# Student-Math-Exam-Performance-Analysis.R

troy

2021-08-12

**Import the data and inspect the data**

```
suppressPackageStartupMessages({
library(dplyr)
library(ggplot2)
library(caret)
library(GGally)
library(tidyr)
library(rmarkdown)
library(leaps)
library(MASS)
library(rpart)
library(rpart.plot)
select <- dplyr::select
})
```

Load the data

```
math <- read.csv("/Users/troy/Desktop/student/student-mat.csv", sep=";")
```

Check out the data type

```
str(math)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ 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      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ 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 : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ 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 ...
```

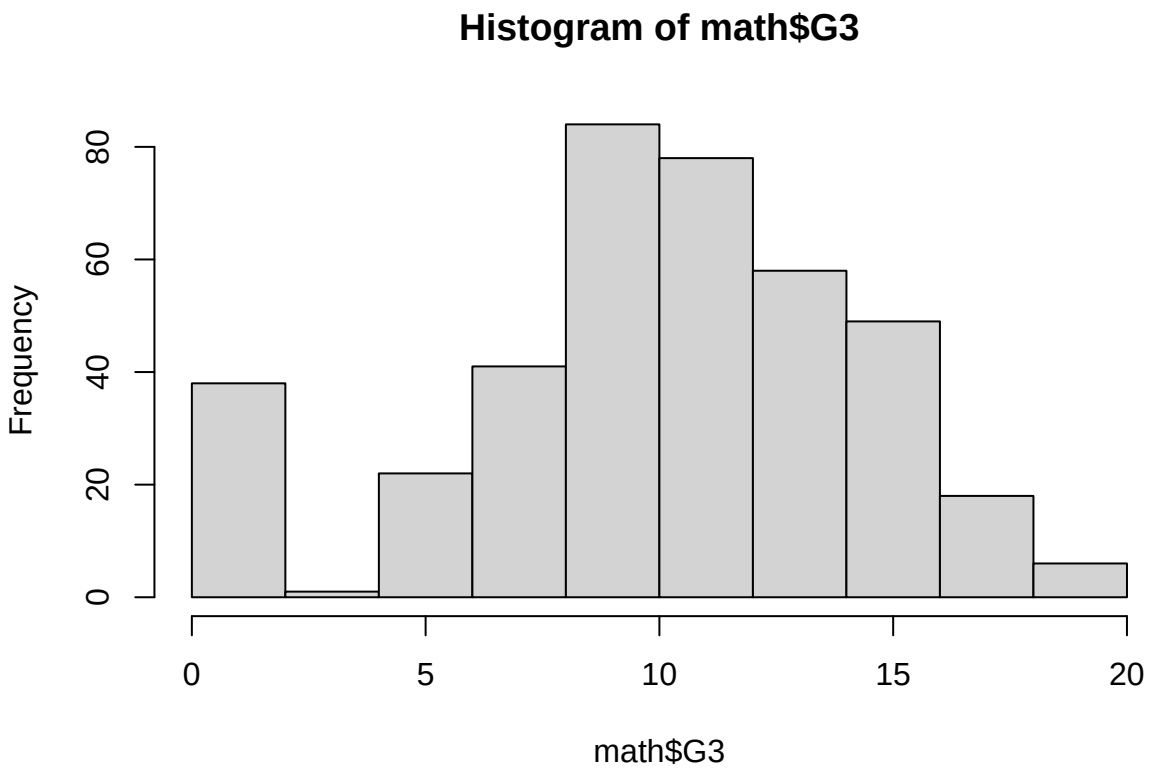
See if there are any null values

```
sapply(math,function(x) sum(is.na(x)))
```

```
##     school        sex        age    address    famsize    Pstatus       Medu
##          0          0          0          0          0          0          0
##       Fedu       Mjob       Fjob     reason   guardian traveltime  studytime
##          0          0          0          0          0          0          0
##   failures  schoolsup     famsup       paid activities    nursery     higher
##          0          0          0          0          0          0          0
##   internet   romantic     famrel   freetime      goout       Dalc       Walc
##          0          0          0          0          0          0          0
##     health   absences         G1         G2         G3
##          0          0          0          0          0
```

Check out the distribution of the G3 score, there are value 0

```
hist(math$G3)
```

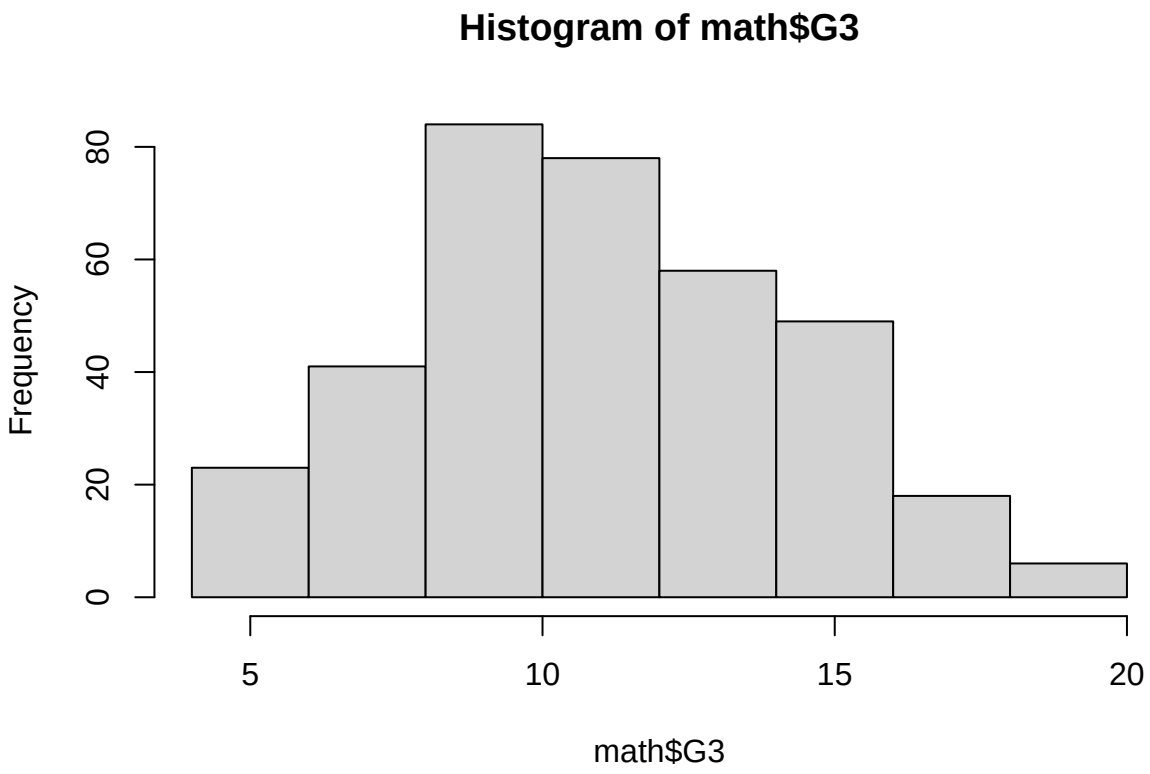



Drop rows where G3 value is 0

```
math <- subset(math,G3 != 0)
```

Check again

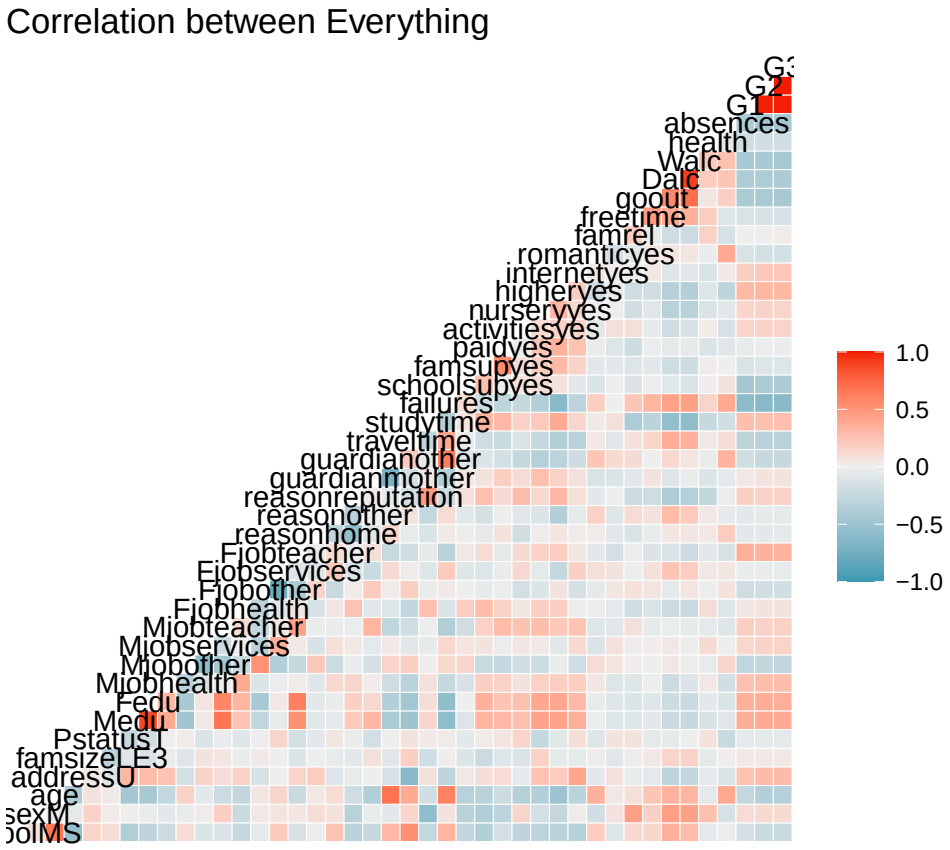
```
hist(math$G3)
```




## Exploratory Data Analysis

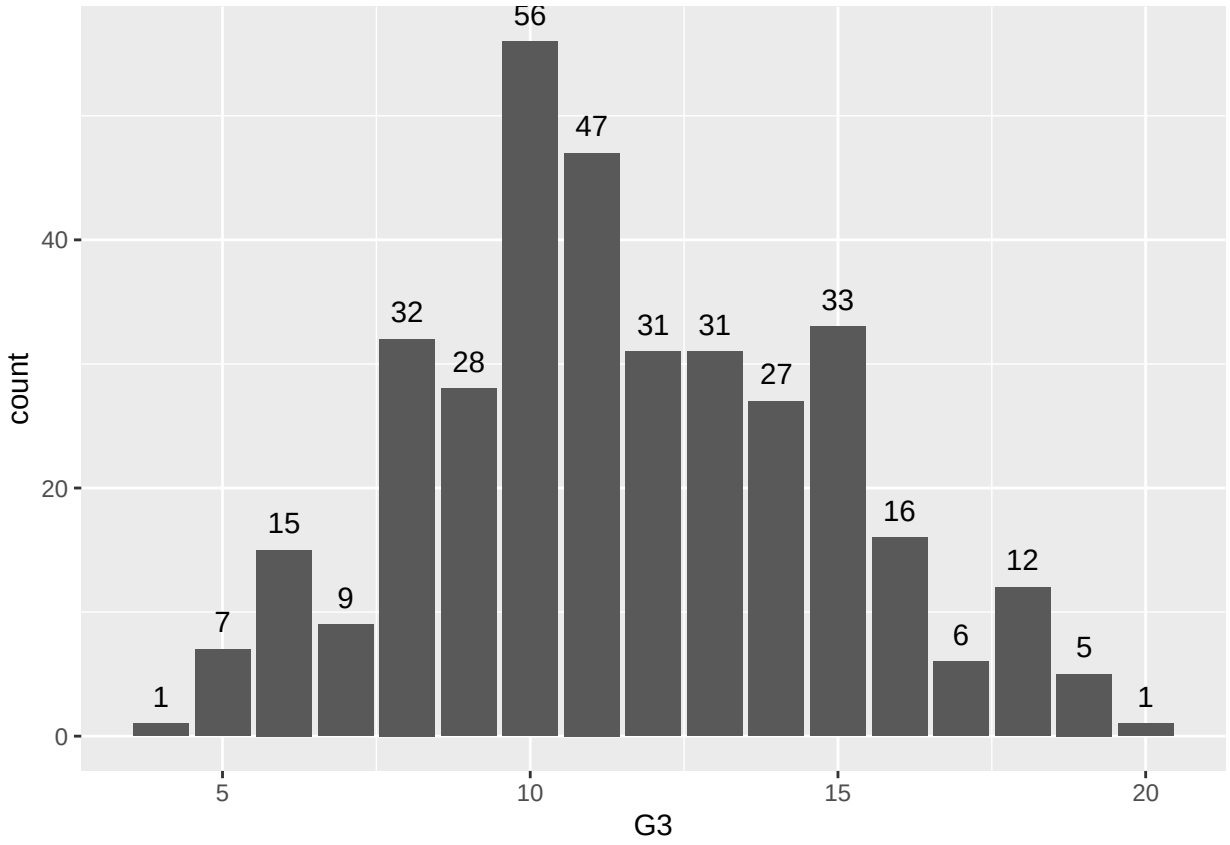
Make dummy variables for correlation matrix

```
dmy <- dummyVars("~.", data=math, fullRank=TRUE)
dummy_math <- data.frame(predict(dmy, newdata=math))
cor(dummy_math, method = "pearson", use = "complete.obs")%>%
  ggcorr(label = F)+ ggtitle("Correlation between Everything")
```



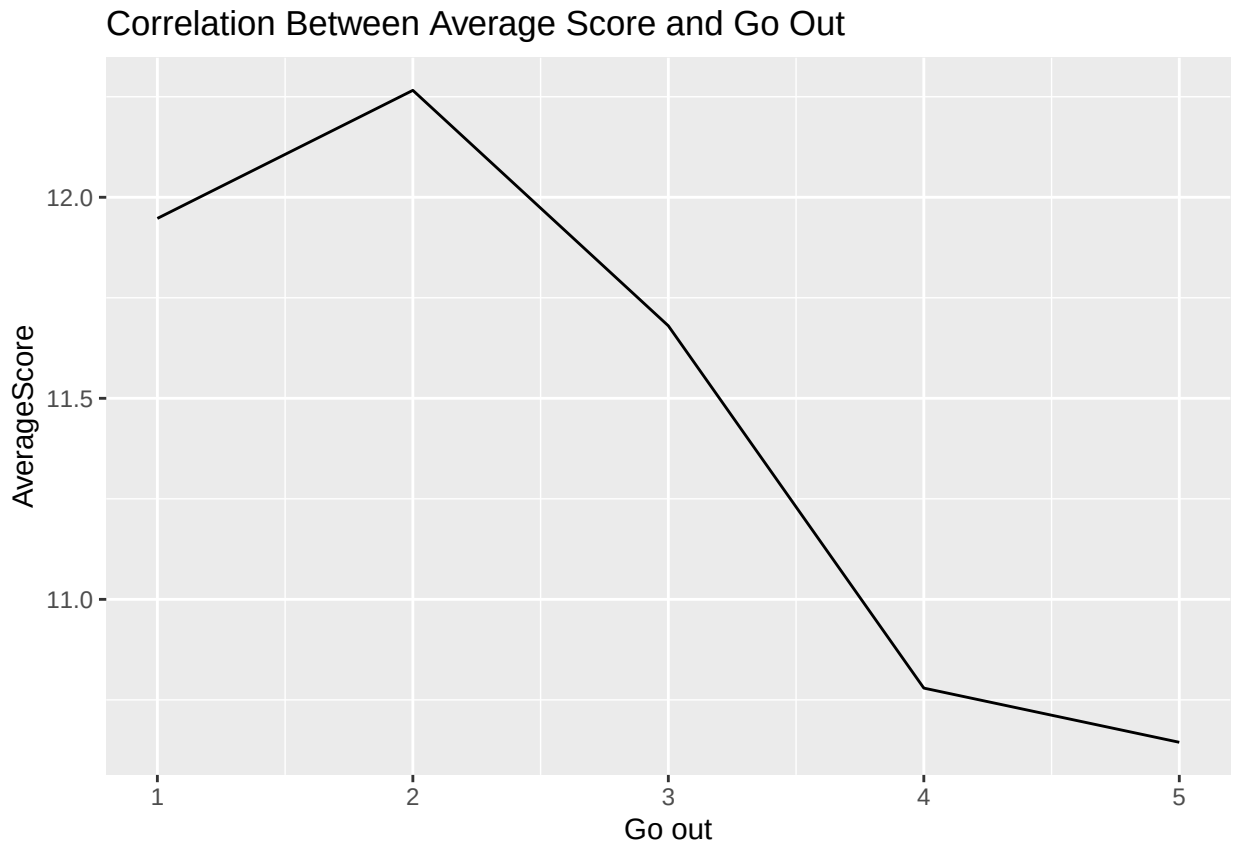

### Distribution of the G3 score

```
math%>%
  ggplot(aes(x=G3))+
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=-0.8)
```

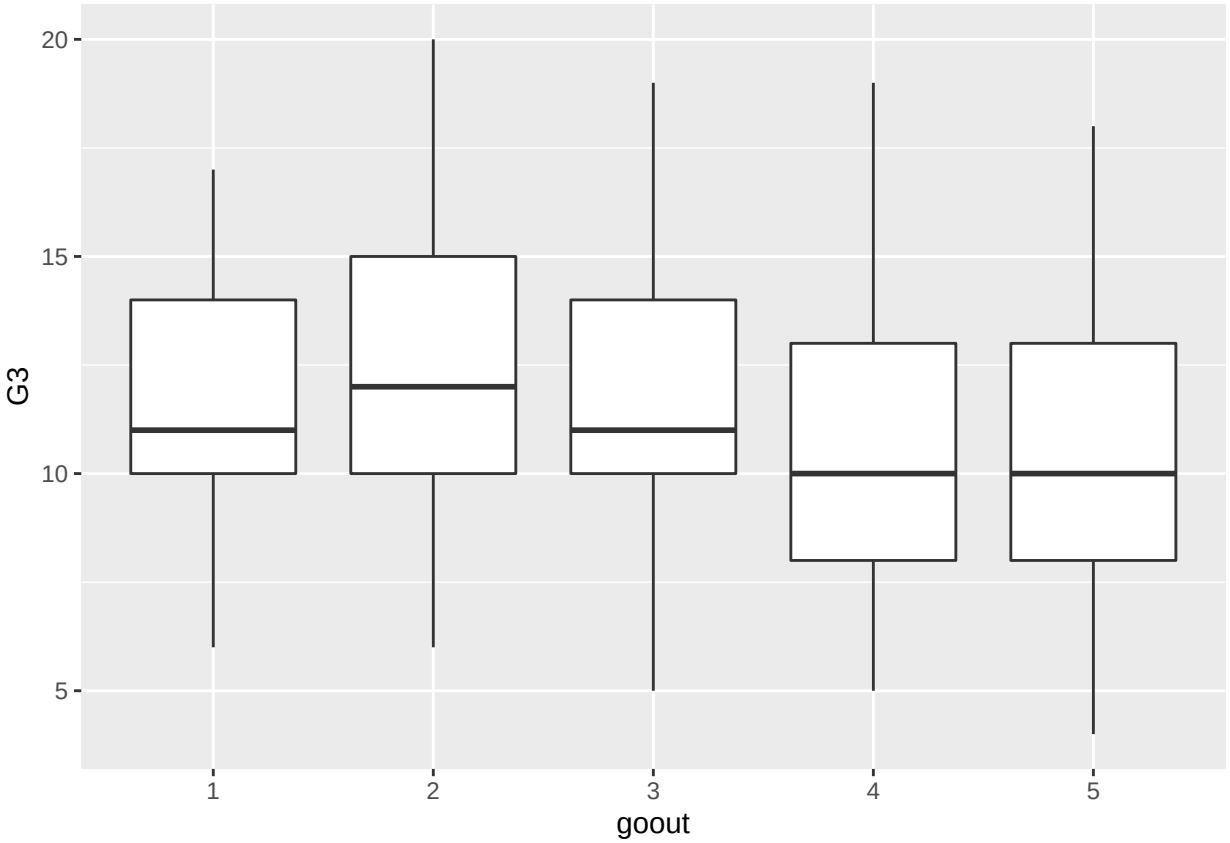



### Relationship between goout and AverageScore

```
math$goout <- as.factor(math$goout)
math_goout <- math %>%
  group_by(goout)%>%
  summarise(AverageScore=mean(G3))%>%
  arrange(desc(AverageScore))
# Make a line graph with `goout` and `AverageScore`
math_goout %>%
  ggplot(aes(x=as.numeric(goout), y=AverageScore))+
  geom_line(stat="identity")+
  labs(x="Go out")+
  ggtitle("Correlation Between Average Score and Go Out")
```

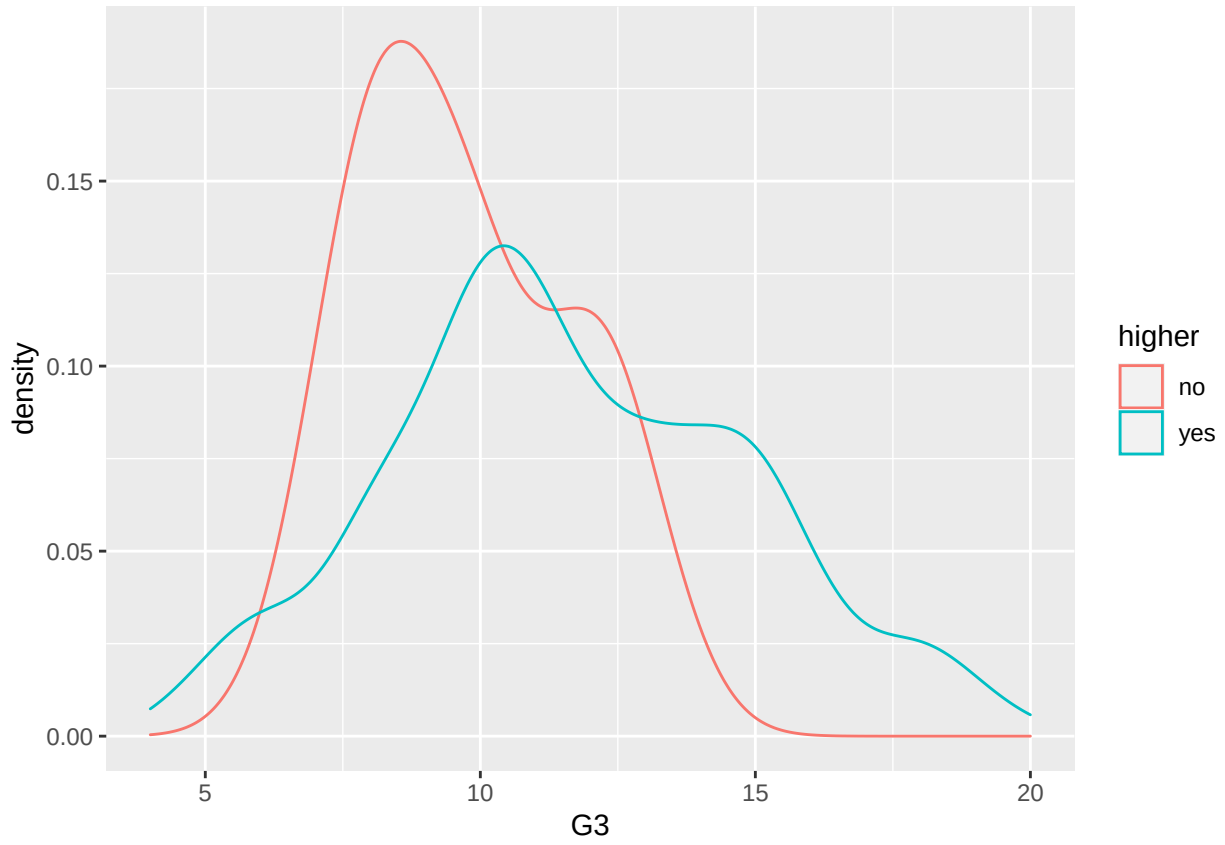



```
math %>%
  ggplot(aes(x=goout, y=G3))+
  geom_boxplot()
```




**higher vs G3**

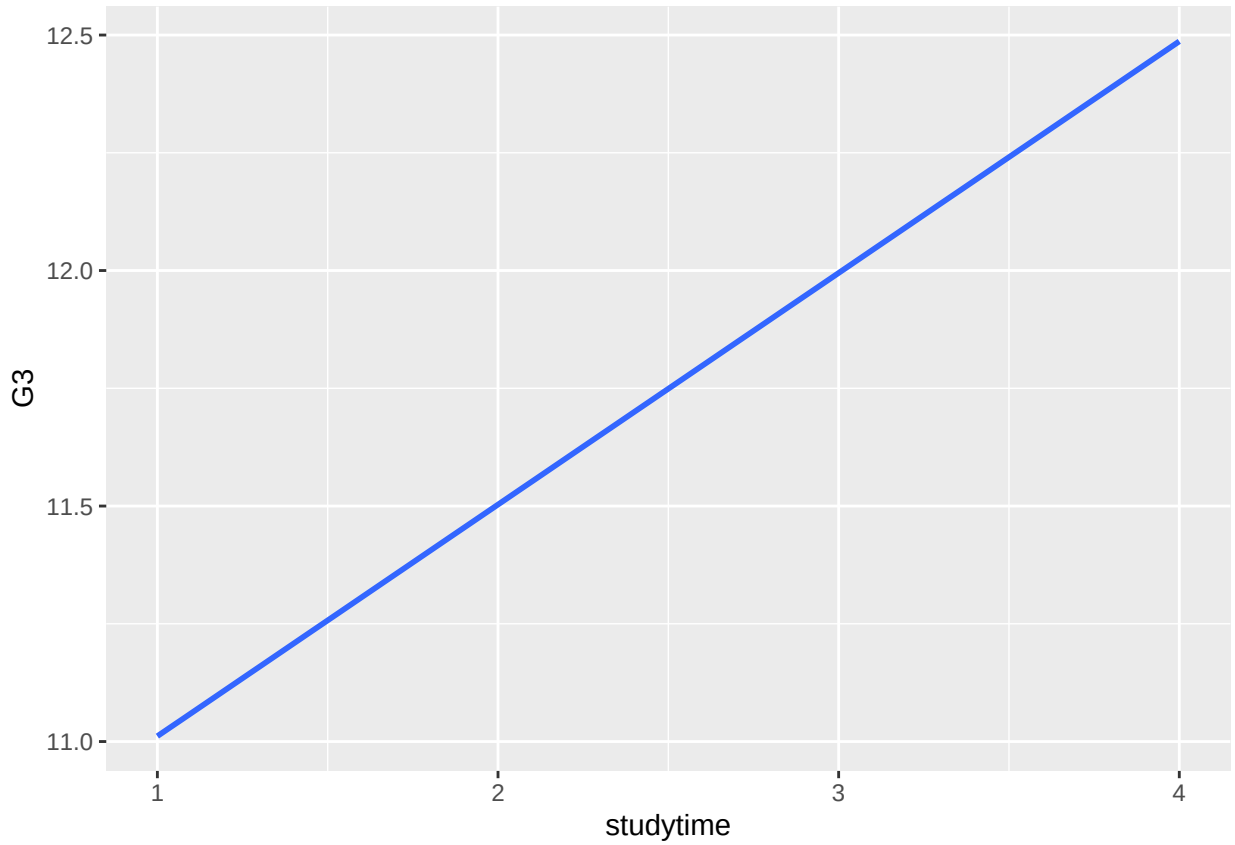
```
math%>%
  ggplot(aes(x=G3))+
  geom_density(aes(color=higher))
```




**studytime vs G3**

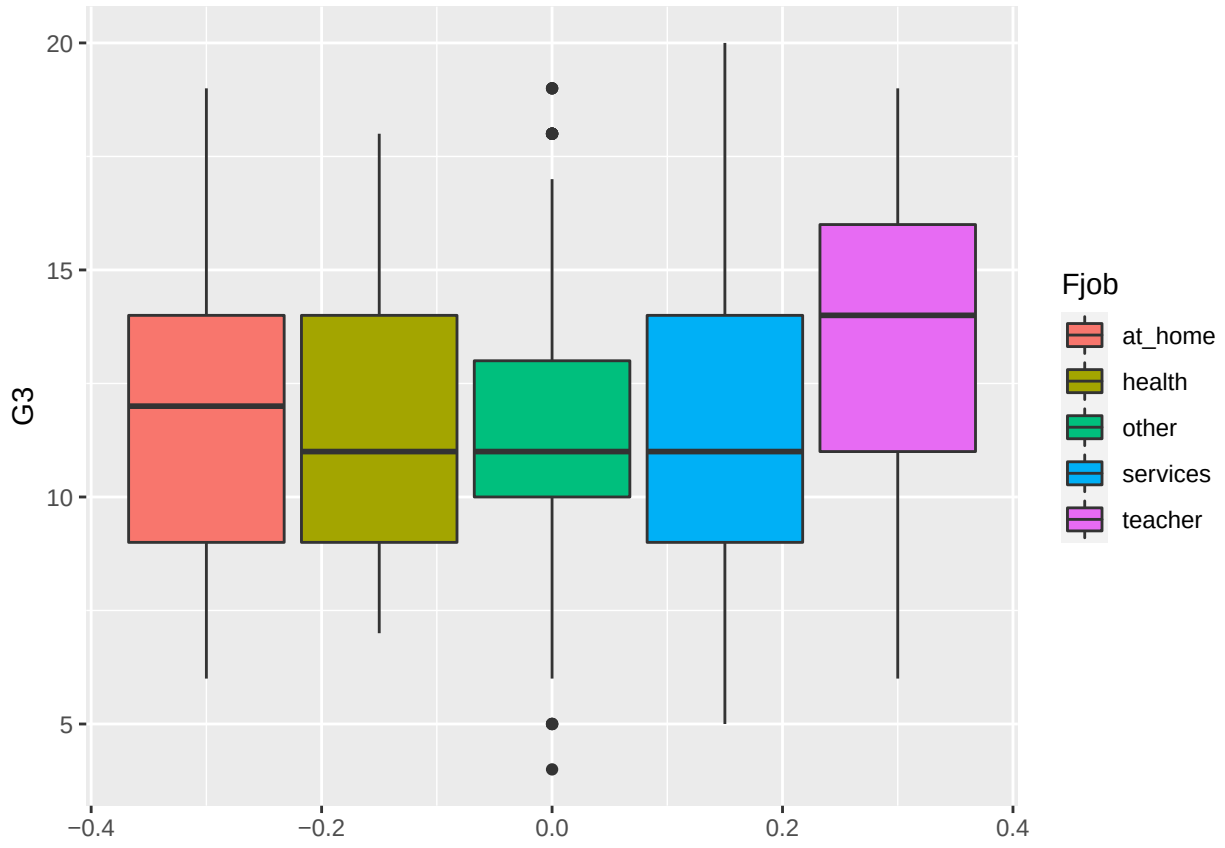
```
math%>%
  ggplot(aes(x=studytime, y=G3))+
  geom_smooth(method="lm", se=F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



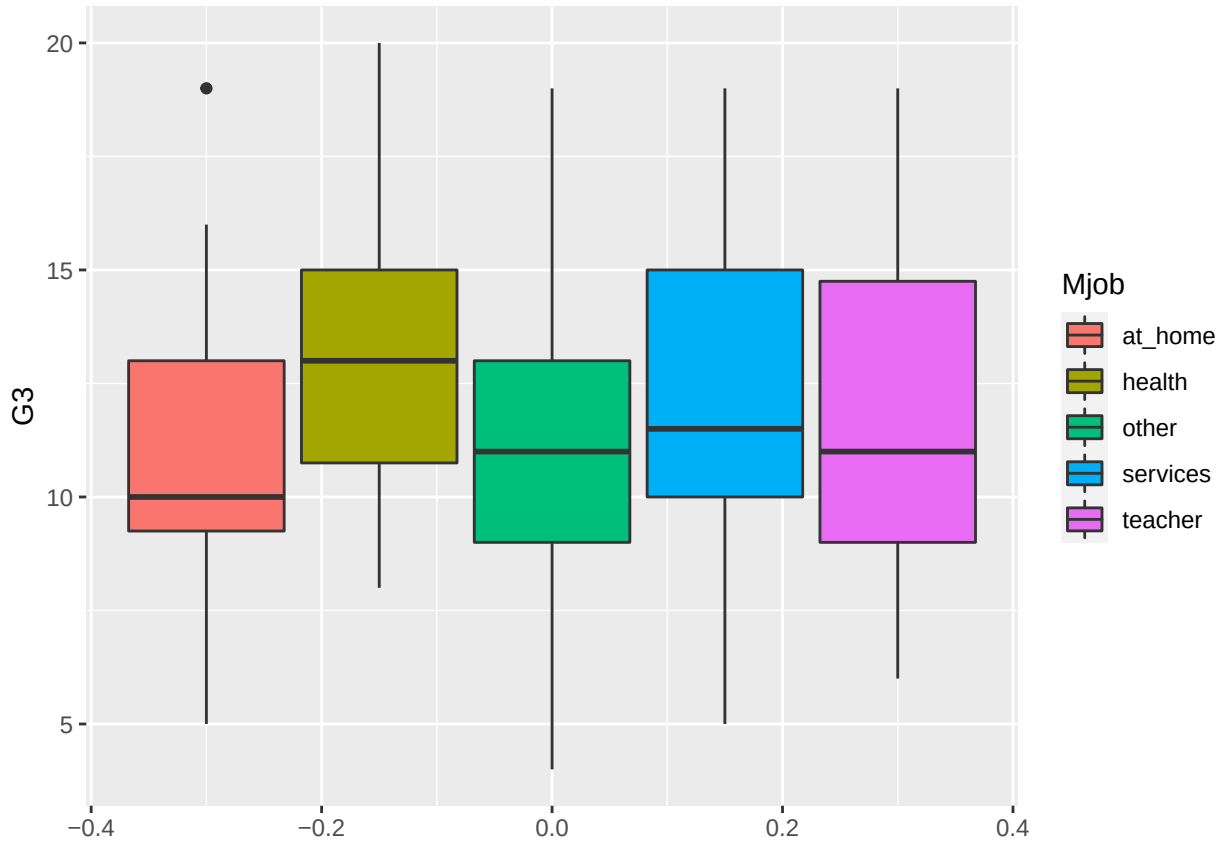

### Fjob vs G3

```
math%>%
  ggplot(aes(y=G3, fill=Fjob))+
  geom_boxplot()
```




**Mjob vs G3**

```
math%>%
  ggplot(aes(y=G3, fill=Mjob))+
  geom_boxplot()
```

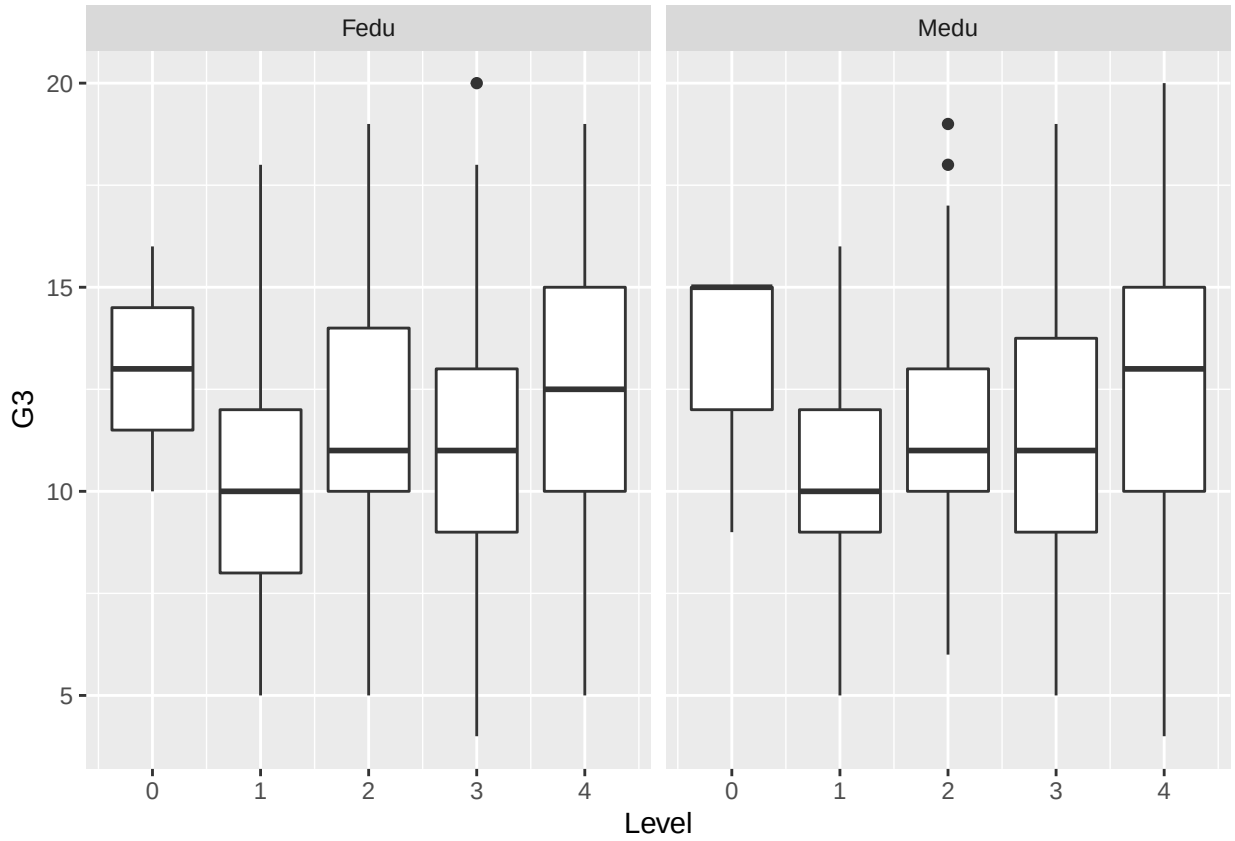



### Medu and Fedu vs G3

```
math_pedu <- math %>%
  gather(Fedu, Medu, key="Gender", value="Level")
math_pedu <- math_pedu%>%select(G3, Gender, Level)

math_pedu[,2] <- as.factor(math_pedu[,2])
str(math_pedu)
```

```
## 'data.frame':    714 obs. of  3 variables:
##  $ G3    : int  6 6 10 15 10 15 11 6 19 15 ...
##  $ Gender: Factor w/ 2 levels "Fedu","Medu": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Level : int  4 1 1 2 3 3 2 4 2 4 ...
```

```
math_pedu%>%
  ggplot(aes(x=Level, y=G3, group=Level))+
  geom_boxplot()+
  facet_wrap(~Gender)
```




```
# Check how many observations in the first level
str(filter(math_pedu, Level ==0))
```

```
## 'data.frame':    5 obs. of  3 variables:
##  $ G3    : int  10 16 9 15 15
##  $ Gender: Factor w/ 2 levels "Fedu","Medu": 1 1 2 2 2
##  $ Level : int  0 0 0 0 0
```

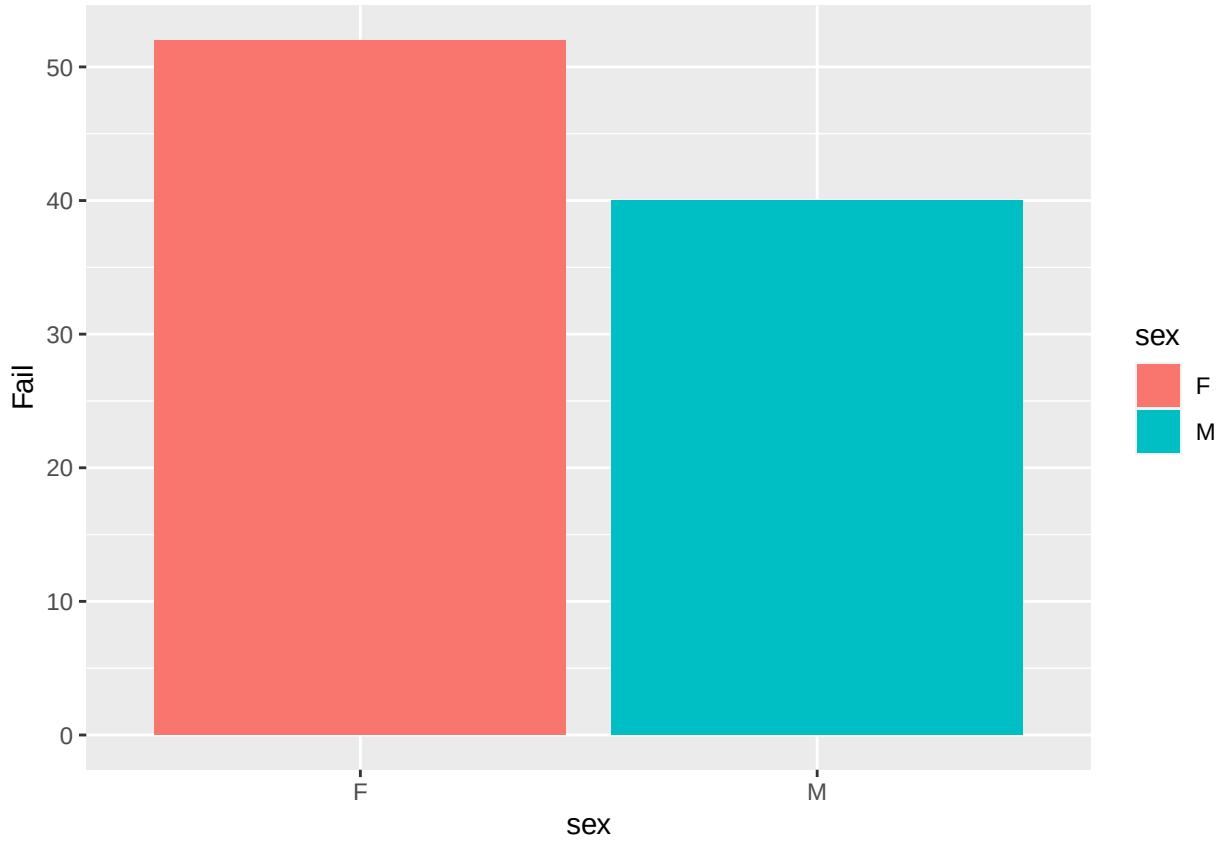
```
str(filter(math_pedu, Level ==1))
```

```
## 'data.frame':    121 obs. of  3 variables:
##  $ G3    : int  6 10 12 10 11 6 16 15 5 14 ...
##  $ Gender: Factor w/ 2 levels "Fedu","Medu": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Level : int  1 1 1 1 1 1 1 1 1 1 ...
```

### fail vs G3 by sex

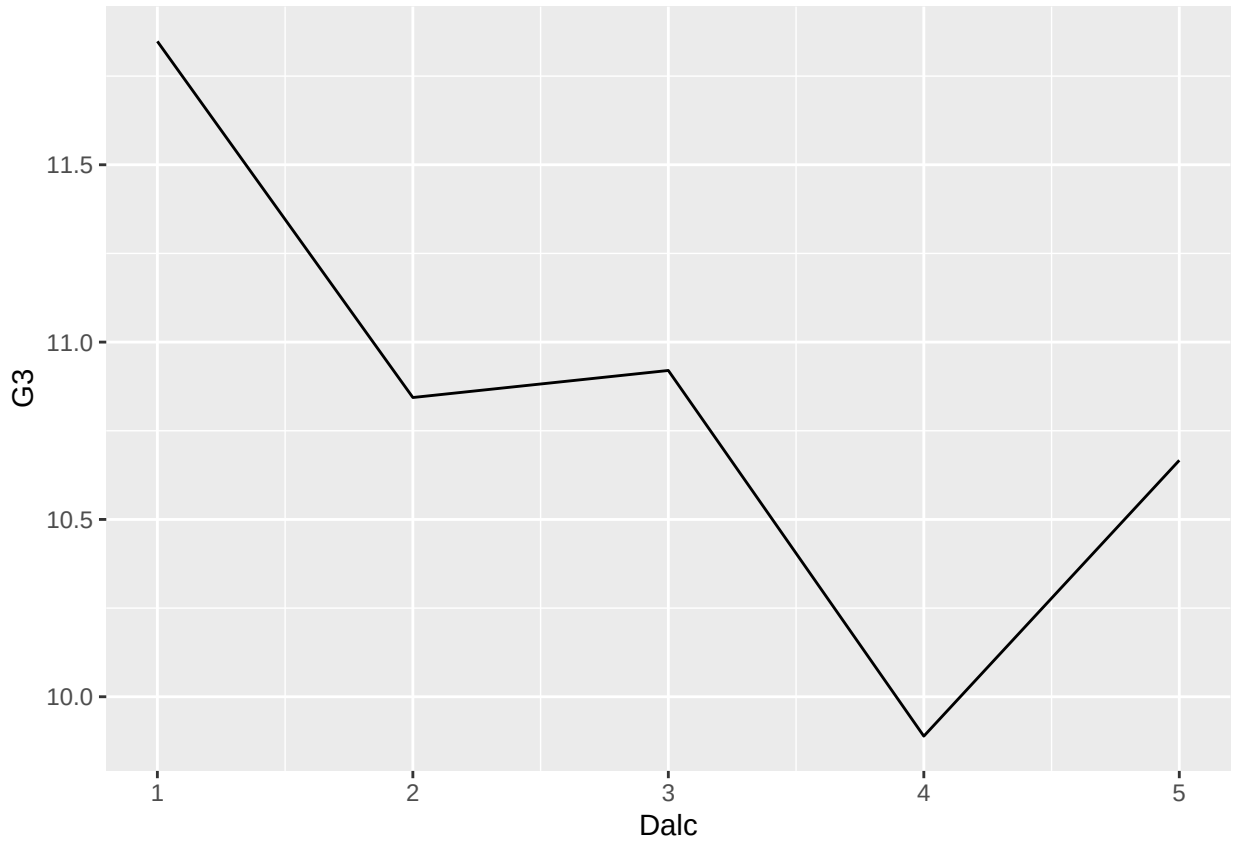
```
math_pf <- math %>%
  mutate(pass=ifelse(G3>=10, 1, 0), fail = ifelse(G3<10,1, 0)) %>%
  group_by(sex) %>%
  summarise(Pass=sum(pass),
            Fail=sum(fail))
```

```
# Let's make a bar chart base on the fail value for male and female
math_pf %>%
  ggplot(aes(x=sex, y=Fail, fill=sex))+
  geom_bar(stat="identity")
```



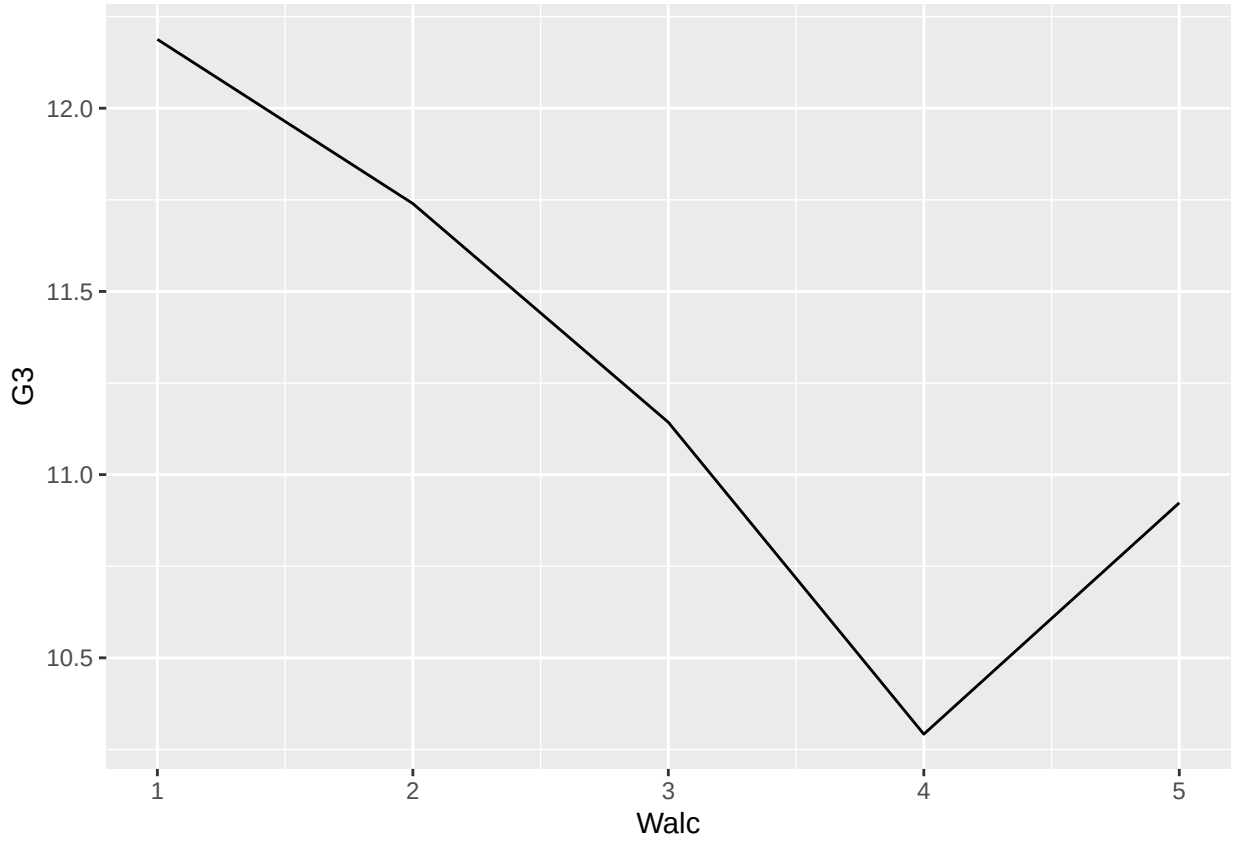

### Dalc vs G3

```
math_dalc <- math %>%
  group_by(Dalc)%>%
  aggregate(G3~Dalc, data=., mean)%>%
  arrange(desc(G3))
# Make a line graph with Dalc and G3
math_dalc %>%
  ggplot(aes(x=Dalc, y=G3))+
  geom_line(stat="identity")
```



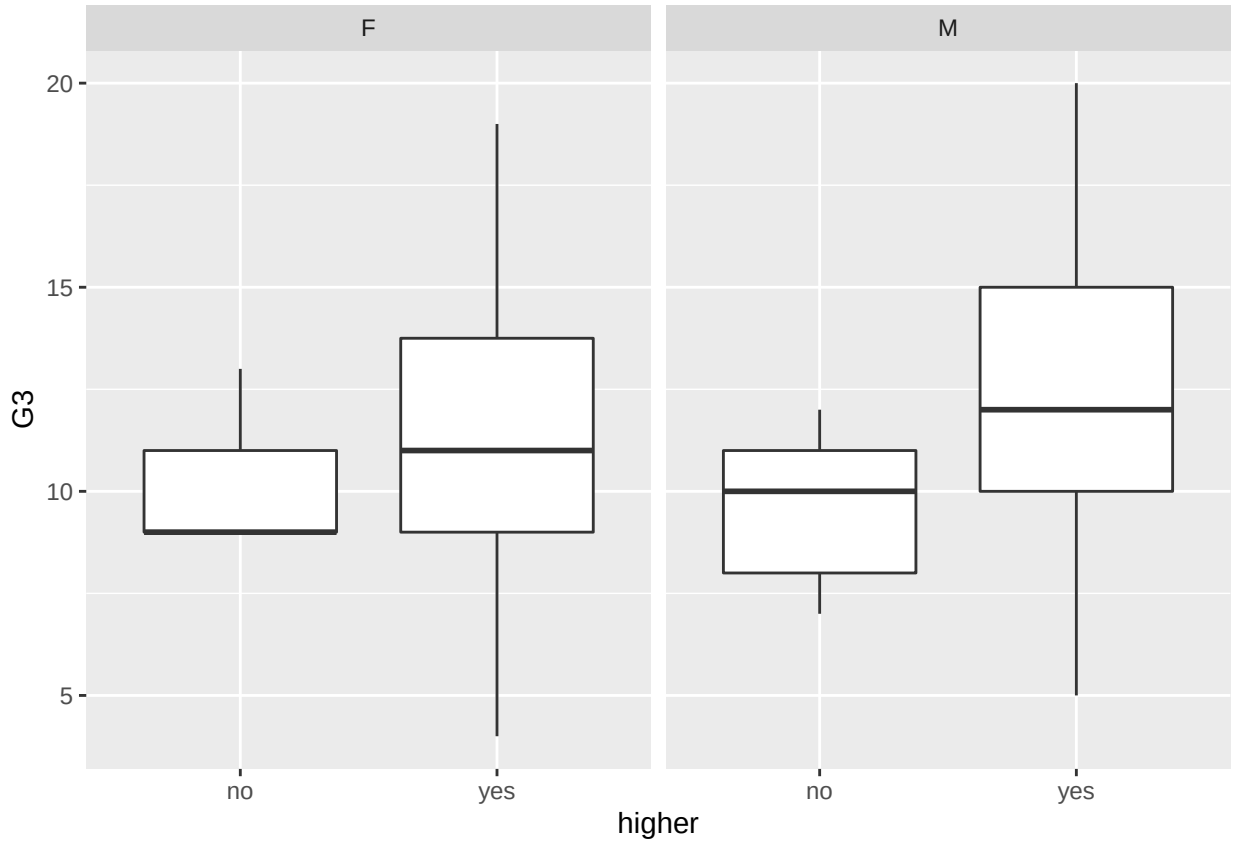

### Walc vs G3

```
math_walc <- math %>%
  aggregate(G3~Walc, data=., FUN = mean) %>%
  arrange(desc(G3))
# Make a line graph with Walc and G3
math_walc %>%
  ggplot(aes(x=Walc, y=G3))+
  geom_line(stat="identity")
```



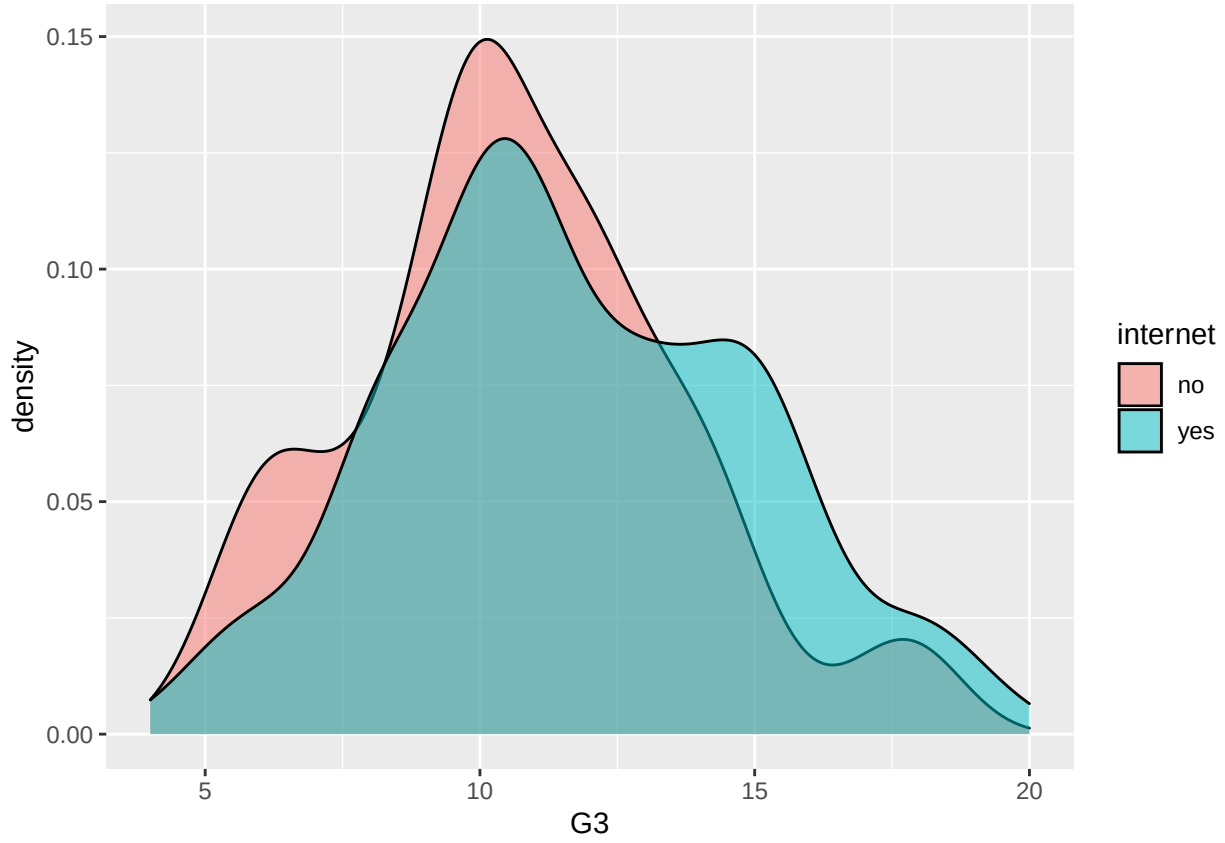

**highter vs G3 by sex**

```
math%>%
  ggplot(aes(x=higher, y=G3))+
  geom_boxplot()+
  facet_grid(~sex)
```




**internet vs G3**

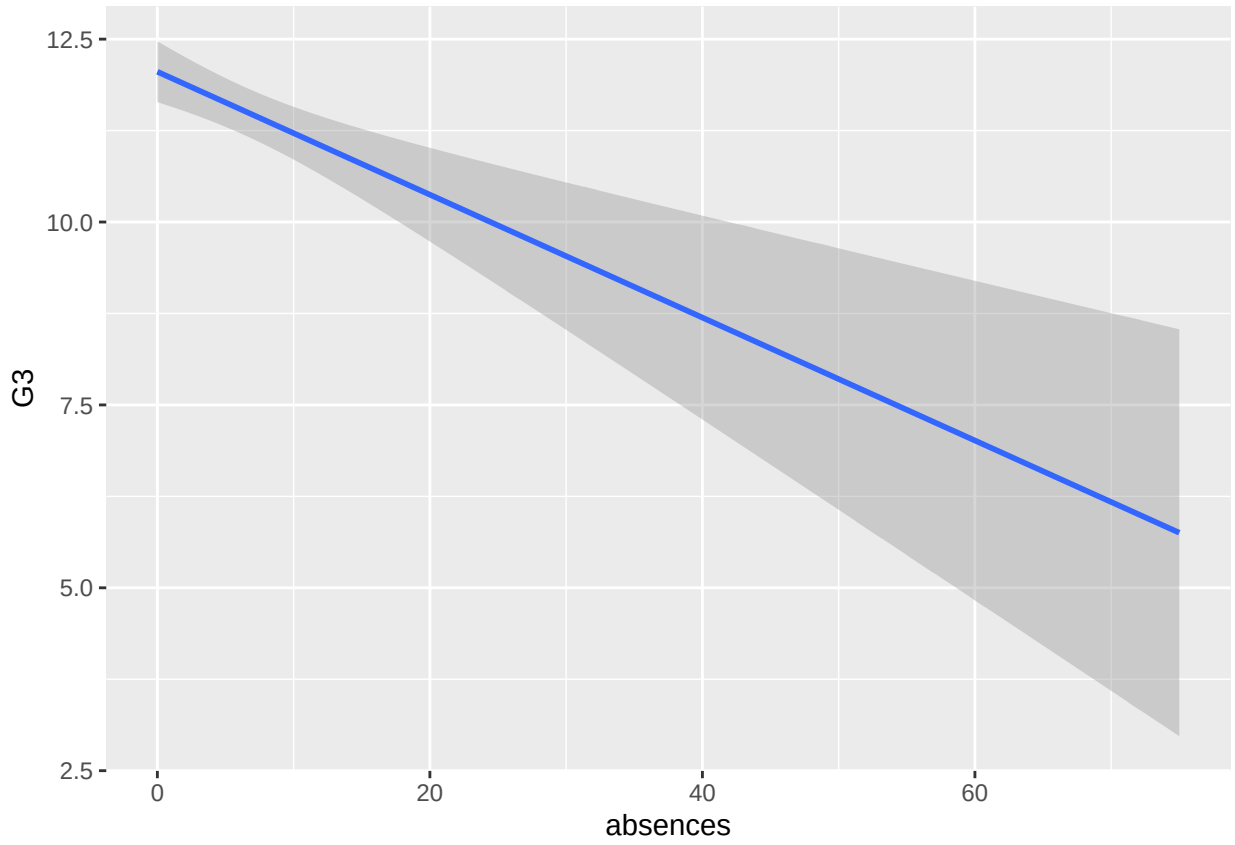
```
math%>%
  group_by(internet)%>%
  ggplot(aes(x=G3, fill=internet))+
  geom_density(alpha=0.5)
```




**absences vs G3**

```
math %>%
  ggplot(aes(x=absences, y=G3))+
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




#Linear Regression

```
set.seed(100)


# Conver chr to factors for the models
for (i in c(1:2, 4:6, 9:12, 16:23 )) {
  math[, i] <- as.factor(math[, i])
}
# Remove Fedu, Walc, G1
math <- subset(math, select=-c(Fedu, Walc, G1))

trainRowNumbers <- createDataPartition(math$G3, p=0.7, list=FALSE)
train_lm <- math[trainRowNumbers,]
test_lm <- math[-trainRowNumbers,]

linear_model <-  lm(G3~., train_lm)
summary(linear_model)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = train_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30581 -0.43529 -0.04585  0.47203  2.04268
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.114643   1.207458   0.095 0.924448
## schoolMS         -0.110684   0.195697  -0.566 0.572275
## sexM             -0.155653   0.123292  -1.262 0.208171
## age               0.022636   0.057149   0.396 0.692441
## addressU          0.216439   0.146256   1.480 0.140401
## famsizeLE3        0.067071   0.125574   0.534 0.593824
## PstatusT         -0.260607   0.189986  -1.372 0.171608
## Medu             -0.134299   0.070186  -1.913 0.057040 .
## Mjobhealth        0.417794   0.273981   1.525 0.128780
## Mjobother        -0.044377   0.180758  -0.246 0.806303
## Mjobservices      0.052873   0.200015   0.264 0.791774
## Mjobteacher       0.482038   0.259975   1.854 0.065110 .
## Fjobhealth        0.440464   0.343392   1.283 0.201009
## Fjobother         0.567356   0.251986   2.252 0.025382 *
## Fjobservices      0.456404   0.260852   1.750 0.081630 .
## Fjobteacher       0.488306   0.332833   1.467 0.143832
## reasonhome        0.186156   0.139453   1.335 0.183349
## reasonother       0.060324   0.204847   0.294 0.768679
## reasonreputation  0.039918   0.146114   0.273 0.784967
## guardianmother   -0.058176   0.135102  -0.431 0.667193
## guardianother    -0.285543   0.256802  -1.112 0.267438
## traveltime        0.060853   0.091734   0.663 0.507820
## studytime         0.002069   0.071943   0.029 0.977081
## failures          0.009076   0.097627   0.093 0.926022
## schoolsupyes     -0.208813   0.176679  -1.182 0.238584
## famsupyes         0.089162   0.119821   0.744 0.457630
## paidyes          -0.227894   0.117881  -1.933 0.054544 .
## activitiesyes     0.143757   0.117481   1.224 0.222444
## nurseryyes       -0.170736   0.137294  -1.244 0.215033
## higheryes         0.069777   0.316678   0.220 0.825819
## internetyes      -0.184483   0.162732  -1.134 0.258223
## romanticyes       0.130361   0.125328   1.040 0.299458
## famrel            0.237197   0.062062   3.822 0.000174 ***
## freetime          0.008640   0.058661   0.147 0.883046
## goout2           -0.361784   0.295815  -1.223 0.222692
## goout3           -0.325317   0.288634  -1.127 0.260984
## goout4           -0.534823   0.299149  -1.788 0.075241 .
## goout5           -0.333026   0.314789  -1.058 0.291296
## Dalc             -0.012890   0.067440  -0.191 0.848599
## health           -0.092961   0.039508  -2.353 0.019544 *
## absences         -0.012499   0.007397  -1.690 0.092580 .
## G2                0.961220   0.020366  47.196  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8023 on 211 degrees of freedom
## Multiple R-squared:  0.9458, Adjusted R-squared:  0.9353
## F-statistic: 89.85 on 41 and 211 DF,  p-value: < 2.2e-16
```

##Evaluate it

```
predict_lm = predict(linear_model, data = test_lm)
results <- cbind(predict_lm, test_lm$G3)
```

```
## Warning in cbind(predict_lm, test_lm$G3): number of rows of result is not a
## multiple of vector length (arg 2)
```

```
colnames(results) <- c('Predicted', 'Real')
results <- as.data.frame(results)

RMSE(results$Predicted, results$Real)
```

```
## [1] 4.546931
```

### Feature selection

```
step.model <- stepAIC(linear_model, direction = "backward",
                      trace = F)
step.model$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + health + absences + G2
##
## Final Model:
## G3 ~ Pstatus + Medu + Mjob + paid + famrel + health + absences +
##     G2
##
##
##            Step Df     Deviance Resid. Df Resid. Dev        AIC
## 1                                     211   135.8212  -73.37870
## 2      - reason  3 1.2364481825       214   137.0576  -77.08594
## 3       - goout  4 2.5357791390       218   139.5934  -80.44782
## 4    - guardian  2 0.6231936497       220   140.2166  -83.32085
## 5        - Fjob  4 3.0163964443       224   143.2330  -85.93593
## 6   - studytime  1 0.0001874887       225   143.2332  -87.93560
## 7    - freetime  1 0.0194920999       226   143.2527  -89.90117
## 8         - age  1 0.0304860347       227   143.2832  -91.84733
## 9    - failures  1 0.0819643439       228   143.3651  -93.70265
## 10     - higher  1 0.0891940357       229   143.4543  -95.54530
## 11    - famsize  1 0.1182371936       230   143.5726  -97.33685
## 12       - Dalc  1 0.2571573463       231   143.8297  -98.88410
## 13     - famsup  1 0.2783768039       232   144.1081 -100.39491
## 14 - activities  1 0.4550050295       233   144.5631 -101.59734
## 15 - traveltime  1 0.5379194160       234   145.1010 -102.65768
```

```
## 16    - address  1 0.5849691601       235   145.6860 -103.63977
## 17   - romantic  1 0.6452500918       236   146.3313 -104.52170
## 18    - nursery  1 0.8016986615       237   147.1330 -105.13938
## 19        - sex  1 0.7257948234       238   147.8587 -105.89442
## 20   - internet  1 0.7560846127       239   148.6148 -106.60398
## 21     - school  1 0.7568093096       240   149.3716 -107.31887
## 22  - schoolsup  1 0.6965134344       241   150.0682 -108.14188
```

```
summary(step.model)
```

```
##
## Call:
## lm(formula = G3 ~ Pstatus + Medu + Mjob + paid + famrel + health +
##     absences + G2, data = train_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15235 -0.41706 -0.06165  0.55269  2.32306
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.588745   0.388522   1.515   0.1310
## PstatusT     -0.271146   0.173433  -1.563   0.1193
## Medu         -0.103202   0.061205  -1.686   0.0931 .
## Mjobhealth    0.414653   0.246202   1.684   0.0934 .
## Mjobother    -0.036961   0.162183  -0.228   0.8199
## Mjobservices  0.035731   0.176866   0.202   0.8401
## Mjobteacher   0.375078   0.228108   1.644   0.1014
## paidyes      -0.220529   0.102381  -2.154   0.0322 *
## famrel        0.235743   0.057735   4.083 6.05e-05 ***
## health       -0.079566   0.036259  -2.194   0.0292 *
## absences     -0.010511   0.006261  -1.679   0.0945 .
## G2            0.963624   0.016781  57.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7891 on 241 degrees of freedom
## Multiple R-squared:  0.9401, Adjusted R-squared:  0.9374
## F-statistic: 344.1 on 11 and 241 DF,  p-value: < 2.2e-16
```

### Evaluate again

```
predict_lm2 = predict(step.model, data = test_lm)
results2 <- cbind(predict_lm2, test_lm$G3)
```

```
## Warning in cbind(predict_lm2, test_lm$G3): number of rows of result is not a
## multiple of vector length (arg 2)
```

```
colnames(results2) <- c('Predicted', 'Real')
results2 <- as.data.frame(results2)

RMSE(results2$Predicted, results2$Real)
```

```
## [1] 4.553732
```

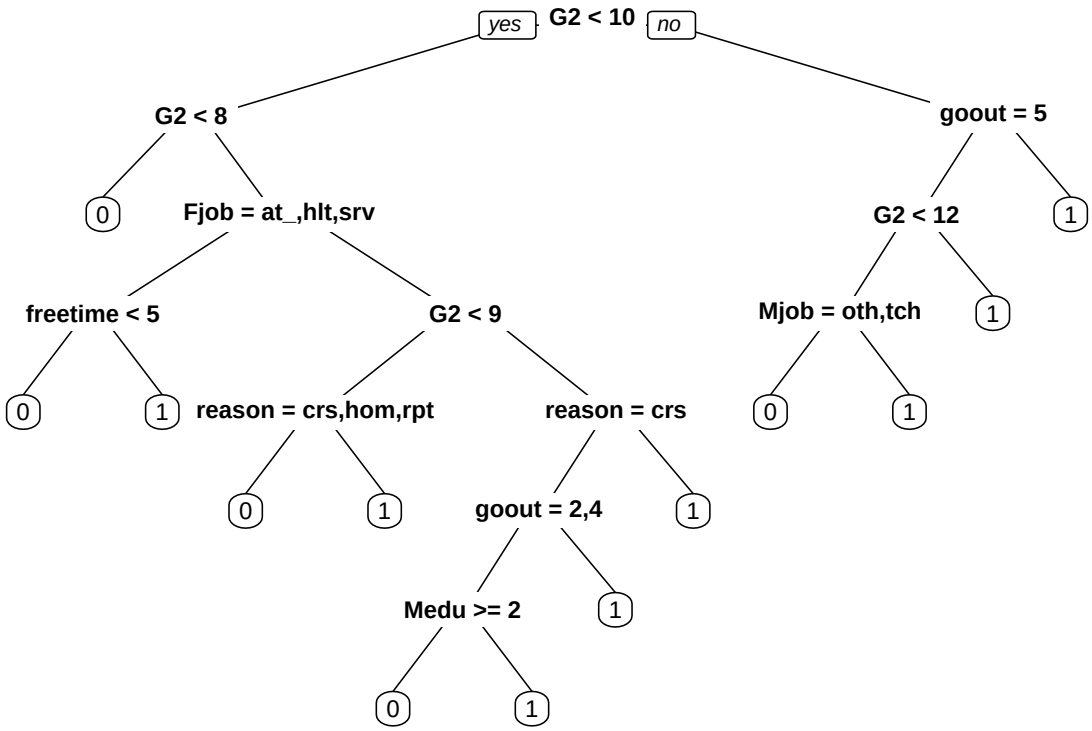
## Decision tree

```
math_tree <- math %>%
  mutate(pass=ifelse(G3>=10, 1, 0))

# Remove G3
math_tree <- subset(math_tree, select=-c(G3))

set.seed(100)
trainRowNumbers <- createDataPartition(math_tree$pass, p=0.7, list=FALSE)
train <- math_tree[trainRowNumbers,]
test <- math_tree[-trainRowNumbers,]
dim(train);dim(test)
```

```
## [1] 250  30
```

```
## [1] 107  30
```

```
tree <- rpart(pass ~ .,
              data = train,
              method = "class", minsplit = 2, minbucket=1)
prp(tree)
```



```r
# Model evaluation

predict_test = predict(tree, newdata = test, type = "class")
confMat2 <- table(test$pass, predict_test)
confMat2
```

```
##    predict_test
##      0  1
##   0 20  9
##   1  6 72
```

```r
accuracy2 <- sum(diag(confMat2))/sum(confMat2)
accuracy2
```

```
## [1] 0.8598131
```

## Findings

Base on the statistical analysis we found that `G2`, `health`, `famrel`, `paidyes` affect `G3` score, It's interesting to see that with more extra paid class, it actually decreases the G3 performance

Some of the important factors that decide if the student will get a above average score are `G2`, `goout`, `Fjob`, `freetime`, `Mjob`. Also a very interesting pattern where if a student has gotten above average score in `G2` then if he or she goes out equal to 5 then he or she will get a above average score in `G3` and for those students who didn't get a above average `G2` score but got a score over 8 then the father's job plays a crucial part those whose father is working in `at_home`, `health`, `service` will have a higher chance achieve a above average G3 score