Anmol Bansal et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 15-18

Tools Used in Data Analysis: A Comparative Study

Anmol Bansal[1] and Dr. Satyajee Srivastava[2]

[1]Galgotias University, Greater Noida, Uttar Pradesh, India
[2]Assistant Professor, Galgotias University, Greater Noida, Uttar Pradesh, India

Abstract. **There have been many statistical tools in the market that are used in data analysis. Each tool is popular in its own feature like cost, visualization, packages, statistics, etc. In this paper, some of the top tools have been taken that are used in data analysis then comparison is done based on some important factors to know the best tool in data science field. The factors that taken into consideration are cost, data handling capabilities, graphical capabilities, big data, etc. Comparison are made on the basis of rating of 1 to 5 and by our own experiences on various data analysis tools. This paper consider the facts that researchers has said in their published papers.**

Keywords: **Data Analysis, Statistical Tools, Data Science Tools, WEKA, SAS, SPSS, R, Python.**

I. INTRODUCTION

The future of almost every sector whether it is business, IT, Medical Sciences, Forensics and many more are revolves around the ability to make predictions and discover patterns in data. Data science is the centre of this revolution. It includes data mining, machine learning, and statistical methodologies to extract knowledge and leverage predictions from data [1]. Data science conceive the "sexiest job of the 21st century" [2]. Data Science is widely used in almost all fields like Forensics, Biology, Botany, Commerce, Medicine, Education, Physics, Chemistry, Bio-Technology, Psychology, Zoology etc.. While doing research in the above fields, the researchers should have knowledge about some of analytical tools which helps them to create the graphs/plots and better conclusions. The most well-known analytical tools are the R, Python, SPSS, WEKA, SAS, etc. [3].

To acquire the methods of data science like predictions, extraction of knowledge, finding the patterns, etc. have several languages/tools in the market. Some are open source whereas some are commercial software. Some are more popular whereas some are less popular. This paper consider the commonly used data analysis tools and compare them with each other to find best.

The value of statistics lies with organizing, transforming and simplifying data. The most well -known Statistical tools are the mean, the arithmetical average of numbers, median and mode, Range, dispersion , standard deviation, inter quartile range, coefficient of variation, etc. Equally important is that the results of these statistical procedures are recorded and can be retrieved. But to sort through all this information, you need the right statistical data analysis tools. This paper gave a brief comparison on Statistical or Statistical tools used in data analytics [3].

II. BACKGROUND

There are 8 factors are used for comparing statistical tools i.e Price, Ease of Learning, Data handling Capabilities, Graphical Capabilities, Usability, Jobs, Customer service and Community and Big data [4].

2.1 Price

There are several data analysis tools that are paid and some are open source. To study whether all features are included or not in case of open source.

2.2 Ease of Learning

Every tool have its own IDE or its own syntax or methods that makes a language easy or difficult to learn.

2.3 Data Handling Capabilities

Data handling describe that how long the data should be kept, and when, how, and who should handle data for storage, sharing, archival, retrieval and disposal purposes [5].

2.4 Graphical Capabilities

A graph can present data in a simple and clear way. The most convenient and popular way of describing data is using graphical presentation. It can illustrate the important aspects of the data to better analysis and presentation of the data.

2.5 Usability

Usability is the degree to which a tool can be used by specified users to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use.

2.6 Jobs

Mostly users have their first priority for learning any tool or language is to get a good job. Every tool have limited jobs depends on the companies that are using that tool.

2.7 Customer service and Community

Community service is important because volunteering teaches people of all ages and backgrounds compassion and understanding. There are opportunities to improve and leave your mark on your global and local community. Volunteering and putting on service events can be used as a way to advocate for causes that you are personally passionate about. Volunteering can also be the avenue to explore areas that you express interest. Not only is community service fun and rewarding, but volunteering looks great on a resume or college application. Sometimes community service is even required for high school graduation [25].

2.8 Big data

Big Data is the large amounts of data that is collected with time and are difficult to analyse using the traditional database system tools [26]. The most important question that arises how the data is stored and processed; most of which is raw, semi structured, and may be unstructured data. Big data platforms are categorized depending on how to store and process them in a scalable, fault tolerant and efficient manner [27]. For handling raw, semi structured and unstructured data, non-relational techniques can be used to produce statistics from big data, or to pre-process big information before it is combined into a data warehouse. Big Data analytics can help to gain ideas and make better choices [28].It is the one of the most important feature in statistical tools that it should handle the big data.

III. BRIEF OVERVIEW OF ANALYTICAL TOOLS

All data analysis tools have in common is the countless debates about why their programming language of choice is better, more advanced, faster, holier etc. In today's data science community, it seems as if these discussions are boundless with advocates of SAS, SPSS, R, Python, Julia, etc. battling and challenging each other on every online medium on the best statistical programming language [6]. This overview gives us the brief knowledge about the tools so that afterwards comparison of these tools can be done and find out which is best among them. The top five statistical tools available for data analytics are briefed as below.

3.1 R

Revolution is a free software programming language and software environment for statistical computing, data analytics, data mining and graphics [7].It consists nearly hundreds of extra "packages" freely available, which provide all sorts of data mining, machine learning and statistical techniques and much more. It has the ability to make a working machine learning program in just 40 lines of code [8].

It focuses on better, user friendly data analysis, statistics and graphical models. R has been used primarily in academics and research. The IDE used in R is RStudio. The popular packages in R are dplyr, ggplot2, readr, stringr, zoo, lattice, caret, etc. CRAN stands for the Comprehensive R Achieve Network. It is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R [9].R can be engaged for statistical and analytics tasks including but not restricted to clustering, regression, time series analysis, text mining, and statistical modelling. R is considered an interpreted language more so than an environment. R supports big data processing with RHadoop. RHadoop connects R to Hadoop environments and runs R programs across Hadoop nodes and clusters [1]. Its broad community provides many graphical utilities such as R Studio.

3.2 Python

Python is a widely used general-purpose, high-level programming language [10][11][12].This language can support different styles of programming including structural and object-oriented. Other styles can be used, too. Coding and debugging is easy to do in python, mainly because of "nice" syntax. The indentation of the code affects its meaning. Python is very flexible for doing something novel that has never been done before. Developers can also use it for scripting a website or other applications. Some features of python are [13]:

- **i.** Python is fast and powerful.
- **ii.** Python supports other technologies
- **iii.** Python is portable
- **iv.** Python is simple
- **v.** Python is open source

Users and admirers of Python—most especially those considered knowledgeable or experienced—are often referred to as Pythonists, Pythonistas, and Pythoneers [14][15].For data scientists with programming knowledge, there are a handful of tools that are particularly suited to the manipulation of data and engineering of features. Python is considered by many to be a particularly useful language for these purposes. In particular, engineering context dependent or temporal features is easier in Python than in Excel or Google Sheets. PyPi is the python package index: it is a repository of python software, consisting of libraries [16].

3.3 SPSS

SPSS stands for Statistical Package for the Social Sciences. In 1979 SPSS jeopardized the University of Chicago's status as a tax-exempt organization. SPSS was acquired by IBM in 2009 for US$1.2 billion [17]. SPSS was made to be easier to use then other statistical software like S-Plus, R or SAS [18]. SPSS is a great tool for non-statisticians since it has a user-friendly interface and easy to use drop down menus. Like Excel, SPSS is known beyond just the data science community. SPSS is primarily a statistical package, and offers a range of statistical tests, regression frameworks, correlations, and factor analyses. SPSS is a versatile package that allows many different types of analyses, data transformations, and forms of output - in short, it will more than adequately serve our purposes. SPSS is by far the easiest to learn. So if you only open a statistical program twice a month SPSS is the way to go [19].

3.4 WEKA

Weka stands for Waikato Environment for Knowledge Analysis, is licensed under the GNU general public license. It first released in 1997. Weka stems from the University of Waikato and is a collection of packages for machine learning and is Java based. It is widely adopted in academic and business and has an active community [20]. Weka provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support pre-processing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by weka [7].It is also suitable for developing new machine learning schemes [21].

3.5 SAS

SAS stands for Statistical Analysis System. The creator of SAS are Jim goodnight and Jim barr. It began at North Carolina State University as a project to analyze agricultural research. Demand for such software capabilities began to

grow, and SAS was founded in 1976 to help customers in all sorts of industries – from pharmaceutical companies and banks to academic and governmental entities. SAS – both the software and the company – thrived throughout the next few decades. Development of the software attained new heights in the industry because it could run across all platforms [22].SAS has strong data handling capabilities. It releases its software updates in a controlled environment, which makes them well tested. SAS is an expensive solution [23]. SAS licenses are priced based on the number of cores in the CPU of the machine the software will be installed on. The minimum configuration is a dual processor server (ie.2 cores). The most stripped down package will include the following Base SAS, SAS STAT, SAS Graph and SAS Access to ODBC. The price of 1 license for a dual core processor for the mentioned package is over 20 lakhs or about $40000.SAS may not take you to long to learn. To become really good at SAS you will need to work through a lot of specifics [24]. It can handle small, medium and large datasets. SAS use many different programs for visualization or Data mining. SAS is slow adapting new techniques.

IV. COMPARISON MATRIX

The comparison matrix formed with the help of Survey shows the tools and factors by which tool most preferable. Based on matrix R offers the best from rest of listed tools; however, each software tool has unique features and strengths. R is best because of its graphical capabilities are great, it is free of cost, it is quite easy to learn and many other features. SAS is close second, SAS is very costly that makes it at second. Tools like Python and SPSS are free of cost and their data handling capabilities are also quite good. However, career in SPSS that is less impressive. It offers very few jobs and salary is not much. Python have jobs and its salary is good comparing to SPSS but it is difficult to learn for new programmers. Similar to SPSS, WEKA's salary is less impressive and it's quite easy to learn. Big data handling capabilities are good comparing to SPSS and graphical capabilities are also good than SPSS. So on the basis of this matrix R comes first on our list than SAS & python both are good and handy tool but SAS is costly. SPSS is good than WEKA. This is hence features wise matrix and have rating 1 to 5 is given below as *table1:*

Table1: Comparison of tools with various factors (Rating 1 to 5)

S.No.	Basis of Comparison	R [4] [16] [19]	SPSS [19]	SAS [4] [19]	Python [4] [16]	WEKA
1	Cost	5	3.5	2.5	5	5
2	Ease of learning	4	4	4.5	3.5	4
3	Data handling capabilities	4	4.5	4	4	4
4	Graphical capabilities	5	3	4	4	3.5
5	Usability	4.5	4	4	4	4
6	Jobs	4.5	3	4	3.5	2.5
7	Customer service support and Community	4	3.5	4	3	3
8	Big data	4.5	2	4	3.5	4

V. CONCLUSION

This paper presents comparison between five analysis tools namely, Python, SPSS, R, SAS and WEKA. This comparison is based upon some factors like cost, ease of learning, data handling capabilities, graphical capabilities, usability, jobs, community and big data. Results based on the ratings that is given on *table1* concluded that R is taking the best analysis tool among all and benefited for Novel method prediction [27]. R have the better graphical representations of result, it is less cost, more usability and better big data handling capabilities. Python have good graphical capabilities but not best. WEKA and SPSS, these both tools are quite on similar ranking but the main difference is their costs. WEKA is totally free whereas there is some price for SPSS. In addition all the five tools are great and have their own good or bad factors but the best is R and what makes it best is its packages.

VI. REFERENCES

[1]. H. Wimmer & L. M. Powell (2015). "*A Comparison of Open Source Tools for Data Science".* Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA. 2167-1508: v8 n3651.

[2]. Davenport, T. H., & Patil, D. (2012). Data scientist. Harvard Business Review, 90, 7076.

[3]. Dr. K. J. Begum & Dr. A. Ahmed, "*The Importance of Statistical Tools in Research Work*". International Journal of Scientific and Innovative Mathematical Research (IJSIMR) Volume 3, Issue 12, December 2015, PP 50-58

[4]. Analytics Vidhya, https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

[5]. Office of Research Integrity, https://ori.hhs.gov/education/products/n_illinois_u/data management/dhtopic.html
[6]. Data Camp, https://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph#gs.9dggy7w
[7]. K. Rangra ,Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014.
[8]. A.Komathi, T.Ramya, M. Shanmugapriya, V. Sarmila, "A Novel Comparative Study on Data Mining Tools", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2016
[9]. CRAN-R, https://cran.r-project.org/
[10]. TIOBE Software Index (2011). "TIOBE Programming Community Index Python". 1
[11]. "Programming Language Trends - O'Reilly Radar". Radar.oreilly.com. 2 August 2006.
[12]. "The RedMonk Programming Language Rankings: January 2011 – tecosystems". Redmonk.com.
[13]. Masoud Nosrati, "Python: An appropriate language for real world programming", World Applied Programming, Vol (1), No
[14]. (2), June 2011.
[15]. Goodger, David. "Code Like a Pythonista: Idiomatic Python".
[16]. "How to think like a Pythonista".
[17]. Data Camp, https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis
[18]. http://www.fundinguniverse.com/company-histories/spss-inc-history/
[19]. Arun menachery, "INTRODUCTION TO SPSS".
[20]. Data Camp, https://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph#gs.N3o9EvM
[21]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.3671
[22]. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition,Morgan Kaufmann, San Francisco(2005).
[23]. SAS, https://www.sas.com/en_us/company-information.html
[24]. https://stats.stackexchange.com/questions/33780/r-vs-sas-why-is-sas-preferred-by-private-companies
[25]. http://analyticstraining.com/2012/pricing-for-analytical-tools-in-india/
[26]. https://www.21stcenturyleaders.org/why-is-community-service-important/
[27]. Marcus R. Wigan, Roger Clarke, "Big Data's BigUnintendedConsequences" Published by the IEEE Computer Society, pp 46-53
[28]. Srivastava S. (2017) Novel Method for Predicting Academic Performance of Students by Using Modified Particle Swarm Optimization (PSO). In: Panigrahi B., Hoda M., Sharma V., Goel S. (eds) Nature Inspired Computing. Advances in Intelligent Systems and Computing, vol 652. Springer, Singapore
[29]. Mr. Mahesh G Huddar, Manjula M Ramannavar, "A Survey on Big Data Analytical Tools", International Journal of Latest Trends in Engineering and Technology (IJLTET) Special Issue - IDEAS-2013.

