chik_analysis_for_madsen.R

joebrew

Sat Nov 8 18:27:59 2014

```
######
# Attach packages
#######
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
##
## The following object is masked from 'package:stats':
##
##     nobs
##
## The following object is masked from 'package:utils':
##
##     object.size
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(maptools)
```

```
## Loading required package: sp
## Checking rgeos availability: TRUE
```

```
library(rgdal)
```

```
## rgdal: version: 0.9-1, (SVN revision 518)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 1.10.1, released 2013/08/26
## Path to GDAL shared files: /usr/share/gdal/1.10
## Loaded PROJ.4 runtime: Rel. 4.8.0, 6 March 2012, [PJ_VERSION: 480]
## Path to PROJ.4 shared files: (autodetected)
```

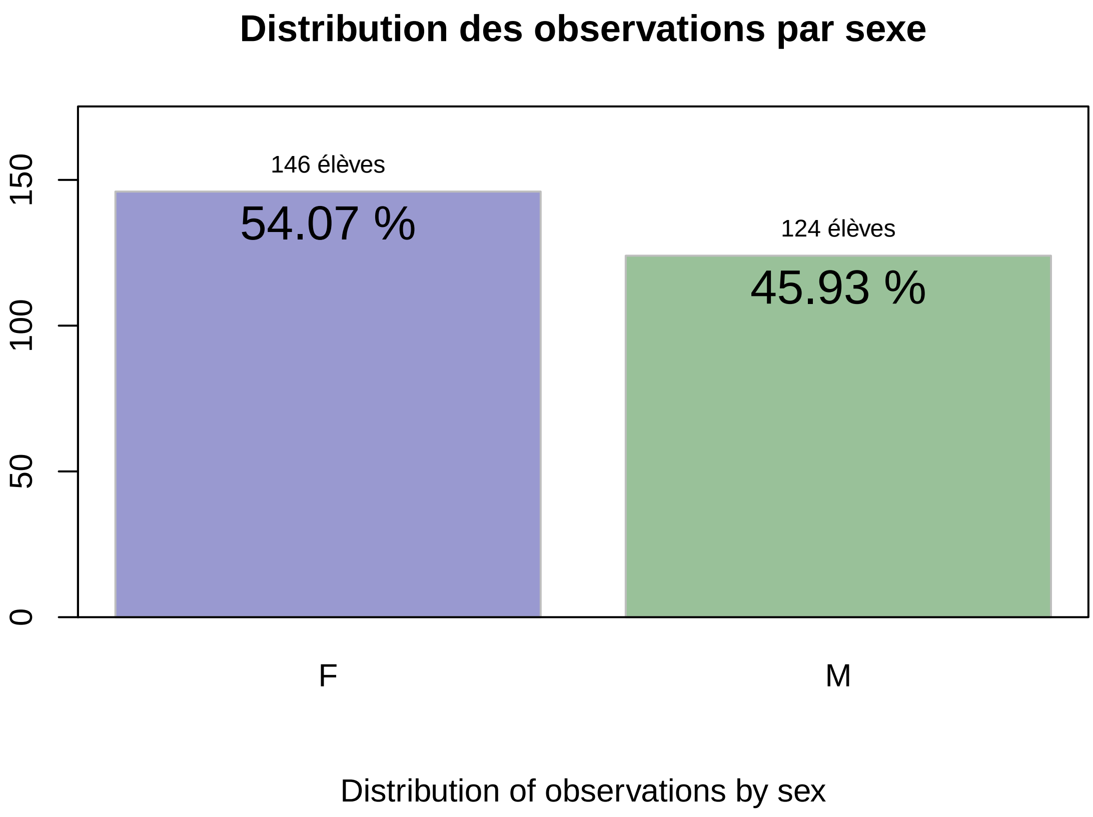
```r
######
# Set working directory to the haiti directories on local machine
######
if(Sys.info()["sysname"] == "Windows"){
  public <- 'C:/Users/BrewJR/Documents/haiti/chik'
  private <- 'C:/Users/BrewJR/Documents/private_data/haiti/chik'
} else {
  public <- '/home/joebrew/Documents/haiti/chick'
  private <- '/home/joebrew/Documents/private_data/haiti/chik'
}
setwd(private)

######
# Read in the 2 datasets sent by Madsen
######
df1 <- read.xls("Clean_Copy of Chikun-V.xlsx",
                stringsAsFactors = FALSE)
df2 <- read.xls("Copy of CHIKV_DATA.xls",
                stringsAsFactors = FALSE)

######
# Clean up column names
######

# lower case all names
names(df1) <- tolower(names(df1))
names(df2) <- tolower(names(df2))

# replace periods with underscores
names(df1) <- gsub("[.]", "_", names(df1))
names(df2) <- gsub("[.]", "_", names(df2))

# clear trailing underscores
names(df1) <- gsub("_(?=_*$)", " ", names(df1), perl=TRUE)
names(df2) <- gsub("_(?=_*$)", " ", names(df2), perl=TRUE)

# clear leading/trailing spaces
names(df1) <- gsub("^\\s+|\\s+$", "", names(df1))
names(df2) <- gsub("^\\s+|\\s+$", "", names(df2))

# clear repeat underscores
names(df1) <- gsub("___|__", "_", names(df1))
names(df2) <- gsub("___|__", "_", names(df2))

######
# Fill missings with NAs
######

# Define function for missing
Missing <- function(var){
  nchar(as.character(var)) == 0
}
```

```r
# Set to NA anything that is missing
for (j in 1:ncol(df1) ){
  x <- df1[,j]
  x[which(Missing(x))] <- NA
  df1[,j] <- x
}

for (j in 1:ncol(df2) ){
  x <- df2[,j]
  x[which(Missing(x))] <- NA
  df2[,j] <- x
}

######
# Format dates
######

# df1
df1$date <- as.Date(df1$date, format = "%Y-%d-%m")
# fill the NAs with the most previous date
for (i in 1:nrow(df1)){
  if(is.na(df1$date[i])){
    correct_dates <- df1$date[1:i] #all the prior dates
    correct_dates <- correct_dates[!is.na(correct_dates)] # remove the NA's
    correct_date <- correct_dates[length(correct_dates)] # take last (most recent)
  } else{
    correct_date <- df1$date[i]
  }
  df1$date[i] <- correct_date
}

df2$date_collected <- as.Date(df2$date_collected, format = "%d-%b-%y")
df2$extraction_date <- as.Date(df2$extraction_date, format = "%d-%b-%y")
df2$pcr_date <- as.Date(df2$pcr_date, format = "%d-%b-%y")

######
# Clean up temperature column
######
df1$temperature <- as.numeric(gsub("°C|\\s", "", df1$temperature))

######
# Clean up localisation column
######
df1$localisation <- gsub("D°|[:]|[.]", "", df1$localisation)
df1$localisation <- gsub("^\\s+|\\s+$", "", df1$localisation)
df1$localisation[which(Missing(df1$localisation))] <- NA

######
# Clean up associatd_symp column
######
df1$associatd_symp <- gsub("D°|D °|°|[:]|[.]", "", df1$associatd_symp)
df1$associatd_symp <- gsub("^\\s+|\\s+$", "", df1$associatd_symp)
df1$associatd_symp[which(Missing(df1$associatd_symp))] <- NA
```

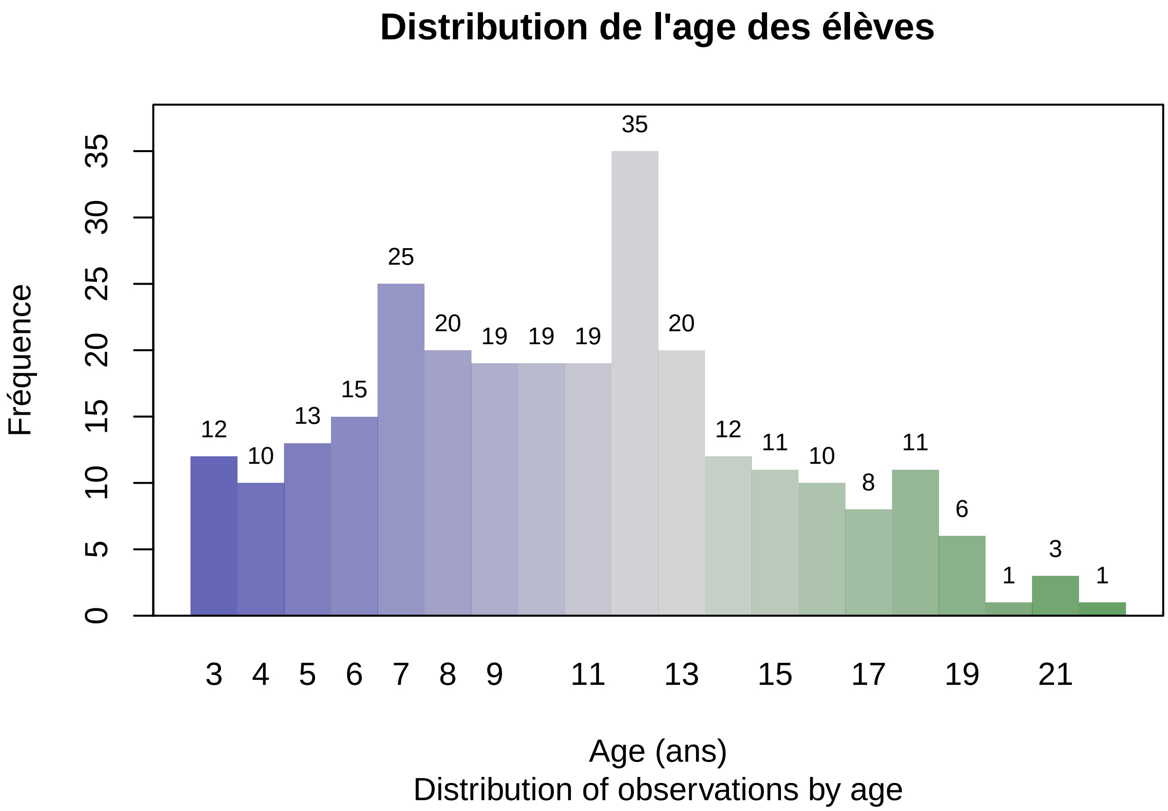
```
######
# Merge by child code
######
df2$child_code <- df2$child_code_no
df2$child_code_no <- NULL

df <- merge(x = df1,
            y = df2,
            by = "child_code",
            all.x = TRUE,
            all.y = FALSE)

####################################################################
# DATA CLEANING DONE
####################################################################

# Madsen's instructions: do an analysis on sex, age, grade, temperature
# age by chikv_rst
# sex by chikv_rst

######
# SEX
######
table(df$sexe)
```

```
##
##   F   M
## 146 124
```

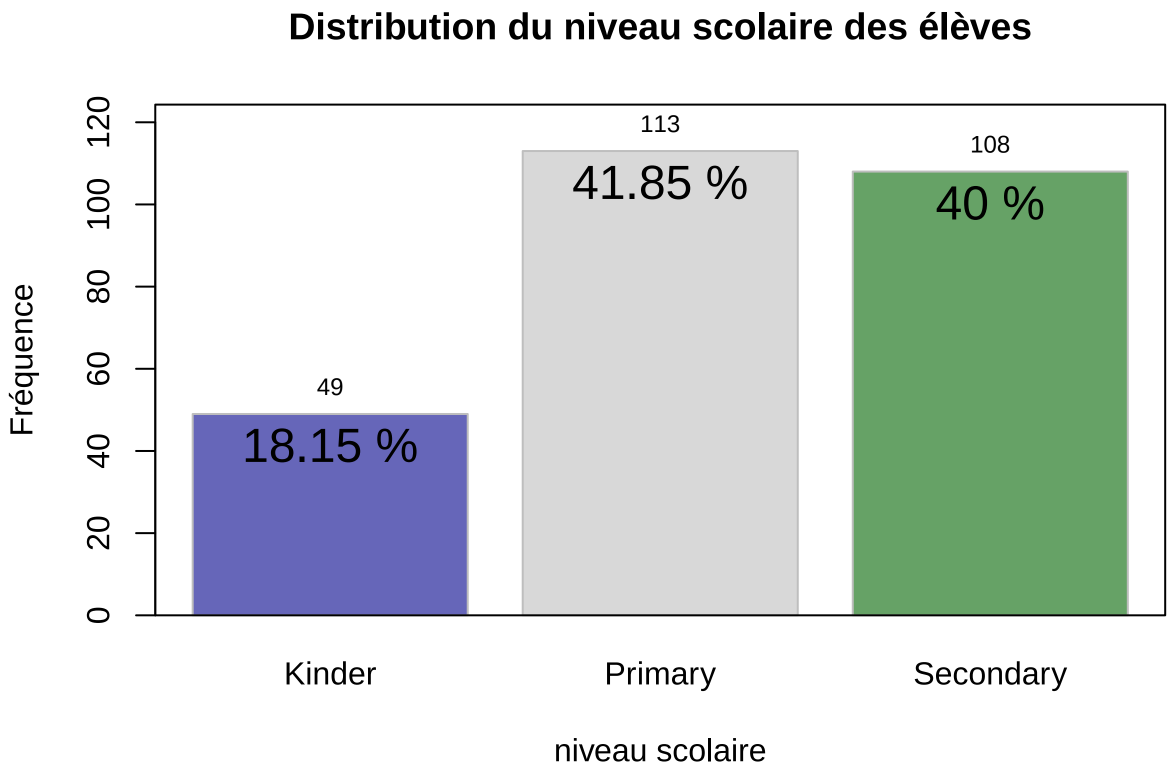
```
mybp <- barplot(table(df$sexe),
                ylim = c(0, max(table(df$sexe))*1.2),
                col = adjustcolor(c("darkblue", "darkgreen"),
                                  alpha.f = 0.4),
                border = "grey")
text(x = mybp[,1],
     y = table(df$sexe),
     pos = 1,
     labels = paste0(100* round(prop.table(table(df$sexe)), digits = 4)," %"),
     cex = 1.5)
text(x = mybp[,1],
     y = table(df$sexe),
     pos = 3,
     labels = paste0(table(df$sexe), " élèves"),
     cex = 0.75)
box("plot")
title(main = "Distribution des observations par sexe")
title(sub = "Distribution of observations by sex")
```



```
######
# AGE
######
ages <- unique(sort(df$age))
my_colors <- adjustcolor(colorRampPalette(c("darkblue", "grey", "darkgreen"))(length(ages)), alpha.f =
mybp <- barplot(table(df$age), col = my_colors,
        xlab = "Age (ans)",
        ylab = "Fréquence",
        ylim = c(0, max(table(df$age)*1.1)),
        main = "Distribution de l'age des élèves",
        border = NA,
        space = 0)
text(x = mybp[,1],
     y = table(df$age),
     pos = 3,
     labels = table(df$age),
     cex = 0.75)
title(sub = "Distribution of observations by age")
box("plot")
```

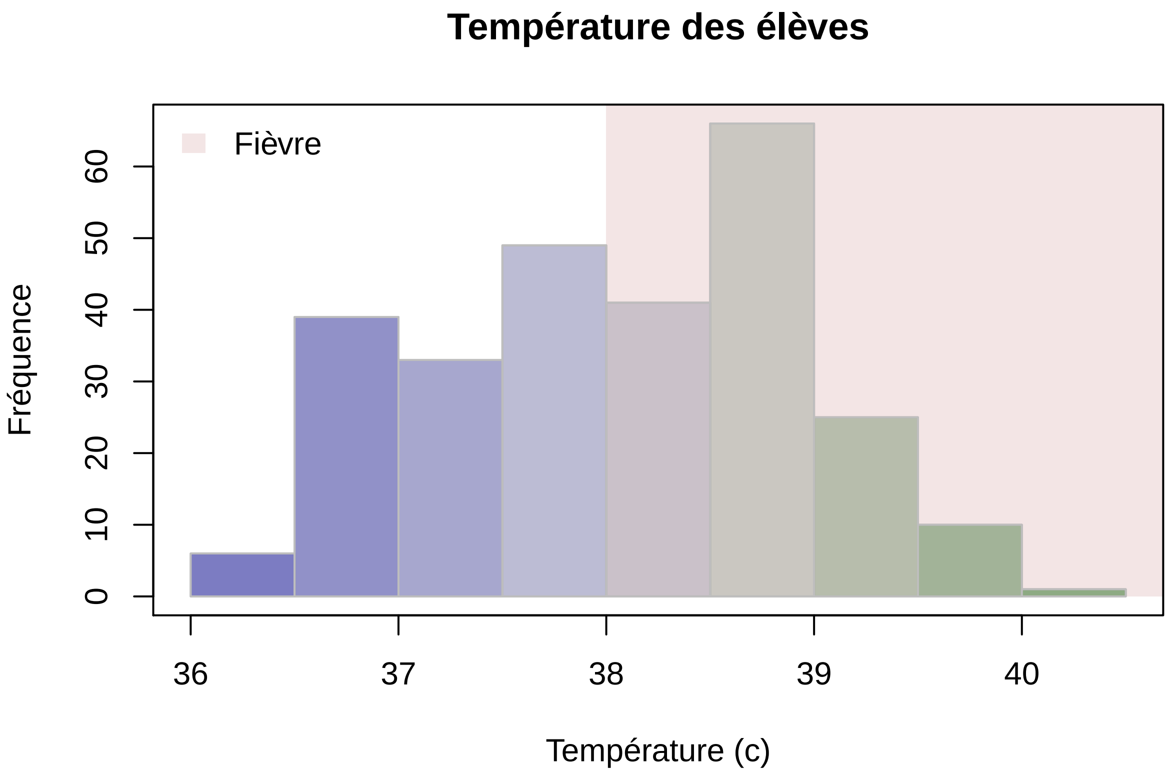


```
######
# GRADE
######
grades <- unique(sort(df$grade))
my_colors <- adjustcolor(colorRampPalette(c("darkblue", "grey", "darkgreen"))(length(grades)), alpha.f =
mybp <- barplot(table(df$grade), col = my_colors,
                xlab = "niveau scolaire",
                ylab = "Fréquence",
                ylim = c(0, max(table(df$grade)*1.1)),
                main = "Distribution du niveau scolaire des élèves",
                border = "grey")
text(x = mybp[,1],
     y = table(df$grade),
     pos = 3,
     labels = table(df$grade),
     cex = 0.75)
text(x = mybp[,1],
     y = table(df$grade),
     pos = 1,
     labels = paste0(100* round(prop.table(table(df$grade)), digits = 4)," %"),
     cex = 1.5)
title(sub = "Distribution of school level by grade")
box("plot")
```

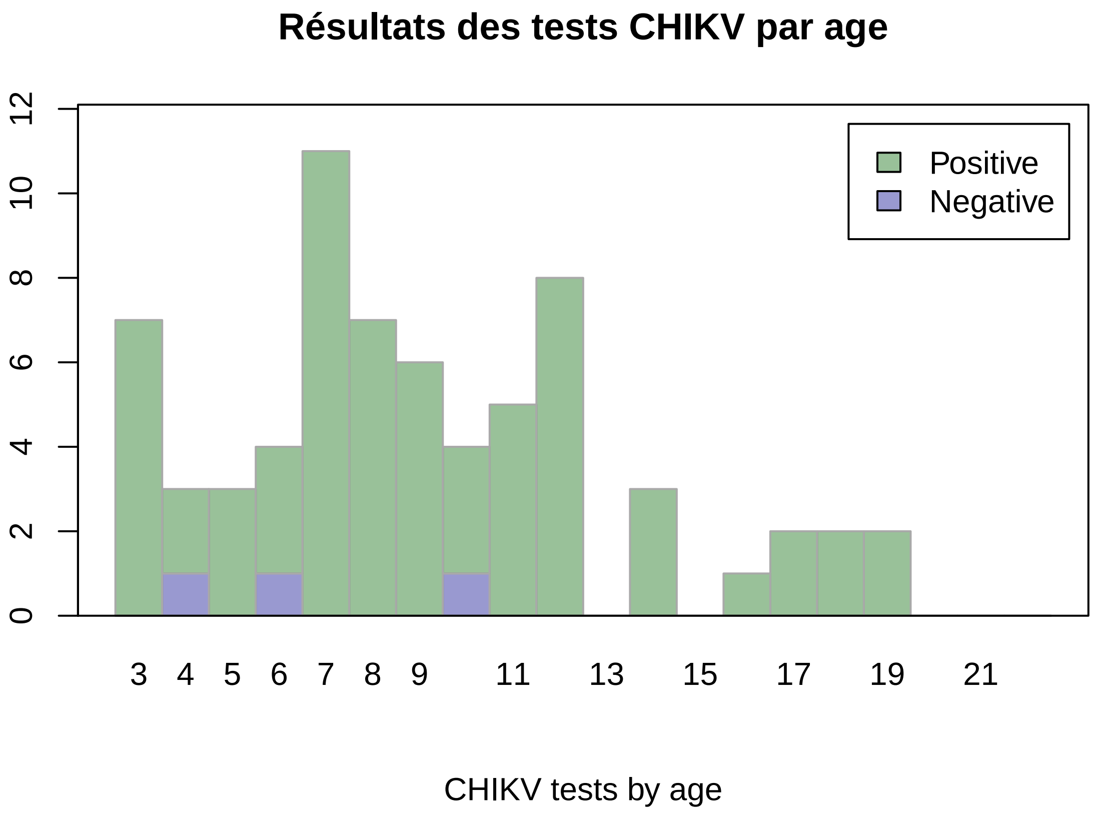


Distribution of school level by grade

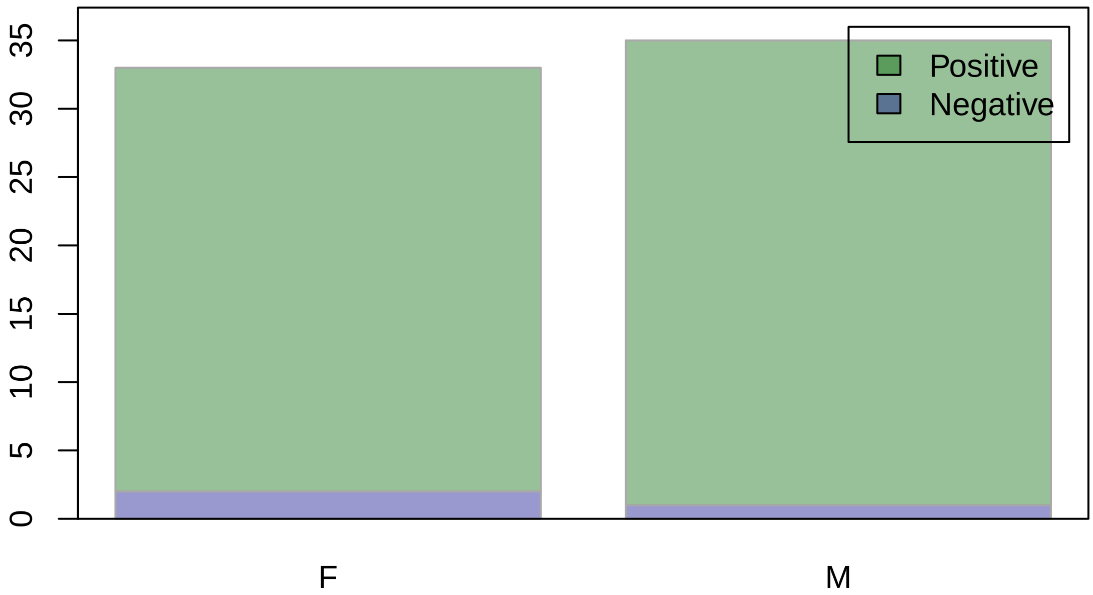
```r
#######
# TEMPERATURE
#######
my_colors <- adjustcolor(colorRampPalette(c("darkblue", "grey", "darkgreen"))(10), alpha.f = 0.3)
myhist <- hist(df$temperature, breaks = 10,
               col = my_colors,
               border = "grey",
               xlab = "Température (c)",
               ylab = "Fréquence",
               main = "Température des élèves")
polygon(x = c(38,50, 50, 38),
        y = c(0, 0, 100, 100),
        col = adjustcolor("darkred", alpha.f = 0.1),
        border = NA)
hist <- hist(df$temperature, breaks = 10,
               col = my_colors,
               border = "grey", add = TRUE)
legend(x = "topleft",
       fill = adjustcolor("darkred", alpha.f = 0.1),
       border = NA,
       bty = "n",
       legend = "Fièvre")
box("plot")
```



```
#######
# age by chikv_rst
#######
x <- table(df$chikv_rst,
      df$age)
barplot(x,
        col = adjustcolor(c("darkblue", "darkgreen"), alpha.f = 0.4),
        legend = TRUE,
        border = "darkgrey",
        space = 0,
        ylim = c(0,max(x)*1.1))
box("plot")
title(main = "Résultats des tests CHIKV par age")
title(sub = "CHIKV tests by age")
```



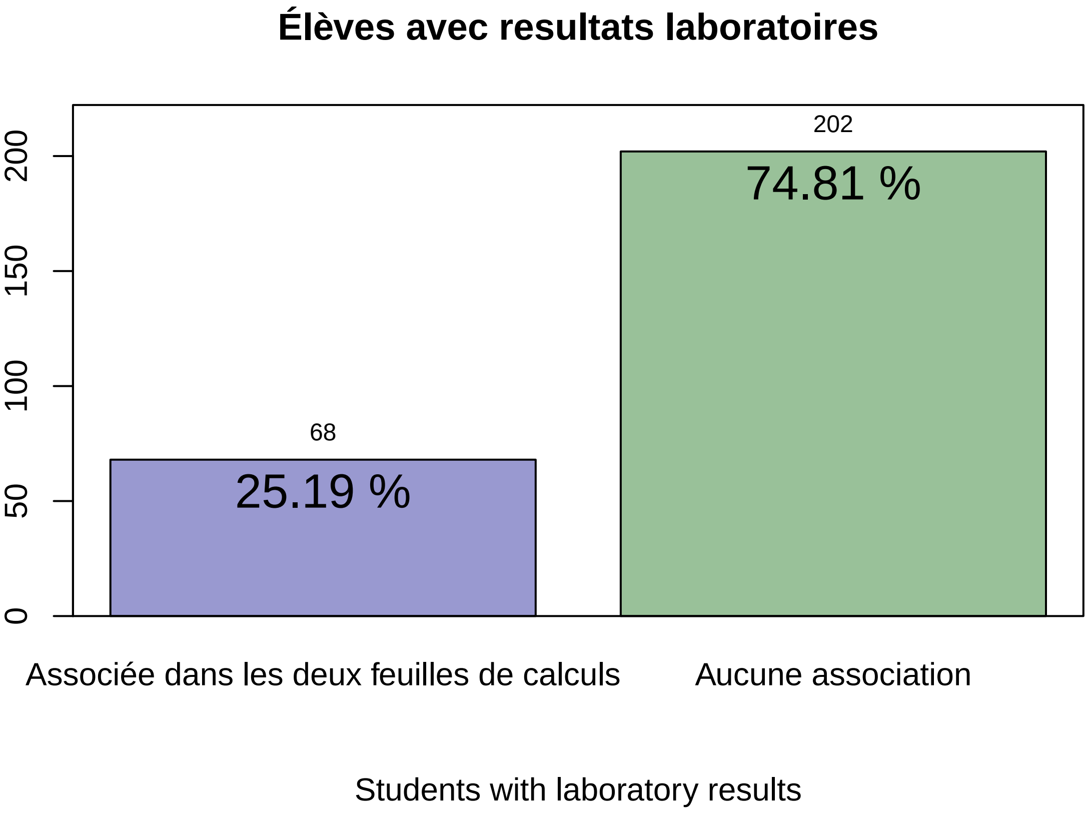
```
#######
# sex by chikv_rst
#######
x <- table(df$chikv_rst,
           df$sex)
barplot(x,
        col = adjustcolor(c("darkblue", "darkgreen"), alpha.f = 0.4),
        legend = TRUE,
        border = "darkgrey",
        ylim = c(0,max(x)*1.1))
box("plot")
title(main = "Résultats des tests CHIKV par sexe")
title(sub = "CHIKV tests by sex")
```



Résultats des tests CHIKV par sexe

CHIKV tests by sex

```
# chisq.test(x)

######
# % MATCHED
######
x <- table(is.na(df$chikv_rst))
bp <- barplot(x,
        names.arg = c("Associée dans les deux feuilles de calculs",
                      "Aucune association"),
        col = adjustcolor(c("darkblue", "darkgreen"), alpha.f = 0.4),
        ylim = c(0, max(table(is.na(df$chikv_rst)))*1.1))
text(x = mybp[,1],
     y = table(is.na(df$chikv_rst)),
     pos = 3,
     labels = table(is.na(df$chikv_rst)),
     cex = 0.75)
text(x = mybp[,1],
     y = table(is.na(df$chikv_rst)),
     pos = 1,
     labels = paste0(100* round(prop.table(table(is.na(df$chikv_rst))), digits = 4)," %"),
     cex = 1.5)
box("plot")
title(main = "Élèves avec resultats laboratoires")
title(sub = "Students with laboratory results")
```


Élèves avec resultats laboratoires
0
50
100
150
200
68
25.19 %
202
74.81 %
Associée dans les deux feuilles de calculs
Aucune association
Students with laboratory results