

# *Exploring and visualizing data with R*

Ecological and Epidemiological Modeling Madagascar, (E2M2)  
ValBio, Ranomafana  
Fianarantsoa, Madagascar  
03-14 January 2020

Hafaliana **Christian** Ranaivoson  
Virology Unit, Institut Pasteur de Madagascar  
Mention of Zoology and Animal Biodiversity  
Faculty of Sciences, University of Antananarivo

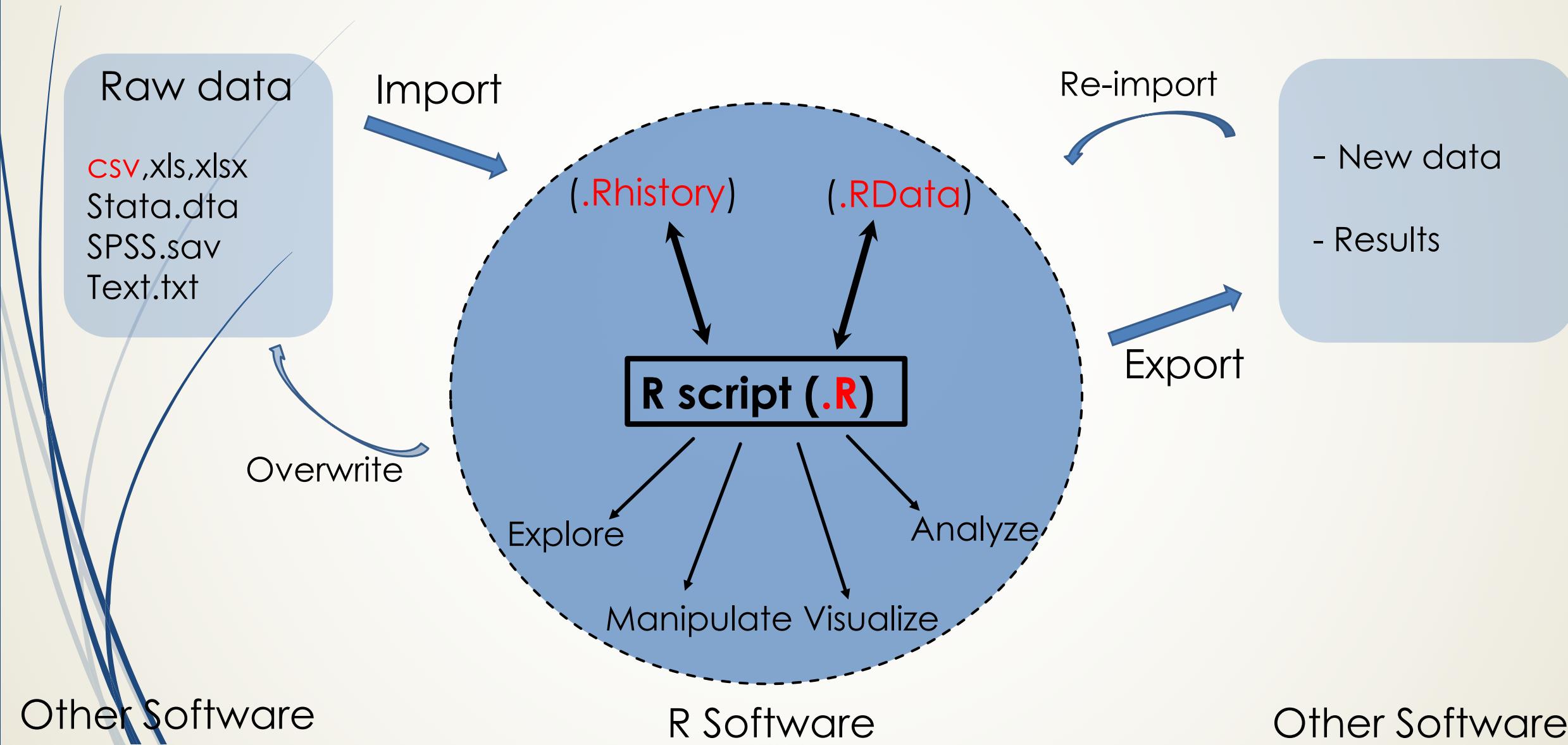
# *Exploring and visualizing data in R*

- ▶ R and RStudio software
- ▶ Importing data
- ▶ Exploring and cleaning data
- ▶ Visualizing data
- ▶ Tutorial

Database: « e2m2\_FB.csv »

# R software ( a statistical tool)

- ▶ It is free!
- ▶ Powerful analysis capability
- ▶ Versatile, flexible, open source



# An interface for R software

The screenshot displays the RStudio interface with several labeled components:

- Script**: The leftmost pane, highlighted with a red border, contains an R script named "e2m2\_2019.R". The code filters a dataset "e2m2" to find bats with forearm lengths less than 56 mm.
- Environment**: The top-right pane shows the global environment with a dataset "e2m2" containing 1200 observations and 8 variables.
- RData**: A purple-highlighted section in the Environment pane indicates the current workspace is "RData".
- Console**: The bottom-left pane shows the R console output, which includes a warning about omitted entries and a list of bat IDs (fb\_66 to fb\_401) filtered by forearm length.
- Browser**: The bottom-right pane lists files in the current directory (C:\Users\Chris\Desktop\2019\_01), including "2019\_01.Rproj", "E2M2.pptx", "e2m2\_2019.R", "e2m2\_FB.csv", and "E2M2-SID-7.pptx".

```

102 e2m2$Forearm
103
104 ## Now filter the content with [...] to show specific data, for example the Id of fruit bats
105 ## which have a forearm length smaller than 56 mm.
106
107 e2m2$id[e2m2$Forearm < 56]
108
109
110 ## You can get the count of bat which have a forearm smaller than 56 mm by using
111 ## length(...) function. which give 130 bats.
112
113
114 length(e2m2$id[e2m2$Forearm < 56])
115
116 ## You are told to print out bat Id that have forearm smaller than 40 mm and their number.
117 ## Try it
118 ## what happened?
119 ## No bats have a forearm shorter than 40mm
120 ## Now use range() function to check it.
121
111:1 # (Untitled) ⇣

```

```

[ reached getOption("max.print") -- omitted 200 entries ]
> e2m2$id[e2m2$Forearm < 56]
[1] "fb_66" "fb_67" "fb_68" "fb_69" "fb_70" "fb_71" "fb_72" "fb_73" "fb_74" "fb_75" "fb_76"
[12] "fb_77" "fb_78" "fb_79" "fb_80" "fb_81" "fb_82" "fb_83" "fb_84" "fb_85" "fb_86" "fb_87"
[23] "fb_88" "fb_89" "fb_90" "fb_91" "fb_92" "fb_93" "fb_94" "fb_95" "fb_96" "fb_97" "fb_98"
[34] "fb_99" "fb_100" "fb_166" "fb_167" "fb_168" "fb_169" "fb_170" "fb_171" "fb_172" "fb_173" "fb_174"
[45] "fb_175" "fb_176" "fb_177" "fb_178" "fb_179" "fb_180" "fb_181" "fb_182" "fb_183" "fb_184" "fb_185"
[56] "fb_186" "fb_187" "fb_188" "fb_189" "fb_190" "fb_191" "fb_192" "fb_193" "fb_194" "fb_195" "fb_196"
[67] "fb_197" "fb_198" "fb_199" "fb_200" "fb_201" "fb_202" "fb_203" "fb_204" "fb_205" "fb_206" "fb_207"
[78] "fb_208" "fb_209" "fb_210" "fb_211" "fb_212" "fb_213" "fb_214" "fb_215" "fb_216" "fb_217" "fb_218"
[89] "fb_219" "fb_220" "fb_286" "fb_287" "fb_288" "fb_289" "fb_290" "fb_291" "fb_292" "fb_293" "fb_294"
[100] "fb_295" "fb_296" "fb_297" "fb_298" "fb_299" "fb_300" "fb_301" "fb_302" "fb_303" "fb_304" "fb_305"
[111] "fb_306" "fb_307" "fb_308" "fb_309" "fb_310" "fb_311" "fb_312" "fb_313" "fb_314" "fb_315" "fb_386"
[122] "fb_387" "fb_388" "fb_389" "fb_390" "fb_391" "fb_392" "fb_393" "fb_394" "fb_401"
>

```

# Importing Data (loading data into R environment)

- Set working directory (Where to put all files?)

```
getwd()  
setwd("Folder path")
```

```
?getwd
```

- Import data (read the data source and pack it within RData)

```
e2m2_FB <- read.csv("e2m2_FB.csv", header=T, stringsAsFactors=F)  
  
View(e2m2_FB)
```

	A	B	C	D	E	F
1	Id	Sex	Forearm	Weight	Age	Date
2	fb_1	f	61.59	34.83	4.8	1/11/2015
3	fb_2	f	62.76	36.06	6.65	1/12/2015
4	fb_3	f	62.94	36.45	6.77	1/13/2015

	A	B	C	D	E	F
1	fb_1	f	61.59	34.83	4.8	1/11/2015
2	fb_2	f	62.76	36.06	6.65	1/12/2015
3	fb_3	f	62.94	36.45	6.77	1/13/2015

	A	B	C	D	E	F
1	fb_1	f	61.59	34.83	4.8	1/11/2015
2	fb_2	f	62.76	36.06	6.65	1/12/2015
3	fb_3	f	62.94	36.45	6.77	1/13/2015

# Exploring and cleaning Data (look at the dataset)

## Data overview

Data frame=  
e2m2\_FBF

Variables= Column (names)

Cases= Row (length)

Id	Sex	Forearm	Weight	Age	Date
fb_1	f	61.59	34.83	4.8	1/11/2015
fb_2	f	62.76	36.06	6.65	1/12/2015
fb_3	f	62.94	36.45	6.77	1/13/2015

Value=  
Contents

```
> dim(e2m2_FBF)  
[1] 100  6
```

How big is the data frame?

```
> names(e2m2_FBF)  
[1] "Id"    "Sex"   "Forearm" "Weight" "Age"   "Date"
```

What are the variables?

# Exploring and cleaning Data (Dive into the dataset)

## Accessing dataset contents (From Outside to Inside!)

```
> e2m2_FB$id  
[1] "fb_1"  "fb_2"  "fb_3"  "fb_4"  
[12] "fb_12" "fb_13" "fb_14" ...  
[100] "fb_100"
```

Data frame > **Variables** > **Contents**

```
> e2m2_FB$id[e2m2_FB$Forearm < 56]  
[1] "fb_64" "fb_65"
```

Dataset name \$ **Variable name**

Filter Contents [...]

Get the Bat Id with Weight > 75

```
> length(e2m2_FB$id[e2m2_FB$Forearm < 56])  
[1] 2
```

Data frame name \$ **Variable name** [Filter]

Get the count with **length(...)**

# Exploring and cleaning Data (look at the dataset structure)

## Variable types and error

**str(...)**

```
> str(e2m2_FB)

$ Id      : chr "fb_1" "fb_2" "fb_3" "fb_4"...
$ Sex     : chr "f" "f" "f" "F" "f" ...
$ Weight  : num 34.8 36.1 36.5 36.6 38.9 ...
$ Age     : chr "one" "6.65" "6.77" "seven" ...
$ Date    : chr "1/11/2015" "1/12/2015"...
```

```
> as.factor(e2m2_FB$Sex)
Levels: f F f m
> as.numeric(e2m2_FB$Age)
Warning message:
NAs introduced by coercion
> as.Date(e2m2_FB$Date,"%m/%d/%Y")
```

**Categorical:**

**Continuous:**

**Time:**

**Binary:**

**Missing Value:**

Factor (n levels)

Numeric (Range)

Date (Range)

logic (T,F)

NA

**as.factor(...)**

**as.Date(...)**

**as.numeric(...)**

"%Y-%m-%d"

Needed format

Value error ->

Missing error ->

-> Re-format

Correct value

Handle NA values

# Exploring and cleaning Data (clean the dataset)

## Correcting Values

(Wrong value <- Right Value)

```
> e2m2_FB$Age[e2m2$Age=="one"] <- "1"  
e2m2_FB$Sex[e2m2$Sex=="F"] <- "f"  
> e2m2_FB$Sex[e2m2$Sex=="f "] <- "f"
```

```
> as.factor(e2m2_FB$Sex)  
Levels: f m  
> as.numeric(e2m2_FB$Age)  
[1] 4.80 6.65 6.77 7.00
```

## Save the format to the variable

```
> e2m2_FBF$Sex <- as.factor(e2m2_FBF$Sex)  
> e2m2_FBF$Age <- as.numeric(e2m2_FBF$Age)  
> e2m2_FBF$Date <- as.Date(e2m2_FBF$Date,"%m/%d/%Y")
```

```
> str(e2m2_FBF)  
$ Id : chr "fb_1" "fb_2" "fb_3" "fb_4" ...  
$ Sex : Factor w/ 2 levels "f","m": 1 1 1 1 2  
$ Age : num 4.8 6.65 6.77 7 8.89 ...  
$ Date : Date, format: "2015-01-11"
```

```
> e2m2_FBF$NewVar <- as.factor(e2m2_FBF$Sex)  
> e2m2_FBF$NewVar <- e2m2_FBF$Forearm/2
```

Or create new variable

# Visualizing Data (Play with data)

## Install and Load Library

```
Install.packages("...")  
Installed.packages()
```

```
> library(dplyr)  
> require(ggplot2)
```

## Data summarizing ("dplyr")

```
> fb_male <- filter(e2m2_FB, Sex=="m")  
> range(fb_male$Forearm)  
> mean(fb_male$Forearm)  
> sd(fb_male$Forearm)
```

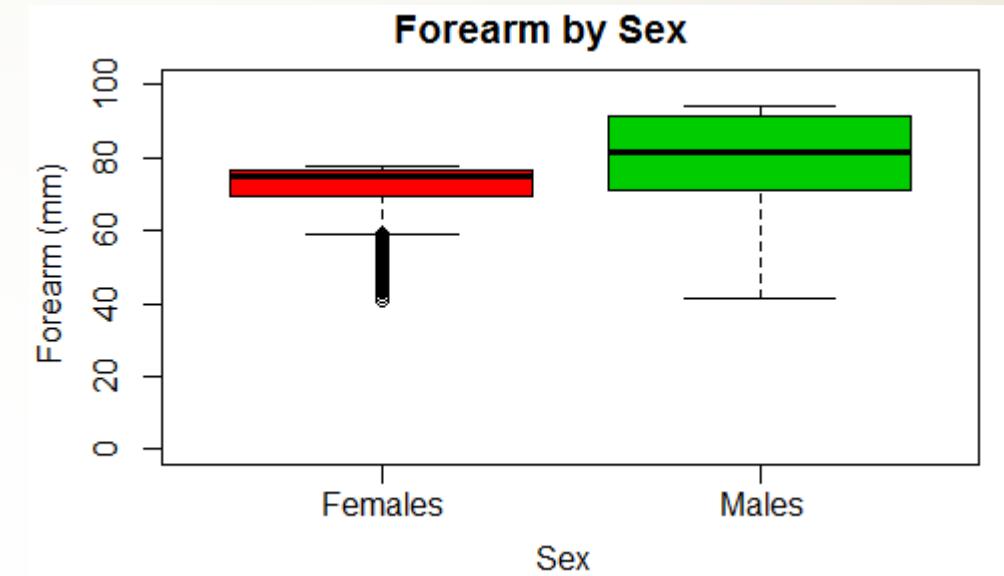
Sex	mean_forearm	sd_forearm	nbr
(fctr)	(dbl)	(dbl)	(int)
1 f	59.6040	11.90278	60
2 m	60.9985	14.12073	40

```
gp_fb <- group_by(e2m2_FB, Sex)  
gp_fb_stat <- summarise(gp_fb,  
  mean_forearm=mean(Forearm,na.rm=T),  
  sd_forearm=sd(Forearm,na.rm=T),  
  nbr=n())
```

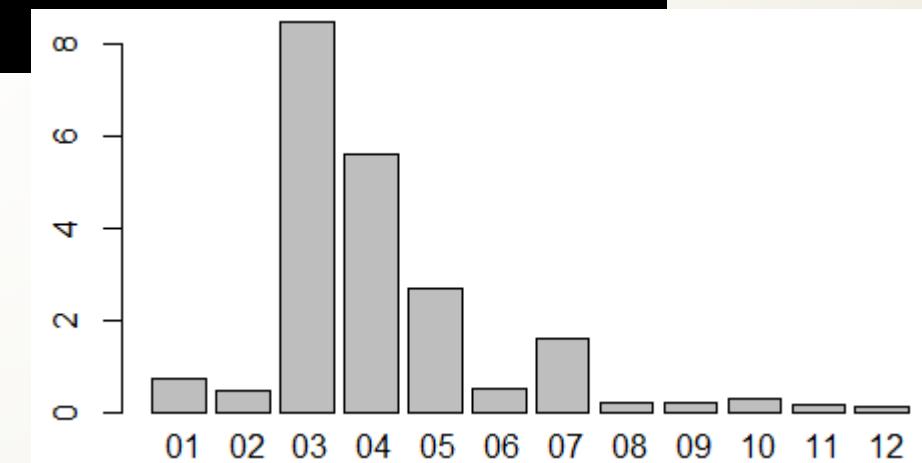
# Visualizing Data (Present de data)

## R base graphical function

```
boxplot(Forearm~Sex,  
        names=c("Females","Males"),  
        col=c(2:3),  
        main="Forearm by Sex",  
        xlab="Sex",  
        ylab="Forearm (mm)",  
        ylim=c(0,100))
```



```
PLoad <- tapply(e2m2$ParLoad,factor(format(e2m2$date,"%m")),mean)  
barplot(PLoad)
```



# Visualizing Data (Present de data)

## R plot() function

```
plot(e2m2$Forearm~e2m2$Age,  
     main = "Forearm/Age",  
     ylab ="Forearm (mm)", xlab = "Age (year)",  
     col=Sex)
```

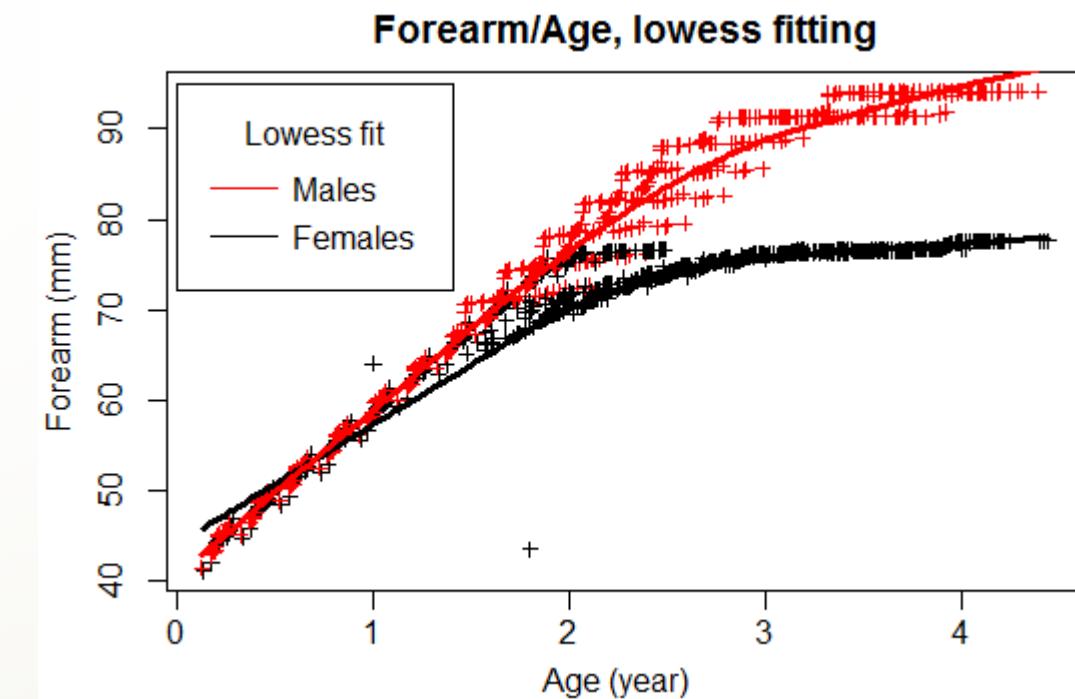
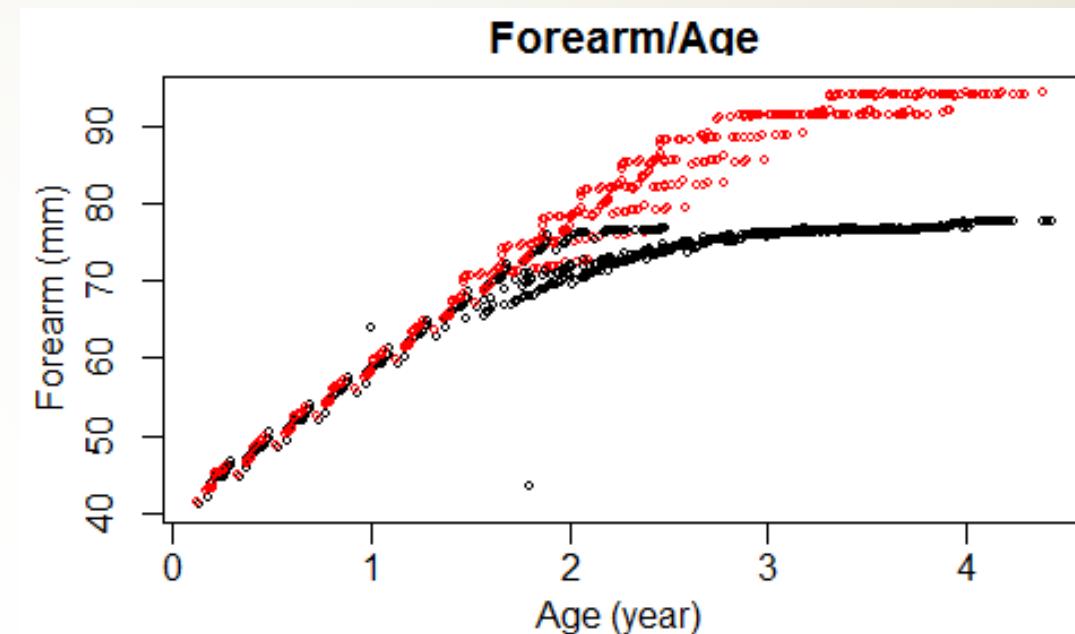
```
fitfem <- lowess(datfem$Forearm~datfem$Age)  
fitmal <- lowess(datmal$Forearm~datmal$Age)
```

```
plot(e2m2$Forearm~e2m2$Age,  
     main="Forearm/Age, lowess fitting",  
     xlab="Age (year)",ylab="Forearm (mm)",  
     type="p",pch=3,cex=0.7,  
     col=e2m2$Sex)
```

```
lines(fitfem,col="black",lwd=3)
```

```
lines(fitmal,col="red",lwd=3)
```

```
legend(x=0, y=95,legend=c("Males","Females"),  
       col=c("red","black"),title="Lowess fit",  
       lty=1,x.intersp = .5,y.intersp = .8)
```



# Visualizing Data (Present de data)

## Data Plotting with (ggplot2)

Base plot

```
ggplot(data,aes(x,y))
```

+ geom

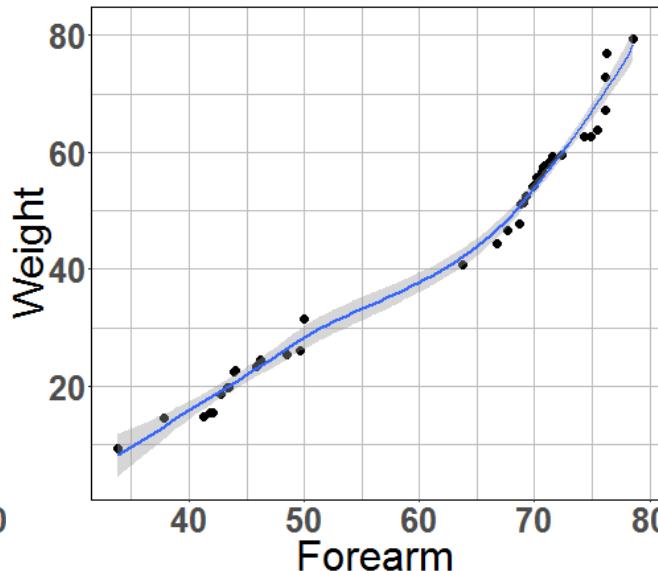
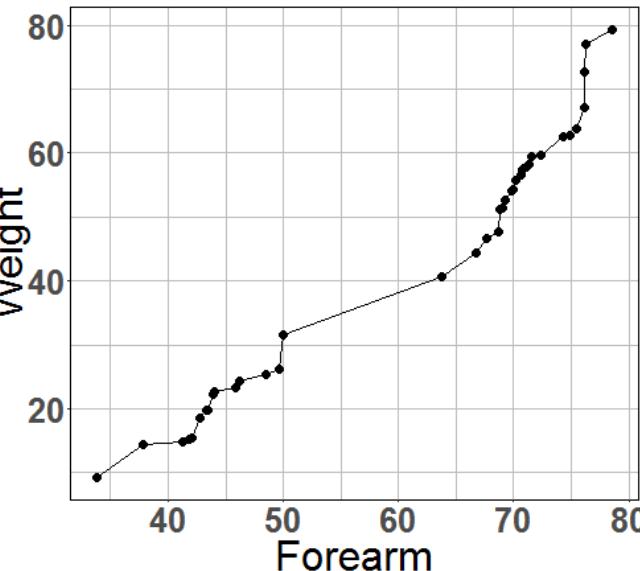
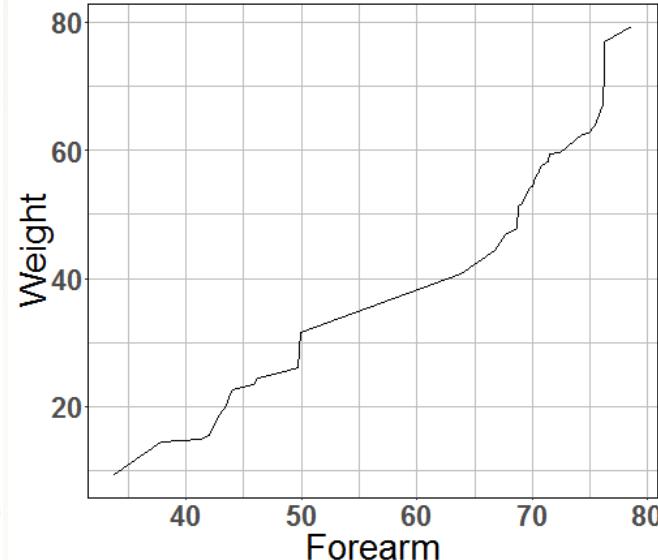
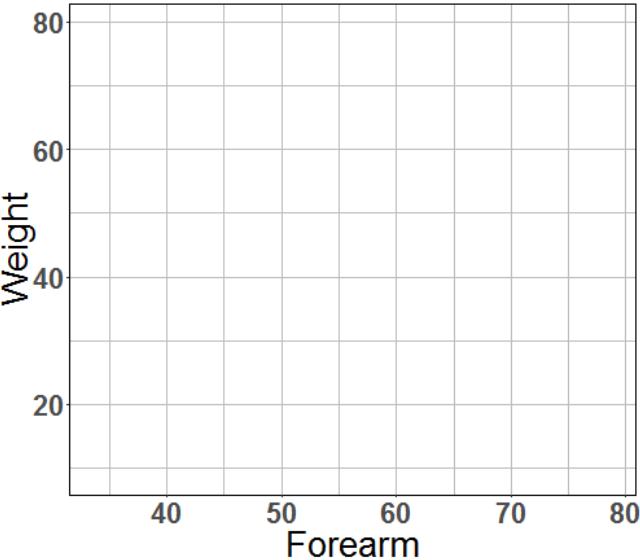
- geom\_line()
- geom\_point()
- geom\_boxplot()
- geom\_bar()

```
ggplot(fb_male,aes(Forearm,Weight))
```

```
+ geom_line()
```

```
+ geom_point()
```

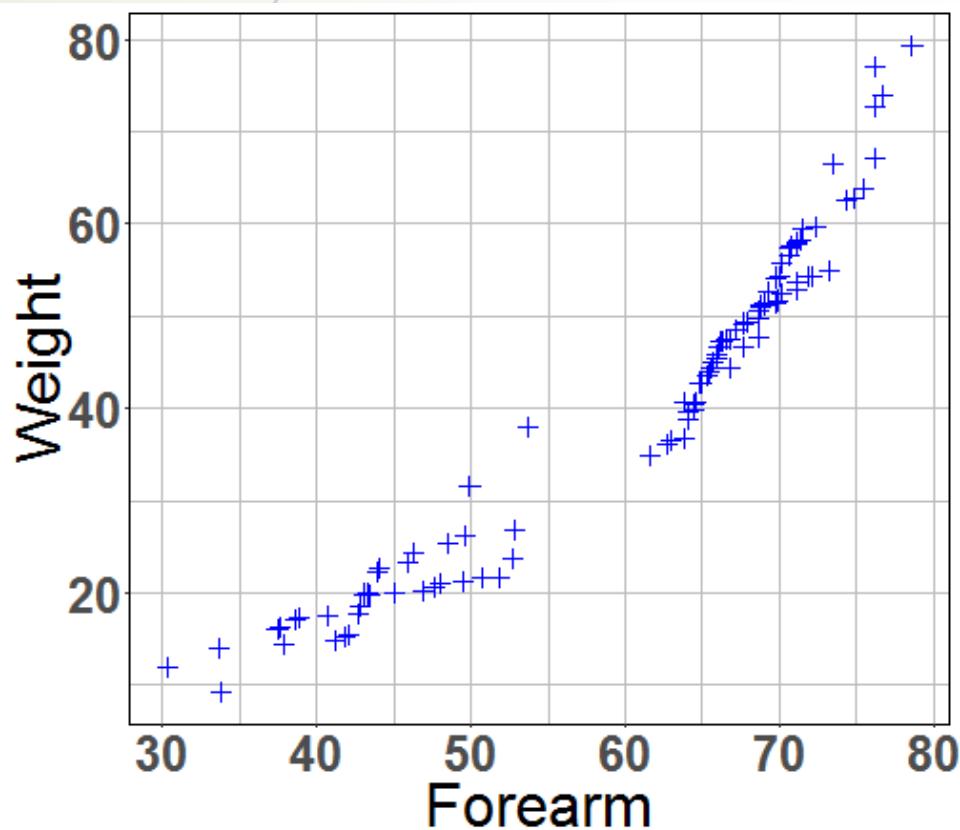
```
+ geom_smooth()
```



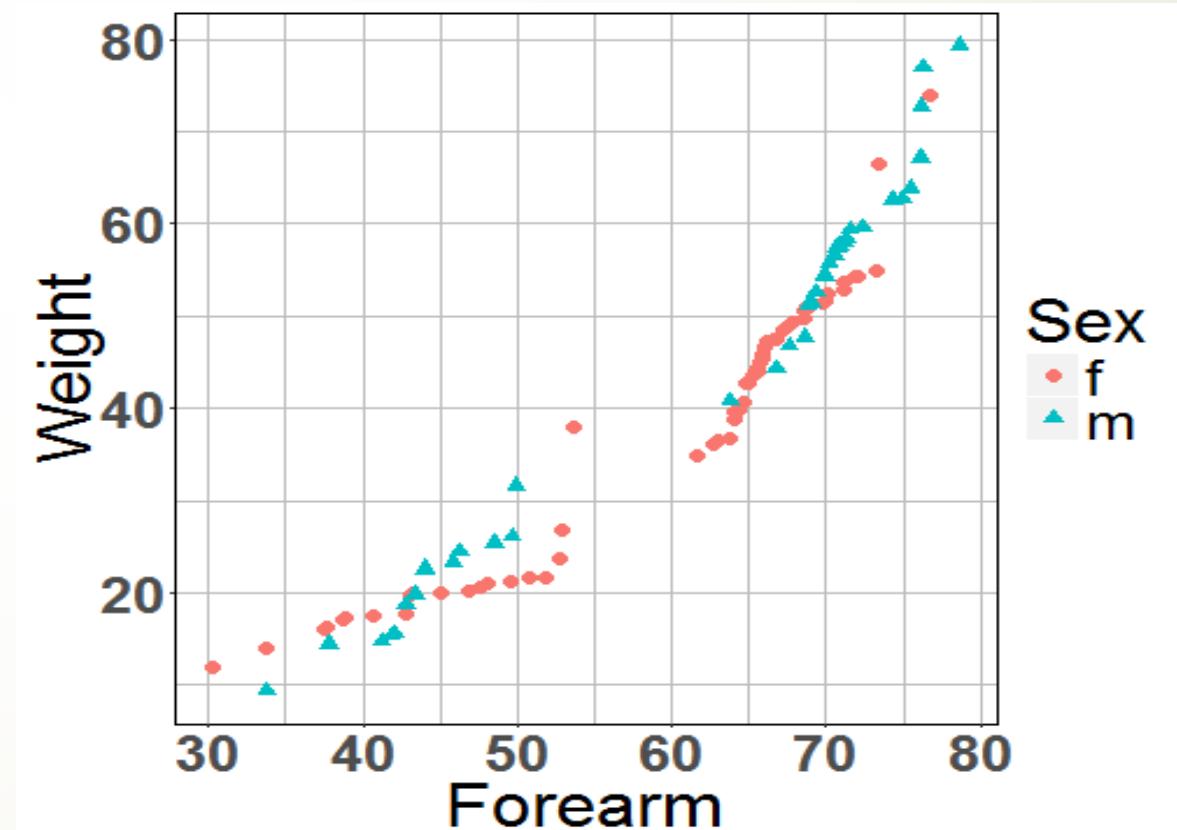
# Visualizing Data (Present de data)

## Mapping aesthetic vs fixed value

```
ggplot(e2m2_FB,aes(Forearm,Weight)) +  
  geom_point(color="blue", shape=3)
```



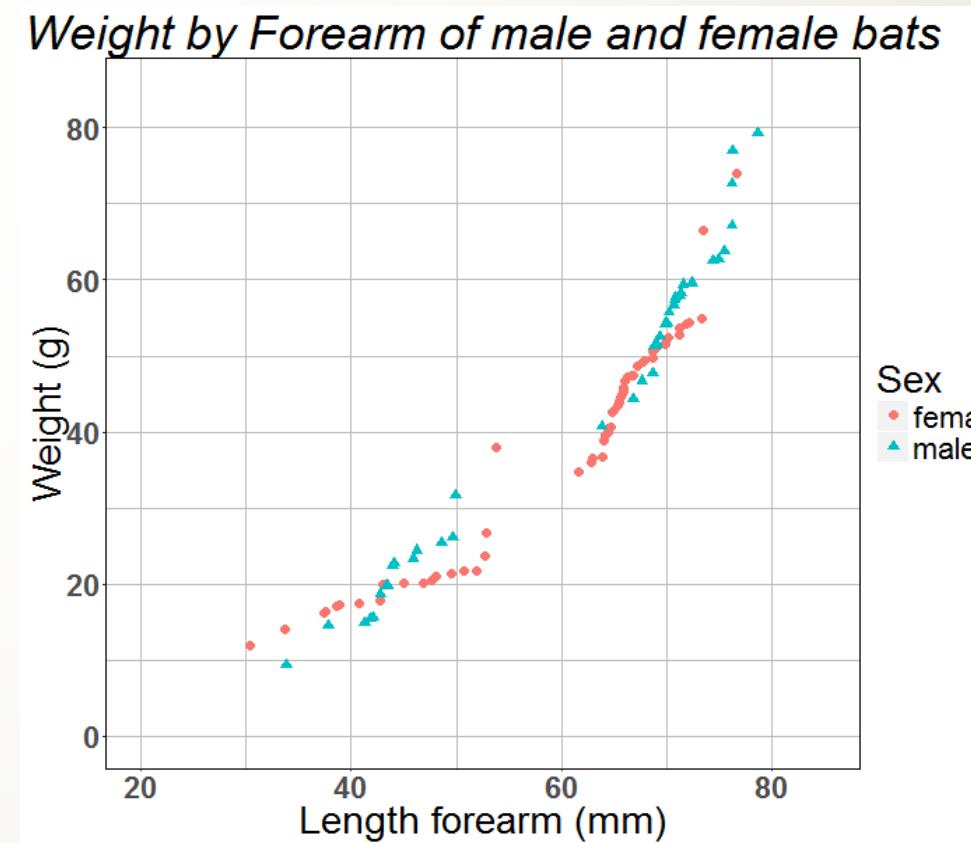
```
ggplot(e2m2_FB,aes(Forearm,Weight)) +  
  geom_point(aes(color=Sex, shape=Sex))
```



# Visualizing Data (Presentar de datos)

## Polish the plot

```
ggtitle("Weight by Forearm of male and female") +  
  scale_x_continuous(name="Length forearm (mm)",  
                     limits=c(20,85)) +  
  scale_y_continuous(name="Weight (g)",  
                     limits=c(0,85)) +  
  scale_color_discrete(name="Sex",  
                       breaks=c("f","m"),  
                       label=c("female","male")) +  
  scale_shape_discrete(name="Sex",  
                       breaks=c("f","m"),  
                       label=c("female","male"))
```



# Conclusion

## R software:

- Powerful data management
- Simple syntax
- Large graphic vocabularies
- Packages to fit needs