INTRODUCTION TO R FOR DATA SCIENCE PART I

##BEING A SHORT TRAINING MATERIAL USED AT OFFA R USERS GROUP MEETING ON 31ST OCTOBER, 2023 AT NO. 15 OLD IRRA GARAGE IRRA ROAD, OFFA, KWARA STATE, NIGERIA

PREPARED BY UDOKANG, ANIETIE EDEM (OGANIZER, ORUG) CHIF LECTURER, STATISTICS DEPARTMENT, THE FEDERAL POLYTECHNIC OFFA, NIGERIA

#Introduction Today we want to look at the world of data science, the different between data scientist and statistician and the role/use of R software in data science. Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge. The world today needs data science more than ever before because people prefer visualization than descriptions. The training will focus also on some aspects of data cleaning which mostly spent in data analysis. #What is R? R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R (<https://www.r-project.org/about.html>). GNU (pronounced guh-noo) is a Unix-like computer operating system developed by the GNU project, Free Software Foundation. R has so many packages that can be installed in R. These packages can be found in Comprehensive R Archive Network (CRAN) repository. It’s a huge repository of R packages that users can easily contribute to.

#Why R? The following points is why R is attractive. *i. R is mainly used when the data analysis tasks require standalone computing or analysis on individual servers.* ii. R focuses on better, user-friendly data analysis, statistics and graphical models. *iii. R has been used primarily in academics and research. However, it’s rapidly expanding into the enterprise market.* iv. Statistical models can be written with only a few lines in R and the same piece of functionality can be written in several ways in R. \*v. Once you know the basics, you can easily learn advanced techniques and R is not hard for experienced programmers.

#How to download R and its Packages Visit <https://cran.r-project.org/> to download and install R version of your choice free of charge. Open the installed R and go to Packages - Install Package(s) – Select CRAN Mirror – Select Package to install and upload. #The R Environment R software environment is used to run code written in R language. This made up of R console and R editor. Go to file – New script to get R-Editor

#R Code The code can be written on the R console or run from R editor or copy and paste. #What is Data Science? This is a process of collecting/recording, storing and analyzing of data, mostly big data, from sources like social media, internet, satellite images, e-commerce sites, healthcare surveys and develop methods to effectively extract useful information for decision making on real life situation. Therefore, data science can also be described as “The application of data-centric computational and inferential thinking to understand the world and solve problems” according to Professor Joseph Gonzalez of University of California, Berkeley.

#What are the Difference between Statisticians and Data Scientists? The audience today are mostly statisticians or familiar with statistical analysis. Let me briefly discuss the difference between statisticians and data scientists.

#Statisticians  
*Deal with small-scale data collection using methods such as surveys, polls, and experiments.* Work on improving one simple model to best fit the data. *Only analysis data.*  
*#Data Scientist* Work on massive data (big data) which involves a lot of time in tasks like large-scale data collection and cleaning. *Try out different methods to create machine learning models, and then they choose the method that results in the best model.* Go beyond data analysis to implement algorithms that process data automatically. This enables data scientists to provide automated predictions and actions. An example is automation that helps with weather forecasts The critical stage of data science is data cleaning. #What is Data Cleaning? When data is collected mostly in raw form it has to be arranged/transformed to have a reasonable structure that can be meaningful and useful. Data cleaning usually takes care of some of the issues as the missing values, the formatting of values, the structure of the data, extracting information from complex values and unit conversion. One of the R packages that handles this aspect is tidyr. To use the package in R, use the library function.

#Creating data frame (df) in R

X<-c(1,3,4,5,6,8,10,12,5,8,9,11)  
Y<-c(2,3,6,8,9,12,23,21,54,30,44,13)  
df<-data.frame(X,Y)  
df

## X Y  
## 1 1 2  
## 2 3 3  
## 3 4 6  
## 4 5 8  
## 5 6 9  
## 6 8 12  
## 7 10 23  
## 8 12 21  
## 9 5 54  
## 10 8 30  
## 11 9 44  
## 12 11 13

#OR

X=c(1,3,4,5,6,8,10,12,5,8,9,11)  
Y=c(2,3,6,8,9,12,23,21,54,30,44,13)  
df<-data.frame(X,Y)  
df

## X Y  
## 1 1 2  
## 2 3 3  
## 3 4 6  
## 4 5 8  
## 5 6 9  
## 6 8 12  
## 7 10 23  
## 8 12 21  
## 9 5 54  
## 10 8 30  
## 11 9 44  
## 12 11 13

#To be sure df is a data frame

print(class(df))

## [1] "data.frame"

print(df)

## X Y  
## 1 1 2  
## 2 3 3  
## 3 4 6  
## 4 5 8  
## 5 6 9  
## 6 8 12  
## 7 10 23  
## 8 12 21  
## 9 5 54  
## 10 8 30  
## 11 9 44  
## 12 11 13

#More examples of data frame

x<-c("a","b","c","d","e","f","g","h","I","j","k")  
f<-c(2,3,5,2,1,4,3,2,5,7,8)  
df<-data.frame(x,f)  
df

## x f  
## 1 a 2  
## 2 b 3  
## 3 c 5  
## 4 d 2  
## 5 e 1  
## 6 f 4  
## 7 g 3  
## 8 h 2  
## 9 I 5  
## 10 j 7  
## 11 k 8

x<-c("a","b","c","d","e","f","g","h","I","j","k")  
f.<-c(2,3,5,2,1,4,3,2,5,7,8)  
y<-c("Ben","Philip","Segun","Ani","Tom","Okon","Jide","Chidi","Betty","Bros","Ken")  
f..<-c(21,23,45,26,19,46,43,62,52,71,88)  
df<-data.frame(x,f.,y,f..)  
df

## x f. y f..  
## 1 a 2 Ben 21  
## 2 b 3 Philip 23  
## 3 c 5 Segun 45  
## 4 d 2 Ani 26  
## 5 e 1 Tom 19  
## 6 f 4 Okon 46  
## 7 g 3 Jide 43  
## 8 h 2 Chidi 62  
## 9 I 5 Betty 52  
## 10 j 7 Bros 71  
## 11 k 8 Ken 88

print(df)

## x f. y f..  
## 1 a 2 Ben 21  
## 2 b 3 Philip 23  
## 3 c 5 Segun 45  
## 4 d 2 Ani 26  
## 5 e 1 Tom 19  
## 6 f 4 Okon 46  
## 7 g 3 Jide 43  
## 8 h 2 Chidi 62  
## 9 I 5 Betty 52  
## 10 j 7 Bros 71  
## 11 k 8 Ken 88

df<-data.frame(rating=1:11,x,f.,y,f..)  
df

## rating x f. y f..  
## 1 1 a 2 Ben 21  
## 2 2 b 3 Philip 23  
## 3 3 c 5 Segun 45  
## 4 4 d 2 Ani 26  
## 5 5 e 1 Tom 19  
## 6 6 f 4 Okon 46  
## 7 7 g 3 Jide 43  
## 8 8 h 2 Chidi 62  
## 9 9 I 5 Betty 52  
## 10 10 j 7 Bros 71  
## 11 11 k 8 Ken 88

#Structuring of the data into character/variable and corresponding values

library(tidyr)  
xf<- tibble(x = c("x1", "x2","x3","x4"),f1 = c(1, 4, 2, 5), f2 = c(3, 4, 6, 2),)  
xf

## # A tibble: 4 × 3  
## x f1 f2  
## <chr> <dbl> <dbl>  
## 1 x1 1 3  
## 2 x2 4 4  
## 3 x3 2 6  
## 4 x4 5 2

xf is the data frame (df)

#Removing missing data

xf<- tibble(x = c("x1", NA,"x3","x4"),f = c(1, NA, 2, 5),)  
xf

## # A tibble: 4 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 <NA> NA  
## 3 x3 2  
## 4 x4 5

drop\_na(xf)

## # A tibble: 3 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 x3 2  
## 3 x4 5

#Replacing missing value with specified value

xf<- tibble(x = c("x1", NA,"x3","x4"),f = c(1, NA, 2, 5),)  
xf %>% replace\_na(list(x = "x2", f = 4))

## # A tibble: 4 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 x2 4  
## 3 x3 2  
## 4 x4 5

#Filling missing values up with the preceding values

xf<-tibble(x=c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10"),f=c(1,4,2, NA,NA,6,7,9,NA,NA),)  
xf

## # A tibble: 10 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 x2 4  
## 3 x3 2  
## 4 x4 NA  
## 5 x5 NA  
## 6 x6 6  
## 7 x7 7  
## 8 x8 9  
## 9 x9 NA  
## 10 x10 NA

xf1 <- xf %>% fill(f, .direction = 'up')  
xf1

## # A tibble: 10 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 x2 4  
## 3 x3 2  
## 4 x4 6  
## 5 x5 6  
## 6 x6 6  
## 7 x7 7  
## 8 x8 9  
## 9 x9 NA  
## 10 x10 NA

#Filling missing values down with the preceding values

xf2 <- xf1 %>% fill(f, .direction = 'down')  
xf2

## # A tibble: 10 × 2  
## x f  
## <chr> <dbl>  
## 1 x1 1  
## 2 x2 4  
## 3 x3 2  
## 4 x4 6  
## 5 x5 6  
## 6 x6 6  
## 7 x7 7  
## 8 x8 9  
## 9 x9 9  
## 10 x10 9

#Splitting columns using “separate” function

GP<- tibble(id = 1:2, x = c("m-360", "f-580"))  
GP

## # A tibble: 2 × 2  
## id x   
## <int> <chr>  
## 1 1 m-360  
## 2 2 f-580

#GP is the data frame (df) representing population by gender

df %>% separate(x, c("gender", "unit"))

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 11 rows [1, 2, 3, 4, 5,  
## 6, 7, 8, 9, 10, 11].

## rating gender unit f. y f..  
## 1 1 a <NA> 2 Ben 21  
## 2 2 b <NA> 3 Philip 23  
## 3 3 c <NA> 5 Segun 45  
## 4 4 d <NA> 2 Ani 26  
## 5 5 e <NA> 1 Tom 19  
## 6 6 f <NA> 4 Okon 46  
## 7 7 g <NA> 3 Jide 43  
## 8 8 h <NA> 2 Chidi 62  
## 9 9 I <NA> 5 Betty 52  
## 10 10 j <NA> 7 Bros 71  
## 11 11 k <NA> 8 Ken 88

#Splitting columns using “strsplit” function

Students<- data.frame(Programme = c("ND Statistics", "HND Computer", "NID Welding", "BTECH Civil"), Pop = c(120, 200, 180, 78) )  
Students

## Programme Pop  
## 1 ND Statistics 120  
## 2 HND Computer 200  
## 3 NID Welding 180  
## 4 BTECH Civil 78

Students[c("Programme","Department")]<- do.call(rbind, strsplit(Students$Programme, " "))  
print(Students)

## Programme Pop Department  
## 1 ND 120 Statistics  
## 2 HND 200 Computer  
## 3 NID 180 Welding  
## 4 BTECH 78 Civil

#Uniting columns using “unite” function

x<-c("B","P","S","A","T","O","J","C","BT","BR","K")  
y<-c("Ben","Philip","Segun","Ani","Tom","Okon","Jide","Chidi","Betty","Bros","Ken")  
df<-data.frame(x,y)  
df

## x y  
## 1 B Ben  
## 2 P Philip  
## 3 S Segun  
## 4 A Ani  
## 5 T Tom  
## 6 O Okon  
## 7 J Jide  
## 8 C Chidi  
## 9 BT Betty  
## 10 BR Bros  
## 11 K Ken

z<-data.frame(x=c("B","P","S","A","T","O","J","C","BT","BR","K"),y=c("Ben","Philip","Segun","Ani","Tom","Okon","Jide","Chidi","Betty","Bros","Ken"))  
df %>% unite("z", x:y, remove = FALSE)

## z x y  
## 1 B\_Ben B Ben  
## 2 P\_Philip P Philip  
## 3 S\_Segun S Segun  
## 4 A\_Ani A Ani  
## 5 T\_Tom T Tom  
## 6 O\_Okon O Okon  
## 7 J\_Jide J Jide  
## 8 C\_Chidi C Chidi  
## 9 BT\_Betty BT Betty  
## 10 BR\_Bros BR Bros  
## 11 K\_Ken K Ken

df %>% unite("z", x:y, remove = TRUE)

## z  
## 1 B\_Ben  
## 2 P\_Philip  
## 3 S\_Segun  
## 4 A\_Ani  
## 5 T\_Tom  
## 6 O\_Okon  
## 7 J\_Jide  
## 8 C\_Chidi  
## 9 BT\_Betty  
## 10 BR\_Bros  
## 11 K\_Ken

#We appreciate your patient and interest in participating in the discussion today. The Offa-R-Users-Group (ORUG) is a place to learn and share knowledge in the use of R. I wish to see you next time. If you are a guest, find time to register as a member to actualize your goal in using R. The ORUG (<https://www.meetup.com/fedpofa-r-users-group/> ) is sponsored by R Consortium and AniKem\_Consult. For any Enquiry Contact the Organizer (WhatsApp: +2349030912602, email: [anietieeu@yahoo.com](mailto:anietieeu@yahoo.com))

#THANK YOU FOR PARTICIPATING IN THIS EVENT