

Data Science Orientation for Engineering Students

Teaching with Open Source Software- R

Norman Gwangwava

Mechanical, Energy and Industrial Engineering
Botswana International University of Science and
Technology
Palapye, Botswana
gwanhwavan@biust.ac.bw

Richard Addo-Tenkorang

Mechanical, Energy and Industrial Engineering
Botswana International University of Science and
Technology
Palapye, Botswana
addotenkorangr@biust.ac.bw

Abstract— This paper highlights uses of R, an open source data analysis software and programming language, in teaching an engineering module (Quality Control and Management). The main objective is to orient engineering students to the booming and lucrative data science career. Data science in engineering is mainly motivated by the need for operations improvement, supply chain management, and personalised production (a.k.a mass customization). The use of R in the mentioned module was therefore very appropriate since the learning objectives are mainly centred on the three areas of data science applications in engineering. The main achievements realised in the module were a 100% semester pass rate. Students were given an online survey questionnaire to solicit data pertaining to their views in using R as platform for introducing data science training. The responses generally indicated average to high satisfaction, practical exposure to statistical quality control tools and data analysis. Students also recorded a relatively high ease of learning of the R system and readiness to tackle practical industry problems. The results of the pilot project inspired future research into industry based applications of R for quality control and management as well as other data analytics demands. The lab sessions during the semester were mainly based on command prompt programming, but industrialists mainly find comfort in graphical based tools and real-time integration with factory floor operations.

Keywords—Big Data; Internet of Things; Data Analytics; Data Science; R Software.

I. INTRODUCTION

Business analytics on demand has become mandatory practice of the modern day corporate world. This has significantly increased the demand for accurate data collection from baseline operations in any business setup. Data and information has long become a central productive and strategic asset, and the success of the organization depends on its ability to gather, produce, maintain and disseminate this information for its benefit and those in its value chain. The engineering field collects huge volumes of data (Big Data) through RFID readers, tags and sensors, and the trend continues to grow explosively as technology advances. The integrated technologies for collecting data are generally called Internet of Things (IoT). Most powerful applications of Big Data and Data science will likely emerge when data from multiple sources is integrated, e.g., material, process, manufacturing, and business

data. The best quality data is usually generated by sensors, which are often arranged in networks. Sensors and Sensor Networks – Sensors resonate in a twofold manner within the areas of research and practice in the field of engineering, particularly mechanical. Many products and systems designed by engineers utilize sensors. Markets call for products and solutions that are more reliable, safer, and more dependable. This in turn creates demand for new and more accurate sensors, which measure parameters that only a few years ago were merely dreamed about. Progress in computing and data science is another factor fueling progress in sensor technology, as the data generated by sensors adds value to products. Engineering training, as well as other scientific disciplines have been emphasising fundamental technical concepts that form the body of knowledge of the specific professional fields. The emanating missing link is now the disparity between technological advancement to collect data about things versus exploitation of the available data to derive value or meaningful information. Data Science skills have since become a mandatory part of the modern day engineering curriculum. The skills are of a multi-disciplinary domain, anchored on computer science, mathematics, human-technology interaction, and business models and operations management. Most of the skills are already incorporated in the traditional curriculum except state-of-the-art machine learning and data analytics.

II. LITERATURE

A. Data Science

Data science is defined as the study of the generalizable extraction of knowledge from data. More people now consider a data scientist as a job role in the big data realm and is different from a statistician or other job roles [1, 2]. Data scientists are also among the most critical skills of the 21st century. Data science is different for conventional science in the sense that “science” extracts knowledge from experiment data whereas data science. Today’s data is mainly in the form of digital traces produced in the regular course of work or play (i.e., data generated or managed by operational support (OS) tools) and is known as operational data (OD). Operational data

has no carefully designed measurement systems. OD exhibits the following attributes;

- Prevalent
 - Massive data from software development
 - Increasingly used in practice
 - Many activities transitioning to a digital domain
- Treacherous - unlike experimental data
 - Multiple contexts
 - Missing events
 - Incorrect, filtered, or tampered with
- Continuously changing
 - OS systems and practices are evolving
 - New OS tools are being introduced in SE and beyond
 - Other domains are introducing similar tools

B. Big Data

Big Data refers to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a reasonable time. Another definition refers to Big Data as the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities – according to zdnet.com. Big data is not just about size. It finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. Big data analytics (BDA) is defined as a holistic approach to manage, process and analyze the “5 Vs” data-related dimensions (i.e., volume, variety, velocity, veracity and value) in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages [3]. It has recently emerged as “the next big thing” in management. Some scholars and practitioners even suggest that BDA is the “fourth paradigm of science” [4, p.34], or even “the next frontier for innovation, competition, and productivity” [5, p.1], or the “new paradigm of knowledge assets” [6, p. 2]. These statements are mainly driven by the pervasive adoption and use of various tools and technologies, including social media (e.g., Facebook, Twitter), mobile devices (e.g., laptops, smartphones), automatic identification technologies enabling the Internet of Things (IoT) (e.g., radio frequency identification (RFID), Bluetooth), and cloud-enabled platforms to support intra- and inter-organizational business processes and achieve a competitive advantage. Some analysts estimate that Twitter users generate more than 250 million tweets per day, while about 50 hours of video are uploaded each minute on YouTube from around the world. The same analysts estimate that Facebook now holds more than 90 billion photos, with over 200 million photos uploaded per day [7].

C. Tools for Big Data and Data Scienc

Python

Python is a powerful, flexible, open-source language that is easy to learn, easy to use, and has powerful libraries for data manipulation and analysis. Its simple syntax is very accessible

to programming novices, and will look familiar to anyone with experience in Mat lab, C/C++, Java, or Visual Basic. For over a decade, Python has been used in scientific computing and highly quantitative domains such as finance, oil and gas, physics, and signal processing. Python is one of the most popular programming languages in the world, ranking higher than Perl, Ruby, and JavaScript by a wide margin. Among modern languages, its agility and the productivity of Python-based solutions are legendary. The future of python depends on how many service providers allow for SDKs in python and also the extent to which python modules expand the portfolio of python apps.

R

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians for developing statistical software and data analysis. According to a survey in 2010, R was the data mining tool used by more data miners (43%) than any other. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity. R is emerging as a defacto standard for computational statistics and predictive analytics. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- An effective data handling and storage facility.
- A suite of operators for calculations on arrays, in particular matrices.
- A large, coherent, integrated collection of intermediate tools for data analysis.
- Graphical facilities for data analysis and display either on-screen or on hardcopy.
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Hadoop

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters.

Currently Hadoop is the standard of many organization in dealing with Big Data [11]. As several technologies associated to Big Data are open source, vendors are showing their interest in integrating these tools with their own products [12]. The Apache Hadoop, an open source framework designed to support distributed parallel processing of large amount of data sets across clusters of computers using simple programming models. It is written in java. It can run on commodity hardware, scaling up from single nodes to thousands of computer, thus forming cluster. Each node in the cluster offers local computation and storage. Hadoop promise high availability to end user. Instead of depending on hardware to

deliver high-availability, its framework itself can identify and figure out single point of failures at the application layer, thus providing a high availability service to its user [8]. It has a number of commercially supported distributions from companies such as Map Technologies, Hortonworks, Cloudera, etc [9].

D. Education and Training Trends in Data Science

There are a wide range of knowledge and skills, or the so called "data scientist's tool box," being recognized as essential for the field. A variety of requirements range from statistics, computer science [5, 7], and business knowledge to creativity, passion [10], and patience.

A poll, conducted by Gallup for the Business-Higher Education Forum in America, revealed that by 2021, 69% of employers expect candidates with data science analytics (DSA) skills to get preference for jobs in their organizations. Yet only 23% of college and university leaders say their graduates will have those skills. Digital infrastructure is being built out in public and private spaces across the US. With it, our capacity to generate and transmit data from and across devices, entities, systems, and sensor networks (embedded in the Internet of Things) is growing exponentially. Capacity, however, is only one of the essential enablers of digitization. Capability is essential, too, and that resides in the skills of people. The market for DSA skills is hot, but it's full of mismatches. There is a growing number of DSA degrees and credentials—since 2010, 303 new accredited DSA programs were launched in the US, a 52% increase overall [13].

There is a far deeper pool of STEM talent, but here, too, it's often unclear how well prepared these job candidates are to use data science and analytics in business pursuits. Meanwhile, business schools have very few programs that include DSA coursework.

DSA is a catalyst for change and innovation within universities not just because it is creating new professional disciplines, but also because it is altering so many established disciplines. As such, DSA can be seen as part of a broader set of trends that are challenging colleges and universities to:

- Change curriculums to meet skill needs
- Structure degrees and credentials for individual paths
- Provide a high-quality education at a reasonable cost
- Partner more frequently with employers
- Invest in their reputations for applied research

Some common trends to curb DSA skills shortage undertaken in the education sector include University-branded MOOCs, bootcamps, and hackathons. These help universities meet some of the challenges,

III. METHODOLOGY

Fundamentals of data science were first incorporated into an already existing semester module for Quality control and management- INDE 312. A course satisfaction survey was then conducted at the end of the semester among all the students who took the module. R platform was chosen as the data

science tool throughout the semester. The survey was aimed at assessing individual perspectives on using R for data science and the actual learning impact on the module.

A. Learning in SPC with R

The survey question sought to gather the depth of understanding on the students on various statistical process control (SPC) tools. The students were asked to rate their understanding on a scale of 1 (not at all) – 5 (very well), due to the use R during lab sessions. The SPC tools that students were exposed to during the semester are;

- a. histogram or stem-and-leaf plot
- b. pareto chart
- c. cause-and-effect diagram
- d. scatter diagram
- e. variables control charts
- f. attributes control charts
- g. process capability analysis

85% of the students revealed a "moderate to very well" understanding of the concepts, enhanced by using R during lab sessions.

B. Data analysis and programming

This inquiry sought to understand the level of data analysis and programming skills that students acquired during the course of the semester as a result of using R. The ranking scale is in the range of 1(nothing) – 5(great deal). The following specific questions follow under this category.

- a. solving quality control problems
- b. writing quality data analysis reports
- c. writing R code to analyse statistical data
- d. working effectively with others

70% responses indicated a moderate to great deal of data analysis and programming skills gained in using R during the course of the semester.

C. Understanding the R Data Science Tool

The survey question sought to evaluate the students' overall understanding and mastery of R. The ranking scale is between 1 (not at all) - 5 (great deal). The category of the survey has many part questions as follows;

- a. good understanding of the main concepts of R and its applications in SPC
- b. understanding the relationship between SPC concepts
- d. understanding of the relevance of R package to real world quality control issues
- e. understanding of the nature of SPC
- f. appreciation of the methods of SPC
- g. ability to think through a problem or argument
- h. confidence in your ability to do statistical process control
- i. feeling comfortable with complex problems
- j. enthusiasm for statistical process control (SPC)

D. Industry Relevance

Students were asked, "How much of what you learned in this class (i.e., knowledge, skills, and other gains) do you think you will REMEMBER and CARRY WITH YOU into other classes or other aspects of your life?" The scale for rating ranged from 1 (nothing) – 5 (great deal). 80% of the class had responses from moderate and great deal, whereas 20% indicated that little of the learned aspects will be passed on to real life situations.

Generally students proved that they gained a great deal of knowledge about data science practice, using R. They showed readiness to take the skills acquired to industry practice.

IV. CONCLUSION

Data science skills are becoming more prominent in the modern day industry practice and the business fraternity. There is a call upon education institutions to be pragmatic in promoting data science oriented curriculum. The good news is that most platforms for data science and analytics are open source, which takes off the huge investment burden on education institutions. The main role demanded now is to invest in curriculum review in order to stay industry relevant. The pilot implementation of fundamental data science skills into an existing module proved to be very relevant and welcome by students. Using R during lab sessions drew interest and continuous engagement upon students, which is a very positive development in linking education and industry practice. Other strategies are needed to further reduce the gap between traditional qualifications oriented education and acquiring skills for industry job practice.

REFERENCES

- [1] Eckerson, W. (2011) "BigDataAnalytics: Profiling the Use of Analytical Platforms in User Organizations," TDWI, September. Available at <http://tdwi.org/login/default-login.aspx?src=%7bC26074AC-998F-431B-BC994C39EA400F4F%7d&qstring=tc%3dassetpg>
- [2] "Research in Big Data and Analytics: An Overview" International Journal of Computer Applications (0975 –8887) Volume 108 –No 14, December 2014
- [3] Fosso Wamba, S., et al., How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 2015. 0(0): p. xx-xx.
- [4] Strawn, G.O., Scientific Research: How Many Paradigms? EDUCAUSE Review, 2012. 47(3): p. 26.
- [5] Manyika, J., et al., Big data: the next frontier for innovation, competition and productivity, 2011, McKinsey Global Institute.
- [6] Hagstrom, M., High-performance analytics fuels innovation and inclusive growth: Use big data, hyperconnectivity and speed to intelligence to get true value in the digital economy. Journal of Advanced Analytics, 2012(2): p. 3-4.
- [7] Sethuraman, M.S., Big Data's Impact on the Data Supply Chain, in Cognizant 20-20 Insights2012, Cognizant: New Jersey.
- [8] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.
- [9] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," interactions, vol. 19, no. 3, pp. 50–59, May 2012
- [10] Ari Banerjee senior analyst, heavy reading, "Big data and advanced analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity," December 2013.
- [11] Hadoop Wiki,"Apache Hadoop", Accessed : <http://wiki.apache.org/hadoop>
- [12] Sheikh Ikhlaq and Dr. Bright Keswani," Computation of Big Data in Hadoop and Cloud Environment ",IOSR Journal of Engineering (IOSRJEN), Vol. 06, Issue 01 (January. 2016)
- [13] Tableau, The State of Data Education 2016 (November 2016)