```r
# --- ------------------------------------------------------------------------
# --- Descriptive Statistics in R with the IRIS dataset
# ---
# --- V1 March   2021, D.Benninger - initial version for BINA-FS21
# --- V2 March   2021, dbe --- some typos corrected
# --- V3 January 2022, dbe --- some minor corrections (qqplot) for CAS BIA12
# ---
# --- Libraries: ggplot2, car
# ---
# --- Data:      iris               >> standard R dataset
# ---
# --- Links
# --- https://www.r-bloggers.com/2020/01/descriptive-statistics-in-r/
# ---
# ---
# --- ------------------------------------------------------------------------

# PACKAGES installieren, falls nicht vorhanden
if(!"ggplot2" %in% rownames(installed.packages())) install.packages("ggplot2")
if(!"car" %in% rownames(installed.packages())) install.packages("car")

# Packages laden
library("ggplot2")
library("car")

# Set WORKING Directory
setwd(choose.dir())


# --- DATA OVERVIEW
# load the iris dataset and renamed it dat
dat <- iris

# --- PREVIEW of the dataset and its structure
# first observations
head(dat)

# ---    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

# structure of dataset
str(dat)

# ---  '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 ...


# --- MINIMUM and MAXIMUM
# -----------------------
min(dat$Sepal.Length)
# ---  [1] 4.3

max(dat$Sepal.Length)
# ---  [1] 7.9

# alternative the range() function
rng <- range(dat$Sepal.Length)
rng
# ---  [1] 4.3 7.9

rng[1] # MIN -- rng = name of the object specified above
# ---  [1] 4.3
rng[2] # MAX
```

```r
# ---  [1] 7.9

# --- RANGE
max(dat$Sepal.Length) - min(dat$Sepal.Length)
# ---  [1] 3.6



# --- MEAN, MEDIAN and QUANTILE
# -----------------------------
mean(dat$Sepal.Length)
# ---  [1] 5.843333

median(dat$Sepal.Length)
# ---  [1] 5.8

quantile(dat$Sepal.Length, 0.5)
# ---  50%
# ---  5.8

# --- 1st and 3rd QUARTILE
# ------------------------
quantile(dat$Sepal.Length, 0.25) # first quartile
# ---  25%
# ---  5.1
quantile(dat$Sepal.Length, 0.75) # third quartile
# ---  75%
# ---  6.4

# --- other QUANTILEs
quantile(dat$Sepal.Length, 0.4) # 4th decile
# ---  40%
# ---  5.6
quantile(dat$Sepal.Length, 0.98) # 98th percentile
# ---  98%
# ---  7.7

# --- Interquartile Range
IQR(dat$Sepal.Length)

# or alternatively with the quantile function
quantile(dat$Sepal.Length, 0.75) - quantile(dat$Sepal.Length, 0.25)
# ---  75%
# ---  1.3



# --- Standard DEVIATION and VARIANCE
# -----------------------------------
sd(dat$Sepal.Length) # standard deviation
# ---  [1] 0.8280661
var(dat$Sepal.Length) # variance
# ---  [1] 0.6856935

# Tipp: compute the standard deviation (or variance) of multiple variables at the
same time
# use lapply() with the appropriate statistics as second argument:
lapply(dat[, 1:4], sd)
# ---  $Sepal.Length
# ---  [1] 0.8280661
# ---
# ---  $Sepal.Width
# ---  [1] 0.4358663
# ---
# ---  $Petal.Length
# ---  [1] 1.765298
# ---
# ---  $Petal.Width
# ---  [1] 0.7622377

#
-------------------------------------------------------------------------------
----
```

```r
# --- SUMMARY
#
# ------------------------------------------------------------------------------
summary(dat)
# ---    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


# Tipp: if you need these descriptive statistics by group use the by() function:
by(dat, dat$Species, summary)
# ---  dat$Species: setosa
# ---    Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
# ---   Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
# ---   1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
# ---   Median :5.000   Median :3.400   Median :1.500   Median :0.200
# ---   Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
# ---   3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
# ---   Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
# ---         Species
# ---   setosa    :50
# ---   versicolor: 0
# ---   virginica : 0
# ---
# ---
# ---
# ---  ------------------------------------------------------------
# ---  dat$Species: versicolor
# ---    Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species
# ---   Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0
# ---   1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50
# ---   Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0
# ---   Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
# ---   3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
# ---   Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
# ---  ------------------------------------------------------------
# ---  dat$Species: virginica
# ---    Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
# ---   Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400
# ---   1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800
# ---   Median :6.500   Median :3.000   Median :5.550   Median :2.000
# ---   Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
# ---   3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
# ---   Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
# ---         Species
# ---   setosa    : 0
# ---   versicolor: 0
# ---   virginica :50
# ---
# ---
# ---



# --- HISTOGRAM
# -------------
hist(dat$Sepal.Length)

# with ggplot2
ggplot(dat) +
  aes(x = Sepal.Length) +
  geom_histogram()

# by default, the number of bins is 30.
```

```r
# you can change this value with geom_histogram(bins = 12) for instance.
ggplot(dat) +
  aes(x = Sepal.Length) +
  geom_histogram(bins = 12)


# --- BOXPLOT
# -----------
boxplot(dat$Sepal.Length)

# Boxplots are even more informative when presented side-by-side
# for comparing and contrasting distributions from two or more groups.
# For instance, we compare the length of the sepal across the different species:
boxplot(dat$Sepal.Length ~ dat$Species)

# with ggplot2
ggplot(dat) +
  aes(x = Species, y = Sepal.Length) +
  geom_boxplot()


# --- SCATTERPLOT
# ---------------
plot(dat$Sepal.Length, dat$Petal.Length)

# with ggplot2
ggplot(dat) +
  aes(x = Sepal.Length, y = Petal.Length) +
  geom_point()

# and (colored) with differentiating the points according to a factor (i.e. species)
ggplot(dat) +
  aes(x = Sepal.Length, y = Petal.Length, colour = Species) +
  geom_point() +
  scale_color_hue()


# --- DENSITY plot
# ----------------
plot(density(dat$Sepal.Length))



# ----------------------------------------
# --- Addtional Graphical Representations
# ----------------------------------------
plot(iris)
ggplot(iris, aes(Petal.Length, Petal.Width, color = Species)) + geom_point()

# --- Attribute Statistical Key Values
# --- Explore Individual Variables
boxplot(iris$Sepal.Length, main="Sepal.Length")
boxplot(iris$Petal.Length, main="Petal.Length")

par(mfrow=c(1,2))
boxplot(iris$Petal.Length, iris$Petal.Width, main="Petal Measures")
boxplot(iris$Sepal.Length, iris$Sepal.Width, main="Sepal Measures")

par(mfrow=c(1,1))
boxplot(iris$Sepal.Length, iris$Sepal.Width, iris$Petal.Length, iris$Petal.Width,
main="IRIS Measures", ylab ="length/width in cm")
boxplot(Sepal.Length ~ Species, data=iris, xlab="Species", ylab="Sepal.Length")

# --- Histograms and Density plots
hist(iris$Sepal.Length, breaks=10)
hist(iris$Sepal.Width, breaks=10)
hist(iris$Sepal.Length+iris$Sepal.Width)

plot(density(iris$Petal.Length))
plot(density(iris$Petal.Width))

# --- SCATTERPLOT
# --- Explore Multiple Variables
```

```r
plot(iris$Sepal.Length,iris$Sepal.Width, xlim=c(0,10), ylim=c(0,10), main="Iris Data")

plot(iris$Sepal.Length, iris$Sepal.Width, col=iris$Species, pch=as.numeric(iris$Species), main="Iris Data Scatterplot")

library(scatterplot3d)
scatterplot3d(iris$Petal.Width,iris$Sepal.Length,iris$Sepal.Width)
scatterplot3d(iris$Petal.Width,iris$Sepal.Length,iris$Sepal.Width, pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])

# --- Smooth SCATTERPLOT
smoothScatter(iris$Sepal.Length,iris$Sepal.Width)

# --- HEAT MAP
distM <- as.matrix(dist(iris[,1:4]), method="euclidean")
heatmap(distM)

# --- Other Visualization (cross reference, species)
pairs(iris)

quickplot(Sepal.Length, Sepal.Width, data=iris, facets= Species ~.,color = Species)

```