Simone Cipriani, **5951907**

Progetto per “**Data Mining and Organization**”

Repository **git** disponibile su: *https://github.com/gesucca/datamining-class-homework*

DATA UNDERSTANDING

vgrfbfb

# set the repo root path before importing data!!!

setwd("C:/simone\_robamia/unifi/datamining-class-homework/")

students <- read.csv("res/TRE ANNI IMMATRICOLATI 2010-2013 PER STUDENTI.csv")

# gather some info about the imported dataset

str(students)

# ADJUST DATA ATTRIBUTES

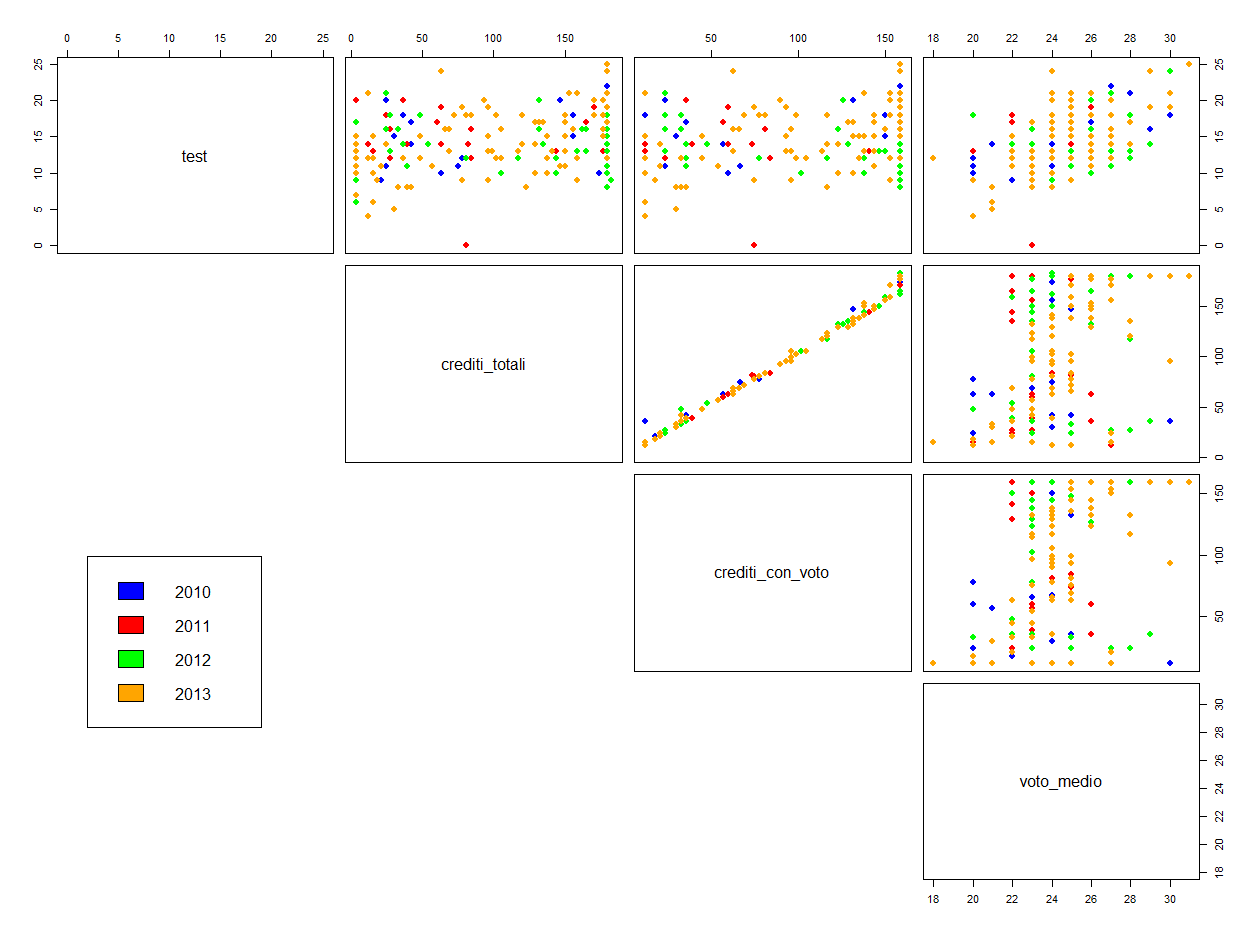
# convert "coorte" to nominal by making it a factor

students[, c(1)] <- sapply(students[, c(1)], factor)

# ok, let's take a look at it now

str(students)

summary(students)



#################

# SCATTER PLOTS #

#################

colors <- c("blue","red", "green", "orange")

coorte\_labels <- students[,1]

coorte\_colors <- colors[as.numeric(coorte\_labels)]

# general attributes

students\_subset1 <- students[,-c(1, 6 : 45)]

pairs(students\_subset1, col = coorte\_colors,lower.panel = NULL,cex.labelsiris=2, pch=19, cex = 1.2)

par(xpd = TRUE)

legend(x = 0.05, y = 0.4, cex = 1,legend = as.character(levels(coorte\_labels)),fill = unique(coorte\_colors))

par(xpd = NA)

# test / avg mark detail

test\_score <- students[,2]

avg\_mark <- students[,5]

plot(jitter(test\_score), jitter(avg\_mark), col = coorte\_colors, pch=19, cex = 1, xlab="Entry Test Score", ylab="Avg. Exam Score")

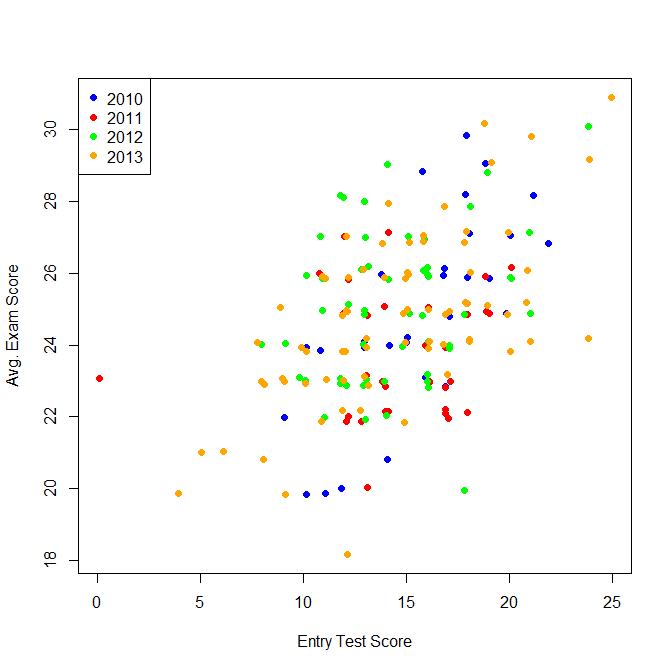
legend(x="topleft", legend = levels(students[,1]), col=colors, pch=19)

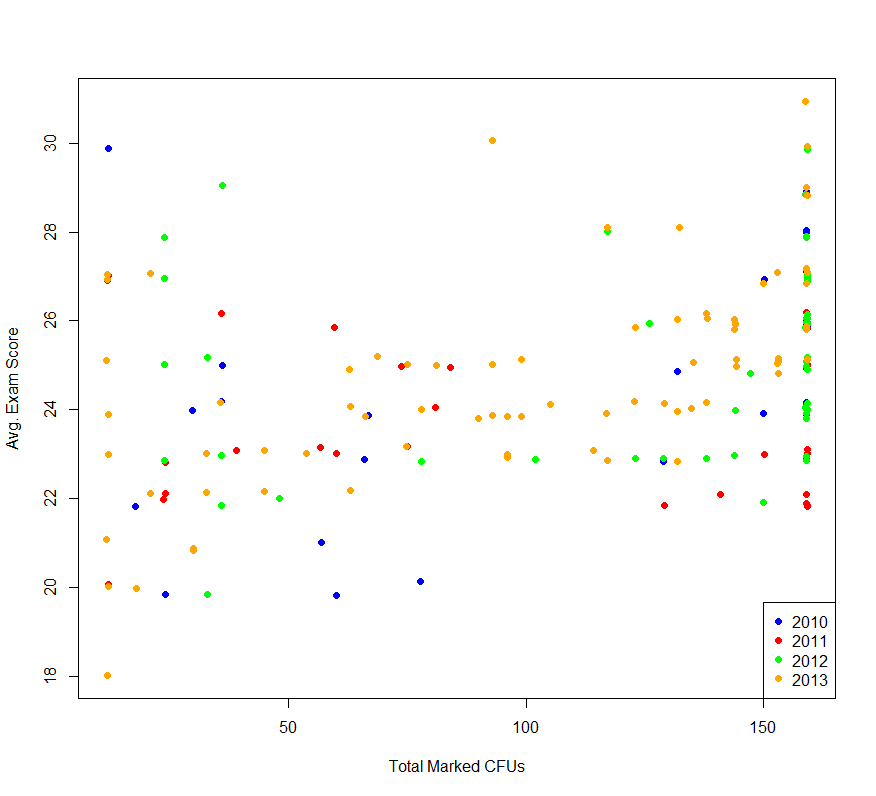
# cfu / avg mark detail

cfu <- students[,4]

plot(jitter(cfu), jitter(avg\_mark), col = coorte\_colors, pch=19, cex = 1, xlab="Total Marked CFUs", ylab="Avg. Exam Score")

legend(x="bottomright", legend = levels(students[,1]), col=colors, pch=19)





# general first year exams performances

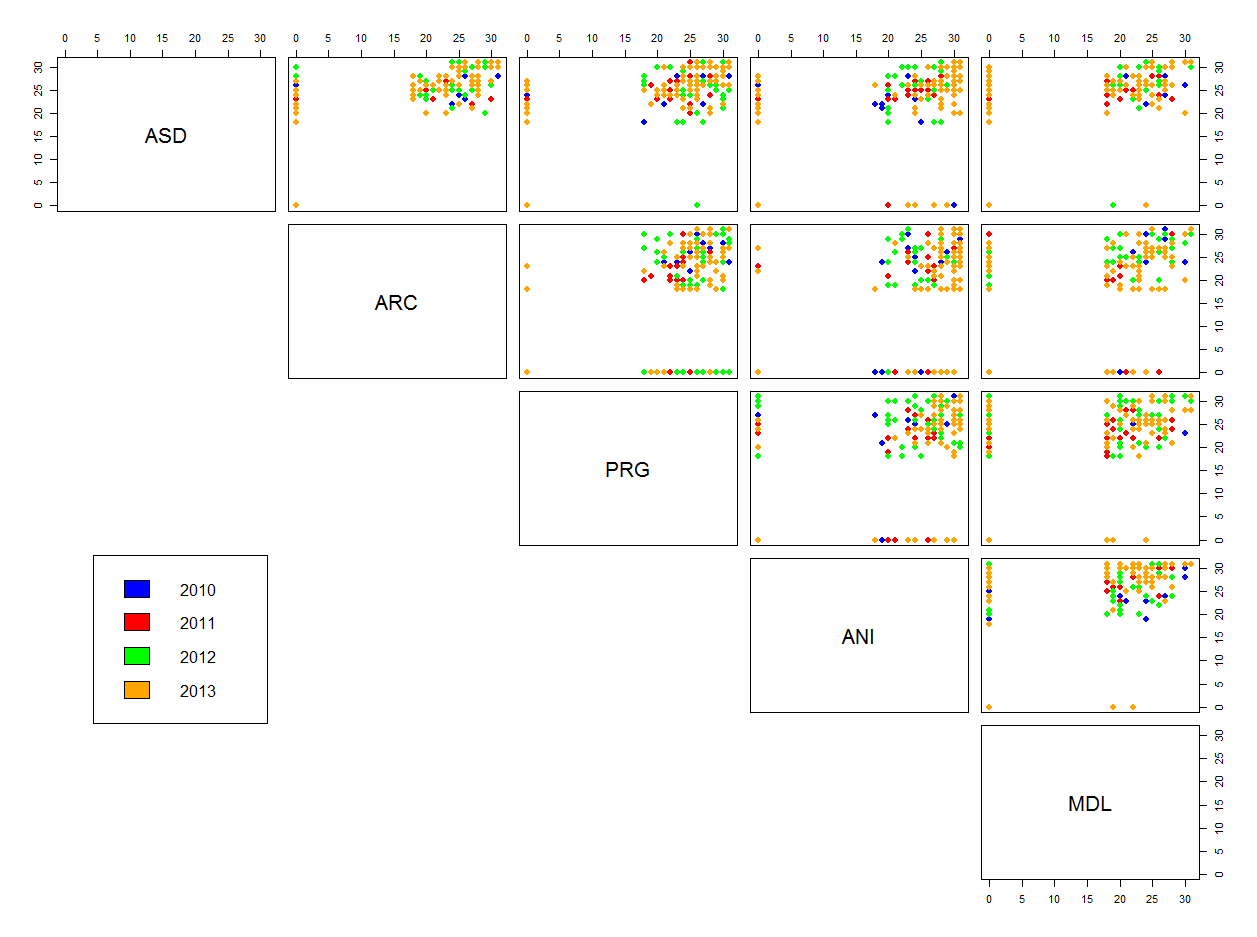
students\_subset2 <- students[,-c(1 : 5, 7, 9, 11, 13, 15 : 45)]

pairs(students\_subset2, col = coorte\_colors,lower.panel = NULL,cex.labelsiris=2, pch=19, cex = 1.2)

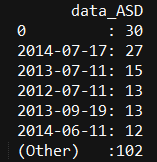
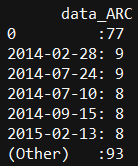
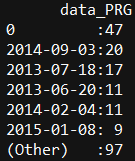
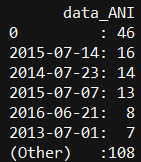
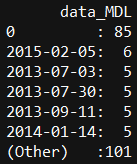
par(xpd = TRUE)

legend(x = 0.05, y = 0.4, cex = 1,legend = as.character(levels(coorte\_labels)),fill = unique(coorte\_colors))

par(xpd = NA))



È possibile notare una caratteristica interessante degli esami del primo anno:

Gli esami meno disputati con successo sono (piuttosto comprensibilmente) Architetture degli Elaboratori e Matematica Discreta e Logica.

Guardiamo in che modo il non superare questi esami ha influito sulla produttività totale:

# first year "difficult" exams / total cfu detail

mdl <- students[,14]

plot(jitter(mdl), jitter(cfu), col = coorte\_colors, pch=19, cex = 1, xlab="M.D.L. Exam Score", ylab="Total Marked CFUs")

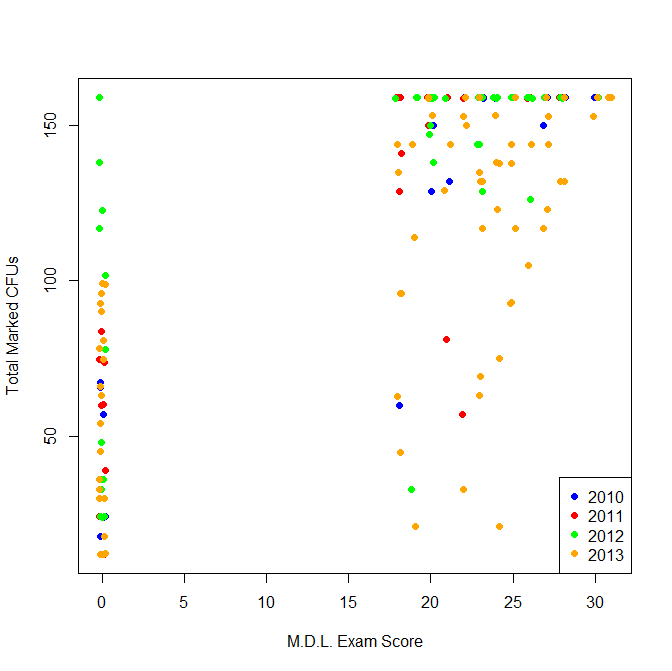
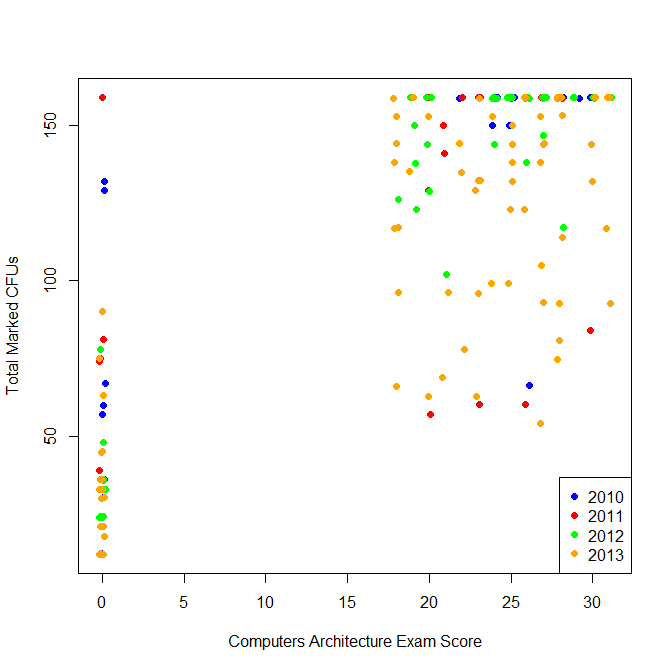
legend(x="bottomright", legend = levels(students[,1]), col=colors, pch=19)

ade <- students[,8]

plot(jitter(ade), jitter(cfu), col = coorte\_colors, pch=19, cex = 1, xlab="Computers Architecture Exam Score", ylab="Total Marked CFUs")

legend(x="bottomright", legend = levels(students[,1]), col=colors, pch=19)

﻿



# computer science exams performances

students\_subset3 <- students[,c(6,8,10,22,24,26,32,36,42)]

pairs(students\_subset3, col = coorte\_colors,lower.panel = NULL,cex.labelsiris=2, pch=19, cex = 0.8)

par(xpd = TRUE)

legend(x = 0.05, y = 0.4, cex = 1,legend = as.character(levels(coorte\_labels)),fill = unique(coorte\_colors))

par(xpd = NA)

# general mathy exams performances

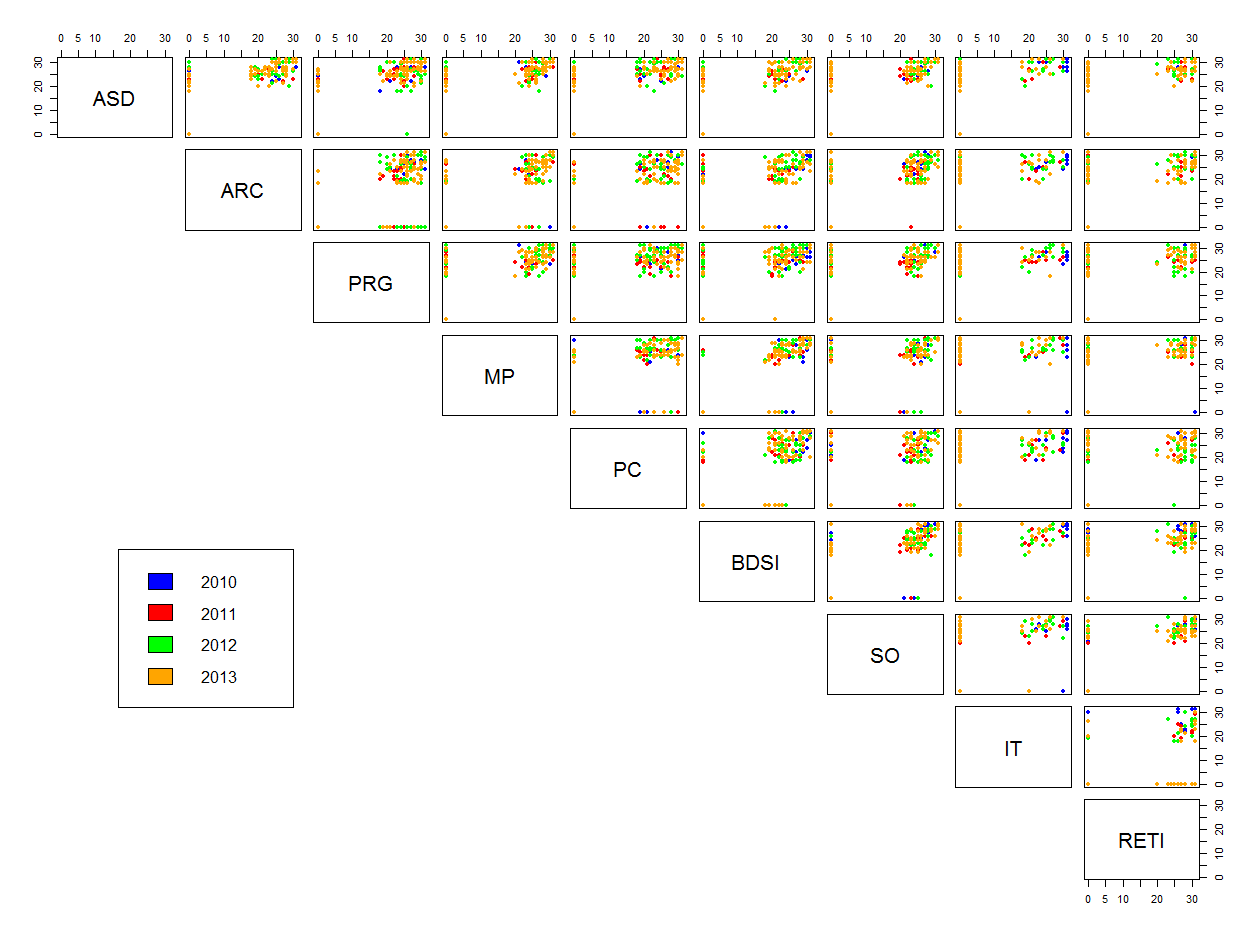
students\_subset4 <- students[,c(12,14,18,20,28,30,34)]

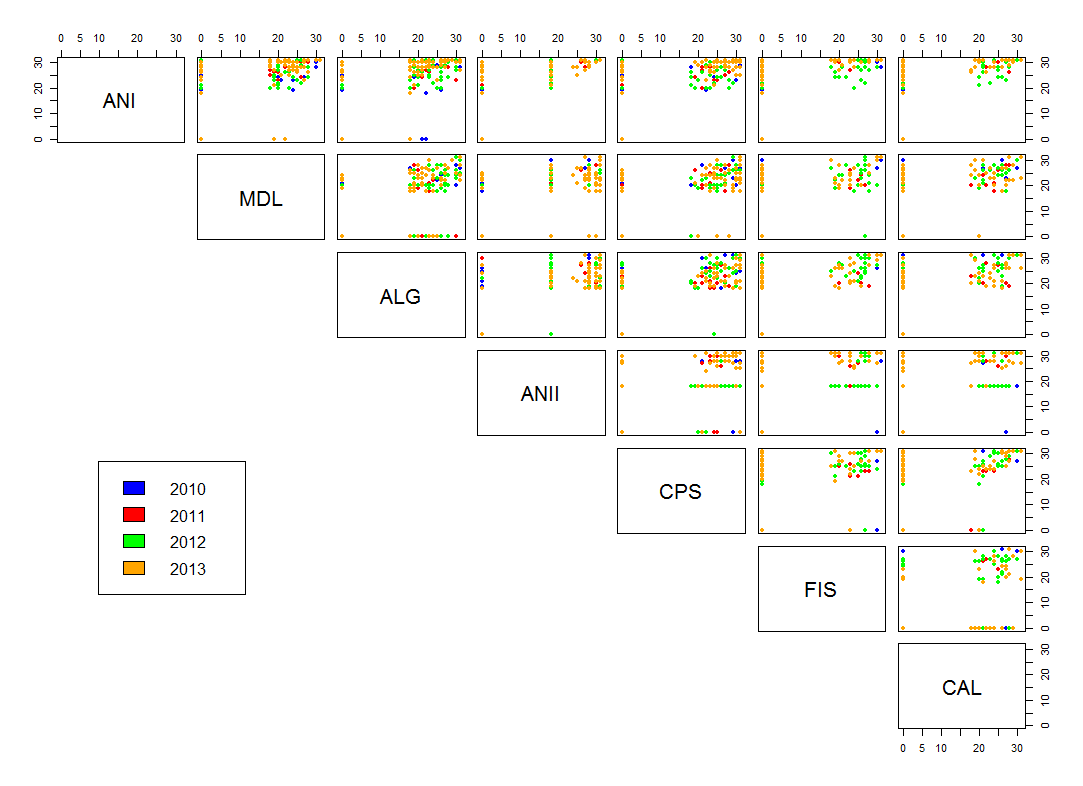
pairs(students\_subset4, col = coorte\_colors,lower.panel = NULL,cex.labelsiris=2, pch=19, cex = 0.8)

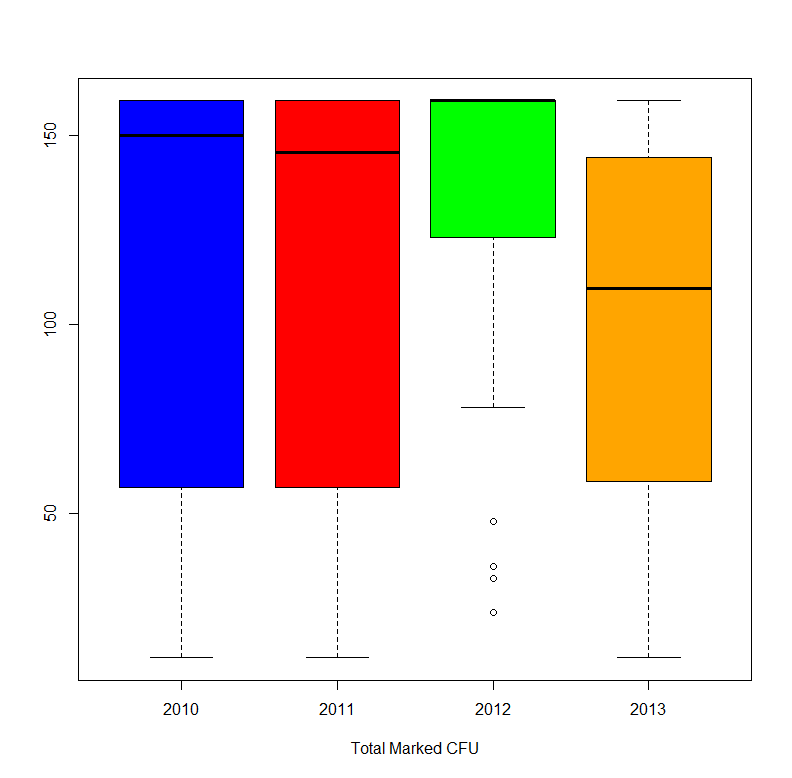
par(xpd = TRUE)

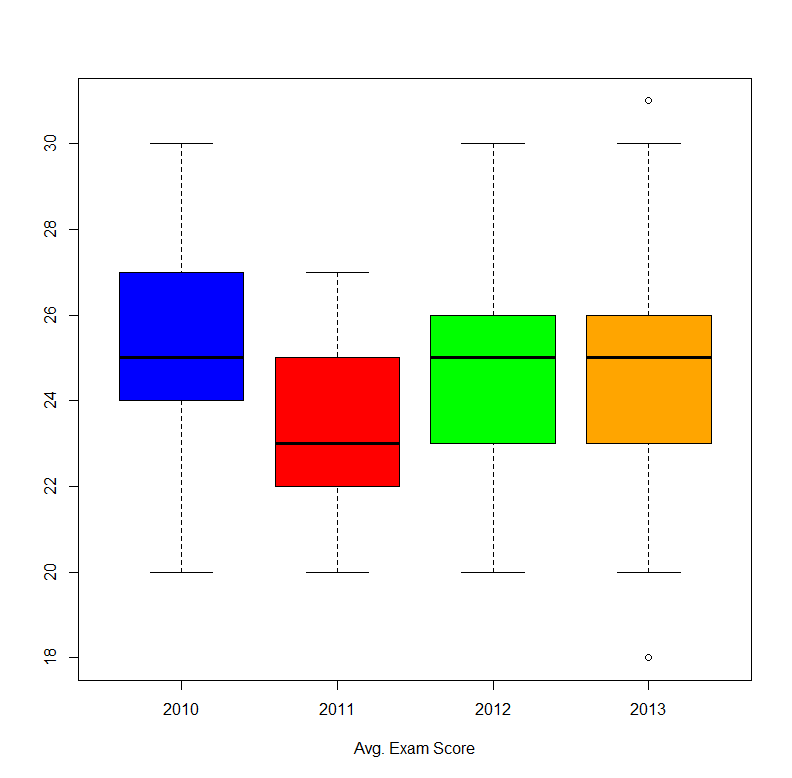
legend(x = 0.05, y = 0.4, cex = 1,legend = as.character(levels(coorte\_labels)),fill = unique(coorte\_colors))

par(xpd = NA)









descriva le caratteristiche del dataset

con calcolo di misure di similarità per vedere se tra i dati esistono particolari correlazioni

tecniche di visualizzazione (in tal caso

preprocessing

algoritmi di clustering che userete.

postprocessing

*Ovviamente tutte queste fasi dovranno essere descritte nella relazione, inserendo i grafici che ritenete più interessanti e commentando i risultati.*

analisi dei risultati