DATA 605 - Final Project

William Outcault

```
require(ggplot2)
require(corrplot)
require(dplyr)
require(MASS)
```

Creating the Distributions

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 (N+1)/2, which is also equal to the standard deviation.

```
set.seed(101)
n <- 10
X <- runif(10000, min=1, max=n)
mn <- (n+1)/2
Y <- rnorm(10000, mean = mn, sd = mn)
x <- median(X)
y <-  as.numeric(quantile(Y)[2])
```

Distribution Probabilites

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.

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

```
a <- round(length(X[X > x]) / length(X[X>y]),4)
b <- round(length(X[X > x & Y > y]) / length(X),4)
c <- round(length(X[X < x & X > y]) / length(X),4)
```

a) = 0.5498; b) = 0.3778; c) = 0.4094

Contingency Table

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.

```
rownames = c('P(X>x)','P(X<=x)','Total')
colnames = c('P(Y>y)','P(Y<=y)','Total')
r1c1 = length(X[X > x & Y > y])
r2c1 = length(X[X <= x & Y > y])
r3c1 = r1c1 + r2c1
```

```r
r1c2 = length(X[X > x & Y <= y])
r2c2 = length(X[X <= x & Y <= y])
r3c2 = r1c2 + r2c2
r1c3 = r1c1 + r1c2
r2c3 = r2c1 + r2c2
r3c3 = r1c3 + r2c3

m <- matrix(c(r1c1,r2c1,r3c1,r1c2,r2c2,r3c2,r1c3,r2c3,r3c3),
            nrow = 3,byrow=TRUE, dimnames=list(rownames,colnames))

A <- (r1c3/10000)*(r3c1/10000)
B <- r1c1/10000


knitr::kable(m)
```

	P(Y>y)	P(Y<=y)	Total
P(X>x)	3778	3722	7500
P(X<=x)	1222	1278	2500
Total	5000	5000	10000

For p(A and B) = p(A) * p(B) we get unequal values however very close.

According to our contingency table $P(X > x, Y > y) = 0.3778$ and $P(X > x) * P(Y > y) = 0.375$

Testing Independence

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?

Hypotheses

H0: The variables are independent, and there is no relationship between variables. H1: The variables are dependent, there is a relationship between variables.

Expected Frequencies

Fisher's exact test should be used given a small sample size (specifically when expected values of the contingecy table falls below 5). Adversely Chi-square test is used when you have a large enough sample size.

Fisher's exact test should not be used for larger sample sizes, over Chi-square tests largely because it is too conservative and can be misleading. However the conservative nature of the Fisher's exact test provides better feedback than using Chi-square test on the same small sample.

We see our frequency counts are well above 5 in our contigency table, therefore we will test using the chisq.test function. It is worth noting that if the sample size is to small, our function will produce a warning about inaccuracy, at which point we would use Fisher's exact test.

Chi-squared test

```
m2 <- m[-3,-3]
chisq.test(m2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m2
## X-squared = 1.6133, df = 1, p-value = 0.204
```

As expected our function did not produce a warning. We see our p-value from the Chi-squared test was greater than our threshold 0.05 therefore we fail to reject our null-hypothesis.

Fisher's exact test

```
fisher.test(m2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  m2
## p-value = 0.204
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.968653 1.163409
## sample estimates:
## odds ratio
##   1.061525
```

Our p-value was the same using the Fisher's exact test. If our sample size was larger however, we would find this test to be computationally impractical.

Advanced Regression for Housing Prices

Descriptive and Inferenctial Statistics

```
train <- read.csv('https://raw.githubusercontent.com/willoutcault/DATA605_Final/master/train.csv',
                  TRUE, ",")
test <- read.csv('https://raw.githubusercontent.com/willoutcault/DATA605_Final/master/test.csv',
                 TRUE, ",")
```

Data Overview

```r
glimpse(train)
```

```
## Observations: 1,460
## Variables: 81
## $ Id           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ MSSubClass   <int> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, ...
## $ MSZoning     <fct> RL, RL, RL, RL, RL, RL, RL, RL, RM, RL, RL, RL, RL, R...
## $ LotFrontage  <int> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, NA, 9...
## $ LotArea      <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, ...
## $ Street       <fct> Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave,...
## $ Alley        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ LotShape     <fct> Reg, Reg, IR1, IR1, IR1, IR1, Reg, IR1, Reg, Reg, Reg...
## $ LandContour  <fct> Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl...
## $ Utilities    <fct> AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllPu...
## $ LotConfig    <fct> Inside, FR2, Inside, Corner, FR2, Inside, Inside, Cor...
## $ LandSlope    <fct> Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl...
## $ Neighborhood <fct> CollgCr, Veenker, CollgCr, Crawfor, NoRidge, Mitchel,...
## $ Condition1   <fct> Norm, Feedr, Norm, Norm, Norm, Norm, Norm, PosN, Arte...
## $ Condition2   <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm,...
## $ BldgType     <fct> 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam,...
## $ HouseStyle   <fct> 2Story, 1Story, 2Story, 2Story, 2Story, 1.5Fin, 1Stor...
## $ OverallQual  <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4,...
## $ OverallCond  <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5,...
## $ YearBuilt    <int> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931,...
## $ YearRemodAdd <int> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950,...
## $ RoofStyle    <fct> Gable, Gable, Gable, Gable, Gable, Gable, Gable, Gabl...
## $ RoofMatl     <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg,...
## $ Exterior1st  <fct> VinylSd, MetalSd, VinylSd, Wd Sdng, VinylSd, VinylSd,...
## $ Exterior2nd  <fct> VinylSd, MetalSd, VinylSd, Wd Shng, VinylSd, VinylSd,...
## $ MasVnrType   <fct> BrkFace, None, BrkFace, None, BrkFace, None, Stone, S...
## $ MasVnrArea   <int> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 30...
## $ ExterQual    <fct> Gd, TA, Gd, TA, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, G...
## $ ExterCond    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ Foundation   <fct> PConc, CBlock, PConc, BrkTil, PConc, Wood, PConc, CBl...
## $ BsmtQual     <fct> Gd, Gd, Gd, TA, Gd, Gd, Ex, Gd, TA, TA, TA, Ex, TA, G...
## $ BsmtCond     <fct> TA, TA, TA, Gd, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ BsmtExposure <fct> No, Gd, Mn, No, Av, No, Av, Mn, No, No, No, No, No, A...
## $ BsmtFinType1 <fct> GLQ, ALQ, GLQ, ALQ, GLQ, GLQ, GLQ, ALQ, Unf, GLQ, Rec...
## $ BsmtFinSF1   <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906,...
## $ BsmtFinType2 <fct> Unf, Unf, Unf, Unf, Unf, Unf, Unf, BLQ, Unf, Unf, Unf...
## $ BsmtFinSF2   <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ BsmtUnfSF    <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140, 134,...
## $ TotalBsmtSF  <int> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952, 991,...
## $ Heating      <fct> GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA,...
## $ HeatingQC    <fct> Ex, Ex, Ex, Gd, Ex, Ex, Ex, Ex, Gd, Ex, Ex, Ex, TA, E...
## $ CentralAir   <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ Electrical   <fct> SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrk...
## $ X1stFlrSF    <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 107...
## $ X2ndFlrSF    <int> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142,...
## $ LowQualFinSF <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ GrLivArea    <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774,...
## $ BsmtFullBath <int> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...
## $ BsmtHalfBath <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ FullBath      <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2,...
## $ HalfBath      <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...
## $ BedroomAbvGr  <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2,...
## $ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ KitchenQual   <fct> Gd, TA, Gd, Gd, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, G...
## $ TotRmsAbvGrd  <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6...
## $ Functional    <fct> Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Min1, Typ, Ty...
## $ Fireplaces    <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0,...
## $ FireplaceQu   <fct> NA, TA, TA, Gd, TA, NA, Gd, TA, TA, TA, NA, Gd, NA, G...
## $ GarageType    <fct> Attchd, Attchd, Attchd, Detchd, Attchd, Attchd, Attch...
## $ GarageYrBlt   <int> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, 1931,...
## $ GarageFinish  <fct> RFn, RFn, RFn, Unf, RFn, Unf, RFn, RFn, Unf, RFn, Unf...
## $ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2,...
## $ GarageArea    <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384...
## $ GarageQual    <fct> TA, TA, TA, TA, TA, TA, TA, TA, Fa, Gd, TA, TA, TA, T...
## $ GarageCond    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ PavedDrive    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ WoodDeckSF    <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, ...
## $ OpenPorchSF   <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 2...
## $ EnclosedPorch <int> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, 176, ...
## $ X3SsnPorch    <int> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ScreenPorch   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0, 0, ...
## $ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ PoolQC        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Fence         <fct> NA, NA, NA, NA, NA, MnPrv, NA, NA, NA, NA, NA, NA, NA...
## $ MiscFeature   <fct> NA, NA, NA, NA, NA, Shed, NA, Shed, NA, NA, NA, NA, N...
## $ MiscVal       <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 7...
## $ MoSold        <int> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3,...
## $ YrSold        <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008,...
## $ SaleType      <fct> WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, New, WD, ...
## $ SaleCondition <fct> Normal, Normal, Normal, Abnorml, Normal, Normal, Norm...
## $ SalePrice     <int> 208500, 181500, 223500, 140000, 250000, 143000, 30700...
```

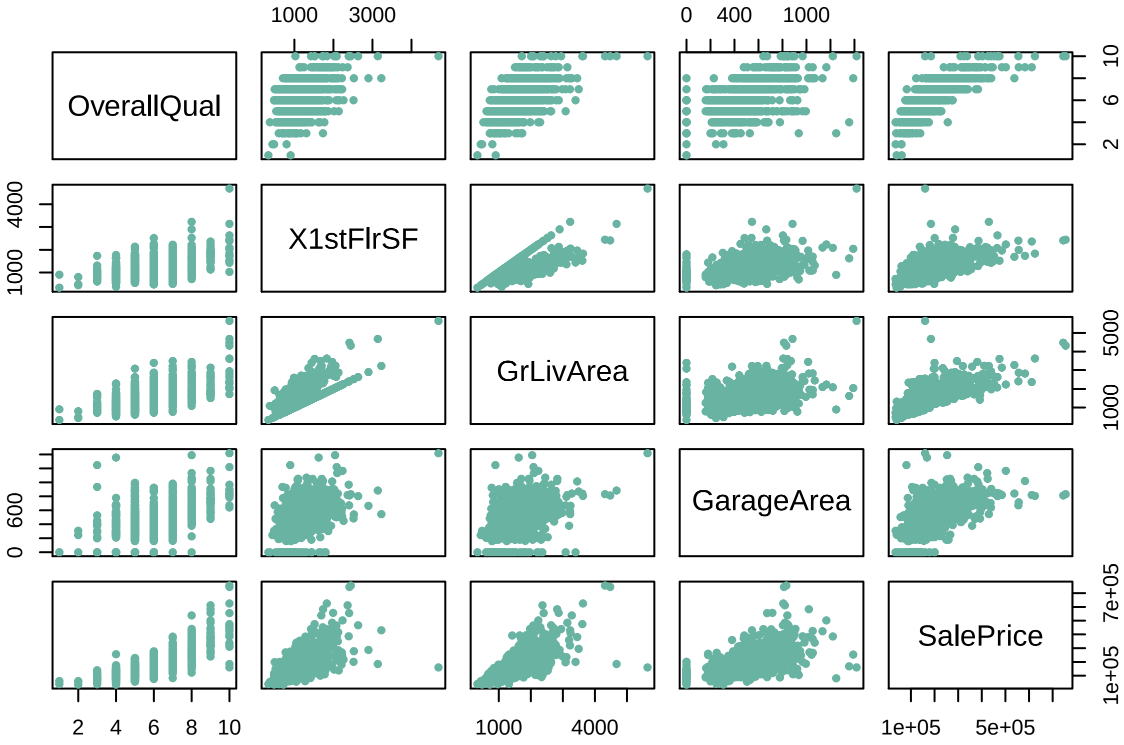
```r
summary(train)
```

```
##        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  Utilities
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##    LotConfig    LandSlope   Neighborhood   Condition1     Condition2
##  Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445
##  CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6
##  FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2
##  FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2
##  Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2
##                            Gilbert: 79   RRAe   :  11   PosA   :   1
##                            (Other):707   (Other):  15   (Other):   2
##    BldgType      HouseStyle   OverallQual      OverallCond      YearBuilt
##  1Fam  :1220   1Story :726   Min.   : 1.000   Min.   :1.000   Min.   :1872
##  2fmCon:  31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Duplex:  52   1.5Fin :154   Median : 6.000   Median :5.000   Median :1973
##  Twnhs :  43   SLvl   : 65   Mean   : 6.099   Mean   :5.575   Mean   :1971
##  TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                1.5Unf : 14   Max.   :10.000   Max.   :9.000   Max.   :2010
##                (Other): 19
##   YearRemodAdd    RoofStyle       RoofMatl     Exterior1st   Exterior2nd
##  Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515   VinylSd:504
##  1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222   MetalSd:214
##  Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220   HdBoard:207
##  Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206   Wd Sdng:197
##  3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108   Plywood:142
##  Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61   CmentBd: 60
##                                (Other):   2   (Other):128   (Other):136
##    MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual
##  BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146   Ex  :121
##  BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634   Fa  : 35
##  None   :864   Median :   0.0   Gd:488    Gd: 146   PConc :647   Gd  :618
##  Stone  :128   Mean   : 103.7   TA:906    Po:   1   Slab  : 24   TA  :649
##  NA's   :  8   3rd Qu.: 166.0             TA:1282   Stone :  6   NA's: 37
##                Max.   :1600.0                       Wood  :  3
##                NA's   :8
##  BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2
##  Fa  :  45   Av  :221     ALQ :220     Min.   :   0.0   ALQ :  19
##  Gd  :  65   Gd  :134     BLQ :148     1st Qu.:   0.0   BLQ :  33
##  Po  :   2   Mn  :114     GLQ :418     Median : 383.5   GLQ :  14
##  TA  :1311   No  :953     LwQ : 74     Mean   : 443.6   LwQ :  46
##  NA's:  37   NA's: 38     Rec :133     3rd Qu.: 712.2   Rec :  54
##                           Unf :430     Max.   :5644.0   Unf :1256
##                           NA's: 37                      NA's:  38
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49
##  Median :   0.00   Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  CentralAir Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  N:  95     FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  Y:1365     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  TotRmsAbvGrd
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100      Min.   : 2.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39      1st Qu.: 5.000
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586      Median : 6.000
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735      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
##
##  Functional    Fireplaces    FireplaceQu   GarageType   GarageYrBlt
##  Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6   Min.   :1900
##  Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870   1st Qu.:1961
##  Min1:  31   Median :1.000   Gd  :380    Basment: 19   Median :1980
##  Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88   Mean   :1979
##  Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9   3rd Qu.:2002
##  Sev :   1   Max.   :3.000   NA's:690    Detchd :387   Max.   :2010
##  Typ :1360                               NA's   : 81   NA's   :81
##  GarageFinish   GarageCars      GarageArea     GarageQual  GarageCond
##  Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3   Ex  :   2
##  RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48   Fa  :  35
##  Unf :605     Median :2.000   Median : 480.0   Gd  :  14   Gd  :   9
##  NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3   Po  :   7
##               3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311   TA  :1326
##               Max.   :4.000   Max.   :1418.0   NA's:  81   NA's:  81
##
##  PavedDrive   WoodDeckSF      OpenPorchSF    EnclosedPorch      X3SsnPorch
##  N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Y:1340     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        PoolQC       Fence      MiscFeature
##  Min.   :  0.00   Min.   :  0.000   Ex  :   2   GdPrv:  59   Gar2:   2
##  1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2   GdWo :  54   Othr:   2
##  Median :  0.00   Median :  0.000   Gd  :   3   MnPrv: 157   Shed:  49
##  Mean   : 15.06   Mean   :  2.759   NA's:1453   MnWw :  11   TenC:   1
##  3rd Qu.:  0.00   3rd Qu.:  0.000               NA's :1179   NA's:1406
##  Max.   :480.00   Max.   :738.000
##
##     MiscVal             MoSold          YrSold       SaleType
##  Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :1267
##  1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   New    : 122
##  Median :    0.00   Median : 6.000   Median :2008   COD    :  43
##  Mean   :   43.49   Mean   : 6.322   Mean   :2008   ConLD  :   9
##  3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   5
##  Max.   :15500.00   Max.   :12.000   Max.   :2010   ConLw  :   5
##                                                       (Other):   9
##  SaleCondition    SalePrice
##  Abnorml: 101   Min.   : 34900
##  AdjLand:   4   1st Qu.:129975
##  Alloca :  12   Median :163000
##  Family :  20   Mean   :180921
##  Normal :1198   3rd Qu.:214000
##  Partial: 125   Max.   :755000
##
```

Using Dplyr's glimpse function allows us to view each columns data type as well as the size of the data frame. We are working with a 1460x81 training set. After scrolling through each feature I found a few that I felt might have a correlation with Sales Price.

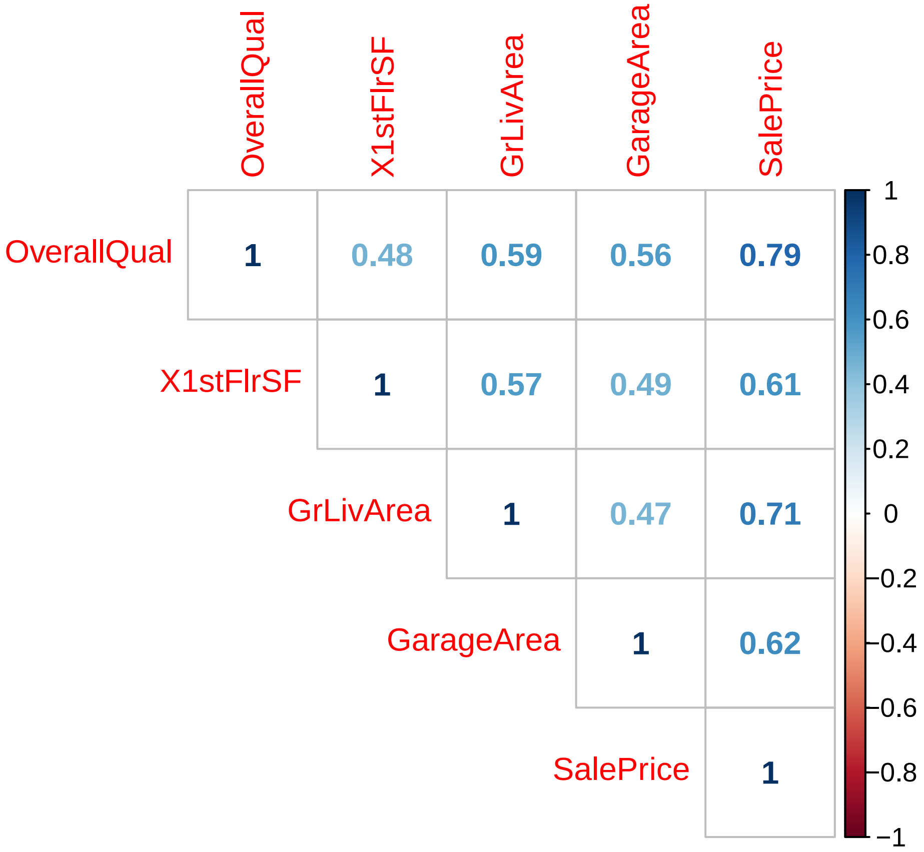
Correlation

```
pairs(train[ , c(18,44,47,63,81)], pch=20, col = "#69b3a2")
```



We notice a positive correlation between out features and our dependent variable, SalePrice. Using these same variables lets visualize a correlation plot.

```
train_sub <- cor(train[, c(18,44,47,63,81)])
corrplot(train_sub, method = "number", type="upper")
```



Out of the independent variables from the correlation plot we see OverallQall with the strongest correlation, and all correlations are positive. Next we will test the significance of each correlation significant.

Hypothesis Testing

We want to test to see if these correlations are significant using the Pearson correlation test. We will use a confidence level of 80% therefore our significance level alpha = 0.20

Null Hypothesis

H0: The variables are independent, and there is no correlation between variables. H1: The variables are dependent, there is a correlation between variables.

Pearson Tests

```r
cor.test(train$SalePrice,train$OverallQual,method = "pearson",conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$SalePrice and train$OverallQual
## t = 49.364, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
```

```
##  0.7780752 0.8032204
## sample estimates:
##       cor
## 0.7909816
```

```r
cor.test(train$SalePrice,train$X1stFlrSF,method = "pearson",conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$SalePrice and train$X1stFlrSF
## t = 29.078, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.5841687 0.6266715
## sample estimates:
##       cor
## 0.6058522
```

```r
cor.test(train$SalePrice,train$GrLivArea,method = "pearson",conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$SalePrice and train$GrLivArea
## t = 38.348, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.6915087 0.7249450
## sample estimates:
##       cor
## 0.7086245
```

```r
cor.test(train$SalePrice,train$GarageArea,method = "pearson",conf.level = 0.80)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$SalePrice and train$GarageArea
## 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
```

Each feature had a p-value less than the significance level so we reject our null hypothesis, therefore we can say the correlation between the listed independent variables and dependent variable is significant.

FWE

```
1-(1-(.2))^4
```

```
## [1] 0.5904
```

Our FWE came in just under 60% which is very high considering we only ran four tests. This is due to our 80% confidence interval, however I am not worried about our chances for a FWE due to the fact that each p-value is extremely low. If we were to change our CI to 95% we would still reject our null hypothesis for each test and reduce our FWE to about 18%.Once again due to our low p-values a type-1 error is very low.

Linear Algebra and Correlation

Precision Matrix

We will calculate the precision matrix by inverting the correlation matrix.

```
library(Matrix)
train_sub_inv <- solve(train_sub)
train_sub_inv
```

```
##             OverallQual  X1stFlrSF   GrLivArea  GarageArea  SalePrice
## OverallQual   2.7480965  0.1115031 -0.19242376 -0.32159236 -1.9044012
## X1stFlrSF     0.1115031  1.7362287 -0.46706793 -0.31007012 -0.6158116
## GrLivArea    -0.1924238 -0.4670679  2.14880584  0.01164084 -1.0947759
## GarageArea   -0.3215924 -0.3100701  0.01164084  1.72833958 -0.6435199
## SalePrice    -1.9044012 -0.6158116 -1.09477590 -0.64351992  4.0564126
```

We notice a high correlation between our selected features and sales price, especially between OverallQual and SalesPrice.

To obtain our identity matrix we will multiply our precision matrix by our correlation matrix, and vice versa. We will also make sure these two results are equal.

```
zapsmall(train_sub_inv %*% train_sub)
```

```
##             OverallQual X1stFlrSF GrLivArea GarageArea SalePrice
## OverallQual           1         0         0          0         0
## X1stFlrSF             0         1         0          0         0
## GrLivArea             0         0         1          0         0
## GarageArea            0         0         0          1         0
## SalePrice             0         0         0          0         1
```

```
zapsmall(train_sub %*% train_sub_inv)
```

```
##             OverallQual X1stFlrSF GrLivArea GarageArea SalePrice
## OverallQual           1         0         0          0         0
## X1stFlrSF             0         1         0          0         0
## GrLivArea             0         0         1          0         0
## GarageArea            0         0         0          1         0
## SalePrice             0         0         0          0         1
```

```r
zapsmall(train_sub %*% train_sub_inv) == zapsmall(train_sub_inv %*% train_sub)
```

```
##             OverallQual X1stFlrSF GrLivArea GarageArea SalePrice
## OverallQual        TRUE      TRUE      TRUE       TRUE      TRUE
## X1stFlrSF          TRUE      TRUE      TRUE       TRUE      TRUE
## GrLivArea          TRUE      TRUE      TRUE       TRUE      TRUE
## GarageArea         TRUE      TRUE      TRUE       TRUE      TRUE
## SalePrice          TRUE      TRUE      TRUE       TRUE      TRUE
```

LU Decomposition

Finally we will perform LU decomposition. Because LU = A by multiplying our lower and upper triangualar matrices it will return our original correlation matrix.

```r
train_sub_lu <- lu(train_sub)
elu <- expand(train_sub_lu)
elu$L
```

```
## 5 x 5 Matrix of class "dtrMatrix" (unitriangular)
##      [,1]       [,2]       [,3]       [,4]       [,5]      
## [1,] 1.00000000          .          .          .          .
## [2,] 0.47622383 1.00000000          .          .          .
## [3,] 0.59300743 0.36680770 1.00000000          .          .
## [4,] 0.56202176 0.28728709 0.09963861 1.00000000          .
## [5,] 0.79098160 0.29638474 0.28569464 0.15864262 1.00000000
```

```r
elu$U
```

```
## 5 x 5 Matrix of class "dtrMatrix"
##      [,1]       [,2]       [,3]       [,4]       [,5]      
## [1,] 1.00000000 0.47622383 0.59300743 0.56202176 0.79098160
## [2,]          . 0.77321086 0.28361970 0.22213350 0.22916790
## [3,]          .          . 0.54430830 0.05423412 0.15550596
## [4,]          .          .          . 0.61491165 0.09755119
## [5,]          .          .          .          . 0.24652324
```

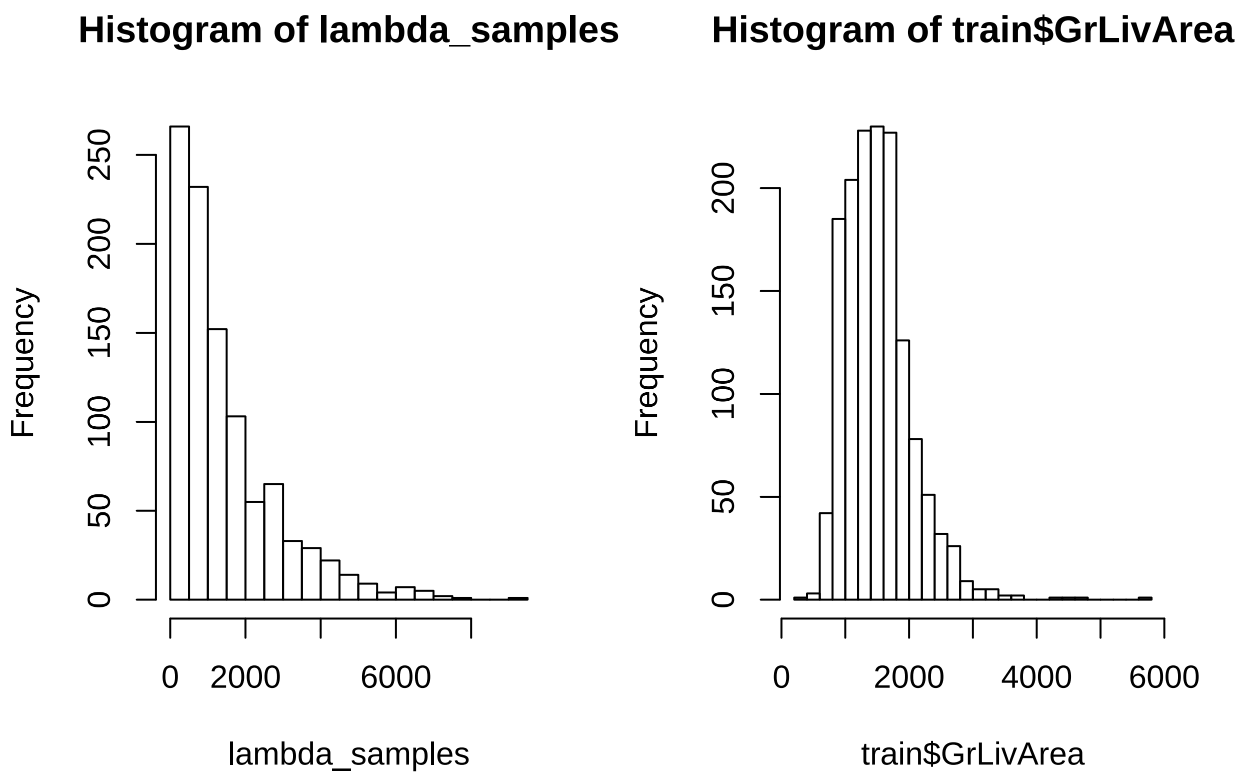
```r
elu$L %*% elu$U == train_sub
```

```
## 5 x 5 Matrix of class "lgeMatrix"
##      [,1] [,2] [,3] [,4] [,5]
## [1,] TRUE TRUE TRUE TRUE TRUE
## [2,] TRUE TRUE TRUE TRUE TRUE
## [3,] TRUE TRUE TRUE TRUE TRUE
## [4,] TRUE TRUE TRUE TRUE TRUE
## [5,] TRUE TRUE TRUE TRUE TRUE
```

Calculus-Based Probability & Statistics

Exponential Distribution

```
fd <- fitdistr(train$GrLivArea, "exponential")
lambda_samples <- rexp(1000, fd$estimate)
par(mfrow=c(1,2))
hist(lambda_samples, breaks=20)
hist(train$GrLivArea, breaks=20)
```



5th and 95th Percentiles

Cumulative Distribution Function

```
qexp(c(0.05, 0.95), rate = fd$estimate)
```

```
## [1]   77.73313 4539.92351
```

Confidence Interval

```
emp <- scale(train$GrLivArea)
me <- qnorm(0.975) * (sd(emp)) / sqrt(length(emp))
c(1 - me, 1 + me)
```

```
## [1] 0.9487054 1.0512946
```

Empirical Distribution

```
quantile(train$GrLivArea, c(0.05, 0.95))
```

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

Regression Model

```
train <- read.csv('https://raw.githubusercontent.com/willoutcault/DATA605_Final/master/train.csv', TRUE
test <- read.csv('https://raw.githubusercontent.com/willoutcault/DATA605_Final/master/test.csv', TRUE,

glimpse(test)
```

```
## Observations: 1,459
## Variables: 80
## $ Id            <int> 1461, 1462, 1463, 1464, 1465, 1466, 1467, 1468, 1469,...
## $ MSSubClass    <int> 20, 20, 60, 60, 120, 60, 20, 60, 20, 20, 120, 160, 16...
## $ MSZoning      <fct> RH, RL, RL, RL, RL, RL, RL, RL, RL, RL, RH, RM, RM, R...
## $ LotFrontage   <int> 80, 81, 74, 78, 43, 75, NA, 63, 85, 70, 26, 21, 21, 2...
## $ LotArea       <int> 11622, 14267, 13830, 9978, 5005, 10000, 7980, 8402, 1...
## $ Street        <fct> Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave, Pave,...
## $ Alley         <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ LotShape      <fct> Reg, IR1, IR1, IR1, IR1, IR1, IR1, IR1, Reg, Reg, IR1...
## $ LandContour   <fct> Lvl, Lvl, Lvl, Lvl, HLS, Lvl, Lvl, Lvl, Lvl, Lvl, Lvl...
## $ Utilities     <fct> AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllPu...
## $ LotConfig     <fct> Inside, Corner, Inside, Inside, Inside, Corner, Insid...
## $ LandSlope     <fct> Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl, Gtl...
## $ Neighborhood  <fct> NAmes, NAmes, Gilbert, Gilbert, StoneBr, Gilbert, Gil...
## $ Condition1    <fct> Feedr, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm...
## $ Condition2    <fct> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm,...
## $ BldgType      <fct> 1Fam, 1Fam, 1Fam, 1Fam, TwnhsE, 1Fam, 1Fam, 1Fam, 1Fa...
## $ HouseStyle    <fct> 1Story, 1Story, 2Story, 2Story, 1Story, 2Story, 1Stor...
## $ OverallQual   <int> 5, 6, 5, 6, 8, 6, 6, 6, 7, 4, 7, 6, 5, 6, 7, 9, 8, 9,...
## $ OverallCond   <int> 6, 6, 5, 6, 5, 5, 7, 5, 5, 5, 5, 5, 5, 6, 6, 5, 5, 5,...
## $ YearBuilt     <int> 1961, 1958, 1997, 1998, 1992, 1993, 1992, 1998, 1990,...
## $ YearRemodAdd  <int> 1961, 1958, 1998, 1998, 1992, 1994, 2007, 1998, 1990,...
## $ RoofStyle     <fct> Gable, Hip, Gable, Gable, Gable, Gable, Gable, Gable,...
## $ RoofMatl      <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg,...
## $ Exterior1st   <fct> VinylSd, Wd Sdng, VinylSd, VinylSd, HdBoard, HdBoard,...
## $ Exterior2nd   <fct> VinylSd, Wd Sdng, VinylSd, VinylSd, HdBoard, HdBoard,...
## $ MasVnrType    <fct> None, BrkFace, None, BrkFace, None, None, None, None,...
## $ MasVnrArea    <int> 0, 108, 0, 20, 0, 0, 0, 0, 0, 0, 0, 504, 492, 0, 0, 1...
## $ ExterQual     <fct> TA, TA, TA, TA, Gd, TA, TA, TA, TA, TA, Gd, TA, TA, T...
## $ ExterCond     <fct> TA, TA, TA, TA, TA, TA, Gd, TA, TA, TA, TA, TA, TA, T...
## $ Foundation    <fct> CBlock, CBlock, PConc, PConc, PConc, PConc, PConc, PC...
## $ BsmtQual      <fct> TA, TA, Gd, TA, Gd, Gd, Gd, Gd, Gd, TA, Gd, TA, TA, T...
## $ BsmtCond      <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ BsmtExposure  <fct> No, No, No, No, No, No, No, No, Gd, No, No, No, No, N...
## $ BsmtFinType1  <fct> Rec, ALQ, GLQ, GLQ, ALQ, Unf, ALQ, Unf, GLQ, ALQ, GLQ...
## $ BsmtFinSF1    <int> 468, 923, 791, 602, 263, 0, 935, 0, 637, 804, 1051, 1...
## $ BsmtFinType2  <fct> LwQ, Unf, Unf, Unf, Unf, Unf, Unf, Unf, Unf, Rec, BLQ...
## $ BsmtFinSF2    <int> 144, 0, 0, 0, 0, 0, 0, 0, 0, 78, 0, 0, 0, 0, 0, 0,...
## $ BsmtUnfSF     <int> 270, 406, 137, 324, 1017, 763, 233, 789, 663, 0, 354,...
## $ TotalBsmtSF   <int> 882, 1329, 928, 926, 1280, 763, 1168, 789, 1300, 882,...
## $ Heating       <fct> GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA, GasA,...
## $ HeatingQC     <fct> TA, TA, Gd, Ex, Ex, Gd, Ex, Gd, Gd, TA, Ex, TA, TA, T...
## $ CentralAir    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ Electrical    <fct> SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrkr, SBrk...
## $ X1stFlrSF     <int> 896, 1329, 928, 926, 1280, 763, 1187, 789, 1341, 882,...
## $ X2ndFlrSF     <int> 0, 0, 701, 678, 0, 892, 0, 676, 0, 0, 0, 504, 567, 60...
## $ LowQualFinSF  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ GrLivArea     <int> 896, 1329, 1629, 1604, 1280, 1655, 1187, 1465, 1341, ...
## $ BsmtFullBath  <int> 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,...
## $ BsmtHalfBath  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ FullBath      <int> 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 2,...
## $ HalfBath      <int> 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0,...
## $ BedroomAbvGr  <int> 2, 3, 3, 3, 2, 3, 3, 3, 2, 2, 2, 2, 3, 3, 2, 3, 3, 3,...
## $ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ KitchenQual   <fct> TA, Gd, TA, Gd, Gd, TA, TA, TA, Gd, TA, Gd, TA, TA, G...
## $ TotRmsAbvGrd  <int> 5, 6, 6, 7, 5, 7, 6, 7, 5, 4, 5, 5, 6, 6, 4, 10, 7, 7...
## $ Functional    <fct> Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ...
## $ Fireplaces    <int> 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1,...
## $ FireplaceQu   <fct> NA, NA, TA, Gd, NA, TA, NA, Gd, Po, NA, Fa, NA, NA, T...
## $ GarageType    <fct> Attchd, Attchd, Attchd, Attchd, Attchd, Attchd, Attch...
## $ GarageYrBlt   <int> 1961, 1958, 1997, 1998, 1992, 1993, 1992, 1998, 1990,...
## $ GarageFinish  <fct> Unf, Unf, Fin, Fin, RFn, Fin, Fin, Fin, Unf, Fin, Fin...
## $ GarageCars    <int> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 3, 3, 3,...
## $ GarageArea    <int> 730, 312, 482, 470, 506, 440, 420, 393, 506, 525, 511...
## $ GarageQual    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ GarageCond    <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, T...
## $ PavedDrive    <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,...
## $ WoodDeckSF    <int> 140, 393, 212, 360, 0, 157, 483, 0, 192, 240, 203, 27...
## $ OpenPorchSF   <int> 0, 36, 34, 36, 82, 84, 21, 75, 0, 0, 68, 0, 0, 0, 30,...
## $ EnclosedPorch <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ X3SsnPorch    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ScreenPorch   <int> 120, 0, 0, 0, 144, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ PoolQC        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Fence         <fct> MnPrv, NA, MnPrv, NA, NA, NA, GdPrv, NA, NA, MnPrv, N...
## $ MiscFeature   <fct> NA, Gar2, NA, NA, NA, NA, Shed, NA, NA, NA, NA, NA, N...
## $ MiscVal       <int> 0, 12500, 0, 0, 0, 0, 500, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ MoSold        <int> 6, 6, 3, 6, 1, 4, 3, 5, 2, 4, 6, 2, 3, 6, 6, 1, 6, 6,...
## $ YrSold        <int> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,...
## $ SaleType      <fct> WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, COD, WD, ...
## $ SaleCondition <fct> Normal, Normal, Normal, Normal, Normal, Normal, Norma...
```

```
summary(test)
```

```
##        Id          MSSubClass        MSZoning     LotFrontage
##  Min.   :1461   Min.   : 20.00   C (all):  15   Min.   : 21.00
##  1st Qu.:1826   1st Qu.: 20.00   FV     :  74   1st Qu.: 58.00
##  Median :2190   Median : 50.00   RH     :  10   Median : 67.00
##  Mean   :2190   Mean   : 57.38   RL     :1114   Mean   : 68.58
##  3rd Qu.:2554   3rd Qu.: 70.00   RM     : 242   3rd Qu.: 80.00
##  Max.   :2919   Max.   :190.00   NA's   :   4   Max.   :200.00
##                                                 NA's   :227
##     LotArea       Street      Alley      LotShape  LandContour  Utilities
##  Min.   : 1470   Grvl:   6   Grvl:  70   IR1:484   Bnk:  54    AllPub:1457
##  1st Qu.: 7391   Pave:1453   Pave:  37   IR2: 35   HLS:  70    NA's  :   2
##  Median : 9399               NA's:1352   IR3:  6   Low:  24
##  Mean   : 9819                           Reg:934   Lvl:1311
##  3rd Qu.:11518
##  Max.   :56600
##
##    LotConfig    LandSlope   Neighborhood   Condition1    Condition2
##  Corner : 248   Gtl:1396   NAmes  :218   Norm   :1251   Artery:   3
##  CulDSac:  82   Mod:  60   OldTown:126   Feedr  :  83   Feedr :   7
##  FR2    :  38   Sev:   3   CollgCr:117   Artery :  44   Norm  :1444
##  FR3    :  10              Somerst: 96   RRAn   :  24   PosA  :   3
##  Inside :1081              Edwards: 94   PosN   :  20   PosN  :   2
##                            NridgHt: 89   RRAe   :  17
##                            (Other):719   (Other):  20
##    BldgType     HouseStyle    OverallQual      OverallCond      YearBuilt
##  1Fam  :1205   1.5Fin:160   Min.   : 1.000   Min.   :1.000   Min.   :1879
##  2fmCon:  31   1.5Unf:  5   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1953
##  Duplex:  57   1Story:745   Median : 6.000   Median :5.000   Median :1973
##  Twnhs :  53   2.5Unf: 13   Mean   : 6.079   Mean   :5.554   Mean   :1971
##  TwnhsE: 113   2Story:427   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001
##                SFoyer: 46   Max.   :10.000   Max.   :9.000   Max.   :2010
##                SLvl  : 63
##   YearRemodAdd    RoofStyle       RoofMatl     Exterior1st   Exterior2nd
##  Min.   :1950   Flat   :   7   CompShg:1442   VinylSd:510   VinylSd:510
##  1st Qu.:1963   Gable  :1169   Tar&Grv:  12   MetalSd:230   MetalSd:233
##  Median :1992   Gambrel:  11   WdShake:   4   HdBoard:220   HdBoard:199
##  Mean   :1984   Hip    : 265   WdShngl:   1   Wd Sdng:205   Wd Sdng:194
##  3rd Qu.:2004   Mansard:   4                  Plywood:113   Plywood:128
##  Max.   :2010   Shed   :   3                  (Other):180   (Other):194
##                                               NA's   :  1   NA's   :  1
##    MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual
##  BrkCmn : 10   Min.   :   0.0   Ex: 55    Ex:   9   BrkTil:165   Ex  :137
##  BrkFace:434   1st Qu.:   0.0   Fa: 21    Fa:  39   CBlock:601   Fa  : 53
##  None   :878   Median :   0.0   Gd:491    Gd: 153   PConc :661   Gd  :591
##  Stone  :121   Mean   : 100.7   TA:892    Po:   2   Slab  : 25   TA  :634
##  NA's   : 16   3rd Qu.: 164.0             TA:1256   Stone :  5   NA's: 44
##                Max.   :1290.0                       Wood  :  2
##                NA's   :15
##  BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2
##  Fa  :  59   Av  :197     ALQ :209     Min.   :   0.0   ALQ :  33
##  Gd  :  57   Gd  :142     BLQ :121     1st Qu.:   0.0   BLQ :  35
##  Po  :   3   Mn  :125     GLQ :431     Median : 350.5   GLQ :  20
##  TA  :1295   No  :951     LwQ : 80     Mean   : 439.2   LwQ :  41
##  NA's:  45   NA's: 44     Rec :155     3rd Qu.: 753.5   Rec :  51
##                           Unf :421     Max.   :4010.0   Unf :1237
##                           NA's: 42     NA's   :1        NA's:  42
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF   Heating     HeatingQC
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0   GasA:1446   Ex:752
##  1st Qu.:   0.00   1st Qu.: 219.2   1st Qu.: 784   GasW:   9   Fa: 43
##  Median :   0.00   Median : 460.0   Median : 988   Grav:   2   Gd:233
##  Mean   :  52.62   Mean   : 554.3   Mean   :1046   Wall:   2   Po:  2
##  3rd Qu.:   0.00   3rd Qu.: 797.8   3rd Qu.:1305               TA:429
##  Max.   :1526.00   Max.   :2140.0   Max.   :5095
##  NA's   :1         NA's   :1        NA's   :1
##  CentralAir Electrical     X1stFlrSF        X2ndFlrSF     LowQualFinSF
##  N: 101     FuseA:  94   Min.   : 407.0   Min.   :   0   Min.   :   0.000
##  Y:1358     FuseF:  23   1st Qu.: 873.5   1st Qu.:   0   1st Qu.:   0.000
##             FuseP:   5   Median :1079.0   Median :   0   Median :   0.000
##             SBrkr:1337   Mean   :1156.5   Mean   : 326   Mean   :   3.543
##                          3rd Qu.:1382.5   3rd Qu.: 676   3rd Qu.:   0.000
##                          Max.   :5095.0   Max.   :1862   Max.   :1064.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath       FullBath
##  Min.   : 407   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:1118   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :1432   Median :0.0000   Median :0.0000   Median :2.000
##  Mean   :1486   Mean   :0.4345   Mean   :0.0652   Mean   :1.571
##  3rd Qu.:1721   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:2.000
##  Max.   :5095   Max.   :3.0000   Max.   :2.0000   Max.   :4.000
##                 NA's   :2        NA's   :2
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex  :105    Min.   : 3.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa  : 31    1st Qu.: 5.000
##  Median :0.0000   Median :3.000   Median :1.000   Gd  :565    Median : 6.000
##  Mean   :0.3777   Mean   :2.854   Mean   :1.042   TA  :757    Mean   : 6.385
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   NA's:  1    3rd Qu.: 7.000
##  Max.   :2.0000   Max.   :6.000   Max.   :2.000               Max.   :15.000
##
##    Functional     Fireplaces     FireplaceQu   GarageType   GarageYrBlt
##  Typ    :1357   Min.   :0.0000   Ex  : 19    2Types : 17   Min.   :1895
##  Min2   :  36   1st Qu.:0.0000   Fa  : 41    Attchd :853   1st Qu.:1959
##  Min1   :  34   Median :0.0000   Gd  :364    Basment: 17   Median :1979
##  Mod    :  20   Mean   :0.5812   Po  : 26    BuiltIn: 98   Mean   :1978
##  Maj1   :   5   3rd Qu.:1.0000   TA  :279    CarPort:  6   3rd Qu.:2002
##  (Other):   5   Max.   :4.0000   NA's:730    Detchd :392   Max.   :2207
##  NA's   :   2                                NA's   : 76   NA's   :78
##  GarageFinish   GarageCars      GarageArea     GarageQual  GarageCond
##  Fin :367     Min.   :0.000   Min.   :   0.0   Fa  :  76   Ex  :   1
##  RFn :389     1st Qu.:1.000   1st Qu.: 318.0   Gd  :  10   Fa  :  39
##  Unf :625     Median :2.000   Median : 480.0   Po  :   2   Gd  :   6
##  NA's: 78     Mean   :1.766   Mean   : 472.8   TA  :1293   Po  :   7
##               3rd Qu.:2.000   3rd Qu.: 576.0   NA's:  78   TA  :1328
##               Max.   :5.000   Max.   :1488.0               NA's:  78
##               NA's   :1       NA's   :1
##  PavedDrive   WoodDeckSF       OpenPorchSF     EnclosedPorch
##  N: 126     Min.   :   0.00   Min.   :  0.00   Min.   :   0.00
##  P:  32     1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00
##  Y:1301     Median :   0.00   Median : 28.00   Median :   0.00
##             Mean   :  93.17   Mean   : 48.31   Mean   :  24.24
##             3rd Qu.: 168.00   3rd Qu.: 72.00   3rd Qu.:   0.00
##             Max.   :1424.00   Max.   :742.00   Max.   :1012.00
##
##    X3SsnPorch       ScreenPorch        PoolArea        PoolQC       Fence
##  Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Ex  :   2   GdPrv:  59
##  1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   Gd  :   1   GdWo :  58
##  Median :  0.000   Median :  0.00   Median :  0.000   NA's:1456   MnPrv: 172
##  Mean   :  1.794   Mean   : 17.06   Mean   :  1.744               MnWw :   1
##  3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000               NA's :1169
##  Max.   :360.000   Max.   :576.00   Max.   :800.000
##
##  MiscFeature    MiscVal            MoSold          YrSold        SaleType
##  Gar2:   3   Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :1258
##  Othr:   2   1st Qu.:    0.00   1st Qu.: 4.000   1st Qu.:2007   New    : 117
##  Shed:  46   Median :    0.00   Median : 6.000   Median :2008   COD    :  44
##  NA's:1408   Mean   :   58.17   Mean   : 6.104   Mean   :2008   ConLD  :  17
##              3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   CWD    :   8
##              Max.   :17000.00   Max.   :12.000   Max.   :2010   (Other):  14
##                                                                 NA's   :   1
##  SaleCondition
##  Abnorml:  89
##  AdjLand:   8
##  Alloca :  12
##  Family :  26
##  Normal :1204
##  Partial: 120
##
```

Formatting the Data

```r
train$SalePrice <- log(train$SalePrice)
test$SalePrice <- 0

asNumeric <- function(x) as.numeric(factor(x))
factorsNumeric <- function(d) modifyList(d, lapply(d[, sapply(d, is.factor)],
                                                    asNumeric))

train <- factorsNumeric(train)
test <- factorsNumeric(test)

train[is.na(train)] <- 0
test[is.na(test)] <- 0

anyNA(train)
```

```
## [1] FALSE
```

```r
anyNA(test)
```

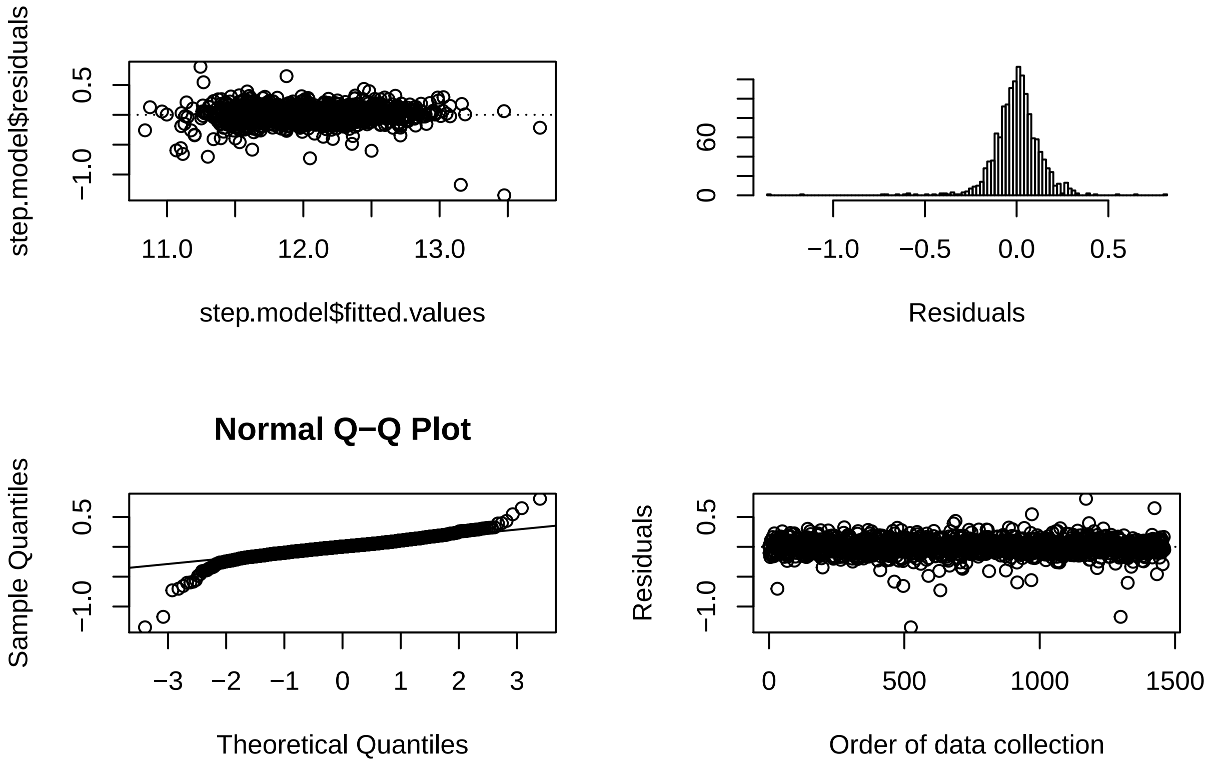
```
## [1] FALSE
```

Training Model

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

step.model <- stepAIC(full.model, direction = "forward",
                      trace = FALSE)
m = summary(step.model)

m$adj.r.squared
```

```
## [1] 0.8868161
```

```r
par(mfrow=c(2,2))

# residuals plot ----------------------------------------------------

plot(step.model$residuals ~ step.model$fitted.values)
abline(h = 0, lty = 3)

# residuals histogram -----------------------------------------------

hist(step.model$residuals,
     xlab = "Residuals", ylab = "", main = "", breaks = 85,
     xlim = c(min(step.model$residuals), max(step.model$residuals)))

# normal probability plot of residuals ------------------------------

qqnorm(step.model$residuals)
qqline(step.model$residuals)

# order of residuals ---------------------------------------------===

plot(step.model$residuals,
     xlab = "Order of data collection", ylab = "Residuals", main = "")
abline(h = 0, lty = 3)
```



```
predictions <- predict(step.model, test, na.action=na.pass)

predictions <- exp(predictions)
```

Kaggle

Kaggle Name: Will Outcault Kaggle Score: 0.14318