DATA 605 - Final

Mia Chen

12/13/2019

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(123)
n <- 6
X <- runif(10000, 1, n) # uniform variable
Y <- rnorm(10000, mean = (n+1)/2, sd = (n+1)/2) # normal variable
```

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) # median of X
x
```

```
## [1] 3.472838
```

```
y <- quantile(Y, 0.25)
y
```

```
##      25%
## 1.171246
```

a. $P(X > x | X > y)$

```
sum(X>x & X>y)/sum(X>y)
```

```
## [1] 0.5186184
```

Given that X is greater than the 25th percentile of Y, the probability that X is greater than its median is 51.86%.

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

```
sum(X>x & Y>y)/length(X)
```

```
## [1] 0.3756
```

Probability that X is greater than its median and Y is greater than its 25th percentile is 37.56%.

c. $P(X < x | X > y)$

```r
sum(X<x & X>y)/sum(X>y)
```

```
## [1] 0.4813816
```

Given that X is greater than the 25th percentile of Y, the probability that X is smaller than its median is 48.14%.

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.

```r
sum_1 <- c(sum(X<x & Y<y), sum(X>x & Y<y), sum(X & Y<y))
sum_2 <- c(sum(X<x & Y>y), sum(X>x & Y>y), sum(X & Y>y))
sum_3 <- c(sum(X<x & Y), sum(X>x & Y), sum(X & Y))

Z <- data.frame(sum_1, sum_2, sum_3)
colnames(Z) <- c("Y < y", "Y > y", "Total")
rownames(Z) <- c("X < x", "X > x", "Total")
Z
```

```
##       Y < y Y > y Total
## X < x  1256  3744  5000
## X > x  1244  3756  5000
## Total  2500  7500 10000
```

P(X>x and Y>y)

```r
3756/10000
```

```
## [1] 0.3756
```

P(X>x)P(Y>y)

```r
(5000/10000)*(7500/10000)
```

```
## [1] 0.375
```

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?

```r
# Contingency table
M <- matrix(c(1256, 1244, 3744, 3756), 2, 2)
M
```

```
##      [,1] [,2]
## [1,] 1256 3744
## [2,] 1244 3756
```

```r
fisher.test(M)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  M
## p-value = 0.7995
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9242273 1.1100187
## sample estimates:
## odds ratio
##   1.012883
```

```r
# Chi-square test
chisq.test(M)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  M
## X-squared = 0.064533, df = 1, p-value = 0.7995
```

Fisher's Exact Test is more appropriate for small sample size. Here, we have sample size of 10000, so it is more appropriate to use the Chi-square test. However, p-value of 0.7995 is the same from both test and it is much greater than 0.05, so we accept the null hypothesis that X and Y are independent.

Problem 2.

Kaggle.com - House Prices: Advanced Regression Techniques competition. https://www.kaggle.com/c/house-prices-advanced-regression-techniques .

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## corrplot 0.84 loaded

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':
##
##     describe

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Load data
data <- read.csv("https://raw.githubusercontent.com/miachen410/DATA605/master/train.csv")

# Data structure
str(data)
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
```

```
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

This dataset contains 1460 observations and 81 variables. The last variable SalePrice is the response variable (dependent variable) that we will be working with in the analysis below.

5 points. 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
# Univariate descriptive statistics
summary(data)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage    
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00  
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00  
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00  
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05  
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00  
##  Max.   :1460.0   Max.   :190.0                  Max.   :313.00  
##                                                  NA's   :259     
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63   
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50   
##  Median :  9478               NA's:1369   IR3: 10   Low:  36   
##  Mean   : 10517                           Reg:925   Lvl:1311   
##  3rd Qu.: 11602                                                
##  Max.   :215245                                                
##                                                                
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1  
##  AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260  
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81  
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48  
##                FR3    :   4              Edwards:100   RRAn   :  26  
##                Inside :1052              Somerst: 86   PosN   :  19  
##                                          Gilbert: 79   RRAe   :  11  
##                                          (Other):707   (Other):  15  
##   Condition2     BldgType      HouseStyle   OverallQual    
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000  
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000  
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000  
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099  
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000  
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000  
##  (Other):   2                 (Other): 19                   
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle   
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13  
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141  
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11  
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286  
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7  
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2  
##                                                              
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea    
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0  
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0  
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0  
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7  
##  ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0  
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0  
##  (Other):   2   (Other):128   (Other):136                 NA's   :8       
##  ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
```

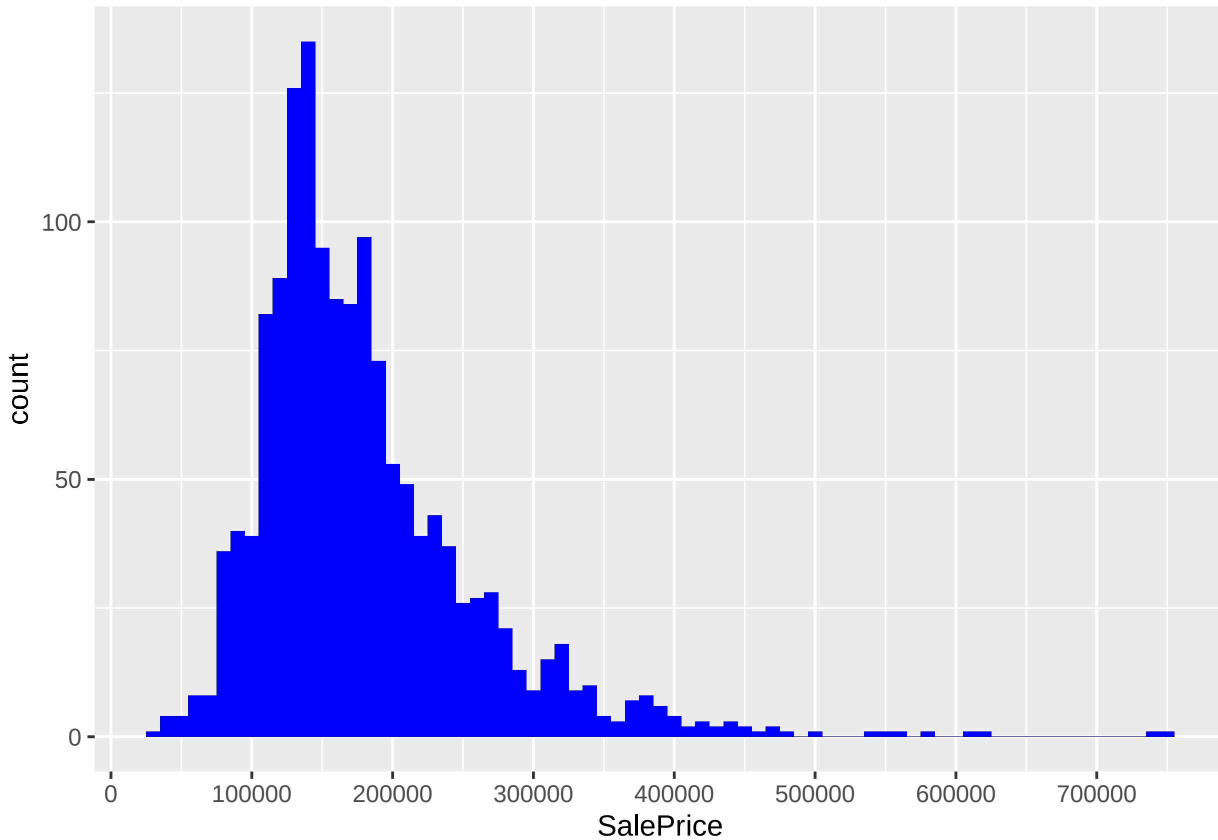
```
##  Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
##  Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
##  TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
##            TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00
##  BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00
##  GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00
##  LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55
##  Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00
##  Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00
##  NA's: 37                      NA's:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   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   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
##  Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
##                   Typ :1360                               NA's   : 81
```

```
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex   :   3
##  1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa   :  48
##  Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd   :  14
##  Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po   :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA   :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
##  NA's   :81
##  GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  Ex   :   2   N:  90      Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Fa   :  35   P:  30      1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Gd   :   9   Y:1340      Median :  0.00   Median : 25.00   Median :  0.00
##  Po   :   7               Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  TA   :1326               3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NA's:  81                Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch       PoolArea        PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex   :   2
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa   :   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Gd   :   3
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  GdPrv:  59   Gar2:    2   Min.   :    0.00   Min.   : 1.000
##  GdWo :  54   Othr:    2   1st Qu.:    0.00   1st Qu.: 5.000
##  MnPrv: 157   Shed:   49   Median :    0.00   Median : 6.000
##  MnWw :  11   TenC:    1   Mean   :   43.49   Mean   : 6.322
##  NA's :1179   NA's:1406    3rd Qu.:    0.00   3rd Qu.: 8.000
##                            Max.   :15500.00   Max.   :12.000
##
##      YrSold        SaleType     SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
```

```r
# Focus on the summary statistics of SalesPrice
summary(data$SalePrice)
```

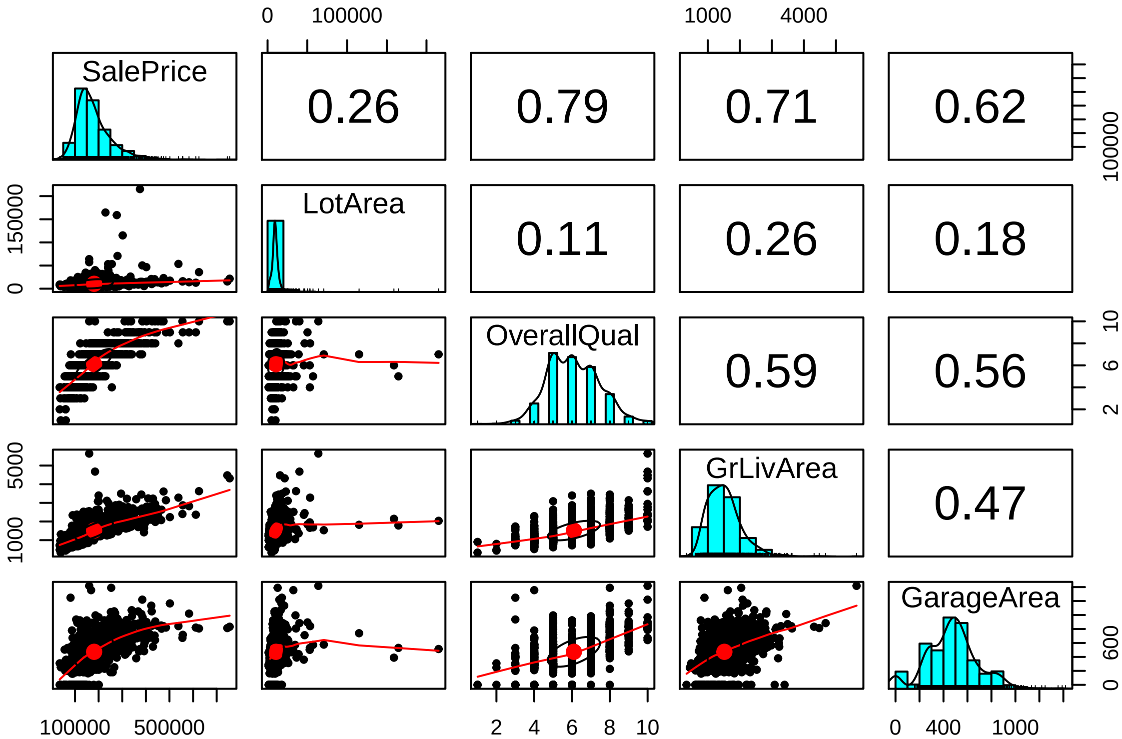
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

```r
# Distribution of SalePrice in a histogram

options(scipen = 5) # setting for not switching xticks to scientific notation

ggplot(data, aes(x = SalePrice)) +
  geom_histogram(fill="blue", binwidth = 10000) +
  scale_x_continuous(breaks = seq(0, 800000, by = 100000))
```




```
# Choosing LotArea, OverallQual, GrLivArea and GarageArea with SalePrice as the correlation testing dat
select_data <- data[, c("SalePrice", "LotArea", "OverallQual", "GrLivArea", "GarageArea")]
```

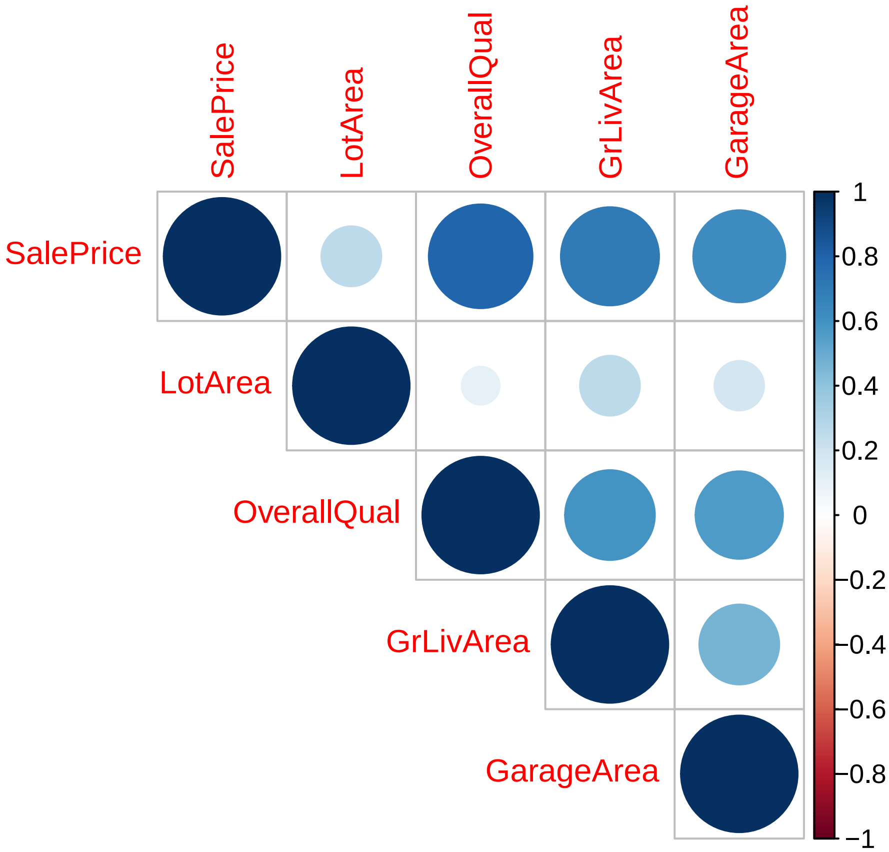
```
# Scatterplot matrix
pairs.panels(select_data, method = "pearson") #correlation method
```




```r
# Correlation Matrix

corr_data <- rcorr(as.matrix(select_data, use = "complete.obs")) # only use observations that have comp
corr_data # display correlation matrix
```

```
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice        1.00    0.26        0.79      0.71       0.62
## LotArea          0.26    1.00        0.11      0.26       0.18
## OverallQual      0.79    0.11        1.00      0.59       0.56
## GrLivArea        0.71    0.26        0.59      1.00       0.47
## GarageArea       0.62    0.18        0.56      0.47       1.00
##
## n= 1460
##
##
## P
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice             0       0           0         0
## LotArea     0                 0           0         0
## OverallQual 0         0                   0         0
## GrLivArea   0         0       0                     0
## GarageArea  0         0       0           0
```

```r
corr_matrix <- cor(select_data, use = "complete.obs")
corrplot(corr_matrix, type = "upper") # visualize the correlation matrix
```



We can see that the correlations are non-zero between the independent variables and the p-values are zero. Therefore, we can reject the null hypotheses that the correlations between each pairwise set of variables is 0. With that said, independent variables OverallQual, GrLivArea and GarageArea and LotArea each has a linear relationship with SalePrice, with OverallQual having the strongest correlation.

5 points. 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
precision_matrix <- solve(corr_matrix)

round(precision_matrix, 2)
```

```
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice        3.97   -0.39       -1.98     -1.19      -0.73
## LotArea         -0.39    1.13        0.34     -0.19      -0.06
## OverallQual     -1.98    0.34        2.84     -0.22      -0.32
## GrLivArea       -1.19   -0.19       -0.22      2.06      -0.06
## GarageArea      -0.73   -0.06       -0.32     -0.06       1.68
```

```
# Multiply correlation matrix by precision matrix
round(corr_matrix %*% precision_matrix, 2)
```

```
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice           1       0           0         0          0
## LotArea             0       1           0         0          0
## OverallQual         0       0           1         0          0
## GrLivArea           0       0           0         1          0
## GarageArea          0       0           0         0          1
```

```
#Multiply precision matrix by correlation matrix
round(precision_matrix %*% corr_matrix, 2)
```

```
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice           1       0           0         0          0
## LotArea             0       1           0         0          0
## OverallQual         0       0           1         0          0
## GrLivArea           0       0           0         1          0
## GarageArea          0       0           0         0          1
```

```
# LU decomposition
Z <- lu.decomposition(corr_matrix)
Z
```

```
## $L
##           [,1]        [,2]      [,3]       [,4] [,5]
## [1,] 1.0000000  0.00000000 0.0000000 0.00000000    0
## [2,] 0.2638434  1.00000000 0.0000000 0.00000000    0
## [3,] 0.7909816 -0.11058789 1.0000000 0.00000000    0
## [4,] 0.7086245  0.08184802 0.1127361 1.00000000    0
## [5,] 0.6234314  0.01710527 0.1946689 0.03685855    1
##
## $U
##      [,1]      [,2]       [,3]       [,4]       [,5]
## [1,]    1 0.2638434  0.7909816 0.70862448 0.62343144
## [2,]    0 0.9303867 -0.1028895 0.07615031 0.01591451
## [3,]    0 0.0000000  0.3629698 0.04091981 0.07065891
## [4,]    0 0.0000000  0.0000000 0.48700546 0.01795032
## [5,]    0 0.0000000  0.0000000 0.00000000 0.59664431
```

```
L <- Z$L
U <- Z$U
```

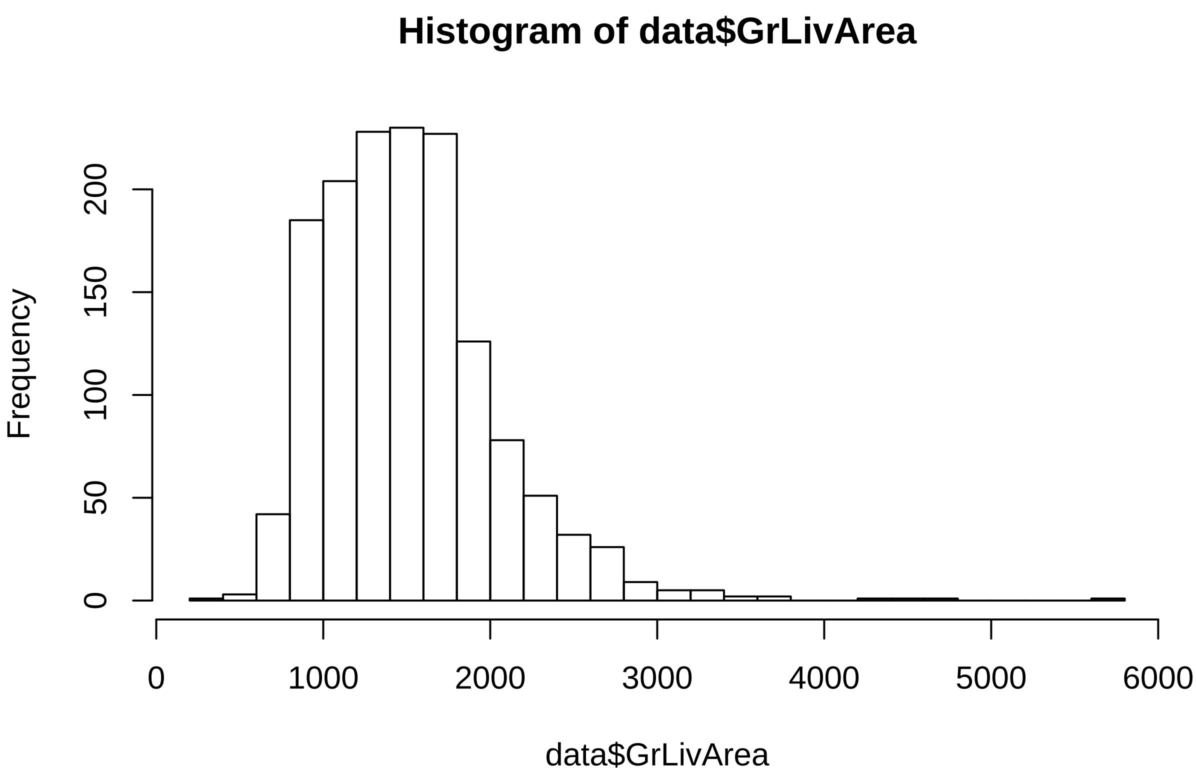
```
# Test if L*U gives us the original correlation matrix
(L %*% U) == corr_matrix
```

```
##             SalePrice LotArea OverallQual GrLivArea GarageArea
## SalePrice        TRUE    TRUE        TRUE      TRUE       TRUE
## LotArea          TRUE    TRUE        TRUE      TRUE       TRUE
## OverallQual      TRUE    TRUE        TRUE      TRUE       TRUE
## GrLivArea        TRUE    TRUE        TRUE      TRUE       TRUE
## GarageArea       TRUE    TRUE        TRUE      TRUE       TRUE
```

5 points. 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.

```
# Fit a variable to exponential distribution

hist(data$GrLivArea, breaks = 30) # GrLivArea is right-skewed
```

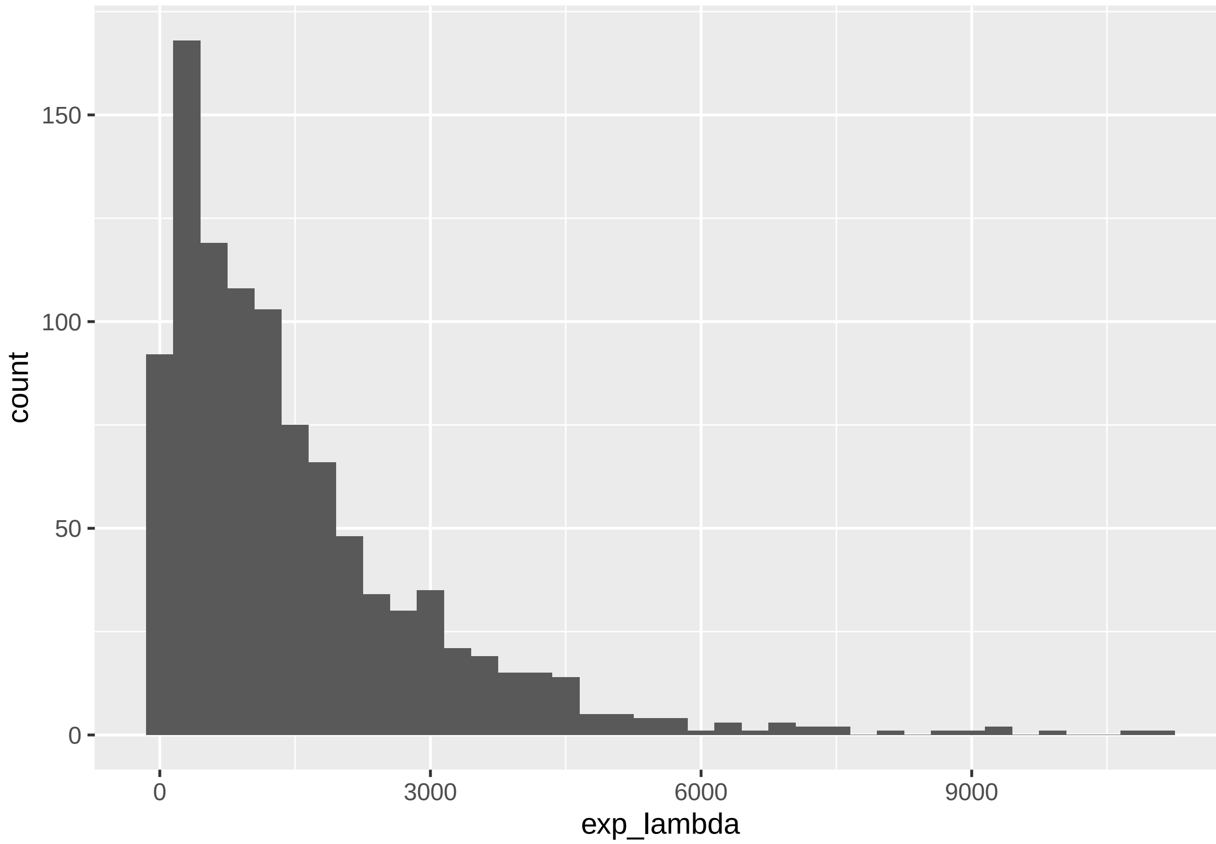



```
fit <- fitdistr(data$GrLivArea, "exponential") # Fit exponential distribution

lambda <- fit$estimate # Find optimal value of lambda
lambda
```

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

```
# Take 1000 sample of the exponential distribution with lambda
set.seed(1234)
exp_lambda <- rexp(1000, lambda)
```

```r
# Plot a histogram
ggplot(as.data.frame(exp_lambda), aes(exp_lambda)) +
  geom_histogram(binwidth = 300)
```




```r
# Find 5th and 95th percentiles of the exponential distribution
qexp(0.05, rate = lambda) # 5th percentile
```

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

```r
qexp(0.95, rate = lambda) # 95th percentile
```

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

```r
# Construct a 95% confidence interval from the empirical data, assuming normality
ci(data$GrLivArea, confidence = 0.95)
```

```
##   Estimate   CI lower   CI upper Std. Error
## 1515.46370 1488.48701 1542.44038   13.75245
```

```r
# The empirical 5th and 95th percentiles
quantile(data$GrLivArea, c(0.05, 0.95))
```

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

The empirical 5th and 95th percentiles are very different from those from the fitted exponential distribution. This suggests that the exponential distribution is not a good fit for the data.

The 95% confidence interval suggests that 95% of the time, we would expect to see a value between 1488.50 and 1542.44. However, this is built upon the mean. For this right-skewed variable, median would be a better representation of the data.

10 points. 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.

```r
# Multiple Linear Regression Model
lm <- lm(SalePrice ~ OverallQual+GrLivArea+GarageArea+OverallQual:GrLivArea+OverallQual:GarageArea+GrLiv

summary(lm)
```

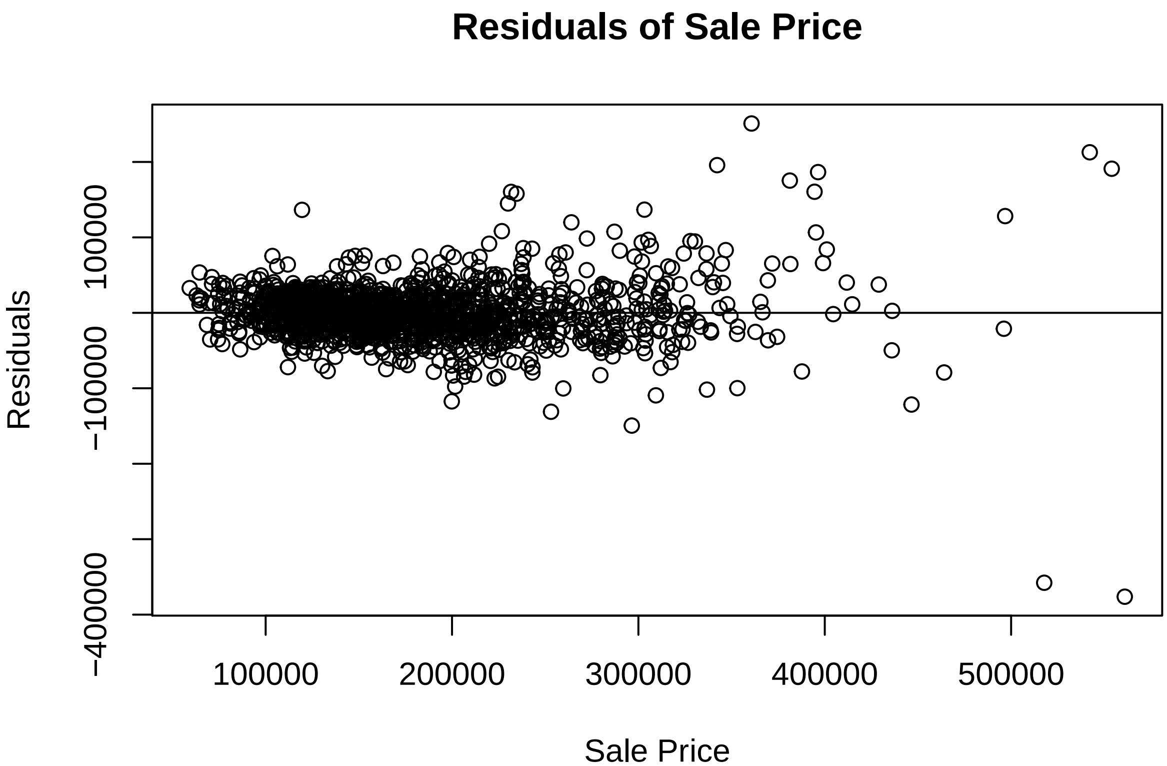
```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageArea +
##     OverallQual:GrLivArea + OverallQual:GarageArea + GrLivArea:GarageArea,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376232  -18777   -1281   17013  250945
##
## Coefficients:
##                            Estimate    Std. Error t value Pr(>|t|)
## (Intercept)             82227.263642  12282.145640   6.695 3.07e-11 ***
## OverallQual            -14415.614721   2720.911401  -5.298 1.35e-07 ***
## GrLivArea                   5.426764      8.504123   0.638    0.523
## GarageArea                 -7.904951     20.911022  -0.378    0.705
## OverallQual:GrLivArea      14.493692      1.478019   9.806  < 2e-16 ***
## OverallQual:GarageArea     39.524778      3.541293  11.161  < 2e-16 ***
## GrLivArea:GarageArea       -0.102236      0.009485 -10.779  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37100 on 1453 degrees of freedom
## Multiple R-squared:  0.7828, Adjusted R-squared:  0.7819
## F-statistic: 872.7 on 6 and 1453 DF,  p-value: < 2.2e-16
```

R^2 explains 78.19% of the variability and p-value is nearly zero, suggests that this relationship is not due to random variation.

Multiple Linear Equation:

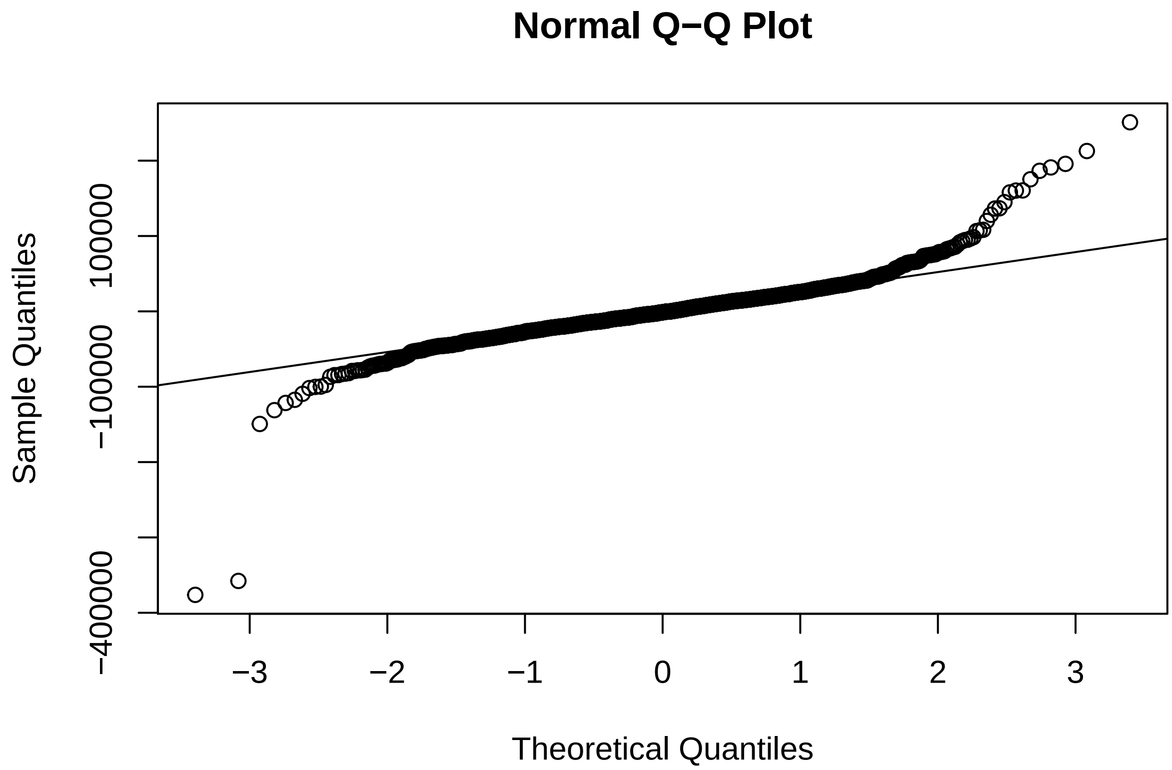
$$SalePrice = 82227.26 - 14415.61 \times OverallQual + 5.43 \times GrLivArea - 7.90 \times GarageArea + 14.49 \times OverallQual \times GrLivArea + 3$$

```r
# Residuals variability plot
plot(fitted(lm), resid(lm),
     xlab = "Sale Price", ylab = "Residuals",
     main = "Residuals of Sale Price")
abline(h = 0)
```



The residual plot seems to meet the constant variability condition with the residuals constantly above and below the zero line, with a few outliners.

```r
# Quantile-Quantile Plot
qqnorm(lm$residuals)
qqline(lm$residuals)
```



There is no significant curvature in the QQ plot; points tend to follow the straight line which suggests there is a linear relationship.