DATA605_Final_Project

Irene Jacob

2021-05-23

Problem 1.

Using R, generate a random variable X that has 10,000 random uniform numbers from 1 to N, where N can be any number of your choosing greater than or equal to 6. Then generate a random variable Y that has 10,000 random normal numbers with a mean of $\mu = \sigma = (N + 1)/2$

```
set.seed(10)

N <- 25  #i am choosing N as 25
n <- 10000

X <- runif(n, 1, N) #random uniform numbers from 1 to N(here 25)

m <- (N+1)/2


Y = rnorm(n, mean=m, sd=m) #random normal numbers

df <- data.frame(x = X, y = Y)

head(df)
```

```
##           x         y
## 1 13.179477 26.150110
## 2  8.362444  9.420517
## 3 11.245784 14.767172
## 4 17.634450 25.644654
## 5  3.043263 15.406521
## 6  6.410479  7.858382
```

Probability

Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the median 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.

```
x <- median(X)

y <- quantile(Y, 0.25)

print(paste("median of the X variable",round(x,4),"1st quartile of the Y variable",round(y,4)))
```

```
## [1] "median of the X variable 12.8626 1st quartile of the Y variable 4.2838"
```

```
#a=X>x
#b=X>y
#P(a|b)=P(ab)/P(b)

b<- length(which(X > y))

ab <- length(which(X > x & X > y))

P <- ab/b

print(paste("The probability is: ",round(P,4)))
```

a. $P(X>x \mid X>y)$

```
## [1] "The probability is:  0.5802"
```

```
P = length(which(X > x & Y > y))/n

print(paste("The probability is: ",round(P,4)))
```

b. $P(X>x, Y>y)$

```
## [1] "The probability is:  0.3743"
```

```
#a=X<x
#b=X>y
#P(a|b)=P(ab)/P(b)

b <- length(which(X > y))

ab <- length(which(X < x & X > y))

P <- ab/b

print(paste("The probability is: ",round(P,4)))
```

c. $P(X<x \mid X>y)$

```
## [1] "The probability is:  0.4198"
```

```
i <- length(which(X < x & Y < y))
j <- length(which(X > x & Y < y))
k <- length(which(X < x & Y > y))
l <- length(which(X > x & Y > y))

tab_df <- matrix(c(i,j,k,l),nrow=2)

rownames(tab_df) <- c('X < x','X > x')
colnames(tab_df) <- c('Y < y','Y > y')

tab_df <- as.table(tab_df)
tab_df
```

Investigate whether P(X>x and Y>y)=P(X>x)P(Y>y) by building a table and evaluating the marginal and joint probabilities.

```
##       Y < y Y > y
## X < x  1243  3757
## X > x  1257  3743
```

```
# Marginal
m1 <- margin.table(tab_df,1)[2] / margin.table(tab_df)
m2 <- margin.table(tab_df,2)[2] / margin.table(tab_df)

print(paste("The probability is: ",m1 %*% m2))    #matrix multiplication
```

```
## [1] "The probability is:  0.375"
```

```
# Joint
joint = tab_df[2,2] / margin.table(tab_df)

print(paste("The probability is: ",joint))
```

```
## [1] "The probability is:  0.3743"
```

The marginal and joint probabilities are almost same that could indicate independancy.

```
df_f_c <- table(X > x, Y > y)

df_f_c
```

Check to see if independence holds by using Fisher's Exact Test and the Chi Square Test. What is the difference between the two? Which is most appropriate?

```
##
##         FALSE TRUE
##   FALSE  1243 3757
##   TRUE   1257 3743
```

```r
fisher.test(df_f_c)  #Fisher's Exact Test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df_f_c
## p-value = 0.764
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.898962 1.079670
## sample estimates:
## odds ratio
##  0.9851765
```

```r
chisq.test(df_f_c)   #Chi Square Test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df_f_c
## X-squared = 0.090133, df = 1, p-value = 0.764
```

Chi Square Test is used for large datasets(it assumes that the size of the sample is large). Fisher's Exact Test is used for small samples(when used for large samples the process is very tedious)

Fisher's Exact Test and the Chi Square Test have large p value(>0.05) which indicates that the independence is true.

Problem 2.

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. https://www.kaggle.com/c/house-prices-advanced-regression-techniques . I want you to do the following.

Descriptive and Inferential Statistics

Provide univariate descriptive statistics and appropriate plots for the training data set. Provide a scatterplot matrix for at least two of the independent variables and the dependent variable. Derive a correlation matrix for any three quantitative variables in the dataset. Test the hypotheses that the correlations between each pairwise set of variables is 0 and provide an 80% confidence interval. Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

```r
t <- read.csv("https://raw.githubusercontent.com/irene908/DATA605/main/train.csv")  #train data
head(t)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
```

```
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0       <NA>     Attchd        2003          RFn          2
## 2          1         TA     Attchd        1976          RFn          2
```

```
## 3          1          TA     Attchd        2001          RFn          2
## 4          1          Gd     Detchd        1998          Unf          3
## 5          1          TA     Attchd        2000          RFn          3
## 6          0        <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

```
summary(t)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.099   Mean   :5.575   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1460        Length:1460        Length:1460
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
##  Exterior2nd         MasVnrType          MasVnrArea      ExterQual
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
```

```
##   HeatingQC          CentralAir         Electrical          X1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
```

```
##                                      Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      X3SsnPorch      ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

```
str(t)
```

```
## '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     : chr  "RL" "RL" "RL" "RL" ...
##  $ 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       : 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" ...
```

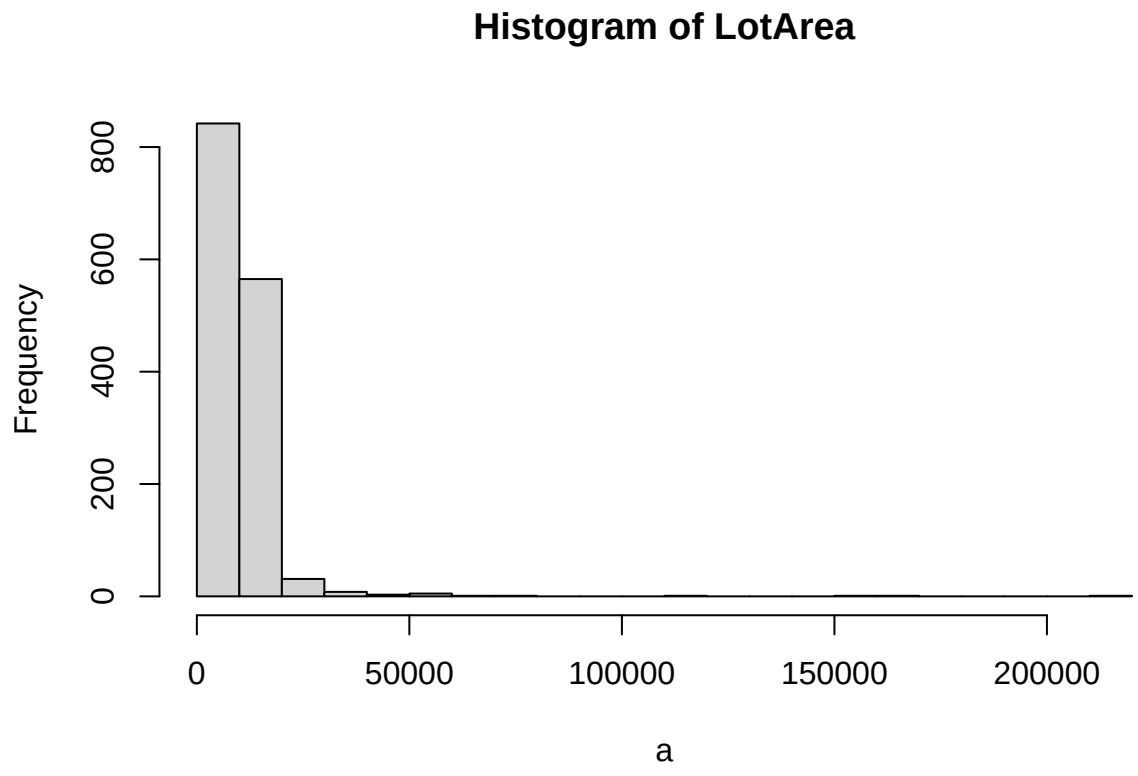
```
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : 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   : int  196 0 162 0 350 0 186 240 0 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   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ 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      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 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  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ 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   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ 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       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA 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     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
a <- t$LotArea
describe(a)
```

```
##    vars    n     mean      sd median trimmed     mad  min    max  range  skew
## X1    1 1460 10516.83 9981.26 9478.5 9563.28 2962.23 1300 215245 213945 12.18
##    kurtosis     se
## X1   202.26 261.22
```

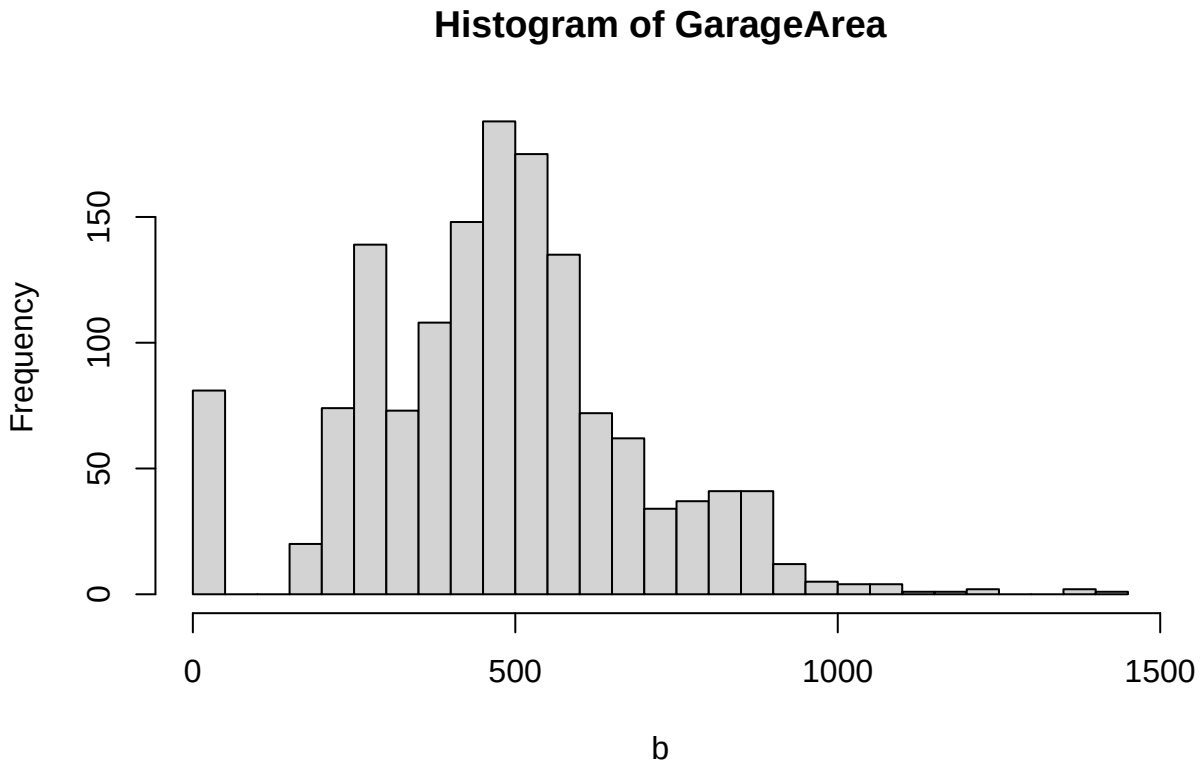
```
hist(a, breaks=30, main = "Histogram of LotArea")
```




```
b <- t$GarageArea
describe(b)
```

```
##     vars    n   mean    sd median trimmed    mad min  max range skew kurtosis
## X1     1 1460 472.98 213.8    480  469.81 177.91   0 1418  1418 0.18      0.9
##      se
## X1 5.6
```

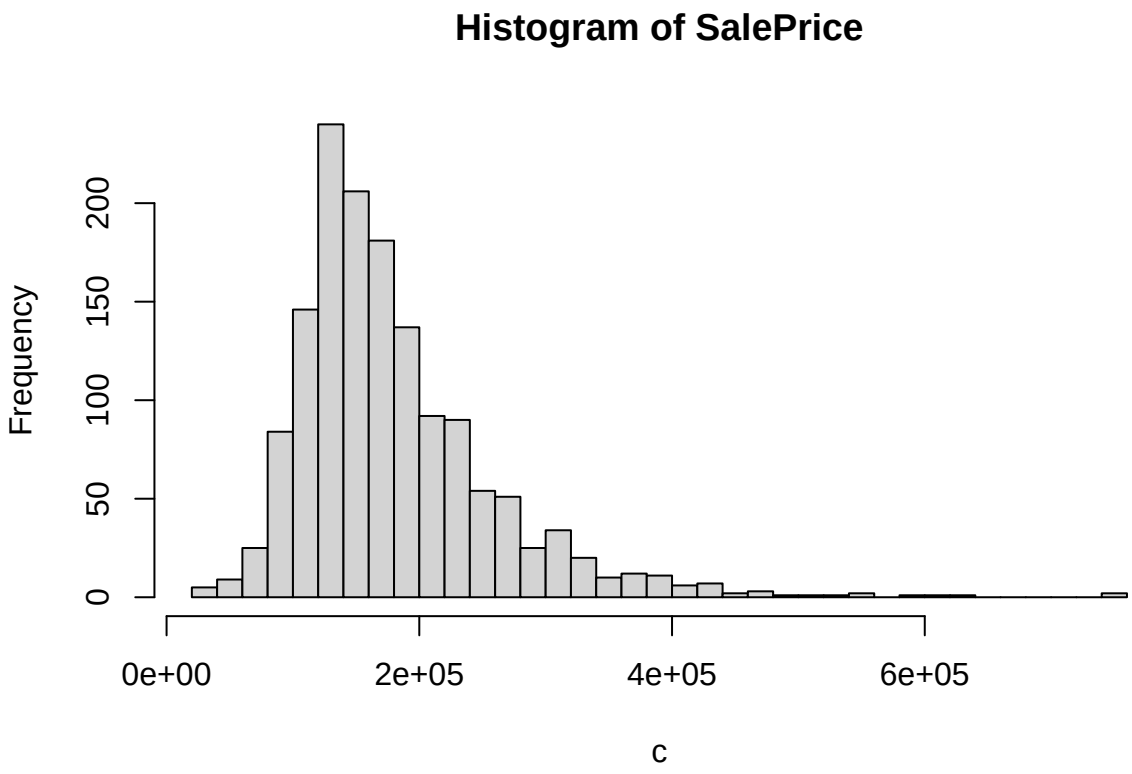
```
hist(b, breaks=30, main = "Histogram of GarageArea")
```




```
c <- t$SalePrice
describe(c)
```

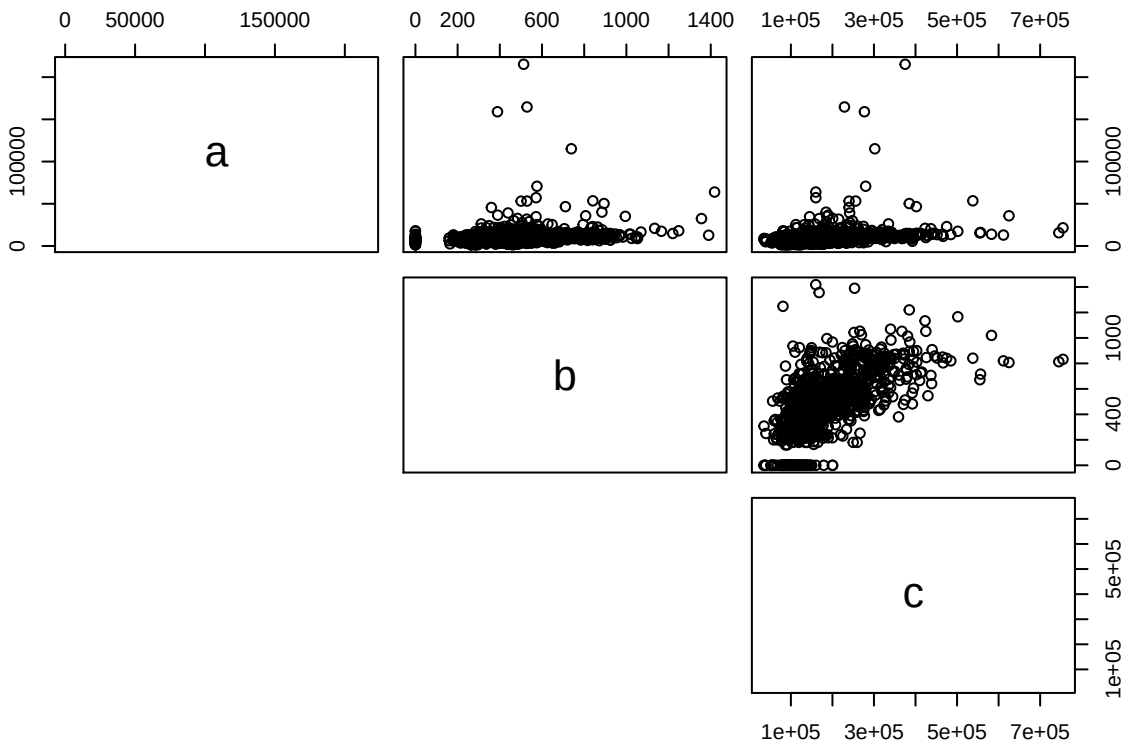
```
##     vars    n     mean      sd median  trimmed     mad   min    max  range skew
## X1     1 1460 180921.2 79442.5 163000 170783.3 56338.8 34900 755000 720100 1.88
##     kurtosis      se
## X1      6.5 2079.11
```

```
hist(c, breaks=30, main = "Histogram of SalePrice")
```

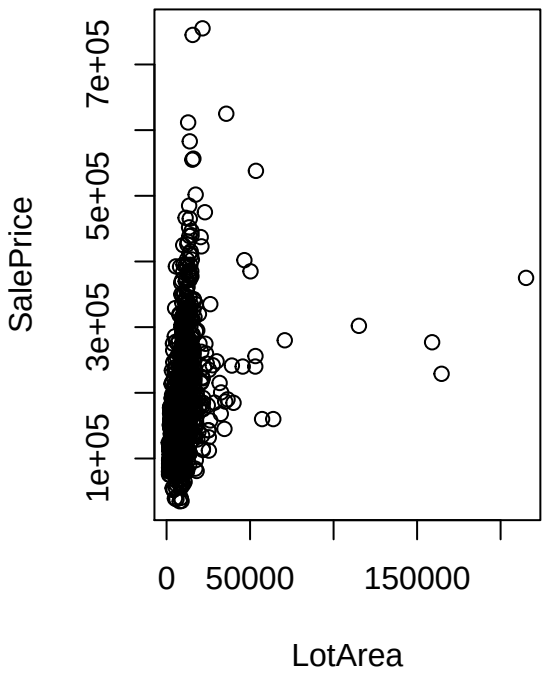


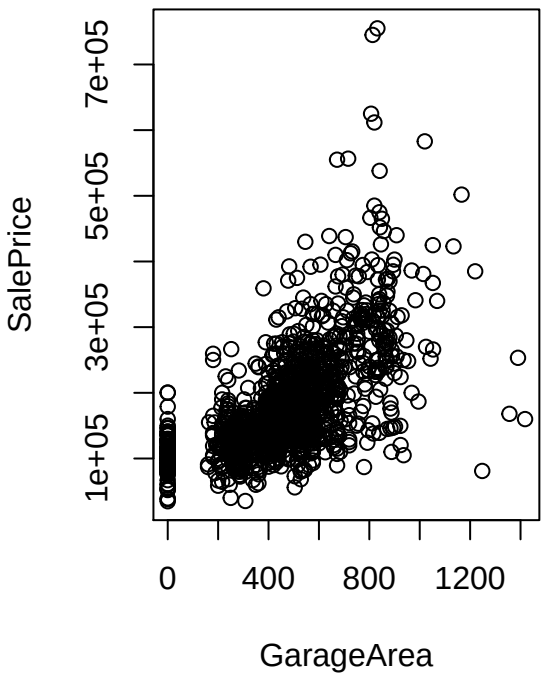
There are 1460 observations in LotArea, GarageArea and SalePrice. The distribution of these observations are all right skewed with few outliers.

```
#scatterplot matrix for at least two of the independent variables and the dependent variable.
pairs(~ a + b + c,lower.panel=NULL, data = t)
```




```
par(mfrow=c(1,2))
plot(a, c, xlab="LotArea", ylab="SalePrice")
plot(b, c, xlab="GarageArea", ylab="SalePrice")
```



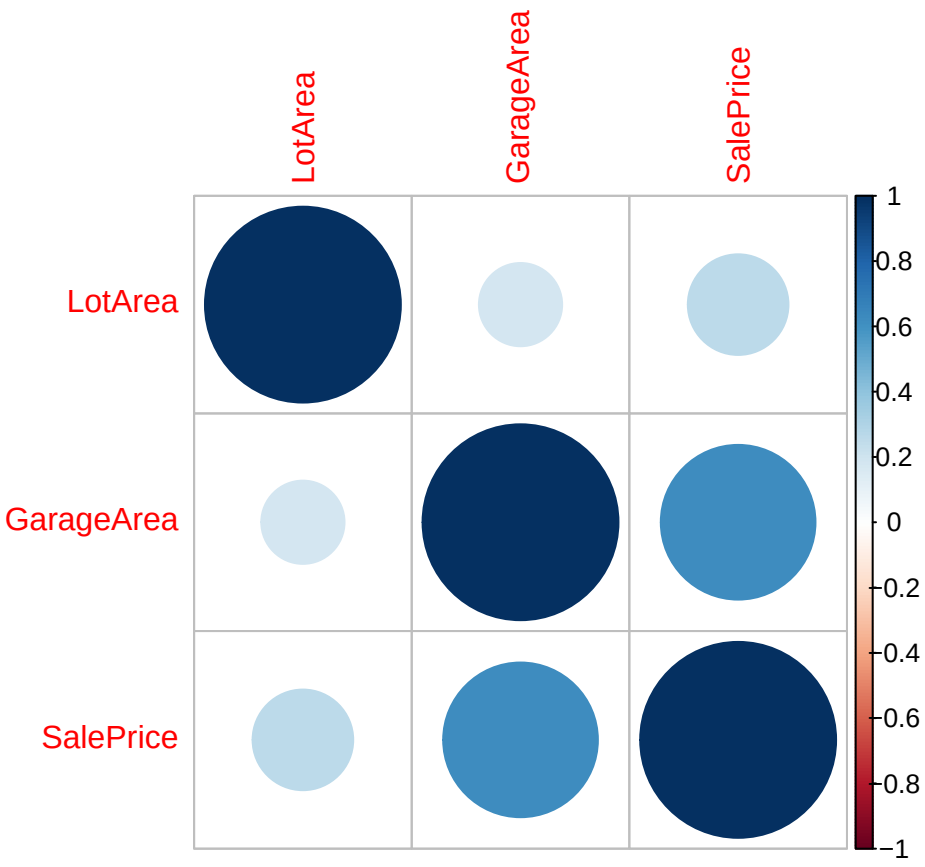


There not much correlation between LotArea and the SalePrice. There is some correlation between GarageArea and the SalePrice.

```
#correlation matrix
df_cor <- t[c("LotArea", "GarageArea", "SalePrice")]
mat_cor <- cor(df_cor, use = "pairwise.complete.obs")
print(mat_cor)
```

```
##              LotArea GarageArea SalePrice
## LotArea    1.0000000  0.1804028 0.2638434
## GarageArea 0.1804028  1.0000000 0.6234314
## SalePrice  0.2638434  0.6234314 1.0000000
```

```
#t %>% select(LotArea, GarageArea, SalePrice ) %>% cor(use="pairwise.complete.obs") %>% corrplot()
corrplot(mat_cor,method="circle")
```



The high correlation between GarageArea and the SalePrice is very clear here.

```
#Test the hypotheses that the correlations between each pairwise set of variables is 0 and provide an 8

#LotArea and GarageArea
cor.test(a, b, method = 'pearson', conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  a and b
## t = 7.0034, df = 1458, p-value = 3.803e-12
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.1477356 0.2126767
## sample estimates:
##       cor
## 0.1804028
```

```
#LotArea and SalePrice
cor.test(a, c, method = 'pearson', conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  a and c
## t = 10.445, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.2323391 0.2947946
## sample estimates:
##       cor
## 0.2638434
```

```
#GarageArea and SalePrice
cor.test(b, c, method = 'pearson', conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  b and c
## t = 30.446, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.6024756 0.6435283
## sample estimates:
##       cor
## 0.6234314
```

Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

The p values are almost 0 for all the above pair-wise comparisons so it is safe to say that the null hypotheses can be rejected. This means that SalePrice has no relation to the other variables.

I would not be worried about the familywise error as the p-values are almost 0 in all the cases.

Linear Algebra and Correlation

Invert your correlation matrix from above. (This is known as the precision matrix and contains variance inflation factors on the diagonal.) Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix. Conduct LU decomposition on the matrix.

```
#precision matrix
mat_pre <- solve(mat_cor)
mat_pre
```

```
##               LotArea  GarageArea  SalePrice
## LotArea     1.07530074 -0.02799273 -0.2662594
## GarageArea -0.02799273  1.63649778 -1.0128585
## SalePrice  -0.26625940 -1.01285847  1.7016986
```

```
mat_cor #correlation matrix
```

```
##              LotArea GarageArea SalePrice
## LotArea    1.0000000  0.1804028 0.2638434
## GarageArea 0.1804028  1.0000000 0.6234314
## SalePrice  0.2638434  0.6234314 1.0000000
```

```r
round(mat_cor %*% mat_pre)  #Multiply the correlation matrix by the precision matrix
```

```
##            LotArea GarageArea SalePrice
## LotArea          1          0         0
## GarageArea       0          1         0
## SalePrice        0          0         1
```

```r
round(mat_pre %*% mat_cor)  #multiply the precision matrix by the correlation matrix
```

```
##            LotArea GarageArea SalePrice
## LotArea          1          0         0
## GarageArea       0          1         0
## SalePrice        0          0         1
```

```r
#LU decomposition on the correlation matrix
cor_lu <- lu.decomposition(mat_cor)
#cor_lu_expand <- expand(cor_lu)

L <- cor_lu$L
U <- cor_lu$U

L #Lower Triangle
```

```
##           [,1]      [,2] [,3]
## [1,] 1.0000000 0.0000000    0
## [2,] 0.1804028 1.0000000    0
## [3,] 0.2638434 0.5952044    1
```

```r
U #Upper Triangle
```

```
##      [,1]      [,2]      [,3]
## [1,]    1 0.1804028 0.2638434
## [2,]    0 0.9674548 0.5758334
## [3,]    0 0.0000000 0.5876481
```

```r
L %*% U
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1804028 0.2638434
## [2,] 0.1804028 1.0000000 0.6234314
## [3,] 0.2638434 0.6234314 1.0000000
```

```r
#LU decomposition on the precision matrix
pre_lu <- lu.decomposition(mat_pre)


p_L <- pre_lu$L
p_U <- pre_lu$U

p_L #Lower Triangle
```

```
##             [,1]       [,2] [,3]
## [1,]  1.00000000  0.0000000    0
## [2,] -0.02603247  1.0000000    0
## [3,] -0.24761389 -0.6234314    1
```

```
p_U #Upper Triangle
```

```
##          [,1]        [,2]       [,3]
## [1,] 1.075301 -0.02799273 -0.2662594
## [2,] 0.000000  1.63576906 -1.0197899
## [3,] 0.000000  0.00000000  1.0000000
```

```
p_L %*% p_U
```

```
##             [,1]        [,2]       [,3]
## [1,]  1.07530074 -0.02799273 -0.2662594
## [2,] -0.02799273  1.63649778 -1.0128585
## [3,] -0.26625940 -1.01285847  1.7016986
```

In both correlation and precision matrix the multiplication of L and U matrix resulted in the original correlation and precision matrix.

Calculus-Based Probability & Statistics

Many times, it makes sense to fit a closed form distribution to data. Select a variable in the Kaggle.com training dataset that is skewed to the right, shift it so that the minimum value is absolutely above zero if necessary. Then load the MASS package and run fitdistr to fit an exponential probability density function. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html). Find the optimal value of λ for this distribution, and then take 1000 samples from this exponential distribution using this value (e.g., rexp(1000, λ)). Plot a histogram and compare it with a histogram of your original variable. Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF). Also generate a 95% confidence interval from the empirical data, assuming normality. Finally, provide the empirical 5th percentile and 95th percentile of the data. Discuss.

```
#selecting LotArea
la<- t$LotArea
min(la)
```

```
## [1] 1300
```

```
skim(t$LotArea)
```

Table 1: Data summary

Name	t$LotArea
Number of rows	1460
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	10516.83	9981.26	1300	7553.5	9478.5	11601.5	215245	

```
#shift it so that the minimum value is absolutely above zero
t_la_shift <- t %>% mutate(LotArea = LotArea - 1300)
skim(t_la_shift$LotArea)
```

Table 3: Data summary

Name	t_la_shift$LotArea
Number of rows	1460
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

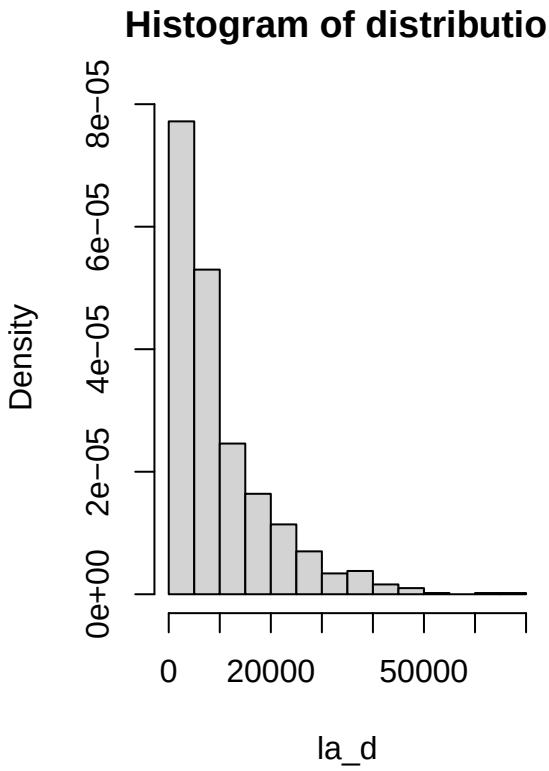
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	9216.83	9981.26	0	6253.5	8178.5	10301.5	213945	

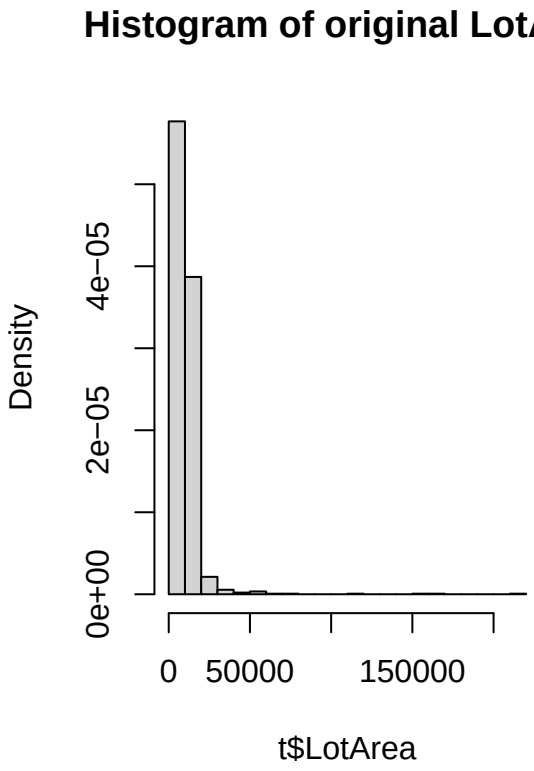
```
#run fitdistr to fit an exponential probability density function.
la_exp <- fitdistr(t_la_shift$LotArea, densfun = "exponential")
la_exp
```

```
##        rate
##   1.084972e-04
##  (2.839501e-06)
```

```
#Find the optimal value of $\lambda$ for this distribution, and then take 1000 samples from this expone
la_d <- rexp(1000, la_exp$estimate)

#Plot a histogram and compare it with a histogram of your original variable.
par(mfrow=c(1,2))
hist(la_d, freq = FALSE, breaks = 20, main = "Histogram of distribution")
hist(t$LotArea, freq = FALSE, breaks = 20, main = "Histogram of original LotArea")
```





The original data's histogram is heavily right skewed whereas the distribution's histogram is less right skewed.

```
#5th and 95th percentiles using the cumulative distribution function (CDF)
quantile(la_d, probs = c(0.05, 0.95))
```

```
##         5%        95%
##   398.6687 30378.5826
```

```
#generate a 95% confidence interval from the empirical data, assuming normality
CI(la_d, ci = 0.95)
```

```
##     upper      mean     lower
## 10569.824  9958.153  9346.483
```

```
#empirical 5th percentile and 95th percentile of the data
la_o <- sample(t$LotArea,  1000, replace=TRUE, prob=NULL)
quantile(la_o, c(.05, .95))
```

```
##       5%      95%
##  3735.00 17500.15
```

Modeling

Build some type of multiple regression model and submit your model to the competition board. Provide your complete model summary and results with analysis. Report your Kaggle.com user name and score.

```
train <- read.csv("https://raw.githubusercontent.com/irene908/DATA605/main/train.csv")  #train data(loa
test <- read.csv("https://raw.githubusercontent.com/irene908/DATA605/main/test.csv") #test data
```

```
#filter out the required attributes from the train set
t <- subset(train, select=c(MSSubClass,MSZoning,LotArea,LotShape,LotConfig,Neighborhood,BldgType,HouseS

#handle missing data
t<- na.omit(t)
```

Multiple Regression

```
t.lm <- lm(SalePrice ~ MSSubClass + MSZoning + LotArea + LotShape + LotConfig + Neighborhood + BldgType


summary(t.lm)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSSubClass + MSZoning + LotArea + LotShape +
##     LotConfig + Neighborhood + BldgType + HouseStyle + OverallQual +
##     OverallCond + YearBuilt + YearRemodAdd + RoofStyle + +MasVnrType +
##     MasVnrArea + ExterQual + BsmtQual + BsmtCond + BsmtExposure +
##     TotalBsmtSF + Heating + HeatingQC + Electrical + X1stFlrSF +
##     GrLivArea + TotRmsAbvGrd + Functional + GarageCars + GarageArea +
##     PavedDrive + WoodDeckSF + OpenPorchSF + MiscVal + MoSold +
##     YrSold + SaleType + SaleCondition, data = t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -358632  -12522      63   11080  243508
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.367e+05  1.316e+06  -0.180 0.857302
## MSSubClass            -1.733e+02  1.029e+02  -1.684 0.092509 .
## MSZoningFV             3.661e+04  1.495e+04   2.449 0.014440 *
## MSZoningRH             2.544e+04  1.492e+04   1.706 0.088316 .
## MSZoningRL             2.746e+04  1.259e+04   2.180 0.029442 *
## MSZoningRM             2.279e+04  1.181e+04   1.929 0.053929 .
## LotArea                4.521e-01  1.029e-01   4.395 1.20e-05 ***
## LotShapeIR2            6.859e+03  5.286e+03   1.298 0.194646
## LotShapeIR3           -3.753e+04  1.064e+04  -3.527 0.000435 ***
## LotShapeReg            1.257e+03  2.019e+03   0.623 0.533561
## LotConfigCulDSac       9.205e+03  4.022e+03   2.288 0.022270 *
## LotConfigFR2          -1.142e+04  5.087e+03  -2.246 0.024881 *
## LotConfigFR3          -1.357e+04  1.561e+04  -0.869 0.384935
## LotConfigInside        1.088e+03  2.191e+03   0.497 0.619542
## NeighborhoodBlueste    1.290e+04  2.371e+04   0.544 0.586569
## NeighborhoodBrDale     1.826e+04  1.352e+04   1.350 0.177218
## NeighborhoodBrkSide    3.364e+03  1.148e+04   0.293 0.769538
## NeighborhoodClearCr    3.012e+02  1.134e+04   0.027 0.978821
## NeighborhoodCollgCr    2.405e+02  8.984e+03   0.027 0.978649
## NeighborhoodCrawfor    2.403e+04  1.041e+04   2.308 0.021156 *
```

```
## NeighborhoodEdwards  -1.653e+04  9.915e+03  -1.667 0.095768 .
## NeighborhoodGilbert  -3.055e+03  9.645e+03  -0.317 0.751501
## NeighborhoodIDOTRR    3.493e+02  1.325e+04   0.026 0.978973
## NeighborhoodMeadowV   9.938e+03  1.277e+04   0.778 0.436593
## NeighborhoodMitchel  -9.850e+03  1.010e+04  -0.975 0.329708
## NeighborhoodNAmes    -5.740e+03  9.545e+03  -0.601 0.547689
## NeighborhoodNoRidge   5.380e+04  1.031e+04   5.217 2.12e-07 ***
## NeighborhoodNPkVill   1.812e+04  1.333e+04   1.360 0.174205
## NeighborhoodNridgHt   3.889e+04  9.243e+03   4.208 2.76e-05 ***
## NeighborhoodNWAmes   -5.128e+03  9.737e+03  -0.527 0.598525
## NeighborhoodOldTown  -1.046e+04  1.189e+04  -0.880 0.379137
## NeighborhoodSawyer   -7.115e+03  1.002e+04  -0.710 0.477841
## NeighborhoodSawyerW   3.679e+03  9.558e+03   0.385 0.700332
## NeighborhoodSomerst   1.135e+04  1.116e+04   1.016 0.309618
## NeighborhoodStoneBr   5.847e+04  1.012e+04   5.779 9.40e-09 ***
## NeighborhoodSWISU    -9.543e+03  1.186e+04  -0.805 0.421253
## NeighborhoodTimber   -4.176e+02  1.002e+04  -0.042 0.966772
## NeighborhoodVeenker   2.191e+04  1.276e+04   1.716 0.086314 .
## BldgType2fmCon        1.124e+04  1.513e+04   0.743 0.457785
## BldgTypeDuplex       -1.531e+04  7.506e+03  -2.040 0.041569 *
## BldgTypeTwnhs        -1.148e+04  1.230e+04  -0.933 0.350932
## BldgTypeTwnhsE       -6.284e+03  1.109e+04  -0.566 0.571173
## HouseStyle1.5Unf      1.254e+04  9.541e+03   1.314 0.189034
## HouseStyle1Story      1.649e+04  5.203e+03   3.168 0.001569 **
## HouseStyle2.5Fin     -1.996e+04  1.282e+04  -1.556 0.119951
## HouseStyle2.5Unf     -7.447e+03  1.051e+04  -0.708 0.478903
## HouseStyle2Story     -6.002e+03  4.150e+03  -1.446 0.148289
## HouseStyleSFoyer      2.042e+04  7.867e+03   2.596 0.009529 **
## HouseStyleSLvl        1.358e+04  6.620e+03   2.052 0.040357 *
## OverallQual           9.721e+03  1.205e+03   8.065 1.66e-15 ***
## OverallCond           4.854e+03  1.028e+03   4.721 2.60e-06 ***
## YearBuilt             2.330e+02  8.052e+01   2.894 0.003866 **
## YearRemodAdd          1.015e+02  6.641e+01   1.528 0.126760
## RoofStyleGable        2.627e+03  1.073e+04   0.245 0.806638
## RoofStyleGambrel      6.152e+03  1.426e+04   0.431 0.666204
## RoofStyleHip          7.059e+03  1.089e+04   0.648 0.516840
## RoofStyleMansard      2.958e+03  1.591e+04   0.186 0.852583
## RoofStyleShed         1.283e+04  2.501e+04   0.513 0.608166
## MasVnrTypeBrkFace     1.038e+04  8.580e+03   1.210 0.226360
## MasVnrTypeNone        1.365e+04  8.636e+03   1.581 0.114160
## MasVnrTypeStone       1.460e+04  9.068e+03   1.610 0.107734
## MasVnrArea            8.214e+00  7.201e+00   1.141 0.254211
## ExterQualFa          -3.604e+04  1.237e+04  -2.915 0.003623 **
## ExterQualGd          -2.403e+04  5.635e+03  -4.265 2.15e-05 ***
## ExterQualTA          -2.771e+04  6.263e+03  -4.424 1.05e-05 ***
## BsmtQualFa           -2.928e+04  7.960e+03  -3.678 0.000245 ***
## BsmtQualGd           -3.085e+04  4.061e+03  -7.596 5.81e-14 ***
## BsmtQualTA           -3.036e+04  4.967e+03  -6.114 1.29e-09 ***
## BsmtCondGd            3.768e+03  6.604e+03   0.571 0.568340
## BsmtCondPo            3.821e+04  3.432e+04   1.113 0.265710
## BsmtCondTA            7.609e+03  5.229e+03   1.455 0.145829
## BsmtExposureGd        1.997e+04  3.688e+03   5.416 7.27e-08 ***
## BsmtExposureMn       -3.770e+03  3.795e+03  -0.994 0.320586
## BsmtExposureNo       -9.457e+03  2.741e+03  -3.450 0.000578 ***
```

```
## TotalBsmtSF           2.915e+00  5.323e+00   0.548 0.583958
## HeatingGasW           6.823e+03  8.094e+03   0.843 0.399379
## HeatingGrav           2.148e+03  1.342e+04   0.160 0.872895
## HeatingOthW          -3.641e+04  2.301e+04  -1.582 0.113843
## HeatingQCFa          -1.376e+03  5.839e+03  -0.236 0.813785
## HeatingQCGd          -4.301e+03  2.589e+03  -1.661 0.096880 .
## HeatingQCPo          -3.263e+04  3.251e+04  -1.004 0.315678
## HeatingQCTA          -3.597e+03  2.480e+03  -1.451 0.147129
## ElectricalFuseF      -5.048e+01  7.770e+03  -0.006 0.994818
## ElectricalFuseP       4.554e+03  2.307e+04   0.197 0.843532
## ElectricalMix        -1.938e+04  4.852e+04  -0.399 0.689675
## ElectricalSBrkr       6.017e+02  3.705e+03   0.162 0.871014
## X1stFlrSF            -1.779e+01  8.317e+00  -2.139 0.032654 *
## GrLivArea             6.950e+01  6.082e+00  11.426  < 2e-16 ***
## TotRmsAbvGrd          1.061e+03  1.061e+03   1.001 0.317193
## FunctionalMaj2       -1.591e+04  1.812e+04  -0.878 0.380126
## FunctionalMin1       -2.121e+03  1.106e+04  -0.192 0.847901
## FunctionalMin2        1.907e+03  1.096e+04   0.174 0.861860
## FunctionalMod         4.067e+03  1.312e+04   0.310 0.756603
## FunctionalSev        -3.993e+04  3.455e+04  -1.156 0.247914
## FunctionalTyp         1.196e+04  9.501e+03   1.259 0.208200
## GarageCars            1.195e+04  2.673e+03   4.469 8.56e-06 ***
## GarageArea           -1.178e+01  9.091e+00  -1.296 0.195128
## PavedDriveP           2.319e+02  6.930e+03   0.033 0.973312
## PavedDriveY           5.521e+03  4.319e+03   1.278 0.201446
## WoodDeckSF            1.463e+01  7.109e+00   2.058 0.039753 *
## OpenPorchSF          -5.521e-02  1.402e+01  -0.004 0.996859
## MiscVal              -9.006e-01  1.736e+00  -0.519 0.603950
## MoSold               -2.819e+02  3.121e+02  -0.903 0.366601
## YrSold               -2.295e+02  6.507e+02  -0.353 0.724327
## SaleTypeCon           3.694e+04  2.262e+04   1.633 0.102717
## SaleTypeConLD         1.579e+04  1.274e+04   1.239 0.215567
## SaleTypeConLI         9.399e+03  1.457e+04   0.645 0.519033
## SaleTypeConLw         8.001e+03  1.524e+04   0.525 0.599611
## SaleTypeCWD           7.400e+03  1.637e+04   0.452 0.651328
## SaleTypeNew           2.554e+04  1.939e+04   1.317 0.188020
## SaleTypeOth           1.674e+04  1.853e+04   0.903 0.366553
## SaleTypeWD            2.864e+03  5.289e+03   0.541 0.588258
## SaleConditionAdjLand  2.410e+04  1.932e+04   1.247 0.212628
## SaleConditionAlloca   1.680e+04  1.206e+04   1.393 0.163977
## SaleConditionFamily  -4.631e+03  7.727e+03  -0.599 0.549074
## SaleConditionNormal   5.316e+03  3.583e+03   1.484 0.138091
## SaleConditionPartial -1.140e+04  1.867e+04  -0.611 0.541395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29830 on 1296 degrees of freedom
## Multiple R-squared:  0.8701, Adjusted R-squared:  0.8585
## F-statistic: 74.85 on 116 and 1296 DF,  p-value: < 2.2e-16
```

The R sqaured value is 0.8701 and the p value is almost 0. This suggests that the model is able to include 87% of the data.

```r
#filter out the required attributes from the test set
test.filter <- subset(test, select=c(MSSubClass,MSZoning,LotArea,LotShape,LotConfig,Neighborhood,BldgTy

#handle missing data
test.filter <- na.omit(test.filter)
```

```r
pred <- predict(t.lm, test.filter)

#summary of prediction
summary(pred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17955  129952  163046  180759  215478  536496
```

```r
#summary of train set SalePrice
summary(t$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  131500  164700  182579  215000  755000
```

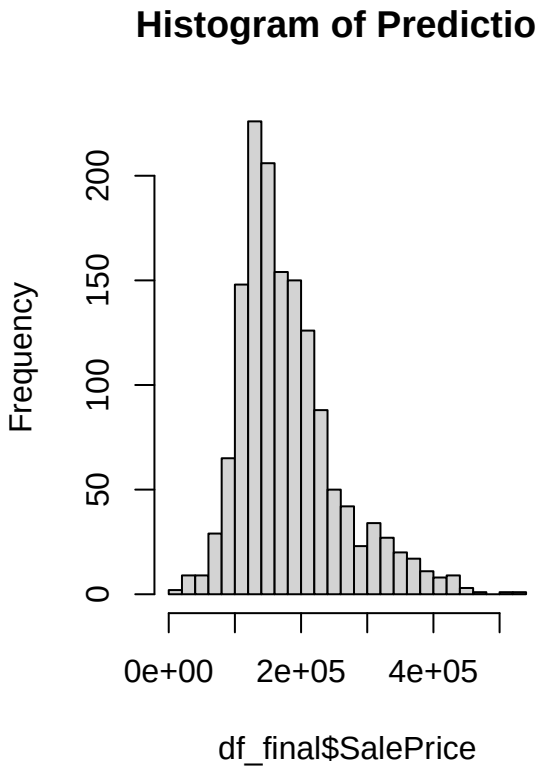
From the summary statistics the 1st Qu,Median,Mean and 3rd Qu looks good.

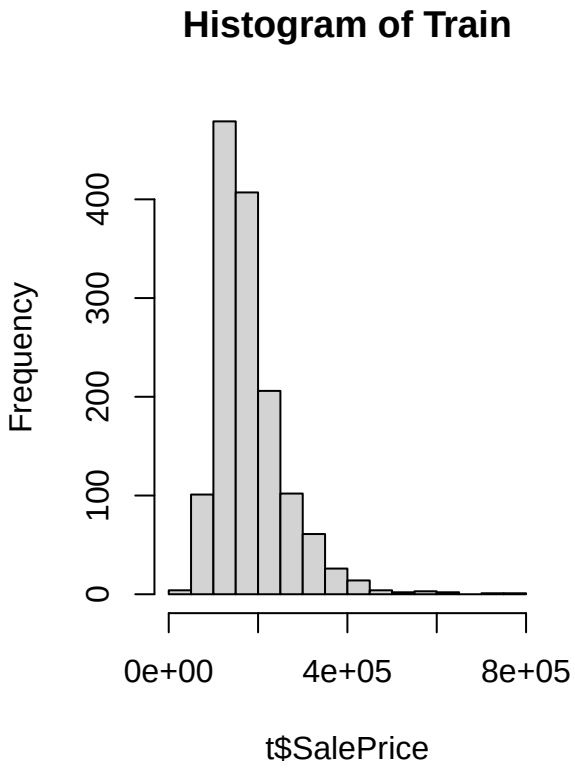
```r
#adding id to the predicted data
df_final <- as.data.frame(cbind(test$Id, pred))
colnames(df_final) = c("Id", "SalePrice")

head(df_final)
```

```
##     Id SalePrice
## 1 1461  112876.6
## 2 1462  146630.6
## 3 1463  162326.1
## 4 1464  177893.1
## 5 1465  227651.7
## 6 1466  172616.3
```

```r
#Histograms of predicted and Train data
par(mfrow=c(1,2))
hist(df_final$SalePrice, breaks=20, main = 'Histogram of Prediction')
hist(t$SalePrice, breaks=20, main = 'Histogram of Train')
```






```
#writing to csv file for kaggle submission
write.csv(df_final, file="IJacob_FinalProject_Kaggle.csv", row.names = FALSE)
```

Kaggle Submission

Username: Irene Jacob 908

Score: 0.52008

```
knitr::include_graphics("C://kaggle_1.JPG")
```


```
knitr::include_graphics("C://kaggle_2.JPG")
```

Name	Submitted	Wait time	Execution
IJacob_FinalProject_Kaggle.csv	2 minutes ago	1 seconds	0 seconds

Complete

Jump to your position on the leaderboard ▾