605 Final

Jose Mawyin

12/14/2019

Final: 605 - Computational Mathematics

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 $\mu = \sigma = (N + 1)/2$.

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.

```
## 'data.frame':    10000 obs. of  3 variables:
##  $ matrix.NA..nrow...10000..ncol...1.: logi  NA NA NA NA NA NA ...
##  $ X                                 : num  5.57 5.69 2.43 5.15 4.21 ...
##  $ Y                                 : num  3.57 4.47 3.81 3.36 3.17 ...
```

```
## The mean of Y is 3.491478
## The median of X or x is 3
## The 1st quartile of the Y or y is 2.815439
```

5 points

a and c are examples of Conditional Probability where:

Conditional Probability: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

b is an example of Joint Probability where:

Joint Probability: $P(A, B) = P(B)P(A|B)$

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

```
## The probability of a is: 0.9427403
```

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

```
## The probability of b is: 0.449475
```

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

```
## The probability of c is: 0.05725971
```

5 points. 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.

A = P(X > x and Y > y)

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

Let's create a contingency table:

```
##       Y < y  Y > y total
## X < x 989    3018  4007
## X > x 1511   4482  5993
## total 2500   7500  10000
```

Let's see if A = B

```
## The probability of A is: 0.4482
## The probability of B is: 0.449475
```

Therefore A is not equal to B

5 points. 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?

Let's see if there is independency between the generated X and Y values using:

Pearson's Chi-squared test

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datatable
## X-squared = 0.33328, df = 1, p-value = 0.5637
```

Fisher's Exact Test

```
##
##  Fisher's Exact Test for Count Data
##
## data:  datatable
## p-value = 0.5558
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8850799 1.0673027
## sample estimates:
## odds ratio
##  0.9720388
```

H_O : Variable A and Variable B are independent.
H_a : Variable A and Variable B are not independent.

For both test we find that the p-value is close to 0.5 that is a lot larger than typical tresholds of 0.05 or 0.1, we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between the two generated variables X and Y.

Both test are appropiate in the case of our small sized contingency table. Fisher's Test performs worse with large datasets. Not in this case.

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.

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?

First, let's see the structure and statistics summary of our dataset:

```
## '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 ...
```

```
describe(house_train)
```

```
##                vars    n      mean       sd   median   trimmed      mad
## Id                1 1460    730.50   421.61    730.5    730.50   541.15
## MSSubClass        2 1460     56.90    42.30     50.0     49.15    44.48
## MSZoning*         3 1460      4.03     0.63      4.0      4.06     0.00
## LotFrontage       4 1201     70.05    24.28     69.0     68.94    16.31
## LotArea           5 1460  10516.83  9981.26   9478.5   9563.28  2962.23
## Street*           6 1460      2.00     0.06      2.0      2.00     0.00
```

```
## Alley*           7   91     1.45     0.50     1.0     1.44     0.00
## LotShape*        8 1460     2.94     1.41     4.0     3.05     0.00
## LandContour*     9 1460     3.78     0.71     4.0     4.00     0.00
## Utilities*      10 1460     1.00     0.03     1.0     1.00     0.00
## LotConfig*      11 1460     4.02     1.62     5.0     4.27     0.00
## LandSlope*      12 1460     1.06     0.28     1.0     1.00     0.00
## Neighborhood*   13 1460    13.15     5.89    13.0    13.11     7.41
## Condition1*     14 1460     3.03     0.87     3.0     3.00     0.00
## Condition2*     15 1460     3.01     0.26     3.0     3.00     0.00
## BldgType*       16 1460     1.49     1.20     1.0     1.14     0.00
## HouseStyle*     17 1460     4.04     1.91     3.0     4.03     1.48
## OverallQual     18 1460     6.10     1.38     6.0     6.08     1.48
## OverallCond     19 1460     5.58     1.11     5.0     5.48     0.00
## YearBuilt       20 1460  1971.27    30.20  1973.0  1974.13    37.06
## YearRemodAdd    21 1460  1984.87    20.65  1994.0  1986.37    19.27
## RoofStyle*      22 1460     2.41     0.83     2.0     2.26     0.00
## RoofMatl*       23 1460     2.08     0.60     2.0     2.00     0.00
## Exterior1st*    24 1460    10.62     3.20    13.0    10.93     1.48
## Exterior2nd*    25 1460    11.34     3.54    14.0    11.65     2.97
## MasVnrType*     26 1452     2.76     0.62     3.0     2.73     0.00
## MasVnrArea      27 1452   103.69   181.07     0.0    63.15     0.00
## ExterQual*      28 1460     3.54     0.69     4.0     3.65     0.00
## ExterCond*      29 1460     4.73     0.73     5.0     4.95     0.00
## Foundation*     30 1460     2.40     0.72     2.0     2.46     1.48
## BsmtQual*       31 1423     3.26     0.87     3.0     3.43     1.48
## BsmtCond*       32 1423     3.81     0.66     4.0     4.00     0.00
## BsmtExposure*   33 1422     3.27     1.15     4.0     3.46     0.00
## BsmtFinType1*   34 1423     3.73     1.83     3.0     3.79     2.97
## BsmtFinSF1      35 1460   443.64   456.10   383.5   386.08   568.58
## BsmtFinType2*   36 1422     5.71     0.94     6.0     5.98     0.00
## BsmtFinSF2      37 1460    46.55   161.32     0.0     1.38     0.00
## BsmtUnfSF       38 1460   567.24   441.87   477.5   519.29   426.99
## TotalBsmtSF     39 1460  1057.43   438.71   991.5  1036.70   347.67
## Heating*        40 1460     2.04     0.30     2.0     2.00     0.00
## HeatingQC*      41 1460     2.54     1.74     1.0     2.42     0.00
## CentralAir*     42 1460     1.93     0.25     2.0     2.00     0.00
## Electrical*     43 1459     4.68     1.05     5.0     5.00     0.00
## X1stFlrSF       44 1460  1162.63   386.59  1087.0  1129.99   347.67
## X2ndFlrSF       45 1460   346.99   436.53     0.0   285.36     0.00
## LowQualFinSF    46 1460     5.84    48.62     0.0     0.00     0.00
## GrLivArea       47 1460  1515.46   525.48  1464.0  1467.67   483.33
## BsmtFullBath    48 1460     0.43     0.52     0.0     0.39     0.00
## BsmtHalfBath    49 1460     0.06     0.24     0.0     0.00     0.00
## FullBath        50 1460     1.57     0.55     2.0     1.56     0.00
## HalfBath        51 1460     0.38     0.50     0.0     0.34     0.00
## BedroomAbvGr    52 1460     2.87     0.82     3.0     2.85     0.00
## KitchenAbvGr    53 1460     1.05     0.22     1.0     1.00     0.00
## KitchenQual*    54 1460     3.34     0.83     4.0     3.50     0.00
## TotRmsAbvGrd    55 1460     6.52     1.63     6.0     6.41     1.48
## Functional*     56 1460     6.75     0.98     7.0     7.00     0.00
## Fireplaces      57 1460     0.61     0.64     1.0     0.53     1.48
## FireplaceQu*    58  770     3.73     1.13     3.0     3.80     1.48
## GarageType*     59 1379     3.28     1.79     2.0     3.11     0.00
## GarageYrBlt     60 1379  1978.51    24.69  1980.0  1981.07    31.13
```

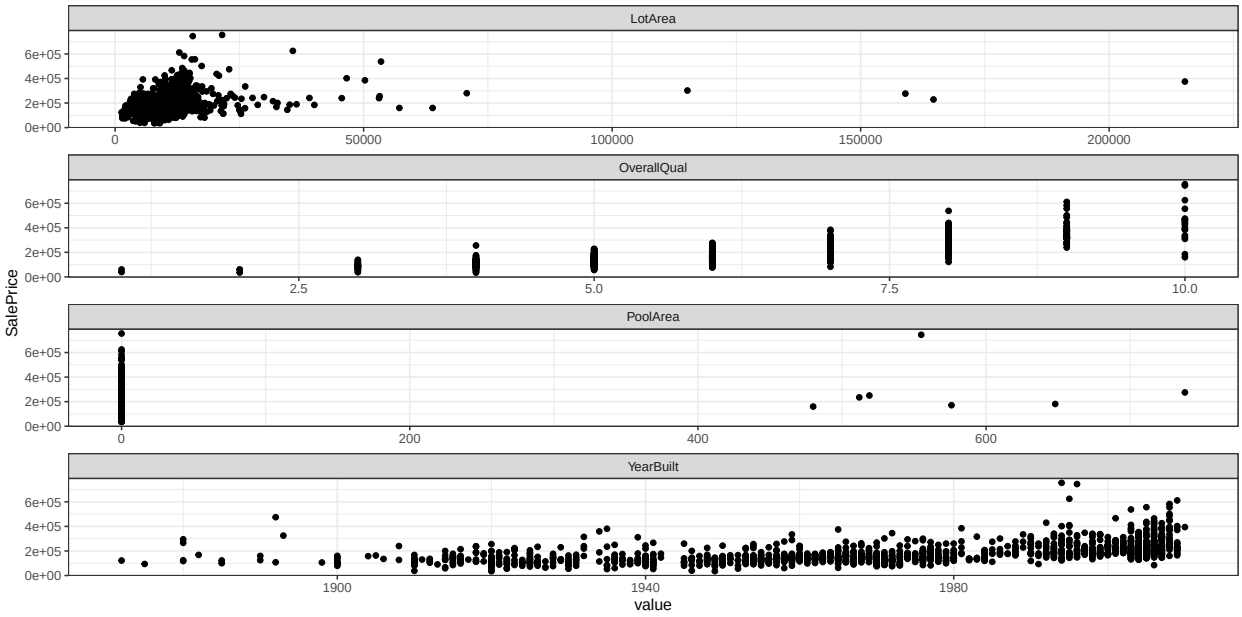
```
## GarageFinish*    61 1379      2.18     0.81      2.0      2.23     1.48
## GarageCars       62 1460      1.77     0.75      2.0      1.77     0.00
## GarageArea       63 1460    472.98   213.80    480.0    469.81   177.91
## GarageQual*      64 1379      4.86     0.61      5.0      5.00     0.00
## GarageCond*      65 1379      4.90     0.52      5.0      5.00     0.00
## PavedDrive*      66 1460      2.86     0.50      3.0      3.00     0.00
## WoodDeckSF       67 1460     94.24   125.34      0.0     71.76     0.00
## OpenPorchSF      68 1460     46.66    66.26     25.0     33.23    37.06
## EnclosedPorch    69 1460     21.95    61.12      0.0      3.87     0.00
## X3SsnPorch       70 1460      3.41    29.32      0.0      0.00     0.00
## ScreenPorch      71 1460     15.06    55.76      0.0      0.00     0.00
## PoolArea         72 1460      2.76    40.18      0.0      0.00     0.00
## PoolQC*          73    7      2.14     0.90      2.0      2.14     1.48
## Fence*           74  281      2.43     0.86      3.0      2.48     0.00
## MiscFeature*     75   54      2.91     0.45      3.0      3.00     0.00
## MiscVal          76 1460     43.49   496.12      0.0      0.00     0.00
## MoSold           77 1460      6.32     2.70      6.0      6.25     2.97
## YrSold           78 1460   2007.82     1.33   2008.0   2007.77     1.48
## SaleType*        79 1460      8.51     1.56      9.0      8.92     0.00
## SaleCondition*   80 1460      4.77     1.10      5.0      5.00     0.00
## SalePrice        81 1460 180921.20 79442.50 163000.0 170783.29 56338.80
##                   min    max  range   skew kurtosis      se
## Id                  1   1460   1459   0.00    -1.20   11.03
## MSSubClass         20    190    170   1.40     1.56    1.11
## MSZoning*           1      5      4  -1.73     6.25    0.02
## LotFrontage        21    313    292   2.16    17.34    0.70
## LotArea          1300 215245 213945  12.18   202.26  261.22
## Street*             1      2      1 -15.49   238.01    0.00
## Alley*              1      2      1   0.20    -1.98    0.05
## LotShape*           1      4      3  -0.61    -1.60    0.04
## LandContour*        1      4      3  -3.16     8.65    0.02
## Utilities*          1      2      1  38.13  1453.00    0.00
## LotConfig*          1      5      4  -1.13    -0.59    0.04
## LandSlope*          1      3      2   4.80    24.47    0.01
## Neighborhood*       1     25     24   0.02    -1.06    0.15
## Condition1*         1      9      8   3.01    16.34    0.02
## Condition2*         1      8      7  13.14   247.54    0.01
## BldgType*           1      5      4   2.24     3.41    0.03
## HouseStyle*         1      8      7   0.31    -0.96    0.05
## OverallQual         1     10      9   0.22     0.09    0.04
## OverallCond         1      9      8   0.69     1.09    0.03
## YearBuilt        1872   2010    138  -0.61    -0.45    0.79
## YearRemodAdd     1950   2010     60  -0.50    -1.27    0.54
## RoofStyle*          1      6      5   1.47     0.61    0.02
## RoofMatl*           1      8      7   8.09    66.28    0.02
## Exterior1st*        1     15     14  -0.72    -0.37    0.08
## Exterior2nd*        1     16     15  -0.69    -0.52    0.09
## MasVnrType*         1      4      3  -0.07    -0.13    0.02
## MasVnrArea          0   1600   1600   2.66    10.03    4.75
## ExterQual*          1      4      3  -1.83     3.86    0.02
## ExterCond*          1      5      4  -2.56     5.29    0.02
## Foundation*         1      6      5   0.09     1.02    0.02
## BsmtQual*           1      4      3  -1.31     1.27    0.02
## BsmtCond*           1      4      3  -3.39    10.14    0.02
```

```
## BsmtExposure*      1      4      3  -1.15    -0.39    0.03
## BsmtFinType1*      1      6      5  -0.02    -1.39    0.05
## BsmtFinSF1         0   5644   5644   1.68    11.06   11.94
## BsmtFinType2*      1      6      5  -3.56    12.32    0.02
## BsmtFinSF2         0   1474   1474   4.25    20.01    4.22
## BsmtUnfSF          0   2336   2336   0.92     0.46   11.56
## TotalBsmtSF        0   6110   6110   1.52    13.18   11.48
## Heating*           1      6      5   9.83   110.98    0.01
## HeatingQC*         1      5      4   0.48    -1.51    0.05
## CentralAir*        1      2      1  -3.52    10.42    0.01
## Electrical*        1      5      4  -3.06     7.49    0.03
## X1stFlrSF        334   4692   4358   1.37     5.71   10.12
## X2ndFlrSF          0   2065   2065   0.81    -0.56   11.42
## LowQualFinSF       0    572    572   8.99    82.83    1.27
## GrLivArea        334   5642   5308   1.36     4.86   13.75
## BsmtFullBath       0      3      3   0.59    -0.84    0.01
## BsmtHalfBath       0      2      2   4.09    16.31    0.01
## FullBath           0      3      3   0.04    -0.86    0.01
## HalfBath           0      2      2   0.67    -1.08    0.01
## BedroomAbvGr       0      8      8   0.21     2.21    0.02
## KitchenAbvGr       0      3      3   4.48    21.42    0.01
## KitchenQual*       1      4      3  -1.42     1.72    0.02
## TotRmsAbvGrd       2     14     12   0.67     0.87    0.04
## Functional*        1      7      6  -4.08    16.37    0.03
## Fireplaces         0      3      3   0.65    -0.22    0.02
## FireplaceQu*       1      5      4  -0.16    -0.98    0.04
## GarageType*        1      6      5   0.76    -1.30    0.05
## GarageYrBlt     1900   2010    110  -0.65    -0.42    0.66
## GarageFinish*      1      3      2  -0.35    -1.41    0.02
## GarageCars         0      4      4  -0.34     0.21    0.02
## GarageArea         0   1418   1418   0.18     0.90    5.60
## GarageQual*        1      5      4  -4.43    18.25    0.02
## GarageCond*        1      5      4  -5.28    26.77    0.01
## PavedDrive*        1      3      2  -3.30     9.22    0.01
## WoodDeckSF         0    857    857   1.54     2.97    3.28
## OpenPorchSF        0    547    547   2.36     8.44    1.73
## EnclosedPorch      0    552    552   3.08    10.37    1.60
## X3SsnPorch         0    508    508  10.28   123.06    0.77
## ScreenPorch        0    480    480   4.11    18.34    1.46
## PoolArea           0    738    738  14.80   222.19    1.05
## PoolQC*            1      3      2  -0.22    -1.90    0.34
## Fence*             1      4      3  -0.57    -0.88    0.05
## MiscFeature*       1      4      3  -2.93    10.71    0.06
## MiscVal            0  15500  15500  24.43   697.64   12.98
## MoSold             1     12     11   0.21    -0.41    0.07
## YrSold          2006   2010      4   0.10    -1.19    0.03
## SaleType*          1      9      8  -3.83    14.57    0.04
## SaleCondition*     1      6      5  -2.74     6.82    0.03
## SalePrice      34900 755000 720100   1.88     6.50 2079.11
```

Is is a dataset with a large number of explanatory variables (80) that are provided to find their relevance in in the response variable - SalePrice of a house.

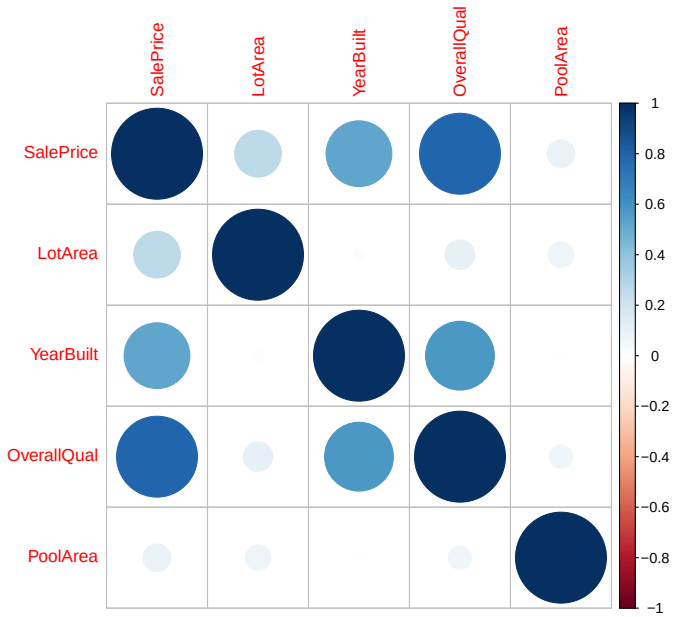
Let's see how some explanatory variables (LotArea,YearBuilt,OverallQual,PoolArea) relate to

SalePrice



Some explanatory variables are discrete whole others are continous.

Let's study the correlation matrix between "SalePrice","OverallQual" and "Fireplaces"



Some of explanatory variables have more statistical significance than others in explaining the response variable.

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.

To find inverse of correlation matrix

```
rownames(corr.matrix) <- c()
colnames(corr.matrix) <- c()
precision.matrix <- solve(corr.matrix)
precision.matrix
```

```
##             [,1]        [,2]        [,3]       [,4]        [,5]
## [1,]  2.97970740 -0.53519673 -0.35409756 -2.0877163 -0.07921778
## [2,] -0.53519673  1.11761624  0.13913329  0.2245489 -0.05696001
## [3,] -0.35409756  0.13913329  1.53244547 -0.6135194  0.06368448
## [4,] -2.08771635  0.22454885 -0.61351942  2.9769937 -0.03845900
## [5,] -0.07921778 -0.05696001  0.06368448 -0.0384590  1.01437853
```

Multiply the correlation matrix by the precision matrix

```
C.product.P <- corr.matrix %*% precision.matrix
C.product.P
```

```
##                 SalePrice  OverallQual Fireplaces
## SalePrice    1.000000e+00 4.163336e-17          0
## OverallQual -3.330669e-16 1.000000e+00          0
## Fireplaces  -1.110223e-16 1.249001e-16          1
```

LU decomposition

```
LU <- expand(lu(C.product.P))
LU
```

```
## $L
## 3 x 3 Matrix of class "dtrMatrix" (unitriangular)
##      [,1]          [,2]          [,3]
## [1,]  1.000000e+00             .             .
## [2,] -3.330669e-16  1.000000e+00             .
## [3,] -1.110223e-16  1.249001e-16  1.000000e+00
##
## $U
## 3 x 3 Matrix of class "dtrMatrix"
##      [,1]         [,2]         [,3]
## [1,] 1.000000e+00 4.163336e-17 0.000000e+00
## [2,]            . 1.000000e+00 0.000000e+00
## [3,]            .            . 1.000000e+00
##
## $P
## 3 x 3 sparse Matrix of class "pMatrix"
##
## [1,] | . .
## [2,] . | .
## [3,] . . |
```

```
L.times.U <- LU[["L"]] %*% LU[["U"]] %>% as.matrix()
L.times.U
```

```
##               [,1]         [,2] [,3]
## [1,]  1.000000e+00 4.163336e-17    0
## [2,] -3.330669e-16 1.000000e+00    0
## [3,] -1.110223e-16 1.249001e-16    1
```

```
C.product.P
```

```
##                 SalePrice  OverallQual Fireplaces
## SalePrice    1.000000e+00 4.163336e-17          0
## OverallQual -3.330669e-16 1.000000e+00          0
## Fireplaces  -1.110223e-16 1.249001e-16          1
```

Multiplying the two component matrices of the LU decomposition we get our initial matrix back.

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.

> First, let's create a dataframe containing only non-factor variables and use these to create a linear regression model with SalePrice as the response variable.

```
##
## Call:
## lm(formula = SalePrice ~ ., data = non_factor_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -442182  -16955   -2824   15125  318183
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.351e+05  1.701e+06  -0.197 0.843909
## Id            -1.205e+00  2.658e+00  -0.453 0.650332
## MSSubClass    -2.001e+02  3.451e+01  -5.797 8.84e-09 ***
## LotFrontage   -1.160e+02  6.126e+01  -1.894 0.058503 .
## LotArea        5.422e-01  1.575e-01   3.442 0.000599 ***
## OverallQual    1.866e+04  1.482e+03  12.592  < 2e-16 ***
## OverallCond    5.239e+03  1.368e+03   3.830 0.000135 ***
## YearBuilt      3.164e+02  8.766e+01   3.610 0.000321 ***
## YearRemodAdd   1.194e+02  8.668e+01   1.378 0.168607
## MasVnrArea     3.141e+01  7.022e+00   4.473 8.54e-06 ***
## BsmtFinSF1     1.736e+01  5.838e+00   2.973 0.003014 **
## BsmtFinSF2     8.342e+00  8.766e+00   0.952 0.341532
## BsmtUnfSF      5.005e+00  5.277e+00   0.948 0.343173
## TotalBsmtSF           NA         NA      NA       NA
## X1stFlrSF      4.597e+01  7.360e+00   6.246 6.02e-10 ***
## X2ndFlrSF      4.663e+01  6.102e+00   7.641 4.72e-14 ***
## LowQualFinSF   3.341e+01  2.794e+01   1.196 0.232009
## GrLivArea             NA         NA      NA       NA
## BsmtFullBath   9.043e+03  3.198e+03   2.828 0.004776 **
## BsmtHalfBath   2.465e+03  5.073e+03   0.486 0.627135
## FullBath       5.433e+03  3.531e+03   1.539 0.124182
## HalfBath      -1.098e+03  3.321e+03  -0.331 0.740945
## BedroomAbvGr  -1.022e+04  2.155e+03  -4.742 2.40e-06 ***
## KitchenAbvGr  -2.202e+04  6.710e+03  -3.282 0.001063 **
## TotRmsAbvGrd   5.464e+03  1.487e+03   3.674 0.000251 ***
## Fireplaces     4.372e+03  2.189e+03   1.998 0.046020 *
## GarageYrBlt   -4.728e+01  9.106e+01  -0.519 0.603742
## GarageCars     1.685e+04  3.491e+03   4.827 1.58e-06 ***
## GarageArea     6.274e+00  1.213e+01   0.517 0.605002
## WoodDeckSF     2.144e+01  1.002e+01   2.139 0.032662 *
## OpenPorchSF   -2.252e+00  1.949e+01  -0.116 0.907998
## EnclosedPorch  7.295e+00  2.062e+01   0.354 0.723590
## X3SsnPorch     3.349e+01  3.758e+01   0.891 0.373163
## ScreenPorch    5.805e+01  2.041e+01   2.844 0.004532 **
## PoolArea      -6.052e+01  2.990e+01  -2.024 0.043204 *
## MiscVal       -3.761e+00  6.960e+00  -0.540 0.589016
## MoSold        -2.217e+02  4.229e+02  -0.524 0.600188
```

```
## YrSold        -2.474e+02  8.458e+02  -0.293 0.769917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36800 on 1085 degrees of freedom
##   (339 observations deleted due to missingness)
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.8034
## F-statistic: 131.8 on 35 and 1085 DF,  p-value: < 2.2e-16
```

Our R-squared value using 37 explanatory varialbles is 0.8095558. Can we do better?

Let's remove those columns with the lowest statistical signifance in explaining our response variable.

And let's run a new linear model.

```
mylm <- lm(SalePrice ~ ., data = non_factor_train2)  #build regression
mysummary <- summary(mylm)  #summarize
mysummary
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = non_factor_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -499711  -16677   -2407   13695  287966
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.464e+05  8.904e+04  -9.506  < 2e-16 ***
## MSSubClass   -1.538e+02  2.605e+01  -5.903 4.44e-09 ***
## LotArea       4.290e-01  1.001e-01   4.285 1.95e-05 ***
## OverallQual   1.884e+04  1.160e+03  16.237  < 2e-16 ***
## OverallCond   5.105e+03  9.332e+02   5.470 5.30e-08 ***
## YearBuilt     3.933e+02  4.521e+01   8.699  < 2e-16 ***
## BsmtFinSF1    1.817e+01  3.898e+00   4.662 3.43e-06 ***
## BsmtUnfSF     8.239e+00  3.645e+00   2.260 0.023948 *
## X1stFlrSF     5.608e+01  4.991e+00  11.237  < 2e-16 ***
## X2ndFlrSF     5.140e+01  4.051e+00  12.689  < 2e-16 ***
## BsmtFullBath  8.627e+03  2.412e+03   3.576 0.000360 ***
## BedroomAbvGr -1.086e+04  1.655e+03  -6.566 7.19e-11 ***
## KitchenAbvGr -1.490e+04  5.066e+03  -2.941 0.003321 **
## TotRmsAbvGrd  5.652e+03  1.221e+03   4.631 3.97e-06 ***
## GarageCars    1.129e+04  1.703e+03   6.627 4.82e-11 ***
## WoodDeckSF    2.688e+01  7.955e+00   3.379 0.000748 ***
## ScreenPorch   5.722e+01  1.687e+01   3.392 0.000713 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35150 on 1443 degrees of freedom
## Multiple R-squared:  0.8064, Adjusted R-squared:  0.8043
## F-statistic: 375.7 on 16 and 1443 DF,  p-value: < 2.2e-16
```

Our new R-squared value using 16 explanatory varialbles is 0.8064105. Not an improvement, but also we have not reduce the fitness of our model by using half the initial number of explanatory variables.

Finally, let's see if our response variables have an exponential component. Bellow, is the list of calculated exponents for the variables we selected in our reduced_variable model.

```
##   MSSubClass      LotArea  OverallQual  OverallCond    YearBuilt
##  -0.71607086  -0.21740785   0.91380627   0.51345663  64.34693729
##   BsmtFinSF1    BsmtUnfSF    X1stFlrSF    X2ndFlrSF BsmtFullBath
##   0.21264255   0.54883601  -0.36306744  -0.06365594  -0.22536066
## BedroomAbvGr KitchenAbvGr TotRmsAbvGrd   GarageCars   WoodDeckSF
##   1.00109483   0.33825177   0.09895533   1.04767560  -0.02926905
##  ScreenPorch    SalePrice
##  -1.28848232  -0.27254159
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train_power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306875  -19027   -3893   13579  361150
##
## Coefficients:
##                 Estimate  Std. Error t value Pr(>|t|)
## (Intercept)   -2.736e+05   1.190e+05  -2.300 0.021605 *
## MSSubClass     1.617e+05   8.414e+04   1.922 0.054813 .
## LotArea       -7.004e+05   1.000e+05  -7.002 3.84e-12 ***
## OverallQual    2.669e+04   1.605e+03  16.624  < 2e-16 ***
## OverallCond    2.383e+04   4.846e+03   4.917 9.77e-07 ***
## YearBuilt     6.210e-227   0.000e+00     Inf  < 2e-16 ***
## BsmtFinSF1     4.241e+03   8.599e+02   4.932 9.10e-07 ***
## BsmtUnfSF      5.579e+01   9.347e+01   0.597 0.550654
## X1stFlrSF     -3.182e+06   3.022e+05 -10.530  < 2e-16 ***
## X2ndFlrSF     -4.381e+04   5.603e+03  -7.819 1.02e-14 ***
## BsmtFullBath  -4.151e+03   1.430e+03  -2.902 0.003758 **
## BedroomAbvGr  -9.779e+03   1.744e+03  -5.608 2.45e-08 ***
## KitchenAbvGr  -5.488e+04   1.881e+04  -2.917 0.003588 **
## TotRmsAbvGrd   5.263e+05   8.685e+04   6.059 1.74e-09 ***
## GarageCars     1.052e+04   1.710e+03   6.152 9.87e-10 ***
## WoodDeckSF     3.105e+01   8.329e+00   3.728 0.000201 ***
## ScreenPorch    5.682e+01   1.778e+01   3.196 0.001423 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36820 on 1443 degrees of freedom
## Multiple R-squared:  0.7875, Adjusted R-squared:  0.7852
## F-statistic: 334.3 on 16 and 1443 DF,  p-value: < 2.2e-16
```

Our new R-squared value using exponential factors with our 16 explanatory varialbles is 0.7875158. Our R-squared value went down! Not an improvement.

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.

First, let's shift the minimum of the Lot Area by 1300 to it starts at zero. Below is the summary of the distributions.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1300    7554    9478   10517   11602  215245
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    6254    8178    9217   10302  213945
```

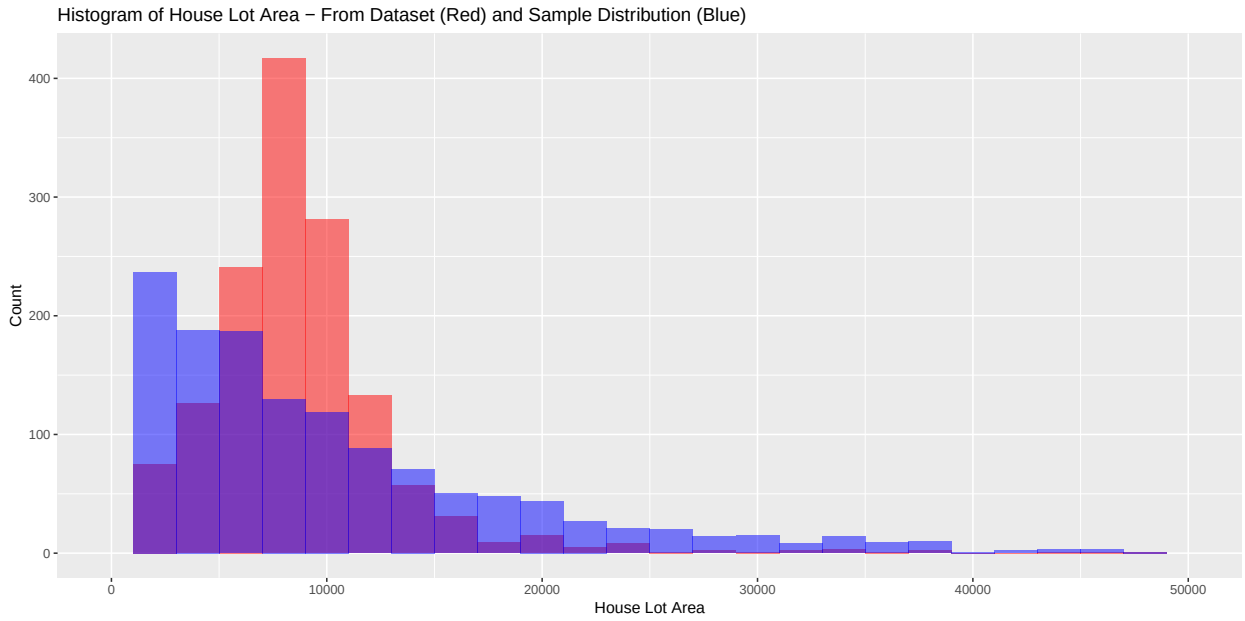
Then we will calculate an exponential fit and the optimal lambda value.

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  house_train$LotArea2
## D = 0.27434, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Where lambda was found to be: 1.084972×10^{-4}

Finally, let's compare both calculated exponential and original distributions from our data.



Our dataset and calculated exponentail distribution do not look anything like. The exponential fit to the our data does not make sense even if our data shows signficant skewness (right skew).

```
hist(house_train$LotArea2, col = "skyblue3", breaks = 40, xlim = c(0,
    50000))
```