sd(XYdf$X)
n <- nrow(XYdf)
se <- qnorm(0.975) * X_std/sqrt(n)
left_interval <- X_mean - se
right_interval <- X_mean + se
left_interval
```

```
## [1] 1488.509
```

```r
right_interval
```

```
## [1] 1542.418
```

SO 95% confidence interval is between 1488.509 and 1542.418

Modeling:

multiple regression model

only a subset of variables were selected by looking at the data that are cleaner and apperently best represent the sale price, following variables were selected.

```r
HouseDF <- DF[, c("LotArea", "Street", "BldgType", "HouseStyle", "OverallQual",
    "OverallCond", "YearBuilt", "YearRemodAdd", "MasVnrType", "ExterQual",
    "BsmtQual", "BsmtCond", "BsmtExposure", "BsmtFinType2", "TotalBsmtSF",
    "HeatingQC", "GrLivArea", "BsmtFullBath", "BsmtHalfBath", "FullBath",
    "HalfBath", "BedroomAbvGr", "KitchenQual", "TotRmsAbvGrd", "GarageArea",
    "PavedDrive", "WoodDeckSF", "OpenPorchSF", "YrSold", "SalePrice")]
```

Remove all 'NA' from the dataset:

```r
HouseDF <- na.omit(HouseDF)
```

generate a regression model

```r
model <- lm(SalePrice ~ ., data = HouseDF)
```

model statistics

```r
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = HouseDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462349  -13478     -99   11700  246442
##
## Coefficients:
##                       Estimate   Std. Error t value          Pr(>|t|)
## (Intercept)       1057682.3682 1358842.0227   0.778          0.436487
## LotArea                 0.4327       0.1006   4.300    0.000018266767249
## StreetPave          24011.1563   15049.2105   1.596          0.110832
## BldgType2fmCon     -13541.3998    6543.1233  -2.070          0.038683
## BldgTypeDuplex     -24228.1645    6160.7239  -3.933    0.000088256630793
## BldgTypeTwnhs      -22525.3073    5580.6729  -4.036    0.000057339237133
```

```
## BldgTypeTwnhsE     -15734.6522    3715.6338  -4.235    0.000024428565099
## HouseStyle1.5Unf    18955.4091    9761.6781   1.942             0.052366
## HouseStyle1Story    17254.7413    3953.9845   4.364    0.000013747435459
## HouseStyle2.5Fin   -28776.9283   12809.8418  -2.246             0.024835
## HouseStyle2.5Unf   -12916.5891   10597.0063  -1.219             0.223098
## HouseStyle2Story    -4924.8119    3750.8554  -1.313             0.189411
## HouseStyleSFoyer     6427.4277    7378.8864   0.871             0.383878
## HouseStyleSLvl      -3377.8242    5520.7393  -0.612             0.540745
## OverallQual         13097.2902    1221.1154  10.726 < 0.0000000000000002
## OverallCond          6018.0505    1034.0256   5.820    0.000000007336228
## YearBuilt             297.4895      69.6958   4.268    0.000021058745033
## YearRemodAdd          -27.1779      70.3954  -0.386             0.699502
## MasVnrTypeBrkFace   15534.5834    8765.8420   1.772             0.076591
## MasVnrTypeNone      14383.9675    8637.0775   1.665             0.096069
## MasVnrTypeStone     18071.2259    9275.6629   1.948             0.051593
## ExterQualFa        -18603.2166   13240.0275  -1.405             0.160229
## ExterQualGd        -16866.7459    6130.0285  -2.751             0.006011
## ExterQualTA        -27036.5873    6790.6466  -3.981    0.000072134572501
## BsmtQualFa         -38817.4665    8092.9519  -4.796    0.000001793647173
## BsmtQualGd         -30677.6561    4212.0288  -7.283    0.000000000000551
## BsmtQualTA         -33221.5955    5150.2973  -6.450    0.000000000154872
## BsmtCondGd           1255.2863    6903.9717   0.182             0.855751
## BsmtCondPo           4015.8708   24861.4763   0.162             0.871700
## BsmtCondTA           6459.6259    5405.6280   1.195             0.232303
## BsmtExposureGd      17841.9832    3813.3235   4.679    0.000003173690952
## BsmtExposureMn       -853.0042    4011.3734  -0.213             0.831635
## BsmtExposureNo      -7826.3857    2849.5206  -2.747             0.006102
## BsmtFinType2BLQ    -10806.7943    9583.5989  -1.128             0.259674
## BsmtFinType2GLQ     -6147.5902   11803.3762  -0.521             0.602568
## BsmtFinType2LwQ     -8317.2602    9125.2029  -0.911             0.362215
## BsmtFinType2Rec     -4121.6952    8887.9456  -0.464             0.642909
## BsmtFinType2Unf     -3392.7542    7797.5501  -0.435             0.663555
## TotalBsmtSF            -7.7892       4.6597  -1.672             0.094830
## HeatingQCFa           -14.3017    5620.2260  -0.003             0.997970
## HeatingQCGd         -3099.0516    2716.2661  -1.141             0.254104
## HeatingQCPo        -28167.9336   33831.0052  -0.833             0.405213
## HeatingQCTA         -2873.9206    2586.4104  -1.111             0.266696
## GrLivArea              59.4186       4.8724  12.195 < 0.0000000000000002
## BsmtFullBath        10835.7826    1971.1601   5.497    0.000000046070301
## BsmtHalfBath         4090.3391    3792.4050   1.079             0.280976
## FullBath             7646.7005    2758.8254   2.772             0.005652
## HalfBath             6777.9417    2612.1113   2.595             0.009566
## BedroomAbvGr        -3662.3620    1710.1682  -2.142             0.032410
## KitchenQualFa      -28916.2275    7919.1711  -3.651             0.000271
## KitchenQualGd      -30221.6000    4515.6785  -6.693    0.000000000032015
## KitchenQualTA      -31030.4954    5082.1425  -6.106    0.000000001333900
## TotRmsAbvGrd         2164.9577    1179.7226   1.835             0.066704
## GarageArea             25.5497       5.6551   4.518    0.000006787546838
## PavedDriveP           610.7726    7299.9203   0.084             0.933332
## PavedDriveY          5617.8075    4394.3639   1.278             0.201323
## WoodDeckSF             14.7707       7.5253   1.963             0.049872
## OpenPorchSF           -25.9672      14.4430  -1.798             0.072414
## YrSold               -801.1455     672.1886  -1.192             0.233530
##
```

```
## (Intercept)
## LotArea            ***
## StreetPave
## BldgType2fmCon     *
## BldgTypeDuplex     ***
## BldgTypeTwnhs      ***
## BldgTypeTwnhsE     ***
## HouseStyle1.5Unf   .
## HouseStyle1Story   ***
## HouseStyle2.5Fin   *
## HouseStyle2.5Unf
## HouseStyle2Story
## HouseStyleSFoyer
## HouseStyleSLvl
## OverallQual        ***
## OverallCond        ***
## YearBuilt          ***
## YearRemodAdd
## MasVnrTypeBrkFace .
## MasVnrTypeNone     .
## MasVnrTypeStone    .
## ExterQualFa
## ExterQualGd        **
## ExterQualTA        ***
## BsmtQualFa         ***
## BsmtQualGd         ***
## BsmtQualTA         ***
## BsmtCondGd
## BsmtCondPo
## BsmtCondTA
## BsmtExposureGd     ***
## BsmtExposureMn
## BsmtExposureNo     **
## BsmtFinType2BLQ
## BsmtFinType2GLQ
## BsmtFinType2LwQ
## BsmtFinType2Rec
## BsmtFinType2Unf
## TotalBsmtSF        .
## HeatingQCFa
## HeatingQCGd
## HeatingQCPo
## HeatingQCTA
## GrLivArea          ***
## BsmtFullBath       ***
## BsmtHalfBath
## FullBath           **
## HalfBath           **
## BedroomAbvGr       *
## KitchenQualFa      ***
## KitchenQualGd      ***
## KitchenQualTA      ***
## TotRmsAbvGrd       .
## GarageArea         ***
```

```
## PavedDriveP
## PavedDriveY
## WoodDeckSF        *
## OpenPorchSF       .
## YrSold
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32370 on 1354 degrees of freedom
## Multiple R-squared:   0.84,  Adjusted R-squared:  0.8331
## F-statistic: 122.5 on 58 and 1354 DF,  p-value: < 0.00000000000000022
```

The Multiple R-squared is 0.84, which is very good, This means 84% variance of the sale price can be explained by predictor variables in the model. F-statistic is 114.8 and p-value is really small. To further improve the model all the variables with p-value greater than .05 will be removed using manual backward selection.

Generate a second model:

```r
model2 <- lm(SalePrice ~ LotArea + BldgType + I(HouseStyle == "1Story") +
    I(HouseStyle == "2.5Fin") + I(BsmtExposure == "Gd") + I(BsmtExposure ==
    "No") + OverallQual + OverallCond + YearBuilt + ExterQual + BsmtQual +
    GrLivArea + BsmtFullBath + FullBath + HalfBath + BedroomAbvGr +
    KitchenQual + TotRmsAbvGrd + GarageArea, data = HouseDF)
```

model statistics

```r
summary(model2)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + BldgType + I(HouseStyle ==
##     "1Story") + I(HouseStyle == "2.5Fin") + I(BsmtExposure ==
##     "Gd") + I(BsmtExposure == "No") + OverallQual + OverallCond +
##     YearBuilt + ExterQual + BsmtQual + GrLivArea + BsmtFullBath +
##     FullBath + HalfBath + BedroomAbvGr + KitchenQual + TotRmsAbvGrd +
##     GarageArea, data = HouseDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -484038  -13201    -314   12037  249468
##
## Coefficients:
##                                 Estimate    Std. Error t value
## (Intercept)                  -652735.64756  115613.28478  -5.646
## LotArea                            0.38208       0.09586   3.986
## BldgType2fmCon                -13402.58510    6458.00947  -2.075
## BldgTypeDuplex                -23775.52864    5741.83924  -4.141
## BldgTypeTwnhs                 -23335.90927    5372.84080  -4.343
## BldgTypeTwnhsE                -16522.70337    3660.82640  -4.513
## I(HouseStyle == "1Story")TRUE   15041.27028    2311.68688   6.507
## I(HouseStyle == "2.5Fin")TRUE  -18464.13232   12173.23723  -1.517
## I(BsmtExposure == "Gd")TRUE     18658.74712    3543.91366   5.265
## I(BsmtExposure == "No")TRUE     -7787.36947    2224.13844  -3.501
## OverallQual                     13039.67124    1184.54045  11.008
## OverallCond                      6347.62821     904.99023   7.014
```

```
## YearBuilt                          345.09475      57.77064   5.974
## ExterQualFa                      -27265.23114   12285.36183  -2.219
## ExterQualGd                      -15510.00104    6058.24663  -2.560
## ExterQualTA                      -25290.69036    6679.38645  -3.786
## BsmtQualFa                       -38805.44672    7753.40612  -5.005
## BsmtQualGd                       -31618.49877    4137.34735  -7.642
## BsmtQualTA                       -34115.22566    5000.48530  -6.822
## GrLivArea                            55.24642       3.94532  14.003
## BsmtFullBath                      10054.88114    1867.86081   5.383
## FullBath                           6331.49812    2689.63453   2.354
## HalfBath                           5256.59446    2352.01875   2.235
## BedroomAbvGr                      -4194.17884    1670.48425  -2.511
## KitchenQualFa                    -28936.60830    7772.65911  -3.723
## KitchenQualGd                    -29636.51510    4467.29726  -6.634
## KitchenQualTA                    -31567.36681    4979.65624  -6.339
## TotRmsAbvGrd                       2181.83011    1155.53422   1.888
## GarageArea                           24.88739       5.52087   4.508
##                                             Pr(>|t|)    
## (Intercept)                       0.0000000199156372 ***
## LotArea                           0.0000707354336654 ***
## BldgType2fmCon                              0.038139 *  
## BldgTypeDuplex                    0.0000367099522610 ***
## BldgTypeTwnhs                     0.0000150581218211 ***
## BldgTypeTwnhsE                    0.0000069213504415 ***
## I(HouseStyle == "1Story")TRUE     0.0000000001071476 ***
## I(HouseStyle == "2.5Fin")TRUE               0.129550    
## I(BsmtExposure == "Gd")TRUE       0.0000001623384859 ***
## I(BsmtExposure == "No")TRUE                 0.000478 ***
## OverallQual                     < 0.0000000000000002 ***
## OverallCond                       0.0000000000036112 ***
## YearBuilt                         0.0000000029476518 ***
## ExterQualFa                                 0.026626 *  
## ExterQualGd                                 0.010568 *  
## ExterQualTA                                 0.000159 ***
## BsmtQualFa                        0.0000006305128657 ***
## BsmtQualGd                        0.0000000000000397 ***
## BsmtQualTA                        0.0000000000133470 ***
## GrLivArea                       < 0.0000000000000002 ***
## BsmtFullBath                      0.0000000859015943 ***
## FullBath                                    0.018710 *  
## HalfBath                                    0.025581 *  
## BedroomAbvGr                                0.012161 *  
## KitchenQualFa                               0.000205 ***
## KitchenQualGd                     0.0000000000467086 ***
## KitchenQualTA                     0.0000000003117401 ***
## TotRmsAbvGrd                                0.059214 .  
## GarageArea                        0.0000071007879346 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 32450 on 1384 degrees of freedom
## Multiple R-squared:  0.8357, Adjusted R-squared:  0.8324 
## F-statistic: 251.4 on 28 and 1384 DF,  p-value: < 0.00000000000000022
```

While manual backward selection did not improve the model based on the R-squared value but the p-value of all of the predictor variables are lower than .05 (except for 'TotRmsAbvGrd' which is close to .05). So any of the models can be used for prediction.

Prediction

```r
testData <- read.csv("test.csv", sep = ",", stringsAsFactors = FALSE)
predictedData_model <- testData
predictedData_model2 <- testData
# modelColumns <- colnames(HouseDF) testDF_model <-
# testData[,colnames(testData) %in% modelColumns]

predictedData_model$salePrice <- predict(model, testData)
predictedData_model2$salePrice <- predict(model2, testData)

Id <- testData$Id
# Kaggle dataset for model1
salePrice <- predictedData_model$salePrice
kaggleData_modelDF <- data.frame(cbind(Id, salePrice))
kaggleData_modelDF[is.na(kaggleData_modelDF)] <- 0
# write.csv(kaggleData_modelDF,'kaggleData_model.csv')

# Kaggle dataset for model2
salePrice <- predictedData_model2$salePrice
kaggleData_modelDF2 <- data.frame(cbind(Id, salePrice))
kaggleData_modelDF2[is.na(kaggleData_modelDF2)] <- 0
# write.csv(kaggleData_modelDF2,'kaggleData_model2.csv')
```

below are two other models created using log transformation. Since the model stats remain almost the same as the above models they were not tested.

```r
numbercolumns <- unlist(lapply(HouseDF, is.numeric))
numDF <- HouseDF[, numbercolumns]
numDF$SalePrice <- NULL
scaledDF <- as.data.frame(log(numDF + 1))
categoryDF <- HouseDF[, !colnames(HouseDF) %in% colnames(scaledDF)]

finalDF <- cbind(categoryDF, scaledDF)

model3 <- lm(SalePrice ~ ., data = finalDF)
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = finalDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -379932  -15173    -657   12525  317215
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       11306380.8 10717062.6   1.055             0.291619
```

```
## StreetPave           17210.2    15364.6   1.120             0.262863
## BldgType2fmCon      -16425.5     6770.3  -2.426             0.015391 *
## BldgTypeDuplex      -30279.1     6300.2  -4.806   0.0000017110049514 ***
## BldgTypeTwnhs        -8158.4     6791.0  -1.201             0.229824
## BldgTypeTwnhsE       -7802.8     4525.7  -1.724             0.084913 .
## HouseStyle1.5Unf     20007.7    10298.1   1.943             0.052242 .
## HouseStyle1Story     10844.7     4181.0   2.594             0.009595 **
## HouseStyle2.5Fin       699.2    12955.1   0.054             0.956967
## HouseStyle2.5Unf     -9785.6    10902.9  -0.898             0.369602
## HouseStyle2Story      4394.8     3850.5   1.141             0.253922
## HouseStyleSFoyer     12604.6     7830.6   1.610             0.107705
## HouseStyleSLvl        -650.9     5762.5  -0.113             0.910079
## MasVnrTypeBrkFace    17725.9     9048.3   1.959             0.050316 .
## MasVnrTypeNone       16079.9     8930.0   1.801             0.071978 .
## MasVnrTypeStone      20239.6     9574.1   2.114             0.034698 *
## ExterQualFa         -17989.1    13721.7  -1.311             0.190083
## ExterQualGd         -25130.8     6309.2  -3.983   0.0000716039974474 ***
## ExterQualTA         -38188.5     6937.1  -5.505   0.0000000441072877 ***
## BsmtQualFa          -40070.7     8390.3  -4.776   0.0000019843703951 ***
## BsmtQualGd          -36314.4     4334.7  -8.378 < 0.0000000000000002 ***
## BsmtQualTA          -38355.3     5294.7  -7.244   0.0000000000007285 ***
## BsmtCondGd            2779.6     7166.6   0.388             0.698187
## BsmtCondPo           19783.1    26253.8   0.754             0.451262
## BsmtCondTA            7262.9     5629.0   1.290             0.197180
## BsmtExposureGd       17887.3     3920.9   4.562   0.0000055261570082 ***
## BsmtExposureMn        -107.0     4164.9  -0.026             0.979508
## BsmtExposureNo       -8192.9     2999.5  -2.731             0.006387 **
## BsmtFinType2BLQ      -8124.1     9937.4  -0.818             0.413772
## BsmtFinType2GLQ      -7110.1    12234.2  -0.581             0.561225
## BsmtFinType2LwQ      -4205.9     9465.3  -0.444             0.656862
## BsmtFinType2Rec      -3012.8     9212.8  -0.327             0.743704
## BsmtFinType2Unf       -288.5     8088.7  -0.036             0.971555
## HeatingQCFa            746.9     5836.8   0.128             0.898190
## HeatingQCGd          -3008.3     2814.5  -1.069             0.285330
## HeatingQCPo         -16962.4    35140.9  -0.483             0.629388
## HeatingQCTA          -1924.0     2681.0  -0.718             0.473096
## KitchenQualFa       -31546.4     8207.9  -3.843             0.000127 ***
## KitchenQualGd       -33185.9     4660.1  -7.121   0.0000000000017305 ***
## KitchenQualTA       -36157.4     5231.4  -6.912   0.0000000000073615 ***
## PavedDriveP          -2707.9     7619.2  -0.355             0.722343
## PavedDriveY           5575.0     4627.1   1.205             0.228462
## LotArea              15541.0     2795.8   5.559   0.0000000327117234 ***
## OverallQual          64752.5     8449.8   7.663   0.0000000000000344 ***
## OverallCond          42618.1     7073.4   6.025   0.0000000021743851 ***
## YearBuilt           441658.7   139319.1   3.170             0.001558 **
## YearRemodAdd         52199.7   143946.5   0.363             0.716936
## TotalBsmtSF          17682.4     4916.8   3.596             0.000334 ***
## GrLivArea            79678.6     8044.8   9.904 < 0.0000000000000002 ***
## BsmtFullBath         15169.5     3023.1   5.018   0.0000005916353624 ***
## BsmtHalfBath          4505.4     5780.0   0.779             0.435828
## FullBath             20979.0     7076.3   2.965             0.003083 **
## HalfBath             12158.1     4048.0   3.004             0.002718 **
## BedroomAbvGr        -20089.0     6157.7  -3.262             0.001132 **
## TotRmsAbvGrd         13070.7     9290.1   1.407             0.159669
```

```
## GarageArea             550.6      761.7   0.723             0.469850
## WoodDeckSF             301.1      387.7   0.777             0.437579
## OpenPorchSF           -748.5      506.3  -1.478             0.139548
## YrSold            -2090084.9  1401007.9  -1.492             0.135973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33580 on 1354 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.8204
## F-statistic: 112.2 on 58 and 1354 DF,  p-value: < 0.00000000000000022
```

```
model4 <- lm(SalePrice ~ LotArea + I(BldgType == "Duplex") + I(HouseStyle ==
    "1Story") + I(BsmtExposure == "Gd") + I(BsmtExposure == "No") +
    OverallQual + OverallCond + YearBuilt + ExterQual + BsmtQual +
    GrLivArea + BsmtFullBath + FullBath + HalfBath + BedroomAbvGr +
    KitchenQual, data = finalDF)
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + I(BldgType == "Duplex") +
##     I(HouseStyle == "1Story") + I(BsmtExposure == "Gd") + I(BsmtExposure ==
##     "No") + OverallQual + OverallCond + YearBuilt + ExterQual +
##     BsmtQual + GrLivArea + BsmtFullBath + FullBath + HalfBath +
##     BedroomAbvGr + KitchenQual, data = finalDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -368048  -15603    -934   12873  320143
##
## Coefficients:
##                                Estimate Std. Error t value
## (Intercept)                    -5979979     865502  -6.909
## LotArea                           19084       2077   9.187
## I(BldgType == "Duplex")TRUE      -21924       5838  -3.755
## I(HouseStyle == "1Story")TRUE     13217       2436   5.425
## I(BsmtExposure == "Gd")TRUE       18269       3633   5.029
## I(BsmtExposure == "No")TRUE       -8896       2329  -3.820
## OverallQual                       70469       8177   8.617
## OverallCond                       42650       6231   6.845
## YearBuilt                        684117     112384   6.087
## ExterQualFa                      -33948      12740  -2.665
## ExterQualGd                      -26542       6272  -4.232
## ExterQualTA                      -39612       6857  -5.777
## BsmtQualFa                       -41703       8049  -5.181
## BsmtQualGd                       -38289       4276  -8.955
## BsmtQualTA                       -39768       5177  -7.682
## GrLivArea                         94182       5564  16.928
## BsmtFullBath                      15131       2856   5.299
## FullBath                          13365       6883   1.942
## HalfBath                           8586       3647   2.354
## BedroomAbvGr                     -14855       5332  -2.786
## KitchenQualFa                    -36894       8054  -4.581
## KitchenQualGd                    -35271       4626  -7.625
```

```
## KitchenQualTA                       -39812       5147  -7.735
##                                              Pr(>|t|)
## (Intercept)                        0.0000000000073961 ***
## LotArea                          < 0.0000000000000002 ***
## I(BldgType == "Duplex")TRUE                   0.00018 ***
## I(HouseStyle == "1Story")TRUE      0.0000000681247929 ***
## I(BsmtExposure == "Gd")TRUE        0.0000005579849243 ***
## I(BsmtExposure == "No")TRUE                   0.00014 ***
## OverallQual                      < 0.0000000000000002 ***
## OverallCond                        0.0000000000114281 ***
## YearBuilt                          0.0000000014826516 ***
## ExterQualFa                                   0.00780 **
## ExterQualGd                        0.0000246689936149 ***
## ExterQualTA                        0.0000000093918540 ***
## BsmtQualFa                         0.0000002527572055 ***
## BsmtQualGd                       < 0.0000000000000002 ***
## BsmtQualTA                         0.0000000000000294 ***
## GrLivArea                        < 0.0000000000000002 ***
## BsmtFullBath                       0.0000001355198441 ***
## FullBath                                      0.05237 .
## HalfBath                                      0.01871 *
## BedroomAbvGr                                  0.00541 **
## KitchenQualFa                      0.0000050442788985 ***
## KitchenQualGd                      0.0000000000000449 ***
## KitchenQualTA                      0.0000000000000197 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33920 on 1390 degrees of freedom
## Multiple R-squared:  0.8197, Adjusted R-squared:  0.8168
## F-statistic: 287.2 on 22 and 1390 DF,  p-value: < 0.00000000000000022
```

Kaggle username: kmehdi2017

Team name: Mehdi Khan

Score for first model: 2.50090

Score for second model:2.15646