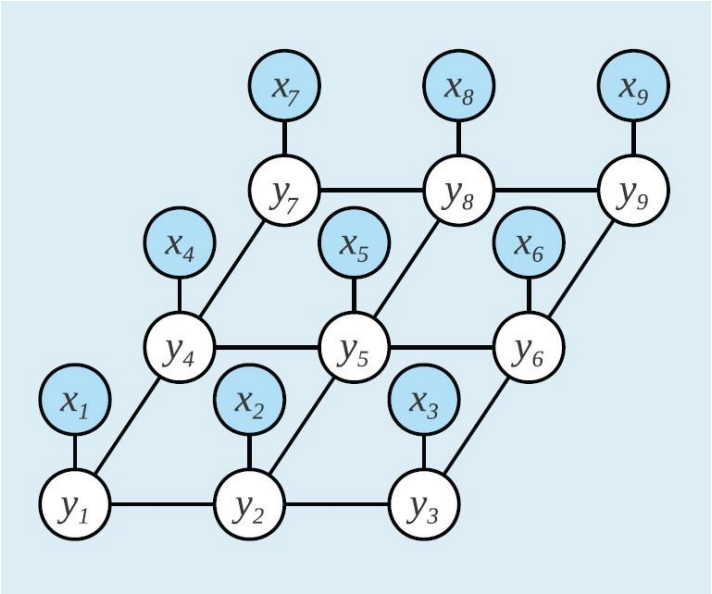
Probabilistic Graphical Models in Bioinformatics

Tutorial 1: Introduction to R

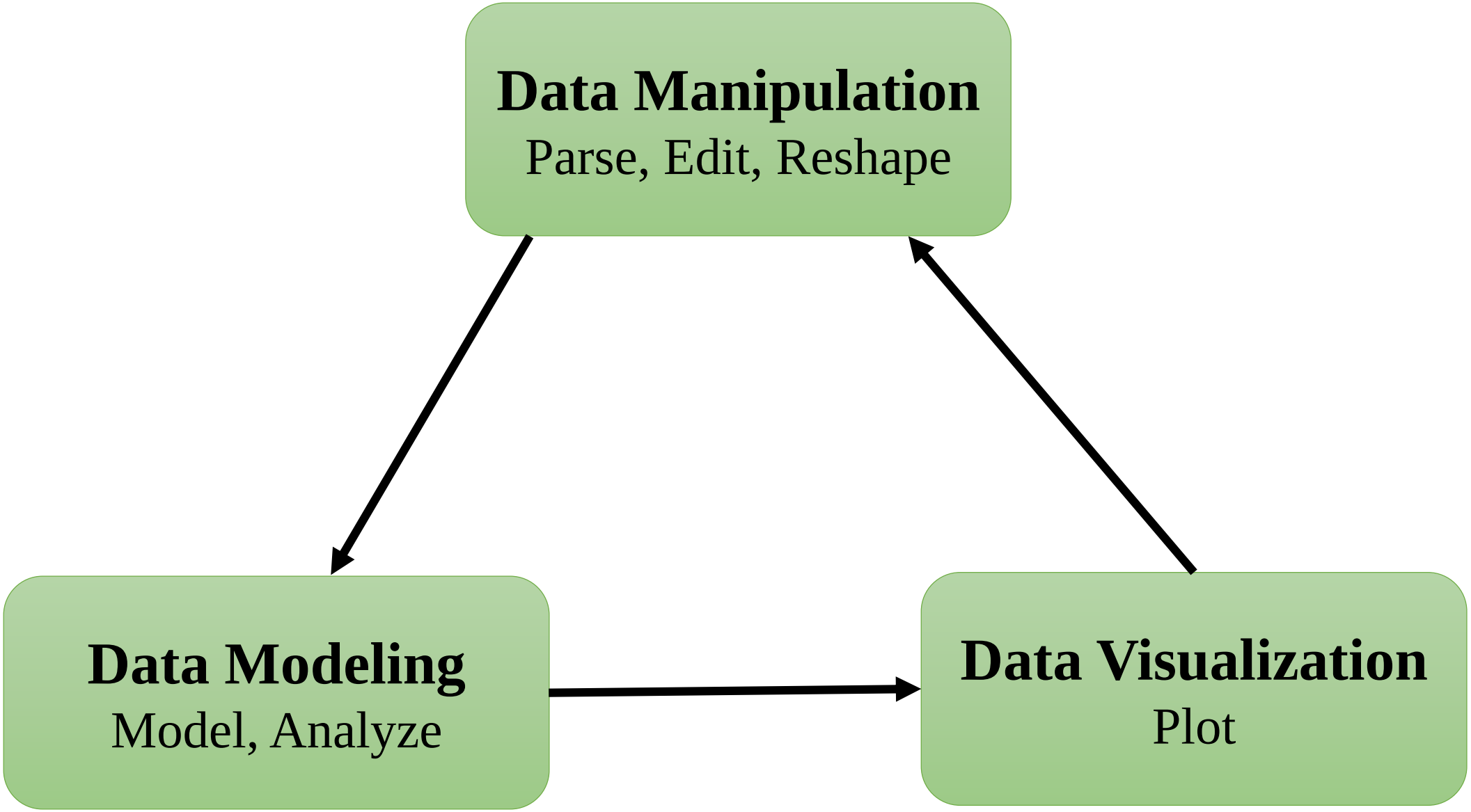


Samaneh Maleknia
Department of Bioinformatics, IBB, University of Tehran

Esfand 6, 1397
(Feb 25, 2019)

Introduction:

- **What is R?**
  Interpreted programming language based on S
  individual statements compiled to machine code
- **Open source**
  freely available for linux, osx, win (http://r-project.org)
  large and active community
- **wide variety of statistical and graphical functions**
  -
    modeling, statistical tests, classification, clustering, easy creation of publication-ready plots
    > 3500 packages providing additional functionality (CRAN, Bioconductor, …)


Data Manipulation
Parse, Edit, Reshape
Data Modeling
Model, Analyze
Data Visualization
Plot

R Basics: Getting Started

- print(“Hello World!”)

- Help

> help(“*”)
> help(exp)
>??matrix

This is a comment

- Assignment operator

> e <- m*c^2

- Display defined objects

  > ls()

Types of data:

- R has several data structures. These include:
  - atomic vector
  - matrix
  - data frame
  - list
  - factors

Types of data:

Just ***one type*** in each vector

- **R has 6 atomic vector types:**
  1. **character**: "a", "swc"
  1. **numeric**: 2, 15.5
  2. **integer**: 2L (the L tells R to store this as an integer)
  3. **logical**: TRUE, FALSE

Example:

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector
vector() # an empty 'logical' (the default) vector
```

"<-" is assigning sign in R

Types of data:

Just ***one type*** in each matrix

- **Matrix:**

matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)

Example:

```
matrix(1:6, nrow = 2, ncol = 3)
x <- 1:3
y <- 10:12
cbind(x, y)
rbind(x, y)
```

Types of data:

- **Data Frame:**

same length for variables

data.frame(..., row.names = NULL, check.rows = FALSE, check.names = TRUE, fix.empty.names = TRUE, stringsAsFactors = default.stringsAsFactors())

Example:

dat <- data.frame(id = letters[1:10], x = 1:10, y = 11:20)

- read.csv() and read.table(), i.e. when importing the data into R.
- create a new data frame with data.frame() function.
- Find the number of rows and columns with nrow(dat) and ncol(dat).

Types of data:

- **list:**

No limitation in type or length of variable

Example:

L <- list(A=c(2,5,3), B=21.3, C="sin", D=dat , G=TRUE)

L$D ➔ data frame dat

•class() - what kind of object is it (high-level)?
•typeof() - what is the object's data type (low-level)?
•length() - how long is it? What about two dimensional objects?
•attributes() - does it have any metadata?

Convert data structure to each other:

•as.matrix(), as.data.frame(), as.list(), . . .
•…

Packages

R packages are a collection of R functions, complied code and sample data. They are stored under a directory called **"library"** in the R environment.

- **Install directly from CRAN**

```
install.packages("Package Name")
```

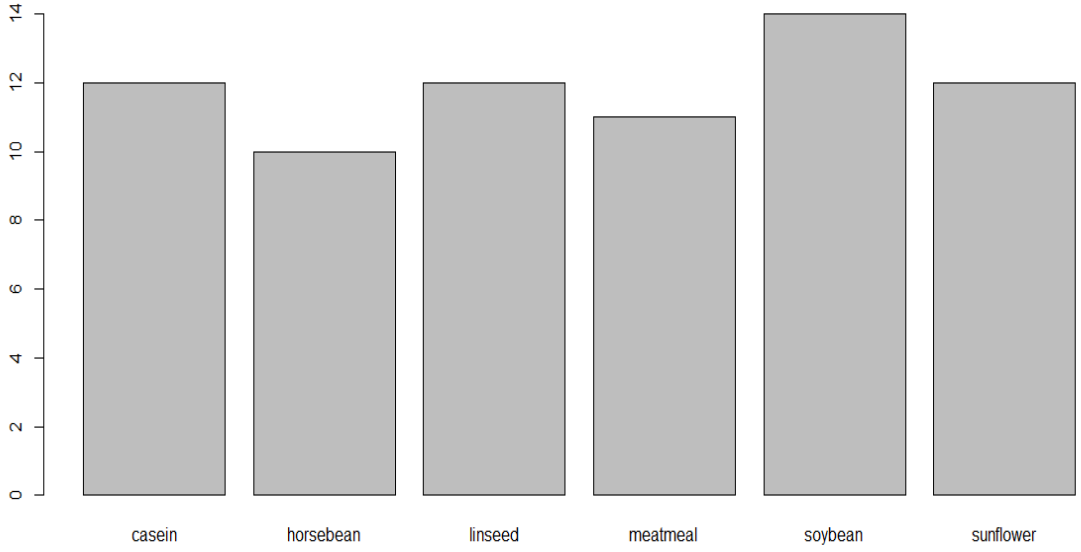
- **Install directly from Bioconductor (example:affy)**

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("affy", version = "3.8")
```

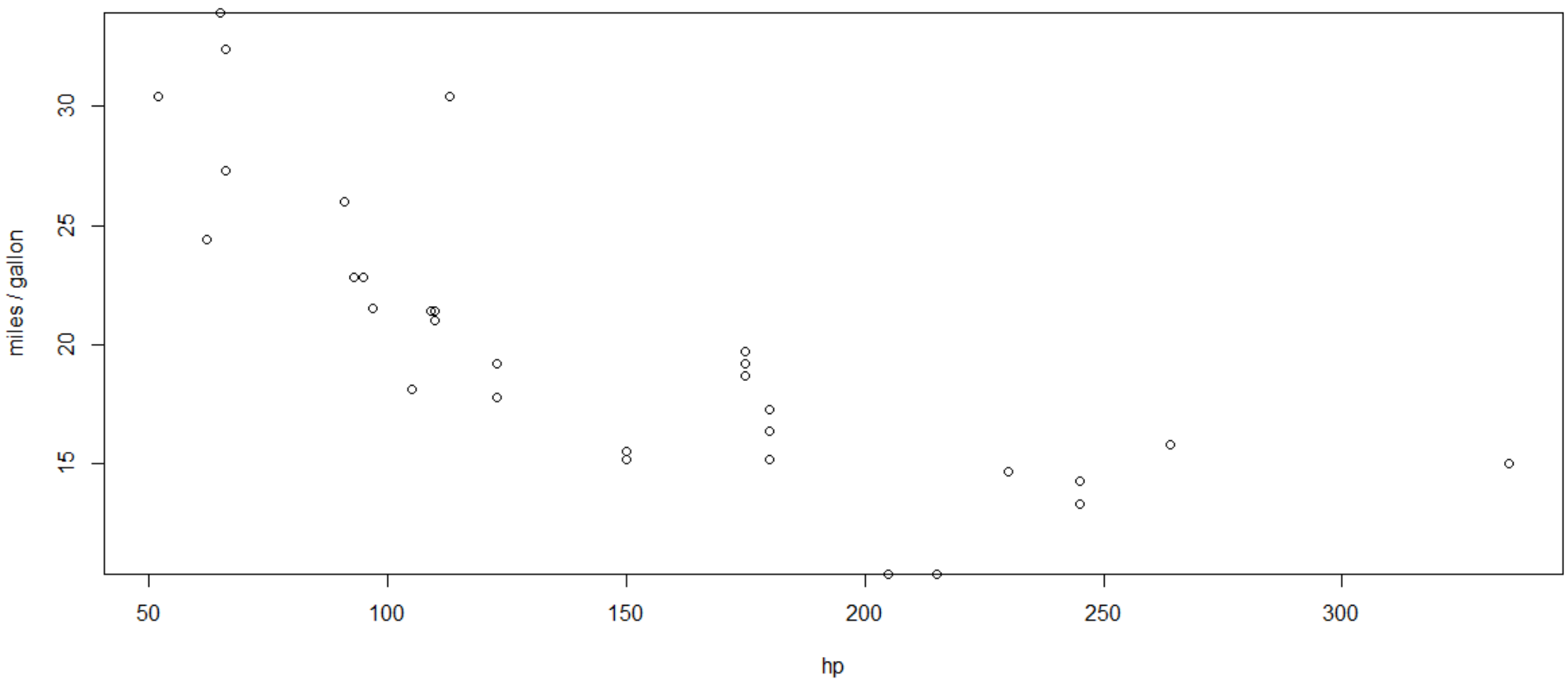
- **Use library(package name)➔ to load and use a package**

Graphical results

```
library(datasets)
?chickwts
chickwts
data(chickwts)
str(chickwts)
plot(chickwts$feed)
```




```
data(mtcars)
str(mtcars)
summary(mtcars)
# Plot Horse Power Against Miles Per Gallon
> plot(mtcars$hp, mtcars$mpg, xlab=”hp", ylab="miles / gallon”)
```


miles / gallon
15
20
25
30
50
100
150
200
250
300
hp

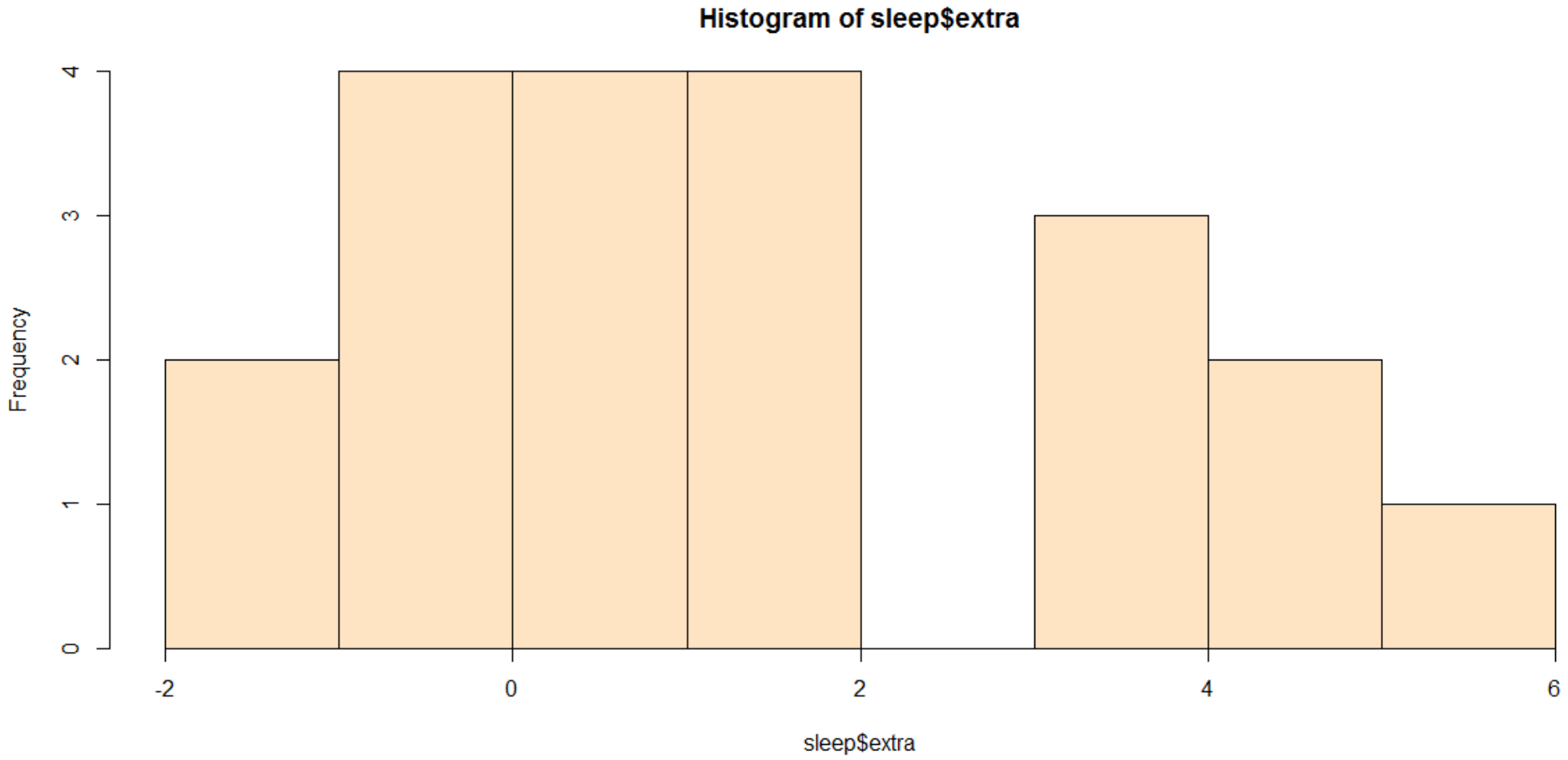
Statistics(correlation)

```
> ?swiss
> data(swiss)
> round(cor(swiss),2) # Rounded to 2 decimals
> cor.test(swiss$Fertility,swiss$Education)
> install.packages("Hmisc")
> library(Hmisc)

#to get correlation matrix and p-value
> rcorr(as.matrix(swiss))

#clean up
> detach("package:Hmisc",unload=TRUE)
> rm(list = ls())
> gc()
```

Statistics(quick plots to check data)

```
> data(sleep)
> #Some quick plots to check data
> hist(sleep$extra, col = "bisque1")
```

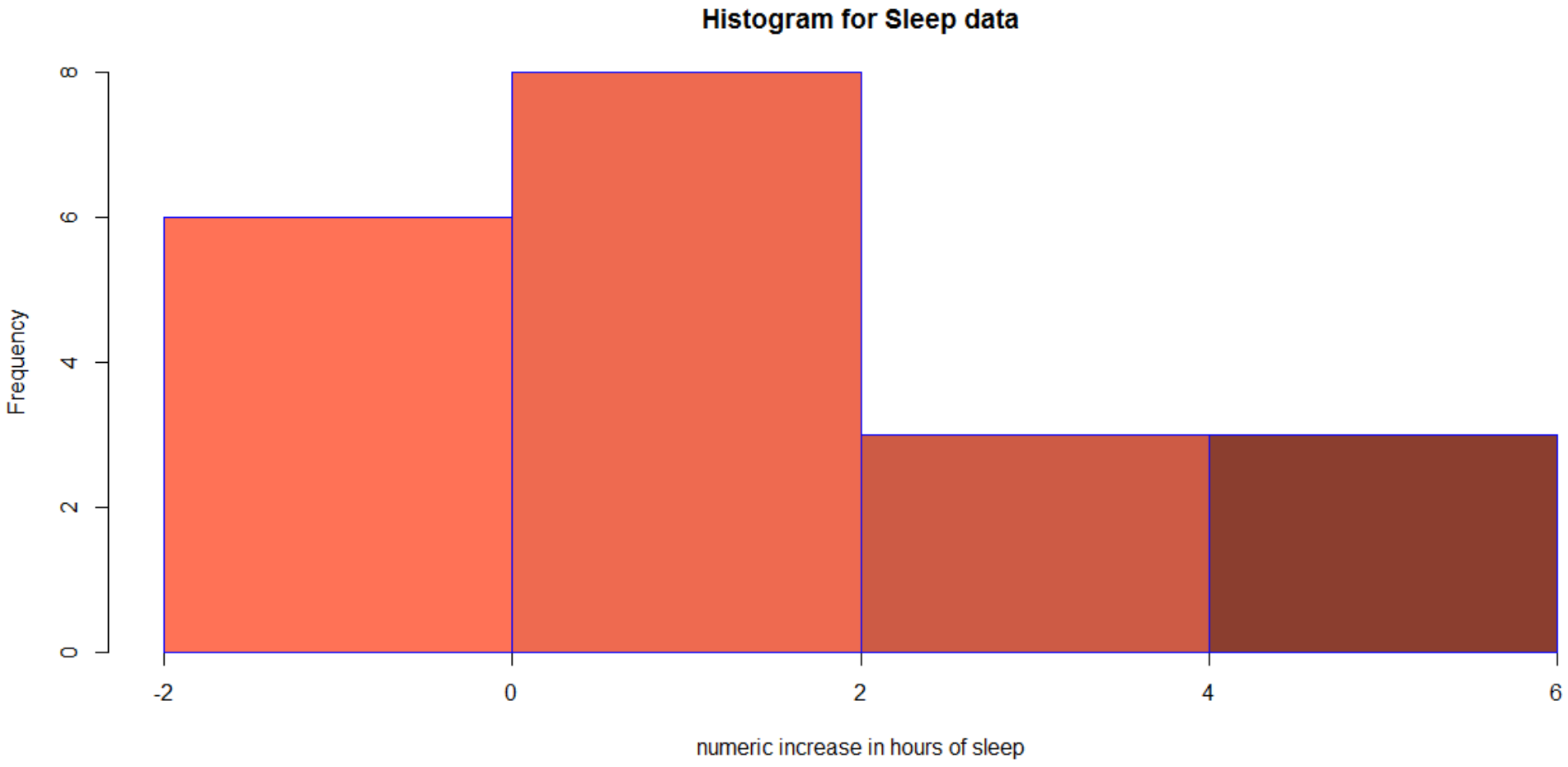
Statistics(quick plots to check data)



Statistics(quick plots to check data)

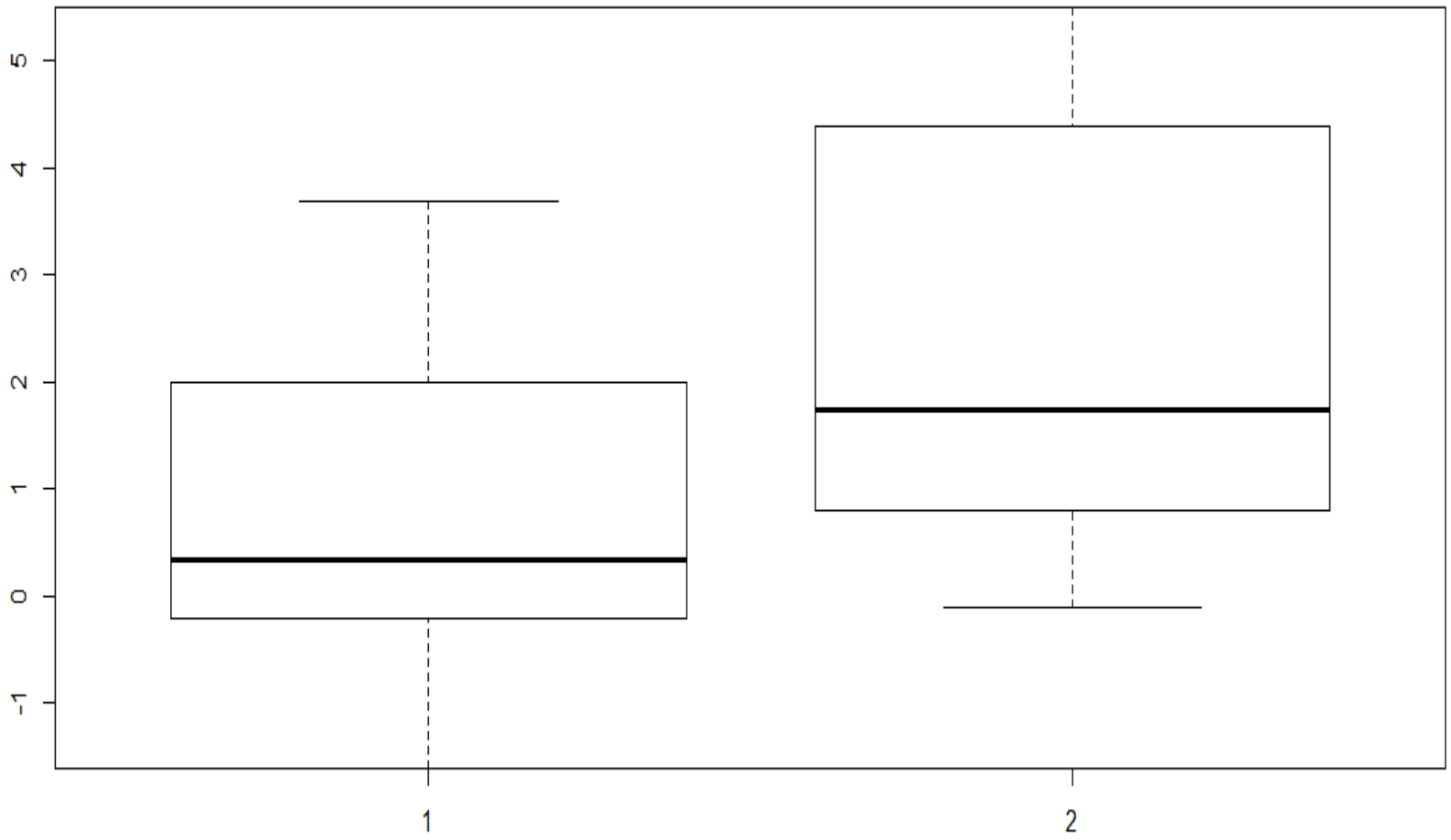
```
> hist(sleep$extra, col = c("coral1","coral2","coral3","coral4"),breaks = 4,
  main="Histogram for Sleep data", xlab="numeric increase in hours of sleep",
border="blue" )
```

Statistics(quick plots to check data)



Statistics(quick plots to check data)

> boxplot(extra ~ group, data = sleep)



Statistics(quick plots to check data)

> t.test(extra ~ group, data = sleep)

```
Welch Two Sample t-test
data: extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-3.3654832 0.2054832
sample estimates:
mean in group 1 mean in group 2
0.75            2.33
```

University of TEHRAN

Thanks for your attention!