Linear Regression

5/18/2019

Pull in the data set

```
df <- read.csv("student-mat.csv", sep = ';')
```

Get to know your data

```
head(df)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
## 4     GP   F  15       U     GT3       T    4    2   health services
## 5     GP   F  16       U     GT3       T    3    3    other    other
## 6     GP   M  16       U     LE3       T    4    3 services    other
##       reason guardian traveltime studytime failures schoolsup famsup paid
## 1     course   mother          2         2        0       yes     no   no
## 2     course   father          1         2        0        no    yes   no
## 3      other   mother          1         2        3       yes     no  yes
## 4       home   mother          1         3        0        no    yes  yes
## 5       home   father          1         2        0        no    yes  yes
## 6 reputation   mother          1         2        0        no    yes  yes
##   activities nursery higher internet romantic famrel freetime goout Dalc
## 1         no     yes    yes       no       no      4        3     4    1
## 2         no      no    yes      yes       no      5        3     3    1
## 3         no     yes    yes      yes       no      4        3     2    2
## 4        yes     yes    yes      yes      yes      3        2     2    1
## 5         no     yes    yes       no       no      4        3     2    1
## 6        yes     yes    yes      yes       no      5        4     2    1
##   Walc health absences G1 G2 G3
## 1    1      3        6  5  6  6
## 2    1      3        4  5  5  6
## 3    3      3       10  7  8 10
## 4    1      5        2 15 14 15
## 5    2      5        4  6 10 10
## 6    2      5       10 15 15 15
```

```
summary(df)
```

```
##  school   sex          age       address famsize   Pstatus      Medu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000
##                   Median :17.0                             Median :3.000
##                   Mean   :16.7                             Mean   :2.749
##                   3rd Qu.:18.0                             3rd Qu.:4.000
##                   Max.   :22.0                             Max.   :4.000
##       Fedu             Mjob           Fjob            reason
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109
##  Median :2.000   other   :141   other   :217   other     : 36  
##  Mean   :2.522   services:103   services:111   reputation:105  
##  3rd Qu.:3.000   teacher : 58   teacher : 29                   
##  Max.   :4.000                                                 
##    guardian     traveltime      studytime       failures      schoolsup
##  father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344  
##  mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51  
##  other : 32   Median :1.000   Median :2.000   Median :0.0000            
##               Mean   :1.448   Mean   :2.035   Mean   :0.3342            
##               3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000            
##               Max.   :4.000   Max.   :4.000   Max.   :3.0000            
##  famsup     paid     activities nursery    higher    internet   romantic 
##  no :153   no :214   no :194    no : 81    no : 20   no : 66   no :263  
##  yes:242   yes:181   yes:201    yes:314    yes:375   yes:329   yes:132  
##                                                                         
##                                                                         
##                                                                         
##                                                                         
##      famrel         freetime          goout            Dalc      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000  
##  Median :4.000   Median :3.000   Median :3.000   Median :1.000  
##  Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481  
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000  
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
##       Walc          health          absences           G1       
##  Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00  
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00  
##  Median :2.000   Median :4.000   Median : 4.000   Median :11.00  
##  Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91  
##  3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00  
##  Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00  
##        G2              G3       
##  Min.   : 0.00   Min.   : 0.00  
##  1st Qu.: 9.00   1st Qu.: 8.00  
##  Median :11.00   Median :11.00  
##  Mean   :10.71   Mean   :10.42  
##  3rd Qu.:13.00   3rd Qu.:14.00  
##  Max.   :19.00   Max.   :20.00
```

Is there any null values

```r
any(is.na(df))
```

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

Structure

Make sure the factors with appropriate levels where imported.

```r
str(df)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Identify what columns have numeric values

```
num.cols <- sapply(df, is.numeric)
```

Coorelation

Build a correlation table

```
cor.data <- cor(df[,num.cols])
```

View the correlation table

```
cor.data
```

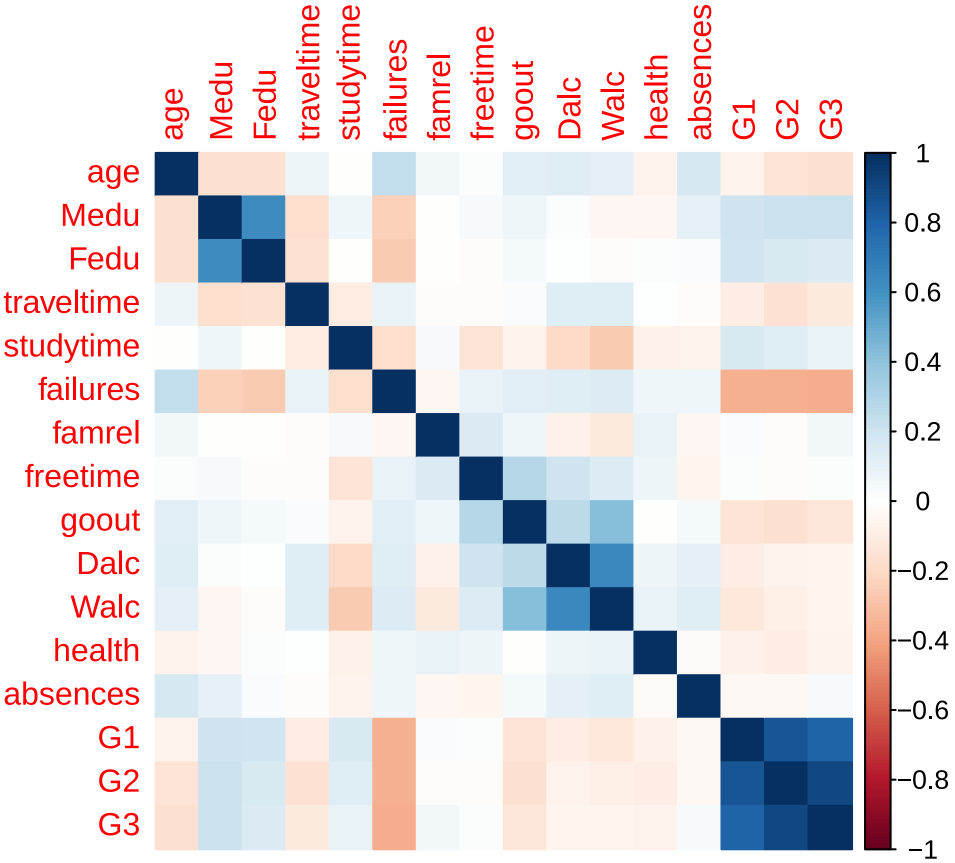
```
##                     age         Medu         Fedu   traveltime
## age         1.000000000 -0.163658419 -0.163438069  0.070640721
## Medu       -0.163658419  1.000000000  0.623455112 -0.171639305
## Fedu       -0.163438069  0.623455112  1.000000000 -0.158194054
## traveltime  0.070640721 -0.171639305 -0.158194054  1.000000000
## studytime  -0.004140037  0.064944137 -0.009174639 -0.100909119
## failures    0.243665377 -0.236679963 -0.250408444  0.092238746
## famrel      0.053940096 -0.003914458 -0.001369727 -0.016807986
## freetime    0.016434389  0.030890867 -0.012845528 -0.017024944
## goout       0.126963880  0.064094438  0.043104668  0.028539674
## Dalc        0.131124605  0.019834099  0.002386429  0.138325309
## Walc        0.117276052 -0.047123460 -0.012631018  0.134115752
## health     -0.062187369 -0.046877829  0.014741537  0.007500606
## absences    0.175230079  0.100284818  0.024472887 -0.012943775
## G1         -0.064081497  0.205340997  0.190269936 -0.093039992
## G2         -0.143474049  0.215527168  0.164893393 -0.153197963
## G3         -0.161579438  0.217147496  0.152456939 -0.117142053
##               studytime    failures       famrel    freetime        goout
## age        -0.004140037  0.24366538  0.053940096  0.01643439  0.126963880
## Medu        0.064944137 -0.23667996 -0.003914458  0.03089087  0.064094438
## Fedu       -0.009174639 -0.25040844 -0.001369727 -0.01284553  0.043104668
## traveltime -0.100909119  0.09223875 -0.016807986 -0.01702494  0.028539674
## studytime   1.000000000 -0.17356303  0.039730704 -0.14319841 -0.063903675
## failures   -0.173563031  1.00000000 -0.044336626  0.09198747  0.124560922
## famrel      0.039730704 -0.04433663  1.000000000  0.15070144  0.064568411
## freetime   -0.143198407  0.09198747  0.150701444  1.00000000  0.285018715
## goout      -0.063903675  0.12456092  0.064568411  0.28501871  1.000000000
## Dalc       -0.196019263  0.13604693 -0.077594357  0.20900085  0.266993848
## Walc       -0.253784731  0.14196203 -0.113397308  0.14782181  0.420385745
## health     -0.075615863  0.06582728  0.094055728  0.07573336 -0.009577254
## absences   -0.062700175  0.06372583 -0.044354095 -0.05807792  0.044302220
## G1          0.160611915 -0.35471761  0.022168316  0.01261293 -0.149103967
## G2          0.135879999 -0.35589563 -0.018281347 -0.01377714 -0.162250034
## G3          0.097819690 -0.36041494  0.051363429  0.01130724 -0.132791474
##                    Dalc        Walc       health    absences          G1
## age         0.131124605  0.11727605 -0.062187369  0.17523008 -0.06408150
## Medu        0.019834099 -0.04712346 -0.046877829  0.10028482  0.20534100
## Fedu        0.002386429 -0.01263102  0.014741537  0.02447289  0.19026994
## traveltime  0.138325309  0.13411575  0.007500606 -0.01294378 -0.09303999
## studytime  -0.196019263 -0.25378473 -0.075615863 -0.06270018  0.16061192
## failures    0.136046931  0.14196203  0.065827282  0.06372583 -0.35471761
## famrel     -0.077594357 -0.11339731  0.094055728 -0.04435409  0.02216832
## freetime    0.209000848  0.14782181  0.075733357 -0.05807792  0.01261293
## goout       0.266993848  0.42038575 -0.009577254  0.04430222 -0.14910397
## Dalc        1.000000000  0.64754423  0.077179582  0.11190803 -0.09415879
## Walc        0.647544230  1.00000000  0.092476317  0.13629110 -0.12617921
## health      0.077179582  0.09247632  1.000000000 -0.02993671 -0.07317207
## absences    0.111908026  0.13629110 -0.029936711  1.00000000 -0.03100290
## G1         -0.094158792 -0.12617921 -0.073172073 -0.03100290  1.00000000
## G2         -0.064120183 -0.08492735 -0.097719866 -0.03177670  0.85211807
## G3         -0.054660041 -0.05193932 -0.061334605  0.03424732  0.80146793
##                     G2          G3
## age        -0.14347405 -0.16157944
## Medu        0.21552717  0.21714750
## Fedu        0.16489339  0.15245694
## traveltime -0.15319796 -0.11714205
## studytime   0.13588000  0.09781969
## failures   -0.35589563 -0.36041494
## famrel     -0.01828135  0.05136343
## freetime   -0.01377714  0.01130724
## goout      -0.16225003 -0.13279147
## Dalc       -0.06412018 -0.05466004
## Walc       -0.08492735 -0.05193932
## health     -0.09771987 -0.06133460
## absences   -0.03177670  0.03424732
## G1          0.85211807  0.80146793
## G2          1.00000000  0.90486799
## G3          0.90486799  1.00000000
```

As you can see it is pretty difficult to see what is happening, so lets visualize it will coorplot.

```
library(corrplot)
```

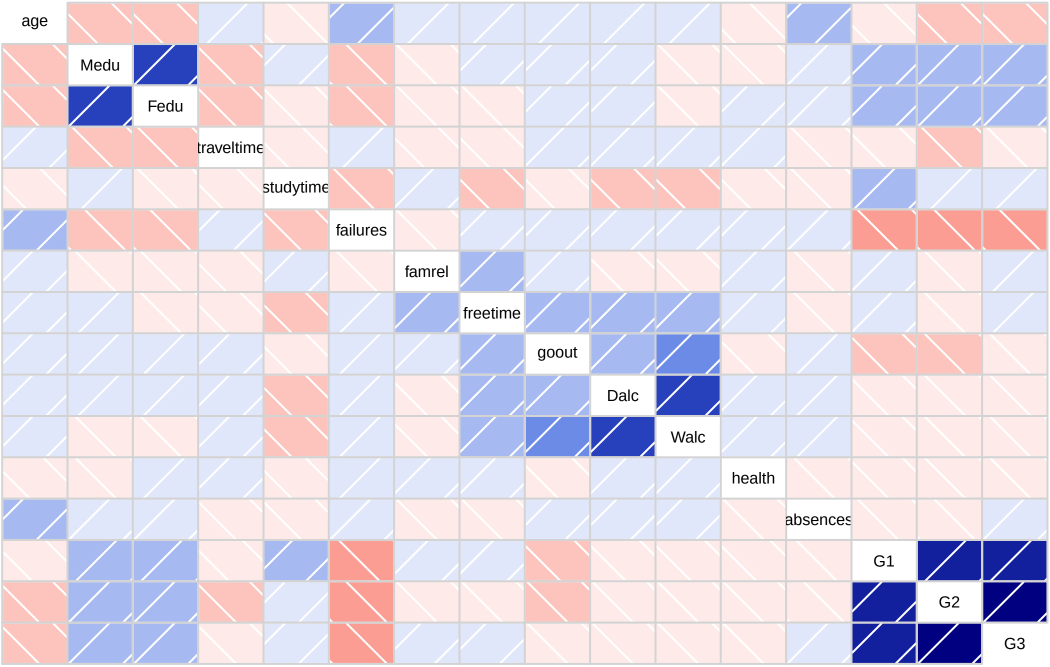
```
## corrplot 0.84 loaded
```

```
corrplot(cor.data, method = "color")
```

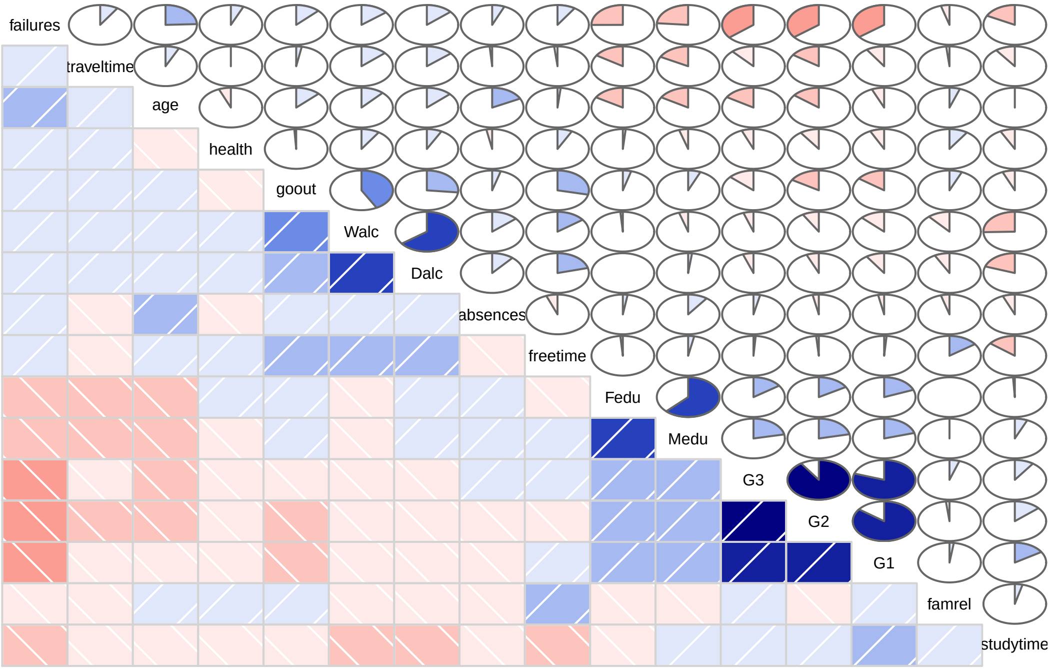


You can also use corrgram for a different perspective.

```
library(corrgram)
corrgram(df)
```



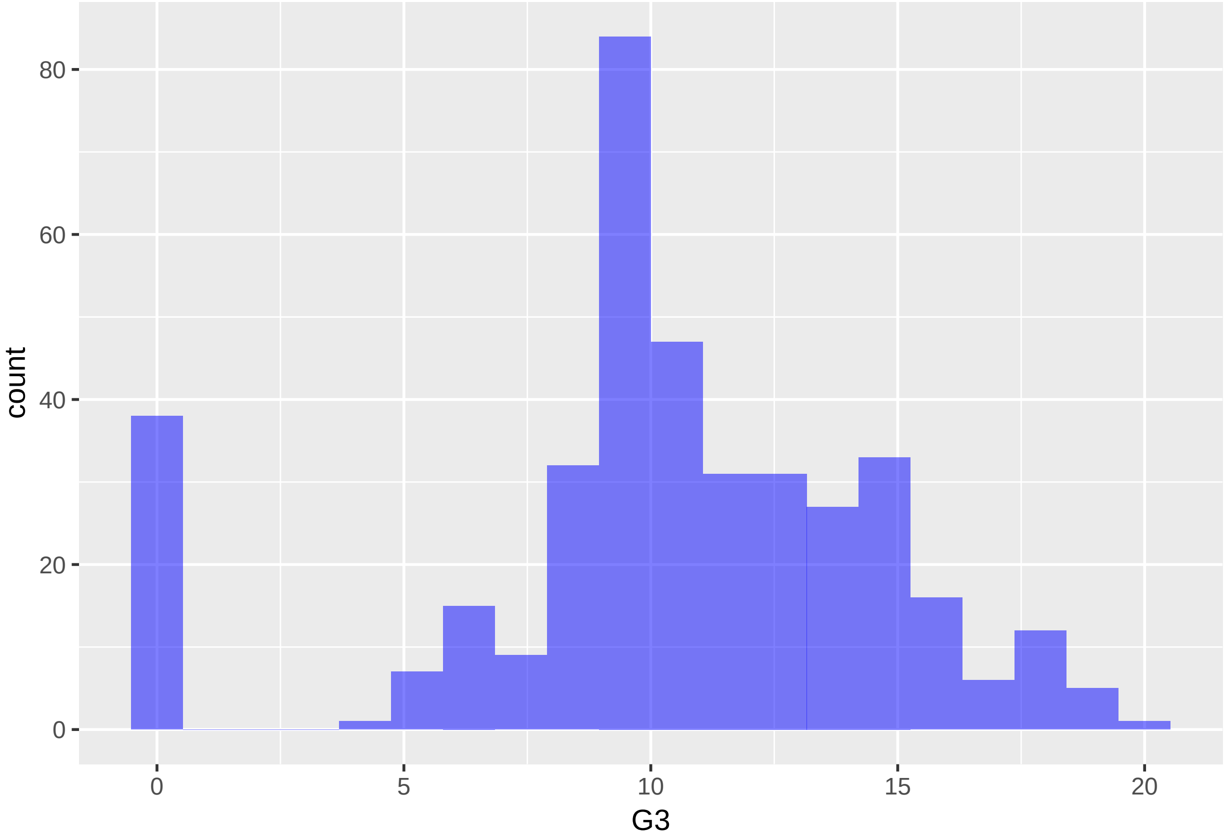
```
corrgram(df, order = TRUE, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt)
```



```
library(ggplot2)
```

Lets look at a histogram of the 3rd quarter test scores (G3) after all this is what we are trying to perdict

```
ggplot(df, aes(G3)) +
    geom_histogram(bins = 20, alpha = .5, fill = "blue")
```


80
60
40
20
0
count
0
5
10
15
20
G3