Short term course on
RESEARCH METHODOLOGY

STATISTICAL ANALYSIS USING R

Sreekanth V K, PhD
School of Management Studies
National Institute of Technology Calicut

University of Calicut
UGC – HUMAN RESOURCE DEVELOPMENT CENTRE

INTRODUCTION TO R

An environment for Statistical Analysis

Short term course on
RESEARCH METHODOLOGY

ABOUT ME

Sreekanth V K, PhD

.

Researcher | Systems Thinking | Free & Open Source Software

- National Institute of Technology Calicut | School of Management Studies
- Indian Institute of Technology Kharagpur | RM School of Engineering Entrepreneurship
- National Institute of Technology Karnataka | Systems and Computer Applications
- Infosys Limited | Trainee, Developer, and Educator
- Mahatma Gandhi University | Electrical and Electronics Engineering

BEFORE WE START..

Prologue.

- Kindly go to the link: **PollEv.com/drsreekanthv000**
- Please fill the survey.
  - Your specialization
  - Coding experience
  - Perception about coding
- Kindly go to Rstudio Cloud (**https://rstudio.cloud/**)
  - Signup and Login

OUTLINE

Introduction to R

- Introduction
- R and RStudio
- Data
- Functions
- Conditional Flow
- Packages
- Distributions
- File handling
- Datasets
- Help and Citation
- Clean Environment

INTRODUCTION

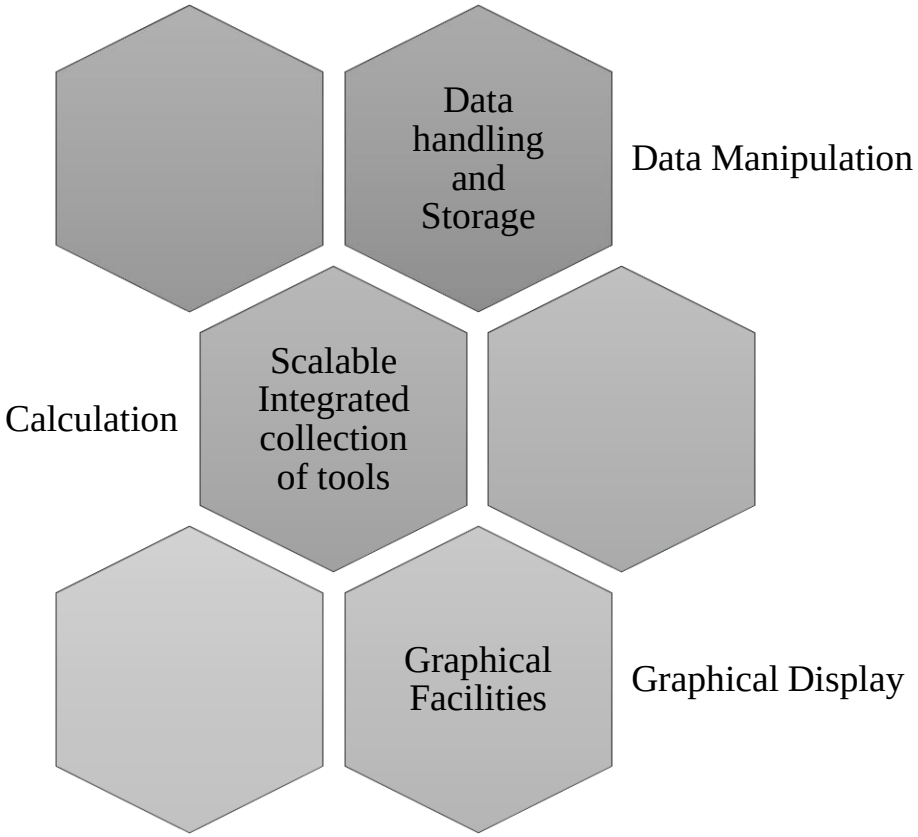
A free and open source environment for statistical computing

- Why do we need a language?
- Why do we need a language for statistics?

R.

a free software environment for statistical computing and graphics.

.



Extensible

- provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

Cross-platform

- compiles and runs on UNIX. LINUX, Windows, and MacOS.

RSTUDIO.

a premier integrated development environment (IDE) for R. .

.

- It is available in open source and commercial editions on the desktop (Windows, Mac, and Linux) and from a web browser to a Linux server running RStudio Server or RStudio Server Pro.

An IDE that was built just for R

- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions

Bring your workflow together

- Integrated R help and documentation
- Easily manage multiple working directories using projects
- Workspace browser and data viewer

Powerful authoring & Debugging

- Interactive debugger to diagnose and fix errors quickly
- Extensive package development tools
- Authoring with Sweave and R Markdown

R AND RSTUDIO.

Live Demo.

DATA

Data in R.

Major data formats

- Type
  - Numeric
  - Characters
  - Date
- Structure
  - Vectors
  - Arrays/Matrices
  - List
  - Data Frame

Naming of variables

- Should start with character
- Can have numerical values in the name.
- Dot(.) and under score (_) are allowed in the variable name. Space and other special characters are not supported
- Needs to be careful as the user defined names can take same names as that of already defined functions, and variables. R does not warn or throw error if we use already existing names.

FUNCTIONS

Functions in R and User-defined functions

```
function(argument-list) {body}
```

Built-in function

- Already existing in the R environment or packages
- Can be directly used by the users

User-defined function

```
function.name <- function(arguments)
{
    computations on the arguments
    some other code
 }
```

CONDITIONAL FLOW

Conditional flow and loops

Conditional Statements

- if
- if…else ….
- if…else if ….else …

- switch statement

Loops

- for (variable in vector){}
- while (condition) { }
- apply, sapply, lapply, and others

R PACKAGES

Packages in Comprehensive R Archived Network (CRAN), and more

From CRAN

- Base
  - Installed but not loaded
- Contributed
  - Need to download, install, and load separately
  - Some useful packages: pacman, dplyr, GGally, ggplot2, ggthemes, ggvis, httr, lubridate, plotly, rio, rmarkdown, shiny, stringr, tidyr
- CRAN (https://cran.r-project.org/)
  - install.packages()
  - installed.packages()
  - library()
  - require()

Other repositories

- Github (https://github.com/trending/r)
  - install.packages("devtools")
  - library(devtools)
  - install_github(<package>)

DISTRIBUTIONS

Statistical Distributions

Distributions in stats package

- p for "probability", the cumulative distribution function (c. d. f.)
- q for "quantile", the inverse c. d. f.
- d for "density", the density function (p. f. or p. d. f.)
- r for "random", a random variable having the specified distribution

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

FILE HANDLING

Reading and Writing Files in R

CSV, Text files, and more

- read.table()
- read.csv(), read.csv2()
- read.delim(), read.delim2()

- **rio**

Other files

- **Microsoft Excel file**
  - `library(readxl)`
  - `df <- read_excel("File.xlsx")`
- **Geo spatial data**
  - `install.packages(c("sp", "rgdal"))`
  - `library(sp)`
  - `library(rgdal)`
  - `neighborhood <- readOGR("data/nynta_20d")`
  - `plot(neighborhood)`

DATA SETS

Some of the Datasets available in R

Built-in datasets and more

- datasets
- ISLR (data set used in the text book – An Introduction to Statistical Learning, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani) [for more details https://statlearning.com/book.html]

- `data()`: Loads specified data sets, or list the available data sets.

Open data sets

- **Neighborhood Tabulation Areas (Formerly "Neighborhood Projection Areas")**
- **`Free Data Sets`**

HELP AND CITATION

To avail help and to give credit to the authors

Help in R

- help()
- ?
- ??

Citation

- citation()
- citation("forecast")
- citation("datasets")

CLEAN ENVIRONMENT

To clean the R environment after use

Remove

- rm()
- detach()

Clean Console

- Ctrl + L **cat("\014")**

ANALYSIS USING R

An environment for Statistical Analysis

Short term course on
RESEARCH METHODOLOGY

BASIC VISUALIZATION

To visualize the data and understand data

Basic Graphics

- plot()
- barplot()
- hist()

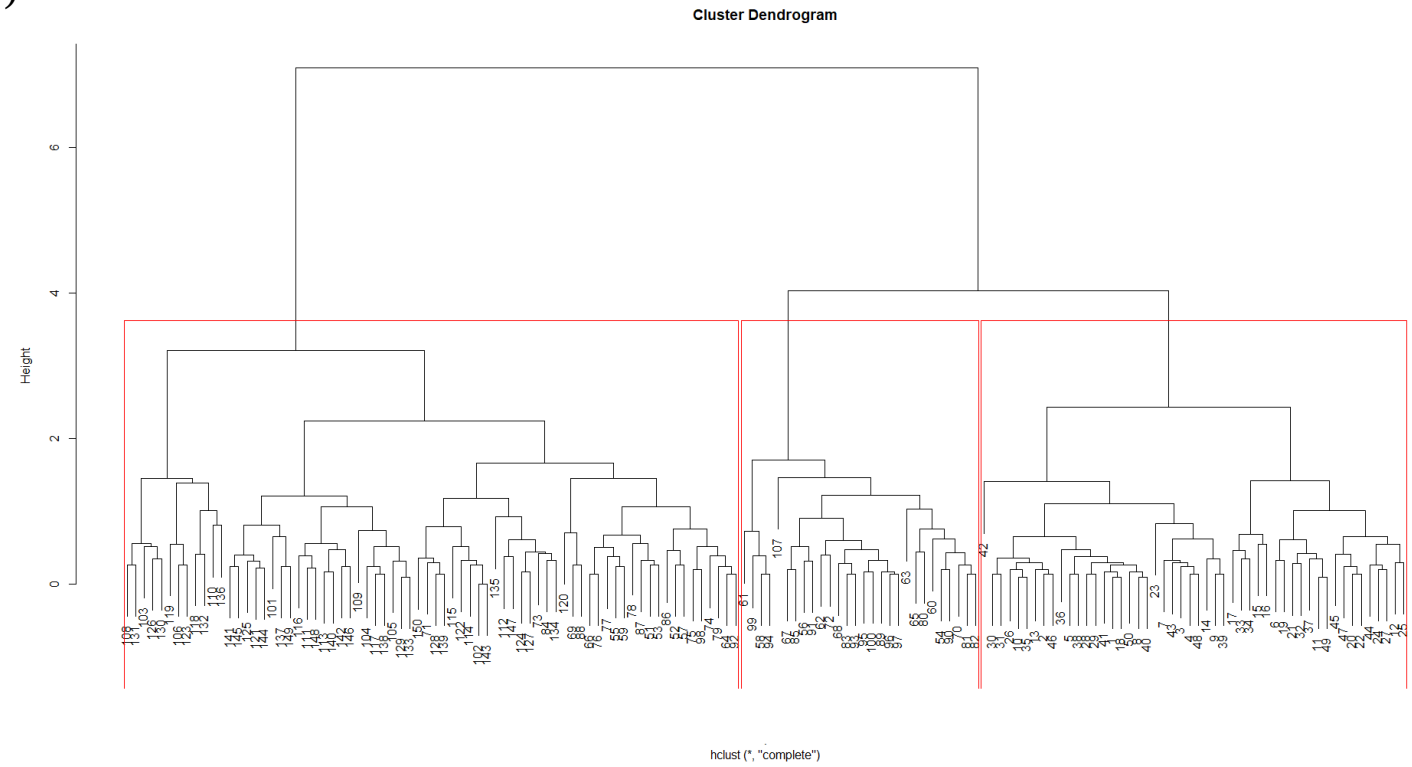
Advanced

- ggplot()

HIERARCHICAL CLUSTERING

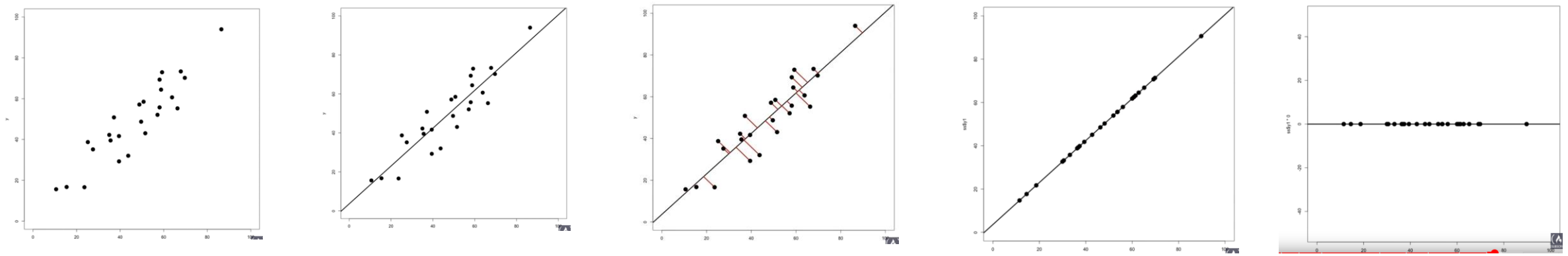
- Which cases are like others.
  - Hierarchical or a given number of clusters (k)
  - Measures of distance
  - Agglomerative or Divisive

- hclust()



PRINCIPAL COMPONENT ANALYSIS

- Used for dimensionality reduction
- prcomp()



Source: https://datalab.cc/tools/r01

REGRESSION

- Linear Regression
  - Step-wise
  - Lasso
  - And many more

- lm()
- glm()

HYPOTHESIS TESTING

- Using the Student's T-test in R
  - Two-Sample T-test with Unequal Variance
  - One-Sample T-testing in R
  - Using Directional Hypotheses in R
- μ-test in R
  - Two-Sample μ-test in R
  - One-Sample μ-test in R
- Correlation and Covariance in R
  - Simple Correlation in R
  - Covariance in R
  - Significance Testing in Correlation Tests

Further learning:

An Introduction to Statistical Learning

with Applications in R. ***https://statlearning.com/book.html***

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

Tutorial Point: ***https://www.tutorialspoint.com/r/index.htm***

Data Analytics Tutorial: ***https://data-flair.training/blogs/data-analytics-tutorial/***

More opportunities for learning:

rstudio::global → https://global.rstudio.com/student/catalog

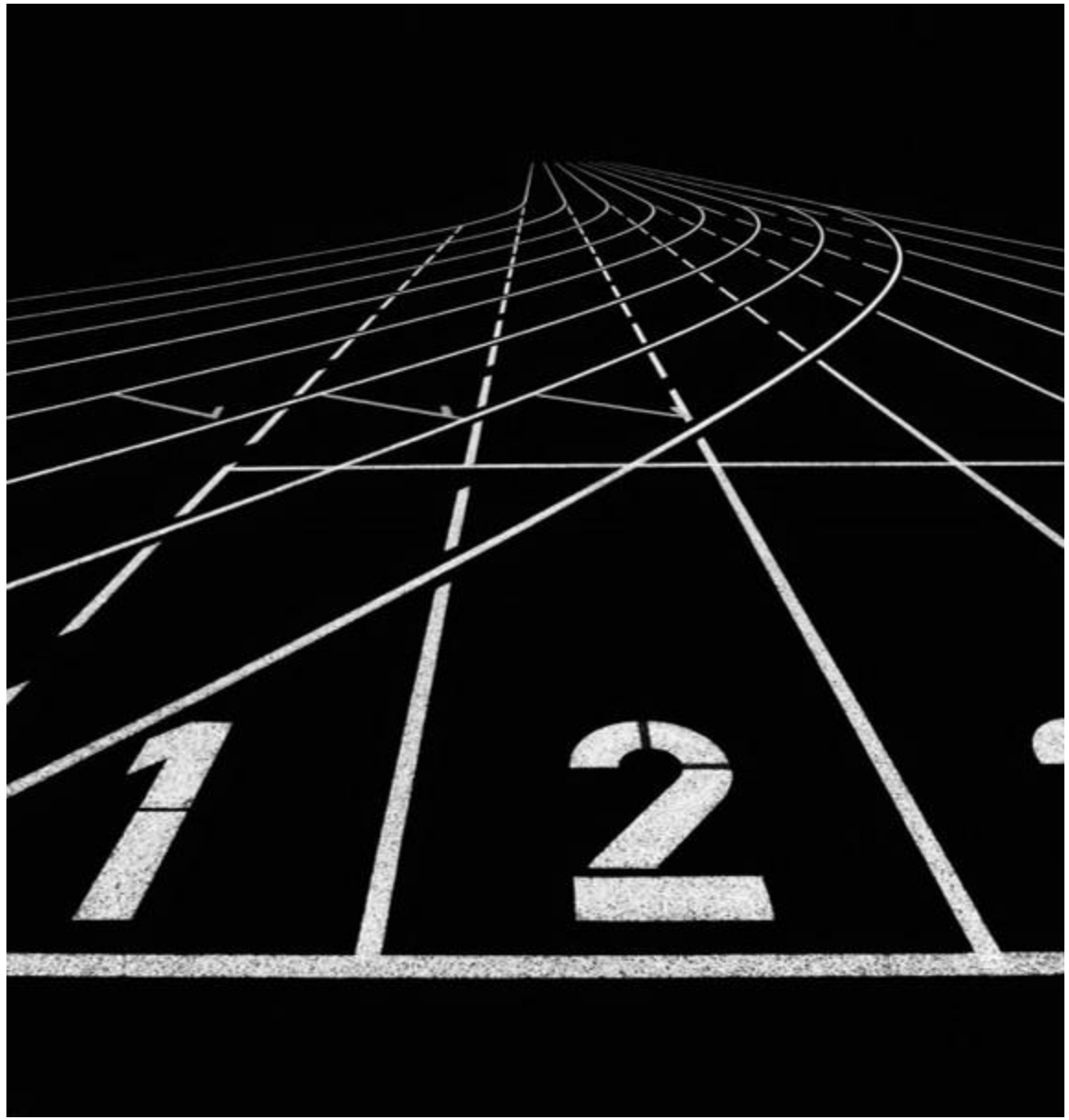
useR! Annual conference: The R User Conference

R User Group

My Github public repo for resources:

https://github.com/sreekanthvk/Statistical-Analysis-Using-R

.



THANK YOU

Dr. Sreekanth V K +91 94747 01658 | +91 98959 25259 sreekanth@nitc.ac.in | sreekanth.vk@outlook.com

@sreekanth.vk:matrix.org (element) https://www.linkedin.com/in/sreekanthvk https://github.com/sreekanthvk