top101 visualization

MarijnJABoer

17/01/2018

```
setwd(
    "~/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/FormicID/stat/f
)

library(ggplot2) # for plotting
library(jpeg)
library(imager) # for reading jpg files
```

```
## Loading required package: plyr
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'imager'
```

```
## The following object is masked from 'package:magrittr':
##
##     add
```

```
## The following object is masked from 'package:plyr':
##
##     liply
```

```
## The following objects are masked from 'package:stats':
##
##     convolve, spectrum
```

```
## The following object is masked from 'package:graphics':
##
##     frame
```

```
## The following object is masked from 'package:base':
##
##     save.image
```

```
library(reshape2) # for using melt()
library(magick) # for image conversion because jpg files are read
```

```
## Warning: package 'magick' was built under R version 3.4.3
```

```
## Linking to ImageMagick 6.9.9.25
## Enabled features: cairo, fontconfig, freetype, fftw, lcms, pango, rsvg, webp
## Disabled features: ghostscript, x11
```

```
top101 <- read.csv('top101.csv') # spreadsheat containing catalognumber, scientific name,
                                 # shot_type and image url
```

Dataset

```
head(top101)
```

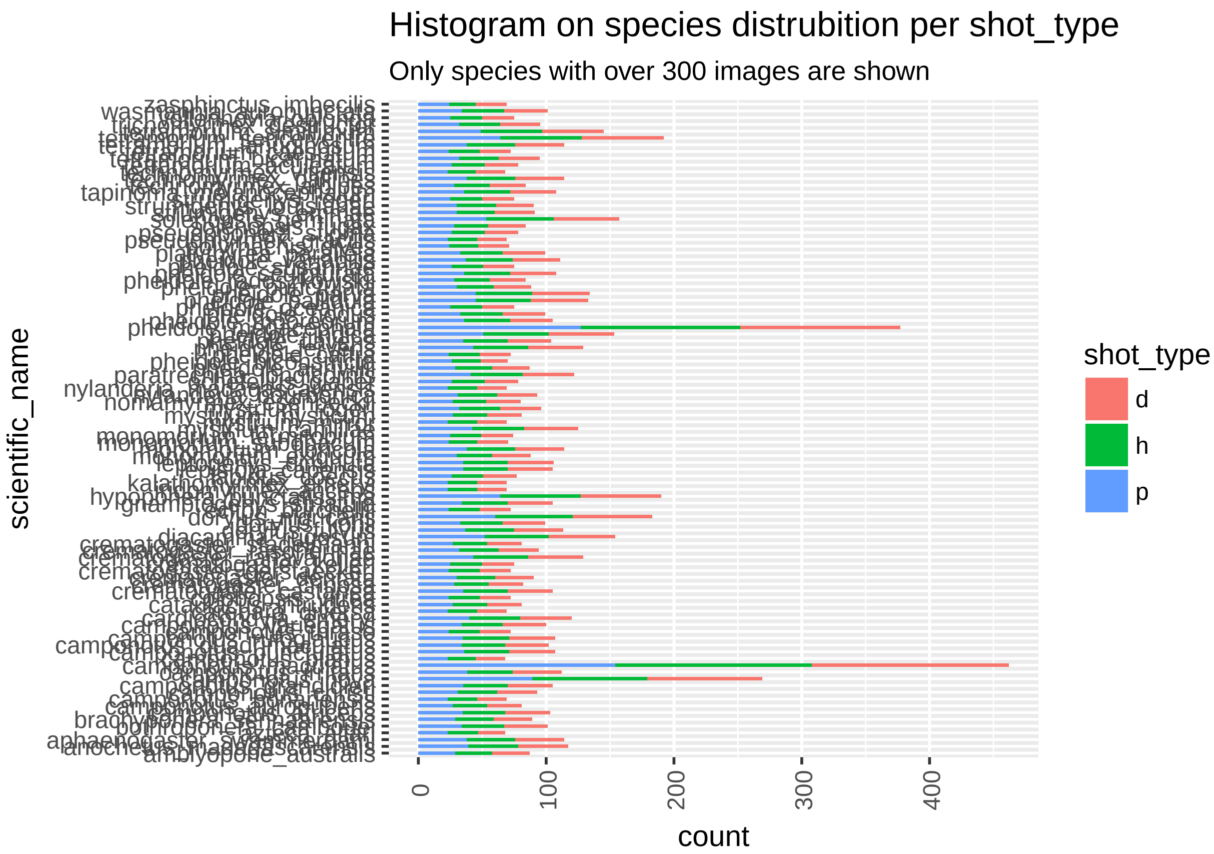
```
##   catalog_number      scientific_name shot_type
## 1  casent0102125 amblyopone_australis         h
## 2  casent0102125 amblyopone_australis         p
## 3  casent0102125 amblyopone_australis         d
## 4  casent0102148 amblyopone_australis         h
## 5  casent0102148 amblyopone_australis         p
## 6  casent0102148 amblyopone_australis         d
##                                                        image_url
## 1 http://www.antweb.org/images/casent0102125/casent0102125_h_1_low.jpg
## 2 http://www.antweb.org/images/casent0102125/casent0102125_p_1_low.jpg
## 3 http://www.antweb.org/images/casent0102125/casent0102125_d_1_low.jpg
## 4 http://www.antweb.org/images/casent0102148/casent0102148_h_1_low.jpg
## 5 http://www.antweb.org/images/casent0102148/casent0102148_p_1_low.jpg
## 6 http://www.antweb.org/images/casent0102148/casent0102148_d_1_low.jpg
```

```
summary(top101)
```

```
##        catalog_number                  scientific_name shot_type
##  anic32-002152:    3   camponotus_maculatus     : 462   d:3409
##  anic32-002153:    3   pheidole_megacephala     : 377   h:3392
##  anic32-002156:    3   camponotus_hova          : 269   p:3424
##  anic32-063120:    3   tetramorium_sericeiventre: 192
##  antweb1008080:    3   hypoponera_punctatissima : 190
##  antweb1008081:    3   dorylus_nigricans        : 183
##  (Other)      :10207   (Other)                  :8552
##                                                          image_url
##  http://www.antweb.org/images/anic32-002152/anic32-002152_d_1_low.jpg:    1
##  http://www.antweb.org/images/anic32-002152/anic32-002152_h_1_low.jpg:    1
##  http://www.antweb.org/images/anic32-002152/anic32-002152_p_1_low.jpg:    1
##  http://www.antweb.org/images/anic32-002153/anic32-002153_d_1_low.jpg:    1
##  http://www.antweb.org/images/anic32-002153/anic32-002153_h_1_low.jpg:    1
##  http://www.antweb.org/images/anic32-002153/anic32-002153_p_1_low.jpg:    1
##  (Other)                                                              :10219
```

Image distribution per species

```
# Ploting image distribution per shot type
g <- ggplot(top101)
g + geom_bar(aes(scientific_name, fill = shot_type), width = 0.5) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.6)) +
    labs(title = "Histogram on species distrubition per shot_type",
         subtitle = "Only species with over 300 images are shown") +
    coord_flip()
```



File size distribution

```r
data_dir <-
    "~/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/FormicID/data/2

files_info <-
    file.info(list.files(
        path = data_dir,
        pattern = ".jpg",
        full.names = TRUE
    ))
head(files_info)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/Europe/Amsterdam'

##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
```

```
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
```

```
##
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
## /Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CNN/8. FormicID/Fo
```

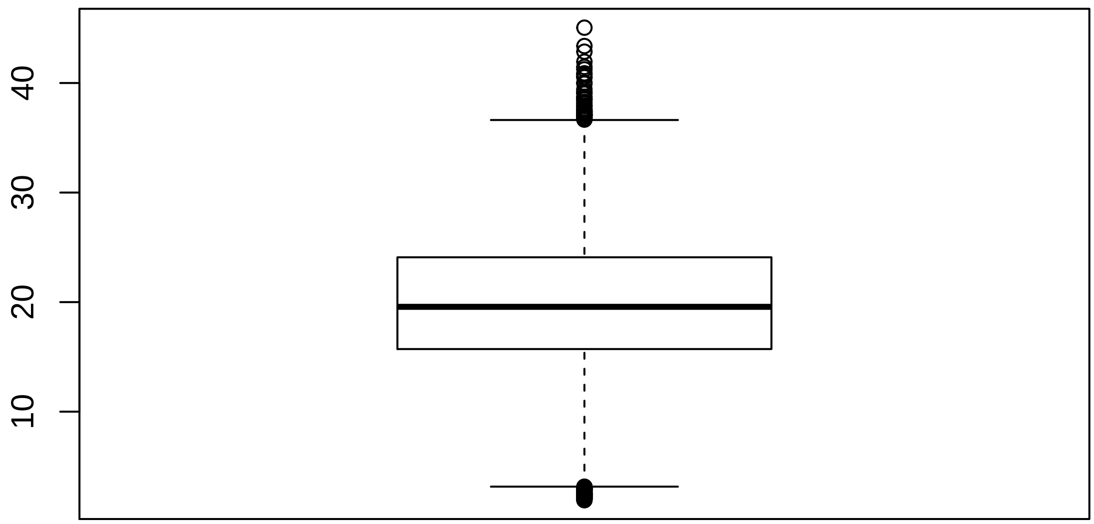
```r
# Convert bytes to kilobytes
files_info$size2 <- files_info$size / 1000

# https://www.statmethods.net/graphs/density.html

# Histogram with normal curve plot
size_kb <- files_info$size2
cat('The mean size is ', mean(size_kb), 'kb')
```

```
## The mean size is  19.87619 kb
```

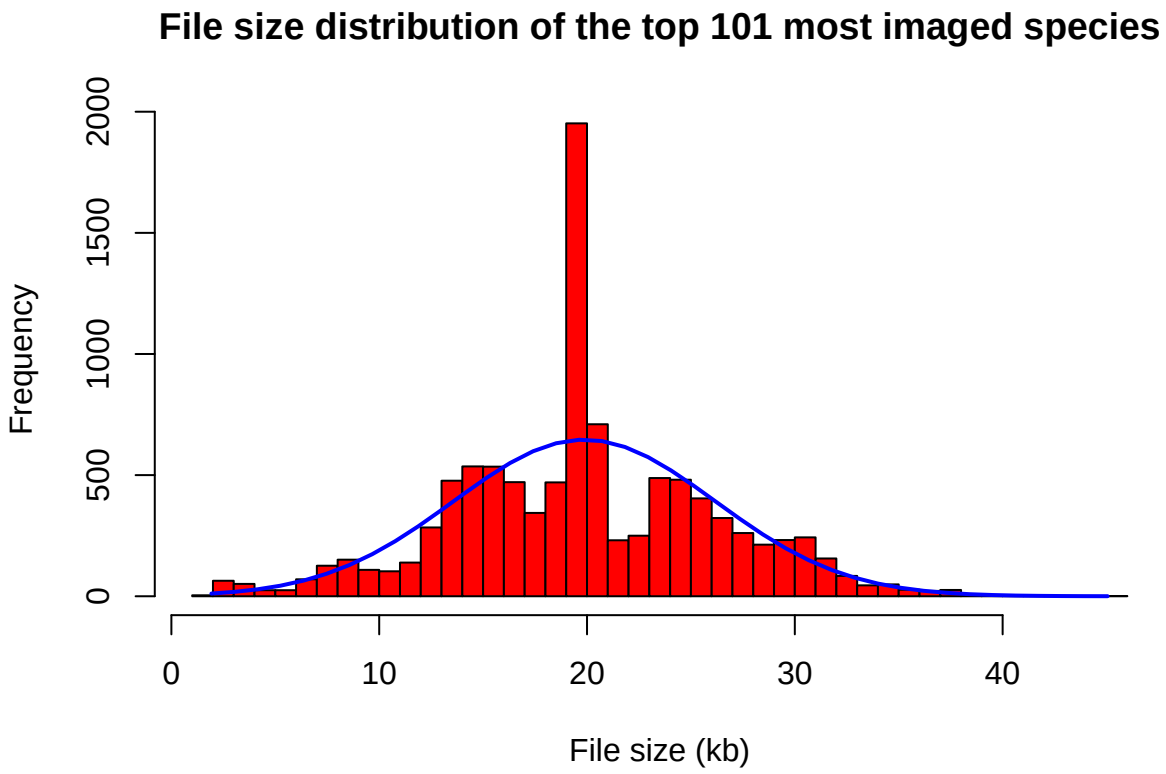
```r
summary(size_kb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.925  15.715  19.569  19.876  24.090  45.051
```

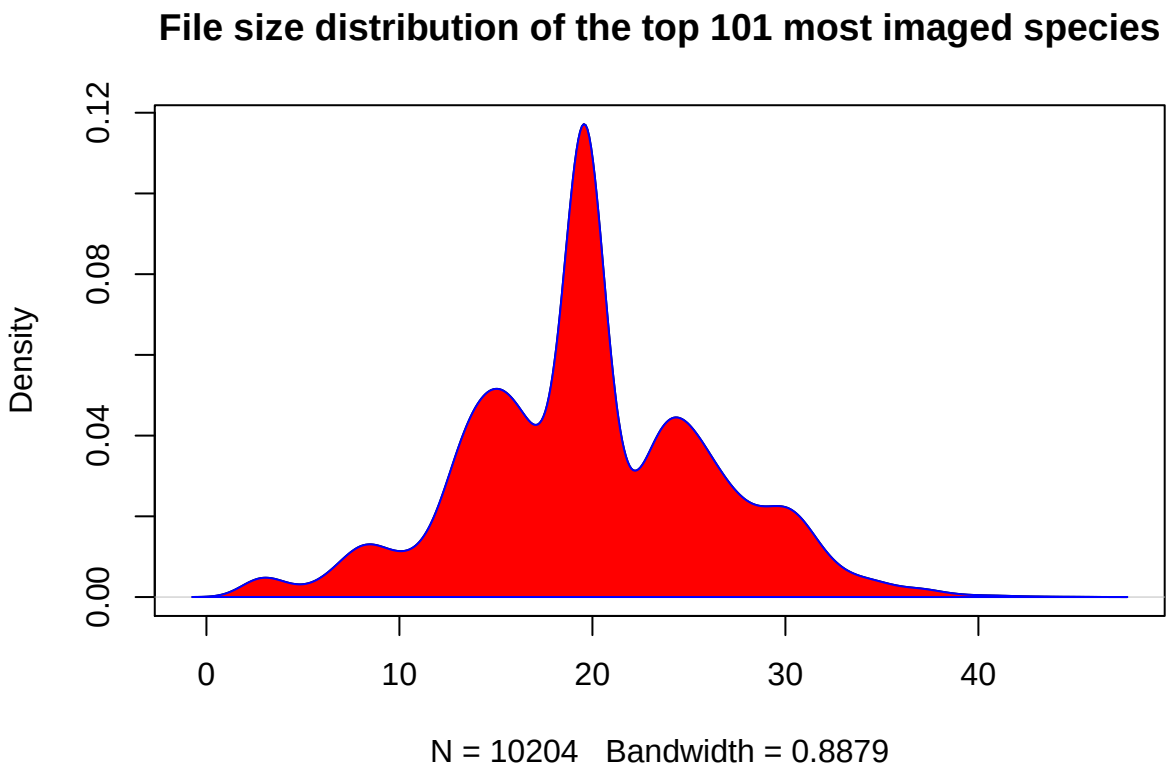
```r
boxplot(size_kb)
```




```r
h <- hist(
    size_kb,
    breaks = 50,
    col = "red",
    xlab = "File size (kb)",
    main = "File size distribution of the top 101 most imaged species"
)
xfit <- seq(min(size_kb), max(size_kb), length = 40)
yfit <- dnorm(xfit, mean = mean(size_kb), sd = sd(size_kb))
yfit <- yfit * diff(h$mids[1:2]) * length(size_kb)
lines(xfit, yfit, col = "blue", lwd = 2)
```




```r
# Kernel density plot
d <- density(files_info$size2)
plot(d, main = "File size distribution of the top 101 most imaged species")
polygon(d, col = "red", border = "blue")
```



Plotting the dimensions of all the images for inspection

```
# Reading all images paths in to a 'list'
images <-
    list.files(path = data_dir,
               pattern = ".jpg",
               full.names = TRUE)
str(images)
```

```
##  chr [1:10204] "/Users/nijram13/Google Drive/4. Biologie/Studie Biologie/Master Year 2/Internship CN
```

```
x <- image_read(images[1])
str(x)
```

```
## Class 'magick-image' <externalptr>
```

```
# Creates an empty list to be filled in the next function
lst2 <- c()

# Returns dimensions, depth, and channels for the images
for (image in images) {
    # x <- image_read(image)
    x <- load.image(image)
    x <- dim(x)
    x <- as.numeric(unlist(x))
    lst2 <- c(lst2, x)
}
```

```
## Error in fun(file): Unsupported file format. Please convert to jpeg/png/bmp or install image magick
```

```
# Converting to a Dataframe and get rid of images without 3 channels (RGB)
df <-
    data.frame(matrix(
        unlist(lst2),
        nrow = 10204,
        byrow = TRUE,
        ncol = 4
    ))
```

```
## Warning in matrix(unlist(lst2), nrow = 10204, byrow = TRUE, ncol = 4):
## data length [10136] is not a sub-multiple or multiple of the number of rows
## [10204]
```

```
colnames(df) <- c('hight', 'width', 'depth', 'channel')
rownames(df) <- images
df_channel_wrong <- subset(df, channel <= 2)
df_good <- subset(df, channel > 2)
df_good_melt <- melt(df_good[1:2])
```

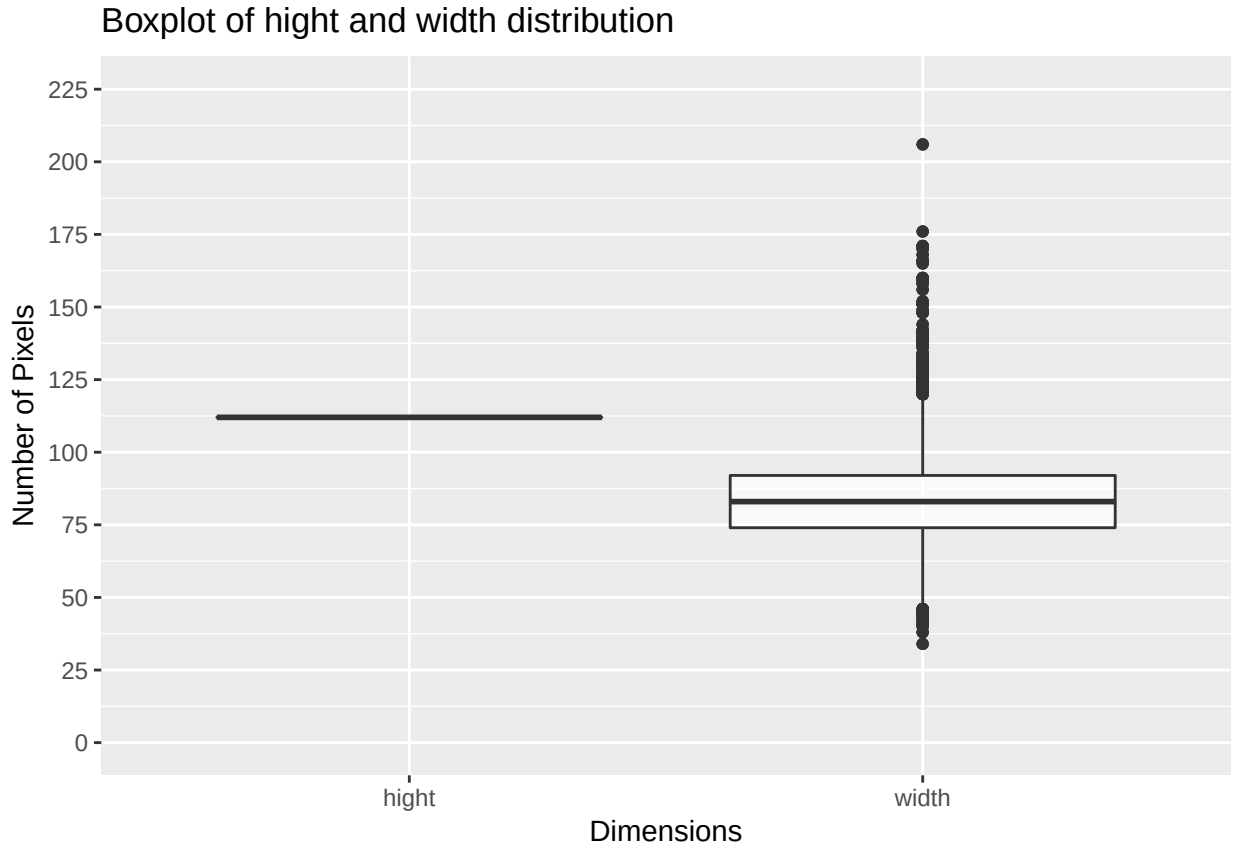
```
## No id variables; using all as measure variables
```

```
summary(df_good_melt)
```

```
##   variable         value
##  hight:10200   Min.   : 34.0
##  width:10200   1st Qu.: 83.0
##                Median :112.0
```

```
##                 Mean   : 98.1
##                 3rd Qu.:112.0
##                 Max.   :206.0
```

```r
# Boxplot of the hight and width distribution
g <- ggplot(df_good_melt)
g + geom_boxplot(aes(x = variable, y = value), alpha = 0.7) +
    scale_y_continuous(name = "Number of Pixels",
                       breaks = seq(0, 225, 25),
                       limits = c(0, 225)) +
    scale_x_discrete(name = "Dimensions") +
    ggtitle("Boxplot of hight and width distribution")
```




```r
# histogram of distribution of the width
h <- hist(
    df_good$width,
    breaks = 50,
    col = "red",
    xlab = "Number of pixels",
    main = "Pixel width distribution of the top 101 most imaged species"
)
# Plotting a normal distribution over the histogram
xfit <- seq(min(df_good$width), max(df_good$width), length = 40)
yfit <-
    dnorm(xfit,
          mean = mean(df_good$width),
          sd = sd(df_good$width))
yfit <- yfit * diff(h$mids[1:2]) * length(df_good$width)
```

```
lines(xfit, yfit, col = "blue", lwd = 2)
```