Introduction to R

Anders K. Krabberød (University of Oslo)

Based on slides by Ramiro Logares (ICM-CSIC, Barcelona)

AB332 – Autumn 2024

History of R

- Originated from S
 - Statistical programming language developed by John Chambers at Bell Labs in the 70s
 - Developed at the same times as Unix
 - Closed source
- First version of R: developed by Robert Gentleman and Ross Ihaka in the mid-1990s
 - Aimed for better statistics software in their Mac teaching labs
 - Open Source alternative
 - R 1.0.0 released in 2000
- Current version 4.2.2
- Developers: international R-core developing team



John Chambers

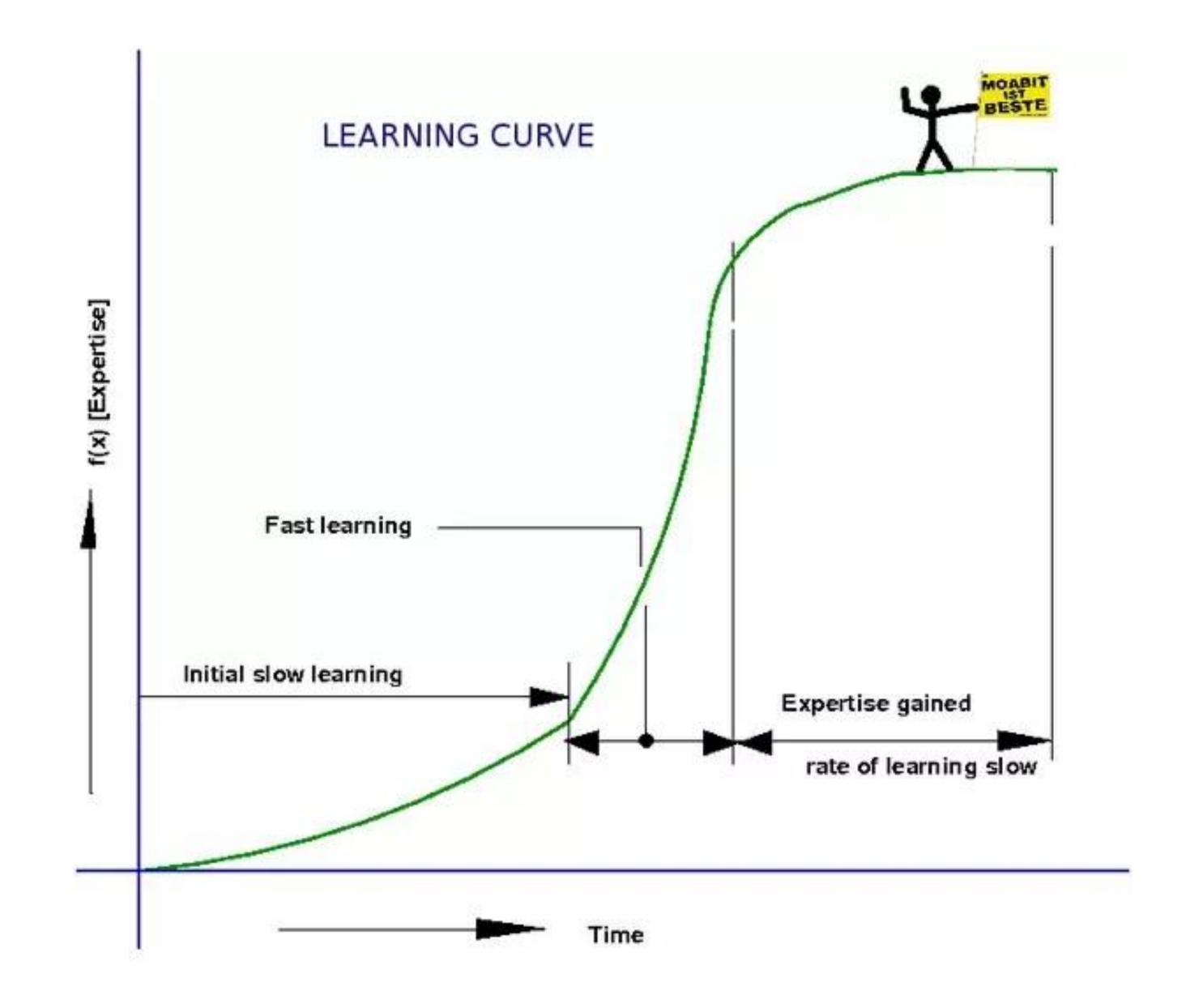




Ross Ihaka

Why learn R?

- Language and environment for statistical computing and graphics
- Open Source (free)
- Cross-platform compatibility
- Community supported
- Great flexibility to do what you want
- Many packages available: ecology, metabarcoding, networks
- Amazing publication quality graphs



Installing R

https://cran.r-project.org/

Download and Install R

Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- · Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2023-06-16, Beagle Scouts) R-4.3.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are <u>available here</u>. Please read about <u>new features and bug fixes</u> before filing corresponding feature requests or bug reports.
- Source code of older versions of R is <u>available here</u>.
- Contributed extension <u>packages</u>

Questions About R

If you have questions about R like how to download and install the software, or what the license terms are, please read our <u>answers to frequently</u> asked <u>questions</u> before you send an email.

Installing R-Studio

- An Integrated Development Environment (IDE): *R-Studio*
 - Set of tools designed to help and be more productive with R
 - Includes a console and syntaxhighlighting editor that supports code execution
 - Can have several open sessions (aka. Projects)
 - Includes a Unix terminal
 - Can run other programming languages (python, bash)

DOWNLOAD

RStudio Desktop

Used by millions of people weekly, the RStudio integrated development environment (IDE) is a set of tools built to help you be more productive with R and Python.

Don't want to download or install anything? Get started with RStudio on <u>Posit Cloud for free</u>. If you're a professional data scientist looking to download RStudio and also need common enterprise features, don't hesitate to <u>book a call with us</u>.

1: Install R

RStudio requires R 3.3.0+. Choose a version of R that matches your computer's operating system.

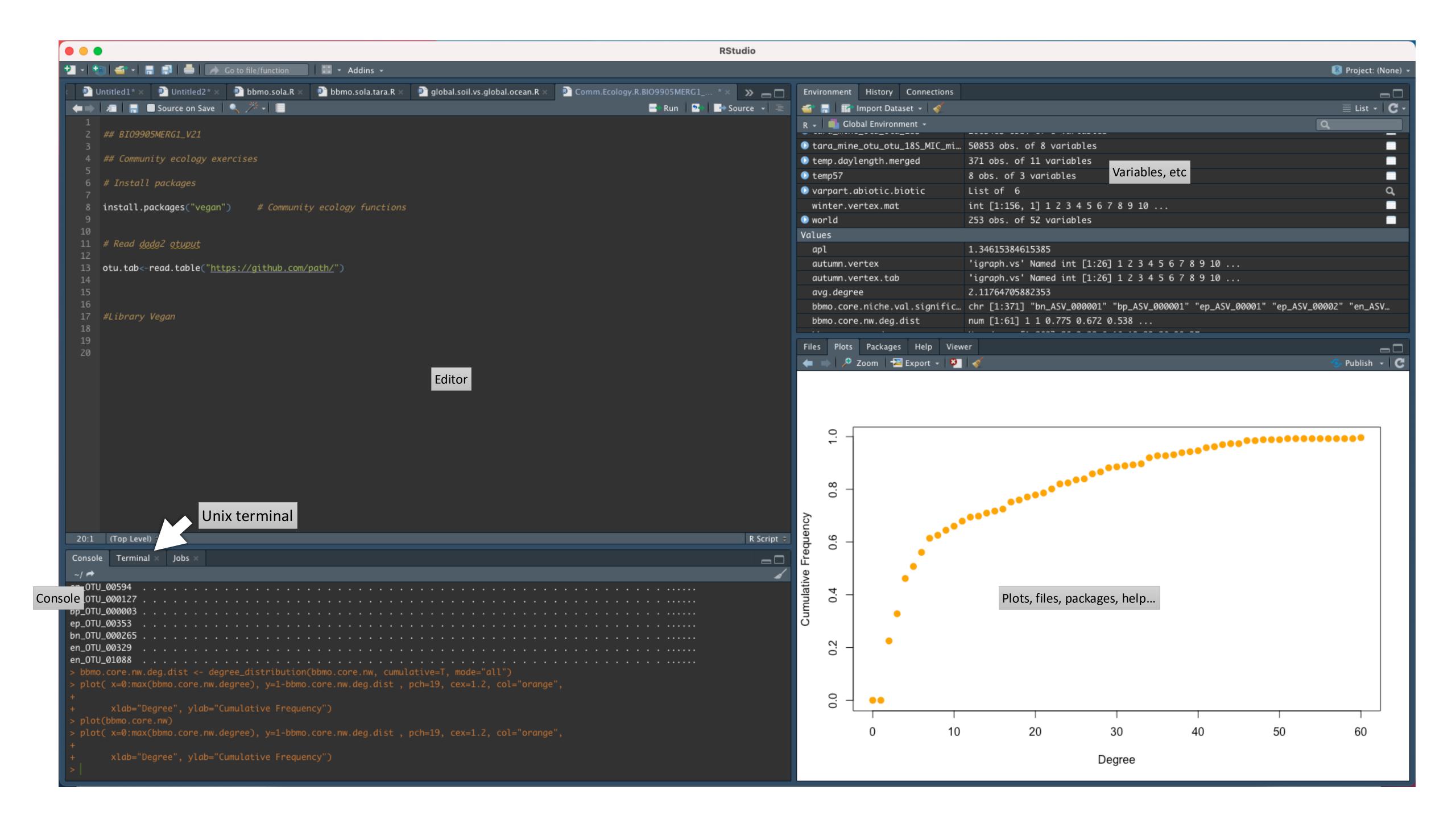
DOWNLOAD AND INSTALL R

2: Install RStudio

DOWNLOAD RSTUDIO DESKTOP FOR MACOS 11+

This version of RStudio is only supported on macOS 11 and higher. For earlier macOS environments, please <u>download</u> a previous version.

Size: 385.88 MB | SHA-256: 25A2CC51 | Version: 2023.09.1+494 Released: 2023-10-17



Presentation format

- The following slides shows some typical R commands and the important data structures
- The idea is that you become familiar with reading code as you will do when working with R-Studio
- After this introduction, you should be able to follow other sections of the course using R

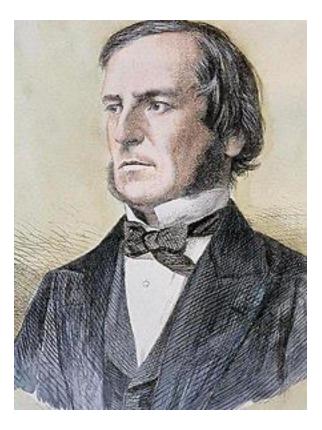


Intro to R

Basic operations.

```
6+6
6-6
6*6
6/6
log(6)
log2(6)
log10(6)
log(6,10)
log(6,3)
6^6
sin(pi/2)
cos(pi/2)
# ...ETC
```

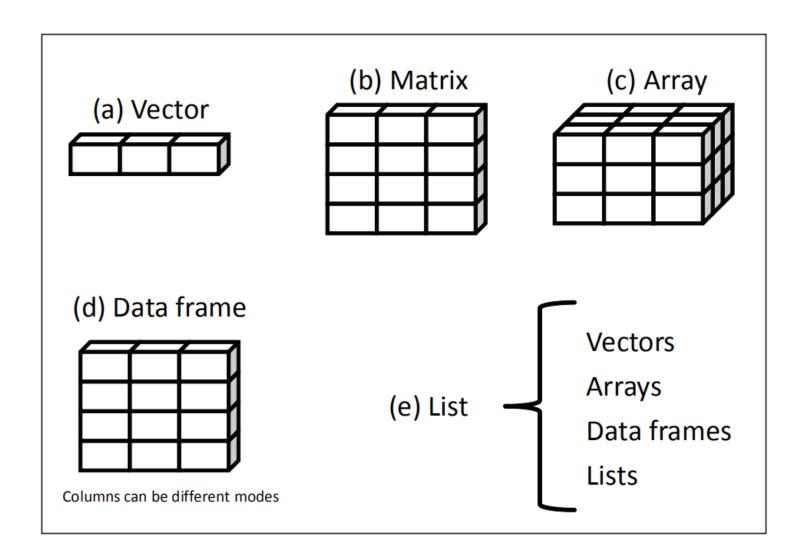
```
#Let's have a look to basic datatypes on which R objects are built
#Numeric: numbers with decimals
mynumber <- 66.6
print(mynumber)
# [1] 66.6
class(mynumber) # use it to know what is the data type
# [1] "numeric"
#Integer: numbers with no decimals
mynumber.int <- as.integer(mynumber)</pre>
# [1] 66
class(mynumber.int)
# [1] "integer"
#Character: can be a letter or a combination of letters enclosed by quotes
mychar <- "bioinfo course"</pre>
print (mychar)
# [1] "bioinfo course"
class(mychar)
#[1] "character"
#Logical: a variable that can be TRUE or FALSE (boolean)
im.true <- TRUE</pre>
print(im.true)
#[1] TRUE
class(im.true)
# [1] "logical"
```



George Boole

Objects and data-types

- Fundamental structures in R
- Objects: Vectors, Lists, Matrices, Arrays, Factors, Data frames
- Data types: numeric, integer, character, logical



Vectors

Objects that are used to store values or other information of the same data type They are created with the function "c()" that will generate a 1D array

```
species <- c(123,434,655,877,986) # we create a numeric vector
class(species)
#[1] "numeric"
length(species) # number of elements in the vector
#[1] 5
species[5] # accessing the fifth element in the vector
#[1] 986
species[1:3]
#[1] 123 434 655
species.names <- c("dog", "lion", "human", "pig", "cow") # we create a character vector
class(species.names)
# [1] "character"</pre>
```

```
species <- c(123,434,655,877,986) # we create a numeric vector

class(species)
#[1] "numeric"
length(species) # number of elements in the vector

#[1] 5
species[5] # accessing the fifth element in the vector

#[1] 986
species[1:3]
#[1] 123 434 655
species.names <- c("dog", "lion", "human", "pig", "cow") # we create a character vector
class(species.names)
# [1] "character"</pre>
```

Factors

```
#Factor: used to refer to a qualitative relationship.

# to generate a factor, we'll use a vector defined with the function c()
myfactor <- factor(c("good", "bad", "ugly", "good", "good", "bad", "ugly"))
print(myfactor)

#[1] good bad ugly good good bad ugly
#Levels: bad good ugly <- NB: levels of the factor
class(myfactor)

#[1] "factor"
levels(myfactor) # this can be used to check the levels of a factor

# [1] "bad" "good" "ugly"
nlevels(myfactor)

# [1] 3
class(levels(myfactor))

# [1] "character"</pre>
```



List

```
# List
# It can contain elements of various data types (e.g.vectors, functions, matrices, another list)
# Example of vectors with three different data types in one list
list1 <- c(1:5) # integer vector
#[1] 1 2 3 4 5
list2 <- factor(1:5) # factor vector</pre>
# [1] 1 2 3 4 5
# Levels: 1 2 3 4 5
list3 <- letters[1:5]</pre>
# [1] "a" "b" "c" "d" "e"
grouped.lists <- list(list1, list2, list3)</pre>
#[[1]]
#[1] 1 2 3 4 5
#[[2]]
#[1] 1 2 3 4 5
#Levels: 1 2 3 4 5
#[[3]]
#[1] "a" "b" "c" "d" "e"
#Accessing elements of a list
grouped.lists[[1]] # accessing the first vector
# [1] 1 2 3 4 5
```

```
#Matrix
#Like a vector, a matrix stores information of the same data type, but different from a vector, it has 2
dimensions.
#syntax: mymatrix <- matrix(vector, nrow=r, ncol=c, byrow=FALSE, dimnames=list(vector rownames,
vector colnames))
# byrow=F indicates that the matrix should be filled by columns
mymatrix <- matrix(seq(1:100), nrow=10, ncol=10, byrow=FALSE, dimnames=list(c(1:10), letters[1:10]))
print(mymatrix)
                  e f q h i
      1 11 21 31 41 51 61
      2 12 22 32 42 52 62 72 82
      3 13 23 33 43 53 63 73 83
      4 14 24 34 44 54 64 74 84
                                 94
      5 15 25 35 45 55 65 75 85
      6 16 26 36 46 56 66 76 86
                                 96
              37 47 57 67
  8
      8 18 28 38 48 58 68 78 88
                                 98
  9
      9 19 29 39 49 59 69 79 89
  10 10 20 30 40 50 60 70 80 90 100
```

#Dataframes

```
# =>NB: this is one of the most common objects in metabarcoding analyses<=
# Generated with the data.frame() function
my.data.frame<-data.frame(</pre>
Name=c("Game of Thrones", "MrRobot", "WestWorld"),
Budget=c(344,59,122),
                                                                 data frame
                                                                                    "R"
Seasons=c(8,4,3),
Audience=c(300, 14, 80),
                                                                                    "S"
Actors=c(221, 56, 90)
print(my.data.frame)
                                                                                    "T"
              Name Budget Seasons Audience Actors
                                                                            numeric
                                                                                  character
#1 Game of Thrones 344
                                8
                                        300
                                               221
#2
          MrRobot 59 4
                                        14
                                               56
#3
         WestWorld 122
                                         80
                                                90
row.names(my.data.frame) <- my.data.frame[,1] # Assign to the row names the names in the first
column
my.data.frame \langle -my.data.frame[,-1] \# Remove the fisrt column
print(my.data.frame)
                  Budget Seasons Audience Actors
                  344
                                       300
# Game of Thrones
                                              221
                                               56
# MrRobot
                      59
                                       14
                                        80
                                               90
# WestWorld
                     122
```

TRUE

FALSE

TRUE

logical

More general than a matrix and can contain different data types

Variables or features are in columns, while observations are in rows

```
class(my.data.frame)
# [1] "data.frame"
ncol(my.data.frame) # Number of columns
# [1] 4
nrow(my.data.frame) # Number of rows
# [1] 3
colnames (my.data.frame) # Column names
# [1] "Budget" "Seasons" "Audience" "Actors"
rownames (my.data.frame) # Name of rows
# "Game of Thrones" "MrRobot" "WestWorld"
colSums (my.data.frame) # Sum values in columns
# Budget Seasons Audience Actors
# 525 15 394
                        367
rowSums (my.data.frame) # We sum the values, even if they make no sense in the example
# Game of Thrones MrRobot
                                    WestWorld
              873
                            133
                                           295
```

Game of Thrones

MrRobot

Seasons Audience Actors

```
rbind(my.data.frame, my.data.frame) # appends dataframes one below the other (column names identical)
                   Budget Seasons Audience Actors
# Game of Thrones
                      344
                                8
                                        300
                                               221
# MrRobot
                                        14
                                               56
# WestWorld
                                        80
                                               90
                                        300
# Game of Thrones1
                                               221
                      344
# MrRobot1
                      59
                                        14
                                               56
                                        80
                                                90
# WestWorld1
                      122
cbind(my.data.frame, my.data.frame) # appends dataframes one next to the other (row names identical)
                   Budget Seasons Audience Actors Budget Seasons Audience Actors
 Game of Thrones
                     344
                                      300
                                              221
                                                                      300
                                                     344
                                                                             221
# MrRobot
                      59
                                       14
                                               56
                                                                              56
                    122
# WestWorld
                                       80
                                                    122
                                                                              90
head(my.data.frame, 2) # Useful to have a look to the beginning of the dataframe (specially useful in big tables)
# Here asking to print only 2 rows
                  Budget Seasons Audience Actors
# Game of Thrones
                     344
                                      300
                                              221
                      59
# MrRobot
                                       14
                                               56
my.data.frame[1:2,2:4] # Useful to look at specific sections of the dataframe
```



```
# Let's generate a dataframe with different data types
my.data.frame.2<-data.frame(</pre>
Name=c("Game of Thrones", "MrRobot", "WestWorld", "Chernobyl"),
Rating=c("Excellent", "Very Good", "Excellent", "Very Good"),
Audience.Restriction=c(TRUE,FALSE,TRUE, FALSE)
print(my.data.frame.2)
                      Rating Audience. Restriction
              Name
# 1 Game of Thrones Excellent
                                               TRUE
     MrRobot Very Good
                                              FALSE
         WestWorld Excellent
                                               TRUE
    Chernobyl Very Good
                                              FALSE
#Rename row names
row.names(my.data.frame.2) <- my.data.frame.2[,1]</pre>
my.data.frame.2 < -my.data.frame.2[,-1] # Remove redundant column 1
                     Rating Audience. Restriction
# Game of Thrones Excellent
                                            TRUE
# MrRobot
                 Very Good
                                           FALSE
# WestWorld
                 Excellent
                                            TRUE
# Chernobyl
             Very Good
                                           FALSE
str (my.data.frame.2) # Let's look at the data types within this dataframe
# 'data.frame':4 obs. of 2 variables:
                         : chr "Excellent" "Very Good" "Excellent" "Very Good"
  $ Rating
 $ Audience.Restriction: logi TRUE FALSE TRUE FALSE
# Variables in this case are characters and logical (TRUE/FALSE)
```

```
#Merge two dataframes based on a pattern
# We will use the series names to merge these dataframes as this is what they have in common

data.frame.large<-merge(my.data.frame, my.data.frame.2, by="row.names") # "by" indicates the column used for merging

# Row.names Budget Seasons Audience Actors Rating Audience.Restriction
# 1 Game of Thrones 344 8 300 221 Excellent TRUE
# 2 MrRobot 59 4 14 56 Very Good FALSE
# 3 WestWorld 122 3 80 90 Excellent TRUE
```

NB: "Chernobyl" was not used, as it was only present in one data frame, but this could be modified



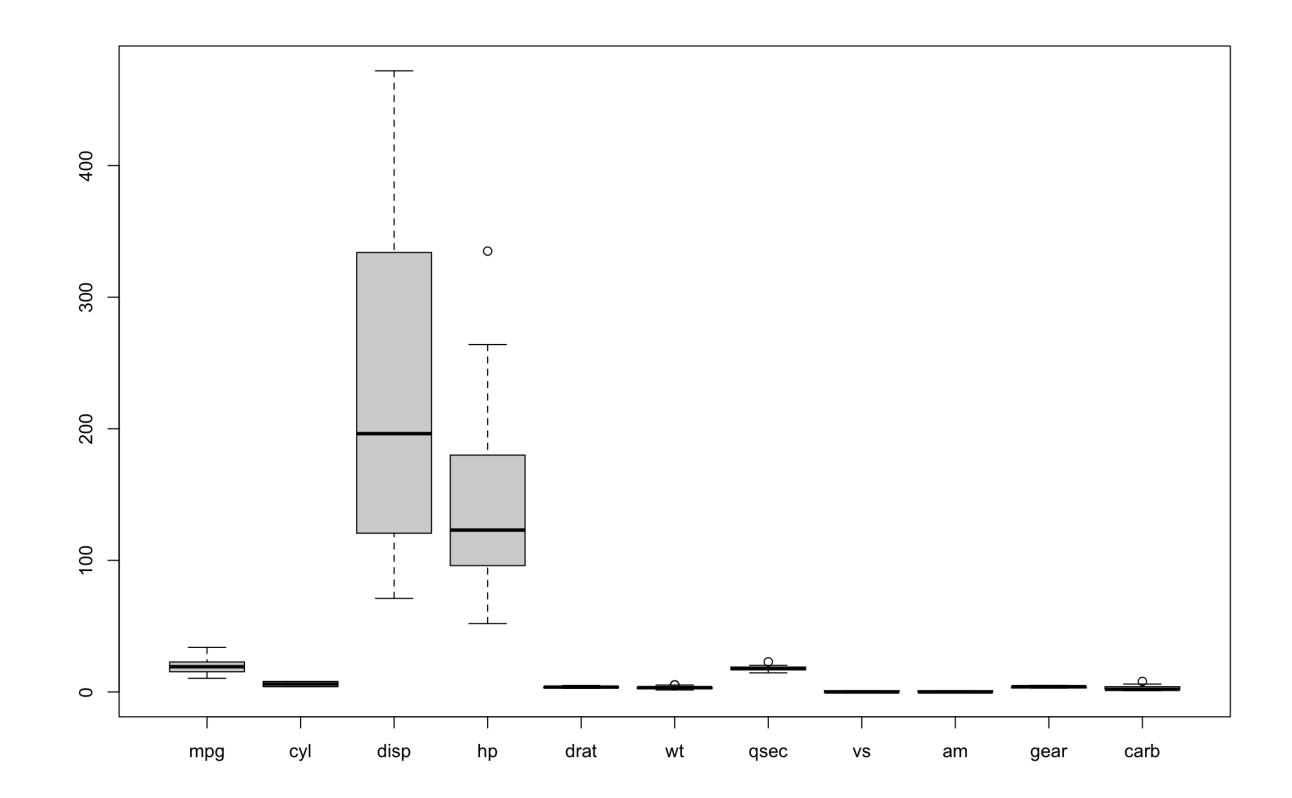
Working with tables or data frames

```
#Useful commands to work with tables or dataframes
getwd() # get working directory
# [1] "/Users/admin"
setwd("path/to/my/directory") # set working directory
my.table<-read.table(file="table.tsv", sep="\t", header=T) # read table; several other options available
dim(my.table) # Table dimensions
nrow(my.table) # Number of rows
ncol(my.table) # Number of columns
colnames (my.table) # Name of columns
rownames (my.table) # Name of rows
colSums(my.table) # Sum of numeric values in columns
rowSums (my.table) # Sum of numeric values in rows
head (my.table) # See table header
t (my.table) # Transpose table
#Table subsetting
# Format: my.table[row, column]
                                         # Get value from row 1, column 2
my.table[1,2]
                                         # Get values from row 1 across all columns
my.table[1,]
                                         # Remove column
my.table$column.name<-NULL
                                         # Remove row 5 and column 2
my.table[-5, -2]
my.table[-(5:10),]
                                         # Remove rows 5 to 10, keep all columns
my.table[,-(which(colSums(my.table)==0))] # Remove columns that sum 0
```

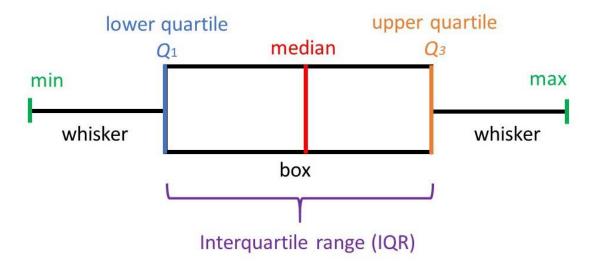
Simple plots

```
#Plotting
data("mtcars") # We load a dataset that comes with R
#The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects
# of automobile design and performance for 32 automobiles (1973 & 74 models).
#Data structure
                      mpg cyl disp hp drat wt qsec vs am gear carb
                            6 160.0 110 3.90 2.620 16.46 0 1
                     21.0
# Mazda RX4
# Mazda RX4 Wag
                     21.0 6 160.0 110 3.90 2.875 17.02 0 1
# Datsun 710
                     22.8 4 108.0 93 3.85 2.320 18.61 1 1
# Hornet 4 Drive
                     21.4 6 258.0 110 3.08 3.215 19.44 1 0 3
# . . .
# [, 1] mpg Miles/(US) gallon
# [, 2] cyl Number of cylinders
# [, 3] disp Displacement (cu.in.)
# [, 4] hp Gross horsepower
# [, 5] drat Rear axle ratio
# [, 6] wt Weight (1000 lbs)
# [, 7] qsec 1/4 mile time
\# [, 8] vs Engine (0 = V-shaped, 1 = straight)
\# [, 9] am Transmission (0 = automatic, 1 = manual)
# [,10] gear Number of forward gears
# [,11] carb Number of carburetors
```

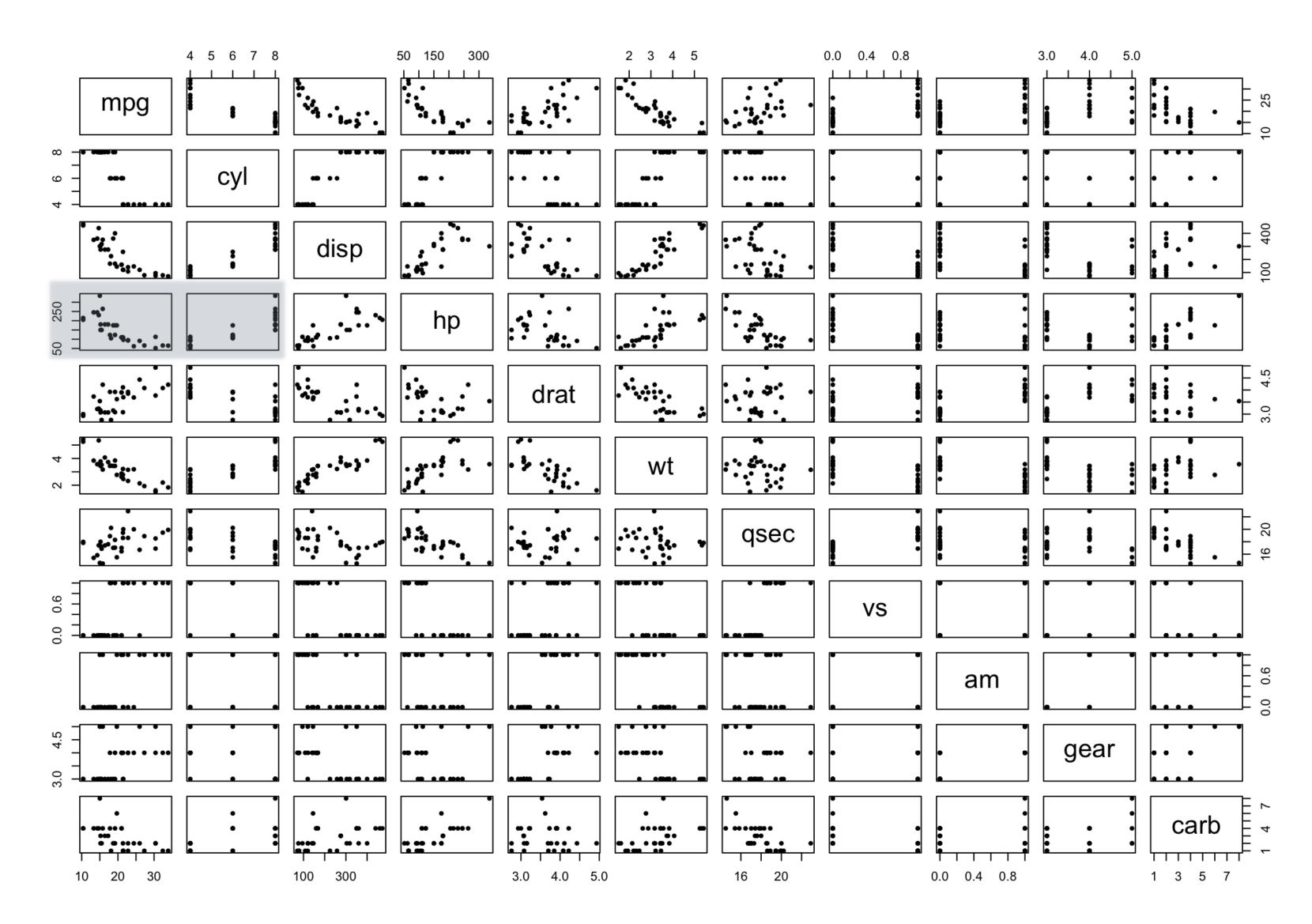




boxplot(mtcars) # make a boxplot of variables across car models

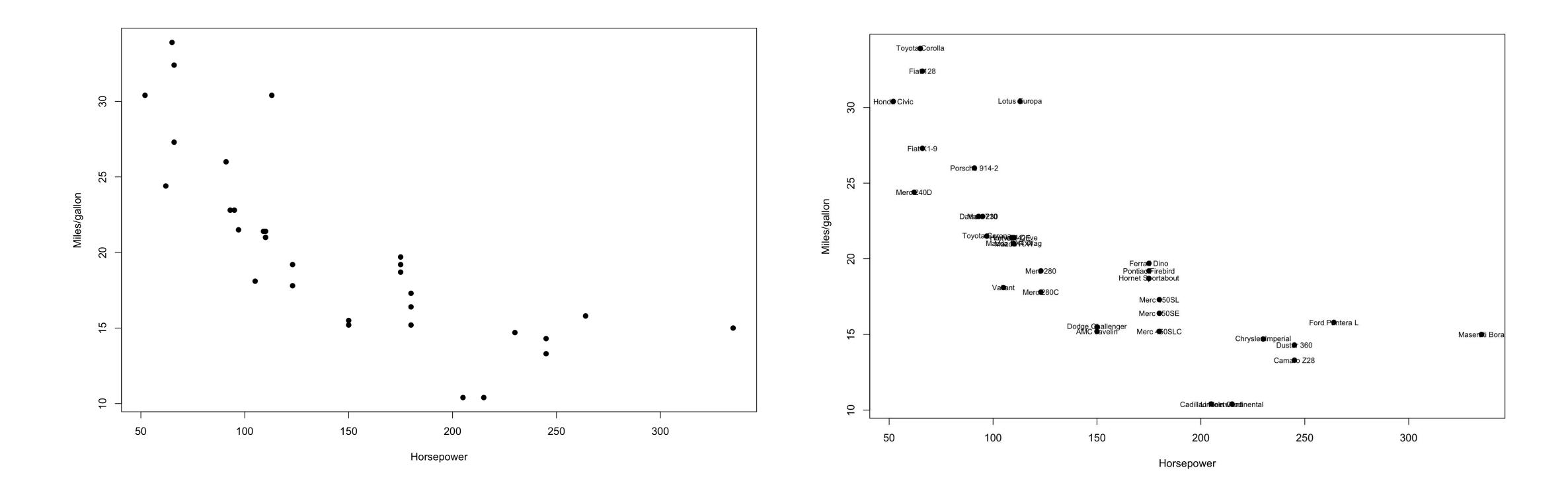


```
14 # [, 1] mpgMiles/(US) gallon
15 # [, 2] cylNumber of cylinders
16 # [, 3] disp Displacement (cu.in.)
17 # [, 4] hp Gross horsepower
18 # [, 5] drat Rear axle ratio
19 # [, 6] wt Weight (1000 lbs)
20 # [, 7] qsec 1/4 mile time
21 # [, 8] vs Engine (0 = V-shaped, 1 = straight)
22 # [, 9] am Transmission (0 = automatic, 1 = manual)
23 # [,10] gear Number of forward gears
24 # [,11] carb Number of carburetors
```

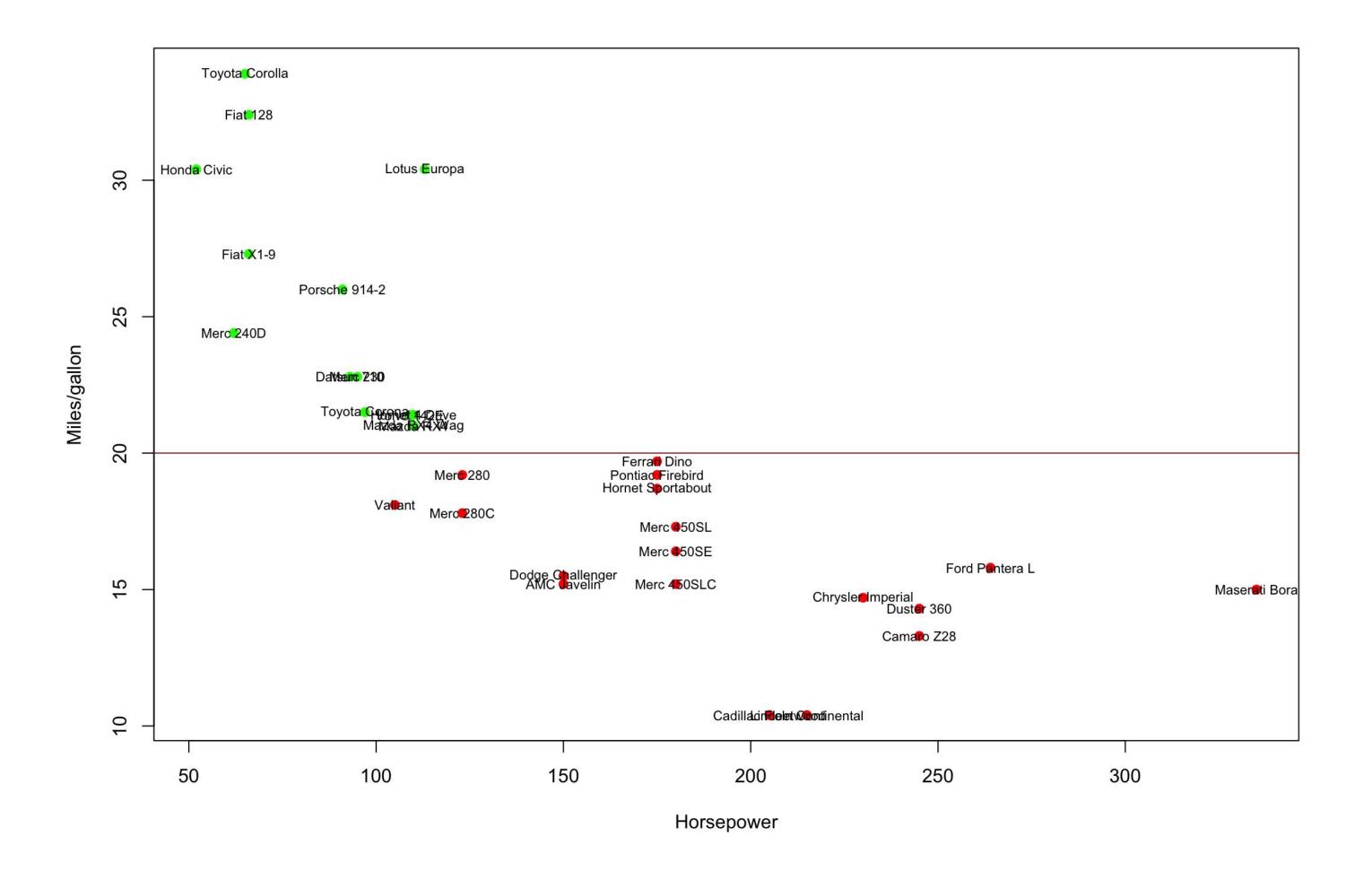


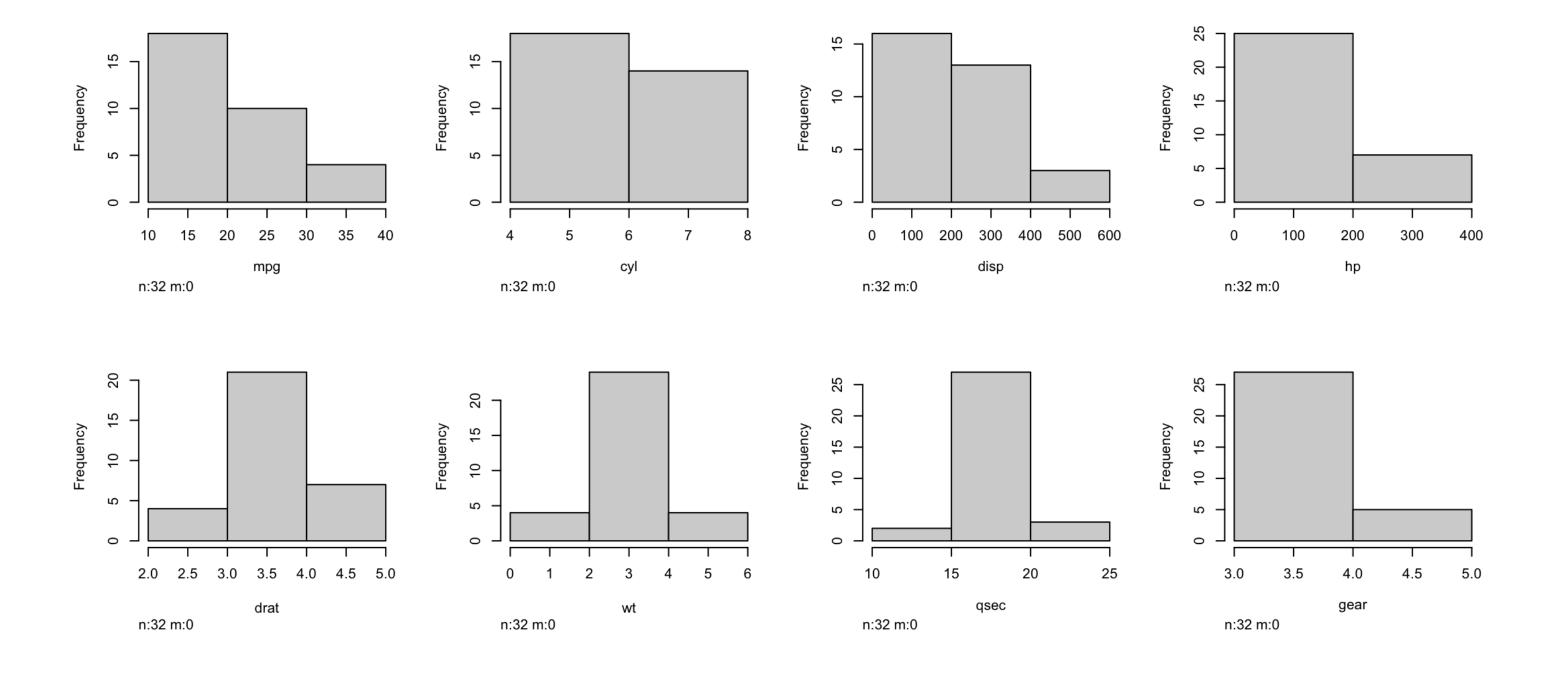
plot(mtcars, pch=19, cex=0.6) # make x-y plots for all variables

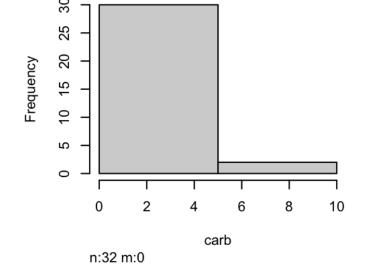
```
14 # [, 1] mpgMiles/(US) gallon
15 # [, 2] cylNumber of cylinders
16 # [, 3] disp Displacement (cu.in.)
17 # [, 4] hp Gross horsepower
18 # [, 5] drat Rear axle ratio
19 # [, 6] wt Weight (1000 lbs)
20 # [, 7] qsec 1/4 mile time
21 # [, 8] vs Engine (0 = V-shaped, 1 = straight)
22 # [, 9] am Transmission (0 = automatic, 1 = manual)
23 # [,10] gear Number of forward gears
24 # [,11] carb Number of carburetors
```



plot(mtcars\$hp, mtcars\$mpg, xlab="Horsepower", ylab="Miles/gallon", pch=19) # we plot horsepower vs. miles per gallon text(mtcars\$hp, mtcars\$mpg, row.names(mtcars), cex=0.7) # we add the car model

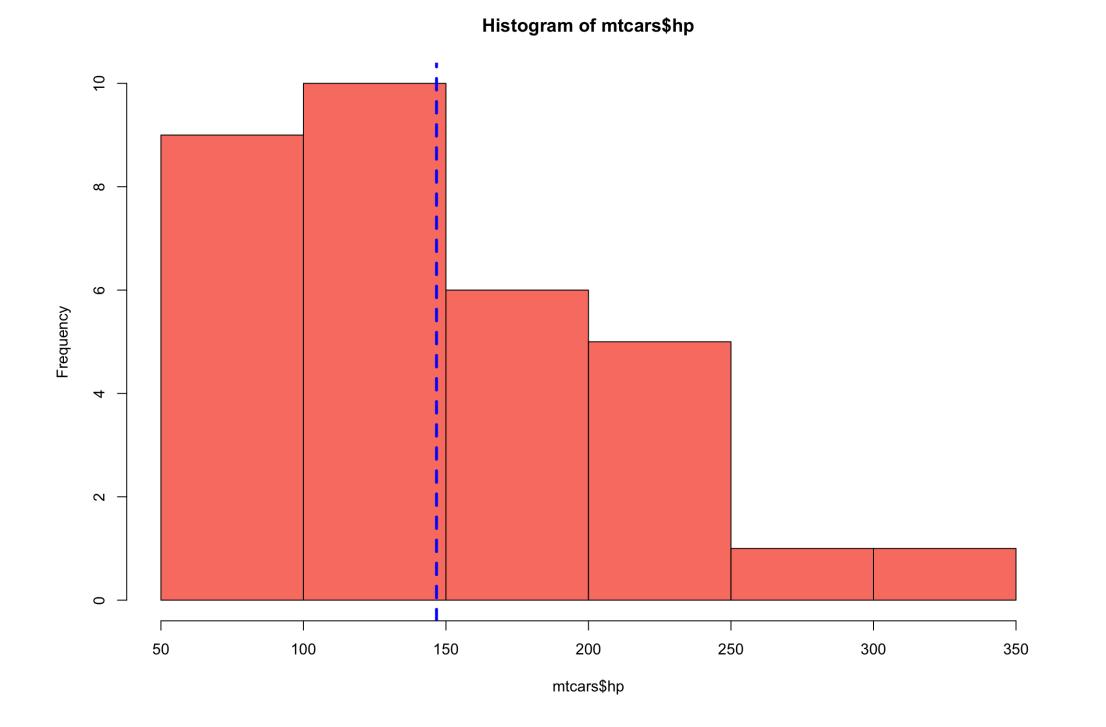


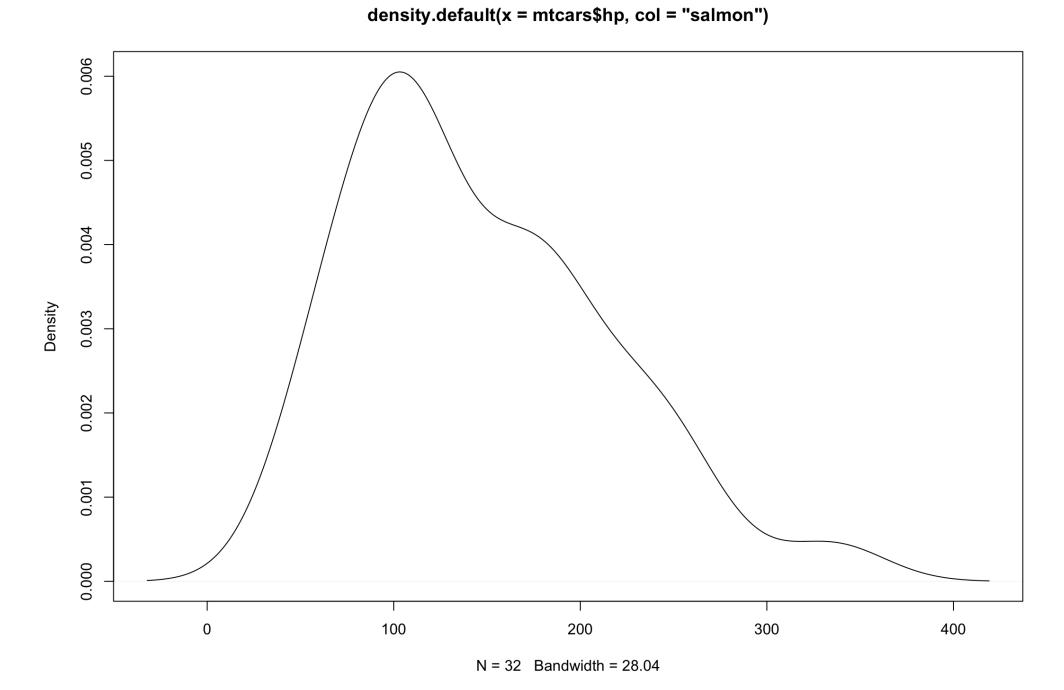




hist(mtcars) # we plot an histogram for the different variables

```
14 # [, 1] mpgMiles/(US) gallon
15 # [, 2] cylNumber of cylinders
16 # [, 3] disp Displacement (cu.in.)
17 # [, 4] hp Gross horsepower
18 # [, 5] drat Rear axle ratio
19 # [, 6] wt Weight (1000 lbs)
20 # [, 7] qsec 1/4 mile time
21 # [, 8] vs Engine (0 = V-shaped, 1 = straight)
22 # [, 9] am Transmission (0 = automatic, 1 = manual)
23 # [,10] gear Number of forward gears
24 # [,11] carb Number of carburetors
```



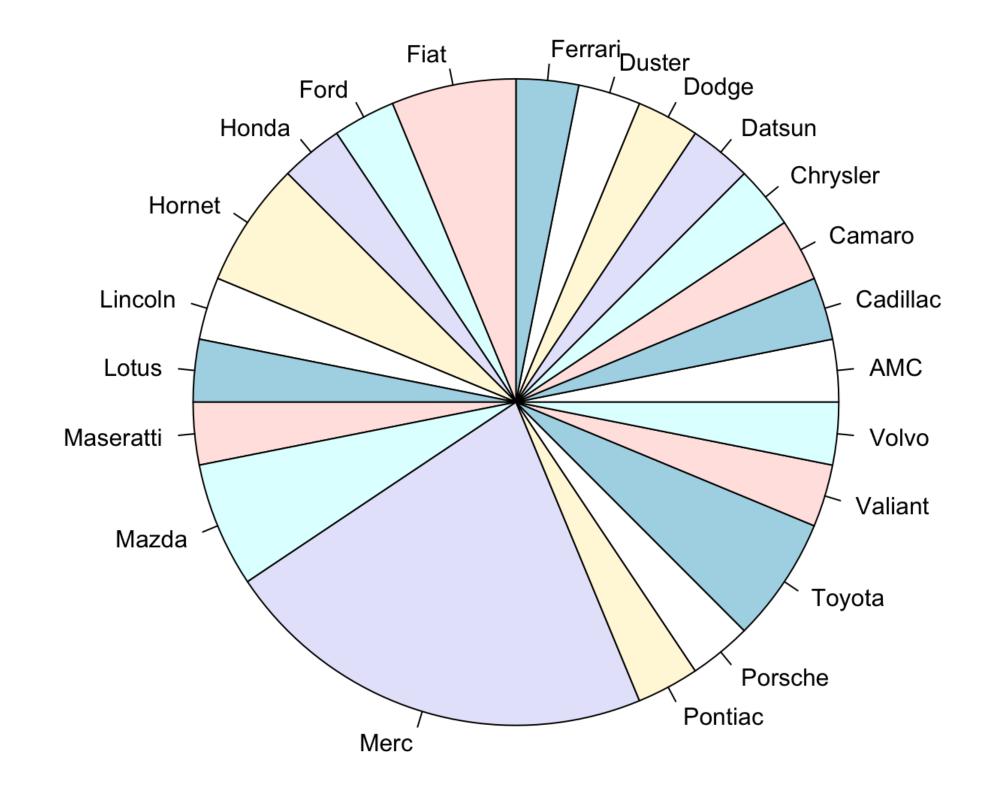


```
hist(mtcars$hp, col="salmon")
abline(v=mean(mtcars$hp), col="blue", lwd=3, lty=2)
plot(density(mtcars$hp))
```

```
brands<-c("Mazda", "Mazda", "Datsun", "Hornet", "Valiant", "Duster", "Merc", "
```

pie(table(mtcars\$brand)) # we make a piechart of the brands

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2 Hornet



Installing and loading packages

```
1 #Installing packages
3 # R has a large repository of packages for different applications
5 install.packages("spaa") # Installs the ecological package spaa
 6 install.packages("vegan") # Installs the community ecology package Vegan with hundreds of functions
7 library("vegan") # load Vegan
8 #Loading required package: permute
9 # Loading required package: lattice
10 \# This is vegan 2.5-7
11
12
13 # Other relevant packages
14
15 install.packages("readr")
                                 # To read and write files
16 install.packages("readxl")
                                 # To read excel files
17 install.packages("dplyr")
                                 # To manipulate dataframes
18 install.packages("tibble")
                                 # To work with data frames
19 install.packages("tidyr")
                                 # To work with data frames
20 install.packages("stringr")
                                 # To manipulate strings
21 install.packages("ggplot2")
                                 # To do plots
22 install.packages("kableExtra") # necessary for nice table formatting with knitr
23 install.packages("tidyverse")
24
25 if (!requireNamespace("BiocManager", quietly = TRUE))
26 install.packages("BiocManager")
27 #BiocManager::install(version = "3.10")
28 BiocManager::install(c("dada2", "phyloseq", "Biostrings"))
29
30 install.packages("devtools")
31 devtools::install github("pr2database/pr2database") # Installs directly from github resources that are not in R repos
32 devtools::install github("GuillemSalazar/EcolUtils") # Installs other tools for ecological analyses
33
34 #Load libraries
35 #### Load libraries ####
36
37 library("dada2")
38 library("phyloseq")
39 library("Biostrings")
40 library("ggplot2")
41 library("dplyr")
42 library("tidyr")
43 library("tibble")
44 library("readxl")
45 library("readr")
46 library("stringr")
47 library("kableExtra")
48 library("tidyverse")
49 #library("pr2database")
```



-The code is available in:

https://github.com/krabberod/UNIS_AB332_2024/