Social and Educational Impacts on Students' Performance

Library installations necesssary for project

```
library(ggplot2)
library(glmulti)
```

```
## Loading required package: rJava
```

```
library(plyr)
library(dplyr)
```

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

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

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

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(mlbench)
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

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

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(bnlearn)
```

```
##
## Attaching package: 'bnlearn'

## The following object is masked from 'package:stats':
##
##     sigma

library(pcalg)

##
## Attaching package: 'pcalg'

## The following objects are masked from 'package:bnlearn':
##
##     dsep, pdag2dag, shd, skeleton

##Installing source for plotting graphics (will be used for Bayesian Ne
twork)
source("https://bioconductor.org/biocLite.R")

## Bioconductor version 3.4 (BiocInstaller 1.24.0), ?biocLite for help

biocLite(c("Rgraphviz","RBGL"))

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.3 (2017-03-06).

## Installing package(s) 'Rgraphviz', 'RBGL'

##
## The downloaded binary packages are in
##  /var/folders/7j/z_jp4f0n7_38zpwtn9l_tw6w0000gn/T//RtmpH6z7zL/downlo
aded_packages

library(bnlearn)
library(Rgraphviz)

## Loading required package: graph

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following object is masked from 'package:bnlearn':
##
##     score

## The following object is masked from 'package:randomForest':
##
##     combine

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:rJava':
##
##     anyDuplicated, duplicated, sort, unique

## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit, which, which.max,
##     which.min

##
## Attaching package: 'graph'

## The following objects are masked from 'package:bnlearn':
##
##     degree, nodes, nodes<-

## The following object is masked from 'package:plyr':
##
##     join

## Loading required package: grid
```

Import student datasets who are enrolled in Portugese and Math Class. Based on the two dataset I found students that were enrolled both classes and will only use these students to conduct analysis.

```
mathclass=read.csv(file="/Users/suzannechung/Desktop/student-mat.csv",s
ep=";",header=TRUE)
portclass=read.csv(file="/Users/suzannechung/Desktop/student-por.csv",s
```

```
ep=";",header=TRUE)

cclass=merge(mathclass,portclass,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))
```

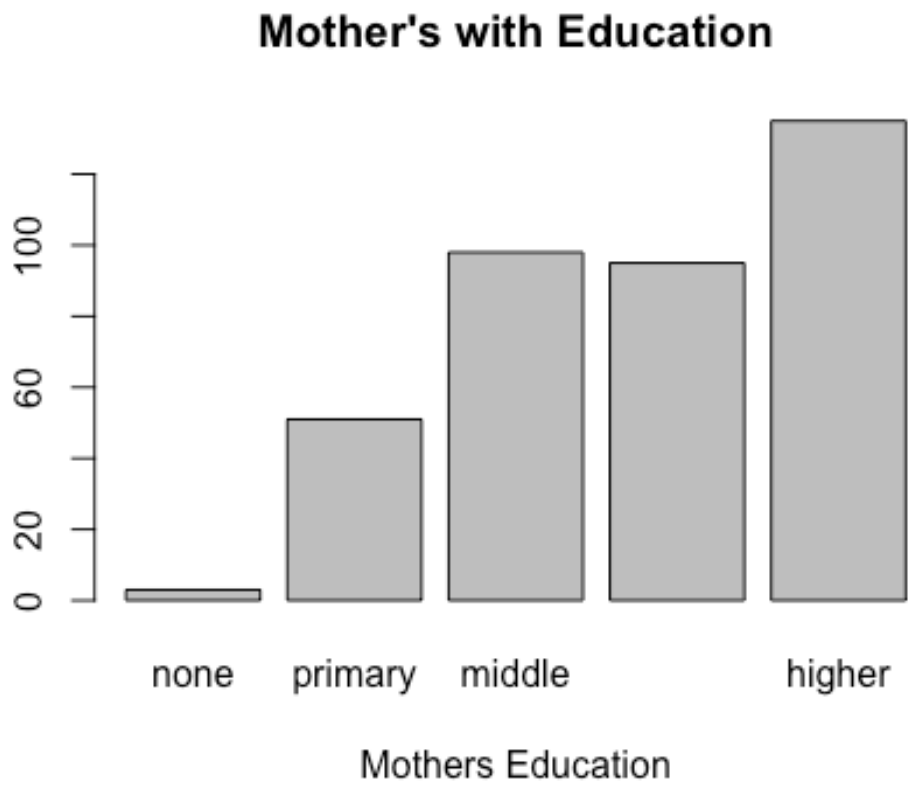
Data Preparation

All attributes have been converted into numeric values.

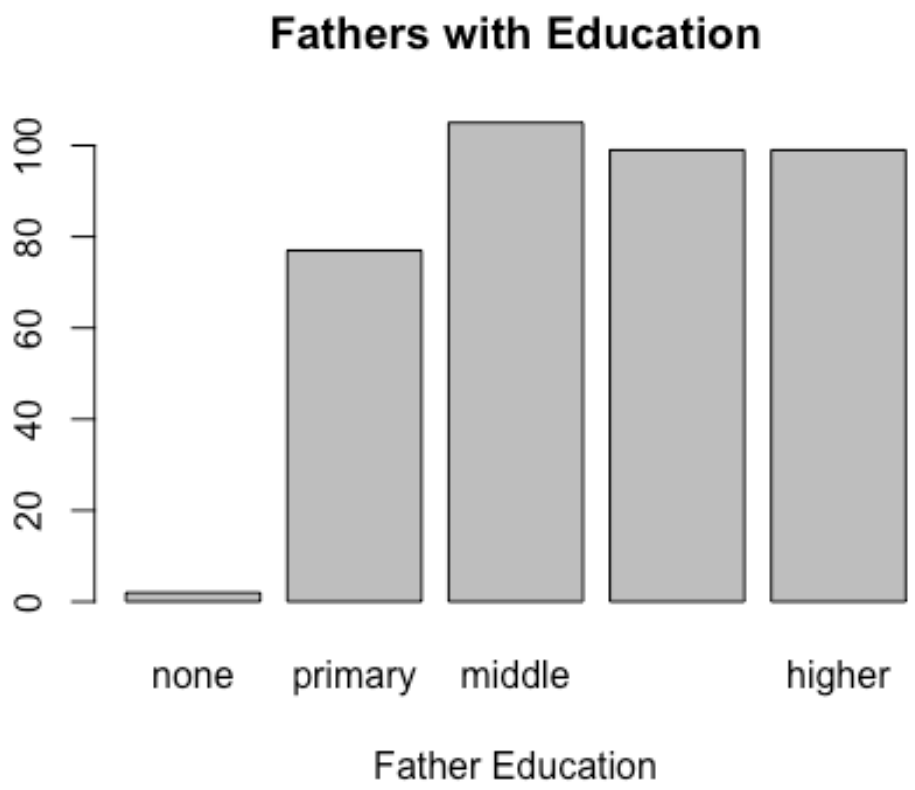
```
cclass[1:53]<- lapply(cclass[1:53], as.numeric)
mathclass[1:33] <- lapply(mathclass[1:33], as.numeric)
```

What are the parents education history? How many received higher education (university/college)?

```
barplot(table(cclass$Medu), names.arg = c("none", "primary", "middle", "secondary", "higher"), xlab = "Mothers Education", main = "Mother's with Education")
```



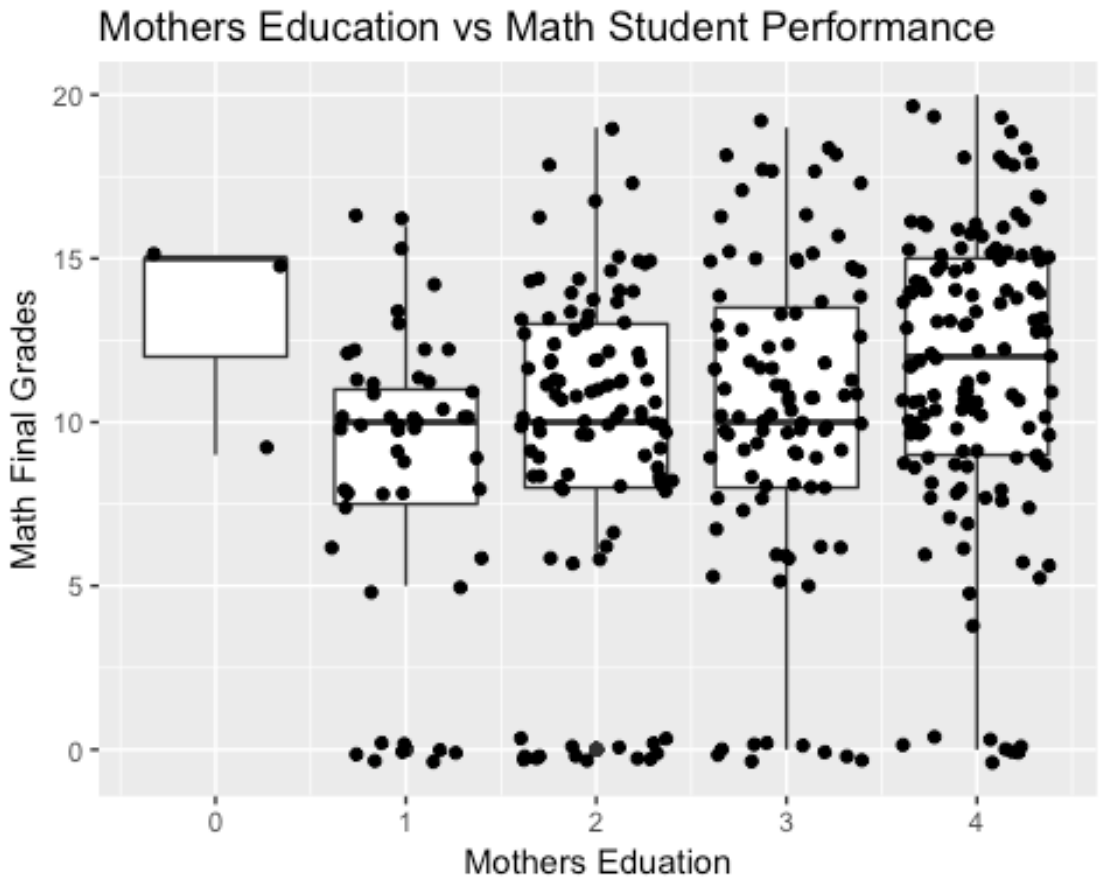

```
barplot(table(cclass$Fedu), names.arg = c("none", "primary", "middle", "secondary", "higher"), xlab = "Father Education",  main = "Fathers with Education")
```



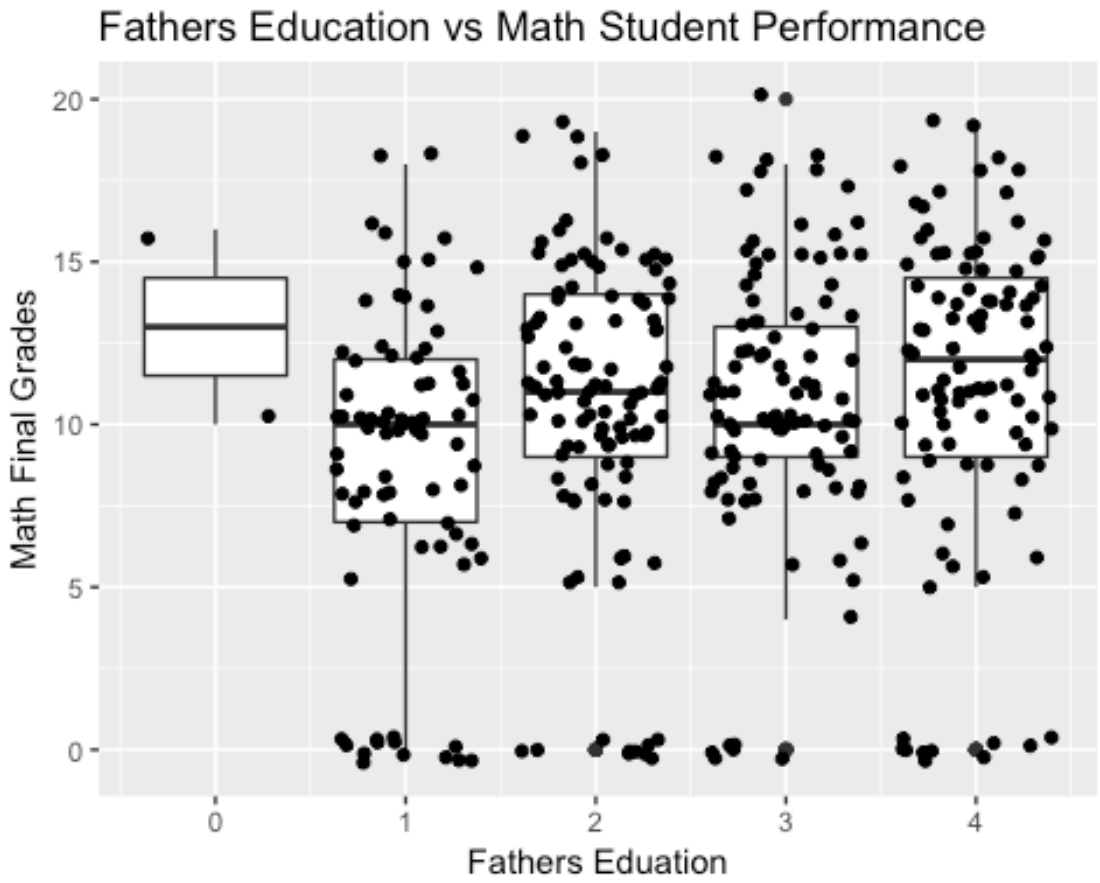
Analysis: Majority of the mothers received either middle education or higher. Whereas the Fathers education seem to be high in all educational classes.

Find the outliers in Math class in relation to Parent's Education and Final grades

```
ggplot(cclass, aes(x=Medu, y=G3.x, group=Medu))+geom_boxplot()+geom_jit
ter()+ xlab("Mothers Eduation")+ylab("Math Final Grades")+ggtitle("Moth
ers Education vs Math Student Performance")
```



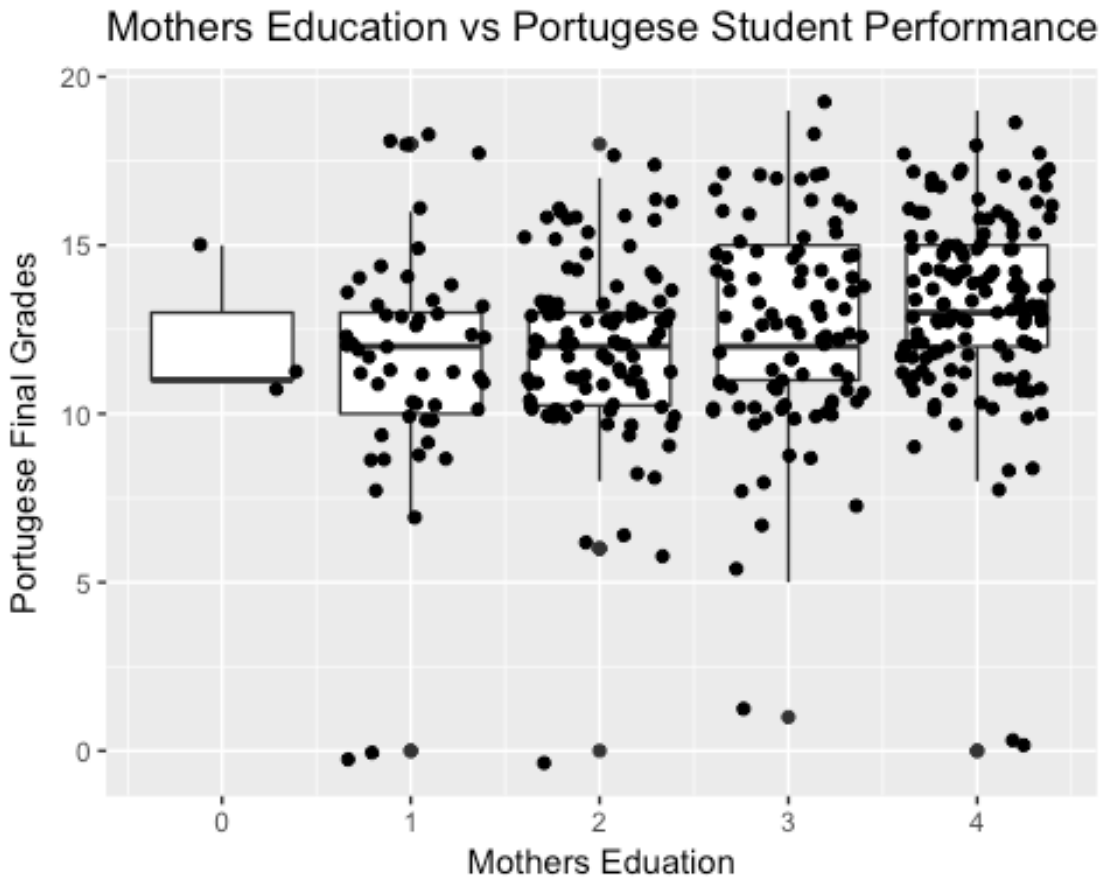

```
ggplot(cclass, aes(x=Fedu, y=G3.x, group=Fedu))+geom_boxplot()+geom_jit
ter()+ xlab("Fathers Eduation")+ylab("Math Final Grades")+ggtitle("Fath
ers Education vs Math Student Performance")
```



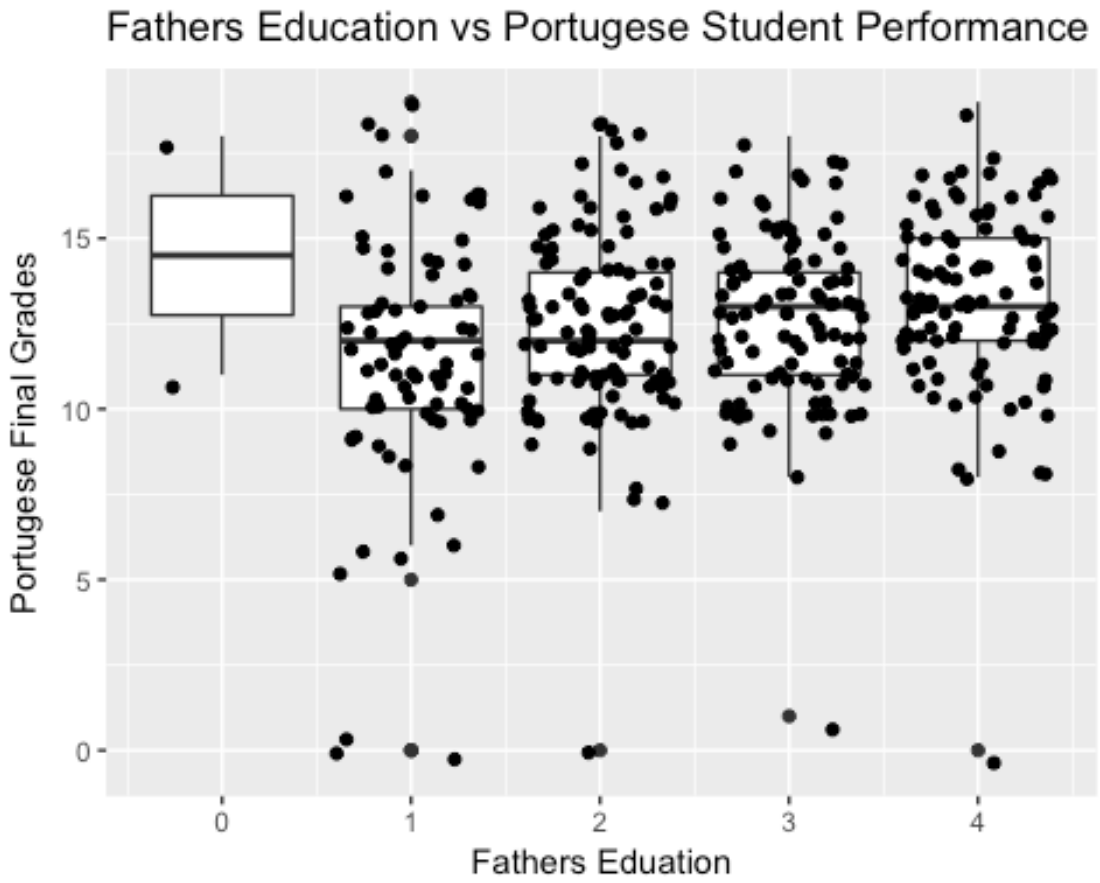
Analysis: As we can see in both plots, we have very distinct outliers in this dataset that will be removed in order to receive optimal results.

Find the outliers in Portugese classin relation to Parent's Education and Final grades

```
ggplot(cclass, aes(x=Medu, y=G3.y, group=Medu))+geom_boxplot()+geom_jit
ter()+ xlab("Mothers Eduation")+ylab("Portugese Final Grades")+ggtitle(
"Mothers Education vs Portugese Student Performance")
```



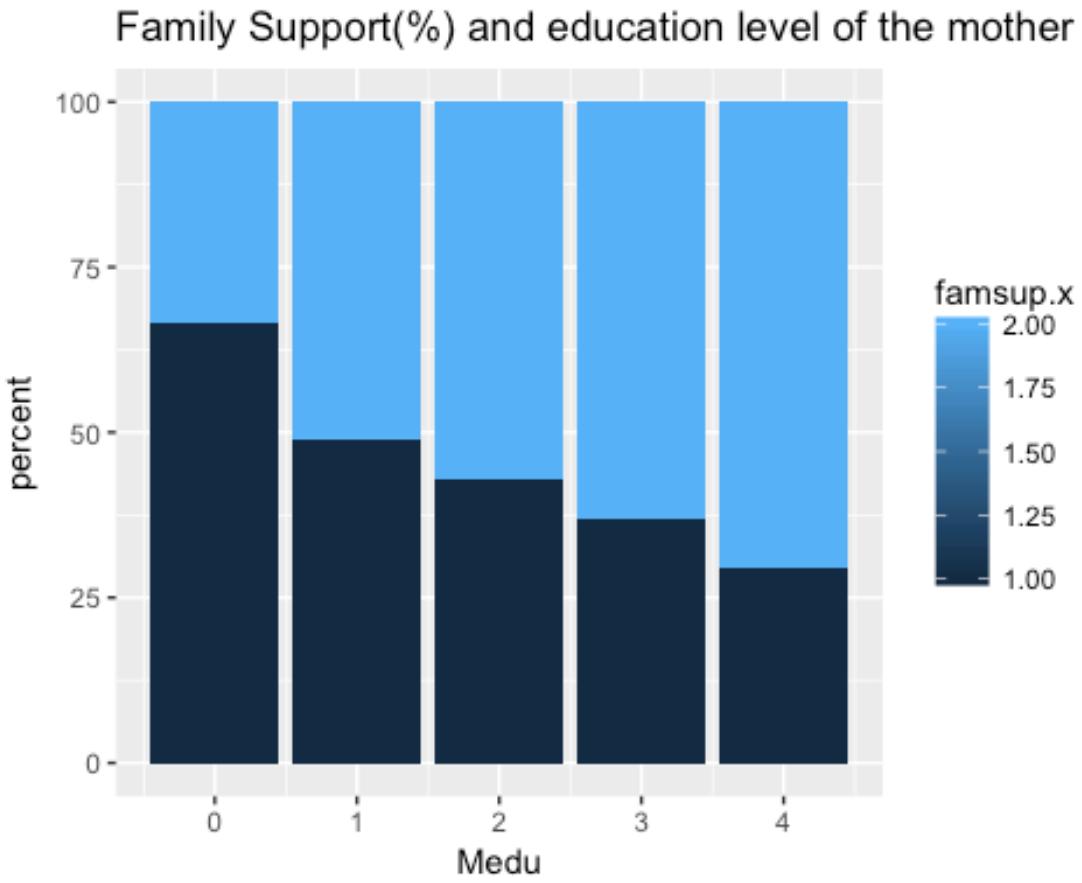

```
ggplot(cclass, aes(x=Fedu, y=G3.y, group=Fedu))+geom_boxplot()+geom_jit
ter()+ xlab("Fathers Eduation")+ylab("Portugese Final Grades")+ggtitle(
"Fathers Education vs Portugese Student Performance")
```



Analysis: Similarly to the math class results, very distinct outliers that will be removed to optimize results in analysis.

Mother education status support students' school in Math class?

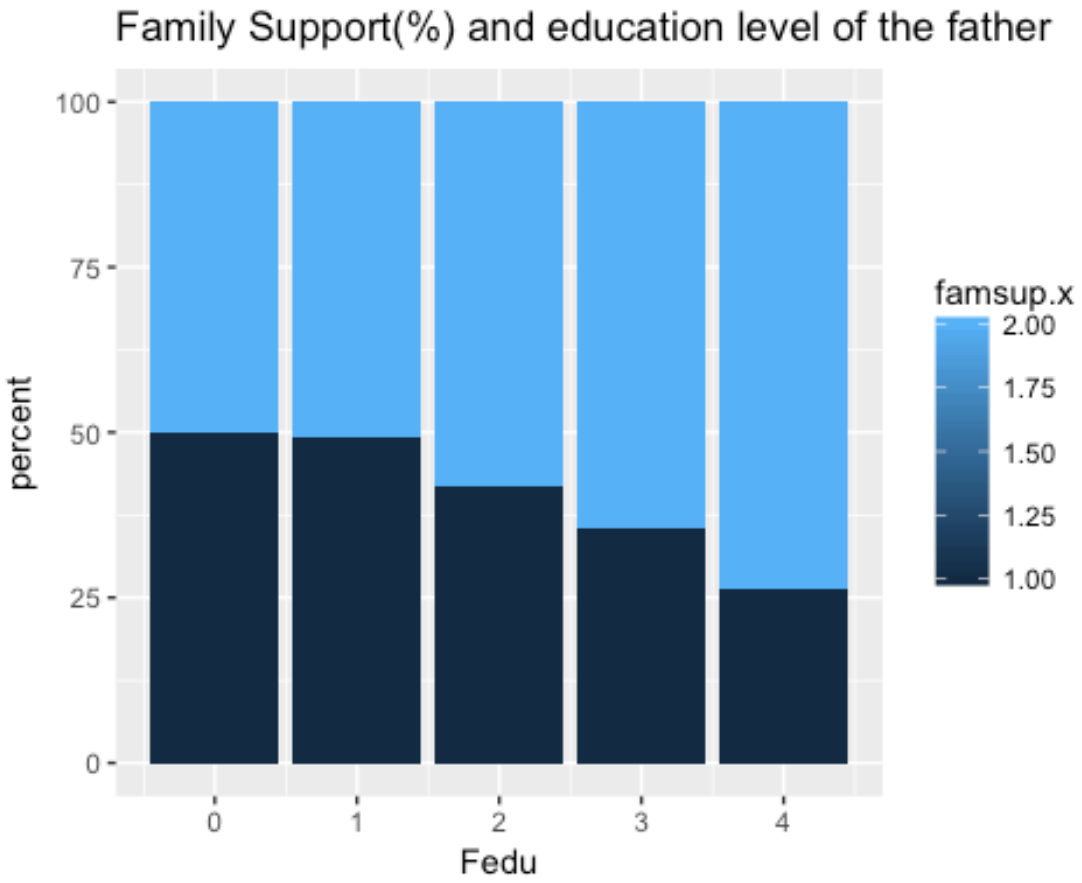
```
cclass %>% group_by(Medu,famsup.x) %>% summarise(n=n()) %>%
  ddply("Medu",transform,percent=n/sum(n)*100) %>%
  ggplot(aes(x=Medu,y=percent,fill=famsup.x))+
  geom_bar(stat="identity")+ggtitle("Family Support(%) and education le
vel of the mother")
```



Analysis: It is evident that the family support increases based on the higher education the mother receives and in return the lower the mothers educational background the lower the family support.

Father education status support students' school in Math class?

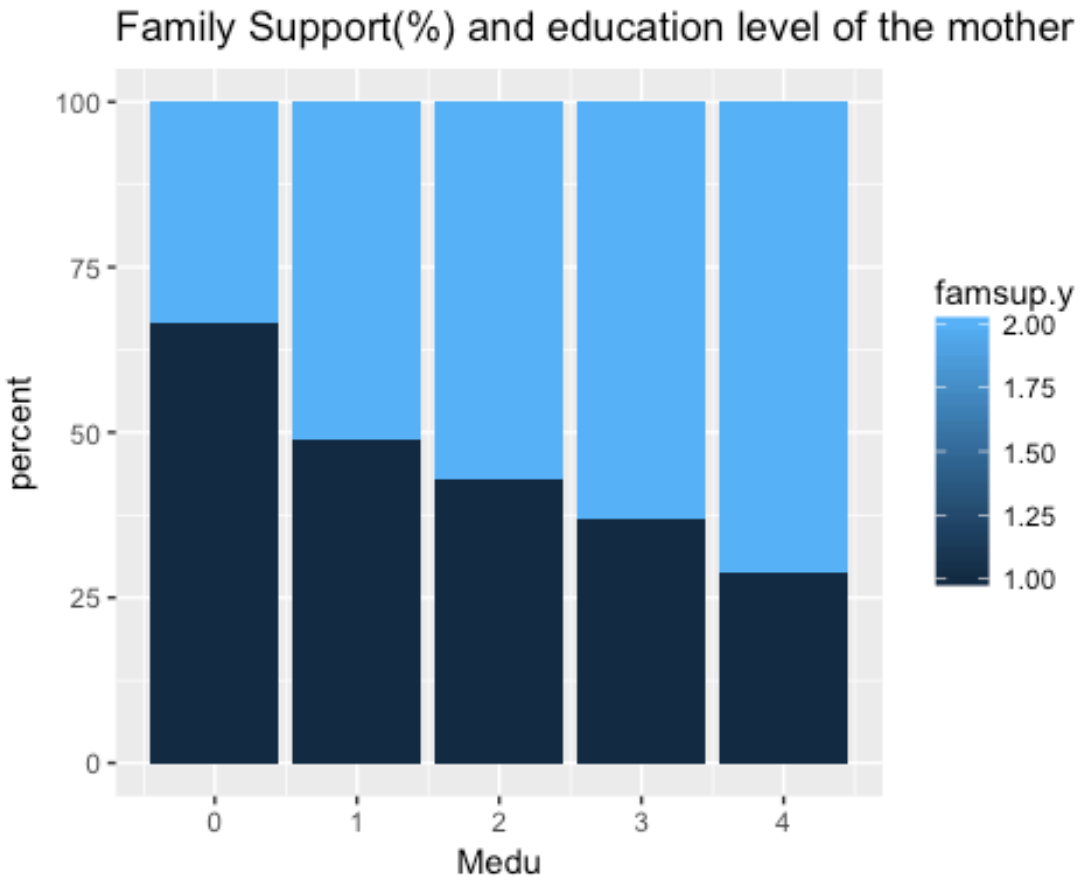
```
cclass %>% group_by(Fedu,famsup.x) %>% summarise(n=n()) %>%
  ddply("Fedu",transform,percent=n/sum(n)*100) %>%
  ggplot(aes(x=Fedu,y=percent,fill=famsup.x))+
  geom_bar(stat="identity")+ggtitle("Family Support(%) and education le
vel of the father")
```



Analysis: It seems that the family support and fathers' education seem to be fairly even with the exception of the father receiver the highest education showing more family support to the students pursuit in school.

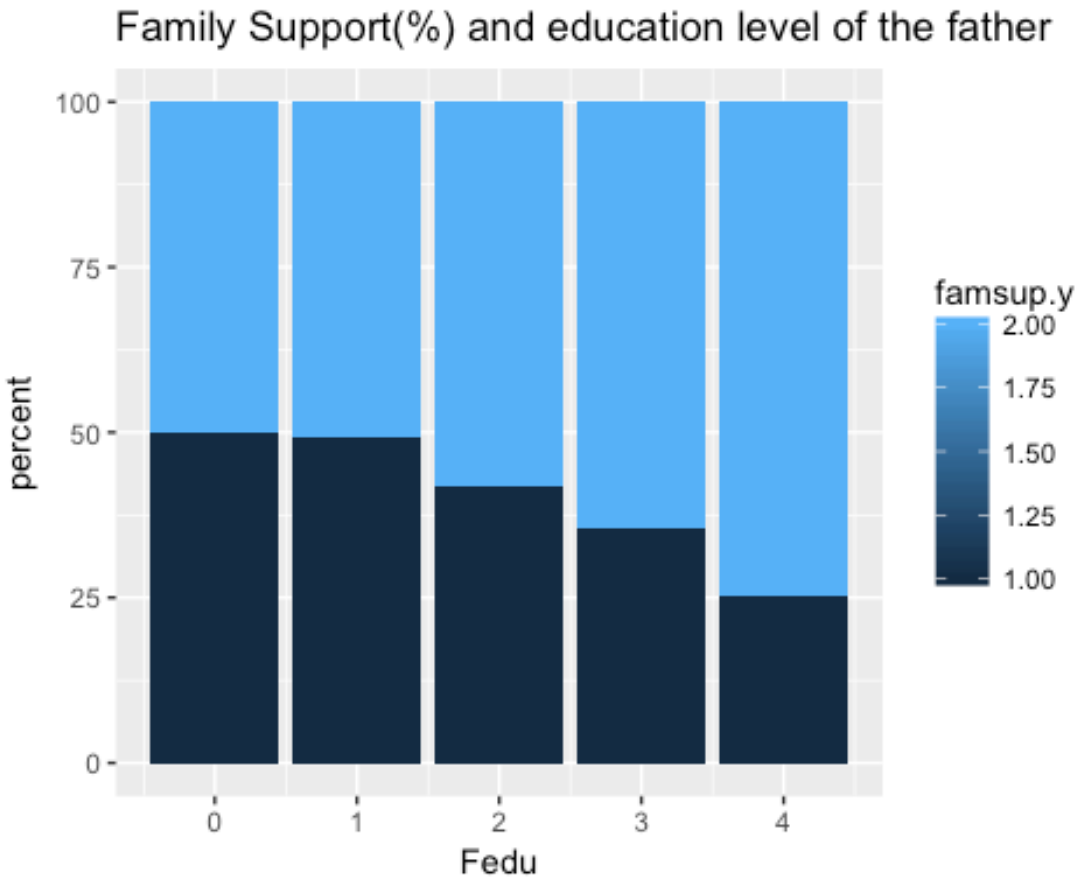
Mother education status support students' school in Portugese class?

```
cclass %>% group_by(Medu,famsup.y) %>% summarise(n=n()) %>%
  ddply("Medu",transform,percent=n/sum(n)*100) %>%
  ggplot(aes(x=Medu,y=percent,fill=famsup.y))+
  geom_bar(stat="identity")+ggtitle("Family Support(%) and education le
vel of the mother")
```



Father education status support students' school in portugese class?

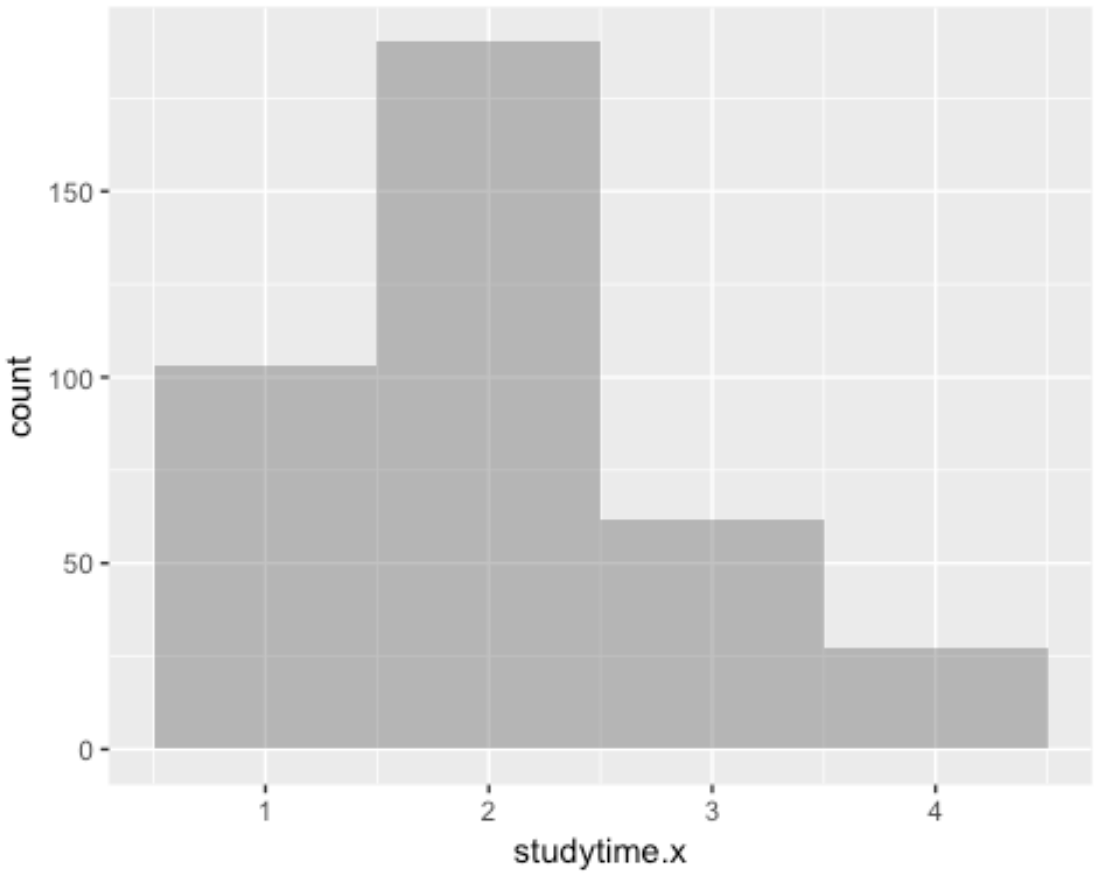
```
cclass %>% group_by(Fedu,famsup.y) %>% summarise(n=n()) %>%
  ddply("Fedu",transform,percent=n/sum(n)*100) %>%
  ggplot(aes(x=Fedu,y=percent,fill=famsup.y))+
  geom_bar(stat="identity")+ggtitle("Family Support(%) and education le
vel of the father")
```



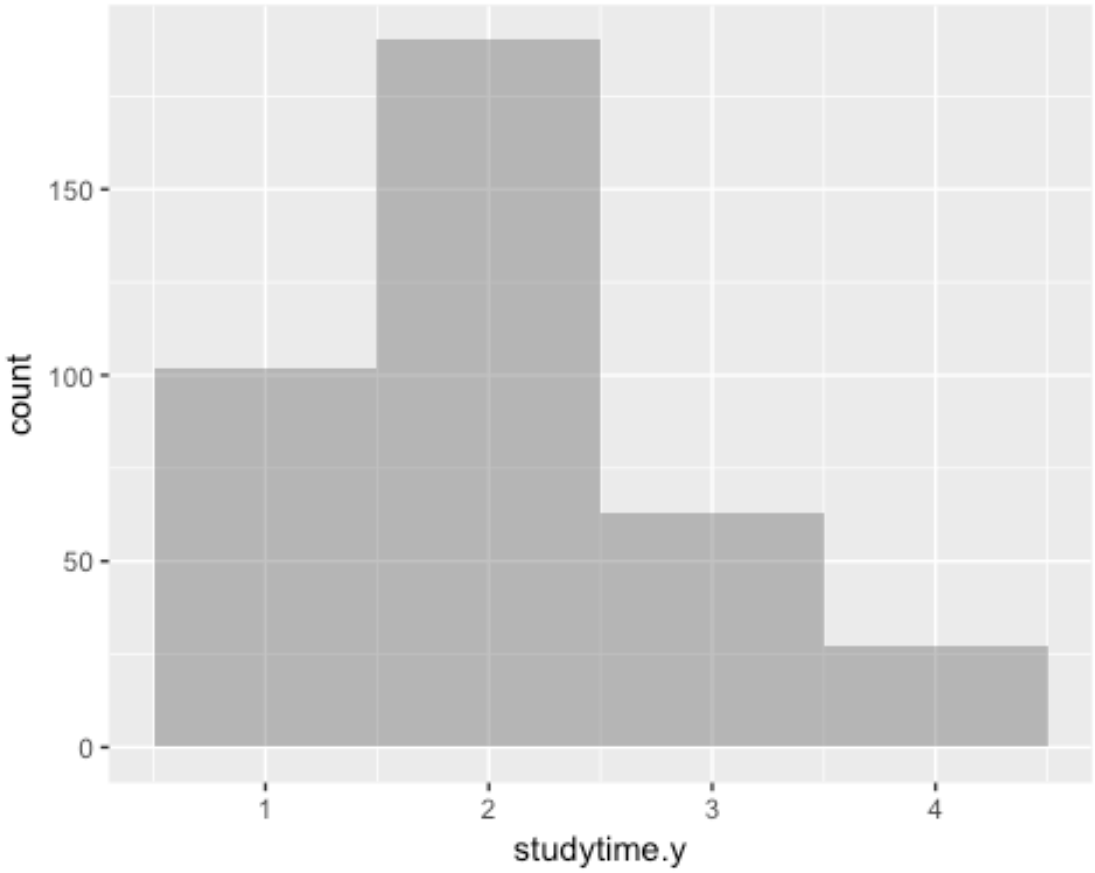
Analysis on Parents education status in relation to students educational goals: No matter what class room they are in math or portuegese the family support is the same in both classes where the mother is more likely to support the student if they have a high educational background whereas the father from low to mid educational level are about 50% to provide educational support.

Student commitment to studying related to amount of failures?

```
ggplot(cclass, aes(x=studytime.x, fill=failures.x)) +
geom_histogram(position="identity", alpha=0.4,binwidth=1.0)
```




```
ggplot(cclass, aes(x=studytime.y, fill=failures.y)) +
geom_histogram(position="identity", alpha=0.4,binwidth=1.0)
```



Analysis:In both classes, the trends are the same where the less you study the higher the failure rate.

Applying simple linear regression

```
mathlm <- lm(G3~., data = mathclass)
summary(mathlm)

##
## Call:
## lm(formula = G3 ~ ., data = mathclass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2675 -0.5346  0.2828  1.0312  4.2503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.66193    2.43410  -0.272 0.785821
## school       0.47055    0.35673   1.319 0.187990
## sex          0.14948    0.22891   0.653 0.514155
## age         -0.20047    0.09549  -2.099 0.036484 *
## address      0.02827    0.26222   0.108 0.914201
## famsize      0.03095    0.22248   0.139 0.889450
```

```
## Pstatus     -0.14999    0.33034  -0.454 0.650066
## Medu         0.12916    0.12933   0.999 0.318628
## Fedu        -0.12667    0.11963  -1.059 0.290371
## Mjob         0.01978    0.09353   0.212 0.832588
## Fjob        -0.11456    0.11787  -0.972 0.331746
## reason       0.07309    0.08347   0.876 0.381771
## guardian     0.08313    0.19626   0.424 0.672130
## traveltime   0.10412    0.15338   0.679 0.497671
## studytime   -0.11183    0.13058  -0.856 0.392331
## failures    -0.20758    0.15172  -1.368 0.172099
## schoolsup    0.49208    0.31227   1.576 0.115939
## famsup       0.15866    0.21972   0.722 0.470695
## paid         0.05743    0.21723   0.264 0.791649
## activities  -0.37216    0.20205  -1.842 0.066314 .
## nursery     -0.22423    0.24797  -0.904 0.366457
## higher       0.09953    0.48441   0.205 0.837325
## internet    -0.20259    0.28054  -0.722 0.470665
## romantic    -0.26683    0.21565  -1.237 0.216759
## famrel       0.35177    0.11153   3.154 0.001745 **
## freetime     0.05260    0.10736   0.490 0.624470
## goout        0.02773    0.10283   0.270 0.787535
## Dalc        -0.17804    0.14675  -1.213 0.225830
## Walc         0.16969    0.11100   1.529 0.127190
## health       0.07055    0.07269   0.971 0.332399
## absences     0.04395    0.01310   3.356 0.000876 ***
## G1           0.18886    0.05939   3.180 0.001600 **
## G2           0.95658    0.05203  18.384  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.888 on 362 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.8302
## F-statistic:  61.2 on 32 and 362 DF,  p-value: < 2.2e-16
```

Analysis: Through a simple linear regression model, we are able to see that the age, family relationship, absences, G1 grades and G2 grades are significant to the final grades G3.

High correlation between each attribute will be removed for math class

```
mathcorrelationMatrix <- cor(mathclass[,1:32])
mathhighlycorrelated <- findCorrelation(mathcorrelationMatrix, cutoff=0
.5)
print(mathhighlycorrelated)

## [1]  7 32 28
```

Analysis: The correlation matrix has demonstrated that Mothers education, Weekend Alcohol consumption, G2 are highly correlated with other attributes in the

dataset. Therefore, I've eliminated these attributes from further analysis. Cut off is set to another over 75% correlated.

Overall results from Explanatory Analysis: I've found through explanatory analysis, both math and portuegese classes show similar or the same trend. Therefore, I've decided to move forward with only analyzing the students enrolled into the math class to narrow my focus on one dataset.

What attributes are important in this dataset by a rank

```
mathcontrol <- trainControl(method="repeatedcv", number=10, repeats=3)
mathmodelknn <- train(G3~., data=mathclass[,c(1:6, 8:27, 29:31, 33)], m
ethod="kknn", preProcess="scale", trControl=mathcontrol)
```

```
## Loading required package: kknn
```

```
##
## Attaching package: 'kknn'
```

```
## The following object is masked from 'package:caret':
##
##     contr.dummy
```

```
gbmIMPknn <- varImp(mathmodelknn, scale = FALSE)
print(gbmIMPknn)
```

Analysis: One of the methods we used to to train and test is the KNN model, we are able to produce the top 20 attributes ranked by importance. However, what does other methods tell us? Same ranks or different?

```
mathcontrol <- trainControl(method="repeatedcv", number=10, repeats=3)
mathmodellm <- train(G3~., data=mathclass[,c(1:6, 8:27, 29:31, 33)], me
thod="lm", preProcess="scale", trControl=mathcontrol)
gbmIMPlm <- varImp(mathmodellm, scale = FALSE)
print(gbmIMPlm)
```

```
## lm variable importance
##
##   only 20 most important variables shown (out of 29)
##
##             Overall
## G1         23.2310
## absences    2.7794
## age         2.7277
## romantic    2.7126
## schoolsup   2.1175
## paid        1.8790
## failures    1.6206
## school      1.5873
## Fedu        1.3574
## famrel      1.3151
## reason      1.2854
```

```
## Pstatus     1.2106
## sex         1.1933
## traveltime  1.1162
## activities  1.1001
## address     0.9618
## studytime   0.8323
## nursery     0.7474
## famsize     0.5386
## internet    0.5359
```

Analysis: Using kknn and lm methods show that the ranking of the attributes are the same.

Feature Selection for Math Class

```
mathcontrol2 <- rfeControl(functions=rfFuncs, method="cv", number=10)
results <- rfe(mathclass[,c(1:6, 8:27, 29:31)], mathclass[,33], sizes=c
(1:20), rfeControl = mathcontrol2)
r1 <-predictors(results)
print(r1)

##  [1] "G1"         "absences"   "failures"   "schoolsup"  "age"
##  [6] "higher"     "romantic"   "guardian"   "goout"      "Mjob"
## [11] "activities" "paid"       "traveltime" "school"     "Pstatus"
## [16] "sex"        "address"    "famsize"
```

Analysis: Based on this Feature Selection method, we see that failures, absences, higher, schoolsup, goout, age, Mjob etc. are the attributes that are used in many combination models that perform with accuracy. Therefore these attributes will be used to create a Bayesian Network.

Creation of a Bayesian Network for the Math Class

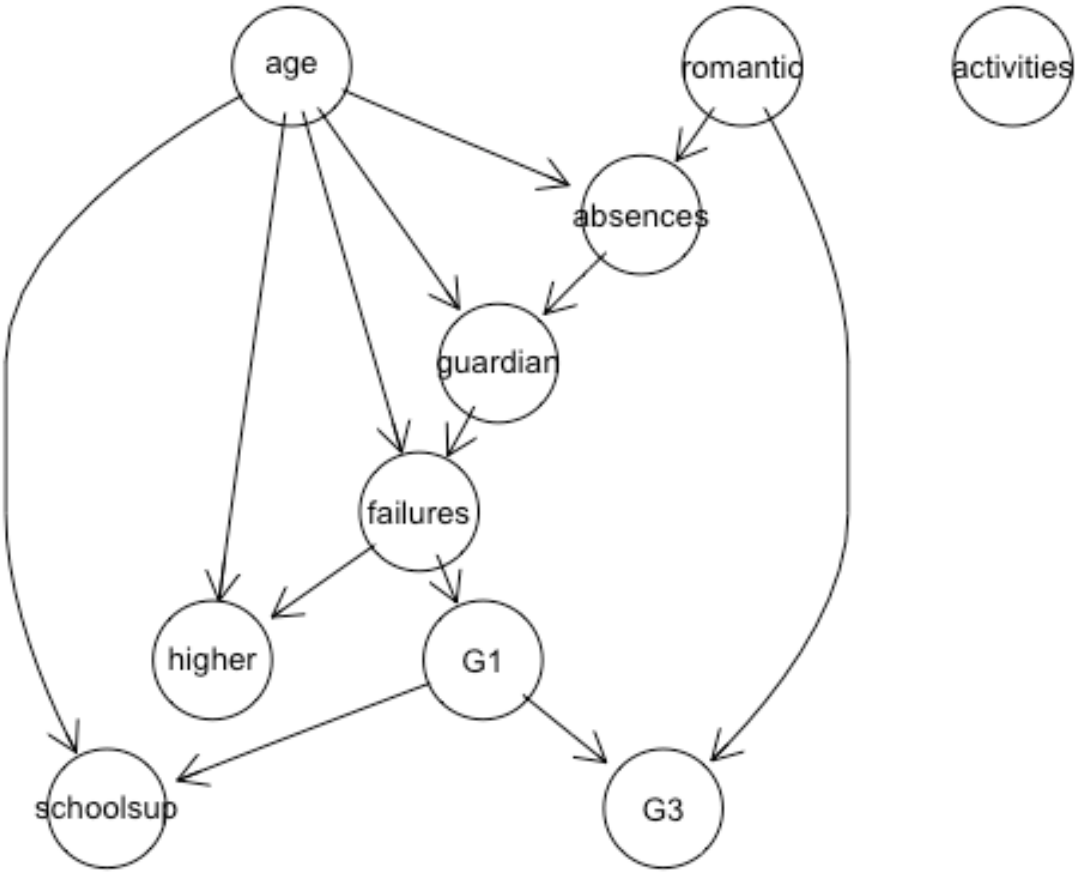
based on the important attributes found from the Feature Selection analysis

```
math_dag <- gs(mathclass[, c(r1,"G3")])

## Warning in FUN(newX[, i], ...): vstructure G1 -> G3 <- romantic is
## not applicable, because one or both arcs are oriented in the opposit
e
## direction.

## Warning in FUN(newX[, i], ...): vstructure age -> absences <- romant
ic
## is not applicable, because one or both arcs are oriented in the oppo
site
## direction.

graphviz.plot(math_dag)
```



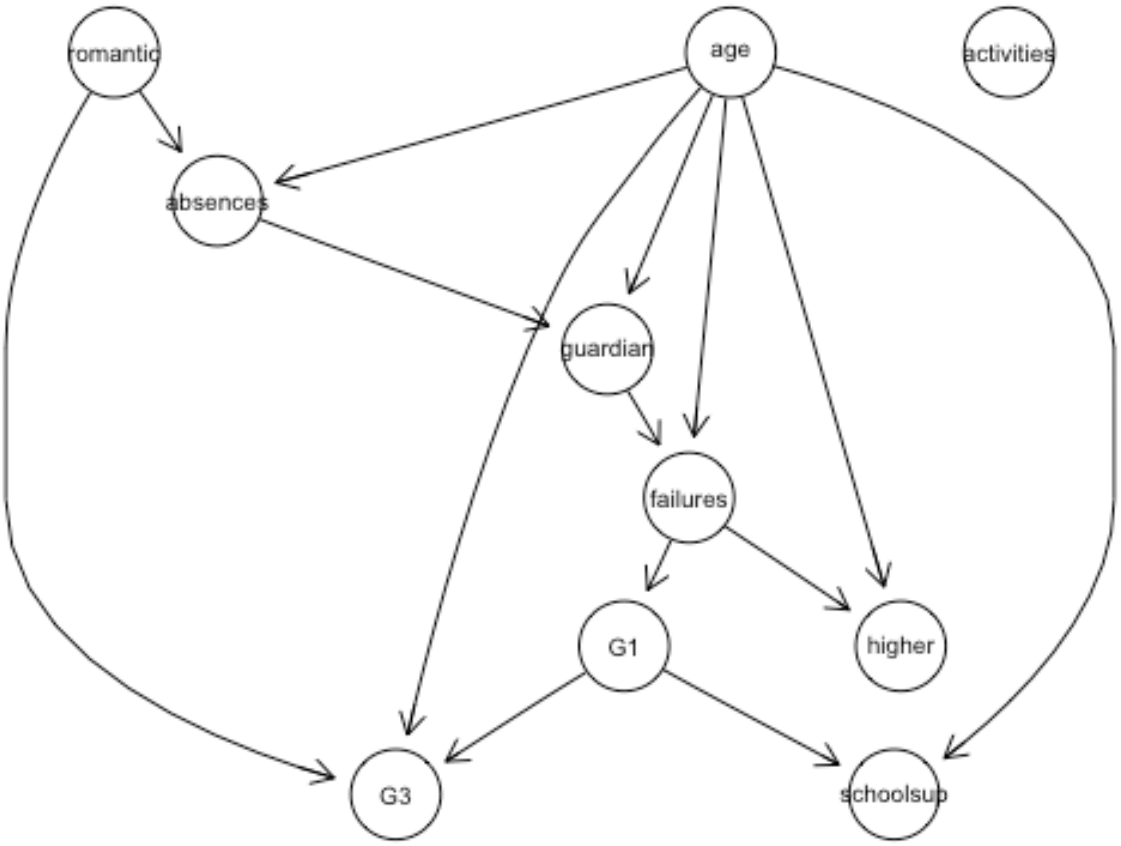

```
##Testing different Bayesian network methods
math_dag2 <- iamb(mathclass[, c(r1,"G3")])

## Warning in FUN(newX[, i], ...): vstructure age -> absences <- romant
ic
## is not applicable, because one or both arcs are oriented in the oppo
site
## direction.

## Warning in FUN(newX[, i], ...): vstructure age -> G3 <- romantic is
## not applicable, because one or both arcs are oriented in the opposit
e
## direction.

## Warning in FUN(newX[, i], ...): vstructure schoolsup -> G1 <- G3 is
## not applicable, because one or both arcs are oriented in the opposit
e
## direction.

graphviz.plot(math_dag2)
```

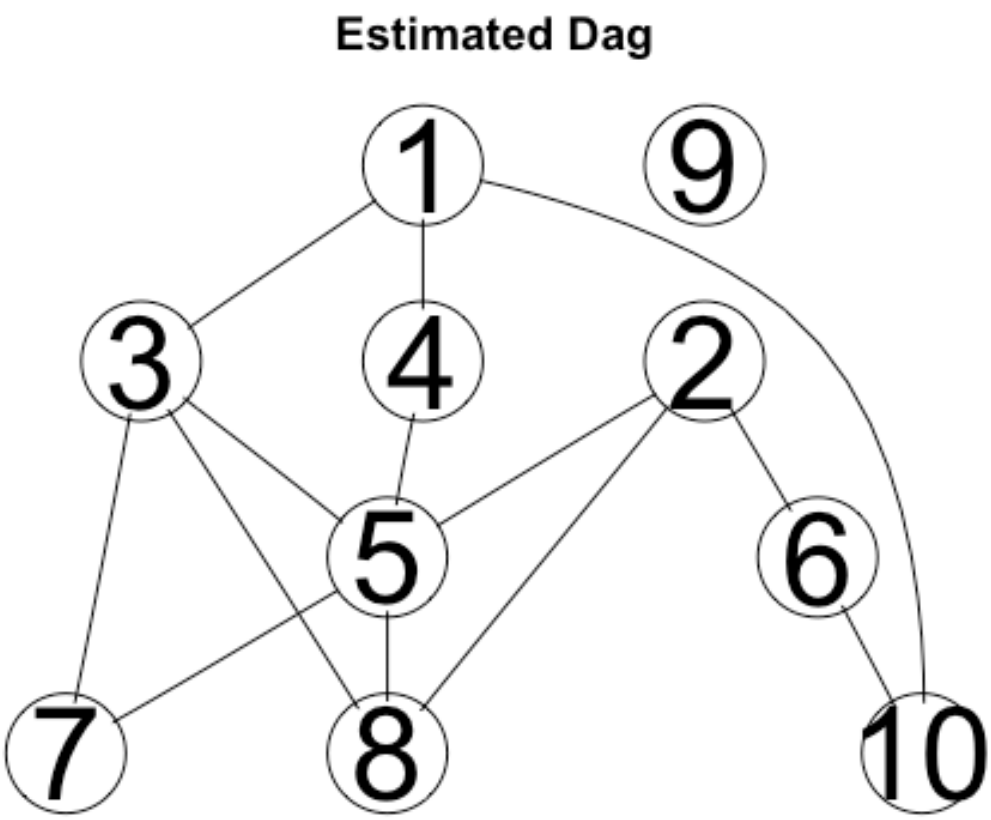


Analysis: Based on the two results of the Bayesian Network, we can conclude that the attribute activities has no significant relationship with other attributes that we focused on. Given our prior analysis that we conducted in the Attribute ranking algorithm, the attribute activities was not listed in the top 20.

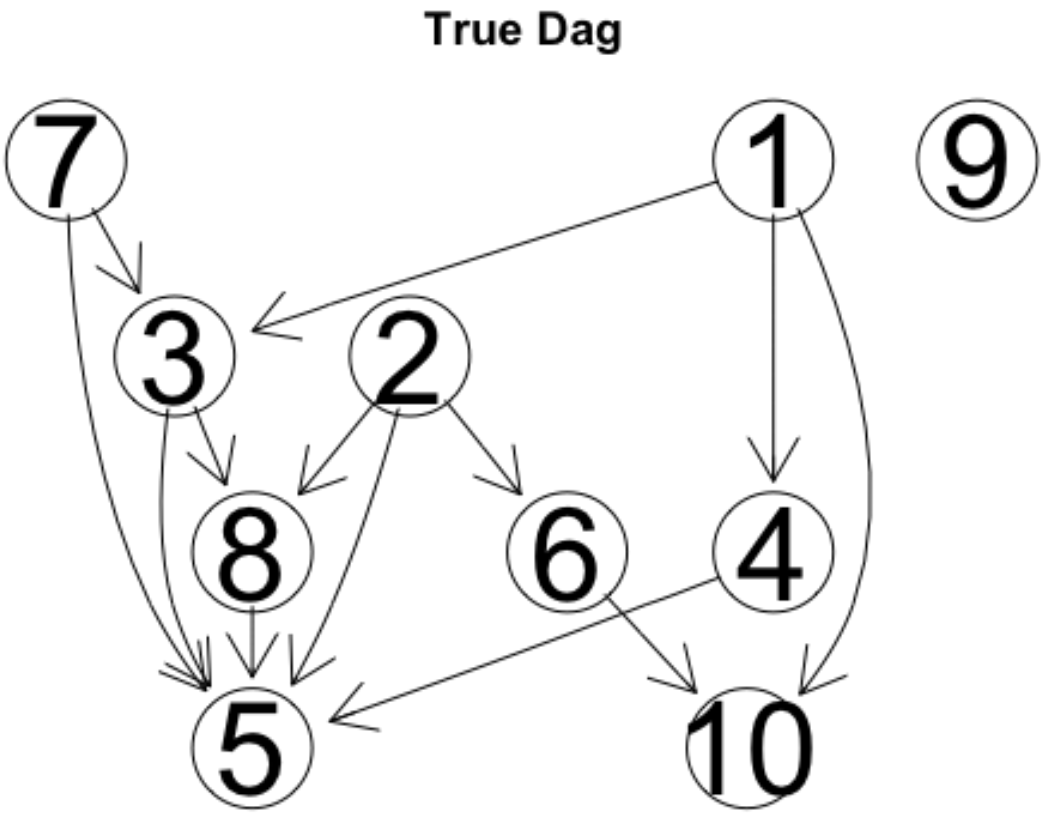
Based on the Bayesian Network, we can see that the students' final grades, G3, are affected by G1, and romantic. However, we see that using iamb method gives us G1, romantic and age.

Casuality of Student final grades

```
suffStat <- list(C = cor(mathclass[, c(r1,"G3")]), n = nrow(mathclass[,
c(r1,"G3")]))
skelpc.fit <- skeleton(suffStat, indepTest = gaussCItest, p = ncol(math
class[, c(r1,"G3")]), alpha = 0.05)
pc.fit <- pc(suffStat, indepTest = gaussCItest, p = ncol(mathclass[, c(
r1,"G3")]), alpha = 0.05)
plot(skelpc.fit, main = "Estimated Dag")
```




```
plot(pc.fit, main = "True Dag")
```



Analysis: We are able to see that the two causes of the students final grading (1) is romantic involvement (6) and their first grading evaluation in the course.