Project2

davidabraham

2/15/2017

Importing in R

```
student.mat<-read.csv('~/Desktop/Shipwreck/student-mat.csv',sep = ",")
head(student.mat)
```

```
##   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
```

A small function to check for missing values within a vector. ## 1. Investigation of the performance in G3

```
na.test <- function (x) {
  output <- any(is.na(x)== TRUE)
return(output)
}
sprintf('Applying the function to every column');
```

```
## [1] "Applying the function to every column"
```

```
apply(student.mat, 2, 'na.test')
```

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

2. What is the impact of age and the sex on performance(G3) ?

First of all, I checked whether there is a difference in performance between boys and girls.

```
gender.dif <- t.test(student.mat$G3~student.mat$sex,var.equal = TRUE)
library(apa)
apa(gender.dif)
```

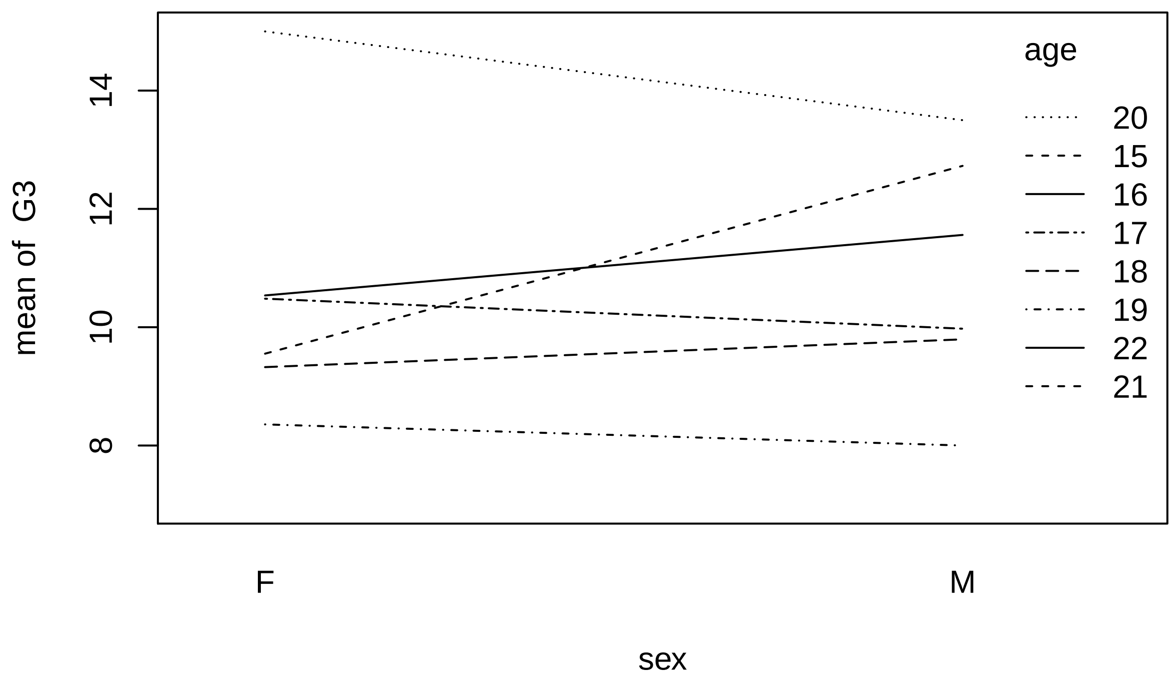
```
## [1] "*t*(393) = -2.06, *p* = .040, *d* = -0.21"
```

The mean values between the genders is not equal.Now, I go a step further and take also the age into consideration.

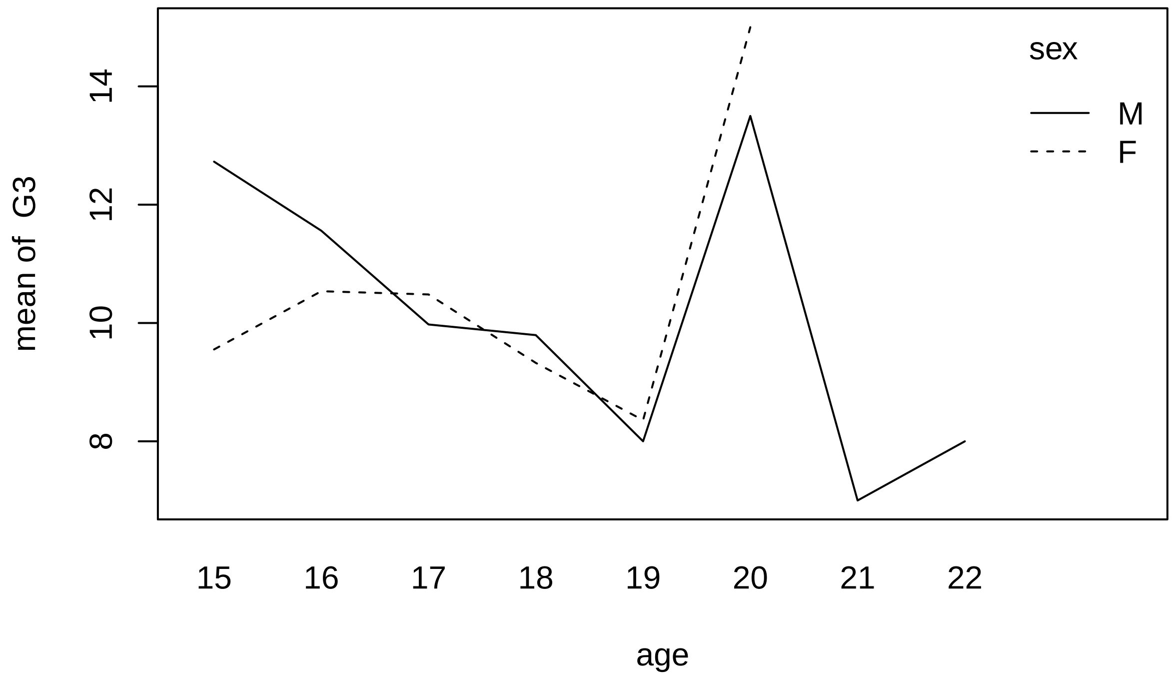
```
summary(with(student.mat, aov(G3 ~ sex + age + sex*age)))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## sex           1     89   88.51   4.385 0.03690 *
## age           1    208  208.24  10.317 0.00143 **
## sex:age       1     81   80.74   4.000 0.04619 *
## Residuals   391   7892   20.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
with(student.mat, interaction.plot(sex, age, G3))
```




```
with(student.mat, interaction.plot(age, sex, G3))
```




```r
table(student.mat$age)
```

```
##
##  15  16  17  18  19  20  21  22
##  82 104  98  82  24   3   1   1
```

```r
student.mat.2 <- subset(student.mat, age < 20)
table(student.mat.2$age)
```

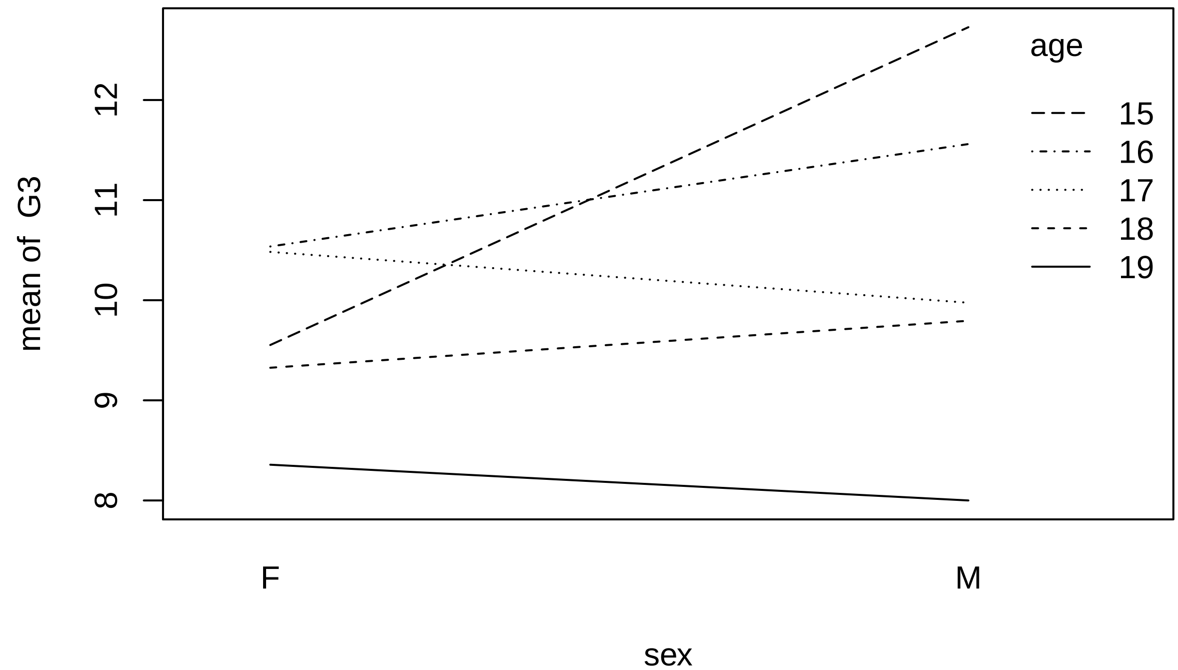
```
##
##  15  16  17  18  19
##  82 104  98  82  24
```

```r
summary(with(student.mat.2, aov(G3 ~ sex + age + sex*age)))
```

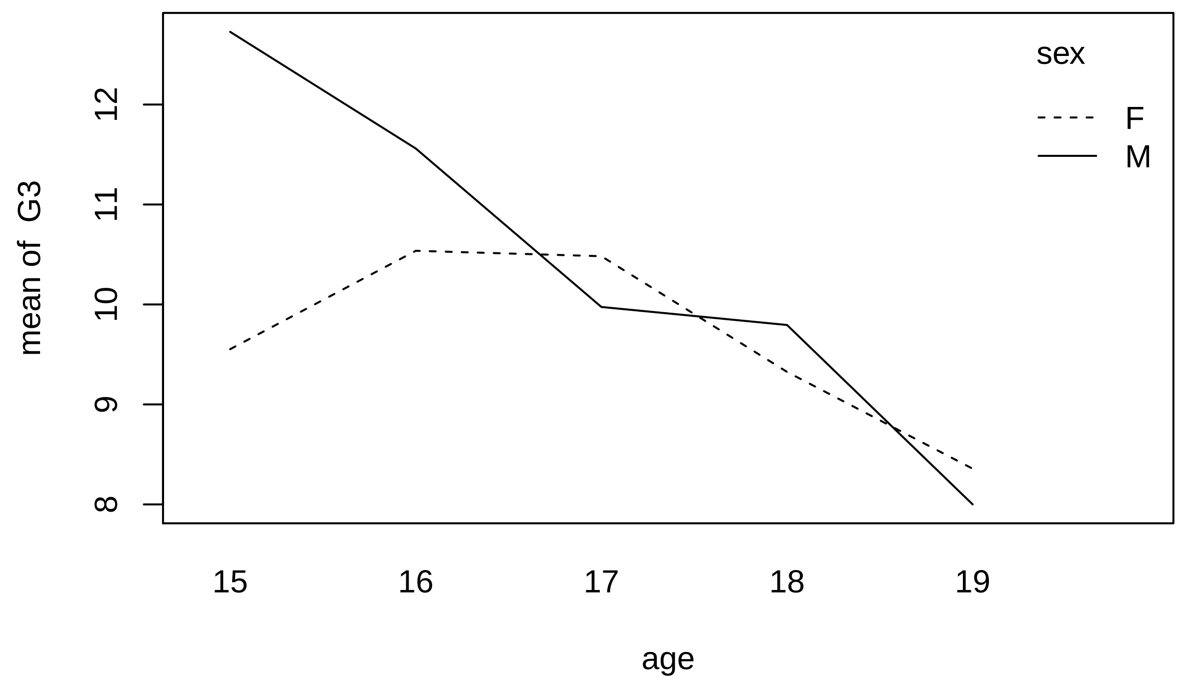
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sex           1     94   93.56   4.665 0.031406 *
## age           1    240  240.21  11.975 0.000599 ***
## sex:age       1     95   95.50   4.761 0.029714 *
## Residuals   386   7743   20.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

While eliminating the outliers, the probability that the treatment means differ became less likely for every factor. Now, I redo the interaction plots, too.

```r
with(student.mat.2, interaction.plot(sex, age, G3))
```




```r
with(student.mat.2, interaction.plot(age, sex, G3))
```



Especially the last plot looks better. But as it seems, the interaction is more difficult to understand. I was wondering why the performance of boys gradually (linearly) decreases when boys grow older and why the performance of girls stays more constant with reference to the age. Looking at the mean values this thought is reinforced.

```r
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
```

```
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
## Warning: package 'ggplot2' was built under R version 3.3.2
## Conflicts with tidy packages ----------------------------------------------
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
student.mat.2 %>%
  group_by(age, sex) %>%
  summarise(
    a.mean = mean(G3)
  )
```

```
## Source: local data frame [10 x 3]
## Groups: age [?]
##
##      age    sex    a.mean
##    <int> <fctr>     <dbl>
## 1     15      F  9.552632
## 2     15      M 12.727273
## 3     16      F 10.537037
## 4     16      M 11.560000
## 5     17      F 10.482759
## 6     17      M  9.975000
## 7     18      F  9.325581
## 8     18      M  9.794872
## 9     19      F  8.357143
## 10    19      M  8.000000
```

Consequently, I checked the correlation between age and the performance for two subsets holding boys and girls separately.

```
cor.test1 <- with(subset(student.mat.2, sex == "M"),
     cor.test(G3, age))
cor.test1
```

```
##
##  Pearson's product-moment correlation
##
## data:  G3 and age
## t = -4.13, df = 181, p-value = 5.535e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4206139 -0.1550039
## sample estimates:
##        cor
## -0.2934623
```

```
apa(cor.test1)
```

```
## [1] "*r*(181) = -.29, *p* < .001"
```

```
cor.test2 <- with(subset(student.mat.2, sex == "F"),
     (cor.test(G3, age)))
cor.test2
```

```
##
##  Pearson's product-moment correlation
##
## data:  G3 and age
## t = -0.93582, df = 205, p-value = 0.3505
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1998139  0.0717874
## sample estimates:
##         cor
## -0.06522112
```

```
apa(cor.test2)
```

```
## [1] "*r*(205) = -.07, *p* = .350"
```

As expected, there is a correlation between age and performance for boys and none for girls. Therefore, I will focus on the boys. I calculate a linear regression analysis between age and performance in order to get further information.

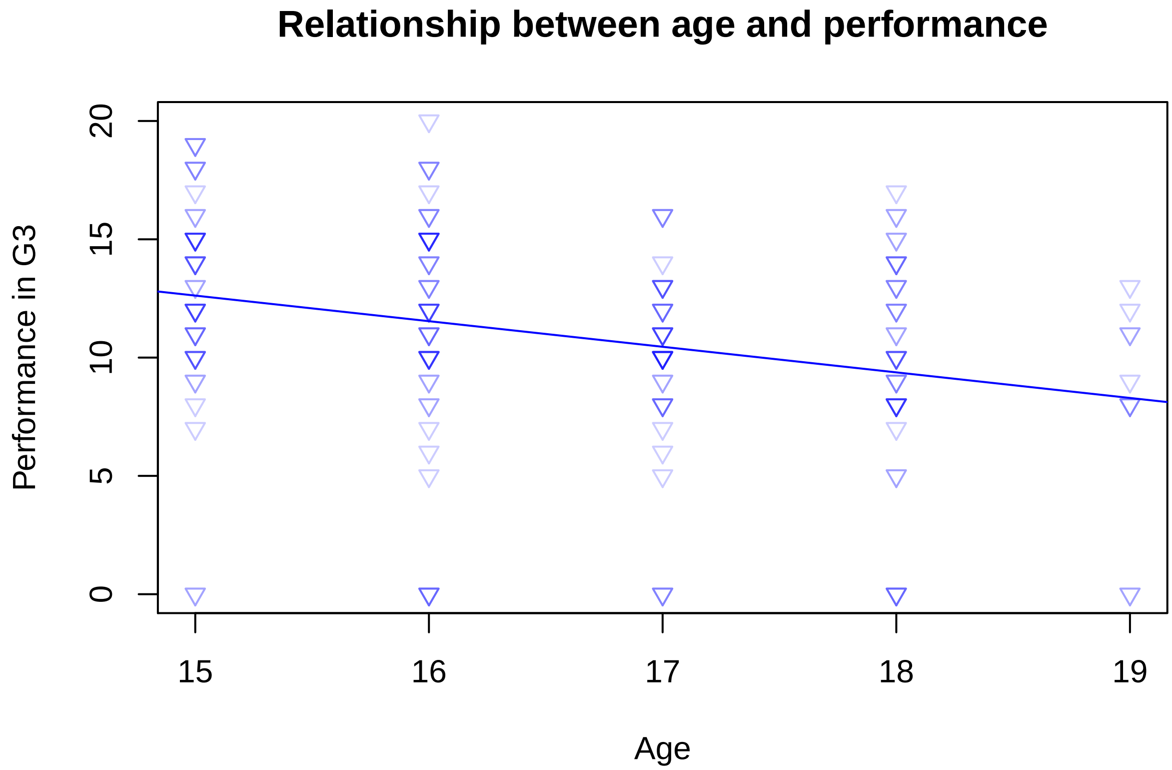
```
with(subset(student.mat.2, sex == "M"),
     summary(lm(G3~ age)))
```

```
##
## Call:
## lm(formula = G3 ~ age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6200  -1.5383   0.4617   2.6252   8.4617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.8464     4.3514   6.629 3.75e-10 ***
## age          -1.0818     0.2619  -4.130 5.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.311 on 181 degrees of freedom
## Multiple R-squared:  0.08612,    Adjusted R-squared:  0.08107
## F-statistic: 17.06 on 1 and 181 DF,  p-value: 5.535e-05
```

The probability that the group means are equal is, of course, the same as in the correlation analysis. However, with the linear regression we can predict values and show a tendency with a regression line. Last but not least, I show the results in a scatter plot:

```
plot(1,
     xlim = c(15, 19),
     ylim = c(0, 20),
     type = "n",
     main = "Relationship between age and performance",
     xlab = "Age",
     ylab = "Performance in G3"
     )

#Now, I fill in the points.
with(subset(student.mat.2, sex == "M"),
```

```r
      points(age,
             G3,
             pch = 25,
             col = alpha("blue", 0.2)
             ))

#Finally, I draw the regression line.
with(subset(student.mat.2, sex == "M"),
     abline(lm(G3 ~ age), col = "blue"))
```



3.What is the relationship between failures and performance with reference to the age?

While eliminating persons older than 20, I recognized that these persons have bad grades. So at first, I checked the correlation between failures and age.

```
cor.test3 <- with(student.mat, cor.test(age, failures))
apa(cor.test3)
```

```
## [1] "*r*(393) = .24, *p* < .001"
```

```
cor.test2 <- with(subset(student.mat.2, sex == "F"),
     (cor.test(G3, age)))
cor.test2
```

```
##
##  Pearson's product-moment correlation
##
## data:  G3 and age
## t = -0.93582, df = 205, p-value = 0.3505
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1998139  0.0717874
## sample estimates:
##         cor
## -0.06522112
```

The results reveal a strong connection between failures and age. This maybe explains why there are people of 22 in a school class. Furthermore, I explored the relationship between age, failures and the performance

```
with(student.mat, summary(aov(G3 ~ age + failures + age*failures)))
```

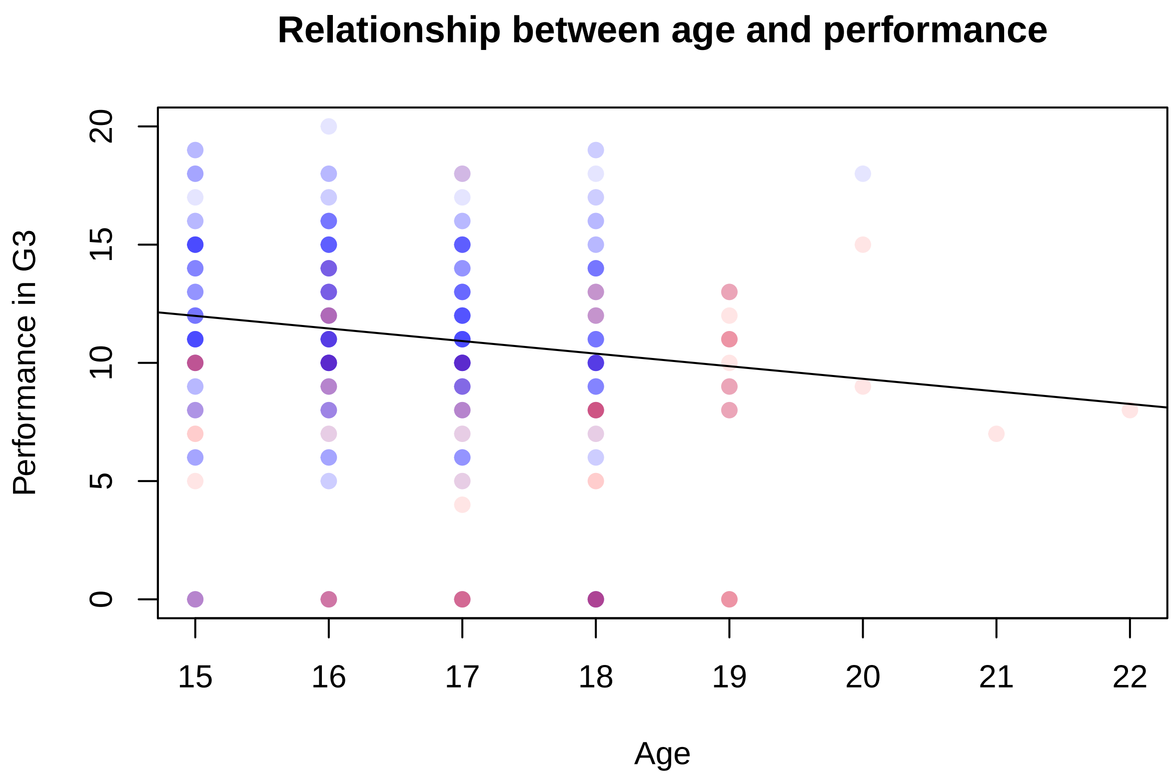
```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## age             1    216   215.9  11.976 0.000598 ***
## failures        1    906   906.2  50.261 6.33e-12 ***
## age:failures    1     98    98.4   5.458 0.019986 *
## Residuals     391   7049    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that all factors are significant. The older a person was and the more failures a person experienced, the more will the performance decrease. Finally, I plot the results:

```
plot(1,
     xlim = c(15, 22),
     ylim = c(0, 20),
     type = "n",
     main = "Relationship between age and performance",
     xlab = "Age",
     ylab = "Performance in G3"
     )

#People with no failures.
with(subset(student.mat, failures == 0),
     points(age,
            G3,
            pch = 21,
```

```
            col = alpha("blue", 0.1),
            bg =alpha("blue", 0.1)
            ))

#People with more than one failure.
with(subset(student.mat, failures > 0),
     points(age,
            G3,
            pch = 21,
            col = alpha("red", 0.1),
            bg =alpha("red", 0.1)
            ))

with(student.mat, abline(lm(G3 ~ age + failures + age*failures)))
```

```
## Warning in abline(lm(G3 ~ age + failures + age * failures)): only using the
## first two of 4 regression coefficients
```



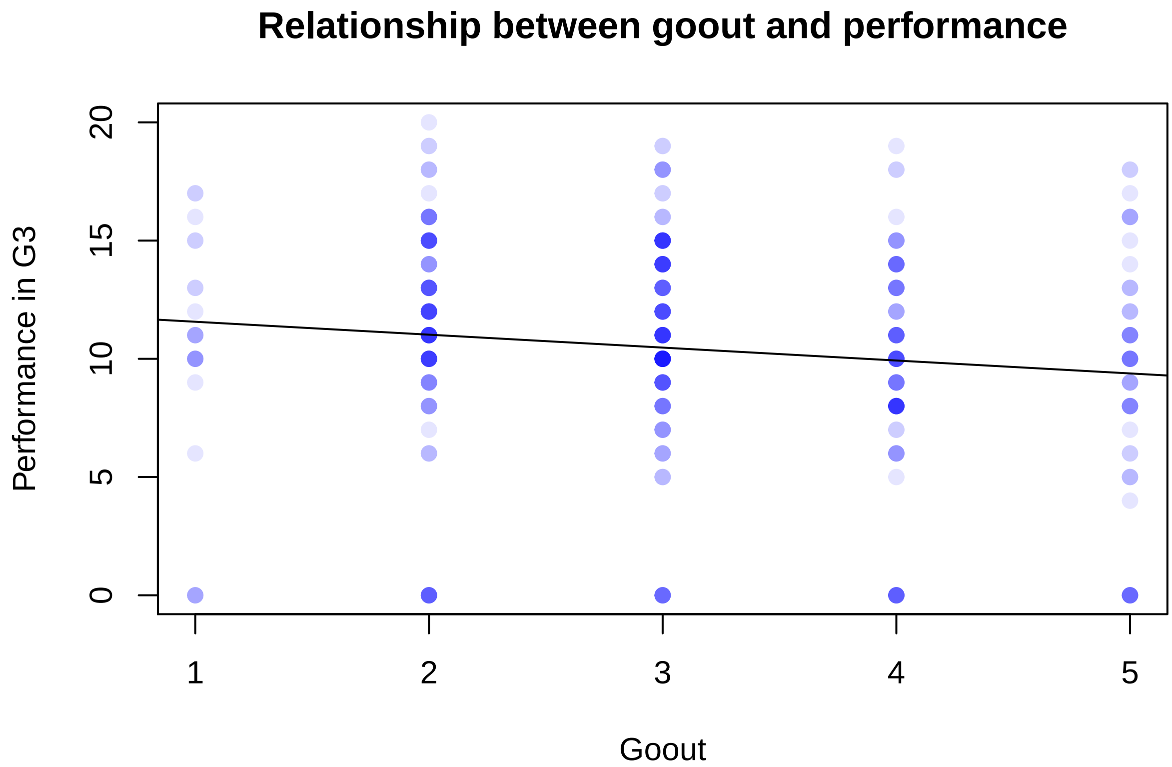
4. Relationship between goout and performance

```r
lm1 <- with(student.mat, summary(lm(G3 ~ goout )))
lm1
```

```
##
## Call:
## lm(formula = G3 ~ goout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5676  -1.9282   0.4324   3.0718   9.0718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.1141     0.6793  17.833  < 2e-16 ***
## goout        -0.5465     0.2057  -2.656  0.00823 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.547 on 393 degrees of freedom
## Multiple R-squared:  0.01763,    Adjusted R-squared:  0.01513
## F-statistic: 7.054 on 1 and 393 DF,  p-value: 0.008229
```

Going out is significantly related to the performance in the math course. I want to visualize this with a scatter plot:

```r
plot(1,
     xlim = c(1, 5),
     ylim = c(0, 20),
     type = "n",
     main = "Relationship between goout and performance",
     xlab = "Goout",
     ylab = "Performance in G3"
     )


with(student.mat,
     points(goout,
            G3,
            pch = 21,
            col = alpha("blue", 0.1),
            bg =alpha("blue", 0.1)
            ))

with(student.mat, abline(lm(G3 ~ goout)))
```



After checking the plot I realized that the mean of the performance is low when the child is rarely going out. This is why I assumed another coherence. At first, I checked for the means:

```
aggregate(
  formula = G3 ~ goout,
  data= student.mat,
  FUN = mean)
```

```
##   goout        G3
## 1     1  9.869565
## 2     2 11.194175
## 3     3 10.961538
## 4     4  9.651163
## 5     5  9.037736
```

The means reveal what I assumed. The first mean is lower than the second or third one. Finally, I expected the relationship between performance and going out to be quadratic. I checked this with a regression analysis.

Conclusion

In comparison to that the result that going out is negatively correlated with your performance in a Math Class is totally intuitive. Additionally, the results revealed that older children which failed once or several times have lower performance rates.

While boys show lower performances when they grow older, girls remain relatively constant.