[DS@USF](mailto:DS@USF) Lab 1

Priyanka Thakur

1/23/2020

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(rio)  
library(Hmisc)

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:rpart':  
##   
## solder

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## combine, src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, round.POSIXt, trunc.POSIXt, units

library(ggpubr)  
my\_data=data.frame(iris)  
colnames(my\_data)=tolower(make.names(colnames(my\_data)))  
attach(my\_data)

## Understanding the dataset

levels(my\_data$species)

## [1] "setosa" "versicolor" "virginica"

str(my\_data)

## 'data.frame': 150 obs. of 5 variables:  
## $ sepal.length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ sepal.width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ petal.length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ petal.width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

head(my\_data)

## sepal.length sepal.width petal.length petal.width species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

summary(my\_data)

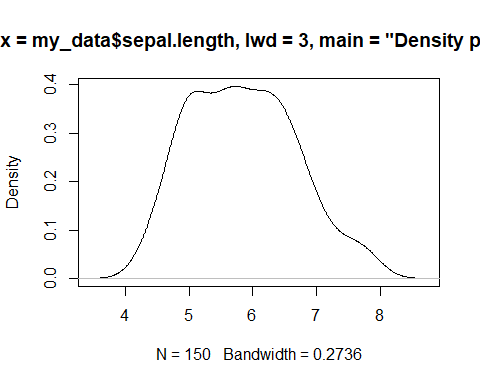
## sepal.length sepal.width petal.length petal.width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

dim(my\_data)

## [1] 150 5

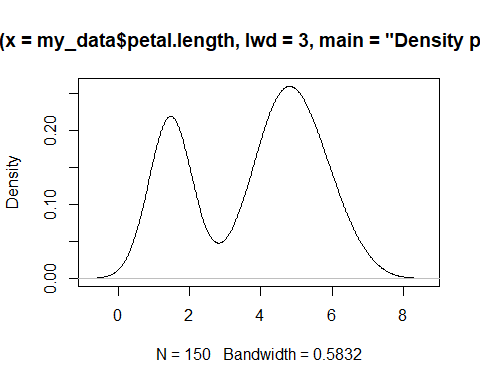
plot(density(my\_data$sepal.length,lwd=3,main="Density plot of sepal length"))

## Warning: In density.default(my\_data$sepal.length, lwd = 3, main = "Density plot of sepal length") :  
## extra arguments 'lwd', 'main' will be disregarded



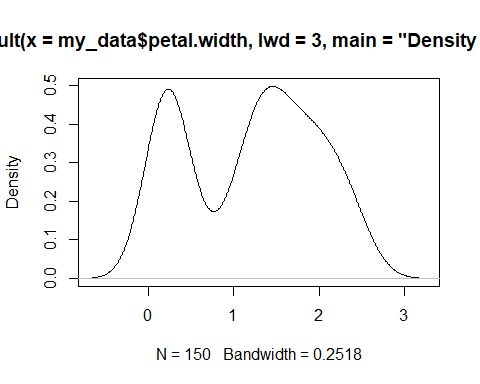
plot(density(my\_data$petal.length, lwd=3, main = "Density plot of petal length"))

## Warning: In density.default(my\_data$petal.length, lwd = 3, main = "Density plot of petal length") :  
## extra arguments 'lwd', 'main' will be disregarded



plot(density(my\_data$petal.width, lwd=3, main = "Density plot of Species"))

## Warning: In density.default(my\_data$petal.width, lwd = 3, main = "Density plot of Species") :  
## extra arguments 'lwd', 'main' will be disregarded



setosa = filter(my\_data, species == "setosa")  
virginica = filter(my\_data, species == "virginica")  
versicolor = filter(my\_data, species == "versicolor")  
count(setosa)

## # A tibble: 1 × 1  
## n  
## <int>  
## 1 50

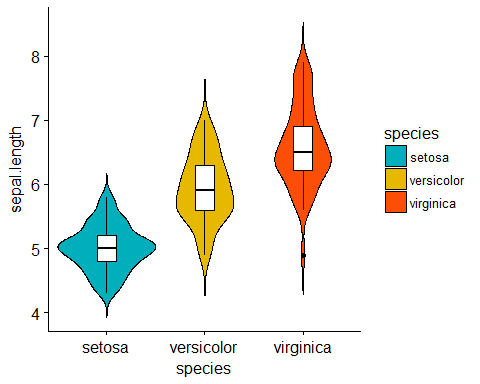
select(my\_data, sepal.length, sepal.width)%>%  
 filter(species == "virginica" & sepal.length <6 & sepal.width <=2.7 )

## sepal.length sepal.width  
## 1 5.8 2.7  
## 2 4.9 2.5  
## 3 5.7 2.5  
## 4 5.8 2.7

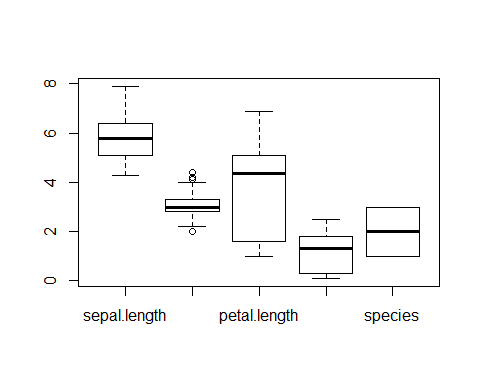
my\_data %>%  
 group\_by(species) %>%  
 summarise (  
 mean\_petal\_length = mean(petal.length),  
 mean\_sepal\_length = mean(sepal.length),  
 median\_sepal\_width = median(sepal.width),  
 median\_pedal\_width = median(petal.width))

## # A tibble: 3 × 5  
## species mean\_petal\_length mean\_sepal\_length median\_sepal\_width  
## <fctr> <dbl> <dbl> <dbl>  
## 1 setosa 1.462 5.006 3.4  
## 2 versicolor 4.260 5.936 2.8  
## 3 virginica 5.552 6.588 3.0  
## # ... with 1 more variables: median\_pedal\_width <dbl>

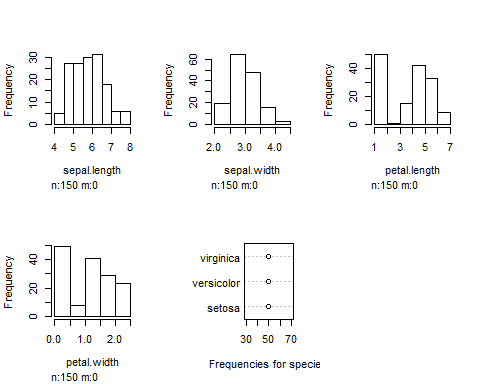
plot\_species = ggviolin(my\_data, x="species" , y="sepal.length", fill = "species", palette = c("#00AFBB", "#E7B800", "#FC4E07"), add = "boxplot" , add.params = list(fill='white'))  
plot\_species



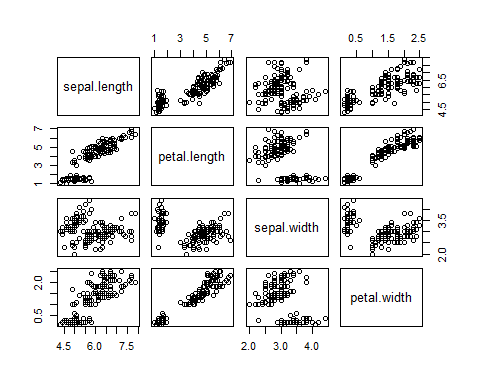
boxplot(my\_data)



hist(my\_data)



some.mydata = subset(my\_data, select = c("sepal.length", "petal.length", "sepal.width","petal.width"))  
plot(some.mydata)



cor(some.mydata)

## sepal.length petal.length sepal.width petal.width  
## sepal.length 1.0000000 0.8717538 -0.1175698 0.8179411  
## petal.length 0.8717538 1.0000000 -0.4284401 0.9628654  
## sepal.width -0.1175698 -0.4284401 1.0000000 -0.3661259  
## petal.width 0.8179411 0.9628654 -0.3661259 1.0000000

xx = rcorr(as.matrix(some.mydata))  
xx

## sepal.length petal.length sepal.width petal.width  
## sepal.length 1.00 0.87 -0.12 0.82  
## petal.length 0.87 1.00 -0.43 0.96  
## sepal.width -0.12 -0.43 1.00 -0.37  
## petal.width 0.82 0.96 -0.37 1.00  
##   
## n= 150   
##   
##   
## P  
## sepal.length petal.length sepal.width petal.width  
## sepal.length 0.0000 0.1519 0.0000   
## petal.length 0.0000 0.0000 0.0000   
## sepal.width 0.1519 0.0000 0.0000   
## petal.width 0.0000 0.0000 0.0000