INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Big data analytics through R programming

Samyak Shah[1]

[1]Student of Computer Engineering Department
K. J. Somaiya College of Engineering
Mumbai, Pin no.400077
samyak.shah@somaiya.edu

Abstract: *Big Data Analytics is the subject that deals with applying various analytical techniques on the big data to know more about that data.Data was already been stored by mostly all organizations but was never put to use .But this trend of data analysis has brought a huge advantage to organizations by giving them knowledge relevant to their business. R programming is used for statistical computing and graphics. According to KDnuggets survey also, R language seems to be most used by people for data analyzing.*

Keywords*: Data Statistics, R programming, Data analytics, Big Data.*

1. INTRODUCTION

Over the last many years, collection and storing of data has become very inexpensive and easy. Many free computing tools available online can be used to deal with entire data from all various sources of science and humanities. This is the beginning of era of Big Data. People debate about the possible advantages and expenses of evaluating data from different social networking sites like Facebook, Pinterest, Path, etc. where many individuals put down digital footprints and leave data. The quantities of data available right now are undoubtedly vast, which is not the most significant characteristic of this new data system. As the kind of the relationality which big data has shown to other data, Big Data is noteworthy? Big Data is no longer just the domain of data scientists, new technologies have made it possible to generate, share, and organize data for people that include humanities and government, educational organisations and motivated individual. [8] Defining Big Data via the Three Vs Amount matters, but other important attributes of big data are data *diversity* and data *speed*. The three Vs of big data (volume, variety, and velocity) comprise an inclusive definition, and they prove that big data is not only about data size.

Data size: It's obvious that data size is the chief characteristic of big data. Mostly big data is defined in terabytes or even petabytes. Yet, big data can also be counted by including accounts, transactions, documents, etc

Data type variety: One of the things that make big data huge is because it's coming from variety of sources. Some of them are Web sources like weblogs, social media, etc. The difference from the past is that too many users are now analyzing big data rather than just storing it and also the way doing it has become more sophisticated.

Data feed velocity: Big data can be described by its speed. You may prefer to think of it as the regularity of generation of data or the regularity of delivery of data. With data coming to you unremittingly in real time, data size gets enormous in a hurry. The challenging part is that the analytics applied on the big data have to make sense of the data and perhaps take action, all this in real time. [4]

2. R PROGRAMMING

R is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. The source code for the R software environment is written primarily in C, FORTRAN, and R. certain aspects discussed here are: [5]

2.2 Arithmetic with R:

R is like a simple calculator in its primary form. For e.g.,

- 3+5
  [1] 8
- 0-4
  [1] -4

Division and multiplication are performed similarly as above. For modulo,

- 762%%8
  [1] 2

2.3 Vectors:

For data analysis, we will make extensive use of vectors. Vector is used to store data in one-dimension and they can be numeric, Boolean as well as character vectors. Example for vector:

A vector <- c(x,y,z) ## A-vector is the name of the vector and x,y,z are its contents.

2.3.2 Naming each vector element:

For e.g.: rain=c(45,30,70,40,10,0,0).

Names(rain) = c("Monday", "Tuesday"," Wednesday", "Thursday" ,"Friday", "Saturday", "Sunday")

This shows that it rained 45mm on Monday, 30mm on Tuesday and so on.

2.3.3 Arithmetic's on Vectors: We have to remember that adding two vectors will mean element-wise vector sum. Hence, to get total sum of vectors we need to take sum of 1st and the 2nd vector.

For e.g., A = c(1,2,3) B=(4,5,6) ⇨ A + B = c (5,7,9)

Sum (A)+sum(B) = 21

2.3.4 Vector Selection:

rain_midweek = rain [2:5] ⇨ ##this depicts rain on Tuesday, Wednesday, Thursday and Friday.

Comparison operators can be applied to vectors also.

For e.g., max_rain = rain>25

max_rain

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Monday	Tuesday	Wednesday
TRUE	FALSE	TRUE
Thursday	Friday	Saturday	Sunday
TRUE	FALSE	FALSE	FALSE

2.4 Matrices:

Construction of matrix: For e.g., Matrix (1:16, by row=TRUE, nrow=4) ## 1:16 are the elements of the matrix , by row=True means the elements would be arranged row-wise and nrow indicates the no of rows.

Naming a matrix: By using row names (matrix_name) and col names (matrix_name) we can name the matrix.

Selecting one or elements from the matrix:

- Sample [1, 2] is used to select from the first row and the second column.
- Sample [1:3, 2:4] will select rows 1, 2, 3 and columns 2,3,4.

If one wants to select all elements of a column or a row, before or after the comma no number is needed:

- Sample [1] selects all elements of the first column.
- Sample [1] selects all elements of the first row.

Arithmetic with matrices: Sample_matrix1 *sample_ matrix2 will create a matrix where each element is the product of the corresponding elements in Sample_matrix1 and sample_ matrix2.

2.5 Plotting graphs:

plot() is a generic function: it does appropriate things for different types of input

scatterplot

- plot(employee$year, employee$salary)

boxplot

- plot(employee$rank, employee$salary)

stacked barplot

- plot(employee$field, employee$rank)

Let us consider an example of baseball player records for the major league. The elements of this database considered were hitter's name, no of times at bat, no of hits, no of home runs, no of runs, no of runs batted in, no of walks, no of years in major league. [3]

Plot(baseball$no.of.times.at.bat,baseball$no.ofwalks)

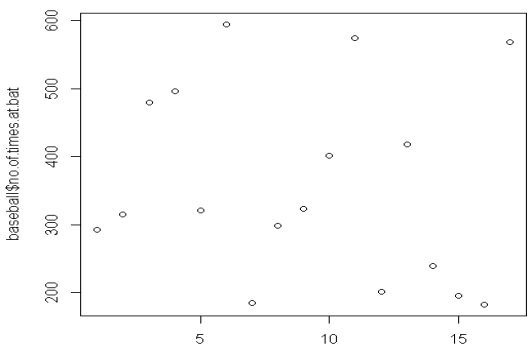

Figure 1: Graph plotted for number of times at bat v/s number of walks

Plot (baseball$no.of.hits,baseball$no.of.runs,type="h")

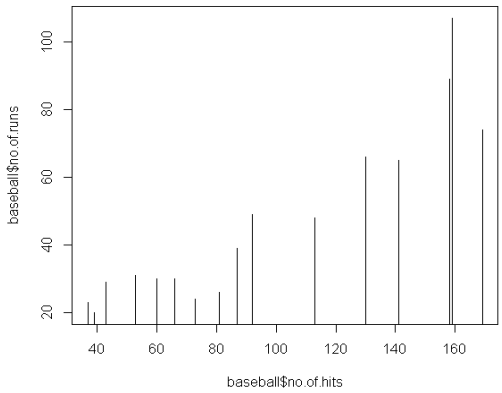

Figure 2: Type H graph: Number of runs v/s number of hits

Plot (baseball$no.of.hits,baseball$no.of.runs,type="n")

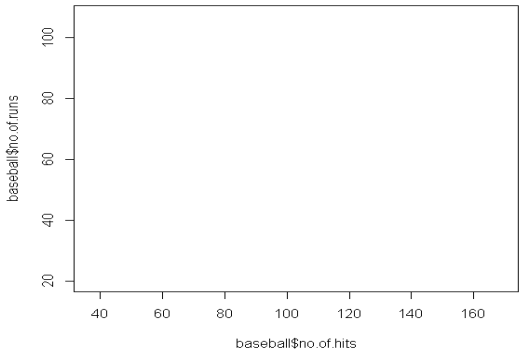

Figure 3: Type N graph: Number of runs v/s number of hits

Plot (baseball$no.of.hits,baseball$no.of.runs,type="l")

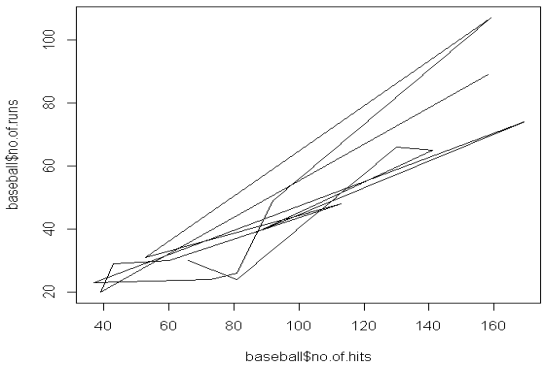

Figure 4: Type L graph: Number of runs v/s number of hits

type= controls how data are plotted. type="n" is not actually ineffective as it looks: it may get plotted for elements added latter.

boxplot(baseball$no.of.runs,baseball$no.of.years.in.major.league, horizontal=TRUE)

horizontal=TRUE plots a boxplot in sideways xlab and ylab are options for labels of x and y axis.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

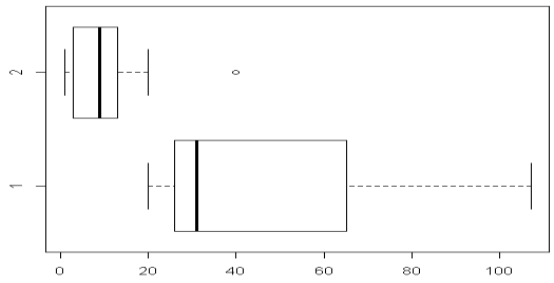

Figure 5: Box plot graph: Number of runs v/s number of years in major league

barplot(baseball$no.of.runs, beside=TRUE, legend=TRUE)

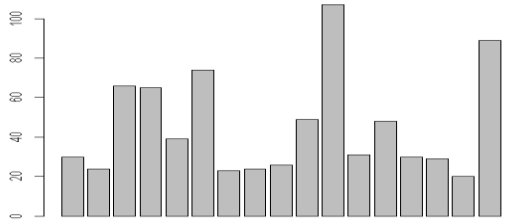

Figure 6: Bar plot graph: Number of runs

3. CONCLUSION

Exploring big data will lead to discovery of business facts an organization never knew. It shows how an organization has changed, also where the opportunities are for new clienteles or budget reductions.. The investigative and exploratory - oriented methods of analytics are appropriate for wisdom from big data. And these analytics methods benefit from the massive data samples produced from big data. But the main point is that big data is an extraordinary enterprise quality that virtues leverage, and analytics provides that leverage Big data should be considered as an opportunity and not a problem. [4]

REFERENCES

1. AMcAfee, E Brynjolfsson," Big Data: The Management Revolution",Harvard Business Review,pp1-9,October 2012.

2. Lavalle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N: "Big data, analytics and the path from insights to value". MIT Sloan Management Review 2011, 52:21-32.

3. Stat library CMU, Available, Statistics Community, http.

4. Philip Russom, "Big Data Analytics", Towi Best Practices Report, Fourth Quarter, 2011.

5. Ross Ihaka and Robert Gentleman," R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics, Vol. 5, No. 3 (Sep., 1996), pp. 299- 314.

6. Shmueli G., and Koppius, O. R. "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly (36:4), pp. 1165-1188. 2011.

7. Hsinchun Chen, Roger H. L. Chiang Veda C. Storey," Business Intelligence and Analytics: from Big Data to Big Impact", MIS Quarterly Vol. 36 No. 4, December 2012.

8. Edd Dumbill,"What is big data? An introduction to the big data landscape.",Available:O'reilly Radar,http://radar.oreilly.com/2012/01/what-is-big-data.html

9. IBM. "What Is Big Data?",Available: IBM Online, 2013. http://www.ibm.com/big-data/us/en/.

10. Shmueli G., and Koppius, O. R.. "Predictive Analytics in Information Systems Research," MIS Quarterly (35:3), pp.553-572, 2011.

11. Mattmann C, Garcia J, Krka I, Popescu D, Medvidovic N: The anatomy and physiology of the grid revisited. In WICSA/ECSA. London, UK: IEEE/IFIP; 2009. Yu J, Buyya R: taxonomy of workflow management systems for grid computing.