History survey in 2003 provided with the “questionr” package. The data set is: hdv2003.

1. **Describe and overview the dataset:**

install.packages("questionr")

library(questionr)

data(hdv2003)

**# assign dataset to a variable d**

d <- hdv2003

**# display size of dataset**

dim(d)

**#display field names**

names(d)

**#show info on data types (including first several value per each field)**

str(d)

**#show attributes of dataset (row, fields, class)**

attributes(d)

**#show first 3 row for each data field**

d[1:3,]

**#show first 10 row of datafield "qualif"**

d[1:10, "qualif"]

**#show summary: min, max, median, etc. per each field**

summary(d)

install.packages("Hmisc")

library(Hmisc)

**#describe fields 1 and 6. Desctiptive info in output depends of data types in each data field**

describe(d[,c(1,6)])

**#show range for datafield "id" (data field must be numerical)**

range(d$id)

**#show quantile for datfield "age" (by default every 25%)**

quantile(d$age)

**#show quantile for datfield "age" set at 0.1, 0.4 and 0.8**

quantile(d$age, c(0.1,0.4,0.8))

**#show var for values in datafield "age"**

var(d$age)

**#build a histogram for values in field "age"**

hist(d$age)

**#plot density distributed values in field "age"**

plot(density(d$age))

**#create a pivot table for values in datafield "occup""**

table(d$occup)

**#build a pie chart for values in field "sex" based on pivot table**

pie(table(d$sex))

**#build a barplot for datafield "age"**

barplot(table(d$age))

**#find a covariance between age and weight. Negative covariance indicates inverse relationship"**

cov(d$age, d$poids)

**#find a correlation between age and weight.Negative correlation is used in statistics to measure**

**#the amount that a change in one variable can affect an opposite change in another variable.**

cor(d$age, d$poids)

**#aggregate columns "sexe" by "age"**

aggregate(sexe ~ age, summary, data = d)

**#create whisker plot on "age" and "occup"**

plot(d$occup, d$age)

**#create whisker plot on "age" and "occup" in color**

with (d, plot(d$occup, d$age, col=age, pch=as.numeric(age)))

**#create 3D plot on data fields: "occup", "age", "sexe"**

install.packages("scatterplot3d")

library(scatterplot3d)

scatterplot3d(d$occup, d$age, d$sexe)

**#create 2 plots of occup vs age: for male and female**

install.packages("ggplot2")

library(ggplot2)

qplot(d$age, d$occup, data=d, facets = sexe ~.)

1. **Explore by creating the following graphics**

**#2-1 – Explore the variable hours to view the number of hours**

**#spent watching television**

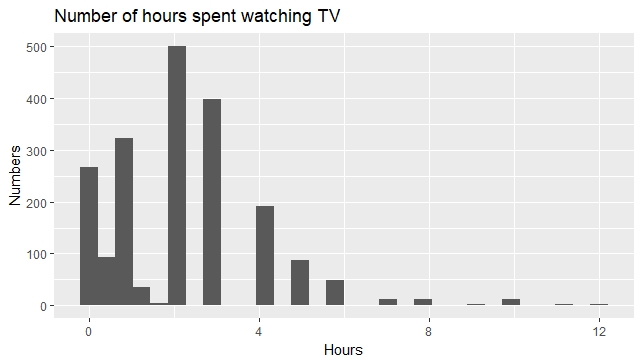
p <- ggplot(d, aes(heures.tv))

p + geom\_histogram()+ labs(x = "Hours", y = "Numbers",

title = "Number of hours spent watching TV")

scale\_x\_continuous(breaks = seq(0, 12, by = 4)) +

scale\_y\_continuous(breaks = seq(0, 500, by = 100))



**#2-2 Modify the graphic to customize the details as follow**

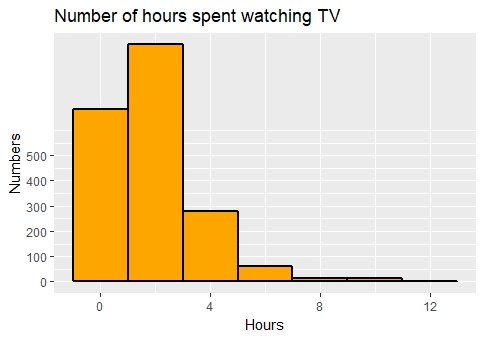
p <- ggplot(d, aes(heures.tv))

p + geom\_histogram(bins = 7, fill="orange", color="black", size=1) + labs(x = "Hours", y = "Numbers",

title = "Number of hours spent watching TV")+

scale\_x\_continuous(breaks = seq(0, 12, by = 4)) +

scale\_y\_continuous(breaks = seq(0, 500, by = 100))



**#2-3 View the density curve for “hours” variable**

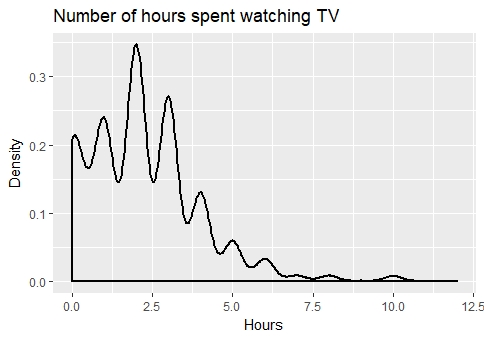
p <- ggplot(d, aes(heures.tv))

p +geom\_density(color="black", size=1) + labs(x = "Hours", y = "Density",

title = "Number of hours spent watching TV")+

scale\_x\_continuous(breaks = seq(0, 12.5, by = 2.5)) +

scale\_y\_continuous(breaks = seq(0, 0.3, by = 0.1))



**#2-4- View the boxplot of the following variables: “hard.rock” and “age”**

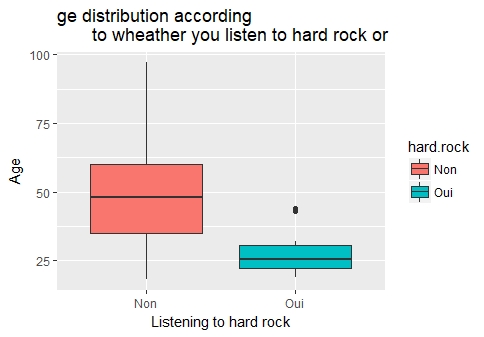
p <- ggplot(d, aes(x=hard.rock, y=age, fill=hard.rock)) +

geom\_boxplot(size=0.5)+

labs(x = "Listening to hard rock", y = "Age", title="ge distribution according

to wheather you listen to hard rock or")

p



**#2-5 Using a bar char view the variable “freres.soeurs”**

ggplot(d, aes(freres.soeurs)) + geom\_bar()+

labs(x = "Number of brothers and sisters", y = "Numbers")

