

R FOR ABSOLUTE BEGINNERS

NUSHRAT KHAN

ALISON BLAINE

JENNIFER GARRETT



<https://www.lib.ncsu.edu/workshops>

LIFE IN A WORLD OF DATA

Imagine yourself stranded in a world of data and you're looking for a better way to process them...



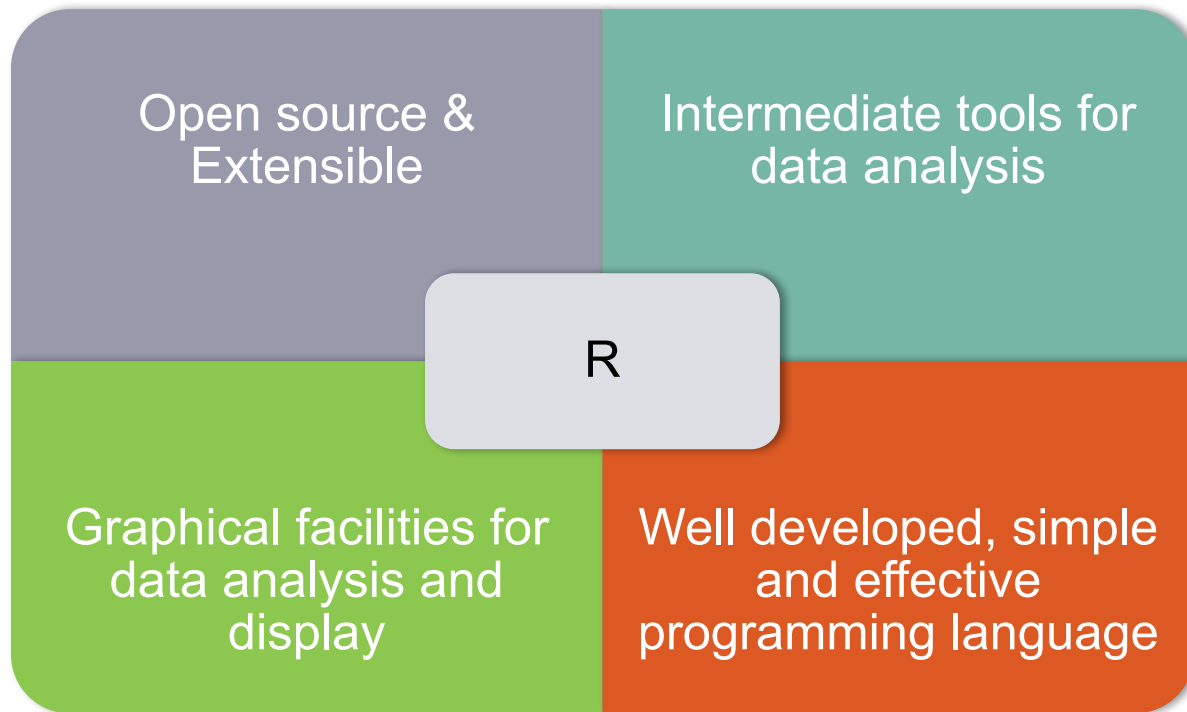
R IS WHAT YOU NEED



R programming language can be your friendly robot that can assist you to do everything with your data!

WHAT IS R?

“Free software environment for statistical computing and graphics.” – R-Project [1]



1. The R Project for Statistical Computing. (n. d.). Retrived from <https://www.r-project.org/>

OTHER STATISTICAL PACKAGES

Some well-known statistical packages include –

- **MATLAB** – Programming language with statistical features
- **Mathematica** – A software package with statistical feature
- **SAS** – Comprehensive statistical package
- **SPSS (Statistical Package for Social Sciences)** – Comprehensive statistical package

WHY USE R?

- All the other software mentioned are proprietary
- Not only a package but also a programming language
- Powerful data handling and storage facility while simple, effective and flexible
- Can write your own package if necessary and make it available for others use



APPLICATIONS OF R



Application Methods



PACKAGES USED

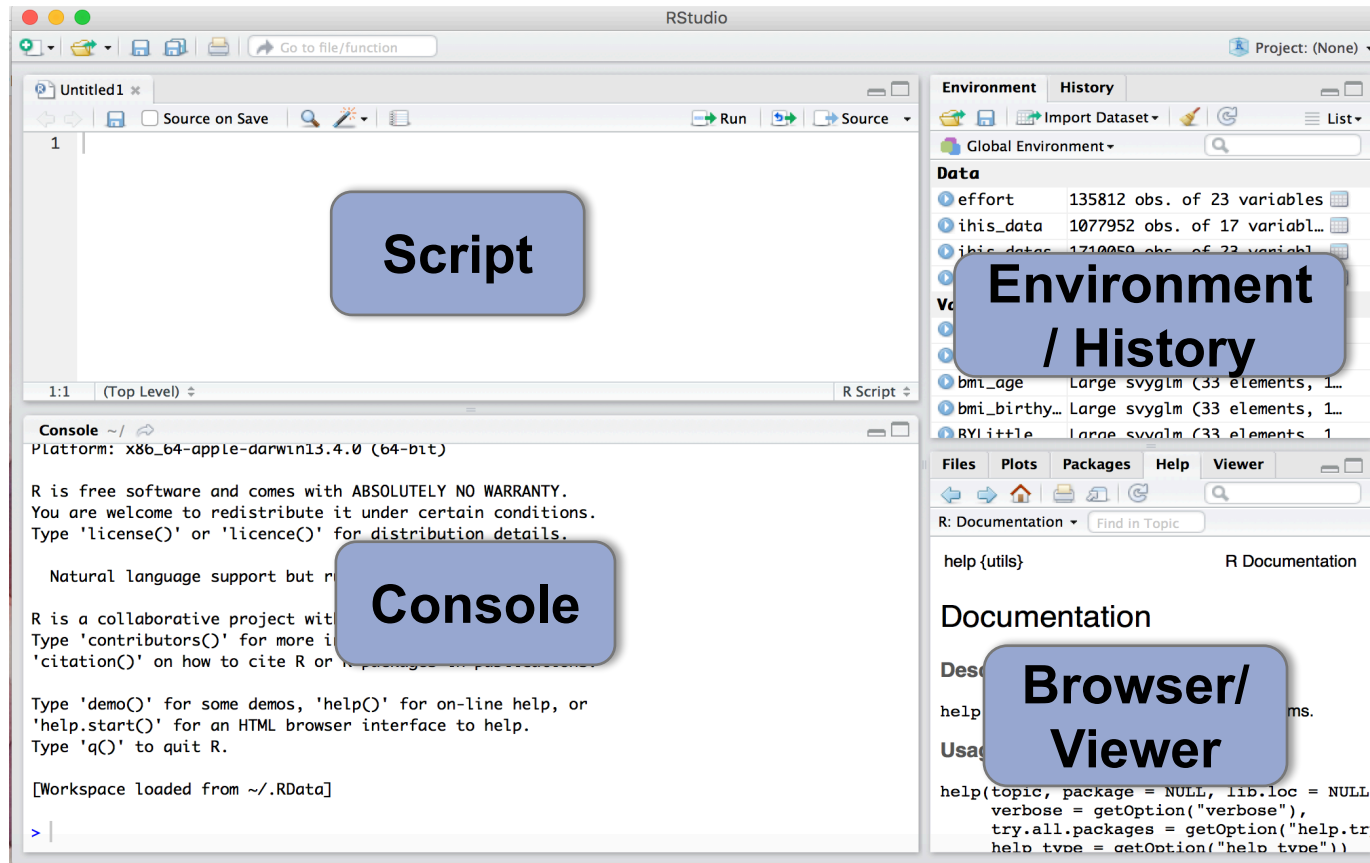
For the hands-on exercises we will use the packages below. These are some of the widely used libraries in R.

- **Hmisc**¹ – Provides numerous functions for data analysis, high level graphics, utility operations etc. We will explore the “describe” function for our exercise.
- **Dplyr**² – Contains many functions to make data manipulation easier, i.e. `filter()`, `arrange()`, `distinct()`.
- **ggplot2**³ – This package allows us to create graphs that are represented by color, symbol, size and transparency. There is a helper function `qplot()` that simplifies complex codes for some standard graphs

1. Overview of Hmisc Library. (n.d.). Retrieved from <http://math.furman.edu/~dcs/courses/math47/R/library/Hmisc/html/Overview.html>
2. Introduction to dplyr. (2016, June 23). Retrieved from <https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html>
3. Quick-R: ggplot2 Graphs. (n.d.). Retrived from <http://www.statmethods.net/advgraphs/ggplot2.html>

ABOUT R-STUDIO

A powerful user interface for R that is free, open source and works in all platforms.



WORKING DIRECTORY

Working directory – Directory of a hierarchical file system

In R Studio we can set our working directory to indicate where we want to get our data from and save our data to.

Method 1 (From the main menu) – Session > Set Working Directory > Choose Directory

Method 2 (On console) – `setwd(directory_path)`

Method 3 (Files browser) – Select directory > Select 'More' from menu > Select 'Set As Working Directory'

More on - [R Studio Support Page](#)

VARIABLES AND FUNCTIONS IN R

What is a variable?

In programming a variable is a value that can change based on the conditions. It can be useful in complex calculation by not having to repeat writing long code.

Example : `x <- c(1,2,5,7)` – here x is a variable that is holding the value of vector c

What is a function?

A function can be defined as a sub program that can be used repeatedly to perform the same task where needed. In R users can write their own functions where necessary.

Example: `f1 <- function(x,y) {x+y}`. So, `f1(1,3)` will return 4.

VARIABLES & FUNCTIONS (CONT'D)

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Shows the 'Data sets in package 'datasets':' list.
- Environment:** Displays the 'Global Environment' with variables 'i' (10) and 'j' (30), and a function 'my_function' (function (x, y)).
- Console:** Shows the execution of R code:

```
> i <- 10
> j <- 30
> my_function <- function(x,y) {x+y}
> my_function(i,j)
[1] 40
>
```
- Files Panel:** Shows the file explorer with a list of files and folders, including '.RData' (2.5 KB, Oct 28, 2016, 11:51 PM) and '.Rhistory' (1.6 KB, Nov 12, 2016, 1:26 PM).

Data sets in package 'datasets':

Variable	Description
AirPassengers	Monthly Airline Passenger Numbers 1949-1960
BJsales	Sales Data with Leading Indicator
BJsales.lead (BJsales)	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
CO2	Carbon Dioxide Uptake in Grass Plants
ChickWeight	Weight versus age of chicks on different diets
DNase	Elisa assay of DNase
EuStockMarkets	Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde	Determination of Formaldehyde
HairEyeColor	Hair and Eye Color of Statistics Students
Harman23.cor	Harman Example 2.3
Harman74.cor	Harman Example 7.4
Indometh	Pharmacokinetics of Indomethacin
InsectSprays	Effectiveness of Insect Sprays
Tobacco	Quarterly Earnings per Tobacco Company

Files Panel:

Name	Size	Modified
.RData	2.5 KB	Oct 28, 2016, 11:51 PM
.Rhistory	1.6 KB	Nov 12, 2016, 1:26 PM
Applications		
Desktop		
Documents		
Downloads		
Library		
Movies		
Music		
Pictures		
Projects		
Public		
VirtualBox VMs		

WORKING WITH DATA

- Create your own data frame by joining multiple vectors (sequence of data elements of the same basic type).
- Load your own datasets
- Work with the sample datasets that comes with R to learn and test
 - To view the list of available datasets run this command in console – `data()`
 - View and download any available dataset from this page - <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>
- For this workshop we will use `airquality` and `mtcars` datasets provided by R

VISUALIZATION WITH R

Migration to the United States by Source Region (1820 - 2006)

Visualization is made pretty easy with R, where most basic ones can be done with the **plot** command.

Types of visualization supported –

Basic Visualization

- Histogram
- Bar/ Line Chart
- Box Plot
- Scatter Plot

Advanced Visualization

- Heat Map
- Mosaic Map
- Map Visualization
- 3D Graphs
- Correlogram

To learn more about visualization with R refer to:

Chang, W. (2012). R graphics cookbook. " O'Reilly Media, Inc." *

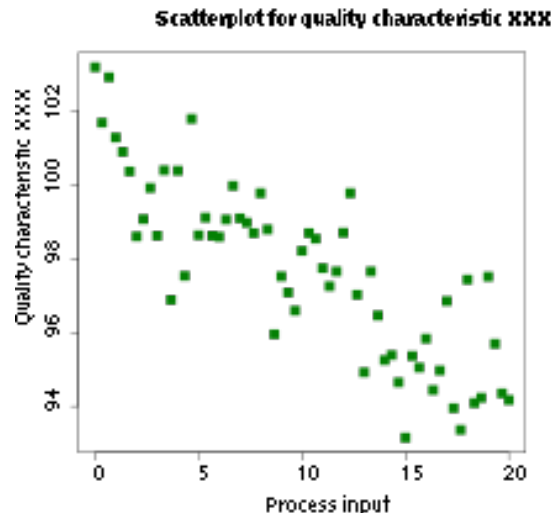
* E-book is accessible from NCSU library, but only one person at a time.

SCATTER PLOT

A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

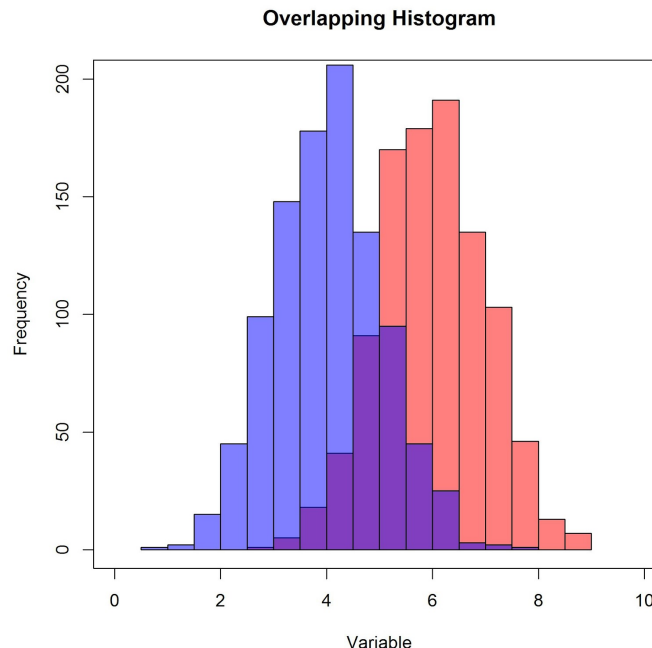
(https://en.wikipedia.org/wiki/Scatter_plot)



HISTOGRAM

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable).

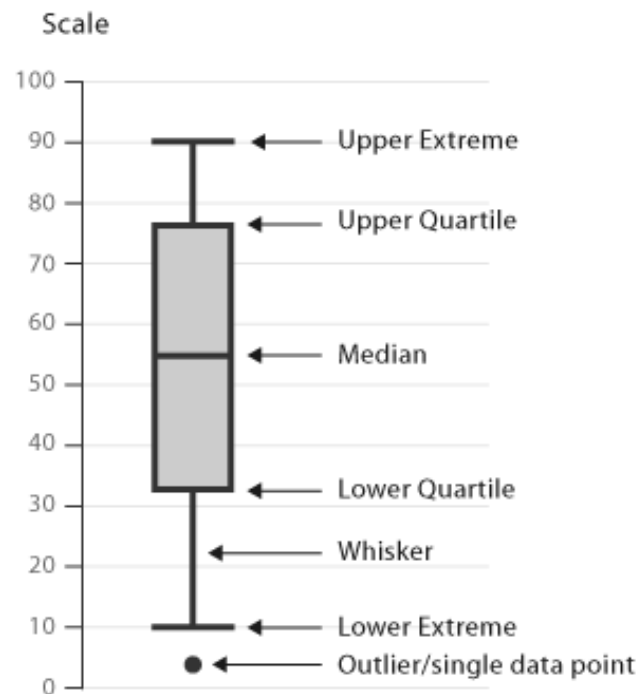
(<https://en.wikipedia.org/wiki/Histogram>)



BOX PLOT

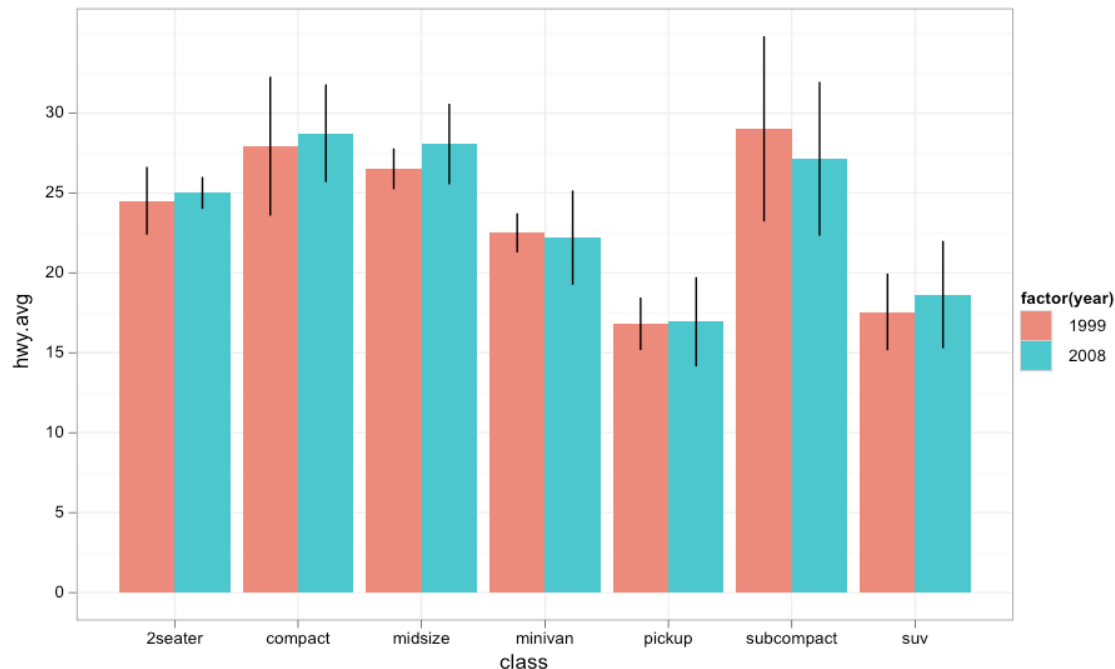
The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

(<http://www.physics.csbsju.edu/stats/box2.html>)



BAR PLOT

A bar chart or bar graph is a chart or graph that presents grouped data with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a Line graph. ([https://en.wikipedia.org/wiki/Bar chart](https://en.wikipedia.org/wiki/Bar_chart))



GET R & R-STUDIO ON YOUR MACHINE

- Open the terminal in your machine and type 'which r'. If R is already installed then it will show the path where it is located. Follow the link below to download R if it is not included.
- R can be downloaded from any of the CRAN mirrors - <https://cran.r-project.org/mirrors.html>. It is available for all types of OS – Windows, Linux and Mac.
- After downloading R, open the package and install it following the installation instructions.
- R Studio can be downloaded from the website - <https://www.rstudio.com/products/rstudio/download3/>
- Install R Studio following the instruction and R can be launched from the console within.

OTHER RESOURCES

- Impatient R – Quick tutorial of R basics for the beginners. Link: <http://www.burns-stat.com/documents/tutorials/impatient-r/>
- R – bloggers – A compiled resource useful articles on R from about 580 blogs. Link: <https://www.r-bloggers.com/>
- A short list of the most useful R commands - <http://www.personality-project.org/r/r.commands.html>
- Learn more advanced topics in depth from this book (freely available) - Wickham, H. (2014). [Advanced R](#). CRC Press.

WORKSHOP MATERIALS

Go to this link:

<http://go.ncsu.edu/rworkshop>