PROMOTING AN INTEREST IN DATA SCIENCE THROUGH A SERIES OF ONLINE WORKSHOPS

Saleha Naghmi Habibullah
Kinnaird College For Women, Lahore, Pakistan
saleha.habibullah@kinnaird.edu.pk

In today's data-rich world, developing countries cannot afford to remain away from the study and practice of data science. This paper presents an account of a series of online workshops initiated by the author in June 2020 from the forum of the Pak Institute of Statistical Training and Research (PISTAR), Lahore, Pakistan, the focus of which has so far been on topics such as 'Tidyverse', 'Machine Learning' and 'Data Science Ideas for Aspiring Statistical Consultants'. The once-a-month online workshop series is exciting for the organizers and the participants alike as it is able to attract participants from various continents of the world, particularly from the developing countries. The congenial atmosphere and high-quality content of the workshops have led to confidence-building and trust that has prompted the launch of preparatory sessions which, together with the workshops, are serving as a vehicle for promoting interest in data science.

INTRODUCTION

In today's world where huge amounts of data are accumulating by the minute, data science is acquiring ever-increasing importance and, in the Western countries, many aspire to become data scientists. However, the situation of the developing countries is very different and there is a need to promote interest in data science as developing countries cannot afford to remain oblivious of the importance of data science. This can be achieved through courses in university programs of study, seminars, webinars, training workshops and other such activities. Hardin et al. (2015) provide a detailed account of the reasons why data science should be incorporated in the statistics curriculum and recommend participation in relevant workshops and seminars to be attended by faculty who are desirous of developing their pedagogical abilities in data science. The authors refer to the 3'd Computation and Visualization Consortium Workshop and NSF Workshops through the UC Davis Data Sciences Initiatives as examples of such training programs.

Training workshops on data science have been held in some of the developing countries too. The International Training Workshop for Developing Countries on Big Data for Science was organized in Beijing, China, in June 2014. Participants included researchers, data managers, and data scientists from ten developing countries including Brazil, Colombia, South Africa, Kenya, Tanzania, Uganda, India, Mongolia, Vietnam and Indonesia. The workshop was jointly organized by CODATA Secretariat, Computer Network Information Center (CNIC) and Chinese National Committee for CODATA (CODATA-China). The objective of the training program was to engage participants with a number of aspects of data science and data management in the era of Big Data. Topics included interdisciplinary applications of data-intensive research, data management policies, cloud computing, visualization and data infrastructure development in the Big Data Age.

Treacy, S (2019) presented an account of an event organized by the TWAS Young Affiliates Network (TYAN), supported by the Elsevier Foundation and held in Akure, Nigeria on 10-14 June, 2019. The event included a symposium with talks by several international experts, and culminated in a workshop in which African researchers received hands-on training in big data and machine learning techniques. The event was created to help Africa adapt and take advantage of an important shift in science: big data and intelligence techniques such as machine learning that are enabling scientists to turn enormous amounts of data into discoveries that have thus far been out-of-reach. More than 70 researchers from 14 different countries (eight of whom were African) participated in the event.

Theobolda, Hancockb and Mannheimerc (2021) describe a research study conducted by them on the design and implementation of a set of data science workshops primarily for environmental science graduate students but also open to the larger academic community. These workshops promote continued learning of the computational tools necessary for working with data, and provide resources for bringing data science into the classroom.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

This paper presents an account of the workshop-series launched in June 2020 from the forum of PISTAR. Ten workshops were completed by April 2021, the first five on Tidyverse, the next four on Machine Learning, and the tenth one on pivot tables in Excel. The months from May to August 2021 were allocated to a set of five workshops on 'Data Science Ideas for Aspiring Statistical Consultants'. The once-a-month PISTAR online workshop series attracts not only national but also international colleagues from various countries of the world, particularly the developing countries and, all workshops held so far being linked with data science in one way or the other, PISTAR is playing a significant role in promoting interest in this emerging discipline among academic and professional statisticians of Pakistan as well as of other developing countries.

PISTAR ONLINE WORKSHOP SERIES

Commencement

The COVID-19 pandemic has had a huge impact on the world and, among many other difficulties, normal educational activities have been hampered. However, with the conviction that learning must go on, in June 2020, the author endeavored to commence a series of online workshops in her capacity as Honorary Director of the Pak Institute of Statistical Training and Research (PISTAR), Lahore, Pakistan.

COVID-19 had hit Pakistan in March 2020 when, all of a sudden, lockdown was announced and educational institutions at all levels were closed down. Whereas the first two/three weeks were those of 'disorientation' and bewilderment, in April/May 2020, some of the leading higher education institutions of the country including Kinnaird College For Women Lahore (KC) started taking their first steps toward the online mode of instruction. At Kinnaird, the month of May was reserved for preparation, and the first day of online classes was 1st June 2020. For a huge majority of the faculty-members, this was their very first experience with online teaching.

It was a matter of one to two days that it occurred to the author that if her KC classes could be conducted online, then the PISTAR training workshops could also be conducted in the same manner. Within the next few hours, it was clear to her that if the workshops are to be online, then there is no need to restrict them to participants from within Pakistan. Emails and WhatsApp messages were sent out in large numbers as a result of which the first international online workshop materialized on June 27, 2020.

Five Workshops on Tidyverse

PISTAR Online Workshop No. 1 was held from 2:00 to 4:00 pm on Saturday, 27th June 2020. The topic of the workshop was "Tidyverse ggplot2" and it was conducted by Prof. Dr. Saleha Naghmi Habibullah, Dr. Sharqa Hashmi and Ms. Zulaikha Mashkoor (faculty-members of the Departments of Statistics at three different higher education institutions of Lahore). Dr. Syed Wasim Abbas of the Bureau of Statistics, Punjab acted as Focal Person and Ms. Kanwal Aslam, MPhil Statistics carried out the responsibilities of Rapporteur.

In addition to national participants, the workshop attracted four international participants, two from Kenya and two from Saudi Arabia. The resource persons conveyed to the participants contents of the workshop through PowerPoint slides and through live demonstration of Tidyverse in R-Studio. Subsequent to an introduction of the set of packages entitled Tidyverse, various functions contained in the ggplot2 such as geom_point(), geom_line(), geom_histogram(), geom_boxplot(), geom_density(), geom_bar(), etc. were introduced. Feedback was requested from the workshop participants in the form of a filled out questionnaire designed by the author. The participants were requested to provide feedback regarding the quality of the workshop.

Online Workshop No. 2 on Tidyverse dplyr was held from 2:00 to 4:30 pm on Saturday, 18th July 2020. The instructors were the same as in Workshop No. 1. The number of international participants increased to eight, four from Kenya, two from Saudi Arabia and two from Iran. The resource persons covered topics such as select(), filter(), logical tests, Boolean operators, arrange(), desc(), pipe operator, summarise(), mean(), max(), min(); summary functions, mutate(), round(), n(), n_distinct(), group_by(), ungroup(), slice(), data frame, tibble, tribble(); tibbles, min_rank, median(rank) and vectorized functions. The participants were also familiarized with the procedure by which they would be able to import their own data into R-Studio. Positive comments received from the participants were a source of satisfaction and encouragement for the author and her Team.

Online Workshops 3, 4 and 5 were held on August 15, September 12 and October 3, 2020 respectively, the topics being 'Tidyverse tidyr', 'Tidyverse Data types' and 'Tidyverse Modelr'. Topics covered include the definition of tidy data, functions such as gather(), spread(), View(), separate(), unite(), missing values, drop_na(), logicals, str_sub(), gss_cat(), levels(), fct_infreq(), fct_rev(), fct_recode(), fct_collapse(), fct_lump(), hms(), ymd(), simple linear regression: lm(), tidy(), glance(), augment(), multiple regression, geom_smooth(), add_predictions(), add_residuals(), spread_residuals() and gather_ residuals().

Prior to Online Workshop No. 3, it occurred to the author that it will be a good idea to invite the international participants to give presentations during the workshops. Mr. Mawora Thomas Mwakudisa from Kenya gave a presentation on an elaborate analysis of rainfall data collected by farmers over a period of one year in Western Kenya in Online Workshop No. 3. Dr. Najla Muhammad Qarmalah from Saudi Arabia gave an insightful talk on 'k-Boxplots for Mixture Data' in Online Workshop No. 4 and Dr. Reza Arabi Belaghi from Iran presented a talk on the utilization of a variety of Tidyverse functions on American Baseball data in Online Workshop No. 5.

Subsequent to the first two workshops, the author contemplated that the appeal of the workshops would be enhanced by inviting eminent statisticians from various countries of the world to render Invited Talks during the workshop sessions. Dr. David Stern (UK), Member ISI Task Force on Capacity Building delivered an interesting and horizon-widening talk on the topic "R's grammar: the Tidyverse!" during Online Workshop No. 3, Dr. Shahjahan Khan (Australia) gave an insightful talk on "Data Revolution and Statistics: Reshaping Society & Science" during Online Workshop No. 4 and, upon being invited again, Dr. David Stern talked about "The Place of Statistics in the Age of Data Science" in the fifth workshop of the series.

Four Workshops on Machine Learning

Subsequent to five successive workshops on Tidyverse, a series of four online workshops on Machine Learning (Levels 1, 2, 3 and 4) was organized by PISTAR on the afternoons of Sundays, 15th November 2020, 13th December 2020, 10th January 2021 and 7th February 2021 (Online Workshops No. 6 to No. 9). The workshops were conducted by Dr. Reza Arabi Belaghi (Resource Person) and Dr. Alireza Safariyan (Co-Resource Person) from Iran. The workshops attracted between 24 and 36 national and between 7 and 9 international participants. The resource persons delivered the contents of the workshop through PowerPoint slides and through live demonstrations of Machine Learning in R-Studio. Topics such as Decision Trees, Random Forests, Boosting Methods, Linear Support Vector Machine (SVN) Linear Regression, Multiple Regression, Types of Regression, Clustering, Principal Component Analysis, Factor Analysis, Artificial Neural Networking and Deep Learning were discussed. Feedback from the participants testified to the quality of the presentations made by the two resource persons and indicated the success of the workshop series.

The series of invited talks that was initiated with effect from Online Workshop No. 3 continued during the four workshops on Machine Learning. Dr. Donna LaLonde of the American Statistical Association (ASA) gave an interesting presentation on "ASA: Promoting the Practice and Profession of Statistics" during Online Workshop No. 6. Dr. Ayse Bilgin (Australia), President-Elect, International Association for Statistical Education (IASE) talked about Machine Learning Applications for Health Problems during Online Workshops No. 7 and No. 8. Dr. Marina Meila (USA) gave an insightful talk on Classic and Modern Data Clustering during Online Workshop No. 9.

Workshop on Pivot Tables for Data Science

PISTAR Online Workshop No. 10 was held from 2:00 to 5:00 pm PKT on Sunday, 10th April 2021. The topic of the Workshop was "Pivot Tables for Data Science" and the resource person was Mr. Muhammad Jawad Mirza from Pakistan. The workshop was attended by eight national and four international participants. The resource person conducted the workshop in two parts. Part 1 included 'Introduction and Importance', 'Data Handling Approaches', 'Lists and Tables in Excel' and 'Pivot Tables - a magical tool: Think, Drag and Drop, and it is Done' whereas Part 2 covered 'Pivot Tables from External Sources', 'If Data size is bigger than Excel Sheet then Where To Go?', 'Using Power Query for Pivot Tables', 'Combining, Importing, Cleaning Data' and 'Using Power Pivot for Big Data'. The contents of the workshop were delivered through live demonstrations in Excel.

The workshop session began with an Invited Talk by Prof. Olawale Awe from Nigeria. The title of the talk was 'Bayesian Dynamic Regression of Climate Data: A Change-Point Analysis'. Towards the end of his presentation, the invited speaker expressed a desire for *collaborations* with reference to climate research.

Five Workshops on Data Science Ideas for Statistical Consulting

PISTAR Online Workshop Nos. 11 to 15 were reserved for a workshop series entitled "Data Science Ideas for Aspiring Statistical Consultants". The five workshops were to be conducted by Prof. Roger Stern and Dr. David Stern from the United Kingdom on 23rd May, 13th June, 04th July, 25th July and 15th August, 2021 (all Sundays). Each workshop was to be held from 3:00 to 6:00 pm Pakistan Standard Time. The first workshop attracted 34 participants, nine of whom were from various countries of the world including one each from Argentina, Nigeria, Kenya, Germany, United Kingdom, Philippines and Australia, and two from Turkey. The workshop started with a presentation on 'Being Statistical Consultants' followed by participatory introduction of participants, introduction to Moodle, resource persons' remarks regarding the importance of adaptability, a presentation entitled 'Is Data Science unifying approaches to working with data?', a brief interactive session on dealing with data, and resource persons' remarks regarding the importance of problem-solving. Towards the end of the session, a number of participants expressed their opinions regarding the quality of the workshop. The positive remarks by participants from various countries were a source of satisfaction for both the resource persons and the workshop organizers.

It occurred to the author that the goal of achieving maximum benefit from the workshop series by Roger and David Stern will not be realized unless and until workshop participants engage in study and practice of materials posted by the resource persons on Moodle. This prompted the launch of relatively informal two-hour-long online *preparatory sessions* convened and conducted by the author herself. Four successful preparatory sessions prior to Workshop No. 3 paved the way for further such meetings, the benefits of which would include not only better understanding of data science but also a congenial atmosphere that would lead to long-term professional interactions.

FEEDBACK FROM PARTICIPANTS

In the end of May 2021, the author administered a questionnaire on participants who had attended a number of online workshops organized by PISTAR during the past one year. The questions as well as the responses to some of the important questions are presented below:

1. Approximately how many PISTAR Online Workshops have you attended?
2. How did you come to know about these workshops?
3. To what extent have these workshops turned out to be a learning experience for you?

- Participant from Argentina: *These workshops have opened a range of new horizons for me, thanks to this learning experience.*
- Participant from Iran: *I have learned a lot of applied materials from the sessions. However, I believe they are somehow quick and should be more focused.*
- Participant from Pakistan: *It is a good learning platform. But my concern is with installing R software...*
- Participant from Pakistan: *The workshops were horizon widening. It provides the platform to learn about the new data analytics techniques used around the globe.*

4. What aspects of the Workshops do you enjoy?

- Participant from Iran: *(a) Diverse learning materials, (b) Very organized workshops,(c) Friendly learning atmosphere, (d) Teachers are very good.*
- Participant from Kenya: *Well thought out presentations, representation of multinationals, active learning, and getting resources to relearn.*
- Participant from Pakistan: *Discussion by expert statisticians.*
- Participant from Egypt: *I like the international nature of the series; both the speakers and the audience. I also liked that you shared the Workshop material with the audience. The selections of speakers and topic were excellent. I sincerely wish to congratulate you on your efforts and success.*

- Participant from Australia: *The aspects I enjoyed is being able to see colleagues from other countries although we are unable to travel to places due to COVID-19.*

5. What aspects of the Workshops do you not enjoy?

- Participant from Kenya: *In some cases it's fast paced, and sometimes the language accent is hard to follow.*
- Participant from Egypt: *Sunday is a working day for us in Egypt. So, if it is possible to do them on Saturday I will be able to attend more of them.*
- Participant from Pakistan: *Sound quality is not good sometimes. R software is not easy to install and run.*
- Participant from Pakistan: *The workshops could be improved by allotting less time for the introduction of the trainers so that more time should be allocated for the learning process.*
- Participant from Australia: *Not enjoy part is the breaks are too little (i.e. 5 or 7 minutes ☹)*

6. Is there anything that makes these workshops different from Workshops organized by other organizations?

- Participant from Kenya: *I can't compare, but I found the experience to be a good return to my monetary investment. And possibly the personalized follow ups by the team makes PISTAR very special to me.*
- Participant from Pakistan: *Selection of resource persons.*
- Participant from Pakistan: *PISTAR online workshops are more regular (than those organized by other organizations), which is commendable.*
- Participant from Australia: *They are different in a sense that they are longer. They are better in a sense that there are more interactions in PISTAR workshops.*

7. What are your suggestions for improvement of PISTAR Workshops in the future?

- Participant from Argentina: *Keep up the enthusiasm.*
- Participant from Iran: *(a) Develop questioner with Likert scale (1-5) and ask questions there, (b) ask the needs of the participants for organizing next workshops, (c) ask the presenters to slow down for better understanding the materials, (d) give homeworks and projects to be fulfilled, and provide the certification after receiving the solved projects. This way will have better learning outcome.*
- Participant from Kenya: *Even cheaper training sessions may possibly be arranged for undergraduate students.*
- Participant from Pakistan: *The workshops should focus more on hands on training because it gives a chance to learn and practice the newly introduced concepts.*
- Participant from Egypt: *The only recommendation that I would suggest is to dedicate one (or even more workshops) to one topic only instead of having more than one topic per Workshop. This would give the speakers to discuss the details that the audience would like to know. Need more emphasis on "why" than on "how".*
- Participant from Pakistan: *Some theoretical elements should be included.*
- Participant from Australia: *Break out rooms in zoom worked really well in the last workshop. It forced people to turn their cameras on and talk to each other. I think future workshops could consider five minute or so small-group discussions to enable networking.*

Additional Remarks:

- Participant from Pakistan: *Thank you so much Mam for providing us this platform to enhance our professional and Educational skills.*
- Participant from Egypt: *The Online International Workshop-Series that you organized were great. I have enjoyed all four workshops that I attended.*

It is a matter of satisfaction for the author and her team that, by and large, the participants found the workshops beneficial and worth their while.

IMPLICATIONS FOR THEORY & PRACTICE

Statistics academicians/professionals living and working in developing countries should go ahead with activities such as those described in this paper without any hesitation or reservations as these are likely to bring together fellow statisticians from various parts of the world, particularly from

countries having similar socio-economic situations. Creation of a welcoming atmosphere in the online workshops is likely to generate feelings of bonding and comradeship, and may lead to long-term professional relationships, collaborative research-work, joint data science projects and the like.

CONCLUDING REMARKS

Whereas in the West, webinars have become a matter of routine, even today, they are not very common in the developing countries. Conduct of online training workshops by a team of statisticians based in a developing country on topics such as data science and machine learning in such a way that the workshops are able to attract participants as well as expert statisticians from various countries of the world, both developed and developing, is special indeed. Particularly, in the time of the COVID-19 pandemic when nations have had to experience lockdown for months on end, the significance of this type of a healthy activity cannot be overemphasized.

ACKNOWLEDGMENT

The author would like to thank her student Aqsa Abid for the valuable assistance provided by her in writing this paper.

REFERENCES

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D.T. & Ward, M.D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician, Vol. 69*(No. 4, Special Issue on Statistics and the Undergraduate Curriculum), pp. 343-353.

Theobold, A.S., Hancock, S.A. & Mannheimer, S. (2021). Designing Data Science Workshops for Data-Intensive Environmental Science Research, *Journal of Statistics and Data Science Education, Vol. 29*(sup1), S83-S94.

CODATA International Training Workshop in Big Data for Science, for Researchers from Emerging and Developing Countries, June 2014. *https://codata.org/events/training-workshops/big-data-science-training-workshop-beijing/*

Treacy, S (2019). Can AI boost Africa's development? *https://twas.org/article/can-ai-boost-africas-development*