R FOR ABSOLUTE BEGINNERS

NUSHRAT KHAN
ALISON BLAINE
JENNIFER GARRETT

NCSU LIBRARIES

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…

NCSU LIBRARIES

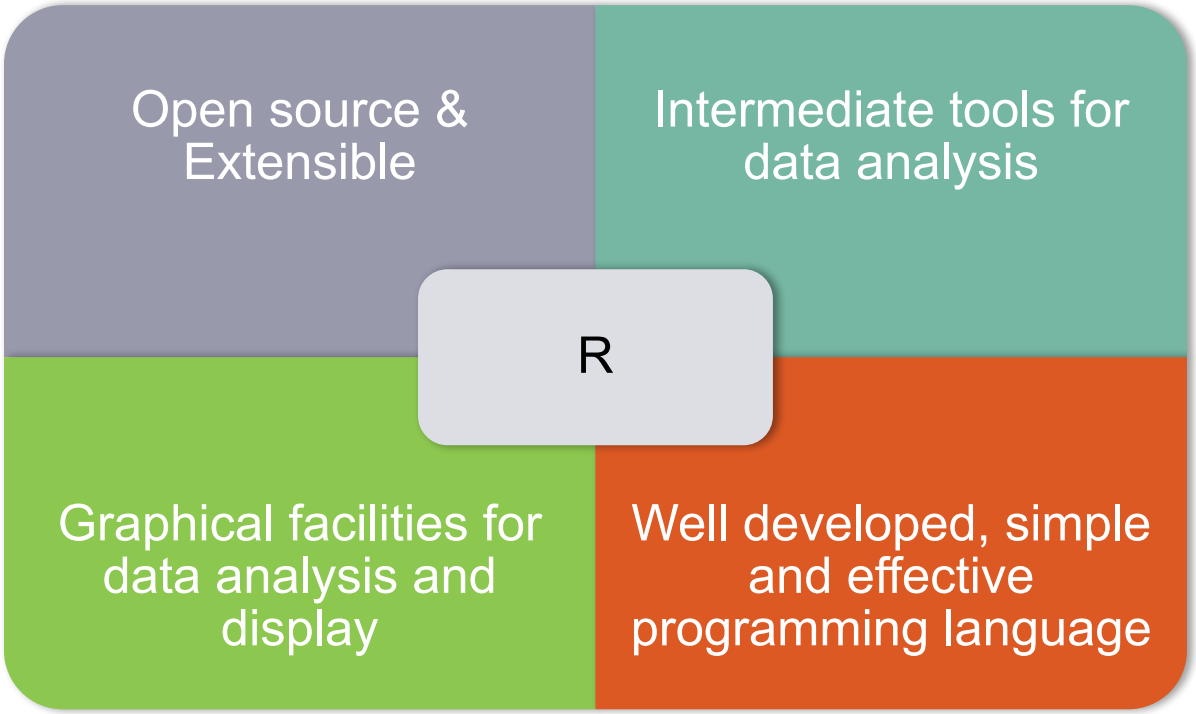
R IS WHAT YOU NEED

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

NCSU LIBRARIES

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/

NCSU LIBRARIES

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**

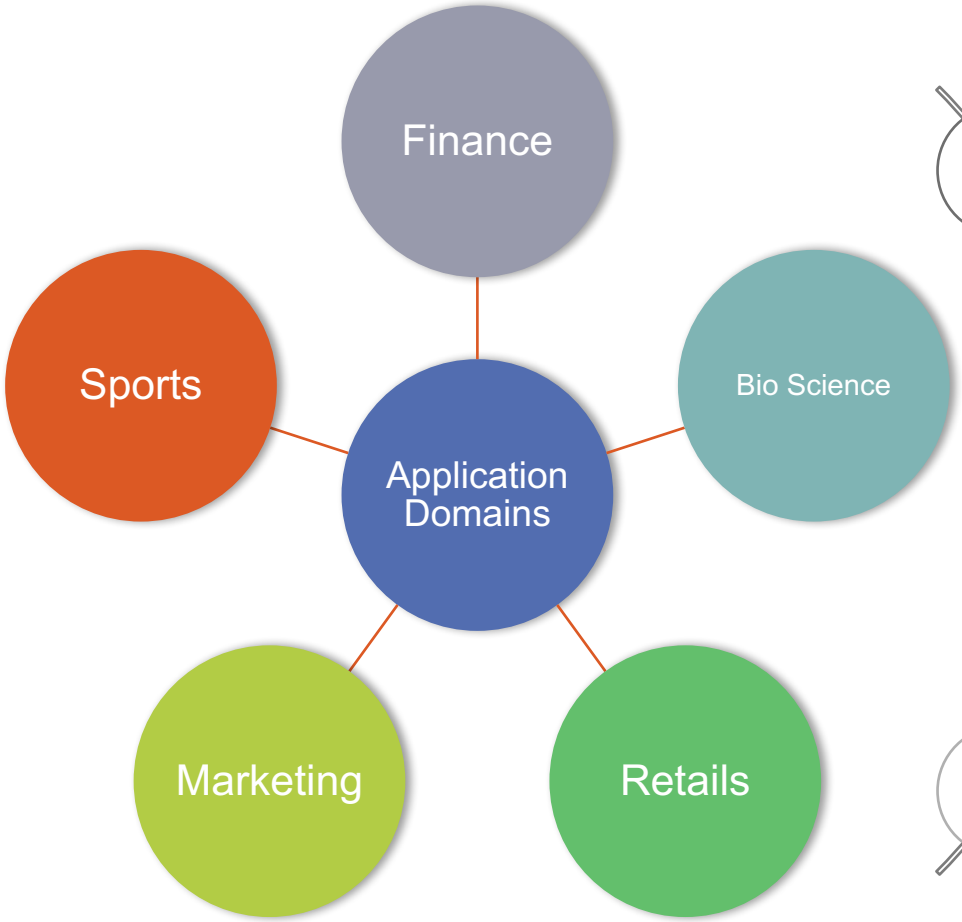
NCSU LIBRARIES

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

NCSU LIBRARIES

APPLICATIONS OF R



Application Methods

- Data Import
- Data Cleaning
- Investigation
- Data Analysis
- Visualization

NCSU LIBRARIES

PACKAGES IN R

Packages are libraries of functions that are built to perform some specific tasks, i.e. create plots.

R is supplied with about eight packages but more can be easily added and extended

To install a new package type in console –

- **install.packages("package_name")**
- **library(package_name)**

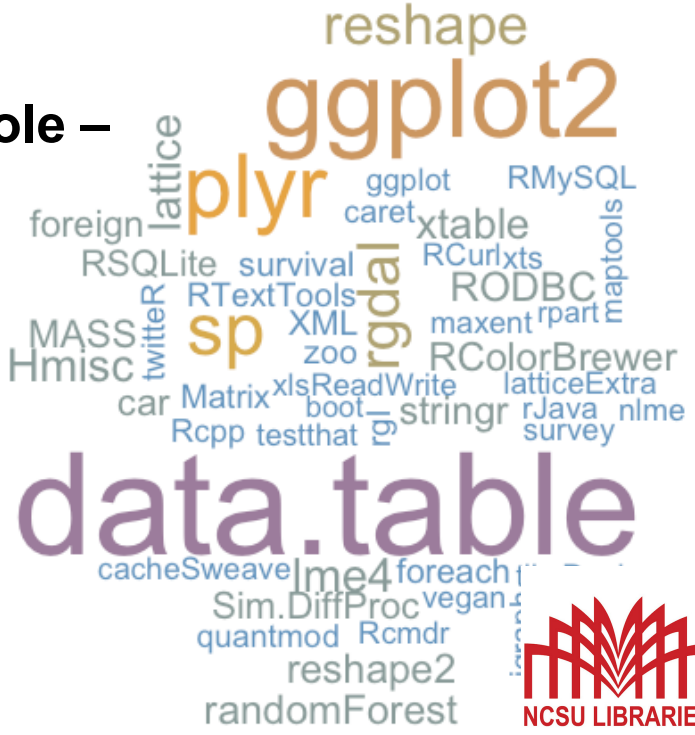


Image from http://bxhorn.com/category/r-packages/

NCSU LIBRARIES

PACKAGES USED

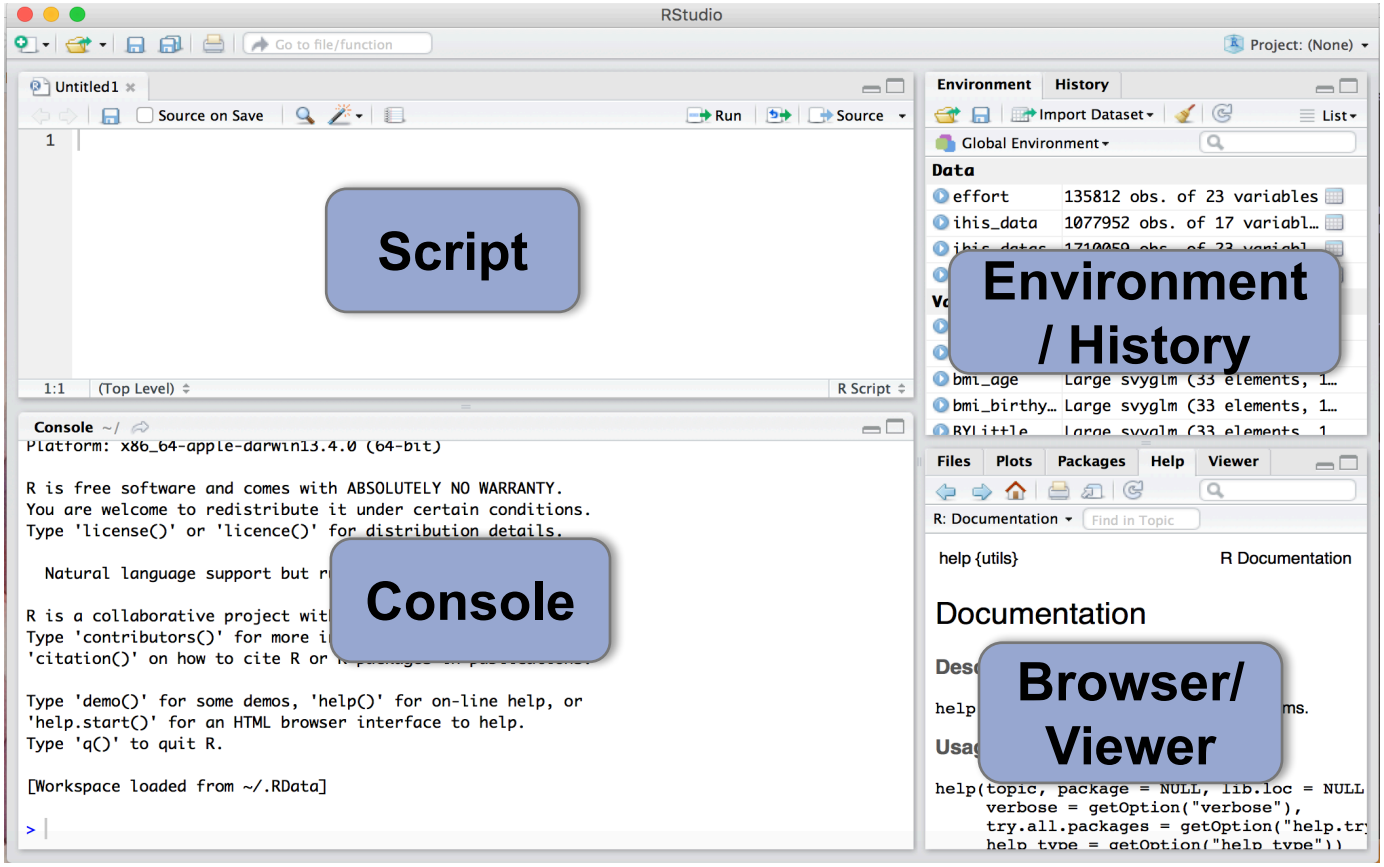
For the hands-on exercises we will use the packages below. These are some of the widely used libraries in R.

- **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.



NCSU LIBRARIES

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'

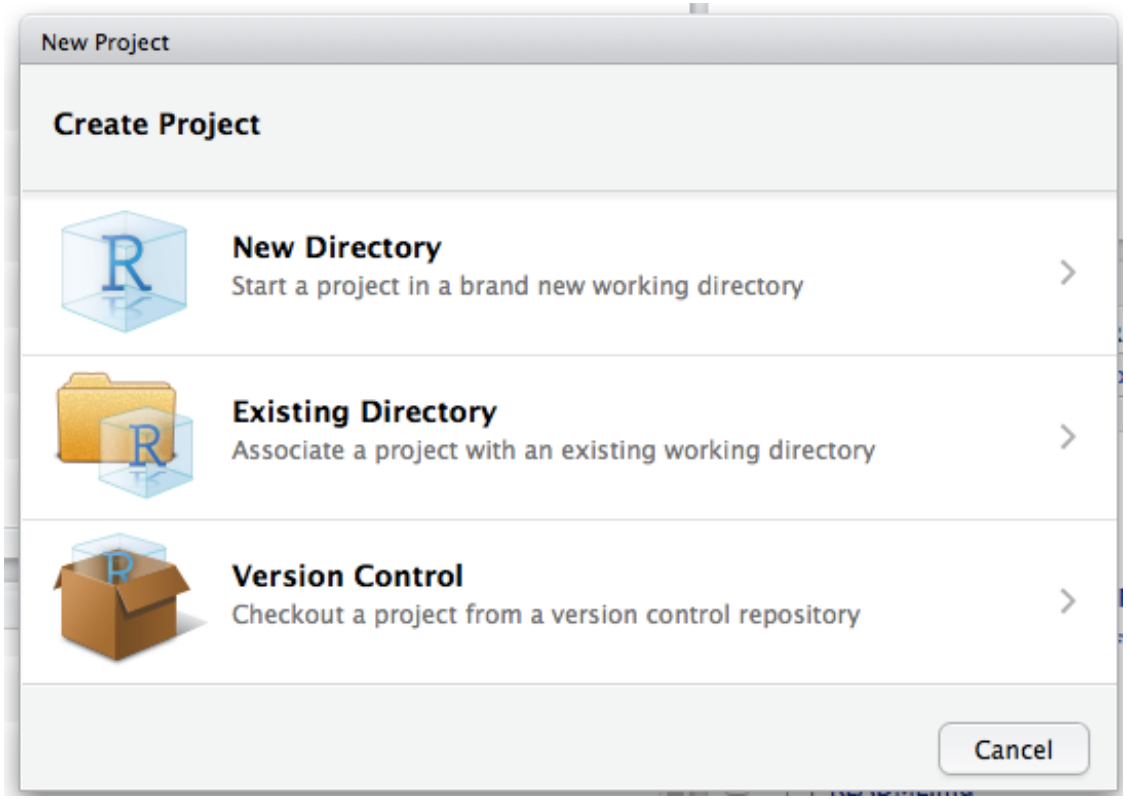
More on - R Studio Support Page

NCSU LIBRARIES

USING PROJECTS IN R STUDIO

A new function that enables dividing work into multiple contexts. Each project can have designated working directory, work space, history, and source documents.

File > New Project



More information - https://support.rstudio.com/hc/en-us/articles/200526207

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

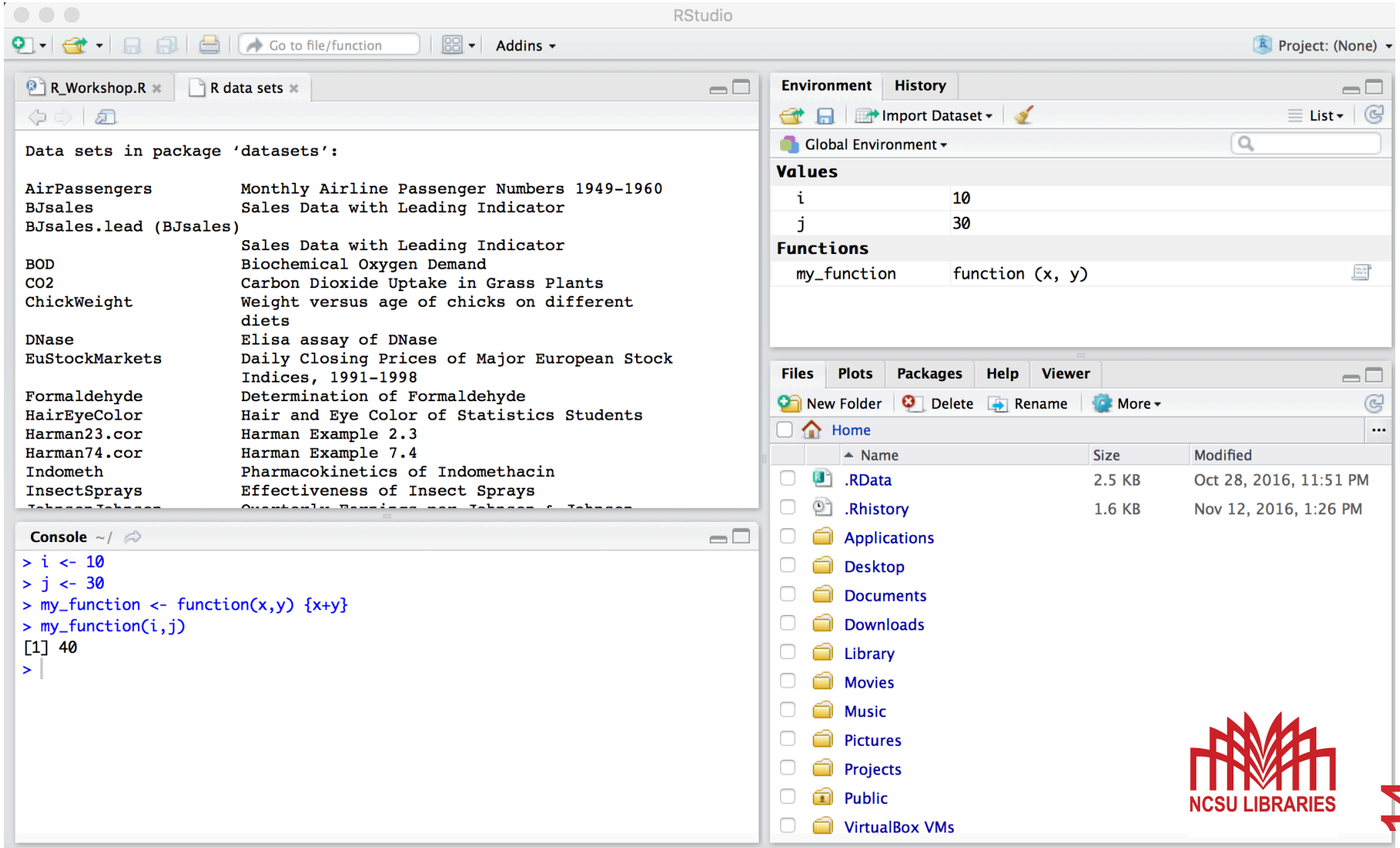
What is a function?

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.

NCSU LIBRARIES

VARIABLES & FUNCTIONS (CONT'D)



NCSU LIBRARIES

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**

NCSU LIBRARIES

VISUALIZATION WITH R

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." *

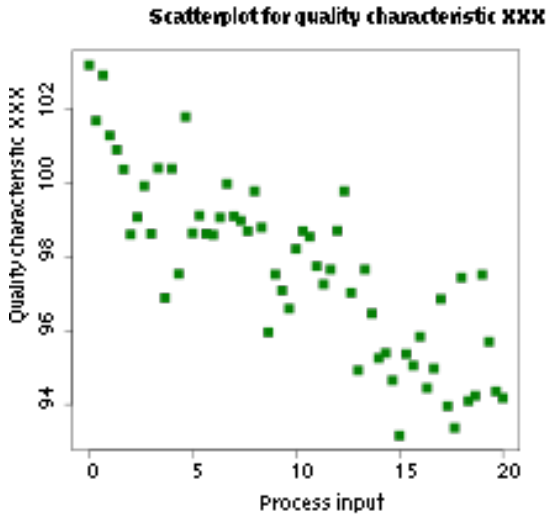
*** E-book is accessible from NCSU library, but only one person at a time.**

NCSU LIBRARIES

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)

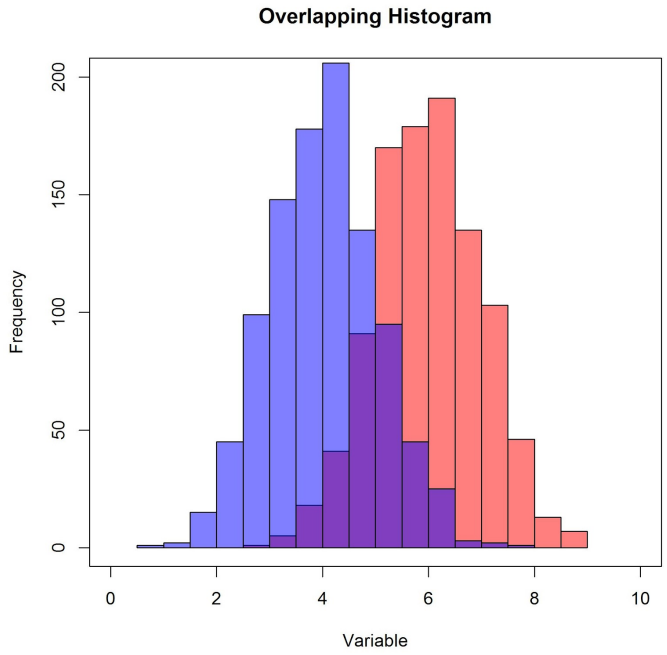


NCSU LIBRARIES

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).

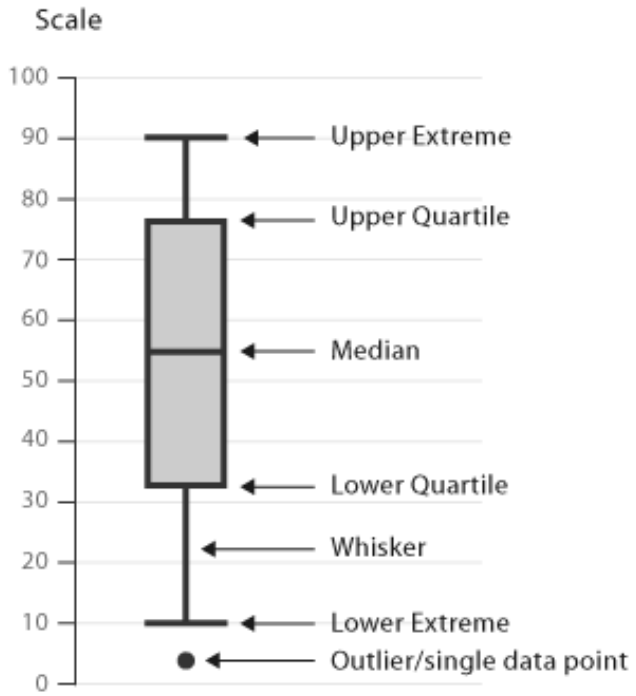
(https://en.wikipedia.org/wiki/Histogram)



NCSU LIBRARIES

BOX PLOT

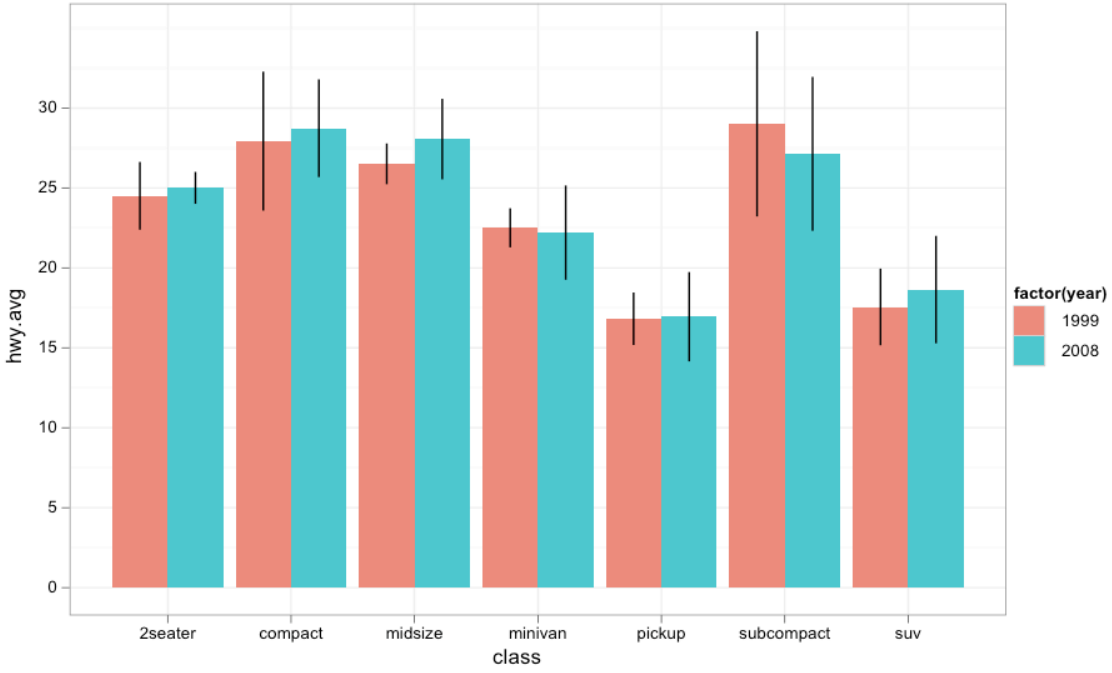
The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.
(http://www.physics.csbsju.edu/stats/box2.html)



NCSU LIBRARIES

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)



NCSU LIBRARIES

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.

NCSU LIBRARIES

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.

NCSU LIBRARIES

WORKSHOP MATERIALS

Go to this link:

http://go.ncsu.edu/rworkshop