Previsão das notas finais de alunos

Silva, Guilherme Aquino

26/10/2021

Criando um modelo para previsão das notas finais de alunos através dos dados disponíveis no dataset **Student Performance Dataset**. Link:
Student Performance Dataset (https://archive.ics.uci.edu/ml/datasets/Student+Performance)

Carregando o dataset

```
df <- read.csv2('estudantes.csv')
```

Explorando os dados

```
View(df)
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      guardian
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273
##  Median :2.000   other   :141   other   :217   other     : 36   other : 32
##  Mean   :2.522   services:103   services:111   reputation:105
##  3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000
##    traveltime      studytime       failures      schoolsup famsup     paid
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181
##  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
##  activities nursery   higher    internet   romantic      famrel
##  no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000
##  yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000
##                                                     Median :4.000
##                                                     Mean   :3.944
##                                                     3rd Qu.:5.000
##                                                     Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health         absences            G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

```
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 ...
```

```r
any(is.na(df)) # verificação de valores NA no dataset
```

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

```r
library(dplyr)
```

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

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

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

Verificando a correlação entre as colunas numéricas

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
col_numericas <- sapply(df, is.numeric) # extraindo as colunas numéricas
length(col_numericas)
```

```
## [1] 33
```

```
?cor
```

```
## starting httpd help server ...
```

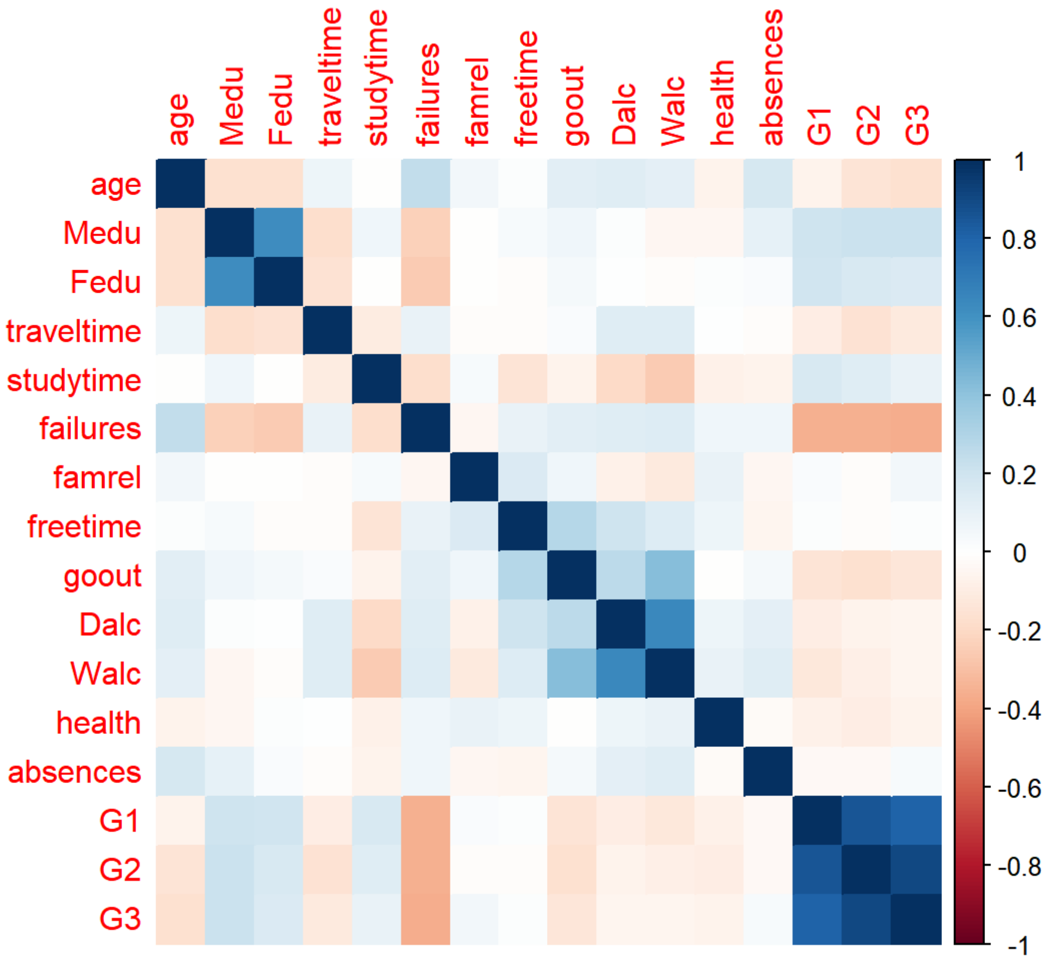
```
##  done
```

```
cor(df[,col_numericas]) # correlação
```

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

```
## failures   -0.36041494
## famrel      0.05136343
## freetime    0.01130724
## goout      -0.13279147
## Dalc       -0.05466004
## Walc       -0.05193932
## health     -0.06133460
## absences    0.03424732
## G1          0.80146793
## G2          0.90486799
## G3          1.00000000
```

```
corrplot(cor(df[, col_numericas]), method = 'color') # plotando a correlação
```



Após a verificação, foi observado que não há nenhuma forte correlação entre as variáveis numéricas.
Chama atenção uma leve correlação positiva entre as variáveis:
- Dalc x Walc
- goout x Walc
- Medu x Fedu
Chama atenção uma leve correlação negativa entre as variáveis:
- failures x G1, G2 e G3
- failures x Medu e Fedu
- studytime x Walc

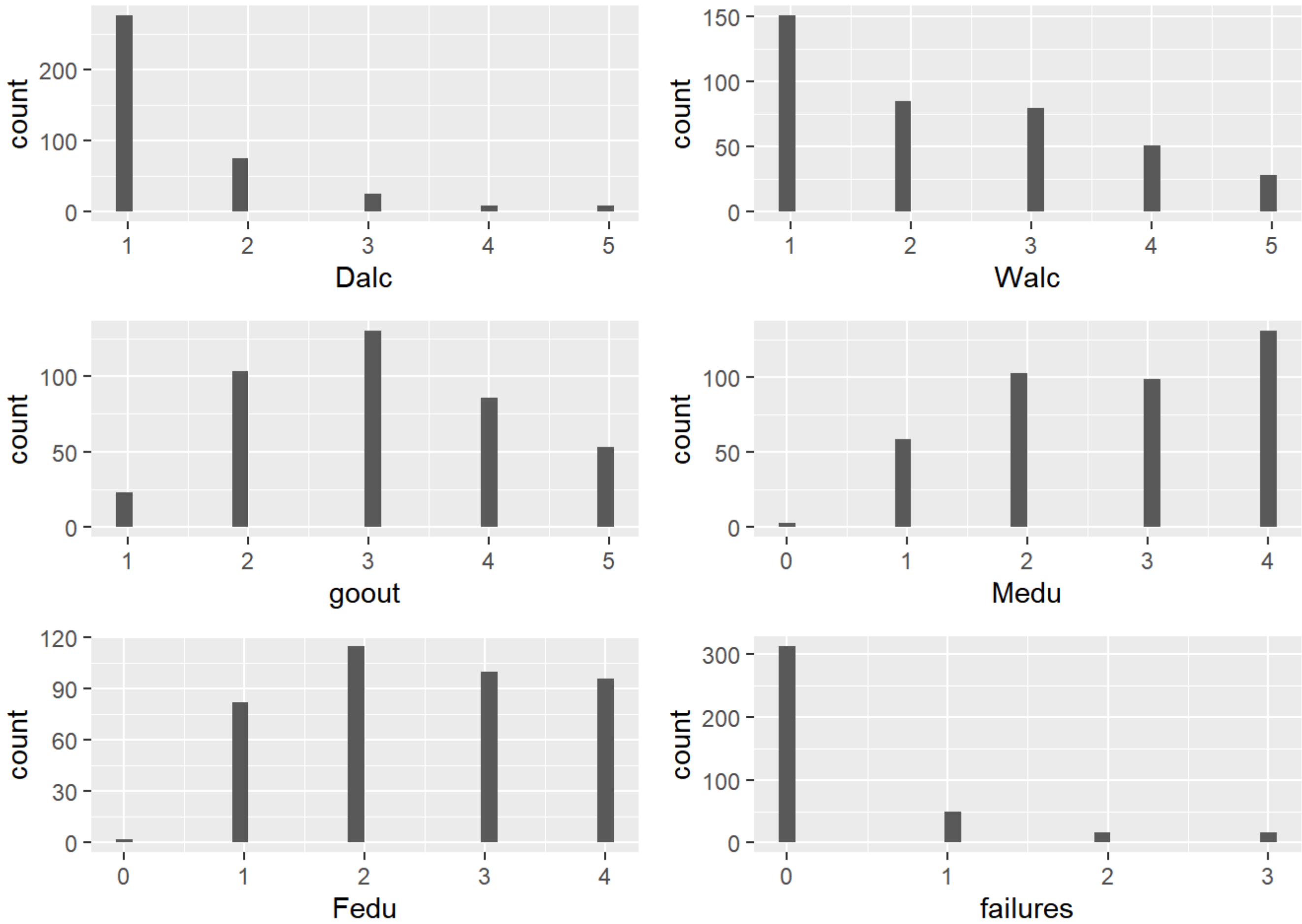
Analisando as variáveis:

```r
library(ggplot2)
library(ggthemes)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

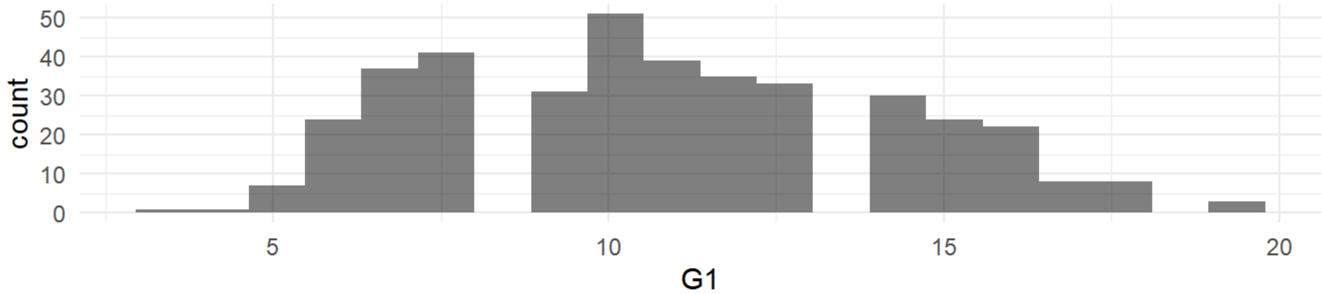
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

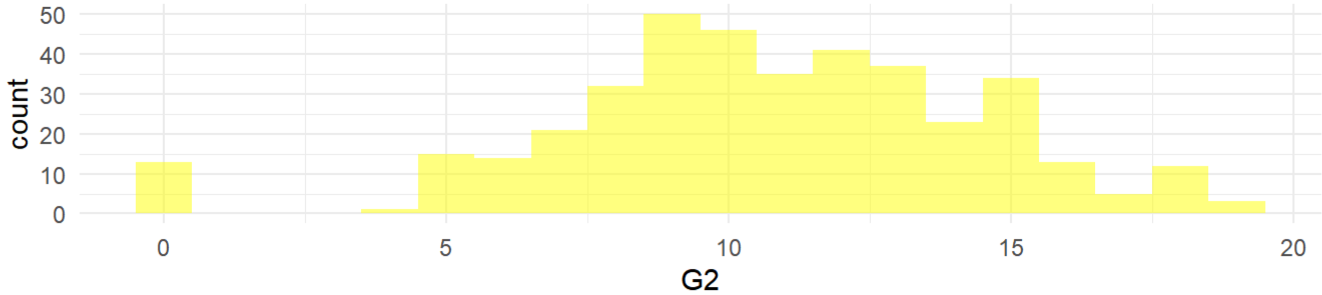
```r
hist1 <- ggplot(df, aes(Dalc)) +
  geom_histogram(bins = 30) # Consumação de Álcool durante de trabalho

hist2 <- ggplot(df, aes(Walc)) +
  geom_histogram(bins = 30) # Consumação de Álcool no final de semana

hist3 <- ggplot(df, aes(x = goout)) +
  geom_histogram(bins = 30) # Frequências de saídas com os amigos

hist4 <- ggplot(df, aes(x = Medu)) +
  geom_histogram(bins = 30) # Escolaridade da mãe

hist5 <- ggplot(df, aes(x = Fedu)) +
  geom_histogram(bins = 30) # Escolaridade do pai

hist6 <- ggplot(df, aes(x = failures)) +
  geom_histogram(bins = 30) # Frequência de reprovações

grid.arrange(hist1, hist2, hist3, hist4, hist5, hist6)
```

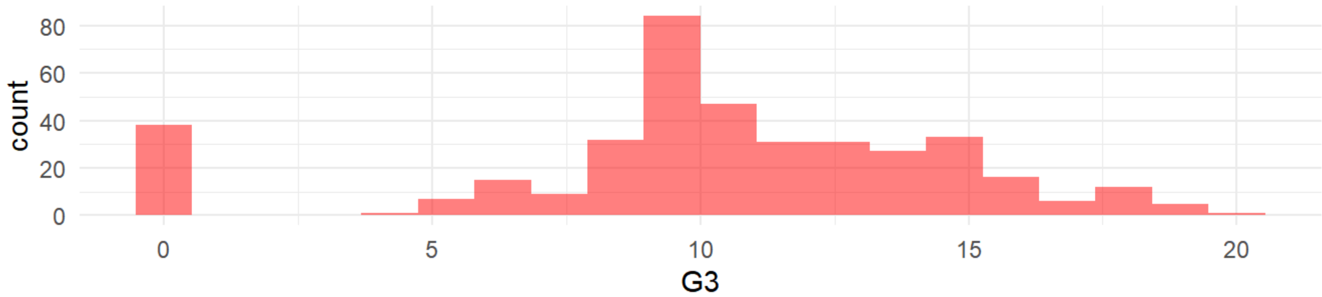


Analisando as variáveis G1, G2 e G3

```
plot1 <- ggplot(df, aes(G1)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'black') +
  theme_minimal()

plot2 <- ggplot(df, aes(G2)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'yellow') +
  theme_minimal()

plot3 <- ggplot(df, aes(G3)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'red') +
  theme_minimal()

grid.arrange(plot1, plot2, plot3, ncol = 1)
```







Obs.: Chama atenção o número de reprovações na 2ª avaliação (G2) e na avaliação final (G3)

Criando as amostras de forma randômica

```
library(caTools)
amostra <- sample.split(df$age, SplitRatio = 0.70)
```

Criando dados de treino

```
treino <- subset(df, amostra == T)
```

Criando dados de teste

```
teste <- subset(df, amostra == F)
```

Criando os modelos

```
modelo_1 <- lm(G3 ~ ., treino)
modelo_2 <- lm(G3 ~ G1 + G2, treino)
modelo_3 <- lm(G3 ~ absences, treino)
modelo_4 <- lm(G3 ~ Medu, treino)
modelo_5 <- lm(G3 ~ Fedu, treino)
modelo_6 <- lm(G3 ~ failures, treino)
modelo_7 <- lm(G3 ~ goout, treino)
modelo_8 <- lm(G3 ~ Walc, treino)
```

Analisando os modelos

```
summary(modelo_1)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7304 -0.6323  0.1848  0.8884  3.1280
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.227755   2.329062  -0.527  0.59859
## schoolMS          0.162977   0.385350   0.423  0.67273
## sexM              0.357555   0.260673   1.372  0.17148
## age              -0.136104   0.111172  -1.224  0.22208
## addressU         -0.010201   0.294007  -0.035  0.97235
## famsizeLE3        0.058476   0.241453   0.242  0.80885
## PstatusT          0.349703   0.398543   0.877  0.38114
## Medu              0.125801   0.158186   0.795  0.42726
## Fedu             -0.085574   0.133002  -0.643  0.52059
## Mjobhealth       -0.177644   0.570733  -0.311  0.75588
## Mjobother        -0.017948   0.351152  -0.051  0.95928
## Mjobservices     -0.037992   0.399021  -0.095  0.92423
## Mjobteacher       0.293866   0.533540   0.551  0.58230
## Fjobhealth        0.260892   0.665625   0.392  0.69545
## Fjobother         0.131189   0.513771   0.255  0.79868
## Fjobservices      0.001531   0.537647   0.003  0.99773
## Fjobteacher      -0.270785   0.649038  -0.417  0.67691
## reasonhome       -0.143984   0.273379  -0.527  0.59891
## reasonother       0.484144   0.382262   1.267  0.20658
## reasonreputation  0.217292   0.293773   0.740  0.46025
## guardianmother   -0.060228   0.271057  -0.222  0.82435
## guardianother     0.071963   0.498570   0.144  0.88536
## traveltime        0.092698   0.170605   0.543  0.58740
## studytime         0.114626   0.151410   0.757  0.44977
## failures         -0.298484   0.181610  -1.644  0.10161
## schoolsupyes      0.584640   0.353612   1.653  0.09960 .
## famsupyes         0.190772   0.242245   0.788  0.43177
## paidyes          -0.208922   0.243993  -0.856  0.39272
## activitiesyes    -0.419891   0.229583  -1.829  0.06868 .
## nurseryyes       -0.166053   0.284588  -0.583  0.56013
## higheryes        -0.114049   0.522480  -0.218  0.82740
## internetyes      -0.435771   0.305220  -1.428  0.15470
## romanticyes      -0.647033   0.241532  -2.679  0.00791 **
## famrel            0.268892   0.126309   2.129  0.03431 *
## freetime         -0.090822   0.120447  -0.754  0.45158
## goout             0.179882   0.113571   1.584  0.11457
## Dalc             -0.014296   0.165961  -0.086  0.93143
## Walc              0.048193   0.127383   0.378  0.70553
## health            0.004724   0.081064   0.058  0.95358
## absences          0.057760   0.016734   3.452  0.00066 ***
## G1                0.151797   0.064808   2.342  0.02000 *
## G2                0.979811   0.053915  18.173  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 235 degrees of freedom
```

```
## Multiple R-squared:  0.8762, Adjusted R-squared:  0.8546
## F-statistic: 40.58 on 41 and 235 DF,  p-value: < 2.2e-16
```

```
summary(modelo_2)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5918 -0.4245  0.2001  0.8271  3.5383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.50989    0.36109  -4.182  3.9e-05 ***
## G1           0.11275    0.05643   1.998   0.0467 *
## G2           1.00870    0.04918  20.511  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.718 on 274 degrees of freedom
## Multiple R-squared:  0.8492, Adjusted R-squared:  0.8481
## F-statistic: 771.7 on 2 and 274 DF,  p-value: < 2.2e-16
```

```
summary(modelo_3)
```

```
##
## Call:
## lm(formula = G3 ~ absences, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3028  -2.3028   0.5854   2.6599   9.6227
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30280    0.34167  30.154   <2e-16 ***
## absences     0.01863    0.03816   0.488    0.626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.415 on 275 degrees of freedom
## Multiple R-squared:  0.0008662,  Adjusted R-squared:  -0.002767
## F-statistic: 0.2384 on 1 and 275 DF,  p-value: 0.6257
```

```
summary(modelo_4)
```

```
##
## Call:
## lm(formula = G3 ~ Medu, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5199  -1.7992   0.4801   2.4801   9.2008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0785     0.6922  11.671  < 2e-16 ***
## Medu          0.8603     0.2370   3.629 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.315 on 275 degrees of freedom
## Multiple R-squared:  0.04571,    Adjusted R-squared:  0.04224
## F-statistic: 13.17 on 1 and 275 DF,  p-value: 0.0003386
```

```
summary(modelo_5)
```

```
##
## Call:
## lm(formula = G3 ~ Fedu, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3808  -1.7330   0.5626   2.6192   9.2670
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7896     0.6584  13.350  < 2e-16 ***
## Fedu          0.6478     0.2418   2.679  0.00782 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.36 on 275 degrees of freedom
## Multiple R-squared:  0.02544,    Adjusted R-squared:  0.0219
## F-statistic: 7.179 on 1 and 275 DF,  p-value: 0.007819
```

```
summary(modelo_6)
```

```
##
## Call:
## lm(formula = G3 ~ failures, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0499  -2.0059  -0.0499   2.9501   8.9501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0499     0.2739  40.336  < 2e-16 ***
## failures     -2.0440     0.3528  -5.793 1.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.17 on 275 degrees of freedom
## Multiple R-squared:  0.1088, Adjusted R-squared:  0.1055
## F-statistic: 33.56 on 1 and 275 DF,  p-value: 1.885e-08
```

```
summary(modelo_7)
```

```
##
## Call:
## lm(formula = G3 ~ goout, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6540  -1.8625   0.5403   2.7346   8.9432
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2511     0.7788  15.731   <2e-16 ***
## goout        -0.5972     0.2376  -2.514   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.367 on 275 degrees of freedom
## Multiple R-squared:  0.02246,    Adjusted R-squared:  0.01891
## F-statistic: 6.319 on 1 and 275 DF,  p-value: 0.01252
```

```
summary(modelo_8)
```

```
##
## Call:
## lm(formula = G3 ~ Walc, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7425  -1.7425   0.2575   3.0213   9.2575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9971     0.5487  20.041   <2e-16 ***
## Walc         -0.2546     0.2077  -1.226    0.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.405 on 275 degrees of freedom
## Multiple R-squared:  0.005434,   Adjusted R-squared:  0.001817
## F-statistic: 1.502 on 1 and 275 DF,  p-value: 0.2214
```

Visualizando as taxas de erro (resíduos) do modelo escolhido

```
res <- residuals(modelo_1)
res <- as.data.frame(res)
res
```

```
##              res
## 2    1.518918e+00
## 3    2.417973e+00
## 5    7.437795e-01
## 6   -1.013247e+00
## 7   -1.004836e+00
## 8    1.074518e+00
## 11   9.473149e-01
## 14   2.316540e-01
## 15   7.044318e-01
## 16  -1.234323e-01
## 17   2.082657e-01
## 18  -6.359023e-01
## 19   6.564640e-01
## 22  -5.465121e-01
## 23   7.219483e-01
## 24  -1.769920e+00
## 25  -6.918323e-01
## 27  -9.447292e-01
## 28  -1.719657e+00
## 29  -2.339764e-01
## 30  -7.970676e-01
## 31   1.379170e+00
## 32   5.230812e-01
## 33   5.576316e-02
## 34   2.500270e+00
## 36  -1.162635e-01
## 37   1.041484e+00
## 38  -4.852092e-02
## 39  -1.060658e+00
## 41   1.852961e+00
## 42  -9.513879e-01
## 43  -6.933400e-01
## 44   3.008871e+00
## 45  -1.178982e+00
## 46  -1.279133e+00
## 47  -8.177435e-01
## 48   3.688401e-02
## 51   9.810602e-02
## 52  -3.491904e-01
## 53  -2.514902e+00
## 54   1.401831e+00
## 56   2.052588e+00
## 57   7.205997e-01
## 58  -1.208862e+00
## 59  -7.407146e-01
## 64  -1.961844e-01
## 65   3.987251e-01
## 68  -1.261449e+00
## 69  -1.287292e+00
## 72  -8.613470e-02
## 74   1.420921e+00
## 76   9.235185e-01
## 77  -1.735322e+00
## 80   1.001439e+00
## 81   9.162620e-01
## 82   1.694974e-01
```

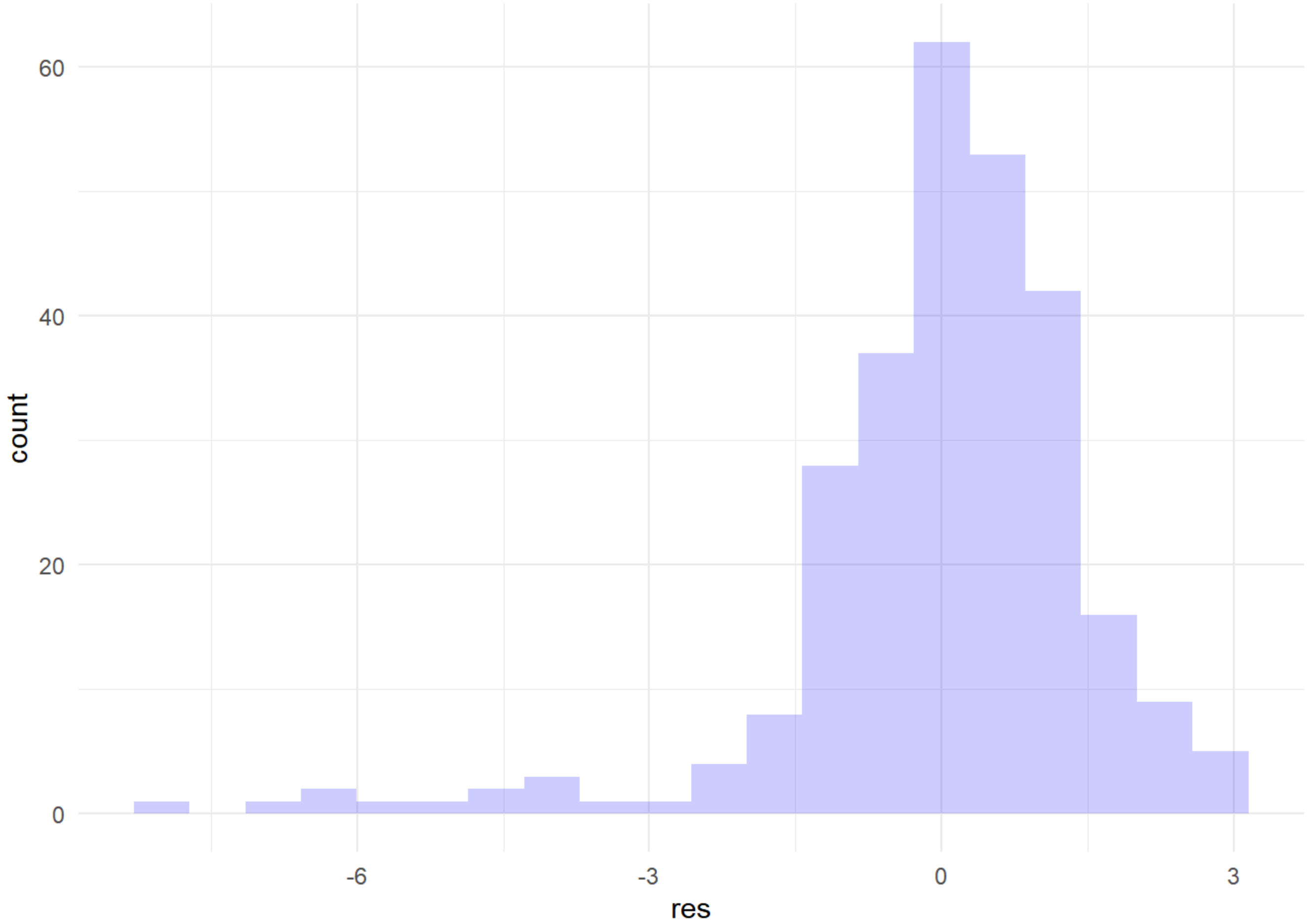
```
## 83  -1.814718e-01
## 84  -3.619274e-01
## 87  -1.294847e+00
## 88  -8.325210e-01
## 89   9.878433e-02
## 92   5.979925e-01
## 93  -4.426742e-01
## 95   1.178716e-01
## 96   1.085826e+00
## 98   1.561994e+00
## 99  -2.149144e-01
## 100 -1.289925e+00
## 102  1.372937e-01
## 103  7.942965e-01
## 106 -1.426166e+00
## 107 -6.300858e-01
## 108 -7.031285e-01
## 109  7.028426e-01
## 110  7.170150e-01
## 111 -5.099237e-01
## 112  5.982709e-01
## 113  8.509454e-01
## 114 -1.361291e+00
## 117  7.592758e-01
## 118 -5.053316e-01
## 119  1.235871e-01
## 120 -5.781030e-01
## 123  1.055679e+00
## 124  1.084042e+00
## 126 -1.936267e+00
## 127  2.822051e+00
## 129 -2.182368e+00
## 131  1.266449e+00
## 133 -3.507849e-01
## 135 -4.104215e-02
## 136  9.525485e-01
## 137 -1.085393e+00
## 139  1.176872e-02
## 140 -6.843009e-01
## 142  1.101789e+00
## 143  6.802652e-01
## 148  5.184333e-01
## 149 -4.443269e+00
## 151 -2.439990e+00
## 152  1.291921e+00
## 153  1.432660e+00
## 154  2.746074e+00
## 156 -4.659960e-01
## 157  1.338419e-01
## 158  2.203630e+00
## 159 -1.475094e+00
## 162 -1.475693e+00
## 163  6.584924e-01
## 164  1.372009e-01
## 165  5.214308e-01
## 166  2.781721e-01
## 167 -6.381591e-01
## 170  7.611915e-01
```

```
## 171 -3.783056e+00
## 172  1.701587e+00
## 174 -4.233863e+00
## 176 -4.887220e-02
## 177 -1.692238e+00
## 178  1.280140e+00
## 179  5.906397e-01
## 180  1.522130e+00
## 181  1.775545e-01
## 182  1.640177e-01
## 183 -2.669124e-01
## 184 -3.128753e+00
## 185 -9.001330e-01
## 187  2.467510e-01
## 188  5.872280e-01
## 191  1.170465e+00
## 192  2.321740e+00
## 193 -1.372321e-01
## 194  1.363428e+00
## 195 -1.903358e-01
## 197  3.881849e-01
## 198  8.865961e-01
## 200  1.816515e+00
## 201 -4.667177e-01
## 202 -1.478588e-01
## 203  9.367106e-01
## 204 -3.503042e-01
## 205  5.070661e-01
## 208  2.240334e+00
## 209  1.200956e+00
## 210  1.242245e+00
## 211  4.842961e-01
## 212  2.956785e-01
## 215 -5.983075e-01
## 217 -1.039989e+00
## 218  2.364894e+00
## 219  1.857133e+00
## 220  1.108278e+00
## 221  5.933360e-01
## 223  1.050679e+00
## 224  4.891300e-01
## 226 -2.681285e-01
## 227 -1.048968e+00
## 229  7.754574e-01
## 230  3.128000e+00
## 231  1.406599e+00
## 232  6.812743e-01
## 233 -7.684488e-01
## 234 -5.519818e-01
## 235 -1.105191e+00
## 237  8.858397e-01
## 238  2.214152e-01
## 241  4.412324e-01
## 242  1.298993e+00
## 243 -3.846300e-02
## 244  3.040347e-01
## 245  1.130169e+00
## 246 -1.298556e+00
```

```
## 247  7.759459e-01
## 248  8.884162e-01
## 249  1.703351e+00
## 250 -1.274092e-01
## 251  7.928787e-01
## 253 -7.665312e-01
## 254 -3.580567e-01
## 255  7.257769e-01
## 256  2.619917e-01
## 257  2.459399e-01
## 259 -7.203798e-01
## 260 -6.462806e+00
## 263 -4.602028e-01
## 265 -7.730367e+00
## 266 -1.240694e+00
## 269  1.008327e+00
## 270  8.254762e-01
## 273  1.847507e-01
## 275  1.028208e+00
## 280 -7.932365e-01
## 281  2.063255e-01
## 284  1.508144e+00
## 285  1.552058e+00
## 286  7.315794e-01
## 287  3.918759e-02
## 288  1.424945e-01
## 291  3.068705e-01
## 292 -5.201979e-02
## 293  1.483245e+00
## 295  3.615182e-01
## 296 -1.018597e+00
## 298  3.210755e-01
## 299  6.904260e-01
## 300  2.363462e+00
## 302 -2.114568e-02
## 303  1.378919e+00
## 304  2.662747e-01
## 306  6.630486e-01
## 308 -2.052346e+00
## 309  8.069881e-01
## 310 -9.298143e-01
## 312  4.669462e-01
## 313  6.403639e-01
## 314  5.431980e-01
## 315  3.866673e-01
## 316 -1.162794e+00
## 317 -6.872915e+00
## 320  1.041464e-01
## 321 -2.066308e-01
## 322  3.124609e-05
## 324  1.149817e+00
## 326 -2.560513e-01
## 327  4.257265e-01
## 328 -9.694952e-01
## 329 -6.717778e-02
## 332  6.342104e-01
## 333  1.365042e+00
## 334 -5.986180e+00
```

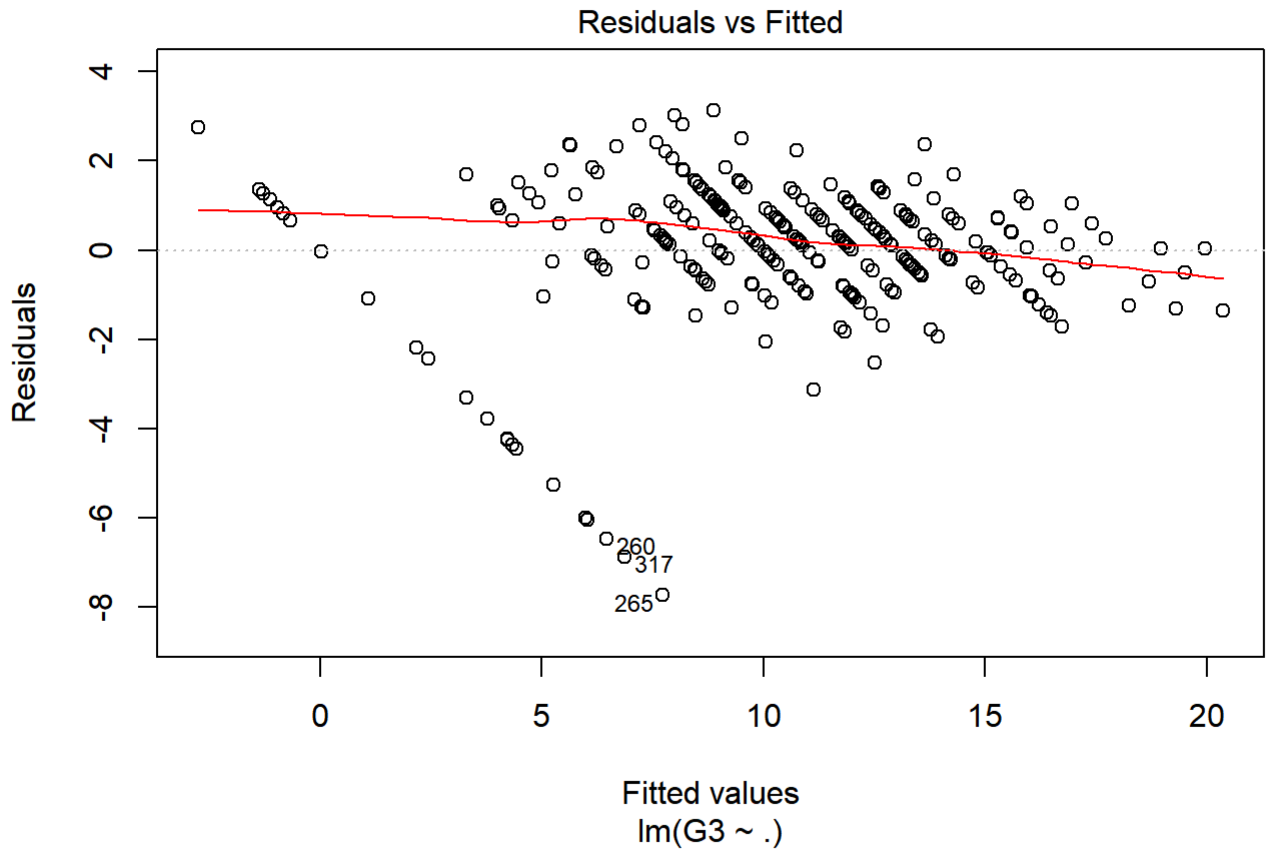
```
## 336 -1.402419e+00
## 337 -1.994945e-01
## 338 -6.043517e+00
## 339  1.193055e+00
## 340  9.748142e-01
## 341  9.446948e-01
## 343  2.009327e-01
## 344 -5.255809e+00
## 345  2.228913e-01
## 346  8.939587e-01
## 348 -1.022518e+00
## 350 -2.554568e-01
## 351  1.748718e+00
## 352 -3.106912e-01
## 353  2.338478e+00
## 354  4.518969e-01
## 355  1.131563e-01
## 357 -4.109333e-01
## 358 -4.116604e-02
## 359  1.238775e+00
## 360 -6.323450e-01
## 361  5.858780e-01
## 362 -7.808241e-01
## 363 -3.304995e-01
## 365  1.109476e+00
## 366 -2.282450e-01
## 367 -1.345421e-01
## 368 -4.236161e+00
## 370 -9.918126e-01
## 372  7.370087e-02
## 373  8.467444e-01
## 374  9.441594e-01
## 376  2.784665e+00
## 377  1.578264e+00
## 378  1.789274e+00
## 379  8.100048e-01
## 380 -6.450189e-01
## 382  1.784861e+00
## 384 -4.352266e+00
## 385 -2.472944e-01
## 386  2.765387e-01
## 388 -3.303454e+00
## 389 -4.403717e-01
## 391  2.059599e-01
## 393 -2.831408e-01
## 394 -1.826362e+00
## 395  7.700352e-01
```

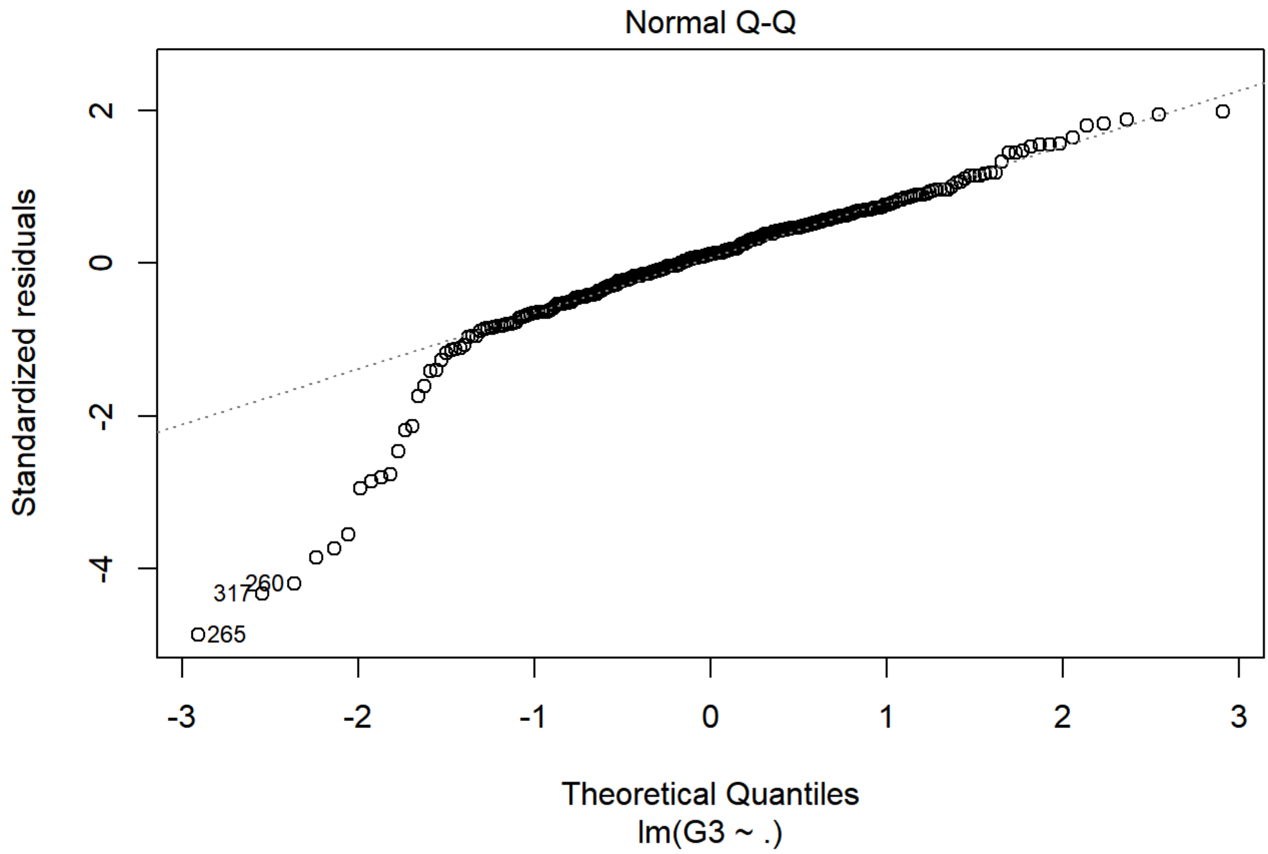
```
ggplot(res, aes(res)) +
  geom_histogram(bins = 20,
                 alpha = 0.20,
                 fill = 'blue') +
  theme_minimal()
```

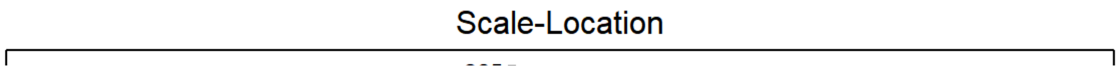


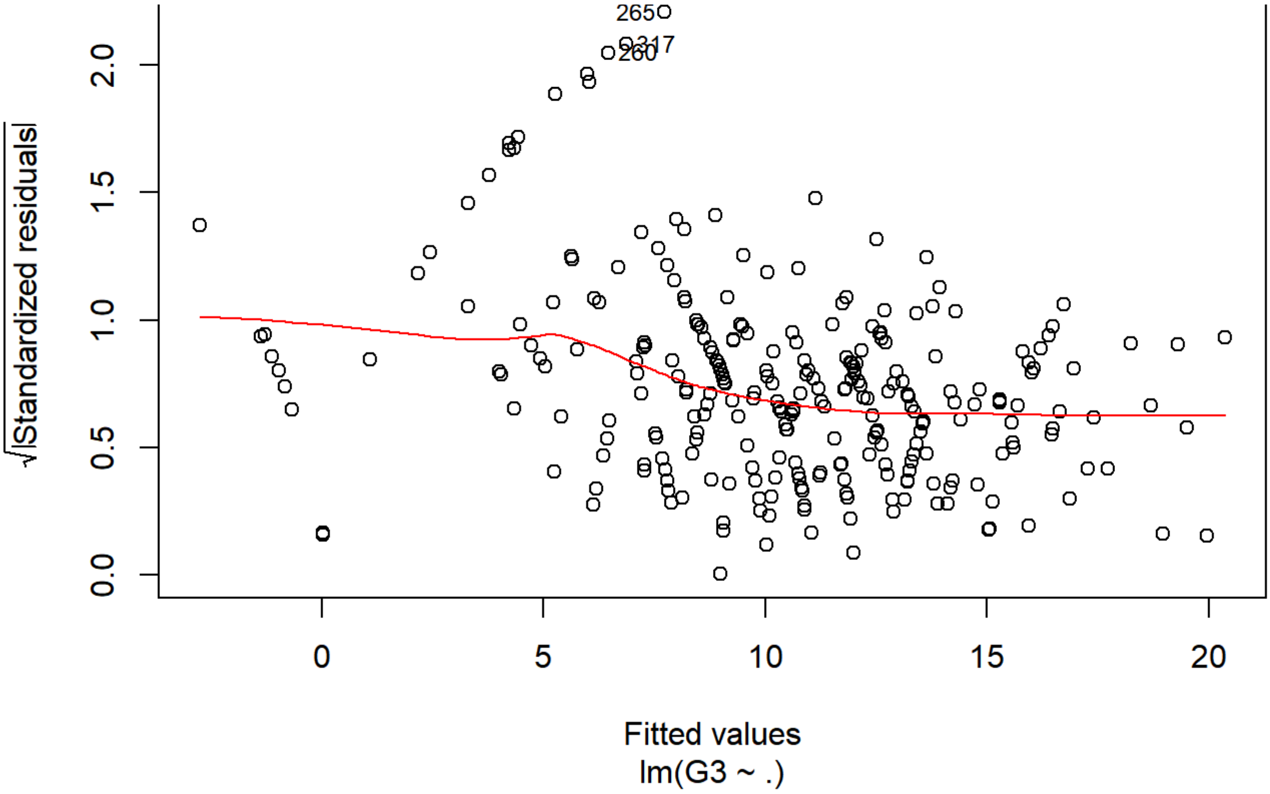
Plot do modelo

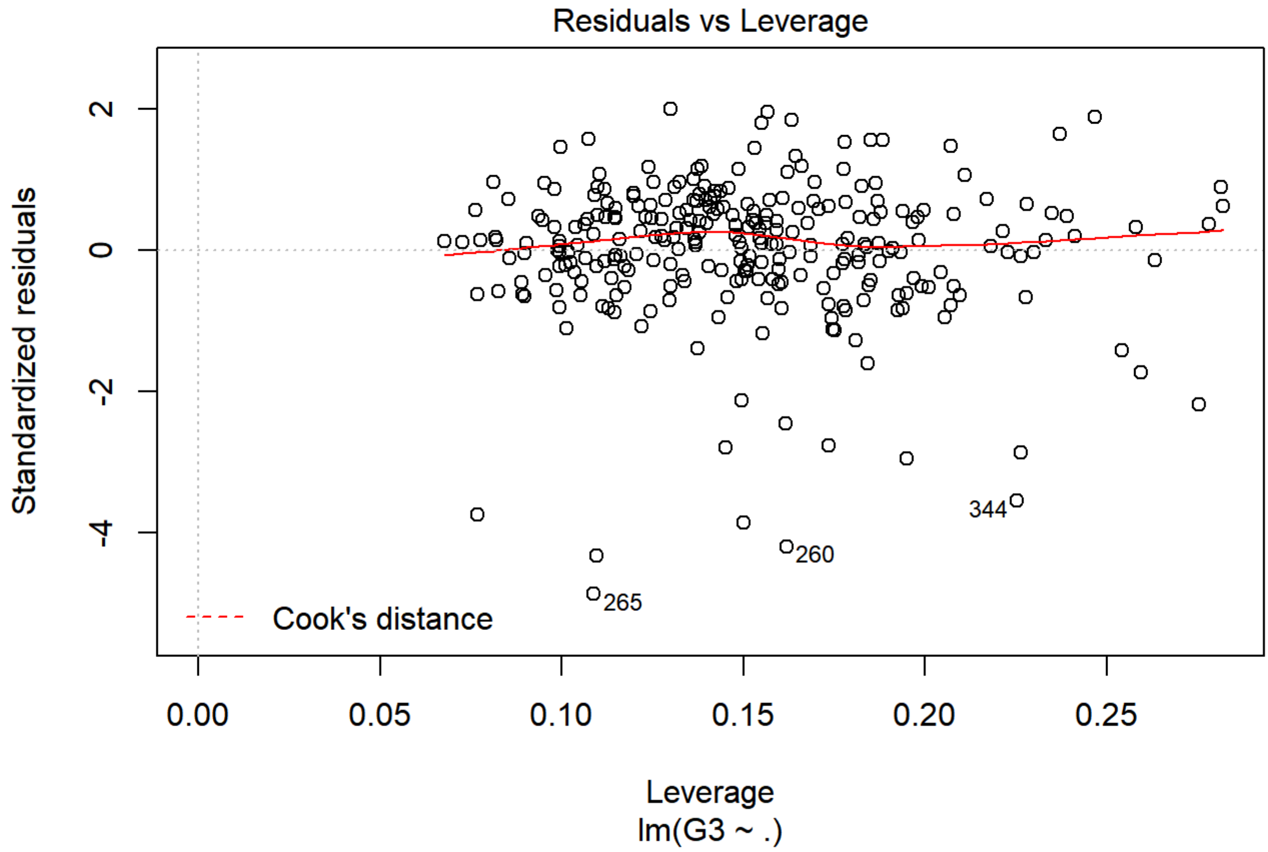
```
plot(modelo_1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(G3 ~ .)
260
317
265


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(G3 ~ .)
317
260
265


Scale-Location





Prevendo as notas finais

```
previsao_G3 <- predict(modelo_1, teste)
as.data.frame(previsao_G3)
```

```
##     previsao_G3
## 1      5.416357
## 4     12.809676
## 9     18.177325
## 10    14.489973
## 12    12.305789
## 13    13.979093
## 20     8.921714
## 21    14.483904
## 26     6.954883
## 35    14.163517
## 40    13.874015
## 49    16.528127
## 50     6.752275
## 55    13.253224
## 60    16.656481
## 61    10.564756
## 62     8.267813
## 63     9.452978
## 66    15.885617
## 67    12.166485
## 70    18.047118
## 71    15.714773
## 73     5.344757
## 75    14.663901
## 78    10.558811
## 79     6.406645
## 85     9.200226
## 86     7.456627
## 90     7.984766
## 91     5.432489
## 94     9.736184
## 97    15.513393
## 101    7.917192
## 104    7.619754
## 105   18.118807
## 115    8.169073
## 116   15.748621
## 121   15.231188
## 122   14.933283
## 125    5.504952
## 128    6.137871
## 130   18.602492
## 132   -1.193772
## 134   10.734784
## 138   -3.005084
## 141    9.317902
## 144   13.691359
## 145   -1.237363
## 146    9.851967
## 147    4.569238
## 150    7.999135
## 155   10.239802
## 160   11.225122
## 161    3.422333
## 168   14.000112
## 169    6.652839
```

```
## 173   11.462053
## 175   10.835683
## 186   12.323515
## 189    5.633352
## 190    8.849261
## 196   13.311146
## 199   19.416395
## 206    8.766419
## 207    4.703358
## 213   12.973333
## 214    6.113886
## 216   15.327512
## 222    3.498275
## 225   13.043979
## 228   11.614790
## 236    9.349652
## 239   10.264566
## 240    6.461054
## 252   11.524314
## 258   10.916905
## 261   18.443909
## 262    7.764029
## 264    8.515096
## 267    8.468895
## 268   10.176129
## 271    8.007773
## 272   13.493822
## 274   12.933667
## 276   12.085373
## 277   11.530267
## 278    9.219253
## 279    7.157277
## 282    9.536078
## 283   12.192482
## 289   14.491990
## 290   13.335779
## 294   19.593569
## 297    7.636752
## 301    9.525256
## 305   14.106046
## 307   17.980909
## 311    7.152792
## 318    8.861262
## 319   10.499834
## 323    9.729981
## 325   14.679918
## 330   13.753284
## 331    7.846424
## 335    8.588259
## 342    9.146283
## 347   15.763441
## 349   13.862317
## 356    8.679958
## 364   13.790483
## 369    9.011544
## 371    4.500805
## 375   19.398188
## 381   14.123547
```

```
## 383   10.295067
## 387    3.596155
## 390    2.559562
## 392   16.530967
```

Comparando os dados previstos com os reais

```
comparacao <- cbind(as.integer(previsao_G3), teste$G3)
class(comparacao)
```

```
## [1] "matrix"
```

```
comparacao <- as.data.frame(comparacao)
colnames(comparacao) <- c("Previsto", "Real")
View(comparacao)
```

Tratando valores negativos

```
tratamento <- function(x){
  if (x < 0) {
    return(0)
  } else{
    return(x)
  }
}

comparacao$Previsto <- sapply(comparacao$Previsto, tratamento)
View(comparacao)
```

Calculando o erro médio

MSE:

```
mse <- mean((comparacao$Real - comparacao$Previsto)^2)
print(mse) # Distancia dos valores previstos para os valores observados
```

```
## [1] 5.915254
```

Calculando R Squared

```
SSE = sum((comparacao$Previsto - comparacao$Real)^2)
SST = sum((mean(df$G3) - comparacao$Real)^2)
```

R-Squared

```
R2 = 1 - (SSE/SST)
R2*100 # Percentual da precisão do modelo criado
```

```
## [1] 75.97239
```