DATA SCIENCE: AN OVERVIEW IN R

##DATA SCIENCE: AN OVERVIEW IN R BEING AN ONLINE SHORT PRESENTATION IN OFFA-R-USERS-GROUP (ORUG) MEETING ON 31ST MAY, 2023 BY UDOKANG, ANIETIE EDEM (OGANIZER, ORUG) CHIF LECTURER, STATISTICS DEPARTMENT, THE FEDERAL POLYTECHNIC OFFA, NIGERIA

##Introduction  Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.  People prefer visualization to descriptions.  Therefore, installing R to your systems shall be done today so that gradually we will get to some applications in subsequent meetings. ##What is Data Science?  This is a process of collecting/recording, storing and analyzing of big data.  Sources of data: Social media, Internet, Satellite images, e-commerce sites, healthcare surveys  Develop methods to effectively extract useful information for decision making on real life situation ##What are the Difference between Statisticians and Data Scientists? #Statisticians  
*Deal with small-scale data.* Work on improving one simple model to best fit the data.  
Only analysis data. #Data Scientist Work on massive data (big data). *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. 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 • The missing values • The formatting of values • The structure of the data overall • Extracting information from complex values • Unit conversion.

An important tool of data science is data visualization.

##What is Data Visualization? *This is a way of bringing data to life that can convey trends and anomalies in the data more efficiently than a written description.* Data visualization is a great way to communicate your predictions and conclusions to other people by using a useful tool as R software and its packages. \*Data visualization can be described as the graphical representation of information and data. By using visual elements like charts, tables, graphs, maps, infographics and dashboards. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

##Examples of Data Visualization: \*Examples in three categories; • Time series data visualization • Interactive data visualization • Static data visualization. ##Examples of Time Series Data Visualisation •Line Chart •Bar Chart •Scatter Chart •Area Chart •Map •Indicator • Pivot Table • Bullet Graph • Box plot • Matrix

##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 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.

#Some R Packages for Data Cleaning: • The Plyr Package • The Stringr Package • The tidyr package • The sqldf package • The janitor package • The splitstackshape package

#Some R Packages for Data Visualization: • Colourpicker • Esquisse • ggplot2 • ggvis • ggforce • Lattice • Plotly • Patchwork • Quantmod • RGL

#TODAY WE HAVE LEARNT ABOUT DATA SCIENCE WITH R AND WHAT WE CAN USE TO ACHIEVE SUCH AS DATA CLEANING AND VISUALISATION. IN SUBSEQUENT MEETINGS WE WILL USE THE R PACKAGES TO RUN CODES TO ACTUALIZE WHAT WE LEARNT. #THANK YOU AND HAVE A NICE DAY