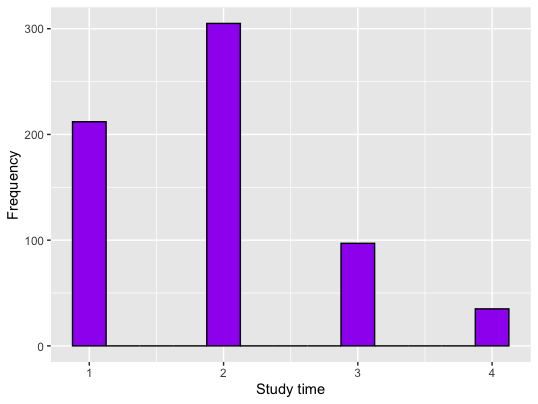
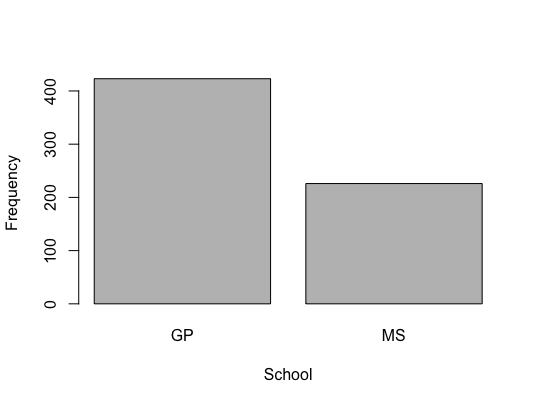
First, we perform some descriptive analysis that is necessary to provide the features of relevant variables and associations in the data. We import the necessary libraries and load the csv file with the dataset in order to perform the analysis. Furthermore, we perform an exploratory analysis through different graphs to describe the data explain how the different variables are correlated with each other.

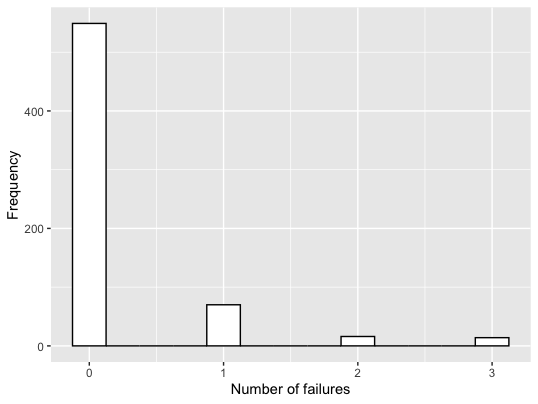
The first histogram (Fig.1) includes the variable “study time” on the x axis and “frequency” on the y axis, explaining how many students study the same number of hours. We can see that the highest frequency of individuals study two hours, while the lowest frequency study four hours.

Fig.1

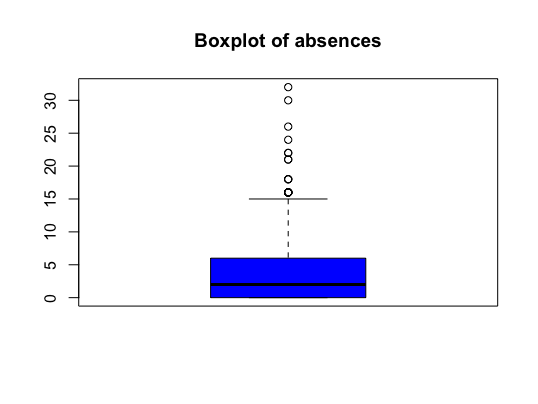
The second histogram (Fig.2) takes into consideration the variables “school” on the x axis and “frequency” on the y axis, highlighting how many students go to either the GP school or MS school. The frequency detected in the GP school, corresponding to 423 students, is higher compared to the frequency in the MS school, corresponding to 226 students.

Fig.2

The third histogram (Fig.3) considers the relationship between the variables “number of failures” on the x axis and “frequency” on the y axis, showing the frequency of student’s failures. The histogram reveals that there is overall a low number of failures between the students.

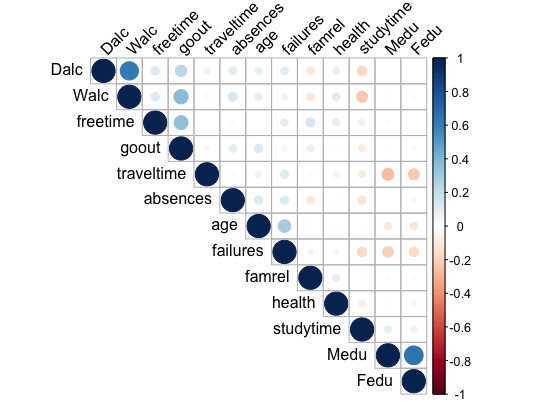
Fig.3

The boxplot of the variable “absences” shows that there are multiple outliers in the frequency of absences.

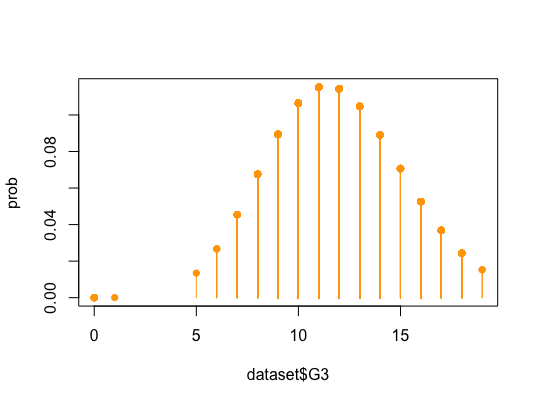
Fig.3

Then, we explored multiple combinations through the contingency tables. The first takes “school” and “study time” as variables, showing the probability that students in the two schools will study a defined number of hours. We tried to see whether there are more relevant differences between probabilities, and we found for instance that the probability that an MS student studies one hour is much higher compared to the GP students. Overall, as we already understood through the previous graph, the highest probability is that a student at both school takes two hours of study time.

Furthermore, we checked the correlation between the variables where the correlation plot (Fig.4) summarized the correlations that can be either positive(blue) or negative (red).

Fig.4

---

Fig.5

To do

#independence test

chisq.test(table(dataset$higher,dataset$sex))

#correlation between numeric variables

#Here we plot the correlation matrix of all numeric vaiables using the corrplot package

num\_variables = select\_if(X, is.numeric)

cor(num\_variables, Y, method = "spearman")

cor\_matr <- cor(num\_variables, method = "spearman")

cor\_matr

library(corrplot)

corrplot(cor\_matr, type = "upper", order = "hclust",

tl.col = "black", tl.srt = 45)

# Does it look like a Poisson?

# Here we create a plot to check whether the outcome variable is Poisson distributed

prob <- dpois(dataset$G3, mean(dataset$G3))

prob

plot(dataset$G3, prob, pch = 16, col = "orange", type = "h")

points(dataset$G3, prob, pch = 16, col = "orange", type = "p")

# From the graph we see it does look like a Poisson distribution

#####POISSON REGRESSION MODEL

# Here we regress the outcome variable G3 on all other variables except G1 and G2 using Poisson regression

mod = glm(Y ~ ., data=X, family=poisson)

summary(mod)

#H Here we perform stepwise variable selection to find the best model

# We finish with an AIC of 3313.1

best\_mod = step(glm(Y ~ ., data=X, family=poisson))

summary(best\_mod)

exp((best\_mod$coefficients))

#deviance goodness of fit

1-pchisq(best\_mod$deviance, best\_mod$df.residual) #good fit

#residuals analysis

#Here we plot the residuals in a scatterplot and then a histogram showing how the residuals are distributed, first usig Pearson, then Deviance residuals

resid <- resid(best\_mod, type = "pearson")

fitted <- fitted(best\_mod)

plot(fitted, resid, main = "Pearson residuals")

hist(resid)

resid <- resid(best\_mod, type = "deviance")

fitted <- fitted(best\_mod)

plot(fitted, resid, main = "Deviance residuals")

hist(resid)

plot(best\_mod, which = 1) # Pearson residuals

##### prediction

#Here we set a random seed for a random train/test split. Then we split up our dataset in 80 % training data and 20 % testing data.

set.seed(23)

tr = sample(nrow(X), nrow(X)\*0.8)

#Here we train our model only using the training data

train\_model <- glm(Y[tr] ~ school + sex + Fedu + studytime + failures +

schoolsup + higher + Dalc + health, data = X[tr,])

summary(train\_model)

#RMSE on the train set

sqrt(mean((Y[tr] - train\_model$fitted.values)^2))

pred <- predict(train\_model, newdata = X[-tr,], type = "response")

pred

# RMSE on the test set

sqrt(mean((Y[-tr] - pred)^2))

#benchmark

mod\_bench = glm(Y ~ 1, data = X, family = "poisson")

summary(mod\_bench)

pr\_bench=predict(mod\_bench,type="response")

mean((pr\_bench - Y)^2)

rmse <- sqrt(mean((Y - pr\_bench)^2))

rmse